Beyond Artificial Reality: Finding and Monitoring Live Events from Social Sensors

CALTON PU, ABHIJIT SUPREM, and RODRIGO ALVES LIMA, School of Computer Science,
Georgia Institute of Technology, Atlanta

AIBEK MUSAEV, Department of Computer Science, University of Alabama, Tuscaloosa

DE WANG, Sunmi US Inc

DANESH IRANI, Google

STEVE WEBB, Web Gnomes

JOAO EDUARDO FERREIRA, Department of Computer Science, University of Sao Paulo, Sao Paulo, Brazil

With billions of active social media accounts and millions of live video cameras, live new big data offer many opportunities for smart applications. However, the main consumers of the new big data have been humans. We envision the research on *live knowledge*, to automatically acquire real-time, validated, and actionable information. Live knowledge presents two significant and diverging technical challenges: big noise and concept drift. We describe the EBKA (evidence-based knowledge acquisition) approach, illustrated by the LITMUS landslide information system. LITMUS achieves both high accuracy and wide coverage, demonstrating the feasibility and promise of EBKA approach to achieve live knowledge.

CCS Concepts: • Computing methodologies → Machine learning; • Information systems → *Information systems applications*;

Additional Key Words and Phrases: Artificial reality, live knowledge, true novelty, concept drift, evidencebased knowledge acquisition, real-time event detection

ACM Reference format:
Calton Pu, Abhijit Suprem, Rodrigo Alves Lima, Aibek Musaev, De Wang, Danesh Irani, Steve Webb, and Joao Eduardo Ferreira. 2020. Beyond Artificial Reality: Finding and Monitoring Live Events from Social Sensors. *ACM Trans. Internet Technol.* 20, 1, Article 2 (March 2020), 21 pages. https://doi.org/10.1145/3374214

This research has been partially funded by the National Science Foundation by CISE's SAVI/RCN (1402266, 1550379), CNS (1421561), CRISP (1541074), SaTC (1564097) programs, an REU supplement (1545173), and gifts, grants, or contracts from Fujitsu, HP, Intel, and Georgia Tech Foundation through the John P. Imlay, Jr. Chair endowment. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation or other funding agencies and companies mentioned above.
Authors' addresses: C. Pu, A. Suprem, and R. A. Lima, School of Computer Science, Georgia Institute of Technology, Atlanta, GA, 30332-0765, USA; emails: calton.pu@cc.gatech.edu, asuprem@outlook.com, rodrigoalveslima@gmail.com; A. Musaev, Department of Computer Science, University of Alabama, Tuscaloosa, AL, USA; email: aibek@cs.ua.edu; D. Wang, Sunmi US Inc, USA; email: jnuwangde@gmail.com; D. Irani, Google, USA; email: danesh@google.com; S. Webb, Web Gnomes, USA; email: steve.webb@gmail.com; J. E. Ferreira, Department of Computer Science, University of Sao Paulo, Sao Paulo, Brazil; email: jef@ime.usp.br.

1533-5399/2020/03-ART2 $15.00
https://doi.org/10.1145/3374214

1 INTRODUCTION

Benchmark Data Sets. The introduction of quantitative performance evaluation on standard benchmark data sets, with well-defined ground truth, enabled precise comparisons among different machine learning (ML) algorithms. Classic examples of such standard data sets include MNIST [40] and CIFAR [41], which are fixed and fully annotated. They are equivalent to the mathematical concept of *universe* in set theory, where only values in the universe are considered. A significant part of ML research is concerned with optimizing ML classifiers and evaluating their performance within fixed standard data sets.

Optimization within Fixed Data Sets. Etzioni [20] has characterized this classic ML approach as "function approximation based on a sample." The evaluation of classifiers restricted to fixed universes has favored ML approaches (e.g., deep learning algorithms) that are optimized and specialized for the target data sets, e.g., LeCun's achieving more than 99% accuracy on MNIST in 1998 [62] using convolutional neural networks. This optimization process also leads to a side effect, called overfitting, where classifier performance degrades significantly when tested with new data from outside the original universe.

Artificial Reality. Fixed data sets are the first examples of bounded environments we call *artificial reality*, universes populated by well-known ground truth. ML classifiers are trained, optimized, and evaluated in artificial realities due to their need for ground truth for evaluation. We recognize that in their own sub-domains, artificial realities are valid sub-models of the actual reality. However, this recognition also creates the question of degree of validity of each artificial reality as sub-model—specifically, how much of the actual reality is the artificial reality able to cover? This coverage question becomes increasingly relevant as the actual reality continues to grow and change as the real world evolves, while artificial realities are defined by the limited ground truth available at their creation.

Evolving Actual Reality. In contrast to the static artificial reality, the explosive growth of big data from the actual reality has been described as "90% of the data in the world today has been created in the last 2 years" [31]. For example, smartphones became the first device to reach 1B deployments in 2012, and they generate huge amounts of data through social media and sensors such as cameras. Twitter reports 500M new tweets/day [29], and Facebook generates 4PB/day of new content [28]. Another example consists of many millions of surveillance video cameras in cities such as London and Beijing. While we recognize the validity and importance of fixed knowledge, e.g., images of apples, the focus of this article is on the new knowledge continuously being generated by the evolving actual reality.

Coverage of Artificial Reality. The recognition of artificial realities as sub-models of the actual reality posits the question of how much of the actual reality a sub-model is able to capture. This coverage question is illustrated by the 2018 fatal accident, when an Uber self-driving car struck and killed a pedestrian. According to the NTSB preliminary report [6] released in November 2019, the main issue was the software (ML sub-model) not considering jaywalkers (humans outside of crosswalks) as high probability events. As the actual reality evolves over time, the initial gap between fixed data sets and actual reality would be expected to widen.

Artificial Novelty. Under the methodological explanation that ground truth in fixed data sets is necessary for precise evaluation, many ML studies have remapped important phenomena in the actual reality into artificial reality, e.g., concept drift [72], which has been studied within fixed data sets by cycling through subsets [73, 74]. The coverage question, i.e., whether classifiers trained within an artificial reality would apply to actual reality, remains unanswered. More serious problems arise when the focus on precisely comparable evaluations result in the exclusion of work on actual reality, which contains incomplete ground truth, as "lacking in evaluation rigor."

True Novelty. It is our contention that artificial reality has served the ML research community well and will continue to be very important, but it is high time to reincorporate actual reality into the universe of ML research. This inclusion can start from the coverage question—specifically, the widening gap between the evolving actual reality and the artificial reality, which is bounded by the original ground truth. Borrowing from signal processing terminology, the gap is filled with both signal and noise. A major research challenge is distinguishing the signal that we call *true novelty* and the "big noise" that surrounds and obscures the true novelty, including random noise, misinformation, and disinformation in many live real data sources. Examples of big noise in Internet applications and social media include: email spam (e.g., Reference [42]), web spam (e.g., Reference [43]), Wikipedia vandalism (e.g., Reference [44]), and social media spam (e.g., References [45–47]).

Risks of Ignoring True Novelty. From a self-contained artificial reality point of view, true novelty would be inconsequential, since it lies outside of the universe of artificial reality. For example, k-fold validation has been considered an acceptable model for many kinds of novelty. We believe this disregard for true novelty and actual reality by extension could explain the failures of AI systems when deployed in actual reality. All the attempted deployments, including the Uber autonomous driving system [6], Microsoft Tay chatbot [8], and Google Flu Trends [1–5], have demonstrated excellent performance within their own artificial reality, but failed when faced with true novelty outside the original universe: pedestrian outside of crosswalk that caused the Uber accident, racial slurs that caused Microsoft Tay inappropriate tweets, and new search terms that caused Google Flu Trends to make more than 100% prediction error in just four years.

Live Knowledge. Just like the ever-changing actual reality containing it, true novelty is being continuously generated. We call *live knowledge* the continuously growing set of validated true novelty to distinguish the long-term challenge from individual snapshots of actual reality. As an example, the problem of finding specific cases of unseen items in retrospectively filtered data sets (which become artificial novelty once the data set is fixed) would not be considered live knowledge due to their disconnection from actual reality after creation. More concretely, just adding pedestrians outside crosswalks would not make the Uber autonomous driving system accident-free. Similarly, just adding racial slurs into Tay's knowledge base would not prevent other kinds of unforeseen inappropriate behavior. Live knowledge requires a methodical and automated approach to filter big noise, find true novelty, and continuously incorporate the new knowledge into a system.

Focus on Event Detection. The issues raised by the recognition of fixed ground truth in artificial reality, true novelty beyond artificial reality, and continuously growing live knowledge are very broad. In this article, we focus on the specific case of factual event detection, with knowable ground truth on the facts. With a concrete example of LITMUS landslide information system [14], we show that live knowledge can be achieved through a judicious integration of complementary live data sources. We hope that such successes can change the perception of necessity for artificial reality

into encouragement, or at least tolerance, for more research efforts on true novelty and live knowledge.

The EBKA Approach. We introduce the evidence-based knowledge acquisition (EBKA) approach to distinguish true novelty from big noise and continuously accumulate live knowledge. EBKA automates the process of recognizing true novelty by integrating complementary data sources using several ML algorithms to handle big noise. The validated true novelty is continuously added to live knowledge through classifier adaptation. The main idea of EBKA is the separation of data sources into three groups: (1) primary sources (social sensors with high coverage, and big noise), (2) deterministic corroborative sources (high reliability, authoritative sources, with relatively low coverage), and (3) probabilistic supporting sources (adding evidence to likely positive cases). By judiciously integrating these different sources, EBKA is able to leverage their strengths to compensate for their limitations.

Event Detection with True Novelty. We built the LITMUS landslide information system [14] to illustrate the EBKA approach to find true novelty and accumulate live knowledge. The primary sources of LITMUS consist of social sensors that include Twitter and Facebook, with corroborative sources (e.g., newspapers) and supporting sources (e.g., NOAA [23]). When processed in real-time, the social sensors present both big noise and true novelty challenges. Applying the EBKA approach, LITMUS integrates the corroborative and supporting sources through a teamed classifier to meet the challenges and achieve high accuracy as well as coverage in the detection and tracking of landslides.

The rest of the article is organized as follows: Section 2 summarizes the related work on the various forms of artificial reality. Section 3 outlines the challenges of finding true novelty and live knowledge. Section 4 highlights the LITMUS landslide information system as a live knowledge real-world application. Section 5 describes the EBKA approach to address big noise and concept drift challenges simultaneously, illustrated by LITMUS. Section 6 suggests future research and development opportunities on live knowledge. Section 7 concludes the article.

2 RELATED WORK ON ARTIFICIAL REALITY

2.1 Fixed Data Sets That Constitute Artificial Reality

The performance of supervised ML algorithms depends critically on the quality of training data: the purer the ground truth, the more accurate the classifier. Fixed data sets are considered to have full ground truth, and thus they became idealized environments in which to test many ML algorithms. There are several alternative ways to concentrate ground truth for evaluation, and we summarize three major variants according to the assumption they make about data sources: Fixed Data, Clustered Cata, and Continuity of sensor source.

Fixed Data Sets as Artificial Reality. We start from a recap of fixed data sets, popularized by TREC [38] data sets for IR, MNIST [40] and CIFAR [41] for ML, and many more [39]. High-quality ground truth data have favored deep learning (DL) algorithms, e.g., LeCun's more than 99% accuracy on MNIST in 1998 [62]. Fixed data sets are the first group of valid testing environments that form an artificial reality, with potentially widening gaps from the actual reality.

Retrospectively Clustered Data Sets. To apply ML algorithms to real-world data sets, an active area of research focuses on retrospectively filtered data sets, e.g., from social media, usually clustered on specific events [69] or a theme. For example, Sakaki et al. [63] studied Twitter reports on earthquakes by filtering out the noise (irrelevant tweets). A survey on this class of studies [69]

mentions examples of noise, including meaningless messages, polluted content, and rumors, all of which negatively affect the performance of ML classifiers. The filtering techniques include unsupervised learning, k-means clustering [84], customized filters for tweets [63, 64], social media analysis [65, 66], and cross-domain classification [88–90].

Clustered Data Sets Become Fixed. Although the clustered data sets typically started from realworld data streams, once created they became fixed data sets. As a result, clustered data sets also belong to artificial reality category. One possible explanation of this transformation is that the majority of current ML algorithms require ground truth for quantitative evaluation. In addition to retrospective clustering of event data (usually from social media), the transformation into fixed data sets also affects several initiatives aimed at real-world data streams, including never-ending learning[18], lifelong learning [19], open set recognition [15], and open world recognition [16, 17]. Their analyses mainly used fixed data sets or retrospectively filtered clusters transformed into fixed data, both in artificial reality.

Continuity in Data Streams. The area of data streaming [67, 68] usually refers to physical sensor data processing, e.g., readings of temperature and atmospheric pressure. Physical sensors in the real world produce time series data and data streaming work often assuming the data come from the same sensors, with predictable variations bounded by physical models of the real world. When the actual reality evolves beyond the known physical models, e.g., the appearance of unprecedented ozone hole over Antarctica since 1979, the discovery was delayed to 1985 [7] due to data assimilation algorithms that filtered out such "physically impossible" data from the Nimbus-7 satellite. Streaming data with filtering based on continuity assumptions from previous known models would fall into the artificial reality category when the actual reality evolves beyond the previous models.

Ground Truth and True Novelty. The dependency of supervised ML algorithms on ground truth (and the dependency of unsupervised algorithms on low noise levels) leads to a confined artificial reality, with three representative groups outlined above: fixed data sets, retrospective clusters, and continuity. While they are able to capture the knowledge within an artificial reality, the coverage question illustrates the gap between an artificial reality and an (evolving) actual reality. This gap will be called true novelty.

2.2 Terms Redefined in Artificial Reality

A major difficulty in true novelty and live knowledge consists of the double meaning of several keywords when their original interpretation from the actual reality becomes restricted by a much smaller artificial reality. An example of this double meaning happened to the term "real-time" in the context of artificial reality, instead of the computer science normal meaning of "real-time." Concretely, "real-time event detection" is part of the title of a highly cited retrospective study [63] in the context of clustered data sets (artificial reality). Their paper uses the term to refer to the relative distance between the timestamp of an event and the timestamps of tweets that enabled their classifier to decide on the event.

A second example is in the area of concept drift [72], which is a real-world problem due to gradual changes in the real world (the actual reality). However, typical papers on concept drift [73, 74] study the drift problem and solutions based on adaptation within the artificial reality of fixed data sets by rotating through subsets. Despite a significant number of papers on concept drift in artificial reality, the gap between the artificial concept drift and the actual concept drift in the real world (the

coverage question) has yet to be addressed. Concept drift will be elaborated in more detail in Section 3.2.

A similar redefinition happened with "novelty," which has different meanings within artificial reality compared to actual/true novelty. Some recent papers have focused on real-time novelty in the actual reality, e.g., TwitterNews [77] and GeoBurst [79], using clustering algorithms related to unsupervised learning. Unfortunately, their tweets appear to be lacking in corroboration, and thus their systems would be vulnerable to disinformation such as fake news.

2.3 Related ML Approaches to Acquiring Knowledge

Some of the ML techniques have an explicit goal of acquiring knowledge. Without entering into the discussion on artificial general intelligence, we mention four such ML techniques here to illustrate their purpose and limitations.

Reinforcement Learning. With the success of AlphaGo [92] and AlphaZero, reinforcement learning [91] has demonstrated super-human capability in well-defined games such as Go. However, their ability to exceed human capability in depth does not address the coverage question. In fact, game-playing programs represent stylized and limited artificial reality, with adaptation to game rule changes as open research challenges. Using the Uber accident example, it is unclear how reinforcement learning would handle unbounded true novelty beyond the specific case of human crossing a road outside of crosswalks.

Transfer Learning. Although more of a knowledge amplification approach instead of new knowledge acquisition, transfer learning (survey by Pan [70], with an update by Weiss [71]) aims at automating the creation of classifiers in the target domain by reusing (parts of) the classifier from a source domain. However, knowledge transfer process based on functional mapping also transfers/maps the limitations of the source. For example, consider a source domain classifier trained within the artificial reality of a fixed data set, or retrospective clustered data, and therefore incapable of detecting true novelty. It is inevitable that the target classifier will inherit the same limitations of the source classifier, within the confines of artificial reality.

Active Learning. In ML, human input has been considered the gold standard in the generation of ground truth. Specifically, active learning [85] uses human experts or crowdsourcing to manually label new training data. There are two general limiting factors of active learning: accuracy problems and (human) resource scarcity. First, the accuracy of human labeling depends heavily on the level of expertise and other human factors such as fatigue, and adding incentives does not necessarily help [86]. Second, human resources remain extremely limited compared to the rapidly growing big data being generated by physical and social sensors. ImageNet [94] illustrates both the success and limitations of human labeling: It has achieved order-of-magnitude improvements in labeled image collection size, but it is unlikely that it can be extended to capture true novelty from exponentially growing new big data.

Automated Machine Learning (AutoML). The many steps involved in typical ML work have spurred the efforts to automate the ML process (AutoML). As described in a recent book [93], the automation has occurred in several areas, including hyperparameter optimization and learning about the search process for the best classifiers, with useful software tools such as Auto-WEKA and Hyperopt-Sklearn. Perhaps as expected, these areas of successful AutoML start from the assumptions of artificial reality and well-defined ground truth, enabling the algorithmic optimization of search process to find the best approximation function.

3 CHALLENGES IN FINDING AND MAINTAINING LIVE KNOWLEDGE

3.1 Finding and Validating True Novelty

The first step in the quest for live knowledge is the automated discovery of true novelty in the midst of big noise that includes random data, misinformation, and disinformation. A traditional way to avoid the big noise challenge is to remap the novelty discovery problem back into artificial reality by making assumptions such as Fixed Data, Clustered Data, and Continuity (Section 2.1). However, these assumptions also preclude true novelty. First, classifiers trained under Fixed Data Sets have inherent difficulties with true novelty beyond the original fixed training data, as shown by Microsoft Tay chatbot. Second, analyses on Clustered Data have difficulties when applied to different clusters and indistinct clusters in true novelty, in addition to fixed data set constraints. Third, algorithms relying on Continuity would disallow the outliers considered by their physical model as noise. This would "throw out the baby with the bath water," since true novelty often appears (at least initially) as outliers.

Live knowledge requires the detection of true novelty whether they arise suddenly or grow gradually over a long period of time. There are similarities and differences in the handling of true novelty over different time scales. This section outlines the problem of short-term true novelty that arises suddenly. In the next section (3.2) the long-term growth of true novelty (a.k.a. concept drift) will be described.

Meaningful Outliers. The first challenge in finding true novelty in social sensors is that the discrete data items from millions of social media accounts are independent of each other. Therefore, there is no Continuity in social channels. Furthermore, some standard statistical assumptions, e.g., all noise being randomly generated with signals following well-behaved distributions such as Gaussian, would discard all outliers as noise. Instead of making Continuity and such statistical assumptions, true novelty detection on social sensor data need to carefully consider potentially meaningful outliers (e.g., first posting of an event), which may become a trend.

Meaningful True Novelty. The excellent accuracy of deep learning (DL) algorithms on fixed data sets reflects their being optimized approximation functions for fixed training data. This optimization also introduces instabilities (e.g., overfitting), leading to DL classifiers considering unseen new data irrelevant due to their being outside of training data. Instead of assuming Fixed Data, true novelty detection on social sensor data needs to carefully consider new data beyond the original training data.

Actionable Real-time Information. The "novelty" contained in retrospectively generated fixed data sets reflects only the ground truth covered by artificial reality. When true novelty arises, e.g., in the Uber accident and Microsoft Tay chatbot, the classifiers trained within artificial reality have reliability issues. To achieve actionable real-time information in actual reality, we need to find true novelty outside the traditional assumptions of Fixed Data or Clustered Data.

3.2 Long-term True Novelty Challenge (Concept Drift)

According to a survey [69], reports on event detection from real social sensors typically have followed the Clustered Data approach by analyzing retrospectively filtered data on large events [63]. This was feasible for events with many tweets (sometimes called bursts). The classifiers and models trained from large clusters on such events have been less successful when applied elsewhere, probably due to the coverage question, and also the differences among the clusters from different

events, e.g., earthquake vs. hurricane. More fundamentally, the long-term contextual changes in social media would have affected the accuracy of classifiers trained from fixed data sets.

The contextual changes have been called concept drift [72], defined as a change (over time) of class conditional probability $p(X,y)$, where X is the set of input variables and y the target variable. Technically, changes in data (both X and y) may change the prior probabilities of classes $p(y)$, the class conditional probabilities $p(X/y)$, and posterior probabilities of classes $p(y/X)$, affecting the prediction. Informally, concept drift is analogous to generation gap, where an "old" classifier has difficulties understanding "young" social postings containing new social slang and jargon that appeared after the older generation training data were created.

Concept Drift Challenge. Concept drift [72] describes the evolution of contextual content in actual reality over a period of time, typically years. Examples include the language used in social media and seasonal changes in scenery. Concept drift affects all ML classifiers trained by fixed training data, but tested over real-world data sets that span a long time. An early example was Google Flu Trends [1], which initially reported very high accuracy (more than 97% in 2009), when predicting flu pandemic areas using (millions of) browser search items associated with the flu. By 2013, the original model's predictions degraded by more than 100% due to changes in the search

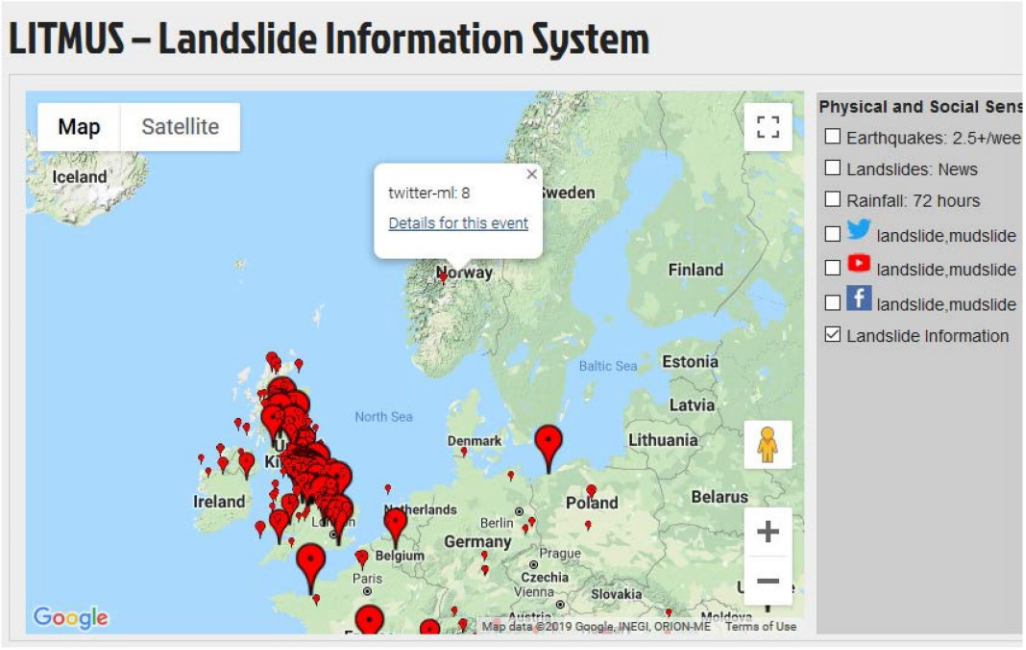

Fig. 1. Sample Screenshot of Landslides Reported by LITMUS.

terms that people used [2–5]. The LITMUS data we collected confirm serious concept drift (see Figure 3). Concept drift represents a major difficulty for the reproducibility of ML classifier performance on evolving social media: strong correlations today may—and almost invariably will— become weaker tomorrow.

4 LIVE KNOWLEDGE ON ACTUAL EVENTS REPORTED IN SOCIAL MEDIA

4.1 Illustrative Application: Live Knowledge on Landslides

Natural disasters are important real events with significant social and economic impact (billions of dollars per year) around the world. Some disasters have dedicated physical sensors for their detection, e.g., earthquakes are measured accurately by USGS Global Seismographic Network (GSN [21]). However, events such as landslides are more difficult to detect physically due to their localized impact, and humans are often the first finders (and responders) of landslides. In recent years, social sensors (e.g., Twitter) have become increasingly important and timely sources of landslide reports, making them a good illustrative live knowledge application.

LITMUS Demo System. LITMUS landslide information system [13] demonstrates effectively the Evidence-Based Knowledge Acquisition (EBKA) approach, described in more detail in Section 5. Figure 1 shows a sample screenshot of the LITMUS demo system, with landslides reported during the last month. Clicking on a pin brings up the "Details" tab, and clicking the Details tab opens a list of relevant tweets (left side of Figure 2). Clicking an item on the list opens the posting itself (right side of Figure 2). LITMUS integrates several primary social sources (Twitter, Facebook, and YouTube), which contain significant noise. In addition, LITMUS utilizes EBKA integration of corroborative sources (e.g., reputable newspapers) and supporting sources (e.g., NOAA [23]) to filter big noise and find true novelty, achieving excellent accuracy and coverage [60, 61] (more details of evaluation in Section 5.5).

Fig. 2. Tweets found on Norway landslides and an example (2019-08-05).

Table 1. LITMUS Data Sets as Illustrative Example of Live Knowledge

LITMUS data	2014 (monthly)	2015 – 17 (monthly)	2018 (monthly)
Relevant Samples	~5K to ~50K	~5K to ~45K	~5K to ~50K
Landslides Found	Hundreds	Hundreds	About a thousand
Positive Example	[Aug. 22, 2016] Train derails in Tokyo after landslide *(URL: photo of derailed train)*		

Negative Example	[Aug. 2016] Tropical Strom #Chanthu dropped 47mm in Tokyo. Moving north fast with landslide threat. @cnntoday @cnni *(URL: radar image of tropical storm Chanthu)*

4.2 Challenges in Finding True Novelty in Social Media

Live Social Sensor Data. LITMUS illustrates well the technical challenges in acquiring live knowledge from live social sensors, since the real-world social sensor data do not follow the common assumptions (discussed in Section 2):

1. No Continuity: Social postings come from millions of different accounts, not time seriesfrom the same sensor;
2. Not Fixed: true novelty arrives continuously from growing social sensor channels;
3. Real-Time: New data must be processed in near real-time for early detection and trackingof landslides of all sizes, not waiting for large events to unfold that enables clustering of data.

Big Noise in LITMUS Data. In the LITMUS data set, 90+% of tweets containing the keywords "landslides/mudslides" refer to non-disaster topics, e.g., landslide victories in elections and sports matches and a popular rock song entitled "Landslide." Table 1 shows the approximate size of LITMUS collected relevant data sets from 2014 to 2018 (top two rows), plus a positive example of

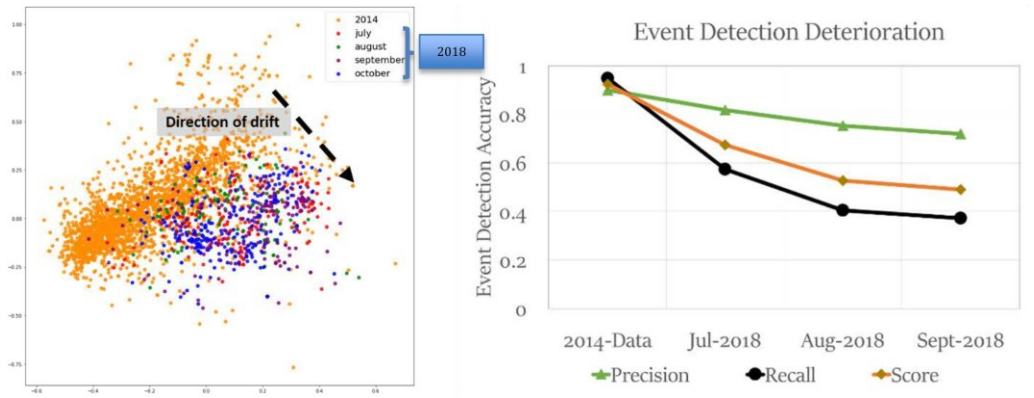

Fig. 3 . Degradation of Static Classifier Due to Concept Drift (from 2014 to 2018).

relevant tweet and a negative example (bottom two rows). Classic ML algorithms are able to handle big noise adequately, without reliance on Continuity, Fixed Data, or Clustered Data assumptions.

Concept Drift in LITMUS Data. The LITMUS data set confirms a clear concept drift and longterm true novelty, with data sets showing monthly oscillations in landslide data as well as true drift on the scale of a few years [61]. The right side of Figure 3 shows the significant accuracy loss of the same classifier (trained with a manually labeled subset of 2014 data) from 2014 (left column) to 2018 (three columns on the right). The left side of Figure 3 shows a direct visualization of terms often used in tweets on landslides. The graph is obtained by converting the tweets to numbers using word2vec [33], followed by dimensional reduction through principal component analysis (PCA) normalized to the interval [–1, +1]. In Figure 3, the light orange dots (upper left) represent the 2014 data, and terms in 2018 have migrated towards the lower right (dark red, green, purple, and blue dots) with a clearly visible drift.

In summary, LITMUS is a good real-world application with big noise and concept drift challenges that cannot be satisfied under artificial reality assumptions such as Fixed Data, Clustered Data, and Continuity. To provide real-time, actionable landslide information, LITMUS needs live knowledge from automated discovery, validation, and incorporation of true novelty. These requirements are achieved through the EBKA approach described in the next section.

5 FROM TRUE NOVELTY TO LIVE KNOWLEDGE THROUGH EBKA

In this article, a *real event* (e.g., a landslide) is characterized by a triple in the space-time continuum: a label (e.g., landslide), a physical location (e.g., Oso, Washington State of USA), and a time window (March 22, 2014). A real event has a meaningful topic label and may have varied (non-zero) sizes in space and time. We are primarily interested in single events at human scales due to the social media reporting. Events at microscopic or astronomical scales are left for future research topics. Most natural disasters fall into the range of interest, including hurricanes and landslides.

5.1 Evidence-Based Knowledge Acquisition (EBKA)

EBKA Information Integration. The main idea of EBKA is to integrate diverse information sources to address both the big noise challenge and the true novelty challenge simultaneously. LITMUS has three kinds of sources. First, the primary sources (social sensors including Twitter, Facebook, and YouTube) have wide coverage, but big noise problems. Second, deterministic corroborative sources (e.g., reputable news reports [26, 27] on landslides) have high reliability, but low

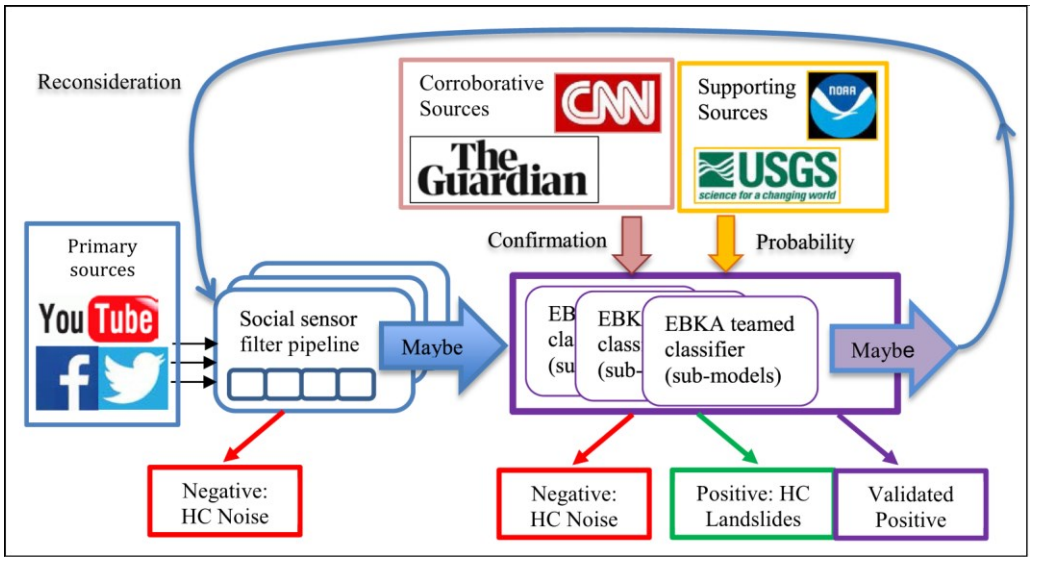

Fig. 4. Evidence-Based Knowledge Acquisition (EBKA) and LITMUS Data Integration.

coverage. Third, probabilistic supporting sources add likelihood estimations with physical sensor data and models. For example, earthquakes (USGS GSN [21]) and rainfall (NASA TRMM [22]) increase the probability of landslides, and NOAA provides a risk model of landslides [23].

EBKA Data Flow and Filtering. Figure 4 illustrates the LITMUS implementation of EBKA. The primary sources appear on the left, with corroborative sources and supporting sources on the top.

The social sensor data (with big noise) go through two filtering stages. The first stage consists of a sequence of ML filters (marked as "social sensor filter pipeline" in the middle left of Figure 4) that use classic ML classification algorithms (e.g., the WEKA toolkit [36]) to filter big noise adequately [42, 48–59].

Capturing True Novelty. The second stage consists of a teamed classifier (to be elaborated in following sections) that incorporates the knowledge acquired from the corroborative and supporting sources. The combination achieves high accuracy on detecting short-term true novelty. On the 2014 landslide data [59], which was improved by deep learning (DL) tools such as Keras [35] and TensorFlow [34], LITMUS achieved about 98% precision and recall [60] for 2014 data. However, addressing the concept drift challenge (Figure 3) required teamed classifiers, elaborated in following sections.

5.2 Teamed Classifiers for Live Knowledge

Teamed Classifier in LITMUS. To avoid confusion with the broad area of ensemble learning [83], we adopt the term *teamed classifier* (also called committee classifier) to denote the LITMUS classifier, which contains several sub-models for a variety of reasons, including location finding (part of the social sensor filter pipeline) and EBKA (middle right of Figure 4). The function of submodels may differ for each source and their weights may vary for each posting. Specific sub-models of interest (e.g., EBKA) will be described in Section 5.4.

Goals of Teamed Classifier Design. Teamed classifiers [67–81] have improved the management of knowledge on complex classification problems such as IBM Watson [37] and restricted versions of concept drift [67]. A simple example of concept drift is the change of scenery due to the four seasons (e.g., snow in winter and flowers in spring). Typically, each sub-model handles a narrow case (e.g., a single season), achieving better decisions with high confidence while maintaining relative simplicity through specialization. As we find validated true novelty, sub-models trained by the new ground truth are added into the teamed classifier. A growing teamed classifier becomes the repository of live knowledge, in addition to the data set. The LITMUS example is explained in Section 5.4.

Optimizing Teamed Classifiers. A major technical challenge in teamed classifiers is the management and integration of specialized sub-models through one or more weight functions that optimize the group decision, e.g., by minimizing the error in classification. As illustration, consider Equation (1) with a multi-dimensional metric state space of real-world events, denoted by ev_k and sub-model x_i evaluating the likelihood of ev_k actually occurring; the potential classification error is calculated as the distance between $decision_i$ (sub-model's estimate of x_i) and actual event occurrence: $err_i(x_i,ev_k) = dist_k(decision(x_i,ev_k), actual(ev_k))$, where $actual(ev_k)$ represents the ground truth on ev_k, and $dist_k$ the amount of error made by sub-model x_i. The objective of the teamed classifier evaluation becomes the search for weight functions $weight_i$ that minimize (or approximate the minimal) total error over the entire team: $total_{erri}(x_i) =_i err_i(x_i,ev_k)$

Equation (1) Optimization Process to Find Best Weight Functions for Ensemble Classifiers

$$\min_{sub-model} err_i(x_i,ev_k)$$

$$i$$

$$n$$

$$= \min_{sub-model} \sum_{i} weight_i * \sum_{k=1} dist_k(decision(x_i, ev_k), actual(ev_k)) .$$

An informal interpretation of Equation (1) is that the error minimization process will converge by giving lower $weight_i$ to sub-models $\{x_i\}$ that make bigger mistakes (large $err_i(x_i, ev_k)$ on the right side of equation).

Equation (1) shows both the strength and weaknesses of common ML assumptions (e.g., Fixed Data Set and Clustered Data). On the positive side, given well-defined ground truth from these assumptions, the optimization converges. However, the ground truth based on known past data would ignore new events due to true novelty challenges (Figure 3). This is the problem to be addressed by EBKA, which is capable of recognizing and acquiring true novelty reliably.

5.3 LITMUS Teamed Classifiers Using EBKA

Corroboration and Support. In the LITMUS teamed classifier, there is a dedicated sub-model for each corroborative and supporting source. For example, there is a sub-model that maintains all CNN.com reports on landslides. A landslide tweet that matches a corroborated location-time (e.g., in the CNN.com sub-model) is considered positively identified. In contrast, a supporting source sub-model (e.g., coincidental heavy rain) only increases the probability of a co-located landslide.

Sources of New Knowledge. The high confidence placed on corroborative sources is due to the publication requirements in reputable news organizations. A typical newspaper requirement consists of confirmed corroboration from multiple independent sources. Newspapers such as *The New York Times* and *The Guardian* have good reputations due to the very low error rates in their articles. If they report an event, it is highly likely to have occurred. The main contribution of corroborative sources to LITMUS (and EBKA more generally) consists of the new knowledge they generate, which is used as ground truth in the selection of training data for new sub-models at team-level adaptation (described in Section 5.4 and Figure 5).

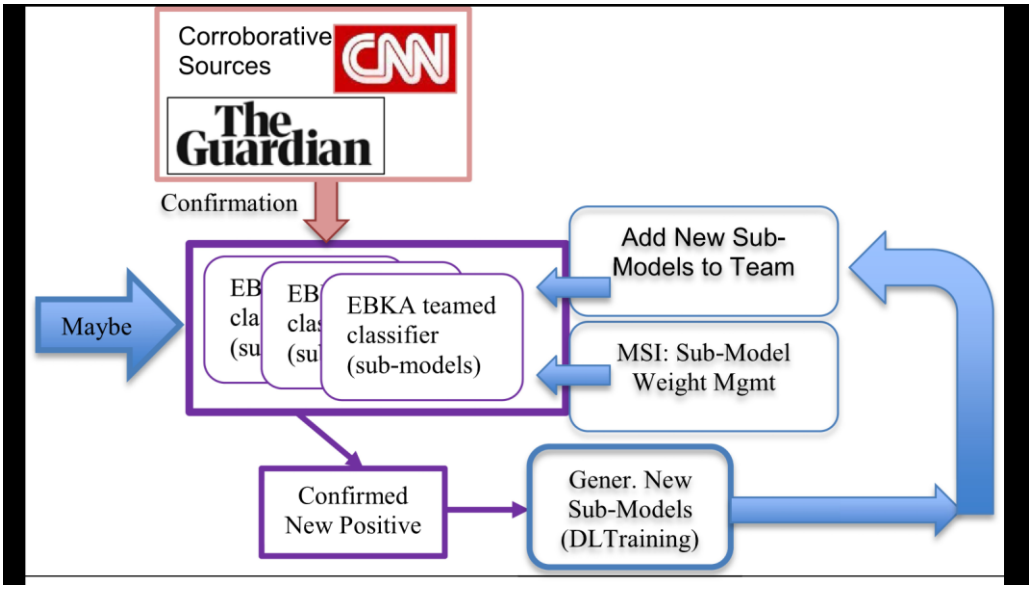

Fig. 5. EBKA Architecture (with LITMUS as Illustration).

Table 2. Improved F-score with EBKA under Imbalanced Learning

	Static (DL)	Adaptive (DL)	Primary Source items	Corroborative items	Percent of Confirmed Data	Improvement on F-score
2014 Data	0.91	0.96	NA	NA	NA	NA
Jul-2018	0.70	0.88	7205	189	2.62%	125.5%
Aug-2018	0.56	0.90	14245	106	0.74%	159.2%
Sep-2018	0.57	0.90	4867	193	3.97%	156.7%
Oct-2018	0.70	0.88	15847	249	1.57%	126.1%
Nov-2018	0.38	0.86	7084	885	12.49%	225.7%
Dec-2018	0.75	0.99	4873	223	4.58%	132.0%

Table 3. Improved Coverage (# landslides) with EBKA and Imbalanced Learning

LITMUS	Jul-18	Aug-18	Sep-18	Oct-18	Nov-18	Dec-18
Static DL	480	644	365	501	508	149
Adaptive DL	480+398	644+681	365+513	501+646	508+772	149+370
Improved Coverage	82.92%	105.75%	140.55%	128.94%	151.97%	248.32%

Limitations of Corroborative Sources. Given their high reliability, a question that arises is whether they can be used as sole sources. The answer is negative, due to delay and limited coverage. First, their requirement for independent corroboration causes some delay before publication (usually hours to days). Second, typical newspapers only publish events of interest to a wide audience, ignoring small events of limited impact. Tables 2 and 3 show that corroborative sources only publish a few percent of all landslides found (and verified) by LITMUS from social sensors.

LITMUS Teamed Classifier Output Categories. Through EBKA, LITMUS teamed classifier divides social postings into one of four categories: (1) very low probability postings, considered definitely irrelevant and marked as HC Negatives (high confidence) in Figure 4; (2) very high probability events (HC Positives); (3) corroborated and confirmed landslides; and (4) uncertain postings with intermediate probability (apparently serious tweets, but insufficient supporting evidence).

Accuracy of Teamed Classifier. The LITMUS teamed classifier is able to achieve high precision and recall (Table 3 and Table 4) by recognizing the clearly irrelevant items (category 1) and clearly relevant items (categories 2 and 3). At the same time, the boundary/uncertain cases (category 4) are sent back to the teamed classifier for reconsideration. As more evidence and corroboration are accumulated by the adaptive teamed classifier (see Section 5.4), a higher confidence decision can be made with better knowledge.

5.4 Automated EBKA through Adaptive Teamed Classifier

Two-level Adaptation of Teamed Classifier. EBKA incorporates new knowledge from corroborative sources at two levels. At the sub-model level, new events from corroborative sources are added into their databases, which increase the number of confirmed events and new knowledge. Similar adaptation happens with supporting sources. At the team level, new sub-models are created with new training data and added to the team. The sub-model adaptation increases the realtime landslide detection capability of LITMUS, and the team-level adaptation handles long-term concept drift.

Sub-model-level Adaptation. In LITMUS, each corroborative sub-model maintains a database of confirmed events, indexed by location-time. As new events are published, they are inserted into the database, providing corroboration to all co-located landslide postings (that match the same location-time). Similarly, each supporting sub-model (e.g., NOAA [23]) maintains a database of landslide probability, indexed by location-time. When a location-time receives an updated probability, co-located landslide postings also update their probabilities. The sub-model adaptation is particularly effective in immediately incorporating supporting evidence, e.g., a new earthquake that occurred in Nepal would increase the probability of landslides occurring in that area during that time.

Team-level Adaptation. To address the long-term concept drift challenge (shown on the left side of Figure 3), EBKA incorporates new knowledge by introducing new sub-models into the teamed classifier. In LITMUS, the social postings on the new confirmed landslides of each month are used as positive training data set (with corresponding negative data) for creating new sub-models. The weighting function is carefully tuned to optimize the impact of new sub-models. The team-level adaptation process is illustrated in Figure 5. Although the corroborative sources only provide a few percent of verified landslides (Table 3), they are sufficient for handling concept drift, as shown on the right side of Figure 3.

EBKA Architecture. Taken together, Figure 4 (adaptive sub-models for classifying each social posting) and Figure 5 (team-level adaptation) illustrate the EBKA architecture to detect new events and capture true novelty. The two-level adaptation handles the different time scales of environmental change: sub-model adaptation for discovering new events (new landslides, from hours to days) and team-level adaptation for handling concept drift (social media language evolution, from months to years).

Better Decisions. A distinctive feature of the EBKA approach consists of the fourth category of output: uncertain postings with intermediary probability. Figure 4 shows that uncertain postings are sent back to the classification process. This recycling make sense for adaptive classifiers, since the decision may change (to better) when new information arrives at corroborative or supporting sources. For example, the first tweet on a new real landslide normally would be sent back for lack of supporting evidence, and its probability would increase by subsequent tweets due to EBKA. The tweets that are irrelevant to landslides would not receive additional support, and eventually they get filtered out.

5.5 Evaluation of EBKA in LITMUS

Accuracy Achieved by LITMUS. We have very encouraging experimental evaluation results in LITMUS that confirm the effectiveness of EBKA approach. Table 3 shows that a modest amount of confirmed data (second column from the right, typically a few percent) is sufficient for improving the F-score of Teamed Classifier by more than 100%. The largest improvement (more than 200% in November 2018) appears to be correlated to the largest percent of confirmed data (12.5%), which indicates more data as well as more research would be warranted.

Improved Coverage. In classic ML, improved accuracy often requires a trade-off in coverage. By acquiring external evidence, the EBKA approach achieves significant improvements in F-score (rightmost column in Table 3) while maintaining wide coverage. Table 4 shows that Adaptive DL classifier achieves a strict superset of Static DL classifier. We believe that the EBKA approach is able to bypass the classic ML trade-off between false positives (FP) and false negatives (FN), because of the additional knowledge from corroborative and supporting sources.

6 FUTURE RESEARCH AND DEVELOPMENTS ON LIVE KNOWLEDGE

6.1 Smarter Applications Enabled by Live Knowledge

Smart Applications within Artificial Reality. Many of current smart (or intelligent) applications have made assumptions similar to those discussed in Section 2 (Continuity, Fixed Data, and Clustered Data). For example, projects that built custom sensors [9] have developed applications such as air monitoring, with both Continuity and Fixed Data assumptions. They work well in the artificial reality of testbed environments but have difficulties with deployment in the open realworld environment. The transition difficulties have often been attributed to scalability issues such as cost or heterogeneity, but we believe that a more fundamental issue is the coverage question: Algorithms that work well within artificial reality may be missing important knowledge about the actual reality, as illustrated by the Uber accident.

Bridging the Gap. From the coverage question point of view, the growth of new big data from the actual reality is much faster than the (mainly human-annotated) ground truth in the artificial reality. While the work under the assumptions of Fixed Data, Clustered Data, and Continuity remain valid within artificial reality, their impact on the actual reality will be reduced by failures arising from the widening gap. Although traditional approaches such as active learning have limitations, we believe that an effective utilization of live knowledge through EBKA could reduce and eventually bridge the gap between smart applications and live real data.

Real-time Incident Detection. Widespread video cameras offer real-time monitoring, but their practical usage has been limited (mostly) to after-the-fact crime investigations and forensics. Similarly, smart transportation (e.g., incident detection and assistance) has relied on NASA-style

command centers (monitored by human operators). Effective acquisition of live knowledge would alleviate the current need for human intervention and enable the next generation of automated smart applications. Examples of live real applications that can benefit substantially from automatically acquired live knowledge include: responsive disaster management (e.g., accurate and timely detection of landslides and sudden rains), efficient transportation (e.g., real-time automated detection of congestions and accidents), and proactive public safety (e.g., real-time crime detection).

Video Event Detection and Tracking. Automated real-time object recognition and tracking for applications such as smart transportation and proactive public safety can be enabled by live knowledge from the live video data. We plan to apply EBKA with object tracking techniques [75–82] on the live video data from USP [11]. An illustrative example of complex incident detection is the case of a stolen vehicle that had its license plate replaced with another from a parked car. The LPR capability would not be able to detect such changes by itself. However, adding the contextual information (make and color of vehicle) to each license plate as live knowledge will enable the real-time detection of a stolen vehicle despite plate switch.

Informative Guidance during Disasters. Navigation systems gained a significant new capability with the launch of Google Maps Live View AR (Augmented Reality) on Pixel in March 2019 and available for Android and iOS smartphones since August 2019. An example of guidance systems during disasters [12] integrated 3D models, crowd behavior videos, and tweets from historical events to show good escape paths from Osaka underground shopping malls that would be quickly inundated by a tsunami caused by earthquake. One of the difficulties with traditional AR is their static view of environment, which may be changed by a disaster, e.g., buildings may have been toppled by an earthquake and roads covered by landslides. Live knowledge from uploaded videos (supported by many social media channels) and public safety apps such as USP Campus [10, 11] can provide fresh views of a changed landscape. Combined with AR navigation, live knowledge can enable the generation of up-to-date or new escape routes during and after disasters.

6.2 Improving EBKA with More Evidence

Looking forward to wider acquisition of live knowledge, particularly research and practice through EBKA, one of the important questions is the availability of reliable sources from which EBKA extracts live knowledge. For example, the reputable news sources used by LITMUS for information gathering on landslides can be reasonably expected to work well for the detection of many real events as they unfold. In addition, there are quite a few more reliable sources that EBKA can draw upon.

Authoritative Sources on Specific Areas. In many specific areas, there are mission-oriented agencies in charge that publish authoritative information on their areas of expertise. In epidemics, for example, the CDC (Centers for Disease Control and Prevention) publishes authoritative information (e.g., Ebola data [25]). For some areas of commercial interest, an increasing number of online services have been improving their accuracy, reliability, and coverage. An example relevant to live knowledge is accurate real-time micro-area weather forecasts with increasingly higher quality service providers such as weather.com and accuweather.com.

Measures of Reputation and Trust. Generally, there are several kinds of measures of reputation, e.g., number of followers for a Twitter account, number of downloads for a YouTube video, and Alexa's top 500 global sites ranking [30], where popularity suggests trust from the crowd. Although there are known threats to these measures (e.g., fake Twitter follower accounts), they may contribute as

supporting evidence. We believe the ongoing research efforts on trust and reputation [87] will help us improve the distinction of reputable sources.

Local Sources for Local Events. As an example of interesting topics in the reputation and trust area, local newspaper reporting of local events tends to have higher reliability [26, 27]. This kind of specialization is supported by EBKA through careful weight assignment as a function of event and source co-locality.

Human Input. Expert input in active learning [85] and general crowdsourcing (e.g., data entry through mobile apps [10, 11]) can add more corroborative and supporting evidence. The EBKA approach can benefit from direct human input (e.g., from mobile apps), particularly since it would not require real-time labeling, allowing for cross-checking and improvement of accuracy and coverage.

6.3 Practical Issues with Live Knowledge Acquisition

EBKA is a principled and practical approach to acquiring live knowledge from real-world social sensors. However, working with real-world sources means resolving some operational issues that can impede access to live knowledge. We mention three examples to illustrate these non-trivial technical issues that are stepping stones and building blocks towards live knowledge.

Instability of Real-world Data Sources. One of the limitations of sensor testbeds is their limited lifespan: They often depreciate when their research budget ends. In contrast, a big advantage of real-world sources such as social media is that they are maintained by other sources. However, live production sources also evolve and change over time outside of our control. For example, Instagram was one of the main sources for the original LITMUS, but an access policy change in June 2016 disallowed public data collection. Another example is the USGS official listing of landslides [24], which shut down in 2016. Flexible adaptation to changes in real data sources is an integral part of live knowledge acquisition process, not just EBKA.

Location-time Determination. An event is defined by its topic and location-time. On the time dimension, typical social sensor postings (e.g., tweets) are timestamped, and it is often reasonable to assume close time proximity to the event reported. In contrast, few tweets contain the GPS location of their origin, and few tweets are sent from the epicenter of an event. Fortunately, many social postings that refer to a real event also include an identifying term on its location, which is used by LITMUS to determine the location-time (primary key) of the event through tools such as CoreNLP [32] and localized software libraries for each country.

Live Knowledge in Multiple Languages. Although this article focused on social sensors in English, the knowledge of the world consists of the union of many languages. The integration of knowledge from multiple languages would enable much better EBKA performance, but such integration still requires significant research [57].

7 LIVE KNOWLEDGE VISION

Because “90% of the data in the world today has been created in the last 2 years” [31], unprecedented opportunities are being created by new big data, including social media, e.g., 500M tweets/day and millions of video cameras in many cities. However, the primary consumers of the explosively growing new big data have been humans. In our view, the ML focus on artificial reality, e.g., through Fixed Data, Clustered Data, and Continuity assumptions, caused the gap between the

artificial reality and the actual reality. While the artificial reality remains a valid research approach on a subset of actual reality, the growing gap demands our attention, as shown by the Uber fatal accident, the Microsoft Tay chatbot misbehavior, and Google Flu Trends shutdown.

Beyond artificial reality, we envision the research and development efforts on *live knowledge*, which automatically acquire real-time, validated, and actionable information for smart applications that must work in the actual reality, including smart transportation and disaster response. Live knowledge contains significant research challenges such as big noise and concept drift. From the new data, we need to distinguish and validate true novelty from random noise, misinformation, and disinformation that derailed the Tay chatbot. For the long term, we need to accumulate true novelty into live knowledge and keep incorporating it into smart applications that work in the actual reality.

To demonstrate the feasibility of achieving live knowledge, we describe the EBKA (evidencebased knowledge acquisition) approach to integrate information and find true novelty in the LITMUS landslide information system. LITMUS integrates three kinds of complementary data sources: primary sources with wide coverage (e.g., tweets on landslides), corroborative sources with high reliability (e.g., news reports), and probabilistic supporting sources (e.g., landslide likelihood model from NOAA). Through EBKA, LITMUS distinguishes true novelty and acquires new knowledge on landslides from this automated integration, and it is independent of the Fixed Data, Clustered Data, and Continuity assumptions. LITMUS achieves both high accuracy and wide coverage through four years of data, demonstrating the feasibility and promise of the EBKA approach towards live knowledge.

REFERENCES

[1] Jeremy Ginsberg, Matthew H. Mohebbi, Rajan S. Patel, Lynnette Brammer, Mark S. Smolinski, and Larry Brilliant. 2009. Detecting influenza epidemics using search engine query data. *Nature*. 457 (7232), 1012–1014.

[2] S. Cook, C. Conrad, A. L. Fowlkes, and M. H. Mohebbi. 2011 Assessing Google Flu trends performance in the United States during the 2009 influenza virus a (H1N1) pandemic. *PLoS ONE* 6, 8 (2011), e23610.86

[3] Google Flu Trends (GTF) failure story. [<https://en.wikipedia.org/wiki/Google_Flu_Trends>]. Retrieved November 9, 2019.

[4] Declan Butler. 2013. When Google got flu wrong. *Nature* 494, 7436 (2013), 155.

[5] David Lazer, Ryan Kennedy, Gary King, and Alessandro Vespignani. 2014. The parable of Google flu: Traps in big data analysis. *Science* 343, 6176 (2014), 1203–1205.

[6] NTSB preliminary report on the Uber fatal accident in Tempe, Arizona. [https://www.ntsb.gov/investigations/AccidentReports/Reports/HWY18MH010-prelim.pdf]. Retrieved November 9, 2019.

[7] Joseph Farman, C. Brian, G. Gardiner, and Jonathan D. Shanklin. 1985. Large losses of total ozone in Antarctica reveal seasonal ClOx/NOx interaction. *Nature* 315, 6016 (1985), 207.

[8] Microsoft Tay chatbot. [<https://en.wikipedia.org/wiki/Tay_(bot)>]. Retrieved November 9, 2019.

[9] Array of Things project at Github [https://arrayofthings.github.io/]. Retrieved November 9, 2019.

[10] Guia USP and Campus USP: mobile apps for users to communicate with campus police and obtain other information. *Available for iPhones (Apple Store) and Android devices (Google Play)*.

[11] J. E. Ferreira, J. A. Visintin, J. Okamoto, and C. Pu. 2017. Smart services: A case study on smarter public safety by a mobile app for University of São Paulo. In *Proceedings of the IEEE SmartWorld Congress*.

[12] Sohei Kojima, Akira Uchiyama, Masumi Shirakawa, Akihito Hiromori, Hirozumi Yamaguchi, and Teruo Higashino. 2017. Crowd and event detection by fusion of camera images and micro blogs. In *Proceedings of the IEEE International Conference on Pervasive Computing and Communications Workshops*.

[13] GRAIT-DM project and the RCN on Real-Time Big Data Analytics for Resilient Infrastructures in Smart and Connected Communities. [https://grait-dm.gatech.edu/]. Retrieved November 9, 2019.

[14] LITMUS landslide information service [https://grait-dm.gatech.edu/demo-multi-source-integration/]. Retrieved November 9, 2019.

[15] Open Set Recognition [<https://www.wjscheirer.com/projects/openset-recognition/>]. Retrieved November 9, 2019.

[16] Open World Machine Learning [<https://www.cs.uic.edu/~liub/open-classification.html>]. Retrieved November 9, 2019.

[17] Bendale Abhijit and Terrance Boult. 2015. Towards open world recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 1893–1902.

[18] T. Mitchell, W. Cohen, E. Hruschka, P. Talukdar, B. Yang, J. Betteridge, A. Carlson, B. Dalvi, M. Gardner, B. Kisiel, and J. Krishnamurthy. 2018. Never-ending learning. *Commun. ACM*, 61, 5 (2018), 103–115.

[19] Bing Liu. 2017. Lifelong machine learning: A paradigm for continuous learning. *Front. Comput. Sci.* 11, 3 (2017), 359– 361.

[20] Etzioni Oren. 2018. Breaking the mold of machine learning: Technical perspective. *Commun. ACM* 61, 5 (2018), 102–102.

[21] USGS Global Seismographic Network [http://earthquake.usgs.gov/monitoring/gsn/]. Retrieved November 9, 2019.

[22] NASA TRMM. Tropical Rainfall Measuring Mission: Satellite monitoring of the intensity of rainfalls in the tropical and subtropical regions. Retrieved on November 9, 2019 from http://trmm.gsfc.nasa.gov/.

[23] NOAA landslide risk predictions for locations with 7-day rainfall: [https://trmm.gsfc.nasa.gov/trmm_rain/Events/latest_7_day_landslide.html]. Retrieved November 9, 2019.

[24] USGS list of landslide events—Landslide Hazards Program. http://landslides.usgs.gov/recent/. Accessed on September 15, 2015. Discontinued in July 2016 and unavailable as of August 2019. Its previous content may have been preserved by the Internet Archive [http://www.archive.org/].

[25] CDC data on Ebola outbreaks [https://www.cdc.gov/vhf/ebola/history/chronology.html]. Accessed on August 8, 2019.

[26] List of Most Trusted News Sources, compiled by Pew Research Center [http://www.pewresearch.org/fact-tank/2014/10/30/which-news-organization-is-the-most-trusted-the-answer-is-complicated/]. Accessed on September 11, 2015. [27] BBC poll on trusted news sources per country, [http://www.globescan.com/news_archives/bbcreut_country.html]. Accessed on September 15, 2015.

[28] Facebook data statistics. [https://www.brandwatch.com/blog/facebook-statistics/] and [https://www.quora.com/ How-many-bytes-does-Facebook-store-every-day]. Retrieved July 25, 2019.

[29] 500M/day tweets on Twitter. [https://www.internetlivestats.com/twitter-statistics/]. Retrieved July 25, 2019.

[30] Alexa's Top 500 Global Sites ranking [https://www.alexa.com/topsites]. Retrieved November 9, 2019.

[31] IBM. 2017. "10 Key Marketing Trends for 2017" [<https://www.ibm.com/downloads/cas/XKBEABLN>]. Retrieved April 8, 2019.

[32] The Stanford Natural Language Processing Group, "Stanford CoreNLP," [http://nlp.stanford.edu/software/corenlp.shtml]. Retrieved January 2, 2015.

[33] Mikolov Tomas, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient estimation of word representations in vector space. *ArXiv Preprint ArXiv:1301*.3781 (2013).

[34] TensorFlow project website [https://www.tensorflow.org/]. Retrieved November 9, 2019.

[35] Keras documentation website [https://keras.io/]. Retrieved November 9, 2019.

[36] WEKA project website [http://www.cs.waikato.ac.nz/ml/weka/]. Retrieved November 9, 2019.

[37] DeepQA Project and Watson Q&A System created by the group at IBM Research [http://researcher.watson.ibm.com/researcher/view_group.php?id=2099]. Retrieved November 9, 2019.

[38] NIST Text Retrieval Conference (TREC) English documents, 2001. http://trec.nist.gov/data/docs eng.html. Retrieved November 9, 2019.

[39] List of data sets for machine learning research [https://en.wikipedia.org/wiki/List_of_datasets_for_machine_learning_research]. Retrieved November 9, 2019.

[40] MNIST (Modified National Institute of Standards and Technology database) [https://en.wikipedia.org/wiki/MNIST_database]. Retrieved November 9, 2019.

[41] CIFAR-10 (Canadian Institute For Advanced Research), labeled subset (60,000 images) of the 80 million tiny images data set, with 10 classes. [https://www.cs.toronto.edu/~kriz/cifar.html]. The associated CIFAR-100 is a superset that contains 100 classes. Retrieved November 9, 2019.

[42] Calton Pu, Steve Webb, Oleg Kolesnikov, Wenke Lee, Richard Lipton. 2006. Towards the integration of diverse spam filtering techniques. In *Proceedings of the IEEE International Conference on Granular Computing*.

[43] De Wang, Danesh Irani, Calton Pu. 2012. A perspective of evolution after five years: A large-scale study of web spam evolution. *Int. J. Coop. Inf. Syst.* 23, 2 (2014).

[44] Qinyi Wu, Danesh Irani, Calton Pu, Lakshmish Ramaswamy. 2010. Elusive vandalism detection at Wikipedia: A text stability-based approach. In *Proceedings of the 19th International Conference on Information and Knowledge Management*.

[45] De Wang, Danesh Irani, and Calton Pu. 2014. SPADE: A social-spam analytics and detection framework. *Soc. Netw. Anal. Mining* 4, 1 (2014).

[46] Danesh Irani, S. Webb, K. Li, and C. Pu. 2011. Modeling unintended personal information leakage from multiple online social networks *IEEE Internet Comput*. 15, 3 (May–June 2011), 13–19.

[47] Jenny Luebbe. 2015. How dirty is social data? An analysis of social spam. *Netw. Insights* (April 1, 2015). [http://www.networkedinsights.com/socialspam/].

[48] Aibek Musaev, De Wang, and Calton Pu. 2014. LITMUS: Landslide detection by integrating multiple sources. In *Proceedings of the 11th International Conference on Information Systems for Crisis Response and Management*.

[49] Aibek Musaev, De Wang, Chien-An Cho, and Calton Pu. 2014. Landslide detection service based on composition of physical and social information services. In *Proceedings of the IEEE International Conference on Web Services*.

[50] Aibek Musaev, De Wang, Saajan Shridhar, and Calton Pu. 2015. Fast text classification using randomized explicit semantic analysis. In *Proceedings of the IEEE International Conference on Information Reuse and Integration for Data Science*.

[51] Aibek Musaev, De Wang, Saajan Shridhar, and Calton Pu. 2015. Toward a real-time service for landslide detection: Augmented explicit semantic analysis and clustering composition approaches. In *Proceedings of the IEEE International Conference on Web Services*.

[52] Aibek Musaev, De Wang, and Calton Pu. 2015. LITMUS: A multi-service composition system for landslide detection. *IEEE Trans. Serv. Comput.* 8, 5 (2015), 715–726.

[53] D. Wang, A. Musaev, and C. Pu. 2016. Information diffusion analysis of rumor dynamics over a social-interaction based model. In *Proceedings of the IEEE 2nd International Conference on Collaboration and Internet Computing*.

[54] I. Tien, A. Musaev, D. Benas, A. Ghadi, S. Goodman, and C. Pu. 2016. Detection of damage and failure events of critical public infrastructure using social sensor big data. In *Proceedings of the International Conference on Internet of Things and Big Data*. 435–440.

[55] Qixuan Hou, A. Musaev, Y. Yang, and C. Pu. 2017. Towards multilingual support of landslides information service. In *Proceedings of the IEEE International Conference on Collaborative and Internet Computing*.

[56] A. Musaev and C. Pu. 2017. Towards multilingual automated classification systems. In *Proceedings of the IEEE 37th International Conference on Distributed Computing Systems*.

[57] A. Musaev, Q. Hou, Y. Yang, and C. Pu. 2017. LITMUS: Towards multilingual reporting of landslides. In *Proceedings of the IEEE 37th International Conference on Distributed Computing Systems*.

[58] A. Musaev, D. Wang, J. Xie, and C. Pu. 2017. REX: Rapid ensemble classification system for landslide detection using social media. In *Proceedings of the IEEE 37th International Conference on Distributed Computing Systems*.

[59] Aibek Musaev and Calton Pu. 2017. Landslide information service based on composition of physical and social sensors. In *Proceedings of the IEEE International Conference on Data Engineering*.

[60] Abhijit Suprem and Pu Calton. 2019. ASSED—A framework for identifying physical events through adaptive social sensor data filtering. In *Proceedings of the 13th ACM International Conference on Distributed and Event-based Systems*.

[61] A. Suprem, A. Musaev, and C. Pu. 2019. Concept drift adaptive physical event detection for social media streams. In *Proceedings of the World Congress on Services. Lecture Notes in Computer Science*, Y. Xia, L. J. Zhang (eds.). Springer, Cham, 11517.

[62] Yann LeCun, Léon Bottou, Yoshua Bengio, and Patrick Haffner. Gradient-based learning applied to document recognition. *Proc. IEEE* 86, 11 (1D998), 2278–2324.

[63] T. Sakaki, M. Okazaki, and Y. Matsuo. 2010. Earthquake shakes Twitter users: Real-time event detection by social sensors. In *Proceedings of the 19th International Conference on World Wide Web*. 851–860.

[64] X. Wang, F. Zhu, J. Jiang, and S. Li. 2013. Real time event detection in Twitter. In *Web-Age Information Management, Vol. 7923, Lecture Notes in Computer Science*, 502–513. Springer Berlin.

[65] K. Radinsky and E. Horvitz. 2013. Mining the web to predict future events. In *Proceedings of the 6th ACM International Conference on Web Search and Data Mining*. 255–264.

[66] M. Kitsuregawa and M. Toyoda. 2011. Analytics for info-plosion including information diffusion studies for the 3.11 disaster. In *Web-Age Information Management,Vol. 6897, Lecture Notes in Computer Science*, 1–1. Springer Berlin.

[67] Jonathan A. Silva, Elaine R. Faria, Rodrigo C. Barros, Eduardo R. Hruschka, Andre C. P. L. F. De Carvalho, and João Gama. 2013. Data stream clustering: A survey. *ACM Comput. Surv.* 46, 1 (2013), 13.

[68] Sergio Ramírez-Gallego, Bartosz Krawczyk, Salvador García, Michał Woźniak, and Francisco Herrera. 2017. A survey on data preprocessing for data stream mining: Current status and future directions. *Neurocomputing* 239 (2017), 39–57.
[69] Atefeh Farzindar and Wael Khreich. 2015. A survey of techniques for event detection in Twitter. *Comput. Intell.* 31, 1 (2015), 132–164.
[70] Pan Sinno Jialin and Qiang Yang. 2009. A survey on transfer learning. *IEEE Trans. Knowl. Data Eng.* 22, 10 (2009), 1345–1359.
[71] Karl Weiss, Taghi M. Khoshgoftaar, and Ding Ding Wang. 2016. A survey of transfer learning. *J. Big Data* 3, 1 (2016), 9.
[72] J. A. Gama, I. Žliobaite, A. Bifet, M. Pechenizkiy, and A. Bouchachia. 2014. A survey on concept drift adaptation. *ACM Comput. Surv*. 46, 4 (2014), 44 1–37.
[73] Sun Yu, Ke Tang, Zexuan Zhu, and Xin Yao. 2018. Concept drift adaptation by exploiting historical knowledge. *IEEE Trans. Neural Netw. Learn. Syst.* 29, 10 (2018), 4822–4832.
[74] Geoffrey I. Webb, Loong Kuan Lee, Bart Goethals, and François Petitjean. 2018. Analyzing concept drift and shift from sample data. *Data Mining Knowl. Disc.* 32, 5 (2018), 1179–1199.
[75] Avidan Shai. 2007. Ensemble tracking. *IEEE Trans. Pattern Anal. Mach. Intell.* 29, 2 (2007).
[76] Helmut Grabner, Michael Grabner, and Horst Bischof. 2006. Real-time tracking via on-line boosting. In *Proceedings of the British Machine Vision Conference* 1, 5 (2006), 6.
[77] Mahmud Hasan, Mehmet A. Orgun, and Rolf Schwitter. 2019. Real-time event detection from the Twitter data stream using the Twitternews+ framework. *Inf. Proc. Manag.* 56, 3 (2019), 1146–1165.
[78] M. Hasan, M. A. Orgun, and R. Schwitter. 2017. A survey on real-time event detection from the Twitter data stream. *J. Inf. Sci.* 44, 4 (2017), 443–463. DOI:http://dx.doi.org/10.1177/0165551517698564 0165551517698564
[79] Chao Zhang, Dongming Lei, Quan Yuan, Honglei Zhuang, Lance Kaplan, Shaowen Wang, and Jiawei Han. 2018. Geoburst+: Effective and real-time local event detection in geo-tagged tweet streams. *ACM Trans. Intell. Syst. Technol.* 9, 3 (2018), 34.
[80] Zdenek Kalal, Krystian Mikolajczyk, and Jiri Matas. 2012. Tracking-learning-detection. *IEEE Trans. Pattern Anal. Mach. Intell.* 34, 7 (2012), 1409–1422.
[81] Qinxun Bai, Zheng Wu, Stan Sclaroff, Margrit Betke, and Camille Monnier. 2013. Randomized ensemble tracking. In *Proceedings of the IEEE International Conference on Computer Vision*. 2040–2047.
[82] Bartosz Krawczyk, Leandro L. Minku, João Gama, Jerzy Stefanowski, Michał Woźniak. 2017. Ensemble learning for data stream analysis: A survey. *Inf. Fusion* 37 (2017), 132–156, Elsevier.
[83] Cha Zhang and Yunqian Ma (eds.). 2012. *Ensemble Machine Learning: Methods and Applications*. Springer Science & Business Media.
[84] K-means clustering. [<https://en.wikipedia.org/wiki/K-means_clustering>].
[85] Burr Settles. 2009. *Active Learning Literature Survey*. Technical report. University of Wisconsin-Madison Department of Computer Sciences.
[86] Panagiotis G. Ipeirotis and Evgeniy Gabrilovich. 2014. Quizz: Targeted crowdsourcing with a billion (potential) users. In *Proceedings of the 23rd International Conference on World Wide Web*. 143–154.
[87] Audun Josang, Roslan Ismail, and Colin A. Boyd. 2007. A survey of trust and reputation systems for online service provisioning. *Dec. Supp. Syst*. 43, 2 (Mar. 2007), 618–644. Elsevier.
[88] E. Lex, C. Seifert, M. Granitzer, and A. Junger. 2010. Efficient cross-domain classification of weblogs. *Int. J. Intell. Comput. Res.* 1, 1 (2010), 36–45.
[89] S. J. Pan, X. Ni, J.-T. Sun, Q. Yang, and Z. Chen. Cross-domain sentiment classification via spectral feature alignment. In *Proceedings of the 19th International Conference on World Wide Web*,. 751–760.
[90] Y. Zhen and C. Li. 2008. Cross-domain knowledge transfer using semi-supervised classification. In *AI 2008: Advances in Artificial Intelligence, Vol. 5360, Lecture Notes in Computer Science*, 362–371. Springer Berlin.
[91] Richard S. Sutton and Andrew G. Barto. 2018. *Reinforcement Learning: An Introduction*. MIT press.
[92] Fei-Yue Wang, Jun Jason Zhang, Xinhu Zheng, Xiao Wang, Yong Yuan, Xiaoxiao Dai, Jie Zhang, and Liuqing Yang. 2016. Where does AlphaGo go: From church-turing thesis to AlphaGo thesis and beyond. *IEEE/CAA J. Autom. Sin.* 3, 2 (2016), 113–120.
[93] Hutter Frank, Lars Kotthoff, and Joaquin Vanschoren. 2019. Automated machine learning-methods, systems, challenges. *Autom. Mach. Learn.* Springer, New York, NY, USA.
[94] ImageNet data set. Retrieved on November 9, 2019 from http://www.image-net.org/.

Received August 2019; revised November 2019; accepted November 2019