ECAI 2024
U. Endriss et al. (Eds.)

doi:10.3233/FAIA240713

Big-Thick Data Generation via Reference and Personal Context Unification

Fausto Giunchiglia[a] **and Xiaoyue Li**[a,*]

[a]Information Engineering and Computer Science, University of Trento, Italy
{fausto.giunchiglia, xiaoyue.li}@unitn.it

Abstract. Smart devices generate vast amounts of *big data*, mainly in the form of sensor data. While allowing for the prediction of many aspects of human behaviour (e.g., physical activities, transportation modes), this data has a major limitation in that it is not *thick*, that is, it does not carry information about the context within which it was generated. *Context* – what was accomplished by a user, how and why, and in which overall situation – all these factors must be explicitly represented for the data to be self-explanatory and meaningful. In this paper, we introduce *Big-Thick Data* as highly contextualized data encoding, for each and every user, both her *subjective* personal view of the world and the *objective* view of an all-observing third party taken as reference. We model big-thick data by enforcing the distinction between *personal context* and *reference context*. We show that these two types of context can be *unified* in many different ways, thus allowing for different types of questions about the users' behaviour and the world around them and, also, for multiple different answers to the same question. We validate the model with a case study that integrates the personal big-thick data of one hundred and fifty-eight University students over a period of four weeks with the reference context built using the data provided by OpenStreetMap.

1 Introduction

Smart devices, e.g., smartphones or smartwatches, allow for the collection of a wide set of large-scale sensor data, e.g., GPS, Bluetooth, WIFI or accelerometer. This type of data, often referred to as (a specific kind of) *big data* [13], has been widely exploited, for instance, in Human Activity Recognition [39], Health Monitoring [36] and Autonomous Vehicles [41]. However, this type of data is often used 'out of context' and this substantially obscures its meaning and, therefore, diminishes its value [14], in particular when trying to understand human behaviour, e.g., one's social or personal life, which is always context-sensitive. *Context* – what was accomplished by a person, how and why, and in which overall situation – all these factors must be explicitly represented for the data to be self-explanatory and meaningful [23]. In particular, these factors become necessary if one wants to use the same dataset for multiple predictions, where the same sensor value may stand for two completely different contextual situations. So, for instance, a professor and a student may be in the same location, e.g., the university, with different purposes, the first for work, the second for study or because looking for a friend, while the former was in the same location during the last week-end because she wanted to collect her tennis racket which she left there on Friday. Meaningful, lifelong human-in-the-loop, human-machine interactions need this level of information richness.

To address the problem of data de-contextualization, we turn to the notion of *Big-Thick Data*. Big-thick data is big data complemented with *thick data*, that is, *observational data about context which allow to reflect upon how and why people do what they do*. We build *Observation Contexts*, as we call them, to represent big-thick data based on two main components, one or more users' *Personal Contexts* and a *Reference Context*. A *personal context*, one for each and every user, encodes the user's *subjective* view of the world, e.g., where she is, what she is doing, why she does it, who she is with, her mood [23]. Personal contexts are different for different users, also when involved in the same activity, and are also different for the same user at different times, this because of their evolving activities. We model personal contexts, in time, as *Personal Big-Thick Data* obtained by integrating *Personal Big-Data*, e.g., sensor data or data from social media, with *user-provided descriptions* of the current situation, for instance, in terms of crowd-sensing [2], human answers to machine questions [24], people's self-reports [44], or information from the phone's personal contacts or agenda [45].

A *reference context* provides a user-independent *objective* all-encompassing view of a third-party observer. It keeps track of the environment within which users are operating, defined in terms of a *reference observation period* and a *reference location*. Examples of reference observation periods are one day, one week or one year. Examples of reference locations are home, the city of Trento, or Italy. Which is the 'right' one depends on the *purpose*. For instance, the reference location could be home if the user is watching the television, the street outside if she is at the window, or Trento if she is driving to the university. Each location determines events and entities, each person entity with its own personal big-thick data. The reference context can be built out of any type of spatio-temporal (big) data, e.g., coordinates, images, labels, as from, e.g., OpenStreetMap (OSM)[1] or the Italian Spatial Data portal[2].

The *observation context* is built out of (a part of) the reference context and (parts of) one or more personal contexts based on shared *identifying information*, e.g., names, identifiers, spatio-temporal coordinates. The idea is to compose the subjective information of personal contexts into the objective perspective of the reference context. We call this process, *context unification*. We implement context unification as a flexible process, which is *configured* as a function of the specific *purpose*, as defined in [26]. Some examples of purpose are,

[*] Corresponding Author.

[1] https://www.openstreetmap.org.

[2] https://www.agid.gov.it/en/data/spatial-data.

for instance, the need of answering a specific query or the need of learning about the behaviour of a certain class of people, where the observation context may be tuned to a specific person or a group of people or to everybody we know is inside the reference location.

The main contributions of this paper are as follows:

1. The notion of *big-thick data* and its operationalization in terms of *observation, personal* and *reference context*;
2. A methodology for the *purpose-driven* generation the observation context via *context unification*;
3. A methodology, that we call *context observation*, for exploiting the observation context with multiple different purposes.

We perform a first assessment of the approach proposed via a case study where we unify a dataset, called SmartUnitn2 (SU2),[3] describing the behaviour of a large sample of university students, with a dataset generated from OSM.[4] The structure of this paper is as follows. Section 2 describes the related work. Section 3 introduces reference context and personal context. Section 4 explains context unification. Section 5 describes context observation. Section 6 provides the case study. Section 7 concludes the paper.

2 Related Work

We organize the section into *big-thick data* and *context*.

Big-Thick Data. The notion of *personal big data* came up with the explosion of digital records extracted from smart devices, mainly via sensors. While it facilitates predictive analytics that far exceed human cognitive capabilities, big data does not support human observation and interpretation in context [3]. Big data is always too poor in its contextualization to describe many interesting aspects of people behaviour, role and motivation for action [7]. This is why Bornakke and Due, in [13], talk of big data as *big thin Data*. Differently from big data, *thick data* is constructed from observations of the context in which human behavior occurs, historically, mainly in the form of people interviews and extensive self-reports [13]. The idea of thick data originated from what in Anthropology are called *thick descriptions* of the world [19], that is, ethnographically collected and analysed observational, contextually rich, *small data*. Despite the small size, thick data is still rich enough in content to enable researchers to understand and reflect upon the scenario within which people act and behave. A major problem remains which relates to the cost of generation of ethnographically generated thick data.

Recent work has suggested the integrated usage of big computational quantitative data and small embodied qualitative thick data [8, 10], with the recent introduction of the concept of *Big-Thick Blending* [13]. However, this work is mainly qualitative. The notion of *big-thick data*, as introduced in this paper, together its computational realization in term of reference and personal context, extends and operationalizes the idea of big thick blending. Figure 1, which is an evolution of Figure 2 in [13], defines big-thick data as the convergence of Big-Thin Data, e.g., Usage Analytics, Sensor data, general Internet-of-Things (IoT) data (Thing Data), Small-Thin Data, e.g., Self-Reports, Small-Thick Data, e.g., Observations, Interviews and Questionnaires, and some first examples of Big-thick data, e.g., Social Media and Ecological-Momentary-Assessment / Experience-Sampling-Method (EMA/ESM) Data.

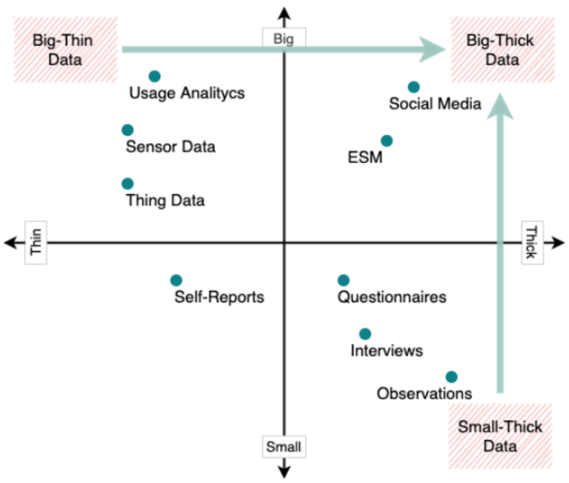

Figure 1: Big-Thick data.

Of course this is just the beginning and there are still important open problems, in particular when one focuses on lifelong, human-in-the-loop human-machine interactions. A crucial issue is how to decrease the cost while increasing the quality of big-thick data collection. For example, users often do not read, or do not answer, or provide wrong answers to machine-asked questions [6], or turn-off their data collection APP. [12] provides a general overview of how we approach the problem. [6] describes first results towards the management of long response times. [11] deals with the problem of mislabelling while [45] describes an early version of an APP which supports the user in providing meaningful answers.

Context. The notion of context has a long history, and has been studied extensively in multiple research areas. The area where it was first introduced is, as far as we know, Knowledge Representation (KR), starting from John McCarthy's Turing Award lecture [33]. Here context was proposed, together with non-monotonic reasoning, as a key element for the formalization of commonsense reasoning. In [28] context and multi-context systems (MCSs) were introduced as personal representations of real world situations, as defined by John McCarthy [34]. The idea of multiple subjective views of the same objective reality was first introduced in [20] using the example of the *magic box*, later recollected in [15]. In [20] a *view* was defined as a partial *representation* of the world consisting of as a set of facts describing the user current perspective. Later on, Brewka et al. [16] used MCSs for the representation of multiple heterogeneous knowledge sources, with information flows allowing for reasoning across multiple contexts. The notion of reference context relates to Guarino and Guizzardi's notion of *scene* [29]. The key difference, using the terminology in [29], is that a scene is a perdurant while the reference context is an endurant. The connection is in the fact that a scene can be seen as the perdurant of the reference context.

In IoT research, sensor data is exploited to enable *context recognition* from sensor time-series data, where context is then used to learn about the user behaviour. In this work, the possible context dimensions include locations, activities, body posture, and more [40]. For example, accelerometer data are used to detect physical activities, e.g., walking or running [5, 37, 40]. A Multi-Layer Perceptron (MLP) that uses multi-modal sensors (smartphone accelerometer, smartwatch accelerometer, phone gyroscope, phone audio, etc.) is described in [40] which allows to simultaneously predict many diverse context labels, e.g., people's body-states, home activities and environment. At the boundary between IoT and KR, Giunchiglia et

[3] A description of the project which generated the SU2 dataset can be found at the link https://datascientia.disi.unitn.it/projects/su2/. This page provides also information about how to download it.

[4] A description of the work described in this paper can be found at the link https://datascientia.disi.unitn.it/projects/su2osm/. This page provides a link to the OSM and SU2 datasets used in the case study.

al. have used context to model personal data streams [23, 27]. The work in this paper builds upon the previous work in KR and IOT, and in particular [23, 27], and it constitutes an attempt towards the generation of big-thick data, up to the quality which is needed in order to support meaningful human-in-the-loop human-machine interactions.

The notion of context has also been extensively studied by HCI community as 'any information that can be used to characterise the situation of an entity' [1, 17]. The underlying intuition was to use context in order to facilitate richer and easier human-machine interactions [18]. For example, the Activity River [4], a personal visualization tool, enables people to visualize historical and contextual data (e.g., activities), flexible planning and logging, etc. The relevance of this work stems from the fact that human-in-the-loop human-machine interactions need the high-quality interfaces and interactions studied and developed by this community.

3 Reference and Personal Context

Let us think of the *world* as an infinite set of, continuously evolving, three-dimensional spatial regions, that we call *locations*, and, inside each location, as an infinite set of, continuously evolving, mono-dimensional temporal regions that we call *events*. Events have a *duration* defined by a *start-time* and an *end-time*. Then, let us take the world as being populated by *entities*, e.g., people, trees, homes, cities, streets, anything which we can think as having a spatial and a temporal extension. We follow [21] and we think of entities as being identified by two key components, each with its own properties, that is *objects*, which define the spatial regions occupied by entities, and *functions*, which define a specific set of entity *expected actions*, with actions being the mechanism by which events change the world. Thus for instance, a *car* and a *bus* are two entities associated to two different objects both performing the same function of a *vehicle*, which is characterized by the action of *carrying* people around. Dually, we can also think of (the body of) a person as an object supporting many functions and corresponding entities and actions, e.g., a student reading a textbook or a driver driving a taxi. Finally, taking an example from [21], an entity implementing the function of a *chicken coop* may consist of as a little wooden house or of the body of an old car. In this latter case the same object was first associated to the entity car and the to entity chicken coop. The key observation is that there is a many-to-many relation between objects and functions, and any such combination defines an entity, not necessarily the same.

We say that, in an object is *partIn*, or *populates*, a location if it is inside the region identifying the location. Similarly, in a given moment in time, an entity is *partIn*, or *populates*, an event if the entity is inside the spatial region of the location of the event and within its temporal region. Entities are associated, for each location and event they populate with, respectively, *(spatial) coordinates* and *(temporal) coordinates*. GPS coordinates and the local time of a time zone, while not being the only possible choice, provide the coordinates for any possible triple <event, location, entity>. Notice that there is no need for locations or events to be positioned with respect to some external spatio-temporal coordinate system. The positioning is only of entities inside locations and events. This captures the intuition that, for instance, when you are at home, what you do depends only on the entities inside home, e.g, the television, and on their evolving state, e.g., the television being turned on, and not, e,g, on the location of the apartment in the city. And similarly for time.

People are entities which have an internal *representation* of the world and use it to reason about it and take action. Following [28], we assume that this mental representation is organized in *contexts* where (quote) '*... a context is a theory of the world which encodes an individual's subjective perspective about it*'. According to [28], contexts are not *situations*, where, following [34], (quote) '*... a situation s is the complete state of the universe at an instant of time*'. In other words, people have partial, possibly incorrect, views of the world while situations establish what is the case, thus providing a single reference point for comparing the contents of contexts. However, situations are not accessible, we can only build mental representations about them. In the following we call *personal contexts* all those contexts, carrying a subjective representation of the world, which satisfy the definition of context from [28], while we call *reference contexts* all those contexts carrying an *objective* representation of situations, as defined in [34]. Notice that here we talk about subjectivity and objectivity in a somewhat limited form, with reference to the fact that people have *partial knowledge* of the world, thus leaving out issues related to the subjectivity of, e.g., opinions or sentiments. Thus, we say that we have *objective knowledge* if everybody in the target audience knows or has the means for knowing about it. Dually, *subjective knowledge* is known only by a few specific subjects while the others may or may not know about it, in the latter case, being able to get to know about it only if told by those who know about it. Thus for instance, my home address is personal knowledge, while the name of the street where I live is objective knowledge.

In the following, we first introduce the reference context C_R (Section 3.1) and then the personal context C_P (Section 3.2).

3.1 The Reference Context

We wrote above that reference contexts carry '... *an objective* representation of situations' and *not* that they carry '... *the objective* representation of situations'. This is because it is impossible to build a complete description of reality. Two objective representations of the same situation may differ in many dimensions, for instance, the level of detail, the level of partiality, the view point, the entities being considered, and so on [21]. We follow an approach where we build the reference context based on a specific *purpose* where, following [26], loosely speaking, the purpose is connected to the specific target use, e.g. answering a specific query or recognizing a specific action. The process is similar to when one has the need of asking multiple questions, for instance: 'what are the people involved in the current event doing now and what will they be doing tomorrow, when in the same location?', 'does the current location usually host similar events?', and, as a result, she focuses on different fragments of her knowledge.

We proceed as follows. We start by choosing a *Reference Spatial Region S*, defined as the set of points (x, y, z) located inside the *boundary* of S, where the boundary defines the *inside / outside S*, with reference to a bigger location that is not considered because not purpose-relevant. Let us assume that S contains a set of spatial regions, perceived as *objects*, $O_1, ..., O_n$, themselves defined by a boundary, an inside and an outside [21], that is

$$O_i \subseteq S \quad \text{with} \quad 1 \leq i \leq n, \tag{1}$$

Let us concentrate on a *Reference Observation Period ΔT_R* where $t \in \Delta T_R$ measures how change happens within S. Then we define a *Spatio-temporal Context C_S* as:

$$C_S(t) = \langle S(t), \{O(t)\}_S \rangle \quad \text{with} \quad t \in \Delta T_R \tag{2}$$

where $\{O(t)\}_S$ is the set of objects O_i satisfying Eq. (1). Intuitively, Eq. (2) tells us that C_S consists of a set of objects located within S, and that both S and the objects O_i's change during ΔT_R. Time variance is a fact of life. Everything continuously changes. This is a major source of subjectivity as the same object looks different at different times. We deal with this problem by requiring that, during ΔT_R we have *Time Invariance.* That is, ΔT_R must be such that the *Reference Location* L_R, that is, the location associated to the spatial object S, and the selected objects in L_R do not change, that is, they keep the same *selected spatial properties*, e.g., position, shape and color. Period, objects and spatial properties are selected as a function of the purpose. We move from C_S to the *Reference Context* C_R as follows. Let e_i^j, with $j = 1, ..., m$ be the j-th entity associated to the object O_i. Then we have:

$$C_R = f(C_S(t)) = \langle L_R, \{e\}_R \rangle \quad \text{with} \quad t \in \Delta T_R \tag{3}$$

where: C_R is time-invariant, $\{e\}_R \subseteq \{e(O)\}_S$ is the set of entities of C_R, with $\{e(O)\}_S$ the set of the entities associated to the objects of C_S, and f is a projection function from C_S to C_R enforcing *objectivity*. The definition of f is up to the modeler, with the proviso that it must satisfy the following set of constraints.

From objects to entities. L_R and $e \in \{e\}_R$ should be chosen to fit the purpose. For instance the region S associated to the Trento spatial region can be thought of as the location represented by a geographical map, or by an administrative map. Similarly the object corresponding to (the body of) a person can be thought, e.g., as a father or as a professor;

Completeness wrt. the users'. C_R should provide enough detail to ground the different subjective views of all the users. Users should be able to determine whether their representation is consistent with that of the reference context and that of any other user;

Localization. C_R should describe the smallest possible location satisfying the previous properties.

Some observations. The first is about how *objectivity* is enforced. *Time invariance* allows to generate shared knowledge which does not change in time. This is a very robust form of objective knowledge, the easiest to manage and scale. This may seem a strong requirement, but notice that most spatial entities, e.g., streets, cities, monuments, the furniture in an apartment, change very rarely. Plus, this approach can be extended to manage what one could call *objective events*, that is events everybody knows about, for instance a concert which has been organized long before its occurrence, and extensively advertised so that everybody knows about it. The idea is to decompose the duration of the observation period ΔT_R into a sequence of smaller periods, each corresponding to a time independent C_R, via a time-aware versioning mechanism. Furthermore, and this is the second key element towards the enforcement of objectivity, when moving *from objects to entities*, only the purpose-relevant functions and actions are selected. Thus, for instance, for the city of Trento we may have multiple reference contexts, one providing information about moving around, one about health related of university related facilities, one about points of interest, and so on [9]. This means, furthermore, allowing for the computation of the relevant subset of spatial relations among the objects in L_S. Some examples are: positioning (e.g., via coordinates), relative positioning (e.g., *Right* or *Above*), proximity, reachability, color, or shape. The request of *Completeness wrt. the users'* allows to enforce a general mechanism for the comparison of the user contexts. It allows the reference context to take the role of an oracle capable of deciding what is true and what is false among the facts stated inside personal contexts. *Localization* is a key requirement for the reference context to work in practice. Thus, for instance, if the focus is what is happening at home, then the reference context should not include entities which are outside home.

3.2 The Personal Context

People are some among the entities inside L_S, each of them with their own unique subjective view of the world, that we formalize as their own personal context C_P. Compared to C_R, C_P has a few distinguishing features, as follows. Let me be a generic person.

$C_P = C_P(me, C_R)$. C_P depends on both me and C_R. Different C_R's may generate different C_P's even for the same me and S;

From inside to outside. While C_R describes the entities which are *inside* S, C_P describes the entities which are *outside* the object of me, still *inside* S;

From no change to change. Differently from C_R, C_P considers also entities which change in time.

We define the *personal context* C_P of me,given C_R, as follows:

$$C_P = \langle C_R, \{E(L_R, t)\} \rangle \quad \text{with} \quad t \in \Delta T_R \tag{4}$$

where $E(L_R, t)$ is a time-varying *event as perceived by* me, involving me, and occurring inside L_R, with $\{E(L_R, t)\}$ being a set of such events. Any two events may occur in sequence or in parallel. These events model how situations, as subjectively perceived by me, evolve in time. Notice how Eq. (4) is the same as Eq. (3) when one substitutes a set of unchanging entities with a set of continuously changing events. This is the formalization, and generalization to a time-variant real world setting, of the idea of *subjectivity as view point* of the *magic box* [20]. That is, there is an unchanging fully known objective reality, modeled by C_R, and multiple time-varying partial perspectives of this reality, each modeled by the set of events in which me is involved.

The first source of subjectivity of $E(L_R, t)$ if the location L where it occurs, with usually $L \subseteq L_R$, and its time interval ΔT, with usually $\Delta T \subseteq \Delta T_R$. Thus, for instance, if L_R is a city, then me may be driving in some street or may be eating at home. Being in a location at a certain time, me does not have access to what is going in the other locations and partially also in that location, if big enough. We capture this request by refining Eq. (4) into Eq. (5) below.

$$C_P = \langle L_R, \{E(L, t)\} \rangle \quad \text{with} \quad t \in \Delta T \subseteq \Delta T_R,\ L \subseteq L_R \tag{5}$$

The second and most important cause of the subjectivity of events involving me is that, as from above: (i) events are the result of the interactions among entities; (ii) these interactions happen because of the actions of entities; and (iii) these actions are motivated by the entities' mutual functions. Thus, for instance, me_1 can be in the *car* of *her friend* me_2, *in front* of the *church*, while *talking* to a *friend* me_3. We model this form of subjectivity by defining $E(L, t)$ as follows.

$$E(L, t) = \langle L, \{\langle e, F_e, A_{F_e}(t) \rangle\} \rangle \text{ with } t \in \Delta T \tag{6}$$

where: $\Delta T \subseteq \Delta T_R$, $L \subseteq L_R$, and where

- $e \in \{e\}_E$ is any *entity*, including me, involved in $E(L, t)$, with $\{e\}_E \subseteq \{e(O)\}_S$ the set of such entities, $\{e\}_E \cup \{e\}_R \neq \emptyset$;
- $F_e = \{f\}_e$ is the set of functions f of any $e \in \{e\}_E$ with respect to any $e \in \{e\}_E$;
- $A_{F_e} = \{a\}_{F_e}$ is the set of actions a of any entity $e \in \{e\}_E$ towards any other $e \in \{e\}_E$, because of a function $f \in F_e$.

We can now merge Eq. 5 and Eq. 6 to obtain the following final characterization of the subjective context C_P of me:

$$C_P = \langle L_R, \{\langle L, \{\langle e, F_e, A_{F_e}(t) \rangle\} \rangle\ \} \rangle \tag{7}$$

with $t \in \Delta T \subseteq \Delta T_R$, $L \subseteq L_R$. That is, looking at the pairs $\langle ... \rangle$ in Fig. 7, the subjective context of me is constructed as follows:

1. select a previously identified spatial region S, for instance the home of me, a museum, the university, or a city;

2. select a previously identified reference context C_R, that is, observation period ΔT_R, reference location L_R and entities $\{e\}_R$;
3. then select a set of me's, each with a corresponding personal context $C_P(me, C_R)$; and,
4. for each C_P select a set of events, each event with its own set of entities and corresponding, objects, functions and actions.

where the four steps above are performed based on a given purpose. Some observations. The first is about how *subjectivity* is modeled. As from above, usually me's do not know about the other me's, this because of their different locations and time periods. But this applies also to the entities involved in the same event. In fact, the functions and also the actions relating people - and entities in general - to one another, are unknown to most me's. Subjectivity arises because of the *diversity of people* and because of the partial knowledge that any me has of the other me's. This is the key intuition behind the idea of big-thick data. The only way to know about context is to ask people. Big data cannot provide information that thick data provide.

The second observation is that we have assumed that, inside events, locations, functions and therefore entities are time-invariant. Their time variance, has to be managed by splitting an event in two or more, following a process similar to that of time-variant C_R's.

The third observation is about *locations*. The same spatial region R can play the function of entity or location. Thus for instance, if me is driving home, then home is the entity that me needs to reach. It is outside the space occupied by me and both me and home are inside the same location, e.g., the city of Trento. However, when me is at home, home is the smallest location outside me. Similarly, when me is in the car, the car is the location where the driving event occurs but it is also the entity taking me home. Any object can act as location: a person's body is the location where COVID-19 operates, a piece of sheet is the location where me is writing, and so on. The discriminating factor is the space granularity of the event.

The fourth observation is about *entities* $e \in \{e\}_E$ and their functions. Any e may be concurrently involved in multiple events, usually from different me's, usually with different functions. The functions of the entities $e \in \{e\}_R$ are known by all me's, those of the entities $e \notin \{e\}_R$, instead, are known only to some me's. Thus, for instance, if me_1 meets me_2 at the University, me_1 will know that the University is her own study place and will not know that it is the work place for me_2, and vice versa.

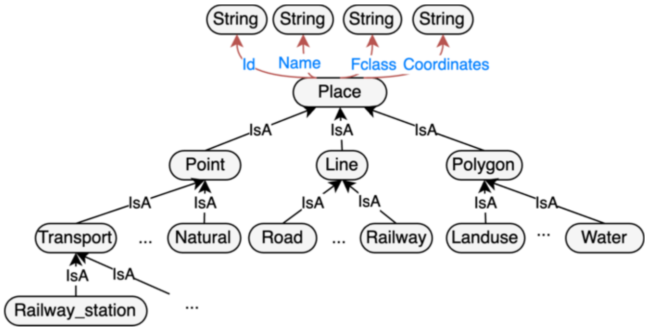

Figure 2: The OpenStreetMap Hierarchy.

The fifth observation is about *actions* $\{a\}_{F_e}$. Differently from functions, actions change frequently in time, with duration $\Delta t \subseteq \Delta T$, and this is why Human Activity Recognition (HAR) is hard. Big-thick data provide extra information, in particular, with respect to the functions that actions are supposed to carry out, see, e.g., [44].

Last but not least, so far we have talked of the functions and actions which characterize how an entity interacts with the outside. However, the behaviour of entities, and humans in particular, is largely influenced by their internals. This is why big-thick data often carry information about people's *internal functions and characteristics*, e.g., personality and procrastination syndrome, that we assume stable in time, and *internal actions*, e.g., mood and tiredness, with usually relevant temporal dynamics [12].

4 Context Unification

Following [23], we represent contexts as Knowledge Graphs (KGs) [30, 31]. Let us consider C_R, see Eq. (3). We exemplify the process of building C_R by formalizing the fragment of the OSM hierarchy in Fig. 2.[5] In Fig. 2, the upper part of the hierarchy, down to the level of `Point`, `Line` and `Polygon`, describes the three types of geometrical features of `places`. Each such feature, in turn, is further refined into a sub-hierarchy of max depth 4, as in `Place < Line < road < major_road < motorway`. The attributes of `Place` must be interpreted as follows: `Id` is its identifier; `Name` is its local name; `Fclass` is the property used to name its class; `Coordinates` is used to store its spatial coordinates.

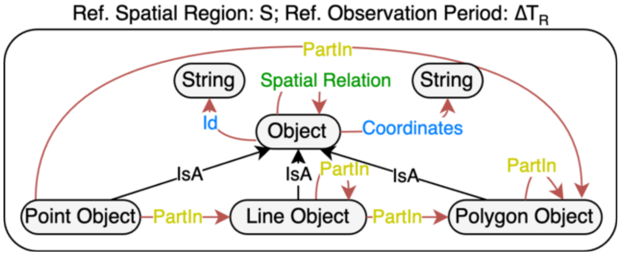

Figure 3: The C_S Teleontology.

We proceed in steps, as follows. If one looks at Fig. 2, it should come intuitive that the upper part encodes the geometrical features of objects and, as such, it should be represented in C_S, while the rest describes the properties, mainly functions, of entities and, as such, should be represented in C_R. Let us start from C_S. The first step is to produce the *Space (Context) Teleontology (STLO)* [6], as represented in Fig. 3. STLO is a KG, representing a generic C_S, as from Eq. (2), where nodes are *object types* (with the added *datatype String*), i.e., sets of objects associated with a set of properties. Nodes are connected by three types of links: (i) the subsumption relation *IsA* rooted in the type *Object* denoting all objects; (ii) *Object Properties* linking two object types, i.e., *Spatial Relation* and *PartIn*, and (iii) two *Data Properties*, i.e., *Id*, *Coordinates*. STLO has four main purposes: (i) to represent all the objects of C_S; (ii) to represent their selected spatial functionality i.e., *Point (Object)*, *Line (Object)* and *Polygon (Object)*; and (iii) to represent spatial containment, i.e. *PartIn*, as from Eq. (1), and (iv) spatial relations, see Section 3. *PartIn* is applied recursively, thus allowing to define locations, as polygons or lines, at any level of spatial granularity, where locations, at the end of the recursion, are populated by point objects. The property *Spatial Relation* can be specialized to more refined relations, e.g., *Near*, *Right* or *North*. Both *PartIn* and *Spatial Relation* can be computed from OSM. Finally, the box around the Teleontology KG represents the region S and observation period ΔT_R of C_R. Implementationally, the information about boxes (all of them, see also below) is encoded as metadata.

[5] Fig. 2 is depicted with reference to the OSM Layered GIS Format, see https://download.geofabrik.de/osm-data-in-gis-formats-free.pdf.

[6] Following [21], the meaning of the word *teleontology* builds on the Greek words *telos* (meaning *end, purpose*) and *logia*, (meaning *a branch of learning*). We use the word teleontology (and teleology, see below) to capture the intuition that a a teleontology is written with a *purpose*. There is no claim of generality beyond the purpose for which it is generated.

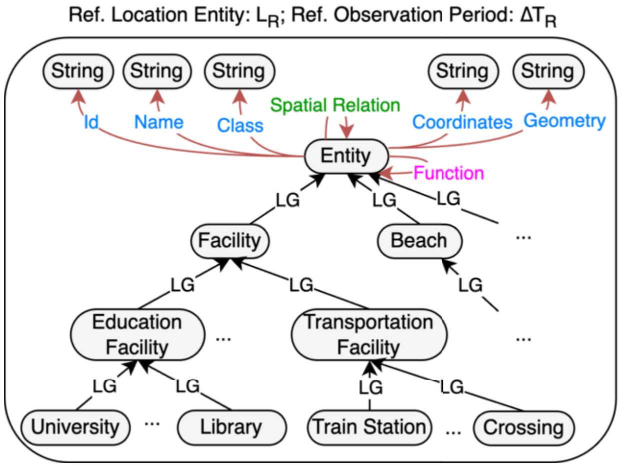

Figure 4: The C_R Teleontology.

Fig. 4 depicts the *Knowledge (Context) Teleontology (KTLO)* representing a generic C_R as derived from STLO in Fig. 3. KTLO is a 'standard' *More/Less General (MG/LG) hierarchy* where nodes are *entity types (etypes)* with root the etype *Entity* and where the lower etypes inherit the properties of the higher etypes. There are three observations. The first is that the the root of KTLO and, therefore each entity, inherits the spatial properties of the root of STLO, where the data property *Geometry* defines whether it is a point, a line or a region. These properties are additional with respect to the ones defining the function(s) of entities, e.g., the type of facility, see the property *Function* in Fig. 4. The second observation is that a point object may be modeled as an entity, a polygon object as a location (i.e., as the label of the surrounding box). Thus L_R, the label of the box in Fig. 4, can be, e.g., *home*, *street* or *city*. An entity with both types of geometry can act as either an entity or a location. The third is that entities are associated a *Name*, and a *Class*. We name entities and not regions, and the same region may take two different names if it changes its function (see above the *chicken coop* example). *Class* is our term for `Fclass` in OSM. The *Class* of an etype can take as value the etype itself or more specific one, for instance we want to say that, e.g., a specific *Facility* is a *Pub*.

Finally, we can now represent C_R as an *Entity Type (eType) Graph (ETG)* [25], that we also call a *Teleology*. ETGs are obtained from KTLO's as follows: (i) select a subset of the etypes of the KTLO; (ii) for each etype, select a subset of its object and data properties, and, finally (iii) eliminate the *LG* relation in KTLO and, for each type, distribute its properties to the lower etypes. The generation of an ETG from a KTLO is the process of selecting what is relevant to the specific purpose, where, here, the purpose is to build an ETG which allows for context unification with the C_P of one or more me's. A fragment of the ETG resulting from Fig. 4 is reported in Fig. 5(a). Fig. 5 reports the ETG of C_R (Fig. 5(a) on the left) together with the ETG of one C_P (Fig. 5(b) on the right) which need to be unified.

The C_P ETGs are constructed following the same process, described above for the C_R ETG (from STLO to KTLO to ETG), with a few key differences that we describe below. Let us compare C_R and C_P ETGs in Fig. 5. The first observation is that C_P and C_R share the same location entity *City*, this because we have assumed that the event does not occur in a sub-location $L \subseteq L_R$. Still we have $\Delta T \subseteq \Delta T_R$. A complete visualization of C_P could have been that me was first reading in the *Library*, then walking towards the *University*, and then taking a class inside the *University*. C_R and C_P share some entity types, e.g., *Library*, namely those entities which, inside C_R are relevant to the activities of me. They also share *Spatial Relations*, and *Coordinates*, which is what allows to position me inside C_R while she moves around. The key difference is that the spatial relations involving me are time-dependent. Both C_R and C_P involve functions, some of which, e.g., *StudyPlaceOf*, define the function of an entity in C_R with respect to me. Notice how functions are triples <entity, function, entity>. Differently from C_R, C_P contains also actions, both external, e.g., *TalkTo* and internal, e.g., *Mood*. We also model actions as triples but, differently from functions, actions have a time duration $\Delta t \subseteq \Delta T$ that sometimes reduce to timestamps.

Given the $n + 1$ EGs the final step is to populate them with the available data, for instance using the approach described in [25, 9]. In the case study in Section 6, C_R is populated with OSM data about *Trentino*, the region where Trento is located, the C_P's with SU2 data. The result is one C_R *Entity Graph (EG)* and n C_P EGs, one each each me. EGs are KGs where nodes are specific entities, e.g., the buildings of Trento or a specific person associated to a me, each associated with its own etype, and links are the properties of the corresponding ETG. Intuitively, EGs are built from ETGs by expanding each and every etype node into all the entities of that etype, e.g., *Restaurant(Biba's)*, *Library(BUC)*, and *me(User73)*, and by adding one link for each specific instantiation of a property. Some observations. The first is that the entities can be associated with spatial relations, e.g., *Near(UniTn, BUC)* in C_R or *Near(User73, User45,* Δt*)* in the *User73*'s EG. The second observation is that, as a consequence of the fact that actions are tagged with a timestamp, for each me we have a timed sequence of C_P EGs, one for each selected duration period. That is, for each me, we have a *stream of Spatio-Temporal* EGs.

Given one reference context and n personal contexts, the next step is to *context-unify* them and build the *Observation Context* C. Let us assume that the EGs of C_R and the C_P's, as represented in Fig. 5, have been constructed, Then, we have

$$C \;=\; C_R \uplus \{C_P\} \tag{8}$$

where $\uplus$ is the *context unification operator* and $\{C_P\}$ is the set, one or more, of personal contexts under consideration. Context unification operates in two macro steps, as follows.

- Unification between C_R and $\{C_P\}$, one C_P at the time;
- Pairwise unification between any two C_P's after they are unified with C_R.

The order of unification is motivated by the fact that the first step objectifies, for what is possible, the contents of C_P with respect to their position and also the functions relating the entities in C_R with the entities in a C_P. This allows to obtain new properties, such as, e.g., *Near(User73, Biba's,* Δt*)*, *RestaurantOf(User73, Biba's,* Δt*)*. Notice that we perform the unification of two personal contexts always with respect the reference context.

Context unification exploits three specific types of unification, as follows.

- *Etype* and *Property Unification (EPU)*. This is typical problem of *ontology / schema alignment*; we follow the approach described in [22, 38];
- *Spatio-Temporal Unification (STU)*. Here the spatio-temporal coordinates of entities are exploited. We have two types of results. The first is the recognition of two entities, for instance two me belonging to two different C_P's, as being the same entity. For this to be the case, the coordinates of the two entities must be the same, modulo approximations, at all times. The second is the spatial relation holding at a certain time between entities, see the examples above;

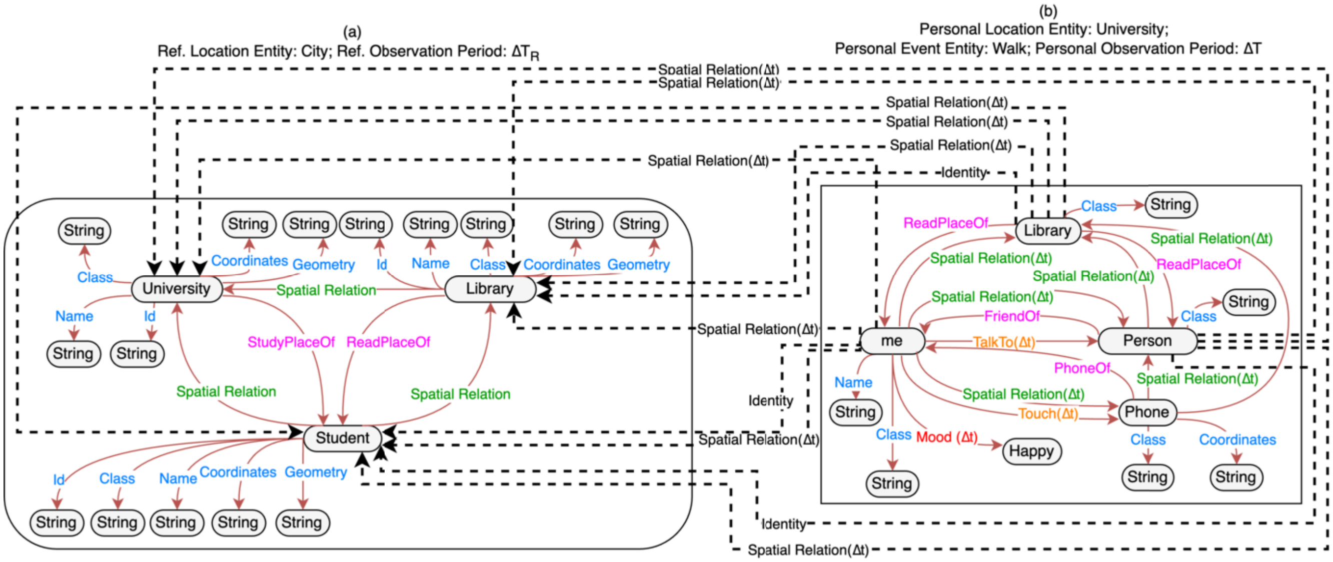

Figure 5: (a) C_R Teleology, and (b) C_P Event Teleology, unified.

- *Entity Unification (EU)*. This is done using specific entity properties, different from spatio-temporal properties, mainly *Name* and *identifier*, but also entity specific attributes [25]. For instance, if available, information about the phone number.

Fig. 5 shows the possible unifications between C_R (Fig. 5(a) and one C_P in Fig. 5(b). For instance, we could learn that *User73*, the specific me in Fig. 5(b) is in *Trento* in certain period, and near *BUC* at a certain hour, that he is talking to *User72* half an hour later, that the two are friends, and so on.

5 Context Observation

The context unification process leaves full flexibility in the selection of the reference context, of the personal contexts, and also what to unify. The result is a spatio-temporal EG which can then be *enquired* in many different ways, for instance, it can be queried like any other KG [27], it can be used to do statistical modelling and reasoning [6], or it can be used to do machine learning, for instance for the machine to learn about the user and thus to enable high quality high value human-machine interactions [11]. Because of this, the approach proposed in this paper can be seen as defining a general purpose *meta-process* which can be used to build big-thick data for the desired purpose. The case study in Section 6 provides a relatively large example of how big-thick data can be generated and then used to do prediction. In this perspective it becomes relevant to classify the possible *observation purposes* into four main groups, as a function of what one is interested in observing. We have the following.

- *Reference (R) enquiries*. The goal here is to know the details of the reference context, for instance as someone would do when getting to a new place for the first time and being in need of finding her way around. R-enquiries are posed *only* to the reference context, independently of the dynamics which may occur inside it. Thus, for instance, possible questions which could be posed to the observation context built in Section 6 are: 'Where is the bus stop near to the bar named Bar Sport?' or 'What are the supermarkets near BiBa's?'.
- *Personal (P) enquiries*. The goal here is to know about what people have done in certain period of time, including also their subjective view of what happened. P-enquiries are only to a single C_P (or streams of C_P's of the same me, see below) with no possibility of reference to entities of C_R which are not part of C_P. For instance, in the case study in Section 6, possible questions which could be posed are 'Which places did me go in a certain period, what she did there, and what was her mood?'. The answers are associated with me's subjective locations (e.g., shopping place), events (e.g., shopping), social interactions (e.g., a seller) and moods (e.g., happy).
- *Personal-Reference (PR) enquiries*. The goal here is to explore how the inside of the reference context evolves as a function of the entities which populate it within a certain period of time. PR-enquiries are posed to the observation context C focusing on C_R and are about its state as a function of the activities of one or more C_P's. Some examples from the case study are: 'How many people are in the Biba's restaurant during week-ends?' and 'How many attractions in Trento have involved me during the last week?'.
- *Reference-Personal (RP) enquiries*. The goal here is to explore the environment around me's and its impact on me. These are enquiries about one or more me posed to the observation context C. Some examples are 'What was your mood when you were in the Coop supermarket?' and 'which friends of me were in the Biba's restaurant during the dinner, yesterday?'.

Roughly speaking, R-enquiries correspond to the 'typical' problems dealt with using big data, while P-enquiries relate to the 'typical' problems dealt with using thick data. PR-enquiries are new types of enquiries, enabled by big-thick data, where one can study how specific elements of objective knowledge can be enriched by the subjective knowledge provided by multiple people, as a function of their behaviour and subjective perspective. Similarly, RP-enquiries are new types of queries, enabled by big-thick data, which allow us to put the subjective behaviour of people under the objective lenses of a third party, possibly also providing an cross-individual inter-subjective view of the world.

6 Case Study

We generate big-thick data by unifying OSM big data with the SU2 thick data. We first build the C_R EG, then the C_P EGs, then the observation context EG, which we then enquire.

6.1 The Reference Context EG

The Geofabrik site[7] provides geographical information worldwide about physical places, that we call OSM `Places`, each associated with multiple features and classes. OSM `Places` are associated with various data properties, some shared by all `Places`, e.g., `id`, `Name`, `Fclass` and `Coordinates`, some others associated only to specific `Places`, e.g., `type` is a property of `Building`, with values, e.g., `apartment` or `church`. We construct the *OSM-Trentino* dataset, in SHP format, by constraining the places' coordinates to lie within the Trentino maximal and minimal latitudes and longitudes.

We construct the C_R from the *OSM-Trentino* dataset using the properties mentioned in the previous sentence. The C_R reference observation period ΔT_R is the period of the SU2 data collection experiment, that is, four weeks from 05-08 22:02:19 to 06-06 21:51:22, where the year is removed for privacy reasons. We define place entities, one for each OSM `Place`, and we compute spatial relations among them, e.g., *Near*, from their coordinates. For instance, a fragment of the *Trentino* C_R EG which described the restaurant of name *Biba's* is represented in Fig. 6(a).

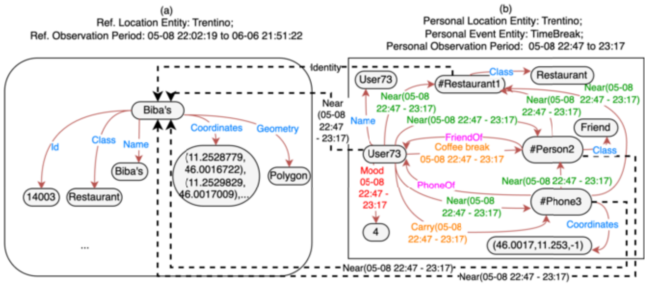
(a)
Ref. Location Entity: Trentino;
Ref. Observation Period: 05-08 22:02:19 to 06-06 21:51:22
(b)
Personal Location Entity: Trentino;
Personal Event Entity: TimeBreak;
Personal Observation Period: 05-08 22:47 to 23:17
Biba's
Id
Class
Name
Coordinates
Geometry
14003
Restaurant
Biba's
(11.2528779, 46.0016722), (11.2529829, 46.0017009),...
Polygon
...
Identity
Near (05-08 22:47 - 23:17)
User73
#Restaurant1
Class
Restaurant
Near(05-08 22:47 - 23:17)
Name
Near(05-08 22:47 - 23:17)
Near(05-08 22:47 - 23:17)
Near(05-08 22:47 - 23:17)
Friend
User73
FriendOf
Coffee break 05-08 22:47 - 23:17
#Person2
Class
Mood 05-08 22:47 - 23:17
PhoneOf
Near(05-08 22:47 - 23:17)
Near(05-08 22:47 - 23:17)
#Phone3
4
Carry(05-08 22:47 - 23:17)
Coordinates
(46.0017,11.253,-1)
Near(05-08 22:47 - 23:17)
Near(05-08 22:47 - 23:17)

Figure 6: An Observation Context EG from the SU2OSM Dataset.

6.2 The Personal Context EGs

The SU2 dataset[8] consists of phone sensor data, time diaries and surveys collected, during a period of four weeks, from 158 students of the University of Trento, using the iLog APP [42, 45] running on their phones. The SU2 dataset is suitably anonymized, it abides by the General Data Protection Regulation (GDPR), and has been utilized in numerous case studies, see, e.g., [43, 44]. The time diaries used in the data collection include various HETUS questions[9]. The questions used in this case study are: *Where*: 'Where are you?'; *What*: 'What are you doing?'; *WithWhom*: 'With whom are you?'; plus, additionally, one question on *Mood*: 'What is your mood?'. The SU2 dataset also includes a wide range of sensor data. Here we use only the GPS. Plus we do not use the information collected during the surveys, before and after the data collection. To summarise, in this case study we use the following data from the SU2 dataset:

- *userid*: the participant identifier, valuing integers from 0 to 157 (the data of some students are considered because the quality was too low);
- *where*: the answer of the `Where` question, a total of 17 possible answers, including 'Home', 'Classroom / Study hall', etc;
- *what*: the answer of the `What` question, a total of 23 possible answers, including 'Sleeping', 'Eating' and 'Studying', etc;
- *withWhom*: the answer of the `WithWhom` question, a total of 9 possible answers, including 'Friend(s)', 'Classmate(s)', etc;
- *mood*: the answer of the `Mood` question, valuing integers from 1 to 5. Higher values mean more positive mood;
- GPS coordinates, i.e., *latitude*, *longitude*, and *altitude*;

GPS data, questions and respective answers are time-stamped, where one timestamp has form 'mm-dd hh:mm:ss'. The frequency of questions is every 1/2 hour or every hour depending on the week. The GPS data has an estimated frequency of once a minute.

We construct the C_P's as follows. We have one me for each participant. This data is formalized into the C_P Event Teleology in Fig. 5(b) in the obvious way, but with a few twists which take into account the specificity of the SU2 dataset. During the overall four weeks of the data collection there were a range of $1063 - 1067$ question batteries per user, $168,095$ in total, each battery involving the questions listed above (and more). We associate to each such battery a C_P associated to a single event $E(L)$ of duration ΔT, the same for both C_P and $E(L)$. We assume that ΔT is the interval time between two questions, viz., half an hour and one hour in the first and second two weeks, respectively, centered in the time of the question. The result is a total of $104,414$ C_P's (participants did not always answer). organized into 158 C_P *streams*, one for each me, with length in the range of $371 - 875$ EGs. We call any element of the stream a *timed EG (or* C_P*)*.

To populate C_P's, one point of attention relates to the answers of the questions *What*, *Where*, and *WithWhom*. The answers to the first question can be directly encoded as actions. However, the encoding of the answers of the last two questions is a little more elaborated. In fact, because of its intended use (mainly within the Social Sciences), HETUS is designed to allow for generic answers, what in the C_P we encode as etypes. Thus, for instance, the me of name *User73* might answer that she is with a friend. Which means that me is with a person entity, whose name is unknown. Thus, we translate the above two answers in the following triples: *FriendOf(#Person2, User73)*, *WithWhom(User73, #Person2,* Δt*)*, *Near(User73, #Person2,* Δt*)*, *Near(User73, #Restaurant1,* Δt*)*, where these anonymous entities have the obvious etypes. The time parameter Δt encodes when that time-variant spatio-temporal property holds. Figure 6 (b) reports a fragment of a *TimeBreak* event involving the participant *User73*. A second point of attention relates to the spatio-temporal position of entities. We computed the coordinates of phones by considering the coordinates in a time window of 10 minutes around the question time. Then, we apply DBSCAN [35] and identify the largest cluster computed by the algorithm. The coordinates are the mean of the coordinates of this cluster.

6.3 EG Unification

Given the contents of C_R, the purpose is to get to know about the physical places where SU2 events occur. To achieve this, we unify the single OSM-Trentino EG with the streams of the SU2 EGs of the 158 me's into the observation EG, that we call SU2OSM. Fig. 6 reports a fragment obtained from the union of the two EGs introduced before. The process proceeds in three steps as follows.

The goal of the *first step* is to perform EPU unification (see Section 4). We have performed this step manually given that we are not interested in full automation. However the task is quite straightforward, given the limited scope of the etypes and properties of the C_R and C_P ETGs. Given the C_R and C_P KTLO's, this task is largely within the reach of the algorithm described in [38]. The *second step* is to establish, using STU unification, inside each timed C_P, the holding of spatial relations between C_P phone entities and C_R place entities. We focus on the spatial relation *Near*. This is achieved, using their

[7] https://download.geofabrik.de

[8] For detailed information about the *SU2* dataset and the process by which has been generated, look at the link reported in footnote[3].

[9] More information about the *HETUS (Harmonized European Time Use Surveys* standard can be found at https://ec.europa.eu/eurostat/web/time-use-surveys.

Table 1: Prediction enquiries.

Prediction Enquiry	Dataset	Purpose Feasibility	Prediction Experiment						
			Target Property	Feature Properties	Best-performance Algorithm	Accuracy	Recall	F1 Score	AUC
E1: Does a place is classified as a residence?	OSM-Trentino	✓	type	name, class	Random Forest	78.23%	0.529	0.491	0.529
	SU2	×	-	-	-	-	-	-	-
	SU2OSM	✓	type	day_of_week, time_of_day, name, class	Random Forest	82.51%	0.714	0.728	0.825
E2: Is a user at a living place?	OSM-Trentino	×	-	-	-	-	-	-	-
	SU2	✓	where	what, withWhom, mood	Decision Tree	87.43%	0.850	0.855	0.930
	SU2OSM	✓		name, class, what, withWhom, mood	Random Forest	94.42%	0.846	0.878	0.949
E3: Is a user in a bank?	OSM-Trentino	×	-	-	-	-	-	-	-
	SU2	×	-	-	-	-	-	-	-
	SU2OSM	✓	class	what, where, withWhom, mood	Decision Tree	90.91%	0.731	0.736	0.923

Note 1. ✓ and × indicate whether E1, E2 and E3 succeed or fail, respectively, in satisfying the purpose feasibility metric of the selected dataset.
Note 2 (E1): *target property* holds if the property *type* (as from OSM) has values ‘apartments’ or ‘house’ or ‘residential’, it does not otherwise; the *day_of_week* and *time_of_day* features are labeled as the weekday (from Monday to Sunday) and the time periods (morning, afternoon, evening and night) based on the time of user answers.
Note 3 (E2): *target property* holds if the property *where* has values ‘Home’, ‘Relatives Home’ or ‘House (friends, others)’, it does not otherwise.
Note 4 (E3): the *target property* holds if the property *class* has value ‘bank’, it does not otherwise.

coordinates, by calculating the distances between phone entities and the closest OSM-Trentino place entities. We assume that two entities are one near the other if their distance is less than 50 meters. For instance, in Fig. 6, *Near(#Phone3, Biba's)* is the result of computing the distance of *Biba's*, the OSM-Trentino place closest to *#Phone3* as being approximately 8.2 meters. This allows us to derive which phone, place and person entities are close to one another. The *third step*, performed using EU unification informed by STU unification, is to establish which specific C_R place entity is the one where me is. The idea is to select among the place entities which are close to me, the one with the proper etype. This allows us to identify a generic C_P place entity, e.g., *#Restaurant1* in Fig. 6, as being a specific C_R entity, which in Fig. 6 is the restaurant of name *Biba's*. In case of multiple places of the correct etype near the phone, we select the closest. Following the process described above, we have recognized a total of 147 out of the $483,981$ OSM-Trentino entities, many of which matched more than once during the four weeks, thus highlighting the student *habits*. In total, 1955 C_P EGs have been unified with the C_R EG (that is, 1.87% of the total number of timed C_P) out of which we have computed 7820 relations linking the C_R EG to the C_P EGs. Notice how this would enable the unification of entities across C_P EGs, allowing us, for instance, to establish when two different me were in the same location at the same time, or how much time a single me spent in the same place in certain periods (e.g., the morning of a specific time).

Table 2: OSM-Trento, SU2 and SU2OSM dataset size.

Dataset	Storage Volume	CR EG	CP EG Streams	Unified EG Streams
OSM-Trentino	156,5 MB	1	0	0
SU2	441,9MB	0	104,414	0
SU2OSM	19,2 MB	1	104,414	1955

Table 2 summarizes the dataset statistics. Notice how the SU2OSM dataset, while carrying more purpose relevant information than the union of C_R and all the C_P's has a size of around $19,2$MB, that is, around the 3% of their cumulative total size ($598,4$MB), this resulting from the fact that a very small minority of the entities in the OS-Trentino dataset are relevant to the current purpose. This provides evidence of the scalability and effectiveness of the proposed approach in the generation of big-thick data.

6.4 *Observation Enquiries*

We concentrate on three distinct binary classification prediction enquiries E1, E2 and E3, as from Table 1, which are illustrative examples of R-, P- and RP-enquiries, respectively (see Section 5). Each enquiry is identified by a *target property* and a set of *feature properties*, see Table 1. Target property and feature properties are the key elements characterizing the *Purpose Feasibility* of a dataset (see column 3 in Table 1). Purpose feasibility is a new boolean metric that indicates whether a dataset has the ability to support a certain purpose. We have tested various algorithms, i.e., *Logistic Regression*, *Decision Tree*, and *Random Forest* following the *5-fold cross-validation* approach [32]. In the evaluation we have used standard metrics, as from Table 1, where AUC stands for *Area Under the Curve*. Table 1 shows the results of prediction experiments for the E1, E2 and E3 with the best-performance algorithm.

There are two main observations. The *first* is about the purpose feasibility of SU2, OSM-Trentino and SU2OSM. OSM-Trentino can only answer E1 (a R question) because it exclusively populates a C_P EG, providing the target and properties for E1. Dually, SU2 can only answer the E2 (a P question), because it exclusively populates streams of C_P EGs. The merged SU2OSM dataset can answer all the proposed enquires because it unifies a C_P EG with the streams of C_P EGs. The *second* is about the prediction results. E1 can be answered both using OSM-Trentino and SU2OSM. However the SU2OSM metrics are better than those of OSM-Trentino, despite the first being much smaller. E2 can be answered both using SU2 and SU2OSM, but, again the SU2OSM metrics are better than those of the SU2 dataset. E3 can be answered only by SU2OSM.

7 Conclusion

The main focus of the line of work described in this paper is the development of AI's which supports humans in their life in the real world, as distinct from the virtual world enabled by the Web. Within this application context, we propose using *big-thick data*, namely, a new type of data which, in the opinion of the authors, are key and most likely necessary for the development of meaningful lifelong human-in-the-loop human-machine interactions.

The main result is an articulation of big-thick data as the result of the flexible integration, we call it *context unification*, of reference context and personal context data. The key element of this type of data is that it is not only machine generated, for instance in the form of IOT or Web data, but it is also *purposely* provided by humans. It is *only* humans who can provide high quality context-aware data. This human contribution can be in the form of the reference context, for instance as provided by motivated volunteers, as it is the case with OSM and open data, but it can also be in the form of personal data carrying detailed information of the *why*, the *what* and the *how* of people's behaviour. Our future work will focus on how to tightly integrate big-thick data generation and machine learning.

Acknowledgements

The first and most important thanks goes to Matteo Busso for inventing the term *big-thick data* and for producing the first version of the picture in Figure 1, reported here essentially unmodified. The work described in this paper could be done thanks to the amazing group of colleagues in the KnowDive Group: Mayukh Bagchi, Simone Bocca, Andrea Bontempelli, Ali Hamza, Leonardo Havier Malcotti, Ivan Kayongo, Alessio Zamboni and Haonan Zhao. This research has received funding from the European Union's Horizon 2020 FET Proactive project "WeNet – The Internet of us", grant agreement No 823783.

References

[1] G. D. Abowd, A. K. Dey, P. J. Brown, N. Davies, M. Smith, and P. Steggles. Towards a better understanding of context and context-awareness. In *Handheld and Ubiquitous Computing: First International Symposium, HUC'99 Karlsruhe, Germany, September 27–29, 1999 Proceedings 1*, pages 304–307. Springer, 1999.

[2] K. Abualsaud, T. M. Elfouly, T. Khattab, E. Yaacoub, L. S. Ismail, M. H. Ahmed, and M. Guizani. A survey on mobile crowd-sensing and its applications in the iot era. *Ieee access*, 7:3855–3881, 2018.

[3] Y. Y. Ang. Integrating big data and thick data to transform public services delivery. 2019.

[4] B. A. Aseniero, C. Perin, W. Willett, A. Tang, and S. Carpendale. Activity river: Visualizing planned and logged personal activities for reflection. In *Proceedings of the International Conference on Advanced Visual Interfaces*, pages 1–9, 2020.

[5] C. Bettini, G. Civitarese, and R. Presotto. Caviar: Context-driven active and incremental activity recognition. *Knowledge-Based Systems*, 196: 105816, 2020.

[6] I. Bison, H. Zhao, and F. Giunchiglia. What impacts the quality of the user answers when asked about the current context? *arXiv preprint arXiv:2405.04054*, 2024.

[7] G. e. Blank. Online research methods and social theory. *The SAGE Handbook of Online Research Methods*, page 537–549, 2008.

[8] A. Blok and M. A. Pedersen. Complementary social science? quali-quantitative experiments in a big data world. *Big Data & Society*, 1(2): 2053951714543908, 2014.

[9] S. Bocca, M. Dragoni, and F. Giunchiglia. iTelos - case studies in building domain specific knowledge graphs. In *ISIC-22 International Semantic Intelligence Conference*, 2022.

[10] T. Boellstorff. Making big data, in theory. *First Monday*, 18(10), 2013.

[11] A. Bontempelli, S. Teso, F. Giunchiglia, and A. Passerini. Learning in the wild with incremental skeptical gaussian processes. In *Proceedings of the 29th International Joint Conference on Artificial Intelligence (IJCAI)*, 2020. arXiv preprint arXiv:2011.00928.

[12] A. Bontempelli, M. R. Britez, X. Li, H. Zhao, L. Erculiani, S. Teso, A. Passerini, and F. Giunchiglia. Lifelong personal context recognition. In *HHAI-22 Workshop on Human-Centered Design of Symbiotic Hybrid Intelligence*, 2022. https://arxiv.org/abs/2205.10123.

[13] T. Bornakke and B. L. Due. Big-thick blending: A method for mixing analytical insights from big and thick data sources. *Big Data & Society*, 5(1):2053951718765026, 2018.

[14] D. Boyd and K. Crawford. Critical questions for big data: Provocations for a cultural, technological, and scholarly phenomenon. *Information, communication & society*, 15(5):662–679, 2012.

[15] G. Brewka and T. Eiter. Equilibria in heterogeneous nonmonotonic multi-context systems. In *AAAI*, volume 7, pages 385–390, 2007.

[16] G. Brewka, S. Ellmauthaler, R. Gonçalves, M. Knorr, J. Leite, and J. Pührer. Reactive multi-context systems: Heterogeneous reasoning in dynamic environments. *Artificial Intelligence*, 256:68–104, 2018.

[17] A. K. Dey. Understanding and using context. *Personal and ubiquitous computing*, 5:4–7, 2001.

[18] A. K. Dey, G. D. Abowd, and D. Salber. A conceptual framework and a toolkit for supporting the rapid prototyping of context-aware applications. *Human–Computer Interaction*, 16(2-4):97–166, 2001.

[19] C. Geertz. Thick description: Toward an interpretive theory of culture. In *The cultural geography reader*, pages 41–51. Routledge, 2008.

[20] C. Ghidini and F. Giunchiglia. Local models semantics, or contextual reasoning= locality+ compatibility. *Artificial intelligence*, 127(2):221–259, 2001.

[21] F. Giunchiglia and M. Fumagalli. Teleologies: Objects, actions and functions. In *ER- International Conference on Conceptual Modeling (ER2017)*, pages 520–534. ICCM, 2017.

[22] F. Giunchiglia and M. Fumagalli. Entity type recognition–dealing with the diversity of knowledge. In *Proc. iInt conference on principles of knowledge representation and reasoning (KRR 2020)*, volume 17, pages 414–423, 2020.

[23] F. Giunchiglia, E. Bignotti, and M. Zeni. Personal context modelling and annotation. In *2017 IEEE Int. Conf. on pervasive computing and communications workshops (PerCom workshops)*, pages 117–122. IEEE, 2017.

[24] F. Giunchiglia, M. Zeni, E. Gobbi, E. Bignotti, and I. Bison. Mobile social media usage and academic performance. *Computers in Human Behavior*, 82:177–185, 2018.

[25] F. Giunchiglia, A. Zamboni, M. Bagchi, and S. Bocca. Stratified data integration. *arXiv preprint arXiv:2105.09432*, 2021.

[26] F. Giunchiglia, S. Bocca, M. Fumagalli, M. Bagchi, and A. Zamboni. Popularity driven data integration. In *Iberoamerican Knowledge Graphs and Semantic Web Conference*, pages 277–284. Springer, 2022.

[27] F. Giunchiglia, X. Li, M. Busso, and M. Rodas-Britez. A context model for personal data streams. In *Asia-Pacific Web (APWeb) and Web-Age Information Management (WAIM) Joint International Conference on Web and Big Data*. Springer, 2022.

[28] F. Giunchiglia et al. Contextual reasoning. *Epistemologia, special issue on I Linguaggi e le Macchine*, 16:345–364, 1993.

[29] N. Guarino and G. Guizzardi. Relationships and events: towards a general theory of reification and truthmaking. In *25th International Conference AIxIA, Genova, Italy, November 29–December 1, 2016, Proceedings XV*, pages 237–249. Springer, 2016.

[30] A. Hogan, E. Blomqvist, M. Cochez, C. d'Amato, G. D. Melo, C. Gutierrez, S. Kirrane, J. E. L. Gayo, R. Navigli, S. Neumaier, et al. Knowledge graphs. *ACM Computing Surveys (Csur)*, 54(4):1–37, 2021.

[31] S. Ji, S. Pan, E. Cambria, P. Marttinen, and S. Y. Philip. A survey on knowledge graphs: Representation, acquisition, and applications. *IEEE Trans.neural networks and learning systems*.

[32] R. Kohavi et al. A study of cross-validation and bootstrap for accuracy estimation and model selection. In *Ijcai*, volume 14, pages 1137–1145. Montreal, Canada, 1995.

[33] J. McCarthy. Generality in artificial intelligence. *Communications of the ACM*, 30(12):1030–1035, 1987.

[34] J. McCarthy and P. J. Hayes. Some philosophical problems from the standpoint of artificial intelligence. In *Readings in artificial intelligence*, pages 431–450. Elsevier, 1981.

[35] E. Schubert, J. Sander, M. Ester, H. P. Kriegel, and X. Xu. Dbscan revisited, revisited: why and how you should (still) use dbscan. *ACM Transactions on Database Systems (TODS)*, 42(3):1–21, 2017.

[36] M. Sheikh, M. Qassem, and P. A. Kyriacou. Wearable, environmental, and smartphone-based passive sensing for mental health monitoring. *Frontiers in Digital Health*, 3:662811, 2021.

[37] Q. Shen, S. Teso, F. Giunchiglia, and H. Xu. To transfer or not to transfer and why? meta-transfer learning for explainable and controllable cross-individual activity recognition. *Electronics*, 12(10):2275, 2023.

[38] D. Shi, X. Li, and F. Giunchiglia. Kae: A property-based method for knowledge graph alignment and extension. *Journal of Web Semantics*, 82:100832, 2024.

[39] W. Sousa Lima, E. Souto, K. El-Khatib, R. Jalali, and J. Gama. Human activity recognition using inertial sensors in a smartphone: An overview. *Sensors*, 19(14):3213, 2019.

[40] Y. Vaizman, N. Weibel, and G. Lanckriet. Context recognition in-the-wild: Unified model for multi-modal sensors and multi-label classification. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, 1(4):1–22, 2018.

[41] D. J. Yeong, G. Velasco-Hernandez, J. Barry, and J. Walsh. Sensor and sensor fusion technology in autonomous vehicles: A review. *Sensors*, 21(6):2140, 2021.

[42] M. Zeni, I. Zaihrayeu, and F. Giunchiglia. Multi-device activity logging. In *Proceedings of the 2014 ACM international joint conference on pervasive and ubiquitous computing: adjunct publication*, pages 299–302, 2014.

[43] M. Zeni, W. Zhang, E. Bignotti, A. Passerini, and F. Giunchiglia. Fixing mislabeling by human annotators leveraging conflict resolution and prior knowledge. *Proc. of ACM IMWUT*, 3(1):32, 2019.

[44] W. Zhang, Q. Shen, S. Teso, B. Lepri, A. Passerini, I. Bison, and F. Giunchiglia. Putting human behavior predictability in context. *EPJ Data Science*, 10(1):42, 2021.

[45] H. Zhao, I. Kayongo, L. Malcotti, and F. Giunchiglia. Human-AI collaborative big-thick data collection. *arXiv preprint arXiv:2404.17602*, 2024.