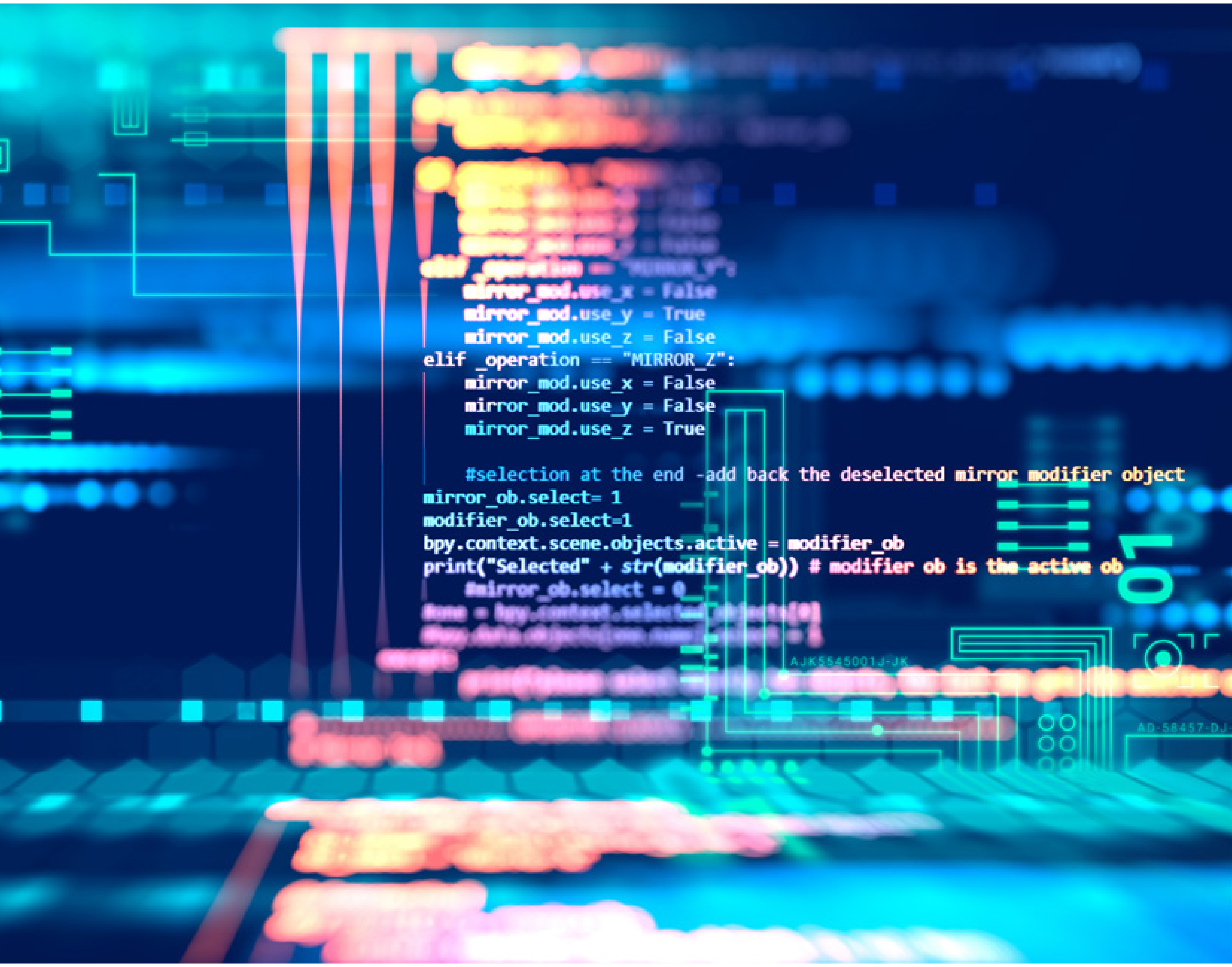
element

Digital Engineering



```
elif _operation == "MIRROR_Y":
    mirror_mod.use_x = False
    mirror_mod.use_y = True
    mirror_mod.use_z = False
elif _operation == "MIRROR_Z":
    mirror_mod.use_x = False
    mirror_mod.use_y = False
    mirror_mod.use_z = True

    #selection at the end -add back the deselected mirror modifier object
mirror_ob.select= 1
modifier_ob.select=1
bpy.context.scene.objects.active = modifier_ob
print("Selected" + str(modifier_ob)) # modifier ob is the active ob
    #mirror_ob.select = 0
```

DATA SCIENCE IN INDUSTRY AND CONTEMPORARY PRACTICES

WHITEPAPER

CONTENTS

element

SUMMARY

Contemporary digital technologies are underpinned by practices such as analytics, which are underpinned in-turn by Data Science.

Data science is a relatively new field of study. It is an interdisciplinary field that uses scientific methods, algorithms, and statistical techniques to extract insightful information, trends, and patterns from large quantities of structured and unstructured data. With the increase in computing power and development of machine learning algorithms, larger sets of data are able to be processed that previously would have been an arduous, if not impossible, task.

Large engineering organisations generate, store and access large amounts of data in various forms such as sensor measurements, site surveys, component drawings, text files, databases, written documents etc. In itself this data has value, but through effective data management can be used to make better business decisions. Data management is the gateway to making sense of the aggregated data, to perform data analytics, data mining, modelling etc, and the insights that this can offer. However, it is the integration of this with decision making tools that provides the opportunity for the full value of the data to be realised, giving the opportunity for improved decision making.

This report focusses on investigating the data science landscape within the nuclear and general engineering sector. A number of case studies are presented which were developed from interviews with industry data science experts. These work to identify common themes regarding data science methods and challenges, as well as providing successful and unsuccessful use cases. Key findings from the interviews are outlined and discussed below.

CHALLENGES

Many of the common challenges faced in industry concerning data and data science can be broadly categorised into 3 categories: culture, security, and technology.

Culture is arguably the most important challenge to overcome when building capability in data science. Many data science methods are different to traditional methods of analysis and large organisations are generally slow to adopt new technologies. Organisations are often resistant to adopting data science because they do not understand the new methods and are therefore reluctant to adopt practices which are sometimes viewed as “black boxes”. Conversely, there may be unrealistic expectations about what can be achieved. Training, collaboration, and engagement with the relevant communities have been shown to help encourage buy-in and promote continuous improvement across industry. “Virtual working groups” have been successfully used to ensure continuous collaboration and learning of data science methods and practises, while involving end-users in the development and deployment of tools has been shown to encourage buy-in and maximise efficiency. All of these factors are especially important considering the high risk of failure associated with data science projects. An efficient approach has been identified to overcome this by using agile “Data Trials” to quickly evaluate the viability of projects.

Security regarding IT and information security can be a major barrier when sharing data. This not only affects getting access to expertise internally and externally, but also in accessing the correct tools for the job.

The major technological failures when undertaking data science in industry are often attributed to the quality and quantity of data. Failure to apply data science is never a failure of data science principles, but rather a failure of the data itself. It is important, then, to ensure large and complete data sets are continually gathered and properly stored to ensure project success. This requires patience and forward planning when installing new equipment, so results cannot be expected immediately. For example, an aerospace and defence company requires at least 6 months of performance data to capture seasonal variations.

ORGANISATION OF DATA SCIENCE ACTIVITIES

The case studies demonstrate successful use of a central data science function as a hub of expertise which can be approached by internal stakeholders who identify potential data science use cases. Data science can be applied to many areas, some often surprising, and so a central function enables resources to be appropriately placed for any potential use case. Data science has been effectively carried out by small teams which involve or work closely with end-users. This can be used to train end-users and models more effectively, and also to encourage trust in data science projects.

An aerospace a defence company employ ‘data engineers’ for more menial tasks of cleaning and storing data. This model allows for more highly skilled data scientists to use their time to focus on extracting meaningful insights from the data.

Data science teams are effective when cross-functional, which has been demonstrated across industry. Teams consisting of engineers with strong software knowledge, mathematicians, and statisticians can successfully undertake high level data science – pure data scientists may not see the wider context:

- A nuclear services provider use a range of expertise, namely statisticians, engineers, mathematicians, chemists, and physicists.
- An aerospace and defence company prefer engineers with good software skills over pure data scientists as engineers are problem oriented.
- A steel manufacturer founded an Advanced Analytics team without pure data scientists, instead successfully redeploying engineers and technical graduates supported by external consultants and experts at a sister plant.

APPLICATIONS OF DATA SCIENCE

Data science has been successfully applied in industry in condition monitoring, process predictive modelling, maintenance/ inspections, and natural language processing. This has been shown to yield operational cost savings and meet regulatory requirements. Data science has been used in aerospace and defence for root cause analysis – an application where it is not widely used in industry.

TOOLS AND WORKING METHODS

A number of tools are commonly used across industry, with most practitioners using both commercial off-the-shelf and in-house developed tools. Python is used by all of the organisations.

Data science operations work effectively using agile sprint methods, or other methods which promote highly adaptable working environments.

DEFINITIONS

The following definitions are used in this document:

Term	Explanation
DS	Data science
ML	Machine Learning
AI	Artificial intelligence
NN	Neural networks
NL	P Natural language processing

CASE STUDIES

A number of case studies are presented which were developed from interviews with industry data science experts.

CASE STUDY 1

Industry: Nuclear Services

Location: UK

Summary:

The organisation possess a wide variety of experience encompassing engineering (civil, structural, and aerospace), applied mathematics, physics, chemistry, and statistics. The bulk of modelling undertaken is physics-based modelling, i.e. CFD and FEA. Data science is a more niche area, and is undertaken primarily by statisticians and engineers.

No one actually has the job title of 'data scientist'. Much of what is referred to as 'data science' is just using applied statistical methods to gain insights on data. Machine learning is an area in which there is currently a lot of activity – it is thought of by some as a silver bullet that can solve every problem, but that is not often the case. A lot of the work undertaken by the Statistics team involves using statistical analysis to inform plant decisions and most of the machine learning falls into this category. Data sets are supplied to the team and they search for trends or engineering correlations that can be used to inform plant decisions.

Methods:

In general, software tools are a mixture of off-the-shelf commercial tools and internally developed tools. These tools are usually developed by engineers who are competent in coding rather than software developers.

Tools used for data science include: Python (Pydata, tensorflow), Fortran, C++, MATLAB, R (used by statisticians).

Team:

Data science projects are undertaken by cross-functional teams which may take expertise from any of the areas the organisation specialises in. Statisticians will almost always be involved where advanced statistical methods are required – there are approximately 10 people (mostly statisticians) who are interested in data science, ML, and AI. Project teams are usually small, consisting of 3 people working over short time periods in a consultancy role. The selection of team members is very much dependent on the nature of the work.

A 'virtual working group' is used to share and grow knowledge for those interested in data science.

Challenges:

A number of challenges were outlined which broadly fit into the following categories:

- **Culture:**

The nuclear sector is slow to adopt new technologies. Proven technologies with many verified examples are usually desired for confidence. Most organisations rely on robust reliable technologies. This doesn't present a major obstacle when using AI and ML as these both fall into a general category of statistics, but there is a reluctance to use newer technologies such as computer vision.

Some people are opposed to new technologies as they believe their roles will become redundant as new technologies are adopted - An internal survey on robotics demonstrated concerns by employees that they will be replaced by robots. When showing proof of concept it is possible to achieve 70% accuracy, but experience has shown customers wanting a 1 in 1 million error rate. This is totally unrealistic and exceeds the current system requirements.

- **Technology:**

The quantity and quality of data available often obstructs data science. This is sometimes a result of a lack of forward planning. When systems were installed, it was not expected that they would be continuously monitored and analysed using digital methods as there was no foresight of any additional value to be taken from data science.

There are challenges using NN as it can be difficult to see why the model has made a particular decision. Using regression models it is easier to visualize the results. The lack of clarity when using NN makes it more difficult to apply in a safety critical environment.

Hardware issues, such as HPC requirements for data science and lack of access to GPU.

- **Security:**

Security poses many challenges when working with data in the nuclear industry, especially regarding getting access to data from clients. Large data sets can be affected by security issues even if only a small fraction of data is sensitive. The sensitive data (which may just be an email address) often cannot be redacted, meaning data cannot be shared or must be subjected to more checks.

IT system accreditation is a challenge which requires significant time and cost to achieve.

Successful Data Science Use Cases:

1. **Materials testing**

The organisation works with highly radioactive materials. In the UK, the waste is reprocessed and what is left is radioactive sludge with complex physical properties. It is necessary to understand how it behaves, such as how physical properties are affected when boiled or evaporated as this can be important if it is pumped. These tests cannot be done with radioactive waste so a non-radioactive chemical analogue is created with similar physical properties. This is used for tests, the results of which are used to find correlations. Regression models are used to correlate the physical properties to the behaviour of the material in the tests.

2. **Natural language processing**

Natural language processing (NLP) is being undertaken which involves extracting information from legacy documents concerning safety reports (NLP has been used in a range of other applications). In the nuclear industry, safety is the top priority and as such detailed records need to be made for any safety incidents concerning plant, faulty equipment, missing PPE, injuries, etc. Records go back decades, and the structure of the data is inconsistent. NLP was used to extract tangible information, thus ignoring the structure of the documents, to identify trends. An example of a finding is an increase in trips in December when there is ice. Data science facilitates the identification of these relationships meaning corrective actions can be implemented.

The nature of accidents means that the data set for this project was very skewed. Non-conformance safety issues are regularly reported, but more serious contamination events may occur once every 10 years. This means the model does not have sufficient data to infer what may cause a contamination event.

3. **Proton spallation**

Neural networks (NN) have been used to try to simulate complex physical systems. Many physics-based models are used which have large run times, and there is a desire to reduce the run times by using a NN. This would also allow engineers to explore and optimise designs more quickly. This was attempted for a proton spallation model which models the complex physical phenomena when a proton is fired at an atom causing the atom to break apart. ML (regression and random forest) was attempted but did not work well. Additionally, polynomial regression did not work as there was no curve to fit all atoms.

CASE STUDY 2

Industry: Aerospace and Defence

Location: Germany

Summary:

An distinct internal data science resource exists within the organization to work as a data innovation catalyst. It exists to deliver untapped value from data, acting as a development hub for new services that improve how the organisation and their customers operate. Data science has been applied very successfully when undertaking condition monitoring and predictive maintenance. The key success criteria when undertaking data science with regards to condition monitoring (or 'engine health monitoring') is to have data of sufficient quality and coverage – for any piece of machinery at least half a year of data is required to capture seasonal variations which affect operation. Therefore, patience is key and results cannot be expected immediately.

Methods:

The true viability of data science projects can only be assessed after the project has started – approximately 33% of all data science projects are abandoned after starting as the data is not good enough. 'Data Trials' are used which encourage a hard and fast approach to investigate the viability of projects. These usually take 3 weeks, although they may take much more or less time. Week 1 focusses on developing an idea for the use case and gathering/cleaning the data, week 2 focusses on developing the idea, and week 3 focusses on building a minimum viable product, e.g. a dashboard or Power Point. When a customer approaches the dedicate data science resource, this method is all that can be promised to demonstrate that the request has been received and to evaluate whether there is enough data to pursue the project further.

Generally, an agile sprint or Kanban approach is used for data science activities.

A number of tools are used for data science, although they are quite restrictive on software use. These tools include: R for statistics, Python, Jupyter Notebooks for sharing knowledge and growing capability, H2O.ai for machine learning (highly recommended), Orange Canvas, Apache airflow (for data pipelines).

For big data management the following tools are used: Oracle, SQL, Elastic surge.

Team:

Data engineers look after databases and undertake activities where data is structured, cleaned, and stored. Data scientists cover everything from CPU architecture to visual story telling. Corporate IT is involved to some degree with data science. A team in India is used which specialise in IT and Mathematics to undertake data engineering activities and some data science. Generally, engineers with lots of software knowledge are preferred for data science projects as they are problem orientated. It has been found that pure data scientists, while competent with maths/ IT and using algorithms, do not understand the context of the data science application, i.e. the relationship between the data and function. An example of this was an aerospace company looking at anomalies in their data caused by step climbs (changes in aeroplane altitude). This was seen as an anomaly to the data scientist but not to engineers with domain expertise.

Challenges:

The organisation as a whole is generally paranoid when dealing with IT and information security and this is usually the biggest obstacle to any data science project. For example, the use of open-source software is difficult to get approved by IT.

Another major challenge is the high risk of failure associated with data science projects– as mentioned above 33% of projects do not progress through the initial stages.

Additionally, the use of agile planning or Kanban methods for data science projects is unfamiliar to traditional corporate working, which means there is a culture difference in how things are done. This can sometimes represent a challenge.

Successful Data Science Use Cases::

1. **Root Cause Analysis**

Most data science is correlation and with large enough data sets insights can be taken. This principle has been applied to root cause analysis. When a failure occurs in a gas turbine engine, large data sets of engine operating data have been used to create a 'stop-motion' sequence of events that lead to the failure. This has been used to highlight previously undiscovered failure mechanisms and therefore reduce the risk of future failures. This method can be applied for almost everything, from finance to logistics, but big data is needed so records cannot be incomplete as they often are.

There are challenges using NN as it can be difficult to see why the model has made a particular decision. Using regression models it is easier to visualize the results. The lack of clarity when using NN makes it more difficult to apply in a safety critical environment.

Hardware issues, such as HPC requirements for data science and lack of access to GPU.

- **Security:**

Security poses many challenges when working with data in the nuclear industry, especially regarding getting access to data from clients. Large data sets can be affected by security issues

Unsuccessful Data Science Use Cases:

1. **Fuel non-compliance**

An attempt was made to investigate non-compliances for fuel and previous investigations were unsuccessful. The internal data science resource was brought in and tried to correlate fuel non-compliances with a range of factors, such as weather, but no correlations existed. It is expected that a large chunk of data was missing, making the use of data science methods unfeasible.

CASE STUDY 3

Industry: Steel manufacturing

Location: UK

Summary:

An Advanced Analytics (AA) team was created in 2018 to realise operational cost savings through the use of data, data analytics, and machine learning in the heavy-end of production. The team initially consisted of just 3 members; process/plant engineers and technical graduates from backgrounds in engineering and pure mathematics. Initial training was given to the team to build capability in analytics. This was supported by a sister plant in the Netherlands, which already had a team of 25-30 people working in AA with additional support from external consultants. In the sister plant, resources were available to send 2-3 AA team members into a works area to form a data science project group. Such resources aren't yet available in the UK, where the team is now slightly larger and have access to a limited amount of external consultancy. The team now widely use machine learning and are venturing into Deep Learning.

Methods:

Data science projects were initially approached using a "Wave" structure, a system in which the whole team would work on one project. This format was unsuccessful as it was inefficient and has since been replaced by the agile sprint approach. The sprint approach makes it easier for regular assessments of the project and to redefine the project scope. This method has decreased project timescales and allowed for greater clarification in circumstances where data science has not been needed. With the sprint method, approximately 26 projects have been completed in the two years with 13 providing tangible results.

Tools used for data science include: Python, KNIME, Spotfire, ProModel.

Team:

Currently, data science project teams are formed of a member of the AA team, who functions as the project manager, and 1-4 process/plant engineers in relevant works areas. This provides expertise in the local system/process in addition to expertise in data analytics, thus bridging any skills gaps. Using this method ensures that there will be buy-in from the team who will be using the data science tools going forward as they have been actively involved throughout.

Challenges:

Building capability within the organisation concerning tracking and using data was a difficult task, and is still ongoing. Many members of the organization are either resistant to change, or simply unfamiliar (or in disagreement with) the potential to use data to achieve operational cost savings. The AA team believe that adding data science capability is well within the reach of many departments and individuals, as the process/project engineers know their area well and therefore have a good understanding of the data and desired outcomes.

Successful Data Science Use Cases::

1. **Controlling hot metal silicon content in the blast furnaces**

This project developed a model to predict the silicone build up in hot metal in the blast furnaces. A huge amount of operational expenditure is allocated to control the thermal performance of the furnaces, with operators likely to conservatively increase temperatures rather than risk cooling below a critical level. Blast furnace operators often adjust the temperature based on the changing amplitude, trends, and frequency of silicon content to reduce of the volatility of furnace conditions. However, the silicon content was determined from chemical analysis of pig-iron samples, meaning there was a delay (up to 6 hours) in feeding back information to improve thermal control. To increase the temperature in the furnace, coke is added. Each addition of coke added costs approximately £30k, and was justified purely on "gut feeling" in absence of good data.

By predicting the current silicon content in the hot metal, as well as its trend in a timely manner, the team aimed to improve the quality of pig iron and reduce the coke rate through the operation, therefore achieving significant cost savings.

Machine learning was used to assess past data (data points every hour for four years) with a 2nd degree polynomial logistic regression model. The model can estimate the current silicon content in the furnaces and the likely silicon content after 2hrs and 4hrs. This enables better management of the coke rate, and has achieved a significant reduction in operating costs. The initial model took into consideration 200 variables but this was reduced to 22 to increase model efficiency. Tests have revealed that the model is approximately 70% accurate. The model is used via a web-based dashboard that the process team can access.

The biggest challenge in producing this model was that the process and management team did not believe data science could achieve such predictions.

2. **Predicting emissions**

This project developed a model to predict flu gas emissions produced by furnaces to better understand pollutant levels and meet statutory requirements. The model uses various process variables and sensor data as inputs. Predicted emissions from the model are supplied to a regulatory body and are accepted as accurate; this gives the plant license to operate. The advanced nature of predictions reduces operational risk as the organisation are able to plan ahead when high emissions are predicted.

Unsuccessful Data Science Use Cases:

1. Scrap material chemical composition

A works area suggested that data science could be used to track and predict the chemical composition of scrap metal to be used in the BOS plant, thus improving understanding of the operating conditions in the BOS vessel. After some review by AA, it was deemed that data science was not an appropriate tool as the data collected was of poor quality and used very small sample sizes, with data from weekly chemical samples allowing anomalies to be over represented. Another factor limiting success was the expertise in the works area – there was a lack of general knowledge/understanding about data science practices.

CASE STUDY 4

Industry: Technical Consultancy

Location: UK

Summary:

The organisation provides technical consultancy for development of data capture and visualization applications and has extensive expertise using data science methods in a range of industries, including nuclear and pharmaceuticals.

Challenges:

A number of challenges were outlined which broadly fit into the following categories:

- **Culture:**

The biggest blockers are typically in managing expectations vs reality. As people, we see some logic behind a task and infer that a machine must be capable of carrying out that task. An example of this concerned a predictive signal for a project on a gas turbine engine. It was identified that the blades were cracking due to vibration. Due to the abundance of operating data available, it was thought the data could assist in root cause analysis. However, upon review of the data, a solution was not clearly visible and a different approach was needed.

Additionally, if the person leading the project isn't experienced in data or isn't able to reformulate their approach based on the data presented, this can present a major challenge in undertaking data science. Preconceived judgment may inform solutions if new methods are not embraced or trusted.

- **Technology:**

A major technological challenge is when the technology being used does not support the data correctly. There is a myriad of database technologies and depending on how they are being used and architected, they either can or cannot support different kinds of machine learning. For example, a simple recursive model can be used on many types of databases but for a deep learning model the data has to be extracted and used in a different way. That can create blockers to training models and accessing data.

More challenges arise when extracting value from data. Using off-the-shelf pre-trained algorithms, it is possible to get 80% of the value of data – the last 20% of value can be disproportionately expensive to realise. If people are not aware of this, they may hit that asymptotic curve where they think they are close to their acceptance criteria but the amount of effort needed to actually reach this criteria may increase exponentially. This can sometimes be overcome by having people involved in the process, rather than using purely computational methods.

Successful Data Science Use Cases:

Nuclear:

1. Part obsolescence using NLP

Due to the length of time machinery is in service in the nuclear industry parts become obsolete in the sense that they are no longer manufactured. The use case for this project was to investigate how to source functionally equivalent parts (pipes, valves, etc.) across different nuclear sites. When new parts arrive, the stock keeper would receive the part and a datasheet and would type in a description of the part into a database. The data was not gathered for any other purpose other than the stock keeper knowing what they had in their database. The data was a classic example of a traditional data source where value is only now being extracted. Problems arose because everyone would do this in a different way, so it was found that the same part would be described in many ways. This made searching through catalogues to find parts very difficult because there were thousands of entries and the search functions were rudimentary, resulting in many different search hits.

It was posited that NLP could be used to take in all database entries and mark similar parts by their description. This could save time on finding the part and help create an index into the different databases such that users can be approximately 80% confident that parts found in other databases are functionally equivalent to the part they are replacing. This would also save a lot of time. For labelling the data, a semi-supervised tool was developed which used the end-user to train the system. This negated the need for a full-time data scientist and allowed the user to do their job whilst improving the algorithm. Python packages which compare strings were leveraged in developing this tool.

2. Predictive maintenance of pumps

Pumps located in highly irradiated areas are not easily accessible for maintenance. When the plant is shut down, either all pumps can be inspected or data can be used to assess which pumps are most degraded and therefore require inspection (other pumps can be inspected if time is available). The latter approach saves days during shut downs. The pumps are old and have been in service for a long time, meaning data prior to any shut down or maintenance work could be interrogated. The amount of work carried out on the pumps was used to determine features for training models and the pumped were classified into categories (degraded, mildly degraded, not very degraded). This informed the basis of prioritization of the inspection work.

Pharmaceuticals:

3. Invoice processing

This use case was identified internally as manually processing invoices was expensive and could be replaced by machine learning. The criteria for success of the project was to get 100% accuracy when automating the process. In reality, reading cells, rows, and columns from paper is a surprisingly challenging problem. Illustrating this point required going through all the Platform-as-a-Service AI tools including AWS, Azure, and GCP. Analysis of a handful of the invoices with these tools showed there were errors in reading columns about 30% of the time. To utilize this, a workflow was built in which some invoice processors were redeployed and some were kept behind to correct where the tool failed and improve results going forward. Over time, this project aims to reach a point where only 5% of the original staff are working to support the tool.

A key consideration of this project concerned edge cases. An incorrect assumption made was that the input data types are static, so historical invoices would be representative of the future invoices. Noise in the model can be created when the printer is changed or when the customer’s name is put in a different place. These edge cases will require some model retraining, which is why a well-planned workflow is critical.

element