**PhUSE EU Connect 2018**

**Paper CD06**

# Disruption Leads to Innovation
# We Must Learn New Things!

Chris Decker, d-Wise, Morrisville, NC USA
Ali Dootson, d-Wise, Manchester, UK

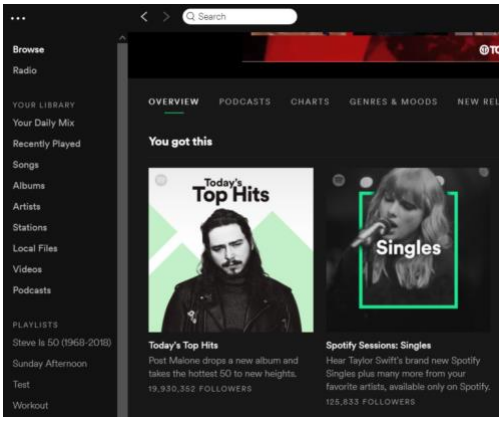
## ABSTRACT

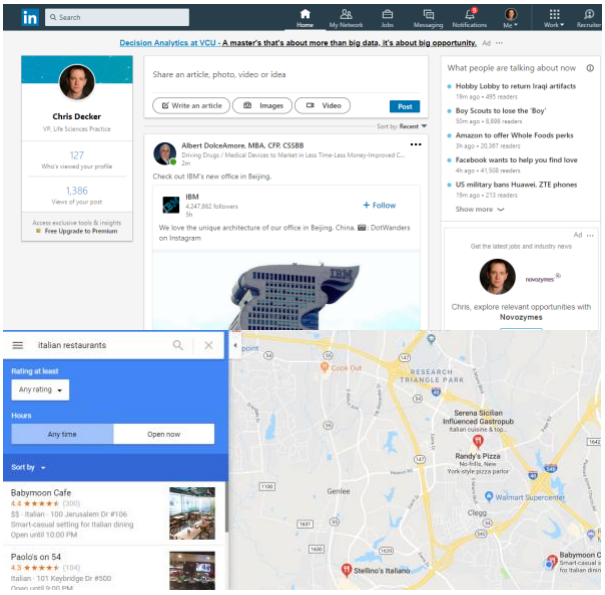
In our personal lives we live in a connected world we take for granted. We are connected through Facebook, Uber, and Alexa with answers at our fingertips. In our clinical trial universe, we still ask people to write standalone subjective specifications, others to write pages of code over many weeks, and yet others to confirm, check, and duplicate just to give us a single number.

If we step back from our what we do today, our current paradigm, how absurd is this? Where could we be in 5 years if we paused and said, "what if…"? or what should I learn in the future? Imagine a world where we have integrity and context embedded within the data, onerous and prescriptive manual steps are replaced by user experience and automation, and users gain insights from the data allowing machines to derive meaning. Does this sound like a bit of fantasy? Well, the reality is that other industries already do it and so can we! Within this presentation, we'll provide you insights into how we could ask Alexa for the answer in the future and what you need to learn to get there.

## INTRODUCTION

In today's world we use various technologies so seamlessly that we take for granted how complex the plumbing is for those solutions. We listen to music from anywhere, find colleagues online, and can locate any kind of restaurant close by.




We use smart phones to text, take selfies, and find information – all of which seem simple. However, under the hood, the neural network of data that provides those capabilities is massive and uses robust technologies that bring that information to your fingertips. We should all be asking ourselves and exploring how we can have this same capability within our clinical data.

This paper will outline the top three roadblocks we face in moving towards a future allowing us to accelerate the work we do, have more confidence in our outputs, and gain true insights from our data. Then we will describe how we might take the steps to reach these lofty goals and initial conceptual ideas on how to get there.

## CURRENT PARADIGM

At a high level, there are three fundamental challenges we face in today's paradigm that stifle our ability to innovate and put a ceiling on our ability to move forward.

### SILOS LIMIT INNOVATION

***"Silos build the wall in people's minds and create the barrier in organizations' hearts." PEARL ZHU***

The first challenge we encounter is more of a people or culture issue than a technology challenge. Within our industry groups of people live in isolated silos producing many manual steps and throw artifacts over the wall with no context or understanding of each other's needs just to get to the end of their process. Today, all the "intelligence" resides in people's heads. People are authoring documents or creating data in silos with no holistic view of the study or program and information is being duplicated and manually transferred from place to place. The wheel is being constantly reinvented. What does this all add up to? – inefficiency, perceived quality that isn't real, extended timelines and an overall façade that you are following a 'standard operating process'.

Many of these silos exist within our ecosystem. Within our own companies we throw things over the wall to each other without enough context. Between regulatory and industry, we create a wall of formality where we are overly cautious about sharing anything. And even with standards organizations the individual teams have evolved from a place of islands leading to disconnected and inconsistent standards.

### BROKEN DATA PARADIGM

***"In the real world, we want to capture the complexity and diversity of the standard of care without forcing it into an artificial structure. The current technology framework used within clinical trials to collect information often makes researchers constrain data to a structure that is not real and has no context…… and forces them into this rigid non-intuitive thought process" – Florence Barkats***

Forty years ago, we started collecting and analyzing data using SAS and defined the SAS data set and transport format as the standard for interoperability and representing clinical information. It was the best choice at the time given what we had available. Unfortunately, it has now become our crutch and severely limits our ability to truly represent our data. The two dimensions below cause us to create more and more data, variables, and standalone siloed specification documents in an attempt to connect the information and put a band aid on a multi-dimensional problem.

*pr.xpt*

| Row | STUDYID | DOMAIN | USUBJID | SPDEVID | PRSEQ | PRLNKID | PRTRT | PRLOC | PRFAST | PRSTDTC |
|---|---|---|---|---|---|---|---|---|---|---|
| 1 | ABC123 | PR | AD01-101 | 22 | 1 | 02 | PET/CT | HEAD | Y | 2012-05-22T09:30:00 |
| 2 | ABC123 | PR | AD01-102 | 22 | 1 | 04 | PET/CT | HEAD | Y | 2012-05-22T08:00:00 |
| 3 | ABC123 | PR | AD01-103 | 44 | 1 | 05 | PET | HEAD | Y | 2012-05-22T09:00:00 |

*ag.xpt*

| Row | STUDYID | DOMAIN | USUBJID | AGSEQ | AGLNKID | AGTRT | AGCAT | AGDOSE | AGDOSEU | AGSTDTC |
|---|---|---|---|---|---|---|---|---|---|---|
| 1 | ABC123 | AG | AD01-101 | 1 | 02 | 18F-Florbetapir | AMYLOID TRACER | 370 | MBq | 2012-05-22T08:40:00 |
| 2 | ABC123 | AG | AD01-102 | 1 | 04 | 11C-PiB | AMYLOID TRACER | 370 | MBq | 2012-05-22T07:20:00 |
| 3 | ABC123 | AG | AD01-103 | 1 | 05 | FDG | GLUCOSE TRACER | 400 | MBq | 2012-05-22T08:30:00 |

*mo.xpt*

| Row | STUDYID | DOMAIN | USUBJID | SPDEVID | MOSEQ | MOLNKID | MOREFID | MOTESTCD | MOTEST | MOORRES | MOORRESU |
|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | ABC123 | MO | AD01-101 | 16 | 1 | 02 | 1234 | VOLUME | Volume | 1.8 | mL |
| 2 | ABC123 | MO | AD01-101 | 16 | 2 | 02 | 1234 | VOLUME | Volume | 1.9 | mL |
| 3 | ABC123 | MO | AD01-101 | 16 | 3 | 02 | 1234 | VOLUME | Volume | 1.7 | mL |
| 4 | ABC123 | MO | AD01-101 | 16 | 4 | 02 | 1234 | THICK | Thickness | 3 | mm |
| 5 | ABC123 | MO | AD01-101 | 16 | 5 | 02 | 1235 | VOLUME | Volume | 864 | mL |

Our two-dimensional data paradigm that we are so tied to is severely limiting and will never provide us the ability to automate and gain insights from our data no matter how many times we try.

We're drowning in an ocean of data without context and overflowing with disconnected documentation. Current data collection does not reflect patient care and is far removed from the source. Data is collected and stored in disparate files, requiring constant monitoring, reviewing, and manual cleaning to ensure medical and scientific integrity. So much time is spent on specifications, data transformations, formatting of results, and validation rather than insight,

leading to prescriptive analysis lacking in scientific curiosity and discovery. We create single use data, analyses, and reports and then put them on a shelf struggling to get answers to simple questions.

#### 20TH CENTURY TECHNOLOGY TO SOLVE 21ST CENTURY PROBLEMS

In our industry, we continue to solve problems today with technology from the last century. What does Windows 1.0, Nokia Talkman, the first Nintendo Entertainment System, and the first Macintosh computer have in common? They were all introduced circa 1985. You know what was also introduced in 1985? – the SAS transport file, a 'technology' that was the foundation of standards development and our core processes for the last 30+ years.

Now, this SAS format is just one example of the technology constraints we have put on ourselves. With the silos described earlier, we have led vendors to build solutions to support those silos, solutions that solve a siloed problem and don't usually integrate with the step before or the step after.

Not only do we not leverage current technology, but we also don't embrace current methodologies like design thinking, user experience, and agile. We pretend to do things in an 'agile' way but often we don't understand what that really means. Think about the app you love to use on your phone…now tell me an app you love to use in your day to day work.

How can we possibly maintain context and relationships in our information with these self-imposed, archaic limitations? We must fundamentally change our current data, people, and technology paradigm to have any chance of significantly helping patients.

## WHERE COULD WE GO?

Let's start by saying that we don't have the golden ticket, the answer to solve world hunger, or the exact path forward to define how we fundamentally change. But we must start somewhere….and below are just a few ideas.

#### BRIDGE THE SILOS

Within our industry we all have our well-defined titles and responsibilities, and in those roles, someone gives you an input, you do your siloed job on that thing, and then throw it over the wall to the next person. What if we were to change all our titles to "Data Scientist", we all had the same title and redirect our focus on doing a task with the mindset of helping the patient.

People use the phrase 'breaking down the silos' when referring to this issue. It's not about breaking down the silos but more about building a bridge; a firm, well architected bridge so that the people who specialize in their niche can do their work and clearly understand what they need from someone and what they need to give someone else.

Functions within Pharma must understand and care what their upstream and downstream customers need, but also must be given the technology and data framework to share that information. Pharma and regulators must not put up walls when discussing challenges…they must share each other's pains, frustrations, and truly collaborate. We have glimpses of this industry collaboration across various organization such as the PhUSE working groups, but we must do more of it.

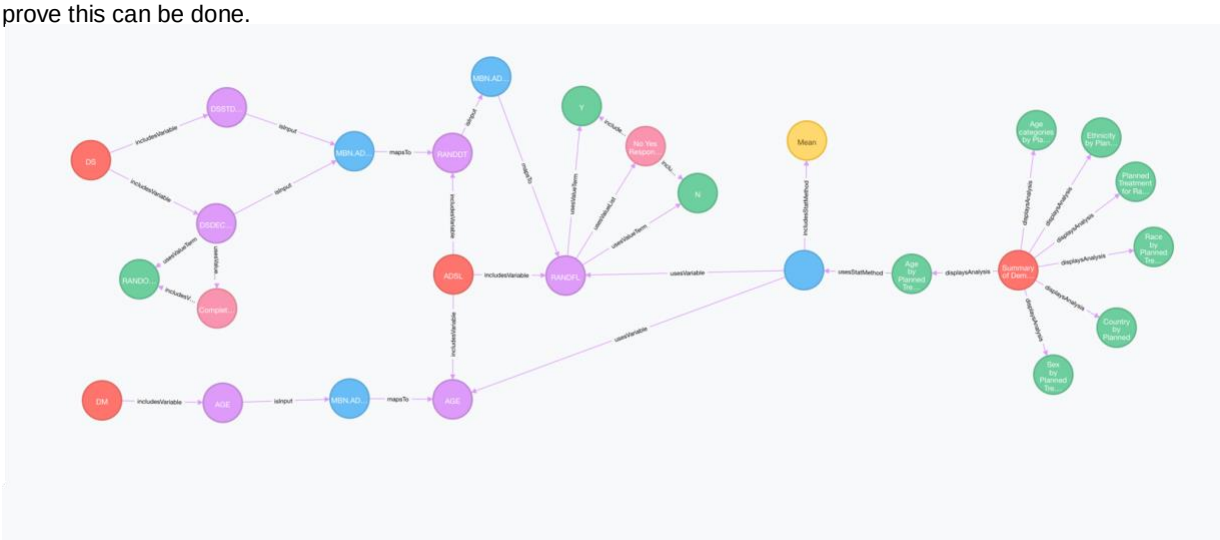
Changing a decades old paradigm won't be easy, but we won't make progress if people keep holding on too tightly to their bit and don't look up to see what is going on around them. The first step is to agree that this is an issue and then clearly define what the bridges could look like.

#### CHANGING THE DATA PARADIGM

In our opinion, the most critical step is to understand that our data paradigm is fundamentally broken. As stated in the quote above from Florence Barkats, we "collect information that constrains the data to a structure that is not real and forces rigid non-intuitive thinking". Currently our paradigm follows a rigid process that removes all context from the data. We collect the data in two dimensions when clinical information is multi-dimensional, and we build isolated specifications that are not really connected to the data.

The first step is for everyone to understand that we must take the data out of two dimensions and co-locate the data and metadata. This is not something that is easy to do overnight but we must have everyone align and understand why this is so important. If we can all align on that concept, then we can start working on the same problem.

The picture below shows an example where we pulled apart a customer's standards and data and linked the information from the analysis backwards to the data collection point with both data and metadata. We were able prove this can be done.




The next step is to define how it can scale and support the variable processes we have within today's workflow. We must move toward a world that has contextual data if we are truly going to automate and gain insights from the data.

**EMBRACE TECHNOLOGY**

The final piece of the puzzle to innovation within our current paradigm forward is to fully embrace that we are squarely stuck in 20th century technology and we must invest time, people, and money into investigating and finding small wins for leveraging today's technology that will allow us to accelerate delivery and gain valuable insights from our data. In the short term, this transition will most likely be painful and we'll most likely be less efficient before we can be more efficient. But it's our commitment to patients that must make us force a path forward.

Recently, I discovered a website online that records my voice by reading 30 sentences. It processes this information, asks me to type in a sentence, and then reads that typed sentence back to me in my own voice. Wow! I know you are asking what this has to do with clinical research, and we won't discuss the ethical views of using artificial intelligence to mimic someone's voice. The key point is not the specific technology, but the fact that this advanced technology exists. Everyone needs to climb out of the box, expand your minds, and think about how technology like this could help us. What if a technology could listen to a physician and populate the EHR which flowed into my clinical data store without ever writing down a word? There are so many possibilities to what we could do if we really embraced the technologies that are available to us today. But let's get practical. Where do we start? How do we introduce these new capabilities while still maintaining a lot of our current process and infrastructure? How do we define a future and then build a bridge to get there?

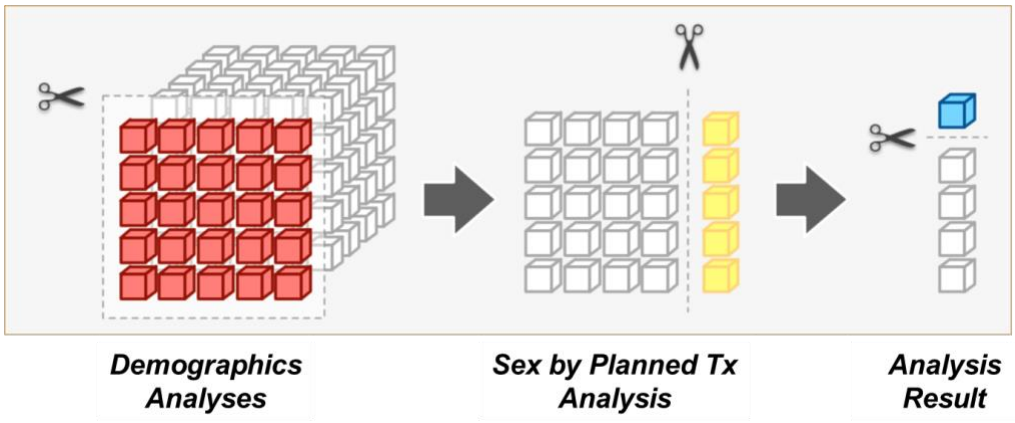
## WHAT IF WE COULD DO THE FOLLOWING?

We must learn from history and use technologies to extract and learn from the plethora of artifacts that exist in our current environments – protocols, specifications, study reports – chocked full of information we can extract, curate, and reuse. We must look at steps in the data flow where we can plug in new capabilities enabling groups of people and making bits of the process more efficient.

We must look at technologies that fix the data paradigm and give us contextual data while still producing and potentially automating our standard artifacts such as SDTM and ADaM in the short term. In parallel, we must engage regulatory agencies to get them on board with submitting data in a structure that provides real traceability and ability to extract knowledge from the data.

**WHAT IF WE COULD…STORE, CURATE, AND REUSE ANALYSIS RESULTS**

What if we could build a repository of analysis results that that can serve a breadth of different stakeholders instead of single use, throw away, paper tables and listings? This use case has been shown within a PhUSE working group




and with several our customers significantly reducing the number of programs and TFLs and providing traceability from the result to the data that supports it.

**WHAT IF WE COULD….AUTOMAGICALLY EXTRACT USABLE CONTENT FROM THE PROTOCOL**

What if a machine could read the contents of a protocol and extract study design information such as inclusion/exclusion criteria, objectives, and the schedule of events curating that information in a database and making


it searchable and usable to downstream processes? What if we could use that information to automatically pick the CRFs that are used for a study?

This can be done now through natural language processing by identifying key components and letting the machines learn from many iterations of processing. We have worked with customers to tackle this concept within clinical study reports under the umbrella of clinical transparency and sharing of reports.

**INTEGRATED SPECIFICATION**

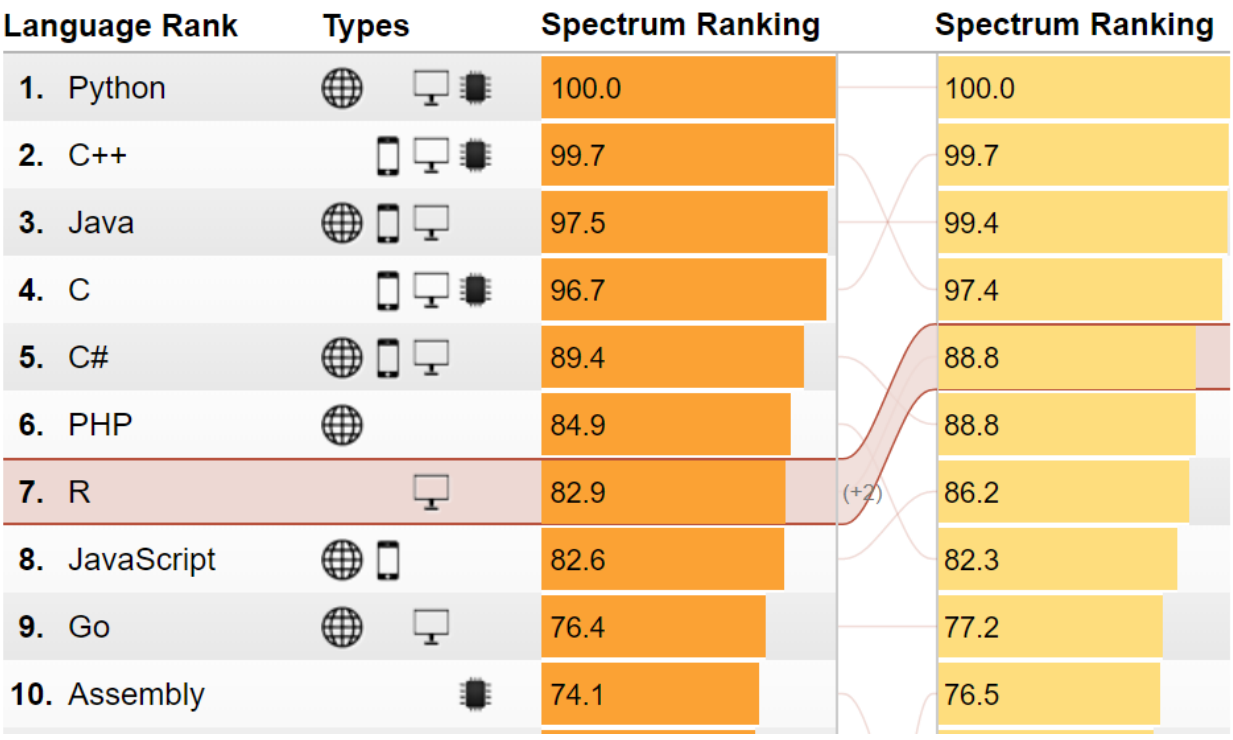
What if we could define our data standards in a way that truly links the information from the analysis through to how it was collected? This would require us to break down our current silos, take our content out of Excel spreadsheets and build a robust model which includes these relationships – contextual data. Again, this has been shown through PhUSE projects and working with our customers. In addition, CDISC is exploring the need to integrate across their standards as the next generation CDISC model.

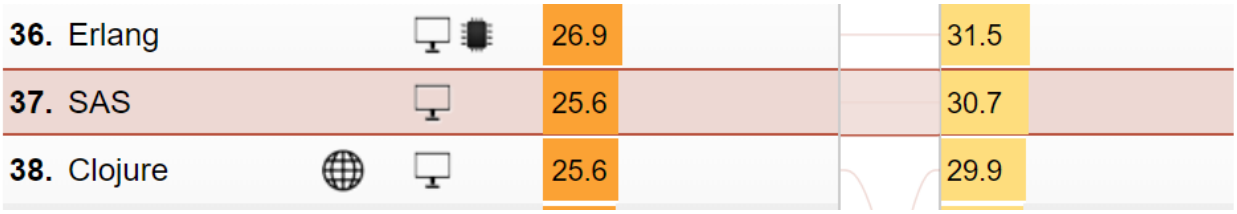

## CAREER DEVELOPMENT IN FOCUS

So, with the changing paradigm in mind how do we approach our careers? Should we stick or twist?

## TOP PROGRAMMING LANGUAGES

The top ranked programming languages are listed below and it's worth noting where SAS ranks amongst the other languages. The list should be taken with a pinch of salt as it is across all types of business, not just clinical trials. But one area it does raise is that there are many programmers out there with varying skills than pure SAS. In more familiar areas, R is one of the top 10 languages and Python is the number one but how does this apply to the clinical trials business? AS with the theme of this paper, you must look beyond what you know and what has been for the last 30 + years. Open source programming is coming, more graduates are entering the business with expertise in R, more data scientists entering the market with varying skills focusing on some new and existing areas of business.






## WHAT DO YOU NEED TO LEARN IN THE FUTURE?

As the data paradigm shifts, as the bridges are built the skills required are going to be shifting too. There will be the need for SAS, of course, but there will be a flexibility in open source analysis strategy that may require our future programming teams to be fluent in more than one language. It opens the doors for other expertise to add value as data scientists in our relatively hard to resource areas. We need to take steps to upskill ourselves, so we can make the necessary changes to the business. We at least need to be prepared to make those leaps to enable that flexibility.

When we think about where to start we need to consider the most flexible of processes. Data processing, analyses and TFLs offer the most inviting place to start adding flexibility to. What the best language to use is not clear, but until we take the starting point, the idea, and make it happen, we will never know. Picking and choosing the right tool for the right job is important to make a successful change and we need to be prepared to add those skills to our portfolios.

1) Programming languages include, but by no means limited to R, Python, C++, Javascript
2) Cloud options that include Amazon Web Services (AWS), Microsoft Azure, Docker amongst others.
3) Linked Data concepts including RDF, graph databases – NEO4J, protégé, ontology definitions
4) Machine Learning - AI, NLP, using Python, training algorithms
5) User Experience, design sprints, agile methodologies to focus on a new way of working

The list above does not nearly cover all the opportunities out there, but merely gives some options that appear to align with a new paradigm. There are several drivers for this change that include licensing, emerging talent, open source advancements, cloud technologies, and a willingness to take notice of these changes. Machine learning, artificial intelligence, graph databases, semantics, ontologies are all buzzwords we hear right now, and these require differing skill sets than the ones we are used to. But we need to break the paradigm, reach out and step towards the new world of programming. Take a course on Python or R. Try to apply it to your job role… Take a PhUSE workshop on linked data or machine learning and take the first steps towards a great big beautiful tomorrow.

## CONCLUSION

As you read this paper, you might be either saying it's full of fairy tales that will never be realized, OR I like the ideas but struggle to understand where we start, OR this is where we need to go, let's charge the hill! No matter what you believe, I hope we can all agree that we find different ways to do our job, collect and analyze contextual data, and help get treatments to patients faster and with more confidence.

To achieve a new paradigm, it requires us to think in new ways and realize what our key unit of measure is – it's not a spreadsheet or a SDTM data set or a paper TFL. It's the analysis, the objective, the key clinical concept we are trying to capture and study. If we can shift our minds to think in a different way we can make major changes and disrupt the industry. What skills would I need to program Alexa to understand an underlying graph database? What do I need to do to empower the future vision? Which classes should I take? The future is out there, we just have to get on board.

## REFERENCES

Analysis Results PhUSE Project: http://www.phusewiki.org/wiki/index.php?title=Analysis_Results_Model
Clinical Trials as RDF: http://www.phusewiki.org/wiki/index.php?title=Clinical_Trials_Data_as_RDF
Interactive: The Top Programming Languages 2018 https://spectrum.ieee.org

## ACKNOWLEDGMENTS

I would like to acknowledge the small, but growing group of people across industry who continue to push these ideas forward. You know who you are and let's keep fighting the good fight.

## RECOMMENDED READING

Linked Data 101: https://en.wikipedia.org/wiki/Linked_data
Moving to a Round World: https://www.lexjansen.com/phuse/2017/ds/DS10.pdf

## CONTACT INFORMATION

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

Chris Decker
d-Wise
919-995-0957
Chris.Decker@d-wise.com
www.d-wise.com