This talk is about providing a Human-Centered Data Science {DS} framework of how to

(i) demystify &
(ii) make {AI} tangible

within a higher educational setting

https://robfvdw.medium.com/the-world-wide-web-ai-safari-b2e4f7f90647

WHY DESMISTFYING AI?

HOGESCHOOL ROTTERDAM

To make sense of the world

> “Sense-making is the way that humans choose between multiple possible explanations of sensory input.”

– Dave Snowden

http://kwork.org/Stars/Snowden/snowden3.html#Simplicity

“The inescapable resurgence of {AI}
on the world wide web {WWW}
— along with the arrival of Internet-of-Things {IoT} —
has expanded the scope of the
digital world into the realm of cybernetics”

Cybernetics studies communication & control of information in living beings + the machines built by humans

======> Feedback & Reinforcement <======

The cybernetic foundation of {AI} explains its insatiable hunger for Big data, with the promise to solve societal challenges ranging from:

Health - Climate Change - Safety up to Cyber Physical Systems {CPSs}: Robotics & Driverless Cars

Today it seems we only
receive ambiguous promises
and paradoxical stories

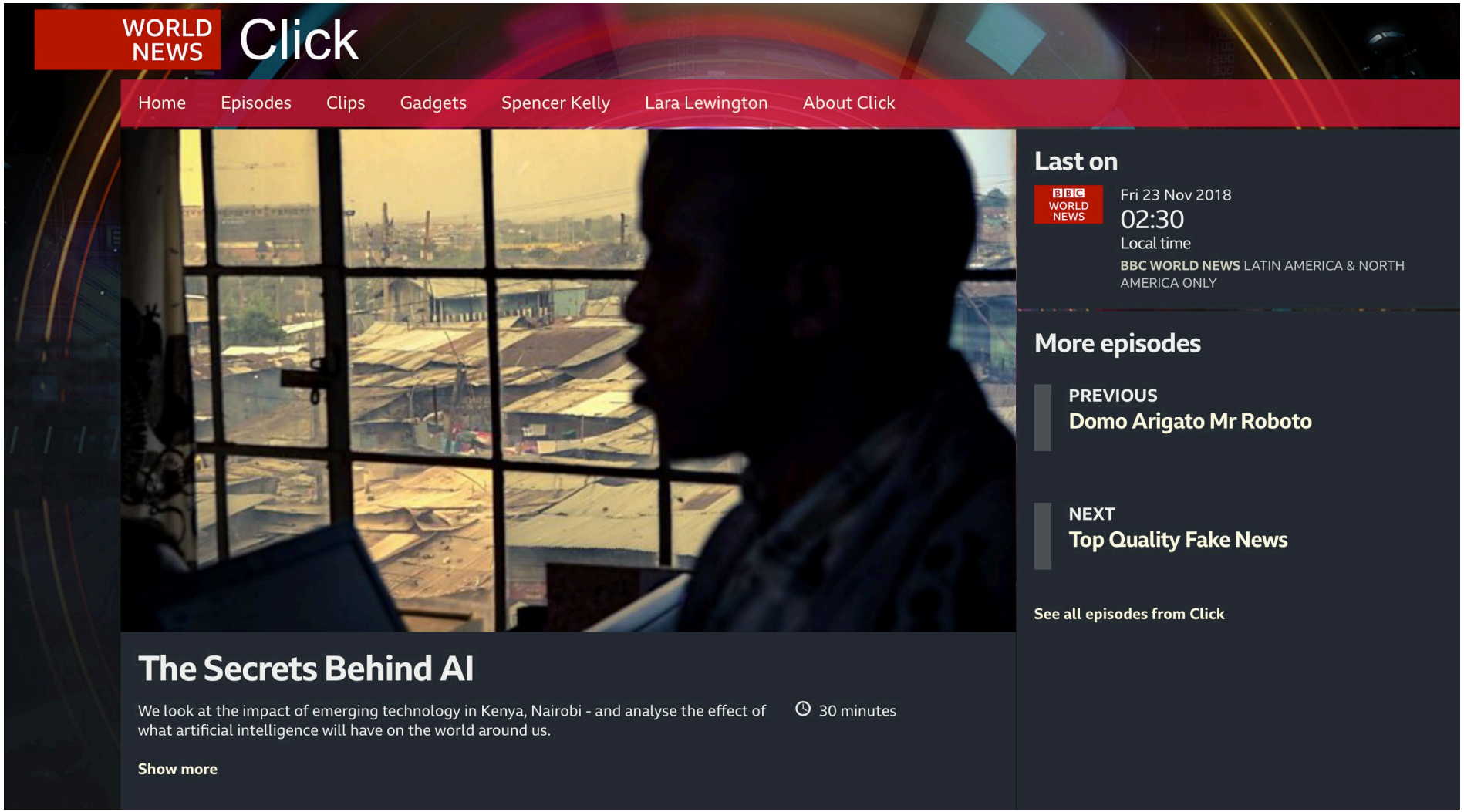
{AI} has revealed itself to us
as a double-edged sward:

Dangerous *yet* Supportive
All-Consuming *yet* Liberating

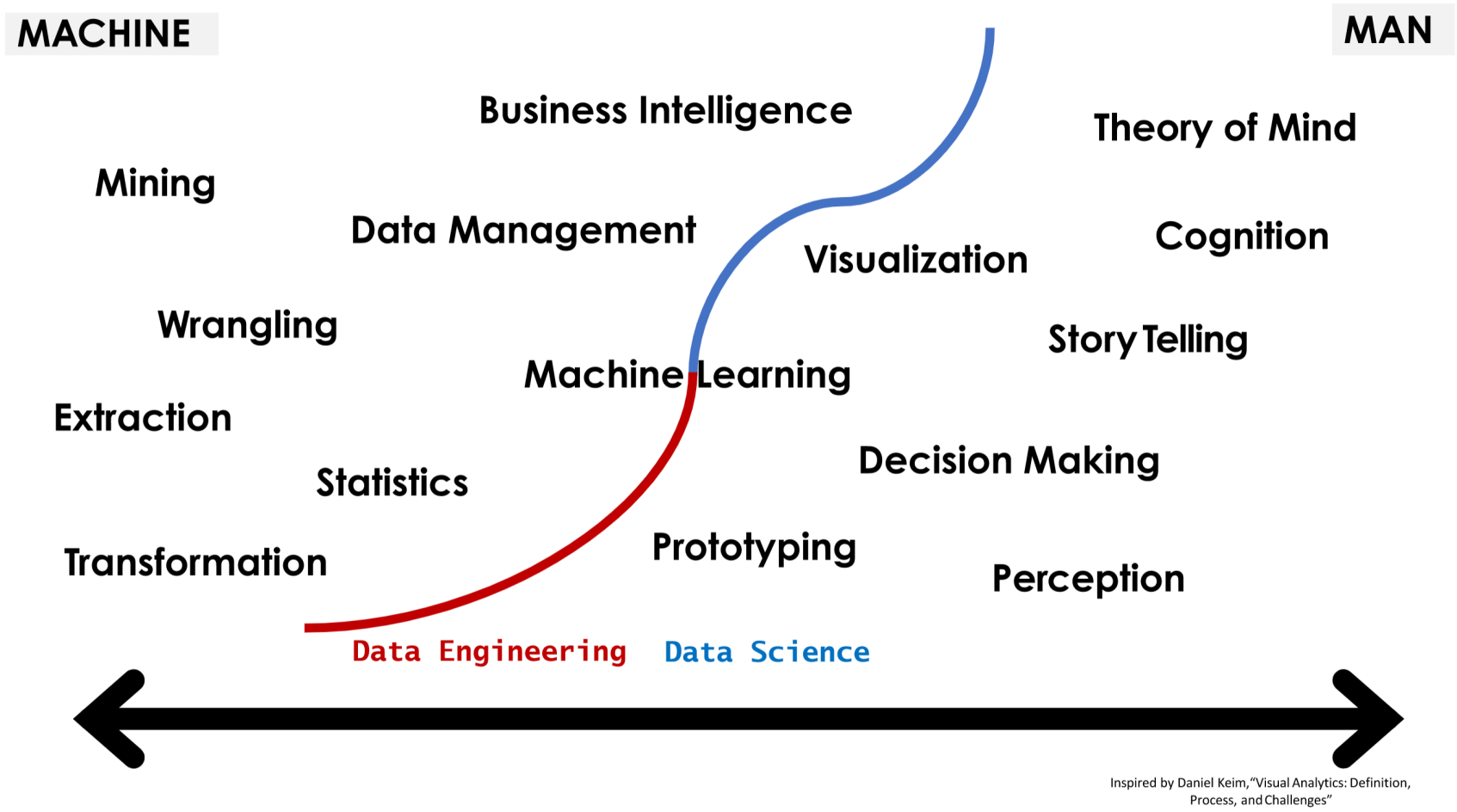
{AI} is dominated by labour intensive, wasteful & costly Brute-Force practices using neural networks

{The Secrets Behind AI}

HOGESCHOOL ROTTERDAM

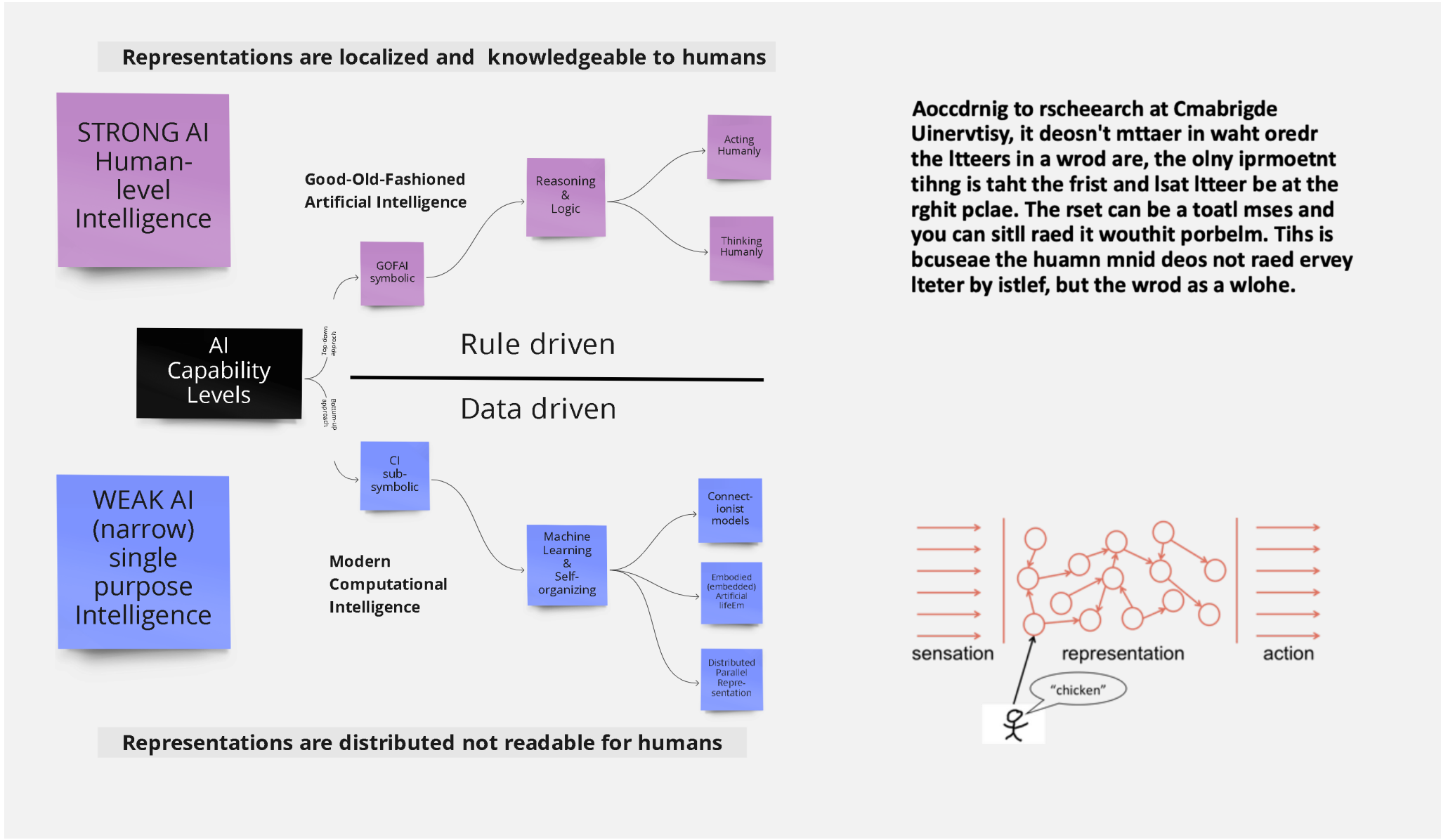


{AI integrates two Scientific Disciplines}



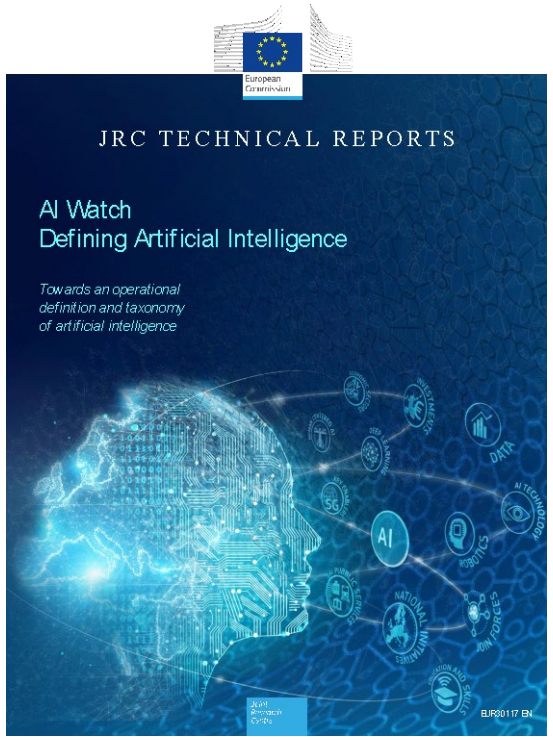
https://robfvdw.medium.com/a-generic-approach-to-data-driven-activities-d85ad558b5fa

{The taxonomy of AI is complex}



Aoccdrnig to rscheearch at Cmabrigde Uinervtisy, it deosn't mttaer in waht oredr the ltteers in a wrod are, the olny iprmoetnt tihng is taht the frist and lsat ltteer be at the rghit pclae. The rset can be a toatl mses and you can sitll raed it wouthit porbelm. Tihs is bcuseae the huamn mnid deos not raed ervey lteter by istlef, but the wrod as a wlohe.

{The taxonomy of AI is complex}



https://publications.jrc.ec.europa.eu/repository/handle/JRC118163

	AI taxonomy	
	AI domain	**AI subdomain**
Core	**Reasoning**	Knowledge representation
		Automated reasoning
		Common sense reasoning
	Planning	Planning and Scheduling
		Searching
		Optimisation
	Learning	Machine learning
	Communication	Natural language processing
	Perception	Computer vision
		Audio processing
Transversal	**Integration and Interaction**	Multi-agent systems
		Robotics and Automation
		Connected and Automated vehicles
	Services	AI Services
	Ethics and Philosophy	AI Ethics
		Philosophy of AI

http://dx.doi.org/10.2760/382730%20

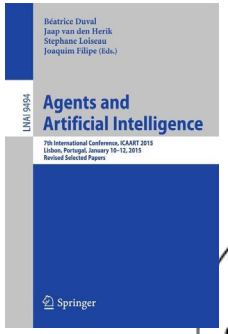
{AI favors Agent-Based Models}

HOGESCHOOL ROTTERDAM

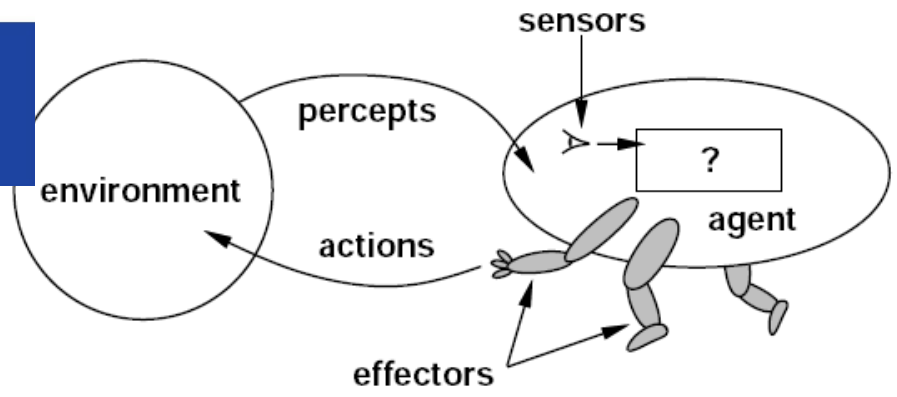
Agents

are abstractions of the real world {models} that can perceive their environment through sensors (input) and act upon that environment through effectors (output), combined with learning capabilities.

As a result, agent behaviour is desirable from an AI-viewpoint



Model



Real-world

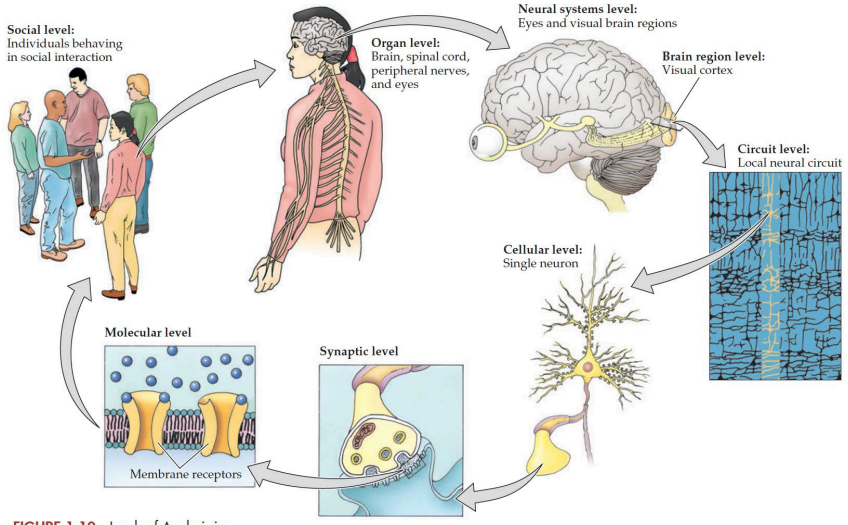
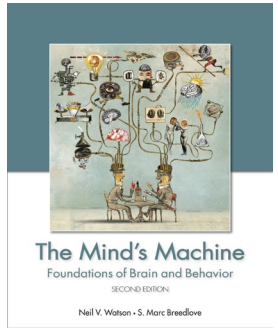


FIGURE 1.10 Levels of Analysis in Biological Psychology

Animal research is an essential part of life sciences research, including biological psychology



https://pdfs.semanticscholar.org/0629/51ac18d311efbfb4a019b1b5c513f1208de5.pdf

{Brute Force}

HOGESCHOOL ROTTERDAM
CREATING

State-of-the-Art {SOTA} {AI} relies
on a brute-force approach:

High-performance GPU or TPU computing power
Black-Box, poorly understood solutions
Training with massive data-sets
Deep Neural Network architecture
Billions of parameters
Lengthy Process & Huge budgets

to produce cognitive abilities
that are on par with Human-Level Performance.

{Chat-Bot Meena}

https://ai.googleblog.com/2020/01/towards-conversational-agent-that-can.html
https://arxiv.org/pdf/2001.09977.pdf

HOGESCHOOL ROTTERDAM
CREATING

Google research 2020:

"Meena is a conversational agent capable of chatting convincingly about any topic that is meaningful to humans"

HPC: 2,048 TPU v3 cores, 16GB DDR5 memory
trained on 40 billion words (61B BPE tokens) [341 GB]
Evolved Transformer NAS
32 decoder layers, deep neural network
2.6 [Billion] parameters
30 days of training, 4[Million] tokens per second
Training cost: 1,500,000$

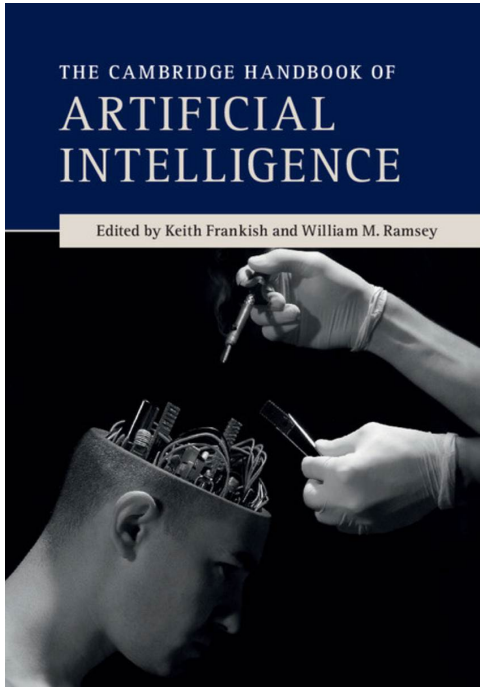
https://cloud.google.com/tpu/pricing

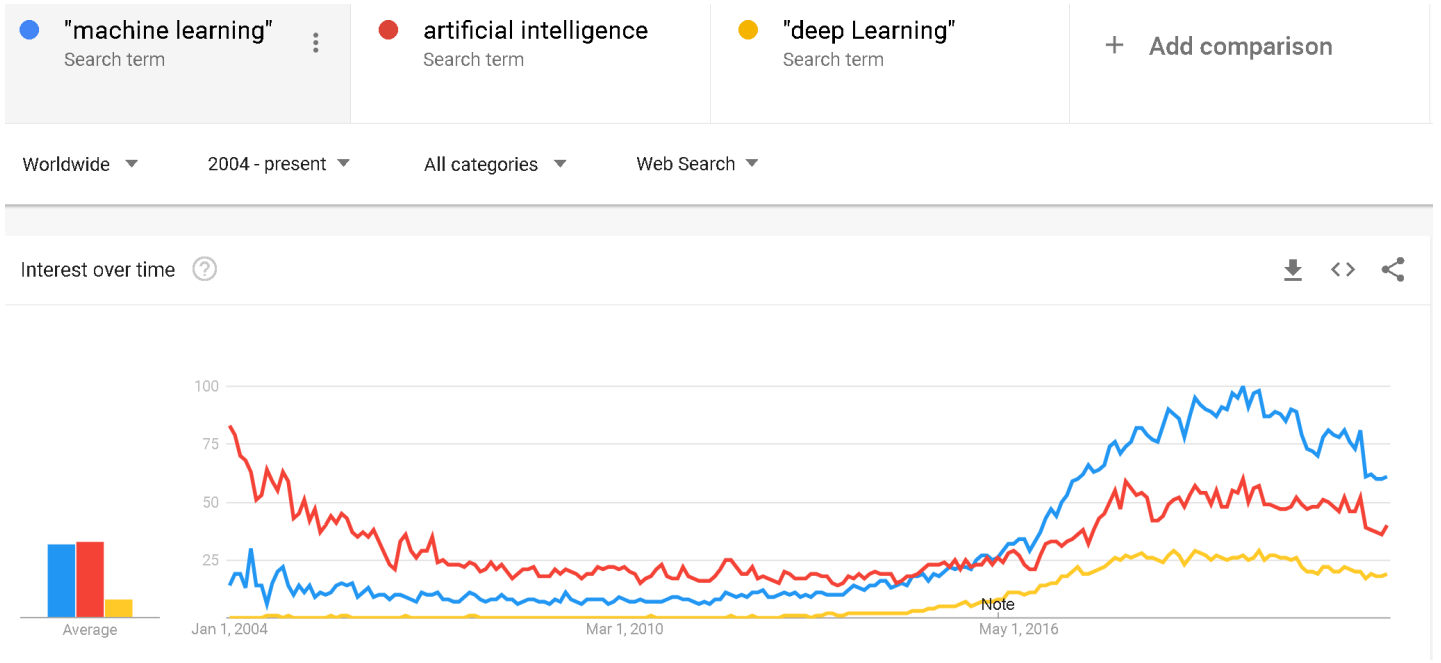
What constitutes a {Deep} Neural Net?

{DL}

HOGESCHOOL ROTTERDAM
CREATING

(Pre-trained) Deep Learning {DL} Neural Networks {NNs} are the most advanced, successful & fastest growing Artificial Intelligence {AI} technology



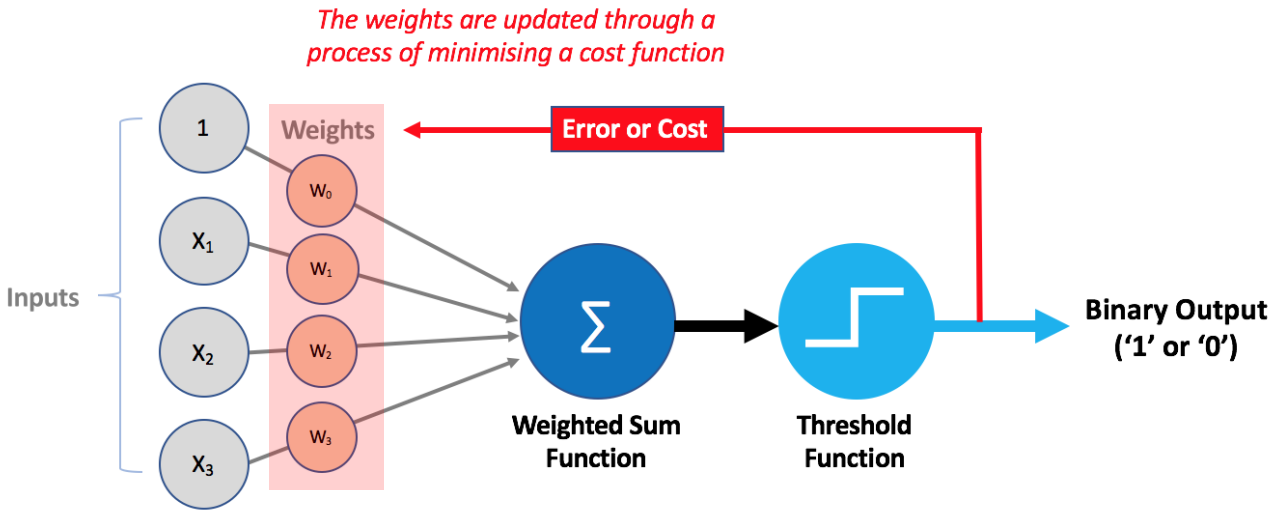


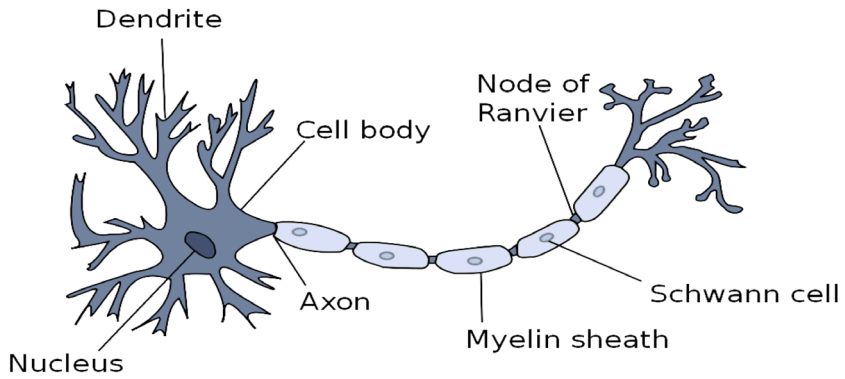
https://trends.google.com/trends/explore?date=all&q=%22machine%20learning%22,artificial%20intelligence,%22deep%20Learning%22

{Artificial Neurons}

Deep Neural Nets {DNNs} harbor vast amounts of "artificial neurons" ➔smallest computational unit⬅

Names for
Artificial Neurons
{unit}
{cell}
{node}
{perceptron}



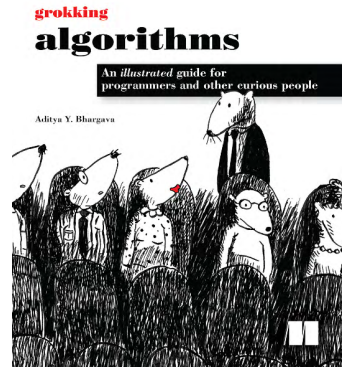


{Algorithm}

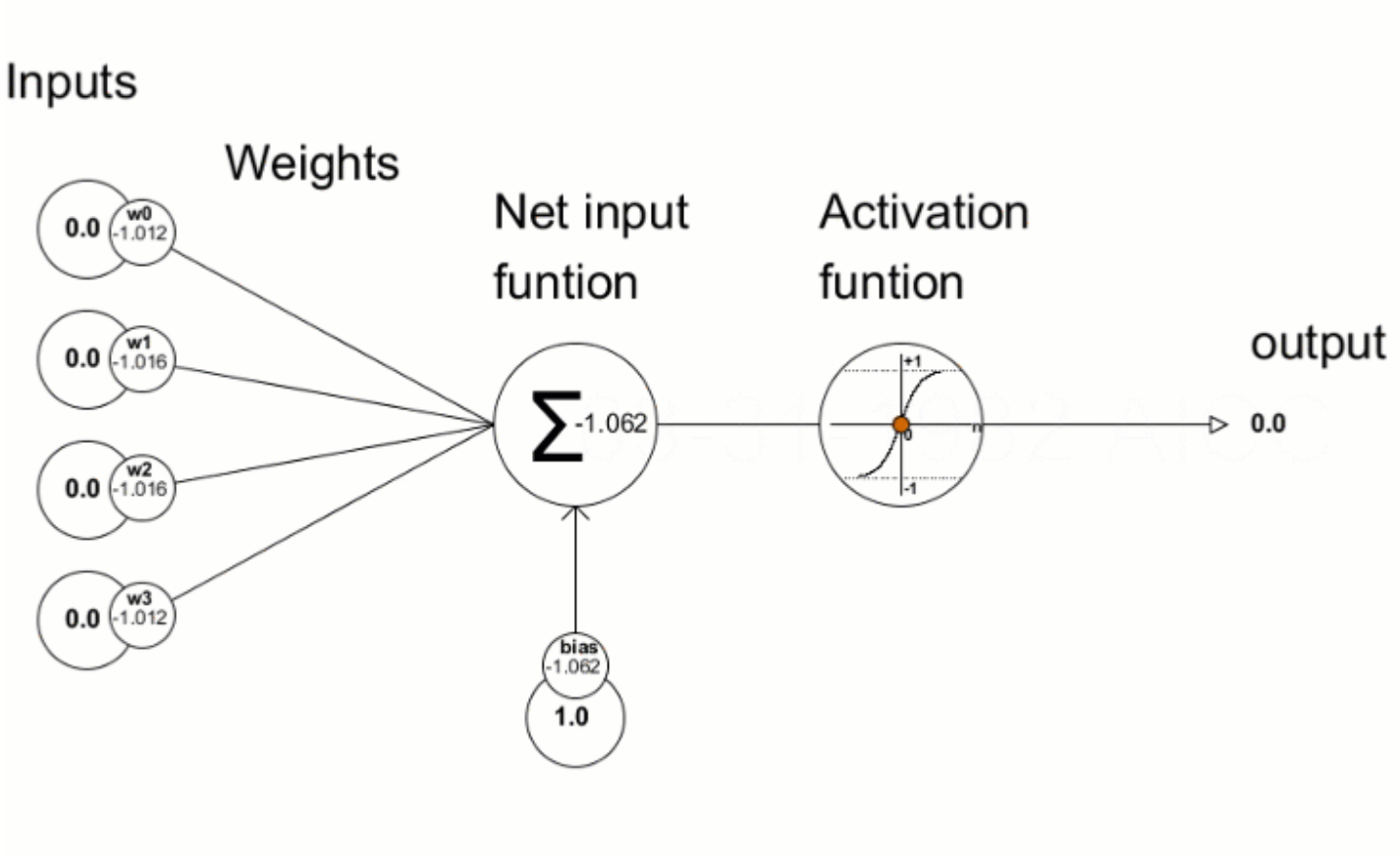
HOGESCHOOL ROTTERDAM

CREATING

Step by step process or recipe
describing
how to solve a problem and/or
complete a task,
which will always give
identical end results



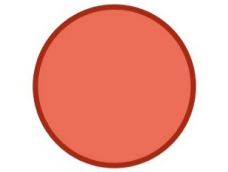
{Artificial Neurons}



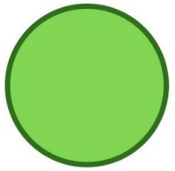
{NN Layers}

HOGESCHOOL ROTTERDAM

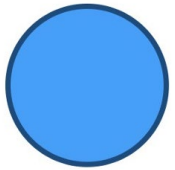
Neural Network {NN} Layer Architecture



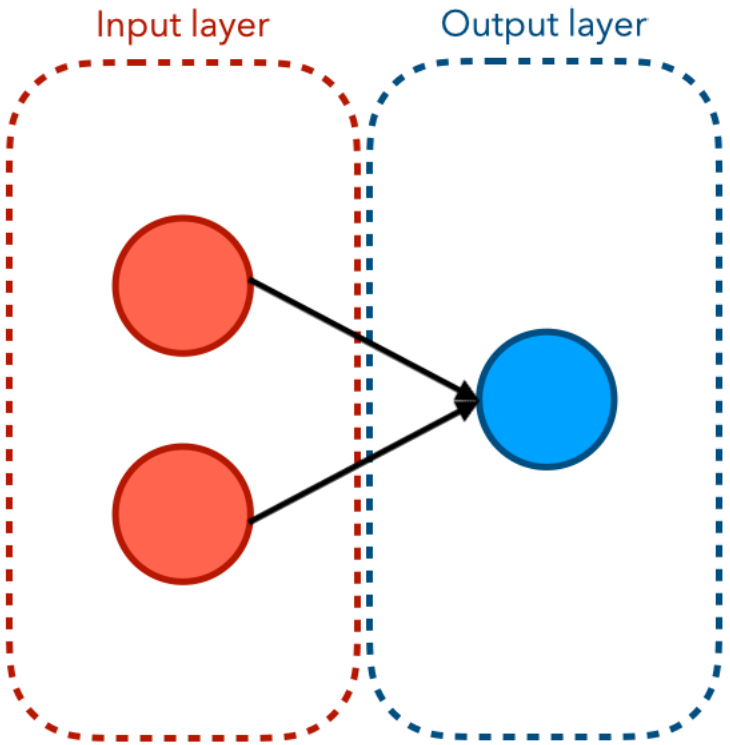
Input neuron



Hidden neuron



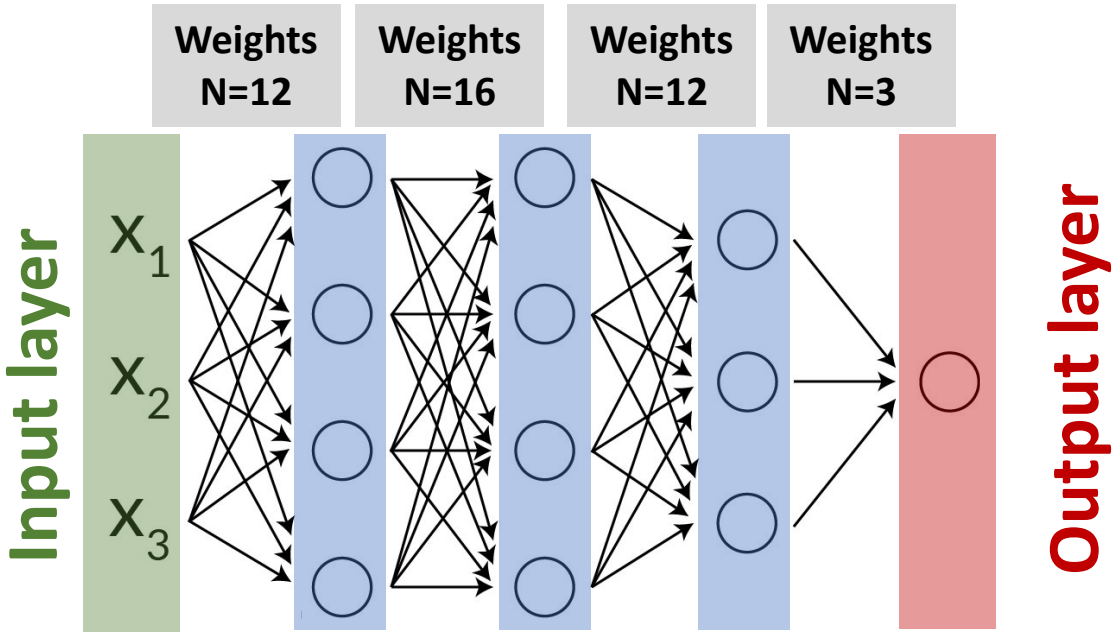
Output neuron



{NN Layers}

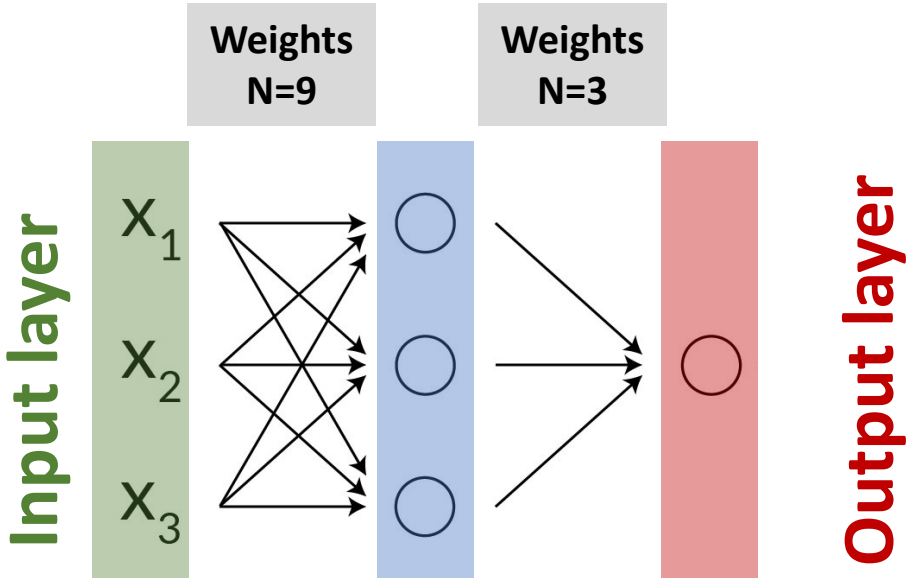
HOGESCHOOL ROTTERDAM

Neural Network {NN} Layer Architecture



hidden layers N>2

Deep network

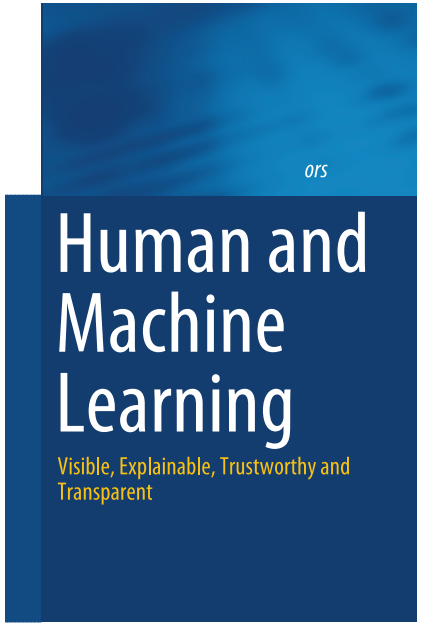


Hidden layer N=1

Shallow network

{Human-in-the-Loop}

HOGESCHOOL ROTTERDAM



AI = ML + TD + HITL

Artificial Intelligence: in contrast to natural intelligence, it is *the ability of computer systems to perform tasks or actions that would normally require a human*

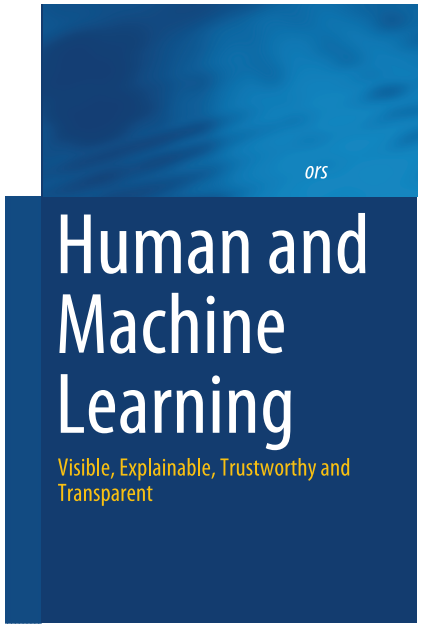
Machine Learning: *the ability of computer systems to use algorithms and statistical models to perform tasks without explicit instruction, through patterns and inferences*

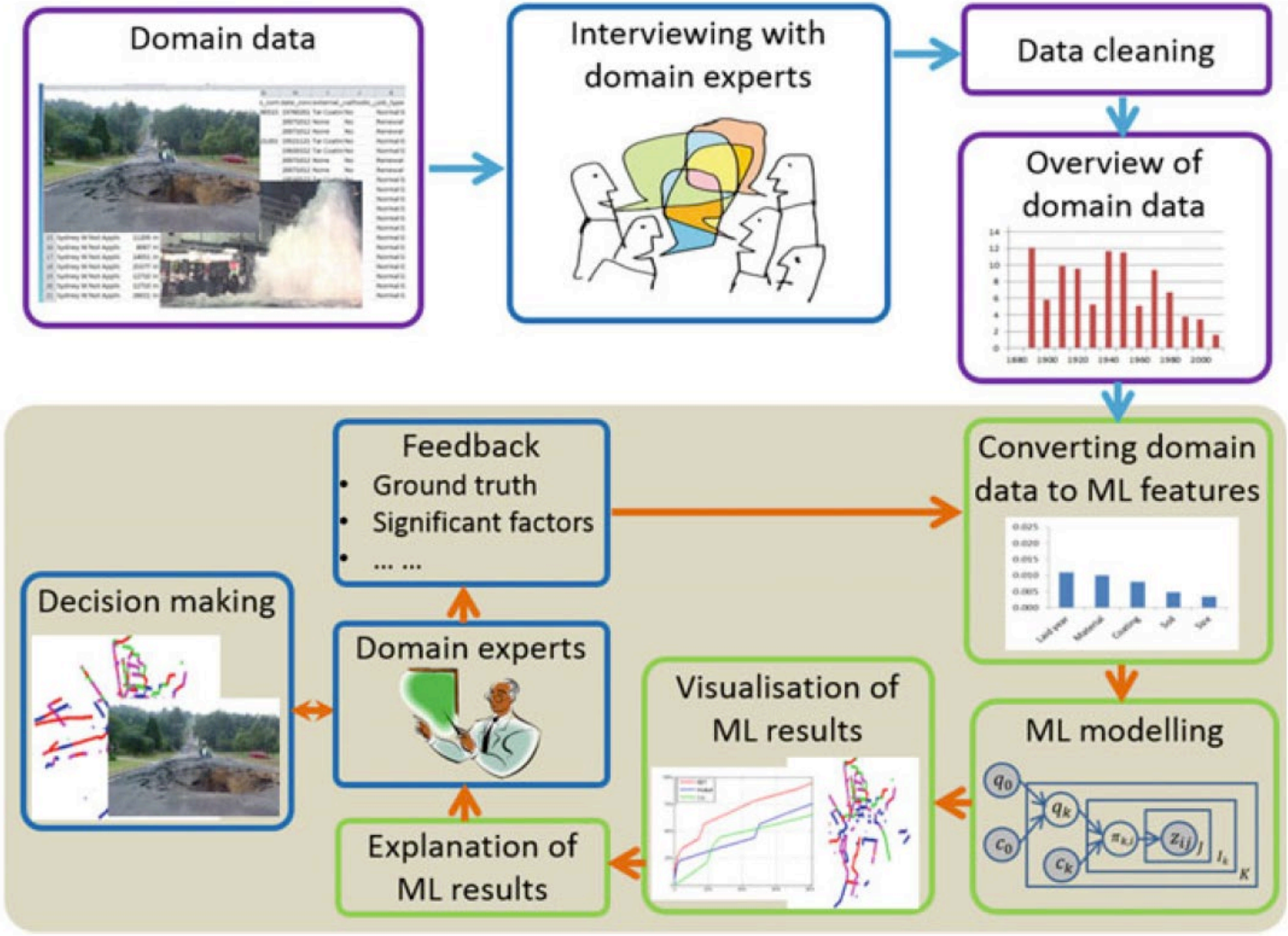
Training Data: *the data used to train a machine learning algorithm to perform a task in supervised machine learning*

Human in the Loop: *the involvement of a human in training a machine learning algorithm*

{Modeling}

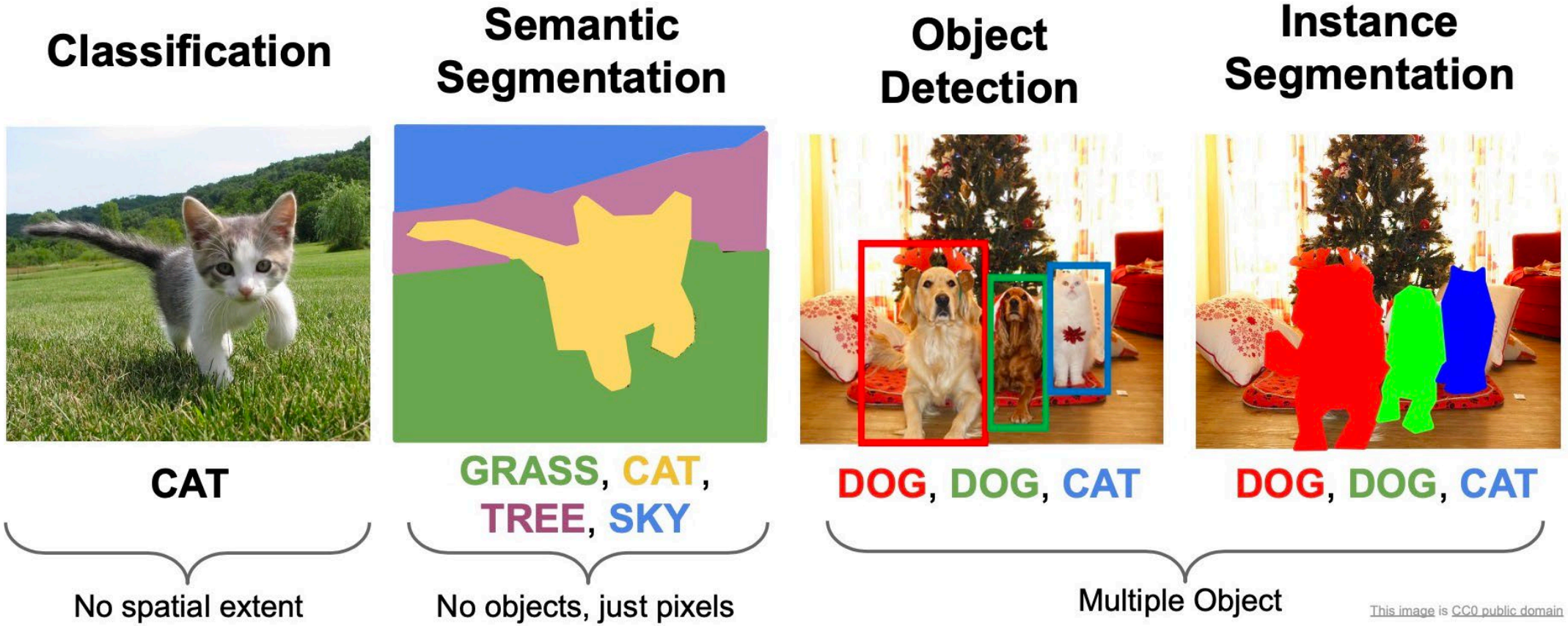
HOGESCHOOL ROTTERDAM





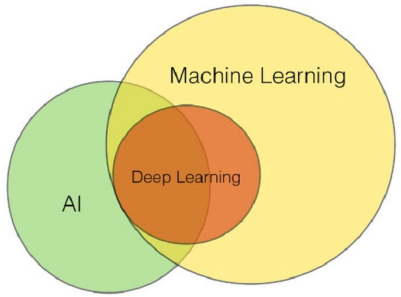
{Modeling}

HOGESCHOOL ROTTERDAM



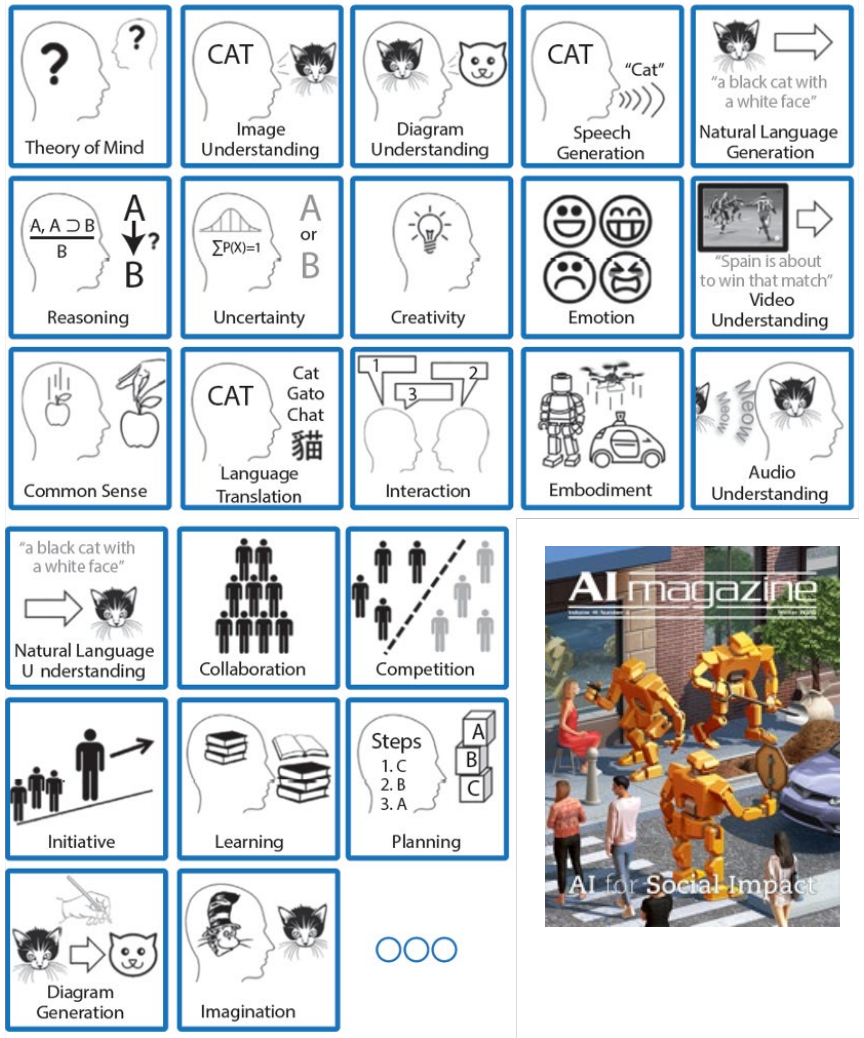
{AI=ML=DL}

AI enabled through {DL} must be understood as any form of Machine Learning {ML} technology mimicking & automatizing tasks which otherwise require



human perception, cognition and/or motor skills

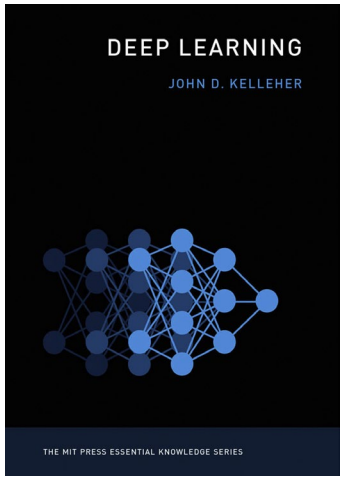
https://doi.org/10.1609/aimag.v37i1.2643

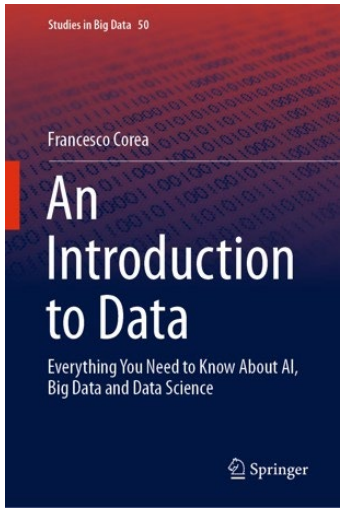


https://robfvdw.medium.com/the-world-wide-web-ai-safari-b2e4f7f90647

{DL + DNNs}

HOGESCHOOL ROTTERDAM
CREATING





Deep learning {DL} must be understood as a major Machine Learning {ML} subdomain:

Crafting Deep Neural Networks {DNNs} that can attain human-level performances on challenging cognitive tasks.

{DNNs} can Recognize Speech or Human Poses & Faces; Translate Text in real time at High Levels of Performance.

{Top-down}

Top-down Encoding Capacity increases by adding hidden layers

HOGESCHOOL ROTTERDAM

What are the limits of deep learning?

The much-ballyhooed artificial intelligence approach boasts impressive feats but still falls short of human brainpower. Researchers are determined to figure out what's missing.

M. Mitchell Waldrop, *Science Writer*

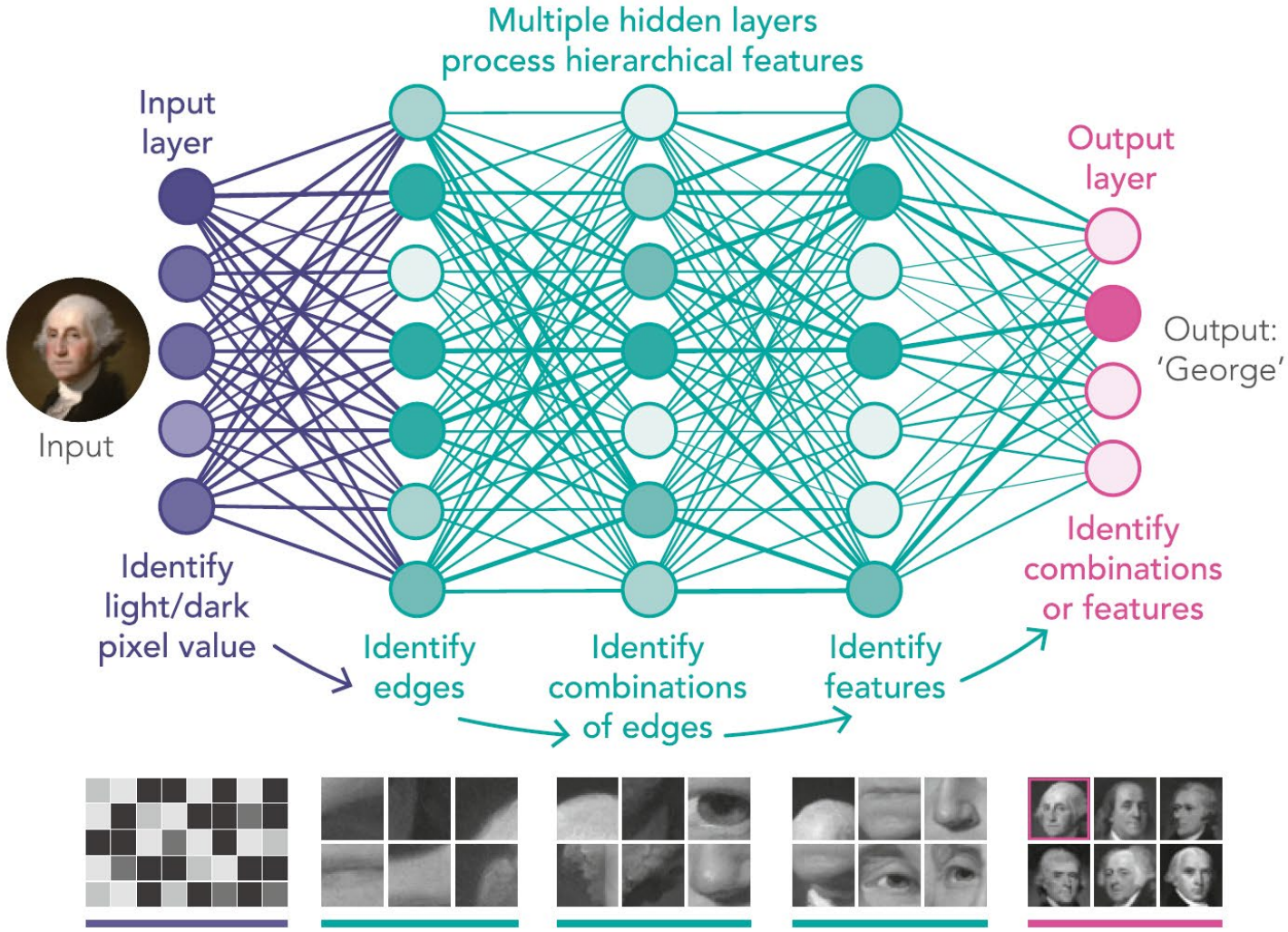
There's no mistaking the image: It's a banana—a big, ripe, bright-yellow banana. Yet the artificial intelligence (AI) identifies it as a toaster, even though it was trained with the same powerful and oft-publicized deep-learning techniques that have produced a white-hot revolution in driverless cars, speech understanding, and a multitude of other AI applications. That means the AI was shown several thousand photos of bananas, slugs, snails, and similar-looking objects, like so many flash cards, and then drilled on the answers until it had the classification down cold. And yet this advanced system was quite easily confused—all it took was a little day-glow sticker, digitally pasted in one corner of the image.

This example of what deep-learning researchers call an "adversarial attack," discovered by the Google Brain team in Mountain View, CA (1), highlights just how far AI still has to go before it remotely approaches human capabilities. "I initially thought that adversarial examples were just an annoyance," says Geoffrey Hinton, a computer scientist at the University of Toronto and one of the pioneers of deep learning. "But I now think they're probably quite profound. They tell us that we're doing something wrong."

That's a widely shared sentiment among AI practitioners, any of whom can easily rattle off a long list of deep learning's drawbacks. In addition to its vulnerability

Apparent shortcomings in deep-learning approaches have raised concerns among researchers and the general public as technologies such as driverless cars, which use deep-learning techniques to navigate, get involved in well-publicized mishaps. Image credit: Shutterstock.com/MONOPOLY919.

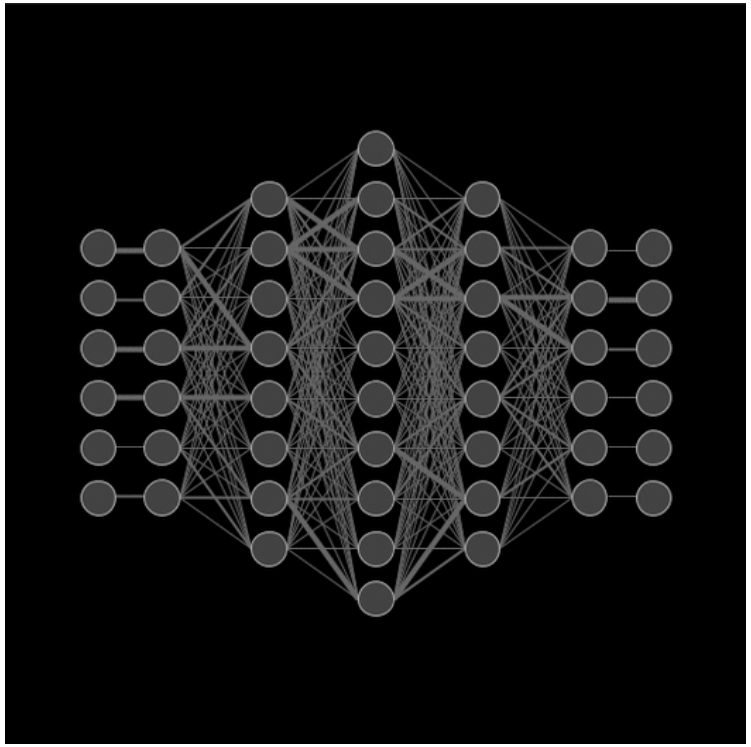
www.pnas.org/cgi/doi/10.1073/pnas.1821594116

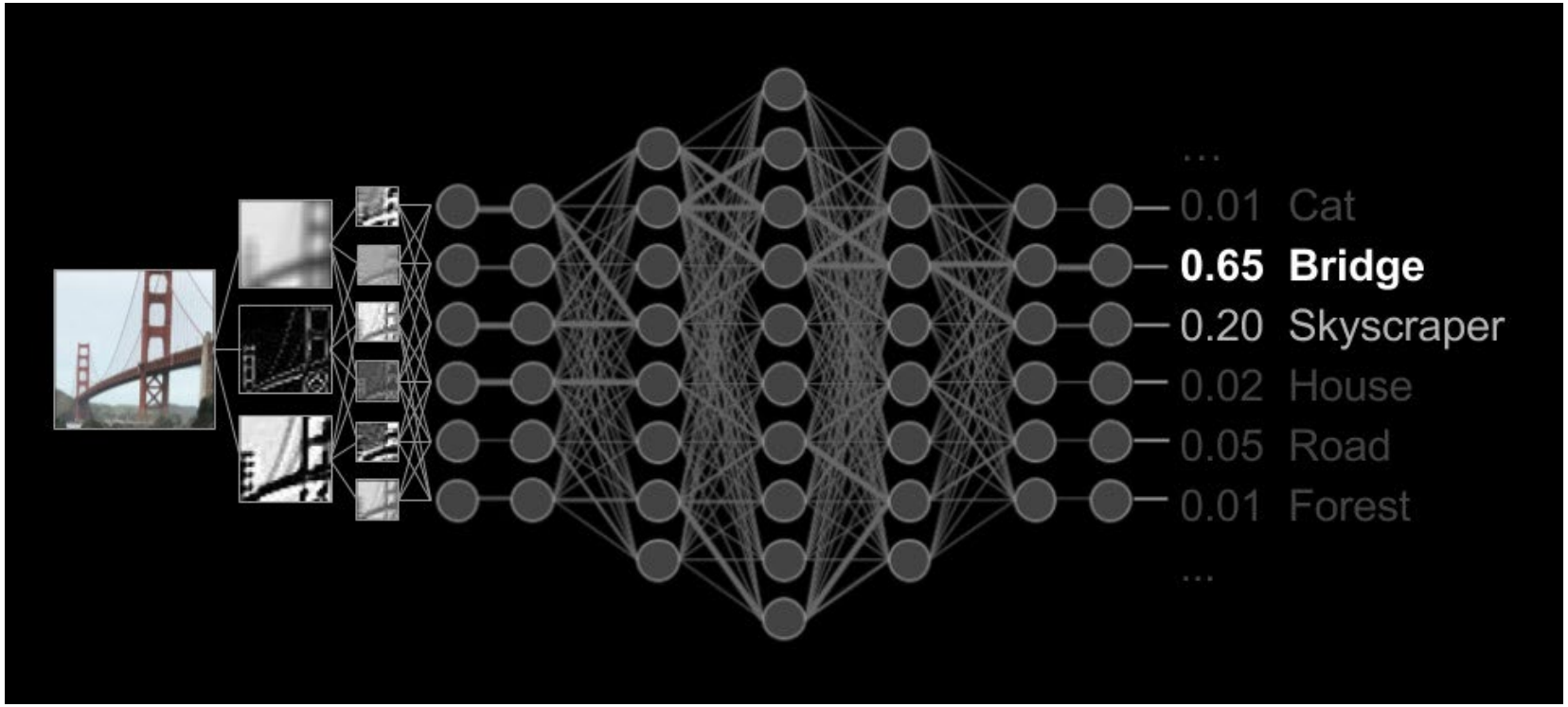


{Top-down}

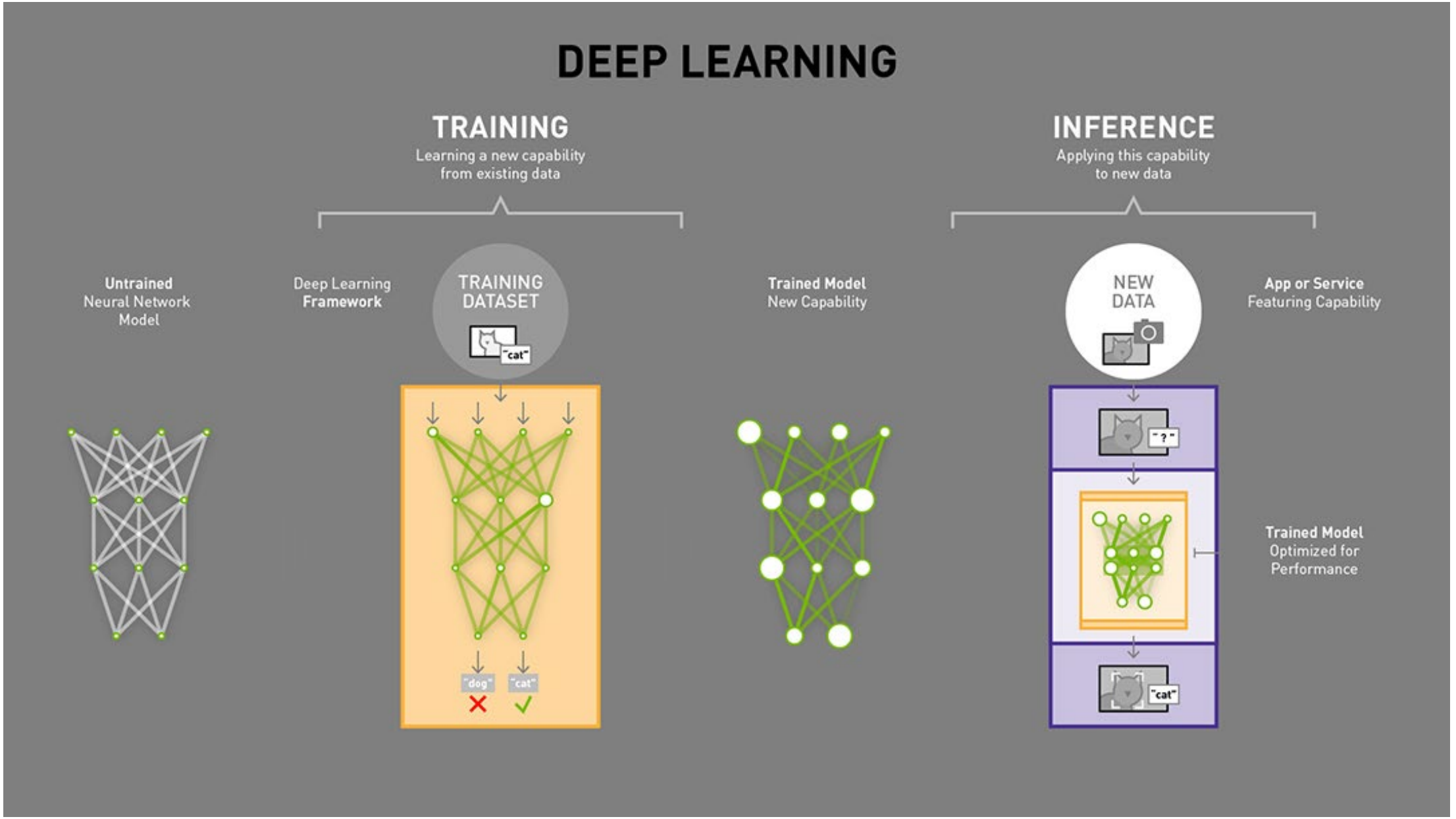
Top-down Encoding Capacity increases by adding hidden layers

HOGESCHOOL ROTTERDAM

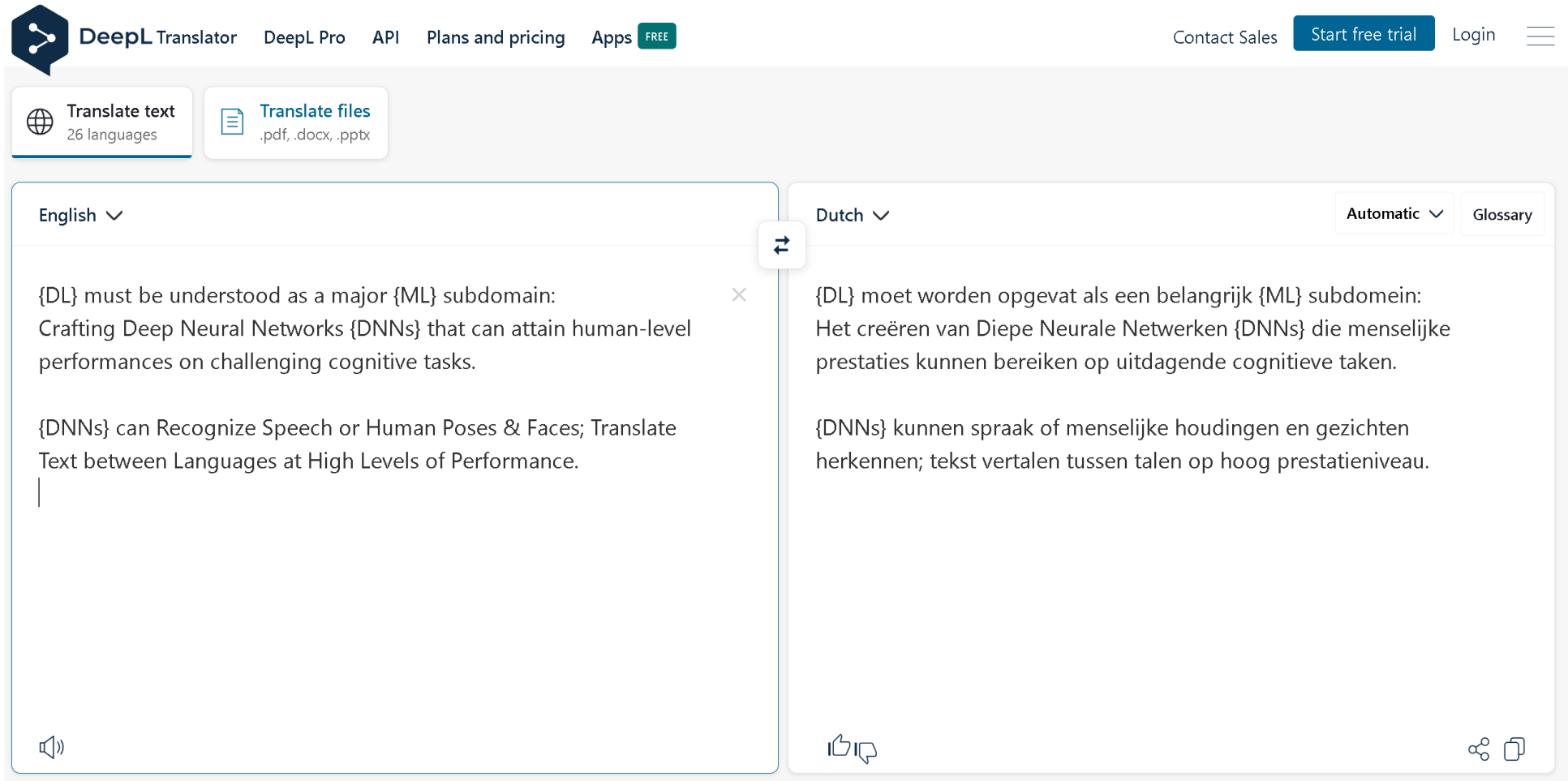




HOGESCHOOL ROTTERDAM



{Human-level performance}



https://www.deepl.com/translator

{Human-level performance}

Probeer Speech to Text uit met deze demo-app, ontwikkeld op basis van onze JavaScript-SDK

Taal

Dutch (Netherlands)

Automatische interpunctie

Spreken

Bestand uploaden

Uw spraakgegevens worden niet opgeslagen

Ontdek hoe u Speech to Text in uw apps en producten gebruikt >

Verken meer aspecten van uw Speech to Text-uitvoer met het programma zonder code in Speech Studio >

Kies de knop Spreken aan de linkerkant en begin met spreken. De spraakservice retourneert herkenningsresultaten terwijl u spreekt. Als u verschillende talen spreekt, kunt u elke taal uitproberen die door de spraakservice wordt ondersteund. U kunt ook bestanden uploaden om de spraakservice voor uw specifieke gebruiksscenario's te testen. Raadpleeg onze documentatie en ontdek hoe u de spraak-naar-tekstfunctie in uw oplossingen inbouwt.

https://azure.microsoft.com/nl-nl/services/cognitive-services/speech-to-text/#features

{DL} represents an {AI} breakthrough

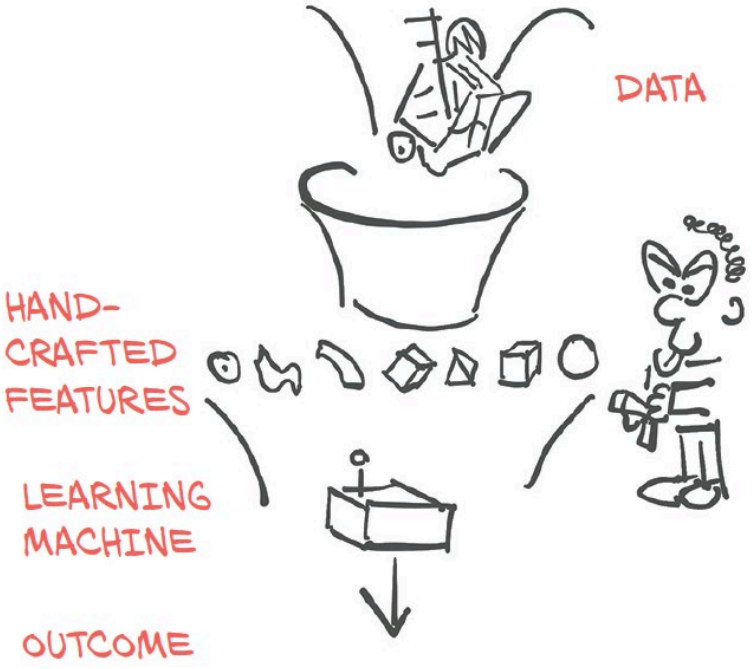
Paradigm-Shift

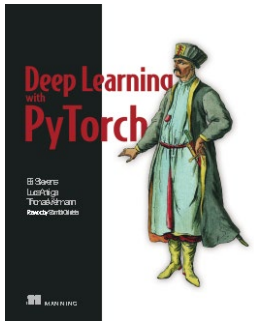
{AI Paradigm-shift}

HOGESCHOOL ROTTERDAM

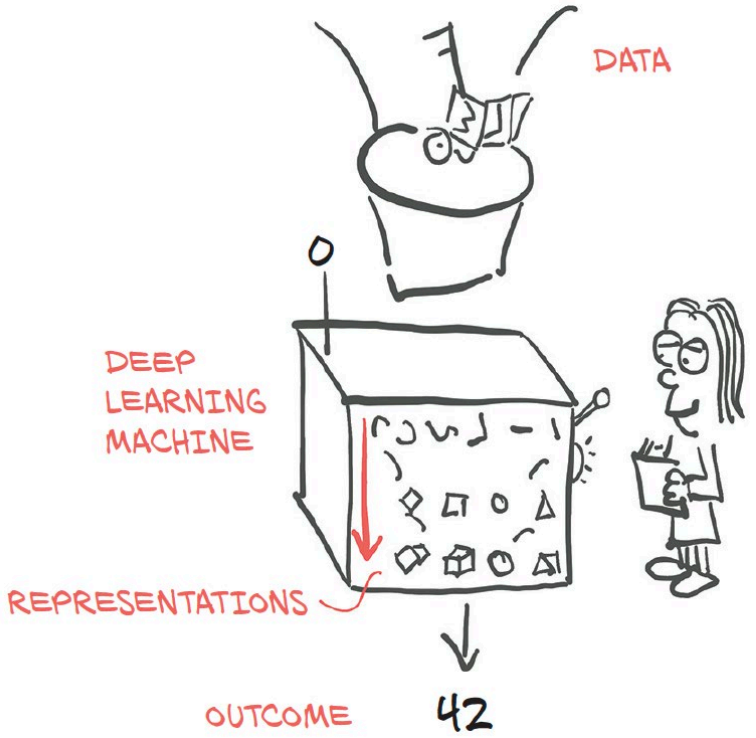
More data, parameters & computing power | Less human-in-the-loop

Machine Learning Paradigm {ML}





Deep Learning Paradigm {DL}

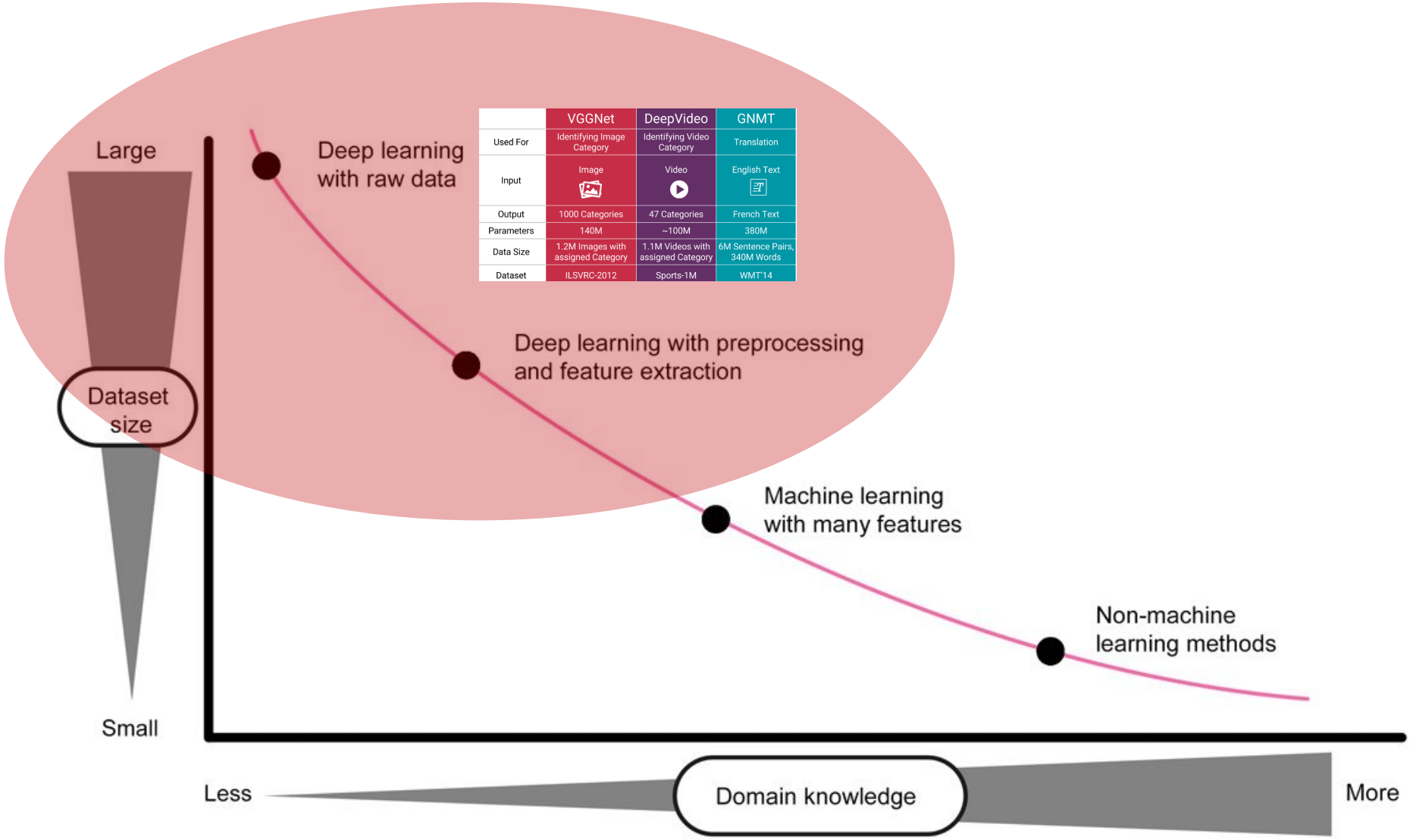


{DL} represents an troublesom Paradigm-Shift

{Big-data}

Big-data is needed to avoid hand-crafted feature extraction

HOGESCHOOL ROTTERDAM
CREATING



	VGGNet	DeepVideo	GNMT
Used For	Identifying Image Category	Identifying Video Category	Translation
Input	Image	Video	English Text
Output	1000 Categories	47 Categories	French Text
Parameters	140M	~100M	380M
Data Size	1.2M Images with assigned Category	1.1M Videos with assigned Category	6M Sentence Pairs, 340M Words
Dataset	ILSVRC-2012	Sports-1M	WMT'14

A Unified Approach to Interpreting Model Predictions

Scott M. Lundberg
Paul G. Allen School of Computer Science
University of Washington
Seattle, WA 98105
slund1@cs.washington.edu

Su-In Lee
Paul G. Allen School of Computer Science
Department of Genome Sciences
University of Washington
Seattle, WA 98105
suinlee@cs.washington.edu

Abstract

Understanding why a model makes a certain prediction can be as crucial as the prediction's accuracy in many applications. However, the highest accuracy for large modern datasets is often achieved by complex models that even experts struggle to interpret, such as ensemble or deep learning models, creating a tension between *accuracy* and *interpretability*. In response, various methods have recently been proposed to help users interpret the predictions of complex models, but it is often unclear how these methods are related and when one method is preferable over another. To address this problem, we present a unified framework for interpreting predictions, SHAP (SHapley Additive exPlanations). SHAP assigns each feature an importance value for a particular prediction. Its novel components include: (1) the identification of a new class of additive feature importance measures, and (2) theoretical results showing there is a unique solution in this class with a set of desirable properties. The new class unifies six existing methods, notable because several recent methods in the class lack the proposed desirable properties. Based on insights from this unification, we present new methods that show improved computational performance and/or better consistency with human intuition than previous approaches.

1 Introduction

The ability to correctly interpret a prediction model's output is extremely important. It engenders appropriate user trust, provides insight into how a model may be improved, and supports understanding of the process being modeled. In some applications, simple models (e.g., linear models) are often preferred for their ease of interpretation, even if they may be less accurate than complex ones. However, the growing availability of big data has increased the benefits of using complex models, so bringing to the forefront the trade-off between accuracy and interpretability of a model's output. A wide variety of different methods have been recently proposed to address this issue [5, 8, 9, 3, 4, 1]. But an understanding of how these methods relate and when one method is preferable to another is still lacking.

Here, we present a novel unified approach to interpreting model predictions.[1] Our approach leads to three potentially surprising results that bring clarity to the growing space of methods:

1. We introduce the perspective of viewing *any* explanation of a model's prediction as a model itself, which we term the *explanation model*. This lets us define the class of *additive feature attribution methods* (Section 2), which unifies six current methods.

[1] https://github.com/slundberg/shap

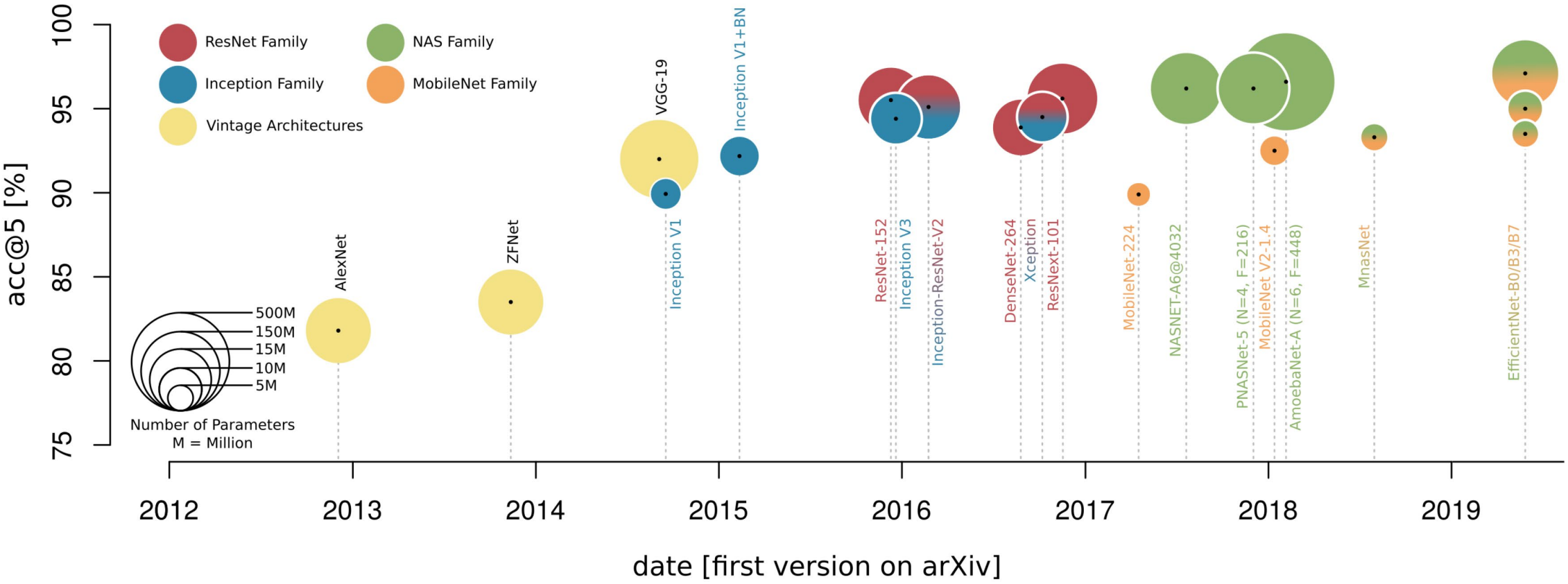
31st Conference on Neural Information Processing Systems (NIPS 2017), Long Beach, CA, USA.

https://proceedings.neurips.cc/paper/2017/file/8a20a8621978632d76c43dfd28b67767-Paper.pdf

{Weights}

Performance increases by adding learnable parameters {weight's}

HOGESCHOOL ROTTERDAM



How to calculate the number of learnable parameters?

https://doi.org/10.3390/rs12101667

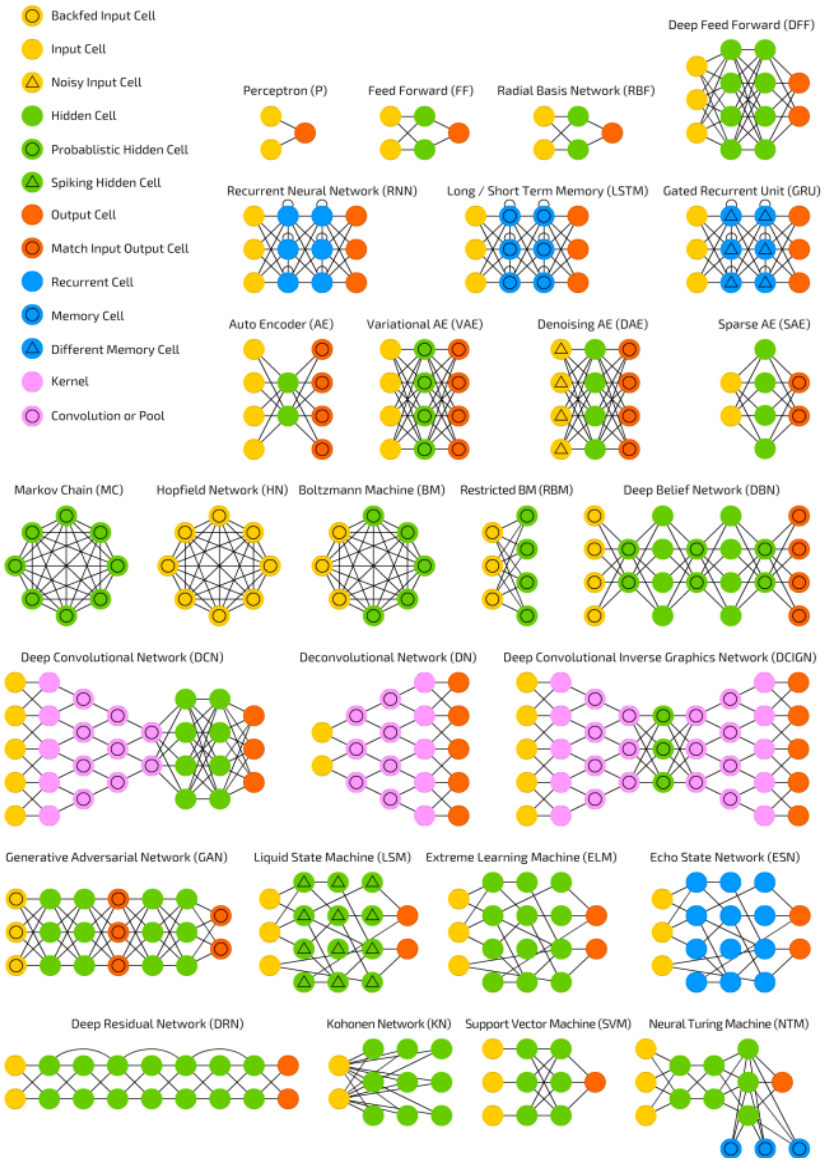
{Topology}

Topology of a neural network refers to the way artificial neurons are connected to form a network.

Form follows function!

The topology of a network determines the degree of perplexity of the tasks it can perform.

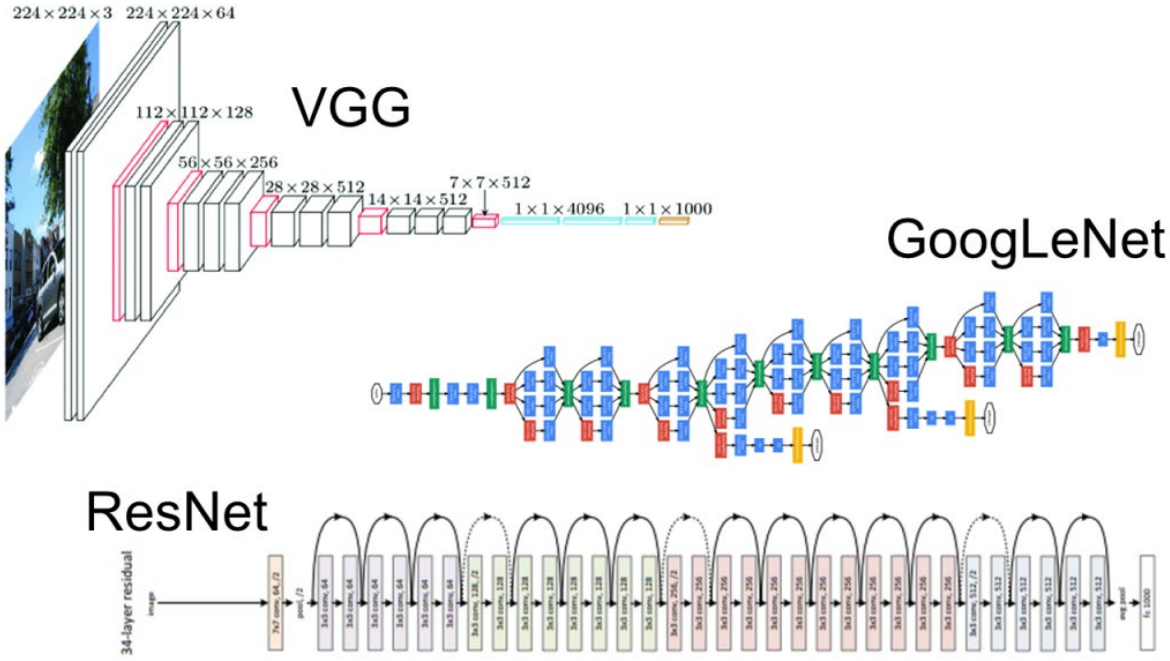
https://pub.towardsai.net/main-types-of-neural-networks-and-its-applications-tutorial-734480d7ec8e



https://www.mdpi.com/2504-3900/47/1/9

{Perplexity}

Toplogical complex Neural Networks Perform Better: have Low Perplexity



Model	Size	Top-1 Accuracy	Top-5 Accuracy	Parameters	Depth
Xception	88 MB	0.790	0.945	22,910,480	126
VGG16	528 MB	0.713	0.901	138,357,544	23
VGG19	549 MB	0.713	0.900	143,667,240	26
ResNet50	98 MB	0.749	0.921	25,636,712	-
ResNet101	171 MB	0.764	0.928	44,707,176	-
ResNet152	232 MB	0.766	0.931	60,419,944	-
ResNet50V2	98 MB	0.760	0.930	25,613,800	-
ResNet101V2	171 MB	0.772	0.938	44,675,560	-
ResNet152V2	232 MB	0.780	0.942	60,380,648	-
InceptionV3	92 MB	0.779	0.937	23,851,784	159
InceptionResNetV2	215 MB	0.803	0.953	55,873,736	572
MobileNet	16 MB	0.704	0.895	4,253,864	88
MobileNetV2	14 MB	0.713	0.901	3,538,984	88
DenseNet121	33 MB	0.750	0.923	8,062,504	121
DenseNet169	57 MB	0.762	0.932	14,307,880	169
DenseNet201	80 MB	0.773	0.936	20,242,984	201
NASNetMobile	23 MB	0.744	0.919	5,326,716	-
NASNetLarge	343 MB	0.825	0.960	88,949,818	-
EfficientNetB0	29 MB	-	-	5,330,571	-
EfficientNetB1	31 MB	-	-	7,856,239	-
EfficientNetB2	36 MB	-	-	9,177,569	-
EfficientNetB3	48 MB	-	-	12,320,535	-
EfficientNetB4	75 MB	-	-	19,466,823	-
EfficientNetB5	118 MB	-	-	30,562,527	-
EfficientNetB6	166 MB	-	-	43,265,143	-
EfficientNetB7	256 MB	-	-	66,658,687	-

Why the simple strategy of scaling up neural networks has been so effective?

[2105.12806] A Universal Law of Robustness via Isoperimetry (arxiv.org)

Keras https://keras.io/api/applications/

{Self-attention}

Transformer DNNs outperform GANs, CNN & RNNs by adding a stacked intermediate neural net topologies that attend to themselves called Transformers

HOGESCHOOL ROTTERDAM

Attention Is All You Need

Ashish Vaswani* Google Brain avaswani@google.com
Noam Shazeer* Google Brain noam@google.com
Niki Parmar* Google Research nikip@google.com
Jakob Uszkoreit* Google Research usz@google.com
Llion Jones* Google Research llion@google.com
Aidan N. Gomez* † University of Toronto aidan@cs.toronto.edu
Łukasz Kaiser* Google Brain lukaszkaiser@google.com
Illia Polosukhin* ‡ illia.polosukhin@gmail.com

Abstract

The dominant sequence transduction models are based on complex recurrent or convolutional neural networks that include an encoder and a decoder. The best performing models also connect the encoder and decoder through an attention mechanism. We propose a new simple network architecture, the Transformer, based solely on attention mechanisms, dispensing with recurrence and convolutions entirely. Experiments on two machine translation tasks show these models to be superior in quality while being more parallelizable and requiring significantly less time to train. Our model achieves 28.4 BLEU on the WMT 2014 English-to-German translation task, improving over the existing best results, including ensembles, by over 2 BLEU. On the WMT 2014 English-to-French translation task, our model establishes a new single-model state-of-the-art BLEU score of 41.8 after training for 3.5 days on eight GPUs, a small fraction of the training costs of the best models from the literature. We show that the Transformer generalizes well to other tasks by applying it successfully to English constituency parsing both with large and limited training data.

1 Introduction

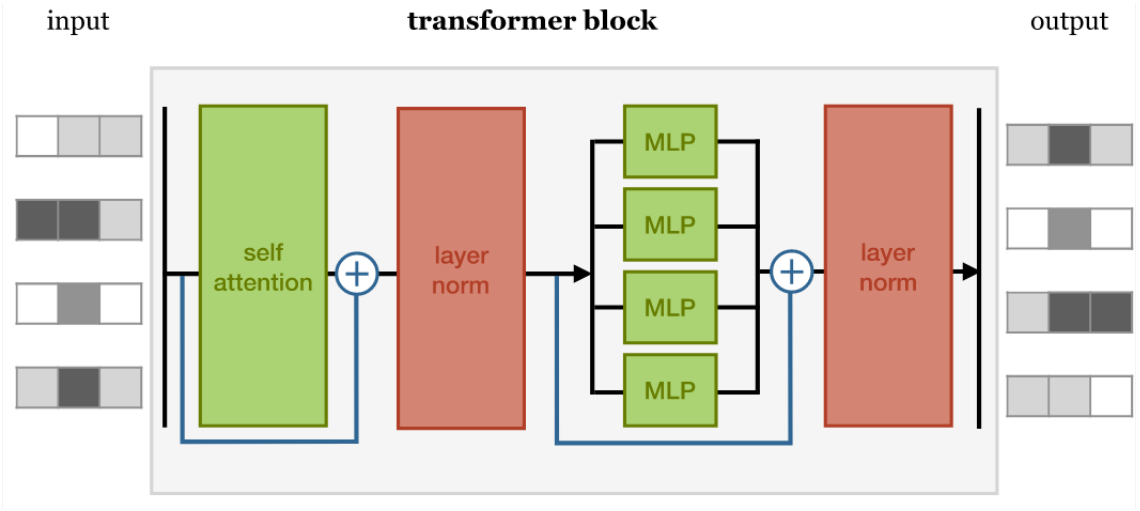
Recurrent neural networks, long short-term memory [13] and gated recurrent [7] neural networks in particular, have been firmly established as state of the art approaches in sequence modeling and

*Equal contribution. Listing order is random. Jakob proposed replacing RNNs with self-attention and started the effort to evaluate this idea. Ashish, with Illia, designed and implemented the first Transformer models and has been crucially involved in every aspect of this work. Noam proposed scaled dot-product attention, multi-head attention and the parameter-free position representation and became the other person involved in nearly every detail. Niki designed, implemented, tuned and evaluated countless model variants in our original codebase and tensor2tensor. Llion also experimented with novel model variants, was responsible for our initial codebase, and efficient inference and visualizations. Lukasz and Aidan spent countless long days designing various parts of and implementing tensor2tensor, replacing our earlier codebase, greatly improving results and massively accelerating our research.

†Work performed while at Google Brain.

‡Work performed while at Google Research.

31st Conference on Neural Information Processing Systems (NIPS 2017), Long Beach, CA, USA.



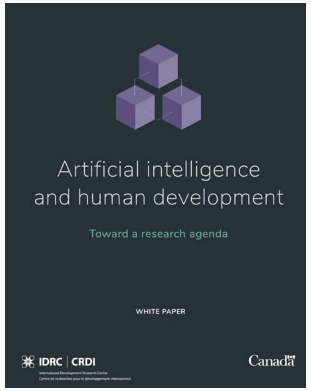
GitHub - pbloem/former: Simple transformer implementation from scratch in pytorch.
Transformers from scratch | peterbloem.nl

Mukhamediev, R. I., Symagulov, A., Kuchin, Y., Yakunin, K., & Yelis, M. (2021). From Classical Machine Learning to Deep Neural Networks: A Simplified Scientometric Review. *Applied Sciences*, *11*(12), 5541. MDPI AG. Retrieved from http://dx.doi.org/10.3390/app11125541

{Skewed}

HOGESCHOOL ROTTERDAM

Big-Data is Inherently Skewed



When artificial intelligence learns from humans, it's bad

James Crowder*

Learned bias

ML algorithms are not biased in and of themselves; they *learn* to be biased. This *algorithmic bias* (Danks and London 2017) has received a great deal of attention. It occurs when the learning algorithm is trained on biased datasets and subsequently "accurately" learns the patterns of bias inherent in the data (see, e.g., Caliskan et al. 2017). In some cases, the learned representations within ML algorithms can even exaggerate these biases (Zhao et al. 2017). Algorithmic bias has two sources: incomplete datasets and datasets that represent biased social phenomena.

Incomplete datasets are those that are not representative of the entire range of potential examples. Consequently, an algorithm trained on an incomplete dataset will perform poorly when given an example that falls outside the scope of the available data. For example, a facial recognition algorithm that is not trained on a wide variety of skin colours might not function accurately for faces of all skin tones. Indeed, one AI researcher with dark skin discovered that an otherwise functional facial recognition algorithm failed to recognize her face unless she put on a white mask[45] . Many similar examples have been reported.[46]

When critical decisions are made based on input from an algorithm trained on a database that is not representative of the entire user population, the results can impact health and well-being. Most clinical trials have highly selective criteria that exclude women (especially pregnant women), the elderly, and those with conditions beyond those being studied. Thus, participants tend to be white males.[47] In some cases, the findings of these studies do not generalize well across the broader population, with outcomes potentially compromised for individuals not represented in the research. As a result, AI algorithms trained on this data are

* bit.ly/2HFetvY

Photo by Marius Masalar

{Disparities}

Big Data causes racial & gender disparities

The Race Gap in Speech Recognition Technology

Researchers from Stanford University found that state-of-the-art ASR systems exhibited significant racial and gender disparity—they misunderstand Black speakers twice as often as white speakers. In the paper, titled "Racial Disparities in Automated Speech Recognition," authors ran thousands of audio snippets of white and Black speakers, transcribed from interviews conducted with 42 white speakers and 73 Black speakers, through leading speech-to-text services by Amazon, Apple, Google, IBM, and Microsoft.

The results suggest that, on average, systems made 19 errors every hundred words for white speakers and 35 errors for Black speakers—nearly twice as many. Moreover, the systems performed particularly poorly for Black men, with more than 40 errors for every hundred words (Figure 2.5.3). The breakdown by ASR systems shows that gaps are similar across companies (Figure 2.5.4). This research emphasizes the importance of addressing the bias of AI technologies and ensuring equity as they become mature and deployed.

TESTINGS on LEADING SPEECH-to-TEXT SERVICES: WORD ERROR RATE by RACE and GENDER, 2019

Source: Koenecke et al., 2020 | Chart: 2021 AI Index Report

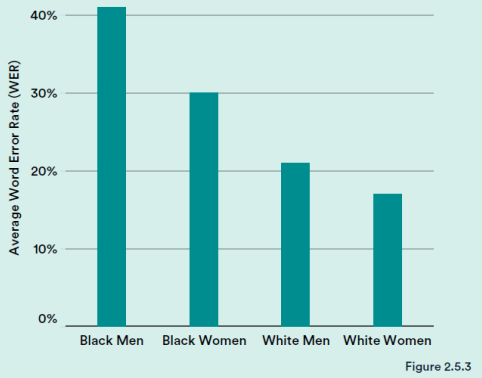


Figure 2.5.3

TESTINGS on LEADING SPEECH-to-TEXT SERVICES: WORD ERROR RATE by SERVICE and RACE, 2019

Source: Koenecke et al., 2020 | Chart: 2021 AI Index Report

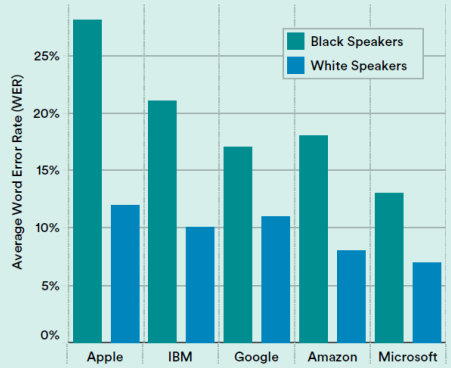


Figure 2.5.4

{Augmentation}

HOGESCHOOL ROTTERDAM

Big Data that is *not* augmented causes Overfitting

SURVEY PAPER Open Access

A survey on Image Data Augmentation for Deep Learning

Connor Shorten* and Taghi M. Khoshgoftaar

*Correspondence:
cshorten2015@fau.edu
Department of Computer and Electrical Engineering and Computer Science, Florida Atlantic University, Boca Raton, USA

Abstract

Deep convolutional neural networks have performed remarkably well on many Computer Vision tasks. However, these networks are heavily reliant on big data to avoid overfitting. Overfitting refers to the phenomenon when a network learns a function with very high variance such as to perfectly model the training data. Unfortunately, many application domains do not have access to big data, such as medical image analysis. This survey focuses on Data Augmentation, a data-space solution to the problem of limited data. Data Augmentation encompasses a suite of techniques that enhance the size and quality of training datasets such that better Deep Learning models can be built using them. The image augmentation algorithms discussed in this survey include geometric transformations, color space augmentations, kernel filters, mixing images, random erasing, feature space augmentation, adversarial training, generative adversarial networks, neural style transfer, and meta-learning. The application of augmentation methods based on GANs are heavily covered in this survey. In addition to augmentation techniques, this paper will briefly discuss other characteristics of Data Augmentation such as test-time augmentation, resolution impact, final dataset size, and curriculum learning. This survey will present existing methods for Data Augmentation, promising developments, and meta-level decisions for implementing Data Augmentation. Readers will understand how Data Augmentation can improve the performance of their models and expand limited datasets to take advantage of the capabilities of big data.

Keywords: Data Augmentation, Big data, Image data, Deep Learning, GANs

Introduction

Deep Learning models have made incredible progress in discriminative tasks. This has been fueled by the advancement of deep network architectures, powerful computation, and access to big data. Deep neural networks have been successfully applied to Computer Vision tasks such as image classification, object detection, and image segmentation thanks to the development of convolutional neural networks (CNNs). These neural networks utilize parameterized, sparsely connected kernels which preserve the spatial characteristics of images. Convolutional layers sequentially downsample the spatial resolution of images while expanding the depth of their feature maps. This series of convolutional transformations can create much lower-dimensional and more useful representations of images than what could possibly be hand-crafted. The success of CNNs has spiked interest and optimism in applying Deep Learning to Computer Vision tasks.

Springer Open

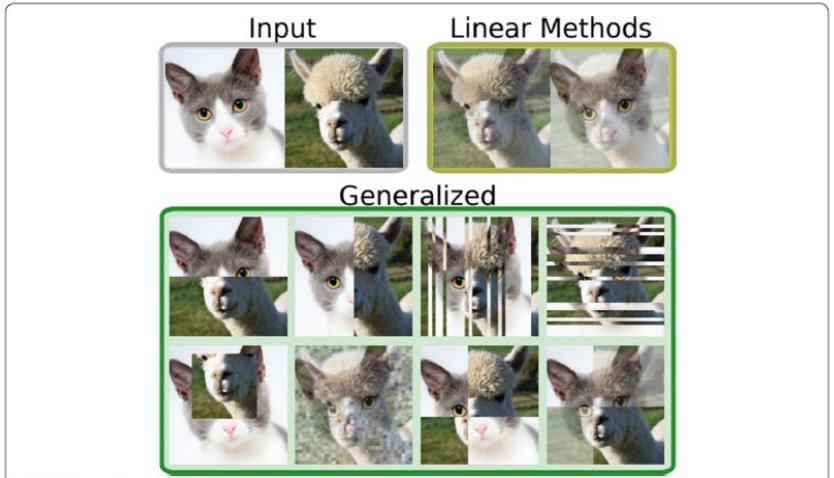
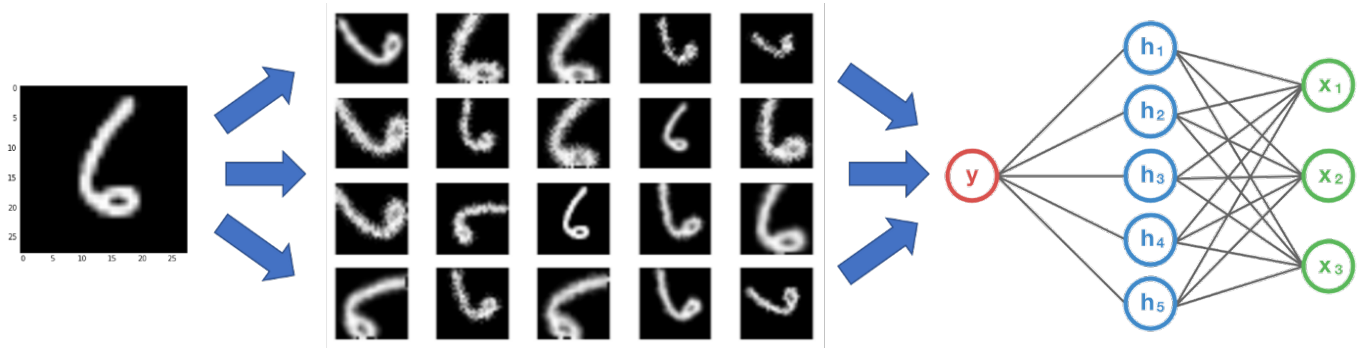


Fig. 9 Non-linearly mixing images [66]



{Regularization}

HOGESCHOOL ROTTERDAM

The process of adding information to adress ill-posed problems: How to increase sparsity of Dense DNNs?

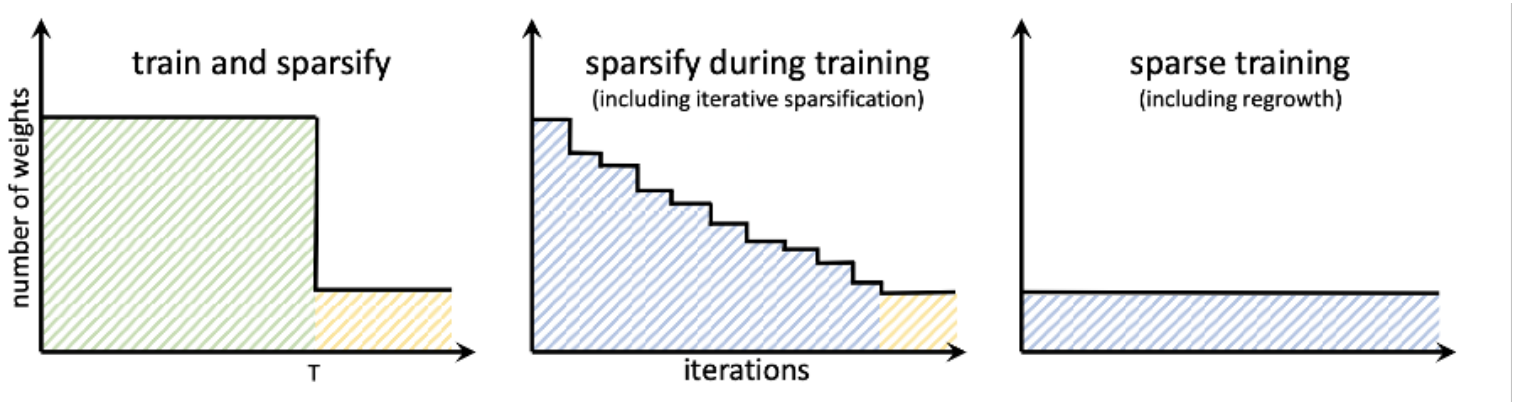
Sparsity in Deep Learning: Pruning and growth for efficient inference and training in neural networks

TORSTEN HOEFLER, ETH Zürich, Switzerland
DAN ALISTARH, IST Austria, Austria
TAL BEN-NUN, ETH Zürich, Switzerland
NIKOLI DRYDEN, ETH Zürich, Switzerland
ALEXANDRA PESTE, IST Austria, Austria

The growing energy and performance costs of deep learning have driven the community to reduce the size of neural networks by selectively pruning components. Similarly to their biological counterparts, sparse networks generalize just as well, if not better than, the original dense networks. Sparsity can reduce the memory footprint of regular networks to fit mobile devices, as well as shorten training time for ever growing networks. In this paper, we survey prior work on sparsity in deep learning and provide an extensive tutorial of sparsification for both inference and training. We describe approaches to remove and add elements of neural networks, different training strategies to achieve model sparsity, and mechanisms to exploit sparsity in practice. Our work distills ideas from more than 300 research papers and provides guidance to practitioners who wish to utilize sparsity today, as well as to researchers whose goal is to push the frontier forward. We include the necessary background on mathematical methods in sparsification, describe phenomena such as early structure adaptation, the intricate relations between sparsity and the training process, and show techniques for achieving acceleration on real hardware. We also define a metric of pruned parameter efficiency that could serve as a baseline for comparison of different sparse networks. We close by speculating on how sparsity can improve future workloads and outline major open problems in the field.

The supreme goal of all theory is to make the irreducible basic elements as simple and as few as possible without having to surrender the adequate representation of a single datum of experience -

Albert Einstein, 1933

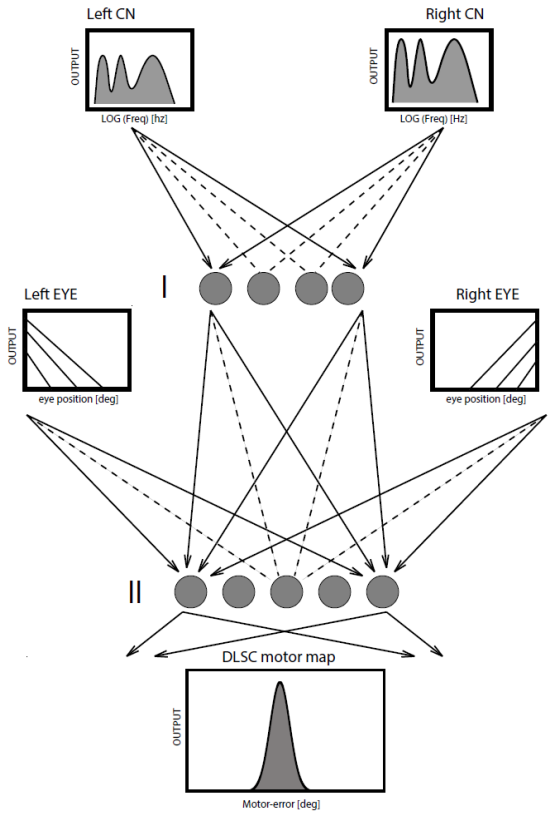


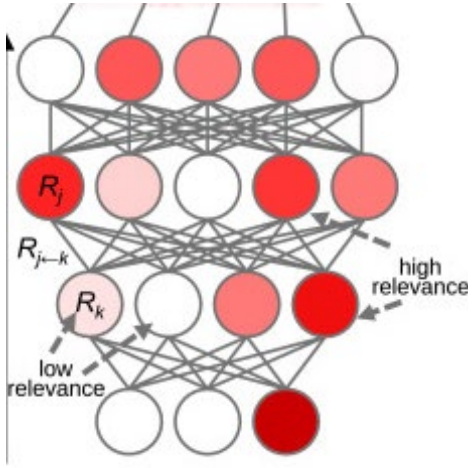
Hoefler, Torsten, Dan Alistarh, Tal Ben-Nun, Nikoli Dryden, and Alexandra Peste.
"Sparsity in Deep Learning: Pruning and growth for efficient inference and training in neural networks."
Journal of Machine Learning Research 22, no. 241 (2021): 1-124. https://doi.org/10.48550/arXiv.2102.00554

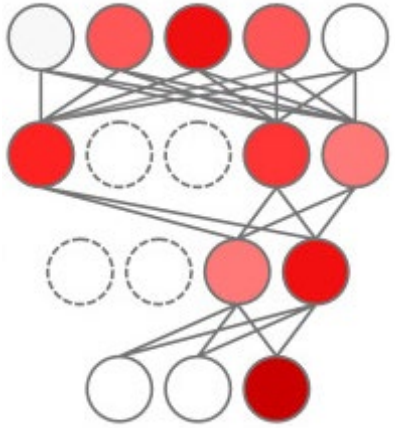
{Pruning}

HOGESCHOOL ROTTERDAM

DNNs are hard to down scale after the training phase, most DNNs are *not* sparse coded







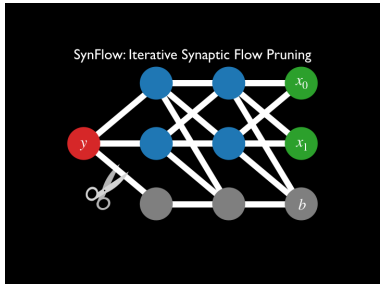
[International Workshop on Biologically Motivated Computer Vision]

Audio-Oculomotor Transformation (2002)
R.F. van der Willigen & Mark von Campenhausen
Part of the [Lecture Notes in Computer Science] book series (LNCS, volume 2525)

Yeom, S.K., Seegerer, P., Lapuschkin, S., Binder, A., Wiedemann, S., Müller, K.R., & Samek, W. (2021). Pruning by explaining: A novel criterion for DNN pruning. Pattern Recognition, 115, 107899.

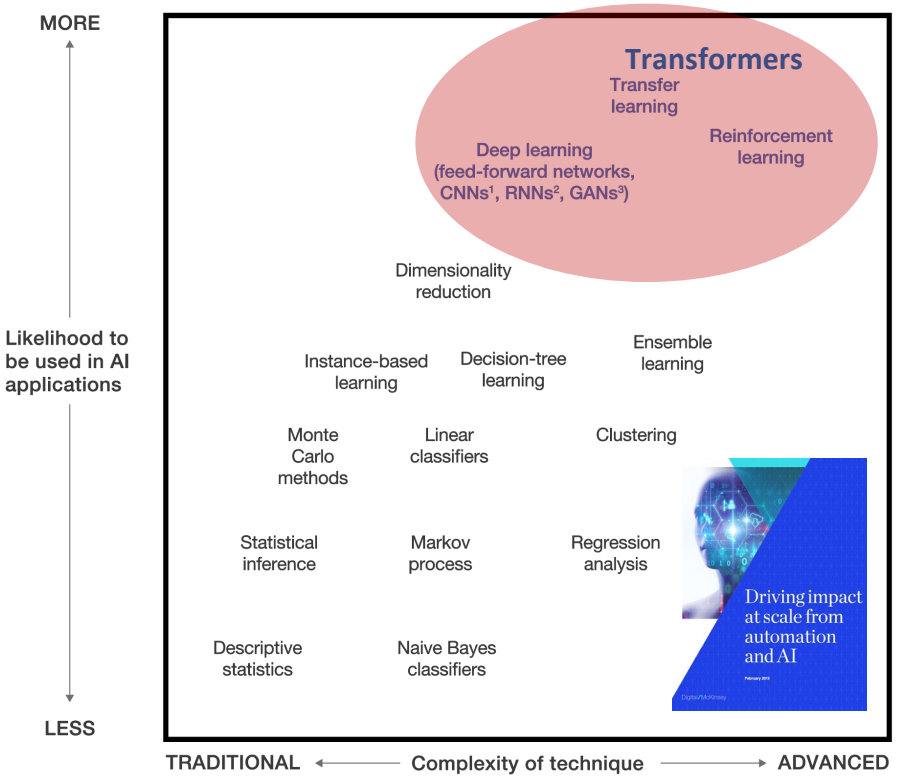
Blalock, D., Gonzalez Ortiz, J. J., Frankle, J., & Guttag, J. (2020). What is the state of neural network pruning? Proceedings of machine learning and systems, 2, 129-146.

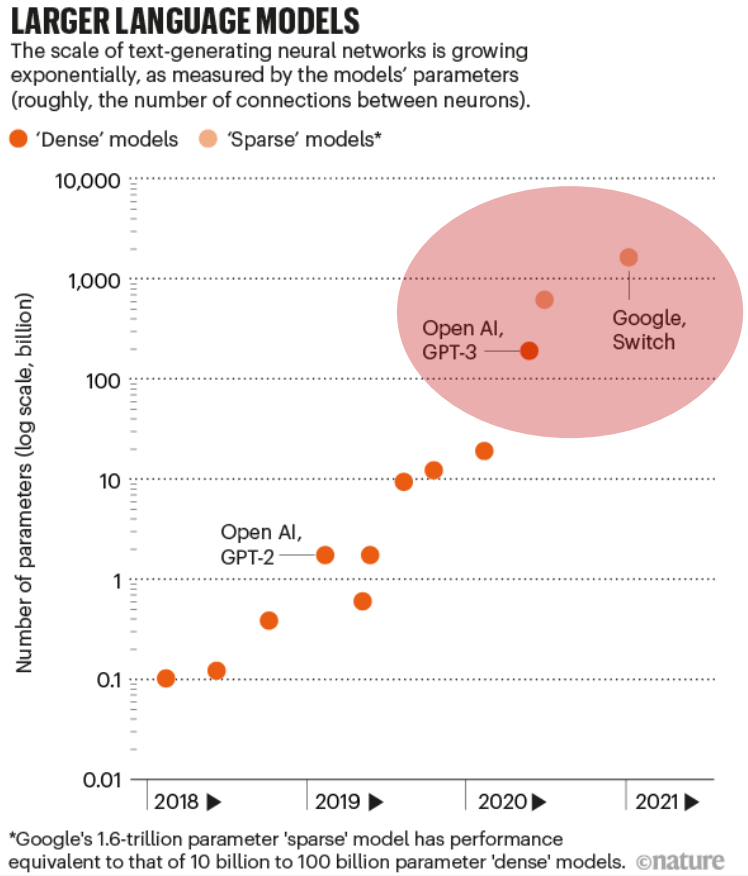
Frankle, J., & Carbin, M. (2018). The lottery ticket hypothesis: Finding sparse, trainable neural networks. *arXiv preprint arXiv:1803.03635*.



{large scale}

Only very large scale {DNNs} are useful [can compete with human performance]

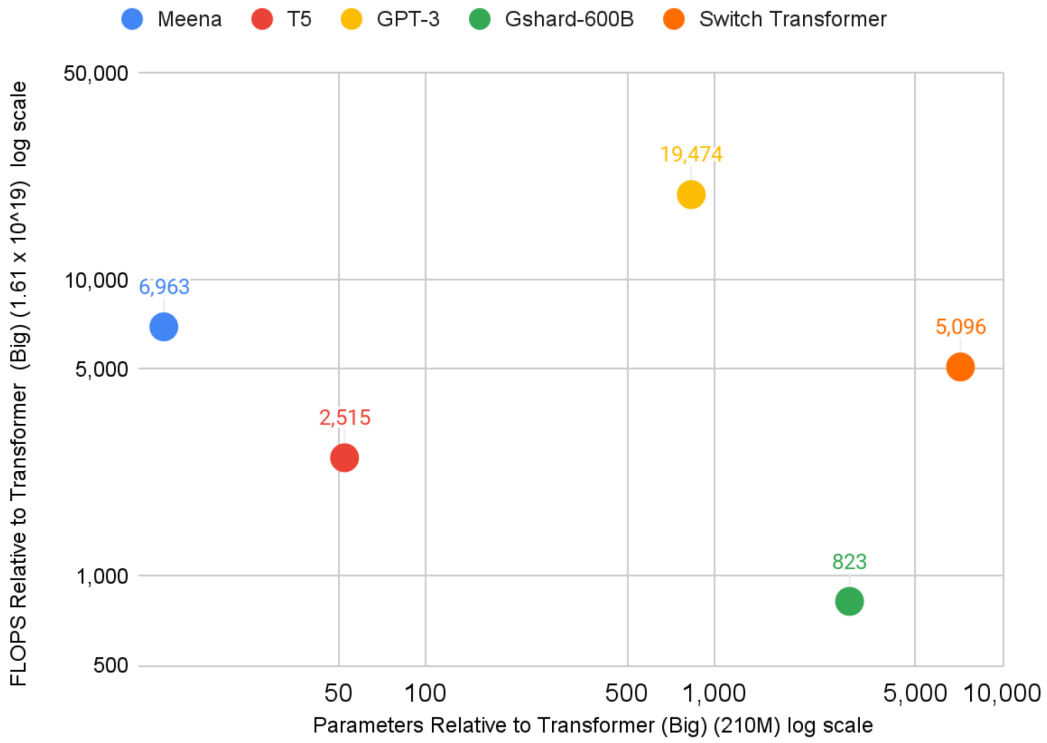


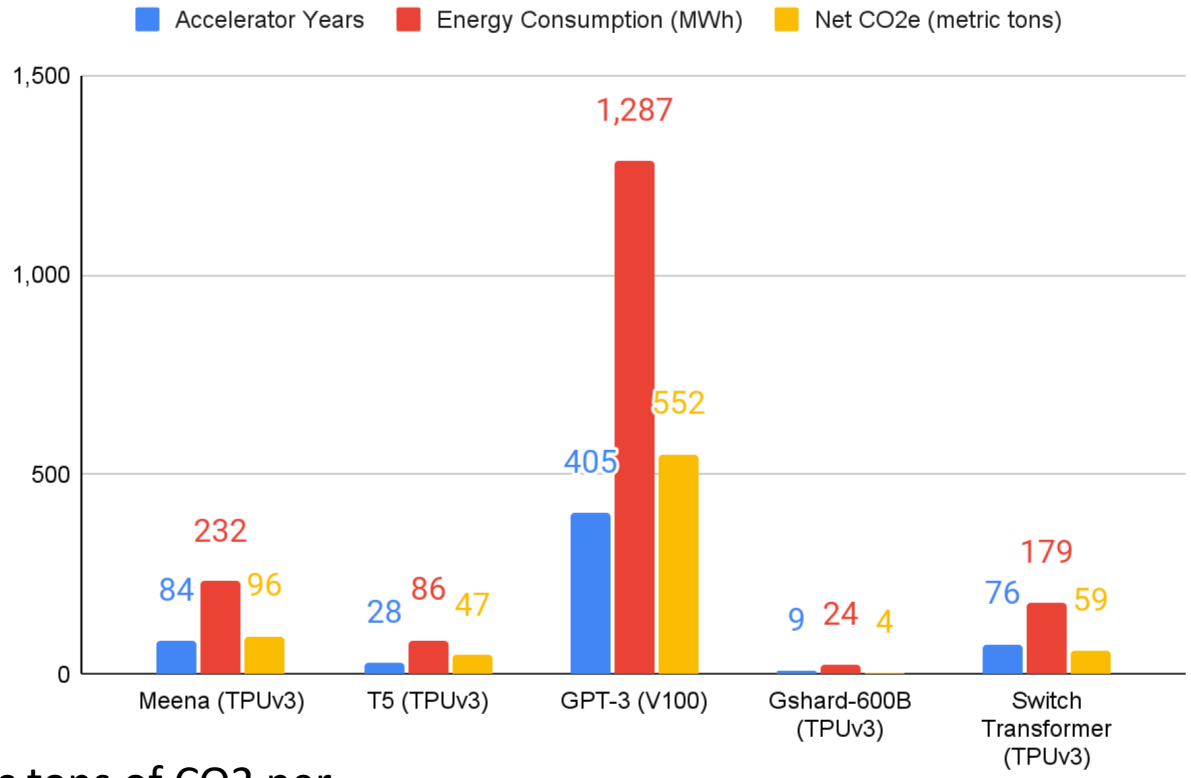


https://www.nature.com/articles/d41586-021-00530-0

{CO2 foot-print}

Training large scale transformer {DNNs} produce massive Carbon Emissions





As of 2007, the average U.S. household emits 20 metric tons of CO2 per year. In comparison to a world average of 4 tons.

Carbon Footprint_CSS09-05_e2021.pdf (umich.edu)

https://arxiv.org/ftp/arxiv/papers/2104/2104.10350.pdf

{computational unsustainability}

HOGESCHOOL ROTTERDAM

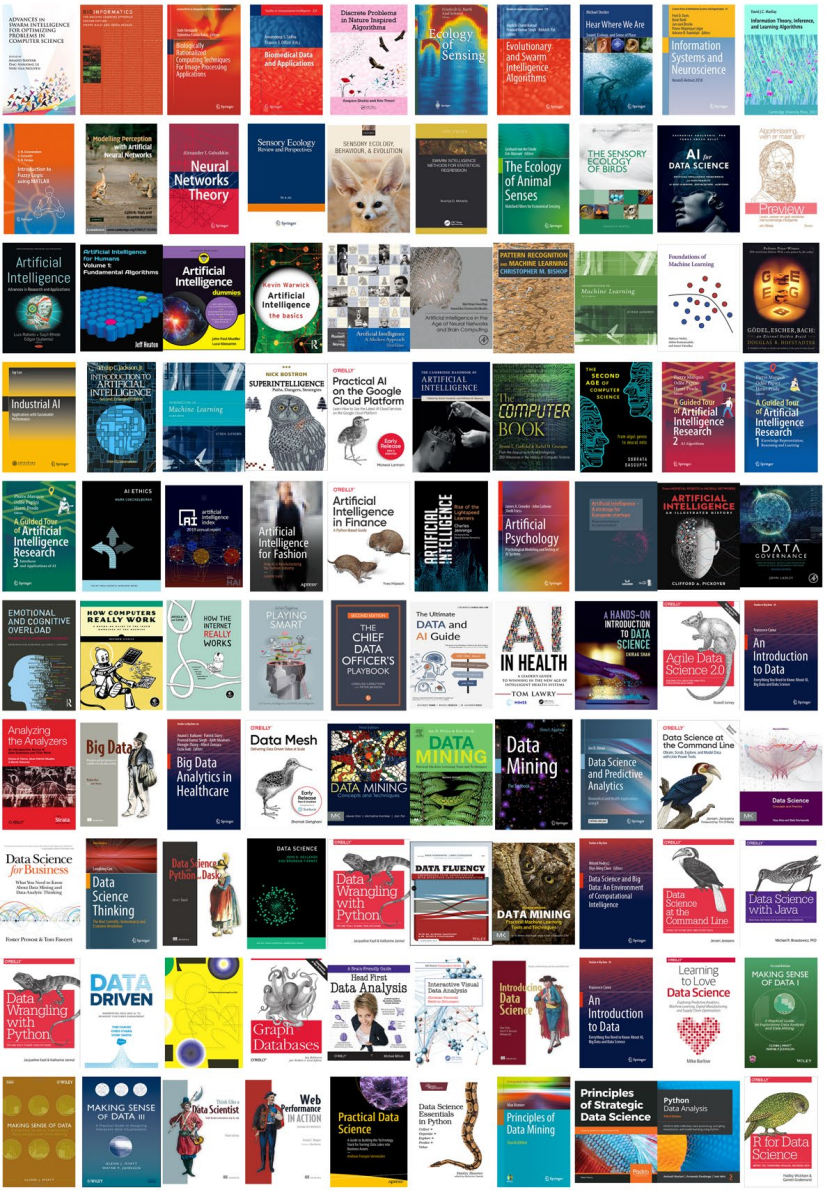
The scale of state-of-the-art {SOTA} –near human level– DNNs
– combined with a blind Brute-Force implementation + post-hoc analysis –
is becoming more and more
computationally unsustainable,
even to the point
that hypernetworks are employed
to help humans to make DDNs work.

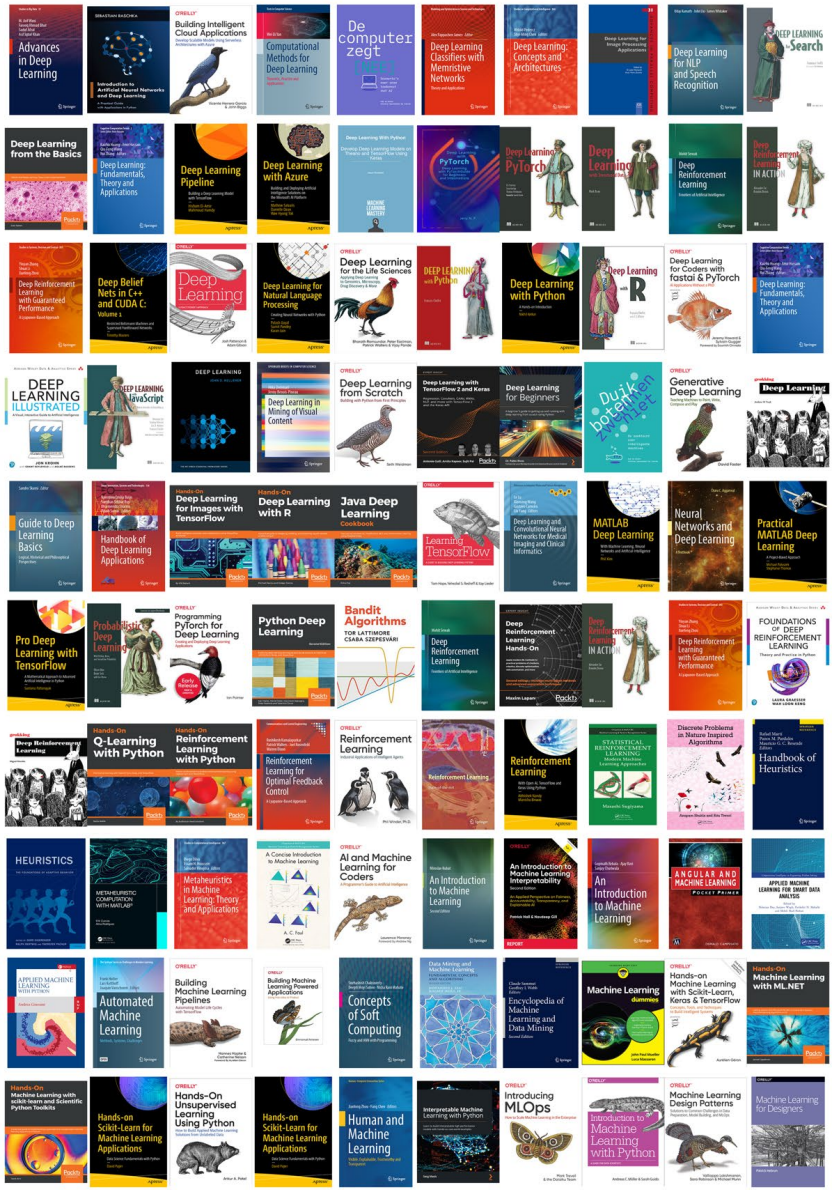
2110.13100v1.pdf (arxiv.org)

https://paperswithcode.com/sota/

{Studied Materials: books}

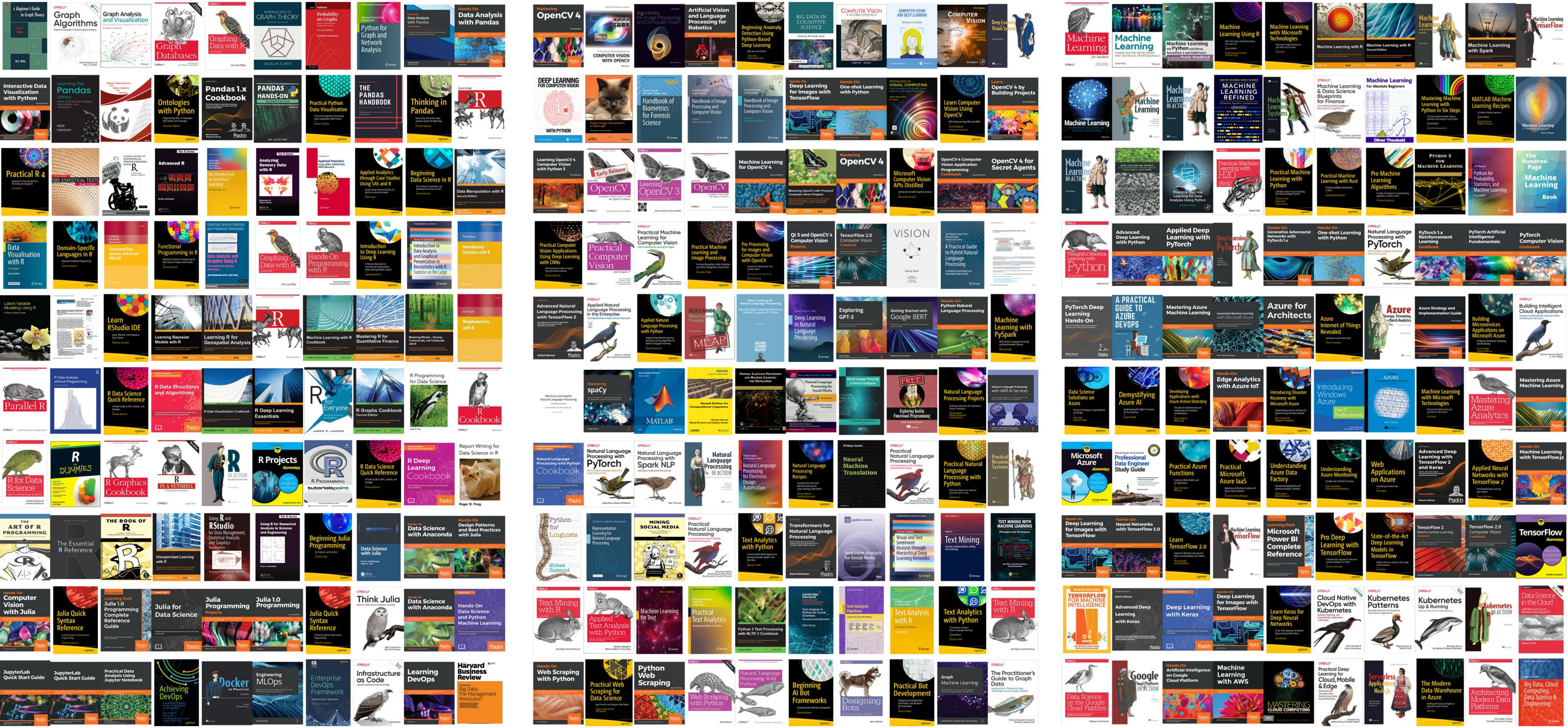
HOGESCHOOL ROTTERDAM





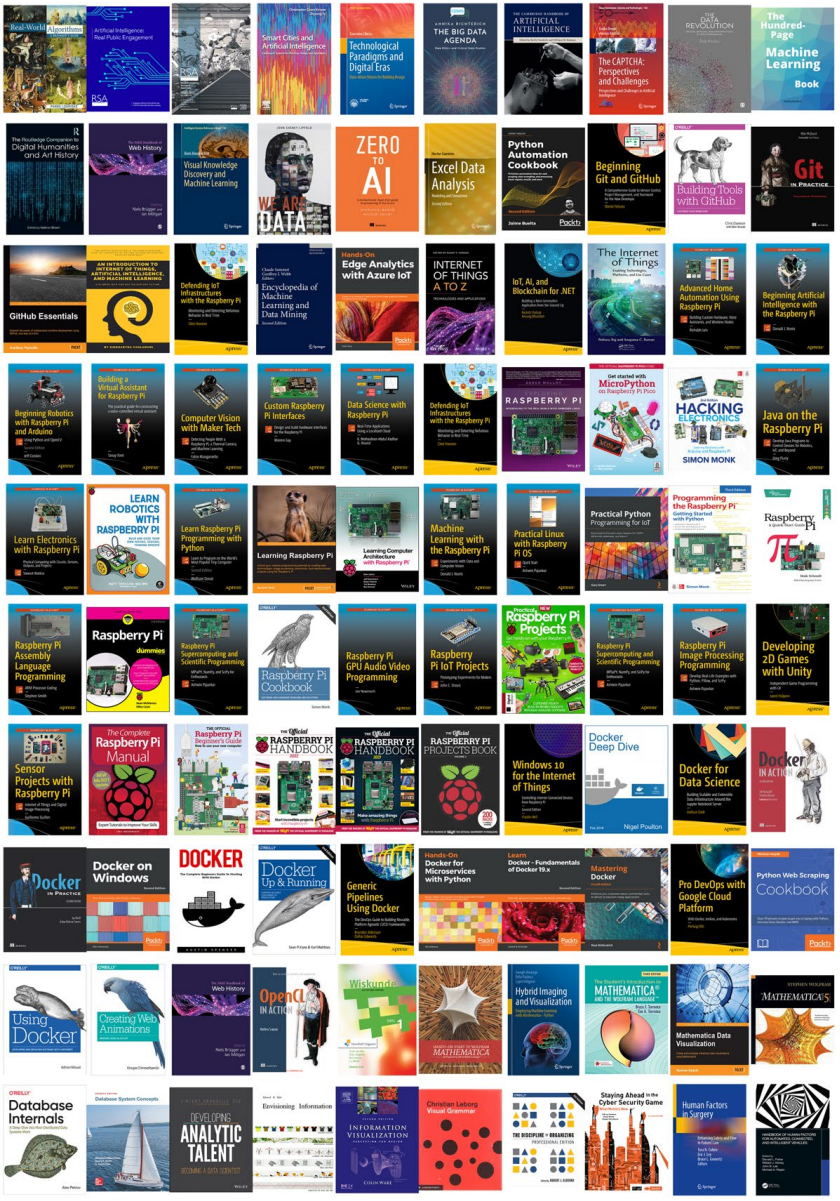
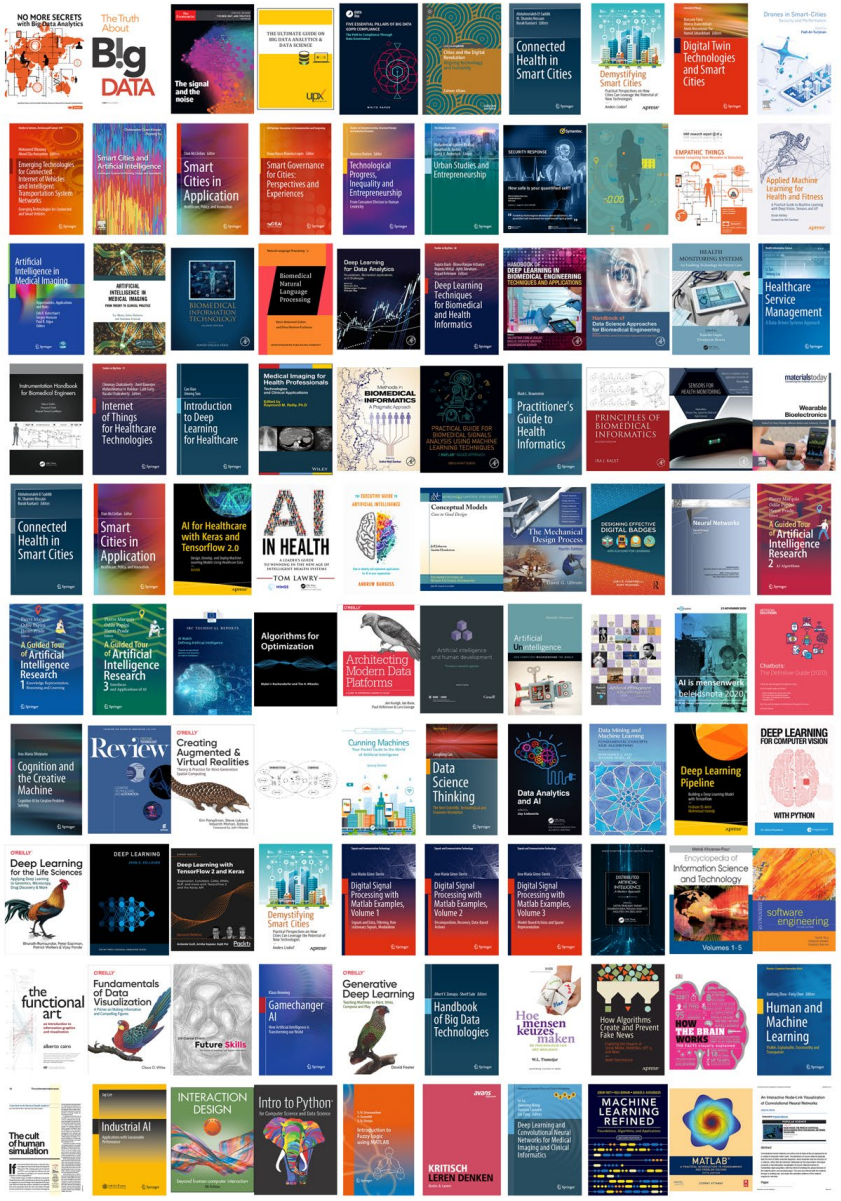
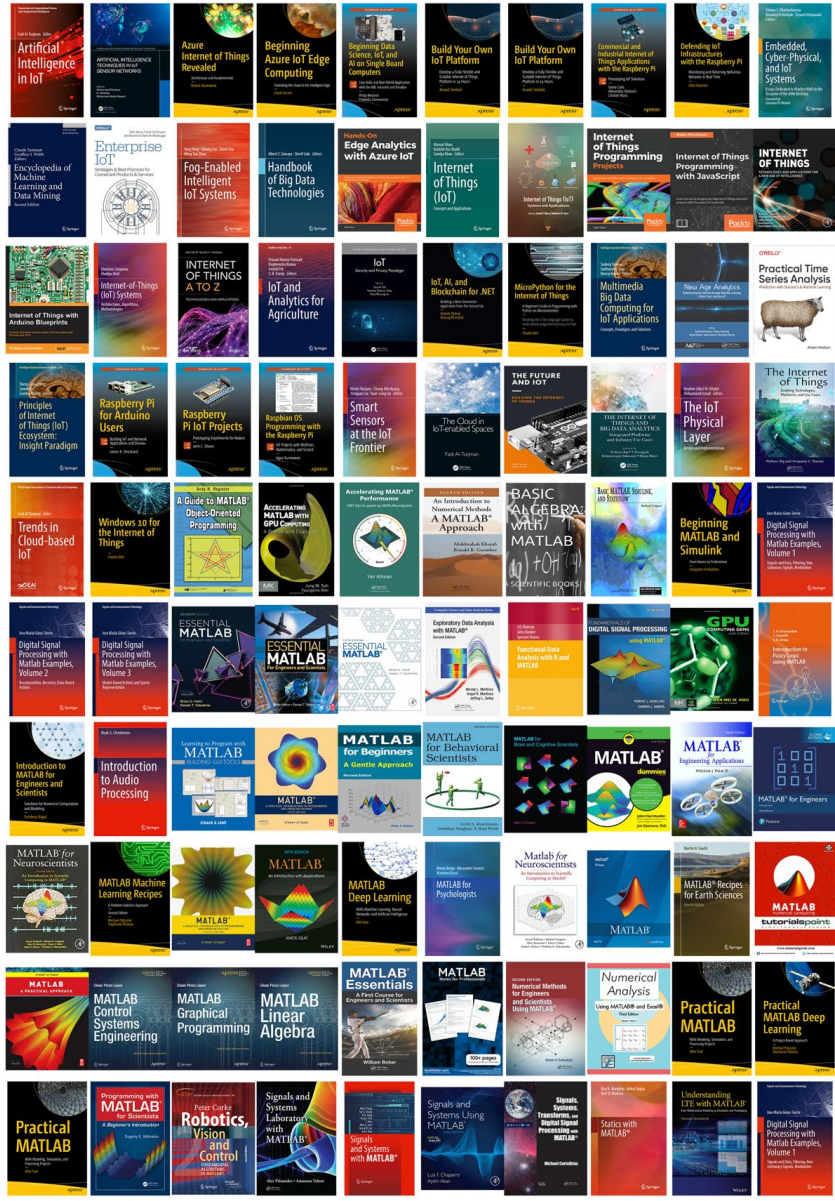
{Studied Materials: books}

HOGESCHOOL ROTTERDAM

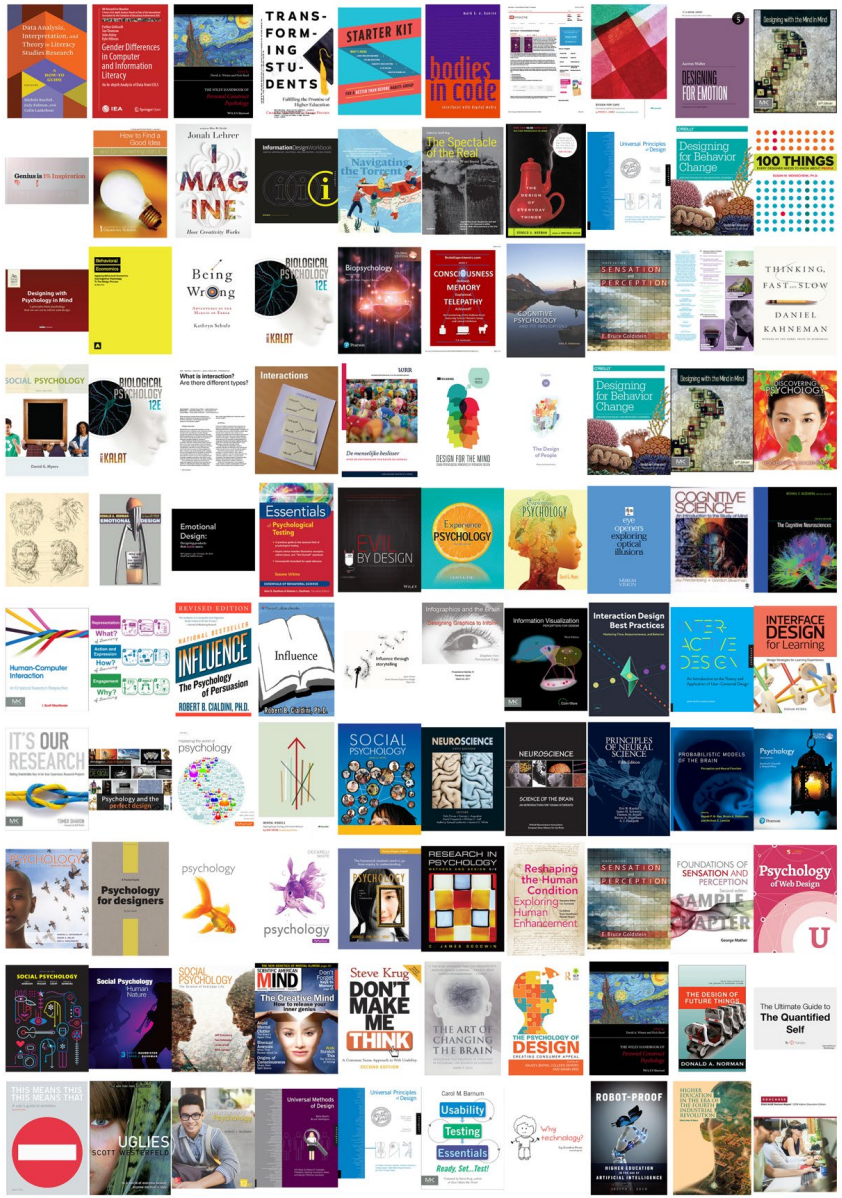


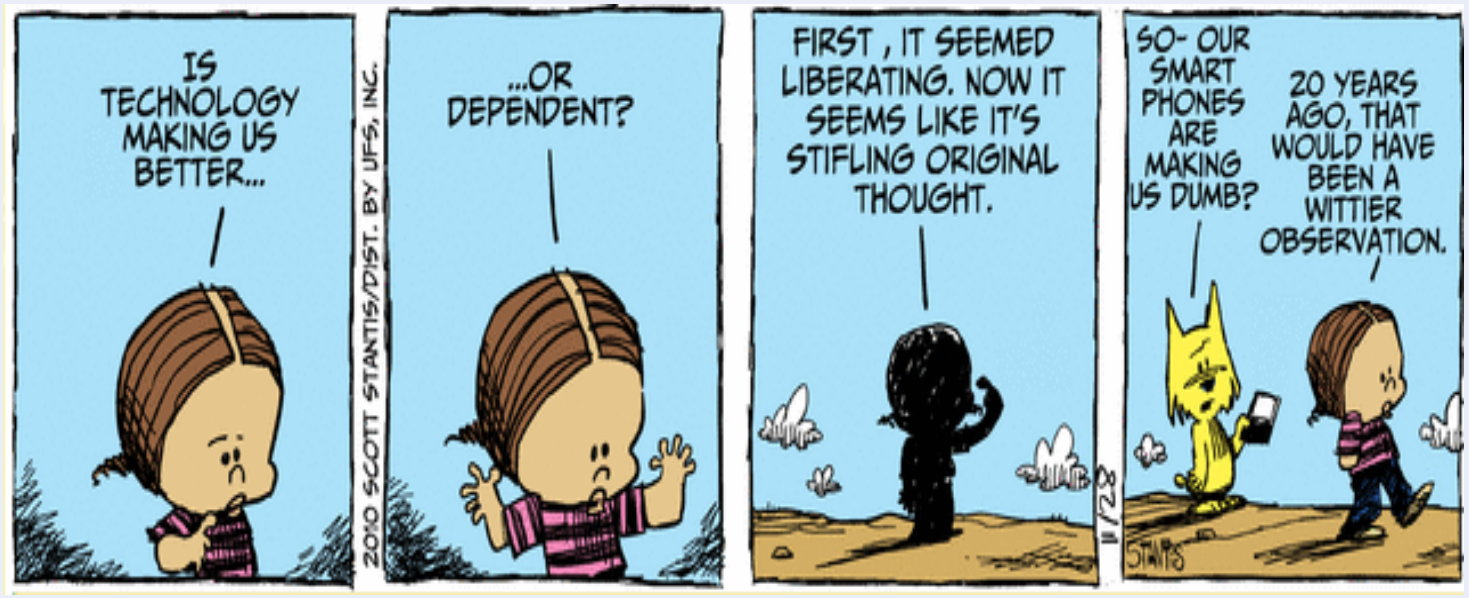
{Studied Materials: books}

HOGESCHOOL ROTTERDAM

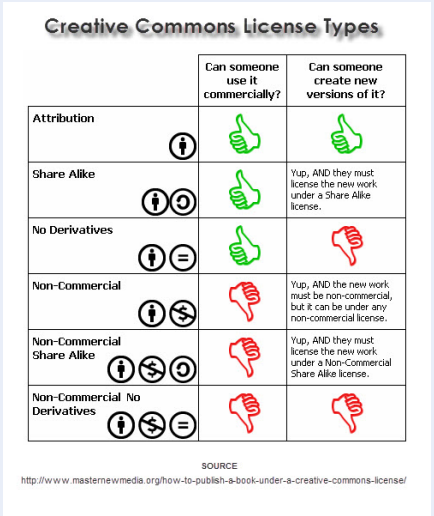




{Studied Materials: books}





This seminar was developed by:
Living-Lab: *AiRA,*
Hub voor Data & Responsible AI
Rob van der Willigen
FEB 2022



Creative Commons License Types

	Can someone use it commercially?	Can someone create new versions of it?
Attribution	👍	👍
Share Alike	👍	Yup, AND they must license the new work under a Share Alike license.
No Derivatives	👍	👎
Non-Commercial	👎	Yup, AND the new work must be non-commercial, but it can be under any non-commercial license.
Non-Commercial Share Alike	👎	Yup, AND they must license the new work under a Non-Commercial Share Alike license.
Non-Commercial No Derivatives	👎	👎

SOURCE
http://www.masternewmedia.org/how-to-publish-a-book-under-a-creative-commons-license/