"And now tonight's business news: irrelevant statistics are up 27.45%, but meaningless figures fell 110%"

Figure 20-1. Useless Analytics
Preconditions of meaningful Analytics

Are precise algorithms and high-quality data all you need?

If you tend to say “yes,” let me try to disappoint you…

Even if you have reliable data and if you use mathematically proven algorithms, the results may not be watertight.

Here are the two main challenges:

- The preconditions for the applicability of an algorithm
- Ambiguous interpretation of rules

Let me use the example of the well-known ANOVA (analysis of variance) from Applied Statistics. This method is used frequently to compare variations of different groups within a sample. I will not go in detail but point out a few potential pitfalls.

Are all preconditions fulfilled?

In general, Data Scientists gratefully accept the provision of mathematical routines through statistics software.

This easy access, however, comes with the risk of those routines getting used thoughtlessly. Preconditions and adequacy often remain unvalidated.

Let’s have a look at a typical example: A one-way analysis of variance. The three critical assumptions for the reliability of such an analysis of variance are

- Response variable residuals are (approximately) normally distributed.
- Variances of populations are equal (“homogeneity of variances”).
- Responses for a given group are independent and identically distributed normal random variables.

Validating those assumptions is not straightforward in real life. As an example, you can determine the residuals based on your sample, but you cannot verify whether they follow a normal distribution. Furthermore, it is difficult to tell whether the variances of different samples are sufficiently equal.
Can the investigator influence the outcome?

The last point leads to a second challenge: Despite the availability of a correctly determined sample and the preciseness of the mathematical formulas, many decisions are left to the investigator. This finding should surprise everybody who considers data science a mathematical, fact-based, undisputable discipline.

Bias comes in two phases: First, in the data preparation phase, then in the data analysis phase.

The data preparation phase, also known as Exploratory Data Analysis (EDA), is an indispensable step to detect data issues early, to exclude first errors, and so on. If done well, it will lead to more meaningful results of the analysis. But it is not a mathematically precisely described phase. Most decisions are left to the investigator. In other words, they are based on experience and gut feeling.

The subsequent data analysis phase, despite the availability of mathematical algorithms, leaves a lot of room to human bias as well.

I will use the ANOVA again, as an example of a scientific approach that leaves a considerable amount of decisions and estimates to the human being.

Let’s have a look at a training text of Zurich University, after describing the execution of an ANOVA:

“(…) The Levene test tests the hypotheses that different groups have the same variances. In case the Levene test is not significant, homogeneous variances can be assumed. If, however, the Levene test was significant, one of the preconditions of the ANOVA would be violated. The ANOVA is assumed to be fault tolerant in case of light violations. Violations are considered unproblematic notably in case of sufficiently big groups of more or less the same size.

If sizes of samples differ significantly, a strong violation of the homogeneity of the variances leads to a distortion of the F test. Alternatively, you could use the Brown Forsythe test or the Welch test which are adjusted F tests. (…)”

¹Website of Universität Zürich, www.methodenberatung.uzh.ch/de/datenanalyse_spss/unterschiede/zentral/evarianz.html. Own translation. Original text: Der Levene-Test prüft die Nullhypothese, dass die Varianzen der Gruppen sich nicht unterscheiden. Ist der Levene-Test nicht signifikant, so kann von homogenen Varianzen ausgegangen werden. Wäre der Levene-Test jedoch signifikant, so wäre eine der Grundvoraussetzungen der Varianzanalyse verletzt. Gegen leichte Verletzungen gilt die Varianzanalyse als robust; vor allem bei genügend grossen und etwa gleich grossen Gruppen sind Verletzungen nicht problematisch. Bei ungleich grossen Gruppen führt eine starke Verletzung der Varianzhomogenität zu einer Verzerrung des F-Tests. Alternativ können dann auf den Brown-Forsythe-Test oder den Welch-Test zurückgegriffen werden. Dabei handelt es sich um adjustierte F-Tests. (…) (called July 18, 2019)
I have underlined all words that leave room for interpretation and bias – an astonishing number!

As an example, take the term “significant” from the preceding text. You may argue: “In the case of statistics, significant and not significant are very precisely determined, aren’t they? Therefore, there is little to no room for misinterpretation.”

Yes, indeed, the border between “significant” and “not significant” is usually precisely defined. This definition, however, is not a result of an objective determination of the best possible (let alone “correct”) threshold.

No matter how you measure significance, it will usually be a continuous function. As a consequence, the degree of significance on both sides of the selected threshold is almost identical, if you stay close to it.²

In other words, “objective significance” does not exist. While there are always good reasons for the chosen separator between “significant” and “not significant,” its final determination remains somewhat arbitrary: You have to draw a line somewhere.

In reality, imagine two random samples taken from the basic population. Not surprisingly, they lead to almost the same analytical result. Now, be conscious of the fact that “almost” may mean that one result is slightly below the threshold, while the other one is above.

How do you proceed in such a case? Will you resist the temptation to select the one sample that supports your original assumption (or your boss’s expectation)? What if somebody else uses the other sample? You come to fundamentally different conclusions, based on very similar statistical results.

This topic is a typical case of a mathematical statement that usually remains unchallenged. I invite you to challenge such statements even if they come from statistical textbooks. At least, you should understand the underlying logic and decisions to limit the human bias.

How about combining both challenges

Sometimes you find both problems combined: The preconditions are listed without explanations, and in a way that allows for biased approaches. A beautiful example can be found on BA-Support.com, again about the ANOVA:

²That means you can get the difference as small as you want by moving both values sufficiently closely toward the threshold. Mathematically, you’d say that no matter how small you set the difference Δ in significance s(x) (x being a real number), there will be a value ε, so that the difference between s(threshold – ε) and s(threshold + ε) is smaller than Δ.
“This test can only use when a number of preconditions are in place. These are that all groups must either contain more than 30 observations or be normally distributed also that there must be comparable variances across groups.

Furthermore, the variable describing the groups should be nominal-scaled while the dependent variable (variable we compare with, or just ‘the other!’) should be interval-scaled. In some textbooks, you will also be able to read that the groups should be similar in size, do not have to worry about in SPSS, as it has a procedure to correct for unequal sample sizes.” (BA-Support, 2019)

Again, the underlined words form an impressive number of cases open to human bias. I invite you to ponder on the following questions:

- Where does the threshold of “30 observations” come from?
- Is there really a good reason to draw a line exactly between 30 and 31 observations – everything below is “bad,” and everything above is “good”?
- What does “comparable” mean? Is this decision left to the investigator, his/her experience and gut feeling?
- “Should” implies that something is recommended but not mandatory. So, what happens if you do not follow this advice?
- Is it wise to trust a piece of software without knowing which logic that software uses to make a decision?

**General limits of AI**

Artificial Intelligence comes with an incredible number of opportunities. But we need to be aware of the limits and how to deal with them.

**Data sources**

The best algorithms lead to wrong results if fed with bad quality data. This should be intuitively clear. However, it is still broadly ignored, as the main focus often lies in obtaining results. And information derived from bad data doesn’t look worse than if it is based on high-quality data.

As a consequence, in many cases, it is only the following activities that are expected from data scientists when it comes to finding data:

- Searching the Web using keywords
- Doing screen scraping where data cannot be downloaded in file format
• Downloading files (if necessary, at second hand)
• Reusing files somebody else had downloaded earlier
• Deriving the meaning of columns, for example, from header names or content

This approach has led to typical patterns of problems that reduce the value of data:

• The age of the data is often unclear. It may have been accurate a while ago but has meanwhile become outdated. (And “a while ago” stands for an ever-decreasing time span!)
• The level of completeness is often unclear: Data may have been filtered, or collection may have stopped prematurely.
• Definitions of columns are often guessed: As most flat files or spreadsheets come without a proper explanation, data scientists are forced to derive the meaning, for example, from header names.
• Bias! It is usually unknown whether a table was created with a specific manipulative purpose in mind. Even if an organization that provides data has a good reputation, human beings are involved in assembling that data. And you cannot tell from the content itself: An honest “5” and a manipulated “5” look exactly the same in a data file.
• Another data scientist may have removed or merged columns that correlate strongly with others, possibly for good reasons in the context of that data scientist’s specific question. But maybe the subtle differences would have been the interesting parts in the subsequent use of the same data to answer different questions.

In response, you may wish to add specific responsibilities to any data scientist’s data acquisition job:

(i) **Know the source**

It is important to obtain knowledge about creators and providers of publicly available data. Without this knowledge, data is as useless as a positive product rating on the Internet without knowing whether the author is the seller or got paid for the rating.
Data search portals such as https://datasetsearch.research.google.com (finally released in 2020) are great. But they are not “the source,” and they don’t give any guarantee of Data Quality.

(ii) **Contact the provider**

In some cases, getting in contact with the original creators or providers of data found on the Internet helps you understand their motives and methodologies.

In the long run, a good relationship may come with access to additional data or even a regular, fruitful exchange between two parties.

(iii) **Understand the history of the data**

Data Quality is not a constant attribute. It changes even if the data itself stays the same. Finding out about a data source’s history of Data Quality helps understand the status quo: Has the data been maintained? Is it under permanent review? Or was it the result of a one-off “fire-and-forget” effort, no matter how diligently it was executed?

(iv) **Clarify the data model**

The data model and the terminology used by a data source will probably differ from how your own organization calls data elements and relationships, and how it defines data structures.

A bit of research at this point helps avoid severe trouble later or even undetected misunderstandings that lead to wrong conclusions.

(v) **Check multiple sources**

Hardly any information is only available through one single source. Searching for multiple sources of the same data is, therefore, a good principle, as it allows you to validate through comparison.

Please consider that different parties often copy from each other. Just as with rumors in your daily life (which appear more plausible if you hear them from different sources), this may result in a false sense of confirmed quality.
(vi) **Objectively classify the data sources**

In order to be able to reproducibly qualify the results of your data science initiatives, you need to classify the quality of each data source. This is also important for future use of the same data in a different context (and maybe by a different data scientist).

Typical attributes are reliability of content, the correctness of the description, and the frequency of updates.

When you determine the level of bias and trustworthiness of a data source, remember that it is not only about the collector of the data. The resulting data may not be based on facts but on a “representative” survey or on estimates.

A survey is a valid means of getting a good understanding of a situation in cases where a full assessment comes with unacceptable effort. But the interviewees might be biased.

If, for instance, you ask 1000 CEOs about their organizations’ maturity level of digitalization, or if you ask 1,000,000 customers about their level of satisfaction, you have a comprehensive sample – but each answer will be biased, and the interviewees’ criteria may differ.

(vii) **Document your data sources**

Information about data sources (both about the data files themselves and where they come from) should be documented (and maintained) together with all other relevant metadata so that the knowledge is systematically made available to all data scientists.

The documentation should contain all known attributes such as the (confirmed!) meaning of columns, the covered moment in time or time period, the year of creation, any applied filter, and the original purpose of the creation of the file.

Furthermore, you may wish to standardize your criteria and attributes so that any two sources can be compared.
Collaboration

If your organization’s policies allow, I encourage all data scientists to work with a broader data community, beyond your own organization. This is not about disclosing internal information but about sharing information about external data sources.

AI algorithms

Let me illustrate some limitations of AI, using a concrete example.

This example is about face recognition, one of the most prominent use cases of neural networks.

I personally am impressed by what face recognition can already achieve. However, using such algorithms to pick suited candidates from a list of job applicants comes with a few challenges that may make you reconsider their suitability.

And I am not even talking about the ethical part – even the service itself cannot be worth the money spent!

In Autumn 2019, Bobby Hellard published an article called “AI and facial analysis used in job interviews for the ‘first time’.” It describes a case of a US organization that developed a software “which analyses the tone of voice, vocabulary and facial expressions in video interviews to determine a candidate’s suitability” (Hellard, 2019).

The algorithm may accidentally be biased. Even worse, it may be biased on purpose. The creators could shape it in a way that their own CVs are rated highest, making themselves look like the most auspicious candidates. In essence, AI allows people to create perfect job opportunities for their own career progression in a very unethical way. After all, it is a black box – nobody else will ever find out…

The worst aspect, as far as I am concerned: Such an approach prevents HR from getting better over time!

Here are a few shortcomings of the underlying AI algorithm. You will find them in other cases as well:

(i) Manifesting errors of the past

People with combinations of attributes that prevented their holders from being selected in the past will not be selected in future. If the selection criteria of the past were suboptimal, so will be the criteria of the future.
(ii) **Inability to consider changes in demand patterns**

The world changes and different profiles are required. How would such a backward-looking algorithm be able to understand changes in demand?

(iii) **Inability to judge the learning level**

Unsupervised learning is key to improvement. In this case, however, the success rate can hardly be validated, particularly not by the algorithm itself. This is not a problem of the algorithm but of the available training data. After all, it takes years until you can tell whether a new hire can be considered successful.

(iv) **Not knowing WHY**

As with many AI algorithms, the algorithm at hand does not tell us why specific profiles have been more successful in the past. To determine the best candidates, you’d need to base your hiring logic on meaningful parameters. An algorithm that cannot distinguish causality from correlation is useless and dangerous.

Don’t get me wrong: We should not assume that the determination of correlation is always insufficient! But we need to be clear about when we need causality and in which cases we can work with correlation.

When it is about finding out whether we should “Do more of A to achieve more of B,” causality is crucial. If, however, we want to validate a statement like “Look out for C to find more of D,” working with correlations may be perfectly okay.

(v) **No insight beyond the training data**

A big challenge of AI is the applicability of results. No AI algorithm can provide reliable insight beyond its training data. If an element is “evaluated” by an algorithm without being covered by the algorithm’s previous training or testing data, the result is useless at best. Such a situation may be acceptable where an algorithm tried to improve the average selection, for example, the number of holes you need to drill in the ground on average until you find particular raw material. You could consider it a success if 50 percent of all cases saw an improvement, while the other 50 percent
were just random – the overall effort would go down. But whenever you are dealing with individuals, improving on average is considered ethically unacceptable in most modern societies.

Why do we need a data-literate HR world? To prevent approaches like this one from becoming part of an HR department’s toolbox. No data manger or privacy officer should have to tell the HR people!

(vi) Legal issues

Organizations that use this kind of algorithms usually stress that it is one evaluation component only and that human beings still do most of the assessment before a hiring decision.

This statement, however, does not apply to those CVs that were sorted out by the algorithm before the first human being had a look. They remain 100 percent processed by software, no matter how much human evaluation is subsequently applied to the other CVs.

That is why this approach is not only ethically questionable, but it is also violating existing laws. Article 22 of GDPR states “The data subject shall have the right not to be subject to a decision based solely on automated processing.”

All in all, you don’t want to lose control over the process (as illustrated in Figure 20-2). The best advice I have read in this context comes from Hanover Recruitment Limited, a British recruitment agency. They state on their website: “Ultimately, you should treat candidate and client data the way you would want your own data to be treated!” (Beatie, 2018).
“Actually, yes, we did let AI choose the shortlist of candidates!...”

Figure 20-2. HR = “Humanoid Resources”?

Human behavior

“I know what I am doing” is what I often hear from Data Scientists. But it is not sufficient for the scientist to know.

Data Science is about credibility. It is a matter of time until two colleagues independently work on the same problem. They will probably use different (variations of) methods. Please expect them to interpret the same situation differently and to come to different results.

These results may not be fundamentally different. But they will at least differ sufficiently even for a layperson from Marketing to question the preciseness suggested by the number of digits after the decimal point.

What helps? Transparency! Share what you are doing, in layman’s terms. Don’t overpromise – you don’t need to! Explain limitations. It doesn’t hurt to say, “This result is only valid if A and B are independent, but we don’t know to which extent they are.”
AI – Quo Vadis?

The game Go is reported to be one of the most complex games on earth. You will have read that, in 2016, DeepMind’s AlphaGo computer defeated South Korea’s Lee Sedol, the reigning world champion. This event was broadly perceived as a massive step in the development of Artificial Intelligence (AI).

I recently read an article from IBM that claimed: “The important outcome from Sedol’s defeat is not that DeepMind’s AI can learn to conquer Go, but that by extension it can learn to conquer anything easier than Go – which amounts to a vast number of things.”

Sounds reasonable? Well, it may not be.

The main problem I have with this statement is the word “easier.” Its comparative degree suggests that we are talking about a one-dimensional scale. In other words, you could conclude that any two intelligence-based performances can be sorted by “ease” (at least if one of the two is “playing Go”).

But how do you define “easy”? Is it easier to defeat the reigning world champion in Go, or is it easier to convince a kidnapper to release all captives and to give up? Is it easier to win Jeopardy! or to convince someone of a different view during a discussion? We obviously face multiple dimensions of human brain performance here.

Ask a computer like AlphaGo to calculate the best next movements of a football player in real time, and (net of hardware constraints) even the author of these lines would turn out to be superior (which says a lot). Whoever has ever watched RoboCup – the world championship in robotics football – knows what I mean…

I don’t know how quickly AI is going to progress. However, one indicator is the investment of money in research. My recent discussions with a lot of organizations suggest that they are moving from the phase of “We need to be part of it by all means” to that of “Show me the value.” The real money will go to AI research that comes with a business case.3

So, if you put yourself in the shoes of an organization, how would you think and act?

A well-managed organization will inevitably ask “How can AI make us more successful?”, considering the usual stakeholder triarchy of customers, shareholders, and employees. Even universities and nonprofit organizations in general are more and more driven by the need to work toward creating tangible value.

3Please don’t focus too much on those public relations–driven “Innovation Hubs” of big corporations where researchers can play around without commercial pressure as long as they publish interesting stories from time to time.
Let us be honest: It will take a huge, long-time investment to have true Artificial Intelligence (in the sense of copying the human brain) add value to our society (or even to single organizations). So, what drives the investments of organizations?

Firstly, organizations generally prefer dedicated specialist solutions for each task. This is not specific to AI. No car manufacturer would want to invest in a vehicle that can transport bulky goods, win a Formula 1 race, and deliver a luxurious convertible feeling at the same time. Three specialized solutions will usually do a better job in their specific areas, at lower overall costs. The fact that you need three solutions for three different tasks is not perceived as something embarrassing, for good reasons.

Look at robotics: While the media is full of humanoid robots that can smile and shake hands, most commercially driven research is still being done on specialized robots which can do precisely one thing in a near-perfect way (and which requires at least substantial reconfiguration before being able to perform another task). The industry has been working on this for decades, with impressive success.

Secondly, developing something further that is extremely far behind existing capabilities (in this case, current human abilities) has a bad business case. Return on investment lies in the distant future, and it is uncertain, after all.

That is why I am reasonably confident that development in AI will continue to focus on areas where AI already delivers a performance superior (or comparable) to that of humans today, through specialized solutions. Mimicking ingenious mechanisms of the human brain, such as a neuron’s way of working, will definitely contribute to the success of this approach.

At the same time, at least in the foreseeable future, I don’t expect any substantial progress in developing clones of human brains that don’t require reprogramming for new challenges.

Does this sound pessimistic? It is not meant to!

I am a great believer in AI’s growing capabilities in complementing the abilities of human beings. God did not design us godlike (which is easy to believe after watching the evening news), so why should we, in turn, invest our precious brainpower in developing humanlike devices?

Recommendations around Analytics

Based on what I have said so far, I would like to propose a set of guidelines that you may ask your entire Analytics team to follow: My “12 Guidelines for Analytics.”
You should feel free to adjust them to your individual findings and priorities—but I strongly recommend having such guidelines as they help achieve a consistent approach across the entire team.

I. Determine the necessary degree of preciseness

Sometimes it is not required to consider all mathematical preconditions.

This is particularly the case if it is only about finding a preoptimized starting point for a second algorithm (through heuristics). In this case, the quality of the starting point has an influence on the amount of the remaining calculation work but not on the final result.

Sometimes, however, convergence is not guaranteed, and often you cannot validate later whether your algorithm plus data worked well. Even if you split your data into training and testing data, the latter may not correctly reflect reality.

But wherever your model is “only” expected to make a process more efficient, where you have a short validation cycle, and where you can measure success easily, you should not insist on a minimum level of accuracy up front.

It doesn’t matter whether you have found the best possible algorithm—it makes things better, its existence is justified. If a process performance was at 70 percent, and you get it to 80 percent, it adds value, despite the fact that 80 percent is far less than 100 percent.

In such a case, you can operationalize a model within days—after you have run it in parallel, and the resulting algorithm delivers improved results.

And once your solution is live, you are in an even better position to further fine-tune it, particularly if you can measure the delta between its current performance and 100 percent.

Putting such progress into charts helps impress your business customers: You can link something your team has done immediately to an improvement of a metric that is relevant to a business stakeholder.

II. Don’t use a formula just because “it works”

Let’s assume you have an interesting case. The question is clear; the base population is already available in your database. You can call a function of your statistics software, and it will return a syntactically (and mathematically) correct result. All input parameters are available. The temptation is high.

But does the formula fit the purpose? Maybe there is a better approach—perhaps even using available business information?
Example? You might have used an Unsupervised Learning routine to cluster your base population. The number of clusters follows scientifically accepted criteria, for example, using the sum of squared distances within each group. Your gut feeling tells you that you have found a good compromise – increasing the number of clusters further would not add significantly to the score you have defined, also taking into consideration the reduction in cluster size that comes with any further clustering. Let me say it in economics language: Your marginal benefit in increasing the number of clusters approximates zero.

So, whatever number of clusters you have finally decided to go for, your decision is based on a blend of data, formulas, gut feeling, and experience. However, it may not have been based on the business background of the underlying population.

This is why data scientists need to be ready to get out of their comfort zone. There is a lot of complementing wisdom on the business side. Discussions with the right business folks may reveal, for instance, that a natural clustering of your base population is already evident from a business perspective, using combinations of well-known attributes.

Using business knowledge may not only be more adequate than applying a broadly accepted yet content-independent method such as an elbow criterion (where you look out for a strong bend along the benefit curve, indicating a significant drop of the marginal gain). It will also make it easier for business folks to understand what you are doing, and your credibility as a practical solver of problems will increase.

### III. Check all preconditions

Is your basic population normal-distributed? Really? Can you guarantee full statistical independence?

You cannot tell from the result whether all preconditions of applying a method were met. And where it is almost impossible to fulfil a precondition, you should at least know to what extent it is fulfilled.

Wherever possible, you should schedule regular validation of whether all preconditions were given. In some cases, you might find that your results were too bad for the preconditions to be true – and sometimes your results corroborate your assumption that all preconditions were fulfilled.

Think of election polls where people coming out of their election cabins are asked which party they have just voted for. The result is then displayed as the first prediction after the polls close. The institute usually selects the polling stations in a way that their joint result of the last election was very close to the outcome. Furthermore, those people need to be randomly selected, and all of them need to voluntarily tell the truth.
You will remember that the results of such polls are usually relatively close to the final outcome of the election. But are they as good as they should be if all the statistical preconditions were fulfilled? The answer is NO. If you do a bit of maths, you will find that, statistically, 98 percent of all polls with fulfilled preconditions would be better than the observed result!

Of course, not all your calculations in a business context can be validated as quickly as this kind of election polls. But if you have the chance to, say, check your forecast of the sales distribution across the product portfolio against the real sales figures later, you should do so, to get a better feeling of the validity of the preconditions.\(^4\)

IV. Be open about the limitations

Both data and algorithms have natural limitations. You might have reasons to take the risk of using them beyond those limitations. But be clear about your choice and motives.

Sometimes you may be forced to perform year-on-year comparisons where the annual data comes from two different sources (e.g., because new software was introduced in between or because the data provider has changed). Please add this information to the presentation of the results, as a warning.

Another typical limitation is the origin of the training data for your AI models. You usually cannot assemble it yourself, particularly if you want to base your analyses on millions of records. As a consequence, you usually end up using the same limited number of publicly available data repositories. Here’s the issue: The fact that your results are in line with those of others does not prove their correctness. It may rather indicate that all of you have been using the same source of data.

Look at the repositories of images for OCR and image recognition: They have usually been assembled in a specific context, and they will probably not work beyond that very context.

There is another challenge that comes with character recognition: No OCR algorithm can safely tell the difference between a 0 (zero) and the capital letter “O” or the difference between the “I” (i.e., the numeral “1” written in Anglo-Saxon style) and the capital letter “I” (if written without serifs). The same image may mean two different things. Even within the family of digits, a European “1” and an Anglo-Saxon “7” may look identical. For a safe distinction, you will also need to consider the context. If you cannot do this, be open about it!

\(^4\)This doesn’t mean that you have to tell the Board that you worked with unfulfilled preconditions.
V. Explain your assumptions

Please have another look at the clustering example earlier and remember the analysis of variance at the beginning of this chapter. How often did you make a human choice?

- Defining a score
- Choosing the metrics of how similarly the elements of a cluster behave
- Balancing the variance within each cluster against the effort that goes up with the number of clusters and so on.

We are far away from the necessary level of unsupervised learning that would allow us to leave all of this to an algorithm that knows autonomously which information to take into consideration. That means it is okay for humans to make decisions.

But you need to explain those decisions. Which assumptions have made you decide the way you decided?

VI. Don’t convey a false impression of preciseness

Imagine someone collecting the passports of a team of 40, recording their ages, and calculating the team’s average age. Let’s assume the resulting average age to be 34 years, rounded.

Now imagine another person ignoring the passports and estimating the age of each team member instead. That person calculates the average age of the team to be 37.575 years.

Which of the two persons provides a more accurate result?

I guess you see what I am aiming at.

Being open about the preciseness of your results is part of building your credibility. Instead of “the likelihood of a customer in this group to repay a loan is 68 percent,” you may wish to state:

“With a probability of 95 percent, the likelihood lies between 65 and 71 percent, assuming that all preconditions of the algorithm are met. In other words, if the true likelihood of that group were 68 percent, on average 19 out of 20 simulations would forecast a likelihood of between 65 and 71 percent. For each precondition that is not fulfilled, the uncertainty increases.”
This statement is harder to digest, but once people get it, you won’t need to repeat it over and over again. You’d achieve the following targets:

- You support your organization on its journey toward a data-literate organization. They don’t need to be able to prove the Bayes theorem, but they should gain a basic understanding of what is possible.
- You manage expectations and protect yourself: People should have a realistic idea of how precise your figures are. And you don’t want to be “proven wrong.”

You always leave room for improvement. More data and more performant computers will generally increase accuracy. This allows you to create a price tag for increased accuracy. If an executive is not happy with a range of six percentage points as in the preceding example, you know what your proposal will have to look like.

VII. Automate data preparation carefully

Automation of data preparation can save a lot of time, and it leads to reproducibility as it reduces the human factor, including personal bias.

Note that this does not guarantee (or even improve) the correctness of the data. Beyond reducing human error and speeding up the process, it just makes the same data look better, and it eases subsequent processing. This, however, may foster a false impression of Data Quality.

VIII. Use DataOps

With the rise of DevOps, that is, having the same teams develop and run a piece of software, people started to think of extending this idea to data topics. Unsurprisingly, this concept got coined DataOps.

DataOps helps set the incentivization properly by avoiding a situation where different parties work toward their own targets only.

Furthermore, the review cycles and the constant monitoring of Data Quality help discover implausible data, for example, through heuristics.

But please resist the temptation to limit this approach to Analytics. Starting DataOps with the preparation of previously unmonitored data for Analytics purposes is too late.
IX. Balance diligently

You will hardly ever find the overall optimum for a problem. The world is simply too complex (pun intended). There are too many influencing parameters, and the different optimization criteria can usually not be weighed objectively.

This situation forces you to compromise – either by further simplifying your models (every model simplifies by definition) or by reducing the list of options to assess (as you easily have more options than there are atoms in the known universe).

You can apply both strategies at multiple levels.

- You can deal with simplification by building a model as complex as technically possible, or you can simplify as much as possible while still obtaining a syntactically valid result.
- Alternatively, you can reduce the number of options through clever usage of algorithms (e.g., I don’t need to assess an entire group of options as soon as I have found one option that is proven to be better than each of the options in that group).
- Furthermore, you can also reduce your list of options before you apply an algorithm, based on nonscientific factors such as your business strategy. For example, you have determined a group of options, and you can prove that the overall optimum is part of that group. But none of these options is in line with what your organization has decided to focus on as part of its product strategy.

But how do you know to which extent you should compromise? Again, you would have to calculate. And, again, you would quickly find out that the determination of the “best” level of compromise is a tremendously complex calculation, far too complex to be executable.

The first challenge would be the measurement of how far you compromise. Insisting on fully considering all parameters would obviously mean zero percent compromise, while a random choice would be 100 percent compromise. But how do you find out where in between any approach would be?

Complexity cannot be overcome by developing more and more sophisticated algorithms. The more sophisticated an algorithm is, the more complex it gets. And the more complex an algorithm is, the less you know how accurate the result is. In other words, you don’t win anything compared to guessing the right balance.
We often think we are forced to go straight from old-fashioned gut feeling to modern data science, and that data science is going to replace gut feeling. This way of thinking suggests that data science is the solution and that it will provide the ultimate answers.

Instead, the best choice may remain a well-balanced approach somewhere in between the two extremes. This, however, might require a change of ambition: From aiming at finding out what “the right business decisions” are toward determining how to apply data-driven approaches in the most adequate (not perfect) way.

And, yes, experience and gut feeling remain part of this exercise – which stresses the importance of being open about the limitations (see Guideline IV “Be open about the limitations”)

X. Exclude emotional factors

Another helpful skill is your ability to distinguish between “rational” and “emotional” decision factors. As we have seen, even the rational factors are deficient. That is why you should at least eliminate emotional factors such as pride or anger.

How do you develop such skills? You can, of course, read books about Emotional Intelligence (and I encourage you to do so), but unlike studying science, you cannot just “read and remember.”

That is why I recommend that you reflect together with others. Be open-minded and expect others to surprise you. Again, unlike science, you will find that two different views may both be acceptable.

Therefore, this approach does not mean adding another piece of knowledge or a rules engine to some knowledge store in your brain – you might rather wish to broaden your basis for future decision-making.

XI. Consider changes outside the model

In traditional modeling, you consciously decide which parameters you want to become part of your model. In Macroeconomics, this is often described as Ceteris Paribus (Latin for “everything else equal”): Whatever you do not want to (or cannot) build into your model is assumed to be constant. Consequently, you are forced to think up front about all aspects that have a potential impact so that you can consciously decide whether to make them part of your model.

In neural networks, you often rely on the model to do this work for you. It is expected to implicitly discover all relevant aspects: If these are part of the training data, they will automatically influence the training process positively.
The missing need to consider all aspects up front comes with a risk, though. It is about the representativeness of the sample data you use to train your model and to validate it afterward.

In fact, you may miss certain aspects when collecting that data. You think your data is representative, but you may have accidentally left out relevant subsets of your overall population.

The most frequent omission is not to consider the time aspect. As a matter of course, any sample was gathered in the past. Well, things may have changed since! The COVID-19 pandemic with its dramatic change in consumer preferences is a good example.

There may be changes to aspects you would not even have thought of, but which may have an impact on what you are assessing. As another example, changes to legislation or scandals made public may have a significant impact on people’s buying preferences as well, so that a sample from before such an incident would not reflect today’s reality.

The issue is that this kind of failure is difficult to spot. You would, as usual, use 70 percent of the sample to train your model and the remaining 30 percent to test your algorithm. As a result, your test data would confirm your model as it reflects the same time in the past as your training data.

The good thing is that you can do something about it. The downside is that this comes with additional effort: You’d have to think about all potentially impacting aspects, even those far away from your core considerations.

It is, in particular, not sufficient to consider all attributes you find in your sample, no matter how reliable that sample may be. You are back to the hard work of deterministic models, to a certain extent: Diligently contemplating about all potential influence factors.

Of course, you can take some mitigative action as well: Use the longest possible time range for your training data and separate it by time period. If you model works similarly with data from different time periods in the past, it seems at least invariant to regular changes (such as fashion trends or the change to a different government) or to the presence of singular events (such as a pandemic, a natural disaster or the Olympic Games). It won’t help in cases, though, where all time periods covered by your training data are before an impactful day X while you try to make forecasts about the time after day X.

Here is my key learning: A good model requires more than a perfectly trained Data Scientist – a broad educational background is almost equally important. This is a strong reason for mixed teams where different skill sets complement each other.
XII. Define success comprehensively

How do you validate whether an AI initiative was successful? Or, in case you are improving a model iteratively, how do you know you are heading in the right direction?

At first glance, it is easy if a measurable business criterion was defined and measured up front. You see improvement, and you can prove that your algorithm was successful, even if you don’t know whether you could have done better.

This judgement may be wrong, and the reason could be an incomplete measurement of the impact beyond the single target criterion you had concentrated on.

Imagine a police station that wants to become proactive in crime avoidance. We can assume that there is enough historical data for them to develop and train an algorithm to determine any (known) person’s probability of committing a crime soon. As a consequence, the police station could implement a policy to put people to jail whose calculated risk of committing a crime exceeds a certain threshold.

Of course, even the most autocratic countries would probably not implement such a concept. But the underlying principle may sound familiar to you as it keeps being applied to other use cases.

Here’s the interesting aspect: In the case described earlier, you would most definitely find that the crime rate will indeed go down! The AI algorithm will identify a lot of truly dangerous persons who would not be able to commit a crime while in jail. The initiative is going to be considered a success, in the name of safety.

But aren’t there a few undesired side effects?

If you don’t consider the entire impact of an initiative in your evaluation, you may miss the consideration of key factors (basic human rights in this case) that can render the entire result negative.

You may not find equally drastic examples within your organization. But you will often discover a similar pattern.

A very simple example is the assessment of different measures to increase revenue. If one of the measures is to apply discounts, you would not just validate the increase in revenue (which is almost certain in this case). Instead, you would also consider the negative impact of the discounts on the margin (and eventually on your organization’s bottom line).

Here’s my recommendation: Determine all areas that are possibly impacted by a data-driven change and consider them in an overall target function up front.
Such a weighted target function also forces the business stakeholders to make up their mind and to set priorities.

Admittedly, translating every parameter to one denominator (most frequently the financial impact on an organization’s bottom line) requires a lot of work. Alas, data can help you do this job, and you are in Data Management!

**Explainable AI (XAI)**

Unlike in the past, where algorithms were built from rules, many modern AI methods are based on learning (thus the expression *Machine Learning*), either from examples or through rewards.

The traditional, **confirmative** approach, as known from Operations Research and matured during the second half of the twentieth century, always starts with the determination of all rules and constraints. These are then translated into a model, to which a suited (mathematically proven) algorithm is applied. This approach comes with two main risks:

1. **Openness to bias and fraud**

   The confirmative approach starts with a set of rules that are sitting outside the precise algorithm. They are usually assembled by people with a vested interest in the underlying issue who may therefore be preoccupied or even working toward the desired outcome.

   As a result, boundaries are omitted, constraints underestimated, or inequations “tuned.” Another significant source of bias is the creation of the target function, that is, the weighting of the different target parameters for the overall normalized target.

   The resulting “problem,” usually an inequation system,\(^5\) can then be calculated with mathematical accuracy – which suggests that the result is indisputably correct. However, if the underlying model has been subject to bias, the result will rather reflect the ambitions of the investigators than an objectively right optimum.

---

\(^5\)Such an inequation system is usually represented by a matrix or by more efficient notations in the case of network optimization tasks where the matrix would be extremely sparse.
(ii) The incompleteness of the model

Each optimization problem has an infinite number of parameters. Any list of nodes, boundaries, and constraints represents a simplification of the true problem.

While in the early days computing power was the main limiting factor, the primary constraint today is the completeness of the underlying business knowledge. If people don’t know all constraints, they cannot add them to the model.

Of course, models are simplifications of reality. However, in Operations Research, you cannot tell up front whether the model is sufficiently representative of the real situation, and the result will not tell you either.

Only if you use the outcome in the real world, you may find out how good the model was. In most cases, however, you will end up with an invalid situation as certain constraints turn out to have been missing in the model. This situation requires a long iterative process that cannot be left entirely to computers.

This situation suggested that the approach be turned around, and to give the parallel development of Artificial Intelligence a second chance: You start with data, followed by the derivation of correlations and causalities.

This transition from rules-based approaches to neural networks means a change in direction:

From

\textbf{wisdom-to-data} (confirmative)

to

\textbf{data-to-wisdom} (exploratory/Bayesian)

You may have guessed that this move has not solved all the problems of mankind. In fact, it has even brought new issues.

Let’s have a look at some striking challenges.

Unknown cause and effect

As we have seen, today’s algorithms work as a black box: You don’t know anymore why a neural network comes to a result. \textit{Causality is unclear}. Neural Networks are all about nodes activated and firing, based on thresholds…
How do you improve an algorithm of which you don’t know how it takes decisions? It is not possible to fine-tune the rules – there are no rules, after all, at least not in a deterministic sense. The only way to improve a neuron network’s quality of decisions is through more and more training. But what if you run out of training data?

And even when the known cases you have kept for validation purposes confirm a sufficiently high success rate, how about false positives?

Imagine an algorithm that successfully recognizes 99 percent of all images with a dog. It may even identify the type of animal on the other images.

But how do you know that, in real life, it will not recognize an (untested) arbitrary color pattern as a dog? The threshold between “It probably is a dog” and “The probability of it being a dog is higher than the probability of it being any other animal” is usually determined through manually set thresholds.

I am almost certain that science will understand these black boxes better over time. But it will be a long and tedious discovery journey, comparable with the decades-long (and still ongoing) process of understanding the human brain itself.

Trust issues

It is usually okay for a patient to know that a pattern recognition algorithm supports the doctor in detecting skin cancer. After all, the results cannot be worse than human judgment alone.

But how about algorithms that may take disadvantageous decisions for single humans?

An AI algorithm doesn’t know anything about the subjects of its calculations. It usually cannot say “Wait, this person filmed by a surveillance camera cannot be Ronald Reagan – it must be a carnival mask!” Instead, we’d need to enrich or modify the data up front, as the algorithm itself is not rule based.

You will easily be able to think of more practical examples with relevance to you and me, such as AI algorithms that calculate credit scores.

Another example of such a dilemma is an algorithm in autonomous driving that may have to decide between two options, both expected to cause casualties. Would you let an algorithm steer your car that may come to the conclusion that your life is of less value than that of a group of people on the road?
Ethical issues

Evidence-based algorithms don’t have a conscience – whereas rules-based algorithms can avoid unethical results by making such outcomes unattractive through the intelligent setting of boundaries or by selecting different weights in the target function.

But how do you find out that an outcome that an AI algorithm considers optimal is, in fact, an unethical outcome?

Remember the HR example under “General Limits of AI” earlier in this chapter: Here we would systematically sort out candidates based on an algorithm of which even the Data Scientists don’t understand why it decides the way it does.

This is not only a violation of GDPR and other data privacy laws, it also discriminates against certain groups of people, without the developers’ intent to do so! In other words, it is not sufficient to feel ethical. A Data Scientist also needs to act ethically, by consciously searching for factors that may have an ethically undesired impact.

Is there a way out?

This question is being asked increasingly frequently, and it has led to the phrase “Explainable AI” (XAI): Understanding what is going on is considered critical to addressing the issues of black-box algorithms, of ethics, and of trust.

This discussion has only started, and there is no satisfying answer yet.

However, the following questions may guide us through the upcoming dialogue:

- How are Data Scientists going to gain the trust of users? Nobody trusts in an algorithm anymore that selects candidates based on experience which may be suboptimal.
- In which cases will users accept the “black box,” and when will they refuse it?
- And what to do about it? Will there be new algorithms that allow for decisions to be tracked back? To which extent will it be possible to derive “rules of life” from the gradient of a learning AI algorithm? Will enhancements to known algorithms such as the recently promoted “Layer-wise Relevance Propagation” (LRP) mature sufficiently to add transparency to existing neural networks? Or will new approaches dominate where transparency comes by design? You might, for instance,
start with subproblems with a lower number of dimensions, where a lower number of aspects can be understood more quickly, and where humans might still be able to understand the smaller number of interdependencies?

- Will it be possible to apply another algorithm to the output of such a “black box” that shows which parameters (or combination of parameters) have led to a yes/no decision?

- If transparency approaches develop, how will it be possible to classify them, so that data scientists will know which algorithm they are allowed to apply in which case?

- Do you think recent initiatives to shape Explainable AI will be successful any time soon? Or will people get used to trusting AI, as they “learned” to trust their IP television device or their Alexa device?

- Where else in our business lives do we face black boxes? Where have we always accepted them as a matter of course, for example, because organizations would not want to be forced to reveal their business secrets?

- If all Data Scientists use the same small number of analytics libraries and sets of training data, will we accept the higher impact if one of these is biased?

- Will we end up having ethical standards that will primarily depend on the impact of an algorithm? Will data scientists be able (or even be obliged) to check the law before selecting or developing the best-suited algorithm for a given problem? Will organizations develop ethical guidelines (“Responsible AI Policy”) to gain trust beyond what the laws demand?

- How complex will it be for an individual Data Scientist to comply with all of these “nontechnical” guidelines? How complex will it be for a Head of Data Analytics to ensure the entire team behaves in a compliant way? Will it be possible to audit compliance?
This discussion is not based on the assumption that we can make all existing AI algorithms sufficiently transparent. But it will certainly become possible for some of these algorithms, and other algorithms may allow for partial transparency. Imagine, for instance, a two-step algorithm that uses the first phase to develop a set of rules based on training with test data. The second phase would then apply the most appropriate rule, which is possible in near real time as the complex calculation happened during the first phase. Such an algorithm can share which rules are selected and applied during the second phase.

You still don’t know why these rules were found to be the most appropriate ones in a particular case, but at least you can check for potential flaws where a rule turns out to be objectively wrong or discriminating.

Finally, a trade-off between effectiveness and “explainability” of algorithms may become both possible and necessary. AI solution providers may be forced to reduce the effectiveness of algorithms in order to comply with laws that ask for a minimum degree of transparency.

Such regulation can be expected to come from the European Union, as can be read from their White Paper on Artificial Intelligence, published in February 2020 (European Commission, 2020). Here, transparency has been determined as one of the Commission’s seven key requirements of their regulatory framework for AI.

On page 15 of their White Paper, the Commission states:

*The lack of transparency (opaqueness of AI) makes it difficult to identify and prove possible breaches of laws, including legal provisions that protect fundamental rights, attribute liability and meet the conditions to claim compensation. Therefore, in order to ensure an effective application and enforcement, it may be necessary to adjust or clarify existing legislation in certain areas [...]*

It is advisable to follow the development carefully, in case your organization falls under EU legislation or wants to be active in any of the EU countries. For ethical reasons, you should consider doing so no matter which situation you are in!