SourceSeer: Forecasting Rare Disease Outbreaks Using Multiple Data Sources

Theodoros Rekatsinas
University of Maryland

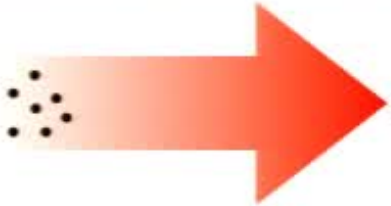
with Saurav Ghosh[1], Sumiko Mekaru[2], Elaine Nsoesie[2], John Brownstein[2], Lise Getoor[3], Naren Ramakrishnan[1]
[1]: Virginia Tech, [2]: Boston Children's Hospital , [3]: UC Santa Cruz

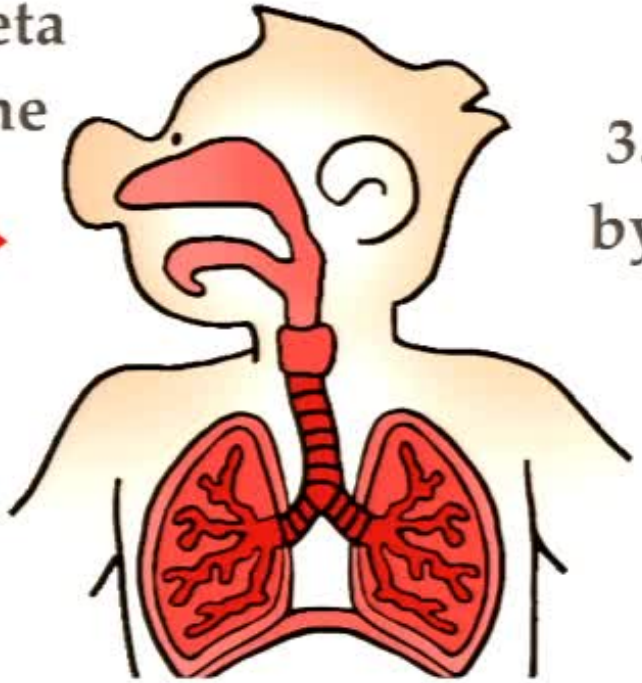
Rare diseases

Hantavirus

1. Virus contained in rodent excreta

2. Infected excreta become airborne





3. Inhaled by humans

4. Acute respiratory distress: serious infection that quickly worsens

Deluge of news feeds on diseases

Track the progression and detect the emergence of outbreaks in realtime

Results for **HANTAVIRUS**

Top / **All**

EDENextData @EDENextDMT · 10m
Katrien Tersago: Getting prepared: Climate based forecating of Puumala **hantavirus** disease risk through space and time in W. Europe

Katy Carter @katyshecooks · 2h
@CoupleCooks the "sleeping sickness!" gah! so mysterious & terrifying! (and, "is" it a **hantavirus**???)

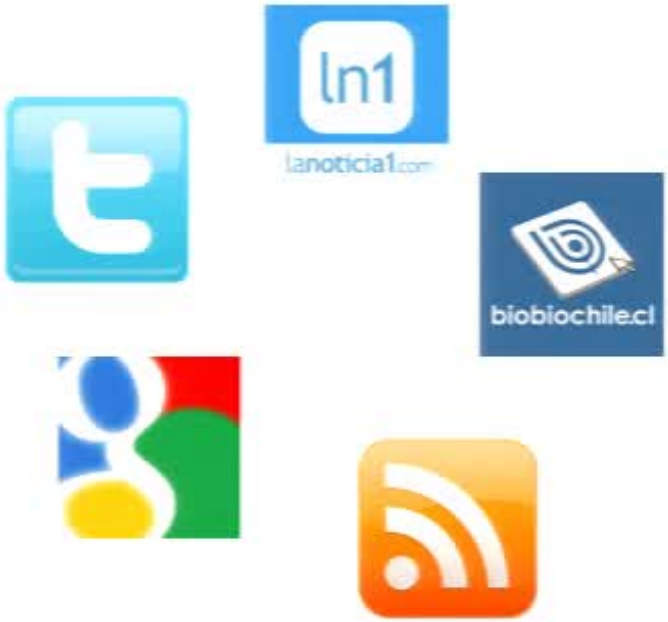
US4USA @US4USA · 3h
@ColoradoNews Colorado: Phillips County man died of **hantavirus**, not flu: A Phillips County, CO man who died th... bit.ly/...

Robert Herriman @bactman63 · 3h
#Colorado: Phillips County man died of #hantavirus, not #flu. 3rd death this year fb.me/2x7e0Bnv1

EDENextData @EDENextDMT · 7h
J-F Cosson: Forest fragmentation & metapopulation dynamics of bank voles govern persistence and spread of **hantavirus** Puumala in W Europe

jerrygenesio @jerrygenesio · 10h
Death in COLORADO confirmed as **HANTAVIRUS** – See naturalunseenhazards.wordpress.com

Analyze related keywords …

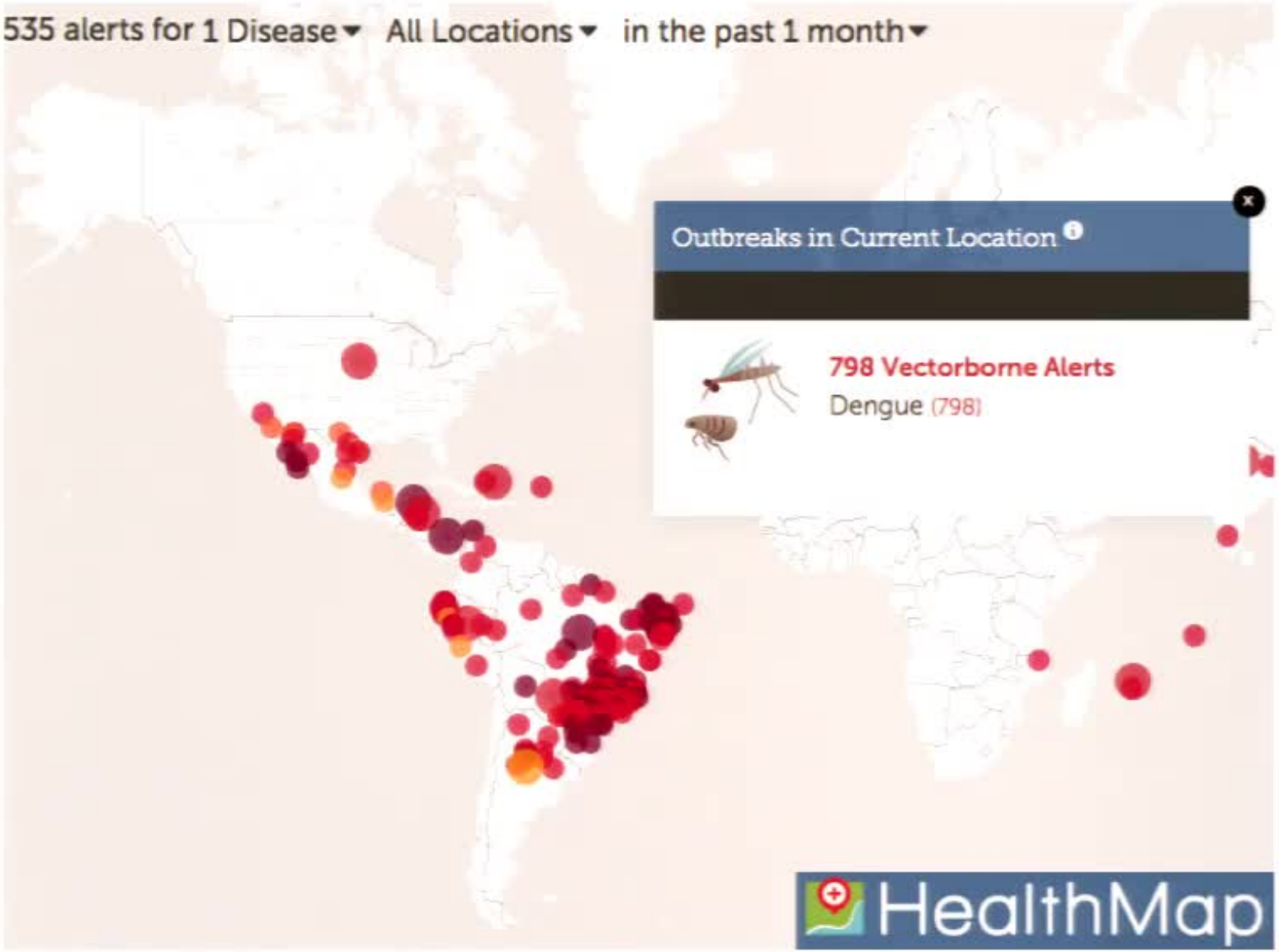


Give me everything you have!!! I need all of it to improve my accuracy!!!

… to forecast the emergence of infectious diseases

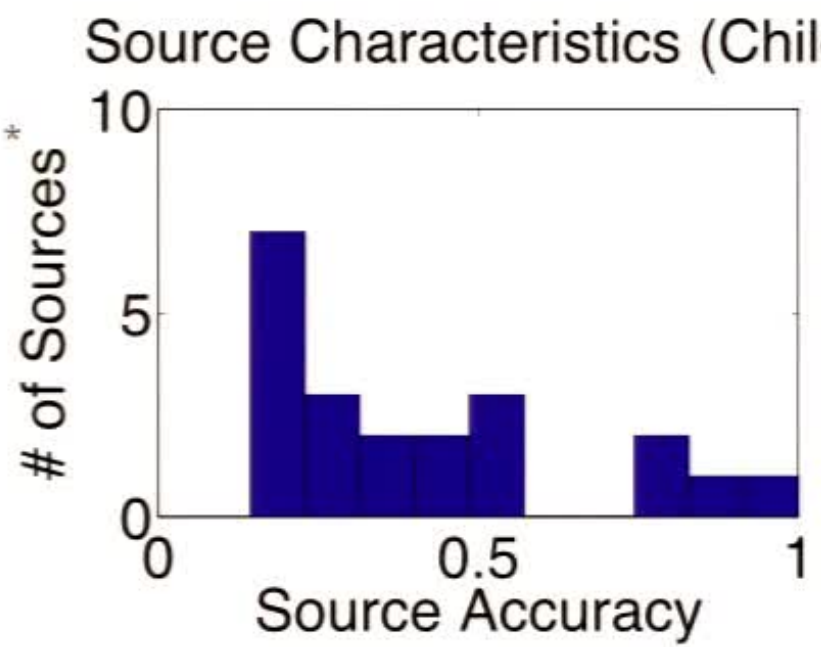
It works !?

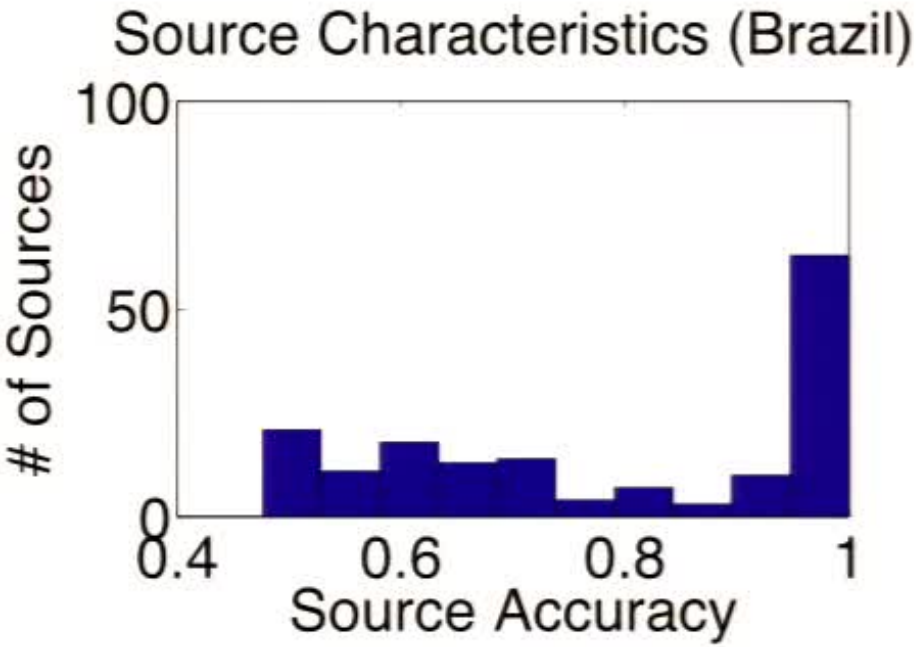
At least for common diseases



What about rare ones?

Other problems???





*Sources are news publishers

Sources exhibit different delays and thus have different accuracies

SourceSeer's approach

Input

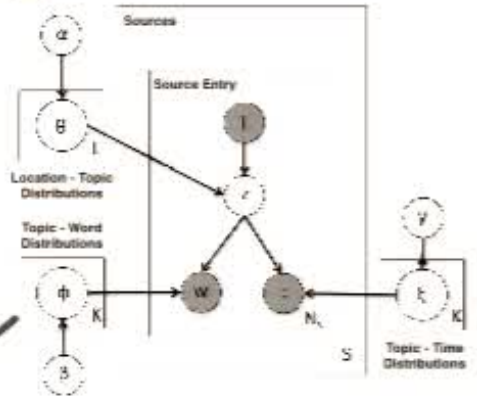
Detect disease topics

biobiochile.cl

ln1

lanoticia1.com

El Repuertero



Does a news paper predict a disease?

V.S.

Fuse expert votes and predict an outbreak

SourceSeer's approach

Forecast Model

Each source is an *expert*

Find the *reliable experts* per location

Intuition: Shouldn't I trust local sources more?

Task of SourceSeer

Each article associated with
a **source**, a **time stamp** and a **location**

Source: *http://www.elrepuertero.cl/*

Time: *Dec 11, 2012 01:12*

Location: *Chaiten, Chile*

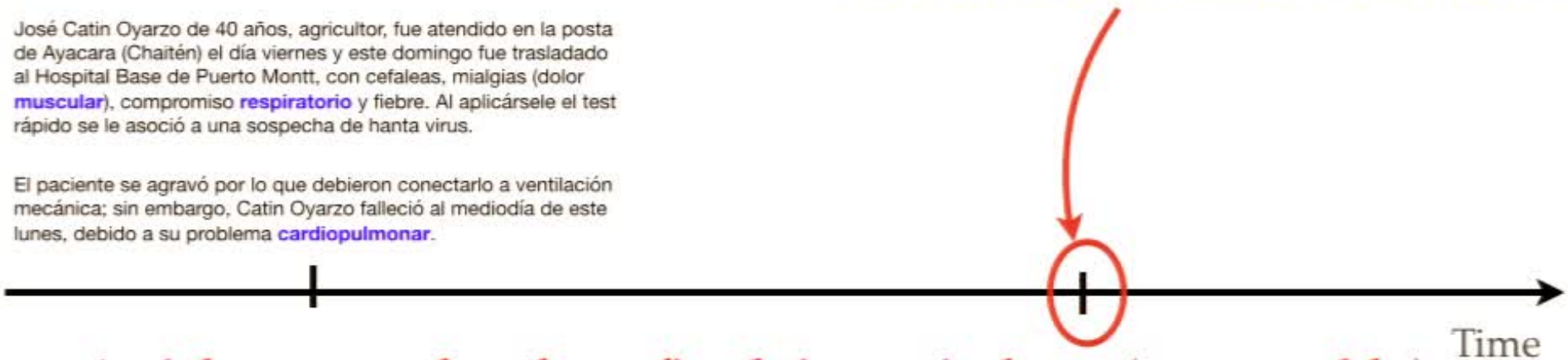
Fallece **paciente**
sospechoso de virus **Hanta**
en Puerto Montt

Este lunes falleció en el **hospital** de Puerto Montt un presunto caso de virus hanta, proveniente del sector rural de Buil, en Chaitén.

José Catin Oyarzo de 40 años, agricultor, fue atendido en la posta de Ayacara (Chaitén) el día viernes y este domingo fue trasladado al Hospital Base de Puerto Montt, con cefaleas, mialgias (dolor **muscular**), compromiso **respiratorio** y fiebre. Al aplicársele el test rápido se le asoció a una sospecha de hanta virus.

El paciente se agravó por lo que debieron conectarlo a ventilación mecánica; sin embargo, Catin Oyarzo falleció al mediodía de este lunes, debido a su problema **cardiopulmonar**.

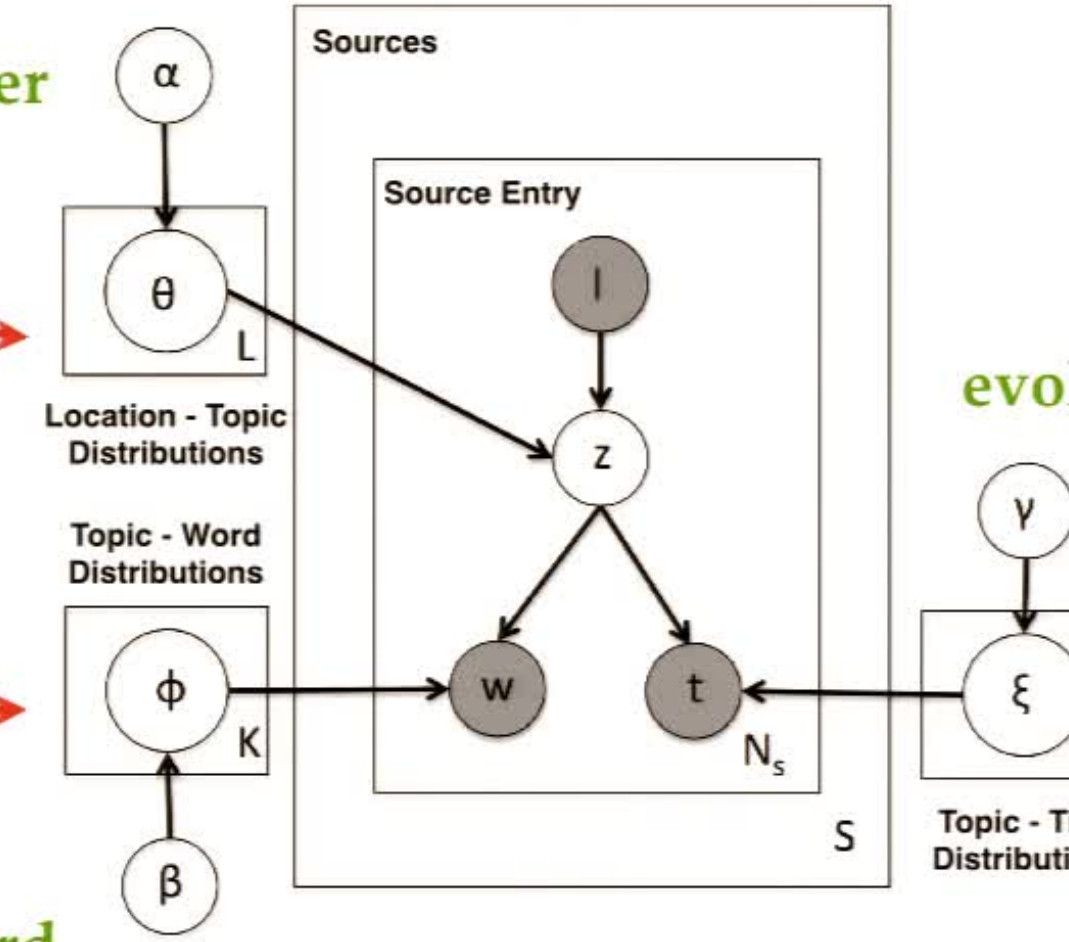
Will an outbreak occur in the next time window?



Articles are updated per fixed time windows (e.g., weekly)

Spatio-temporal topic model

Detect **topic prominence per location**



Track **topic evolution over time**

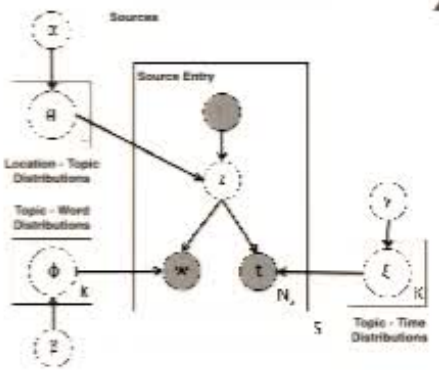
Detect **word correlations**

Forecast outbreaks: Single source

Our model detects **disease topic z**

t: future time point

For a **location l**, how **similar** is the content of **source s** to topic **z**?



Relevance(s, z; l, t)

=

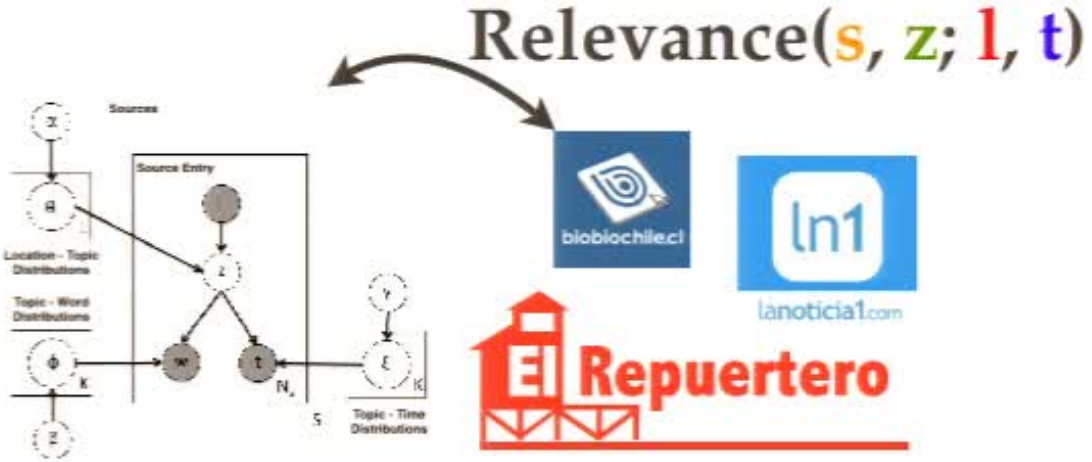
CosineSim ($F_{s,l,t}$, F_z)

F_z: representative document for topic

$F_{s,l,t}$: estimated content of source

Statistical models to estimate F_z and $F_{s,l,t}$

Forecast outbreaks: Single source



Intuition: Spikes in relevance indicate **anomalies**, i.e., potential outbreaks!

Detect anomalies using OCSVMs

for each source-location pair

features are topic relevances

Forecast outbreaks: Multiple sources

Each source-location OCSVM is an individual expert!

Fuse predictions using **weighted majority voting**

Combine expert accuracies to find prediction's **confidence score**

Multiplicative weights to learn accuracies

Delayed ground truth after outbreaks occur

Experimental highlights

Goal

Forecast **Hantavirus** outbreaks in Latin America

HealthMap

Input data: corpus from June`12 to Mar`14 with mentions to **Dengue**, **Avian Flu**, **Swine Flu** and **Hantavirus**

ProMED mail

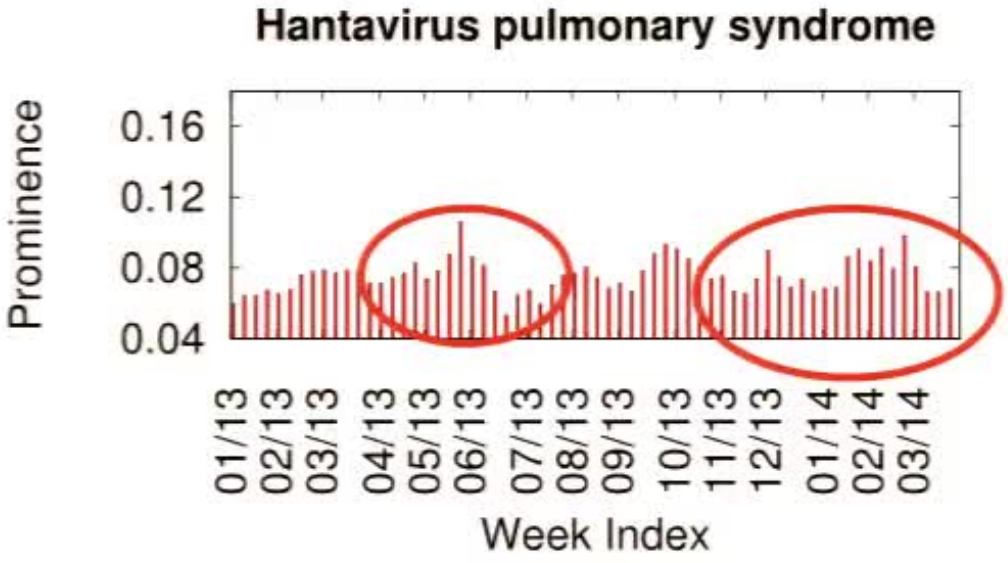
Evaluation: gold standard report provided by analysts Jan`13 to Mar`14

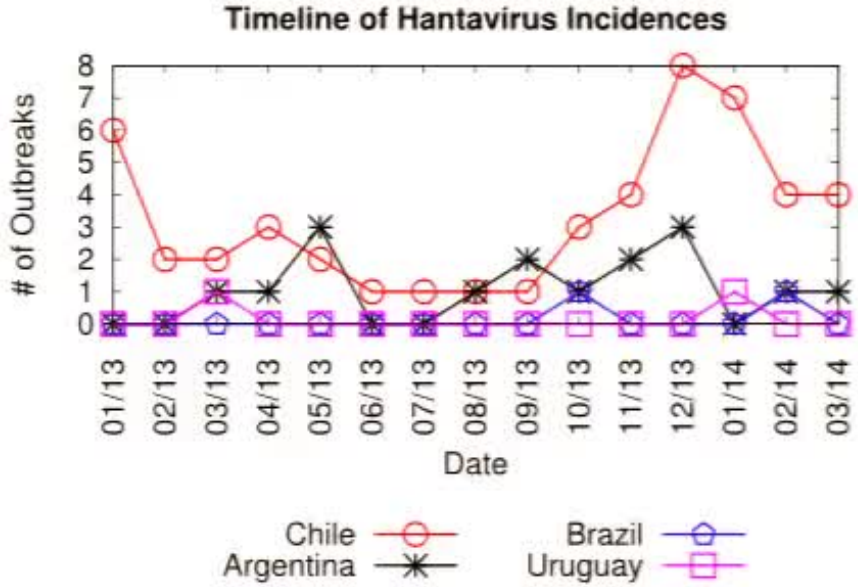
Topic discovery

Hantavirus pulmonary syndrome		Dengue fever	
virus	0.0468	dengue	0.2095
epidemia	0.0443	aegypti	0.0166
enfermos	0.0066	agua	0.0137
hanta	0.0068	mosquitos	0.0058
viral	0.0038	agricultura	0.0019
territorio	0.0027	respiratoria	0.0018
pneumonia	0.0014	rurales	0.0006
sangre	0.0014	agropecuano	0.0006
ratones	0.006	hemorragias	0.0005
cariopulmonar	0.0002	suero	0.0004

SourceSeer detects topics related to **all diseases (both rare and common)** present in the corpus

Temporal pattern discovery





Prominence spikes aligned with occurrences

Forecasting outbreaks

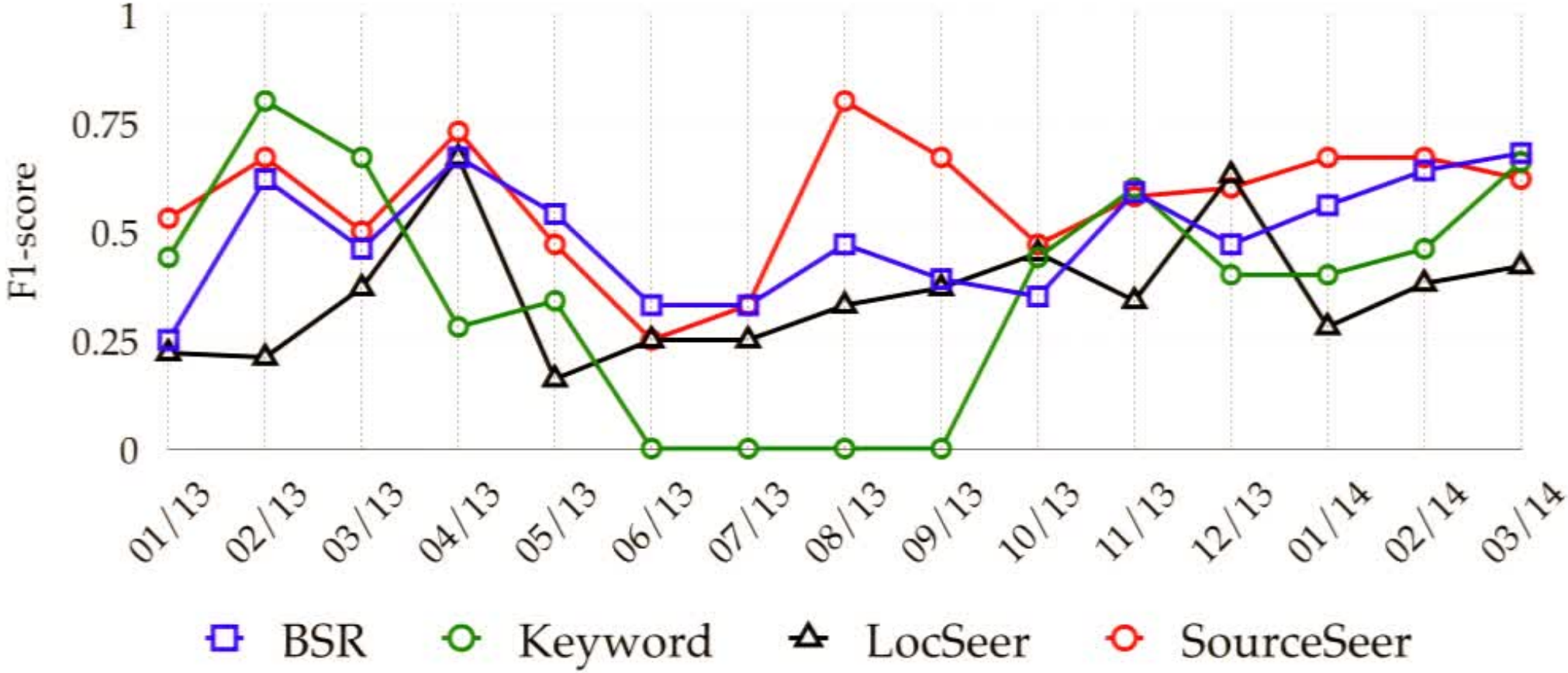
Algorithms

SourceSeer: source based predictions
LocSeer: topic model only; ignore source focus and quality
Keyword: keyword based technique plus OCSVM for anomalies
BSR: fixed rate of occurrence; uses ground truth

Evaluation metrics

F1-score: considers only if an occurrence was **correctly predicted per country per week** (**ignores location and exact day**)
Quality: accuracy with respect to **exact** location (i.e., state) and day (**We forecasted an outbreak but was the place and time correct?**)

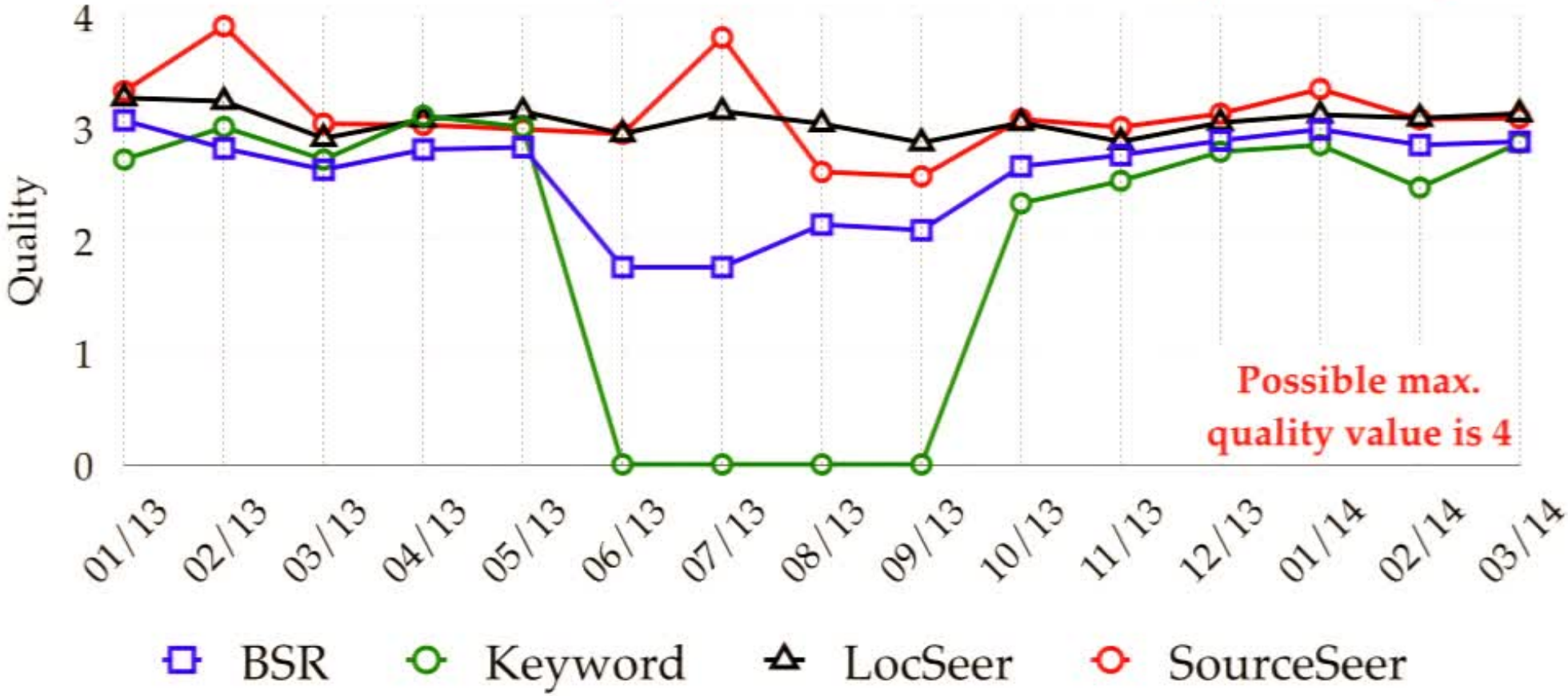
Is SourceSeer more accurate?



SourceSeer more **accurate** than baselines

Statistical significance: Wilcoxon signed-rank test

What about prediction quality?



Many times SourceSeer forecasts correctly at the state level!

Avg. lead-time ~ 8 days

Conclusions

Analyze **open source indicators** to forecast rare-disease outbreaks

Introduced SourceSeer that combines **spatio-temporal topic models** with multi-source **anomaly detection**

SourceSeer forecasts outbreaks more **accurately** at a **finer spatial granularity** with a **better lead-time** than baselines

Conclusions

Analyze **open source indicators** to forecast rare-disease outbreaks

Introduced SourceSeer that combines **spatio-temporal topic models** with multi-source **anomaly detection**

SourceSeer forecasts outbreaks more **accurately** at a **finer spatial granularity** with a **better lead-time** than baselines

Supported by

IARPA
OSI, D12PC00337

Delivered to you by

EMBERS
Early Model-Based Event Recognition with Surrogates

Thank you!

thodrek@cs.umd.edu