IMPROVING PROJECT TIMELINES USING AI / ML TO DETECT FORECASTING ERRORS

By

David Goldberg

B.Sc. Biomedical Engineering, Technion - Israel Institute of Technology, 2014

SUBMITTED TO THE MIT SLOAN SCHOOL OF MANAGEMENT
AND THE ELECTRICAL ENGINEERING AND COMPUTER SCIENCE DEPARTMENT
IN PARTIAL FULFILLMENT OF THE REQUIREMENTS FOR THE DEGREES OF

MASTER OF BUSINESS ADMINISTRATION
AND
MASTER OF SCIENCE IN ELECTRICAL ENGINEERING AND COMPUTER SCIENCE

IN CONJUNCTION WITH THE LEADERS FOR GLOBAL OPERATIONS PROGRAM AT THE

MASSACHUSETTS INSTITUTE OF TECHNOLOGY
JUNE 2019

Author: Signature redacted

MIT Sloan School of Management
and the Electrical Engineering and Computer Science Department
May 10, 2019

Certified by: Signature redacted

Roy Welsch
Eastman Kodak Leaders for Global Operations Professor of Management
Professor of Statistics and Data Science
Thesis Supervisor

Certified by: Signature redacted

Patrick Henry Winston
Ford Professor of Artificial Intelligence and Computer Science
Thesis Supervisor

Accepted by: Signature redacted

Maura Herson
Assistant Dean, MBA Program
MIT Sloan School of Management

Accepted by: Signature redacted

Leslie A. Kolodziejski
Chair of the Committee on Graduate Students
Department of Electrical Engineering and Computer Science

This page intentionally left blank.

IMPROVING PROJECT TIMELINES USING AI / ML TO DETECT FORECASTING ERRORS

By

David Goldberg

Submitted to the MIT Sloan School of Management and the Electrical Engineering and Computer Science Department on May 10, 2019 in partial fulfillment of the requirements for the degrees of Master of Business Administration and Master of Science in Electrical Engineering and Computer Science in conjunction with the Leaders for Global Operations Program

ABSTRACT

This project focuses on the creation of a novel tool to detect and flag potential errors within Amgen's capacity management forecast data, in an automated manner using statistical analysis, artificial intelligence and machine learning. User interaction allows the tool to learn from experience, improving over time. **While the tool created here focuses on a specific set of Amgen's data, the framework, approach and techniques offered herein can more broadly be applied to detect anomalies and errors in other sets of data from across industries and functions.**

By detecting errors in Amgen's data, the tool improves data robustness and forecasts, which drive decisions, actions and ultimately results. Flagging and correcting this data allows for overcoming errors, which would otherwise damage the accurate allocation of Amgen's human resources to activities in the drug pipeline, ultimately hampering Amgen's ability to develop drugs for patients efficiently.

A user interface (UI) dashboard evaluates the tool's performance, tracking the number of errors correctly identified, the accuracy rate, and the estimated business impact. To date the tool has identified 893 corrected errors with a 99.2% accuracy rate and an estimated business impact of $77.798M optimized resources. Using the paradigm of intelligent augmentation (IA), this tool empowers employees by focusing their attention and saving them time. The tool handles the human-impossible task of sifting through thousands of lines and hundreds of thousands of data points. The human user then makes decisions and takes action based on the tool provided output.

***Keywords*:** anomaly detection, error detection, automation, artificial intelligence, machine learning, project management, intelligent augmentation.

Thesis Supervisor: Roy Welsch
Title: Eastman Kodak Leaders for Global Operations Professor of Management, Professor of Statistics and Data Science

Thesis Supervisor: Patrick Henry Winston
Title: Ford Professor of Artificial Intelligence and Computer Science

This page intentionally left blank.

ACKNOWLEDGEMENTS

First, and foremost I would like to express my sincere gratitude to my thesis advisors. Without Guy Schwartz's committed support, energy and never-ending stream of excellent ideas and advice, this work quite literally would have been impossible. Roy and Patrick, thank you for your advice and input throughout this journey, nudging me back on track as needed.

Thank you to the incredible Amgen family, who welcomed me in, trusted me and pushed me to do more. Thank you to Dollie Grajczak who moves mountains for her LGOs and has the warmest heart. Thank you Cleat Jerden for your ability to instill calm and focus. Thank you Leigh Hunnicutt, the ultimate LGO role model. Thank you Aine Hanly and Jerry Murry for the opportunity to work on interesting side projects. Thank you Roger Hart, one of the smartest people around and someone who tailor makes opportunities and experiences. Thank you to the extended Amgen LGO family for your advice, insights and friendship. Thank you to all the interns for our shared experience, especially the "Off-Cycle Mafia" Emi, Hillary, Martin and Yucen.

Thank you to the LGO program and staff, especially Patty who is the glue that keeps us from falling apart. To my fellow LGO students, I am in awe of your accomplishments and I learn from you daily – thank you.

Last and most importantly, thank you to my amazing family. Traversing the globe, you have given constant support throughout the many adventures and you are always there when needed. Without you, I would not be able to do what I do. Together we laugh, cry, and live. This work is every bit as much your accomplishment as it is mine. Thank you.

There are many unnamed-here others that were also part of this journey. I thank you sincerely.

This page intentionally left blank.

The Author wishes to acknowledge the Leaders

for Global Operations Program for its support of this work.

You can do it,

Only you can do it,

You cannot do it alone.

If you take care of your people, you will never be alone.

- - Patrick Henry Winston

TABLE OF CONTENTS

This page intentionally left blank.

1 Introduction

In any complex discovery project, project management works to minimize cost and to maximize speed for the project's successful completion. In Amgen's continuous quest to improve drug development timelines, innovative approaches such as data-science can help to ensure an operational advantage. This section describes the specific project motivations, which led to the commission of this work, and then goes on to define the Problem and Thesis Statement, before concluding with the Thesis Overview.

1.1 PROJECT MOTIVATION: Taking forecasting to the next level

Models implemented by Amgen's Process Development organization have historically allowed for forecasting drug development resources about 8 quarters into the future. Forecasting is a critical ability, which allows the company to prepare and adapt in efficient ways, ultimately lowering costs. For example, earlier and more accurate forecasting could allow for the prevention of bottlenecks or the allocation and hiring of people with needed skills. There are several opportunities for Amgen to improve existing models:

1) Improve input data integrity - using machine learning in conjunction with established rules one can flag for inspection inputs that deviate from what is defined as 'normal'. Examples of such deviations include a mistyped input (25 FTE instead of 2.5 FTE) or a faulty assumption (5 FTEs for a job that typically needs 2 FTE). The ability to successfully flag deviant inputs, while controlling for false positives, can be based on established rules from the current 'sources of truth' (SMEs and documented process norms) and additionally from using statistical methods on current vetted forecasts.
2) Gathering more complete and useful data to guide decision making – having less missing data and data that is more accurate, will allow Amgen to better understand

opportunity costs and strategic drivers. Amgen is fortunate to have more projects available than can be executed. By understanding the qualitative and quantitative considerations one can construct an algorithm that optimizes for "value". "Value" could include considerations for NPV, contractual requirements, strategic importance, etc. This algorithm could propose, which projects to work on and how best to include a new project with minimum 'damage' to the projects currently in progress. One example use case would be deciding, which few hundred work streams should we delay/cancel in order to prioritize a specific drug's development; currently this would involve gathering several experts in a room to talk it out, without a clear way to know if their intuition has led to the best decision.

3) Creating a usable tool for the business – being able to construct a tool that the business implements and finds useful will allow for real impact. Such a tool will need to have a good user interface and design. The tool should help to support employee's needs, providing them with relevant and timely information. To this aim it will be important to understand how best to develop such a tool, gaining insights and support from those who will ultimately use the tool.

4) Mapping automatically the skillsets of people – there is the potential to create a mechanism to understand what skills employees have and could develop based on the projects and work they have been involved with thus far, using existing logged system data on employees work times and types. Examples might include, if an employee has worked in a package for X time, they have this skill and they have the potential to learn this other skill. A database on existing and potential skills within the company could be created for use in project demand and supply planning. This

planning will help to pinpoint bottlenecks, identify training opportunities, and prevent shortages of people with specific skills.

In Summary, Amgen's Process Development can realize large benefits by improving and updating their existing approach to forecasting around the various drug development projects in their drug pipeline. These benefits can be realized in reduced cost, but also in shorter development time, which drives earlier and increased revenues, and more importantly allows for patients to receive treatments sooner.

1.2 PROBLEM STATEMENT: Automating Data Robustness

Amgen Process Development has recently taken the heroic step of consolidating huge amounts of their forecast planning data into a centralized system which throughout this work will be referred to as FSM (Forecast System Management – an assumed name) where various users are able to interact with this data in different ways, including planning and risk mitigation. With this progress, it has become clear that the user generated planning data is far from perfect and includes several errors that have the potential to delay drug development. Due to the critical nature of accurate forecasting, it is imperative to improve the existing data and create forecasts that are more robust. Human screening of the tens of thousands of line items constituting the planning data for errors and issues is unreliable and time consuming, therefore automated approaches should be explored as a valuable aid in this endeavor.

1.3 THESIS STATEMENT: Does data-science hold the key?

Can a data-science approach, including artificial intelligence (AI) and machine learning (ML), aid in ensuring data robustness for project management activities (specifically for Amgen's capacity management forecast data)?

1.4 THESIS OVERVIEW: TL;DR[1] Yes! The tool works!

This thesis focuses on the creation of a novel tool to detect and flag errors in Amgen's capacity management forecast data. User interaction allows the tool to learn from experience, improving over time.

By detecting errors in Amgen's data, the tool improves data robustness and forecasts, which drive decisions, actions and ultimately results. Flagging and correcting this data allows for overcoming errors, which would otherwise damage the accurate allocation of Amgen's human resources to activities in the drug pipeline, ultimately hampering Amgen's ability to develop drugs for patients efficiently.

Success of this endeavor is evaluated using a UI dashboard, which tracks the number of errors correctly identified, the accuracy rate, and the estimated business impact. To date the tool has identified 893 corrected errors with a 99.2% accuracy rate and an estimated business impact of $77.798M non-optimized resources. Using the paradigm of intelligent augmentation (IA), this tool empowers employees by focusing their attention and saving them time. The tool handles the human-impossible task of sifting through thousands of lines and hundreds of thousands of data points. The human user then makes decisions and takes action based on the tool provided output.

---

[1] TL;DR – Too Long; Didn't Read. Slang for 'the main point is'.

2 Background

The Biopharma industry is high risk, high reward industry and Amgen is a strong player, that strives to continuously improve and excel. Steps that allow Amgen to better plan future work, help to reduce risk and accelerate drug development. One critical component that drives this success is better quality planning and forecast data. Improving this data is a major focus of this project. This section provides a background to the biopharma industry and Amgen Inc. It goes on to discuss Amgen's organizational structure and wherein the work in this paper took place. Contextual information on Amgen's portfolio resource forecasting is discussed to provide some additional insight into the motivation for this work.

2.1 INDUSTRY OVERVIEW: Biotech is high tech and knowledge intensive

The pharmaceutical industry is comprised of companies engaged in researching, developing, manufacturing and distributing drugs for human or veterinary use. New drugs have an enormous positive influence on global health, prosperity and economic productivity by saving lives, increasing life spans, reducing suffering, preventing surgeries and shortening hospital stays. Advances in medicine have eliminated deadly diseases and have brought other life-threatening conditions under control. Drug therapy is now an integral part of nearly every facet of healthcare, and new breakthroughs promise to revolutionize the treatment of non-communicable diseases. (Department of Commerce, USA)

Biologics (biotech drugs, biological drugs, biopharmaceuticals) include a wide range of products such as vaccines, therapeutic proteins, blood and blood components, tissues, etc. In contrast to chemically synthesized drugs, which have a well-defined structure and can be thoroughly verified, biologics are derived from living material (human, animal, microorganism or plant) and are vastly larger and more complex in structure. Biologic medicines are revolutionizing the treatment of cancer and autoimmune disorders and are critical to the future of the industry. (Department of Commerce, USA)

The biopharma industry is characterized by being high tech and knowledge intensive, for example it includes expertise such as gene cloning, purification, and genetic engineering. Large upfront investment costs are needed for research and development and these costs are amplified by the low rate of success of drugs reaching the market. The average estimated cost to bring a drug to market is $2 billion. The drug development lifecycle is exceptionally long at approximately 10 years, and as mentioned there is a high risk of failure along the way, including throughout the various stages of the FDA trials. These difficulties are counterbalanced by the enormous potential revenues generated from a successful drug and the associated temporary monopoly created by its patent. The payback period of a successful drug is about 2-3 years and returns can exceed 10 times the original very large investment.

These large revenues have fueled accelerated growth in the biopharma industry over the last 30 years, but as the industry has matured, major firms are beginning to balance revenue growth with reducing costs and becoming more efficient. As of 2018, major players in the pharma and biopharma space include Pfizer, Merck, Johnson & Johnson, Roche, Sanofi, Novartis, AbbVie, AstraZeneca, Gilead Sciences, and Amgen.

2.2 AMGEN: A world leader enabled by innovative technologies

Amgen is one of the world's leading biotechnology companies. A biotechnology innovator since 1980, Amgen has grown to be one of the world's leading independent biotechnology companies, has reached millions of patients around the world and is developing a pipeline of medicines with breakaway potential. Amgen strives to serve patients by transforming the promise of science and biotechnology into therapies that have the power to restore health or save lives. (Amgen, Inc., 2018)

Amgen is committed to unlocking the potential of biology for patients suffering from serious illnesses by discovering, developing, manufacturing and delivering innovative human therapeutics. Amgen's belief—and the core of their strategy—is that innovative, highly differentiated medicines that provide large clinical benefits in addressing serious diseases are medicines that will not only help patients, but also will help reduce the social and economic burden of disease in society today. (Amgen, Inc., 2018)

Amgen has a presence in approximately 100 countries and regions worldwide and their innovative medicines have reached millions of people in the fight against serious illnesses. Amgen currently focuses on six therapeutic areas: cardiovascular disease, oncology, bone health, neuroscience, nephrology and inflammation. (Amgen, Inc., 2018).

With a market cap of $130.088 billion (Yahoo Finance, January 2018), Amgen had $22.8 billion in revenue in 2017 (Amgen, Inc financial statements). Amgen's approx. 20,000 employees are working on at least 40 drugs in the pipeline, five of which are biosimilars (amgenpipeline.com, 2018). Amgen currently manufactures 17 drugs, detailed in Table 1-1 (Amgen, Inc., 2018).

Table 1. Amgen's Commercial Products.

Drug	Therapeutic Area
Aimovig® (erenumab-aooe)	For the preventive treatment of migraine in adults
Aranesp® (darbepoetin alfa)	For the treatment of anemia caused by kidney failure or chemotherapy.
BLINCYTO® (blinatumomab)	For the treatment of Philadelphia chromosome-negative relapsed /refractory B cell precursor acute lymphoblastic leukemia
Corlanor® (ivabradine)	For the treatment of chronic heart failure
Enbrel® (etanercept)	For the treatment of autoimmune diseases such as rheumatoid arthritis
EPOGEN® (epoetin alfa)	For the treatment of anemia
IMLYGIC® (talimogene laherparepvec)	For the treatment of unresectable recurrent melanoma
KYPROLIS® (carfilzomib)	For the reatment of multiple myeloma
Neulasta® (pegfilgrastim)	Treatment to decrease the chance of infection by febrile neutropenia in patients receiving chemotherapy
NEUPOGEN® (filgrastim)	Treatment for slow white blood cell recovery following chemotherapy
Nplate® (romiplostim)	For the treatment of thrombocytopenia in patients with chronic immune (idiopathic) thrombocytopenic purpura
Parsabiv® (etelcalcetide)	For the treatment of secondary hyperparathyroidism in adults with chronic kidney disease on hemodialysis
Prolia® (denosumab)	For the treatment of postmenopausal women with osteoporosis at high risk for fracture
Repatha® (evolocumab)	For the treatment of high cholesterol
Sensipar®/Mimpara® (cinacalcet)	For the treatment of secondary hyperparathyroidism and hypercalcemia in parathyroid carcinoma patients
Vectibix® (panitumumab)	For the treatment of colorectal cancer
XGEVA® (denosumab)	For the prevention of skeletal-related events in patients with bone metastases from solid tumors

Core to Amgen's business is developing these drugs as safely and quickly as possible, while balancing tight timelines, limited resources and financial investments. Effective project management and planning of this complex task is imperative to Amgen's success. Utilizing innovative technology such as data science is a key enabler for Amgen to achieve this goal.

2.3 ORGANIZATIONAL STRUCTURE: The cross-functional impact of this work

In this section, I will describe Amgen's organizational structure (as of 2018) with emphasis on the arms of the organization that I was most involved with. This description is useful for understanding the impact and scope of this work and is also helpful in interpreting the data that is referred to later in this paper.

2.3.1 The Amgen organization

Amgen has approximately 20,000 employees and three major functions, including Research & Development (R&D), Operations and Commercial. Within Operations there are five distinct sub functions Quality, Engineering, Supply Chain, Manufacturing and Process Development. The work referred to in this thesis primarily involves activities in Process Development, which I will explore in more detail below.

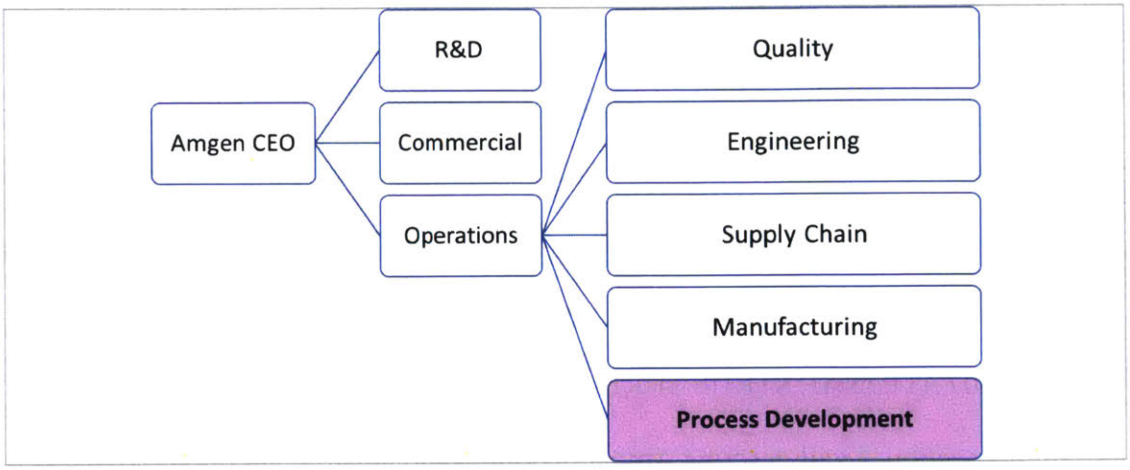


Figure 1. The five sub functions of Amgen Operations.

2.3.2 Process Development (PD)

Amgen aims to deliver therapies to patients as quickly as possible. A key change to achieve this goal, was the creation of an integrated Process Development (PD) organization operating under a single lead. This structure combines Process Development functions from Research and Development and Operations into a unified organization within Operations under the leadership of Jerry Murry, senior vice president.

The redesign of this complex and diverse network has increased efficiency and speed, creating a more cohesive organization generating value for Amgen—ultimately helping Amgen reach the clinic and market faster. Throughout this evolution, investing in technology and preserving PD's culture of scientific innovation has remained a core guiding principle as the organization implemented an integrated, adaptable approach across Amgen's entire product lifecycle.

PD includes approximately thousands of employees who take scientific development (from R&D) and turn it into a product approved for commercialization, after which the product is handed off to the larger Operation's sub-functions for scaled up production. PD is a mini

Operations in its own right, encompassing activities that cover Quality, Engineering, Supply Chain, and Manufacturing.

The PD organizational structure includes a number of functions, each with their own functional analyst (FA – an assumed name). These FAs own their own portion of the dataset (3.2.1 Input Data) I have been working on and as such, interactions with these FAs and their data was an important part of this work. Additional meetings with the VPs leading these PD functions provided valuable insight and context to how my work affected and aided their functions. These meetings also served to generate stakeholder buy-in and engagement, which was imperative for later user acceptance.

2.4 PORTFOLIO RESOURCE FORECASTING: Past standardization and alignment have allowed for automation opportunities

In this section I will elaborate on PD's portfolio resource forecasting, to provide further context and motivation for the work described in this paper. I will show that although, great strides have been made already, there is still much to be done to help improve the decision making around which drugs are developed by Amgen and when.

2.4.1 History

Historically, each team in PD had their own, siloed resource plan complete with separate nomenclatures and defined frameworks for estimating resources that would be need in drug development activities. Consolidating these different plans into an overarching plan was a great challenge. PD set out to standardize and align these processes and templates so that consolidation could be more streamlined. The "Flip", constructing and transitioning to these templates, was a big step, which was completed in 2017.

The next step was to combine these inputs, looking for misalignments of resources and timelines, such as over-demand for a resource at a point in time across projects. The solution was to 'hack it out' in meetings in Q2 2017 and to align on the master plan by typically taking a bottom up approach, rolling upward and checking for issues.

Results were pretty good with a forecast until the end of 2018 established. However, the broader goal is to have a long term (10 year) forecast. The longer-term forecast would leverage the aforementioned models and templates, and would be tagged onto the shorter (1-2 year) forecast, with these plans being regularly updated as time goes on.

2.4.2 Forecast System Management (FSM)

With nomenclature now aligned, an additional opportunity to streamline the process presented itself. Guy Schwartz set out to build a system to consolidate and hold the various resource plans. The solution was FSM, a database that held the resource plans for the different functions. These plans could be entered directly or imported from a template Excel file. With all the information in one place, consolidation was practically automatic and countless hours were saved. Additional features such as scenario analysis and visualization were also incorporated as the system evolved.

2.4.3 The Opportunity

FSM facilitated better creation, access and uses of the resource planning, which in turn better facilitated decision making around which drugs should be developed and when – dictating when drugs would ultimately reach patients. With this progress, it had become clear that the user generated planning data was far from perfect and included several errors that have the potential to delay drug development. Due to the critical nature of accurate forecasting, it is imperative to improve the existing data, creating more robust forecasts. Human screening of the tens of thousands of line items constituting the planning data for errors and issues is unreliable and time consuming, therefore automated approaches could serve as a valuable aid in this endeavor.

This page intentionally left blank.

3 Approach and Methodology

With the "need" formally established I set out to build a tool to automate the review of the vast amounts of data in the existing FSM system. The goal was to understand the input data, create a tool architecture to detect and flag the errors, and finally to serve this to the end users in a useful way. In this section I will describe the data and tool architecture in detail, and how they fit together. I will then go on to explain the tool's anomaly detection in greater detail and explore how the user interacts with the tool.

3.1 SYSTEM FRAMEWORK: How the pieces fit together

Taking a systems and signals approach, the framework I decided to use was one of input-transformation-output. The input is the FSM data and the output is the tool generated data that specifies which input data was flagged as being anomalous, as well as context as to why it was flagged. The core of this work was the creation of the error detection tool, which I created as a series of modules. On either side, the system needs to interact with end users through a user interface (UI). FSM already had a UI and I was faced with the dilemma of creating a new flexible, shiny app or using the existing FSM interface. Motivated by user adoption, I ultimately decided to forego creating a new app, and instead used the existing FSM interface, adding additional screens for the new information, where needed.

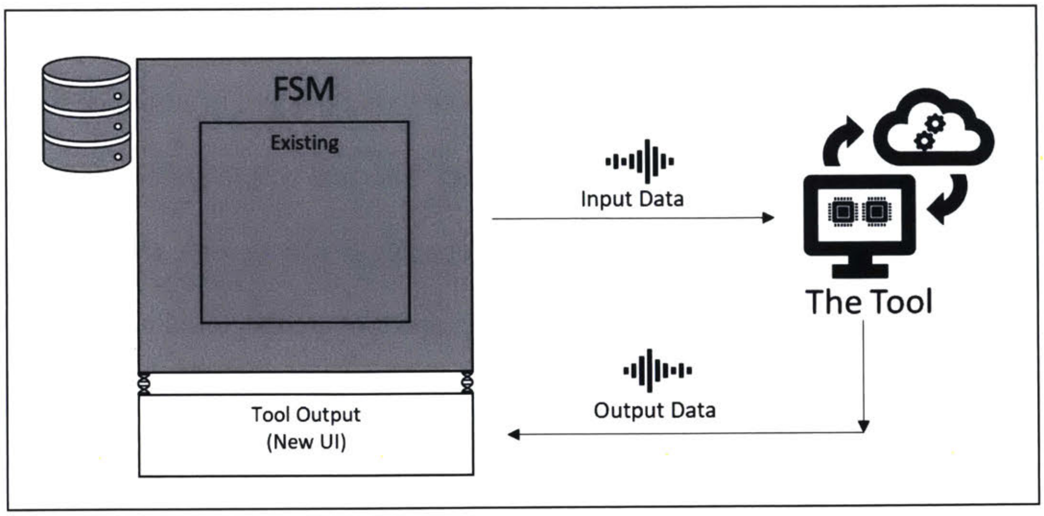


Figure 2. The system architecture: Input from FSM is passed to the anomaly detection tool. Output is passed back to FSM into a user interface added within FSM.

3.2 DATA AND TOOL ARCHITECTURE: Input, output and the tool.

In this section I will explore the input data and output data in more detail, discussing their compositions and elaborating on the final tool architecture that transforms the input into the output.

3.2.1 Input Data

The source data from this work is pulled directly from the FSM system, specifically for large molecule (biologics) drugs. The data is comprised of a database of line items with various fields. There are around ten-thousand line items, each with approximately 60 data points. Excluding the identity (ID) field there are two other types of fields making up these data points:

1. Metadata Fields – These fields provide context as to which type of drug is being worked on, when and by which teams. The specific fields can be seen in the "Imported Fields" image.
2. Quarterly FTE Data – These fields are numeric fields signifying how many FTEs have been allocated for this line item of work for each quarter. Quarters run from

2018 Q1through 2030 Q4. The values here are of type decimal and can also be 0, signifying no work being allocated for that period.

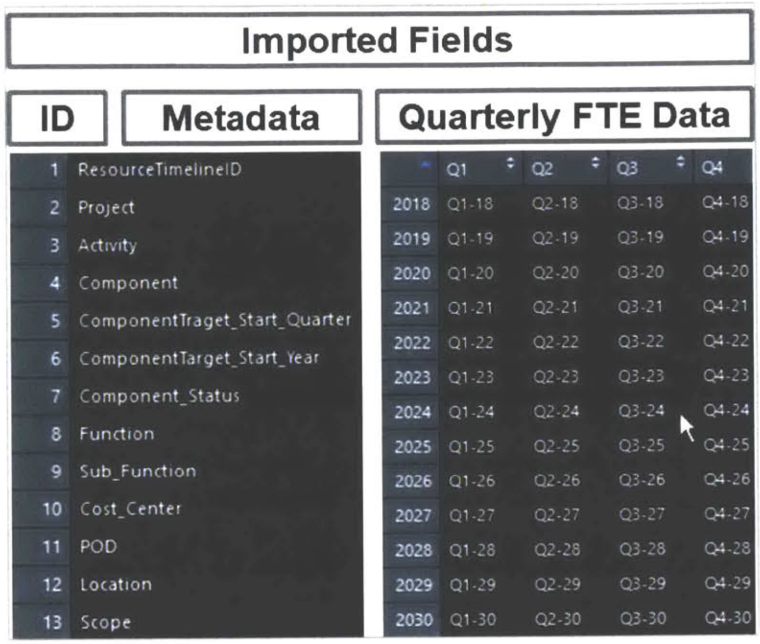


Figure 3. The input data from FSM consists of ID, Metadata and Quarterly FTE Data field types.

3.2.2 Output Data

Through a series of steps and functions the raw input data is converted into several calculated fields that will help inform and direct end-user decisions and interaction. These calculated fields include:

1. Attributes – calculated attributes are more informative than the raw data and are used down the line for comparison to detect anomalies. The attributes calculated are:
   i. Duration – The number of quarters from the first time FTEs are assigned until the last time FTEs are assigned i.e. non-zero FTE allocation.
   ii. Total resources – The integral (sum) of all FTEs assigned across quarters

iii. Maximum resources – The peak number of FTEs assigned across quarters

2. Attribute Stats – Basic statistical attributes are calculated across similarly clustered line items, to aid in the detection of anomalies. The upper limit, lower limit, median and mean are calculated for these groups.
3. Flags – Line items are marked as being potentially anomalous. A line item may have more than one flag. These flags serve first to draw the user's attention to a potential issue and then to aid in the user's investigation and analysis thereof. A line item may be marked with the following types of flags.

   i. Logical – These flags are based on the pure quarterly data. There are several expert system rules, which have been implemented based on the insights of subject matter experts (SMEs). One may define these rules as expert system artificial intelligence (AI) rules. The rules checked here include duplicate entries, overly similar entries, discontinuities and zero-demand allocations. Each logical rule has an associated flag. (For more on the logical anomalies detected, see 3.3.1 Logical anomalies: AI expert system rules)

   ii. Statistical – These flags are based on the Duration, Total Resources and Maximum Resources attributes, which in turn are built from both the metadata (for clustering) and the quarterly data. Each attribute value is compared to its statistical attribute limits and if found to be a statistical outlier, the relevant flag(s) are marked. (For more on the statistical anomalies detected and how outliers are defined, see 3.3.2 Statistical anomalies: Catching mathematical deviants)

iii. Relationship – Even more strongly based on the metadata, the relationship-based flags are built from machine learned relationships between the different activities in the various drug-projects. Relationships rules are machine learned across the data set, such as if two activities usually exist together or when they begin/end. Once learned, these rules are tested against to flag for potential issues. The initial rules implemented by this tool are 'existence', and four project management relationships: start-start, start-finish, finish-start, and finish-finish. (For more on the relationship anomalies detected, see 3.3.3 Relational Anomalies: Machine learned relationships)

iv. Any – A simple check to see if any other flag has been assigned. This aids with various logistical operations and calculations down the line.

3.2.3 Final tool Architecture

The final tool architecture is comprised of 6 primary modules and an additional supporting module (module A) that provides a measurement of how the tool is performing. The modules are:

1. Input – Data is provided to the system via interfacing with the existing FSM database. The data is comprised of the cross functional FTE forecast data input by the FAs and associated metadata
2. Classify – Data entries are clustered based on the metadata fields so that we can "compare like with like". Practically this refers to ensuring that the forecast data

for drugs with similar attributes is being compared with one another e.g. large molecule vs small molecule.

3. Attributes – Based on the underlying data relevant attributes are calculated such as duration of work, total work etc. These attributes will be used for comparison across the segmented data to detect anomalies.
4. Detection – With the attributes calculated, the system runs tests to detect anomalies, which are potential issues, that will be flagged for the end users' attention. This step also calculates and stores baseline "normal" attributes for comparison, which can aid in the end users' later investigation
5. Flagging – With anomalies detected the system internally marks which entries should be brought to the end users' attention and for which reason(s).
6. User Interface (UI) – Until this point the system has completed its core technical purpose and now needs to serve the information to the end users in a user-friendly manner. The user interactions at this point, also provide feedback for future iterations.

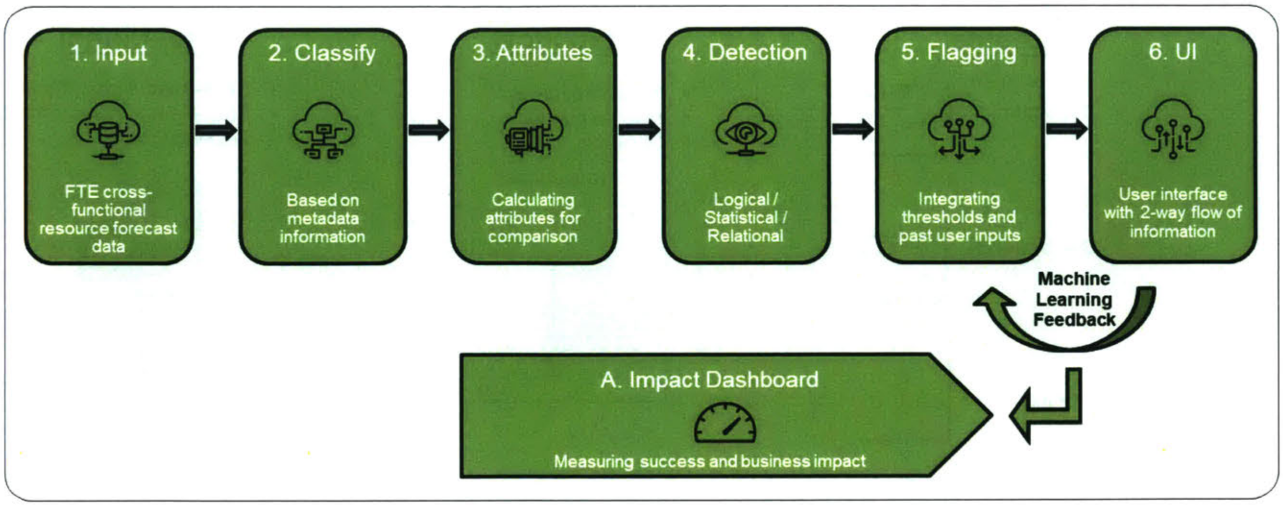


Figure 4. The final tool architecture is comprised of 6 primary modules and an additional supporting module (module A) that provides a measurement of how the tool is performing.

A. Impact Dashboard – The impact dashboard measures the results of the system and provides a range of metrics to the business so that impact and success can be monitored.

3.3 THE ANOMALIES DETECTED: Logical, Statistical and Relational

Consistent with the anomaly flags associated with the output data, there are three types of anomalies detected: Logical, Statistical and Relational.

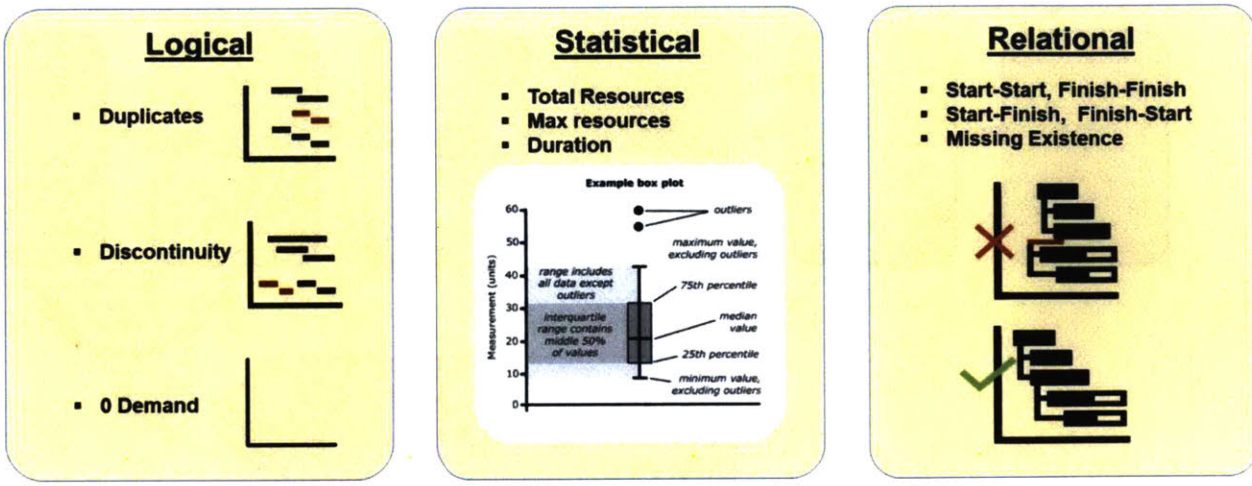


Figure 5. An overview of the types of anomalies detected by the tool.

3.3.1 Logical anomalies: AI expert system rules

Based on subject matter experts' insight, these AI expert system logical-rules detect line items that are too similar, exact duplicates, discontinuous, or have zero demand. Using relatively simple algorithms, these logical-rules can be tested directly on the underlying data.

- Line items that are too similar or exact duplicates are suspected of erroneously being introduced into the system through import or copy paste operations. These line items usually need to be updated or removed entirely.
- Line items that are discontinuous are suspected of erroneously being generated from time shifts or an accidental zero demand for a quarter in the middle of otherwise continuous work. Typically work activities should be planned and done in continuous blocks.
- Line items with zero demand throughout fundamentally do not make sense. Having a line item implies that some work needs to be done, which is then contradicted by having no FTE assigned to do that work. Upon further investigation there are two potential causes for this anomaly. First, this may be an artefact in the data that needs to be deleted, originally generated for various

reasons. Alternatively, this is a placeholder created for later update with meaningful values but may have been missed by the end user.

3.3.2 Statistical anomalies: Catching mathematical deviants

Statistical – Outliers are detected based on a statistical analysis of the calculated attributes within their relative clustered groups. Over time as the data changes (becoming more and more accurate), the calculation for outliers updates, leading to improved and tighter outlier detection. Similarly, if entries once thought of as outliers become the new normal and enough of these are accepted by users as 'okay' in the system, the system will automatically adjust to no longer flag these as outliers. In this way, the tool adapts over time.

For this tool, the outlier definition uses the standard box-plot definition: outliers are considered to be values at a distance exceeding one and a half times the interquartile range (IQR) above the 75th percentile or below the 25th percentile. The IQR is defined as the distance between the 75th percentile to the 25th percentile.

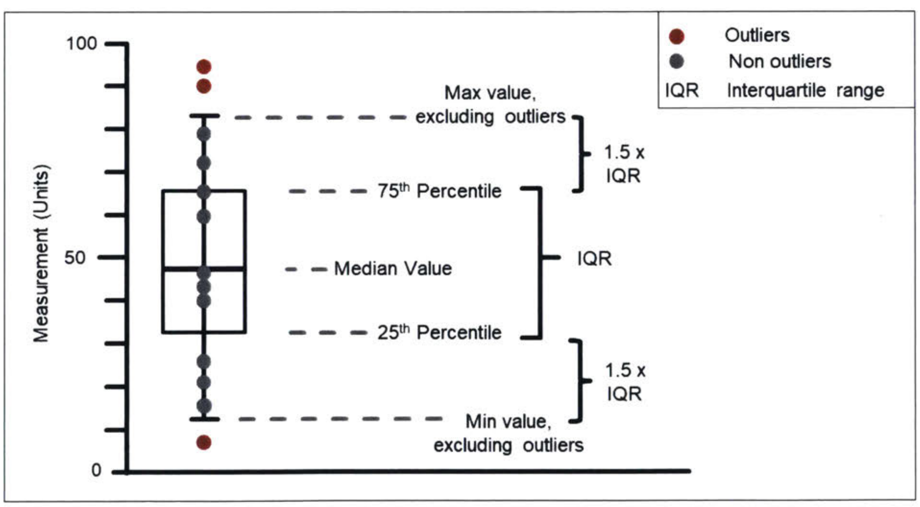


Figure 6. The boxplot method for defining outliers.

The calculated attributes considered in this detection step are Duration, Total Resources and Maximum Resources across the quarterly FTE resource allocation data. The cluster grouping used for attribute sample comparison is according to the unique combination of 'Component', 'POD', 'Function', and 'Sub-Function for each line item of data. As an example of why this grouping is important, one could imagine that the same activity called "regulatory filing documentation preparation" across different function or sub-functions may have different measures of what would be considered a normal duration but should be relatively standard within this granular grouping.

3.3.3 Relational Anomalies: Machine learned relationships

Using a kind of machine learning it is possible to discern (across the drug development projects) certain relationship rules that exist between specific activities. This machine learning refers to the iterative process of comparing multiple data elements to one another in a brute force and statistical way to detect rules across the massive amount of data we are analyzing. By changing hyper-parameters, such as the threshold for rule detection, the sensitivity of this approach can be adjusted. The initial relationships detected by the tool include the 'existence' relationship and the four project management relationships: start-start, start-finish, finish-start, and finish-finish.

- Existence – Activities are checked to see if they typically occur together across different projects.
- Start-start – Activities are checked to see if they typically start together.
- Start-finish – Activities are checked to see if the first activity typically starts as the second activity finishes.
- Finish-start – Activities are checked to see if the first activity typically finishes as the second activity starts.
- Finish-finish – Activities are checked to see if they typically start together.

The threshold used to decide if a rule should be established was set at 95% (nineteen in twenty). That is to say, if a specific relationship is detected 95% of the time, it is deemed strong enough to establish a rule. **The lack of a rule being established does not signify that a relationship does not exist; rather it signifies that the rule is not strong enough that we would be worried when it is not satisfied.**

An example with numbers: "Pre-pivotal Cell Line Development" and "Pre-pivotal Bioprocess Development" occurred ('existed') together 35 times out of 35 project instances of each activity, thus a bidirectional existence rule was established. Here every time "Pre-pivotal Cell Line Development" occurred (35 times), so too "Pre-pivotal Bioprocess Development" occurred (and visa versa).

In a different situation, one might see unidirectional existence relationships signifying that if activity X occurs, we expect activity Y to occur too, but if activity Y occurs we do <u>not</u> necessarily expect that activity X needs to occur. For example, here, imagine Y occurs 35 times, each time with a corresponding X in the same project, creating an IF X THEN Y relationship. However, X occurs 70 times with a corresponding Y occurring still only 35 times, thus IF Y <u>NOT Necessarily</u> X. In this case, a unidirectional relationship would exist.

Building on the above example, a case of a relationship non-detection (neither unidirectional or bidirectional relationships) is as follows. "Pre-pivotal Cell Line Development" and "Pre-pivotal Bioprocess Development" started together 29 times out of 35 times each, thus although somewhat common, the mutual occurrences were not strong enough for a start-start relationship rule to be established.

3.4 USER INTERFACES: The UI and UX design

An important accomplishment of this work is the roll out of the tool to the live environment where end users can interact and benefit from the tool. Users have two primary ways of interacting with the tool: "in process" and "batch review". Additionally, a "Metrics" screen provides measurements of the tools performance.

3.4.1 "In process" interaction: convenience is key

Until now, users such as FAs have developed established processes of doing their work. To aid users in a streamlined way the tool incorporates an additional field for each line simply stating whether this line item has been assigned a flag of having potential errors. It is important to emphasize that the users would otherwise already be looking at this screen and these fields during their work and this intervention has been designed to be minimally disruptive and convenient

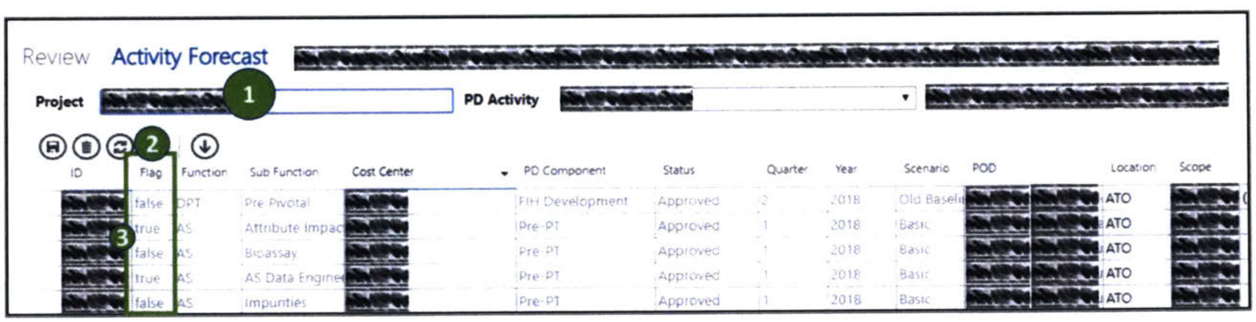


Figure 7. A typical review of forecast data in FSM. (1) The user selects the drug project and receive the corresponding forecast line items. (2) A new 'Flag' column has been added. True signifies the line item has been flagged for having a potential error, signifying the users review is needed. (3) By clicking on the associated ID number, a dashboard with contextual information will pop up to help the user evaluate the flag and take an action.

At this point the user has been made aware of a potential issue(s) for certain activities for the chosen drug and the user can deep dive into this information by clicking on the ID number of the line item to bring up a contextual dashboard, including relevant metadata, calculated fields, relative metrics, and color-coded flags. Using all this information in the dashboard, as well as any outside information the user might have, the user is well equipped to take an action to rectify the flag. The actions available to the user in the FSM system are to "Edit", "Ignore" or "Delete" the flagged line item.

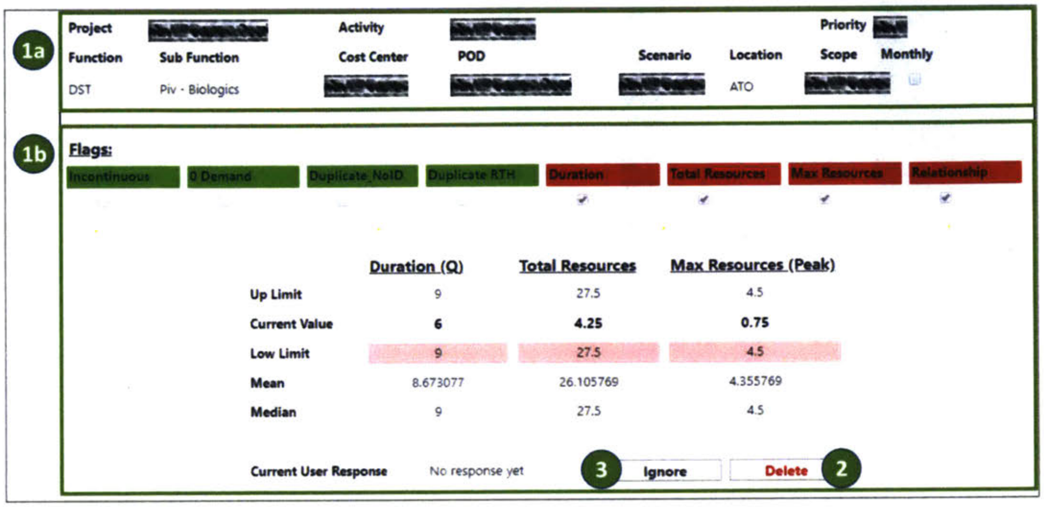
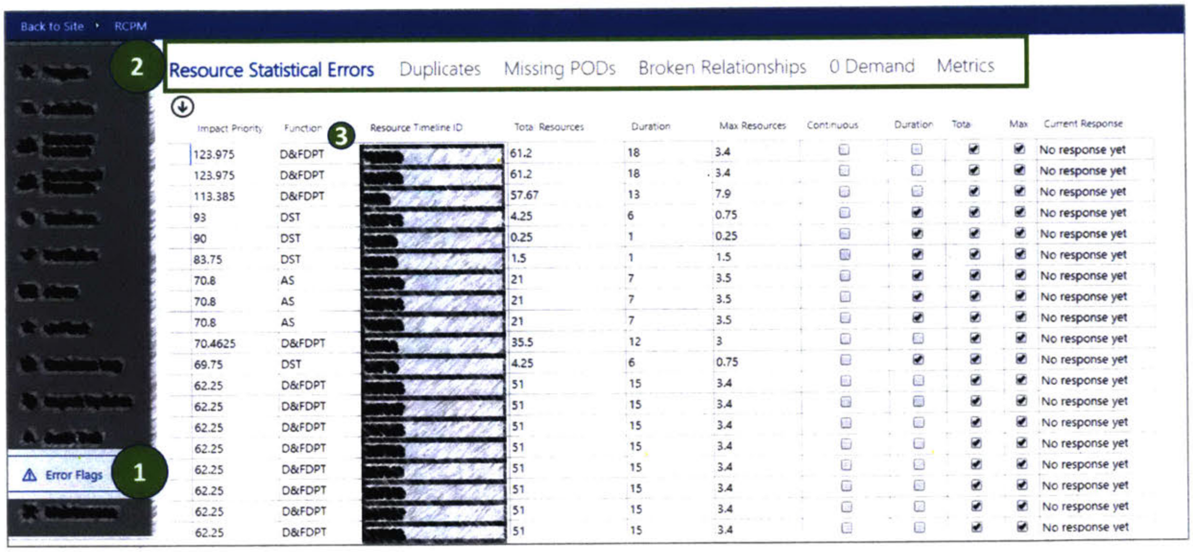


Figure 8. The contextual dashboard. (1) Relevant contextual information is displayed including metadata (1a) and information on the flags (1b). After evaluating the information, the user can take action by editing the line item data, deleting this line item (2) or ignoring this line item's flags (3)

3.4.2 "Batch review" interaction: efficient review

A new "Error Flags" screen has been incorporated into the FSM system, which can be reached by clicking on the tab with this name. Clicking on this option brings up a choice of interfaces that the user can decide to interact with to more efficiently evaluate error flags in the system. By grouping similar errors together, the user can evaluate the common issues more quickly and with the context of other similarly flagged errors. The interfaces include errors batched by:

- Statistical anomalies (duration, total and max all on one tab-screen)
- Logical anomalies (duplicates and zero-demand each on their own tab-screen)
- Relational anomalies (existence on the "Missing PODs" tab and the four project management relationships all on the "Broken Relationships" tab)



Impact Priority	Function	Resource Timeline ID	Total Resources	Duration	Max Resources	Continuous	Duration	Total	Max	Current Response
123.975	D&FDPT	[redacted]	61.2	18	3.4	☐	☐	☑	☑	No response yet
123.975	D&FDPT	[redacted]	61.2	18	3.4	☐	☐	☑	☑	No response yet
113.385	D&FDPT	[redacted]	57.67	13	7.9	☐	☐	☑	☑	No response yet
93	DST	[redacted]	4.25	6	0.75	☐	☑	☑	☑	No response yet
90	DST	[redacted]	0.25	1	0.25	☐	☑	☑	☑	No response yet
83.75	DST	[redacted]	1.5	1	1.5	☐	☑	☑	☑	No response yet
70.8	AS	[redacted]	21	7	3.5	☐	☑	☑	☑	No response yet
70.8	AS	[redacted]	21	7	3.5	☐	☑	☑	☑	No response yet
70.8	AS	[redacted]	21	7	3.5	☐	☑	☑	☑	No response yet
70.4625	D&FDPT	[redacted]	35.5	12	3	☐	☐	☑	☑	No response yet
69.75	DST	[redacted]	4.25	6	0.75	☐	☑	☑	☑	No response yet
62.25	D&FDPT	[redacted]	51	15	3.4	☐	☐	☑	☑	No response yet
62.25	D&FDPT	[redacted]	51	15	3.4	☐	☐	☑	☑	No response yet
62.25	D&FDPT	[redacted]	51	15	3.4	☐	☐	☑	☑	No response yet
62.25	D&FDPT	[redacted]	51	15	3.4	☐	☐	☑	☑	No response yet
62.25	D&FDPT	[redacted]	51	15	3.4	☐	☐	☑	☑	No response yet
62.25	D&FDPT	[redacted]	51	15	3.4	☐	☐	☑	☑	No response yet
62.25	D&FDPT	[redacted]	51	15	3.4	☐	☐	☑	☑	No response yet
62.25	D&FDPT	[redacted]	51	15	3.4	☐	☐	☑	☑	No response yet

Figure 9. The batch review screens in the Error Flags tab. (1) The new Error Flags tab. (2) Errors are batched by type and can be viewed on their corresponding screens. (3) The Function column is especially useful for filtering, allowing the FA to view just the errors they own.

Conveniently the tool allows the user to click on a desired ID and once again bring up the same contextual dashboard that we saw in the "in process" interaction. Once again with the relevant information in hand the user can take the appropriate "Edit", "Ignore" or "Delete" action.

3.4.3 User feedback: "Edit", "Ignore" or "Delete"

The three possible actions that a user can take to resolve a flagged line item are "Edit", "Ignore" or "Delete"

- Edit – If the user believes that updates should be made, they would go into the line item and edit it, for example changing or shifting resource allocation.
- Delete – In extreme cases the user can delete the item immediately by clicking the 'Delete' button. A user might delete a line item due to zero demand or duplication for example.

- Ignore – If the user deems this line item as incorrectly flagged, the user has the option to notify the system and prevent this line item from being flagged (for the same reason) in future iterations of the tool. By clicking on 'Ignore' the system effectively learns that this line item is incorrectly flagged, and will prevent it from being flagged in the future. Changes to this line item or if it is flagged for different reasons will cause its ignore status to be reset and it can be brought to the attention of the end user again if need be.

The action taken by the user is automatically logged by the system and used to track metrics on how the tool is performing. The user is also able to add a comment to any line item to provide any additional feedback or contextual information that might aid in the improvement of the tool.

3.4.4 Impact Priority: Focus effort

A system defined "Impact Priority" metric has been incorporated and made available to the user on the various dashboards. This Impact Priority is a relative measure of severity and helps guide the user to the most impactful items first, prioritizing their list of errors. The numeric value of the calculated impact priority does not hold any intrinsic meaning but rather can be used as a relative measure.

The impact priority is calculated as follows:

$$Impact\ Priority = Severity\ \times Deviation$$

Where,

$$Severity = Number\ of\ flags = \sum_{i=1}^{n} F_i\,, \qquad F \in \{0,1\}$$

And,

$$Deviation =$$

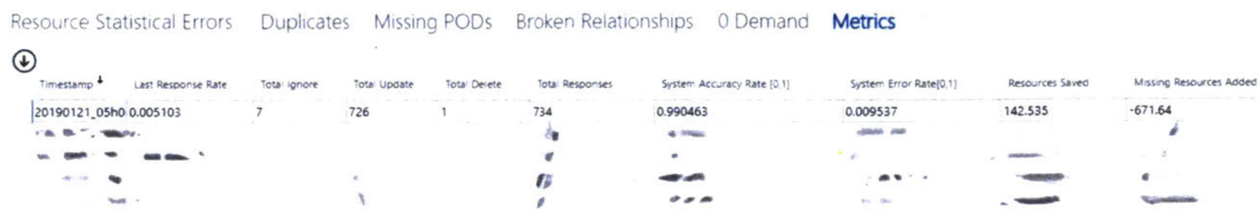
$$= How\ far\ the\ total\ number\ of\ resources\ is\ outside\ the\ statistical\ limits\ for\ that\ cluster\ group$$

$$= \max(0, \quad Total\ Resources - UpLimit, \quad LowLimit - Total\ Resources)$$

The impact priority helps to raise for attention entries that have multiple flags (seemingly very wrong) and grossly deviating from what would be considered as normal.

3.4.5 Tool Metrics: Measuring performance

The final tab on the Error Flags screen is the Metrics Tab. Performance metrics reflected here include the last response rate since the tool ran, totals for how many items were marked as ignores, updated and deleted, and the tool's accuracy and error rates. The metrics also include fulfilled and potential FTE impact that that the tool has detected. Here, 'fulfilled' refers to flagged anomalies that were subsequently updated and 'potential' refers to unresolved flagged anomalies in the system. The impact of the tool for both fulfilled and potential impact are broken down into the resources saved (overstaffed) and missing (understaffed), as these issues do not necessarily net out. Absolute and non-absolute summations are also provided to assist in evaluating the current state (robustness) of the FSM system with its current level of detected anomalies.



Resource Statistical Errors | Duplicates | Missing PODs | Broken Relationships | 0 Demand | Metrics

Timestamp	Last Response Rate	Total Ignore	Total Update	Total Delete	Total Responses	System Accuracy Rate [0,1]	System Error Rate[0,1]	Resources Saved	Missing Resources Added
20190121_05h0	0.005103	7	726	1	734	0.990463	0.009537	142.535	-671.64

Figure 10. The Metrics screen reports on the performance of the tool

3.5 TECHNICAL DETAILS: The tech stack

In this section I will detail some of the technical components of this project that were critical for its success. The technical components chosen here, may be substituted for other similar components by someone working on their own tool and associated problem.

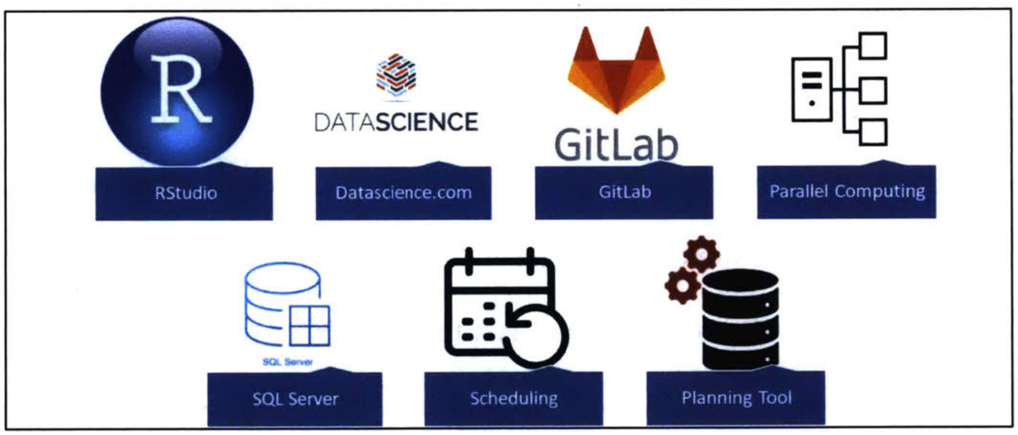


Figure 11. Notable technical components of this work.

3.5.1 Choosing the R programming language

Going into this work, I debated between coding in Python or R, both programming languages with which I had some experience. Ultimately, I went with R, as it is well suited for data analysis and statistical computing, and had the added benefit of R Shiny integration, a package that makes building standalone apps easy. R Shiny was initially appealing, as I believed I would build my own beautiful app but in the end the needs of the project meant that integration into the existing FSM system was a better solution than developing another separate tool for end users to manage. The ease of data manipulation within R was especially helpful during the exploratory phase of this project, especially for gaining an initial understanding of the data. Overall, R was a great fit for this work, but Python could have achieved the same goals. Python is arguably as technically able as R (maybe more) and seems to be better known by data-scientists, including those at Amgen.

3.5.2 GitLab for version control and code hosting

GitLab is a web-based repository management system, allowing for the sharing, tracking and management of changes to files and code throughout the DevOps lifecycle. Amgen has their own internal GitLab system ensuring code is secure and able to interface with other internal Amgen tools and systems.

3.5.3 The datascience.com platform

Oracle's datascience.com offers a powerful enterprise data science platform that enables data-science teams to organize work, access data and computing resources, and run code in the cloud. This platform added value in two main ways: first, it allowed for hosting and running my code from the cloud and executing this code on a scheduled basis. Second, the computing resources available allowed for the ability to spin up virtual machines much more powerful than my laptop, facilitating parallel-computing, which was extremely valuable in the development and implementation of this work. More on these two benefits below. While datascience.com was extremely valuable, there are other similar options available such as Databricks.

3.5.4 The benefits of Parallel Computing

With over ten thousand data-lines, with associated quarterly data and metadata, compute times were high for the various logical and statistical checks being done but through using matrix operations compute times were still manageable. Once relational rules, which were brute force iterative comparisons, were implemented compute times skyrocket. This could be a killer for development, as waiting hours between each code edit to test the results is not an option. For the implemented code, this is also problematic as minimizing downtime is a priority.

With datascience.com's on demand cloud-computing option I was able to spin up powerful virtual machines with multiple cores (up to 40) to run my code at an accelerated pace. Implementing parallel-computing meant recoding some of my work according to parallel principles to be able to benefit from the additional cores processing in parallel but it was well worth it. The benefits from parallel computing saved me 100s of computing hours over the course of this work. For an illustrative example, the implemented code now runs in a little over 5 minutes as opposed to a little under 3 hours without parallel computing.

3.5.5 Scheduling for automation

With the desire to leave Amgen with a tool fully in place and continuing to deliver benefits, I knew I needed to find a solution to the question of who would take the steps to run the tool, especially with all the other responsibilities they might have. The solution was clear: This project was about automation, so the running of this tool also needed to be automated. The ability to schedule the running of code from within datascience.com was critical for this automation. This scheduling not only runs the code but also spins up the needed computing power on demand according to schedule and then relinquishes it after, which is important for cost management. The tool is now scheduled to run on Sunday evenings at 11:59pm to minimize down time.

3.5.6 Database interfacing with SQL Server and Planning Tool

FSM is currently built on Planning Tool, an internal Database Management System (DBMS), although there are potential plans to migrate from this solution in the future. In order to interact with underlying data, I needed to install and use SQL Server Open Database Connectivity (ODBC) drivers. These drivers were important for creating connections to pull and push data from the database, especially from within the datascience.com Linux environment.

4 Results and Discussion

In this section I discuss the primary results of this work including going live with the tool, using change management to encourage user adoption and the statistics around the tools performance since going live a few months ago.

4.1 IMPLEMENTATION AND GOING LIVE: The tool comes to life

A major success of this work is that the created tool was implemented and incorporated into the business. More than a theoretical exercise, this tool is in use and adding value to Amgen.

The tool went live in September 2018 and has been scheduled to run automatically from a cloud server weekly on Sunday nights at 11:59pm. The time was chosen to minimize down time of FSM. Using parallel computing the tool runs in a little over 5 minutes (as opposed to a little under 3 hours without parallel computing).

4.2 ENCOURAGING USER ADOPTION: Change management is key

Creating a tool for users is not enough if it remains unused. To this goal, I have worked with a change management mindset throughout my time at Amgen both to facilitate the creation of something useful and to encourage users to interact with and thus reap the benefits of the implemented tool.

4.2.1 Development phase

In an effort to increase user adoption of the tool, steps were taken throughout the development phase of the tool. During the design phase, meetings were held with both management and end-users from across functions to align on the purpose of the tool and how it could be used. Meetings with management facilitated their buy-in as well as cooperation from their teams. Understanding how users went about their existing work was paramount for establishing how best to include the tool and its output in their workflow. Specifically understanding that an additional and external application would likely be rejected, and therefore integration into the existing FSM system would be key.

4.2.2 Rollout

In preparation for the rollout training materials and quick-guides were prepared. A hands-on, in person training session was conducted for the users at the end of August. The session went through the theory and applicable changes to the system, complete with examples of how to interact. Time was allocated for the users to begin interacting with the system and to ask questions. Overall users were impressed but hesitant to commit to using the tool.

4.2.3 Ongoing

To increase engagement and awareness, once the tool was live I began sending updates on the performance metrics of the tool every few weeks, including the user interaction information, tool accuracy and impact. I also included modest targets to try to influence the users to interact with the tool more for example: "Please take an action to each check at least 5 FLAGGED line items this week". Included in these email updates was a quick-guide presentation, designed to lower the barrier to using the tool by including a refresher on the main points of navigating and using the tool.

4.3 TOOL PERFORMANCE: Initial impact is impressive

As of January 2019, the live tool has performed with an accuracy of 99.22% (an error rate of ~1%). The business impact of the accurately detected anomalies is 60.48 FTE years saved and 250.71 FTEs missing resources added. Potential savings and missing resources detected in the system (yet to be reviewed) are 853.65 FTEs and 67.32 FTEs respectively. Using internal Amgen estimates the resulting impact of the corrected resources is $77.798M optimized so far. Similarly, potential impact currently detected by the tool and awaiting user feedback is $230.243M (an average of ~$18M/year over the 13 years analyzed).

As of January 2019, the live tool has been run approximately 20 times. Each time the tool runs, it analyzes approximately ten-thousand line items (described in 3.2.1 Input Data). Initially the tool flagged potential issues in about 35% of line-items, and over time with user input and changes this number has slowly decreased (currently 32%), improving trust in the data.

4.4 MEASURING IMPACT: Discussion on the performance calculations

4.4.1 "True wins" vs "soft wins"

The primary role of this tool is to detect errors and flag them for end users, who should then consider the flag and other provided information to take action (3.4.3 User feedback: "Edit", "Ignore" or "Delete"). If user indeed took an action that they would not have taken otherwise (or taken that action earlier) due to interaction with the tool this would be a "true win". Measuring a true win is difficult and cumbersome for end users, and thus "soft wins" are instead considered.

A soft win occurs when the tool correctly identifies an error, signified by the user taking an appropriate action to update the flagged field, however one is unsure if the user was indeed helped by the tool or if the user made their updates regardless of the tools help. The reason this might be challenging, is because the tool may be passively racking up wins, unfairly taking credit for changes and the resulting FTE impact. With this challenge in mind, Amgen colleagues and I feel comfortable counting soft wins because 1) The did indeed identify an error correctly and 2) even if not considered, input from the tool should have been used to expedite and assist in the update process.

The remaining asterisk here is that wins will be overemphasized over losses (incorrectly flagged items) when user interaction with the tool is low. To address this issue, I am supporting Amgen in its effort to survey current user interaction and encourage more interaction with the tool moving forward. This asterisk may explain the very high (~99%) accuracy rate of the tool to date, and it would not be surprising if this accuracy fell somewhat as interaction increased.

4.4.2 FTE deviation: Missing and excess resources

When considering the impact of a win, a simple delta calculation is performed between the total number of resources before and after the correction. This change in number of resources represents how many FTEs were saved or added. This means that although all wins are considered for the accuracy score, FTE impact is only considered when the total number of resources changes. For example, if one deletes a zero-demand error or make a time shift based on a relationship flag, this would not be represented in the FTE impact calculations. Thus, it is difficult to capture the full (non-FTE) impact of the tool on improved planning and forecasts.

When considering the potential non-optimized resources remaining in the system a more nuanced approach was used. The calculation calculates the deviance of the current total number of resources from what would be considered (normal) within the limits i.e. the change needed to no longer be considered an outlier. This measurement effectively only capture the potential FTE impact of line-items flagged due to total resource outliers. Similar to above but even more so, this understates the potential impact of these errors.

4.4.3 Imperfect but useful metrics

Despite not being perfect, the metrics calculated are useful, providing approximations for both accuracy and FTE deviation. The accuracy represented in the system is an upper bound and will become more and more valuable as user interaction increases. The FTE deviation is a lower bound, helping us assess impact in FTEs and later in optimized dollars.

This page intentionally left blank.

5 Literature Review

Inspiration for much of this work was found in the literature. Below are some of the pieces that most shaped my approach to this work, especially with regard to anomaly detection.

There are a number of methods for conducting anomaly detection. In our data we are dealing with multiple unlabeled and potentially changing sets, so a supervised approach would be not be feasible. Knorr et al. [1] highlights the importance of identifying distance based outliers for data-mining and Petrovskiy [2] surveys outlier detection algorithms in data mining. Domingues et al. [3] conducts a comparative evaluation of unsurpervised machine learning outlier detection algorithms. Ahmed et al. [4] describes the need for clustering for dealing with contextual anomalies. Angiulli and Fassetti [5] use induction from examples to assist with identifying domain specific outliers. Loureiro et.al [6] illustrates how outlier detection can be used on real world data, flagging suspicious entries for further human review, saving time. Mansur and Sap [7] address the importance of statistics and potential weaknesses with various approaches. Moumena and Guessoum [8] use boxplots for fast anomaly detection in real world data. Beck et al. [9] speaks to the value and the importance of automating data integrity, with an example from the pharma industry.

This page intentionally left blank.

6 Recommendations and Contributions

In this section, I will outline recommendations for future work and actionable next steps that will continue to grow the value and success of this work for Amgen and PD. I will go on to describe the major contributions achieved by this work, which include the key takeaways for today's business leaders. I discuss these topics covering the narrower viewpoint of the tool and the data, but also a more general paradigm shift for the world and expanding research.

6.1 FUTURE WORK: Opportunities to expand impact and research

With the success of this work, several opportunities to further this work are in discussion. There is always the opportunity for incremental benefits to the existing tool such as improving user interaction, work processes, code efficiency and maintenance, and interpretability of results. More exciting however is expanding the work done here to new frontiers in terms of scope or research possibilities.

6.1.1 Expanding within FSM data

The easiest and most natural step for expansion is to expand the data set upon which the created tool acts. Currently the dataset acting as input for the tool is restricted to large molecule (LM) drug projects in the pipeline. This is just a subset of the total data in FSM. By expanding the data scope, the same benefits can be realized across a larger set of data, creating more accurate forecast planning for Amgen.

The path forward would need to consider implications of introducing data that may have its own definitions of normal for the various attributes and relationships. In principle, a similar grouping mechanism could occur at a higher level (drug-type) to ensure differentiation, but opportunities for combining data sets should also be investigated.

With the framework and methodology in place, this future work would be considered as a tactical execution and less of continued research.

6.1.2 Activity tracker, a cousin of FSM

Within Amgen opportunities have arisen to expand this work to include "Activity Tracker" data. Activity Tracker has a similar data structure connecting work activities to projects but instead of forecasting, Activity Tracker records actuals of work performed. An anomaly detection tool similar to the one created here could be used to detect anomalies in actual work performed. Flagged instances of these deviations could then be brought to a human to conduct a root cause analysis, learning lessons and improving efficiency for the future or perhaps preventing further complications and delays in real time.

Another opportunity would be to combine Activity Tracker data with FSM to detect where Amgen forecasts do not align with actuals. In this way, forecasts could become more accurate and reliable. Similarly, actual work done that doesn't match forecasts, could once again be flagged so that leadership can be made aware of potential impacts on drug development timelines in real time.

6.1.3 Anomaly detection, more general applications

Data robustness or accuracy is a common issue across industries and functions. The techniques and approach described herein serve as a basis for creating anomaly detection tools for anyone concerned with data integrity. Some of the anomaly detection approaches can be used to complement existing detection of abnormal behavior. Possible examples within and outside of Amgen include: predictive maintenance, penetration detection in cyber security, sales forecasting, process monitoring. etc.

6.1.4 Machine learned project management relationships, a new frontier for research

The machine learned relationships that this tool generated were very exciting. In an effort to contextualize the machine learned relationships, an inspection was conducted with Amgen's project management. Many of the detected relationships made sense but there were also those that were unexpected. Further investigation is needed to understand if this methodology could be used to detect all known relationships and furthermore possibly detect unknown relationships and dependencies that might exist across functions, or work activities that traditionally have not been considered connected. Detection of these new dependencies may provide interesting insights into project activity relationships and dependencies, which could prevent delays from black-swan type events or encourage leadership to consider other organizational structures or improved communication between relevant groups.

There are simple tweaks and tests, which can be investigated. For example, tweaking the threshold for new rules, or adjusting the clustering level for considering which activities should be compared to one another. There are also more complicated but fascinating areas of research to be explored.

The types of machine learned relationships tested here are very narrow. Only five relationships were considered (3.3.3 Relational Anomalies: Machine learned relationships) but there are many more possibilities. Each activity has been compared to each other activity, but further dimensions could be explored for example if two activities, X and Y exist, do we expect Z to also exist?

Higher dimensions can also be explored, similarly. Activities were tested for project management activities such as two activities starting together. This should be expanded for checks of fixed staggered starts or windows of time for example if two activities X and Y always start within three quarters of one another, we would want to investigate in the case of activities starting six quarters apart. There are many other possible relationships that come to mind, including negator relationships, such as if X exists we expect Y to not exist. There is an abundance of research to be explored in this area.

Once machine learned relationships have been defined and detected, there are additional opportunities. Machine learned detected relationships could also be used to autogenerate project management pert charts or timeline schedules. In one such application, I envision schedule timelines being constructed automatically from operational data (for example in the airline industry) for understanding bottlenecks or for competitive intelligence.

This entire work on relationship detection flips the conventional approach of using what we know to define rules and ensure compliance in operations. Instead here the goal is use existing data (operational or otherwise) to learn the rules and relationships that exist in the underlying data for research purposes. Of course, once detected these rules can easily be used to detect and enforce compliance as was done in this work.

6.2 PD AND AMGEN RECOMMENDATIONS: The "to-do" list

As seen in the above future work, opportunities to further benefit from this effort are plentiful and possibly overwhelming. Here I will outline specific actionable recommendations for Amgen and PD.

6.2.1 Ensure User interaction

The most important recommendation is for PD to encourage user interaction with the tool and to gather feedback from end users. This is of course important, because if unused the value of the tool goes unrealized. It is also important because user interaction and feedback will help define what has been done well and what needs to be improved. These lessons learned can be carried forward for future iterations and new applications of the tool and its developmental approach.

To encourage user interaction several actions should be taken:

- Leadership sponsored surveys should be sent out periodically gauging the interaction with the tool, perceived performance and value of the tool, and soliciting feedback on the positives and negatives of the tool. A sample survey is included in the appendix (Figure 12). These results should be collected and reported to both leadership and end users to ensure alignment on and engagement with the tool. Sequential survey results can be compared monitor tool engagement trends and for responsive action.
- Training session and office hours should be conducted to refresh existing or teach new users how to best interact with the tool. Similarly, users should be assigned an expert user who they can turn to with questions or for coaching.
- Leadership should set and track defined goals for use of the tool. One suggested goal is that each end user (FA) review at least five flags weekly. This modest goal can be completed in about a minute and will ensure flags are continuously being dealt with, leading the data and the system to improve. An added benefit of this weekly goal is that users will remain familiar with the tool and its benefits if they have some sort of minimal interaction on a regular basis.

6.2.2 Develop Expertise for maintenance and updates

One of the accomplishments of this work is the fact that the created tool was implemented in such a way that it could continue to run and add value own its own. Having said that, it would not be unexpected for other Amgen systems to undergo changes that might necessitate updates to, or maintenance of the tool. Additionally, user requested, or leadership specified changes may also dictate changes needed to the system. It will be important for Amgen's PD to develop the necessary technical expertise to carry out maintenance of and updates to the tool as needed.

6.2.3 Move to expand this work internally

With lessons learned from user interaction and development of the necessary technical expertise, Amgen will be ready to explore options for expanding this work and gaining further benefits. Some examples of these expansions are discussed above (6.1 FUTURE WORK: Opportunities to expand impact and research). The relatively quick wins will be expanding the tool's scope of FSM data and exploring integration of Activity Tracker data. Already today conversations can begin with various parts of the company to share this success story. These conversations will help see where and how this success can be replicated in other areas. The exchange of knowledge may provide valuable insights to help improve the tool developed here.

6.3 MAJOR CONTRIBUTIONS: Changing Amgen, changing the world

Beyond the impressive accomplishments with regard to accuracy, error detection and recovery of non-optimized dollars (4.3 TOOL PERFORMANCE: Initial impact is impressive), there are several overarching accomplishments of this work.

6.3.1 Showed that AI technologies can have enormous impact on resource planning

With advanced technologies such as AI and ML making their way into industry, companies such as Amgen are looking for ways to capitalize on this opportunity. This project serves as a roadmap on how individuals and teams can use today's available infrastructure and open source software, to implement and benefit from AI and ML in potentially any part of the business. This work provides a framework for how (even) individuals can identify an issue and then roll out their own grassroots tactical AI and ML projects across Amgen (and other companies). More broadly data science teams, which Amgen is already creating, can work to roll out similar and even more expansive projects across the organization, using this work as a proof of concept. This work illustrates how a modular approach, utilizing parallel and cloud computing can efficiently add value to an organization.

6.3.2 Demonstrated the power of IA (Intelligent Augmentation) to complement AI

Globally, leaders and employees rightfully debate the future of the workforce as artificial intelligence creeps into current human performed tasks. **Shifting the paradigm from AI to IA, this project demonstrates a symbiotic approach and the power of human-computer collaboration to achieve something that was not possible before**. There are jobs that people perform better and there are jobs that computers perform better. It is our responsibility to create tools, such as the one created here, to aid humans in their work.

Sifting through thousands of lines and hundreds of thousands of data points is not feasible for us humans, but with intelligent augmentation, we can use advanced technologies to focus our attention and then do what we do best (and what the machine is unable to do alone): make decisions based on an array of, possibly untracked, contextual information. Understanding how best to prepare the world for the upcoming changes that automation and artificial intelligence will bring is of a global priority and the IA paradigm, along with change management, is an important part of the solution.

6.3.3 Implemented an end-to-end solution that will save time and money

This work created an end-to-end solution for Amgen's PD helping operations and solving a very real problem, namely improving data accuracy that drives decision making for millions of dollars of spend annually. The end-to-end nature of this work, means that the tool is currently in place providing ongoing benefit to Amgen, without demanding additional (possibly unavailable) support. Looking at this another way, this work has real impact beyond a theoretical study that otherwise could have been filed away at the end of the project, essentially disappearing. As true for all Amgen tools and systems, when Amgen evolves e.g. increasing the scope of this work, changing other systems, etc. Amgen may need to decide if they will continue to allocate the support needed to keep this effort alive or even to expand it.

7 Summary

Advanced technologies such as AI and ML have begun to trickle down to industry, yet many organizations and employees are unsure of how to handle the associated benefits and risks. Using the paradigm of IA (intelligent augmentation), this paper illustrated the successful development of a tool created to empower employees by focusing their attention and saving them time. The tool handles the human-impossible task of sifting through thousands of lines and hundreds of thousands of data points to flag potential issues for the human end-user. The human end-user is then able to leverage the provided flags and contextual information needed to make decisions and take appropriate action. The initial results and metrics reviewed are very promising and tell the story of a company continuously improving their data.

This symbiotic IA mindset was integral for developing and rolling out a tool that would be used by the organization and that would leave a lasting impact. The approach and methodology for rolling out similar data science tools for Amgen and other companies has been discussed in detail. It is my hope that the lessons learned and shared in this paper will serve to guide and inspire others in the abundant, continued work needed to incorporate these advanced technologies across the world.

As a student in business and engineering, I strive to bridge worlds. It seems clear to me that the future belongs to those who can successfully combine the skills of both machines and humans into something more than the sum of their parts. Some people believe machines will soon take over as our overlords, others believe that general purpose AI is impossible and that humans are best. My belief is that we need to appreciate and enhance the aspects of both human and machine. To do more than we can imagine today, because…

Sometimes, we are better together.

This page intentionally left blank.

GLOSSARY

Table 2. Glossary of terms

Term	Explanation
AI	Artificial Intelligence
CMC-LCM	chemistry, manufacturing and controls, life cycle management
DBMS	Database Management System
DevOps	Development and Operations
FA	Functional Analyst (anonymized name)
FSM	Forecast System Management (anonymized name)
FTE	Full-time equivalent / Full-time employee
IA	Intelligent Augmentation
LGO	Leaders for Global Operations
LM	Large Molecule
NPV	Net Present Value
ODBC	Open Database Connectivity
PD	Process Development
R&D	Research and Development
SME	Subject Matter Expert
SVP	Senior Vice President
TL; DR	Too Long; Didn't Read
UI	User Interface
UX	User Experience
VP	Vice President

This page intentionally left blank.

REFERENCES

[1] E. M. Knorr, R. T. Ng, and V. Tucakov, “Distance-based outliers: algorithms and applications.,” *VLDB J. Int. J. Very Large Data Bases*, vol. 8, no. 3/4, p. 237, Feb. 2000.

[2] M. Petrovskiy, “Outlier Detection Algorithms in Data Mining Systems,” vol. 29, Jan. 2003.

[3] R. Domingues, M. Filippone, P. Michiardi, and J. Zouaoui, “A comparative evaluation of outlier detection algorithms: Experiments and analyses,” *Pattern Recognit.*, vol. 74, pp. 406–421, Feb. 2018.

[4] M. Ahmed, A. N. Mahmood, and M. R. Islam, “A survey of anomaly detection techniques in financial domain,” *Future Gener. Comput. Syst.*, vol. 55, pp. 278–288, Feb. 2016.

[5] F. Angiulli f. angiulli@dimes. unical. i. and F. Fassetti f. fassetti@dimes. unical. i., “Exploiting domain knowledge to detect outliers.,” *Data Min. Knowl. Discov.*, vol. 28, no. 2, pp. 519–568, Mar. 2014.

[6] A. Loureiro, L. Torgo, and C. Soares, “Outlier Detection Using Clustering Methods: a Data Cleaning Application,” in *In Proceedings of the Data Mining for Business Workshop*, 2004.

[7] Mansur and Noor Md. Sap, “Outlier-Detection-Technique-in-Data-Mining-A-Research-Perspective.” 2005.

[8] A. Moumena and A. Guessoum, “Fast anomaly detection using Boxplot rule for multivariate data in cooperative wideband cognitive radio in the presence of jammer.,” *Secur. Commun. Netw.*, vol. 8, no. 2, pp. 212–219, Jan. 2015.

[9] B. Beck bernd. beck@boehringer-ingelheim. co., D. Seeliger, and J. Kriegl, “The impact of data integrity on decision making in early lead discovery.,” *J. Comput. Aided Mol. Des.*, vol. 29, no. 9, pp. 911–921, Sep. 2015.

This page intentionally left blank.

LIST OF FIGURES

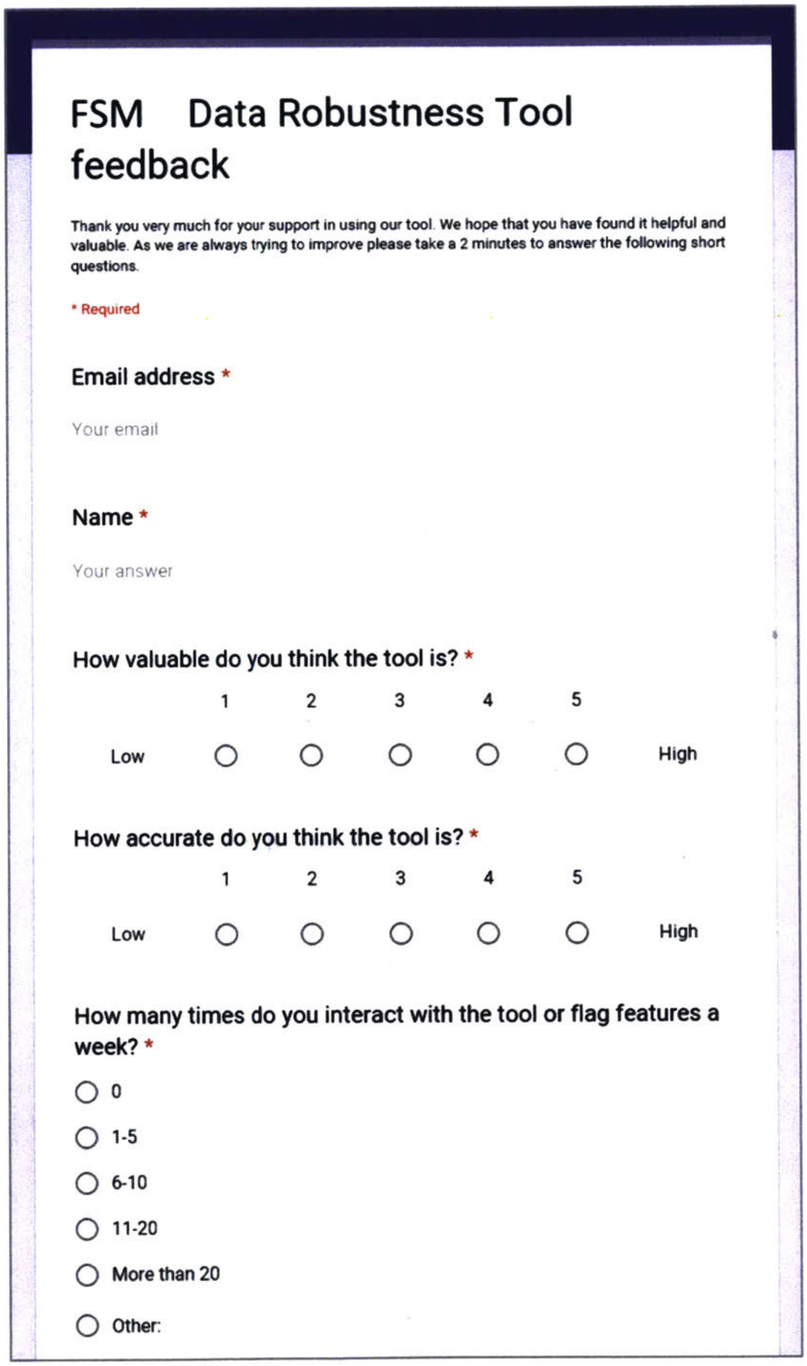
This page intentionally left blank.

LIST OF TABLES

This page intentionally left blank.

APPENDIX

Figure 12. Sample feedback survey (image 1 of 2)



FSM Data Robustness Tool feedback

Thank you very much for your support in using our tool. We hope that you have found it helpful and valuable. As we are always trying to improve please take a 2 minutes to answer the following short questions.

* Required

Email address *

Your email

Name *

Your answer

How valuable do you think the tool is? *

	1	2	3	4	5	
Low	○	○	○	○	○	High

How accurate do you think the tool is? *

	1	2	3	4	5	
Low	○	○	○	○	○	High

How many times do you interact with the tool or flag features a week? *

- ○ 0
- ○ 1-5
- ○ 6-10
- ○ 11-20
- ○ More than 20
- ○ Other:

Figure 13. Sample feedback survey continued (image 2 of 2)

How many flags do you review weekly? *

- ○ 0
- ○ 1-5
- ○ 6-10
- ○ 11-20
- ○ More than 20
- ○ Other:

What is preventing you from reviewing more flags? *

- ☐ I don't have time
- ☐ I don't know how
- ☐ I don't find the tool valuable
- ☐ I have reviewed all my flags
- ☐ I don't interact with RCPM
- ☐ Other:

What do you find valuable in the tool? *

Your answer

What do you wish could be improved with the tool? *

Your answer

Any other feedback?

Your answer

SUBMIT

Page 1 of 1