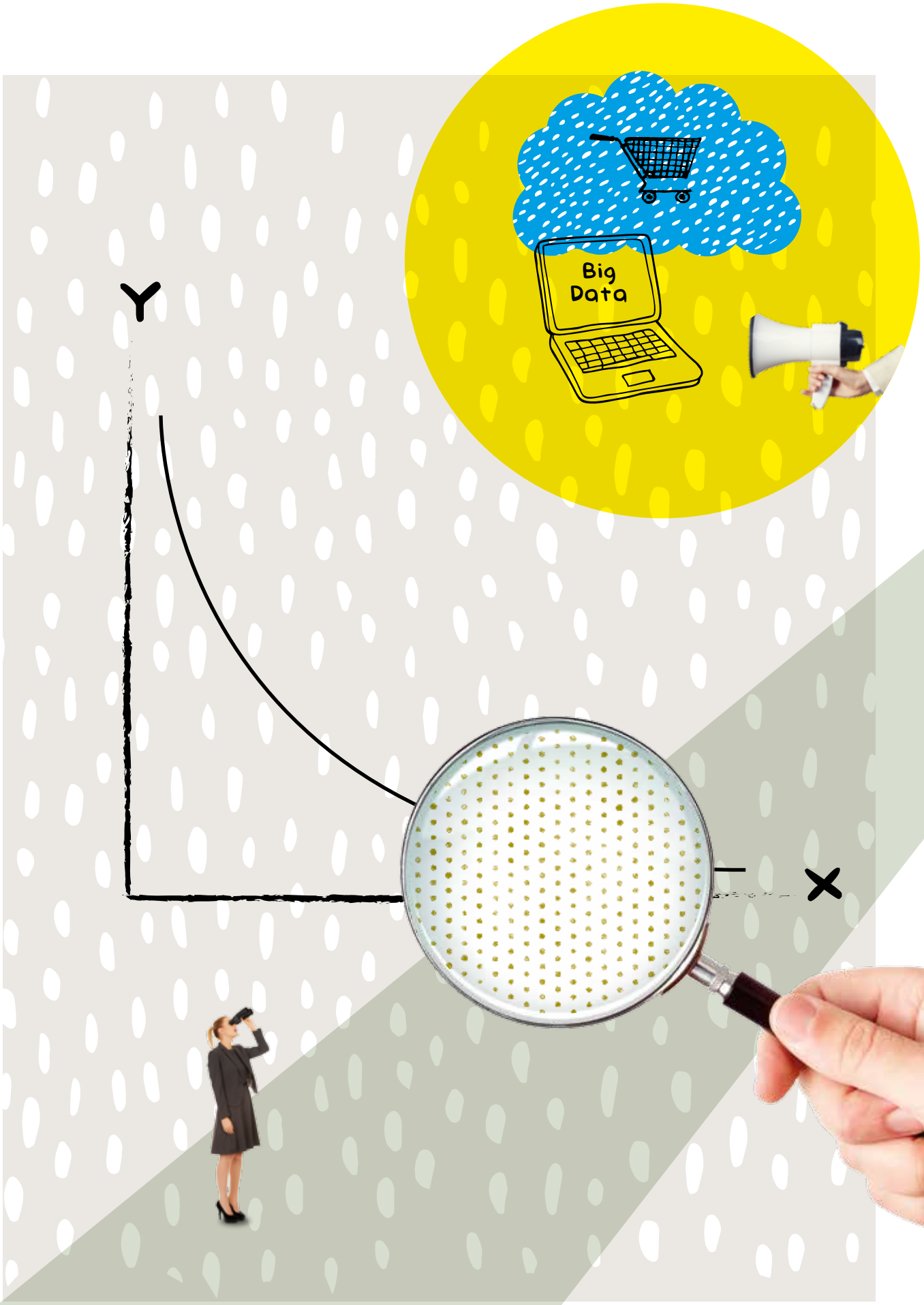

Big
Data
Y
X

Big Data in Market Research: Why More Data Does Not Automatically Mean Better Information

Volker Bosch

KEYWORDS

Big Data, Market Research, Information, Representativeness, Data Integration, Data Imputation

THE AUTHOR

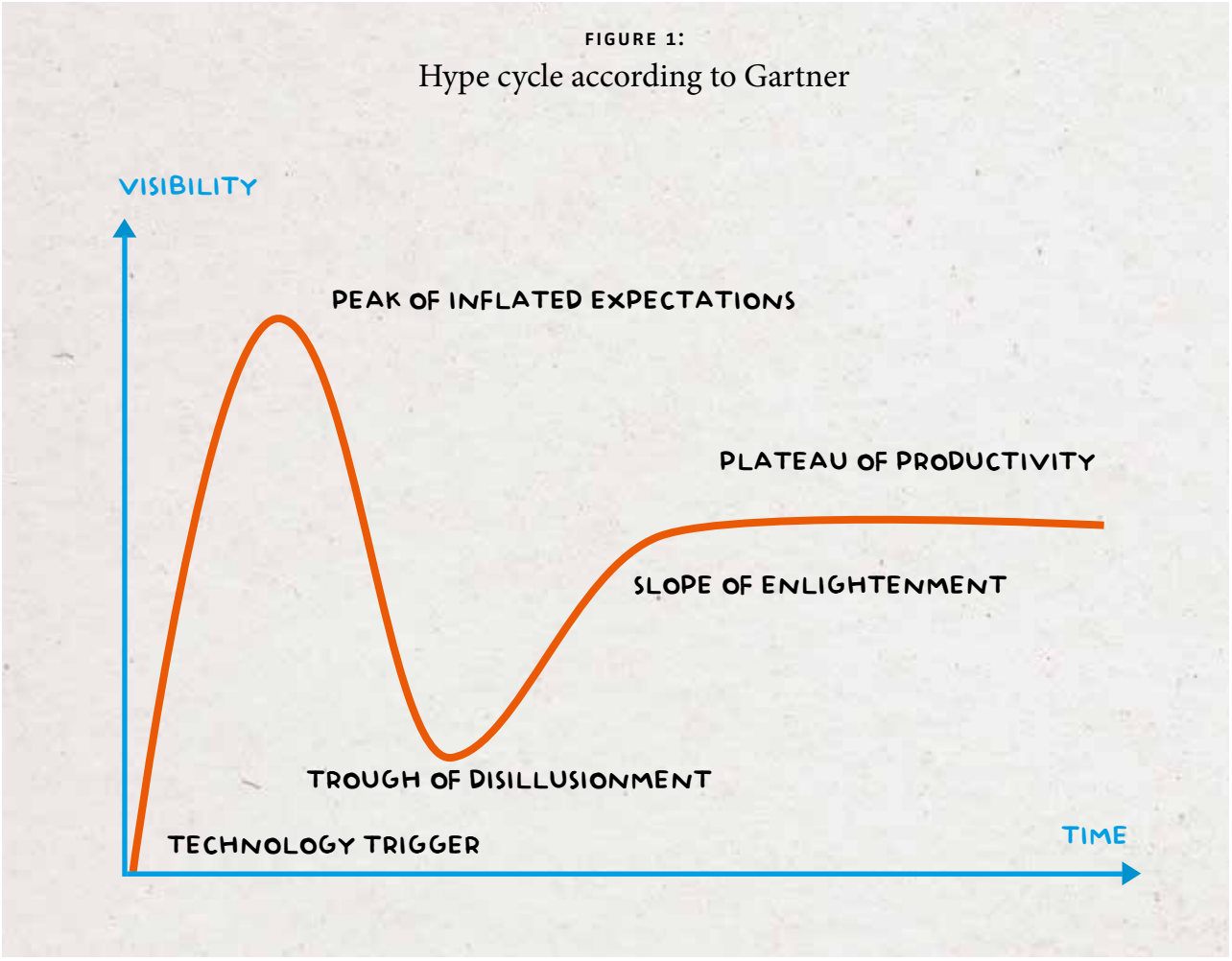
Volker Bosch,
Head of Marketing & Data Sciences,
GfK SE, Nuremberg, Germany;
volker.bosch@gfk.com

Everyone is Talking about Big Data /// "Big data" is a megatrend, although not everyone means the same thing when they talk about it. Generally speaking, it involves enormous amounts of data, which are generated almost automatically with the help of the latest technological developments, and examines how this data can be converted into useful information. Big data has raised big expectations, but the profitable monetization of data usually turns out to be much more complex than anticipated. That new technological developments are no guarantee for a start-to-finish victory is nothing new, though. Over time, many conform to a typical pattern which was first described by the consulting firm Gartner in 1995 and has since been referenced in numerous publications as the "hype cycle" (Figure 1).

At this time, big data has probably passed the "peak of inflated expectations" and is still yet to cross the "trough of disillusionment" before reaching the "plateau of productivity". But what are realistic expectations for big data? And what does big data mean for the market research industry?

Big Data Conquers Market Research /// Big data will change market research at its core in the long term. While other trends such as neuromarketing have not been able to gain a substantial foothold, big data business models will assume a central role in the value chain. The consumption of products and media can be logged electronically more and more, making it measurable on a large scale. In some areas of market research, big data is already established today, with social media analytics and the use of cookie data to measure internet coverage being two prominent examples. The use of panels for the passive measurement of media consumption through the internet, television, and radio also falls under

FIGURE 1:
Hype cycle according to Gartner



big data. But what additional benefits does big data provide compared to traditional market research data?

Big Data = Passive Measurement /// The 4V definition describes the core characteristics of big data: volume, velocity, variability (of the data structure) and (questionable) veracity. However, 4V doesn't tell the whole story, because the origin of the data is especially decisive. New sensor technologies and processing architectures enable entirely new possibilities for gathering and processing information. We are dealing with a fundamental paradigm shift – in traditional market research, data is collected actively, e.g. through human interaction or interviews. With big data, by contrast, the information no longer needs to be processed by slow, limited-capacity, mistake-prone and emotional human brains for a dataset to be created. As a result, passive measurement is the actual driver of the efficiency of big data in market research. It creates economies of scale that were formerly the stuff of dreams.

»
Passive measurement is the actual driver of the efficiency of big data in market research. It creates economies of scale that were formerly the stuff of dreams.
«

Defining big data based as passive measurement does not necessarily mean massive amounts of data. Equipping just a few measured units with sensors can already create datasets that are hard to manage. Examples include the software-based measurement of internet behavior in a panel or equipping shopping carts with RFID technology to transmit the precise location and purchase history of a customer in a supermarket. Perhaps it would be more appropriate to speak of "new data" than "big data."

Is Twice as Much Data Worth Twice as Much? /// The size of a typical big data dataset leads to the false assumption that it provides a correspondingly large amount of information. From an organizational perspective that is absolutely correct, but from a statistical perspective, it is wrong, because "information" in statistical analysis is defined as the reduction of uncertainty. But twice the data does not mean twice the accuracy, only an improvement by a factor of 1.4, as measured by the confidence interval of a sample. Marginal utility declines significantly as data amounts increase. Ignoring declining marginal utility will almost certainly result in overestimating the value of big data for market research and overlooking its actual benefits.

»

Big data can be used like a microscope to see structures that would appear blurred with conventional market research or even be entirely unrecognizable.

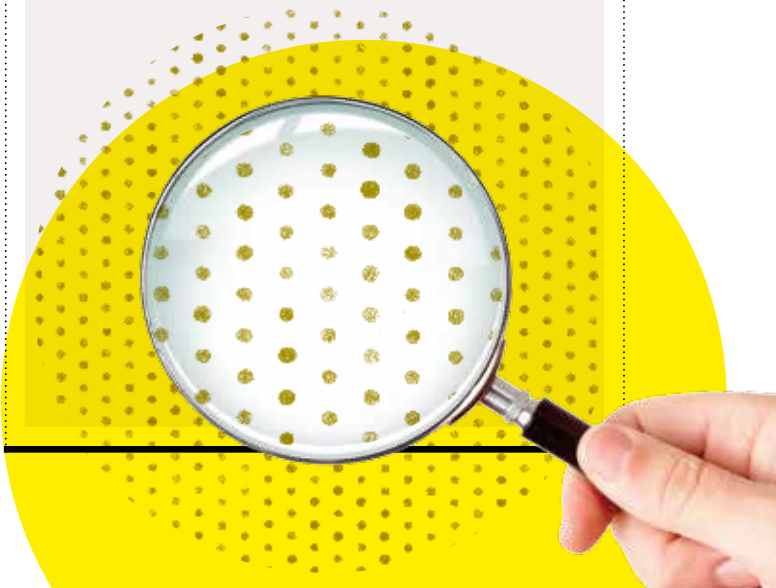
«

In visual terms, a larger amount of data results in greater statistical resolution, enabling structures of finer granularity to be described with statistical validity. Examples are smaller target groups, websites in the "long tail" of the internet, or rare events. Big data can be used like a microscope to see structures that would appear blurred with conventional

{ *Box 1* }

MORE DATA ≠ BETTER DATA: THE LITERARY DIGEST FARCE

In 1936, the leading news magazine in the 1930s, the Literary Digest, delivered a very clear prediction for the outcome of the presidential elections. It was based on an extensive mail and telephone survey using sources available at the time, the telephone book and a list of car owners. 2.4 million citizens participated and a clear victory for Alf Landon, the challenger to the incumbent F.D. Roosevelt, was predicted. For the standards of the day, that was certainly big data, even though there were no methods for passive measurement in 1936. But based on a much smaller sample of about 50,000 persons, George Gallup predicted the exact opposite. After analyzing the non-representative subsamples of telephone and car owners, he predicted that the Literary Digest forecast would be wrong. He was vindicated, and shortly after this glaringly wrong prediction, the Literary Digest had to cease publication.



market research or even be entirely unrecognizable. In other words, the declining marginal utility is mitigated by the fact that with big data, structures of the finest granularity can be defined in the midst of the statistical noise.

Big Data Must Be Scientifically Evaluated in Market Research /// In direct marketing, with CRM systems, or in intelligence agencies, it mainly comes down to describing individual characteristics. In market research, by contrast, the aim is to find valid, generalizable statements based on scientific standards. When analyzing the use of products and media by populations and their segments, it must also be possible to describe statistical errors. This has a decisive impact on the applied algorithms and processes. Statistical methods, data integration, weighting, variable transformations, and issues of data protection present much larger challenges than was previously the case with conventional datasets. In particular, the three following challenges must be overcome.

> ***Challenge 1: Big Data is (Almost Always) not Representative*** /// Massive amounts of data do not necessarily result in good data and more does not automatically mean better. Big data can easily tempt us to fall into the same "more is better" trap that the Literary Digest fell into back in 1936 (see box 1). It is a truism of market research that sample bias cannot be reduced by more of the same and that the representativeness problem remains. Consequently, traditional topics of sampling theory like stratification and weighting are highly topical in the age of big data, and must be reinterpreted. Rarely is it possible to measure all units of interest and avoid bias. For example, the scope of interpretation for social media analysis is restricted because the silent majority usually cannot be observed. This may explain why social media data often behaves unexpectedly in predictive models. By using smart algorithms, however, it is possible to achieve astounding precision with non-representative digital approaches, as for example in the elections in the United States in 2012 and in Great Britain in 2015.

> ***Challenge 2: Big Data is (Almost Always) Flawed*** /// The passive measurement of behavior along with its proxies and the high level of technology being used may lead to the false assumption that practically no measurement error exists and that data can be further processed without hesitation. But that is very rarely the case. Such technologies are highly complex and often not designed for use in market research. Big data must be processed with very complex and therefore error-prone software and many measurement errors arise. In addition, the internet ecosystem is subject to constant updates (in the best case) or to changes in technology (in the worst case): Internet Explorer yields to Edge, HTML5 replaces the old HTML4, http pages turn into https, or Flash is no longer supported. In measuring internet behavior in the GfK Cross-Media-Link panel, we observed how browser updates, technological upgrades, changes in website behavior, and end-of-life systems can lead to a measurement failure. If updates occur unannounced and unexpectedly, emerging measurement gaps might even be noticed (too) late.

Things become even more difficult with systems that were originally constructed for another purpose, for example, if mobile internet use is measured by a mobile network operator and not by the market researcher directly. This is referred to as network-centric measurement, in contrast to user-centric measurement in a panel or site-centric measurement using cookies. Data processing capacities in such systems primarily serve to maintain telephone or internet service and for billing. Market research requirements were not even a factor in the original design at all. Therefore, so-called "probes" must be laboriously installed in order to retrieve the relevant information. Control over data quality is limited. Undetected data blackouts frequently occur because the primary tasks of the system take priority and no error routines have been installed for other requirements. GfK discovered this the hard way in its "Mobile Insights" project.

> ***Challenge 3: Big Data (Almost Always) Lacks Important Variables*** /// The biggest market research challenge from a methodological point of view is the limited data depth of big data. Despite the sometimes overwhelming amount of data in terms of observed units, the number of measured variables is low or critical variables are missing. By contrast,

{ Box 2 }

DATA IMPUTATION

In a traditional data matrix, the columns represent variables and the rows represent observation units like people or households. Variables observed in the census are available for all units, while other variables, for example sociodemographic characteristics, can only be collected in a subset, e.g. the panel.

In image data, gray values represent the measured values of a variable (Figure 2). In the example, 75% of the data or image points are missing. Only a few randomly selected rows (panel members=donors) and columns (census data=common variables) are fully observed. In order to ensure that an algorithm cannot use pure image information (the physical proximity of image points), rows and columns are randomly sorted. As a result, the image behaves like a market research dataset and the data can be processed accordingly.

The missing values are filled in by imputation. Many algorithms exist and all work with different assumptions regarding the statistical properties of data. As a common feature they learn from donors how the common variables relate to the specific variables to be transferred and they fill the data gap for the recipients using this knowledge. In the big data context, imputation is particularly difficult because large quantities of data must be processed and finding the optimal model is usually too costly. In addition, the data rarely follows a multivariate normal distribution or other well-defined distributions.

This is why the Marketing & Data Sciences department at GfK developed the "linear imputation" process. It requires a minimum of theoretical assumptions and delivers good results, even for highly non-linear data structures (as in the image) by using local regression models.

In the image example, the quality of the imputation can be judged immediately if the matrix is sorted back to its original order (Figure 3, middle picture).

FIGURE 2:

Data imputation using an image file with random sorting of rows and columns

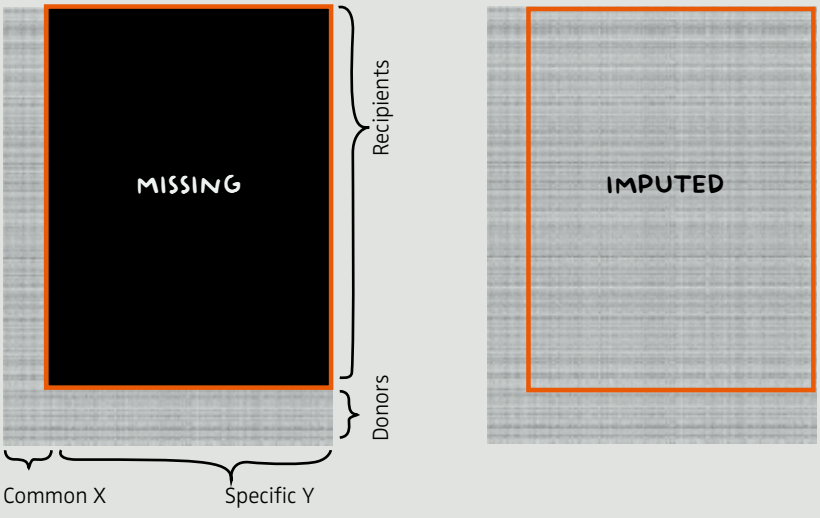
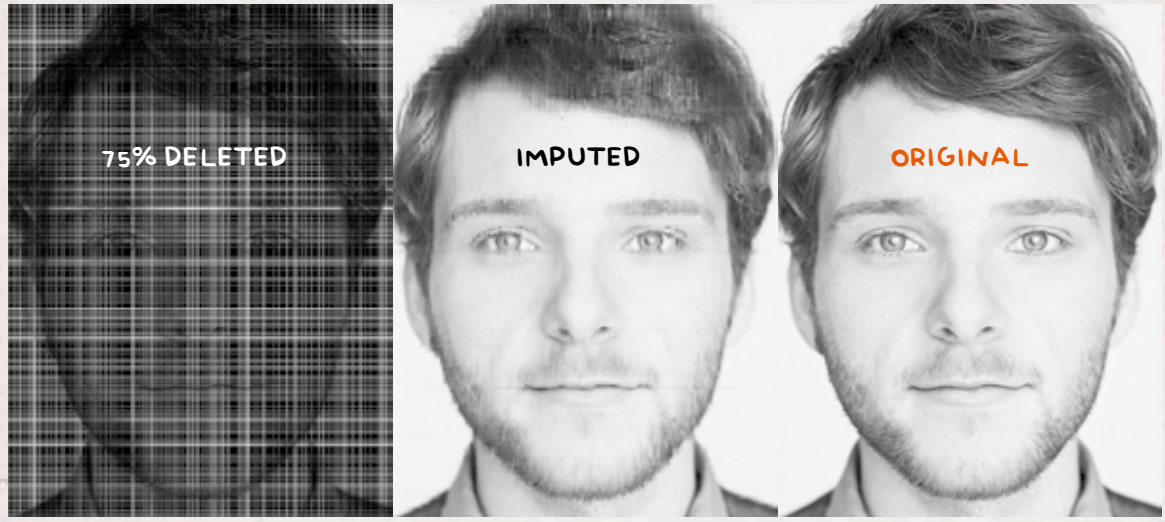


FIGURE 3:

The original sequence of rows and columns was restored following imputation



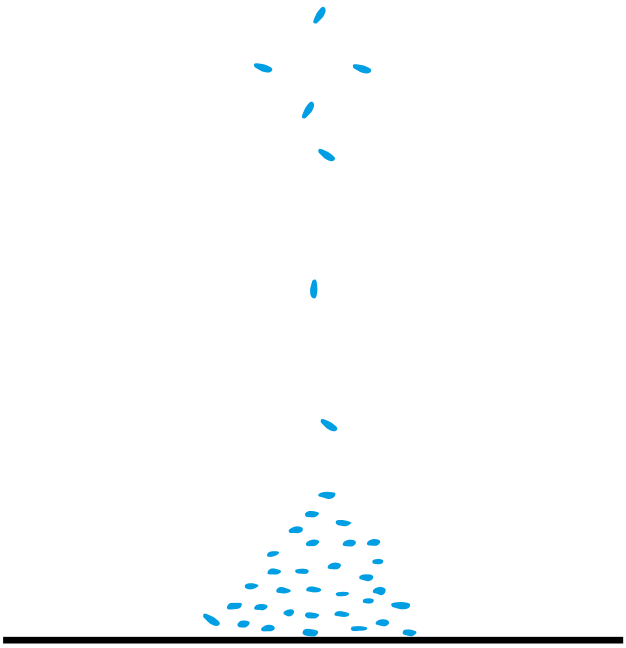
in traditional opinion research the relevant variables are optimized for a specific subject and can be very extensive. Internet coverage research based on cookies or network-centric data illustrates this. Even if almost the whole population is reached, like in a census, critical information is missing, such as sociodemographic data. Therefore the value of the collected data is limited and important evaluations such as target group or segment-specific analyses cannot be conducted. The missing information can only be filled in using statistical data imputation. This requires an additional data source with the additional variables, for example a panel. The source must also contain the variables of the big data dataset. Imputation is a statistical procedure that is anything but trivial. Box 2 describes the underlying logic using an image dataset, which is handled like market research data.

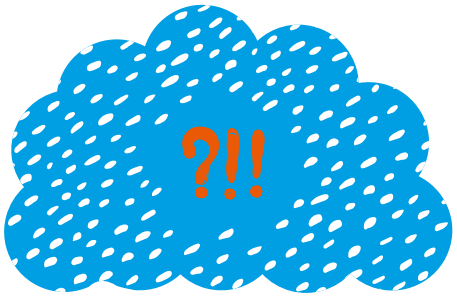
However, imputation is not a tool that lets information be gathered by magic. Statistics do not create information, only observation does. Statistics makes structures visible. Therefore, imputation is an instrument to "transport information" and the higher the observed data correlates with the data to be imputed, the better it works.

More Value through More Data: Also in Market Research

/// Big data poses special challenges for market research. It is by no means sufficient to master the technologies for processing large amounts of data, or to engage in pure "data science". It is also necessary to develop in-house market research algorithms, which can be applied to the new data and successfully address the three challenges of representativeness, measurement errors, and statistical data integra-

»
Despite the sometimes overwhelming amount of data in terms of observed units, the number of measured variables is low or critical variables are missing.
«





tion. Therefore, the young discipline of "data science" must be fused with the classic field of "marketing science" to help market research expand its core business successfully.

And at least as far as applications in market research are concerned, big data is well on its way to the plateau of productivity in the hype cycle of new technologies.

/.

FURTHER READING

Fenn, Jackie (1995):
The Microsoft System Software Hype Cycle Strikes Again

Gaffert P., Bosch V., Meinfelder, F. (2016):
"Interactions and squares. Don't transform, just impute!," Conference Paper, Joint Statistical Meetings, Chicago

http://www.ibmbigdatahub.com/infographic/four-vs-big-data

http://fivethirtyeight.blogs.nytimes.com/2012/11/10/which-polls-fared-best-and-worst-in-the-2012-presidential-race/?_r=0