Paper SP07

Mine the Gap: How the Role of Data Scientist Fills a Need in the Pharmaceutical Industry

Michael S Rimler, FMD K&L, Cincinnati, Ohio, US
Jorine Putter, Grünenthal GmbH, Aachen, Germany

SUMMARY OF PAPER

Attend any industry conference, pharmaceutical or otherwise, and you are bound to hear the new buzzwords and catch phrases driving innovative thought at that time. In the new world of 'big data', one such expression that continues to float to the top is 'data science' or the 'data scientist'. Although this field/role may be well defined in industries such as marketing and finance, it has not yet found a home in the pharmaceutical industry ('pharma'), and least outside of the largest pharmaceutical companies.

The objective of this paper is to present ideas on how the data scientist can make positive enterprise-wide contributions to small- to mid-sized pharmaceutical companies, whether directly employed, contracted via a CRO vendor, or even as an independent contractor. One of the key characteristics of the data scientist in other industries is that one person is performing all data science activities. We propose that the pharma either develop this "one person" skillset or create multifaceted teams to perform the analyses. Without this advancement, we don't believe that pharma can truly overcome the universal paradigm shift across industries and leverage it to be fit for the future. For the purposes of this paper, we will refer to the combination of traditional analyses in pharma and applications of data science as "pharma data science".

We aim to motivate members of the pharmaceutical industry to think about how best we can achieve this and ensure that we stay as current as possible in a very fast changing "data and technology" world. We will discuss the skills and knowledge that a pharma data scientist would ideally have, as well as the types of cross-functional projects that leverage the pharma data scientist's expertise. For example, the pharma data scientist may search for unidentified efficacious subpopulations, analyze preclinical expression pattern detection, or undergo data mining and clustering analyses.

SECTION 1: WHAT IS DATA SCIENCE?

If the data scientist practices data science, what exactly is the field of *data science*? Wikipedia defines *data science* as an "interdisciplinary field about scientific methods, processes, and systems to extract knowledge or insights from data in various forms, either structured or unstructured, similar to data mining."[1] In the early days of data science, the typical analyses were the standard descriptive and inferential statistical analyses, with some effort into predictive analyses. The types of analyses were limited, as was the computing power available to perform them. Today, the analyses available to process data has expanded rapidly, in parallel with significant improvements in processing power, allowing for more computationally intensive analyses to be performed. Contemporary complex analyses include data mining techniques, machine learning, deep learning, neural networks, and cluster analyses.

In an article from January 2017, data scientist Raj Bandyopadhyay offers a framework for breaking down the data science process.[2] We find this process to be quite helpful, both in identifying the skills required of the successful pharma data scientist, as well as dissecting potential projects that some might undertake. We will apply this framework to selected examples of pharma specific problems that could be addressed with data science techniques. The steps outlined by Bandyopadhyay are as follows:

1. Frame the problem
2. Collect the raw data needed to solve the problem
3. Process the data
4. Explore the data
5. Perform in-depth analysis
6. Communicate results of the analysis

In section 5, we apply this framework to example projects that the pharma data scientist might undertake, to understand more fully the scope of work required of the role.

SECTION 1.1: ACADEMIC TRAINING FOR DATA SCIENCE

Many major universities now offer Master's degrees in data science or related fields such as business analytics, including Columbia, NYU, Northwestern, Georgia Tech, Syracuse, UC Berkeley, and Texas A&M. Even institutions that do not offer a master's degree proper, may still offer a curriculum that results in a certification of completion. The curricula of the certificate programs have less requirements, but still cover the core topics of a typical data science program. In addition, some universities, e.g., NYU and Yale, offer doctoral level degrees in data science.

The curricula in the master's programs are varied, but all required or elective coursework typically falls into a few general categories. After the new student completes primer courses such as overviews of data science or practical introductions to computer programming and/or statistics, the degree coursework can generally be characterized into statistical analyses, courses related to computer science, and algorithm courses for data science. Coursework related to statistical analyses include topics such as Probability and Statistics for Data Science, Statistical Inference and Modeling, Predictive Analytics, Time Series and Forecasting, Logistic Regression, Linear Algebra, Survival Analysis, and Optimization. Coursework related to computer science includes advanced programming courses in SAS, Excel, R, Python, SQL, Enterprise Miner, Hadoop, Tableau, Java, Visual Analytics, Google, Spark, Pig, or Hive. It also includes systems courses such as Computer Systems for Data Science, Cloud Computing, Parallel Computer Architecture and Programming, Machine Learning, and Deep Learning. Finally, the algorithm courses for data science include courses such as Algorithms for Data Science, Data Mining, Text Mining, Big Data, and Data Visualization. In addition, students may be able to specialize their degree through elective courses such as Database Systems, Econometrics, Bioinformatics, Natural Language and Processing, Deep Learning in Semantics, Optimization Based Data Analysis, and Optimization and Computer Linear Algebra. Most programs end with a capstone project, geared toward preparing the graduate for applying their coursework to real-world industry challenges.

There are 3 primary threads of master's degrees in data science. Some programs focus on one thread (e.g., business analytics), whereas others will offer all three. The ability to focus may either be implicit through the student's ability to choose elective courses, or explicit through the choice of a particular track of study offered by the program. For example, Georgia Tech offers all three tracks, as described below:[3]

- **Business:** Understanding the practice of using analytics in business and industry. How to understand, frame, and solve problems in marketing, operations, finance, management of information technology, human resources, and accounting in order to develop and execute analytics projects within businesses.
- **Computational:** Understanding the practice of dealing with so-called "big data". How to acquire, preprocess, store, manage, analyze, and visualize data arriving at high volume, velocity, and variety.
- **Analytical:** Understanding the quantitative methodology of descriptive, predictive, and prescriptive analytics. How to select, build, solve, and analyze models using methodology such as parametric and non-parametric statistics, regression, forecasting, data mining, machine learning, optimization, stochastics, and simulation.

For the aspiring data scientist, there are plenty of well-structured academic programs which provide training fundamental to the requirements of the role. Although these programs equip the student with the tools to analyze a problem, the successful data scientist must also be able to initially frame the problem, interpret the results, and read them back to stakeholders with non-technical language.

SECTION 1.2: DATA SCIENCE ORGANIZATIONS AND CONFERENCES

In addition to academic programs for data science, there are numerous organizations dedicated to the practice and advancement of data science and analytics. A simple internet search for 'data science conference' results in multiple sites listing the 'best' data science conferences that are a 'must' for any practitioner. We've selected six conferences that tend to appear on multiple 'top' lists:

Table 1. Data Science Conferences

Data Science Conference	URL
Strata	https://conferences.oreilly.com/strata
Predictive Analytics World (PAW)	https://www.predictiveanalyticsworld.com
KDD (Knowledge Discovery and Data Mining)	http://www.kdd.org
The Data Science Conference	https://www.thedatascienceconference.com
IEEE International Conference on Data Mining	http://icdm2018.org
Open Data Science Conference	https://odsc.com

There are many others including (but definitely not limited to) Data Science Pop-up, MLConf, Enterprise Data World, useR!, and Joint Statistical Meetings.

Strata is presented by O'Reilly and Cloudera and was formerly referred to as Strata + Hadoop World. The conference claims to be 'where cutting-edge science and new business fundamentals intersect—and merge.' Predictive Analytics World (PAW)

has separate conferences that focus on business, government, financial, manufacturing, and healthcare. PAW refers to themselves as 'the leading cross-vendor conference series covering the commercial deployment of machine learning and predictive analytics.' In 2018, PAW is hosting a 'Mega-PAW' event, featuring concurrent sessions from each area. Incidentally, Mega-PAW is being held June 3-7, 2018 in Las Vegas, coincident with PhUSE US Connect 2018. Although not directly related to pharmaceutical drug development, a few of the sessions held in the Healthcare track at Mega-PAW 2018 include

- From Big Data to Big Impact: Redefining Type 2 Diabetes
- Using Machine Learning to Detect Errors in Medical Data
- Reducing Hospitalization with Predictive Modeling in Patients with Kidney Failure

KDD (Knowledge Discovery and Data mining) self-describes as 'the premier interdisciplinary conference bringing together researchers and practitioners from data science, data mining, knowledge discovery, large-scale data analytics, and big data.' The Open Data Science Conference is a self-purported 'safe-space', free of vendors, sponsors, and recruiters. Each of these conferences are targeted toward the field of data science broadly, and hence few provide a focus toward healthcare or pharmaceuticals.

The Innovation Enterprise[4] appears to be the primary source for the intersection of pharmaceutical research and data science. In 2018, Innovation Enterprise is hosting summits for Big Data and Analytics for Pharma in London (March), Philadelphia (June), San Francisco (October), and New York (December).

Key themes from the 2018 London Summit included:

- New Commercial Model of Pharma
- Localization of Pharma
- Drug Research and Development
- Pre-clinical Discovery and Prevention

The London summit included presentations from Novartis, GSK, Deloitte, AstraZeneca, Johnson & Johnson, and Shire Pharmaceuticals, to name a few. Topics ranged from the general ("Bayesian machine learning: when you need to know that you don't know", AstraZeneca) to applied ("How Data From Audits, Investigations and Inspections May be Analysed to Facilitate Optimisation of the Performance of Clinical Trials", Johnson & Johnson).

Key themes for the 2018 Philadelphia Summit include:

- Patient-Centricity - utilizing all available data, often from a variety of different/disparate sources, in order to personalize treatment to the patient
- Use of Real World Data/Real World Evidence - data to identify gaps in patient care, improving quality of care and patient outcomes
- Real World Data Usage in Clinical Trials & what trials may look like in the future
- Data Opportunities for Commercial Analytics – customer targeting/segmentation and customer engagement

The growing presence of conferences that intersect data science and pharmaceutical R&D indicate the value that data science can bring to the pharmaceutical industry. Although we primarily see large pharma engaging in such activities, the mid-size pharma company may still see a positive return on investment from building out a data science group. Furthermore, CROs (or even independent contractors) may also be able to serve small- to mid-sized pharma companies with specialized groups that support analyzing the questions which data science is best positioned to address.

SECTION 2: WHAT IS THE DATA SCIENTIST?

The data scientist is conventionally understood to be someone that elicits insights from data using highly technical methods, ultimately presenting them to the layperson in an understandable way. They often excel in analytical fields such as mathematics, programming, and computer science. The data scientist must exhibit a level of curiosity to identify and investigate problems that may not have been previously considered. They must also be creative in generating solutions using data from a variety of sources and formats. Finally, they must be able to take sophisticated algorithms, which may be processing billions or trillions of data points, tease out the relevant results, and present it to the lay person in a meaningful and insightful way. This ability to think and work technically, but communicate non-technically, is a key characteristic of the successful data scientist. The closest analogy within the pharmaceutical industry is the SAS programmer's transition from hands-on programming to leading and managing projects, but this still fails to capture the full extent of the data scientist's role.

The data scientist will pose questions, devise solutions employing the necessary analytical tools, interpret the results, tease out the noise, and present it in language that is easily digestible to the non-technical stakeholder. The work is often cross-functional, cross-platform, and pulls insights from data in a seemingly unconventional way. Therefore, the typical statistical programmer or biostatistician within a pharmaceutical company or CRO may require additional training to fully function in the data scientist role.

The job market is seemingly overflowing with opportunities for data scientists in general. Quick searches for 'data scientist' on job listings sites such as Indeed.com or Glassdoor.com yield a variety of roles across many industries such as finance, insurance, and technology. One such search, additionally including the term *pharmaceutical*, resulted more than a handful of job listings, mostly from large pharmaceutical drug developers. Roughly half fit the description of the true data scientist. The other half represented role descriptions that seemed like the traditional data manager role, but with the title data scientist. However, there is an indication that pharmaceutical companies, at least larger ones, are beginning to capitalize on the value of the data scientist's capabilities.

SECTION 3: WHERE DOES DATA SCIENCE FIT IN THE PHARMACEUTICAL INDUSTRY?

In the current era of big data and the Internet of Things (IoT), the amount of data available for analysis has increased exponentially. Using any kind of data and eliminating the noise surrounding insights is a daunting challenge. Organizations such as Google and Facebook have dedicated groups focused on developing and optimizing different data integration methodologies, technologies, and storage mechanisms for speed and quality. The increased growth and traction that data mining, machine learning, artificial intelligence, and blockchain technologies are gaining in many industries is an indication that this is the right time to get these technologies into our daily lives within pharma.

Traditionally, pharma has utilized data scientists, which mainly consisted of biostatisticians and biostatistical programmers, to ensure that sound statistical principles were applied to clinical trials. The ultimate goal of such endeavors was to gain regulatory approval of their compound based on clinical significance, substantiated by statistical significance. Over time, a need grew to apply statistical principles in the design of clinical trials to generate greater flexibility and apply adaptive designs. Researchers then developed the need to extrapolate known effects to see what inferences could be made within different populations or subgroups. This innovation led to data driven decision making, which ultimately spans almost every corner of preclinical, clinical, and commercial pharmaceutical drug development. In each of these areas, one could argue the merits for applying data science expertise.

Historically in clinical research, it has typically sufficed to design preclinical and clinical trials to answer a limited set of questions (mostly providing either informative and/or confirmatory information) and move to the next steps in the clinical development plan (CDP) answering the next questions. In our current era, we face increased regulatory, ethical, patient and payer pressure to do a smarter, better, and/or cheaper CDP. Thus, the question arises, how can we do this? There are several areas that can be explored by performing data & text mining, predictive modeling, machine learning, and even deep learning. Some areas to explore with these techniques can include mechanisms of action, optimal delivery methods, disease understanding, pathway detection, genetics, benefits-risk profile, operational delivery, patient populations, standard of care, health economics, payer models, market access, marketing, real-world data, social media, digital apps, and wearables. Even with this short list, there exist a plethora of data sources, both structured and non-structured, as well as countless analysis possibilities. The combination of heterogeneous data sources and analysis techniques generates a complexity that may not be well-handled by siloed business units such as Research and Development, Commercial, and Biometrics. Furthermore, often these divisions may be further divided into early, late, and mature divisions.

The role of data science in the pharmaceutical industry is to answer questions related to business analytics, efficient and economical drug development, or improving patient outcomes through discovering new efficacious and safe administrations. Data science can provide fast, fit-for-purpose, easily understandable knowledge/insights to questions spanning multiple data source(s), by applying advanced and highly sophisticated data integration, exploitation, and analysis methodologies without losing the traceability of the data to the insight. Data science can also provide business insights based on pattern detection within vast amounts of data that is otherwise too large or too complex to visualize.

SECTION 4: WHAT DOES THE PHARMACEUTICAL DATA SCIENTIST LOOK LIKE?

At the high level, the pharmaceutical data scientist does not look much different from data scientists in other industries. The successful pharma data scientist will possess the ability to formulate a scientific problem, translate it into a hypothesis, and then the hypothesis into unbiased analysis. Like other industries, the successful pharma data scientist must be capable with various programming languages (R, Python, etc.) and understand various computing environments (cloud analytical services, hosted environments, etc.) and different kinds of databases and data access layers (linked graph databased, NoSQL, relational databases, unstructured data, streaming data, etc.). They must also have a solid foundation of analytical methods required to keep the problems feasible (e.g., data integration, data visualization).

However, the successful *pharmaceutical* data scientist must also have knowledge of the pharmaceutical industry and available data sources to keep the problems meaningful to their stakeholders. As well, they will be able to effectively translate the analysis into scientific terms, i.e. know how to tell the "story" of the results. Perhaps the most distinct difference from other industries is pharma's requirement for validated, unbiased, good clinical practice (ICH GCP including patient/subject safety, data integrity and data traceability), which may not be as stringently regulated in other industries. Another difference is the required understanding of detailed data privacy rules and regulations, specific to regulatory authority, with regards to

patient/subject individual level data. This, in turn, leads to understanding how to successfully and accurately anonymize/de-identify individual level data points, as well as correctly utilize de-identified individual level patient data.

Currently, large pharmaceutical companies are building out roles pharma data science that encompass the data science work similar to other industries. Perhaps small, single-compound developers do not have a large enough need, but mid-size developers may indeed benefit from the analytical prowess of the pharma data scientist. Indeed, even CROs may begin to staff these as functional areas to serve smaller clients with such analyses, spreading the fixed costs of building out a new department (in terms of staffing, software, and hardware), over multiple projects contracted through many sponsors. The independent pharma data scientist may even be able to serve clients for one-off analyses, using their skills to frame and analyze unique problems as a selling point during the recruitment phase.

If a pharmaceutical company (or CRO) seeks to staff pharma data scientists, they could do so either by promoting from within or hiring from outside company or industry. It is likely that the most efficient method for starting a pharma data science department/program would be to bring in an outsider (even outside the industry), rather than rely on internal training on data science. If you are creating a group, you need a leader with data science experience. There will be plenty of people within the organization that can contribute to the departmental design with firm-specific knowledge and industry knowledge. This knowledge is very important, but to be most effective quickly, data science knowledge should not be relied upon to develop organically. Still, external training for staff will be critical so that you can promote from within as the department grows.

SECTION 5: POTENTIAL PROJECTS FOR THE PHARMACEUTICAL DATA SCIENTIST

Potential projects for the pharmaceutical data scientist include:

- Detection of efficacious subpopulations
- Data Mining & Clustering
- Preclinical expression pattern detection
- Safety signal detection methods
- Model selection or nonparametric methods
- Adaptive study designs

In this section, as a thought exercise, we analyze the first two potential projects by applying Bandyopadhyay's framework.

SECTION 5.1: DETECTION OF EFFICACIOUS SUBPOPULATIONS

Definition

A part of a population that shares one or more additional attributes is called a subpopulation. Different subpopulations may yield different descriptive statistics. Similarly, one may often estimate parameters more accurately if one separates out subpopulations. An integrated analysis could be performed to examine the consistency of the observed treatment effect across individual trials for demographic subpopulations and for other relevant subpopulations. Differences among subpopulations that are consistent across trials and of a meaningful size from a clinical perspective may generate hypotheses that can be tested in future trials. In some cases, where results are persuasive, it is important to describe differences in effects in subpopulations.

Frame the problem

Does my compound show different results in different subpopulations? Which subpopulations are of interest? Do I have enough data to perform a meaningful analysis?

Collect the raw data needed to solve the problem

Identify which individual trials possess the relevant attributes you may want to explore further. These attributes can be demographic factors (e.g., age, sex, race, region, ethnicity, etc.) or predefined/relevant intrinsic or extrinsic factors (e.g., disease severity, prior treatment, concomitant illness, substance abuse, body weight, renal or hepatic function, etc.).

Process the data

If required, prepare pooled datasets containing all information required as defined in your analysis plan.

Explore the data

Perform basic descriptive analysis on your chosen subpopulations to determine whether you possess enough data and/or whether you have adequate power to assess statistical differences. Perform some data mining or exploratory analysis to identify some further potential subpopulations of interest.

Perform in-depth analysis

Populations consisting of subpopulations can be modeled by mixture models, which combine the distributions within subpopulations into an overall population distribution. Even if subpopulations are well-modeled by given simple models, the overall population may be poorly fit by these models – poor fit may be evidence for existence of subpopulations. For example, given two equal subpopulations, both normally distributed, if they have the same standard deviation and different means, the

overall distribution will exhibit low kurtosis relative to a single normal distribution – the means of the subpopulations fall on the shoulders of the overall distribution. If sufficiently separated, these form a bimodal distribution, otherwise it simply has a wide peak. Further, it will exhibit over dispersion relative to a single normal distribution with the given variation. Alternatively, given two subpopulations with the same mean and different standard deviations, the overall population will exhibit high kurtosis, with a sharper peak and heavier tails (and correspondingly shallower shoulders) than a single distribution.[5]

When conducting a pooled analysis, the following analyses could be considered:[6]

- For the larger subpopulations (e.g., sex subgroups), there should be a side-by-side summary and comparison of individual trial results
- For ease of visual comparison, results could be presented as forest plots and as tabular summaries showing point estimates and confidence intervals
- Forest plots can be used to display subset results in large outcome trials and meta-analyses of trials to assess dichotomous cardiovascular endpoints, (i.e., event: yes/no)
- Forest plots can also be used to show subset analyses of pooled data from trials designed to ameliorate the symptoms of a disease or condition
- A quantitative analysis of data pooled across trials generally should be shown
- Consideration should be given to conducting analyses that both include and omit trials that failed to show an effect

Communicate results of the analysis
Differences among subpopulations that are consistent across studies, and of a meaningful size from a clinical perspective results, should be communicated in a layperson manner. Differences that are not consistent and/or not of a meaningful size from a clinical perspective, should be communicated with a layperson statement to clearly indicate that the differences may not be clinically significant.

SECTION 5.2: DATA MINING & CLUSTERING

Definition
Wikipedia defines data mining as “the process of discovering patterns in large data sets involving methods at the intersection of machine learning, statistics, and database systems. It is an essential process where intelligent methods are applied to extract data patterns. It is an interdisciplinary subfield of computer science. The overall goal of the data mining process is to extract information from a data set and transform it into an understandable structure for further use. Aside from the raw analysis step, it involves database and data management aspects, data pre-processing, model and inference considerations, interestingness metrics, complexity considerations, post-processing of discovered structures, visualization, and online updating. Data mining is the analysis step of the "knowledge discovery in databases" process, or KDD.”[7]

Frame the problem
What knowledge can you discover from the database? As per the Cross Industry Standard Process for Data Mining (CRISP-DM), there are six phases to consider when KDD is considered:[8]

1. Business understanding
2. Data understanding
3. Data preparation
4. Modeling
5. Evaluation
6. Deployment

Collect the raw data needed to solve the problem
Using the six steps defined as per CRISP-DM, you need to collect all relevant data with regards to question 1 & 2, business and data understanding, of the database(s) in question.

Process the data
Determine which data preparation steps are required. A typical source for data is a data mart or data warehouse. It is recommended to perform pre-processing and data cleaning activities. Before you can perform data mining activities on a dataset, it needs to be pre-processed to obtain the target set. For optimal results, the target set should be cleaned, thus observations that contain noise or missing values should be removed.

Explore the data
When working with large target sets, this step might require you to take a smaller subset of the data. Once you have a dataset of a size which can be handled easier, start exploring the data to get a better understanding of the data, if not yet present.

Perform in-depth analysis
This section corresponds to the “modeling” section mentioned above.

Cluster analysis

Often exploratory data mining analysis starts with clustering or cluster analysis. Wikipedia defines clustering or cluster analysis as "the task of grouping a set of objects in such a way that objects in the same group (called a cluster) are more similar (in some sense) to each other than to those in other groups (clusters)."[9]

Given that clustering is often approached very different between researchers, cluster analyses differs significantly, and their algorithms more so. Typical cluster models include:[9]

- Connectivity models: for example, hierarchical clustering builds models based on distance connectivity.
- Centroid models: for example, the k-means algorithm represents each cluster by a single mean vector.
- Distribution models: clusters are modeled using statistical distributions, such as multivariate normal distributions used by the expectation-maximization algorithm.
- Density models: for example, DBSCAN and OPTICS defines clusters as connected dense regions in the data space.
- Subspace models: in bi-clustering (also known as co-clustering or two-mode-clustering), clusters are modeled with both cluster members and relevant attributes.
- Group models: some algorithms do not provide a refined model for their results and just provide the grouping information.
- Graph-based models: a clique, that is, a subset of nodes in a graph such that every two nodes in the subset are connected by an edge can be considered as a prototypical form of cluster. Relaxations of the complete connectivity requirement (a fraction of the edges can be missing) are known as quasi-cliques, as in the HCS clustering algorithm.
- Neural models: the most well-known unsupervised neural network is the self-organizing map and these models can usually be characterized as similar to one or more of the above models and including subspace models when neural networks implement a form of Principal Component Analysis or Independent Component Analysis.

Clusters are distinguished as: [9]

- Hard clustering: each object belongs to a cluster or not
- Soft clustering: each object belongs to each cluster to a certain degree (for example, a likelihood of belonging to the cluster)

There are also finer distinctions possible, for example: [9]

- Strict partitioning clustering: each object belongs to exactly one cluster
- Strict partitioning clustering with outliers: objects can also belong to no cluster, and are considered outliers
- Overlapping clustering (also: alternative clustering, multi-view clustering): objects may belong to more than one cluster; usually involving hard clusters
- Hierarchical clustering: objects that belong to a child cluster also belong to the parent cluster
- Subspace clustering: while an overlapping clustering, within a uniquely defined subspace, clusters are not expected to overlap

Other common modeling activities[9]

- Anomaly detection (outlier/change/deviation detection) – The identification of unusual data records, that might be interesting or data errors that require further investigation.
- Association rule learning (dependency modelling) – Searches for relationships between variables. For example, a supermarket might gather data on customer purchasing habits. Using association rule learning, the supermarket can determine which products are frequently bought together and use this information for marketing purposes. This is sometimes referred to as market basket analysis.
- Classification – is the task of generalizing known structure to apply to new data. For example, an e-mail program might attempt to classify an e-mail as "legitimate" or as "spam".
- Regression – attempts to find a function which models the data with the least error that is, for estimating the relationships among data or datasets.
- Summarization – providing a more compact representation of the data set, including visualization and report generation.

Evaluation / Results Validation

An extremely important step during KDD, is the validation of the patterns detected. Often patterns produced from data mining can seem to be significant, however the patterns can't be reproduced on a new target set. One such a problem can be due to overfitting of the model. As a model is often "trained" on a specific training set, it is recommended to run the model on a wider target set or another set of data, validation set, that was not used as a training set. Wikipedia defines a training set as "A

dataset of examples used for learning, that is to fit the parameters (e.g., weights) of, for example, a classifier" and a validation set as "A set of examples used to tune the hyperparameters (i.e. the architecture) of a classifier."[10]

Deployment
Once your model has passed the evaluation step, it is now ready to be deployed on productive data.

Communicate results of the analysis
Once results/patterns have been validated, the results should be communicated in a layperson manner to the business.

SECTION 6: CONCLUSION

We believe the pharma data scientist can make positive enterprise-wide contributions to small- to mid-sized pharmaceutical companies, whether directly employed, contracted via a CRO vendor, or even as an independent contractor. Either by developing the traditional "one person" skillset or creating multifaceted teams, members of the pharma community can ensure that we stay as current as possible in the very fast changing "data and technology" landscape. The pharma data scientist, in many respects, is quite similar to a data scientist in finance or marketing. However, the pharma data scientist must also deliver validated, unbiased analyses based in good clinical practice. As well, they must possess an understanding of detailed data privacy rules and regulations, across various jurisdictions, with regards to patient/subject individual level data.

CONTACT INFORMATION

Your comments and questions are valued and encouraged. Contact the authors at:

Michael S Rimler
FMD K&L Inc.
1300 Virginia Dr., Ste 408
Fort Washington, Pennsylvania, US, 19034
Email: michael.rimler@klserv.com

Jorine Putter
Grünenthal GmbH
Zieglerstr 6
Aachen, 52078, Germany
Email: Jorine.putter@grunenthal.com

REFERENCES

[1] https://en.wikipedia.org/wiki/Data_science

[2] https://medium.springboard.com/the-data-science-process-the-complete-laymans-guide-to-what-a-data-scientist-actually-does-ca3e166b7c67

[3] https://analytics.gatech.edu/curriculum

[4] https://theinnovationenterprise.com/summits/

[5] https://en.wikipedia.org/wiki/Statistical_population

[6] https://www.fda.gov/downloads/drugs/guidances/ucm079803.pdf

[7] https://en.wikipedia.org/wiki/Data_mining#Process

[8] https://en.wikipedia.org/wiki/Cross_Industry_Standard_Process_for_Data_Mining

[9] https://en.wikipedia.org/wiki/Cluster_analysis

[10] https://en.wikipedia.org/wiki/Training,_test,_and_validation_sets