Information & Security

P. Lenk, M. Street, et al.
vol. 47, no. 2 (2020): 157-171
https://doi.org/10.11610/isij.4711

Published since 1998

ISSN 0861-5160 (print), ISSN 1314-2119 (online)

Research Article

Data Science as a Service: The Data Range

Peter Lenk [a] (✉), *Michael Street,* [a] *Ivana Ilic Mestric,* [a] *Arvid Kok,* [a] *Giavid Valiyev,* [a] *Philippe Le Cerf,* [b] *Barbara Lorincz* [b]

[a] *Service Strategy & Innovation NATO C & I Agency, The Hague, Netherlands*

[b] *Gartner Belgium Bvba, Zaventem, Belgium*

ABSTRACT:

As with many new disciplines, in many organisations data science is being embraced in a piecemeal way with many parts of organisations creating special purpose environments designed to answer specific problems, fragmenting the overall capacity and knowledge base. Often vendors selling proprietary approaches, potentially creating lock-in, fuel these isolated solutions. This article's main contribution is a 'Data Science as a Service (DSaaS)' model, where common elements required to conduct data science are abstracted and gathered into a logical layered, service-based architecture. This way, each element of the organisation can utilise the services it needs to progress its work, use specific solutions or share common tool sets, share results in a 'model zoo,' share data sets, share best practices and benefit from common, robust high-performance infrastructure and tools. With such an approach, it is possible to cluster data science skill sets and provide critical mass where needed. The proposed approach also facilitates a charge-back business model, where data science services are costed and charged to internal organisational elements or external customers in a measured, pay-as-you go way.

ARTICLE INFO:

RECEIVED: 14 JUNE 2020
REVISED: 14 SEP 2020
ONLINE: 22 SEP 2020

KEYWORDS:

data science, artificial intelligence, machine learning, big data, data engineer, data engineering, data range, data science as a service, charge back

✉ E-mail: peter.lenk@ncia.nato.int

Context / Motivation

Data Science, along with its constituent technologies of Artificial Intelligence (AI), Machine Learning (ML), Big Data Analytics (BD), etc., has found utility across many industries and areas of government. Five of the world's ten most valuable companies – Amazon, Alphabet, Alibaba, Facebook, Tencent – all owe their value in large measure to data and exploitation of that data.[1] In government, we are seeing that better collection, management and exploitation of data is also bringing benefit.

Defence generates huge volumes of data, ranging from reconnaissance data, communications data, cyber defence data, human resources data, biometric data, procurement data, Command and Control (C2) data, open source data, social media, policies, directives, and so on. How the military capitalises on the value of this data is likely to affect the future social, political and economic landscapes of the world.

Just as data can be used to improve things, it can also be used in negative ways. A recent article in the BBC reported a Facebook executive as stating that Facebook helped President Trump win the 2016 election "because he ran the single best digital ad campaign I've ever seen from any advertiser. Period." [2] We are also aware of the serious allegations of meddling in the 2016 Brexit referendum in the UK through the potentially illegal use of social media data to identify and target voters that might be swayed to favour leaving the European Union.[3] These tools are powerful and can be put to sinister purposes if appropriate policies and governance are not in place. In 2017 Vladimir Putin stated:[4] "the one who becomes the leader in this sphere will be the ruler of the world." The Covid-19 pandemic has shown that failure to scrutinise data sources and the results of data analysis can put lives in danger on a huge scale.[5]

Several nations, inside NATO and beyond, are actively adopting AI within defence while recognising much work remains to be done:[6]

> As numerous studies have made clear, the Department of Defense (DoD) must integrate artificial intelligence and machine learning more effectively across operations to maintain advantages over increasingly capable adversaries and competitors. Although we have taken tentative steps to explore the potential of artificial intelligence, big data, and deep learning, I remain convinced that we need to do much more, and move much faster, across DoD to take advantage of recent and future advances in these critical areas.

In NATO as well, we are seeing an increase in the awareness of what these technologies might bring, as well as new interest in developing tools to support situational awareness and decision making at all levels. NATO has recently established a 'Data Board' to develop policy in relation to data and data science. A number of elements of NATO have begun to experiment and develop 'Minimum Viable Products'; that is, limited scope prototypes that address real challenges that might be solved with these technologies. These early adopters have raised awareness of what these technologies can bring and have learned how to prepare and work with their data assets.

The issue we now face is one of fragmentation; fragmentation in the sense that many of these communities are going in their own directions and there is little sharing of data, resources or approaches. In these early adoption stages of data science technologies, it is unlikely that strong centralised governance is the right approach. The level of understanding of the issues needs to reach a suitable level of maturity to guarantee that policies support, rather than unduly constrain, the pursuit of data science; a large degree of flexibility is considered necessary in order to keep an entrepreneurial approach alive. While policy is needed regarding data ownership and ethical use issues, we should not overly constrain how things are done. We need approaches that bring communities together voluntarily, by offering resources that can be shared (which may require some form of standardisation) and by creating ecosystems that extend addressable resource pools.

There appears to be an efficiency that can be created by abstracting common resources and delivering these to the community, as a service, thus preventing unnecessary duplication of effort.

We will need to have a big data analytics / AI / ML capability of sufficient scale to transform the overwhelming quantities of data into actionable information in meaningful timeframes. Due to the vast quantities of data, there will be no possibility to employ sufficient human resources to accomplish this so we will become dependent on technology to assist in this effort. Initially this will be to automate the routine and mundane tasks, operating at scale across large and varied data sets, leaving more difficult analyses to scarce and highly qualified / experienced human operators.[7] Over time, as trust deepens and data loads become ever larger, the AI / ML agents and algorithmic approaches will take on more and more tasks. Machine learning will mean that these algorithmic approaches will improve over time, enabling more and more difficult tasks to be delegated to machines, freeing up the humans to focus more and take on new roles.

One important aspect is that the defence community will need a capacity to conduct data science on sensitive data at a classified level. While we can make use of commercial public cloud infrastructure and tools to develop unclassified solutions, NATO is often constrained to run in a disconnected mode because of the nature of the data and questions to be answered.

Whereas the NIST Big Data Reference Architecture (NBDRA) [8] (illustrated in Figure 1) provides a good starting point for this purpose as it identifies many of the underlying functions that are needed, our model frames the problem in a service context, exposing the functions needed as a set of services and adds the people element needed to conduct data science.

In this paper we propose a Data Science as a Service (DSaaS) model, a layered model of resources needed to conduct data science and provide it as a service. We propose that the resources are exposed as a set of interlinked services that can be consumed by practitioners within and without the organisation. This should facilitate a reuse / shared use of resources, provisioning of higher capability resources than might be afforded by any single segment of the organisa-

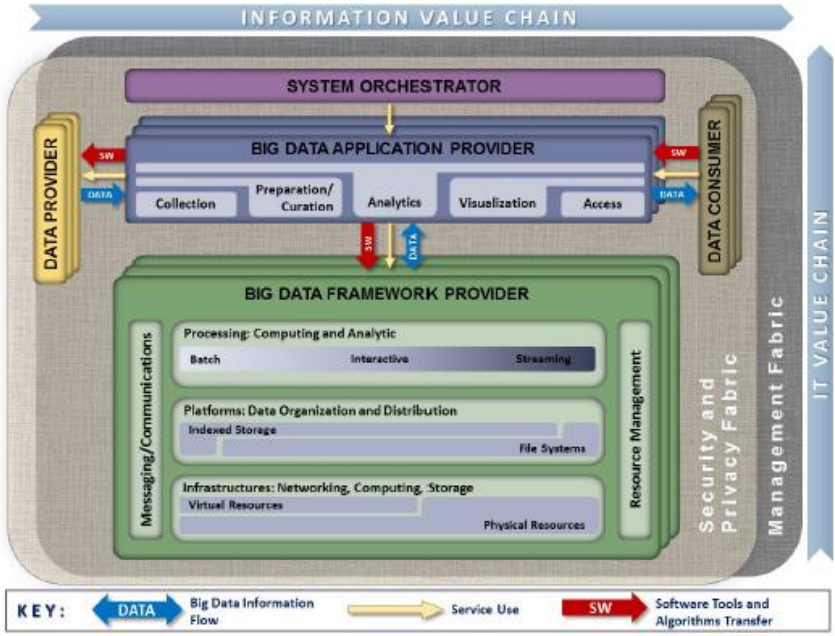

Figure 1: NIST Big Data Reference Architecture (NBDRA).

tion, and a sharing of knowledge and results. The model draws loosely on the NIST Cloud Computing Reference Architecture [9] and the NIST Big Data Interoperability Framework.[10]

The paper is organised into four sections, this introduction being the first. The second section provides more explanation of the problem that we are addressing, the third delves into the proposed solution and the final section presents conclusions and topics for further work.

Scoping the Problem – Use Cases

The NATO Communication and Information Agency (NCIA) is NATO's Information and Communication Technology (ICT) service provider. As such, we are developing an environment to support the Alliance's data science needs. While data that is not sensitive can be processed in the public cloud, sensitive data needs to be carefully handled and is subject to many regulations that assure its confidentiality, integrity and availability. For this reason, it is not currently possible to use the public cloud for certain types of data sets. As a result, NATO will need its own environment where such work can be undertaken and can support a wide range of users and use cases. In allusion to a rifle range, where soldiers learn, practice and perfect the art of using firearms, the data science environment has been referred to as a *data range*, where staff can learn, practice and perfect data science.

Such a data range is expensive, requiring costly technology, scarce skills and careful data curation in order to provide an effective capability. This has been evident in the embryonic data science environment developed to date at the NCIA. As a fully 'customer funded' agency, NCIA can only provide capabilities and services that users want and are willing to pay for. This provides transparency and accountability to the nations of the Alliance. Therefore, the Data Range needs to be architected in such a way that easily facilitates a charge back mechanism, so that users can pay for data science services on a consumption basis. This has led us to pursue a concept where Data Science resources and capabilities are exposed as a set of services; i.e., Data Science as a Service (DSaaS).

Use Cases

We have identified a number of use cases that the Data Range has to be able to support. A non-exhaustive set of use cases are illustrated in Figures 2 through 6.

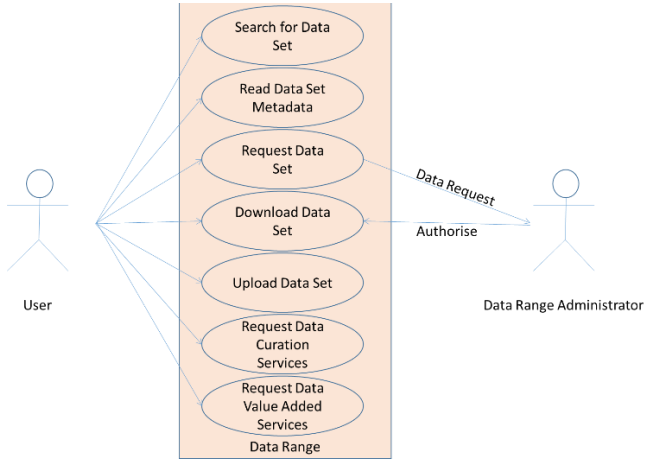

Figure 2: Data Use Cases.

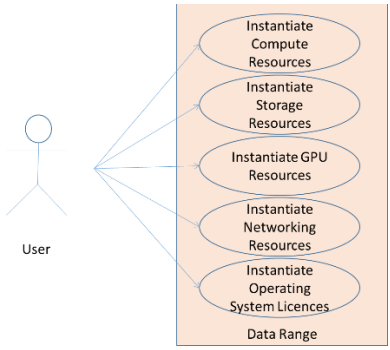

Figure 3: Data Science Infrastructure Use Cases.

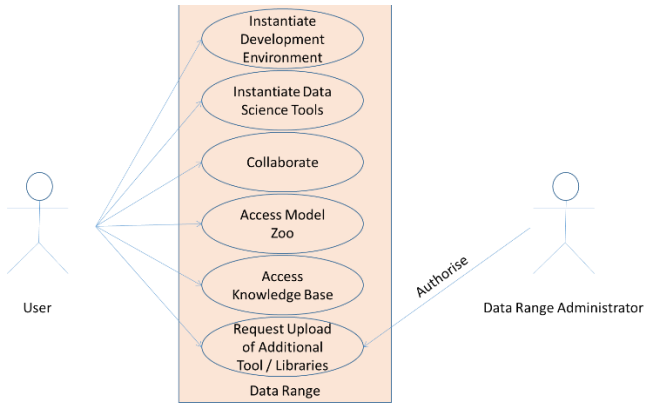

Figure 4: Data Science Platform Use Cases.

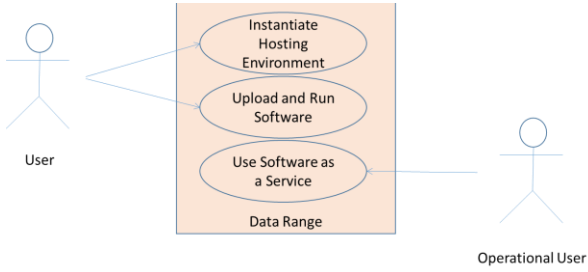

Figure 5: Data Science Software Use Cases.

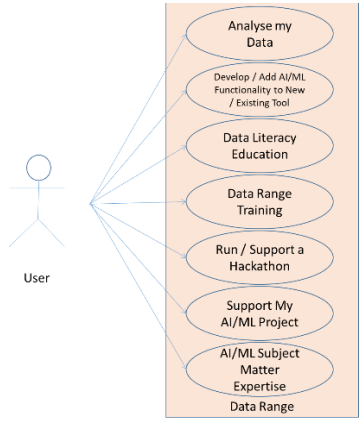

Figure 6: Data Science as a Service Use Cases.

Data Science as a Service Model

In order to implement this vision of a family of services, together delivering DSaaS, we propose a layered model, consisting of five service layers and two supporting pillars, as illustrated in Figure 7. The rest of this section defines the services envisioned in each layer and pillar. Not all services are defined in equal detail at this time, the development of a fully documented service catalogue is still work in progress.

Critically, our proposed model allows for access by the users and eco-system to services at each layer in the stack. This provides flexibility allowing users to avail themselves of as much or as little of the Data Range as needed, depending on where they feel it adds value to their specific problem. Another design driver was that the Data Range be extensible so that data, services and even specialised infrastructure, where these offer benefit, can be introduced into the environment by trusted and authorised parties and can be consumed by authorised users.

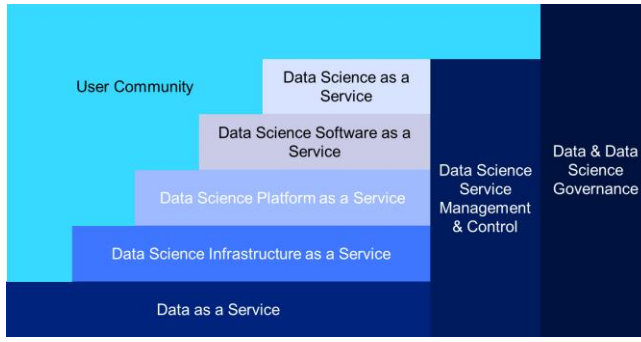

Figure 7: A layered DSaaS model.

Data and Data Science Governance

The set of services in this pillar deliver governance over both data holdings and data science services. Each organisation needs to consider, what standards will be used to store, retrieve and access data; what is permissible to be done with the data; who owns data; etc. This pillar also needs to provide direction regarding prioritisation of work to be conducted, given the reality of finite resources.

- Data governance – authority that sets and enforces the rules such as
  - Data ownership policy and directives
  - Data Access Management
  - Data usage policies and directives
  - Data accountability
  - Technical standards
- Data Science Service Governance

This layer should be in part implemented by an external [to the Data Range] Board, or other governing body to determine organisational issues such as appropriate access and usage and ownership of data, and partially internally to direct the tactical direction of the Data Range, ensuring that policy is being applied and work is being prioritised appropriately.

Data Science Services Management and Control

Services are required that in turn allow the management of the data and the data range services and permit appropriate controls to be executed. These service management and control services should follow the guidance of the latest version Information Technical Infrastructure Library (ITIL) including:

- User Provisioning Services
- Service Provisioning Services
- Service Orchestration Services
- Service Management Services
- Service level monitoring / KPI measurement services
- Service consumption monitoring and charge back services
- Data Science Service Catalogue services defining what services the Data Range offers, what they cost, how they are requested, what service levels are provided, etc.

This layer will require staff with specialised but widely available IT Service Management and Control and knowledgeable of ITIL processes.

Data as a Service (DaaS)

This layer provides the service interfaces that provide a trusted way to add or access data in the Data as a Service (DaaS) layer. Data collection services; that is, mechanisms for adding data to the lake, or even services for collecting data from an exercise, operation, or other event, will be needed. Curatorial services will enhance the data by providing indexing, quality indicators, provenance, metadata, etc.

We anticipate that one of the biggest values that the Data Range will bring is the creation of a library of curated, labelled and unique NATO and coalition operational data, as well as a place others can bring and share data. The Data Range will need to deal with static, non-real-time data, as well as either live or recorded streaming types of data.

Policies defined by the governance layer that facilitate the incorporation of data into the Data Range, rather than inhibiting or prohibiting it, will be required. Standard data formats, where they exist and are useful will be followed; however, the data should generally be introduced in a raw format to ensure that information is not lost.

It is not envisioned that there will be a single centralised repository of data inside the Data Range, rather the concept is to provide a single entry point where users can locate data and access the appropriate meta-data describing

the characteristics of the data, such as: provenance, security classification, access constraints, data usage restrictions, points of contact, etc. Data will of necessity, especially in scenarios where bandwidth is challenged, be distributed; however, it should still be considered as part of the Data Range data holdings. Commercial or open source data should be brokered through this layer to ensure consistency, avoid duplication and provide economies of scale to data acquisition. These data services should not be equated to an archive; one group of consumers may be archivists, but they normally have different needs than data scientists and may require a more processed and structured set of data.

Access management mechanisms will be of high importance to ensure that trusted users are authorised to access those specific items of data they are allowed to access, and are prohibited from accessing other data sets and services. Services will also be required to authorise upload of data sets, verify that such data sets are in line with the provided metadata and also to assess the veracity of data and protect against data poisoning. These will rely on Data Science Infrastructure as a Service (DSIaaS) services that can store and retrieve data at rates that are in line with the needs of the source or of the subsequent layers.

We have clustered the services in this layer into four groups: Data Ingestion & Storage Services, Data Curation Services, Data Access Management and Data Security Services, and Data Value Added Services.

Data Ingestion & Storage Services

- Data ingestion services
- Data storage services
- Data archival services
- Data sharing services – sensitive data 'drop box'

Data Curation Services

- Data library services providing a single-entry point for all data
    - Data cataloguing services
    - Data engineering services
    - Data labelling services
    - Data quality and veracity evaluation services
    - Data configuration management services
    - Data provenance services
    - Data cleaning services
    - Metadata tagging services
    - Releasable data set creation services
- Data brokering services for open source and commercially sourced data

Data Access Management & Data Security Services

- Managed data access services
    - Ensuring persons and institutions are authorised to access data;

  - Ensuring that persons and institutions can only access data they are authorised to access; and
  - Ensuring that accessed data is only utilised for authorised purposes.
- Cyber security services providing data loss prevention, confidentiality, data integrity, etc.

Value Added Services

- Data collection planning services
- Data collection execution services
- Data conversion services
- Data replay service (e.g. for video and other streamed sources)

In order to implement the services in this layer, specialised personnel skills will be needed such as Data Engineering, Data Protection and Enterprise Architecture in order to make the appropriate data accessible.

Data Science Infrastructure as a Service (DSIaaS)

Data science demands specialised hardware to support the intensive processing required to machine learning and AI train models. This hardware generally fits in the category of High-Performance Computing (HPC). The Data Science as a Service (DSIaaS) layer exposes the needed hardware as a set of infrastructure services, similar to normal cloud Infrastructure as a Service (IaaS), but augmented with additional capabilities needed to support data science.

In particular, data science demands high performance processors, with multiple cores, large memory resources and perhaps most importantly access to specialised Graphical Processing Units (GPU); that is, specialised computer processors, originally designed for performing demanding graphical calculations. GPUs are well adapted to the data science problem as they provide a highly parallelised architecture, providing hundreds or even thousands of individual computer cores, and thus provide training results in reasonable timeframes.

This layer also needs to provide connectivity to end user locations, as well as connectivity to federated data sources forming a part of the larger virtualised data library.

The set of services at this layer will virtualise all these physical resources and provide them to the user in measurable quantities. It is intended that this be done in accordance with the cloud principles of resource pooling, on-demand, self-service, elasticity, with high reliability and scalability.

- Tiered storage services – online, near-line, off-line
- Computing services – Central processing unit cores and memory
- GPU services – GPU cores required
- Networking services
- Operating system licensing services

This layer will depend on relatively usual personnel skills as required for Infrastructure as a Service provision.

Data Science Platform as a Service (DSPaaS)

Data Science Platform as a Service (DSPaaS) provides a library of software tools, exposed as services, allowing for all phases of the exploitation of the data. These tools will include common open-source tool sets, NATO-owned tool sets or commercial tool sets. It will also allow for eco-system partners to bring in custom tools for hosting in the environment where these bring specific capability or benefit. Included in the library will be data preparation e.g. Extract, Transform, Load (ELT) tools, normalisation tools, machine learning tools, big data tools, artificial intelligence tools, s, etc.

In order to support the eco-system, the DSPaaS will also include 'value add' tools such as collaboration spaces, chat services, knowledge bases, etc.

Importantly, the DSPaaS will include a library of components and pre-trained models, etc., that can be used by other analysts. This 'model zoo' will facilitate sharing of existing components and models, bringing cost efficiency while greater exposure and scrutiny should increase confidence in these models.

Generally, environments would be bundled and provided to users as a pre-configured package for their use.

- Development environments for generating software code
    - Python
    - R
    - Etc.
- Data science tools services
    - Extract, Transform, Load (ETL) tools
    - Open source libraries, such as Tensor Flow, PyTorch, Keras, etc.
    - Open source data science tools e.g. KNIME, etc.
    - Commercial data science products
    - Visualisation tools
    - Add additional tool set / libraries
    - Etc.
- Collaboration services
    - Presence
    - Chat
    - Voice
- Model Zoo Services
    - Algorithms as a service
    - Search for Models
    - Extract Models
- Knowledge Base Services

- Search
- Access
- Edit

In order to provide these tools, some specialised knowledge will be needed to configure environments and make available all the needed libraries and tools to allow data scientists to do their work. These environments will need to be patched and constantly upgraded, as the domain is evolving very quickly.

Data Science Software as a Service (DSSaaS)

The Data Science Software as a Service (DSSaaS) layer will provide a protected production environment where data science applications can be hosted in a way that they can be exposed to users to test implementations and concepts but without endangering operational systems. While this is not foreseen as a full production environment, it will need to be sufficient to host Minimum Viable Product (MVP) implementations of decision aids and other data science applications for extensive periods of time. This will allow the tools to be trialled in exercises and even in operations before going into full production. This layer can also support function developed in house to provide custom *self-service analytics*.

While the environment will be hosted in the DSIaaS layer, and makes use of the DaaS layer, it is considered as a separate layer as it will bring different demands in terms of connectivity, availability, and support.

The services in this layer will depend largely on the DSIaaS layer and on the skills of the personnel there to keep the infrastructure running and providing the needed services. However, for each application that is hosted in this DSSaaS layer, expertise will be required in order to deal with any issues, problems or changes that arise. The level of expertise required will depend on the service level required.

Data Science as a Service (DSaaS)

The Data Science Services will provide a variety of value added services that support exploitation of data. The DSaaS layer differs from the other layers as it is primarily a set of services that rely on specialised data science and domain subject matter experts (SME) that make use of the other layers within the Data Range. They will use the tools provided in the DSPaaS layer, which will in turn access data in the DaaS layer and process that data in the DSIaaS layer.

Specialised data scientists will provide many of the services such as data preparation services, big data analytics services, ML/AI services etc., as requested by outside users. Besides data scientists, important to this layer will be access to individuals with domain knowledge that can bring understanding of the technology and data used in the domain as well as the problem space. The services can be ‘standard’ conducting of repetitive analysis of data, as it arrives, providing perhaps situational awareness or entity extraction, or it can be one off, providing studies or answers to specific questions.

We have clustered the services in this layer into two groups, that cover the services currently envisaged, but as this layer relies on SMEs, additional services are likely to be requested and provided.

Education Services

- Data Range training services for data scientists – specific training to help data scientists understand the services provided by the Data Range and how to request and make use of these
    - For Data range staff
    - For 'citizen data scientists' embedded in the business
    - Eco-system partners
- Data literacy training services
    - Data and data science awareness training services for senior staff and decision makers
    - Data and data science awareness for account managers, to properly represent what services are offered and what can be accomplished in the data range

Data Science Consulting Services

- Data Science SME Services
    - Provide AI/ML Subject Matter Expertise
    - Provide AI/ML and Domain Subject Matter Expertise.
    - Develop / Support development of project / procurement documents for AI/ML projects
    - Support AI/ML project implementation
    - Hackathon Services
- Data Analysis Services
    - Analysis of specific Data Science questions
- Tool building / enhancing services
    - Development / enhancement of custom AI/ML tools that can then be used by an end-user to enhance their performance
    - Development / enhancement of self-service analytics functions

The services in this layer will be provided by a core team of data scientists, supplemented by partners from the eco-system as needed. It can also be that much of the work in this layer is conducted by SMEs embedded in the business, only reaching out occasionally to the core team.

Eco-Systems

Key to the success of the Data Range will be the ability to tap into an eco-system that supports and exploits it. This will add to not only the breadth of the technology base, but also to the solution base. We will not only seek technology

from this eco-system that can be integrated into the Data Range, but more and more for complete elements of the overall end-to-end service, where a supplier provides specialised functionality and algorithms as a service. Eco-systems also need to be seen as a part of the overall workforce strategy, enriching the organic force with greater diversity of skills and backgrounds that can be called in when needed.

To enable eco-systems to function, clear and open Application Programming Interfaces (APIs) that can be used by the eco-system members to both inject and pull information [within their permissions] to the Alliance's advantage. These will also enable the ingestion of data from relevant sources of information.

We recognise that the coalition partners, nations and other community members are integral parts of the Enterprise and are eco-system partners, providing and consuming information from the NATO Data Range. This is again where standards, and APIs, such as those brought through efforts like Federated Mission Networking (FMN) [11] or the NATO Core Data Framework [12] will be essential.

In our model the eco-system includes the end-users as they have an essential role in the Data Range. They are the ones with the questions to answer, insights to be found, processes to support and decisions to make. For mature processes, over time, we may provide automated tooling to deliver answers; in less mature areas, or one offs, the data science functions become more iterative, requiring a close interaction throughout the process to ensure solutions meets user expectations.

Conclusions and Future Work

The paper has presented a proposed architecture and set of services to implement a Data Range, providing Data Science capabilities as a set of services which can operate on classified data and use cases. This model provides a scalable layered approach that can be evolved over time to ensure that it stays relevant.

A detailed service catalogue, following this architecture and principles is in development; as is a major uplift in the computing infrastructure to support the Data Range. From there, very quickly, we will build up both the platform and data service layers. The data layer will initially by implemented as a centralised instance, using data sets that are already available at NCIA, with federation of other data sources taking some time to accomplish. Nevertheless, we would wish to include as many data sets in the library index system as possible, early on, to ensure that people can discover them and find appropriate ways of accessing them. At the platform level, focussing initially on open source tool sets will allow a useful and powerful capability to be created rapidly. At later stages, business models for hosting commercial tools will be developed and implemented as needed. In this way we hope to rapidly achieve an initial Data Range capability, and grow and expand it over time.

Additional work is ongoing to fully develop the Data Range service catalogue, including pricing of the services, and to develop or acquire tools that can support the proposed services.

References

1 Tim Harford, "The Man Who Got Rich on Data – Years before Google," *BBC Article*, 8 January 2020, https://www.bbc.com/news/business-50578234.

2 "Facebook Ad Campaign Helped Donald Trump Win Election, Claims Executive," *BBC*, 8 January 2020, https://www.bbc.com/news/technology-51034641.

3 Constitution Unit, "Independent Commission on Referendums," Report of the Independent Commission on Referendums, London, July 2018, pp. 178-179, https://www.ucl.ac.uk/constitution-unit/sites/constitution-unit/files/ICR_Final_Report.pdf

4 "Vladimir Putin Warns Whoever Cracks Artificial Intelligence Will 'Rule the World'," Vladimir Putin quoted in *The Mail Online*, 1 September 2017, https://hotcopper.com.au/threads/vladimir-putin-warns-whoever-cracks-artificial-intelligence-will-rule-the-world.3659445/.

5 "Surgisphere: Governments and WHO Changed Covid-19 Policy Based on Suspect Data from Tiny US Company," *Guardian*, 3 June 2020, https://www.theguardian.com/world/2020/jun/03/covid-19-surgisphere-who-world-health-organization-hydroxychloroquine.

6 Robert Work, Deputy Secretary of Defense Memorandum, "Establishment of an Algorithmic Warfare Cross-Functional Team (Project Maven)" *govexec*, 26 April 2017, https://www.govexec.com/media/gbc/docs/pdfs_edit/establishment_of_the_awcft_project_maven.pdf.

7 Ivana Ilic Mestric, Arvid Kok, Giavid Valiyev, Michael Street, Peter Lenk, Mihaela Racovita, and Filipe Vieira, "Extracting Value from NATO Data Sets through Machine Learning and Advanced Data Analytics," *IST-178 Specialists meeting on Big data challenges: situational awareness and decision support*, Budapest, October 2019.

8 US Department of Commerce, National Institute of Standards and Technology, "NIST Big Data Interoperability Framework," NIST Special Publication 1500-6r2, Version 3, October 2019.

9 US Department of Commerce, National Institute of Standards and Technology, "NIST Cloud Computing Reference Architecture," Version 1, Special Publication 500-292, September 2011.

10 US Department of Commerce, National Institute of Standards and Technology, "NIST Big Data Interoperability Framework," NIST Special Publication Series 1500, Version 3, October 2019.

11 "Federated Mission Networking," *Wikipedia*, 2020, https://en.wikipedia.org/wiki/Federated_Mission_Networking.

12 M. Nguyen, "NATO Core Data Framework (NCDF)," *Emerging Technologies Enabling Cross-COI Information Sharing,* 2015.