Controlling False Discoveries During Interactive Data Exploration

Zheguang Zhao Lorenzo De Stefani Emanuel Zgraggen Carsten Binnig
Eli Upfal Tim Kraska
Department of Computer Science, Brown University
{firstname_lastname}@brown.edu

ABSTRACT

Recent tools for interactive data exploration significantly increase the chance that users make false discoveries. They allow users to (visually) examine many hypotheses and make inference with simple interactions, and thus incur the issue commonly known in statistics as the "*multiple hypothesis testing error*." In this work, we propose a solution to integrate the control of multiple hypothesis testing into interactive data exploration systems. A key insight is that existing methods for controlling the false discovery rate (such as FDR) are not directly applicable to interactive data exploration. We therefore discuss a set of new control procedures that are better suited for this task and integrate them in our system, QUDE. Via extensive experiments on both real-world and synthetic data sets we demonstrate how QUDE can help experts and novice users alike to efficiently control false discoveries.

1. INTRODUCTION

"Beer is good for you: study finds that suds contain anti-viral powers" [DailyNews 10/12]. "Secret to winning a Nobel Prize? Eat more chocolate" [Time, 10/12]. "Scientists find the secret of longer life for men (the bad news: Castration is the key)" [Daily Mail UK, 09/12]. "A new study shows that drinking a glass of wine is just as good as spending an hour at the gym" [Fox News, 02/15].

In recent years there has been an explosion of data-driven discoveries as the ones cited above. While some of these are likely to be legitimate, there is an increasing concern that a large amount of current published research findings may actually be false [20].

In this paper we make the case that the rise of interactive data exploration (IDE) tools has the potential to worsen this situation further. Commercial systems such as *Tableau* or research prototypes such as *Vizdom* [8], *Dice* [23] or *imMens* [26], aim to empower domain experts and novice users alike to discover complex relationships and trends from data in an entirely visual manner. Unfortunately these systems often ignore even the most basic statistical rules. We recently performed a user study and asked people to explore the U.S. Census data [25] using such an interactive data exploration tool.[1] Within minutes, all participants were able to derive multiple insights, such as "*people with a Ph.D. earn more than people with a lower educational degree*." However, none of the participants used a statistical procedure to determinate whether the visually observable differences in the histogram is actually meaningful (i.e., "*statistically significant*"). Further, none of the users considered that the data exploration consisting of multiple attempts to find interesting insights would considerably increase the risk of observing seemingly significant results by chance.

This problem is well known in the statistics community and referred to as the "*multiple comparisons problem*" or "*multiple hypothesis error*" and it states that the more hypothesis tests an analysts performs, the higher is the chance that apparently significant phenomenon (i.e., a "*discovery*") is actually observed just by chance. Let us assume an analyst tests 100 potential correlations each with significance level $\alpha = 0.05$. Assume further that 10 of the correlation are in fact true, and that our test has a statistical power (i.e, the likelihood to discover a real correlation) of 0.8; all very common values for a statistical testing. In this setting, the user would find ≈ 13 correlations of which 5 ($\approx 40\%$) are "bogus."

One way to lower the probability of incurring any false discovery among all the tests — known as the *family-wise error rate* (FWER) — is to use a multiple hypothesis correction procedure such as the Bonferroni correction [6]. Unfortunately, a well-known drawback of the Bonferroni correction and of may others FWER control procedures is that they lead to a significant decrease of the statistical power; the chance to detect truly significant phenomenons. Furthermore, in the context of interactive data exploration we need to cope with the ulterior complication that the hypotheses are generally unknown upfront, hence rendering any static procedures, such as the Bonferroni, correction unsuitable.

Another crucial challenge in modeling data exploration lies in the fundamental question of "*what should be considered as a hypothesis test when users interactively explore the data*." Suppose a user sees a visualization, which shows no difference in salaries between men and women based on their education, and then decides based on this insight to look at salary differences between married men and women. Should the first, the second or both visualization be considered as a hypothesis test? The answer in most cases is *both*, as the analyst probably implicitly made a conclusion based on the first visualization, which then led to her next exploration step.

However, if she considers this visualization just as a descriptive statistic of the current dataset, and makes no inference based on it (i.e. it did not influence the decision process and no inference is made by it), then it should not be considered as a hypothesis test.

[1]The results were gathered by analyzing (in retrospective) the think-aloud protocols of various user studies including the study described in [36] and entailed in total over 50 participants.

SIGMOD '17, May 14–19, 2017, Chicago, IL, USA.

ISBN 978-1-4503-4197-4/17/05…$15.00

DOI: http://dx.doi.org/10.1145/3035918.3064019

The difference is subtle and usually very hard to understand for non-expert users, while it might have a profound impact on the false discovery a user makes.

In this paper, we present the first end-to-end system, QUDE (short for Quantifying Uncertainty in Data Exploration), to automatically control the risk of false discovery for visual, interactive data exploration. We propose a user interface and an initial set of meaningful default hypotheses, i.e., the "*null hypotheses*," to control the ratio of false discoveries without interrupting the exploration process (Section 3). In Section 4 we discuss the control procedures based on the family-wise error rate (FWER), and explain why they are too pessimistic for interactive data exploration, and why the more modern criterion of controlling the false discovery rate (FDR) is better suited. The challenge of FDR, however, is that the standard techniques, such as the Benjamini-Hochberg procedure [2], are not incremental and require computing the p-values of all the hypotheses a priori before determining which hypotheses are significant. This clearly constitutes a problem for interactive data exploration where hypotheses are created incrementally. The recent α-*investing* technique [14] proposes an incremental procedure to control a variant of FDR, the marginal FDR (mFDR), together with a special-case investing strategy, Best Foot Forward. However, our experiment shows that this technique does not work well in interactive data exploration, because it was designed for a rather limited scenario where hypotheses are clustered. Thus, in Section 5 we propose new α-investing techniques that are designed specifically for interactive data exploration. Finally, we implement these ideas in QUDE and demonstrate how the system controls false discovery for experts and novice users alike using generated and real-world data.

It should be noted, that multiple hypothesis control is perhaps one of the most difficult and thorny issues facing modern statistics. Despite extensive work that explored a variety of approaches and techniques, there is *no single perfect solution*, thus neither do we claim one in this work. Instead, our main contribution is to combine and extend recent advances in statistics to build the first functional system which automatically assists the users in recognizing and controlling the false discovery during interactive data exploration. In summary, we make the following detailed contributions:

- We establish a connection between data visualizations and multiple hypothesis testing. We point out the risk of incurring a high number of false discoveries unless visual exploration systems employ corrections for multiple hypothesis testing.
- We propose a model of setting default (i.e., null) hypotheses during the interactive data exploration.
- We present QUDE, a novel system which automatically controls the multiple hypothesis error in visual data exploration.
- We discuss inadequacies of the existing multiple hypothesis control methods for interactive data exploration;
- Based on these observations, we develop new α-investing rules to control the marginalized false discovery rate (mFDR) that are designed specifically for interactive data exploration.
- We use Markov chain simulation and real-world datasets and workflows to show that our methods indeed achieve automatic control of false discovery and have significantly higher power than other techniques.

2. A MOTIVATIONAL EXAMPLE

To motivate the various aspects of multi-hypothesis control during data exploration we present a use case that is inspired by Vizdom [8]. Similar workflows however can be achieved with other systems like Tableau [17], imMens [26] or Dice [23].

Suppose Eve is a researcher at a non-profit organization and is working on a project relevant to a specific country. She obtained a new dataset containing census information and is interested in getting an overview of this data and extracting new insights.

She first considers the "*gender*" attribute and observes that the dataset contains the same number of records for men and women (Figure 1 A). She then moves on to a second visualization, displaying the distribution of people who earn above or below $50k a year. Eve links the two charts so that selections in the "*salary*" visualization filter the "*gender*" attribute. She notices that for salaries above $50k, the "*gender*" distribution is skewed towards men, and infers that men have higher salaries than women (B). After creating a third visualization for "*gender*" with, conversely, salaries lower than $50k (dashed line indicates inversion of selection), she confirms her finding "*Women are predominately earning less than $50k*" (C).

Eve now wants to understand what influences salaries and creates a chain of visualizations for people who have PhD degrees and are not married (D). Extending this chain using "*salary*" appears to suggest that this sub-population contains many high-earners (E). By selecting the high-earners and extending the chain with two "*age*" visualizations, she compares the age distribution of unmarried PhDs earning more than $50k to those making less. To verify that the observed visual difference is actually *statistically significant* she performs a t-test by dragging the two charts close to each other (F).

While the example contains only one hypothesis test *explicitly* initiated by the user, we argue that without accounting for other *implicit hypothesis tests* there is a significant increase of risk that the user may observe a false phenomenon during similar scenarios of data exploration. This opens up new important questions: why and when should visualizations be considered statistical hypothesis tests? How should these tests be formulated?

2.1 Hypothesis Testing

In this paper, we focus on the widely used frequentist approach. That is, to determine whether the relationship between two observations formalized as a "*research hypothesis*" or "*alternative hypothesis*' $\mathcal{H}$ is statistically relevant (i.e., not a product of data noise) we analyze its corresponding "*null hypothesis*" H which states no such relationship. The *testing procedure* will then calculate the p-value, which denotes the probability of observing an outcome at least as extreme as the one that was actually observed in the data, under the assumption that the null hypothesis H is true. If the p-valueassociated to the null hypothesis H is less than or equal to a priori chosen *significance level* α (commonly 0.05 or 0.01), the test suggests that the observed data is *inconsistent* with the null hypothesis which must thus be rejected. Respectively, if the p-valueis larger than the significance level, the null hypothesis H is accepted. This procedure guarantees for a single test, that the probability of a "*false discovery*" (also known as "*false positive*" or "*Type I error*") – wrongly rejecting the null hypothesis of no effect – is at most α. This does not imply that the alternative hypothesis is true; it just states that the observed data has the likelihood of $p \leq \alpha$ if the null hypothesis is true. In contrast, the *statistical power* is the probability that the test correctly rejects the null hypothesis H.

While the frequentist approach to hypothesis testing has been criticized [21, 28] and there has been work in developing alternative approaches, such as Bayesian tests [4], it is still widely used in practice and we consider it a good first choice to build a system which automatically controls the multiple hypothesis error as it has two advantages: (1) Novice users are more likely to have experience with standard hypothesis testing than the more demanding Bayesian testing paradigm. (2) The frequentist inference approach does not require to set a sometimes hard-to-determine *prior* as it is the case with Bayesian tests.

Figure 1: An example Interactive Data Exploration Session

2.2 Visualizations as Hypotheses

A visualization per-se shows a descriptive statistic (e.g., the count of women or the count of men) of the dataset and is not a hypothesis. It is reasonable to assume that in step A of Figure 1 the user just looks at the gender distribution and simply acknowledges that the census surveys roughly the same amount of women and men. However, it becomes a hypothesis if the user draws a conclusion/inference based on the information. For example, if the user assumed that there should be more men than women in the data and therefore considering the fact that there is an equal amount as an insight. The notion of a visualization being considered as a hypothesis becomes even clearer in step (B) and (C) of the example workflow. When looking at the visualization in (B) in isolation, it just depicts a descriptive statistic. But once the user makes any inference and/or bases further exploration on an insight extracted from this visualisation, then it should be considered an hypothesis. We believe that user inference in data exploration is ubiquitous and important. First, the human analytical reasoning and sense-making process is inherently non-linear [29, 33]. The future actions are influenced by new knowledge the user discovered in previous observations. Second, while susceptible to certain types of biases [10], the human visual system is highly optimized at registering differences in visual signals and detecting patterns [7]. An average user is very likely drawn to the changes between the gender distribution of step (A) and step (B) and might therefore infer that women earn less than men and potentially flag this as an interesting insight that deserves more investigation. This is illustrated in step (C) where the user now further drills down and visually compares the distribution of gender filtered by salary. We qualitatively confirmed this notion through a formative user study where we manually coded user-reported insights, following a think-aloud protocol similar to the one proposed in [16]. In this study we observed that users tend to pick up on even slight differences in visualizations and regard them as insights and users predominantly base future exploration paths on previously inferred insights.

We conclude two things: (1) most of the time users indeed treat visualizations as hypotheses, though there are exceptions, and (2) they often (wrongly) assume that what they see is statistical significant. The latter is particularly true if the users do not carefully check the axis on the actual count. For example, if a user starts to analyze the outliers of a billion record dataset and makes the conclusion that mainly uneducated whites are causing the outliers, the subset of the data she is referring to might be comparable small and the chance of randomness might be much higher. As part of visual data exploration tools, users often explore sub-populations, and while the original dataset might be large, the sub-population might be small. Thus, we argue that every visualization as part of a interactive data exploration tool should be treated as a hypothesis and that users should be informed about the significance of the insights they gain from the visualization. At the same time, a user should have the choice to declare a visualization as just descriptive.

2.3 Heuristics for Visualization Hypotheses

A core question remains: what should the hypothesis for a visualization be. Ideally, users would tell the system every single time what they are thinking so that the hypothesis is adjusted based on their assumed insight(s) they gain from the visualization. However, this is disruptive to any interactive data exploration session. We rather argue that the system should use a good default hypothesis, the user can modify (or even delete) if she so desires. For the purpose of this work, we mainly focus on histograms as shown in Figure 1 and acknowledge that there exist many other visualizations, which we consider as future work. We derived the following heuristics from two separate user studies where we observed over 50 participants using a IDE tool to explore various datasets.

1. *Every visualization without any filter conditions is not a hypothesis (e.g., step A in Figure 1) unless the user makes it one.* This is reasonable, as users usually first gain a general high-level impression of the data. Furthermore, in order to make it an hypothesis, the user would need to provide some prior knowledge/expectation, for example as discussed before, that he expected more men than women in the dataset.
2. *Every visualization with a filter condition is a hypothesis with the null hypothesis that the filter condition makes no difference compared to the distribution of the whole dataset.* For example, in step B of Figure 1 the null hypothesis for the distribution of men vs. women given the high salary class of over $\$50k$ would be that there is no difference compared to the equal distribution of men vs. women over the entire dataset (the visualization in step A). This is again a reasonable assumption as the distribution of an attribute given others is only interesting, if it shows some different effect compared to looking at the whole dataset.
3. *If two visualization with the same but some negated filter conditions are put next to each other, it is a test with the null hypothesis that there is no difference between the two visual-*

ized distributions, which supersedes the previous hypothesis. This is the case in step C: given that the user looks explicitly at the distribution of males vs females given a salary over and under $\$50k$ is a strong hint from the user, that he wants to compare these two distributions.

As with every heuristic it is important to note, that the heuristic can be wrong. Therefore it is extremely important to allow the user to overwrite the default hypothesis as well as delete default hypothesis if one really just acted as a descriptive statistic or was just generated as part to a bigger hypothesis test. Furthermore, there exist of course other potential null hypothesis. For example, in our workflow we assume by default that the user aims to compare distributions, which requires a χ^2-test. However, maybe in some scenarios comparing the means (i.e., a t-test) might be more appropriate as the default test. Yet, studying in detail what a good default null hypothesis is dependent on the data properties and domain, is beyond the scope of this paper.

2.4 Heuristics Applied to the Example

For our example in Figure 1 the resulting hypothesis could be as follows: Step A is not an hypothesis based on rule 1 as it just visualizes the distribution of a single attribute over the whole dataset. Step B is the hypothesis m_1 if the distribution of gender is different given a salary over $\$50k$. Step C supersedes the previous hypothesis and replaces it with an hypothesis m_1' if the gender distribution between a salary over and under $\$50k$ is different, which is a sightly different question. Step D creates a hypothesis m_2 if the marital status for people with PhDs is different compared to the entire dataset, whereas step-E generates a hypothesis m_3 if there is a different salary distribution given not married people with a PhD. By studying the age distribution in step F the system first generated a default hypothesis m_4 that the distribution of the ages is different given a PhD and being not married for different salary classes. However, the user overwrites immediately the default hypothesis with an hypothesis m_4' about the average age. Furthermore, as the previous visualizations in step D and E might just have been stepping stones towards creating m_4 the user might or might not delete hypothesis m_2 and m_3. However, if the insights our user gained from viewing the marital status, etc., influenced her to look at the age distribution, she might want to keep them as hypothesis.

While this is clearly a simple example, it succeeds in highlighting the general issue. Not every insight the user gains (e.g., the insight that women earn less) is explicitly expressed as a test. At the same time, the more the user explores the data the higher the chance that she finds something which looks interesting, but is actually just effect of noise in the data. In the example above, by the time the user actually performs its first test (step F), she implicitly already tested at least one other hypothesis and potentially even four others. Assuming a targeted p-valueof $\alpha = 0.05$, the chance of a false discovery therefore increased to $1-(1-\alpha)^2 = 0.098$ for two hypothesis and up to $1-(1-\alpha)^4 = 0.185$ for four hypothesis. While the question of what should count as an hypothesis is highly dependent on the user and can never be fully controlled by any system, we can however, enable the system to make good suggestions and help users to track the risk of making false discoveries by chance. Furthermore, this short workflow also demonstrates that hypotheses are built by adding but also by removing attributes. As we will discuss later, there exist no good method so far to control the risk of making false discoveries for incremental sessions like the ones created by interactive data exploration systems. We present new methods for interactive data exploration in Section 5.

Finally, it should be noted, that the same problems also exist with exploratory analysis using SQL or other tools. However, we argue that the situation is becoming worse by the up-rise of visual exploration tools, like Tableau, which allow to test more hypothesis in a shorter amount of time.

3. THE QUDE USER INTERFACE

As argued in the previous section, user feedback is essential in determining, tracking and controlling the right hypothesis during the data exploration process. With QUDE we created a system that applies our heuristic automatically to all visualizations. We designed QUDE 's user interface with a few goals in mind.

First, the user should be able to see the hypotheses the system assumed so far, their p-values, effect sizes and if they are considered significant and should be able to change, add or delete hypotheses at any given stage of the exploration.

Second, hypotheses rejection decisions should never change based on future user actions unless the user explicitly asks for it. We therefore require an incremental procedure to control the multiple hypothesis risk that does not change its rejection decisions even if more hypothesis tests are executed. For example, the system should not state that their is a significant age difference for not married highly educated people, and then later on revoke its assessment just because the user did more tests. More formally, if the system determined which hypotheses $m_1...m_n$ are significant (i.e., it rejects the null) or not and the user changes the last hypothesis or adds an hypothesis m_{n+1}, which should be the most common cases, the significance of hypotheses $m_1..m_n$ should not change. However, if the user might change, delete, or add hypothesis $k \in 1,..,n$, depending on the used procedure we might allow that the significance of hypotheses m_{k+1} to m_n might have to change as well.

Third, individual hypothesis descriptions should be augmented with information about how much data n^{H1} the user has to add, under the assumption that the new data will follow the current observed distribution of the data, to make an hypothesis significant. While sounding counter-intuitive, as one might (wrongly) imply, it is possible to make any hypothesis true by adding more data, calculating this value is in some fields already common practice. For example, in genetics scientist often search (automatically) for correlations between genes and high-level effects (like cancer). If such a correlation is found, often because of the multiple hypothesis error the chance of a true discovery is tiny (i.e., the p-valueis too high). In that case the scientist works backwards and estimates how much more genes she has to to sequence in order to make the hypothesis relevant, expecting that the new data (e.g., gene sequences) follow the same distribution of the data the scientist already has. However, if the effect was just produced by chance, the new data will be more similar to the distribution of the null hypothesis and the null will not be rejected. The required value is generally easy to calculate or approximate, and are highly valuable for the end-user. A small value for n^{H1} in relation to the number of totally tested hypotheses might be an indication that the power (i.e., the chance to accept a true alternative hypothesis) of the test was not sufficiently large.

Finally, users should be able to bookmark important hypotheses. As our system uses default hypotheses, there might be more hypotheses generated by the system than those that the user actually intends to test. It may be too cumbersome for the user to correct every time to convey his true intentions. Further, some hypotheses might be more important than others; (e.g. the hypotheses the user would like to include in a presentation or show to her boss).

A key question is what is the expected number of false discoveries among those important discoveries. Figure 2 shows the current interface design of QUDE with a risk controller, which incorporates the above ideas, running on a tablet. The user interface features

Figure 2: The QUDE User Interface

an unbounded 2D canvas where chains of visualizations (such as the one shown in Figure 1) can be laid out in a free form fashion. A "risk-gauge" on the right-hand side of the display (Figure 2 (A)) serves two purposes: it gives users a summary of the underlying procedure (e.g., the budget for the false discovery rate set to 5% with current remaining wealth of 2.5%; both explained in the next two sections) and it provides access to a scrollable list of all the hypothesis tests (implicit and explicit) that have been execute so far. Each list entry displays details about one test and its results. Textual labels describe the null and alternative hypothesis and color coded p-valuesindicate if the null hypothesis was rejected or accepted (green for rejected, red for accepted). Furthermore, it visualizes the distribution of null hypothesis and alternative hypothesis and shows its difference, included an indication of its color coded effect size (D). Tap gestures on a specific item allow users to change things like the default hypothesis or the type of test. Additionally other information such as an estimation of the size of an additional data n^{H1} that could make the observation significant can be displayed in each item. In the example this information is encoded through a set of small squares (B, C) where each square indicates the amount of data that is in the corresponding distribution. In (B) the five red squares tells us that we need 5x the amount of data from the distribution under null to flip this test form rejected to accepted or conversely in (C) 11.5x the amount of data from the alternative-distribution to rejected this hypothesis. Finally, we allow to mark important hypotheses by tapping the "star" icons (E).

4. BACKGROUND

The previous section described how we convey the multiple hypothesis error to the user and ask for user feedback to derive the right hypothesis to be considered. In this section we describe different techniques that allow to control the risk of incurring in false discoveries and we discuss they appropriateness for the IDE setting. The notation used through the paper is summarized in Appendix A.

We consider a setting, in which we evaluate the statistical relevance of hypotheses from a set $\mathcal{H} = \mathcal{H}_1, \mathcal{H}_2, \ldots, \mathcal{H}_m$, created incrementally by an IDE system in a streaming fashion. In order to verify whether any such hypothesis $\mathcal{H}_j$ corresponds to an actually statistically significant phenomenon, we consider its corresponding null hypothesis H_j. Using the appropriate statistical test (e.g., the *t-test* or the $\mathcal{X}^2$*-test*), we evaluate p-valueof H_j. Our testing procedure the uses this value to determinate whether to *accept* (resp., *reject*) a null hypothesis H_j which in turn corresponds to rejecting (resp., accepting) the corresponding *alternative hypothesis* (or *research hypothesis*) $\mathcal{H}_|$. The hypothesis according to which all null hypotheses are true is referred as the "*complete*" or "*global*" null hypothesis.

The set of null hypotheses rejected by a statistical test is called "*discoveries*"and is denoted as R. Among these we distinguish the set of *true discoveries* S, and the set of *false discoveries* or *false positives* V where $|V|+|S|= |R|$. False discoveries are commonly referred to as Type 1 errors. Null hypotheses corresponding to true discoveries are called *false null hypotheses*, whereas the others are referred as *true null hypotheses*.

4.1 Hold-Out Dataset

A plausible method to handle the multiple hypothesis error is to split the original dataset D into an exploration D_1 and a validation D_2 dataset [37]. D_1 is then used for the data exploration process, while the validation dataset is used to re-test all hypotheses in order to *validate* the results of the first phase. Next we provide examples to clarify why, albeit useful, the hold-out approach does not provide a solution for the multiple hypothesis testing problem.

Let us consider a null hypothesis H, and let p_D denote its associated p-valuewhen H is evaluated with respect of the entire dataset D. Lets assume we perform a test with significance-level α. In this case the probability of wrongly rejecting H is at most α Suppose now that we randomly split the dataset into two datasets D_1 and D_2. For the same null hypothesis H we evaluate the p-valuesp_{D_1} and p_{D_2} each obtained by evaluating H on D_1 or D_2 respectively. We then run a test with significance-level α (like the one discussed above) for each of the datasets. We then decide to reject H if it has been rejected by *both* the testing procedures operating on the datasets D_1 and D_2. Given that both the procedures operating on D_1 and D_2 have significance-level α, the probability that the overall procedure ends up rejecting H is at most α^2.

For the common value of $\alpha = 0.05$, the chance of a Type I error is thus reduced to $p_D = 0.0025$, which is good news. Rather than fully handling the multiple hypothesis problem, we have however only just lowered of the threshold for rejecting the null hypothesis (i.e., the significance level of the test).

This fact appears clearly in the following scenario. Suppose that the user wants to evaluate multiple hypotheses (e.g., 25) rather than just one. Assuming that these hypotheses, and their p-valuesare independent, the probability of observing at least one erroneous rejection using the test technique based on the use of the holdout dataset would be: $p_f = 1 - (1 - p_D)^{25} \approx 0.06$, which is higher than the desired α significance level. If the user would re-test 100 hypotheses on the validation dataset, p_f increases to ≈ 0.22. As it can be seen by the examples, the hold-out dataset helps to reduce the chance of a false discovery as it lowers the chance of a false positive, but does not control the multi-hypothesis error.

Unfortunately, this technique also significantly reduces the power of the testing procedure. Consider the following example scenario in which we aim to compare the means M_1 and M_2 of two samples one drawn from a population with expected value $\mu_1 = 0$ and the other from a population with $\mu_2 = 1$, both having standard deviation $\sigma = 4$. In order to determinate weather the observed difference between M_1 an M_2 is actually statistically significant, we test the null hypothesis "*there is no significant difference between* μ_1 *and* μ_2" using the *one-sided t-test* and a sample composed by 500 records from each population. Given the properties of the t-test (see [13]), the statistical power of our test would be 0.99, and the probability of erroneously accepting the null hypothesis would be at most 0.01.

Suppose now that we divide the original dataset into a training dataset for exploration and one for validation each composed by 250 records. The statistical power for each of the individual t-test

executed on the two dataset is now lowered to 0.87, due to the reduction of the amount of data being used in the individual tests. Further, recall that the procedure based on the holdout set rejects a null hypothesis only if said hypothesis is rejected by both sub-tests. This implies that the actual *overall* power of the testing procedure is $0.87 \cdot 0.87 \approx 0.76$, which is significantly lower than the 0.99 achieved by the test which uses the entire data.

In general, approaches based on hold-out datasets are considered inferior compared to testing over the entire dataset. In some scenarios, such as building machine learning models, hold-out datasets might even be the only possibility to test a model or tune parameters. In those cases, a hold-out approach (e.g., "*k-fold cross-validation*") should be considered as test on its own and, as recent work suggests [9, 24, 30], should be controlled for the multiple hypothesis error. It is however important to remark that in our work we aim to predict guarantees on the statistical significance of the statistical predictors which are instead not achievable using prediction-driven approaches such as cross-validation.

4.2 Family-Wise Error Rate (FWER)

Traditionally, frequentist methods for multiple comparisons testing focus on correcting for modest numbers of comparisons. A natural generalization of the significance level to multiple hypothesis testing is given by the *Family Wise Error Rate* (FWER). Given a family of hypotheses, the FWER denotes the is the probability of making at least one type I error when rejecting the corresponding null hypotheses:

$$FWER = \Pr(V \geq 1) = 1 - \Pr(V = 0) \tag{1}$$

If $FWER \leq \alpha$, that is the FWER is *controlled at level* α, we have that the probability of even one Type I error in evaluating a family of hypotheses is at most α.

We say that a procedure controls the FWER *in the weak sense*, if the FWER control at level α is guaranteed only when *all* null hypotheses are true (i.e. when the complete null hypothesis is true). We say that a procedure controls the FWER *in the strong sense*, if the FWER control at level α is guaranteed for any configuration of true and non-true null hypotheses (including the global null hypothesis).

Bonferroni Correction: The Bonferroni correction is simple test procedure that allows to control FWER over multiple hypotheses [6]. Let α be the critical threshold for the test. The value of α is usually selected at 0.01 or 0.05.

Let p_i the p-valuestatistic associated with the null hypothesis H_i. When testing m distinct null hypotheses using the Bonferroni correction, a null hypothesis H_i is rejected if $p_i \leq \alpha/m$. The Bonferroni procedure thus achieves control of the FWER at level α.

Using the Bonferroni correction requires knowledge of the total number of hypotheses being evaluated. This constraint make it not applicable in our IDE setting where the set of hypotheses is incrementally generated by the user over multiple data exploration steps. A possible alternative approach would be to use a variation of the Bonferroni correction, according to which the j-th null hypothesis H_j is rejected if $p_j \leq \alpha \cdot 2^{-j}$. It is possible to show that this procedure indeed controls FWER at level α as $j \to \infty$ while not requiring knowledge of m. The crucial drawback of this approach is however given by the fact that the acceptance threshold decreases exponentially with respect to the number of hypotheses, thus resulting in a high number of false negatives.

The main common issue with all FWER techniques is that the power of the test significantly decreases as m increases due to the corresponding decrease in the acceptance threshold (α/m in the original Bonferroni or $\alpha/2^i$ in the sequential variant). While some alternative testing procedures such as those of Vĭdák [34], Holm [19], Hochberg [18], and Simes [35] offer more power while controlling FWER, the achieved improvements are generally minor (see [32] for a review of several of these techniques).

4.3 False Discovery Rate (FDR)

In [2] Benjamini and Hochberg proposed the notion of *False Discovery Rate (FDR)* as a less conservative approach to control errors in multiple hypotheses tests which achieves a substantial increase in the power of the testing procedure.

FDR-controlling procedures are designed to control the expected ratio of false discoveries among all discoveries returned by a procedure. In particular, the FDR of a statistical procedure is defined as:

$$FDR = E\left[Q\right] = E\left[\frac{V}{R}|R > 0\right] P(R > 0). \tag{2}$$

If we define FDR to be zero when $R = 0$, we can simplify 2 to:

$$FDR = E\left[\frac{V}{R}\right] \tag{3}$$

We say that a testing procedure controls FDR at level α if we have $FDR \leq \alpha$. Designing a statistical test that controls for FDR is not simple, as the FDR is a function of two random variables that depend both on the set of null hypotheses and the set of alternative hypotheses. The standard technique to control the FDR is the *Benjamini-Hochberg procedure*(BH), which operates as follows: let $p_1 \leq p_2 \leq \ldots \leq p_m$ be the sorted order of the the p-valuesfor the m tested null hypotheses. To control FDR at level α (for independent null p-values) determine the maximum k for which $p_k \leq \frac{k}{m} \cdot \alpha$, and reject the null hypotheses corresponding to the p-values$p_1, p_2, \ldots, p_k$.

Interestingly, under the complete null hypothesis, controlling the FDR at level α guarantees also "*weak control*" over the FWER: $FWER = P\left(V \geq 1\right) = E\left(\frac{V}{R}\right) = \text{FDR} \leq \alpha$. This follows from the fact that the event of rejecting at least one true null hypothesis $V \geq 1$ is exactly the event $V/R = 1$, and the event $V = 0$ is exactly the event $V/R = 0$ (recall $V/R = 0$ when $V = R = 0$). The concept of FDR control is thus relatively easy to convey to the user as under complete random data, the chance of one or more false discoveries is at most α as in FWER. However, FDR does not ensure control of the FWER if there are some true discoveries to be made (i.e., it does not ensure "*strong control*" of the FWER). In Appendix B we provide experimental comparison between FDR and FWER.

Because of its increased power, FDR appears to be a better candidate than FWER in the context interactive data exploration, where usually a larger number of hypotheses are to be considered. Unfortunately, both the original *Benjamini-Hochberg procedure* and its variation for dealing with dependent hypotheses [3] are not incremental as they require knowledge of the total number of hypotheses being tested (similar to what was discussed for the Bonferroni correction) and of the sorted list of all the p-valuescorresponding to each null hypothesis being evaluated.

An adaptation of the FDR technique to a setting for which an unspecified number of null hypotheses are observed incrementally was recently presented in [15]. The main idea behind the Sequential FDR procedure is to convert the arbitrary sequence of p-valuescorresponding to the null hypotheses observed on the stream into an sequence of increasing p-values akin to the one generated by the classical Benjamini-Hochberg procedure. The natural application for this technique is the progressive refinement of a model by considering additional features. That is, it starts constructing a model for the data with something known and general. The user then proceeds to refine the model by determining the most significant features.

One drawback of the Sequential FDR method, is given by the fact that the order according to which the hypotheses are observed

on the stream heavily influences the outcome of the procedure. For example, if an hypothesis with high p-valueis observed among the first in the stream, this will harm the ability of the procedure of rejecting following null hypotheses, even if they have low p-value(see discussion in [15]). This aspect makes Sequential FDR not applicable for data exploration system for which the user is likely to explore different "*avenues*" of discovery rather than focusing on the specialization of a model.

4.4 Other Approaches

Although for most practical applications, FDR controlling procedures constitute the *standard de facto* for multiple hypothesis testing [12], many other techniques have been presented in the literature. Among them, Bayesian techniques are particularly noteworthy. Alternative solutions based on decision theory using "*Bayesian FDR*" are discussed in [4]. However, as often the case with Bayesian approaches, the computational cost for these procedures when applied to large datasets are significant, and the results are highly dependent on the prior model assumptions.

Another approach is correcting for the multiplicity through simulations (e.g., the *permutation test* [31]) in order to experimentally evaluate the probability of an observation in the null distribution. This approach is also not practical in large datasets because of the large number of different possible observations and the need to evaluate very small p-values for each of these distributions [22].

In this paper, we elect to use a family of multiple hypothesis testing procedures know as α-investing introduced in [14] and then generalized in [1]. These procedures are especially interesting for the incremental and interactive nature of interactive data exploration. The details of α-investing and its application to our setting is extensively discussed in the next section.

5. INTERACTIVE CONTROL

One drawback of the Sequential FDR procedure [15] and any adaptation of the FWER controlling techniques to the streaming setting lies in the fact that decisions regarding the rejection or acceptance of previously considered null hypotheses could potentially be overturned in latter stages due to new hypotheses being considered. Although statistically sound, this fact could appear counter intuitive and confusing to the user. The only way to adopt the Sequential FDR procedure to data exploration would be to batch all the hypotheses and only present the final decisions after the exploration halts. Thus although it is incremental, Sequential FDR is *not* suited for interactive data exploration.

In order to have both incremental and interactive multiple hypothesis control, we consider a different approach for multiple hypothesis testing based on the "α-investing " testing procedure originally introduced by Foster and Stine in [14]. Similarly to Sequential-FDR , this procedure does not require explicit knowledge of the total number of hypotheses being tested and can therefore be applied in the hypothesis streaming setting. α-investing presents however several crucial differences with respect to both traditional and sequential FDR control procedures.

In the following, we first introduce the general outline of the procedure as presented in [14], then we discuss several strategies (called policies) that we have developed for interactive data exploration.

5.1 Outline of the Procedure

In α-investing , the quantity being controlled is not the classic FDR but rather an alternative quantity called "*marginal FDR*" (mFDR):

$$mFDR_\eta(j) = \frac{E\left[V(j)\right]}{E\left[R(j)\right] + \eta} \tag{4}$$

where j denotes the total number of tests which have been executed, while $V(j)$ (resp., $R(j)$) denote the number of false (resp., total) discoveries after j tests using the α-investing procedure.

We say that a testing procedure controls $mFDR_\eta$ at level α if $mFDR_\eta(j) \leq \alpha$. The parameter η is introduced in order to weight the impact of cases for which the number of discoveries is limited. Common choices for η are $1, (1-\alpha)$, whereas the procedure appears to lose in power for values of η close to 0 [14].

Under the complete null hypothesis we have $V(j) = R(j)$ hence $mFDR_\eta(j) \leq \alpha$ implies that $E\left[V(j)\right] \leq \alpha\eta/(1-\alpha)$. If we chose $\eta = 1-\alpha$ then $E\left[V(j)\right] \leq \alpha$, and we can thus conclude that control of the $mFDR_{1-\alpha}$ at level α implies weak control fo the FWER at level α [14]. We refer the reader to the original paper of Foster and Stine [14] for an extensive discussion on the relationship between $mFDR$ and the classic FDR. A generalization of the α-investing procedure was later introduced in [1]. The α-investing procedure does not in general require any assumption regarding the independence of the hypotheses being tested, although opportune corrections are necessary in order to deal with possible dependencies. In our analysis, we however assume that all the hypotheses and the corresponding p-valuesare indeed independent (see also Section 7).

Intuitively the α-investing procedure works by assigning to each hypothesis test a budget α' from an initial "*α-wealth*." If the p-valueof the null hypothesis being considered is above α' the null hypothesis is accepted and some budget is lost, otherwise it is rejected and some exploration budget is gained. More formally, we denote as $W(0)$ the initial α-wealth assigned to the testing procedure. If the goal of the testing procedure is to control $mFDR_\eta$ at level α, then we shall set $W(0) = \alpha \cdot \eta$. Here, η is commonly set to $(1-\alpha)$. We denote as $W(j)$ the amount of "*available α-wealth*" after j tests have being executed.

Each time a null hypothesis H_j is being tested, it is assigned a budget $\alpha_j > 0$. Let p_j denote the p-valueassociated with the null hypothesis H_j. This hypothesis is rejected if $p_j \leq \alpha_j$. If H_j is rejected than the testing procedure obtains a "*return*" on its investment $\omega \leq \alpha$. Instead, if the null hypothesis H_j is accepted, $\alpha_j/(1-\alpha_j)$ alpha wealth is deducted from the available α-wealth:

$$W(t) - W(t-1) = \begin{cases} \omega & \text{if } p_j \leq \alpha_j, \\ -\frac{\alpha_j}{1-\alpha_j} & \text{if } p_j > \alpha_j \end{cases} \tag{5}$$

The testing procedure halts when the available α-wealth reaches 0. At that point in time, the user should stop exploring to guarantee that $mFDR \leq \alpha$. This is obviously something not desirable from the perspective of the user. We discuss this problem and potential solutions in Section 5.8.

The budget α_j which can be assigned to a test must be such that regardless of the outcome of the test, the available α-wealth available after the test is not negative $W(j) \geq 0$, hence $\alpha_j \leq W(j-1)/(1-W(j-1))$. Further we impose that $\alpha_j < 1$. While this constraint was not explicated in [14], it is indeed necessary for the correct functioning of the procedure. Setting $\alpha_j = 1$ would lead to the potential deduction of an infinite amount of α-wealth, violating the non negativity of $W(j)$. Setting $\alpha_j > 1$ would instead lead to having a positive increase of the available α-wealth *regardless* of the outcome of the test. In our analysis we will however assume that all the hypotheses being considered are indeed independent and their associated p-valuesare independent as well.

We refer as "*α-investing rule*" to the policy according to which available budget is assigned to the hypotheses that needs to be tested. Furthermore, in [14] it was shown that any α-investing policy for which $W(0) = \eta \cdot \alpha$, $\omega = \alpha$, and which obeys the rule in (5), controls the $mFDR$ at level α, for $\alpha, \eta \in [0, 1]$. The freedom of assigning to each hypothesis a specific level of significance independent of the order, and the possibility of "*re-investing*" the wealth

obtained by previous rejections constitute great advantages with respect to the Sequential FDR procedure.

5.2 α-Investing for Data Exploration

While it is relatively straightforward to devise investing rules, it is difficult determinate a priori the "*best way to invest*" the available $alpha$-wealth. If α_j budged assigned to the hypothesis is too low, the statistical power of every test is greatly reduced and there is a high chance of losing the invested budget even if the null hypothesis being considered is indeed false. On the other hand, if the α_j assigned is too high, the entire α wealth might be quickly exhausted and the testing procedure has to halt. A policy is most likely to be successful if it can exploit some knowledge of the actual data distribution and the data exploration setting.

Another complication involves the statistical analysis of the individual tests that provide the p-values. To show that a testing procedure controls $mFDR$, we require that conditionally on the prior j - 1 outcomes (denoted as R_i), the level of the test of H_j must not exceed α_j:

$$P(R_j = 1|R_{j1}, R_{j2}, ..., R_1) \leq \alpha_j. \tag{6}$$

Note that the tests do not need to be independent, though independence is the easiest setting that satisfies (6). Furthermore, the condition is only on the outcome of the previous tests, not on the values of their statistic (see discussion in Section 7).

In [14] a specialized strategy, *Best Foot Forward*, is designed for a specific situation where true discoveries are highly clustered. However for interactive data exploration we need to consider wider situations where the structure of true discoveries may not always be clustered but have other properties. We further model these general exploration settings in Section 8.

In the remainder of this section we propose different α-investing policies particularly suited for interactive data exploration. Each policy captures a different exploration strategy and aims to exploit different possible properties of the data and the exploration settings.

All our policies assign to each hypothesis a strictly positive budget $\alpha_j > 0$ as long as any α-wealth is available. If $p_j \leq \alpha_j$, the null hypothesis H_j is rejected (i.e., it is considered a discovery). Vice versa, if $p_j > \alpha_j$ is *accepted*. The current α-wealth $W(j)$ is then updated according to the rule in (5) and because of it controls $mFDR$ at level α as shown in [14].

5.3 β-Farsighted Investing Rule

Like with real investment, the question is if one should invest short or long-term. β-farsighted policies are aimed at preserving wealth wealth over long exploration sessions, while effectively using the available α-wealth for the evaluation of new hypotheses. Given $\beta \in [0, 1)$, we say that a policy is β-farsighted if it ensures that regardless of the outcome of the j-th test at least a fraction β of the current α-wealth $W(j-1)$ is preserved for future tests, that is for $j = 1, 2, \ldots$:

$$\begin{aligned} W(j) &\geq \beta W(j-1), \\ W(j) - W(j-1) &\geq (\beta - 1)W(j-1) \end{aligned} \tag{7}$$

An example of β-farsighted is given in Investing Rule 1. Indeed, this procedure achieves control of the $mFDR_\eta$ at level α.

Different choices for the parameter $\beta \in [0, 1)$ characterize how conservative the investing policy is. If there is high confidence on the first observed hypotheses being true discoveries, small values of beta (i.e., 0.25) would be more effective. Vice versa, high values of β (i.e. 0.9) ensure that even if the first hypotheses are true null, a large part of the α-wealth is preserved.

We say that an α investing policy is "*thrifty*" if it never fully commits its available α-wealth. The described β-farsighted is indeed

Investing Rule 1 β-farsighted

1: $W(0) = \eta\alpha$
2: **for** $j = 1, 2, \ldots$ **do**
3: $\quad \alpha_j = \min\left(\alpha, \frac{W(j-1)(1-\beta)}{1+W(j-1)(1-\beta)}\right)$
4: $\quad$ **if** $p(H_j) < \alpha_j$ **then**
5: $\quad\quad W(j) = W(j-1) + \omega$
6: $\quad$ **else**
7: $\quad\quad W(j) = W(j-1) - \frac{\alpha_j}{1-\alpha_j} = \beta W(j-1)$
8: $\quad$ **end if**
9: **end for**

thrifty. While the procedure will never halt due to the available α-wealth reaching zero, after a long series of acceptance of null hypotheses the available budget may be reduced so much that it will be effectively impossible to reject any more null hypotheses.

Although these policies may appear wasteful as there is no reward for wealth which has not been invested, they are aimed to preserve some of their current budget for future tests in case the hypotheses considered in the beginning of the testing procedure are not particularly trustworthy.

This investing rule is therefore particular suited for scenarios were the total number of false discoveries in long exploration sessions, potentially across multiple users, should be controlled.

While β-farsighted, is a generalization of the "*Best-foot-forward policy*" in [14], different values of β allow for higher flexibility.

5.4 γ-Fixed Investing Rule

A different *non-thrifty* procedure assigns to each hypothesis the same budget α^*. We call γ-*fixed* a procedure that assigns to each null hypothesis a fixed budget α_j equal to a fraction of the initial α-wealth $W(0)$, that is $\alpha^* = W(0)/(W(0) + \gamma)$, as long as any α-wealth is available.

The details of the γ-fixed procedure controlling $mFDR_\eta$ at level α can be found in the procedure for Investing Rule 2. Note that we

Investing Rule 2 γ-fixed

1: $W(0) = \eta\alpha$
2: $\alpha^* = \frac{W(0)}{\gamma + W(0)}$
3: **while** $W(j-1) - \frac{\alpha^*}{1-\alpha^*} \geq 0$, for $j = 1, 2, \ldots$ **do**
4: $\quad$ **if** $p(H_j) < \alpha^*$ **then**
5: $\quad\quad W(j) = W(j-1) + \omega$
6: $\quad$ **else**
7: $\quad\quad W(j) = W(j-1) - \frac{\alpha^*}{1-\alpha^*} = W(j-1) - \frac{W(0)}{\gamma}$
8: $\quad$ **end if**
9: **end while**

define α^* as $W(0)/(\gamma + W(0))$ to ensure that the subtraction of the wealth is constantly $W(0)/\gamma$. Different choices for the parameter γ characterize how conservative the investing policy is. If there is high confidence on the first observed hypotheses being actual discoveries small values of γ (i.e. 5,10,20) would make more sense. Vice versa a high value of γ, e.g., 50, 100, ensures that even if the first hypotheses are true null, a large part of the α wealth is preserved.

5.5 δ-Hopeful Investing Rule

In a slight variation of the γ-fixed investing rule, we say that a policy is δ-hopeful if the budget is assigned to each hypothesis "*hoping*" that at least one of the next δ hypotheses will be rejected. Each time a null hypothesis is rejected the budget obtained from the rejection is re-invested when assigning budget over the next δ null hypotheses. γ-fixed and δ-hopeful operate by spreading the amount of α-wealth over a fixed number of hypotheses (either γ or δ), δ-hopeful is however "*less conservative*" than γ-fixed as

it always operates by investing *all* currently available α-wealth over the next δ hypotheses. The details of the δ-fixed procedure controlling $mFDR_\eta$ at level α can be found in the procedure for Investing Rule 3.

Investing Rule 3 δ-hopeful

$W(0) = \eta\alpha$
$\alpha^* = \frac{W(0)}{\delta + W(0)}$
$k^* = 0$
while $W(j-1) - \frac{\alpha^*}{1-\alpha^*} \geq 0$, for $j = 1, 2, \ldots$ **do**
  if $p(H_j) < \alpha^*$ **then**
    $W(j) = W(j-1) + \omega$
    $\alpha^* = \min\left(\alpha, \frac{W(j)}{\delta + W(j)}\right)$
    $k^* = j$
  else
    $W(j) = W(j-1) - \frac{\alpha^*}{1-\alpha^*} = W(j-1) - \frac{W(k^*)}{\alpha^*}$
  end if
end while

5.6 ϵ-Hybrid Investing Rule

Because α-investing allows contextual information to be incorporated, the power of the resulting procedure is related to how well the design heuristic fits the actual data exploration scenario. For example, when the data exhibits more randomness, the γ-fixed rule tends to have more power than the δ-hopeful rule. Intuitively, the α-wealth decreases when testing a true null hypothesis, because the expectation of the change of wealth is negative when the p-value is uniformly distributed on $[0, 1]$. Thus the initial α-wealth is on average larger than the α-wealth available at subsequent steps. Furthermore, since the γ-fixed rule invests a constant fraction of the initial wealth, the power tends to be larger than δ-hopeful. Vice versa, when the data is less random, we expect the power of the γ-fixed rule to become lower than that of the δ-hopeful rule. This is due to the fact that in this setting more significant discoveries tend to keep the subsequent α-wealth high, potentially even higher than the initial wealth. We further study this difference in Section 8.

Investing Rule 4 ϵ-hybrid

$W(0) = \eta\alpha$
$k^* = 0$
$H_d = []$ // Sliding window of size d
while $W(j-1) > 0$, for $j = 1, 2, \ldots$ **do**
  if Rejected$(H_d) \leq |H_d|\epsilon$ **then**
    $\alpha_j = \frac{W(0)}{\gamma + W(0)}$
  else
    $\alpha_j = \min\left(\alpha, \frac{W(k^*)}{\delta + W(k^*)}\right)$
  end if
  if $W(j-1) - \frac{\alpha_j}{1-\alpha_j} \geq 0$ **then**
    if $p(H_j) < \alpha_j$ **then**
      $W(j) = W(j-1) + \omega$
      $k* = j$
      $H_d[j] = R_j = 1$
    else
      $W(j) = W(j-1) - \frac{\alpha_j}{1-\alpha_j}$
      $H_d[j] = R_j = 0$
    end if
  end if
end while

In order to have a robust performance in terms of power and false discovery rate, we design ϵ-hybrid investing rule that adjust the α_j assigned to the various tests based on the estimated data randomness. Our estimation of the randomness of the data is based on the ratio of rejected null hypotheses over a sliding window H_d constituted by the last d null hypotheses observed on a stream. We then compare this ratio with a "*randomness threshold*" $\epsilon \in (0, 1)$ and we conclude whether the data exhibits high randomness or not. The procedures is outlined in Investing Rule 4.

5.7 Investment based on Support Population

An important intuition relative to the computation of the p-valuesis that is most likely to observe high p-valuesfor hypotheses which rely on a small number of data points, that is for hypothesis with low "*support population*." It is thus only natural to pursue a strategy which adjusts the budget of each hypothesis based on its support population, assigning lower α-wealth budget to hypotheses which are more likely to exhibit higher p-value. In this section we discuss how to *bias* the amount budget assigned to each hypothesis so that hypotheses with more support data receive more "*trust*" (in terms of budget) from the procedure.

Let us denote as $|n|$ the total amount of data being used and by $|j|$ the available data for testing the j-th null hypothesis H_t. A simple way of correcting the assignment of the budget α_j in any of the previously mentioned hypothesis is to assign to the test of the hypothesis $\alpha_j f(\frac{|j|}{|n|})$. Depending on the choice of $f(\cdot)$ the impact of the correction may be more or less severe. Some possible choices for $f(\cdot)$ would be $f(\frac{|t|}{|n|}) = \left(\frac{|t|}{|n|}\right)^\psi$ for possible values of $\psi = 1, 2/3, 1/2, 1/3, \ldots$. Note, that this support-size dependent *bias* idea can be used with any of the previously described strategies. In the following, we show the ψ-support policy applied to the γ-fixed rule in Investing Rule 5.

Investing Rule 5 ψ-support

$W(0) = \eta\alpha$
α^* is set by an α-investing rule
while $W(j-1) > 0$, for $j = 1, 2, \ldots$ **do**
  $\alpha_j = \alpha^* \left(\frac{|t|}{|n|}\right)^{\frac{1}{2}}$
  if $W(j-1) - \frac{\alpha_j}{1-\alpha_j} \geq 0$ **then**
    if $p(H_j) < \alpha_j$ **then**
      $W(j) = W(j-1) + \omega$
    else
      $W(j) = W(j-1) - \frac{\alpha_j}{1-\alpha_j}$
    end if
  end if
end while

5.8 What Happens If the Wealth is 0?

Among all our proposed investing policies, only β-farsighted is "*thrifty*" in that it never fully commits all of the available α-wealth. Still, the available wealth for β-farsighted could eventually become extremely small, to the point that hypotheses can be harder to reject. All the remaining procedures are "*non-thrifty*" and can thus reach zero α-wealth, in which case the user should stop exploring because no more hypotheses can be rejected.

Theoretically, a vanishing α-wealth indicates higher uncertainty from the current data and hypotheses, and thus it is reasonable to restrain further exploration. On the other hand, it is only natural to wonder if it would be possible for the user to "*recover*" some of the lost α-wealth and thus continuing the testing procedure. One possible way would require the user to reconsider and possibly overturn some of the previous decisions on whether to reject or accept some null hypotheses using alternative testing procedures (i.e., the Benjamini-Hochberg procedure). There are however several challenges to be faced when pursuing this strategy: 1) great care has to be put on haw to combine results from different testing procedures (i.e., control of FDR for a subsets of hypotheses and control of *mFDR* for a distinct subset of hypotheses) and 2) testing hypotheses for a second time given the outcomes of other test implies a clear (and strong) dependence between the outcome of the tests and the p-valueassociated with the null hypotheses being considered. Therefore, depending on the context such control could only be

achieved given additional assumptions about the level of control or would require adding additional data or the use of a hold-out dataset. We aim to study this problem in detail as part of future work.

6. MOST IMPORTANT DISCOVERIES

In Section 3 we argued that the user should be able to "*mark* the important hypotheses (e.g., the ones she wants to include in a publication). This is particularly important as QUDE uses default hypotheses, which the user might consider as less important. In the following we show that if these "*important discoveries*" are selected from all the discoveries given by a testing procedure that controls FDR at level α independently of their p-values, then the FDR for the set of important discoveries is controlled at level α as well.

THEOREM 1. *Assume that we executed a collection of hypothesis tests with a rejection rule that controls the FDR at α. Assume that the procedure rejected a nonempty set of null hypotheses R, and let $V \subseteq R$ be the set of false discoveries. If the hypothesis tests are independent then for any subset $R' \subseteq R$ we have $E[\|V \cap R'\|/\|R'\|] \leq \alpha$.*

PROOF. Let $p_1, \ldots, p_{\|R\|}$ be the p-valuesof the rejected hypotheses. Since the rejection rule controls the FDR at α we have

$$\sum_{i=1}^{\|R\|} \frac{i}{\|R\|} P(\|V\| = i \mid P_1 = p_1, \ldots, P_r = p_r) \leq \alpha \tag{8}$$

Assume that $\|V\| = i$. The p-valuesof true null hypotheses in V are i.i.d. uniformly distributed in $[0, 1]$. The set V is uniformly distributed among all the i-subsets of R. Let $p_i^V, \ldots p_{\|V\|}^V$ be the p-valuesof V, and let $p'_1, \ldots, p'_{\|R'\|}$ be the p-valuesof a subset of rejected hypotheses $R' \subseteq R$, then:

$$E[\|V \cap R'\| \mid \|V\| = i] = E[\|\{p'_1, \ldots, p'_{\|R'\|}\} \cap \{p_1^V \ldots P_{\|V\|}^V\}\| \mid \|V\| = i] = i\frac{\|R'\|}{\|R\|}. \tag{9}$$

Combining equations (8) and (9) we get:

$$E\left[\frac{\|V \cap R'\|}{\|R'\|}\right] = \sum_{i=1}^{\|R\|} E\left[\frac{\|V \cap R'\|}{\|R'\|} \,\middle|\, \|V\| = i\right] P(\|V\| = i \mid P_1 = p_1, \ldots, P_r = p_r)$$

$$= \sum_{i=1}^{\|R\|} \frac{1}{\|R'\|} i \frac{\|R'\|}{\|R\|} P(\|V\| = i \mid P_1 = p_1, \ldots, P_r = p_r) \leq \alpha \tag{10}$$

□

Consider a set R' of important discoveries selected independently of the p-valuesof the corresponding tests from a larger set of discoveries R for which then $mFDR$ is controlled at level α. Using a proof similar to the one discussed in Theorem 1 it is possible to show that the $mFDR$ of R' is controlled at level α as well. This is an important result, as it implies that the user can select the important discoveries from a larger pool of discoveries while maintaining the control of FDR (or mFDR) at level α.

7. LIMITATIONS AND OPPORTUNITIES

QUDE is the first system that automatically controls the multiple hypothesis error during visual data exploration and as such, it is important to understand the assumptions and limitations of our current approach and the opportunities for future work.

Visualizations: Currently we only automatically derive a null-hypothesis for histograms. While we believe that many other visualizations (e.g., line charts, heatmaps, etc.) have natural null-hypothesis associated to them, exploring them remains future work.

Sequential Dependencies: As formulated in Section 5.2, our α-investing procedures assume that the p-value of each test is computed in the sample space conditioned on the outcome of previous tests. How restrictive is this assumption? Significant part of any exploration process is done in a sequential process, in which features, or variables, are selected one at a time. Once we selected the first $k-1$ features, we test for the significance of adding the k feature to the current model. This process, which corresponds to testing nested hypothesis, trivially satisfies the condition formulated by (6). When not testing for nested hypothesis we need to be more careful with sequential dependency. Ideally, we want to test mutually independent features, or correct for the dependencies, but this may not be feasible. In practice, features have to be highly correlated to significantly distort the outcome of the process, and we can identify highly correlated features by computing the correlation coefficient in the data (with no testing). While modern statistics does not provide a fully analytical solution for the problem, our experiments show that independence assumption is a reasonable approximation for non-adversarial users and provides a best-effort attempt considering that often the only alternative is to leave the user in the dark.

Is it Possible to "Game" the System? For example, could a user boost her α-wealth for risky test by testing trivial hypothesis first ? While the short answer is "yes", it is not really gaming the system. mFDR controls the ratio of false over all discoveries and adding more trivial true hypotheses simply increases the denominator.

8. EXPERIMENTAL EVALUATION

In this section, we evaluate QUDE in different data exploration settings to answer the following questions: (1) how the different α-investing rules perform in different exploration scenarios, (2) how different parameters change the performance of the rules, and (3) how to select the parameters.

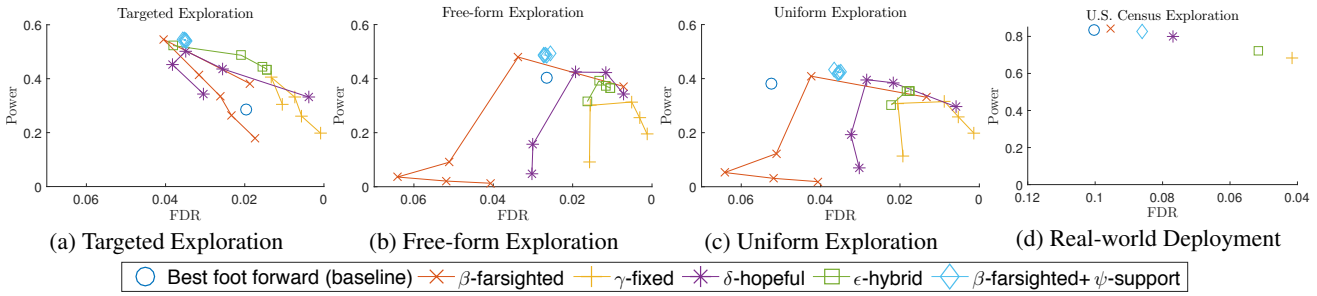
8.1 Exploration Settings

The data exploration process can vary significantly depending on how the hypotheses are structured. For example, in some settings the explorer may not start with any particular set of questions as target, but gradually develops interests in certain aspect of the data. Such settings are predominant in interactive data exploration. On the other hand, the exploration may be structured around a clear subject, such as understanding gravity, and the exploration tends to progress from easier questions towards harder questions. Such cases arise frequently in data-driven scientific studies. In other cases the mining of insights might usually less structured.

To account for varying structures of the data exploration, we formulate three different data exploration scenarios, namely, the *targeted exploration*, the *free-form exploration*, and the *uniform exploration*. We use Markov processes to simulate the data exploration process under these models as a stream of hypotheses, and identify the suitability of different α-investing rules.

We model the data exploration as divided into two phases of different distributions of random noise. Concretely, we construct two two-state Markov chains, X_a and X_b, where the states correspond to ground truth labels, namely true or false null hypotheses. The two Markov chains have stationary distributions π_a and π_b. We start the process at one chain and then switch to the other chain at a preset time. If the Markov chain generates a significant hypothesis, it draws data points from two normal random variables with different means (the difference is sampled from the intervals with boundaries 5/4, 5/2, 15/4, 5), whereas for an insignificant hypothesis (i.e. true null), it samples from two zero-mean random variables. The hypothesis tests are all t-tests. Afterwards, another uniform random variable is used to determine the number of records used per test, which simulates the selectivity of a histogram filter chain such as in Figure 1.

Finally, to demonstrate real-world applicability, we deploy the



(a) Targeted Exploration (b) Free-form Exploration (c) Uniform Exploration (d) Real-world Deployment

Figure 3: The Power-FDR curves with varying parametrization (Table 1) in different data exploration scenarios.

α-investing rules with the parameters obtained from the models to U.S. Census dataset [25] using an authentic user workflow [36].

8.2 Targeted Exploration

The user in the target exploration may have a well-defined set of questions to ask, and thus explores the data in a more structured manner. She starts with simpler questions such as "is there a salary difference between men and women," perhaps for confirmatory purposes. As data conforms to her initial questions, she dives deeper into the more complicated questions such as "does the state impact the salary between men and women." Note that the structure of the hypotheses progresses in such a way that the significant insights tend to cluster more from the beginning than towards the end. Thus in the corresponding Markov process, we set the first phase of exploration with lower distribution of random noise than the second phase. Concretely, the stationary distributions of the true nulls are set to $\pi_a(0) = 0.25$ and $\pi_b(0) = 0.75$. The switching time is set to 30% of the process.

Figure 3 shows the performance power vs. *realized* FDR of the different α-investing rules in this setting. Notice, that we plot FDR in decreasing order. Thus, in general the top-right corner means better: more true discoveries (i.e., power) with fewer false discoveries (i.e., realized FDR). For each rule we vary the parameters according to Table 1 from more aggressive to more conservative settings, resulting in the Power-FDR curve. For the ϵ-hybrid and ψ-support strategies we keep $\delta = \gamma = 64$ and $\beta = 0.9$. We implement the Best foot forward [14] as a baseline; it does not allow for parametrization and is thus a single point.

Rule	Parameter	Aggressive $\longleftrightarrow$ Conservative
β-farsighted	β	0.1, 0.3, 0.5, 0.7, 0.9, 0.99
γ-fixed	γ	8, 16, 32, 64, 128
δ-hopeful	δ	8, 16, 32, 64, 128
ϵ-hybrid	ϵ	0.1, 0.3, 0.5, 0.7, 0.9
ψ-support	ψ	1/2, 1/4, 1/6, 1/8, 1/10

Table 1: Parameters for Power-FDR curves in Figure 3

Figure 3a shows that all proposed α-investing rules can outperform the original Best foot forward rule by achieving more power. Which policy to use depends on the context information from the user. For the thrifty policies, i.e., policies which allow an unbounded number of exploration steps, β-farsighted at $\beta = 0.9$ achieves 1.9x higher power than the baseline Best foot forward, while having slightly higher but still bounded error rate at 0.05. For a more conservative approach, increasing β to 0.99 reduces error by more than 50% while still having 1.3x (second to the top data point) higher power than the Best foot forward. Varying ψ-support shows very little impact and its augmented version of β-farsighted achieves the highest power at lower realized FDR than other thrifty strategies.

If the exploration steps are known or can be approximated a priori, non-thrifty policies can be used to further reduce the error rates. Among the non-thrifty policies, δ-hopeful is ideal if the user has an understanding on how many exploration steps she expects in the targeted exploration. It achieves up to 2x the power of Best foot forward and never drops below it. The δ-hopeful also nicely visualizes the impact of being too optimistic versus too pessimistic. For example, $\delta = 8$ and $\delta = 128$ achieve similar power but $\delta = 128$ has a much lower FDR.

ϵ-hybrid offers the best combination of power and error rate with sharp drop of FDR while maintaining the power within 10% of its peak. It offers similar error rate as the baseline Best foot forward, but has 1.7x higher power. This shows the appealing aspect of ϵ-hybrid that it combines the high power of δ-hopeful and the low error of γ-fixed.

8.3 Free-form Exploration

In free-form exploration, the explorer does *not* start with a specific set of questions in mind, but rather starting with creating an initial understanding by exploring different aspects of the dataset before narrowing down to certain aspect that is interesting. In terms of hypothesis testing, the fraction of significance insights tends to be lower at the beginning, but increases towards the end because hypotheses become more based upon the previously discovered as "interesting" findings. To simulate this effect, we set the first phase of the Markov process with higher distribution of random noise ($\pi_a(0) = 0.75$) than the second phase ($\pi_b(0) = 0.25$) with a total of 128 hypotheses and switching time at 70% of the process.

In this scenario the non-thrifty policies do not perform as well if the parameters are set too lower than the number of expected exploration steps. However, if set correctly, e.g., at about half of the number of expected steps as $\delta = 64$, δ-hopeful has a similar power as Best foot forward but with a much lower error rate.

For unbounded exploration, β-farsighted with $\beta = 0.9$ achieves 1.18x power over Best foot forward, while adding ψ-support reduces its error rate by 20%, making it the overall best strategy.

8.4 Uniform Exploration

The last data exploration model does not build upon how the hypotheses are structured or ordered, and therefore represents an average case. Specifically, the significant insights during the data exploration process are observed uniformly at random.

Interestingly, the Power-FDR curves of the policies in Figure 3(c) look similar to the ones in the free-form exploration in (b). One difference is that the baseline Best foot forward performs worse, having similar power but almost 2x the error rate at close to 0.05. This demonstrates the downside of having a specialized policy as Best foot forward, which optimizes for a special case where the significant insights are highly clustered.

For an exploration with or without known expected length, our α-investing rules have parametrization that outperform the baseline Best foot forward in error rate. As for power, the β-farsighted with ψ-support offers the highest power. With information of expected exploration length, the ϵ-hybrid achieves almost the same power

but has 60% lower error rate than the baseline Best foot forward, making it the most balanced strategy.

8.5 Real-world Deployment

To test the applicability of the optimal parameters in the models, we use the U.S. Census dataset [25] with 10,000 records and derived a real exploration workflow from the user study in [36]. The workflow contains 117 hypotheses, similar to the configuration of our models. To generate ground truth labels, we run Bonferroni procedure [6] with family-wise error rate set to 10^{-20} on the Census dataset as the population. We then sample 10% of the Census dataset to run the user workflow with $\alpha = 0.05$.

We do not vary the parameters for this experiment , but instead pick the best overall settings from the Markov simulation with $\beta = 0.9$, $\delta = 64$, $\gamma = 64$, $\epsilon = 0.5$, $\psi = 1/4$.

Figure 3(d) shows a similar trend: our proposed α-investing rules achieves lower error rate than the baseline Best foot forward. If the expected exploration length is unknown, the β-farsighted with ψ-support is better. Otherwise with known expected exploration length, the non-thrifty policies such as γ-fixed, δ-hopeful and ϵ-hybrid achieve significantly less error than the thrifty policy β-farsighted. The ϵ-hybrid combines the higher power of δ-hopeful and lower error rate of γ-fixed and is the best-balanced strategy.

Finally, note that the Best foot forward overachieves on the real-world dataset than the simulation because of the way we generate the ground truth; the Bonferroni procedure creates a more benign setting for this strategy (note that statisticians usually only use simulations to eliminate this bias).

8.6 Discussion

To conclude, if the expected number of hypotheses can be estimated a priori, ϵ-hybrid provides high power with significantly lower error rate than the baseline Best Foot Forward. Otherwise if the exploration is unbounded, β-farsighted is the best policy. Overall the β-farsighted with ψ-support achieves the highest power often at a lower realized FDR and presents a good choice independent of the exploration scenario.

Finally, it should be noted that while some of the power/FDR improvements appear to be minor, they can have profound statistical impact in practice, such as determining which discoveries can be deemed scientific, how much more data has to be collected, or what exploration path the user takes.

In Appendix B we further compare QUDE's dynamic scheme against the static FDR method [2], the FWER control procedure [6], and the scheme without multiple hypotheses control to illustrate QUDE's overall safety and efficiency.

9. RELATED WORK

There has been surprisingly little work in controlling the number of false discoveries during data exploration even. This is especially astonishing as the same type of false discovery can also happen with traditional analytical SQL-queries. To our knowledge this is the first work to achieve an automatic control in tracking the user steps.

Most related to this work are all the various statistical methods for significance testing and multiple hypotheses control. Early works tried to improve the power of the Family Wide Error Rate using adaptive Bonferroni procedures such as Sĭdák [34], Holm [19], Hochberg [18], and Simes [35]. However, all these methods lack power in large scale multi-comparison tests.

The alternative False Discovery Rate measure was first proposed by Benjamini and Hochberg [2], and soon became the statistical criteria of choice in the statical literature and in large scale data exploration analysis for genomic data [27]. In the original FDR method, *all* hypotheses have to be collected and sorted by their p-values before determining the significance of each test. The Sequential FDR procedure [15] does not require the sorting of all the hypotheses, but still require to calculate *all* the tests before their corresponding significance can finalize. These procedures cannot determine the final signifance of each test incrementally and hence are not applicable to intearctive data exploration. The interactive data exploration motivated the study of interactive and adaptive techniques, such as α-investing [14], which can be applied in scenarios where hypotheses arrive sequentially and the testing procedure needs to decide "on the fly" whether to accept or reject each of the hypotheses before testing the next one, while maintaining a bound on the FDR. Depending on the observed order of hypotheses, Sequential FDR can overturn previously accepted hypotheses into rejections based on the subsequent hypotheses.

α-investing procedure also has revisiting policies that can potentially overturn previous decisions. The implication is that these procedures are incremental but non-interactive, because they require observing all the hypotheses before finalizing the decisions. However, it is often infeasible to obtain all the possible hypotheses a priori. Therefore our work concerns α-investing procedure with policies that are both incremental and interactive. In addition, none of the work addresses the issue on how to automatically integrate these techniques as part of an data exploration tool.

In a recent paper [11], Dwork et al. introduce a new adaptive testing procedure for streams of hypotheses which exploits concepts and techniques from differential privacy. Although this technique can reliably test up to m adaptively chosen hypotheses it has also several practical drawbacks: the computationally efficient version of the procedure requires the size of the available sample to be proportional to $\sqrt{m}$ and knowledge of the amount of hypotheses being tested is required.

In [5] Blum and Hardt presented "*the Ladder*", an algorithmic test procedure that, given a training dataset, reliably evaluates the quality of different versions a model by adaptively tuning the parameters. While this approach does not address the general issue raised in [2, 15, 14, 11], it shows good performance in the practical context of parameter tuning for machine learning.

10. CONCLUSION AND FUTURE WORK

In this paper we presented the first automatic approach to controlling the multiple hypothesis problem during data exploration. We showed how the QUDE systems integrates user feedback and presented several multiple hypothesis control techniques based on α-investing, which control *mFDR*, and are especially suited for controlling the error for interactive data exploration sessions. Finally, our evaluation showed that the techniques are indeed capable of controlling the number of false discoveries using synthetic and real world datasets. We consider this work as an important first step towards more sustainable discoveries in a time where the importance of data analysis is more pervasive than ever. We strongly believe that our work constitutes a departing point for a wealth of important research topics such as creating and evaluating other types of default hypothesis, developing new testing procedures (e.g., for interactive Bayesian tests) and investigating techniques to recover from cases where the testing procedure runs out of α-wealth.

11. ACKNOWLEDGMENTS

This research is funded in part by the Intel Science and Technology Center for Big Data,DARPA Award 16-43-D3M-FP-040, NSF CAREER Award IIS-1453171, NSF Award IIS-1514491, NSF Award IIS-1562657, Air Force YIP AWARD FA9550-15-1-0144, and gifts from Google, VMware, Mellanox, and Oracle.

12. REFERENCES

[1] E. Aharoni and S. Rosset. Generalized α-investing: definitions, optimality results and application to public databases. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 76(4):771–794, 2014.

[2] Y. Benjamini and Y. Hochberg. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the royal statistical society. Series B (Methodological)*, pages 289–300, 1995.

[3] Y. Benjamini and D. Yekutieli. The control of the false discovery rate in multiple testing under dependency. *Annals of statistics*, pages 1165–1188, 2001.

[4] D. A. Berry et al. Bayesian perspectives on multiple comparisons. *Journal of Statistical Planning and Inference*, 82(1–2), 1999.

[5] A. Blum and M. Hardt. The ladder: A reliable leaderboard for machine learning competitions. *arXiv preprint arXiv:1502.04585*, 2015.

[6] C. E. Bonferroni. *Teoria statistica delle classi e calcolo delle probabilita*. Libreria internazionale Seeber, 1936.

[7] A. Burgess, R. Wagner, R. Jennings, and H. B. Barlow. Efficiency of human visual signal discrimination. *Science*, 214(4516):93–94, 1981.

[8] A. Crotty, A. Galakatos, E. Zgraggen, C. Binnig, and T. Kraska. Vizdom: Interactive analytics through pen and touch. *Proceedings of the VLDB Endowment*, 8(12):2024–2027, 2015.

[9] J. Demšar. Statistical comparisons of classifiers over multiple data sets. *J. Mach. Learn. Res.*, 7:1–30, Dec. 2006.

[10] E. Dimara, A. Bezerianos, and P. Dragicevic. The attraction effect in information visualization. *IEEE Trans. Vis. Comput. Graph.*, 23(1), 2016.

[11] C. Dwork, V. Feldman, M. Hardt, T. Pitassi, O. Reingold, and A. L. Roth. Preserving statistical validity in adaptive data analysis. In *STOC*, pages 117–126. ACM, 2015.

[12] B. Efron and T. Hastie. *Computer Age Statistical Inference*, volume 5. Cambridge University Press, 2016.

[13] R. Fisher. *The design of experiments*. Oliver and Boyd, Edinburgh, Scotland, 1935.

[14] D. P. Foster and R. A. Stine. α-investing: a procedure for sequential control of expected false discoveries. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 70(2):429–444, 2008.

[15] M. G. G'Sell et al. Sequential selection procedures and false discovery rate control. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 78(2), 2016.

[16] H. Guo, S. Gomez, C. Ziemkiewicz, and D. Laidlaw. A case study using visualization interaction logs and insight. *IEEE Trans. Vis. Comput. Graph.*, 2016.

[17] P. Hanrahan. Analytic database technologies for a new kind of user: the data enthusiast. In *SIGMOD*, 2012.

[18] Y. Hochberg. A sharper bonferroni procedure for multiple tests of significance. *Biometrika*, 75(4):800–802, 1988.

[19] S. Holm. A simple sequentially rejective multiple test procedure. *Scandinavian journal of statistics*, pages 65–70, 1979.

[20] J. P. A. Ioannidis. Why most published research findings are false. *Plos Med*, 2(8), 2005.

[21] H. Jeffreys. *The theory of probability*. OUP Oxford, 1998.

[22] M. I. Jordan. The era of big data. *ISBA Bulletin*, 18(2), 2011.

[23] N. Kamat, P. Jayachandran, K. Tunga, and A. Nandi. Distributed and interactive cube exploration. In *Data Engineering (ICDE), 2014 IEEE 30th International Conference on*, pages 472–483. IEEE, 2014.

[24] R. Kohavi. A study of cross-validation and bootstrap for accuracy estimation and model selection. In *Proceedings of the 14th International Joint Conference on Artificial Intelligence - Volume 2*, IJCAI'95, pages 1137–1143, San Francisco, CA, USA, 1995. Morgan Kaufmann Publishers Inc.

[25] M. Lichman. UCI machine learning repository, 2013.

[26] Z. Liu, B. Jiang, and J. Heer. immens: Real-time visual querying of big data. In *Computer Graphics Forum*, volume 32, pages 421–430. Wiley Online Library, 2013.

[27] J. H. McDonald. *Handbook of Biological Statistics*. Sparky House Publishing, Baltimore, Maryland, USA, second edition, 2009.

[28] J. Neyman and E. L. Scott. Consistent estimates based on partially consistent observations. *Econometrica: Journal of the Econometric Society*, pages 1–32, 1948.

[29] P. Pirolli and S. Card. The sensemaking process and leverage points for analyst technology as identified through cognitive task analysis. In *Proceedings of international conference on intelligence analysis*, volume 5, pages 2–4, 2005.

[30] P. Refaeilzadeh, L. Tang, H. Liu, and M. T. ÖZSU. *Cross-Validation*, pages 532–538. Springer US, Boston, MA, 2009.

[31] M. Schemper. A survey of permutation tests for censored survival data. *Communications in Statistics-Theory and Methods*, 13(13):1655–1665, 1984.

[32] J. P. Shaffer. Multiple hypothesis testing. *Annual review of psychology*, 46, 1995.

[33] Y. B. Shrinivasan and J. J. van Wijk. Supporting the analytical reasoning process in information visualization. In *Proceedings of the SIGCHI conference on human factors in computing systems*, pages 1237–1246. ACM, 2008.

[34] Z. Šidák. Rectangular confidence regions for the means of multivariate normal distributions. *Journal of the American Statistical Association*, 62(318):626–633, 1967.

[35] R. J. Simes. An improved bonferroni procedure for multiple tests of significance. *Biometrika*, 73(3):751–754, 1986.

[36] E. Zgraggen, A. Galakatos, A. Crotty, J.-D. Fekete, and T. Kraska. How progressive visualizations affect exploratory analysis. *IEEE Trans. Vis. Comput. Graph.*, 2016.

[37] A. F. Zuur, E. N. Ieno, and C. S. Elphick. A protocol for data exploration to avoid common statistical problems. *Methods in Ecology and Evolution*, 1(1):3–14, 2010.

APPENDIX

A. SYMBOL TABLE

The following Table 2 summarizes the important symbols and notations used in this paper.

B. SUPPLEMENTAL EXPERIMENTS

We provide additional experimental results to illustrate the improved safety QUDE provides over the scheme without any multiple comparisons control, and the higher power over the static control methods, including the false discovery rate (FDR) control procedure such as Benjamini-Hochberg (BHFDR) [2], and the family-wise error rate (FWER) control procedure such as Bonferroni [6]. In the following experiments we use the exploration settings as in 8.1 with 64 user observations, but with varying ratio of randomness.

H	The set $\{H_1, \ldots, H_m\}$ of null hypotheses.
$\mathcal{H}$	The set $\{\mathcal{H}_1, , \ldots, \mathcal{H}_m\}$ of corresponding alternative hypotheses.
R	The number of null hypotheses rejected by the testing procedure (i.e., the discoveries).
V	The number of erroneously rejected null hypotheses (i.e., false discoveries, false positives, Type I errors).
S	The number of correctly rejected null hypotheses (i.e., true discoveries, true positives,).
$R(j)$	The number of discoveries after j tested hypotheses.
$V(j)$	The number of false discoveries after j tested hypotheses.
$S(j)$	The number of false discoveries after j tested hypotheses.
m	The number of tested hypotheses.
p_j	The p-valuecorresponding to the null hypothesis H_j.
$W(0)$	Initial wealth for the α-investing procedures.
$W(j)$	Wealth of the α-investing procedures after j tests.
α	Significance level for the test with $\alpha \in (0, 1)$.
η	Bias in the denominator for $mFDR_\eta$.

Table 2: Notation Reference

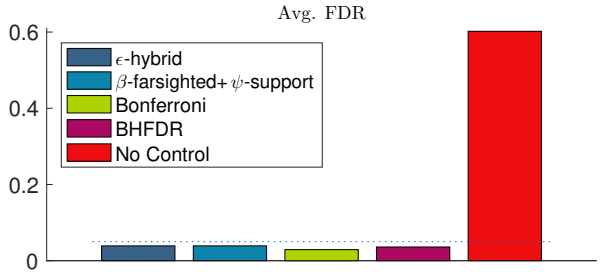


Figure 4: Avg. False Discovery Rate on Random Data

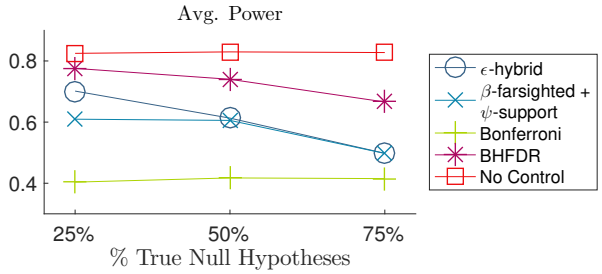


Figure 5: Avg. Power on Data with Varying Uncertainty

B.1 Safety against Uncertainty

We further discuss the impact of false discovery on completely random data to show that in this worst case QUDE guarantees the bounded error rate, providing the same guarantee as the static FDR control procedure BHFDR and the FWER control procedure Bonferroni. On the other hand, the scheme without multiple comparison control is highly error-prone.

We use t-tests on all true null hypotheses generated from normal random variables with the same mean. The *average* FDR is based on 1000 repetitions, where for each repetition the *realized* FDR is either 1 or 0. As shown in Figure 4, the scheme without control results in as high as 60% false insights for a moderate number of 64 user observations. By contrast, methods with multiple hypotheses control achieve the guaranteed error rate as low as 5%.

B.2 FDR versus FWER

We compare the two multiple hypothesis control targets, FDR and FWER, to show that FDR (and its variant mFDR) provides the best trade-off in that it has much higher statistical power while having the same worst-case error bound than the FWER. This comparison motivates our choice of FDR as the control target for interactive data exploration.

On completely random data, controlling FDR also implies controlling FWER [2]. This can be seen in Figure 4 that both of our representative thrifty and non-thrifty α-investing rules achieve the same error rate as the Bonferroni at 0.05.

With varying degree of uncertainty as in Figure 5, the power of the Bonferroni remains only as low as 40%, whereas the power of the (m)FDR control procedures all achieve significantly higher power. With less randomness, i.e. lower fraction of true null hypotheses, both of QUDE's representative thrifty and non-thrifty α-investing rules achieve close to the static FDR control method BHFDR and the scheme without control.

As the uncertainty increases, the power of our α-investing rules maintain higher than the FWER control procedure. The static method BHFDR has slightly better power than our α-investing rules because it operates offline where it collects and sorts all hypotheses based on their final p-values, which however limits its usability on dynamic settings; whereas our α-investing rules removes this limitation to support interactive data exploration without significant loss of power while with the similar guarantee on the error rate.

Finally, although the scheme without control achieves slightly higher power, it is at the expense of much higher error rates, such as about 100X on random data. By contrast, QUDE represents the optimal trade-off between the power and the risk.