Machine Learning for Development: Challenges, Opportunities, and a Roadmap

Daniel B. Neill, Ph.D.
Center for Urban Science and Progress
New York University
E-mail: daniel.neill@nyu.edu

This talk is partially based on: M. de Arteaga, W. Herlands, D.B. Neill, and A. Dubrawski, "Machine Learning for the Developing World", *ACM Transactions on Management Information Systems*, 2018.

Machine learning for development

- ML methods have the potential to contribute greatly to human welfare by addressing numerous problems in the developing world.
  - Agriculture, education, governance, poverty reduction, microfinance, human rights, public safety, healthcare, disease surveillance, disaster response, etc…
- Approach 1: Data-driven policy analysis.
  - Analysis of existing data (combining multiple noisy, incomplete sources, deciding what new data to collect)
- Approach 2: Incorporation of ML into deployed information systems to improve public services.
  - For use by local governments, NGOs, etc.
  - Need to be able to identify and respond to emerging events, trends, and patterns on a shorter time scale (disease outbreaks, civil unrest, etc.)

What data might be available?

- **Survey data**- typically historical rather than current, coarse spatial resolution, often incomplete, noisy, and suffering from sampling biases.
- **Satellite imagery**- climate data, wildfires, land use, urbanization, access to electric lighting, migratory or displaced populations.
- **Cell phone** data of various types:
  - Calls and SMS text messages.
  - Location and movement data.
  - “Financial” data: cell-phones for mobile banking (MPESA); cell-phone airtime used as informal currency.
- **Internet data**- penetration much lower; usage patterns different, e.g. public kiosks; low but rapidly increasing smart phone penetration
  - **Social media** content reveals what people want and need, how they spend their time, and how they interact.

In general, richer data enables responsiveness to emerging events and patterns, <u>but</u> such data sources may not be available in **poorer** and more **rural** areas.

ML4D application examples

Broad areas: social, economic, environmental, institutional, health

- Apply standard ML prediction methods
- Goal: Target long-term interventions to those who need them most, to improve conditions and reduce risks

McBride and Nichols (2015) perform **poverty targeting,** predicting which households are living on less than one dollar per day, using observable household characteristics. They show that random forests outperform linear regression on USAID Poverty Assessment Tools data.

Knippenberg, Jensen, and Constas (2018) predict **food insecurity** at the household level using random forests and penalized regression.

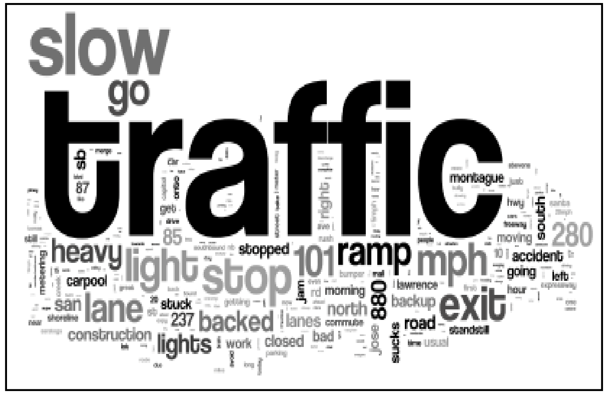
Tien Bui et al. (2012) predict **landslide susceptibility** using decision trees, naïve Bayes, and support vector machine classifiers.

Mwebaze et al. (2010) learn causal Bayesian networks for **famine prediction** in Uganda, while Okori and Obua (2011) use support vector machines, k-nearest neighbors, naïve Bayes, and decision trees.

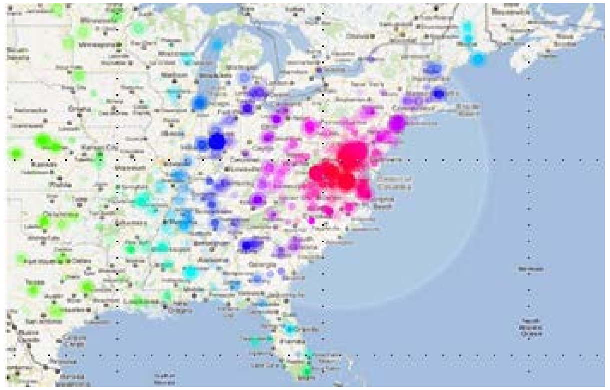
Event Detection and Prediction from Social Media

Protest in Mexico, 7/14/2012

2012 Washington D.C. Traffic



Tweet Map for 2011 VA Earthquake



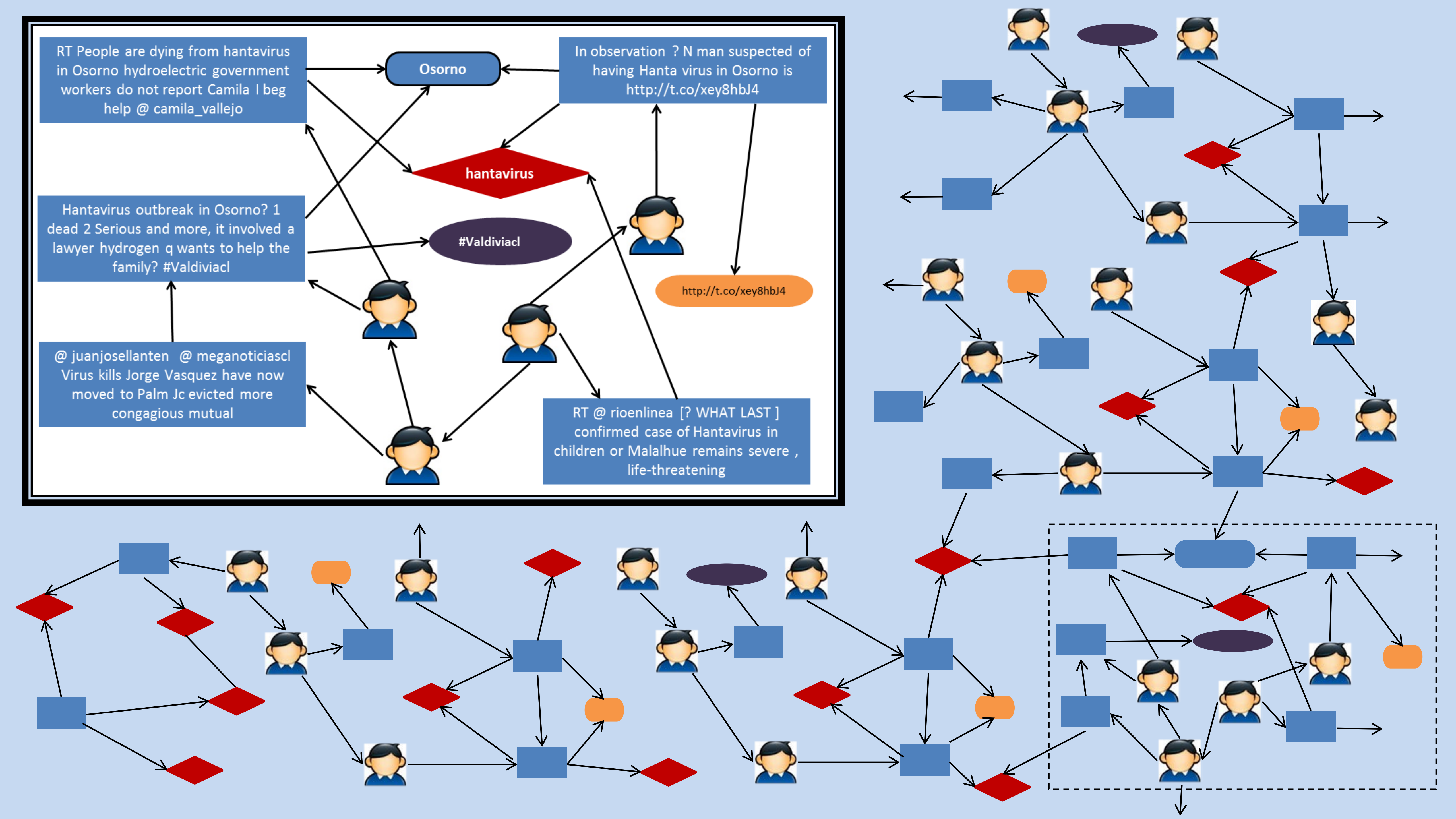
Social media is a real-time “sensor” of large-scale population behavior, and can be used for early detection of emerging events...

… but it is very complex, noisy, and subject to biases.

Chen and Neill (2014) developed a new Twitter event detection approach:

“Non-Parametric Heterogeneous Graph Scan” (NPHGS)

Applied to: civil unrest prediction, rare disease outbreak detection, and early detection of human rights events.


RT People are dying from hantavirus in Osorno hydroelectric government workers do not report Camila I beg help @ camila_vallejo
Osorno
In observation ? N man suspected of having Hanta virus in Osorno is http://t.co/xey8hbJ4
hantavirus
Hantavirus outbreak in Osorno? 1 dead 2 Serious and more, it involved a lawyer hydrogen q wants to help the family? #Valdiviacl
#Valdiviacl
http://t.co/xey8hbJ4
@ juanjosellanten @ meganoticiascl Virus kills Jorge Vasquez have now moved to Palm Jc evicted more congagious mutual
RT @ rioenlinea [? WHAT LAST] confirmed case of Hantavirus in children or Malalhue remains severe , life-threatening

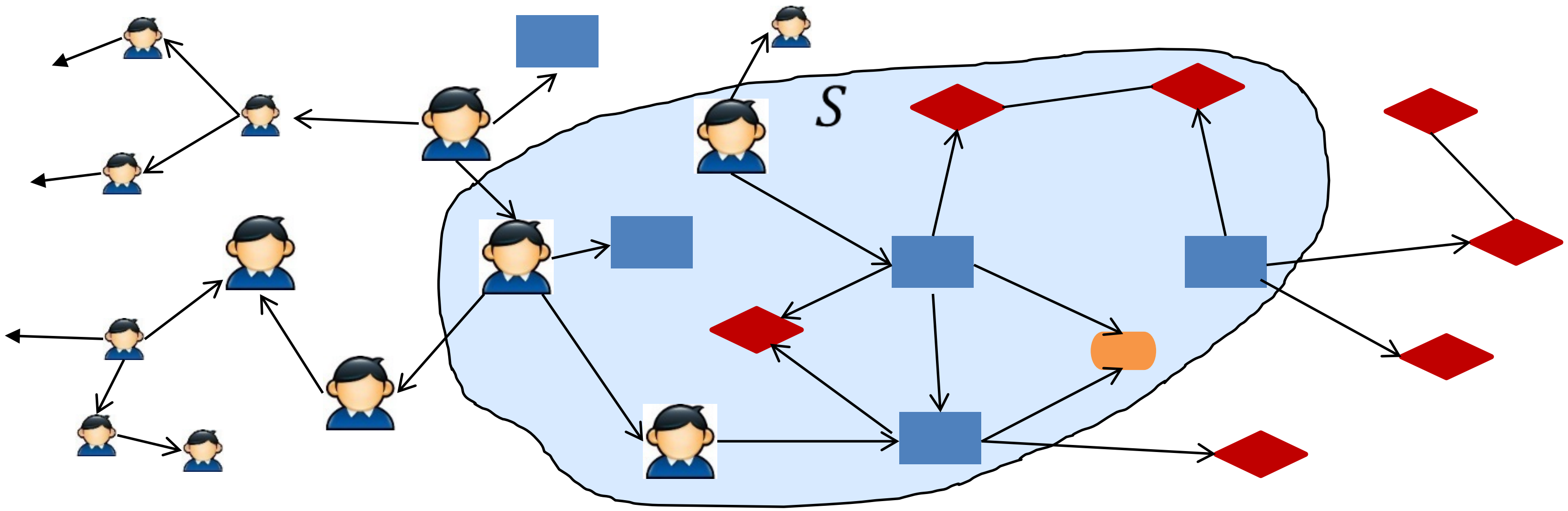
Step 1: Evaluate each node in the Twitter graph

Each node (user, tweet, location, etc.) reports a value measuring how **anomalous** it is for the current hour or day.

Object Type	**Features**
User	# tweets, # retweets, # followers, #followees, #mentioned_by, #replied_by, diffusion graph depth, diffusion graph size
Tweet	Klout, sentiment, replied_by_graph_size, reply_graph_size, retweet_graph_size, retweet_graph_depth
City, State, Country	# tweets, # active users
Term	# tweets
Link	# tweets
Hashtag	# tweets

Features $\xrightarrow[\text{calibration}]{\text{empirical}}$ Individual p-value for each feature $\xrightarrow{\text{min}}$ Minimum empirical p-value for each node $\xrightarrow[\text{calibration}]{\text{empirical}}$ Overall p-value for each node

Step 2: Find the most anomalous subgraphs



$$S^\star = \underset{S \in V:\, S \text{ is connected}}{\operatorname{argmax}} F(S)$$

This step allows us to find groups of nodes (users, keywords, tweets, hashtags, etc.), that are most anomalous when considered collectively.

Using gold standard data from Chile’s Ministry of Health, we demonstrated that NPHGS outperforms existing state-of-the-art methods for detecting emerging outbreaks of **hantavirus**, with respect to timeliness of detection and spatial accuracy.

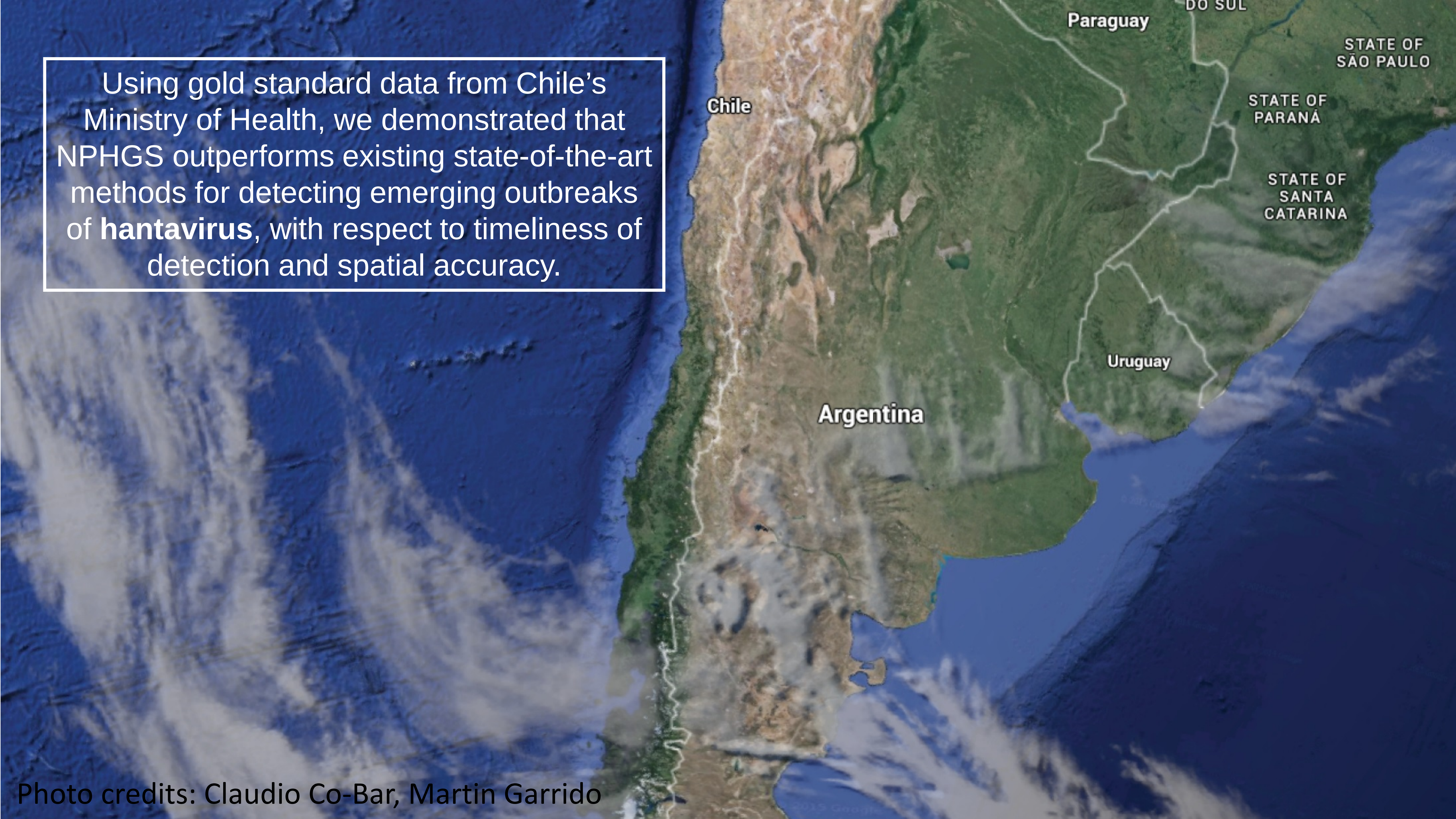


Photo credits: Claudio Co-Bar, Martin Garrido


Detected Hantavirus outbreak, 10 Jan 2013
Locations
Users
Keywords
Hashtags
Links
Videos
Temuco and Villarrica, Chile
First news report: 11 Jan 2013
Chile
Paraguay
DO SUL
STATE OF SÃO PAULO
STATE OF PARANÁ
STATE OF SANTA CATARINA
Uruguay
Photo credits: Claudio Co-Bar, Martin Garrido

"Word cloud" from tweets

Defining ML4D as a field of study

(De Arteaga, Herlands, Neill, and Dubrawski, 2018)

Problem	Solution
1. Applications and data geographically constrained to developing countries. 2. Problem concerns a critical development area for the region of interest.	4. Solution substantially uses ML as an integral element.
3. Problem and/or contextual elements are such that existing or plausible solutions in developed regions are not viable, and proposed solution effectively addresses these differences.	

The specific challenges of development problems (and data) may require us to use existing ML techniques in novel ways or to develop new ML methodology.

Some practical challenges

- Poverty (individual, business, government)- even "inexpensive" solutions may be impossible w/o subsidies.
- Illiteracy, lack of education/training → user interface challenges.
- Diversity of languages → need for machine (or human) translation.
- Lack of power and communication infrastructure: low Internet penetration; frequent outages of electricity and connectivity require specialized solutions; rapidly increasing use of cell phones.
- Migratory/transitory populations; weak transportation infrastructure.
- Corrupt government, misuse of funds, low rule of law.
- Cultural differences, racial/religious/tribal conflicts, mistrust of authorities, outsiders, and top-down solutions.
- Challenges of field research: remote locations, security concerns, need to partner with local governments/NGOs.
- Low amount and quality of collected data → robust analyses needed.
- Challenges of measuring and sustaining impact of new technologies.

Challenge #1: Low data quality

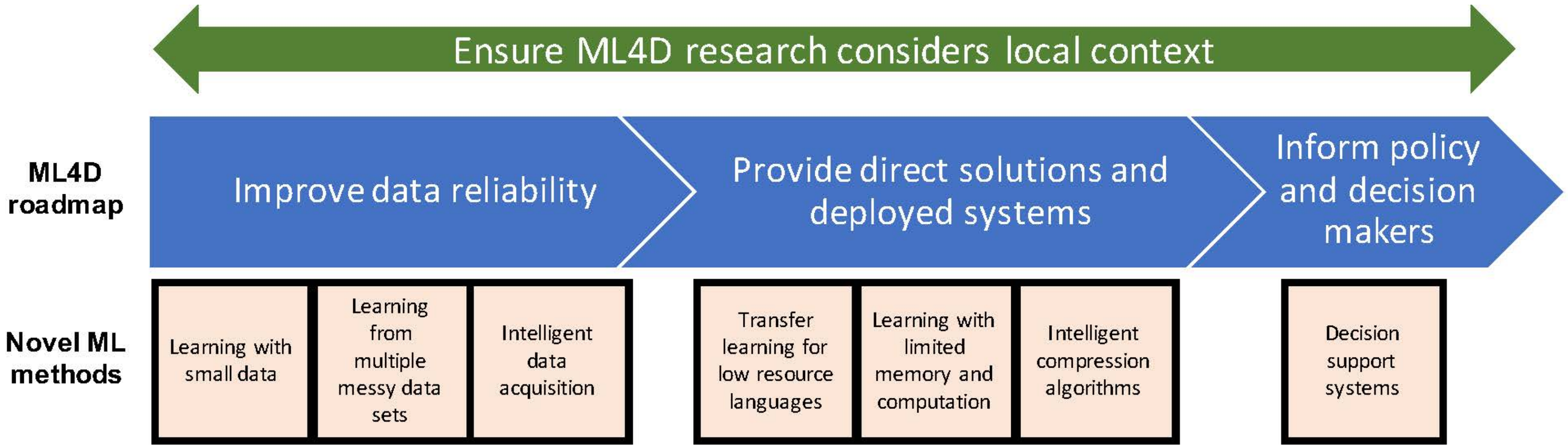
- Missing data (no values for some record-attribute pairs)
  - Data can be missing completely at random (easiest), missing at random, or missing not at random (hardest).
- Noisy data (incorrect/altered values for some record-attribute pairs)
  - Easiest case: i.i.d., Gaussian, additive noise
  - Often not true in practice: dependent errors, anomalous values, noise distribution unknown (must be inferred).
- Systematic biases in data (known? unknown? how to infer?)
  - Convenience sampling, selection bias, reporting bias, false info.
- Different sources report different, often conflicting data
  - Each source has its own limitations/biases.
  - How to integrate information and obtain an accurate “big picture”?
  - Extreme case- crowdsourcing!

Challenge #2: Which data to collect?

- Data collection in the developing world is often difficult and expensive, thanks to logistical challenges and lack of existing infrastructure.
- Available data may be **sparse** or **nonexistent**:
  - Need ML methods that can deal with low quantity of data (e.g. by incorporating models, priors, distributional assumptions)…
  - … and/or clever workarounds (e.g. use of non-traditional data sources which can be more easily collected).
- Typical problem: which data to collect given limited resources and costliness of acquisition?
  - **Active learning** problem: given unlabeled data, which points should we ask an oracle to label? Goal: maximum accuracy with minimum query cost.

A roadmap for ML4D

(De Arteaga, Herlands, Neill, and Dubrawski, 2018)

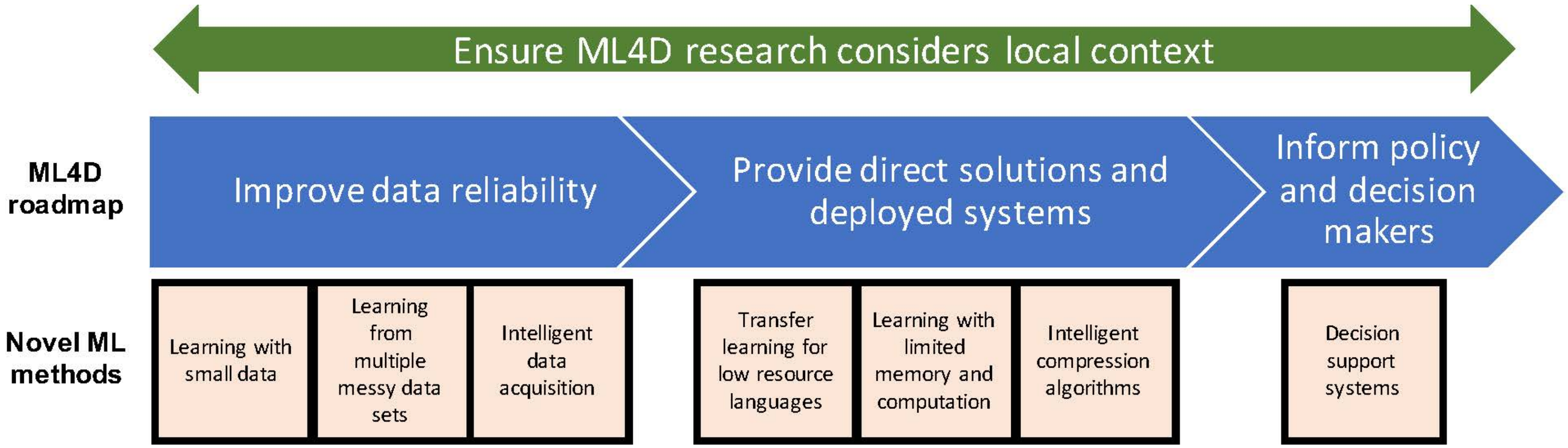


Common features and challenges across many development problems suggest the value of ML solutions, including difficult data, uncertain outcomes, many confounding variables, and costliness of data collection.

“Difficult” data could be biased, incomplete, noisy, or otherwise “messy”; either overly massive or insufficient; unstructured (text, images) or have complex, heterogeneous network structure (e.g., online social media).

A roadmap for ML4D

(De Arteaga, Herlands, Neill, and Dubrawski, 2018)



Common features and challenges across many development problems suggest the value of ML solutions, including difficult data, uncertain outcomes, many confounding variables, and costliness of data collection.

Cleaning (structured) data

Structuring unstructured data

Prioritizing data for collection

Conclusions

Computational and data limitations in the developing world are often lamented as deterrents for use of ML tools in these regions.

These ML4D challenges should instead be considered as inspiring and framing cutting-edge research questions and new ML paradigms.

By considering how ML and development studies can reinforce each other, ML4D researchers have the opportunity to create cutting-edge ML methods while addressing critical issues in the developing world.

Thanks for listening!

More details on my web site:
http://www.cs.nyu.edu/~neill

Or e-mail me at:
daniel.neill@nyu.edu