Mining photo-sharing websites to study ecological phenomena

Haipeng Zhang, Mohammed Korayem, David Crandall
School of Informatics and Computing
Indiana University, Bloomington, USA

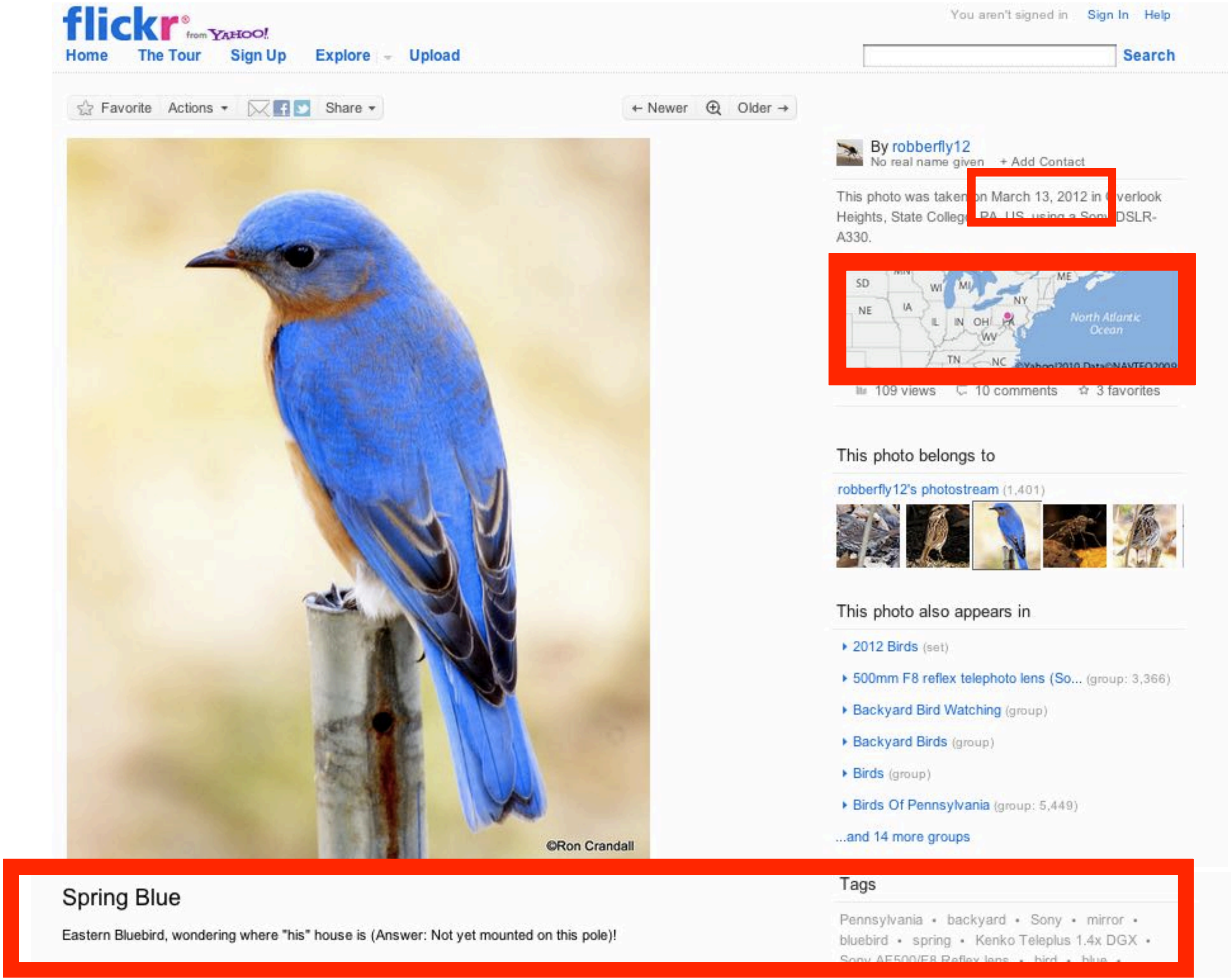
Gretchen LeBuhn
Department of Biology
San Francisco State University, San Francisco, USA

Social photo sharing websites

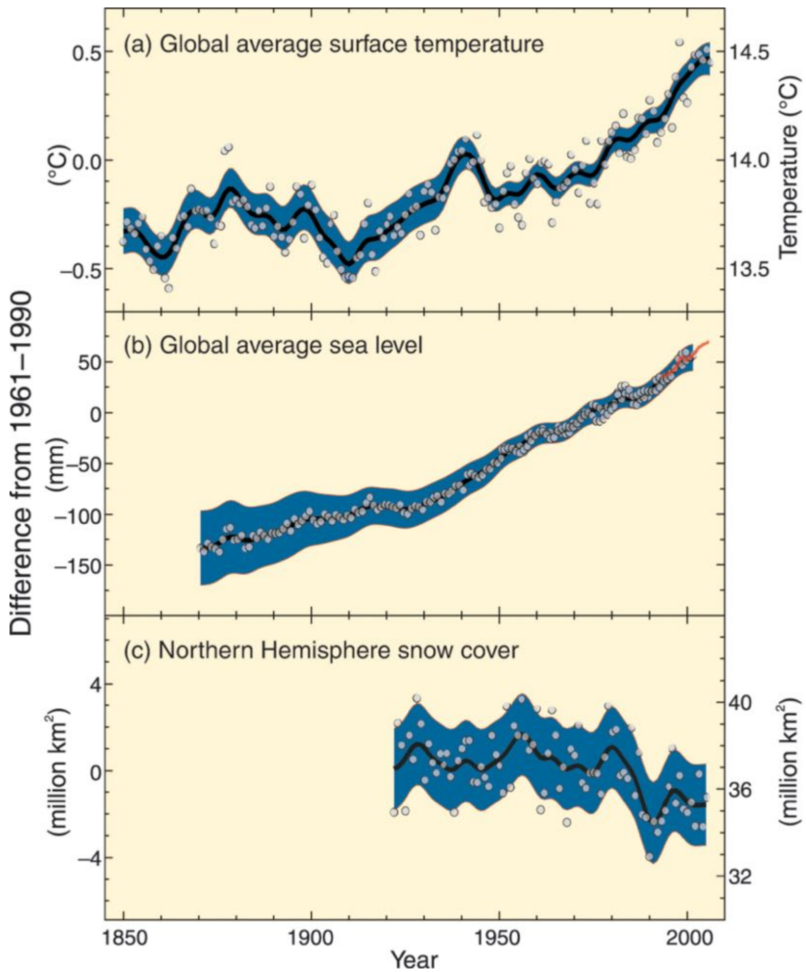
6+ **billion** photos

100+ **billion** photos


Snow
Cloud cover
Wildlife
Foliage
Flowers


flickr from YAHOO!
Home
The Tour
Sign Up
Explore
Upload
You aren't signed in
Sign In
Help
Search
Favorite
Actions
Share
← Newer
Older →
©Ron Crandall
By robberfly12
No real name given
+ Add Contact
This photo was taken on March 13, 2012 in Overlook Heights, State College, PA, US, using a Sony DSLR-A330.
SD
NE
IA
WI
MI
IL
IN
OH
WV
NY
ME
PA
TN
NC
North Atlantic Ocean
109 views
10 comments
3 favorites
This photo belongs to
robberfly12's photostream (1,401)
This photo also appears in
2012 Birds (set)
500mm F8 reflex telephoto lens (So... (group: 3,366)
Backyard Bird Watching (group)
Backyard Birds (group)
Birds (group)
Birds Of Pennsylvania (group: 5,449)
...and 14 more groups
Spring Blue
Eastern Bluebird, wondering where "his" house is (Answer: Not yet mounted on this pole)!
Tags
Pennsylvania • backyard • Sony • mirror •
bluebird • spring • Kenko Teleplus 1.4x DGX •
Sony AF500/F8 Reflex lens • bird • blue •

Need for ecological data



[IPCC2007]

How is nature changing due to global warming?

- **Plot-based studies:** Fine-grained information but only at a few locations, and labor-intensive
- **Aerial surveillance:** Continental-scale information, but only useful for some phenomena

Our paper

- Can we observe nature by mining photo websites?

- We study two phenomena: **snow** and **vegetation cover**
  - Estimate geo-temporal distributions at continental scale, using ~150 million photos from Flickr (via public API)
  - Analyze geo-tags, timestamps, text tags, visual content
  - Evaluate techniques for estimation in crowd-sourced data
  - Compare to data from weather stations and satellites

Related Work

- Crowd-sourced observational data, e.g.:
  - Estimating public mood from Twitter [Bollen11]
  - Predicting product sales from Flickr tags [Jin10]
  - Estimating spread of flu from search queries [Ginsberg09]
  - Monitoring forest fires from Twitter [DeLongueville09]

- Volunteer-based citizen science

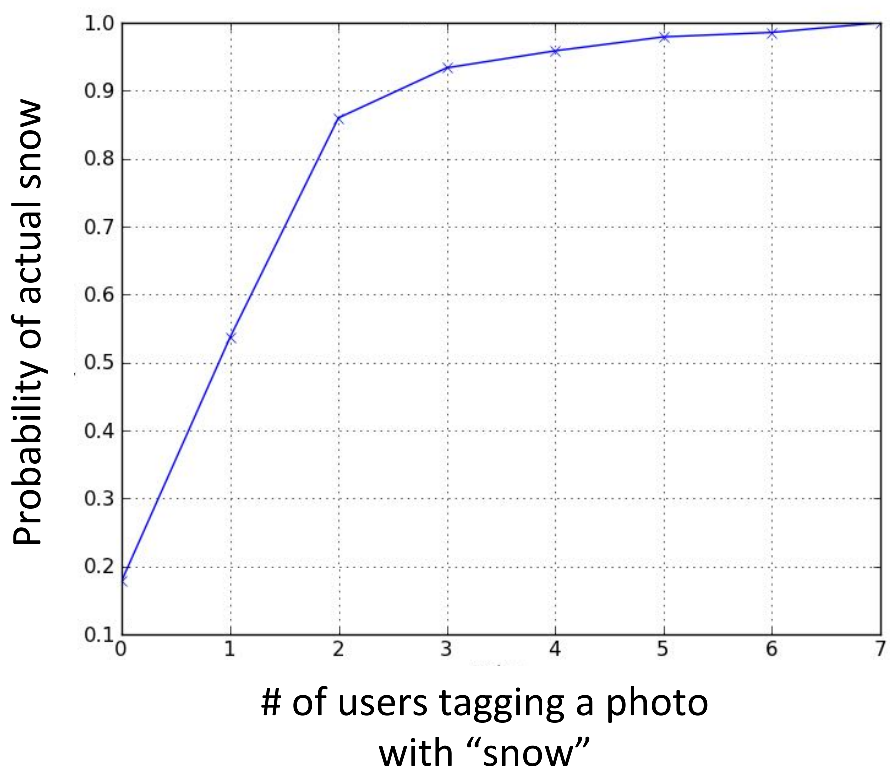
The Great Sunflower Project

Challenges

- Incorrect geotags and timestamps
- Difficult to recognize image content automatically
- Text tags helpful but noisy
  - Some tags are completely incorrect, others are misleading
- Dataset biases
  - Many more photos in cities than rural areas
  - People more likely to take photos of the unusual
- Misleading image data
  - e.g. zoos, ski slopes, synthetic images, etc.

Combining evidence

- Photos by different people are (almost) independent observations, with uncorrelated noise



A simple model

- Suppose we're interested in some object *X* (e.g. snow)
  - Specifically, whether *X* was present at a given time and place
  - Let *s* denote the event that a given user takes a picture of *X*
  - Assume *s* depends on presence of *X*:

$P(s \mid X)$ = probability of taking picture of X, given X was present

Could be factored into: Probability of seeing X, probability of taking photo, probability of uploading to Flickr, …

$P(s \mid \overline{X})$ = probability of taking picture of X, given X was <u>not</u> present

Bad timestamps or geotags, misleading image content, …

A simple model

- Suppose *m* users took photos of *X*, and *n* users did not
  - Using Bayes law,

$$P(X|s^m,\bar{s}^n) = \frac{P(s^m,\bar{s}^n|X)P(X)}{P(s^m,\bar{s}^n)} \qquad P(\bar{X}|s^m,\bar{s}^n) = \frac{P(s^m,\bar{s}^n|\bar{X})P(\bar{X})}{P(s^m,\bar{s}^n)}$$

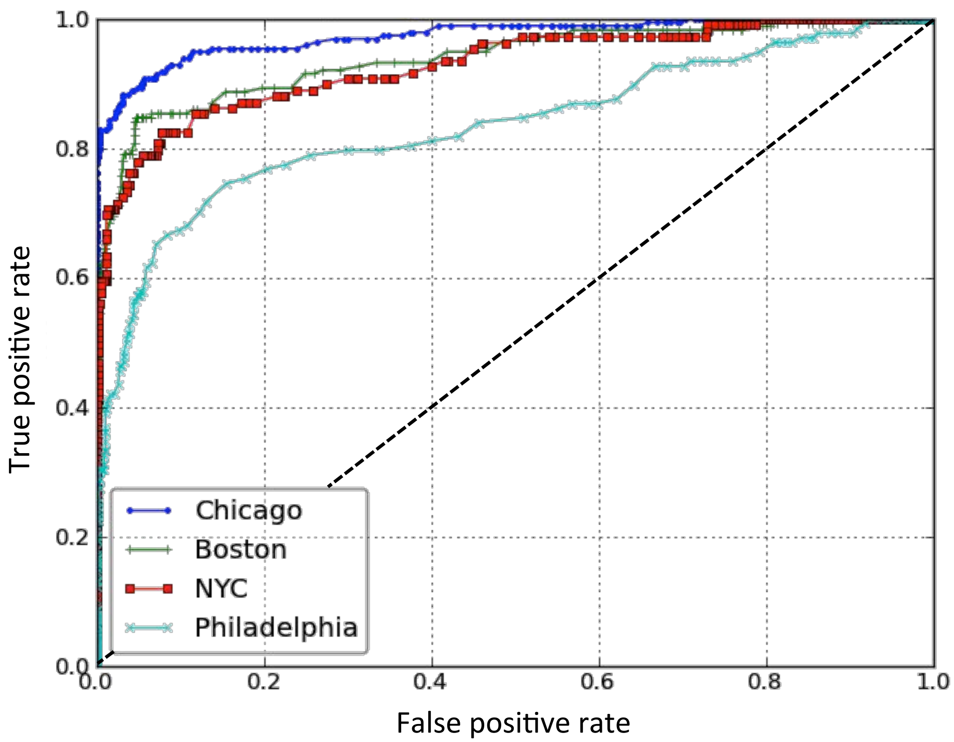
  - Assuming each user acts independently (conditioned on X),

$$\frac{P(X|s^m,\bar{s}^n)}{P(\bar{X}|s^m,\bar{s}^n)} = \frac{P(X)}{P(\bar{X})}\left(\frac{P(s|X)}{P(s|\bar{X})}\right)^m \left(\frac{1-P(s|X)}{1-P(s|\bar{X})}\right)^n$$

  - High or low ratio means high or low probability of X; ratio near 1 means low confidence either way

Snow estimation in cities

- Estimate daily snow cover (presence or absence)
  - Predict using Flickr photo tags, compare to ground truth from National Weather Service historical data
  - Estimate parameters on 2007-2008, test on 2009-2010.



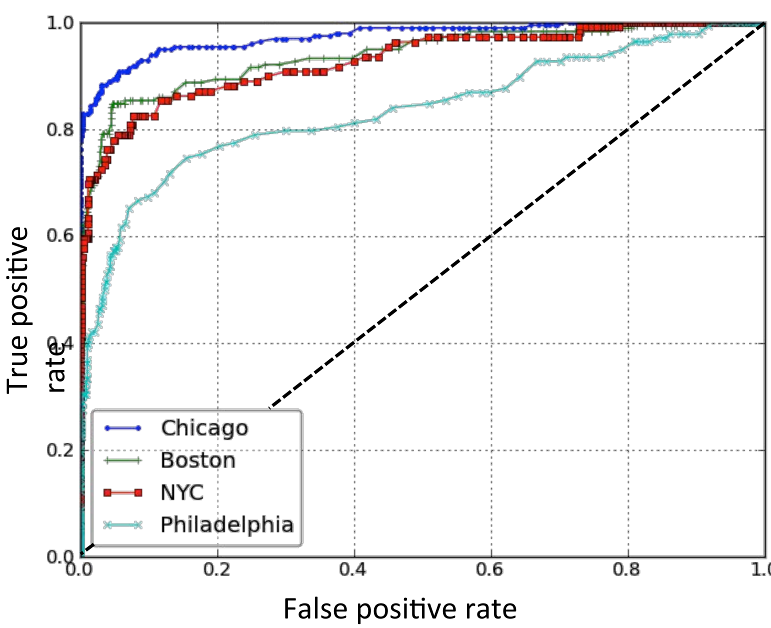
Tag set (hand-selected):
{snow, snowy, snowing, snowstorm}

Model parameters (estimated from training data):
P(s|snow) = 17.12%
P(s|no snow) = 0.14%

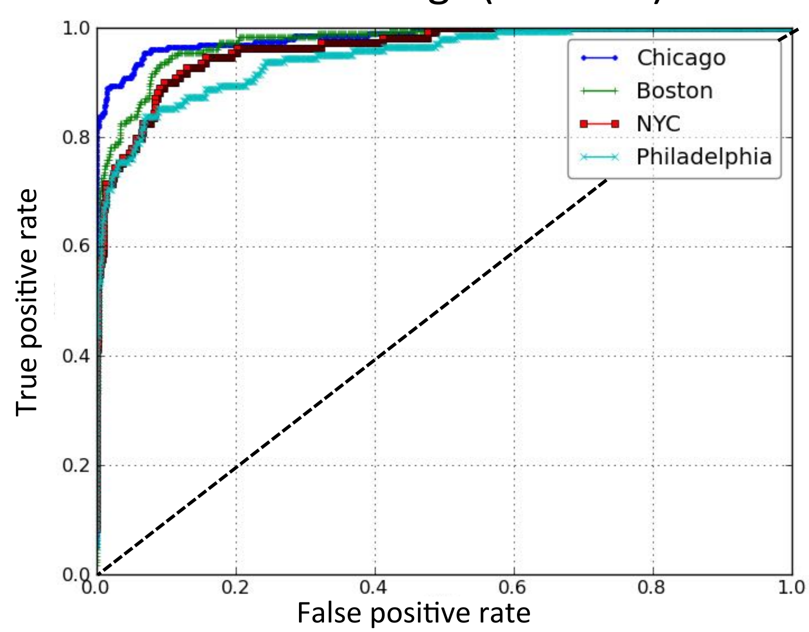
Learning relevant tags

- Find tags that correlate well with snow cover in GT
  - Feature vector for each day is histogram of number of people that used each tag; labels are snow/no snow from GT
  - Train on 2007-2008 data, test on 2009-2010 data
  - Increases classification accuracies significantly:

Hand-selected tags



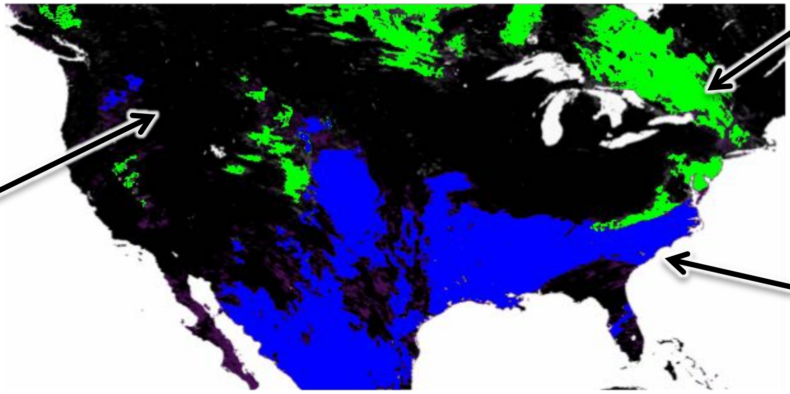
Learned tags (via SVM)

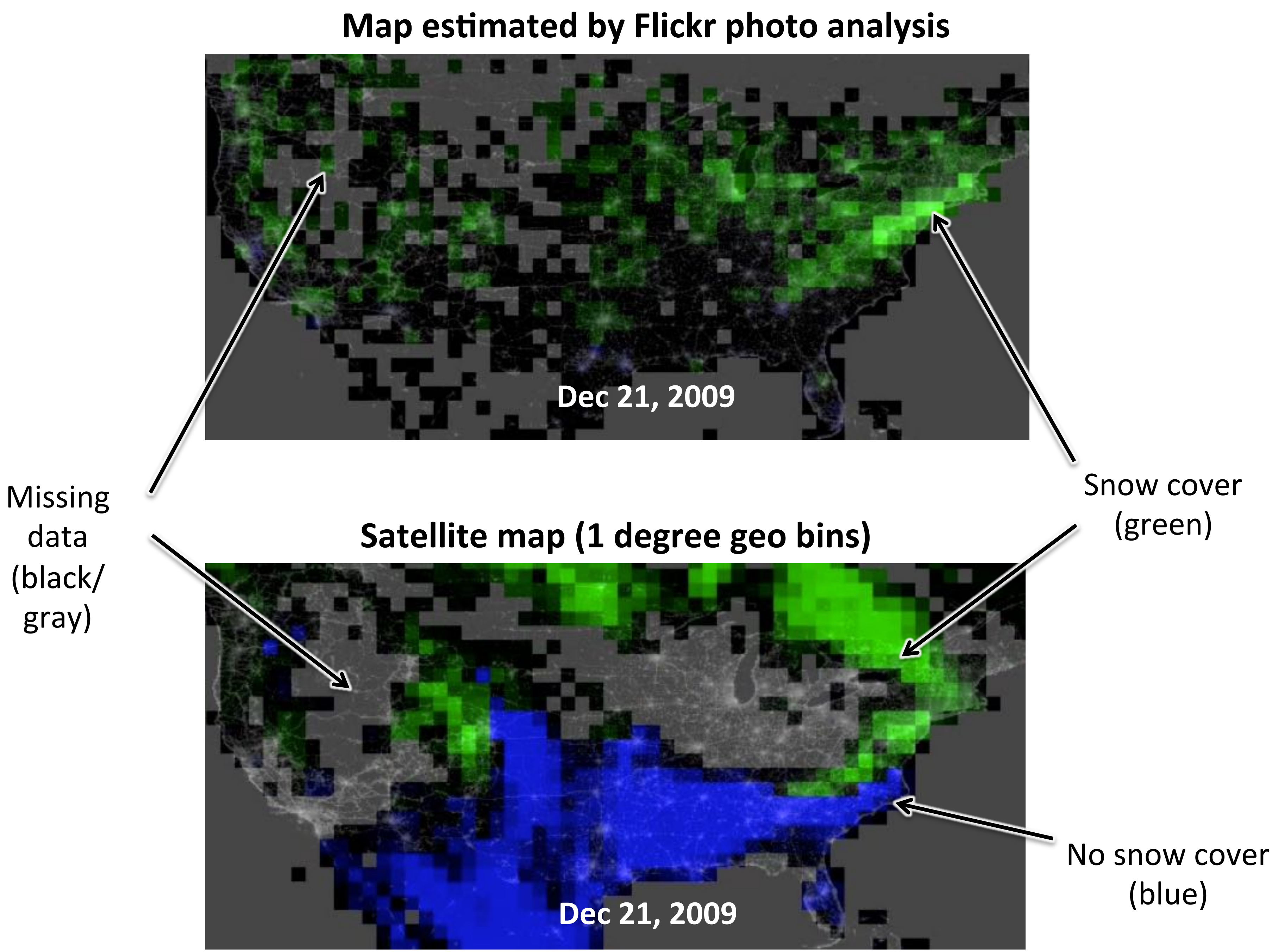


Continental-scale observation

- Estimate snow cover on each day at each place in North America
  - For each geographic bin of size 1° x 1°
  - Use ground truth data from Terra satellite

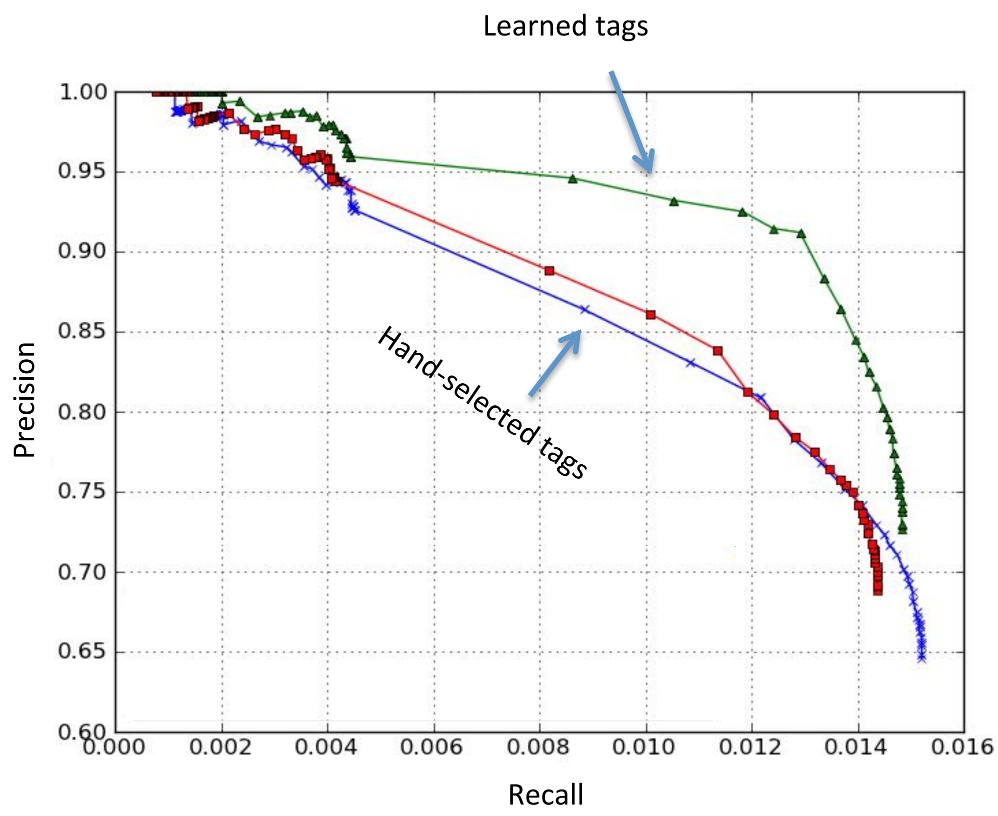
NASA Terra




Map estimated by Flickr photo analysis
Dec 21, 2009
Missing data (black/ gray)
Snow cover (green)
Satellite map (1 degree geo bins)
Dec 21, 2009
No snow cover (blue)

Continental-scale estimation

- Predict presence of snow on each day for each geo bin
  - ~35 million total decisions



Visual features

- Color and texture features similar to GIST [Torralba03]
  - Divide image into array of 4x4 cells; in each cell compute mean color value (in CIELab space) and mean gradient energy

Color channels

Image

Gradient magnitude

Visual features

- Color and texture features similar to GIST [Torralba03]
  - Divide image into array of 4x4 cells; in each cell compute mean color value (in CIELab space) and mean gradient energy

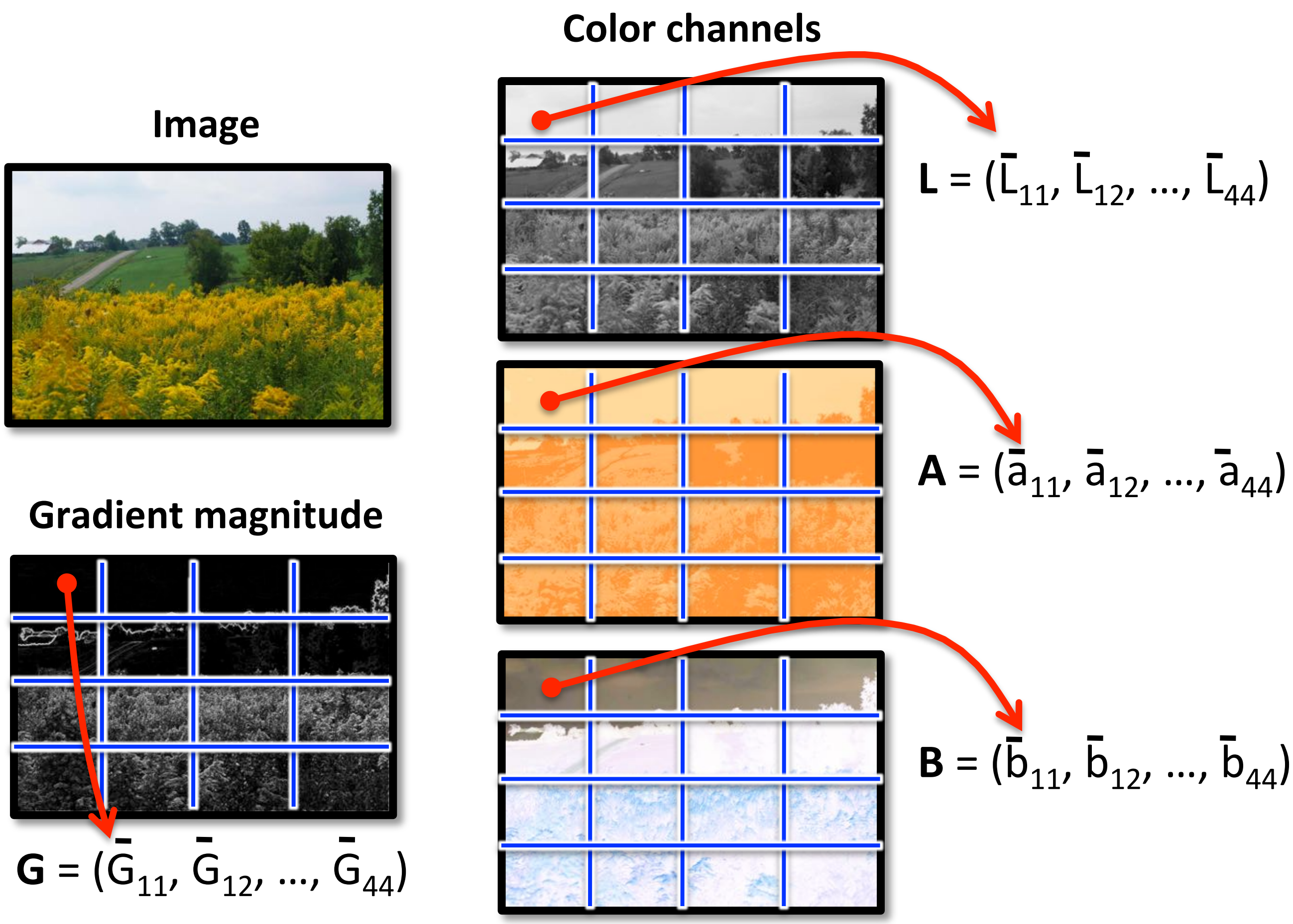


Image descriptor is concatenation of **L**, **A**, **B**, and **G** (64 dimensions); then learn SVM classifier

Classification with visual features

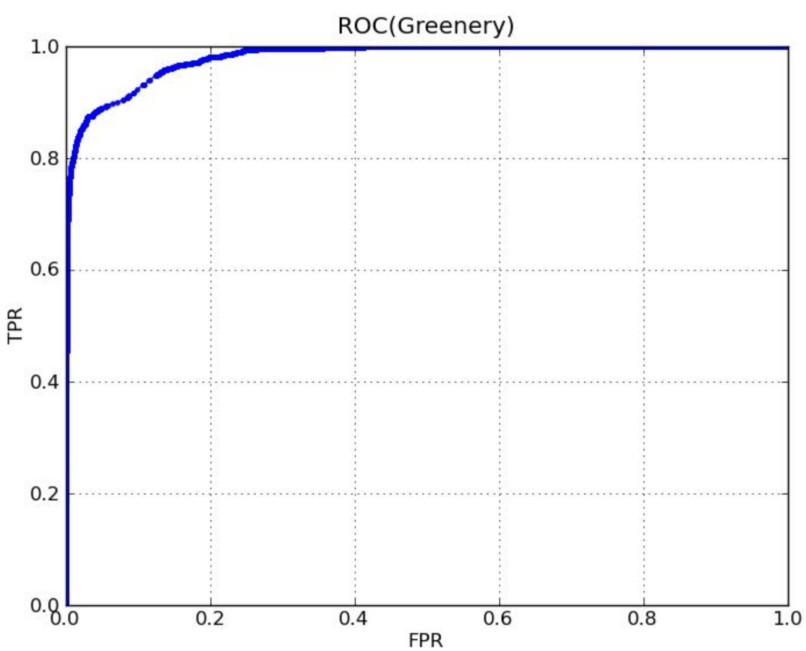
- Vision yields modest (~3%) improvement in precision

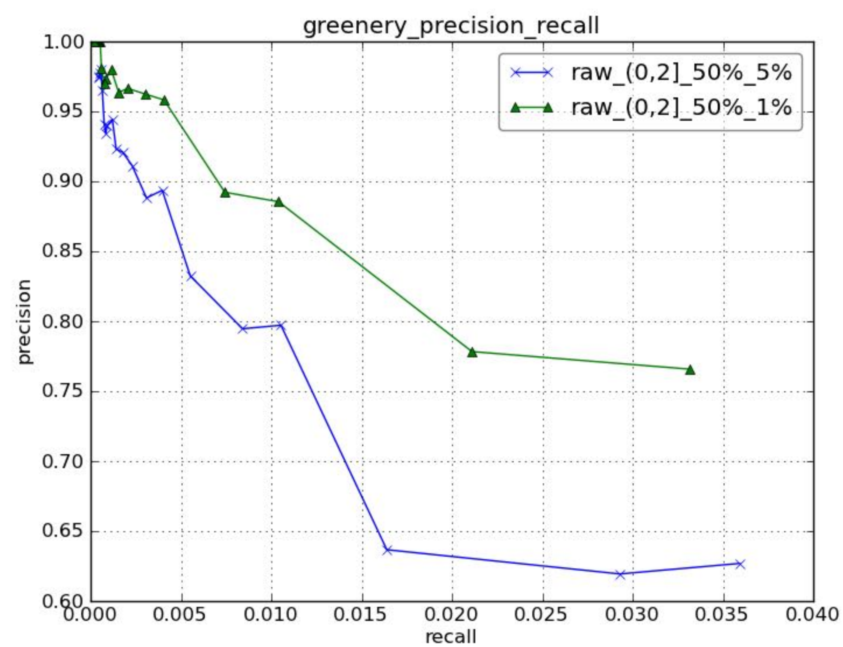
Correctly classified as non-snow:

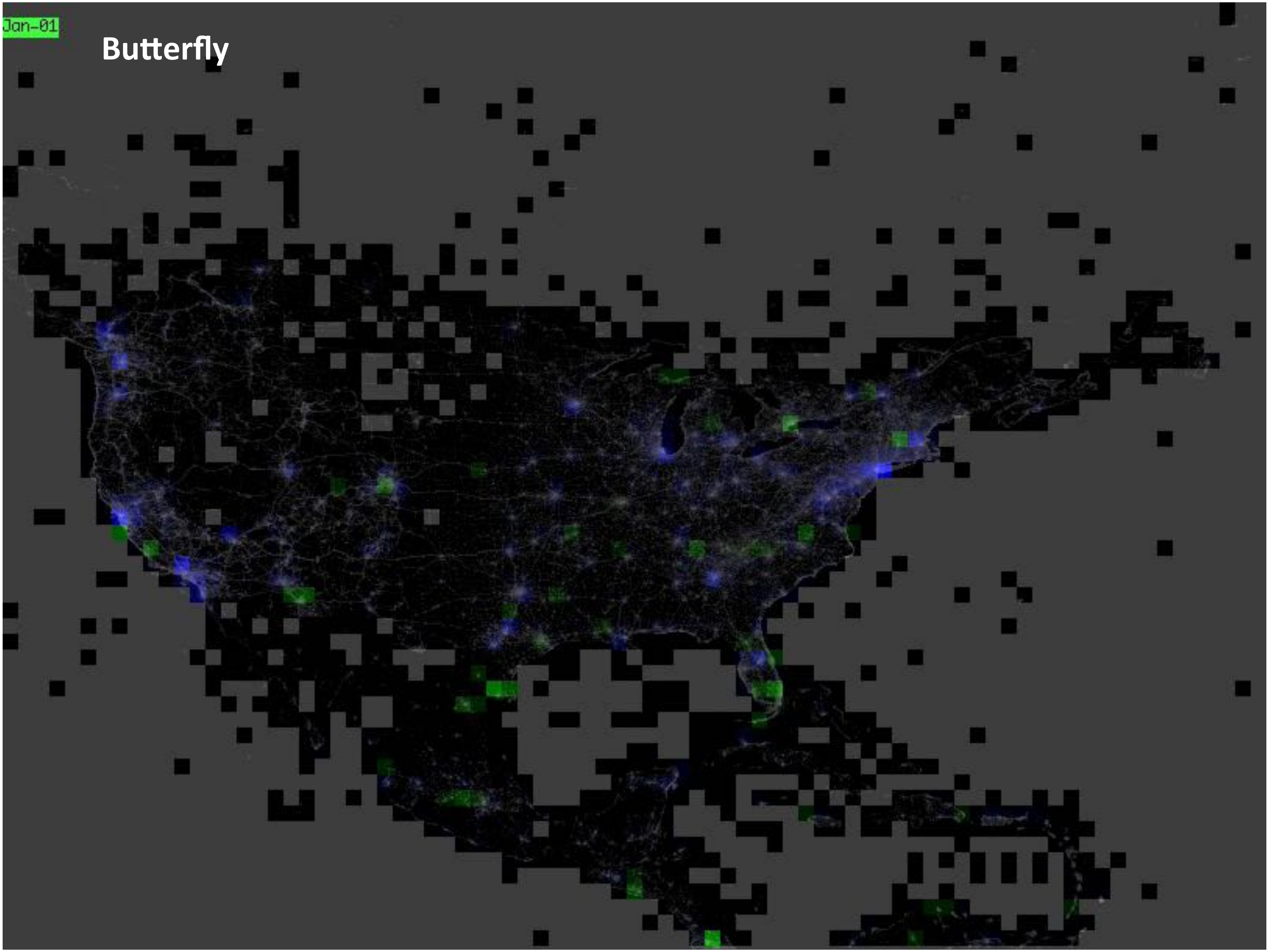
Incorrectly classified as snow:

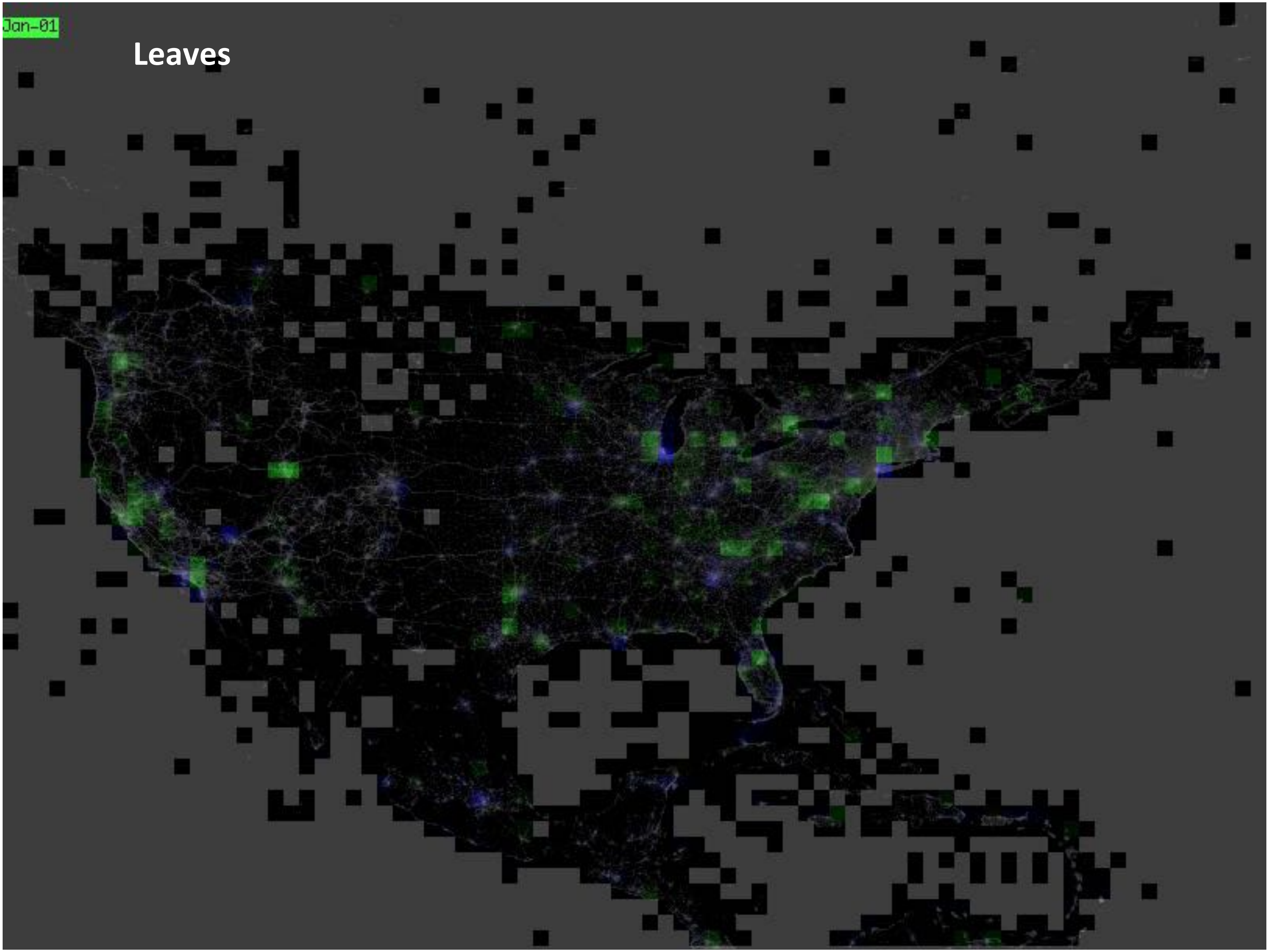
Estimating vegetation cover

- We also estimate vegetation cover (greenery index) on a continental scale
  - Again using ground truth data from Terra satellite






Jan-01
Butterfly

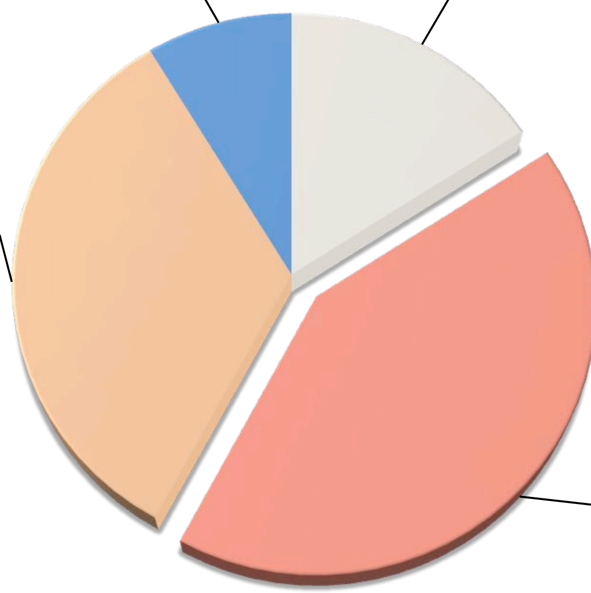

Jan-01
Leaves

Conclusion

- We propose to observe the natural world through mining public photos from online social sharing sites
  - Hundreds of billions of images available
  - But noise, bias, content extraction are challenges
- We study two phenomena, snow cover and vegetation
  - Using geo-tags, time stamps, text tags, and visual features
  - Use ground truth from satellites to measure estimation accuracy
- Future work
  - More sophisticated computer vision techniques
  - Combine our noisy, sparse data with biologists' noisy, sparse data
  - Study other phenomena, like migration patterns of wildlife, distributions of blooming flowers, etc.

Thank you!

False positives



(Total of N=1,855 photos)