A Conceptual Model for Data Analysis Highlights

Panos Vassiliadis[1], Veronika Peralta[2], Patrick Marcel[3], Dimos Gkitsakis[1], Angeliki Dougia[4] and Faten El Outa[2]

[1]Univ. Ioannina, Ioannina, Greece
[2]University of Tours, Blois, France
[3]University of Orleans, Orleans, France
[4]P&I Hellas, Ioannina, Greece; work done while with the Univ. Ioannina
[2]University of Orleans, Orleans, France

Abstract

We introduce a conceptual model for highlights to support automated data analysis and storytelling. Highlights reveal key facts, of high significance, that are hidden in the data with which a data analyst works. The model builds on the concepts of Holistic and Elementary Highlights, along with their context, constituents and interrelationships, whose synergy can identify internal properties, patterns and key facts in a dataset being analyzed. We also report how the related literature fits within the model, as well as a first implementation of it.

Keywords

Highlights, Data exploration, Data storytelling

1. Introduction

Data analysis concerns the processing of large volumes of data for the identification of important information hidden in them, or derivable from them. Meliou et al., [1] offer an excellent classification of analysis tasks as (a) descriptive analytics, that report properties and extract patterns from the underlying data, (b) diagnostic analytics, conducting analyses (e.g., causal inference) to explain the reasons behind the observed state of affairs, (c) prescriptive analytics, mainly solving optimization problems to derive the best configuration of parameters for optimal solutions to complex problems, and, (d) predictive analytics, mainly using historical data and simulation models to predict future trends as well as the effects of interventions. The results of such analysis tasks are first to be extracted by the analysts, interpreted and evaluated for the significance, compiled into a data-based story of "what – why – how - next" and finally be communicated, in humanly understandable format, to broader audiences of non-technical users and decision-makers via data narration methods.

As, more and more, in their attempt to extract meaningful results, data analysts have to overcome the time pressure, the learning curve of analytics tools, and the growing volume and variety of data, there is unavoidably a growing need for *automating* the process of (a) efficiently and (b) effectively discovering significant characteristics of the underlying data sets, that are subsequently used to derive decisions and actions. Already, several efforts towards automating the results of analytical questions have been made [2, 3, 4, 5, 6, 7]. The typical terminology for these automatically extracted answers are *findings*, also called *insights*, and *data facts*. These concepts are data-oriented and concern the identification of parts of the data space that demonstrate interesting properties (typically a pattern or relationship between its data). When, for example, a time series demonstrates a trend, or a histogram follows a certain distribution, then the data demonstrate a phenomenon that may, or may not, be interesting to the analyst. *The key contribution of this paper is the introduction of a richer conceptual model for highlights, which, much like findings, isolate the parts of the data that make the existence of an*

ER2025: Companion Proceedings of the 44th International Conference on Conceptual Modeling: Industrial Track, ER Forum, 8th SCME, Doctoral Consortium, Tutorials, Project Exhibitions, Posters and Demos, October 20-23, 2025, Poitiers, France
pvassil@cs.uoi.gr (P. Vassiliadis); first.last@univ-tours.fr (V. Peralta); first.last@univ-orleans.fr (P. Marcel); dgkits@cs.uoi.gr (D. Gkitsakis); adougia@pi-ag.com (A. Dougia); first.last@univ-tours.fr (F. E. Outa)
0000-0003-0085-6776 (P. Vassiliadis); 0000-0002-9236-9088 (V. Peralta); 0000-0003-3171-1174 (P. Marcel); 0009-0006-9559-819X (D. Gkitsakis); 0000-0002-8153-4252 (F. E. Outa)

interesting phenomenon true, but also, come with structured extra information with respect to (i) their roles, relationships and provenance, as well as, (ii) the necessary information that explains why these data are of importance. But first, we will illustrate the case with an example.

Working Example. Assume we have a relational database with product sales, and for reasons of comprehension, we structure it as a cube. The fundamental source of information is a fact table *SalesFact(ProductId, TimeId, CityId, PromotionId, CustomerId, Sales, Costs, UnitSales, AvgSalesPerUnit, Profit)*, with all the monetary measures expressed in thousands of Euros. Around the aforementioned fact table, we also have several lookup dimension tables, namely *Products, Time, Cities, Promotions, Customers*, joinable to the fact table via the respective Id attributes (i.e., we have a PK-FK relationship between fact and dimension tables for each Id attribute).

Assume a query that selects the total sum of sales of the product *Wine* for the 2nd quarter of 2023 in Greece, grouped by month and city (Table 1). Observe the following *highlights* of the result set:

- The city *Athens* dominates all other cities: for every month, the sales of Athens are higher than the sales of every other city.
- The month *May 2023* dominates all other months: for every city, the sales of May are higher than the sales of other months.
- The city of *Athens* is a mega-contributor to the total sales: the sales of *Athens* are 75% of the total sales.
- If one observes the time-series of the marginal sales per month, there is no trend or seasonality; however there is a unimodality in the time-series: sales rise, reach a peak (in *May*), and then drop.

Month	*Athens*	*Rhodes*	*Chania*	*Thera*	*Total*
April 2023	500	50	85	80	*715*
May 2023	1000	70	90	120	*1280*
June 2023	600	65	70	70	*805*
Total	*2100*	*185*	*245*	*270*	***2800***

Table 1
Reference Example

The following (automatically derivable) textual summary reports the discovered highlights grouped around the main *characters* of Athens and May:

> *In terms of geography, Athens dominates all other cities, in every month. In fact, Athens is a mega-contributor to total sales, by contributing 75% of all sales. In terms of time, the progression in time shows a peak in May; in fact, May dominates all other months in terms of total sales. No trend or seasonality were detected.*

Besides data analysis, highlights are of capital importance for data narration. When a data narrative is ultimately constructed (via a process also known as data storytelling, which falls outside the scope of this paper), these highlights will provide the basis for automatically generating text, chart annotations or choosing of visualizations (e.g., type, style, position). Highlight constituents, in particular main *Characters* (as Athens or May 2023 in the example) and *Measure values* (e.g., Athens total of 2100K€, representing 75% of total sales), may also guide the choice of specialized visual artifacts (e.g., colors or effects).

To support data analysts in the retrieval, understanding and communication of highlights, we first need to handle two important problems: (a) to clarify the involved concepts in the production of highlights, and (b) to present a unifying conceptual model that covers a large number of automatically extracted highlights and allows to uniformly handle them via automated tools under common syntax and semantics.

Contribution. *The main contribution of this paper lies in the provision of a comprehensive and precise conceptual model for highlights, along with their constituents and their interrelationships, with the goal to*

help system builders implement tools and algorithms that facilitate the automated extraction, representation, and exploitation of highlights in data, for data analysis and storytelling purposes.

Outline. In Section 2, we survey related work. In Section 3, we introduce the model for highlights. We first describe supporting concepts and introduce the main concepts of *Character* and *Measure Value*, and then define highlights, organizing them as (a) *Holistic Highlights*, which are properties of the entire dataset being examined, and, (b) *Elementary Highlights* which concern individual *Characters*, or sets of them, that play a crucial role to *Holistic Highlights*. In Section 4, we discuss the relationship of the proposed model with the state of the art. In Section 5, we discuss the practical usage of the proposed model. Finally, in Section 6 we conclude with a summary and ideas for future work.

2. Related Work

Data analysis. As already mentioned, data analysis can be (a) descriptive, (b) diagnostic, (c) prescriptive, and, (d) predictive [1]. Traditionally, statistics, data mining and business intelligence offered descriptive and diagnostic analysis support. In recent years, these traditional techniques have been complemented with Exploratory Data Analysis (EDA), where users are interactively analyzing datasets to gain insights [8, 9, 10, 11, 12, 13, 14]. Supporting this task can be done, e.g., by generating EDA notebooks using deep learning [15] which supposes having access to lots of former analyses, or by pre-analyzing datasets for computing highlights [3, 16, 5]. New tools combine LLM with data exploration techniques [7, 14]. EDA is similar to *Discovery-Driven Exploration* (DDE) of data cubes [17], essentially motivated by explaining unexpected data in the result of a cube query. Gkesoulis et al. [2] demonstrated how to enrich query answering with a short data movie that provides highlights for the results of an OLAP query, albeit with hard-coded highlights and without a general model. Many recent tools propose automating not only data exploration but the whole data narration process [4, 6, 18, 19, 7].

Defining highlights. There is no clear consensus on the terminology used for the case of highlights. *We believe that a clarification of the related concepts is one of the contributions of this paper.* The term *insight* is well adopted in the data management and data visualization communities [16, 5], and other terms are also used, like *discoveries* [17], *data facts* [4] or *findings* [20, 21]. To the extent that insight is "the act or result of apprehending the inner nature of things or of seeing intuitively" according to Merriam-Webster, and, hence, a perception-based concept, we prefer to adopt the term *highlights* for the discoveries made in data, and, the assessment of their significance.

In the conceptual model of [20], highlights are the more striking, surprising or relevant facts of a query result. Similar definitions are given in VOOL [22] (interesting facts in a cube) and DAISY [23] (a table containing interesting data). In these works, however, the structure of highlights is not detailed. In a large number of works [3, 16, 4], and most notably MetaInsight [5], a highlight is characterized by the subset of the data space that generates, possibly with a group-by clause, a type of the pattern that is identified over a measure of this data sub-space, and a score of its importance. These definitions are fairly similar in other tools: Calliope [6], Erato [18], Notable [19] and InsightPilot [7].

Interestingness of highlights. Characterizing meaningful highlights in data has attracted a lot of attention, since the seminal works on discovery driven exploration [17, 24] and knowledge discovery in databases [25]. Often, this characterization takes the form of *interestingness* scores for retrieved data [26] or *patterns* [27, 28]. Scoring a highlight is used to express its importance, or interestingness, for the user. As explained in [26, 29], interestingness is manifold: scores can be computed for different dimensions of interestingness. In the taxonomy proposed in [12], following [30], interestingness of highlights can be characterized with human, system or data metrics. Chen et al. [21] stress the importance of the relationships between highlights for the selection of the more important ones, while other authors exploit their *significance*, defined in terms of statistical tests [31, 3, 32, 33, 34, 4], Shapley values [35], or information theory [6].

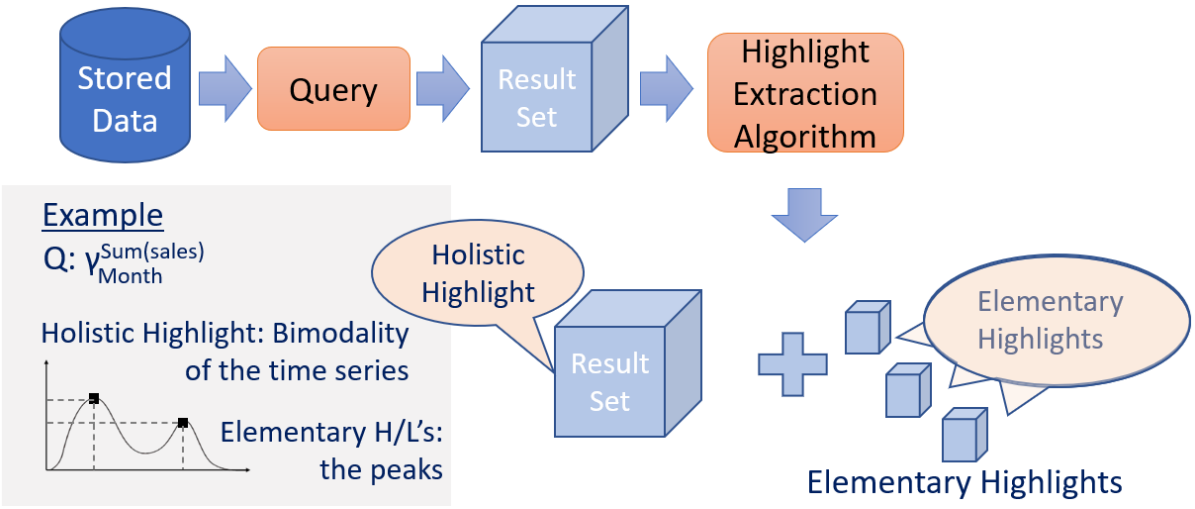

Figure 1: The generic process of highlight extraction

3. A model for highlights

In general, the process of automated highlight extraction (Fig. 1) applies a set of Highlight Extraction Algorithms over a dataset, which can be, e.g., the result of a query. These algorithms operate like pattern-matching testers, checking whether the data abide by a certain pattern or not. Examples of possible questions such algorithms might ask are: (a) is there a bimodality in a time-series produced as a query result, and if yes, which are the peaks?, (b) if we breakdown the total sum of sales by product type, is there a "mega-contributor" product type with more than 40% of total sales, and if yes, who is it?, (c) assuming we sum total sales per month and product is there any month that systematically outperforms all other months for all products, and if yes, who is it?

Given a dataset, highlights are important properties of the entire dataset, extracted via the respective algorithms, along with the necessary data and statistical properties, as answers to the above questions. In this Section, we present a model for highlights, with Fig. 2 serving as its visual depiction.

3.1. Supporting concepts

A **Dataset** is a set of **Facts**, i.e., structured observations of a domain, under a **Schema**. Schemata define a template internal structure for the facts of the dataset, and include a set of **Features**, each with a domain of values. For the purpose of our model, features are of two kinds: **Measure Types**, for measurable quantities of facts, and, **Character Types** for contextual, lookup dimensions of facts. In our reference example, the query result is a Dataset with $Month$ and $City$ as its Character Types, and $TotalSales$ as its Measure.

Characters are the main entities of the domain being modeled (which in the OLAP terminology would be named as *Dimension Members*) and belong to the domain of Character Types. A Character Type comes with the following characteristics: (a) an Id that uniquely identifies each Character, (b) a $Description$ that provides a textual, human-relatable description of the Character, and (c) a set of $Characteristic\ Properties$, that each Character Type carries with it. For example, assume that the Character Type $City$, has the structure $City(Id, Description, Area, Population)$. A possible character would be *Athens* $\langle$ *101, "Athens Metropolitan Area", 2928* Km^2*, 3.7M* $\rangle$. A **Measure Value** is an instance of a Measure Type and belongs to its domain.

To be able to discriminate between attributes and values playing different roles in the same highlight, we need to beforehand introduce the concept of a **Role** as a template that annotates a class with specific attributes, specifically: (a) a unique *name*, (b) an accompanying *textual description*. The *Main Measure Role* and the *Explanator Role*, introduced in the sequel, are incarnations of roles, each with a name and textual information that describe their function and discriminate them from other Roles.

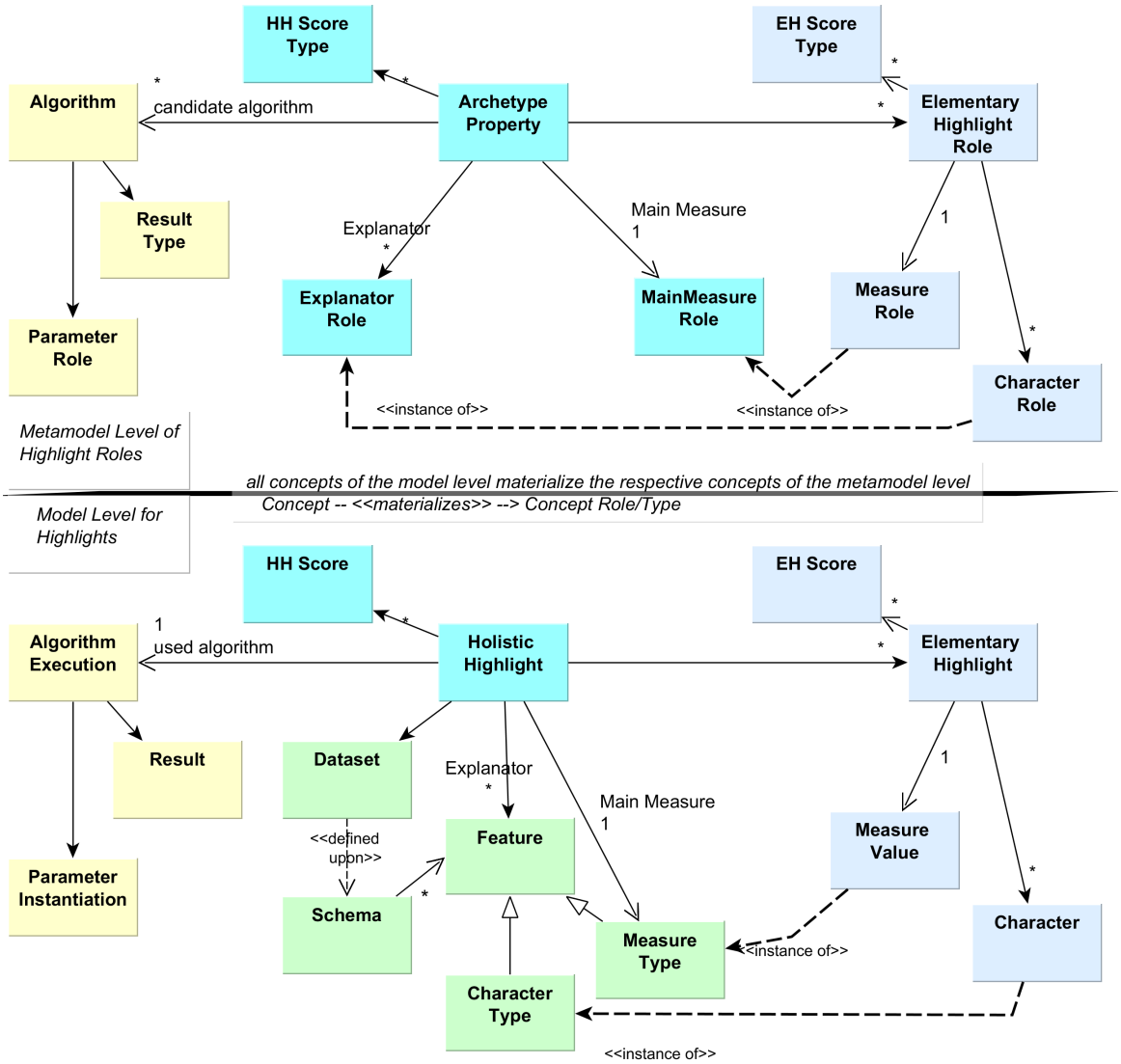

Figure 2: The metamodel for highlights. The materialize relationships between homologous concepts are omitted for clarity

3.2. Holistic Highlights

We start our discussion at the Metamodel level, which involves abstract concepts, not bound to any specific dataset. We will handle how the abstract concepts are materialized into discoveries of significance for a specific dataset at the Model level, introduced in the sequel.

An **Archetype Property** is a specific pattern of internal relationships that can appear between the contents of a dataset. We assume an extensible set of Archetype Properties, each with well-known semantics.

Examples of Archetype Properties include the existence of a trend in a timeseries, the correlation of two Measure Types, the existence of mega-contributors, etc. One can think of Archetype Properties as *hypotheses* that are to be verified when tested over a specific dataset. As we will see in the sequel, whenever a data set supports an Archetype Property, a Holistic Highlight is born as a testimony for the validation of the hypothesis (Fig. 3).

Since we work in the area of data analysis, each Archetype Property operates over a Measure Type of a dataset. The **Main Measure Role** of the Archetype Role is a placeholder that captures the details and constraints of the studied Measure Type. For example, when testing for mega-contributors, where a certain data point contains a large percentage of a marginal sum, we need a Main Measure that is

additive, such that the summation is allowed.

Similarly, for the Archetype Property to eventually create a highlight, certain Features of a data set need to be reserved for specific roles. We reserve an **Explanator Role** for each such Feature. For example, when studying a time series trend, apart from the Main Measure being studied, we need to specify a feature playing the role of a sorter that defines the time ordering of the measure values (equivalently: what would be in the x-axis of a line chart, if the y-axis would be the measure). For instance, a $Sales$ dataset can have several time attributes, $OrderDate$, $DispatchDate$, $ArrivalDate$, while the testing of the trend needs to be done with respect to one of them.

To test the hypothesis that an Archetype Property poses, we need a family of generic, parameterizable Highlight Extraction Algorithms, or **Algorithm**s for short, that are to be applied over incoming datasets. For example, algorithms like Shapiro-Wilk or Kolmogorov-Smirnov test the normality of the distribution of a measure; Pearson, Spearman or Kendall assess correlation; Difference-Precomputation or Boolean-based tests can compute Uni/Bi-modality, etc. In all these cases, there are several candidate, schema-agnostic algorithms to be applied over any possible Dataset, in order to check an Archetype Property.

All these algorithms have a set of parameters that have to be fixed whenever executed, as well as a result. At the meta-level, parameters are modeled as **Parameter Roles**, and results as **Result Types**. Assume, for example, a linear regression algorithm that can optionally take a pre-specified intercept as parameter: in this case, the Algorithm's signature at the meta-level contains the respective Parameter Role. Result Types need the actual Results to have been introduced, so we defer their discussion for the next paragraph.

Whenever an entire Dataset is tested for the existence of an Archetype Property in its contents via a specific algorithm, the resulting verdict of the algorithm's execution along with important facts and their interrelationships that verify the Archetype Property constitute a Holistic Highlight.

A **Holistic Highlight** is a significant, structured testimony for the existence of an Archetype Property over a specific dataset that is automatically tested via a dedicated algorithm and characterized accordingly.

Holistic Highlights characterize an entire dataset with respect to an Archetype Property and not specific data points. A couple of illustrative examples follow:

1. *Distribution.* The *distribution of the values* of a certain Measure Type, say M, follows the *Normal distribution* model. The test was performed by a Shapiro-Wilk test and the p-value is 10^{-4}.
2. *Correlation.* The *correlation* of a Measure Type, say M, with another Measure Type M', is characterized as *significant*. The test was performed via a Kendall test and the tau-value is 0.83.

A Holistic Highlight materializes an *Archetype Property* over a *Dataset*. The rest of the elements of a Holistic Highlight are structured and instantiated in direct correspondence to the elements of the Archetype Property they testify. Hence, a Holistic Highlight concerns a **Main Measure**, i.e., a specific Measure Type of its Dataset, and a set of **Explanators**, i.e., specific Features. Moreover, a Holistic Highlight comes with an **Algorithm Execution**, (i.e., the execution of a specific *Algorithm* among the *Algorithms* of its Archetype Property), with each *Parameter Role* being assigned a concrete **Parameter Instantiation**, and producing a concrete Result.

The **Result** of an Algorithm Execution is a structured testimony that tells us whether the Archetype Property exists or not. To define results, we first need to introduce an illustrative example. Assume we run a Simple Linear Regression over a dataset measuring the sales of cities as a function of the population in thousands. The result of an algorithm assessing what is the best linear relationship possible, will produce a *technical result* containing, among others:

- The constituents of the resulting *Model*, specifically, an *Intercept* and a *Slope*.
- A set of *auxiliary metrics* characterizing the result; in the case of an SLR these will include, the MSE, R^2 and a p-value.

At the Metamodel Level						At the Model Level	
The **Archetype Property**	to be tested over **Main Measure Role** *<constraint>*	wrt **Explanator**<supporting txt> *<constraint>*	via candidate **Algorithms** *(parameters omitted)*	having **Result Type**	and **Elementary Highlight Roles**	for measure **Main Measure**	wrt **Explanator**
Normal Distribution	*Any measure*	-- --	Shapiro-Wilk, Kolm.-Smirnov	- {W, D} Statistic - p-value		`SumSales`	
Correlation	*Any measure*	"wrt measure" *Any measure*	Pears., Spear., Kendall	- {r,ρ,τ}, - p-v, df, 95%Conf.		`SumSales`	`cityPopulation`
Regression formula	*Any measure*	"wrt feature" *Any Feature*	SimpleRegr	- Intercept, Slope - R2, p-value, MSE		`SumSales`	`cityPopulation, month`
Top-k values	*Any measure*	*Any measure*	topK check	k	Top-k points, i=1..k	`SumSales`	
Trend, Seasonality	*Any measure*	"over" *Time-related Char.Type*	STL	- {season, trend, rem.} - #iterations		`SumSales`	`month`
Uni(bi) modality	*Any measure*	"over" *Time-related Char.Type*	DiffCheck, BoolCheck, …	Exists?	Point(s) of peak(s)	`SumSales`	`month`
Peer-domination	*Additive*	"over" *{Char.Type}**	Peer Dom Check	Exists?	Dominant Characters	`SumSales`	`city, month`
Total sum Mega-Contribution	*Additive*	"over" *{Char.Type}**	Mega Contrib	Exists?	Mega-contributor fact	`SumSales`	`city, month`
Surprise	*Any measure*	"over" *{Char.Type}**	3PastYears Extrapolator	Exists?	Surprising fact	`SumSales`	`city, month`

Figure 3: Examples of holistic highlights

Moreover, based on these results, we can have a *qualitative assessment* of the result – e.g., when $p - value < 0.05$ and $R^2 > 0.7$ we can call the hypothesis of a linear relationship as being verified. *Conceptually, since a result is the automated, data-based assessment for the existence of an Archetype Property in a data set, we can report the result of this assessment – equiv., the verification of whether the Archetype Property holds, and to what extent– via a predicate.* In fact, we can have two predicates, a Simple Qualitative Report and a Detailed Report:

- A *Simple Qualitative Report* is a generic predicate, common to all Result Types of the form: $ArchetypeProperty(Dataset, Measure, Explanators) : Boolean$, e.g., $SLR(Sales, SumSales, CityPop) \rightarrow True$
- A *Detailed Report* customizes the predicate according to the model of the algorithm and extends the list of generic variables with the technical results: $ArchetypeProperty(\ldots, Intercept, Slope, R^2, p, MSE) : Boolean$, e.g., $SLR(Sales, SumSales, CityPop, 756, 0.5, 0.87, 4e^{-6}, 900) \rightarrow True$

To support Results at the Metamodel level, **Result Types** come with (a) a static Simple Quantitative Report, common to all Result Types, (b) a *Model*, which internally holds all the essential attributes for the result quantification (e.g., an intercept and a slope for the SLR), (c) a set of *Auxiliary Metrics*, each with a name and a definition.

Finally, Holistic Highlights are assessed and labeled for their significance. The significance of a result can be evaluated via an *extensible* palette of (a) characteristics directly measurable from a query result, or, (b) session-level characteristics – for both cases, see e.g., the families of *Novelty*, *Relevance*, *Peculiarity* and *Surprise* [36, 37] as well as *Commonness* and *Exception* [5]. Any significance facet can be assessed either via an arithmetic score, or an enumerated label, but in any case, it should have an ordinal domain, such that different highlights can be compared to each other. To this end, we introduce (a) at the Metamodel level, **Score Types** having names with well known semantics and ordinal domains, and (b) at the Model level, **Scores**, such that the individual highlights can be annotated with $< ScoreType, ScoreValue >$ pairs.

A possible textual description of a Holistic Highlight is as follows:

> The $< ArchetypeProperty >$ for $< MainMeasure >$ in $< Dataset >$, tested via $< Algorithm\ Instantiaton >$ and supported by $\{Explanator\}\star$, results in $< Result >$ with $\{< HHScoreType >=< HHScore >\}\star$.

Last but not least, a Holistic Highlight can include a set of details in the form of *Elementary Highlights*, to be discussed in the sequel.

3.3. Elementary Highlights

3.3.1. Model level

In some cases, like for instance in the case of modality peaks, top-k values, or peer-domination, it is possible that there exist *Facts*, i.e., specific members of the Dataset, that play specific roles to facilitate the verification of the Archetype Property in the context of a Holistic Highlight. Depending on the Archetype Property, these can be its important details, without which any reporting is useless. For instance, an analyst needs to know at which time point the time series reaches a peak. In other cases, e.g., in the cases of a distribution check, or the correlation of two Measure Types, these details are not present.

An **Elementary Highlight** is a fact, determined by the combination of a set of contextualizing **Character**s demonstrating a behavior measured by a **Measure Value**, which plays an important role to the interpretation of a Holistic Highlight. The fundamental difference from Holistic Highlights is that instead of annotating the entire Dataset with an Archetype Property, Elementary Highlights refer to specific data points. Two examples (see also Fig. 4):

The **Character set**	*over character types*	with a **Measure value**	For measure type	serves as the **Elementary Highlight Role**	with **EH Score Type** *<formula>*	having **EH Score**
<April 20022>	Month	68	SumSales	1st peak of unimodality	*Peak Rank Significance htan(measureValue / avgValue)*	80%
<June 2022>	Month	62	SumSales	2nd peak of bimodality	*Peak Rank Significance htan(measureValue / avgValue)*	50%
<Athens, June 2022>	City, Month	54	SumSales	Top-2 point	*Peak Rank Significance htan(measureValue / avgValue)*	50%
<Athens>	City	54	SumSales	Peer-dominator over siblings	*Pct dominated peers (% peers with less value)*	100%
<June 2022>	Month	32	SumSales	Peer-dominator over siblings	*Pct dominated peers (% peers with less value)*	80%
<June 2022, Athens>	City, Month	12	SumSales	Mega-contributor of the total sum	*Pct of total sum (%total sum)*	25%
<Athens>	City	2100	SumSales	Mega-contributor of the marginal sum	*Pct of marginal sum (%marginal sum)*	75%
<April 2022, Rhodes>	City, Month	10	SumSales	Surprising fact	*DiffFromExpected Abs Diff Ratio(actual, extrapolated)*	28%

Figure 4: Examples of elementary highlights

- *Top-k.* Character $Athens$, with a Measure value of 2100 for $TotalSales$ is in the *top-k* Facts for a dataset. Its *rank*=1 is a score that denotes how high this particular fact is in the list of top-k facts.
- *Unimodality peak.* The time-series of $Athens$ has a unimodality *peak* for the Fact determined by the combination of Characters $city$=$Athens$ and $month$=2023-05 and a *Measure Value* of 1000 for Total Sales.

Scores are handled in a symmetrical manner to Holistic Highlights.
A possible textual description of an Elementary Highlight is as follows:

> The combination of characters $\{< Character >\}*$ with value $< Measure\ Value >$
> serves as $< Elementary\ Highlight\ Role >$
> with $< ScoreType >$ = $< ScoreValue >$.

3.3.2. Metamodel level

In full accordance with the Model Level, each Archetype Property comes with a set of characteristic Elementary Highlight Roles. An **Elementary Highlight Role** has a set of identifying **Character Role**s and a **Measure Role**, for the specific Elementary Highlights to materialize. Moreover, it carries **Score Types** with formulae for individual scores.

4. Discussion

Who are the beneficiaries of the model? A first contribution of the proposed conceptual model is that it clarifies both the concepts and the terminology for data storytelling, for every stakeholder

Our Model	Top-K Insights	Quick Insights	DataShot	Meta Insight	Calliope, Erato	Notable	Insight Pilot	VOOL	DAISY
Holistic highlight	Insight	Insight	Data fact	Basic data pattern, Meta-Insight	Data fact	Data fact	Insight	Insight	Insight
Archetype Property	Insight type (2)	Type (12)	Type (11)	Type (11)	Type (10)	Type (6)	Type		
Result			Parameter	Commonness & exceptions		Parameter			
Algorithm Exec.	Extractor (4)			Evaluate function				Module (2)	
Score	Significance, Impact	Score (impact * signific.)	Importance Score	Impact	Importance	Importance & intention score	Diversity	Interest, Coverage	Interestingness
Dataset	Subspace	Subspace	Context	Subspace	Subspace	Subspace	Subspace	Cube	Table
Main Measure		Measures	Measures	Measure	Measure	Measures	Measure		
Explanator	Dividing dimension	Breakdown	Breakdown	Breakdown	Breakdown	Dimension	Breakdown		
Elementary highlight				Highlight				Components (facts)	
Elementary H/L Role									
Characters			Focus		Focus	Focus	Property		
Measure value									
Score									

Figure 5: Coverage of highlights concepts in the literature. Gray font is used when concepts are proposed but are not part of the highlight. For each tool, we report the original concept names. For archetype properties and algorithms, we also indicate the number of implementations (when available).

involved. Concerning *data analysts*, the conceptual model allows the structuring of important parts of the problem in a way that is exploitable later: as soon as Archetype Properties, Characters and important Measure Values become part of a structured solution, the data analyst can think on the problem in terms of them (e.g., "What are the main Archetype Properties hiding in my data? Who are the main Characters in these data?"). Concerning *tool builders*, it is absolutely feasible to direct the automation of algorithm execution and result structuring along the concepts of our model; once this automated extraction and representation of highlights is achieved, their exploitation for storytelling purposes is straightforward (see Fig. 3 and 4).

How realistic is the proposed model? Fig. 3 and 4 report on a large number of typical data analysis algorithms. The fact that, for all these heterogeneous algorithms, there is a straightforward translation to a common, structured result, immediately exploitable for data storytelling purposes, testifies in favor of the model proposed in this paper.

Furthermore, as shown in Fig. 5, the Holistic and Elementary Highlights proposed in this paper cover the ones reported in the literature for tools automatically producing highlights. All of them deal with a dataset (typically a query result) and, excepting VOOL and DAISY, they distinguish at least one measure and a *breakdown* dimension (there are no other types of explanators). Many highlight types are proposed, and there is a set of underlying extraction algorithms, albeit not thoroughly modeled. Similarly, all tools score highlights, but scores are frequently not part of the highlight. Highlight models are typically not modeled, or, mostly limited to one parameter. Finally, although many tools distinguish a subset of breakdown values as *focus*, or a set of important facts, which correspond to Elementary Highlights, no work provides details on them.

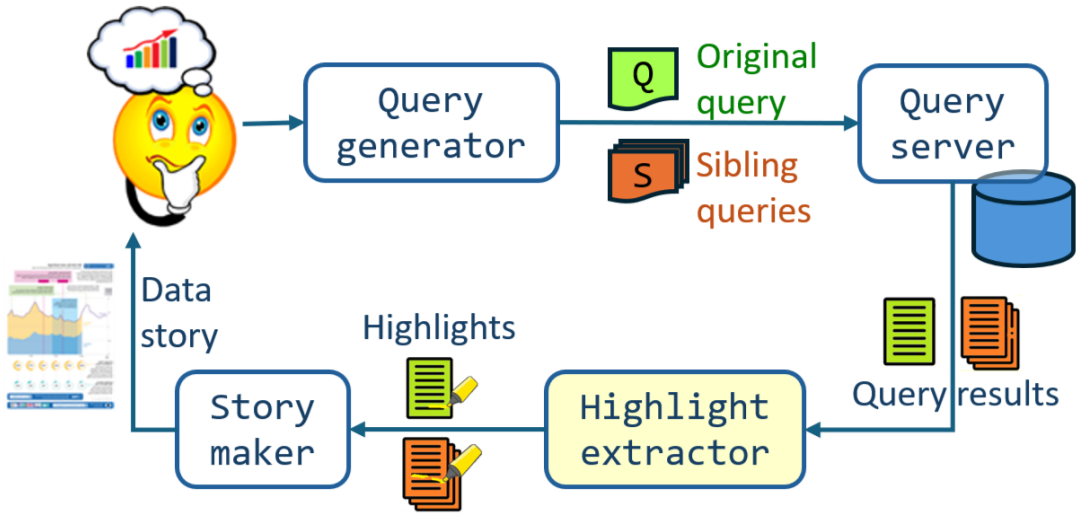

Figure 6: Data story generation, with emphasis to the highlight extraction part

5. Application of the model

We have implemented our model in two tools: The first implementation involves a data profiler tool[1], that automatically profiles the columns of submitted datasets for their descriptive statistics, histograms, correlations, decision trees, outliers and dominance patterns, along with the respective highlights. The second implementation involves the extension of a Business Intelligence system[2], with a new subsystem that answers time series queries with a data story based on highlights. In this section we describe the second tool, as an illustration of the application of our model.

The analyst is allowed to define a time-related query as a chart that needs to be constructed, and the system responds by (a) obtaining the result of the query, (b) enriching its result with (b1) results of auxiliary queries and (b2) highlights from the highlight extraction algorithms, applied over the query results, and, (c) ultimately, returning all the interesting findings wrapped as a data story composed of text and graphical representations.

The main steps of the process (see Figure 6) are the following:

1. Using the Query generator GUI, the analyst specifies the intended query Q as a chart, which can be a bar-chart, line-chart or scatter-plot. This means that practically, the query will have two grouper attributes (dimension levels) that will be visualized in the two axes of the chart, and, consequently an aggregated measure. Filters are also parts of the specification. The Query generator module automatically generates a set of auxiliary queries S in order to contextualize and assess the results of the original query (see next).
2. All the queries are executed by the Query server module, over the underlying data set, and their results are produced.
3. The results can be immediately visualized, but most importantly, they are passed through a set of *Highlight extractors*. This is the part that is mostly related to our deliberations in this paper. Highlights include the existence of trend, unimodality and bimodality in time-series query results (where one grouper is time), the existence of a strong linearity in the results as demonstrated by a strong linear regression score, and others.
4. A Story-maker module receives query results and highlights, compares and combines them as a data story with automatically generated charts and text (the explanation of this part falls outside the scope of this paper).

[1]Available at https://github.com/DAINTINESS-Group/Pythia
[2]Available at https://github.com/DAINTINESS-Group/DelianCubeEngine

Queries. We consider several query classes over cubes in our setup. The most relevant query class is *Time-series queries*. For this query class, the first grouper (obligatorily) concerns a level of a time-related dimension (e.g., month, year, decade). This allows the result to be a timeseries of aggregate measurements for the second grouper (and, thus visualized as a line-chart, with time in the horizontal axis). The relationship of groupers with filters can be arbitrary. The general form of the query is

$$q = \gamma^{agg(M)}_{T.L1,D2.L2}(\sigma_{\phi}(C)), \phi : \bigwedge D_i.L_i = v_i$$

where C is a cube name, σ_{ϕ} is the selection operator applying the conjunctive selection condition ϕ to C, D_i refers to a cube dimension, L_i to a dimension level, v_i to a value in $dom(D_i.L_i)$, M is a measure, agg is the aggregate function applied to it, and γ is the group-by operator with the two grouper levels as subscripts (the first being time-related) and the aggregated measure as superscript.

A special case of time-series queries occurs when the second grouper has been pinned to a single value, i.e., $\phi : D_2.L_2 = v$. In other words, the first grouper concerns time, and the second grouper includes a filter at the same level as the grouper, resulting in a single timeseries as the query result.

The construction of *sibling queries* exploits the dimension hierarchies of the involved cubes to contrast the result of the original query to the results of "peer" queries. To avoid long formalities, here, we restrict ourselves to an indicative example and refer the interested reader to [2] for a rigorous definition and discussion. In the example of the introduction, the selection condition of the original query restricts the country to $Greece$ and the time to $2023 - Qtr2$. We assume that all such values pertain to dimension hierarchies, creating trees of values with ancestor and child values: e.g., the mother of $Greece$ (at the $Country$ level) is $Europe$ (at the $Continent$ level) and the mother of $2023 - Qtr2$ is 2023. To generate siblings, we need to construct two auxiliary queries, one per filter: the first is based on contrasting Greece to all its siblings, i.e., $children(mother(Greece))$ = $children(Europe)$ at the country level (all European countries), whereas the second contrasts $2023 - Qtr2$ to its siblings, i.e., all the quarters of 2023. Both (a) the results, and, (b) the highlights of sibling queries can be later contrasted for similarities and exceptions during the story making process.

Highlights. We test query results for archetype properties via a set of highlight extraction algorithms, applied to both the original and the auxiliary queries. We have implemented the following highlight types, along with their respective extraction algorithms, albeit specifically tailored for aforementioned classes of time-series queries:

- *AbsoluteTrend*, to detect whether a time-series is absolutely monotonically increasing (uptrend) or decreasing (downtrend) or none of them.
- *KendallBased*, to detect whether a timeseries is monotonically increasing (uptrend) or decreasing (downtrend) based on the Kendall tau coefficient.
- *Contributor*, to detect whether a timeseries has a value at the x-axis that contributes more than 50% *(Mega contributor)* in the produced results.
- *Modality*, to detect whether a Unimodality or Bimodality shape governs the timeseries. A timeseries is considered unimodal when it forms a U-shaped valley or peak shape. A timeseries is considered bimodal when it has two distinct peaks or modes in its distribution
- *Regression*, to perform a Simple Linear Regression.

Score. To determine the importance of results of the model we introduce a score function for each algorithm producing a score in the range [-1,1]:

- *AbsoluteTrend score*: the score equals to -1 if there is an absolute downtrend, 0 if there is no absolute trend, and 1 if there is an absolute uptrend.
- *KendallBased score*: $score = abs(\tau)$ where τ is Kendall's coefficient.
- *Contributor score*: The score is the ratio of the maximum measure of a grouper's values over the sum of measures across all grouper instances.
- *Modality score*: The following steps determine the modality score for the time series:

  - Divide the series into segments from the start to each point where the sign of $y_i - y_{i-1}$ changes.
  - For each segment, we compute its score as $score = \frac{|maxValue|-|minValue|}{|maxValue|}$ (actually with an offset to avoid zero divisions).
  - Finally, we compute the score for the timeseries by calculating the average of the scores from all segments: $score = \frac{1}{N-1} \sum_0^{N-1} score(segment)$, where N the number of segments.
- *Regression score*: The formula $score = 1 - |tanh(MSE)|$ (normalized Mean Square Error via hyperbolic tangent) of the Simple Linear Regression.

Score-based pruning. To alleviate the user from the burden of going through all the model results, we need to identify *significant highlights* with a significant score, i.e., above a pre-determined threshold. As the score range is always between [-1,1], and we do recognize that some negative scores are important (for example downtrend), we employ the absolute value of every score as the criterion of its significance. In all our deliberations, we employ a threshold $\theta = 0.5$ for the absolute score, above which a model is considered to be an important highlight.

As a side note, with respect to the characterizations of interestingness along different dimensions, all the aforementioned scores are related to *peculiarity*: the more intensely present a property is, the more extreme the score the data receive. Thus, subsets of the data space stand out with a higher score than the rest of the data, because of their high adherence to the archetype property (correlation, modality, mega-contributor presence, etc).

In summary. An observant reader will have already underscored how all these inherently different algorithms (regression, modalities, contributors) are all modeled within the same tool and with the same highlights model. The handling of this heterogeneity via a single, common, and, as this section shows, feasible model, is one of the main contributions of this paper.

6. Conclusions

In this paper, we have presented a model that facilitates (a) the modeling, (b) the clarification of terminology, (c) the automation of the production of *highlights*, i.e., *structured* characterizations of subsets of the data space that are worth reporting due to their support of archetype statistical properties of interest, demonstrated as *phenomena*. We have introduced the most important entities of the domain of highlight extraction and discussed their inter-relationships. *Holistic highlights* are global properties that pertain to an entire set of facts, whereas *elementary highlights* are their constituents that identify facts that play a particular role for the holistic highlight to take place. We have also demonstrated that (a) frequently encountered archetype properties are nicely covered by our modeling and (b) the highlight structure facilitates narratives straightforwardly.

The evaluation of highlight interestingness in a fully automated way, such that we can rank and prune highlights in an even more precise and context-aware fashion is an open research issue. Structuring data stories in an efficient way, by taking advantage of the complementarity, overlap, discrepancy, or other properties of a set of highlights is another open research issue.

Declaration on Generative AI

The author(s) have not employed any Generative AI tools.

References

[1] A. Meliou, A. Abouzied, P. J. Haas, R. R. Haque, A. L. Mai, V. Vittis, Data management perspectives on prescriptive analytics (invited talk), in: ICDT, 2025.

[2] D. Gkesoulis, P. Vassiliadis, P. Manousis, Cinecubes: Aiding data workers gain insights from OLAP queries, Inf. Syst. 53 (2015) 60–86.

[3] B. Tang, S. Han, M. L. Yiu, R. Ding, D. Zhang, Extracting top-k insights from multi-dimensional data, in: SIGMOD, 2017.

[4] Y. Wang, Z. Sun, H. Zhang, W. Cui, K. Xu, X. Ma, D. Zhang, Datashot: Automatic generation of fact sheets from tabular data, IEEE Trans. Vis. Comput. Graph. 26 (2020) 895–905.

[5] P. Ma, R. Ding, S. Han, D. Zhang, Metainsight: Automatic discovery of structured knowledge for exploratory data analysis, in: SIGMOD, 2021.

[6] D. Shi, X. Xu, F. Sun, Y. Shi, N. Cao, Calliope: Automatic visual data story generation from a spreadsheet, IEEE Trans. Vis. Comput. Graph. 27 (2021) 453–463.

[7] P. Ma, R. Ding, S. Wang, S. Han, D. Zhang, Insightpilot: An llm-empowered automated data exploration system, in: EMNLP'2023, 2023.

[8] S. Idreos, O. Papaemmanouil, S. Chaudhuri, Overview of data exploration techniques, in: SIGMOD, 2015.

[9] O. B. El, T. Milo, A. Somech, ATENA: an autonomous system for data exploration based on deep reinforcement learning, in: CIKM, 2019.

[10] A. Personnaz, S. Amer-Yahia, L. Berti-Équille, M. Fabricius, S. Subramanian, DORA THE EXPLORER: exploring very large data with interactive deep reinforcement learning, in: CIKM, 2021.

[11] T. De Bie, L. D. Raedt, J. Hernández-Orallo, H. H. Hoos, P. Smyth, C. K. I. Williams, Automating data science, Commun. ACM 65 (2022) 76–87.

[12] P. Marcel, V. Peralta, S. Amer-Yahia, Data narration for the people: Challenges and opportunities, in: EDBT, 2023.

[13] S. Amer-Yahia, Intelligent agents for data exploration, VLDB Endow. 17 (2024) 4521–4530.

[14] T. Lipman, T. Milo, A. Somech, T. Wolfson, O. Zafar, LINX: A language driven generative system for goal-oriented automated data exploration, in: EDBT, 2025.

[15] O. B. El, T. Milo, A. Somech, Automatically generating data exploration sessions using deep reinforcement learning, in: SIGMOD, 2020.

[16] R. Ding, S. Han, Y. Xu, H. Zhang, D. Zhang, Quickinsights: Quick and automatic discovery of insights from multi-dimensional data, in: SIGMOD, 2019.

[17] S. Sarawagi, R. Agrawal, N. Megiddo, Discovery-driven exploration of OLAP data cubes, in: EDBT, 1998.

[18] M. Sun, L. Cai, W. Cui, Y. Wu, Y. Shi, N. Cao, Erato: Cooperative data story editing via fact interpolation, IEEE Trans. Vis. Comput. Graph. 29 (2023) 983–993.

[19] H. Li, L. Ying, H. Zhang, Y. Wu, H. Qu, Y. Wang, Notable: On-the-fly assistant for data storytelling in computational notebooks, in: CHI, 2023.

[20] F. E. Outa, M. Francia, P. Marcel, V. Peralta, P. Vassiliadis, Towards a conceptual model for data narratives, in: ER, 2020.

[21] S. Chen, J. Li, G. L. Andrienko, N. V. Andrienko, Y. Wang, P. H. Nguyen, C. Turkay, Supporting story synthesis: Bridging the gap between visual analytics and storytelling, IEEE Trans. Vis. Comput. Graph. 26 (2020) 2499–2516.

[22] M. Francia, E. Gallinucci, M. Golfarelli, S. Rizzi, VOOL: A modular insight-based framework for vocalizing OLAP sessions, Inf. Syst. 129 (2025) 102496.

[23] J. Xing, X. Wang, H. V. Jagadish, Data-driven insight synthesis for multi-dimensional data, VLDB Endow. 17 (2024) 1007–1019.

[24] S. Sarawagi, User-adaptive exploration of multidimensional data, in: VLDB, 2000.

[25] R. Agrawal, R. Srikant, Fast algorithms for mining association rules in large databases, in: VLDB, 1994.

[26] P. Marcel, V. Peralta, P. Vassiliadis, A framework for learning cell interestingness from cube explorations, in: ADBIS, 2019.
[27] L. Geng, H. J. Hamilton, Interestingness measures for data mining: A survey, ACM Comput. Surv. (2006).
[28] T. D. Bie, Subjective interestingness in exploratory data mining, in: IDA, 2013.
[29] T. Milo, A. Somech, Automating exploratory data analysis via machine learning: An overview, in: SIGMOD, 2020.
[30] Y. Patil, S. Amer-Yahia, S. Subramanian, Designing the evaluation of operator-enabled interactive data exploration in VALIDE, in: HILDA, 2022.
[31] H. Guo, S. R. Gomez, C. Ziemkiewicz, D. H. Laidlaw, A case study using visualization interaction logs and insight metrics to understand how analysts arrive at insights, IEEE Trans. Vis. Comput. Graph. (2016).
[32] E. Zgraggen, Z. Zhao, R. C. Zeleznik, T. Kraska, Investigating the effect of the multiple comparisons problem in visual analysis, in: CHI, 2018.
[33] M. Joglekar, H. Garcia-Molina, A. G. Parameswaran, Interactive data exploration with smart drill-down, IEEE Trans. Knowl. Data Eng. (2019).
[34] A. Giuzio, G. Mecca, E. Quintarelli, M. Roveri, D. Santoro, L. Tanca, INDIANA: an interactive system for assisting database exploration, Inf. Syst. (2019).
[35] D. Deutch, A. Gilad, T. Milo, A. Somech, Explained: Explanations for EDA notebooks, Proc. VLDB Endow. (2020).
[36] D. Gkitsakis, S. Kaloudis, E. Mouselli, V. Peralta, P. Marcel, P. Vassiliadis, Assessment methods for the interestingness of cube queries, in: DOLAP, 2023.
[37] D. Gkitsakis, S. Kaloudis, E. Mouselli, V. Peralta, P. Marcel, P. Vassiliadis, Cube query interestingness: Novelty, relevance, peculiarity and surprise, Inf. Syst. 123 (2024) 102381. URL: https://doi.org/10.1016/j.is.2024.102381. doi:10.1016/J.IS.2024.102381.