Limits and Prospects of Big Data and Small Data Approaches in AI Applications

Ivan Kraljevski, Constanze Tschöpe, Matthias Wolff

Introduction

The renaissance of artificial intelligence (AI) in the last decade can be credited to several factors, but chief among these is the ever-increasing availability and miniaturization of computational resources. This process has contributed to the rise of ubiquitous computing via popularizing smart devices and the Internet of Things in everyday life. In turn, this has resulted in the generation of increasingly enormous amounts of data.

The tech giants are harvesting and storing data on their clients' behavior and, at the same time, introducing concerns about data privacy and protection. Suddenly, such an abundance of data and computing power, which was unimaginable a few decades ago, has caused a revival of old and the invention of new machine learning paradigms, like Deep Learning.

Artificial intelligence has undergone a technological breakthrough in various fields, achieving better than human performance in many areas (such as vision, board games etc.). More complex tasks require more sophisticated algorithms that need more and more data. It has often been said that data is becoming a resource that is more valuable than oil; however, not all data is equally available and obtainable. Big data can be described by using the "four Vs"; data with immense velocity, volume, variety, and low veracity. In contrast, small data do not possess any of those qualities; they are limited in size and nature and are observed or produced in a controlled manner.

Big data, along with powerful computing and storage resources, allow "black box" AI algorithms for various problems previously deemed unsolvable. One could create AI applications even without the underlying expert knowledge, assuming there are enough data and the right tools available (e.g. end-to-end speech recognition and generation, image and object recogni-

tion). There are numerous fields in science, industry and everyday life where AI has vast potential. However, due to the lack of big data, application is not straightforward or even possible. A good example is AI in medicine, where an AI system is intended to assist physicians in diagnosing and treating rare or previously never observed conditions, and there is no or an insufficient amount of data for reliable AI deployment. Thus, both big and small data concepts have limitations and prospects for different fields of application. This paper[1] will try to identify and present them by giving real-world examples in various AI fields.

Although the concept of non-human intelligence is as old as early humanity and can be witnessed in thousands of gods and mythological creatures, the idea of a machine capable of acting and behaving like a human being originated as early as antiquity. Mythical beings and objects were imbued with the ability to have intelligence and wisdom, and to be emotional. Throughout history, we have witnessed many attempts to create mechanical (much later also electrical) machines (such as the "automaton" Talos, the brazen guardian of Crete in Greek mythology) relatively successful in imitating human abilities. From early attempts to formalize human-level intelligence, cognition, and reasoning, from early antique on through medieval philosophers and up to 17th-century scholars, the physical symbol system hypothesis (PSSH) (Russell/Norvig 2009) laid the foundation of modern AI. Later, in the 20th century, advances in mathematics, especially logic, proved that any form of mathematical reasoning could be automated despite its limits.

The birth of AI as a modern research discipline is considered to be in the 1950s, when scientists from different research fields started to articulate the idea of the artificial brain.Since then, progress in AI has been primarily determined by technological developments and varying public interests. Because of unrealistic expectations, the field survived two "AI Winters." However, recent advances in computer science, particularly in machine learning, fueled by ever-increasing computing and storage performance, has led to an astronomical increase in research studies and previously unimaginable achievements. Artificial intelligence is increasingly becoming part of people's everyday lives and is achieving, in some respects, superhuman abilities. This has sparked some old and some new fears about the ethics and dangers of artificial intelligence.

1 The article was finished in March 2021.

To define artificial intelligence, we should first explain human intelligence. Finding a simple definition of human intelligence is seemingly impossible. In Legg et al. (2007), the authors succeeded in compiling 70 informal definitions, which is, as far as we know, the most extensive and most referenced collection thus far. The definitions could be nuanced into a more straightforward and a more general one: "Intelligence is a very general mental capability that, among other things, involves the ability to reason, plan, solve problems, think abstractly, comprehend complex ideas, learn quickly and learn from experience." (Gottfredson 1997)

Artificial intelligence could be categorized as one strictly demonstrated by machines (Artificial General Intelligence), such as that implemented in Unmanned Aerial Vehicles (UAV), autonomous driving systems or remote surgery. The second one imitates bio-inspired natural intelligence (Artificial Biological Intelligence) as in humans (humanoid robots) and animals (evolutionary algorithms, artificial neural networks, immune systems, biorobotics and swarm intelligence).

For a long time, the central paradigm to solve general artificial intelligence problems was the symbolic approach, known as "good old-fashioned AI." It was mainly based on the idea that intelligence can be emulated by manipulating symbols. The principal methodologies were based on formal logic or pure statistics (neat) and anti-logic (scruffy). The most prominent example of the latter is ELIZA, the first natural language processing (NLP) program. Without any formal knowledge, ad-hoc rules and scripts were optimized by humans until human-machine-like conversational behavior was achieved.

With the emergence of larger and more affordable computing memory necessary to store knowledge, AI scientists started building knowledge-based expert systems in the form of production rules, e.g. a sequence of "if-then" statements. Symbolic AI failed to fulfill expectations, primarily due to the combinatorial explosion, scalability and the fact it was only successful in solving very simple rather than real domain problems. It slowly faded into the first AI winter triggered by the Lighthill report (McCarthy 1974), which lasted until the 1980s. Approaches based on cybernetics and brain simulation experienced a similar fate, having been abandoned even earlier in the 1960s.

Consequently, in that period, other approaches appeared that were not based on symbolic reasoning and specific knowledge representations. Embodied intelligence in robotics (from control theory) promotes the idea that

body aspects, such as perception and movement, are required for higher intelligence. Furthermore, soft computing approaches (such as fuzzy systems and evolutionary algorithms) provide approximate solutions for problems that are deemed unsolvable from the point of view of logical certainty.

Statistical AI employs sophisticated mathematical apparatuses, including information and decision theory (Bayesian), hidden Markov models (HMM) and artificial neural networks (ANN), achieving better performance in many practical problems without the need for semantic understanding of the data. The shift from symbolic to statistical-based AI was a consequence of its limitations, and it is considered to be diverting away from explainable AI.

It is worth mentioning that the advances also influenced the appearance and disappearance of different information technology paradigms. Larger and more affordable computer memory enabled knowledge-based expert systems. The development of digital electronic components such as metal-oxide-semiconductors (MOS) and very-large-scale integration (VLSI) provided even more computational power, which led to a revival in artificial intelligence neural networks.

Artificial Neural Networks

The computational model for artificial neural networks appeared in the 1940s and was a biologically inspired attempt to replicate how the human brain works. The basic units are the artificial neurons which can be organized and interconnected in various ways and topologies. The artificial neuron represents a simple mathematical function where multiple inputs are separately weighted, summed and passed through a threshold to activate one or more other neurons. In this sense, an artificial neural network represents a directed weighted graph (Guresen/Kayakutlu 2011).

There are many ways that ANNs could be organized using different types of neurons and various connectivity patterns of neuron groups (layers). Commonly, ANNs have zero or few (shallow) or more hidden layers (deep) between the input and the output layer, where extremes like a single (e.g. perceptron) and unlayered networks are also possible. An artificial neural network is known as “feed-forward” when each neuron in a layer connects to all of (fully connected) or a group of neurons (pooling) from the previous layer. Other connectivity patterns are also possible, such as fully recurrent neural networks, Hopfield networks, Boltzmann machines and self-organizing maps (SOM).

The network parameters, such as connection weights, bias and activation thresholds, are estimated in a learning or training process. Sample observations in the form of input data with the corresponding desired outputs are necessary for supervised learning. The network's parameters are adapted to minimize the difference between the expected and the predicted outcomes. The difference is quantified with a loss or cost function and backpropagated to compute the gradient of the weights. The procedure is repeated, either with stochastic (each input produces weight change) or batch (accumulated error in a batch of inputs produces the weight change) learning mode until the desired accuracy is reached, or there is no more improvement observed in the loss function (Goodfellow et al. 2016).

Deep Learning

ANNs have had a complicated history since the very first emergence of the concept. The hype cycle for this technology is atypical, with many ups and downs. The initial excitement about the perceptron (1957) disappeared quickly because only linear separability was possible; this was mitigated with the multilayer perceptron (1969), which again raised expectations. However, this lasted only until the appearance of support vector machines (1998), which provided better performance than ANNs while also boasting better theoretical background and understandability.

Things changed again in 2006 after overcoming crucial numerical issues, like the problem of "vanishing" or "exploding" gradients. Due to the availability of more powerful computers, it was suddenly feasible to use architectures with increased numbers and sizes of hidden layers (deep neural networks), capable of outperforming the support vector machines (SVMs) in many different classification tasks (Temam 2010). Deep neural networks (DNN) are architectures with many hidden layers made of different cell types that can be combined and provide better data representation (Schmidhuber 2015). For instance, convolutional layers could accept external multidimensional data (e.g. images), learn the distinctive features and pass their output to the recurrent layers to capture spatial or time dependencies, which will serve the output to a fully connected layer to classify the image into one of the expected categories (e.g., cats or dogs).

The size of such architectures quickly increased from thousands to billions of trainable parameters, making the training quite challenging (Rajbhandari et al. 2019) while providing a significant increase in classification

performance. Just to mention a few notable examples in the field of NLP, there are the Bit-M (Kolesnikov et al. 2019) with 928M, Megatron-LM from NVIDIA (Shoeybi et al. 2019) with 8.3B, Turing-NLG from Microsoft (Rasley et al. 2020) with 17B and GPT-3 from OpenAI (Brown et al. 2020) with 175B parameters. The latter, which is considered the largest artificial language model, was dwarfed by the Google Brain team model, which has a staggering 1.6 trillion parameters (Fedus et al. 2021).

The training and optimizations of large and performant models have many implications. First, there is the financial costs to train such a model. The costs can range from a couple of thousand to several million dollars (e.g. for the 350GB large Open AI GPT-3 model, the costs are around $12M). Second, there is an issue with the availability of proper computing infrastructure. Supercomputers with multiple instances of specialized hardware (GPUs and TPUs) are not affordable or accessible to most AI communities (researchers, startups and casual enthusiasts). Third, one must consider the levels of energy and power consumption. Strubell et al. (2019) showed that the yearly power consumption of the tech giants equals that of the entire United States, whereas on a global level, 3-7% of electrical energy is consumed by computing devices (Joseph et al. 2014). About 50% of the electrical energy consumption in data centers is used only for cooling (Meijer 2010). In turn, energy and power consumption directly influence the carbon dioxide footprint. For instance, the carbon emissions (CO_2 emission in lbs.) to train one model on a GPU with tuning and experimentation (78,468) surpasses the average annual emissions of two cars (36,156 per car) (Strubell et al. 2019).

Finally, we come to the data that is collected, processed and stored in data centers, data that are in a reinforcing loop with deep learning. Deep learning requires an abundance of data for the training of large models. On the other hand, the tremendous growth of generated data allows applying even more complex algorithms and creating even larger models, which, when used in consumer devices or similar appliances, in turn generates even more data.

The Data

The "digital transformation" and the advances in computing and communication technology are the driving factors for data creation and consumption growth. Only in the last two years, more data was created than in the previous entirety of human history. The most recent Global DataSphere forecast

published by IDC predicts more than 59 zettabytes of data creation and consumption in a year, doubling the amount of two years before and surpassing that of the previous 30 years together (Völske et al. 2021).

In so-called "Surveillance capitalism" (Zuboff 2015), consumer devices' users voluntarily provide sensitive data to the tech giants in return for free applications and services, not fully aware that their behavioral data is transformed into a commodity. The data security and data protection aspects are regulated in most countries, like the General Data Protection Regulation (EU) law in European Union. Despite that, in recent years, there have been numerous breaches committed by the tech corporations, where personal data was used for social media analytics in business intelligence or political marketing (notable cases involve AOL, Facebook and Cambridge Analytica).

The value of data and the ability to govern data in an organization is considered an essential asset. Nowadays, almost no business decision could be made without valuable knowledge extracted from a waste amount of structured or unstructured data. On the other hand, there are still areas where the volume and data collection rate are the opposite of what we have described. Natural and technological processes that last long and have lower dynamics provide data at a much slower rate. Therefore, applying artificial intelligence in such cases requires substantially different approaches than those relying on an abundance of data.

Artificial Intelligence Applications

We would like to illustrate the recent advances in artificial intelligence with examples where AI achieved near-human and superhuman abilities. Those achievements are a direct result of having sophisticated machine learning algorithms, computational power, and of course, a vast amount of data. According to Haenlein and Kaplan (2019), there are three evolutionary stages of AI, namely artificial narrow intelligence – ANI, artificial general intelligence – AGI and artificial super intelligence – ASI and three different types of AI systems: analytical AI, human-inspired AI and humanized AI.

Although the big corporations (such as Google or Tesla) are revolutionizing the ways in which artificial intelligence interacts with humans and the environment, all the existing solutions could be described as "Weak AI" (artificial narrow intelligence). In distinction to general AI (AGI), weak AI strives to solve a specific task. In its essence, it is just a sophisticated algo-

rithm responding to sensory inputs or user behavior. We usually see those in everyday life, in our smartphones, smart home devices, personal assistants. They are merely machine learning algorithms that adapted themselves and improved their skills according to user's feedback and behavior. Such systems are not getting smarter by themselves, and their usability and prediction performance always depend on the available data. For instance, a personal voice assistant can learn the speech and language specificities of its user. However, it cannot handle requests that are not foreseen or easily understand a new dialect or language. Although deep learning and big data contributed to ANI achieving impressive results in some application areas, closing in on or even surpassing human abilities, currently, we are still far from developing an AGI (Fjelland 2020).

When comparing artificial intelligence with human abilities, there were two events in recent human history in which AI competed against humans in zero-sum games, that have drawn the whole world's attention and focus on AI. The first event that gained broader public attention was the chess computer Deep Blue's (IBM) match against the reigning world champion in chess in 1997, Garry Kasparov. Playing chess better than humans was always considered one of the "Holy Grail" milestones in AI. Chess requires everything that makes up human intelligence: logic, planning and creativity. A computer beating a human would prove the supremacy of AI over human abilities. Using supercomputer power with custom-build hardware (Hsu et al. 1995) to allow evaluation of millions of possible moves in a second, the machine won only one game in the tournament. Although Deep Blue, being merely a fine-tuned algorithm that exhibits intelligence in a limited domain (Hsu 2002), was nothing like AI nowadays, it passed the chess Turing test, which was a historic moment.

Almost twenty years afterward, in 2017, a similar event took place. AlphaGo (DeepMind), a computer that plays the game Go, beat a human professional player on equal terms. Go is considered much more difficult for computers to play against humans because of the almost infinite combinations of board positions and moves. Traditional approaches, such as heuristics, tree search and traversal (Schraudolph et al. 1994), applied for chess were not applicable in this case. Neither would a pure deep learning approach be feasible due to the data requirements to train a network for all the possible outcomes. Instead, an algorithm that combines tree search techniques (Monte-Carlo) with machine learning was devised. Historical games (30 million moves from 160,000 games) were used for supervised training

and computer self-play for reinforced learning of "value networks" for board position evaluations and "policy networks" for move selection (Silver et al. 2016).

Although trained to mimic human players, during the famous game 2 in the match against the professional Korean Go player Lee Sedol, AlphaGo made the famous move 37 (Metz 2016), considered by Go experts as a "unique" and "creative" move that no human player would play. That emerged from the algorithm's underlying objective to maximize the probability of winning, regardless of the margin, even losing some points, in contrast with human players. That showed that AlphaGo exhibited AI-analytical rather than human-inspired AI abilities. The original AlphaGo system went through several transformations. It was further developed into MuZero (Schrittwieser et al. 2020), an algorithm that can learn to play games without knowing the game rules and achieved superhuman performance in Atari Games, Go, Chess and Shogi.

Natural Language Processing

Natural Language Processing is another milestone in the advancements towards AGI. With the greater human dependency on computers to communicate and perform everyday tasks, it is gaining more attention and importance. NLP is one of the most prominent and widespread areas where ANI excels, and there are many examples of NLP in action in our everyday lives at home or work.

NLP allows machines to understand human language conveyed by spoken and written words or even gestures, which are tremendously diverse, ambiguous, and complex. It is quite problematic for the NLP to achieve near-human performance due to the nature of language itself, with many obstacles in the form of unstructured language data, less formal rules and a lack of a realistic context.

NLP solutions are employed to analyze a large amount of natural language data to perform numerous tasks in different applications: written text and speech recognition and generation; morphological, lexical and relational semantics; text summarization, dialogue management, natural language generation and understanding, machine translations, sentiment analysis, question answering and many, many others. Despite the recent advances, language processing is still presented with plenty of unsolved problems.

NLP shares a history timeline with AI in its progression from symbolic, statistical approaches to the recent neural and deep learning-based approaches (Goldberg 2016). The revival of neural networks in the form of deep learning, along with the ever-increasing availability of unstructured textual online content (big data), propelled research in NLP, introducing new algorithms and models (Young et al. 2018).

It is pretty challenging to build a state-of-the-art NLP model from scratch, particularly for a specific language domain or task, due to the lack of the required and appropriate data. Therefore, transfer learning by leveraging extensive and generalized pre-trained NLP models is the typical approach to solve a specific problem by fine-tuning a dataset of a couple of thousand examples which saves time and computational resources.

The pre-trained NLP models opened a new era, and we will mention some of the most recent state-of-the-art (SotA) models that can close the gap to human performance in many different NLP tasks. The Bidirectional Encoder Representations from Transformers or BERT (Devlin et al. 2018) developed at Google is an enormous transformer model (340M parameters) trained on a 3.3-billion-word corpus not requiring any fine-tuning to be applied to specific NLP tasks. The Facebook AI team introduced RoBERTa (Liu et al. 2019) as a replication study of BERT pretraining with more data, better hyperparameters and training data size optimizations. To anticipate the continuous growth of the models' size, the need for computational resources and to provide acceptable performance for downstream tasks (chatbots, sentiment analysis, document mining, and text classification), the A Lite BERT (ALBERT) architecture was introduced by Google (Lan et al. 2019). XLNet was developed by the researchers at Carnegie Mellon University, which has a generalized autoregressive pretraining method that outperforms BERT on tasks like question answering, natural language inference, sentiment analysis and document ranking (Yang et al. 2019). The OpenAI GPT-3 model (Brown et al. 2020) is an up-scaled language model surpassing in size the Turing-NLG (Rosset 2020), which aims to avoid fine-tuning (zero-shot) or at least to use a few-shot approach. It is an autoregressive language model with 175B parameters and trained on all-encompassing data that consist of hundreds of billions of words.

All the top-performing NLP models created by the tech giants have in common their data-hungry complex deep learning architectures, trained on supercomputing systems with thousands of CPUs and GPUs (TPUs), consequently creating colossal models. Despite breaking records in numerous

benchmarks, they are still making trivial mistakes (e.g. in machine translation); therefore, SotA NLP models are yet to close the gap on the AGI.

Speech Technologies

NLP applications have been present for a long time, but the arrival of virtual personal assistants with AI technology brought the AI experience much closer to ordinary users. Smart personal assistant technology encompasses many AI-capable technologies, and its increasing adoption and popularity are due to the recent advances in speech technologies.

The first modern smart assistant is Siri (Apple), which appeared in 2011 on the smartphone as standard software. Google introduced its personal assistant Google Assistant (2016), a follower of Google Now, and recently Google Duplex, which can communicate in natural language and became famous for demonstrating robocalling to a hair salon (Leviathan/Matias 2018). A few years later, after the appearance of Siri, AI personal assistants went mainstream when Amazon introduced Alexa Echo, a smart loudspeaker. Alexa is capable of voice interaction, understanding questions and queries and providing answers, controlling smart home systems and many other skills whose number is steadily increasing thanks to the large developer community. Microsoft's voice assistant Cortana has also become a standard feature on their Windows operating systems. Recognizing and understanding human speech and interacting in a human voice are considered AI benchmarks to be mastered on the route towards AGI.

Automatic speech recognition (ASR) has a long history. Like NLP, its crucial milestones correspond with AI's general advances, but practical and widespread application has been possible only for the last two decades. The technology experienced breakthroughs, rapidly dropping the word error rates, capitalizing on deep learning and big data. Suddenly, a massive amount of speech data became available, as consumers using voice queries provided already quasi-transcribed speech in many different styles and environments.

For deep learning, the paradigm "There is no data like more data" (Bob Mercer at Arden House 1985) is the key to success. Now it is possible to train end-to-end speech recognition systems with raw speech signals as input and the transcriptions in the form of a sequence of words as outputs. During training on tens of thousands of hours of speech, the neural network model learns the optimal acoustic features, phono-tactic and linguistic rules,

syntactic and semantic concepts. There is no need for algorithms, linguistics, statistics, error analysis, or anything else that requires human expertise (black-box approach) as long as there are enough data.

The current SotA speech recognition systems demonstrate impressive performance on standard recognition tasks (Librispeech, TIMIT). The most recent framework from the Facebook research team, the wav2vec 2.0 (Baevski et al. 2020) (trained on 960 hours of speech), achieved a word-error-rate (WER) of 1.8/3.3 percent on "test-clean/other" of Librispeech and phoneme-error-rate of 7.4/8.3 percent on "dev/test" on TIMIT. In Zhang et al. (2020), WERs of 1.4/2.6 percent on the LibriSpeech test/test-other sets are achieved. Such impressive results are achieved on standard speech corpora in restricted conditions, domain, and language, and cannot necessarily be directly translated to real-world situations. The study (Georgila et al. 2020) shows that the current SotA of off-the-shelf speech recognizers perform relatively poorly in domain-specific use cases under noisy conditions. In summary, it thus appears that despite the impressive achievements, speech recognition technology has yet to close the gap, and it is expected to pass the Turing test for speech conversation in 40 years (Huang et al. 2014). However, history has already proven that this could happen much sooner than expected.

Computer Vision

Computer vision (CV) is a field where, like no other, some of the AI challenges are considered to be completely solved. Improvements in the CV hardware in terms of computing power, capacity, optics and sensor resolutions paved the way for high-performance and cost-effective vision systems (cameras, sensors and lidar) to consumer devices. Coupled with sophisticated video-processing software and social networks, the smartphone becomes the ideal crowd-sourcing platform for generating image and video data. Traditional CV methods were replaced with end-to-end systems that, like speech recognition, avoided the need for expert analysis and fine-tuning while achieving greater accuracy in most of the tasks (O'Mahony et al. 2019).

The traditional CV task is image recognition, where the objective is to recognize, identify or detect objects, features and activities. The image recognition finds the 2D or 3D positions of an object and classifies it in some category. On the other hand, identifying an object also recognizes a particular instance (such as a person's face or a car license plate). Simultaneously, de-

tection provides information on whether some object or condition is present or not in the observed scene (anomalies in manufacturing industries, medical conditions, obstacles detection), which could also be consistently recognized and identified. Other major CV tasks are motion analysis (same as image recognition but over time), scene reconstruction (extracting a 3D model from the observed scene) and image restoration (such as denoising, upscaling, colorization).

The SotA deep learning algorithms for image recognition use convolutional neural networks (CNNs), leveraging the increase in computing power and data available to train such networks (Krizhevsky et al. 2012). Since 2011 (Ciresan et al. 2011), CNNs have regularly achieved superhuman results in standard computer vision benchmarks (MNIST, CIFAR, ImageNet), which influence industry (such as health care, transport and many other areas). ImageNet (Deng et al. 2009) is the database commonly used for benchmarking in object classification (Russakovsky et al. 2015). It consists of millions of images and object classes. Currently, the new state-of-the-art top-1 accuracy of 90.2% on ImageNet (Pham et al. 2020) is achieved by the method of pseudo labels. A pre-trained teacher network generates pseudo labels on unlabeled data to teach a student network, and it is continuously adapted by the student's feedback, which generates better pseudo labels.

There are various CV applications, like image super-resolution (Grant-Jacob et al. 2019), face and image recognition in consumer devices, search engines and social networks. In healthcare, medical image analysis, such as MRI, CT scans and X-rays, allows physicians to understand and interpret images better by representing them as 3D interactive models. Autonomous vehicles, like UAV, submersibles, robots, cars and trucks, employ many of computer vision's general tasks. Many car producers already offer autonomous driving as a feature in their products despite it being far from perfect. Human pose and gesture recognition finds application in real-time sports and surveillance systems, augmented reality experience, gaming and improving the life of hearing and speech impaired people (sign language recognition). In agriculture, CV is used in farm management, animal and plant recognition, and in monitoring and predictive maintenance in the manufacturing industry, in the military for battle scene analysis, assisting law enforcement, education and many others. All in all, however, it must be noted that even if CV achieves human capabilities in pattern matching, it still cannot extract and represent the vast amount of human experience to understand the whole context and the hidden object dependencies.

Healthcare

There are so many AI applications in healthcare and medicine that it would be impossible to mention all of them in the space given. Advancements of AI in healthcare are among the most beneficial uses to humankind and recently gained more attention with focused research and development in combating the COVID19 pandemic.

AI in healthcare has a long history; it has been strongly influenced by the discoveries and advances in medicine and computing technology in the last fifty years. Deep learning enabled the replication of human perceptual processes by natural language processing and computer vision. Computing power resulted in faster and better data processing and storage, leading to an increase in electronic health record systems, which offer the required security and privacy level. In turn, this made the data necessary for machine learning and AI in healthcare affordable and available, a prerequisite for rapid advancement in this area.

AI can transform the healthcare industry and make it more personal, predictive and participatory. However, the greatest challenge for AI is ensuring its adoption in daily clinical practice. For now, its potential has primarily been demonstrated in carefully controlled experiments, and few of the AI-based tools (medical imaging) have been approved by regulators for use in real hospitals and doctors' offices (Davenport/Kalakota 2019).

Many AI-enabled technologies are applicable in healthcare. The core applications can be divided into: disease diagnostics and treatment, medical devices, robotic-assisted surgery, medical imaging and visualization, administrative tasks, telemedicine, precision medicine and drug discovery (Rong et al. 2020). However, the most promising area of application is clinical data management, which aims to provide better and faster health services. Some examples are: collection, storing and mining of medical records in the form of demographic data, medical notes, electronic records from medical devices, physical examinations, clinical laboratory and image data (Jiang et al. 2017). With NLP, it is possible to analyze a massive amount of unstructured medical data, even entire healthcare systems, and discover unknown and hidden patterns in medical condition diagnostics, therapies and drug treatments, creating insights for healthcare providers in making better clinical decisions.

Limits and Prospects of Big Data and Small Data

As already presented in the previous section on major AI applications, deep learning advances are the main driving force of AI technology. However, recently it has become apparent that deep learning has its limitations. The critical voices lately raised the concern that we face another AI winter, primarily due to AI scientists' and companies' unrealistic and ambitious promises. Besides that, since deep learning and AI have a significant influence on our lives and societies, the question of understanding the underlying principles and limitations becomes ever more critical. We need to understand how deep learning works and, when it fails, why it failed. For instance, some AI systems achieved the superhuman ability to recognize objects. However, even a slight divergence from the training data renders the predictions unusable. Deep learning models can recognize an image because they learned not the object, but other features, such as the background. In other applications, like face recognition, the results are strongly biased against minority groups and gender because they were not part of the training data, and so on.

The main issue is that given enough (big) data, the deep learning neural network always tries to find the simplest possible solution to the problem; no matter the task, it just might find a "shortcut" solution (Geirhos et al. 2020). One solution is to provide even more data that include adversarial examples. Another is to combine symbolic AI with deep learning by introducing hard-coded rules (hybrid systems), which are as good as the data or the expertise they are based on.

On the other hand, AI must learn from less data, as we humans do. In some cases, employing deep learning would be overkill, as the traditional machine learning techniques can often solve a problem much more efficiently. Therefore, it is essential to challenge the "black-box" approach in deep learning and understand the underlying principles, providing interpretable and explainable AI.

Big vs. Small Data

One of the drawbacks of using AI algorithms is the need and consumption of an enormous amount of data. In many real-world use-cases, employing data-driven AI methods is not feasible because of the limited amount of available data. The desired accuracy is not achievable due to various datasets'

constraints, mainly when they are high-dimensional, complicated or expensive. An example would be time-consuming data collection where a specific industrial system does not produce enough data in a foreseeable period to train an AI system. Also, in personalized medicine (Hekler et al. 2019), big data approaches are impossible. All such cases are considered examples of "small" as a counterpart to "big" data. So far, we have shown that most state-of-the-art AI applications would not be possible at all if there were no big data.

To define small data (Kitchin/McArdle 2016), we shall first present definitions of big data (Boyd/Crawford 2012; Hilbert 2016) and try to give the complementary definitions of the big data attributes while keeping in mind that there will be substantial overlap without a clear distinction between them. Although there is no consensus in the interpretation of definitions of big data (Ward/Barker 2013; Favaretto et al. 2020) showed that most of their interviewees agreed on the following: "Big Data are vast amounts of digital data produced from technological devices that necessitate specific algorithmic or computational processes to answer relevant research questions."

Most small and medium enterprises deal with little and highly structured data. These could be in the form of transaction logs, invoices, customer support or business reports and email communications. The volume of such data is relatively small, mostly averaging a few GBs. To get business intelligence from such data by machine-learning requires smart approaches that are suitable for small data. Most of the definitions about data encompass big data's key properties, which are popularly described by the "Vs" attributes. It started with the 3 "Vs" and quickly escalated to more than a few dozen (McAfee et al. 2012; Patgiri/Ahmed 2016).

The "Vs"

The first definition of big data according to the 3 "Vs" appeared in 2001 (Laney et al. 2001), with Volume, Variety, Velocity, and later was expanded by IBM to include Veracity. Additional to these four fundamental "Vs", many others appeared over the years, further enhancing the definitions of big data properties. Here, we will mention only the most prominent ones and try to contrast them against small data traits:

- *Volume.* This attribute presents the sheer quantity of the data, where the volume contains various data types. Social networks, e-commerce, IoT

and other internet services or applications generate vast data that traditional database systems cannot handle. By contrast, in small data, the volume is much smaller, making it easily accessible and easy to understand. For instance, think of an Excel table containing production figures in a company that are easily readable and interpretable by the CEOs.

- *Velocity.* This represents data generation rate (such as from the Internet of Things, social media), analysis and processing to satisfy specific standards or expectations. The data flow is massive and continuous. As an illustration, the velocity of big data can be expressed by the velocity of the data produced by user searches in real time. For instance, Google processes more than "80,000 search queries every second",[2] and these figures have already become obsolete while you were reading this paragraph. On the other hand, the small data accumulation is relatively slow, and the data flow is steady and controlled.
- *Variety.* Big data comes as structured, semi-structured and unstructured data with any combination of these, such as in documents, emails, texts, audio and video files, graphics, log files, click data, machine and sensor data. In contrast, small data is structured and in a well-defined format collected or generated in a controlled manner (tables, databases, plots).
- *Veracity.* This attribute of big data refers to quality, reproducibility and trustworthiness. Reproducibility is essential for accurate analysis, and veracity refers to the provenance or reliability of the data source, its context, and the importance of the analysis based on it. Knowing the veracity of the data avoids risks in analysis and decisions based on the given data. The quality of big data cannot be guaranteed. It can be messy, noisy and contain uncertainty and errors, meaning that rigorous validation is required so that data can be used. Small data, on the other hand, is produced in a controlled manner. It is less noisy and possesses higher quality and better provenance.
- *Value*. The value of data is hard to define; usually, it denotes the added value after producing and storing the data, involving significant financial investments. The data is more valuable when various insights can be extracted and the data can be repurposed. It has a lower value when it has limited scope and cannot be reused for different purposes, which mostly corresponds with small data collected for a specific task.

2 www.worldometers.info

- *Validity*. This is another aspect of veracity. It is the guarantee for data quality, authenticity and credibility. It indicates how accurate and correct the data is for the intended use. Big data has quantity which almost always results in a lack of quality. In such cases, substantial effort is necessary to preprocess and clean the data before it can be used at all. Consistent data quality, standard definitions and metadata are small data qualities because they are defined before the collection process begins.
- *Variability*. This characteristic refers to the fact that the meaning of the data can continuously be shifting depending on the context in which they are generated, which is in contrast to inflexible generation, as is the case in small data.
- *Volatility*. Data durability determines how long data is considered valid and how long it should be stored before it is considered irrelevant, historical or no longer valuable. Due to the volume and the velocity, this attribute is getting more important because it is directly related to storage complexity and expenses. Because of that, small data is inheritably easier to handle and could be archived much longer.
- *Viability*. The data should reflect the reality of the target domain and the intended task. The entire system could be fully captured, or of just being sampled. Using relevant data will provide robust models that are capable of being deployed and active in production. As we already saw, there are many examples where AI applications fail due to the mismatch of the training data and actual real data. Big data originate from broad and diverse sources, and it is not always possible to filter out domain-specific information. For instance, an NLP system built on textual content arising from popular books will fail in tasks where medical records, which are difficult to collect due to patient privacy protection concerns, are supposed to be processed.
- *Vulnerability*. While any security breach in systems storing big data has enormous consequences (e.g. leaked credit card numbers, personal accounts information), the damage is relatively low and limited in the case of small data.
- *Visualization*. Because of the technical challenges in storage and computing resources, different approaches must make the big data's insights visible and understandable. Simple charts, graphs and plots suitable for small data are not feasible for exploratory analysis in big data because of the volume and the number of data points with complex relationships. It is necessary to decrease data size before actual graphical rendering,

feature extraction and geometric modeling can be implemented. It is possible to employ clustering, tag clouds and network diagrams for unstructured data, correlation matrices, heat maps, treemaps, histograms, box, whisker plots, etc. (Olshannikova et al. 2016).

- *Virality*. This describes how quickly the data is spread or broadcasted from a user and picked up and repeated by other users or events.
- *Viscosity*. This is a measure of how difficult it is to work with the data, and it could be described as the lag between the data occurrence and the following conversion from data to insights. It appears primarily due to different data sources, integration of data flows and the subsequent processing phase.
- *Versatility*. This describes the extensionality and scalability of the data, adding or changing new attributes easily and rapidly expanding in size, against the data which is difficult to administer and have limited extensionality and scalability.
- *Vagueness*. This concerns the interpretation issues with the results being returned. The meaning of the produced correlations is often very unclear and misinterpreted as causation. Regardless of how much data is available, better or more accurate results are not possible. Small data has the advantage that it is easier to interpret and comprehend by humans, making it less prone to misinterpretations.
- *Vocabulary*. This is the ontology or the language used to describe an analysis's desired outcome, specific definitions and relationships with other terms. Big data has complex and unknown relationships; consequently, the employed ontology to describe these intricacies is complex.
- *Venue*. This refers to different locations and arrangements where data is processed, like multiple platforms of various owners with different access and formatting requirements. It could be distributed in a cloud (data warehouses, data lakes) or locally at the customers' workstations.

We can summarize all the big data traits and provide a more straightforward definition of them as "Enormous amount of rapidly generated and unstructured data that is too complex for human comprehension." Whereas, in contrast, a popular explanation of small data has been given as "data that connects people with timely, meaningful insights (derived from big data and/or 'local' sources), organized and packaged – often visually – to be accessible, understandable, and actionable for everyday tasks." (Small Data Group 2013).

In the end, there are complex relationships between big and small data where no clear line could be drawn. Big data could simultaneously be small data and vice versa. Small data could always be extracted from big data, it could become big data by extending it, or it can exist independently (Huang/Huang 2015).

Approaches to Big and Small Data

As no clear line can be drawn between big and small data, the same applies to choosing the approaches for handling and processing the data. Which method should be selected depends strongly on the data properties (see the Vs), the specific objective, the intended application task or problem. Sometimes, having a good sample of small data and applying traditional machine learning paradigms provide better results than employing big and noisy data with deep learning. Some expert knowledge and expertise are still necessary to make the right choice. Also, defining simple guidelines that consider the common traits of small data could help provide answers. Here, AI will play a more critical role in discovering an appropriate machine learning approach (AutoML) for a given problem (He et al. 2021).

The data attributes, some of which are already known before the collection or acquisition, some of which are discovered in the exploratory analysis phase, implicitly define the appropriate methodology. Common issues with small data could be identified according to the above-mentioned "V"-attributes as unlabeled, insufficient data, missing data, rare events and imbalanced data.

Unlabeled data are, by some of their attributes (volume, velocity and veracity), closer to big than small data. On the other hand, they exhibit small data properties like limited scope, weak relationality and inflexible generation. Insights of unlabeled data are unknown; the data contain a variety of novel outputs. Processing and analyzing unlabeled data require substantial effort involving organizational, technical and human resources. That is even more pronounced in specific tasks because the data type and acquisition are restricted to the task.

Insufficient amount of data. This is the most common case in small data. The data volume is relatively small and limited due to expensive collection (Krizhevsky et al. 2009) or generation and the low velocity or creation rate. However, sometimes the data are abundant but with low veracity, where the data points are not capturing the domain sufficiently. With insufficient data,

the features cannot be well discriminated and standard machine learning algorithms cannot provide proper modeling. The data are complete and fully labeled; however, due to the small number or to the low veracity of the data points, the boundaries between the target categories are less defined, and the ML algorithms will not generalize well for unseen data.

Missing data. Data can be missing due to technical issues during data collection (e.g. sensor failure) or human factors, e.g. flawed experimental or data collection procedure. It can also be interpreted as missing feature values or missing class samples. Preprocessing, consolidation or generation of missing samples or attributes is required to utilize such data.

Rare events. Rare or low probability events, known as outliers or anomalies, differ from the noise that arises from the random variance. They could result from natural variations, system behavior changes, faulty measurement equipment or foreign origin. Due to a large amount of "normal" data, detecting rare events or anomalies is usually relevant for big data and its property of veracity. However, very few and occasional observations contain valuable information of high interest for the intended task; hence, data with anomalies also fit the definition of small data.

Imbalanced data. Even when sufficient data exist, the data count with the desired features could be minimal because of uneven and non-uniform sampling of the domain. Therefore, the approaches usually applied in big data would be unacceptable and not provide the expected performance level. In many real-world applications, the minority class is more important because the information of interest belongs to this class, such as medical diagnostics, detection of banking fraud, network intrusion and oil spills.

Conclusions

We have tried to identify and present the limitations and prospects of big and small data by providing real-world examples and presenting SotA in different AI applications, which has led us to the following results: In the era of big data, small data is gaining more significance. Small data approaches promote AI's democratization, where small and medium enterprises can create tailored AI solutions without the need for massive data storage and computing infrastructure. Often, small data is a result of big data mining and analysis, transformed into smart data, which is more accessible, interpretable, actionable and provides the distilled insights of the big data.

The sheer volume of small data restricts many recent SotA approaches in machine and deep learning, indicating suitable algorithms for small data, combining expert knowledge with the black box approaches into hybrid AI systems. Looking forward, it seems clear that AI is developing towards a system able to learn and create an efficient AI system for the given data and tasks, which could open up machine learning and AI to non-experts.

References

Baevski, Alexei, Henry Zhou, Abdelrahman Mohamed, and Michael Auli (2020). "wav2vec 2.0: A framework for self-supervised learning of speech representations." In: *arXiv.org*. URL: https://arxiv.org/abs/2006.11477.

Boyd, Danah, and Kate Crawford (2012). "Critical questions for big data: Provocations for a cultural, technological, and scholarly phenomenon." In: *Information, communication & society* 15.5, pp. 662–679.

Brown, Tom B., Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal et al. (2020). "Language models are few-shot learners." In: *arXiv.org*. URL: https://arxiv.org/abs/2005.14165.

Ciresan, Dan Claudiu, Ueli Meier, Jonathan Masci, Luca Maria Gambardella, and Jürgen Schmidhuber (2011). "Flexible, high performance convolutional neural networks for image classification." In: *Twenty-second international joint conference on artificial intelligence.*

Davenport, Thomas, and Ravi Kalakota (2019). "The potential for artificial intelligence in healthcare." In: *Future healthcare journal* 6.2, p. 94.

Deng, Jia, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Fei-Fei Li (2009). "Imagenet: A large-scale hierarchical image database." In: *2009 IEEE conference on computer vision and pattern recognition*. IEEE, pp. 248–255.

Devlin, Jacob, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova (2018). "Bert: Pre-training of deep bidirectional transformers for language understanding." In: *arXiv.org*. URL: https://arxiv.org/abs/1810.04805.

Favaretto, Maddalena, Eva de Clercq, Christophe Olivier Schneble, and Bernice Simone Elger (2020). "What is your definition of Big Data? Researchers' understanding of the phenomenon of the decade." In: *PloS one* 15.2, e0228987.

Fedus, William, Barret Zoph, and Noam Shazeer (2021). "Switch Transformers: Scaling to Trillion Parameter Models with Simple and Efficient Sparsity." In: *arXiv.org*. URL: https://arxiv.org/abs/2101.03961.

Fjelland, Ragnar (2020). "Why general artificial intelligence will not be realized." In: *Humanities and Social Sciences Communications* 7.1, pp. 1–9.

Geirhos, Robert, Jörn-Henrik Jacobsen, Claudio Michaelis, Richard Zemel, Wieland Brendel, Matthias Bethge, and Felix A. Wichmann (2020). "Shortcut learning in deep neural networks." In: *Nature Machine Intelligence* 2.11, pp. 665–673.

Georgila, Kallirroi, Anton Leuski, Volodymyr Yanov, and David Traum (2020). "Evaluation of Off-the-shelf Speech Recognizers Across Diverse Dialogue Domains." In: *Proceedings of the 12th Language Resources and Evaluation Conference*, pp. 6469–6476.

Goldberg, Yoav (2016). "A primer on neural network models for natural language processing." In: *Journal of Artificial Intelligence Research* 57, pp. 345–420.

Goodfellow, Ian, Yoshua Bengio, and Aaron Courville (2016). "6.5 Back-propagation and other differentiation algorithms." In: *Deep Learning*, MIT Press, pp. 200–220.

Gottfredson, Linda S. (1997). "Mainstream science on intelligence: An editorial with 52 signatories, history, and bibliography." In: *Intelligence* 24.1, pp. 13–23.

Grant-Jacob, James A., Benita S. Mackay, James A.G. Baker, Yunhui Xie, Daniel J. Heath, Matthew Loxham et al. (2019). "A neural lens for super-resolution biological imaging." In: *Journal of Physics Communications* 3.6, p. 65004.

Guresen, Erkam, and Gulgun Kayakutlu (2011). "Definition of artificial neural networks with comparison to other networks." In: *Procedia Computer Science* 3, pp. 426–433.

Haenlein, Michael, and Andreas Kaplan (2019). "A brief history of artificial intelligence: On the past, present, and future of artificial intelligence." In: *California management review* 61.4, pp. 5–14.

He, Xin, Kaiyong Zhao, and Xiaowen Chu (2021). "AutoML: A Survey of the State-of-the-Art." In: *Knowledge-Based Systems* 212, p. 106622.

Hekler, Eric B., Predrag Klasnja, Guillaume Chevance, Natalie M. Golaszewski, Dana Lewis, and Ida Sim (2019). "Why we need a small data paradigm." In: *BMC medicine* 17.1, pp. 1–9.

Hilbert, Martin (2016). "Big data for development: A review of promises and challenges." In: *Development Policy Review* 34.1, pp. 135–174.

Hsu, Feng-hsiung (2002). *Behind Deep Blue: Building the computer that defeated the world chess champion*. Princeton University Press.

Hsu, Feng-hsiung, Murray S. Campbell, and A. Joseph Hoane Jr (1995). "Deep Blue system overview." In: *Proceedings of the 9th international conference on Supercomputing*, pp. 240–244.

Huang, Po-Chieh, and Po-Sen Huang (2015). "When big data gets small." In: *International Journal of Organizational Innovation (Online)* 8.2, p. 100.

Huang, Xuedong, James Baker, and Raj Reddy (2014). "A historical perspective of speech recognition." In: *Communications of the ACM* 57.1, pp. 94–103.

Jiang, Fei, Yong Jiang, Hui Zhi, Yi Dong, Hao Li, Sufeng Ma et al. (2017). "Artificial intelligence in healthcare: past, present and future." In: *Stroke and vascular neurology* 2.4.

Joseph, Siny, Vinod Namboodiri, and Vishnu C. Dev (2014). "A MarketDriven framework towards environmentally sustainable mobile computing." In: *ACM SIGMETRICS Performance Evaluation Review* 42.3, pp. 46–48.

Kitchin, Rob, and Gavin McArdle (2016). "What makes Big Data, Big Data? Exploring the ontological characteristics of 26 datasets." In: *Big Data & Society* 3.1. URL: https://doi.org/10.1177%2F2053951716631130.

Kolesnikov, Alexander, Lucas Beyer, Xiaohua Zhai, Joan Puigcerver, Jessica Yung, Sylvain Gelly, and Neil Houlsby (2019). "Big transfer (bit). General visual representation learning." In: *arXiv.org*. URL: https://arxiv.org/abs/1912.11370 p. 8.

Krizhevsky, Alex, and Geoffrey Hinton (2009). *Learning multiple layers of features from tiny images*. Technical Report. University of Toronto.

Krizhevsky, Alex, Ilya Sutskever, and Geoffrey E. Hinton (2012). "Imagenet classification with deep convolutional neural networks." In: *Advances in Neural Information Processing Systems* 25, pp. 1097–1105.

Lan, Zhenzhong, Mingda Chen, Sebastian Goodman, Kevin Gimpel, Piyush Sharma, and Radu Soricut (2019). "Albert: A lite bert for self-supervised learning of language representations." In: *arXiv.org*. URL: https://arxiv.org/abs/1909.11942.

Laney, Doug et al. (2001). "3D data management: Controlling data volume, velocity and variety." In: *META group research note* 6.70, p. 1.

Legg, Shane, Marcus Hutter et al. (2007). "A collection of definitions of intelligence." In: *Frontiers in Artificial Intelligence and applications* 157, p. 17.

Leviathan, Yaniv, and Yossi Matias(2018). "Google Duplex: an AI system for accomplishing real-world tasks over the phone.". URL: https://ai.googleblog.com/2018/05/duplex-ai-system-for-natural-conversation.html

Liu, Yinhan, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen et al. (2019). "Roberta: A robustly optimized bert pretraining approach." In: *arXiv.org*. URL: https://arxiv.org/abs/1907.11692.

McAfee, Andrew, Erik Brynjolfsson, Thomas H. Davenport, D.J. Patil, and Dominic Barton (2012). "Big data: the management revolution." In: *Harvard business review* 90.10, pp. 60–68.

McCarthy, John (1974). "Professor Sir James Lighthill, FRS. Artificial Intelligence: A General Survey." In: *Artificial Intelligence* 5.3, pp. 317–322 (https://doi.org/10.1016/0004-3702(74)90016-2).

Meijer, Gerhard Ingmar (2010). "Cooling energy-hungry data centers." In: *Science* 328.5976, pp. 318–319.

Metz, Cade (2016). "In two moves, AlphaGo and Lee Sedol redefined the future." In: *Wired*, March 16 (Available online at http://www.wired.com, checked on 6/28/2021).

O'Mahony, Niall, Sean Campbell, Anderson Carvalho, Suman Harapanahalli, Gustavo Velasco Hernandez, Lenka Krpalkova, Lenka et al. (2019). "Deep learning vs. traditional computer vision." In: *Science and Information Conference*. Springer, pp. 128–144.

Olshannikova, Ekaterina, Aleksandr Ometov, Yevgeni Koucheryavy, and Thomas Olsson (2016). "Visualizing big data." In: *Big Data Technologies and Applications*. Ed. by Borko Furht and Flavio Villanustre, Cham: Springer, pp. 101–131.

Patgiri, Ripon, and Arif Ahmed (2016). "Big data: The v's of the game changer paradigm." In: *2016 IEEE 18th International Conference on High Performance Computing and Communications; IEEE 14th International Conference on Smart City; IEEE 2nd International Conference on Data Science and Systems (HPCC/SmartCity/DSS)*, pp. 17–24. URL: https://doi.org/10.1109/HPCC-SmartCity-DSS.2016.0014.

Pham, Hieu, Qizhe Xie, Zihang Dai, Quoc V. Le (2020). "Meta pseudo labels." In: *arXiv.org*. URL: https://arxiv.org/abs/2003.10580.

Rajbhandari, Samyam, Jeff Rasley, Olatunji Ruwase, and Yuxiong He (2019). "Zero: Memory optimization towards training a trillion parameter models." In: *arXiv.org*. URL: https://arxiv.org/abs/1910.02054.

Rasley, Jeff, Samyam Rajbhandari, Olatunji Ruwase, and Yuxiong He (2020). "Deepspeed: System optimizations enable training deep learning models with over 100 billion parameters." In: *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pp. 3505–3506.

Rong, Guoguang, Arnaldo Mendez, Elie Bou Assi, Bo Zhao, and Mohamad Sawan (2020). "Artificial intelligence in healthcare: review and prediction case studies." In: *Engineering* 6.3, pp. 291–301.

Rosset, Corby (2020). "Turing-nlg: A 17-billion-parameter language model by Microsoft." In: *Microsoft Research Blog* 2, p. 13.

Russakovsky, Olga, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma et al. (2015). "Imagenet large scale visual recognition challenge." In: *International journal of computer vision* 115.3, pp. 211–252.

Russell, Stuart, and Peter Norvig (2009). *Artificial Intelligence: A Modern Approach*, 3rd ed. Edition, USA: Prentice Hall Press.

Schmidhuber, Jürgen (2015). "Deep learning in neural networks: An overview." In: *Neural networks* 61, pp. 85–117.

Schraudolph, Nicol N., Peter Dayan, and Terrence J. Sejnowski (1994). "Temporal difference learning of position evaluation in the game of Go." In: *Advances in Neural Information Processing Systems*, p. 817.

Schrittwieser, Julian, Ioannis Antonoglou, Thomas Hubert, Karen Simonyan, Laurent Sifre, Simon Schmitt et al. (2020). "Mastering atari, go, chess and shogi by planning with a learned model." In: *Nature* 588.7839, pp. 604–609.

Shoeybi, Mohammad, Mostofa Patwary, Raul Puri, Patrick LeGresley, Jared Casper, and Bryan Catanzaro (2019). "Megatron-lm: Training multi-billion parameter language models using model parallelism." In: *arXiv.org*. URL: https://arxiv.org/abs/1909.08053.

Silver, David, Aja Huang, Chris J. Maddison, Arthur Guez, Laurent Sifre, George van den Driessche et al. (2016). "Mastering the game of Go with deep neural networks and tree search." In: *Nature* 529.7587, pp. 484–489.

Strubell, Emma, Ananya Ganesh, and Andrew McCallum (2019). "Energy and policy considerations for deep learning in NLP." In: *arXiv.org*. URL: https://arxiv.org/abs/1906.02243.

Temam, Olivier (2010). "The rebirth of neural networks." In: *SIGARCH Computer Architecture News* 38.3, p. 349. URL: https://doi.org/10.1145/1815961.1816008.

Völske, Michael, Janek Bevendorff, Johannes Kiesel, Benno Stein, Maik Fröbe, Matthias Hagen, and Martin Potthast (2021). "Web Archive Analytics." In: *INFORMATIK 2020: Gesellschaft für Informatik*. Ed. by Ralf H. Reussner, Anne Koziolek, and Robert Heinrich, Bonn, pp. 61–72.

Ward, Jonathan Stuart, and Adam Barker (2013). "Undefined by data: a survey of big data definitions." In: *arXiv.org* URL: https://arxiv.org/abs/1309.5821.

Yang, Zhilin, Zihang Dai, Yiming Yang, Jaime Carbonell, Ruslan Salakhutdinov, and Quoc V Le (2019). "Xlnet: Generalized autoregressive pretraining for language understanding." In: *arXiv.org*. URL: https://arxiv.org/abs/1906.08237.

Young, Tom, Devamanyu Hazarika, Soujanya Poria, and Erik Cambria (2018). "Recent trends in deep learning based natural language processing." In: *IEEE Computational Intelligence Magazine* 13.3, pp. 55–75.

Zhang, Yu, James Qin, Daniel S. Park, Wei Han, Chung-Cheng Chiu, Ruoming Pang et al. (2020). "Pushing the Limits of Semi-Supervised Learning for Automatic Speech Recognition." In: *arXiv.org*. URL: https://arxiv.org/abs/2010.10504.

Zuboff, Shoshana (2015). "Big other: surveillance capitalism and the prospects of an information civilization." In: *Journal of Information Technology* 30.1, pp. 75–89.