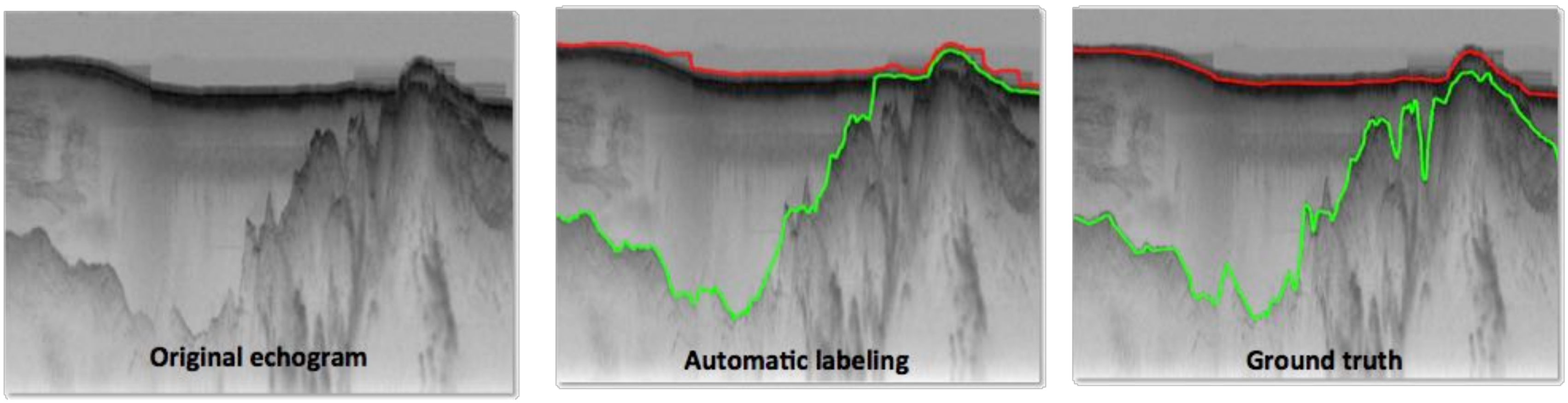
Computer vision meets high-performance computing

David Crandall
School of Informatics and Computing
Indiana University
Bloomington, Indiana

SPIDAL work

- Radar informatics (with CRESIS)



- High-performance abstractions for large-scale image analysis and computer vision
  - Find connections between computer vision on consumer photos, with medical imaging, GIS, etc.

facebook

flickr

photobucket

Panoramio
from Google

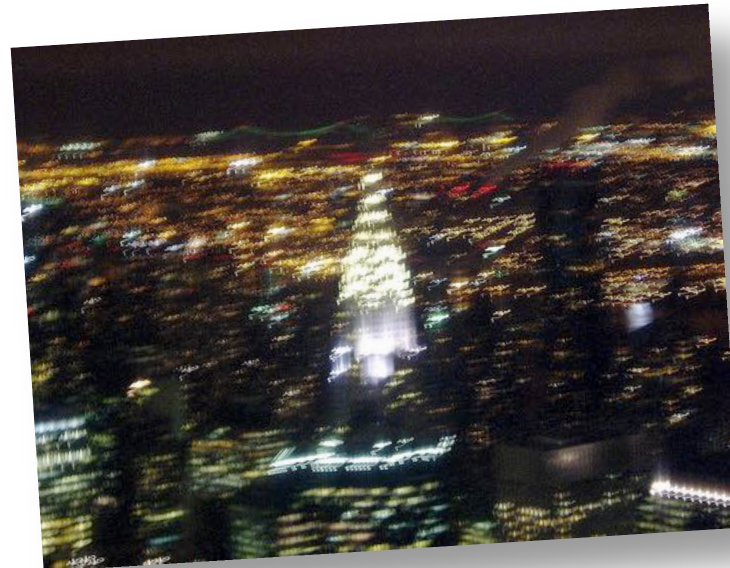
Instagram
Fast beautiful photo sharing

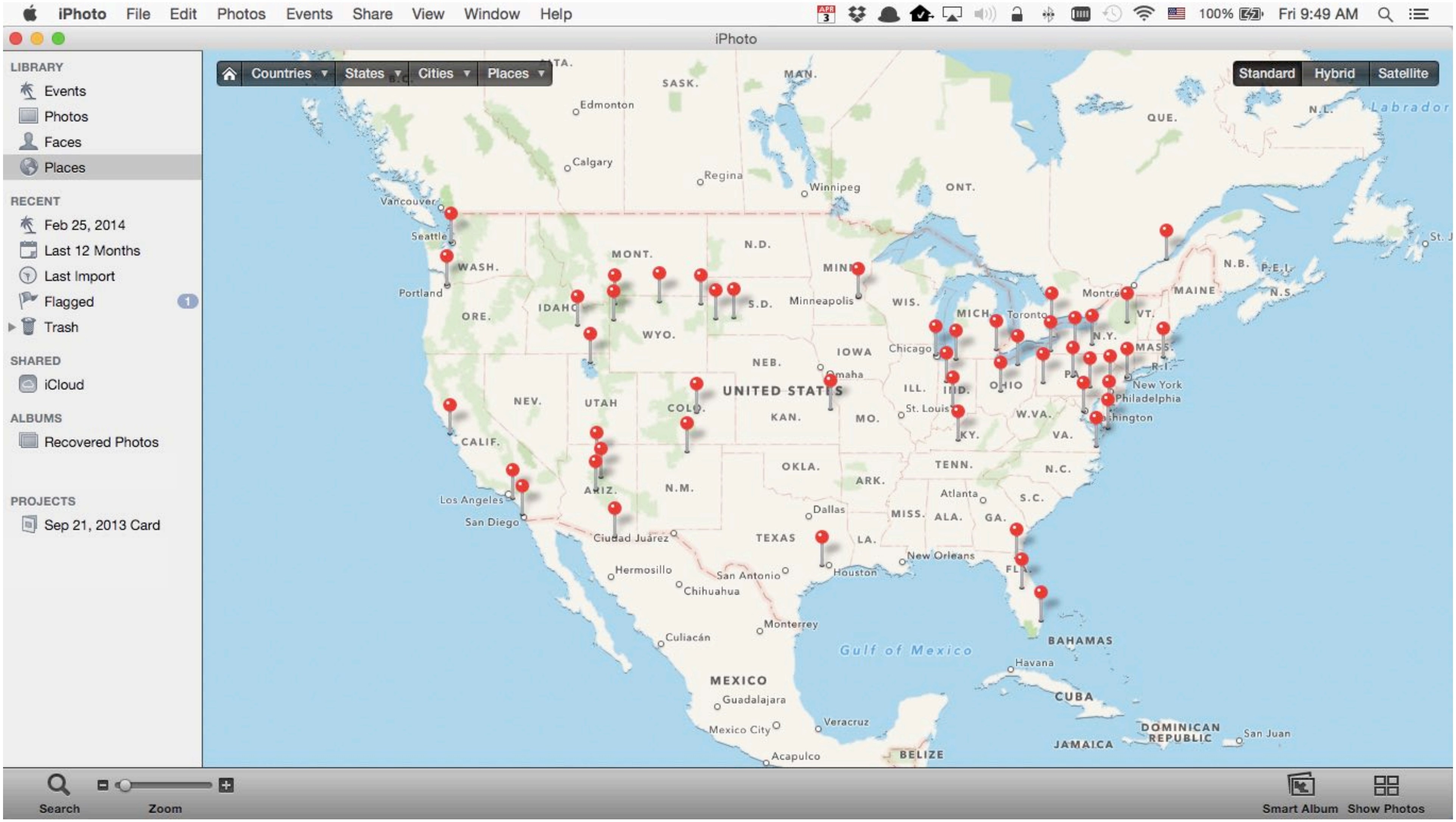
FOTOLOG

Computational patterns in vision

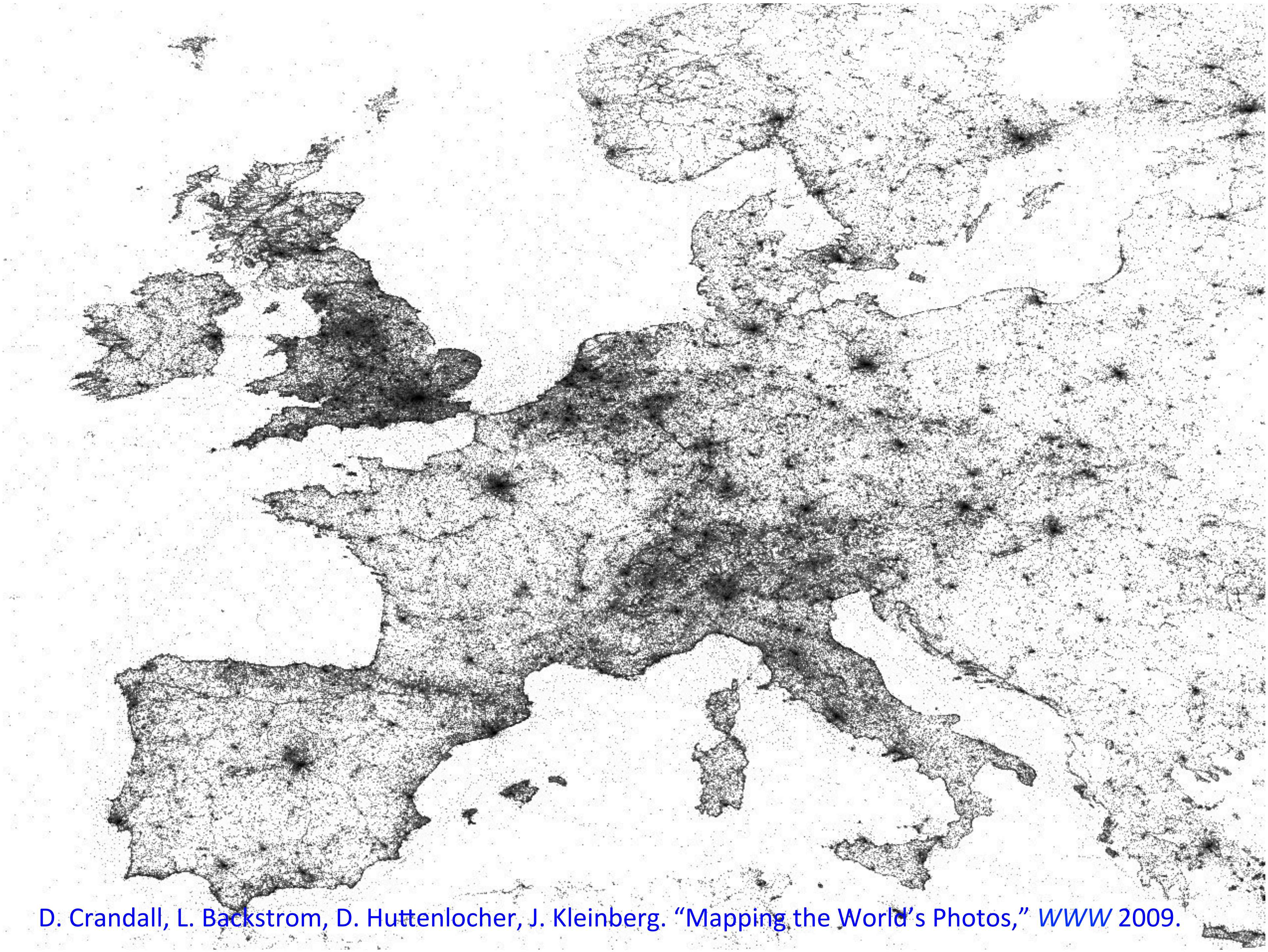
1. Single image tasks (e.g. feature extraction)
   - # of images may be large, but easily parallelizable
2. Image matching (e.g. recognition, clustering)
   - Evaluating distances between many high-dimensional vectors
3. Iterative algorithms (e.g. learning)
   - Few, but long-running iterations (e.g. k-means)
   - Lightweight, but many iterations (e.g. neural net backprop)
4. Inference on graphs (e.g. reconstruction, learning)
   - Small graphs with huge label spaces (e.g. pose detection)
   - Large graphs with small label spaces (e.g. resolving stereo)
   - Large graphs with large label spaces (e.g. reconstruction)

Visual geolocation: where was the photo taken?

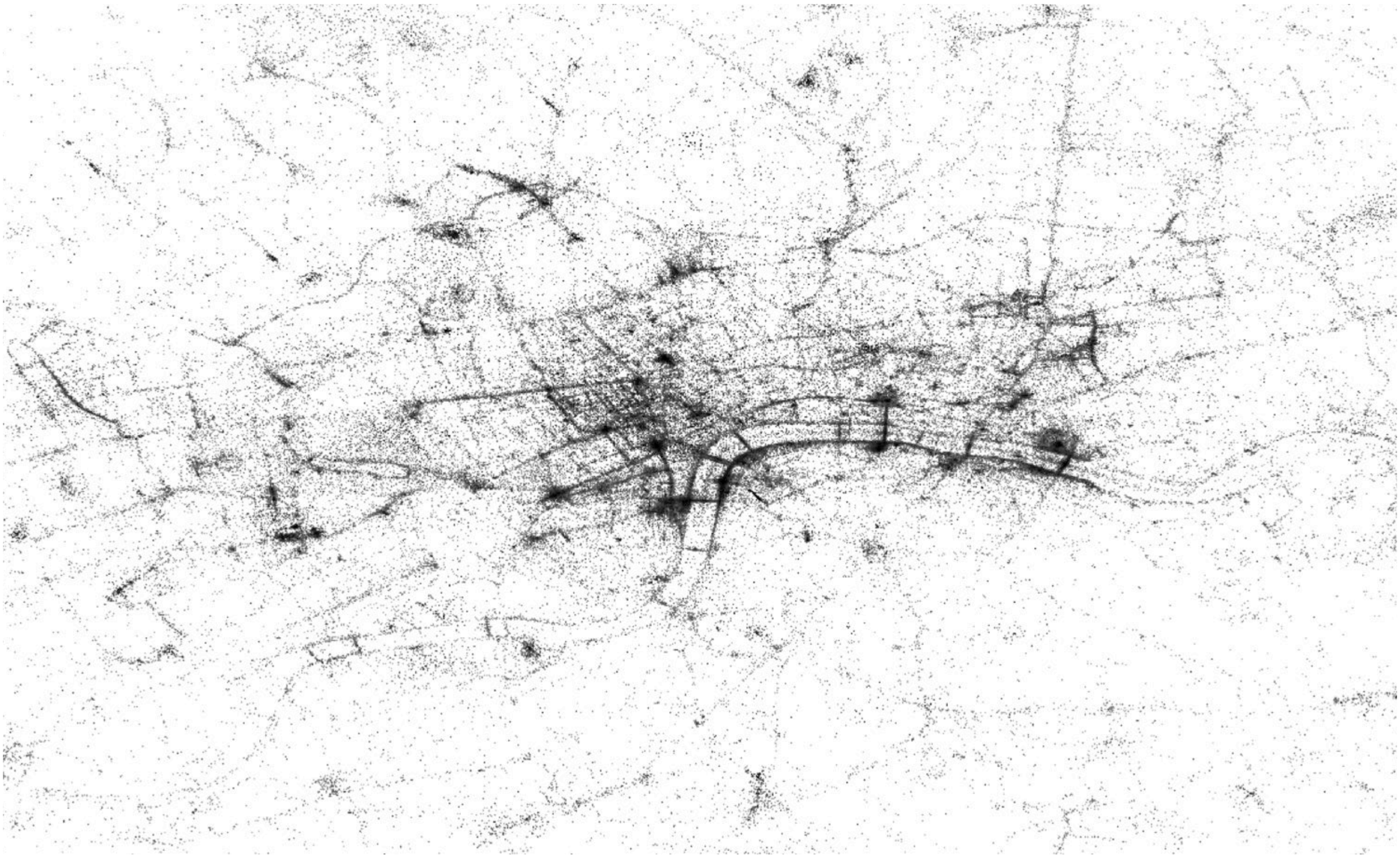



iPhoto File Edit Photos Events Share View Window Help
100%
Fri 9:49 AM
iPhoto
LIBRARY
Events
Photos
Faces
Places
RECENT
Feb 25, 2014
Last 12 Months
Last Import
Flagged
Trash
SHARED
iCloud
ALBUMS
Recovered Photos
PROJECTS
Sep 21, 2013 Card
Countries
States
Cities
Places
Standard
Hybrid
Satellite
UNITED STATES
MEXICO
Gulf of Mexico
BAHAMAS
CUBA
JAMAICA
DOMINICAN REPUBLIC
BELIZE
Search
Zoom
Smart Album
Show Photos


TEAM
30

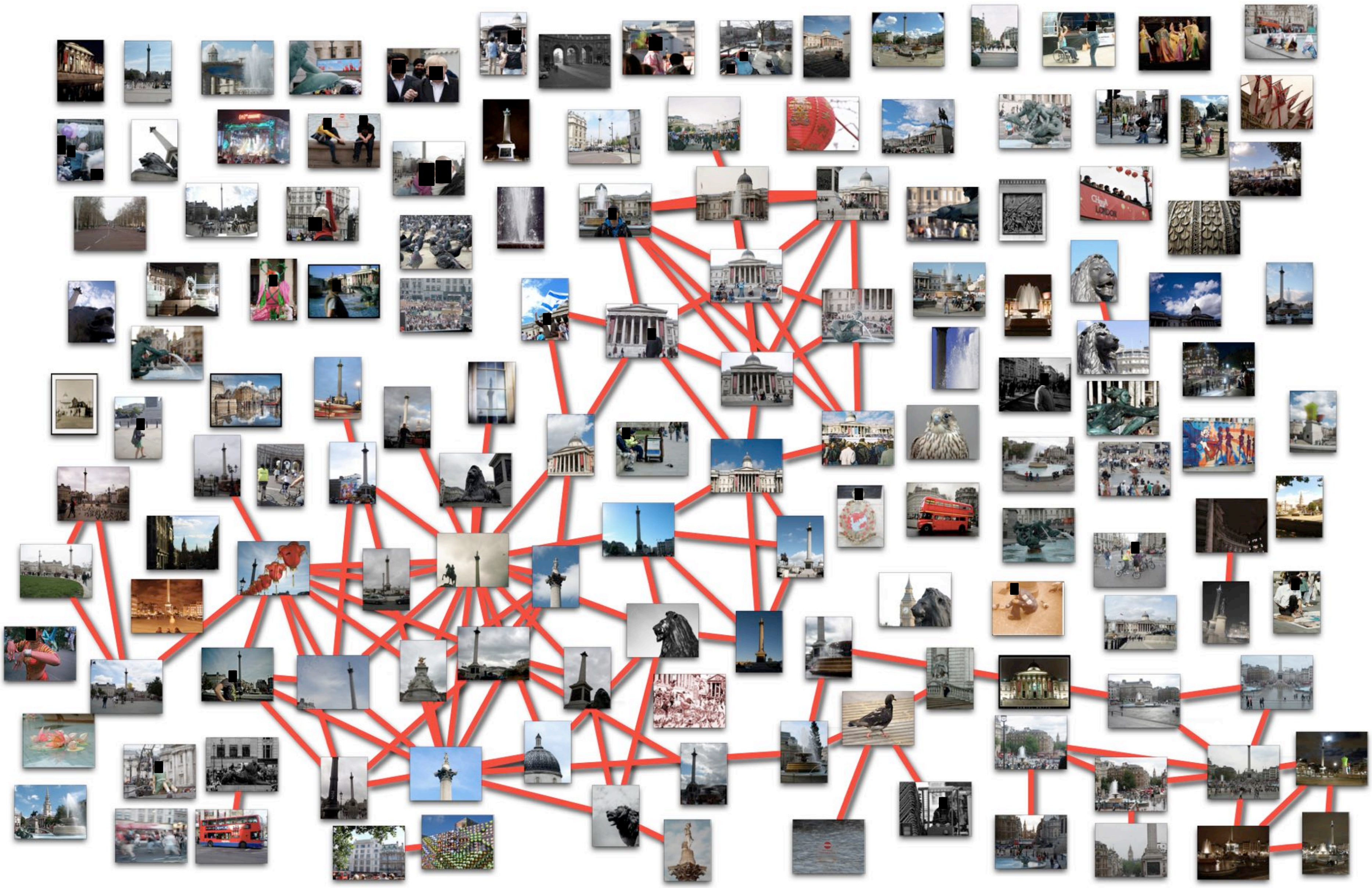


D. Crandall, L. Backstrom, D. Huttenlocher, J. Kleinberg. “Mapping the World’s Photos,” *WWW* 2009.



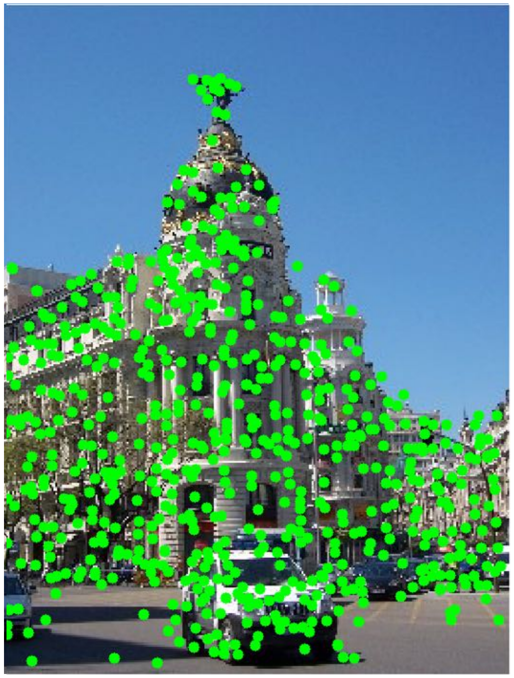
D. Crandall, L. Backstrom, D. Huttenlocher, J. Kleinberg. "Mapping the World's Photos," *WWW* 2009.

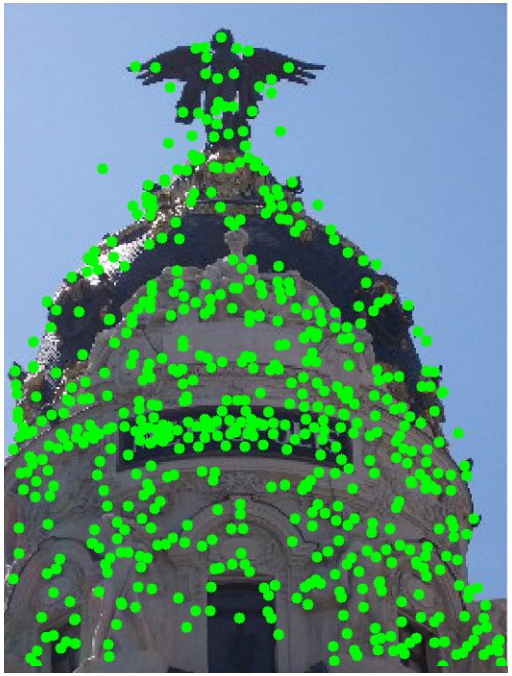
Image similarity graphs

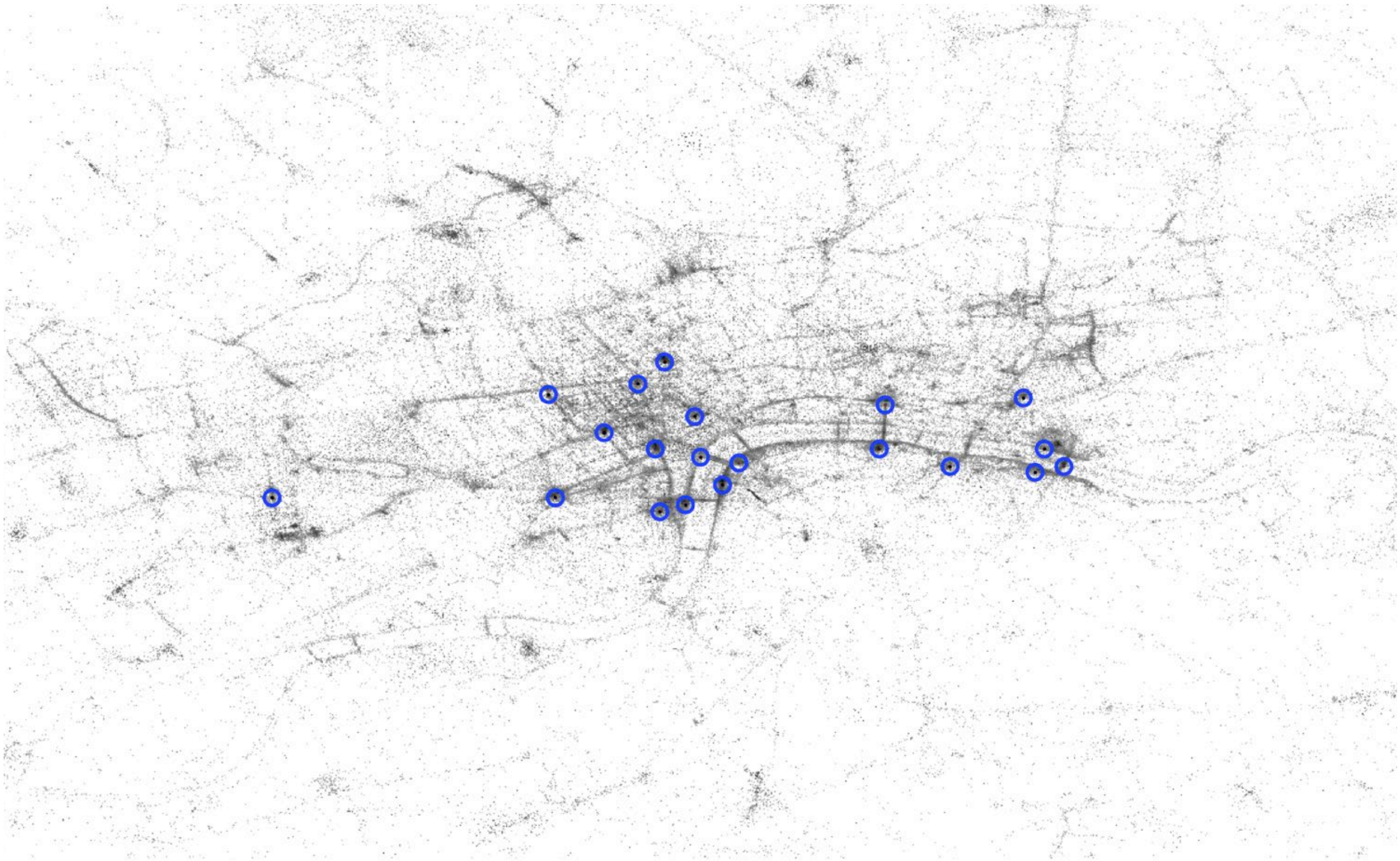


Measuring image similarity

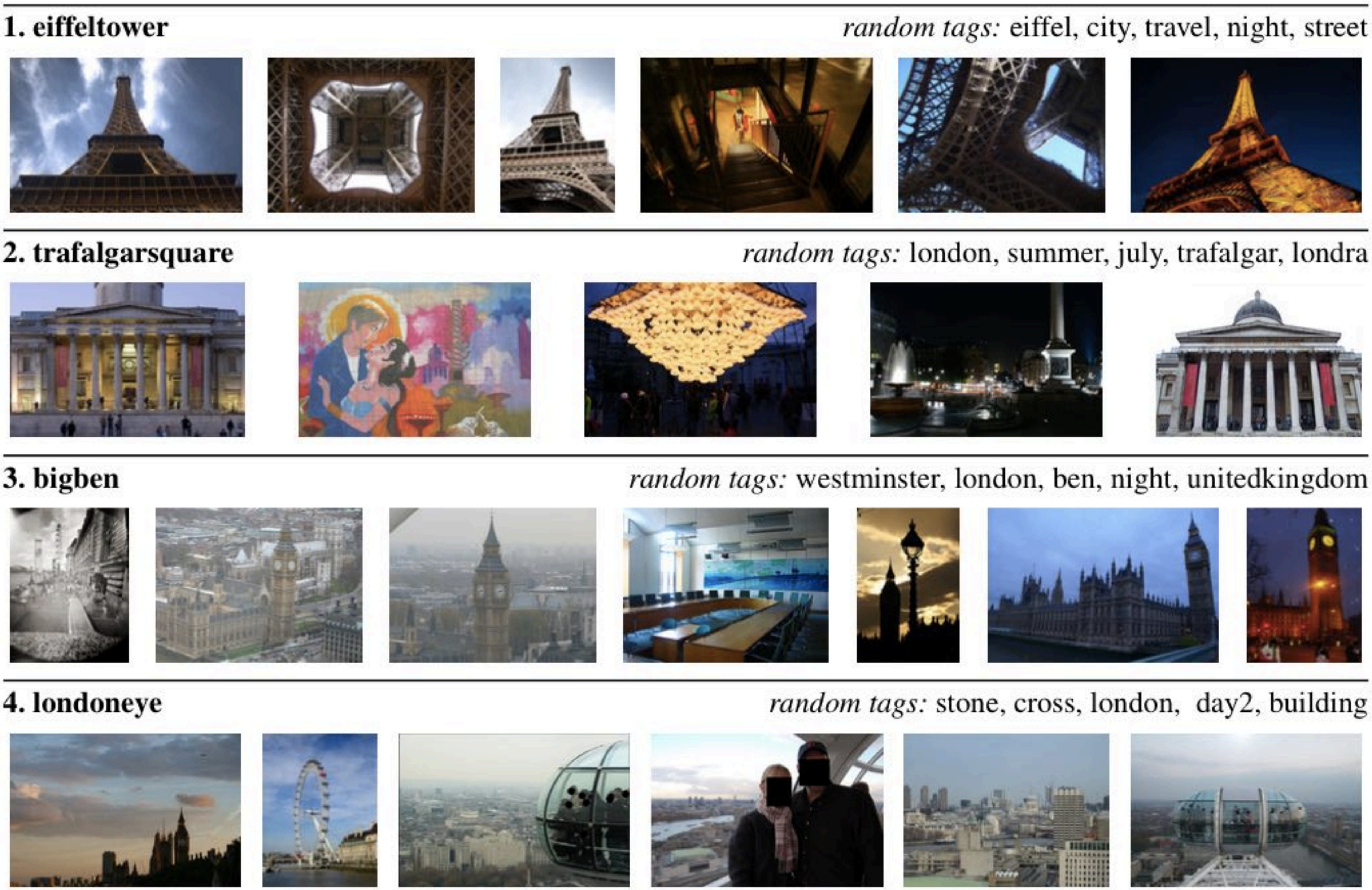
- We use SIFT to extract interest point descriptors [Lowe04]
  - Compute an invariant descriptor for each interest point
  - ~1000 interest points per image, 128-dimensional descriptors
  - To compare 2 images, count number of “matching” descriptors







D. Crandall, L. Backstrom, D. Huttenlocher, J. Kleinberg. "Mapping the World's Photos," *WWW* 2009.


1. eiffeltower
random tags: eiffel, city, travel, night, street
2. trafalgarsquare
random tags: london, summer, july, trafalgar, londra
3. bigben
random tags: westminster, london, ben, night, unitedkingdom
4. londoneye
random tags: stone, cross, london, day2, building

Landmark classification

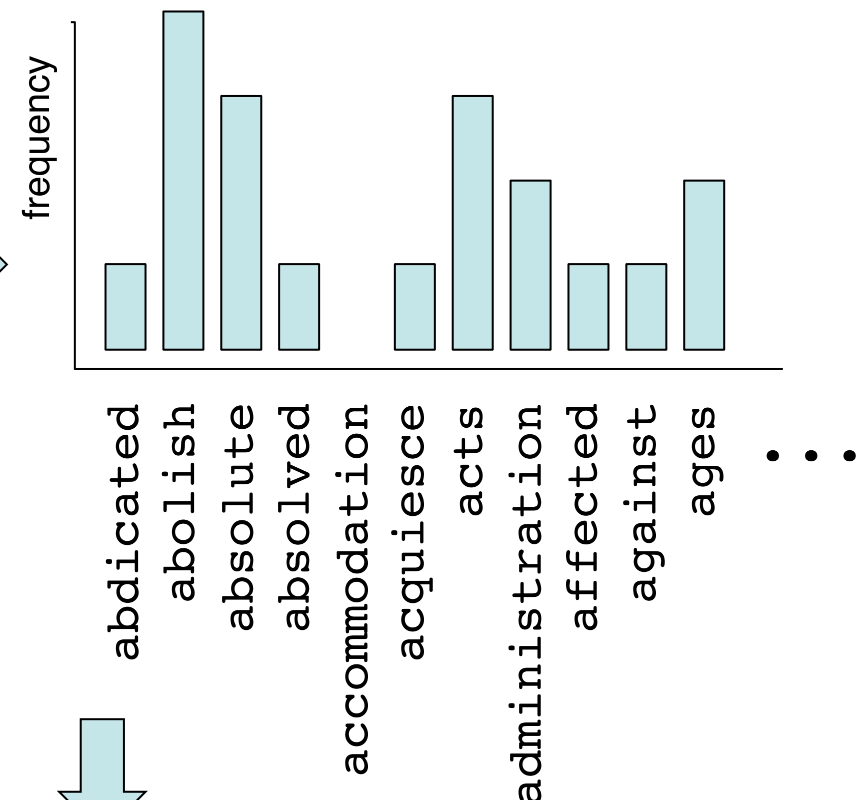
- Our task: given a photo known to be taken at one of *n* landmarks, identify the correct landmark
  - Define classes based on data-driven "hotspots" of photo activity
- For training, use ~100 million geo-tagged Flickr photos
  - Geo-tags give us (noisy) ground truth labels
- For testing, use separate set of millions of Flickr photos
- Approach based on "bag of visual words" models

Y. Li, D. Crandall, D. Huttenlocher. "Landmark recognition in large-scale image collections," *ICCV* 2009.

Vector space model

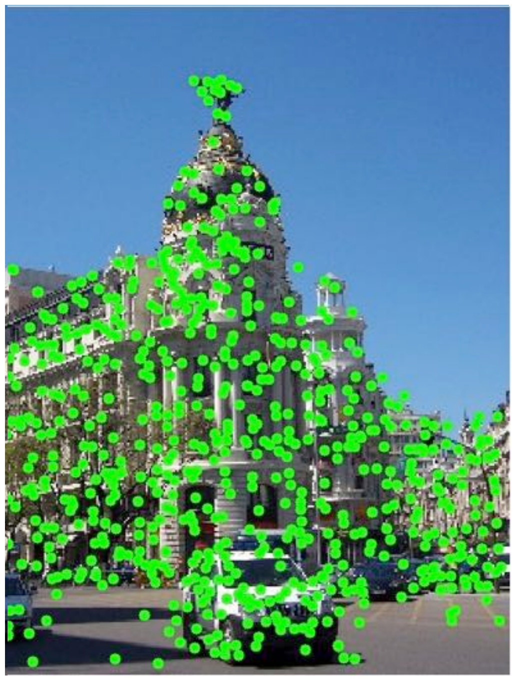
- Represent a document as a histogram over word frequency

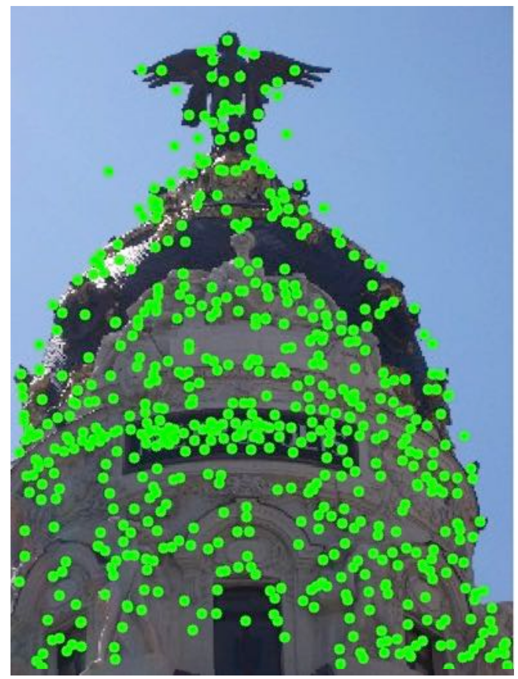
When in the Course of human
events, it becomes necessary
for one people to dissolve
the political bands which
have connected them with
another, and to assume among
the powers of the earth, the
separate and equal station to
which the Laws of Nature and
of Nature's God entitle them,
a decent respect to the…



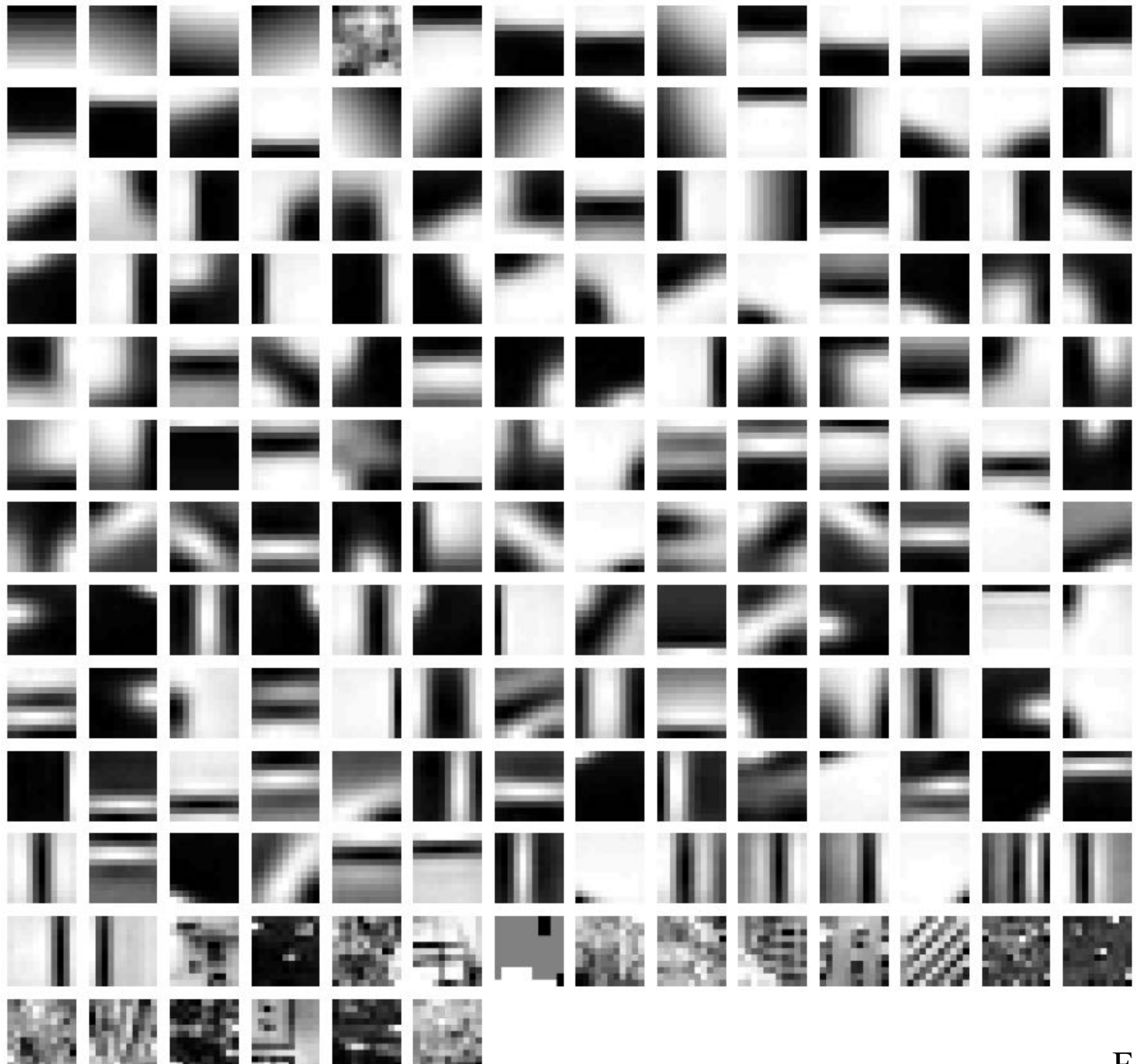
Encode mathematically as a vector: (1,4,3,1,0,1,3,2,1,1,2 …

Find “interest points”





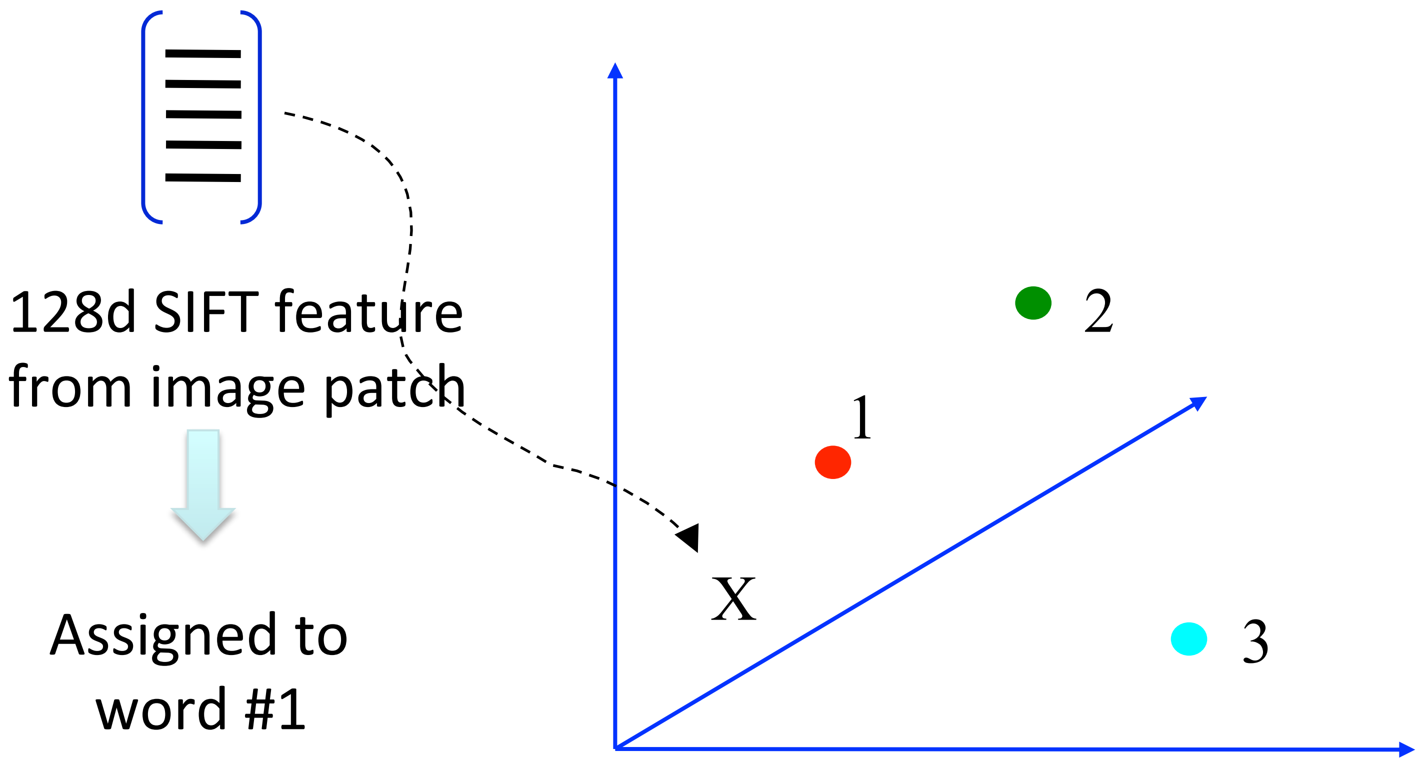
Build a "visual vocabulary"



Fei-Fei et al. 2005

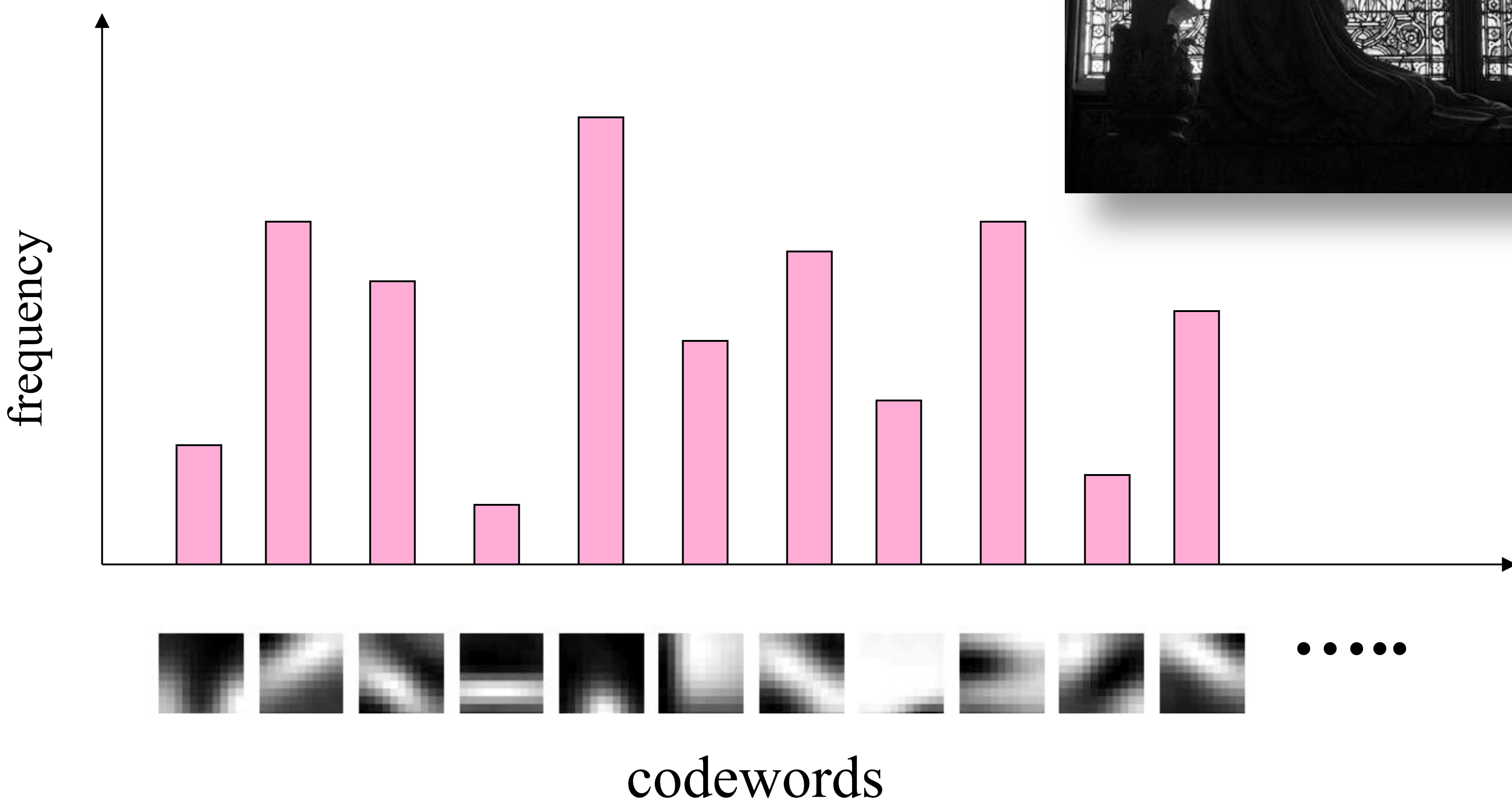
Map features to words

- Given a feature in a new image, assign it to the closest visual word in the clustered “vocabulary”



Adapted from slide by J. Sivic

Compute visual word histogram for each image



Apply machine learning

- Given feature vectors from many labeled images, learn a model of a landmark
  - E.g. using a Support Vector Machine (SVM)

Landmark classification results

Categories	Random baseline	Images - BoW visual	Images - BoW text	Images - BoW vis+text
Top 10 landmarks	10.00	57.55	69.25	80.91
Landmark 200-209	10.00	51.39	79.47	86.53
Landmark 400-409	10.00	41.97	78.37	82.78
Human baseline	10.00	68.00	—	76.40
Top 20 landmarks	5.00	48.51	57.36	70.47
Landmark 200-219	5.00	40.48	71.13	78.34
Landmark 400-419	5.00	29.43	71.56	75.71
Top 50 landmarks	2.00	39.71	52.65	64.82
Landmark 200-249	2.00	27.45	65.62	72.63
Landmark 400-449	2.00	21.70	64.91	69.77
Top 100 landmarks	1.00	29.35	50.44	61.41
Top 200 landmarks	0.50	18.48	47.02	55.12
Top 500 landmarks	0.20	9.55	40.58	45.13

Y. Li, D. Crandall, D. Huttenlocher. “Landmark recognition in large-scale image collections,” *ICCV* 2009.

Classifying photo streams

3:35pm

Alcatraz, SF bay?
Ellis Island, NYC?

8:03pm

Piazza San Marco, Venice?
Sather Tower, Berkeley?

9:27pm

Bay Bridge, SF bay?
Geo Wash Bridge, NYC?

Classifying photo streams

3:35pm

Alcatraz, SF bay?
~~Ellis Island, NYC?~~

8:03pm

~~Piazza San Marco, Venice?~~
Sather Tower, Berkeley?

9:27pm

Bay Bridge, SF bay?
~~Geo Wash Bridge, NYC?~~

- Model as a Hidden Markov Model, learn parameters via Structured SVMs, do fast inference with Viterbi algorithm

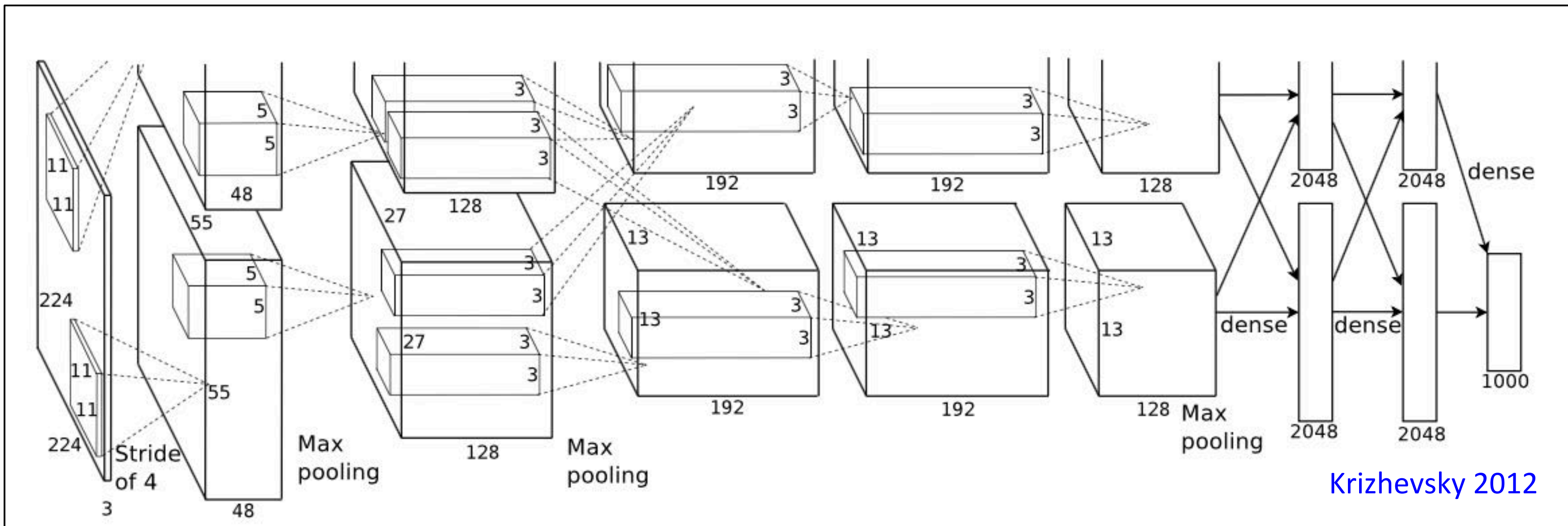
Landmark classification results

Categories	Random baseline	Images - BoW visual	Images - BoW text	Images - BoW vis+text	Photo streams visual	Photo streams text	Photo streams vis+text
Top 10 landmarks	10.00	57.55	69.25	80.91	68.82	70.67	82.54
Landmark 200-209	10.00	51.39	79.47	86.53	60.83	79.49	87.60
Landmark 400-409	10.00	41.97	78.37	82.78	50.28	78.68	82.83
Human baseline	10.00	68.00	—	76.40	—	—	—
Top 20 landmarks	5.00	48.51	57.36	70.47	62.22	58.84	72.91
Landmark 200-219	5.00	40.48	71.13	78.34	52.59	72.10	79.59
Landmark 400-419	5.00	29.43	71.56	75.71	38.73	72.70	75.87
Top 50 landmarks	2.00	39.71	52.65	64.82	54.34	53.77	65.60
Landmark 200-249	2.00	27.45	65.62	72.63	37.22	67.26	74.09
Landmark 400-449	2.00	21.70	64.91	69.77	29.65	66.90	71.62
Top 100 landmarks	1.00	29.35	50.44	61.41	41.28	51.32	62.93
Top 200 landmarks	0.50	18.48	47.02	55.12	25.81	47.73	55.67
Top 500 landmarks	0.20	9.55	40.58	45.13	13.87	41.02	45.34

Y. Li, D. Crandall, D. Huttenlocher. "Landmark recognition in large-scale image collections," *ICCV* 2009.

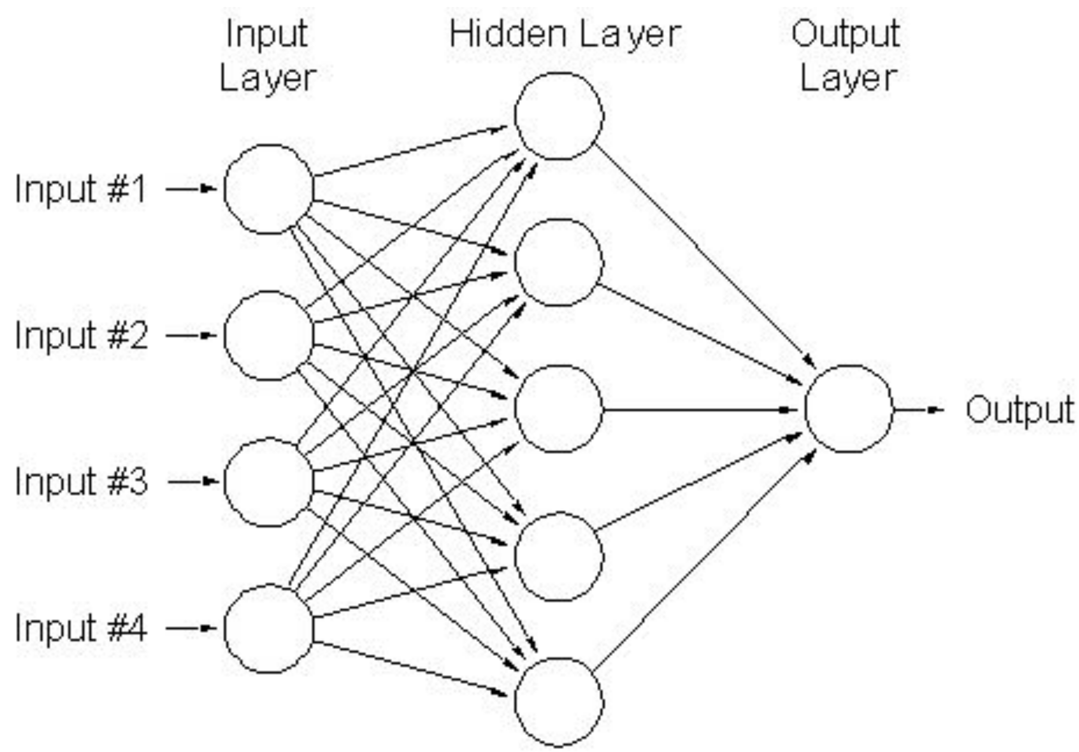
Deep learning

- A breakthrough in Artificial Intelligence
  - Learn low-level features and high-level classifier simultaneously, e.g. using Convolutional Neural Networks



Krizhevsky 2012

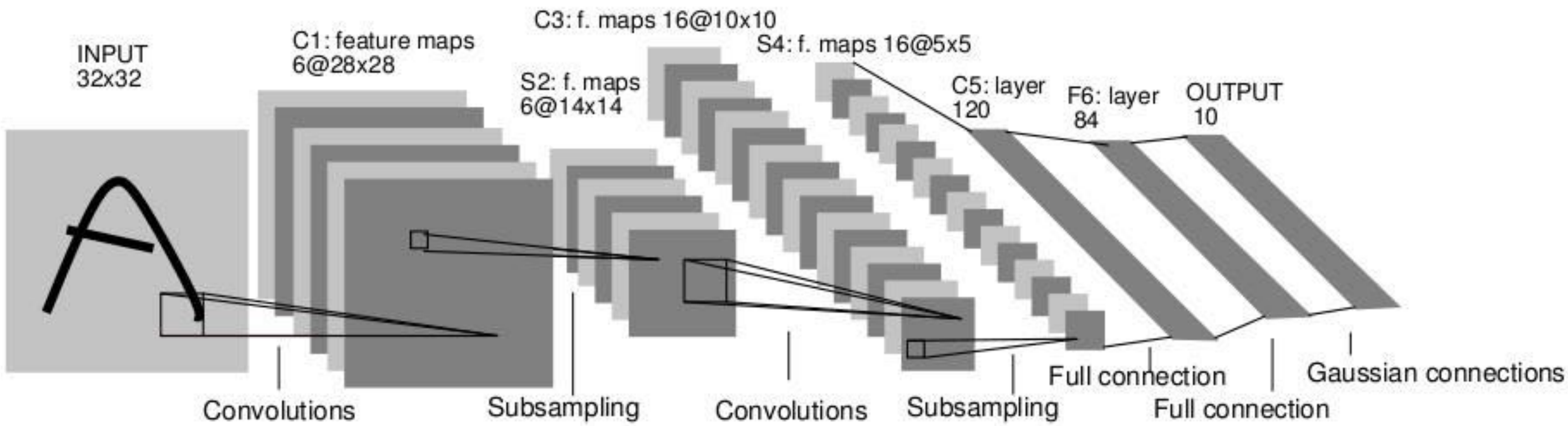
Background: Multi-Layer Neural Networks



- Each neuron calculates a non-linear function of the dot product of its inputs with a weight vector

Adapted from slide by R. Fergus

Convolutional Neural Network

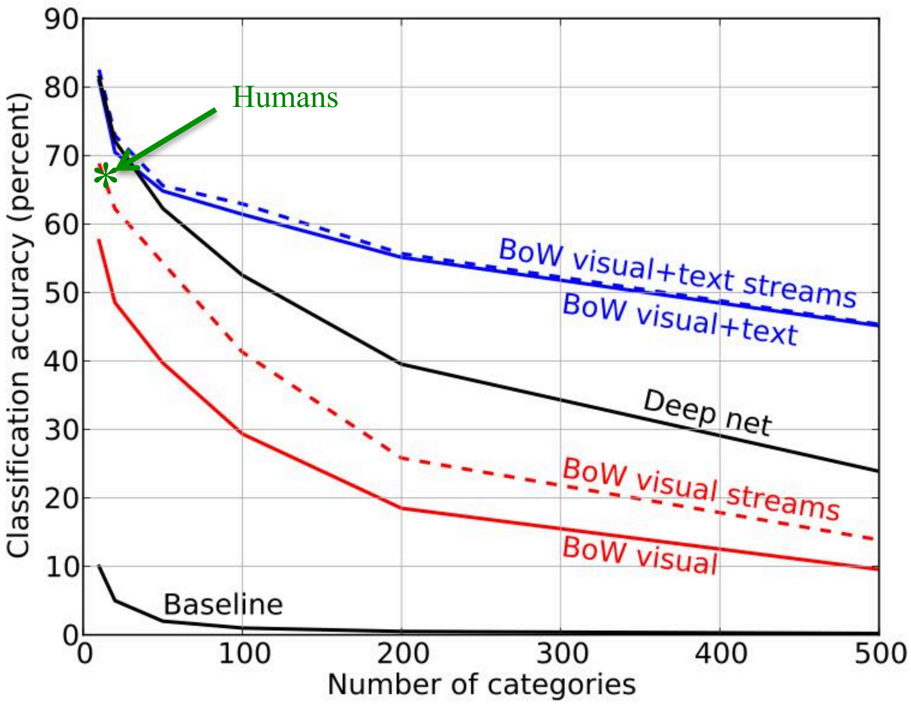


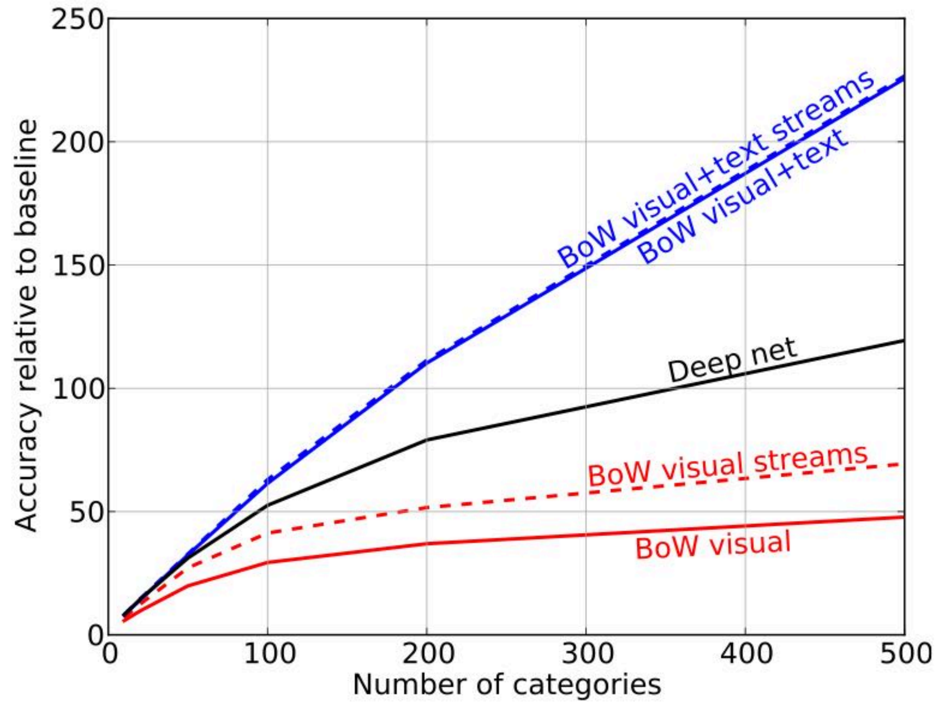
Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner, Gradient-based learning applied to document recognition, Proceedings of the IEEE 86(11): 2278–2324, 1998.

Landmark classification results

Categories	Random baseline	Images - BoW visual	Images - BoW text	Images - BoW vis+text	Photo streams visual	Photo streams text	Photo streams vis+text	Images - deep visual
Top 10 landmarks	10.00	57.55	69.25	80.91	68.82	70.67	82.54	81.43
Landmark 200-209	10.00	51.39	79.47	86.53	60.83	79.49	87.60	—
Landmark 400-409	10.00	41.97	78.37	82.78	50.28	78.68	82.83	—
Human baseline	10.00	68.00	—	76.40	—	—	—	68.00
Top 20 landmarks	5.00	48.51	57.36	70.47	62.22	58.84	72.91	72.10
Landmark 200-219	5.00	40.48	71.13	78.34	52.59	72.10	79.59	—
Landmark 400-419	5.00	29.43	71.56	75.71	38.73	72.70	75.87	—
Top 50 landmarks	2.00	39.71	52.65	64.82	54.34	53.77	65.60	62.28
Landmark 200-249	2.00	27.45	65.62	72.63	37.22	67.26	74.09	—
Landmark 400-449	2.00	21.70	64.91	69.77	29.65	66.90	71.62	—
Top 100 landmarks	1.00	29.35	50.44	61.41	41.28	51.32	62.93	52.52
Top 200 landmarks	0.50	18.48	47.02	55.12	25.81	47.73	55.67	39.52
Top 500 landmarks	0.20	9.55	40.58	45.13	13.87	41.02	45.34	23.88

Landmark classification results





Some random failures

	(a)	(b)	(c)	(d)	(e)
Correct:	Trafalgar Square	London Eye	Trafalgar Square	Notre Dame	Trafalgar Square
Predicted:	Colesseum	Eiffel Tower	Piazza San Marco	Eiffel Tower	Empire State Building

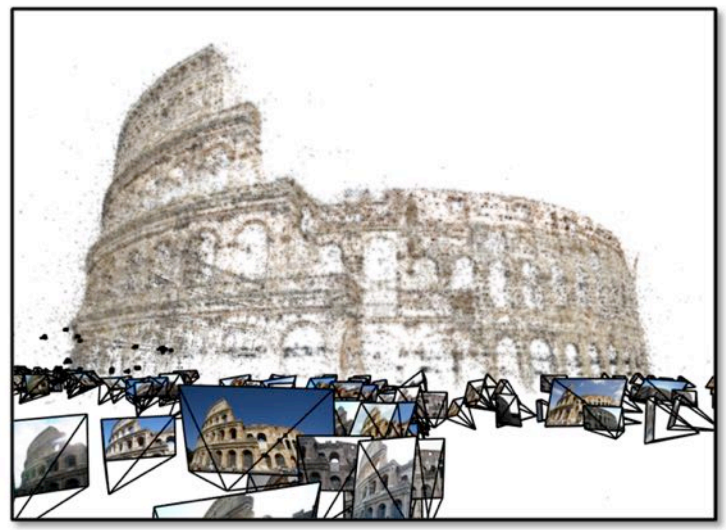
	(f)	(g)	(h)	(i)	(j)
Correct:	Tate Modern	Big Ben	Notre Dame	Louvre	Piazza San Marco
Predicted:	Louvre	Piazza San Marco	Big Ben	Notre Dame	London Eye

Building 3D reference models

If we had a 3D model, we could geo-locate images very precisely.

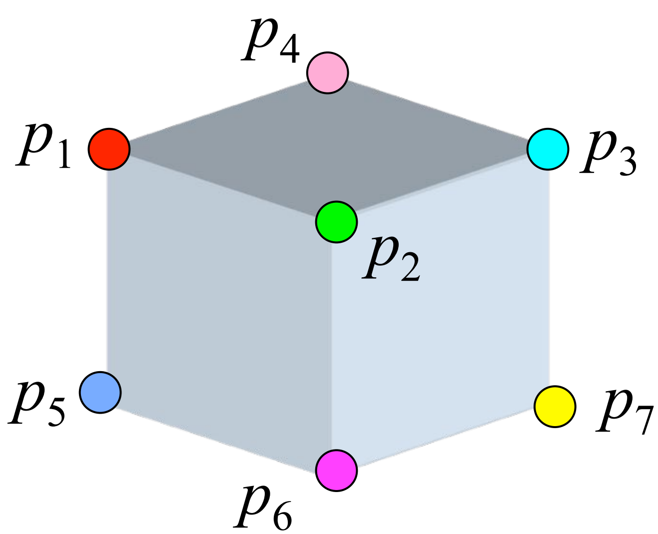
If we had precise geo-locations for photos, we could build a 3D model.

So we have to do both simultaneously…

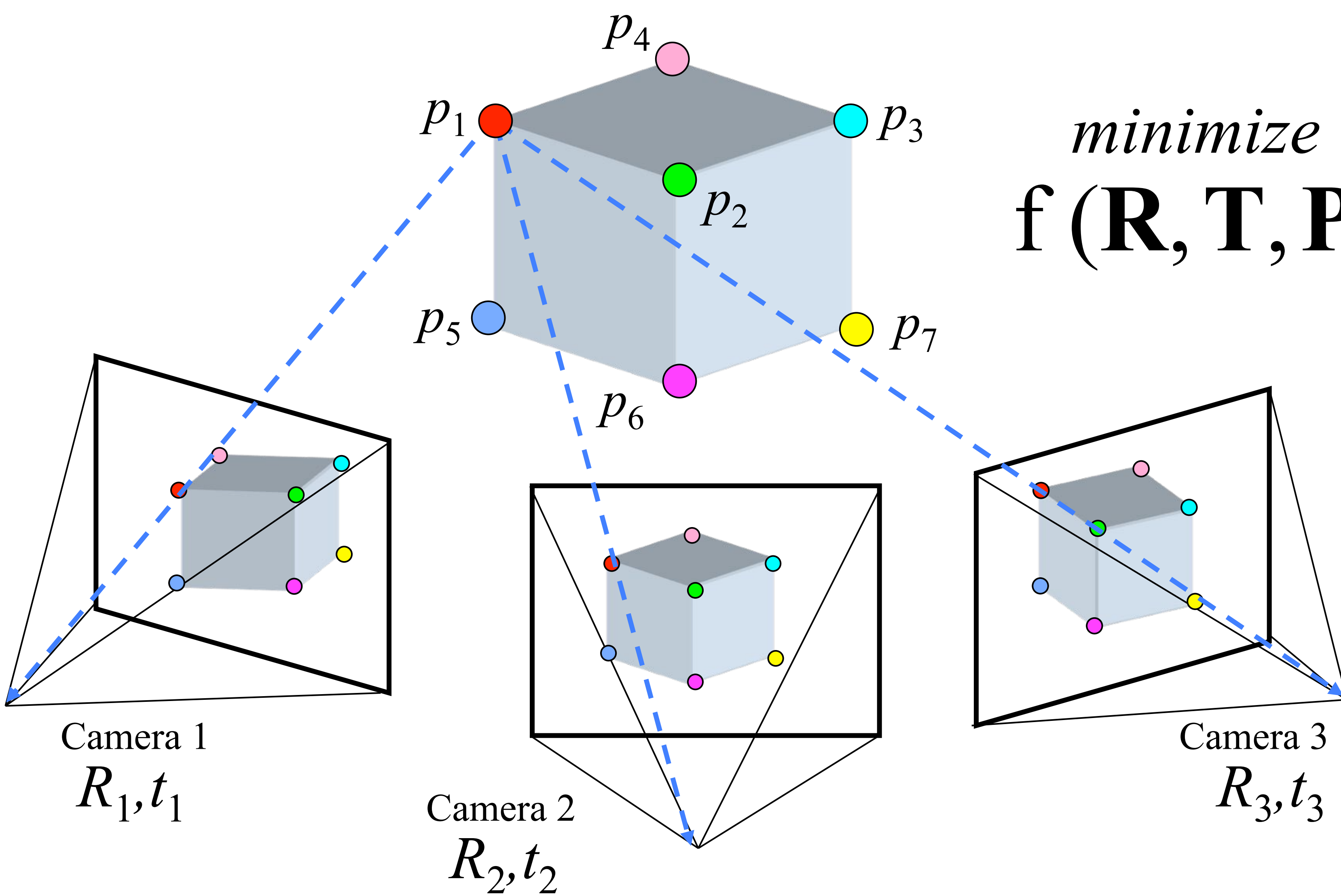


[Snavely06]

Solving for scene structure and camera poses



Solving for scene structure and camera poses



minimize

$$\mathrm{f}\,(\mathbf{R}, \mathbf{T}, \mathbf{P})$$

Structure from motion on unstructured photo sets

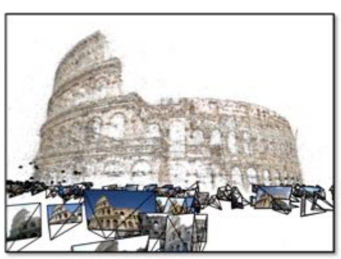
Feature point detection

Image matching
Find scene points seen by multiple cameras

Initialization
Robustly estimate camera poses and/or scene points

Discrete optimization over viewing directions
NLLS refinement of rotations
Discrete optimization over 2D translations
NLLS refinement of 3D translations

Bundle adjustment
Solve for camera poses R, T and scene structure P



D. Crandall, A. Owens, N. Snavely, D. Huttenlocher, "SfM with MRFs: Discrete-Continuous Optimization for Large-scale Structure from Motion," *PAMI,* December 2013.

Our approach

- View SfM as inference over a Markov random field, solving for all camera poses at once

- **Vertices** are cameras (or points)
- Both **pairwise** and **unary** constraints
- **Inference problem:** label each image with a camera pose, such that constraints are satisfied

Our approach

- View SfM as inference over a Markov random field, solving for all camera poses at once

  - Combines **discrete** and **continuous** optimization:
    - **Discrete optimization** (loopy belief propagation) with robust energy functions used to find good initialization
    - **Continuous optimization** (bundle adjustment) used to refine

Reconstruction video

http://www.cs.indiana.edu/~djcran/combined-movies.m4v

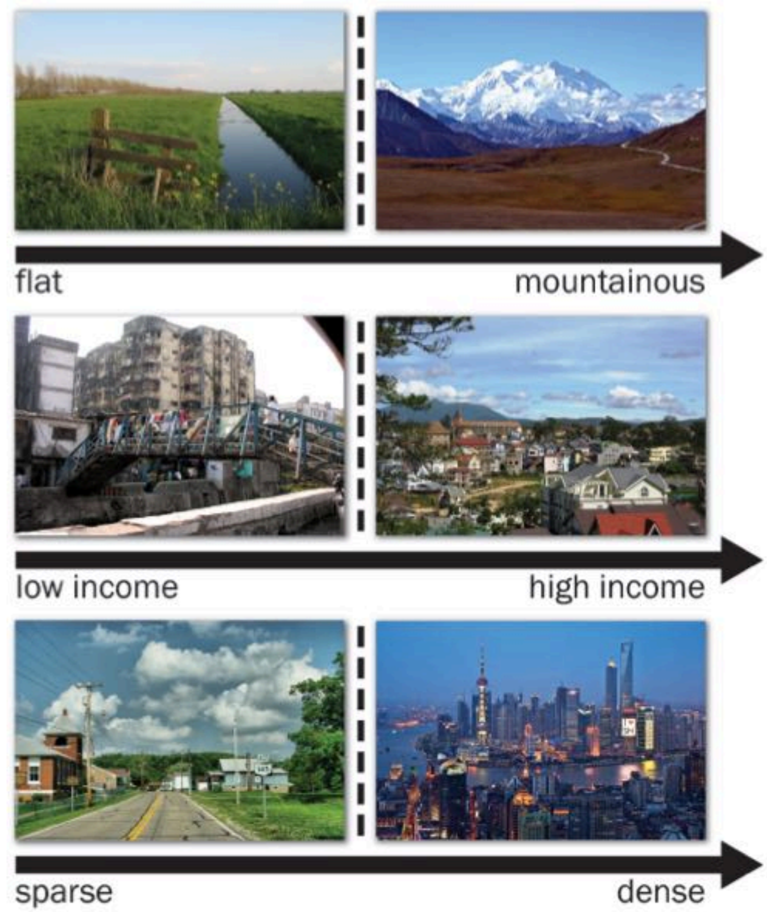
Median geotag accuracy from **GPS**: **15.5m**
Median geotag accuracy from **3D reconstruction**: **1.16m**

D. Crandall, A. Owens, N. Snavely, D. Huttenlocher, "SfM with MRFs: Discrete-Continuous Optimization for Large-scale Structure from Motion," *PAMI,* December 2013.


But what about the rest of the world?

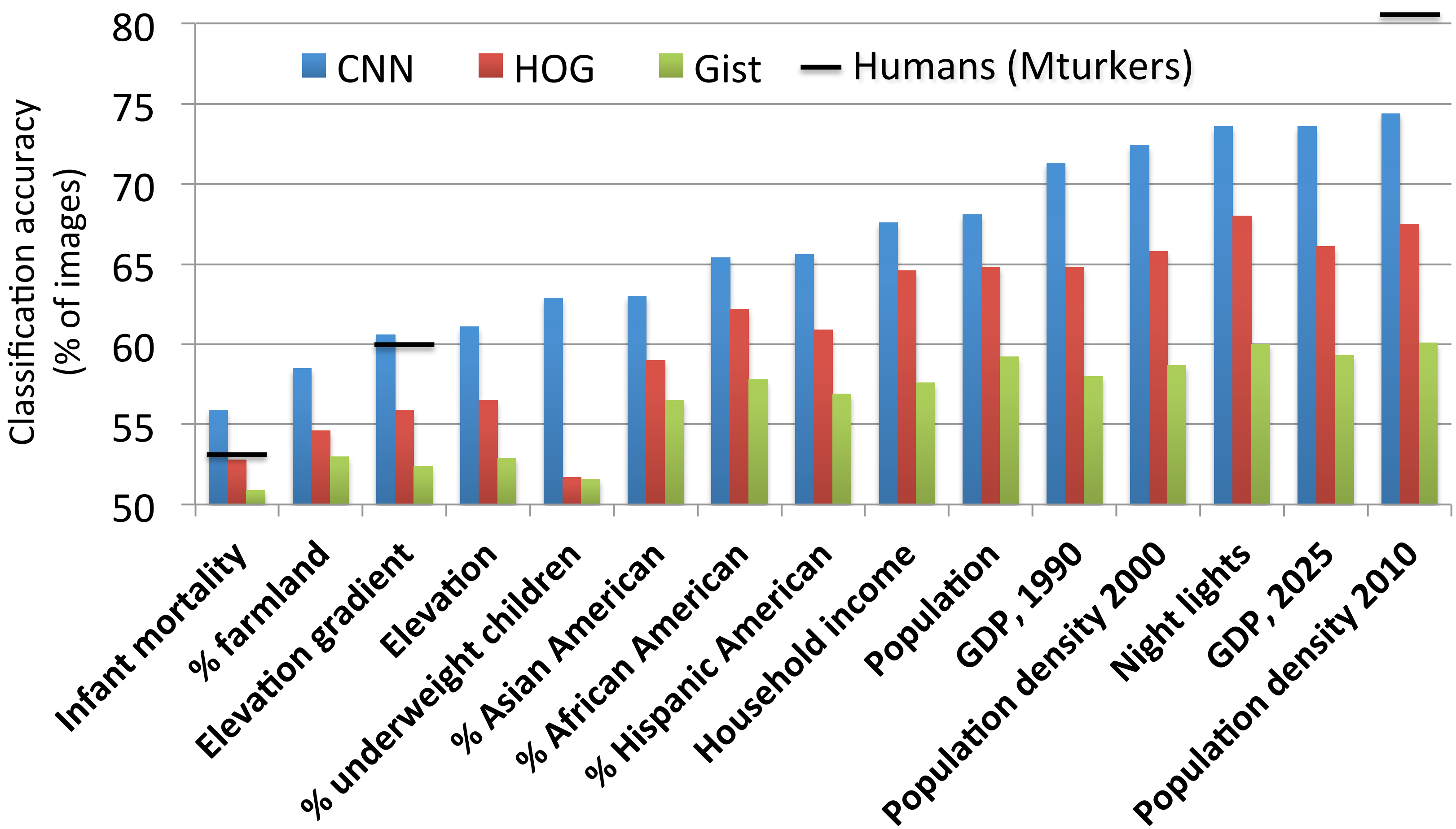
Recognizing geo-spatial attributes

- Can we recognize *attributes* of the place where a photo was taken?
  - Then use public GIS maps to narrow down the possible places
- Use geotagged images from Flickr, cross-referenced with GIS maps
- Compare deep learning with traditional visual features



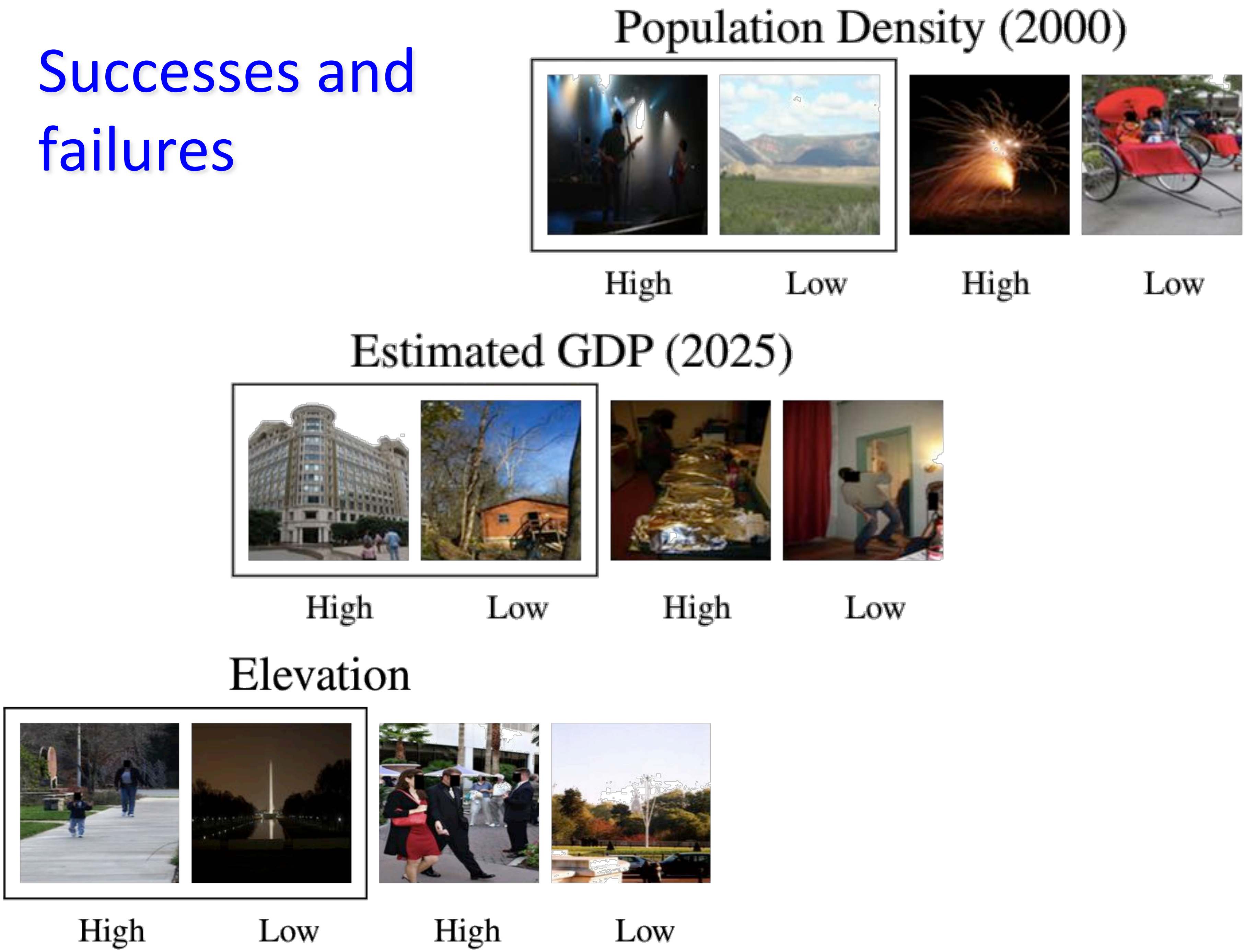
S. Lee, H. Zhang, D. Crandall. "Predicting geo-informative attributes in large-scale image collections using convolutional neural networks," *WACV* 2015.

Deep learning for geo-informative attribute detection



S. Lee, H. Zhang, D. Crandall. "Predicting geo-informative attributes in large-scale image collections using convolutional neural networks," *WACV* 2015.

Successes and failures



Computational patterns in vision

1. Single image tasks (e.g. feature extraction)
   - # of images may be large, but easily parallelizable
2. Image matching (e.g. recognition, clustering)
   - Evaluating distances between many high-dimensional vectors
3. Iterative algorithms (e.g. learning)
   - Few, but long-running iterations (e.g. k-means)
   - Lightweight, but many iterations (e.g. neural net backprop)
4. Inference on graphs (e.g. reconstruction, learning)
   - Small graphs with huge label spaces (e.g. pose detection)
   - Large graphs with small label spaces (e.g. resolving stereo)
   - Large graphs with large label spaces (e.g. reconstruction)

For more information about these projects, please visit:

http://vision.soic.indiana.edu/

Thanks to:

- *Funders:* NSF CAREER, IARPA, Google
- *Collaborators:* Dan Huttenlocher, Apu Kapadia, Yunpeng Li, Noah Snavely
- *Students:* Sven Bambach, Mohammed Korayem, Stefan Lee, Andrew Owens, Rob Templeman, Jingya Wang, Haipeng Zhang

This work was supported in part by the Intelligence Advanced Research Projects Activity (IARPA) via Air Force Research Laboratory, contract FA8650-12-C-7212. The U.S. Government is authorized to reproduce and distribute reprints for Governmental purposes notwithstanding any copyright annotation thereon. Disclaimer: The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies or endorsements, either expressed or implied, of IARPA, AFRL, or the U.S. Government.