**On Data Quality**

* Data pubblicazione 17 settembre 2017

****

[**Roberto V. Zicari**](https://www.linkedin.com/in/roberto-v-zicari-087863)

**SeguiRoberto V. Zicari**

University Professor- Goethe University Frankfurt

I have interviewed a number of Data Scientists and asked them questions on the subject of **Data Quality.** I listed their reply, below. Perhaps they are useful for your work. Take what you think is relevant for you and leave the rest. And if you wish you can quote some of them in your publications (all interviews are listed with the relevant links below)

R.

**Jeff Saltz**: <http://www.odbms.org/2017/08/qa-with-data-scientists-jeff-saltz/>

*Q. How do you ensure data quality?*

Data quality is a subset of the larger challenge of ensuring that the results of the analysis are accurate or described in an accurate way. This covers the quality of the data, what one did to improve the data quality (ex. remove records with missing data) and the algorithms used (ex. were the analytics appropriate). In addition, it includes ensuring an accurate explanation of the analytics to the client of the analytics. As you can see, I think of data quality is being an integrated aspect of an end-to-end process (i.e., not a “check” done before one releases the results)

*Q. How do you evaluate if the insight you obtain from data analytics is “correct” or “good” or “relevant” to the problem domain?*

With respect to being relevant, this should be addressed by our first topic of discussion – needing domain knowledge. It is the domain expert (either the data scientist or a different person) that is best positioned to determine the relevance of the results.However, evaluating if the analysis is “good” or “correct” is much more difficult, and relates to our previous data quality discussion. It is one thing to try and do “good” analytics, but how does one evaluate if the analytics are “good” or “relevant”? I think this is an area ripe for future research. Today, there are various methods that I (and most others) use. While the actual techniques we use vary based on the data and analytics used, ensuring accurate results ranges from testing new algirhtms with known data sets to point sampling results to ensure reasonable outcomes.

**Yanpei Chen:**[**http://www.odbms.org/2017/08/qa-with-data-scientists-yanpei-chen/**](http://www.odbms.org/2017/08/qa-with-data-scientists-yanpei-chen/)

*Q. How do you evaluate if the insight you obtain from data analytics is “correct” or “good” or “relevant”?*

Here’s a list of things I watch for:

• Proxy measurement bias. If the data is an accidental or indirect measurement, it may differ from the “real” behavior in some material way.

• Instrumentation coverage bias. The “visible universe” may differ from the “whole universe” in some systematic way.

• Analysis confirmation bias. Often the data will generate a signal for “the outcome that you look for”. It is important to check whether the signals for other outcomes are stronger.

• Data quality. If the data contains many NULL values, invalid values, duplicated data, missing data, or if different aspects or the data are not self-consistent, then the weight placed in the analysis should be appropriately moderated and communicated.

• Confirmation of well-known behavior. The data should reflect behavior that is common and well-known. For example, credit card transaction volumes should peak around well-known times of the year. If not, conclusions drawn from the data should be questioned.

My view is that we should always view data and analysis with a healthy amount of skepticism, while acknowledging that many real-life decisions need only directional guidance from the data.

**Manohar Swamynathan:**[**http://www.odbms.org/2017/05/qa-with-data-scientists-manohar-swamynathan/**](http://www.odbms.org/2017/05/qa-with-data-scientists-manohar-swamynathan/)

*Q. How do you ensure data quality?*

Looking at basic statistics (central tendency and dispersion) about the data can give good insight into the data quality. You can perform univariate and multivariate analysis to understand the trends and relationship within, between variables. Summarizing the data is a fundamental technique to help you understand the data quality and issues/gaps. Below figure maps the tabular and graphical data summarization methods for different data types. Note that this mapping is the obvious or commonly used methods, and not an exhaustive list.

*Q. How do you know when the data sets you are analyzing are “large enough” to be significant?*

Don’t just collect a large pile of historic data from all sources and throw it to your big data engine. Note that many things might have changed over time such as business processes, operating condition, operating model, and systems/tools. So be cautious that your historic training dataset considered for model building should be large enough to capture the trends/patterns that are relevant to the current business problem, otherwise your model might be misleading. Let’s consider an example of a forecasting model which usually have three components i.e. seasonality, trend and cycle. If you are building a model that considers external weather factor as one of the independent variable, note that some parts of USA have seen comparatively extreme winters post 2015, however you do not know if this trend will continue or not. In this case you would require minimum of 2 years data to be able to confirm the seasonality trend repeats, but to be more confident on the trend you can look up to 5 or 6 years historic data, and anything beyond that might not be the actual representation of current trends.

**Jonathan Ortiz:**[**http://www.odbms.org/2017/04/qa-with-data-scientists-jonathan-ortiz/**](http://www.odbms.org/2017/04/qa-with-data-scientists-jonathan-ortiz/)

*Q. How do you ensure data quality?*

The world is a messy place, and, therefore, so is the web and so is data. No matter what you do, there’s always going to be dirty data lacking attributes entirely, missing values within attributes, and riddled with inaccuracies. The best way to alleviate this is for all data users to track provenance of their data and allow for reproducibility of their analyses and models. The open-source software development philosophy will be co-opted by data scientists as more and more of them collaborate on data projects. By storing source data files, scripts, and models on open platforms, data scientists enable reproducibility of their research and allow others to find issues and offer improvements.

*Q. How do you evaluate if the insight you obtain from data analytics is “correct” or “good” or “relevant” to the problem domain?*

I think “good” insights are those that are both “relevant” and “correct,” and those are the ones you want to shoot for. As I wrote in Q2, always have a baseline for comparison. You can do this either by experimenting, where you actually run a controlled test between different options and determine empirically which is the preferred outcome (like when A/B testing or using a Multi-armed Bandit algorithm to determine optimal features on a website), or by comparing predictive models to the current ground truth or projected outcomes from current data. Also, solicit feedback about your results early and often by showing your customers, clients, and domain experts. Gather as much feedback as you can throughout the process in order to iterate on the models.

**Anya Rumyantseva:**[**http://www.odbms.org/2017/03/qa-with-data-scientists-anya-rumyantseva/**](http://www.odbms.org/2017/03/qa-with-data-scientists-anya-rumyantseva/)

*Q. How do you ensure data quality?*

Quality of data has a significant effect on results and efficiency of machine learning algorithms. Data quality management can involve checking for outliers/inconsistences, fixing missing values, making sure data in columns are within a reasonable range, data is accurate etc. All can be done during the data pre-processing and exploratory analysis stages.

*Q. How do you evaluate if the insight you obtain from data analytics is “correct” or “good” or “relevant” to the problem domain?*

I would suggest constantly communicating with other people involved in a project. They can relate insights from data analytics to defined business metrics. For instance, if a developed data science solution decreases shutdown time of a factory from 5% to 4.5%, this is not that exciting for a mathematician. But for the factory owner it means going bankrupt or not!

**Dirk Tassilo Hettich:**[**http://www.odbms.org/2017/03/qa-with-data-scientists-dirk-tassilo/**](http://www.odbms.org/2017/03/qa-with-data-scientists-dirk-tassilo/)

*Q. How do you ensure data quality?*

Understanding the data at hand by visual inspection. Ideally, browse through the raw data manually since our brain is a super powerful outlier detection apparatus. Do not try to check every value, just get an idea of how the raw data actually looks! Then, looking at the basic statistical moments (e.g. numbers and boxplots) to get a feeling how the data looks like. Once patterns are identified, parsers can be derived that apply certain rules to incoming data in a productive system.

*Q. How do you know when the data sets you are analyzing are “large enough” to be significant?*

Very important! I understand the question like this: how do you know that you have enough samples? There is not a single formula for this, however in classification this heavily depends on the amount and distribution of classes you try to classify. Coming from a performance analysis point of view, one should ask how many samples are required in order to successfully perform *n*-fold cross-validation. Then there is extensive work on permutation testing of machine learning performance results. Of course, Cohen’s d for effect size and or p-statistics deliver a framework for such assessment.

Not to make too much of advertisement, but I wrote exactly about this [article](http://journal.frontiersin.org/article/10.3389/fnins.2016.00244/full) in Section 2.5.

**Wolfgang Steitz:**

*Q. How do you ensure data quality?*

It’s good practice to start with some exploratory data analysis before jumping to the modeling part. Doing some histograms and some time series is often enough to get a feeling for the data and know about potential gaps in the data, missing values, data ranges, etc. In addition, you should know where the data is coming from and what transformations it went through. Ones you know all this, you can start filling the gaps and cleaning your data. Eventually there is even another data set you want to take into account. For some model running in production, it’s a good idea to automate some data quality checks. These tests could be as simple as checking if the values are in the correct range or if there are any unexpected missing values. And of course someone should be automatically notified if things go bad.

*Q. How do you evaluate if the insight you obtain from data analytics is “correct” or “good” or “relevant” to the problem domain?*

Presenting results to some domain experts and your customers usually helps. Try to get feedback early in the process to make sure you are working in the right direction and the results are relevant and actionable. Even better, collect expectations first to know how your work will be evaluated later-on.

**Paolo Giudici:**[**http://www.odbms.org/2017/03/qa-with-data-scientists-paolo-giudici/**](http://www.odbms.org/2017/03/qa-with-data-scientists-paolo-giudici/)

*Q. How do you ensure data quality?*

For unsupervised problems: checking the contribution of the selected data to between groups heterogeneity and within groups homogeneity; For supervised problems: checking the predictive performance of the selected data.

*Q. How do you evaluate if the insight you obtain from data analytics is “correct” or “good” or “relevant” to the problem domain?*

By testing its out-of-sample predictive performance we can check if it is correct. To check its relevance, the insights must be matched with domain knowledge models or consolidated results.

*Q. What are the typical mistakes done when analyzing data for a large scale data project? Can they be avoided in practice?*

Forget data quality and exploratory data analysis, rushing to the application of complex models. Forgetting that pre-processing is a key step, and that benchmarking the model versus simpler ones is always a necessary pre requisite.

*Q. How do you know when the data sets you are analyzing are “large enough” to be significant?*

When estimations and/or predictions become quite stable under data and/or model variations.

**Andrei Lopatenko:**[**http://www.odbms.org/2017/03/qa-with-data-scientists-andrei-lopatenko/**](http://www.odbms.org/2017/03/qa-with-data-scientists-andrei-lopatenko/)

*Q. How do you ensure data quality?*

Data quality is not enough, it must be automatically checked. In real world applications it rarely happens that you get data once. Frequently you get a stream of data. If you build an applications about local business, you get a stream of data from provided of data about businesses. If you build an ecommerce site, then you get regular data updates from merchants, and other data providers. The problem is that you can almost never be sure in data quality. In most cases data are dirty.

You have to protect your customers from dirty data. You have to work to discover what problems with data you might have. Frequently problems are not trivial. Sometimes you can see them browsing data directly, frequently toy can not.

For example, in case of local business latitude longitude coordinates might be wrong because provided has a bad data geocoding system. Sometime you do not see problems with data immediately, but only after using them for training some models, where errors are accumulated and lead to wrong results and you have trace back what was wrong.

To ensure data quality once I understand what problems may happen, I build data quality monitoring software. At every step of data processing pipelines I embed tests, you may compare them with unit tests for traditional software development which checks quality of data. They may check total amount of data, existence or non existence of certain values, anomalies in data, compare data to data from previous batch and so on. It required significant error to build data quality tests, but it pays back, they protect from errors in data engineering, data science, incoming data, some system failures , it always pays back. From my experience almost every company build a set of libraries and c[ode alike](http://blog.cloudera.com/blog/2015/07/how-to-do-data-quality-checks-using-apache-spark-dataframes/) to ensue data quality control. We did it in Google, we did it in Apple, we did it Walmart. In the Recruit Institute of Technology we work on [Big Gorilla tools set](http://www.biggorilla.org/) , which will include our open source software and references to other open source software which may help companies build data quality pipelines

*Q. How do you evaluate if the insight you obtain from data analytics is “correct” or “good” or “relevant” to the problem domain?*

Most frequently companies have some important metrics which describe company business. It might be the average revenue per session, the conversion rate, precision of the search engine etc. And your data insights are as good as they improve this metrics. Assume in e-commerce company, the main metrics is average revenue per session (ARPS). And you work on a project of improving extraction of a certain item attribute, for example, from non-structured text.

Questions to ask yourself, will it help to improve ARPS by improving search because it will increase relevance for queries with color intents or faceted queries by color, or by providing better snippets, or by still other means. When one metric does not describe company business and many numbers are needed to understand it. Your data projects might be connected to other metrics. But what’s important is to connect your data insight project to metrics which are representative of company business and improvement of these metrics will be as a significant impact to the company business. Such connection makes a good project.

*Q. What are the typical mistakes done when analyzing data for a large scale data project? Can they be avoided in practice?*

Typical mistake – assuming that data are clean. Data quality should be examined and checked.

 **Mike Shumpert:**[**http://www.odbms.org/2017/03/qa-with-data-scientists-mike-shumpert/**](http://www.odbms.org/2017/03/qa-with-data-scientists-mike-shumpert/)

*Q. How do you ensure data quality?*

On the one hand, one of the basic tenets of “big data” is that you *can’t* ensure data quality – today’s data is voluminous and messy, and you’d better be prepared to deal with it. As mentioned before, “dealing with it” can simply mean throwing some instances out, but sometimes what you think is an outlier could be the most important information you have.

So if you want to enforce at least *some* data quality, what can you do? It’s useful to think of data as comprising two main types: transactional or reference. Transactional data is time-based and constantly changing – it typically conveys that something just happened (e.g., customer checkouts), although it can also be continuous data sampled at regular intervals (e.g., sensor data). Reference data changes very slowly and can be thought of as the properties of the object (customer, machine, etc.) at the center of the prediction.

Both types of data typically have predictive value: *this amount at this location was just spent (transactional) by a platinum-level female customer (reference) – is it fraud?* But the two types often come from different sources and can be treated differently in terms of data quality.

Transactional data can be filtered or smoothed to remove transitory outliers, but the problem domain will determine whether or not any such anomalies are noise or real (and thus very important). For example, the $10,000 purchase on a credit card with a typical maximum of $500 is one that deserves further scrutiny, not dismissal.

But reference data can be separately cleansed and maintained via [Master Data Management](https://en.wikipedia.org/wiki/Master_data_management)(MDM) technology. This ensures there is only one version of the truth with respect to the object at the core of the prediction and prevents nonsensical changes such as a customer moving from gold status to platinum and back again within 30 seconds. Clean reference data can then be merged with transactional data on the fly to ensure accurate predictions.

Using an [Internet of Things](https://en.wikipedia.org/wiki/Internet_of_things) (IoT) example, consider a predictive model for determining when a machine needs to be serviced. The model will want to leverage all the sensor data available, but it will also likely find useful factors such as the machine type, date of last service, country of origin, etc. The data stream coming from the sensors usually will not carry with it the reference data and will probably only provide a sensor id. That id can be used to look up relevant machine data and enrich the data stream on the fly with all the features needed for the prediction.

One final point on this setup is that you do not want to go back to the original data sources of record for this on-the-fly enrichment of transactional data with reference data.

You want the cleansed data from the MDM system, and you want that stored in memory for high-performance retrieval.

**Romeo Kienzler:**[**http://www.odbms.org/2017/03/qa-with-data-scientists-romeo-kienzler/**](http://www.odbms.org/2017/03/qa-with-data-scientists-romeo-kienzler/)

*Q. How do you ensure data quality?*

This is again a vote for domain knowledge. I have someone with domain skills assess each data source manually. In addition I gather statistics on the accepted data sets so some significant changes will raise an alert which – again – has to be validated by a domain expert.

Q. How do you evaluate if the insight you obtain from data analytics is “correct” or “good” or “relevant” to the problem domain?

I’m using the classical statistical performance measures to assess the performance of a model. This is only about the mathematical properties of a model. Then I check with the domain experts on the significance to their problems. Often a statistically significant result is not relevant for the business. E.g. if I tell you that a bearing will break with 95% probability within the next 6 months might not really help the PMQ ([Predictive Maintenance](https://en.wikipedia.org/wiki/Predictive_maintenance) and Quality) guys. So the former can be described as “correct” or “good” whereas the latter as “relevant” maybe.

**Elena Simperl:**[**http://www.odbms.org/2017/02/qa-with-data-scientists-elena-simperl/**](http://www.odbms.org/2017/02/qa-with-data-scientists-elena-simperl/)

*Q. How do you ensure data quality?*

It is not possible to “ensure” data quality, because you cannot say for sure that there isn’t something wrong with it somewhere. In addition, there is also some research which suggests that compiled data are inherently filled with the (unintentional) bias of the people compiling it. You can attempt to minimise the problems with quality by ensuring that there is full provenance as to the source of the data, and err on the side of caution where some part of it is unclassified or possibly erroneous.

One of the things we are researching at the moment is how best to leverage the wisdom of the crowd for ensuring quality of data, known as crowdsourcing. The existence of tools such as [Crowdflower](https://www.crowdflower.com/) makes it easy to organise a crowdsourcing project, and we have had some level of success in image understanding, social media analysis, and data integration Web. However, the best ways of optimising cost, accuracy or time remain to be determined and are different relative to the particular problem or motivation of the crowd one works with.

*Q. How do you evaluate if the insight you obtain from data analytics is “correct” or “good” or “relevant” to the problem domain?*

This question links back to a couple of earlier questions nicely. The importance of having good enough domain knowledge comes into play in terms of answering the relevance question. Hopefully a data scientist will have a good knowledge of the domain, but if not then they need to be able to understand what the domain expert believes in terms of relevance to the domain. The correctness or value of the data then comes down to understanding how to evaluate machine learning algorithms in general, and using domain knowledge to apply to decide whether the trade-offs are appropriate given the domain.

**Mohammed Guller:**[**http://www.odbms.org/2017/02/qa-with-data-scientists-mohammed-guller/**](http://www.odbms.org/2017/02/qa-with-data-scientists-mohammed-guller/)

*Q. How do you ensure data quality?*

It is a tough problem. Data quality issues generally occur upstream in the data pipeline. Sometimes the data sources are within the same organization and sometimes data comes from a third-party application. It is relatively easier to fix data quality issues if the source system is within the same organization. Even then, the source may be a legacy application that nobody wants to touch. So you have to assume that data will not be clean and address the data quality issues in your application that processes data. Data scientists use various techniques to address these issues. Again, domain knowledge helps.

*Q. How do you evaluate if the insight you obtain from data analytics is “correct” or “good” or “relevant” to the problem domain?*

This is where domain knowledge helps. In the absence of domain knowledge, it is difficult to verify whether the insight obtained from data analytics is correct. A data scientist should be able to explain the insights obtained from data analytics. If you cannot explain it, chances are that it may be just a coincidence. There is an old saying in machine learning, “if you torture data sufficiently, it will confess to almost anything.”

Another way to evaluate your results is to compare it with the results obtained using a different technique. For example, you can do backtesting on historical data. Alternatively, compare your results with the results obtained using incumbent technique. It is good to have a baseline against which you can benchmark results obtained using a new technique.

**Natalino Busa:**

*Q. How do you ensure data quality?*

I tend to rely on the “wisdom of the crowd” by implementing similar analysis using multiple techniques and machine learning algorithms. When the results diverge, I compare the methods to gain any insight about the quality of both data as well as models. This technique works also well to validate the quality of streaming analytics: in this case the batch historical data can be used to double check the result in streaming mode, providing, for instance, end-of-day or end-of-month reporting for data correction and reconciliation.

*Q. How do you evaluate if the insight you obtain from data analytics is “correct” or “good” or “relevant” to the problem domain?*

Most of the time I interact with domain experts for a first review on the results. Subsequently, I make sure than the model is brought into “action”. Relevant insight, in my opinion, can always be assessed by measuring their positive impact on the overall application. Most of the time, as human interaction is part of the loop, the easiest method is to measure the impact of the relevant insight in their digital journey.

**Vikas Rathee:**

*Q. How do you ensure data quality?*

Data quality is very important to make sure the analysis is correct and any predictive model we develop using that data is good. Very simply I would do some statistical analysis on the data, create some charts and visualize information. I also will clean data by making some choice at the time of data preparation. This would be part of the feature engineering stage that needs to be done before any modeling can be done.

*Q. How do you evaluate if the insight you obtain from data analytics is “correct” or “good” or “relevant” to the problem domain?*

Getting Insights is what makes the job of a Data Scientist interesting. In order to make sure the insights are good and relevant we need to continuously ask ourselves what is the problem we are trying to solve and how it will be used. In simpler words, to make improvements in existing process we will need to understand the process and where the improvement is requirement or of most value. For predictive modeling cases, we need to ask how the output of the predictive model will be applied and what additional business value can be derived from the output. We also need to convey what does the predictive model output means to avoid incorrect interpretation by non-experts.

Once the context around a problem has been defined and we proceed to implement the machine learning solution. The immediate next stage is to verify if the solution will actually work. There are many techniques to measure the accuracy of predictions i.e. testing with historic data samples using techniques like k-fold cross validation, confusion matrix, r-square, absolute error, MAPE (Mean absolution percentage error), p-value etc. We can choose from among many models which show most promising results. There are also ensemble algorithms which generalize the learning and avoid being over fit models.

**Christopher Schommer:**[**http://www.odbms.org/2017/01/qa-with-data-scientists-christopher-schommer/**](http://www.odbms.org/2017/01/qa-with-data-scientists-christopher-schommer/)

*Q. How do you ensure data quality?*

To keep a data quality is mostly an adaptive process, for example, because provisions of national law may change or because the analytical aims and purposes of the data owner may vary. Therefore, the ensuring of a data quality should be performed regularly, it should be consistent with the law (data privacy aspects and others), and should be commonly performed by a team of experts of different education levels (e.g., data engineers, lawyers, computer scientists, mathematicians).

*Q. How do you evaluate if the insight you obtain from data analytics is “correct” or “good” or “relevant” to the problem domain?*

In my understanding, an insight is already a valuable/evaluated information, which has been received after a detailed interpretation and which can be used for any kind of follow-up activities, for example to relocate the merchandise or to deeper dig in clusters showing a fraudulent behavior.

However, it is less oportune to rely only on statistical values: an association rule, which shows a conditional probability of, e.g., 90% or more, may be an “insight”, but if the right-hand side of the rule refers to a plastic bag only (which is to be paid (3 cents), at least in Luxembourg), the discovered pattern might be uninteresting.

**Slava Akmaev:**[**http://www.odbms.org/2017/01/qa-with-data-scientists-slava-akmaev/**](http://www.odbms.org/2017/01/qa-with-data-scientists-slava-akmaev/)

*Q. How do you evaluate if the insight you obtain from data analytics is “correct” or “good” or “relevant” to the problem domain?*

In a data rich domain, evaluation of the insight correctness is done either by applying the mathematical model to new “unseen” data or using cross-validation. This process is more complicated in human biology. As we have learned over the years, a promising cross-validation performance may not be reproducible in subsequent experimental data. The fact of the matter is, in life sciences, laboratory validation of computational insight is mandatory. The community perspective on computational or statistical discovery is generally skeptical until the novel analyte, therapeutic target, or biomarker is validated in additional confirmatory laboratory experiments, pre-clinical trials or human fluid samples

**Jochen Leidner:**[**http://www.odbms.org/2017/01/qa-with-data-scientists-jochen-leidner/**](http://www.odbms.org/2017/01/qa-with-data-scientists-jochen-leidner/)

*Q. How do you ensure data quality?*

There are a couple of things: first, make sure you know where the data comes from and what the records actually mean. Is it a static snapshot that was already processed in some way, or does it come from the primary source. Plotting histograms and profiling data in other ways is a good start to find outliners and data gaps that should undergo imputation (filling of data gaps with reasonable fillers). Measureing is key, so doing everything from inter-annotator agreement of the gold data over training, dev-test and test evaluations to human SME output grading consistently pays back the effort.

*Q. How do you evaluate if the insight you obtain from data analytics is “correct” or “good” or “relevant” to the problem domain?*

There is nothing quite as good as asking domain experts to vet samples of the output of a system. While this is time consuming and needs preparation (to make their input actionable), the closer the expert is to the real end user of the system (e.g. the customer’s employees using it day to day), the better.

**Claudia Perlich:**[**http://www.odbms.org/2016/11/qa-with-data-scientists-claudia-perlich/**](http://www.odbms.org/2016/11/qa-with-data-scientists-claudia-perlich/)

*Q. How do you ensure data quality?*

The sad truth is – you cannot. Much is written about data quality and it is certainly a useful relative concept, but as an absolute goal it will remain an unachievable ideal (with the irrelevant exception of simulated data …).

First of, data quality has many dimensions.

Secondly – it is inherently relative: the exact data can be quite good for one purpose and terrible for another. Third, data quality is a very different concept for ‘raw’ event log data vs. aggregated and processed data. Finally, and this is by far the hardest part: you almost never know what you don’t know about your data. In the end, all you can do is your best! Scepticism, experience, and some sense of data intuition are the best sources of guidance you will have.

*Q. How do you evaluate if the insight you obtain from data analytics is “correct” or “good” or “relevant” to the problem domain?*

First of, one should not even have to ask whether the insight is relevant – one should have designed the analysis that led to the insight based on the relevant practical problem one is trying to solve! The answer might be that there is nothing better you can do than status quo. That is still a highly relevant insight! It means that you will NOT have to waste a lot or resources. Taking negative answer into account as ‘relevant’ – if you are running into this issue of the results of data science not being relevant you are clearly not managing data science correctly. I have commented on this here: [What are the greatest inefficiencies data scientists face today?](https://www.quora.com/What-are-the-greatest-inefficiencies-data-scientists-face-today)

Let’s look at ‘correct’ next. What exactly does it mean? To me it somewhat narrowly means that it is ‘true’ given the data: did you do all the due diligence and right methodology to derive something from the data you had? Would somebody answering the same question on the same data come to the same conclusion (replicability)? You did not overfit, you did not pick up a spurious result that is statistically not valid, etc. Of course you cannot tell this from looking at the insight itself. You need to evaluate the entire process (or trust the person who did the analysis) to make a judgement on the reliability of the insight.

Now to the ‘good’. To me good captures the leap from a ‘correct’ insight on the analyzed dataset to supporting the action ultimately desired. We do not just find insights in data for the sake of it! (well – many data scientists do, but that is a different conversation). Insights more often than not drive decisions. A good insight indeed generalizes beyond the (historical) data into the future. Lack of generalization is not just a matter of overfitting, it is also a matter of good judgement whether there is enough temporal stability in the process to hope that what I found yesterday is still correct tomorrow and maybe next week. Likewise we often have to make judgement calls when the data we really needed for the insight is simply not available. So we look at a related dataset (this is called transfer learning) and hope that it is similar enough for the generalization to carry over. There is no test for it! Just your gut and experience …

Finally, good also incorporates the notion of correlation vs. causation. Many correlations are ‘correct’ but few of them are good for the action one is able to make. The (correct) fact that a person who is sick has temperature is ‘good’ for diagnosis, but NOT good for prevention of infection. At which point we are pretty much back to relevant! So think first about the problem and do good work next!

**Ritesh Ramesh:**[**http://www.odbms.org/2016/11/qa-with-data-scientists-ritesh-ramesh/**](http://www.odbms.org/2016/11/qa-with-data-scientists-ritesh-ramesh/)

*Q. How do you ensure data quality?*

Data Quality is critical. We hear often from many of our clients that ensuring trust in the quality of information used for analysis is a priority. The thresholds and tolerance of data quality can vary across problem domains and industries but nevertheless data quality and validation processes should be tightly integrated into the data preparation steps. Data scientists should have full transparency on the profile and quality of datasets that they are working with and have tools at their disposal to remediate fixes with proper governance and procedures as necessary. Emerging data quality technologies are embedding leverages machine learning features to detect proactive data errors and make data quality a business-user friendly and an intelligent function more than ever it has been for years

*Q. How do you evaluate if the insight you obtain from data analytics is “correct” or “good” or “relevant” to the problem domain?*

Many people view Analytics and Data science as some magic crystal ball into the future events and don’t realize that it is just one of many probable indicators for successful outcomes – If the model predicts that a there’s an 80% chance of success, you also need to read it as there’s still a 20% chance of failure. To really assess the ‘quality’ of of insights from the model you may start with the below areas –

1) Assess whether the model makes reasonable assumptions on the problem domain and takes into account all the relevant input variables and business context – I was recently reading an article on a U.S. based insurer who implemented an analytics model that looked for number of unfavorable traffic incidents to assess risk on the vehicle driver but they missed out on assigning weights to the severity of the traffic incident. If your model makes wrong contextual assumptions – the outcomes can backfire

2) Assess whether the model is run on a sufficient sample of datasets. Modern scalable technologies have made executing analytical models on massive amounts of data possible.

More data the better although every problem does not need large datasets of the same kind

3) Assess where extraneous events like macroeconomic events, weather, consumer trends etc. are considered in the model constraints. Use of external data sets with real time API based integrations is highly encouraged since it adds more context to the model

4) Assess the quality of data used as an input to the model. Feeding wrong data to a good analytics model and expecting it to produce the expected outcomes is an unreasonable expectation. The stakes are higher in high regulatory environments where minimal error in the model might mean millions of dollars of lost revenues or penalties Even successful organizations who execute seamlessly in generating insights struggle to “close the loop” in translating the insights into the field to drive shareholder value. It’s always a good practice to pilot the model on a small population, link its insights and actions to key operational and financial metrics, measure the outcomes and then decide whether to improve or discontinue the model

**Richard J Self:**[**http://www.odbms.org/2016/11/qa-with-data-scientists-richard-j-self/**](http://www.odbms.org/2016/11/qa-with-data-scientists-richard-j-self/)

*Q. How do you ensure data quality?*

Data Quality is a fascinating question. It is possible to invest enormous levels of resource into attempting to ensure near perfect data quality and still fail.

The critical question should, however, start from the Governance perspective of questions such as:

1. What is the overall business Value of the intended analysis?
2. How is the Value of the intended insight affected by different levels of data quality (or Veracity)?
3. What is the level of Vulnerability to our organisation (or other stakeholders) if the data is not perfectly correct (see J Easton of IBM comment above) in terms of reputation, or financial consequences?

Once you have answers to those questions and the sensitivities of your project to various levels of data quality, you will then begin to have an idea of just what level of data quality you need to achieve. You will also then have some ideas about what metrics you need to develop and collect, in order to guide your data ingestion and data cleansing and filtering activities.

Q. How do you evaluate if the insight you obtain from data analytics is “correct” or “good” or “relevant” to the problem domain?

The answer to this returns to the Domain Expert question. If you do not have adequate domain expertise in your team, this will be very difficult.

Referring back to the USA election, one of the more unofficial pollsters, who got it pretty well right observed that he did it because he actually talked to real people. This is domain expertise and Small Data.

All the official polling organisations have developed a total trust in Big Data and Analytics, because it can massively reduce the costs of the exercise. But they forget that we all lie un-remittingly on line. See the first of the “All Watched Over by Machines of Loving Grace” documentaries at <https://vimeo.com/groups/96331/videos/80799353> to get a flavour of this unreasonable trust in machines and big data.

————

* ConsigliaOn Data Quality
* Commenta
* CondividiCondividi On Data Quality

[Segnala](https://www.linkedin.com/pulse/data-quality-roberto-v-zicari)

SeguiRoberto V. Zicari

****

[**Roberto V. Zicari**](https://www.linkedin.com/in/roberto-v-zicari-087863)

University Professor- Goethe University Frankfurt

* [**39 articoli**](https://www.linkedin.com/today/author/roberto-v-zicari-087863)

**6 commenti**

****

[**Ernesto Zumelzu**](https://www.linkedin.com/in/ernesto-zumelzu-a92b7611)

Full Professor, Researcher, Management of Universidad Austral de Chile

Very interesting opinions at a time when each day is valued more data to transform them into information. An important aspect of scientific activity is not to commit too much to expected data. I think the quality of the data is when it is reproducible and statistically validated in samples.

Consiglia

Rispondi

5 giorni

****

[**Ron S. Kenett**](https://www.linkedin.com/in/ronkenett)

Senior Research Fellow at Samuel Neaman Institute, the Technion, Haifa

This bottom up contribution is very nice. None of the interviews bring out the concept of information quality. The path from numbers -> data (numbers with a context) -> information (data with analysis) -> insights, is indeed blurred in the mind of many data scientists (and data science programs). For a comprehensive treatment of information quality (InfoQ), including reference to our new book, papers and presentations, see http://infoq.galitshmueli.com, for what experts say about our book see https://www.linkedin.com/pulse/what-experts-say-information-quality-potential-data-analytics-kenett