The Data-Centric Lab – A pharmaceutical perspective

Dennis Della Corte[1,*] and Karen A. Della Corte[2]

[1] Department of Physics and Astronomy, Brigham Young University, Provo UT 84602, USA
[2] Public Health Nutrition, Paderborn University, 33098 Paderborn, Germany
Dennis.dellacorte@byu.edu

Abstract
The pharmaceutical industry is on the brink of entering into the digital age, yet still suffers from fundamental misconceptions and outdated IT systems that inhibit its progress. Four key criteria are identified that have enabled labs to reach the post-modern stage, which are insights generation through advanced analytics, automatic communication through machine to machine interfaces, removal of boundaries for an open lab, and novel means of ensuring trust through automatic submissions. Further progress in these four areas will enable the pharmaceutical laboratory to enter the digital age. Unfortunately, historical roadblocks in the form of an application-centric mindset have so far stifled progress. However, initiatives that supported other industries on their path into the digital age are introduced and evidences for the benefits of the digital age are provided. These span from advanced analytics, data-centric architecture, metadata supported communication, knowledge assisted submissions, to digital maturity models. It is concluded that executives and lab staff within Pharma need to transition to a data-centric world view to reap all the benefits of the digital age for faster, better, and cheaper drug development.

Keywords: Data Centricity, FAIR Data, Digitalization, Pharma, Advanced Analytics

1 Introduction

The digitalization of the pharmaceutical industry is expected to enhance customer experience, streamline operations, revolutionize existing business models, and disrupt the entire industry [1]. While most Pharma companies have realized this and launched diverse IT lighthouse projects to drive digitalization, much of the promised benefits have not yet been materialized. Central reasons for the difficulty to derive benefits from digitalization efforts rest in the historical development of the pharmaceutical laboratory. The history of the laboratory can illuminate key areas that marked the transition between different eras of time including the early-modern, modern, and post-modern ages. Understanding these areas can shed light on where Pharma needs to focus today in order to reach the digital age successfully. Further, a historical review can help identify root causes that impede the progress of current initiatives. This review unveils that a focus on specific devices and applications gave rise to an application-centric IT

infrastructure with very limited flexibility and scalability. Once these root causes are understood it will become feasible to investigate the benefits of digitalization in more detail and to provide guidance to shift from an application-centric to a data-centric organization.

This report is structured into five sections. To begin, an extended background section is provided that reviews the development of the lab in a historical context and thereby identifies four key criteria that marked its progress throughout time. Next, the modern-day roadblocks are identified that currently impede progress on the path to the final goal of digitalization within the pharmaceutical laboratory. Following this, a discourse into the tangible benefits of digitalization of the four criteria is offered. Further, the key take-away learnings of other more digitally advanced industries are presented, including how this information can be leveraged to help the pharmaceutical lab. In conclusion, a set off current trends and observations is provided that proposes advised entrance points for Pharma to enter into the digital age.

2 Background

To offer some explanatory background, a laboratory refers to all accommodations in which natural phenomena and processes are explored by means of tools and instruments. Lab venues, experiments, measurements and practices have been defined to exemplify moral values of objectivity, embodying prevailing ideals of a competence meritocracy and transparency through publication while promoting discovery and reward [2]. A lab has never just been a space of knowledge production, it has also always been a place of illustrating, recording and documenting [3].

The key criteria that marked the transition between the aforementioned ages include the following: insights, communication, boundaries, and trust. Insights correspond to the means by which the majority of new knowledge is generated in a certain age. It is the predominant mode of operation that researchers and scientists turned to when they attempted to describe the world. Communication refers to the mode of transmitting insights between individuals and organizations. It corresponds to the preferred choice of conveying and preserving knowledge. Boundaries mark the structural, physical, or imposed delineations that separate the laboratory from the rest of the world. They represent the mental model of what a researcher would call her laboratory. Trust deals with the means by which a laboratory establishes its credibility. It corresponds to the requirements imposed by the general public of a certain age that researchers need to meet in order for their findings to be accepted as validated knowledge. All of these criteria have evolved and need to adapt further in order for the pharmaceutical lab to enter the digital age. Table 1 provides an overview that will be expanded on in the following paragraphs.

Table 1 Overview of four important criteria that mark transition between different ages

CRITERION	EARLY MODERN	MODERN	POST MODERN	DIGITAL AGE
Insights	Observations	Analysis	Meta-Analysis	Advanced analytics
Communication	Oral	Paper	Digital files	Machine-to-machine
Boundaries	Homes of noblemen	National institutes	Global organizations	Open labs
Trust	Honor-based	Peer-review	Regulatory Approvals	Automatic quality checks

At the beginning of the early modern period most insights were generated through tedious and precise observations. A famous example, representative of the time, is the affluent aristocrat Tycho Brahe [4] who recorded observations in privately funded spaces to satisfy his own curiosity. Others, like alchemists who attempted to turn lead into gold for their own gain, would refrain from making any written notes of their observations so as to protect findings that the world today would call intellectual property (IP). This separation between the curious individual and the financially motivated researcher could be called the first instance of laboratories that can be classified as either academic or industrial. At this age, the boundaries of both types of labs were restricted to the homes of those that could afford them. The distinction between academic and industrial labs is important because the different goals and aims mandate different practices in order to secure the trust of the general society in the respective results. Trust in early modern labs was mainly formed depending upon their location and the social moral code that advocated for not bearing false witness and highly valued individual honor [3]. In retrospect, we can identify four core challenges that made it difficult to advance into the modern area. First, the limited scope of insight generation did not yet benefit from the mathematical tools that would allow rigorous analysis of observations. Second, the limited interest to communicate alchemical findings substantially stagnated progress in chemistry. Third, the restricted access to laboratory spaces only allowed an exclusive group to contribute to laboratory work. Fourth, a lack of checks and balances by the general public to ensure the level of proven science caused superstitions to be elevated.

The laboratory moved successfully into the modern area by embracing some fundamental changes of the time. One great revolution that paved the way for the modern era is attributed to Francis Bacon and Robert Boyle, who reinvented communication, as they suggested that all laboratory findings should be recorded on paper in a fashion that allows others to understand and reproduce results [2]. The growing corpus of scientific observations and the arrival of mathematical tools like calculus gave rise to a new mode of insight generation, that is, analysis. The ability to interpolate trends and to make accurate predictions was altering the way scientists thought about nature and the deterministic view of the universe with general laws was being constructed. The ability to not only observe but analyze for correlations and causality was at the root of the major breakthroughs of the modern era. In this age,

laboratories left the homes of the rich and become general practices through the establishment of local academies of sciences. As a result, trust was no longer naturally bestowed on each scientist based on honor, rather the result of a peer review process by expert groups was initiated. This peer review led in many cases to an improvement of the vetting process for new insights. In the modern era, the writing of scientific findings and publication in local academies of science eventually became the norm. With the beginning of industrialization, three specific shifts occurred. First, academic labs became increasingly institutionalized, with the goals to educate citizens, foster high ideals, and generate knowledge and trust through peer review processes [2]. Second, within industry educated citizens found new employment in labs devoted to the creation of patents rather than publications. Third, the lab turned from a site of character realization in the early modern period to a site of establishing scientific facts by demystified professionals. While this era made great progress, certain limitations hindered the laboratories from reaching its full potential. Insights were mainly generated as pen and paper exercises and absence of computers strongly narrowed the ability to compare multiple data sets. The national institutes of sciences did not yet agree to a universal scientific language or vocabulary, which led to error-prone integrations between the dominant languages of Latin, French, German, and English. Finally, industrial organizations had a difficult time protecting their IP because patent rights were just beginning to become established and enforcement on a global scale was difficult to ensure.

The pharmaceutical laboratories entered the post-modern era as they embraced the new technologies of the day. The advent of 'Big Science' measured the amount of electricity a research endeavor consumed. The growing automation of laboratory equipment coupled with computational devices increased the IT footprint of the laboratory drastically. With the ability to connecting findings and observations from different sources it was possible to perform meta-analyses to generate new insights. With the statistical advancements and the advent of meta-analyses, researchers had and continue to have the ability to leverage multiple studies on the same subject by pooling results together in order to derive novel conclusions. For the pharmaceutical lab the evidence of such analyses led to the introduction of rational drug discovery that replaced much of the high throughput screening activities. As computers began to be ubiquitously used, the need to represent insights in electronic formats increasing, giving eventual rise to improved communications through email and the internet. At this time laboratories started to become integral parts of worldwide networks that were no longer defined by walls of buildings. With the emergence of global enterprises and the corresponding desire to market products worldwide, it became necessary for countries to protect their populations from fraudulent products through regulatory approvals. This was especially important for the pharmaceutical market and the drugs it produced. With this sharpened focus on regulatory approval and global marketing, a very clear separation between academic and industrial labs was observed. Instead of publishing articles in scientific journals, the researchers in industrial laboratories were increasingly interested in getting patents recognized so as to have commercial control of the processes and products involved in their research [3]. Within 'Big Science', laboratories became increasingly reliant on electronic devices leading to the introduction of digital files, which in itself revolutionized and disrupted already well-established and complex workflows. The promise of improved efficiency through

automation motivated the introduction of ever-growing numbers of electronic devices. The digital files produced by these devices brought business benefits but also became an obligation that had to be managed and maintained. The broad range of disparate digital file types with rapid technological life cycles made it necessary to migrate or integrate between formats. As a result of the increasing amounts of digital information and the need for management solutions the new and lucrative market of laboratory information technology emerged [5]. Many of the barriers that currently impede the evolution of the laboratory from the post-modern to the digital age are connected to the drivers and motives of this market and the resulting conflicts of interest.

The historical review has detailed how the pharmaceutical lab benefited through the ages by advancing in four key areas. With the inclusion of new methodologies for insight generation, technologies for communication, broadening of boundaries, and novel ways of ensuring trust, Pharma companies were enabled to become global players that could benefit the lives of others on unprecedented scales. Today the world is entering into the next transition phase, moving from the post-modern era into the digital age. Whether this transition will be successful for the pharmaceutical market depends on how well this industry adapts a data-centric business model. Some industries, like those of technology and finance, have already made headway into the digital age, while the Pharma industry currently lags behind. There are specific reasons for this underdevelopment which are unique to Pharma and are elucidated in the next section. The key benefits that the digital age has to offer the Pharma industry include altogether new ways of generating insights by leveraging technologies epitomized under the buzzwords of 'advanced analytics' [6, 7]. Further, the arrival of smart interfaces between IT systems allows for seamless communication of highly complex laboratory procedures, allowing for the reproduction and exchange of information with hitherto unknown efficiency. With mobile technology and an ever-growing pool of publicly available data, labs are no longer defined by boundaries. All data sources should be incorporated to generate prime results. Especially for Pharma, the ability to engender trust through full product lifecycle representations offers opportunities to improve the speed of regulatory approval and reduce the effective time to market. With all these benefits to be won, we ask what is holding the Pharma industry back from entering the digital age?

This paper sets out to clearly identify the current challenges and opportunities of the pharmaceutical industry to transition successfully into the digital age. The current roadblocks of the post-modern laboratory, inherited by its complex history, will be described and a solution proposal presented. Where possible, concrete examples from the pharmaceutical industry, past and present, will be included to make abstract ideas presented tangible to a broader audience.

3 Roadblocks on the path into the digital age

The transition from one age into the next is marked, as presented in Table 1, by the distinct advancement in four key areas. The first being the ability to draw new insights from data, second the ability to communicate scientific results effectively, third the physical boundaries of the laboratories, and fourth the means by which trust is enabled. While other industries, especially the tech and financial industries, have made great

progress on the path towards the digital age, pharmaceutical laboratories are delayed in their progress. Having identified these for key areas it is now possible to investigate the root causes for Pharma's slow progress into the digital age.

When it comes to training, there was a clear disjoint until recently between how IT personnel versus laboratory staffs are trained. Most of the current laboratory heads are excellent chemists and/or biologists, yet rather few of them have strong backgrounds in data science. From their perspective, a laboratory device is a means to produce measurements that can become part of a report. The quality and speed at which these reports are generated by the devices is of highest importance. Vendors of the electronic devices therefore also develop the necessary control software to operate the machines and the analysis software to make the results easily accessible for their purposes. Whether the results from different software programs are readable or interoperable between themselves is not generally relevant to the vendors. In this closed world perspective, a laboratory is equipped with a machine suited for a certain task and leverages whatever software came with it, regardless of how incompatible that software is with other machines in the lab. When multiple discordant devices are taken into operation at the same lab complications start to emerge. Results from different software solutions have to be made comparable and this requires a process of standardization or harmonization. Common solutions for this in the past have been either paper-based records into which specific measurements were recorded, or spread sheet software solutions that could aggregate manually entered information [8]. With the onset of 'Big Science' such time-intensive and tedious processes are certainly not feasible and new solutions are required.

The lacking interoperability of the primary software solutions creates the demand for a second layer of IT complexity, a unifying layer that would be able to reduce manual effort. Many eager start-up companies attempted to deliver electronic notebooks (ELNs) or laboratory information management systems (LIMS) to general labs but the poor quality of software limited the usability in academic labs [8]. The authoritarian structure and the larger budgets within the pharmaceutical area, however, made it possible for academically rejected software to gain a foothold in industry [8]. Through standard operation procedures (SOP) and expensive customization projects the first generations of ELNs and LIMS went productive. As some of these products started to gain market shares a wave of consolidation was kicked off, in which device vendors acquired many successful tech companies to increase their own offerings. This typically resulted in an increased interoperability between the vendors machines and the previously device-agnostic solutions. While this may appear at first sight to be a benefit, it is in fact one of the great dilemmas of the market. After acquiring a specific product, the device vendors frequently neglected the necessary investments to improve the software and rather focused on their strong marketing and sales networks to propagate it through the industry. The ability to integrate different data sources was frequently excluded from the core software functionality and left as an integration task for service solution providers. As a result, most Pharma companies became and are currently burdened by legacy IT systems that have no real support and development and only limited ability to connect to newer devices and machines in the lab. In order to improve the connection between devices and data analysis tools, Pharma companies have launched countless projects to interface and integrate data. The internal resources are often IT specialists with limited understanding and appreciation for laboratory

workflows. External IT consultants seek to deliver best solutions for the data integration problems, but rarely suggest fundamental changes that would solve root causes and reduce future demands. The costly interplay between device vendors, internal IT specialists, and external integration consultants as well as the limited scope of data layers or applications lead to the projected cost of the laboratory equipment service market to reach USD 17.7 billion by 2024 from USD 9.5 million in 2019 [5].

The hard question to face is what fraction of these astronomical figures will actually solve scientific questions of the pharmaceutical lab, like creation of personalized medicine for currently untreatable diseases, and what fraction of this cost will only attempt to solve self-inflicted problems like the interoperability of IT platforms. We predict that the latter will consume the most resources unless a fundamental shift begins now in the mindsets of scientists running the laboratories supported by an executive level strategy that focusses on data-centricity. This necessary mind shift must be accompanied by the acceptance that the digital age will change current business models and workflows fundamentally. With such needed change, digitalization is expected to enhance customer experience, streamline operations, revolutionize existing business models, and disrupt the entire industry [1].

With this background information now highlighted, it is possible to state what changes need to take place and discuss the anticipated challenges. In order to improve insight generation toward the digital age, advanced analytics needs to be empowered. One core driver of advanced analytics is deep learning models, which offer a flavor of machine learning that requires immense amounts of high-quality training data [9]. The application-centric design of the past has been to create inaccessible data silos or unharmonized data lakes that make the tasks of leveraging existing information for the training of machine learn models very difficult. As was recently mused by a Pharma executive, “Data scientists spend 70% of their time cleaning data, and the remaining 30% complaining about this” [10, 11]. The preliminary problem is initially due to the IT layers introduced which are difficult to interoperate. But a more fundamental problem was revealed by the failed data lake projects of the last decade [12]. These projects have shown that even if data is centralized, it does not provide a useful basis for insight generation unless it is communicated correctly. The way digital files were communicated in the past was very tightly connected to the source system producing them. This tight connection created a dependency on source systems to open and understand the digital file. Only when the linkage between source and file is broken and the dataset sufficiently enriched with descriptive metadata can a digital file be meaningful by itself. The progress made in semantic technologies provides a possible solution to this communication challenge. Once the pharmaceutical laboratory agrees on a clearly defined ontological model [13]. it can start to fully leverage machine-to-machine communication through Representational State Transfer Application Programming Interfaces (REST-APIs) [14]. The ability to freely communicate between machines will further enable and create the conditions for an open lab that rapidly interconnects with live data streams from Internet of Things (IoT) devices, smartphones with apps that capture customer information, and global news and trends. The ability to harvest such large pools of information will enable better decision making and improve the time to market. Finally, with improving technological abilities, the pharmaceutical industry is subject to increasing regulatory demands that can efficiently be adhered to if data is represented in a self-sufficient manner [15]. These changes will

not be driven by device vendors, IT startups, integration consultants, or even internal IT departments unless they are motivated by tangible business benefits as expressed by the laboratory owners. Additional insights that enable improved decision-making processes and best practices are shared in the following section.

4 Opportunities and Business Benefits of Entering the Digital Age

The previous sections have identified four key areas in which the pharmaceutical lab needs to progress in order to reap the benefits of the digital age. Here, a more detailed description of these benefits is provided along with tangible recommendations that can support decision makers within Pharma. A final subsection is added to provide a meaningful entry point for Pharma to engage in the cross-industry digitalization dialogue.

4.1 New Insight Generation - Advanced analytics

Current estimates from leading pharmaceutical companies evaluate that 50% of their researchers' time is spent on administrative tasks related to issues arising from inaccessible data. This could be reduced significantly with a better data management strategy in place. The ultimate goal for these companies is to produce better drugs at a faster rate and with the smallest amount of investment [16]. Accessibility of data and the ability to leverage the information contained therein is a key enabler in reaching this goal [17]. The Findable-Accessible-Interoperable-Reusable (FAIR) data principles [18] were designed to support this endeavor and are expected to reduce the number of time-intensive open queries that are frequently encountered due data management issues.

Many decisions have to be made during the drug development process, from target identification, to candidate selection and clinical trial priorities. The availability of additional data during these decision-making processes is expected to improve the attrition rate which is the single greatest problem faced by drug developers. The attrition rate measures the percentage of drug candidates that do not make it to the market due to issues with safety, kinetics, potency, intellectual property protection, or other factors. Today, only about 1 in 8 compounds entering clinical development in the Pharma industry is eventually approved for marketing [16]. Table 2 provides an overview of the costs associated with the drug development process. It takes on average 12 years and costs on approximately $2.6 billion on average from target identification to market approval of a new drug [16]. For neurological drugs this time span is 18 years, with 20-year patent protection rights. The time to generate revenue from a novel drug is therefore severely limited and any improvements in the development costs or speed would increase the number of economically feasible drug candidates and the period of profitability substantially.

Data-driven approaches have had tremendous impact in the past, as can be illustrated by two famous examples. The first example involved the replacement of

large screening efforts in the 1970s by what is coined to be “rational drug design”, the process of leveraging structural information of drug targets and compounds to decrease the experimental research cost [19]. The other example is Lipinski’s ‘rule of five’ [20] which decreased the attrition rate from 40% to 10% in only 10 years by filtering out candidates with poor absorption or permeation behavior through a simple heuristic derived from former experimental studies [16]. These heuristics identified 4 measurements whose outcomes correlate strongly with pharmacokinetic and bioavailability properties of experimental drugs and by enforcing these measurements to come in multiples of five many unfavorable drug candidates could be ruled out [21]. Additionally, a noteworthy trend was the replacement of traditional bench work with in-silico experiments to identify novel drug targets and pathways through deep learning and simulations [9]. Forward looking, data-driven insights and information-guided experiment replacements are expected to deliver similar improvements in the near future for the pharmaceutical companies that master digitalization.

Table 2 Adapted from: SM Paul et al. Nature Reviews: Drug Discovery, 2010. Costs are capitalized based on 11% cost of capital and in 2010 dollars

	Target to Hit	Hit to Lead	Lead Optim.	Non-Clinical	Phase 1	Phase 2	Phase 3	Sub to Launch
Number per Launch	24.3	19.4	14.6	12.4	8.6	4.6	1.6	1.1
P(TS)	80%	75%	85%	69%	54%	34%	70%	91%
Cycle time (yrs)	1.0	1.5	2.0	1.0	1.5	2.5	2.5	1.5
Cost/launch ($mil)	$94	$166	$414	$150	$273	$319	$314	$48

4.2 Clearer Communication – Data-centric architecture

The list of complications stemming from the application-centric architecture in most pharmaceutical companies is lengthy, consisting of legal risks, storage costs for legacy systems, migrations, integrations, search across siloes, and the interpretation of insufficiently described data sets. One of the most prominent problems is the necessary expenditure to maintain and develop communication channels between application focused software solutions. This expense is approximately 35-65% of the entire IT budget [22]. Additionally, the number of failed IT harmonization projects grows faster than the number of productive solutions, with two out of three projects failing according to a recent survey [22]. According to Gartner, 85% of big data projects will not be leverageable because they will continue to reside in silos of technology or location [23]. Potentially the biggest threats to established Pharma are highly agile start-ups and biotechs that are not burdened by a tradition of thousands of data silos that curtail progress. These and other dangers and threats around the application-centric design are well expressed in the so-called Data-Centric Manifesto [22], an online proclamation that has found signatories from major players in Fintech, Oil & Gas, and Pharma. This manifest makes the core statement that the root cause for the messy state of information architecture it the prevailing application-centric mindset that gives applications priority

over data. The remedy for this is a data-centric mindset. Here, data-centricity means literally to place data at the center of your business, no longer allowing it to be an appendage or side project. This means data is more important than the machines that produce it and is managed like an asset [24].

A data-centric architecture has the goal to eliminate the issues described above. To successfully implement a data-driven organization, data must be managed as an asset to increase its value and maintain it over time. Data-centricity is the mentality of putting data in the center of the informatics solution design. The value of data assets is realized by harmonizing and interconnecting data across internal and external data sources and enriching it with contextual metadata and long-term reusable representations. Recent progress in software development has given rise to a new breed of digital integration hub technology. The first of which is now available as a validated solution for Pharma, and many more are expected to enter the market soon. Pharma can benefit from this development and leverage such an off-the-shelf digital hub to support a data-centric architecture strategy. These data hubs typically adhere to the FAIR data principles, which allows any primary or secondary use application such as an ELN, LIMS or AI/ML platform to find and securely access any dataset provided by all data sources. Interoperability and reuse together lower the cost of data integration and eliminate the need for point-by-point data parsing and formatting. This process liberates the data from the application constraints and eases efforts for data reuse and results in a single platform for integration and data sharing. Additionally, such digital integration hubs lead to the elimination of duplicated data, data migration projects, as well as the costs and delays in data extraction, transformation, and loading into AI/ML platforms. Table 3 provides an overview of the benefits that stem from data-centric solutions compared to application-centric problems.

Table 3 Adapted from http://datacentricmanifesto.org/principles

Application-Centric Problems	Data-Centric Solutions
Exorbitant, often prohibitive, cost of change.	Reasonable cost of change.
Data is tied up in applications because applications own data.	Data is an open resource that outlives any given application.
Every new project comes with a big data conversion project.	Every new project taps into existing data stores.
Data exists in wide variety of heterogeneous formats, structures, meaning, and terminology.	Data is globally integrated sharing a common meaning, being exported from a common source into any needed format.
Data integration consumes 35%-65% of IT budget.	Data integration will be nearly free.
Hard or impossible to integrate external data with internal data.	Internal and external data readily integrated.

As an example, Figure 1 illustrates one digital integration hub [25] showing how data is typically loaded into application databases found in application-centric architectures. As a result of this process, data is represented in a proprietary data model as provided by the application vendor. As a Pharma company typically owns multiple data management solutions, this approach will create a set of incompatible data silos that cannot be used for holistic analysis such as product lifecycle management, control and

data-driven regulatory filing without further data migration efforts. Deployment of a digital integration hub can be the backbone of any digitalization strategy and quickly bring the benefits from digitally mature companies to those attempting to catch up.

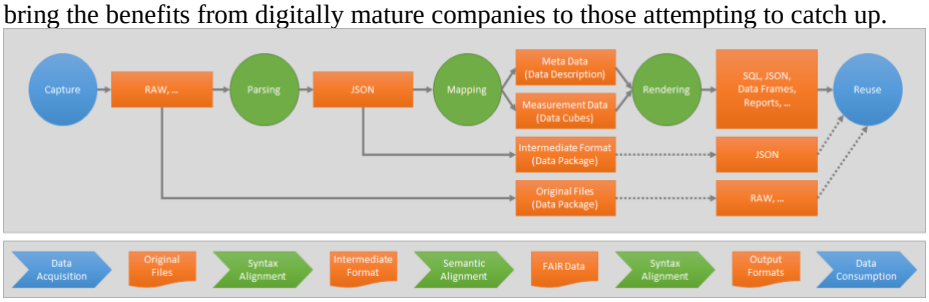

Fig 1. Data flow in digital integration hub to ensure reusability of captured information.

4.3 Remove all boundaries - Improved collaboration

It takes more than a village to develop a successful drug. The tasks involved span from target identification, validation, finding new molecules, screening assays, data on drug-like characteristics, development tools, efficacy measures, to technologies that improve efficiency of trial completion. To enable successful delivery of these tasks, the pharmaceutical industry is supported by a subsidiary market of contract research organizations and manufacturing organizations. Collaboration across disciplines and between preclinical and clinical studies is necessary for a successful drug discovery and development program [16]. The largest challenges in these collaborations are the exchanges of either precise scientific protocols in the form of a method definition or the interpretation of received analytical measurements. The risk of misinterpretation was in the past mitigated by in-person trainings and descriptions that accompany each study. Despite these efforts, common problems still remain such as challenges of ontology mismatches, entity resolution, and correlation confusion. This does not need to be the case and can be avoided because entering the digital age holds the promise of providing formal foundations for service interoperation that address issues such as providing a syntax-independent metamodel and semantics. These new semantic technologies can enable faithful modeling of parallel interactions between multiple parties.[24] Such models make seamless communication without loss of information possible. The creation of the necessary ontologies is an industry-wide endeavor and many opportunities exist to support this important charge [26].

4.4 Trust in the digital age - Regulatory compliance

A common problem of the market-driven industries of today is the pressure to retain the same headcount of employees and to reduce the time that people spend with data, while at the same time the amounts of data and regulations consistently increase. Within the pharmaceutical industry two important tactics can be deployed to solve the conundrum. First, the barriers between data silos have to be opened for machine-to-machine communication and second, all communications must be centered around the

end goal of an electronic submission. While pdf-based electronic Common Technical Documents (eCTDs) [27] and electronic Trial Master Files (eTMFs) [28] have been broadly accepted as document-based electronic submission formats, U.S. Food and Drug Administration (FDA) has recently announced a new initiative that will take electronic submissions to the next level. This initiative is called Knowledge-aided Assessment & Structured Applications (KASA) and focusses on capturing and managing knowledge during the entire lifecycle of a drug product for improved submission quality and speed [15].

While some companies might consider KASA as another regulatory hurdle on the way to market, it also holds some opportunities to reshape the flow of data that may dramatically improve the drug discovery process. Recent IT-driven activities within Pharma have focused on establishing a central data repository or data lake. Unfortunately, there are only limited case studies to be found that suggest true business benefits of the process of collecting massive amounts of data without a governing concept. To some in the industry 'Big Data' is dead [29], but rather than focusing on missed opportunities it is time to redefine 'Good Data'. KASA will be able to support this redefinition as it places similar emphasis on the quality of data according to the FAIR data principles. Of even higher importance, KASA establishes rules and algorithms for risk assessment, control, and communication. It enables computer-aided analyses of applications to compare regulatory standards and quality risks across applications and facilities. It additionally provides a structured assessment that minimizes text-based narratives and summarization of provided information. The goal is that captured knowledge will aide in accessing critical information and will lead to more objective decision making. All of these concepts should not only be developed by FDA, but rather embraced by Pharma, as they will bring direct business benefits as side effects of compliance and faster time to market through optimized review processes.

4.5 Learning from other industries - Digital maturity models

As the pressure on different industries to enter into the digital age has mounted, their resulting experience can be gathered and applied to the pharmaceutical industry [30]. The financial industry has made substantial progress in their business workflows to enable digital age technology. Whole consortia were formed to bring together academic insights and industrial experience to develop Digital Management Maturity Models (DMMM) which can benchmark the digital maturity of a company. Recent history has shown that companies with increased digital maturity outperform laggards in every industry [31]. Being able to access the digital maturity against a DMMM provides detailed insights into which actions an organization should undertake to make headway into the digital age. These maturity models used by other industries capture the cultural, organizational, and technological changes necessary to achieve successful transition towards a data-centric organization. From the various available models (e.g. DAMA-DMBOK, DCAM, CMMI CERT-RMM, IBM DGC Maturity Model, Stanford DG Maturity Model, Gartner's EIM Maturity Model) the Electronic Data Management (EDM) council has the most comprehensive model. EDM's Data management CApability maturity Model (DCAM) was developed together with a quickly growing number of board member companies [32-34]. DCAM based assessment enables the

classification of digital maturity on a scale from 1 to 6, with the biggest gap between level 3 (developmental) and level 4 (defined). Many financial institutes managed to drive their maturity to levels 4 to 6, while most of the pharmaceutical industry still struggles at the conceptual level 2. This struggle is a multifaceted problem for which one core challenge has been identified as the mismatch between system responsibilities of the business functions and IT organizations within large Pharma. The first step in addressing this challenge will be the development of a better understanding of the core competencies and organizational changes, e.g., the introduction of a chief data officer, that are required to take a model like DCAM into operation. The established cross-industry platforms, like EDM, provide the ideal training ground and communication platform to achieve these objectives. It is therefore recommended that Pharma engages on decision maker level with groups like EDM to openly discuss and share the problems and solutions of the industry.

The reports on companies embarking on the journey into the digital age are increasing. McKinsey recently reported that a leading pharmaceutical company enforced a platform strategy that will reduce the number of applications from 4,000 to 1,000 [17]. Further, the formation of cross industry consortia and initiatives like Pistoia [35] and the Allotrope Consortium [36] elevate the challenge of successful digitalization from individual companies to the industry level. First market ready solutions are emerging from these activities, such as method exchange databases and first in class validated enterprise digital integration hubs like ZONTAL Space [25].

5 Conclusion

History has shown that the lab progressed in four key areas from one age to the next. To enter the digital age, the pharmaceutical lab needs to innovate in areas of insight generation, communication, boundaries, and trust. Unfortunately, a legacy of IT systems that stems from the self-serving niche market of laboratory equipment services has placed major roadblocks in the way. However, once the cost of ownership of such an application-centric infrastructure is realized, the pharmaceutical industry can strongly benefit from established digitalization pathways and begin to reap the benefits of entering the digital age. A mind shift is required that places data itself at the center of the value chain instead of the data sources [37]. In this context Pharma can learn from digital maturity models and join established cross-industry working groups.

The field of digitalization spans all industries, therefore some limitations to the scope of this work were necessary. We limited our detailed analysis to the history f the laboratory and current developments in the specific domain of the pharmaceutical laboratory. While some of the presented ideas transcend a specific target group, this study has a rather sharp focus on the audience of decision makers within the pharmaceutical industry. The forward-looking statements with regard to business value of digital activities will need to be sustained by future business cases and proofs of concept.

Powered by a data-centric architecture, the benefits of entering the digital age will impact the ability to make better decisions during the drug discovery process through advanced analytics, increased regulatory compliance, shorter time to market,

and improved internal and external collaboration. The time is ripe for Pharma to take full advantage of the available resources and to enter the digital age.

References

1. Chircu, A.M., E. Sultanow, and L.D. Sözer, *A reference architecture for digitalization in the pharmaceutical industry.* INFORMATIK 2017, 2017.
2. Kohler, R.E., *Lab history: reflections.* Isis, 2008. **99**(4): p. 761-768.
3. Schmidgen, H., *The Laboratory Laboratory.* 2011.
4. Thoren, V.E. and J.R. Christianson, *The lord of Uraniborg: a biography of Tycho Brahe*. 1990: Cambridge University Press.
5. Webpage, *Laboratory Equipment Service Market Report,* https://www.marketsandmarkets.com/Market-Reports/laboratory-equipment-service-market-171213101.html. last accessed 2020/07/14.
6. Barton, D. and D. Court, *Making advanced analytics work for you.* Harvard business review, 2012. **90**(10): p. 78-83.
7. Bose, R., *Advanced analytics: opportunities and challenges.* Industrial Management & Data Systems, 2009.
8. Giles, J., *Lab-management software and electronic notebooks are here—and this time, it's more than just talk.* Nature, 2012. **481**.
9. Rifaioglu, A.S., et al., *Recent applications of deep learning and machine intelligence on in silico drug discovery: methods, tools and databases.* Briefings in bioinformatics, 2019. **20**(5): p. 1878-1912.
10. Suda, B., *2017 Data Science Salary Survey: Tools, Trends, what Pays (and what Doesn't) for Data Professionals*. 2017: O'Reilly Media.
11. Sulo, J., *Data Cleaning and Feature Engineering.*
12. Webpage, *https://www.dataversity.net/is-it-time-to-drain-the-data-lake/* last accessed 2020/07/14.
13. Staab, S. and R. Studer, *Handbook on ontologies*. 2010: Springer Science & Business Media.
14. Masse, M., *REST API Design Rulebook: Designing Consistent RESTful Web Service Interfaces*. 2011: " O'Reilly Media, Inc.".
15. Lawrence, X.Y., et al., *FDA's new pharmaceutical quality initiative: Knowledge-aided assessment & structured applications.* International journal of pharmaceutics: X, 2019. **1**: p. 100010.
16. Mohs, R.C. and N.H. Greig, *Drug discovery and development: Role of basic biological research.* Alzheimer's & Dementia: Translational Research & Clinical Interventions, 2017. **3**(4): p. 651-657.
17. Webpage, *The power of platforms to reshape the business,* https://www.mckinsey.com/business-functions/mckinsey-digital/our-insights/digital-blog/the-power-of-platforms-to-reshape-the-business. last accessed 2020/07/14.
18. Wilkinson, M.D., et al., *The FAIR Guiding Principles for scientific data management and stewardship.* Scientific data, 2016. **3**(1): p. 1-9.
19. Gane, P.J. and P.M. Dean, *Recent advances in structure-based rational drug design.* Current opinion in structural biology, 2000. **10**(4): p. 401-404.
20. Lipinski, C.A., *Lead-and drug-like compounds: the rule-of-five revolution.* Drug Discovery Today: Technologies, 2004. **1**(4): p. 337-341.
21. Lipinski, C.A., et al., *Experimental and computational approaches to estimate solubility and permeability in drug discovery and development settings.* Advanced drug delivery reviews, 1997. **23**(1-3): p. 3-25.

22. Webpage, *http://www.datacentricmanifesto.org/.* last accessed 2020/07/14.
23. Webpage, *https://www.techrepublic.com/article/85-of-big-data-projects-fail-but-your-developers-can-help-yours-succeed/* last accessed 2020/07/14.
24. Hull, R. *Data-Centricity and Services Interoperation*. in *International Conference on Service-Oriented Computing*. 2013. Springer.
25. Della Corte, D., et al., *Library eArchiving with ZONTAL Space and the Allotrope Data Format.* Digital Library Perspectives, 2020.
26. Pachl, C., et al., *Overview of chemical ontologies.* arXiv preprint arXiv:2002.03842, 2020.
27. Cartwright, A.C., *The Electronic Common Technical Document.* International journal of pharmaceutical medicine, 2006. **20**(3): p. 149-158.
28. Roy, K., *Electronic Trial Master Files.* Applied Clinical Trials, 2009: p. 16.
29. Buckee, C., *Improving epidemic surveillance and response: big data is dead, long live big data.* The Lancet Digital Health, 2020. **2**(5): p. e218-e220.
30. Legner, C., T. Pentek, and B. Otto, *Accumulating Design Knowledge with Reference Models: Insights from 12 Years' Research into Data Management.* Journal of the Association for Information Systems, 2020. **21**(3): p. 2.
31. Westerman, G., et al., *The Digital Advantage: How digital leaders outperform their peers in every industry.* MITSloan Management and Capgemini Consulting, MA, 2012. **2**: p. 2-23.
32. Baolong, Y., W. Hong, and Z. Haodong. *Research and application of data management based on Data Management Maturity Model (DMM)*. in *Proceedings of the 2018 10th International Conference on Machine Learning and Computing*. 2018.
33. Hoecker, J., et al., *These Models Need Enterprise Data Management!* Journal of Information Technology Education: Discussion Cases, 2017. **6**(1).
34. Ge, J., et al., *Research on the maturity of big data management capability of intelligent manufacturing enterprise.* Systems Research and Behavioral Science, 2020.
35. Simon Thornber, C.O.D., C.S. Kallesøe, and J. Wise, *The Pistoia Alliance.* The Sequence Service Project. GIT Laboratory Journal, Trends in Drug Disocvery Business, 2011: p. 1-3.
36. Vergis, J.M., et al., *Unlocking the power of data.* LC GC NORTH AMERICA, 2015. **33**(4): p. 270-+.
37. Ratnasamy, S., et al., *Data-centric storage in sensornets with GHT, a geographic hash table.* Mobile networks and applications, 2003. **8**(4): p. 427-442.