BIG DATA, SMART DATA AND BIG ANALYSIS

WHAT CAN THE PETRO-INDUSTRIES LEARN FROM BIG PHARMA AND THE ALLOTROPE FOUNDATION, AND WHERE SHOULD THE FUTURE LIE?

At first glance the Petro industries may not seem to have much in common with the Pharma industries. Petro produces vast quantities of bulk fuels, refined and petrochemical products usually from large continuous processes which at the minimum specification level must be fungible between suppliers. Conversely, in Pharma individual companies produce patented, pharmacologically active pure compounds in small quantities usually from batch processes.

However, when it comes to the analytical world they would seem to have more in common. Both use a wide range of elemental and molecular characterisation techniques either as standard methods, to demonstrate product compliance and quality, or bespoke methods in research, development and problem solving activities. As the sophistication of analytical methods has developed the volume and complexity of the analytical data produced has increased dramatically, for example through recent developments in comprehensive chromatographic techniques and high resolution, high mass accuracy, mass spectrometry (1)(2). In addition, much of this data is information rich at the molecular level and can offer new opportunities for developing structure-property relationships to improve process efficiency and develop new products with differentiated performance. Two examples of this could be

- The detailed molecular information on crudes and process stream composition provided by Petroleomic approaches (3) when combined with historical process data could lead to new routes to produce and value crudes, optimise refining strategy and identify key components leading to processing issues such as fouling and corrosion.
- By combining detailed major minor and trace component analysis of formulated fuels or lubricants with test data from rig and engine performance tests and consumer trends using advanced data analytics it should be possible to identify key chemical components and additives which generate differentiated performance and therefore premium prices and margins in the market.

The general hypothesis is: If I have more data at my fingertips – then I will have more answers

But, this is not necessarily the case as real-world data is messy data, filled with inconsistencies, potential biases, and noise. Therefore, to be able to effectively mine this wealth of data to generate increased performance we need an innovative approach to "Big Data" to generate "Smart Data" with a view to achieving value through "Big Analysis".

So, if we consider the current situation in typical petro industry laboratories around analytical data we can identify several key constraints including:-

- Data Silos, even between laboratories in the same company
- Incompatible instruments and software systems, often with proprietary 3rd party data formats
- Legacy architectures are brittle and rigid with low connectivity.
- Critical knowledge resides in people's heads, little common vocabulary
- Data schemas are not explicitly understood
- Lack of common vision and language between business units and scientists
- The Petro-Industries deals with products which are complex molecular mixtures often with many thousands of differentially active components

All of this this makes it almost impossible to move forward to a world of smart data and big analysis unless these constraints can be overcome.

The 4 Vs of Big Data

Big data exponents often refer to the 4Vs of big data – namely Volume, Velocity, Variety and Veracity as illustrated in Figure 1.

Volume and Velocity are normally the areas covered by conventional big data analytics and have shown dramatic development over the past few years thanks to the technology developments of many companies such as Amazon, Google, Netflix, etc. Data Variety on the other hand speaks to the increasing types of data sources available to companies for analytics – something that continues to grow at a rapid rate (e.g., image data, video data, unstructured text data). Data Variety speaks to data complexity and here semantic technologies have a clear advantage due to their graph-based ability to connect various data on a conceptual and class-based level. Veracity refers to data uncertainty and abounds in scientific and experimental data. Here statistical and probability analysis is required with mathematical clustering techniques providing clear advantages – often referred to in contemporary circles as Data Science.

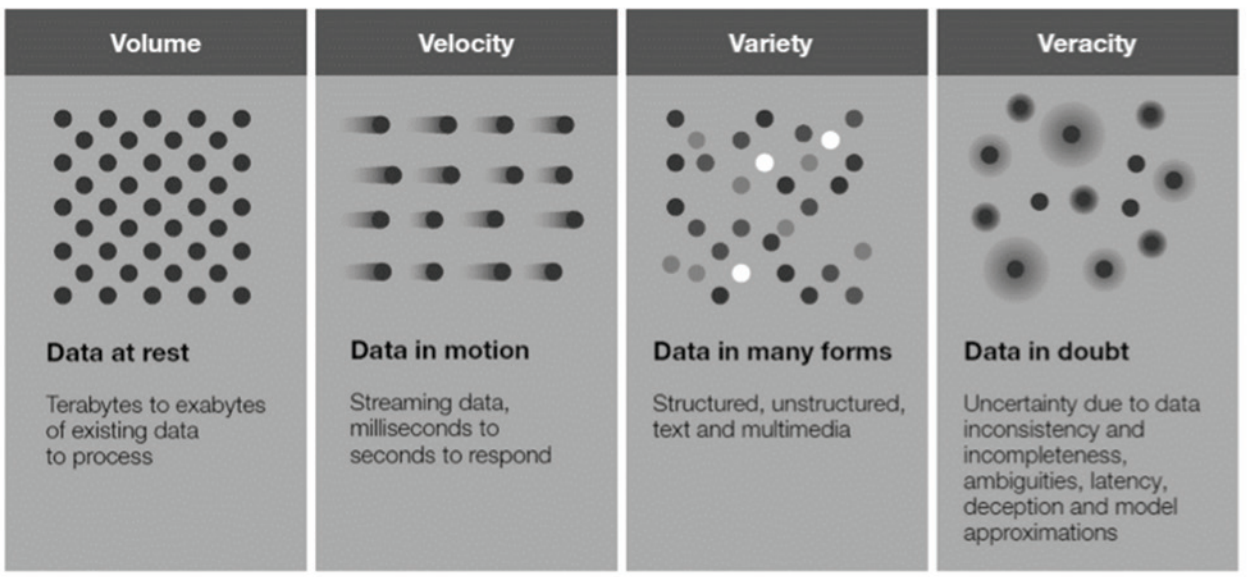

Figure 1: The four Vs of big data

Data Variety, Big Pharma and Allotrope

In pharma analytical laboratories a key source of the variety of data pertains to the plethora of instruments from a range of vendors and the variety of file types they generate, which were traditionally proprietary and owned by laboratory instrument vendors. So, if you have an HPLC or GC from one vendor then devices from other vendors cannot easily exchange data in the same lab because they produce incompatible data formats. This problem is exacerbated across large organisations who must manage multiple labs at multiple sites. To overcome this problem, the pharma industry established the Allotrope Foundation. The Allotrope Foundation website states:-

"Founded in 2012, Allotrope Foundation is an international consortium of pharmaceutical, biopharmaceutical, and other scientific research-intensive industries that is developing advanced data architecture to transform the acquisition, exchange, and management of laboratory data throughout its complete lifecycle. Its first initiative is the development of the Allotrope Framework for analytical data, consisting of a standard data format, class libraries for interfacing with applications, and semantic capabilities to support standardised, structured metadata. Allotrope aims to make the intelligent analytical laboratory a reality – an automated laboratory where data, methods and hardware components are seamlessly shared among disparate platforms, and where one-click reports can be produced based on data generated by any analytical instrument, and data integrity is built-in by design. Allotrope's vision of an intelligent analytical laboratory will be realised through the creation of an "ecosystem" in collaboration and consultation with vendors and the scientific community. Our shared mission is to develop innovative approaches to improve data access, interoperability, and data integrity through standardisation which ultimately serves as a key enabler of data-driven innovation." (www.allotrope.org).

Allotrope captures metadata in the Data Description Layer, together with raw data from the instrument (contained in the Data Package), as well as Data Cubes that can contain array data such as 2D or 3D chromatograms in a standardised file type called the Allotrope Data Format (ADF) (see Figure 2). All of this is contained within a large-scale HDF5 Container (ADF files can be petabytes in size based on the amount of raw data from a specific run). ADF files are generic and can serve as an industry-standard that do not fall prey to the issues of proprietary file types generated by specific instrument vendors.

This variety of data is certainly not unique to pharma and although in general the analytical methods and chemistries used by the petro-based industries differ significantly from those in the pharma-based industries they use the same instruments from the same vendors and therefore suffer from the same issues. So, if the petroleum and petrochemical industry want to embrace "Big Data" and move to "Smart Data" and "Big Analysis" they too will have to tackle the subject of analytical data variety and veracity at an early stage. This begs the question as to whether they will choose to "reinvent the wheel" or is the time right for the industry to get together and look at where developments emerging from Allotrope could equally be applied for them.

How do you Integrate all the Vs and where does the future lie?

Integrating the 4 V's allows one to understand the connections between Big Data amounts (Volume), the speeds at which it must be queried or processed (Velocity), the types of data and their logical connections (Variety) and the probabilities and uncertainties surrounding the data (Veracity). Combining this together into a common approach is what we call "Big Analysis" because industries need to think about not only how they store and manage their data, but more specifically how they can utilise it for business decisions. Big Analysis combines two basic types of technologies together (see Figure 3), namely mathematical technologies on the one hand (top row of Figure 3) and semantic technologies on the other hand (bottom row of Figure 3). Mathematics treats data in terms of probabilities and statistics and is most commonly associated with traditional data science. Here one utilises bottom up approaches from within the data to cluster data points, extract important features algorithmically, trend data over time, etc. Semantics, on the other hand, provides a logical way of representing data from a top-down perspective where one applies logic, concepts and rules of inference. Normally, these items are considered in isolation in computer science, where calculations and data science is purely mathematical, but semantics provides metadata modelling of class-entity-relationship structures. The approach of Big Analysis considers their association and overlap. In this manner, Big Analysis can compute over entity-level data derived from a posteriori (experiential) knowledge directly gleaned from the data sources or a priori (pre-experiential) knowledge which represent the most basic logical classification schemas that people use to form background beliefs and metaphysical understandings of classes of things and relationships between them. Taken together, Big Analysis provides a way to understand both the mathematical and logical structures of one's data, providing various ways to interrogate and manipulate the data for analytics, search, pattern analysis, etc. using automated methods of various types.

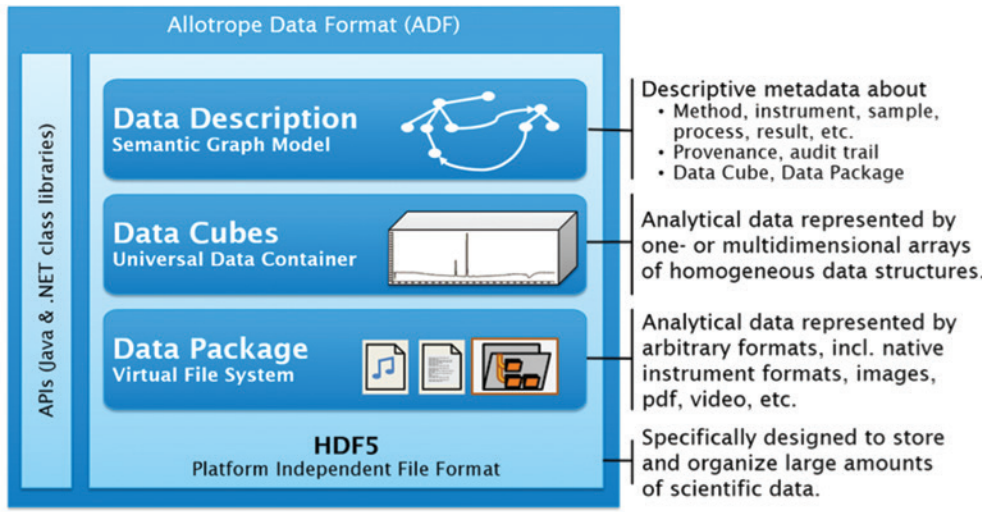

Figure 2: Allotrope Data Format (ADF) Overview

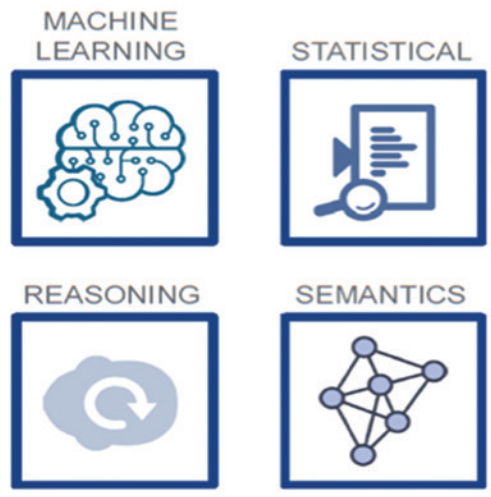

Figure 3: Combining Mathematical Techniques (top) with Logical Techniques (bottom) to form Big Analysis

Many of the issues we currently see with legacy data sources across the petrochemical space is that databases were built to store data, to lock it away into a vault for safe keeping. Many of these systems were not designed for easy retrieval of data, let alone repurposing of it over time for a growing set of use cases and business needs. Likewise, we see the same approach historically with Data Warehouses, which were often built by massaging data into certain schemas that matched to specific business questions or problems at a certain time. Given the changing variety of data already discussed, it is imperative that future systems be far more malleable and offer users an easier way to run a wide range of analytics for an ever-increasing set of business purposes. Big Analysis looks at how companies want to utilise their data across different business units and across different virtual and physical locations. Therefore, this approach attempts to understand a variety of use cases from the beginning and store the data in a flexible and well annotated way so that analytics is easier to perform now and in the future.

One key area that relates to Big Analysis is the concept of FAIR Data, initially adopted by the Open Source scientific community (4). FAIR Data is an approach to make data more Findable, Accessible, Interoperable, and Reusable. Like the Data Description Layer of the ADF file discussed previously, FAIR data embraces semantic principles whereby data is logically organised, annotated, and carefully modelled using sound logical and metaphysical principles. FAIR data, in turn, provides an ability to build Data Catalogs where existing (i.e., legacy) data sources can be more easily described via their metadata, making them available for search and analytics. A Data Catalog provides a large enterprise with a means to know where its data sources are, what they contain, who has access to them, how they are governed, security modules, and which connectors (e.g., APIs) are needed for access. Data that can be understood in this manner can be more easily integrated into large-scale, cloud-based applications and available for a variety of applications and analytics.

The future of data analysis is rapidly moving towards cloud-based approaches where data will reside in a combination of both public and private clouds (e.g., servers running in a cloud service such as Amazon, Google, Microsoft, etc., in combination with on-site managed servers). Analytics are growing at a rapid pace, requiring large amounts of various data to be brought together in order to solve complex problems and understand complex situations. Analytics of this nature require a better technology infrastructure than has historically been provided as more and more companies are looking for ways to utilise their vast data sources more efficiently and effectively for valuable business decision-making processes. Understanding data Variety and Veracity moves the discussion from just amounts of data to types of data and the statistical significance of it. This amounts to a cognitive shift in what data means and how it can be used for analysis. Organising existing data into FAIR-ified data catalogues allows companies to take advantage of the investments they have made into their relational database systems, business analytics tools, cloud tables, etc., because those data sources can now be more easily located and used across the enterprise. Analytics systems that can then utilise the plethora of new and existing data sources can provide unprecedented abilities for petrochemical companies to use their data for improved business decisions.

Conclusions

The time is ripe for the petroleum-based industries to embrace the concept of "Big Analysis". At the moment many individual companies will already be chasing the benefits and looking at how they can make better use of the vast quantities of compositional data they generate on their feedstocks and formulated products. A large amount of data currently exists inside large petrochemical companies and more data is being created daily. Harnessing that data and providing new ways to utilise it for ever-increasing complex analytics could provide companies with unprecedented capabilities. Petro-chemical companies must begin to understand that their data is an extremely valuable asset, as valuable as crude itself, since the ability to optimise, cut costs, drive new innovations, virtualise experiments, create analytical pipelines, and answer complex business questions relies on the use of data. Big pharma companies have realised some time ago that there will be clear winners and losers in their competitive space, based on who has the ability to wrangle and utilise their data, in order to make better decisions at important stage gates, and drive new areas of innovation. Petrochemical companies could follow suit in a similar manner by understanding their data and being able to better utilise it in more complex ways – ultimately lending a form of controlled and specialised artificial intelligence for the petrochemical space. This is achievable in the next few years for those companies who embrace newer technology approaches to: understand their data sources, apply FAIR principles to them, understand the 4 V's of Big Data, apply the approach of Big Analysis (statistics + semantics), and bring together human and machine representations for improved analytics. As a first step the time is right for the industry to look at where developments emerging from Allotrope could equally be applied for them.

References

(1) Selective ionisation and affordable high resolution mass spectrometry is revolutionising molecular characterisation in the petroleum and petrochemical industries: Tom Lynch, Petro Industry News, August/September 2018

(2) The versatility of time-of-flight mass spectrometry for the investigation of petroleum derived matrices – from middle distillates toward vacuum residues: Thomas Gröger, Maximilian Jennerwein, Uwe Käfer and Ralf Zimmermann. Petro Industry News June/July 2018

(3) Petroleomics: Chemistry of the underworld: Alan G. Marshall and Ryan P. Rodgers, PNAS November 25, 2008. 105 (47) 18090-18095; (https://doi.org/10.1073/pnas.0805069105)

(4) The FAIR Guiding Principles for Scientific Data Management and Stewardship https://www.nature.com/articles/sdata201618

Author Contact Details

- **Tom Lynch CSci, CChem FRSC, Independent Analytical Consultant, Cricket House, High St, Compton, Newbury, RG20 6NY**
  Email: tomlynch.lynch@ btinternet.com
- **Eric Little, Chief Data Officer, OSTHUS Inc. 1990 W New Haven Ave, Suite 301, Melbourne, FL 32904, USA**
  Email: eric.little@osthus.com