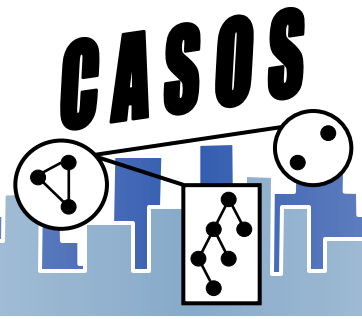


An exploration into user-defined keyword sampling from Twitter during disasters

Kenneth Joseph | Kathleen M. Carley

kjoseph | kathleen.carley @cs.cmu.edu

isr institute for SOFTWARE RESEARCH

Carnegie Mellon

Center for Computational Analysis of Social and Organizational Systems
http://www.casos.cs.cmu.edu/

The "search" problem

- Less than .5% of Tweets sampled in a recent paper were found to be "actionable" during a disaster
  - A needle in a haystack (Munro, 2011)
- To make matters worse:
  - You don't get all the Tweets you're interested in
    - The needle is in the haystack, but you only get to search through a few bales
  - What you're interested in is constantly changing
    - The needle has legs
- Human volunteers really shouldn't be spending their time combing through Twitter
- If you're a computer, you have to be smart about how you search

The "post-search" problem

- Many have taken the approach of pulling tweets with heavily-used search terms (e.g. Haiti) and then analyzing these as a representative sample to study the uses of Twitter during a disaster
  - E.g. Oh et al, 2010; Hughes and Palen, 2009;Medoza et al, 2010; Starbird and Palen, 2011; Munro, 2010

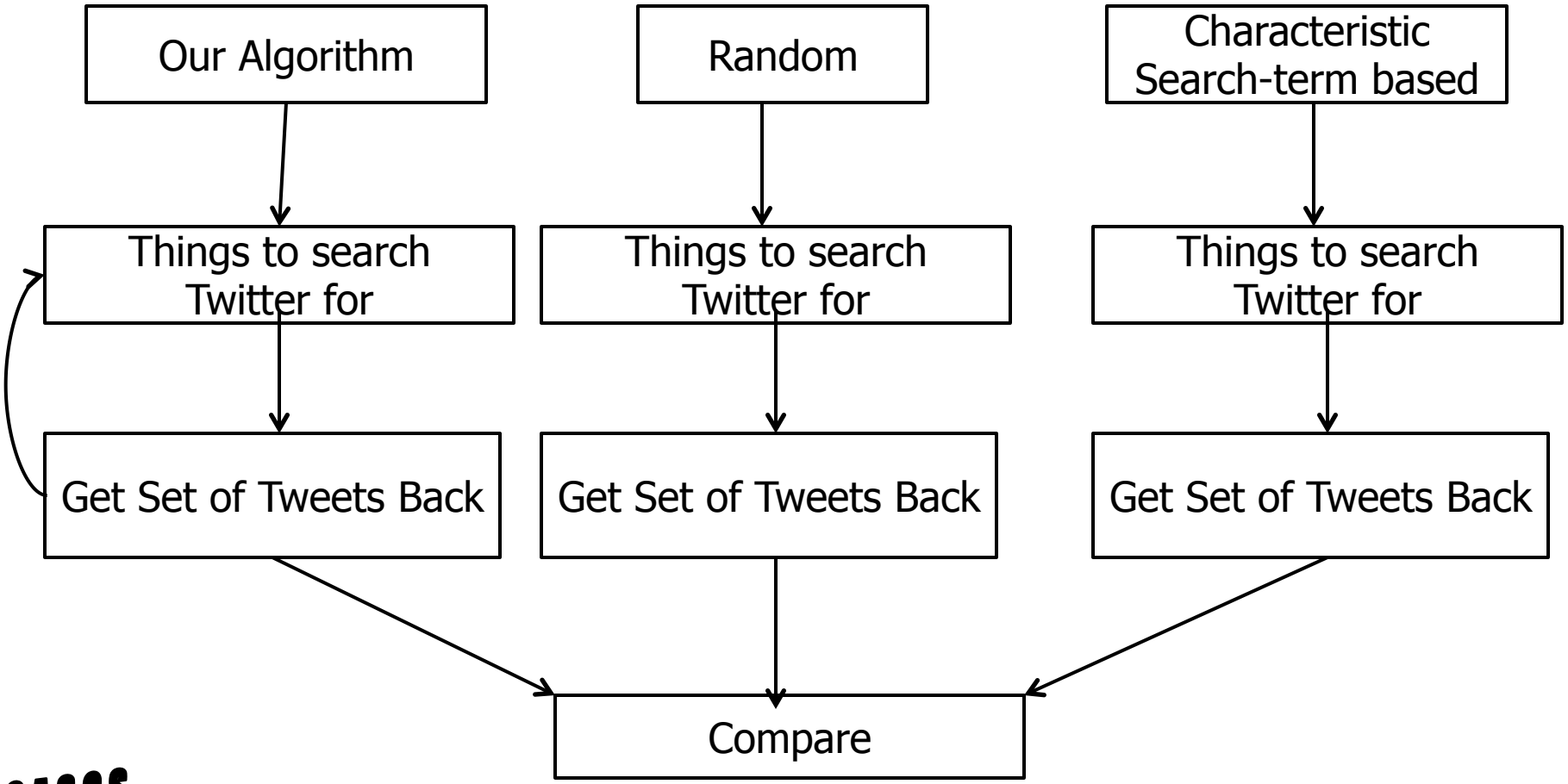
- A question which has not been asked is, is this an appropriate sampling methodology?

- Just because there is a lot of data, does that mean we can sample, more or less, any way we want?

The research questions

- **RQ1:** can we find Tweets that are "useful" for crisis workers in "real-time"?
- **RQ2:** how different are the tweets we find from those containing characteristic search terms
- **The data**
  - Gardenhose (about 15% of all Tweets at the time) from the time of through three weeks after the Haitian earthquake of 2010 - ~100M tweets
  - We simulate real-time by only allowing ourselves to look at data from previous time periods

Experiment


Our Algorithm
Random
Characteristic Search-term based
Things to search Twitter for
Things to search Twitter for
Things to search Twitter for
Get Set of Tweets Back
Get Set of Tweets Back
Get Set of Tweets Back
Compare

The Algorithm – General Idea

- Develop a method which looks for Tweets that might be "useful" to crisis workers without any work on their part

- How? Find search terms that are relevant and recent using other, more reliable data
  - Recent: our method uses reports from Ushahidi workers to create a dynamic set of search terms to account for concept drift and the differing stages of disasters
  - Relevant: It uses simple unsupervised (boosting) techniques to determine search terms which return relevant results

Ushahidi

- Ushahidi – a crowdsourced platform for mapping situational awareness during a disaster
- Ushahidi got the data from:
  - Mission 4636
  - Mainstream media outlets
  - Twitter
- What did Ushahidi do?
  - Get coordinates, plot incidents on map, and at some points, communicate with Coast Guard

What can we use?

- Categories
  - Ushahidi-volunteers set them, based on a general set of categories for disasters from Red Cross
- Location Data – e.g. “Port-au-Prince”
- The full text description
- The title
  - A more refined version of the text

Algorithm Updating Search Terms

Language Model	Search Term Set, t=0	New Incoming Ushahidi Report	Search Term Set, t=1
Title	{}	"Need Water"	{need, water}
Category	{}	"Water shortage"	{water, shortage}
Location	{}	"PaP"	{PaP}
Content	{}	"we are in need of water, please help"	{need, water, help}

The Algorithm
Search Term contextualization

{…. "Haiti", "Earthquake", "Help",…}

$$contextualization(q_i) = \frac{relevance(q_i) + temporalWeight(q_i)}{2}$$

$$relevance(q_i) = \frac{n(q_i, C_+)}{n(q_i, C_+) + \sqrt{n(q_i, C_-)}} \qquad temporalWeight(q_i) = \frac{lastTimeSeenInAReport}{currentTime}$$

The Algorithm
Tweet Score

Oh boy, will there ever be a water shortage in my tomato garden this year!

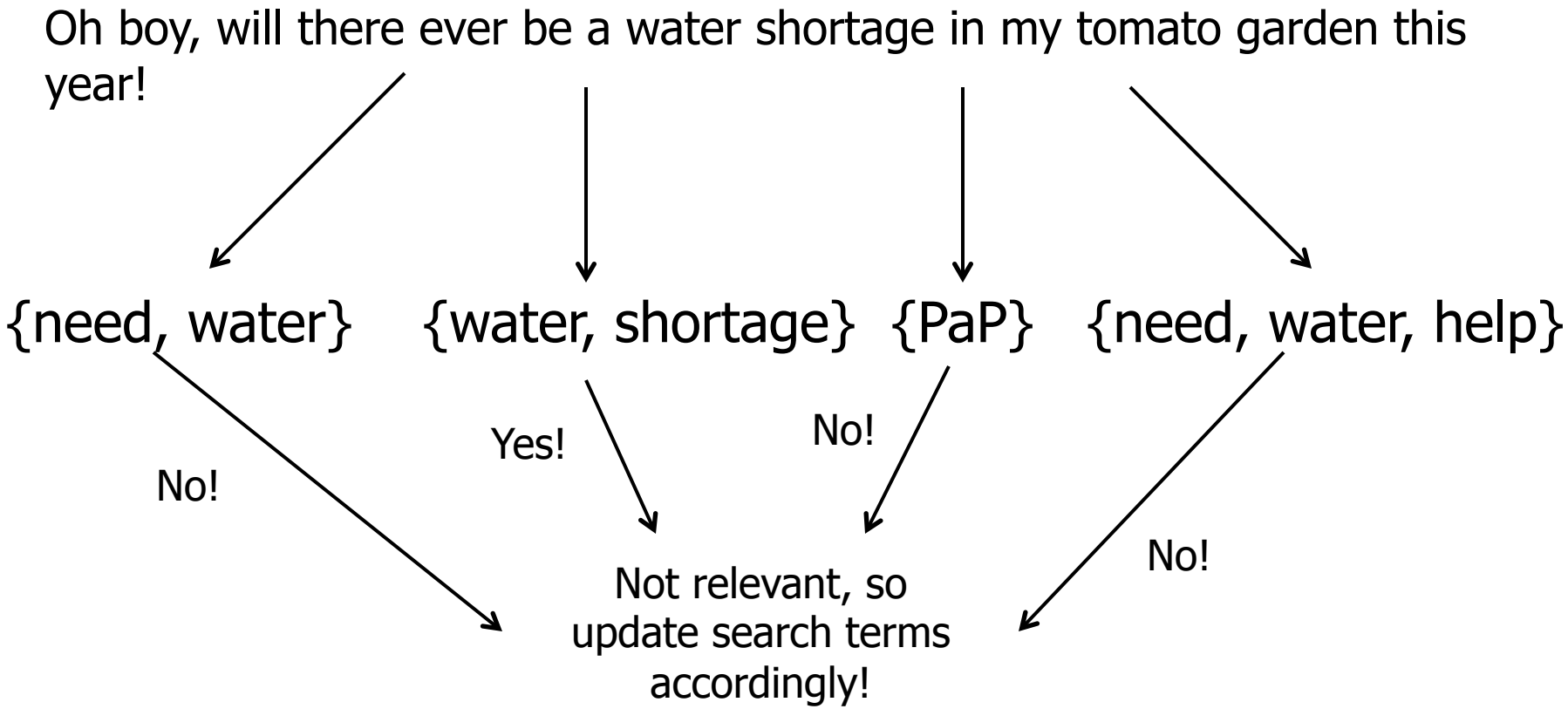
{water, shortage}

$$p(q_i|t) = \frac{n(q_i,t)}{|t|}(1-\gamma) + \gamma\frac{n(q_i,T)}{|T|}$$

$$p(q_i|t) = \frac{n(q_i,t)}{|t|}(1-\gamma) + \gamma\frac{n(q_i,T)}{|T|}$$

$$score(t) = \prod_{q_i \epsilon (Q \epsilon t)} \sqrt{contextualization(q_i) * p(t|q_i)}$$

Algorithm Combining Models



$$relevance(q_i) = \frac{n(q_i, C_+)}{n(q_i, C_+) + \sqrt{n(q_i, C_-)}}$$

Sampling Algorithm - Random

- Randomly sample ~same amount of tweets
- .0013% of entire stream
- Allows us to see what our model might be doing well on just because even the dumbest possible model does well on it (it happened…)

Carnegie Mellon
isr institute for SOFTWARE RESEARCH

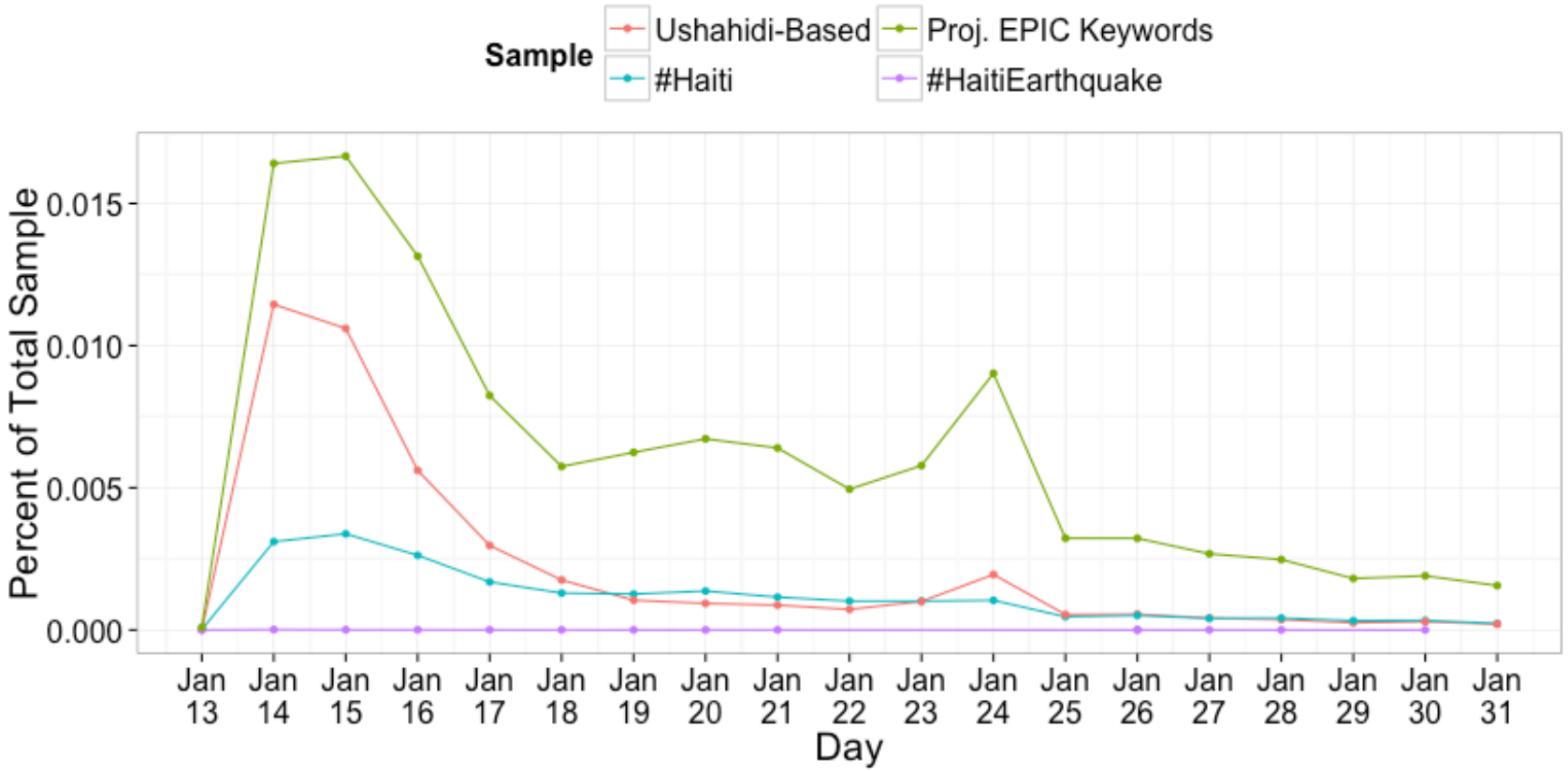
Sampling Algorithm Characteristic Search Terms

Work	Search Terms	Sample Size
Munro (2011); Munro and Manning (2012)	#haiti	40,000
Oh et al. (2010)	#haitieathquake	962
Verma et al. (2011); Vieweg (2012)	haiti, earthquake, quake, shaking, tsunami, ouest, Port-au-Prince, tremblement, tremblement de terre (but had to contain "haiti")	4M
Sarcevic et al. (2012)	earthquake, Port-au-Prince, Ouest, tsunami, haiti, and tremblement	3.28M

Proj. EPIC Keyword set

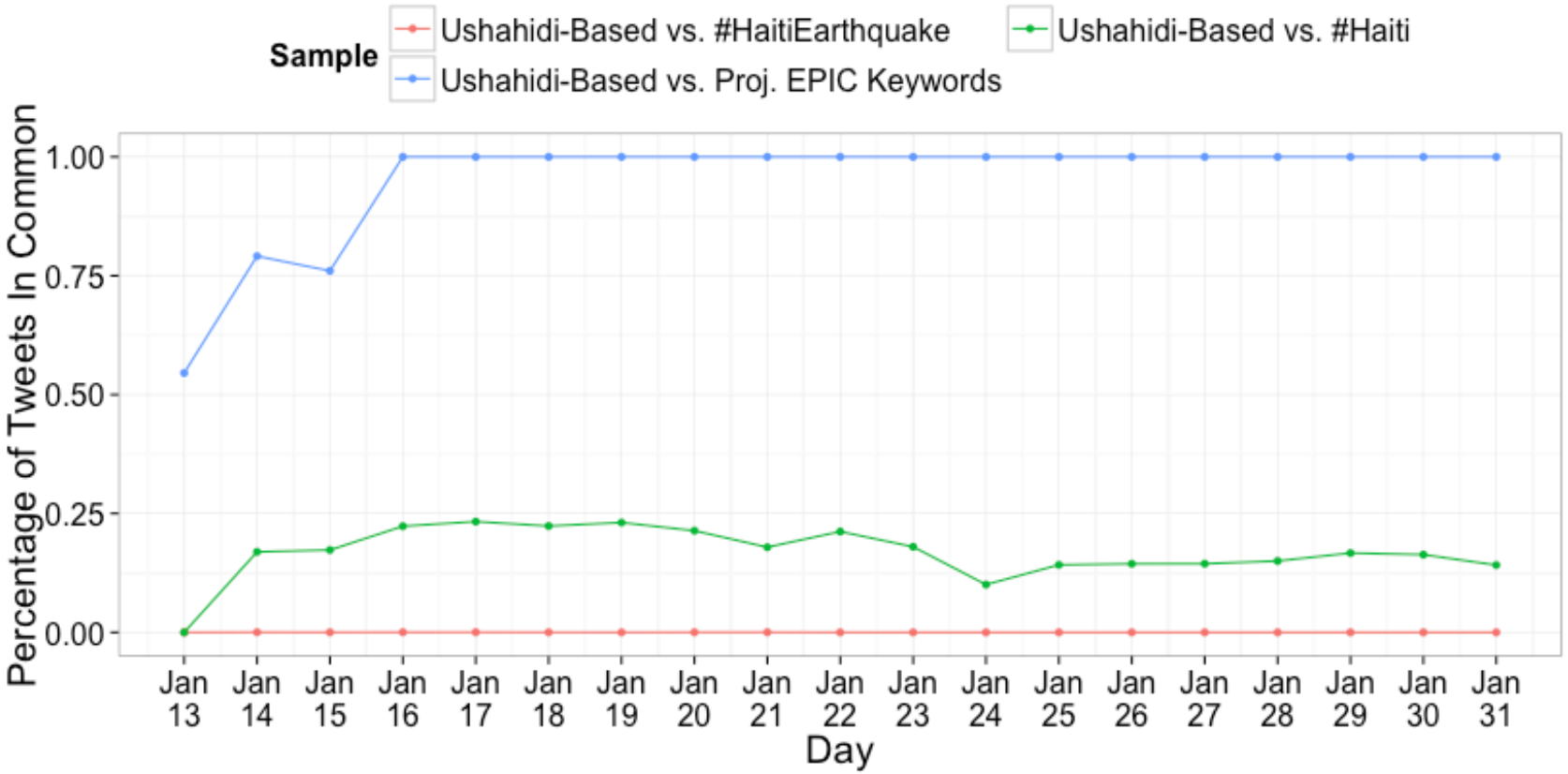
Carnegie Mellon
isr institute for SOFTWARE RESEARCH

Results
Percent of all Tweets selected



Results

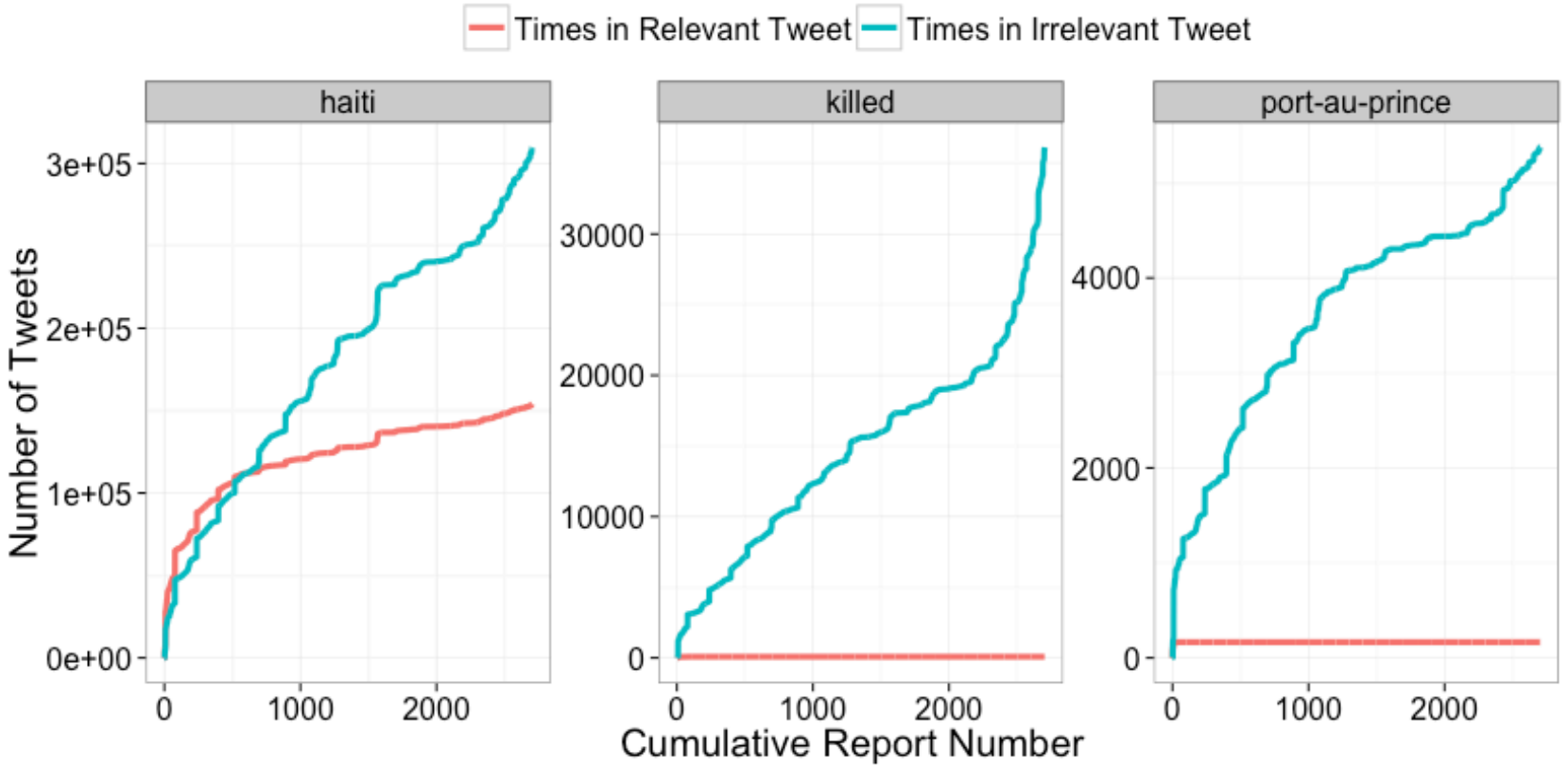
Percent of Our Model's Tweets Captured



Results
Matching to Proj. EPIC

Words	Field	Relevant	Not Relevant	% Reports	Cont.
haiti	Location	152275	302068	0.22	1
haiti	Title	152207	300428	0.02	0.98
haiti	Content	152207	300428	0.08	0.95
earthquake	Content	1318	62151	0.03	0.88
quake	Content	756	18934	0.01	0.85
port-au-prince	Location	165	5293	0.19	0.84
port-au-prince	Content	165	5293	0.07	0.83
port-au-prince	Title	165	5293	0.02	0.81
earthquake	Title	1318	62151	<3	0.80
tremblement	Content	13	225	<3	0.68
ouest	Location	0	104	<3	0.46
ouest	Title	0	5	<3	0.46
quake	Title	0	16553	<3	0.45
ouest	Content	0	69	<3	0.35
earthquake	Category	0	19816	<3	0.28
shaking	Content	5	5617	<3	0.04

Results
The good, the ???, and the ugly



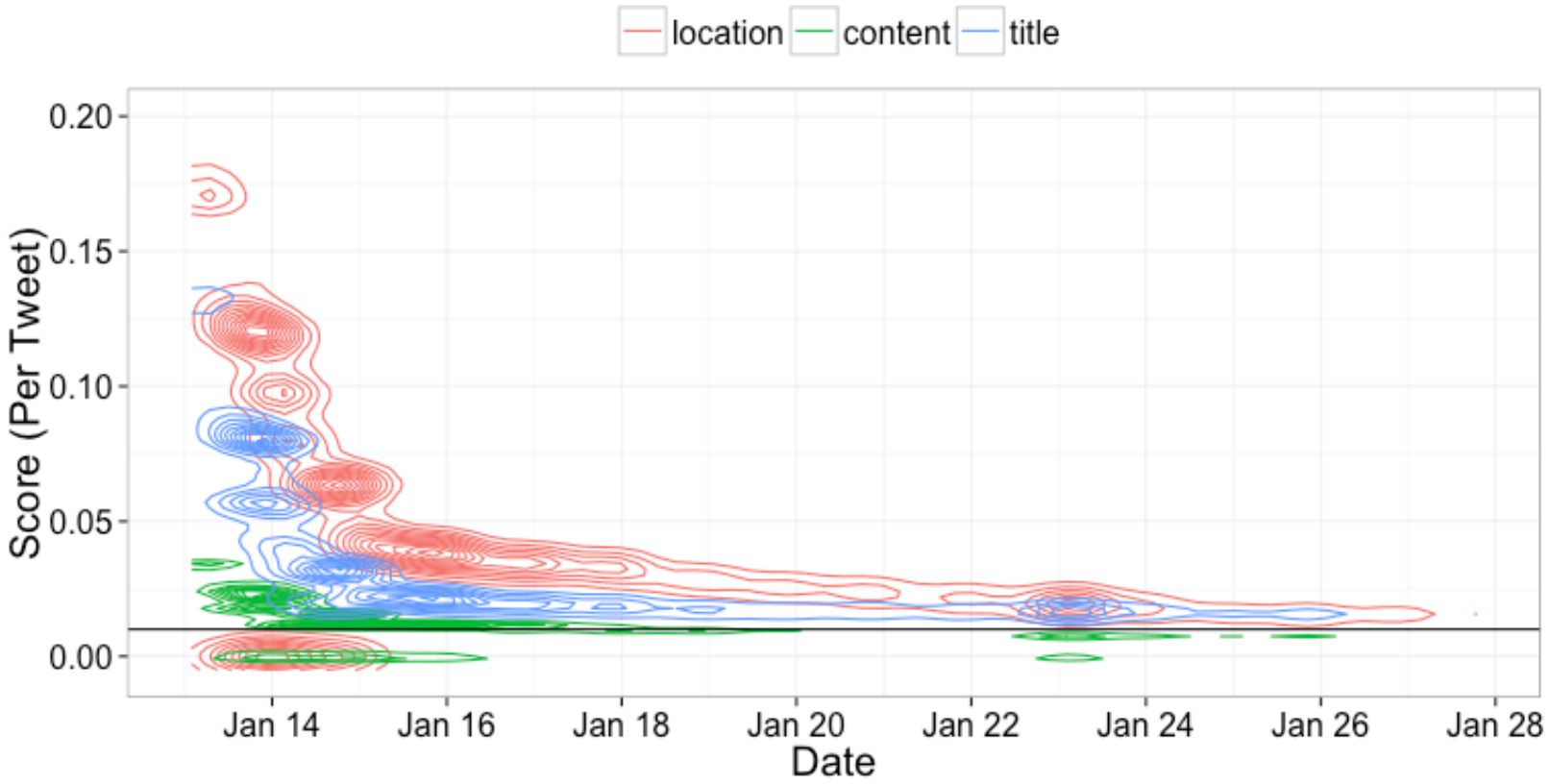
Conclusion

- Much still to be done
- Solid evidence that the model is performing well, however
- Allows Twitter to be searched in real-time, which no direct human interaction
- Validates previous work

Thanks!

- Questions?

Future Work


location
content
title
0.20
0.15
0.10
0.05
0.00
Score (Per Tweet)
Jan 14
Jan 16
Jan 18
Jan 20
Jan 22
Jan 24
Jan 26
Jan 28
Date

Results
Hashtag Comparison