Shifting the Frame: The Labors of ImageNet and AI Data

Alex Hanna
Director of Research
Distributed AI Research Institute

The data that transformed AI research—and possibly the world

Image: AP Photo/Jeff Chiu

Stanford professor and Google Cloud chief scientist Fei-Fei Li changed everything.

The ImageNet Dataset

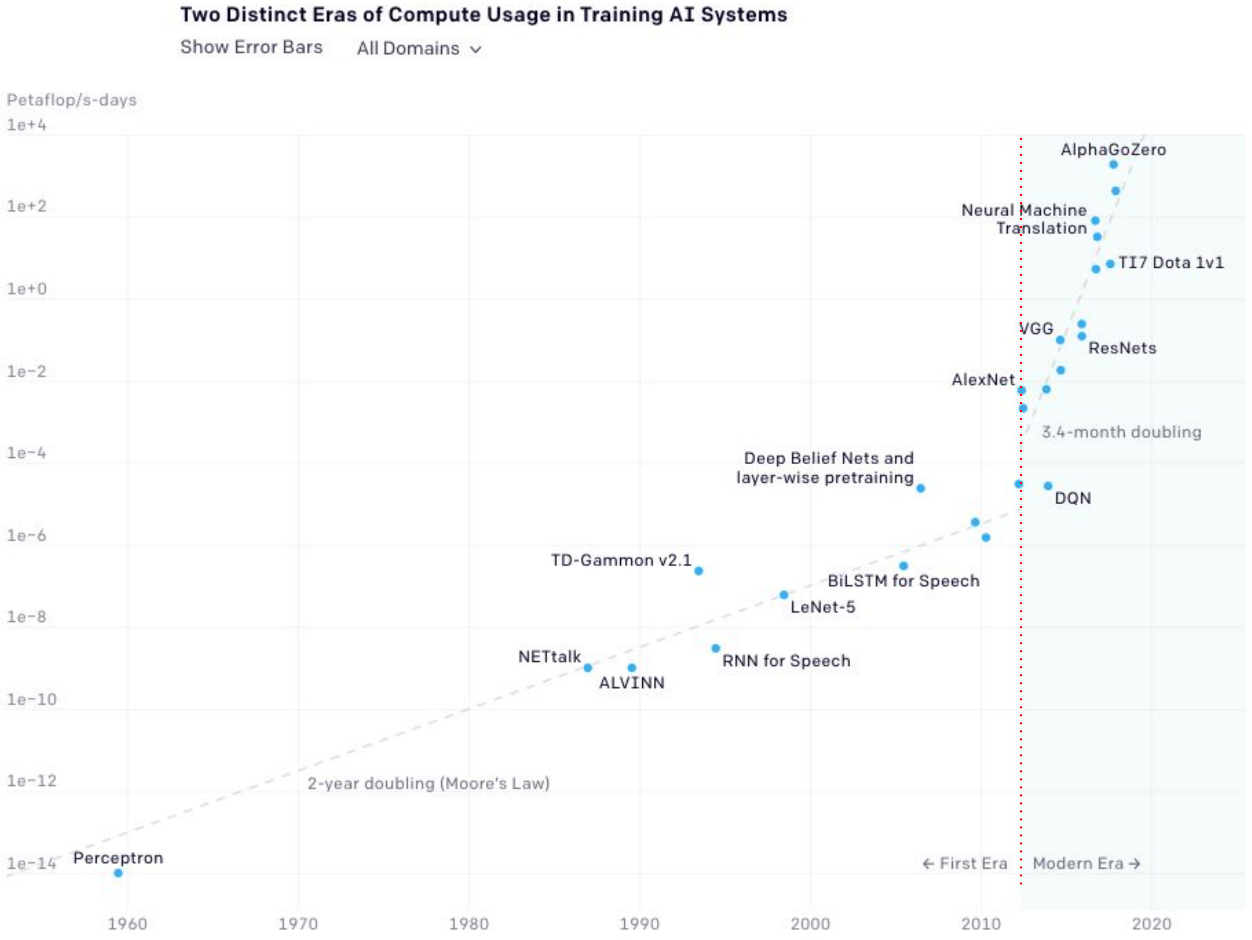
Effort to "map out the entire world of objects"

Over 14 million images
Over 20,000 categories

Regarded as a key benchmark

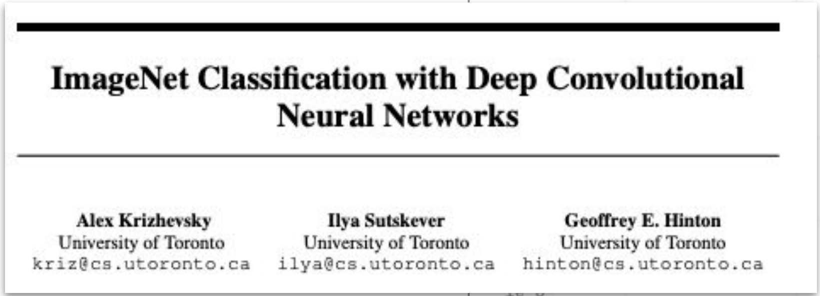
http://www.image-net.org

The ImageNet Challenge



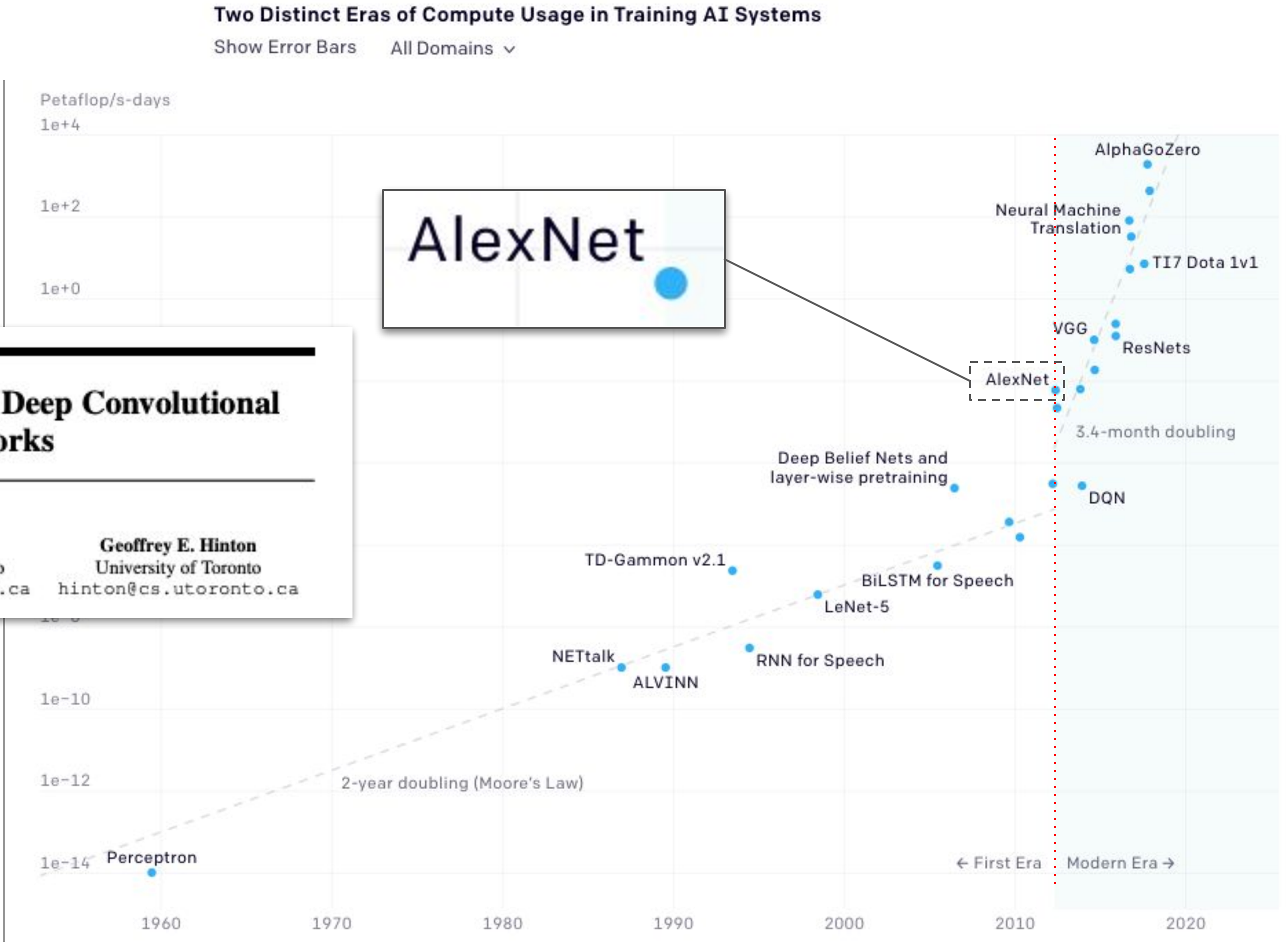
OpenAI. 2018. "AI + Compute."

The ImageNet Challenge



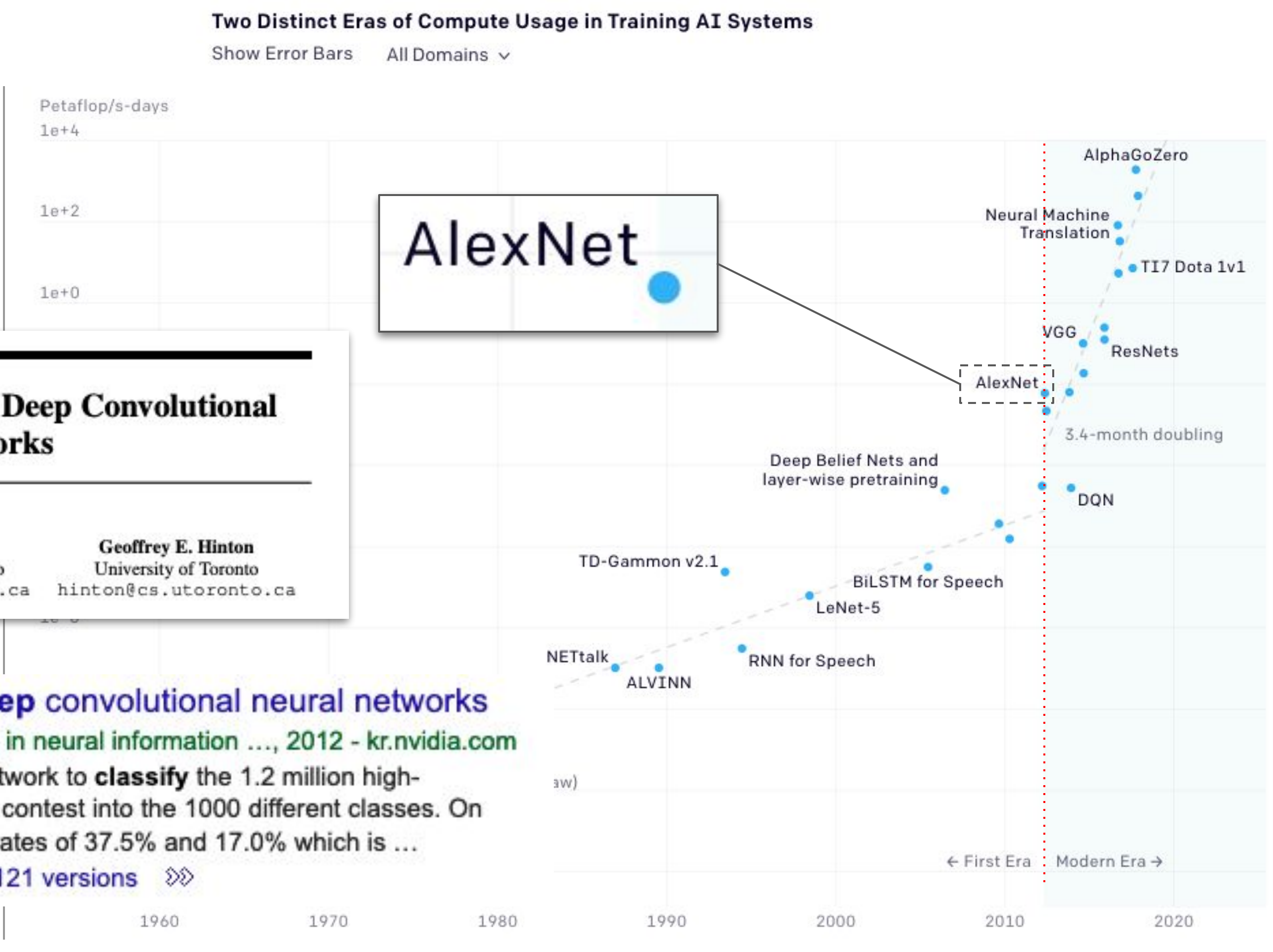
ImageNet Classification with Deep Convolutional Neural Networks

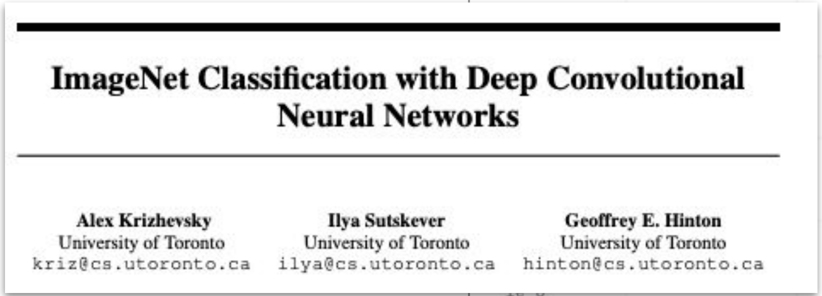
Alex Krizhevsky	Ilya Sutskever	Geoffrey E. Hinton
University of Toronto	University of Toronto	University of Toronto
kriz@cs.utoronto.ca	ilya@cs.utoronto.ca	hinton@cs.utoronto.ca



OpenAI. 2018. "AI + Compute."

The ImageNet Challenge





ImageNet Classification with Deep Convolutional Neural Networks

Alex Krizhevsky
University of Toronto
kriz@cs.utoronto.ca

Ilya Sutskever
University of Toronto
ilya@cs.utoronto.ca

Geoffrey E. Hinton
University of Toronto
hinton@cs.utoronto.ca

[PDF] **Imagenet classification** with **deep** convolutional neural networks

A Krizhevsky, I Sutskever, GE Hinton - Advances in neural information ..., 2012 - kr.nvidia.com

We trained a large, **deep** convolutional neural network to **classify** the 1.2 million high-resolution images in the **ImageNet** LSVRC-2010 contest into the 1000 different classes. On the test data, we achieved top-1 and top-5 error rates of 37.5% and 17.0% which is ...

☆ 99 Cited by 77971 Related articles All 121 versions »

OpenAI. 2018. "AI + Compute."

Papers with Code

CIFAR-10

The CIFAR-10 dataset (Canadian Institute for Advanced Research, 10 classes) is a subset of the Tiny Images dataset and consists of 60000 32x32 color images. The images are labell...

11,018 PAPERS • 69 BENCHMARKS

ImageNet

The ImageNet dataset contains 14,197,122 annotated images according to the WordNet hier-archy. Since 2010 the dataset is used in the ImageNet Large Scale Visual Recognition...

10,614 PAPERS • 102 BENCHMARKS

Papers with Code

CIFAR-10

The CIFAR-10 dataset (Canadian Ins
the Tiny Images dataset and consist

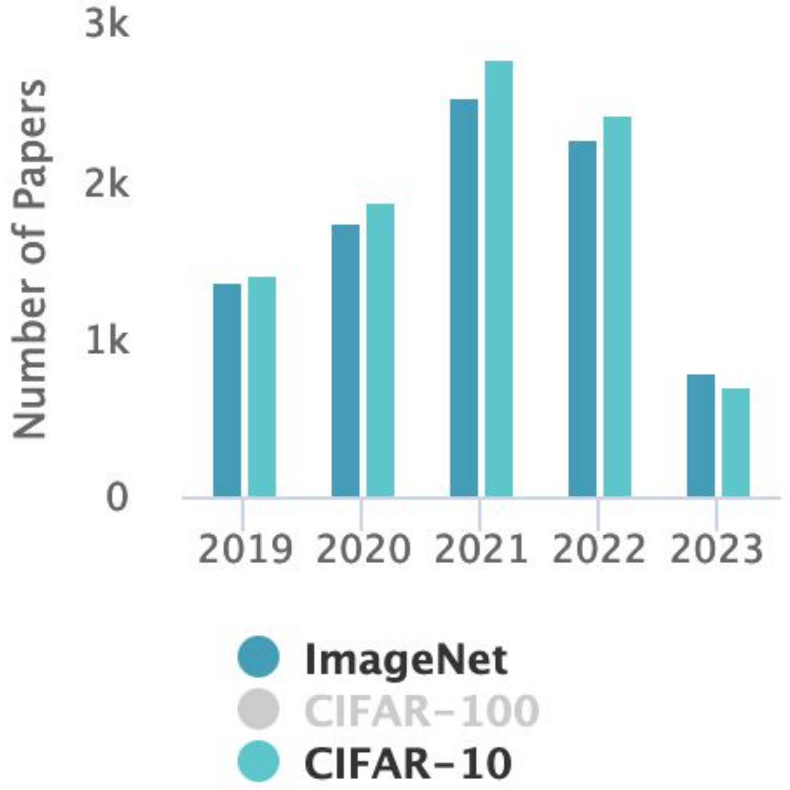
11,018 PAPERS • 69 BENCHMARKS

ImageNet

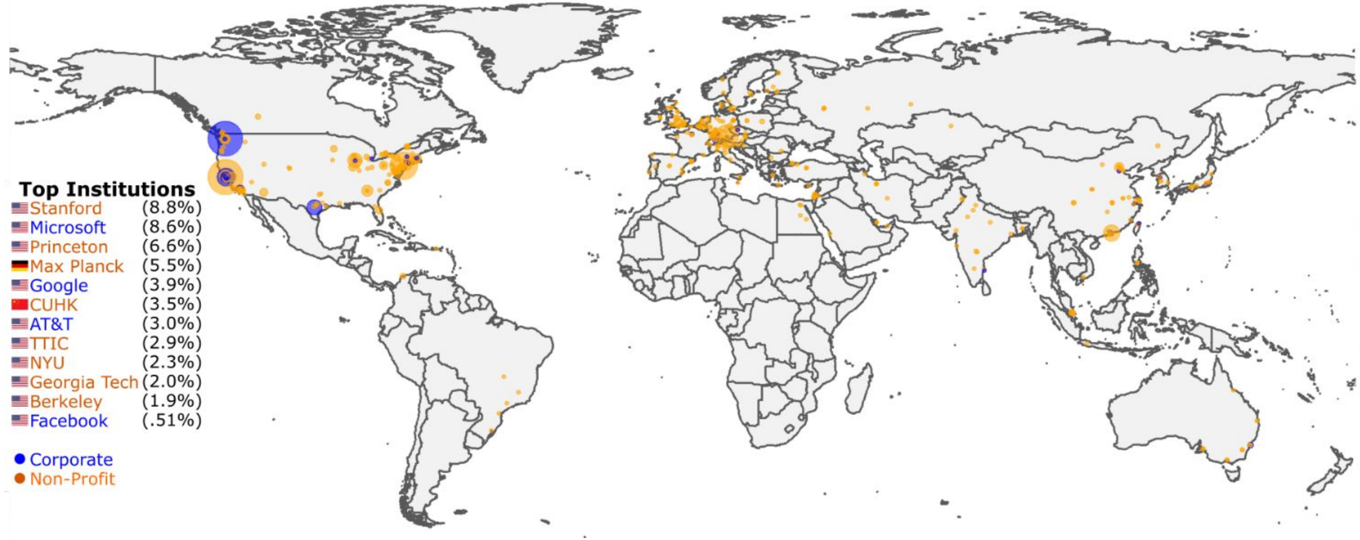
The ImageNet dataset contains 14,1
archy. Since 2010 the dataset is use

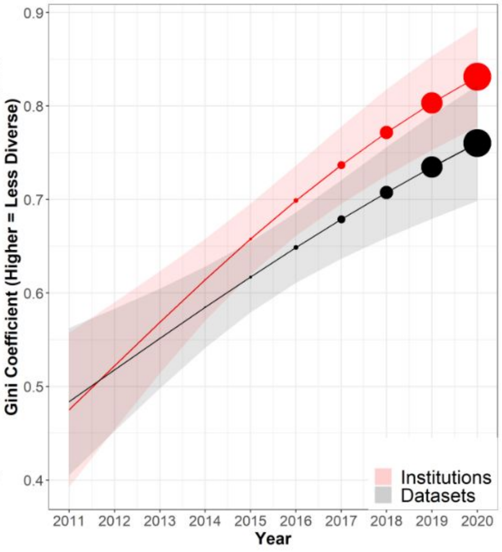
10,614 PAPERS • 102 BENCHMARKS

Usage



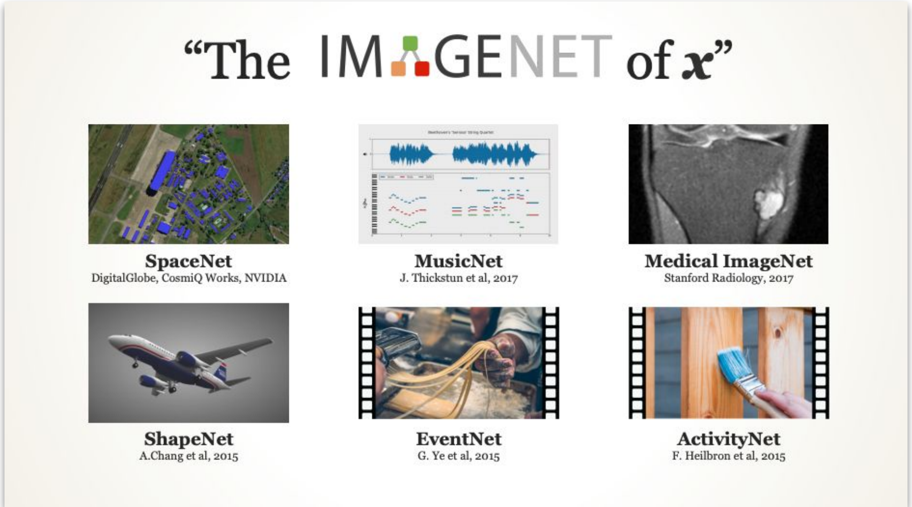
Concentration of Benchmark Creation





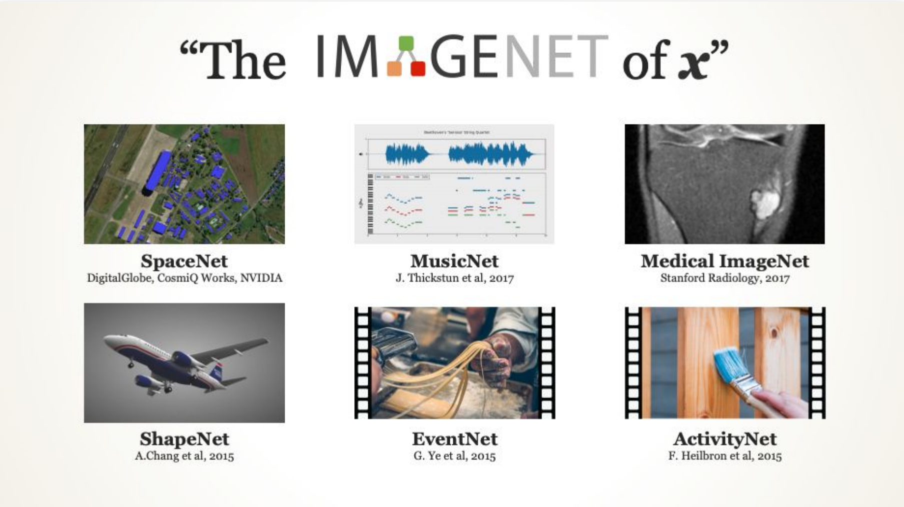
Koch et al. 2021. "Reduced, Reused and Recycled: The Life of a Dataset in Machine Learning Research." *NeurIPS (Data and Benchmark Track).*

“The Unreasonable Effectiveness of Data”



Denton, Hanna, Amironesei, et al. 2021. “On the Genealogy of Machine Learning Datasets: A Critical History of ImageNet.” *Big Data & Society.*

“The Unreasonable Effectiveness of Data”



The Steep Cost of Capture

Meredith Whittaker, New York University

Denton, Hanna, Amironesei, et al. 2021. “On the Genealogy of Machine Learning Datasets: A Critical History of ImageNet.” *Big Data & Society.*

Agenda

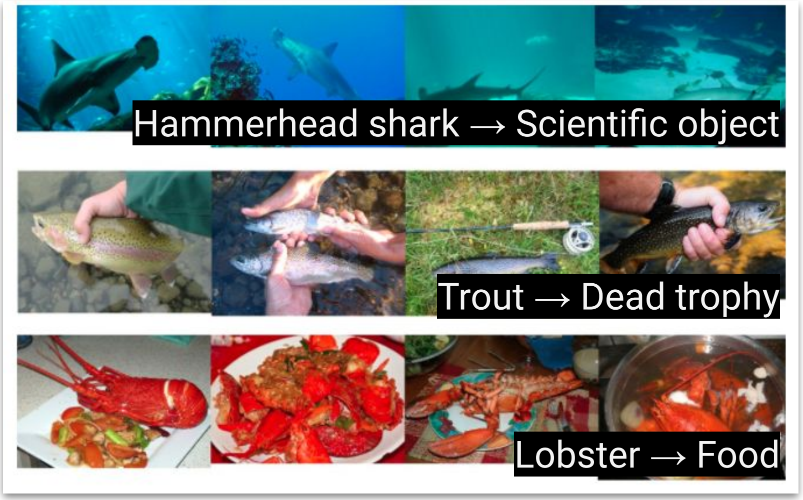
Ontologies Upon Ontologies

“Heroes Emerge Only in Times of Great Need!”

Being a Data Subject in Data-ful Times

Datasets as Infrastructure

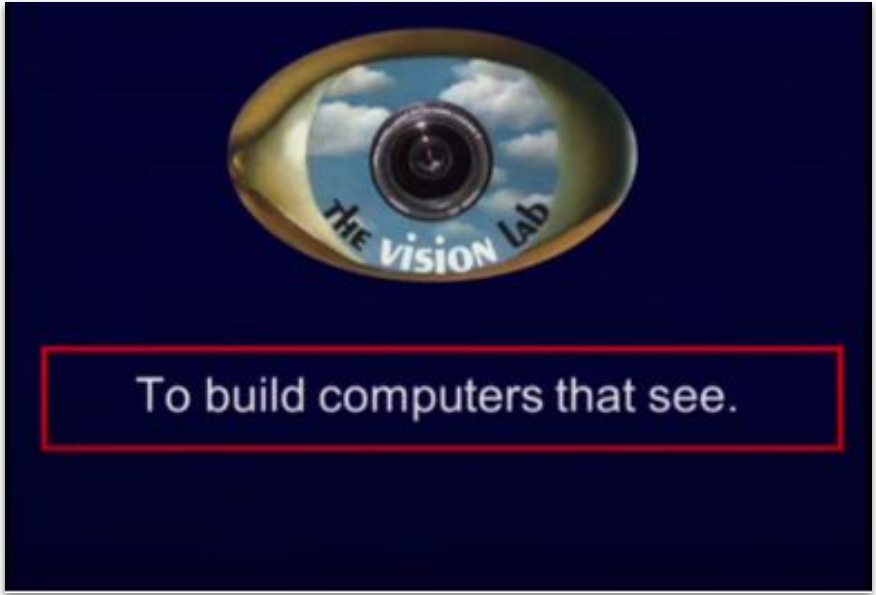
1) **Datasets determine what a model learns**
2) Datasets benchmark algorithms
3) Datasets serve as model organisms
4) Datasets provide methodological grounds for model development in industry contexts



Malevé. 2019. An Introduction to Image Datasets.

Denton and Hanna et al. 2020. "Bringing the People Back In: Contesting Benchmark Machine Learning Datasets." PAML Workshop, *ICML*.

Computational Construction of Meaning and Understanding



Denton, Hanna, Amironesei, et al. 2021. "On the Genealogy of Machine Learning Datasets: A Critical History of ImageNet." *Big Data & Society*.

What is WordNet?

Any opinions, findings, and conclusions or recommendations expressed in this material are those of the creators of WordNet and do not necessarily reflect the views of any funding agency or Princeton University.

When writing a paper or producing a software application, tool, or interface based on WordNet, it is necessary to properly **cite the source**. Citation figures are critical to WordNet funding.

About WordNet

WordNet® is a large lexical database of English. Nouns, verbs, adjectives and adverbs are grouped into sets of cognitive synonyms (synsets), each expressing a distinct concept. Synsets are interlinked by means of conceptual-semantic and lexical relations. The resulting network of meaningfully related words and concepts can be navigated with the **browser** . WordNet is also freely and publicly available for **download**. WordNet's structure makes it a useful tool for computational linguistics and natural language processing.

WordNet Search - 3.1

- WordNet home page - Glossary - Help

Word to search for: pembroke welsh corgi Search WordNet

Display Options: (Select option to change) Change

Key: "S:" = Show Synset (semantic) relations, "W:" = Show Word (lexical) relations

WordNet Search - 3.1
- WordNet home page - Glossary - Help

Word to search for: pembroke welsh corgi Search WordNet

Display Options: (Select option to change) Change

Key: "S:" = Show Synset (semantic) relations, "W:" = Show Word (lexical) relations

Noun

- S: (n) Pembroke, **Pembroke Welsh corgi**
  - ***direct hypernym*** / *inherited hypernym* / *sister term*
    - S: (n) corgi, Welsh corgi
      - S: (n) dog, domestic dog, Canis familiaris
        - S: (n) canine, canid
          - S: (n) carnivore
            - S: (n) placental, placental mammal, eutherian, eutherian mammal
              - S: (n) mammal, mammalian
                - S: (n) vertebrate, craniate
                  - S: (n) chordate
                    - S: (n) animal, animate being, beast, brute, creature, fauna
                      - S: (n) organism, being
                        - S: (n) living thing, animate thing
                          - S: (n) whole, unit
                            - S: (n) object, physical object

Where do you get the definitions for WordNet? (short answer)

> Our lexicographers write them.

Where do you get the definitions for WordNet? (long answer)

> From the foreword to **WordNet: An Electronic Lexical Database**, pp. xviii-xix:
>
> People sometimes ask, "Where did you get your words?" We began in 1985 with the words in Kučera and Francis's Standard Corpus of Present-Day Edited English (familiarly known as the Brown Corpus), principally because they provided frequencies for the different parts of speech. We were well launched into that list when Henry Kučera warned us that, although he and Francis owned the Brown Corpus, the syntactic tagging data had been sold to Houghton Mifflin. We therefore dropped our plan to use their frequency counts (in 1988 Richard Beckwith developed a polysemy index that we use instead). We also incorporated all the adjectives pairs that Charles Osgood had used to develop the semantic differential. And since synonyms were critically important to us, we looked words up in various thesauruses: for example, Laurence Urdang's little "Basic Book of Synonyms and Antonyms" (1978), Urdang's revision of Rodale's "The Synonym Finder" (1978), and Robert Chapman's 4th edition of "Roget's International Thesaurus" (1977) -- in such works, one word quickly leads on to others. Late in 1986 we received a list of words compiled by Fred Chang at the Naval Personnel Research and Development Center, which we compared with our own list; we were dismayed to find only 15% overlap.
>
> So Chang's list became input. And in 1993 we obtained the list of 39,143 words that Ralph Grishman and his colleagues at New York University included in their common lexicon, COMLEX; this time we were dismayed that WordNet contained only 74% of the COMLEX words. But that list, too, became input. In short, a variety of sources have contributed; we were not well disciplined in building our vocabulary. The fact is that the English lexicon is very large, and we were lucky that our sponsors were patient with us as we slowly crawled up the mountain.

Where do you get the definitions for WordNet? (short answer)

> Our lexicographers write them.

Where do you get the definitions for WordNet? (long answer)

> From the foreword to **WordNet: An Electronic Lexical Database**, pp. xviii-xix:
>
> People sometimes ask, "Where did you get your words?" We began in 1985 with the words in Kučera and Francis's Standard Corpus of Present-Day Edited English (familiarly known as the Brown Corpus), principally because they provided frequencies for the different parts of speech. We were well launched into that list when Henry Kučera warned us that, although he and Francis owned the Brown Corpus, the syntactic tagging data had been sold to Houghton Mifflin. We therefore dropped our plan to use their frequency counts (in 1988 Richard Beckwith developed a polysemy index that we use instead). We also incorporated all the adjectives pairs that Charles Osgood had used to develop the semantic differential. And since synonyms were critically important to us, we looked words up in various thesauruses: for example, Laurence Urdang's little "Basic Book of Synonyms and Antonyms" (1978), Urdang's revision of Rodale's "The Synonym Finder" (1978), and Robert Chapman's 4th edition of "Roget's International Thesaurus" (1977) -- in such works, one word quickly leads on to others. Late in 1986 we received a list of words compiled by Fred Chang at the Naval Personnel Research and Development Center, which we compared with our own list; we were dismayed to find only 15% overlap.
>
> So Chang's list became input. And in 1993 we obtained the list of 39,143 words that Ralph Grishman and his colleagues at New York University included in their common lexicon, COMLEX; this time we were dismayed that WordNet contained only 74% of the COMLEX words. But that list, too, became input. In short, a variety of sources have contributed; we were not well disciplined in building our vocabulary. The fact is that the English lexicon is very large, and we were lucky that our sponsors were patient with us as we slowly crawled up the mountain.

BROWN CORPUS MAUNAL

BROWN CORPUS MAUNAL

LIST OF SAMPLES

A01	Atlanta Constitution	Political Reportage
A02	Dallas Morning News	Political Reportage
	Chicago Tribune	Political Reportage
A03	Chicago Tribune	Political Reportage
A04	Christian Science Monitor	Political Reportage
A05	Providence Journal	Political Reportage
A06	Newark Evening News	Political Reportage
A07	New York Times	Political Reportage
A08	Times-Picayune, New Orleans	Political Reportage
A09	Philadelphia Inquirer	Political Reportage
	Chicago Tribune	Political Reportage
A10	Oregonian, Portland	Political Reportage
A11	Sun, Baltimore	Sports Reportage
A12	Dallas Morning News	Sports Reportage
A13	Rocky Mountain News	Sports Reportage
	Dallas Morning News	Sports Reportage.
A14	New York Times	Sports Reportage.
A15	St. Louis Post-Dispatch	Sports Reportage
A16	Chicago Tribune	Society Reportage
A17	Rocky Mountain News	Society Reportage
	Dallas Morning News	Society Reportage
A18	Philadelphia Inquirer	Society Reportage
	Times-Picayune, New Orleans	Society Reportage

BROWN CORPUS MAUNAL

LIST OF SAMPLES

A01	Atlanta Constitution	
A02	Dallas Morning News	
	Chicago Tribune	
A03	Chicago Tribune	
A04	Christian Science Monitor	
A05	Providence Journal	
A06	Newark Evening News	
A07	New York Times	
A08	Times-Picayune, New Orleans	
A09	Philadelphia Inquirer	
	Chicago Tribune	
A10	Oregonian, Portland	
A11	Sun, Baltimore	
A12	Dallas Morning News	Sports Reportage
A13	Rocky Mountain News	Sports Reportage
	Dallas Morning News	Sports Reportage.
A14	New York Times	Sports Reportage.
A15	St. Louis Post-Dispatch	Sports Reportage
A16	Chicago Tribune	Society Reportage
A17	Rocky Mountain News	Society Reportage
	Dallas Morning News	Society Reportage
A18	Philadelphia Inquirer	Society Reportage
	Times-Picayune, New Orleans	Society Reportage

D01	William Pollard	Physicist and Christian
D02	Schubert Ogden	Christ Without Myth
D03	Edward E. Kelly	Christian Unity in England
D04	Jaroslav Pelikan	The Shape of Death
D05	Perry Miller	Theodore Parker: Apostasy With in Liberalism
D06	A Howard Kelly	Out of Doubt into Faith
D06B	Shirley Schuyler	Not as the World Giveth
D06C	Nathanael Olson	Are You in Orbit?
D07	Peter Eldersveld	Faith Amid Fear
D08	Schuyler Cammann	The Magic Square of Three
D09	Eugene E. Golay	Organizing the Local Church
D10	Huston Smith	Interfaith Communication: The Contemporary Scene
D11	Paul Ramsey	War & the Christian Conscience

BROWN CORPUS MAUNAL

LIST OF SAMPLES

A01	Atlanta Constitution
A02	Dallas Morning News
	Chicago Tribune
A03	Chicago Tribune
A04	Christian Science Monitor
A05	Providence Journal
A06	Newark Evening News
A07	New York Times
A08	Times-Picayune, New Orleans
A09	Philadelphia Inquirer
	Chicago Tribune
A10	Oregonian, Portland
A11	Sun, Baltimore
A12	Dallas Morning News
A13	Rocky Mountain News
	Dallas Morning News
A14	New York Times
A15	St. Louis Post-Dispatch
A16	Chicago Tribune
A17	Rocky Mountain News
	Dallas Morning News
A18	Philadelphia Inquirer
	Times-Picayune, New Orleans

D01	William Pollard	Physicist and Christian
D02	Schubert Ogden	Christ Without Myth
D03	Edward E. Kelly	Christian Unity in England
D04	Jaroslav Pelikan	The Shape of Death
D05	Perry Miller	Theodore Parker: Apostasy With in Liberalism
D06	A Howard Kelly	Out of Doubt into Faith
D06B	Shirley Schuyler	Not as the World Giveth
D06C	Nathanael Olson	Are You in Orbit?
D07	Peter E	
D08	Schuyl	
D09	Eugene	
D10	Huston	
D11	Paul R	

F01	Rosemary Blackmon	How Much Do You Tell When You Talk?
F02	Glenn Infield	America's Secret Poison Gas Tragedy
F03	Nathan Rapport	I've Been Here Before
F04	Ruth F. Rosevear	North Country School Cares for the Whole Child
F05	Richard S. Allen	When Fogg Flew the Mail
F06A	Alice Ho Austin	Let's Discuss Retirement
F06B	Harold P. Winchester	What It Means to be Creative
F07A	Marvin Sentnor and Stephen Hult	How to Have a Successful Honeymoon
F07B	Ho Walter Yoder	Attitudes Toward Nudity
F08	Philip Reaves	Who Rules the Marriage Bed?
F09A	David Martinson	Fantastic Life & Death of the Golden Prostitute.
F09B	Isel D. Rugget	When It Comes to Carpets
F10	Jack Kaplan	Therapy by Witchcraft
F11	Lillian Pompian	Tooth-Straightening Today
F12	Marian Neater	New Methods of Parapsychology.
F13	Orlin J. Scoville	Part-time Farming
F14	Harold Rosenberg	The Trial and Eichmann
F15	John A. O'Brien	Let's Take Birth Control Out of Politics

src

brown.py

brown.py > ...

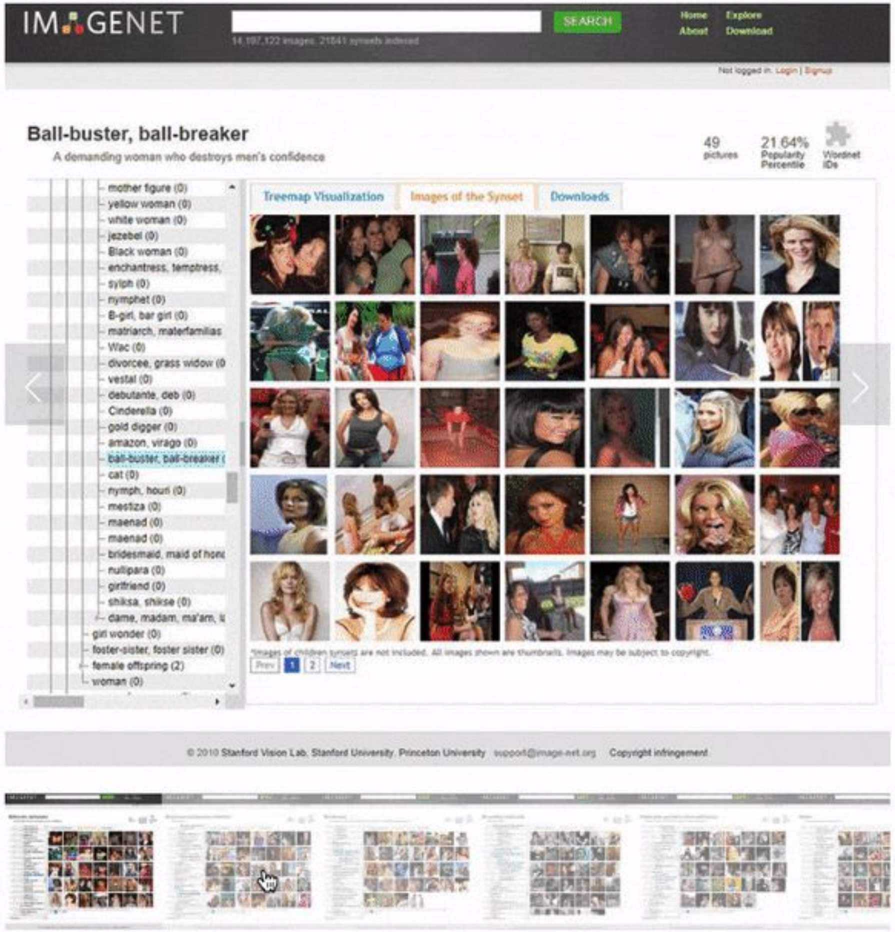
```python
import nltk
from nltk.corpus import brown

## Who Rules the Marriage Bed?
[' '.join(x) for x in brown.sents(fileids=['cf08'])]

```

OUTPUT DEBUG CONSOLE PROBLEMS TERMINAL python3.8

```
In [44]: [' '.join(x) for x in brown.sents(fileids=['cf08'])]
    ...:
Out[44]:
['In tradition and in poetry , the marriage bed is a place of unity and harmony .',
 'The partners each bring to it unselfish love , and each takes away an equal share of pleasure and joy .',
 'At its most ecstatic moments , husband and wife are elevated far above worldly cares .',
 'Everything else is closed away .',
 'This is the ideal .',
 'But marriage experts say that such mutual contribution and mutual joy are seldom achieved .',
 'Instead one partner or the other dominates the sexual relationship .',
 'In the past , it has been the husband who has been dominant and the wife passive .',
 'But today there are signs that these roles are being reversed .',
 'In a growing number of American homes , marriage counselors report , the wife is taking a commanding role
 in sexual relationships .',
 'It is she who decides the time , the place , the surroundings , and the frequency of the sexual act .',
 'It is she who says aye or nay to the intimate questions of sexual technique and mechanics -- not the husb
```



Selections from the "Person" classes, ImageNet

excavating.ai
(Crawford and Paglen, 2020)

WordNet Search - 3.1

- [WordNet home page] - [Glossary] - [Help]

Word to search for: closet queen [Search WordNet]

Display Options: (Select option to change) [Change]

Key: "S:" = Show Synset (semantic) relations, "W:" = Show Word (lexical) relations

Display options for sense: (gloss)

Noun

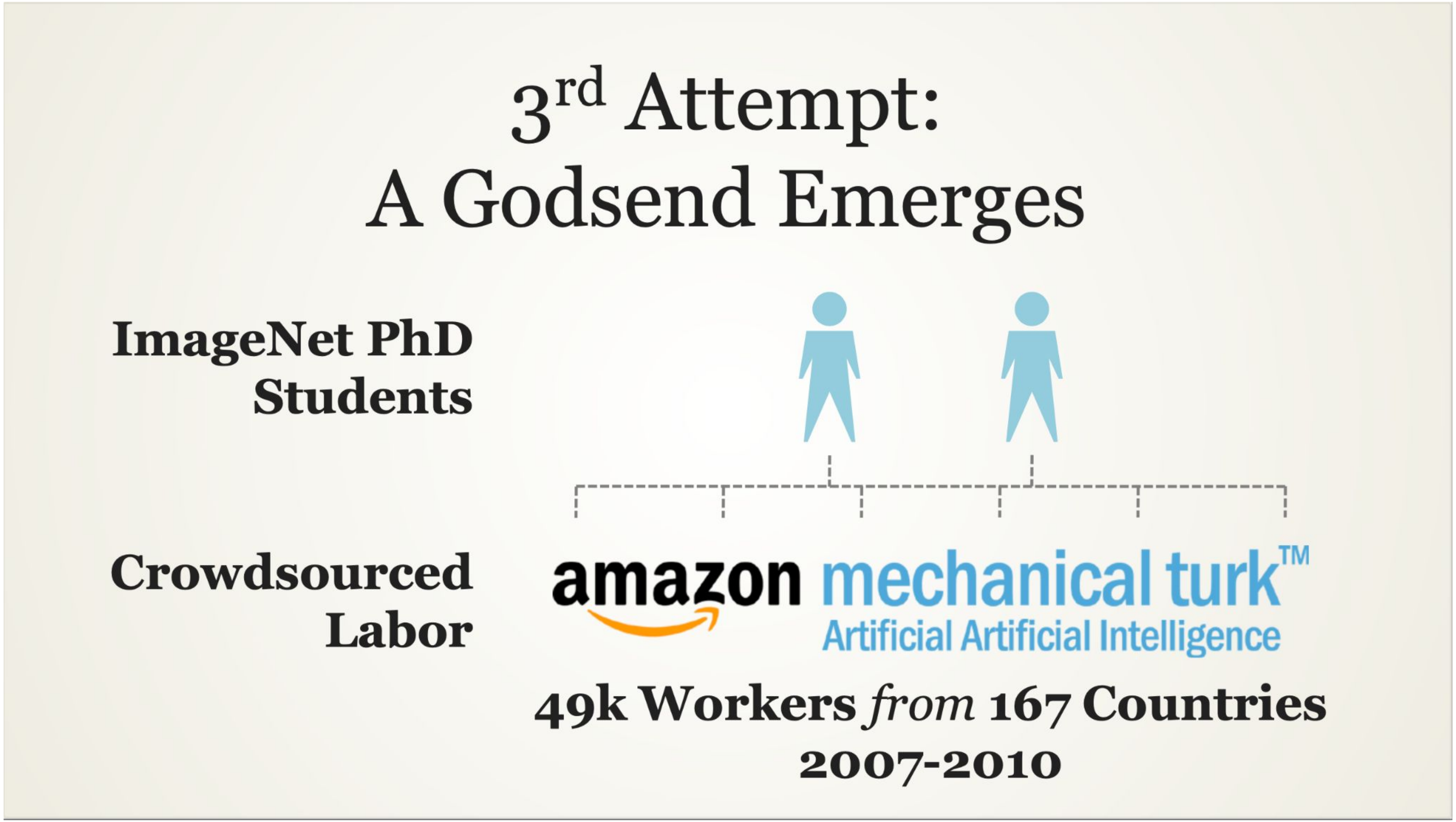
- S: (n) **closet queen** (a negative term for a homosexual man who chooses not to reveal his sexual orientation)
  - ***direct hypernym*** / *inherited hypernym* / *sister term*
    - S: (n) homosexual, homophile, homo, gay (someone who is sexually attracted to persons of the same sex)
      - S: (n) person, individual, someone, somebody, mortal, soul (a human being)
        - S: (n) organism, being (a living thing that has (or can develop) the ability to act or function independently)
          - S: (n) living thing, animate thing (a living (or once living) entity)
            - S: (n) whole, unit (an assemblage of parts that is regarded as a single entity)
              - S: (n) object, physical object (a tangible and visible entity; an entity that can cast a shadow)
                - S: (n) physical entity (an entity that has physical existence)
                  - S: (n) entity (that which is perceived or known or inferred to have its own distinct existence (living or nonliving))
        - S: (n) causal agent, cause, causal agency (any entity that produces an effect or is responsible for events or results)
          - S: (n) physical entity (an entity that has physical existence)
            - S: (n) entity (that which is perceived or known or inferred to have its own distinct existence (living or nonliving))

Agenda

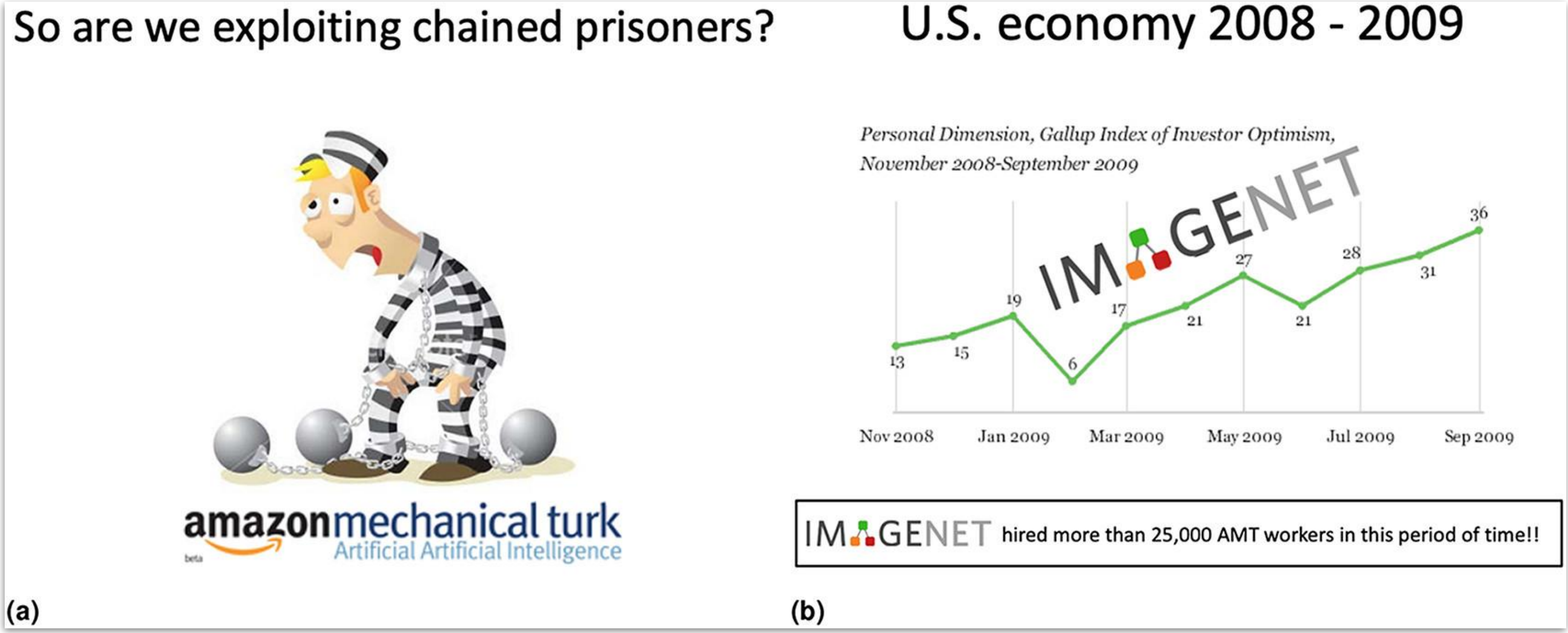
Ontologies Upon Ontologies

“Heroes Emerge Only in Times of Great Need!”

Being a Data Subject in Data-ful Times



Fei-Fei L 2017. Imagenet: Where have we gone? Where are we going?



Denton, Hanna, Amironesei, et al. 2021. “On the Genealogy of Machine Learning Datasets: A Critical History of ImageNet.” *Big Data & Society.*

Mimi Ọnụọha. 2019. “The Future Is Here!”

Turkopticon

REQUESTER LIST REVIEWS ABOUT RULES FAQ BLOG REVIEW DONATE

Tweets by @turkopticon

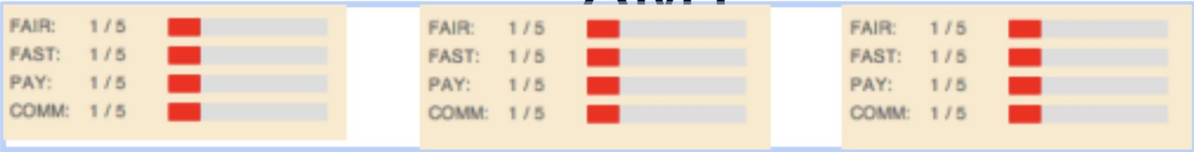
Our mission is to organize mutual aid, resources, and advocacy to improve conditions for all people using Amazon's Mechanical Turk (AMT) platform while striving to make this work a good job for all.

Turkopticon was founded by Lilly Irani and Six Silberman as a review website to provide AMT workers with a space to share information about bad requesters and tasks. This forum also serves as a safety net. Workers are still coming to the website ten years later to check requesters' reviews before they accept tasks. **BUT, WAIT! THERE IS MORE!**

In 2019 it was decided that Turkopticon could be more than a review site. Platform workers came together to form a team with Lilly and Six. Graduate students and tech workers stepped forward to help with software needs and more. **We are so excited to be able to say that we are now a worker-led non-profit organization!**

ImageNet AMT A3C8DPCYBZH618 Averages » Requester on AMT » Review Requester » Contact the requester »	FAIR: 4 / 5 FAST: 3 / 5 COMM: NO DATA PAY: 1 / 5	*Play a game of identifying birds and get paid! and *Look at car images and identify cars. - Both of them awful underpaid. I think this requester is mocking us. I did just two hits but no way to continue wasting so much time for him. Oct 17 2012 | [redacted] |
ImageNet AMT A3C8DPCYBZH618 Averages » Requester on AMT » Review Requester » Contact the requester »	FAIR: 1 / 5 FAST: 1 / 5 COMM: 1 / 5 PAY: 1 / 5	One of the all time worst requesters on AMT has changed their requester ID to run from all the negative reviews. DO NOT BE FOOLED THIS IS THE SAME ImageNet AMT. See the other profile for real reviews. http://turkopticon.differenceengines.com /ADLB46N5J8NCP Oct 16 2012 | [redacted] | Aha, now I understand! That's it, "Old habits die hard" ! Oct 18 2012 | [redacted]
ImageNet AMT A3C8DPCYBZH618 Averages » Requester on AMT » Review Requester » Contact the requester »	FAIR: NO DATA FAST: NO DATA COMM: NO DATA PAY: 1 / 5	Wow... they expect you to do 6 pages of tedious work identifying specific years & models of cars for $0.05! I started working on one & returned it after I had already wasted 10 mins on their qualification test that is included with your first HIT. Sep 28 2012 | [redacted] |

We got terrible reviews on AMT



“apparently the virtually indistinguishable cars are super distinguishable to them”

“I paid very close attention to the pictures and let the requester know that I feel these rejects were unfair. Will update when I get communication back.”

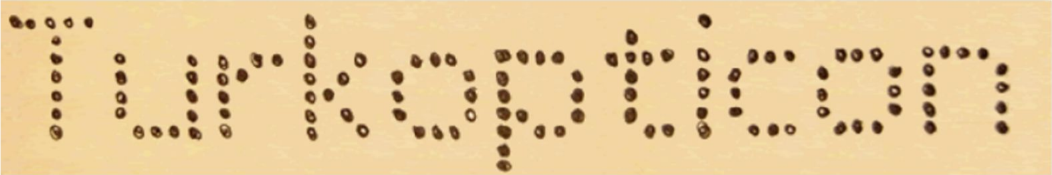
Gebru et al. 2017. “Predicting Demographics Using 50 Million Images.” Dissertation defense.

Mass rejections

admin   December 21, 2020   1 Comment

First, allow me to introduce myself. I'm Sherry, lead organizer for Turkopticon. I have been a Turker for going on 6 years. You may ask why did I finally decide to step up and work to make Turk better? Two words: MASS REJECTION. My personal story involves having a great history of finding good new requesters and having plenty of wiggle room in my approval rating so I decided to work for a new requester in the same hopes of finding good work. That is when my Mturk experience drastically changed and I knew I had to stand up.

Amazon: Stop the Mass Rejections!



Amazon Mechanical Turk Workers have been organizing for years to prevent mass-rejections on the platform. These mass rejections are unfair, hurt workers, and must end. Turkopticon has organized a mass rejection protection platform that **must** be adopted to protect workers!

Agenda

Ontologies Upon Ontologies

“Heroes Emerge Only in Times of Great Need!”

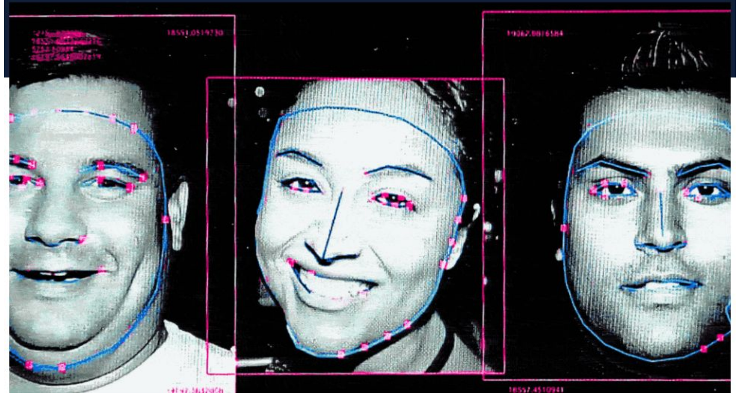
Being a Data Subject in Data-ful Times

Smith, 2019, IBM Research Releases 'Diversity in Faces' Dataset to Advance Study of Fairness in Facial Recognition Systems

IBM Research Releases ‘Diversity in Faces’ Dataset to Advance Study of Fairness in Facial Recognition Systems

Facial recognition's 'dirty little secret': Millions of online photos scraped without consent

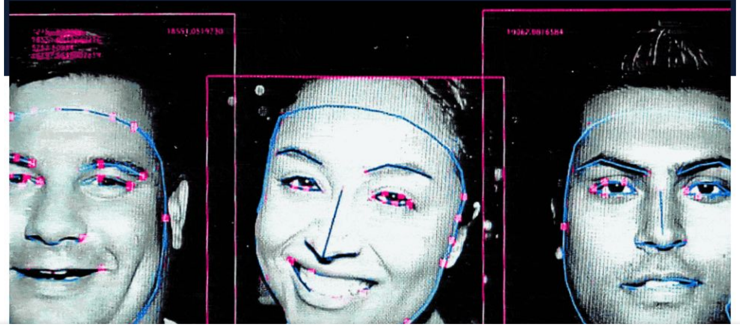
People's faces are being used without their permission, in order to power technology that could eventually be used to surveil them, legal experts say.



IBM Research Releases 'Diversity in Faces' Dataset to Advance Study of Fairness in Facial Recognition Systems

Facial recognition's 'dirty little secret': Millions of online photos scraped without consent

People's faces are being used without their permission, in order to power technology that could eventually be used to surveil them, legal experts say.



How Photos of Your Kids Are Powering Surveillance Technology

Millions of Flickr images were sucked into a database called MegaFace. Now some of those faces may have the ability to sue.

By Kashmir Hill and Aaron Krolik

IN THE UNITED STATES DISTRICT COURT
FOR THE NORTHERN DISTRICT OF ILLINOIS
EASTERN DIVISION

STEVEN VANCE, for himself and others similarly situated,	)	
Plaintiff,	)	Case No. _________________
v.	)	
	)	**CLASS ACTION COMPLAINT**
INTERNATIONAL BUSINESS MACHINES CORPORATION,	)	**JURY TRIAL DEMANDED**
Defendant.	)	**INJUNCTIVE RELIEF DEMANDED**

CLASS ACTION COMPLAINT

Plaintiff STEVEN VANCE, on behalf of himself and all other similarly situated individuals ("Plaintiff"), by and through his attorneys, brings this Class Action Complaint against Defendant INTERNATIONAL BUSINESS MACHINES CORPORATION ("IBM") and alleges the following:

IN THE UNITED STATES DISTRICT COURT
FOR THE NORTHERN DISTRICT OF ILLINOIS
EASTERN DIVISION

STEVEN VANCE, for himself and others)

for third parties to connect the individual whose biometrics were collected to other photos in which they appeared and other individuals appearing in those photos with them, subjecting them to increased surveillance, stalking, identity theft, and other invasions of privacy and fraud.

CLASS ACTION COMPLAINT

Plaintiff STEVEN VANCE, on behalf of himself and all other similarly situated individuals ("Plaintiff"), by and through his attorneys, brings this Class Action Complaint against Defendant INTERNATIONAL BUSINESS MACHINES CORPORATION ("IBM") and alleges the following:

[Excavating “Excavating AI”: The Elephant in the Gallery](#) (Lyons, 2020)

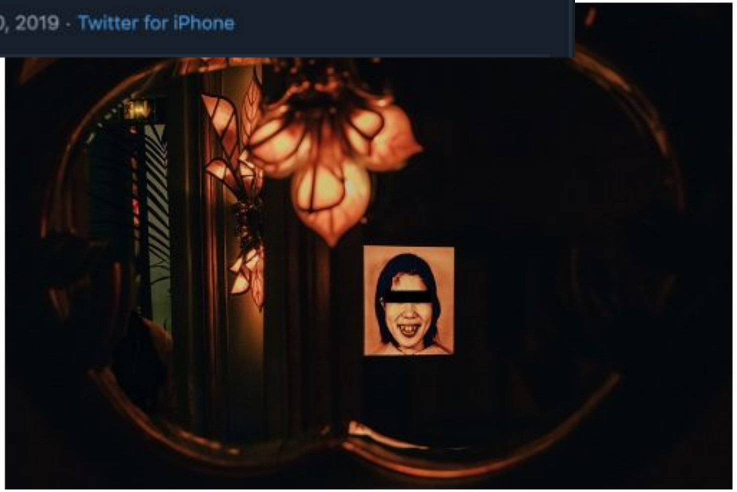


Figure 2: “Making Faces” exhibition at Maxim’s, Prada Mode Paris.

[On Lacework: watching an entire machine-learning dataset]
(Pipkin, 2020)

Agenda

Ontologies Upon Ontologies

“Heroes Emerge Only in Times of Great Need!”

Being a Data Subject in Data-ful Times

The cultural labor of AI data maintenance

Thank You

Alex Hanna

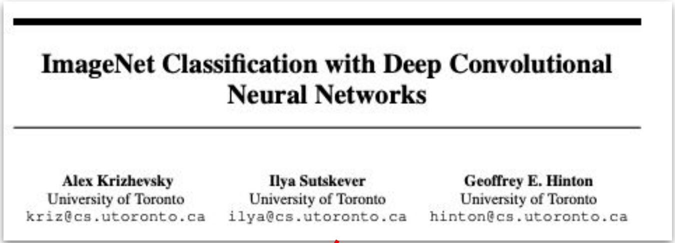
alex@dair-institute.org

@alexhanna

@alex@dair-community.social

Datasets as Infrastructure

1) Datasets determine what a model learns
2) **Datasets benchmark algorithms**
3) Datasets serve as model organisms
4) Datasets provide methodological grounds for model development in industry contexts

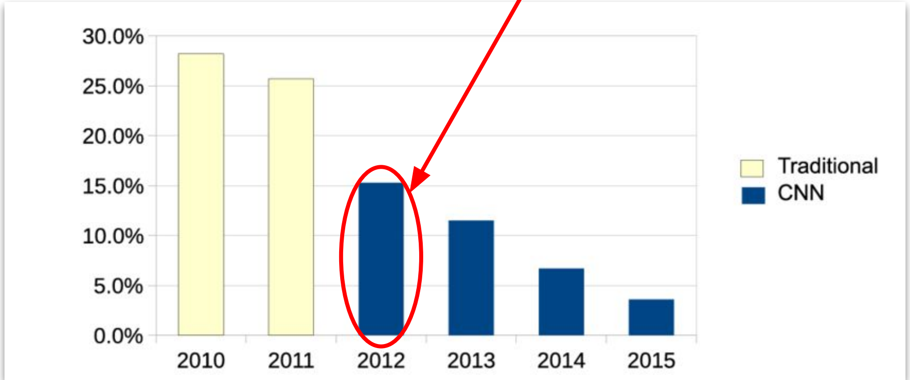


ImageNet Classification with Deep Convolutional Neural Networks

Alex Krizhevsky
University of Toronto
kriz@cs.utoronto.ca

Ilya Sutskever
University of Toronto
ilya@cs.utoronto.ca

Geoffrey E. Hinton
University of Toronto
hinton@cs.utoronto.ca



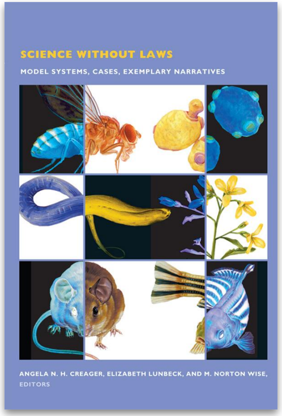
Denton and Hanna et al. 2020. "[Bringing the People Back In: Contesting Benchmark Machine Learning Datasets]." PAML Workshop, *ICML*.

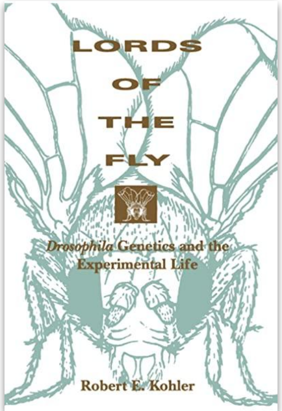
Datasets as Infrastructure

1) Datasets determine what a model learns
2) Datasets benchmark algorithms
3) **Datasets work as model organisms**
4) Datasets provide methodological grounds for model development in industry contexts

```
# Load ImageNet data
import tensorflow_datasets as tfds
ds = tfs.load('imagenet2012', split = 'train')

# Load neural network with ResNet50 architecture trained on ImageNet data
from tensorflow.keras.applications.resnet50 import ResNet50
model = ResNet50(weights='imagenet')
```

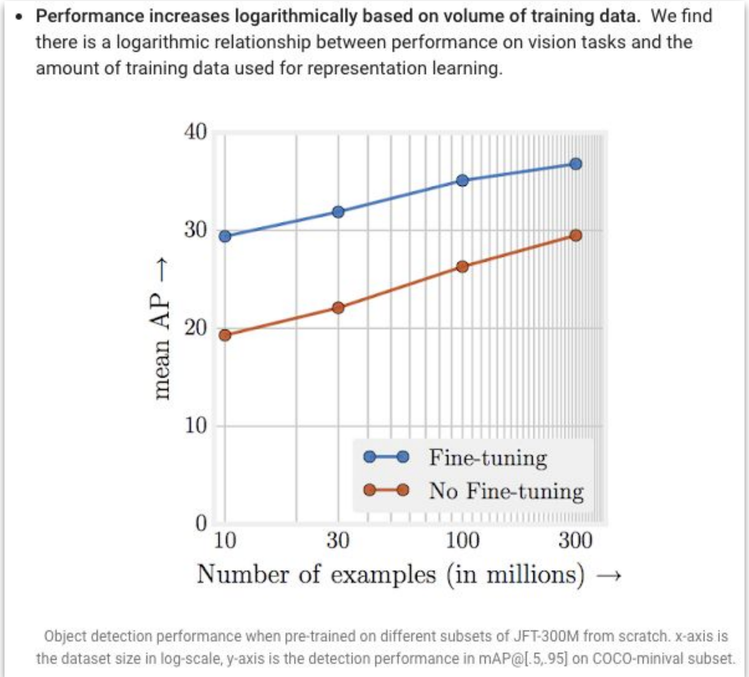




Denton and Hanna et al. 2020. "[Bringing the People Back In: Contesting Benchmark Machine Learning Datasets]." PAML Workshop, *ICML*.

Datasets as Infrastructure

1) Datasets determine what a model learns
2) Datasets benchmark algorithms
3) Datasets work as model organisms
4) **Datasets provide methodological grounds for model development in industry contexts**



Sun et al. 2017. [Revisiting the Unreasonable Effectiveness of Data](#).

Denton and Hanna et al. 2020. "[Bringing the People Back In: Contesting Benchmark Machine Learning Datasets](#)." PAML Workshop, *ICML*.