

Data analysis and visualization (DAV)

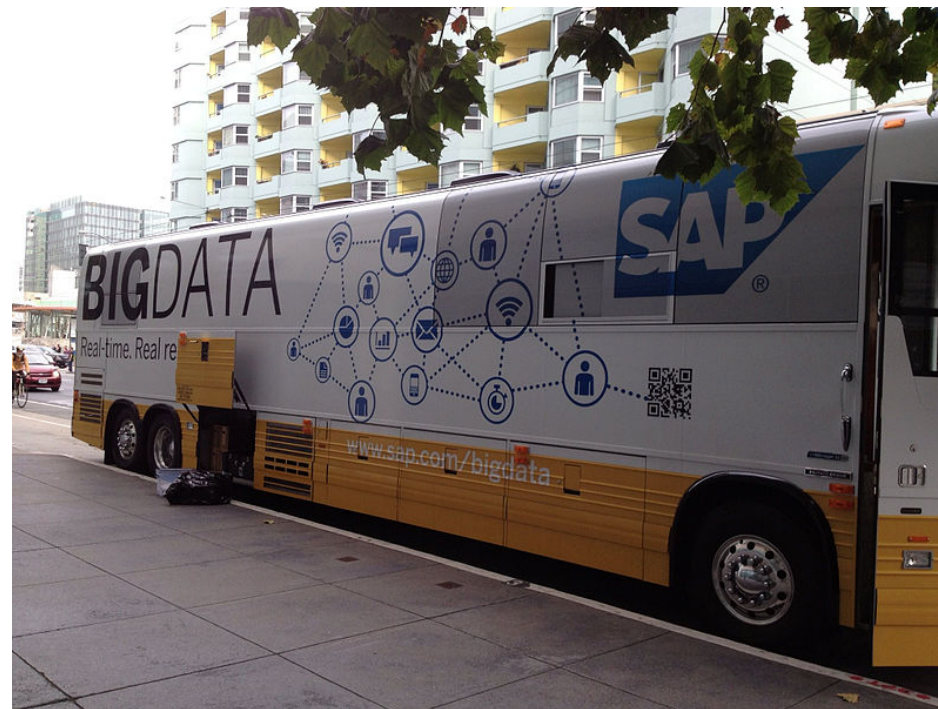
Lecture 08

Łukasz P. Kozłowski

Warsaw, 2025

Data analysis is a process of inspecting, cleansing, transforming and modeling data with the goal of discovering useful information, informing conclusion and supporting decision-making.

Wikipedia



BIGDATA

Datasets (classic examples)

In few slides we will list
the most commonly used datasets
(subjective and incomplete list)

Datasets (classic examples)

Iris dataset



Iris setosa



Iris versicolor



Iris virginica

Datasets (classic examples)

Iris dataset - contains a set of 150 records under five attributes: petal length, petal width, sepal length, sepal width and species.

Dataset Order ↕	Sepal length ↕	Sepal width ↕	Petal length ↕	Petal width ↕	Species ↕
1	5.1	3.5	1.4	0.2	<i>I. setosa</i>
2	4.9	3.0	1.4	0.2	<i>I. setosa</i>
3	4.7	3.2	1.3	0.2	<i>I. setosa</i>

51	7.0	3.2	4.7	1.4	<i>I. versicolor</i>
52	6.4	3.2	4.5	1.5	<i>I. versicolor</i>
53	6.9	3.1	4.9	1.5	<i>I. versicolor</i>

100	5.7	2.8	4.1	1.3	<i>I. versicolor</i>
101	6.3	3.3	6.0	2.5	<i>I. virginica</i>
102	5.8	2.7	5.1	1.9	<i>I. virginica</i>
103	7.1	3.0	5.9	2.1	<i>I. virginica</i>



Iris setosa



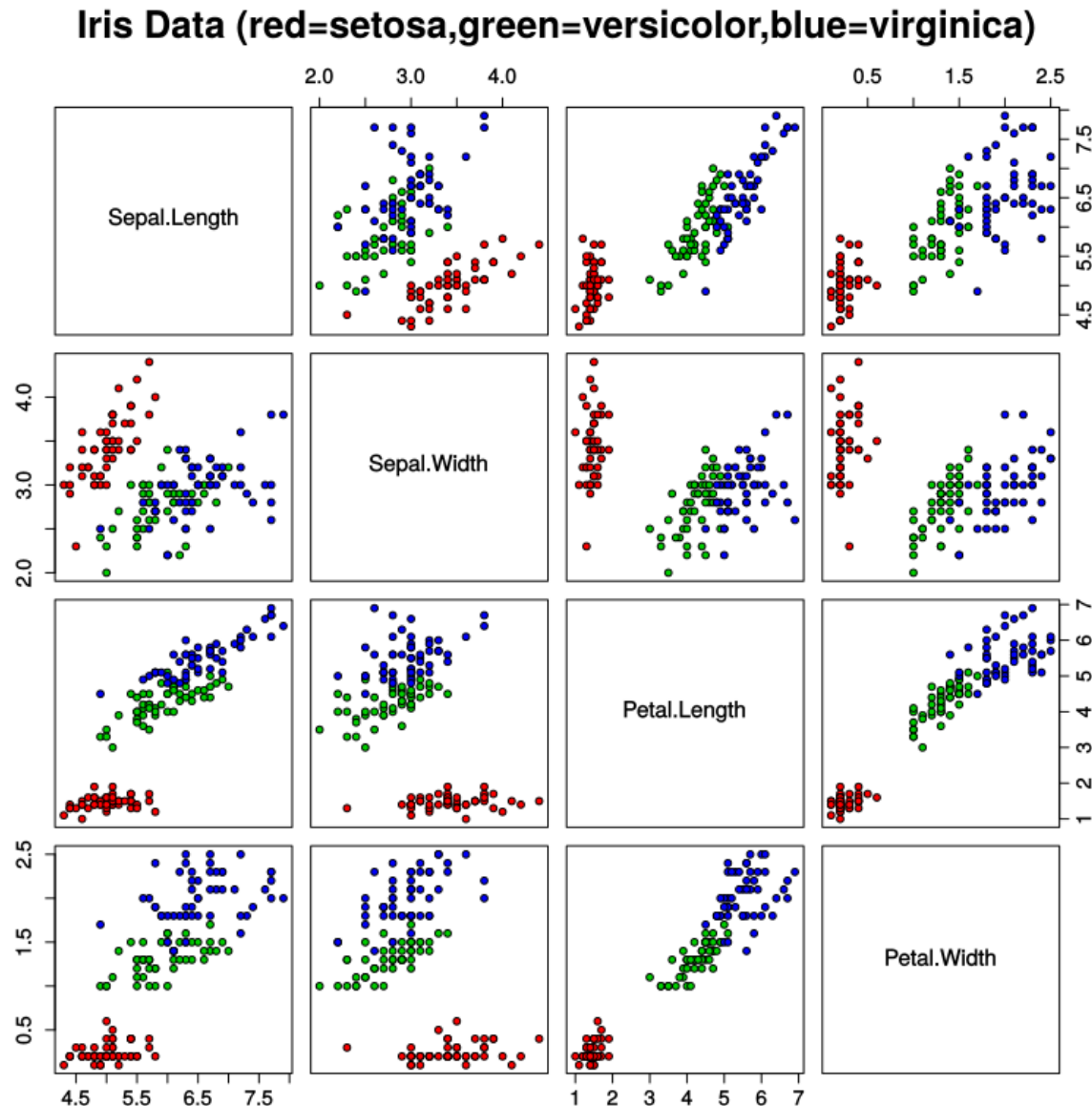
Iris versicolor



Iris virginica

Datasets (classic examples)

Iris dataset is also a classic dataset from which all tutorials in statistics & machine learning starts



Iris setosa



Iris versicolor



Iris virginica

Datasets (classic examples)

Iris dataset is also a classic dataset from which all tutorials in statistics & machine learning starts

```
from sklearn.datasets import load_iris  
  
iris = load_iris()  
iris
```

This code gives:

```
{'data': array([[5.1, 3.5, 1.4, 0.2],  
                [4.9, 3. , 1.4, 0.2],  
                [4.7, 3.2, 1.3, 0.2],  
                [4.6, 3.1, 1.5, 0.2],
```

You will find it in all statistics and ML libraries in any programming language



Iris setosa



Iris versicolor



Iris virginica

Datasets (classic examples)

Wine Dataset: 178 samples, 13 chemical features (e.g., alcohol, malic acid), 3 classes (wine types). Ideal for multiclass SVM testing due to its numerical features and moderate size. Available via UCI Machine Learning Repository or scikit-learn.

Breast Cancer Wisconsin Dataset: 569 samples, 30 features (e.g., cell nucleus measurements), 2 classes (benign/malignant). Great for binary SVM classification with real-world medical data. Available via UCI or scikit-learn.

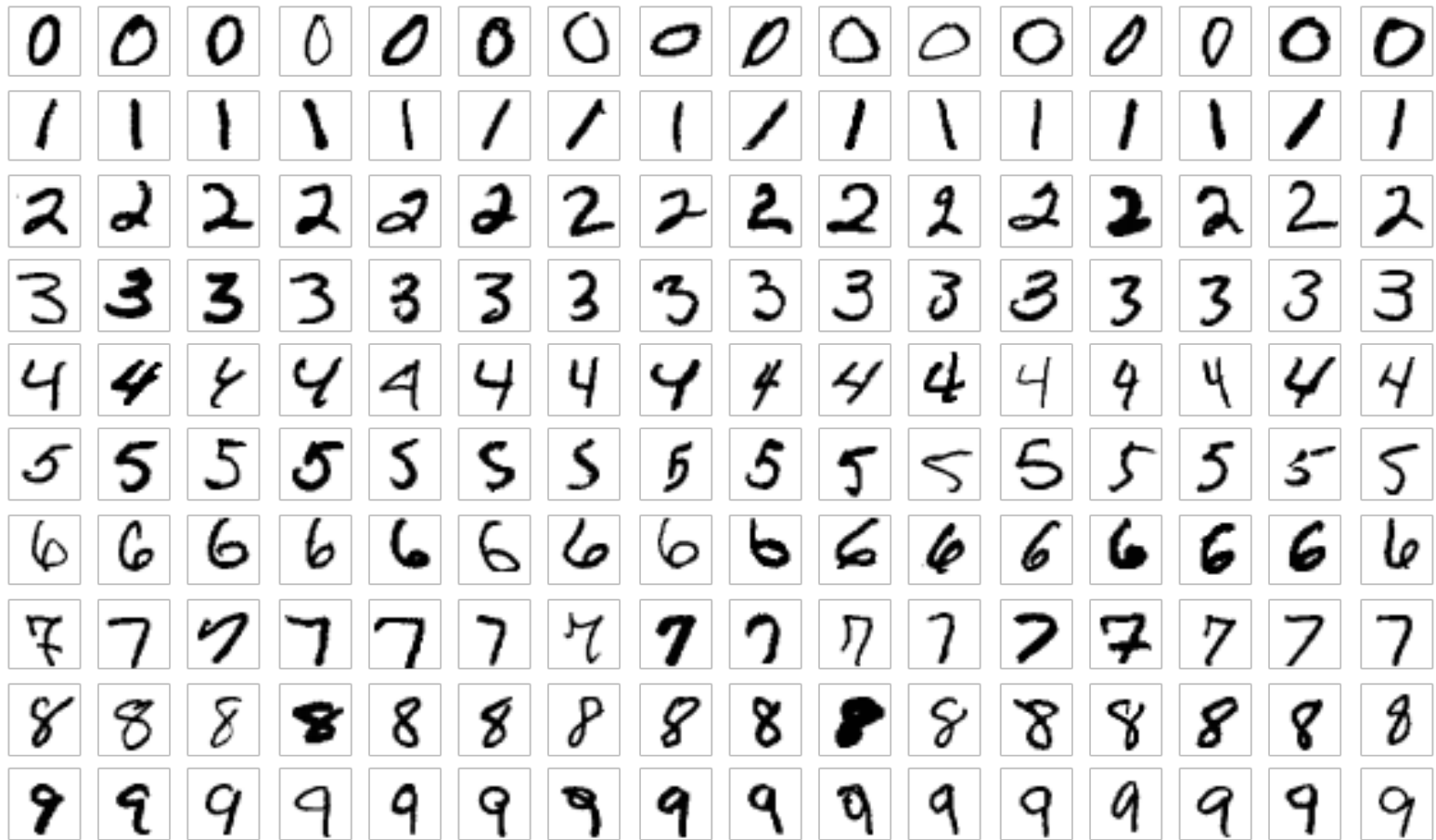
Digits Dataset: 1,797 samples, 64 features (8x8 pixel values), 10 classes (digits 0-9). A step up in complexity from Iris, good for multiclass SVM and image-based classification. Available via scikit-learn.

Sonar Dataset: 208 samples, 60 features (sonar signal returns), 2 classes (rock/metal). Small but high-dimensional, perfect for testing SVM's ability to handle feature-rich data. Available via UCI.

Glass Identification Dataset: 214 samples, 9 features (chemical composition), 6 classes (glass types). Suitable for multiclass SVM with a small but challenging dataset. Available via UCI.

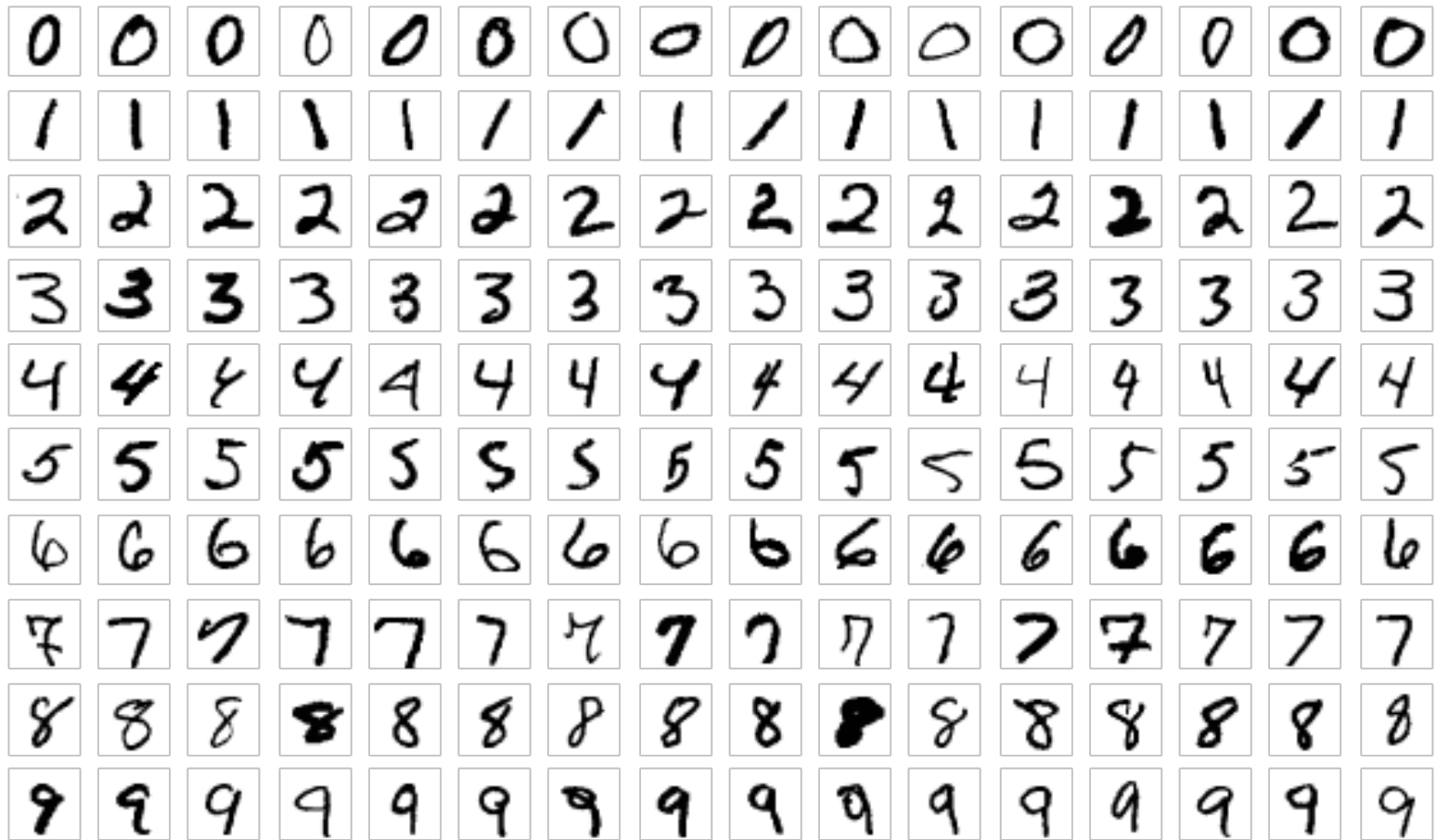
Datasets (classic examples)

MNIST dataset - handwritten digits



Datasets (classic examples)

MNIST dataset - handwritten digits



60,000 training images and 10,000 testing images

28x28 pixel bounding box and anti-aliased, and grayscale levels

Datasets (classic examples)

MNIST dataset - handwritten digits

Type ↕	Classifier	Error rate (%) ↕
Linear classifier	Pairwise linear classifier	7.6 ^[8]
Decision stream with Extremely randomized trees	Single model (depth > 400 levels)	2.7 ^[20]
K-Nearest Neighbors	K-NN with non-linear deformation (P2DHMDM)	0.52 ^[21]
Boosted Stumps	Product of stumps on Haar features	0.87 ^[22]
Non-linear classifier	40 PCA + quadratic classifier	3.3 ^[8]
Random Forest	Fast Unified Random Forests for Survival, Regression, and Classification (RF-SRC) ^l	2.8 ^[24]
Support-vector machine (SVM)	Virtual SVM, deg-9 poly, 2-pixel jittered	0.56 ^[25]
Deep neural network (DNN)	2-layer 784-800-10	1.6 ^[26]
Deep neural network	2-layer 784-800-10	0.7 ^[26]
Deep neural network	6-layer 784-2500-2000-1500-1000-500-10	0.35 ^[27]
Convolutional neural network (CNN)	6-layer 784-40-80-500-1000-2000-10	0.31 ^[15]
Convolutional neural network	6-layer 784-50-100-500-1000-10-10	0.27 ^[28]
Convolutional neural network	Committee of 35 CNNs, 1-20-P-40-P-150-10	0.23 ^[10]
Convolutional neural network	Committee of 5 CNNs, 6-layer 784-50-100-500-1000-10-10	0.21 ^{[17][18]}
Random Multimodel Deep Learning (RMDL)	10 NN-10 RNN - 10 CNN	0.18 ^[19]
Convolutional neural network	Committee of 20 CNNs with Squeeze-and-Excitation Networks ^[29]	0.17 ^[30]

Commonly used for more serious ML algorithms evaluation

Default dataset for deep learning

Datasets (classic examples)

MNIST dataset - handwritten digits

Type ↕	Classifier	Error rate (%) ↕
Linear classifier	Pairwise linear classifier	7.6 ^[8]
Decision stream with Extremely randomized trees	Single model (depth > 400 levels)	2.7 ^[20]
K-Nearest Neighbors	K-NN with non-linear deformation (P2DHMDM)	0.52 ^[21]
Boosted Stumps	Product of stumps on Haar features	0.87 ^[22]
Non-linear classifier	40 PCA + quadratic classifier	3.3 ^[8]
Random Forest	Fast Unified Random Forests for Survival, Regression, and Classification (RF-SRC) ^l	2.8 ^[24]
Support-vector machine (SVM)	Virtual SVM, deg-9 poly, 2-pixel jittered	0.56 ^[25]
Deep neural network (DNN)	2-layer 784-800-10	1.6 ^[26]
Deep neural network	2-layer 784-800-10	0.7 ^[26]
Deep neural network	6-layer 784-2500-2000-1500-1000-500-10	0.35 ^[27]
Convolutional neural network (CNN)	6-layer 784-40-80-500-1000-2000-10	0.31 ^[15]
Convolutional neural network	6-layer 784-50-100-500-1000-10-10	0.27 ^[28]
Convolutional neural network	Committee of 35 CNNs, 1-20-P-40-P-150-10	0.23 ^[10]
Convolutional neural network	Committee of 5 CNNs, 6-layer 784-50-100-500-1000-10-10	0.21 ^{[17][18]}
Random Multimodel Deep Learning (RMDL)	10 NN-10 RNN - 10 CNN	0.18 ^[19]
Convolutional neural network	Committee of 20 CNNs with Squeeze-and-Excitation Networks ^[29]	0.17 ^[30]

Commonly used for more serious ML algorithms evaluation

Default dataset for deep learning

Datasets (classic examples)

MNIST-like dataset

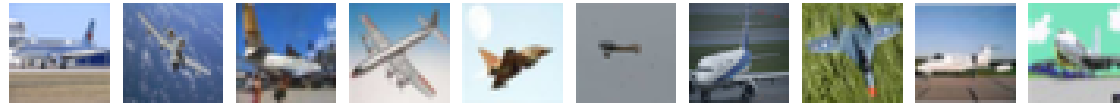
- **EMNIST: 814,255 images, handwritten digits & letters, 28x28 grayscale. Subsets: Digits (280K), Letters (145K). Via NIST/TensorFlow**
- **Kuzushiji-MNIST: 70,000 Japanese Kuzushiji characters, 28x28 grayscale. Larger sets: Kuzushiji-49 (270K, 49 classes), Kuzushiji-Kanji (140K, 3,563 classes). On GitHub**
- **IAM Handwriting Database: 657,000+ word images, English, variable size, grayscale. For text recognition. Via University of Bern**
- **CASIA-HWDB: ~3.9M Chinese character images, 7,356 classes, grayscale, variable size. ~1.2M isolated characters. Via Chinese Academy of Sciences**
- **Omniglot: 1.6M+ character images, 50 alphabets, 105x105 grayscale (often 28x28). Via GitHub.**

Datasets (classic examples)

CIFAR-10 & CIFAR-100

The CIFAR-10 dataset consists of 60,000 32x32 colour images in 10 classes, with 6000 images per class. There are 50,000 training images and 10,000 test images

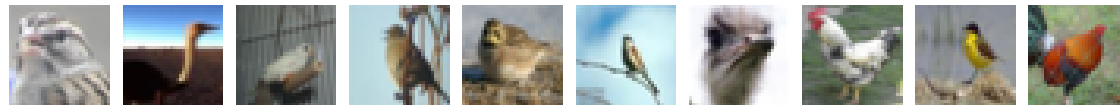
airplane



automobile



bird



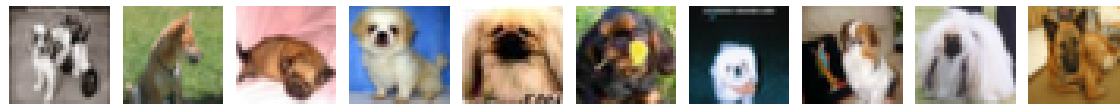
cat



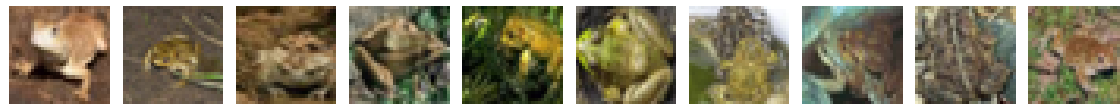
deer



dog



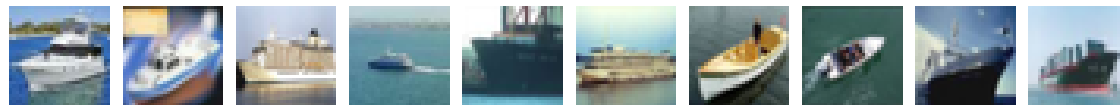
frog



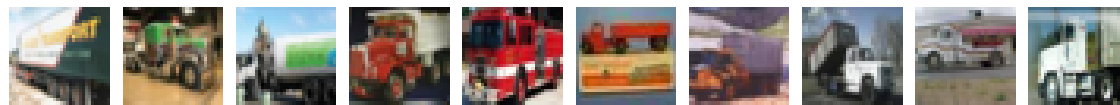
horse



ship

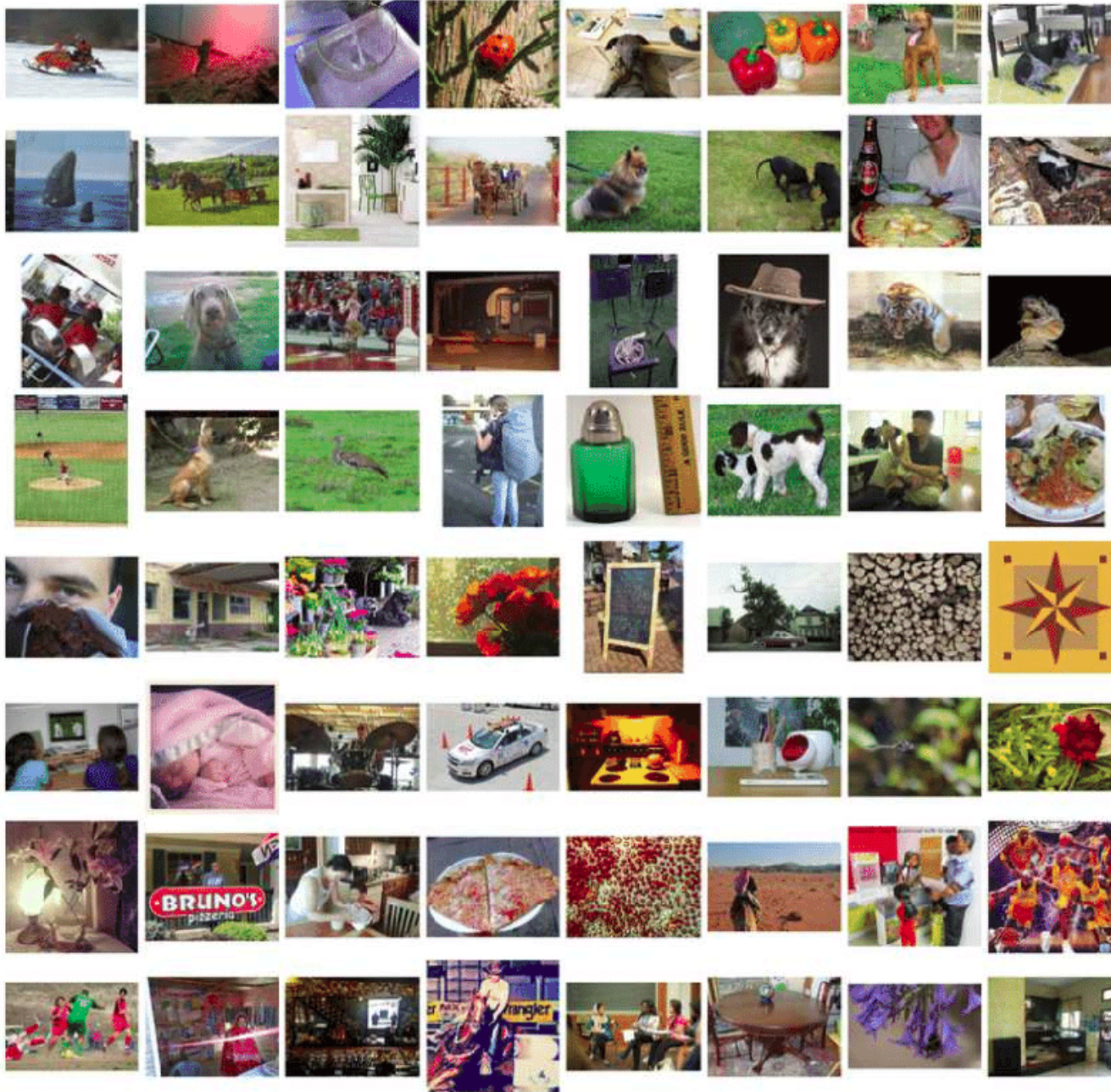


truck



Datasets (classic examples)

ImageNet – classic deep learning test dataset



Licensing: ImageNet does not own the copyright of the images, it only compiles an accurate list of web images for each synset of WordNet. For this reason, ImageNet is available for use under terms that facilitate both **academic** and **non-commercial research**, with specific [guidelines for usage and attribution](#).

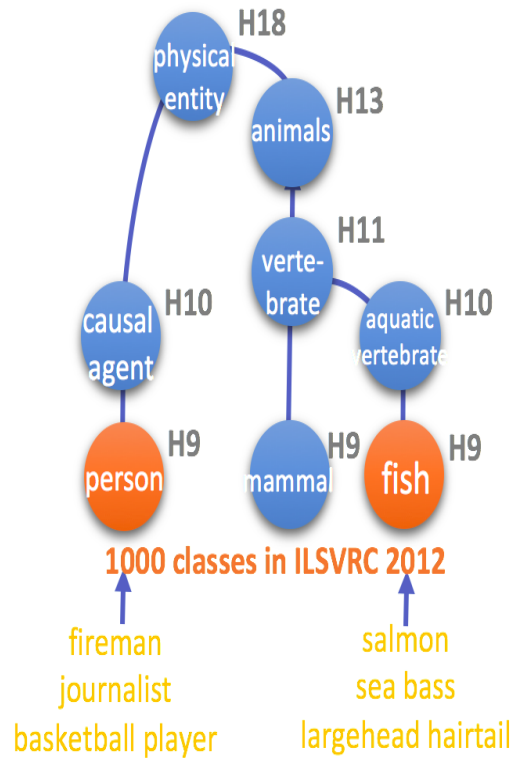
Official Site: <http://www.image-net.org/>



Datasets (classic examples)

ImageNet – classic deep learning test dataset

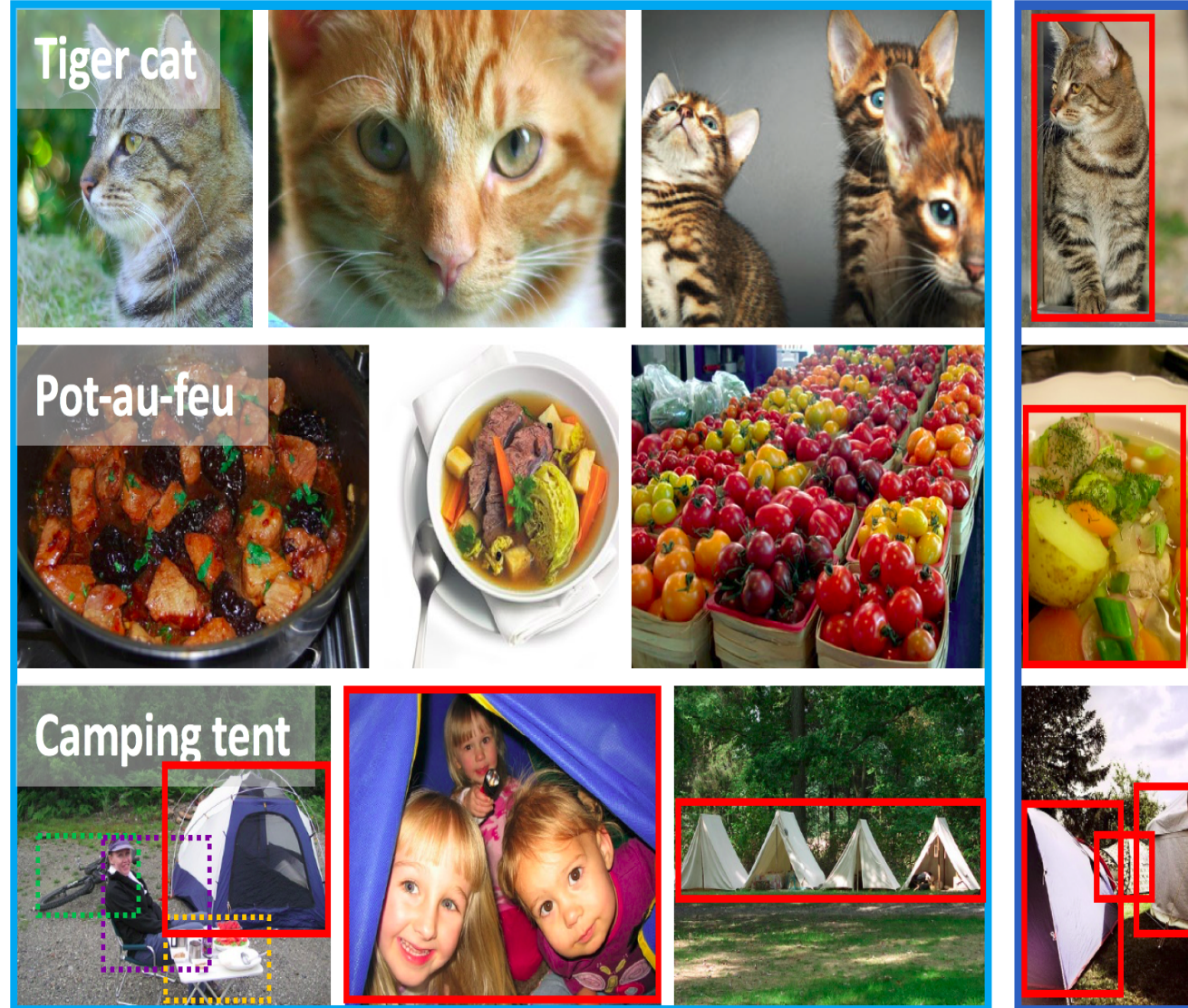
WordNet Class Hierarchy



The extended classes beyond ILSVRC 2012

-  Ground truth (provided or annotated)
-  Training set (2686 classes)
-  Test set (1/10 amount of training)

Extended ImageNet Classification Dataset (EIC)

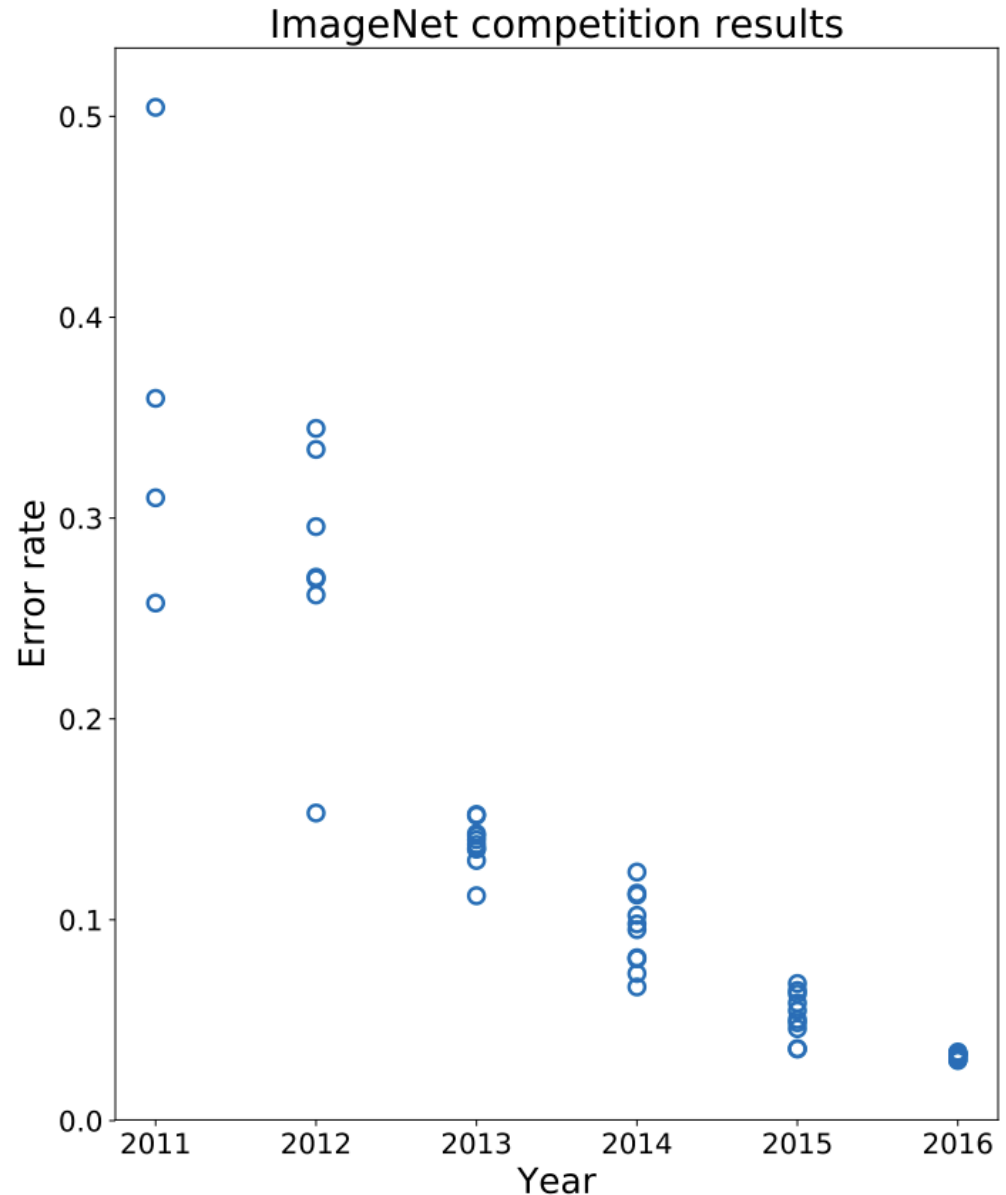


Datasets (classic examples)

ImageNet – classic deep learning test dataset

More than 14 million images have been hand-annotated to indicate what objects are pictured and in at least one million of the images, bounding boxes are also provided. ImageNet contains more than 20,000 categories.

Current algorithm have <5% error rate



LAION-5B: Massive Vision-Language Dataset

- **~5.85B image-text pairs for multimodal AI training**
- **Split: 2.32B English, 2.26B multilingual (100+ languages), 1.27B language-agnostic**
- **Sourced from Common Crawl web archive**
- **Text from alt-text, captions, and web content**
- **Emphasizes diversity in visuals and concepts**

Access Metadata: Download from LAION's official site
(<https://laion.ai/laion-5b/>) or Hugging Face (laion/laion5b dataset)



Datasets (classic examples)

WATERMARK_STRICT annotation guide

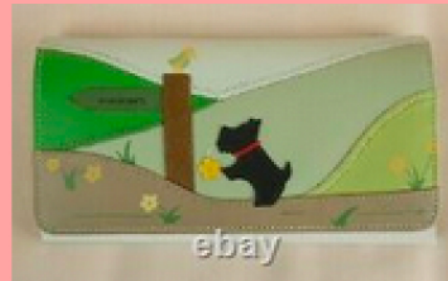
Classic watermark



Advertisement and watermark



Brand text covering image, watermark looks



No trademark/brand, but "100%" watermark characteristics



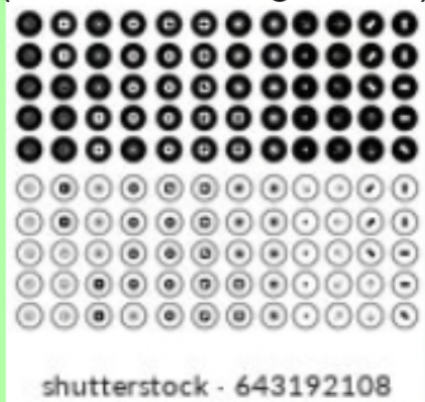
WATERMARK

WATERMARK

WATERMARK

WATERMARK

Text not covering image (also containing ®, TM)



Subtle text not covering main parts of the image



Logo



Advertisement



NO WATERMARK NO WATERMARK NO WATERMARK NO WATERMARK

Datasets (classic examples)

IMDB Reviews

A binary sentiment classification containing substantially more data than previous benchmark datasets. We provide a set of **25,000** highly polar movie reviews for training, and **25,000** for testing.

The movie poster for "The Matrix" (1999), showing Keanu Reeves, Laurence Fishburne, and other cast members in a dark, futuristic setting.

The Matrix (1999) Top 500

R 136 min - [Action](#) | [Adventure](#) | [Sci-Fi](#) - [31 March 1999 \(USA\)](#)

★★★★★ 8.7/10

Users: (448,475 votes) 3,342 reviews **Critics: 248 reviews**

Metascore: **73**/100 (based on 35 reviews from Metacritic.com)

A computer hacker learns from mysterious rebels about the true nature of his reality and his role in the war against its controllers.

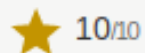
Directors: [Andy Wachowski](#), [Lana Wachowski](#)

Writers: [Andy Wachowski](#), [Lana Wachowski](#)

Stars: [Keanu Reeves](#), [Laurence Fishburne](#) and [Carrie-Anne](#)



sentiment

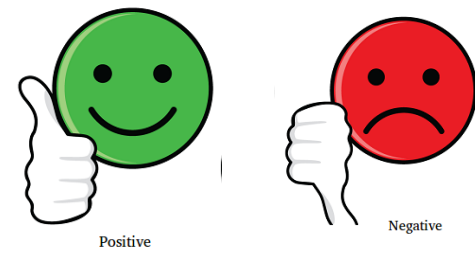


The benchmark for all sci-fi films to come

[mambubukid](#) 19 September 2000

The story of a reluctant Christ-like protagonist set against a baroque, MTV backdrop, The Matrix is the definitive hybrid of technical wizardry and contextual excellence that should be the benchmark for all sci-fi films to come.

Datasets (classic examples)



IMDB like datasets

Yelp Reviews: ~8.5M reviews of businesses, rated 1-5 stars. Great for multi-class sentiment analysis or regression. Available via Yelp's dataset page.

Amazon Reviews: ~83M product reviews (1996-2018), with ratings 1-5 stars, across categories like electronics, books, etc. Ideal for large-scale sentiment or topic modeling. Accessible via Stanford SNAP or UCSD archives.

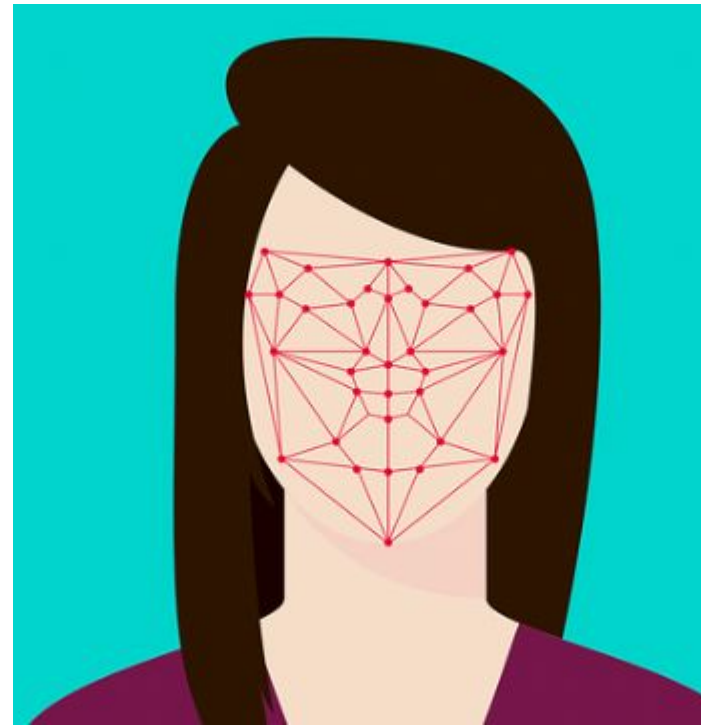
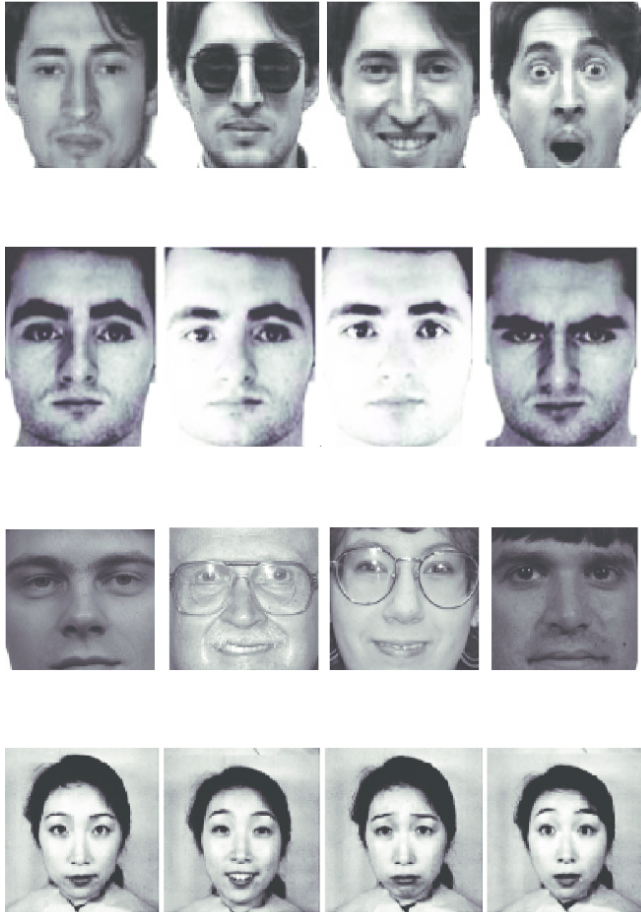
Twitter Sentiment140: ~1.6M tweets, labeled for sentiment (positive/negative/neutral). Focused on short-text sentiment analysis, unlike IMDB's longer reviews. Available via Sentiment140 website.

Reddit Comments (Pushshift): ~1B comments (2005-2023), often used for sentiment or topic analysis. No direct labels, but can be processed for sentiment. Available via Pushshift archives or BigQuery.

TripAdvisor Reviews: ~1.3M hotel reviews, with ratings 1-5 stars. Useful for sentiment and aspect-based analysis (e.g., service, location). Accessible via Kaggle or Webis group.

Datasets (classic examples)

FERET (facial recognition technology)



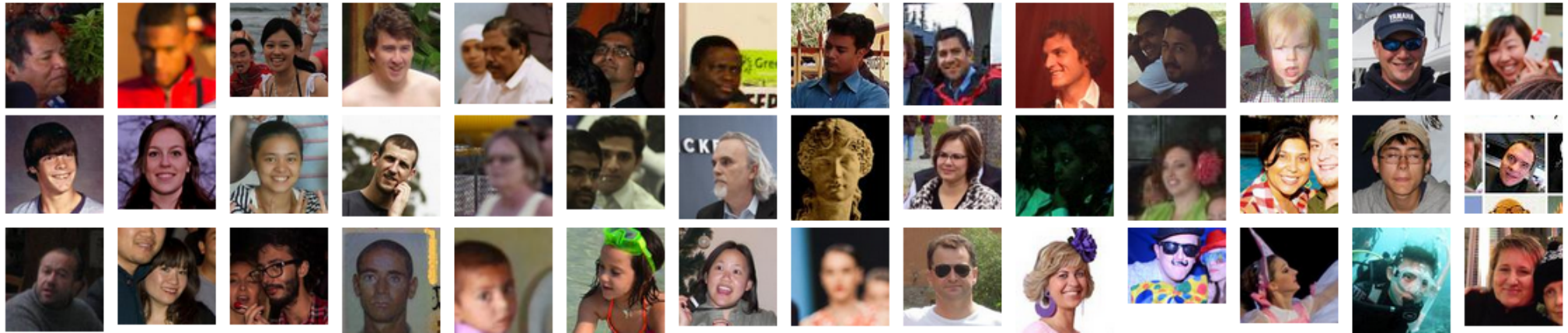
FACE RECOGNITION

11338 images of 1199 individuals in different positions and at different times

Datasets (classic examples)

MegaFace and MF2: Million-Scale Face Recognition

The MegaFace challenge has concluded, reaching a benchmark performance of over 99%. Because its goals have been met, and ongoing maintenance of this platform would require considerable administrative effort, MegaFace is being decommissioned and MegaFace data are no longer being distributed.



Distractors

1 Million Photos

690,572 Unique Users

Training Set

4.7 Million Photos

672,057 Unique Identities

7 Mean photos / person (3 min, 2469 max)

Test Sets

FaceScrub Celebrities

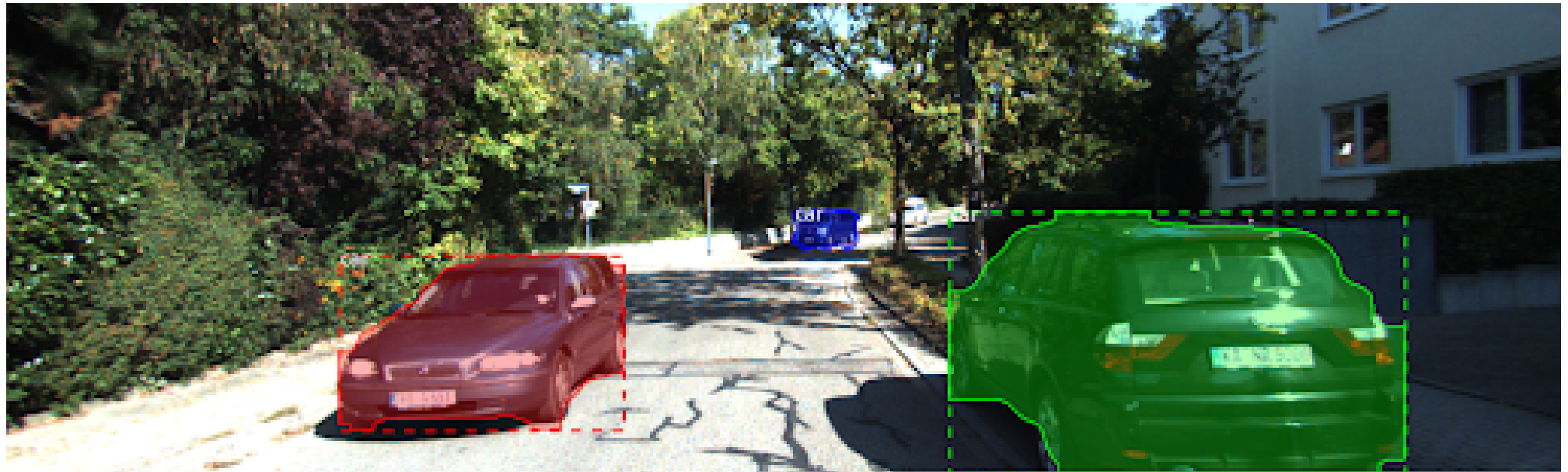
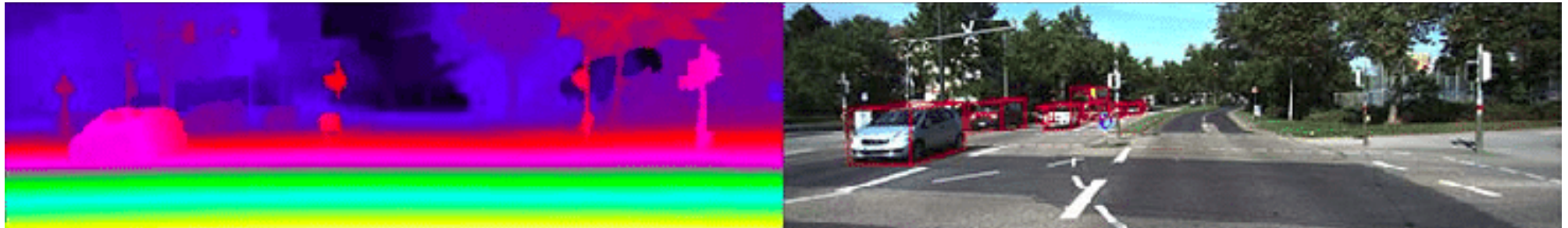
FGNet Age-invariant non-celebrities

More appropriate for deep learning

Datasets (classic examples)

KITTI Vision Benchmark Dataset

>100 GB of data from Autonomous vehicles driving through a mid-size city
captured images of various areas using cameras and laser scanners



Datasets (classic examples)

Youtube-8M

Large-scale video dataset that was announced in Sept 2016 by Google group. The labels per video are organized into 24 top-level verticals.

6.1 Million
Video IDs

350,000
Hours of Video

2.6 Billion
Audio/Visual Features

3862
Classes

3.0
Avg. Labels / Video

YouTube | 8M

Dataset

Explore

Download

Workshop

About

Vertical

All

Filter

Entities

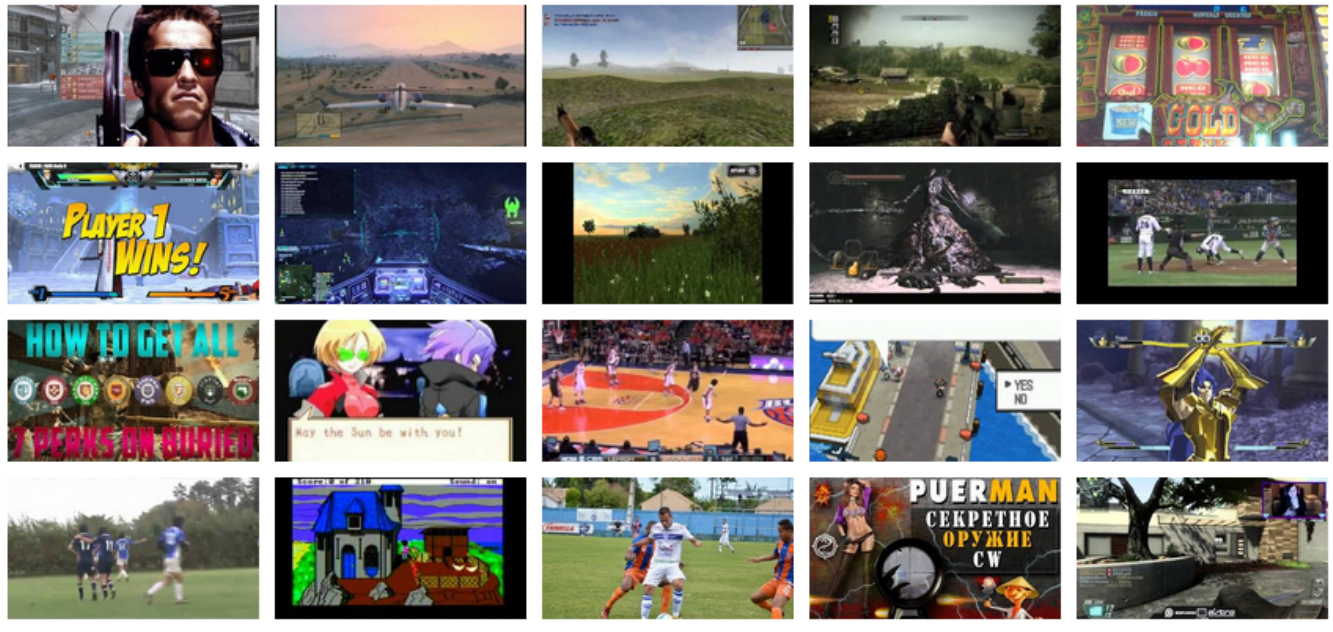
Games (788288) Video game (539945)

Vehicle (415890) Concert (378135)

Musician (286532) Cartoon (236948)

Performance art (203343) Car (200813)

Dance (181579) Guitar (156226)



Datasets (classic examples)

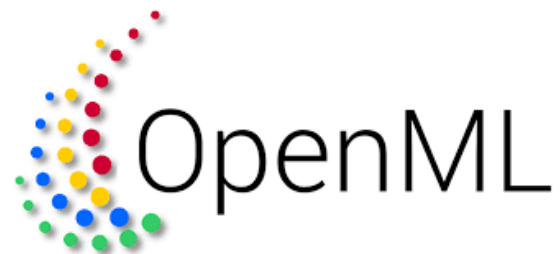
For many standard tasks/areas there is plenty of datasets to try/use

Many new datasets are generated each day

Yet, datasets for many specific areas are missing and very expensive to generate (e.g. medicine related data like types of the cancer using specific medical equipment)

Check:

https://en.wikipedia.org/wiki/List_of_datasets_for_machine-learning_research





ChatGPT



LLM

ChatGPT



The 72 tokens

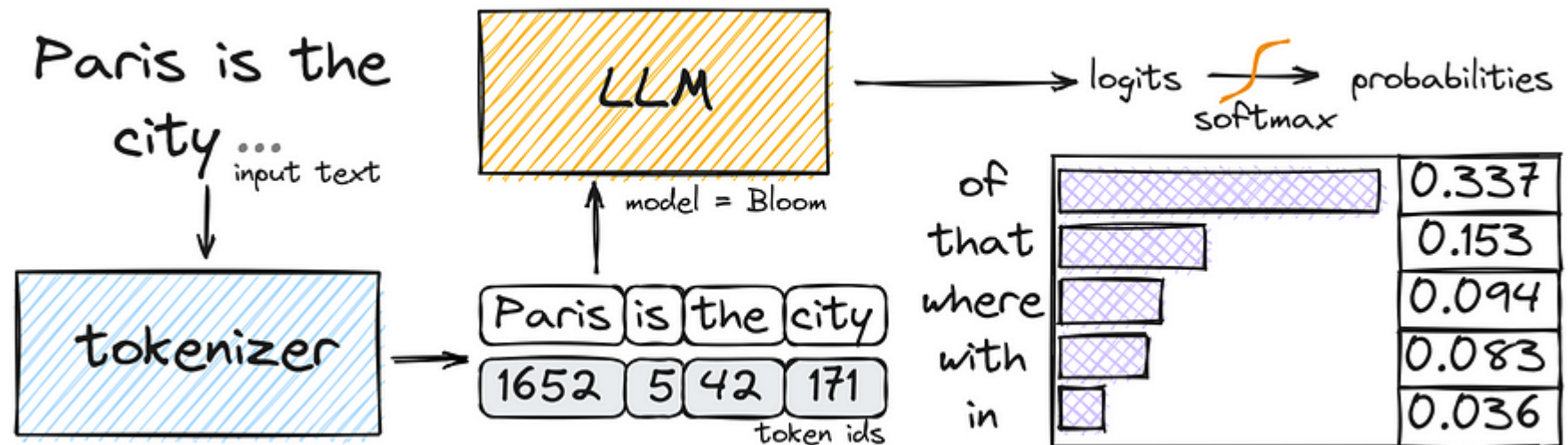
This tool allows you to visualize the tokens of a text prompt or tokenization models of the various Google Cloud Vertex AI PaLM - are also counted, and hovering over them will indicate their internal code of this application is available on GitHub.

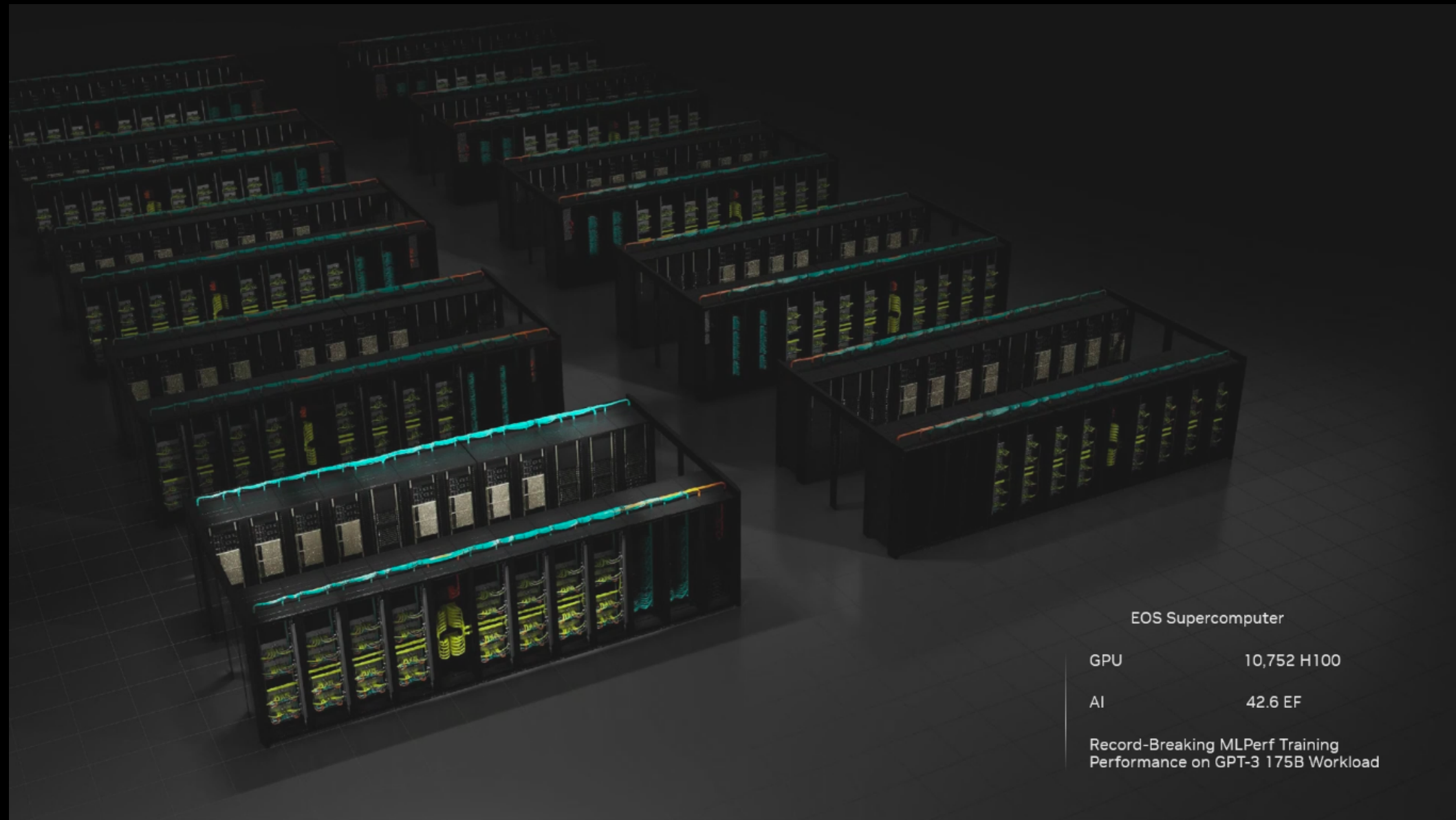
ChatGPT



The 72 tokens

This tool allows you to visualize the tokens of a text prompt or tokenization models of the various Google Cloud Vertex AI PaLM - are also counted, and hovering over them will indicate their internal code of this application is available on GitHub.





EOS Supercomputer

GPU 10,752 H100

AI 42.6 EF

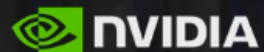
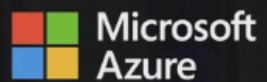
Record-Breaking MLPerf Training
Performance on GPT-3 175B Workload



Azure ND H100 V5 Supercomputer

10,752 H100 GPUs w/ 400Gb/s IB

Record-Breaking MLPerf Training
Performance on GPT-3 175B
Workload Running on 10,000+ GPUs



[🏠](#) > [Części PC](#) > [Karty graficzne](#) > [Serwer GPU](#) > [NVIDIA Hopper](#) > [NVIDIA HGX H100 640GB - 935-24287-0001-000](#)

NVIDIA HGX H100 640GB - 935-24287-0001-000



Graphics Engine: Hopper Memory size: 640 GB Number of tensor cores: 5 performance: 268 TFLOP

Wszystkie u GPU architektury NVIDIA Ampere, NVIDIA Grace Hopper, NVIDIA Ada Lovelace i podlegają **NCNR (Non-Cancellable, Non-Returnable) wynoszącemu 52 tygodnie**. Ponadto sankcjom dla niektórych krajów i konieczne jest udokumentowanie końcowego klienta.

Możemy dostarczyć te karty GPU bezpośrednio i z indywidualną ceną B2B **się z nami i zapytaj o cenę.**

Kod towaru	214.166385
Part number	935-24287-0001-000
Supermicro Part No.	GPU-NVHGX-H100-LC-88-EX2
Producent	NVIDIA
Dostępność	Na zamówienie
Gwarancja	36 miesięcy
Waga	2 kg
Cena zawiera wszystkie obowiązujące opłaty	

1 130 107 PLN bez VAT
1 390 032 PLN z VAT

1 szt.



Dodaj do koszyka



Ustaw watchdoga



Dodaj do porównania produktów



Kliknij tutaj, aby...



COLOSSUS



■ CONSTRUCTION TIME
122 DAYS

NUMBER OF GPUS

TOTAL MEMORY BANDWIDTH

NETWORK BANDWIDTH PER SERVER

STORAGE CAPACITY

200K GPUs

194 Petabytes/s

3.6 Terabits/s

>1 Exabyte

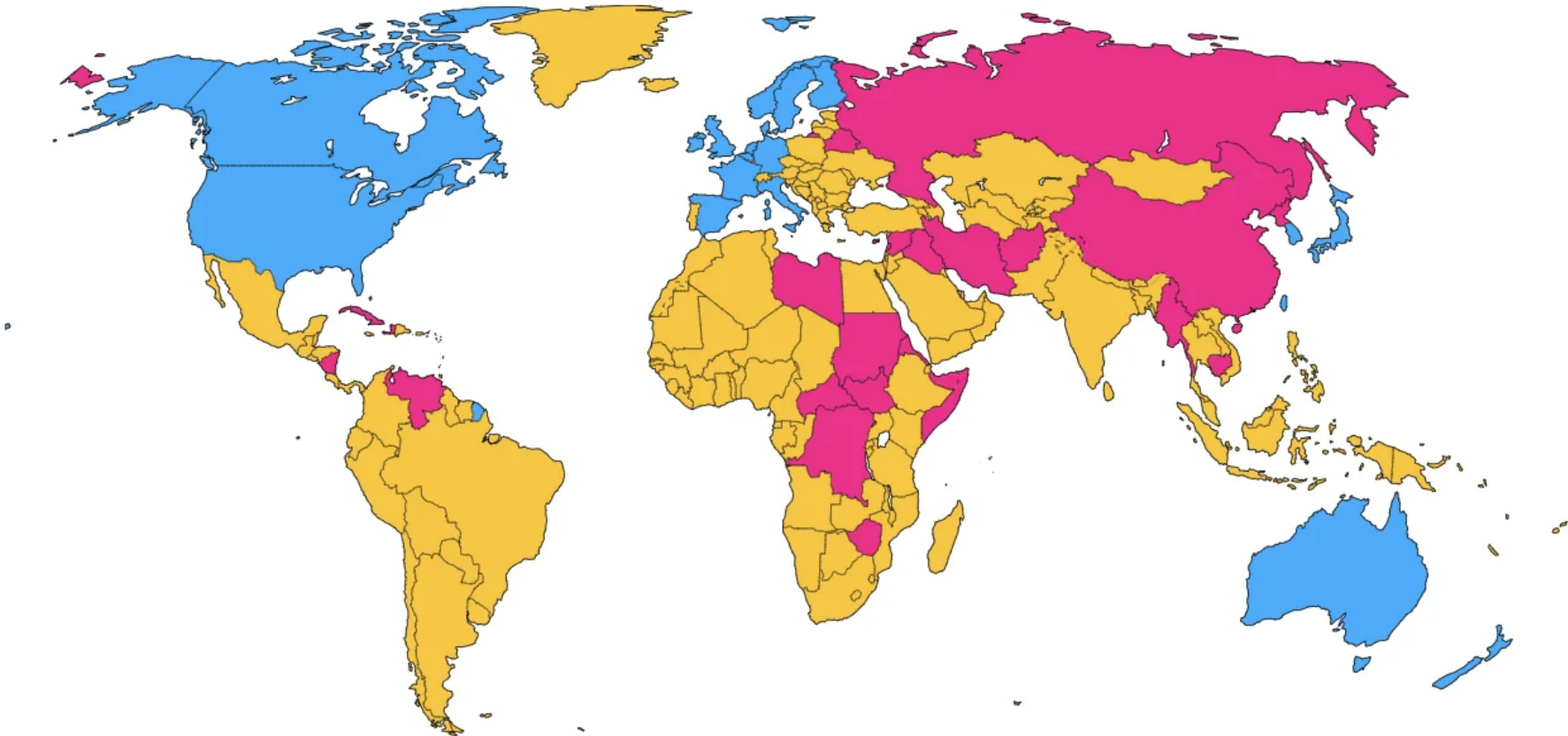
Uproar as Biden Administration Limits AI Chip Shipments

By Sally Ward-Foxton 01.15.2025 0

US to Curb Global Chip Shipments

Most markets will face new restrictions on data center development

■ Tier 1 (Most permissive) ■ Tier 2 ■ Tier 3 (Most restrictive)



Source: Bloomberg reporting

Note: Mapped data show level of restrictions on chip shipments for distinct markets

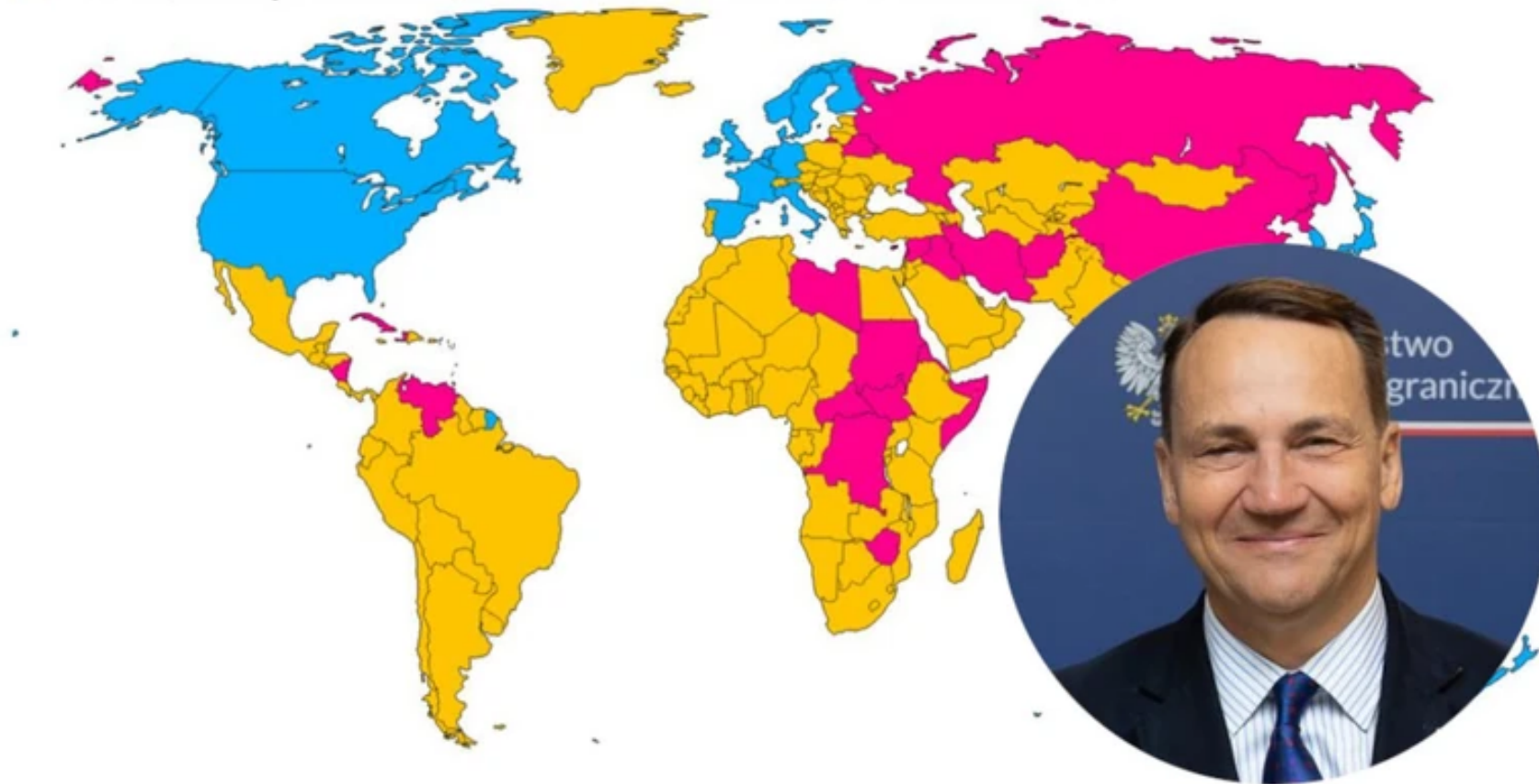
Uproar as Biden Administration Limits AI Chip Shipments

By Sally Ward-Foxton 01.15.2025 0

US to Curb Global Chip Shipments

Most markets will face new restrictions on data center development

■ Tier 1 (Most permissive) ■ Tier 2 ■ Tier 3 (Most restrictive)



Source: Bloomberg reporting

Note: Mapped data show level of restrictions on chip shipments for distinct markets

Bloomberg

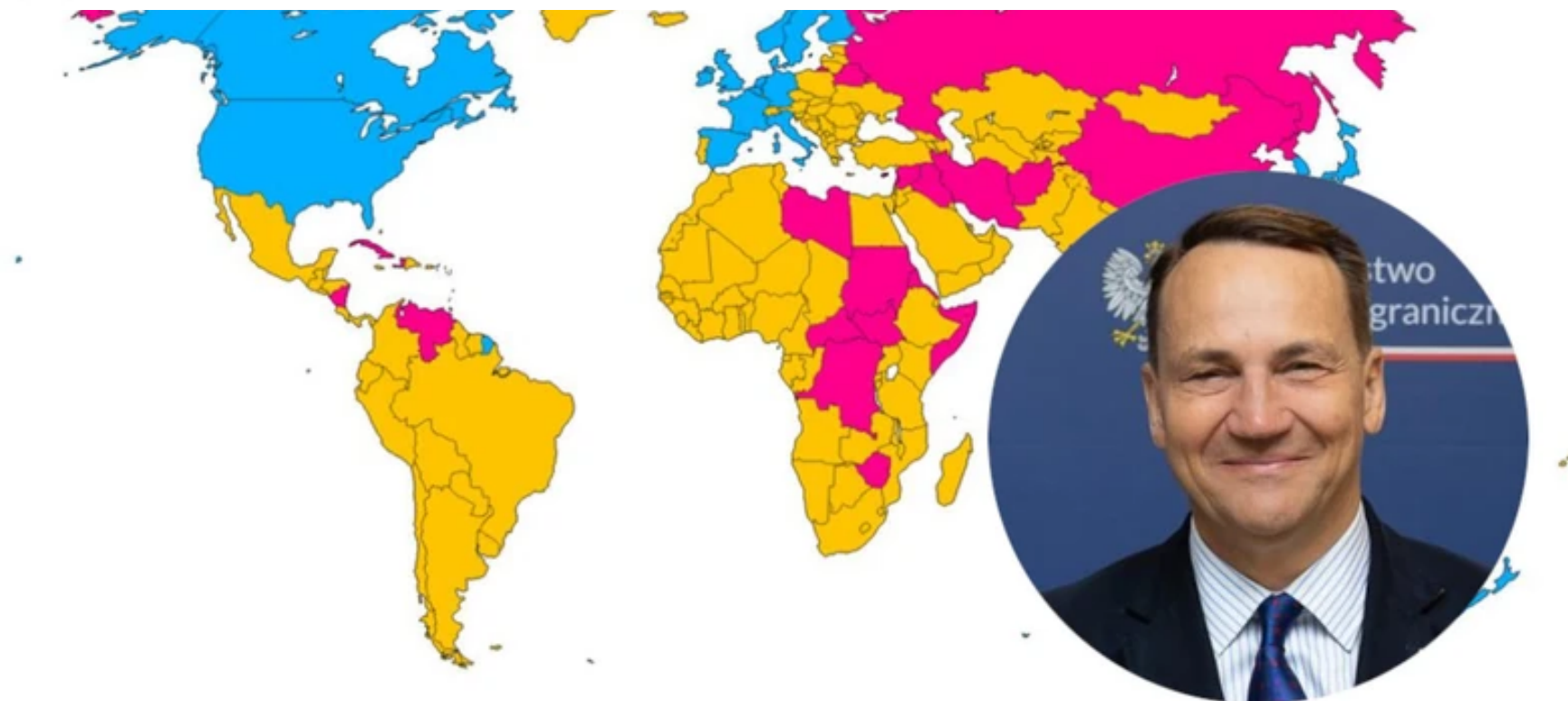
Ministerstwo Cyfryzacji odpowiada na restrykcje dot. sprzedaży GPU nałożone na Polskę przez USA



Joanna Żabnicka



15 stycznia 2025



Source: Bloomberg reporting

Note: Mapped data show level of restrictions on chip shipments for distinct markets

Bloomberg

Ministerstwo Cyfryzacji odpowiada na restrykcje dot. sprzedaży GPU nałożone na Polskę przez USA



Joanna Żabnicka



Piotr Sankowski • Obserwujesz

Associate Professor at the University of Warsaw and CSA at MIM Solutions

21 h • 🌐

...

Jednak ta mapka okazała się prawdą! A my w Polsce nie powinniśmy myśleć o jakiś globalnych zastosowaniach AI.

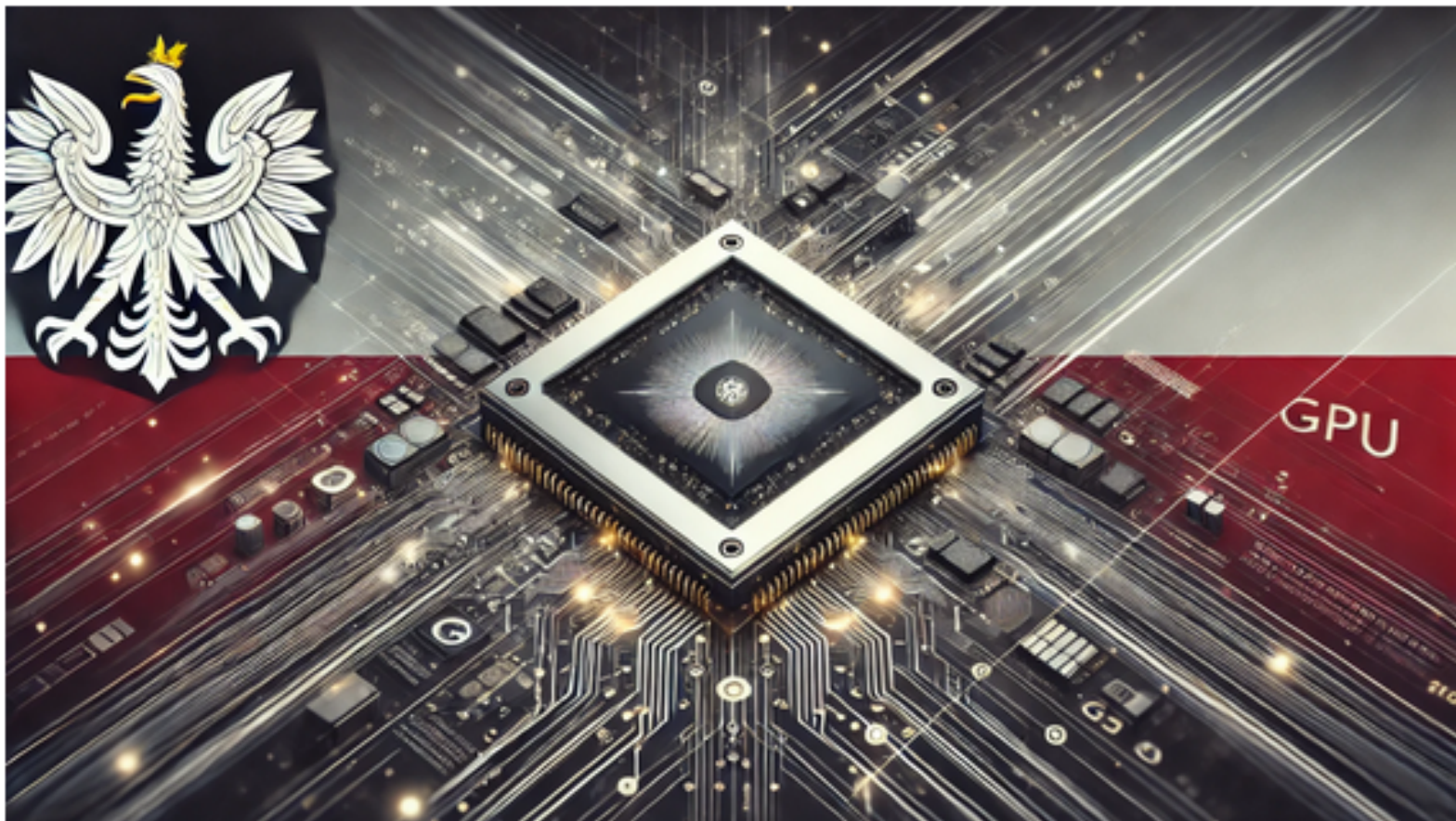
Na podstawie decyzji administracji prezydenta Bidena, Polska jest objęta restrykcjami w imporcie najnowszych chipów NVIDIA. Ostateczny dokument jest dostępny tutaj: <https://lnkd.in/dy7SsM5s>, choć pewnie lepiej spojrzeć na jego podsumowanie tutaj: https://lnkd.in/dnN4B_dq. Reguły te na pierwszy rzut oka wyglądają może niezbyt strasznie, bo limit 50 tys kart jaki dostała Polska może brzmieć jak dużo dla nas, a i jeszcze może zostać podniesiony do 100 tys.

Jednakże jak ktoś wie co się dzieje w rozwoju AI i, że teraz forpoczta walczy o stworzenie superinteligencji. To wydźwięk tego dokumentu jest jednoznaczny - żółte kraje powinny zapomnieć o rozwoju własnej superinteligencji. A limity mają im wystarczyć tylko do zwykłych zastosowań AI bo "ten limit zapewnia, że amerykańska technologia jest dostępna do obsługi zagranicznych rządów, dostawców usług medycznych i innych lokalnych przedsiębiorstw". Czyli broń boże nie powinniśmy myśleć o jakiś globalnych AI. Żałuję, że ten wpis jest tak negatywny, ale ja osobiście nie rozumiem jak w tej całej konfiguracji geopolitycznej znaleźliśmy się w drugiej kategorii państw.

Source: Bloomberg reporting

Note: Mapped data show level of restraints on distinct markets





Co nowa regulacja sprzedaży chipów dla AI oznacza dla Polski?



Marek Magryś

Acting Director at Cyfronet



<https://www.linkedin.com/pulse/co-nowa-regulacja-sprzeda%C5%BCy-chip%C3%B3w-dla-ai-oz>

Obecnie mamy w kraju zainstalowane ok. **2 000 GPU** w centrach KDM i wg. moich szacunków nie więcej niż drugie tyle w pozostałych jednostkach naukowych i w przemyśle. Najbliższe plany budowy Fabryk AI powinny pozwolić podwoić tę liczbę, co w połączeniu z innymi inwestycjami, np. w przemyśle czy w administracji publicznej, oznaczałoby, że w najlepszym razie będziemy mieć w tym roku w kraju ok. 10 000 najszybszych GPU, co stanowi 20% limitu.

Prawda jest jednak nieco inna, bo kiedy wczytamy się głębiej w regulację znajdziemy tam zapis, że **zamówienia do ok. 1700 GPU nie wymagają zgody i nie wliczają się w krajowy limit**. Wszystkie obecne zamówienia na wysokowydajne GPU w Polsce były mniejsze niż ta liczba; **Helios** zainstalowany w Academic Computer Centre CYFRONET AGH ma **464 takie chipy**. W praktyce więc, obecnie wykorzystany przez nasz kraj limit to 0%, który być może dobieje do 8-10% do końca roku dzięki inwestycjom w Fabryki AI.

Trump's AI Push: Understanding The \$500 Billion Stargate Initiative

By [Garth Friesen](#), Contributor. ⓘ Specialist in global markets, economics and a...

[Follow Author](#)

Jan 23, 2025, 08:00am EST

Initial equity funders



Key initial technology partners



Trump Announces the
Largest **\$500 Billion** Stargate
AI Infrastructure Project In
History



Megalodon: Efficient **LLM** Pretraining and Inference with Unlimited Context Length

In comparison with **LLAMA2**, MEGALODON achieves better efficiency than Transformer in the scale of 7 billion parameters and 2 **trillion** training tokens.

buff.ly/4aZ4eQb

MEGALODON: Efficient LLM Pretraining and Inference with Unlimited Context Length

Xuezhe Ma^{*,*} Xiaomeng Yang^{*,*} Wenhan Xiong^{*,*} Beidi Chen^{*,*} Lili Yu^{*,*}

Hao Zhang[†] Jonathan May[‡] Luke Zettlemoyer[‡] Omer Levy[‡] Chunting Zhou^{*,*}

^{*}AI at Meta

^{*}Carnegie Mellon University

[†]University of Southern California

[‡]University of California San Diego

Meta releases Llama 3, the most powerful open **LLM** model

Context length = 8,192 tokens

Training data size = 15 **trillion** tokens

Training data sources = Publicly available, over 30 languages

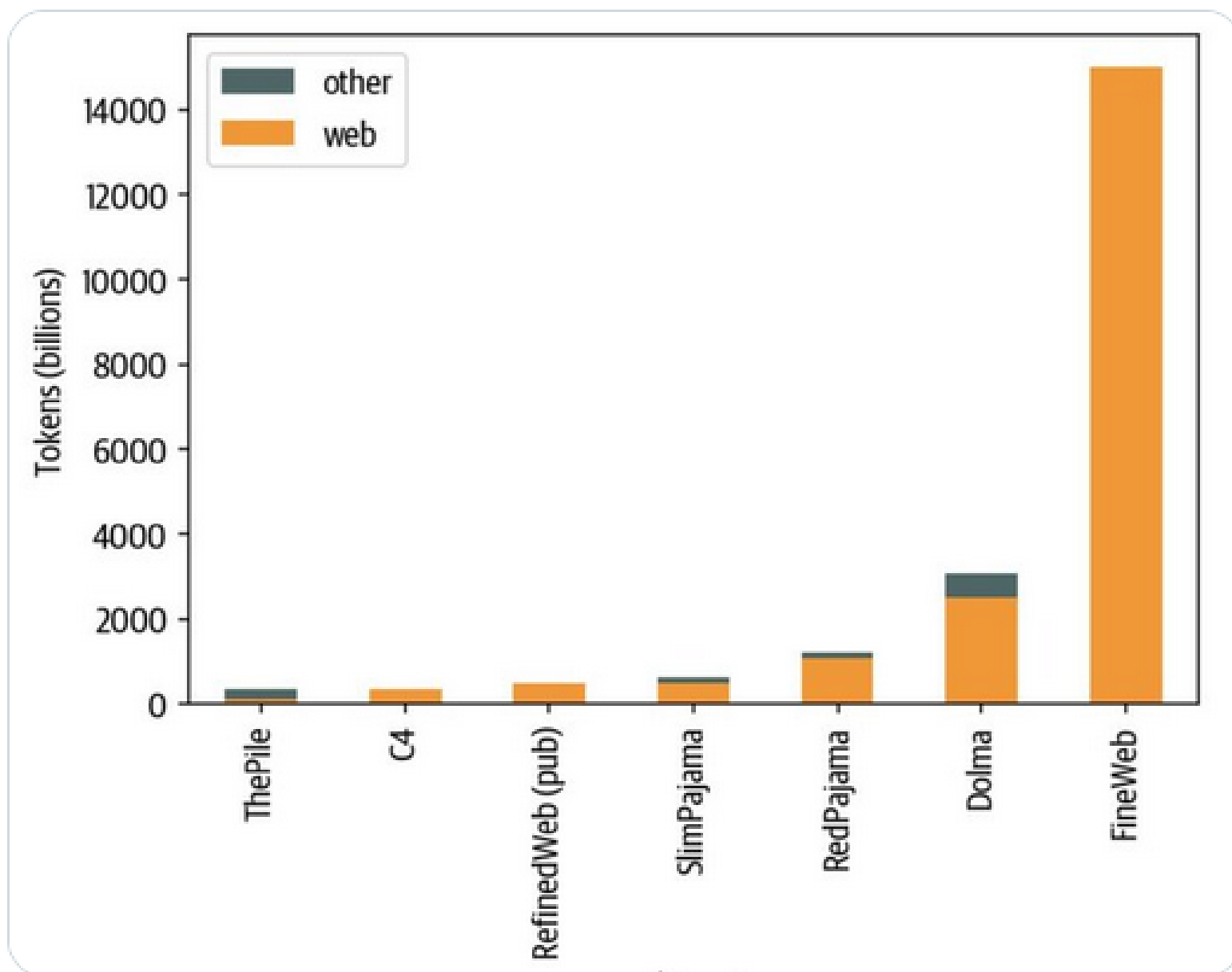
llama.meta.com/llama3/

	Meta Llama 3 8B	Gemma 7B - It Measured	Mistral 7B Instruct Measured
MMLU 5-shot	68.4	53.3	58.4
GPQA 0-shot	34.2	21.4	26.3
HumanEval 0-shot	62.2	30.5	36.6
GSM-8K 8-shot, CoT	79.6	30.6	39.9
MATH 4-shot, CoT	30.0	12.2	11.0

	Meta Llama 3 70B	Gemini Pro 1.5 Published	Claude 3 Sonnet Published
MMLU 5-shot	82.0	81.9	79.0
GPQA 0-shot	39.5	41.5 CoT	38.5 CoT
HumanEval 0-shot	81.7	71.9	73.0
GSM-8K 8-shot, CoT	93.0	91.7 11-shot	92.3 0-shot
MATH 4-shot, CoT	50.4	58.5 Minerva prompt	40.5

The largest dataset with 15T+ tokens for **LLM** training

Llama 3 models are trained on over 15 **trillion** tokens. Seems impossible to get such a huge dataset? It's not! FineWeb is a dataset with 15 **trillion+** tokens of cleaned and deduplicated English **web** data from CommonCrawl dumps.





BIELIK

BIELIK-11B-v2

Large Language Model

Bezpieczne przetwarzanie

Pełna kontrola

Kompaktowa moc

Porozmawiaj z Bielikiem

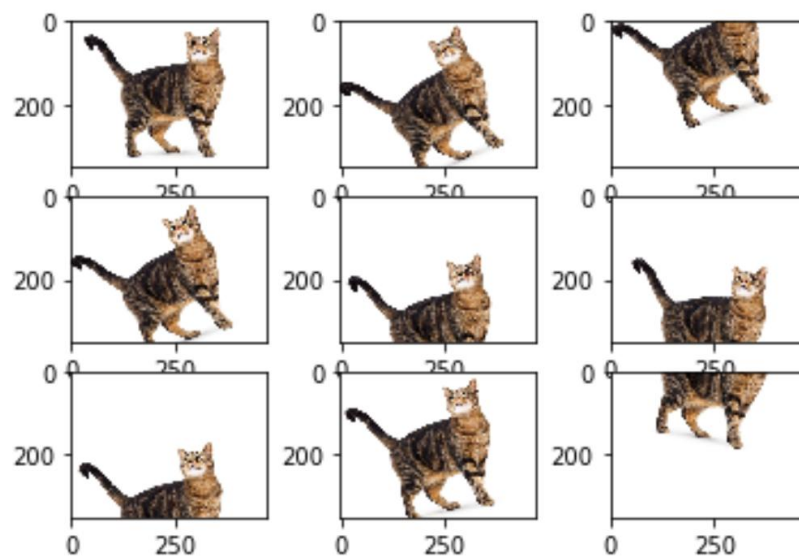
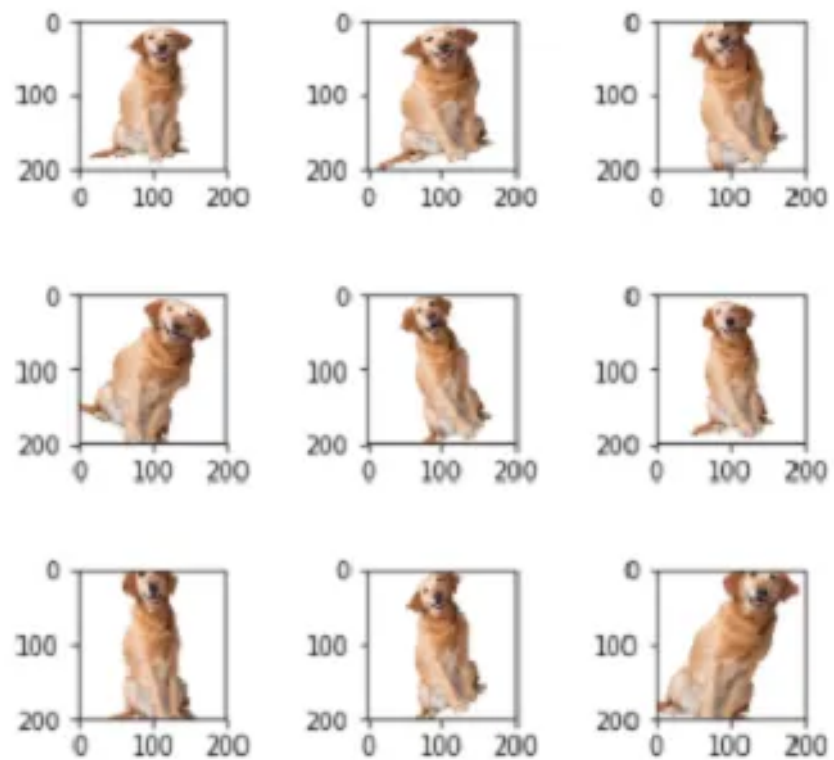
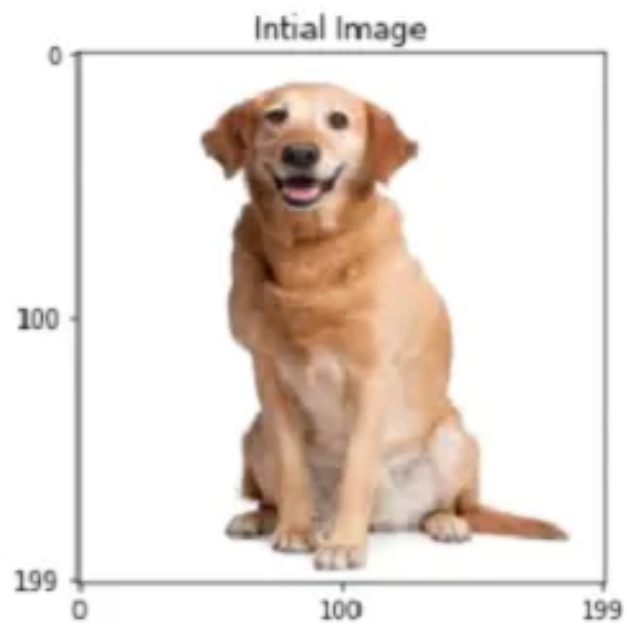


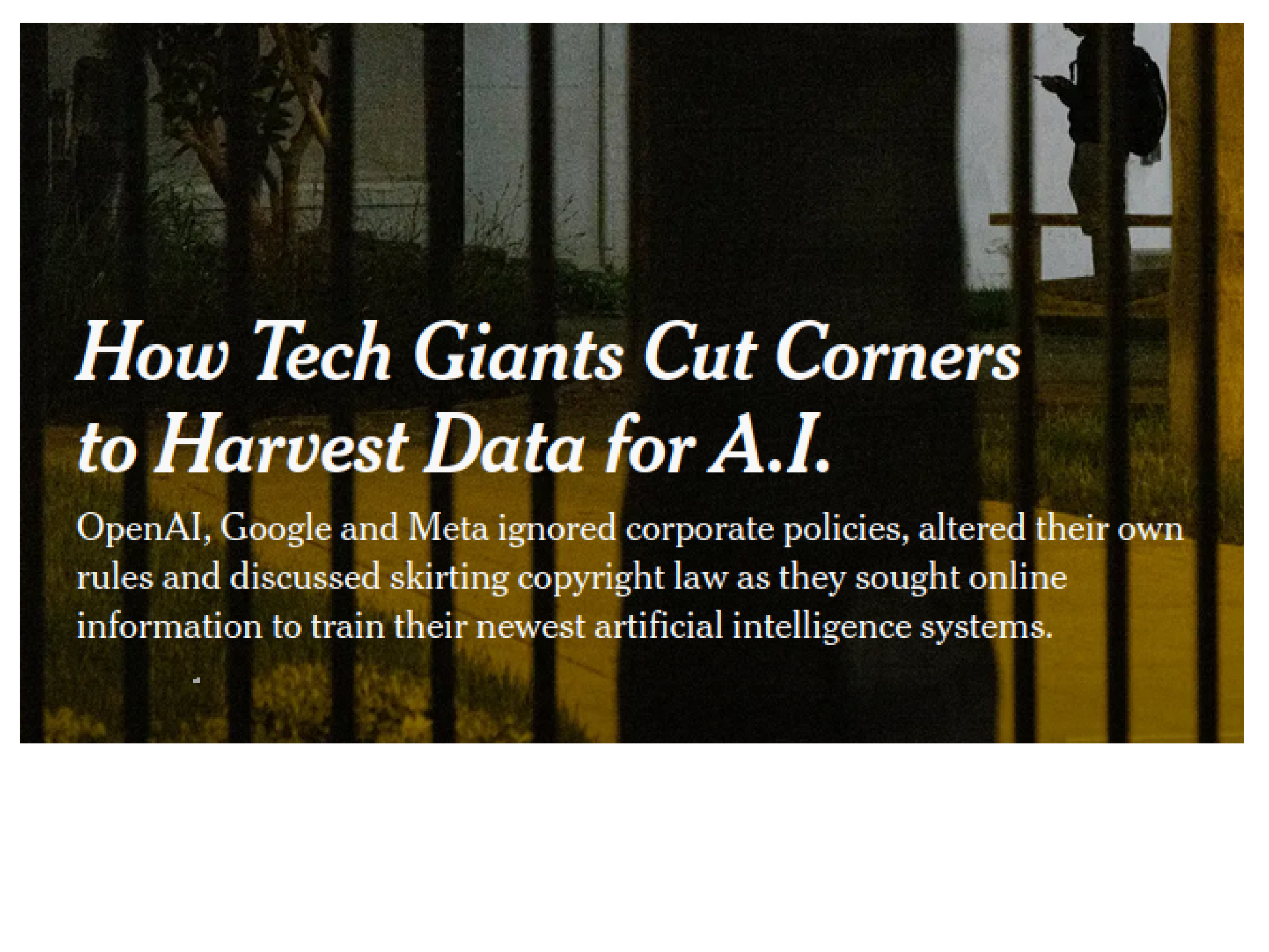
<https://bielik.ai/>

DATA AUGMENTATION



A technique of artificially increasing the training set by creating modified copies of a dataset using existing data. It includes making minor changes to the dataset or using deep learning to generate new data points.





How Tech Giants Cut Corners to Harvest Data for A.I.

OpenAI, Google and Meta ignored corporate policies, altered their own rules and discussed skirting copyright law as they sought online information to train their newest artificial intelligence systems.

ONE HUNDRED EXAMPLES OF GPT-4 MEMORIZING CONTENT FROM THE NEW YORK TIMES

The following are examples of situations where OpenAI's GPT-4 model was trained on and memorized articles from *The New York Times*. Each example focuses on a single news article. Examples were produced by breaking the article into two parts. The first part of the article is given to GPT-4, and GPT-4 replies by writing its own version of the remainder of the article. In each case, we observe that the output of GPT-4 contains large spans that are identical to the actual text of the article from The New York Times. For each example, we provide the following:

- The URL of the online version of the article.
- The prompt that was given to GPT-4. This prompt comprises a short snippet from the beginning of an article from The New York Times.
- The response from GPT-4. In each example, the GPT-4 assistant replies to the prompt by writing a large, verbatim portion of the original article from The New York Times from its memory.
- The original end of the article, as it appears on NYTimes.com.

Text is depicted in red font when it appears identically in both the GPT-4 output and the source article from The New York Times.

Output from GPT-4:

exempted it from regulations, subsidized its operations and promoted its practices, records and interviews showed.

Their actions turned one of the best-known symbols of New York — its yellow cabs — into a financial trap for thousands of immigrant drivers. More than 950 have filed for bankruptcy, according to a Times analysis of court records, and many more struggle to stay afloat.

“Nobody wanted to upset the industry,” said David Klahr, who from 2007 to 2016 held several management posts at the Taxi and Limousine Commission, the city agency that oversees medallions. “Nobody wanted to kill the golden goose.”

New York City in particular failed the taxi industry, The Times found. Two former mayors, Rudolph W. Giuliani and Michael R. Bloomberg, placed political allies inside the Taxi and Limousine Commission and directed it to sell medallions to help them balance budgets and fund key initiatives.

During that period, much like in the mortgage lending crisis, a group of industry leaders enriched themselves by artificially inflating medallion prices. They encouraged medallion buyers to borrow as much as possible and ensnared them in interest-only loans and other one-sided deals that often required borrowers to pay hefty fees, forfeit their legal rights and give up most of their monthly incomes.

When the market collapsed, the government largely abandoned the drivers who bore the brunt of the crisis. Officials did not bail out borrowers or persuade banks to soften loan

Actual text from NYTimes:

exempted it from regulations, subsidized its operations and promoted its practices, records and interviews showed.

Their actions turned one of the best-known symbols of New York — its signature yellow cabs — into a financial trap for thousands of immigrant drivers. More than 950 have filed for bankruptcy, according to a Times analysis of court records, and many more struggle to stay afloat.

“Nobody wanted to upset the industry,” said David Klahr, who from 2007 to 2016 held several management posts at the Taxi and Limousine Commission, the city agency that oversees cabs. “Nobody wanted to kill the golden goose.”

New York City in particular failed the taxi industry, The Times found. Two former mayors, Rudolph W. Giuliani and Michael R. Bloomberg, placed political allies inside the Taxi and Limousine Commission and directed it to sell medallions to help them balance budgets and fund priorities. Mayor Bill de Blasio continued the policies.

Under Mr. Bloomberg and Mr. de Blasio, the city made more than \$855 million by selling taxi medallions and collecting taxes on private sales, according to the city.

But during that period, much like in the mortgage lending crisis, a group of industry leaders enriched themselves by artificially inflating medallion prices. They encouraged medallion buyers to borrow as much as possible and ensnared them in interest-only loans and other one-sided deals that often required them to pay hefty fees, forfeit their legal rights and give up most of their monthly incomes.

The Times Sues OpenAI and Microsoft Over A.I. Use of Copyrighted Work

Millions of articles from The New York Times were used to train chatbots that now compete with it, the lawsuit said.



Share full article



1.3K



OpenAI Seeks to Dismiss Parts of The New York Times's Lawsuit

The artificial intelligence start-up argued that its online chatbot, ChatGPT, is not a substitute for a New York Times subscription.

OpenAI claims New York Times 'hacked' ChatGPT to build copyright lawsuit

In a filing Monday, OpenAI claims a 'hired gun' took 'tens of thousands of attempts to generate the highly anomalous results'



Pulitzer-winning authors join OpenAI, Microsoft copyright lawsuit

By Blake Brittain

December 21, 2023 1:22 AM GMT+1 · Updated 4 months ago



Writer and Hollywood Reporter editor Julian Sancton first filed the proposed class-action lawsuit last month. The case is one of several that have been brought by groups of copyright owners including authors John Grisham, George R.R. Martin and Jonathan Franzen against OpenAI and other tech companies over the alleged misuse of their work in AI training.

The copyright case against AI art generators just got stronger with more artists and evidence



Credit: VentureBeat made with OpenAI DALL-E 3 via ChatGPT

Stable Diffusion

https://en.wikipedia.org/wiki/Stable_Diffusion



Stable Diffusion

https://en.wikipedia.org/wiki/Stable_Diffusion



Stable Diffusion

https://en.wikipedia.org/wiki/Stable_Diffusion

Dall-E2



Midjourney



Stable Diffusion 2



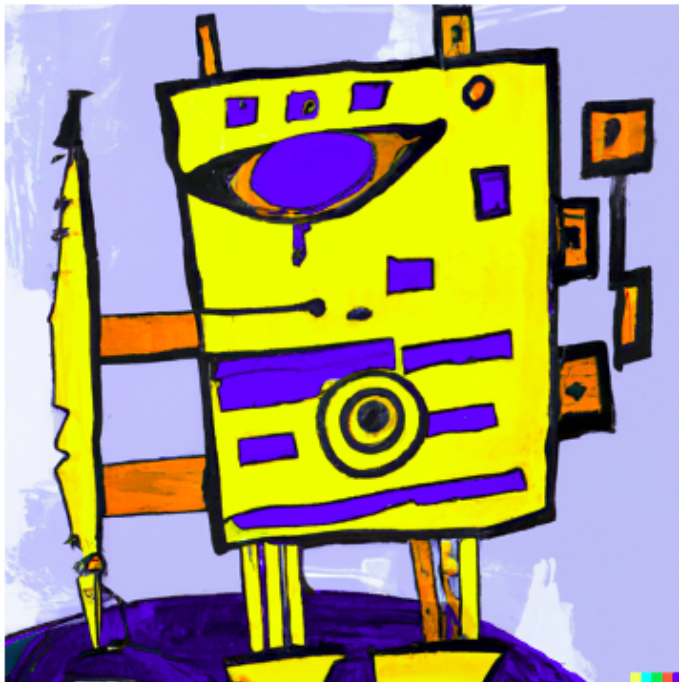
Prompt: green forest, blue sky, fog, clouds, Ultra HD nature photography

<https://miroslawmamczur.pl/dall-e2-vs-midjourney-vs-stable-diffusion/>

Stable Diffusion

https://en.wikipedia.org/wiki/Stable_Diffusion

Dall-E2



Midjourney



Stable Diffusion 2



Prompt: yellow transformer robot on alien planet in Pablo Picasso style

<https://mirosławmamczur.pl/dall-e2-vs-midjourney-vs-stable-diffusion/>

Stable Diffusion

https://en.wikipedia.org/wiki/Stable_Diffusion

Dall-E2



Midjourney



Stable Diffusion 2



Prompt: Close-up of Viking King emerging from wet black mud

<https://mirosławmamczur.pl/dall-e2-vs-midjourney-vs-stable-diffusion/>

Stable Diffusion

https://en.wikipedia.org/wiki/Stable_Diffusion

Dall-E2



Midjourney



Stable Diffusion 2



Prompt: green forest, blue sky, fog, clouds, Ultra HD nature photography

<https://miroslawmamczur.pl/dall-e2-vs-midjourney-vs-stable-diffusion/>

The copyright case against AI art generators just got stronger with more artists and evidence

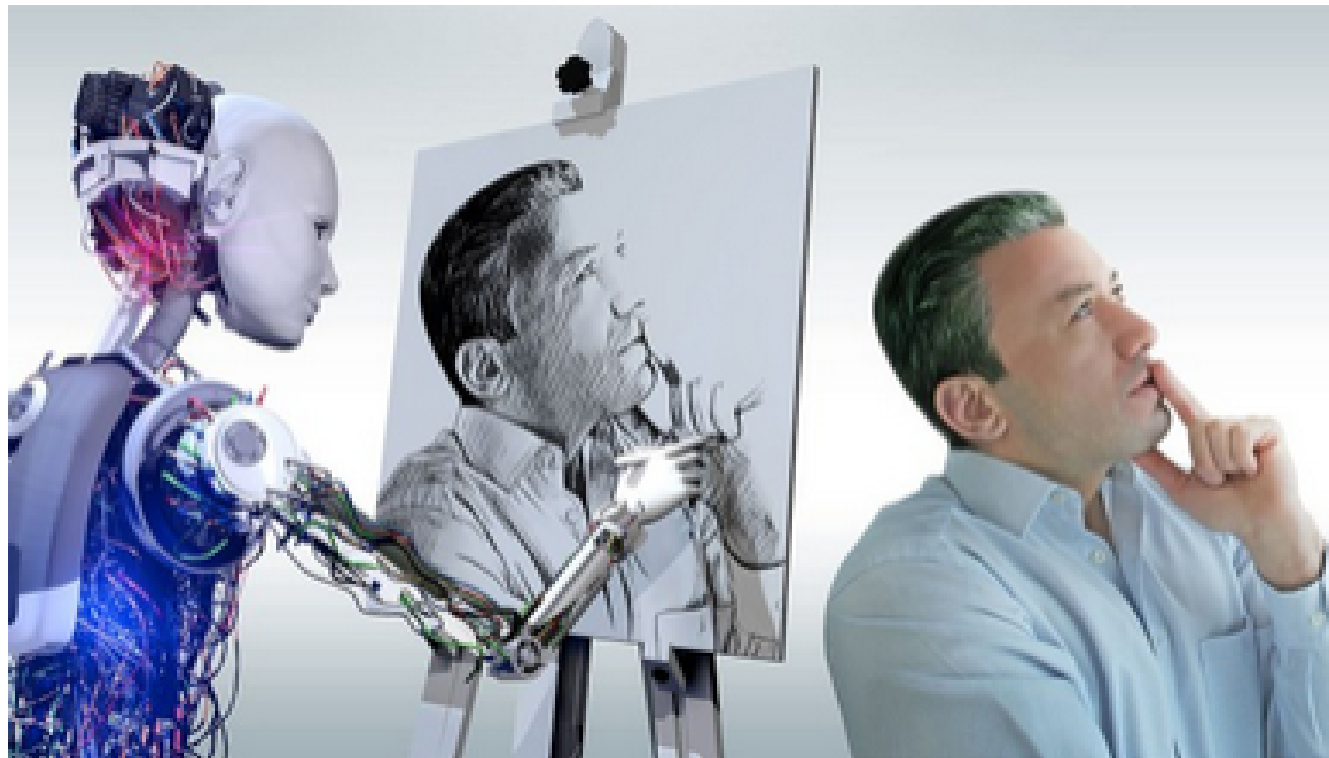


Credit: VentureBeat made with OpenAI DALL-E 3 via ChatGPT

Artists Lose First Round of Copyright Infringement Case Against AI Art Generators

While a federal judge advanced an infringement claim against Stability AI, he dismissed the rest of the lawsuit.

BY WINSTON CHO  OCTOBER 30, 2023 4:57PM



ADVERTISEMENT

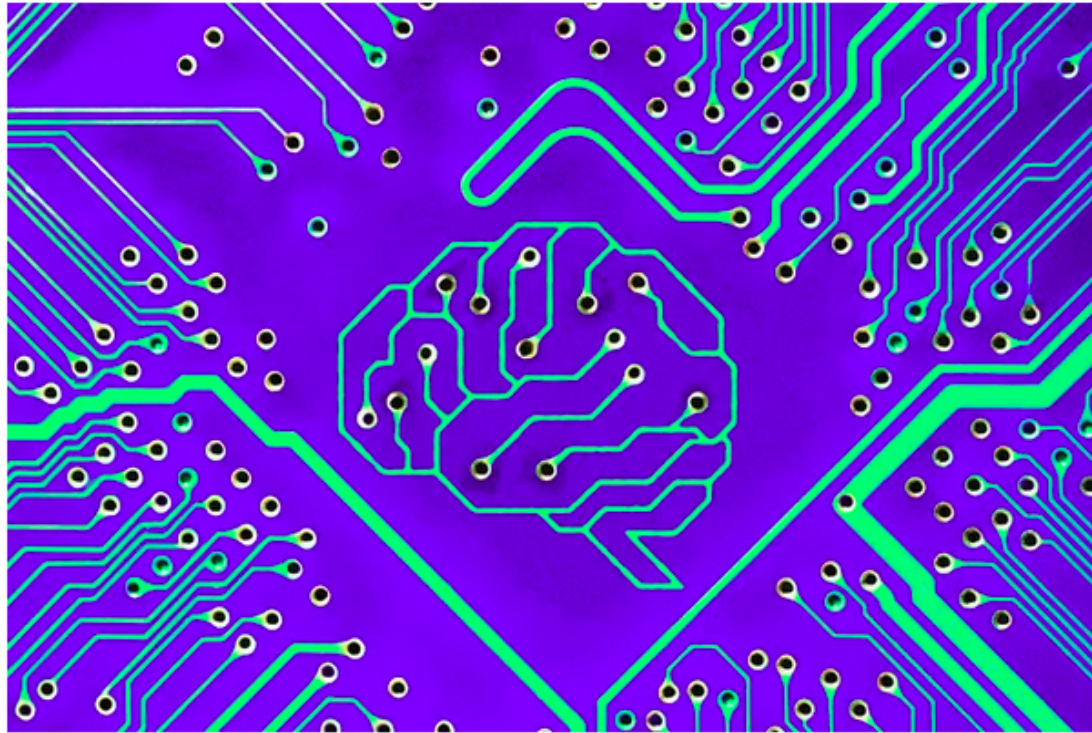
THR NEWSLETTERS

Sign up for THR news straight to
your inbox every day

PHOTO

<https://www.hollywoodreporter.com/business/business-news/artists-copyright-infringement-case-ai-art-generators-1235632929/>

OpenAI transcribed over a million hours of YouTube videos to train GPT-4



Cath Virginia / The Verge | Photos from Getty Images

/ A New York Times report details the ways big players in AI have tried to expand their data access.

By [Wes Davis](#), a weekend editor who covers the latest in tech and entertainment. He has written news, reviews, and more as a tech journalist since 2020.

Apr 6, 2024, 10:29 PM GMT+2



45 Comments (45 New)

If you buy something from a Verge link, Vox Media may earn a commission. [See our ethics statement.](#)

YOUTUBE WARNS OPENAI TRAINING AI WITH ITS CREATORS' VIDEOS IS A “CLEAR VIOLATION”



GitHub Users Want to Sue Microsoft For Training an AI Tool With Their Code



Copilot



Microsoft



Buys



GitHub

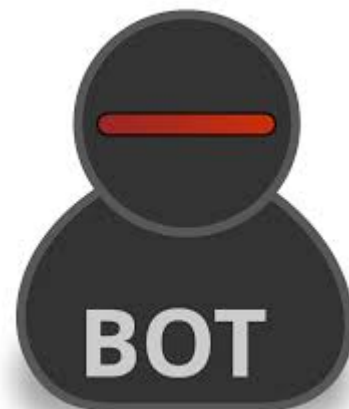
Wikipedia Faces Flood of AI Bots That Are Eating Bandwidth, Raising Costs

'Our content is free, our infrastructure is not,' the nonprofit that runs Wikipedia says, warning that AI scrapers are wasting the site's bandwidth.

Michael

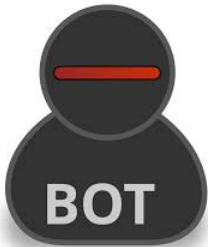
Kan

By [Michael Kan](#) April 2, 2025



Wikipedia and Wikidata Realtime Edit Stats

[Read](#)



Listening to a [Server-Sent Events stream](#) based on Wikipedia and Wikidata edits

☒ Consider Wikidata (It may take some time until Wikidata and all Wikipedias show up)

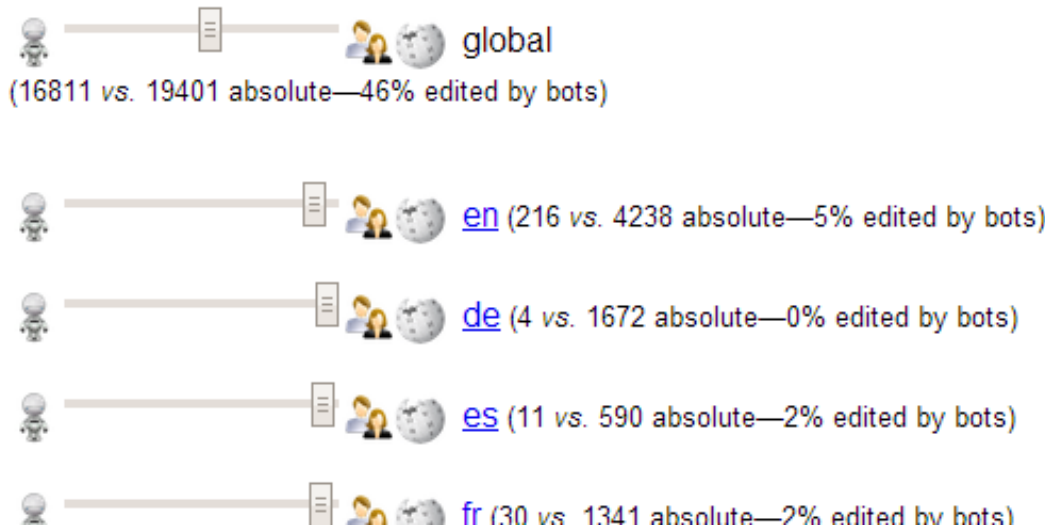
**Bots vs.**

**Wikipedians**

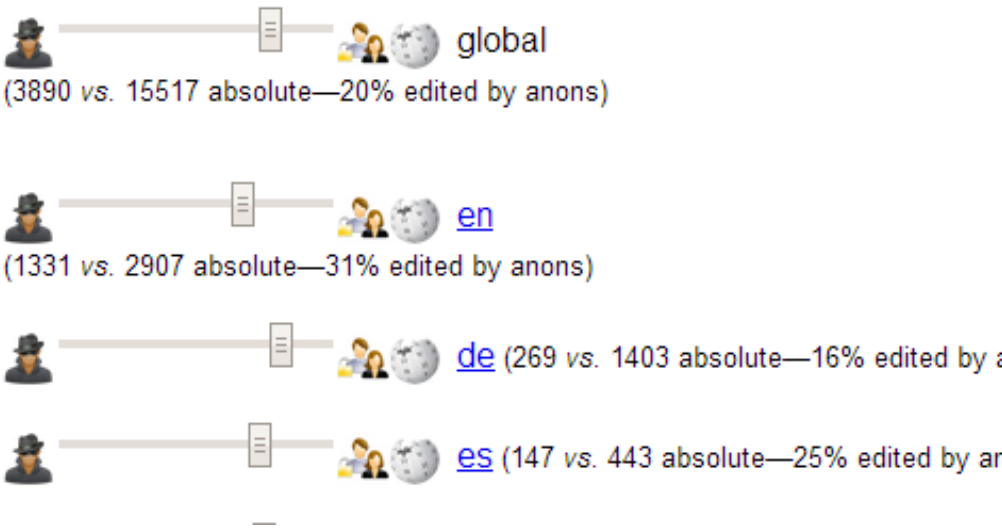
**Anon Humans vs.**

**Logged-In Humans**

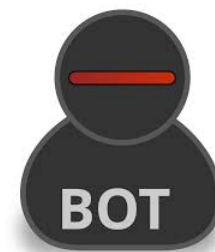
Who edits more?



Who edits more?



Internet Traffic Breakdown (2023)



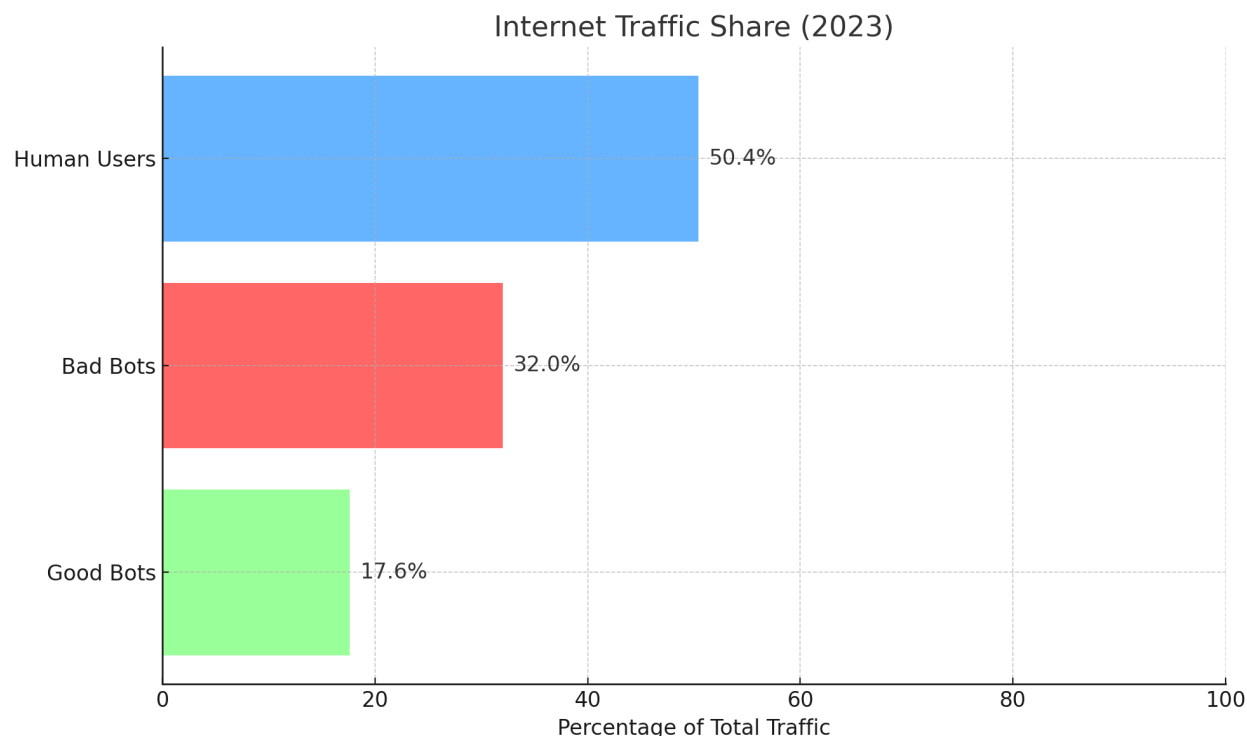
According to the 2024 Imperva Bad Bot Report:

Human Users: 50.4% of total internet traffic

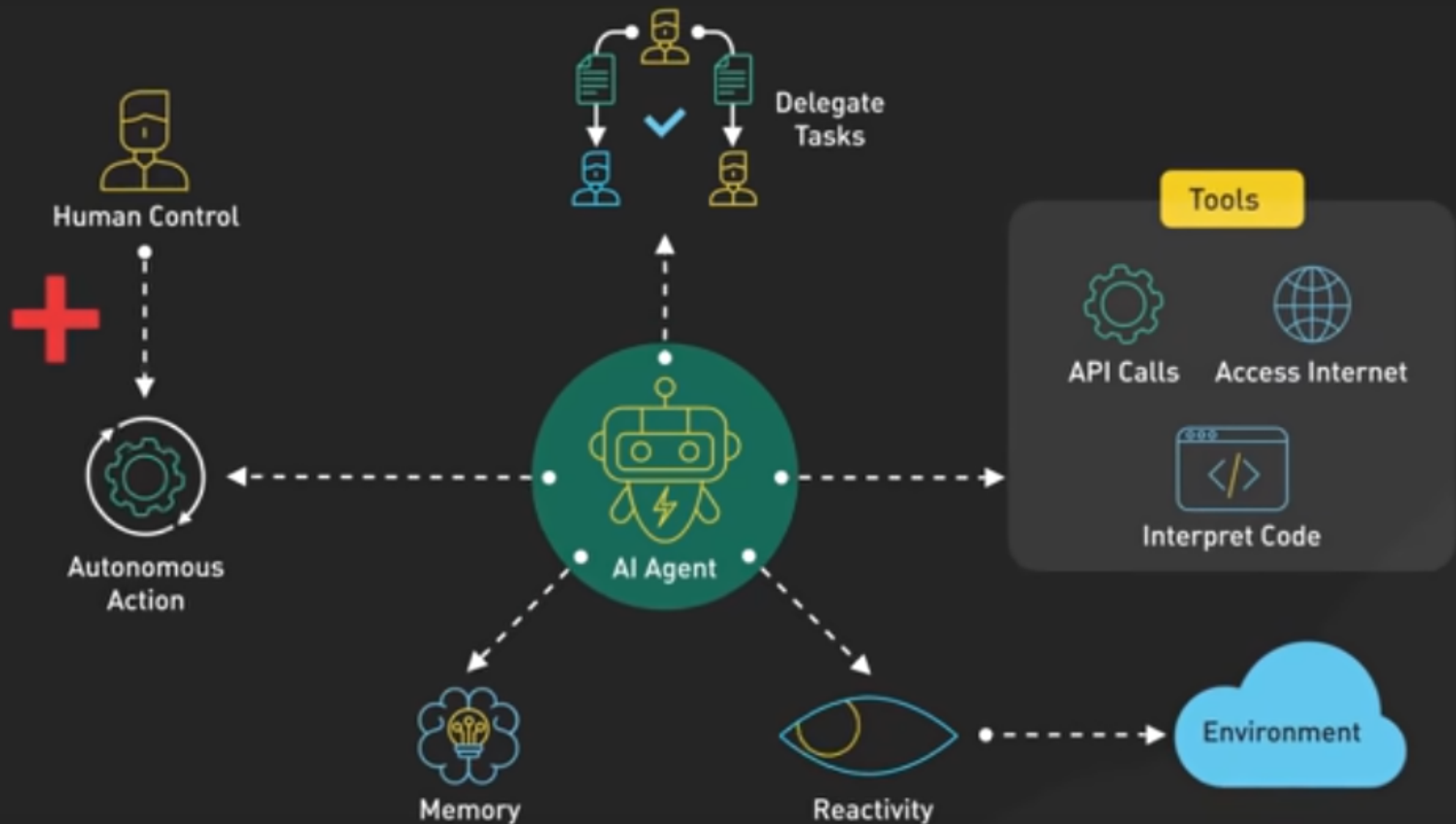
Bots: 49.6% of total internet traffic

Good Bots: 17.6%

Bad Bots: 32.0%



AI Agents



Thank you for your time
and
See you at the next lecture

Any other
questions & comments

l.kozlowski@mimuw.edu.pl