FERRET: REVIEWING TABULAR DATASETS FOR MANIPULATION

Devin Lange University of Utah
Shaurya Sahai University of Utah
Jeff M. Phillips University of Utah
Alexander Lex University of Utah

ABSTRACT

How do we ensure the veracity of science? The act of manipulating or fabricating scientific data has led to many high-profile fraud cases and retractions. Detecting manipulated data, however, is a challenging and time-consuming endeavor. Automated detection methods are limited due to the diversity of data types and manipulation techniques. Furthermore, patterns automatically flagged as suspicious can have reasonable explanations. Instead, we propose a nuanced approach where experts analyze tabular datasets, e.g., as part of the peer-review process, using a guided, interactive visualization approach. In this paper, we present an analysis of how manipulated datasets are created and the artifacts these techniques generate. Based on these findings, we propose a suite of visualization methods to surface potential irregularities. We have implemented these methods in Ferret, a visualization tool for data forensics work. Ferret makes potential data issues salient and provides guidance on spotting signs of tampering and differentiating them from truthful data.

1 Introduction

Data manipulation is an unfortunate reality of the scientific publication process. Like plagiarism, it is an unethical attempt to game the system, usually to further academic careers. The effects of falsified data in research vary. Manipulated data and the resulting incorrect claims can mislead scientists who want to build on the incorrect knowledge or lead to actions not based on evidence. Manipulated data can even lay a faulty foundation for a whole area of research, leading to years of wasted effort by researchers. At worst, incorrect and dishonest findings can result in the inappropriate application of knowledge in society, with potentially severe consequences, such as the harmful treatment of patients. In a far-ranging Alzheimer's scandal [Pil22], image and numerical data were suspected to have been manipulated in what was considered one of the most important publications on the topic. Based on this — now considered false — knowledge, drugs were developed and even FDA-approved, exposing patients to potentially useless medication while foregoing alternative treatments and causing side effects. Pharmaceuticals have also invested "millions of dollars, or even billions" [Pil22] based on the manipulated findings. Unlike plagiarism — which can be discovered by checking articles against other published sources — falsified data is difficult to detect.

Plagiarism checks are now part of the editorial process of many conferences and journals. However, in several high-profile data manipulation cases, scientists have had seemingly productive careers, and only after a single case of misconduct surfaced did the community critically scrutinize their whole academic record to find many instances of wrongdoing [Vig20].

Besides urging individuals to refrain from such activity, how can we prevent or at least mitigate this problem? To address this issue, we look to peer review, a cornerstone of the scientific process. Although peer review has known flaws, the premise of peer review is that experts can verify the soundness of the research and increase the quality of published works. So why is fabricated data not caught in this step of the publishing pipeline? There are many factors: reviewers may assume a good-faith effort by their peers and are not looking for falsified data. In addition, combing through data to find signs of malpractice is difficult and time-consuming, especially when reviewers are not educated on what to look for and have no tools that can help at their disposal. Also, checking data requires that the data is made available to the reviewers and, subsequently, the readers, a practice gaining momentum with the open science movement but still far from universally adopted [Har18].

Existing tools that help find cases of data fabrication tend to focus on finding duplicated regions in images. The goal of our work is to equip editors, reviewers, and scientists with the knowledge and tools to make the investigation of fabricated tabular data feasible. In particular, the tools we provide are designed to aid and enhance human judgment,

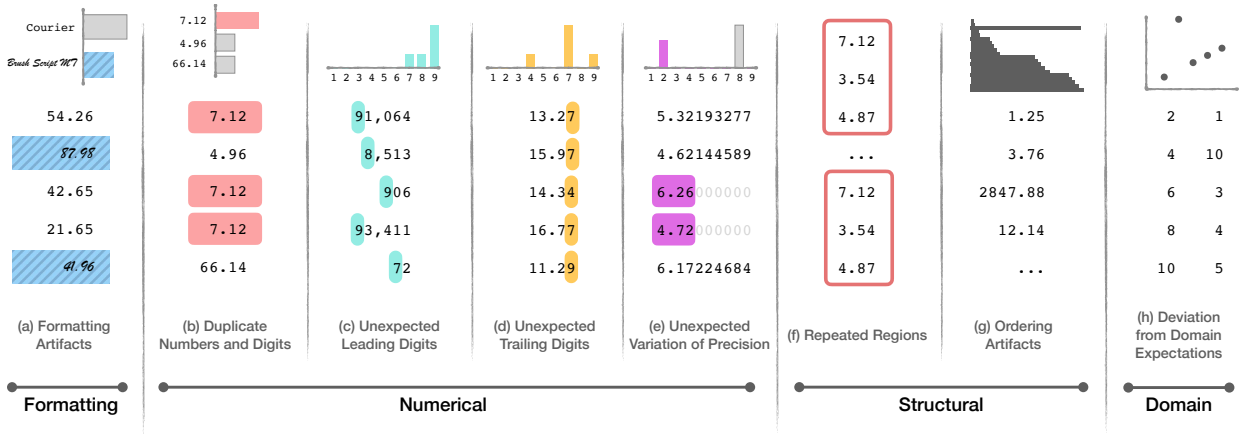


Figure 1: Artifacts of manipulation we discovered when analyzing manipulated datasets. The artifacts range from (a) unexpected formatting; to numerical issues such as (b) duplicates, unusual distributions of (c) leading or (d) trailing digits, and (e) variations in precision; to structural issues such as (f) repeated regions and (g) artifacts associated with sorting and ordering items; and to (h) the unrealistic relationships in the data.

as scientific data can be noisy and extremely varied, and alternatives leveraging automated statistical analysis can potentially encourage false accusations.

Our work has two primary contributions: first, we identify common artifacts of data manipulation using a combination of analyzing datasets known to be manipulated and interviews with researchers investigating fraudulent datasets. Second, we propose an array of design principles and visualization methods to saliently surface these artifacts, thereby enabling experts to easily and confidently identify fraudulent datasets.

As an additional contribution, we have developed Ferret, a prototype in which we have implemented these visualizations. In addition to these different visualization designs, we also include guidance on interpreting the results directly in the tool. Since artifacts can arise both from falsified and truthful data, it is important for users of the tool to have guidance on interpreting the results without being prescriptive in how they use the tool.

We evaluate our methods and our tool using case studies from a series of known fraudulent datasets, demonstrating that these patterns become evident by leveraging Ferret. Finally, we discuss the ethics and the potential for abuse of our approaches.

2 Related Work

We are unaware of research on using interactive tools to detect manipulated tabular data, but approaches have been published on detecting duplications in more general cases, detecting errors in spreadsheets, and detecting manipulations with numerical methods, which we discuss in this section.

2.1 Detecting Duplicated Data

Data duplication and data manipulation share some commonalities because copying and pasting parts of a dataset is a common approach. The duplication of data, broadly speaking, is of interest in many domains and for many types of data. The detection of text plagiarism is an active research field [FMG19]. Even though plagiarism detection is not a solved problem, progress has been made, as is evident by the use of plagiarism-checking tools in many journals' review processes.

Detecting software plagiarism is a similar problem. The most widely used tool for determining software similarity is MOSS [SWA03]. The authors of MOSS argue that it should not be used as an automated tool but rather as a way to surface potentially questionable data to reviewers, which is consistent with our motivations. Duplication of software can also occur when programmers copy/paste regions of code and then modify them. Detecting such copies is of interest to software engineers [BKA*07]. Similarly, in spreadsheet programs, an analyst may copy and paste a table, and just like in code, when one is updated, copies possibly should be as well. Hence, methods to detect such copies exist [HSPv13, ZDZ*20]. Although detecting copies in tables shares some similarities with detecting manipulation in datasets, the structure of intentional clones compared to duplicated regions due to data manipulation cannot be expected to be identical. Also, duplication is only one of the many artifacts, as we discuss in Section 5.

Some forms of image manipulation create duplicated regions, such as the use of Adobe Photoshop's Clone Stamp tool. Image manipulation through duplication is a common problem in science [Bik22]. Even though much of the work on identifying manipulations remains manual, recent work relies on machine-learning techniques such as CNNs [WWZ*19, LH19, BNTZ20, YLL*20, BCM*21, DCH*23, KNY*22]. CNNs are effective for images,

where large datasets can be acquired or generated. However, tabular datasets are more varied in their structure. More importantly, the context associated with the data is also critical for interpretation. A tabular dataset could be completely plausible given one context and obviously manipulated in another. Such contextual understanding is difficult to encode in machine learning models today, especially with limited data, as is the case for manipulated tabular datasets. Therefore, we believe that a human-in-the-loop approach is needed when detecting manipulation in tabular data.

2.2 Detecting Errors in Spreadsheets

Detecting unintentional errors in spreadsheets is a well-researched problem [PBL08, Boc16]. Basic patterns of errors have been categorized [SNR17]. The detection of formula errors [BBZ18] is useful when working with spreadsheets. Detecting structures that can lead to errors in tables ([CCLC16]) shares commonalities with our work since table clones are one of these structures. Beyond these structures, other methods for detecting errors in spreadsheets exist [JSHW14, KSJ*21, LWX*19b, LWX*19a, HXJ*20]. However, unintentional errors do not always produce the same artifacts as intentional manipulations, so such techniques will not identify all manipulated datasets.

2.3 Detecting Manipulation With Numerical Methods

An alternative to our interactive visual system is to inspect data for statistical anomalies. Rules such as Benford's law [Nig12, Mil15], which states that the leading digit is more likely to be 1 and then 2 than the large digits 7, 8, 9, have been used in domains such as accounting [DHP04]. In these settings, financial fraud has been spotted by observing that Benford's Law was violated over a series of transactions; not as many had leading digits of 1 or 2 as suspected. Statistical hypothesis tests can be leveraged to assess the statistical significance of deviation from this expected distribution of leading digits [Mil15, NM09]. However, to apply Benford's law, assumptions must be made on the background (null) distribution. In particular, the distribution must span multiple orders of magnitude, which applies to some data, e.g., in astronomy or finance, but not in many others. For instance, the time in seconds to run a mile in a professional competition will almost surely start with a 2 (the current world record is 223 seconds).

Similar concerns exist in applying any statistical hypothesis testing method to look for anomalous patterns in data. All these methods start with an assumed background (null) distribution and look for a fixed type of pattern that may deviate from it. However, the *choice* of the background distribution requires domain knowledge and human judgment. Hence, one should not automatically or generically apply tests such as those for Benford's law. In the tabular datasets we analyzed, we rarely could apply such tests. As a result, we decided not to include statistical tests because we believe that they would lead to numerous false-positives (claimed detection when a wrong background distribution was assumed).

Beyond statistical tests, there are numerical tests that do not check for statistically unlikely data but rather numerically impossible data. Notably, StatCheck [RNE16] checks for internal consistency of statistical measures. StatCheck is used in some peer-review processes. However, it has received criticism due to concerns for its accuracy and its automated testing of papers [Cha17]. The convenience of these automatic systems carries the risk that they will be used without providing authors the opportunity to respond to claims. Furthermore, if the raw dataset has been manipulated before a correct statistical analysis is run, StatCheck cannot identify any errors. Another algorithm for detecting manipulation is described in Park et al.'s work [PSL21] on detecting (and recovering) integer data when it has been multiplied by a nonintegral real number and has been rounded. Such techniques can be useful in the right situation, but they are limited to identifying a narrow set of problems.

2.4 Visualization Systems

Many systems visualize tabular data, but without a focus on detecting manipulations. The Table Lens [RC94] and Taggle [FGS*20] inspired our tabular layouts. Similarly, Domino [GGL*14] and SMARTexplore [BBS*18] link tabular data to visualizations. However, our focus is on custom visualizations and descriptions specifically designed to expose artifacts. Most closely related is Taco [NSH*18], a system for comparing similar datasets, but Taco cannot be used to find patterns of similarity within a single table.

3 Methods

To further understand how datasets can be manipulated, we collected datasets with known issues predominantly associated with retracted publications. To identify fraudulent datasets, we leveraged a database collected by the Retraction Watch Project, a website that tracks retractions in their database and disseminates them through blog-style articles [Ora10], through community feedback on social media, and through interviews with two researchers who have investigated and reported evidence of data falsification. In total, we identified 10 datasets, with strong evidence that some manipulation occurred on them, summarized in Table 1.

We obtained a complete version of the Retraction Watch Database [The18] through a special request to the database curator. Since these papers were retracted for various reasons, we filtered to papers that included "Falsification/Fabrication of Data" as one of the reasons for retraction, resulting in 1161 candidate papers. Next, we manually examined the papers to find fraudulent datasets by reading the official reason for retraction and checking the retracted publication for any references to public data. After checking 103 papers, we found only a single tabular

dataset with signs of manipulation. We hypothesize that this low success rate is due to a focus on manipulated images in the database and because authors who manipulate data are incentivized to not publish it.

We then elicited help through social media. Using this approach, we identified four datasets associated with retracted papers. All four datasets also have an associated blog post where the evidence for manipulation and process of investigation has been posted. Search for the paper titles in the Retraction Watch Database revealed that these papers are in the dataset but were not flagged with "Falsification/Fabrication of Data". Three of the four include a flag related to data, such as "Error in Data," "Unreliable Data," and "Concerns/Issues About Data." The fourth paper, which is a preprint, has only a flag of "Notice - Limited or No Information." We suspect that these less serious classes may have been used due to an abundance of caution by editors.

We also interviewed two researchers who have investigated and reported evidence of data falsification. These interviews provided us with two additional datasets. One of these was in the Retraction Watch Database, again without the "Falsification/Fabrication of Data" flag but with other flags related to data. The other paper was not in the Retraction Watch Database as of Oct 28, 2022. The interviews also introduced us to patterns of manipulations and approaches for data fabrication that these experts had encountered. For example, we had not considered checking the plausibility of the data in a larger, domain-specific context. The interviews also provided additional context for how analysts search for anomalies.

To find common patterns of artifacts across datasets, we performed a primary analysis of the data in Excel and Ferret. We also reviewed existing discussions of anomalies in the data in published works, blog posts, and online forums such as PubPeer.

Name	Status	Statement	Domain	Blog
DS-Priming	R	[Edi16]	Mrkt.	[Cha21] [PRA*16]
DS-Driving	R	[Edi21]	Psy.	[SSN21]
DS-Covid	W	[Law21]	Med.	[Bro21]
DS-Gaming	R	[SKV*20]	Med.	[Bro20]
DS-Spider-P	R	[LMD*20]	Bio.	
DS-Spider-E	R	[LP20]	Bio.	[Las20]
DS-Spider-I	R	[LMP20]	Bio.	
DS-Glioma	R	[Wan19]	Med.	
DS-Fly	C	[EB21]	Bio.	[Aut20]
DS-Fish	R	[Tho22]	Bio.	[Ens21]

Table 1: Table of datasets associated with retracted or withdrawn papers. Clicking on the dataset name will open Ferret with the dataset loaded. The *Status* column indicates whether a paper was retracted (R), withdrawn (W), or has earned an expression of concern (C). References in the *Statement* column link to the retraction statement. References in the *Blog* column link to blog posts that discuss how the data was manipulated.

4 Datasets Overview

As described in the previous section, we collected datasets that contain data manipulations associated with retracted papers. All datasets are listed in Table 1. Here we briefly introduce a subset of these datasets, and how they were likely manipulated so that it is easier to understand the artifacts present in the datasets.

DS-Driving This dataset comes from a retracted study on honesty in the field of psychology. One experiment asked participants to report the odometer mileage of their car both before and after some period of time. It appears that the "after" column was generated by adding a random number between 0 and $50,000$ to the "before" number. In addition, half of the rows also appear to be generated by adding a small amount of noise to the original values.

DS-Gaming In this study, a survey was sent over email asking about video gaming habits, demographic information, and sleeping habits. The paper contains a table with summary statistics that include duplicate regions.

DS-Spider-E This study measured the "boldness" of spiders by recording how long it will take spiders to reemerge from their enclosure after a simulated predator attack. The dataset includes a large number of duplicates, as well as repeated regions.

DS-Fly In this study, the sizes of flies were measured, as well as the distance they traveled. Both measurements include values that have a high degree of precision, with roughly 16 digits after the decimal point, as well as values with a precision of two.

5 Artifacts of Manipulation

The act of manipulating or completely fabricating a dataset can leave behind signs: We call these signs artifacts of manipulation. As shown in Fig. 1, we have organized these artifacts into four common categories: **formatting**, relating to how the data appears in the data files; **numerical**, relating to patterns of numbers and digits in and across columns; **structural**, relating to patterns that appear when analyzing multiple rows or columns together; and **domain**, relating to patterns that show impossible or implausible effects in the data given the meaning of the data. The types of artifacts we found in our ten datasets are summarized in Figure 2. It is important to note that since these artifacts are derived from a limited collection of manipulated datasets, the list is only a starting point. To track artifacts that may be discovered in the future, we have created a living document of artifacts and invite others to suggest changes. Also, the presence of artifacts is not always an indication of wrongdoing — they can be produced by a

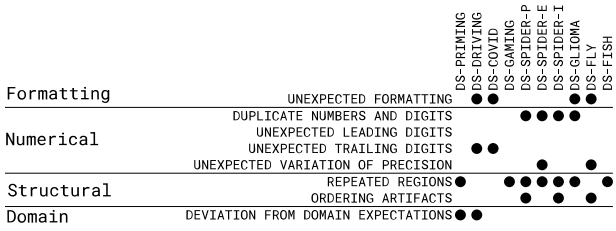


Figure 2: An overview of which datasets exhibit which artifacts.

valid data processing step or be an artifact of the data collection methodology. Finally, it is sometimes difficult to distinguish intentional wrongdoing from honest mistakes while working with data. We discuss the implications of our visualization design in Section 6 and the necessary care in our section on broader impacts (Section 10).

5.1 AR-Formatting: Formatting Artifacts

Spreadsheet tools like Excel or Google Sheets allow users to format the appearance of the data, including choosing a *font*; *font size*; methods of text emphasis such as *bold*, *italics*, and *underlining*; and background *colors*. In addition, users can select a *data format*. For example, changing a cell to a date format will alter how the cell is displayed without changing the underlying information. These formats can be flexibly chosen for cells, columns, or rows, and combinations thereof. Formatting is typically consistent and logical in inconspicuous data. However, the occurrence of odd patterns of formats can hint at manipulation, as illustrated in Fig. 1a. For instance, in the study on honesty (**DS-Driving**), it appears that data was copied to a temporary file, where the values were modified, and later copied back into the master spreadsheet. Notably, these two files seem to have used different fonts, so in the final dataset, exactly half of the rows contained text using the font *Calibri* and the other half using *Cambria*. These rows were interspersed, likely due to shuffling or sorting the table after augmenting it. In total, we found evidence of strange formatting in three of the datasets we collected (**DS-Driving**, **DS-Covid**, **DS-Fly**).

This type of artifact, however, could also appear in authentic data, for example, when assembling a dataset from multiple data sources. Whether or not such a pattern is a sign of manipulation will depend on details, such as whether a whole column has a different format (likely not suspicious), or whether individual cells are formatted differently (possibly suspicious).

5.2 Numerical

The variety of possible numerical artifacts left behind by bad actors is considerable. Here we describe common types that we have observed. All these artifacts occur both in individual columns and across columns.

5.2.1 AR-Duplicate: Duplicate Numbers and Digits

This artifact describes cases when (whole) numbers or sequences of digits (parts of numbers) are repeated more frequently than expected (Fig. 1b). Encountering duplicate numbers or digits can suggest that data was copied and pasted or manually entered. When measuring a natural phenomenon, there is typically variation in the data, either from differences in the signal being measured or from noise introduced by the tools used to measure the signal. For a specific number of values sampled from a distribution at a specified precision, a certain number of duplicate values can be expected. With more duplicates than expected, a few possible explanations can be suggested. First, the underlying distribution could be different than expected. For instance, a narrow Gaussian distribution would result in more duplicates than a wider one. Next, low precision generally would make duplicates more likely.

A common cause of duplicate numbers and sequences of digits that may seem suspicious at first, but is typically innocent, is high-precision duplicates caused by converting measurements. For example, converting fractions to decimals could introduce duplicates with seemingly high precision. If an experiment recorded the length of an animal in inches as integers, but in a subsequent step, the data was converted to feet using decimals, we would expect that the resulting decimals have values with high precision, such as 0.33333333 and 0.41666667. In this case, the number 0.33333333 may appear more often than naively expected and an n-gram of digits, such as 3s, or 6s may appear frequently.

Another common cause of duplicates that is likely innocent is thresholding or reaching a maximum value. In many scientific experiments, there is a terminating condition, such as a maximum time of the experiment or a score corresponding to a maximum achievable value. For example, whereas the spider datasets (**DS-Spider-E**) are definitely manipulated, they also measured only a time period of ten minutes, recorded as 600 seconds, and that maximum threshold was reached often. Hence, the frequent occurrence of 600 in such a dataset is likely inconspicuous.

Duplicate numbers can also appear when a dataset is manipulated by copying items or by manually inventing numbers. Humans are bad at generating random numbers [TLB14, SSBW12, FSK08] and random sequences of digits. When humans simulate the process of sampling from a distribution by repeatedly typing numbers, they tend to produce patterns (duplications) that often can be distinguished from collected data. In addition, sequences of digits appear more frequently in fabricated sets of numbers. For example, 54.23 and 23.54 are not duplicate numbers, but they do contain duplicate digit sequences, 54 and 23. We observed a suspicious amount of duplicated numbers and digits in four of our datasets (**DS-Spider-E**, **DS-Spider-P**, **DS-Spider-I**, and **DS-Glioma**).

5.2.2 AR-Leading: Unexpected Leading Digits

Benford's law [Nig12, Mil15] (also discussed in Section 2.3) is an expected pattern of the first digits of numbers in a dataset (Figure 1c). In short, it states that in datasets that span multiple orders of magnitude, the most frequent first digit should be a one, followed by a two, then a three, and so on. For example, in a dataset of the number of people living in cities and villages, we would expect more cities with 100,000–199,999 inhabitants (leading digit 1) than cities with 900,000–999,999 inhabitants (leading digit 9). We have included this artifact in our collection since checking for violations of Benford's law is a known technique for unearthing fabricated data. However, none of the scientific datasets in our collection spreads densely over such multiple orders of magnitude; hence, we did not identify this pattern.

5.2.3 AR-Trailing: Unexpected Trailing Digits

We have also found it useful to examine the last digit of numbers (Figure 1d). In some situations, the last digit of a collection of measurements might represent a randomly sampled uniform distribution. In other situations, different patterns would be expected. For example, in a list of prices for grocery-store products, an increased frequency in the digit nine would be expected since prices ending with 99 are strategically selected to make a product appear cheaper. On the other hand, if people are asked to provide an estimate for a value, we expect a final digit of zero to be more frequent than other digits. For example, if participants at a large event were asked how many people attended, we would expect an answer of $15,000$ to be much more common than $14,872$. We consider a trailing digits artifact to be a mismatch between the expected pattern of the last digit and the pattern observed in the data or an unexplained inconsistency of trailing digits between parts of a dataset.

In **DS-Driving**, two columns represent values where drivers are asked to give the mileage of their car. In one column, this rounding effect — showing a large amount of numbers ending with zeros — is present. In the other column, the trailing digits follow a uniform distribution. The retracted manuscript does not describe any difference in data collection between these two columns that might explain the difference in pattern between the final digits.

Although this example could be identified by shifting the decimal place and performing an analysis on precision, this is not always the case. The frequency of numbers ending in nine would not be noticeable in a precision analysis. Furthermore, a precision analysis of **DS-Covid** would not catch a strange pattern where even trailing digits occur more frequently than odd digits. Vice versa, not all precision artifacts are noticeable through a trailing digit analysis. For instance, the variance of precision of the stopwatch example is independent of the frequency of different trailing digits.

5.2.4 AR-Precision: Unexpected Variation of Precision

Data formatting in spreadsheet programs can also obfuscate data, leading to numerical artifacts that may not be evident in the source spreadsheet. In particular, this obfuscation can occur with the precision of numbers. Numerical data may record a varying number of digits after the decimal places. However, if the data is formatted as a *number*, the default in Excel is to show two digits after the decimal place. We assume inconspicuous data has similar precision for similar observations. Time measured with a stopwatch, for example, would typically have a precision of up to $1/100$ of a second. Most numbers should have two digits after the decimal points, a few with one digit (e.g., exactly 3.1 seconds), and even fewer with no digits (3 seconds). Manipulated data may have extremely varied precision Fig. 1e. This variation could happen if data is recorded (or generated) with a high degree of precision, then manually manipulated to change some values. Such a difference may not be apparent in a spreadsheet program when two digits are displayed. Alternatively, some authentic data could be collected with limited precision, and a function with high precision could be used to generate the rest. However, such a phenomenon could also arise innocently, e.g., when converting between fractions and decimals, as explained earlier. We have observed unexplained varied precision in **DS-Fly**.

5.3 Structural

Beyond the frequency of data or attributes of data, the structure of data can also play a role in detecting manipulation. Structural patterns are concerned with both the value of measurements and the order of the observations in the data file.

5.3.1 AR-Regions: Repeated Regions

Whereas six duplicate numbers may be considered a weak signal of manipulation, two identical sequences of six numbers are a much stronger one. We consider a region to consist of multiple cell values that have a spatial relationship in a spreadsheet, as illustrated in Figure 1f. Regions include nearby cells, vertically or horizontally, and may include gaps. Repeated regions can be artifacts of manipulation. Although some repeated regions could be caused by how the data is collected, such an innocuous structure is likely obvious. For manipulated data, regions are likely copied and pasted multiple times, either accidentally or as a convenient way to augment a dataset. In addition to simply copying and pasting regions, parts of the region are sometimes modified manually, resulting in similar regions with gaps. We saw this type of artifact in seven of our ten datasets, making it the most common artifact (**DS-Priming**, **DS-Gaming**, **DS-Spider-E**, **DS-Spider-I**, **DS-Spider-P**, **DS-Fish**).

5.3.2 AR-Ordering: Ordering Artifacts

It is natural for ordering artifacts to exist in authentic datasets. For instance, if multiple observations are recorded over time, we would expect that time increases throughout the dataset. Our interviews revealed that some experts consider it a good practice to avoid changing the order of a dataset. However, re-sorted data is not uncommon or automatically suspicious.

The ordering of the data can still indicate manipulation, as illustrated in Figure 1g. For example, if a bad actor wants to show that an experimental condition has an effect on the weight of animals, they might sort the data based on weight. Then, they might modify values at the distribution's tails — altering the data to match their hypothesis. This approach is economical since changing the extreme values will have the largest effect on aggregate measurements. However, this approach can leave behind ordering artifacts. This kind of dataset where a column is nearly sorted is one example of an ordering artifact. If the order is reset after modifications, such a pattern might be difficult to detect. However, if the data is reset by sorting on a column with duplicates (e.g., by a categorical value), then the effects of sorting on weight before the reset will still be seen within the groups. This kind of ghost sorting is another variation of an ordering artifact. A different order artifact exists in **DS-Fly**. Here one column has a mixture of high and low precision (**AR-Precision**). Additionally, the cells with low precision do not appear to be randomly interspersed throughout the rows, but rather appear in a repeated structured way.

5.4 AR-Domain: Deviation from Domain Expectations

The artifacts discussed so far are visible in the data's formatting, structure, or values. However, authors may use more sophisticated techniques, such as scripts that randomly sample from a distribution, to generate fabricated data. Manipulation in these datasets may be difficult or even impossible to detect. In these situations, more sophisticated techniques are required to find the artifacts.

Single-Dimensional. For single-dimension data, we often have prior knowledge about how that data should look, at least in the aggregate. For instance, many natural measurements, such as the height of humans, will exhibit a normal distribution. We consider drastic variations from these expectations, such as a uniform distribution occurring when a normal distribution is expected or a normal distribution with an obviously clipped tail, to be a single-dimensional domain artifact.

Relational. With prior knowledge of how data should look and some scripting ability, it may be possible to fabricate a column of data that is indistinguishable from authentic data. However, such fabrication becomes increasingly difficult as more columns are included. For example, if an experiment records the height, weight, age, sex, lab values, etc, it would be more difficult to generate the data while ensuring all relationships remain plausible. In other words, scripts that generate data would have to ensure that height correlates with weight, age, sex, etc. — it is insufficient for each column to pull from its underlying distribution independently.

Relational artifacts can be more nuanced than a missing correlation. In **DS-Driving**, for example, it appears that some data has been programmatically generated by adding a random value between 0 and $50,000$ to create a new column for $13,488$ rows. When comparing the relationship of the cars' mileages before and after a period of time, the miles driven in this period are uniformly distributed between 0 and $50,000$, with many drivers close to the $50,000$, violating an assumption of a smooth distribution.

6 Visualization Design Principles

Analyzing datasets for manipulation is a difficult and potentially fraught endeavor. A claim of manipulation, even during the review process, is a serious accusation and should be levied with caution. Hence, we believe domain experts must maintain agency while analyzing a dataset. Analysis tools should support experts by providing guidance without being prescriptive. To realize this sentiment, we developed design principles to guide our development of Ferret, a visualization tool for reviewing tabular datasets for manipulation. In this section, we introduce these design principles, and we describe the particulars of Ferret in the next section. Some of our guidelines are related to general visualization guidelines, such as Shneiderman's Mantra (overview first, zoom and filter, details on demand) [Shn96], yet we provide more specific guidance for the use case of detecting manipulations in datasets.

6.1 Framing the Exploration

One early observation we made is that the breadth of artifacts of manipulation is significant and that domain knowledge about the data is often necessary to make accurate judgments. Hence, we argue that a human, ideally with domain expertise, is needed to discern whether an artifact is the result of manipulation. However, we also learned that typical exploration of data through visualizations and statistical tests might be insufficient without knowledge of what to look for. In other words, "until you know what to look for, the patterns are not obvious" [Las20]. As a result, our first design principle is to provide guidance by describing artifacts of manipulation and why they may exist (including benign explanations), and providing salient visualizations of artifacts. At the same time, we avoid being prescriptive, for example, by describing why a dataset is manipulated or recommending a particular analysis or statistical test.

This principle is manifested in Ferret in several ways: First, Ferret lists and explains the different types of artifacts (Fig. 3 a and b). The introductory text gives advice on how to spot an artifact but also lays out common benign causes

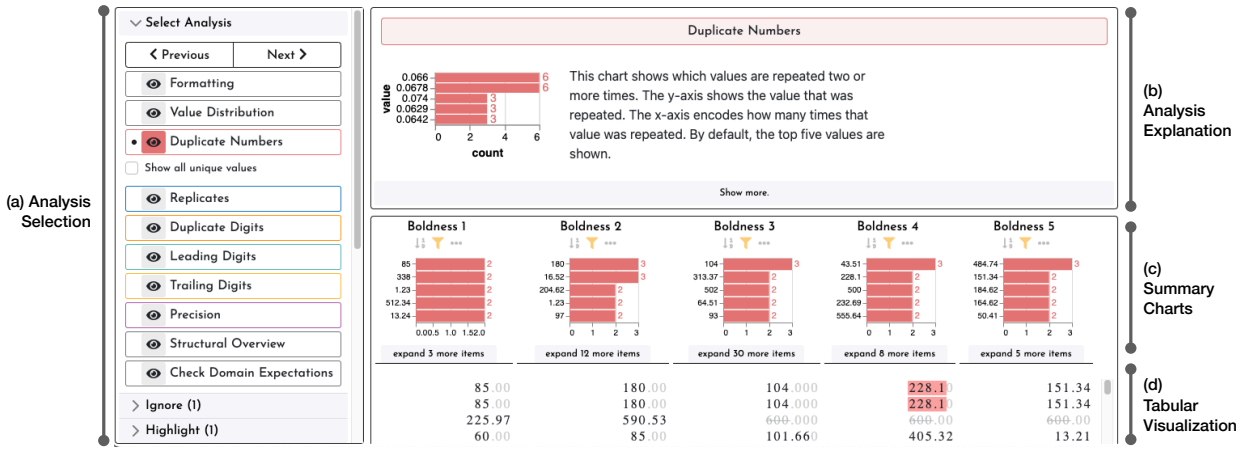


Figure 3: Overview of the Ferret visualization tool. (a) The *Analysis Selection* panel gives quick access to the available analysis modes. (b) An explanation is provided for each analysis mode to help users understand artifacts of manipulation and guard against false positives. (c) The *Summary Charts* display aggregate information for each column in the dataset. (d) The *Tabular Visualization* gives access to the raw tabular data with relevant data surfaced through highlighting and rearranging. Values can be highlighted (such as 228.1) and ignored (600).

of those artifacts. Second, Ferret provides visualizations to identify and confirm suspicious patterns (Fig. 3 c and d). Finally, Ferret refrains from using statistical tests to identify issues. Our argument for not using tests is that most tests would be valid only under narrow circumstances (such as a specific type of distribution), and that the danger of inappropriately using a test outweighs the benefits.

6.2 Make Artifacts Salient

Our next design principle is to make artifacts of manipulation salient. Since our first design guideline necessitates an expert human to investigate the data, our goal is to make that investigation more efficient by quickly exposing artifacts of manipulation. This principle is best illustrated with an example: It can be difficult to notice things like the difference between Calibri and Cambria fonts or between 11.9 and 12-point font sizes in traditional spreadsheet software. In Ferret, cells with deviating formatting are highlighted with a distinct background color and pattern (Fig. 8a) so that differences are salient. Similarly, spreadsheet tools will often round decimals in their display to two digits. Ferret will display all the digits recorded and aligned at the decimal point (Fig. 5a).

6.3 Use Overview and Details

A well-designed visual overview handles large datasets and helps analysts quickly spot suspicious patterns. At the same time, making the raw tabular data a first-class citizen within the visualization is essential. Only access to the raw data enables an analyst to confirm their suspicion or identify a benign explanation. In other words, any overview visualizations should be tightly integrated with a visualization of the details. If an interesting feature is noticed in an overview, it should be possible to query for details and easily see the rows generating that feature, as shown in Fig. 3d. Conversely, if an interesting pattern is found by inspecting the raw tabular data, it should be easy to switch to the overview visualization and observe that pattern from a higher vantage point (Fig. 5b and Fig. 8b).

6.4 Leverage Interactivity

While investigating artifacts of manipulation, interactive sorting and filtering is essential. **Sorting** by different columns provides many ways to view the data, and combining this ability with different visual encodings can reveal interesting patterns, such as alternating fonts Fig. 8b. **Filtering** is useful for focusing/excluding specific items. However, unlike most systems, ignoring only the values of specific cells (in contrast to filters that remove a row from a dataset) is more useful for detecting manipulations. For instance, in the case where values are clamped to an upper bound (**DS-Spider-E**), there may be many duplicates. Such duplicates will affect the analysis of several artifacts. Excluding those frequent values from the analysis is a convenient way of running the visualizations on the remaining data without excluding entire rows (Fig. 3d).

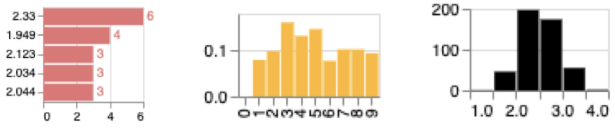
7 Visualizations in Ferret

Ferret is based on the design principles described to surface artifacts of manipulation. Ferret provides various visualizations for different aspects, yet some visual encodings can be used for multiple patterns. At the heart of Ferret is a tabular visualization technique [RC94, FGS*20] combining spreadsheet-like raw values with graphical marks, with a series of custom visual encodings, enriched by a set of supplementary views.

Formatting Ferret uses dedicated visual encodings for formatting artifacts within the tabular visualization and considers font styling and emphasis as well as the *data format*. Ferret does not use the styling of the source, since the exact formatting is usually immaterial for detecting manipulations. Instead, our encoding emphasizes the differences in formatting: The most frequent combination is assigned the default white background. All other unique combinations of formats are assigned a background color and texture/pattern (see Figure 8a). We chose to use five patterns and seven colors (35 combinations) because the number of unique combinations can exceed the number of reasonably distinguishable colors. When a cell is selected, the exact formatting parameters and a count for the number of cells that share the same formatting are listed.

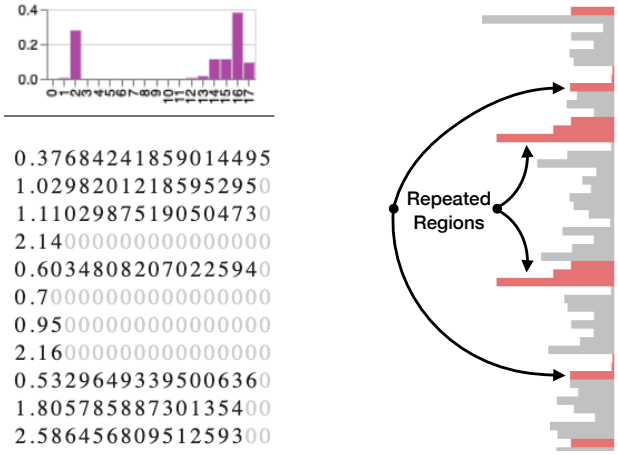
Summary Charts: Counts, Proportions, and Distributions We use histograms and bar charts to visualize distributions (how values in a column are distributed), counts (how often a number is duplicated), and proportions of values (what percentage of numbers has a precision of 2). To view the **counts of values**, we use horizontal bar graphs (Fig. 4a), which is useful for visualizing the count of duplicates and duplicate digits **AR-Duplicate**. These graphs can contain long labels, which is well suited for a horizontal layout. In Fig. 3c, the duplicate numbers of one of the spider datasets (**DS-Spider-E**) are shown at the top of the five numerical columns. The duplicate digits chart works analogously; instead of visualizing duplicated whole numbers, it shows duplicated sequences of digits (2- or 3-grams). To view the **proportion of values** with certain properties, Ferret shows vertical bar charts, where each bar shows a percentage of the property on the overall column (Fig. 4b). We use proportion bar charts to show the frequency of trailing and leading digits (**AR-Leading** and **AR-Trailing**), as well as the frequency of precisions (**AR-Precision**). Finally, we use a histogram to show the **distribution of values** (Fig. 4c), which is useful for sanity checks and alignment with domain expectations (**AR-Domain**).

Tabular Visualization These summary visualizations are tightly integrated with the tabular visualization. Using the summary charts, values can be selected or filtered. In Fig. 3 the number 600 has been filtered out, which removes it from the bar chart, and strikes it out in the tabular view. The value 228.1 has been selected, which highlights it in red.



(a) Count **(b)** Proportion **(c)** Distribution

Figure 4: Different summary visualizations available in Ferret.



(a) Precision **(b)** Structural

Figure 5: Visualizations for precision and structural artifacts. (a) A precision artifact (**AR-Precision**) is visible in **DS-Fly** through the proportion chart and the tabular visualization. (b) Repeated regions (**AR-Regions**) are visible for **DS-Gaming** using the overview.

Figure 5a shows another example of tight integration between the summary visualization on top and the tabular visualization below. The bar chart shows the proportions of different levels of precision, and the tabular visualization below shows data, highlighting the precision through alignment.

Structural Visualization For large tables, scrolling through the full dataset can be cumbersome, and raw numbers do not show structural effects well (**AR-Structural**). The table overview mode [FGS*20] in Ferret solves this problem by reducing the cell height to at least one pixel, maximizing the number of rows visible on the screen (see Figures 5b and 8b). In overview mode, exact values are elided, and graphical representations are shown.

Domain Visualizations Finally, Ferret includes a suite of domain visualizations to help reviewers test the data for deviations from their domain expectations. Ferret supports scatterplots (Fig. 7), faceted strip plots (Fig. 6), violin charts, bar charts, and parallel coordinate plots. Although these visualizations share similarities with general visualization exploration tools, we believe their inclusion is useful due to convenience and the guidance the tool provides.

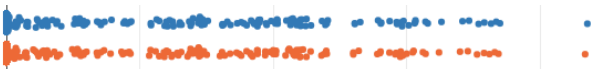


Figure 6: Strip-plot for miles driven in **DS-Driving** faceted by font. The data rendered appears to be duplicated with minor noise added.

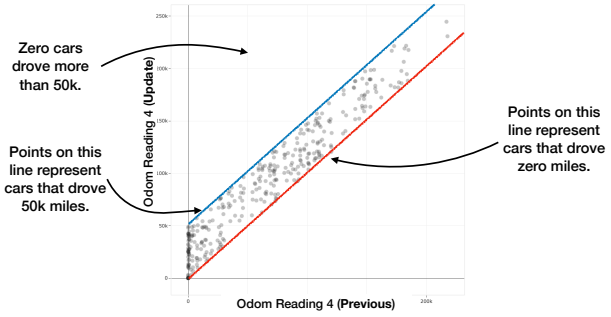


Figure 7: Scatterplot used to analyze a deviation from the domain expectation (**AR-Domain**). The x-axis corresponds to odometer readings taken at the beginning of the study, and the y-axis corresponds to readings taken at a later time, as recorded in **DS-Driving**. Note that the distribution of miles driven seems uniform, up to a hard cut-off after $50,000$ miles, indicated by the blue line.

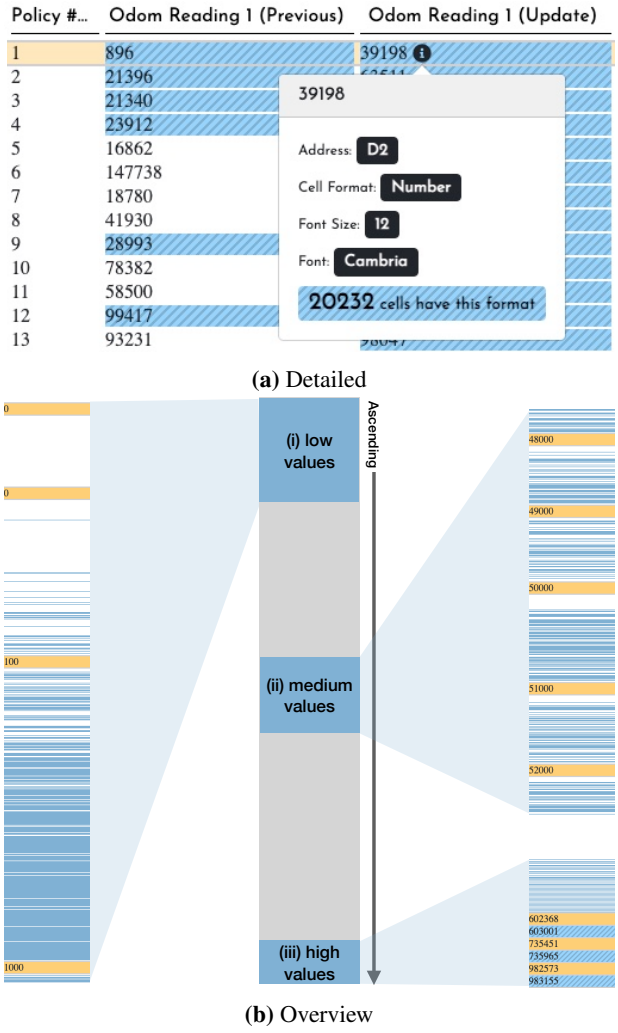
8 Implementation

Ferret is open source and implemented as a front-end web application. The code is available at `https://github.com/visdesignlab/ferret`, and a demo of the tool is available at `https://ferret.sci.utah.edu/`. The summary charts are built with Vega-Lite [SMWH17]. The table is built on top of LineUp [GLG*13] and Taggle [FGS*20] and uses custom code for cell rendering. The general visualizations are implemented with React and Plot.ly [Inc15]. Ferret uses excel.js [ed22] to load and process Excel files, which can be uploaded by users.

9 Case Study

In this section, we demonstrate the utility of the classification of artifacts, our design guidelines, and the Ferret prototype. We include case studies for all ten datasets in the supplementary material.

As our primary case study, we analyze the driving dataset (**DS-Driving**) and recreate and expand upon the analysis in a blog post that led to the paper's retraction [Edi21]. The post discusses four anomalies and provides two hypotheses to explain them. A similar analysis with Ferret unearths the same and some additional anomalies, sometimes using different kinds of visualizations, that support the claims of the blog post. Upon loading a dataset, Ferret displays the Formatting visualization (**AR-Formatting**), also showing instructions on when to use it and how to read it. For the driving dataset (**DS-Driving**), suspicious formatting in the second (*Odom Reading 1 (Previous)*) and the third (*Odom Reading 1 (Update)*) columns is immediately obvious. Mixed formatting within a single column, as seen in Fig. 8a, is unusual and suspicious. The second column has a mixture of Calibri and Cambria fonts. If this formatting was the only artifact found in a dataset, an editor could ask the authors for an explanation. However, further investigation reveals additional irregularities. Switching



(a) Detailed

(b) Overview

Figure 8: Visualizing formatting artifacts with color and patterns for **DS-Driving**. (a) The most frequent type of formatting is not highlighted (white background). All other formatting combinations are assigned a unique pattern/color combination. A tool-tip shows the formatting details on demand. (b) Showing structural patterns related to formatting. The pull-outs i-iii are taken from a large column, illustrated schematically in the center. Low values (i) are formatted in Cambria (white), and (ii) medium values alternate between Calibri and Cambria (blue), with Cambria clusters of round numbers. High values (iii) alternate between the fonts.

to the overview mode allows a faster review of the table's $13,488$ rows. The pattern of seemingly random mixes of fonts continues throughout the column. Yet, sorting the data reveals several patterns, shown in Fig 8b.

First, most values less than 100 in this column are in Calibri font, and all rows with a value of zero are in Calibri . Conversely, values between 100 and 1000 are predominately Cambria . For the remaining data, the two fonts are interspersed, except for certain regions, where

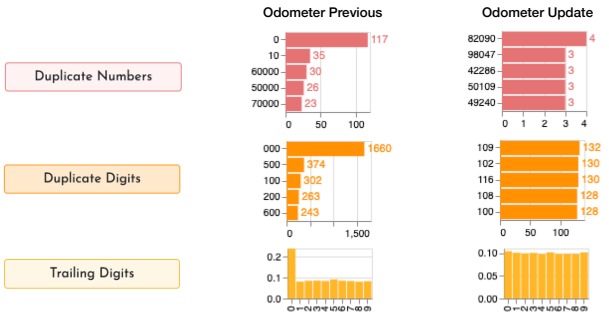


Figure 9: Rounding effects are clearly present in the duplicate numbers, duplicate digits, and leading digit frequency charts for the initial odometer reading (left column) but suspiciously absent for the follow-up reading (right column) in **DS-Driving**.

Calibri dominates. Inspecting the values reveals they are duplicate round numbers, such as $75,000$. Since these values represent self-reported car mileage, the data makes sense if people estimate the mileage of their car. Suspiciously, these rounding effects are not visible for the values in Cambria font, suggesting that the data collection method for the two fonts diverges.

Finally, the high values (Fig. 8b) alternate perfectly between Calibri and Cambria . Closer inspection reveals that every value styled in Calibri font has a corresponding Cambria value that is within 1000 miles. This pattern suggests data was copied and a random number between 1 and 1000 was added. Visualizing this column (Fig. 6), reveals that the two datasets are extremely similar.

A different approach to analyzing this dataset is to look at rounding effects. Fig 9 reveals tell-tale signs of rounding using the duplicate numbers, duplicate digits, and trailing digit frequency charts for the first column (the initial odometer reading). However, the second column (the follow-up reading after some time has passed) does not show any rounding effects.

To further explore this difference, we visualize the relationship between these two variables with a scatterplot. Fig. 7 shows that the miles driven never exceed $50,000$, and the distribution of miles driven is uniform between zero and $50,000$, an unlikely distribution for this dataset, supporting a hypothesis made in the blog that the odometer readings in the updated column were generated by adding a random number between 0 and $50,000$.

10 Discussion and Broader Impacts

With the goal of increasing the trustworthiness of scientific research, our work collects manipulated datasets, categorizes artifacts of manipulation, designs visualization methods to explore them, and prototypes a tool to make those artifacts salient. We observe that manipulated datasets tend to present multiple artifacts simultaneously, which can be spotted with different techniques offered by Ferret. Hence, we believe that our approach of providing multiple visualizations that are easy to step through and interpret is a robust method for spotting artifacts and minimizing risks. That said, several potential unintended consequences from our work could affect researchers and society in general.

False Positives. One concern is the possibility that our methods indicate that data has been manipulated when, in reality, it has not. To address this concern, we suggest that when artifacts are identified, they should be used as a means of discussion with the authors, not as indisputable evidence of wrongdoing. This concern is also one of the reasons we believe statistical tests or summary reports would be treacherous, as they might reduce the nuance and complexity of the topic to simplistic answers. Still, a tool such as Ferret can make it easier to levy accusations against authors. An overly zealous individual could cause harm if they place too much confidence in individual artifacts of manipulation and do not give authors opportunities to respond. In the worst case, bad actors could use a tool like Ferret to maliciously target individuals. To remedy this problem, we suggest that Ferret should predominantly be deployed for general checks as part of the review process or when there are reasons to suspect wrongdoing with a paper.

Shaming. Our hope in collecting references to the manipulated datasets in this paper is that it will be a resource for others interested in investigating data manipulation. However, our work may lead to additional unwanted attention for the authors of these datasets. To minimize the potential impact of our actions, we have published only datasets that come with an official retraction or an expression of concern from the publishing journal.

Security Theater. Reviewers and editors are often volunteers; hence, limiting their workload is an important consideration, especially if the additional work would be akin to useless "security theater". Similar concerns can be raised about plagiarism checkers, yet they have detected numerous cases of plagiarism. We also attempt to make Ferret easy to use to avoid unnecessary burdens. However, conducting a cost-benefit analysis in a trial run with a selected journal is a logical next step.

Abuse. Knowledge about artifacts of manipulation and the existence of tools to identify them may help bad actors avoid detection of their misconduct. Experience with plagiarism detection tools shows that they continue to catch manipulation. Although we cannot ensure that abuse will not happen, we hope that the burden of "engineering" a dataset that does not raise suspicion is so high that bad actors may conclude that manipulation is not worth the risk.

Data Sharing. Using tools like Ferret may disincentivize authors to submit data with their manuscripts for fear of being unjustly accused of manipulation. Many journals and conferences already require the publication of data. Some researchers may choose to publish with journals that do not. We hope that the scientific community can meet

this challenge by (a) carefully using tools like Ferret and (b) more broadly endorsing open science practices.

11 Conclusion

We believe our work will help future reviewers "ferret out" manipulations in tabular datasets. Knowing what artifacts of manipulation to look for will help analysts focus their search. Our design guidelines will aid in the development of tools for performing data forensics. Finally, Ferret is a first step toward instantiating this knowledge in a tool. Due to the adversarial nature of catching data manipulation, designing a single static tool is likely impossible. However, we believe our approach, which emphasizes the importance of the human-in-the-loop, is robust to changes in future manipulation techniques.

12 Acknowledgments

The authors wish to thank Holger Stitz, Michael Pühringer, and the LineUp authors for their support using the library; the Retraction Watch Project for access to their database; the reviewers for their feedback; Zach Cutler and Jack Wilburn for technical help; Dr. Ken A. Thompson and another anonymous interview participant for their time and expertise; and the Visualization Design Lab for feedback. This work was supported by NSF IIS 1751238 and CCF-2115677.

References

[Aut20] AUTHOR A.: PubPeer discussion of "Host-parasitoid evolution in a metacommunity", Aug. 2020.

[BBS*18] BLUMENSCHEIN M., BEHRISCH M., SCHMID S., BUTSCHER S., WAHL D. R., VILLINGER K., RENNER B., REITERER H., KEIM D. A.: SMARTexplore: Simplifying High-Dimensional Data Analysis through a Table-Based Visual Analytics Approach. In *Proceedings of IEEE Conference on Visual Analytics Science and Technology (VAST '18)* (2018), IEEE, pp. 36–47. doi:10.1109/VAST.2018.8802486.

[BBZ18] BAROWY D. W., BERGER E. D., ZORN B.: ExceLint: Automatically Finding Spreadsheet Formula Errors. *Proceedings of the ACM on Programming Languages 2*, OOPSLA (Oct. 2018), 1–26. arXiv:1901.11100, doi:10.1145/3276518.

[BCM*21] BONETTINI N., CANNAS E. D., MANDELLI S., BONDI L., BESTAGINI P., TUBARO S.: Video Face Manipulation Detection Through Ensemble of CNNs. In *2020 25th International Conference on Pattern Recognition (ICPR)* (Jan. 2021), IEEE Computer Society, pp. 5012–5019. doi:10.1109/ICPR48806.2021.9412711.

[Bik22] BIK E.: Science Has a Nasty Photoshopping Problem. *The New York Times* (Oct. 2022).

[BKA*07] BELLON S., KOSCHKE R., ANTONIOL G., KRINKE J., MERLO E.: Comparison and Evaluation of Clone Detection Tools. *IEEE Transactions on Software Engineering 33*, 9 (Sept. 2007), 577–591. doi:10.1109/TSE.2007.70725.

[BNTZ20] BARNI M., NOWROOZI E., TONDI B., ZHANG B.: Effectiveness of Random Deep Feature Selection for Securing Image Manipulation Detectors Against Adversarial Examples. In *ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (May 2020), pp. 2977–2981. doi:10.1109/ICASSP40776.2020.9053318.

[Boc16] BOCK A. A.: A Literature Review of Spreadsheet Technology. *IT University of Copenhagen Technical Report Series* (2016), 35.

[Bro20] BROWN N.: Some issues in a recent gaming research article: Etindele Sosso et al. (2020), Apr. 2020.

[Bro21] BROWN N.: Some problems in the dataset of a large study of Ivermectin for the treatment of Covid-19, July 2021.

[CCLC16] CHEUNG S.-C., CHEN W., LIU Y., CHANGXU: CUSTODES: Automatic Spreadsheet Cell Clustering and Smell Detection Using Strong and Weak Features. In *2016 IEEE/ACM 38th International Conference on Software Engineering (ICSE)* (May 2016). doi:10.1145/2884781.2884796.

[Cha17] CHAWLA D.: Controversial software is proving surprisingly accurate at spotting errors in psychology papers. https://www.science.org/content/article/controversial-software-proving-surprisingly-accurate-spotting-errors-psychology-papers, Nov. 2017.

[Cha21] CHARLTON A.: RETRACTED ARTICLE: Why money meanings matter in decisions to donate time and money. https://openmkt.org/blog/2021/retracted-article-why-money-meanings-matter-in-decisions-to-donate-time-and-money/, July 2021.

[DCH*23] DONG C., CHEN X., HU R., CAO J., LI X.: MVSS-Net: Multi-View Multi-Scale Supervised Networks for Image Manipulation Detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence 45*,

3 (Mar. 2023), 3539–3553. doi:10.1109/TPAMI.2022.3180556.

[DHP04] DURTSCHI C., HILLISON W., PACINI C.: The effective use of benford's law to assist in detecting fraud in accounting data. *Journal of forensic accounting 5*, 1 (2004), 17–34.

[EB21] EDITORIAL BOARD P. B.: Editor's Note on: Host–parasitoid evolution in a meta-community. *Proceedings of the Royal Society B: Biological Sciences 288*, 1947 (Mar. 2021), 20210505. doi:10.1098/rspb.2021.0505.

[ed22] EXCEL.JS DEVELOPERS T.: ExcelJS. exceljs, Dec. 2022.

[Edi16] EDITORS J.: Retraction Note to: Why money meanings matter in decisions to donate time and money. *Marketing Letters 27*, 2 (June 2016), 409–409. doi:10.1007/s11002-016-9401-6.

[Edi21] EDITORS J.: Retraction for Shu et al., Signing at the beginning makes ethics salient and decreases dishonest self-reports in comparison to signing at the end. *Proceedings of the National Academy of Sciences 118*, 38 (Sept. 2021), e2115397118. doi:10.1073/pnas.2115397118.

[Ens21] ENSERINK M.: Sea of doubts. *Science 372*, 6542 (May 2021), 560–565. doi:10.1126/science.372.6542.560.

[FGS*20] FURMANOVA K., GRATZL S., STITZ H., ZICHNER T., JARESOVA M., LEX A., STREIT M.: Taggle: Combining overview and details in tabular data visualizations. *Information Visualization 19*, 2 (2020), 114–136. doi:10.1177/1473871619878085.

[FMG19] FOLTÝNEK T., MEUSCHKE N., GIPP B.: Academic Plagiarism Detection: A Systematic Literature Review. *ACM Computing Surveys 52*, 6 (Oct. 2019), 112:1–112:42. doi:10.1145/3345317.

[FSK08] FIGURSKA M., STAŃCZYK M., KULESZA K.: Humans cannot consciously generate random numbers sequences: Polemic study. *Medical Hypotheses 70*, 1 (Jan. 2008), 182–185. doi:10.1016/j.mehy.2007.06.038.

[GGL*14] GRATZL S., GEHLENBORG N., LEX A., PFISTER H., STREIT M.: Domino: Extracting, Comparing, and Manipulating Subsets Across Multiple Tabular Datasets. *IEEE Transactions on Visualization and Computer Graphics (InfoVis'14) 20*, 12 (2014), 2023–2032. doi:10.1109/TVCG.2014.2346260.

[GLG*13] GRATZL S., LEX A., GEHLENBORG N., PFISTER H., STREIT M.: LineUp: Visual Analysis of Multi-Attribute Rankings. *IEEE Transactions on Visualization and Computer Graphics (InfoVis '13) 19*, 12 (2013), 2277–2286. doi:10.1109/TVCG.2013.173.

[Har18] HAROZ S.: Open Practices in Visualization Research : Opinion Paper. In *2018 IEEE Evaluation and Beyond - Methodological Approaches for Visualization (BELIV)* (Oct. 2018), pp. 46–52. doi:10.1109/BELIV.2018.8634427.

[HSPv13] HERMANS F., SEDEE B., PINZGER M., VAN DEURSEN A.: Data clone detection and visualization in spreadsheets. In *2013 35th International Conference on Software Engineering (ICSE)* (San Francisco, CA, USA, May 2013), IEEE, pp. 292–301. doi:10.1109/ICSE.2013.6606575.

[HXJ*20] HUANG Y., XU C., JIANG Y., WANG H., LI D.: WARDER: Towards effective spreadsheet defect detection by validity-based cell cluster refinements. *Journal of Systems and Software 167* (Sept. 2020), 110615. doi:10.1016/j.jss.2020.110615.

[Inc15] INC. P. T.: Collaborative data science. Plotly Technologies Inc., 2015.

[JSHW14] JANNACH D., SCHMITZ T., HOFER B., WOTAWA F.: Avoiding, finding and fixing spreadsheet errors – A survey of automated approaches for spreadsheet QA. *Journal of Systems and Software 94* (Aug. 2014), 129–150. doi:10.1016/j.jss.2014.03.058.

[KNY*22] KWON M.-J., NAM S.-H., YU I.-J., LEE H.-K., KIM C.: Learning JPEG Compression Artifacts for Image Manipulation Detection and Localization. *International Journal of Computer Vision 130*, 8 (Aug. 2022), 1875–1895. doi:10.1007/s11263-022-01617-5.

[KSJ*21] KOCH P., SCHEKOTIHIN K., JANNACH D., HOFER B., WOTAWA F.: Metric-Based Fault Prediction for Spreadsheets. *IEEE Transactions on Software Engineering 47*, 10 (Oct. 2021), 2195–2207. doi:10.1109/TSE.2019.2944604.

[Las20] LASKOWSKI K. L.: What to do when you don't trust your data anymore – Laskowski Lab at UC Davis, Jan. 2020.

[Law21] LAWRENCE J.: Why Was a Major Study on Ivermectin for COVID-19 Just Retracted?, July 2021.

[LH19] LI H., HUANG J.: Localization of Deep Inpainting Using High-Pass Fully Convolutional Network. In *2019 IEEE/CVF International Conference on Computer Vision (ICCV)* (Seoul, Korea (South), Oct. 2019),

IEEE, pp. 8300–8309. doi:10.1109/ICCV.2019.00839.

[LMD*20] LASKOWSKI K. L., MODLMEIER A. P., DEMARCO A. E., COLEMAN A., ZHAO K., BRITTINGHAM H. A., MCDERMOTT D. R., PRUITT J. N.: Retraction: Persistent social interactions beget more pronounced personalities in a desert-dwelling social spider. *Biology Letters 16*, 2 (Feb. 2020), 20200062. doi:10.1098/rsbl.2020.0062.

[LMP20] LASKOWSKI K. L., MONTIGLIO P.-O., PRUITT J. N.: Retraction: Individual and Group Performance Suffers from Social Niche Disruption. *The American Naturalist 195*, 2 (Feb. 2020), 393–393. doi:10.1086/708066.

[LP20] LASKOWSKI K. L., PRUITT J. N.: Retraction: Evidence of social niche construction: Persistent and repeated social interactions generate stronger personalities in a social spider. *Proceedings of the Royal Society B: Biological Sciences 287*, 1919 (Jan. 2020), 20200077. doi:10.1098/rspb.2020.0077.

[LWX*19a] LI D., WANG H., XU C., SHI F., MA X., LU J.: WARDER: Refining Cell Clustering for Effective Spreadsheet Defect Detection via Validity Properties. In *2019 IEEE 19th International Conference on Software Quality, Reliability and Security (QRS)* (July 2019), pp. 139–150. doi:10.1109/QRS.2019.00030.

[LWX*19b] LI D., WANG H., XU C., ZHANG R., CHEUNG S.-C., MA X.: SGUARD: A Feature-Based Clustering Tool for Effective Spreadsheet Defect Detection. In *2019 34th IEEE/ACM International Conference on Automated Software Engineering (ASE)* (Nov. 2019), pp. 1142–1145. doi:10.1109/ASE.2019.00122.

[Mil15] MILLER S. J.: *Benford's Law*. Princeton University Press, 2015.

[Nig12] NIGRINI M. J.: *Benford's Law: Applications for forensic accounting, auditing, and fraud detection*, vol. 586. John Wiley & Sons, 2012.

[NM09] NIGRINI M. J., MILLER S. J.: Data diagnostics using second-order tests of benford's law. *Auditing: A Journal of Practice & Theory 28*, 2 (2009), 305–324.

[NSH*18] NIEDERER C., STITZ H., HOURIEH R., GRASSINGER F., AIGNER W., STREIT M.: TACO: Visualizing Changes in Tables Over Time. *IEEE Transactions on Visualization and Computer Graphics 24*, 1 (Jan. 2018), 677–686. doi:10.1109/TVCG.2017.2745298.

[Ora10] ORANSKY A. I.: Why write a blog about retractions?, Aug. 2010.

[PBL08] POWELL S., BAKER K., LAWSON B.: A critical review of the literature on spreadsheet errors. *Decision Support Systems 46* (Dec. 2008), 128–138. doi:10.1016/j.dss.2008.06.001.

[Pil22] PILLER C.: Potential fabrication in research images threatens key theory of Alzheimer's disease. *Science 377*, 6604 (2022), 358–363.

[PRA*16] PASHLER H., ROHRER D., ABRAMSON I., WOLFSON T., HARRIS C. R.: A Social Priming Data Set With Troubling Oddities. *Basic and Applied Social Psychology 38*, 1 (Jan. 2016), 3–18. doi:10.1080/01973533.2015.1124767.

[PSL21] PARK T., SONG H., LEE S. J.: Detecting and Recovering Integer Data Manipulated by Multiplication With a Nonintegral Real Number and a Rounding Operation. *IEEE Access 9* (2021), 57149–57164. doi:10.1109/ACCESS.2021.3071794.

[RC94] RAO R., CARD S. K.: The table lens: Merging graphical and symbolic representations in an interactive focus + context visualization for tabular information. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (New York, NY, USA, Apr. 1994), CHI '94, Association for Computing Machinery, pp. 318–322. doi:10.1145/191666.191776.

[RNE16] RIFE SC., NUIJTEN MB., EPSKAMP S.: Statcheck: Extract statistics from articles and recompute p-values [web application], 2016.

[Shn96] SHNEIDERMAN B.: The Eyes Have It: A Task by Data Type Taxonomy for Information Visualizations. In *Proceedings of the IEEE Symposium on Visual Languages (VL '96)* (1996), pp. 336–343. doi:10.1109/VL.1996.545307.

[SKV*20] SOSSO F. A. E., KUSS D. J., VANDELANOTTE C., JASSO-MEDRANO J. L., HUSAIN M. E., CURCIO G., PAPADOPOULOS D., ASEEM A., BHATI P., LOPEZ-ROSALES F., BECERRA J. R., D'AURIZIO G., MANSOURI H., KHOURY T., CAMPBELL M., TOTH A. J.: Retraction Note: Insomnia, sleepiness, anxiety and depression among different types of gamers in African countries. *Scientific Reports 10*, 1 (June 2020), 9256. doi:10.1038/s41598-020-66798-w.

[SMWH17] SATYANARAYAN A., MORITZ D., WONGSUPHASAWAT K., HEER J.: Vega-Lite: A

Grammar of Interactive Graphics. *IEEE Transactions on Visualization and Computer Graphics 23*, 1 (Jan. 2017), 341–350. doi:10.1109/TVCG.2016.2599030.

[SNR17] SUKHOBOK D., NIKOLOV N., ROMAN D.: Tabular Data Anomaly Patterns. In *2017 International Conference on Big Data Innovations and Applications (Innovate-Data)* (Aug. 2017), pp. 25–34. doi:10.1109/Innovate-Data.2017.10.

[SSBW12] SCHULZ M.-A., SCHMALBACH B., BRUGGER P., WITT K.: Analysing Humanly Generated Random Number Sequences: A Pattern-Based Approach. *PLOS ONE 7*, 7 (July 2012), e41531. doi:10.1371/journal.pone.0041531.

[SSN21] SIMONSOHN U., SIMMONS J., NELSON L.: [98] Evidence of Fraud in an Influential Field Experiment About Dishonesty. https://datacolada.org/98, Aug. 2021.

[SWA03] SCHLEIMER S., WILKERSON D. S., AIKEN A.: Winnowing: Local algorithms for document fingerprinting. In *Proceedings of the 2003 ACM SIGMOD International Conference on Management of Data* (New York, NY, USA, June 2003), SIGMOD '03, Association for Computing Machinery, pp. 76–85. doi:10.1145/872757.872770.

[The18] THE CENTER FOR SCIENTIFIC INTEGRITY: The Retraction Watch Database, 2018.

[Tho22] THORP H. H.: Editorial Retraction. *Science 377*, 6608 (Aug. 2022), 826–826. doi:10.1126/science.ade2691.

[TLB14] TOWSE J. N., LOETSCHER T., BRUGGER P.: Not all numbers are equal: Preferences and biases among children and adults when generating random sequences. *Frontiers in Psychology 5* (Jan. 2014), 19. doi:10.3389/fpsyg.2014.00019.

[Vig20] VIGLIONE G.: 'Avalanche' of spider-paper retractions shakes behavioural-ecology community. *Nature 578*, 7794 (Feb. 2020), 199–200. doi:10.1038/d41586-020-00287-y.

[Wan19] WANG S.: Retraction: Glioma Gene Therapy Using Induced Pluripotent Stem Cell Derived Neural Stem Cells. *Molecular Pharmaceutics 16*, 9 (Sept. 2019), 4088. doi:10.1021/acs.molpharmaceut.9b00837.

[WWZ*19] WANG S.-Y., WANG O., ZHANG R., OWENS A., EFROS A.: Detecting Photoshopped Faces by Scripting Photoshop. In *2019 IEEE/CVF International Conference on Computer Vision (ICCV)* (Oct. 2019), IEEE Computer Society, pp. 10071–10080. doi:10.1109/ICCV.2019.01017.

[YLL*20] YANG C., LI H., LIN F., JIANG B., ZHAO H.: Constrained R-Cnn: A General Image Manipulation Detection Model. In *2020 IEEE International Conference on Multimedia and Expo (ICME)* (July 2020), IEEE Computer Society, pp. 1–6. doi:10.1109/ICME46284.2020.9102825.

[ZDZ*20] ZHANG Y., DOU W., ZHU J., XU L., ZHOU Z., WEI J., YE D., YANG B.: Learning to detect table clones in spreadsheets. In *Proceedings of the 29th ACM SIGSOFT International Symposium on Software Testing and Analysis* (New York, NY, USA, July 2020), ISSTA 2020, Association for Computing Machinery, pp. 528–540. doi:10.1145/3395363.3397384.