VizCertify: A framework for secure visual data exploration

Lorenzo De Stefani, Leonhard F. Spiegelberg, Eli Upfal
Department of Computer Science
Brown University
Providence, United States of America
{lorenzo, lspiegel, eli}@cs.brown.edu

Tim Kraska
CSAIL
Massachusetts Institute of Technology
Cambridge , United States of America
kraska@mit.edu

***Abstract*—Recently, there have been several proposals to develop visual recommendation systems. The most advanced systems aim to recommend visualizations, which help users to find new correlations or identify an interesting deviation based on the current context of the user's analysis. However, when recommending a visualization to a user, there is an inherent risk to visualize random fluctuations rather than solely true patterns: a problem largely ignored by current techniques.**

In this paper, we present *VizCertify*, a novel framework to improve the performance of visual recommendation systems by quantifying the statistical significance of recommended visualizations. The proposed methodology allows to control the probability of misleading visual recommendations using both classical statistical testing procedures and a novel application of the Vapnik Chervonenkis (VC) dimension towards visualization recommendation which results in an effective criterion to decide whether a recommendation corresponds to a true phenomenon or not.

***Index Terms*—Visualization, Recommendation Systems, FWER, Data Exploration, Visual Recommendation**

I. Introduction

Recently many visual recommendation engines [1]–[8] have been proposed to help savvy and unsavvy users with data analysis. While some recommendation engines (e.g., [2], [8], [4]) aim to recommend better visual encodings, others (e.g., SeeDB [1], DeepEye [7], MuVE [5], or VizDeck [6]) aim to automatically recommend entirely new visualization to help users finding interesting insights.

While the latter are significantly more powerful, they also significantly increase the risk of finding false insights. This happens whenever a visualization is used as a device to represent statistical properties of the data. Consider a user exploring a dataset containing information about different wines. After browsing the data for a while, she creates a visualization of wines ranked by origin showing wines from France to be apparently higher rated. If her only takeaway is, that in *this particular dataset* wines from France have a higher rating, there is no risk of a false insight. However, it is neither in the nature of users to constrain themselves to such thinking [9], nor would visualizing the data be insightful then. Rather, users most likely would infer that French wines are *generally* rated higher; *generalizing* their insight to all wines. Statistically savvy users will now test whether this generalization is actually statistically valid using an appropriate test. Even more technically savvy users will also consider other hypothesis they tried and adjust the statistical testing procedure to account for the multiple comparisons problem. This is important since every additional hypothesis, explicitly expressed as a test or implicitly observed through a visualization, increases the risk of finding insights which are just spurious effects.

The issue with visual recommendation engines: What happens when visualization recommendations are generated by one of the above systems? First, the user does not know if the effect shown by the visualization is actually significant or not. Even worse, she can not use a standard statistical method and simply test the effect shown in the visualization for significance. Visual recommendation engines are potentially checking hundreds of thousands of visualizations for their interesting-factor. As a result, by testing large quantities of visualizations it is very likely that a system will find something "*interesting*" regardless of whether the observed phenomenon is actually statistically valid or not. A test for significance for the recommended visualization should therefore consider the whole testing history.

Why not a holdout? Advocates of visual recommendation engines usually argue that visual recommendations systems are meant to be hypothesis generation engines, which should always be validated on a separate hold-out dataset. While this is a valid method to control false discoveries, it is has several unpractical implications: (1) *None* of the found insights from the exploration dataset should be regarded as an actual insight before they are validated. This is clearly problematic if one observation may steer towards another during the exploration. (2) Splitting a dataset into an exploration and a hold-out set can significantly reduce the power (i.e., the chance to find actual true phenomena). (3) The hold-out needs to be controlled for the multi-hypothesis problem unless the user only wants to use it exactly once for a single test.

In this paper, we present VizCertify, a first framework to make visual recommendation engines "*safe*".

Scope: As already pointed out in [10], [11], even defining a hypothesis test for a visualization is notoriously hard. We therefore, decided to focus on the visual recommendation technique proposed by SeeDB for histograms as it uses a clear semantic for what "interesting" means and rumored to be implemented in a widely-used commercial product. SeeDB [1]

978-1-7281-4493-1/19/$31.00 ©2019 IEEE
DOI 10.1109/DSAA.2019.00039

makes recommendations based on a reference view; it tries to find a visualization whose underlying data distribution is different from the one's the user is currently shown (i.e., the implicit test performed by SeeDB is for difference in sub-populations). However, our techniques can be extended to other visualization frameworks and hypothesis generation tools (e.g., Data Polygamy [12]) provided that the chosen "interestingness" criterion can be evaluated by a statistical test.

Contributions: The core idea of VizCertify is to *estimate* the interest of candidate recommendations and automatically adjusts the significance value based on the search space to account for the multiple-hypothesis pitfall. With this work we make the following contributions:

- We formalize the process of making visualization recommendations as statistical hypothesis testing.
- We propose a method based on the use of VC dimension, which allows controlling the probability of observing false discoveries during the visualization recommendation process. VizCertify allows control of the *Family Wise Error Rate* (FWER) at a given control level $\delta \in (0, 1)$. Our method provides *finite sample* FWER control whereas classical statistical approaches (i.e., the *chi-squared test*) ensure *asymptotic* FWER control.
- We evaluate the performance of our system, in comparison with SeeDB via extensive experimental analysis.

The remainder of this paper is organized as follows: In Section II we give a definition of the visualization recommendation problem in rigorous probabilistic terms. In Section III we discuss possible approaches for the visualization recommendation problem, and highlight why they suffer due to having to account for a high number of statistical tests. In Section IV we introduce our VizCertify approach to overcome these problems. In Section V we present an extensive experimental evaluation of the effectiveness of VizCertify.

II. Background

In the following, we first provide an intuitive example on how SeeDB makes recommendations (Section II-A) before formalizing SeeDB's technique and its connection to hypothesis testing.

A. A SeeDB Recommendation Example

SeeDB makes recommendations based on a reference visualization, referred to as a *reference view*, by adding or modifying filter conditions to find sub-populations of the data, referred to as *target views*, which if visualized in the same way show the largest deviation from the reference view. For example, consider the reference view in Figure 1a over a survey dataset we conducted with $2,644$ participants on Amazon Mechanical Turk for 35 (mostly unrelated) multiple-choice questions. It shows that the majority do not believe Astrology. Based on this visualization, SeeDB's top 3 recommended visualization are shown in Figure 1b-1d. SeeDB proposes various ways to measure the difference between two visualizations to rank the visualization. For this experiment we configured SeeDB to use the Chebyshev distance. For example, Figure 1b shows that people who prefer cheese-flavored potato chips are more likely to believe in Astrology.

However, the visualizations alone do not answer the question, if the shown difference is actually statistically significant or not. In contrast, VizCertify would automatically mark the reference view in Figure 1a and 1c as statistically significant, but disregard the visualizations in Figure 1b and 1d. Important to note is, that the top-ranked visualization, Figure 1b, is not considered statistically significant, but the second one is; so higher rank does not necessarily imply higher statistical significance.

Obviously, detecting statistically significant differences is challenging as it depends on the data size, the number of tested hypothesis by the user and the recommendation system. Furthermore, while it is easy to define what a test means with a two-bar chart histogram, it is much harder for visualizations with multiple bars. Finally, it has to be pointed out that in the case of SeeDB increasing the data size does not necessarily mitigate the problem as SeeDB automatically adds filter conditions to create ever smaller sub-populations.

B. Problem set-up

In this work we assume that $\mathcal{D}$ (e.g., our survey data) consists of n records chosen uniformly at random and independently from a universe Ω. We can imagine Ω in the form of a two-dimensional, relational table with N rows and m columns, where each column represents a *feature* or *attribute* of the records.

We refer to $\mathcal{D}$ as the *sample dataset* or the *training dataset* and denote by $\mathcal{F}_{\mathcal{D}}$ the probability distribution that generated the sample $\mathcal{D}$. Alternatively, one can consider $\mathcal{D}$ as a sample of size n from a distribution $\mathcal{F}_{\mathcal{D}}$ that corresponds to a possibly infinite domain. As discussed in Section I rather than enabling users to only interpret visualizations for *some particular data set*, we want to make sure that a system provides them with statistical guarantees when they try to generalize their results over Ω. While not universally verified, these assumptions do indeed hold for many realistic datasets (e.g., for a controlled survey as in our running example).

C. Visualizations

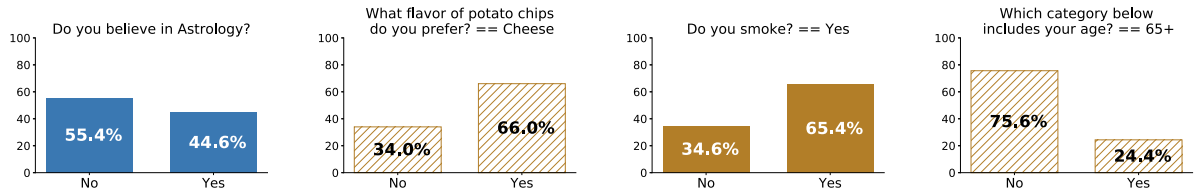
As outlined previously our focus is on bar graphs. To simplify the presentation, we restrict our discussion in the following on categorical features/attributes, but it is possible to use our techniques with continuous features by standard binning techniques. Thus, we can define SeeDB's recommended visualization as:

Definition 1: A bar graph visualization V is a tuple $\big(\mathcal{D}, F, X, Y, \mathrm{AGG}\big)$ which describes the result of a *query* of the form

```
SELECT X, AGG(Y) FROM D WHERE F GROUP BY X
```

which can be plotted as a bar graph. The aggregate `AGG` is partitioned according to the values of a discrete feature `X`, after filtering the records of the input dataset $\mathcal{D}$ by predicate `F`.

Again for simplicity we focus in the following on `COUNT(Y)` aggregates. However, our approach can be ex-



(a) Reference View (b) SeeDB Top 1 (c) SeeDB Top 2 (d) SeeDB Top 3

Fig. 1: An example of SeeDB [1] and VizCertify on survey data. VizCertify only recommends (a) and (c).

tended with minor modifications to the *average* **AVG**(Y) aggregate.

Visualizations as distributions: Thus, the bar graph obtained by a query of the above form can be seen as a histogram of the *probability mass function* (pmf) of a *discrete* random variable X corresponding to the feature X conditioned (or filtered) on the predicate (or event) F. This correspondence between visualizations and distributions provides us with a *natural criterion* to compare visualizations by evaluating their *statistical difference*.

To close the connection to statistical testing, we further define the *support* of a visualization V as the number of records of D which satisfy the predicate F, and denote it as $|V|$. The *selectivity* of a visualization V, denoted as γ_V, is defined as the fraction of records which satisfy F:

$$\gamma_V := |\mathcal{D}|F|/|\mathcal{D}|. \tag{1}$$

The aggregate **COUNT**(Y) corresponds to the number of records which satisfy the query predicate F grouped according to the values of the feature X, henceforth referred as the "*group-by*" feature.

Extensions While our definition seems restrictive to SeeDB and certain aggregations, it our results can easily be extended to other types of visualizations (e.g., *heat maps*) and systems, and in some cases do not require any modifications. For example, without any significant changes our techniques can be used to make Data Polygamy [12], an automatic correlation finder, "safe". Similarly, it is possible to extend our techniques to other types of counting-based visualizations (e.g., *heat maps*). In contrast, we do not consider Min or Max aggregates, which are supported by SeeDB as useful as they (obviously) can not represent *statistically significant* behavior of distributions. However, if they would be expressed as *conditional expectations* (e.g., aggregates in the form $Y|Y > c$ for some constant $c \in \mathbb{R}$), our framework can in fact handle them.

D. Visualization recommendations

An user starts with a reference visualization V_1 for which visualizations $\{V_q\}_{q=2,\ldots,Q}$ shall be recommended that are interesting with respect to V_1. A candidate visualization V_q is defined to be interesting with respect to a visualization V_1 if and only if V_q and V_1 have a different distribution of the group-by feature X under the predicates associated with V_1 and V_q, respectively. Consistently, SeeDB defines the best recommendations as the ones, which have the highest difference with respect to V_1 among all eligible candidates $\{V_q\}_{q=2,\ldots,Q}$. Note that we constrain candidates to share the same group-by feature with the reference visualization to guarantee that the visualizations are in the same user context.

A visualization V_2 is an eligible candidate for recommendation with respect to a reference visualization V_1 if the two are "*different enough*". That is, if their distance $d(V_1, V_2)$ is larger than some given threshold value ϵ:

$$\mathcal{V}_2 \text{ interesting w.r.t } \mathcal{V}_1 \iff d(V_1, V_2) > \epsilon.$$

In its simplest form, ϵ may be zero. However it makes more sense to define ϵ in terms of a minimum visual distance ϵ_V required by a user to spot a difference [13] when shown both visualizations. That is, a recommended visualization is only interesting if it shows a strong visual difference usually for a subset of the currently selected data.

In this work, we use the Chebyschev distance to decide whether two visualizations are different. Given two pmfs $\mathcal{D}_1$ and $\mathcal{D}_2$ over the same support set $\mathcal{X} = \{x_1, x_2, \ldots, x_n\}$, the Chebyschev distance between $\mathcal{D}_1$ and $\mathcal{D}_2$ is defined as:

$$d\left(D_1, D_2\right) := \max_{x \in \mathcal{X}} |\Pr_{\mathcal{D}_1}(x) - \Pr_{\mathcal{D}_2}(x)|, \tag{2}$$

Here, $\Pr_{\mathcal{D}_1}(x)$ denotes the probability of a random variable taking value x according to the distribution $\Pr_{\mathcal{D}_1}(x)$.

We justify this choice of distance measure for two visualizations by its ability to capture the maximum difference between pairs of corresponding columns of the two bar plots (or histograms).

III. Statistically safe visualizations

While the statistical pitfalls of exploratory data analysis are well understood and documented ([14], [15]) the connection to visualizations only has been rigorously studied recently [11], [16], [17].

A first crucial demand for a system that provides users with visualizations unveiling *interesting relationships* amongst visualizations is to provide tools allowing the analyst to ascertain that phenomena being observed are actually *statistically relevant*. Furthermore, results displayed need to exhibit characteristics that are *non-random* and *visually intelligible*. That is, a user looking at two visualizations should both be able to understand that they are different and why they are different without worrying whether visual features are due to missing support or random noise.

Thus in probabilistic terms the question becomes how likely it is that the visualization is also interesting with respect to the underlying distribution $\mathcal{F}_\mathcal{D}$ for a dataset $\mathcal{D}$ available to the visualization itself. Recalling the correspondences drawn in Section II-C, each query with a filter predicate $\mathtt{F}$ corresponds to an event E over Ω. Consider a visualization of a histogram of a (discrete and finite) variable X conditioned on an event $E \neq \emptyset$, denoted $X|E$ as defined in Definition 1. The true values (in $\mathcal{F}_\mathcal{D}$) for $k \in \operatorname{dom} X$ are given by $p_k := \mathbb{P}(X = k|E)$. Given a dataset $\{X_1, \ldots, X_n\}$ we estimate the p_k values as

$$\hat{p}_k := \frac{\sum_{i=1}^{n} \mathbb{1}_{\{X_i=k,\ X_i \in E\}}}{\sum_{i=1}^{n} \mathbb{1}_{\{X_i \in E\}}}. \tag{3}$$

If in $\mathcal{D}$ the histogram of $X|E$ is visually different from the histogram of X, what can we tell about the difference between the histograms in $\mathcal{F}_\mathcal{D}$? Here, we say that the difference between two visualizations $\mathcal{V}_1$ and $\mathcal{V}_2$ is *statistically significant* if and only if the difference observed between the two in the finite sample $\mathcal{D}$ is due to a difference between the two histograms with respect to $\mathcal{F}_\mathcal{D}$. The recommendation problem is thus the task of verifying that visualizations flagged as interesting with respect to $\mathcal{D}$ *generalize* to interesting visualizations with respect to the *true* underlying distribution $\mathcal{F}_\mathcal{D}$.

A. Classical statistical testing

In the *classical* statistical testing setting, our problem could be formulated either as a *goodness-of-fit* test or as a *homogeneity* test. Example of classical goodness-of-fit (resp., homogeneity) tests include the single sample (resp., two sample) χ^2-test for discrete distributions or the Kolmogorov-Smirnov test for continuous random variables. However, there are major difficulties in applying *classical* standard statistical tests to the visualization problem.

First, depending on the input data the correct test needs to be selected. For example, when using a χ^2-test over discrete attributes, each bucket must not be empty. A general rule of thumb to make sure estimates are reliable is to have at least 5 samples per bucket. Further, there need to be enough samples to actually use the χ^2-test. Else, Fisher's exact test should be used for small sample sizes. In addition to each test being only applicable to certain input data, they generally consider *different* notion of *difference*, and, hence, interest. Such nonhomogeneity may considerably hinder the user's ability to connect the results of the test to the notion of a *significant visual difference*.

A second issue has to do to with opportunely defining the hypotheses to be tested so that they allow recognizing visual differences in a meaningful way. Consider for this a t-test that essentially compares whether the observed mean resembled the expected mean. Naturally, a consequence is that if they differ the candidate query should get recommended. This may however lead to many wrong recommendations merely because the null hypothesis used is too *simple* and gets rejected too often.

Third, classical tests, such as the χ^2-test, only offer *asymptotic guarantees*. That is due to the fact the validity of the test hinges on the fact that the p-value of true null-hypotheses is uniformly distributed between 0 and 1. While this holds as the size of the size of the available sample grows, such assumption does not generally holds for samples of finite, small, size. For skewed distributions or queries that return only a small number of rows this is problematic. For example, when using a χ^2-test for a heavily skewed discrete distribution with a high number of *degrees of freedom* (e.g., > 20), a high number of samples are required for the the observed test statistic to converge to the χ^2 distribution. In contrast, our VizCertify method ensures rigorous finite sample FWER control with no further assumption of the test statistics of true null hypotheses.

Lastly, while it might be tempting to combine statistical testing with a selection heuristic based on the distance measure introduced in (2), this would not eliminate the risk of false discoveries. While the initial statistical tests identify some candidate visualizations as *different* from the reference, they do not provide guarantees on the statistical significance of such difference (see the examples in Section V-C).

B. Recommendation validation via estimation

Rather in VizCertify we use the sample dataset $\mathcal{D}$ to obtain *approximations* of the visualization according to the entire global sample space Ω. Consider a single histogram visualization $\mathcal{V}_1$, and assume it is comprised of K bars, one for each of the K possible values of the chosen group-by feature X. Let $p_{\mathcal{V}_1}(x_1), \ldots, p_{\mathcal{V}_1}(x_k)$, denote the normalized bars corresponding to $\mathcal{V}_1$, i.e. $\sum_{k=1}^{K} p_{\mathcal{V}_1}(x_k) = 1$. Note that such bars denote the probability of a randomly chosen record from Ω for the event $X = x_i$ conditioned on the fact that a record satisfies the predicate associated with $\mathcal{V}_1$. Using $\mathcal{D}$ and Equation 3 we estimate $p_{\mathcal{V}_1}(x_i)$ with $\hat{p}_{\mathcal{V}_1}(x_i)$. In order to provide guarantees for these estimates, it is necessary to bound the maximum difference between the correct and estimated sizes of bars in the normalized histograms.

In particular for a given δ (i.e., our level of control for *false positive recommendations*) we want to compute the minimum value $\epsilon \in (0, 1)$ such that $\Pr_\mathcal{D}\left(|p_{\mathcal{V}_1}(x_i) - \hat{p}_{\mathcal{V}_1}(x_i)| > \epsilon\right) < \delta$. In addition, ϵ quantifies the accuracy of an estimation $p_{\mathcal{V}_1}(x_i)$ obtained by its empirical counterpart $\hat{p}_{\mathcal{V}_1}(x_i)$.

Let F denote the predicate associated with our visualization $\mathcal{V}_1$. We denote as $\Omega|F$ (resp., $\mathcal{D}|F$) the subset of Ω (resp., $\mathcal{D}$) which is composed by those records that satisfy the predicate F. Given X the value $p_{\mathcal{V}_1}(x_i)$ (resp., $\hat{p}_{\mathcal{V}_1}(x_i)$) corresponds to (resp., is computed as) the *relative frequency* of records such that $X = x_i$ in $\Omega|F$ (resp., $\mathcal{D}|F$) which is expressed in the following fact:

Fact 1: Let $\mathcal{D}$ be an uniform random sample of Ω composed by m records. For any choice of predicate, as specified in Definition 1, the subset $\mathcal{D}|F$ is a uniform random sample of $\Omega|F$ of size $|\mathcal{D}|F|$.

As a direct consequence $\hat{p}_{\mathcal{V}_1}(x_i)$ is an *unbiased estimator* for $p_{\mathcal{V}_1}(x_i)$.

$$E_\mathcal{D}\left[\hat{p}_{\mathcal{V}_1}(x_i)\right] = p_{\mathcal{V}_1}(x_i). \tag{4}$$

In order to bound the estimation error $|p_{\mathcal{V}_1}(x_i) - \hat{p}_{\mathcal{V}_1}(x_i)|$ it is therefore sufficient to bound the *deviation from expectation* of the empirical estimate $\hat{p}_{\mathcal{V}_1}(x_i)$. Chernoff-Bounds [18] yield

$$\Pr_{\mathcal{D}}\left(|p_{\mathcal{V}_1}(x_i) - \hat{p}_{\mathcal{V}_1}(x_i)| > \epsilon\right) \leq e^{-2|\mathcal{D}|F|\epsilon^2}. \quad (5)$$

which can be rewritten to

$$\Pr_{\mathcal{D}}\left(|p_{\mathcal{V}_1}(x_i) - \hat{p}_{\mathcal{V}_1}(x_i)| > \epsilon\right) \leq e^{-2\gamma_{\mathcal{V}_1} n\epsilon^2} \quad (6)$$

using the *selectivity of a visualization* $\gamma_{\mathcal{V}_1}$ and $n = |\mathcal{D}|$. Equation (6) implies that a higher selectivity of a visualization leads to a better estimate and vice versa. Even though the method based on an application of the Chernoff bound appears to be very useful and practical, it is important to note that it only offers guarantees on the quality of the approximation of a single bar for a single visualization. While it is in general possible to combine multiple applications of the Chernoff bound, the required correction leads to a quick and marked decrease of the quality of the bound. For a visualization $\mathcal{V}_1$ composed by K bars a bound on all bars would be

$$\Pr_{\mathcal{D}}\left(\max_{i=1,\ldots,K} |p_{\mathcal{V}_1}(x_i) - \hat{p}_{\mathcal{V}_1}(x_i)| > \epsilon\right) \leq K e^{-2\gamma_{\mathcal{V}_1} n\epsilon^2}$$

through the *union bound* [18]. While tolerable for small K, the performance decreases for larger K leading potentially to a complete loss of significance of the bound itself.

C. Adaptive Multi-Comparisons

If we let a recommendation system explore an unlimited number of possible visualizations, it will eventually find an "*interesting*" one, even in random data. To avoid this, one may test every visualization recommendation on an independent sample that has not been used during the exploration and recommendation process yet. However, this solution is not practical for a system that explores many possible visualizations. Moreover, it is desirable for a system to use the entirety of the available data because it would allow to discover all interesting insights and bolsters the confidence of any statistical method being used in addition.

Assume that in our exploration of possibly interesting visualizations we tried ℓ different visualization patterns, and we computed for each of these patterns a bound h_i, $i = 1, \ldots, \ell$, on the probability that the corresponding observation in the sample $\mathcal{D}$ does not generalize to the distribution with respect to Ω. It is tempting to conclude that the probability that none of the ℓ visualizations generalize is bounded by $\sum_{i=1}^{\ell} h_i$. Unfortunately, this probability is actually much larger when the choice of the tested visualization depends of the outcome of prior tests. This phenomenon is often referred to as Freedman's paradox [19] and the only known practical approach to correct for it is to sum the error probability of all possible tests, not only the tests actually executed[1]. Note that standard statistical techniques for controlling the Family-Wise- Error-Rate (FWER) or the False Discovery Rate (FDR) require that the collection of tests is fixed independent of the data and therefore do not apply to an interactive or adaptive exploration scenario.

A possible solution would be fixing a-priory the set of possible visualizations to be considered (e.g., those obtained by predicates combining at most k features). Let M denote the size of such set. By recommending visualization that pass the individual visualization test with confidence level $\leq \alpha/M$ we are guaranteed, by the union bound, that the probability that any of our recommendations does not generalize is bounded by α. As we show in the experiments section this method is only effective for relatively small search space, as for high values of M the individual tests become extremely conservative with a consequent overall loss of statistical power.

[1]Theoretical methods, such as differential privacy [20] claim to offer an alternative method to address this issue. In practice however, the signal is lost in the added randomization before it becomes practical.

IV. Statistical Guarantees Via Uniform Convergence Bounds

In order to overcome the challenge of multiple comparisons, we propose to use Vapnik-Chernovenkis (VC) dimension to constraint the visualizations toonly statistically valid ones. To our knowledge, this is the first use of VC-dimensions for visualizations or recommendation systems in general.

A. VC-Dimension Background

The Vapnik-Chernovenkis (VC) dimension is a measure of the complexity or expressiveness of a family of indicator functions (or equivalently a family of subsets) [21]. Formally, VC-dimension is defined on *range spaces*:

Definition 2: A *range space* is a pair (X, R) where X is a (finite or infinite) set and R is a (finite or infinite) family of subsets of X. The members of X are called *points* and those of R are called *ranges*.

Note that both X and R can be infinite. Consider now a projection of the ranges into a finite set of points A:

Definition 3: Let (X, R) be a range space and let $A \subset X$ be a finite set of points in X.

1) The *projection* of R on A is defined as

$$P_R(A) = \{r \cap A \ : \ r \in R\}.$$

2) If $P_R(A) = 2^{|A|}$, then A is said to be *shattered by* R.

The VC-dimension of a range space is the cardinality of the largest set shattered by the space:

Definition 4: Let (X, R) be a range space. The *VC-dimension* of (X, R), denoted $\mathsf{VC}(X, R)$ is the maximum cardinality of a shattered subset of X. If there are arbitrary large shattered subsets, then $\mathsf{VC}(X, R) = \infty$.

Note that a range space (X, R) with an arbitrarily large (or infinite) set of points X and an arbitrary large family of ranges R can have bounded VC-dimension (see section IV-B). VC-dimension, allows to characterize the sample complexity of a learning problem as it is allows to obtain a tradeoff between the number of sample points being observed by a learning algorithm and the performance achievable by the algorithm itself.

Consider a range space (X, R), and a fixed range $r \in R$. If we sample uniformly at random a set $S \subset X$ of size

$m := |S|$ we know that the fraction $\frac{|S \cap r|}{|S|}$ rapidly converges to the frequency of elements of r in X. A finite VC-dimension implies an explicit upper bound on the number of random samples needed to achieve such convergence within a predefined error bound (known as *uniform convergence property*). For a formal definition we need to distinguish between finite X, where we case estimate the sizes r, and infinite X, where we estimate $Pr(r)$, the frequency of r in a uniform distribution over X.

Definition 5 (Absolute approximation): Let (X, R) be a range space and let $0 \leq \epsilon \leq 1$. A subset $S \subset X$ is an absolute ϵ-approximation for X iff for all $r \in R$ we have that for finite $S \subseteq X$,

$$\left| \frac{|r|}{|X|} - \frac{|S \cap r|}{|S|} \right| \leq \epsilon. \tag{7}$$

[22] establishes an interesting connection between the VC dimension of a range space (X, R) and the number of samples which are necessaries in order to obtain absolute ϵ-approximations of X itself.

Theorem 1 (Sample complexity [22]): Let (X, R) be a range-space of VC-dimension at most d, and let and $0 < \epsilon, \delta < 1$. Then, there exists an absolute positive constant c such that any random subset $S \subseteq X$ of cardinality

$$|S| \geq \frac{c}{\epsilon^2} \left(d + \log_2 \delta^{-1} \right) \tag{8}$$

is an ϵ-approximation for X with probability at least $1 - \delta$. The constant c was shown experimentally [23] to be at most 0.5. Indeed, we use $c = 0.5$ in our experimental evaluation.

B. Statistically Valid Visualization using VC

To apply the uniform convergence method via VC dimension to the visualization setup, we consider a range space (Ω, R), where Ω is global domain, and R consists of all the possible subsets of X that can be selected by visualizations predicates. That is, R includes all the subsets that correspond to *any* bar for *any* visualization which can be selected using the appropriate predicate filter. Given a choice of possible allowed predicates, we refer to the associate set of ranges as the "*query range space*" and denote it as Q.

The VC dimension of a query range class is a function of the type of select operators (i.e., $>, <, \geq, \leq, =, \neq$) and the number of (non-redundant) operators allowed on each feature in the construction of the allowed predicates. Note that depending on the *domain* of the selected features and the complexity according to which the predicate filters can be constructed, the number of possible predicates may be infinite. In order to use the VC-approach it is however sufficient to efficiently compute a finite *upper bound* of the VC-dimension of the set of *allowed predicates*. In order to use the results from the previous section, we have to ensure that the sample $\mathcal{D}$ provides an ϵ-approximation for the values $p_{\mathcal{V}}$ for all the visualizations being part of the query range space Q. To accomplish this, we introduce the following, well known, property:

Fact 2: Let (X, R) be a range space of VC dimension d. For any $X' \subseteq X$, the VC-dimension of (X', R) is bounded by d.

In conjunction with Theorem 1 this synthesizes:

Lemma 1: Let (Ω, Q) denote the range space of the queries being considered with VC dimension bounded by d, and let $\delta \in (0, 1)$. Let $\mathcal{D}$ be a random subset of Ω. Then there exists a constant c, such that with probability at least $1 - \delta$ for any filter F defined in Q we have that the subset $\mathcal{D}|F$ is an ϵ_F-approximation of $\Omega|F$ with:

$$\epsilon_F \geq \sqrt{\frac{c}{|D|F|} (d + \log_2 \delta^{-1})}.$$

Proof: Fact 1 ensures that given the dataset $\mathcal{D}$, for any choice of a predicate F we have that $\mathcal{D}|F$ is a random sample of $\Omega|F$. Therefore regardless of the specific choice of the predicate, we have that the VC dimension of the reduced range $(\Omega|F, Q)$ is bounded by d. From Theorem 1 we have that if:

$$|\mathcal{D}| \geq \frac{c}{\bar{\epsilon}^2} \left(d + \log_2 \delta^{-1} \right) \tag{9}$$

then $\mathcal{D}|F$ is an $\bar{\epsilon}$ approximation for the respective set $\Omega|F$. ■

Lemma 1 provides us an efficient tool to evaluate the quality of our estimations $\hat{p}_{\mathcal{V}}$ of the actual ground truth values $p_{\mathcal{V}}$ for any choice of predicate associated with the visualization. In particular, Lemma 1 verifies that the quality decreases gradually the more *selective* the predicate associated with a visualization is. That is, the smaller the cardinality of $|\mathcal{D}|F|$, the higher the *uncertainty* $\bar{\epsilon}$.

Corollary 1: Let $\mathcal{D}$ be a random sample from Ω, and let Q be a query range space with VC dimension bounded from above by d. For any visualization with $\mathcal{V} \in Q$ and for any value $\delta \in (0, 1)$ we have that

$$\Pr\{\left| p_{\mathcal{V}}(X = x_i) - \hat{p}_{\mathcal{V}(X=X_i)} \right| \geq \bar{\epsilon}\} < \delta, \tag{10}$$

where

$$\bar{\epsilon} \geq \sqrt{\frac{c}{|D|F|} \left(d + \log_2 \frac{1}{\delta} \right)}, \tag{11}$$

F denotes the predicate associated with the visualization $\mathcal{V}$ and and X denotes the group-by feature being considered.

C. The VizCertify validation criteria

Consider now a given reference visualization $\mathcal{V}_1$ and a candidate recommendation $\mathcal{V}_2$, both using X as the group-by feature, were the domain of X has K values (i.e., $\mathrm{dom}(X) = \{x_1, \ldots, x_K\}$). With probability $1 - \delta$ the empirical estimates of the normalized columns are accurate within $\bar{\epsilon}$. Let $\bar{\epsilon}_1$ (resp., $\bar{\epsilon}_2$) denote the uncertainty such that with probability of at least $1 - \delta$ we have $\|p_{\mathcal{V}_1}(x_i) - \hat{p}_{\mathcal{V}_1}(x_i)\| \leq \bar{\epsilon}_1$ and $\|p_{\mathcal{V}_2}(x_i) - \hat{p}_{\mathcal{V}_2}(x_i)\| \leq \bar{\epsilon}_2$ according to Lemma 1. Thus, if $|\hat{p}_{\mathcal{V}_1}(x_i) - \hat{p}_{\mathcal{V}_2}(x_i)| > \bar{\epsilon_1} + \bar{\epsilon_2}$ we can conclude that with probability of at least $1 - \delta$ we have $p_{\mathcal{V}_1}(x_i) \neq p_{\mathcal{V}_2}(x_i)$. This leads to:

Theorem 2: For any given $\delta \in (0, 1)$, VizCertify ensures FWER control at level δ while offering visual recommendations if

$$\max_{x_i \in \mathrm{dom}\, X} |\hat{p}_{\mathcal{V}_1}(x_i) - \hat{p}_{\mathcal{V}_2}(x_i)| > \max\{\bar{\epsilon}_1 + \bar{\epsilon_2}, \epsilon_{\mathcal{V}}\}$$

with $\bar{\epsilon}_{1,2}$ denoting the uncertainty given by Corollary 1 and a threshold $\epsilon_{\mathcal{V}} \geq 0$ denoting visual discernability.

Proof: VizCertify recognizes two visualizations to be *statistically different* (and hence, interesting) when the most different pair of corresponding columns differs by no more than the error in its estimates. Due to the uniform convergence bound ensured by the application of VC dimension, the probabilistic guarantees of this control hold simultaneously for all possible pairs of reference and candidate recommendation visualizations. ■

This VC dimension based approach is *agnostic* to the adaptive nature of the testing as it accounts *preemptively* for *all* possible evaluations of pairs of visualizations. Theorem IV-C allows us thus to design the following algorithm for the visual recommendation problem:

Algorithm 1 VizCertify: Recommendations with VC dimension

Input: Starting visualization $\mathcal{V}_1$, query space Q, sample dataset $\mathcal{D}$, FWER target control level $\delta \in (0, 1)$.
Output: A set of statistically safe recommendations Y.

$Y \leftarrow []$ ▷ Empty list of recommendations
$X \leftarrow$ the group-by feature being considered.
$F_{\mathcal{V}_1} \leftarrow$ the predicate associated with $\mathcal{V}_1$.
$\bar{\epsilon}_1 \leftarrow \frac{d+\log_2 \delta^{-1}}{2|D|F_{\mathcal{V}_1}|}$ ▷ Uncertainty in $\mathcal{V}_1$ approx.
for all $\mathcal{V}' \in Q$ **do**
    $F_{\mathcal{V}'} \leftarrow$ the predicate associated with $\mathcal{V}'$.
    $\bar{\epsilon}' \leftarrow \frac{d+\log_2 \delta^{-1}}{2|D|F_{\mathcal{V}'}|}$
    $dist \leftarrow \max_{x_i \in \mathrm{dom}(X)} |\hat{p}_{\mathcal{V}_1}(x_i) - \hat{p}_{\mathcal{V}_2}(x_i)|$
    $interest \leftarrow dist - (\bar{\epsilon}_1 + \bar{\epsilon}_2)$
    if $dist \geq \max\{\bar{\epsilon}_1 + \bar{\epsilon}_2, \epsilon_V\}$ **then**
        $Y.append([\mathcal{V}', interest]$
return sort Y according to interest value (and uncertainty).

Our VizCertify approach can be reworked to resemble a two-sample test, when we assume that there is uncertainty in the reconstruction of both the reference $\mathcal{V}_1$ and the candidate $\mathcal{V}_2$. In some scenarios, the reference visualization may be not have any uncertainty (e.g., when using a flat distribution as reference). In this case, it is sufficient to recommend $\mathcal{V}_2$ if and only if

$$\max_{x_i \in \mathrm{dom}(X)} |\hat{p}_{\mathcal{V}_1}(x_i) - \hat{p}_{\mathcal{V}_2}(x_i)| > \bar{\epsilon}_2.$$

D. VC dimension of the Query Range Space

For practical implementation it is sufficient to bound the VC dimension of the class of queries being considered. Since features are assumed to be equipped with a natural metric, constraints on values of a certain feature can be expressed using operators $\geq, \leq, =$ and $\neq$. This corresponds to selecting intervals (either open or close) of the possible values of a feature. For each feature, these clauses are *connected* by means of logical or $\vee$. We characterize the complexity of such connection by the minimum number of *non-redundant* open and close intervals of the value. In particular we say that a connection of intervals is *non-redundant* is there is no connection of fewer intervals that selects the same values. The VC dimension of this class of queries can then be characterized according to:

Lemma 2: Let $\mathcal{Q}$ denote the class of query functions such that each query is a conjunction of connections of disjunctive clauses on the value of distinct features. The VC dimension of Q is then:

$$VC(Q) = \sum_{i=1}^{m} 2\alpha_i + \beta_i, \tag{12}$$

where α_i (resp., β_i) denotes the maximum number of non-redundant closed (resp., open) intervals of values corresponding to the connection of constraints regarding the value of the i-th feature, for $1 \leq i \leq m$.

Proof: The proof is by induction on i: in the base case we have $i = 1$. In this case, the VC dimension of $\mathcal{Q}$ corresponds to the VC dimension of the union of α_1 closed intervals and β_1 open intervals on the line. By a simple modification of well-known "*folklore*" textbook result according to which the VC dimension of the union of j closed intervals on $\mathbb{R}$ is $2j$, we have that $\mathcal{Q}$ it has VC dimension at most $2\alpha_1 + \beta_1$. Let us now inductively assume that the statement holds for $i > 1$. In order to conclude the proof we shall verify that it holds for $i + 1$ as well.

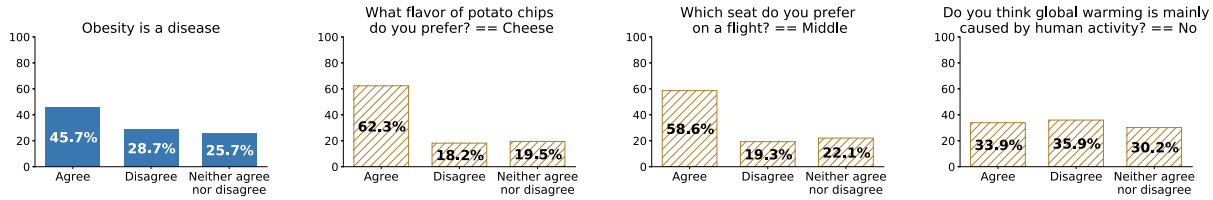
Assume towards contradiction that there exists a set X of $\sum_{j=1}^{i+1} 2\alpha_j + \beta_j$ points that can be shattered by $\mathcal{Q}$. From the inductive hypothesis, we have that for any subset of X with more than $\sum_{j=1}^{i} 2\alpha_j + \beta_j$ cannot be shattered by the family of query functions which can express constraints only on the features $1, 2, \ldots, i$. Without loss of generality let X' denote one of the maximal subsets of X which can be shattered using only the constraints on the features $1, 2, \ldots, i$. Recall that the queries in $\mathcal{Q}$ are constituted by logical conjunctions (i.e., "`and`") of connections (i.e.,"`or`" statement) of constraints on a feature. Hence, for any function in $\mathcal{Q}$ if any of the $i + 1$ connections are such that they assume value "`false`", then the query will not select such point *regardless* of the value of the remaining i connections being conjuncted.

Consider any assignment π of $\{0, 1\}$ to the points in X' and let r_π the range which realizes such shattering.

If r_π would assign to any point in $X \setminus X'$ value "0", then, according to the structure of the queries, no constraint on the $(i + 1)$-th feature would allow to assign to it value "1", and, hence, it would not be possible to shatter X.

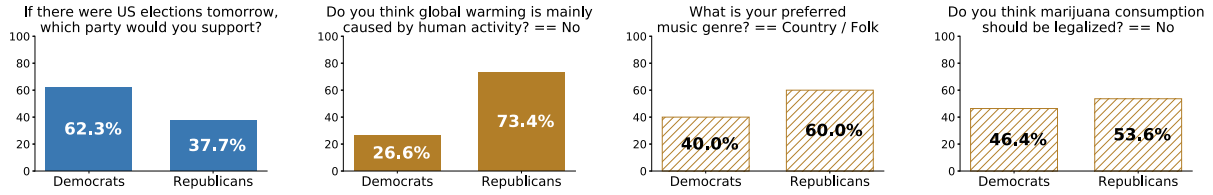
Note that for any assignment π of $\{0, 1\}$ to the points in X', there may may not exists two ranges r_1 and r_2 such that based solely on constraints on the first i features, on would assign "0" to a point in $X \setminus X'$ and the other would assign "1" to the same point. If that would be the case, then i would be possible to shatter $\sum_{j=1}^{i} 2\alpha_j + \beta_j$ points using just constraints on the first i features and this would violate the inductive hypothesis.

Without loss of generality, in the following we can therefore assume that for any assignment π of $\{0, 1\}$ to the points in X' the ranges that realize such assignment just based on the first i features would assign "1" to all the points in $X \setminus X'$. This implies that the shattering of the points in $X \setminus X'$ relies *solely* on the constraints on the values of the $i + 1$-th feature.



(a) Reference View (b) SeeDB Top 1 (c) SeeDB Top 2 (d) SeeDB Top 3

Fig. 2: Example 1: the Top-3 SeeDB recommendations for the reference (a) do not pass the VizCertify control.



(a) Reference View (b) SeeDB Top 1 (c) SeeDB Top 2 (d) SeeDB Top 3

Fig. 3: Example 2: out of the Top-3 SeeDB recommendations for the reference (a) only (b) passes the VizCertify control.

Consider now the points in $X \setminus X'$, according to our assumption $|X \setminus X'| = 2\alpha_{i+1} + \beta_{i+1}$. As discusses in the base of the induction, it is not possible to shatter $2\alpha_{i+1} + \beta_{i+1}$ points using just α_{i+1} (resp., β_{i+1}) closed (resp., open) intervals on the $(i+1)$-th dimension.

Hence, it is not possible to shatter X and we have a contradiction. ■

E. Query complexity vs. minimum selectivity

When exploring the space of possible recommendations by growing a filter condition one clause at a time (i.e., multiple non-trivial clauses are added), with more claused added the number of records selected by the predicate will decrease. Therefore, it is reasonable to start evaluating *simpler* predicate filters first and then proceed *depth-first* by adding more and more clauses. While reasonable, this procedure will likely explore still a large number of queries. However, most of the filters obtained by composing a high number of filters will yield visualizations supported by a few sample points which are intrinsically *unreliable*.

Our VC dimension approach recognize this fact and can be also leveraged to limit the search space. As discussed in Section IV-E, the lower the selectivity γ_F, the higher the uncertainty $\bar{\epsilon}$. From Corollary 1 it follows

$$\bar{\epsilon} \geq \sqrt{\left(d + \log_2 \delta^{-1}\right)(2n\gamma)^{-1}} \tag{13}$$

Since $\bar{\epsilon} \leq d(\mathcal{V}_1, \mathcal{V}_2) \leq 1$, this implies that all visualizations with selectivity

$$\gamma \leq \left(d + \log_2 \delta^{-1}\right)(2n)^{-1} \tag{14}$$

are not going to be recommended as interesting. As a direct consequence, Equation 14 allows to *prune* the search space by eliminating from the exploration queries which are "*not worth to be considered*" possible recommendations.

By taking into consideration the selectivity of a candidate visualizations, our method *automatically adjusts* the threshold of interest for each candidate visualization. From a different perspective this may be also used to limit the structure (i.e., the VC dimension) of all queries being considered when ensuring that each candidate visualization, which differs from the reference by at least θ, is marked as a safe recommendation. When $\gamma_{1,2}$ are the selectivities for the two visualizations, the maximum VC dimension guaranteeing these requirements can be obtained from (13) as:

$$d \leq \theta^2 \min\{\gamma_1, \gamma_2\} n - \log_2(\delta^{-1}).$$

V. Experiments

In this section, we show how our framework can be applied towards both real data (i.e. the collected survey data) and synthetic data.

A. Anecdotal examples

To illustrate how VizCertify restricts the recommended visualizations, we used again the survey data from Figure 1a. Our first example shows that a system without statistical control may lead the user to false insights due to random noise in the sample. Consider, the reference view in Figure 2a, which shows the believe of our participants in "Obesity being a disease" and SeeDB's top recommendations in Figure 2a to 2d, which either emphasize the reference view (more people agreeing or disagreeing). Just looking at the filter conditions, it is rather obvious that all of them should have very little impact on "if people believe in obesity as a result" and VizCertify would not recommend any of them.

In contrast, in our second example we look at Democrats vs Republican supporters (reference view is shown in Figure 3a and the top 3 recommendations in Figure 3b-3d. Here, the top visualization would be recommended (Figure 3b),

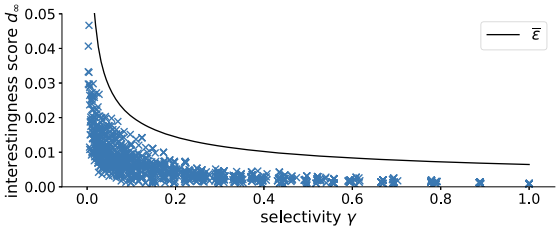


Fig. 4: Blue dots represent the interest scores of all evaluated visualizations. The $\bar{\epsilon}$ curve denotes the threshold for recommendation using VC dimension = 4 to achieve $FWER \leq 0.05$. The lower the VC dimension the more the $\bar{\epsilon}$ curve takes the form of an "L".

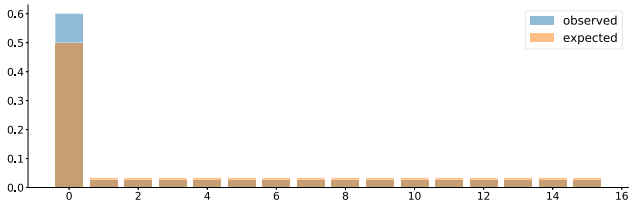


Fig. 5: Comparing two close distributions that however should not be recommended since the visual difference criterion according to the VC dimension approach is not met.

which intuitively makes sense, but the other two visualizations are not recommended by VizCertify. Especially, Figure 3d would probably count as very interesting, as republicans and democrats are equally split in the support of marijuana, but VizCertify considers it as not statistical significant given the survey data size. Thus, if a non-savvy journalist would have used SeeDB over this dataset, VizCertify could have potentially prevented a very questionable news headline.

Note, that it is not the case, that the highest ranked visualizations are necessarily the most statistical significant ones as our leading example in Figure 1a to Figure 1d already demonstrated.

B. Random data leads to no discoveries

A meaningful baseline for any safe visual recommendation system is to make sure that random data does not lead to any recommendations. To demonstrate that the VC approach will not recommend any false positives, we generated a synthetic dataset with uniformly distributed data. $100,000$ samples were generated in total with the first column being selected as aggregate and the other 3 columns as features.

The aggregate is uniformly distributed over $\{1,2,3,4\}$ and each of the 3 features are uniformly distributed over $\{1,...,9\}$. With simple predicates (i.e. a queries formed from $\leq$ clauses solely) there are 1331 visualizations to be explored (a dummy value of $+\infty$ was used in the queries to make a feature active or not. E.g. consider a query of the form $(X_1 \leq 8) \wedge (X_2 \leq +\infty) \wedge (X_3 \leq 3)$. In this query, feature X_2 has no effect on the rows returned since $(X_1 \leq 8) \wedge (X_2 \leq +\infty) \wedge (X_3 \leq 3) \equiv (X_1 \leq 8) \wedge (X_3 \leq 3)$. Note that using $+\infty$-values in the clauses does not change the VC dimension.). As a reference, a uniform distribution over $\{1,2,3,4\}$ was chosen. This means, that the expected support of any visualization is at least $10^5/9^3$ samples which is a fair amount to estimate 4 bars. When not accounting for the multiple comparison problem p-values below the threshold of $\alpha = 0.05$ occur inevitably. A system without FWER guarantees would classify them thus as false positives. Using Bonferroni (or other comparable corrections) remedies at the cost of incurring a noticeable loss in statistical power.

In comparison, the lowest ϵ the VC approach guarantees is $\epsilon_{\min} = 0.0059$. As discussed in IV-B the required threshold $\bar{\epsilon}$ to be met by the Chebychev norm induced distance measure d_∞ depends on the selectivity γ of the query. The necessity of this can be observed in Figure 4 too. With the interestingness scores (distances) being lower than the curve defined by $\bar{\epsilon}$ for all queries in Figure 4 the VC approach does not recommend any false positives in this experiment. Using different distributions instead of the uniform one showed comparable results.

C. Statistical Testing vs. VC approach

While classic statistical testing in the form of a χ^2-test is correct building block for a VRS, in some situations a χ^2-test is unable to spot meaningful visual differences which would however be recognized by our VizCertify approach. Assume we had a query that yielded $m = 1,200$ out of $n = 10,000$ samples and a perfect estimator for the true distribution function of the reference and the query distribution which shall be distributed as in Figure 5. A χ^2-test would yield a p-value of of $2.54 \cdot 10^{-5}$ implying that they are different when no more than 1967 visualizations under Bonferroni's correction are tested. However, at a VC dimension of 10 ($\delta = 0.05$) the required $\bar{\epsilon}$ must be at least 0.22 which is nearly twice as high as the 0.1 difference at the first bar as shown in Figure 5. Thus, the VC approach would not select this visualization as being significantly different enough given the modest sample size. The χ^2 test would recommend this visualization though since it only spots that there is a difference but not whether the difference is significant enough.

Such scenarios occurr in practice especially due to outliers in the data (e.g., for one feature value there are only 1-5 samples that would lead without any correction to a correct recommendation). Though a heuristic may ignore visualizations with less than 5 samples, this would come at the cost of ignoring rare phenomena and while using an arbitrary threshold.

This reinforces that a VRS using the χ^2-test would correctly identify two visualizations being different but can not guarantee a meaningful difference in terms of a distance which is crucial to build usable systems without luring the user into a false sense of security. One may argue that filtering out visualizations after having performed statistical testing would remedy this (which may work in practice when the interest score is high enough), but then there was no guarantee that the distances observed are statistically guaranteed.

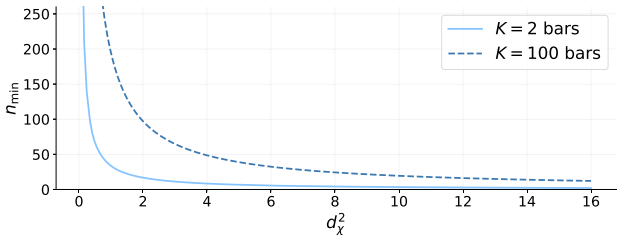


Fig. 6: Chi-square distance d_{χ^2} and minimum number of samples $n_{\min}$ required to reject H_0 with Bonferroni correction, $\alpha' = 0.05$ and 10^6 queries.

Furthermore we want to underscore the point that the Chi-squared test is indeed a very powerful test but that the correct estimation of the distribution dominates the selectivity. I.e., when we guarantee that the estimates for the probability mass function are close enough to the true values, a testing procedure like χ^2-test will even under a million possible hypothesis only need a small number of samples to spot a difference between two distributions. We thereby define the required number of point estimates to be in the range of $2 \leq K \leq 100$ bars as meaningful.

In Figure 6 it is shown that even low values for the χ^2-distance d_χ^2 only require queries with hundreds of samples to be identified correctly.

VI. Related Work & Conclusion

[24] introduced the VC approach to provide ϵ-approximations for the selectivity of queries. Whereas they also consider joins in addition to multi-attribute selection queries, by restricting to AND conjunctions over multiple attributes as used naturally in OLAP we were able to lower the required VC dimension.

Recent work [25] introduced the problem of group-by queries leading to wrong interpretations, specifically in the case when **AVG** aggregates are used. To remedy this, the notion of a biased query is introduced. However, they do not account for the multiple comparison problem and also have no significant distance notion.

[10] introduced various control techniques for interactive data exploration scenarios. Whereas it accounts for the multiple comparison problem, it does not solve the problem of pointing out a statistical different enough distance between two visualizations.

[1] provides an approach to effectively compute visualizations over an exponential search space by using reuse of previous results and approximate queries. Visualizations are recommended by treating group-by results as normalized probability distributions and using various distance measures between two probability distributions to yield a ranking in order to recommend top-k interesting visualizations. The authors found that the actual choice of the distance did not really alter results, which does not come at a great surprise given their relations as pointed out in [26].

Conclusion: In this work, we demonstrated why visual recommendation systems require techniques to prevent users from making false discoveries. We further proposed a novel way to control false discoveries for visual recommendations systems based on VC dimensions.

As described in [27] zooming into particular interesting regions of the data is a key task performed by users in the data exploration setting. Our technique provides a simple and effective methodology which can be applied to a wide range of data. We believe our VC approach can be easily extended to allow for more complicated query types such as these.

Acknowledgment

This work was supported by NSF Award RI-18134446.

References

[1] M. Vartak, S. Madden, A. Parameswaran, and N. Polyzotis, "Seedb: automatically generating query visualizations," *Proceedings of the VLDB Endowment*, vol. 7, no. 13, pp. 1581–1584, 2014.

[2] K. Wongsuphasawat, Z. Qu, D. Moritz, R. Chang, F. Ouk, A. Anand, J. Mackinlay, B. Howe, and J. Heer, "Voyager 2: Augmenting visual analysis with partial view specifications," in *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems*. ACM, 2017, pp. 2648–2659.

[3] J. Seo and B. Shneiderman, "A rank-by-feature framework for interactive exploration of multidimensional data," *Information Visualization*, vol. 4, no. 2, pp. 96–113, Jul. 2005. [Online]. Available: http://dx.doi.org/10.1057/palgrave.ivs.9500091

[4] J. Mackinlay, P. Hanrahan, and C. Stolte, "Show me: Automatic presentation for visual analysis," *IEEE Transactions on Visualization and Computer Graphics*, vol. 13, no. 6, pp. 1137–1144, Nov 2007.

[5] H. Ehsan, M. A. Sharaf, and P. K. Chrysanthis, "Muve: Efficient multi-objective view recommendation for visual data exploration," in *2016 IEEE 32nd International Conference on Data Engineering (ICDE)*, May 2016, pp. 731–742.

[6] A. Key, B. Howe, D. Perry, and C. Aragon, "Vizdeck: self-organizing dashboards for visual analytics," in *Proceedings of the 2012 ACM SIGMOD International Conference on Management of Data*. ACM, 2012, pp. 681–684.

[7] X. Qin, Y. Luo, N. Tang, and G. Li, "Deepeye: An automatic big data visualization framework," *Big Data Mining and Analytics*, vol. 1, no. 1, pp. 75–82, March 2018.

[8] D. Moritz, C. Wang, G. L. Nelson, A. H. Lin, M. Smith, B. Howe, and J. Heer, "Formalizing visualization design knowledge as constraints: Actionable and extensible models in draco," in *IEEE Trans. Visualization and Comp. Graphics (Proc. InfoVis)*, 2018.

[9] E. Zgraggen, Z. Zhao, R. C. Zeleznik, and T. Kraska, "Investigating the effect of the multiple comparisons problem in visual analysis," in *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems, CHI 2018, Montreal, QC, Canada, April 21-26, 2018*, 2018, p. 479. [Online]. Available: http://doi.acm.org/10.1145/3173574.3174053

[10] C. Binnig, L. D. Stefani, T. Kraska, E. Upfal, E. Zgraggen, and Z. Zhao, "Toward sustainable insights, or why polygamy is bad for you," in *CIDR 2017, 8th Biennial Conference on Innovative Data Systems Research, Chaminade, CA, USA, January 8-11, 2017, Online Proceedings*, 2017. [Online]. Available: http://cidrdb.org/cidr2017/papers/p56-binnig-cidr17.pdf

[11] Z. Zhao, L. D. Stefani, E. Zgraggen, C. Binnig, E. Upfal, and T. Kraska, "Controlling false discoveries during interactive data exploration," in *Proceedings of the 2017 ACM International Conference on Management of Data, SIGMOD Conference 2017, Chicago, IL, USA, May 14-19, 2017*, 2017, pp. 527–540. [Online]. Available: http://doi.acm.org/10.1145/3035918.3064019

[12] F. Chirigati, H. Doraiswamy, T. Damoulas, and J. Freire, "Data polygamy: The many-many relationships among urban spatio-temporal data sets," in *Proceedings of the 2016 International Conference on Management of Data*, ser. SIGMOD '16, New York, NY, USA, 2016, pp. 1011–1025. [Online]. Available: http://doi.acm.org/10.1145/2882903.2915245

[13] M. K. Stern and J. H. Johnson, "Just noticeable difference," *The Corsini Encyclopedia of Psychology*, pp. 1–2, 2010.
[14] D. Russo and J. Zou, "How much does your data exploration overfit? controlling bias via information usage," *arXiv preprint arXiv:1511.05219*, 2015.
[15] J. Taylor and R. J. Tibshirani, "Statistical learning and selective inference," *Proceedings of the National Academy of Sciences*, vol. 112, no. 25, pp. 7629–7634, 2015.
[16] A. Kim, E. Blais, A. Parameswaran, P. Indyk, S. Madden, and R. Rubinfeld, "Rapid sampling for visualizations with ordering guarantees," *Proceedings of the VLDB Endowment*, vol. 8, no. 5, pp. 521–532, 2015.
[17] S. Chaudhuri, R. Motwani, and V. Narasayya, "Random sampling for histogram construction: How much is enough?" in *Proceedings of the 1998 ACM SIGMOD International Conference on Management of Data*, ser. SIGMOD '98. New York, NY, USA: ACM, 1998, pp. 436–447. [Online]. Available: http://doi.acm.org/10.1145/276304.276343
[18] M. Mitzenmacher and E. Upfal, *Probability and Computing: Randomized Algorithms and Probabilistic Analysis*. New York, NY, USA: Cambridge University Press, 2005.
[19] P. M. Lukacs, K. P. Burnham, and D. R. Anderson, "Model selection bias and freedman's paradox," *Annals of the Institute of Statistical Mathematics*, vol. 62, no. 1, p. 117, May 2009. [Online]. Available: https://doi.org/10.1007/s10463-009-0234-4
[20] C. Dwork, V. Feldman, M. Hardt, T. Pitassi, O. Reingold, and A. Roth, "The reusable holdout: Preserving validity in adaptive data analysis," *Science*, vol. 349, no. 6248, pp. 636–638, 2015.
[21] V. N. Vapnik and A. Y. Chervonenkis, "On the uniform convergence of relative frequencies of events to their probabilities," in *Measures of complexity*. Springer, 2015, pp. 11–30.
[22] S. Har-Peled and M. Sharir, "Relative (p, ε)-approximations in geometry," *Discrete & Computational Geometry*, vol. 45, no. 3, pp. 462–496, 2011.
[23] M. Löffler and J. M. Phillips, "Shape fitting on point sets with probability distributions," *CoRR*, vol. abs/0812.2967, 2008. [Online]. Available: http://arxiv.org/abs/0812.2967
[24] M. Riondato, M. Akdere, U. Çetintemel, S. B. Zdonik, and E. Upfal, "The vc-dimension of queries and selectivity estimation through sampling," *CoRR*, vol. abs/1101.5805, 2011. [Online]. Available: http://arxiv.org/abs/1101.5805
[25] B. Salimi, J. Gehrke, and D. Suciu, "Hypdb: Detect, explain and resolve bias in olap," *arXiv preprint arXiv:1803.04562*, 2018.
[26] A. L. Gibbs and F. E. Su, "On choosing and bounding probability metrics," *International statistical review*, vol. 70, no. 3, pp. 419–435, 2002.
[27] B. Shneiderman, "The eyes have it: A task by data type taxonomy for information visualizations," in *Visual Languages, 1996. Proceedings., IEEE Symposium on*. IEEE, 1996, pp. 336–343.