Grounding Vision in the Real World

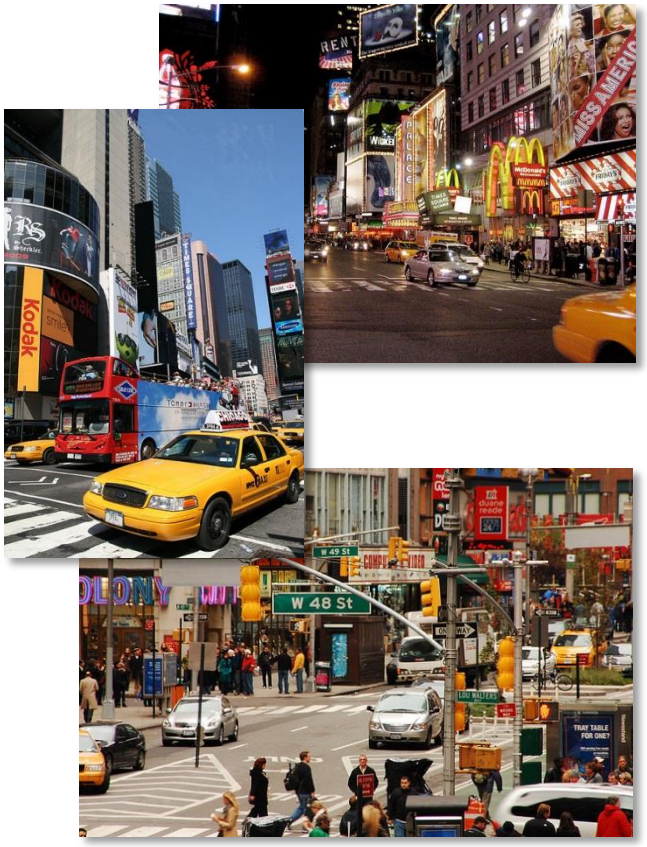
Noah Snavely
Google Inc., Cornell University

RSS RGB-D Workshop, 2014

The Age of Exapixel Image Data

- Trillions of photos available online
- Millions uploaded every hour
- Interconnected
- The Internet is becoming a **living visual record** of our world

What can we do with all this data?



- Use images to understand the world
  - Building the ultimate world map
  - Changes in cities over time
  - Traffic patterns
  - Surprising events
  - Forensics: what happened, when?
- **Challenge**: data is extremely unstructured

flickr from YAHOO!

Home You Organize & Create Contacts Groups Astrometry.net

Favorite Actions Share ← Newer Older →

NGC 2264 / Christmas Tree cluster / Cone nebula

(reprocessed)

Hello, this is the blind astrometry solver. Your results are:
(RA, Dec) center:(100.227235684, 9.93850271956) degrees
(RA, Dec) center (H:M:S, D:M:S):(06:40:54.537, +9:56:18.610)
Orientation:146.00 deg E of N

Pixel scale:3.17 arcsec/pixel

Parity:Reverse ("Left-handed")
Field size :72.02 x 52.33 arcminutes
Your field contains:
The star 15Mon
NGC 2264 / Christmas Tree cluster / Cone nebula

View in World Wide Telescope

If you would like to have other images solved, please submit them to the astrometry group.

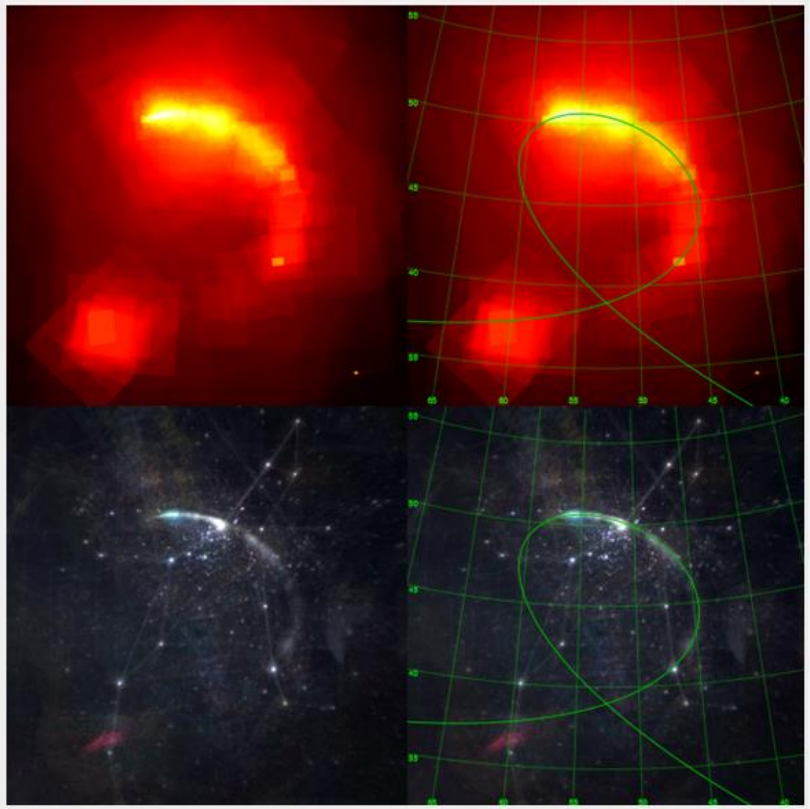
The Physics arXiv Blog

Astronomers Calculate Comet's Orbit Using Amateur Images From The Web

Amateur astrophotographs posted online represent a massive untapped resource. Now astronomers have worked out how to mine it

KFC 04/04/2011

5 COMMENTS



Lang and Hogg, **Searching for comets on the World Wide Web: The orbit of 17P/Holmes from the behavior of photographers**

Grounding photos in the real world

Photos

Place + Time

Image understanding

Prior world knowledge
(geographic / GIS data)

IM2GPS: **estimating geographic information from a single image**

James Hays and Alexei A. Efros
Carnegie Mellon University

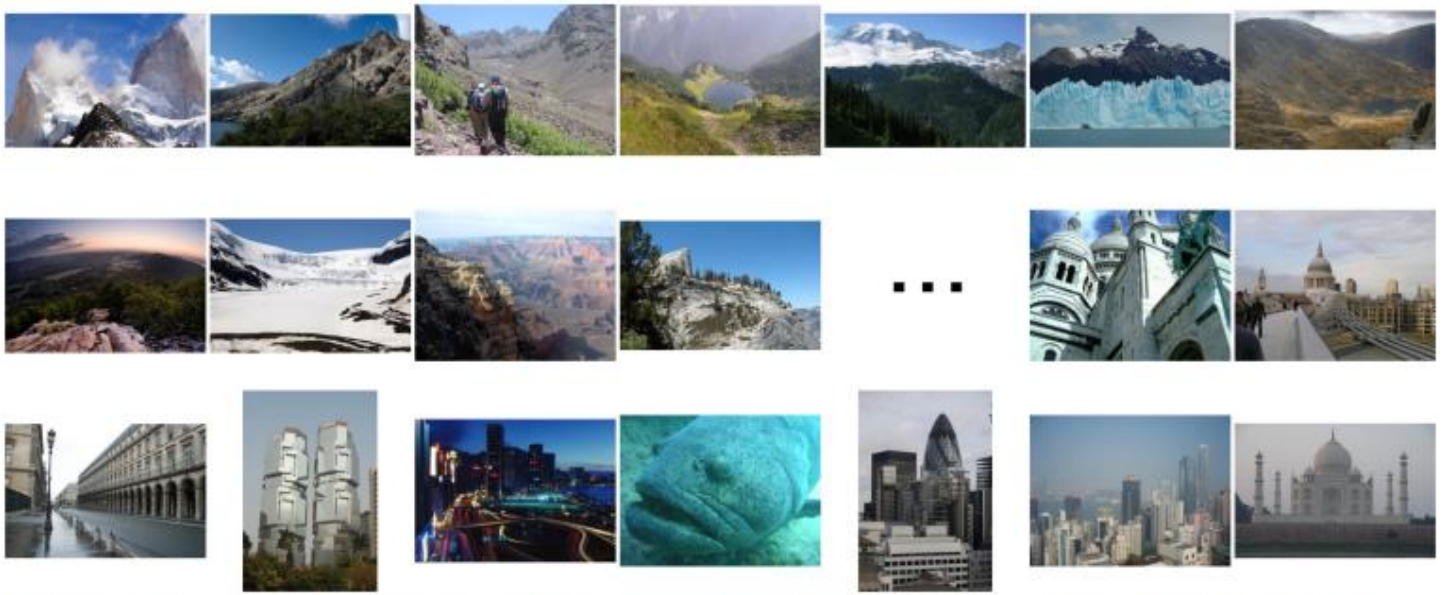


Figure 9. In scanline order, the test cases with the largest and smallest estimated elevation gradient.

Deep Photo: Model-Based Photograph Enhancement and Viewing

Johannes Kopf
University of Konstanz

Boris Neubert
University of Konstanz

Billy Chen
Microsoft

Michael F. Cohen
Microsoft Research

Daniel Cohen-Or
Tel Aviv University

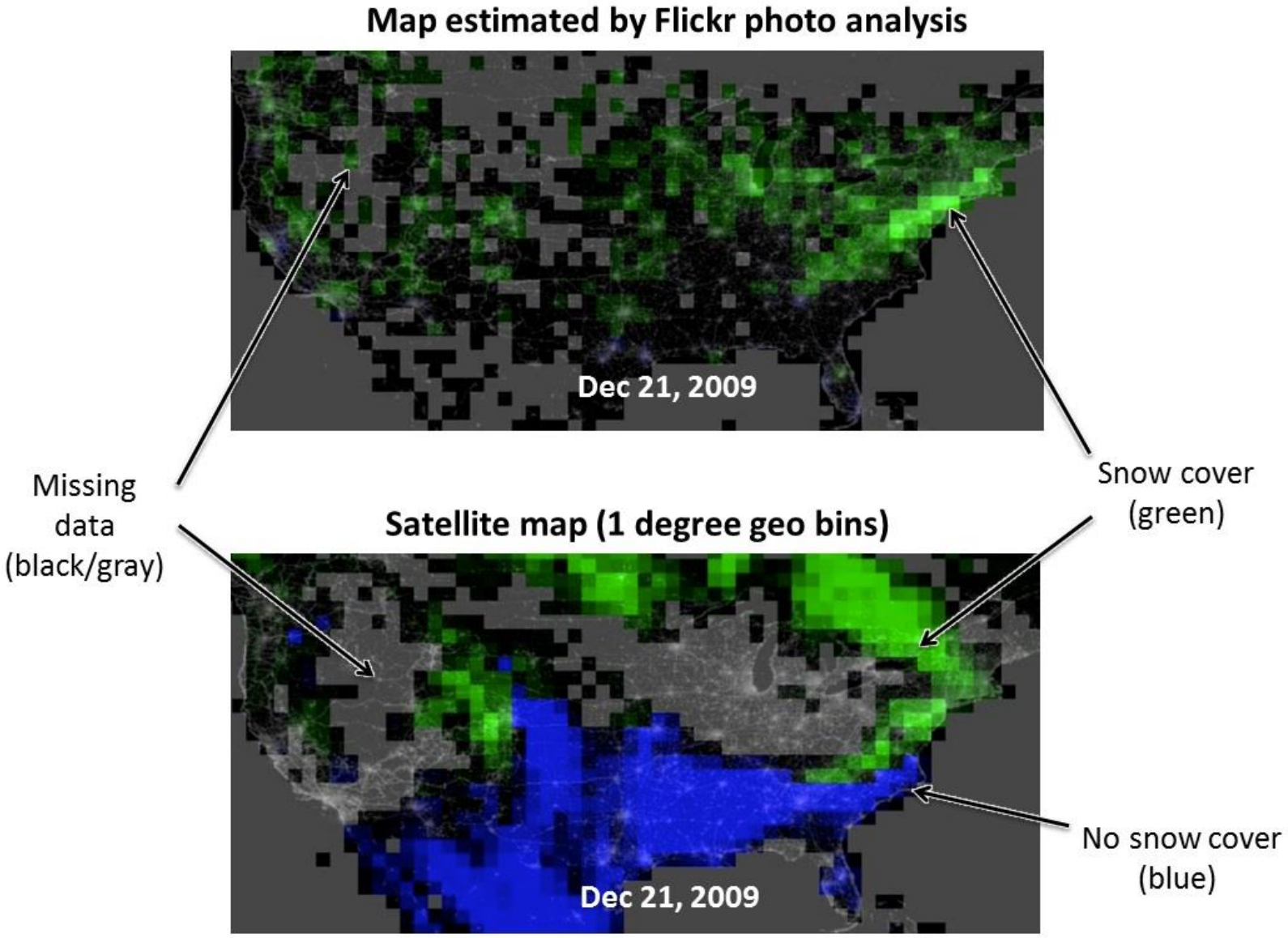
Oliver Deussen
University of Konstanz

Matt Uyttendaele
Microsoft Research

Dani Lischinski
The Hebrew University

Original | Dehazed | Relighted | Annotated

[Kopf, et al., SIGGRAPH Asia 2008]

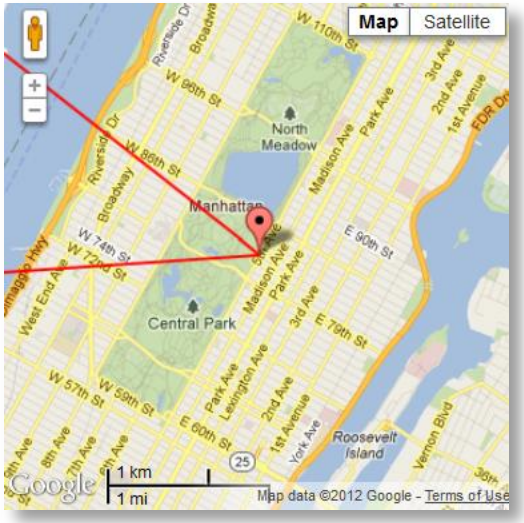


Snow cover estimation from Flickr photos

[Zheng, Korayem, Crandall, LeBuhn, WWW 2012]

Calibrating the distributed camera

- For any photo on the web



- Where was it taken? In what direction?
- What time was it taken?
- What is visible in the image? Where?
- **Thesis**: Geometry is key to this problem

Sensor data alone is too noisy

Place

Location recognition

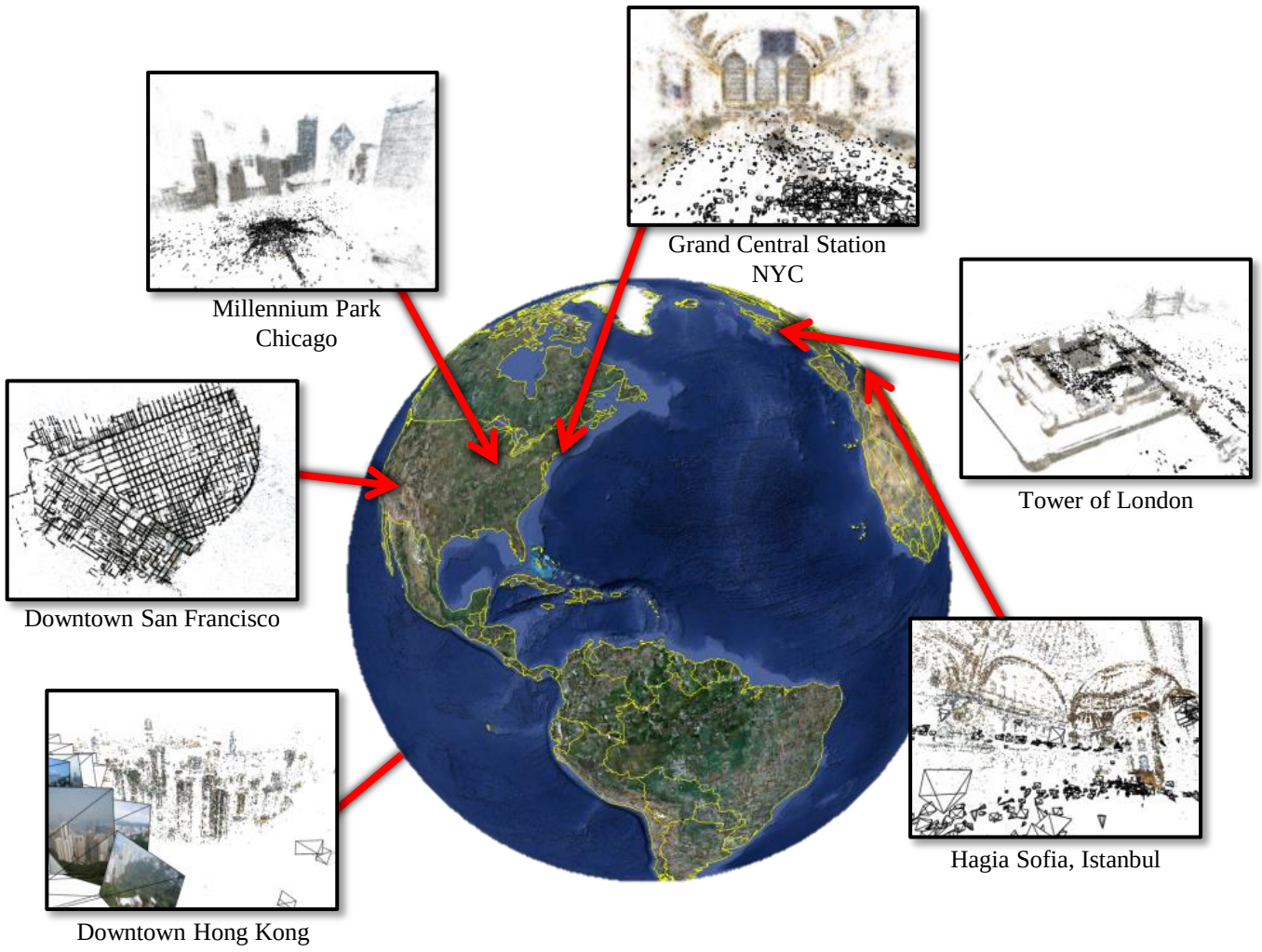
Image-based

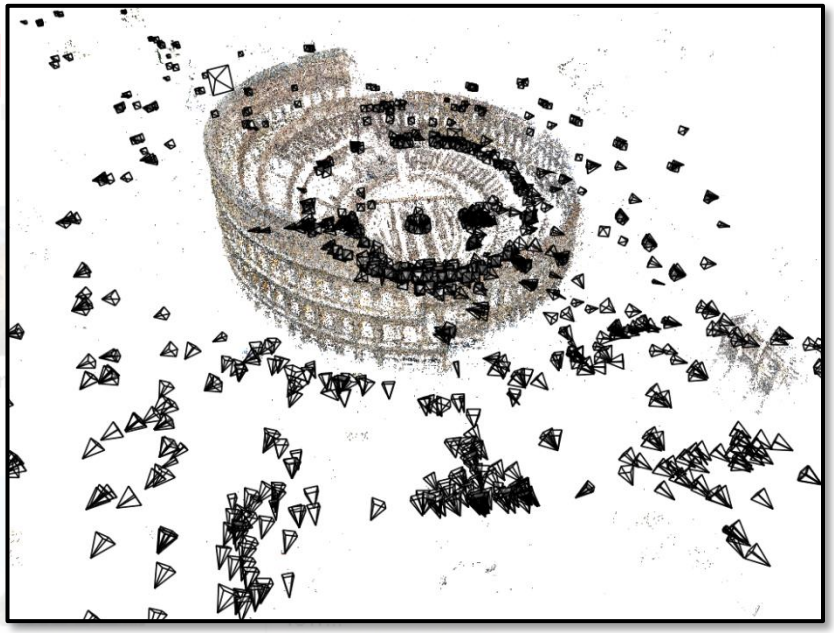
[Schindler, Brown, Szeliski '06]
[Hays & Efros '08]
[Kalogerakis *et al.* '09]
[Li, Crandall, Huttenlocher '09]
[Knopp, Sivic, Pajdla, '10]

Geometry-based

[Li, *et al.* '10]
[Sattler & Leibe '11]
[Lim *et al.*, '11]
[Li, Snavely, Huttenlocher, Fua '12], …

A Database of 3D Geometry





[Snavely, Seitz, Szeliski, 2006]

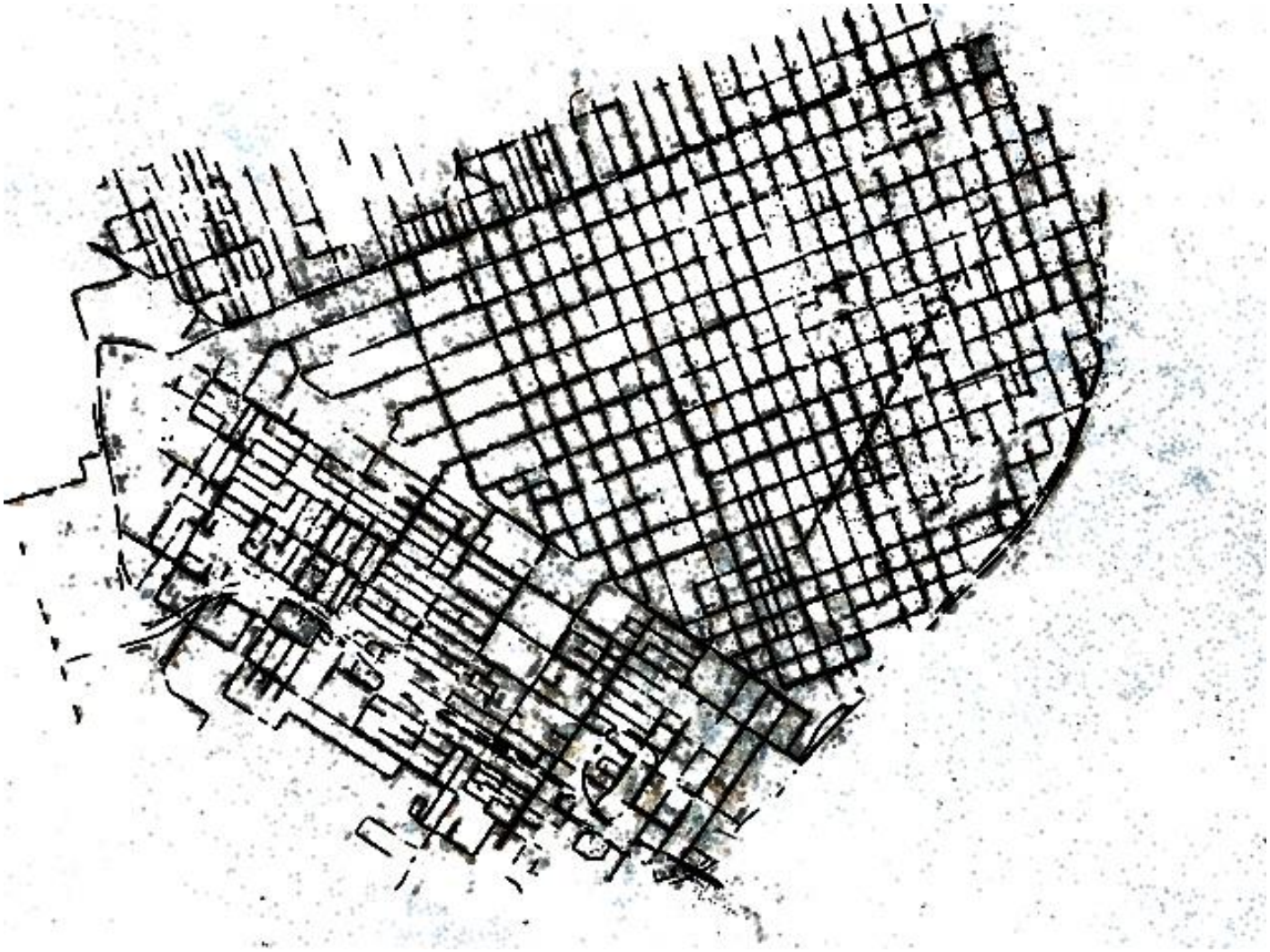
Dubrovnik, Croatia

[*Building Rome in a Day*, Agarwal, Snavely, Simon, Seitz, Szeliski, ICCV 2009]

[Crandall, Backstrom, Huttenlocher, and Kleinberg. WWW09]

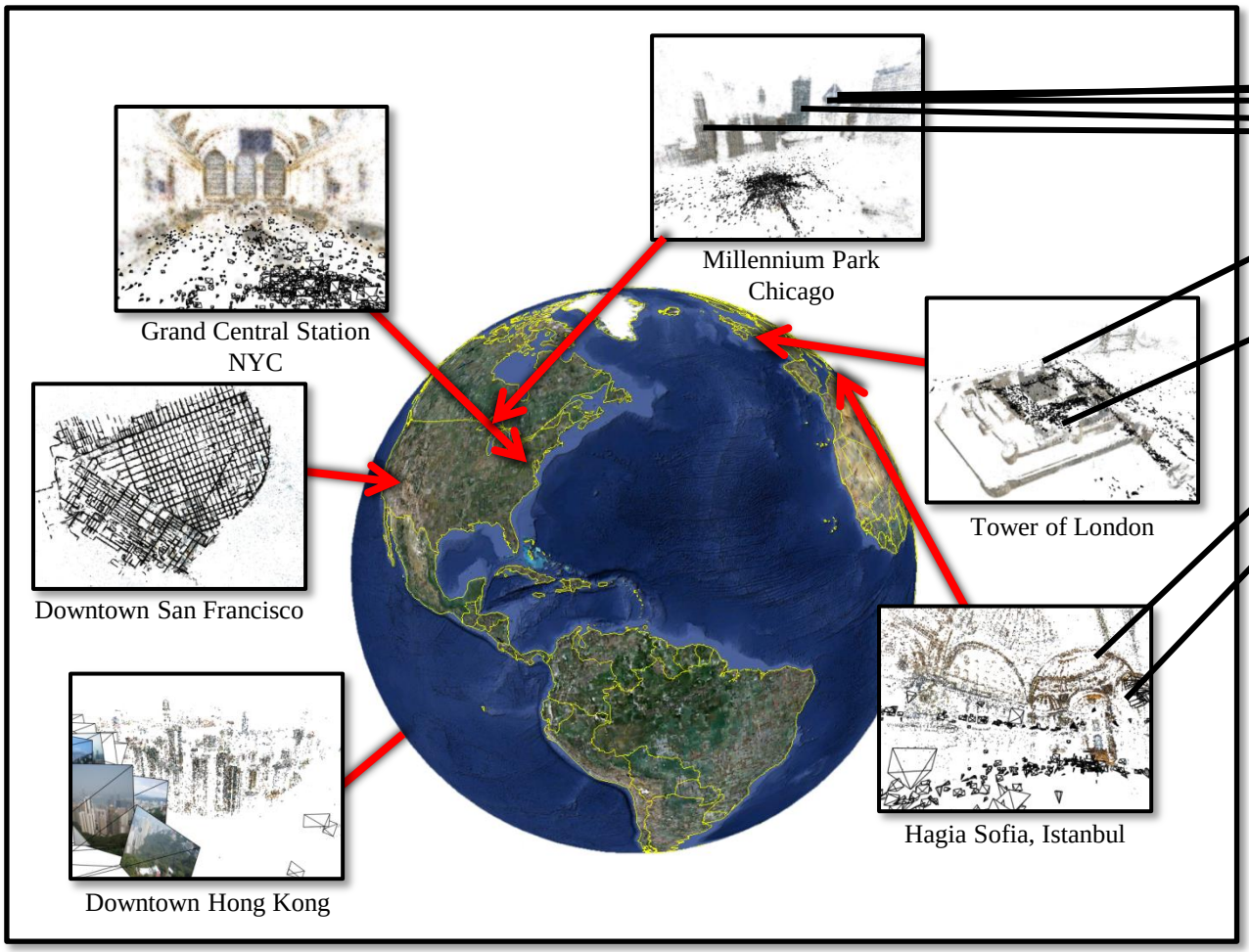
NAVTEQ SF Street View Dataset

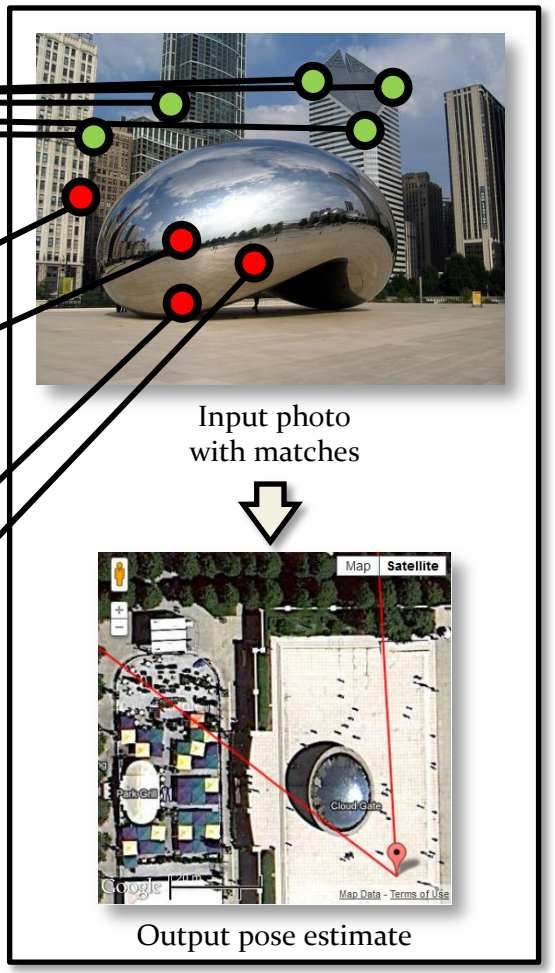
Chen et al. City-scale landmark identification on mobile devices.[CVPR 2011]



Model of San Francisco

World-wide Pose Estimation





Matching becomes challenging as # of points grows very large

[Li, Snavely, Huttenlocher, Fua. ECCV 2012]

Very large search problem

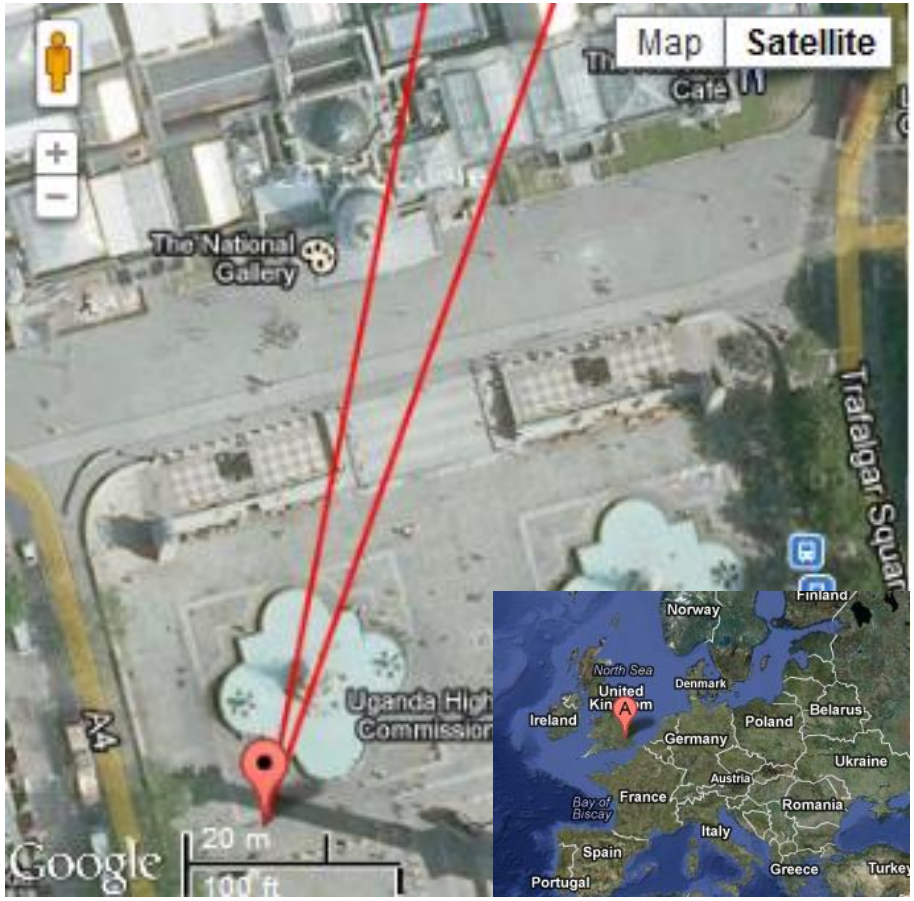
- Largest model we've created:
  - About 500M 3D points
  - … from several million images
- Each 3D point has 1 or more SIFT descriptors
  - We index these using standard kd-trees
- Finding good matches at this scale is challenging
  - We have some new tricks to estimate location quickly and robustly

[Li, Snavely, Huttenlocher, Fua. ECCV 2012]

Example results

Input Photo

Estimated Camera Pose



latitude: 51.5079 deg
longitude: -0.1283 deg
altitude: 0.718 m

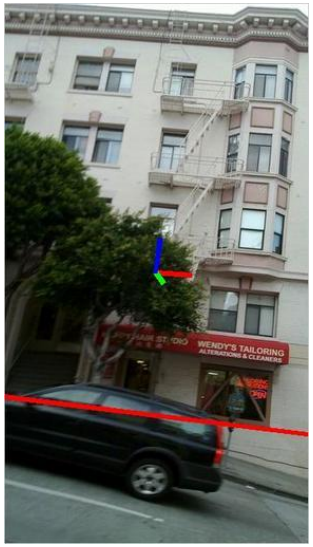
zenith: 82.2991 deg
azimuth: -8.7291 deg
roll: -0.0391 deg
focalLength: 1610.01 px

Machu Picchu, Peru

Times Square

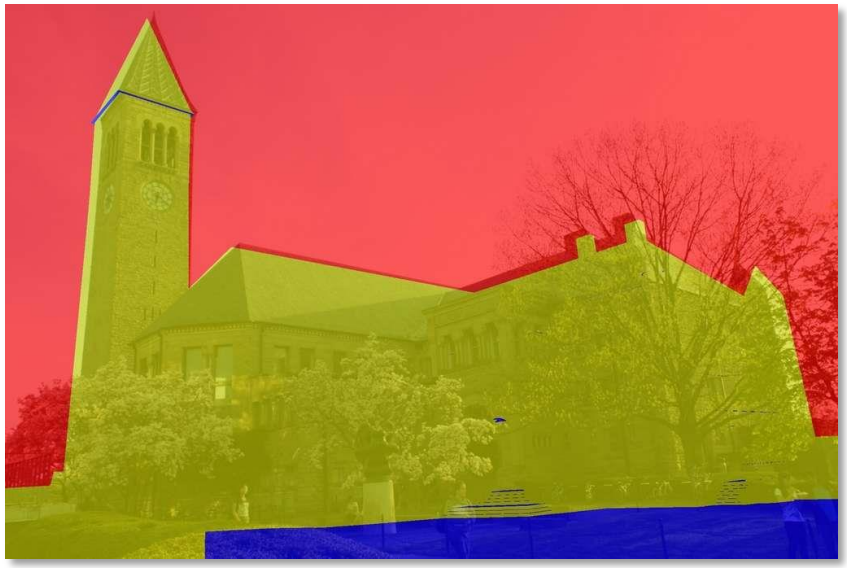
Corner of Beach and Jones (San Francisco)

Sutter St.



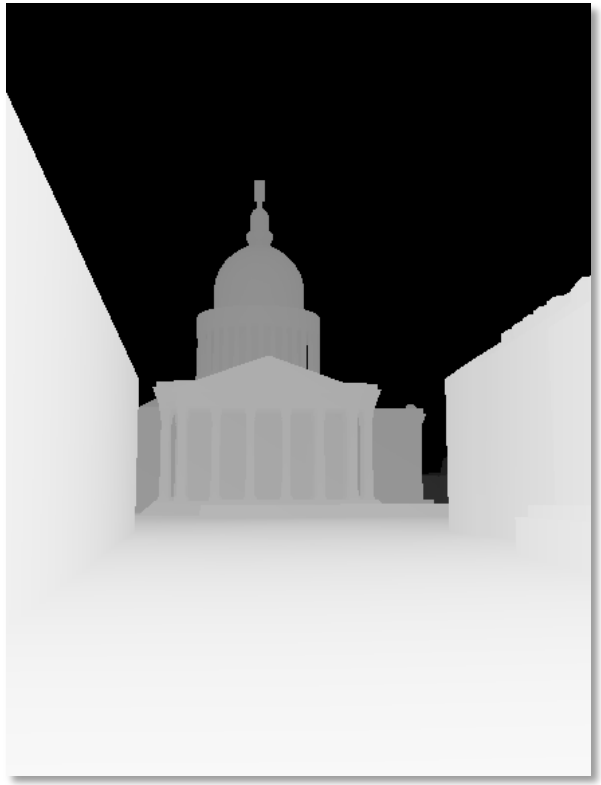
Pine St.

“Pixel-accurate” alignment



3D world model rendered from estimated viewpoint



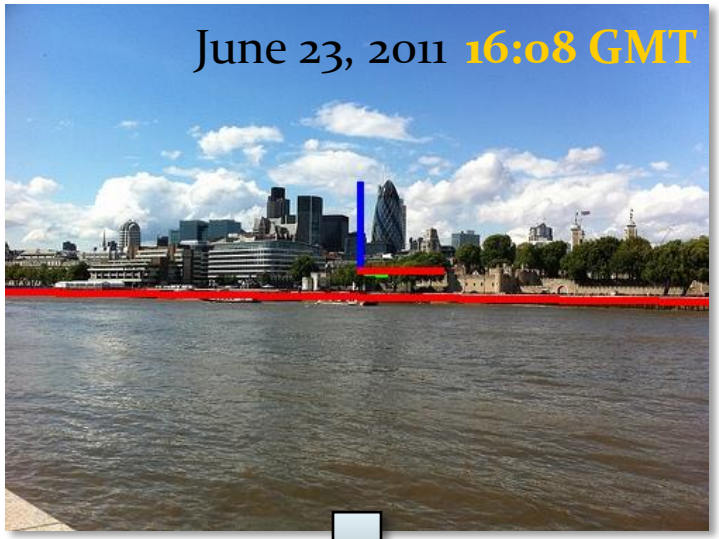


Creates a “stage” for the action to take place in

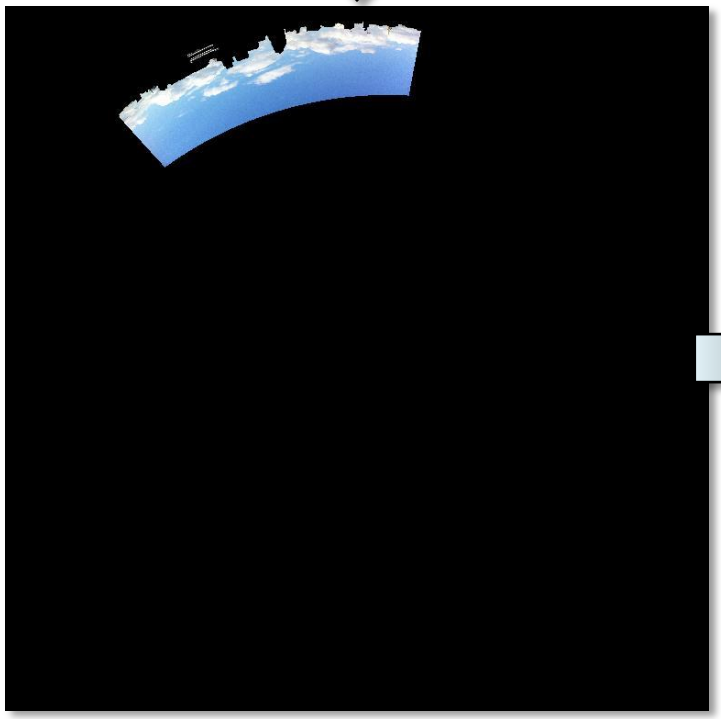
Time

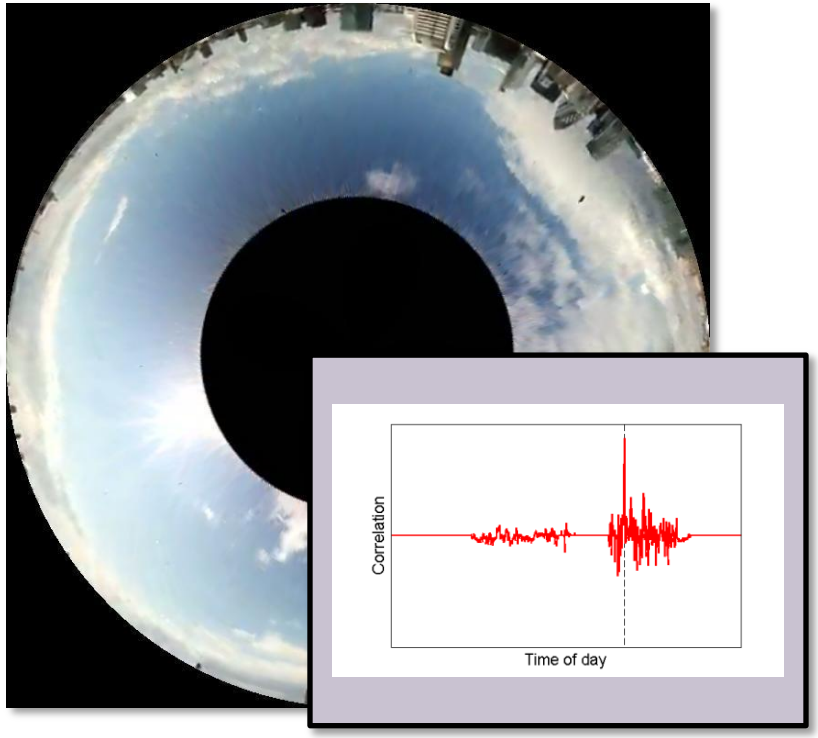
The Monument, London

[credit: Chris Meighs-Andrews]


June 23, 2011 16:08 GMT

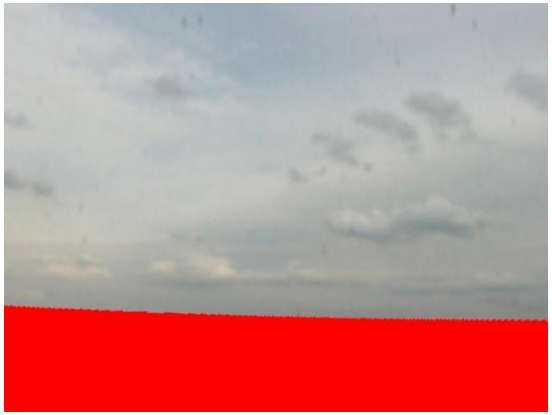

Map
Satellite
London Fenchurch Street
Eastcheap
Tower Gateway
Tower Hill
Lower Thames St
Byward St
A100
Tower of London
The Scoop at More London
Google
Map Data - Terms of Use



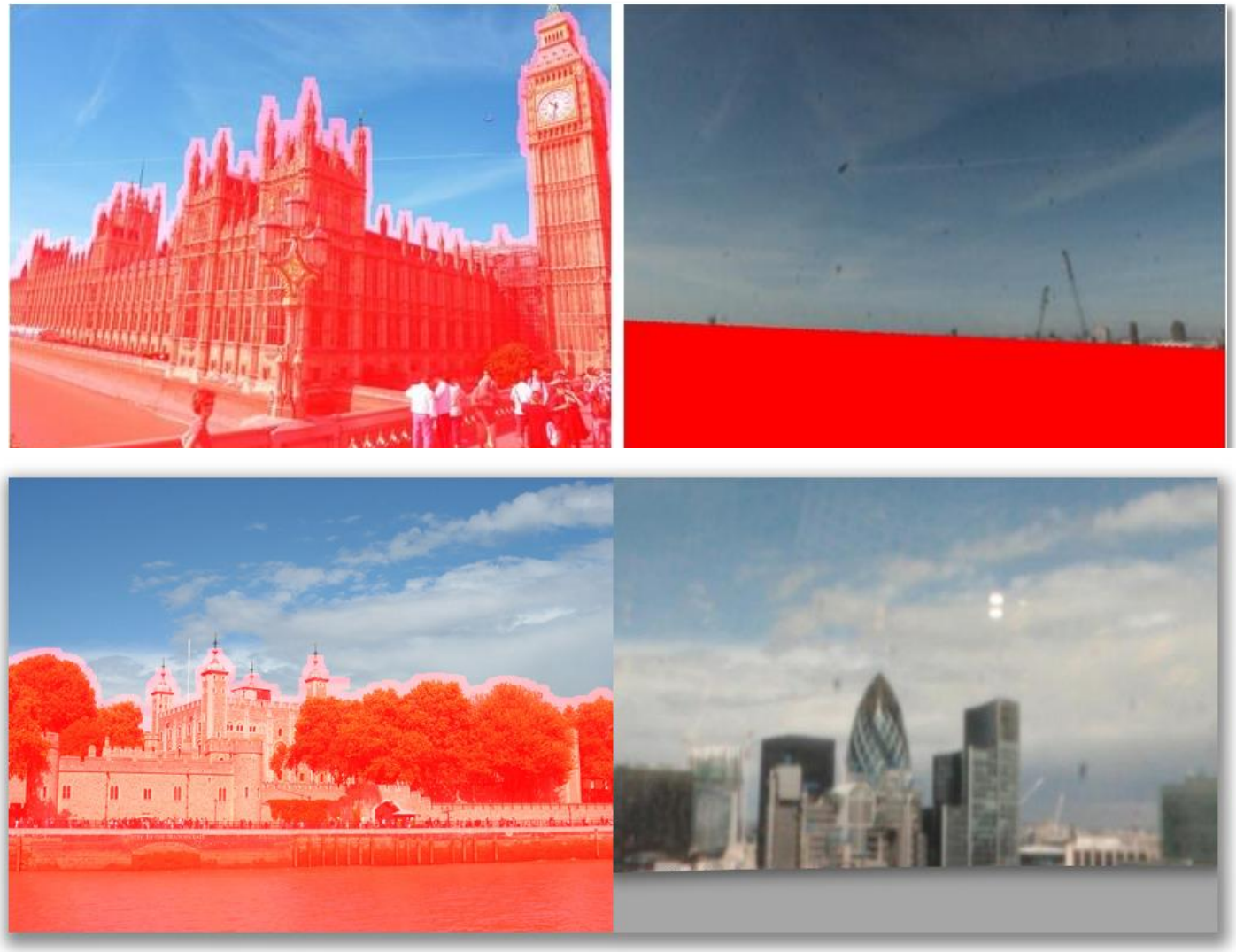

Correlation
Time of day

Input photo

Best matching webcam frame (cropped)



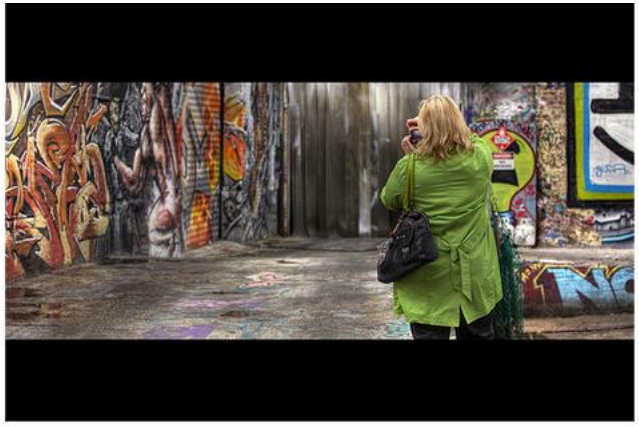
[joint work with Chun-Po Wang]



For images that contain clouds and within ~0.75km of the webcam, we can identify correct time stamps (+- 5 minutes) ~40% of the time

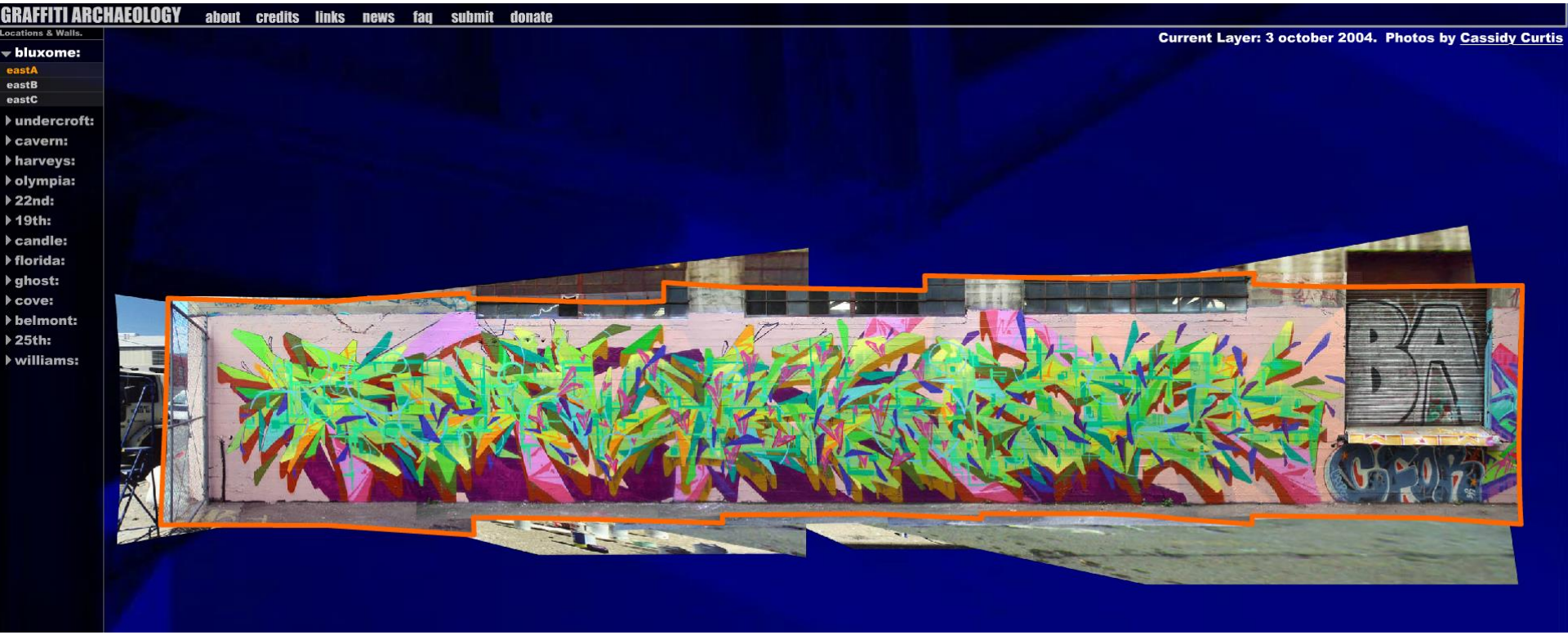
Scenes are dynamic


Times Square

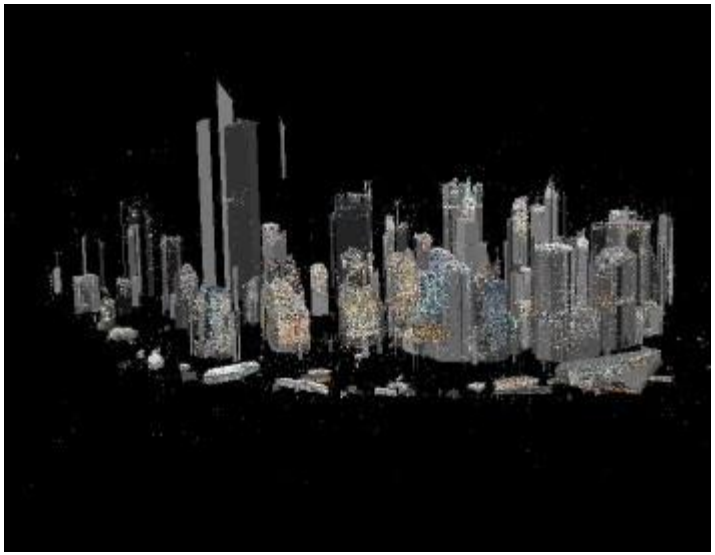


5pointz, Brooklyn

Graffiti Archaeology



4D Cities



[Frank Dellaert, Grant Schindler, et al.]

Scene Chronology

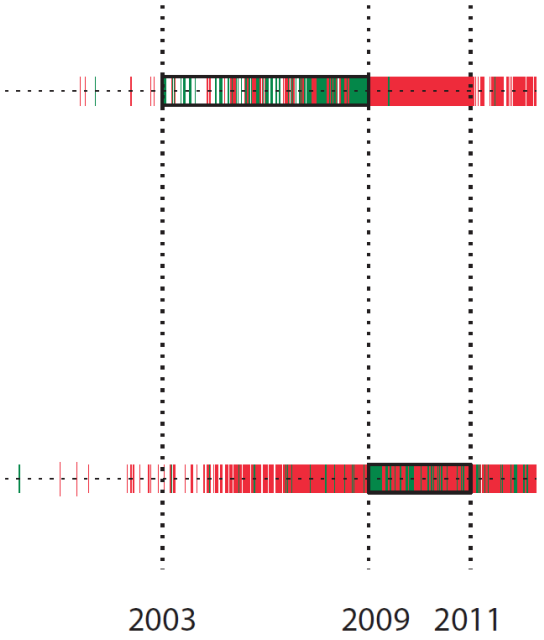
Step 1: Download photos from Flickr

Step 2: Reconstruct a single 3D model with all times mixed up together

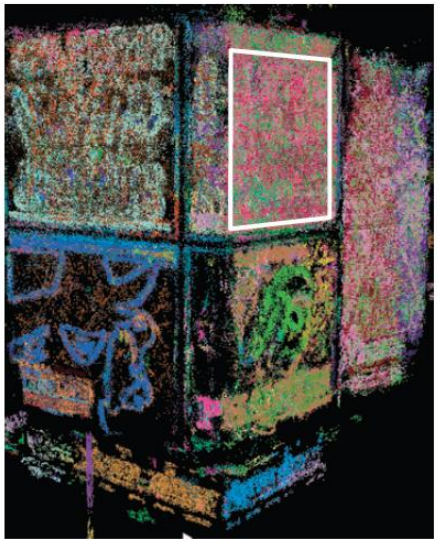
Step 3: Recover the chronology of the scene

Kevin Matzen and Noah Snavely, **Scene Chronology**, ECCV 2014

Single 3D Model (from ~100,000 images)



Per-Point Time Observations



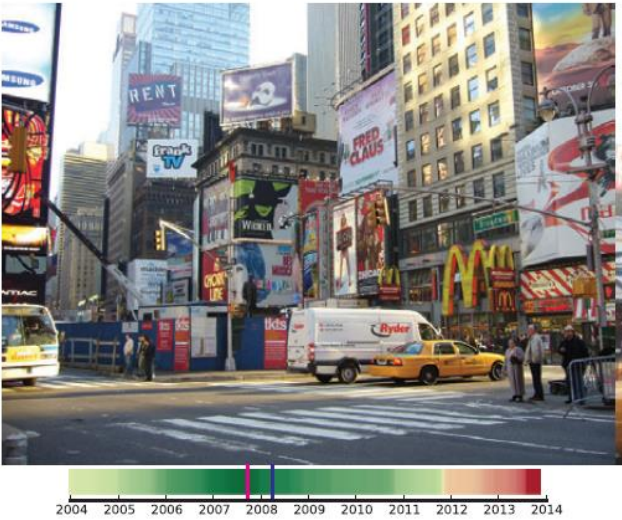
Plane-Time point clustering

Reconstructed structure across time

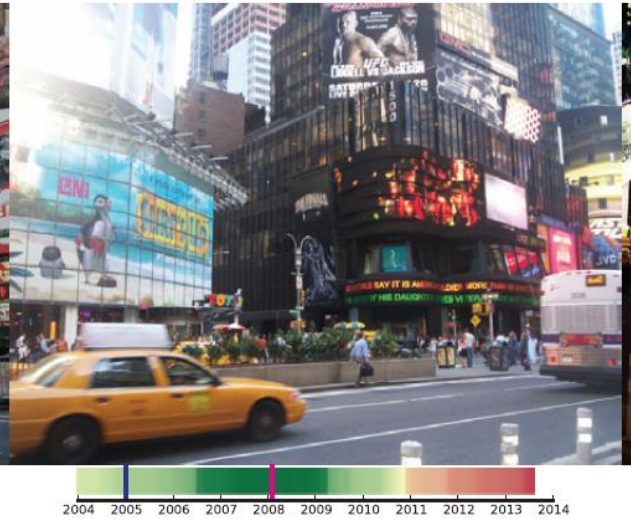
Video

Re-time-stamping

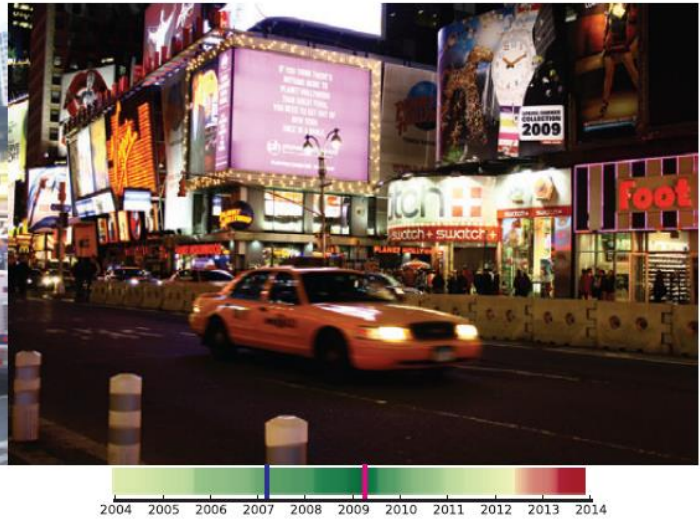
~~March 2008~~ - Late 2007



~~January 2005~~ - Early/Mid 2007



~~April 2007~~ - 2008/2009



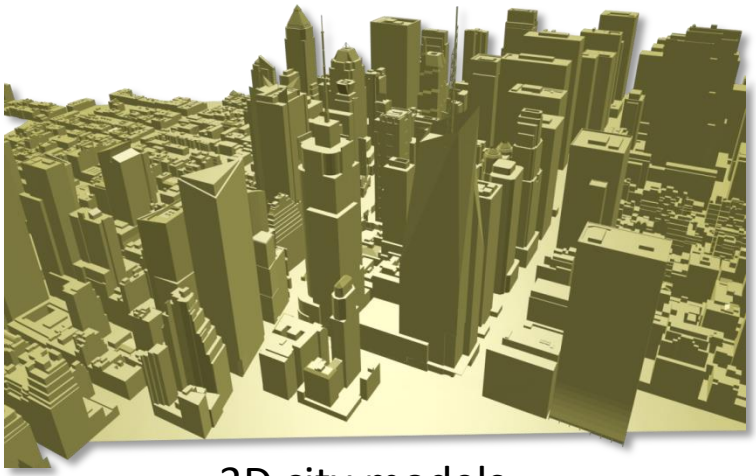
Blue: original timestamp

Red: our predicted timestamp

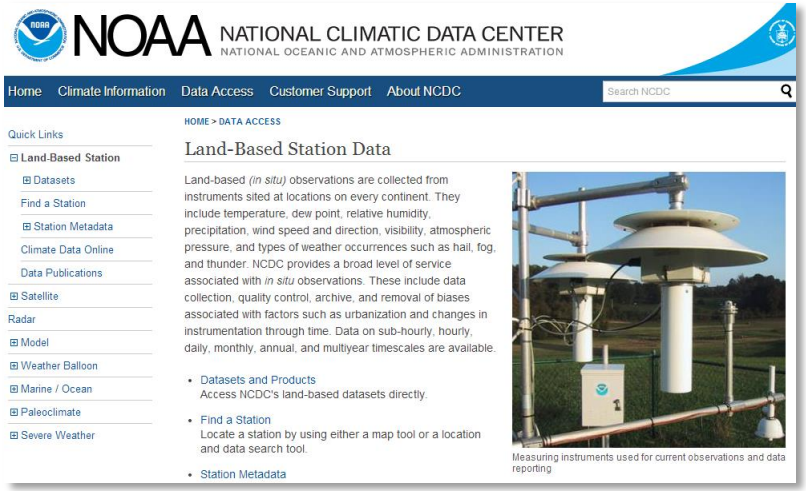
Geography

Grounding vision in the world

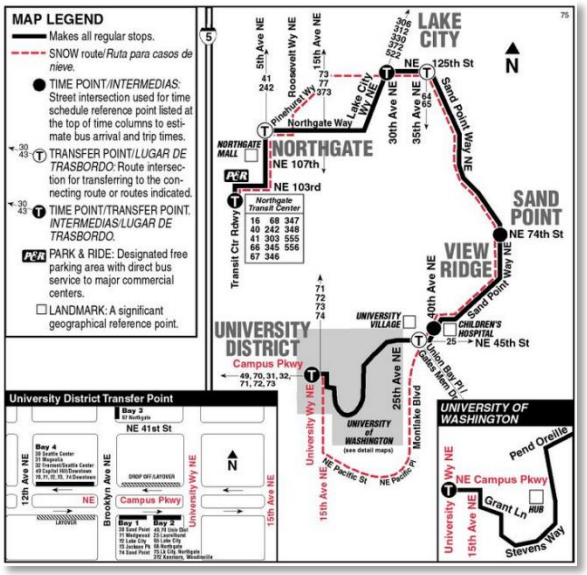
OpenStreetMap



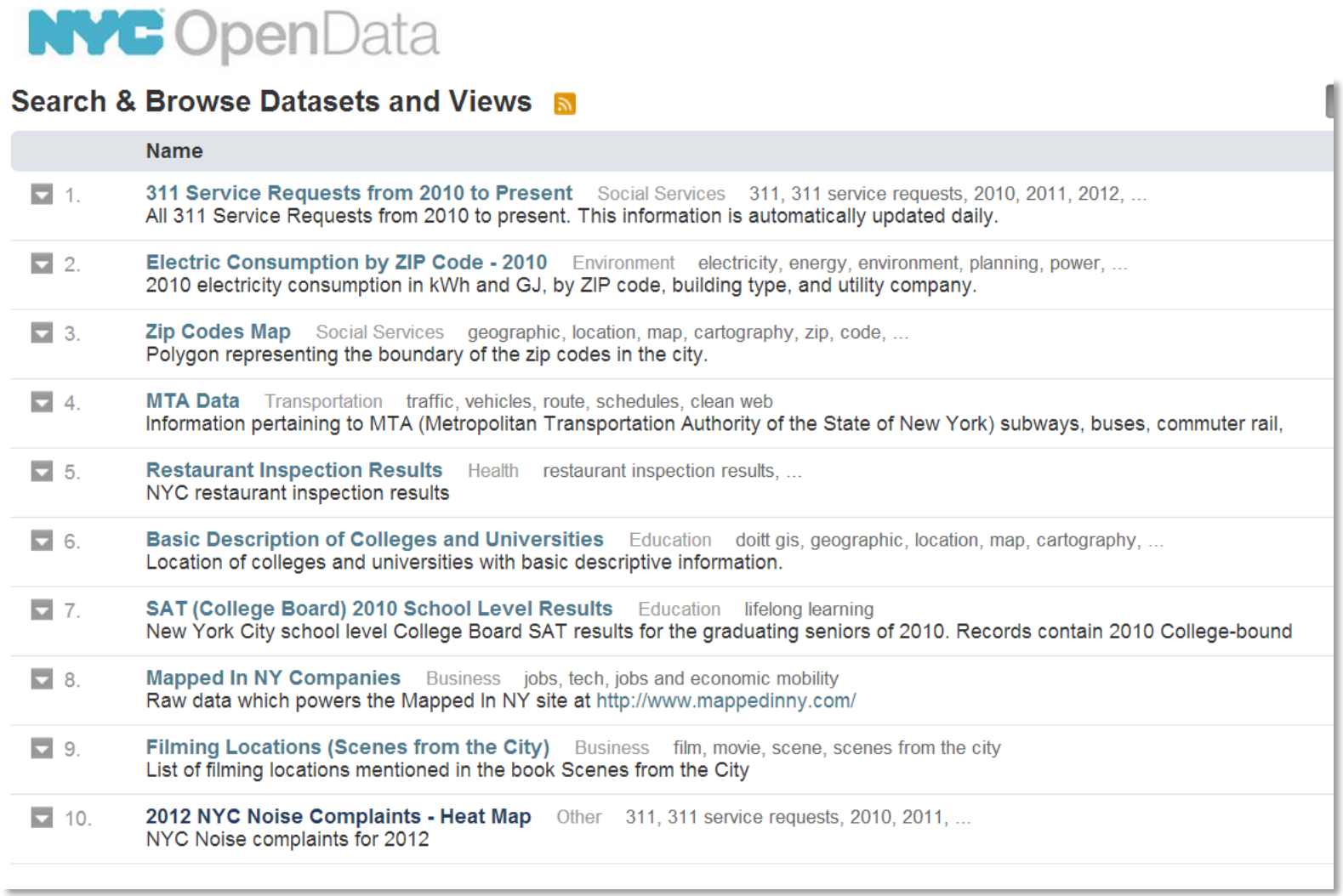
3D city models



Weather data



Bus schedules



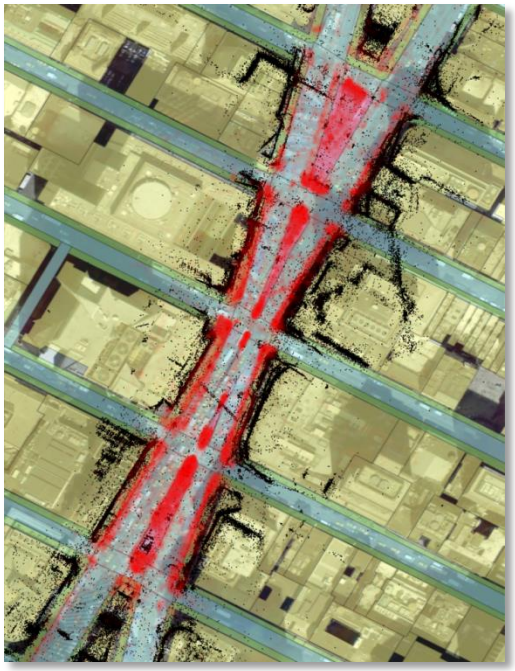
https://nycopendata.socrata.com (https://data.sfgov.org/, https://data.seattle.gov/, …)

Grounding vision in the world

- Which direction is north?
- What is the shape of the buildings?
- What was the weather like?
- Where are streets?
- What is the #51 bus schedule in Rome?

Goal: Integrate images into this ecosystem of geographic data

First steps: NYC3DCars



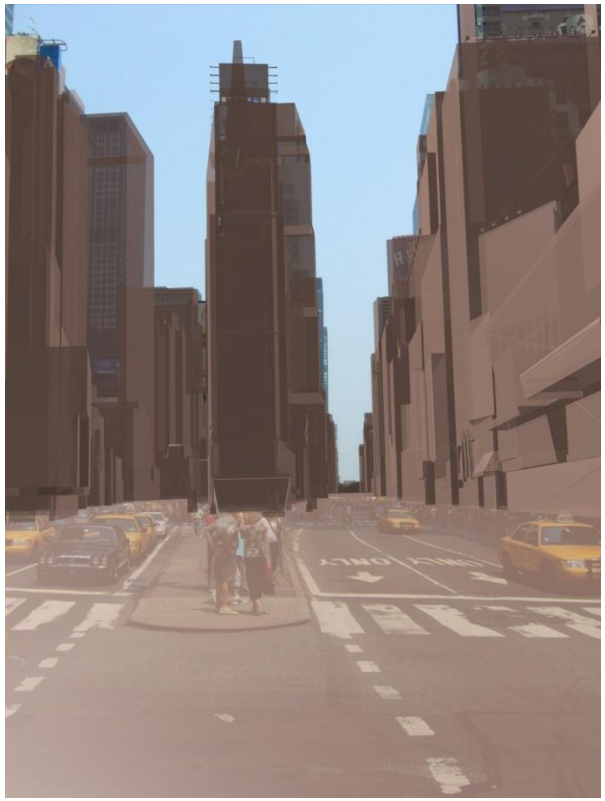
[Kevin Matzen and Noah Snavely, ICCV 2013]

NYCOpenData Roadbeds

Vision grounded in the real world

Input photo

Overlayed GIS data
(roads / sidewalks / medians)

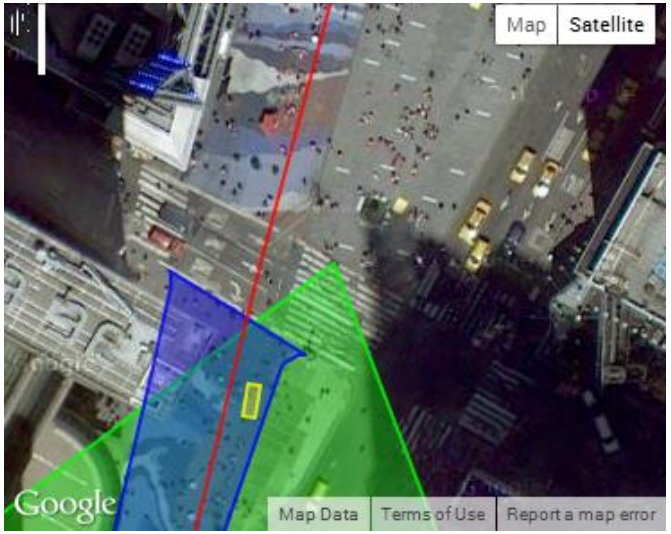


Overlayed Google Earth models

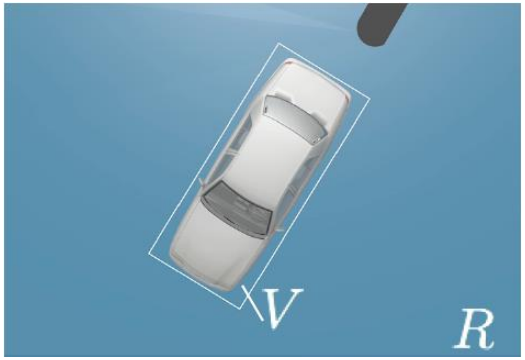
Annotated 3D Vehicles

Video

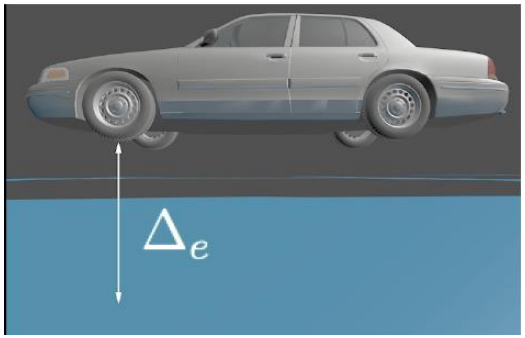
3D Detection



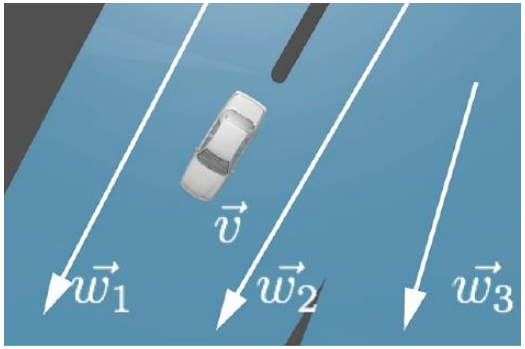
Appearance score



Ground coverage score

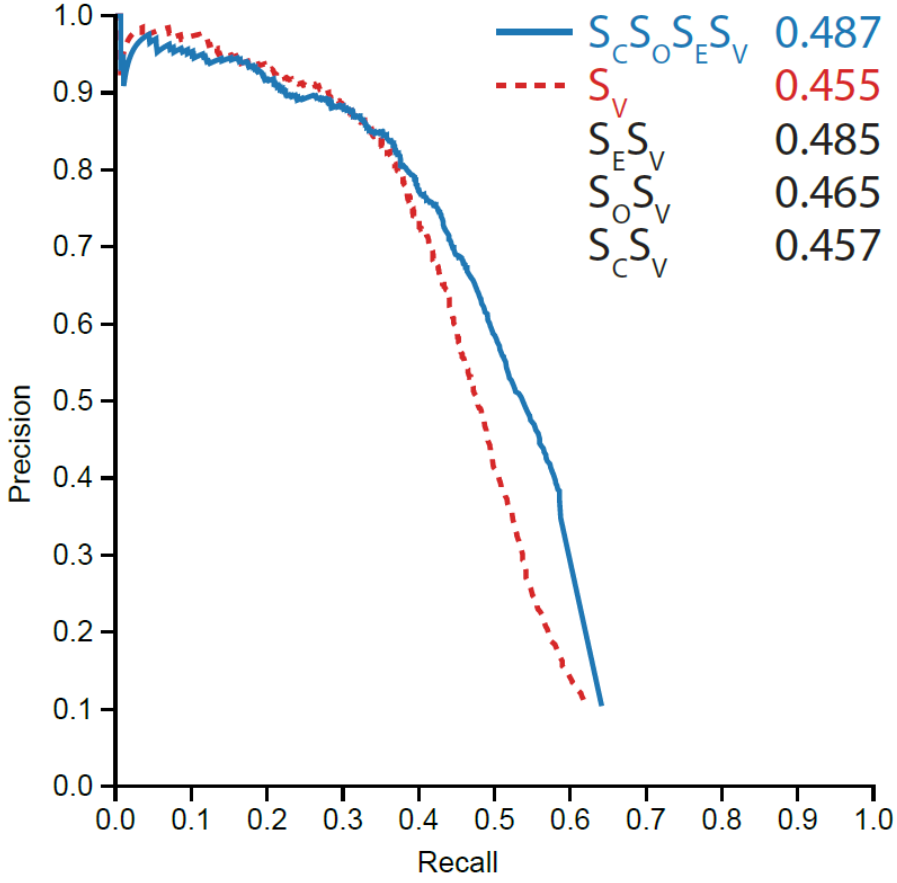


Elevation score

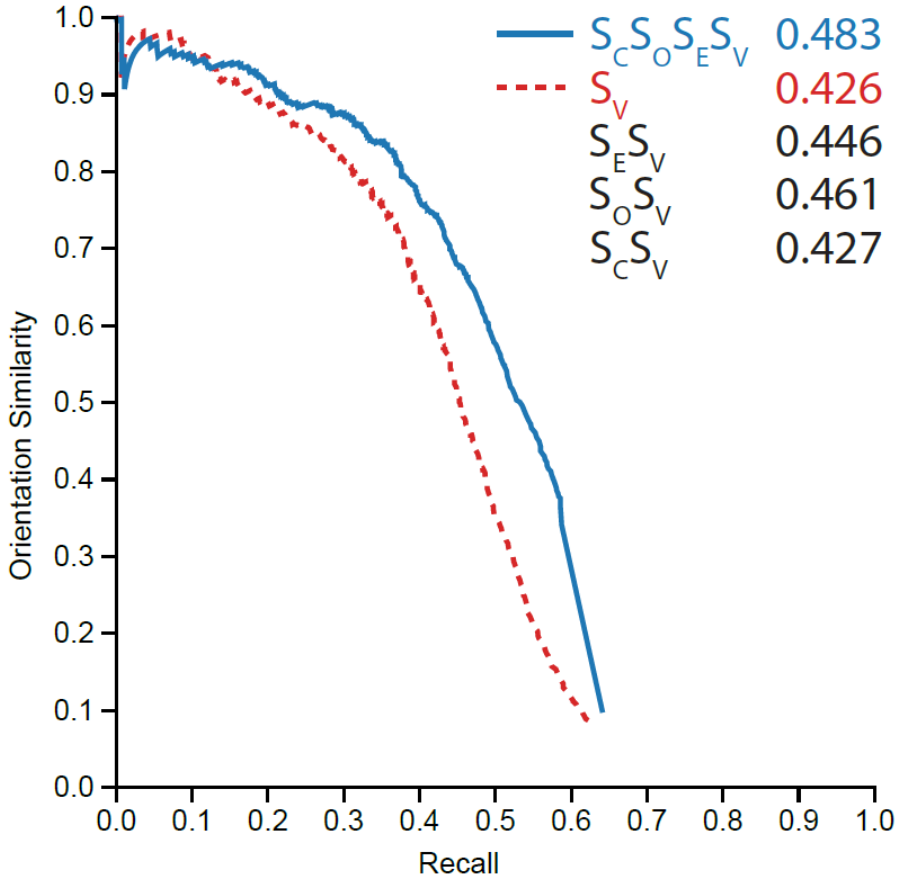


3D orientation score

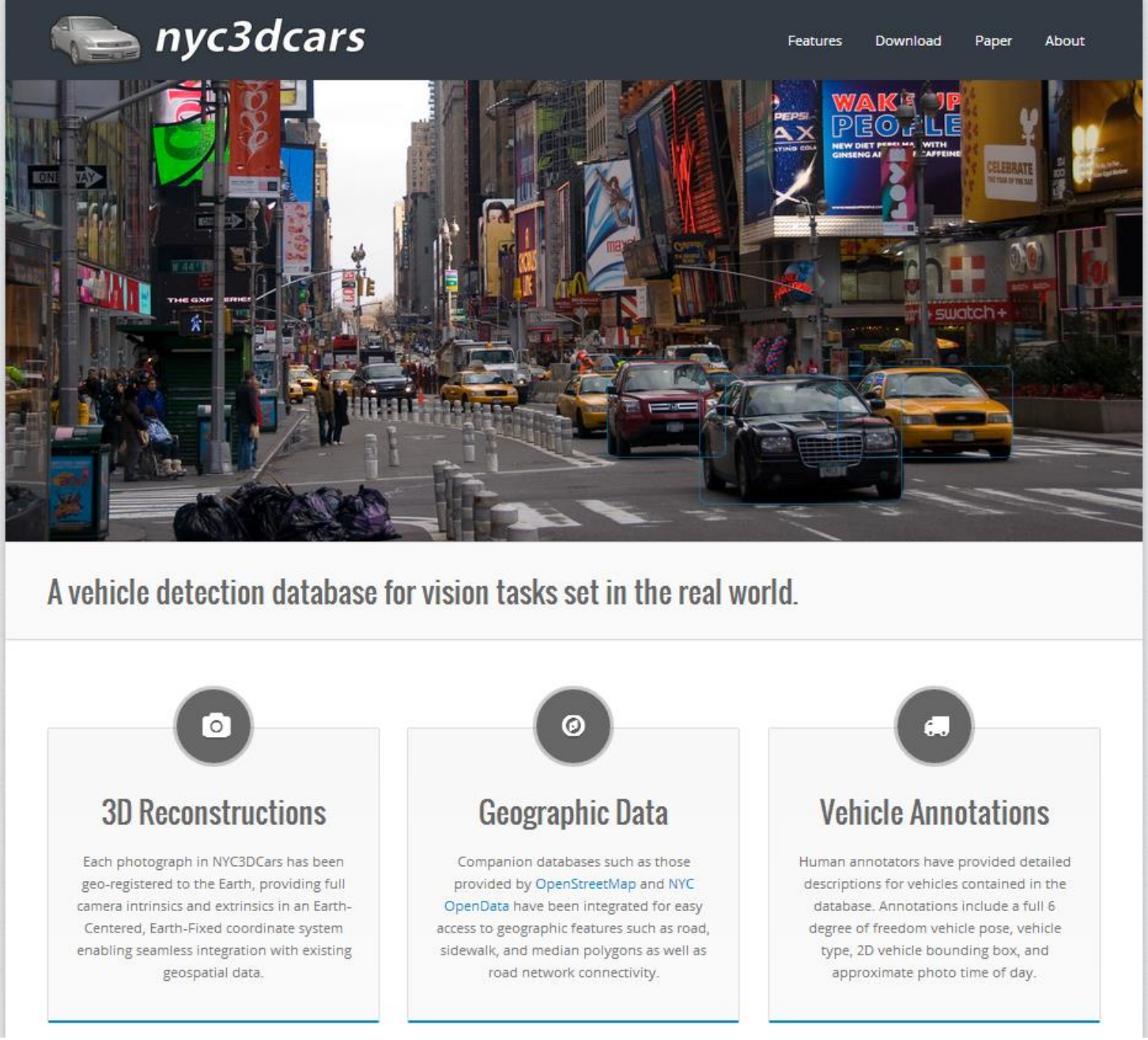
Results



Precision / Recall



Orientation similarity / Recall



http://nyc3d.cs.cornell.edu/

Summary

- Massive image collections are a rich source of information about our world
- In vision we should be leveraging all available data: images, depth, **geographic context**
- Lots of interesting challenges

Acknowledgements

- Sean Bell
- Daniel Cabrini Hauagge
- Kevin Matzen
- Andrew Owens
- Chun-Po Wang
- Kyle Wilson

- Yunpeng Li
- Dan Huttenlocher
- David Crandall
- Kavita Bala

- National Science Foundation
- Intel Center for Science and Technology – Visual Computing
- Amazon AWS for Education

Thank you!

More information at
http://www.cs.cornell.edu/~snavely/