EAS 504 Applications of Data Science: Industry Overview
Assignment 3

Name: Aboli Rawanhate
UB number: 50374341

Q1. Describe the market sector or sub-space covered in this lecture

Mr. Ram Narasimhan's talk on the application of data science in industries such as aviation, healthcare, mining, power distribution, water, oil and gas, wind, rail, manufacturing, power generation, and others covered this market area. The industrial internet is being shaped by factors such as the internet of things, intelligent machines, Big Data and Analytics, and domain knowledge, according to him. As a result, focusing the above-mentioned sectors solely on data completely transformed them.
The outcome-oriented application of mathematics and physics-based analyses and models to real-world challenges in industrial operations is known as industrial data science. And the tools and methods needed to do so on a consistent and large scale. It focuses on determining what can be done to maximize equipment usage, increase equipment life, reduce maintenance, optimize fleets, optimize businesses, and manage contracts.
Three basic components of Industrial Data Science are:

1. Physics/Engineering-based models - It needs less data so it is powerful but is difficult to scale and maintain.
2. Empirical, heuristic rules and insights: It is easy to understand and experts knowledge is captured
3. Data-driven techniques – ML, statistics, optimization and visualization: Here the data is less, biased to parameter space of normal operation and easy to scale and maintain.

Q2. What data science related skills and technologies are commonly used in this sector?

In Industrial Data Science, time series data is commonly used to crunch data for highlighting (summarization), piecewise decomposition (segmentation), prediction, anomaly detection, indexing, clustering, and classification. Component failure models are built using a data-driven and probabilistic approach, as well as stochastic forecasting. For quality control of machines, unsupervised algorithms are used with Big Data techniques. Engine maintenance dates are calculated using analytics-based maintenance solutions.
Understanding the problem (Descriptive Analysis), condition-based model maintenance (Correlations Analysis), predicting future variables (Predictive Analysis), and optimizing with recommendations are all possible using machine learning algorithms (Prescriptive Analysis) In Industrial Data Science, there are a number of different data science-related skills and technologies, some of them are below:

- Clustering for detecting Anomaly
- Machine Learning and Artificial Intelligence
- Theoretical and Applied Statistics

- Prognostic Systems & Methods
- Image processing and Analysis
- Sensor and Signal processing
- Control and Optimization
- Process Engineering and Operation Science
- PCA and feature engineering
- Knowledge Discovery and Contextual Analysis
- Physical and Expert based modeling
- Verification, Validation and Metrology

Q3. How are data and computing related methods used in typical workflows in this sector? Illustrate with an example.

Development and Deployment are two steps in the Industrial Data Science pipeline. Data is acquired from assets such as client demands, use-cases, and test cases during the development process. The data is then examined using domain expertise and expert views. A preliminary hypothesis is established. Models are created and assessed based on the hypothesis. The models that pass the test cases are advanced to the deployment step.

The models will be tested with new data and the business results will be reviewed during the deployment phase. According to the demands of the client, more data is found, gathered, and integrated. Active learning algorithms are used, and the results are transmitted back to the development phase for further refinement and improvement.

Detailed workflow:

1. Workout:
    - Customer needs defined
    - Use cases identified
    - Data status established
    - OpMechs Established
    - Core Team Established
    - Test Cases Ided for model v&v
    - UAT criteria established
2. Data Exploration:
    - Data understanding
    - Domain Knowledge
    - Hypothesis Defined
    - MVP scoped & Iteration Cadence & Phasing Agreed with stakeholders
3. Analytic Development:
    - Analysis options developed, scope refined as agreed with customer
    - Models developed, evaluated, refined and verified
    - Additional data identified, collected, integrated as needed and agreed
    - Model validation against test cases
4. Analytic Hardening:
    - Solution scaled for target environment

    - SOPs established for process integration
    - Solutions Deployed, tested in target environment
    - Model translated into predix-ready or customer framework-ready state
5. User Acceptance Testing:
    - Solution tested by customer to UAT criteria
6. Project Closure:
    - Ongoing Support Established as agreed w/customer
    - Project artifacts & documentation integrated into repositories

Example: Time series data handling- In time series, the relationship to the timestamp must be preserved at all times during the development phase, the observations cannot be scrambled, and each observation is reliant on its neighbor. Cross validation cannot be done in the usual method, and a random sample cannot be used as a validation set. We can derive from time series data using computing techniques. For example, we can crunch data to convey highlights, piecewise decompose it to get features, predict tn+1 value, detect anomaly, find similar patterns and time series, find natural groups in time series data, and assign records to classes.

Q4. What are the data science related challenges one might encounter in this domain?

- The data is of poor quality. With noise, we obtain pictures, diagrams, and signal data. The information gathered from assets is unstructured and in an improper format. To get it into a suitable format, a lot of feature engineering is required. Even so, the outcomes might be negligible.
- The most significant part of solving an IDS problem is domain expertise. Because data science is a combination of machine learning and other analytical techniques, customer assistance is required, as data scientists may not be experts in all fields. Lack of communication might have unfavorable consequences.
- The fact that there is a class imbalance and a small number of failure data points makes it difficult to train effective models for industrial data science challenges.
- Noise in signal data can be treated as valid data points and hence model development will be affected.
- Data amount varies greatly - since the variables are limited, data must be gathered from other sources. Design specifications, maintenance records, and operational data are examples of these sources. It can be unreliable at times.
- We can’t sample random points for validation in cross validation for Time series data as it will be out of sequence. So we need to carefully sample it.
- Classification of data points is challenging. As the records depend on its neighbors it is difficult to know the labels of classes they belong to.
- Models created for a certain element in an industry might soon become obsolete if those variables change. Unusual circumstances can cause models to be faulty.
- Clients frequently believe that Data Science is the answer to all of their problems. Customers compel the Data scientist to employ a certain method or approach merely for the sake of name recognition. After identifying the problem and doing data exploration, we must pick a solution approach.

- The problem we are solving is important to the client and worth working on is also the main factor to consider.

Process integration becomes crucial as a result of all of these considerations. When employing data science approaches in the industrial realm, the asset lifespan is crucial.

Q5. What do you find interesting about the nature of data science opportunities in this domain?

Data Science possibilities abound in Industrial Data Science, but they do need some domain knowledge. Connectivity and analysis can help revolutionize industrial processes.

- It is used in the healthcare industry for asset management, patient safety, patient flow, and network cooperation.
- It is used in the oil and gas sector to determine overall equipment dependability, system optimization, field production, and pipeline insights.
- Grid delivery optimization and improved meter insights are used in power distribution.
- Resource optimization used to match equipments to industry needs
- In mining industries IDS is used to maximize mine performance
- It is used in manufacturing to improve the production process and ensure product safety.
- Aviation industries uses Industrial data science to optimize fuel consumption, navigation insights and flight synchronization, risk insights
- Industrial data science is also utilized to improve the efficiency and reliability of water plan operations, as well as to calculate locomotive uptime, maintenance, and parts management for a variety of machines.
- During severe weather, machine learning algorithms may be used to determine power outages at high resolution across the country..
- Anomaly detection to predict the failure
- Prediction of how long will the equipment or product remain operational
- Demand forecasting to retain continuous supply for the needs
- Insurance domain can have prediction of time until claim useful
- Retail industry needs information of when will the customer buy the next product
- To detect failure post-mortems - What was the root cause of failure
- Process optimization to improve asset performance.

Q6. i. Discuss some of the characteristics of industrial data science problems alluded to in the talk and how they differ qualitatively from data science problems in other domains.

- The fundamental distinction between Industrial Data Science challenges and other Data Science problems is that the processing activities of the industry carry a very high level of risk. We may look at areas like telecommunications, banking, investing, and accounting services, all of which have a lot of data and are looking for increased productivity and profits. Other industries, such as power, water, rail, and oil & gas, are concerned with data and how it will help people. These examples use diverse

approaches, but they all fall under the umbrella of industrial data science. However, the Data Science sector operates in a different way.

- In Industrial, false positives can be costly as we have to stop operation.However in data science we do not face much issue with the false positives.
- Because the machines in IDS are so large and failure occurs so seldom, the frequency of occurrences is so low that our model treats it as noise, making it impossible to train the model appropriately. The frequency of events is large/ enormous in terms of click-through and media acquisition.
- In Industrial Data Science, there are a lot of diversity factors. Machine failure can be caused by a variety of factors. Because recommendation systems and media buy precise information about the viewer and present material accordingly, variety is limited.
- The amount of data in Industrial Data Science varies greatly depending on the sort of problem we're dealing with and the gear that's involved. The amount of data in click-through and media buy is incredibly huge.
- Expert opinion and domain expertise are essential. However, planning takes place at the start of the development process, and data science professionals are not included. As a result, there is a communication gap.
- Customers usually want the finest outcomes, which is why sophisticated machine learning approaches such as Predictive Analysis and Deep Learning are so popular. However, selecting a strategy is totally dependent on the issue specification and the available data.
- Data should be simple and easy to understand. The fundamental problem in Industrial Data Science, on the other hand, is the increasingly scattered nature of industrial data sets. When data is collected from a variety of sources, it is often unstructured and presented in a disjointed manner. As a result, it complicates access to, integration with, and exchange of industrial data. As a result, many businesses can acquire data but are unable to adequately extract and evaluate it.
- In Industrial data science, time series models are majorly used, however websites like facebook, google use graph theory and networks to determine our interests.

Q6. ii. MCQ:

Q1. A

Q2. B

Q3. D

Q4. E

Q5. E