A Framework for Generating Summaries from Temporal Personal Health Data

JONATHAN J. HARRIS, Rensselaer Polytechnic Institute
CHING-HUA CHEN, IBM Research
MOHAMMED J. ZAKI, Rensselaer Polytechnic Institute

21

Although it has become easier for individuals to track their personal health data (e.g., heart rate, step count, and nutrient intake data), there is still a wide chasm between the collection of data and the generation of meaningful summaries to help users better understand what their data means to them. With an increased comprehension of their data, users will be able to act upon the newfound information and work toward striving closer to their health goals. We aim to bridge the gap between data collection and summary generation by mining the data for interesting behavioral findings that may provide hints about a user's tendencies. Our focus is on improving the explainability of temporal personal health data via a set of informative summary templates, or "protoforms." These protoforms span both evaluation-based summaries that help users evaluate their health goals and pattern-based summaries that explain their implicit behaviors. In addition to individual-level summaries, the protoforms we use are also designed for population-level summaries. We apply our approach to generate summaries (both univariate and multivariate) from real user health data and show that the summaries our system generates are both interesting and useful.

CCS Concepts: • **Information systems → Summarization**; *Data analytics*; *Personalization*; Recommender systems; • **Applied computing →** *Consumer health*; Health care information systems; Health informatics;

Additional Key Words and Phrases: Linguistic data summarization, time-series analysis, sequence mining, natural language summaries, protoforms, personal health data

ACM Reference format:
Jonathan J. Harris, Ching-Hua Chen, and Mohammed J. Zaki. 2021. A Framework for Generating Summaries from Temporal Personal Health Data. *ACM Trans. Comput. Healthcare* 2, 3, Article 21 (July 2021), 43 pages.
https://doi.org/10.1145/3448672

1 INTRODUCTION

Smartphone apps and personal fitness devices have made it increasingly easy for users to collect and monitor their personal health data. Whereas some of this data requires active entry by the user (e.g., dietary behaviors), other types of data are passively and continuously collected (e.g., physical activity, heart rate, and location). The

Authors' addresses: J. J. Harris and M. J. Zaki (corresponding author), Rensselaer Polytechnic Institute, 110 8th St, Troy, NY 12180; emails: harrij15@rpi.edu, zaki@cs.rpi.edu; C.-H. Chen, IBM Thomas J Watson Research Center, 1101 Kitchawan Rd, Yorktown Heights, NY 10598; email: chinghua@us.ibm.com.

2637-8051/2021/07-ART21 $15.00
https://doi.org/10.1145/3448672

increased ease of data collection in the personal health domain has inspired the quantified-self movement, where motivated individuals record almost every aspect of their lives, including mental and physical health. Likewise, users with chronic conditions regularly use their own health information for health decision making [40], and ineffective interpretation of one's data may adversely affect how they take their medications, what they eat, how they exercise, and even how they socialize [34].

However, there are people who fit neither of the aforementioned groups and simply wish to live a healthier lifestyle. For such people, it is widely reported that fitness devices and health apps experience a high abandonment rate. Although there are many reported reasons for this, technology-related reasons include the lack of desired features such as *notifications* or *decision support*. Furthermore, if the user-perceived value of the data is low, this can create a feedback loop where such a perception increases the chance of erroneous or sparse data being recorded, which in turn lowers the utility of the data and leads to further user disengagement [10]. A key challenge for most users is often the lack of meaningful interpretation of the health data [9].

This concept can also be applied to personal health data. For those who may wish to improve or maintain their health, it is important for them to gain more insight into their own health logs to help them reach their personal health goals, and to assess whether their efforts are bringing them closer to those goals. However, individuals often find it challenging to understand their own health data, especially when they record multiple types of data over a long period of time. For example, in the quantified-self community, structured recording of daily activities and outcomes is practiced regularly. A key hurdle for this community is the extraction of high-level information from the sea of data and to interpret that information in a meaningful way [9]. This hurdle is also commonly reported among patients living with chronic conditions [34], who use data for daily decision making on medication dosages, food intake, and other behaviors. Ineffective interpretation of one's data may affect the subsequent decision-making process and anticipated health outcomes. The high frequency of data usage by those populations makes it impractical to rely solely on medical professionals to interpret their data. Automated methods to support data interpretation is therefore an urgent need.

Today, a common approach to obtaining expert-generated information on improving or maintaining health is through a search query online or through contact with a health expert. Although searching for health information on the Internet works for the general case, it often lacks the personalization required to accommodate individual needs. In particular, every person has a different health experience, as exemplified by the uniqueness of the data collected by the person's health apps; human health experts may be able to relate an individual's data to general health knowledge, but they are expensive to engage with and there are not enough of them. Therefore, health consumers are often left on their own to bridge the gap between the sea of general health knowledge and the sea of personal health data. Addressing this gap via automation requires a combination of methods for anticipating and understanding an individual's needs, providing an answer or recommendation for meeting that need, and, importantly, providing an explanation for that recommendation. Black box approaches that generate recommendations from data without explanation may be acceptable in some domains (e.g., manufacturing, advertising); however, this is rarely the case when it comes to personal health and healthcare. We believe that an important aspect of data-driven recommendation involves explaining how the data itself is being interpreted, and how it can be used to support explanations of downstream algorithms to produce a recommendation.

The main motivation of our work is the need of individuals (who wish to improve their health) to better understand their past behaviors based on their personal health data that may be inhibiting them from reaching their health goals. With additional comprehension via a natural language summary and a refined focus on key aspects of their health data, they will have the ability to take action by making appropriate changes to their routine. We address this problem by creating a framework that provides individuals with *personalized natural language summaries* based on behavioral patterns found within their time-series data. Generating explainable summaries from personal health data is a challenging task. Within the field of summarization, there are three main approaches when it comes to linguistic summary generation: probabilistic/statistical, neural, and rule-based methods [43]. Whereas state-of-the-art probabilistic/statistical and neural methods generate the sentences automatically, the textual output of these approaches is of lower quality than that of the rule-based approach [43]. Our work, therefore, utilizes a rule-based approach inspired by the principles presented in the work of Zadeh [49–51] and the database summarization approach in the work of Kacprzyk et al. [23]. Most existing methods within the linguistic summarization community do not handle time-series data; the few approaches that do either generate longer narratives of the data [15] or summaries of simpler trends [22], such as whether the trend is increasing or concave. Our unsupervised approach takes advantage of a more comprehensive set of summaries along with the use of time-series data mining methods to generate more meaningful summaries.

Our work focuses on improving the explainability of personal health data by generating temporal summaries in natural language from time-series health data. We propose a comprehensive framework to generate summaries that can help users evaluate their personal health data, and compare their data against general health guidelines or goals. In particular, we propose a systematic classification of summary types that cover a wide range of applications in personal health, including evaluation-based summaries that help users evaluate their health goals, and pattern-based summaries that explain their "hidden" behaviors. Our approach extracts temporal patterns from data and generates clear and concise summaries. In particular, our summaries are based on a categorical (or symbolic) representation of time-series data (via **Symbolic Aggregate approXimation (SAX)** [28]), combined with frequent sequence pattern mining (via SPADE [52]) and categorical clustering (via Squeezer [18]), allowing us to generate understandable descriptions of hidden and implicit trends (both within and across multiple time series) that are not obvious from the raw data. To generate these summaries, we employ the linguistic summarization approach [48, 51] that relies on our proposed time-series *protoforms* to describe comprehensible natural language findings in personal health data. A protoform is a sentence prototype or template with placeholders or blanks that are automatically chosen to reflect trends and patterns supported by the data. For example, consider the protoform *On ⟨quantifier⟩ ⟨sub-time window⟩ in the past ⟨time window⟩, your ⟨attribute⟩ was ⟨summarizer⟩*, where the blanks (represented as ⟨*blank*⟩) are of different types and must satisfy different constraints. An example summary generated from this protoform is *On most of the days in the past week, your step count was high.* Here, the ⟨quantifier⟩ is "most of the." Each summary explains a particular pattern for an attribute (or a set of attributes) to help users make sense of their own data.

It is important to note that our work significantly extends existing approaches for both linguistic summarization and time-series data mining. For example, whereas existing time-series data mining methods extract patterns and trends, we extend them by generating understandable summaries. Likewise, existing linguistic summarization approaches primarily focus on tabular and non-temporal data, whereas our approach is especially focused on time-series data. Furthermore, our framework extends current time-series summarization approaches [20–23, 45] both in terms of diversity of summaries that can be produced and what patterns can be found within the data via mining. With the generation of natural language summaries from both univariate and multivariate temporal personal health data, our system can provide important clues to a better understanding of users' general behavior and can facilitate actionable changes to fix areas where they may be falling short of their health goals. To summarize, our work makes the following significant contributions:

- We propose a comprehensive framework of informative time-series protoforms to produce both evaluation-based and pattern-based summaries using time-series data mining methods. In particular, we provide a systemic classification of summary types to be applied to the temporal personal health domain.
- We generate meaningful natural language summaries from both univariate and multivariate time-series data to highlight hidden patterns found within and between multiple variables. Our approach also illustrates summary provenance via charts highlighting the appropriate data and/or pattern that support the summaries. A preliminary user evaluation confirms the usefulness and comprehensibility of our summaries.

- We showcase and evaluate the usefulness of the summaries on real user data including food nutrient logs (using the MyFitnessPal dataset [44]) and fitness data (using the Insight4Wear [35] dataset). We highlight interesting summaries obtained from users' nutrient intake, heart rate, and step count data.
- We show that our framework is general and can be applied to different domains in addition to personal health. We illustrate the generalizability by showing summaries from weather and stock market data.

2 RELATED WORKS

2.1 Data-to-Text Generation

In general, data-to-text generation methods include statistical [26] and neural [25] machine translation, and rule-based linguistic summarization [6]. Neural and statistical methods rely on the use of automation to produce natural language summaries, relying on measures like the BLEU score [33] to evaluate these summaries (the BLEU score measures the correspondence between the algorithm output and the reference human sentences). Rule-based methods typically use semantically meaningful templates or protoforms to generate their output and thus obviate the need for measures like the BLEU score. Instead, they rely on human evaluation to judge this utility, but they can use objective measures such as significance, frequency, and other metrics [6] to judge the quality of the summary output. Van der Lee et al. [43] compared the performance and text quality between rule-based, neural, and statistical methods. Their main and important conclusion is that rule-based methods generally perform faster and produce higher text quality, although the manual creation of the sentence prototypes is time-intensive. Furthermore, they observe that rule-based methods are generally restricted to more simple situations and may be less useful in more complex cases. However, statistical- and learning-based methods avoid manual creation of prototypes but are generally lacking in performance and text quality. It is important to note a significant drawback of the supervised neural and statistical text generation methods: they typically need a large set of training pairs containing the input data and the desired natural language summary, which is often not available. Given that text quality is extremely important within the personal health domain, and given the lack of training data comprising input raw time-series data and their corresponding natural language health summaries, we follow the rule-based linguistic paradigm.

2.2 Time-Series Summarization

Our work builds upon and extends linguistic database summarization methods [20–23, 45] that rely on the concept of protoforms and fuzzy logic [6, 51] to summarize data. Linguistic summarization methods have also been applied to time-series data in various domains, such as elderly care [46], physical activity tracking [37], driving simulation environments [13], deforestation analysis [11], human gait study [1], periodicity detection [31], time-series forecasting [24], and generation of longer temporal "narratives" from neonatal intensive care data via the use of a neonatal ontology [15]. Other work includes the use of genetic algorithms [7] to generate linguistic summaries from time series, and those that place emphasis on simple trends (e.g., increasing, concave) [22]. In contrast to these works, we propose a more comprehensive set of summaries, and unlike all previous time-series summary-based works, we also apply data mining to discover interesting patterns across multiple variables to produce more interesting summaries.[1]

More recent work on time-series summaries includes that of Murakami et al. [32], who use a neural encoder-decoder model to generate natural language summaries in the financial domain, and that of Aoki et al. [2], who extend the model to multiple external factors (e.g., relationships between the Nikkei and Dow Jones stock market data). For training, Murakami et al. [32] pair a time series with a market comment that aligns with it. They used 16,276 headlines gathered from Nikkei Quick News to train their model. Aoki et al. [2], in addition, use 5-minute

[1]While there is work on summaries spanning multiple variables in the context of neonatal intensive care data [15], that work is based on a neonatal ontology and does not perform multivariate temporal pattern mining as in our work.

charts of seven stock market indices from Thomson Reuters DataScope Select[2] as an external resource; however, summaries generated by both of these methods are limited to relatively simple conclusions (e.g., a continual rising trend). As such, neural network–based methods suffer from several drawbacks, such as the aforementioned lack of high-quality text [43], dependence on large training data and/or supervision (which is not available for our personal health domain), and lack of ability to explain patterns directly from raw temporal data. In contrast, our system is unsupervised and generates summaries that explain interesting patterns and trends that are not immediately apparent (based on pattern mining and clustering).

The closest work to ours is that of Guimarães and Ultsch [17], in which the authors generate linguistic descriptions of multivariate data via feature extraction, primitive pattern extraction via neural networks, and rule generation [42]. In their work, they use unsupervised neural networks to find primitive patterns of events in time series with natural language names assigned to the primitive patterns in a semi-automated manner. An example summary of an event from their work is the following: ***If*** "no airflow without snoring" ***is more or less simultaneous*** "no chest and abdomen wall movements without snoring," where phrases (e.g., "no") resulting from fuzzy-membership functions are paired with measured attributes (e.g., "airflow", "chest wall movements"). In our work, we mine frequent sequences to generate more interesting if-then pattern summaries, as well as cluster-based pattern summaries. Our framework is also able to provide many more informative summaries based on the comprehensive set of univariate and multivariate protoforms.

2.3 Time-Series Data Mining

There are many works on time-series data mining, reviewed by Batyrshin and Sheremetov [5], including the construction of rules based on patterns found in the data [12], using derivatives to describe the concavity/convexity of trends [8], identification of pre-determined patterns using shape descriptors [4], transformation of time series into state intervals to create association rules [19], generating reports about stocks [36], and so on. It is important to note that these approaches find temporal patterns or rules based on shapes and trends, but they do not generate explanations. In contrast, our work tries to explain the important patterns via temporal natural language summaries.

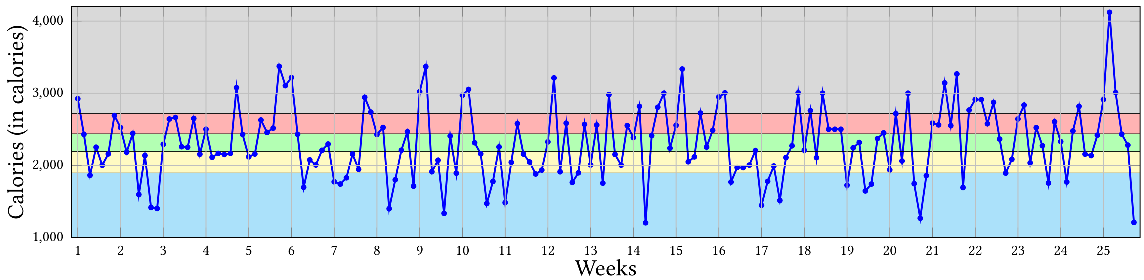
Our work utilizes the SAX [28] approach to discretize time-series data into a symbolic sequence from which we can mine patterns and trends. SAX is very effective for motif discovery, dimensionality reduction, and other data mining tasks on time-series data. Other related work includes Symbolic Fourier Approximation [38], which is more data adaptive; ABBA [14] that aims to better preserve shapes via adaptive polygonal chain approximation and mean-based clustering; and cSAX [27], which incorporates complexity invariance within SAX. In the future, we plan to explore these alternative methods in terms of their effect on the mined patterns for summary generation.

3 TEMPORAL SUMMARIES FOR PERSONAL HEALTH DATA

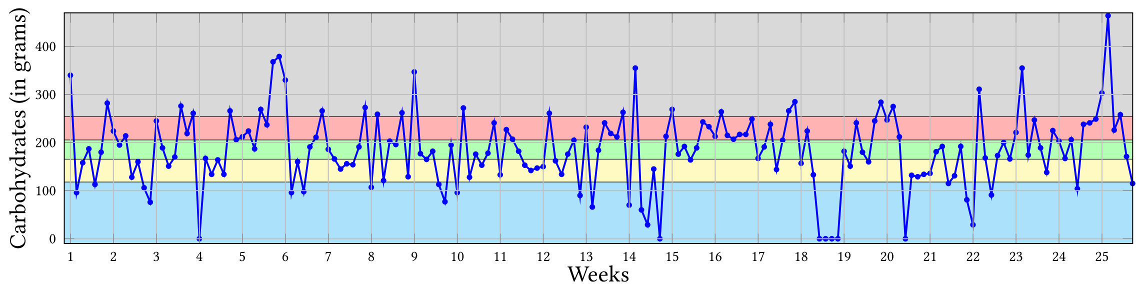
We begin by defining the basic concepts we will use in the remainder of this article:

- ***Protoform* (*P*):** A sentence prototype (or template) that can be used to generate a natural language summary.
- ***Summarizer* (*S*):** A conclusive phrase for a summary.
- ***Quantifier* (*Q*):** A word or phrase that specifies how often the summarizer S is true.
- ***Attribute* (*A*):** A variable of interest.
- ***Time window* (*TW*):** A time window of interest.
- ***Sub-time window* (*sTW*):** A time window at a smaller granularity than TW.
- ***Qualifier* (*R*):** A word or phrase that adds more specificity to a summary.

[2] https://hosted.datascope.reuters.com/DataScope/.



(a) Calorie Intake Data



(b) Carbohydrate Intake Data

Fig. 1. Calorie (a) and carbohydrate (b) intake data for a user from the MyFitnessPal dataset [44]. The different colored regions correspond to the different summarizers—very low, low, moderate, high, very high—from bottom to top.

Given a set of quantifiers $\boldsymbol{Q}$, a set of summarizers $\boldsymbol{S}$, a specified time window granularity TW and sub-time window granularity sTW, a set of protoforms $\boldsymbol{P}$, and a set of time series $\boldsymbol{T}$ for a corresponding set of attributes $\boldsymbol{A}$, we generate natural language summaries of behavioral patterns found in temporal personal health data. For example, consider the summary *On most of the days in the past week, your calorie intake was high*, generated from the protoform *On Q sTW in the past TW, your A was S*, the quantifier Q (*most of the*) represents how often the finding is found to be true in the data, the attribute A (*calorie intake*) represents the variable of interest, and the summarizer S (*high*) represents the conclusion from the data. Here, the time window TW is *weeks* and the sub-time window sTW is *days*. We use different protoforms to generate more complex summaries describing interesting patterns within and across variables.

Running example. To instantiate summaries that illustrate each of our protoforms, we will consider real user data from the MyFitnessPal dataset [44], particularly data on their intake of calories and carbohydrates for a period of about 6 months (actually 174 days). The corresponding time-series data is plotted in Figure 1. The five horizontal ranges within the charts each correspond to the range of values for each summarizer in $\boldsymbol{S}$. For example, any data point within the top-most (gray) range can be described as "very high." These regions are found using SAX [28], which will be explained later in Section 4. Assume that the user is interested in finding patterns in a weekly time window and has the goal to limit his or her calorie and carbohydrate intake for a *2,000-calorie* diet. Table 1 shows some of the possible quantifiers ($\boldsymbol{Q}$), and also the attributes of interest (i.e., Calories, Carbohydrates) and the (sub-) time window values (i.e., day, week) for our running example. For the remainder of this article, we use the data in Figure 1 as a running example to explain the various protoforms.

3.1 Protoform Hierarchy

We seek to automatically generate a diverse set of summaries of time-series data related to a user's personal health. It is important that we have a diverse set to allow us to take into account the various ways we can look at

Table 1. Variable Assignments

Q	none of the, almost none of the, some of the, half of the, more than half of the, most of the, all of the
A	calorie intake, carbohydrate intake
TW	weekly granularity
sTW	daily granularity

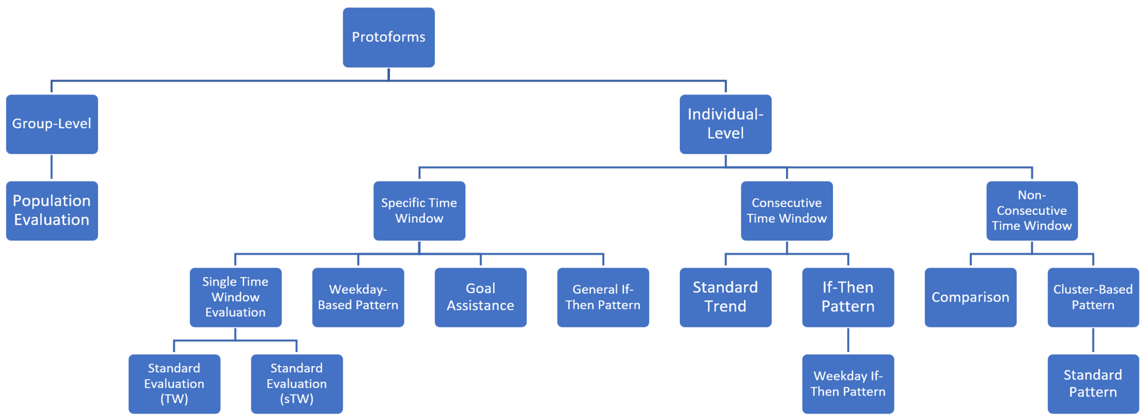


Fig. 2. Hierarchy of protoform or summary types. Protoforms can be group-level or individual-level. Within the individual level, we propose three types of summaries: (1) based on a specific time window, (2) comparing two consecutive time windows, and (3) comparing any two time periods. Each of these is further divided into more specific summary types.

a user's data. Each kind of protoform our system generates highlights a different aspect from the user's data that may be helping or hurting their efforts to reach their personal health goals. As shown in Figure 2, we propose a number of different summary types that are applicable to a wide range of personal health scenarios, and are meant to be both useful and comprehensive. In particular, we propose three types of individual-level summaries: (1) specific time window summaries, which look at trends within a specified time window; (2) consecutive time window–based summaries that compare two successive time periods; and (3) non-consecutive time window–based summaries that compare different time periods. We also propose a group-level summary that is designed to see the patterns in a population of users. In addition, these summaries can be augmented with goals or guidelines to better help the user. These protoforms are equally applicable to quantified-selfers or general users who want to understand their personal data. When users look at their own data, they may try to look for patterns in the data that correlate with their daily routine. These patterns reflect their behaviors and can provide clues as to what aids or hampers progress toward their health goals.

In the following sections, we will explore various protoform types where each protoform $p \in \boldsymbol{P}$ has a corresponding set of summarizers $\boldsymbol{S}$. Each summary template or protoform requires a set of quantifiers and a unique set of summarizers as appropriate placeholders. Table 2 enumerates the different types of summarizers ($\boldsymbol{S}$), whereas Table 1 shows the quantifiers ($\boldsymbol{Q}$). For each summary type, we will provide univariate and multivariate examples using the data from the running example in Figure 1, using both calorie and carbohydrate intake as input variables. In addition to showing the natural language summaries, for better explainability, we display the provenance of the data supporting the summaries generated using the various protoform types. In particular, we automatically generate corresponding time-series charts showing where the discovered patterns were found

Table 2. Summarizers by Protoform Type

Protoform Type	Possible Summarizers
`Standard Evaluation`	very low, low, moderate, high, very high
`Standard Evaluation (w/ goal)`	reached, did not reach
`Goal Assistance`	increase, decrease
`Day-Based Pattern`	very low, low, moderate, high, very high
`Standard Trend`	increased, decreased, stayed the same
`If-Then Pattern`	very low, low, moderate, high, very high
`Comparison`	higher, lower, about the same
`Comparison (w/ goal)`	better, not do as well, about the same
`Cluster-Based Pattern`	rose, dropped, stayed the same

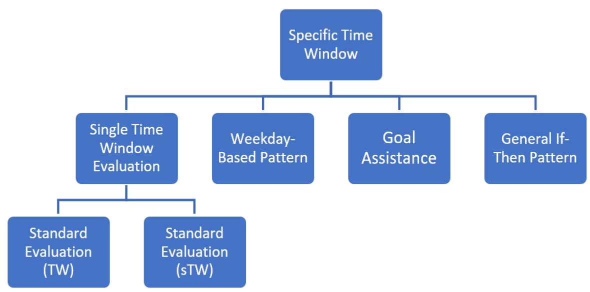


Fig. 3. Specific Time Window Hierarchy.

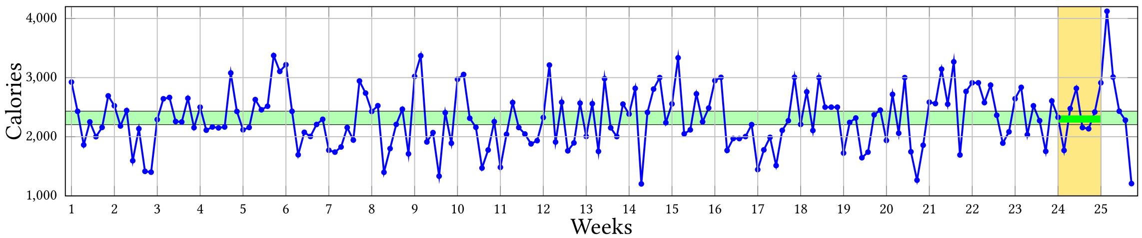
in the data, which can be a great aid in understanding the summaries. Furthermore, we describe the summaries from the perspective of users who are exploring their data and looking for summaries from the simple to the more complex, which also motivates the need for each summary type.

3.2 Specific Time Window Summaries

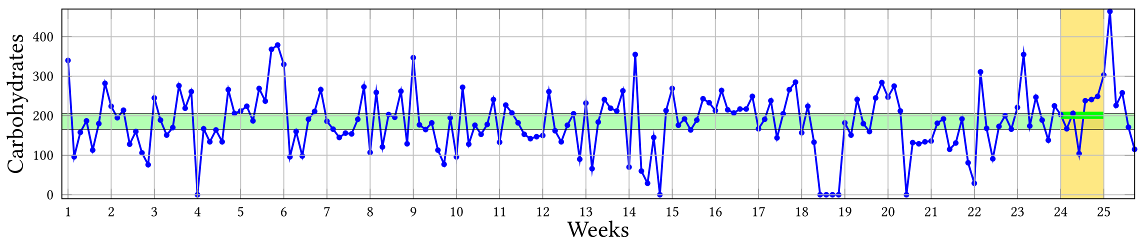
When looking for behavioral patterns in a time series, users may want to search for patterns within a specific time window and evaluate themselves within that time period. Since TW is set to a weekly granularity and sTW is set to a daily granularity in our running example, this user would be looking at a particular week (or days within that week) in her intake data so she can evaluate their progress. The summaries we generate for these patterns are called *specific time window summaries*; this sub-hierarchy of summaries is shown in Figure 3. Specific summary types include standard evaluation summaries, those that elicit the particular weekdays, those that aid goal assistance, and those based on if-then patterns, all within a given time window (e.g., within the past week).

3.2.1 Standard Evaluation Summaries. Standard evaluation summaries are descriptions of evaluations made over the specified time window by pairing the standard evaluation summarizers from Table 2 with the "best" quantifier from Table 1. These summaries contain conclusions drawn from both TW and sTW and use summarizer set $\boldsymbol{S} = \{very\ low, low, moderate, high, very\ high\}$.

Suppose the user is interested in knowing how well she has been doing for the past week. To find this information in $\boldsymbol{T}$, the she may compare the past week with other weeks in the data. Our framework can generate standard evaluation summaries at the TW granularity with this protoform:



(a) Calorie Intake Data



(b) Carbohydrate Intake Data

Fig. 4. Summary provenance—standard evaluation (TW granularity). The horizontal green range denotes "moderate," and the green segment denotes the weekly average. The vertical gold range denotes the week of interest.

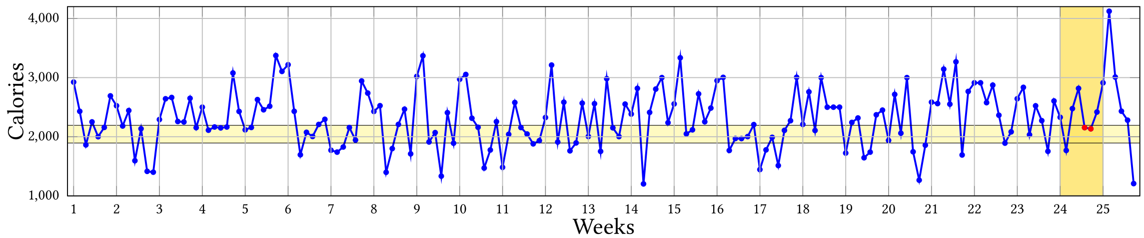
> **Standard Evaluation Protoform** (TW granularity): In the past full ⟨time window⟩, your ⟨attribute 1⟩ has been ⟨summarizer 1⟩, . . . , and your ⟨attribute n⟩ has been ⟨summarizer n⟩.
>
> **Univariate Example** (TW granularity): In the past full **week**, your **calorie intake** has been **moderate**.
> **Multivariate Example** (TW granularity): In the past full **week**, your **calorie intake** has been **moderate** and your **carbohydrate intake** has been **moderate**.

Here, n is the number of attributes. When users receive these summaries on the TW (weekly) granularity, they are able to evaluate their past full week as a whole relative to other weeks in their data. The provenance of the preceding example summaries are shown in Figure 4, where the green range represents the summarizer "moderate" and the gold vertical range represents the time window of interest, which corresponds to the last *full* week, namely week 24.

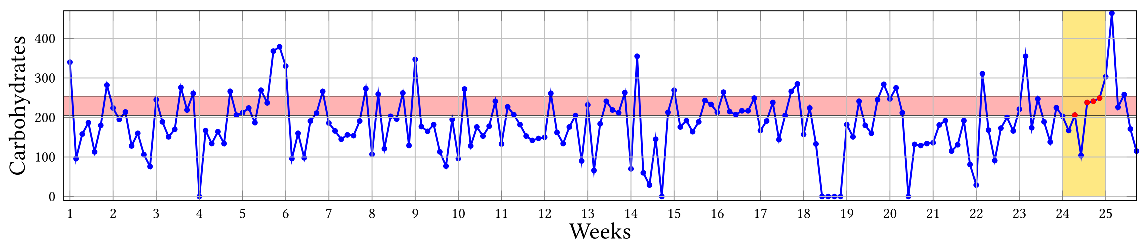
Suppose the user wants more detail on what happened during the past week? We can switch to the sTW (daily) sub-time window granularity by looking at summaries modeled by the following protoform:

> **Standard Evaluation Protoform** (sTW granularity): On ⟨quantifier⟩ ⟨sub-time window⟩ in the past ⟨time window⟩, your ⟨attribute 1⟩ was ⟨summarizer 1⟩, . . . , and your ⟨attribute n⟩ was ⟨summarizer n⟩.
>
> **Univariate Example** (sTW granularity): On **some of the *days*** in the past **week**, your **calorie intake** has been **low**.
> **Multivariate Example** (sTW granularity): On **some of the *days*** in the past **week**, your **calorie intake** has been **low** and your **carbohydrate intake** has been **high**.

With these summaries, the user gains the knowledge that their calorie intake was actually low on some of the days in the past week. Figure 5 shows specific days (red points) that support the summary. In these charts, the yellow range represents the "low" summarizer, whereas the red range represents the "high" summarizer.



(a) Calorie Intake Data



(b) Carbohydrate Intake Data

Fig. 5. Summary provenance—standard evaluation (sTW granularity). The (horizontal) yellow region denotes "low," and the red region denotes "high." The vertical range in gold highlights the week of interest.

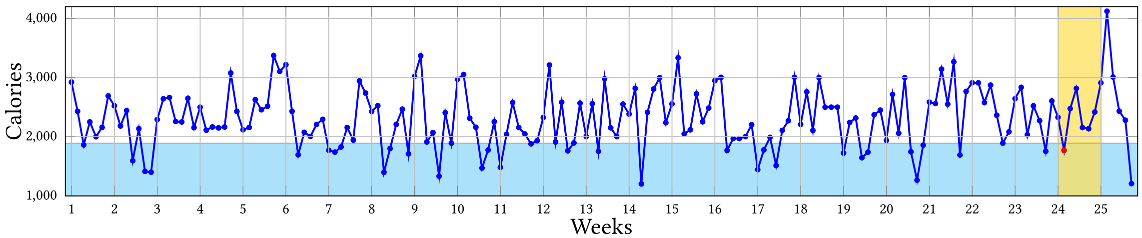
The vertical range in gold represents the time period of focus with relevant data points in red. Users can use these summaries to examine their behavior on those days, and try to get closer to their goals. The multivariate example implies that there may be a behavioral pattern between the user's calorie and carbohydrate intake.

Often, the user may be interested in specific conditions under which a pattern manifests itself. The following protoform can be used when enhanced with a qualifier, which adds more context:

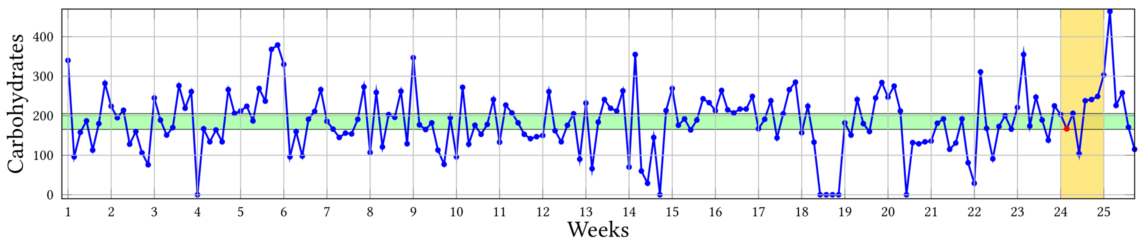
> **Standard Evaluation Protoform** (w/ qualifier): On ⟨quantifier⟩ ⟨sub-time window⟩ in the past ⟨time window⟩⟨qualifier⟩, your ⟨attribute $n + 1$⟩ was ⟨summarizer $n + 1$⟩, . . .
>
> **Multivariate Example** (sTW granularity w/ qualifier): On **all of the *days*** in the past **week**, ***when your calorie intake was very low***, your **carbohydrate intake** was **moderate**.

For this summary, the user can clearly see a behavioral pattern that occurred in the past week. Whenever she had a very low calorie intake in the past week (the qualifier), her carbohydrate intake was moderate. She can use this summary to lower her carbohydrate intake if she chooses to eat similar foods as on the day(s) she had a very low calorie intake. These summaries enable the user to comprehend how well she has performed in specific aspects (e.g., her calorie intake) within a specified time window. She can also look at Figure 6 to verify the multivariate summary. In these charts, the blue range represents the "very low" range, whereas the green range represents the "moderate" range. The vertical range represents the data the summary is describing, and the data points that agree with the summary are in red.

3.2.2 Day-Based Pattern Summaries. These summaries focus on patterns in the user's behavior in terms of certain attributes during "named" days of the week (e.g., Mondays). After receiving the preceding standard evaluation summaries, our user may wonder how she typically performs on certain days of the week. It is possible that she perform better for her health goals on certain days. The following protoform can be used with $\boldsymbol{S}$ = {*very low*, *low*, *moderate*, *high*, *very high*}:



(a) Calorie Intake Data



(b) Carbohydrate Intake Data

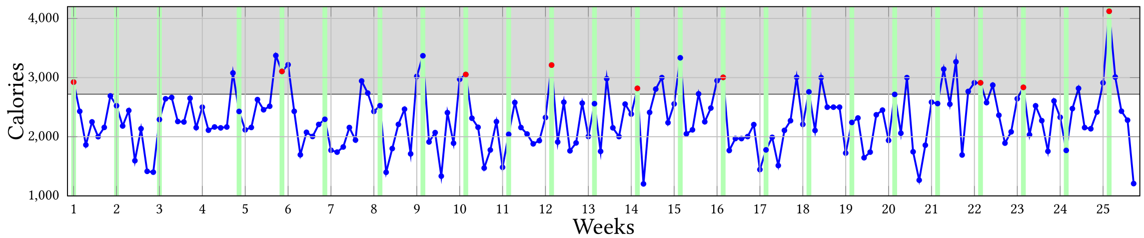
Fig. 6. Summary provenance—standard evaluation (sTW granularity w/ qualifier). The cyan region corresponds to "very low" and green to "moderate." The vertical region in gold denotes the week of interest.

> **Day-Based Pattern Protoform:** Your ⟨attribute 1⟩ tends to be ⟨summarizer 1⟩, . . . , and your ⟨attribute *n*⟩ tends to be ⟨summarizer *n*⟩ on ⟨specified day⟩.
>
> **Univariate Example:** Your **calorie intake** tends to be **very high** on **Sundays**.
> **Multivariate Example:** Your **calorie intake** tends to be **very high** and your **carbohydrate intake** tends to be **very high** on **Sundays**.

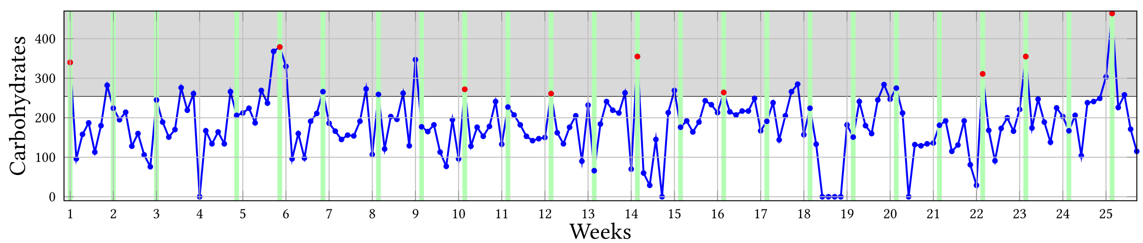
According to the preceding summaries, the user does not perform well on Sundays; both calorie and carbohydrate intake are typically very high on Sundays. Using these conclusions, the user can monitor how she usually eats on that day and take preventive action. Figure 7 illustrates the multivariate summary. In the charts, the gray range represents the "very high" range. The vertical green bars represent the day of the week specified.

3.2.3 Goal Assistance Summaries. Goal evaluation can be added to any summary type to evaluate a certain attribute against a goal or a guideline. How would users evaluate their progress toward their goals? If our user wishes to evaluate how well she limits their carbohydrate and calorie intake in a specific week, this protoform can be used:

> **Goal Evaluation Protoform:** On ⟨quantifier⟩ ⟨sub-time window⟩ in the past ⟨time window⟩, you ⟨summarizer 1⟩ your goal to keep your ⟨attribute 1⟩ ⟨goal 1⟩, . . . , and you ⟨summarizer 1⟩ your goal to keep your ⟨attribute *n*⟩ ⟨goal *n*⟩.
>
> **Univariate Example:** On **most of the *days*** in the past **week**, you **did not reach** your goal to keep your **calorie intake *low***.
> **Multivariate Example:** On **some of the *days*** in the past **week**, you **did not reach** your goal to keep your **calorie intake *low*** and you **reached** your goal to keep your **carbohydrate intake *low***.



(a) Calorie Intake Data



(b) Carbohydrate Intake Data

Fig. 7. Summary provenance—day-based pattern. The data points supporting the conclusion are in red. Specific days of interest are shown as green line segments. The gray region denotes "very high."

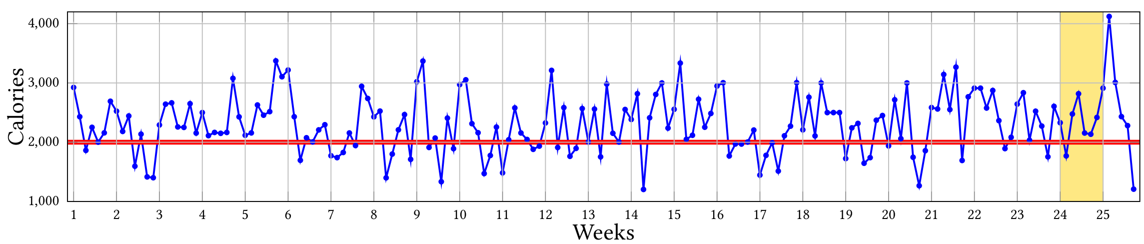
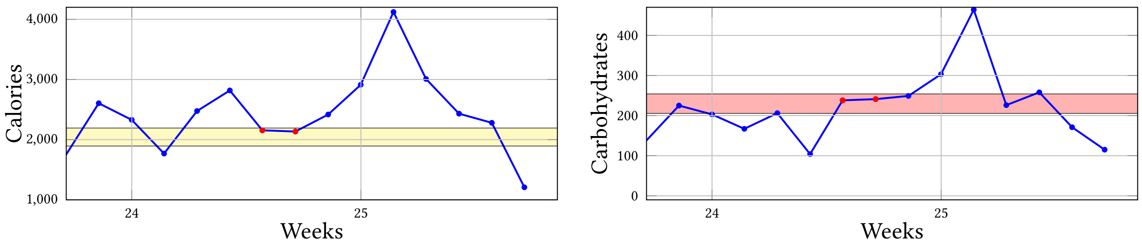


Fig. 8. Summary provenance—goal evaluation/assistance. The red line denotes the goal of 2,000 calories, and the vertical region in gold denotes the week of interest.

This protoform is similar to the one used for the standard evaluation summary at the sub-time window granularity but with the summarizer set $\boldsymbol{S} = \{reached, \textit{did not reach}\}$. These summaries can be used to realize that they fail to reach their calorie intake goal. On the bright side, they have some days where they reach their carbohydrate intake goals. These goals can be extracted from official health guidelines such as the ADA Lifestyle guidelines or suggested by health physicians [3]. Figure 8 illustrates the univariate summary. In this chart, the horizontal red line represents the calorie intake goal and the vertical range represents the data the summary is describing.

In addition to goal evaluation summaries, we also provide goal assistance summaries that not only evaluate users' progress toward a goal but are also constructed to assist the users if they seem to be struggling. These summaries evaluate the user's data for multiple attributes against guidelines that are less defined, such as certain diets. The system must also determine which attributes to mention in the final summary without making the summary too lengthy. Goal assistance summaries can be thought of as a combination of goal evaluation summaries. This time, the set of summarizers is $\boldsymbol{S} = \{increase, decrease\}$. If the users wish to receive a more direct summary of what they should be working on, we can use this protoform:



(a) Calorie Intake Data

(b) Carbohydrate Intake Data

Fig. 9. Summary provenance—general if-then pattern. The yellow region denotes "low," and red denotes "high."

Goal Assistance Protoform: In order to better follow the ⟨goal⟩, you should ⟨summarizer 1⟩ your ⟨attribute 1⟩, ⟨summarizer 2⟩ your ⟨attribute 2⟩, . . . , and ⟨summarizer *n*⟩ your ⟨attribute *n*⟩.

Univariate Example: In order to better follow the **2000-calorie diet**, you should **decrease** your **calorie intake**.
Multivariate Example: In order to better follow the **2000-calorie diet**, you should **decrease** your **calorie intake** and **increase** your **carbohydrate intake.**

Looking at the preceding summary, the user may actually want to increase her carbohydrate intake while lowering her calorie intake based on how she performed last week. This is an expected output, as the *2,000-calorie* diet recommends a higher amount of carbohydrates while the user wishes to limit her carbohydrate intake. It may be best for the user to switch instead to a *low-carbohydrate eating plan*. The provenance illustration for the univariate summary is the same as in Figure 8.

3.2.4 General If-then Pattern Summaries. These summaries find possible correlations between multiple variables pertaining to a user's behavior over the entire time window. What if the user wishes to find a possible correlation between a certain behavior and an inhibiting action she takes? The following protoform generates a summary that describes this correlation:

General If-Then Pattern Protoform: In general, if your ⟨attribute 1⟩ is ⟨summarizer 1⟩, . . . , and your ⟨attribute *n*⟩ is ⟨summarizer *n*⟩, then your ⟨attribute $n + 1$⟩ is ⟨summarizer $n + 1$⟩, . . . , and your ⟨attribute $n + m$⟩ is ⟨summarizer $n + m$⟩

Example: In general, if your **calorie intake** is **low**, then your **carbohydrate intake** is **high**.

These summaries have summarizer set $\boldsymbol{S} = \{$*very low*, *low*, *moderate*, *high*, *very high*$\}$. Figure 9 verifies the multivariate summary, focusing on the last 2 weeks of data. The yellow range represents the "low" range, and the red range represents the "high" range. The data points that agree with the summary are in red.

3.3 Consecutive Time Window Summaries

After searching within specific time windows to find behavioral patterns, our framework allows the user to move on to comparisons between time windows. Naturally, the user would start with consecutive time windows, or time windows that are next to each other. With TW set to a weekly granularity and sTW set to a daily granularity, our user will find patterns between consecutive weeks and consecutive days. The summaries that follow are referred to as consecutive time window summaries, and their inter-relationships are shown in Figure 10.

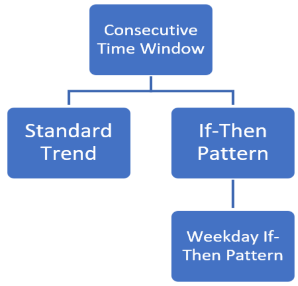


Fig. 10. Consecutive time window hierarchy.

3.3.1 Standard Trend Summaries. Suppose the user wishes to know how she performs from day to day. Looking at the data, how often does her calorie intake increase or decrease between days? We can use a standard trend summary to see this.

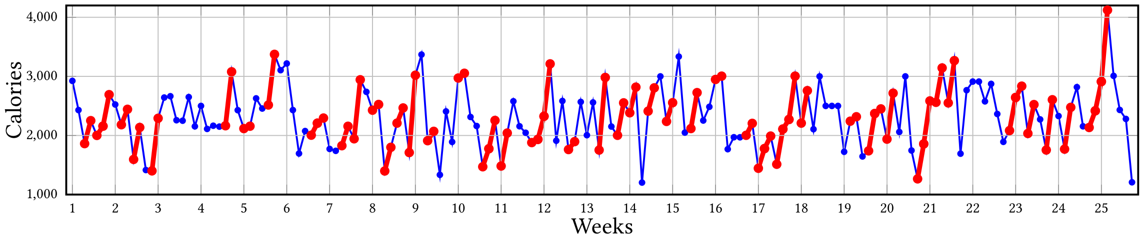
Standard trend summaries describe trends from one sub-time window to the next. These summaries can be used to describe a user's tendency between two consecutive sub-time windows and use summarizer set $\boldsymbol{S}$ = {*increased*, *decreased*, *stayed the same*}.

> **Standard Trend Protoform:** ⟨Quantifier⟩ time, your ⟨attribute 1⟩ ⟨summarizer 1⟩, . . . , and your ⟨attribute n⟩ ⟨summarizer n⟩ from one ⟨sub-time window⟩ to the next.
>
> **Univariate Example: Half of the** time, your **calorie intake *increases*** from one **day** to the next.
> **Multivariate Example: Some of the** time, your **calorie intake *increases*** and your **carbohydrate intake *increases*** from one **day** to the next.

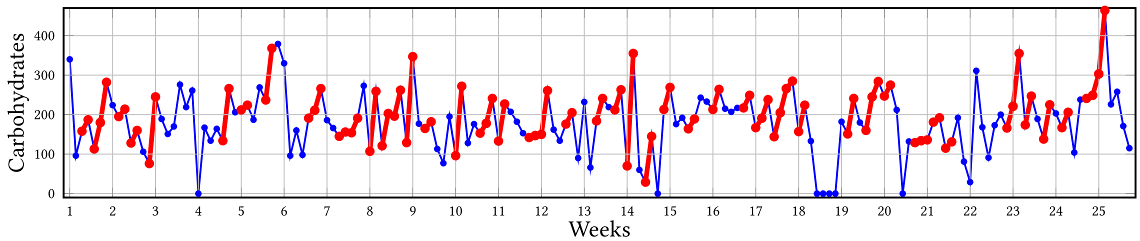
These two summaries allow our user to know that there is around a 50% chance that her calorie intake will increase the next day, and that there is a relatively smaller chance that her calorie intake and her carbohydrate intake will both increase the next day. Although similar to standard evaluation summaries (which evaluate the attribute on each day), here we evaluate the attribute between 1 day and the next; these summaries are ratio-based and span the entire dataset instead of a specified time window. Figure 11 verifies the multivariate summary. In the charts, the red line segments indicate where the pattern is found.

3.3.2 If-Then Pattern Summaries. What if the user wishes to know more about how her past and current behaviors predict the trends in the near future? For answering this, we propose if-then pattern summaries that provide more interesting patterns based on frequent sequence mining [52]. These patterns span multiple consecutive sub-time windows and are of variable length, constrained by the size of the time window. They use summarizer set $\boldsymbol{S}$ = {*very low*, *low*, *moderate*, *high*, *very high*}. The protoform is as follows:

> **If-Then Pattern Protoform:** There is ⟨confidence value⟩ confidence that, when your ⟨attribute 1⟩ is ⟨summarizer 1:1⟩, then ⟨summarizer 2:1⟩, . . . , then ⟨summarizer m:1⟩, . . . , and your ⟨attribute n⟩ is ⟨summarizer 1:n⟩, then ⟨summarizer 2:n⟩, . . . , then ⟨summarizer m:n⟩, your ⟨attribute 1⟩ tends to be ⟨summarizer $(m + 1)$:1⟩, . . . , and your ⟨attribute n⟩ tends to be ⟨summarizer $(m + 1)$:n⟩ the next ⟨time window⟩.
>
> **Univariate Example:** There is **100%** confidence that, when your **calorie intake** follows the pattern of being **moderate**, your **calorie intake** tends to be **very low** the next **day**.



(a) Calorie Intake Data



(b) Carbohydrate Intake Data

Fig. 11. Summary provenance—standard trend. Red line segments illustrate the pattern.

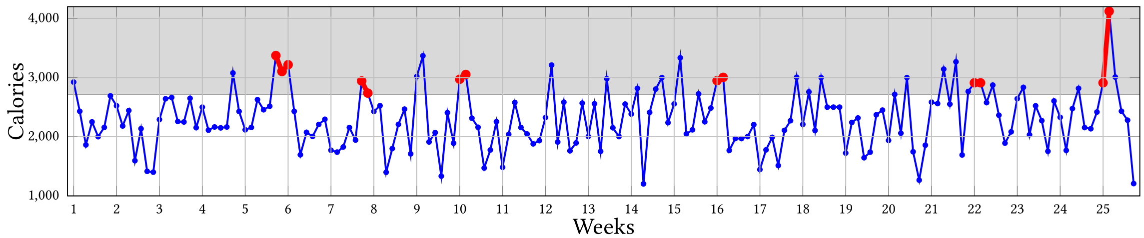
Multivariate Example: There is **100%** confidence that, when your **calorie intake** follows the pattern of being **very high**, your **calorie intake** tends to be **very high** and your **carbohydrate intake** tends to be **very high** the next **day**.

In the preceding protoform, m represents the number of summarizers per attribute and n represents the total number of attributes. From these summaries, the user can conclude that her calorie intake is typically very low after having a moderate intake the previous day. However, if she has a very high calorie intake, then both calorie and carbohydrate intake remains very high the next day as well. Figure 12 verifies the multivariate summary. In the charts, the gray range represents the "very high" range. The data points that agree with the summary are in red.

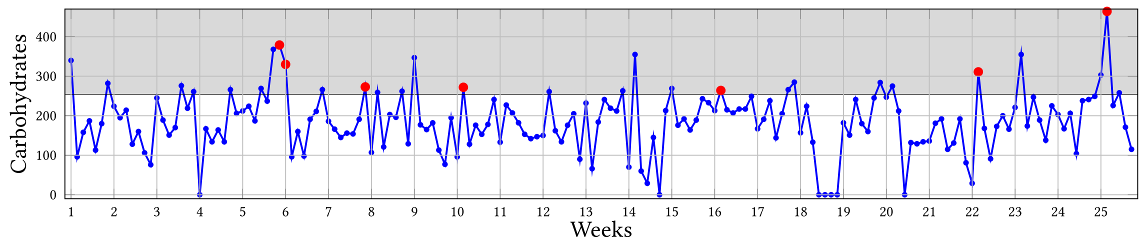
How about if the user wants to see these behavioral patterns pertaining to days of the week? If-then pattern summaries can also be made dependent on the day of the week, via the protoform:

Day If-Then Pattern Protoform: There is ⟨confidence value⟩ confidence that, when your ⟨attribute 1⟩ is ⟨summarizer 1:1⟩ on a ⟨day 1:1⟩, then ⟨summarizer 2:1⟩ on a ⟨day 2:1⟩, . . . , then ⟨summarizer m:1⟩ on a ⟨day m:1⟩, . . . , and your ⟨attribute n⟩ is ⟨summarizer 1:n⟩ on a ⟨day 1:n⟩, then ⟨summarizer 2:n⟩ on a ⟨day 2:n⟩, . . . , then ⟨summarizer m:n⟩ on a ⟨day m:n⟩, your ⟨attribute 1⟩ tends to be ⟨summarizer $(m+1)$:1⟩ the next ⟨day $(m+1)$:1⟩, . . . , and your ⟨attribute n⟩ tends to be ⟨summarizer $(m+1)$:n⟩ the next ⟨day $(m+1)$:n⟩.

Univariate Example: There is **100%** confidence that, when your **calorie intake** follows the pattern of being **very high** on a **Saturday**, your **calorie intake** tends to be **very high** the next **Sunday**.
Multivariate Example: There is **100%** confidence that, when your **calorie intake** follows the pattern of being **very high** on a **Saturday**, your **calorie intake** tends to be **very high** the next **Sunday** and your **carbohydrate intake** tends to be **very high** the next **Sunday**.



(a) Calorie Intake Data



(b) Carbohydrate Intake Data

Fig. 12. Summary provenance—if-then pattern. The gray region denotes "very high." Red points denote points supporting the pattern.

In the preceding protoform, *m* represents the number of summarizer-day pairs per attribute and *n* represents the total number of attributes. The user can observe that if she has Saturdays with very high calorie intake, she typically consumes a lot of calories and carbohydrates on Sunday. This can allow the user to make changes in her weekend diet. Figure 13 verifies the multivariate summary. In the charts, the gray range represents the "very high" range. The data points that agree with the summary are in red, and the vertical green bars show the specified day of the week.

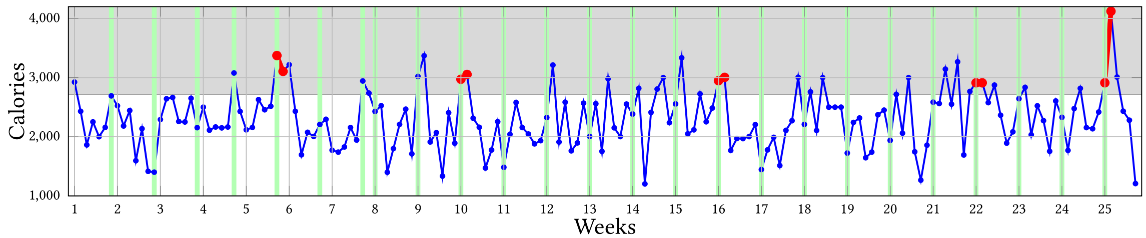
3.4 Non-Consecutive Time Window Summaries

Having examined the patterns that can be found between consecutive time windows, the user may try to find patterns across time windows that are not consecutive. Perhaps the past week she had was similar to another week that occurred in an earlier month. Summaries explaining these types of patterns are called *non-consecutive time window summaries*. These summaries look at time windows that do not necessarily have to be consecutive; they compare discovered trends found in one time window with those of another time window in the data. The inter-relationships for these summaries are shown in Figure 14.

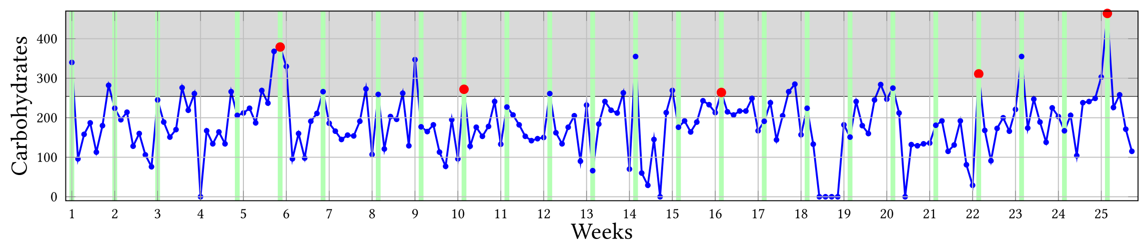
3.4.1 Comparison Summaries. What if the user wants to make comparisons between her most recent week of logging and a week in a much earlier part of her data? Comparison summaries provide comparisons between any two different time windows to help users evaluate their behavioral differences. These summaries use summarizer set **S** = {*higher*, *lower*, *about the same*}. The protoform is as follows:

Comparison Protoform: Your ⟨attribute 1⟩ was ⟨summarizer 1⟩, . . . , and your ⟨attribute *n*⟩ was ⟨summarizer *n*⟩ on ⟨time window 1⟩ ⟨number 1⟩ than they were on ⟨time window 2⟩ ⟨number 2⟩.

Univariate Example: Your **calorie intake** was **about the same** in **week** ***24*** than it was in **week** ***12***.
Multivariate Example: Your **calorie intake** was **about the same** and your **carbohydrate intake** was **about the same** in **week** ***24*** than they were in **week** ***12***.



(a) Calorie Intake Data



(b) Carbohydrate Intake Data

Fig. 13. Summary provenance—day if-then pattern. Red points support the summary, and green line segments denote the days of interest.

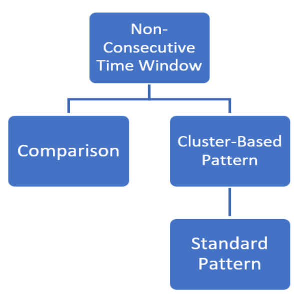


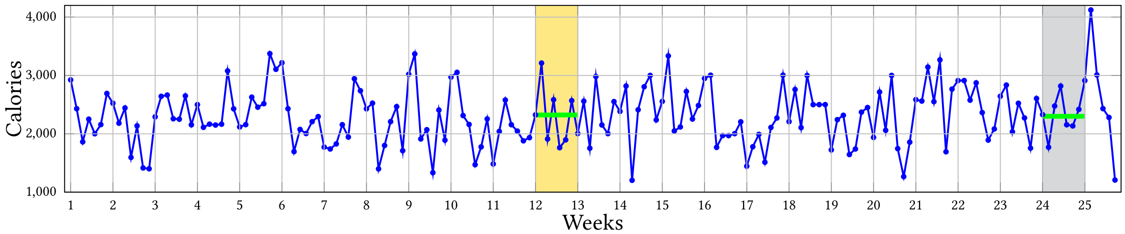
Fig. 14. Non-Consecutive Time Window Hierarchy.

Looking at the past week and another week earlier in the data, the user can see that her intakes of calories and carbohydrates of this past week were about the same as they were 4 months before. Figure 15 verifies the multivariate summary. In the charts, the gray vertical range represents the past week, whereas the golden vertical range represents the week it is compared against.

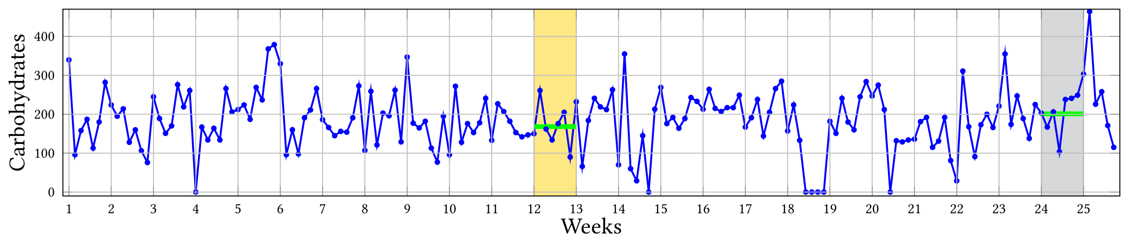
Comparison summaries can also be enhanced with a goal using summarizer set $\boldsymbol{S} = \{$*better*, *not do as well*, *about the same*$\}$. We display the protoform next:

> **Goal Comparison Protoform:** You did ⟨summarizer 1⟩ overall with keeping your ⟨attribute 1⟩ ⟨goal 1⟩, . . . , and you did ⟨summarizer *n*⟩ overall with keeping your ⟨attribute *n*⟩ ⟨goal *n*⟩ in ⟨time window 1⟩ ⟨number 1⟩ than you did in ⟨time window 2⟩ ⟨number 2⟩.

> **Univariate Example:** You did **about the same** overall with keeping your **calorie intake *low*** in **week *24*** than you did in **week *12***.



(a) Calorie Intake Data



(b) Carbohydrate Intake Data

Fig. 15. Summary provenance—comparison. The gray vertical region is the past full week, and the gold region is the week being compared against. The green line segments denote the week-based summarizer, which is "moderate."

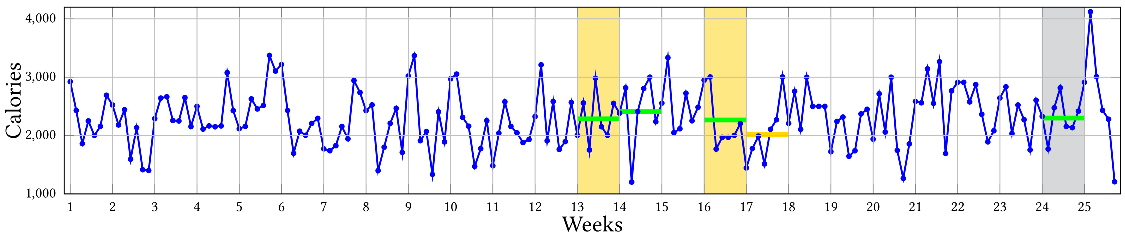
> **Multivariate Example:** You did **about the same** overall with keeping your **calorie intake *low*** and you did **about the same** overall with keeping your **carbohydrate intake *low*** in **week *24*** than you did in **week *12***.

Figure 15 illustrates the multivariate summary. In the charts, the gray vertical range represents the past week, whereas the yellow vertical range represents the week it is compared against. The green segments denote the average nutrient level for that week.

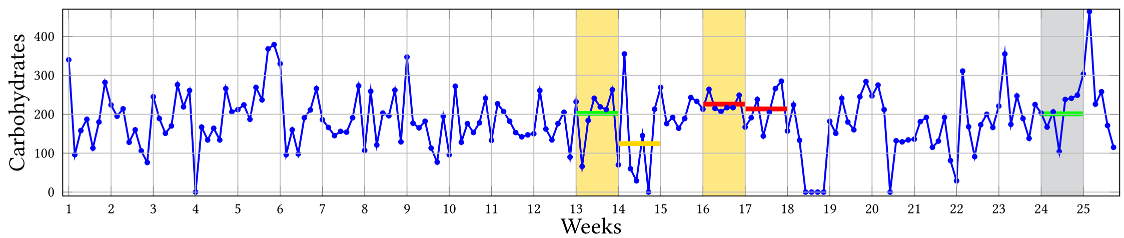
3.4.2 Cluster-Based Pattern Summaries. The users may also want to predict how they will act the following week based on their behavior in the past week. One method to achieve this would be to find other weeks most similar to this past one and to see what happened in the weeks that followed them. We use cluster-based pattern summaries to display these patterns.

These summaries factor in all of the other time windows that are similar to the time window in question, resulting in a cluster. For example, if we are looking at the current week, our system will factor in every other week that has a similar representation (using the Squeezer [18] clustering algorithm). These summaries use summarizer set $\boldsymbol{S} = \{rose, dropped, stayed\ the\ same\}$. In addition to a protoform, we also add a description of the preceding week:

> **Preceding Time Window Description Protoform:** In ⟨time window⟩ ⟨week number⟩, your ⟨attribute 1⟩ was ⟨summarizer 1:1⟩, then ⟨summarizer 2:1⟩, . . . , then ⟨summarizer m_1:1⟩, . . . , and your ⟨attribute n⟩ was ⟨summarizer n:1⟩, then ⟨summarizer n:2⟩, . . . , then ⟨summarizer m_n:n⟩.
> **Cluster-Based Pattern Protoform:** During ⟨quantifier⟩⟨time window (plural)⟩ similar to ⟨time window⟩ ⟨week number⟩, your ⟨attribute 1⟩ ⟨summarizer 1⟩, . . . , and your ⟨attribute n⟩ ⟨summarizer n⟩ the next ⟨time window⟩.



(a) Calorie Intake Data



(b) Carbohydrate Intake Data

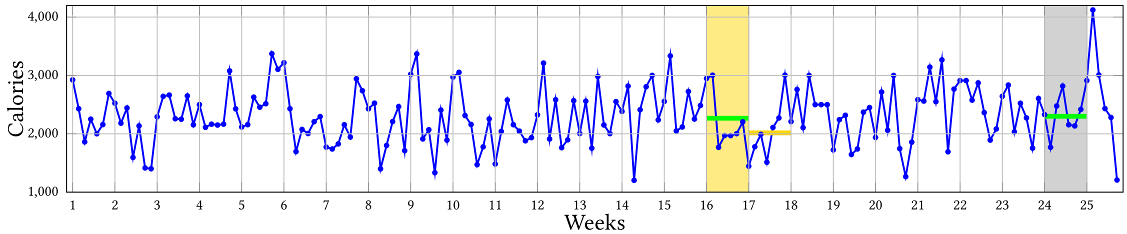
Fig. 16. Summary provenance—cluster-based pattern. The gray region is the past full week and the golden regions are the weeks similar to it. The different colored line segments denote the summarizers at week level: yellow is "low," green "moderate," and red "high."

> **Univariate Example:** In **week** ***24***, your **calorie intake** was **moderate**, then **very low**, then **high**, then **very high**, then **low**, then **moderate**. During **more than half of the** ***weeks*** similar to **week** ***24***, your **calorie intake** ***dropped*** the next **week**.
>
> **Multivariate Example:** In **week** ***24***, your **calorie intake** was **moderate**, then **very low**, then **high**, then **very high**, then **low**, then **moderate** and your **carbohydrate intake** was **moderate**, then **high**, then **very low**, then **high**. During **half of the** ***weeks*** similar to **week** ***24***, your **calorie intake** ***dropped*** and your **carbohydrate intake** ***stayed the same*** the next **week**.

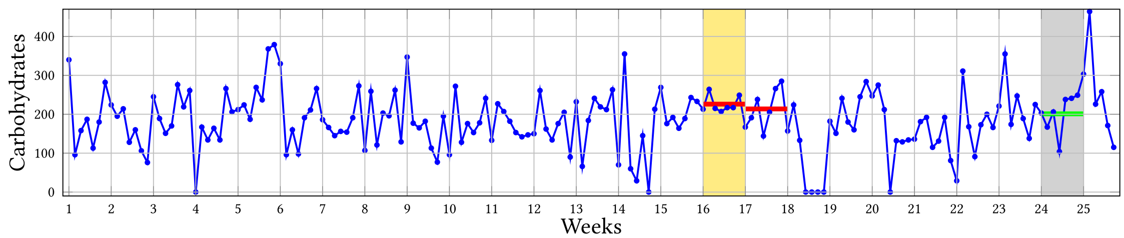
Here, m_i is the number of summarizers for attribute i, and n is the number of attributes. Note that the quantifier is calculated from the cluster alone instead of the entire dataset. We can see that in every summary of this type, the description of the time window comes first. The description is then followed by the actual protoform. From these summaries, the user in our running example is able to know how exactly her past week went for each nutrient. The user can also conclude that her calorie intake will likely drop the next week, whereas her carbohydrate intake has around a 50% chance of staying the same. Figure 16 illustrates the multivariate summary. In the charts, the gray vertical range represents the past week, whereas the golden vertical ranges represent the weeks considered similar to the past week.

The user may also wish to focus on the most recent week. Despite the conclusions stated by the cluster-based pattern summaries, it is possible that the user has not behaved this way recently. Cluster-based pattern summaries can also be used for what we call a *standard pattern protoform*:

> **Standard Pattern Protoform:** The last time you had a ⟨time window⟩ similar to ⟨time window⟩ ⟨number⟩, your ⟨attribute 1⟩ ⟨summarizer 1⟩, . . . , and your ⟨attribute n⟩ ⟨summarizer n⟩ the next ⟨time window⟩.
>
> **Univariate Example:** The last time you had a **week** similar to **week** ***24***, your **calorie intake** ***dropped*** the next **week**.



(a) Calorie Intake Data



(b) Carbohydrate Intake Data

Fig. 17. Summary provenance—standard pattern. The gray region is the past full week, and the golden region is the most recent week similar to it. The different colored line segments denote the summarizers at week level: yellow is "low," green "moderate," and red "high."

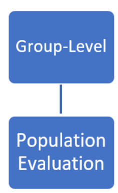


Fig. 18. Group-Level Hierarchy.

> **Multivariate Example:** The last time you had a **week** similar to **week** ***24***, your **calorie intake** ***dropped*** and your **carbohydrate intake** ***stayed the same*** the next **week**.

The user can now see from the standard pattern summaries that (for the most part) the behavior found by the cluster-based pattern summaries still repeats in the most recent week. This reinforces the user's motivation to work toward changing or maintaining her behavior during the next week. Figure 17 verifies the multivariate summary.

3.5 Group-Level Summaries

Moving away from the point of view of a user, suppose we have researchers or clinicians who wish to evaluate an entire population of users or patients. What if they want to know how the entire user population is faring as a whole or how a particular user compares to other users (where such data is available)? We can use group-level summaries to find this answer.

Population evaluation summaries. Our system currently generates population evaluation summaries for nutrient intake data (Figure 18), which has further sub-types described in the following. Our approach can summarize the study population as a whole using the individual summaries previously generated by our system. If the user wishes to know how they compare against other users in terms of their calorie and carbohydrate intake in the past week, this protoform can be used:

> **Population Evaluation Protoform:** ⟨Quantifier 1⟩ participants in this study had a ⟨summarizer 1⟩ ⟨attribute 1⟩, a ⟨summarizer 2⟩ ⟨attribute 2⟩, . . . , and a ⟨summarizer *n*⟩⟨attribute *n*⟩ ⟨sub-protoform⟩.
> **Univariate Example (Standard Evaluation (TW)): Some of the** participants in this study had a **moderate *calorie intake* in the past full week**.
> **Multivariate Example (Standard Evaluation (TW)): Some of the** participants in this study had a **high *calorie intake*** and a **very high *carbohydrate intake* in the past full week**.

Here we define ⟨sub-protoform⟩ as a portion of an actual summary used to describe a number of users in the dataset. This "sub-protoform" identifies the summary type the population has been evaluated on and the conclusion found. The user in our running example can use these summaries to know that at least some of the users in the study did better at managing their calorie intake in the past full week. Now, the user may be more motivated to make changes in their diet for the upcoming week.

There are also special cases of the group-level protoforms for some of the aforementioned protoform types, namely the cluster-based pattern, standard pattern, and (day) if-then pattern protoforms. For the cluster-based pattern protoform, we have the following:

> **Population Evaluation Protoform (Cluster-Based Pattern):** After looking at clusters containing ⟨time window (plural)⟩ similar to this past one, it can be seen that ⟨quantifier⟩ participants with these clusters may see ⟨summarizer 1⟩ in their ⟨attribute 1⟩, . . . , and ⟨summarizer *n*⟩ in their ⟨attribute *n*⟩ next ⟨time window⟩.
> **Univariate Example (Cluster-Based Pattern):** After looking at clusters containing **weeks** similar to this past one, it can be seen that **some of the** participants with these clusters may see **a rise** in their **calorie intake** next **week**.
> **Multivariate Example (Cluster-Based Pattern):** After looking at clusters containing **weeks** similar to this past one, it can be seen that **almost none of the** participants with these clusters may see **a rise** in their **calorie intake** and **little to no change** in their **carbohydrate intake** next **week**.

For the standard pattern protoform, we have the following:

> **Population Evaluation Protoform (Standard Pattern):** Based on the most recent ⟨time window⟩ similar to this past one, it can be seen that ⟨quantifier⟩ participants may see ⟨summarizer 1⟩ in their ⟨attribute 1⟩, . . . , and ⟨summarizer *n*⟩ in their ⟨attribute *n*⟩ next ⟨time window⟩.
> **Univariate Example (Standard Pattern):** Based on the most recent **weeks** similar to this past one, it can be seen that **some of the** participants may see **a drop** in their **calorie intake** next **week**.
> **Multivariate Example (Standard Pattern):** Based on the most recent **weeks** similar to this past one, it can be seen that **almost none of the** participants may see **a rise** in their **calorie intake** and **little to no change** in their **carbohydrate intake** next **week**.

Finally, for the if-then pattern protoform, we have the following:

> **Population Evaluation Protoform (If-Then Pattern):** For ⟨quantifier⟩ participants in this study, it is true that when your ⟨attribute 1⟩ is ⟨summarizer 1:1⟩, then ⟨summarizer 2:1⟩, . . . , then ⟨summarizer m:1⟩, . . . , and your ⟨attribute n⟩ is ⟨summarizer 1:n⟩, then ⟨summarizer 2:n⟩, . . . , then ⟨summarizer m:n⟩, their ⟨attribute 1⟩ tends to be ⟨summarizer $(m + 1)$:1⟩, . . . , and their ⟨attribute n⟩ tends to be ⟨summarizer $(m + 1)$:n⟩ the next ⟨time window⟩.

Univariate Example (If-Then Pattern): For **all of the** participants in this study, it is true that when their **calorie intake** follows the pattern of being **very high**, their **calorie intake** tends to be **high** the next **day**.
Multivariate Example (If-Then Pattern): For **all of the** participants in this study, it is true that when their **calorie intake** follows the pattern of being **low**, their **calorie intake** tends to be **moderate** and their **carbohydrate intake** tends to be **moderate** the next **day**.

The population evaluation protoform for the day if-then pattern protoform differs in the same way as the regular if-then pattern protoform.

4 SUMMARY GENERATION AND MINING

Having described the different protoforms and concrete examples of summaries on the real user data, we now outline our summary generation approach.

Representing time series as symbolic sequences. To find interesting discoveries from time-series data, such as frequent patterns and anomalistic behavior, we first represent the raw time-series data in symbolic form. To achieve this, we use the SAX symbolic representation [28] for each time series, which also makes it easier to represent the time series at different granularities.

The symbols, in particular, are letters from some alphabet. Provided an alphabet size n and the time window size, SAX z-normalizes the raw data of each time series to a zero mean with a standard deviation of 1. It then uses Piecewise Approximate Aggregation to reduce the dimensionality of each time series, depending on the time window size. This reduction allows the ability to easily switch between temporal granularities. After the data is projected onto its principle components and normalized, SAX generates n equiprobability bins based on the standard Gaussian distribution with each segment represented by its corresponding bin symbol. Figure 19 shows the SAX representation of both calorie and carbohydrate intake data for our example user. The SAX representation at the daily level can be read by simply associating each colored region in the figure with the corresponding letter from the alphabet comprising a, b, c, d, and e (corresponding to "very low," "low," "moderate," "high," and "very high," respectively). The SAX representations at the week level correspond to the similarly colored line segments. For example, as observed in Figure 19b), the week-level categorical sequence for carbohydrates is "cbdbeccccccccbdddaccbbdce."

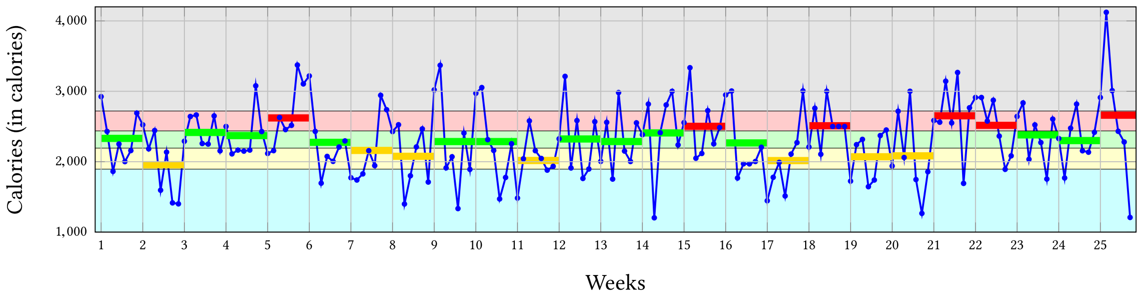
4.1 Pattern Mining

We employ two different types of pattern mining approaches, based on clustering and frequent sequence mining.

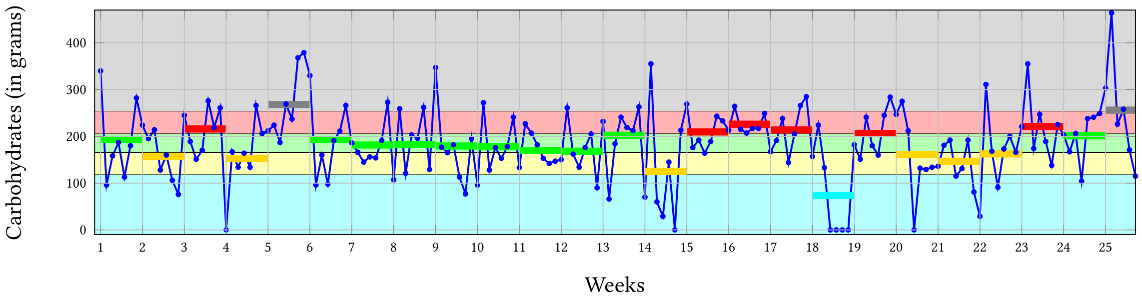
Cluster-based patterns. After partitioning the sub-time window SAX representation into time window tuples (e.g., chunking a string of days into weeks), we combine multiple time series into multivariate symbolic sequences. For example, if one variable has the SAX sequence "abacbbc" and another the sequence "bccabcc," then the combined multivariate sequence is "a-b, b-c, a-c, c-a, b-b, b-c, c-c," where the symbols in the corresponding positions have been combined into an "event." We then group these combined sequences into clusters. For clustering, we use Squeezer [18], which is an online clustering algorithm for categorical data that only needs a similarity threshold s to find clusters. For each tuple t, Squeezer assigns it to an existing cluster or creates a new cluster based on the similarities between t and the existing clusters, using threshold s. A sampling-based approach is used to determine s. We sample a fraction f of the window tuples and calculate the average similarity between each pair of tuples in the sample, using

$$sim(T_i; T_j) = |\{A_k | T_i.A_k = T_j.A_k, 1 \leq k \leq n\}|, \tag{1}$$

where T_i and T_j are tuples, A_k is the SAX symbol at index k, and n is the time window size. In other words, the similarity is based on the number of matching symbols at corresponding positions. We repeat this process $1/f$



(a) Calorie Intake Data: SAX week sequence ‘cbccdcbbccbcccdcbdbbddccd’



(b) Carbohydrate Intake Data: SAX week sequence ‘cbdbecccccccbdddaccbbdce’

Fig. 19. SAX representation for the calorie (a) and carbohydrate (b) intake data for a user from the MyFitnessPal dataset [44]. The different colored regions correspond to the different letters (a, b, c, d, and e) that we map to different summarizers (very low (cyan), low (yellow), moderate (green), high (red), and very high (gray), respectively). Each horizontal segment represents the SAX symbol for each week, whereas individual data points can be seen to belong to the different regions. The SAX representation (sequence over the letters a through e) at the weekly granularity is shown for both calorie and carbohydrate intake.

times (e.g., if we sample $f = 0.2$ or 20% of the tuples, we repeat the sampling $1/f = 5$ times), and set a as the mean of all average pairwise similarities. Finally, we set $s = a + 1$, as suggested by He et al. [18].

Each cluster now contains non-consecutive time windows that have been grouped together by similarity. From these clusters, we can use the history of the attributes involved to “predict” what may happen in the time window following the one we are interested in. Typically, we choose the most recent time window in an attempt to “predict” the future, although it may also be beneficial to use another time window. If we were to use a time window other than the most recent one, we can extract the expected result for the following time window. In short, if we have a time window TW_i, we should be able to use the result of similar time windows to see the expected outcome of the following time window.

For each cluster, we pair each tuple with the tuple that follows it (e.g., pair each week in a cluster with the week following it). Next, we replace the tuples, which are at the sub-time window level (sTW), with the time window–level (TW) SAX symbols. These time window–level pairs are used to generate cluster-based pattern summaries. To describe a pattern, we map the letters (typically, in the last full week) to their corresponding summarizers.

Frequent sequence mining for if-then patterns. To generate if-then pattern summaries, we employ frequent sequence mining over the symbolic SAX temporal data using SPADE [52]. Frequent means that the pattern appears more than a user-specified value called *minimum support*. The method outputs all of the frequent sequence patterns found in the data.

For each frequent sequence, we map each of its prefixes to the following suffixes. For instance, if "abca" is a frequent sequence, then we consider the pairs: ("a," "bca"), ("ab," "ca"), and ("abc," "a"). A similar approach is taken for multivariate data. Next, we generate confidence values (or conditional probability) of observing the suffix given the prefix, given as

$$P(\text{suffix}|\text{prefix}) = \frac{\text{count}(\text{prefix} + \text{suffix})}{\text{count}(\text{prefix})}, \tag{2}$$

where $count(seq)$ is the frequency of the sequence seq. We use a minimum confidence threshold to retain only those frequent if-then patterns of the form "If {prefix}, then {suffix}" with the highest confidence values. Finally, these patterns are used to generate summarizers to be presented to the user.

4.2 Summary Generation

To generate summaries, we fill in the blanks of protoforms presented in Section 3 using summarizers from Table 2 and quantifiers from Table 1. The data we look at is also modified depending on the summary type. For instance, for standard evaluation, the data is the past full week of the data.

As there are many possible combinations of summarizers and quantifiers for each attribute, we choose a combination that is "most appropriate" based on the *average membership function* for a summarizer S and a quantifier Q. We denote μ_S as the membership function value for summarizer S in a time window TW. The membership value μ_S will either have the value of 1 or 0, based on whether the value v for attribute A of the time window follows the conclusion implied by the summarizer. For example, when evaluating a goal for calorie intake where a user wishes to eat at most 2,000 calories a day, the possible summarizers would be "reached" or "did not reach," according to Table 2 (for the standard evaluation summary with a goal). In this case, a value v less than or equal to 2,000 would imply that the user "reached" the goal, whereas a value v greater than 2,000 would imply that the user "did not reach" the goal.

From the μ_S for each time window, we calculate the aggregated average

$$r_S = \frac{1}{n}\sum_{i=1}^{n} \mu_S(y_i), \tag{3}$$

where y_i is a data point in the time-series data and n is the size or length of the time series. This fraction r_S indicates the percentage of the dataset that agrees with the summarizer S.

Once we obtain the r_S value for each summarizer S, we use this value to determine the best quantifier for each summarizer. We employ the use of trapezoidal membership functions [48] μ_Q to calculate how well r_S fits each quantifier. As μ_S is the membership function of a summarizer S, μ_Q is the membership function of a quantifier Q. For example, in Figure 20 (in brown), for the quantifier "most of the," we have

$$\mu_Q(r_S) = \begin{cases} 4r_S - 2 & 0.5 < r_S < 0.75 \\ 1 & 0.75 \leq r_S \leq 0.9 \\ -10r_S + 10 & 0.9 < r_S < 1 \\ 0 & \text{otherwise.} \end{cases} \tag{4}$$

We have defined membership functions for each possible quantifier based on the approach of Kacprzyk et al. [23]. They were designed to create trapezoidal functions that match with the values we believe best fit each quantifier. The membership functions for the different quantifiers are plotted in Figure 20.

Once we have the best quantifier for each summarizer, we will have k quantifier-summarizer candidate pairs. The pair that contains the quantifier with the highest μ_Q will be chosen for the summary, whereas the value of μ_Q eventually becomes the summary's *truth value*. When the most appropriate summarizer and quantifier are found, they are used within the protoform or template to generate the summary. As a result, we generate a list of candidate summaries, paired with a truth value using the μ_Q value of the quantifier within the summary [50].

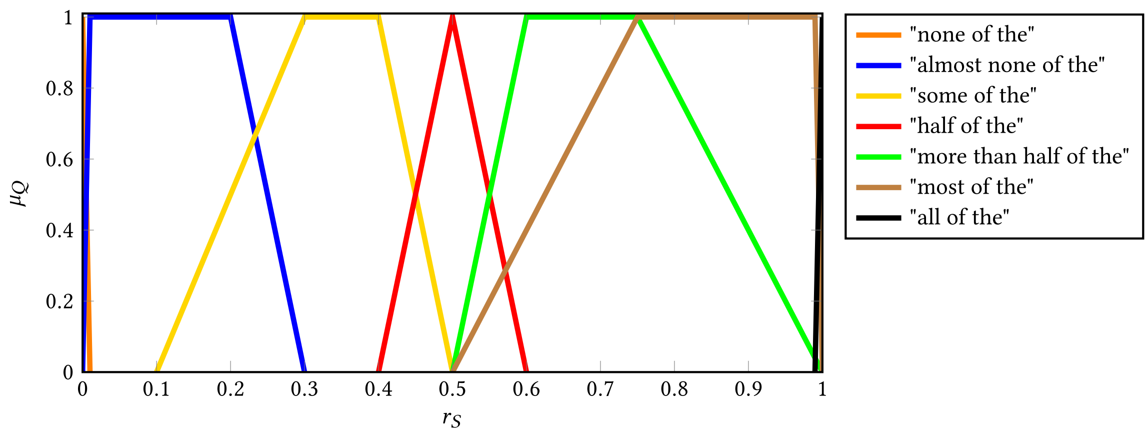


Fig. 20. Quantifier Membership Functions.

Finally, we choose the summary with the highest truth value μ_Q, breaking ties by selecting the summary with the quantifier that implies the largest amount (following Yager's approach for text quality [48]).

4.2.1 Summary Metrics. To evaluate the summaries generated by our system, we use five evaluation metrics, which help measure the Gricean maxims [16]: the maxims of quality, quantity, relevance, and manner. These maxims are well-known pragmatic rules that are known to improve communication of information to humans [6], especially for natural language summaries. The evaluation metrics we use on our summaries are the degrees of truth, imprecision, covering, appropriateness, and coverage, along with the length quality [6, 23, 47]. For all of the metrics, they range from 0 to 1 in value, where 1 is the ideal value.

Degree of truth (T_1). First and foremost, we want the summaries that our framework generates to convey the degree of truth. We use natural language to summarize how often a finding may be true in the data. We use fuzzy quantifiers to describe the frequencies of certain behaviors that best fit the percentage found in the data, although the truthfulness of the overall summary may not be absolute. Zadeh's degree of truth [50] determines which summaries are actually true statements. We use this degree to measure to what extent our summaries follow the maxim of quality, which states how true a summary is and how much evidence supports it. As discussed earlier, Equation (3) calculates the percentage of the dataset that supports the summarizer S. Then, we calculate the μ_Q for the quantifier in each quantifier-summarizer pair via Equation (4), which in fact represents a summary's truth value. For the remainder of this article, we will refer to the degree of truth as $T_1 = \mu_Q$.

Degree of imprecision (T_2). Also known as the degree of fuzziness, the degree of imprecision measures how useful a summary is. It is highly possible that a summary is generated that has a high degree of truth but is also a statement that is not useful, such as "All winter days are cold."

Recall that r_S indicates the fraction of the dataset that agrees with the summarizer S. In our summary generation approach, we keep track of percentages r_{S_j} of each possible summarizer S_j. To calculate the degree of imprecision, we use the following equation:

$$T_2 = 1 - \sqrt[m]{\prod_{j=1 \ldots m} r_{S_j}}, \tag{5}$$

where m is the number of possible summarizers for the protoform type. Here, we compute the geometric mean of the percentages of agreement over the possible summarizers S_j. For the case where a summary is obvious, every summarizer S_j would have a membership value $\mu_{S_j} > 0$ for every (sub-)time window. For example, if every day were snowy and cold, then it would be unwise to output a summary describing this trend. When we subtract the geometric mean from 1, the degree of imprecision represents the extent to which the summary is useful.

Degree of covering (T_3). The degree of covering measures how many points in a user's query are covered by the summary. When we refer to the user's query, we are referring to the subset d of the dataset D that is used to create the summary. We wish to know how often the summary's conclusion (the summarizer S) is true within the subset d. Within this domain, the degree of covering can be expressed as the r_S of the summarizer S of the summary restricted to d. When we generate summaries, we specify time windows to look for particular trends depending on the protoform type. In this way, the time window specification is the "query," whereas the time window itself is the subset d of D. We already base the ratio r_S off of how often summarizer S is true in d so it in fact represents the degree of covering:

$$T_3 = r_S, \tag{6}$$

where S is the best summarizer for the summary. Although this degree is just the ratio of the subset d that agrees with the summary, it is useful to see the actual percentage of agreement alongside the summary. In cases where the quantifier's definition is more fuzzy (i.e., the range of the trapezoidal membership function is especially large), it may be useful to know the exact percentage. For example, if a summary uses the "some of the" quantifier, the trapezoidal function corresponding to this quantifier ranges from an agreement of 10% to 50% of d. Even if the quantifier is not guaranteed to be chosen unless the percentage is between 30% and 40% (see Figure 20), it is still useful to have the ability to know what "some of the" actually means.

Degree of appropriateness (T_4). The degree of appropriateness [23] also helps avoid trivial multivariate summaries. The degree's value represents how interesting and unexpected a finding in the summary may be. We use this degree to measure to what extent our summaries follow the maxim of quantity, which states how much information should be conveyed in a summary. When communicating with a human user, it is important that our summaries avoid providing (1) too much information to easily process when reading the summary or (2) too little information to fully comprehend the findings implied and to act upon those findings.

To calculate the degree of appropriateness of a summary, the summary is split into K sub-summaries by attribute. For each sub-summary, the percentage r_k of the data where the membership value is $\mu_{S_k} > 0$ is calculated, with r_k given as

$$r_k = \frac{1}{n} \sum_{i=1}^{n} \mu_{S_k}(y_i). \tag{7}$$

Afterward, the product

$$r^* = \prod_{k=1}^{K} r_k \tag{8}$$

of the percentages r_k is calculated. Finally, the absolute difference between r^* and the summary's degree of covering T_3, given as

$$T_4 = |r^* - T_3|, \tag{9}$$

yields the degree of appropriateness. It should be noted that the degree of appropriateness will be 0 for any univariate summary since this degree requires relations between two or more variables.

This degree is mainly used to analyze relations in the data. For example, if a user has a high calorie intake on 50% of the days and a low carbohydrate intake on 50% of the days, one may expect that the user has a high calorie intake and a low carbohydrate intake on 25% of the days. This intuition corresponds to the product of ratios r^* presented earlier. If, however, the actual percentage of days differs from 25%, then we can say that the outcome is unexpected and the difference represents the extent to which the outcome differs from what was expected. In terms of the level of informativeness, if the degree of appropriateness is 0, it is possible that the summary states too much where the finding is too precise or too little where the finding is too vague.

Degree of coverage (T_5). We also want the summaries to be relevant to the user. It would not be very useful to receive a summary that is not relevant to a user's context or situation. The maxim of relevance states how relevant the summary should be. It is calculated by using the degree of coverage (not to be confused with the degree of covering), which determines whether the conclusion made by the summary is supported by enough data [47]. If the summary is not supported by enough data, then it may not be worth stating.

We can use the ratio r_S to find the percentage of the data that agrees with the summary. The degree of coverage [47] is given as

$$T_5 = f(r_S) = \begin{cases} 0 & r_S \leq r_1 \\ \dfrac{2(r_S - r_1)}{(r_2 - r_1)^2} & r_1 < r_S < \dfrac{r_1 + r_2}{2} \\ 1 - \dfrac{2(r_S - r_1)}{(r_2 - r_1)^2} & \dfrac{r_1 + r_2}{2} \leq r_S < r_2 \\ 1 & r_S \geq r_2 \end{cases} \tag{10}$$

In the preceding equation, Wu et al. [47] use values of 0.02 and 0.15 for r_1 and r_2, respectively. The definition of this function creates a trapezoidal membership function (e.g., see Figure 20). Where r_S lies on this curve determines how relevant the finding is. Intuitively, this is a fuzzy measure of covering.

Length quality (T_6). Finally, we want for our summaries to be concise. The maxim of manner states how clear the summary should be. The more words a user has to read, the less invested that user will be in what the summary means. In addition, the summary may become more difficult to understand [6]. As the way we convey information is extremely important within the personal health domain, we must evaluate the conciseness as well as the comprehensiveness of our summaries. We calculate the length measure [23] as follows:

$$T_6 = 2(0.5^{cardS}) \tag{11}$$

In this case, S is the set of summarizers included within the summary. This function generates an exponentially decreasing curve in the number of summarizers so that the higher the number of conclusions within a summary, the lower the length quality.

5 EXPERIMENTS

We ran experiments on multiple datasets to analyze the different types of summaries we generate. In particular, we use real data from the *MyFitnessPal food log dataset* [44], which consists of 587,187 days of food log data across 9,900 users over a course of up to 180 days. Each entry logs a user's food items with nutrient information, daily totals, and the user's nutrient goals. We also use user health data from *Insight4Wear* [35], which is a quantified-self/life-logging app, with about 11.5 million records of information. It provides data gathered from mobile devices that track step count, heart rate, and user activities for around 1,000 users.

For all of the example summaries reported earlier in the article, we used a default alphabet size of $n = 5$, a time window of 7 days, a minimum support of 20%, and a minimum confidence of 80%, which comprise the default parameter values. We explore the use of different sets of input parameters later in this section. In the following, we provide the results of our framework's summary generation on real user data and also show quantitative results in terms of evaluation metrics. It is important to note that existing systems for summary generation are either not publicly available or they do not handle time-series data; therefore, a direct comparison is not feasible. Nevertheless, we qualitatively showcase how our framework compares to other state-of-the-art works on temporal data from stock market and weather domains. However, for reproducibility, our implementation is open source and can be downloaded from https://github.com/harrij15/TemporalSummaries.

Table 3. Univariate Individual-Level Summaries for Calorie Intake Data

Protoform Type	Summary	T_1	T_2	T_3	T_4	T_5	T_6
`Standard Evaluation (TW)`	In the past full week, your calorie intake has been moderate.	N/A	N/A	1	0	1	1
`Standard Evaluation (sTW)`	On some of the days in the past week, your calorie intake has been low.	0.93	0.81	0.29	0	1	1
`Standard Evaluation + Goal`	On most of the days in the past week, you did not reach your goal to keep your calorie intake low.	1	0.65	0.86	0	1	1
`Comparison`	Your calorie intake was about the same in week 24 as it was in week 12.	N/A	N/A	1	0	1	1
`Comparison + Goal`	You did about the same overall with keeping your calorie intake low in week 24 than you did in week 12.	N/A	N/A	1	0	1	1
`Standard Trend`	Half of the time, your calorie intake increases from one day to the next.	0.71	0.84	0.53	0	1	1
`Cluster-Based Pattern`	In week 24, your calorie intake was moderate, then very low, then high, then very high, then low, then moderate. During more than half of the weeks similar to week 24, your calorie intake dropped the next week.	1	1	0.6	0	1	0.02
`Standard Pattern`	The last time you had a week similar to week 24, your calorie intake dropped the next week.	N/A	N/A	1	0	1	1
`If-Then Pattern`	There is 100% confidence that, when your calorie intake follows the pattern of being moderate, your calorie intake tends to be very low the next day.	1	0.68	0.32	N/A	1	0.5
`Day If-Then Pattern`	There is 100% confidence that, when your calorie intake follows the pattern of being very high on a Saturday, your calorie intake tends to be very high the next Sunday.	0.7	0.76	0.24	N/A	1	0.5
`Day-Based Pattern`	Your calorie intake tends to be low on Tuesdays.	0.9	0.81	0.28	0	1	1
`Goal Assistance`	In order to better follow the 2000-calorie diet, you should decrease your calorie intake.	N/A	N/A	N/A	N/A	N/A	1

5.1 Summary Generation

We show summaries on calorie and carbohydrate intake from the MyFitnessPal food log dataset and heart rate data from Insight4Wear. All summaries are generated using the default input parameters.

5.1.1 Calorie and Carbohydrate Intake: MyFitnessPal Food Logs. Based on the calorie and carbohydrate data shown in Figure 1, we display three lists of summaries generated for our user: for calorie intake (univariate), for carbohydrate intake (univariate), and for summaries handling both calorie and carbohydrate intake (multivariate). For each list, there are also corresponding group-level summaries evaluated on 389 users (15,915 summaries) from the food log dataset. We selected users that have logged at least 175 days. In total, our system generates 113 summaries.

Calorie intake: univariate summaries. With the calorie intake data from Figure 1(a), we can use the summaries to draw a picture of how the user usually handles her calories, what kinds of conclusions we can draw from this data, and how our user compares to the rest of the study population. Our system produces 19 individual-level summaries using 11 protoforms and 16 group-level summaries using 5 protoforms. To avoid repetition,

Table 4. Univariate Group-Level Summaries for Calorie Intake Data

Protoform Type	Summary	T_1	T_2	T_3	T_4	T_5	T_6
`Standard Evaluation (TW)`	Some of the participants in this study had a moderate calorie intake in the past full week.	1	0.84	0.37	0	1	1
`Standard Evaluation (sTW)`	Almost none of the participants in this study had a high calorie intake on more than half of the days in the past week.	1	0.98	0.06	0	0.21	1
`Standard Evaluation + Goal`	Some of the participants in this study reached their goal to keep their calorie intake low on all of the days in the past week.	0.84	0.87	0.42	0	1	1
`Comparison`	Some of the participants in this study had a higher calorie intake in week 12 than they did in week 24.	1	1	0.3	0	1	1
`Comparison + Goal`	Some of the participants in this study did not do as well with keeping their calorie intake low in week 12 as they did in week 24.	1	1	0.31	0	1	1
`Standard Trend`	More than half of the participants in this study increase their calorie intake from one day to the next half of the time.	0.91	1	0.59	0	1	1
`Cluster-Based Pattern`	After looking at clusters containing weeks similar to this past one, it can be seen that some of the participants with these clusters may see a rise in their calorie intake next week.	1	0.67	0.34	0	1	1
`Standard Pattern`	Based on the most recent weeks similar to this past one, it can be seen that half of the participants may see little to no change in their calorie intake next week.	0.95	0.69	0.5	0	1	1
`If-Then Pattern`	For all of the participants in this study, it is true that when their calorie intake follows the pattern of being very high, their calorie intake tends to be high the next day.	1	0	1	0	1	0.5
`Day If-Then Pattern`	For all of the participants in this study, it is true that when their calorie intake follows the pattern of being high on a Tuesday, their calorie intake tends to be moderate on a Wednesday.	1	0	1	0	1	0.5
`Day-Based Pattern`	Some of the participants in this study tend to have a low calorie intake on Mondays.	0.81	1	0.26	0	1	1
`Goal Assistance`	All of the participants in this study have been given advice to decrease their calorie intake.	1	0	1	0	1	1

11 representative individual-level summaries are shown in Table 3 and 9 group-level summaries are shown in Table 4.

From the individual-level summaries, it becomes apparent that the user is struggling with their calorie intake. The standard evaluation (TW) and goal evaluation summaries explain how our user has struggled in the past week. We can gather from the comparison and goal comparison summaries that the user is also performing worse than the week before, so she is getting further from her health goals. From the standard trend, the cluster-based pattern, and the standard pattern summaries, our user can also see that she will most likely do even worse the next week unless they make changes to her usual routine. To make changes, the user can look at the standard evaluation (sTW), the (day) if-then pattern, and the day-based pattern summaries to closely look at their behavioral tendencies and see when she did things right. Looking at the group-level summaries, the user seems to be performing as well as some of the other users when comparing the summaries, although performing fairly worse than the average user.

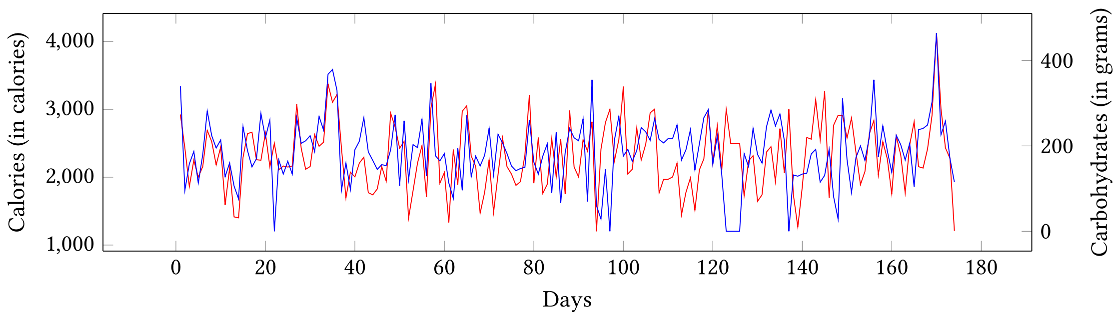


Fig. 21. Calorie (red) and carbohydrate (blue) intake data (superimposed for multivariate analysis).

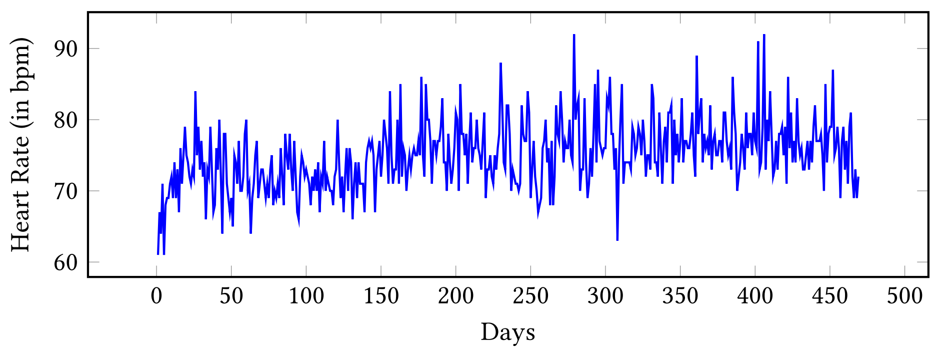


Fig. 22. Heart Rate Data Snippet (Insight4Wear).

When looking at the evaluation metrics in these tables, we can see that some of the evaluation metrics do not apply to all of the individual-level summary types (labeled as N/A). For the degrees of truth T_1 and imprecision T_3, there are certain individual-level summary types that do not use a ratio-based method on a subset d of the dataset D. The goal assistance summary depends only on the average value of the last week's values to draw conclusions about how well the user followed a certain diet in the past full week. In light of this, only the length quality metric is applicable for this summary type.

Carbohydrate intake: univariate summaries. For carbohydrate intake data from Figure 1(b), our system produces 19 individual-level summaries using nine protoforms and 14 group-level summaries using six protoforms. The conclusions we can draw from the carbohydrate intake are very similar to what we drew from the calorie intake (using the same protoforms). At the group level, the user can observe that most of the other users struggled to reach their daily carbohydrate intake goals in the past week. We omit the detailed results since they are qualitatively similar to the calorie intake case.

Calorie and carbohydrate intake: multivariate summaries. What if there is a correlation between the calorie and carbohydrate intake (see Figure 1) for this particular user? We can find out by looking at the multivariate summaries. Figure 21 shows both calories and carbohydrates superimposed by day for our example user. Although it is easy to see that they are correlated, it is nevertheless hard to discern common trends and patterns directly from the raw multivariate time-series data. In contrast, for the joint calorie and carbohydrate intake data, our system produces 27 individual-level summaries using 11 different protoforms and 18 group-level summaries using 8 different protoforms. Representative individual-level summaries (13 of them) are shown in Table 5, and group-level summaries (10 of them) are shown in Table 6. It appears that our user is performing much better

Table 5. Multivariate Individual-Level Summaries for Calorie and Carbohydrate Intake

Protoform Type	Summary	T_1	T_2	T_3	T_4	T_5	T_6
`Standard Evaluation (TW)`	In the past full week, your calorie intake has been moderate and your carbohydrate intake has been moderate.	N/A	N/A	1	0	1	0.5
`Standard Evaluation (sTW)`	On some of the days in the past full week, your calorie intake has been low and your carbohydrate intake has been high.	0.93	1	0.29	0.12	1	0.5
`Standard Evaluation (sTW) w/ qualifier`	On all of the days in the past week when your calorie intake was very low, your carbohydrate intake was moderate.	1	1	1	0.96	0.99	0.5
`Standard Evaluation + Goal`	On some of the days in the past week, you did not reach your goal to keep your calorie intake low and you reached your goal to keep your carbohydrate intake low.	0.71	1	0.43	0.06	1	0.5
`Comparison`	Your calorie intake was about the same and your carbohydrate intake was about the same in week 24 than they were in week 12.	N/A	N/A	1	0	1	0.5
`Comparison + Goal`	You did about the same overall with keeping your calorie intake low and you did about the same overall with keeping your carbohydrate intake low in week 24 than you did in week 12.	N/A	N/A	1	0	1	0.5
`Standard Trend`	Some of the time, your calorie intake increases and your carbohydrate intake increases from one day to the next.	1	1	0.32	0.06	1	0.5
`Cluster-Based Pattern`	In week 24, your calorie intake was moderate, then very low, then high, then very high, then low, then moderate and your carbohydrate intake was moderate, then high, then very low, then high. During half of the weeks similar to week 24, your calorie intake dropped and your carbohydrate intake stayed the same the next week.	1	1	0.5	0.25	1	0
`Standard Pattern`	The last time you had a week similar to week 24, your calorie intake dropped and your carbohydrate intake stayed the same the next week.	N/A	N/A	1	0	1	0.5
`If-Then Pattern`	There is 100% confidence that, when your calorie intake follows the pattern of being very high, your calorie intake tends to be very high and your carbohydrate intake tends to be very high the next day.	0.7	0.76	0.24	N/A	1	0.25
`Day If-Then Pattern`	There is 100% confidence that, when your calorie intake follows the pattern of being very high on a Saturday, your calorie intake tends to be very high the next Sunday and your carbohydrate intake tends to be very high the next Sunday.	1	0.8	0.2	N/A	1	0.25
`Day-Based Pattern`	Your calorie intake tends to be very low and your carbohydrate intake tends to be very low on Mondays.	1	1	0.04	0.02	0.05	0.5
`Goal Assistance`	In order to better follow the 2000-calorie diet, you should decrease your calorie intake and increase your carbohydrate intake.	N/A	N/A	N/A	N/A	N/A	0.5

Table 6. Multivariate Group-Level Summaries for Calorie and Carbohydrate Intake

Protoform Type	Summary	T_1	T_2	T_3	T_4	T_5	T_6
Standard Evaluation (TW)	Almost none of the participants in this study had a very high calorie intake and a high carbohydrate intake in the past full week.	1	0.98	0.02	0	0	0.5
Standard Evaluation (sTW)	Almost none of the participants in this study had a very high calorie intake and a high carbohydrate intake on some of the days in the past week.	1	1	0.03	0	0	0.5
Standard Evaluation (sTW) w/ qualifier	Some of the participants in this study had a very low carbohydrate intake, when they had a very low calorie intake on all of the days in the past week.	0.95	1	0.29	0	1	0.5
Standard Evaluation + Goal	Some of the participants in this study reached their goal to keep their calorie intake low and did not reach their goal to keep their carbohydrate intake low on all of the days in the past week.	0.98	0.97	0.3	0	1	0.5
Comparison	Almost none of the participants in this study had a similar calorie intake and a higher carbohydrate intake in week 11 than they did in week 23.	1	1	0.01	0	0	0.5
Comparison + Goal	Almost none of the participants in this study did about the same with keeping their calorie intake low and about the same with keeping their carbohydrate intake low in week 11 as they did in week 23.	1	1	0.03	0	0	0.5
Standard Trend	Most of the participants in this study increase their calorie intake and increase their calorie intake from one day to the next some of the time.	1	1	0.81	0	1	0.5
Cluster-Based Pattern	After looking at clusters containing weeks similar to this past one, it can be seen that almost none of the participants with these clusters may see a rise in their calorie intake and little to no change in their carbohydrate intake next week.	1	0.91	0.16	0	1	0.5
Standard Pattern	Based on the most recent weeks similar to this past one, it can be seen that some of the participants may see little to no change in their calorie intake and little to no change in their carbohydrate intake next week.	0.72	0.91	0.24	0	1	0.5
If-Then Pattern	For all of the participants in this study, it is true that when their calorie intake follows the pattern of being low, their calorie intake tends to be moderate and their carbohydrate intake tends to be moderate the next day.	1	0	1	0	1	0.125
Day If-Then Pattern	For all of the participants in this study, it is true that when their calorie intake follows the pattern of being high on a Tuesday, their calorie intake tends to be moderate on a Wednesday and their carbohydrate intake tends to be moderate on a Wednesday.	1	0	1	0	1	0.125
General If-Then Pattern	For most of the participants in this study, it is true that when they had a very low carbohydrate intake, they had a very low calorie intake.	1	0.79	0.78	0	1	0.5
Day-Based Pattern	More than half of the participants in this study tend to have a very low calorie intake and a very low carbohydrate intake on Fridays.	1	1	0.66	0	1	0.5
Goal Assistance	More than half of the participants in this study have been given advice to increase their calorie intake.	1	0.93	0.7	0	1	1

Table 7. Mapping of Heart Rate Data Using SAX and Meaningful Ranges

SAX		Meaningful Ranges	
Symbol	Summarizer	Value Range	Summarizer
a	very low	0–50	abnormally low
b	low	50–60	low
c	moderate	60–110	within range
d	high	110–120	high
e	very high	120 and above	abnormally high

with carbohydrate intake when compared to her performance with her calorie intake. The rest of the users seem to perform well on Mondays.

5.1.2 Heart rate: Insight4Wear. Figure 22 shows a snippet of heart rate data for one user that spans over 400 days. For creating the corresponding categorical sequence, SAX binning is not ideal because heart rate data has very little temporal variation. Looking at the data, the data points are strictly between 60 and 100 beats per minute (bpm). Due to the lack of temporal variation, the SAX representations for each granularity will be heavily affected. The letters chosen for each day or week will make data points seem to differ greatly when in reality the differences are minimal. For instance, the standard evaluation summaries at the daily and weekly granularities both state the user's average daily heart rate to be "low," even though the heart rate is within a healthy range and only differs slightly from other data points. As heart rate is more about staying within a healthy range, it is better to create our own discretization for this particular dataset. Thus, given that the default equiprobable bins used in SAX are actually not ideal for heart rate data, we generate meaningful ranges for a heart rate; these are shown and contrasted with the SAX symbols in Table 7. As can be seen, a different set of summarizers is used as well.

Our system produces 21 summaries using nine different protoforms, with representative summaries shown in Table 8. Unlike the user for the calorie intake study, this user does not have trouble satisfying the goal of keeping his heart rate within range. For example, the if-then pattern summary suggests that whenever the user has 6 days of "within range" behavior, the heart rate on the following day also remains within range. This study also showcases the robustness of our framework, since we can generate meaningful summaries by simply changing the symbolic mappings and adjusting the summarizers.

5.2 Effect of Choosing Different Parameters

In this section, we will explore different parameter values when running our system on our user's calorie intake data.

5.2.1 Time Window. We tried different time windows to use in order to look for more patterns. What if our user wished to look at months instead of weeks? How about the entire time frame? For our earlier experiments, we used a weekly time window with a daily sub-time window. We re-ran our experiments for a monthly time window with a daily sub-time window, and for no time window (where the entire time frame is evaluated).

When we switched to the monthly granularity, the system produced 17 individual-level summaries using six protoforms and 13 group-level summaries using five protoforms. The change in time window affects every summary type except the standard trend and the day-based pattern summaries since they do not depend on the input time window. The output also does not contain if-then pattern summaries. This is not very surprising since the calorie intake data contains only 6 months of data (174 days), and thus there is not enough data to extract meaningful or frequent monthly patterns. The results are also very different with the group-level summaries, although

Table 8. Summaries for Average Daily Heart Rate

Protoform Type	Summary	T_1	T_2	T_3	T_4	T_5	T_6
`Standard Evaluation (TW)`	In the past full week, your heart rate has been within range.	N/A	N/A	1	0	1	1
`Standard Evaluation (sTW)`	On all of the days in the past week, your heart rate has been within range.	1	1	1	0	1	1
`Standard Evaluation (sTW) + Goal`	On all of the days in the past week, you reached your goal to keep your heart rate within range.	1	1	1	0	1	1
`Comparison`	Your heart rate was lower in week 67 than it was in week 33.	N/A	N/A	1	0	1	1
`Comparison + Goal`	You did about the same overall with keeping your heart rate within range in week 67 than you did in week 33.	N/A	N/A	1	0	1	1
`Standard Trend`	Half of the time, your heart rate increases from one day to the next.	0.56	0.74	0.46	0	1	1
`Cluster-Based Pattern`	In week 67, your heart rate was within range. During more than half of the weeks similar to week 67, your heart rate dropped the next week.	1	1	0.67	0	1	0.5
`Standard Pattern`	The last time you had a week similar to week 67, your heart rate dropped the next week.	N/A	N/A	1	0	1	1
`If-Then Pattern`	There is 100% confidence that, when your heart rate follows the pattern of being within range, your heart rate tends to be within range the next day.	1	0	1	N/A	1	0.5
`Day If-Then Pattern`	There is 100% confidence that, when your heart rate follows the pattern of being within range on a Saturday, your heart rate tends to be within range the next Sunday.	1	0.33	0.67	N/A	1	0.5
`Day-Based Pattern`	Your heart rate tends to be within range on Wednesdays.	1	1	1	0	1	1

the same summary types are present since the set of group-level types is derived from the set of individual-level summary types. As for the summaries themselves, we see a difference in the conclusion between the standard evaluation (sTW) summaries at the weekly and monthly granularities; in particular, we observe that the days in the past week are not very representative of the calorie intake for the entire month.

When we remove the time window, all summaries evaluate the entire time frame. The system produces 10 summaries using three protoforms for both individual-level and group-level summary output. The only summary types that work without a time window are the standard evaluation (sTW), goal evaluation, standard trend, and day-based pattern types. These summary types can be used for the entire dataset where no time window needs to be specified. Similar outcomes are observed for group-level summaries.

5.2.2 Alphabet Size. The alphabet size determines the number of letters we use to discretize the time-series data. The chosen default alphabet size is 5, which allows our framework to use letters "a" through "e" in the alphabet. What if our user wanted her summaries to be more/less precise? When we change the alphabet size, we may be able to find different patterns as the data points will be assigned different letters. Additionally, we will have different sets of summarizers for the summaries we generate. Table 9 displays the number of individual- and group-level summaries found using different alphabet sizes: 3, 5, and 7.

Table 9. Number of Summaries Generated per Alphabet Size

Alphabet Size	Individual Level	Group Level
3	22	16
5	19	16
7	17	16

Table 10. Minimum Support and Confidence Thresholds

Minimum Support	Minimum Confidence	If-Then Pattern Summaries (#)
0	0	118
0	0.2	83
0	0.5	5
0	0.8	5
0.2	0	15
0.2	0.2	15
0.2	0.5	3
0.2	0.8	3

Overall, we can see a decrease in the number of individual-level summaries whenever the alphabet size increases. This may reflect a decrease in the number of if-then pattern summaries found as more letters are used to create the symbolic representation. To select an ideal alphabet size for the summarization framework, we must consider both the precision of the symbolic representation of the data, as well as the quality of the summaries we generate. We can follow the discussion outlined by Lin et al. [28], who evaluate the alphabet size using the tightness of lower bound metric. When calculating this metric for each alphabet size (from 3 to 10), they observed that the bounds are typically weakest when the alphabet size is smaller. They also explain that it may be intuitive to use larger alphabet sizes to better represent the data, although there can be spatial concerns with a larger alphabet size. Therefore, the recommended range (according to the authors) for a chosen alphabet size is between 5 and 8 to balance the tightness of the lower bound and the amount of space used. However, looking at the quality of our summaries, we typically use a different summarizer for each different letter we use. It is possible that the meaning of the summarizers can become increasingly ambiguous as the number of summarizers rises. Although this has not been fully tested, one can imagine the difficulty of finding 10 distinct ways to describe a data point within the range of "high" and "low." Combined with the conclusions from Lin et al. [28], choosing alphabet sizes of 5 or 7 may be the ideal choice for our scenario.

5.2.3 Minimum Support and Confidence Thresholds. These thresholds mainly control the output of the if-then pattern summaries. The default thresholds are 20% for minimum support and 80% for minimum confidence. What if our user wanted patterns that occur more or less frequently? We re-ran our experiments for different minimum support and confidence thresholds (e.g., 20%, 50%, or 80%). We did not find any if-then patterns with a minimum support of 50% and 80%, and thus we show results for the lower support threshold. Note that a minimum support of 0 means that we consider all patterns that occur at least once in the data. As we can see from the results in Table 10, all 118 of the frequent patterns found for our user's calorie intake data in Figure 1(a) occur less than half of the time (since no sequences reached the 50% threshold). Only 15 sequences surpass the 20% support threshold, whereas only 5 sequences surpass the 50% confidence threshold. For the calorie intake data in particular, the default thresholds filter out most of the discovered if-then patterns to just three patterns.

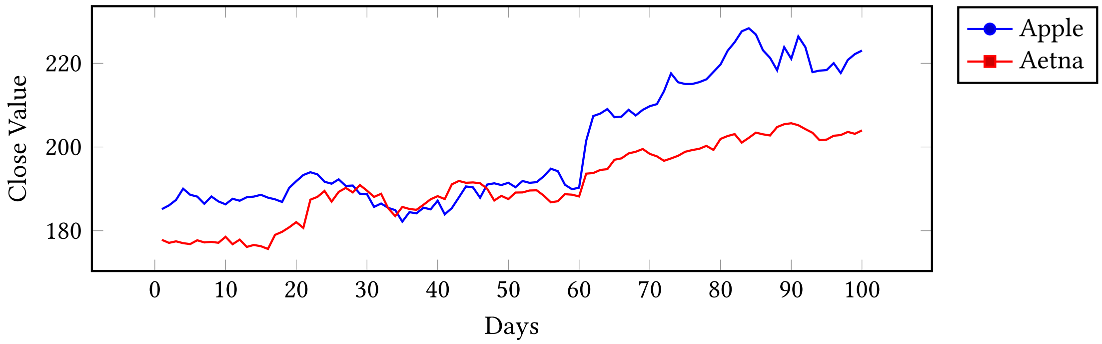


Fig. 23. Close Value Data for Apple and Aetna.

These results beg a question: how many if-then patterns should we be generating? Ideally, we should show only some of the if-then pattern summaries to users, since we do not want to overwhelm them. This implies the need to be able to prioritize and select the if-then patterns that are best to show to a user. At the same time, we believe all relevant patterns should be extracted and used to create a health profile for a user, which can be useful for other kinds of analysis. For example, an if-then pattern that is infrequent could turn into an anomaly later on that may prove useful to mention to a user.

5.3 Generalizabilty of Our Framework: Application to Weather and Stock Data

Although there is a lack of open source or publicly available automatic natural language summarization systems in the personal health domain, we qualitatively compare our summary output versus other systems. We decided to look at the stock market [2] and weather [29] domains, which also demonstrates the generalizability of our approach to time-series data from other domains besides personal health.

Stock market data. In the stock market domain, Aoki et al. [2] extended a neural encoder-decoder model created by Murakami et al. [32] to generate comments about the Nikkei stock market. They generate summaries about the general trend of the stock market time-series ticker data, such as *Nikkei turns lower as yen's rise hits exporters* and *Nikkei Stock Average opens at a high price after Dow Jones Industrial Average closes at a high price.* Their main extension is the ability to handle multivariate input. For our work, we can apply our protoform-based approach to stock market data gathered using the REST API from AlphaVantage [41]. With this API, we retrieved a snippet of 100 days of Apple's and Aetna's stock market data beginning from May 2018, as plotted in Figure 23. Our system is able to provide more insights as shown in Table 11. It generated a total of 242 multivariate summaries, which were slightly modified (protoform-wise) to match the stock market data. We find patterns that cannot be as easily seen, and our summaries say a lot more about the data. The protoform-based approach also has better performance in terms of how quickly the summaries are generated.

Weather data. In the weather domain, the SUMTIME system [39] proposes a general time-series summarization model. They focus their efforts on the weather domain, where they describe the forecast of the next 12 to 24 hours in natural language. An example summary is "W 8-13 backing SW by mid afternoon and S 10-15 by midnight," which describes wind direction and speed. Using our framework, we generated summaries describing the average temperature and the average wind speed tracked by the weather station at the Huntsville International Airport in Huntsville, Alabama. This data was provided by the **National Centers for Environmental Information (NCEI)** [30]. We used two datasets, one containing a year of daily data between March 1, 2018, and March 1, 2019, and the other containing a day of hourly data for January 1, 2010. We display figures for temperature and wind speed daily data in Figures 24(a) and (b). Our framework generates 52 summaries at the weekly (TW) granularity, some of which can be found in Table 12. For the hourly data, we display average temperature and wind speed

Table 11. Apple (AAPL) and Aetna (AET): Stock Market Summaries (AlphaVantage)

Protoform Type	Summary	T_1	T_2	T_3	T_4	T_5	T_6
`Standard Evaluation (TW)`	In the past full week, the AAPL close value has been very high and the AET close value has been very high.	N/A	N/A	1	0	1	0.5
`Standard Evaluation (sTW)`	On all of the days in the past week, the AAPL close value has been very high and the AET close value has been very high.	1	1	1	0	1	0.5
`Standard Evaluation (sTW) w/ qualifier`	On all of the days in the past week when the AAPL close value was very high, the AET close value was very high.	1	1	1	0	1	0.5
`Comparison`	The AAPL close value was higher and the AET close value was higher in week 14 than they were in week 7.	N/A	N/A	1	0	1	0.5
`Standard Trend`	Some of the time, the AAPL close value increases and the AET close value increases from one day to the next.	1	1	0.34	0.004	1	0.5
`Cluster-Based Pattern`	In week 14, the AAPL close value was very high and the AET close value was very high. During more than half of the weeks similar to week 14, the AAPL close value stayed the same and the AET close value stayed the same the next week.	1	1	0.67	0	1	0.125
`Standard Pattern`	The last time you had a week similar to week 14, the AAPL close value stayed the same and the AET close value stayed the same the next week.	N/A	N/A	1	0	1	0.5
`If-Then Pattern`	There is 100% confidence that, when the AAPL close value follows the pattern of being high, the AET close value tends to be high, then high the next day.	1	0.8	0.2	N/A	1	0.25
`Day If-Then Pattern`	There is 100% confidence that, when the AET close value follows the pattern of being very low on a Thursday, the AAPL close value tends to be low the next Friday and the AET close value tends to be very low the next Friday.	1	0.8	0.2	N/A	1	0.125
`General If-Then Pattern`	In general, if the AAPL close value is very low, then the AET close value is very low.	0.55	1	0.45	0.41	0.7	0.5
`Day-Based Pattern`	The AAPL close value tends to be very low and the AET close value tends to be very low on Mondays.	1	1	0.05	0.02	0.13	0.5

data in Figures 24(c) and (d). We generate 11 summaries in total with the multivariate summaries shown in the following Table 13. We can see that some of the summary types do not show as we stick to mainly sub-time window conclusions. Some of the protoforms were also modified to account for the change in granularity.

These two examples showcase the generalizability of our time-series summary generation framework. They demonstrate that the framework is generic and can be applied to different domains with slight changes to the protoforms and the vocabulary (e.g., quantifiers, summarizers).

5.4 Evaluation via User Study

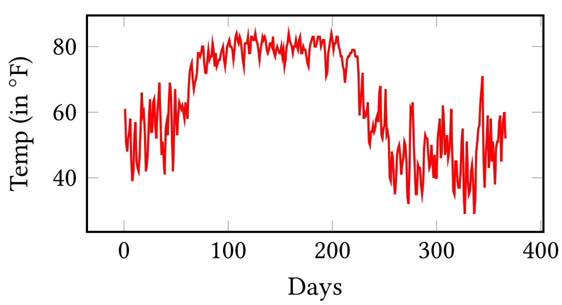
Having looked at various types of summaries produced by our system, along with quantitative metrics, we wanted to evaluate the efficacy of the summaries in terms of the understandability (readability and comprehensiveness), the usefulness of our output summaries, and how well they align with the source data or provenance.

Table 12. Huntsville, Alabama: Temperature and Wind Speed Summaries (NCEI)

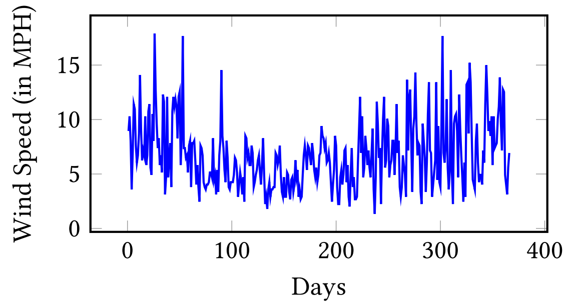
Protoform Type	Summary	T_1	T_2	T_3	T_4	T_5	T_6
`Standard Evaluation (TW)`	In the past full week, the average temperature has been low and the average wind speed has been high.	N/A	N/A	1	0	1	0.5
`Standard Evaluation (sTW)`	On some of the days in the past full week, the average temperature has been low and the average wind speed has been very low.	0.93	1	0.29	0.04	1	0.5
`Standard Evaluation (sTW) w/ qualifier`	On all of the days in the past full week when the average temperature was very low, the average wind speed was low.	1	1	1	0.98	0.99	0.5
`Comparison`	The average temperature was lower and the average wind speed was higher in week 52 than they were in week 26.	N/A	N/A	1	0	1	0.5
`Standard Trend`	Some of the time, the average temperature increases and the average wind speed increases from one day to the next.	0.88	0.94	0.28	0.04	1	0.5
`Cluster-Based Pattern`	In week 52, the average temperature was low, then very low, then low and the average wind speed was very high, then moderate, then very high, then low, then very low. During all of the weeks similar to week 52, the average temperature stayed the same and the average wind speed stayed the same the next week.	1	1	1	0	1	0.002
`Standard Pattern`	The last time there was a week similar to week 52, the average temperature stayed the same and the average wind speed stayed the same the next week.	N/A	N/A	1	0	1	0.5
`If-Then Pattern`	There is 100% confidence that, when the average temperature follows the pattern of being very low, the average temperature tends to be very low and the average wind speed tends to be very high the next day.	0.92	0.79	0.21	N/A	1	0.25
`Day-Based Pattern`	The average temperature tends to be very low and the average wind speed tends to be very low on Wednesdays.	1	1	0.06	0.17	0.17	0.5

Table 13. Huntsville, Alabama: Hourly Temperature and Wind Speed Summaries (NCEI)

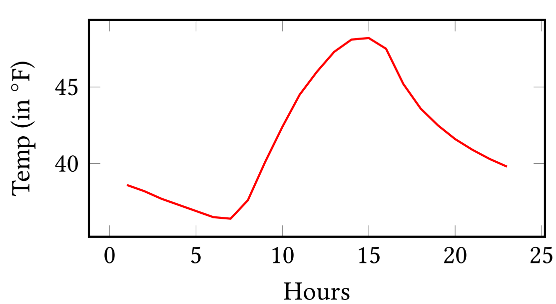
Protoform Type	Summary	T_1	T_2	T_3	T_4	T_5	T_6
`Standard Evaluation (sTW)`	During almost none of the hours in the past day, the average temperature was very low and the average wind speed was very low.	1	1	0.43	0.39	0.95	0.5
`Standard Evaluation (sTW) w/ qualifier`	During most of the hours in the past day, the average temperature was low, the average wind speed was low.	1	1	0	0.09	0.95	0.5
`Comparison`	The average temperature was lower and the average wind speed was lower in hour 23 than they were in hour 11.	N/A	N/A	1	0	1	0.5
`Standard Trend`	During some of the day, the average temperature decreases and the average wind speed decreases from one hour to the next.	1	1	0.3	0.07	1	0.5



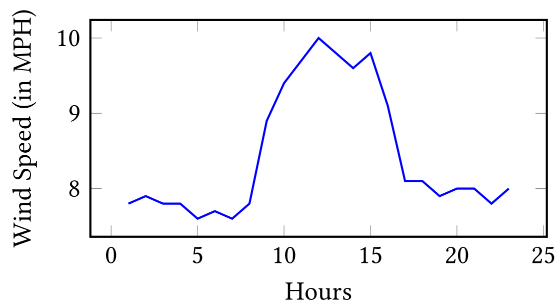
(a) Average Temperature (Daily)



(b) Average Wind Speed (Daily)



(c) Average Temperature (Hourly)



(d) Average Wind Speed (Hourly)

Fig. 24. Average daily temperature and wind speed data for Huntsville, Alabama (for 1 year and for 1 day).

To perform this evaluation, we designed a small user study comprised of 11 participants (eight male and three female) aged 21 to 25 years. As this user study is preliminary, we only excluded potential recruits who were too familiar with our work. All participants had varying levels of fluency in English and knowledge pertaining to personal health. For our study, the participants were asked to evaluate our summaries on three subjective metrics: readability/comprehensiveness, usefulness, and data alignment. Each individual completed a unique survey (i.e., each survey used a different dataset), where they were first presented with a single time-series chart of calorie intake data, similar to Figure 1(a), and a scenario. The scenario matches the running example within the article where a user is following a 2,000-calorie diet. The horizontal colored ranges were also described to the user. In fact, all of the figures presented earlier correspond to similar charts and provenance shown to the participants for various protoforms. After reading the scenario, each participant was asked to provide a textual description of the chart. We requested this preliminary description to analyze what patterns non-experts typically look for and how they describe those patterns. After the description was provided, each participant was shown a number of representative summaries generated for the calorie intake data, as well as corresponding charts displaying the provenance of the discovered patterns. Presented with one summary at a time, each participant was asked to evaluate each summary over the three aforementioned metrics. Each metric can be given a score from 1 (strongly disagree) to 5 (strongly agree) by the participants based on their agreement with the following statements:

(1) This summary is readable and comprehensible.
(2) This summary is useful to me and my goals.
(3) This summary aligns well with the data.

Table 14. Human Evaluation Scores

Summary Type	Readability/ Comprehensiveness	Usefulness	Data Alignment
Standard Evaluation (TW granularity)	3.9	3.55	3.09
Standard Evaluation (sTW granularity)	4.42	3.75	4
Standard Evaluation (sTW granularity w/ qualifier)	4.29	3.65	4.53
Evaluation Comparison	4	3.67	3.92
Goal Comparison	4	3.73	4.09
Goal Evaluation	4.55	4.45	4.18
Standard Trends	4.5	2.5	4
Cluster-Based Pattern	2.73	2.91	2.55
Standard Pattern	4.5	3.38	3.38
If-Then Pattern	4.21	2.86	2.07
Day If-Then Pattern	3.82	3.18	3
General If-Then Pattern	4.33	3.4	3.8
Goal Assistance	4.36	3.36	4
Day-Based Pattern	4.5	4	3.33
Overall	4.13	3.48	3.58

After evaluating the univariate summaries, the participants were asked to provide another description of the data, to capture how the freeform responses change after having seen our summaries. Finally, another variable was added in to display multivariate summaries and the process was repeated. Each participant evaluated 14 summaries in total and was exposed to most, if not all, of the summary types. Some participants did not receive every summary type if the representative summary for the summary type (chosen using a weighted sum of the objective evaluation metrics T_1 through T_6 from Section 4.2.1) was determined to have been redundant or trivial. In this case, an additional representative summary from another summary type served as a replacement.

The results for our user study are presented in Table 14. We show the overall averages of the scores given to the subjective metrics (displayed at the bottom), as well as the average scores within the summary types. Looking at the results, we can see that the participants had a fairly strong agreement overall with the readability and the comprehensiveness of our summaries at a score of 4.13 out of 5. As for usefulness and data alignment, there is still agreement on average when it comes to the usefulness of our summaries (3.48 out of 5) and how well they align with the presented charts (3.58 out of 5). When looking at the averages within the summary types, it can be seen that the goal evaluation summaries had the best scores in all three categories. The standard trends summaries scored lowest on usefulness, although they have high scores in the other categories. Over all of the three aspects, the cluster-based pattern summaries received the lowest scores. We believe that the description of the week before the actual summary may make it too complex of a read. It is also possible that the pattern itself is not easily comprehensible. The if-then pattern summaries also receive lower scores in usefulness and data alignment, which may also stem from pattern complexity. These results suggest that although our framework generates understandable summaries that are useful, and there is great value in displaying the provenance for illustrating the supporting data, there is still scope for further improvement of the more complex summary types like the cluster-based and if-then patterns, both in terms of the natural language text and the accompanying provenance charts. This will provide fruitful directions for our future work.

We also received a lot of helpful descriptions and feedback from the participants that we will use to improve our system in the future. From their descriptions, we are able to determine possible future patterns to find in time-series data, such as variance, consistency, and general drops/rises. It was interesting to find that some of the participants' initial descriptions aligned with patterns our summaries describe, such as the standard evaluation (w/ qualifier) summaries.

6 CONCLUSION

We presented a system to automatically generate summaries of a user's personal health data. Unlike most previous approaches that either focus on tabular, textual, or relatively simple trend summaries, we mine interesting patterns from symbolic representations of numeric temporal data and propose a comprehensive set of useful summaries that cover a wide range of scenarios. We showcase our work using real user data. Our system is designed to extract comprehensible summaries to better guide users toward their goals. To the best of our knowledge, this is the first comprehensive and systematic approach to generate natural language summaries from time-series personal health data via protoforms. There is *no current system* that can automatically extract patterns and clusters from time-series data and present them to a user in an explainable manner in natural language. In fact, our approach is also *generic* and *extensible* to other domains outside of the personal health domain.

It is important to note that our main contribution in this article is the comprehensive framework for the generation of useful and informative summaries of time-series data. We conducted a preliminary user study that confirms our approach is indeed effective and useful. However, it also elucidated aspects that need improvement. In the future, we aim to analyze how our summaries ultimately impact the behavior of users via a larger user study. This will allow us to focus on protoforms that are the most interesting and helpful to users, along with the most comprehensible ways to put these findings into words.

One limitation of our work is that our system is protoform- or template-based. In the future, we seek to automate the summarization process where the use of protoforms is no longer needed while retaining the system efficiency and summary readability. For example, we can extract temporal shapes and relationships to better summarize a time series by describing where interesting shapes (e.g., certain spikes or drops) occur within and across time series. These descriptions could be seen as creating a narrative about a time series within a specific window when applied to the personal health domain. We also aim to utilize deep learning to automatically generate summaries based on the shapes found in the time series.

ACKNOWLEDGMENTS

This work is supported by IBM Research AI through the AI Horizons Network, and was conducted under the auspices of the RPI-IBM Health Empowerment through Analytics Learning and Semantics (HEALS) project.

REFERENCES

[1] Alberto Alvarez-Alvarez and Gracian Trivino. 2013. Linguistic description of the human gait quality. *Engineering Applications of Artificial Intelligence* 26, 1 (2013), 13–23.

[2] Tatsuya Aoki, Akira Miyazawa, Tatsuya Ishigaki, Keiichi Goshima, Kasumi Aoki, Ichiro Kobayashi, Hiroya Takamura, and Yusuke Miyao. 2018. Generating market comments referring to external resources. In *Proceedings of the International Conference on Natural Language Generation.*

[3] American Diabetes Association. 2019. 5. Lifestyle management: Standards of medical care in diabetes—2019. *Diabetes Care* 42, Suppl. 1 (2019), S46–S60.

[4] James Baldwin, Trevor P. Martin, and Jonathan M. Rossiter. 1998. Time series modelling and prediction using fuzzy trend information. In *Proceedings of the International Conference on Soft Computing and Information Intelligent Systems.*

[5] Ildar Z. Batyrshin and Leonid B. Sheremetov. 2008. Perception-based approach to time series data mining. *Applied Soft Computing* 8, 3 (2008), 1211–1221.

[6] Fatih Emre Boran, Diyar Akay, and Ronald R. Yager. 2016. An overview of methods for linguistic summarization with fuzzy sets. *Expert Systems with Applications* 61 (2016), 356–377.

[7] Rita Castillo-Ortega, Nicolás Marín, Daniel Sánchez, and Andrea Tettamanzi. 2011. Linguistic summarization of time series data using genetic algorithms. In *Proceedings of the Conference of the European Society for Fuzzy Logic and Technology.*

[8] Jarvis T.-Y. Cheung and George Stephanopoulos. 1990. Representation of process trends—Part I. A formal representation framework. *Computers & Chemical Engineering* 14, 4 (1990), 495–510.

[9] Eun Kyoung Choe, Nicole B. Lee, Bongshin Lee, Wanda Pratt, and Julie A. Kientz. 2014. Understanding quantified-selfers' practices in collecting and exploring personal data. In *Proceedings of the ACM Conference on Human Factors in Computing Systems*.

[10] James Codella, Chohreh Partovian, Hung-Yang Chang, and Ching-Hua Chen. 2018. Data quality challenges for person-generated health and wellness data. *IBM Journal of Research and Development* 62, 1 (Jan. 2018), Article 3, 8 pages.

[11] Patricia Conde-Clemente, Jose M. Alonso, Eldman O. Nunes, Angel Sanchez, and Gracian Trivino. 2017. New types of computational perceptions: Linguistic descriptions in deforestation analysis. *Expert Systems with Applications* 85 (2017), 46–60.

[12] Gautam Das, King-Ip Lin, Heikki Mannila, Gopal Renganathan, and Padhraic Smyth. 1998. Rule discovery from time series. In *Proceedings of the ACM SIGKDD Conference on Knowledge Discovery and Data Mining*.

[13] Luka Eciolaza, Martin Pereira-Farina, and Gracian Trivino. 2013. Automatic linguistic reporting in driving simulation environments. *Applied Soft Computing* 13, 9 (2013), 3956–3967.

[14] Steven Elsworth and Stefan Guttel. 2020. ABBA: Adaptive Brownian bridge-based symbolic aggregation of time series. *Data Mining and Knowledge Discovery* 34 (2020), 1175–1200.

[15] Albert Gatt, François Portet, Ehud Reiter, Jim Hunter, Saad Mahamood, Wendy Moncur, and Somayajulu Sripada. 2009. From data to text in the neonatal intensive care unit: Using NLG technology for decision support and information management. *AI Communications* 22, 3 (Aug. 2009), 153–186.

[16] Herbert Paul Grice. 1967. Logic and conversation. In *Studies in the Way of Words*, Paul Grice (Ed.). Harvard University Press, 41–58.

[17] Gabriela Guimarães and Alfred Ultsch. 1999. A method for temporal knowledge conversion. In *Advances in Intelligent Data Analysis*, David J. Hand, Joost N. Kok, and Michael R. Berthold (Eds.). Springer, 369–380.

[18] Zengyou He, Xiaofei Xu, and Shengchun Deng. 2002. Squeezer: An efficient algorithm for clustering categorical data. *Journal of Computer Science and Technology* 17 (2002), 611–624.

[19] Frank Höppner. 2001. Learning temporal rules from state sequences. In *Proceedings of the IJCAI Workshop on Learning from Temporal and Spatial Data*.

[20] Janusz Kacprzyk and Anna Wilbik. 2008. Linguistic summarization of time series using fuzzy logic with linguistic quantifiers: A truth and specificity based approach. In *Proceedings of the International Conference on Artificial Intelligence and Soft Computing*. 241–252.

[21] Janusz Kacprzyk, Anna Wilbik, and Slawomir Zadrozny. 2008. Linguistic summarization of time series using a fuzzy quantifier driven aggregation. *Fuzzy Sets and Systems* 159, 12 (2008), 1485–1499.

[22] Janusz Kacprzyk, Anna Wilbik, and Slawomir Zadrozny. 2010. An approach to the linguistic summarization of time series using a fuzzy quantifier driven aggregation. *International Journal of Intelligent Systems* 25, 5 (May 2010), 411–439.

[23] Janusz Kacprzyk, Ronald R. Yager, and Slawomir Zadrozny. 2002. *Fuzzy Linguistic Summaries of Databases for an Efficient Business Data Analysis and Decision Support*. Springer, Boston, MA, 129–152. DOI:https://doi.org/10.1007/0-306-46991-X_6

[24] Katarzyna Kaczmarek-Majer and Olgierd Hryniewicz. 2019. Application of linguistic summarization methods in time series forecasting. *Information Sciences* 478 (2019), 580–594.

[25] Guillaume Klein, Yoon Kim, Yuntian Deng, Jean Senellart, and Alexander M. Rush. 2017. OpenNMT: Open-source toolkit for neural machine translation. arxiv:1701.02810.

[26] Philipp Koehn, Hieu Hoang, Alexandra Birch, Chris Callison-Burch, Marcello Federico, Nicola Bertoldi, Brooke Cowan, et al. 2007. Moses: Open source toolkit for statistical machine translation. In *ACL Companion Volume: Demo and Poster Sessions*.

[27] Xuan-May Le, Tuan Tran, and Hien Nguyen. 2020. An improvement of SAX representation for time series by using complexity invariance. *Intelligent Data Analysis* 24 (2020), 625–641.

[28] Jessica Lin, Eamonn J. Keogh, Li Wei, and Stefano Lonardi. 2007. Experiencing SAX: A novel symbolic representation of time series. *Data Mining and Knowledge Discovery* 15 (2007), 107–144.

[29] Walter Maner and Sean Joyce. 1997. WXSYS Weather Lore + Fuzzy Logic = Weather Forecasts. Retrieved March 13, 2021 from https://www.researchgate.net/publication/237546595_WXSYS_Weather_Lore_Fuzzy_Logic_Weather_Forecasts

[30] Matthew J. Menne, Imke Durre, Bryant Korzeniewski, Shelley McNeal, Kristy Thomas, Xungang Yin, Steven Anthony, et al. 2020. Global Historical Climatology Network Daily (GHCN-Daily), Version 3. Retrieved March 13, 2021 from https://www.ncei.noaa.gov/

[31] Gilles Moyse and Marie-Jeanne Lesot. 2016. Linguistic summaries of locally periodic time series. *Fuzzy Sets and Systems* 285 (2016), 94–117.

[32] Soichiro Murakami, Akihiko Watanabe, Akira Miyazawa, Keiichi Goshima, Toshihiko Yanase, Hiroya Takamura, and Yusuke Miyao. 2017. Learning to generate market comments from stock prices. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics*.

[33] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: A method for automatic evaluation of machine translation. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics*.

[34] Elizabeth Peel, Margaret Douglas, and Julia Lawton. 2007. Self monitoring of blood glucose in type 2 diabetes: Longitudinal qualitative study of patients' perspectives. *BMJ* 335, 7618 (Sept. 2007), 493.

[35] Reza Rawassizadeh, Elaheh Momeni, Chelsea Dobbins, Joobin Gharibshah, and Michael Pazzani. 2016. Scalable daily human behavioral pattern mining from multivariate temporal data. *IEEE Transactions on Knowledge and Data Engineering* 28, 11 (Nov. 2016), 3098–3112.

[36] Ehud Reiter and Robert Dale. 2000. *Building Natural Language Generation Systems*. Cambridge University Press.

[37] Daniel Sanchez-Valdes, Alberto Alvarez-Alvarez, and Gracian Trivino. 2016. Dynamic linguistic descriptions of time series applied to self-track the physical activity. *Fuzzy Sets and Systems* 285 (2016), 162–181.

[38] Patrick Schäfer and Mikael Högqvist. 2012. SFA: A symbolic Fourier approximation and index for similarity search in high dimensional datasets. In *Proceedings of the 15th International Conference on Extending Database Technology*. 516–527.

[39] Somayajulu G. Sripada, Ehud Reiter, Jim Hunter, and Jin Yu. 2003. Generating English summaries of time series data using the Gricean maxims. In *Proceedings of the 9th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. 187–196.

[40] Si Sun and Kaitlin L. Costello. 2018. Designing decision-support technologies for patient-generated data in type 1 diabetes. In *AMIA Annual Symposium Proceedings*. 1645–1654.

[41] Romel Torres. 2019. Alpha Vantage. Retrieved March 13, 2021 from https://github.com/RomelTorres/alpha_vantage

[42] A. Ultsch. 1993. Knowledge extraction from self-organizing neural networks. In *Information and Classification*, Otto Opitz, Berthold Lausen, and Rüdiger Klar (Eds.). Springer, Berlin, Germany, 301–306.

[43] Chris van der Lee, Emiel Krahmer, and Sander Wubben. 2018. Automated learning of templates for data-to-text generation: Comparing rule-based, statistical and neural methods. In *Proceedings of the International Conference on Natural Language Generation*.

[44] Ingmar Weber and Palakorn Achananuparp. 2016. Insights from machine-learned diet success prediction. In *Proceedings of the Pacific Symposium on Biocomputing*.

[45] Anna Wilbik and Uzay Kaymak. 2015. Linguistic summarization of processes—A research agenda. In *Proceedings of the Conference of the International Fuzzy Systems Association and the European Society for Fuzzy Logic and Technology*.

[46] Anna Wilbik, James M. Keller, and Gregory L. Alexander. 2011. Linguistic summarization of sensor data for eldercare. In *Proceedings of the IEEE International Conference on Systems, Man, and Cybernetics*.

[47] Dongrui Wu, Jerry M. Mendel, and Jhiin Joo. 2010. Linguistic summarization using IF-THEN rules. In *Proceedings of the International Conference on Fuzzy Systems*. 1–8.

[48] Ronald R. Yager. 1982. A new approach to the summarization of data. *Information Sciences* 28, 1 (1982), 69–86.

[49] Lotfi A. Zadeh. 1975. The concept of a linguistic variable and its application to approximate reasoning—I. *Information Sciences* 8, 3 (1975), 199–249.

[50] Lotfi A. Zadeh. 1983. A computational approach to fuzzy quantifiers in natural languages. *Computers & Mathematics with Applications* 9, 1 (1983), 149–184.

[51] Lotfi A. Zadeh. 2002. A prototype-centered approach to adding deduction capability to search engines—The concept of protoform. In *Proceedings of the IEEE Symposium on Intelligent Systems*.

[52] Mohammed J. Zaki. 2001. SPADE: An efficient algorithm for mining frequent sequences. *Machine Learning* 42, 1 (Jan. 2001), 31–60.

Received March 2020; revised November 2020; accepted January 2021