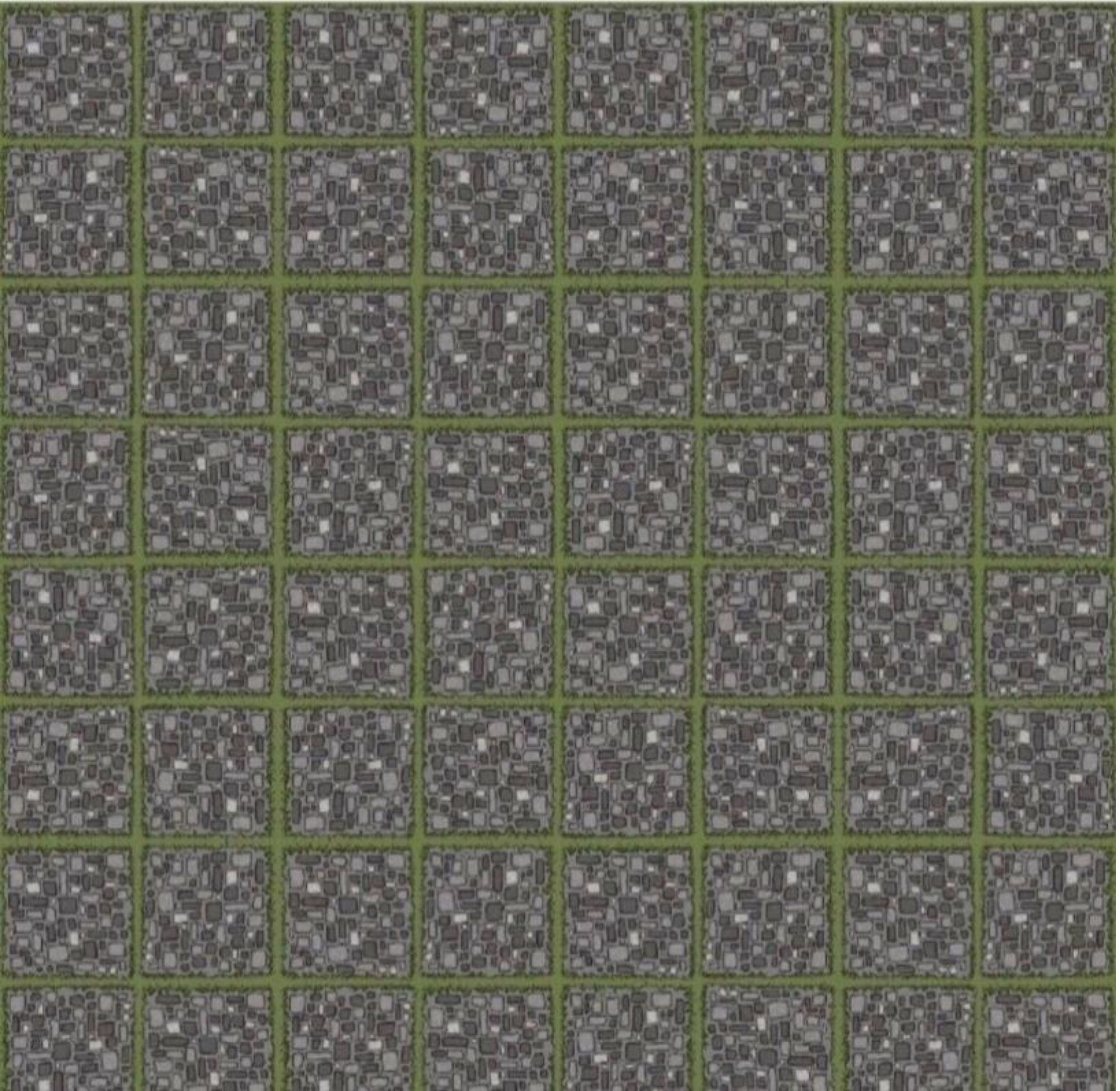
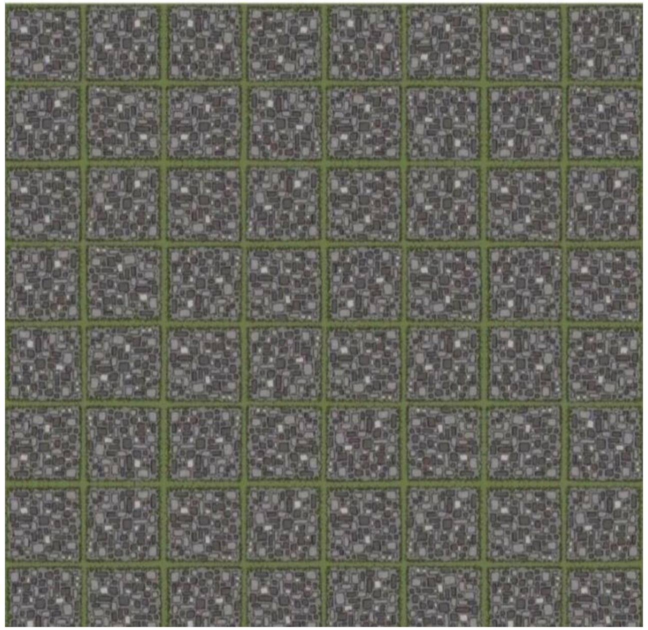
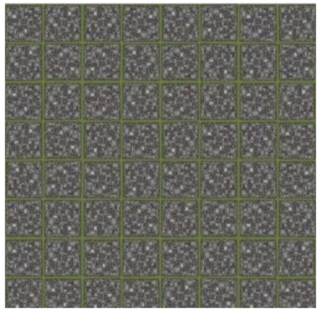
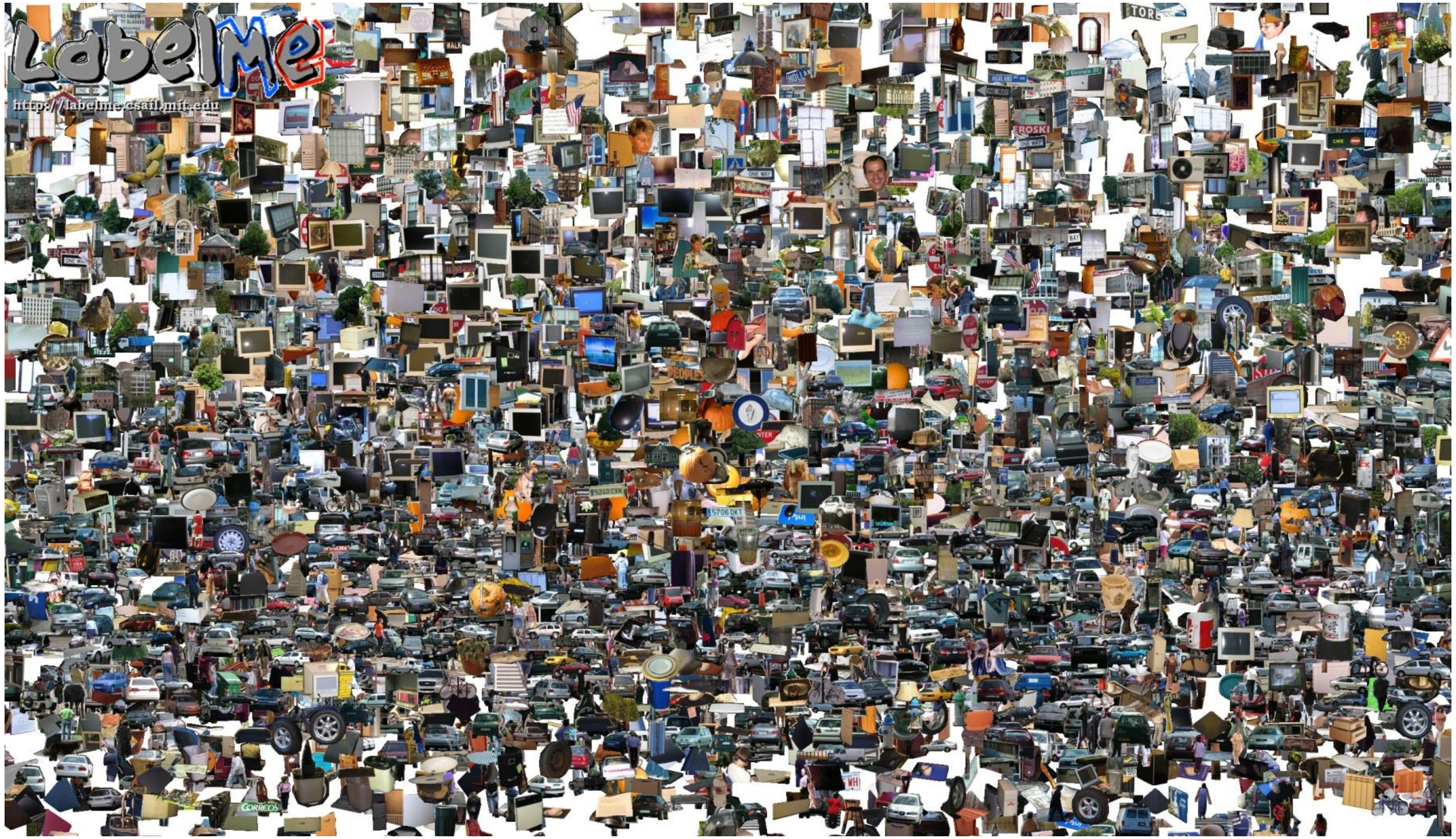






Big Data: Opportunities of Scale



Computer Vision

James Hays

Many slides from James Hays, Alyosha Efros, and Derek Hoiem

Graphic from Antonio Torralba

Outline

Opportunities of Scale: Data-driven methods

- The Unreasonable Effectiveness of Data
- Scene Completion
- Image Geolocation

Computer Vision Class so far

- The geometry of image formation
  - Ancient / Renaissance
- Signal processing / Convolution
  - 1800s, but really the 50's and 60's
- Hand-designed Features for recognition, either instance-level or categorical
  - 1999 (SIFT), 2003 (Video Google), 2005 (Dalal-Triggs), 2006 (spatial pyramid bag of words)
- Learning from Data
  - 1991 (EigenFaces) but late 90's to now especially

What has changed in the last 20 years?

- The Internet
- Crowdsourcing
- Learning representations from the data these sources provide (deep learning)
- The inevitable Moore's-law-esque increase in compute that allows large scale deep learning

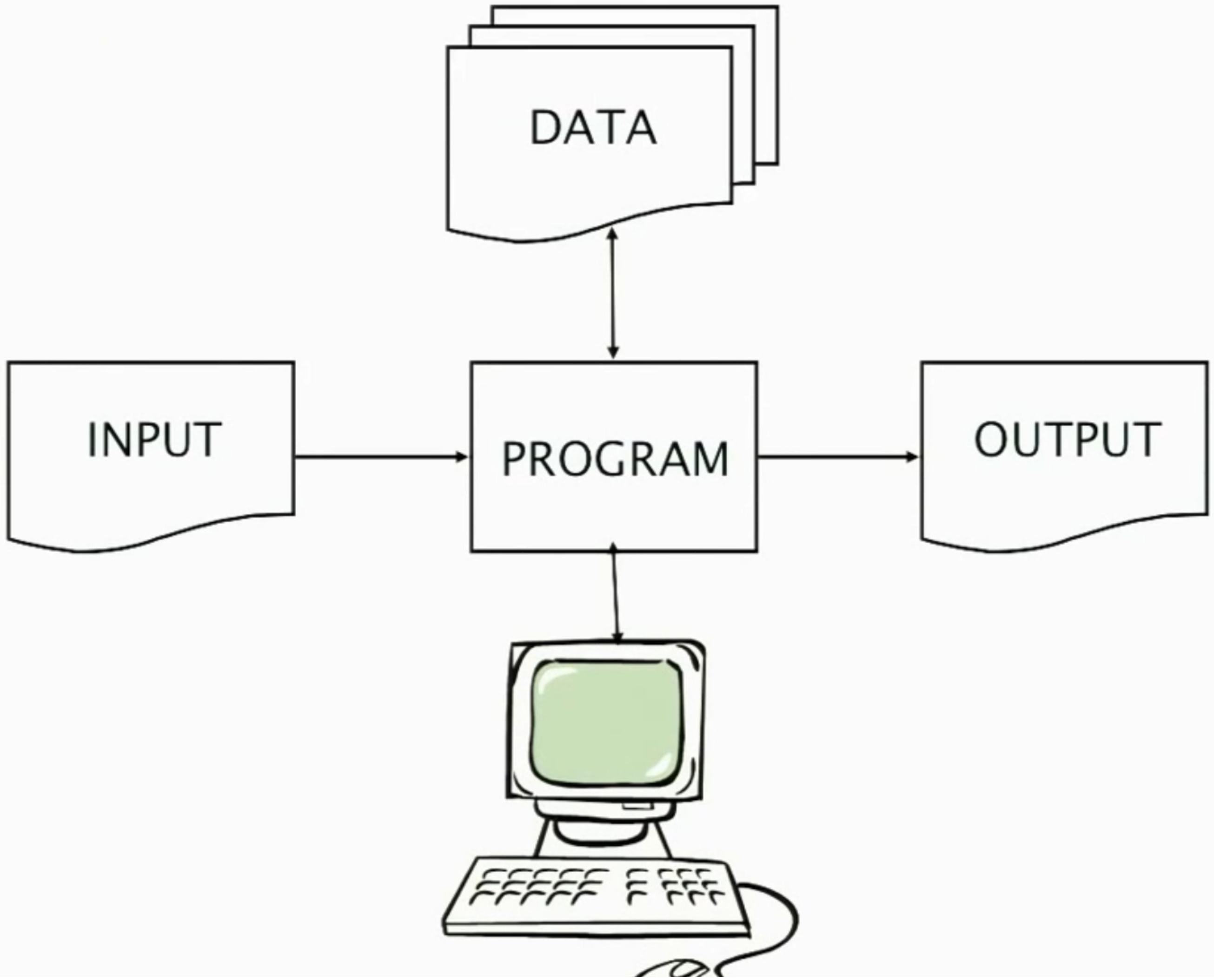
Google and massive data-driven algorithms

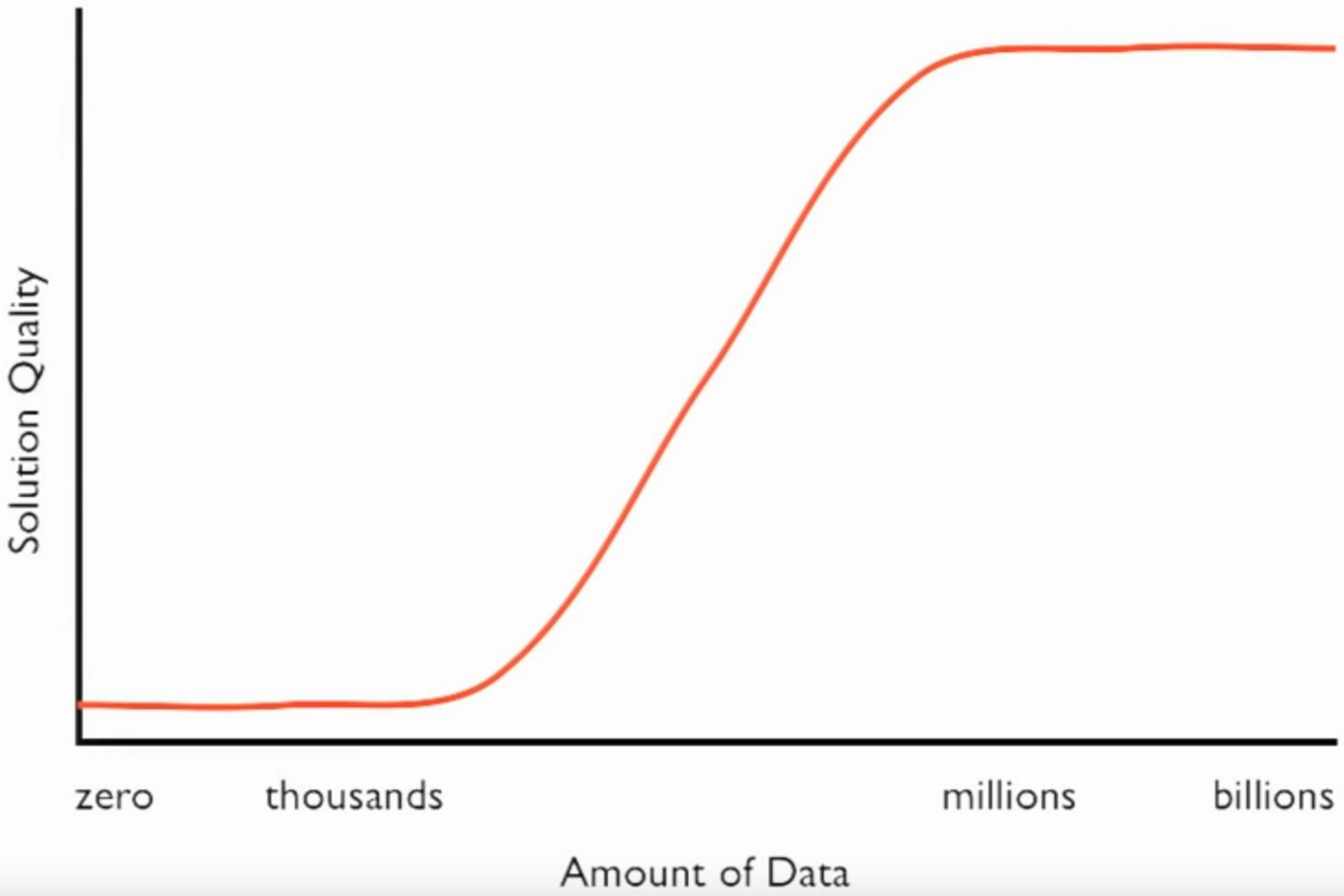
A.I. for the postmodern world:

– all questions have already been answered…many times, in many ways

– Google is dumb, the "intelligence" is in the data

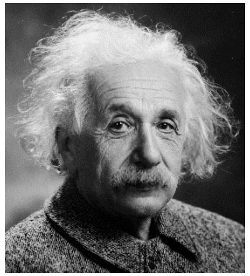
https://youtu.be/yvDCzhbjYWs?t=24
Watch until 9:42


DATA
INPUT
PROGRAM
OUTPUT


Solution Quality
zero
thousands
millions
billions
Amount of Data

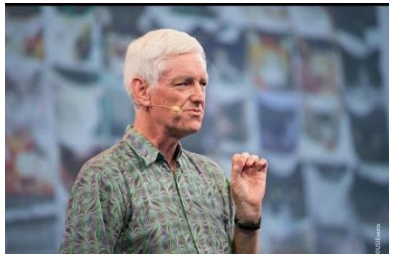
The Unreasonable Effectiveness of Math

- **“**The miracle of the appropriateness of the language of mathematics…” **Eugene Wigner**



- “The most incomprehensible thing about the universe is that it is comprehensible.” **Albert Einstein**

- “There is only one thing which is more unreasonable than the unreasonable effectiveness of mathematics in physics, and this is the unreasonable ineffectiveness of mathematics in biology.” **Israel Gelfand**



- “We should stop acting as if our goal is to author extremely elegant theories, and instead embrace complexity and make use of the best ally we have: the unreasonable effectiveness of data.” **Peter Norvig**

Chinese Room, John Searle (1980)

If a machine can convincingly simulate an intelligent conversation, does it necessarily understand? In the experiment, Searle imagines himself in a room, acting as a computer by manually executing a program that convincingly simulates the behavior of a native Chinese speaker.

Most of the discussion consists of attempts to refute it. "The overwhelming majority," notes *BBS* editor Stevan Harnad,“ still think that the Chinese Room Argument is dead wrong." The sheer volume of the literature that has grown up around it inspired Pat Hayes to quip that the field of cognitive science ought to be redefined as "the ongoing research program of showing Searle's Chinese Room Argument to be false.

Yann LeCun

October 23 at 9:58pm ·

Questions from the piece:

Q1. Does the Chinese Room argument prove the impossibility of machine consciousness?
A1: Hell no. ... See More

Can Machines Become Moral?

The question is heard more and more often, both from those who think that machines cannot become moral, and who think that to believe otherwise is a dangerous illusion, and from those who think that machines must become moral,...

BIGQUESTIONSONLINE.COM | BY DON HOWARD

You and 156 others 30 Comments 20 Shares

Like Comment Share

Big Idea

- Do we need computer vision systems to have strong AI-like reasoning about our world?
- What if invariance / generalization isn’t actually the core difficulty of computer vision?
- What if we can perform high level reasoning with brute-force, data-driven algorithms?

Scene Completion

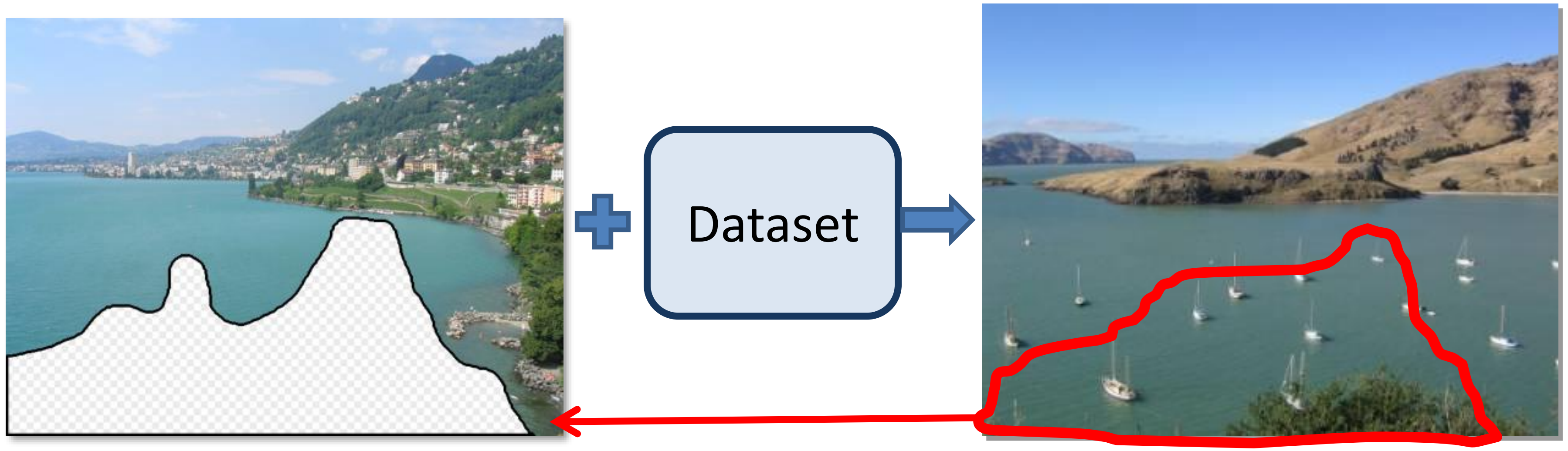
[Hays and Efros. Scene Completion Using Millions of Photographs. SIGGRAPH 2007 and CACM October 2008.]

Selected as one of SIGGRAPH's "Seminal papers" in 2023.

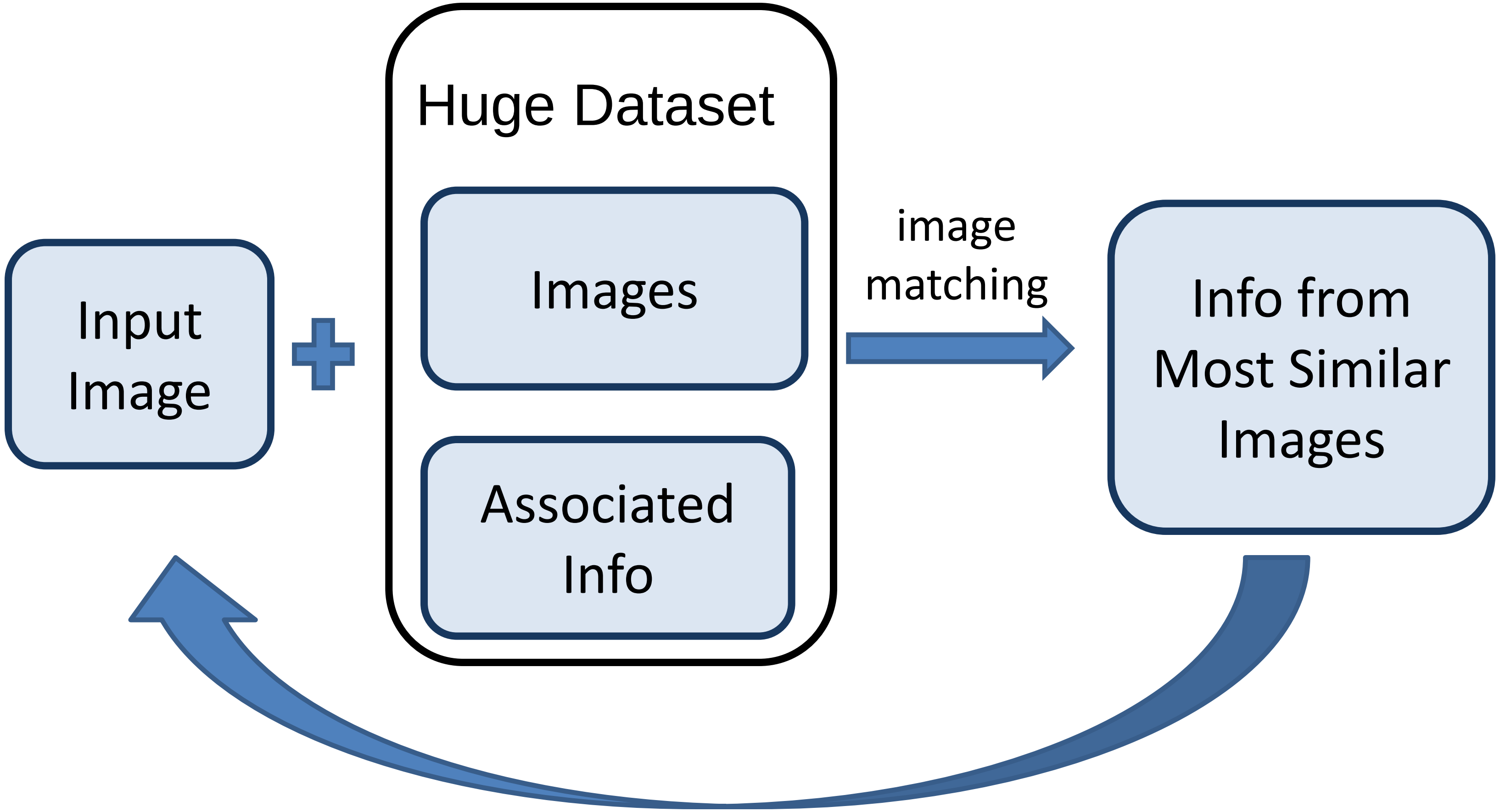
http://graphics.cs.cmu.edu/projects/scene-completion/

How it works

- Find a similar image from a large dataset
- Blend a region from that image into the hole

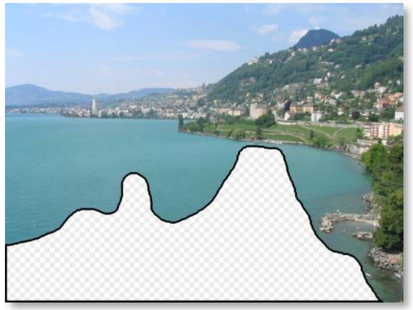


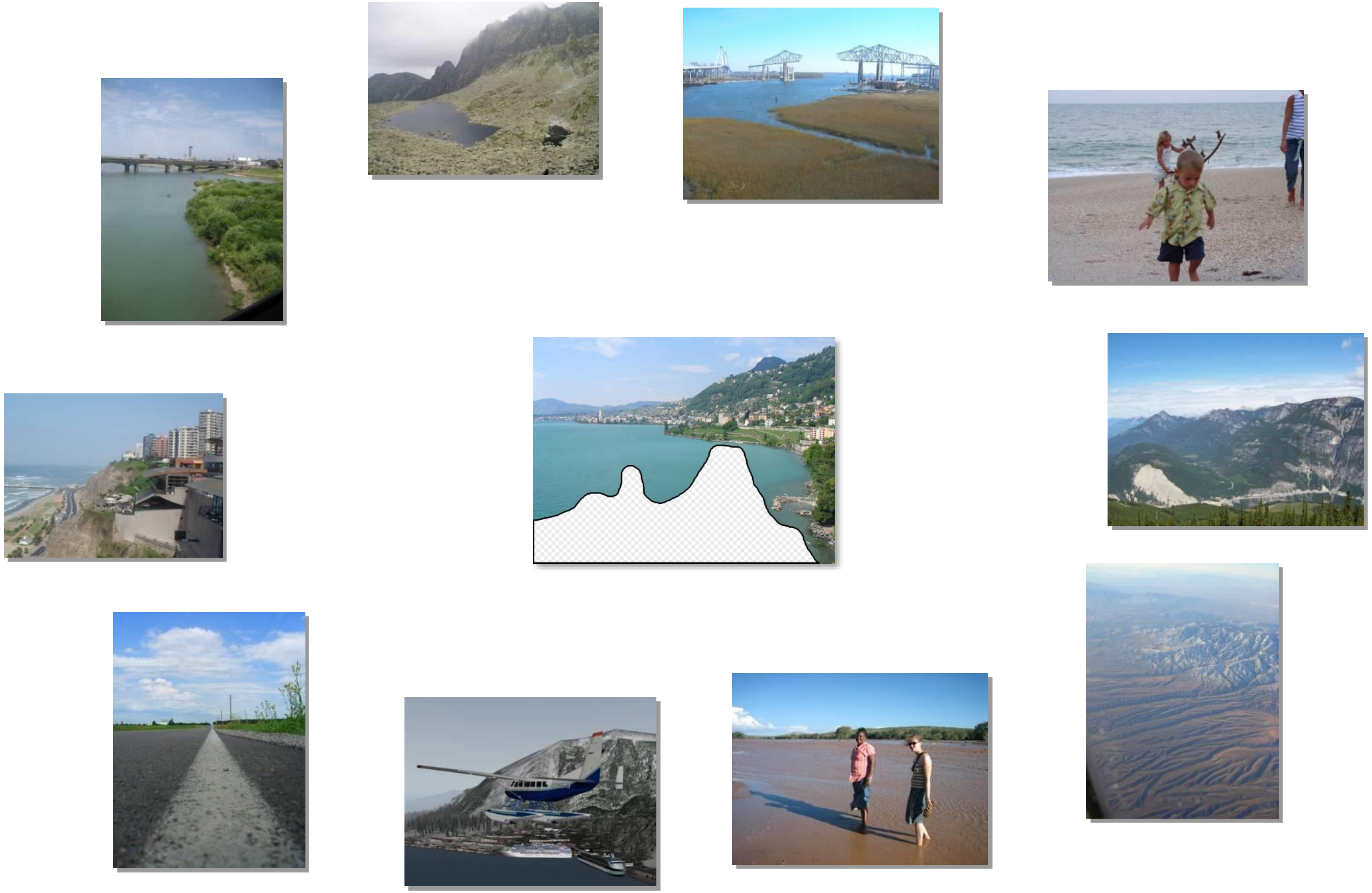
General Principal



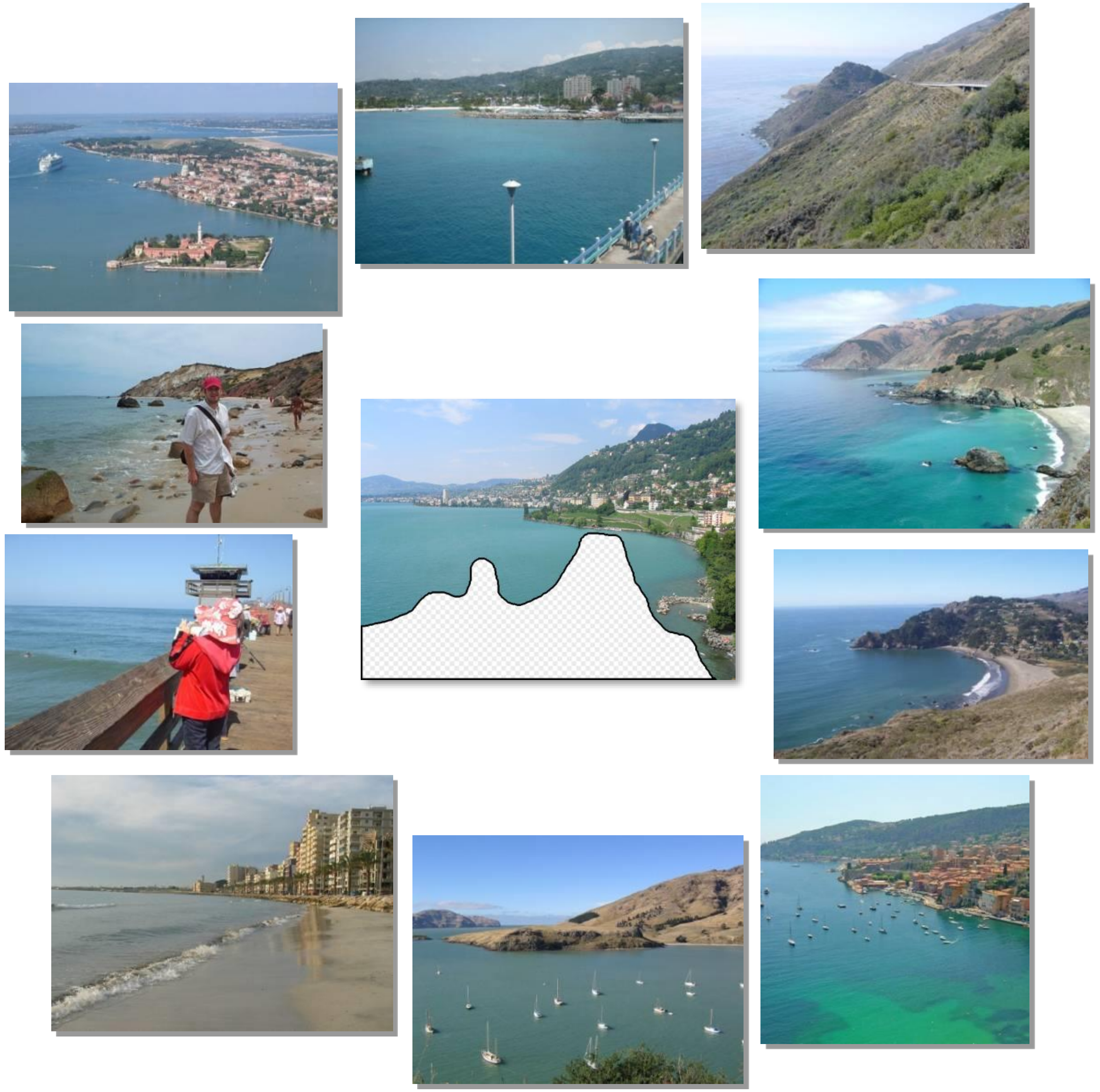
Hopefully, If you have enough images, the dataset will contain very similar images that you can find with simple matching methods.

How many images is enough?





Nearest neighbors from a
collection of 20 thousand images



Nearest neighbors from a collection of 2 million images

Image Data on the Internet

- Flickr (as of Sept. 19th, 2010)
  - 5 billion photographs
  - 100+ million geotagged images
- Facebook (as of 2009)
  - 15 billion

http://royal.pingdom.com/2010/01/22/internet-2009-in-numbers/

Image Data on the Internet

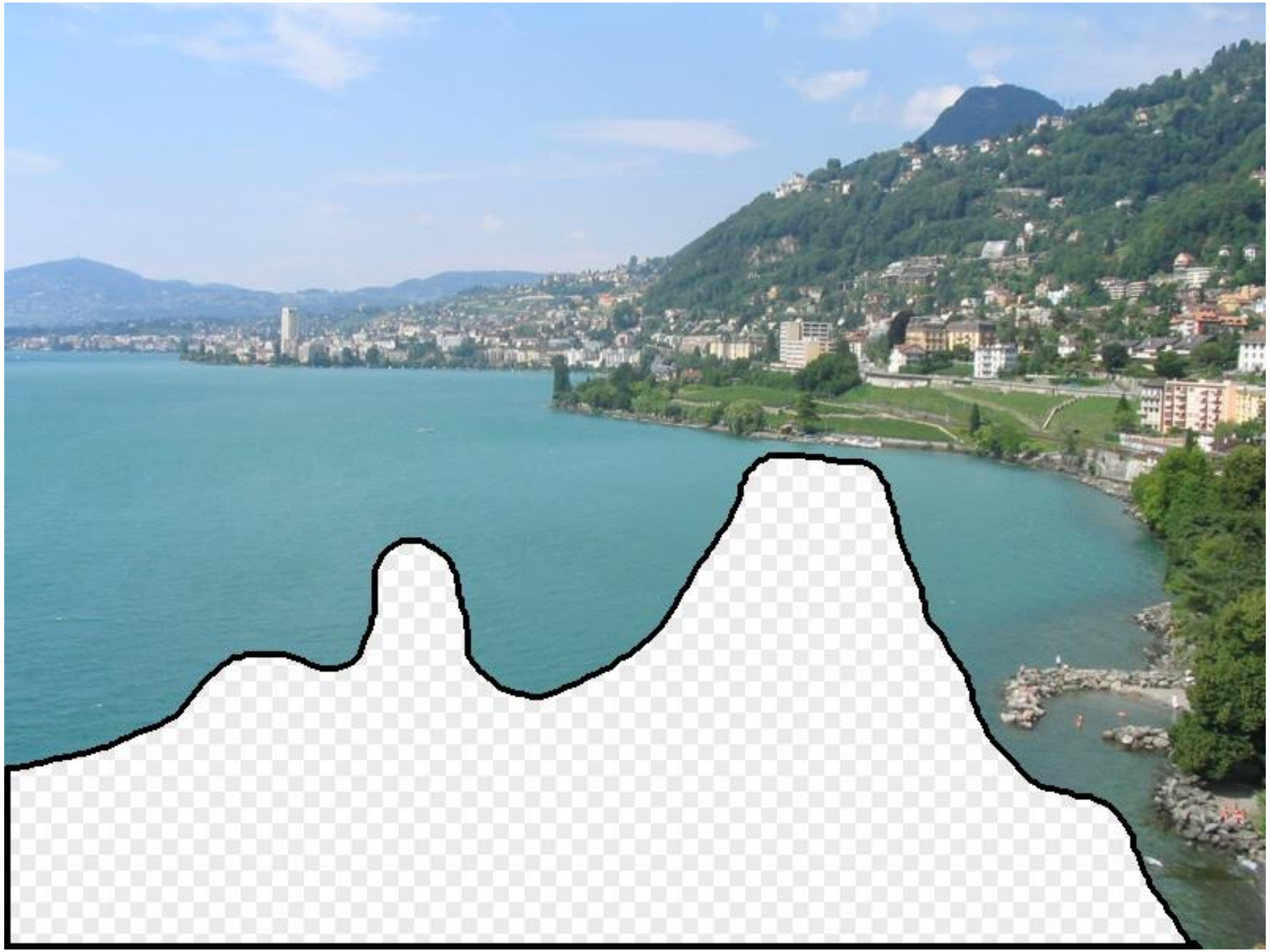
- Flickr (as of Nov 2013)
  - 10 billion photographs
  - 100+ million geotagged images
  - 3.5 million a day
- Facebook (as of Sept 2013)
  - 250 billion+
  - 300 million a day
- Instagram
  - 55 million a day

Scene Completion: how it works

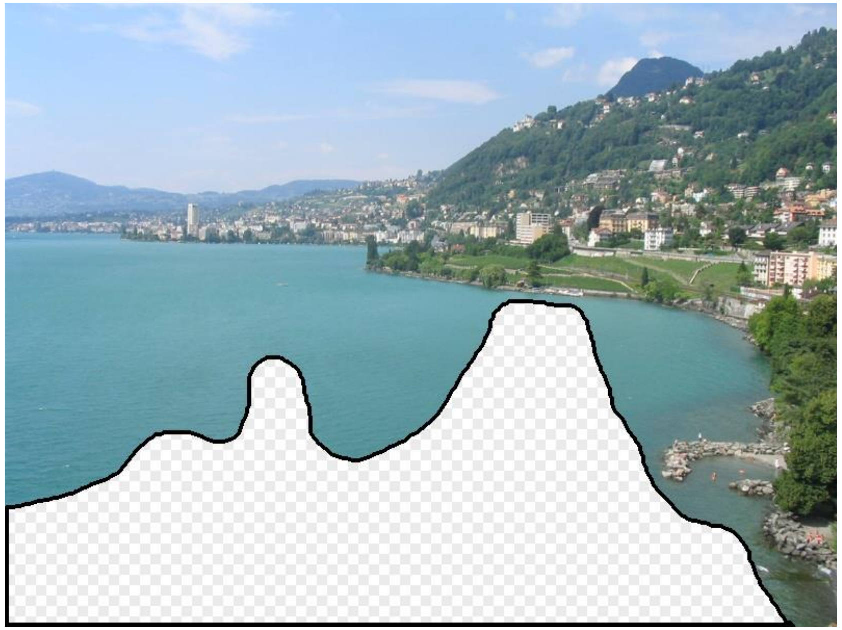
[Hays and Efros. Scene Completion Using Millions of Photographs. SIGGRAPH 2007 and CACM October 2008.]

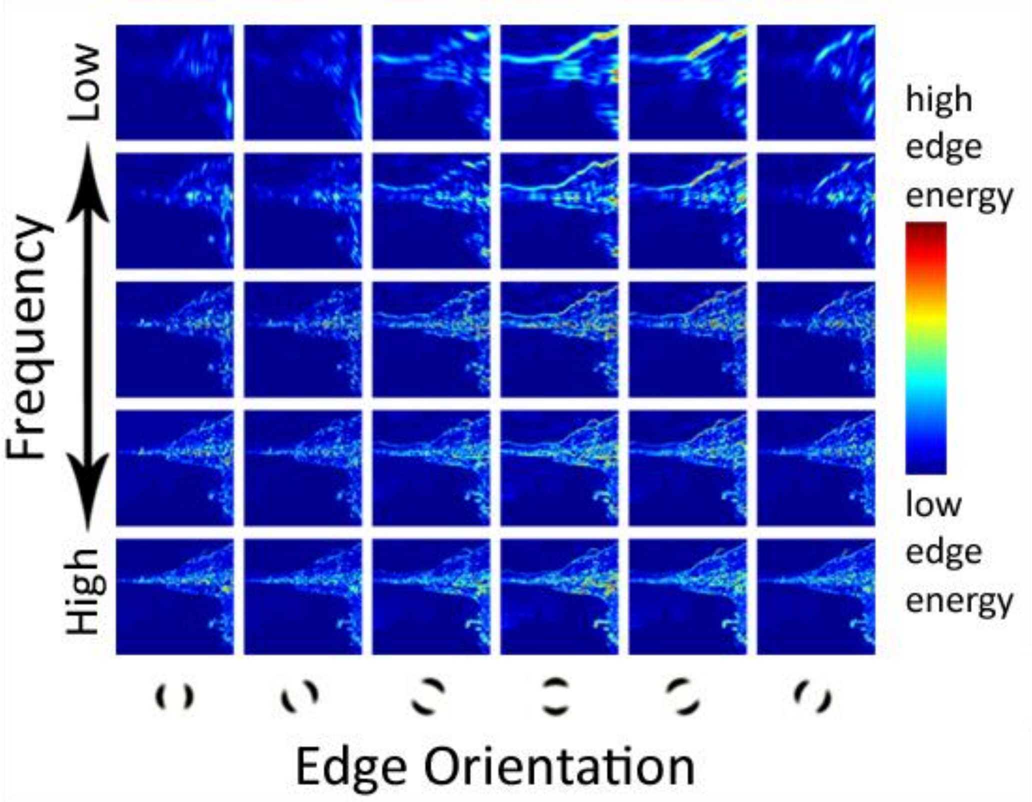
The Algorithm

Scene Matching

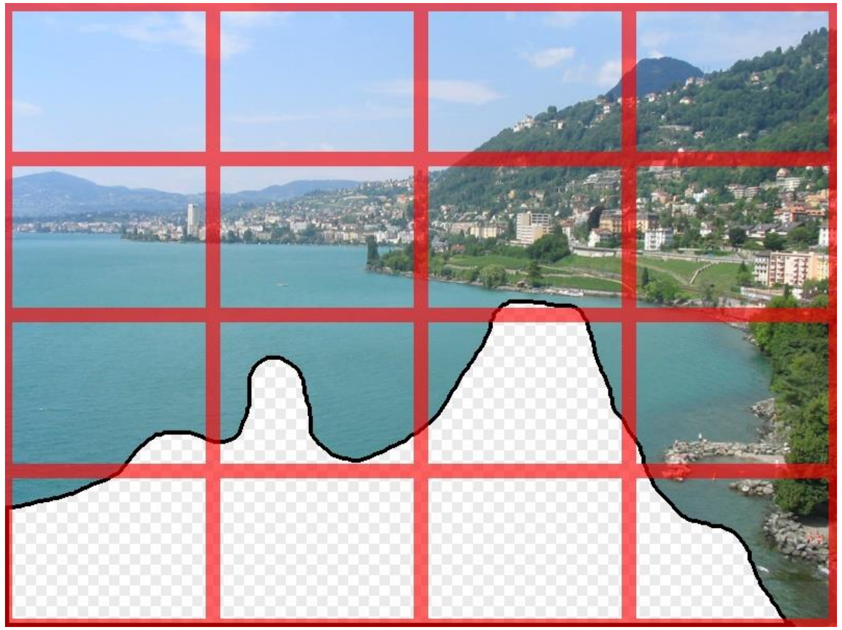


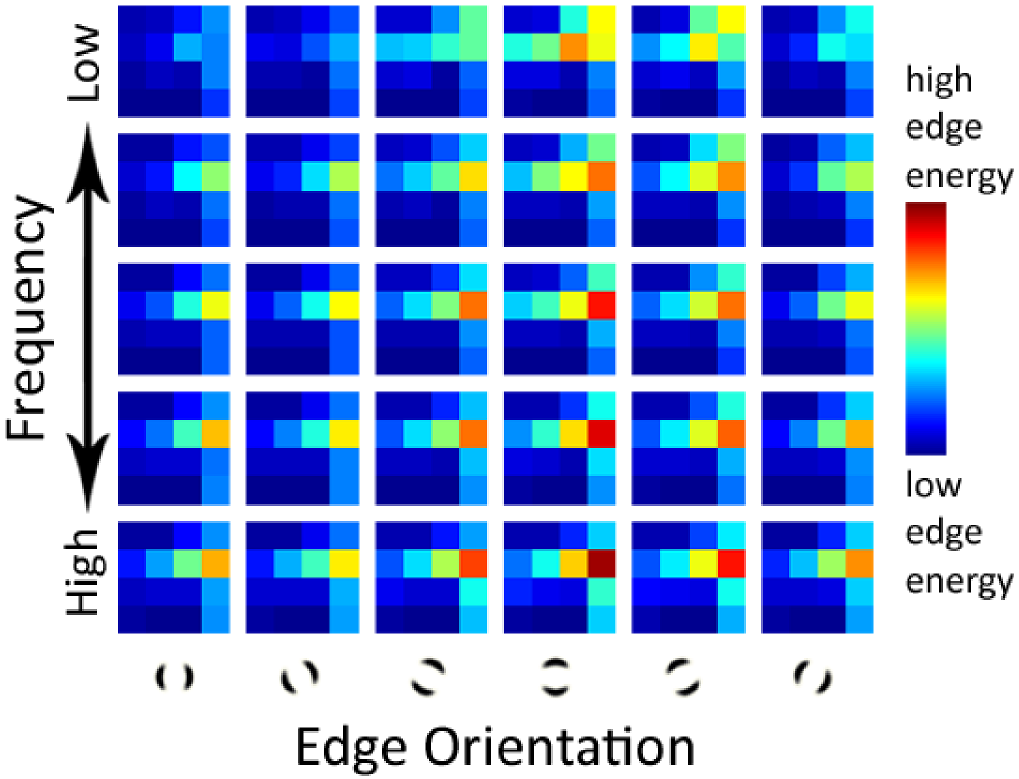
Scene Descriptor





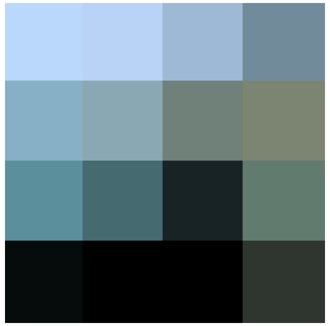
Scene Descriptor

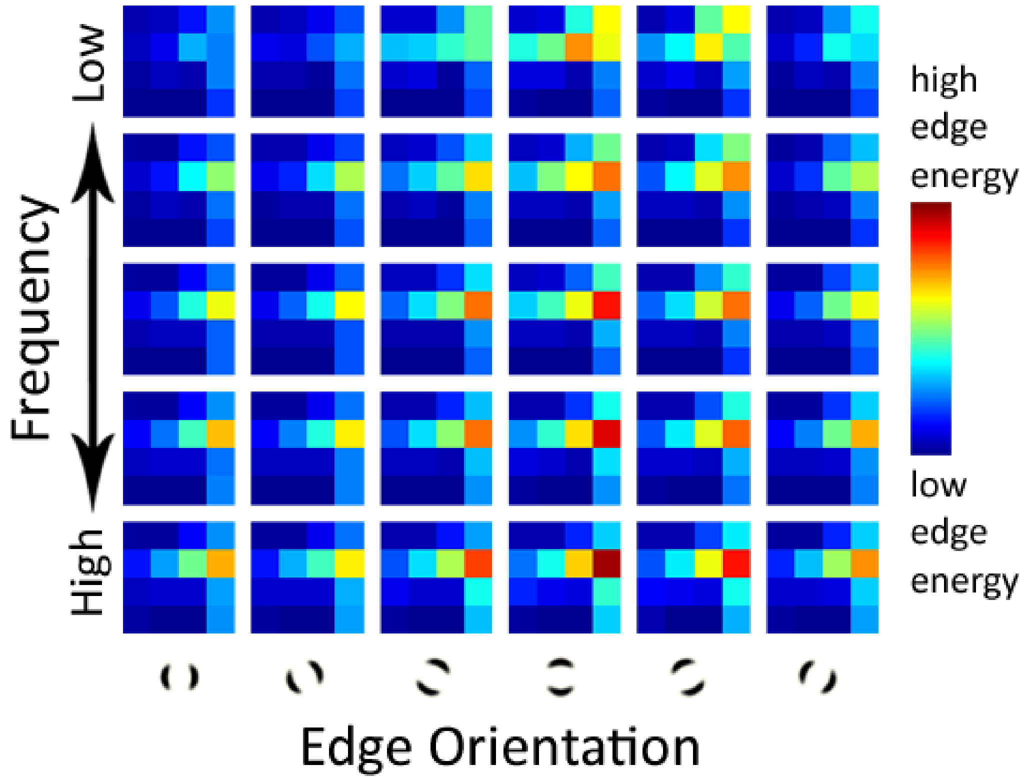




Scene Gist Descriptor
(Oliva and Torralba 2001)

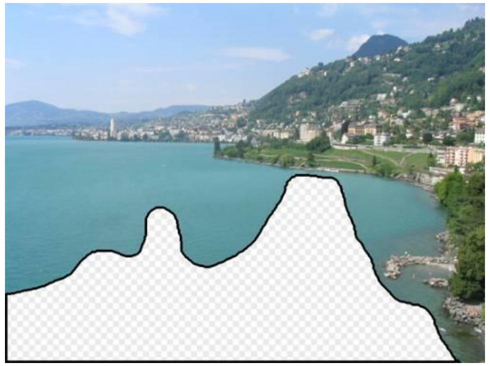
Scene Descriptor

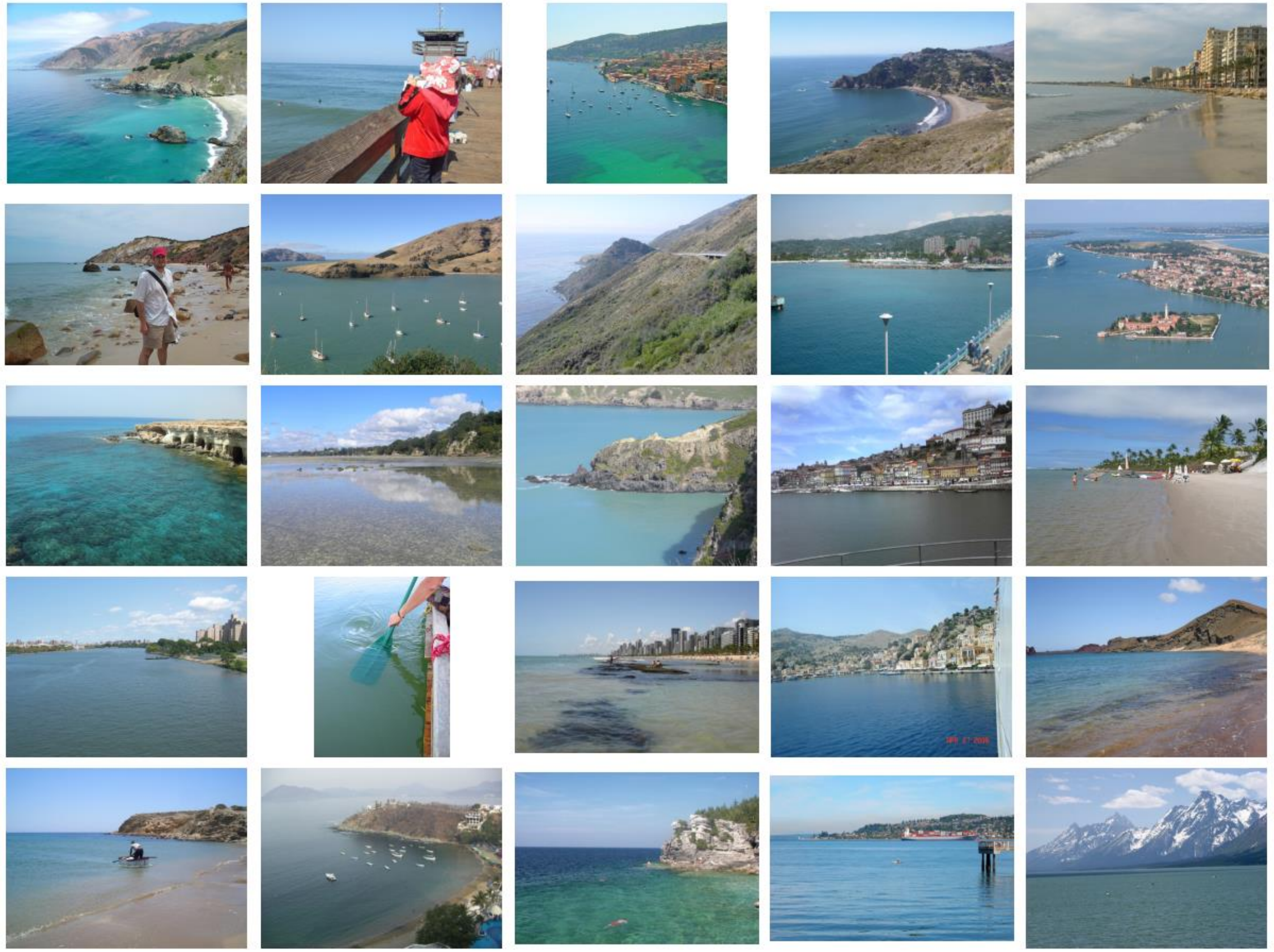




Scene Gist Descriptor
(Oliva and Torralba 2001)

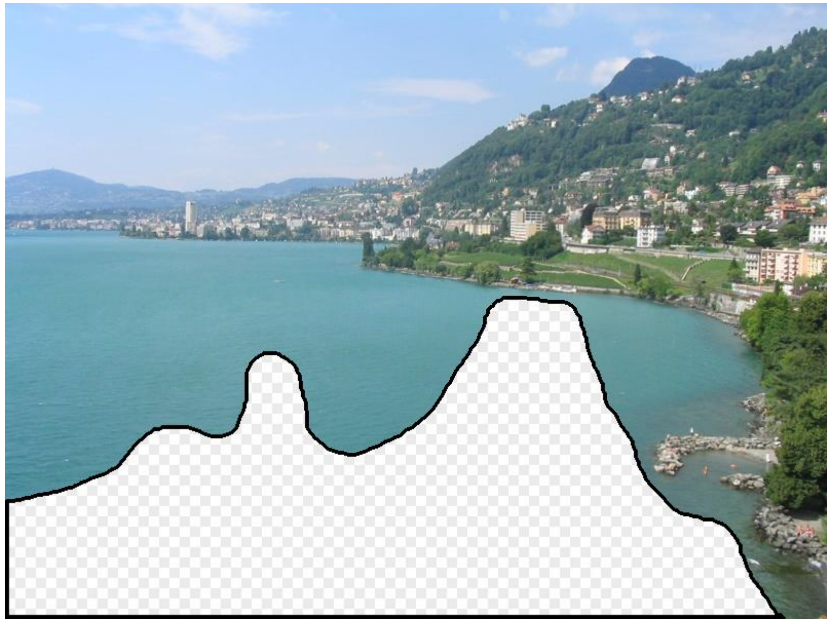
2 Million Flickr Images





… 200 total

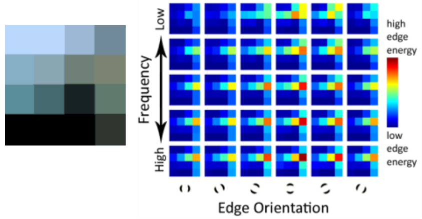
Context Matching



Graph cut + Poisson blending

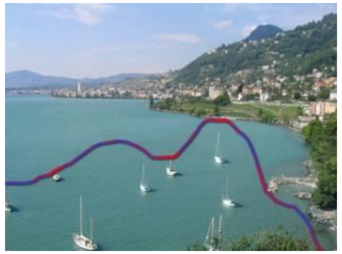
Result Ranking

We assign each of the 200 results a score which is the sum of:

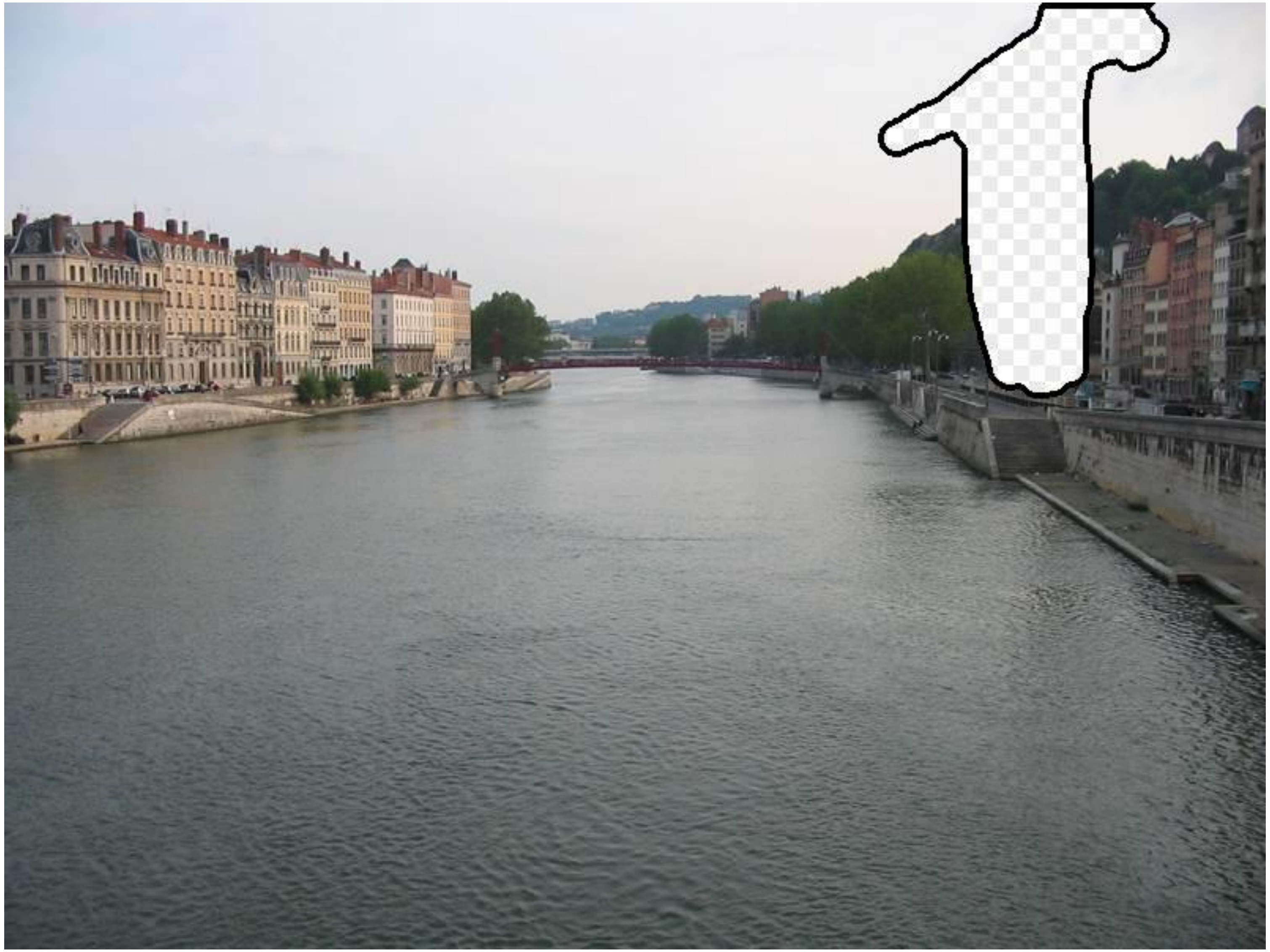


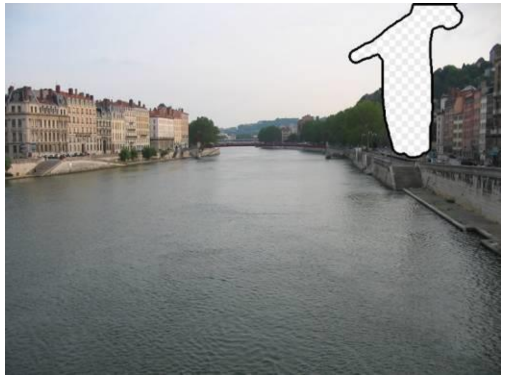
The scene matching distance

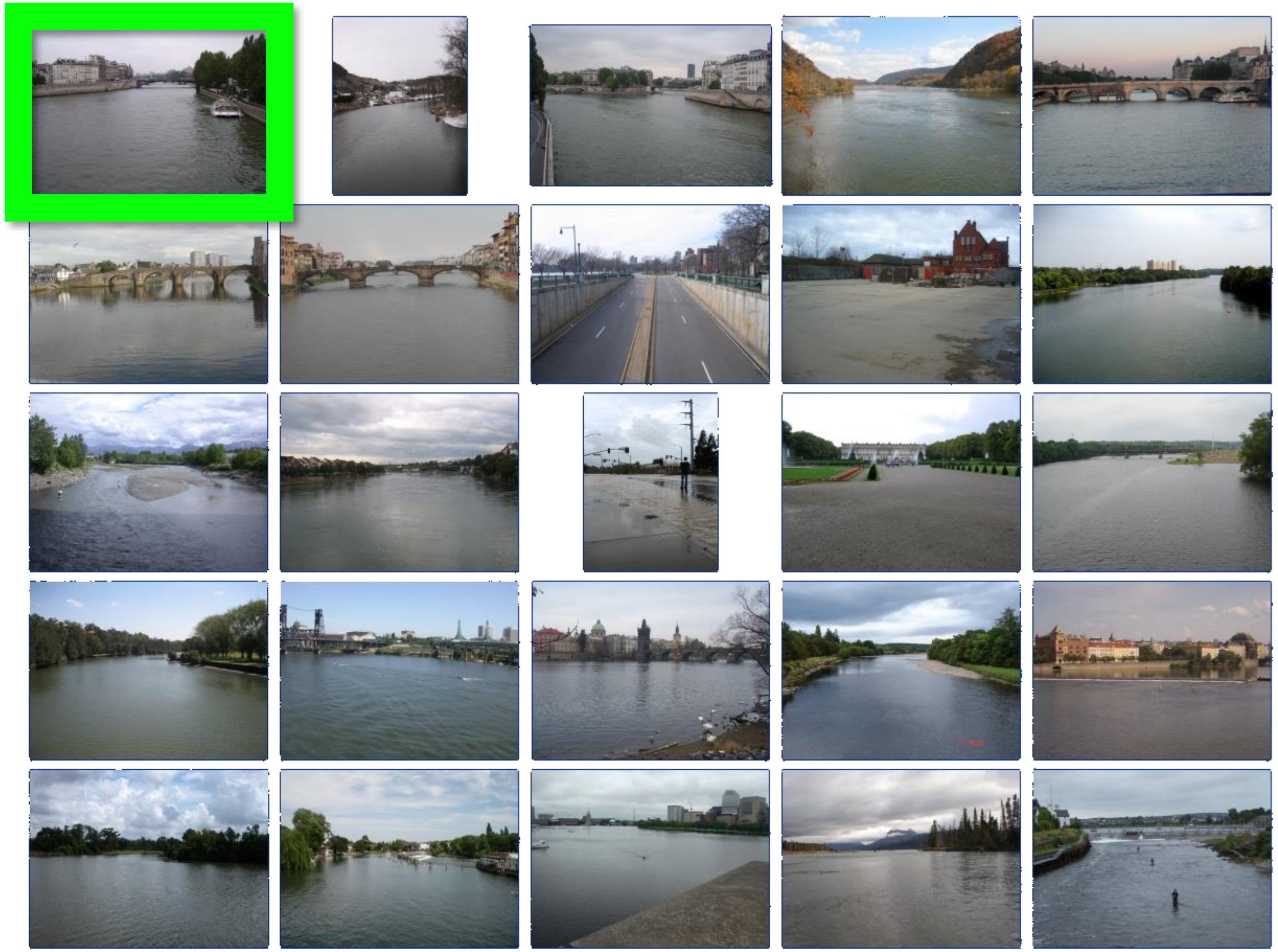
The context matching distance (color + texture)



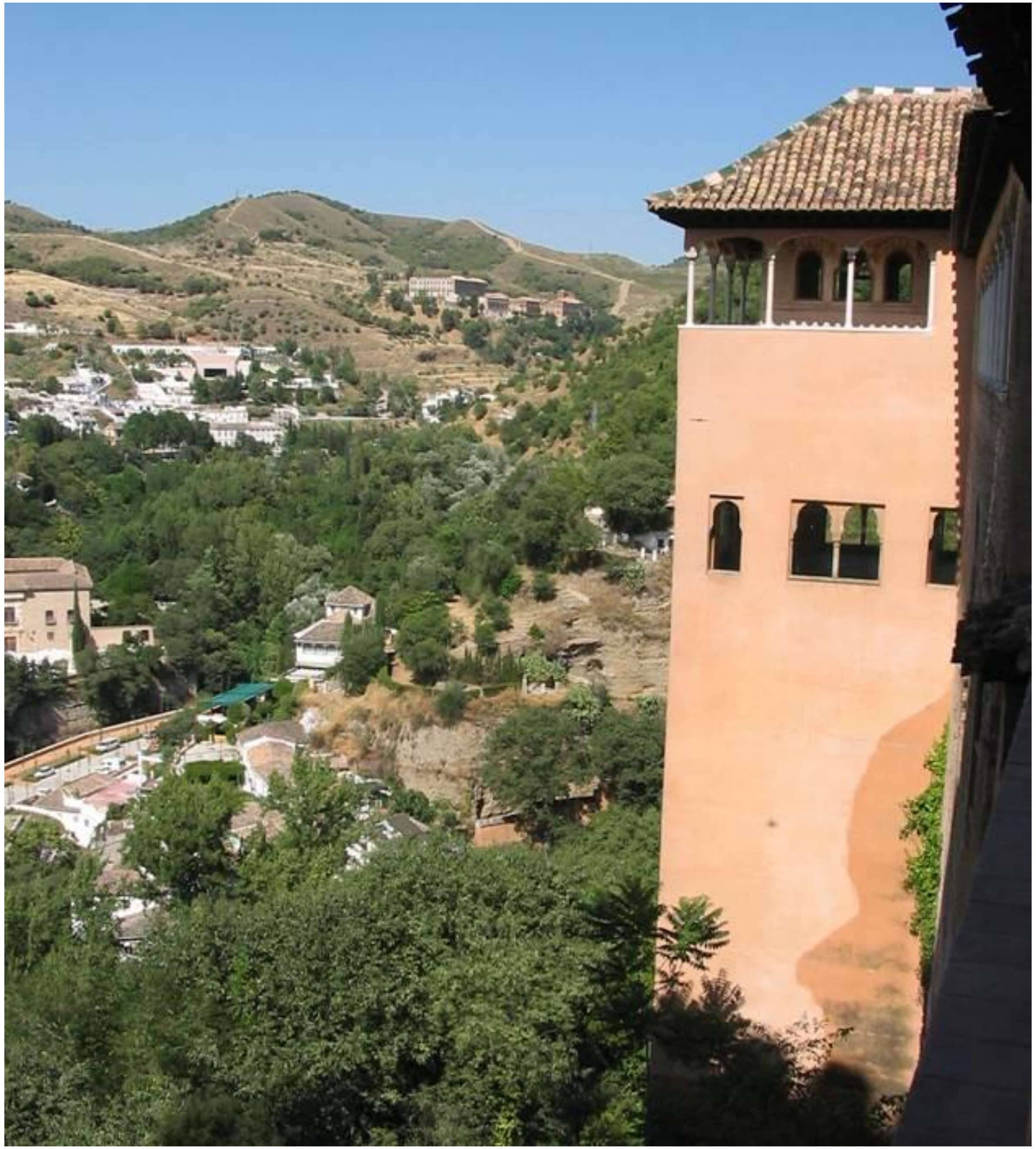
The graph cut cost

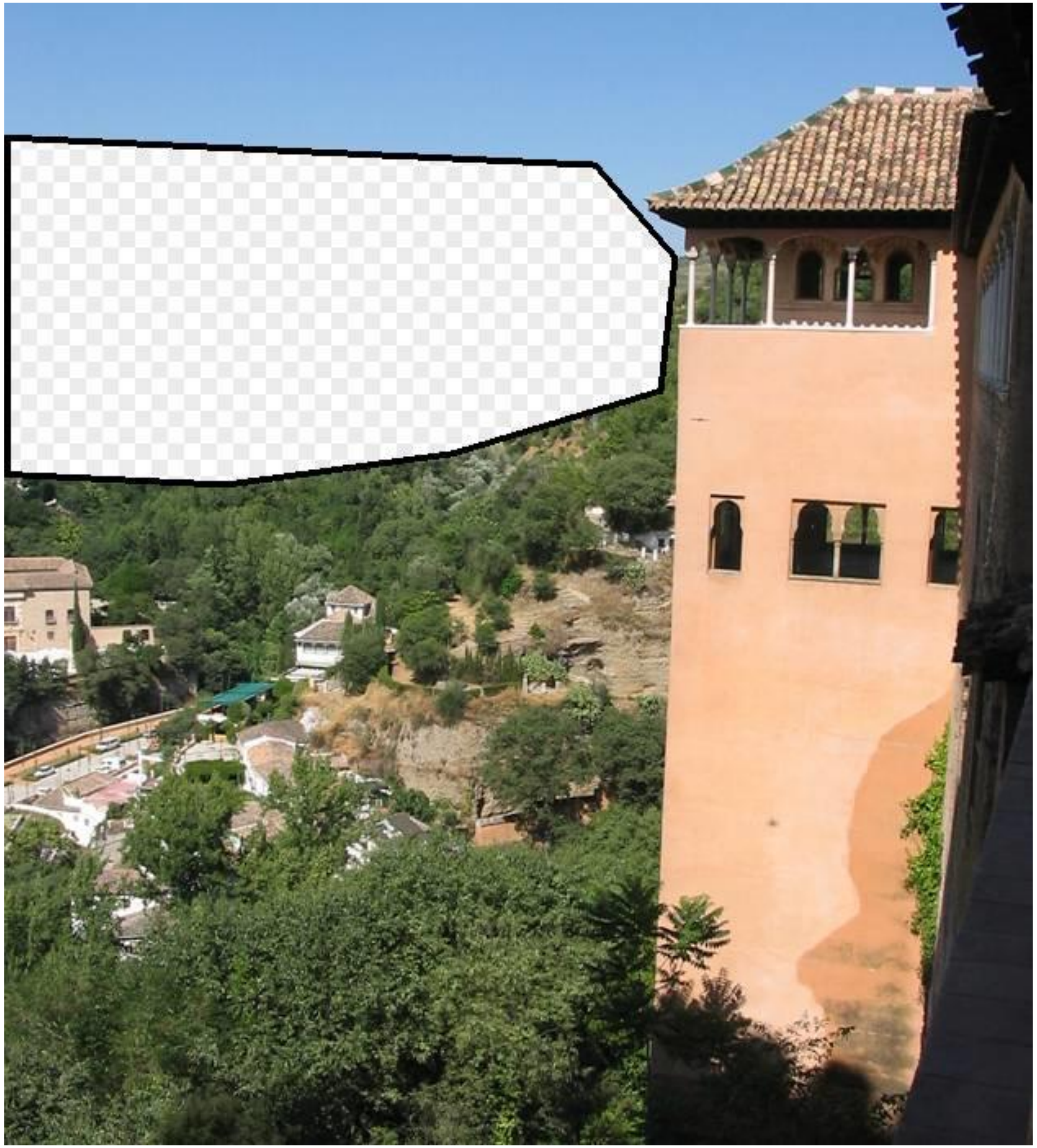






… 200 scene matches



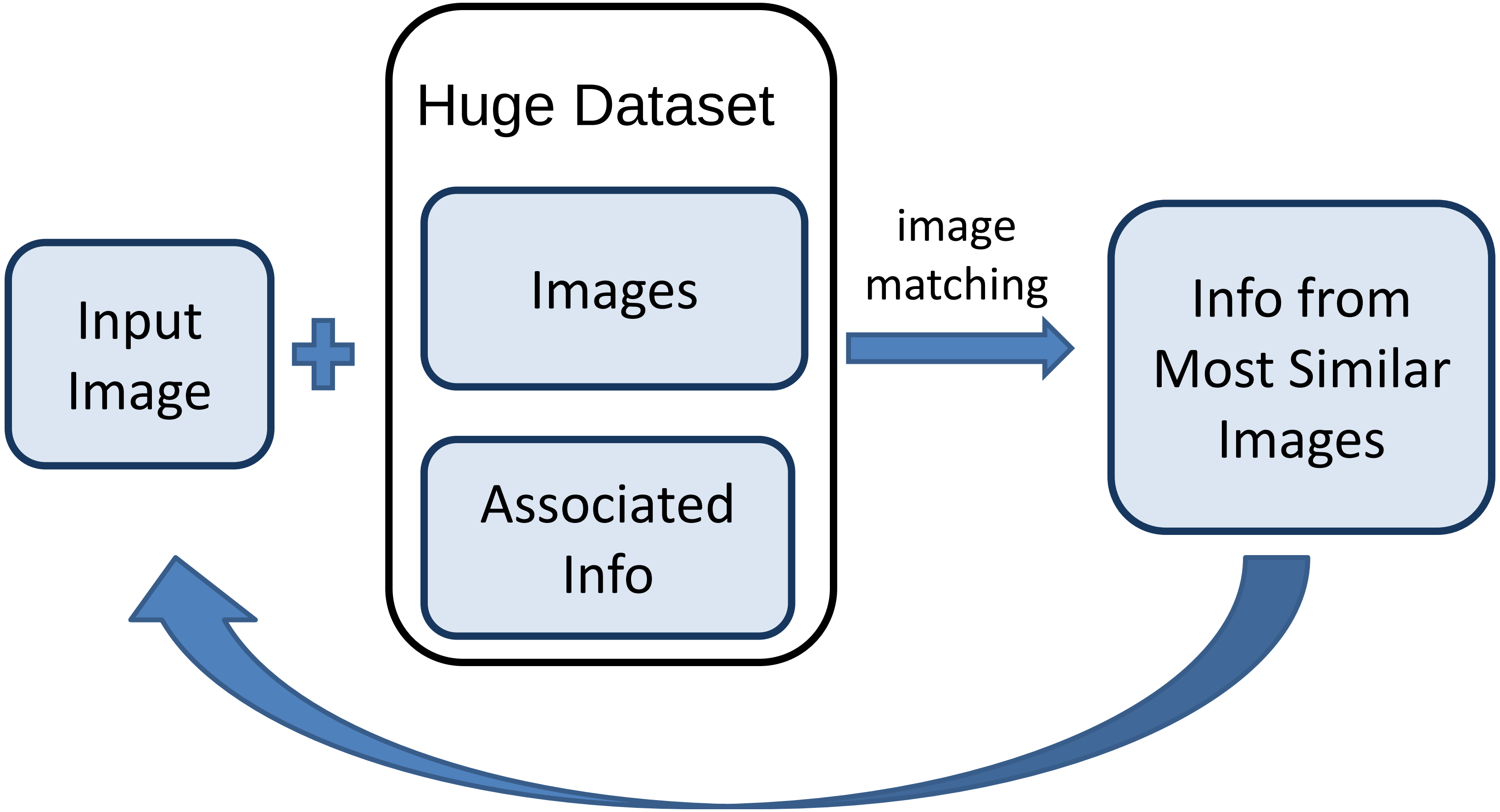


Outline

Opportunities of Scale: Data-driven methods

- The Unreasonable Effectiveness of Data
- Scene Completion
- Image Geolocation

General Principal



Hopefully, If you have enough images, the dataset will contain very similar images that you can find with simple matching methods.

Kosta Derpanis
@CSProfKGD

This reminded me of @jhhays and Efros' large-scale image geolocalization work

This Geography Genius Can Figure Out Exactly Where a Photo Was Shot
Tom Davies (AKA GeoWizard) is a human photo geotagger. He can figure out exactly where an outdoor photo was shot by studying it carefully.
petapixel.com

11:08 PM · Mar 4, 2021 from Toronto, Ontario · Twitter for iPhone

3 Likes

https://www.geoguessr.com/

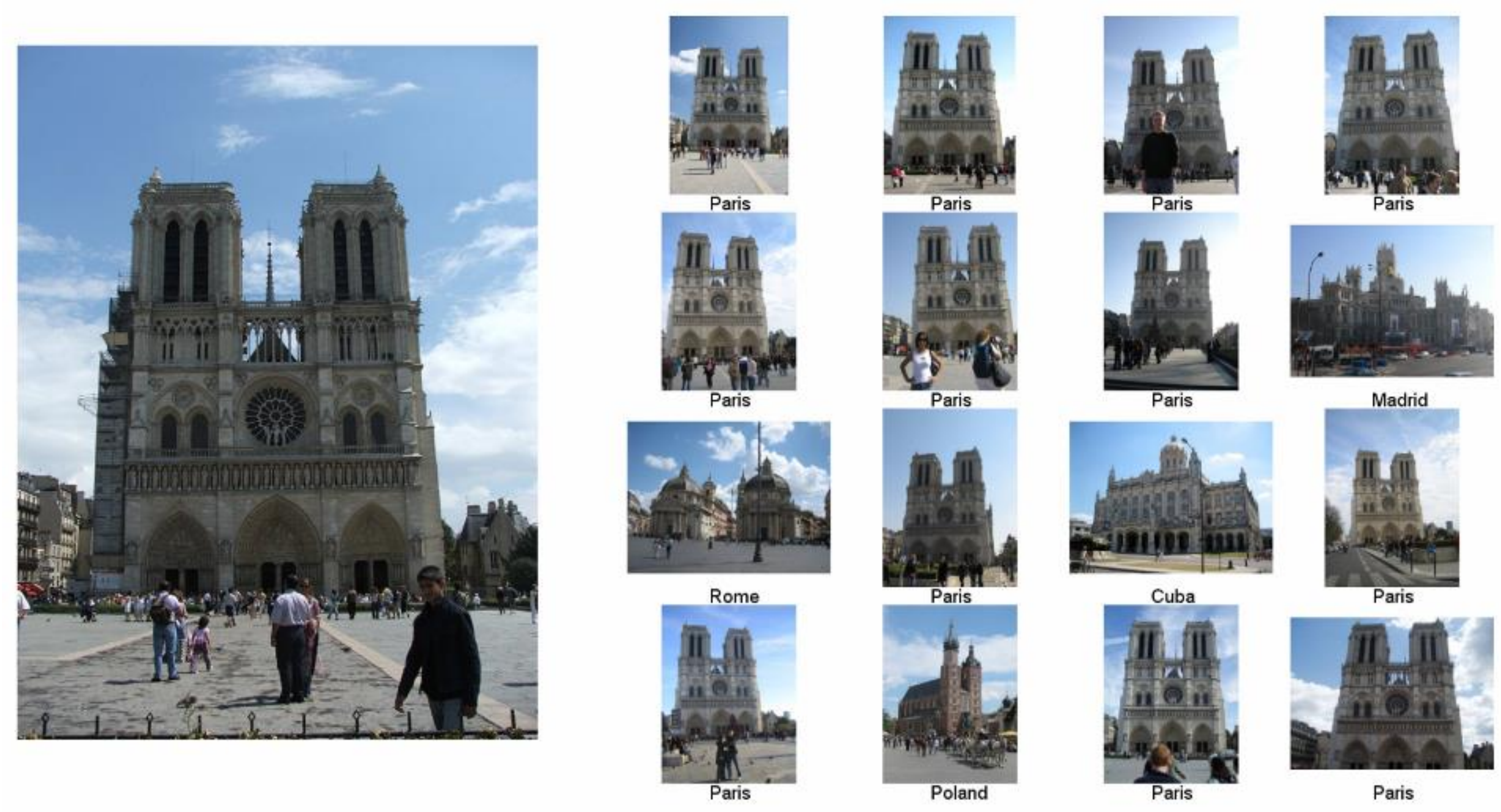
https://www.youtube.com/c/GeoWizard/videos

im2gps (Hays & Efros, CVPR 2008)

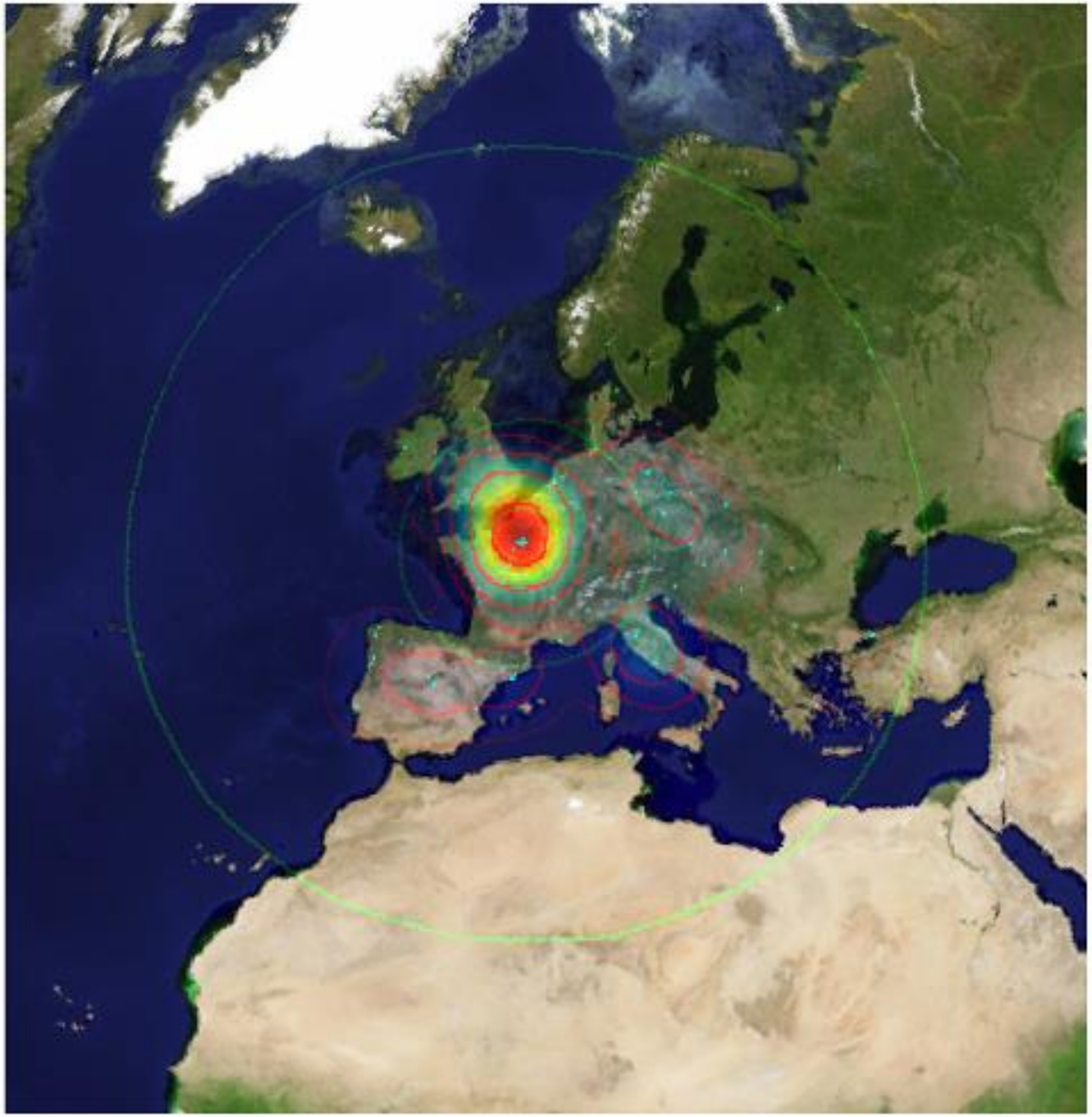
6 million geo-tagged Flickr images

http://graphics.cs.cmu.edu/projects/im2gps/

Where is this photo?

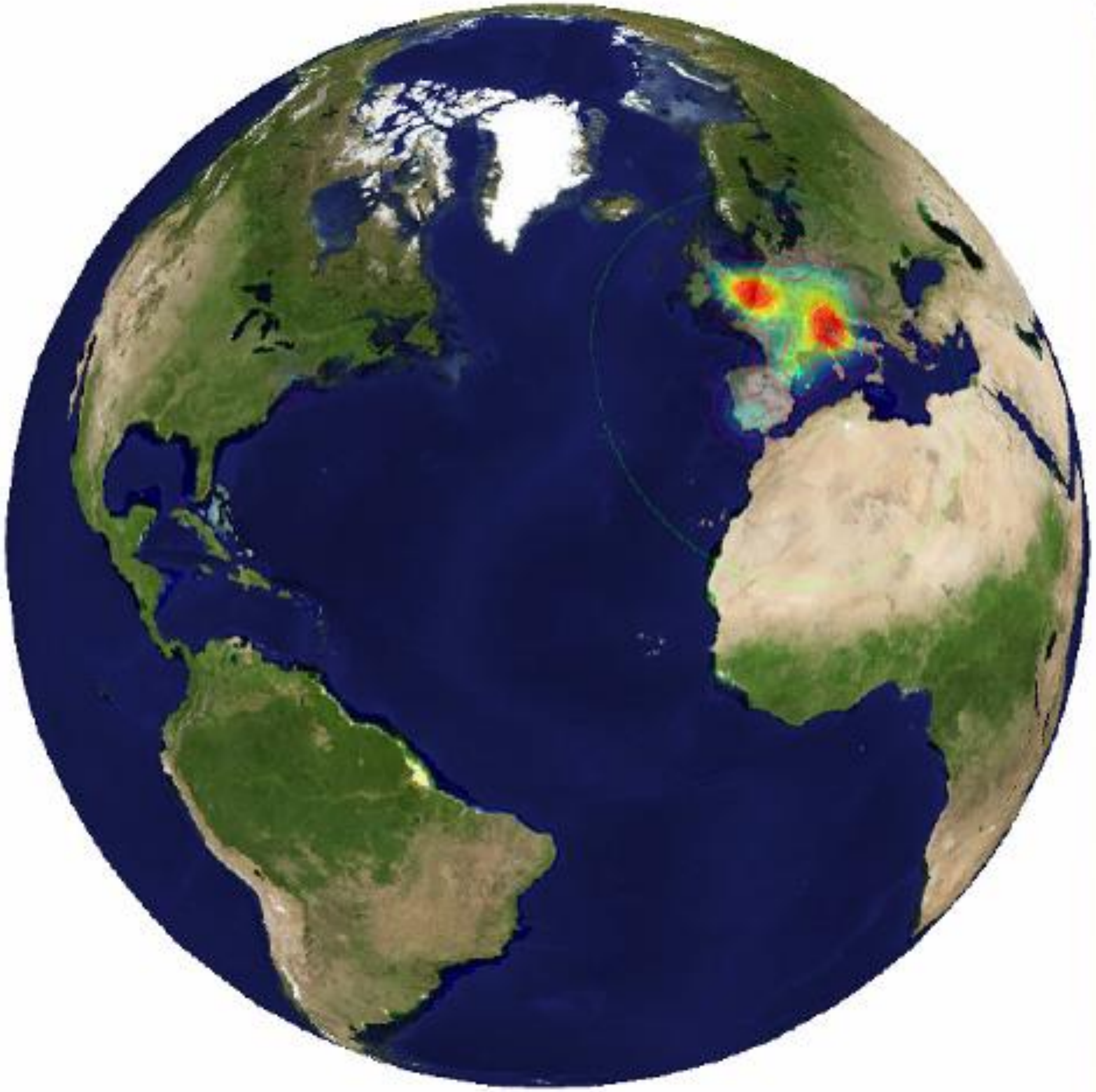


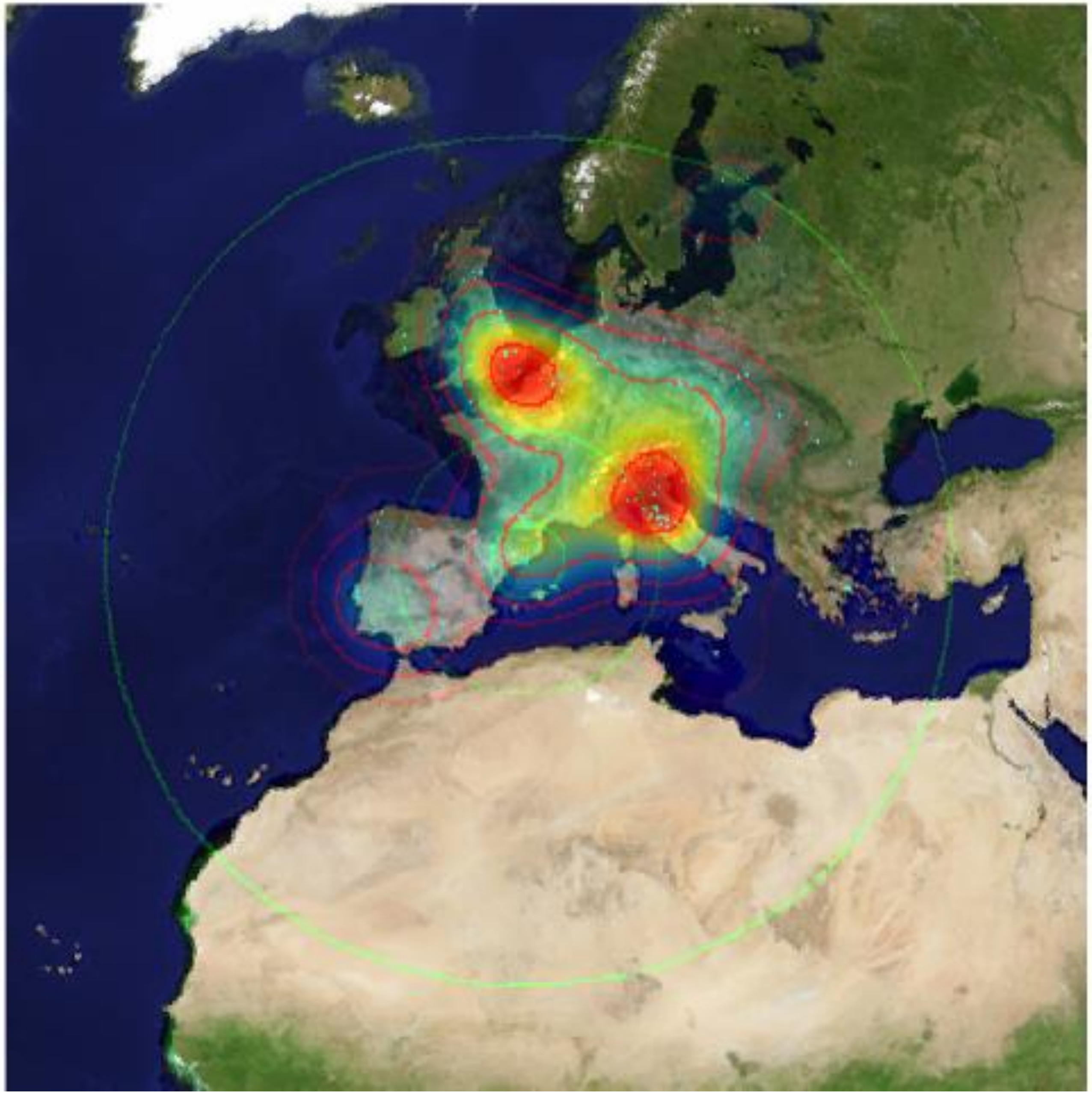
Nearest Neighbors according to gist + bag of SIFT + color histogram + a few others



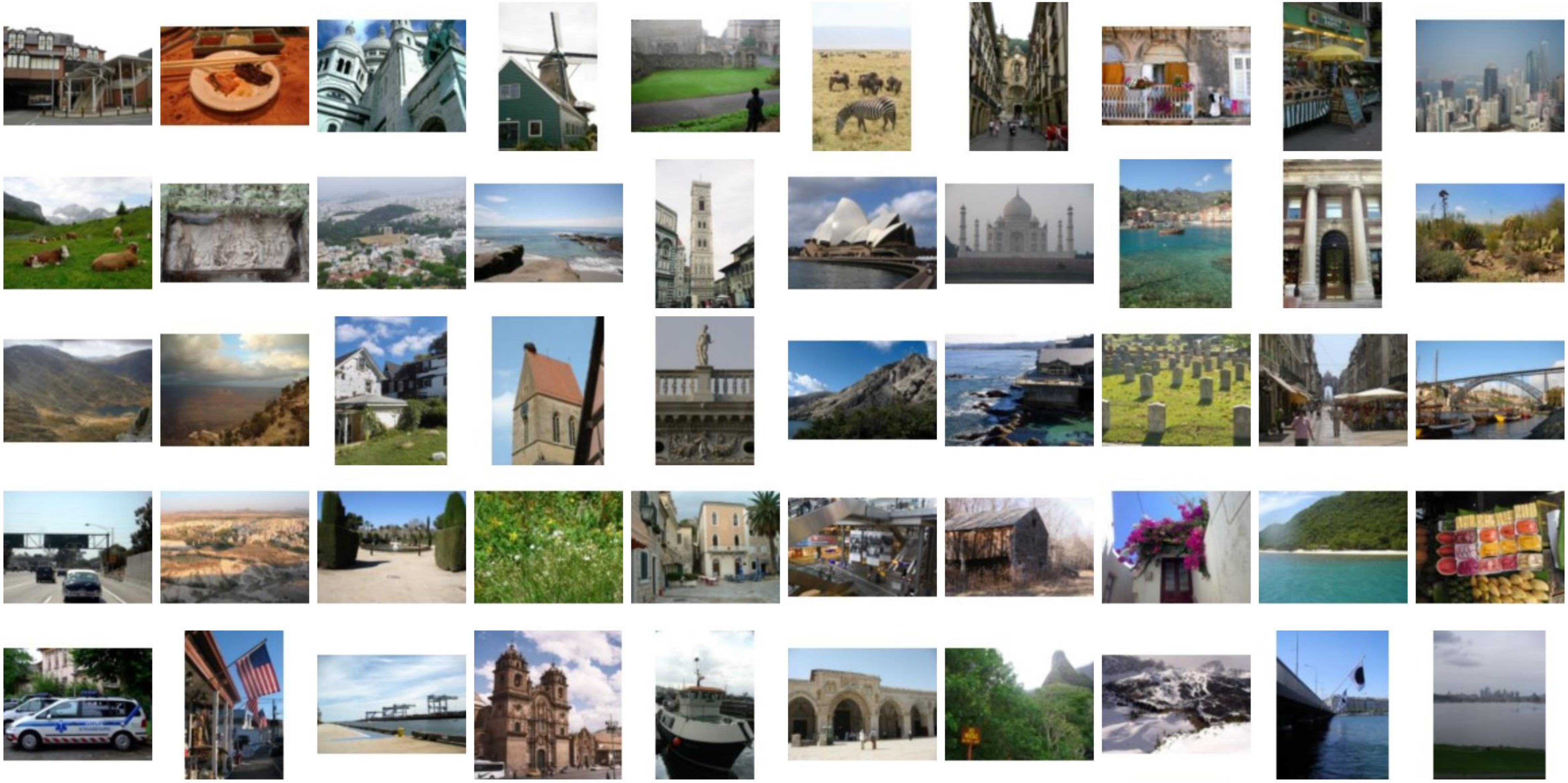
Where is this photo?

Nearest Neighbor Scenes

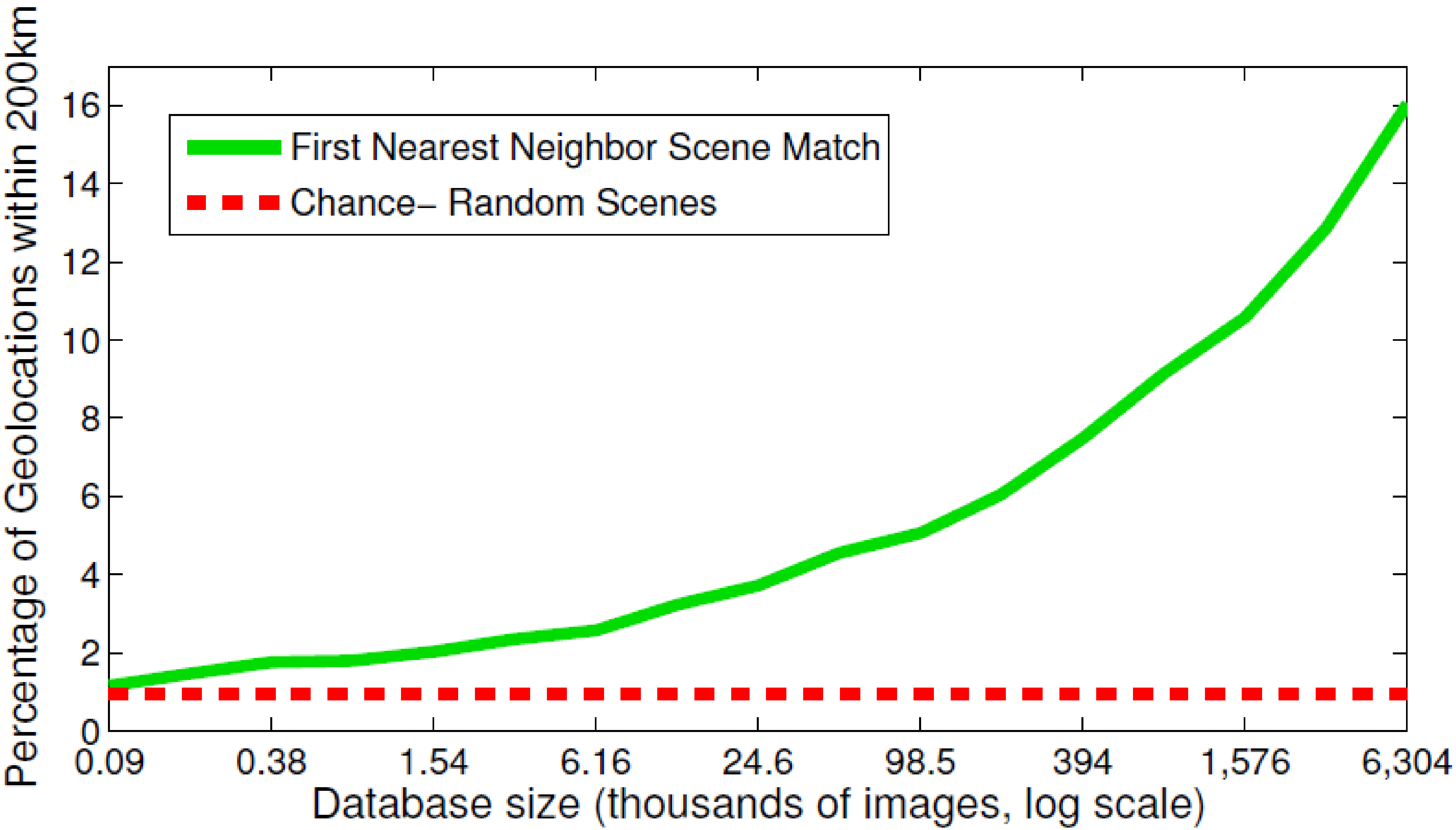




Test Set of 237 Touristy Photos



Effect of Dataset Size



PlaNet - Photo Geolocation with Convolutional Neural Networks

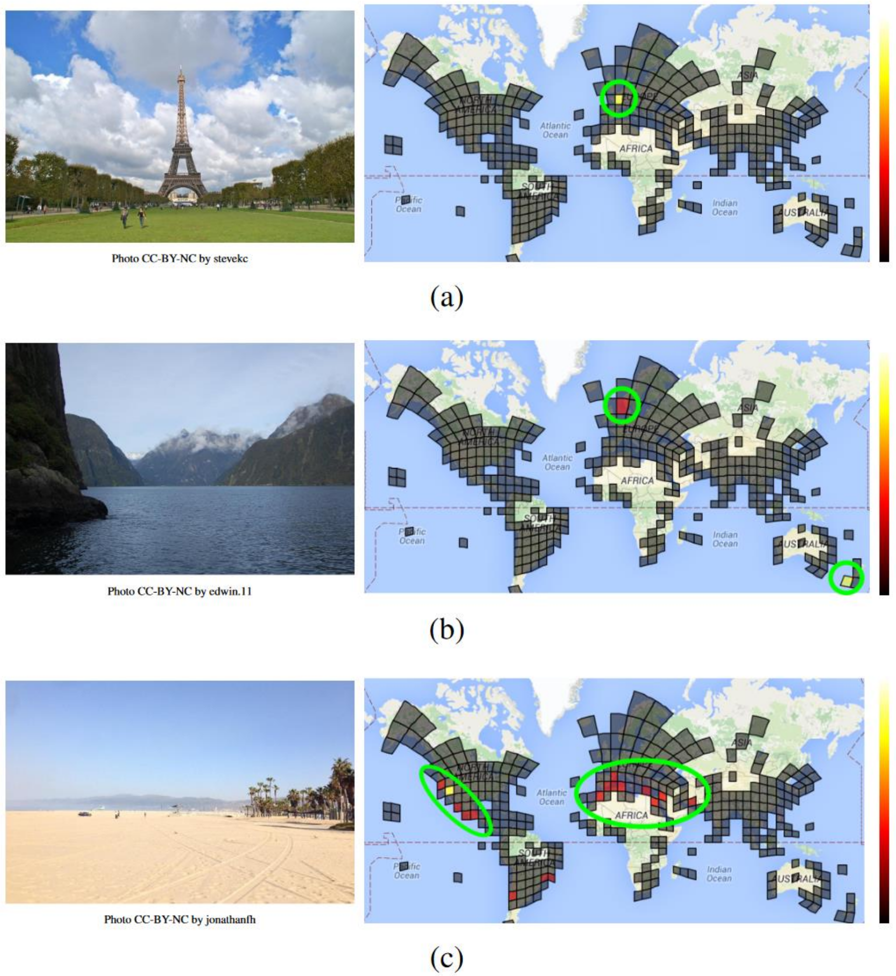
Tobias Weyand
Google
weyand@google.com

Ilya Kostrikov
RWTH Aachen University
ilya.kostrikov@rwth-aachen.de

James Philbin
Google
philbinj@gmail.com

Abstract

Is it possible to build a system to determine the location where a photo was taken using just its pixels? In general, the problem seems exceptionally difficult: it is trivial to construct situations where no location can be inferred. Yet images often contain informative cues such as landmarks, weather patterns, vegetation, road markings, and architectural details, which in combination may allow one to determine an approximate location and occasionally an exact location. Websites such as GeoGuessr *and* View from your Window *suggest that humans are relatively good at integrating these cues to geolocate images, especially en-masse. In computer vision, the photo geolocation problem is usually approached using image retrieval methods. In contrast, we pose the problem as one of classification by subdividing the surface of the earth into thousands of multi-scale geographic cells, and train a deep network using millions of geotagged images. While previous approaches only recognize landmarks or perform approximate matching using global image descriptors, our model is able to use and*



(a)

(b)

(c)

Revisiting IM2GPS in the Deep Learning Era.
Nam Vo, Nathan Jacobs, James Hays. ICCV 2017

		Street	City	Region	Country	Cont.
	Threshold (km)	1	25	200	750	2500
	Human*			3.8	13.9	39.3
2008	Im2GPS [9]		12.0	15.0	23.0	47.0
2009	Im2GPS [10]	02.5	21.9	32.1	35.4	51.9
2016	PlaNet [36]	08.4	24.5	37.6	53.6	**71.3**
2017	[L] 7011C	06.8	21.9	34.6	49.4	63.7
2017	[L] kNN, σ=4	**12.2**	**33.3**	**44.3**	**57.4**	**71.3**
2017	... 28m database	**14.4**	**33.3**	**47.7**	**61.6**	**73.4**

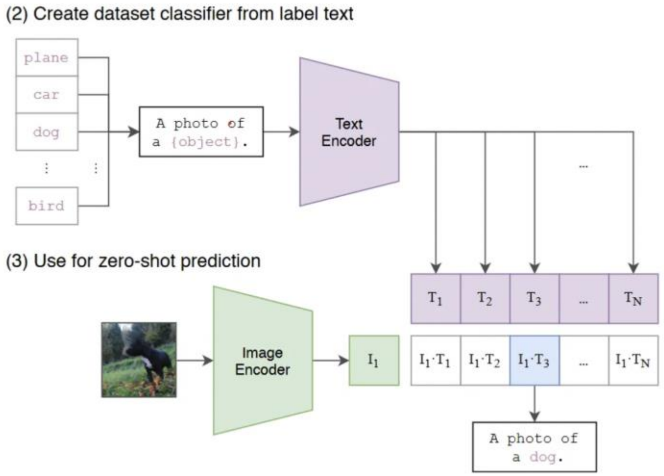
Geolocation Overview

- Bespoke Image Geolocation Approaches
  - Im2gps (2008)
  - PlaNet and im2gps revisited (2016 and 2017)
- Can Large Vision-Language Models geolocate images?
  - CLIP (2021)
  - GeoGuessr
  - Pigeon (2023)
- Can Large *Generative* Vision-Language Models geolocate images?

Are “Foundation” Models Good at Geolocation?

Learning Transferable Visual Models From Natural Language Supervision

Alec Radford [*1] Jong Wook Kim [*1] Chris Hallacy [1] Aditya Ramesh [1] Gabriel Goh [1] Sandhini Agarwal [1] Girish Sastry [1] Amanda Askell [1] Pamela Mishkin [1] Jack Clark [1] Gretchen Krueger [1] Ilya Sutskever [1]



	1km	25km	200km	750km	2500km
ISNs[a]	**16.9**	**43.0**	**51.9**	**66.7**	**80.2**
CPlaNet[b]	16.5	37.1	46.4	62.0	78.5
CLIP	13.9	32.9	43.0	62.0	79.3
Deep-Ret+[c]	14.4	33.3	47.7	61.6	73.4
PlaNet[d]	8.4	24.5	37.6	53.6	71.3

OpenAI’s CLIP (2021) is strong using only 1 million reference images


CUP

ROUND 6 OF 10 - 3.5X DAMAGE
CANADA
Long'Yugan
Лонгъюган
Yagel'nyi
Ягельный
Тарко-Сале
Gubkinsky
Vanzevat
Ванзеват
Sorum
Сорум
Priozernyi
Приозерный
Tol'ka
Толька
Verkhnekazymskii
Верхнеказымский
Sosnovka
Сосновка
Muravlenko
Муравленко
Khanymei
Ханымей
Polnovat
Beloyarsky
Белоярский
Numto
Нумто
Khalyasavei
Халясавэй
Tolka
Толька
Kikkiakki
Киккиакки
Ratta
Ратта
Noyabrsk
Ноябрьск
Vyngapurovskiy
Вынгапуровский
Kholmy
Холмы
Verkhne-Tazovsky State Nature Reserve
Верхне-Тазовский государственный природный заповедник
Nizhnesortymskiy
Нижнесортымский
Priobye
Приобье
Когалым
Nyagan
Нягань
Raduzhny
Радужный
Pokachi
Novoagansk
Новоаганск
1251
320
TheFungusAmongUs
Consus
231 km
DISTANCE FROM LOCATION
25 km
4282
ROUND RESULT
x3.5
4916


Ozero Pil'tanlor
оз. Пильтанлор
Lyantor
Лянтор
Fedorovskii
Федоровский
Chernorechenskiy
Черноречѐнский
Peschanyi
Песчаный
Lyamina
Лямина
Bannyy
Банный
Ob River
Ob River
Ob River
Ob River
Tundrino
Тундрино
Saigatina
Сайгатина
Bely Jar
Городское
поселение
Белый Яр
Surgut
Сургут
Rechnik
Речник
Shirokovo
Широково
Singapai
Google


Lyantor
Лянтор

https://www.plonkit.net/russia

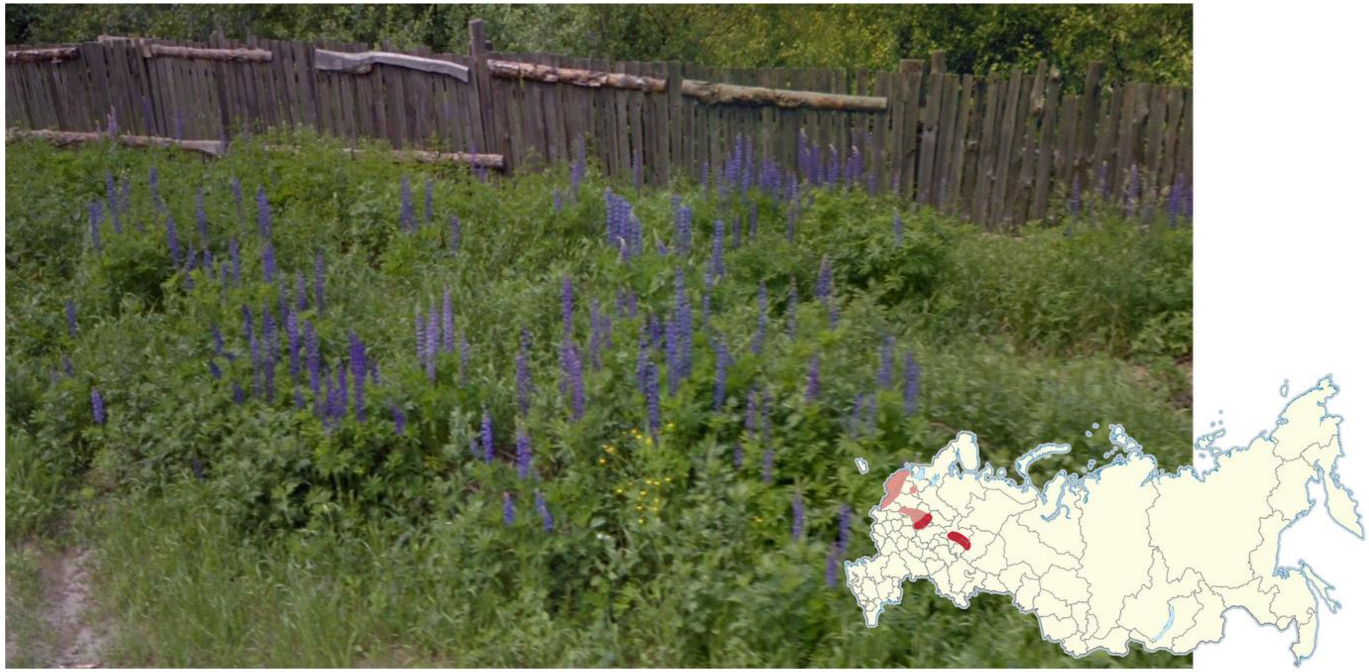
Sandy roadsides are common in Khanty-Mansi and Yamalo-Nenets and adjacent subjects, as well as areas around Nizhny Novgorod on the Volga river. Other notable areas are Karelia, Murmansk, and western Sakha. Beware, however, that sandy roadsides can less commonly be found near rivers in other regions.

Red soil is common in the highlighted areas, notably around Izhevsk and Perm, much of Arkhangelsk Oblast, Leningrad Oblast, and Pskov Oblast, and near Volgograd and Astrakhan. Note that this map is by no means exhaustive; red soil can be found almost anywhere in the country near water or iron mines.

Birches very close together, as well as forests consisting of **only birches**, are indicative of areas east of the Urals, most commonly between Chelyabinsk Oblast and Novosibirsk Oblast.

Siberian larches are one of the dominant tree species in much of eastern Russia, recognized by their unique **needle-like leaves**. Generally speaking, they become more prevalent the further east you go in the country, as well as at high elevations.

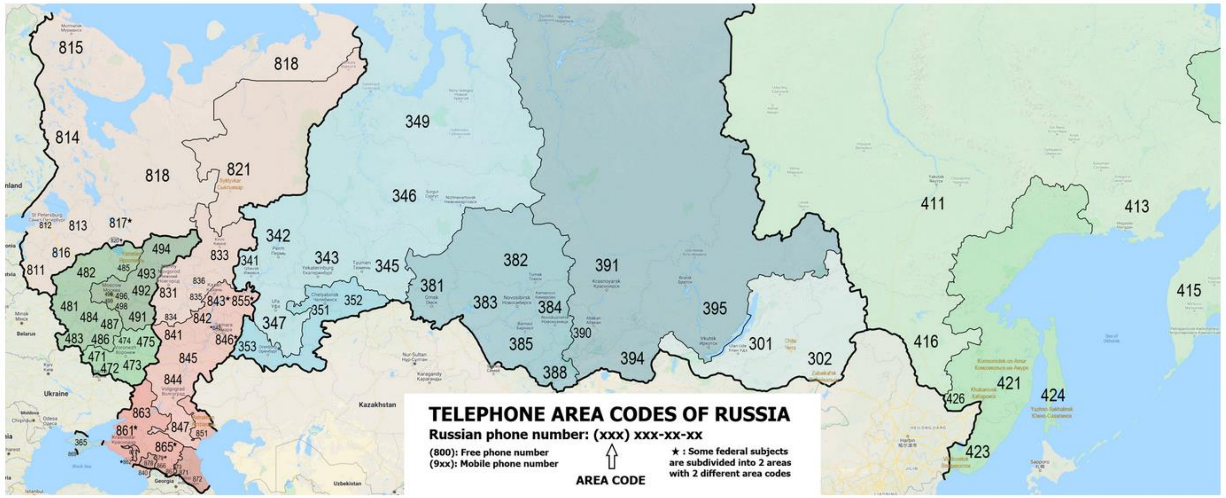
Sunflowers are common along the border with Ukraine as well as more eastern oblasts like Ulyanovsk, Samara, and northern Orenburg.



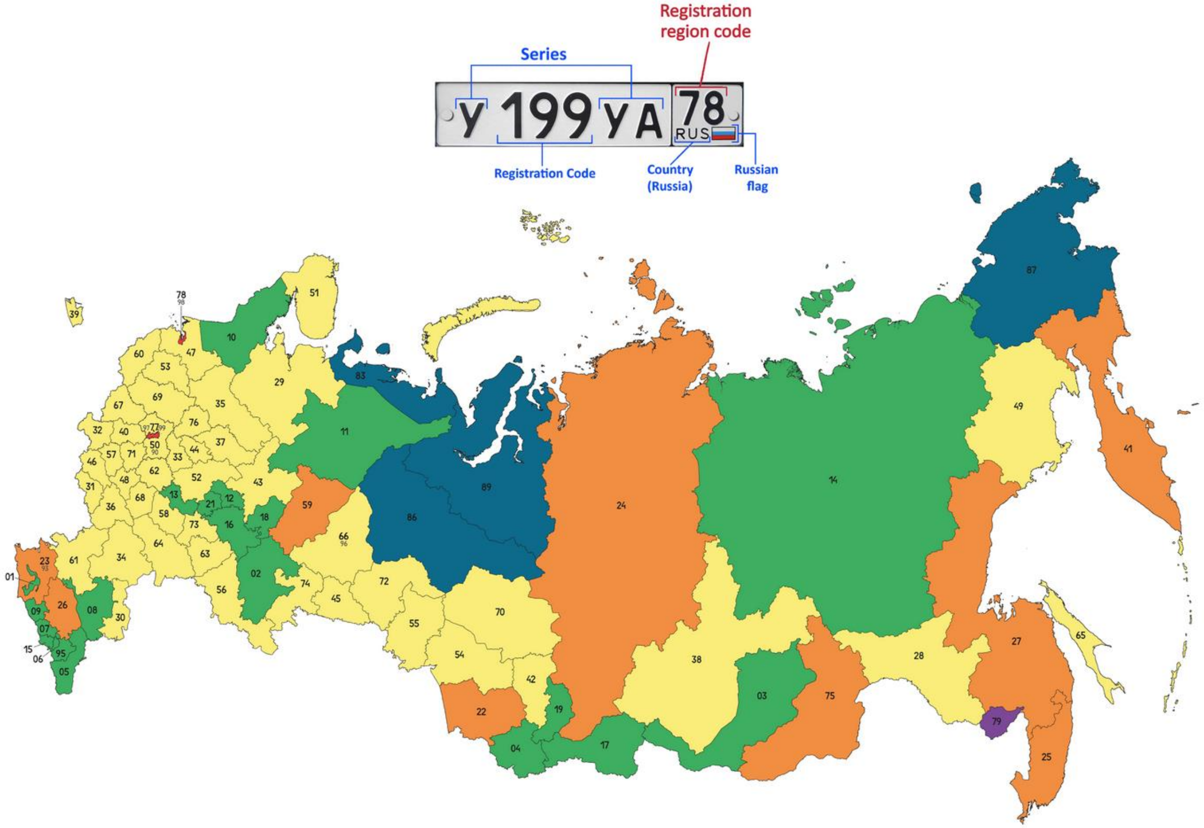
Blue-pod lupines appear very commonly in northern Vladimir Oblast, eastern Ivanovo Oblast, and southwestern Kostroma Oblast. It can also be found less commonly elsewhere in Kirov Oblast and towards the Baltics.

Grassy fields, with **bushy vegetation**, in early spring Generation 4 coverage is typical for Dagestan. The landscape can either be completely flat or mountainous. These flowers are also quite common in the Generation 4 Dagestan coverage.

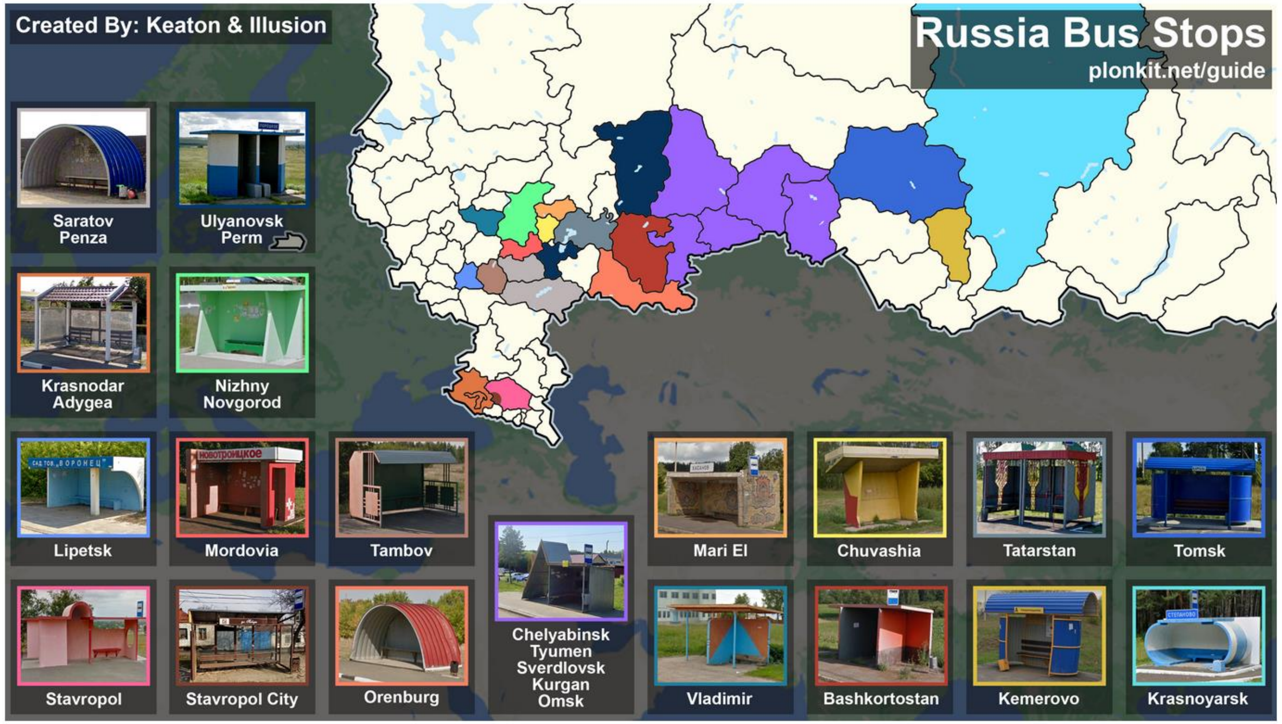
The **Caucasus mountain range** is one of the largest mountain ranges in Russia. The tallest mountain in Russia, Mount Elbrus, can be found on the border of the Kabardino-Balkarian Republic and Karachay-Cherkessia.



This is a map of Russian **area codes**. Notably, area codes starting with 8 are in the west, codes starting with 3 are fairly central and codes starting with 4 are either east or around Moscow.



In Generation 3 coverage you will somewhat commonly find unblurred licence plates, featuring a **regional code** on the right side. The codes are generally ordered alphabetically within each type of federal subject, starting at republics and ending with autonomous okrugs. Therefore, the Republic of Adygea will be represented by 01, and the Amur Oblast by 28, both being the first alphabetical subjects of republics and oblasts respectively. If you encounter a three digit code, the second and third digit will form the regional code, in this case 123 becomes 23, for Krasnodar Krai. You may also find the codes written out on the back of trucks and vans.



These are the bus stops unique to specific federal subjects in Russia. Notably common and memorable ones include Krasnoyarsk Krai, Chuvashia, Tatarstan, and Mari El Republic.

Japanese-made cars with the **steering wheel on the right** become more common the further east you go, generally starting around Novosibirsk.

Buildings built almost entirely of **red brick** are mostly found south, but other notable exceptions include Magnitogorsk, Orsk, and Omsk.

While mosques can be found everywhere in Russia, they are by far most common in areas with a Muslim majority, mainly in much of south Russia as well as Tatarstan and Bashkortostan.

The A-type short antenna has its highest ridge on the right. Notable areas for this antenna are around Kaluga, Orenburg, and Krasnodar.

The B-type short antenna has its highest ridge on the left. This antenna is wide-ranging, but it is most notably found near Nizhny Novgorod, Elista, and Yekaterinburg and Tyumen.

PIGEON: PREDICTING IMAGE GEOLOCATIONS

PREPRINT

Lukas Haas
Department of Computer Science
Stanford University
lukashaas@cs.stanford.edu

Michal Skreta
Department of Computer Science
Stanford University
michal.skreta@stanford.edu

Silas Alberti
Department of Electrical Engineering
Stanford University
salberti@stanford.edu

Chelsea Finn
Department of Computer Science
Stanford University
cbfinn@cs.stanford.edu

Benchmark	**Method**	**Median Error** km	**Distance (% @ km)** *Street* **1 km**	*City* **25 km**	*Region* **200 km**	*Country* **750 km**	*Continent* **2,500 km**
IM2GPS (Hays & Efros, 2008)	PlaNet (Weyand et al., 2016)	> 200	8.4	24.5	37.6	53.6	71.3
	CPlaNet (Seo et al., 2018)	> 200	16.5	37.1	46.4	62.0	78.5
	ISNs(M,f^*,S_3) (Müller-Budack et al., 2018)	> 25	16.9	43.0	51.9	66.7	80.2
	Translocator (Pramanick et al., 2022)	> 25	19.9	48.1	64.6	75.6	86.7
	GeoDecoder (Clark et al., 2023)	$\sim$ **25**	**22.1**	**50.2**	**69.0**	80.0	89.1
	PIGEOTTO (Ours)	70.5	14.8	40.9	63.3	**82.3**	**91.1**

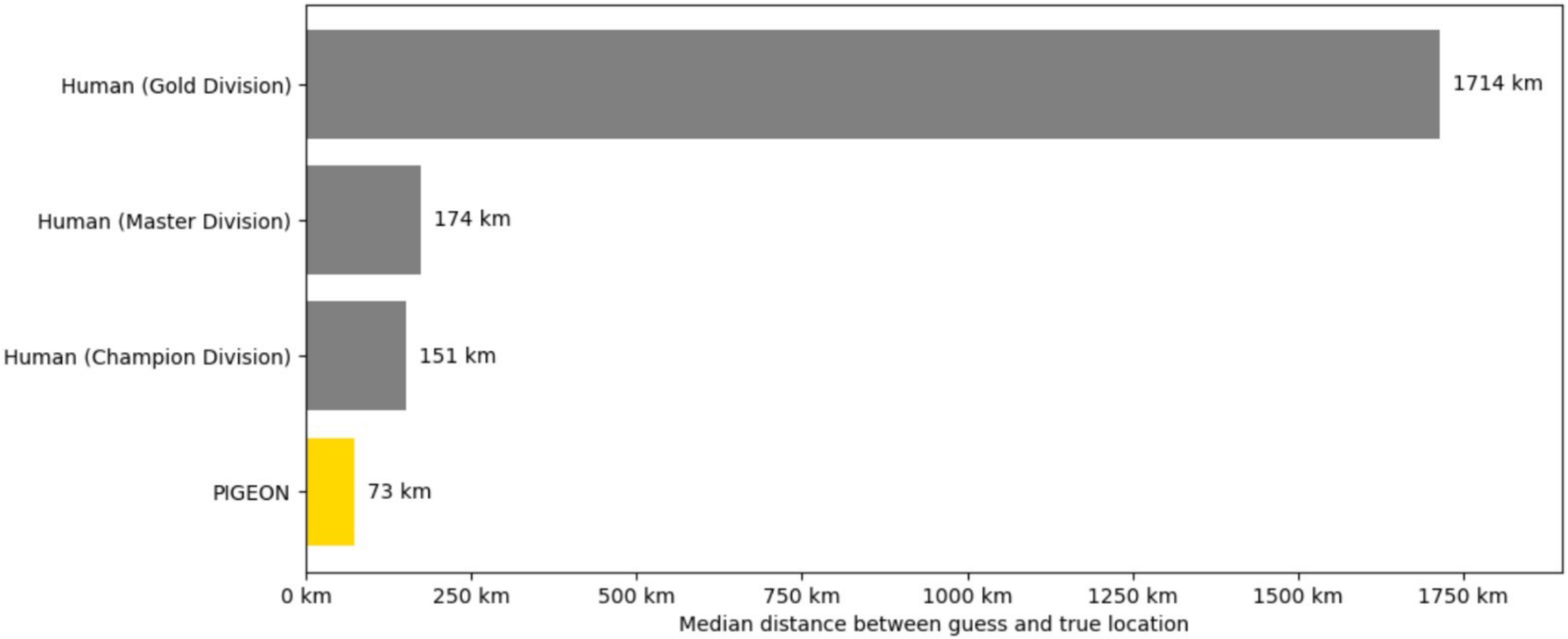


Figure 4: Geolocalization error of PIGEON against human players of various in-game skill levels across 458 multi-round matches. The Champion Division consists of the top 0.01% of players. PIGEON's error is higher than in Table 1 because Geoguessr round difficulties are adjusted dynamically, increasing with every round.

https://www.youtube.com/watch?v=ts5lPDV--cU geospy.ai/

https://geospy.ai

GeoSpy AI v0.1.6

Country: United States
State: Washington
City: Seattle
Explanation: The photo was taken from an airplane flying over Seattle. The clouds in the background and the green landscape below suggest that the photo was taken in the Pacific Northwest. The shape of the Puget Sound and the location of the mountains in the background confirm that the photo was taken in Seattle.
Coordinates: 47.6062° N, 122.3321° W

Outline

- Bespoke Image Geolocation Approaches
  - Im2gps (2008)
  - PlaNet and im2gps revisited (2016 and 2017)
- Can Large Vision-Language Models geolocate images?
  - CLIP (2021)
  - GeoGuessr
  - Pigeon (2023)
  - Geospy
- Can Large *Generative* Vision-Language Models geolocate images?

GPT-4V System Card

Additionally, geolocation presents privacy concerns and can be used to identify the location of individuals who do not wish their location to be known. Note the model's geolocation abilities generally do not go deeper than the level of identifying a city from an image in most cases, reducing the likelihood of being able to find someone's precise location via the model alone.

https://openai.com/index/gpt-4v-system-card/

“Least to Most” prompting of GPT-4V

Please provide your speculative guess for the location of the image at the country, city, neighborhood, and exact location levels. You must provide reasoning for why you have selected the value for each geographical level...

Work by Ethan Mendes, Yang Chen, James Hays, Sauvik Das, Wei Xu, and Alan Ritter

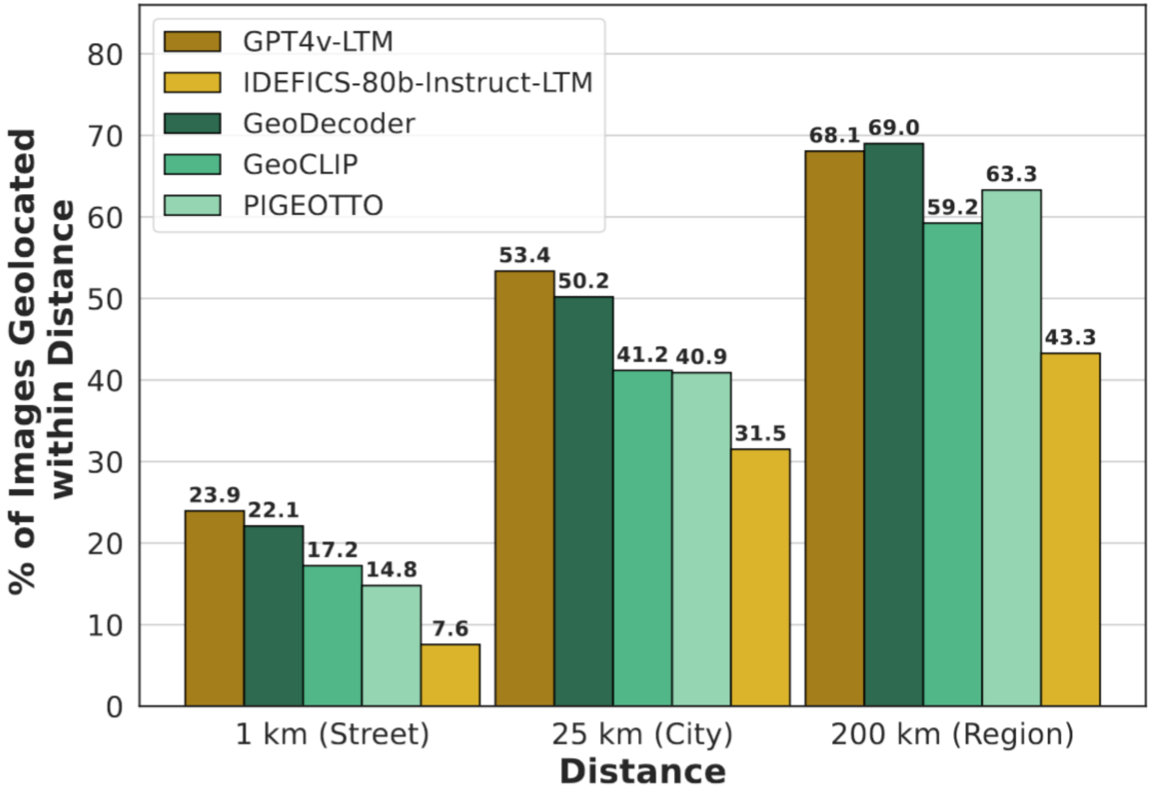
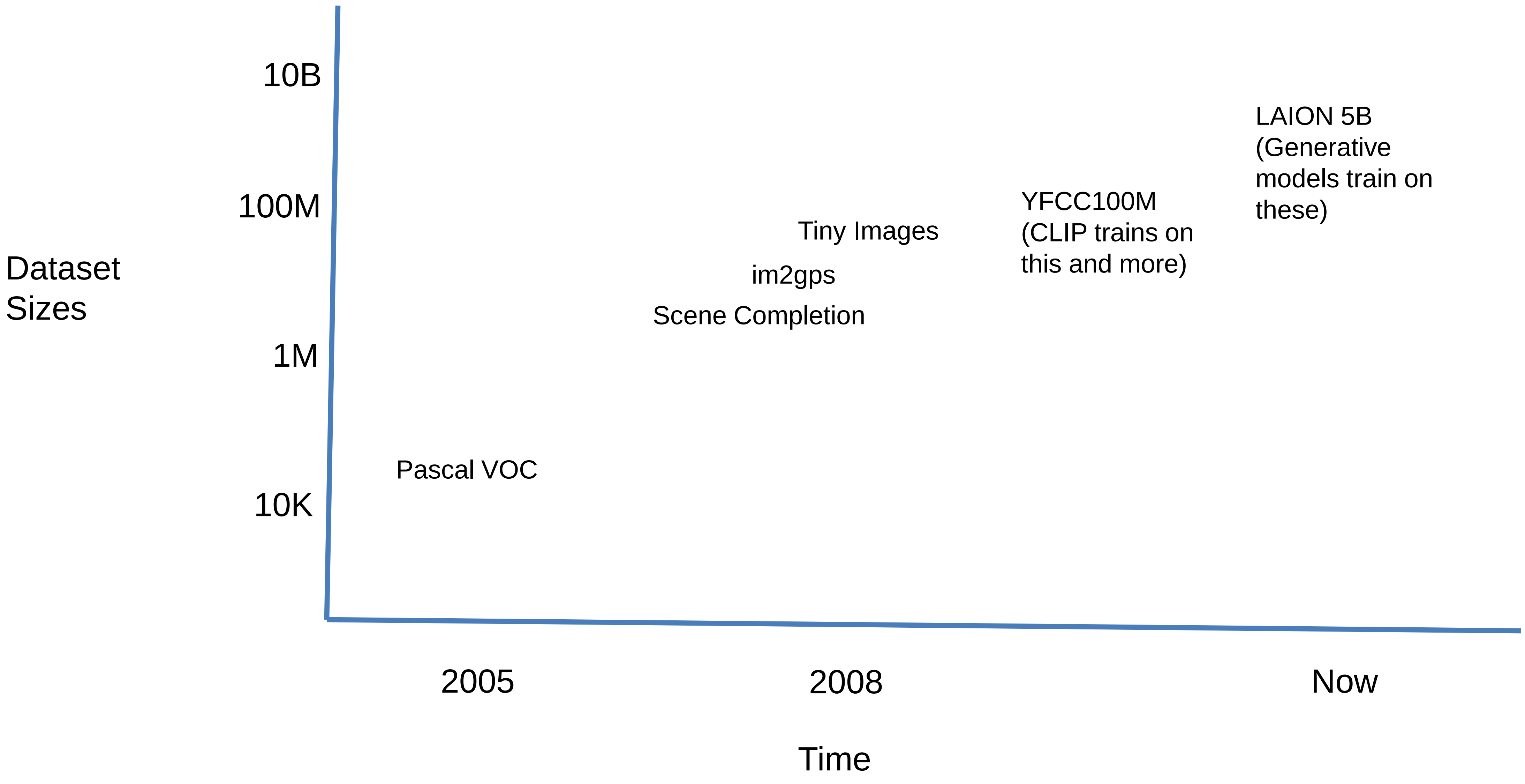


Figure 2: GPT-4v with geographical least-to-most (LTM) prompting performs well on the IM2GPS (Hays and Efros, 2008) benchmark compared to the state-of-the-art geolocation models GeoDecoder (Clark et al., 2023), GeoCLIP (Vivanco Cepeda et al., 2024), and PIGEOTTO (Haas et al., 2023). GPT-4v also has the lowest median distance error of 13 km.

Concern: GPT could have memorized the testing data

Dataset Sizes through Time



Revisiting our Big Idea

- Do we need computer vision systems to have strong AI-like reasoning about our world?
  - For some tasks, yes. For most tasks, probably not.
- What if invariance / generalization isn't actually the core difficulty of computer vision?
  - Generalization is still a fundamental, hard task.
- What if we can perform high level reasoning with brute-force, data-driven algorithms?
  - Combinatorics tells us we can't naively brute force our way very far.

Summary

- With billions of images on the web, it's often possible to find a close nearest neighbor
- In such cases, we can shortcut hard problems by "looking up" the answer, stealing the labels from our nearest neighbor. For example, simple (or learned) associations can be used to synthesize background regions, colorize, or recognize objects
- But we can't really "brute force" computer vision. Still, it's nice to get an intuition for the size of "image space".