Machine Learning for Zoonotic Emerging Disease Detection

Xiaojin Zhu, Jun-Ming Xu, Christine M. Marsh, Megan K. Hines, F. Joshua Dein University of Wisconsin-Madison & USGS

Background

We might have had an earlier identification of West Nile virus ten years ago, had people reported that they were seeing dead crows in their backyard.

75% of emerging pathogens are zoonotic, exchanging between humans and other animals. Examples include Avian Influenza, SARS, and West Nile Virus. Unfortunately, even in developed countries there are no comprehensive systems for wildlife disease surveillance because

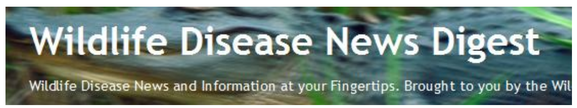
1. Nobody owns wildlife (unlike pets or livestock)
2. It is not clear where and how to report.

Our Position

With machine learning, we can form such a wildlife disease surveillance system with 4 layers:

1. **The experts** at the Wildlife Health Center (USGS) who can provide definite diagnosis on a small number of cases.

2. **Local news reports** often contain significant wildlife incidents.



3. **Citizen scientists** can be organized to observe and report wildlife to professionals via specific channels.

4. **Social media** users acting as incidental observers.

Dead armadillo on the side of the road with a buzzard picking at it; what a lovely sight on my trip to work. :P

Machine Learning Challenges

1. **Unsupervised, semi-supervised, weakly supervised learning:**
   - expert data is labeled with diagnosis;
   - citizen scientists and social media data is not;
   - many features can be missing (only 1% tweets has coordinates)
   - many features contain noise (tweets usually not generated at the same time & place as the event)
2. **Cost-sensitive active learning and ranking:**
   - experts' effort is limited
   - citizen scientists and social media data forms the pool
3. **Topic detection and tracking**
   - identifying new wildlife health events
4. **Social networks**
   - "advertizing" to selected individuals who most likely will become citizen scientists and influence others to do the same
5. **Natural language processing**
   - citizen scientists and social media data can be noisy
   - categorization: real wildlife events or not
   - Information extraction: what, when, where, how
   - multilingual processing
6. **Bias correction**
   - human presence bias
   - human psychological bias
7. **Sparse signal recovery**
   - wildlife events tend to be spatially and temporally sparse
8. **Computer vision**
   - citizen scientists and social media data may contain pictures and video
   - help identifying species and environment
9. **Developing countries**
   - lacking in all 4 layers
   - rapid adoption of technology (e.g. cell phones) is changing that

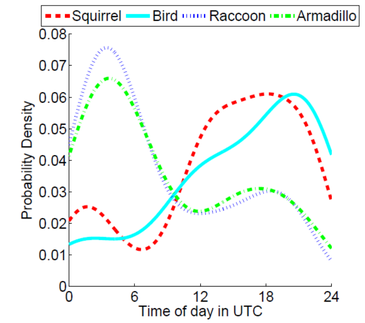
Example: Roadkill

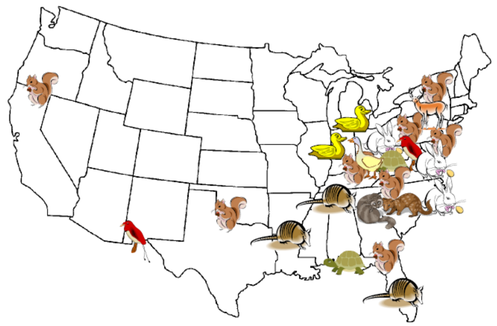
Sample tweets:

- *oh god me and hannah just saw a dead squirrel by the road and screamed... awful*
- *she ran over a kangaroo yesterday at the new side of epping?! i never knew kangaroos are anywhere near the city! just heard crazy incidents*
- *I'm 78% sure I ran over a dolphin with a jet ski today. Shut your mouth hippies, Earth day is over. Next time I'm aiming for the manatees.*

Shallow natural language processing: twitter stream API; pattern match "ran over" or "dead … road"; extract species, time, location. ~120 roadkill tweets a day.

Animal	Frequency	Animal	Frequency
SQUIRREL	22	ARMADILLO	3
RABBIT	11	FOX	3
BIRD	11	RAT	2
SKUNK	10	TURKEY	2
SNAKE	9	GOOSE	2
TURTLE	7	OPOSSUM	2
DUCK	6	MOUSE	2
FROG	5	BEAVER	1
DEER	5	CHIPMUNK	1
RACCOON	4	BEAR	1





Acknowledgment

We thank the workshop organizers for helpful comments. Research supported in part by Great Lakes-Northern Forests CESU Agreement 07HQAG0150.

 jerryzhu@cs.wisc.edu