In partnership with:

GSI TECHNOLOGY

RE · WORK

The Challenges, Successes, Progression & Failures of Processing in AI

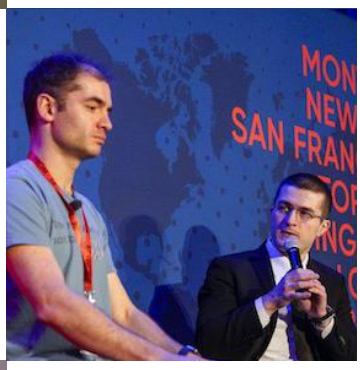
WHITEPAPER | 2021

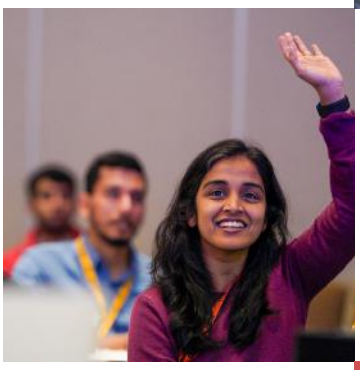
About RE•WORK

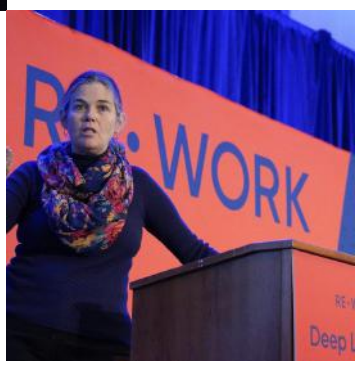
As well as creating digital content, RE•WORK is first and foremost the global leader in AI and Deep Learning events. RE•WORK creates and organizes globally renowned summits, workshops and dinners, as well as virtual events, bringing together the brightest minds in AI from both industry and academia. At each RE•WORK event, we combine the latest technological innovation with real-world applications and practical case studies. You can learn from global pioneers and industry experts, and network with CEOs, CTOs, data scientists, engineers and researchers disrupting their industries with AI. We also provide an analysis of current trends and innovations, through podcasts, white papers and video interviews. Additionally, we have an extensive on-demand video library of presentations from world-leading experts in AI. We cover topics such as Deep Learning, Machine Learning, AI in Healthcare, Women in AI, AI in Finance, Reinforcement Learning, Computer Vision, Autonomous Vehicles, Conversational AI, AI for Good, Responsible AI and more.

This White Paper is sponsored by GSI Technology, Inc. Founded in 1995, GSI Technology, Inc. is a leading provider of SRAM semiconductor memory solutions. They recently launched radiation-hardened memory products for extreme environments and the Gemini® APU, a memory-centric associative processing unit designed to deliver performance advantages for diverse AI applications. The Gemini APU's architecture features parallel data processing with two million-bit processors per chip.

The massive in-memory processing reduces computation time from minutes to milliseconds, even nanoseconds. Gemini excels at large (billion item) database search applications, like facial recognition, drug discovery, Elasticsearch, and object detection. Gemini's scalable format, small footprint and low power consumption, make it an ideal solution for edge applications where rapid, accurate responses are critical.







Contributors

Adebunmi Elizabeth Odefunso
Software Engineer & ML Practitioner | Purdue University

Adebunmi Odefunso is a machine learning engineer with years of experience in data collection, engineering and analysis. She is experienced in database programming and administration. She loves data and likes it being well used. She currently works on building computer vision models and making quality datasets for computer vision studies available.

Mark Wright
Director of Marketing | GSI Technology

Having completed his BSc in Electrical Engineering and the University of Toronto, Mark went on to work as a senior design engineer, later working at various smart technology firms, applying his technical knowledge to both their product development and marketing. Mark is currently Director of Marketing at GSI Technology where he works on Similarity Search and AI in Aerospace.

Rosana de Oliveira Gomes
Lead ML Engineer | Omdena

Rosana is a PhD in Astrophysicist, with a 10 years background in academic research. Currently she is transitioning careers into Data Science and Artificial Intelligence for social good, looking for opportunities to apply her knowledge into the nonprofit and humanitarian sector. In the Omdena community, Rosana has participated in Artificial Intelligence challenges, implementing data-driven solutions to real world problems. In particular, she was one of the managers in the Omdena project presented in this webinar, in which a team of 35 persons build an application for cyclone response, identifying what items need to be supplied when a disaster takes place. (Also includes contributions from: **Harini Suresh, PhD Researcher, MIT; Erum Afzal, ML Engineer, Omdena**).

.

Shaina Raza
AI Researcher | Ryerson University

Shaina Raza is an AI researcher, Computer science Advisor and Instructor who is finishing PhD in Computer Science from Ryerson University, Canada. Her doctoral research is on developing algorithms for the news recommender systems, with focus on various machine learning and the cutting-edge deep learning algorithms. She also developed her own algorithms to address the unique challenges in the news recommenders. Currently, she is focussing on the AI4Good (artificial intelligence for good) to bring the technology towards the social good. For this she is implementing algorithms to address the 17 SGD goals of the United Nations.

Shivam Mathura
Director of Strategy | COTA Inc

Shivam studied Mathematics and Computer Science at New York University, later working at that same University for four years as a research assistant in game theory and other AI-based fields. Shivam now uses his computer science knowledge to lead the strategy at healthcare-based company, COTA.

Additional Contributing Authors:

Kemal Akkaya, Arjuna Madanayake, Udara De Silva, and **Sravan Pulipati**
Florida International University, Miami, FL

Josep M. Jornet, Kaushik Chowdhury, Francesco Restuccia, and **Tommaso Melodia**
Northeastern University, Boston, MA

Soumyajit Mandal and **John Shea**
University of Florida, Gainesville, FL

Aditya Dhananjay
Pi Radio, Brooklyn NY

Jay Dawani and **Vassil Dimitrov**
Lemurian Labs, Toronto, ON

Introduction

Data makes the world go round, or at the very least, is the underlying heartbeat of all Artificial Intelligence (AI) and Machine Learning Development. Whilst 2020 was somewhat of a testing and tumultuous year, the development of AI infrastructure and the collection of data for this purpose continued, at a faster pace than previously seen. Albeit behind the scenes, 2021 could be the year in which we see rollout of AI in societal settings. None of this, however, would be possible without the large swathes of data collected, which in itself, could potentially be a bigger challenge than creating the models themselves.

The key components of a data science project, and the different kinds of challenges associated with them, play an important role in identifying data limitations. Availability, cost, privacy, ethics and processing data collections all stand in the way of wide scale development at all industry levels, as well as rollout for consumer use. It is widely believed that we will not be experiencing 100 years of progress in the 21st century, but rather closer to 20,000 years. That is dependent, however, on the large amounts of data needed to be collected. Therefore it is only through consistent experimentation that the future potential of machines can be met.

The stark reality is that we have moved through a 'generation of big data' to daily generation, from the many mobile applications, messages sent, and the 3.5 billion daily searches. One of the key challenges to this acceleration is in the data exchange required between processors' and memory, as well as data transfer and storage capabilities. Furthermore, data availability is not the same as data integrity, data retention or data reliability.

While all these concepts have some similarities, they are also very different from one another. Hence, potential scarcity of usable and good quality data could create a world in which building suitable models that work cross-industry is so close, yet so far.

From ElasticSearch to Fake news and Edge-computing to Data Limitation, the following white paper takes the concept of data and addresses a variety of both accelerating and restricting factors, as well as discussing relevant industry developments and their effect on the current state of data.

> *One of the biggest limitations to workload acceleration is the limitation in data exchange required between processors' and memory."*
>
> -Mark Wright, GSI Technology

Contents

Data Limitations in Common Industry and Non-Profit Applications

Rosana de Oliveira Gomes, Lead Machine Learning Engineer, Omdena; Harini Suresh, PhD Researcher, MIT; Erum Afzal, ML Engineer, Omdena.

The scope of Artificial Intelligence (AI) is getting broader day by day. Data is becoming increasingly more necessary to solve ongoing real-world problems, and the same is true for the challenges that come with them. Data challenges are highly related to the definition of a problem, data types, as well as availability, and privacy, among other topics.

In this chapter, we discuss the most common data challenges in the private and non-profit sectors, as well as the currently trending solutions in the AI industry. In order to illustrate the challenges of dealing with different data types, we present case studies involving topics such as mobility and sustainability, amongst others.

Background:

The key components of a data science project, and the different kinds of challenges associated with them, play an important role in identifying data limitations. In particular, project setting, data collection and preprocessing are the initial stages of a project in which practitioners find most of the data challenges.

Having a clear problem statement is not only crucial to avoid data limitations, but also to ensure success as a whole. This entails having clear goals, from which a team can brainstorm on data usage. In some cases, the team has data available and needs to figure out what insights can be extracted from it. In other cases, data is not available and a data collection phase becomes necessary.

In a data collection stage, the most common limitations are:

1 *Availability & Cost:* knowing where to find available and relevant sources is not an easy task. The most common options to overcome this challenge are the usage of open source data or acquiring data from data mining companies, which naturally come with a cost. Another possibility is the introduction of artificial data, as discussed later.

2 *Privacy & Ethics:* another great concern after finding data is to guarantee that the privacy of individuals is not compromised. In the same way, concerns about data misuse in terms of ethics, such as public manipulation [1] and bias against gender or ethnicity has led to efforts on data policy and regulation [2]. In particular, the European Union has adopted data protection regulations which are valid for all member countries [3].

Finally, when it comes to preprocessing, each kind of data has its own challenges, going from resolution in images to a slang in text. In the next section, we introduce some of the most common data types and later discuss specific limitations in use cases, along with possible solutions.

Different Types of Data:

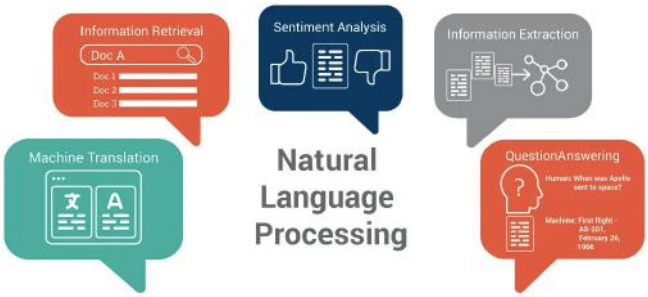
The digital transformation alongside adaptation of AI in all spheres of life has made data become a thorny subject as termed in a Frobe article [4].

Whether it is for forecasting, analysis or automation, all depends upon the good quality of clean data and dataset.

Features in a dataset have their own types, categories or classifications that being unique in nature bring their own kind of challenges for each category. It is crucial to understand what, when, and why which kind of data is used in order to get better insights from it. In what follows, a few specific classification of data types is discussed along with their challenges and appropriate solutions.

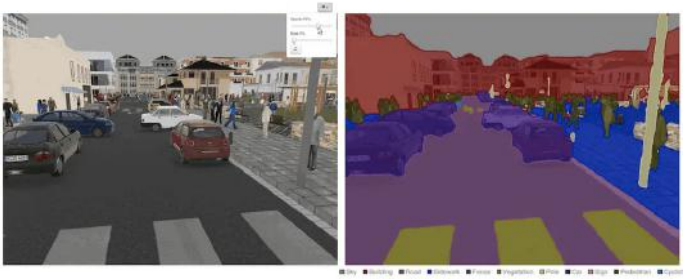
Numerical: the most common type of data, it usually appears in the form of charts and tables with sequences of discrete or continuous numerical values. Numerical data is applied to all sorts of problems being forecast the most common, for example, estimating consumer expenses in finance. Limitations associated with numerical data include statistical significance and interpretation. In order to properly interpret numerical data, it is necessary to have enough data able to provide proper distribution suitable for statistical analysis of the problem. Accounting for thorough units and data scaling checking during preprocessing, along with assigning proper error metrics for models are good practices for ensuring a valid interpretation of numerical data.

Text: normally in the form of words, sentences and paragraphs. It can be collected as text reviews, discussions on social media, or simply private text. Some of the challenges with text data are their raw and unstructured form. Nowadays, text processing libraries such as Natural Language Toolkit (nltk) [5] are extremely powerful when it comes to get insight from the textual information.



Source: [https://sigmoidal.io/machine-learning-terms/]

Image: Frequently used for object identification or classification tasks. Regarding object detection and classification, common challenges are identifying different objects that belong to the same class (such as chairs with different shapes), along with objects in different scales or perspectives. A solution for such challenges is the use of pretrained models, such as YOLO [6], which is already trained in a massive data-set named Common Object in Context (COCO) with 1.5 million object instances and 80 object categories [7]. In the case of detecting specific objects which are not part of pretrained datasets, a second option is to manually annotate images with annotation tools [8]. Another challenge related to image data include several types of image extensions which are not always compatible with models. For that, one can use computer vision libraries which help convert the extensions [9].



Source: [https://sigmoidal.io/machine-learning-terms/]

Geospatial: Applying machine learning algorithms in real-time and having the benefit of identifying the location comes with the usage of geospatial data. Geospatial data finds its application beyond Earth, in a different planet where the location/presence of water bodies is vital for the existence of human life. Geospatial data is simply a combination of satellite imagery associated with geographical information. Geospatial data is commonly used to tackle problems which require specific visual perspective, such as natural disasters. While accessing this data is challenging considering the cost factor, a major problem faced is the resolution level of the image making it difficult to run a few models. However, the introduction of GANs [10] have come to the rescue where the resolution of the tiled image can be increased by altering the pixels leading to the smoothening of edges and images.

Sound: advances in speech recognition are making sound data become a popular data type in industry. Applications range from personal assistants and dictation softwares to medical applications in transcriptions and ultrasound analysis [11]. Regarding speech recognition, suitable training data still poses a significant limitation on model accuracy, as availability of data with a broad range of accents and voice tones is still lacking overall in industry. A solution currently ongoing for such a challenge is collecting more diverse and unbiased voice data from different areas of the world [12]. A more state of the art limitation about sound data comes from background noise, being a current topic of broad research in the field [13] and having as possible solution the training of background that can be subtracted from the raw data during preprocessing.

Data Creation

The necessity of the generation of artificial data is based on the cause and effect of data limitations. The cause of data challenges can be the availability, amount and cost of data. Data creation helps solve these limitations.

Explicitly addressing the issue of data availability: scarcity of data makes it impossible to build suitable performance models. Furthermore, when there is little data available, a considerable amount of work has to be invested on expanding the dataset. Now when it comes to applying AI to social impact and democratizing data-driven solutions, the cost of data plays an important role in what can be done. Data can be unaffordable for nonprofit organizations or local institutions, which are usually the ones tackling social issues.

Possible solutions to overcome data scarcity and high costs are by the creation of new data. Deepfake, popular innovative AI technology, can be used for data augmentation based on existing data to increase its quantity [14], especially for artificial image, sound or video data. When dealing with text data, a possibility to overcome data availability is by a system to collect data from resources which do not require any cost. A simple form of doing so is through surveys and on a higher technical level, data can also be collected by chatbots. Survey forms can have queries/labels and can be shared across the world for humans to enter data. Likewise, a simple chatbot with basic input data and queries can be built to extract more relevant information from users. However, such data creation techniques raise the topic of privacy and ethics once again, and care needs to be taken to make data anonymized and safe.

The proof of concept is the final part of the artificial data generation, through the verification of the model performance.

Another common form of text data collection is web scraping, in which one performs keyword similarity mining from online sources such as social media like Twitter, Facebook, Reddit, etc.

A common limitation associated with this practice is the issue of disparate sources [15], as the data collected will differ in structure, length and description. An important step on preprocessing web scraped data is to balance the data according to the parameters mentioned above. Another important task is checking the use licensing of such data to avoid legal conflict.

However, in the case of supervised learning this is directly related with the quality of the data labeling. When alternative sources of data are used, it is difficult to mine labeled datasets. In order to overcome such issues, annotations help crafting the labeled dataset with a bit of human effort. Annotations range from text to image data with compact user-interfaces provided by annotation tools [16, 17]. When humans label the data and generate a labeled dataset, it can be used as a base source of data in order to build reliable models.

Case Studies

Three cases studies are discussed below, in order to better illustrate the data limitations in the case of text, video and geospatial data.

Addressing vulnerability with NLP

Data Type: Text

Case: Identifying online abuse

Description: Most online communications happen via private chat, comments on the content shared at social media or in forums. In such contexts, predators attack the persons through abusive chat. NLP models can help predicting the risk of abuse in online environments [18].

Limitation: Data privacy and availability. Personal stories and messages in online forums are most of the time not publicly available. Public text associated to abuse such as comments and reviews is usually biased, given that reported content tends to be removed from websites. Moreover, identifying meaning of slang and non-text features, as emojis, is also a common NLP challenge.

Solution: Collaboration across sectors in order to provide access to anonymized data from online forums and platforms where predatory behavior is common, such as for example gaming scenarios. Acquisition of more survivor stories through anonymized surveys can help identifying typical language and non-text features in abuse stories and contextual information, to help building robust models.

Computer Vision for Emergency Response

Data Type: Geospatial

Case: Emergency Response

Description: satellite images provide a privileged perspective of locations on Earth during emergencies. In the case of natural disasters or human conflict, common road accesses may be blocked and satellite images provide a way to capture the current status of the region and help providing faster help. DataKind has applied object detection to identify the number of refugees in a camp for an awareness campaign [20].

Limitation: image resolution / cost of data
Solution: GANs for improving resolution [10]/ possible use of drones for data collection on closer perspective.

Computer Vision Applications for Autonomous Driving

Data Type: Video

Case: Self-driving Cars

Description: AI steps into the shoes of human brains and drives the car by automating the knowledge/learning gathered from humans. It is responsible for automatically detecting humans, lanes and traffic signals.
Limitation: While considering safety on building self-driving cars, the volume, diversity and accuracy of the training data has to be kept in mind

Solution: The data is collected from Radar, Lidar and Camera increasing the redundancy factor where each model can succeed with one type at the least. The accuracy of the data can be improved with annotations focussed on critical 3D data labeling. [17, 19]

Conclusion

Data Science and AI are among the most powerful tools in the tech industry in the 21st century. Behind all the advances AI can bring to society, there is data driving all the learning processes of models and tools used by it. Therefore, data limitations play a critical role on how the whole field of Data Science evolves and it is by overcoming these challenges that important advances are made.

Data limitations can be strongly problem dependent. Each field of AI is developing its own libraries to deal with preprocessing of text, image, among others. It is possible that the preprocessing stage for specific data types becomes automated by standard industry practices in the near future.

As the need for more data in more diverse forms becomes more clear to several sectors, such as speech and face recognition, additional open source and collective collaborations on data collection and labelling are expected [21]. In a similar fashion, there have been several efforts on data policy and regulations across the globe by governments [2,3] and institutions [22]. The topic of AI ethics has gained significant impulse in 2020 and is expected to continue growing as more diverse communities start to have more access to data-powered technologies.

References

[1] Facebook-Cambridge Analytica data scandal

[2] Information Privacy Law

[3] General Data Protection Regulation

[4] Types of Data by Forbes

[5] Natural Language Toolkit

[6] Yolo model

[7] Coco Dataset

[8] Label Box

[9] Solaris Package

[10] Ultra-dense GAN for satellite imagery super-resolution

[11] Deep Learning for Medical Ultrasound Analysis

[12] Forward Artificial Intelligence for All

[13] Dealing with Noise Problem in Machine Learning Datasets

[14] Data Augmentation with GANs

[15] Omdena - Disparate Data Sources

[16] Machine Learning Accelerated Data Annotation

[17] Challenges & Solutions for 3D LiDAR Annotation & 3D Data Sets in 2020

[18] Omdena Online Abuse of Children

[19] Training Data for Autonomous Driving

[20] Data Kind Refugees Project

[21] Open Data Kit

[22] AI for Peace - Human Rights and Democracy

Convergence of ElasticSearch, ANN and Compute-in-Memory

Mark Wright, Director of Marketing, GSI Technology

We are in the midst of technological convergences that will disrupt industries and provide new industry opportunities. One such alignment is the use of approximate nearest neighbor (ANN) with in-memory acceleration processing to provide near real-time responses for billion-scale elastic search operations.

Elasticsearch is a search engine that takes JSON requests for document searches and provides JSON data as results. The elastic search data format is a document which is structured data encoded in JSON. While Elasticsearch started as a search engine for text, the database can be any type of data with each document having a unique ID and a data type. The structure is "schema-free" allowing the documents to be defined to whatever the user needs and further flexible in what and how they are indexed for searching. In different examples of elastic search databases, documents can be:

- Pictures: used to identify consumer pictural search requests or similar interests.
- Network data logs: used to identify network intrusion, anomalies, or load imbalances.
- Product receipts: used to identify customer purchasing patterns and improve stock rotation.
- Network architecture: used for automatic sharding and replication.
- Text documents: used to find specific literary instances.
- Text documents with one-to-many mappings: used for computer assisted translation.

Elasticsearch was designed to be distributed and thus scalable in infrastructure, and flexible for local server, remote server, or cloud-based operation. With an open and restful API structure, the extensible search engine is easily used with plugins. One such plugin is from GSI technology that adds the following improvements: hardware accelerated ANN, the use of vectors for multi-modal search, and merging score results.

Elasticsearch counts on its distributed computing support for scalability, and its fast speeds are in the order of seconds for million-scale database searches. Because of its distributed nature and sharding support, Elasticsearch supports duplication of data for parallelizing the search and speeding up search for larger databases. Core Elasticsearch uses a computationally heavy exhaustive match (match all) which slows it down or makes it very expensive in duplicate hardware to support large-scale database search. Approximate Nearest Neighbor (ANN) search is a technique that can be used to increase the database size that can be searched by first looking for similarity in common groupings then doing the final search within those one or more groupings. While ANN provides a methodology for Elasticsearch to support very large databases, such as those at billion-scale entries and above, ANN is compute exhaustive also and has been a challenge to accelerate due to the constraint of moving the databases between GPU or CPU cores.

> ***The ability to use the GSI Elasticsearch plug-in without the requirement to reindex documents to support vector search is a huge simplification for customers."***
>
> - Pat Lasserre, Director of Strategic Sales and Business Development, GSI Technology.

One of the biggest limitations to workload acceleration is the limitation in data exchange required between processors' and memory. A major drawback of the Von Neumann architecture used in modern processors is the overhead of data transfer between processors and storage. The CPU must fetch data for every operation it does. This architecture is even more inefficient in an offload acceleration environment. The performance of such systems is limited by the speed at which data can be exchanged via memory by the host requesting the operations and also compute engines performing the operations.

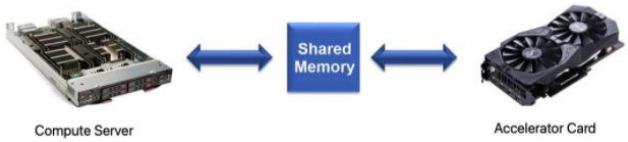


Figure 1. *Constant Data Transfer Reduces Performance of Accelerating Servers.*

Architectures that reduce the flow of data from the memory are being studied or evaluated to reduce the Von Neumann bottleneck. The Von Neumann bottleneck is particularly egregious when you're dealing with memory intensive artificial intelligence applications. The operation of AI-related applications depends on the fast and efficient movement of massive amounts of data in memory. Trained data-bases need to be loaded into working memory and vectorized input queries then processed and also loaded for comparison functions to operate.

One proven technology that is already in the market is the Associative Processing Unit or APU. The beauty of in-memory acceleration is that storage itself becomes the processor. This is not a massive array of processing cores with cache memory close by, but a memory array with compute units built into the read-line architecture. Thus, the APU is differentiated by having the memory array capable of accelerating compute. This type of "accelerated" processor has been shown to accelerate performance by orders of magnitude while reducing workload power consumption of standard servers.

> ***On-prem or cloud based, ES with ANN addresses savings in DB platforms which are very costly from plant, network, server, and energy perspectives."***
>
> - Avidan Akerib, VP of Associative Computing, GSI Technology

> ***This is a solution to database search for which GPU adoption is very low."***
>
> - Avidan Akerib, VP of Associative Computing, GSI Technology

Combining Elasticsearch, ANN, and APU acceleration provides less latency and more queries per second. It also provides capability for billion-scale database search support. Running AWS Elasticsearch or using AWS compute resources and running Elastic.co Elasticsearch equally accelerate through the use of the built-in API plugin capability. The figure following shows use with the additional step of the AWS search sending the ANN lookup for APU acceleration done as a transparent-to-the-user step.

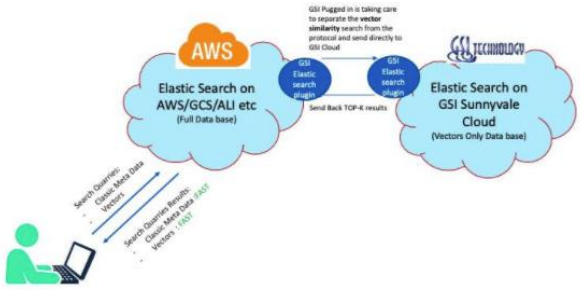


Figure 2. *Transparent Elasticsearch Cloud Acceleration using Plug-ins.*

Performance for standard GIST-960-euclidean and SIFT-128-euclidean are shown. This test is for a 1M entry training database so ANN was not required or used. A flat search acceleration was used to improve times but also provided a system power consumption savings coming in at 240W total, versus 325W without the APU (test run locally and not in cloud). All cases use a single Intel Xeon E5-2680 v3 2.5GHz CPU as the host.

Engine	QPS	Latency [ms]	P95 Latency [ms]	Recall@10
Elastic 7.6	3.29	303.96	337.89	1
Vespa 7.190.14	9.14	109.33	148.90	1
w/1APU	15.8	63.3	64.7	1

Figure 3. *SIFT-128-euclidean vs SIFT-128-APU Results.*

Engine	QPS	Latency [ms]	P95 Latency [ms]	Recall@10
Elastic 7.6	0.57	1752.74	1850.74	1
Vespa 7.190.14	1.32	756.61	955.63	1
w/1APU	8.1	123	135	1

Figure 4. *GIST-960-euclidean vs GIST-960-APU Results*

Preliminary results for Deep1B performance show 40mS P95 search latency in a system of 1 Intel 5115 Gold Xeon CPU and 2 APU's. Vectors are DEEP1B with 96 features. Recall@10 is better than 0.9. The ANN is split into 1M clusters with 1000 records per cluster. Getting results in less than internet latency from such a large database: what new opportunities can this convergence enable?

> ***Retail fashion are a particular driver of mul-timodal search because they rely heavily on visual search since style is often difficult to describe using text."***
>
> - Pat Lasserre, Director of Strategic Sales & Business Development. GSI Technology

Further Reading

Scalable Semantic Vector Search with Elasticsearch

The Limitations & Advances of Data Availability

Adebunmi Odefunso, Software Engineer & ML Practitioner, Purdue University

What is Data Availability?

It is no longer new that we are in the era of big data. We all generate it daily. Every tap on one mobile application, every message sent etc. becomes a part of big data. About 3.5 billion searches data is everywhere, it is locked up, or scattered everywhere that it is not so available for use.

Data availability is having access to needed data at the time it is needed. This definition focuses on three important aspects of data availability.

A) Accessibility
B) On-demand
C) Mission-centered data

Accessibility is very close to availability in that it is concerned about being able to retrieve data. Accessibility is about ease of use of data. Decision making by organizations is guided by information retrieved from data, not having access to the data that will provide this data is a challenge. On-demand is about the timeliness of the data. Data is said to be available if one can access and use it at the time it is needed. Mission-centered is about having access to the data that will provide information that meets the need of the user/researcher. So, if I need network traffic data for time series study, but the data I got does not have time features, even if all other information is needed, the data is not available for me.

What Data Availability is Not

Data availability is not the same as data integrity, data retention or data reliability.

While all these concepts have some similarities or meeting points, they are different from one another. Available data may not be accurate, complete, or reliable.

Limitations of data availability.

Data availability has a lot of challenges which has been a bottleneck for researchers and organizations alike, including:

1. Data compatibility
2. Storage failure
3. Server/Network failure
4. Cost
5. Poor data quality

Data compatibility: Data types or formats required on different platforms or applications can be different. This affects the availability of data in these scenarios. Another compatibility issue happens with legacy data, these were usually in a strict format that are hard to work with now. However, there are now several data transformation platforms available as well.

Storage failure: This is about the limitation of storage spaces and devices. No matter how good a storage device is, there will be a time they age away. And in situations where the cloud is being utilized as it is common today, once you grow your data storage to a size, there are no more storage spaces and thus your data availability is affected.

Server/Network failure: When the network crashes, access to data is impeded. Data storage and retrieval is also impeded.

Cost: In this age of cloud computing, all the benefits of the cloud comes with an additional cost. It takes a lot of budgeting and planning to maintain having data available and accessible every time it is needed. The various cloud companies have tiers of storage that makes cutting down costs possible. Albeit; the cost of this service can limit data availability.

Poor data quality: While data quality is not a determinant of data availability, it can impede data availability. Data with very poor quality can not be used to make any informed decision, thus it is only occupying storage space, but it is not available. The quality of data collected can be affected by the framework behind the collection points. Poor quality data will not only waste storage space, it will also waste cost and effort, thus it is important that the window of data collection by organizations be well structured to make data available for use.

The Solutions and Future of Data Availability

Technology keeps changing face frequently and new solutions to problems that seem very tough to overcome are springing up every day. New technologies are rising by the day on data format conversion; the future of this will be automation. Automated data format conversion is the future of data format transformation. This will make data highly available and accessible.

High quality data starts with a high-performance data processing pipeline. The data collection or entry point must be made flexible and sensitive. Early detection of errors at collection can save the system. Poor quality data is mostly due to poor software implementations, system level issues or poor collection pipeline design. When the entry level is made sensitive/responsive, when something is lacking in data, there will be early notification. This will reduce errors in data and thus leads to better data availability. On the cloud and in telecommunications, network resiliency is getting better by the day and the future for this is brighter. There are various activities going on to improve resilience that will have a ripple effect on data availability.

Data storage for the future is the hybrid cloud. The cloud computing framework is moving towards removing boundaries and making data storage and accessibility effective. The use of hard drives and flash drives for data storage is becoming less costly, better resilient and more agile daily. This progress sets the ground for a better data availability.

The future of cost incurred in data availability either on premises or in the cloud will be highly affordable in the near future. Cloud companies have different storage arrangements that make saving cost possible. Knowledge is also expanding daily for data pipeline management for saving costs. The future for data availability is positive and exciting.

Data Roadblocks in ML & AI

Shivam Mathura, Director of Strategy, COTA inc

With the convergence of data proliferation and advancements in computational power, artificial intelligence (AI) has already transformed the world around us. However, there is an ever-growing list of models leading to harmful outcomes due to their increased complexity, lack of interpretability, lack of reproducibility, and increased bias. Facial analysis algorithms have demonstrated higher rates of error for people of color and implementation by law enforcement agencies reinforce these disparities at every stage of the justice system [1]. Hiring practices using natural language processing (NLP) to score applicants in the tech industry penalized candidates through encoded gender bias, amplifying patterns in the existing gender employment gap [2]. A predictive algorithm for resource allocation used cost of care as a proxy for healthcare needs. The output reflected the history of inequality by requiring Black patients to be considerably sicker than White patients in order to qualify for the same level of coverage [3]. Translating theory into successful, responsible AI systems is constrained by a multitude of real-world factors. The foremost consideration is the data foundation upon which the models rest. Machine learning models reflect the data they are built on, and so a failure to overcome the data challenges within the development lifecycle can lead to poor performance and even poorer results.

As large-scale datasets become more widely available, it's important to make sure the dataset is appropriate to address the questions of interest. The dataset should include information on potential confounding variables to mitigate bias and increase interpretability of the model. Confounders are variables that affect both the inputs and outputs of the model and can therefore lead to spurious correlations between inputs and outputs. Since confounders are context-dependent, domain knowledge is necessary for identifying the right variables to include for modeling.

In addition, the dataset should be large enough and diverse enough to avoid making spurious correlations and to support the conclusions drawn from the models. The larger the dataset, the higher the probability that one will find an event of interest even if the event is spurious. Rigorous statistical modeling, including power calculations, can help reduce the rate of such false discoveries. Acquiring a dataset with the appropriate sample size will often require the aggregation of multiple data sources. Each data source has its own potential biases, based on how the data was collected and processed. Harmonizing the different data sources requires extensive characterization and analysis but will ultimately lead to less biased and more predictive models.

Over time, datasets will continue to grow and evolve, posing a specific challenge to models created on older datasets. Data drift refers to changes in the data infrastructure and architecture over time, while concept drift refers to the change in relationship between input and output variables over time. The speed and magnitude of both affect model performance, with faster and more drastic changes in new datasets leading to worse performance. Anticipating how these relationships will change and periodically retraining the model with updated datasets can mitigate both issues.

Dataset volume describes both the number of samples and the number of attributes (or features) for each sample. While increasing dataset size and complexity are important for making new discoveries, they also pose distinct challenges.

As the dataset size (both in terms of sample size and feature dimension) increases, the runtime and resources needed to train machine learning algorithms also increases. Unfortunately, the scaling factor is usually not linear, meaning that doubling of dataset sample size can require more than twice the computational resources. This can render certain algorithms ineffective and unusable with large datasets, requiring a comprehensive approach to data storage and processing and careful model considerations in order to begin analysis.

Increases in the feature dimension can also affect model performance through the curse of dimensionality. With more attributes than samples, it becomes quite easy to overfit to the noise in the training data and decrease model performance on out of sample datasets. In addition, many machine learning algorithms are based on measuring similarity (or distance) between data samples, which is complicated by large feature dimensions. Dimension reduction algorithms can help avoid these problems by selecting only the most pertinent features. Machine learning algorithms can be sensitive to class imbalance, an issue that arises in datasets with non-uniform distribution of samples across classes. Minority classes will be harder to classify, leading to decrease in overall model performance. The challenge of class imbalance is further amplified in large datasets. Therefore, it is important to make sure that all data classes are represented as equally as possible.

Similarly, the dataset needs to be evaluated for its quality. In the context of data quality, completeness is the degree to which the dataset includes expected values. The level of completeness required can vary by model and should be considered before using the source as an input.

In much the same way that the sample size must meet the minimum requirements for statistical validity, the dataset needs to meet the minimum requirement for the availability of features. Virtually all data sources will have gaps, but understanding the difference between what values are present, what values are null, missing, or unusable, and what percentage of values is expected can identify potential errors in data collection and if steps are required to clean or correct them.

Other aspects like the consistency and accuracy of the data sources are equally important when it comes to data quality and preparation. Consistency refers to the ability of the input sources to adhere to the logical rules that define them. This includes rules like valid values for each attribute, valid relationships between attributes, and each feature having the same data type or format. Not only should each source follow its own internal rules, but the aggregation of sources should not create any logical conflicts.

Accuracy can be complex to assess, as it is meant to measure the degree to which the dataset is correct or truly models what it is intended to capture. It can be quantified by observing the conformity to expectations of completeness and consistency, and if the source is subject to duplicates, conflicts, or invalidity. Accuracy should also be informed by the provenance of the data: the origins, the transformations, the derivations, the metadata, and the complete audit trail from generation to destination. A deeper understanding of the data lineage and its context within the domain to build a good model.

As machine learning and artificial intelligence becomes more ubiquitous, the tools and techniques to overcome these challenges will also advance. While there are many important steps to building ethical and effective ML/AI solutions, it often begins - and ends - with the data.

References:

[1] How is Face Recognition Surveillance Technology Racist?

[2] Amazon scraps secret AI recruiting tool that showed bias against women

[3] Dissecting racial bias in an algorithm used to manage the health of populations

Processing limitations on Enterprise AI - Is GPT-3 the ultimate solution?

Shaina Raza, PhD Candidate, Advisor Computer Science, Ryerson University, Toronto, Canada

In this paper, we discuss the wonders of Artificial Intelligence (AI) and the limitations of AI on Enterprise. We elaborate this with an example of the latest AI model GPT-3, a model that has surprised the world by all the wonders it can do. For example, it can replace the computer programmers, coders, doctors, typists, futurists and industrialists through its capabilities as a sole model. But at the same time, we must think of the computational cost, hardware/software liabilities and the unforeseeable biases that could be created as a result of training such high-level models.

In 2001, the futurist Ray Kurzweil said in his article 'The Law of Accelerating Returns' [1] that technological change is exponential. We will not be experiencing 100 years of progress in the 21st century, instead it will be more like 20,000 years of progress. Today we are seeing these changes in the IT world. Machine intelligence surpassing human intelligence. Recently, the launch of GPT-3 (Generative Pre-trained Transformer 3) [2] by OpenAI, a company co-founded by Elon Musk, has surprised the world. The GTP-3 is a computer-based model that uses deep learning to produce human-like-text. The GPT-3 can respond to any text that a person types into the computer with a new piece of text that is appropriate to the context. The GPT-3 is the most powerful language model ever produced in history. With almost 175 billion parameters, the GPT-3 model can replace human coders, programmers, data analysts, data scientists, web developers, human operators, doctors like radiologists and many industrialists.

However, it is unreasonable to expect GPT-3 to be all good. There will be some limitations on processing while using the enterprise Artificial Intelligence.

The GPT-3 works miraculously and comes from using an extraordinarily larger model. The cost of GPT-3 comes with an immense increase in the model parameters and the data size that is beyond the capacity of normal to high-tech computer systems. Even the GPT-3 paper's core message was less about its performance on benchmarks, and more about the discovery of solving the highly complex tasks that have never been performed before by any other model. The impressive text generation by GPT-3 can solely be attributed to the massive computational power, scale, and the number of resources used for training the model. This means, the Enterprises AI needs the growing layers of cyberspace to deploy a GPT-3 model.

The GPT-3 algorithms also go through massive amounts of data to recognize patterns and draw conclusions. These models are trained with labeled data that rely on scenarios the model will encounter in the real world. For example, the doctors must tag each x-ray to denote if a tumor is present and what type it is. Only after reviewing thousands of x-rays, can an AI correctly label new x-rays on its own. This collection and labeling of data is an extremely time-intensive process for humans.

One should also not expect the GPT-3 to be a universal solution. The Artificial Intelligence used in this model may be excellent at pattern recognition, but one can't expect it to operate on a higher level of consciousness.For example, if we ask the GPT-3 to enter someone's home and make a cup of coffee, this includes finding the coffee grinds, locating a mug, identifying the coffee machine,

adding water and hitting the right buttons. This is referred to as artificial general intelligence where AI makes the leap to simulate human intelligence.
Another major limitation of GPT-3 is its algorithmic bias, which is known to have biases towards gender, race, and religion. This arises from biases in training exponential amounts of data that reflects the societal views and opinions. It further reinforces the fact that the GPT-3 is not a standalone intelligent system. Like any form of new technology, there can be a significant cost of purchase and a need for on-going maintenance and repair of GPT-3. The model will also require regular upgrades in order to adapt to the continually changing business environment. The return on investment needs to be carefully considered by a company before going ahead and implementing the GPT-3 system.

Creativity remains a vital component of a successful marketing campaign. The AI generated models can make decisions but are not necessarily creative.
Unlike machines, humans can think and feel, which often guides their decision making when it comes to being creative. Obviously, the AI can assist in terms of helping to determine what the consumer is likely to click on - from colour preferences to style and price. But when it comes to originality and creative thinking, a machine simply cannot compete with the human brain. We still need both humans and machines.

AI and ML are evolving technologies. Today's limitations are tomorrow's successes. The key to success is to continue to experiment and find where we can add value to the organization. Although we should recognize AI's limitations, we shouldn't let it stand in the way of the revolution. This paper has concentrated the limitations of AI to the Enterprise. There are some real advantages of AI to the Enterprise which we have already mentioned, however, ultimately, Artificial Intelligence is only going to become more and more efficient and effective.

Keeping ourselves aware of the current limitations of AI helps ensure we do not set ourselves up for unrealistic expectations.

References:

[1] Kurzweil, Ray. "The law of accelerating returns." In Alan Turing: Life and legacy of a great thinker, pp. 381-416. Springer, Berlin, Heidelberg, 2004.

[2] Brown, Tom B., Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan et al. "Language models are few-shot learners." arXiv preprint arXiv:2005.14165 (2020).

AI in 6G Wireless Communication Networks

Kemal Akkaya, Arjuna Madanayake, Udara De Silva, & Sravan Pulipati, Florida Int. University; Josep M. Jornet, Kaushik Chowdhury, Francesco Restuccia, & Tommaso Melodia, Northeastern University; Soumyajit Mandal & John Shea, University of Florida; Aditya Dhananjay, Pi Radio; Jay Dawani & Vassil Dimitrov, Lemurian Labs.

Emerging 6G wireless communication networks will require cross-cutting developments in mm-wave and sub-THz wireless technologies, software defined radio (SDR), software defined networks (SDN), AI, and machine learning (ML) across all layers of the networking protocol stack. 6G systems will be based on concepts such as dynamic spectrum access (DSA) and cognitive radio (CR) and will enable unprecedented capacity and network access, while providing the enabling technologies for augmented/virtual reality (AR/VR), multi-agent robotics, wireless IoT, autonomy, and more. The expected explosion in network complexity and size of 6G networks will necessarily require an unprecedented level of real-time inference and autonomous intelligence to manage cross-layer decision-making at the physical layer and above. Fundamental research on next-generation wireless, such as cognitive beamforming for spectrum-awareness in 6G networks and RF fingerprinting of wireless IoT devices will require access to existing state of the art test-beds such as Colosseum and PAWR, as well as the development of entirely new test-beds. This chapter explores new and exciting ideas on the convergence of AI/ML and mm-wave 6G wireless network research towards next-generation data networks, robotics, autonomy, wireless IoT, and industry 4.0.

Next-Generation Wireless Networks with AI and SDN

Over the last few decades, network systems have become an indispensable part of our lives, serving our basic and crucial needs. In particular, the Internet evolved from an information service to a major utility service providing lifeline services to people, organizations, businesses, and governments. COVID-19 further accelerated and clinched this reality: network systems will continue to grow, particularly on the wireless side, and become ubiquitous with diverse components in almost every domain. This continuous growth of network systems in terms of scale and capabilities come with pressing challenges on how to manage, control and maintain such a huge and complex system of systems without compromising its safety, security and efficiency since the scale and complexity are beyond the limits of manual control. We believe that after being managed manually with teams of administrators, emerging network systems are at a critical conjunction where they will need to develop some sort of agency (i.e., automation) to be able to self-address many of the above challenges as existing approaches outstretch their limits. Similar to all biological systems, network systems also need the ability to adapt to continuous changes in regard to the way they are used, secured and managed. If this transformation does not occur now, current network systems will eventually become overwhelmed and under-performing, thus risking the fulfilment of many of the lifeline services upon which society depends.

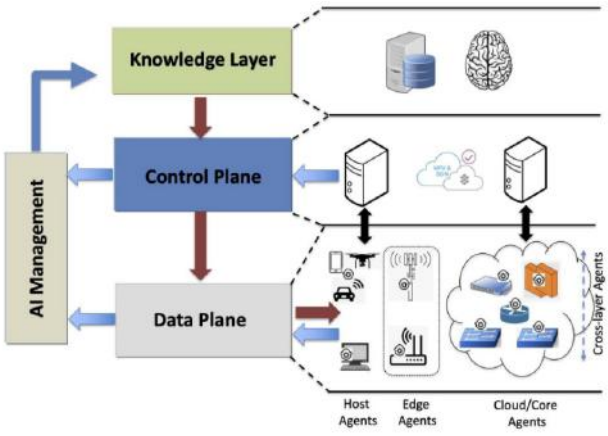


Fig. 1. *AI-based SDN for next-generation wireless networks.*

Intelligence in SDNs: Advancements in AI and machine learning (ML) provide excellent mechanisms to tackle these transformational challenges as we seek to bring agency and self-decision making to next generation wireless network systems that will automatically help addressing scalability, manageability and security. 5G networks already include a service-based architecture in the network core that significantly differentiates itself from the 4G design. This architecture calls for heavy utilization of SDN and network function virtualization (NFV) to enable flexibility and elasticity targeted for different application needs [1, 2]. The upcoming 6G networks will take this approach further by supplementing SDN and NFV with AI/ML capabilities, resulting in a knowledge layer. As seen in Fig. 1, the envisioned architecture will bring self-management and self-security that will take next-generation 6G systems to another level. Specifically, the knowledge layer will act as a centralized brain that can make decisions by collecting data from both the control and data planes managed by SDN. Data collection will be achieved by utilizing an agent-based system at the data plane. AI/ML will be integrated within every layer of both the hardware and software, making this a comprehensive AI-enabled approach that spans end-user devices, base-stations, radio-access networks, and the core network as is detailed in the next sections.

Motivation from the DARPA Spectrum Collaboration Challenge

The RF spectrum is a resource that is shared among a diverse set of users that are distributed across space, often mobile, and have diverse and often time-varying quality-of-service requirements for their traffic. Most existing spectrum usage policies follow one of two approaches. In the first, a frequency band is allocated by a regulatory body to a specific user over some geographic area.

This approach often results in inefficiencies because the user does not fully occupy the entire band over the entire space all the time. In the second approach, the spectrum is open to any unlicensed user that follows certain power and bandwidth requirements. In the US, this is primarily the Industrial, Scientific, and Medical (ISM) bands, such as the 2.4 GHz and 5 GHz bands that are used for WiFi. This approach, too, often results in inefficiencies because users interfere with each other and do not coordinate or broker their use of the spectrum. This motivates the use of spectrum-sharing techniques, in which users work together to deliver their traffic over a set of shared frequency bands.

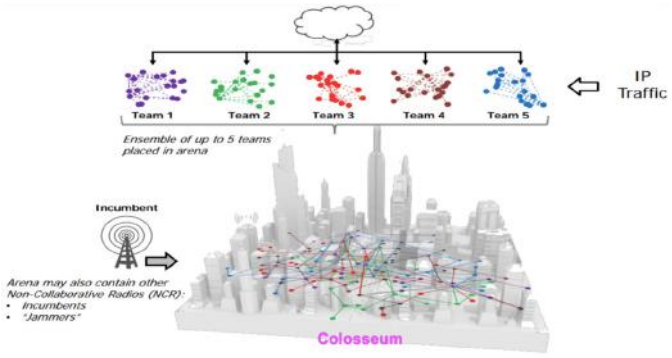


Fig. 2. *Real-time physics modeling on DARPA Colosseum for AI-enabled SDRs.*

The DARPA Spectrum Collaboration Challenge (SC2)

The AI-based SD Grand Challenge was run by the Defense Advanced Research Projects Agency (DARPA) from 2016-2019. The SC2 started with 90 teams from academia and industry, as well as a few individuals, both from the US and around the world, competing to win up to $3.5 million in prizes. DARPA describes it as "the first-of-its-kind collaborative machine-learning competition to overcome scarcity in the radio frequency (RF) spectrum." Contributor Shea of the University of Florida (UF) was co-lead of Team GatorWings, which won the SC2 championship.

During the SC2, teams developed intelligent communication algorithms and implemented them on software-defined radios (SDRs). Teams had to work completely independently and were barred from exchanging any information about their team's algorithms or strategies. The performance of teams' intelligent radio designs were evaluated in mixed cooperative-competitive matches in which the real radio transmissions were sent over an emulated channel using Colosseum, billed by DARPA as the world's largest channel emulator. As shown in Fig. 2, up to 5 teams of 10 radios each would be placed in an emulated physical environment, along with other scenario radios, such as jammers and incumbents that cannot tolerate significant interference. Teams could collect performance and spectrum data that could be used to train ML algorithms from their radios during various scrimmages, preliminary events, and special freeplay jobs.

Fig. 3. *Snapshot of a visualization generated during the SC2 Championship Event.*

In October 2019, the SC2 Championship Event was held at Mobile World Congress in Los Angeles. During this event, the final ten teams competed in a wide range of radio scenarios with different types of traffic, mobility, and incumbents. Fig. 3 shows a snapshot of a visualization from the last match of the Championship Event. The bottom right shows a terrain map along with color-coded and numbered markers indicating positions of each node for each of five teams.

In the top right is a spectrogram, which shows the radio spectrum over time and how it is being used by each of the teams. Finally, on the left are the cumulative scores of the teams, along with a bar graph indicating progress toward a 35k "finish line" score that would end the match. SC2 demonstrated that spectrum sharing can be achieved among teams that utilize diverse radio algorithms and spectrum-access strategies and that handle diverse traffic. In fact, in the SC2 Championship Event, teams demonstrated the capability to drastically increase spectrum efficiency through multiple teams using the same spectrum at the same time through careful allocation across transmitter-receiver pairs and use of robust communication schemes.

Implementation of Intelligent Radio Algorithms

Building an intelligent radio network requires a diverse set of talents and skills. Many radio algorithms are inherently real-time in nature, and robust algorithms that rely on advanced signal processing and error-control coding schemes are very computationally intensive. In addition, many ML algorithms are also computationally intensive, especially during the training phase. Thus, teams working in this field must be skillful in implementing problems across FPGAs, graphics processing units (GPUs), and general-purpose processors (GPPs). For instance, the communication algorithms in Team GatorWings' SC2 radio were implemented across an FPGA and GPP, and also included a highly-parallelized Viterbi decoder on the GPU. A block diagram illustrating the fundamental structure of Team GatorWings' SC2 radio and the assignment to processing resources is shown in Fig. 4.

Intelligent SDR Platform: ML algorithms leveraged the use of a GPU, especially during the training phase. The FPGA was integrated onto National Instruments

USRP X310 radios that were used in the competition, and low-level signal processing including frequency, mixing, filtering, up/downsampling, and spectrum sensing was implemented using a RF Network on Chip (RFNoC) framework. The remaining communication stack (such as equalization, modulation, coding, and resource allocation) and online ML algorithms were almost fully-custom and implemented in multi-threaded C++. ML algorithms for optimization of spectrum sharing used PyTorch, and TensorFlow was used for spectrum understanding. Future work on AI-enabled wireless will require similar interdisciplinary teams with a diversity of software and hardware expertise.

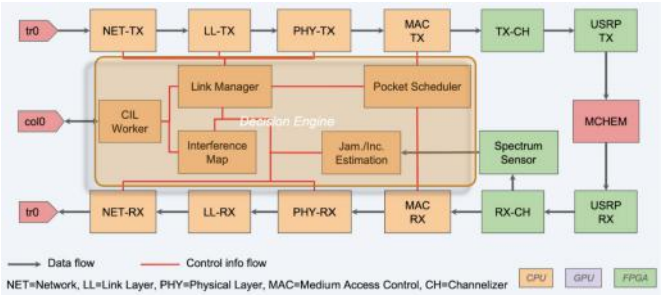


Fig. 4. *Structure of Team GatorWings' SC2 Radio and Resource Assignment.*

Frontier Communications at MM-Wave and THZ

Expanding Spectrum: Irrespective of the development of novel spectrum sharing methods, the need to provide higher data rates to an ever-increasing number of wirelessly connected devices will continue to motivate the exploration of higher frequency bands for communication. Currently, in 5G networks, the lower-end of the millimeter-wave spectrum (under 100 GHz) has been adopted. Moving forward, in 6G systems, frequencies above 100 GHz will enter the game [3, 4], not only to support data-rates in excess of 1 Terabit-per-second (Tbps) in dense networks, but also as the enabler of innovative sensing systems.

The very large bandwidth available at terahertz-band frequencies (from 300 GHz to 10 THz), which comes in the form of multiple hundreds of GHz-wide windows, only interrupted by the presence of sharp water absorption lines, provides RF communication systems with opportunities usually only available to optical systems. The short wavelengths (less than 1 millimeter) and the meaningful photon energy of THz signals enable both precise localization and non-damaging material identification. Moreover, communications and sensing are not two separate processes, but benefit from being integrated: after all, the topology of the shared medium, including the presence of different types of obstacles and the composition of the air itself, ultimately determines the best communication and networking strategies to follow, thus making THz communications a new playing field for AI algorithms and data-driven ML approaches.

The resulting applications of the millimeter-wave and terahertz-band spectrum are plenty. On the ground, ultra-broadband point-to-point links can serve as backhaul links with capacities comparable to that of wired optical fiber systems but at a fraction of the cost, bringing the benefits of 6G to rural areas or in emergency zones (e.g., after an Earthquake or tsunami). Similarly, joint communications and sensing can be utilized to enable ultra-low latency high-capacity communication and sensing in networks of autonomous vehicle networks. Up in the sky, the opportunities are even larger, because blocking obstacles are less likely and, moreover, the lower precipitable water content at higher altitudes leads to even larger channel bandwidths. Large swarms of unmanned autonomous systems, simultaneously collecting and sharing hyper-spectral data for (literally) in-the-cloud data processing are not out of the realm of possibilities.

Networks of autonomous satellites (i.e., multi-agent systems) with terahertz radios in Earth orbit or even around the moon, Mars, or Venus, can simultaneously extract scientific data while serving as a 6G space-borne communication infrastructure [5].

Intelligent Spectrum: AI will play a key role across the system, from hardware to software and at the intersection of the two. For example, AI can be utilized to learn and compensate for the highly non-linear responses of the transmitter and receiver, which result from the frequency multiplication and amplification chains needed to generate terahertz signals with meaningful power. Beyond that, AI can be leveraged to design new robust and secure communications strategies (e.g., waveforms, modulation and coding techniques) that exploit such non-linear behavior. Similarly, AI will be needed to efficiently estimate the communication channel and compensate for it. Among others, due to the very small wavelength of terahertz frequency signals, any minor (e.g., millimetric) change in the network topology or propagation medium can lead to major changes in the channel impulse response. An AI-driven joint communication and sensing framework for channel estimation and equalization becomes a must to enable reliable communications. Ultimately, AI can also be utilized to control the many knobs (learnable parameters) related to hardware by implementing ML algorithms within millimeter-wave and terahertz transceivers.

AI for Wireless Transceivers

Cognitive Radio: Traditional spectrum licensing models cannot satisfy the exponentially-growing demands for wireless spectrum. However, moving to dynamic sharing models that allow additional users to exploit the unused spatio-temporal regions ("white spaces"') within existing bands, such as those demonstrated during SC2, is intrinsically challenging because spectrum access patterns are stochastic and fluctuate on multiple timescales. Moreover, existing spectrum sensors cannot monitor the spatio-temporal domain on such timescales due to the very high data rates (e.g., 100 Gbps for Nyquist-rate digitization of a single 5 GHz bandwidth beam at 10-bit resolution), resulting in incomplete spectral awareness. These awareness gaps also result in security vulnerabilities that can be exploited by malicious users. Addressing these spectrum usage challenges requires a new generation of DSA algorithms that can improve effective channel capacity, link latency, and user-perceived data throughput, thus enabling rapid growth in wireless applications such as autonomous vehicles, AR/VR, and industrial automation.

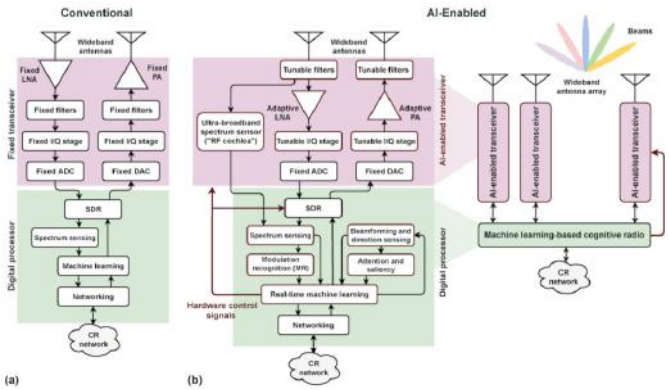


Fig. 5. *(a) Conventional, and (b) AI-Enabled Wireless Transceiver Architectures.*

Realizing DSA algorithms suffer from two key challenges. Firstly, while ML is promising for improving DSA, existing ML-based efforts are limited to software-driven approaches that do not consider relevant RF front-end properties (bandwidth, tunability, linearity, noise, and dynamic range). Secondly, existing efforts are focused on the "legacy" sub-6 GHz bands where spectral crowding is currently a major problem, and do not consider the unique properties of the emerging mm-wave and sub-THz bands where most future wireless developments will occur. AI-enabled mm-wave wireless transceivers can fill this gap by closely

integrating AI algorithms with programmable RF front-ends to enable spectrum awareness and DSA. The ML-based spectrum monitoring and analysis capabilities of such cognitive radio (CR) receivers enables fine-grained spatio-temporal DSA algorithms that increase channel capacity and wireless data rates within both licensed and unlicensed mm-wave bands. Such capabilities also improve spectral situational awareness for detecting anomalous or malicious spectrum usage. Fig. 5 compares the architectures of conventional and AI-enabled transceivers.

Spectral Attention: Ultra-broadband spectrum monitoring (consisting of feature extraction followed by ML, as shown in Fig. 5(b)) is required for spatio-temporal DSA. The ML algorithms can use a selective attention mechanism (analogous to similar concepts in natural language processing and computer vision) to address the "fire hose" challenge of how to effectively process large volumes of 3D spectrum data (varying in frequency, direction, and time) to detect useful white spaces. The chosen DSA parameters (operating frequencies, directions, time slots, modulation types, and transmit power levels) are then used to program the receiver front-end, which includes reconfigurable components such as tunable antennas, low-noise amplifiers (LNAs), and filters. This closed-loop adaptation process enables "hardware-in-the-loop" self-tuning and self-healing to achieve high-level goals, such as autonomously maximizing the received signal-to-noise and interference ratio (SINR) and thus the channel capacity.

Adaptive Circuits: A key challenge for AI-enabled receivers is the identification of RF signals with unknown modulation, center frequency, and bandwidth. This is critical in both sub-6 GHz and mm-wave bands due to the presence of undesired interferers (blockers) close to the weak signals of interest. This challenge can be addressed using programmable RF front-ends that maximize the under algorithmic control. SINR of the receiver for particular frequency bands under algorithmic control. For example, a programmable passive RF band-pass filter (BPF) before the LNA can provide adaptive blocker rejection under the control of a deep belief network (DBN)-based modulation recognition (MR) algorithm [6]. Spectro-temporal RF features can then be efficiently extracted from the selected bands by biologically-inspired real-time spectrum analyzers (known as "RF cochleas") [7], as shown in Figs. 5(b) and 6. Low-complexity space-time array processing algorithms can then implement spectral awareness and attention based on the extracted features. Finally, ML-enabled signal detection and identification algorithms select desired signals from the attended regions while rejecting unwanted ones (such as blockers).

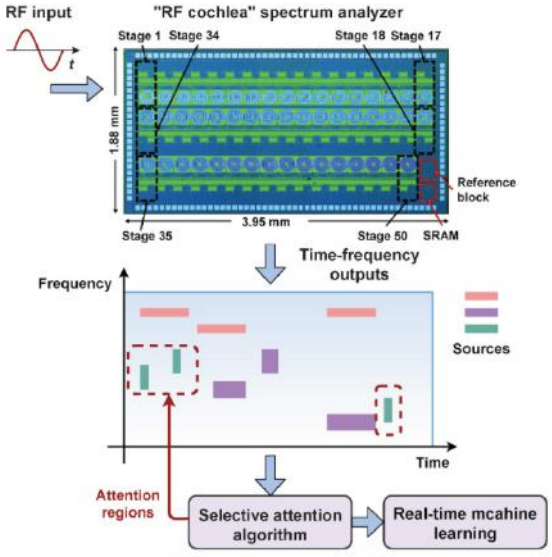


Fig. 6.*AI-driven selective attention mechanism using an "RF Cochlea" Spectrum Analyzer.*

Adaptive Computing: AI-enabled wireless transceivers require real-time computing platforms. A key requirement for such real-time systems is consistent and low latency for handling high-data-rate I/Q samples - either those received from high speed analog-to-digital converters (ADCs) or those being transmitted to digital-to-analog converters (DACs). Emerging applications require reconfigurable hardware like FPGAs to deploy AI

algorithms on these samples at 6G bandwidths. Xilinx RF SoCs are promising for such applications since they integrate parallel high-speed ADCs and DACs with a fully-reconfigurable FPGA fabric and ARM Cortex processors (see Fig. 7). These devices can efficiently process high-speed data streams by using a combination of custom ML hardware accelerators and software code running on the ARM cores, thus allowing compute-intensive AI on the edge.

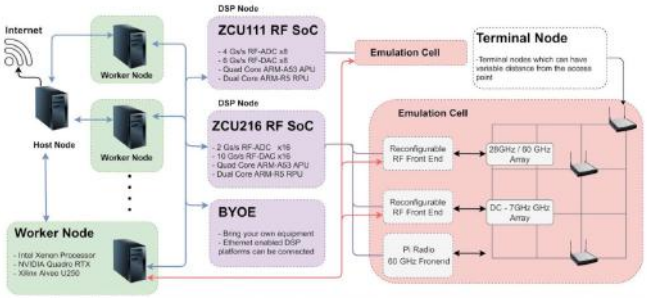


Fig 7. Block Diagram of an Xilinx RF SoC-based Reconfigurable Computing Platform for AI-enabled Wireless Transceivers.

Compute acceleration using Xilinx's Vitis AI flow enables a new RF SoC-based software solution for running 6G AI algorithms in real-time. Acceleration is introduced by the built-in deep learning processing unit (DPU), which can perform multiply-and-accumulation (MAC) and non-linear activation functions in parallel. This accelerated environment is suitable for superfast ML inference at low latency, as required for AI-enabled mm-wave and sub-THz transceivers.

Spectrum Sharing by Design: Today, spectrum sharing operators must rely on cloud-based spectrum access systems (SAS), which statically determine if the spectrum is available for a specific time period over a geographical region. However, it is easy to see that this centralized manual approach lacks scalability, and it does not allow for fine-grained real-time spectrum management. In other words, today spectrum sharing is not completely efficient, for both operators and spectrum owners alike. Conversely, a scalable and effective solution would be to let wireless devices opportunistically discover which spectrum sub-bands are currently available among ongoing licensed transmissions. This approach would severely boost spectrum usage without the need of central coordination. In other words, we need a closed-loop approach where both the transmitter's spectrum access and the receiver's waveform demodulation strategies are self-optimized in real time for maximum performance and minimum interference with licensed users. We need to enable the transition of spectrum sharing from a nice-to-have feature that few devices can perform, to a widespread technology that IoT devices will utilize "by design," as an integral part of their operations. However, achieving this goal is extremely daunting due to the strict real-time constraints of wireless domains, the resource-constrained nature of wireless devices, and the unpredictable nature of the wireless spectrum. The proposed AI-enabled wireless transceivers can provide a solution to these problems.

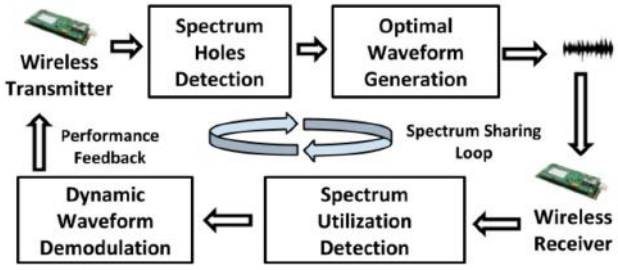


Fig. 8. Block Diagram of a Spectrum-Sharing-By-Design Wireless System.

Figure 8 shows the key operations of a spectrum-sharing-by-design wireless system that utilizes the proposed AI-enabled transceiver architecture. The approach is data-driven in nature, and leverages feedback from the receiver to continuously adjust the transmitter's strategy according to the ongoing performance. First, the IoT transmitter performs spectrum sensing to learn not only the current location of spectrum "holes," but also learn to automatically associate a given spectrum hole with a set of waveform parameters, such as the modulation

/coding scheme (MCS) and bandwidth, so as to maximize spectrum utilization in a variety of channel conditions. The
receiver, on the other hand, will (i) identify the relevant transmission among other concurrent licensed waveforms; and (ii) learn to identify the current transmission's parameters and demodulate the transmission accordingly. The receiver performance -- in terms of throughput, bit error rate (BER) or similar -- will be sent as feedback to improve the performance of the transmitter's algorithms.

AI for Radio Identification and Hardware-Level Security

As has already been noted, today's wireless spectrum is exceptionally crowded. According to the latest Ericsson's mobility report, there are now 5.7 billion mobile broadband subscriptions worldwide, generating more than 130 exabytes per month of wireless traffic. In addition to the clear commercial need, 5G and IoT have been identified as critical technologies for national security. The recent report "The Global Innovation Sweepstakes: A Quest to Win The Future," by the Atlantic Council's Scowcroft Center for Strategy and Security, identified 5G/IoT as one of a few cutting-edge technologies that are shaping an unprecedented technological revolution that will have far-reaching social, economic, and geostrategic consequences.

Radio Frequency ML Systems: Traditional spectrum sharing approaches based on fixed frequency allocations have led to fracturing and poor utilization. Thus, the wireless community has heavily invested in various DSA modes, although economic, regulatory, and enforcement issues have resulted in slow traction towards actual deployment. Specifically considering the target wireless application areas that will be radically transformed by city-scale deployment of small form-factor IoT devices, DSA does not explicitly involve considerations of security or energy savings. Finally, traditional authentication mechanisms that rely heavily on cryptography-based algorithms and protocols are not well-suited to the IoT, as they are usually too computationally expensive to be run on tiny, energy-constrained devices.

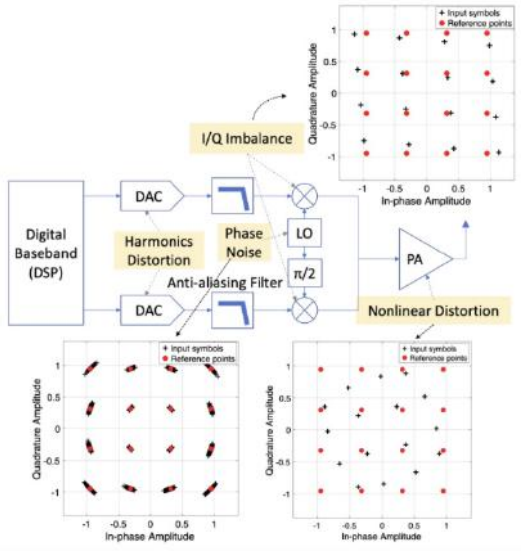


Fig. 9. Typical transceiver chain with various sources of RF impairments.

Clearly, to realize the promise of IoT, new clean-slate approaches are needed to achieve efficient spectrum utilization by (i) reducing transmission overhead in a way that is independent of specific standards; (ii) enable device authentication using immutable characteristics of a device, beyond cryptographic energy-hungry techniques, and (iii) lessen, if not eliminate, control signaling. In our vision, devices will learn small-scale hardware-level imperfections of their neighbors, which typically manifests in phase noise, I/Q imbalance, frequency and sampling offset, and harmonic distortions [8], to identify and classify specific transmitters, protocols and protocol-settings. Because of these unique impairments, a number of wireless devices operating on the same baseband signal will necessarily transmit two slightly different RF waveforms. We can then obtain a "fingerprint" of the wireless device by estimating the RF impairments on the received waveform and associating them to a given device [9].

We show in Fig. 9 a block diagram of the transceiver chain and the main causes for shifts in a 16-QAM constellation, with the red and black points showing expected and observed complex-valued I/Q samples. These shifts impart a unique fingerprint to a given device and signal. *The power of ML, specifically convolutional neural networks (CNNs), can be used to identify devices and modulation types based on these fingerprints. Furthermore, this capability can be integrated into wireless protocol design to increase spectrum efficiency in ways that are resilient to adversarial action.*

Conceptually, the approach is simple: CNN-based radio fingerprinting is used to eliminate device-identifying MAC ID fields contained in the header. CNN architectures suitable for implementation within embedded systems further reduce the need for periodic control signaling by the transmitter, e.g., to indicate the choice of modulation. Deep learning goes beyond "shallow" neural networks by autonomously extracting extremely complex features thanks to the very large number of parameters (typically 106 or more). Thus, deep neural networks can analyze unprocessed I/Q samples without the need of application-specific and computational-expensive feature extraction and selection algorithms [10]. Earlier work has explored the limits of RF fingerprinting using ML by analyzing a dataset of 400 GB of signal traces from over 10,000 radio transmitters as part of a DARPA-funded effort. These transmitters used off-the-shelf (COTS) 802.11b/g/n WiFi, and Automatic Dependent Surveillance-Broadcast (ADS-B) standards. This concept has also been used to demonstrate fingerprinting of 5G base stations deployed in the POWDER PAWR community-scale testbed in Salt Lake City, UT [11], thus motivating a push towards the concept of "shared infrastructure as a resource" for UAVs [12], a segment that several industry estimates predict to grow to $17 Billion by 2024 [13].

Deep learning has many other applications in RF identification, for example detection and automated classification of fully-autonomous drones using micro-doppler radar signatures [14,15].

AI for Spectrum Management

There is still a lot of work to do to improve the efficiency of spectrum sharing and to make it more practical for use in commercial and government systems. Optimizing the use of the radio spectrum across users, time, frequency, and space results in high-dimensional, distributed, and non-convex optimization problems. While conventional approaches break down under these conditions, ML techniques may offer viable approaches. From a single-team perspective, they can be modeled as reinforcement learning (RL) problems, but the problems are moreappropriately modeled as stochastic games or multi-agent RL (MARL) problems, for which the scientific literature and commercial developments are still quite limited. Adding in the effect of multiple users/teams significantly increases the dimensionality of the problem, which makes collecting enough training data problematic.

AI in Spectrum Sharing:

For users to accept the use of ML solutions to these problems, we require mechanisms to ensure that the ML solutions offer better performance than conventional approaches, offer reasonable actions in the presence of new inputs, and are secure against manipulation and hostile attacks. One of the key developments of SC2 was the use of an interchange language, developed by the teams, that allowed the intelligent radio networks to exchange information that would enable spectrum sharing. For SC2, this information included GPS coordinates of radios, spectrum usage plans, and performance measures.

In addition, performance metrics were determined by DARPA, and the ultimate match scores were created according to DARPA's metrics and under DARPA's control. Future commercialization will require new approaches that enable information interchange among users, that protect the privacy of the information exchanged, and the incentivize users to provide correct information. Working on applying ML to spectrum sharing problems currently requires a lot of domain-specific knowledge, so new approaches are needed to expose the fundamental resource allocation problems in a way that they can be worked on by the broader community of ML researchers. The use of new ML-based spectrum sharing techniques will require forward thinking by regulatory bodies, industry, academics, and standards bodies.

AI for Software Defined Radio (SDR):

The use of physics-based simulation models for training AI algorithms is a crucial component of 6G wireless systems. Colosseum is currently the world's largest (256x256) RF-Channel emulator for cloud-based RF research and development, and consists of two overarching constituent components: (i) a pool of 128 SDR resources, which defines a common platform upon which to build radio experiments; and (ii) a massive channel emulator (MCHEM) with 128 additional SDRs that emulate the interactions of radio waves in the physical world with sufficient veracity so that from any one radio's perspective, it appears to be operating in an open-air environment (see Fig. 10). The radio resource pool consists of off-the-shelf Ettus Research USRP X310 SDRs mated with high performance rack servers. There are no channel emulators currently in the market capable of supporting the computation and bandwidth needed to compute the interactions of hundreds of radios in real-time. As such, Colosseum has been custom designed and built using an extensive and dedicated GPU and FPGA processing hardware infrastructure that will be leveraged in the upcoming OpenBeam standard.

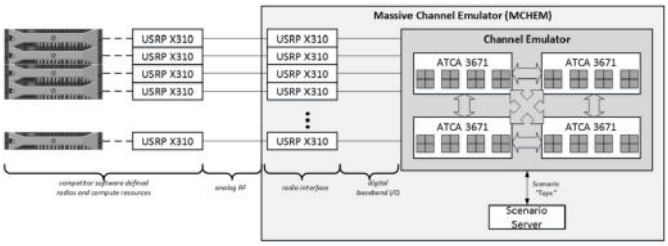


Fig. 10. *Colosseum architecture, showing the user (or competitor) X310 USRPs and the additional USRPs in the massive channel emulator (MCHEM).*

Accelerating Wave Propagation: Colosseum serves as both a development and test environment for researchers, originally conceived to test AI-enabled radio design ability to autonomously manage the spectrum with the following features: (i) Large-scale: Colosseum can connect 128 SDRs (each with two daughterboards/antennas giving a total of 256 transmit chains) in a realistic RF environment mimicking real world multipath. Furthermore, each user/competitor SDR host can access a dedicated NVIDIA K40m GPU. Each ATCA-3671 in the channel emulator features four user-programmable Virtex-7 690T FPGAs, giving a total of 64 such FPGAs, which can be utilized for realizing massive MIMO. (ii) Full-mesh: Colosseum is constructed as a "full-mesh," such that every radio is able to hear every other radio through a unique RF channel, as well as being connected to a shared 10G Ethernet plane via 18 switches. (iii) Wideband: Colosseum can emulate the wireless interactions across a bandwidth of 80MHz each with 2 UBX daughterboards. (iv) Neighborhood-sized: Colosseum is able to model an area of a neighborhood–approximately 1 sq. km.

Thus, Colosseum is naturally suited for repeatable, reliable testing of AI-enabled massive-MIMO technology and other distributed antenna systems suitable for future 6G deployments.

The 128 SDRs can be synchronized for phase-coherent operation using 19 hierarchically arranged clock distribution systems.

Next-Generation 6G Wireless Testbeds:

6G Testbeds for Fast Computing: While Colosseum is a great shared resource for testing 6G deployments, it does not replace the need for low-cost wireless test-beds that can be deployed by individual research groups. There have been many efforts from industry and academia towards creating affordable SDRs that enable 6G experimentation for the research community. Of these, a notable example from industry has been launched by Pi-Radio, a start-up funded by the US National Science Foundation (NSF) SBIR/STTR programs. The Pi-Radio SDR features a fully-digital 60 GHz transceiver (Fig. 11) that allows the radio to beamform in several directions simultaneously. This system operates in the unlicensed V-band (57-64 GHz), and has 4 independent channels. The Xilinx RF SoC-based ZCU111 board is the chosen baseband subsystem for this SDR, thus allowing ample compute capacity for wireless algorithms as well as for running ML and AI code using low-level languages (such as C). Apart from the FPGA and ARM cores, this powerful RF SoC also features soft-decision forward error correction (SD-FEC) blocks in hard silicon. The system operates over 2 GHz of real-time bandwidth. Critically, the ZCU111 can also be interfaced with reprogrammable AI/ML accelerators (such as the Xilinx Versal ACAP system) through fiber-optic links with throughputs on the order of hundreds of gigabits per second and sub-microsecond latency. We believe that the Pi Radio SDR offers entry-level experimentation capabilities to emerging AI- and ML-enabled 6G systems operating in the 60 GHz band.

Fig. 11. *Pi Radio 60 GHz array SDR for 6G experimentation, consisting of 4 transmit antennas, 4 receive antennas, and Xilinx ZCU-111 RF SoC compute platform for SDR and ML applications.*

References:

[1] Y. Cheng, J. Geng, Y. Wang, J. Li, D. Li, and J. Wu, "Bridging machine learning and computer network research: a survey," CCF Transactions on Networking, vol. 1, 11 2018.

[2] J. Xie, F. R. Yu, T. Huang, R. Xie, J. Liu, C. Wang, and Y. Liu, "A survey of machine learning techniques applied to software defined networking (sdn): Research issues and challenges," IEEE Communications Surveys & Tutorials, vol. 21, no. 1, pp. 393–430, 2018.

[3] I. F. Akyildiz, J. M. Jornet and C. Han, "TeraNets: Ultra-broadband Communication Networks in the Terahertz Band," IEEE Wireless Communications Magazine, vol. 21, no. 4, pp. 130-135, August 2014.

[4] I. F. Akyildiz, J. M. Jornet and S. Nie, "A New CubeSat Design with Reconfigurable Multi-Band Radios for Satellite Communication in Dynamic Spectrum Frequencies," Ad Hoc Networks (Elsevier) Journal, vol. 86, pp. 166-178, April 2019.

[5] Y. Wang, X. Tang, G. J. Mendis, J. Wei-Kocsis, A. Madanayake, and S. Mandal. "AI-Driven Self-Optimizing Receivers for Cognitive Radio Networks." In 2019 IEEE Cognitive Communications for Aerospace Applications Workshop (CCAAW), pp. 1-5. IEEE, 2019.

[6] Y. Wang, G. J. Mendis, J. Wei-Kocsis, A. Madanayake, and S. Mandal. "A 1.0-8.3 GHz Cochlea-Based Real-Time Spectrum Analyzer With Δ-Σ-Modulated Digital Outputs." IEEE Transactions on Circuits and Systems I: Regular Papers (2020).

[7] S. Riyaz, K. Sankhe, S. Ioannidis, and K. R. Chowdhury, "Deep Learning Convolutional Neural Networks for Radio Identification," IEEE Communications Magazine, vol. 56, no. 9, September, 2018.

[8] T. Jian, B. C. Rendon, E. Ojuba, N. Soltani, Z. Wang, K. Sankhe, A. Gritsenko, J. Dy, K. R. Chowdhury, and S. Ioannidis, "Deep Learning for RF Fingerprinting: A Massive Experimental Study," IEEE Internet of Things Magazine, vol. 3, no. 1, pp. 50-57, April 2020.

[9] K. Sankhe, M. Belgiovine, F. Zhou, L. Angioloni, F. Restuccia, S. D'Oro, T. Melodia, S. Ioannidis, and K. R. Chowdhury, "No Radio Left Behind: Radio Fingerprinting Through Deep Learning of Physical-Layer Hardware Impairments," IEEE Trans. on Cognitive Communications and Networking, Special Issue: Evolution of Cognitive Radio to AI-enabled Radio and Networks, vol. 6, no. 1, Oct. 2019.

[10] K. Sankhe, M. Belgiovine, F. Zhou, L. Angioloni, F. Restuccia, S. D'Oro, T. Melodia, S. Ioannidis, and K. R. Chowdhury, "No Radio Left Behind: Radio Fingerprinting Through Deep Learning of Physical-Layer Hardware Impairments," IEEE Trans. on Cognitive Communications and Networking, Special Issue: Evolution of Cognitive Radio to AI-enabled Radio and Networks, vol. 6, no. 1, Oct. 2019.

[11] G. Reus-Muns, D. Jaisinghani, K. Sankhe and K. R. Chowdhury, "Trust in 5G Open RANs through Machine Learning: RF Fingerprinting on the POWDER PAWR Platform," IEEE Globecom, 7-11 December 2020, Taipei, Taiwan.

[12] N. Soltani, G. Reus-Muns, B. Salehi, J. Dy, S. Ioannidis, and K. R. Chowdhury, "RF Fingerprinting Unmanned Aerial Vehicles with Non-standard Transmitter Waveforms," IEEE Transactions on Vehicular Technology, accepted, Nov. 2020.

[13] GlobeNewswire, "Commercial Drone Market." Website: https://www.globenewswire.com/news-release/2018/02/28/1401040/0/en/Commercial-Drone-Market-to-hit-17bn-by-2024-Global-Market-Insights-Inc.html, 2020.

[14] G. J. Mendis, J. Wei-Kocsis and A. Madanayake, "Deep Learning Based Radio-Signal Identification With Hardware Design," in IEEE Transactions on Aerospace and Electronic Systems, vol. 55, no. 5, pp. 2516-2531, Oct. 2019, doi: 10.1109/TAES.2019.2891155.

[15] G. J. Mendis, Jin Wei and A. Madanayake, "Deep learning cognitive radar for micro UAS detection and classification," 2017 Cognitive Communications for Aerospace Applications Workshop (CCAA), Cleveland, OH, 2017, pp. 1-5, doi: 10.1109/CCAAW.2017.8001610.

Concluding Remarks

Whilst we should not let the limitations noted previously stand in the way of what many consider a pending AI revolution, both understanding and communicating areas in which restrictions exist can ensure realistic timelines and expectations for AI to be integrated into everyday life. It should be noted that, prior to the injection of data, high-performance data processing pipelines must be in place, as the injection of high-quality data into a less than effective engine would only slow the move to market in tech organisations. As discussed, the main limitations in regard to data, are availability, cost, privacy, ethics, storage and quality. Knowing where to source data is not an easy task, this is, in fact, the most common challenge to overcome in general. Often organizations use data mining companies for this which creates the hurdle of cost but also then creates the possible avenue of artificial data.

Privacy and the ethical implications should also be a key consideration for every organization. Good quality data can sometimes come at a cost, however, the privacy of individuals is imperative. Whilst there are currently no ethical guidelines blanketing the whole of AI, it must be in the forefront of company values to acknowledge such issues as gender bias in data, ethnicity bias and public manipulation. Data compatibility is also another common challenge, with variations in format, platform and application. For example, in legacy data, which requires a strict format and environment.

It must also be recognized that with development in capabilities, comes the need for a variation in data collection. With the advancement of facial recognition, for example, there needs to be a huge degree of data, accuracy and data labelling, which would need to be done at an incredibly fast rate.

Through this, however, we will see developments of libraries to deal with preprocessing of text and image, amongst others, in each individual field. Over time, it then becomes possible that the preprocessing stage for specific data types becomes automated by standard industry practices. Following on from this, it is hoped that these libraries will be replaced with automated data format conversion, making data highly available and accessible to all, regardless of budget or similar restrictions.

Therefore, data limitations play a critical role on how the whole field of Data Science evolves and it is by overcoming these challenges that important advances are made.

> ***We are in the midst of technological convergences that will disrupt industries and provide new industry opportunities."***
>
> -Mark Wright, GSI Technology

Additional Reading

Thank you to the whitepaper sponsor, GSI Technology

Founded in 1995, GSI Technology, Inc. is a leading provider of SRAM semiconductor memory solutions. The Company recently launched radiation-hardened memory products for extreme environments and the Gemini® APU, a memory-centric associative processing unit designed to deliver performance advantages for diverse AI applications. The Gemini APU's architecture features parallel data processing with two million-bit processors per chip. The massive in-memory processing reduces computation time from minutes to milliseconds, even nanoseconds. Gemini excels at large (billion item) database search applications, like facial recognition, drug discovery, Elasticsearch, and object detection. Gemini's scalable format, small footprint and low power consumption, make it an ideal solution for edge applications where rapid, accurate responses are critical.

Further reading:

Scalable Semantic Vector Search with Elasticsearch