Interestingness measures for Exploratory Data Analysis: a survey

Alexandre Chanson[1 [0000−0001−9195−5950]], Nicolas Labroche[1 [0000−0002−2794−2124]], Patrick Marcel[2 [0000−0003−3171−1174]], Verónika Peralta[1 [0000−0002−9236−9088]], and Panos Vassiliadis[3 [0000−0003−0085−6776]]

[1] LIFAT - University of Tours, Blois, France
[2] LIFO - University of Orléans, France
[3] University of Ioannina, Hellas

Abstract. Exploratory Data Analysis (EDA) is the tedious activity of interactively analyzing a dataset to extract insights. Many approaches aiming at supporting EDA were recently proposed. They all rely on interestingness measures to score the importance of insights. This paper surveys and categorizes the different interestingness measures proposed in the literature for approaches aiming at automating EDA. The lessons learned from this survey allow to point out promising research directions.

Keywords: EDA · Insights · Interestingness measures

1 Introduction

Exploratory Data Analysis (EDA) is the notoriously tedious activity of Data Science consisting of interactively analyzing a dataset to gain insights, for "exposing the unanticipated" [34]. According to De Bie et al. [3] EDA poses the greatest challenges for automation, since background knowledge and human judgment are the keys to success. EDA is close to discovery-driven analysis [25–28] that guides the exploration of a datacube by providing users with interestingness values for measuring the peculiarity of the cells in a data cube, with the use of statistical models, e.g., based on the maximum entropy principle, and leveraging the intrinsic structure of multidimensional information. As we will see, EDA does not adhere to the multidimensional model, and the measures proposed go beyond the peculiarities of cube cells.

Recently many approaches were proposed to support EDA, including approaches to automatically generate EDA sessions, often defined as maximization problems (see, e.g., [33, 10, 5, 36]). At the heart of each approach is the quantification of the importance of an *insight*, i.e., a piece of valuable information, for the user analyst, by means of one or more *interestingness measures*. While interestingness measures have been reviewed in several domains (see e.g., [13, 6] for pattern mining or [17] for recommender systems), a survey and organization of interestingness measures tailored for EDA has not been done yet.

This paper fills this gap. We review the measures proposed for EDA, and propose a classification using 6 dimensions from the literature (see e.g., [15]).

This paper appeals to various readers and needs, including: the analyst interested in finding an off-the-shelf EDA system, the researcher looking to devise new EDA support approaches, or the system designer willing to combine measures from different approaches.

The paper is organized as follows. Section 2 details the organization of the survey. Section 3 reviews the measures proposed. Section 4 discusses combinations of interestingness measures. Section 5 discusses lessons learned and perspectives.

2 Categorization of Interestingness Measures

This section explains how we organize the survey. We define what insights are and explain that the interestingness of insights is inherently multidimensional.

2.1 What Are Insights?

Insights are properties or patterns of a subset of a dataset that signify the presence of an interesting relationship among the data participating to the insight. So, practically, an insight is:

- *scoped* by a set of data, typically a subset of a dataset,
- *defined* by the existence of a pattern, or property of the data (e.g., the existence of a peak in a distribution for a particular time point; the existence of seasonality, or the increase along a period of time, or drop in an otherwise steady series, in a time series),
- *computed* via an appropriate algorithm that verifies the presence or absence of the related property,
- *quantified* via a score that measures the degree of the presence of the pattern in the data (e.g., the support of an association rule, the Kendall τ score of the correlation of two measures, etc).

To detect insights, query mechanisms are often used to isolate potential data subsets that are either (a) evaluated on their own, or (b) contrasted to each other for the fulfillment of the insight's defining property. Table 1 lists the insight data subset generation mechanisms most frequently found in the recent literature:

- The Group-by/filter form is the result of an aggregate query over an (often multidimensional) dataset. It was popularized with discovery-driven exploration of datacubes.
- The Sibling group form corresponds to the data of a one-dimensional slice in a multidimensional space.
- The Comparison form corresponds to two series of numbers being compared.

Examples of insights are: a rising trend in yearly sales [33], a factor being relevant to the difference on a given disease between two locations [19], a month having minimum sales for some location [18]. We note that insights can be spurious, i.e., resulting from random data and a particular aggregation [37]. Therefore many works insist that insights should be statistically significant [37, 8, 33].

Table 1. Some popular insight forms

Insight forms	Salient contributions
Group-by / filter	[26, 14, 35, 10, 12]
Sibling groups	[33, 8, 18]
Comparison	[37, 11, 30, 5]

While any piece of data can be an insight, practically, the presence of an insight in a subset of the data, because of the existence of a pattern in these data, separates them from the rest of the dataset as interesting, or at least, potentially interesting for the analyst. What comes out as interesting for an analyst is, however, not immediately obvious. In general, the interestingness of an insight can be quantified via **an interestingness score**. Again, the semantics behind the interestingness score can be diverse; in the sequel, we try to organize these semantics along a principled framework.

2.2 Interestingness is a Multidimensional Notion

Two main approaches are used to capture the insights' interestingness: (i) the definition of heuristic measures and (ii) machine learning. Many heuristic measures were proposed, each capturing a different facet of the broad concept. However, as reported in [21], there is no single measure that consistently outperforms the others, interestingness being often subjective and changing dynamically [32]. This is why some works resort to machine learning (e.g., [10]) to dynamically select interestingness measures (and often combine them) or to model the users' interest with active-learning or learning-to-rank techniques. **We deliberately focus on heuristic definitions because they help understanding the nature of interestingness** in many ways – most importantly, as they are able to explain why a particular insight is proposed to the user.

Patil et al. [22] propose to evaluate EDA approaches using 3 categories of metrics: human, system and data.

- Human: quantitative and qualitative measures to evaluate user satisfaction (through questionnaires, tracking, etc.)
- System: measures evaluating the resources (memory, latency) consumed by the system. TPC benchmarks abound with this type of measures.
- Data: measures proposed to qualify an interesting property or pattern for a subset of the data in a dataset, often called insight, highlights, findings, discoveries, data facts, etc. [26, 14, 35, 10, 21].

In this work we focus on the third one, specific to EDA. Earlier works addressed the classification of these criteria [13, 20, 15], in particular contexts (pattern mining, data cube exploration) without actually reviewing and analyzing the measures proposed. We start by explaining what interestingness dimensions are.

We adopt a multidimensional view point, and propose 6 dimensions for characterising insight interestingness: *peculiarity, novelty, relevance, surprise, diversity* and *presentation* (they are defined in next paragraphs). These dimensions are inspired by the seminal work for cube exploration [13], where the authors review interestingness measures for results of OLAP queries, and by recent works [20, 15] reworking such classification, and proposing interestingness aspects for datacubes, grounded by human behavior studies[4].

These dimensions are orthogonal, and have the advantage of clearly indicating what is needed to compute interestingness. We describe them hereafter, providing the signatures of the functions implementing their evaluation, and highlighting what is contrasted to generate interestingness.

- $Peculiarity(i, D)$: The peculiarity of insight i indicates whether data of i is different and not in accordance to other data. An insight is contrasted to other data for commonalities or differences. Therefore, peculiarity depends on the dataset D where i comes from.
- $Novelty(i, H)$: The novelty of novelty of insight i indicates whether i is new and previously unseen. Thus, an insight is contrasted to a user's history, and novelty depends on the history H of data seen before i.
- $Relevance(i, g)$: The relevance of insight i indicates whether i is related to the overall analysis intention of the user, expressed in the user's exploration goal. Therefore, relevance depends on the user's goal g.
- $Surprise(i, b)$: The surprise of insight i indicates whether i contradicts and revises the user's previous beliefs. Therefore, surprise depends on the belief b of the user.
- $Diversity(i, C)$: The diversity of insight i indicates whether i covers various classes of the underlying data. Such classes may represent the user's targeted groups where a fair coverage is desirable (e.g., the values of a sensitive attribute like gender). Therefore, diversity depends on the coverage of user's targeted classes C;
- $Presentation(i)$: The presentation of insight i indicates the difficulty for understanding i. This includes (but is not limited to its conciseness). Therefore, presentation depends on i itself.

Papers selection The papers reviewed in this survey were chosen based on the following considerations:

- we focus on approaches automating EDA proposed by the data management community;
- note that EDA approaches were already surveyed in the data management community [16]. The focus was slightly different (how to store and access data, how to interact with a data system to enable users and applications to quickly figure out which data parts are of interest). We mostly chose papers posterior to that survey, since they attach more importance to notions of interestingness and insights;

[4] [13] proposes *peculiarity, surprise, diversity* and *presentation* (some of them with different names) and [20, 15] propose *relevance, novelty, peculiarity* and *surprise*.

– some less recent but influential papers were included nonetheless (e.g., [7]) if they are key to understand important concepts;
– we chose papers pertaining to the most popular form of insights and pertaining to heuristic definitions of interestingness (see Sections 2.1 and 2.2).

3 The Variety of Interestingness Measures for EDA

We review the measures proposed, according to the dimensions introduced in the previous section. For each dimension, we identify refinements based on the semantics of the measures proposed. We also indicate the importance of the dimension in helping building EDA explorations.

3.1 Peculiarity

Peculiarity allows to **quantify the importance of an insight among its peer data by evaluating how deviant, or, common the data of the insight are compared to the rest of the dataset**. The measures defined in this dimension concern either (i) the **outlierness**, or (ii) the **typicality** of the insights. Using this dimension, an analyst, or a recommendation system, can **steer the exploration to phenomena** (trends, outliers, etc.) or to **better represent the dataset**.

Outlierness. The *outlierness* of an insight quantifies its interestingness based on its difference with a broader set of data to which it is contrasted. Sintos et al. [31] measure the extent of the incorrectness of a value in a dataset (practically measuring the amount of false information of two values before and after a data cleaning procedure). Gkitsakis et al. [15] compute the outlierness of a newly posed cube query by aggregating the distances between the data of the query and past data retrieved. Many approaches consider the distribution of data [33, 37, 8, 12, 1, 5]. In [12], outlierness is measured using the difference in z-scores of the data obtained in two consecutive exploration steps. A recent trend is to turn insights into hypothesis testing [33, 37, 8, 5], which allows to: (i) use the p-value for the insight significance, (ii) define false discoveries (type-1 errors, e.g., visualizations supporting a non-significant insight) and false omissions (type-2 errors, e.g., visualizations not supporting a significant insight), (iii) define credibility (e.g., percentage of visualizations supporting an insight). However, since the risk of type-1 error increases as more than one hypothesis are considered at once, a correction is needed to ensure reporting only non-spurious insights [37].

Typicality. Measuring the *typicality* of the insight consists of quantifying to what extent the subject of an insight can represent the *entire dataset* [33, 8, 18]. In most cases, anti-monotonic conditions are checked to prune insights. For instance: if the subject of insight A is a superset of the subject of insight B, then the impact of A should be no less than the impact of B. The *market share* measure used in [33, 8] is defined as the ratio between the sum of values of the insights and the sum of all data.

3.2 Novelty

Novelty characterizes insights in terms of **being new observations** (or operations). Using this dimension allows to **make the exploration go further** or **make it focused**.

In its simplest expression, novelty is measured as a Boolean indicating whether some data have already been seen [12]. In [15], novelty is computed as the fraction of new data brought by a current query compared to the data retrieved by that query and the previous ones, either per se, or in different degrees of granularity. In [23], curiosity is measured as a function of the number of times a result is encountered (being inversely proportional to it).

3.3 Relevance

This dimension characterizes insights in terms of **fulfilling a user's goal** or being *familiar and coherent* to the user. The measures defined in this dimension concern (i) **goal fulfillment**, (ii) **familiarity** or (iii) **coherency**. Using this dimension allows to make explorations **connected to the analyst's interests**.

Goal fulfillment. Gkitsakis et al. [15] distinguish two ways goals are declared: (a) explicit, directly stated by the user under the form of selection predicates over the dataset, or, (b) implicit, i.e., goal is approximated and estimated by the system. In case (a), the relevance of data computed by a query is measured as the fraction of data from the dataset it covers (i.e., the data used by the query) that overlaps with the user's goal. In case (b), the goal is inferred from the user's history (queries sent in the past), and relevance is measured as in case (a). The basic idea of the approach is openness: any other means of deriving a goal for the analyst can be plugged into the mechanisms, while retaining the fundamental essence of a goal, which is coarsely speaking, a "fence" that isolates the relevant subset of the data space (within the exploration goal) from the irrelevant one.

Familiarity. In [23], a familiarity measure is defined as the concentration ratio of target data in a set. It is implemented as a variant of the Jaccard index between data encountered during the exploration and a given target set of familiar data. This measure is expected to increase as the EDA session goes on, to avoid over-exploiting a set of familiar objects.

Coherency. The *coherency* of an insight contrasts the insight with other insights obtained in the *exploration session*, to check whether a given EDA operation is coherent at a certain point. For instance, in [10] heuristic classification rules are used to express general properties on the input dataset semantics (e.g., if the user focuses on flight delays, aggregating on the "departure-delay time" column is preferred). Some other works express coherency as a distance between exploration actions (separate from their definition of interestingness) and measure how coherent a sequence of actions is as a whole. For instance, in [5] a weighted Hamming distance of relational query parts is used.

3.4 Surprise

This dimension allows to characterize insights in terms of **how distant they are from the user's expectations**. The measures defined in this dimension express a **distance to expectations**. Using this dimension allows to **steer explorations to data showing unexpected values**.

A formal framework for defining measures of surprise has been introduced by De Bie for exploratory data mining [7]. Using an information-theoretic approach, the framework consists of quantifying the interactive exchange of information between data and user, accounting for the *user's prior belief state*. Approximating the belief that the user would attach to the result being expected is modeled as a background distribution, namely, a probability measure over the exploration results. This background distribution is updated after each result is presented to the user. Chanson et al. [4], propose a way to measure subjective interestingness for exploratory OLAP, inspired by De Bie's work [7]. The user belief is inferred based on the user's past interactions over a data cube, the cube schema and the other users' past activities. This belief is expressed by a probability distribution over all the query parts potentially accessible to the user. Surprise is then measured as in De Bie's work.

In the seminal work of Sarawagi [26], belief (i.e., expected values) is computed using maximum entropy principle, and Kullback-Leibler divergence is used to measure surprise. Gkitsakis et al. [15] distinguish two ways to account for beliefs: (a) expected values are provided by the user, or, (b) expectations are registered by annotating the expectation for a value to appear via a probability of appearance. In case (a), the surprise is measured using a distance function between actual data and expected data. In case (b), surprise for a given value is measured as the sum of the probabilities of all values that are different.

3.5 Diversity

Diversity characterizes insights in terms of **their coverage of population classes**. Using this dimension allows to **make the exploration more representative of the underlying dataset**.

Simple versions of diverstiy measures have been proposed. In [10], a diversity measure is introduced to encourage the analysis of different parts of the dataset. It is computed as the minimal Euclidean distance between the current observation and all the previous displays obtained. Francia et al. [12] also measure diversity[5] as the proportion of values that have not been seen frequently, presented in models (e.g., clustering) extracted from the insight. In [36], diversity is measured as the pairwise difference between insights. In [24] the authors use a pairwise Jacquard similarity to measure diversity within their sub tables.

3.6 Presentation

This dimension characterizes either **how compact** the insight is when presented to a user, or **the amount if information** the insight displays. The measures

[5] Called surprise in [12], but reclassified here since it does not refer to a user's belief.

Table 2. What interestingness dimensions are combined (left part) and how they are combined (right part)

Contribution	Rel.	Nov.	Pec.	Sur.	Div.	Pre.	ratio	product	[weighted] sum
ATENA [9, 10]	✓		✓		✓	✓			✓
B.I.lief [4]				✓					
Calliope [29]			✓					✓	
Cube Query Int. [15]	✓	✓	✓	✓					
DataShot [35]			✓						✓
Describe [12]		✓	✓		✓				✓
DORA [23]	✓	✓							✓
EDA4Sum [36]		✓	✓		✓				✓
Forsied [7]				✓		✓	✓		
Metainsight [18]			✓			✓		✓	
Quickinsights [8]			✓					✓	
SubTab [24]			✓		✓				✓
TAP-Comparisons [5]			✓			✓		✓	
Top-k insights [33]			✓					✓	

defined in this dimension concern either (i) the **compactness**, or (ii) the **descriptional complexity** of the insights. Using this dimension allows to **favor insights being both informative and easy to understand**.

Descriptional complexity. Descriptional complexity measures how complex it is for a human to assimilate an insight [7]. For instance, the complexity of a set of values can be the number of elements in a set.

Conciseness Conciseness measures how compact is an insight. For instance, when presenting aggregated results over a set of tuples, the ratio of tuples to groups or a function thereof can be used as a rough estimate of the chosen groups ability to summarize large quantity of information (tuples) [10, 5]. Conciseness can also be defined as a measure of entropy of the insight, acting for a proxy to the human effort necessary for its assimilation [18]. In this later form it also fits the definition of *descriptional complexity* of [7].

4 Combining Interestingness Dimensions

This section shows how interestingness dimensions are combined. Usually, insights are scored based on more than one dimension, to account for goal, history, or belief, or combinations thereof. Table 2 (left) indicates which dimensions of interestingness are commonly used together. Peculiarity is the most frequent dimensions used. Noticeably, there is no consensual approach as how dimensions are combined. For instance, a ratio is used in [7], a weighted sum is used in [10], and a product is used in [5]. This is summarized in Table 2 (right).

5 Conclusion

This paper surveys interestingness measures proposed to support Exploratory Data Analysis. The main lesson learned is that **no definitive measures or combinations of interestingness dimensions** have already been proposed. Some dimensions, like **peculiarity, attracted lots of attention** while others, like diversity, relevance, or surprise, that confront insights with the user's goals or beliefs, much less so.

This survey opens several research directions:

- development of new interestingness measures: the analyst is at the center of the data exploration activity, and measures tailored for personalized or collaborative EDA [2] are still to be proposed,
- formalizing the desirable properties of interestingness measures: in the spirit of what was done for pattern mining [13], the properties of interestingness measures will provide a fine understanding of how measures should be combined,
- contextualizing interestingness dimensions: a typology of EDA sessions is yet to be done. This will enable the characterization of what interestingness measures are required at what step of a given type of EDA session.

References

1. F. Abuzaid, P. Kraft, et al. DIFF: a relational interface for large-scale data explanation. *VLDB J.*, 30(1):45–70, 2021.
2. S. Amer-Yahia, P. Marcel, et al. Data narration for the people: Challenges and opportunities. In *EDBT*, pages 855–858. OpenProceedings.org, 2023.
3. T. D. Bie, L. D. Raedt, et al. Automating data science. *Commun. ACM*, 65(3):76–87, 2022.
4. A. Chanson, B. Crulis, et al. Profiling user belief in BI exploration for measuring subjective interestingness. In *DOLAP*, volume 2324 of *CEUR Proceedings*, 2019.
5. A. Chanson, N. Labroche, et al. Automatic generation of comparison notebooks for interactive data exploration. In *EDBT*, pages 2:274–2:284, 2022.
6. V. Dadvar, L. Golab, et al. Exploring data using patterns: A survey. *Inf. Syst.*, 108:101985, 2022.
7. T. De Bie. Subjective interestingness in exploratory data mining. *IDA*, 8207:19–31, 2013.
8. R. Ding, S. Han, et al. QuickInsights: Quick and automatic discovery of insights from multi-dimensional data. In *Proceedings of SIGMOD*, pages 317–332, 2019.
9. O. B. El, T. Milo, et al. ATENA: an autonomous system for data exploration based on deep reinforcement learning. In *CIKM*, pages 2873–2876, 2019.
10. O. B. El, T. Milo, et al. Automatically generating data exploration sessions using deep reinforcement learning. In *SIGMOD*, pages 1527–1537, 2020.
11. M. Francia, M. Golfarelli, et al. Assess queries for interactive analysis of data cubes. In *EDBT*, 2021.
12. M. Francia, P. Marcel, et al. Enhancing cubes with models to describe multidimensional data. *Inf Syst Front*, 24(1), 2021.

13. L. Geng and H. J. Hamilton. Interestingness measures for data mining: A survey. *ACM Comput. Surv.*, 38(3):9, 2006.
14. D. Gkesoulis, P. Vassiliadis, et al. Cinecubes: Aiding data workers gain insights from OLAP queries. *Inf. Syst.*, 53:60–86, 2015.
15. D. Gkitsakis, S. Kaloudis, et al. Cube query interestingness: Novelty, relevance, peculiarity and surprise. *Information Systems*, 123:102381, 2024.
16. S. Idreos, O. Papaemmanouil, et al. Overview of data exploration techniques. In *SIGMOD*, pages 277–281. ACM, 2015.
17. M. Kaminskas and D. Bridge. Diversity, serendipity, novelty, and coverage: A survey and empirical analysis of beyond-accuracy objectives in recommender systems. *TiiS*, 7(1):2:1–2:42, 2017.
18. P. Ma, R. Ding, et al. MetaInsight: Automatic discovery of structured knowledge for exploratory data analysis. In *Proceedings of SIGMOD*, pages 1262–1274, 2021.
19. P. Ma, R. Ding, et al. Xinsight: Explainable data analysis through the lens of causality. *Proc. ACM Manag. Data*, 1(2), June 2023.
20. P. Marcel, V. Peralta, et al. A framework for learning cell interestingness from cube explorations. In *ADBIS*, pages 425–440. Springer, 2019.
21. T. Milo and A. Somech. Automating exploratory data analysis via machine learning: An overview. In *SIGMOD*, 2020.
22. Y. Patil, S. Amer-Yahia, et al. Designing the evaluation of operator-enabled interactive data exploration in VALIDE. In *HILDA@SIGMOD*, pages 4:1–4:7, 2022.
23. A. Personnaz, S. Amer-Yahia, et al. DORA THE EXPLORER: exploring very large data with interactive deep reinforcement learning. In *CIKM*, 2021.
24. K. Razmadze, Y. Amsterdamer, et al. SubTab: Data exploration with informative sub-tables. In *SIGMOD*, pages 2369–2372, 2022.
25. S. Sarawagi. Explaining differences in multidimensional aggregates. In *Proceedings of VLDB*, pages 42–53, 1999.
26. S. Sarawagi. User-adaptive exploration of multidimensional data. In *VLDB*, pages 307–316, 2000.
27. S. Sarawagi, R. Agrawal, et al. Discovery-driven exploration of OLAP data cubes. In *EDBT*, volume 1377 of *LNCS*, pages 168–182, 1998.
28. G. Sathe and S. Sarawagi. Intelligent rollups in multidimensional OLAP data. In *Proceedings of VLDB*, pages 531–540, 2001.
29. D. Shi, X. Xu, et al. Calliope: Automatic visual data story generation from a spreadsheet. *TVCG*, 27(2):453–463, 2021.
30. T. Siddiqui, S. Chaudhuri, et al. COMPARE: accelerating groupwise comparison in relational databases for data analytics. *VLDB*, 14(11):2419–2431, 2021.
31. S. Sintos, P. K. Agarwal, et al. Selecting data to clean for fact checking: Minimizing uncertainty vs. maximizing surprise. *Proc. VLDB Endow.*, 12(13):2408–2421, 2019.
32. A. Somech, T. Milo, et al. Predicting "what is interesting" by mining interactive-data-analysis session logs. In *EDBT*, 2019.
33. B. Tang, S. Han, et al. Extracting top-k insights from multi-dimensional data. In *SIGMOD*, 2017.
34. J. W. Tukey. *Exploratory Data Analysis*. Addison-Wesley, 1977.
35. Y. Wang, Z. Sun, et al. Datashot: Automatic generation of fact sheets from tabular data. *TVCG*, 26(1):895–905, 2020.
36. B. Youngmann, S. Amer-Yahia, et al. Guided exploration of data summaries. *Proc. VLDB Endow.*, 15(9):1798–1807, 2022.
37. E. Zgraggen, Z. Zhao, et al. Investigating the effect of the multiple comparisons problem in visual analysis. In *Proceedings of CHI*, page 479, 2018.