

SYMPOSIA PAPER

Exploratory Analysis and the Expected Value of Experimentation

Colin Klein 

School of Philosophy, The Australian National University, Canberra, Australia
Email: colin.klein@anu.edu.au

(Received 13 March 2023; revised 06 July 2023; accepted 13 August 2023; first published online 15 September 2023)

Abstract

It is increasingly easy to acquire a large amount of data about a problem before formulating a hypothesis. The idea of *exploratory data analysis* (EDA) predates this situation, but many researchers find themselves appealing to EDA as an explanation of what they are doing with these new resources. Yet there has been relatively little explicit work on what EDA is or why it might be important. I canvass several positions in the literature, find them wanting, and suggest an alternative: exploratory data analysis, when done well, shows the *expected value of experimentation* for a particular hypothesis.

1. Introduction

Exploratory data analysis (EDA) has something of a mixed reputation. On the one hand, most researchers readily admit to doing something they call exploratory data analysis. EDA is increasingly common in the era of big datasets, especially when there is more data available than there are questions to be asked. New preregistration protocols encourage researchers to establish exploratory protocols alongside confirmatory ones. And there is a long-standing belief that exploratory and confirmatory data analysis are fruitfully, perhaps ineliminably, paired. On the other hand, exploratory data analysis is also blamed for many scientific ills. Critics such as Wagenmakers et al. (2012) suggest that exploratory analysis is nothing more than confirmatory analysis done sloppily.

Fuelling this debate is the lack of an accepted theory of what EDA actually amounts to. Everyone is clear that EDA differs from confirmatory data analysis. But it is surprisingly hard to figure out *how*.

At the heart of this, I suggest, is what I will call the *Central Puzzle* of EDA. Everyone agrees that EDA has an important role in generating hypotheses. Merely generating hypotheses is easy enough—idle speculation isn't the hard part of science. What we want are *good* hypotheses. Yet what could a good hypothesis be, except one that *might be true*? But if the goal of EDA is to generate hypotheses that might be true, then it

seems like it's doing the same thing that confirmatory data analysis is doing. Hence the Central Puzzle: what's the difference?

I suspect the lack of an answer has also contributed to the somewhat seedy reputation of EDA. We know the aims of confirmatory data analysis: to confirm or disconfirm scientific hypotheses. There is a voluminous technical literature on whether particular methods (like null hypothesis statistical testing) are an appropriate means to that end. By contrast, EDA is often presented in a way that makes it sound like a freewheeling, laid-back cousin of proper confirmatory analysis, one with vague plans and low standards.

Beyond rehabilitating its reputation, EDA also needs a proper theory for its own sake. There are also whole manuals devoted to the *methods* of EDA. Yet a positive characterisation of the *aims* of EDA is also necessary. Without one, it can be difficult to know whether one is doing EDA well or poorly. That means that it can be hard to evaluate whether a new proposed method is a good one—for surely being a good method for EDA consists, at least in part, in doing a good job at serving the aims of EDA.

I think there's a straightforward enough answer to the Central Puzzle. I suggest that EDA informs us about *expected value of experimentation*. This gives a neat solution the Central Puzzle. My account is meant to be partly rational reconstruction and partly rehabilitation: it's not always what people have had in mind, but it is what they should have in mind. I will then consider three other (more or less skeptical) suggestions in the literature, and show that my account captures what's good about them while avoiding obvious issues. I conclude with a discussion about how a more formal treatment of EDA's aims might be developed.

2. The expected value of experimentation

When we think of investigating hypotheses, we usually think of getting information that bears on their truth. Yet it is also possible to get information about a hypothesis that tells you not about whether it is true, but rather about how likely it is that you are to get useful information about truth if you perform the right kinds of tests.

A toy case first. You're a detective investigating a high-profile murder. You think the deed was done by either Tony or Louie. A wiretap will stretch your resources; you can't afford to tap both, and you're not sure if it's worth tapping either. Your informant knows nothing about who did the hit, but can give you useful information about their reputation. Louie is a loudmouth: he frequently brags about his crimes. Tony, on the other hand, is notoriously tight-lipped about his misdeeds.

Given this information, you decide it's reasonable to tap Louie. Why? Not because you've gotten evidence about Louie's guilt. The informant doesn't know anything about that. Your other evidence may lead you to think that Tony is the more likely culprit. What you have learned is something about the usefulness of performing a certain kind of test. That is, you've learned something about the *expected value of experimentation* (EVE): the likelihood that if you perform an experiment, the effort will be repaid.

Of course, if you had infinite money, you could tap both. But resource limitations demands what Good calls "Type II rationality," where the reasonable thing to do depends in part on "the expected time and effort taken to think and do calculations."¹

¹ See Good (1983, 16). For an explicit link between Type II rationality and EDA, see also Good (1983, 291).

Scientific experimentation is also a paradigm of Type II rationality. Scientific experiments take resources—time, grant money, postdoc-hours—and different experiments compete for access to resources (Klein 2018). Hence, performing any particular experiment comes with opportunity costs: testing one hypothesis is failing to test many others. Which hypothesis should be tested is a function of how likely that hypothesis is to be true, how costly it is to test the hypothesis, and how likely it is that the test will get a definitive answer for one's pains.

Often the information needed for estimating the expected value of future experiments can come from past experiments. As Tukey (1993, 4) notes, "... the process of formulating the question and planning the data acquisition will, often, have to depend on exploratory analyses of previous data." This might include information about the testing process itself (we hoped we could run three phases of surveys, but the data are so messy from phase one that we know we will need more subjects in phase two, and so have to scotch phase three ...). Sometimes past experiments are part of a hypothesis-testing procedure, but also give information about the future value of tests by ruling out some possibilities. Before sending my sparkly ore sample off to have an expensive assay for gold content, I might first see if it dissolves in strong nitric acid. If it does, it's only fool's gold, and further tests are pointless.

Such tests can form part of an iterated strategy. Consider the use of Bayesian search theory (Stone 1976), best known for its use in looking for underwater wrecks. To find (e.g.) a lost submarine, one divides the search area into a grid, estimating for each square the chance that the wreck is present there, the chance that the wreck will be found if you look for it, and the cost of searching in that square. The optimal strategy at each step (in the single-wreck case) is simply to search the square with the best expected utility (Assaf and Zamir 1985). However, each failed search triggers an update of the grid parameters—most obviously, the chance of finding the wreck, but also the cost of search as well as anything new you've learned about your model. A failed search is still useful, because it can update your estimation of where the next most promising places might be.

In sum, all manner of evidence—observational or experimental—can be relevant to determining EVE, and in many cases this is something that is assessed in an ongoing answer alongside confirmation of particular hypotheses.

3. EDA and the expected value of experimentation

Let's return to exploratory data analysis. I propose that the aim of EDA is, fundamentally, to estimate the expected value of experimentation. EDA is distinguished from confirmatory analysis by (i) using data specific to the problem, and (ii) focusing on all of the aspects that go into estimating EVE except those that depend on our estimation of the truth of the hypothesis. The first condition distinguishes it from other things that might go into EVE but which are not data-specific (e.g., checking Edmund Scientific for the cost of reagents), while the second distinguishes it from confirmatory analysis.

One advantage of this view is that it neatly solves the Central Puzzle. Exploratory data analysis is worth doing because it gives us information about what hypotheses to test—but that comes from all of the *other* things that go into estimating EVE aside from truth. That shows why EDA is particularly valuable in young fields, or in fields

that use very large datasets. When it is easy to come up with hypotheses but costly to properly test them, EDA guides us towards the tests worth doing. That said, EDA can play a role even in mature fields, because estimating EVE is always an important step in allocating scientific resources.

This is a broad hypothesis about the aims of EDA, but I think that, having adopted it, one can see it play out in specific methods. Consider Tukey's frequent advocacy for graphical methods. Tukey gives an example of Raleigh's investigation on the density of nitrogen, and the differing values he obtained for nitrogen obtained via removing oxygen from air and that estimated via other means (Tukey 1977, 49ff). A dot plot reveals two groups of results, separated by a distance clearly more than noise. This is not definitive of anything, and indeed does not yet specify any particular hypothesis (in fact, the gas obtained from air was not pure nitrogen; the result led to the discovery of argon). It doesn't even conclusively establish a difference: the setup was not meant to test for such a difference, after all. Rather, it suggests that something is going on, and that *whatever is happening is striking enough to be worth further investigation*.

The example of Raleigh's measurements also brings up a common theme that runs throughout many exploratory analyses: the estimation of potential effect sizes. Visual methods are useful precisely because they make certain kinds of relationships salient: dot plots for clustering, box-and-whisker plots for skewed distributions, log plots for exponential relationships, and so on (Tukey 1977). Indicators of effect size "jump out at you" upon appropriate graphical presentation. Estimation of effect size is a crucial ingredient for evaluating EVE, because it determines statistical power of designs and therefore how much data will need to be collected for a proper experiment. When hunting for novel hypotheses, we also want to find effects that are large enough that they'll be easy to see again.

Note that large effects in preliminary data can be misleading: flukes occur with surprising regularity, especially if you are not controlling for them. So the fact that an apparent effect is large need not be a reason to believe it is real. Rather, it is a reason to believe that, *if it is a real effect*, it will be easy to find again.

Of course, as practitioners such as Tukey insist, EDA comprises a set of techniques that *may* be used rather than a recipe for success. That is for many reasons, but one simple one is that estimating EVE is itself dependent on many factors, often murky. Further, we often satisfice when we hunt for hypotheses to test. I don't need the best possible hypothesis to test. The space of possible hypotheses is very large. Really, what I need is a hypothesis that is likely to repay effort and be interesting. So there is incentive to proliferate the methods of EDA to include anything that might be usefully brought into service of estimating EVE.

More subtly, this view on EDA also makes clear why many of the techniques that fall under its traditional umbrella have the relatively simple form that they do. In a retrospective look, Tukey (1993) noted that many of the techniques in his classic work involved graphical visualisation—"an emphasis on *seeing* results" (5)—but also very simple arithmetical methods involving as little probability as possible. That makes sense. When one is looking for patterns that might be worth testing, one generally wants to avoid complex procedures. Not just because they might confuse with false hope (though that too), but because simple patterns that jump out visually are more likely to be meaningful the less processing one has to do to reveal them.

4. Advantages of the proposal

Thinking of EDA as fundamentally aiming to estimate the expected value of experimentation has a number of advantages. I think it also does better than several rival proposals in the (admittedly small) philosophically oriented literature on EDA, while at the same time capturing what seems attractive about those proposals. I consider in turn what I take to be the three main rivals.

4.1. EDA as hypothesis generation

The phrase “exploratory data analysis” was popularised by the work of Tukey (1969, 1977, 1993), who also did quite a bit to introduce graphical techniques for investigating data. Sometimes these were an aid to more compact representation of descriptive statistics and first-pass sanity checking of data.² On this view the primary reason to use EDA is as an aid to *generating hypotheses* (Tukey 1969, 1993; Good 1983). Displaying things graphically can help reveal patterns in the data, and these patterns are the nuclei of new research questions.

It is common to claim that EDA generates hypotheses. Yet there are three problems that arise when you try to pin down this idea further. First, how hypotheses are meant to be generated is a bit of a mystery. Tukey emphasised the “procedure-oriented” nature of exploratory analysis and the extent to which these techniques were “things that can be tried, rather than things that ‘must’ be done” (1993, 7). Hartwig and Dearing (1979, 10) similarly speak of EA as a “state of mind” or a “certain perspective” that one brings to the data (Hartwig and Dearing 1979, 3; see also Jebb et al. 2017, 257). This is hardly the precision one hopes for: why some techniques rather than others do the trick seems idiosyncratic, or at best, as Good (1983) suggests, a question for neuroscience.

Second, it’s not clear how we rule out anything at all on this account. Is having a couple of beers at the pub a form of EDA?³ A lot of scientists seem to come up with good hypotheses that way. But if that counts, we’ve hardly shaken EDA’s wild-and-woolly reputation.

Third, and most importantly, this does not sufficiently distinguish EDA from confirmatory analysis. Recall that the Central Puzzle arose precisely because “hypothesis generation” would appear to be pointless unless we’re generating hypotheses that might be true. So we have not really distinguished the aims of EDA and those of confirmatory analysis.

By contrast, thinking of EDA as fundamentally concerned with evaluating EVE gives it a more focused aim. Sometimes we do want to brainstorm hypotheses, and that may well be a largely unconstrained process. Yet the role of EDA is not fundamentally for suggesting hypotheses. Rather, it suggests hypotheses that would

² My data science colleagues sometimes use “exploratory analysis” in this weaker sense, as a collection of methods for data preprocessing and first-pass sanity checking—removing outliers, noting obvious collection problems, and so on. This strikes me as an area that is as much craft and skill as anything; nobody claims that there is a good theory behind it, only a lot of experience about the very many ways data collection can go wrong.

³ I ask rhetorically, but consider Good’s (1983, 290) remark that “One of the devices for making discoveries is to sleep on a problem. There is no reason why this device should not be used for EDA.”

also be worth your time to test. So my account captures why EDA and hypothesis generation go hand in hand, without identifying one with the other.

4.2. EDA as second-rate confirmation

In contrast with Tukey's optimism, there are real skeptics about EDA. One increasingly common view is to suggest that the Central Puzzle does not actually have a resolution. Instead, EDA is just confirmatory analysis done poorly (Wagenmakers et al. 2012; Nosek and Lakens 2014). So, for example, exploratory data analysis might be what happens when you fail to specify hypotheses in advance, or test multiple hypotheses without adequate correction, or choose a lenient alpha for significance testing, or exploit freedom in parameter choice, or whatever. (Rubin (2017) gives a useful review of the statistical end of this debate.)

As you might imagine, such a view is rarely offered as a *defence* of EDA. Recent versions usually come in the context of hand-wringing around the replication crisis in psychology.⁴ Nor is the charge completely unfair. It is probably the case that at least some abuses are papered over by suggesting that an analysis is merely exploratory, where that is code for taking advantage of parameter freedom.

Yet as a general diagnosis, this strikes me as altogether too pessimistic a conclusion. Some researchers are sloppy, but the charge can hardly be sustained against the likes of Tukey or Good. Charity demands that we prefer a model where authors who appeal to EDA are not simply covering up their sins as researchers. Nor does such a position really explain why there is a separate, and quite extensive, tradition of techniques associated with EDA. Such pessimism may be a final resting place were we unable to find a better theory. But I have already suggested that we can do better.

That said, thinking of EDA as estimating EVE also makes it easy to see why EDA can be mistakenly substituted for confirmatory analysis. In a large dataset with many comparisons to be made, reducing significance levels might reveal effects that are interesting, relatively large, but also noisy. So they do not pass an appropriately rigorous test. These hypotheses are ones worth further testing: if the noted effect really is large (one might reason) it should show through cleanly on subsequent, more targeted, tests.

However, it would be very easy to confuse this sensible claim with a claim that there is a strong chance of the hypothesis being *true*. And of course this need not be the case. Lax significance levels also open you up to large flukes that will be eliminated upon closer inspection. Again, what one learns from this sort of EDA is that *if* there is an effect, it will be easy enough to find that it will be worth the resources to test further. But this is a subtle distinction, and one that's easy to elide in the pursuit of publication. As Wagenmakers et al. (2012, 634) warn, "Academic deceit sets in when this does not happen and partly exploratory research is analysed as if it had been completely confirmatory." My account explains why this is tempting, but also shows why it is not inevitable.

⁴ See, e.g., Nosek and Lakens (2014); Szollosi and Donkin (2021). The American Psychological Association (2020, §3.6) guidelines are clearly meant to be defensive in this regard.

4.3 EDA as exploratory experimentation

Third and finally, EDA is sometimes linked to so-called *exploratory experimentation* (Burian 1997; Steinle 1997; Franklin-Hall 2005; Feest and Steinle 2016). Exploratory experimentation, as Feest and Steinle (2016, 9) put it, is “a type of experiment that consists in systematic variation of experimental variables with the aim of phenomenologically describing and connecting phenomena in a way that allows a first conceptual structure of a previously poorly structured research field.” Franklin-Hall (2005) notes that exploratory experimentation is especially important in cases where experimenters use “wide” instruments that “allow scientists to assess many features of an experimental system” (896). These include techniques like fMRI or broad DNA microarrays (898)—classic cases of systems that generate large amounts of data, often in the absence of particular hypotheses. Furthermore, Steinle notes that an important part of exploratory experimentation is the stabilisation of phenomena (Steinle 1997). This idea, that experimentation is meant to draw out an underlying pattern in noisy data (Bogen and Woodward 1988; Feest 2011), seems quite similar to some standard claims about EDA.

While EDA and exploratory experimentation have many similar features, I think it is important to keep them distinct. For one, EDA can be done in the absence of experiment. We can (and do) do exploratory analysis on pre-existing convenience datasets. Exploratory experimentation is usually discussed in the context of immature fields where hypotheses have yet to be formulated or concepts stabilised. EDA finds plenty of use in mature fields that are not in flux. Finally, exploratory experimentation is clearly aimed at finding out the truth, which makes it difficult to appeal to for a solution to the Central Puzzle.

Nevertheless, there is an important confluence between the two ideas. The fields in which exploratory experimentation is most common are *also* often the ones in which the expected value of particular experiments will be hard to determine. The use of wide instruments can also generate large amounts of relatively unstructured data, and the techniques of EDA can be useful in investigating that to find hypotheses worth testing. Conversely, as noted above, well-chosen exploratory experiments can generate data for EDA, creating a virtuous cycle. So while the two concepts do not reduce to one another, thinking about EVE helps illuminate how they are related.

5. Conclusion: Reviving a logic of discovery

Thinking of EDA in terms of expected value of experimentation opens up a number of interesting possible avenues of elaboration. The obvious connection would be to formal theories about the expected value of search and of gaining more information. I mentioned one such formal treatment above, that of Bayesian search theory. There is also a rich tradition in economics of techniques for estimating the expected value of information (Arrow 1984) that might be brought to bear.

I would like to conclude, however, with a slightly more ambitious proposal. Simon (1973) suggests that discovery—here considered as the generation of useful hypotheses—ought to be seen as governed by internal norms. These norms (like the norms of chess strategy) do not give step-by-step directions for discovery, but do provide guidances as to practices that might be better or worse for one’s aims.

Simon credits the idea of a logic of discovery to Hanson,⁵ suggesting that Hanson's focus on visual pattern-matching has wrongly led people to assume that Hanson was merely concerned with psychology. In fact, says Simon, if we think of one of the core operations for discovery as pattern detection, then computers can do discovery too—discovery is not a mysterious, purely psychological, affair. Simon thus postulates that from this point of view, “[t]he normative theory of discovery processes can be viewed as a branch of the theory of computational complexity” (Simon 1973, 477). In other words, what we are looking for when we theorise about discovery are algorithms or heuristics that give effective ways to extract patterns from data, given our interests and goals. Those patterns (such as ones suggesting a large effect size) can then guide future experimentation.

Simon gives a few toy examples of such procedures, noting several simple algorithms for finding repeating patterns in strings. Simon was writing in 1973, when the science of pattern detection was very much in its infancy. Modern machine learning is fantastic at detecting patterns in data. And indeed, many of the places you encounter exploratory claims these days are precisely those that involve the application of machine learning techniques to very large datasets.

Now, the nice thing about following Simon's suggestion is that there is a fair bit of meta-discussion about particular techniques and what they do or don't show under different conditions. Take clustering algorithms, for example. Clustering is in many ways a natural, multi-dimensional extension of Tukey's dot-plot of Raleigh's data. There are a broad number of clustering algorithms available. Broadly speaking, there is good agreement about the background assumptions of each, the relative usefulness in different cases, and (importantly) their failure modes where they will produce wrong or uninterpretable results.⁶

As Von Luxburg et al. (2012) note, the goodness of clustering algorithms cannot be established independently of a particular project or set of ends. Relative to such ends, however, there are plenty of norms that we can establish—thereby establishing a concrete logic of discovery, in Simon's sense. The nice thing about this perspective is that we can both give mathematically precise formulations of the conditions under which certain techniques would be expected to work well, *and* how we might use the techniques when we're not sure that the conditions actually obtain. Hence the hypothesis-generating function of (e.g.) clustering techniques can be performed in a way that is truly exploratory. Furthermore, many of the standard bits of advice for the use of such techniques—such as separating a training and a test set, or looking for robustness of a result over multiple permutations—can be seen as ways to ensure that one finds hypotheses that can later be fruitfully tested.

Far from a second-rate, norm-free warmup to real science, then, the techniques of EDA can be seen as governed by their own set of norms. They work just in case they give researchers an accurate estimate of the expected value of future confirmatory experiments. But, equally importantly, confirmatory experiments provide the foundation for better estimates of EVE. By viewing EDA as a distinct but respectable practice, we thus go a long way towards showing how it can be done better.

⁵ See Hanson (1958), especially 71ff.

⁶ For a stock example, see the scikit-learn documentation for clustering algorithms at <https://scikit-learn.org/stable/modules/clustering.html>.

References

- American Psychological Association. 2020. *Publication Manual of the American Psychological Association 2020: The Official Guide to APA Style*. Seventh edition. Washington, DC: American Psychological Association.
- Arrow, Kenneth J. 1984. *The Economics of Information*. Cambridge, MA: Harvard University Press.
- Assaf, David and Shmuel Zamir. 1985. "Optimal Sequential Search: A Bayesian Approach." *The Annals of Statistics* 13 (3):1213–21.
- Bogen, James and James Woodward. 1988. "Saving the Phenomena." *The Philosophical Review* 97 (3):303–52. <https://doi.org/10.2307/2185445>.
- Burian, Richard M. 1997. "Exploratory Experimentation and the Role of Histochemical Techniques in the Work of Jean Brachet, 1938–1952." *History and Philosophy of the Life Sciences* 19 (1):27–45.
- Feest, Uljana. 2011. "What Exactly Is Stabilized when Phenomena Are Stabilized?" *Synthese* 182 (1):57–71.
- Feest, Uljana and Friedrich Steinle. 2016. "Experiment." In *The Oxford Handbook of Philosophy of Science*, edited by Paul Humphreys. Oxford: Oxford University Press.
- Franklin-Hall, Laura R. 2005. "Exploratory Experiments." *Philosophy of Science* 72 (5):888–99. <https://doi.org/10.1086/508117>.
- Good, Irving J. 1983. "The Philosophy of Exploratory Data Analysis." *Philosophy of Science* 50 (2):283–95.
- Good, Irving J. 1983. "Twenty-Seven Principles of Rationality." In *Good Thinking: The Foundations of Probability and its Applications*, 15–19. Minneapolis, MN: University of Minnesota Press.
- Hanson, Norwood R. 1958. *Patterns of Discovery: An Inquiry into the Conceptual Foundation of Science*. Cambridge: Cambridge University Press.
- Hartwig, Frederick and Brian E. Dearing. 1979. *Exploratory Data Analysis*. Thousand Oaks, CA: Sage Publications.
- Jebb, Andrew T., Scott Parrigon and Sang E. Woo. 2017. "Exploratory Data Analysis as a Foundation of Inductive Research." *Human Resource Management Review* 27 (2):265–76. <https://doi.org/10.1016/j.hrmr.2016.08.003>.
- Klein, Colin. 2018. "Mechanisms, Resources, and Background Conditions." *Biology & Philosophy* 33 (36): 1–14. <https://doi.org/10.1007/s10539-018-9646-y>.
- Nosek, Brian A. and Daniël Lakens. 2014. "Registered Reports: A Method To Increase the Credibility of Published Results." *Social Psychology* 45 (3):137–41. <https://doi.org/10.1027/1864-9335/a000192>.
- Rubin, Mark. 2017. "Do *p* Values Lose their Meaning in Exploratory Analyses? It Depends How You Define the Familywise Error Rate." *Review of General Psychology* 21 (3):269–75. <https://doi.org/10.1037/gpr0000123>.
- Simon, Herbert A. 1973. "Does Scientific Discovery Have a Logic?" *Philosophy of Science* 40 (4):471–80.
- Steinle, Friedrich. 1997. "Entering New Fields: Exploratory Uses of Experimentation." *Philosophy of Science* 64:S65–S74.
- Stone, Lawrence D. 1976. *Theory of Optimal Search*. New York: Elsevier.
- Szollósi, Ábel and Chris Donkin. 2021. "Arrested Theory Development: The Misguided Distinction Between Exploratory and Confirmatory Research." *Perspectives on Psychological Science* 16 (4):717–24. <https://doi.org/10.1177/1745691620966796>.
- Tukey, John W. 1969. "Analyzing Data: Sanctification or Detective Work?" *American Psychologist* 24 (2): 83–91.
- Tukey, John W. 1977. *Exploratory Data Analysis*. Reading, MA: Addison-Wesley.
- Tukey, John W. 1993. "Exploratory Data Analysis: Past, Present and Future." Technical Report 302 (Series 2). Research Triangle Park, NC: US Army Research Office.
- Von Luxburg, Ulrike, Robert C. Williamson and Isabelle Guyon. 2012. "Clustering: Science or Art?" *Proceedings of Machine Learning Research* 27:65–79.
- Wagenmakers, Eric-Jan, Ruud Wetzels, Denny Borsboom, Han L. J. van der Maas and Rogier A. Kievit. 2012. "An Agenda for Purely Confirmatory Research." *Perspectives on Psychological Science* 7 (6):632–8. <https://doi.org/10.1177/1745691612463078>.

Cite this article: Klein, Colin. 2024. "Exploratory Analysis and the Expected Value of Experimentation." *Philosophy of Science* 91 (5):1109–1117. <https://doi.org/10.1017/psa.2023.116>