How to Stem the Tide of Data Deluge to Produce Better Science in Biopharma R&D

Today, being a scientist is tough, especially if we consider biopharma R&D through the lens of generating new insight and knowledge. As a community, we now create more data than ever before, and we want to analyse it with more speed and fidelity than previously possible. Data are prolific. They are the essence of scientific research and the discovery currency that we turn into the tangible asset of knowledge through analysis. Data fall into a number of categories: data we produce are endpoint-driven and generated for a specific purpose; contextual data add colour to explain our findings; some can be inherently difficult to interpret; and some data are simply the 'noise' within which we hope to find that elusive 'signal'. Each of these categories requires tools and techniques to sift and process the raw data into the refined result we want to consider. What we can say with certainty is: today there are a lot of diverse data to analyse and using our expert judgement to assess their scientific merit can take significant effort and will require a specific strategy to succeed.

This data backdrop raises some interesting questions for us to consider: How do we do analyse data effectively in the future? What protection mechanisms do we need to put in place to stop us all from floundering? Which methods will generate useful insight against this big data backdrop?

In this article, we will try to dissect this problem and identify some ways to help scientists stem the data deluge.

The Difference Between Data and Knowledge

Data is the lowest tier of the classical Data, Information, Knowledge, Wisdom (DIKW) pyramid (Ackoff, 1989) and forms the basis of our knowledge and wisdom. We interpret 'data' to generate a tangible insight that we then call 'knowledge' but this distinction is often missed, and the terms of data and knowledge are frequently used interchangeably. Simply having data does not create knowledge and creating large amounts of data does not automatically generate lots of knowledge. Instead, more thought and more tools are needed to help us to ascend this value chain.

What Causes Data Deluge?

Global scientific output doubles every nine years (Noorden, 2014) and this combines with the recent estimate from IBM that over 90% of the world's data have been produced in the last four years (IBM, 2013). The Big Data explosion is well-documented and truly a part of modern scientific research.

So if we believe these estimates to be true, generating data is not necessarily a bottleneck. On a practical level, most scientific instrumentation can produce many channels of data for processing and further analyses. However, scientific knowledge generation from these data sources can be hampered by various limiting factors, each of which become more impactful as the volume and diversity of data increase. Each of these factors can work alone or in concert. If not controlled, however, these factors can combine to reduce the ability to produce information, and therefore scientific knowledge.

Looking at this in detail, we can quickly generate a shortlist of limiting factors that fall into two broad categories, classified as *environmental* and *cultural* factors.

Environmental factors can be defined as attributes that relate to *the process* of generating knowledge from data. They can be generic and not necessarily specific to bio-pharma R&D per se but they affect what we do in the lab.

Conversely, cultural factors, when generating scientific insight, can arise *as a consequence* of how our science training directs us to look at data and are oftentimes institutionally/historically-based in their origin.

Environmental Factors	Cultural Factors
Geographically distributed teams	Appropriate expertise in data analysis techniques
Speed and volume of data creation	Data alignment between discrete datasets
Authenticity of data	

Fig. 1: Some factors contributing to the data deluge

- **Geographically distributed teams** – the establishment of FIPNet (Kaitin, 2010), a strategy for a collaborative network of information professionals, addresses the fact that co-workers may not reside in the same geography or time zone anymore. Science is now a team sport comprised of multiple players and organisations. Coordination of these players at scale can be difficult to manage with only traditional human efforts. These distributed teams require systems that help them to collaborate to generate and store the institutional knowledge generated (Leven & Teburi, 2016).

- **Speed and volume of data creation** are becoming hot topics across the industry. The speed and data intensity of some instrumentation and techniques are an order of magnitude greater in some instances. More data are generated more quickly than before (Rudd, 2017).

- **Authenticity of the data** – data provenance is a key principle of good science. Steps to ensure unadulterated data can be affected by emerging technologies such as blockchain and other techniques. These principles are employed across many sectors but gain greater importance particularly with clinically-relevant data sources (Shute, 2018).

- **Availability of expertise in data analysis** – Recent reports from the Association of British Pharmaceutical Industries indicate that the life sciences industry is facing a data analysis skills shortage (ABPI, 2015). This raises the question of how all scientists become data-analysis savvy when the volume of data is so large? Given the range of informatics expertise, support from additional systems is needed to support both analysis experts and non-experts.

- **Alignment of data and its challenges** – Initiatives such as openphacts (https://www.openphacts.org/) and FAIR principles for data stewardship (Wilkinson, Micheal, & al, 2016) mean that comparison of scientific data in a standardised and accessible manner is critical to making comparisons. Performing this standardisation at scale poses a challenge to scientists less comfortable with data analysis.

A Solution?
Fortunately, all is not lost. We have at our disposal a combination of technological solutions that we can employ to overcome these knowledge generation challenges. To deal with this, scientific data analysis software capable of meeting the data deluge challenge is an important component in the scientist's arsenal. In selecting the right solution, scientists should consider the following three attributes:

1) Enterprise-wide data analysis platforms
Analysis of scientific data requires a sequence of reproducible data reduction steps to convert the raw data into actionable knowledge following a pre-determined (but with the ability to be flexible) scientific workflow. Increasing data volumes and the complexity of the analysis will mean that manual techniques to perform data processing are becoming insufficient to keep up with and keep track of what has been done.

Implementing enterprise workflow data analysis platforms enables a scientist to automate the data analysis process and is critical to success. For these platforms to work effectively, they should provide transparency in the analysis process being performed and provide an audit trail to support the scientist. Selecting the right platform early can significantly improve efficiency of data processing.

Furthermore, the knowledge generated in any single workflow and the associated contextual meta-data also used by other players in the distributed team. Therefore, a large research team may have different uses for the generated data (different research foci), and it becomes imperative for research groups to use data analysis systems that can store the organisation's (the enterprise) accumulated knowledge for sharing while satisfying the individual scientist's workflow needs.

2) Use the right 'science aware' data analysis platforms
Each scientist should critically assess what is adding value in the data processing workflow and where to possibly eliminate or automate non-value add steps. Oftentimes, simple data manipulation can take up a significant proportion of a poorly defined analysis workflow. This penalty may be acceptable when data volumes are low but once a threshold in data production is reached this proves unsustainable. A science-aware and full-featured data analysis platform can remove this bottleneck in two ways: 1) Automate non-value add tasks (e.g. data parsing or reformatting); 2) Provide baked-in best-of-breed analysis techniques that are most efficient. Other key attributes to look for when performing this selection are platforms with open application programming interfaces (APIs), which enable seamless integration into a mixed environment and flexible deployment mechanisms that support changing organisational needs

3) Consider automated decision support
Modern data analysis platforms can now find signals without human intervention at a speed that matches the data generated. Sophisticated data reduction techniques such as machine learning and artificial intelligence are becoming more commonplace and should be embraced by bench scientists as a tool to maintain their research velocity and not seen as a replacement for the human scientist. Moreover, these techniques will become sophisticated not only in analysis capabilities but in suggesting items of interest that would otherwise have been missed in a manual analysis process.

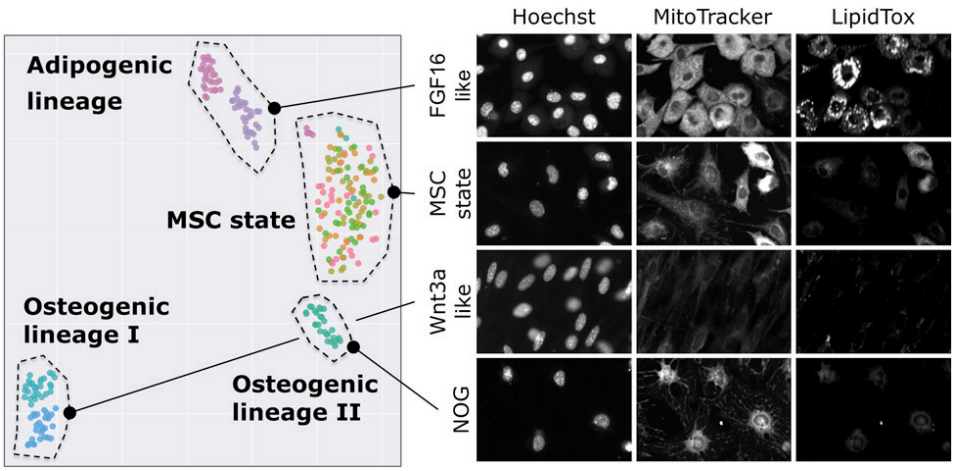

Fig 2. 2D visualisation of cellular phenotypes derived from a high-content screening. Osteogenic controls split in two groups (Wnt3a, TGFb1 vs. NOG) while adipogenic controls (FGF16 like) group together. This phenotypic grouping on this similarity map is confirmed by visual inspection of underlying measured images. The combination of similarity maps and quick visual review allows for a very quick and efficient exploration and definition of phenotype groups. The procedure takes many weeks with a classical HCS approach and just a few minutes with deep learning. Figure taken from Siegismund et al., poster presentation (SLAS 2018).

Conclusion: The Approach We Need

These data volume problems are not new and have been emerging over time. The key difference today is that due to our ever-connected world, the business of science is a team endeavour with researchers across labs, countries and continents contributing to the scientific discovery on a particular subject, with research performed at a scale previously never imagined. Aligning the needs of a scientific research organisation with the temporal nature of the data generated adds an extra dimension to overcome.

With the amount of data and subsequent knowledge being generated at a faster rate than ever before, tracking this knowledge explosion will require researchers to embrace change through smart thinking and adaptation. To stem data deluge requires greater reliance on smart automation and analytic tools, which will ensure future-proofed data analysis that incorporates protection mechanisms to drive meaningful insights and knowledge.

REFERENCES

1. ABPI. (2015). Bridging the skills gap in the biopharmaceutical industry. The pharmaceutical journal, The Pharmaceutical Journal, Vol 295, No 7883, online | DOI: 10.1211/PJ.2015.20200064.
2. Ackoff, R. (1989). From data to wisdom. Journal of Applied Systems Analysis, pp. vol 16 pgs 3-9.
3. IBM. (2013). https://www.ibm.com/blogs/insights-on-business/consumer-products/2-5-quintillion-bytes-of-data-created-every-day-how-does-cpg-retail-manage-it/. Retrieved from IBM.com: https://www.ibm.com/blogs/insights-on-business/consumer-products/2-5-quintillion-bytes-of-data-created-every-day-how-does-cpg-retail-manage-it/
4. Kaitin, K. (2010). Deconstructing the Drug Development Process: The New Face of Innovation. Clinical pharmacology and therapeutics., 87(3):356-361. doi:10.1038/clpt.2009.293.
5. Leven, O., & Teburi, K. (2016, June). An Antidote to R&D Antipatterns. Genetic Engineering and Biotechnology News.
6. Noorden, R. V. (2014, May). global-scientific-output-doubles-every-nine-years. Retrieved from http://blogs.nature.com/: http://blogs.nature.com/news/2014/05/global-scientific-output-doubles-every-nine-years.html
7. Rudd, J. (2017, May). 7 Data Challenges in the Life Sciences. Retrieved from technology networks: https://www.technologynetworks.com/informatics/lists/7-data-challenges-in-the-life-sciences-288265
8. Shute, R. (2018, Feb 7). Scientific Data Doctoring: Could Blockchain Technology Help Stamp It Out? Retrieved from www.bio-itworld.com: http://www.bio-itworld.com/2018/02/07/scientific-data-doctoring-could-blockchain-technology-help-stamp-it-out.aspx
9. Wilkinson, M., Micheal, D., & al, e. (2016). The FAIR Guiding Principles for scientific data management and stewardship. Nature: Scientific data, Article number: 160018 (2016)DOIdoi:10.1038/sdata.2016.18.

Kevin Teburi

Kevin Teburi, managing director of Genedata Ltd., is passionate about the role of scientific informatics in drug discovery and life science research and biotechnology. Prior to Genedata, he led informatics research teams at AstraZeneca.

Twitter: @kevinteburi