Sampling Techniques to Improve Big Data Exploration

Julian A. Ramos Rojas Mary Beth Kery Stephanie Rosenthal Anind Dey

Carnegie Mellon University, School of Computer Science, Pittsburgh PA, US

Abstract

The success of Big Data relies fundamentally on the ability of a person (the data scientist) to make sense and generate insights from this wealth of data. The process of generating actionable insights, called data exploration, is a difficult and time-consuming task. Data exploration of a big dataset usually requires first generating a small and representative data sample that can be easily plotted and viewed, managed and interpreted to generate insights. However, the literature on the topic hints at data scientists only using random sampling with regular sized datasets and it is unclear what they do with Big Data. In this work, we first show evidence from a survey that random sampling is the only technique commonly used by data scientists to quickly gain insights from a big dataset despite theoretical and empirical evidence from the active learning community that suggests benefits of using other sampling techniques. Second, to evaluate and demonstrate the benefits of other sampling techniques, we conducted an online study with 34 data scientists. These scientists performed a data exploration task to support a classification goal using data samples from more than 2 million records of editing data from Wikipedia articles, generated using different sampling techniques. The study results demonstrate that sampling techniques other than random sampling can generate insights that help to focus on different characteristics of the data, without compromising quality in a data exploration.

Keywords: Visual Knowledge Discovery, Data Filtering, Human-Computer Interaction

Index Terms: CCS→ Human-centered computing→ Human computer interaction (HCI) → Empirical studies in HCI.

1. Introduction

The expectations for Big Data are high and varied; they range from improved and more efficient healthcare [14] to crime-spot emergence prediction [26]. While Big Data algorithms continue to improve, its success relies upon the ability of a data scientist to detect patterns, determine useful features and visualizations and generate actionable insights from the data. In this article, we refer to data scientists as individuals whose job consists on extracting knowledge from large volumes of data in various forms using statistics, data mining or machine learning. While recent studies of data scientists examine their entire process from attaining data to reporting results [8,9], in this work we focus only on *data exploration* tasks that often occur throughout the process. Data exploration consists mostly of tasks that help data scientists understand their data by gaining insights into the phenomena being modeled, assess the quality of the data, and find or create features that improve model accuracy. Data exploration includes the acts of creating graphs and plots, estimating statistics, transforming the data, finding anomalies, *etc*. These steps however are not always feasible for data scientists to perform on entire Big Data datasets. A single analysis can take hours to compute on a Big Data dataset [6], and in some cases it is beyond the capabilities of the available computing infrastructure [9,14]. With a high cost in time and resources to analyze the dataset, this restricts a data scientist's ability to perform data exploration. In order to overcome these challenges, a data scientist can create smaller data samples to be able to perform data exploration in a timely manner. These small data samples must be statistically sound in order to offer an unbiased and correct representation of the full dataset and, as a result, some data scientists have reported avoiding or limiting their use to avoid introducing bias [8,11].

Nonetheless, sampling seems to be an important approach for exploring large datasets. However, to our knowledge there is no evidence in the literature about how data scientists use sampling techniques with Big Data, nor how those sampling techniques affect the quality or focus of their insights. Based on the lack of tools for sampling Big Data available today, we hypothesize that data scientists are using random sampling, if they are sampling their data at all despite evidence from the active learning community on the advantages of using a variety of sophisticated sampling methods [5,24]. A study of data scientists also suggests that using multiple sampling strategies on the same dataset would enable more effective evaluation of datasets [9].

We first surveyed 22 data scientists who work on large datasets and confirmed that data scientists are using only random sampling or pseudo-random sampling. However we also found that if multiple sampling techniques were available to them, a majority of participants believed that this would decrease the time it takes to explore their data and make insights, and also improve the quality of the insights. Based on this finding and the theoretical and empirical evidence in favor of using multiple different active learning sampling methods to improve classification performance [2,4,7,12], we hypothesized that *data scientists would also benefit from data exploration on smaller but better selected data samples generated from sampling techniques other than random sampling*.

To test our hypothesis we developed large-scale versions of popular sampling algorithms used in active learning [24] – density sampling, uncertainty sampling, and query by committee - and setup a study to investigate the quality of insights generated by data scientists as they explored data sampled using these algorithms. The results of our online study with 34 data scientists indicate that data scientists can produce insights of comparable quality with any of our sampling techniques, but that the *content* of the insights varies across the different techniques. This indicates that a data scientist may be able to make a broader range of insights about their data by using a range of sampling tools rather than their current practice of only relying on random sampling. We conclude that data scientists should use multiple sampling techniques separately or even in conjunction with each other to generate a broad range of insights about their datasets.

* ingenia@cs.cmu.edu, † mkery@cs.cmu.edu

‡ srosenthal@sei.cmu.edu, § anind@cs.cmu.edu

IEEE Symposium on Large Data Analysis and Visualization 2017
October 1–6, Phoenix, Arizona, USA
978-1-5386-0617-9/17/$31.00 ©2017 IEEE

Copyright 2017 IEEE. All Rights Reserved. This material is based upon work funded and supported by the Department of Defense under Contract No. FA8702-15-D-0002 with Carnegie Mellon University for the operation of the Software Engineering Institute, a federally funded research and development center. DM17-0520

Authorized licensed use limited to: b-on: UNIVERSIDADE DE AVEIRO. Downloaded on March 14,2023 at 16:02:51 UTC from IEEE Xplore. Restrictions apply.

2. Related Work

Given the reliance on data scientists to build Big Data models, understanding their processes for exploring and analyzing data and the ways in which we can improve those processes is paramount to the success of Big Data. In the next section we review the related work on Big Data related studies and sampling techniques from the Active Learning literature.

2.1 Data Exploration Process

Many studies have focused on the processes of data scientists as they attempt to attain, clean, understand, model, and visualize their data [8,9]. For example, Kandel *et al*. [9] define the general process followed by data scientists performing data analysis as follows: *Discover*: Tasks necessary to acquire a dataset; *Wrangle*: Preprocessing tasks executed to get the data into a desired format; *Profile*: Set of tasks that guarantee the quality of the data; *Model*: Tasks accomplished to obtain information from the dataset; and *Report*: Final set of results. Underlying many of these processes is the task of data exploration, often called exploratory data analysis. Performing data exploration refers to gaining an understanding of the dataset with the objective of generating hypotheses, testing assumptions, supporting the selection of statistical methods and providing a basis for further data collection [25]. Tasks commonly executed during data exploration include but are not limited to creating summary statistics, histograms or other visualizations, and determining the data distribution. Data exploration is an iterative and ongoing process [8], as new analyses lead to new data cleaning procedures which in turn lead to new assumptions to be checked and analyzed.

However, data exploration and visualization of Big Data datasets is computationally challenging. A common way to deal with Big Data datasets for data exploration is to use feature selection to create two-dimensional plots. However, rendering a plot for millions of data points is still time consuming and sometimes impossible [10]. Similarly, any data transformation or computation of any kind of statistic becomes time consuming when the data points number in the millions [6,21]. To overcome this challenge, it is necessary to *sample* Big Data datasets and the use of multiple sampling techniques may be necessary to avoid bias in the data [9].

2.2 Data Sampling

An evaluation of common data science software and packages shows that random sampling is frequently the only supported sampling technique to use for the large-scale datasets ([15,19,20], [3] implements other techniques but not for large scale data). Drawing from active learning (*e.g.,* [5,7,17,24]), which relies heavily on its ability to select the most informative data points, it is apparent that there are many different sampling algorithms that select data points with specific properties that can be used in classification related problems. Empirical and theoretical [2] results from the active learning community show that these basic querying strategies outperform passive learning using random sampling strategies [2,4,7,12] and are specially good at dealing with imbalanced datasets [7] and maintaining fast rates of error decay [2,4,7,12].

Although there are many different variants of active learning sampling techniques, we are focused mainly on common techniques [24]: Query By Committee, Density and Uncertainty Sampling. Other sampling methods like stratified random sampling were discarded because they require prior knowledge of the strata of the dataset or the population and this may not be an option for data exploration when the dataset or the population are not well known.

Query By Committee (QBC): The QBC approach involves maintaining a committee or set of classifiers that are all trained on a labeled dataset. Each classifier (or committee member) is then allowed to vote (predict the label) on the labels of input data points. The most informative data points are considered to be those about which the classifiers most disagree [24].

Uncertainty Sampling: This sampling technique [12] selects data points for which there is a low posterior probability $P(\hat{y}|x)$ (high uncertainty), where $\hat{y}$ is the label assigned and x is the data point.

Density Sampling: In this method, the feature space is divided into a grid and points are picked probabilistically, relative to the number of points in the grid cell. This technique usually does not select data points in low probability regions, resulting in a low number of selected outliers.

We hypothesize that *data scientists can create better insights from data sampled using different techniques (just as active learning algorithms can)*.

2.3 Insights

Data exploration, although an important process, is meaningless if it does not produce a tangible result that can be used for data modeling, or any other high level purpose. We refer to this main output from data exploration as an insight. In this article, we use the canonical definition of insight as found in the Merriam-Webster dictionary: *The act or outcome of grasping the inward or hidden nature of things or of perceiving in an intuitive manner.* Insights have mainly been used in the visualization community as a way to evaluate visualizations [18]. These evaluations are different from typical visualization benchmarks which are closed and specific to finding specific properties of a dataset and are usually biased towards a specific line of thought [18]; instead insight-based evaluations are open ended as they are designed to let the data scientist more organically discover whatever she can on the data without being biased by specific questions [18]. In this work we use insight-based evaluation as a way to evaluate the effectiveness of sampling techniques for data scientists to learn about a dataset. Other evaluations were considered like measuring the discovery rate of outliers or other types of observations. However, this kind of evaluation is not a good fit for data exploration because it is very constrained and specific to the task in hand, while data exploration is an open ended process.

Given the lack of work examining the sampling techniques that Big Data data scientists use (if any), as well as their potential to affect data scientists' insights, we first deployed a survey to understand current practice.

3. Survey Of Big Data Data Scientists

Kandel *et al.* note that the data scientists they interviewed were wary of using data sampling because of the bias it could introduce into their analysis [9]. Similarly, Lin *et al.* [13] state that sampling for Big Data is easy to get wrong, is contrary to the goal of Big Data, and is inaccurate to the point where they suggest to simply use as much data as possible and run experiments at scale.

While we agree that using all available data is important when computing models, it is challenging for a data scientist to explore and understand very large sets of features and observations at once. With the goal of understanding the state of the practice used to perform data exploration on Big Data, we created an online survey illustrated in Figure 1. We announced our survey in multiple online locations that would attract individuals with some Big Data experience. The locations included Kaggle forums (https://www.kaggle.com/general/24584), Big Data and Data

Authorized licensed use limited to: b-on: UNIVERSIDADE DE AVEIRO. Downloaded on March 14,2023 at 16:02:51 UTC from IEEE Xplore. Restrictions apply.

science special interest groups on LinkedIn in topics like: Robotics, Computer vision, data science, Big Data, Machine Learning and email lists of students who had successfully taken a doctoral-level machine learning course at our institution. Study participants were given a $5 gift card to compensate for their time. To filter out participants who were not data scientists, we asked them to check whether their job could be described as the extraction of knowledge from large volumes of data in various forms using statistics, data mining or machine learning.

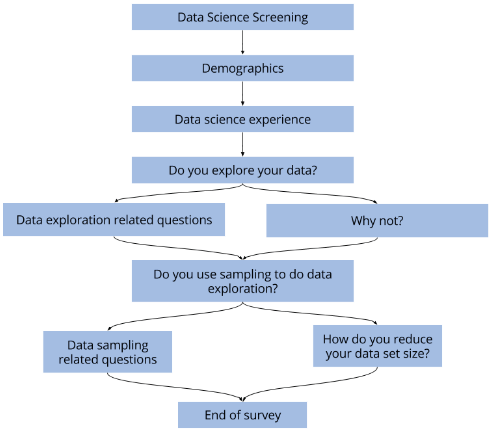

Data Science Screening
Demographics
Data science experience
Do you explore your data?
Data exploration related questions
Why not?
Do you use sampling to do data exploration?
Data sampling related questions
How do you reduce your data set size?
End of survey

Figure 1. Survey structure

3.1 Demographics

35 people completed our survey, of which 31 passed our screening question about their experience with Big Data. From these 31 survey responses, 22 were valid while the rest were discarded due to bogus responses (participants that tried to get paid without actually properly answering the survey). We report the findings from these 22 valid surveys. More than half (55%) of our participants had an age of 30-39, 30% were 20 – 29 years old and 15% were 40 years or older. 75% of our participants were male. 70% of our participants had an education level of Master's degree or higher which is not surprising given our recruitment method. The remaining 30% was uniformly distributed between bachelors' degrees, current master's students and other. They worked at companies like Microsoft, Google, and IBM, and their data science experience was very diverse as shown in Figure 2.

3.2 Responses

We analyzed our survey responses focusing on understanding our respondents' typical dataset and their process for exploring that data. 21 of the 22 respondents reported to have analyzed datasets with a million data points or bigger and 7 had analyzed 10 million data points or larger. More than half of our participants had worked with datasets with 1000 to 1 billion features. These responses show that the majority of our respondents have had some exposure to a large dataset. To understand their data exploration process, we asked two kinds of questions: multiple choice and open ended. We manually coded the open-ended text responses and identified the most prevalent themes. First, we note that all of our participants reported exploring their dataset before doing anything else.

During their data exploration process, 63% of the participants answered that they use data sampling. Of these 63%, 50% use random sampling, 33% stratified sampling and 16% perform sampling by hand (*i.e.*, manually selecting data points by looking through the data). Of the participants that used sampling, we then assessed their responses about the quality and bias that their sampling might introduce. Most respondents decided what the size of the sample should be without the guidance of measures like variance of the dataset. Participants did not evaluate their samples for quality. We note that the quality of a sample *can* be shown by generating many samples of the same size, and then plotting all of them to check if they give a consistent distribution.

Finally, the survey asked participants about their perceived utility of other sampling techniques. 65% of our participants believed that using multiple sampling techniques could *improve the quality* of their insights and 71% thought it would *decrease the time* spent on data exploration.

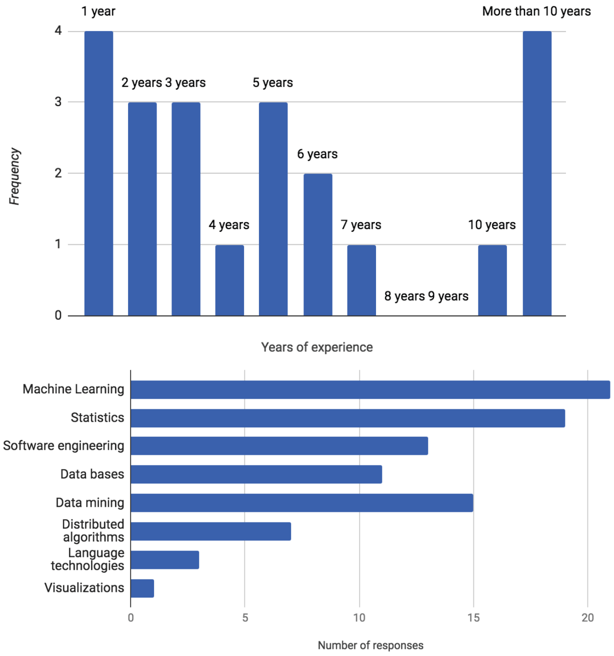


Figure 2. Data science demographics of survey respondents. Top: Years of experience, Bottom: Data science relevant skills

3.3 Conclusions

We found that a majority of our surveyed data scientists use data sampling, though mostly random sampling, stratified sampling and sampling by hand. There are a variety of reasons that could explain these results including lack of tools available in common data science toolkits or knowledge about the existence of different sampling techniques. Additionally, statistics classes typically only cover random sampling for data exploration. Other sampling techniques are introduced for active learning, but it may not be apparent that those techniques could be used within data exploration. However, our data scientists did agree that other sampling algorithms may help them more efficiently generate insights as well as improve the quality of their insights. In order to evaluate whether their (and our) hypothesis is true, we now present our study to compare the insights generated using different data sampling algorithms on the same large dataset.

4. DATA SAMPLING INSIGHTS STUDY

While random sampling selects individual data points through an unbiased selection, the final sample may not be a representative subset of the data. For example, from a random sample it is hard for a data scientist to determine which data points are common and which are anomalous. Random sampling may fail to select anomalous points that the data scientist would prefer to evaluate. Given that common tasks for data exploration include finding natural clusters of data as well as anomalies, it is possible that using

Authorized licensed use limited to: b-on: UNIVERSIDADE DE AVEIRO. Downloaded on March 14,2023 at 16:02:51 UTC from IEEE Xplore. Restrictions apply.

alternative sampling techniques that bias their selection in predefined ways would help practitioners focus their efforts on the types of points they are interested in. In order to understand this tradeoff in practice, we setup an online study in which data scientists (our participants) were assigned to a data sample generated by one of four different sampling techniques. Then we asked our participants to understand the data sample and write insights that could help them build a classifier. Next we describe the details of the study.

4.1 Study Design

Our main hypothesis was that sampling techniques with different selection biases would generate insights with a quality on par with random sampling but focused on different aspects of the data. We chose a between-subjects study design, with one of four sampling techniques in each condition: uncertainty, density, query by committee, and random. Random sampling represents our control condition based on our finding that this is what data scientists use under normal circumstances. In order to ensure that participants generate high-quality insights, as defined by Chis North [18], we used an open-ended study protocol that allows the participants of a study to explore the dataset, as opposed to a study or benchmark in which tasks and steps are predefined and focus only on specific aspects of a data analysis process like finding outliers. Our main analysis measures insight content and quality of the insights produced in each condition.

The participants' task was to explore the features of a provided dataset containing the editing habits of Humans and AI Bots on Wikipedia, using basic interactive visualizations, such as scatterplots and histograms that we provided. As they explored the data, they were asked to create insights that they might use to build a Bot detector for finding unregistered Bots. Specifically, we told them: "try to understand the dataset to a level where you could build a bot detector. We want you to formalize this understanding in the form of notes that we will call insights". We collected all of their insights as well as all of their interactions with our visualizations.

4.2 Participants

We recruited participants using the same online forums as before: Kaggle, Big Data and Data science special interest groups on LinkedIn in topics like: Robotics, Computer vision, data science, Big Data, Machine Learning, and institutional email lists. To participate in our study, participants had to be 18-80 years of age, must have studied machine learning or data science for 1 or more semesters at the University level and/or worked in Data Science or Machine Learning for the last 6 months. We compensated participants with a $20 gift card for their time.

4.3 Materials

The data samples used in the study were generated using four different sampling techniques as described in the related work: random, density, uncertainty and query by committee. This entire dataset, samples and code to generate the samples are available online at https://github.com/ubicomp-lab/big-data-sampling.

Table 1. Features computed from the Wikipedia editing log data set

Features	Most important to detect	Relevancy	Number
Per-user mean of inserts	Humans	High	0
Per-user mean of deletes	Neutral	High	1
Per-user mean of changes	Bot	High	2
Per-user std dev of inserts	Not sure	High	3
Per-user std dev of deletes	Not sure	High	4
Per-user std dev of changes	Not sure	High	5
Per-user-per-page mean of inserts	Humans	Medium	6
Per-user-per-page mean of deletes	Neutral	High	7
Per-user-per-page mean of changes	Bots	Medium	8
Per-user-per-page std dev of inserts	Not sure	Low	9
Per-user-per-page std dev of deletes	Not sure	Low	10
Per-user-per-page std dev of changes	Not sure	Low	11
Per-user largest single add	Humans/Bots	Medium	12
Per-user largest single delete	Humans/Bots	Medium	13
Per-user largest single change	Humans/Bots	Medium	14
Per-user most frequent hour edited	Humans/Bots	Medium	15
Per-user std dev of most frequent hour edited	Bots	Medium	16
Per-user total edits	Bots	Medium	17
Per-user total unique pages edited	Bots	Medium	18
Per-user avg total edits per page	Bots	Medium	19
Per-user std dev total edits per page	Not sure	Low	20
Per-user avg minor revisions	Bots	Medium	21
Per-user std dev minor revisions	Neutral	High	22
Per-user avg time between edits	Humans/Bots	Medium	23
Per-user std dev time between edits	Humans/Bots	High	24
Per-user mean time edit from last rev	Not sure	High	25
Per-user std time edit from last rev	Not sure	High	26

The data that we sampled from is a Wikipedia editing-log dataset for 2.22 million different users. For each user there are 27 different feature values that summarize editing behaviors during the user's total time editing on Wikipedia (as of May 2014). The features were chosen based on prior work [1] and are listed in Table 1. The users in the dataset are either Bots or Humans. Bots are meant to fix grammar or punctuation errors, find and revert malicious changes, and other maintenance tasks, while Human users contribute the bulk of the content of a Wikipedia article. Although Bots are required to register with Wikipedia (at the time of the study and in our dataset there were 500 Bots registered), there may be others that are editing the website without registering. This means that users that we would think are Human users may actually be Bots. The different samples generated contain all of the 500 registered Bots plus a sample generated from the remaining dataset. This was done to guarantee that our participants had the highest possible number of examples of Bots and to follow closely a real-world task of creating a bots classifier with a small labeled dataset. In total the 4 different samples (including the 500 bots) had data sizes of: 6015 users for random, 5948 for density, 6099 for uncertainty and 5542 for query by committee. The different sizes are due to the randomness inherent to each of the sampling techniques. Each sample is about 0.26% of the original dataset. All participants in the same study condition received the same sample of the data.

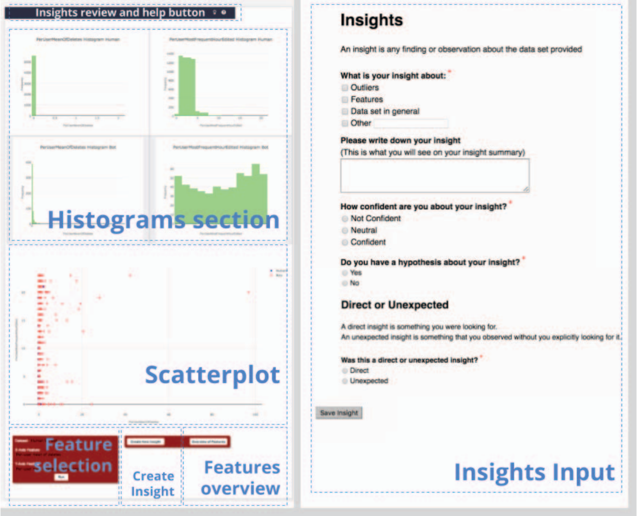
4.4 Online Data Exploration User Interface

To run our online study, we developed and hosted a study website on our own server. Our website did not record any personal information that could identify any of our participants like name or IP address with the exception of email address, which we used to send the gift card for participating in the study. However, we recorded the location of the mouse pointer every minute during the study to help identify participants that did not complete the study appropriately (*e.g.*, did not use the visualization support to generate insights). Our website was composed of the following webpages: Consent, demographics survey, education, quiz (Figure 4), task explanation, navigation help (Figure 5 left), visualizations (Figure 3. Left: Visualizations webpage screenshot, Right: Insights input screen left) and insight writing (Figure 5 right).

The webpage containing the visualizations of the sampled data is shown in Figure 3. In this webpage, the participant selects two features (from the 27 available) to display. The feature value distributions for Bots and Humans is shown in separate histograms

Authorized licensed use limited to: b-on: UNIVERSIDADE DE AVEIRO. Downloaded on March 14,2023 at 16:02:51 UTC from IEEE Xplore. Restrictions apply.

as shown in Figure 3. Under these histograms, it was displayed a scatter plot of the two features selected. There are a total of 351 possible combinations of pairs of features that our participants could look at. To help them decide which pairs of features to focus on, we provided the top 10 scatter correlation plots for different pairs of features that had the closest value to 0, -1 and +1. All of the visualizations were created using plotly.js [22].



Insights

An insight is any finding or observation about the data set provided

What is your insight about: *
□ Outliers
□ Features
□ Data set in general
□ Other

Please write down your insight
(This is what you will see on your insight summary)

How confident are you about your insight? *
○ Not Confident
○ Neutral
○ Confident

Do you have a hypothesis about your insight? *
○ Yes
○ No

Direct or Unexpected

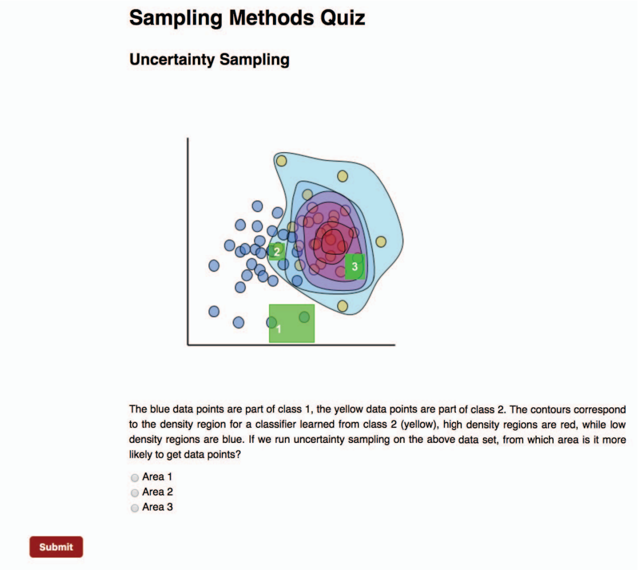
A direct insight is something you were looking for.
An unexpected insight is something that you observed without you explicitly looking for it.

Was this a direct or unexpected insight? *
○ Direct
○ Unexpected

Save Insight

Figure 3. Left: Visualizations webpage screenshot, Right: Insights input screen

On the insight submission page shown in Figure 3 (right), in addition to writing and submitting insights, the participants were required to select what the insight was about (outliers, features, general, other), indicate her confidence in the insight, specify the hypothesis associated with the insight (if applicable) and specify whether the insight was expected or unexpected (participants were told that an expected insight is one the participant was looking for, and an unexpected insight is one the participant observed without explicitly looking for it). We collected this information to better characterize the generated insights, as described in [23].



Sampling Methods Quiz

Uncertainty Sampling

The blue data points are part of class 1, the yellow data points are part of class 2. The contours correspond to the density region for a classifier learned from class 2 (yellow), high density regions are red, while low density regions are blue. If we run uncertainty sampling on the above data set, from which area is it more likely to get data points?

○ Area 1
○ Area 2
○ Area 3

Submit

Figure 4. Uncertainty sampling quiz, participants were shown a visualization of a dataset and then asked which of the data points will be more likely to be selected by the sampling method.

Participants could also click on a help button on the top right corner of the website as shown in Figure 5 (left), which displays the description of the task and features, a screenshot of the visualization webpage with usage hints (Figure 5) and a depiction of the sampling technique used to generate the sample

4.5 Study Protocol

Figure 5 (right) outlines the study structure. After consenting to participate in the study, participants were assigned at random to one of the 4 conditions. Next they responded to a demographics survey, followed by an explanation of the sampling condition they were assigned to. To check the participant's understanding of their assigned sampling technique, the following screen contained a one-question quiz. Participants had two attempts to answer correctly, and were removed from the study if they failed both. Participants who passed were briefed on the exploratory data analysis task they would be working on, as well as a description of the Wikipedia dataset. Participants were then given the task instructions to explore the data and make insights that would be useful for making a Bot detector. Finally, participants were shown how to use the different visualizations available and how to record insights in the platform as shown in Figure 5 (left). Participants had up to 1.5 hours to complete the study, after which the website closed automatically. To be eligible for payment, participants were required to enter at least 4 insights. After 30 minutes, a button appeared on the screen allowing participants to submit their insights and finish the study.

5. Data Analysis and Results

To understand the differences in the insights produced with each sampling technique, we measured insight quality and content. We describe our method for preprocessing our insights, assessing quality and content, and our analysis results.

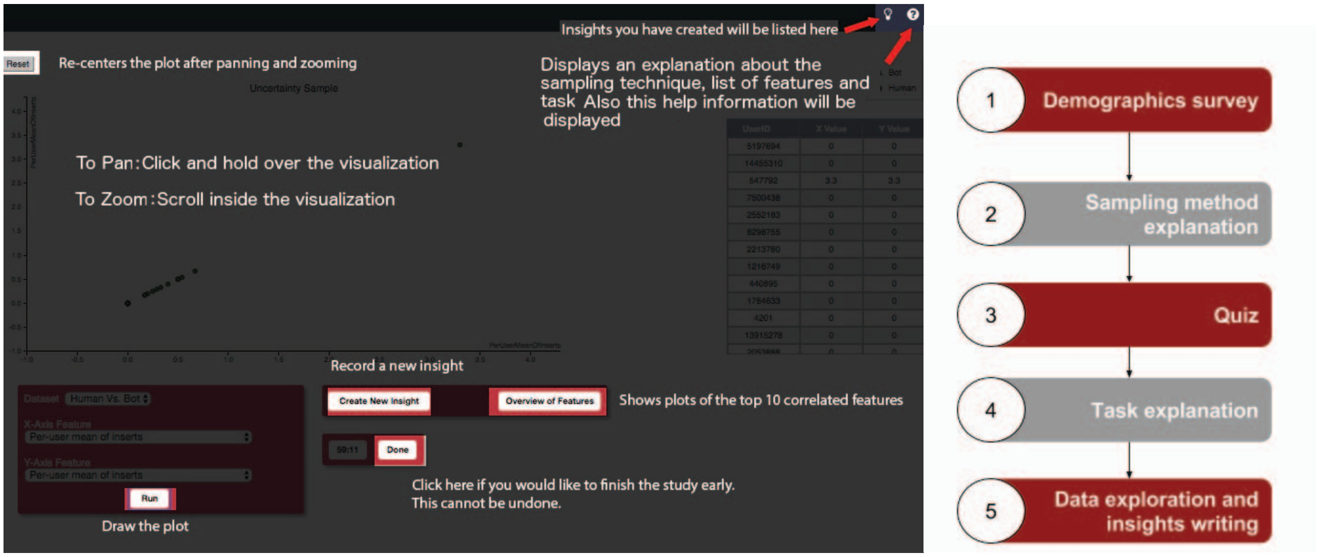
Table 2. Participants and insights per condition

Total number of participants	Total number of insights	Condition
10	43	Random
9	41	Densitv
9	44	Uncertaintv
6	29	Ouerv bv committee

5.1 Preprocessing

We used several measures to avoid participants from cheating. First, participants had to take a quiz to assess that they understood how the data sampling technique worked. If they failed the quiz twice, we stored a cookie on their computer to impede them from using the same computer to re-take the study. Our second measure was to constrain navigation in the website by forcing them to move step-by-step through the study. Without this measure, participants could skip the quiz simply by looking at the html code and redirecting to the next page. We also checked the insights for valid responses. Some participants wrote only unintelligible insights, wrote the same insight with slight variations multiple times, or wrote insights without interacting with the visualizations, which was revealed through the mouse movement logs recorded by our website. After using the above measures, from a total of 40 participants we discarded 6 users from our final dataset. From the 34 remaining participants, we filtered out unreadable insights, which contained unparseable text and duplicate insights. From a total of 174 insights, only 17 were excluded leaving a total of 157

Authorized licensed use limited to: b-on: UNIVERSIDADE DE AVEIRO. Downloaded on March 14,2023 at 16:02:51 UTC from IEEE Xplore. Restrictions apply.



Insights you have created will be listed here

Re-centers the plot after panning and zooming

Displays an explanation about the sampling technique, list of features and task Also this help information will be displayed

To Pan:Click and hold over the visualization

To Zoom:Scroll inside the visualization

Record a new insight

Shows plots of the top 10 correlated features

Click here if you would like to finish the study early. This cannot be undone.

Draw the plot

Figure 5. Left: Help webpage where is shown to the participant how to interact with the visualization website, Right: Structure of the study

insights. Table 2 shows a summary of the quantity of insights and participants across conditions.

5.2 Insights content analysis

The content analysis is divided into two parts: self-evaluation and feature differences. In the self-evaluation, we summarized and compared the responses to the supplementary questions that participants provided when they wrote each insight (kind of insight, confidence, *etc*.) as described in section 4.4. In the feature evaluation, we coded each insight by the data features that were mentioned explicitly or implicitly. We compare the distribution of features mentioned between the conditions (normalized by the number of insights per condition). Additionally, we applied a *tf-idf* transformation [11] to obtain a measure of importance of the features mentioned in each condition as shown in Figure 6 (Top).

Table 3. Participants selected categories for each of their insights.

Condition	Features	Outliers	General
Random	63%	18%	19%
Density	73%	12%	14%
Uncertainty	48%	37%	16%
Query by committee	80%	10%	10%

Finally, to understand the differences across conditions we identify the idiosyncratic features as defined by Zhao *et al.* [27], by measuring the Euclidean distance of the features per condition to the average feature weights across all conditions. This distance then is used to rank the features for each condition. This approach allows us to understand which feature weights help differentiate each condition.

5.2.1 Results

We first compared the participants' responses to the self-evaluation questions. In general, we can see that the participants' self-reported answers about their insights were different for each condition (Table 3). When selecting whether they were confident in their insights, participants in the Density condition were more confident than in any other condition by a large margin: 90%, compared to 65% for random, 63% for uncertainty and 51% for QBC. Participants also had differences in their ability to generate a hypothesis about their data. Participants in the QBC condition did not have a hypothesis 86% of the time compared to 65% for uncertainty, 46% for density and 44% for random. Finally, participants focused on different categories of insights (Table 3).

In particular, participants in the uncertainty sampling condition made a large number of insights about outliers compared to the other conditions, which is unsurprising given the bias in the sample towards outliers.

For the second part of the content analysis, we compared the features that were mentioned in the insights to determine whether participants were focusing on similar or different parts of the dataset. Figure 6 (top) shows the relative importance of insights that contain each feature in the dataset. The values in the horizontal axis correspond to the feature number as shown in Table 1. Each condition's top 3 features are shown in Table 4. We compared each pair of conditions using the Spearman's rank correlation coefficient to determine if they had similar importance.

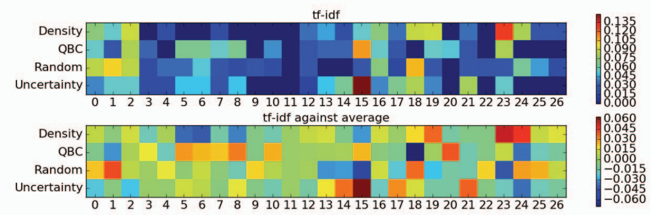


Figure 6. Top: Insights-Feature importance across the different conditions after normalizing and applying the tf-idf transform, dark blue indicates least important and dark red indicates very important. Bottom: Difference of insights-feature importance from the mean calculated across all conditions: highlights those features that are above or below the mean. Dark blue indicates lower than average importance, dark red indicates higher than average importance.

Table 4. Top 3 most representative features for each condition. The (-) sign indicates whether that feature has lower than average importance or (+) higher

Condition	1st	2nd	3rd
Density	(23) Per-user avg time between edits (+)	(24) Per-user std dev time between edits (+)	(6) Per-user-per-page mean of inserts (-)
Query-by-committee	(18) Per-user total unique pages edited (-)	(20) Per-user std dev total edits per page (+)	(24) Per-user std dev time between edits (-)

Authorized licensed use limited to: b-on: UNIVERSIDADE DE AVEIRO. Downloaded on March 14,2023 at 16:02:51 UTC from IEEE Xplore. Restrictions apply.

Random	(15) Per-user most frequent hour edited (-)	(23) Per-user avg time between edits (-)	(1) Per-user mean of deletes (+)
Uncertainty	(15) Per-user most frequent hour edited **(+)**	(21) Per-user avg minor revisions **(+)**	(24) Per-user std dev time between edits **(-)**

No correlations were found (Table 5), indicating that participants using each sampling algorithm focused on different features in the dataset.

The differences in features are also apparent upon inspection of Figure 6 bottom. For example, the importance of features in the Density condition for features 5,6 and 8 is lower than average, while the importance of these features for QBC for these features is higher than average. We observed the similar pattern for random and uncertainty for features 13-15. This finding indicates that by using two or more techniques in conjunction during data exploration, data scientists could potentially focus on more of the features rather than a small subset of them.

Table 5. Spearman's rank correlation for each pair of conditions

Condition	**Condition**	***Correlation***	***p-value***
Density	Query-by-Committee	0.05	0.77
Density	Random	-0.008	0.964
Density	Uncertainty	0.21	0.28
Query-by-Committee	Random	0.08	0.65
Query-by-committee	Uncertainty	-0.11	0.57
Random	Uncertainty	-0.04	0.8

5.3 Insights quality analysis

Given that the content of the insights appears to be different for each sampling technique, we then focused our analysis on whether the quality of the insights was reduced with the different samples. Two human raters assessed each insight for three quality metrics: Complexity, Depth and Relevancy. The average of each quality metric was calculated for each insight, and statistical tests were used to compare the values across conditions. We used the following rubric to rate each insight for each of the three quality measures and provide example insights from the study. This rubric is based on Saraiya *et al.* [23] characterization of insights .

Complexity: Number of features used (explicit and implicit).

Low complexity: Two or fewer features

> *"Bots tend to spend less average time between posts compared to the total changes made by the user"*

Medium Complexity: Three to four features

> *"Bots make less but larger deletes per page. Humans make more but smaller deletes, Bots make single deletes per page, humans make more deletes per page"*

High Complexity: More than four features

> *"For majority of the features like number of edits, inserts, deletes have much lesser values for bots than that of humans. This is expected as humans are prone to error and changing their minds but bots are not."*

Depth: An insight is deeper if it builds on previous insights [23].

Low depth: Usually this kind of insight does not have a hypothesis associated with it, and is mainly descriptive without any conclusion or hypothesis.

> *"An abnormally high number of bots average use at 4 am in the morning"*

Medium depth: Great detail but poor hypothesis or poor detail with great hypothesis.

> *"I found the graph Per-user Per-Page Mean Of Inserts vs Per-user Largest-Single-Add splits human and bots well. The bots tend to add things randomly large but in few times. While human tend to have more times to think so they have larger mean of inserts."*

High depth: This kind of insight is very detailed and has a closing hypothesis.

> *"This is a clear difference we have with bots: Working hours. In this graph (most freq. hour of inserts vs most freq hour edited), we see that human data peak at 9PM, whereas bots seem to have a more uniform distribution. We sleep during 12-8AM and work from 9AM-5PM typically, so around 8-9 PM is a natural time to do most edits for humans! For bots, they don't have sleeping habits that I know of, or 9-5 work hours, so it's more uniform for them. Hypothesis, Bots don't sleep or go to work 9-5, and hence the uniform distribution for their changes / edits frequency throughout the day."*

Relevancy: Relevancy is a quantitative measure of insight quality in which we compare the features mentioned in each insight to the pre-computed importance of each feature to a Bot/Human detector classifier. The relevancy of each feature can be seen in Table 1, and they were determined as follows: High relevancy features are those that are mentioned in the literature as being useful at detecting bots vs. humans, Medium are features we determined to be useful, and low importance features were not found in the literature and do not appear to be useful to the task.

Two members of the research team independently rated every insight. To address rating disagreements, a scores difference was calculated and then used to rank the conflicts. The scores with highest difference were discussed until reaching agreement and until a high Cohen's-Kappa for all the factors was achieved: Complexity *k=0.84*, Depth *k=0.91*, Relevancy *k=0.91*. After rating every insight for the 3 quality measures, we transform the quality measures to numerical values: High = 3, Medium=2 and Low =1. We then calculate the mean to produce a quality score for each insight.

5.3.1 Results

In order to compare the quality of the insights generated between conditions, we used the Kruskal-Wallis H test, since this test does not assume normality of the data.

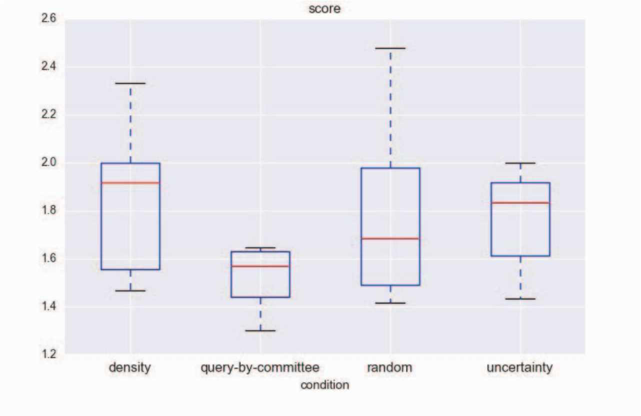


Figure 7. Insights Quality distribution for all the conditions

There was no significant difference in the quality of the insights for the four conditions (*H=2.86, p=0.41*). This can be confirmed visually as shown in Figure 7. It is worth mentioning, however, that these results are based on a small sample size that is within the limits of use of the Kruskal-Wallis test [16].

Density has the highest median value (1.9), followed by uncertainty (1.8), random (1.6) and QBC (1.5). However, not surprisingly, there was high variance in the quality scores.

Kruskal-Wallis H tests were also computed for each quality factors separately across conditions (Figure 8) but there was not a significant difference: Complexity (*statistic=2.43, p=0.48*), Depth *(statistic=1.4, p=0.70)* and Relevancy (*statistic=7.7, p=0.051),* which was marginally significant. In general, although we cannot state that the quality of the insights across conditions is the same,

Authorized licensed use limited to: b-on: UNIVERSIDADE DE AVEIRO. Downloaded on March 14,2023 at 16:02:51 UTC from IEEE Xplore. Restrictions apply.

the box plots show that they are of comparable in overall quality and for the individual quality factors. We conclude that while the content of the samples was different between the sampling conditions, the quality of the insights across different conditions was not significantly different.

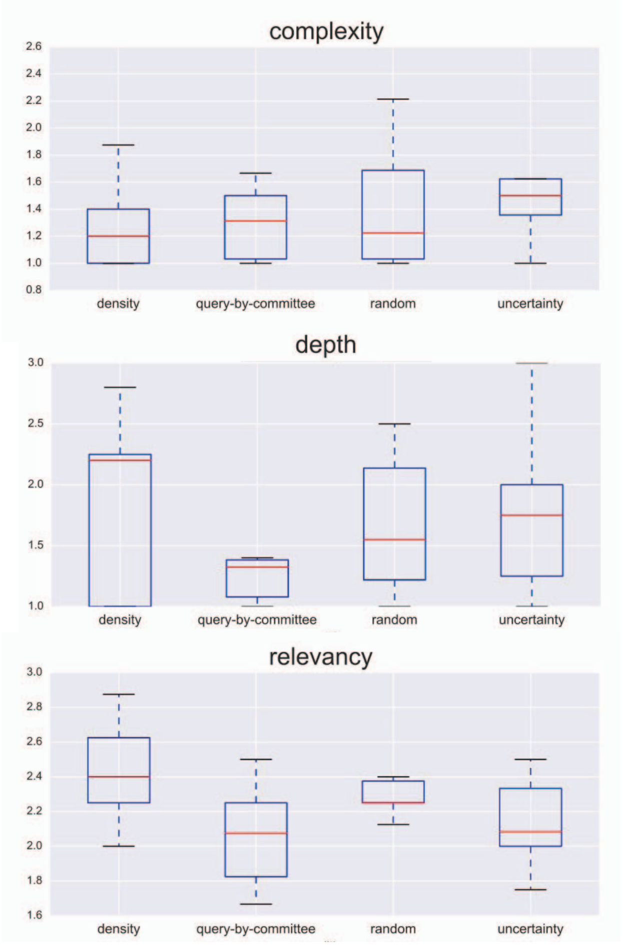


Figure 8. Insights quality measurements distribution across conditions

6. Discussion

Our analysis of insights quality revealed that all of the conditions have comparable levels of quality including the individual quality measures of Depth, Relevancy and Complexity. Independent of the condition, the quality of the insights is not significantly affected positively or negatively by using density, QBC or uncertainty sampling compared to random sampling. It is interesting that even techniques like uncertainty and QBC that are purposefully selecting data points that lie in conflicting regions of the data do not hinder data scientists from producing quality insights on par with random sampling. The characteristics of individual sampling techniques can be seen in the content analysis results. For example, participants using density sampling felt much more confident about their insights than in any other condition by a large margin. We attribute this to the natural tendency of density sampling to filter out outliers and pick data samples from high-density regions in the dataset. The consequence of this is that users may have felt like the data was more homogenous than it actually is as the sampling approach concealed the more extreme values. Despite this, the quality of the insights was not affected and although not significantly different, density has the highest median depth and relevancy scores, which in both cases are higher than 2 (the medium value). Relevancy in this case is especially important because it is an indirect measure of correctness of the insight, meaning that the features described in the insights by participants in the density condition were recognized in the literature as being useful for detecting bots and humans. This is a meaningful result given the context of our study: Our participants were from a diverse background in skills and work experience, and yet they were able to pinpoint the most important features for the task in a very limited amount of time and with no previous experience with this specific dataset.

We also found that density sampling produces insights with a very low importance for per-page related features (features 6 to 11). These features are particularly interesting because they break down inserts, deletes and changes for each Wikipedia article highlighting more specific behaviors. This further shows that density sampling helps data scientists focus on general trends and even features that better describe those trends.

QBC sampling produced insights with higher than average importance for features 5 to 8 and 10. Most of these features are “per user per page” based features. This is a very interesting finding that shows how density sampling and QBC could be complementary: density sampling highlights population trends while QBC highlights more specific trends.

Uncertainty sampling produced insights with higher than average importance for features 13-15, 17, 18, 21; these are all features focused on outliers since uncertainty sampling does sample from regions where the certainty of the classifier is low. This is further supported by the participants’ answers to the question: *What is your insight about?* Participants answered 37% of the time that their insight was about outliers (with the next highest condition being at 20%). Uncertainty sampling is complementary to random sampling, *i.e.* random sampling had lower importance values for features 13-15, 17 and 21.

In general as shown in Figure 6, the different sampling techniques have different importance weights for each of the features in the data set and some of them are complementary to each other. Our study validates that different sampling techniques can generate insights that help to focus a data scientist on different aspects of a dataset, without loss of quality, when compared to the most commonly used sampling technique, random sampling.

6.1 Limitations

As a remark about our study design and survey, one of the biggest challenges was finding participants. Although there was a monetary compensation, the motivation for the participants to participate was very low. Despite reaching out to online groups with thousands of members, talking to students taking advanced machine learning courses and reaching alumni from related programs, we had a very low turnout. From study pilots we received feedback from the participants stating the payment was low when compared with the time spent and the effort required. As an example, professional data scientists are paid an average of $US60 per hour in the U.S. Graduate students, although paid much less have very limited time, which conflicted with our 1.5 hour long study. We hope that with these results we encourage the voluntary participation of data scientists in this kind of study, as their input is very valuable and necessary for creating and understanding new methods for exploratory data analysis of Big Data. We also found that QBC had the highest number of participants removed due to cheating and unintelligible insights. Despite conditions in the experiment being exactly the same, this cheating behavior cannot be attributed to the

Authorized licensed use limited to: b-on: UNIVERSIDADE DE AVEIRO. Downloaded on March 14,2023 at 16:02:51 UTC from IEEE Xplore. Restrictions apply.

sampling technique alone and a follow up study with an interview maybe required to find out the reason.

7. CONCLUSION AND FUTURE WORK

As our results show, not only do multiple sampling techniques produce insights with a quality comparable to insights produced with current practices (random sampling) but also they help data scientists focus on specific aspects of a dataset. These results follow the theoretical properties of the different sampling techniques. For example, density sampling filters out outliers and helps focusing on general trends, QBC looks at data points that lie at the boundary of two classes, hence helping to focus on features that do not capture general trends, while uncertainty focuses on outliers and features that highlight them. We also found that these sampling techniques are complementary and should be used together. For example, density and QBC sampling together could produce richer insights than either one alone; similarly, random sampling and uncertainty sampling are complementary to each other.

More generally, our results support claims in the literature [5,16,19] for using multiple models and sampling strategies in data science. In addition to our own results, we argue for using not just one but all of the sampling algorithms to generate a richer understanding of the dataset. In our future work, we will test the hypothesis that data scientists could generate insights with higher quality and richer contents than with any individual sampling technique alone. For future work, an aspect of the results that deserves more investigation is the estimation of possible biases caused by the goal task. In this study we used classification as the goal task, and although we believe our results generalize to other goals tasks there might be aspects of other goals tasks that are sensitive to the different sampling methods used.

REFERENCES

1. B Thomas Adler, Luca De Alfaro, Ian Pye, and Vishwanath Raman. 2008. Measuring Author Contributions to the Wikipedia. *Proceedings of the 4th International Symposium on Wikis*: 15:1–15:10. http://doi.org/10.1145/1822258.1822279

2. Rui Castro, Rebecca M. Willett, and Robert D Nowak. 2005. Faster Rates in Regression Via Active Learning. *UW-Madison Technical Report ECE-05-3*: 1–46.

3. Yu-An Chung, Shao-Chuan Lee, Yao-Yuan Yang, Tung-En Wu, and Hsuan-Tien Lin. 2015. Pool-based Active Learning in Python. Retrieved from https://github.com/ntucllab/libact

4. Ido Dagan and Sean P. Engelson. 1995. Committee-Based Sampling For Training Probabilistic Classifiers. *In Proceedings of the Twelfth International Conference on Machine Learning*: 150–157. http://doi.org/10.1.1.30.6148

5. Pinar Donmez, Jaime G Carbonell, and Paul N Bennett. 2007. Dual Strategy Active Learning. *Machine Learning ECML 2007*: 116–127. http://doi.org/10.1007/978-3-540-74958-5_14

6. Danyel Fisher, Igor Popov, Steven Drucker, and Monica Schraefel. 2012. Trust Me, I'm Partially Right: Incremental Visualization Lets Analysts Explore Large Datasets Faster. http://doi.org/10.1145/2207676.2208294

7. C Lee Giles. 2007. *Learning on the Border : Active Learning in Imbalanced Data Classification*.

8. S. Kandel, J. Heer, C. Plaisant, et al. 2011. Research directions in data wrangling: Visualizations and transformations for usable and credible data. *Information Visualization* 10, 4: 271–288. http://doi.org/10.1177/1473871611415994

9. Sean Kandel, Andreas Paepcke, Joseph M Hellerstein, and Jeffrey Heer. 2012. Enterprise data analysis and visualization: An interview study. *Visualization and Computer Graphics, IEEE Transactions on* 18, October: 2917–2926.

10. Daniel a. Keim, Florian Mansmann, and Hartmut Ziegler. 2006. Challenges in Visual Data Analysis. *Information Visualization*, IV 2006: 9–16. http://doi.org/10.1109/IV.2006.31

11. Jure Leskovec, Anand Rajaraman, and Jeffrey David Ullman. 2014. *Mining of massive datasets*. Cambridge university press.

12. David D. Lewis and Jason Catlett. 1994. Heterogeneous uncertainty sampling for supervised learning. *Proceedings of the 11th international conference on machine learning (ICML'94)*: 148–156. http://doi.org/10.1016/B978-1-55860-335-6.50026-X

13. Jimmy Lin and Dmitriy Ryaboy. 2013. Scaling big data mining infrastructure: the twitter experience. *ACM SIGKDD Explorations Newsletter* 14, 2: 6–19. http://doi.org/10.1145/2481244.2481247

14. James Manyika, Michael Chui, Brad Brown, et al. 2011. *Big data: The next frontier for innovation, competition, and productivity*.

15. Hall Mark, Frank Eibe, Holmes Geoffrey, Pfahringer Bernhard, Reutemann Peter, and H. Witten Ian. 2009. The WEKA Data Mining Software: An Update. *SIGKDD Explorations* 11, 1.

16. John H Mcdonald. 2014. *Handbook of Biological Statistics*. Sparky House Publishing, Baltimore,Maryland. Retrieved from http://udel.edu/~mcdonald/

17. H T Nguyen and a Smeulders. 2004. Active learning using pre-clustering. *Proceedings of the 21st International Conference on Machine Learning (ICML'04)*: 623–630. http://doi.org/10.1145/1015330.1015349

18. C. North. 2006. Toward measuring visualization insight. *IEEE Computer Graphics and Applications* 26, 3: 6–9. http://doi.org/10.1109/MCG.2006.70

19. Travis E Oliphant. 2007. SciPy: Open source scientific tools for Python. *Computing in Science and Engineering* 9: 10–20. http://doi.org/10.1109/MCSE.2007.58

20. F Pedregosa, G Varoquaux, A Gramfort, et al. 2011. Scikit-learn: Machine Learning in {P}ython. *Journal of Machine Learning Research 12*, 2825–2830.

21. Robert Pienta, James Abello, Minsuk Kahng, Duen Horng, and Chau Georgia. Scalable Graph Exploration and Visualization: Sensemaking Challenges and Opportunities.

22. Plotly Technologies Inc. 2015. Collaborative data science.

Authorized licensed use limited to: b-on: UNIVERSIDADE DE AVEIRO. Downloaded on March 14,2023 at 16:02:51 UTC from IEEE Xplore. Restrictions apply.

Plotly Technologies Inc. Retrieved from https://plot.ly

23. Purvi Saraiya, Chris North, and Karen Duca. 2005. An Insight-based Methodology for Evaluating Bioinformatics Visualizations. *IEEE Transactions on Visualization and Computer Graphics* 11, 4: 443–456. http://doi.org/10.1109/TVCG.2005.53

24. Burr Settles. 2010. Active Learning Literature Survey. *Machine Learning* 15, 2: 201–221. http://doi.org/10.1.1.167.4245

25. John W. (John Wilder) Tukey. 1977. *Exploratory data analysis*. Addison-Wesley Pub. Co.

26. Xiaofeng Wang, Matthew S Gerber, and Donald E Brown. 2012. Automatic Crime Prediction Using Events Extracted from Twitter Posts. In *Social Computing, Behavioral - Cultural Modeling and Prediction: 5th International Conference, SBP 2012, College Park, MD, USA, April 3-5, 2012. Proceedings*, Shanchieh Jay Yang, Ariel M Greenberg and Mica Endsley (eds.). Springer Berlin Heidelberg, Berlin, Heidelberg, 231–238. http://doi.org/10.1007/978-3-642-29047-3_28

27. Sha Zhao, Julian Ramos, Jianrong Tao, et al. 2016. Discovering different kinds of smartphone users through their application usage behaviors. *Proceedings of the 2016 ACM International Joint Conference on Pervasive and Ubiquitous Computing - UbiComp '16*, ACM Press, 498–509. http://doi.org/10.1145/2971648.2971696

Authorized licensed use limited to: b-on: UNIVERSIDADE DE AVEIRO. Downloaded on March 14,2023 at 16:02:51 UTC from IEEE Xplore. Restrictions apply.