

**The First International Workshop on Web Science and Information Exchange  
in the Medical Web (MedEx 2010)**

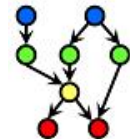
19th World Wide Web Conference WWW-2010  
26-30 April 2010: Raleigh Conference Center, Raleigh, NC, USA

# **Animal Disease Event Recognition and Classification**

**Svitlana Volkova, Doina Caragea, William H. Hsu, Swathi Bujuru**

Laboratory for Knowledge Discovery in Databases  
Department of Computing and Information Sciences  
Kansas State University

Sponsor: K-State National Agricultural Biosecurity Center (NABC)



# Agenda

---

- ▶ Overview
- ▶ Related Work
- ▶ Methodology
  - ▶ Entity Recognition
  - ▶ Event Sentence Classification
  - ▶ Event Tuple Generation
- ▶ Experiment
- ▶ Summary and Future Work



# Animal Infectious Disease Outbreaks

---

- ▶ influence on the travel and trade



- ▶ cause economic crises, political instability;



- ▶ diseases, zoonotic in type can cause loss of life.



# Animal Disease-related Data Online

---

## Structured Data

- ▶ Official reports by different organizations:
  - ▶ state and federal laboratories, bioportals;
  - ▶ health care providers;
  - ▶ governmental agricultural or environmental agencies.



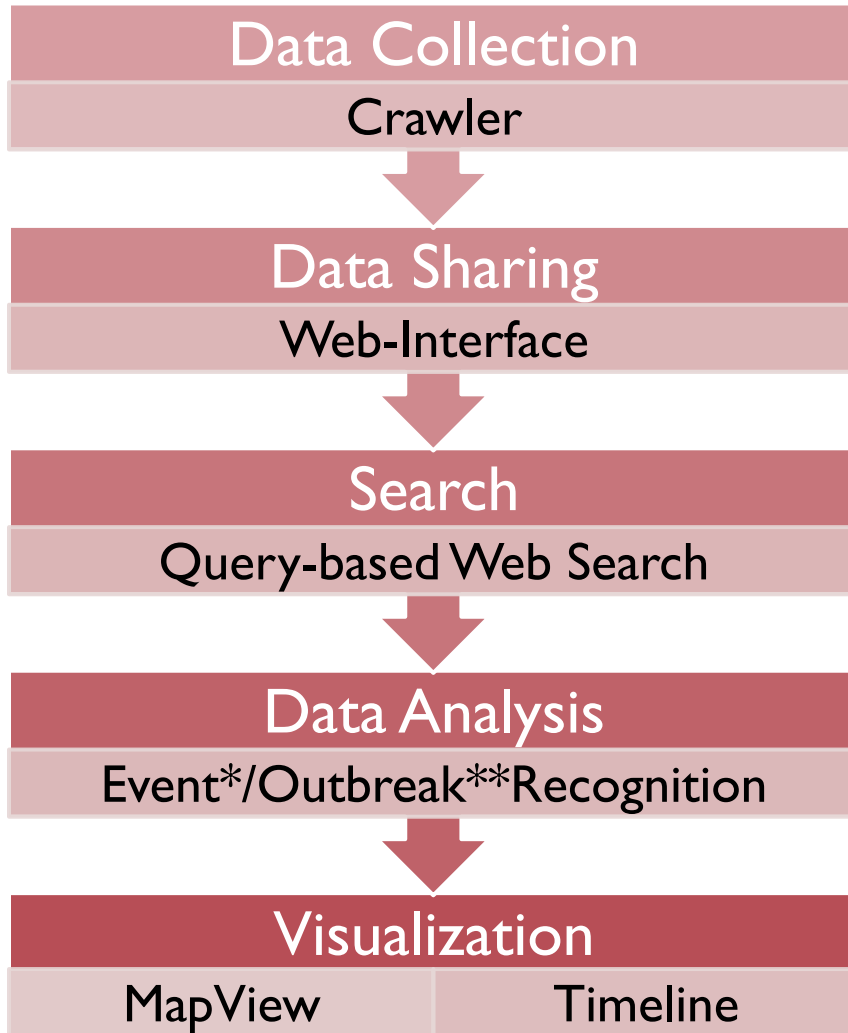
## Unstructured Data

- ▶ Web-pages
- ▶ News
- ▶ E-mails (e.g., ProMed-Mail)
- ▶ Blogs
- ▶ Medical literature (e.g., books)
- ▶ Scientific papers (e.g., PubMed)



# Infectious Disease Informatics (IDI)

---



Event example:

“On **12 September 2007**, a new **foot-and-mouth disease** outbreak was confirmed in **Egham, Surrey**”

\*Event is an occurrence of a disease within a particular time and space range.

\*\*Outbreak is a set of disease-related events which are constrained in space and have temporal overlap



# Problem Statement

---

- ▶ How to recognize animal disease-related events from unstructured web documents?
- ▶ How to accurately extract event-related entities?
  - ▶ Disease Names, Viruses, Serotypes
  - ▶ Locations
  - ▶ Species
  - ▶ Dates
- ▶ How to classify event-related and non-related sentences?
- ▶ How to define the confirmation status of the event?
- ▶ How to generate true event tuples?



# Related Work

---

- ▶ BioCaster - <http://biocaster.nii.ac.jp/>
  - + crawls news and uses ontology pattern matching approaches to recognize disease-location-verb pairs;
  - does not classify events as suspected or confirmed.
- ▶ HealthMap - <http://healthmap.org/en>
  - + crawls data from Google News and the ProMED-Mail;
  - manually supported web system.
- ▶ Pattern-based Understanding and Learning System (PULS) - <http://sysdb.cs.helsinki.fi/puls/jrc/all>
  - does not classify events and does not report past outbreaks.
- ▶ Our approach:
  - ▶ Consider such event attributes as: disease, date, location, species and confirmation status;
  - ▶ Classify events into two categories such as: suspected or confirmed
  - ▶ Identify events in the historical data, e.g. medical literature, papers.

# Methodology

---

- ▶ **Step 1.** Entity recognition from raw text.
- ▶ **Step 2.** Sentence classification from which entities are extracted as being related to an event or not; if they are related to an event we classify them as confirmed or suspected.
- ▶ **Step 3.** Combination of entities within an event sentence into the structured tuples and aggregation of tuples related to the same event into one comprehensive tuple.



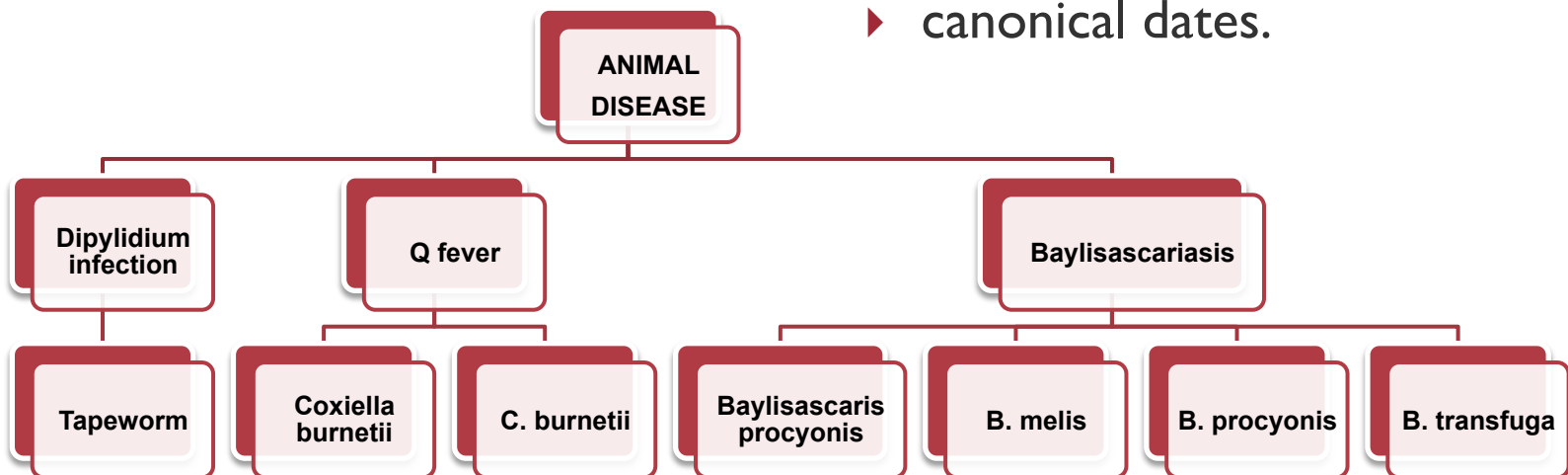
# Domain Meta-data

## Domain-specific knowledge

- ▶ Medical ontology
  - ▶ diseases, serotypes, and viruses.

## Domain-independent knowledge

- ▶ Location hierarchy
  - ▶ names of countries, states, cities;
- ▶ Time hierarchy
  - ▶ canonical dates.



# Step1: Entity Recognition

---

- ▶ Locate and classify atomic elements into predefined categories:
  - ▶ **Disease names:** “foot and mouth disease”, “rift valley fever”;  
**viruses:** “picornavirus”; **serotypes:** “Asia-1”;
  - ▶ **Species:** “sheep”, “pigs”, “cattle” and “livestock”;
  - ▶ **Locations** of events specified at different levels of geo-granularity: “United Kingdom”, “eastern provinces of Shandong and Jiangsu, China”;
  - ▶ **Dates** in different formats: “last Tuesday”, “two month ago”.



# Entity Recognition Tools

---

## ▶ Animal Disease Extractor\*

- ▶ relies on a medical ontology, automatically-enriched with synonyms and causative viruses.

## ▶ Species Extractor\*

- ▶ pattern matching on a stemmed dictionary of animal names from Wikipedia.



## ▶ Location Extractor

- ▶ Stanford NER Tool\*\* (uses conditional random fields);
- ▶ NGA GEOnet Names Database (GNS)\*\*\* for location disambiguation and retrieving latitude/longitude.

## ▶ Date/Time Extractor

- ▶ set of regular expressions.

\*KDD KSU DSEx - <http://fingolfin.user.cis.ksu.edu:8080/diseaseextractor/>

\*\*Stanford NER - <http://nlp.stanford.edu/ner/index.shtml>

\*\*\*GNS - <http://earth-info.nga.mil/gns/html/>

## Step 2: Event Sentence Classification

- ▶ Constraint: True events should include a disease name together with a status verb (eliminate event non-related sentences).
  - ▶ “Foot and mouth disease **is**<sub>[V]</sub> a highly pathogenic animal disease”.
- ▶ Confirmed status verbs “*happened*” and verb phrases “*strike out*”
  - ▶ “On 9 Jun 2009, the farm's owner **reported**<sub>[V]</sub> symptoms of FMD in more than 30 hogs”.
- ▶ Suspected status verbs “*catch*” and verb phrases “*be taken in*”
  - ▶ “RVF is **suspected**<sub>[V]</sub> in Saudi Arabia in September 2000”.
- ▶ Extend the Initial List with synonyms from **GoogleSets\*** and **WordNet\*\***

Status	IN-V	IN-VP	GS-V	GS-VP	WN-V	WN-VP
Suspected	7	1	55	2	37	10
Confirmed	7	1	55	13	48	9

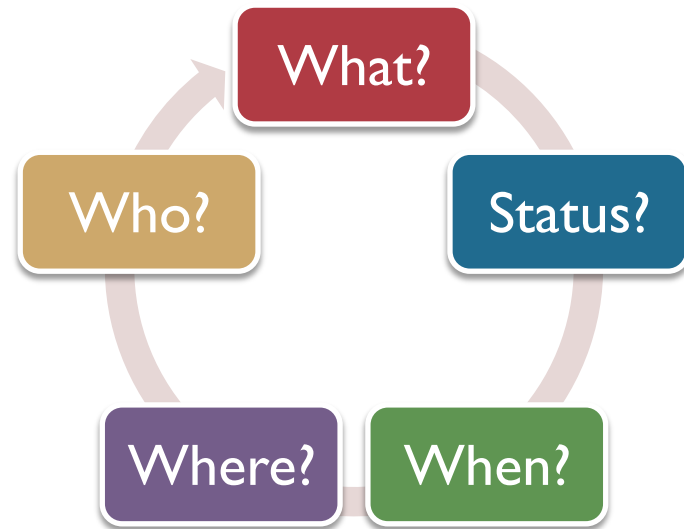
\*GoogleSets - <http://labs.google.com/sets>

\*\*WordNet - <http://wordnet.princeton.edu/>

# Step 3: Event Tuple Generation

## ▶ Event attributes:

- ▶ disease
- ▶ date
- ▶ location
- ▶ species
- ▶ confirmation status



## ▶ Event tuple:

- ▶  $\text{Event}_i = \langle \text{disease}; \text{date}; \text{location}; \text{species}; \text{status} \rangle =$   
 $\langle \text{FMD}, 9 \text{ Jun } 2009, \text{Taoyuan}, \text{hog}, \text{confirmed} \rangle$

## ▶ Event tuple with missing attributes:

- ▶  $\text{Event}_j = \langle \text{FMD}, ?, ?, ?, \text{confirmed} \rangle$

# Event Recognition Workflow



## Step 1: Entity Recognition

**Foot-and-mouth disease**<sub>[DIS]</sub> on **hog**<sub>[SP]</sub> farm in **Taoyuan**<sub>[LOC]</sub>.

Taiwan's TVBS television station reports that agricultural authorities confirmed **foot-and-mouth disease**<sub>[DIS]</sub> on a **hog**<sub>[SP]</sub> farm in **Taoyuan**<sub>[LOC]</sub>. On **9 Jun 2009**<sub>[DT]</sub>, the farm's owner reported symptoms of **FMD**<sub>[DIS]</sub> in more than 30 **hogs**<sub>[SP]</sub>. Subsequent testing confirmed **FMD**<sub>[DIS]</sub>. Agricultural authorities asked the farmer to strengthen immunization. The outbreak has not affected other farms. Authorities stipulated that the affected **hog**<sub>[SP]</sub> farm may not sell pork for 2 weeks.



## Step 2: Sentence Classification

**YES** 1. **Foot-and-mouth disease**<sub>[DIS]</sub> on **hog**<sub>[SP]</sub> farm in **Taoyuan**<sub>[LOC]</sub>.

**YES** 2. Taiwan's TVBS television station **reports** that agricultural authorities **confirmed** **foot-and-mouth disease**<sub>[DIS]</sub> on a **hog**<sub>[SP]</sub> farm in **Taoyuan**<sub>[LOC]</sub>.

**YES** 3. On **9 Jun 2009**<sub>[DT]</sub>, the farm's owner **reported** symptoms of **FMD**<sub>[DIS]</sub> in more than 30 **hogs**<sub>[SP]</sub>.

**YES** 4. Subsequent testing **confirmed** **FMD**<sub>[DIS]</sub>.

**NO** 5. Agricultural authorities asked the farmer to strengthen immunization.

**NO** 6. The outbreak has not affected other farms.

**NO** 7. Authorities stipulated that the affected **hog**<sub>[SP]</sub> farm may not sell pork for 2 weeks.



## Step 3a: Tuple Generation

E<sub>1</sub> = <**Foot-and-mouth disease**, ?, **Taoyuan**, **hog**, ?>

E<sub>3</sub> = <**FMD**, **9 Jun 2009**, ?, **hog**, **reported**>

E<sub>2</sub> = <**Foot-and-mouth disease**, ?, **Taoyuan**, **hog**, **confirmed**>

E<sub>4</sub> = <**FMD**, ?, ?, ?, **confirmed**>



## Step 3b: Tuple Aggregation

E = <**disease**, **date**, **location**, **species**, **status**> = <**Foot-and-mouth disease**, **9 Jun 2009**, **Taoyuan**, **hog**, **confirmed**>



---

**Algorithm 1 Entity Recognition, Sentence Classification and Tuple Generation**

---

Input: Set of web documents  $D$

Output: Set of extracted events  $e_k \in E$  for each document  $d_j \in D$

```
foreach document  $d_j \in D$  do
   $S = \text{TokenizeToSentences}(d_j)$ ;
  foreach sentence  $s_i \in S$  do
     $disease = \text{ExtractDiseaseEntities}(s_i)$ ;
    if  $disease \neq \emptyset$  then
       $status = \text{ExtractConfirmationStatus}(s_i)$ ;
      if  $status \neq \emptyset$  then
         $date = \text{ExtractDateEntities}(s_i)$ ;
         $location = \text{ExtractLocationEntities}(s_i)$ ;
         $species = \text{ExtractSpeciesEntities}(s_i)$ ;
      else
        skip sentence  $s_i$ ;
      end;
    else
      skip sentence  $s_i$ ;
    end;
  end;
   $E = \text{GenerateTuples}(disease, date, location, species, status)$ ;
   $e_k = \text{AggregateTuples}(E)$ ;
end.
```

---

Tuple aggregation is based on the set of rules:

- disease name is the main attribute for the event tuple;
- optimally combine the event tuples around a disease name (event tuple should have max number of event attributes);
- if tuple has a new disease, then form the next tuple.

## **Rule-based event recognition approach**

Step 1.  
Entity Recognition

Step 2.  
Event Sentence  
Classification

Step 3.  
Event Tuple  
Generation &  
Aggregation



# Experiment

---

- ▶ ~100 event-related documents
  - ▶ Foot-and-mouth disease (FMD)
  - ▶ Rift valley fever (RVF)
- ▶ Manually created 2 sets of summaries for 100 docs
- ▶ DUCView Pyramid Scoring Tool\* – Score [0..1]
  - ▶ relies on multiple summaries to assign the significance weights to summarization content units (i.e., entities)
  - ▶ to compare automatically generated event tuples with entities from human summaries.



$$\text{Score}_i = \langle w_d \text{disease}; w_t \text{date}; w_l \text{location}; w_s \text{species}; w_c \text{status} \dots \rangle,$$

subject to disease + status = 2



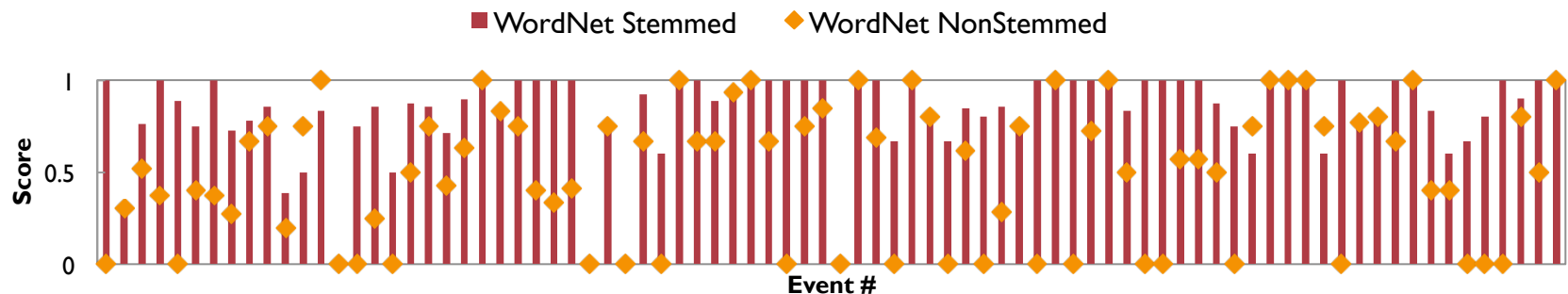
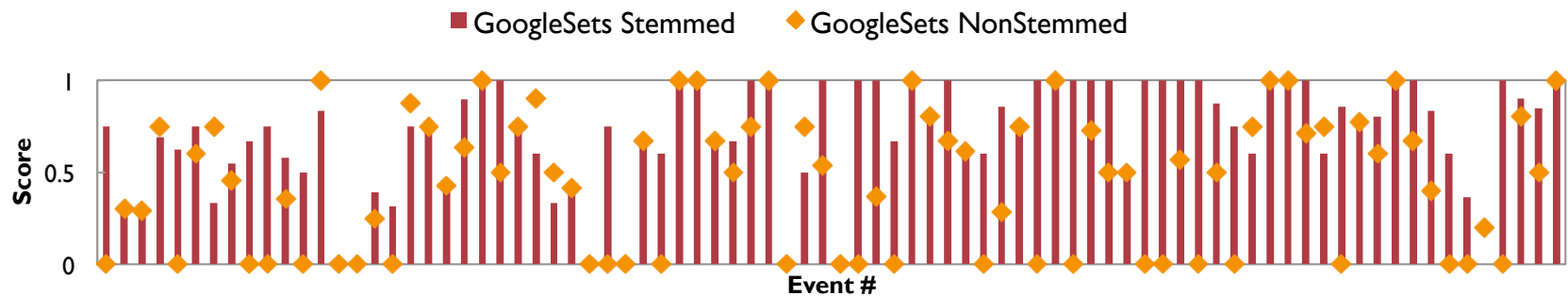
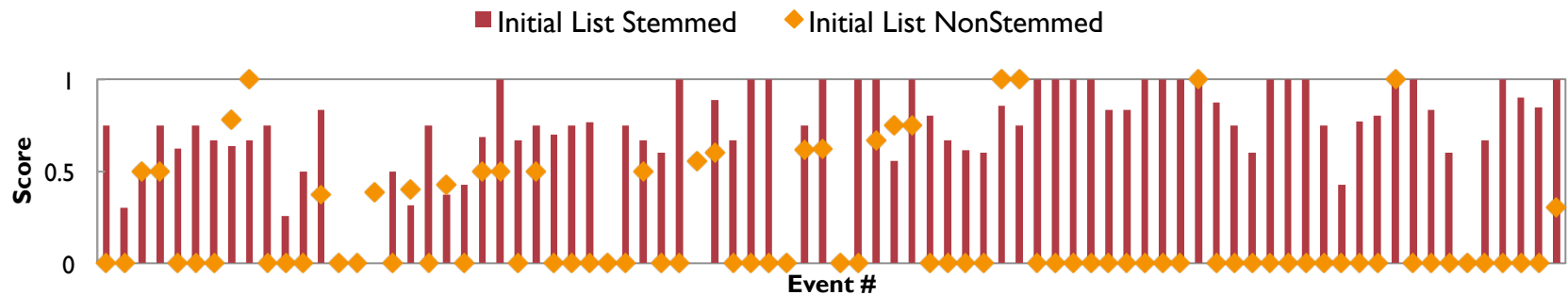
# Event Score Distribution by Range

- ▶ We interpret the Pyramid score values as an event extraction accuracy:
  - ▶ # of unique contributing entities (TP);
  - ▶ # of entities not in the summary (FP);
  - ▶ # of extra contributing entities from summary (FN).
  - ▶ multiple summaries – majority voting for annotation.

Score	IN-NS	GS-NS	WN-NS	IN-S	GS-S	WN-S
Low [0 - 0.3]	73%	43%	38%	19%	18%	13%
Medium [0.31 - 0.7]	18%	27%	29%	27%	30%	13%
High [0.71 - 1]	9%	30%	33%	54%	52%	74%
Mean	0.17	0.40	0.45	0.64	0.65	0.75



# Experimental Results



# Summary & Future Work

---

- ▶ Sentence-based Event Recognition and Classification
- ▶ Entity and Confirmation Status Extraction Methods
  - ▶ apply several lists of verbs for confirmation status extraction: initial list, lists augmented using GoogleSets and WordNet
- ▶ Pyramid method and DUCView Tool for scoring automatically generated event tuples.

## Future Work:

- ▶ Deeper syntactic analysis of the sentence
- ▶ Co-reference resolution
- ▶ Multilingual event recognition and classification



# Thank you!

