**Data Science – Trading Houses case**

I choose to write this little note to give my view on what data science is in Commodities, as there is no closed definition today, and to show through examples some real case applications for Commodities Trading Houses.

This trendy field englobes numerous concepts, and data science is often viewed as long and costly projects for “geeks". However, it can be experienced with agile methods through small projects, to have a quick impact in distinct parts of the business. It provides the tools to reach the objective of business optimisation: saving time and costs, bringing rigor in some aspects, and can even turn some support functions into profit center.

***Definition – Data at the center***

(*Wikipedia*) [[1]](#endnote-1)Data science is an interdisciplinary field about scientific methods, processes, and systems to extract knowledge or insights from data in various forms, either structured or unstructured.

This introduction from the free encyclopedia gives a broad overview of what we call data science. This term became a buzzword after the Harvard Business Review called Data Scientist the “sexiest job of the 21st century”, but we see that there is no real academic definition of that. What we can take away from Wikipedia is that data is center stage, with an objective of extracting knowledge.

It is the practical intersects of several fields like Computer science, Economics, Statistics, etc. that gives it its definition in day to day life. Often compared with statistics, which can be considered somewhat as its “parent”, for me the main difference is based on its usage:

* Statistics are used more frequently to look at the past, for explanation
* Data science, is used for forecasting power

So, this new field, even though it is trending through many industrial sectors is, in my opinion, under exploited within the commodity trading industry: the companies do not use the data they are sitting on. It is often considered to be within the scope of the market intelligence analysts, or economists, to deal with. But the potential usage of it is much wider.

An article from *The Ecomomist* [[2]](#endnote-2)recently shed some light on data: according to The Economist, data is now more valuable than what oil used to be. Companies know they are sitting on a gold mine of information, but as of today, they are often not able to exploit it due to processes, organization failure or lack of technical knowledge.

Within trading houses, the information is often reported through phone calls or unstructured emails. So, even if the information travels quickly via these ways of communication, the data is not collected for further analysis. This does not even take into account that even before collecting the data, the first step we should be thinking about is structuring it.

***Concrete use cases***

The skills of a data scientist are distributed among three pillars: Mathematics, Coding and Business knowledge. The examples listed below are all based on authentic experiences and are an illustration of how the combination of the three main skills can help empower a business.

Processes improvement & cost reduction

A significant part of internal processes in a trading house are manually handled by people. This increases costs and errors, and often make processes dependent on the people running them. A part of the workforce, generally with an elevated level of education, is employed to do these repetitive tasks. In my opinion, they should be allocated to tasks with more value added for the company, leaving the simple ones to computers.

The automation doesn’t belong directly to this data science world, it is more a co-product, but it is still a part of the skills of a data science team.

*Example:*

This process improvement could be implemented for position reporting: I do not know how all trading houses are monitoring their cash and derivatives positions, but generally it is quite archaic.

In the best case there is an extraction process from a global system, reporting positions in Excel files, and then an email is manually sent to the appropriate trader for him to check if what he has in mind is matching with the system. Errors are common, as humans are implicated in all parts of this process. I have personally seen someone joining a department and completely changing the daily activities of 50% of them, making some processes more automatic using more powerful programming languages.

Risk management

Estimating and managing risk is critical for trading houses. Apart from the processes perspective, there is also often a lack of information regarding physical prices. Even if ongoing prices can be checked through brokers to estimate a mark to market value, the issue is more with historical data, used to compute classic risk metrics. There are ways to keep running risk analysis without the real data.

*Example:*

The classic example of need of information is when a company takes a position in a less liquid market: let’s say barley in Ukraine. Apart from in house data, if they are collected, there is no readily available data to estimate the risk from management’s standpoint. In situations such as these, many hypothesis need to be made in order to compute a few metrics.

One solution to overcome the available data is to use alternative techniques. In the example above, one could collect data related to Ukrainian Currency, Futures Markets, Cash prices available of similar commodities, and perform a Principal Component Analysis (PCA) to estimate the risk factors. A PCA is a statistical procedure that converts a set of observations into a set of linearly uncorrelated variables, which are the factors we want to highlight. Again, this kind of procedure do not solve for the company, but it can help to better understand the risk of that particular position.

Market Intelligence

Market intelligence is the first example of the good use of data science techniques which comes to mind. Developing econometric models to analyze markets. The idea is to be able to make decisions based on analysis, or more structured analysis at least.

*Example:*

The United States Department of Agriculture proposed a yield model [[3]](#endnote-3)based on weather. The model description is available in the references.

The model is a basic multilinear regression model and has been developed by the USDA after some droughts in the US to help the government analyze the impact of climate on different crop production. I spent the beginning of my career working on this topic, using the ideas of back testing and cross validation to give better confidence in my own forecasts, reaching better accuracy than what has been proposed by USDA. At the end of the day, all models are wrong, but some are useful. They give confidence in our analysis, and therefore using state of the art technique can help increase confidence levels.

Investment & Corporate Finance

A model is a simplification of reality, and financial models are heavily used when considering asset transactions or M&A. The classic methods for a valuation in the economic theory are the comparables and cash flow approaches, detailed below:

* Comparables: This is a direct approach, when looking at a specific metric within a set of “comparables”. In mining, it will typically be the (Value of Assets) / (Tones of minerals as Reserve). This kind of metric allows us to isolate factors, which then help determine price, and are generally reported in big tables in banks statement, for example.
* Cash Flow: Accounting analysis, making growing assumptions regarding several parts of the business.

Analysts help understand reality by simplifying assumptions and using machine learning techniques to help improve their models and provide more accuracy.

***Back to the data***

All these examples are interesting and useful ways to give more confidence in the tasks performed by the different parts of the business. However, it is clear that these techniques and approaches need a way to quickly collect information. As previously mentioned, data is generally not stored, and the information flows through broker reports, internal emails, or simple phone calls. In some cases, this is not collected. In the best cases, reports are only archived. Such is the current state of affairs for data collection.

An ideal flow of data flow within a company would look like something like this:

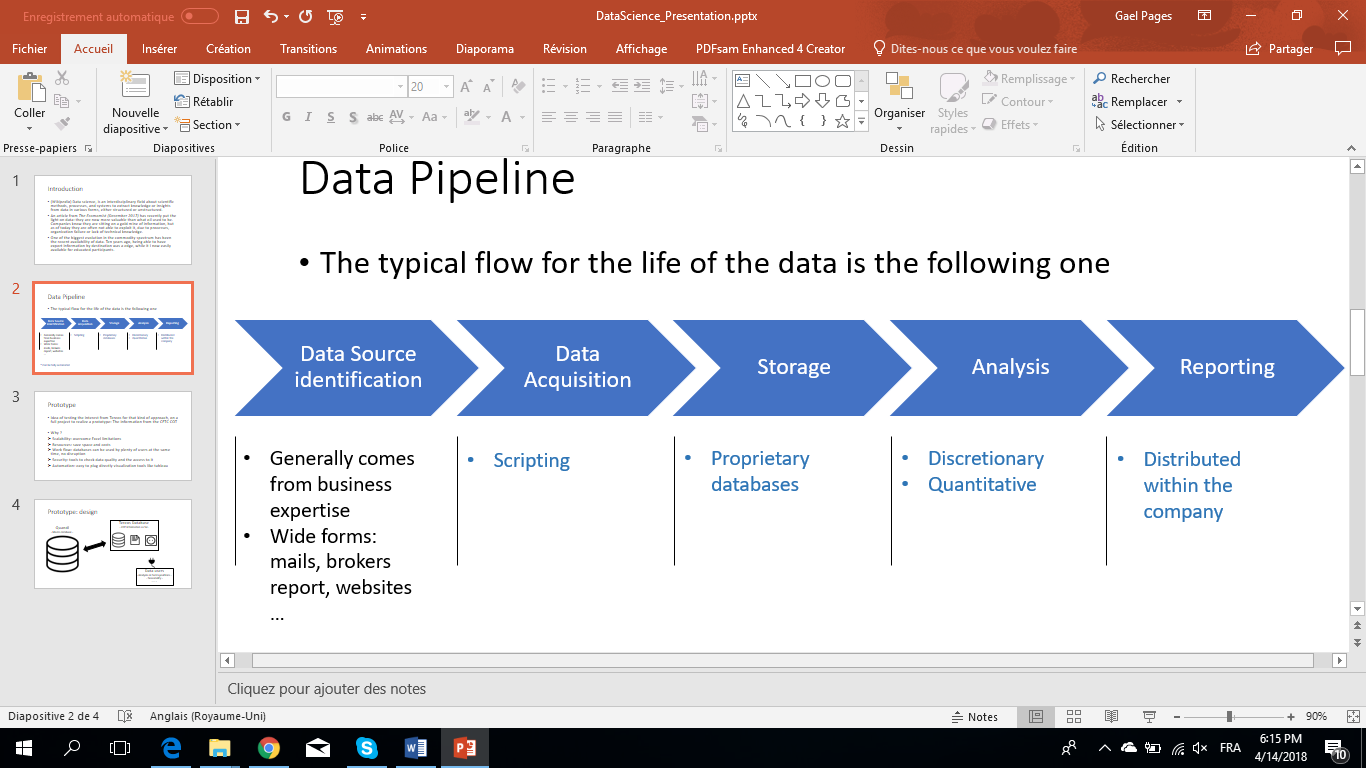


Figure 1: Ideal information pipeline

The empirical reality is very different, and there are few organisms in companies in charge of collecting and structuring this information pipeline. From my point of view, this should be the first task of data scientist teams in a company when starting their activities. Even if it has no measurable benefit in the short term, it is clearly a worth investment for the company in the long run.

In my opinion, structuring all information flow in this way is useful for the business, it could also turn the support function into a profitable business. Indeed, the first companies who are going to move on this area will be able to sell their data to other traders. Who can say that they have collected and structured relevant information in the last few years? One of the biggest evolutions in the commodities world has been the availability of data, for educated participants. The information that was giving an edge 10 years ago is no longer true today. However, companies are still sitting on this proprietary information without being able to exploit it. Structuring it and selling it to systematic actors could generate quite consequential profits.

***References***

1. Wikipedia – Data Science - <https://en.wikipedia.org/wiki/Data_science> [↑](#endnote-ref-1)
2. The Economist - <https://www.economist.com/news/leaders/21721656-data-economy-demands-new-approach-antitrust-rules-worlds-most-valuable-resource> [↑](#endnote-ref-2)
3. USDA - https://www.usda.gov/oce/forum/past\_speeches/2013\_Speeches/Westcott.pdf [↑](#endnote-ref-3)