SHEVA: A Visual Analytics System for Statistical Hypothesis Exploration

Vicente Nejar de Almeida
Instituto de Informática, UFRGS
Porto Alegre, Brazil
vicente.almeida@inf.ufrgs.br

Eduardo Ribeiro
Universidade Federal do Tocantins
Palmas, Brazil
uft.eduardo@mail.uft.edu.br

Nassim Bouarour
Univ. Grenoble Alpes
Grenoble, France
nassim.bouarour@univ-grenoble-alpes.fr

João Luiz Dihl Comba
Instituto de Informática, UFRGS
Porto Alegre, Brazil
comba@inf.ufrgs.br

Sihem Amer-Yahia
CNRS, Univ. Grenoble Alpes
Grenoble, France
sihem.amer-yahia@univ-grenoble-alpes.fr

ABSTRACT

We demonstrate SHEVA, a System for Hypothesis Exploration with Visual Analytics. SHEVA adopts an Exploratory Data Analysis (EDA) approach to discovering statistically-sound insights from large datasets. The system addresses three longstanding challenges in Multiple Hypothesis Testing: (i) the likelihood of rejecting the null hypothesis by chance, (ii) the pitfall of not being representative of the input data, and (iii) the ability to navigate among many data regions while preserving the user's train of thought. To address (i) & (ii), SHEVA implements significance adjustment methods that account for data-informed properties such as coverage and novelty. To address (iii), SHEVA proposes to guide users by recommending one-sample and two-sample hypotheses in a stepwise fashion following a data hierarchy. Users may choose from a collection of pre-trained hypothesis exploration policies and let SHEVA guide them through the most significant hypotheses in the data, or intervene to override suggested hypotheses. Furthermore, SHEVA relies on data-to-visual element mappings to convey hypothesis testing results in an interpretable fashion, and allows hypothesis pipelines to be stored and retrieved later to be tested on new datasets.

PVLDB Reference Format:
Vicente Nejar de Almeida, Eduardo Ribeiro, Nassim Bouarour, João Luiz Dihl Comba, and Sihem Amer-Yahia. SHEVA: A Visual Analytics System for Statistical Hypothesis Exploration. PVLDB, 16(12): 4102 - 4105, 2023.
doi:10.14778/3611540.3611631

PVLDB Artifact Availability:
The source code, data, and/or other artifacts have been made available at https://github.com/Vicente-Nejar-de-Almeida/sheva.

This work is licensed under the Creative Commons BY-NC-ND 4.0 International License. Visit https://creativecommons.org/licenses/by-nc-nd/4.0/ to view a copy of this license. For any use beyond those covered by this license, obtain permission by emailing info@vldb.org. Copyright is held by the owner/author(s). Publication rights licensed to the VLDB Endowment.
Proceedings of the VLDB Endowment, Vol. 16, No. 12 ISSN 2150-8097.
doi:10.14778/3611540.3611631

1 INTRODUCTION

The ability to make sound discoveries has become a central concern in extracting insights from large datasets [2, 10, 15]. Multiple Hypothesis Testing is a method of choice as it enables simultaneous test hypotheses on several data regions. However, it raises several new challenges: as the number of data regions increases, (i) the likelihood of rejecting the null hypothesis by chance increases, as well as (ii) the likelihood of missing some data regions and returning non-representative results and (iii) breaking the train of thought of the user due to navigating in a large number of data regions. To address these challenges, we built SHEVA, a System for Hypothesis Exploration with Visual Analytics. SHEVA adopts an Exploratory Data Analysis (EDA) approach to discovering statistically-sound insights from large datasets in a stepwise fashion.

A hypothesis is verified when *data regions* identified by some *filters*, have statistically *similar, higher, or lower aggregated values* (for instance, mean, variance, or distribution on data regions). Examples of hypotheses on the MovieLens dataset [7] are "Male reviewers' mean rating for Action movies is greater than 3.5" (one-sample hypothesis) and "The average rating of females in Nebraska for long movies is higher than that of females in Hawaii" (two-sample hypothesis). This is referred to as the alternative hypothesis that states the desired test and complements the null hypothesis. Therefore, the null hypothesis is said to be rejected by selected regions.

As the number of hypotheses increases, the likelihood of making a false discovery also increases [9]. We address that using p-value corrections [5, 12] such as the Bonferroni Family-Wise Error Rate (FWER) and the Benjamini-Yekutieli False Discovery Rate (FDR) [1].

In analyzing data for actionable insights, it's preferable to have a limited number of data regions that are both valid concerning a statistical test and representative of the data. Still, the sole focus on hypothesis significance may lead to exploring a small portion of the data of interest. For instance, in the case where the input is male users, only 25-30 year-old engineers who live in California may be returned. To remedy that, we build on the approach proposed in [2] which combines data-informed properties such as coverage with significance correction methods. We extend that approach to include data-informed properties such as novelty of data regions.

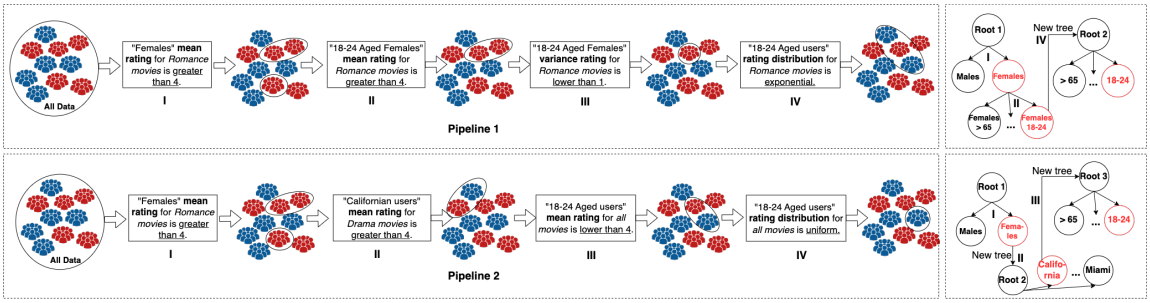

Figure 1: Two hypothesis exploration pipelines. The first pipeline preserves the user's train of thought by enforcing meaningful transitions between hypotheses. The second pipeline breaks the train of thought by "jumping" between unrelated regions.

As the number of hypotheses to be tested increases, preserving the user's train of thought becomes challenging. For instance, a user testing a hypothesis on the average rating of females for Drama movies would benefit from an iterative testing process where transitions between different hypotheses are semantically related. We illustrate this in Figure 1. To address this, we guide users by recommending hypotheses in a stepwise fashion following a hierarchy.

In SHEVA the user's train of thought follows a forest induced by splitting data regions into different attributes. This is accomplished by using Deep Reinforcement Learning to train an agent to generate a forest of decision trees [13, 14] where the notion of gain includes adjusted statistical significance and data-informed dimensions such as coverage and novelty. Furthermore, users may also intervene to override suggested hypotheses. SHEVA relies on appropriate data-to-visual element mappings to convey hypothesis testing results in an interpretable fashion. Finally, generated hypothesis pipelines may be stored and retrieved later to be tested on new datasets. We demonstrate SHEVA with two datasets: a composition of MovieLens 20M [6] and IMDB [8] datasets , and the MatMat dataset [11].

2 THE SHEVA SYSTEM

2.1 Framework

We define a data region d as a conjunction of predicates of the form (attribute = value), referred to as its label. We denote $\mathcal{D}$ the entire data region space generated from the input dataset O. All elements belonging to the same data region satisfy its label. Data regions can be organized into a forest of trees hierarchically induced by their labels (see example depicted in Figure 1). We define $children(d, a)$ - the children of a region d - as all regions defined by $d.label \wedge p$, where p is a predicate expressed on an attribute a.

Table 1 illustrates the hypotheses SHEVA supports using examples of movie ratings along with the type of test that is relevant for each request. For example, the request for H_2 encapsulates both null and alternative hypotheses as well as dimensions highlighted in Table 1. H_1-H_3 use mean to aggregate data region ratings and require different tests. H_1 shows the case of a one-sample t-test. Input data is all movie ratings by students. Data regions such as "Students in California" or "Young students" are generated, and their average rating is compared to a reference value (here, 3.5) with a one-sample test. Data regions that reject the null hypothesis

Table 1: Examples of hypotheses in SHEVA with data regions, aggregate, dimension, and statistical test operator

H_1	Student data regions whose rating mean is greater than 3.5	One-sample t-test
H_2	Female data regions whose rating mean is lower than Male data regions within the same period	Two-sample Welch's test
H_3	Male data regions whose rating mean changes between 2 Seasons	Two-sample paired t-test
H_4	Data regions whose rating variance for Comedy movies is greater than 1	One-sample variance Chi-square test
H_5	Data region pairs whose rating variance for 70's movies differs in the Spring	Two-sample variance F-test
H_6	Data regions whose yearly rating distribution does not follow a Gaussian distribution	One-sample Kolmogorov-Smirnov test
H_7	Data region pairs whose rating distribution for Drama movies differs in the same season	Two-sample Kolmogorov-Smirnov test

"Students whose rating average is equal to 3.5" and satisfy the alternative hypothesis "Students whose rating average is greater than 3.5" are returned. The case of a two-sample test is shown in H_2 and H_3. We use variance as an aggregation in H_4 and H_5 and rely on one-sample Chi-square test and two-sample F-test respectively. It starts with all rating records for movies in the 70's and returns pairs of data regions whose rating variance for those movies differs in the Spring. The last two types, H_6 and H_7, compare rating distributions using one-sample and two-sample Kolmogorov-Smirnov tests.

2.2 Data-Informed Hypothesis Satisfaction

A hypothesis test considers two hypotheses that contain opposing viewpoints. The null hypothesis H_0 usually states that data region aggregates are identical. The alternative hypothesis H_a states a claim that contradicts H_0 and corresponds to desired samples (in our case, user data regions). The decision can either be "reject H_0" if the sample favors the alternative hypothesis or "do not reject H_0" if the sample is insufficient to reject the null hypothesis.

The standard protocol to compute p-values of each candidate data region first verifies the normality and independence of each sample [3]. When comparing two means with a two-sample t-test, a value of 0.05 for α indicates a 5% risk of concluding that a difference

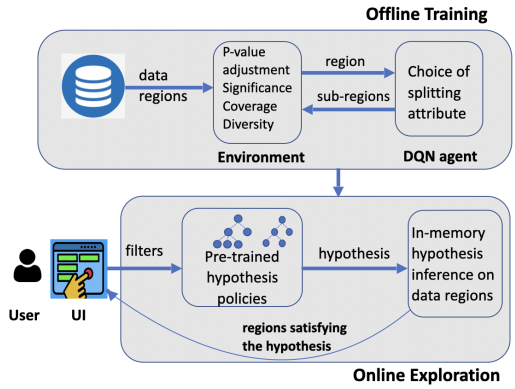

Figure 2: SHEVA architecture. Hypothesis policies are trained offline with rewards that combine statistical significance with data-informed dimensions (coverage and novelty).

exists between the two aggregated values when there is none [4]. A data region d is said to satisfy a hypothesis h with respect to a significance p-value correction methods such as FWER and FDR if and only if $\exists\, a \,|children(d, a)$ that satisfy the hypothesis h.

The SHEVA framework is designed to support various types of statistical tests. Each test depends on the size of the data region, its members (paired or unpaired), and the aggregation function AGG.

2.3 Guided Hypothesis Testing

In SHEVA, we use hypothesis testing to guide users in discovering meaningful insights from a large dataset. In this work, we propose to do that iteratively following Exploratory Data Analysis (EDA) [14]. We model the hypothesis exploration process as a Markov Decision Process and aim to learn a policy that seeks to identify the best (data region, hypothesis) pair to test at each step. For this purpose, we define a reward function that combines hypothesis significance with the data-informed dimensions of coverage and novelty.

The state space observed by the reinforcement learning agent is represented by embedding vectors of all data regions. There are two types of actions: (i) exploiting one of the data regions in the current state (diving into the depth of the same tree) or exploring a new data region (starting a new tree), (ii) choosing a new hypothesis to apply next. Applying the selected hypothesis to the exploited/explored data region results in a new set of data regions in the next state.

The reward $\mathcal{R}: \mathcal{S} \times \mathcal{A} \mapsto \mathbb{R}$ is the instantaneous reward of taking an action a_t at state s_t. More formally, $r(s_t, a_t) = w_1 \cdot sig(s_{t+1}) + w_2 \cdot cov(s_{t+1}) + w_3 \cdot nov(s_{t+1})$ where w_1, w_2, and w_3 are non-negative, $w_1 + w_2 + w_3 = 1$, $sig(.)$ is the significance of the new set of data regions, $cov(.)$ its coverage of the input region, and $nov(.)$ its novelty with respect to previously seen data regions. SHEVA online exploration uses pre-trained hypothesis policies that differ according to the weights w_1, w_2, and w_3, as illustrated in Figure 2.

3 SHEVA USER INTERFACE

The SHEVA user interface is responsible for supporting hypothesis exploration in large datasets. It comprises three main sections: **control parameters**, **current pipeline**, and **current operator results** (Figure 3). Additionally, the **current pipeline** section is divided in two subsections: **table view**, and **tree view**. The **control parameters** are positioned on top of the interface and are composed of selection dropdowns and control buttons. The **dataset selection** dropdown allows users to select the dataset (such as MovieLens). The **hypothesis policy selection** dropdown allows users to select the available pre-trained hypothesis policies (*e.g.,* power-only, coverage-only, etc.) for the currently selected dataset. Finally, three **exploration control buttons (start/stop, next, and reset)** allow the user to engage the exploration.

The current results of the exploration process are returned in the **current pipeline** and **current operator results** display areas. The hypothesis pipeline results are displayed in the left panel of the interface, composed of two complementary visualizations: **table view**, and **tree view**. The **table view** displays the pipeline of hypotheses formulated, stacked vertically, with the most recent on top. Each hypothesis is inside a **hypothesis box**, which contains the hypothesis formulated in natural language, the number of data regions generated, and statistical values (FDR, power, coverage, etc). The hypothesis box also displays an icon identifying whether the current hypothesis is derived from an exploration (side-by-side circles) or exploitation (two concentric circles) operation.

In the **tree view**, hypotheses are shown in a collapsible tree-based visualization. Nodes represent data regions, and edges represent hypotheses. Each node describes a data region in a user-friendly manner, such as "Movies released in the '60s (4726 users, 185 movies)", while edges represent a short mathematical formulation of a statistical hypothesis (for example, $\mu > 2.5$). Furthermore, when the mouse hovers over a node, a tooltip is shown describing the statistical values of the hypothesis applied to that data region.

The pipeline of hypotheses displays a sequence of hypotheses formulated by SHEVA and their statistical values, which are essential for the user to evaluate the quality of the hypotheses. Associated with the current hypothesis, SHEVA displays on the right panel of the interface the **current operator**, which shows the input data region and the data regions returned by the operator (which are the resulting regions from applying a hypothesis to the input data region). Data regions are summarized using a **bar plot** for each attribute (*e.g.,* ratings, genders, genres, occupations, ages, etc.).

The exploration begins when the user selects the dataset of choice and the pre-trained hypothesis policy (*e.g.,* power-only). To initiate the hypothesis generation, the user may press the **start button** or the **next button**. The **start button** signals SHEVA to continuously generate hypotheses and stack them in the **current pipeline** results area. After each hypothesis is generated, the system pauses briefly to allow the user to analyze the hypothesis generated, and then proceeds automatically to generate the following hypothesis. When this exploration process starts, a **stop button** replaces the **start button**. It allows the user to stop the exploration. The current hypothesis and charts describing the data region's different attributes can be further explored at this point.

The user may also engage the exploration by pressing the **next button**. In this case, SHEVA generates one hypothesis and then stops. This allows for an increased control of the exploration process. Furthermore, the **next button** contains a dropdown, which enables the user to suggest the next statistical hypothesis to be used.

In the **current pipeline** section, the user may select a previous hypothesis. In the **current operator results** section, any of the

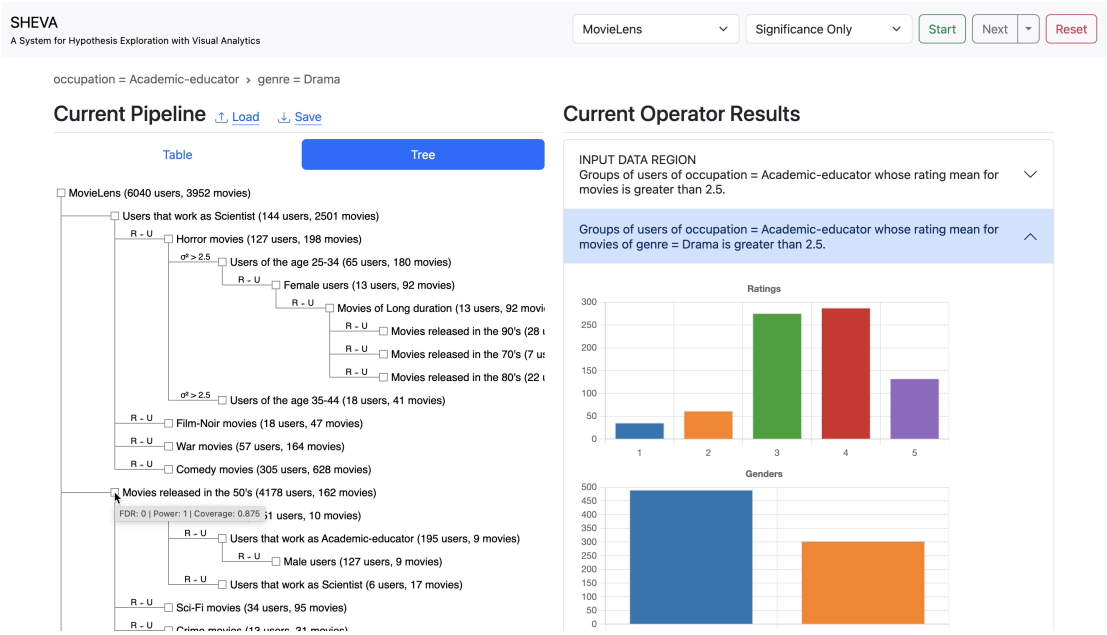

Figure 3: SHEVA user interface has three sections: control parameters, current hypothesis pipeline and current operator results.

available data regions may be selected. By selecting a data region within a specific hypothesis, the user informs SHEVA to generate a statistical hypothesis within that data region (exploitation operation). If no sufficiently good hypothesis with regards to statistical significance and data-informed properties mentioned previously, SHEVA simply performs an exploration operation.

4 DEMONSTRATION SCENARIO

In our demo, we showcase how SHEVA can be used as a tool to guide data exploration. To start, we first select a dataset of interest and an exploration policy (for instance, Significance Only).

To initiate the exploration, we click on start, and wait while the underlying reinforcement learning agent generates several statistically significant hypotheses. We then click on stop, to carefully analyze the generated set of hypotheses. Since the agent explores many regions of the dataset within a reasonably small number of iterations, this set will contain a diverse group of statistically-sound insights, automatically generated by the system.

The default view of the current pipeline of hypotheses is the table view. In this view, the user can select a hypothesis of interest, and visualize the input and output data regions of that hypothesis in the current operator results section. Alternatively, the tree view allows a better understanding of the hierarchy of generated hypotheses.

There are two options to continue the exploration from an output data region. The first option is to apply an alternative hypothesis using the dropdown in the next button (this does not guarantee that the alternative hypothesis of choice will be used, as there may not be any significant regions generated by this hypothesis). A second option is to change the reinforcement learning policy and resume the exploration. Finally, once we are satisfied with the generated hypotheses, we can download the pipeline by clicking on save.

REFERENCES

[1] Yoav Benjamini and Daniel Yekutieli. 2001. The control of the false discovery rate in multiple testing under dependency. *Annals of Statistics* 29 (2001), 1165–1188.

[2] Nassim Bouarour, Idir Benouaret, and Sihem Amer-Yahia. 2022. Significance and Coverage in Group Testing on the Social Web. In *WWW '22: The ACM Web Conference 2022, Virtual Event, Lyon, France, April 25 - 29, 2022*. 3052–3060.

[3] David Colquhoun. 2014. An investigation of the false discovery rate and the misinterpretation of p-values. *Royal Society Open Science* 1, 3 (2014), 140216.

[4] Giovanni Di Leo and Francesco Sardanelli. 2020. Statistical significance: p value, 0.05 threshold, and applications to radiomics—reasons for a conservative approach. *European radiology experimental* 4, 1 (2020), 1–8.

[5] Sander Greenland, Stephen J. Senn, Kenneth J. Rothman, John B. Carlin, Charles Poole, Steven N. Goodman, and Douglas G. Altman. 2016. Statistical tests, P values, confidence intervals, and power: a guide to misinterpretations. *European Journal of Epidemiology* 31, 4 (2016), 337–350.

[6] F Maxwell Harper and Joseph A Konstan. 2015. The movielens datasets: History and context. *ACM Trans. on Interactive Intelligent Systems (tiis)* 5, 4 (2015), 1–19.

[7] F Maxwell Harper and Joseph A Konstan. 2016. The movielens datasets: History and context. *ACM Transactions on Interactive Intelligent Systems* 5, 4 (2016), 19.

[8] IMDb. 2021. IMDb Dataset Details. https://www.imdb.com/interfaces/. Accessed: January 20, 2023.

[9] Mohieddin Jafari and Naser Ansari-Pour. 2019. Why, when and how to adjust your P values? *Cell Journal (Yakhteh)* 20, 4 (2019), 604.

[10] Doris Lee, Himel Dev, Huizi Hu, Hazem Elmeleegy, and Aditya G. Parameswaran. 2019. Avoiding drill-down fallacies with *VisPilot*: assisted exploration of data subsets. In *International Conf. on Intelligent User Interfaces*. 186–196.

[11] MatMat. 2023. MatMat dataset. https://github.com/adaptive-learning/matmat-web/. Accessed: July 14, 2023.

[12] Rosa J. Meijer and Jelle J. Goeman. 2016. Multiple Testing of Gene Sets from Gene Ontology: Possibilities and Pitfalls. *Briefings Bioinform.* 17, 5 (2016), 808–818.

[13] Aurélien Personnaz, Sihem Amer-Yahia, Laure Berti-Équille, Maximilian Fabricius, and Srividya Subramanian. 2021. Balancing Familiarity and Curiosity in Data Exploration with Deep Reinforcement Learning. In *Fourth Workshop in Exploiting AI Techniques for Data Management.* 16–23.

[14] Aurélien Personnaz, Sihem Amer-Yahia, Laure Berti-Équille, Maximilian Fabricius, and Srividya Subramanian. 2021. DORA THE EXPLORER: Exploring Very Large Data With Interactive Deep Reinforcement Learning. In *Conference on Information and Knowledge Management*, Gianluca Demartini, Guido Zuccon, J. Shane Culpepper, Zi Huang, and Hanghang Tong (Eds.). ACM, 4769–4773.

[15] Zheguang Zhao, Lorenzo De Stefani, Emanuel Zgraggen, Carsten Binnig, Eli Upfal, and Tim Kraska. 2017. Controlling False Discoveries During Interactive Data Exploration. In *International Conference on Management of Data, SIGMOD*. ACM, 527–540.