A Pragmatic Approach to Biases in Visual Data Analysis

Toni Verbeiren
VDA-lab, KU Leuven
ESAT-STADIUS, KU Leuven
ExaScience Life Lab
iMinds Medical IT

Ryo Sakai
VDA-lab, KU Leuven
ESAT-STADIUS, KU Leuven
iMinds Medical IT

Jan Aerts
VDA-lab, KU Leuven
ESAT-STADIUS, KU Leuven
iMinds Medical IT

ABSTRACT

Visual biases and more generally cognitive biases are a part of human life. Often to the frustration of the rational decision makers we aspire to be. Research into these biases has sparked a recent burst in interest, and more and more people are aware of possible pitfalls. In this paper, we argue that the consequences of biases during data analysis have to be considered rather than the occurrences themselves. In applying this, we distinguish between (visual) analysis for exploration and validation. Especially the latter turns out to be hard in some cases, indicated by a qualitative measure we call validation cost. Examples are provided of situations with a high validation cost and the role of visualization is discussed in these cases. For cases with a low validation cost, we argue that biases leading to false positives are far better than trying to avoid biases and ending up with false negatives.

Categories and Subject Descriptors

H.1.2 [**Information Systems**]: Models and Principles, Human information processing; I.5.0 [**Pattern Recognition**]: General; I.2.8 [**Artifical Intelligence**]: Problem Solving, Control Methods, and Search—*Heuristic methods*

General Terms

HUMAN FACTORS, MEASUREMENT

1. INTRODUCTION

The human body possesses an extremely intricate device for capturing visual stimuli: a combination of a) the eye that captures millions of pixels with b) a neur(on)al structure that converts these visual signals into relevant stimuli for different parts of the brain and body to process. The latter is largely based on fast heuristics, where a lot of information is reduced or removed because it is not deemed relevant. The irony is that the heuristics that allow us to deal with huge amounts of (visual) information without being overwhelmed also tend to leave us prey to mistakes and (in general) cognitive biases. The DECISIVe workshop deals with the question of how to avoid these biases in visualizations.

One of the first to study the biases of the brain was Francis Bacon in his Novum Organum. Much later, cognitive and behavioural sciences have embraced large scale social experiments since the ground-braking work by Kahneman and Tversky [9]. Studying those biases gives deeper insight in the workings of the human brain and can teach us how to avoid some of the issues. Visualization experts have taken these and other sources (e.g., Gestalt laws) in order to distil general rules and guidelines on how to encode certain types of data [19].

The reasons why visualizations are used in the course of an analysis are manifold. In general, though, two important reasons can be distilled: (1) to visually check or validate models and assumptions (e.g. QQ-Plot), and (2) for hypothesis generation, i.e. finding patterns during exploratory data analysis.

Point (1) has been shown empirically to be very effective [14]. Point (2) can be very effective indeed but is also prone to the ironic paradox described above: the human mind is a powerful pattern seeking device, in some cases seeing patterns that aren't really there.

The terms cognitive bias and heuristic are typically used as being synonymous to *errors* [9]. A long standing debate in behavioural sciences is whether some of the *shortcuts* the human brain employs really are *mistakes*.

Our proposed alternative definition for *heuristic* is more nuanced: a shortcut or bias in decision making, such that the effects of an error are limited. Not only is this definition more nuanced than the usual use of the term, it is also more pragmatic because it includes the notion of the effect of a bias.

This alternative definition is similar to the concept of fast and frugal heuristics and ecological rationality as used by Gigerenzer et.al. [7]. In their work, arguments are made in favour of heuristics to be more effective than complex statistical models, depending on the measure of effectiveness [6, 5]. When this measure of effectiveness is based on the outcome of the heuristic, it corresponds to our definition above.

In this paper, we provide a pragmatic or heuristics based approach to the occurrence of (cognitive) biases and other types of errors in visualization and data analysis. The central concept is the focus on the effect or impact of an event rather than its occurrence, as described in section 2. This concept is applied to the occurrence of biases and mistakes in two stages of (visual) data analysis: the exploratory phase and the confirmatory phase.

Applied to the occurrence of (visual) biases in *exploratory* *analysis*, we argue in section 3 that one should weigh the costs of validation against the (opportunity) cost of trying to debias the visualization. In section 4, we discuss the effect of biases in visualization and statistics for *confirmatory analysis*, where it is tempting to turn to statistical validation. It turns out that such a validation is not always straightforward, as indicated in section 5. We expand on the feasibility of validation by introducing the concept of *validation cost* and give examples in section 6.

At this point, we are left with a paradox. On the one hand we argue in favour of allowing for biases during exploration, mainly because false positives will be filtered out by the later statistical validation. On the other hand we describe situations where such validation is hard or even impossible. This paradox can be resolved by again focussing on the effects of an event, rather than its occurrence as applied in section 7. It turns out that a proper risk analysis based on the effects of biases and mistakes can resolve the paradox. In section 8, we provide an opinionated view on how visualization may play a role in these situations with high validation cost.

2. OCCURRENCE VERSUS IMPACT

Central to the opinion presented in this paper is the notion of the difference between the *occurrence* of a phenomenon and the exposure to it or the *effect* it has. The concept has been applied to a variety of domains [10, 18], but to our knowledge not yet within visual analysis.

We illustrate the point using a basic example: It is widely accepted that people living in a region where deadly snakes reside tend to react unconsciously and intuitively to the form of a snake [13]. In some cases, the reaction is triggered by a wooden stick lying on the ground, which could be regarded as a bias or mistake. Luckily, the impact is only a sudden boult of fear until the true identity of the object is established. So the effect of a wrong (biased) reaction is harmless.

The four possible outcomes of a snake encounter can be summarized in a table (aka confusion table) that lists the *consequence* of observing a snake (yes/no) when a snake is effectively present (yes/no). From an evolutionary point of view, the bias of observing a snake when there is no snake (*false positive*, *type I error*) is far better than not observing a snake when there is one (*false negative*, *type II error*), without reference to the underlying base rate probability of encountering a snake in the first place.

We note that this corresponds well to the working definition of a heuristic: the heuristic is such that the worst possible consequence (death) is unlikely to occur. This kind of heuristic or rule of thumb is omnipresent in daily life. Most strangers we meet on the street are perfectly sane and nice people, but still most parents will tell their young children not to trust strangers. This is a case of an enormous amount of false positives in favour of one *real* false negative.

3. VISUAL BIASES IN EXPLORATORION

Roughly speaking, we distinguish 2 phases in a decision making process based on data: 1) *exploratory analysis* and 2) *confirmatory analysis*.

In the course of an exploratory analysis, we come up with hypotheses to later verify these hypotheses in a confirmatory analysis or validation. Generally speaking, the confirmatory phase consists of a statistical test and it is the exploration that requires the most effort.

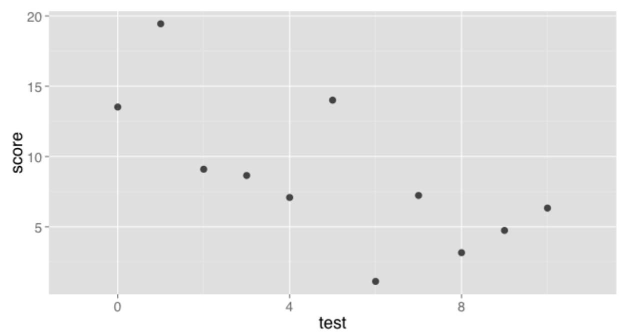

Figure 1: Scatterplot of the example of the student's results on 11 consecutive tests.

In what follows, we consider the example of the scores of a student on 11 consecutive exams:

13.5, 19.4, 9.1, 8.7, 7.1, 14, 1.1, 7.2, 3.2, 4.7, 6.3

The numbers are artificially generated (see further) and could signify other systems like stocks, expression values, temperature readouts. We refer to the scatterplot in Figure 1 for a visual representation of the data.

Given the student scores above and based on Figure 1 we might hypothesise that a downward trend exists. This trend may or may not be *really* present in the data until tested using appropriate statistical methods in the confirmatory phase.

A false positive during analysis means that we notice a downward trend where this cannot be objectively shown. A false negative during analysis means that *some* pattern is present in the data, but we do not observe it, at least not using the given visual encoding or representation.

Especially in exploratory analysis, visualization plays an important role, just because a) we are good at spotting patterns and generating hypotheses in this way and b) a statistical test is often enough to (dis)prove the hypothesis.

Both false positives and false negatives may be the result of a bias. As mentioned already, humans tend to see patterns that are not really there, which leads to false positives. We argue that this is a good thing for science and society in general. In our opinion, it is far worse for the sake of development to miss patterns that are present in the data than to see a pattern where there is none. Especially so when checking the (false) hypothesis is relatively easy and cheap, e.g., by using a simple statistical test. In other words, the finding of false positives is encouraged during exploration.

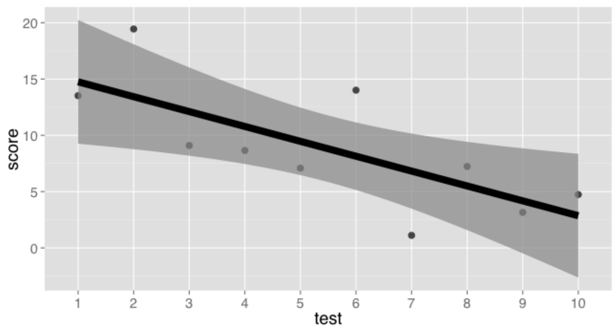

Figure 2: The data from Figure 1 with addition of the 95% confidence interval for linear regression.

4. VISUAL BIASES IN CONFIRMATION

Statistical validation is usually easy and fast. Let us illustrate this by means of the example above. Say you are making a bet for a considerable amount of money that the student's 12th exam will turn out to have a score above 10. Suddenly the downward trend that we noticed becomes important because there's money involved, and we should start to consider a kind of validation of our hypothesis.

In Figure 2, we present the same data as before, but with the 95% confidence interval added by means of a simple linear regression. Based on this, we can conclude that a student score above 10 is extremely unlikely and we can thus be relatively confident of a betting strategy that makes use of this.

This is how these kinds of questions would typically be handled in data analysis or statistics. And the nice thing is that the result of the regression analysis nicely confirms our perception of the tendency in the data. The introduction of a simple statistical validation ensures that possible biases are mitigated.

5. FOOLED BY THE ARGUMENT

In the above, we argue that:

1) Biases and consequently false positives are not a concern during exploration when considering the *impact* of the bias rather than the *occurrence* of the bias.
2) The impact of a possible bias can easily be mitigated by means of a statistical validation during confirmation analysis.

The above conclusion may seem convincing, but it is flawed. Let us illustrate this by means of the example we used earlier. We used a statistical test to back our intuition that a downward trend is apparent in the student scores.

The dataset was generated by the authors as a set of points drawn independently from a normal distribution with mean 13 and variance 10. The probability of the student having a score bigger than 10 on the next exam is 62%. Much higher than the probability estimated from the linear regression method.

The primary reason for this discrepancy is that the data size is too small for linear regression to be valuable. Truth be told, for the sake of the argument we drew a lot of random numbers and selected a range of 11 datapoints with a downward trend. Increasing the confidence interval, one notices that a horizontal line (no downward trend) is one of the possible outcomes of the analysis, but not for 95% significance. And even then, it still means that a score of 10 is very unlikely.

In other words, even the simple example of 11 points turns out not to be simple at all.

It would be unfair to conclude that we fell prey to a visual bias in this case, since even the statistical modelling approach led us to the same (wrong) conclusion. It simply indicates that whenever statistical validation is involved, we make assumptions about the problem which may be wrong which in turn means that modelling may be hard and thus validation is hard.

In what follows we introduce a qualitative measure for the feasibility of (statistical) validation and give other examples of situations where validation is hard.

6. VALIDATION COST

We notice that the main difficulty with data analysis is largely in validation of the possible hypotheses. The world around us is uncertain and a proper validation is often not possible or feasible: there is a cost associated with validation and confirmation that may be too high to bear. The *cost* here is a generic term that refers to economic, emotional and other factors involved.

We illustrate this with the following examples.

6.1 Experiment setup

In medical and cognitive tests, double blind studies are used in order to draw statistically relevant conclusions. Such a study however requires considerable effort from the researchers involved. What if a mistake in the experiment's setup is found during the experiment?

The experience with the process of drug discovery and approval teaches us that validation costs can be very high. Many chemical compounds are potential drugs, but some of them can be toxic. The process of drug discovery is very lengthy, and rightly so, but thus also very costly. One does not want to risk people's lives in an attempt to cure a disease.

6.2 Modelling errors

We have encountered a practical example of modelling errors above in the example of the student's test scores. One of the reasons the modelling approach was not working, is the lack of sufficient data points. With insufficient data, a rigorous statistical analysis is not possible.

6.3 Fat tails

In many real-life cases, statistics can be gathered, but the underlying probability distributions turn out to be fat-tailed [4]. This means that extreme events are not all that exceptional. The problem is that statistically valid claims require lots of data to sample from, which is data we do not generally have.

This in turn makes modelling extremely hard or even impossible. It is almost as if there will never be sufficient data for validating the model.

From an abstract point of view, one could argue that the financial crisis was a consequence of modelling (or better, validation) errors due to the fat tailed nature of the probabilities involved [12]. It has proven that the potential negative impact of an event in the banking sector may have a huge impact on our societies.

For the same reason, visualization for confirmatory analysis in these kind of situations is impossible: Imagine a stock trader looking at a screen with a trace of the stock movements for the last 20 years or so. Depending on this graph he or she makes a decision on whether to sell or to buy. Neither the visualization, nor the models discussed earlier will be able to provide a definite answer.

6.4 Undesirable consequences

In some cases false positives may lead to undesirable consequences, even if validation is possible [20]. Say a cheap and quick test reveals a certain disease with a high rate of false positives. Although no harm is done by the test, and a better (but more expensive) test can disprove the false results, this situation leads to increased anxiety which may be harmful to the patient and its relatives. This is sometimes referred to as the nocebo-effect [2].

6.5 Ethics

There are cases where assessing the possible impact is costly, unethical or when the impact cannot reasonably be estimated. Many of life's important questions fall into this category. It's not possible to raise the same child 10 times in order to find out which approach to parenting is the better one. Randomized trials are an attempt to resolve this, but will never be able to provide you with a definite answer to a simple question like: *Should I marry this person?*

7. MODELLING VS RISK ASSESSMENT

We are aware of the fact that we ended up in a paradox: We have claimed that false positives during exploration should not be feared because statistical validation will filter them out. But then we noticed that this validation in many cases turns out to be hard.

This brings us back to our original idea and look at the impact of a bias or, in general, mistake. Applied to the earlier example, consider the following: depending on the money we put in the wager, the impact may be small or severe. The money involved defines the outcome more than anything else. A proper *risk analysis* is in order which lets us conclude that betting is fine as long as the amount of money we bet is reasonable, less so the error-prone estimated probability of the event.

In our first attempt to make a reasonable bet, we attempted to model the student's scores using linear regression but it turned out that this was the wrong approach and could have led to losses in a real bet.

When dealing with uncertainty, generally speaking, these two approaches can be followed: (1) Devise the (very best) model that describes the data and make a prediction about the future, or (2) come up with a reasonable bet, *regardless* of the outcome, similar to a risk analysis.

In the modelling approach, we try to get a grip on the abstract probabilities. In a risk analysis approach, we consider the real-life consequences (impact) of the bet, even if we are relatively confident of the outcome. This is like making sure nuclear reactors can withstand earthquakes (even if those are extremely rare) or your house is insured against fire or flood (even though the probability of those events occurring is low).

In our example, even if we are very confident about the outcome, we should not enter into competitions that can ruin us. As a matter of fact, in general, we should prefer to make many mistakes with small repercussions rather than one mistake with big repercussions [17].

False hypotheses generated during data analysis, whether they be based on visual exploration or other techniques have a small impact. It is when we want to validate the hypotheses that it becomes harder. Betting all our money on a false positive result can mean bankruptcy.

We argue therefore to keep in mind the risks associated with bad outcomes, however unlikely they may appear to be. Please note that perception and biases in perception become less relevant in this context.

Given the examples provided in section 6 it appears that a heuristic can be distilled from this: *the higher the validation cost, the higher the possible negative impact of the outcome.*

There is a spectrum of possible outcomes ranging from little annoyances to extremely severe. The irony is that the harder it is to assess the impact, the bigger the potential harm. Bankruptcy, death, etc. are outcomes that provide no fallback scenario.

In fact, another powerful heuristic may be derived from this observation: *Don't bet your life on models in situations with a high validation cost.* Or in other words, the higher the validation cost, the more emphasis we should put on risk analysis rather than (statistical) modelling.

8. VISUALIZATION WITH A CAUSE

In situations with a low validation cost, the effect of biases turned out to be minimal. As a consequence, we should avoid spending more effort in debiasing the visualization than it would be to statistically validate or invalidate a possible false positive.

In situations with a high validation cost, we have seen that we wouldn't even be able to point out the false positives because statistical validation is hard or even impossible. By

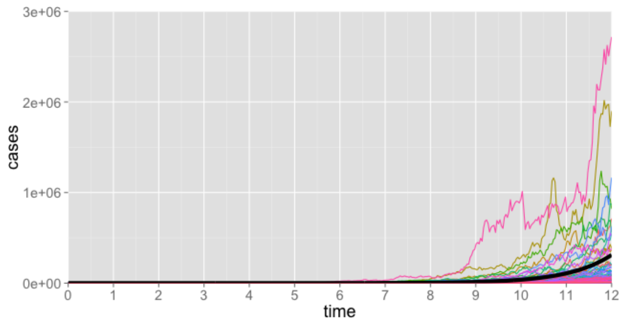

Figure 3: An illustration of the effects of uncertainty in a model for the spreading of a virus. The coloured time series represent possible trajectories. The black line corresponds to the evolution without added multiplicative noise.

focussing on the impact of possible outcomes, however, a qualitative risk analysis can often be done. As a consequence, rather than trying to debias the analysis of an event, we should consider using visualizations in order to illustrate the possible outcomes of the event.

We illustrate this by means of an example. The recent Ebola epidemic in West Africa has raised a lot of discussions, especially in the US, on how to treat people travelling from infected regions. The debate was mostly concerned with the risks of a US epidemic. Stories and visualizations have been created in an effort to make sure the general public would not panic [11, 3].

Most of these analyses are irrelevant in the light of the worst-case outcome of an infection in an urban area, especially if you take into account the multiplicative and fat-tailed nature of the probability distribution of an outbreak [1, 15]. The probability of an epidemic may be small, but not as small as one might think based on a simple visualization or analysis. Moreover, statistical validation is next to impossible. The impact of such an unlikely adverse event though is tremendous.

Efforts to create visualizations should therefore focus, not on choosing sides or estimating the probabilities of an outbreak, but rather on creating *awareness* of the possible impact of such (admittedly rare) events. A simple visualization, showing the difference between a stochastic process governed by a fat tailed probability distribution versus a thin tailed one is sufficient and far more effective than any numerical or analytical argument.

In Figure 3 we illustrate this by means of a very simple simulation using a simple model of virus spreading based on [16]. A hundred random geometric Brownian motion paths have been generated from a very simple exponential growth model that is drawn in black. At every time step, the average growth rate (taken to be 12.65 or a doubling period of 20 days) is modified using a random normally distributed variable with variance 2. In other words, the number of infections per person is allowed to vary in time. For a discussion of factors that may influence transmission rate, we refer to [1].

As it turns out, the majority of paths result in less infected cases than the fixed exponential model. In our simulation, only 13 paths out of 100 end up worse than the fixed model and the average number of infections after one year by adding the random noise is around half of the total infections for the fixed model. However, there is a large difference in outcome for some of the extreme paths. The reason for this discrepancy is the fact that as a result of adding multiplicative noise to the growth rate, the resulting probability distribution of outcomes is fat-tailed. Large deviations are therefore more likely, as can be observed. Using this simple model, we immediately get an intuitive understanding of the dangers of modelling the spreading of viruses without understanding the risks that are associated with small errors or uncertainties in the model.

In other words, just as regulators start to make use of similar arguments used in this paper [8], we should encourage visualization designers to also focus on impact or consequences of events.

9. CONCLUSION

Cognitive and visual biases occur all the time, usually without people being aware of them. The brain's visual cognition system is so good at distinguishing patterns, it sometimes recognizes too many of them.

In this position paper we argue that instead of focussing on the biases themselves, it makes more sense to look at their consequences. We have seen that in an exploratory analysis, cognitive biases (i.e. noticing non-existing patterns) may lead to false positives. We argue that for the sake of development and research, false positives are far better than false negatives. In other words, the cost of attempting to debias a (visual) analysis during exploration should be weighed against the cost of invalidating false positives.

In a validation phase, visualization may lead us astray but more rigorous methods have an even chance of resulting in errors, mainly because of the difficulty of correctly validating hypotheses. In this case we argue for a risk analysis approach where again the impact of possible adverse events is assessed rather than their theoretical probabilities.

Putting too much emphasis on trying to avoid errors due to visual biases may distract us from the larger and more fundamental picture: The world around us is uncertain and we will never be able to be sure. The challenge is to avoid mistakes with big impact. Visualizations may be used effectively for creating awareness about this possible impact, rather than focussing on the occurrences or their estimated probabilities.

Acknowledgments

We thank Geoffry Ellis and the reviewers for stimulating questions and comments. This project was funded by KU Leuven program financing PFV/10/016 SymBioSys, IWT

O&O ExaScience Life, Intel, Janssen Pharmaceutica and H2020-RIA Virogenesis.

10. REFERENCES

[1] Y. Bar-Yam. Risking it all: Why are public health authorities not concerned about ebola in the us? part i. fat tailed distributions. http://arxiv.org/abs/1411.1472.

[2] F. Benedetti, M. Lanotte, L. Lopiano, and L. Colloca. When words are painful: Unraveling the mechanisms of the nocebo effect, 2007.

[3] P. Bump. Want to feel better about ebola? this (massive) chart should do the trick. http://www.washingtonpost.com/news/the-fix/wp/2014/10/16/want-to-feel-better-about-ebola-this-massive-chart-should-do-the-trick/.

[4] R. M. Cooke, D. Nieboer, and J. Misiewicz. Fat-tailed distributions: Data, diagnostics, and dependence. http://www.rff.org/News/Features/Pages/Understanding-Fat-Tailed-Distributions-and-What-They-Mean-for-Policy.aspx, 2011.

[5] A. Dijksterhuis, M. W. Bos, A. van der Leij, and R. B. van Baaren. Predicting soccer matches after unconscious and conscious thought as a function of expertise. *Psychological science*, 20(11):1381–7, Nov. 2009.

[6] G. Gigerenzer. Gut Feelings. In *Gut Feelings: the intelligence of the unconscious*, pages 3–20. 2008.

[7] T. P. Gigerenzer G. *Simple Heuristics That Make Us Smart*. New York, NY: Oxford University Press, 1999.

[8] S. Jha. The antifragile cdc. http://thehealthcareblog.com/blog/2014/10/20/the-antifragile-cdc/.

[9] D. Kahneman. *Thinking, Fast and Slow*. 2011.

[10] S. Makridakis and N. Taleb. Living in a world of low levels of predictability. *International Journal of Forecasting*, 25:840–844, 2009.

[11] D. McCandless. Visualised: how ebola compares to other infectious diseases. http://www.theguardian.com/news/datablog/ng-interactive/2014/oct/15/visualised-how-ebola-compares-to-other-infectious-diseases.

[12] A. Murphy. An analysis of the financial crisis of 2008: Causes and solutions.

[13] A. Ohman. Has evolution primed humans to "beware the beast"? *Proceedings of the National Academy of Sciences of the United States of America*, 104(42):16396–7, Oct. 2007.

[14] E. Soyer and R. M. Hogarth. The illusion of predictability: How regression statistics mislead experts. *International Journal of Forecasting*, 28(3):695–711, July 2012.

[15] N. Taleb. Ebola, including model error. https://drive.google.com/file/d/0B8nhAlfIk3QIdlJNSXo1dkMzOVU/view.

[16] N. Taleb. Ebola, including model error. https://drive.google.com/file/d/0B8nhAlfIk3QIdlJNSXo1dkMzOVU/view.

[17] N. Taleb, R. Read, R. Douady, J. Norman, and Y. Bar-Yam. Precautionary principle. *under revision*, 2014.

[18] N. N. Taleb. *Antifragile: Things That Gain from Disorder*, volume No. 203. 2012.

[19] C. Ware. *Information visualization: perception for design*. 2004.

[20] S. Woloshin and L. M. Schwartz. How a charity oversells mammography. *BMJ*, 345, 2012.