Refactoring Earthquake-Tsunami Causality and Messaging via Big Data Analytics: The Transformative Potential of Credible Tweets

L. Ian Lumb
Science, York University & Univa Corporation
Toronto & Markham, Ontario, Canada
Email: ianlumb@yorku.ca, ilumb@univa.com

James R. Freemantle
225 Lloyd Ave
Newmarket, Ontario L3Y 5L4, Caanda
Email: james.freemantle@rogers.com

***Abstract*—By including credible data extracted from the Twitter social networking service, the study of earthquakes and tsunamis is legitimately transformed into a Big Data Analytics problem. The challenge of establishing geophysically credible tweets is considered first through a combination of graph analytics and knowledge representation, and subsequently via Machine Learning. Although there remains cause for optimism in augmenting scientific data with that derived from social networking, further research is required to provide utility in practice. The motivation for success remains strong, as establishing a causal relationship between earthquakes and tsunamis remains problematical, and this in turn complicates any ability to deliver timely messaging that could prove life-critical.**

I. Motivation

The 26 December 2004 Indian Ocean earthquake produced an absolutely devastating tsunami, whereas the 28 March 2005 earthquake did not [1], despite similarities in Richter-scale magnitude *and* tectonic setting [2]. Much more recently, a tsunami advisory was issued for coastal regions of Japan's southernmost island of Kyushu on 16 April 2016, following a sizable earthquake [3]; 46 minutes later, and following escalating interim updates, the advisory was (surprisingly) lifted as the threat of a tsunami was no longer evident [4]. These events illustrate that the cause-effect relationship between earthquakes and tsunamis *remains* uncertain [5].[1] Thus the primary purpose here is to refactor earthquake-tsunami causality as both a challenge and opportunity for mission-critical Big Data Analytics - a refactoring that ultimately results in more timely messaging (e.g., [8]).

After briefly reviewing the scientific context for earthquakes and tsunamis in §II, the transformative inclusion of data available from social media results in recasting earthquake-tsunami causality as a problem in Big Data Analytics in §III and §IV; conclusions follow in §V.

[1]Assuming the presence of an appropriately distributed array of deep-ocean tsunami detection buoys and a forecasting model, however, far-field estimates of tsunami propagation (pre-computed) and coastal inundation (real-time) have proven to be extremely accurate (e.g., [6]). Dense, *in situ* tsunameter networks also appear promising in tsunami forecasting through the real-time assimilation of their data into tsunami wavefield models, without a need for detailed earthquake source parameters [7].

II. The Science and its Traditional Data

With bias towards whole-mantle convection [9], the context provided for earthquakes by geophysical fluid dynamics can be summarized as follows: The lithosphere (i.e., Earth's crust) comprises the upper thermal boundary layer of a convective circulation that spans the full, ∼2900 km depth of Earth's mantle. The natural build up of stress in Earth's mantle is such that releases in the form of earthquakes occur primarily at one of the three types of junctions between tectonic plates and (less frequently) at epicentral locations interior to plates (e.g., [10]). Because thermal convection at conditions appropriate for Earth's mantle is nondeterministic, uncertainty is inherent in any effort that aims to predict the hypocentric location of an earthquake in space and time.[2]

Earth-based (e.g., gravimeters, seismometers, tidal gauges, tsunameters) and remotely sensing (e.g., satellite altimeters and GPS) scientific instruments comprise the primary sources of *objective* data on earthquakes and tsunamis. While it is certain that there is earthquake-tsunami data available from numerous objective sources for corroborative analyses (e.g., joint inversions to elucidate earthquake parameters), the need for Big Data Analytics is hardly pressing - in other words, Big Data's original three [13] to six [14] Vs are barely justifiable. In the following section (§III), however, the inclusion of data from social media completely disrupts this perspective and does justify re-framing as a Big Data problem.

III. Credible Data from Twitter?

As Figure 1 indicates, a Perl script (e.g., Listing 1) can readily extract geophysically relevant data (e.g., earthquake events in Chile and Italy) via keyword matching [5]. The script, however, also extracts instances of the keyword "earthquake" that do not have any relevance in the geophysical context of

[2]Of course, in extremely well instrumented locales (e.g., parts of the San Andreas Fault), measures of accumulated stress can reduce this uncertainty somewhat. In the current tsunami-centric context for considering earthquakes, however, oceanic-based tectonic settings impede the ongoing, in-situ monitoring of stress-related measures. Remotely sensed crustal dynamics data, that makes use of GPS technologies for example (e.g., [11], [12]), also has the potential to monitor stress build up on an ongoing basis (even in oceanic settings).

```
Created at: Wed Jun 04 20:29:33 +0000 2014
5.0 earthquake! Thu Jun 05 02:04:27 GMT+09:00 2014 near 84km SW of Iquique, Chile
  http://t.co/mmFokGQWT7 #earthquake
Created at: Wed Jun 04 20:30:13 +0000 2014
The #earthquake continues: Latest via @Spectator_CH /@YouGov -#Labour 36 #Tories 32%, LD 8%,
  #Ukip 14%.  Implied Labour majority- 42 .
Created at: Wed Jun 04 20:31:35 +0000 2014
#terremoto ML 2.7 CENTRAL ITALY: Magnitude ML 2.7 Region CENTRAL ITALY Date time 2014-06-04
  20:01:33.9 UTC... http://t.co/Y141Ovu6kP
Created at: Tue Jun 10 12:22:34 +0000 2014
RT @TheRock: Just wrapped a massive post earthquake scene for SAN ANDREAS. To the hundreds of
  background actors/extras.. THANK U for all yo...
```

Fig. 1. Geophysically relevant and irrelevant tweets example after [5]. (Note that indents implied continued lines.)

earthquakes and tsunamis - e.g., the political and entertainment references indicated in Figure 1.

At this point, it is instructive to conduct a '6V Test' on the tweets captured in Figure 1- i.e., assess the findings thus far against the litmus test for Big Data's 6Vs [14]:

- **Volume** Although the Figure 1 tweets span only a few days, it is reasonable to expect that data volumes *could* be substantial. For example, a significant earthquake could generate data (almost) globally, as its impact is 'felt'. Significant earthquake events are typically accompanied by aftershocks that could continue to generate tweets for days or longer.
- **Variety** Other than the `#earthquake` hashtag used by the Perl script (Listing 1), the data retrieved is various and unstructured - in direct contrast to the constrained, semi-structured scientific data [15] typical of the instruments alluded to previously (§II). In addition to unstructured text, through use of URLs, tweets encompass additional variety through the inclusion of links; again, some of these may be geophysically relevant - e.g., a photograph may directly convey the draw-down of water on a beach that precedes a tsunami [16], or through damage an indirect indication of earthquake intensity.
- **Velocity** Significant events have the potential to 'go viral' in the Twittersphere. Clearly, in such cases, the rate at which data is generated could be of considerable velocity. Assuming it can be suitably extracted via calls to the Twitter API, Twitter's trending data could be particularly useful in quantifying the velocity of earthquake-related events.
- **Veracity** According to [14]: "Big Data Veracity refers to the biases, noise and abnormality in data." As the sample presented in Figure 1 resoundingly demonstrates, this is a pressing concern for any effort that seeks to 'make sense' of data extracted from Twitter. From a geophysical perspective, politically or entertainment-charged tweets are rightly noise, whereas use of colloquialisms like "quake" point to a much more subtle need for disambiguation.
- **Validity** As the name implies, validity is concerned with matters such as accuracy and correctness. It is fair to state that Lumb & Freemantle [5] placed most of their emphasis on this aspect of the social networking data typified by Twitter. As Gupta et al. [17] demonstrated, 'faked photographs' of natural disasters, are also within the scope of validity.
- **Volatility** As social-networking services such as Snapchat [18] (literally) illustrate, data can have a (very) finite lifetime. Of course, the (inherent) volatility of social-networking data is orthogonal to that of traditional scientific data where ensuring preservation is a priority (e.g., [19]). Accessed in the June 2014 time frame, historical streams of Twitter data dating back four years were deemed to be quite valuable by Lumb & Freemantle [5].

In summarizing the '6V Test' then, Twitter serves as a *tipping point* in the present effort to ultimately elucidate earthquake-tsunami causality, as it transforms an existing scientific data problem into a Big Data problem.[3]

```
use Net::Twitter::Lite::WithAPIv1_1;
my $nt = Net::Twitter::Lite::WithAPIv1_1->new(
  consumer_key        => 'xxx...xxx',
  consumer_secret     => 'xxx...xxx',
  access_token        => 'xxx...xxx',
  access_token_secret => 'xxx...xxx',
  ssl => 1
);
my $result = $nt->search("earthquake");
for my $status(@{$result->{statuses}} ) {
  print "$status->{text}\n";
}
```

Listing 1. A Perl script protoype that extracts keyword matches to the string "earthquake" from streamed Twitter data via an API call [5]

Lumb & Freemantle [5] indicated that iterative processing of data extracted from Twitter via Giraph [22] would lend *contextual* credibility by 'truthing' tweets in time, location and intensity; thus complementing the 'inherent' credibility derived through of TweetCred [23]. Scoring, in this sense, is depicted by the feedback loop in Figure 2. Ontologies can be leveraged, for example, to disambiguate confusion amongst terms - in other words, to express the semantic equivalence between "earthquake" and "quake". With additional semantic context, ontologies are key to enabling the distinction between geophysically inclined use of "earthquake" from other uses

[3]And as if the data deluge offered up by Twitter is not convincing enough, there is considerably more objective data rapidly becoming available through 'consumer seismometers' such as MyShake - smart-phone software that makes use of built-in accelerometers to detect and distinguish signals of potential interest in real time [21].

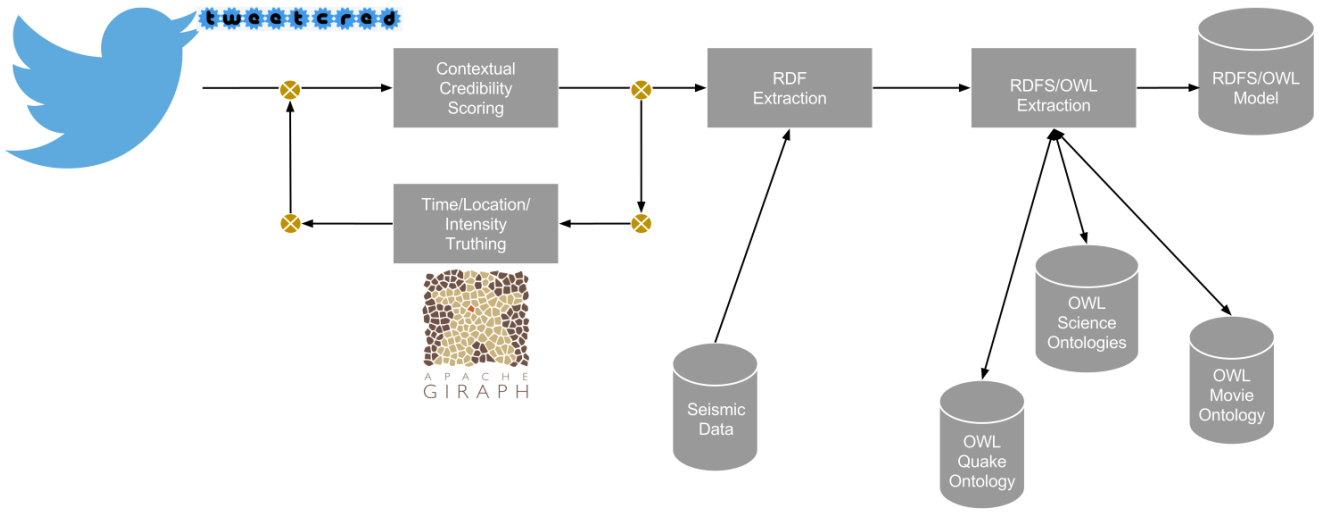

Fig. 2. Iteratively enhanced via graph analytics involving Giraph, 'more credible' data extracted from Twitter populates a knowledge-representation framework developed elsewhere ([15], [20]). Adapted from [5].

(e.g., political, or entertainment involving movies and gaming). Thus Lumb & Freemantle [5] also sought to leverage a previously developed framework for knowledge representation [15] that would allow for the ingestion of discipline-specific ontologies (e.g., from science alongside gaming and movies) into a knowledge-representation model ultimately expressed using Resource Description Framework Schema (RDFS) [24] and/or OWL [25]. Credible data extracted from Twitter, as well as scientific data (e.g., seismic data), would also be used to populate this model. Because the model preserves relationships represented as triples via the Resource Description Framework (RDF) [26], the ultimate objective here is enhanced earthquake-tsunami knowledge.

Although a knowledge-representation model for credible tweets remains of interest in the current context of earthquakes and tsunamis, the following section (§IV) introduces consideration of another analytics approach that also appears to be promising.

IV. Deep Learning from Social Media?

From the successful detection of geological faults in seismic data [27], to support for full waveform inversion [28], to cost-sensitive improvements in the productivity of unconventional petroleum reservoirs [29], Machine Learning (e.g., [30]) is already delivering impressive results in exploration geophysics. And these outcomes appear to reflect a groundswell of interest in Machine Learning (e.g., [31]) that cuts across disciplinary boundaries, but especially emphasizes 'Deep Learning' - "a modern refinement of 'machine learning', in which computers teach themselves tasks by crunching large sets of data" [32].

Of course, the idea of applying Machine Learning to data extracted from social-media sources (such as Twitter) is not a new one. And although even a hastily typed seach-engine query will rapidly validate this point, it becomes quite clear that considerable effort has been devoted to sentiment analysis in the case of data extracted from Twitter (e.g., [33]).

Although there *might* ultimately be a benefit associated with taking sentiments into account, attention was focused on making use of support for Machine Learning in Apache Spark ([34], [35]) to implement the well-established task of classifying text (e.g., Example 11-1 of [36]) as follows:

1) **Represent data** Data extracted from Twitter (along the lines of Figure 1) was manually curated into 'ham' (i.e., geophysically relevant tweets) and 'spam' (i.e., all other tweets) for training purposes. The curated collections of 140-character ham and spam tweets were represented as in-memory strings through use of Spark's implementation of Resilient Distributed Datasets (RDDs) [37].
2) **Extract features** Using spaces as the delimiter, each 'word' of each tweet was represented numerically as a 'feature' for the purpose of Machine Learning. Thus, RDDs of tweets were converted into vectors. By emphasizing the frequency with which a word was used, MLlib's `HashingTF` populates feature vectors.
3) **Develop model object** A model object was developed next through application of a classification algorithm to the vectorized features. Logistic regression via MLlib's `LogisticRegressionWithSGD` (i.e., Stochastic Gradient Descent, SGD) was used here for the purpose of the classification algorithm.
4) **Evaluate model** Through use of a minimal set of strictly curated tweets (i.e., only five ham and nine spam) from 2014, the model object was used to correctly evaluate the 2016 tweet `2.1 magnitude #earthquake. 89km E of Bear Creek, Alaska http://earthquaketrack.com/quakes/-2016-05-03-01-54-15-utc-2-1-9 ...` as ham using a prediction capability built into MLlib.

Comprehensive experiments with classifying `#earthquake` tweets via Machine Learning are currently underway, and results will be reported elsewhere.

V. Conclusions

In the case of tsunamis, timely messaging is of *indisputable* value [8]. Unfortunately, the ability to fully realize this service operationally is hampered in theory (i.e., scientifically established earthquake-tsunami causality, §II) and in practice (i.e., incomplete and uneven observations, §II). Subjective (e.g., Twitter) and objective (e.g., MyShake) data, from social-media sources and citizen-science efforts, respectively, can augment traditional datasets gathered using a wide variety of scientific instruments (§III). By systematically improving the credibility of data extracted from Twitter, for example, additional observations are rendered available (§III); beyond Twitter, other prospects (e.g., Instagram [38]) exist for extracting data and placing it through a credibility enhancing process.[4] Although past efforts employed a combination of graph analytics and semantics, Machine Learning clearly requires serious consideration (§IV). By ignoring the semantically rich content of tweets (e.g., Twitter metadata such as IDs, hastags and URLs) and offering no means for disambiguating semantically equivalent terms (e.g., "earthquake" and "quake"), however, it appears useful to employ Machine Learning in tandem with knowledge representation via RDF and OWL. If ultimately proven successful, approaches like this will certainly serve as examples of mission-critical Big Data Analytics.

References

[1] T. Lay *et al.*, "The Great Sumatra-Andaman Earthquake of 26 December 2004," *Science*, vol. 308, no. 5725, pp. 1127–1133, 2005.

[2] E. L. Geist, V. V. Titov, and C. E. Synolakis, "Tsunami: Wave of Change," *Scientific American*, vol. 294, no. 1, pp. 56–63, 2006.

[3] Japan Meteorological Agency. Tsunami Advisory, Issued at 01:27 JST, 16 Apr. 2016. [Online]. Available: http://www.jma.go.jp/en/tsunami/-focus_03_20160416012755.html

[4] ——. Tsunami Advisories Lifted, Issued at 02:14 JST, 16 Apr. 2016. [Online]. Available: http://www.jma.go.jp/en/tsunami/-focus_03_20160416021401.html

[5] L. I. Lumb and J. R. Freemantle, "Towards Earthquake-Tsunami Causality via Data Science: Giraph-Derived Credibility Scores for Data from Twitter," presented at the High Performance Computing Symposium (HPCS), 2014. [Online]. Available: http://2014.hpcs.ca/-program/#posters

[6] Y. Wei *et al.*, "Real-time experimental forecast of the Peruvian tsunami of August 2007 for U.S. coastlines," *Geophysical Research Letters*, vol. 35, no. 4, pp. n/a–n/a, 2008, l04609.

[7] T. Maeda *et al.*, "Successive estimation of a tsunami wavefield without earthquake source data: A data assimilation approach toward real-time tsunami forecasting," *Geophysical Research Letters*, vol. 42, no. 19, pp. 7923–7932, 2015.

[8] NOAA National Tsunami Warning Center - Message Definitions. [Online]. Available: http://wcatwc.arh.noaa.gov/?page=message_definitions

[9] W. Peltier, "Mantle convection and viscosity," *Physics of the Earth's Interior*, vol. 361, p. 431, 1980.

[10] F. D. Stacey, P. M. Davis *et al.*, *Physics of the Earth*. Wiley New York, 1977.

[11] J. Murray-Moraleda, "GPS: applications in crustal deformation monitoring," in *Extreme Environmental Events*. Springer, 2011, pp. 589–622.

[12] D. Melgar *et al.*, "Local tsunami warnings: Perspectives from recent large events," *Geophysical Research Letters*, 2016.

[13] E. Dumbill. (2012) Volume, Velocity, Variety: What You Need to Know About Big Data. [Online]. Available: http://www.forbes.com/sites/oreillymedia/2012/01/19/volume-velocity-variety-what-you-need-to-know-about-big-data/print/

[14] I. Bhandar. (2013) Beyond Volume, Variety and Velocity is the Issue of Big Data Veracity. [Online]. Available: http://insidebigdata.com/2013/09/12/beyond-volume-variety-velocity-issue-big-data-veracity/

[15] L. I. Lumb and K. D. Aldridge, "Grid-enabling the Global Geodynamics Project: automatic RDF extraction from the ESML data description and representation via GRDDL," in *High-Performance Computing in an Advanced Collaborative Environment, 2006. HPCS 2006. 20th International Symposium on*. IEEE, 2006, pp. 29–29.

[16] C. Klettner *et al.*, "Draw-down and run-up of tsunami waves on sloping beaches," *Proceedings of the Institution of Civil Engineers-Engineering and Computational Mechanics*, vol. 165, no. 2, pp. 119–129, 2012.

[17] A. Gupta *et al.*, "Faking Sandy: Characterizing and identifying fake images on Twitter during hurricane Sandy," in *Proceedings of the 22nd international conference on World Wide Web companion*. International World Wide Web Conferences Steering Committee, 2013, pp. 729–736.

[18] Snapchat. [Online]. Available: https://www.snapchat.com

[19] GGP Home Page. [Online]. Available: http://www.eas.slu.edu/GGP/-ggphome.html

[20] L. I. Lumb *et al.*, "Annotation modeling with formal ontologies: Implications for informal ontologies," *Computers & Geosciences*, vol. 35, no. 4, pp. 855–861, 2009.

[21] Q. Kong *et al.*, "MyShake: A smartphone seismic network for earthquake early warning and beyond," *Science Advances*, vol. 2, no. 2, 2016. [Online]. Available: http://advances.sciencemag.org/content/2/2/e1501055

[22] Apache Giraph. [Online]. Available: http://giraph.apache.org/

[23] A. Gupta and P. Kumaraguru, "Credibility ranking of tweets during high impact events," in *Proceedings of the 1st Workshop on Privacy and Security in Online Social Media*. ACM, 2012, p. 2.

[24] RDF Schema 1.1. [Online]. Available: https://www.w3.org/TR/rdf-schema/

[25] OWL. [Online]. Available: www.w3.org/2004/OWL/

[26] RDF. [Online]. Available: https://www.w3.org/RDF/

[27] C. Chen, T. Clee, L. Huang, and Y. Yan, "Applying Big Data Analytics to Seismic Interpretation," presented at the Rice University Oil & Gas HPC Conference, 2016. [Online]. Available: http://sched.co/5sVA

[28] G. Wenes, "Machine Learning Support for Full Waveform Inversion," presented at the Rice University Oil & Gas HPC Conference, 2016. [Online]. Available: http://sched.co/5sV8

[29] D. Hohl, L. Lu, X. Wang, and M. Wu, "SweetSpot Identification Using Machine Learning for Unconventionals," presented at the Rice University Oil & Gas HPC Conference, 2016. [Online]. Available: http://sched.co/5sVA

[30] T. M. Mitchell *et al.*, "Machine learning. WCB," 1997.

[31] D. Micci-Barreca. (2016) "Big Data" reaches plateau while interest in Machine Learning grows. [Online]. Available: http://eliteanalytics.com/big-data-reaches-plateau-while-interest-in-machine-learning-grows/

[32] T. Economist. (2015) Rise of the machines. [Online]. Available: http://www.economist.com/news/briefing/21650526-artificial-intelligence-scares-peopleexcessively-so-rise-machines

[33] M. S. Neethu and R. Rajasree, "Sentiment analysis in Twitter using machine learning techniques," in *Computing, Communications and Networking Technologies (ICCCNT),2013 Fourth International Conference on*, July 2013, pp. 1–5.

[34] X. Meng *et al.*, "MLlib: Machine Learning in Apache Spark," *arXiv preprint arXiv:1505.06807*, 2015.

[35] Apache Spark MLlib. [Online]. Available: http://spark.apache.org/mllib/

[36] H. Karau, A. Konwinski, P. Wendell, and M. Zaharia, *Learning Spark: Lightning-Fast Big Data Analysis*. " O'Reilly Media, Inc.", 2015.

[37] M. Zaharia *et al.*, "Resilient distributed datasets: A fault-tolerant abstraction for in-memory cluster computing," in *Proceedings of the 9th USENIX conference on Networked Systems Design and Implementation*. USENIX Association, 2012, pp. 2–2.

[38] Instagram. [Online]. Available: https://www.instagram.com/

[4] As experience here with Twitter makes clear, the availability of a well-documented API is absolutely crucial in enabling progress.