

datasets, benchmarks, and competitive testing

Ben Recht

Generalization in Machine Learning

Given: i.i.d. sample $S = \{z_1, \dots, z_n\}$ from dist D

Goal: Find a good predictor function f

$$R[f] = \mathbb{E}_z \text{loss}(f; z)$$

Population risk
(test error)

unknown!

$$R_S[f] = \frac{1}{n} \sum_{i=1}^n \text{loss}(f; z_i)$$

Empirical risk
(training error)

Minimize using SGD!

Generalization error: $R[f] - R_S[f]$

How much empirical risk underestimates population risk

We can optimize R_S ...

When is it a good proxy for R ?

Fundamental Theorem of Machine Learning

$$R[\hat{f}] = (R[\hat{f}] - R_S[\hat{f}]) + R_S[\hat{f}]$$

population
risk

generalization
error

training
error

- small training error implies risk $\approx$ generalization error
- zero training error *does not imply* overfitting

Holdout method

Given: i.i.d. sample $T = \{z_1, \dots, z_n\}$ and $H = \{z'_1, \dots, z'_m\}$ from dist D

Goal: Find a good predictor function f

$$R_T[\hat{f}] = \frac{1}{n} \sum_{i=1}^n \text{loss}(f; z_i)$$

(training error)

Minimize using SGD!

$$R_H[\hat{f}] = \frac{1}{m} \sum_{j=1}^m \text{loss}(f; z'_j)$$

(test/holdout error)

Look judiciously!

Look at k models:

$$\text{Generalization error: } R[\hat{f}] - R_H[\hat{f}] \leq C \sqrt{\frac{\log(k)}{m}}$$

Holdout method (caveats)

Given: i.i.d. sample $T = \{z_1, \dots, z_n\}$ and $H = \{z'_1, \dots, z'_m\}$ from dist D

$$R_T[\hat{f}] = \frac{1}{n} \sum_{i=1}^n \text{loss}(\hat{f}; z_i)$$

(training error)

Minimize using SGD!

$$R_H[\hat{f}] = \frac{1}{m} \sum_{j=1}^m \text{loss}(\hat{f}; z'_j)$$

(test/holdout error)

Look judiciously!

Look at k models:

$$R[\hat{f}] - R_H[\hat{f}] \leq C \sqrt{\frac{\log(k)}{m}}$$

- Unbiased test set assumption: “Training set” and “test set” are selected randomly from the same distribution
- “Iron vault” assumption: “Secret” hold out set (no adaptive fitting or cheat fitting)
- Statistical power assumption: Size of hold out set is sufficient for “test set error” to estimate “population error”

Machine Learning's Open Dirty Secret

Given: i.i.d. sample $T = \{z_1, \dots, z_n\}$ and $H = \{z'_1, \dots, z'_m\}$ from dist D

Goal: Find a good predictor function f

Myth

$$R_T[\hat{f}] = \frac{1}{n} \sum_{i=1}^n \text{loss}(f; z_i)$$

(training error)

Minimize using SGD!

$$R_H[\hat{f}] = \frac{1}{m} \sum_{j=1}^m \text{loss}(f; z'_j)$$

(test/holdout error)

Only look once!!!

Reality

$$\begin{array}{ll} \text{minimize} & R_H[\hat{f}] \\ \text{subject to} & R_T[\hat{f}] \leq \epsilon \\ & f \in \mathcal{F} \end{array}$$

$$\mathcal{F} = \left\{ \begin{array}{l} \text{functions computable} \\ \text{before the heat death} \\ \text{of the universe} \end{array} \right\}$$

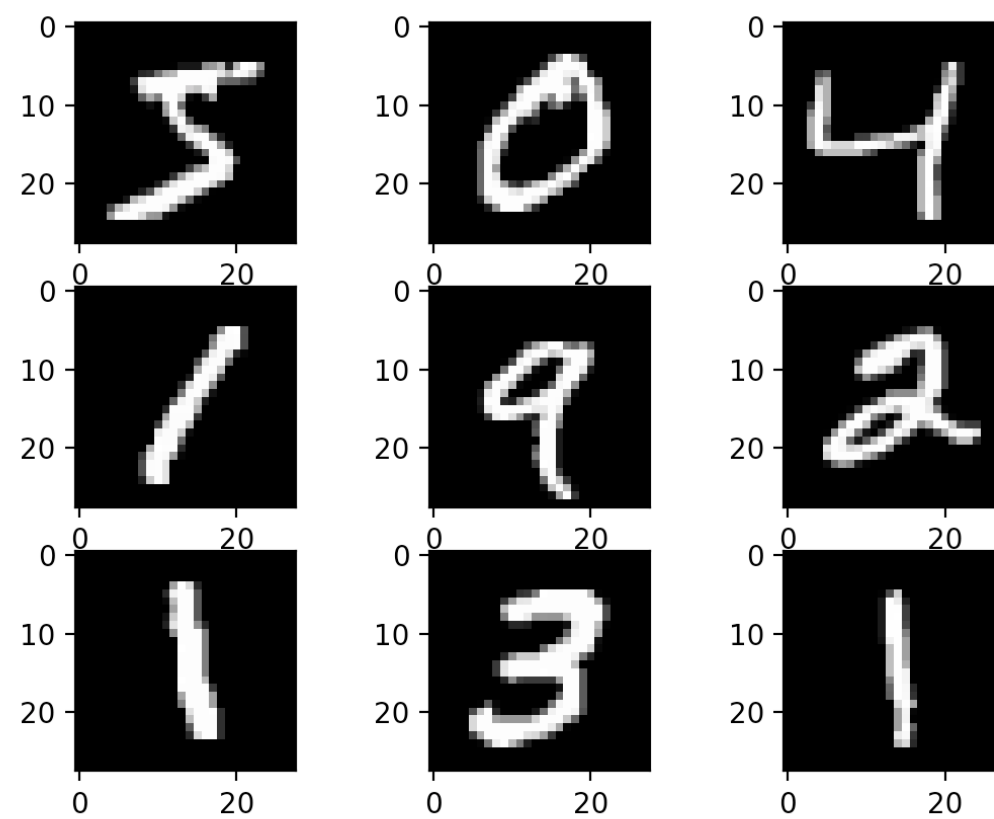
We never update the train and test sets.

List of datasets for machine-learning research

From Wikipedia, the free encyclopedia

These [datasets](#) are applied for [machine-learning](#) research and have been cited in [peer-reviewed](#) academic journals. Datasets are an integral part of the field of machine learning. Major advances in this field can result from advances in learning [algorithms](#) (such as [deep learning](#)), computer hardware,

- Earliest handwriting data set is 1991
- Only dataset before 1986 is IRIS (1936)



Frictionless Reproducibility

- Shared Data
- Re-executability
- *Competitive Testing*

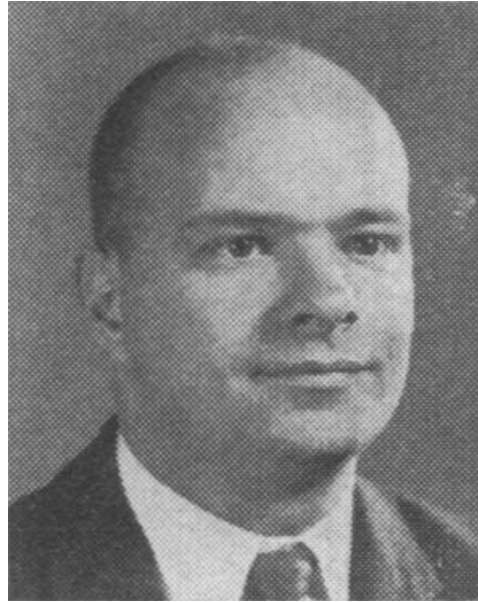
Why is it ok to train on the test set?

What can we learn from this practice?

Data Science at the Singularity

by David Donoho

Published on Jan 29, 2024



April 4, 1961

W. H. HIGHLEYMAN
CHARACTER RECOGNITION SYSTEM

2,978,675

Filed Dec. 10, 1959

4 Sheets-Sheet 1

FIG. 1

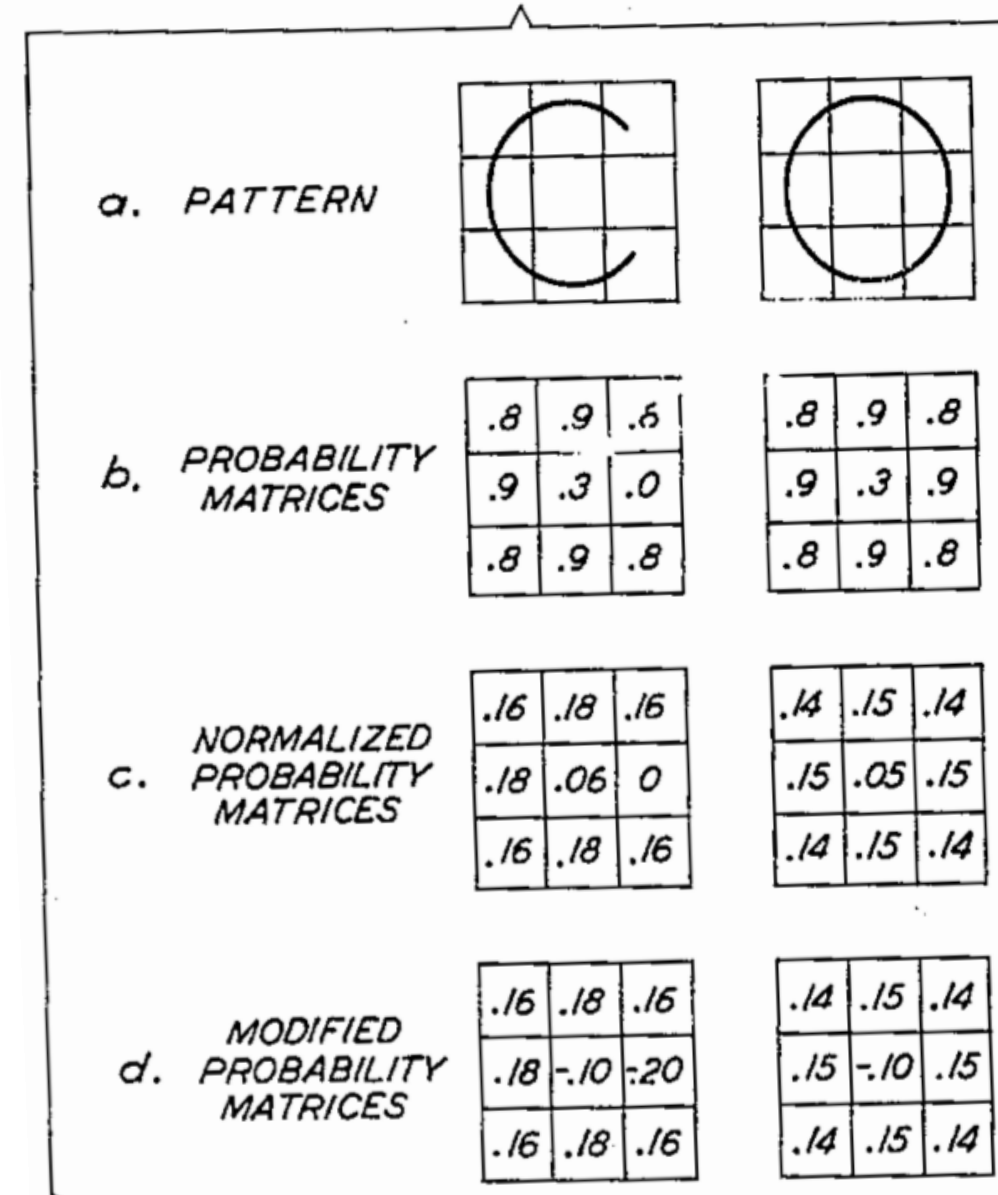


FIG. 5

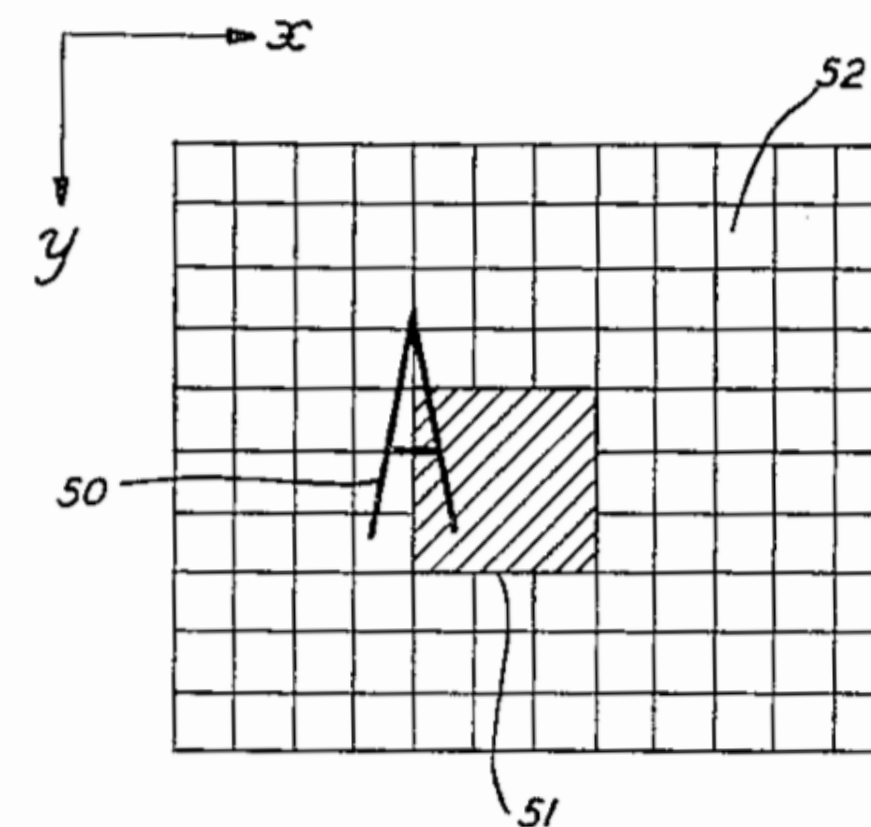
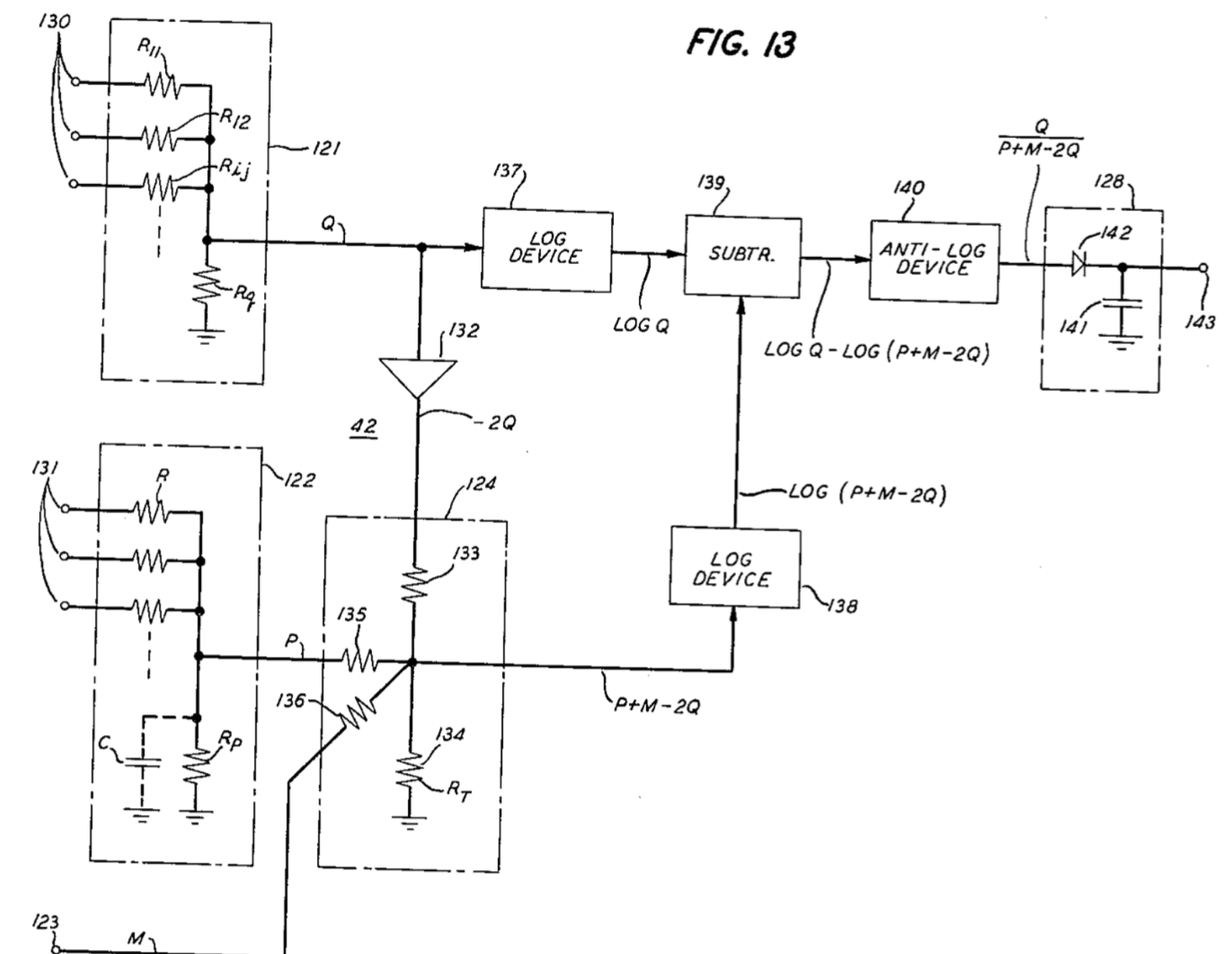
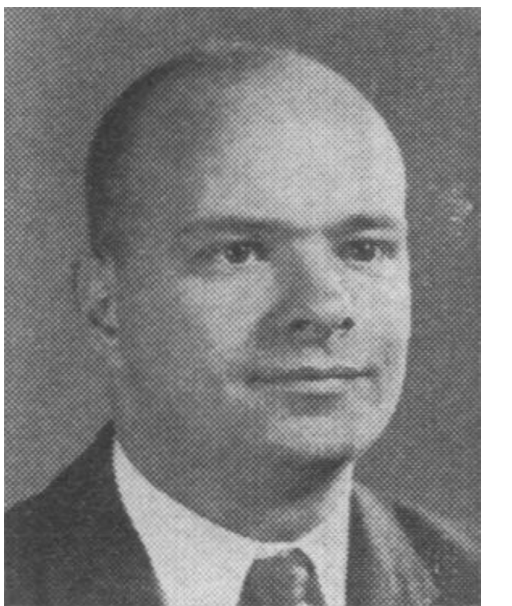


FIG. 13

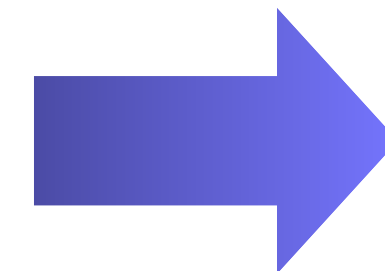
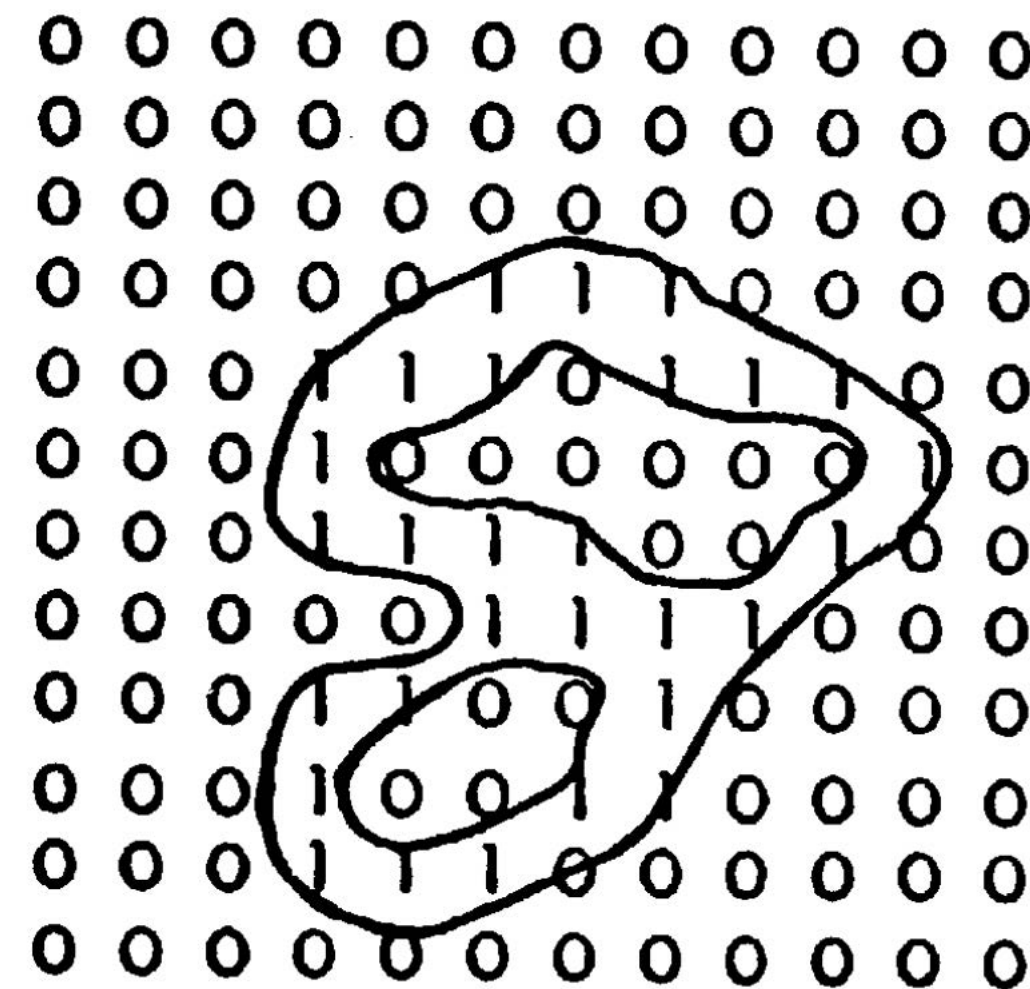
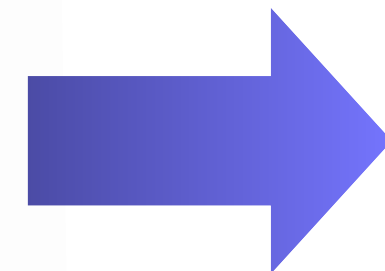
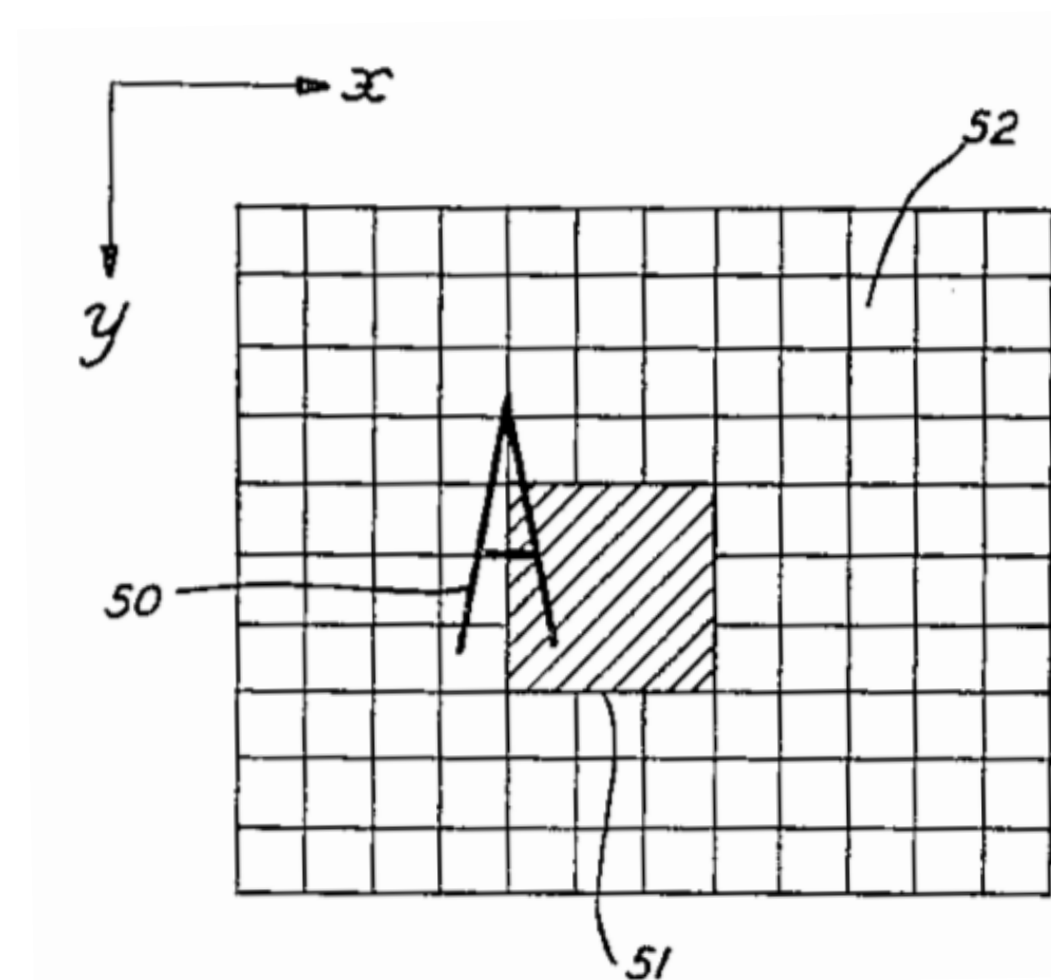


Convolutions have been with us since the beginning!
(And are discussed in part 2 of Duda and Hart)

Bill's Wild Idea



- Building hardware in analog that computes an end-to-end algorithm is hard.
- Why not build a *simulator*?



1 2 3 4 5 6 7 8 9 0

WHH

1 2 3 4 5 6 7 8 9 0

LHm

1 2 3 4 5 6 7 8 9 0

ERD

1 2 3 4 5 6 7 8 9 0

SD

1 2 3 4 5 6 7 8 9 0

NLS

1 2 3 4 5 6 7 8 9 0

BMT

1 2 3 4 5 6 7 8 9 0

RMR

1 2 3 4 5 6 7 8 9 0

STH

1 2 3 4 5 6 7 8 9 0

JRD

1 2 3 4 5 6 7 8 9 0

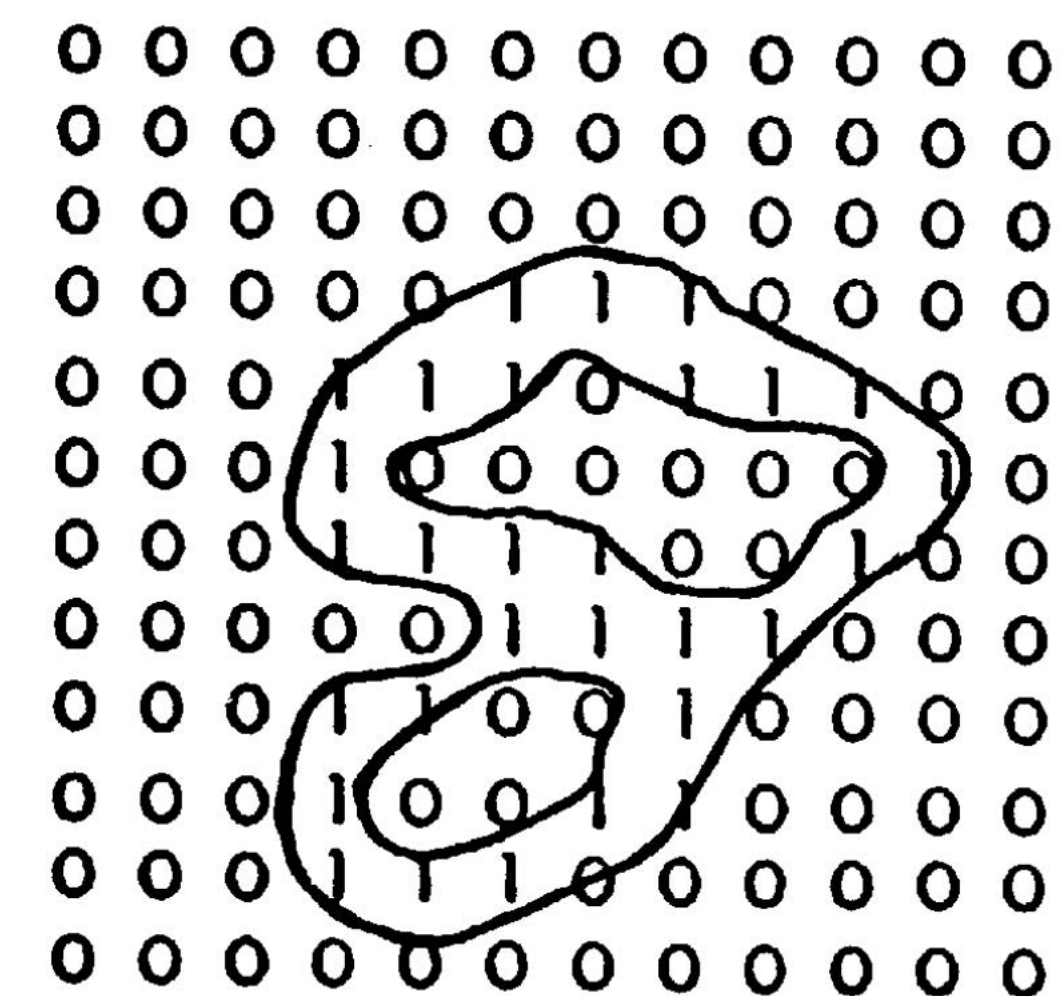
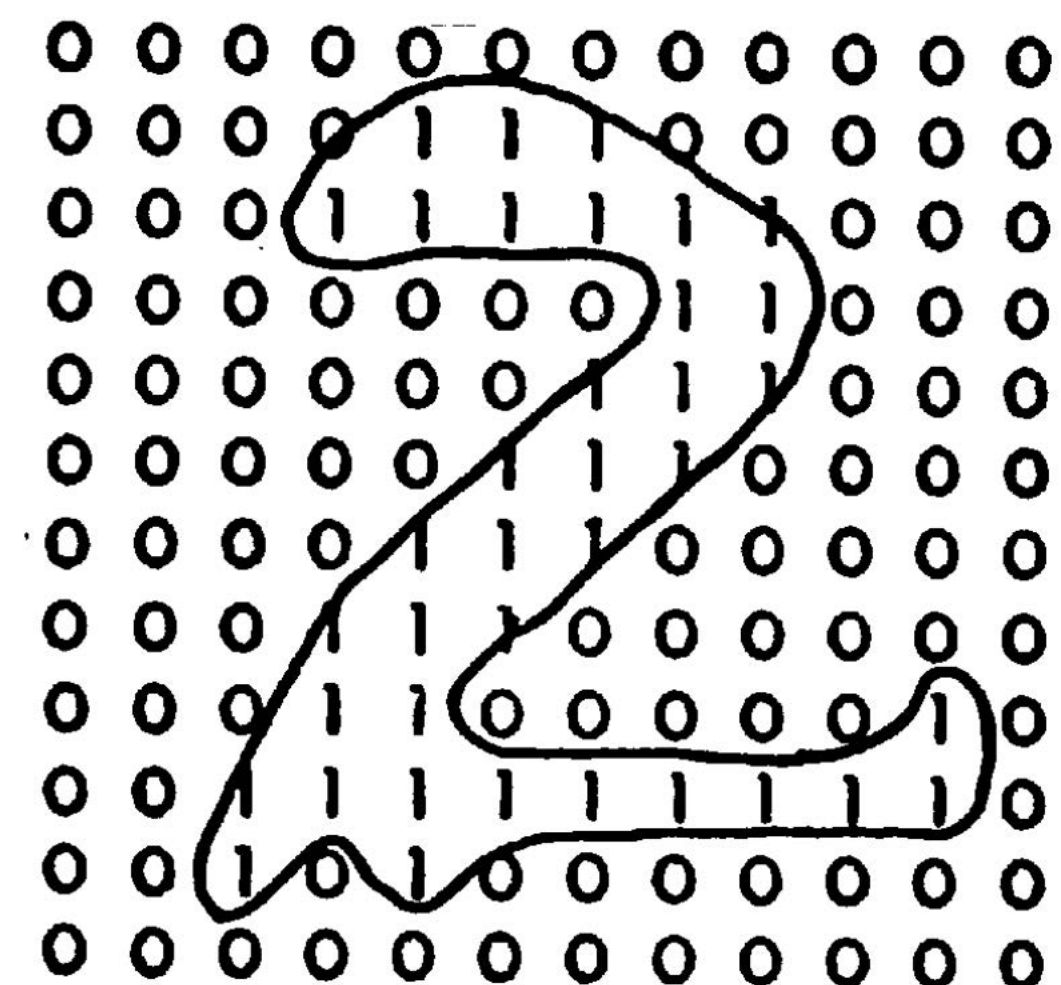
SCC

1 2 3 4 5 6 7 8 9 0

M.G.F

1 2 3 4 5 6 7 8 9 0

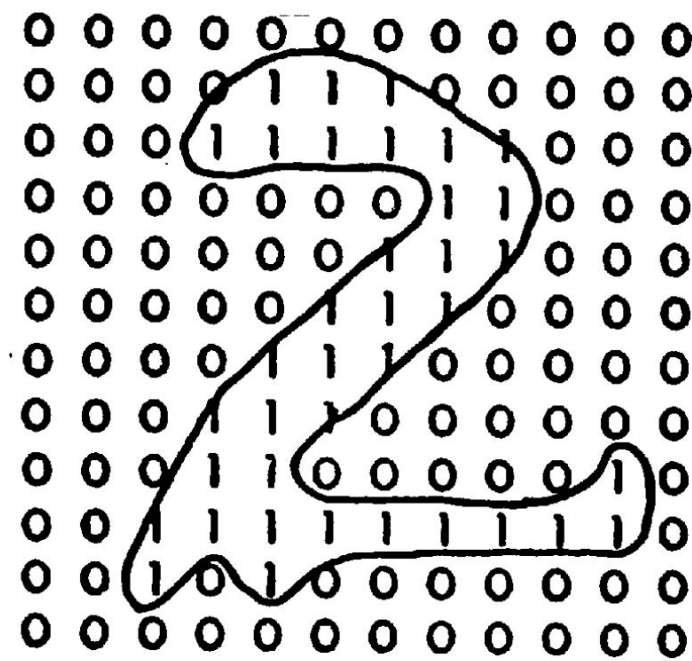
FWL



A B C D E F G H I
J K L M N O P Q R
S T U V W X Y Z
[blank] a b c d e f g h

The birth of modern machine learning

1 2 3 4 5 6 7 8 9 0



- Highleyman tests on Bledsoe's data, reports lower numbers than expected.
- Uses 40 alphabets for training, 10 for testing.
- Bledsoe retorts "send me teh data, ur doing it wrong!"
- Hihgleyman copies the data and sends it to Bledsoe.
- Others get jealous, so he sends it to Chow and SRI too.
- Everyone uses Highleyman's train-test split.
- Eventually he offers to make copies for whomever wants them.



The Design and Analysis of Pattern Recognition Experiments

By W. H. HIGHLEYMAN

(Manuscript received March 2, 1961)

A popular procedure for testing a pattern recognition machine is to present the machine with a set of patterns taken from the real world. The proportion of these patterns which are misrecognized or rejected is taken as the estimate of the error probability or rejection probability for the machine. In Part I, this testing procedure is discussed for the cases of unknown and known a priori probabilities of occurrence of the pattern classes. The differences between the tests that should be made in the two cases are noted, and confidence intervals for the test results are indicated. These concepts are applied to various published pattern recognition results by determining the appropriate confidence interval for each result.

In Part II, the problem of the optimum partitioning of a sample of fixed size between the design and test phases of a pattern recognition machine is discussed. One important nonparametric result is that the proportion of the total sample used for testing the machine should never be less than that proportion used for designing the machine, and in some cases should be a good deal more.

Leaderboards!

All characters

	Test Error
Bledsoe	60%
Chow	42%
MDH	32%
EMNIST-Balanced	9.4%

60x more data!

Would expect 7x
improvement in test error.

Digits Only

	Test Error
MDH	12%
MNIST	0.6%

150x more data!

Would expect 12x
improvement in test error.

Do a little better than this with
similar architecture of MDH.

Experiment was a failure

- Highleyman: “Although it was accepted as legitimate research by Brooklyn Polytechnic Institute, the thesis did not achieve its goals because of a lack of computer resources and reliable categorization algorithms.”
- Bledsoe: “it will be necessary to have a much larger sample (perhaps 1000 alphabets) before one can decide with any certainty how successful the n-tuple method will read characters with this much variability.”
- MDH: The error rates were “still far too high to be practical...larger and higher-quality datasets are needed for work aimed at achieving useful results... An array size of at least 20X20 is needed, with an optimum size of perhaps 30X30.”

Remarkably prophetic predictions!

And then winter came...

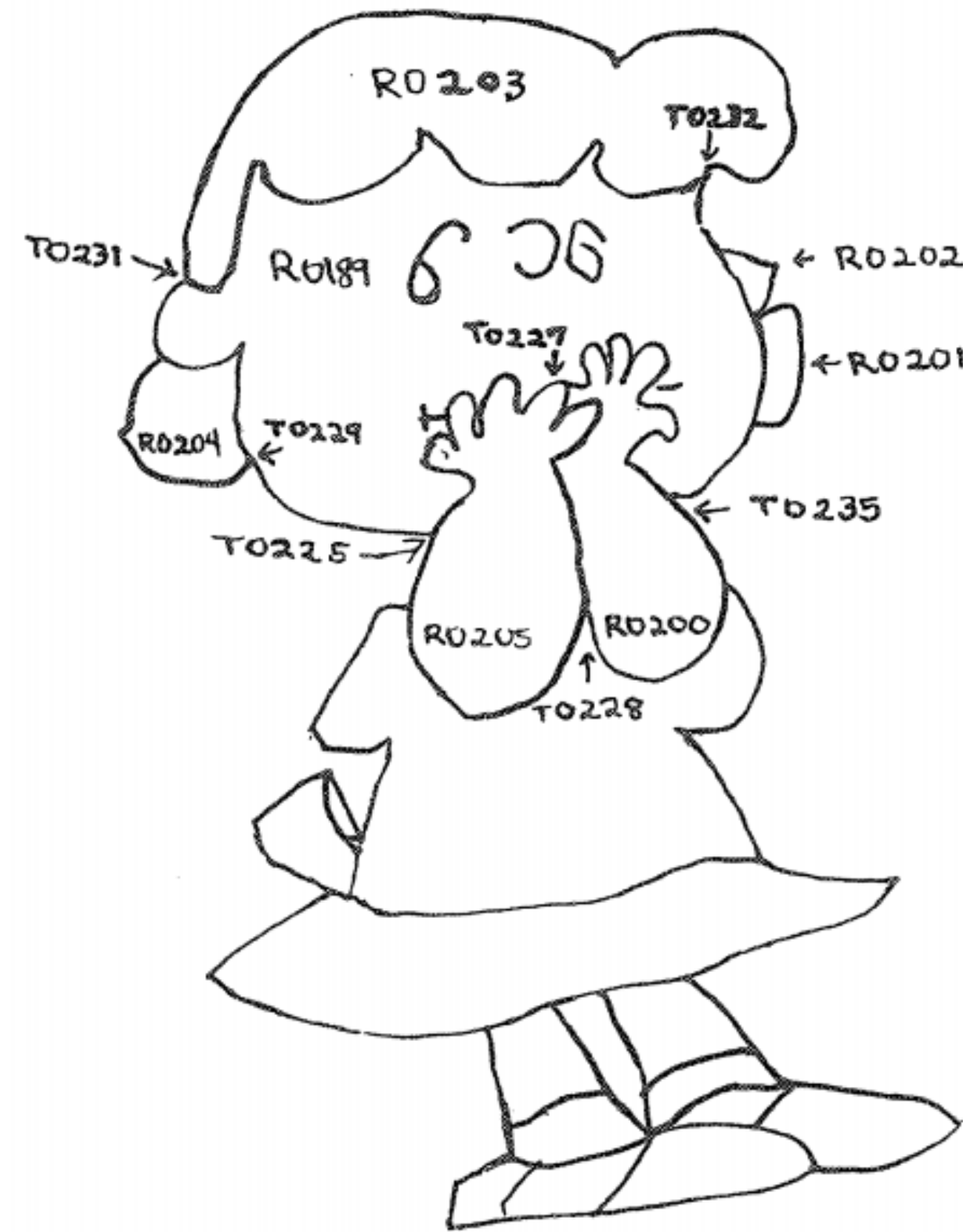


Figure 1

Example Pattern Recognition Data Set circa 1976

Herb Simon, plenary at ICML I, 1981:

2.4.3 Perceptrons

A final “classical” example (this is a negative example to prove my point) is the whole line of Perceptron research and nerve net learning [Rosenblatt, 1958]. A Perceptron is a system for classifying objects (that is, a discovery and learning system) that computes features of the stimulus display, then attempts to discriminate among different classes of displays by computing linear additive functions of these features. Functions producing correct choices are reinforced (receive increased weight), those producing incorrect choices have their weights reduced. I have to conclude (and here I don’t think I am in the minority) that this line of research didn’t get anywhere. The discovery task was just so horrendous for those systems that they never learned anything that people didn’t already know. So they should again strengthen our skepticism that the problems of AI are to be solved solely by building learning systems.

Papers at ICML I are almost not recognizable as machine learning papers:

A *maximally-specific* conjunctive generalization is the most detailed (most specific) description that is true of all of the known objects in the class. Since specific descriptions list many facts about the class, the maximally-specific conjunctive generalization is the longest conjunctive generalization that still describes all of the training instances.

N(eur)IPS 1987

- MURPHY: A Robot that Learns by Doing
- How Neural Nets Work
- Encoding Geometric Invariances in Higher-Order Neural Networks
- Performance Measures for Associative Memories that Learn and Forget
- An Optimization Network for Matrix Inversion
- Constrained Differential Optimization
- Introduction to a System for Implementing Neural Net Connections on SIMD Architectures

⋮

How did machine learning become machine learning again?

- Funding Pressures.
- FTP became widely available.

TIMIT



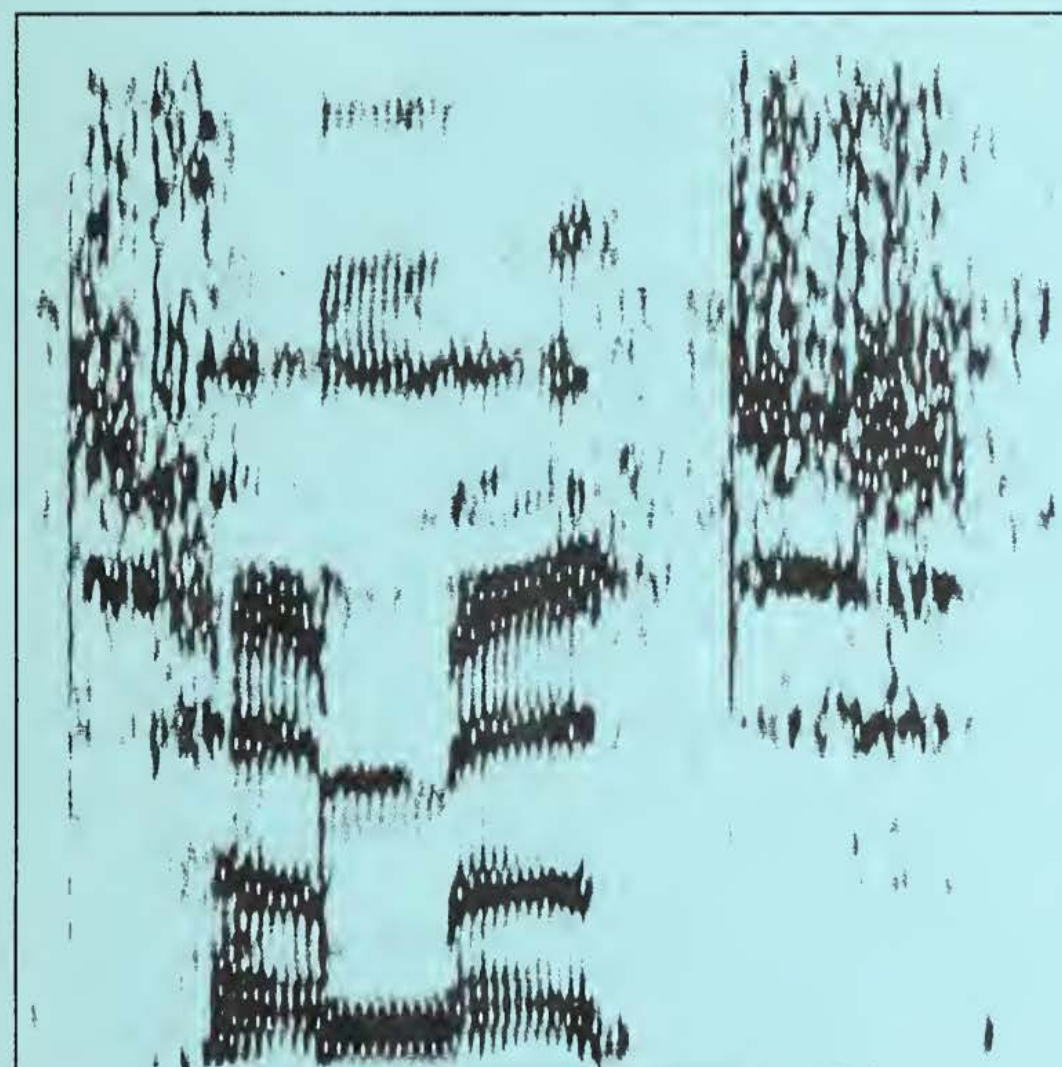
NIST
PUBLICATIONS

NISTIR 4930

DARPA TIMIT

Acoustic-Phonetic Continuous Speech Corpus CD-ROM

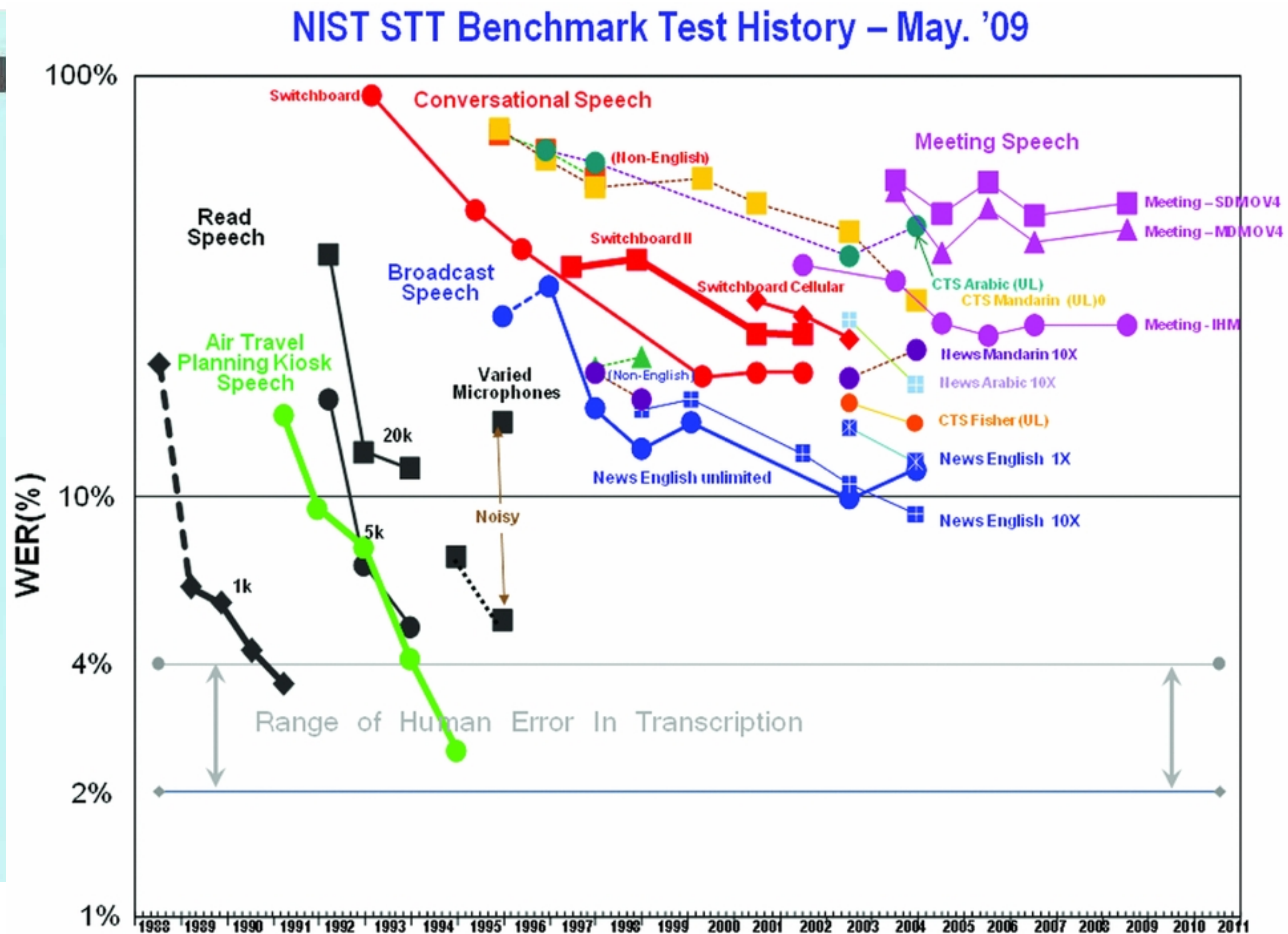
NIST Speech Disc 1-1.1



John S. Garofolo
Lori F. Lamel
William M. Fisher
Jonathan G. Fiscus
David S. Pallett
Nancy L. Dahlgren

U.S. DEPARTMENT OF COMMERCE
Technology Administration
National Institute of Standards
and Technology
Computer Systems Laboratory
Advanced Systems Division
Gaithersburg, MD 20899

CD-ROM Released October, 1990
Documentation Published February, 1993



Many attribute to PM Charles Wayne

UCI



- Made available by FTP in 1987
- Developed by then PhD student David Aha: “I was determined to create and share it, both because I wanted to use the datasets for my own research and because I thought it was ridiculous that the community hadn’t fielded what should have been a useful service.”
- "By ICML-95, the problems ‘caused’ by the repository had become popularly espoused. For example, at that conference Lorenza Saitta had, in an invited workshop that I co-organized, passionately decried how it allowed researchers to publish dull papers that proposed small variations of existing supervised learning algorithms and reported their small-but-significant incremental performance improvements in comparison studies.”

How did machine learning become machine learning again?

- Funding Pressures.
- FTP became widely available.
- Much easier to compare on same train-test splits than hand curating new data sets...
- Hence the “data-set-as-benchmark” paradigm was reborn.

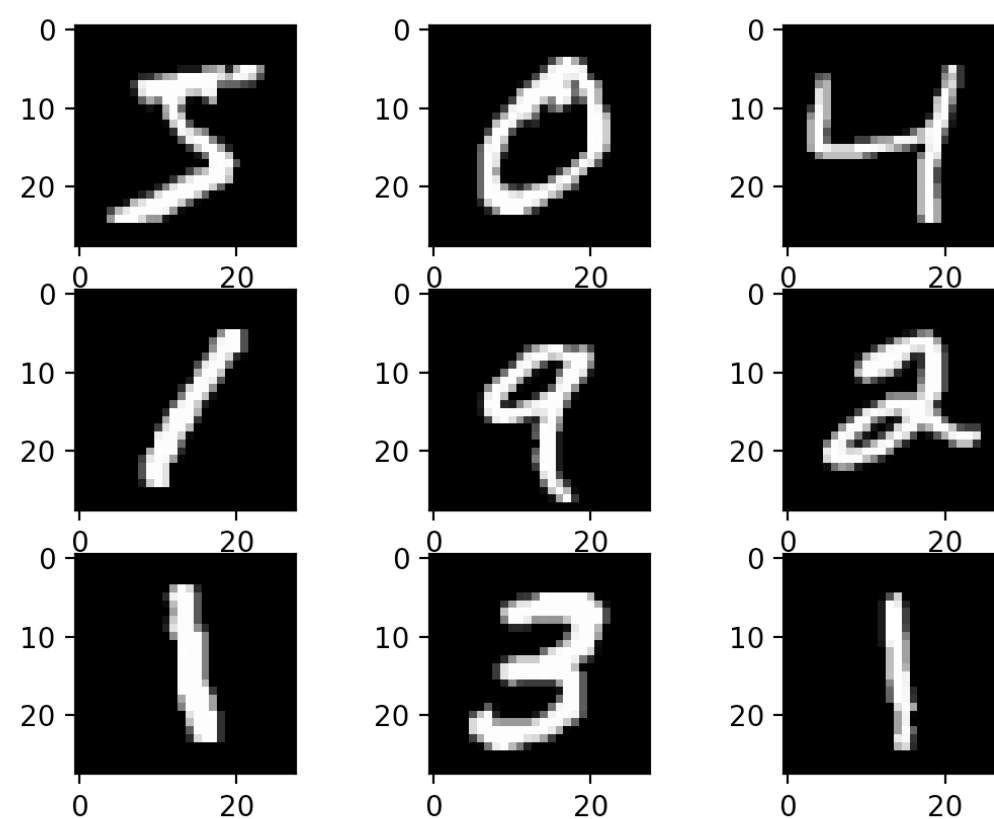
Why is this idea so sticky?

List of datasets for machine-learning research

From Wikipedia, the free encyclopedia

These [datasets](#) are applied for [machine-learning](#) research and have been cited in [peer-reviewed](#) academic journals. Datasets are an integral part of the field of machine learning. Major advances in this field can result from advances in learning [algorithms](#) (such as [deep learning](#)), computer hardware,

- Earliest handwriting data set is 1991 - no sign of Highleyman's lost data from 1959.
- Only dataset before 1986 is IRIS (1936)



1956 - Rosenblatt's Perceptron
1959 - Highleyman's Data creation
1962 - Novikoff Mistake bound
1962 - Highleyman on train-test split
1968 - First Issue of Pattern Recognition
1973 - Duda and Hart 1st Edition
1974 - Vapnik and Chervonenkis 1st book
1979 - First Issue of IEEE PAMI
1980 - First ICML Workshop
1984 - Classification and Decision Trees
1985 - RFC 595 FTP released
1985 - Invention of CD ROM
1986 - Snowbird workshop proposes NIPS
1986 - Elm released
1986 - First issue of Machine Learning
1986 - DARPA program creates TIMIT
1987 - First NIPS conference
1987 - UCI ftp site launched
1988 - TIMIT released
1989 - First issue of Neural Computation
1989 - Pine client released
1990 - TIMIT released on CD-ROM
1994 - MNIST creation
1998 - MNIST release (i.e., the singularity)

UCI Adult

- Predict whether income exceeds \$50K/yr based on census data.
- Features: age, workclass, fnlwgt, education, marital-status, occupation, relationship, race, gender, capital-gain, capital-loss, hours-per-week worked, native-country
- 48,842 examples. Standard train-test split into 32K/16K.
- Extraction was done by Barry Becker from the 1994 Census database.
- <http://www.census.gov/ftp/pub/DES/www/welcome.html>
- “A set of reasonably clean records was extracted using the following conditions:
((AAGE>16) && (AGI>100) && (AFNLWGT>1)&& (HRSWK>0))”

fnlwgt?

The weights on the CPS files are controlled to independent estimates of the civilian noninstitutional population of the US. These are prepared monthly for us by Population Division here at the Census Bureau. We use 3 sets of controls. These are:

1. A single cell estimate of the population 16+ for each state.
2. Controls for Hispanic Origin by age and sex.
3. Controls by Race, age and sex.

We use all three sets of controls in our weighting program and "rake" through them 6 times so that by the end we come back to all the controls we used.

The term estimate refers to population totals derived from CPS by creating "weighted tallies" of any specified socio-economic characteristics of the population.

People with similar demographic characteristics should have similar weights. There is one important caveat to remember about this statement. That is that since the CPS sample is actually a collection of 51 state samples, each with its own probability of selection, the statement only applies within state.

Wut?

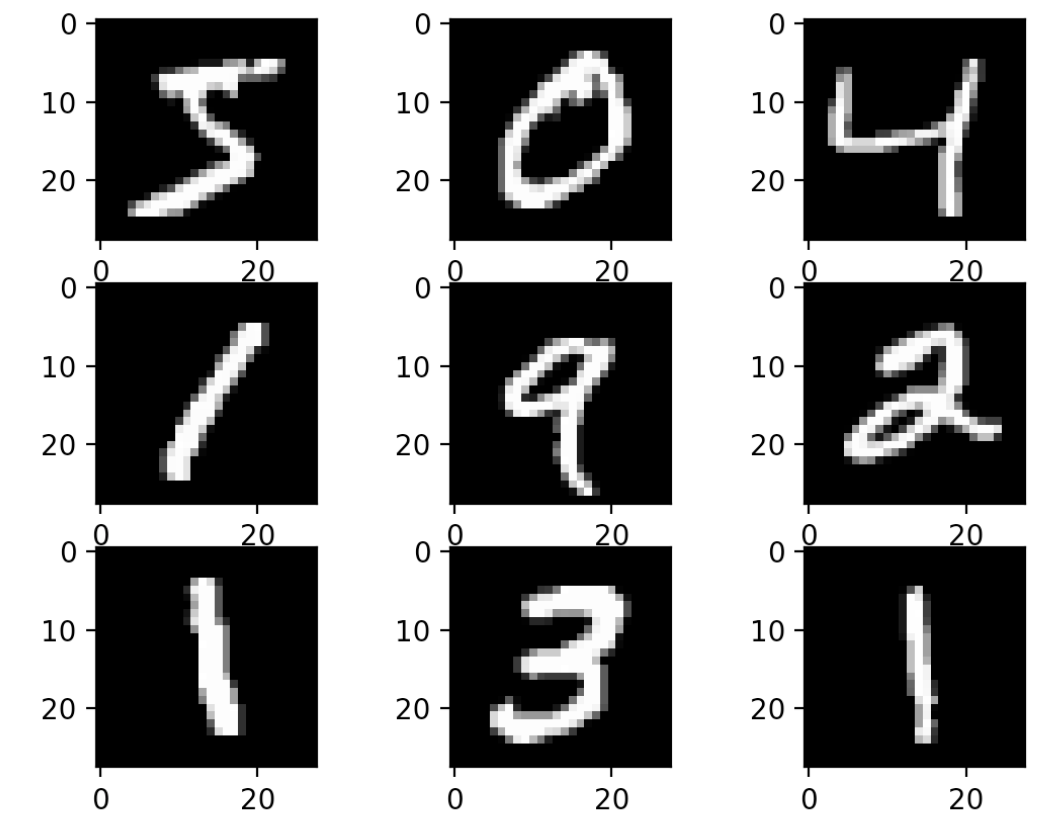
UCI Leaderboard 1994

- Blind leaderboard chasing of a completely meaningless task.
- One of the most popular “tabular” data sets.
- Worse, was adopted by the fairness community as a benchmark because it contained race and gender.

Algorithm	Error
C4.5	15.54
C4.5-auto	14.46
C4.5 rules	14.94
Voted ID3 (0.6)	15.64
Voted ID3 (0.8)	16.47
T2	16.84
IR	19.54
NBTree	14.10
CN2	16.00
HOODG	14.82
FSS Naive Bayes	14.05
IDTM (Decision table)	14.46
Naive-Bayes	16.12
Nearest-neighbor (1)	21.42
Nearest-neighbor (3)	20.35
OCI	15.04

MNIST

- 60,000 handwritten digits
- 10,000 test digits
- 28 x 28 in grayscale
- Released in 1998 to show convnets were better than everything else
- “It is a good database for people who want to try learning techniques and pattern recognition methods on real-world data while spending minimal efforts on preprocessing and formatting.”



Train-test split

- The original NIST data had the property that training and test data came from *two different* populations.
 - ~2,000 American Census Bureau employees,
 - ~500 American high school students.
- That was the original train-test split.
- MNIST combined these, and shuffled them both into train test.
- Exact procedure to derive MNIST from NIST was lost.

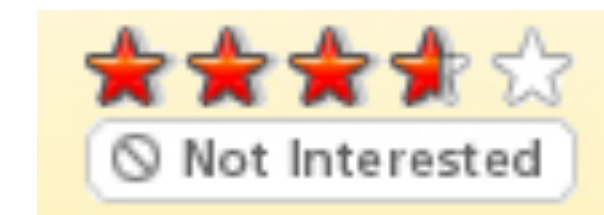
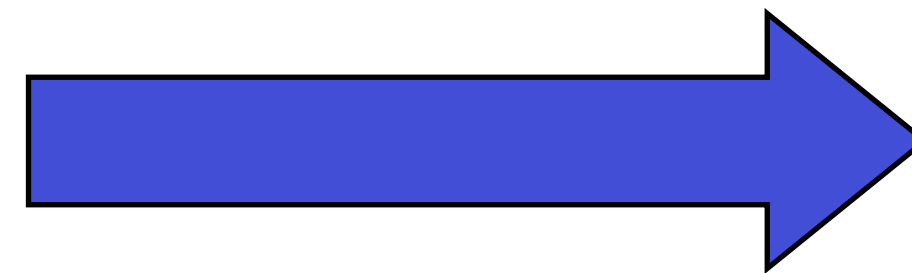
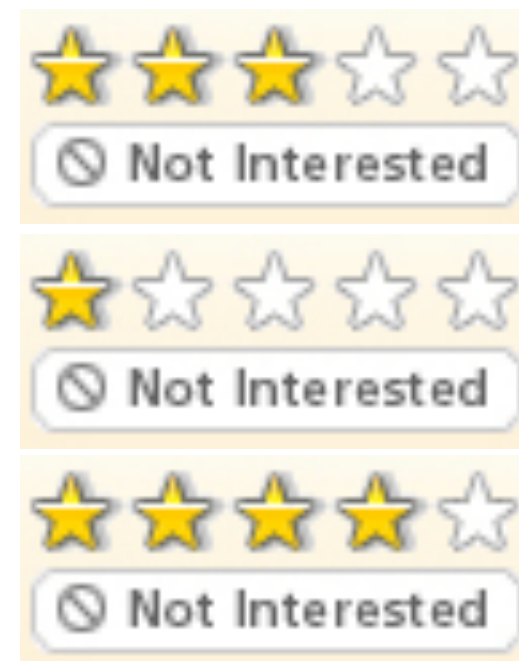
MNIST Leaderboard (RIP)

Non-Linear Classifiers			
40 PCA + quadratic classifier	none	3.3	LeCun et al. 1998
1000 RBF + linear classifier	none	3.6	LeCun et al. 1998
SVMs			
SVM, Gaussian Kernel	none	1.4	
SVM deg 4 polynomial	deskewing	1.1	LeCun et al. 1998
Reduced Set SVM deg 5 polynomial	deskewing	1.0	LeCun et al. 1998
Virtual SVM deg-9 poly [distortions]	none	0.8	LeCun et al. 1998
Virtual SVM, deg-9 poly, 1-pixel jittered	none	0.68	DeCoste and Scholkopf, MLJ 2002
Virtual SVM, deg-9 poly, 1-pixel jittered	deskewing	0.68	DeCoste and Scholkopf, MLJ 2002
Virtual SVM, deg-9 poly, 2-pixel jittered	deskewing	0.56	DeCoste and Scholkopf, MLJ 2002
Neural Nets			
2-layer NN, 300 hidden units, mean square error	none	4.7	LeCun et al. 1998
2-layer NN, 300 HU, MSE, [distortions]	none	3.6	LeCun et al. 1998
2-layer NN, 300 HU	deskewing	1.6	LeCun et al. 1998
2-layer NN, 1000 hidden units	none	4.5	LeCun et al. 1998
2-layer NN, 1000 HU, [distortions]	none	3.8	LeCun et al. 1998
3-layer NN, 300+100 hidden units	none	3.05	LeCun et al. 1998
3-layer NN, 300+100 HU [distortions]	none	2.5	LeCun et al. 1998
3-layer NN, 500+150 hidden units	none	2.95	LeCun et al. 1998
3-layer NN, 500+150 HU [distortions]	none	2.45	LeCun et al. 1998
3-layer NN, 500+300 HU, softmax, cross entropy, weight decay	none	1.53	Hinton, unpublished, 2005

If anyone wants to get to 1.18 error with two lines of code, go here:
github.com/benjamin-recht/mnist_l_pt_2.git

Netflix Prize

- One million big ones!
- Given 100 million ratings on a scale of 1 to 5, predict 3 million ratings to highest accuracy



How to fill in the blanks?

- 17770 total movies
- 480189 total users
- Over 8 billion total ratings

Netflix Prize

Leaderboard

Rank	Team Name	Best Score	% Improvement	Last Submit Time
--	No Grand Prize candidates yet	--	--	--
Grand Prize - RMSE <= 0.8563				
--	No Progress Prize candidates yet	--	--	--
Progress Prize - RMSE <= 0.8625				
1	When Gravity and Dinosaurs Unite	0.8675	8.82	2008-03-01 07:03:35
2	BellKor	0.8682	8.75	2008-02-28 23:40:45
3	Gravity	0.8708	8.47	2008-02-06 14:12:44
Progress Prize 2007 - RMSE = 0.8712 - Winning Team: KorBell				
4	KorBell	0.8712	8.43	2007-10-01 23:25:23
5	acmehill	0.8720	8.35	2008-03-02 05:08:12
6	Dan Tillberg	0.8727	8.27	2008-03-02 08:42:29
7	basho	0.8729	8.25	2007-11-24 14:27:00
8	Just a guy in a garage	0.8740	8.14	2008-02-06 12:16:40
9	BigChaos	0.8748	8.05	2008-03-01 17:26:06
10	Dinosaur Planet	0.8753	8.00	2007-10-04 04:56:45
•		•	•	•
•		•	•	•
•		•	•	•
50	amgl	0.8897	6.49	2007-12-23 18:44:03
51	Remco	0.8899	6.46	2007-04-04 06:16:56
52	mxlg	0.8900	6.45	2007-12-23 18:54:46
53	JustWithSVD	0.8900	6.45	2008-02-14 16:17:54
54	flyingswallo	0.8900	6.45	2008-02-28 09:56:20
55	OS_001	0.8901	6.44	2008-02-29 05:53:11
56	Bozo_The_Clown	0.8902	6.43	2007-09-06 17:24:48

Netflix Prize

Leaderboard

Mixture of
hundreds of
models, including
gradient descent

Gradient descent
on low-rank
parameterization

Rank	Team Name	Best Score	% Improvement	Last Submit Time
--	No Grand Prize candidates yet	--	--	--
Grand Prize - RMSE <= 0.8563				
--	No Progress Prize candidates yet	--	--	--
Progress Prize - RMSE <= 0.8625				
1	When Gravity and Dinosaurs Unite	0.8675	8.82	2008-03-01 07:03:35
2	BellKor	0.8682	8.75	2008-02-28 23:40:45
3	KorBell	0.8708	8.47	2008-02-06 14:12:44
RMSE 2007 - RMSE = 0.8712 - Winning Team: KorBell				
4	KorBell	0.8712	8.43	2007-10-01 23:25:23
5	acmehill	0.8720	8.35	2008-03-02 05:08:12
6	Dan Tillberg	0.8727	8.27	2008-03-02 08:42:29
7	basho	0.8729	8.25	2007-11-24 14:27:00
8	Just a guy in a garage	0.8740	8.14	2008-02-06 12:16:40
9	BigChaos	0.8748	8.05	2008-03-01 17:26:06
10	Dinosaur Planet	0.8753	8.00	2007-10-04 04:56:45
...	...	...	...	...
50	amgl	0.8897	6.49	2007-12-23 18:44:03
51	Remco	0.8899	6.46	2007-04-04 06:16:56
52	mxlg	0.8900	6.45	2007-12-23 18:54:46
53	JustWithSVD	0.8900	6.45	2008-02-14 16:17:54
54	fluffy	0.8900	6.45	2008-02-28 09:56:20
55	Bozo_The_Clown	0.8901	6.44	2008-02-29 05:53:11
...	...	...	...	...
...	Bozo_The_Clown	0.8902	6.43	2007-09-06 17:24:48

Controversy

- By correlating public IMDB profile with ratings database, could find movies in private watch list.
- “De-anonymization” by Narayanan and Shmatikov (<https://arxiv.org/abs/cs/0610105>)
- Potentially could reveal sensitive information.
- Lawyers smelled blood.
- Netflix pulled the data.
- Now no useful datasets exist to benchmark recommender systems.
- It is left for industry to innovate itself and also to “protect user privacy”
- Is that what we want?

Wordnet

- Lexical database of the English language. WordNet groups nouns into cognitive synonyms, called synsets. The words car and automobile fall into the same synset.
- WordNet provides a hierarchical tree structure according to a super-subordinate relationship between synsets. The synset for *chair*, for example, is a child of the synset for *furniture* in the wordnet hierarchy.

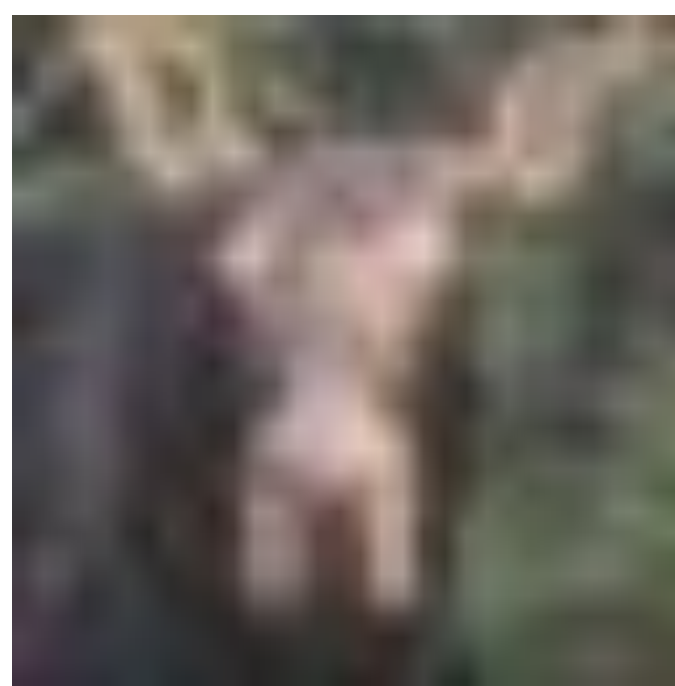
Tiny Images Dataset

- 79,302,017 images
- 32x32x3
- 400GB
- 75,062 non-abstract nouns (WordNet)
- Collected by [Torralba, Fergus, Freeman'08]
- Collected via queries to image search engines



CIFAR10

- 10 classes: airplane, car, bird, cat, deer, dog, frog, horse, ship, truck.
- 32×32 color images
- Used to prototype models for imagenet.
- (It is not true that something that is good on one is good on the other.)



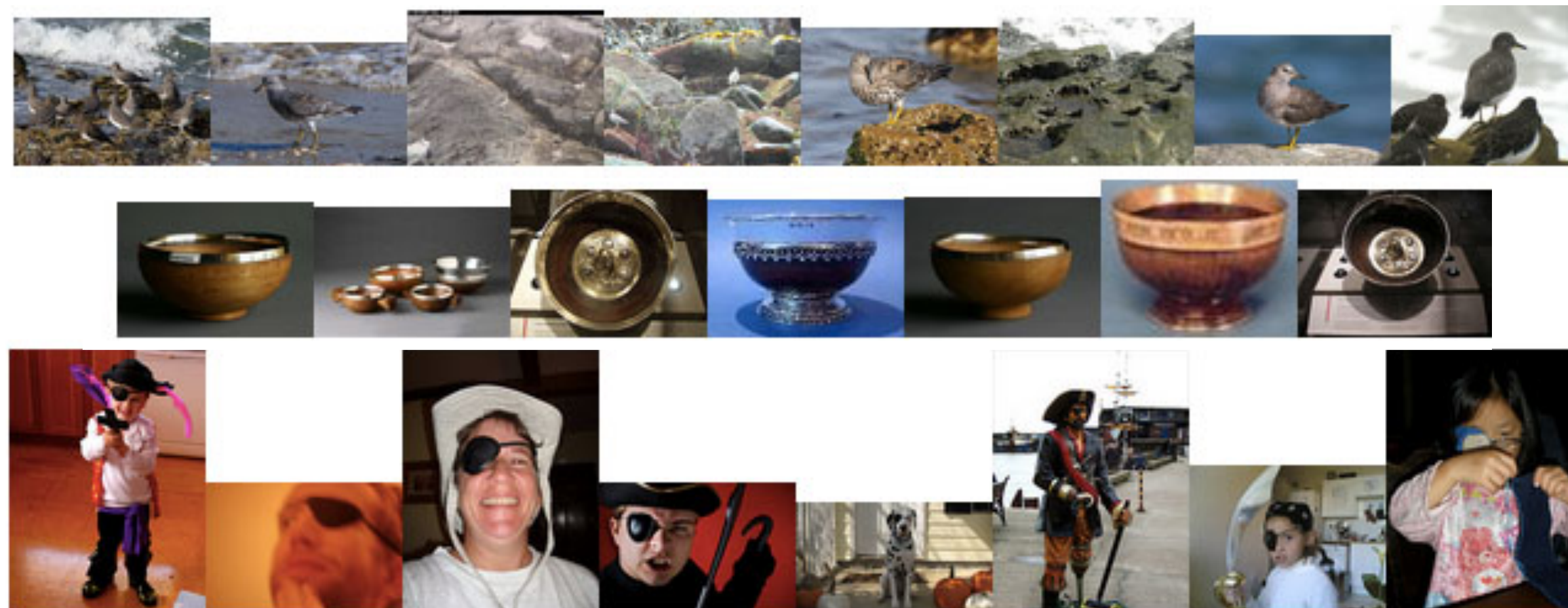
CIFAR-10 State of the Art

Year	Model	Test accuracy	
2009	Raw pixels	37.3%	
2009	RBM	64.8%	
2011	Random features	79.6%	
2012	AlexNet	88.5%	} Deeep networks
2014	VGG	92.8%	
2015	ResNet	93.5%	
2016	Wide ResNet	95.9%	
2017	Shake Shake	97.1%	



IMAGENET

- Introduced in [Deng, Dong, Socher, Li, 2009]
- organized according to the “WordNet hierarchy”
- 1.2 million training images, 50k validation images
- RGB color images with around 500 × 400 pixels
- 1,000 classes (about 150 dog breeds)



This research study is being conducted by Ben Recht, Rebecca Roelofs, Ludwig Schmidt, and Vaishaal Shankar from UC Berkeley. For questions about this study, please contact ludwig@berkeley.edu and roelofs@cs.berkeley.edu. In this study, we will ask you to indicate whether given images belong to a certain object category. Occasionally, the images may contain disturbing or adult content. We would like to remind you that participation in our study is voluntary and that you can withdraw from the study at any time.

Which of these images contain at least one object of type

bow

Definition: a weapon for shooting arrows, composed of a curved piece of resilient wood with a taut cord to propel the arrow

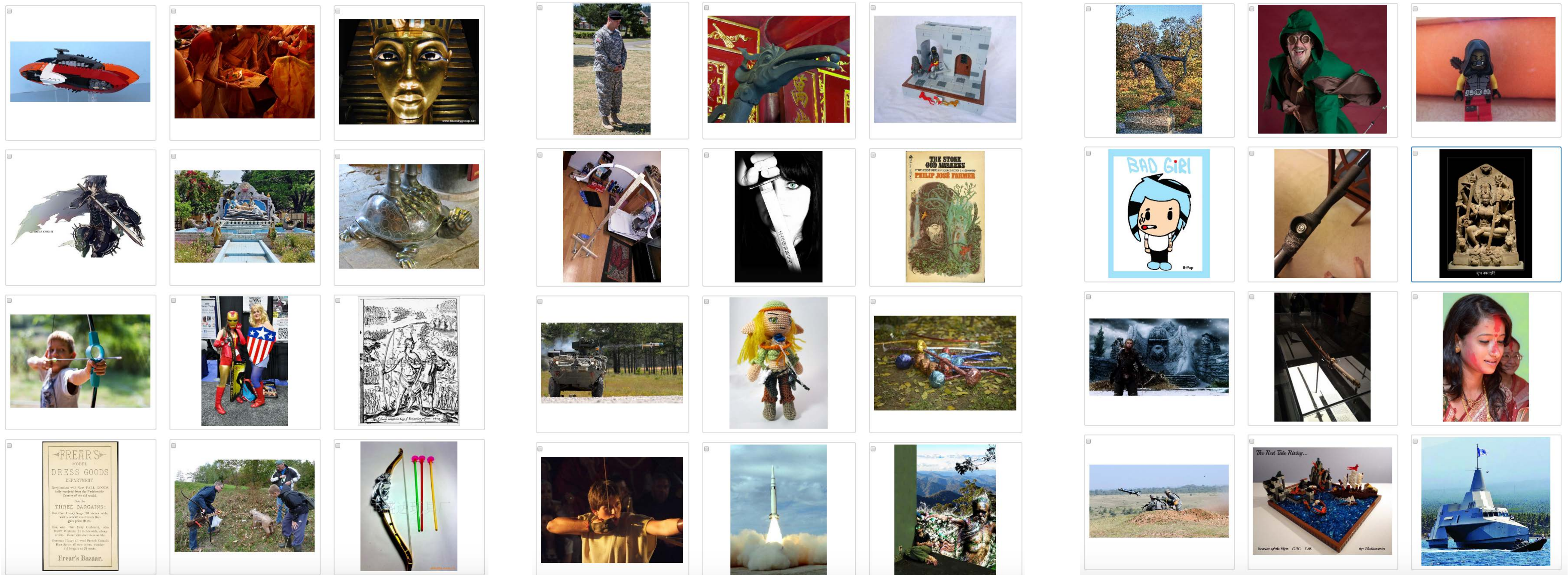
Task:

For each of the following images, check the box next to an image if it contains at least one object of type *bow*. Select an image if it contains the object regardless of occlusions, other objects, and clutter or text in the scene. Only select images that are photographs (**no drawings or paintings**).

Please make accurate selections!

If you are unsure about the object meaning, please also consult the following Wikipedia page(s): https://en.wikipedia.org/wiki/Bow_and_arrow

If it is impossible to complete a HIT due to missing data or other problems, please return the HIT.



Submit

This research study is being conducted by Ben Recht, Rebecca Roelofs, Ludwig Schmidt, and Vaishaal Shankar from UC Berkeley. For questions about this study, please contact ludwig@berkeley.edu and roelofs@cs.berkeley.edu. In this study, we will ask you to indicate whether given images belong to a certain object category. Occasionally, the images may contain disturbing or adult content. We would like to remind you that participation in our study is voluntary and that you can withdraw from the study at any time.

Which of these images contain at least one object of type

bow

Definition: a weapon for shooting arrows, composed of a curved piece of resilient wood with a taut cord to propel the arrow

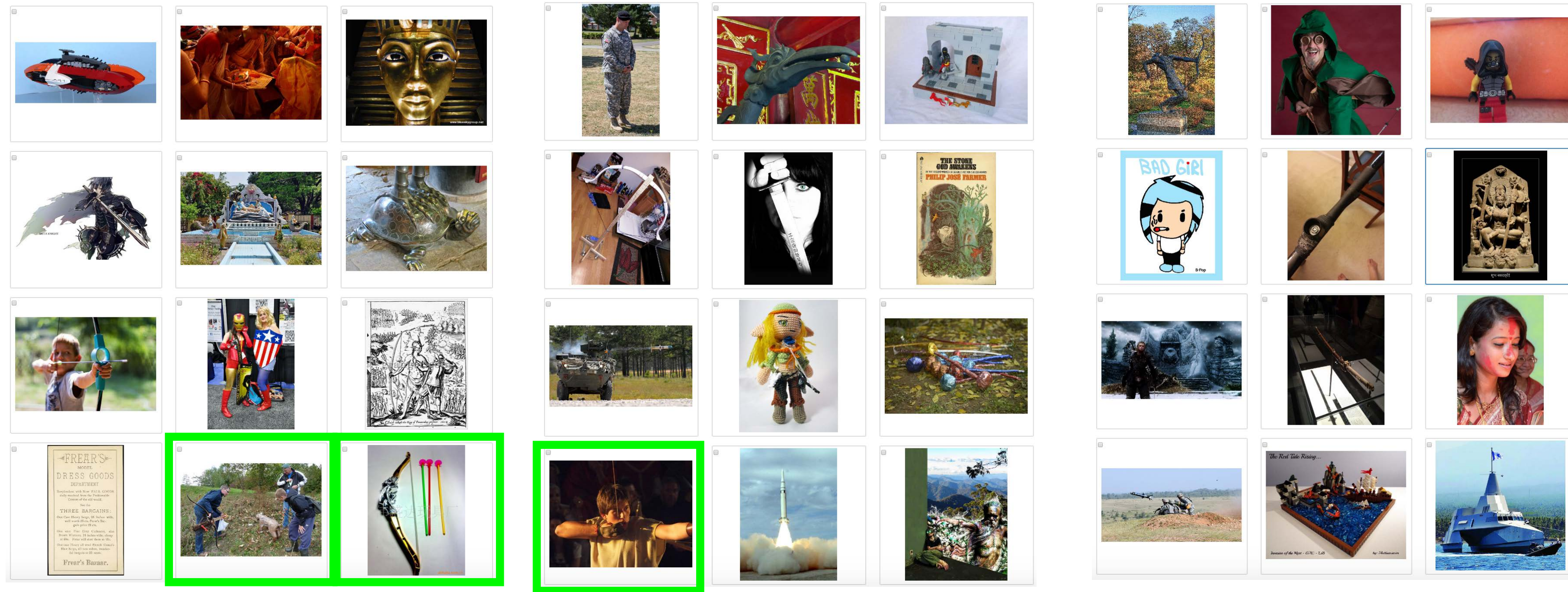
Task:

For each of the following images, check the box next to an image if it contains at least one object of type *bow*. Select an image if it contains the object regardless of occlusions, other objects, and clutter or text in the scene. Only select images that are photographs (**no drawings or paintings**).

Please make accurate selections!

If you are unsure about the object meaning, please also consult the following Wikipedia page(s): https://en.wikipedia.org/wiki/Bow_and_arrow

If it is impossible to complete a HIT due to missing data or other problems, please return the HIT.



Submit



1.0



0.7



0.5



0.2

ImageNet Inconsistencies

Mushroom vs. Gyromitra



ILSVRC2012_val_00023237.JPEG

Subset Relationships

Wood Rabbit



Problematic Images

n02641379 gar, garfish, garpike, billfish, *Lepisosteus osseus*

Gloss: primitive predaceous North American fish covered with hard scales and having long jaws with needlelike teeth

Synsets are not synonyms

Sunglass

a convex lens that focuses the rays of the sun; used to start a fire



ILSVRC2012_val_00030556.JPEG

Redefined Classes

Magpie



ILSVRC2012_val_00035348.JPEG

Drawings or Paintings

ILSVRC2012_val_00033112.JPEG



ILSVRC2012_val_00029666.JPEG

Near Duplicates

Why report top-5 accuracy?



Acorn Squash, Butternut Squash, Spaghetti Squash, Grocery Store

Multiple correct classes



Great Pyrenees



Kuvasz

Classes that are impossible
to distinguish

Ground Truth



Bernese
Mountain Dog



Greater Swiss
Mount Dog



Appenzeller

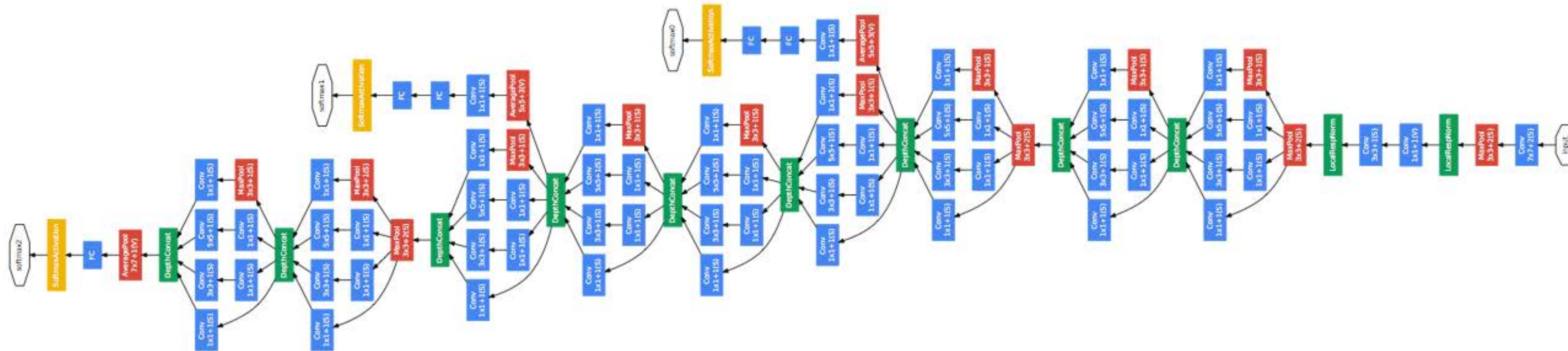


EntleBucher

ImageNet Competition



- AlexNet
- VGG
- InceptionNet
- ResNet
- Everyone still loves to show off performance on ImageNet as meaningful of progress.



The Benchmark Paradigm

- Has driven machine learning research for 40 years
- Massive progress made through competitive testing
- What is progress?
- What is in these data sets?
- How to think about issues of consent, privacy, copyright, etc.?

1956 - Rosenblatt's Perceptron
1959 - Highleyman's Data creation
1962 - Novikoff Mistake bound
1962 - Highleyman on train-test split
1968 - First Issue of Pattern Recognition
1973 - Duda and Hart 1st Edition
1974 - Vapnik and Chervonenkis 1st book
1979 - First Issue of IEEE PAMI
1980 - First ICML Workshop
1984 - Classification and Decision Trees
1985 - RFC 595 FTP released
1985 - Invention of CD ROM
1986 - Snowbird workshop proposes NIPS
1986 - Elm released
1986 - First issue of Machine Learning
1986 - DARPA program creates TIMIT
1987 - First NIPS conference
1987 - UCI ftp site launched
1988 - TIMIT released
1989 - First issue of Neural Computation
1989 - Pine client released
1990 - TIMIT released on CD-ROM
1994 - MNIST creation
1998 - MNIST release (i.e., the singularity)

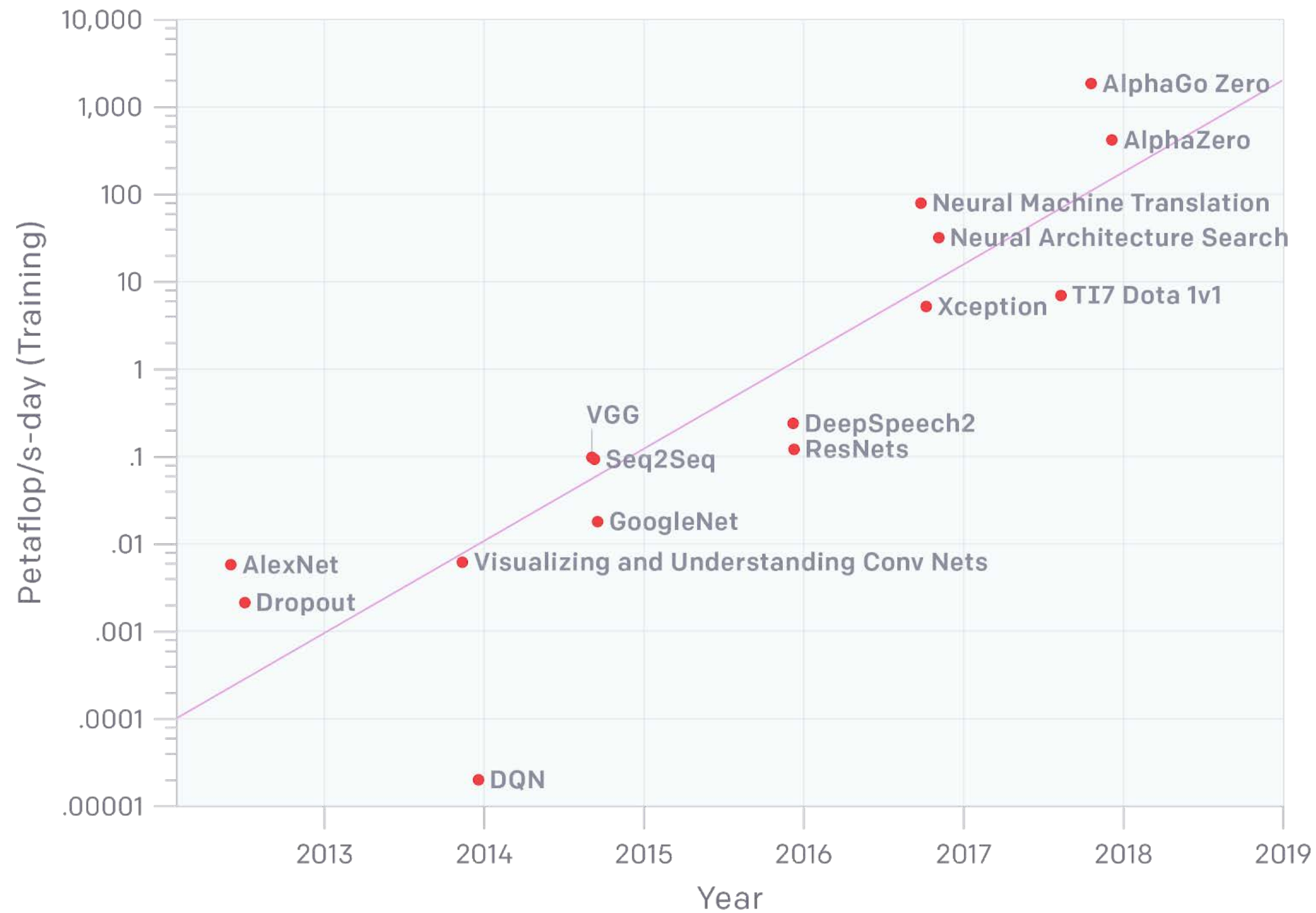
We use the same algorithms and same theory.

All that has changed: our computers are faster
and our data is larger.

That this was necessary was predicted before
1970 but took 50 years to come to fruition.

What does the next 50 years hold for
machine learning?

AlexNet to AlphaGo Zero: A 300,000x Increase in Compute



G. Brockman

Timeline of Machine Learning Evaluation

Antiquity Era

Pre-1990s

- *Highleyman's data* (1959)
- *UCI Machine Learning Repository* (1987)
- *TIMIT* (1988)

Middle Ages

1990s - 2019

- MNIST (1998-2012)
- Pascal VOC (2005-2012), CIFAR (2008)
- ImageNet (2009-2017)
- Netflix Prize (2006-2009)
- SQuAD (2016)
- GLUE (2018) / SuperGLUE (2019)

Modern Era

2020 - Now

- MMLU (2021)
- ARC-Challenge (2019)
- GSM8k (2021)
- HumanEval (2022)
- BIG Bench (2022)
- Chatbot Arena (2024)
- Dynabench (2021)
- SWE-bench Verified (2024)
- ... more



1980

- Checkers (1949)
- Chess (1966)
- Eleusis (card game) (1983)

1990

- Backgammon (1992)
- Jeopardy! (2011)
- Atari Learning Environment (ALE) (2012)
- Go (2015)
- OpenAI Gym (2016); AI Safety Gridworlds (2017)
- Dota2, StarCraft II (2019)
- Poker (2019)

2000

2010

2020

2030

- Diplomacy (2022)
- ...more

Test set saturation/obsolescence

- How long can a benchmark be useful for?
- ImageNet competition
- GLUE

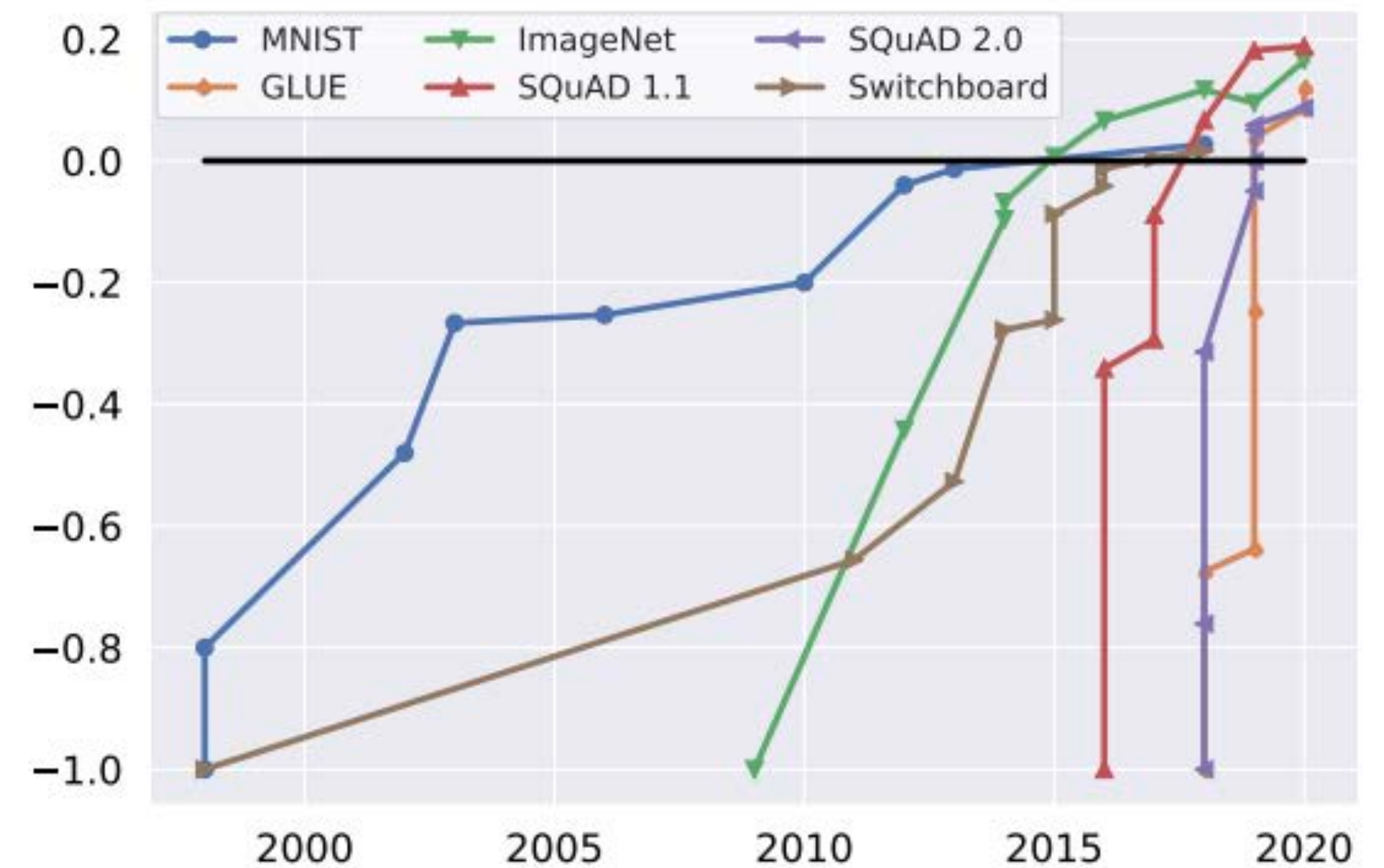


Figure 1: Benchmark saturation over time for popular benchmarks, normalized with initial performance at minus one and human performance at zero.

“Polymorphic” era: How do you evaluate a model when there is no specified task?

GPT-2 (2019)

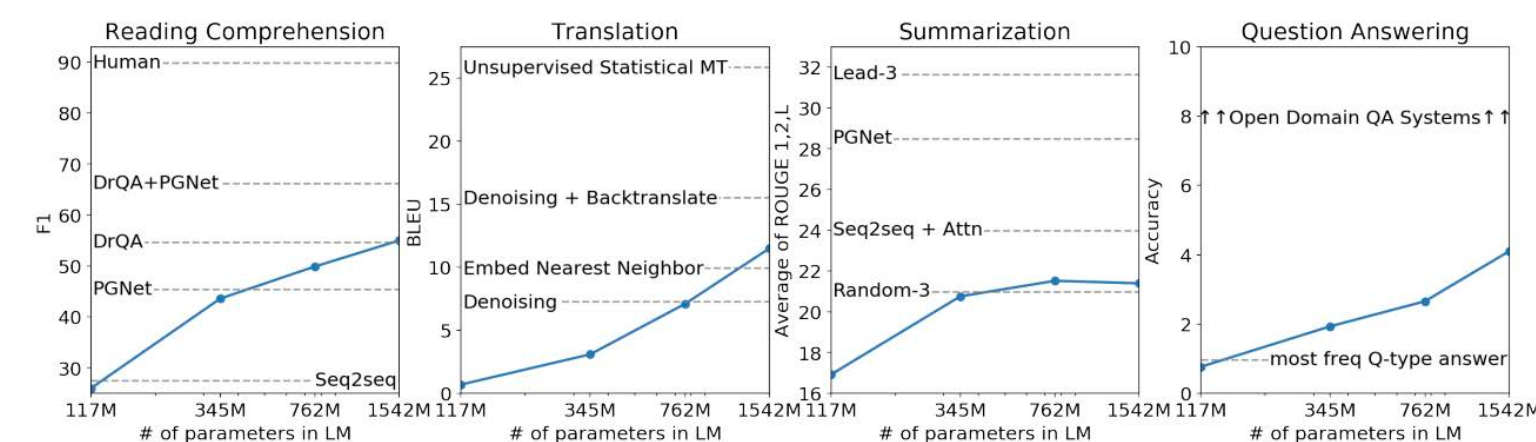


Figure 1. Zero-shot task performance of WebText LMs as a function of model size on many NLP tasks. Reading Comprehension results are on CoQA (Reddy et al., 2018), translation on WMT-14 Fr-En (Artetxe et al., 2017), summarization on CNN and Daily Mail (See et al., 2017), and Question Answering on Natural Questions (Kwiatkowski et al., 2019). Section 3 contains detailed descriptions of each result.

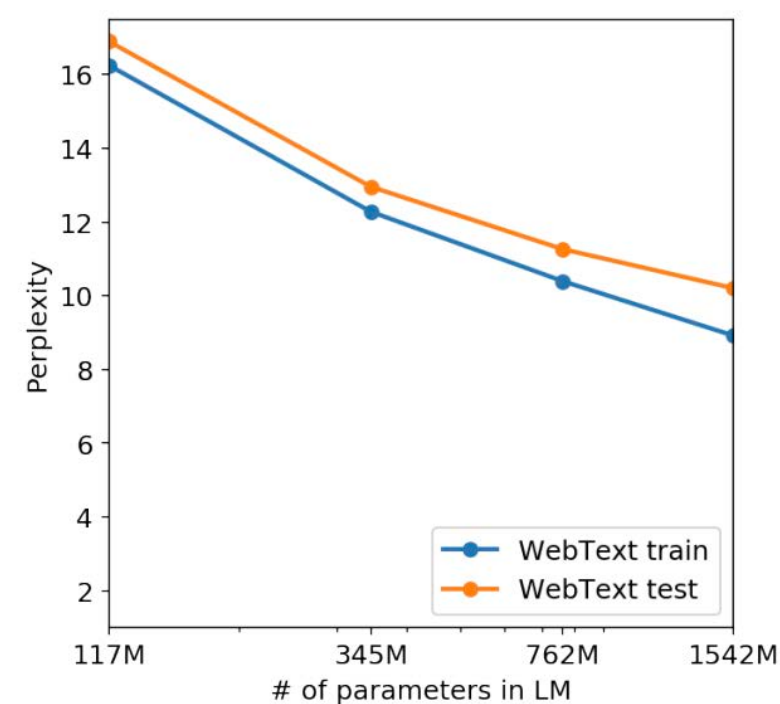


Figure 4. The performance of LMs trained on WebText as a function of model size.

GPT-3 (2020)

	SuperGLUE Average	BoolQ Accuracy	CB Accuracy	CB F1	COPA Accuracy	RTE Accuracy
Fine-tuned SOTA	89.0	91.0	96.9	93.9	94.8	92.5
Fine-tuned BERT-Large	69.0	77.4	83.6	75.7	70.6	71.7
GPT-3 Few-Shot	71.8	76.4	75.6	52.0	92.0	69.0

	WiC Accuracy	WSC Accuracy	MultiRC Accuracy	MultiRC F1a	ReCoRD Accuracy	ReCoRD F1
Fine-tuned SOTA	76.1	93.8	62.3	88.2	92.5	93.3
Fine-tuned BERT-Large	69.6	64.6	24.1	70.0	71.3	72.0
GPT-3 Few-Shot	49.4	80.1	30.5	75.4	90.2	91.1

Table 3.8: Performance of GPT-3 on SuperGLUE compared to fine-tuned baselines and SOTA. All results are reported on the test set. GPT-3 few-shot is given a total of 32 examples within the context of each task and performs no gradient updates.

Took the most popular
linguistics inspired
“BERTology”
benchmarks

[Winograd-Style Tasks;
lots of entailment, NLI,
QA, etc.]

GPT-4 (2022)

	GPT-4 Evaluated few-shot	GPT-3.5 Evaluated few-shot	LM SOTA Best external LM evaluated few-shot	SOTA Best external model (incl. benchmark-specific tuning)
MMLU [49] Multiple-choice questions in 57 subjects (professional & academic)	86.4% 5-shot	70.0% 5-shot	70.7% 5-shot U-PaLM [50]	75.2% 5-shot Flan-PaLM [51]
HellaSwag [52] Commonsense reasoning around everyday events	95.3% 10-shot	85.5% 10-shot	84.2% LLaMA (validation set) [28]	85.6 ALUM [53]
AI2 Reasoning Challenge (ARC) [54] Grade-school multiple choice science questions. Challenge-set.	96.3% 25-shot	85.2% 25-shot	85.2% 8-shot PaLM [55]	86.5% ST-MOE [18]
WinoGrande [56] Commonsense reasoning around pronoun resolution	87.5% 5-shot	81.6% 5-shot	85.1% 5-shot PaLM [3]	85.1% 5-shot PaLM [3]
HumanEval [43] Python coding tasks	67.0% 0-shot	48.1% 0-shot	26.2% 0-shot PaLM [3]	65.8% CodeT + GPT-3.5 [57]
DROP [58] (F1 score) Reading comprehension & arithmetic.	80.9 3-shot	64.1 3-shot	70.8 1-shot PaLM [3]	88.4 QDGAT [59]
GSM-8K [60] Grade-school mathematics questions	92.0% * 5-shot chain-of-thought	57.1% 5-shot	58.8% 8-shot Minerva [61]	87.3% Chinchilla + SFT+ORM-RL, ORM reranking [62]

Table 2. Performance of GPT-4 on academic benchmarks. We compare GPT-4 alongside the best SOTA (with benchmark-specific training) and the best SOTA for an LM evaluated few-shot. GPT-4 outperforms existing LMs on all benchmarks, and beats SOTA with benchmark-specific training on all datasets except DROP. For each task we report GPT-4’s performance along with the few-shot method used to evaluate. For GSM-8K, we included part of the training set in the GPT-4 pre-training mix (see Appendix E), and we use chain-of-thought prompting [11] when evaluating. For multiple-choice questions, we present all answers (ABCD) to the model and ask it to choose the letter of the answer, similarly to how a human would solve such a problem.

Standardized test
questions

Practical, task specific
datasets; perplexity

GPT-4

obsessed with “standardized testing”

Exam	GPT-4	GPT-4 (no vision)	GPT-3.5
Uniform Bar Exam (MBE+MEE+MPT)	298 / 400 (~90th)	298 / 400 (~90th)	213 / 400 (~10th)
LSAT	163 (~88th)	161 (~83rd)	149 (~40th)
SAT Evidence-Based Reading & Writing	710 / 800 (~93rd)	710 / 800 (~93rd)	670 / 800 (~87th)
SAT Math	700 / 800 (~89th)	690 / 800 (~89th)	590 / 800 (~70th)
Graduate Record Examination (GRE) Quantitative	163 / 170 (~80th)	157 / 170 (~62nd)	147 / 170 (~25th)
Graduate Record Examination (GRE) Verbal	169 / 170 (~99th)	165 / 170 (~96th)	154 / 170 (~63rd)
Graduate Record Examination (GRE) Writing	4 / 6 (~54th)	4 / 6 (~54th)	4 / 6 (~54th)
USABO Semifinal Exam 2020	87 / 150 (99th - 100th)	87 / 150 (99th - 100th)	43 / 150 (31st - 33rd)
USNCO Local Section Exam 2022	36 / 60	38 / 60	24 / 60
Medical Knowledge Self-Assessment Program	75 %	75 %	53 %
Codeforces Rating	392 (below 5th)	392 (below 5th)	260 (below 5th)
AP Art History	5 (86th - 100th)	5 (86th - 100th)	5 (86th - 100th)
AP Biology	5 (85th - 100th)	5 (85th - 100th)	4 (62nd - 85th)
AP Calculus BC	4 (43rd - 59th)	4 (43rd - 59th)	1 (0th - 7th)
AP Chemistry	4 (71st - 88th)	4 (71st - 88th)	2 (22nd - 46th)
AP English Language and Composition	2 (14th - 44th)	2 (14th - 44th)	2 (14th - 44th)
AP English Literature and Composition	2 (8th - 22nd)	2 (8th - 22nd)	2 (8th - 22nd)
AP Environmental Science	5 (91st - 100th)	5 (91st - 100th)	5 (91st - 100th)
AP Macroeconomics	5 (84th - 100th)	5 (84th - 100th)	2 (33rd - 48th)
AP Microeconomics	5 (82nd - 100th)	4 (60th - 82nd)	4 (60th - 82nd)
AP Physics 2	4 (66th - 84th)	4 (66th - 84th)	3 (30th - 66th)
AP Psychology	5 (83rd - 100th)	5 (83rd - 100th)	5 (83rd - 100th)
AP Statistics	5 (85th - 100th)	5 (85th - 100th)	3 (40th - 63rd)
AP US Government	5 (88th - 100th)	5 (88th - 100th)	4 (77th - 88th)
AP US History	5 (89th - 100th)	4 (74th - 89th)	4 (74th - 89th)
AP World History	4 (65th - 87th)	4 (65th - 87th)	4 (65th - 87th)
AMC 10 ³	30 / 150 (6th - 12th)	36 / 150 (10th - 19th)	36 / 150 (10th - 19th)
AMC 12 ³	60 / 150 (45th - 66th)	48 / 150 (19th - 40th)	30 / 150 (4th - 8th)
Introductory Sommelier (theory knowledge)	92 %	92 %	80 %
Certified Sommelier (theory knowledge)	86 %	86 %	58 %
Advanced Sommelier (theory knowledge)	77 %	77 %	46 %
Leetcode (easy)	31 / 41	31 / 41	12 / 41
Leetcode (medium)	21 / 80	21 / 80	8 / 80
Leetcode (hard)	3 / 45	3 / 45	0 / 45

Table 1. GPT performance on academic and professional exams. In each case, we simulate the conditions and scoring of the real exam. We report GPT-4’s final score graded according to exam-specific rubrics, as well as the percentile of test-takers achieving GPT-4’s score.

TECH

OpenAI announces GPT-4, claims it can beat 90% of humans on the SAT

PUBLISHED TUE, MAR 14 2023•1:42 PM EDT | UPDATED TUE, MAR 14 2023•2:32 PM EDT



Kif Leswing
@KIFLESWING

SHARE [f](#) [X](#) [in](#) [✉](#)

Here’s how GPT-4 scored on the GRE, LSAT, AP English, and other exams



J.D. Capelouto

Updated Mar 15, 2023, 11:52am PDT TECH NORTH AMERICA

©CBS NEWS ChatGPT bot passes law school exam



ChatGPT may be smart enough to graduate law school

TECH

OpenAI announces GPT-4, claims it can beat 90% of humans on the SAT

PUBLISHED TUE, MAR 14 2023•1:42 PM EDT | UPDATED TUE, MAR 14 2023•2:32 PM EDT



Kif Leswing
@KIFLESWING

SHARE [f](#) [X](#) [in](#) [✉](#)

Here's how GPT-4 scored on the GRE, LSAT, AP English, and other exams



JD Canelotto

INSIDER

Researchers just tested ChatGPT on the same test questions as aspiring doctors – and found the AI was 'comfortably within the passing range'

Daily **Mail**

The AI doctor will see you now: ChatGPT passes gold-standard US medical exam — as researchers hail moment as 'milestone for artificial intelligence'

GPT-4

Demo-based, strange anthropomorphization going on

GPT-4

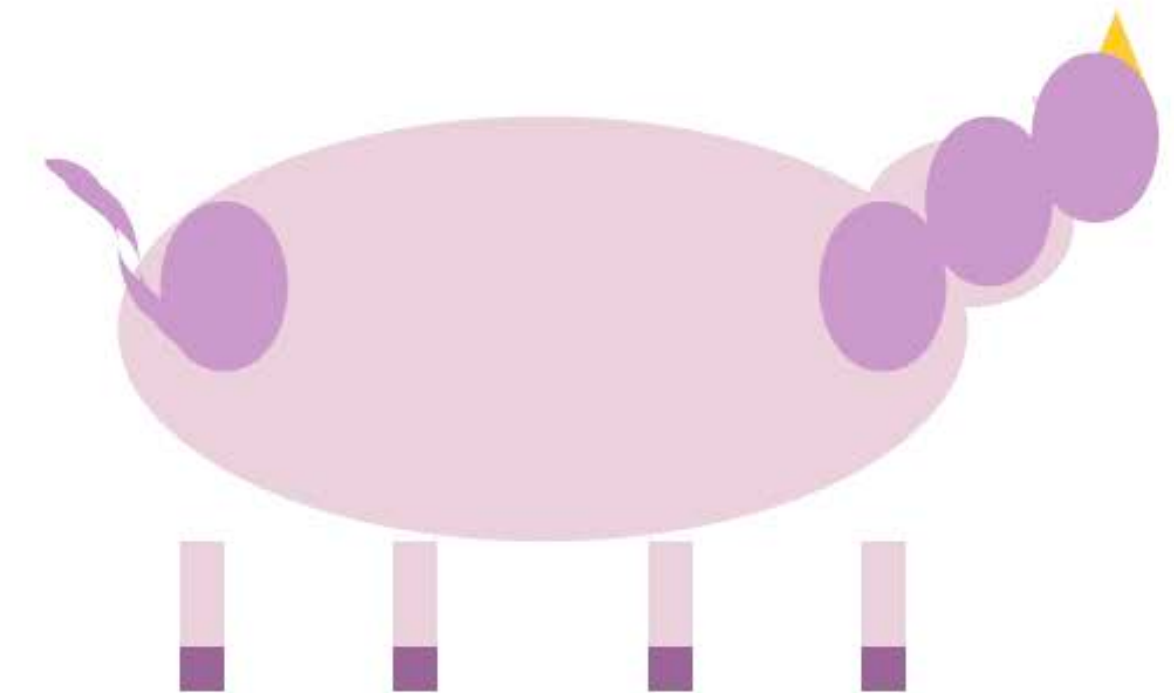
Prompt: Can you write a proof that there are infinitely many primes, with every line that rhymes?

GPT-4:

Yes, I think I can,
Though it might take a clever plan.
I'll start by noting Euclid's proof,
Which shows that primes aren't just aloof.
Assume we have a finite list,
Of primes, and that none have been missed.
Multiply them all together,
And add one, just to be clever.
The result will be a number,
That has no prime factors, oh wonder!
But every number has a prime,
That divides it, it's just a matter of time.
So we've found a contradiction,
And our finite list needs eviction.
There must be infinitely many primes,
And that's the end of my rhyming lines.s

Prompt: Draw a unicorn in TiKZ.

GPT-4: [Produces $\text{\LaTeX}$ compiling to following picture.]



Upcoming, new “benchmarking” trends...

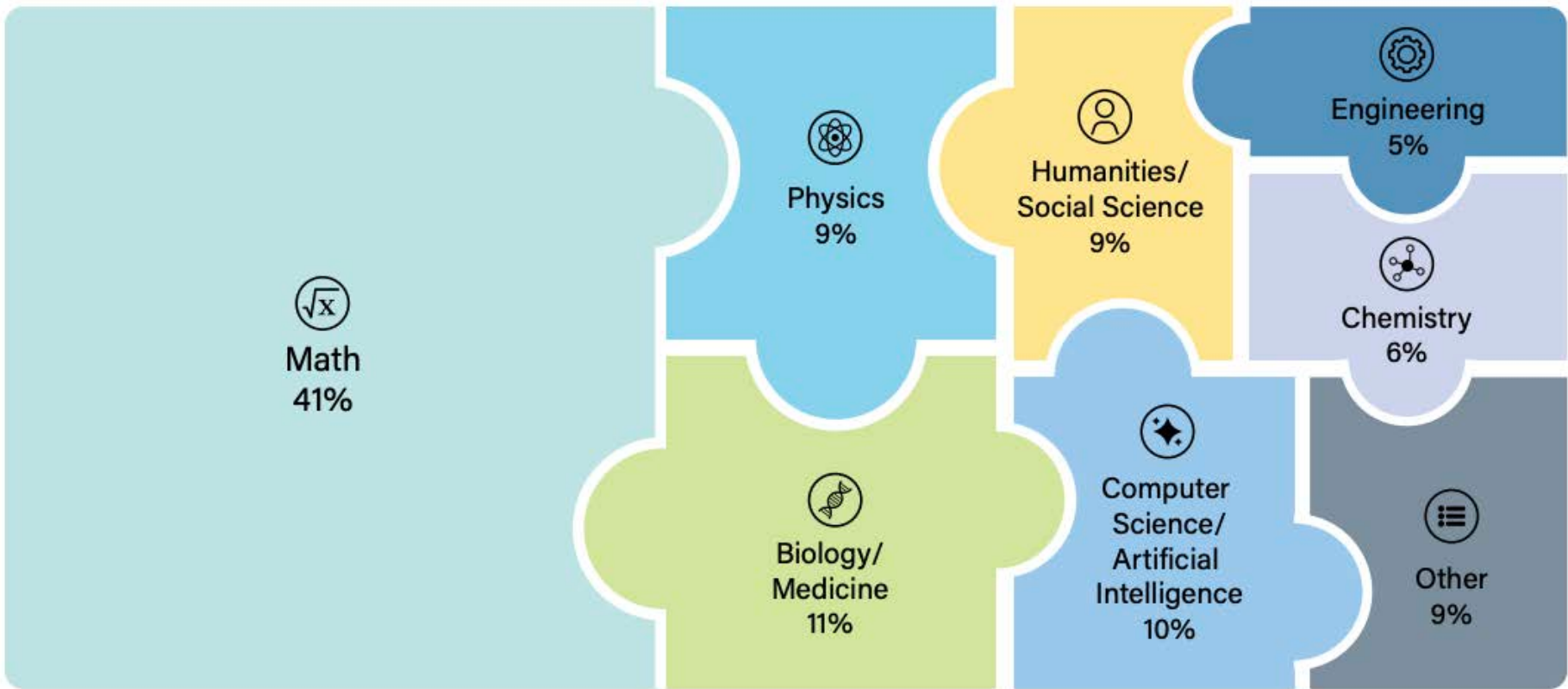
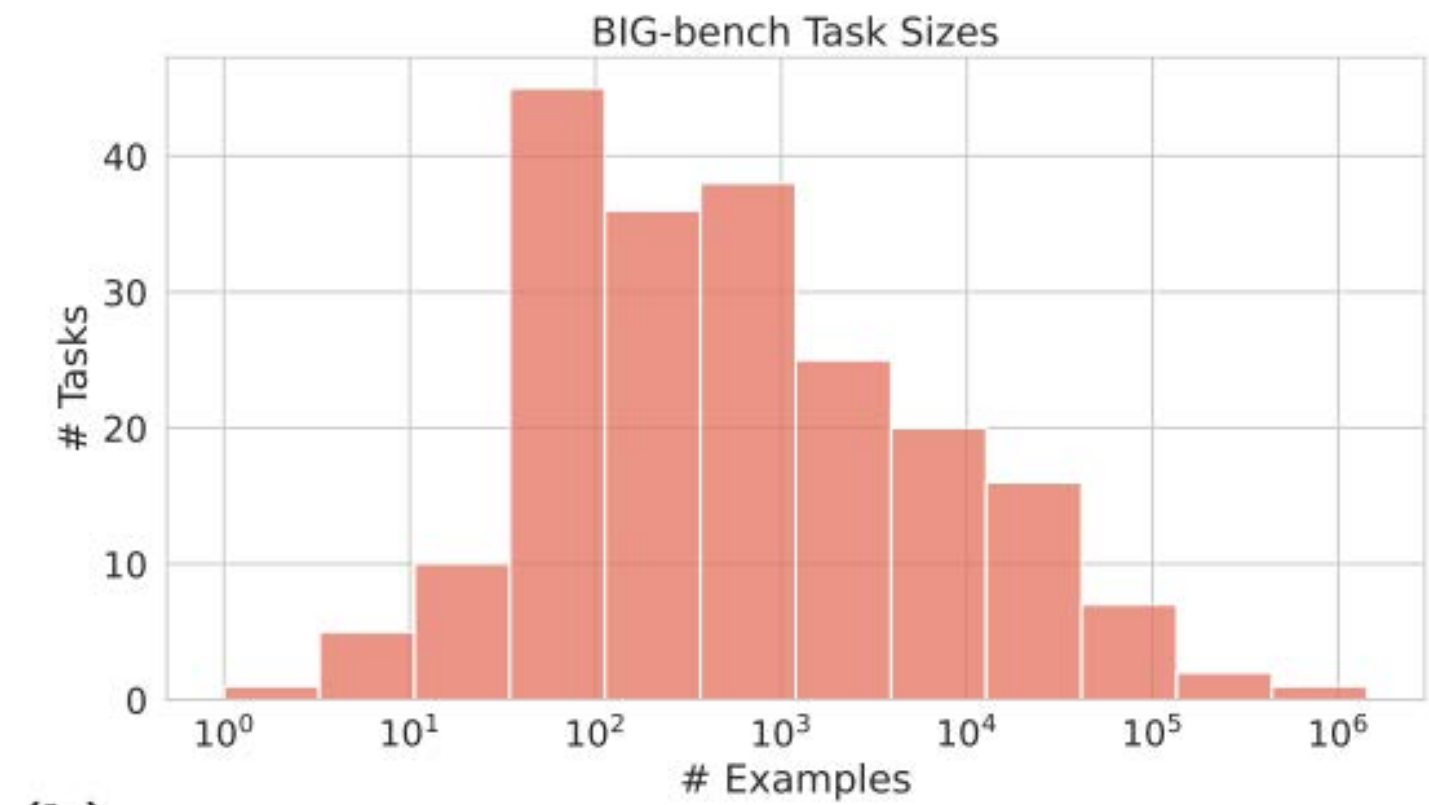


Figure 3: HLE consists of 2,700 exam questions in over a hundred subjects, grouped into high level categories here. We provide a more detailed list of subjects in Appendix B.3.

Humanity’s last exam



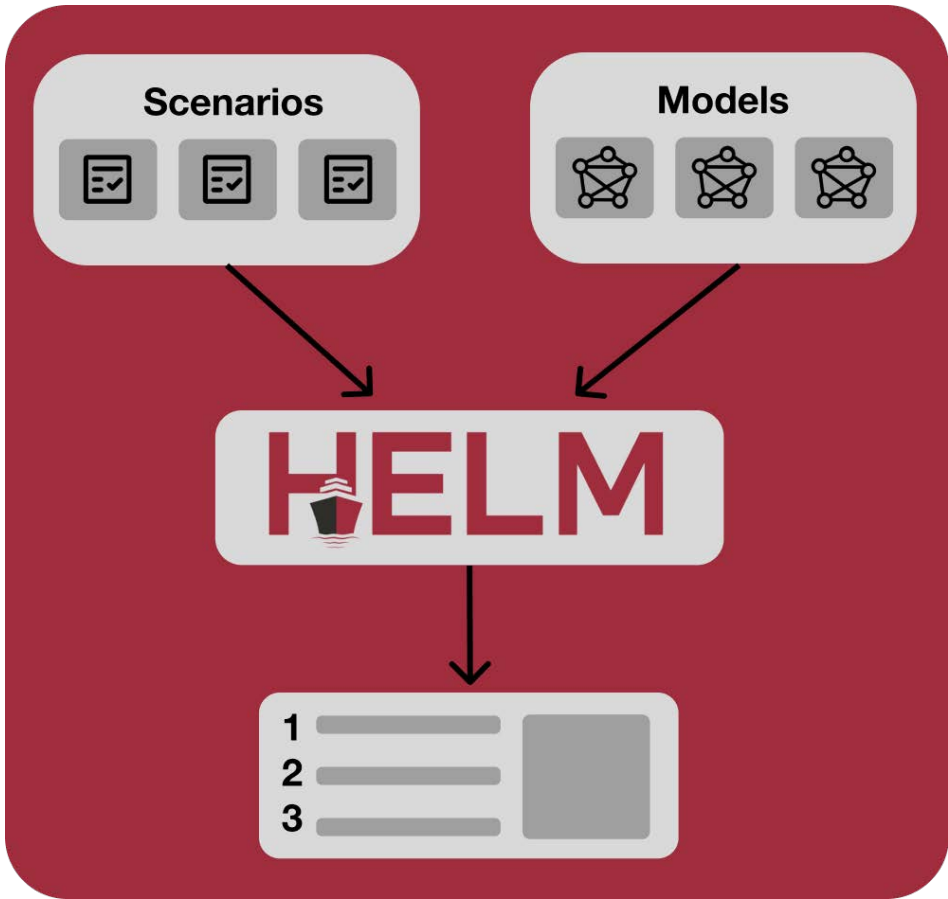
BIG-Bench



FrontierMath

A math benchmark testing the limits of AI

In collaboration with OpenAI: [Learn more](#)





Open LLM Leaderboard Archived

Comparing Large Language Models in an open and reproducible way

1956 - Rosenblatt's Perceptron
1959 - Highleyman's Data creation
1962 - Novikoff Mistake bound
1962 - Highleyman on train-test split
1968 - First Issue of Pattern Recognition
1973 - Duda and Hart 1st Edition
1974 - Vapnik and Chervonenkis 1st book
1979 - First Issue of IEEE PAMI
1980 - First ICML Workshop
1984 - Classification and Decision Trees
1985 - RFC 595 FTP released
1985 - Invention of CD ROM
1986 - Snowbird workshop proposes NIPS
1986 - Elm released
1986 - First issue of Machine Learning
1986 - DARPA program creates TIMIT
1987 - First NIPS conference
1987 - UCI ftp site launched
1988 - TIMIT released
1989 - First issue of Neural Computation
1989 - Pine client released
1990 - TIMIT released on CD-ROM
1994 - MNIST creation
1998 - MNIST release (i.e., the singularity)

We use the same algorithms and same theory.

All that has changed: our computers are faster
and our data is larger.

That this was necessary was predicted before
1970 but took 50 years to come to fruition.

What does the next 50 years hold for
machine learning?