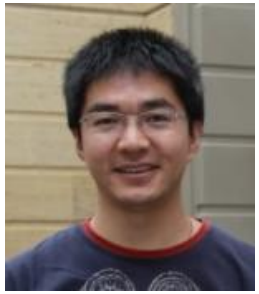
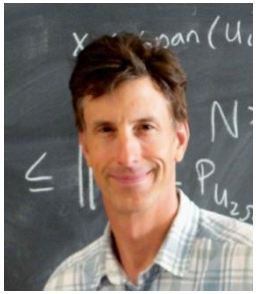
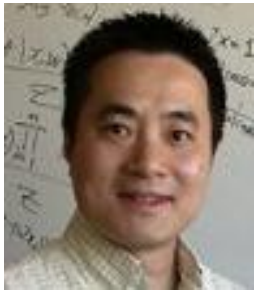
ECML-PKDD 2012

THE UNIVERSITY of WISCONSIN MADISON

Socioscope:
Spatio-Temporal Signal Recovery from Social Media



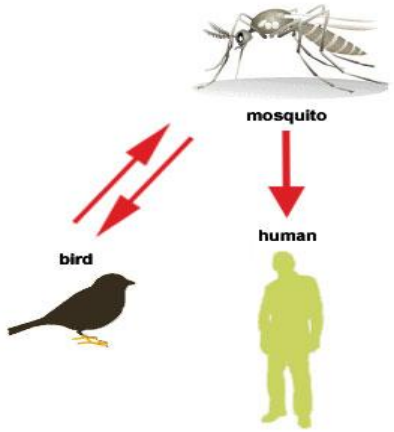



Jun-Ming Xu†, Aniruddha Bhargava*, Robert Nowak*, Xiaojin Zhu†*

†Department of Computer Sciences
*Department of Electrical and Computer Engineering
University of Wisconsin-Madison

Spatio-temporal Signal: When, Where, How Much

Public Health



"**100** dead robins found in **New York last Friday**"

Transportation Safety

"**16** deer got run over by cars in **Wisconsin last month** "

Direct instrumental sensing is difficult and expensive

Humans as Sensors

"I saw a dead crow on its back in the road. It was a bit SPLAT! I thought it had fallen out of the sky."

Created at: 2012-09-26 17:35:23

Location: Madison, WI

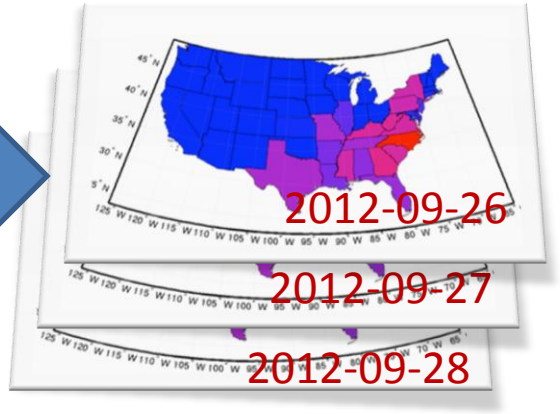
Humans as Sensors

Text Pre-process

Time	Location
2012-09-26 17:35:23	Wisconsin US
2012-09-27 12:17:52	N/A
2012-09-27 08:28:12	(-98.24, 23.22)
…	

Socio-scope

Spatio-temporal signal



Socioscope is not for hot trending topics. Instead we want to precisely recover the intensity of pre-defined target phenomenon.

Challenges of Using Humans as Sensors

Keyword doesn't always mean event

"I was just told I look like dead crow."

"Don't blame me if one day I treat you like a dead crow."

Human sensors aren't under our control

"You are such a 'lazy sensor.' Stop watching Olympic Games! Go to the forests and count the dead birds for us! Now!"

Location stamps may be erroneous or missing

3% have GPS coordinates: *(-98.24, 23.22)*

47% have valid user profile location: *"Bristol, UK"*, *"New York"*

50% don't have valid location information

"Hogwarts," "In the traffic..blah," "Sitting On A Taco"

Socioscope: Problem Definition

Input:

A list of time and location stamps of the target posts.

Time	Location
2012-09-26 17:35:23	New York US
2012-09-27 12:17:52	N/A
2012-09-27 08:28:12	(-98.24, 23.22)
…	

Output: $f_{s,t}$

Intensity of target phenomenon at location *s* (e.g., New York) and time *t* (e.g., 0-1am)

		Time (*t*)		
		0-1am	1-2am	2-3am
Location (*s*)	California	f(1,1)	f(1,2)	f(1,3)
	New York	f(2,1)	f(2,2)	f(2,3)
	Washington	f(3,1)	f(3,2)	f(3,3)

Why Simple Estimation is Bad

$f_{s,t} = x_{s,t}$, the count of target posts in bin (*s, t*)

Justification: MLE of the model $x \sim Poisson(f)$

However,

- Population Bias

  Even $f_{s,t} = f_{s',t'}$, if more users in (*s, t*) , then $x_{s,t} > x_{s',t'}$
- Imprecise location

  Posts without location stamp, noisy user profile location
- Zero/Low counts

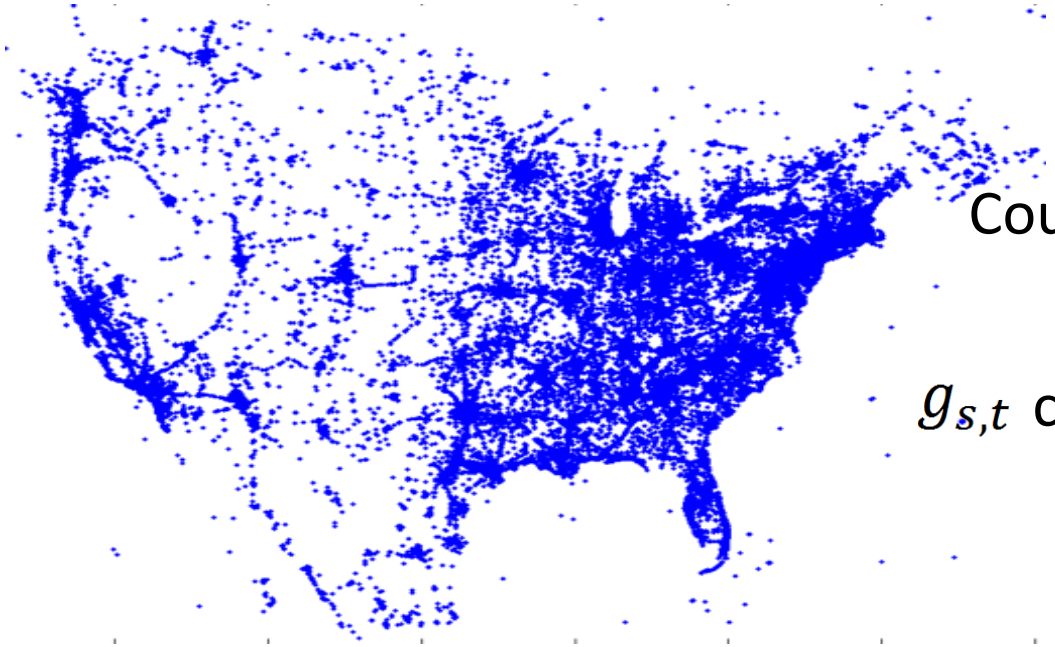
  If no tweeters in Antarctica, does it mean no penguins there?

Correcting Population Bias

Social media user activity intensity $g_{s,t}$

$$x \sim Poisson(\eta(f, g))$$

Link function (target post intensity) $\eta(f, g) = f.g$



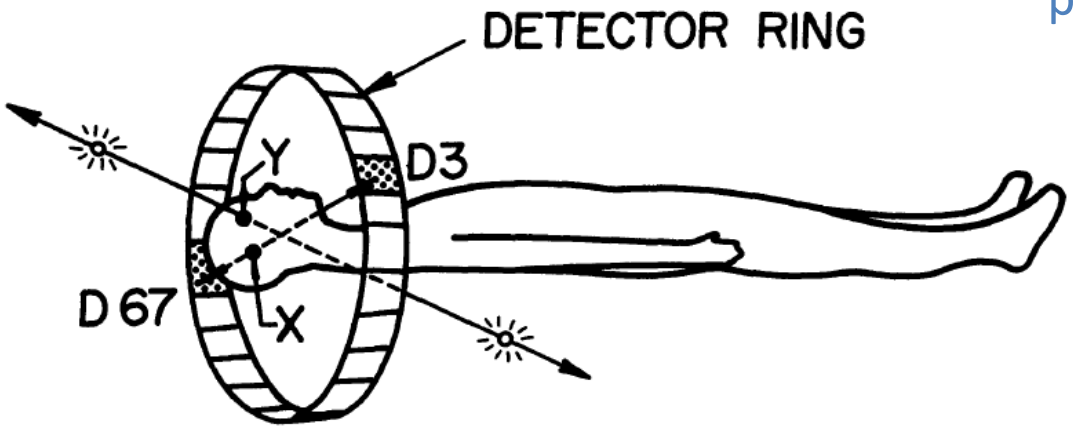
Count of all posts

$$z \sim Poisson(g)$$

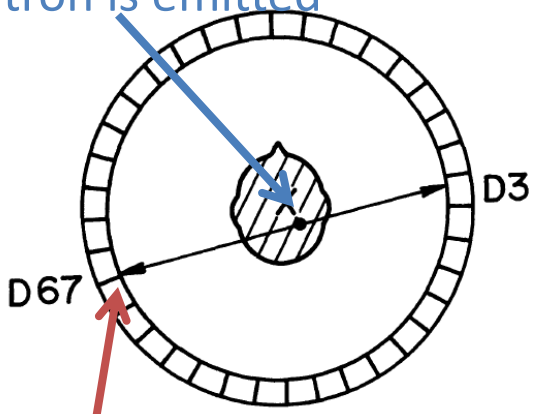
$g_{s,t}$ can be accurately recovered

Handling Imprecise Location

Positron Emission Tomography (PET)



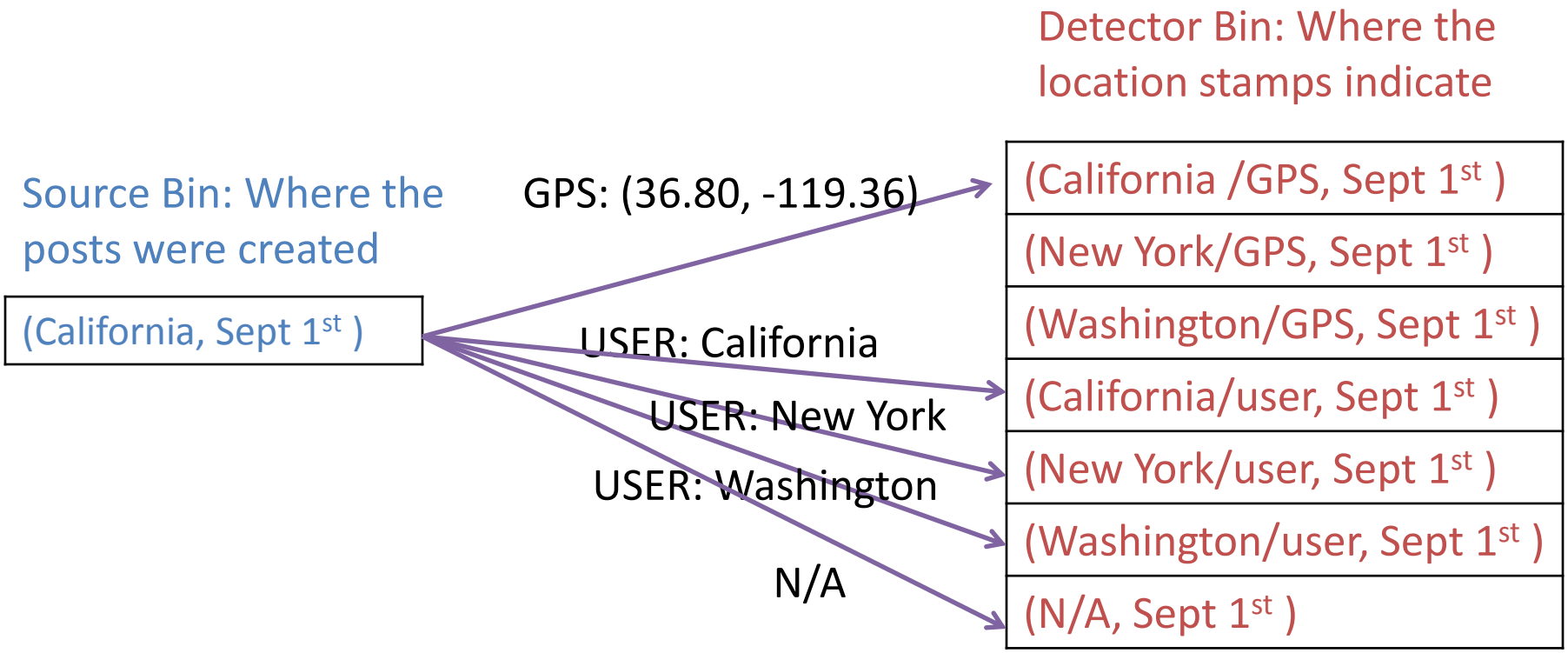
Source Bin (voxel in brain): Where positron is emitted



Detector Bin (detector ring): Where positron is detected

[Reproduced from Vardi *et al*(1985), A statistical model for positron emission tomography]

Handling Imprecise Location (cont.)



Handling Imprecise Location (cont.)

Fraction of posts with GPS coordinates

Fraction of posts without location stamps

.03	0	0
0	.03	0
0	0	.03
.37	.1	.01
.08	.3	.01
.02	.07	.45
.5	.5	.5

Probability that user was in California, but profile location is New York

x

Source Bin: Where the posts were created

(California, Sept 1st)
(New York, Sept 1st)
(Washington, Sept 1st)

Intensity $\eta(f, g)$

=

Detector Bin: Where the location stamps indicate

(California /GPS, Sept 1st)
(New York/GPS, Sept 1st)
(Washington/GPS, Sept 1st)
(California/user, Sept 1st)
(New York/user, Sept 1st)
(Washington/user, Sept 1st)
(N/A, Sept 1st)

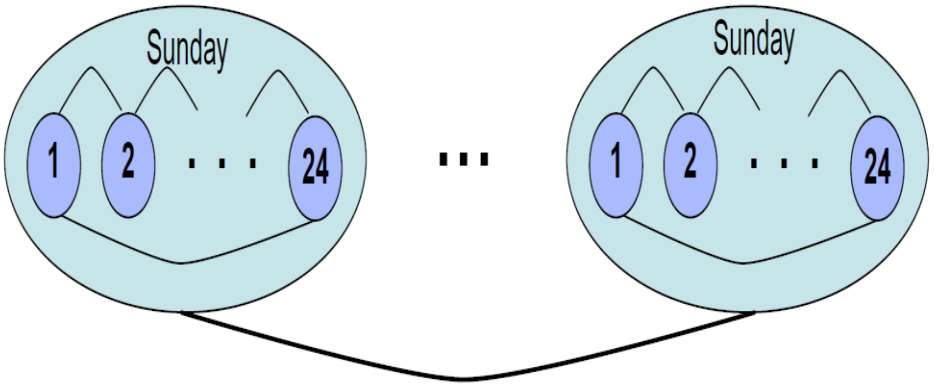
Intensity $h_i = \sum_{j=1}^{n} P_{ij}\eta(f_j, g_j)$

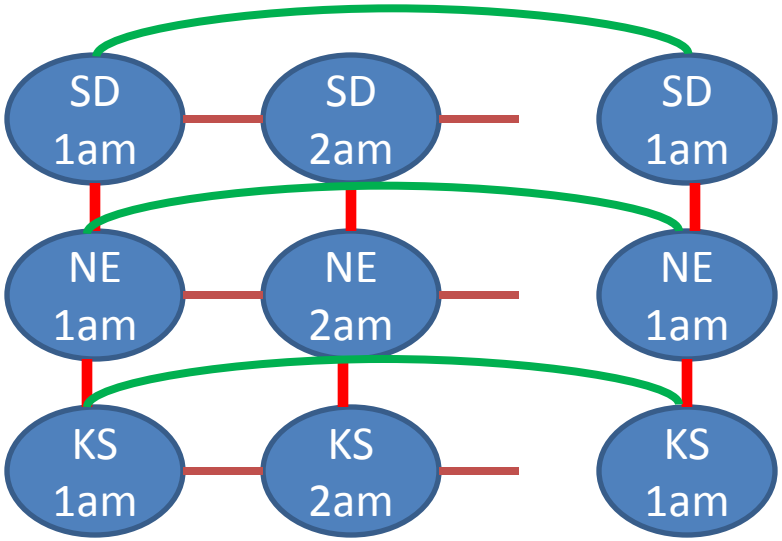
$x_i \sim Poisson(h_i)$

Handling Zero/Low Counts

Spatial Smoothness

Temporal Smoothness





Weight Matrix W

$$D_{jj} = \sum_{k=1}^{n} W_{jk}$$

Graph Laplacian $L = D - W$

Regularizer $\Omega(f) = \frac{1}{2} \log f^T L \log f$

Optimization and Tuning

$$\min_{\theta} -\sum_{i=1}^{m} (x_i \log h_i - h_i) + \lambda \Omega(\theta)$$

$$\theta_j = \log f_j \qquad h_i = \sum_{j=1}^{n} P_{ij} f_j g_j$$

Quasi-Newton method (BFGS)

Cross-Validation

- Data-based and objective approach to regularization
- Sub-sample events from the total observations

Theoretical Consideration

How many posts do we need to obtain reliable recovery?

$$\text{If } x \sim Poisson(h), \text{ then } E\left[\left(\frac{x-h}{h}\right)^2\right] = h^{-1} \approx x^{-1}$$

more counts, less error

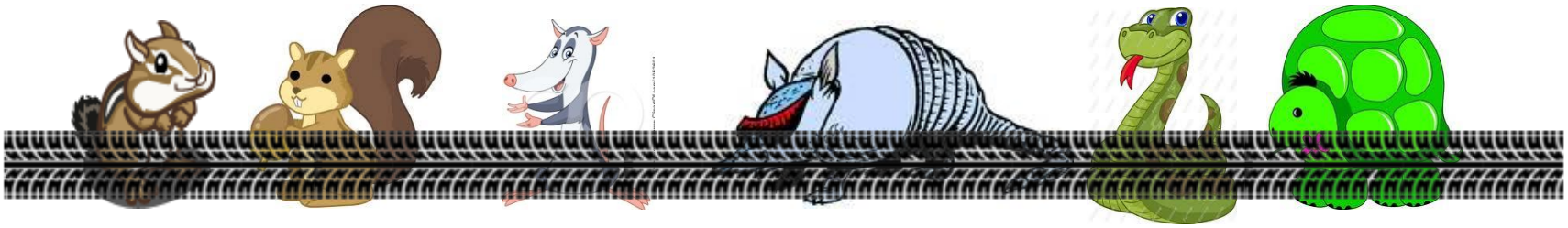
Theorem 1. *Let f be a Hölder α-smooth d-dimensional intensity function and suppose we observe N events from the distribution Poisson(f). Then there exists a constant $C_\alpha > 0$ such that*

$$\inf_{\widehat{f}} \sup_{f} \frac{\mathbf{E}[\|\widehat{f} - f\|_1^2]}{\|f\|_1^2} \geq C_\alpha N^{\frac{-2\alpha}{2\alpha+d}},$$

Best achievable recovery error is inversely proportional to N with exponent depending on the underlying smoothness

Case Study: Roadkill

The intensity of roadkill events within the continental US



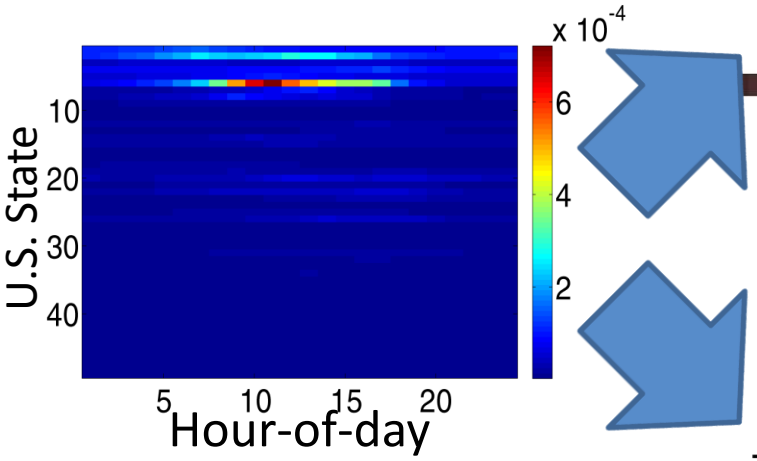
Spatio-Temporal resolution:

State: 48 continental US states, hour-of-day: 24 hours

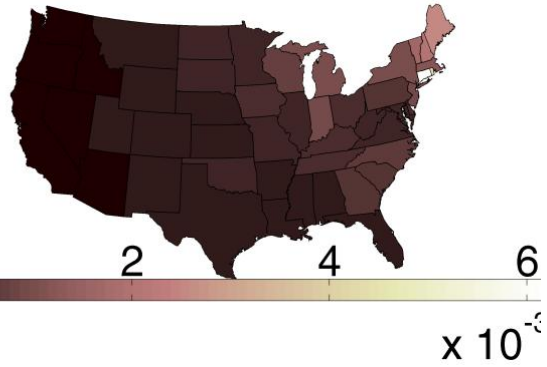
Data source: Twitter

Text classifier: Trained with 1450 labeled tweets. CV accuracy 90%

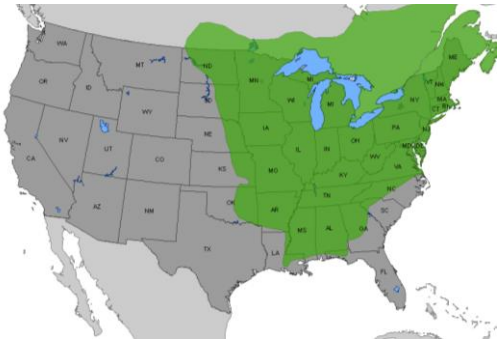
Chipmunk Roadkill Results



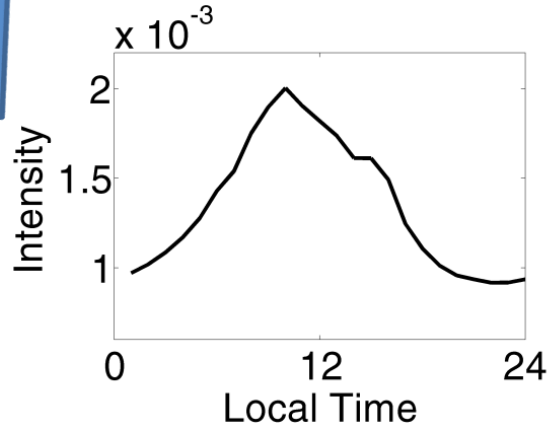
Spatio-temporal distribution
Recovered by Socioscope



Socioscope – Spatial



Range Map

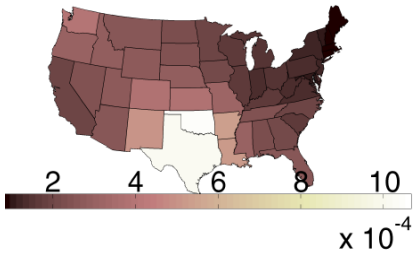


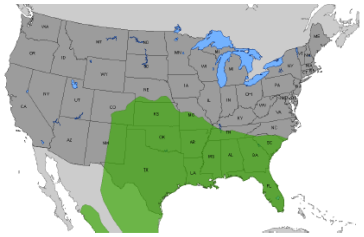
Socioscope – Temporal

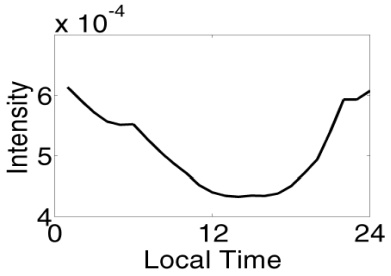
"Chipmunks are diurnal."

Activity Pattern

Roadkill Results on Other Species

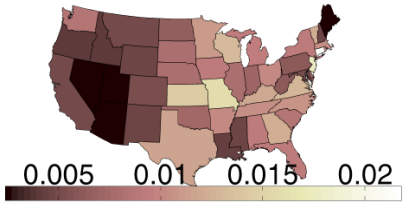


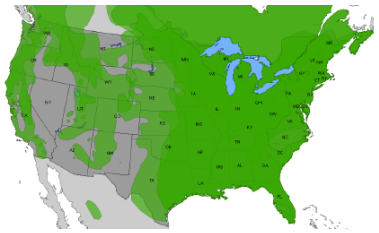


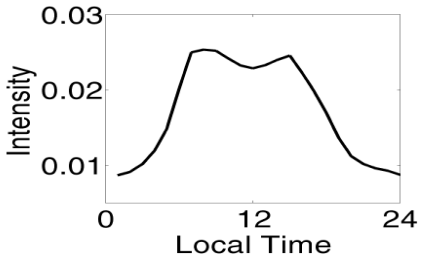


"Armadillos are nocturnal"

Armadillos

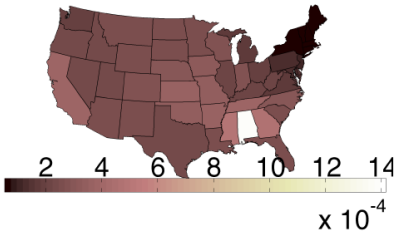


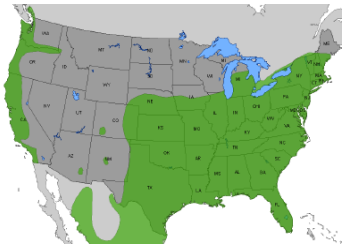


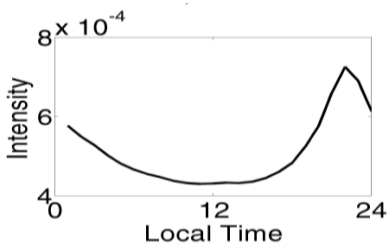


"Most squirrels are diurnal"

Squirrels







"Opossums are nocturnal"

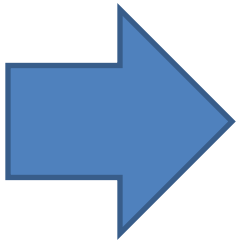
Opossums

Future Work

Incorporate text classification confidence in the input

Time	Location
2012-09-26 17:35:23	Wisconsin US
2012-09-27 12:17:52	N/A
2012-09-27 08:28:12	(-98.24, 23.22)
…	

Target post only



Text Classifier Confidence	Time	Location
0.9	2012-09-26 17:35:23	Wisconsin US
0.2	2012-09-26 17:38:33	N/A
0.6	2012-09-27 12:17:52	N/A
0.05	2012-09-27 13:13:28	(-105.24, 35.82)
0.7	2012-09-27 08:28:12	(-98.24, 23.22)
	…	

Future Work (cont.)

Handle the time delay and spatial displacement between the target event and the generation of a post

"*So the pigeon I ran over yesterday must have some bird friends in high places. Car is full of bird shit*."

"Ran over a chipmunk on my way 2 work this morning ☹"

Incorporate Psychology factors :

Will you post a tweet about running over a …?

Thanks!

We thank Megan K. Hines for her guidance on wildlife.

This work is supported in part by Global Health Institute at the University of Wisconsin-Madison.