When a data scientist drowns in the Data Lake

When companies hire a data scientist -putting apart the hype and the *"we hire data scientist because everybody else is doing that"*-, the expectation on her/him goes along these lines: **"I sit on a pile of data and I want you to generate all the insights I need to steer my company"**... But without a proper enabling groundwork, you are going to probably find the body of the data scientist on the shore of your data lake... the poor guy who inevitably drown due to irresponsible or inexistent data management.

The refinery without crude

Using the *"data is the new oil"* mantra -which I'm starting to hate-, it would be similar to: "*Somewhere in this area I got a lot of crude oil; do whatever you guys do to create fuel"*. Well, it is exactly what a refinery does, isn't it? Not quite... a lot of work is required before the distillation process starts in the refinery: a first phase of crude extraction, exploration for quality and volume, creation of a drilling rig, well evaluation and completion... After that, the process of taking the crude to the refinery starts: transportation in oil tankers or pipes, lightering, etc. You can have a look at Adventures in energy ... really well explained and fun to read.

Back in our data world, it is important that companies don't oversees what comes before the data science from data to insights distillery process, which I call **Data Science Enabling**.

If this enabling work has not taken place yet in your company, you end up having your poor data scientists new hires inevitably drowning in your data lake -if you have one- or in one of your siloed data reservoirs.

After "drowning", these guys can either leave the company to join a better refinery, or try their best as "data engineers", trying to dig out data from somewhere, distill it as far as they can to potentially come with insights that are not usable, because without the proper raw material you cannot produce any combustible.

Drowning in a data lake: know the symptoms

I've been leading several data scientists teams over the course of the years and I'm a data scientist myself -whatever it means-. When you lead a data scientist team, you are accountable for the results and you are supposed to do whatever it takes to get your team members in a position of delivering them (a.k.a. you cannot let them "drown" by making sure the right material reaches in the right way and the right pace the refinery).

But how do you know ***Data Science Enabling*** is yet to be done? your data scientists tell you that:

? I don't know which data sources are there or which ones are relevant for me.

Often happens that you are working with a data source, and well advanced in the process, somebody comes around the corner with a much better data source with information you have started to infer... Knowing which data sources are

available upfront saves time, resources and contributes to the quality of the results

? I don't know where the data is available:

- I don't know how to access the data
  Typical example, you are said that we have this wonderful data source in a proprietary database... you ask to get access and nobody can help there... no API and no services on top... what do you do?
- I don't have the tools (or I can't install it in my company laptop)
  Accessing corporate data is subject to security and data protection policies -the way it should be-. Often, you are given company equipment with a lot of restrictions (no Admin rights, you cannot install anything, etc). Yet, you need your tools to start making sense of the data... so you need help to navigate this hurdle.
- I don't have the rights
  Obviously, nobody can have access to everything. Yet the process of granting access to a particular cut of the data is often not well documented, technically not possible or it is part of a tedious ever lasting request process that take years.
- I don't know whether I'm indeed supposed to access a given data source
  Who can access what is clearly not well defined. Processes for granting temporary access to a particularly sensitive source (e.g.: via Non-Disclosure Agreements, etc) are often accountability-orphans. Those who want to help can't, those who can't don't care.
- I don't know whether I can copy, modify, persist, etc the data
  Once you get access to a particular sensitive source, there are most probably guidelines with Do's and Don'ts, but often unclear.

? I don't know the meaning of the data fields and the correspondence to business information

- I see a lot of Id's I can't connect to anything
  Often I'm delivered with just "facts" tables, but I miss the dimensions... so I end up with many funny named Ids I can't do anything with, but I feel they play an important role.
- I don't know how to aggregate my data into broader entities according to the business standards
  Often business taxonomies are not supplied. Often they don't even exists or worse, there are several versions hanging around in the company that are not quite compatible. Which one to use becomes a *Russian roulette* decision

? I don't know if I'm reinventing the wheel

- I don't know if somebody faced and solved the same problem or a highly related problem
  In big corporations, it is not rare that you are in the middle of a project and you get to know that somebody has already or is still trying to solve the same issue... but you get to know it by chance... there's no system to check for this information (a.k.a: knowledge management just missing)
- I don't know if when somebody claimed to have solved the same problem, it is true.
  Or even worst, your project get stopped or challenged, because somebody put on a power point, that they have already done that... but when you scratch the surface, there's nothing behind.

- I don't know the quality (MAPE, MAE, accuracy error) of the approach chosen by somebody who solved the same problem.
  But let's say there is something done... often quality metrics are missing, so if your method is better or worst remains unknown, because the existing solution does not provide any quality metric.

? I really have issues understanding the data

- I'm missing the business annotations giving the feedback on how to solve the problem or to explain the data
  Data consumers can enrich the data the best, because they have business context (e.g.: during the release week-end, we don't see any orders... is it a tracking issue? was the weather was so good? difficult to guess, there was a release, that's why annotations are a must have)
- I'm don't know whom I can talk about the data or where the process picture is
  Processes built up as different statuses in the data can't be easily understood... The data scientist can infer the process by identifying combinations of statuses with a timestamp but again this is to certain extent guesswork. The responsible for the process can be of much more help.

? I don't know if the data I'm getting access to is kept up-to-date or is complete

- I know it is just a sample, but I don't know how the sample has been taken
  Sometimes Data Scientist are given a dump of data, sometimes somebody just took a sample. Data sampling is per se a prolific research area with thousands of papers written every year. Also the way you create a data sample defines the entire data science process.
- I don't know by when I'm expecting fresher data
  To prevent overfitting / underfitting a model fresher data can be of great help. Not knowing when new data is going to arrive, enormously challenges the data science job.

? I don't know whether the data is consistent along the time line

- I don't know if any algorithm of data correction has been applied
  It is not unusual, that the data presents gaps... sometimes, somebody correct these gaps, but the remedy can be worst than the problem if not properly done (e.g.: interpolating the sales of a bank-holiday). Not knowing it, might lead the data scientists to draw wrong conclusions.

- I don't know the reason why there are gaps in the data (if any)
  A release, a bank holiday, a system outage, a change in the logs, or just something looking like a gap, that is not a gap... without proper documentation it is difficult to

- I don't have an indicator for the completeness of the data
  This is similar to the sampling, only forcibly done by the measuring system. Let's say you are analyzing an incidents log and the application registers only 75% of the incidents... Knowing that would be precious if you are tasked with creating an early warning system, don't you think so?

? I don't know whether the data is consistent with other data sources

- I don't know in which other additional data sources is the same information available
  Let's say your model is based on one particular data source, which is slightly inconsistent with another one (you didn't even know it existed)... your model is going to be conflicting with the findings in some other parts of the organization, and your existence in the company turned into hell.
- I don't know whether different users have the same access to a data source or each one has their own copies -in which case I don't know if they are 100% aligned-
  This is another aspect... local copy hardly ever updated... or slightly modified... snapshots you need to identify, because your local copy might become completely inconsistent over the time

? I don't know in which platform I can run my analysis or publish the results

- I don't know if my laptop can cope with so much information
  We all know that... data is getting big :) but apparently sometimes companies think that most of the tasks can be done in your own PC... Certainly many of them can, but not all!
- When I'm ready with my analysis, I don't know where to deploy it within the context of a data product
  Nowadays, good data scientists don't just provide the results of their analysis... they go beyond that and create data products. A data product is a piece of software which needs to be hosted on a platform and needs to be fed with fresh data.
- I don't have any environment with the tooling and the best practices processes of standard software development
  Software development best practices also apply to data products: version control, repositories, continuos integration, testing, etc. Probably none of these components have been made available for the data scientists to properly work.

What can be done before your data scientists start to sink?

Make companies aware of the need for proper data management. The role of a Data Quality Manager needs to be understood, well staffed and empowered.

Before a data scientist is thrown in the middle of the data ocean, there are some criteria that shall be fulfilled. In the picture below, I provide a quite useful check list where the most relevant Data Readiness criteria are listed and waiting for getting a tick on their boxes.

Sticking to this list can be literally a life-saver or the guarantee for success making the most of your data!