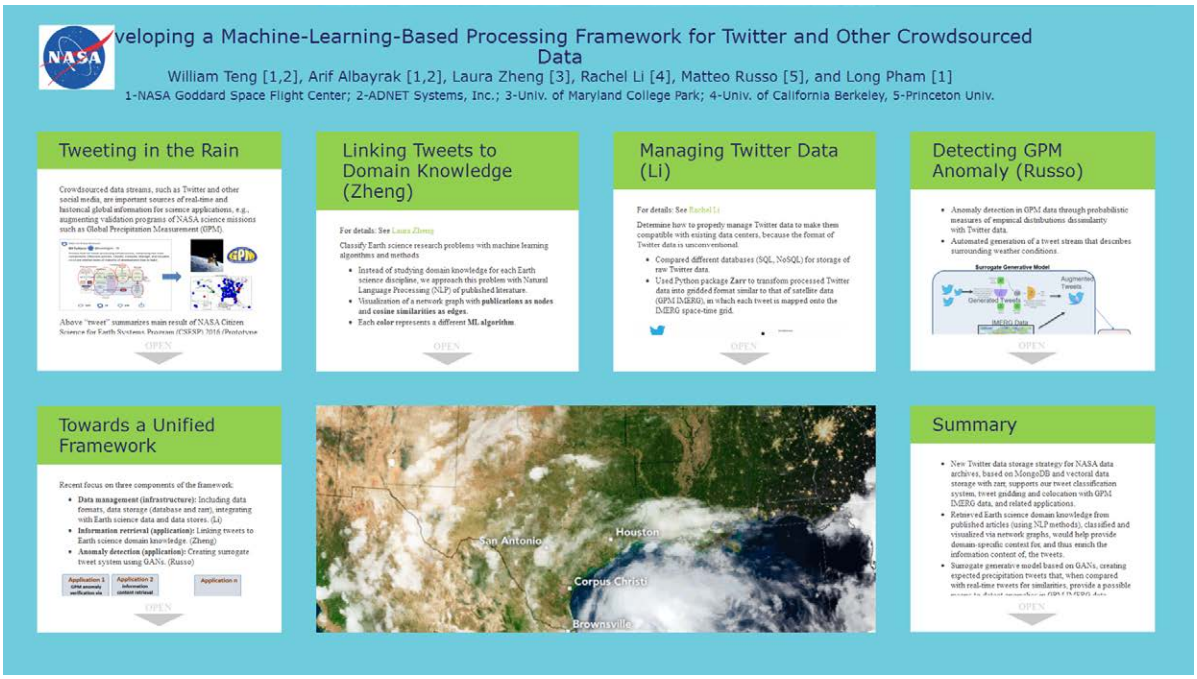
Developing a Machine-Learning-Based Processing Framework for Twitter and Other Crowdsourced Data



William Teng [1,2], Arif Albayrak [1,2], Laura Zheng [3], Rachel Li [4], Matteo Russo [5], and Long Pham [1]

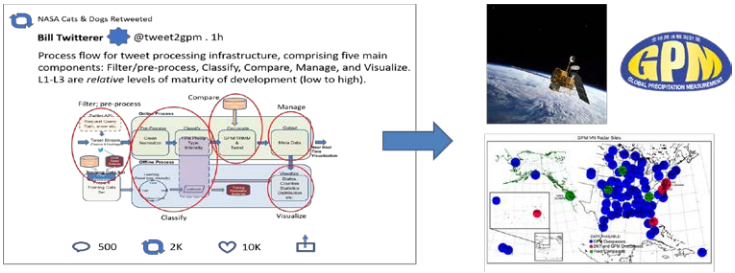
1-NASA Goddard Space Flight Center; 2-ADNET Systems, Inc.; 3-Univ. of Maryland College Park; 4-Univ. of California Berkeley, 5-Princeton Univ.

PRESENTED AT:



TWEETING IN THE RAIN

Crowdsourced data streams, such as Twitter and other social media, are important sources of real-time and historical global information for science applications, e.g., augmenting validation programs of NASA science missions such as Global Precipitation Measurement (GPM).



Above “tweet” summarizes main result of NASA Citizen Science for Earth Systems Program (CSESP) 2016 (Prototype Phase).

Why machine learning (ML)? Why ML-based processing infrastructure?

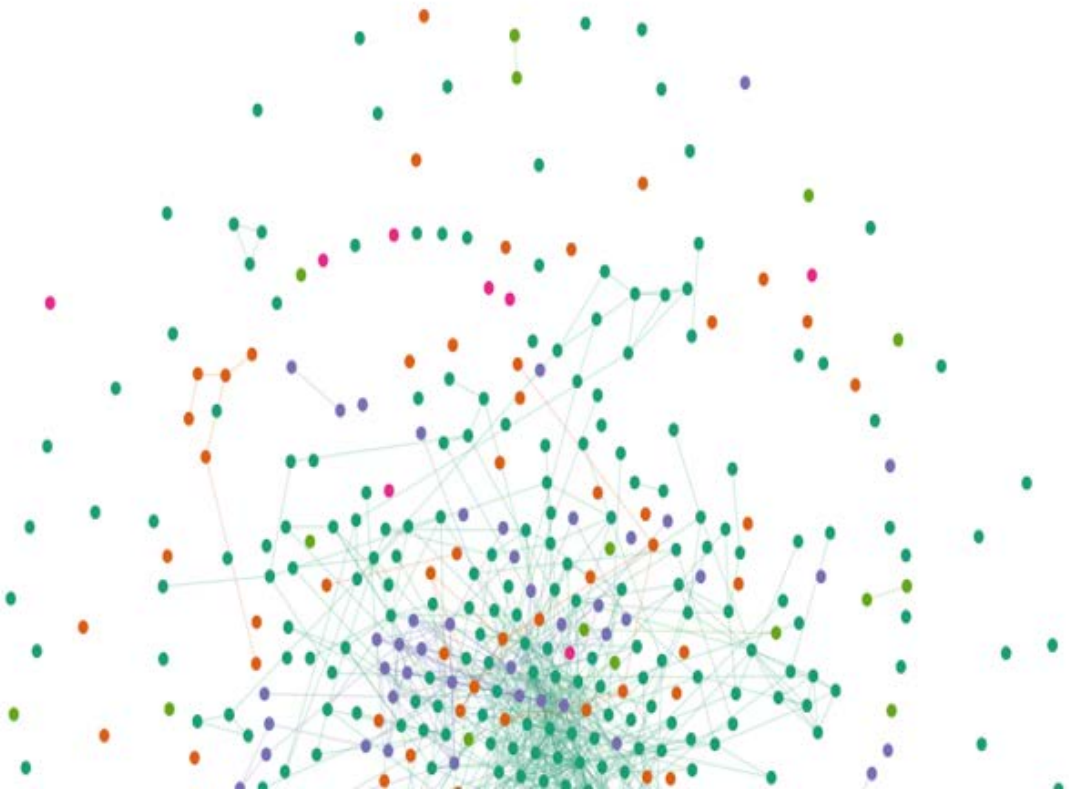
- Velocity – 500 million tweets/day [1]
- Volume – ~2 TB/day
- Value – More than sum of individual tweets
- Variety – Structured/unstructured; text, images, videos, emojis
- Veracity – Active/passive tweets; enriching the Twitter stream

LINKING TWEETS TO DOMAIN KNOWLEDGE (ZHENG)

For details: See Laura Zheng

Classify Earth science research problems with machine learning algorithms and methods

- Instead of studying domain knowledge for each Earth science discipline, we approach this problem with Natural Language Processing (NLP) of published literature.
- Visualization of a network graph with **publications as nodes** and **cosine similarities as edges**.
- Each **color** represents a different **ML algorithm**.



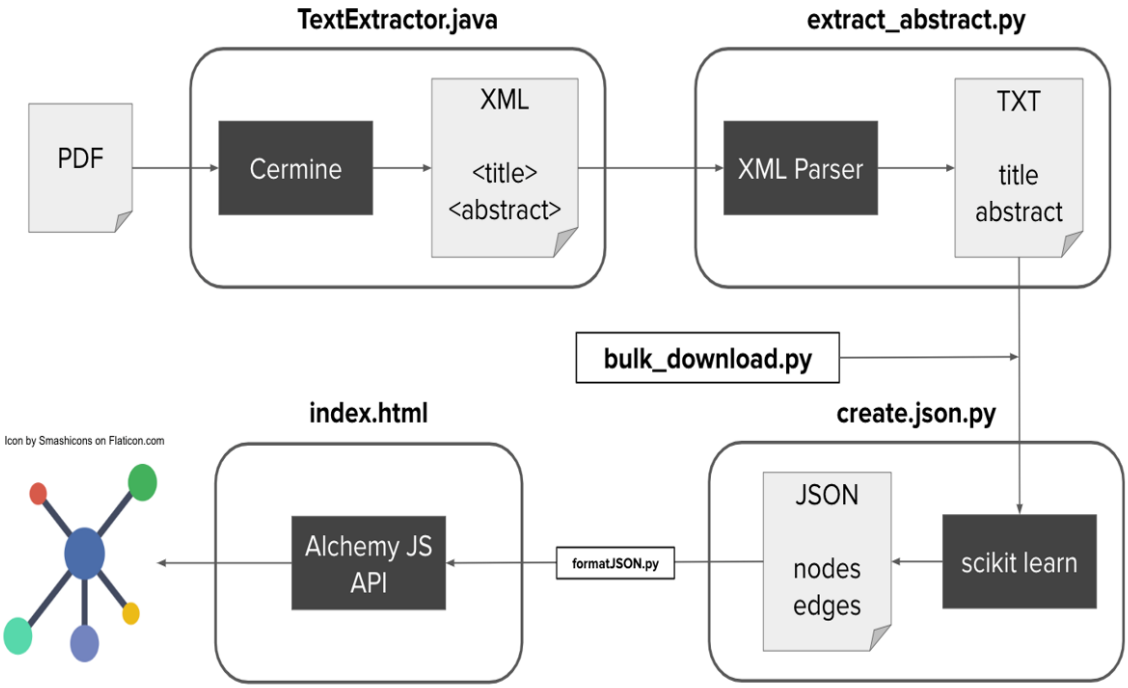
- **Tfidf** represents each publication as a vector.

$$\mathrm{tfidf}(t, d, D) = \mathrm{tf}(t, d) \cdot \mathrm{idf}(t, D)$$

$$f_{t,d} \Big/ \sum_{t' \in d} f_{t',d}$$

$$\log\left(\frac{N}{1 + n_t}\right) + 1$$

- Visualization workflow

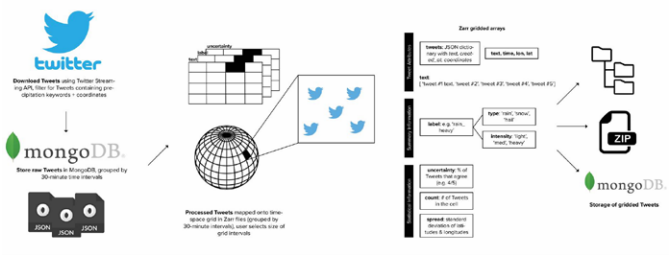

TextExtractor.java
extract_abstract.py
PDF
Cermine
XML
<title>
<abstract>
XML Parser
TXT
title
abstract
bulk_download.py
index.html
create.json.py
Icon by Smashicons on Flaticon.com
Alchemy JS API
formatJSON.py
JSON
nodes
edges
scikit learn

MANAGING TWITTER DATA (LI)

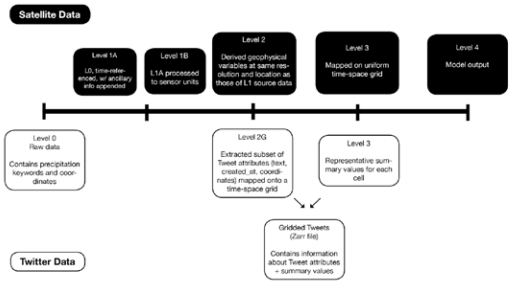
For details: See Rachel Li

Determine how to properly manage Twitter data to make them compatible with existing data centers, because the format of Twitter data is unconventional.

- Compared different databases (SQL, NoSQL) for storage of raw Twitter data.
- Used Python package **Zarr** to transform processed Twitter data into gridded format similar to that of satellite data (GPM IMERG), in which each tweet is mapped onto the IMERG space-time grid.



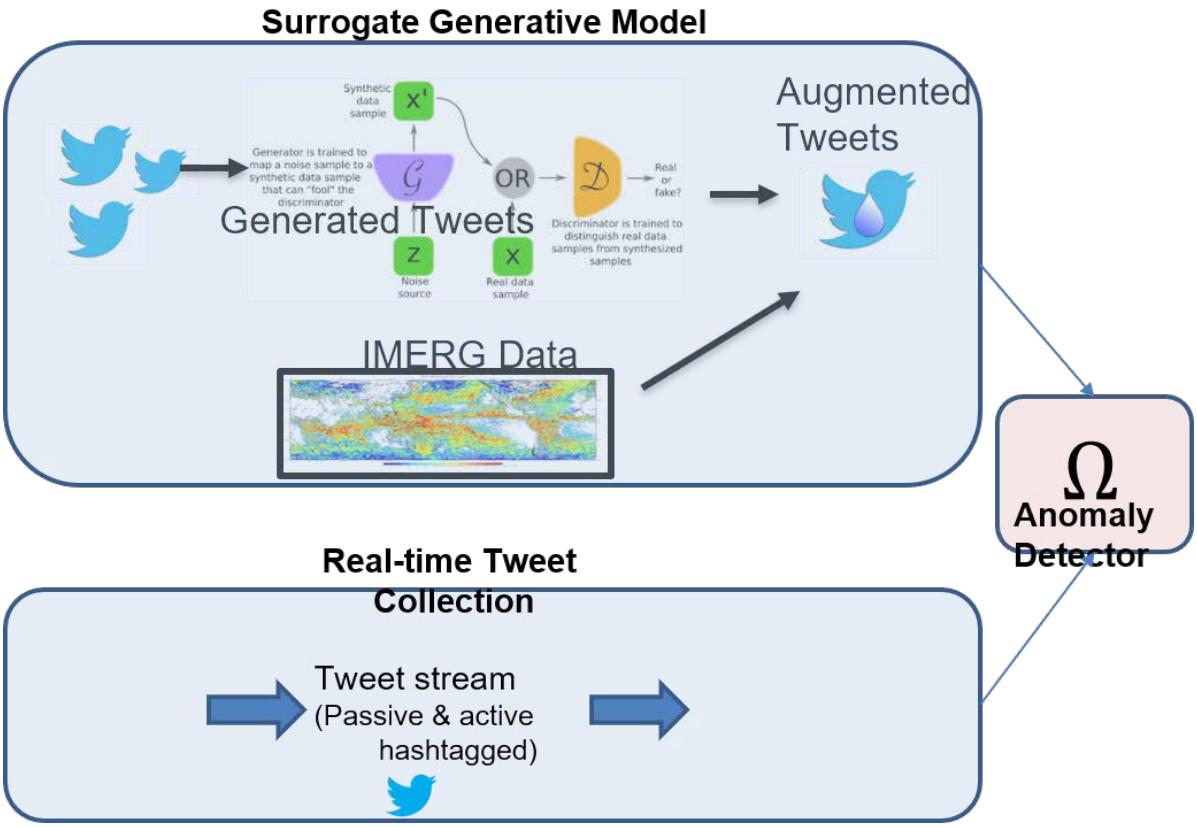
- Selected **MongoDB** for the storage of raw Twitter data due to its flexibility, ability to be accessed by multiple users, and querying functionality
- Developed a pipeline for downloading, storing, and gridding Twitter data.



Notional mapping of Twitter data to equivalent NASA EOSDIS data processing levels for the case of tweets binned to the Global Precipitation Measurement (GPM) Integrated Multi-satellitE Retrievals for GPM (IMERG) grid.

DETECTING GPM ANOMALY (RUSSO)

- Anomaly detection in GPM data through probabilistic measures of empirical distributions dissimilarity with Twitter data.
- Automated generation of a tweet stream that describes surrounding weather conditions.



- Obtain 4 precipitation classes: No, light, medium, and heavy rain for generated and collected tweets.
- Compose gridded distribution of precipitation for generated and collected tweets.
- Measure dissimilarity between the two distributions.

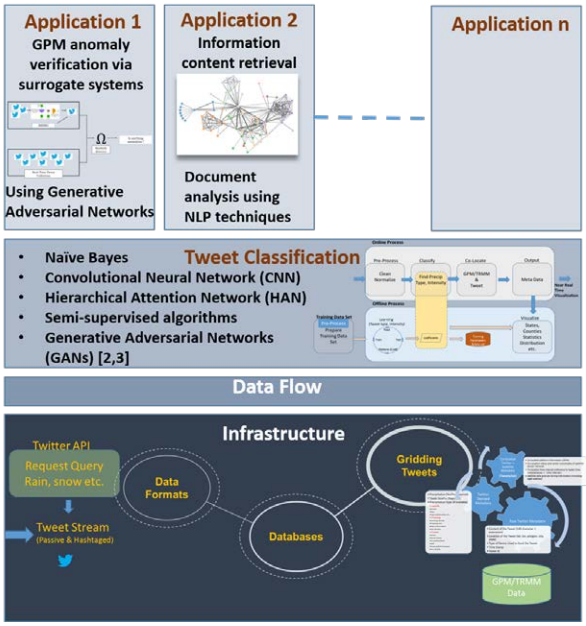
Using these similarity measures, we obtain the following results:

Event/Measure	cos	ρ	JS	Jaccard	W
Typical	0.88	0.89	0.83	0.88	0.91
Extreme	0.67	0.67	0.70	0.74	0.67

TOWARDS A UNIFIED FRAMEWORK

Recent focus on three components of the framework:

- **Data management (infrastructure):** Including data formats, data storage (database and zarr), integrating with Earth science data and data stores. (Li)
- **Information retrieval (application):** Linking tweets to Earth science domain knowledge. (Zheng)
- **Anomaly detection (application):** Creating surrogate tweet system using GANs. (Russo)



SUMMARY

- New Twitter data storage strategy for NASA data archives, based on MongoDB and vectoral data storage with zarr, supports our tweet classification system, tweet gridding and colocation with GPM IMERG data, and related applications.
- Retrieved Earth science domain knowledge from published articles (using NLP methods), classified and visualized via network graphs, would help provide domain-specific context for, and thus enrich the information content of, the tweets.
- Surrogate generative model based on GANs, creating expected precipitation tweets that, when compared with real-time tweets for similarities, provide a possible means to detect anomalies in GPM IMERG data.

References

[1] Twitter statistics

[2] Goodfellow et al., 2014, Generative adversarial nets, in Proc. Neural Information Processing Systems 2014.

[3] Goodfellow et al., 2016, Deep learning, Adaptive computation and machine learning series, Cambridge: MIT Press.

Acknowledgments: This work was supported, in part, by

- NASA GES DISC, for the internships of Rachel Li and Matteo Russo
- ADNET Systems, Inc., for the internship of Laura Zheng

and built on the project,

- “Mining Twitter data to augment NASA GPM validation” funded by the NASA Citizen Science for Earth Systems Program.

Sorry but time is up!

SWITCH TEMPLATE

