An ML-Powered Human Behavior Management System

Sihem Amer-Yahia, Reynold Cheng, Mohamed Bouadi, Abdelouahab Chibah, Mohammadreza Esfandiari, Jiangping Zhou, Nan Zhang, Eric Lau, Yuguo Li, Xiaolin Han, et al.

An ML-Powered Human Behavior Management System

Sihem Amer-Yahia*, Reynold Cheng+, Mohamed Bouadi*, Abdelouahab Chibah*,
Mohammadreza Esfandiari*, Jiangping Zhou+, Nan Zhang+, Eric Lau+, Yuguo Li+,
Xiaolin Han+, Shivansh Mittal+
*Univ. Grenoble Alpes, CNRS, France,
{firstname.lastname}@univ-grenoble-alpes.fr
+University of Hong Kong, ckcheng@cs.hku.hk,
{zhoujp, liyg, zhangnan, ehylau, xiaolinh, shivansh}@hku.hk

Abstract

Our work aims to develop novel technologies for building an efficient data infrastructure as a backbone for a human behavior management system. Our infrastructure aims at facilitating behavior modeling, discovery, and exploitation, leading to two major outcomes: a behavior data management back-end and a high-level behavior specification API that supports mining, indexing and search, and AI-powered algorithms that provide the ability to extract insights on human behavior and to leverage data to advance human capital. We discuss the role of ML in populating and maintaining the back-end, and in exploiting it for human interest.

1 Introduction

We make a case for building a human behavior management system, where human behavior is a first class citizen and is mined, queried and managed over time. While several efforts have focused on studying human behavior in large scale population studies, in mining customer purchase patterns, or on the social Web, there is no single architecture that provides the ability to mine, query and manage behavior. Such a system would encourage reproducibility and enable several applications that benefit humans. In particular, the ability to model individual and collective behavior enables to offer new functionalities that let everyone can share and discover all kinds of assets, and combine assets to advance human capital. Assets can be composed into learning strategies that everyone can use to propose their skills, acquire new skills, or enhance existing ones. The proposed research will contribute to designing and developing approaches that leverage ML for building an effective and efficient human behavior management back-end that represents the variety of behaviors and caters to human needs. To enable this work, two novel and challenging research axes need to be developed in parallel: a database system to manage human behavior and a set of ML-powered algorithms that populate and maintain that database, as well as approaches to search and leverage behavior and assets with the goal of studying human behavior and advancing their capital.

To design a human behavior database, we need to capture human factors and populate the database by mining behavioral patterns over time. We propose to leverage approaches that estimate human factors [56] and mine

Bulletin of the IEEE Computer Society Technical Committee on Data Engineering

behaviors at the individual and collective levels [34, 46, 49]. The highly dynamic nature of human factors and behavior renders the maintenance of such a database quite challenging. We propose to leverage ML approaches to (i) learn behavior change rates, (ii) develop adaptive approaches to cater to humans, and (iii) choose back-end maintenance strategies accordingly. Existing work that leverages ML for data management [62] is nascent and covers the use of ML for query optimization [40] or for database indexing [38]. Additionally, the ability to search and leverage behavior and assets with the goal of studying human behavior and advancing their capital requires the development of novel ML-powered algorithms.

To study human behavior, we plan to develop a flexible tool for the on-demand discovery of behavioral patterns and their exploration over time. Such functionality benefits a variety of stakeholders. It would enable the design of robust population studies, marketing strategies, and promotional campaigns. Data scientists, social scientists, Web application designers, marketers, and other domain experts, need a single destination to explore behavioral changes. To query human behavior, we will leverage work on querying time series. Systems like Qetch [47] and ShapeSearch [60] provide the ability to express powerful shape queries through sketches, natural language, and regular expressions with flexible time span and amplitude. However, those queries are solely shape-based and cannot be used to search for change induced by humans of interest. We will hence develop approaches to combine shape queries with querying humans and change intensity.

For leveraging assets to advance human capital, we will make use of recent work on virtual marketplaces [29]. In physical workplaces, learning strategies include scaffolding where assets are combined in alternating difficulty levels, and collaboration where humans learn from their interactions with higher-skilled peers [24, 37, 42]. In virtual marketplaces, a few studies focused on humans' ability to improve their skills by completing tasks [32], and how affinity between humans can be used to form teams that collaborate to produce high quality contributions while also improving skills [29]. Our observation is that optimizing for human capital should be seen as a multi-objective problem that accounts for several goals and quality control, cost reduction, and human effort. We will develop a framework to observe humans and leverage ML techniques to learn their abilities and adaptively suggest appropriate assets to them for advancing their capital while accounting for other goals.

2 Motivating examples

Example 1 (Mining and querying behavior.): Our first example motivates the need for sophisticated mining approaches to discover and model evolving human behavior. We consider two typical examples: mining customer behavior in retail and mining the behavior of citizens riding public transportation. Mining behavior can help identify purchase patterns in the first case, and specific rider groups in the second, e.g., familiar stranger groups who use public transportation during the same period. In both examples, evolving behavior captures changing habits such as customers reacting to promotional offers, or riders changing groups when in transit. The ability to query behavior changes is useful for people in charge to analyze usage trends. In the riders case, this would help them identify trends of daily travel populations over time and see their behaviors change when special events occur (e.g., during the COVID-19 crisis). Extracted insights must be made available through a powerful querying interface to instruct public policies (e.g., new COVID-19 measures). Additionally, they can be used to train behavior change models that will instruct data maintenance. By learning different change models, we will enable ML-powered back-end updates.

Example 2 (Assets for advancing capital.): Assets such as online courses or facts to be verified, can be used for advancing human capital and the system must help humans improve their skill and knowledge either individually or collaboratively (a.k.a. peer learning). Assume we have 3 humans: Mary, John and Sarah. Mining their behavior in an online course system would help determine their skill level (1 for Mary who is a novice, 3 for John who has an intermediate level, and 5 for Sarah who is an expert). Given the courses they consumed so far, our goal is to assign to Mary a batch of $k = 5$ assets that maximize her learning. Mary needs the support of John or Sarah to consume intermediate level assets. She cannot consume hard assets even with help. By exposing Mary

to her peers' contributions, her learning potential is likely to increase [27]. Assigning over-challenging assets to her may result in frustration, and assigning under-challenging assets may lead to boredom.

In another example, we have a set of fact-checking tasks. Each task constitutes a collaborative asset, to be consumed by 12 individuals with varying skills. Each pair of individuals has an affinity that reflects how effectively they can collaborate based on their socio-demographics. Therefore, there are $\binom{12}{2}$ pairs of affinities forming a graph. One goal here is to divide the humans into 3 equi-sized groups of 4 members each so that peer learning is maximized.

3 Our system

3.1 Building a human behavior management database

The goal of this axis is to develop an integrated data model to represent human behavior, encompassing human factors and asset dimensions, and to leverage ML in maintaining the database. To make behavior and assets usable, we need to develop an API for populating, maintaining and accessing behavior and assets, including a declarative query language for expressing complex conditions on human factors and behavior and assets. Through these objectives it will be possible to offer querying human behavior and assets as a simple service promoting their reusability. Additionally, we will investigate performance aspects of this language and propose mechanisms for its efficient evaluation. The database needs to store raw human/asset data but also the results of extracting insights such as mining behavioral change of user groups. The evolving nature of human factors makes this particularly challenging. It is therefore necessary to apply ML approaches to learn behavioral change models and leverage them at maintenance time, when new raw data and new insights need to be stored in the back-end. This ML-powered approach will be compared against traditional batch and incremental maintenance approaches.

Objectives:

1. Design a model to capture, represent and manage human factors, human behavior and assets. Two kinds of factors must be considered: people-specific ones such as socio-demographic attributes, skill, reputation/trust, and motivation; and collaborative factors such as affinity and interaction models.

2. Design a model to capture and store extracted insights on individual and group behavior.

3. Investigate indexing mechanisms to retrieve and query behavioral change and assets efficiently.

4. Develop ML-powered algorithms for updating and managing evolving human behavior at individual and group levels.

3.2 Leveraging assets

This axis explores the study and querying of human behavior over time and the development of approaches for leveraging assets in different applications. In particular, we will focus on leveraging assets for advancing human capital. In our first endeavor, we will design queries that express behavior-aware change primitives. Our queries will be sent to the backend and need to be expressive enough to query behavior shapes and changes. This would require defining new scoring semantics that combine shape matching and intensity of change. Existing algorithms to match shape queries operate in time series and rely on splitting time using a fixed-size window and matching the query to each region in the window. We will leverage drift detection approaches on data streams [34] to handle time in a dynamic fashion.

In our second endeavor, we will study asset assignment, expecting that appropriate assignments will have a positive impact on the inherent learning capability of humans and on their overall performance. We focus on

a common class of assets, "Knowledge assets" in Bloom's taxonomy of educational objectives [17, 39] such as image classification, text editing, labeling, fact checking, and speech transcription asset. A common problem we will tackle is: given a human h and a set of unconsumed assets, which sequence of k assets will maximize h's learning potential? Here, learning potential is the maximum possible improvement in h's skill.

We adopt a model where contributions from other humans are made visible to the current human. Several studies showed that humans learn better when contributions from higher-skilled humans are shown to them [25, 26, 35, 36]. Our challenges are: (1) how to choose an appropriate batch of k assets where a human can see others' contributions, (2) how to order the chosen k assets appropriately so that the human's skill improvement is maximized. Our approach must enabke both individual and peer learning. We will leverage work in online critiquing communities,[1] social Q&A sites,[2] and crowdsourcing platforms[3] that investigate how collaboration can promote knowledge and skill improvement of individuals. In particular, we propose to explore how affinity between group members improves peer learning and address modeling, theoretical, and algorithmic challenges. We will build on our recent work for algorithmic group formation with affinities for peer learning [29].

Objectives for querying behavior:

1. Formalize an algebra that captures behavior evolution over time.

2. Develop a framework that given raw human/asset data, extracts groups and their behavior, stores them as insights in the database and represents change using our algebra.

3. Build a visual interface to query behavior with powerful conditions on behavior shape and change intensity.

Objectives for learning:

1. Formalize the *learning potential* of a human for an asset and choose k assets that maximize the total learning potential. This formalization should capture individual learning and peer learning. There are two theories underlying our framework. First, Zone of Proximal Development (ZPD) [65] is a well-known theory that defines three zones of assets with different skill improvements; (1) A learnable zone that contains assets a person can learn how to consume when assisted by a teacher or peer with a higher skillset, (2) a flow/comfort zone of assets that are easy and can be consumed with no help, and (3) a frustration zone of assets that a learner cannot consume even with help. Second, the Flow theory [21] states that people are able to immerse themselves in doing things whose challenge matches their skills. In [15], the authors claim that to improve skills, the assets should be either in the flow/comfort zone, or in the learnable zone on the condition that there is some "scaffolding" to help humans consume assets that are a bit more challenging for them. This results in skill improvement (the dotted line). Our formalization builds on that and defines the learning potential for both individual assets (mainly in the flow/comfort zone) and collaborative assets (mainly in the learnable zone).

2. Devise learning strategies which interleave individual assets and collaborative assets. We will study their impact on humans' performance and skills. Previous work found that the order of assets impacts quality and completion time [22, 18]. For instance, assets could be provided in no particular order, or grouped and presented in alternating difficulty levels.

3. Propose adaptive and iterative asset search methods that take a human h, and assigns to h, at each iteration, a batch of k assets according to a learning objective. This approach may give rise to multi-objective problems.

[1] https://movielens.org/
[2] http://quora.com/
[3] https://www.figure-eight.com/

4 Modeling human behavior

Our model must capture humans, assets, and human behavior and its evolution over time. Figure 1 shows a two-level E/R diagram that represents human and asset data and extracted insights. It will serve as a basis for the design of our backend.

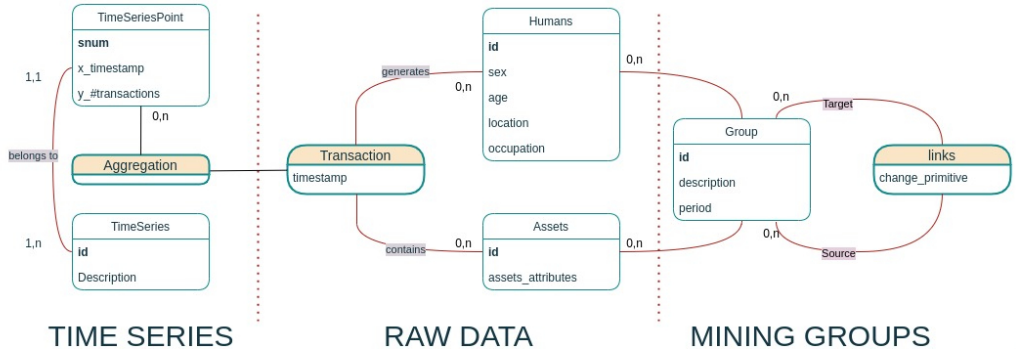


Figure 1: E/R diagram representing raw human/asset data and extracted insights.

The above framework could be used to discover and model evolving human behavior of citizens riding public transportation, and see how their behaviors change when special events occur. Here, we aim to study passenger behaviors during the COVID-19 crisis. COVID-19 is characterized by a broad spectrum of disease presentation, from asymptomatic or subclinical infections to severe diseases and deaths. A person infected with SARS-CoV-2 virus who has not yet developed symptoms or confirmed by laboratory testing would likely maintain normal social activities. Recent studies show that infected cases are contagious for asymptomatic and pre-symptomatic cases, and continuous to be contagious for at least a week [1, 2], providing ample opportunity for transmitting SARS-CoV-2 through public transportation. Recently, we have performed a study in Hong Kong with the Mass Transit Railway (MTR) Corporation, which is the only railway and subway service provider in Hong Kong. The MTR railway network covers areas inhabited by more than 70% of the local population [4]. About 50% of total number of rider trips in Hong Kong (or 4.5 million riders) are made through MTR [5]. Thus, the local population mobility in Hong Kong is well represented by the MTR traveling population. Courtesy of MTR, we have obtained *all* the entry and exit data of of anonymous riders (e.g., time and station of entry, ticket type (kid/adult/elder)) in the first four months (i.e., 1 January to 30 April) of 2020, which is also a coronavirus outbreak period in Hong Kong.

We have performed some initial analysis of the MTR data. Figure 2 shows the change of daily population in the first three months of 2020 (here, *octopus* refers to the most popular smart card in Hong Kong; *ticket* means the single entry ticket). We can see that the daily MTR traveling population is reduced by more than 40% after the end of Chinese New Year holiday on Jan 28, which can be related to the lockdown of Hubei, China on Jan 23, just before the holiday. Also, the MTR traveling population in weekends is more than 10% less than that in workdays. Another observation is that while the number of MTR passengers is generally decreasing, the number of new confirmed cases increases significantly.

Rider type discovery. The MTR riders could physically encounter one another, and temporally share different facilities or spaces such as stations, platforms, elevators, and carriages. They can concurrently share a crowded and small carriage for an extensive amount of time. We would also like to study different group behaviors of MTR riders in the period of January to March 2020, during which the outbreak of coronavirus occurs. As discussed in [67], these riders include:

- "Someone-like-you": they are groups of riders who share the same trip and stay close to each other, and simultaneously share trajectories for at least one trip.

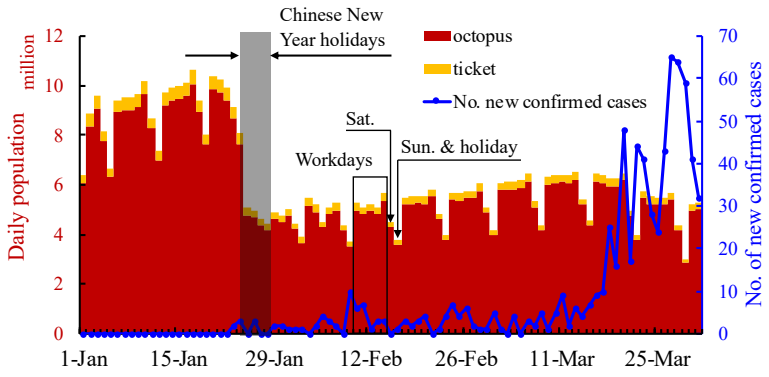


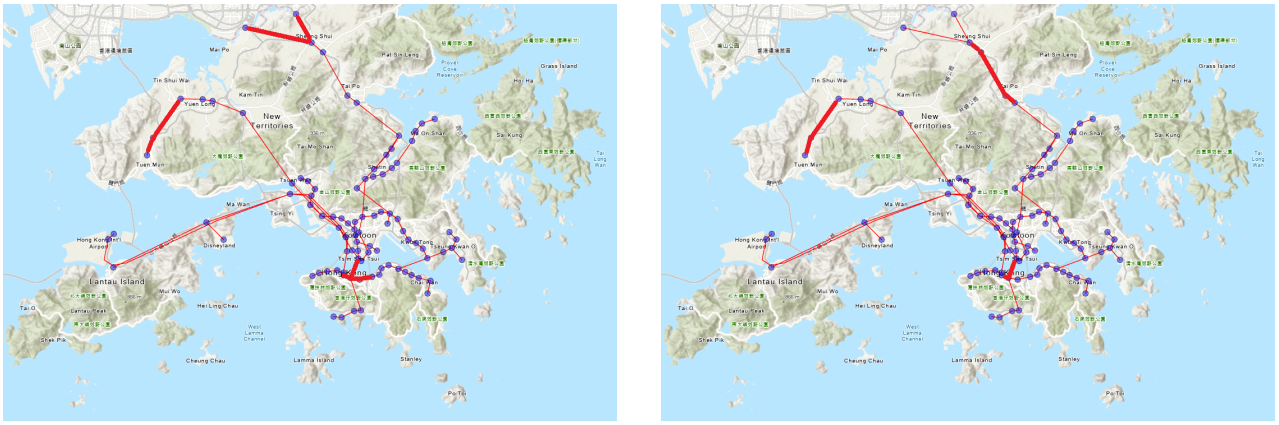
Figure 2: MTR daily population and number of infected cases in Jan-Mar 2020.

- Sensor riders, who have the most physical contacts with other riders at stations and carriages. If they are "super spreaders", they can infect many people.
- Extreme riders, who have to endure the longest journeys or who have to ride the MTR train most frequently. They can have a high risk of contracting coronavirus or respiratory diseases.
- Choice riders, who quickly change their travel patterns given the COVID-19 situation, for example, due to outbreak in certain districts, or new government policies for work and classes.

We aim to discover the above riders through mining the MTR data. The riders found can be useful to understand whether they have a high risk of spreading or contracting coronavirus. They can be useful for simulating scenarios and testing rider control and intervention policy. Moreover, they can be used for issuing warning or further actions. For example, in a "someone-like-you" group, if any member of this group contracts contagious viruses, other members who are close contacts of the patient should be notified and monitored as soon as possible. We will develop data mining solutions to identify rider types from the MTR data. For example, to obtain "someone-like-you" riders, one way is to group the rider information according to the stations and the times they entered and exited. Each group consists of riders who are estimated to take the same train. We will study methods for finding all these groups, and perform extensive analysis, including the estimation of average group size over weekdays and weekends, and examining the station-pairs that exhibits large group sizes. Figure 3 shows the spatial-temporal patterns of the average "someone-like-you" group sizes along different MTR trips during workdays in Hong Kong before Jan 24 and after Jan 28 in 2020. The thickness of the line indicates the size of the rider group. We can see that that the trips with the largest "someone-like-you" groups in the two time periods are not the same. We will compare the rider types across the first four months of 2019 and 2020.

Due to the gigantic number of riders, the discovery of rider types can be time-consuming. We will develop fast database algorithms and indexes to discover riders efficiently. We will extend the TopPI algorithm [44], developed by co-I Amer-Yahia, in order to mine rider groups from millions of records in sub-seconds. We will also extend the work on exploring datasets with rating maps to enable group comparison [11].

Our initial studies show that the data in hand is very promising (e.g., find out the risks of contracting coronavirus for different groups of passengers, and riders' travel behavior arising from the anxiety of being infected in public transport). However, without a scalable platform for human behavior analysis, it is difficult to perform advanced analysis and simulation. In particular, the MTR data, more than 60GB, contains more than 1 billion entry/exit transactions of 8 million riders. We would like to perform "microscopic" analysis (e.g., do riders have a close contact?), and also see how the disease is spread in a crowded train with insufficient ventilation. As the MTR data spans 4 months, it is also interesting to examine how the behaviors of passengers change with time. Existing practices and solutions in the public health and transportation fields are not scalable



Workdays on and before Jan 24 **Workdays after Jan 28**

Figure 3: Spatio-temporal patterns of "someone-like-you" riders in MTR.

to handle such a large amount of data (e.g., it takes more than 5 hours to perform a task to analyze rider behavior for a single day's data). This is simply too slow in face of the crisis that we are facing. The proposed system would be able to support analysis on the group behavior of citizens taking public transportation. Our results would help to monitor population transmission potential, and guide the appropriate level of social distancing measures [3].

5 Leveraging assets for learning

We consider a set of humans $\mathcal{H}$ and a set of assets $\mathcal{A}$. Humans in $\mathcal{H}$ consume assets in $\mathcal{A}$ at different times, either together or separately. The term asset is general enough to represent recommendations on the social web, courses in an online teaching system, tasks in crowdsourcing, etc. We denote $\mathbf{A}_h^t$ a batch of (possibly ordered) assets consumed by a human h at time t. The learning potential of a human h who consumes a batch of assets $\mathbf{A}$ at time $t+1$ depends on several factors: 1) the effort of h to consume assets in $\mathbf{A}$, 2) h's performance factor, 3) other humans' performance when learning collaboratively, and 4) the affinity between h and other humans. This gives rise to two problems: individual learning and collaborative learning.

We can define learning as the problem of individual asset assignment as follows: Given a human h and the batches of assets consumed by h up to iteration i: $\mathbf{A}_h^1 \ \ldots \ \mathbf{A}_h^t$, find a batch $\mathbf{A}$ of at most k assets to assign to human h at time $t+1$ such that:

$$\underset{\mathbf{A}}{\operatorname{argmax}}\, \mathit{learning}(h, \mathbf{A})$$

$\mathit{learning}(h, \mathbf{A})$ is a function that captures the learning potential of h who consumes the set of assets $\mathbf{A}$. Our problem is to determine the right batch of assets to provide to a human at time t. Our problem can be seen as a variant of the Knapsack Problem [19]. Our items are assets and each asset has a value (in our case v is $\mathit{learning}(w, t)$) and a weight, we want to find k assets that maximize the sum of values $\sum v_i$ under a capacity constraint k. What makes our problem simple is that the weight is equal to 1 which yields a top-k solution. Additionally, as the value of assigning an asset to a human depends on the human and evolves over time as other humans consume assets, we need to account for that dynamicity in the asset assignment process.

We can also define a variant of our problem where assets are consumed collaboratively: Given a set $H = \{h_1, ..., h_n\}$ of humans with their corresponding skill values h_i^s, our goal is to form a grouping $\mathcal{G}$ that

contains k equi-sized groups $g_1, g_2, ..., g_k$ and that maximizes two objective functions, aggregated learning potential (*LP*) and aggregated affinity (*Aff*) between humans in the same group. More formally:

$$\underset{\mathcal{G}}{\text{maximize}} \quad \sum_{i=1}^{k} LP(g_i), \sum_{i=1}^{k} Aff(g_i) \;\text{ s.t. }\; |\mathcal{G}| = k,\; |g_i| = \frac{n}{k} \tag{1}$$

where $LP(g_i)$ (resp. $Aff(g_i)$) refers to any of the learning potential (resp. affinity).

Since the two objectives are incompatible with one another, our problem qualifies as *multi-objective*. In [29], we present approximation algorithms that find a feasible grouping (that maximizes learning potential) and offer provable constant approximation for affinities. We plan to build on that work.

Whenever a human consumes a collaborative asset, the asset's metadata is updated. Additionally, assets are grouped by difficulty level and by the number of remaining humans. A human's performance factor and skill also need to be updated as humans consume assets. We assume that a human's skill improves monotonically: the skill level remains the same or increases as time passes [48, 64] and is updated as they consume more assets. An ML approach that observes humans and revisits their attributes is warranted.

6 Related work

6.1 Mining customer behavior

Several approaches were proposed to track the evolution of customer groups over time [34, 46], and detect behavioral changes using pattern mining. Starting from a transactional database with customer demographics, the RFM score [49] is used to create customer groups according to their purchase frequency and spending. Association rules are extracted for consecutive time periods and compared with a custom similarity measure that leads to four changes: emerging, added, perished and unexpected patterns.

In [45], customer groups are built to reflect long-term patterns such as seasonal effects, and short-term patterns such as attractiveness of promotions. Customer purchases are captured with a non-homogeneous Poisson Process. The aim is to identify for a given product and period, k latent overlapping customer groups. Model parameters are learned with an Expectation–Maximization algorithm. Experiments on an Australian supermarket show that long-term and short term patterns provide insights such as "If the demand for a product category is seasonal, the category will have more U-shape and inverse U-shape patterns".

We plan to enable the application of different approaches for mining behavioral change. More importantly, the result of behavior mining will be stored in our back-end and readily available for querying.

6.2 Peer learning

Social science has a long history of studying non-computational aspects of computer-supported collaborative learning [20, 23]. With the development of online educational platforms (such as, Massive Open Online Courses or MOOCs), several parameters were identified for building effective teams: (1) individual and group learning and social goals, (2) interaction processes and feedbacks [61], (3) roles that determine the nature and group idiosyncrasy [23]. To the best of our knowledge, the closest to our work are [6, 7, 9], where quantitative models are proposed to promote group-based learning, albeit without affinity.

Our work is grounded in social science and takes a computational approach to the design of scalable solutions for peer learning with guarantees.

Many papers in crowdsourcing report that making other humans' contributions visible improves skills during task completion [27, 41, 35, 36, 26]. We will use this kind of indirect communication among humans in our collaborative assets. Group formation in online communities has been studied primarily in the context of task

assignment [12, 13, 43, 16, 54]. The problem is often stated as: given a set of individuals and tasks, form a set of groups that optimize some aggregated utility subject to constraints such as group size, maximum workload etc. Utility can be aggregated in different ways: the sum of individual skills, their product, etc [13]. Group formation is combinatorial in nature and proposed algorithms solve the problem under different constraints and utility definitions (e.g., [43]).

Our work studies computational aspects and formulates optimization problems to find the best assets for a human. In particular, we leverage expressive multi-objective formulations to optimize more than one goal and form groups with the goal of maximizing peer learning under different affinities.

6.3 Querying change

Visual querying tools [50, 52, 57, 59, 66] help search for time series containing a desired shape by taking as input a sketch. Most of these tools perform precise point-wise matching using measures such as Euclidean distance or DTW. A few others enable flexible search and define a scoring function to capture how well a time series matches a sketch. Tools like TimeSearcher [14] let users apply soft or hard constraints on the x and y range values via boxes or query envelopes, but do not support other shape primitives beyond location constraints. Qetch [47] supports visual sketches and a custom similarity metric that is robust to distortions in the query, in addition to supporting a "repeat" operator for finding recurring patterns. ShapeSearch [60] enables expressive shape queries in the form of sketches, natural-language, and visual regular expressions. Queries are translated into a shape algebra and evaluated efficiently. Symbolic sequence matching papers approach the problem of pattern matching by employing offline computation to chunk trendlines into fixed length blocks, encoding each block with a symbol that describes the pattern in that block [10, 31, 33, 8, 58]. Among those, Shape Definition Language (SDL) [10] encodes search blocks using "up", "down", and "flat" patterns, much like Qetch and ShapeSearch, and supports a language for searching for patterns based on their sequence or the number of occurrences. A few other visual time series exploration tools such as Metro-Viz [28] and ONEX [53] support additional analytics tasks such as anomaly detection and clustering.

Our work is complementary to ShapeSearch and Qetch. We will extend Qetch with the ability to detect and score finer changes for a set of humans.

7 Conclusion

Emerging Big Data applications, such as learning the behaviors of citizens taking public transportation and providing them with assets for advancing human capital, necessitates the development of a scalable ML-driven human behavior management system. Such a system not only enables the development of human-behavior learning applications, but also provides insights on how to properly enable updates, which is important to these applications in which new data are generated at high speeds. An immediate direction is to study ML-driven data maintenance, which governs data update strategies based on machine learning approaches.

References

[1] He X, Lau EHY, Wu P, Deng X, Wang J, Hao X, Lau YC, Wong JY, Guan Y, Tan X, Mo X, Chen Y, Liao B, Chen W, Hu F, Zhang Q, Zhong M, Wu Y, Zhao L, Zhang F, Cowling BJ, Li F, Leung GM. Temporal dynamics in viral shedding and transmissibility of COVID-19. *Nat Med*. 2020 May;26(5):672-675.

[2] Emery JC, Russell TW, Liu Y, Hellewell J, Pearson CA; CMMID COVID-19 Working Group, Knight GM, Eggo RM, Kucharski AJ, Funk S, Flasche S, Houben RMGJ. The contribution of asymptomatic SARS-CoV-2 infections to transmission on the Diamond Princess cruise ship. *Elife*. 2020 Aug 24;9:e58699.

[3] Buckee CO, Balsari S, Chan J, et al. Aggregated mobility data could help fight COVID-19. *Science*. 2020;368(6487):145-146.

[4] Transport and Housing Bureau. Railway development strategy. *HKSAR Government*, 2014.

[5] Transport Department. Public transport strategy study. *HKSAR Government*, 2017.

[6] Rakesh Agrawal, Behzad Golshan, and Evangelos Papalexakis. Toward data-driven design of educational courses: A feasibility study. *EDM*, 2016.

[7] Rakesh Agrawal, Behzad Golshan, and Evimaria Terzi. Grouping students in educational settings. In *SIGKDD*, 2014.

[8] Rakesh Agrawal, King-Ip Lin, Harpreet S. Sawhney, and Kyuseok Shim. Fast similarity search in the presence of noise, scaling, and translation in time-series databases. In Umeshwar Dayal, Peter M. D. Gray, and Shojiro Nishio, editors, *VLDB'95, Proceedings of 21th International Conference on Very Large Data Bases, September 11-15, 1995, Zurich, Switzerland*, pages 490–501. Morgan Kaufmann, 1995.

[9] Rakesh Agrawal, Sharad Nandanwar, and Narasimha Murty Musti. Grouping students for maximizing learning from peers. In *EDM*, 2017.

[10] Rakesh Agrawal, Giuseppe Psaila, Edward L. Wimmers, and Mohamed Zaït. Querying shapes of histories. In Umeshwar Dayal, Peter M. D. Gray, and Shojiro Nishio, editors, *VLDB'95, Proceedings of 21th International Conference on Very Large Data Bases, September 11-15, 1995, Zurich, Switzerland*, pages 502–514. Morgan Kaufmann, 1995.

[11] Sihem Amer-Yahia, Sofia Kleisarchaki, Naresh Kumar Kolloju, Laks V. S. Lakshmanan, Ruben H. Zamar. Exploring Rated Datasets with Rating Maps. In Proceedings of the 26th International Conference on World Wide Web, WWW 2017, Perth, Australia, 2017, 1411–1419, 2017.

[12] Aris Anagnostopoulos et al. Power in unity: forming teams in large-scale community systems. In *CIKM*, 2010.

[13] Aris Anagnostopoulos et al. Online team formation in social networks. In *WWW*, 2012.

[14] A Aris, A Khella, P Buono, B Shneiderman, and C Plaisant. Timesearcher 2. *Human-Computer Interaction Laboratory, Computer Science Department, University of Maryland*, 2005.

[15] Ashok R Basawapatna, Alexander Repenning, Kyu Han Koh, and Hilarie Nickerson. The zones of proximal flow: guiding students through a space of computational thinking skills and challenges. In *Proceedings of the ninth annual international ACM conference on International computing education research*, pages 67–74, 2013.

[16] Senjuti Basu Roy et al. Task assignment optimization in knowledge-intensive crowdsourcing. *VLDA*, 2015.

[17] Benjamin S Bloom. Taxonomy of educational objectives. vol. 1: Cognitive domain. *New York: McKay*, pages 20–24, 1956.

[18] Carrie J Cai, Shamsi T Iqbal, and Jaime Teevan. Chain reactions: The impact of order on microtask chains. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*, pages 3143–3154. ACM, 2016.

[19] Chandra Chekuri and Sanjeev Khanna. A polynomial time approximation scheme for the multiple knapsack problem. *SIAM J. COMPUT*, 38(3):1, 2006.

[20] Elizabeth G Cohen. Restructuring the classroom: Conditions for productive small groups. *Review of educational research*, 1994.

[21] Mihaly Csikszentmihalyi. *Beyond boredom and anxiety: The experience of play in work and games.* Jossey-Bass, 1975.

[22] Peng Dai, Jeffrey M Rzeszotarski, Praveen Paritosh, and Ed H Chi. And now for something completely different: Improving crowdsourcing workflows with micro-diversions. In *Proceedings of the 18th ACM Conference on Computer Supported Cooperative Work & Social Computing*, pages 628–638. ACM, 2015.

[23] Thanasis Daradoumis et al. Supporting the composition of effective virtual groups for collaborative learning. In *ICCE*. IEEE, 2002.

[24] Leo J De Vin, Lasse Jacobsson, JanErik Odhe, and Anders Wickberg. Lean production training for the manufacturing industry: Experiences from karlstad lean factory. *Procedia Manufacturing*, 11:1019–1026, 2017.

[25] Mira Dontcheva, Robert R Morris, Joel R Brandt, and Elizabeth M Gerber. Combining crowdsourcing and learning to improve engagement and performance. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pages 3379–3388, 2014.

[26] Shayan Doroudi, Ece Kamar, and Emma Brunskill. Not everyone writes good examples but good examples can come from anywhere. In *Proceedings of the AAAI Conference on Human Computation and Crowdsourcing*, volume 7, pages 12–21, 2019.

[27] Steven Dow, Anand Kulkarni, Scott Klemmer, and Björn Hartmann. Shepherding the crowd yields better work. In *Proceedings of the ACM 2012 conference on computer supported cooperative work*, pages 1013–1022, 2012.

[28] Philipp Eichmann, Franco Solleza, Nesime Tatbul, and Stan Zdonik. Visual exploration of time series anomalies with metro-viz. In Peter A. Boncz, Stefan Manegold, Anastasia Ailamaki, Amol Deshpande, and Tim Kraska, editors, *Proceedings of the 2019 International Conference on Management of Data, SIGMOD Conference 2019, Amsterdam, The Netherlands, June 30 - July 5, 2019*, pages 1901–1904. ACM, 2019.

[29] Mohammadreza Esfandiari, Dong Wei, Sihem Amer-Yahia, and Senjuti Basu Roy. Optimizing peer learning in online groups with affinities. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, KDD 2019, Anchorage, AK, USA, August 4-8, 2019*, pages 1216–1226, 2019.

[30] Sarah Evans et al. More than peer production: Fanfiction communities as sites of distributed mentoring. In *CSCW*, 2017.

[31] Christos Faloutsos, M. Ranganathan, and Yannis Manolopoulos. Fast subsequence matching in time-series databases. In Richard T. Snodgrass and Marianne Winslett, editors, *Proceedings of the 1994 ACM SIGMOD International Conference on Management of Data, Minneapolis, Minnesota, USA, May 24-27, 1994*, pages 419–429. ACM Press, 1994.

[32] Ujwal Gadiraju and Stefan Dietze. Improving learning through achievement priming in crowdsourced information finding microtasks. In *Proceedings of the Seventh International Learning Analytics & Knowledge Conference*, pages 105–114. ACM, 2017.

[33] Minos N. Garofalakis, Rajeev Rastogi, and Kyuseok Shim. SPIRIT: sequential pattern mining with regular expression constraints. In Malcolm P. Atkinson, Maria E. Orlowska, Patrick Valduriez, Stanley B. Zdonik, and Michael L. Brodie, editors, *VLDB'99, Proceedings of 25th International Conference on Very Large Data Bases, September 7-10, 1999, Edinburgh, Scotland, UK*, pages 223–234. Morgan Kaufmann, 1999.

[34] Daniel Kifer, Shai Ben-David, and Johannes Gehrke. Detecting change in data streams. In *(e)Proceedings of the Thirtieth International Conference on Very Large Data Bases, VLDA 2004, Toronto, Canada, August 31 - September 3 2004*, pages 180–191, 2004.

[35] Juho Kim. *Learnersourcing: improving learning with collective learner activity*. PhD thesis, Massachusetts Institute of Technology, 2015.

[36] Masaki Kobayashi, Hiromi Morita, Masaki Matsubara, Nobuyuki Shimizu, and Atsuyuki Morishima. An empirical study on short-and long-term effects of self-correction in crowdsourced microtasks. In *Sixth AAAI Conference on Human Computation and Crowdsourcing*, 2018.

[37] Martijn Koops and Martijn Hoevenaar. Conceptual change during a serious game: Using a lemniscate model to compare strategies in a physics game. *Simulation & Gaming*, 44(4):544–561, 2013.

[38] Tim Kraska, Alex Beutel, Ed H. Chi, Jeffrey Dean, and Neoklis Polyzotis. The case for learned index structures. In *Proceedings of the 2018 International Conference on Management of Data, SIGMOD Conference 2018, Houston, TX, USA, June 10-15, 2018*, pages 489–504, 2018.

[39] David R Krathwohl and Lorin W Anderson. *A taxonomy for learning, teaching, and assessing: A revision of Bloom's taxonomy of educational objectives*. Longman, 2009.

[40] Ani Kristo, Kapil Vaidya, Ugur Çetintemel, Sanchit Misra, and Tim Kraska. The case for a learned sorting algorithm. In *Proceedings of the 2020 International Conference on Management of Data, SIGMOD Conference 2020, online conference [Portland, OR, USA], June 14-19, 2020*, pages 1001–1016, 2020.

[41] Suna Kyun, Slava Kalyuga, and John Sweller. The effect of worked examples when learning to write essays in english literature. *The Journal of Experimental Education*, 81(3):385–408, 2013.

[42] Susanne P Lajoie and Alan Lesgold. Apprenticeship training in the workplace: Computer-coached practice environment as a new form of apprenticeship. *Machine-mediated learning*, 3(1):7–28, 1989.

[43] Theodoros Lappas, Kun Liu, and Evimaria Terzi. Finding a team of experts in social networks. In *SIGKDD*, 2009.

[44] Vincent Leroy, Martin Kirchgessner, Alexandre Termier, Sihem Amer-Yahia. TopPI: An efficient algorithm for item-centric mining. Inf. Syst. Volume 64, 104–118, 2017.

[45] Ling Luo, Bin Li, Irena Koprinska, Shlomo Berkovsky, and Fang Chen. Discovering temporal purchase patterns with different responses to promotions. In *Proceedings of the 25th ACM International on Conference on Information and Knowledge Management*, page 2197–2202, New York, NY, USA, 2016. ACM.

[46] Ling Luo, Bin Li, Irena Koprinska, Shlomo Berkovsky, and Fang Chen. Tracking the evolution of customer purchase behavior segmentation via a fragmentation-coagulation process. In *Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence, IJCAI-17*, pages 2414–2420, 2017.

[47] Miro Mannino and Azza Abouzied. Qetch: Time series querying with expressive sketches. In Gautam Das, Christopher M. Jermaine, and Philip A. Bernstein, editors, *Proceedings of the 2018 International Conference on Management of Data, SIGMOD Conference 2018, Houston, TX, USA, June 10-15, 2018*, pages 1741–1744. ACM, 2018.

[48] Julian John McAuley and Jure Leskovec. From amateurs to connoisseurs: modeling the evolution of user expertise through online reviews. In *Proceedings of the 22nd international conference on World Wide WeA*, pages 897–908, 2013.

[49] R Miglausch John. Thoughts on rfm scoring. *The Journal of Database Marketing*, 8(1):7, 2000.

[50] Matt Mohebbi, Dan Vanderkam, Julia Kodysh, Rob Schonberger, Hyunyoung Choi, and Sanjiv Kumar. Google correlate whitepaper. 2011.

[51] Vicente Rodríguez Montequín et al. Using myers-briggs type indicator (MBTI) for assessment success of student groups in project based learning. In *CSEDU*, 2010.

[52] P. K. Muthumanickam, K. Vrotsou, M. Cooper, and J. Johansson. Shape grammar extraction for efficient query-by-sketch pattern matching in long time series. In *2016 IEEE Conference on Visual Analytics Science and Technology (VAST)*, pages 121–130, 2016.

[53] Rodica Neamtu, Ramoza Ahsan, Charles Lovering, Cuong Nguyen, Elke A. Rundensteiner, and Gábor N. Sárközy. Interactive time series analytics powered by ONEX. In Semih Salihoglu, Wenchao Zhou, Rada Chirkova, Jun Yang, and Dan Suciu, editors, *Proceedings of the 2017 ACM International Conference on Management of Data, SIGMOD Conference 2017, Chicago, IL, USA, May 14-19, 2017*, pages 1595–1598. ACM, 2017.

[54] Habibur Rahman et al. Task assignment optimization in collaborative crowdsourcing. In *ICDM*, 2015.

[55] Habibur Rahman et al. Worker skill estimation in team-based tasks. *PVLDA*, 2015.

[56] Habibur Rahman, Saravanan Thirumuruganathan, Senjuti Basu Roy, Sihem Amer-Yahia, and Gautam Das. Worker skill estimation in team-based tasks. *Proceedings of the VLDB Endowment*, 8(11):1142–1153, 2015.

[57] Kathy Ryall, Neal Lesh, Tom Lanning, Darren Leigh, Hiroaki Miyashita, and Shigeru Makino. Querylines: Approximate query for visual browsing. In *CHI '05 Extended Abstracts on Human Factors in Computing Systems*, CHI EA '05, page 1765–1768, New York, NY, USA, 2005. Association for Computing Machinery.

[58] Hagit Shatkay and Stanley B. Zdonik. Approximate queries and representations for large data sequences. *CoRR*, abs/1904.09262, 2019.

[59] Tarique Siddiqui, Albert Kim, John Lee, Karrie Karahalios, and Aditya G. Parameswaran. zenvisage: Effortless visual data exploration. *CoRR*, abs/1604.03583, 2016.

[60] Tarique Siddiqui, Paul Luh, Zesheng Wang, Karrie Karahalios, and Aditya G. Parameswaran. Shapesearch: A flexible and efficient system for shape-based exploration of trendlines. In David Maier, Rachel Pottinger, AnHai Doan, Wang-Chiew Tan, Abdussalam Alawini, and Hung Q. Ngo, editors, *Proceedings of the 2020 International Conference on*

Management of Data, SIGMOD Conference 2020, online conference [Portland, OR, USA], June 14-19, 2020, pages 51–65. ACM, 2020.

[61] Ivan Srba and Maria Bielikova. Dynamic group formation as an approach to collaborative learning support. *TLT*, 2015.

[62] Ion Stoica. Systems and ML: when the sum is greater than its parts. In *Proceedings of the 2020 International Conference on Management of Data, SIGMOD Conference 2020, online conference [Portland, OR, USA], June 14-19, 2020*, page 1, 2020.

[63] Mohammad Taheri, Nasser Sherkat, Nick Shopland, Dorothea Tsatsou, Enrique Hortal Nicholas Vretos, Christos Athanasiadis, and Penny Standen. Adaptation and personalization principles based on mathisis findings. In *Public report on Managing Affective-learning THrough Intelligent atoms and Smart InteractionS project*, 2017.

[64] Kazutoshi Umemoto, Tova Milo, and Masaru Kitsuregawa. Toward recommendation for upskilling: Modeling skill improvement and item difficulty in action sequences. In *2020 IEEE 36th International Conference on Data Engineering (ICDE)*, pages 169–180. IEEE, 2020.

[65] Lev Vygotsky. Zone of proximal development. *Mind in society: The development of higher psychological processes*, 5291:157, 1987.

[66] Martin Wattenberg. Sketching a graph to query a time-series database. In Marilyn M. Tremaine, editor, *CHI '01 Extended Abstracts on Human Factors in Computing Systems, CHI Extended Abstracts '01, Seattle, Washington, USA, March 31 - April 5, 2001*, pages 381–382. ACM, 2001.

[67] J. Zhou, Y. Yang, H. Ma, and Y. Li. Familiar strangers in the big data era: An exploratory study of Beijing metro encounters. *Cities*, 97:102495, 2020.