4-2014

Leveraging Social Media and Web of Data for Crisis Response Coordination

Carlos Castillo

Fernando Diaz

Hemant Purohit
Wright State University - Main Campus, purohit.5@wright.edu

Leveraging Social Media and Web of Data to Assist Crisis Response Coordination

SDM-2014 Tutorial

Carlos Castillo, Qatar Computing Research Institute, Qatar
Fernando Diaz, Microsoft Research, NYC, USA
Hemant Purohit, Ohio Center of Excellence in Knowledge-enabled Computing (Kno.e.sis), Wright State Univ, USA

Check Tutorial site for latest slides: http://knoesis.org/hemant/present/sdm2014

Introduction

- Carlos Castillo, QCRI
  - *Social Computing, Information Credibility*

- Fernando Diaz, MSR
  - *Temporal Information Retrieval, Crisis Informatics*

- Hemant Purohit, Kno.e.sis, Wright State U
  - *Computational Social Science, Crisis Response Coordination*
  - *NSF SoCS project for improving crisis response by social media*

Outline

- A.) Introduction
  - Role of Web 2.0 data during Crisis
  - Data and General Challenges
- B.) Specific Problems, Methods & Future Research
  - Event Detection
  - Data Collection
  - Information Classification
  - Structured Data Extraction from Unstructured
  - Event Summarization
  - Hybrid Systems: Human+Machine Computing
  - Mining for Actions: Coordination and Decision Making
- C.) Conclusion

Outline

- A.) Introduction
  - Role of Web 2.0 data during Crisis
  - Data and General Challenges
- B.) Specific Problems, Methods & Future Research
  - Event Detection
  - Data Collection
  - Information Classification
  - Structured Data Extraction from Unstructured
  - Event Summarization
  - Hybrid Systems: Human+Machine Computing
  - Mining for Actions: Coordination and Decision Making
- C.) Conclusion

Scope

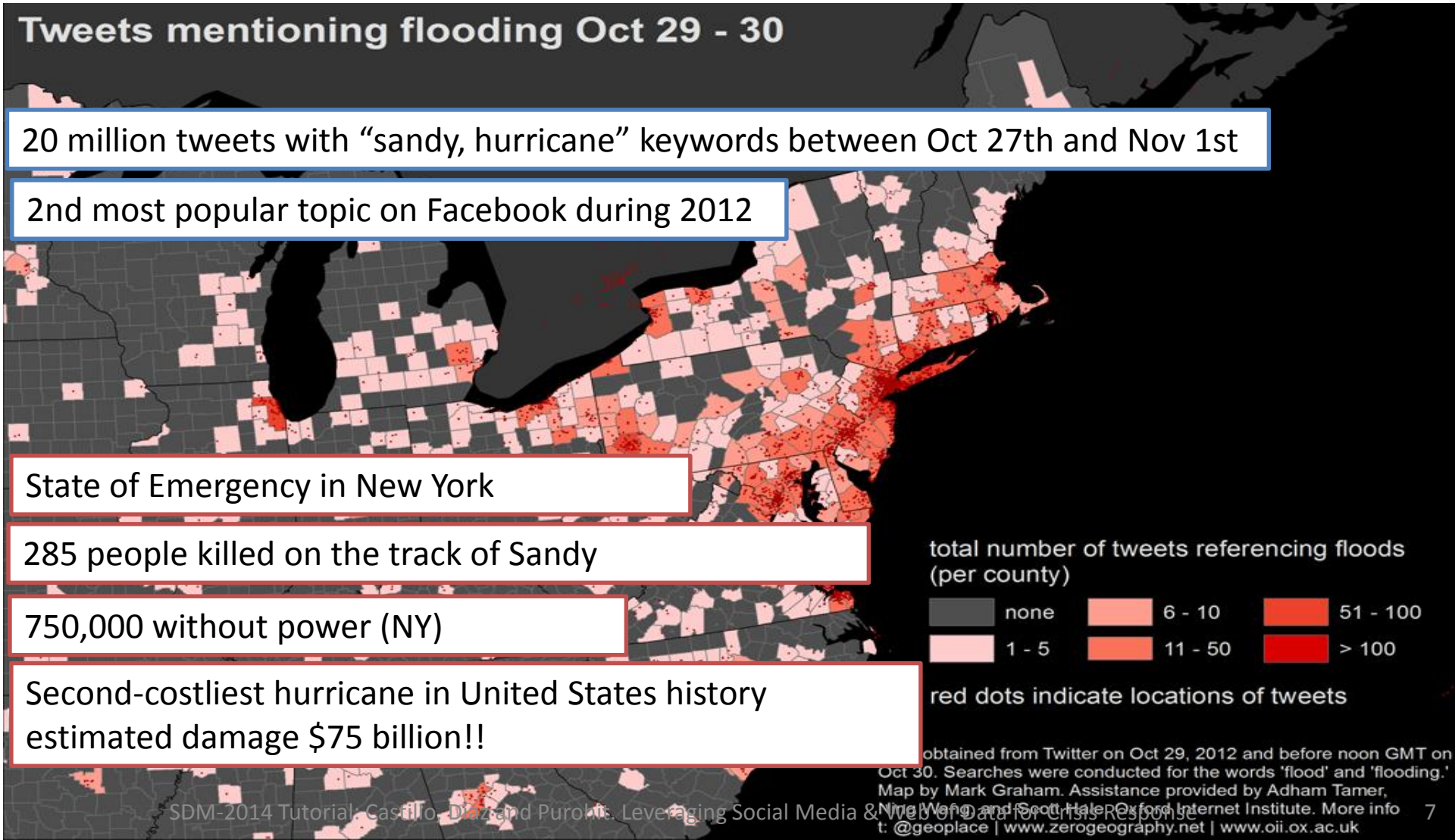
- What we are doing
  - Data Focus:
    - Social media in Emergency Management (SMEM), especially Twitter
    - LOD—Linked Open Data (e.g., GeoNames, Dbpedia)
  - Mining Focus:
    - Exemplary DM problem characterization in crisis domain
    - Application and gaps of existing DM methods in enriching information to support time-critical decisions
- What we are *not* doing
  - Proposing New Algorithms
  - Covering all the problems of Crisis Data Analytics

Predecessor: ICWSM-13 Tutorial

- Purohit, H., Castillo, C., Meier, P., Sheth, A. (2013). Crisis Mapping, Citizen Sensing and Social Media Analytics- Leveraging Citizen Roles for Crisis Response Coordination. In *ICWSM Tutorials.*
  - Extensive Domain knowledge
  - Crisis Mapping and Citizen Sensing
  - Tools and gaps in the state of the art
  - Application of computing methods

- Tutorial details: http://www.knoesis.org/hemant/present/icwsm2013

Motivation: SM Role



20 million tweets with “sandy, hurricane” keywords between Oct 27th and Nov 1st

2nd most popular topic on Facebook during 2012

State of Emergency in New York

285 people killed on the track of Sandy

750,000 without power (NY)

Second-costliest hurricane in United States history
estimated damage $75 billion!!

Motivation: Example

- Short video
  - Dr. Patrick Meier, lead coordinator for digital volunteer effort to assist social media data filtering during Yolanda (Philippines) typhoon, 2013

Video link:
http://www.youtube.com/watch?v=Hj-cW-2XUwU

Motivation: Usability

- Social Media is to assist, ***NOT TO REPLACE*** the existing Emergency Response Coordination

- Appreciation of leveraging *Web of Data*
  - Andrej Verity, UNOCHA Info Mgmt officer: During ICCM-2013 keynote speech, showed the value of augmented product for decision making, based on inputs of Digital Humanitarian Network (DHN)– a crisis-map based on 'crowd/volunteer-filtering' of Twitter and Web data for Philippines typhoon, 2013
    - Keynote: https://www.youtube.com/watch?v=yrwrJS4dwQc (see from 19th min.)

Motivation: Usability

- Appreciation by leads

- "Improved Access to Technology Can Save Lives in Emergencies", Red Cross, 2013
- Blog on SMEM by domain expert from FEMA:http://thinkdisaster.com/2014/01/25/q-how-reliable-is-social-media-during-a-disaster-a-very/

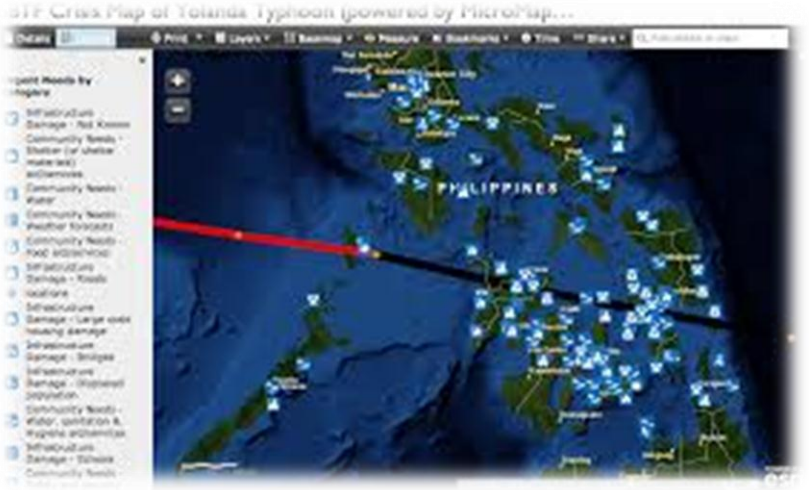
Concepts: Citizen Sensing

- Citizen ‘sensing’ aka observations
  - Report events
  - Express opinions
  - Share experience, etc.

- Characteristics:
  - Faster medium for information sharing
  - Potential for assisting decision making & coordination via reported situation for needs/damage

Concepts: Crowdsourcing & Crisis Mapping

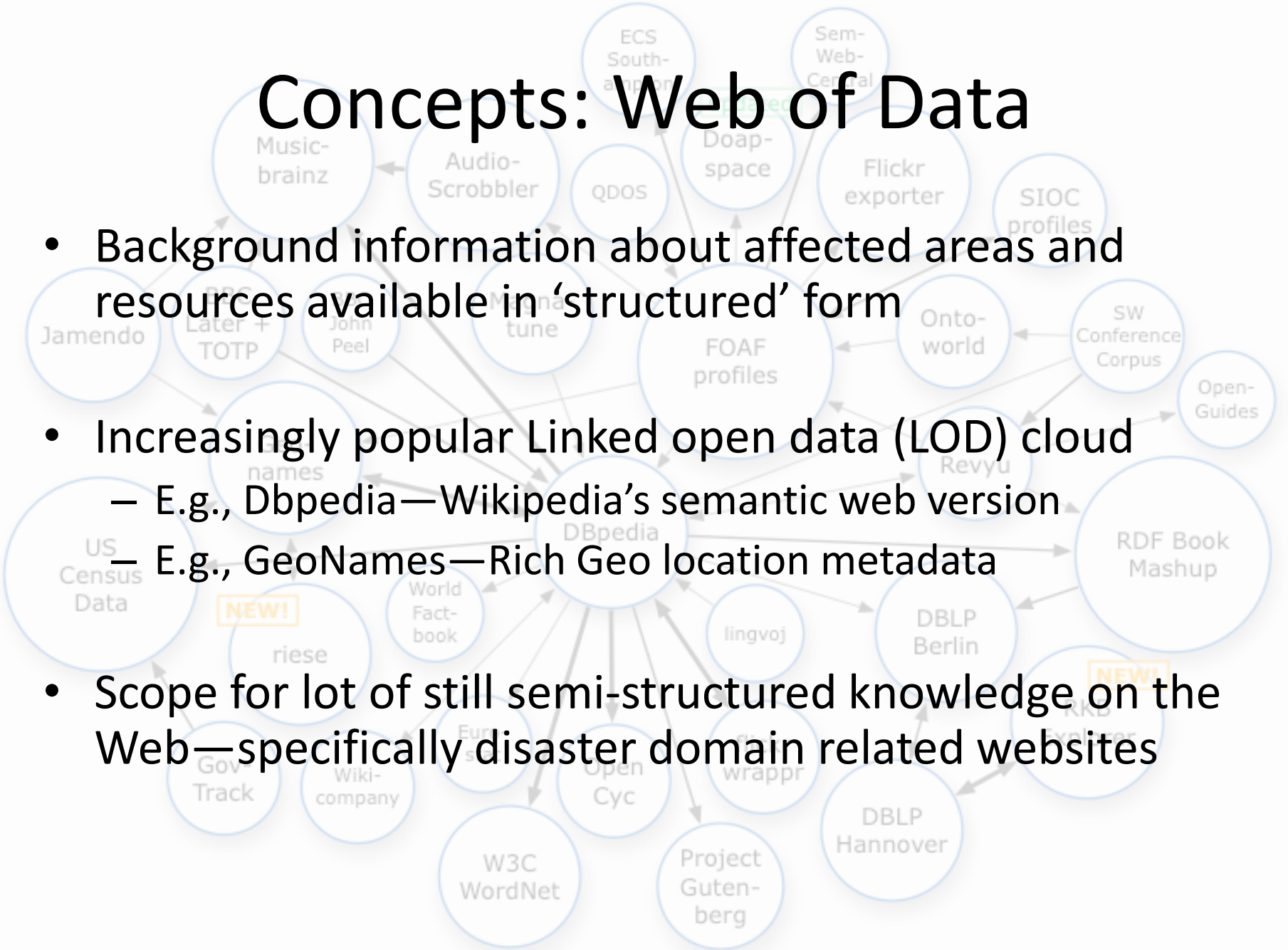
- Geo-locating information on the map
  - Inputs via Crowdsourcing in general



- Characteristics:
  - Too much focused on the 'data creation'
  - Undermined use of data in the higher level analytics

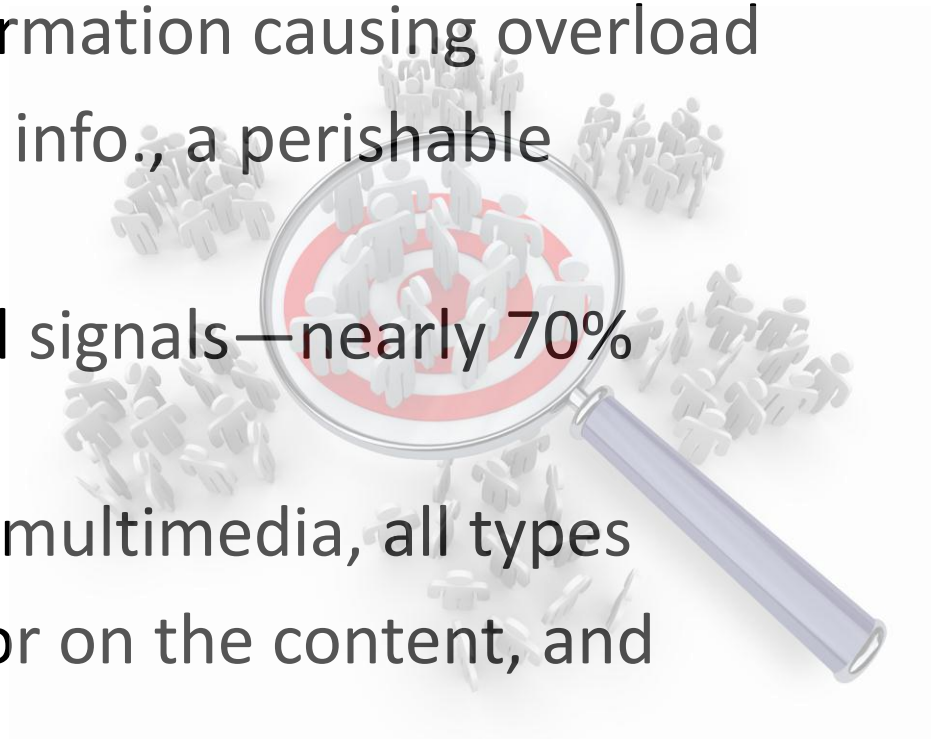
Concepts: Web of Data

- Background information about affected areas and resources available in ‘structured’ form
- Increasingly popular Linked open data (LOD) cloud
  - E.g., Dbpedia—Wikipedia’s semantic web version
  - E.g., GeoNames—Rich Geo location metadata
- Scope for lot of still semi-structured knowledge on the Web—specifically disaster domain related websites



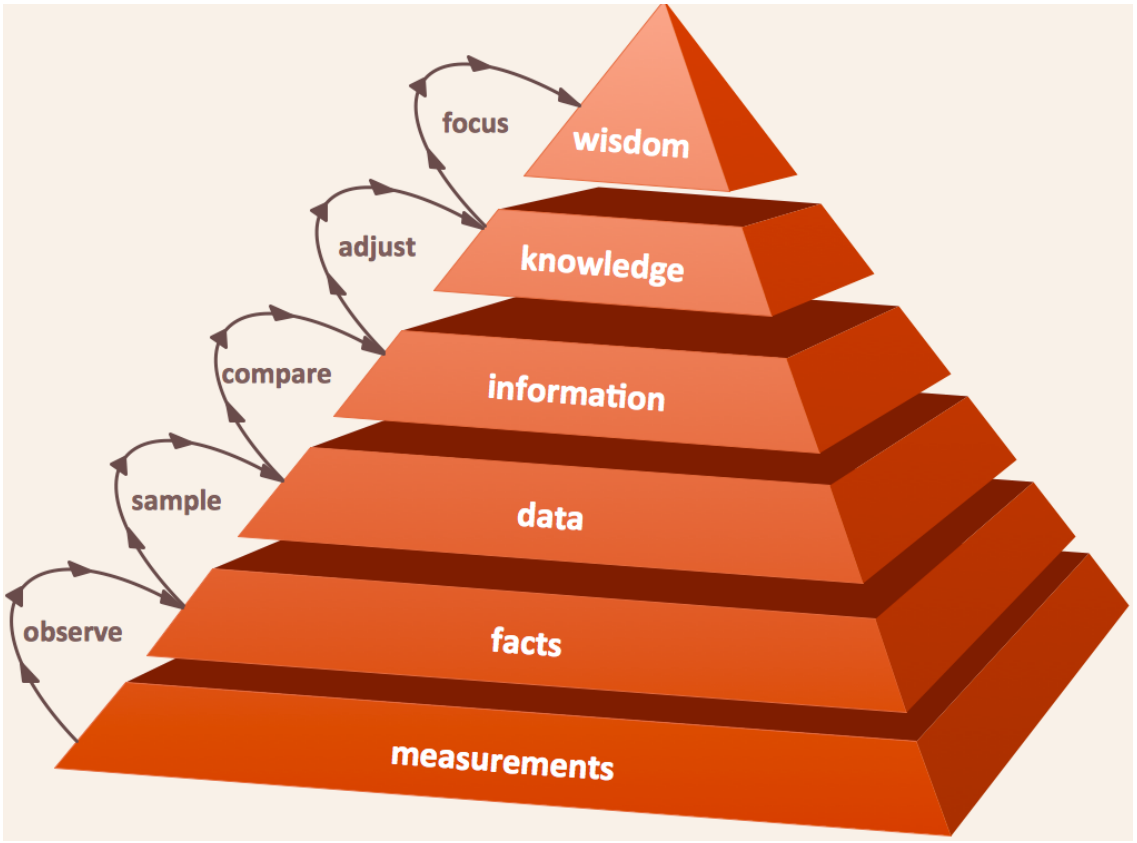
General Challenges: Social Media

- Social media during crisis is "big data"
  - Scale: voluminous information causing overload
  - Velocity: fast incoming info., a perishable commodity
  - Redundancy: amplified signals—nearly 70% repeated
  - Heterogeneity: text to multimedia, all types
  - Verifiability: Trust factor on the content, and sources



General Challenges: DIKW paradigm

- Lack of 'good enough' data quality support for actionability
- Expectation vs. Data & Computing Support



http://www.conceptdraw.com/How-To-Guide/dikw-pyramid

General Challenges: DIKW paradigm

- Data collection and filtering:
  - Bias vs. Coverage problem

- Continuous evolution in the topics causes ‘less’ informative data, eventually ‘less’ effective awareness for actions

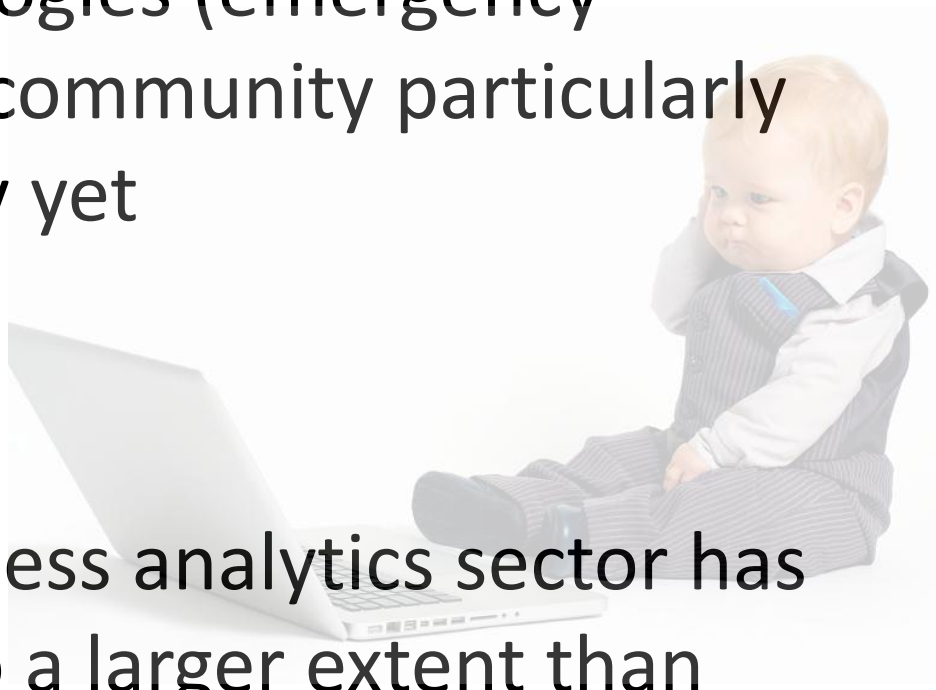
General Challenges: DIKW paradigm

- heterogeneous data aggregation

- For enhanced situational awareness, multiple data sources are available
  - Social media, news, blogs, background knowledge from Wikipedia, existing data sources
  - Concept Normalization a big challenge

General Challenges: Tech Adoption

- Users of these technologies (emergency responders) are not a community particularly oriented to technology yet



- For instance, the business analytics sector has adopted technology to a larger extent than the humanitarian analytics sector!

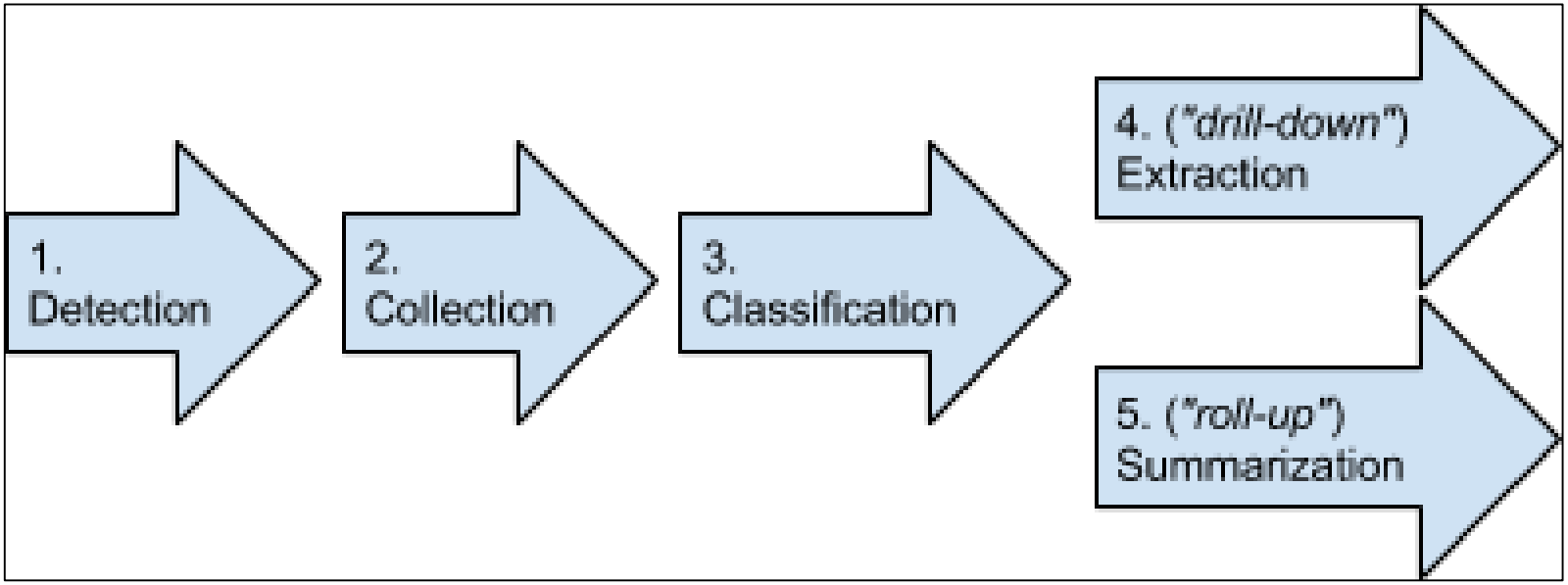
Outline

- A.) Introduction
  - Role of Web 2.0 data during Crisis
  - Data and General Challenges
- B.) Specific Problems, Methods & Future Research
  - Event Detection
  - Data Collection
  - Information Classification
  - Structured Data Extraction from Unstructured
  - Event Summarization
  - Hybrid Systems: Human+Machine Computing
  - Mining for Actions: Coordination and Decision Making
- C.) Conclusion

Our Focus

- For each chosen problem space:
  - Problem Description and Challenges
  - Available Data Characteristics
  - Current Research Methods
  - Potential Future Directions

Process Workflow



Crisis Detection: Problem Definition

- Input
  - set of crisis types
  - stream(s) of data
- Output
  - low latency signal of crisis event onset

Challenges

- Generalizability
  - Language use will be (very) dependent on the type, location, and specifics of a specific crisis event.
- Granularity
  - Crises can be very local and of limited consequence
  - a fire at a store across the street is important to me and others in my neighborhood but not to others
- Latency
  - Systems must make decisions as soon as possible after the event onset.

Available Data Characteristics

- Traditional media
  - advantages
    - clean text
    - (somewhat) reliable sources
    - good coverage of high profile events
  - disadvantages
    - higher latency
    - poor coverage of granular events
- Social media (e.g. Twitter, Facebook)
  - advantages
    - lower latency
    - good coverage of granular events
  - disadvantages
    - noisier text
    - (somewhat) unreliable sources
    - access
- Wikipedia
  - advantages
    - self-regulated content
    - (somewhat) clean text
  - disadvantages
    - (somewhat) higher latency
    - poor coverage of granular events

Available Data Characteristics

- GDELT
  - all news events ever
- Twitter Events Corpus
  - 120 million tweets, with relevance judgments for over 500 events.
- TREC Temporal Summarization
  - crisis events from 2012 aligned with TREC KBA Corpus
- Topic Detection and Tracking
  - news events aligned with standard LDC Corpora

Current Research Methods

- Topic Detection and Tracking
  - unsupervised clustering of a news stream
- Outlier/burst detection
  - detect rise in activity in a stream
- Newsworthy query detection
  - detect news-oriented queries in a query stream.

Current Research Methods (cont.)

- Traditional Media
  - "Updating Users about Time Critical Events", ECIR 2013
- Multimedia social data
  - "Supporting Crisis Management via Sub-event Detection in Social Networks", WETICE 2012
- Text Social Media
  - "Event detection over twitter social media streams", VLDB 2013

Current Research Methods (cont.)

- Detecting earthquake magnitude from Twitter
  - "Earthquake shakes Twitter users: real-time event detection by social sensors", WWW 2010
  - "Twitter earthquake detection: earthquake monitoring in a social world", Annals of Geophysics 2011
- Predicting flu trends from query logs
  - "Infodemiology: tracking flu-related searches on the web for syndromic surveillance", AMIA 2006
  - "Using Internet Searches for Influenza Surveillance", CID 2008
  - "Detecting influenza epidemics using search engine query data", Nature 2009

Potential Future Directions

- fine-grained event
- multiple text streams
- integrating non-text streams

Outline

- A.) Introduction
  - Role of Web 2.0 data during Crisis
  - Data and General Challenges
- B.) Specific Problems, Methods & Future Research
  - Event Detection
  - Data Collection
  - Information Classification
  - Structured Data Extraction from Unstructured
  - Event Summarization
  - Hybrid Systems: Human+Machine Computing
  - Mining for Actions: Coordination and Decision Making
- C.) Conclusion

Data Collection: Problem Definition

- Input
  - crisis description (query)
  - stream(s) of data
- Output
  - items relevant to the crisis description

Available Data Characteristics

- Twitter Events Corpus
  - 120 million tweets, with relevance judgments for over 500 events.
- TREC Temporal Summarization
  - crisis events from 2012 aligned with TREC KBA Corpus
- Topic Detection and Tracking
  - news events aligned with standard LDC Corpora

Current Research Methods

- Information filtering
  - track information related to an arbitrary topic over time (e.g. alerts)
- Topic tracking
  - track information related to a news topic over time
- TREC Knowledge base acceleration
  - track information related to an entity over time
- TREC Microblog
  - track microblog posts related to an arbitrary topic over time

Potential Future Direction

- Current: (Specific to social media)
  - use small set of keywords, hashtags
  - all geotagged posts within an area of interest
- Future:
  - adaptive language models for tracking
  - detecting cross-crisis behavior

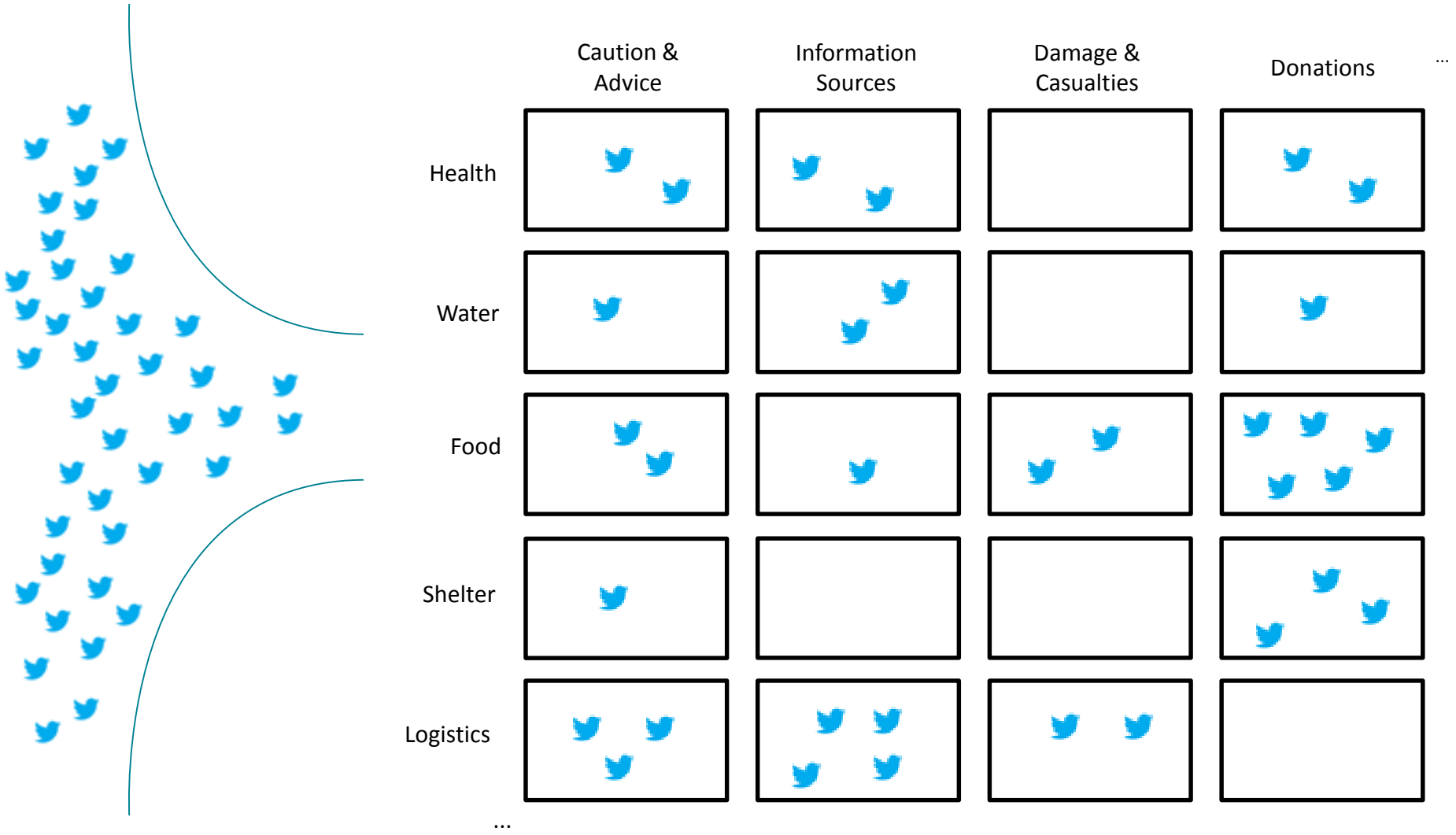
Outline

- A.) Introduction
  - Role of Web 2.0 data during Crisis
  - Data and General Challenges
- B.) Specific Problems, Methods & Future Research
  - Event Detection
  - Data Collection
  - Information Classification
  - Structured Data Extraction from Unstructured
  - Event Summarization
  - Hybrid Systems: Human+Machine Computing
  - Mining for Actions: Coordination and Decision Making
- C.) Conclusion

Information classification for situational awareness

- Problem Description and Challenges
  - Simple classification problem
  - Short text with little context
  - Imbalanced classes
    - Categories that are important but rare

Multiple ways of classifying tweets



Available Data Characteristics: What do people tweet about?

- Sarah Vieweg's PhD Thesis @ UC Boulder
  - People tweet about their social, built, and physical environment
- Social environment
  - **Advice: Information Space**; Animal Management ; Caution; Evacuation; **Fatality**; General Population Information; Injury; Missing; Offer of Help; **Preparation**; Recovery; Report of Crime; Request for Help; Request for Information; Rescue; Response: Community; **Response: Formal**; Response: Miscellaneous; Response: Personal; Sheltering; Status: Community/Population; Status: Personal
- Built environment
  - **Damage**; Status: Infrastructure; Status: Personal Property; Status: Public Property
- Physical environment
  - General Area Information; General Hazard Information; Historical Information; Prediction; **Status: Hazard**; **Weather**

Boldfaced classes were found to be particularly frequent across 4 disasters in her thesis

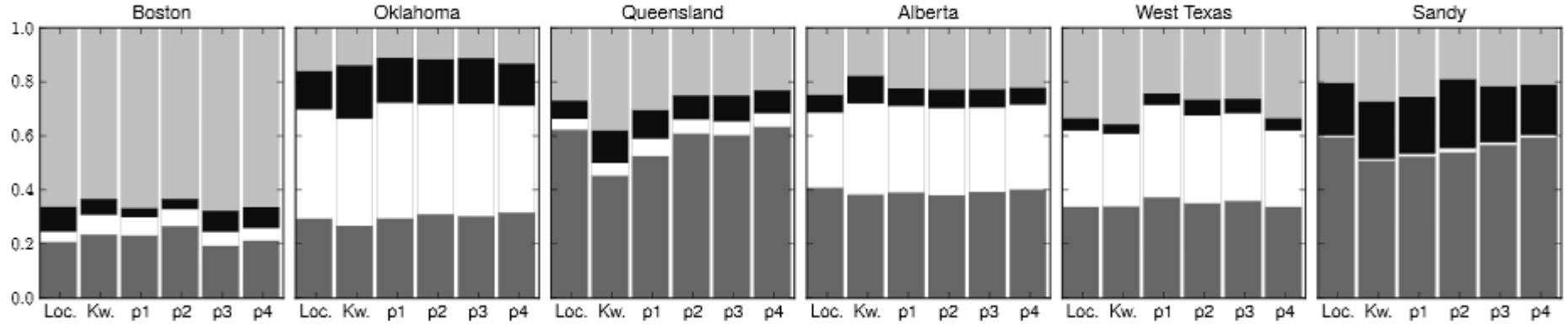
Available Data Characteristics: What do people tweet about?

Example: Joplin Tornado 2011

Caution and advice	50%	A siren heard	12%
		A tornado/thunderstorm warning issued/lifted	42%
		A tornado sighting/touchdown	30%
		Other	16%
Information source	18%	Webpage information source	44%
		Photo information source	16%
		Video information source	20%
		Other	18%
Donation	16%	Money	38%
		Equipment	2%
		Shelter	2%
		Volunteers	2%
		Blood	2%
		Other	54%
Causalities & damage	10%	People dead	44%
		Infrastructure damage	10%
		People injured	2%
		Both people and infrastructure damage	2%
		Not specified but damage	34%
		Unknown	8%
Unknown	6%		

Imran et al. ISCRAM, 2013

Available Data Characteristics: What do people tweet about?

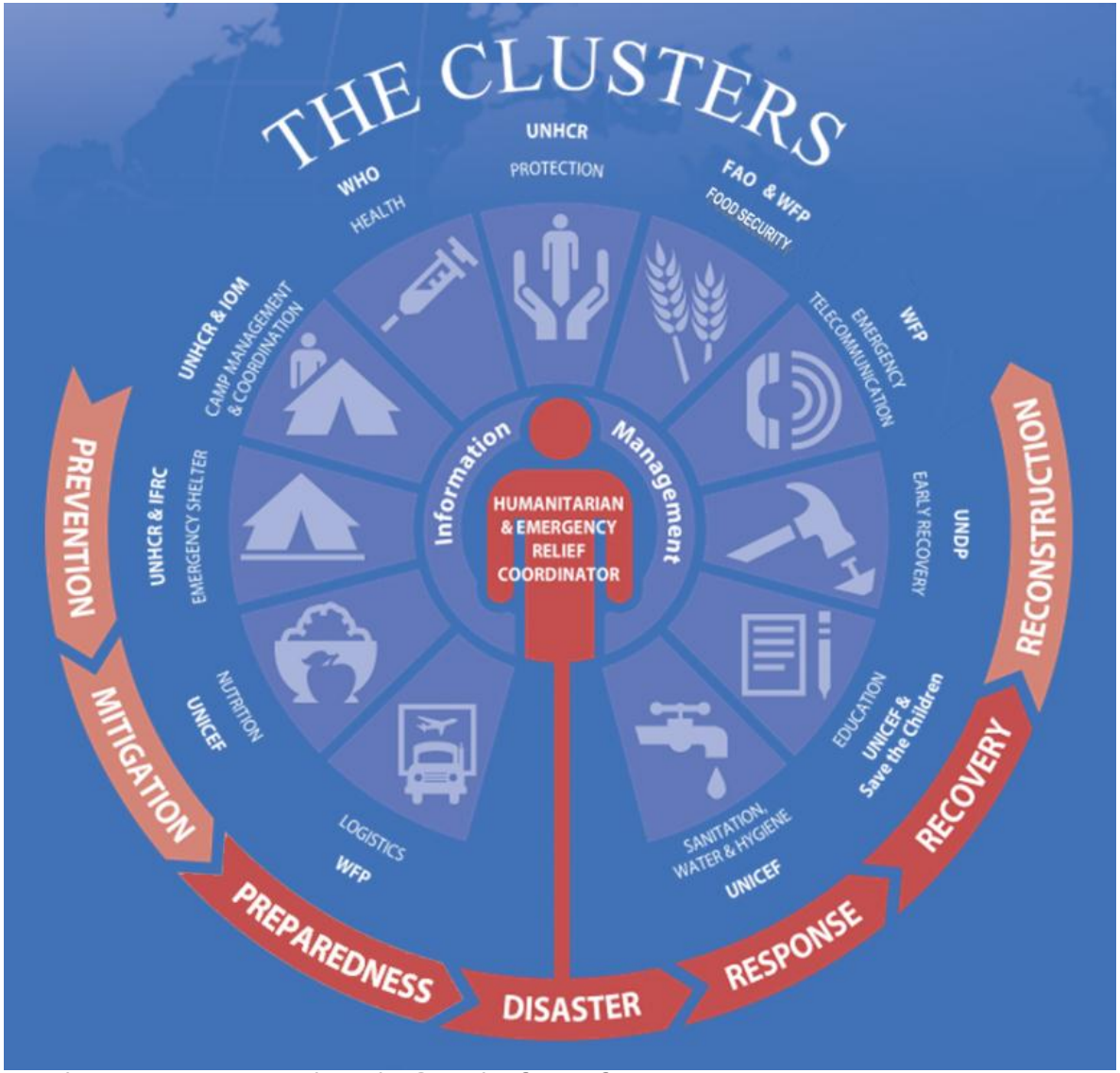


- Prevalence of different categories depends on the disaster and on the data collection method
- Large variety of messages
- Figure from [Olteanu et al. under review]

Current Research Methods: UN Clusters

UN

"Clusters" of agencies



Note e.g.

food security != nutrition,
camp. mgmt != shelter

Current Research Methods: supervised classification

- See e.g. Imran et al. SWDM, 2013 and references therein
- Important finding: model built for one disaster can not be blindly re-used for another
  - Model transfer methods must be used

A key problem when using human labelers for classification

- Taxonomy must strike a compromise between:
  - **Categories labelers can understand**: if the categories cannot be understood by the coders, they will be slower and less consistent
  - **Categories that make sense to agencies**: they must reflect to a large extent how they see the world
  - (When doing supervised learning) **Categories for which an automatic classifier can be reliable**: if differences are too subtle, we may require huge amounts of training data

Potential Future Directions

- Model transfer across disasters
  - Bootstrapping classifiers for a crisis based on previous crises
- Sampling smaller categories
  - Methods for sampling smaller categories for training, e.g. donations is a small categories, donations of blood even smaller: how can we train for that category?
- Conjecture: good classification methods will change the behavior of users
  - E.g. if organizations start to monitor Twitter to create lists of missing people, there will be more users who will report missing people through Twitter

Outline

- A.) Introduction
  - Role of Web 2.0 data during Crisis
  - Data and General Challenges
- B.) Specific Problems, Methods & Future Research
  - Event Detection
  - Data Collection
  - Information Classification
  - Structured Data Extraction from Unstructured
  - Event Summarization
  - Hybrid Systems: Human+Machine Computing
  - Mining for Actions: Coordination and Decision Making
- C.) Conclusion

Extraction of structured data from unstructured text

- Problem Description: similar to textual feature extraction problem
  - Create structured records from unstructured social media text
  - To leverage semantics of the data, link structured records to existing knowledge, e.g., geo-locations

- Challenges
  - Informal language and short length text presents lack of context for existing NLP based methods

Available Data Characteristics

- Social Media:
  - Informal natural language text
  - Helpful platform features such as #hashtags
  - Partial structured metadata along with short messages improves the context, e.g., sensor observations, embedded URLs descriptions, etc.

- Knowledge-bases:
  - Better structured (Linked Open Data (LOD) datasets), and descriptive factual data (Open Gov Data initiative)
    - E.g., GeoNames for locations, Dbpedia for entities

Current Research Methods

- Feature extraction based on predefined categorical attributes: classification methods
  - e.g., message class: bribery/violence etc. (Ushahidi DSSG Project 2013), nature of message- intention: demand/supply (Purohit et al., 2014); etc.

- Crowd-supported feature creation for predefined categories
  - Tweak-the-Tweet project: Structured syntax for helping identify specific information (location, need, etc.) in the text message (Starbird et al., 2010)
    - E.g., *#haiti #name Altagrace Pierre #need help #loc Delmas 14 House no. 14.*

- Mining semantics: Entity/geo-location spotting in the text to extract as structured facets
  - Using Knowledge-bases in LOD (e.g., DBpedia)
  - Using Named Entity Recognition techniques (e.g., Stanford tagger)
  - More details in survey by (Bontcheva & Rout, 2012)

Potential Future Directions

- Design of agreeable structural feature schema across the datasets: Data Normalization
- Modeling inference of potential structural annotations from the knowledge-bases
  - e.g., 'Staten Island' in a message under event of Hurricane Sandy can imply semantic implications of the message for whole south-west NY region
- Anonymization via privacy preserving extraction
  - e.g., anonymity of phone numbers

Outline

- A.) Introduction
  - Role of Web 2.0 data during Crisis
  - Data and General Challenges
- B.) Specific Problems, Methods & Future Research
  - Event Detection
  - Data Collection
  - Information Classification
  - Structured Data Extraction from Unstructured
  - Event Summarization
  - Hybrid Systems: Human+Machine Computing
  - Mining for Actions: Coordination and Decision Making
- C.) Conclusion

Event Summarization: Problem Definition

- Input
  - stream(s) of data
  - query
- Output
  - relevant, novel, comprehensive, timely updates

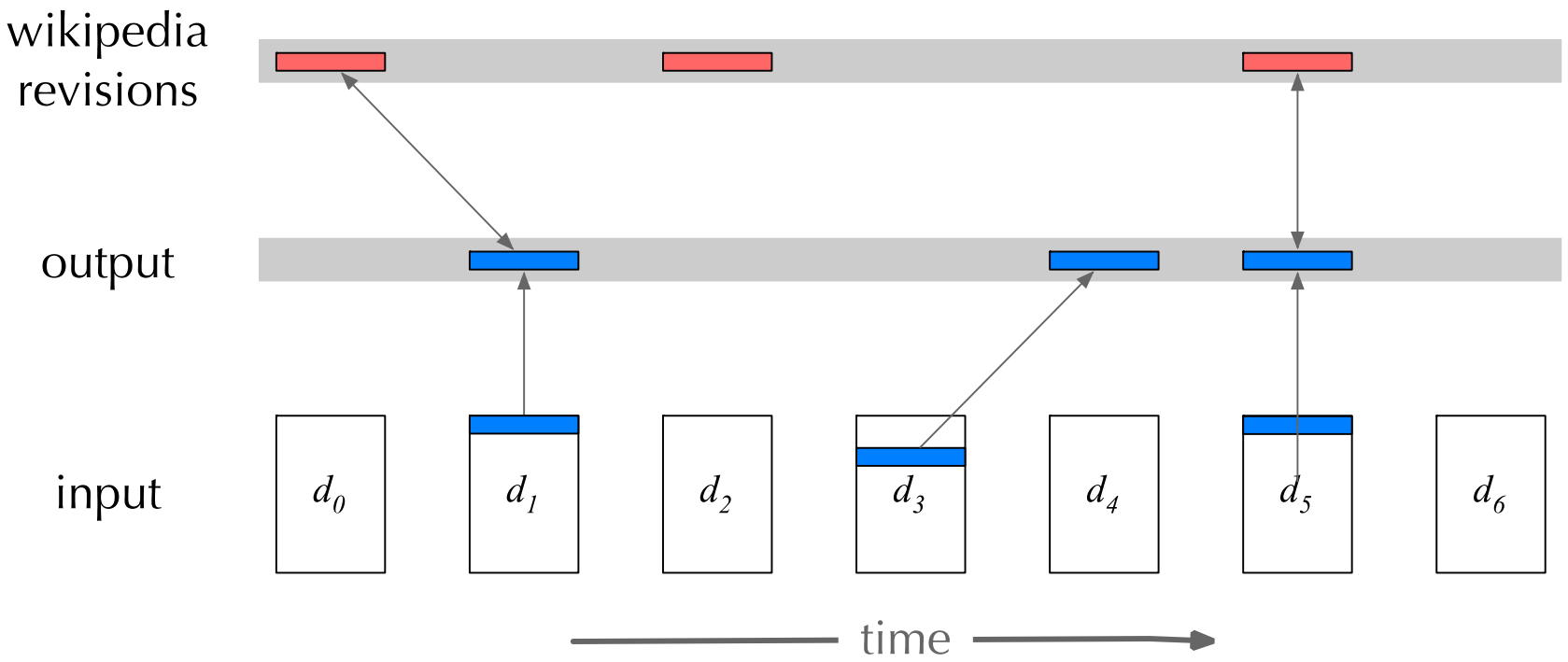
Example: 2011 Tōhoku Earthquake

2:46 PM	Magnitude 8.9 earthquake 231 miles northeast of Tokyo, Japan at a depth of 15.2 miles.
	Quake is fifth largest in the world (since 1900) and the largest quake ever to hit Japan.
3:00 PM	Pacific Tsunami Warning Center issues tsunami warning for the Pacific Ocean from Japan to the U.S. west coast. Tsunami alerts sound in more than 50 countries and territories.
3:30 PM	Wall of water up to 30 feet high washes over the Japanese coast.
7:39 PM	Casualty reports begin to come in. Kyodo News Service reports at least 32 dead.
8:15 PM	Japanese government declares emergency for nuclear power plant near Sendai, 180 miles from Tokyo. Japan has 54 nuclear power plants.
9:35 PM	4 nuclear power plants closest to the quake are shut down.
10:29 PM	Cooling system at Fukushima nuclear report are reported not working: Authorities say they are “bracing for the worst”.

Goal

- to develop algorithms which detect sub-events with **low latency**.
- to develop algorithms which **minimize redundant information** in unexpected news events.
- to model information **reliability** in the presence of a dynamic corpus.
- to understand and address the sensitivity of text summarization algorithms in an online, sequential setting.
- to understand and address the sensitivity of information extraction algorithms in dynamic settings.

Event Summarization: Overview



Available Data Characteristics

- KBA2013
  - July 2012-January 2013
  - web, news, (twitter, facebook)
  - NLP annotations (e.g. segmentation, coref)
  - noisy timestamps (possibly ~1-2 hours late)
  - evaluation on `all sources' and `twitter only'

Available Data Characteristics

- Desired properties
  - timestamped text `nugget'
  - low latency w.r.t. when nugget was known
  - standard method for determining importance
- Approach
  - nuggets semi-automatically derived from Wikipedia revision history.

Measures

- **Precision**: fraction of system updates that match any Gold Standard update.
- **Recall**: fraction of Gold Standard updates that are matches by the system.
- **Novelty**: fraction of system updates which did not match the same Gold Standard update.
- **Timeliness**: difference between the system update time and the matched Gold Standard update time.

Current Research Methods

- **Model**: explicitly predict the relevant and novelty of each sentence as indexed.
  - **Features**
    - **Stationary**: function of the query and the sentence/document content (e.g. sentence position, query match).
    - **Nonstationary**: function of accumulated document/decisions (e.g. LexRank, running similarity to summary).
- **Candidates**
  - **Title**: only consider article title for summary.
  - **Title+Body**: consider title+body.

Current Research Methods: Exemplary Method

TemporalSummarization($\mathbf{S}, \mathcal{C}, q, t_s, t_e$)

```
     S          ▷ Participant system.
     C          ▷ Time-ordered corpus.
     q          ▷ Event keyword query.
     t_s        ▷ Event start time.
     t_e        ▷ Event end time.
 1   U ← {}
 2   S.Initialize(q)
 3   for d ∈ C
 4         do
 5            S.Process(d)
 6            t ← d.Time()
 7            if t ∈ [t_s, t_e]
 8               then
 9                    U_t ← S.Decide()
10                    for u ∈ U_t
11                          do
12                             U.Append(u, t)
13   return U
```

Current Research Methods: Exemplary Results

(a) Sentences selected from the first 10 sentences of the document

features	$\mathrm{E}_s[\mathrm{P}]$	$\mathrm{E}[\delta\mathrm{P}]$	$\mathrm{P}(\tilde{\mathcal{S}})$	$\mathrm{E}_s[\mathrm{R}]$	$\mathrm{E}[\delta\mathrm{R}]$	$\mathrm{R}(\tilde{\mathcal{S}})$	$\mathrm{AUC}_{\mathrm{R}(\tilde{\mathcal{S}})}$
stat.	$0.4468^{n,a}$	0.2144^{a}	0.2156^{a}	0.0101^{n}	0.0041^{a}	0.2894^{a}	$0.2532^{n,a}$
non-stat.	0.5282^{s}	0.2855^{a}	0.2814^{a}	$\mathbf{0.0163}^{s}$	0.0056^{a}	0.2846^{a}	$0.2521^{s,a}$
all	$\mathbf{0.5548}^{s}$	$\mathbf{0.4136}^{s,n}$	$\mathbf{0.4129}^{s,n}$	0.0133	$\mathbf{0.0128}^{s,n}$	$\mathbf{0.3496}^{s,n}$	$\mathbf{0.3034}^{s,n}$

(b) Sentences selected from title of the document

features	$\mathrm{E}_s[\mathrm{P}]$	$\mathrm{E}[\delta\mathrm{P}]$	$\mathrm{P}(\tilde{\mathcal{S}})$	$\mathrm{E}_s[\mathrm{R}]$	$\mathrm{E}[\delta\mathrm{R}]$	$\mathrm{R}(\tilde{\mathcal{S}})$	$\mathrm{AUC}_{\mathrm{R}(\tilde{\mathcal{S}})}$
stat.	0.5400^{n}	0.3004^{a}	0.2950^{a}	0.0123^{n}	0.0062^{a}	$\mathbf{0.3233}^{n,a}$	$\mathbf{0.2810}^{n,a}$
non-stat.	$0.4546^{s,a}$	0.2272^{a}	0.2340^{a}	$0.0097^{s,a}$	0.0052^{a}	0.2549^{s}	0.2143^{s}
all	$\mathbf{0.5459}^{n}$	$\mathbf{0.4097}^{s,n}$	$\mathbf{0.4067}^{s,n}$	$\mathbf{0.0132}^{n}$	$\mathbf{0.0102}^{s,n}$	0.2772^{s}	0.2425^{s}

Potential Future Work

- Personalized summaries
  - e.g. `sandy updates near NYC' vs `sandy updates NJ'
- Topical summaries
  - e.g. `infrastructure damage related to sandy'
- Synthesizing multiple streams

Outline

- A.) Introduction
  - Role of Web 2.0 data during Crisis
  - Data and General Challenges
- B.) Specific Problems, Methods & Future Research
  - Event Detection
  - Data Collection
  - Information Classification
  - Structured Data Extraction from Unstructured
  - Event Summarization
  - Hybrid Systems: Human+Machine Computing
  - Mining for Actions: Coordination and Decision Making
- C.) Conclusion

Hybrid human-automatic mining during crises: context

- Getting a "crowd" of workers to participate is almost never easy
- Motivating workers is a huge problem
- Except during large-scale, highly-publicized crises!
  - Digital volunteers become rapidly available and "just want to help"
  - There are existing digital volunteering communities, e.g. the Stand-by Task Force
- Great opportunity for us ☺

Problem definition

- Given an application, what is the optimal way of integrating human processing with automatic processing of social media data?
- Optimal in what sense?
  - As an automatic processing system: high throughput, low latency, high load adaptability (response to load changes), etc.
  - As a system involving crowdsourcing: high quality, low cost, high engagement of users, etc.

Challenges

- Too much data for the volunteers alone
  - E.g. max. rate in Hurricane Sandy 2012 was ~200 tweets/second
- Problem typically too difficult for automatic systems alone
  - Usage of linguistic pragmatics: people assume a shared context which is not within reach of computers

Challenges (cont.)

- High variety of volunteers
  - Different backgrounds, skills (e.g. languages), commitment (minutes, hours, days), familiarity with the crisis' context, conception of priorities, understanding of the tasks, etc.
  - Opportunities for digital vandalism (a minor issue in practice, but a concern nevertheless)

Current Research Methods

- Central concept: Human co-Processing Unit (HPU)
  - Analogous to a GPU (Graphics co-Processing Unit) Davis et al. CVPRW 2010.
- Key method: crowdsourcing work quality assurance
  - Keep track of worker's performance
  - See e.g. Tutorial by Matthew Lease at SDM 2013.
- Focus: classification tasks
  - Well-defined, easy to set-up/explain, easy to evaluate

Current Research Methods (cont.)

- Franklin et al. SIGMOD 2011 ("CrowdDB")
  - Creates crowdsourcing tasks on-demand
  - Given an information need, generates a *query plan* that minimizes crowdsourcing calls
- Artikis et al. EDBT 2014 (road traffic management)
  - Candidate events (accidents, congestions) generated through physical sensors
  - Generate and send questions to citizens when unsure

Crowdsourced Stream Processing

- Composition of automatic and manual processing elements
- There are typical design patterns that appear in many applications [Imran et al. under review] e.g.:
  - Quality assurance loops: human processing elements do the work, automatic processing elements check for consistency
  - Process-verify: work is done automatically, humans check low-confidence or borderline cases
  - Online supervised learning: humans train the machine to do the work automatically

Crowdsourced Stream Processing (cont.)

- Key design question (application-dependent):
  - Should humans review every element, or should the system be able to run sometimes without humans
- Example CSP in the humanitarian domain
  - http://aidr.qcri.org/

Note about volunteers

- The fact that they want to work for free doesn't mean they will continue to work no matter what
- Context and informed consent of volunteers
  - Worker shouldn't be left wondering "What am I doing here? For whom am I working?"
- Task has to be engaging, and all the design principles of crowdsourcing tasks apply!
  - Task has clear instructions, it is simple and fast, we give feedback to the workers and receive feedback from them, we are **present** to manage the process as it unfolds.

Potential future directions

- Going beyond classification into higher-level tasks: extraction, summarization, etc.
- Reduce task set-up time, e.g. allow crowdsourcing workers collaborate to create the coding manual on their own

Outline

- A.) Introduction
  - Role of Web 2.0 data during Crisis
  - Data and General Challenges
- B.) Specific Problems, Methods & Future Research
  - Event Detection
  - Data Collection
  - Information Classification
  - Structured Data Extraction from Unstructured
  - Event Summarization
  - Hybrid Systems: Human+Machine Computing
  - Mining for Actions: Coordination and Decision Making
- C.) Conclusion

Mining for Actions: Coordination and Decision Making

- Challenges
  - Response coordination actions are complex, involving human intelligence for decision making
    - e.g., optimized response and resource allocation
  - Instead of creating fully automatic systems, DM can '*assist*' coordination teams
    - challenges for ground truth design & evaluation
  - *Assist* for identifying information of key functions to support actions
    - e.g., extract demand-supply of resource needs, data abstraction for damage assessment

Problems

- Intent Mining for demand-supply extraction
  - Demand-supply extraction and match (Purohit et al., 2014); Problem-aid pair identification and match (Varga et al., 2013)
- Intent Matching for demand-supply of resources
  - Likewise Stable Roommate problem (Irving, 1985)
  - Beyond Question-Answering system setting (Bian et al., 2008)
  - Similar to relevance and ranking in dating systems (Diaz et al., 2010)
- Information Aggregation and Abstraction
  - Bridging bottom-up and top-down views
  - Computing High level indicators from low level signals (text, multimedia, sensors)- e.g., semantic perception (Henson et al., 2013), abstraction (Saitta and Zucker, 2013)
    - E.g., Damage Assessment

Available Data Characteristics

- Social media driven unstructured text
  - Transformation to semi-structured form requires first identifying relevant attributes and their extraction
- Highly noisy data: Informal language, slang, jokes, sarcasm, spam, and a very low percentage (below 5%) of demand-supply intentions
  - Imbalance class distribution
  - Ambiguity in expressing intentions confuses classifiers
- Missing data (geo-location, specificity of needs, etc.)
  - Presents challenge for contextualization during intent matching

Focus: Assisting Donation Coordination

- Many people want to donate during disasters, but they are not informed or engaged from the response coordinator side
- Waste occurs due to resources being over- or under-supplied
- ***Goal**: understanding what is needed and what is offered by social media users*

Piles of donated clothes to be managed as a 'second disaster' after Hurricane Sandy
- NPR, Jan 2013
- http://www.npr.org/2013/01/09/168946170/thanks-but-no-thanks-when-post-disaster-donations-overwhelm

Problem: Identifying & Matching Demand with Supply

How to volunteer, donate to Hurricane Sandy: <URL>

If you have clothes to donate to those who are victims of Hurricane Sandy …

Red Cross is urging blood donations to support those affected <URL>

Needs to get something, suggests scarcity:

REQUEST (demand)

I have TONS of cute shoes & purses I want to donate to hurricane victims …

Does anyone know how to donate clothes to hurricane #Sandy victims?

Does anyone know of community service organizations to volunteer to help out?

Offers or wants to give, suggests abundance:

OFFER (supply)

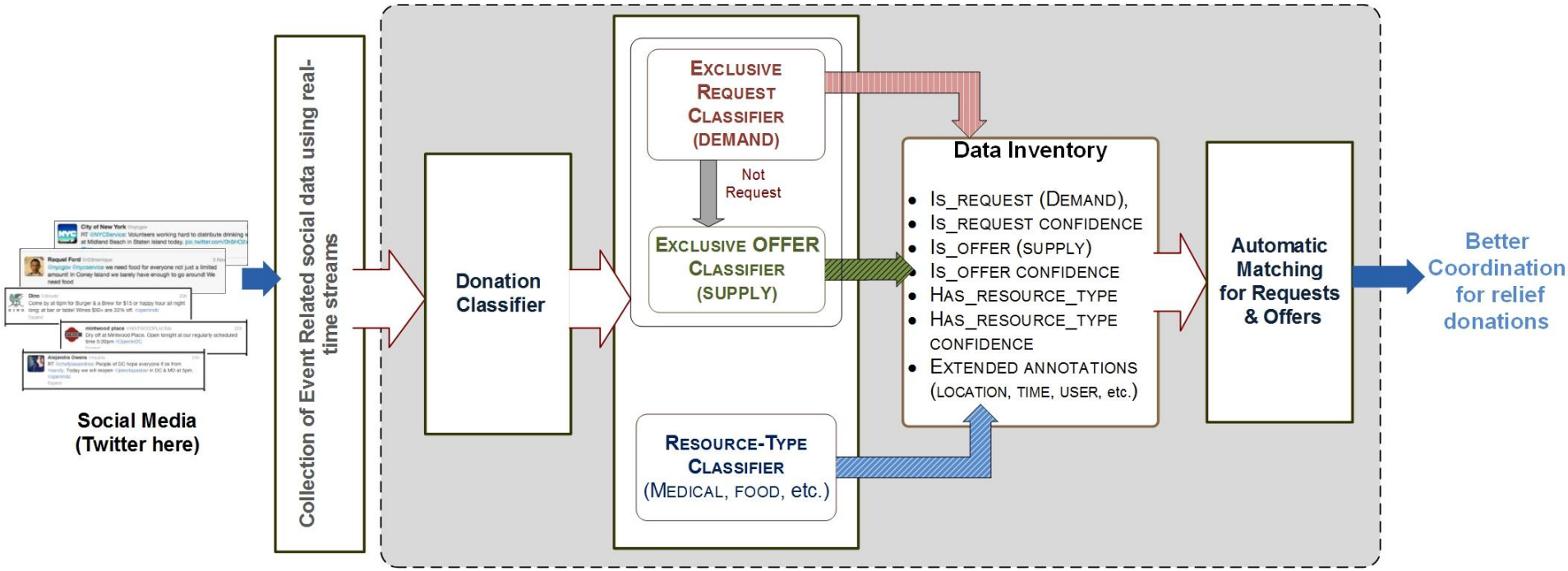
Current Research Methods

- Qualitative: Human review process (high trust factor): e.g., recovers.org, aidMatrix
  - Demand and Supply information collection via registration

- Semi-automatic: a hybrid approach to first collect demand-supply via registration, e.g., CrisisMatch (https://crisismatch.herokuapp.com/)
  - First, seek registration for potential volunteer tasks
  - Second, allocate tasks via optimization algorithm

- Automatic match recommendation
  - Supervised learning method for identifying and matching demand-supply (Purohit et al., 2014)
  - NLP based method for identifying and linking problem-aid messages (Varga et al., 2013)

Illustration: A supervised learning approach



Purohit, H., Castillo, C., Diaz, F., Sheth, A., & Meier, P. (2014). Emergency-relief coordination on social media: Automatically matching resource requests and offers. First Monday, 19(1). doi:10.5210/fm.v19i1.4848

Demand-supply identification and representation: core & facets

Rotary collecting **clothing and other donations** in New Jersey <URL>

- **Core** of the phrase is the "what"
- **Other facets** may include "who", "where", "when", etc.

Corresponding data item in the semi-structured inventory:

{ *source*: "Twitter", *author*: "@NN", *text*: "Rotary collecting clothing and other donations in New Jersey <URL>", *donation-info*: { *donation-type*: "Request", *donation-type-confidence*: 0.8, *donation-organization*: "Rotary", *donation-item*: "clothing and other donations", *donation-location*: "New Jersey" }, … }

Feature types

- Demand/Supply/Resource type Identification:
  - Word N–gram, with standard text mining pre–processing operations
  - Regex-based additional binary features, based on patterns provided by experts
    - E.g., \b(shelter|tent city|warm place|warming center|need a place|cots) \b
- Demand-Supply Matching:
  - Prediction probabilities for demand (request), supply (offer) and resource type in the prior steps
  - Text similarity between vectors of candidate demand-supply pair of messages

Statistics showing Imbalance Distribution of Resource types

Initial number of tweets	4,904,815	100%
Classified as donation-related	214,031	4.4%
Donation-related:		
Classified as exclusively req./off.	23,597 (11%)	100%
Requests (exclusively)	21,380	91%
Offers (exclusively)	2,217	9%
Exclusively requests/offers:		
Classified into a resource type	23,597 (100%)	100%
Money	22,787	97%
Clothing	34	0.1%
Food	72	0.3%
Medical	76	0.3%
Shelter	78	0.3%
Volunteer	550	2%

*Design choice intentionally made as high precision, low recall. Because to really help actions, we need to focus on specific behavior rather than generic. (Purohit et al., 2014)

Some example matches

- Pair 1:
  - Anyone know of volunteer opportunities for hurricane Sandy? Would like to try and help in anyway possible (OFFER)
  - RT @Gothamist: How To Volunteer, Donate To Help Hurricane Sandy Victims http://t.co/fXUOnzJe (REQUEST)
- Pair 2:
  - I want to send some clothes for hurricane relief (OFFER)
  - Me and @CeceVancePR are coordinating a clothing/food drive for families affected by Hurricane Sandy. If you would like to donate, DM us. (REQUEST)

Table 6: Evaluation of matching results, using root mean square error (RMSE) and average precision (AP), where *RMSE is smaller the better and AP is higher the better.* (Purohit et al., 2014)

Method with feature set	Root mean squared error (RMSE)	Average precision (AP)
Text–similarity only (*Baseline*)	0.394 ± 0.08	0.162 ± 0.08
Features except text–similarity, but including request–offer prediction probabilities	0.388 ± 0.08	0.279 ± 0.10
All features	0.383 ± 0.08	0.207 ± 0.09

Potential Future Directions

- Enrich item representation for specificity: capacities of resource needs
  - E.g., #of shelter beds available, #of shelter beds required
  - Using background knowledge of existing resources from Web of Data (e.g., Information about region's shelters from government data—a potential for future Open Data Gov & LOD initiatives)
- Hybrid approach to overcome data sparsity and information verification
  - E.g., Budget of K crowdsourcing calls, which items to annotate?
  - E.g., How much trust in the provided source of demand?
- Use of geographical vs. informational context in matching
  - E.g., distance for volunteering, methods for online donations
- Design of a continuous querying system in the real-world
- User vs. group-based demand/supply matches

Outline

- A.) Introduction
  - Role of Web 2.0 data during Crisis
  - Data and General Challenges
- B.) Specific Problems, Methods & Future Research
  - Event Detection
  - Data Collection
  - Information Classification
  - Structured Data Extraction from Unstructured
  - Event Summarization
  - Hybrid Systems: Human+Machine Computing
  - Mining for Actions: Coordination and Decision Making
- C.) Conclusion

Data Sharability and Integration

- There are obstacles towards creating benchmark collections
- "Companies don't have much commercial incentive to analyze their data in ways that won't make them money, and we shouldn't expect them to. But it would be great if we could find a way to open up more of that data to people who will." [Twitter's data grant and the proprietary data conundrum, D. Harris Feb 2014]
- Twitter access policies
  - 180 tweet searches (100 results each) every 15 min.
  - 15 graph queries (followers OR followees of a user) every 15 min.

Data Shareability (cont.)

- Twitter data sharing policies have softened in the last year
  - Currently: 100K tweets or unlimited (tweet-id, author-id) pairs
  - Before: unlimited (tweet-id, author-id) pairs
  - Before that: nothing
- "Re-hydration" (tweet-id to tweet conversion) requires querying Twitter's API: 180/15 min. = 17,280 /day
  - **Very little** considering that large crises have in the order of a few million tweets per day.
- GNIP/Datasift did not provide paid rehydration queries on large datasets (by Nov. 2013)
  - GNIP/Datasift offer queries by time, geo, or keywords.

Data Shareability (cont.)

- Need data? Ask to the groups that are active on this!
  - Amit Sheth in Kno.e.sis, Wright State Univ
  - Leysia Palen in Univ of Colorado, Boulder
  - Ed Fox in Virginia Tech
  - Patrick Meier in QCRI
  - Many others

Data Integration

- Challenge of designing common data schema
  - HXL – currently sponsored by UN OCHA : http://hxl.humanitarianresponse.info/ns/index.html
  - W3C group on bridging gap of data organization-centric and application-centric approaches for ontology design: http://w3.org/community/emergency/

Improvement for DM methods: requirements

- Performs streaming computation
  - Social media is most useful early on a disaster, focus is on obtaining results early => streaming computation is necessary
- Allows to incorporate human computation
  - There are volunteers that can help
- Incorporates methods from NLP/IR to classify/extract/summarize information

Improvement for DM methods: a few current problems

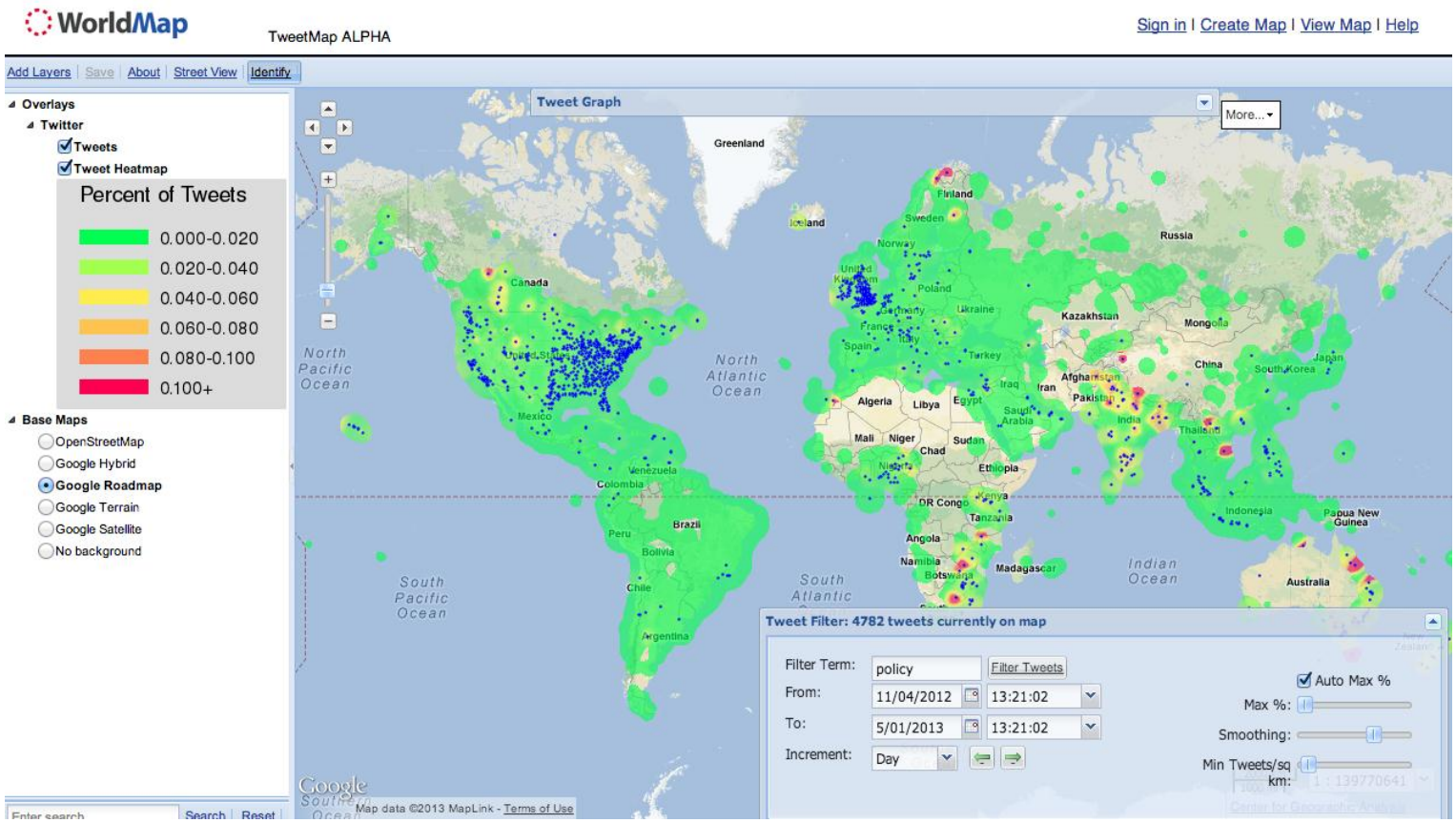
- Capturing a "complete" collection from social media
  - Messages evolve during a crisis
  - E.g., Continuous Semantics approach (Sheth et al., 2010)
- Aggregating data from a variety of sources on the web including news and blogs
  - Important information may be present on web pages but not on the tweets
- Incorporate background knowledge, e.g. geographical information, list of known media sources, etc.

Improvement for DM methods: a few current problems (cont.)

- Rapidly identify knowledgeable (about the crisis) and trustworthy users
  - Specially eye-witness accounts (e.g., On-ground twitterers identification, Starbird et al., 2012)
  - They may not be "influential" in a general sense, and they may not be very active before the crisis
- Time-sensitive methods
  - Create evolving summaries of a crises
  - Time-aware information extraction methods

Improvement for DM methods: a few current problems (cont.)

- Deal with large-scale data, e.g. MapD



Design principles for new tools to help in humanitarian actions

- Identifying target consumers of the designed systems
- Co-design with target consumers, instead of standalone software design based on given requirements
- Conceptualize as socio-technical systems, where humans can assist in computing
- Empirical evaluation through actions, instead of simply visualization of analyses

Principle 1: Explicitly identify target users

- Identifying target consumers of the designed systems.
  - Examples of such consumers
    - what are their backgrounds?
    - where do they work?
    - what is easy for them to do?
    - what is difficult for them to do?

- This may not be a homogeneous groups
  - Identify profiles

Target users: examples

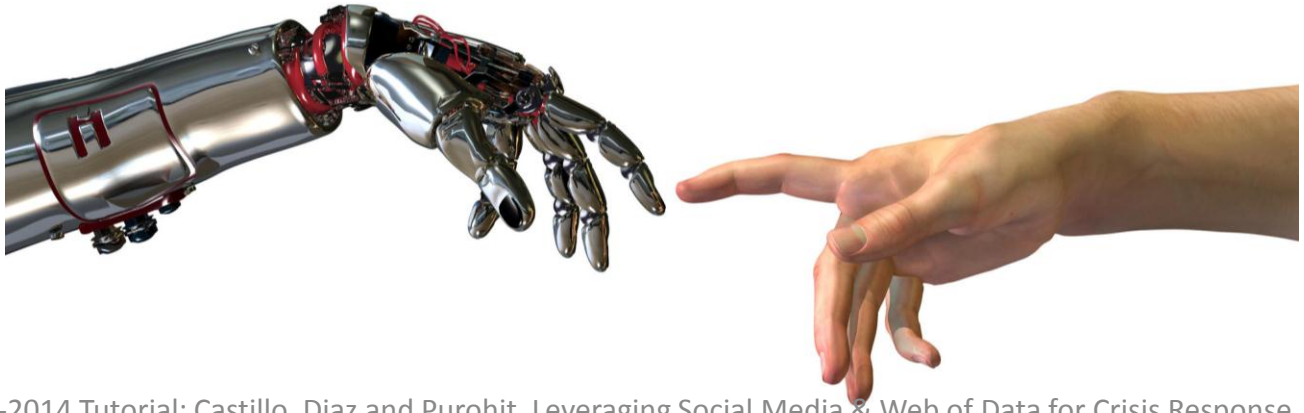
- Headquarters Humanitarians
  - Policy, Information Products, Coordination
- Field Humanitarians
  - Logistics, Relief, Coordination
- Digital Humanitarians
  - Information Collection
  - Analysis

Principle 2: Engage users in co-design

- Do not let humanitarians offload requirements and then leave
- We want them to co-design with us
- This requires effective tools for communication
  - e.g. wireframe designs, user stories, etc.

Principle 3: Socio-technical systems

- Conceptualize the system as hybrid (human and computer intelligence) from the beginning
- Improve response in a continuous fashion
- We want users to be part of the operation of the systems themselves



Principle 4: Empirical evaluation through actions

- We want systems that look good and are easy to use
- We do not evaluate based on looks
- **Are the actions of users better than those of non-users?**

Everybody is needed to join hands!

- Interdisciplinary research is not easy to execute
- But an unidirectional approach will create only more gaps in the research-to-practice pipeline.

Acknowledgements

- Special thanks to our colleagues: Prof. Amit Sheth (Kno.e.sis, WSU) and Dr. Patrick Meier (QCRI)
- NSF for SoCS project grant IIS-1111182: Social Media Enhanced Organizational Sensemaking in Emergency Response at Kno.e.sis, Wright State and Ohio State
  - Profs. Amit Sheth & Srinivasan Parthasarathy (CS, OSU), Valerie Shalin & John Flach (Pyschology, WSU)
- Mohammad Imran at QCRI
- Alexandra Olteanu at EPFL
- Respective image sources
- And the Crisis Computing community!

Thanks and Questions

- Questions, discussion and Feedback:
  - Tweet us: @ChaToX , @fdiaz_msr, @hemant_pt
  - Mail us: chato@acm.org , hemant@knoesis.org, fdiaz@microsoft.com

- Detailed references
  - Tutorial site: http://knoesis.org/hemant/present/sdm2014
  - Also check humanitarian computing bibliography: http://humanitariancomp.referata.com/wiki/Welcome