



CSCI 3360 | Spring 2024 Data Science I

Jin Sun, PhD
School of Computing

Week 3: Data

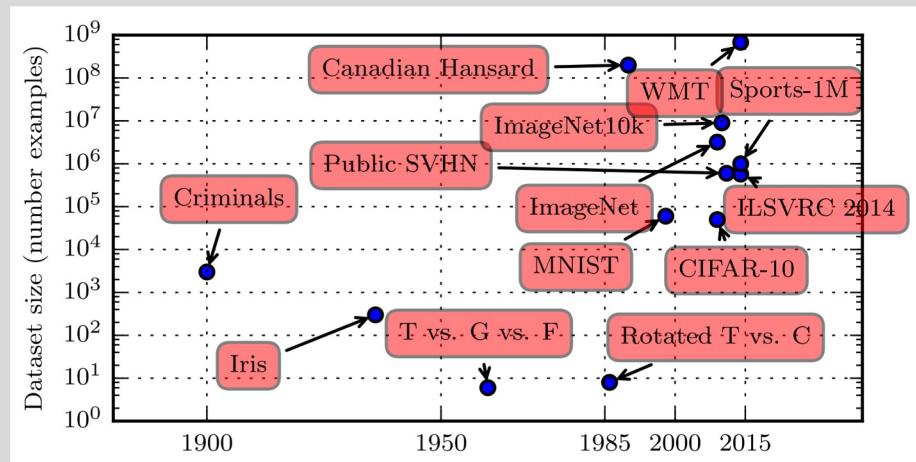
Outline

- The importance of data
- How to build a dataset
- Tools

Main Reasons Behind Deep Learning's Success



Hardware



Data

Dataset used to train GPTs



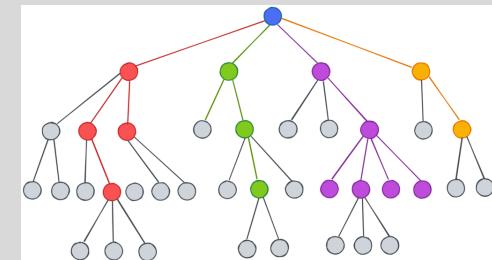
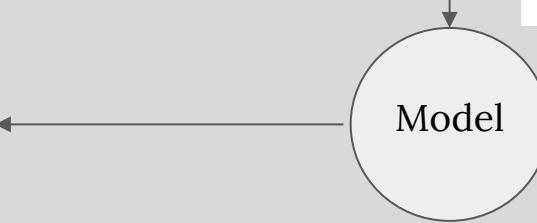
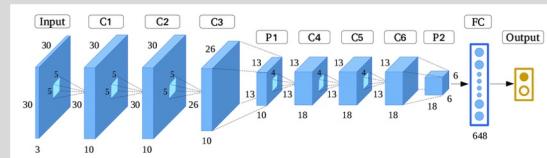
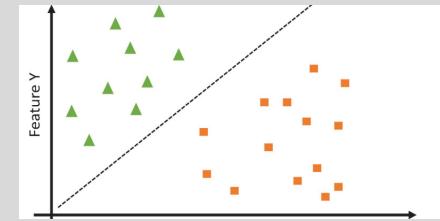
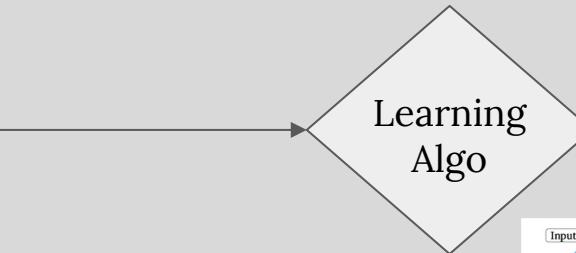
The Data ▾ Resources ▾ Community ▾ About ▾ Search ▾ Contact Us

[Stats](#)

Common Crawl
maintains a **free, open**
repository of web crawl
data that can be used by
anyone.

Common Crawl is a 501(c)(3) non-profit founded in 2007.

Data in a learning system pipeline



Importance of Dataset

Datasets are used as benchmarks to compare learning systems

IM²GENET Large Scale Visual Recognition Challenge 2017 (ILSVRC2017)

DET LOC VID Team information

Legend:
Yellow background = winner in this task according to this metric; authors are willing to reveal the method
White background = authors are willing to reveal the method
Grey background = authors chose not to reveal the method
Italics = authors requested entry not participate in competition

Object detection (DET)^[top]

Task 1a: Object detection with provided training data

Ordered by number of categories won

Team name	Entry description	Number of object categories won	mean AP
BDAT	submission4	85	0.731392
BDAT	submission3	65	0.732227
BDAT	submission2	30	0.723712
DeepView(ETRI)	Ensemble_A	10	0.593084
NUS-Qihoo_DPNs (DET)	Ensemble of DPN models	9	0.656932
KAISTNIA_ETRI	Ensemble Model5	1	0.61022
KAISTNIA_ETRI	Ensemble Model4	0	0.609402
KAISTNIA_ETRI	Ensemble Model2	0	0.608299
KAISTNIA_ETRI	Ensemble Model1	0	0.608278
KAISTNIA_ETRI	Ensemble Model3	0	0.60631
DeepView(ETRI)	Single model A using ResNet for detection	0	0.587519

Importance of Dataset

Datasets are used as benchmarks to compare learning systems

Russian–English			
#	score	range	system
1	0.583	1	AFRL-PE
2	0.299	2	ONLINE-B
3	0.190	3-5	ONLINE-A
	0.178	3-5	PROMT-HYBRID
	0.123	4-7	PROMT-RULE
	0.104	5-8	UEDIN-PHRASE
	0.069	5-8	Y-SDA
	0.066	5-8	ONLINE-G
4	-0.017	9	AFRL
5	-0.159	10	UEDIN-SYNTAX
6	-0.306	11	KAZNU
7	-0.487	12	RBMT1
8	-0.642	13	RBMT4

English–Russian			
#	score	range	system
1	0.575	1-2	PROMT-RULE
	0.547	1-2	ONLINE-B
2	0.426	3	PROMT-HYBRID
3	0.305	4-5	UEDIN-UNCNSTR
	0.231	4-5	ONLINE-G
4	0.089	6-7	ONLINE-A
	0.031	6-7	UEDIN-PHRASE
5	-0.920	8	RBMT4
6	-1.284	9	RBMT1

French–English			
#	score	range	system
1	0.608	1	UEDIN-PHRASE
2	0.479	2-4	KIT
	0.475	2-4	ONLINE-B
	0.428	2-4	STANFORD
3	0.331	5	ONLINE-A
4	-0.389	6	RBMT1
5	-0.648	7	RBMT4
6	-1.284	8	ONLINE-C

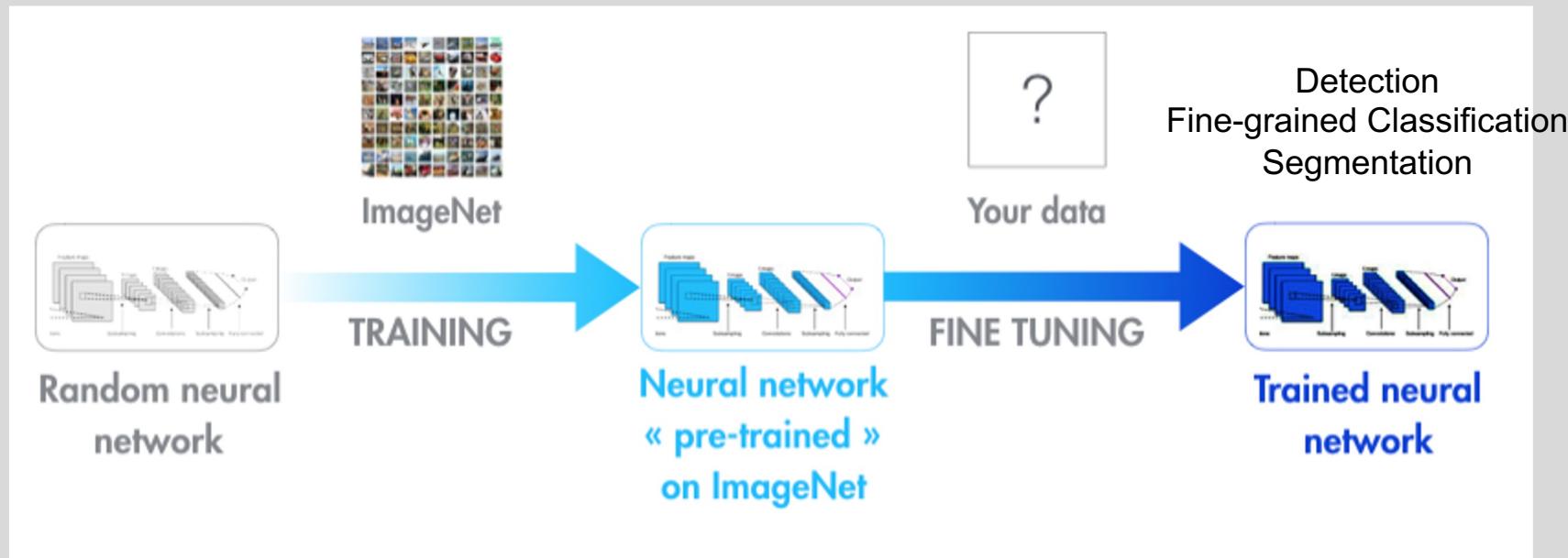
Hindi–English			
#	score	range	system
1	1.326	1	ONLINE-B
2	0.559	2-3	ONLINE-A
	0.476	2-4	UEDIN-SYNTAX
	0.434	3-4	CMU
3	0.323	5	UEDIN-PHRASE
4	-0.198	6-7	AFRL
	-0.280	6-7	IIT-BOMBAY
5	-0.549	8	DCU-LINGO24
6	-2.092	9	IIT-HYDERABAD

English–French			
#	score	range	system
1	0.327	1	ONLINE-B
2	0.232	2-4	UEDIN-PHRASE
	0.194	2-5	KIT
	0.185	2-5	MATRAN
	0.142	4-6	MATRAN-RULES
	0.120	4-6	ONLINE-A
3	0.003	7-9	UU-DOCENT
	-0.019	7-10	PROMT-HYBRID
	-0.033	7-10	UA
	-0.069	8-10	PROMT-RULE
4	-0.215	11	RBMT1
5	-0.328	12	RBMT4
6	-0.540	13	ONLINE-C

English–Hindi			
#	score	range	system
1	1.008	1	ONLINE-B
2	0.915	2	ONLINE-A
3	0.214	3	UEDIN-UNCNSTR
4	0.120	4-5	UEDIN-PHRASE
	0.054	4-5	CU-MOSES
5	-0.111	6-7	IIT-BOMBAY
	-0.142	6-7	IPN-UPV-CNTXT
6	-0.233	8-9	DCU-LINGO24
	-0.261	8-9	IPN-UPV-NODEV
7	-0.449	10-11	MANAWI-H1
	-0.494	10-11	MANAWI
8	-0.622	12	MANAWI-RMOOV

Importance of Dataset

Datasets are used to learn a general purpose prior



<https://medium.com/owkin/transfer-learning-and-the-rise-of-collaborative-artificial-intelligence-41f9e2950657>

Importance of Dataset

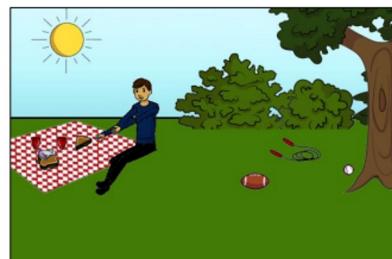
New datasets inspire novel algorithms and research problems



What color are her eyes?
What is the mustache made of?



How many slices of pizza are there?
Is this a vegetarian pizza?



Is this person expecting company?
What is just under the tree?



Does it appear to be rainy?
Does this person have 20/20 vision?

Visual Q&A

Who is wearing glasses?

man



woman



Where is the child sitting?

fridge



arms

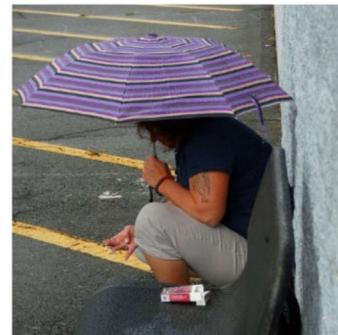


Is the umbrella upside down?

yes



no



How many children are in the bed?

2



1



Importance of Dataset

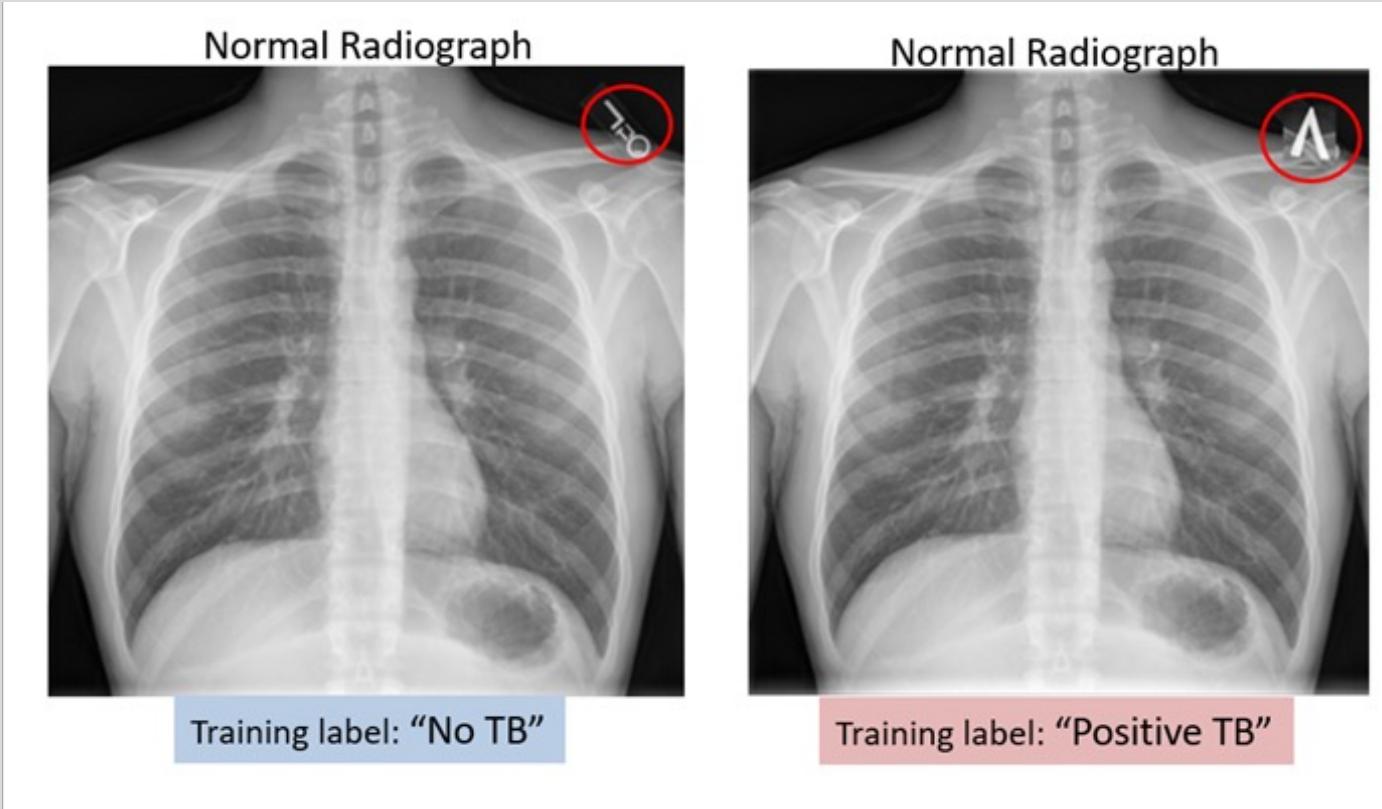
New datasets inspire novel algorithms and research problems



Recommendation Systems

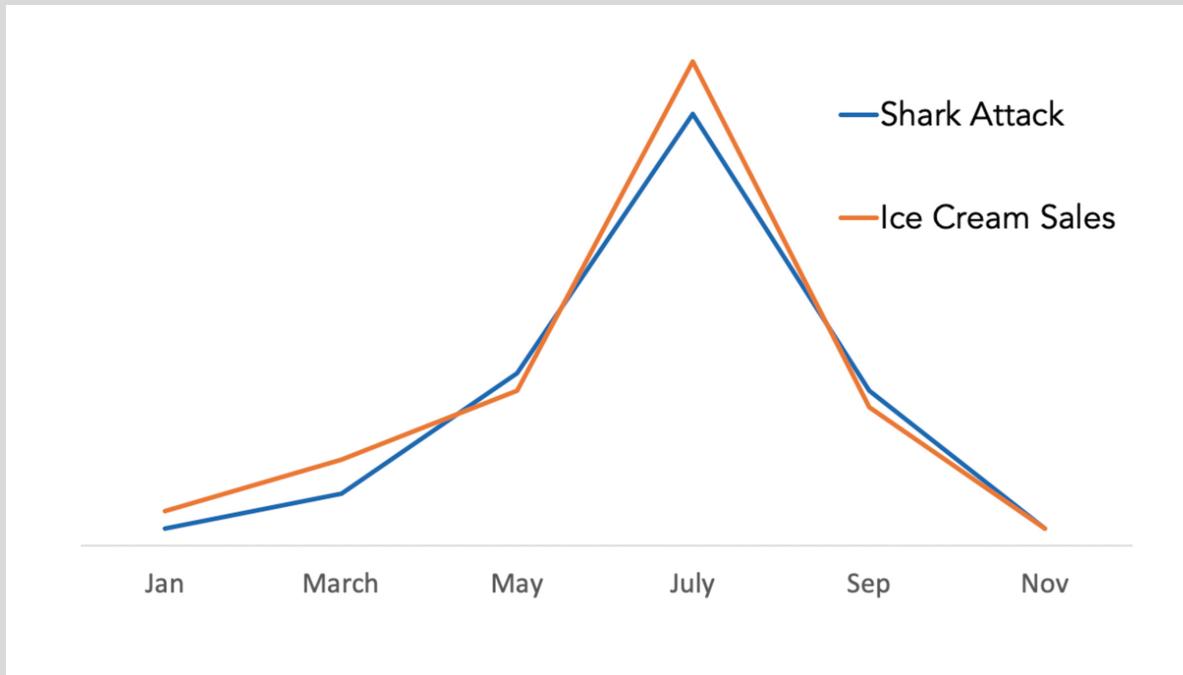
Music, books, videos
Online shopping
Financial
Online dating
...

You can have a wrong model from wrong data



Correlation vs Causation

Car allergic to vanilla ice cream?



Formulation of learning:

$$\min_f \mathbb{E}_{X,Y}[\mathcal{L}(Y, f(X))] \quad \text{Expected error}$$

$$\min_f \frac{1}{n} \sum_{i=1}^n [\mathcal{L}(Y_i, f(X_i))] \quad \text{Sample mean}$$

$$\min_{f \in \mathcal{F}} \frac{1}{n} \sum_{i=1}^n [\mathcal{L}(Y_i, f(X_i))] \quad \text{Search over restricted class of functions}$$

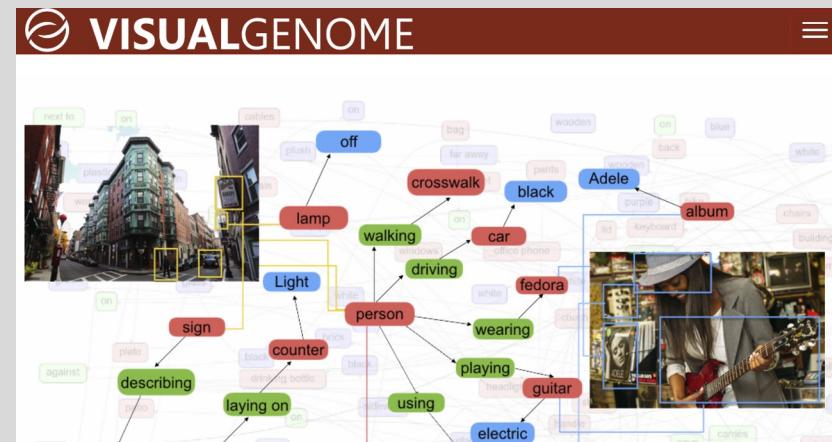
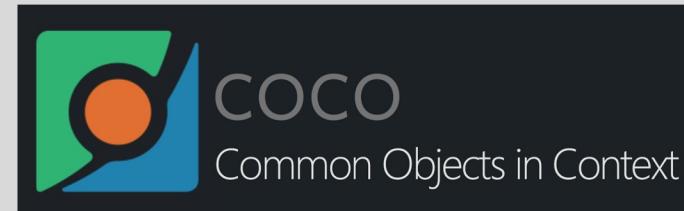
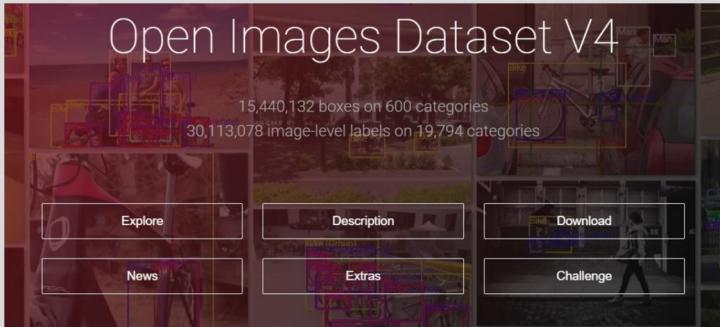
Data distribution is central:

$$X_i, Y_i \sim p(X, Y)$$

Outline

- The importance of data
- How to build a dataset
- Tools

Existing Dataset - Vision



LAION

Projects

Team

Blog

Notes

Press

About

FAQ

Donations

Privacy Policy

Dataset Requests

Impressum



LAION-5B: A NEW ERA OF OPEN LARGE-SCALE MULTI- MODAL DATASETS

by: Romain Beaumont, 31 Mar, 2022

We present a dataset of 5.85 billion CLIP-filtered image-text pairs, 14x bigger than LAION-400M, previously the biggest openly accessible image-text dataset in the world - see also our [NeurIPS2022 paper](#)

Authors: Christoph Schuhmann, Richard Vencu, Romain Beaumont, Theo Coombes, Cade Gordon, Aarush Katta, Robert Kaczmarczyk, Jenia Jitsev

Backend url:
<https://knn5.laion>

Index:
laion_5B

[Clip retrieval](#) works by converting the text query to a CLIP embedding, then using that embedding to query a knn index of clip image embeddings

Display captions Display full

french cat


french cat


french cat


How to tell if your feline is french. He wears a b...


Hilarious pics of funny cats! [funnycatsgif.com](#)



Existing Dataset - Natural Language

IMDB Reviews (Sentiment Analysis)

1 Billion Word Language Model Benchmark (Language Modeling)

WordNet (Database for English ‘synsets’)

Google Books Ngrams

Existing Dataset - Others

[HealthData.gov](#) (Health Care)

[OASIS brain images](#)

[Data.gov](#) (agriculture, climate, ecosystems, public safety...)

[Kaggle Dataset](#)

Why Build Your Own Dataset

Variation

Existing datasets do not contain enough variety.

E.g., non-traditional lighting and poses.

Annotation

Existing datasets do not provide the information you need.

E.g., no object segmentation masks in ImageNet.

Case I:

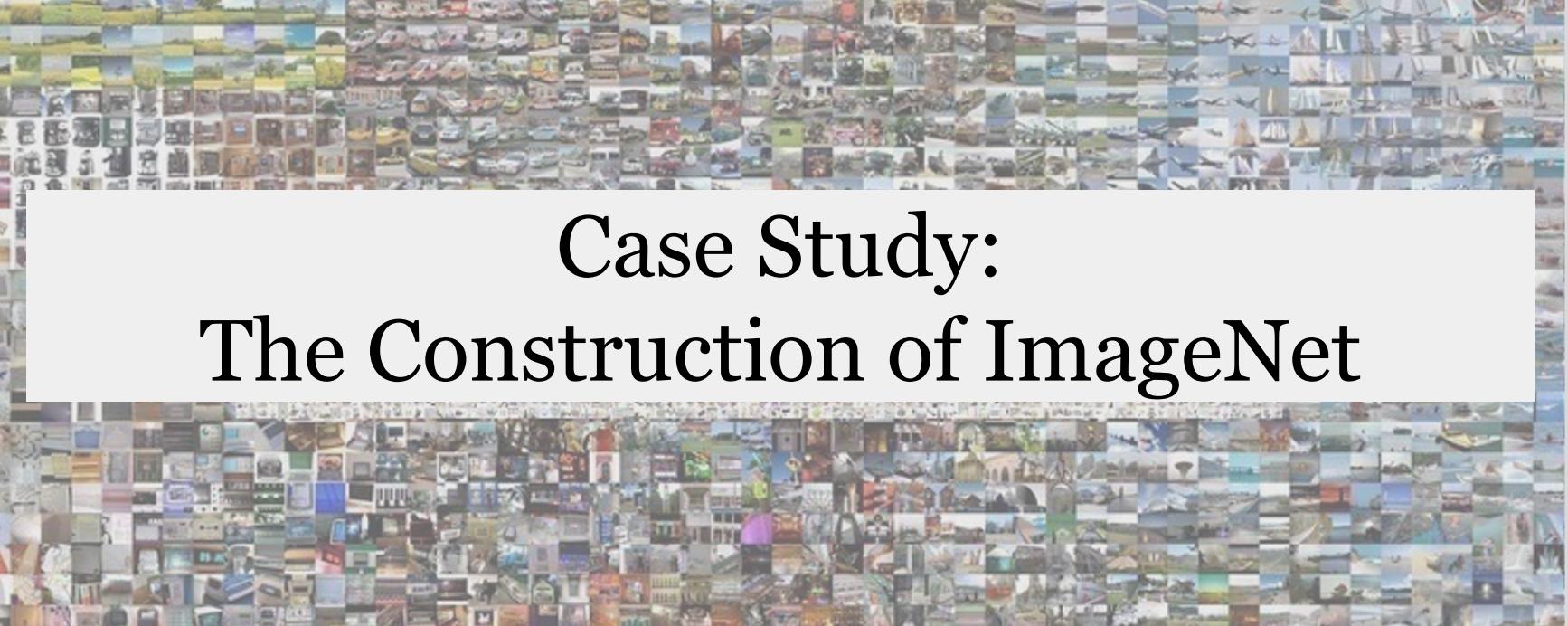
The construction of ImageNet.

Case II:

The construction of COCO.

Case III:

The construction of Waymo.



Case Study: The Construction of ImageNet

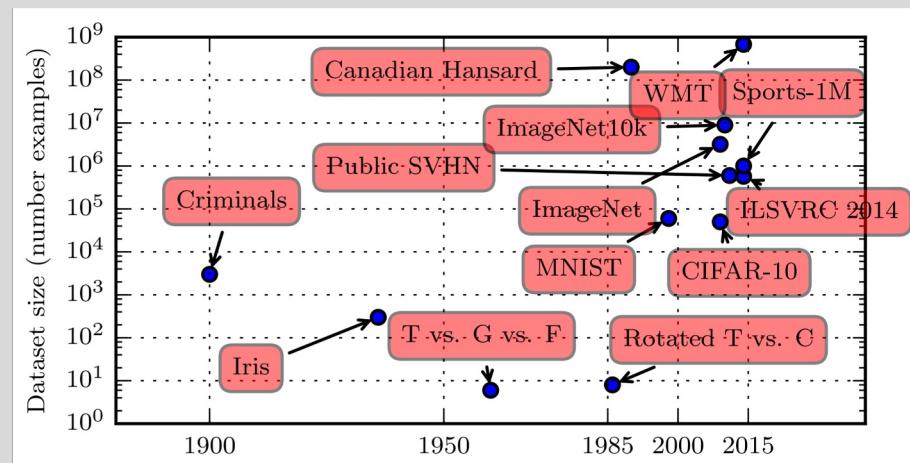
Purpose of the dataset

“We believe that a large-scale ontology of images is a critical resource for developing advanced, large-scale content-based image search and image understanding algorithms, as well as for providing critical training and benchmarking data for such algorithms.”

Before ImageNet, computer vision datasets have hundreds or thousands samples.

ImageNet has over 3 Million (paper version).

It is a visionary move as we now know better network models are obtained from more data.



How to build a dataset to capture the visual world

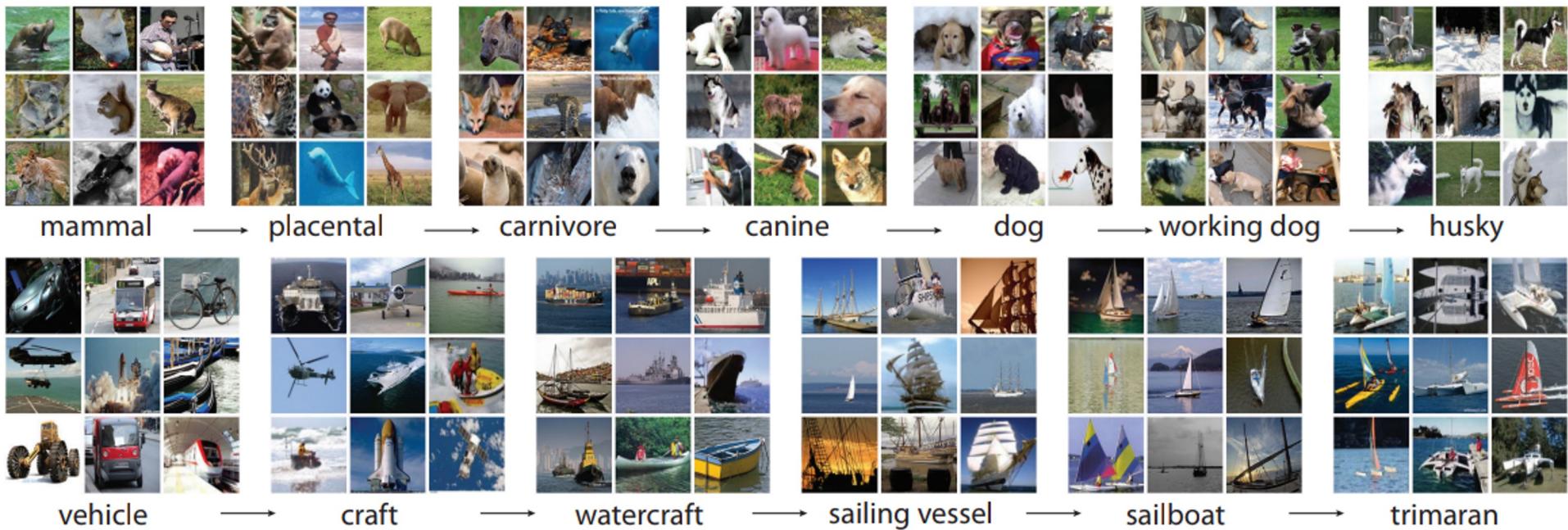


How to build a dataset to capture the visual world



A hierarchical structure that describes concepts in the world.

ImageNet Hierarchy



Construction

Collecting candidate images

Goal: About 500-1000 images per synet (i.e., visual concepts).

Perform keyword based image search. (10% accuracy)

Collected 10k per synet.

Some tricks:

- Expand search queries

- Multiple languages

Construction

Cleaning images

Human verification by Amazon MTurk.

For each task, an image obtained by searching is presented with the definition of the synet.

The user is asked to see if they are consistent.

Construction

Cleaning images

Majority vote over the categories and confidence.

More confidence score is needed for finer classes.

				
User 1	Y	Y	Y	
User 2	N	Y	Y	
User 3	N	Y	Y	
User 4	Y	N	Y	
User 5	Y	Y	Y	
User 6	N	N	Y	

#Y	# N	Conf Cat	Conf BCat
0	1	0.07	0.23
1	0	0.85	0.69
1	1	0.46	0.49
2	0	0.97	0.83
0	2	0.02	0.12
3	0	0.99	0.90
2	1	0.85	0.68

Dataset examples

Case Study: The Construction of COCO Dataset

COCO Dataset Statistics

What is COCO?



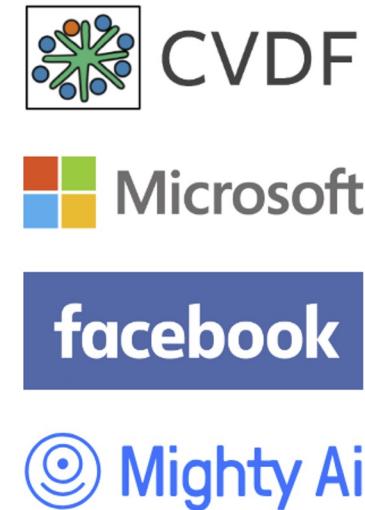
COCO is a large-scale object detection, segmentation, and captioning dataset. COCO has several features:

- ✓ Object segmentation
- ✓ Recognition in context
- ✓ Superpixel stuff segmentation
- ✓ 330K images (>200K labeled)
- ✓ 1.5 million object instances
- ✓ 80 object categories
- ✓ 91 stuff categories
- ✓ 5 captions per image
- ✓ 250,000 people with keypoints

Collaborators

- Tsung-Yi Lin Google Brain
Genevieve Patterson MSR, Trash TV
Matteo R. Ronchi Caltech
Yin Cui Cornell Tech
Michael Maire TTI-Chicago
Serge Belongie Cornell Tech
Lubomir Bourdev WaveOne, Inc.
Ross Girshick FAIR
James Hays Georgia Tech
Pietro Perona Caltech
Deva Ramanan CMU
Larry Zitnick FAIR
Piotr Dollár FAIR

Sponsors

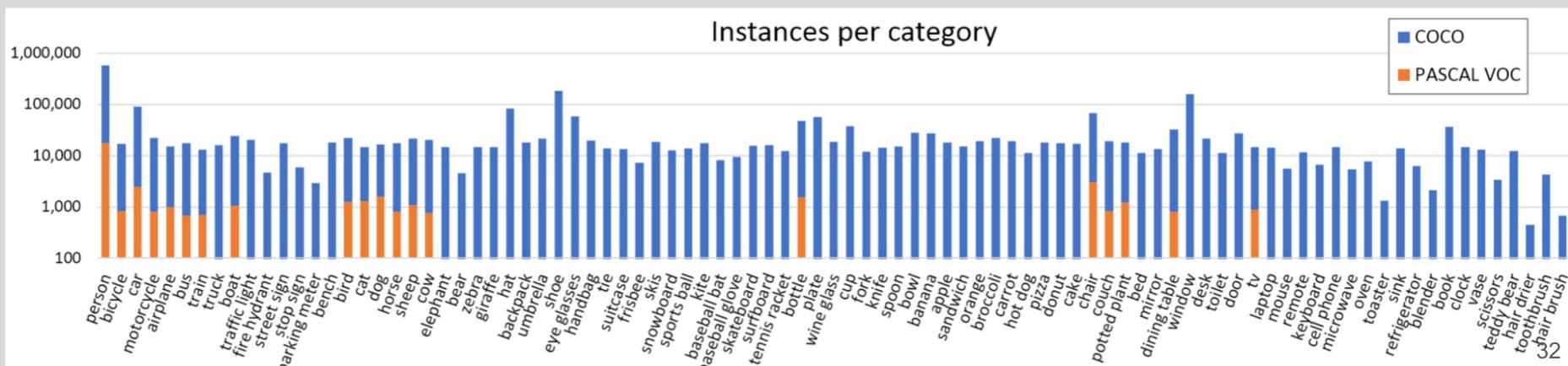


Data Collection

Identify Object Categories

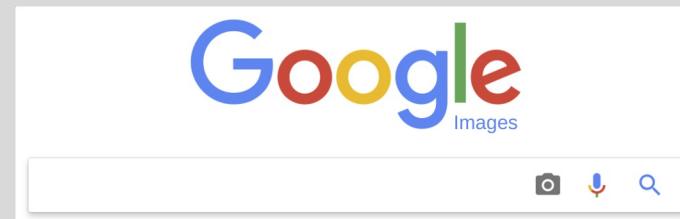
PASCAL VOC + frequently used words for objects + survey on 4-8 years old children = 272 candidates

Voting to get final categories: 91.



Data Collection

Collect Images For Each Object Category



Iconic Images

Data Collection

Collect Images For Each Object Category 328,000 images in total.



A screenshot of the Flickr website interface. At the top, there is a navigation bar with the Flickr logo, 'Explore', 'Create', and 'Get Pro' buttons, and a search bar containing the placeholder 'Photos, people, or groups'. Below the navigation bar, there are three tabs: 'Photos' (which is selected), 'People', and 'Groups'. To the right of the tabs is a color palette consisting of a row of colored squares followed by a larger set of decorative patterns. Underneath the tabs, there are two dropdown menus: 'Any license' and 'SafeSearch on'. At the bottom of the page is another search bar with the same placeholder and a blue 'Search' button.

Non-Iconic Images

Data Annotation

Crowdsourcing to label over 2.5 million object instances in 300K+ images.

Annotation Pipeline



(a) Category labeling

8 Workers Per Image

~20k Worker Hours



(b) Instance spotting

8 Workers Per Image

~10k Worker Hours

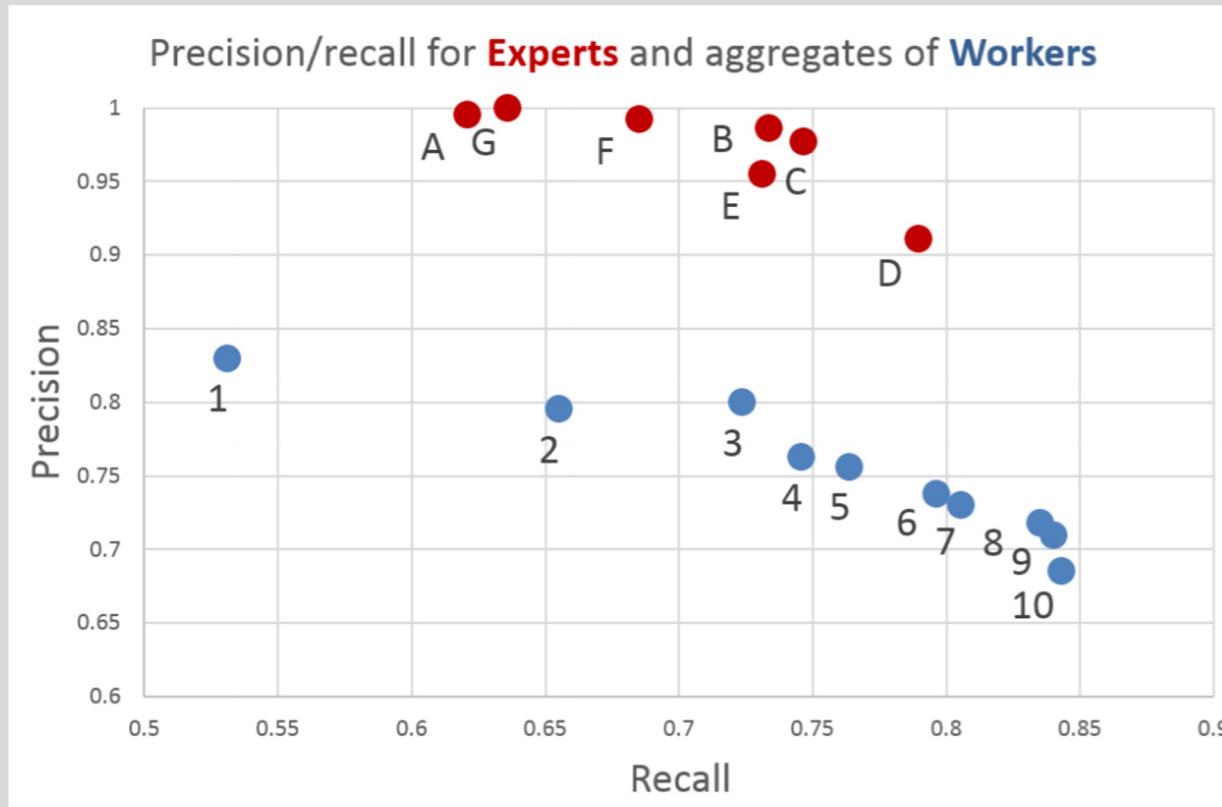


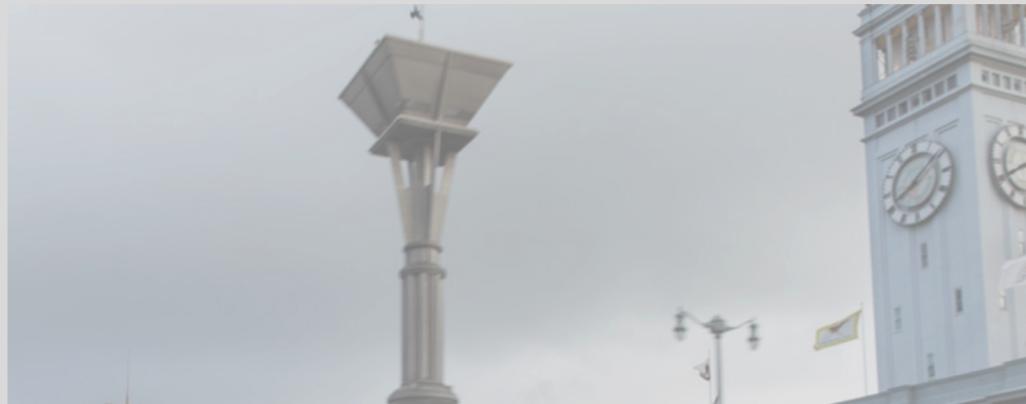
(c) Instance segmentation

Only 1 worker per image.

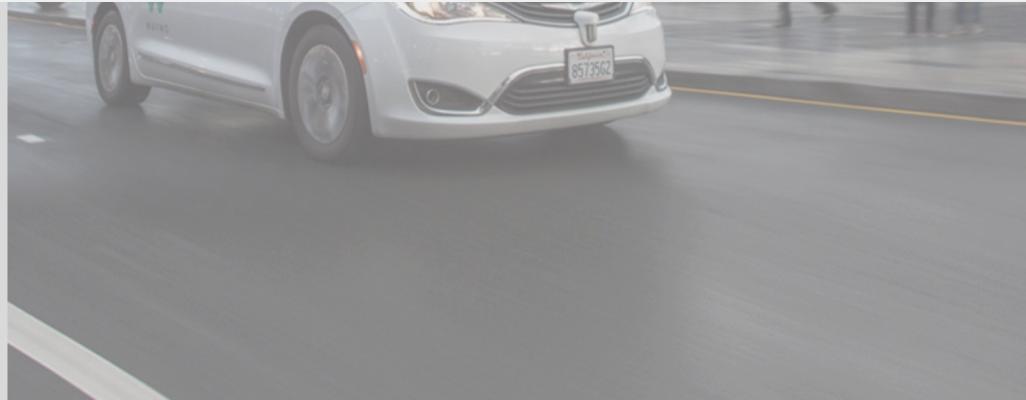
Training stage enforced.

Data Verification



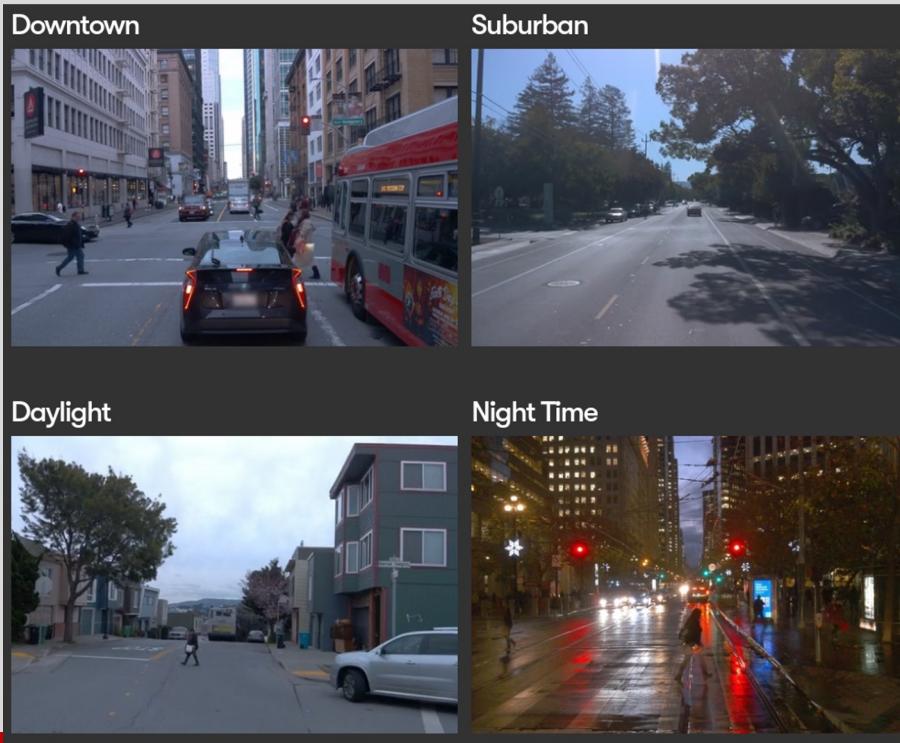


Case Study: The Construction of Waymo Dataset

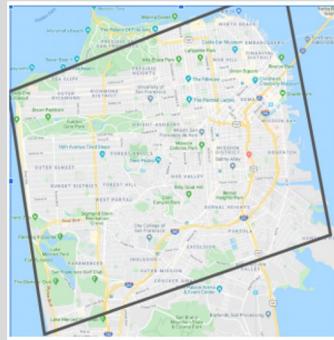
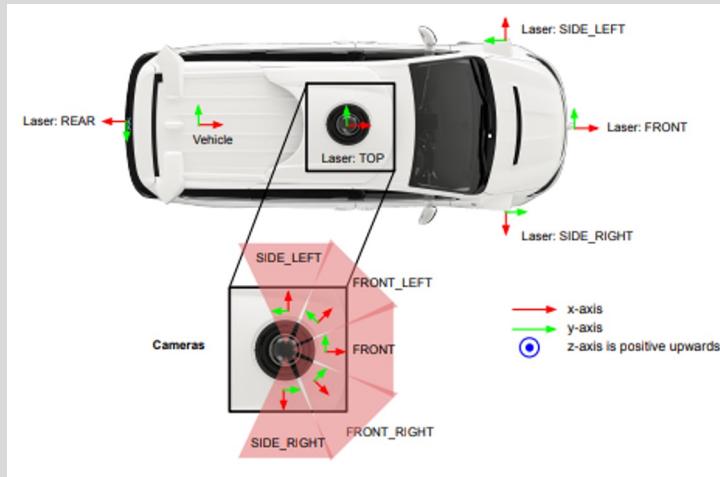


Purpose of the dataset

Existing self-driving dataset is limited in the scale and variation of the environments.

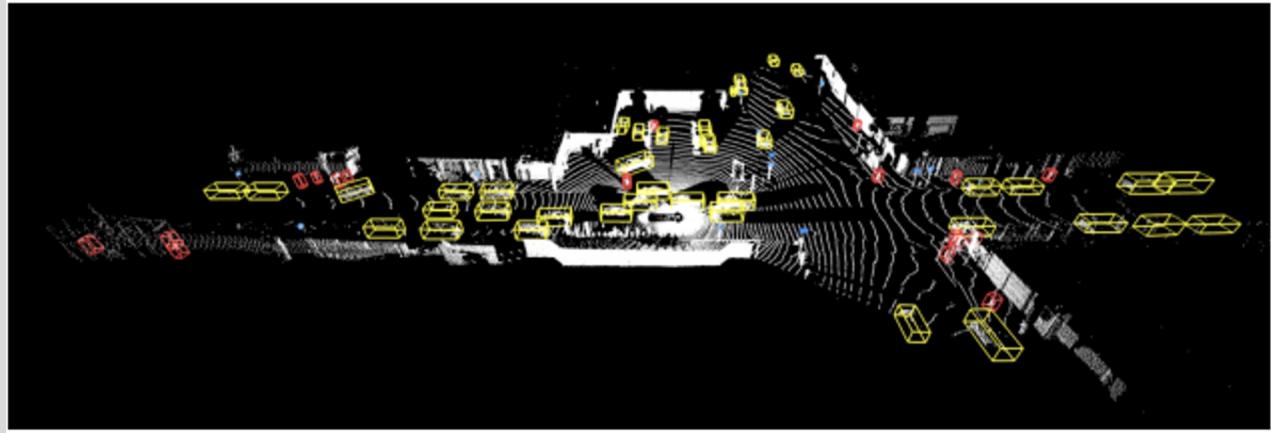


Setup



1150 scenes, each is a 20-second long video.

Labeling



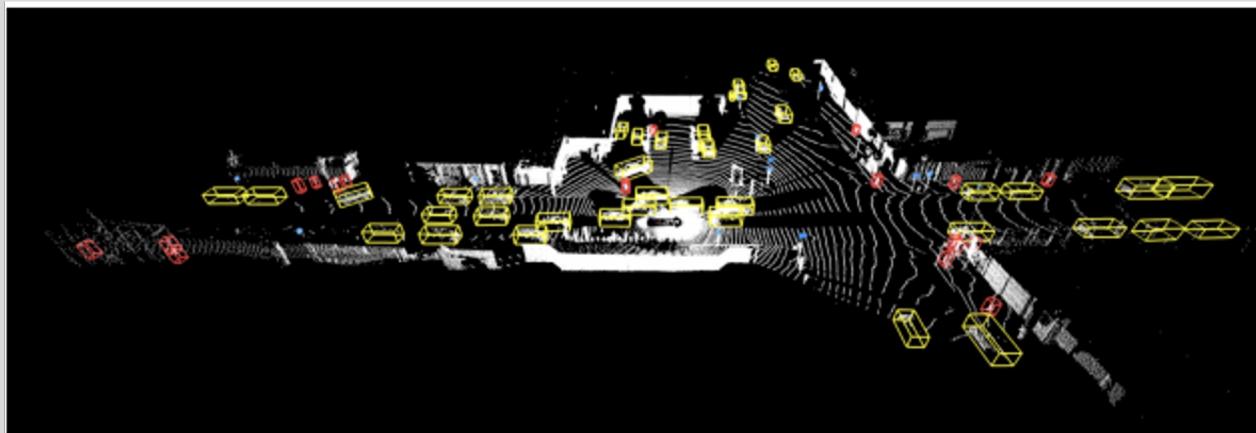
Camera and vehicle poses.

3D bounding cuboids of four categories: pedestrian, car, sign, cyclist.

2D bounding boxes in images.

Identity of objects.

Labeling



	Vehicle	Pedestrian	Cyclist	Sign
3D Object	6.1M	2.8M	67K	3.2M
3D TrackID	60K	23K	620	23K
2D Object	7.7M	2.1M	63K	–
2D TrackID	164K	45K	1.3K	–

Summary: things need to be considered

Purpose and design:	Do you have to build your own? What do you want?
Where to get data:	Web-crawling? Collect in real-life?
How many data to collect:	Within your resource limits, more the better.
Tools to use:	Download, collect, label.
How to verify the collected data: labeler good?	Is your data good? Is your labeler good?

Legal issues on data collection



A Startup Is Suing Facebook, Princeton For Stealing Its AI Data

The legal wars over AI training data are just beginning.

hiQ Labs, Inc. v. LinkedIn Corp., No. 17-16783 (9th Cir. 2019)

Annotate this Case

Justia Opinion Summary

The Ninth Circuit affirmed the district court's grant of a preliminary injunction in favor of hiQ, a data analytics company, prohibiting LinkedIn, a professional networking website, from denying hiQ access to publicly available LinkedIn member profiles.

The panel held that the district court did not abuse its discretion in concluding that hiQ currently has no viable way to remain in business other than using LinkedIn public profile data for its Keeper and Skill Mapper services, and that hiQ therefore has demonstrated a likelihood of irreparable harm absent a preliminary injunction. The panel also held that the district court's determination that the balance of hardships tips sharply in hiQ's favor was not illogical, implausible, or without support in the record; hiQ raised serious questions regarding the merits of its tortious interference with contract claim and LinkedIn's legitimate business purpose defense; hiQ also raised a serious question regarding whether state law causes of action were preempted by the Computer Fraud and Abuse Act; and the district court's conclusion that the public interest favors granting the preliminary injunction was appropriate.

[Collapse Summary](#)

Outline

- The importance of data
- How to build a dataset
- Tools

Tools

Data Source

Google/Bing Search, Flickr, Instagram, Google Map/Streetview, Satellite

Visual Annotation

[VGG Image Annotator](#), [Video Annotation Tool](#), [Scalabel](#)

Or Build your own (HTML+JS)

Crowdsourcing

Amazon MTurk

Amazon Mechanical Turk Tutorial



On Demand

Over 500K workers, 24x7

Speed

Work is done in parallel

Scalable

No minimum project size

Qualification

Set prerequisite to workers

Amazon MTurk Marketplace

The screenshot shows the Amazon MTurk Marketplace interface. At the top, there's a navigation bar with the MTurk logo, tabs for 'HITs', 'Dashboard', and 'Qualifications', a search bar labeled 'Search All HITs' with a magnifying glass icon, and a 'Filter' button. Below the navigation is a secondary header with 'All HITs' and 'Your HITs Queue' buttons.

The main content area is titled 'HIT Groups (1-20 of 2106)'. It features a table with columns: Requester, Title, HITs, Reward, Created, and Actions. The 'Actions' column contains buttons like 'Accept & Work', 'Preview', and 'Qualify'.

A tooltip is visible over the second HIT listing, which is from 'Amazon Requester Inc. - C'. The tooltip displays the requester's name, the hit title in French, and a detailed description of the hit: '[French language proficiency required] Questionnaire sur la relativité des produits aux intérêts (répondre par oui ou non)'. The tooltip also shows the hit ID, reward (\$0.01), creation time (7h ago), and preview link.

Requester	Title	HITs	Reward	Created	Actions
Amazon Requester Inc. - C	[French language proficiency requir...	61,046	\$0.01	17h ago	Preview <button>Accept & Work</button>
Amazon Requester Inc. - C	[日本語能力が必	[French language proficiency required] Questionnaire sur la relativité des produits aux intérêts (répondre par oui ou non)	\$0.01	7h ago	Preview <button>Accept & Work</button>
Amazon Requester Inc. - C	Product to Interest Audit (single yes/...		\$0.01	1h ago	Preview <button>Accept & Work</button>
Amazon Requester Inc. - C	[dominio del idioma español requiri...	27,670	\$0.01	21h ago	Preview <button>Accept & Work</button>
Amazon Requester Inc. - C	[Proficiência no idioma português br...	19,719	\$0.01	20h ago	Preview <button>Accept & Work</button>
Crowdsurf Support	Transcribe up to 35 Seconds of Med...	17,485	\$0.05	3m ago	Preview <button>Qualify</button>
TC Research	Find the Email for These Mental He...	13,896	\$0.12	5d ago	Preview <button>Accept & Work</button>
UnSpun Opinions	Opinion Survey	12,180	\$0.50	1m ago	Preview <button>Accept & Work</button>
KronoPin	Find the Website Address for a Con...	11,846	\$0.03	2/23/2018	Preview <button>Qualify</button>
Assistive Technology Rese	1 minute survey: Smart speakers at ...	10,577	\$0.15	3d ago	Preview <button>Accept & Work</button>

MTurk Concepts

Requesters

Person creates tasks for Workers to work on.

Human Intelligence Tasks (HITs)

HIT is a single, self-contained task.

Assignment

Multiple Workers can be assigned to a single HIT.

A Worker can only accept a HIT once and submit one assignment per HIT.

Workers

Person completes assignments.

Approval and Payment

After assignment submission, if you approve the work, the HIT reward is drawn from your MTurk account.

Qualification

Anyone can register as a worker. You can set qualification types such as approval rate to control the quality of submissions.

Common Use Cases

Image/Video Processing

MTurk is well-suited for processing images. While difficult for computers, it is a task that is extremely easy for people to do. In the past, companies have used MTurk to:



Tag objects found in an image to improve your search or advertising targeting

Review a set of images to select the best picture to represent a product



Audit user-uploaded images or videos to moderate content

Classify objects found in satellite imagery

Data Verification and Clean-up

Companies with large online directories or catalogs are using MTurk to identify duplicate entries and verify item details. Examples of this have included:



Removing duplicate content from business listings

Identifying incomplete or duplicate product listings in a catalog



Verifying restaurant details such as phone numbers or hours of operation

Converting unstructured data about locations into well-formed addresses

Common Use Cases

Information Gathering

The diversification and the scale of the MTurk workforce allows you to gather a breadth of information that would be almost impossible to do otherwise such as:



Allowing people to ask questions from a computer or mobile device about any topic and have Workers return the results



Filling out market research or survey data on a variety of topics



Writing content for websites



Finding specific fields or data elements in large legal and government documents

Data Processing

Companies take advantage of the power of the MTurk workforce to understand and intelligently respond to different types of data including:



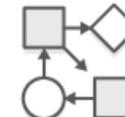
Audio editing and transcription



Human powered translation services



Rating the accuracy of results for a search engine



Categorizing information to match a given schema or taxonomy

Example: Data Labeling Using MTurk

1. Setup

Python and Boto3 (AWS SDK).

2. Accounts

AWS and MTurk (Also need to link the two).

Purchasing Prepaid HITs.

3. Creating Tasks

Define a HIT and its reward.

4. Retrieving Results

Verify result, Add a bonus

No coding needed:
[Tutorial 1](#)

Command line approach:
[Tutorial 2](#)

Things to keep in mind

1. Turkers are humans, not robots.

Try to make reasonable tasks, don't expect they finish something a normal people won't do.

1. It is a real-world labor market.

Turkers have their own forums and communities. They check and compare your rate with others.

1. Check the quality of your labels!

You don't want to pay money for useless labels.

Data checklist

- Problem domain
- Where can you get the data?
- How many data can you get?
- Data format
- Define the learning problem
- Do you need additional annotation?
- How do you evaluate?