Visualization Model Validation via Inline Replication

Information Visualization
XX(X):1–29

DOI: 10.1177/ToBeAssigned
www.sagepub.com/

David Gotz[1], Wenyuan Wang[1], Annie T. Chen[2], and David Borland[3]

Abstract

Data visualizations typically show a representation of a dataset with little to no focus on the repeatability or generalizability of the displayed trends and patterns. However, insights gleaned from these visualizations are often used as the basis for decisions about future events. Visualizations of retrospective data therefore often serve as "visual predictive models." However, this visual predictive model approach can lead to invalid inferences. In this paper, we describe an approach to visual model validation called *Inline Replication* (IR). IR is closely related to the statistical techniques of bootstrap sampling and cross-validation and, like those methods, provides a nonparametric and broadly applicable technique for assessing the variance of findings from visualizations. This paper describes the overall IR process and outlines how it can be integrated into both traditional and emerging "big data" visualization pipelines. It also provides examples of how IR can be integrated into common visualization techniques such as bar charts and linear regression lines. Results from an empirical evaluation of the technique and two prototype IR-based visual analysis systems are also described. The empirical evaluation demonstrates the impact of IR under different conditions, showing that both (1) the level of partitioning, and (2) the approach to aggregation have a major influence over its behavior. The results highlight the tradeoffs in choosing IR parameters, but suggest that using $n = 5$ partitions is a reasonable default.

Keywords

Visual Analytics, Information Visualization, Replication, Validation, Prediction

[1]University of North Carolina, Chapel Hill, NC USA
[2]University of Washington, Seattle, WA USA
[3]RENCI, University of North Carolina, Chapel Hill, NC USA

Corresponding author:
David Gotz, School of Information and Library Science, University of North Carolina at Chapel Hill Manning Hall, 216 Lenoir Drive Chapel Hill, NC 27599 USA.
Email: gotz@unc.edu

Prepared using sagej.cls *[Version: 2015/06/09 v1.01]*

Introduction

Visualization tools are most often designed to depict the entirety of a dataset–subject to a set of filters applied to focus the analysis–as accurately as possible. In this typical pattern, the goal is to provide the user with an accurate understanding of *all of the data* in the underlying dataset that matches the active set of filters. This ethos was captured, perhaps most famously, in Shneiderman's Visual Information Seeking Mantra: *overview first, zoom and filter, then details-on-demand.*[1] Variations of this basic approach have since been adopted in most modern visualization systems.

The foundations for these systems are visual mappings that specify a graphical representation for the underlying data. For small and low-dimensional datasets, these mappings can be direct (e.g., a scatter plot for a small 2D dataset). As problems grow in data size or dimensionality, algorithmic data transformation methods can be used to filter, manipulate, and summarize raw data into a more easily visualized form.

On top of these mappings, interactive controls are often provided to give users even more flexibility to filter or zoom to specific subsets of data. These interactions can be linked to more detailed information about data objects, such as via levels-of-detail or multiple coordinated views. The result, when well designed, is an effective visual interface for data exploration and insight discovery.

For this reason, these steps form the core stages of the canonical visualization pipeline.[2;3] This approach can be enormously informative, and has led to advancements in how people seek to understand information across a wide range of domains, such as helping computer users navigate through large file systems, visualizing medical records to help doctors understand patient histories, and visualizing weather data to identify regions most impacted by a recent storm.

Critically, these visualization use cases are all *retrospective* in nature. More specifically, they employ visualizations that attempt to faithfully report data exactly as it was observed. Users aim to see an overview of the entirety of a given dataset. If a user applies constraints to focus the visual investigation (e.g., via zoom and filter), the visualization is expected to show the full set of data that satisfies the applied constraints.

In many visualization scenarios, however, users are in fact more interested in conducting *prospective analysis*: using visualizations of these historical data to reason about future or not-yet-observed data. For example, medical experts examining data for a cohort of patients might be most interested in what treatments would work best for a future patient with similar characteristics. Visualizations of historical sports statistics are often used to inform strategic decisions that are used in upcoming competitions. Financial visualization tools are often used to inform future investment decisions. In each of these use cases, visualizations of historical data are used to inform future decisions. For such prospective analysis tasks, these retrospective visual depictions of the data are often used, in essence, as naïve *visual predictive models*, with the assumption that visualizations of historical data can be used to infer or predict future observations. In many ways, these visual models are analogous to statistical models developed during statistical analyses, from which users often also attempt to predict future observations.

In many cases retrospective representations are indeed very informative. However, just like the underlying descriptive statistics that such visualizations often depict,[4] traditional retrospective visualizations often provide insufficient evidence for making predictive inferences, even as the visual depiction itself might be especially suggestive for making such inferences. In many cases, the trends and patterns that a visualization of retrospective data presents to a user may be artifacts of noise and expected randomness within the underlying data. For users to make valid inferences or predictions based on historical data, a more nuanced understanding of the data being visualized is required.

This critical gap between (a) retrospective visualization designs and (b) the predictive requirements of many users has been recognized within the visualization community.[5] Some have attempted to bridge this gap by adding support for inferential statistics within the visualization. Typically, this approach combines carefully designed statistical models with visualizations of the model's results. For example, visualizations can be instrumented to estimate and display uncertainty, confidence intervals, or statistical significance. Alternatively, predictive modeling methods can be used to generate additional data, with the predictions themselves being incorporated into the visualization. These systems go beyond traditional descriptive reporting, but typically require a careful and sometimes onerous focus on modeling, including estimating underlying statistical distributions, which is incompatible with many applications for which a more accurate assessment of the repeatability of a given visualization would be useful.

This paper presents *Inline Replication* (IR), an alternative approach to enabling inferential interpretation that is designed to overcome the above challenges. We introduce a partition function within the visualization pipeline to produce multiple *folds* for each visualized data subset. Metric functions are applied to each fold, and an aggregation function combines the individual measures prior to visualization. Interaction techniques enable users to examine both aggregate and individual fold metric values.

Our method is motivated by the bootstrap sampling and cross-validation techniques used widely in the statistics and machine learning communities. The IR approach is nonparametric, making it easy to apply and use generically within a visualization system without arduous modeling or assumptions about distribution parameters. IR integrates easily with the standard visualization pipeline, and is also ideal for use in large-scale visualization systems where progressive or sample-based approaches are required. Finally, our method provides users with validation information that is both intuitive and easy to interpret.*

The remainder of this paper is organized as follows. It begins with a review of related work, then describes the details of the IR methodology. After that we present an experiment that aims to test the accuracy of such method against traditional statistical analysis. We then share example results from a variety of proof-of-concept systems that include the IR technique. These examples range from simple bar charts to more sophisticated interactive visualizations of large-scale event data collections.[7] The paper concludes with a discussion of limitations and outlines key areas for future work.

Related Work

The IR approach to visual model validation is informed by advances in several different areas of research. These include the topics of uncertainty, predictive visualization, and progressive or incremental visualization. Also relevant are visualization systems that utilize inferential statistics methods and conceptual models of the visualization pipeline.

Visualization of Uncertainty

The visualization of uncertainty has been an active research area within the visualization community for many years. Studies have explored the problem from many perspectives, including developing taxonomies that have examined types of uncertainty[8] as well as visualization methods for conveying

*This article includes material described in a pre-print that has released by the authors via arXiv[6]. This version of the article is heavily revised and contains an entirely new evaluation section describing and analyzing results from a set of empirical experiments to test the behavior of IR under various conditions.

uncertainty.[9] In addition, there have been many efforts to formally study alternative methods for depicting uncertainty measures[10–12] through user studies that explore the perceptual understanding of various uncertainty representations. However, these studies focus on the visual representation rather than methods for determining the degree of uncertainty.

Perhaps most relevant to the IR approach proposed in this paper is work that has focused on estimating uncertainty via measures of entropy within a dataset rather than by using carefully constructed statistical models.[13] Compatible with IR, this work proposes using entropy as a non-parametric measure of uncertainty for categorical data which does not require formal modeling nor make assumptions about specific distributions within the data. IR provides a broader and more general framework for this approach, within which the entropy metric could be easily adopted.

In other work, the distinction between the "visualization of uncertainty" and "the uncertainty of visualization" has been highlighted.[14] The latter is a related but separate concept from traditional uncertainty visualization. Such work highlights that the rendered graphics of a visualization can convey a sense of authority which may not be warranted, even when the underlying data itself is considered to be beyond reproach. This challenge is a key motivation for IR, especially with respect to the confidence of the user in making predictions based on this unwarranted authority, as outlined in the discussion presented in the Section "Visualization as a Predictive Model."

Finally, we distinguish IR's focus on variation of data over partitions to challenges related to non-representative samples. Non-representiveness and selection bias are important threats to visualization validity,[15] and recent research has proposed methods to address these challenges within the context of interactive exploratory visualization.[16;17] IR can complement these methods, but focuses on issues of variation rather than representativeness.

Predictive Visual Analytics

Visualization has long been used to support predictive analysis tasks. However, most often, the "prediction" is performed by users reviewing historical data and making assumptions about what might happen in the future for similar situations. In fact, the relatively limited history of work on visualizations that incorporate more formal predictive modeling methods was the topic for a workshop at a recent IEEE VIS Conference.[5]

The work that does exist in this area has often focused on model development and evaluation rather than supporting end users' predictive analysis tasks. For example, BaobabView[18] supported interactive construction and evaluation of decision trees. More recent work has focused on building and evaluating regression models.[19] This method, like ours, adopts a partition-based approach to avoid making structural assumptions about the data. However, the focus on building regression models leads to an overall workflow that is very different from the proposed IR approach.

Others have focused on visualizing the output produced by predictive models. For example, Gosink et al. have visualized prediction uncertainty based on formalized ensembles of multiple predictors.[20] This approach, however, requires careful modeling to develop the predictors, including the specification of priors that enables the Bayesian method that they propose.

Outside the visualization literature, where novel visual or interaction methods are not a concern, predictive features are typically visualized using traditional statistical graphics, for example, systems that visually prioritize and threshold p-values to rank features for prediction (e.g., Sipes et al.[21]). Such methods are fully compatible with the IR process proposed in this paper.

Progressive/Incremental Visualization

Model overfitting and other sampling challenges are common to "Big Data" visualizations that rely on progressive or incremental techniques (e.g., Fisher et al.[22] and Stolper et al.[23]). Initial samples are small, grow over time, and can change in distribution as time proceeds. Some have addressed this challenge by including confidence intervals along with partial sets of query results.[24] However, relying on the query platform to assess confidence in data subsets does not easily support interactive zoom and filter operations after the query, because these changes in visual focus do not necessarily result in new queries that generate new result sets. Moreover, these papers do not propose methods for computing confidence intervals, but rather, assume that such data will be provided by the the database.

Inferential Statistics

Statistical inference is a discipline with a very long and distinguished history. Most relevant to the IR method described in this paper are challenges related to statistical significance and null hypotheses, and in particular Type 1 and Type 2 errors. Type 1 errors refer to improper rejections of the null hypothesis which lead to conclusions that are not real effects, while Type 2 errors refer to falsely retaining the null hypothesis which can lead to assumptions that a true effect is false.[25]

These types of errors are of critical concern in high-dimensional exploratory visualization where computational methods can quickly assess vast numbers of dimensions for statistical significance. Statistical correction methods have been proposed to reduce Type 1 errors,[26] but arguments have also been made against such approaches. Those arguments suggest that parameterized models or assumptions of "default" null hypotheses don't match real world situations where distributions are rarely straightforward or independent. Suggesting that these correction methods are the wrong approach for exploratory work, Rothman argues that "scientists should not be so reluctant to explore leads that may turn out to be wrong that they penalize themselves by missing possibly important findings."[27]

This tension is present in many interactive exploratory systems which make it easy to generate vast numbers of potential hypotheses. As a result, a wide range of methods have been proposed for modeling measures of confidence or significance.[28–30] These efforts, however, typically rely on formal statistical methods that make assumptions about distributions and variable independence. For example, confidence intervals have some conceptual similarities to IR. However, calculating a confidence interval requires assumptions about underlying distributions of the data and knowledge of key distribution parameters such as mean and variance. Such approaches are problematic for exploratory visualizations which enable users to rapidly apply filters or constraints that can quickly change the underlying assumptions. The IR method we propose enables users to visually assess the reliability of hypotheses, providing a high degree of flexibility. Similar approaches that rely on user judgment have been shown to be quite effective.[31;32]

Models of the Visualization Pipeline

The traditional visualization pipeline model describes the process of transforming raw data first to an analytical abstraction, then to a visualization abstraction, and finally to a rendered graphic for interaction.[2;3] We add partitioning and aggregation stages to support the IR approach. As we will describe, a special case of the IR model (with just one partition) is equivalent to the traditional model. By extending the canonical pipeline, our work has similarities with Correa et al.'s paper describing pipeline extensions for an uncertainty framework focused on the data transformation process.[33] However, unlike

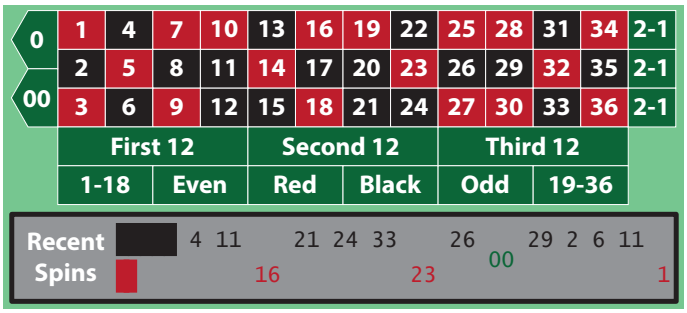


Figure 1. Roulette wheels allow users to bet on "black" or "red" squares. Casinos often display a simple visualization of "recent spins" to provide gamblers with a false sense of predictive knowledge. In this example, the display shows a recent preponderance of black numbers with the implication to gamblers that this may influence future spins of the wheel.

Correa at al., IR proposes a different set of extensions which adopt an empirical approach rather than assuming a Gaussian or any other formal error model.

Visualization as a Predictive Model

Visualization design is often conceptualized as a mechanism for reporting. This retrospective approach is so ubiquitous that terms such as *prediction*, *forecast*, and *inference* cannot be found within the indices of many leading visualization texts from the past 25 years.[34–36]

Many visualization *consumers*, however, use graphical representations of historical data as the basis for decisions about future performance. This is done even when the underlying data and transformations do not support such prospective conclusions. Despite potentially fatal flaws in terms of generalizability and repeatability, retrospective visualizations are in essence being used as predictive models.

The tendency to assume predictive power in visualization can be seen, for example, in modern casinos. Roulette wheels commonly include an electronic display[37] which shows the table's recent history. Assuming a fair table, "red" and "black" numbers should be equally likely to appear. However, as illustrated in Figure 1, the history provided to gamblers is not sufficiently long to learn if the table has any systemic bias.

Why then is the gambler presented with a simple visualization of the history? The data is visualized to provide gamblers with a false sense of knowledge; to suggest to a hesitant gambler that a bet is an informed decision rather than a random choice. A gambler may infer that the recent streak of black suggests more black spins will soon appear. Alternatively, the gambler may infer an imminent return to red. To the casino, it does not matter what predictive inference is drawn as long as it provides a false confidence that leads to increased betting.

It is tempting to dismiss this scenario as one in which the gambler should be more informed about basic statistics. The small sample size and the independence of each roulette spin should make it clear that the display is not especially informative. However, relatively sophisticated users performing visual analysis of data from more complex systems can make similarly poor predictive assessments on the basis of visual representations that don't properly convey the underlying limits of their predictive power.

For example, consider a business analyst attempting to learn about why sales are declining, or a physician using historical patient data to compare treatment efficacy. In these complex real-world cases, in which it is essentially impossible to fully understand the underlying statistical processes, it is natural for analysts to turn to visualization as a predictive model for their problem. Visualization enables these users to see what has happened and, based on trends or patterns in the visual representation, make assumptions about what will happen in the future.

However, just as the casino gambler draws inference from a not-so-meaningful visualization, these power users can be led to make poor predictions on the basis of visualizations that are essentially "overfit" models based on poor visual representations of the underlying process. This problem has even been documented in highly quantitative fields such as epidemiology, where public health analysts have had trouble discounting statistics from small sample sizes when visualized.[38]

Issues of poor sampling and overfitting are especially problematic during exploratory visualization in which users can interactively apply arbitrary combinations of filters to produce new ad hoc subsets of data for visualization. Such systems are at greater risk of generating misleading visualizations that occur "by chance" rather than due to real properties of the underlying problem.[39] The same is true for visualization systems that use sampled or progressive queries to address issues of scale.

The potential for this sort of "visual model overfitting" is analogous to the overfitting problem in more traditional modeling tasks. In the machine learning community, this is addressed in part by cross-validation, a widely used technique for assessing the quality and generalizability of a model.[40] Rather than relying on a single model solution, cross-validation methods create and compare multiple solutions, one for each of several partitions of a dataset (often called "folds"). This enables an assessment of model repeatability, with models that work consistently across partitions considered more trustworthy. Similarly, bootstrap sampling techniques in statistics[41] produce multiple samples from a single set of observations in order to derive better estimates of the original sample's statistical properties.

If one considers—as we argue here—that a visualization is often used as a form of predictive model, then validation becomes a critical guard against problems associated with visual model overfitting. When a visualization is zoomed and filtered to focus on a specific subset, is the visual representation repeatable, or is it due to chance variation? Are the conclusions drawn from the visualization generalizable? The IR method outlined in the next section seeks to help answer such questions by embedding an approach similar to cross-validation and bootstrap sampling within the visualization pipeline so that each new view produced during user interaction can be evaluated for validity.

Inline Replication

Inline Replication (IR) is an approach to visualization in which the dataset associated with each visualized measurement is partitioned into multiple subsets, or *folds*, processed independently to calculate derived statistics or metrics, then aggregated back together to be rendered in a visualization. This partitioned approach embeds an automated and non-parametric workflow for data replication within the visualization pipeline, as illustrated in Figure 2. The result is that visualizations based on IR can provide users with important information about the repeatability of observed visual trends, reducing the likelihood of certain types of erroneous conclusions.

The IR pipeline begins with the same initial step as a traditional visualization pipeline. A set of query or filter constraints is first applied to a primary data source D to produce a focused dataset $d \subset D$. The data

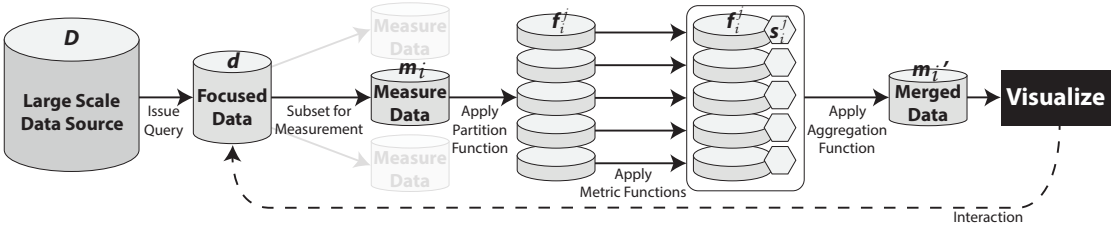


Figure 2. The Inline Replication (IR) visualization pipeline sends each derived measure's subset of data (m_i) through a partition function to create multiple folds (f_j) prior to mapping and visualization. A metric function is applied to each fold independently, and an aggregation function recombines the folds to form an aggregate measure (m_i') for subsequent visualization and interaction.

in d is then organized into subsets for which statistical measurements are calculated, creating measure-specific data subsets, m_i. For example, a visualization pipeline configured to generate the inset bar chart in Figure 1, showing the distribution between black and red spins for a roulette wheel, would include the subset m_{recent} containing data for the most recent spin results (black or red). If the visualization included multiple bar charts (e.g., past 10 spins, past 100 spins, and past 1,000 spins) then multiple subsets m_i would be defined because each requires the calculation of a distinct set of measurements.

Traditionally, the data for each m_i would immediately be processed to compute the measurements required for visualization (e.g., the fraction of spins resulting in black, and the fraction of spins resulting in red). Those measures would then be mapped to visual properties of objects within the visualization (e.g., the size of each bar in the bar chart).

The IR pipeline, however, behaves differently. Each m_i is first partitioned into a set of distinct folds, F_i, where each fold $f_i^j \in F_i$ is analyzed independently via a metric function. The results are then aggregated to form a merged dataset m_i'. It is this merged representation of the measures that is mapped to the visual representation and rendered to the screen for interaction using methods designed to convey the repeatability of the visual model across each of the folds.

This section provides an overview of the IR pipeline, focusing on the three functions at the core of the design: the *partition* function, the *metric* function, and the *aggregation* function. It then describes the IR approach to visual display and interaction, and concludes with a discussion of useful variations to the core design.

Partition Function

Conceptually, the partition function is designed to subdivide the data in a given measure-specific subset m_i into multiple independent partitions, or folds. These folds are used as the basis for calculating each measurement. Later in the IR process, derived measures (e.g., proportions or statistical significance) will be calculated for each fold.

Formally, we define the $partition$ function as an operator that subdivides a measure-specific set of data m_i into n folds such that each fold $f_i^j \subset m_i$.

$$Partition(m_i, n) \rightarrow \{f_i^0, f_i^1, \cdots, f_i^j, \cdots, f_i^{n-1}\} \quad (1)$$

This function is applied to the raw data in m_i, prior to any other aggregating transformations (such as the summation in the roulette example). Following an approach inspired by k-fold cross-validation,[40] the baseline partition function creates n folds that are disjoint, approximately equal in size, and randomly partitioned such that:

$$\bigcup_{j=0}^{n-1} f_i^j = m_i \tag{2}$$

As discussed previously, multiple folds are created with the goal of supporting repeated calculations for each measure. Increasing the value of n to produce more folds increases the replication factor. However, higher n values also produce smaller $|f_i^j|$. If n is too large for a given m_i, the folds may be too small to compute useful measures. Therefore, n can be dynamically determined so as to require a minimum fold size. If m_i represents a "large enough" subset of data, it will produce a full set of folds. If, however, m_i is too small for the minimum fold size, fewer than n folds will be produced. The threshold for "large enough" depends on many factors, including the specific metrics that will be calculated.

Partitioning with $n = 1$ results in the *identity partition function* where $f_i^0 = m_i$ regardless of the size of m_i. Because no partitioning is performed, an IR process using the identity partition function produces results that are identical to a traditional visualization pipeline: a single metric is calculated and visualized. In this way, the traditional approach to visualization can be seen as a special case of the IR process containing only one fold.

Choosing an appropriate n is necessarily a compromise between increased replication and smaller sample size. We can look to the machine learning community for guidance, however, where empirical studies have shown that there is no meaningful benefit for values of n over 10.[40] Moreover, as datasets grow larger in many fields, smaller sample size becomes less of a concern.

Finally, there are certain conditions (e.g., very small datasets with little data to partition, or very large datasets where sampled queries are required) where the basic formulation for the partition function can be problematic. Variations to the partitioning process, designed to help address these challenges of scale, are discussed in Section "Variations."

To illustrate the partitioning process, consider the roulette example from earlier in this paper. The example bar chart showing the fraction of spins resulting in black or red is based on a single measure-specific subset of data m_{recent}. The $partition$ function would be applied to this subset to create a set of multiple folds, each of which would contain a subset of the recent spin results. For example, $Partition(m_{recent}, 5)$ would produce a set of five folds, each containing results from roughly one-fifth of the overall set of recent spins.

Metric Function

The folds produced during partitioning are sent to a *metric function* which is applied independently to each fold as illustrated in Figure 2. The metric function computes derived statistics s_i^j for each fold f_i^j. These derived statistics (one s_i^j for each of the n folds f_i^j) can then be aggregated and compared during the visualization rendering process.

The specific measures computed by the metric function are application specific, but could range from simple descriptive statistics (e.g., sums, averages) to more complex analyses (e.g., classification, regression). Generally speaking, metric functions produce the same derived values that would normally

be computed as part of a more traditional visualization process. The key difference in IR is that the metrics are computed n times for each m_i (once for each fold), where traditionally such values would be computed just once. We also note that multiple statistics might be computed for each fold. For example, a system might compute both the average and a linear regression for each fold.

For example, consider the roulette use case described earlier. The metric function in this example would compute the fraction of spins resulting in black and red in each fold f_i^j. This fraction is the same measure that the original bar chart is designed to display. However, with the IR approach, the metric is calculated for each of the five folds produced by $Partion(m_{black}, 5)$.

An actual implementation of IR using a similar "fraction of the population" metric function is discussed in Section "Use Cases." However, more sophisticated systems may adopt more advanced measures. For example, correlation statistics, p-values, metrics of model "fit", and regression lines are all compatible with the IR approach. Examples of IR using linear regression, correlation, and statistical significance testing are also described in Section "Use Cases."

Aggregation Function

The metric function produces a set of statistical measures s_i^j corresponding to the set of n folds f_i^j that are produced by the partition function. These multiple measures could, in principle, be visualized directly. However, this would introduce complexity to the visual representation. Therefore, prior to visualization, the multiple s_i^j metrics can be aggregated to a single representation m_i', inverting the partition process to produce a simple aggregate summary. We note that the detailed s_i^j metric values are retained in this process and can be made visible through the visualization via *unfolding* as described in the next section. As illustrated in Figure 2, the values are combined via an *aggregation function* which we define as follows.

$$Aggregate(s_i^j) \rightarrow m_i' \tag{3}$$

A variety of aggregation algorithms can be employed, with different approaches appropriate to different types of metrics. For example, for count-based metrics which capture the frequency of data items in each fold, a summation across all folds might be the most appropriate because a sum of counts for each fold provides an accurate total for the overall data subset m_i. For a metric that captures a mean or rate, averaging the values across all folds may be most appropriate. For categorical metrics, meanwhile, such as those produced by classification algorithms, a "majority vote" aggregation method[42] can be applied to capture the most frequently assigned category. The same voting approach can be used when aggregating thresholded measures (e.g., tests of statistical significance) across all folds. This approach is demonstrated in Section "DecisionFlow2."

The aggregation function produces a single aggregate measure s_i which is derived from the set of fold-specific measures s_i^j. These are combined to form a merged data representation $m_i' = (s_i, \{s_i^j\})$ which is then used in the mapping and rendering process of the final visualization.

As a concrete example, consider again the roulette scenario. The metric function described previously computed the fraction of spins resulting in black and red numbers for each of the five folds created by $Partition(m_{recent}, 5)$. The aggregate rates for both colors are obtained by averaging the five fold-specific rates. The overall average, along with the individual values for each fold, are combined to form m'_{recent}.

Visual Display And Unfolding of Partition Data

Once aggregation has been performed, the merged data m'_i are mapped to their corresponding visual marks and displayed as part of the visualization, as shown in the final step in Figure 2. The IR approach to visualizing m'_i has two components, which correspond to the two distinct types of information in the merged data structure: (a) the aggregate statistical measure, s_i, and (b) the individual fold measures, $\{s_i^j\}$.

First, an initial visualization is created using only the aggregate measures. The process for this stage is similar to a traditional visualization pipeline. The aggregate measures are mapped to visual properties of the corresponding graphical marks, which we call *aggregate marks*. These marks are then rendered to the screen for display and interaction. In the roulette example, for instance, the aggregate data for black and red spin rates (produced by the $Aggregate$ function) can be used to generate a basic bar chart that is identical to what is shown in Figure 1.

Second, an IR visualization enables aggregate marks to be *unfolded*. An unfolding operation—typically triggered by a user interaction event such as selection or brushing—augments the aggregate marks with a visualization of the individual fold statistics that contribute to the aggregate measures. In the ongoing roulette example, the fold data would show the variation in proportion of spins that result in black and red numbers across each of the $n = 5$ independent folds. Additional examples from our experimental prototypes are described in Section "Use Cases."

Discussion

The ability to unfold aggregate measures into repeated measurements is a central contribution of the IR approach. By graphically depicting the repeatability of a particular measure across multiple folds, IR provides users with important and easy-to-interpret cues as to the variability of a given measure. Traditional visualization methods do not convey this information, meaning it is often not considered when predictive conclusions are made by users.

Another benefit of IR comes from the aggregation function. In particular, embedding within the visualization pipeline an ability to aggregate categorical values such as statistical significance classification can lead to more accurate results. Repeated measures combined with voting-based aggregation can, for instance, reduce the exposure to Type 1 errors when looking for statistically significant p-values. For example, a statistically significant ($p < 0.05$) run of black spins on the roulette wheel is less likely to occur "by chance" across a majority of n folds than it is across a single group of spins. This is a major benefit for exploratory visualization techniques that enable users to visually "mine" through large numbers of variables searching for meaningful correlations. The results of an experiment exploring how IR can help reduce Type 1 errors with respect to correlation are presented in Section "Empirical Experiments."

Variations

Following the traditional approach to k-fold cross-validation, the baseline $Partition$ function defined in Section "Partition Function" specifies that the constructed folds are disjoint, randomly partitioned, and exhaustive (Equation 2). However, relaxing these constraints leads to several valuable variations to the baseline IR procedure.

Partial Partitioning. Relaxing the requirement of Equation 2 enables the creation of partitions that do not contain all data points within m_i. For very large datasets, this can enable approximate analyses

that use only a subset of the available data. This approach provides significant performance benefits for metric functions that have poor scaling properties, and enables IR to work directly with recently proposed techniques for progressive visualization.[22;23]

Partitioning With Replacement. Relaxing the requirement that all folds are disjoint enables partitioning with replacement. Similar to bootstrap resampling,[41] this approach enables the same data point to be included in multiple folds (or even multiple times within the same fold). When using replacement, the dataset in m_i becomes a sample distribution from which the partitioning algorithm can generate a larger population. This is especially useful for small datasets—a frequent occurrence in exploratory visualization in which multiple filters can be quickly applied—because the larger generated population can enable the IR process to run with less concern about producing fold sizes that are too small.

Incremental Partitioning. A number of progressive or sampled methods have been proposed in recent years to address the challenges of "Big Data" visualization.[23;24] In these approaches, the full dataset m_i is often never fully retrieved. To utilize an IR approach in these cases, an incremental partitioning process is needed. During this process, data points should be distributed to folds as they are retrieved such that all n folds are kept roughly equal in size. This will enable IR to work with continuing improvement in metric quality as more data arrives. However, we note that IR will not overcome selection bias that may be introduced as part of the progressive query process. Therefore, the determination of a progressive sampling order that is both representative and balanced remains a critical concern.

Empirical Experiments

A series of simulated data experiments were conducted to empirically measure the behavior of IR in conditions where–unlike many real-world situations–accurate ground truth information about data distributions was available. The experiments focused on evaluating the impact of IR with respect to Type 1 and Type 2 errors for a common type of analysis: correlations between two datasets.

Data Generation

To simulate empirical experiments under varying conditions, two different types of base data distributions were created as illustrated in Figure 3. These base distributions include a unimodal test case with simple unimodal normal distributions–the typical assumption for traditional statistics–and a multimodal test case designed to test how IR behaves when the unimodal assumption is broken–a common occurrence in real-world datasets.

A second factor in the empirical experiments was the amount of known correlation in a dataset. The experiment tested four levels of correlation: strong correlation (Pearson's r = 0.61), moderate correlation (r = 0.25), weak correlation (r = 0.1) and completely random data with no correlation (r = 0.0).

This two-by-four experimental design (two types of base distributions, four levels of correlation) resulted in the creation of eight artificial datasets–eight D in the IR nomenclature–as summarized in Table 1. Each D consists of 5,000 (x,y) data points drawn from the corresponding base distributions (unimodal vs. multi-modal) to achieve the pre-determined level of correlation. These datasets served as the starting point for the empirical experiments described below.

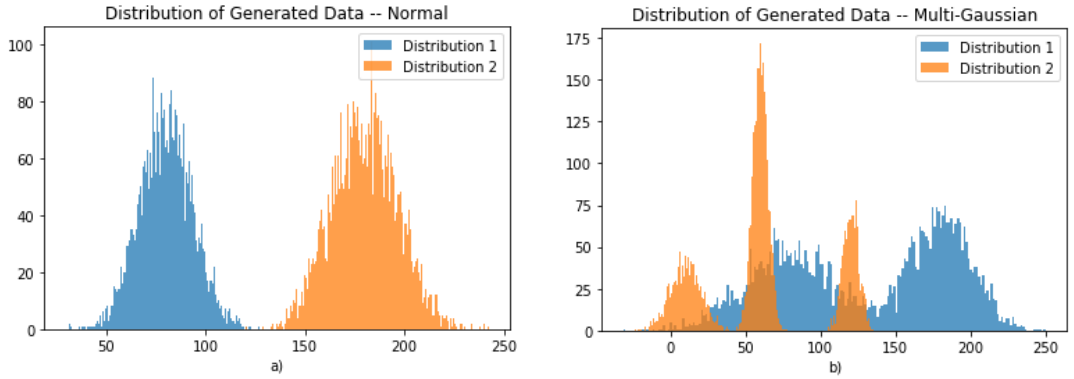


Figure 3. Empirical experiments measured IR behavior under two different conditions: (a) data drawn from simple unimodal normal distributions; and (2) data drawn from a more complex condition of two non-aligned multi-modal Gaussian distributions. This figure illustrates the distributions from which the eight datasets used in the empirical experiments were drawn.

Table 1. The empirical evaluation utilized eight artificially generated datasets corresponding to the two-by-four experimental condition design of the experiments (two distribution types, four correlation levels).

Pearson Correlation	Unimodal	Multi-modal
Strong Correlation (r=0.61)	D_{uni}^{strong}	D_{mult}^{strong}
Medium Correlation (r=0.25)	D_{uni}^{med}	D_{mult}^{med}
Weak Correlation (r=0.1)	D_{uni}^{weak}	D_{mult}^{weak}
No Correlation (r=0.0)	D_{uni}^{none}	D_{mult}^{none}

Experiments

For each D, 10,000 randomized IR experimental trials were conducted for each of the $4x5x3 = 60$ experimental conditions (4 settings for the number of IR folds n, 5 settings for the p-value threshold in the metric function, and 3 settings for the type of aggregation function). The results of these trials were then analyzed to characterize how IR behaves under the various conditions.

In each trial, 1,000 (x,y) data points were randomly sampled from D to create a focused data subset d_i where $i \in [0, 9999]$. We note that while each D has a known correlation value between x and y, the correlation for each random subset d_i varies due to the random sampling process. Figure 4 shows a summary of the correlations found in the 10,000 d_i for each of the eight D in our experiments. As expected, the mean correlation for each set of 10,000 d_i was aligned with the correlation value present in the corresponding D. Moreover, because we aim to test only a single statistic (correlation) in each trial, we treat each d_i as a single measurement-specific subset m_i in the IR pipeline.

A first set of experiments was designed to explore three different parameters of the IR process. First, IR was applied to each m_i with four different n values (1, 3, 5, and 7) during partitioning. When IR is used with $n > 1$, the data are randomly split into n disjoint folds. Second, the metric function applied in the experiments was a significance test for Pearson correlation for the x and y values in each m_i. This

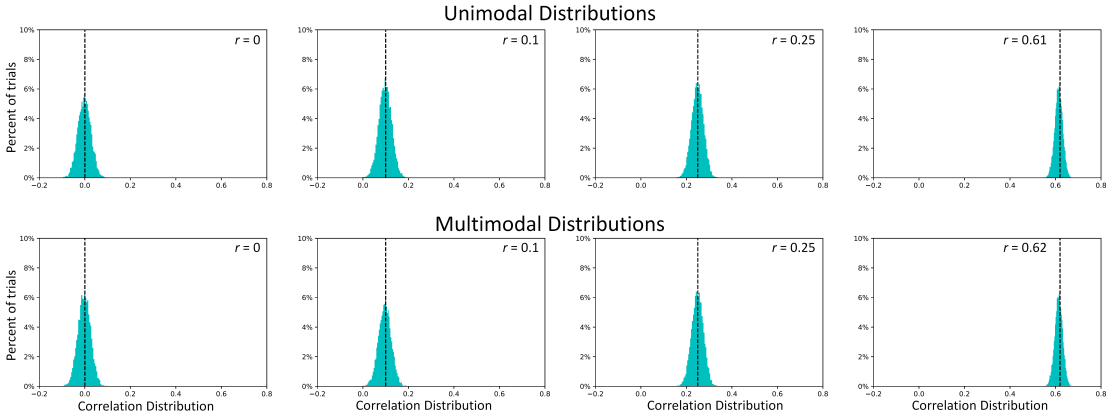


Figure 4. For each of the eight artificially generated datasets D, we created 10,000 randomly sampled subsets (d_i) of size 1,000 to use in our experiments. Due to the random sampling process the correlation present in each d_i varied slightly from the correlation in D. Each plot in this figure shows a histogram of the correlation values for the 10,000 d_i, with the correlation for the corresponding D shown as a dotted black line.

was chosen because it is a commonly used metric in real-world systems, including the DecisionFlow2 system described in Section "Use Cases." The metric calculated the Pearson correlation for each fold and computed a p-value against the null hypothesis, h_0, that the data points in the fold are not correlated. Tests were performed at five different threshold levels (0.005, 0.01, 0.02, 0.03, and 0.04). Finally, three different aggregation methods were used to generate m'_i for each m_i: (1) *any-pass*, in which we reject h_0 if any fold passes the p-value threshold test, (2) *majority vote*, in which we reject h_0 if more than half of the folds pass the test, and (3) *all-pass*, in which we reject h_0 only if all folds pass the test.

A second set of experiments was designed to characterize IR's behavior in response to different numbers of data points in d_i. The previously described experiments maintained a constant size of 1,000 for each d_i while (1) varying the number of folds, and (2) drawing samples for D with different levels of correlation. In the second round of experiments, the number of folds was kept constant at $n = 5$ and the samples were all drawn from a D with a correlation of $r = 0.25$. However, in these experiments the size of d_i was varied with sizes of 100, 500, 1,000, 1,500, and 2,000. A similar set of *any-pass*, *majority-vote*, and *all-pass* experiments were performed for both the unimodal and multimodal distributions.

Results and Discussion

The results of the experimental trials outlined above are illustrated in Figures 5 and 6 for the unimodal and multi-modal distributions, respectively. The charts reflect the 4x5x3 experimental design: the number of folds (n) is shown along the x-axis of each individual chart; the p-value threshold is represented by differently colored lines; and the three aggregation methods are represented by the three columns. The eight D map to the eight rows (four in each figure). The y-axis in all charts shows the number of the 10,000 trials that are flagged as having significant correlation in the aggregate measure.

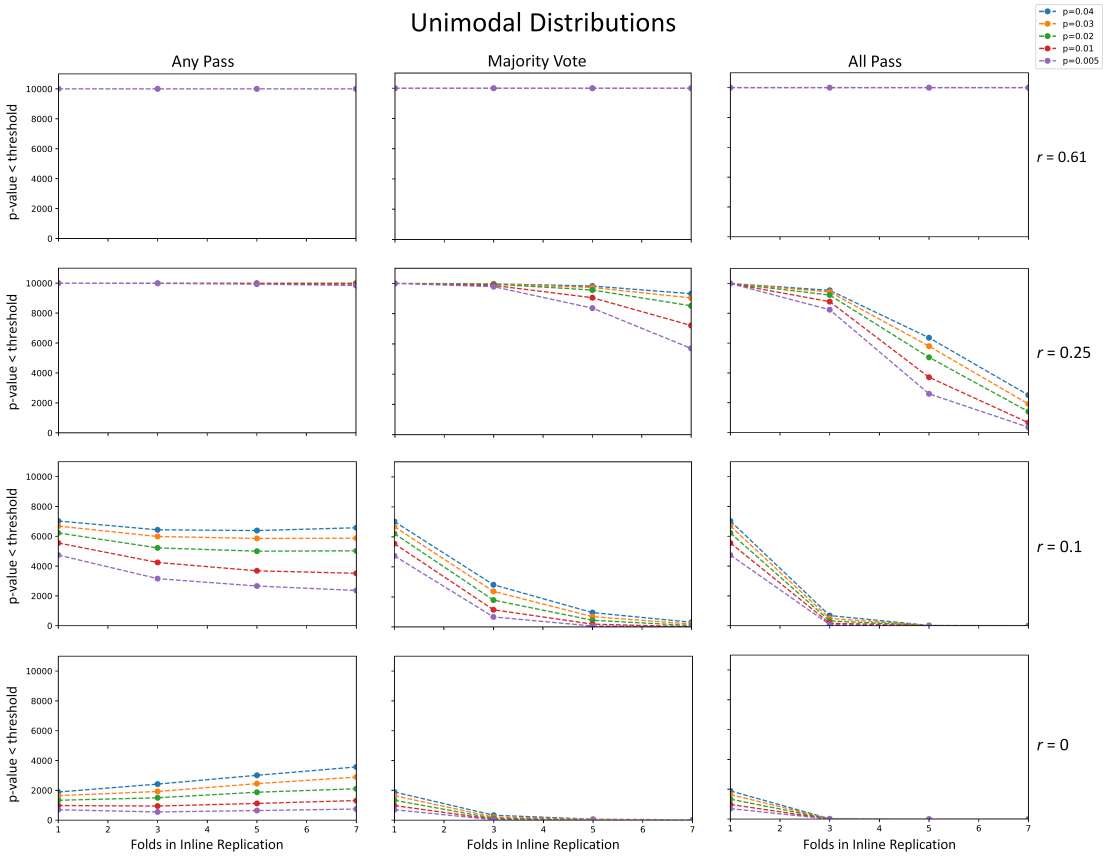


Figure 5. Results of significance tests on correlation for 10,000 trials of subsets of 1,000 (x, y) data points sampled from datasets of 5000 (x, y) data points from unimodal normal distributions with differing correlations (0.61, 0.25, 0.1 and 0). The number of trials with p-values below different threshold values (0.04, 0.03, 0.02, 0.01, and 0.005) for inline replication using different numbers of folds (1, 3, 5, and 7) are shown for three different aggregation methods (any pass, majority vote, and all pass).

The charts in Figure 5 show the results for the unimodal experiments. As might be expected, when the correlation was high ($r = 0.61$), all trials are shown to pass the significance threshold regardless of aggregation method or number of folds. This demonstrates that high correlation is successfully found with zero false negatives. When the level of correlation in the dataset is moderate ($r = 0.25$), the number of trials passing the significance threshold for a given aggregation methods decreases as the number of folds (n) increases. This demonstrates that the n parameter serves as a control over the sensitivity of the system to flag results as significant.

There is a similar reduction as n increases for correlations of 0.1 and 0 for all conditions except under the the "any pass" aggregation method. This result may seem counter intuitive at first. However, in this

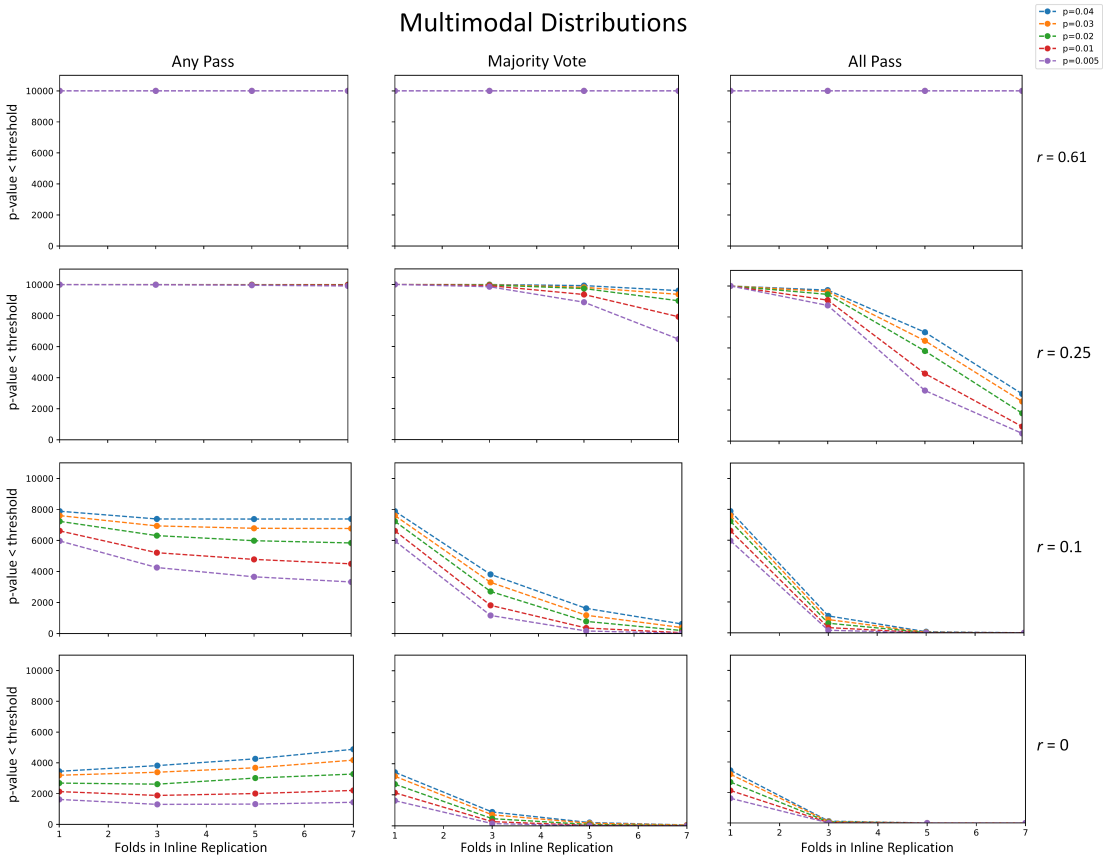


Figure 6. Results of significance tests on correlation for 10,000 trials of subsets of 1,000 (x, y) data points sampled from datasets of 5000 (x, y) data points from multimodal normal distributions with differing correlations (0.61, 0.25, 0.1 and 0). The number of trials with p-values below different threshold values (0.04, 0.03, 0.02, 0.01, and 0.005) for inline replication using different numbers of folds (1, 3, 5, and 7) are shown for three different aggregation methods (any pass, majority vote, and all pass).

case, the only correlation present in a given trial is the result of random noise since the dataset from which the sample is drawn has zero correlation by design. Increasing the number of folds results in an increase in the number of statistical tests. This in turn allows for a higher chance of noise producing a result that appears significant. Because the "any pass" aggregation method marks any result as significant as long as at least one fold was marked as significant, an increase in n results in an increase in false positives. This result is critical. Using an aggregation method such as "any pass" essentially eliminates the benefits of IR because it enables any single fold result to determine the aggregate measure result. This means that results need not replicate across folds, increasing the chance for false positive results.

As expected, trials that used the stricter "majority vote" and "all pass" aggregation methods had fewer samples exceed the significance threshold when compared to "any pass." As the correlation level decreased, the "all pass" aggregation method returned the fewest significant results. This reflects the stricter replication requirements for that approach. This trend is evident across all permutations of the experiment, suggesting that IR behaves predictably, providing system designers with multiple controls (the number of folds n, and the aggregation method) over the tradeoff between false positives and false negatives.

We note that the results of these experiments confirm that the "any pass" aggregation method should not be used for IR as it results in an increase of false positives. The results are reported to contrast with the other aggregation options tested in the experiments ("all pass" and "majority vote"), and to highlight that some aggregation functions are counter-productive.

Comparing the data from the unimodal experiments to the multi-modal results in Figure 6, the results are nearly identical. All of the observations noted above with respect to the unimodal results also apply in the multi-modal case. This suggests that one of the primary design goals for IR has been met: that IR is robust to the underlying data distribution, without any assumption of normality.

The results of the second set of experiments, which focused on the impact of sample size on IR behavior, are shown in Figure 7. As expected, larger sample sizes produced more reliable detection of the correlations at all correlation levels. Using a stricter p-value thresholds (e.g., $p < 0.01$), or a stricter aggregation functions (e.g., "all pass") resulted in fewer trials producing significant results. This result confirms that IR behaves as desired, exhibiting the same response to increases in samples size as traditional statistical tests.

Overall, the experimental results present a clear and familiar picture of the tradeoffs that exist in the parameterization of IR (e.g., the number of folds (n), and the choice of aggregation functions) when it comes to limiting the number of false positives vs. false negatives. Higher n and stricter aggregation methods result in a higher threshold for reporting meaningful results. This tradeoff suggests that there is no "always optimal" set of parameters for IR. However, the results suggest that using $n = 5$ seems to provide some balance between the two extremes, while "majority vote" aggregation seems to offer a similar compromise. In particular, this combination exhibited desired behavior in both low and high correlation experiments as illustrated in Figure 6.

Use Cases

The IR approach is compatible with a broad range of visual metaphors and interaction models, from basic charts to more sophisticated exploratory visual analysis systems. To demonstrate this flexibility and to explore the impact of adopting an IR pipeline, we developed two prototype IR systems: (a) a reference prototype to study IR in isolation, and (b) a sophisticated visual analysis system to examine IR in the context of a more complex analysis environment.

Reference Prototype

We developed a reference IR implementation as part of a simplified visual analysis prototype with the goal of exploring the IR parameter space in isolation, without concern for the more complex interactions that are part of a real-world application such as the one described in Section "DecisionFlow2." The prototype supports two basic visualization metaphors: (a) bar charts and (b) scatter plots with linear

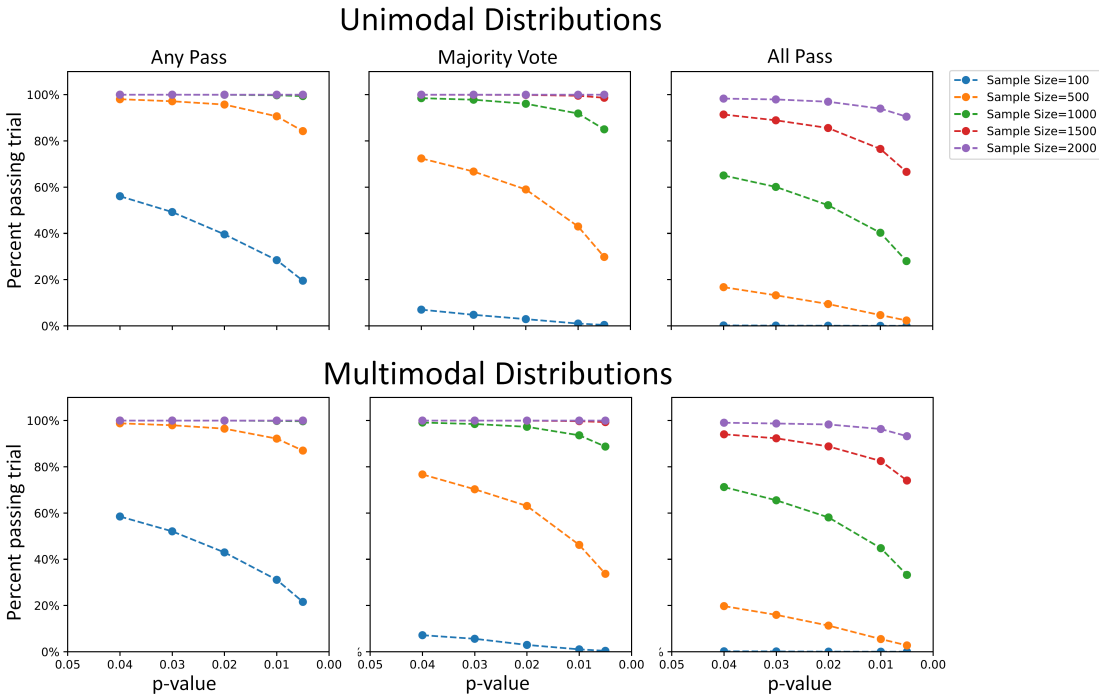


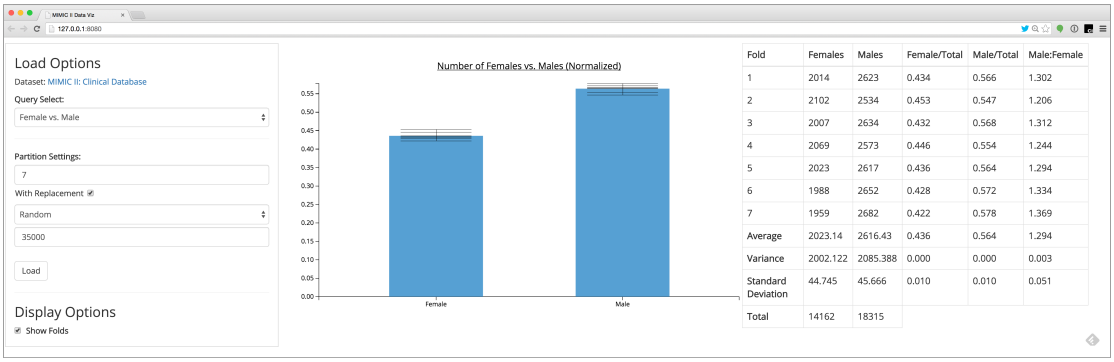
Figure 7. Results of significance tests on correlation for 10,000 trials of subsets of data points sampled from unimodal and multi-modal distributions with correlation of $r = 0.25$. The number of trials with p-value thresholds (0.04, 0.03, 0.02, 0.01, and 0.05) for inline replication using a constant number of folds ($n = 5$) is shown for different sample sizes of d_i (100, 500, 1,000, 1,500, and 2,000).

regression lines. The prototype was tested using a dataset of electronic medical data containing over 40,000 intensive care unit (ICU) stays.[43]

The prototype interface, shown in Figure 8, contains three panels. In the center is the visualization canvas itself. The left panel enables users to issue queries and control key parameters of the IR process. Options include the number of folds (n) for the partition function, the use of sampling with replacement, support for random or ordered incremental sampling, and controls to unfold the merged statistics and show individual folds within the visualizations. The right panel shows detailed descriptive statistics for both the individual folds (s_i^j) and the aggregation (s_i).

Figure 9 shows a series of bar charts rendered using the IR prototype to visualize the gender distribution across three subpopulations from the ICU database. This example is directly analogous to the roulette wheel bar chart example introduced in Section "Visualization as a Predictive Model," as both summarize the distribution of a binary variable in a given population.

The top row of charts in Figure 9 shows the aggregate gender distribution for each of the three populations. The charts show a relatively similar distribution across all three populations, with a moderate



Fold	Females	Males	Female/Total	Male/Total	Male:Female
1	2014	2623	0.434	0.566	1.302
2	2102	2534	0.453	0.547	1.206
3	2007	2634	0.432	0.568	1.312
4	2069	2573	0.446	0.554	1.244
5	2023	2617	0.436	0.564	1.294
6	1988	2652	0.428	0.572	1.334
7	1959	2682	0.422	0.578	1.369
Average	2023.14	2616.43	0.436	0.564	1.294
Variance	2002.122	2085.388	0.000	0.000	0.003
Standard Deviation	44.745	45.666	0.010	0.010	0.051
Total	14162	18315			

Figure 8. The IR-based prototype shown here was developed to test the proposed pipeline and to explore the parameter space with two baseline visualization types: bar charts and linear regression lines. The left panel shows the query and IR controls, the middle panel shows the visualization space, and the right panel shows detailed descriptive statistics computed for both the aggregate representation and the individual folds.

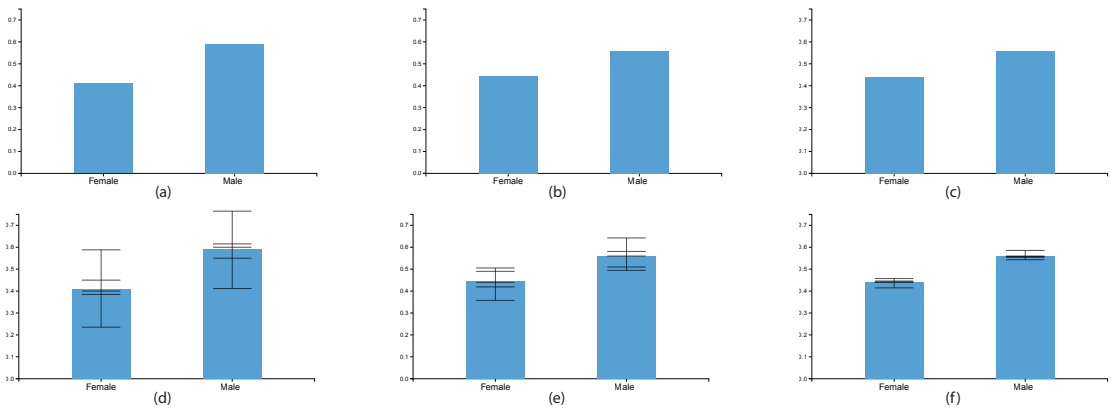


Figure 9. Six charts produced by the IR prototype system. The top three charts (a-c) show the gender distribution for three different sets of ICU patients. The relatively similar bar charts suggest that the underlying populations are comparable. However, when the same populations are visualized with 5 folds (d-f), a different story appears. The charts now clearly demonstrate that we know less about the population visualized in the left column than we do about the population on the right. In this case, the difference is due largely to the size of the respective populations.

increase in female representation moving from panel (a) to (b) to (c). The bar chart shows the gender breakdown in each population quite clearly. However, there is no indication of the distribution's stability across different groups of patients. Consumers of the visualization are left to assume that the bar charts provide an accurate depiction.

Panels (d-f) in Figure 9 show the exact same populations as panels (a-c), respectively. However, these views incorporate measures computed for multiple folds ($n = 5$) using IR. These unfolded views provide

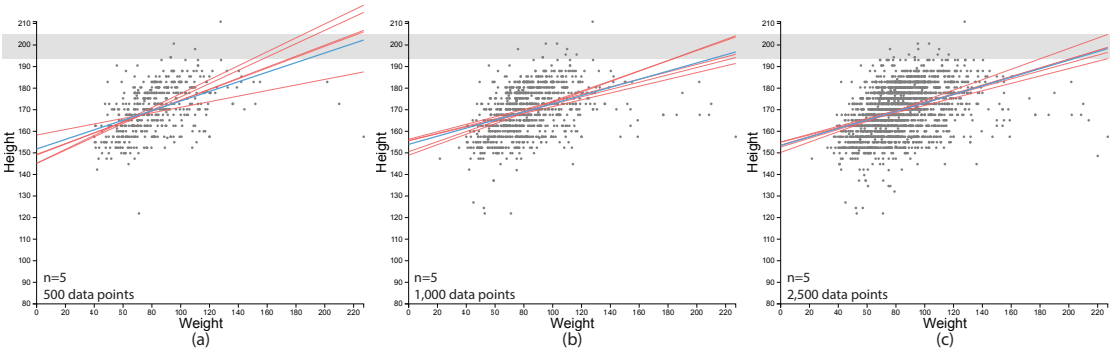


Figure 10. Weight versus height distribution for patients admitted to a neonatal intensive care unit. Simulating the results from a progressive visualization system, this figure shows both the raw data and best fit regression line (shown in blue) for (a) 500 patients, (b) 1,000 patients, and (c) 2,500 patients. In all three cases, the IR pipeline has computed a regression across five folds, shown in red. The decreasing spread across the red regression lines conveys the expected—but often overlooked—change in variation between folds as the sample size increases. The gray band across all three charts has been added to this figure to emphasize these differences and reflects the variation across folds in (c) at the maximum observed weight.

a more accurate picture regarding the repeatability of the gender distributions. In particular, we see from Figure 9(d) that the population visualized in the left column of the figure is not very predictable. Meanwhile, far less variation across folds is visible in Figure 9(f). In this case, the major difference is the size of the respective populations, which are 100, 1,000, and 10,000 for panels (d), (e), and (f), respectively. The expected variance within each group is a critical factor for interpretation, but is invisible in the original bar charts and easily overlooked even by expert users.[38] The IR approach makes this information more visible without requiring assumptions about the underlying distributions.

Figure 10, meanwhile, shows three screenshots of the IR prototype displaying data from the ICU dataset using scatter plots with linear regression lines. In this case, the examples show data for populations of neonates, with weight mapped to x position and height mapped to y position. A linear regression model was calculated in all three cases using the IR pipeline with $n = 5$. The five regression lines, one for each fold, are visible ("unfolded") in the visualizations as red lines. In addition, an aggregate best-fit linear model is shown in blue.

These screenshots show how IR helps convey uncertainty during progressive analysis, using the incremental sampling feature of the prototype to vary the number of samples while keeping all IR parameters constant. In Figure 10(a), only 500 patients are included in the scatter plot. As captured by the varying slopes between the five red lines, there is relatively large disagreement across folds in the linear models they produce. This uncertainty would be invisible in a similar plot rendered without the folds.

As expected, the spread between the individual fold regression lines decreases as more patients are retrieved by the incremental query feature. For example, Figure 10(b) shows the same visualization with the same $n = 5$ folds. However, this version includes data for 1,000 patients. The larger sample size results in increased stability across the folds. Part (c) of the same figure shows the same visualization

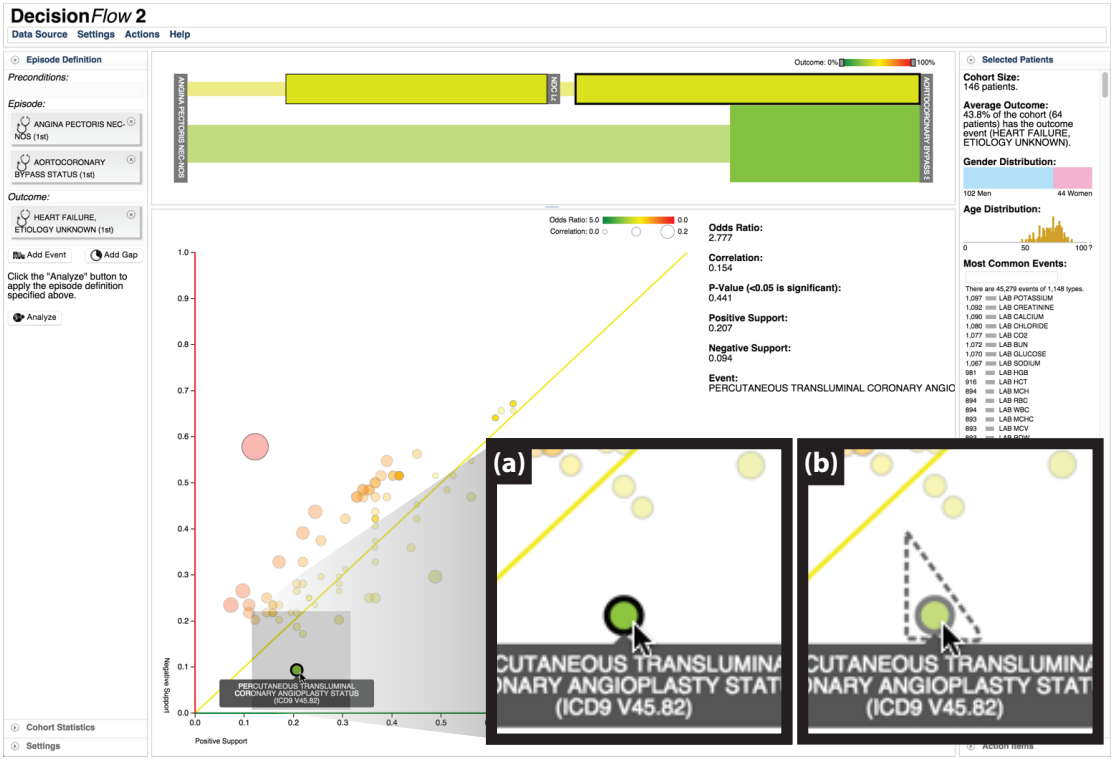


Figure 11. The DecisionFlow2 visual analytics system is shown here displaying medical event data using the Inline Replication (IR) process outlined in this paper. The data in this example have been analyzed using five folds, without replacement. The inset subfigures show (a) an initial visualization of the aggregation function's results for a particular medical event, and (b) a more detailed "unfolded" representation showing the variation in positive and negative support as observed across the five folds produced by the partition function. Figures 12 and 13 show how differences in the unfolded representation can help inform users during an analysis.

with 2,500 patients. We see little improvement in agreement across folds compared to 1,000 patients, suggesting that the rate of further gains in agreement will slow down.

As previously stated, the improvement in agreement as sample size increases is expected. However, as evidenced by the "recent history" charts at casino roulette tables and the other examples referenced throughout this paper, visualizations are often assumed to be accurate, without taking into account issues of sample size or variation. This use case shows that IR can effectively convey this variation in the data without the need for careful modeling, and in a non-parametric way that avoids assumptions about the underlying distributions.

DecisionFlow2

To test IR within a more fully-featured exploratory visual analysis environment, we developed DecisionFlow2, a new IR-based version of our existing visual analysis system for high-dimensional temporal event sequence data.[7] A screen capture of the DecisionFlow2 interface is shown in Figure 11.

Original DecisionFlow Design The original version of DecisionFlow made heavy use of p-values to help users identify event types that had a statistically significant correlation to a user-specified outcome measure. When visualizing medical data, for example, this approach enables users to find types of medical events (such as specific diagnoses, medications, and procedures) that—when appearing in a particular pattern in a patient's history—are associated with better or worse medical outcomes.

An interactive timeline at the top of the screen enables users to segment a cohort of event sequences based on the presence of so-called "milestone" events. For a given subgroup, DecisionFlow visualizes statistics for the potentially thousands of different event types that occur between milestones, with the goal of helping users identify good candidates for new milestones. DecisionFlow conveys the event type statistics via an interactive bubble chart similar to the one seen in Figure 11.

In the bubble chart, each event type is represented by a circle whose x-axis position is determined by its positive support (the fraction of "good outcome" event sequences that contain the event type). Similarly, each circle's y-axis position is determined by its negative support (the fraction of "bad outcome" sequences with the event type). Circle size and color encode correlation and odds ratio, respectively. Importantly, circles representing event types whose presence correlates significantly ($p < 0.05$) with outcome are drawn with a distinct border to make it easier for users to visually distinguish between expected variation and potentially meaningful associations.

Design Adaptation for IR In the IR-based DecisionFlow2 system developed for this paper, a similar bubble chart design is used to visualize the event type statistics. However, rather than showing data for measures computed for the overall population m_i, the circles encode aggregate measures computed by IR. For example, Figure 11 shows the system with a bubble chart focused on a subset of data containing 45,278 individual events with 1,148 distinct event types. The support values (used to position the circles) and other measures were all computed across 5 folds.

The aggregate view (without showing the unfolded data) in Figure 11(a) looks essentially identical to the original DecisionFlow design. This is as intended, with the goal of making IR compatible with typical visualization designs. However, while the visual encoding is similar, the number of statistically significant correlations scores is reduced. In particular, a number of event types that were labeled as statistically significant in the original design were no longer found to be significant once majority-voting across the five folds was used to determine which event types were significant. This makes the visualization system more selective in rejecting the null hypothesis. The result is a reduction in the likelihood of Type 1 errors, which are a common problem in high-dimensional exploratory analysis. More detailed results and discussion are provided in Section "Results and Analysis."

Another important part of the IR-based DecisionFlow2 is the ability to unfold the aggregate statistics for each event type. Users can unfold an event type by hovering the mouse pointer over the corresponding circle. For example, after hovering the mouse pointer for a few seconds over the circle shown in Figure 11(a), the unfolded representation shown in Figure 11(b) is added to the visualization.

As this example shows, the DecisionFlow2 displays the unfolded data as a convex region drawn around the original circle and outlined with a dashed border. This region corresponds to the convex

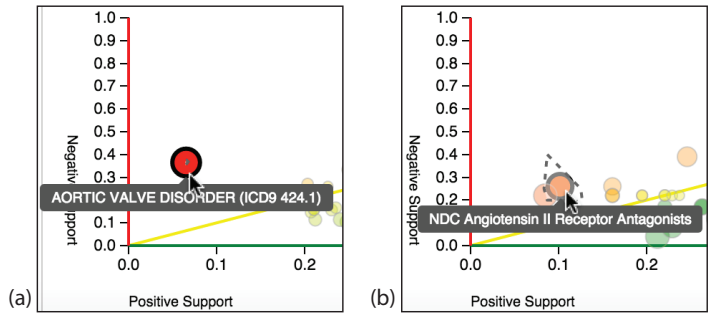


Figure 12. In general, (a) smaller differences between folds are seen when sample sizes are larger, while (b) higher levels of variation are seen for smaller sample sizes.

hull determined by the (x, y) locations for each of the n folds that contribute to the aggregate measures that determine the position of the original circle. In other words, the size and position of the unfolded region represent the variation across folds in both the positive and negative support measures. Smaller unfolded regions indicate that the values have little variation across folds. Larger unfolded regions, such as the one shown in Figure 11(b), suggest a high degree of variation between folds and therefore lower confidence in the repeatability of the aggregate measure.

The typical behavior observed when utilizing the IR-based implementation of DecisionFlow2 is shown in Figure 12. Part (a) of the figure shows an event type from a very large subset of data that shows very limited variation across folds. This is represented by the very small unfolded region located near the center of the red circle just above the mouse pointer. Part (b) of the figure, meanwhile, shows an event type with much higher variation (the dashed unfolded region surrounding the mouse pointer). This figure, visualizing data from a smaller sample size, demonstrates what one might expect: findings based on smaller sample sizes tend to have more variability and therefore should typically be given less weight in a decision making process.

However, this very critical difference is not observable via the original bubble chart. The size of the data corresponding to each bubble is made available elsewhere in the user interface for users who consciously seek it out, but the implications of the differences in data size are left to the user's imagination. It is only through the unfolding process that the visualization itself conveys the difference in confidence that users should place in one view versus the other.

Moreover, it is critical to note that the size of the dataset is not the sole determinant of repeatability for a given measure across folds. Major differences in measure values can be seen even for similarly sized datasets. For example, Figure 13 shows three different event types from the exact same subset of event sequences. While the number of event sequences was the same for each type, the association between *ACE Inhibitors* (center panel) and the user-defined outcome (eventual diagnosis with heart failure) was far more consistent across folds.

Results and Analysis The IR-based DecisionFlow2 prototype provides visual feedback regarding the variation in positive and negative support. As previously described, the system also uses IR to assess the statistical significance of each event type's correlation with patient outcome. For a given event type, correlation coefficients and p-values are computed for each fold, then aggregated via majority-vote.

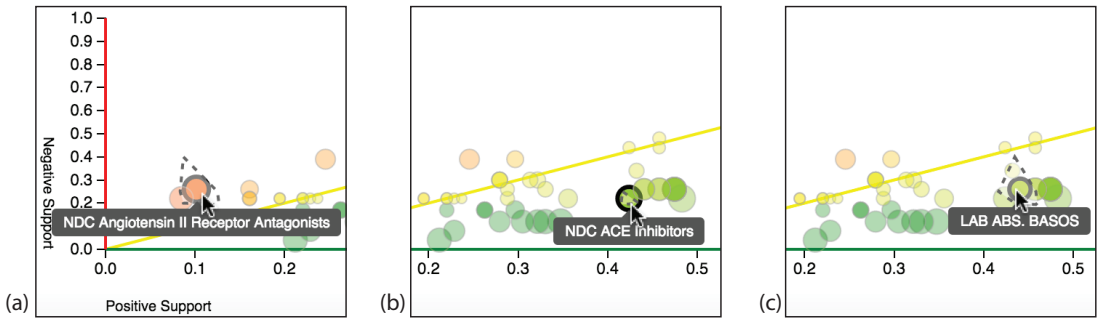


Figure 13. Even with the same sample size, different measures can have different levels of repeatability across folds. In this example, both (a) and (c) show relatively high levels of variability, while the small unfolded region in (b) suggests that the relationship between outcome and ACE Inhibitors was fairly consistent across all five folds. All three views were calculated using identical sample sizes.

Event types with more than $n/2$ folds showing statistically significant correlation are displayed in the visualization with a distinct visual encoding.

To better understand the impact of IR and the choice of n on the visualized results, we conducted a quantitative experiment in which we compared performance for a sample user interaction sequence under various conditions. More specifically, we experimented repeatedly by performing the exact same exploratory analysis steps using DecisionFlow2, using the exact same input data, varying only the number of folds. The experiment was conducted at three partition settings: $n = \{1, 3, 5\}$.

In all three cases, the input dataset consisted of event data from the medical records of 2,899 patients containing 1,074,435 individual medical events. These timestamped events contained 3,631 distinct medical event types: specific diagnoses, lab tests, or medication orders that were present in the patients' records. Of the 3,631 distinct event types, 381 were deemed prevalent enough by the DecisionFlow2 system to be the target of correlation analysis within the metric function. The same threshold was used across all three partition settings, enabling us to compare analysis results across the exact same control conditions.

The results of our analysis are shown in Table 2. With $n = 1$, the DecisionFlow2 system flagged statistically significant results in the same way as in the original paper.[7] Using a threshold of $p < 0.05$, 144 statistically significant event types were detected. When n was increased to three, the numbers were reduced dramatically. Only 50 of the original 144 statistically significant event types remained after applying a majority-vote aggregation algorithm. Of those 50, only 43 were significant across all three folds. For $n = 5$, the number of significant event types was even smaller. The stricter requirements for replication resulted in just 24 event types being flagged as significant given a majority-vote aggregation algorithm, and just 15 event types were significant across all 5 folds.

As expected—and as intended—the number of statistically significant findings is reduced as n grows from one to five. There are two primary reasons for this reduction. First, because each condition is applied to the same set of event sequences for the same patients, the partition size is smaller as n increases. The smaller number of patients reduces the statistical power for each partition. The expected impact of this is higher p-values and fewer statistically significant findings. With the ever-growing size of datasets in

Table 2. A comparison of statistically significant findings in three different IR configurations with DecisionFlow2 applied to the same data. The number of event types flagged as significantly associated with outcome was largest for $n = 1$. This setting corresponds to a traditional visualization approach with no partitioning. Larger n values dramatically reduced the number of significant findings.

Number of Folds	$n = 1$	$n = 3$	$n = 5$
Unanimously Significant	144	43	15
Majority Significant	144	50	24
At Least One Significant	144	56	29
Total Number of Measurements Made		381	
Total Number of Event Types		3,631	

many applications, however, the impact on statistical power due to partitioning should be minimal in many use cases. At the same time, the majority vote aggregation function requires that a significant level be repeatedly observed across multiple partitions (2 for $n = 3$, or 3 for $n = 5$). This reduces the likelihood of random variation being misinterpreted.

While statistical significance based on p-value thresholds has known limitations to medical research and beyond,[44] it is a widely used metric in exploratory visualization because it enables a rough filtering of data to manage visual complexity and the user's analytic attention. Follow-up analysis of any discovered insights is required. For this reason, reducing Type 1 errors becomes critical for modern visual analysis applications where vast numbers of data points can be tested and prioritized for user analysis. As the results presented here show, IR applies a higher bar for statistical significance, which has the potential to limit unsupported conclusions from the data in cases where users make quick predictive assessments directly from a visualization. It can also save significant effort in cases where follow-up analysis is performed by reducing the number of falsely generated hypotheses.

Discussion of Limitations

The IR approach is designed to embed the process of replication directly within the visualization pipeline, providing a non-parametric approach to calculating and visualizing the repeatability of derived measures. As the examples in Section "Use Cases" demonstrate, the approach can be effective when applied to a variety of different measures and visual metaphors. However, there are limitations to IR that must be acknowledged.

First, the proposed approach does nothing to combat selection bias or other problems in the creation of the original dataset. Any systemic sampling biases in the original data will be present across all folds created by the partitioning algorithm. Therefore, even measures that generalize well across multiple partitions are not necessarily generalizable to entirely new datasets.

Second, the IR approach is not truly predictive in nature. While information about the ability of various measures to replicate across multiple folds can be useful in vetting potential conclusions, findings uncovered via IR should be considered hypotheses that require testing using more rigorous methods when important decisions are to be made.

In particular, hypothesis testing often requires the collection and analysis of new data to fully understand the conditions under which a given insight holds true. Our method does not replace this

step. Instead, IR helps reduce the number of Type 1 errors, which can lower the number of conclusions that need testing. However, IR does not eliminate the necessity of a post-hypothesis validation process.

Conclusion

Traditional data visualizations show retrospective views of existing datasets with little to no focus on prediction or generalizability. However, users often base decisions about future events on the findings made using these visualizations. In this way, visualization can be considered to be a visual predictive model that is subject to the same problems of overfitting as traditional modeling methods. As a result, visualization users can often make invalid inferences based on unreliable visual evidence.

This paper described an approach to visual model validation called *Inline Replication* (IR). Similar to cross-validation and bootstrap resampling techniques, IR provides a nonparametric and broadly applicable approach to visual model assessment and repeatability. The IR pipeline was defined, including three key functions: the partition function, the metric function, and the aggregation function. In addition, methods for visual display and interaction were discussed. The article reported results from empirical experiments that capture how IR performs under different conditions, providing insights into how the choice of IR parameters impacts performance. Finally, two uses cases were described, including a new IR-based implementation of a previously-published exploratory visual analytics system. The use cases demonstrated the successful compatibility of IR with a variety of visual metaphors and derived measures.

While the results presented in this paper are promising, they represent only one step in a growing effort to bring high repeatability and predictive power to visualization-based analysis systems. There are many areas for future work including: improved techniques for detecting and conveying issues related to missing data, techniques for addressing and visually warning users regarding selection bias, and improved methods for conveying the degree of compatibility between a given statistical model's assumptions and the actual underlying data.

Acknowledgements

This article is based in part upon work supported by the National Science Foundation under Grant No. 1704018. We thank Brandon A. Price for his contributions to the software development process for the reference prototype described in the Use Cases section of this article.

References

1. Shneiderman B. The eyes have it: a task by data type taxonomy for information visualizations. In *, IEEE Symposium on Visual Languages, 1996. Proceedings*. pp. 336–343. DOI:10.1109/VL.1996.545307.
2. Card S, Mackinlay J and Shneiderman B. *Readings in information visualization: using vision to think*. 1999. ISBN 1558605339.
3. Chi EH. A taxonomy of visualization techniques using the data state reference model. In *IEEE Symposium on Information Visualization, 2000. InfoVis 2000*. pp. 69–75. DOI:10.1109/INFVIS.2000.885092.
4. Ostle B and others. Statistics in research. *Statistics in research* 1963; (2nd Ed).
5. Perer A, Bertini E, Maciejewski R et al. IEEE VIS 2014 Workshop on Visualization for Predictive Analytics. http://predictive-workshop.github.io/, 2014.

6. Gotz D, Price BA and Chen AT. Visual model validation via inline replication. *CoRR* 2016; abs/1605.08749. URL http://arxiv.org/abs/1605.08749. 1605.08749.
7. Gotz D and Stavropoulos H. DecisionFlow: Visual Analytics for High-Dimensional Temporal Event Sequence Data. *IEEE Transactions on Visualization and Computer Graphics* 2014; Early Access Online. DOI: 10.1109/TVCG.2014.2346682.
8. Skeels M, Lee B, Smith G et al. Revealing uncertainty for information visualization. *Information Visualization* 2010; 9(1): 70–81.
9. Potter K, Rosen P and Johnson CR. From quantification to visualization: A taxonomy of uncertainty visualization approaches. In *Uncertainty Quantification in Scientific Computing*. Springer, 2012. pp. 226–249.
10. MacEachren A, Roth R, O'Brien J et al. Visual Semiotics and Uncertainty Visualization: An Empirical Study. *IEEE Transactions on Visualization and Computer Graphics* 2012; 18(12): 2496–2505. DOI: 10.1109/TVCG.2012.279.
11. Sanyal J, Zhang S, Bhattacharya G et al. A User Study to Compare Four Uncertainty Visualization Methods for 1d and 2d Datasets. *IEEE Transactions on Visualization and Computer Graphics* 2009; 15(6): 1209–1218. DOI:10.1109/TVCG.2009.114.
12. Tak S, Toet A and van Erp J. The Perception of Visual Uncertainty Representation by Non-Experts. *IEEE transactions on visualization and computer graphics* 2013; DOI:B42BEEBA-E27C-47E6-B973-0E02E9C3003A.
13. Potter K, Gerber S and Anderson EW. Visualization of uncertainty without a mean. *Computer Graphics and Applications, IEEE* 2013; 33(1): 75–79.
14. Brodlie K, Osorio RA and Lopes A. A review of uncertainty in data visualization. In *Expanding the Frontiers of Visual Analytics and Visualization*. Springer, 2012. pp. 81–109.
15. Borland D, Wang W and Gotz D. Contextual Visualization: Making the Unseen Visible to Combat Bias During Visual Analysis. *IEEE Computer Graphics and Applications* 2018; 38(6).
16. Gotz D, Sun S and Cao N. Adaptive Contextualization: Combating Bias During High-Dimensional Visualization and Data Selection. In *Proceedings of the 21st International Conference on Intelligent User Interfaces*. IUI '16, New York, NY, USA: ACM. ISBN 978-1-4503-4137-0, pp. 85–95. DOI:10.1145/2856767.2856779. URL http://doi.acm.org/10.1145/2856767.2856779.
17. Gotz D, Sun S, Cao N et al. Adaptive Contextualization Methods for Combating Selection Bias During High-Dimensional Visualization. *ACM Trans Interact Intell Syst* 2017; 7(4): 17:1–17:23. DOI:10.1145/3009973. URL http://doi.acm.org/10.1145/3009973.
18. van den Elzen S and van Wijk J. BaobabView: Interactive construction and analysis of decision trees. In *2011 IEEE Conference on Visual Analytics Science and Technology (VAST)*. pp. 151–160. DOI:10.1109/VAST.2011.6102453.
19. Muhlbacher T and Piringer H. A Partition-Based Framework for Building and Validating Regression Models. *IEEE Transactions on Visualization and Computer Graphics* 2013; 19(12): 1962–1971. DOI:10.1109/TVCG.2013.125.
20. Gosink L, Bensema K, Pulsipher T et al. Characterizing and visualizing predictive uncertainty in numerical ensembles through bayesian model averaging. *Visualization and Computer Graphics, IEEE Transactions on* 2013; 19(12): 2703–2712.
21. Sipes NS, Martin MT, Reif DM et al. Predictive models of prenatal developmental toxicity from ToxCast high-throughput screening data. *Toxicological Sciences* 2011; : kfr220.
22. Fisher D, Drucker SM and Konig AC. Exploratory Visualization Involving Incremental, Approximate Database Queries and Uncertainty. *IEEE Computer Graphics and Applications* 2012; 32(4): 55–62.

23. Stolper C, Perer A and Gotz D. Progressive Visual Analytics: User-Driven Visual Exploration of In-Progress Analytics. *IEEE Transactions on Visualization and Computer Graphics* 2014; 20(12): 1653–1662. DOI: 10.1109/TVCG.2014.2346574.
24. Fisher D, Popov I, Drucker S et al. Trust Me, I'm Partially Right: Incremental Visualization Lets Analysts Explore Large Datasets Faster. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. CHI '12, New York, NY, USA: ACM. ISBN 978-1-4503-1015-4, pp. 1673–1682. DOI:10.1145/2207676.2208294.
25. Sheskin DJ. *Handbook of Parametric and Nonparametric Statistical Procedures: Third Edition*. CRC Press, 2003. ISBN 9781420036268.
26. Stoline MR. The Status of Multiple Comparisons: Simultaneous Estimation of All Pairwise Comparisons in One-Way ANOVA Designs. *The American Statistician* 1981; 35(3): 134–141. DOI:10.2307/2683979.
27. Rothman KJ. No adjustments are needed for multiple comparisons. *Epidemiology (Cambridge, Mass)* 1990; 1(1): 43–46.
28. Chen H, Zhang S, Chen W et al. Uncertainty-aware Multidimensional Ensemble Data Visualization and Exploration. *IEEE Transactions on Visualization and Computer Graphics* 2015; : 1–1DOI:10.1109/TVCG.2015.2410278.
29. Feng D, Kwock L, Lee Y et al. Matching Visual Saliency to Confidence in Plots of Uncertain Data. *IEEE transactions on visualization and computer graphics* 2010; 16(6): 980–989. DOI:10.1109/TVCG.2010.176.
30. Wu Y, Yuan GX and Ma KL. Visualizing Flow of Uncertainty through Analytical Processes. *IEEE Transactions on Visualization and Computer Graphics* 2012; 18(12): 2526–2535. DOI:10.1109/TVCG.2012.285.
31. Majumder M, Hofmann H and Cook D. Validation of Visual Statistical Inference, Applied to Linear Models. *Journal of the American Statistical Association* 2013; 108(503): 942–956. DOI:10.1080/01621459.2013.808157.
32. Wickham H, Cook D, Hofmann H et al. Graphical inference for infovis. *IEEE Transactions on Visualization and Computer Graphics* 2010; 16(6): 973–979.
33. Correa C, Chan YH and Ma KL. A framework for uncertainty-aware visual analytics. In *Visual Analytics Science and Technology, 2009. VAST 2009. IEEE Symposium on*. IEEE, pp. 51–58.
34. Munzner T. *Visualization Analysis and Design*. Har/psc edition ed. Boca Raton: A K Peters/CRC Press, 2014. ISBN 9781466508910.
35. Tufte ER. *Envisioning Information*. Cheshire, Conn.: Graphics Pr, 1990. ISBN 9780961392116.
36. Ware C. *Information visualization: perception for design (interactive technologies)*. Morgan Kaufmann, 2004.
37. Cammegh Limited. Cammegh - the world's finest roulette wheel. http://www.cammegh.com/product.php?product=displays, 2015. URL http://www.cammegh.com/product.php?product=displays.
38. Sutcliffe A, Bruijn Od, Thew S et al. Developing visualization-based decision support tools for epidemiology. *Information Visualization* 2014; 13(1): 3–17. DOI:10.1177/1473871612445832.
39. Zgraggen E, Zhao Z, Zeleznik R et al. 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 '18, New York, NY, USA: ACM. ISBN 978-1-4503-5620-6, pp. 479:1–479:12. DOI:10.1145/3173574.3174053. URL http://doi.acm.org/10.1145/3173574.3174053.
40. Kohavi R. 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, San Francisco, CA, USA: Morgan Kaufmann Publishers Inc. ISBN 1-55860-363-8, pp. 1137–1143.

41. Efron B. Bootstrap Methods: Another Look at the Jackknife. *The Annals of Statistics* 1979; 7(1): 1–26.
42. Lam L and Suen SY. Application of Majority Voting to Pattern Recognition: An Analysis of Its Behavior and Performance. *Trans Sys Man Cyber Part A* 1997; 27(5): 553–568. DOI:10.1109/3468.618255.
43. Goldberger AL, Amaral LAN, Glass L et al. PhysioBank, PhysioToolkit, and PhysioNet Components of a New Research Resource for Complex Physiologic Signals. *Circulation* 2000; 101(23): e215–e220. DOI: 10.1161/01.CIR.101.23.e215.
44. Goodman SN. Toward Evidence-Based Medical Statistics. 1: The P Value Fallacy. *Annals of Internal Medicine* 1999; 130(12): 995–1004. DOI:10.7326/0003-4819-130-12-199906150-00008.