Taylor & Francis
Taylor & Francis Group

Behaviour & Information Technology

ISSN: 0144-929X (Print) 1362-3001 (Online) Journal homepage: https://www.tandfonline.com/loi/tbit20

Towards insight-driven sampling for big data visualisation

Moeti M. Masiane, Anne Driscoll, Wuchun Feng, John Wenskovitch & Chris North

To cite this article: Moeti M. Masiane, Anne Driscoll, Wuchun Feng, John Wenskovitch & Chris North (2019): Towards insight-driven sampling for big data visualisation, Behaviour & Information Technology, DOI: 10.1080/0144929X.2019.1616223

To link to this article: https://doi.org/10.1080/0144929X.2019.1616223

Published online: 16 May 2019.

BEHAVIOUR & INFORMATION TECHNOLOGY
https://doi.org/10.1080/0144929X.2019.1616223

Taylor & Francis
Taylor & Francis Group

Towards insight-driven sampling for big data visualisation

Moeti M. Masiane, Anne Driscoll, Wuchun Feng, John Wenskovitch and Chris North

Virginia Tech, Blacksburg, VA, USA

ABSTRACT

Creating an interactive, accurate, and low-latency big data visualisation is challenging due to the volume, variety, and velocity of the data. Visualisation options range from visualising the entire big dataset, which could take a long time and be taxing to the system, to visualising a small subset of the dataset, which could be fast and less taxing to the system but could also lead to a less-beneficial visualisation as a result of information loss. The main research questions investigated by this work are what effect sampling has on visualisation insight and how to provide guidance to users in navigating this trade-off. To investigate these issues, we study an initial case of simple estimation tasks on histogram visualisations of sampled big data, in hopes that these results may generalise. Leveraging sampling, we generate subsets of large datasets and create visualisations for a crowd-sourced study involving a simple cognitive visualisation task. Using the results of this study, we quantify insight, sampling, visualisation, and perception error in comparison to the full dataset. We use these results to model the relationship between sample size and insight error, and we propose the use of our model to guide big data visualisation sampling.

ARTICLE HISTORY
Received 4 May 2018
Accepted 1 May 2019

KEYWORDS
Visualisation; insight; big data; sampling; error

1. Introduction

Due to the large amounts of data involved, big data visualisation authors are presented with the challenge of navigating a trade-off spectrum. Visualising big data presents a 'need for speed' (Wang, Wang, and Alexander 2015) as well as a need for accuracy. On one end of the spectrum is the option to visualise the entire dataset with the benefit of generating a highly accurate visualisation; however, this option also has high latency. On the other end of the spectrum is the option to visualise a smaller subset of the data, which not only has the advantage of reducing latency, but also has the effect of reducing the accuracy of the visualisation. The decision of where to land in this spectrum depends on several factors, including the system resources available, the objective of the visualisation, and the context of the visualisation application. These issues are particularly salient in scientific simulation visualisation applications, where generating larger datasets via simulation takes substantial time.

We propose that the decision of where to land in the trade-off spectrum should be left to the visualisation end-user, as the determination between speed and accuracy is a fluid one that could change even between successive runs of the same visualisation. This could happen when speed and accuracy demands change due to factors such as an iterative work flow or the uncovering of previously unknown information during prior runs of the visualisation. Ensuing questions become (i) how do visualisation authors enable system users to make this decision and (ii) how do system users know if they are making the right decision? We believe that the answers to the questions above lie in sampling and providing interactive visualisation accuracy feedback to the end user.

When controlling the accuracy versus speed trade-off, users need the ability to make an informed decision. A feedback measure is needed to help inform the user regarding the impacts of using a given accuracy or speed value. An obvious choice for a measure that can lead to informed decision making would be a statistical measure like confidence interval. However, such measures can be difficult to relate to visualisation. For example, when one is attempting to identify regions of high wind velocity in a geoscience visualisation, an accuracy value of 30% is not intuitive. Are the glyphs positioned inaccurately by 30%? Are the depictions of the wind velocity inaccurate by 30%? Since the purpose of visualisation is insight (Card, Mackinlay, and Shneiderman 1999), the use of an insight based metric is more intuitive, but statistical measures are more suitable for describing the performance of sampling operations. What is needed is an approach that takes statistical measures and brings them to the visualisation user domain.

CONTACT Moeti M. Masiane moeti@vt.edu Virginia Tech, Blacksburg, VA 24061, USA

Our ultimate goal is to provide a model that allows for the speedup of big data visualisation. Any speedup of data visualisation is pointless if the visualisation provides no insights. At the same time, any visualisation is wasteful if it provides insights that exceed human insight processing capabilities. What is needed is efficient big data visualisation speedup that does not provide excessive insights. Any big data visualisation speedup should be evaluated based on its benefits in terms of insight. If sampling provides big data visualisation speedup, and one can determine the relationship between sample size and insight quality, then one can use human provided insight requirements to determine an efficient sample size. This work models the relationship between insight quality and sample size. A model of the human insight versus sample size relationship in conjunction with a slider for the input of human insight requirements can then be used to speed up big data visualisation. In addition to visualisation speedup, this approach is based on interactivity and this has well-known benefits (Holzinger 2013). Our approach places primary focus on the human in the loop of big data visualisation. The human dictates the insight quality, and our model allows for this requirement to be converted to an efficient speedup requirement that provides just enough required insight quality. We work to understand how sample size impacts insight accuracy; in particular, how does sample size impact the accuracy of human estimates of the mean made from viewing histogram visualisations?

In this work, we provide a measure of the human-perceived mean error that relates the effects of sampling to insights obtained from histogram visualisations of big data. We proceed to relate our measure of insight to the well-known statistical measure of standard error of the mean. This enables a visualisation user to control the trade-off using an insight measure that they understand. The system then converts this insight measure into a statistical parameter that is better suited for our sampling algorithm. We provide a model that allows for the speedup of big data histogram visualisations by providing sample sizes that meet user-provided insight requirements. This work represents an initial step towards using insight-driven sampling to speed up the visualisation of big data. Future applications will involve more complex datasets, visualisations, and user tasks.

1.1. Contributions

As a result of our study of insight error associated with the analysis of histogram visualisations, we make the following contributions:

C 1. Measure and define the relationship between insight accuracy and sample size in histograms (see Section 6.6)

C 2. Using simple histogram visualisation-based tasks, we demonstrate the relationships between insight error and (i) its components, (ii) sample size, and (iii) standard error of the mean (see Sections 6.3, 6.4, 6.5, 6.8.2, 6.8.3 and 6.8.4)

> C 2a. We measure and define the relationship between sample accuracy, histogram visualisation accuracy, perception error, and insight error (see Section 6.8.1)

C 3. We use standard error to predict insight accuracy (see Section 6.8.5)

C 4. We provide a model that allows for the speedup of big data visualisation by providing sample sizes that meet user-provided insight requirements (see Sections 6.8.7)

2. Background and related work

The use of sampling introduces accuracy and uncertainty considerations into information visualisation and these could lead to error in any decision making that occurs as a result of viewing such visualisations. In other words, insight levels generated from a visualisation can be impacted by sampling. Any visualisation system that introduces error brings about the question of trust to system users. Questions like 'how accurate is this visualisation?' and 'how does the error impact my results?' need to be addressed by system authors.

A visualisation system that no one trusts provides no value, even if it provides valuable insights. The related accuracy, uncertainty, and error information that can impact insight levels need to be presented to end users of these visualisations so as to develop and maintain trust between users and visualisation applications, and to enable users to know the limitations of the visualisation application. For example, error bars can be placed on the bins of a big data histogram visualisation to provide feedback of the uncertainty associated with a visualisation (Zgraggen et al. 2017). Other research has been conducted into these areas (Chen et al. 2015; Sacha et al. 2016; Liu et al. 2017; Grtler et al. 2018).

2.1. Insight

In this paper, we use Saraiya et al. and Choe et al. definition of insight that describes insight as an ' individual observation about the data, a unit of discovery ' (Saraiya, North, and Duca 2004; Choe, Lee, and et 2015). Examples of insight that can be obtained from a numeric dataset include the mean, median, mode, and range. In

addition to the definition of insight is the idea that insight can have various levels that can be quantified. To explain the idea of quantifying levels of insight, we consider two human migration datasets. The first contains human migration data sampled monthly for a decade, while the other contains the same information sampled daily for a decade. The latter is likely to have a higher level of insight because it enables a more detailed understanding of migration trends, beyond the information that can be obtained from the former dataset.

Measuring the level of insight can be done qualitatively and quantitatively (Rojas et al. 2017; North 2006; Chang et al. 2009). We use a quality measure that is based on the error associated with the insight derived from a visualisation in comparison to a ground truth.

2.1.1. Measuring insight

The idea of measuring insight is one with major implications on the effectiveness of visualisation approaches. Yi et al. research the idea of how insight is gained in information visualisation systems (Yi et al. 2008). During the analysis of their findings, they allude to the idea of higher level insights, meaning that some insights are better than others. The main focus of their work is to identify the processes involved in the formulation of insight. Our work, on the other hand, is concerned with measuring different levels of insight and defining their relationship to sample sizes. Our use of levels of insight has the same intuition as the use of a Likert scale in the Insight orientation scale (Gori et al. 2015).

Rojas et al. measure insight in an attempt to evaluate the effectiveness of various sampling techniques on big data (Rojas et al. 2017). Similarly to our work, they compare the results of visualisation-based tasks to a ground truth and use the results of the comparison to evaluate the insights. They propose that insights can be evaluated quantitatively and qualitatively, and they use open-ended tasks to do so. Their objective is different from ours in that they are concerned with the performance of sampling techniques while our approach is mainly focused on measuring and quantifying insight. Their choice of more complex user tasks makes the evaluation heavily reliant on human analysts, and this introduces bias to the analysis. Our analysis, on the other hand, is objective because it relies on simple tasks that have an objective ground truth.

2.2. Effects of sampling

Sampling as a method for reducing the size of big data is effective, but it also leads to information loss. This loss of information is important because it introduces error into the decision-making process of big data analytics. In the information visualisation world, we care about this error because we are interested in how this error impacts the insights that we generate from a visualisation, which in turn impacts decision making.

2.2.1. Effects on statistics

The effects of sampling can be measured and described using statistical measures like confidence interval and z-values for how close a sample statistic is to the population parameter. As mentioned earlier, the challenge with statistical measures is that they do not necessarily translate to visualisation-based measures. For example, consider a scenario (S1) where one samples data and generates a visualisation of quantitative data. If statistical tests show that the mean of the sample data has a relative error of 25%, does this mean that insights generated from a visualisation of this sample will be 25% inaccurate? What does it mean to have insights that are 25% inaccurate? The answers to these questions depend on the visualisation, user task, and amount of data being visualised. If one is visualising the mean of the data as a bar in a chart, and the insight being generated is the height of the bar, the inaccuracy of the insight generated could be close to 25%, and that could be substantial. On the other hand, if the data consists of a more complex visualisation of wind speeds at a given time and place, and the insight being generated is the identification of regions where gusts exceed a certain threshold, the insight inaccuracy associated with a 25% error could be negligible. What is needed is a visualisation-based measure that is applicable to a wide range of statistical measures, visualisations, user tasks, and datasets. A measure of insight level that quantifies what the end user learns from a visualisation would meet this requirement.

2.2.2. Effects on insight

The idea of a measure of the level of insight when evaluating the performance of a visualisation is intuitive and for that reason we propose the use of insight error as an instrument for feedback during the system user controlled 'accuracy versus speed trade-off' process. While leveraging sampling to improve visualisation performance, the aim is to avoid compromising the visualisation quality. That being said, we also need to avoid superfluous quality if it negatively impacts performance. This is the same idea behind approximate query processing (Lin et al. 2018; Kulessa et al. 2018), and the same idea in reverse that is used in privacy preservation, where sampling rate is used to reduce the insight level in cases where the insight in question is a person's identity (Xiao et al. 2018).

Through measuring the insight error associated with various sample sizes, we can present the visualisation user with system controls that allow the selection of an arbitrary amount of insight error, which in turn will select the sample size and the resultant visualisation speed. The underlying idea is that gaining insights takes precedence over processing the entire dataset (Kaisler et al. 2013), but efficiency matters. If a system user can generate quality insights from a small subset of the data, then the user should be made aware of that information. The goal is for the system user to make an informed decision as to his or her desired amount of insight error based on how the insight error impacts the speed of the visualisation.

We adopt a qualitative approach with the idea of using error to quantify such insights. The delta between an insight value and the ground truth is the error that can be used to compare similar insights and thus signify the insight level (Figure 1). As in many evaluations of visualisations, we begin by using simple user tasks in such evaluations (North 2006). The intuition is that better visualisations result in better insight levels, and as a result, methods for evaluating visualisations are in essence methods for evaluating levels of insight associated with a visualisation. An example of better visualisations leading to better insights in the context of histograms could involve two histograms of normally distributed data, where one has 2 bins and the other has 10. The former could give the inaccurate insight of the data being uniformly distributed, while the latter could show a more accurate representation of the underlying distribution. Following this example, an even higher number of bins would give an even better insight into the distribution of the data being visualised.

Using a crowd study, we investigate these relationships. Leveraging simple tasks consisting of estimating the mean of a visualised sample, we compare the results of the simple tasks to the ground truth generated using statistical methods and assign error to each estimation. This error represents the quantification of the quality of the insight generated from the visualisation. The error values associated with the different sample sizes over a wide range of iterations allow us to measure the effect of sampling on insight. This in turn allows us to present insight error as intuitive feedback to the visualisation system users as they decide on an appropriate sample size for their visualisation. This insight level measure is not only appropriate for a visualisation context, but is also backed by sound statistical concepts.

2.2.3. Sampling in big data visualisation

Sampling is used widely in big data visualisation (Park, Cafarella, and Mozafari 2016; Hong, Hwa, and Kyung 2018). While random sampling is mostly used (Rojas et al. 2017), cluster (Nguyen and Song 2016) and systematic (Berres et al. 2017) sampling are also used to reduce big data for the purpose of visualisation (Leetaru 2019). In imMens (Liu, Jiang, and Heer 2013), Liu et al. explore the use of sampling in visualisation and provide references of other works that utilise sampling to reduce the volume of big data. Liu et al. argue that various types of sampling have disadvantages that include the need for preprocessing, the exclusion of outliers, and in the case of random sampling they could still result in a sample that is too big to visualise. As a result, they choose binning as a method for big data volume reduction. Their arguments against sampling can be summarised as (i) sampling introduces error and (ii) big data sampling is not guaranteed to result in small data. We agree with their arguments, but do not agree that these points are hindrances to the visualisation of big data. Our approach with regard to their first argument is to present the sampling error to the visualisation system user and allow the user to control that error. With regard to

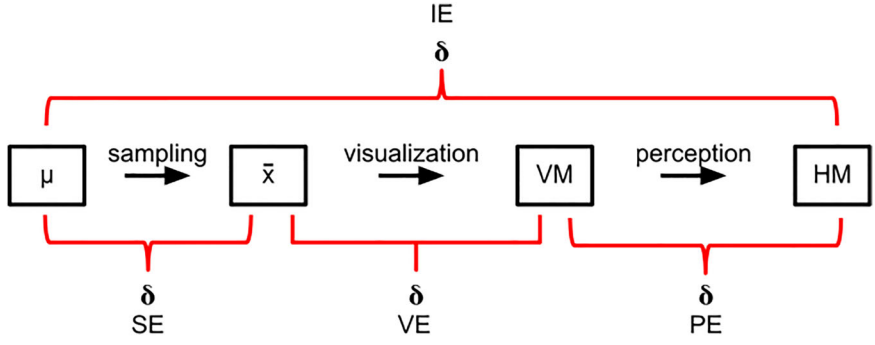

Figure 1. Error flow diagram. The delta (δ) between the population mean (μ) and the sample mean ($\bar{x}$) is the sampling error (*SE*), while the delta (δ) between $\bar{x}$ and the visualisation mean (*VM*), which is the mean that can be calculated from the visualisation, is the visualisation error (*VE*). The delta (δ) between *VM* and the human mean (*HM*), which is the mean that is perceived by the human, is the perception error (*PE*) and the delta between μ and *PE* is our insight error (*IE*).

their second argument, we also propose to give the visualisation user control of the sample size along with relevant information that will allow the user to make an informed decision. The visualisation system user is empowered to make an informed decision on the best position to place the slider for the size and performance (speed) trade-off.

Ruan et al. (Ruan et al. 2017) use a novel sampling technique that can be thought of as random sampling with weighting to reduce the volume of big data in their intrusion detection system's big data visualisation. Their novel sampling technique is aimed at reducing the sampling error by allocating higher weights to data instances with fewer occurrences in the dataset. After applying their technique, they acknowledge that the sampling error still exists, but do not attempt to quantify this error. We on the other hand do not attempt to produce an error free sample. We expect error to exist, hypothesise that it can be controlled, provide the system user with the means to control it and provide feedback regarding the effects of adjusting the error on the visualisation.

2.3. Uncertainty in visualisation

Pang et al. surveyed and classified uncertainty visualisation methods (Pang, Wittenbrink, and Lodha 1997). They identified three sources of uncertainty in the visualisation pipeline as (i) data gathering, (ii) transformation and (iii) visualisation. They proceed to create a taxonomy and classification of uncertainty visualisation techniques. While they are concerned with how uncertainty is visualised, we are concerned with its existence and how it impacts insight. We incorporate this uncertainty into our insight level measures, which we can present to the end user so they can use it to control the accuracy versus speed trade-off.

Tolerating uncertainty for the effect of reaping performance benefits extends beyond visualisation applications to systems and applications that provide data to visualisation systems. IDEA and Google's Dremel are two such systems (Galakatos et al. 2017; Melnik et al. 2010). IDEA treats current and previous query responses as random variables, and uses standard error based metrics to determine the accuracy of current results. Based on this accuracy, IDEA determines whether or not to rewrite an issued query and provides the most accurate response to the visualisation system. Unlike our approach, IDEA does not provide accuracy-related user feedback. Similar to our work and IDEA, Dremel uses the idea of a trade-off between accuracy and speed. Some of their queries return approximate results. They query many tables and use a quality parameter to determine the percentage of relevant tables to query. Their work does not state how they select this quality parameter, but they mention that they sacrifice accuracy for performance. The difference between the leveraging of uncertainty in these works and ours is that we offload the decision concerning the quality of the sample onto the visualisation system user because we believe that this decision is dependent on fluid factors that each user can evaluate best.

Progressive visualisation is another technique that leverages uncertainty to improve visualisation speeds (Fisher et al. 2012; Turkay et al. 2017). Similar to our work, progressive visualisation answers the question of how much data is enough to make analytic decisions? We agree that for a visualisation there is not one answer and as a result, the user is in a better position to know based on the visualisation context. In progressive visualisation, data is loaded using small batch processes and visualised incrementally until the visualisation system user stops it or all the data is loaded (Fekete 2015; Moritz et al. 2017). The general idea is that the visualisation will converge to a solution that is satisfactory to the user. We argue that the strength of this approach does not lie in increasing the speed of a visualisation but lies in keeping the user engaged while a satisfactory visualisation is being generated.

Relying on the user to decide when the visualisation is satisfactory is based on the assumption that the user knows what they are looking for and will be able to identify it as soon as they see it. In exploratory tasks, this may not be the case. We follow the approach of supplying what we deem an appropriate amount of data to satisfy a user-controlled insight level and let the user explore the data. We believe that this is a better approach for scenarios where the user is unsure of what they expect to find. That being said, the idea of progressive refinement is one with many valid applications, and it can be used in tandem with our approach. For example, the progressive refinement technique is applicable in cases where the sample size proposed by our approach is large, or in scenarios where the user prioritises accuracy over speed but still needs to remain engaged with the system.

3. Methodology

The ultimate goal of this work is to allow the speed up of big data visualisation by allowing system users to decide the size of the sample to visualise in order to obtain an arbitrary insight quality. We realise this is a challenging

task, and hence we explore it in the context of histograms as a proof of concept. Our insight quality is measured in terms of insight error (IE). In order to measure and predict IE, we gather data using a crowd study and use it to test our hypothesised IE component errors and their behaviour. To do this we measure IE as the normalised difference between a human estimate and the ground truth. We learn and model the relationships between IE and its component errors, with sample size. We also learn the parameters of an IE predictive model using its component errors, and those for predicting IE using standard error of the mean (SEOM). SEOM is used as a proxy for IE component errors because it is a well-studied statistical measure and due to our hypothesised relationship between SEOM and the component errors. We proceed to evaluate our models using data both seen and unseen during training.

3.1. Research questions

This research seeks to identify the effect of data sample size on visualisation insight quality. Insight quality is measured in terms of insight error. In particular, this work addresses the following research questions:

R 1. How does insight error (*IE*) and its components (i) sampling error (*SE*), visualisation error (*VE*), and perception error (*PE*) behave as a function of sample size? (see Sections 6.3, 6.5, 6.6, and 6.8.3)
R 2. How do *SE*, *VE*, and *PE* contribute to total *IE*? (see Section 6.8.1)
R 3. How well can a statistical measure that is a function of sample size be used to predict each ***IE*** component? (see Sections 6.8.2, 6.8.3, and 6.8.4)
R 4. Can a well-known statistical measure that is a function of sample size be used to estimate *IE* in place of the *IE* components? (see Section 6.8.5)
R 5. Can we predict sample size as a function of *IE*? (see Section 6.8.7)

3.2. Hypotheses

In an attempt to understand the relationship between sample size and insight error, we focus our research on studying this relationship in the context of insights gained from histogram visualisations. We chose to use histograms in our study because they are commonly used in exploratory data analysis (Macke et al. 2018) and because they are easy to understand visualisations that highlight important data attributes like distribution, range, mean, and other central tendency properties. We also focus on the user task of visually estimating the mean of the data, as an initial simple representative of many types of insights that users want to gain from histogram visualisations. Our hope is that our results will serve as a proof of concept and starting point to generalise to other insight-related visualisation tasks. With respect to the research questions stated above, we hypothesise the following:

H 1. Using the error between estimates generated from a histogram visualisation of sampled data and actual parameters from the full dataset, we can measure the insight error and its components. We expect insight error to decrease with an increase in sample size with a behaviour similar to that of exponential decay.

> H 1a. The insight error measure depends on the quality of the sample, the effectiveness of the histogram visualisation, and the perception abilities of the end-user viewing the visualisation.

H 2. Calculating the accuracy of statistics derived from our samples in comparison to population parameters will give us a metric that behaves similarly to a well-known error measure like standard error that is a function of sample size.
H 3. The relationship between standard error and size is similar to that of our component errors and size. We should be able to use standard error to predict our component errors.
H 4. Standard error can be used to predict the insight error.

> H 4a. Our insight error and standard error quality measures are positively correlated.

H 5. Given our insight metric, one can predict a corresponding sample size.

The ability to predict our insight error measure from standard error of the mean means that we can employ this approach of using insight error to drive sampling, leading to the ability to visualise large datasets without having to run a user study for each dataset being analysed. We can use standard error of the mean to determine insight level as long as the dataset shares the same properties as a dataset that has previously been studied.

4. Insight model

We hypothesise that the insight error can be measured (H1), and that the measure is impacted by sampling, visualisation, and perception effects (H1a). Using a crowd-sourced study, we measure these error

components, and then subsequently define a multiple linear regression model for the relationship between the insight error and perception, sampling, and visualisation effects, before finally learning the parameters for our model. Using the insight error to predict the sample size, which in turn controls the perception, sampling, and visualisation effects, will allow us to provide visualisation end users with a slider that they can use to control the accuracy versus speed trade-off. Such a slider would allow the user to input an arbitrary insight level, and our model would then provide a sample size that meets this user provided insight requirement, thus providing a statistically sound speedup of the visualisation of big data. However, since we are unable to know the insight error components for datasets that we have not studied, it is important to find a proxy for insight error or any of its components that is easy to calculate. We use standard error of the mean as our proxy, and we use our study results to learn the relationship between insight error, its components, and our proxy. Our slider will allow the visualisation system end user to control the insight level, which will control the sample size and subsequently the visualisation, perception, and sampling errors.

4.1. Assumptions

In this work, we assume that samples are being drawn from a finite population. We foresee cases where these samples would be drawn from samples of larger or infinite populations, and in those cases our samples would be considered as subsamples. We expect our approach to still be applicable in those cases.

4.2. Definitions

In our investigation of insight error ***(IE)***, we hypothesised the existence of components of ***IE*** and defined the relationships of these components. The components are (i) sampling error (*SE*), (ii) visualisation error (*VE*), and (iii) perception error (*PE*).

Central to the definitions of our components is the concept of error, which is the difference between the expected and observed values. Due to the different magnitudes of our expected values as a result of using various populations and samples with different ranges and means, using error values as described above can lead to misleading insights. To mitigate this issue, we normalise our errors based on the range of the underlying sample. We use the range of the data that our participants see to normalise their responses onto a common scale.

4.2.1. Population and sample

$$d_N = \{x \mid x \in R\} \tag{1}$$

$$d_N \sim \alpha(\mu,\sigma) \tag{2}$$

$$d_n \subseteq d_{\mathrm{N}} \tag{3}$$

$$d_n \sim \alpha(\bar{x},s) \tag{4}$$

A population (d_N) consists of N real numbers (Equation 1). The population has a distribution α, a mean μ, and a standard deviation σ (Equation 2). A sample is a subset of the population (Equation 3), and it has a distribution α, a mean $\bar{x}$, and a standard deviation s (Equation 4).

4.2.2. Histogram

$$\text{histogram} = \{\text{bin} \mid \text{bin has a width and frequency}\} \tag{5}$$

$$f{:}d_{\mathrm{n}} \rightarrow \text{histogram} \tag{6}$$

A histogram consists of bins and each bin has a width and a frequency (Equation 5). Using D3.js, we generate a histogram from a sample (Equation 6).

4.2.3. Population, sample, visualisation, and human means

$$\mu = \frac{1}{N}\sum_{i=1}^{N} x_{\mathrm{i}}, \text{ where } N = \text{population cardinality} \tag{7}$$

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_{\mathrm{i}}, \text{ where } n = \text{sample cardinality} \tag{8}$$

$$VM = \frac{\sum_{i=1}^{b} \text{frequency}(\text{bin}[i]) * \frac{\min(\text{bin}[i]) + \max(\text{bin}[i])}{2}}{\sum_{i=1}^{b} \text{frequency}(\text{bin}[i])}, \tag{9}$$

where b = number of bins in the visualisation
HM = mean estimate provided by a human looking at a histogram visualisation

The population mean (μ) is calculated as the average of the values of all instances in the population (Equation 7), while the sample mean ($\bar{x}$) is calculated as the average of the values of all instances in the sample (Equation 8). The human mean (HM) is the estimate of the mean provided by a human as a result of analysing a histogram (Equation 10). The visualisation mean (VM) is an objective calculation of the mean of the bins in a histogram as a result of summing the products of the middle x-value and the height of each bin and dividing this sum by the sum of all bin frequencies (Equation 9).

4.2.4. Sampling error (SE)

Sampling introduces error, and as a result, the visualisation of sampled data can only lead to generated insights that are at best as good as the quality of the sample. If as the result of sampling bias, all the data from our S1 scenario (Section 2.2) happens to fall to the left of the mean, an accurate estimate of the population mean using this sample will fall to the left of the true population mean (Figure 2). The impact of the quality of the sample on the insight error generated from a visualisation of that sample is referred to as the sampling error (*SE*).

$$SE = \frac{|\mu - \bar{x}|}{\text{range(sample)}} \tag{11}$$

Sampling error is defined as the absolute difference between the actual population mean and the sample mean, normalised by the range of the sample (Equation 11). The *SE* for a given sample size is a distribution of errors for that sample size. For the purpose of sample size prediction, we represent this distribution with the mean of the observed errors for that sample size.

4.2.5. Visualisation error (VE)

After the sample is chosen in our S1 scenario (Section 2.2), certain considerations need to be made before the data is visualised. Decisions like the need for data transformations, the type of visualisation used, the data encoding schemes, the type of data scales, the need for data binning, etc. contribute to the effectiveness of the visualisation. For example, binning data in a histogram could lead to a loss of accuracy based on the width of the bin. In other words, wider bins group a wider range of data and this results in a loss of accuracy due to the generalisation of the binned data. Two different visualisations of the same data could lead to different levels of accuracy, which in turn could lead to different levels of insight generated by the same end user. The impacts of the visualisation choices on the insight levels generated are referred to as the visualisation error (*VE*).

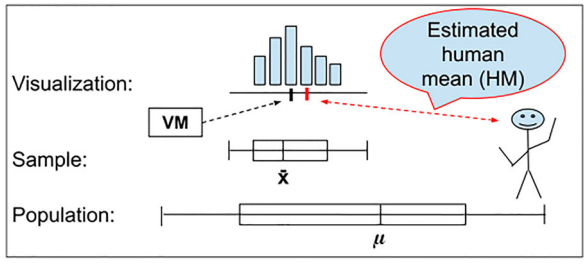

Figure 2. Component errors. If a sample drawn from a population is visualised and the visualisation is used to estimate the true population mean (μ) an accurate human estimate of the mean can only fall within the range of the sample. The difference between the sample mean ($\bar{x}$) and population mean **(*μ*)** is the sampling error, while the difference between the objective mean of the visualisation (*VM*) and the sample mean is the visualisation error. Any difference between the visualisation mean and the perceived mean of a human (*HM*) viewing the visualisation is the perception error. Insight error which is a measure of the how well the human estimates the actual mean is impacted by the quality of the sample, visualisation, and human perception.

$$VE = \frac{|VM - \bar{x}|}{\text{range(sample)}} \tag{12}$$

Visualisation error is defined as the absolute difference between the sample mean and the mean of the visualisation, which can be calculated using the middle value of each bin width and its height, normalised by the range of the sample (Equation 12). Similar to the sampling error, the *VE* for a given sample size is a distribution of errors for that sample size. For the purpose of sample size prediction, we represent this distribution with the mean of the observed errors for that sample size.

4.2.6. Perception error (PE)

The visualisation created in our S1 scenario (Section 2.2) could be interpreted differently by different end users. This is due to the variation in human perceptive capabilities. We take this variation into consideration when hypothesising our model and refer to this variation as the perception error (*PE*).

$$PE = \frac{|VM - HM|}{\text{range(sample)}} \tag{13}$$

Perception error is defined as the absolute difference between the estimate of the mean provided by a human and the mean of the visualisation, which can be calculated using the middle value of each bin width and its height, normalised by the range of the sample (Equation 13). Our *PE* is also a distribution of errors as opposed to a single error value and for the same reasons as those given for the *SE* and *VE*, we represent this distribution with the mean of the observed errors for a given sample size.

4.2.7. Insight error (IE)

Insight error (*IE*) is a measure of how accurate the insights that are gathered from a visualisation of sampled data are in relation to the ground truth. Our *IE* is the difference between an estimate made by a user viewing a visualisation (*HM*) and the actual population parameter (μ).

$$IE = \frac{|\mu - HM|}{\text{range(sample)}} \tag{14}$$

Insight error is defined as the absolute difference

between the estimate of the mean provided by a human and the actual population mean, normalised by the range of the sample (Equation 14). For the same reasons as those given for the *SE*, *VE*, and *PE*, we can think of *IE* as the mean of a distribution of errors as opposed to a single error value.

4.3. Model details

We define our model of *IE* as consisting of the *SE*, *VE*, and *PE* and hypothesise that our model is a multiple linear regression model that can be used to predict the insight level. We hypothesised that our model would follow the following form:

$$IE = \beta_0 + \beta_1 * x_1 + \beta_2 * x_2 + \beta_3 * x_3 + \varepsilon,$$
where $\beta_{0..3}$ are parameters and $x_{1 \cdots 3}$ are regressors (15)

Insight error can be modelled as a function of our sampling, visualisation and perception error regressors and parameters that can be learnt from training data (Equation 15). Using data gathered from a crowd study, we learn the parameters of our model and evaluate the model (Section 6.8.1).

5. Crowd study

Our study used a blocked design with replication. The treatments consisted of dataset and sample size combinations and we blocked by study participant. With 20 datasets and 20 sample sizes, we had a total number of 400 treatments. Due to the large number of treatments, we went with a design that limited our treatments to 20 by limiting each participant to seeing each dataset once and seeing each sample size once. This reduced the number of tasks each participant had to complete and avoided learning that would impact the responses of future tasks based on information learned from prior interactions with a dataset. Participants were randomly assigned to one of the 20 treatments. Our treatment design consisted of a table with an axis of 20 values representing each of our 20 datasets and an axis of 20 values each representing one of our 20 sample sizes. Each of the 20 rows of 20 values within our table represented a block and each block consisted of the order and combination of the treatments each participant encountered. Each treatment had 3 replications and the 60 participants were randomly assigned to blocks.

5.1. Participants

We had two trial studies before conducting our final study, as we fine-tuned the procedure of the study. Our final study had 60 participants whose identity, sex, and age are unknown because they were crowdworkers. They were paid an average of about $3 for about 15 minutes of work. Their online identities were unique, guaranteeing that there were no repeat participants. One of the user tasks was only used for quality control, while the other 20 were included in our study results. This quality control task in conjunction with our clear instruction stating that we would not pay for unsatisfactory results and our criteria for crowd workers who had never had their work rejected for bad quality in the past ensured that we got quality results. We provided participants with an email to submit any feedback and only heard from 1 who did not receive a confirmation code after submitting his results.

5.2. Data

We inspected various probability distribution plots and created a collection of datasets to use based on the shapes of their distribution curves, with the intention of providing heterogeneity amongst our datasets. We generated 20 synthetic datasets using the R programming language distribution functions. The distributions of the datasets we selected varied (Table 1), and within the same distribution, the means, ranges of values, skewness, and kurtosis were also varied.

5.3. Procedure

Only Amazon Turkers who had never had their work rejected for poor quality on Amazon Web Services (AWS) were allowed to participate in the study. Participants were each provided with a URL that led them to a page with a survey. The survey, created using HTML, D3.js, and CSS, showed a sequence of histograms using data sampled from a larger population. For each task, participants were asked to estimate the mean of the data shown in the histogram (Figure 3). After each estimation task, the task was repeated with a different histogram. To control the quality of responses, participants were informed that quality of results would be reviewed and payment would only be issued for results exceeding a

Table 1. Distribution of the datasets that are used in the study. Each dataset is sampled and all participants encounter each dataset once and each sample size once.

Distribution	# of Datasets
Chi-Square	3
Exponential	3
Left Skewed	5
Normal	3
Right Skewed	5
Uniform	1

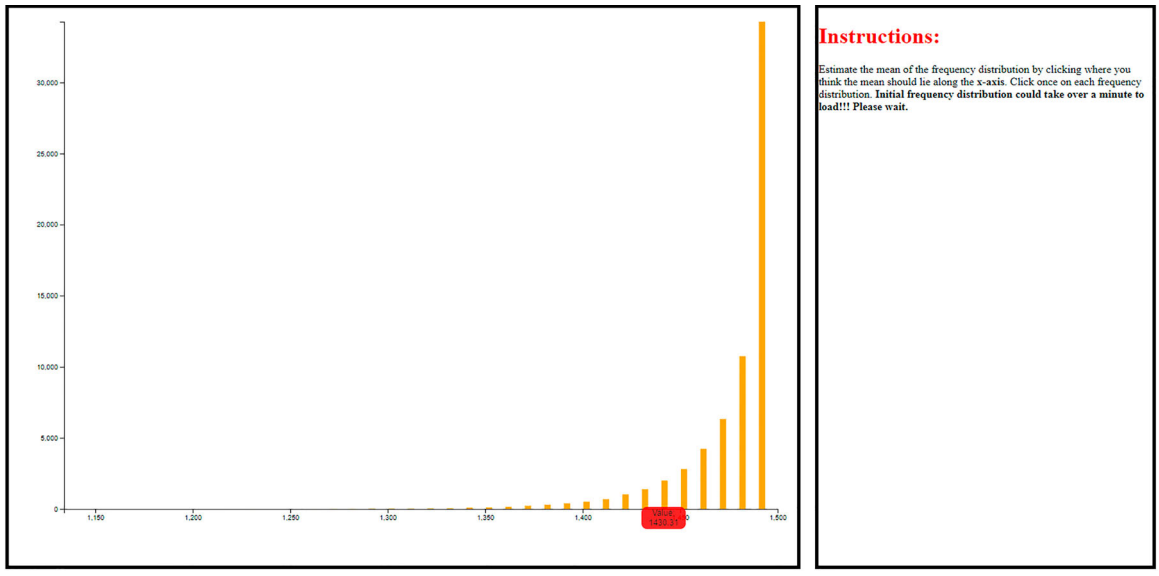

Figure 3. A screenshot of our survey showing a single task. The participant is shown a histogram and asked to estimate the mean value. Participants can select any value in the histogram. Values coincide with the current mouse X position and the red label gives feedback as to the value that is currently selected. An estimation is made using a simple mouse click. Due to the fixed bin width, our visualisation like any other introduces error (*VE*) between the mean of the sample (x) and the mean that can be calculated from the histogram (*VM*). We capture this error (*VM*) and account for it in our model.

certain quality, and a quality control histogram was added to each treatment.

There were 21 tasks for each participant, and one of these tasks was for quality control. These tasks consisted of the basic visualisation task of estimating the mean of the visualised histogram. All of the tasks consisted of an identical question about a histogram visualisation, but the data being visualised was different for each task. The question asked for each task was 'Estimate the mean of the histogram by clicking on where you think the mean should lie along the x-axis'. The quality control task consisted of a histogram with two bars whose mean was easy to estimate. None of the responses showed a large error on the quality control estimate, and thus none were rejected. The highest error was a normalised root mean square deviation value of 0.0126.

To prevent any learning effects, participants saw each dataset once. Using random sampling with replacement, each dataset was sampled at 20 sample sizes, and each participant saw each sample size once. Sample sizes of 2^n were used, where n was an integer in the range [1,20]. A treatment was considered to be a histogram visualisation of a combination of a dataset and a sample size. We randomised the order that treatments were presented to each participant. Participants were randomly assigned IDs in the order that they signed up to participate in the study.

After each histogram loaded, participants would see a label on the histogram showing an estimated mean value along the x-axis for the current cursor position. Moving the cursor would update the x-value that would also move to stay aligned with the cursor position. Any location on the histogram could be clicked to provide an estimated mean value, including the space between bars for tasks that consisted of sparsely populated histogram bars. The selected value, as well as the properties of the histogram including the dataset name, size, standard deviation, sample size, sample mean and population name, mean and standard deviation, were stored in a MariaDB database along with the user ID and a per-user 13-character hexadecimal code that was provided to the participant at the end of the survey in order to confirm survey completion. Each participant had to enter their unique completion code into Amazon Mechanical Turk in order to receive payment.

6. Results and evaluation

We hypothesised that our insight metric was composed of several errors (H1a) and that standard error of the mean (H2) could be used to predict our insight metric (H4). We observed and modelled each of these errors in relation to sample size and standard error of the mean, and generated both a predictive model for insight error given its component errors and a predictive model for sample size when given arbitrary insight and perception errors. We evaluated each of these models based on their performance on data that had been observed in training and also on their performance on data instances

specifically withheld for testing that the model had never processed before.

While building our model, we were interested in the impact that the distribution of the underlying data would have on our model. To investigate this, we used two methods: (i) we added a distribution predictor to our model and observed the behaviour of the R^2 and Cp values when this predictor was added to and removed from the model and (ii) we used ANOVA to determine if the skew of the underlying distribution had an impact on the insight levels.

6.1. Impact of data distribution on insight error

Initially, we built an MLR model that used our three component errors, and we observed an R^2 value of 0.9953 and a Cp value of 2.5689. We added a distribution predictor for the right, left, normal, and uniform skews of our datasets. When we added the distribution predictor, the R^2 value did not change, but the Cp value improved slightly to 4.3861. This modest improvement did not seem to warrant the use of a more complicated model, but to be sure we decided to conduct an additional ANOVA test.

This ANOVA test was conducted to determine if the insight levels for our four distributions that were blocked by sample size were significantly different at an $\alpha = 0.05$ significance level. The null hypothesis was that there was no difference in means of the samples drawn from the four distributions that were right skewed, left skewed, normal, and uniform, while the alternative hypothesis was that at least one of the means was different. We analysed this data as a Balanced Incomplete Block Design (BIBD), as the number of datasets that we used from each distribution was uneven. We had more *IE* information about left and right skewed distributions than we had about normal and uniform distributions. BIBD ANOVA mitigates this lack of balance by using LS means as opposed to means in the analysis. We obtained an *F*-value of 0.6415 and a critical *F*-value ($df_1 = 3$ and $df_2 = 57$) of 2.76643794. Since our *F*-value was less than our critical *F*-value, we fail to reject H_0 and conclude that the population distribution does not lead to significantly differences in insight levels. As a result, our models did not include predictors for distribution of the populations from which the samples were drawn.

6.2. Relationship between standard error and sample size

Having hypothesised that our insight metric would behave similarly to standard error of the mean (H4a) with values that grow inversely proportional to sample size, we experimentally reproduced the behaviour of standard error (Figure 4). Standard error showed exponential decay in relation to sample size, as expected.

6.3. Relationship between sampling error (SE) and sample size

SE was one of the component errors of our insight metric (H1a). We reproduced this error for all sample sizes and observed its behaviour. As expected, the error between the sample mean and the population mean was inversely proportional to the sample size. The error shows exponential decay behaviour (Figure 5). As the size of the sample increases, the mean of the sample becomes closer to the mean of the population. The quality of the sample improves with an increase in sample size. The sampling error behaves similar to standard error of the mean.

6.4. Relationship between visualisation error (VE) and sample size

Our *VE* also shows exponential decay behaviour in relation to an increase in sample size (Figure 5). This is a result of the bin width in our histograms becoming more representative of the data within each bin as the amount of data increases. Our study shows that the more data one adds to a histogram, the more accurate the histogram becomes at representing the data. As a result, our visualisation error behaves similar to standard error of the mean.

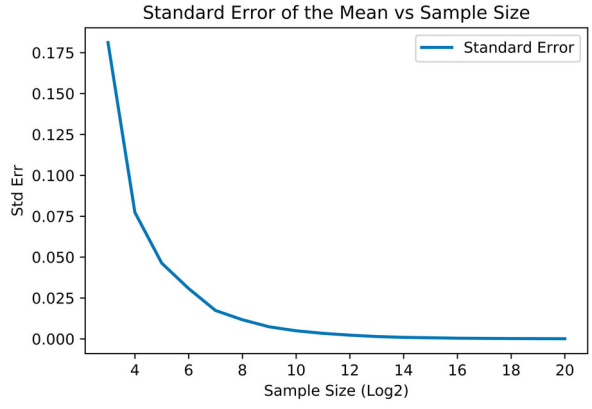

Figure 4. Observed standard error of the mean (SEOM) values plotted against sample size (Log_2) showing exponential decay. The smallest two sample sizes have been excluded from the figure to highlight the behaviour of SEOM for larger sample sizes. The SEOM values are normalised with the sample range as the denominator.

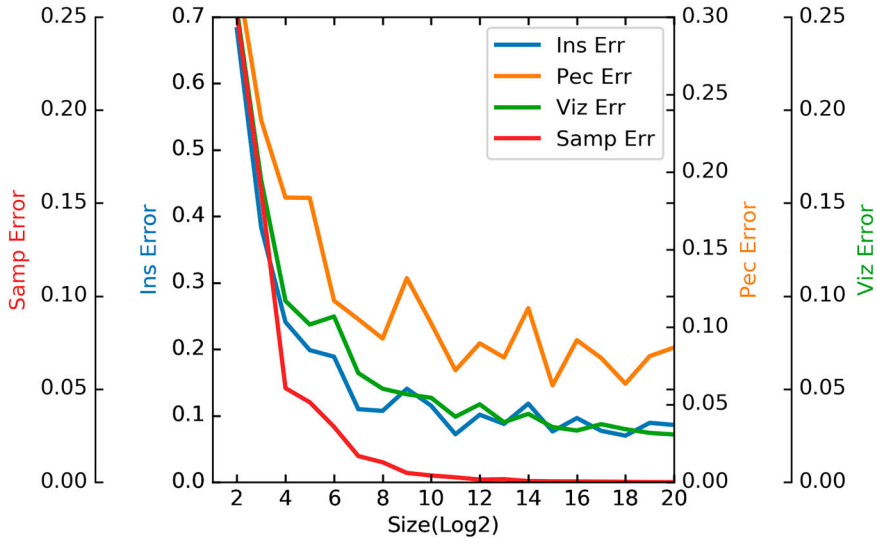

Figure 5. Observed Insight, Sampling, Visualisation, and Perception Error values plotted against sample size (Log_2) showing exponential decay. The smallest sample size has been excluded from the figure to highlight the behaviour of errors for larger sample sizes. These values are normalised with the sample range as the denominator.

6.5. Relationship between perception error (PE) and sample size

As expected, our *PE* decreases with an increase in sample size and exhibits the same exponential decay as our other component errors (Figure 5). This tells us that people understand histograms better as more data is added to the histogram, but there are diminishing returns. After some point, the rate of decrease of *PE* cools off even though the rate of adding data does not. In the world of big data, this could mean large amounts of processing power being used for no added benefit.

6.6. Relationship between insight error (IE) and sample size

Our *IE* behaved as expected in relation to sample size (Figure 5). This confirmed our hypothesis that *IE* would behave similar to standard error (H3). This relationship shows that as more data is added to a histogram, the quality of insights gained from the histograms improve even though there are diminishing returns in the relationship. This diminishing effects relationship has to be considered when one is visualising big data. As more data is added at a high computing cost, while potentially increasing application latency, there is a possibility that this could be for no added insight benefit.

6.7. Model parameter learning

In order to make the best decision on the appropriate sample size needed to attain an arbitrary *IE*, a relationship that accounts for the *IE* component errors is needed. This can be attained by learning the parameters of an MLR model that uses the *IE* component errors as predictors. Using the results of our study, we learned the parameters of models describing the relationship between sample size and our various errors, and evaluated these models. We split our crowd study results into a training and testing dataset using an 80–20 ratio. The data was randomised before splitting. 80% of the data was used to train our models while 20% was held out for evaluating the model. We did not create a validation set, but relied on the visualisation of the relationships between learned values and the actual observed values to help us decide on the best model to use.

6.8. Model evaluation

Due to the small error values that were close to zero, forecast error metrics like Mean Absolute Percentage Error that would seem intuitive to evaluate the performance of our model were unsuitable for this study. Such evaluation metrics are percentage based, and in our case resulted in large numbers due to very small denominators. As a result, we argue that plotting our model predictions provides sufficient evaluation.

6.8.1. Using component errors to predict insight error

Following R2, we proceeded to use all component errors to model insight error and our results were successful.

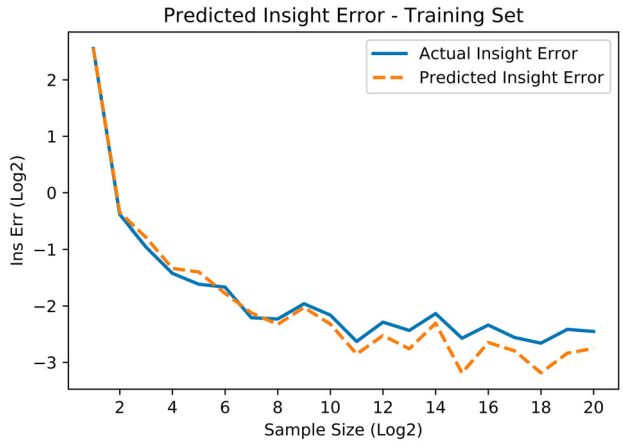

Figure 6. Forecasting insight error (*IE*) given insight error component errors from the training set. The actual errors seen in the training dataset **(solid)**, and those predicted by our model **(dashed)**.

Our learned parameters for our relationship were:

$$IE = -0.0474158 + 1.05784676 * PE + 0.7535806 * VE + 0.8509646 * SE$$

During the learning of these model parameters, we added a predictor for the distribution of the underlying data, but this predictor did not significantly improve the Rsq and Cp values for our model, nor did it largely degrade the performance. From this, we determined to keep the simpler model, deciding that the distribution of the dataset was not a requirement for predicting *IE* from component errors. The results of this model on the training data were good (Figure 6) and using the model on the held out testing set also showed excellent results (Figure 7). As successful as this model was, we

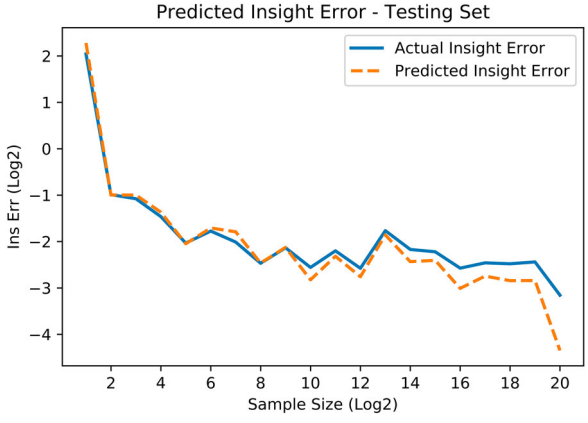

Figure 7. Forecasting insight error (*IE*) given insight error component errors on the testing set. The actual errors seen in the testing dataset **(solid)**, and those predicted by our model **(dashed)**.

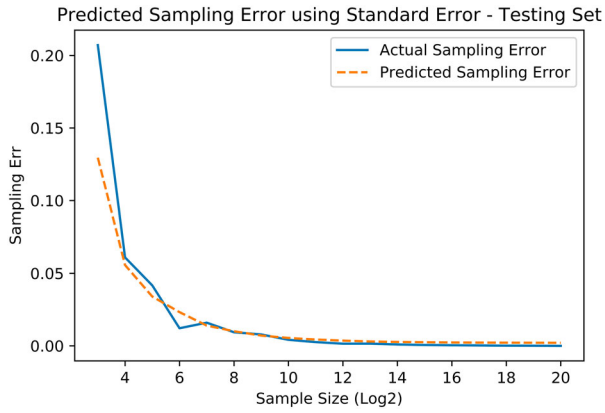

Figure 8. Forecasting sampling error (*SE*) when given standard error of the mean from the testing set. The smallest two sample sizes have been excluded from the figure to highlight the behaviour of SE for larger sample sizes. The actual errors seen in the testing dataset **(solid)** and those predicted by our model **(dashed)**.

sought a model that did not require a prior knowledge of the component errors. Such a model could be created by learning the relationship between the component errors and standard error, and using standard error as a proxy for these component errors.

6.8.2. Using standard error of the mean to predict sampling error

Modelling the relationship between sampling error and the standard error of the mean on the training dataset was very successful (Figure 8). The relationship was linear and the parameters for this model were:

$$SE = 0.097 * \text{std. error}^2 + 0.086 * \text{std. error} + 0.002 \quad (17)$$

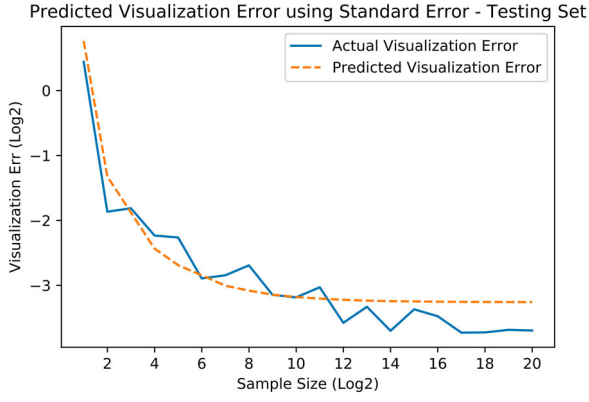

Figure 9. Forecasting visualisation error (*VE*) given standard error of the mean from the testing set. The actual errors seen in the testing dataset **(solid)** and those predicted by our model **(dashed)**.

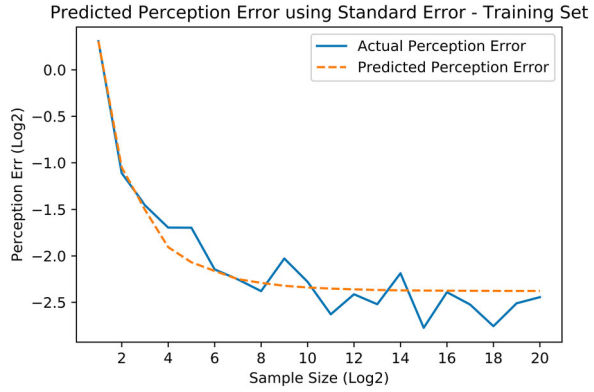

Figure 10. Forecasting perception error (*PE*) given standard error of the mean from the training set. The actual errors seen in the training dataset **(solid)** and those predicted by our model **(dashed)**.

6.8.3. Using standard error of the mean to predict visualisation error

Predicting *VE* using the training set also produced good results (Figure 9). Our linear model predicted values that exhibited a behaviour that was similar to the actual observed *VE*. The parameters for this model were:

$$VE = -0.0472*\text{std. error}^2 + 0.6405*\text{std. error} + 0.0384 \quad (18)$$

6.8.4. Using standard error of the mean to predict perception error

Our linear model to predict perception error based on the sample size was very successful on the training data (Figure 10) and our variables were positively correlated. However, the results on the unseen data (Figure 11) showed a need for improvement. It seems that we either need more training data or to try a different mode. More investigation on modelling how different people understand visualisations is needed in the future. The parameters for this model were:

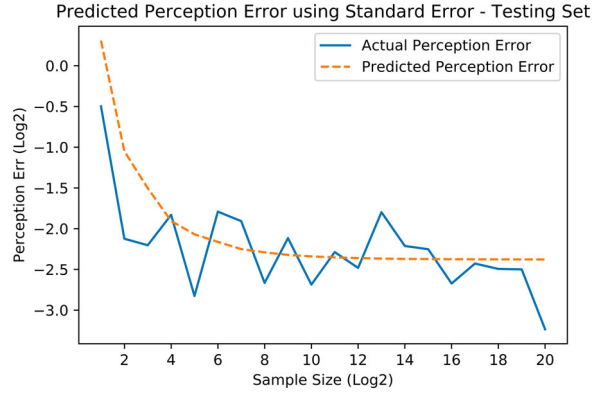

Figure 11. Forecasting perception error (*PE*) given standard error of the mean from the testing set. The actual errors seen in the testing dataset **(solid)** and those predicted by our model **(dashed)**.

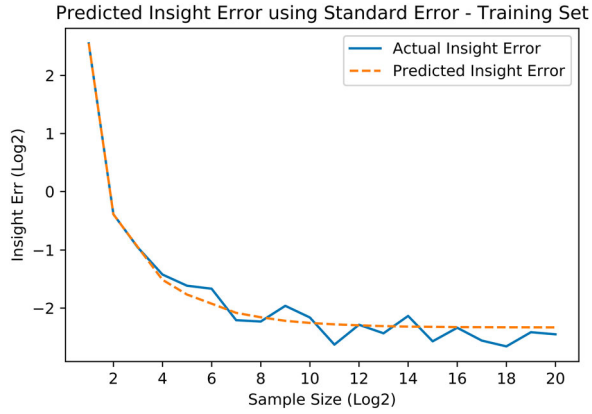

Figure 12. Forecasting insight error (*IE*) using standard error of the mean as a proxy for our component errors from the training set. The actual errors seen in the training dataset **(solid)** and those predicted by our model **(dashed)**.

$$PE = -0.07157*\text{std}^2 + 0.0.7297*\text{std. error} + 0.09271 \quad (19)$$

6.8.5. Using standard error of the mean to predict insight error

As hypothesised (H4) and also due to the fact that each of our component errors were positively correlated with the standard error of the mean, we built a model for the relationship for predicting insight error using standard error of the mean as a proxy for our component errors. This model showed good results for the training

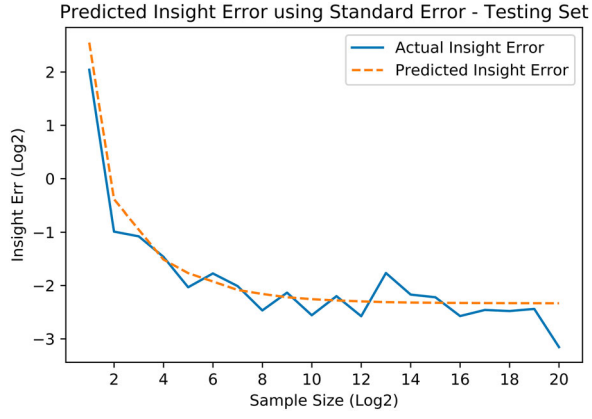

Figure 13. Forecasting insight error (*IE*) using standard error of the mean as a proxy for our component errors on data that our model had not seen in training. The actual errors seen in the testing dataset **(solid)** and those predicted by our model **(dashed)**.

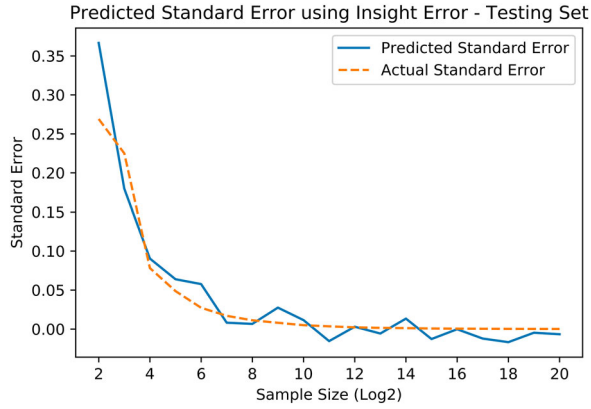

Figure 14. Forecasting standard error of the mean given insight error (*IE*) on data that our model had not seen in training. The smallest sample size has been excluded from the figure to highlight the behaviour of standard error of the mean for larger sample sizes. The actual errors seen in the testing dataset **(solid)** and those predicted by our model **(dashed)**.

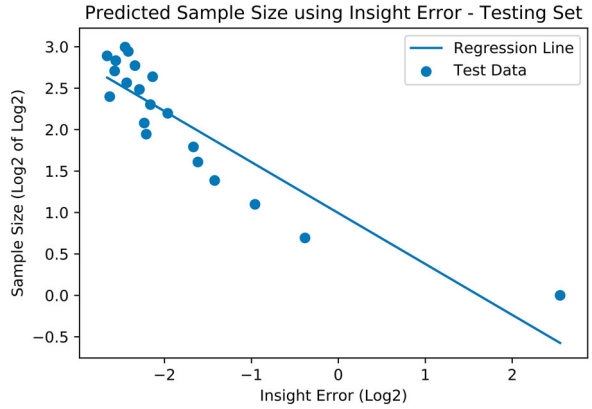

Figure 15. Forecasting sample size given insight error (*IE*) on data that our model had not seen in training. The actual errors seen in the testing dataset (circles) and our regression line (solid line).

(Figure 12) and testing (Figure 13) sets. The parameters for this model were:

$$IE = 1.595 * \text{std. error} + 0.097 \tag{20}$$

6.8.6. Using IE to prediction standard error of the mean

Having determined the linear relationships between standard error and our component errors, we proceeded to model the relationship between standard error of the mean and *IE*. Our model performed very well for both our training and testing (Figure 14) datasets. The parameters for our model were:

$$\text{std. error} = 0.6269 * IE - 0.06067 \tag{21}$$

6.8.7. Using IE to predict sample size

We learn the parameters for an MLR for predicting the sample size using *IE*. Modelling the relationship between *IE* and the sample size would allow us to accomplish our main goal of providing a model for the relationship between a sample size and a user provided insight measure. We modelled this relationship successfully using the log of *IE* and the log of sample size and had great results for both data that our model had seen in training and data that it had not been seen in training (Figure 15). Our R^2 for our testing dataset was 0.81. This means that our model can accurately determine the sample size required to meet user provided insight requirements. The parameters for our model were:

$$\log_2(\log_2(\text{size})) = -0.61 * \log_2(IE) + 0.99 \tag{22}$$

7. Applications

Our model allows us to predict the sample size that would give us an arbitrary insight error. For example, if we have scientists visualising exabytes of data, they can use our application to guide their work in terms of how much error they are willing to accept in order to speed up their workflow. They would simply use a slider to input an arbitrary value for *IE*, see a feedback visualisation that shows the associated visualisation speed for the given *IE* value, and our model would calculate the associated sample size (n) for the selected *IE*, providing the visualisation application with the corresponding n and rendering the visualisation of the sampled data.

We are also able to provide the sampling, visualisation, perception, and insight errors, given an arbitrary sample size. An example of an application for this would be one where a scientist knows the sample size that is required to run their visualisation within a given time but would like to know the impacts of using that sample size. He or she would provide our model with the sample size and our model would provide the corresponding sampling, visualisation, perception, and insight errors. This would give the scientist an objective measure of the uncertainty associated with the results he or she would get from the given sample.

Our approach can also be used to save time and money for scientists running ensemble simulations. Scientific simulation can take a long time to execute and ensemble simulation requires many simulation runs. Lowering the runtime of each simulation would save time and money due to power, cooling and per CPU licensing costs associated with scientific simulation (Adhinarayanan 2015; Borghesi et al. 2018). For

example, given a scenario (S2) where a scientist is using histograms to gain insight on the distribution of ensemble data objects produced by an ensemble simulation. Using a slider based on our model, one can reduce the runtime of each simulation. The scientist would input an expected insight error measure. This measure would determine the amount of data required by the visualisation application in order to provide the required insight level. The data amount provided by our model would in turn be used to guide the grid size for each simulation run in order to produce the required data. The result of this approach would be a faster runtime for each simulation that results in time and money savings as a result of lower power, cooling, and licensing costs.

Scenario S2 (above) could be extended to cases where our scientist is visualising the resultant ensemble data objects in order to determine if the complete simulation input parameter space is being evenly covered (Dahshan and Polys 2018). Using a spatial visualisation of the mapping of the resulting ensemble data objects to the input space, one could determine input parameter spaces that need more or less coverage. Given the cost and long runtime of each simulation, the scientist might determine that a lowering of insight accuracy is worth the time and cost savings. As a result, running the simulation faster, despite increased insight error, at an acceptable insight error rate would result in large runtime benefits.

Our model could also bring interactivity and its benefits to ensemble and other scientific simulation applications that produce big data and have long runtimes. An example of a simulation with long runtimes that could benefit from our approach is one where the results are analysed using a geographic visualisation. In cases where this simulation takes days or weeks to run, it would be wasteful to have to wait for days before determining that the incorrect set of parameters has been used. Using our model to decide on a quick runtime to zero in on the right input parameters to use before running the simulation at a high accuracy could save days or months in the simulation work flow.

8. Discussion and future work

This work implements the idea of providing a solution that is focused on the human. It presents a model that is driven by human insight requirements that are presented in terms of insight error. The model allows us to calculate the insight error as a function of sampling, perception, and visualisation errors, and visualisation, perception, and sampling errors as functions of sample size. This means that a visualisation end user who selects an arbitrary insight and perception level from a slider is in essence selecting the sample size. This approach can be used to sample big datasets using human insight and perception levels as the primary consideration.

Most big data analytic approaches focus on automating the analytics due to the well-known human limitations associated with the processing of time sensitive large amounts of data. The consensus it seems is to get the humans out of the way of the faster more accurate computing machines. The result is usually either one that works extremely well and is used widely, or one that does not seem to do the job well and is avoided by end users. In both these cases, the humans using the solution have very little understanding of the inner workings of the solution. The problem with these approaches is that when things go wrong, the results can be catastrophic because very few people are able to catch early telltale signs of something going wrong (O'Neil 2017). This work is a contribution to an alternate human-in-the-loop approach that focuses on the human and uses faster compute machines to aid the human.

The benefit of our approach is that it is based on the understanding that users have highly variable needs and as a result does not try to model these needs. An approach that tries to model user needs introduces a lot of bias. Our approach is centred on the user and allows for users to adjust the model's involvement intuitively each time the user's needs change. For example, in a situation where a visualisation user is exploring big data and is willing to accept a low fidelity visualisation, one can input a high *IE* value into the model and as a result visualise a small sample in a short amount of time with a clear understanding of how it will impact the insights drawn from the visualisation. If the user feels that he or she has a firm understanding of the big data and would like to generate a high fidelity visualisation even though it will take a long time, he or she could input a low *IE* value and generate the visualisation. Our approach provides this flexibility based on the user's needs at a given time.

One could argue against normalising our *IE* using the sample range. We chose this approach for two reasons: (i) all interactions with the data by study participants were done using the sample, so it makes sense to normalise using a sample statistic, and (ii) during the study design, we chose depth over breadth when investigating the relationship between IE and its component errors as a function of sample size. We chose to have one sample per sample size for each population, with multiple repetitions of the same samples as opposed to many different samples for each sample size with no repetitions. This constraint was as a result of having 400 treatments and a need to avoid learning. This design decision limited our knowledge of the behaviour of *IE* and its components across populations. As a result, our *IE*

measure can be thought of as a relative measure similar to the widely used Cp value (Gilmour 1996) that can be used to compare models using the same data.

Our IE measure allows one to compare the *IE* from different sample sizes and determine the difference of *IE*s, when *IE* is increasing, decreasing, or not changing as a function of sample size. Converting our *IE* to an absolute measure ***(IE')*** that can be used to compare different models using different data would entail running an additional study with more samples and less repetitions or more samples and participants. We plan to conduct such a study in the future. However, since we had learned the relationship between our component errors and sample size, we were able to generate a proof-of-concept absolute ***IE'*** measure by leveraging sampling and the data from our user study. Using a single population, we generated a hundred samples for each sample size (Figure 16). For each sample, we predicted the human estimates using the average perception error for that sample size learned from our crowd study. We then learned an absolute ***IE'*** normalised by the range of the population. Using the variance of this absolute ***IE'*** error with a log transformation gave us a fairly linear model that can be used to predict the ***IE'*** given a sample size and vice versa (Figure 17).

In order to produce solutions that humans understand, trust and can contribute to their accuracy, we need to understand human strengths and capabilities when it comes to data analytics and processing. Even though more work is needed to understand how people perceive visualisations, this work contributes to that caused by quantifying human understanding of visualised information and provides a model that leverages our understanding of human understanding to provide visualisations that efficiently present information at arbitrary workload levels.

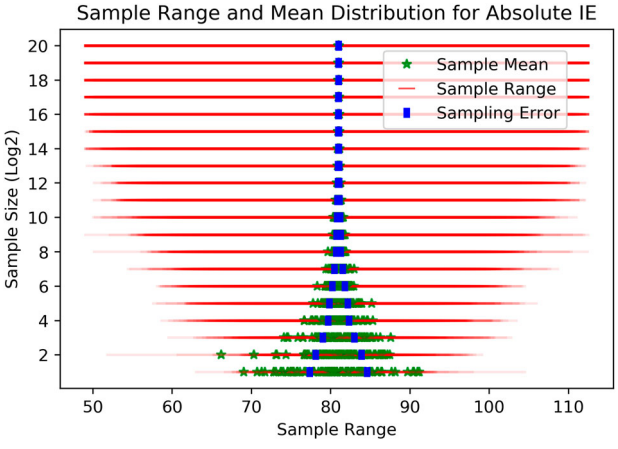

Figure 16. Samples generated to learn simulated absolute insight error (***IE'***). Sample ranges are visualised in red lines with low opacity. Dark red areas show repeated coverage. Sample means are visualised using green asterisks. The blue markers represent the expected sampling error (*SE*).

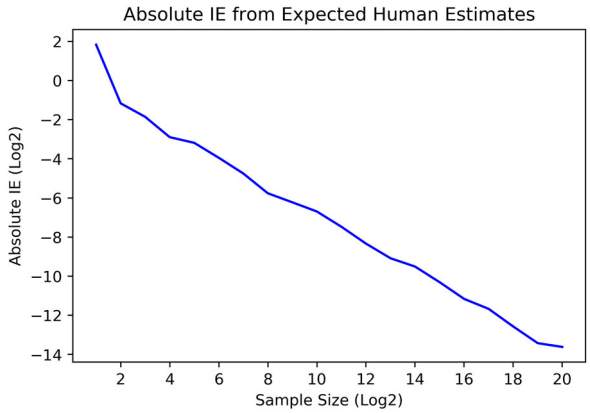

Figure 17. Absolute insight error (***IE'***) after log transformation as a function of sample. This measure is generated using expected human estimates from the crowd study as well as a hundred random samples for each sample size. This ***IE'*** measure is normalised by the range of the population and is robust to different populations.

8.1. Limitations

The major limitation of this work is that it is based on simple tasks and only on histogram visualisation that can be seen as providing little real world benefit. Even though a histogram is a simple visualisation, its strength lies in its simplicity. In a glance, one can understand the range, mean, distribution, mode, etc. of a big dataset with very little training. We plan to address these limitations in future work by focusing on more complex visualisations and user tasks. That being said, the use of simple tasks is beneficial because it allows us to focus on the main deliverable of this work, which is building a model that relates human insight to sample sizes without getting caught up in other details and confounding effects associated with complex tasks. Other limitations include the use of synthetic datasets as opposed to real world data. We plan to address these limitations in future work. Some might frown on the use of R^2 values to evaluate MLR models, but we feel that the use of these values in conjunction with other measures like Cp values and error visualisation provides adequate model evaluation.

Additional limitations include the need for an improvement in our modelling of human perception of visualisations, and the fact that our data consisted of only numerical data. Most big datasets are composed of high dimensional data of a variety of data types. Even though one can argue that a lot of data mining

and analytics approaches tend to transform this data into a numeric format, it is important that we study such data and see how our model perform on it. We plan to address these area of limitation in the future as we apply our approach to real world data. The use of a relative IE metric is another limitation that needs to be addressed in the future. In this work, we provided a proof of concept absolute IE metric (Section 8), but we plan to produce one that is extensively evaluated in the future.

8.2. Future work

This work provides a relative measure of IE. As noted earlier, this measure cannot be used to compare IE from different datasets. In future work, we plan to address this limitation by providing a global IE measure that can be used across datasets. We also plan to add more evaluation of the human component in perceiving visualised data and to evaluate our approach with case studies involving the application of this approach to real world challenges like those encountered in information security and geoscience domains. These domains typically rely on advanced hardware to process and visualise big data. We think our approach could help extend the reach of these approaches. Additionally, we intend to explore the perception effect on insight in greater detail to determine if there are other dependencies like stress, workload, and time of day for example. Questions like 'does the time of day impact how end users understand big data visualisation?'. Such questions have an impact in domains with a low error tolerance like information security where adversaries can leverage high stress periods to bypass security that relies on an analyst viewing a security visualisation application. A potential visualisation that leverages this information would adjust itself based on these external conditions that could impact insight. We also intend to investigate the effects of different datasets, visualisations, and user tasks on our model. Will different visualisations and user tasks also produce an exponential decay relationship between *IE* and sample size?

9. Conclusion

In this work, we used a simple task to run a crowd study that quantified human insight levels. Using the results of the study we generated a model that can be used to predict the sample size when given an arbitrary insight level. This allows for the efficient visualisation of big datasets by using just the right amount of data needed to meet a given insight requirement. We evaluated our model and reduced it to a form that only has one variable. Our model relies heavily on the use of standard error of the mean as a proxy for our features and this allows the generation of model parameters without needing to run a study for every new dataset. Our approach allows visualisation authors to create an interface with a slider, through which the end user selects an arbitrary insight level that in turn predicts the required sample size to generate a visualisation that would produce the required insight level. We also generated a model that can provide the impact of an arbitrary sample size on sampling, visualisation, perception, and insight errors. This model is valuable for use cases when one wants to put results from a visualisation of sampled data into perspective. We studied and defined the relationship between insight accuracy and sample size, defined and demonstrated the relationship between insight error, its component errors and a well-known statistical measure, displayed the behaviour of insight error and its component errors as a function of sample size, and provided a model that allows for the speedup of big data visualisation based on user provided insight levels.

Disclosure statement

No potential conflict of interest was reported by the authors.

Funding

This work is supported in part by the National Science Foundation via grant #DGE-1545362, UrbComp (Urban Computing): Data Science for Modeling, Understanding, and Advancing Urban Populations.

References

Adhinarayanan, V. 2015. "On the Greenness of In-Situ and Post-Processing Visualization Pipelines." *2015 IEEE International Parallel and Distributed Processing Symposium Workshop*, May, 880–887.

Berres, Anne Sabine, Vignesh Adhinarayanan, Terece Turton, Wu Feng, and David Honegger. Rogers. 2017. *A Pipeline for Large Data Processing Using Regular Sampling for Unstructured Grids*. Technical Report. Los Alamos National Lab.(LANL), Los Alamos, NM.

Borghesi, Andrea, Andrea Bartolini, Michela Milano, and Luca Benini. 2018. "Pricing schemes for energy-efficient HPC systems: Design and exploration." *The International Journal of High Performance Computing Applications* 109434201881459. http://dx.doi.org/10.1177/1094342018814593.

Card, Stuart K., Jock D. Mackinlay, and Ben Shneiderman. 1999. "Using vision to think." *Readings in information visualization*, 579–581. Morgan Kaufmann Publishers Inc.

Chang, Remco, Caroline Ziemkiewicz, Tera Marie Green, and William. Ribarsky. 2009. "Defining Insight for Visual Analytics." *IEEE Computer Graphics and Applications* 29 (2): 14–17.

Chen, H., S. Zhang, W. Chen, H. Mei, J. Zhang, A. Mercer, R. Liang, and H. Qu. 2015. "Uncertainty-Aware Multidimensional Ensemble Data Visualization and Exploration." *IEEE Transactions on Visualization and Computer Graphics* 21 (9): 1072–1086.

Choe, Eun Kyoung, Bongshin Lee, and M. C. Schraefel. 2015. "Characterizing Visualization Insights from Quantified Selfers' Personal Data Presentations." *IEEE Computer Graphics and Applications* 35 (4): 28–37.

Dahshan, Mai, and Nicholas Polys. 2018. "Making Sense of Scientific Simulation Ensembles." *Poster presented at SC 2018*, Dallas, Texas, Nov. https://sc18.supercomputing.org/proceedings/tech_poster/poster_files/post165s2-file3.pdf.

Fekete, Jean-Daniel. 2015. "ProgressiVis: A Toolkit for Steerable Progressive Analytics and Visualization." *1st Workshop on Data Systems for Interactive Analysis*, 5.

Fisher, Danyel, Igor Popov, Steven Drucker, and M. C. Schraefel. 2012. "Trust Me, I'm Partially Right: Incremental Visualization Lets Analysts Explore Large Datasets Faster." *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '12, New York, NY, 1673–1682. ACM. http://doi.acm.org/10.1145/2207676.2208294.

Galakatos, Alex, Andrew Crotty, Emanuel Zgraggen, Carsten Binnig, and Tim Kraska. 2017. "Revisiting Reuse for Approximate Query Processing." *Proc. VLDB Endow.* 10 (10): 1142–1153. https://doi.org/10.14778/3115404.3115418

Gilmour, Steven G. 1996. "The Interpretation of Mallows's C_p-statistic." *The Statistician* 45: 49–56.

Gori, A., G. Craparo, M. Giannini, Y. Loscalzo, V. Caretti, D. La Barbera, G. M. Manzoni, 2015. "Development of a New Measure for Assessing Insight: Psychometric Properties of the Insight Orientation Scale (IOS)." *Schizophrenia Research* 169 (1–3): 298–302.

Grtler, J., C. Schulz, D. Weiskopf, and O. Deussen. 2018. "Bubble Treemaps for Uncertainty Visualization." *IEEE Transactions on Visualization and Computer Graphics* 24 (1): 719–728.

Holzinger, Andreas. 2013. "Human–Computer Interaction and Knowledge Discovery (HCI-KDD): What is the benefit of bringing those two fields to work together?" *International Conference on Availability, Reliability, and Security*, 319–328. Springer.

Hong, Seong E., J. Kim Hwa, and J. Cha Kyung. 2018. "Big Data Preliminary Analysis: A Framework for Easier Data Sharing and Discovery." *International Information Institute (Tokyo). Information* 21 (2): 755–763. http://login.ezproxy.lib.vt.edu/login?url=https://search.proquest.com/docview/2038677630?accountid=14826

Kaisler, Stephen, Frank Armour, J. Alberto Espinosa, and William Money. 2013. "Big Data: Issues and Challenges Moving Forward." *2013 46th Hawaii International Conference on System Sciences (HICSS)*, 995–1004. IEEE.

Kulessa, Moritz, Alejandro Molina, Carsten Binnig, Benjamin Hilprecht, and Kristian. Kersting. 2018. "Model-based Approximate Query Processing." *arXiv preprint arXiv:1811.06224*.

Leetaru, Kalev. 2019. "The Big Data Revolution will be Sampled: How 'Big Data' Has Come To Mean 'Small Sampled Data'." *Forbes*.

Lin, Qingwei, Weichen Ke, Jian-Guang Lou, Hongyu Zhang, Kaixin Sui, Yong Xu, Ziyi Zhou, Bo Qiao, and Dongmei Zhang. 2018. "BigIN4: Instant, Interactive Insight Identification for Multi-Dimensional Big Data." *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, 547–555. ACM.

Liu, L., A. P. Boone, I. T. Ruginski, L. Padilla, M. Hegarty, S. H. Creem-Regehr, W. B. Thompson, C. Yuksel, and D. H. House. 2017. "Uncertainty Visualization by Representative Sampling from Prediction Ensembles." *IEEE Transactions on Visualization and Computer Graphics* 23 (9): 2165–2178.

Liu, Zhicheng, Biye Jiang, and Jeffrey. Heer. 2013. "imMens: Real-time Visual Querying of Big Data." *Computer Graphics Forum* 32 (3pt4): 421–430. https://onlinelibrary.wiley.com/doi/abs/10.1111/cgf.12129

Macke, Stephen, Yiming Zhang, Silu Huang, and Aditya Parameswaran. 2018. "Adaptive Sampling for Rapidly Matching Histograms." *Proc. VLDB Endow.* 11 (10): 1262–1275https://doi.org/10.14778/3231751.3231753

Melnik, Sergey, Andrey Gubarev, JingJing Long, Geoffrey Romer, Shiva Shivakumar, Matt Tolton, and Theo. Vassilakis. 2010. "Dremel: Interactive Analysis of Web-scale Datasets." *Proceedings of the VLDB Endowment* 3 (1-2): 330–339.

Moritz, Dominik, Danyel Fisher, Bolin Ding, and Chi. Wang. 2017. "Trust, but Verify: Optimistic Visualizations of Approximate Queries for Exploring Big Data." *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems*, CHI '17, New York, NY, 2904–2915. ACM. http://doi.acm.org/10.1145/3025453.3025456.

Nguyen, T. T., and I. Song. 2016. "Centrality Clustering-based Sampling for Big Data Visualization." *2016 International Joint Conference on Neural Networks (IJCNN)*, July, 1911–1917.

North, Chris. 2006. "Toward Measuring Visualization Insight." *IEEE Computer Graphics and Applications* 26 (3): 6–9.

O'Neil, Cathy. 2017. *Weapons of Math Destruction: How Big Data Increases Inequality and Threatens Democracy*. New York: Broadway Books. https://www.worldcat.org/title/weapons-of-math-destruction-how-big-data-increases-inequality-and-threatens-democracy/oclc/1005213790.

Pang, Alex T, Craig M. Wittenbrink, and Suresh K. Lodha. 1997. "Approaches to Uncertainty Visualization." *The Visual Computer* 13 (8): 370–390.

Park, Yongjoo, Michael Cafarella, and Barzan Mozafari. 2016. "Visualization-Aware Sampling for Very Large Databases." *2016 IEEE 32nd International Conference on Data Engineering (ICDE)* http://dx.doi.org/10.1109/ICDE.2016.7498287.

Rojas, Julian A Ramos, Mary Beth Kery, Stephanie Rosenthal, and Anind. Dey. 2017. "Sampling Techniques to Improve Big Data Exploration." *2017 IEEE 7th Symposium on Large Data Analysis and Visualization (LDAV)*, 26–35. IEEE.

Ruan, Zichan, Yuantian Miao, Lei Pan, Nicholas Patterson, and Jun. Zhang. 2017. "Visualization of Big Data Security: a Case Study on the KDD99 Cup Data Set." *Digital Communications and Networks* 3 (4): 250–259.

Sacha, D., H. Senaratne, B. C. Kwon, G. Ellis, and D. A. Keim.. 2016. "The Role of Uncertainty, Awareness, and Trust in Visual Analytics." *IEEE Transactions on Visualization and Computer Graphics* 22 (1): 240–249.

Saraiya, Purvi, Chris North, and Karen. Duca. 2004. "An Evaluation of Microarray Visualization Tools for Biological Insight." *IEEE Symposium on Information Visualization, 2004 (INFOVIS'04)*, 1–8. IEEE.

Turkay, C., E. Kaya, S. Balcisoy, and H. Hauser. 2017. "Designing Progressive and Interactive Analytics Processes for High-Dimensional Data Analysis." *IEEE Transactions on Visualization and Computer Graphics* 23 (1): 131–140.

Wang, Lidong, Guanghui Wang, and Cheryl Ann. Alexander. 2015. "Big Data and Visualization: Methods, Challenges and Technology Progress." *Digital Technologies* 1 (1): 33–38.

Xiao, Fengjun, Mingming Lu, Ying Zhao, Soumia Menasria, Dan Meng, Shangsheng Xie, Juncai Li, and Chengzhi Li. 2018. "An Information-aware Visualization for Privacy-Preserving Accelerometer Data Sharing." *Human-Centric Computing and Information Sciences* 8 (1): 13. https://doi.org/10.1186/s13673-018-0137-6

Yi, Ji Soo, Youn-ah Kang, John T. Stasko, and Julie A. Jacko. 2008. "Understanding and Characterizing Insights: How do People Gain Insights Using Information Visualization?" *Proceedings of the 2008 Workshop on BEyond time and errors: novel evaLuation methods for Information Visualization*, 4. ACM.

Zgraggen, Emanuel, Alex Galakatos, Andrew Crotty, Jean-Daniel Fekete, and Tim. Kraska. 2017. "How Progressive Visualizations Affect Exploratory Analysis." *IEEE Transactions on Visualization & Computer Graphics* 23 (8): 1977–1987.