# Enriching the Twitter Stream

## Increasing Data Mining Yield and Quality Using Machine Learning

NASA/Goddard EARTH SCIENCES DATA and INFORMATION SERVICES CENTER (GES DISC)

AGU December 2018 NH43B-2988

**Arif Albayrak[1,2], William Teng[1,2], John Corcoran[3], Sky C. Wang[4], Daniel Maksumov[5], Carlee Loeser[1,2], and Long Pham[1]**

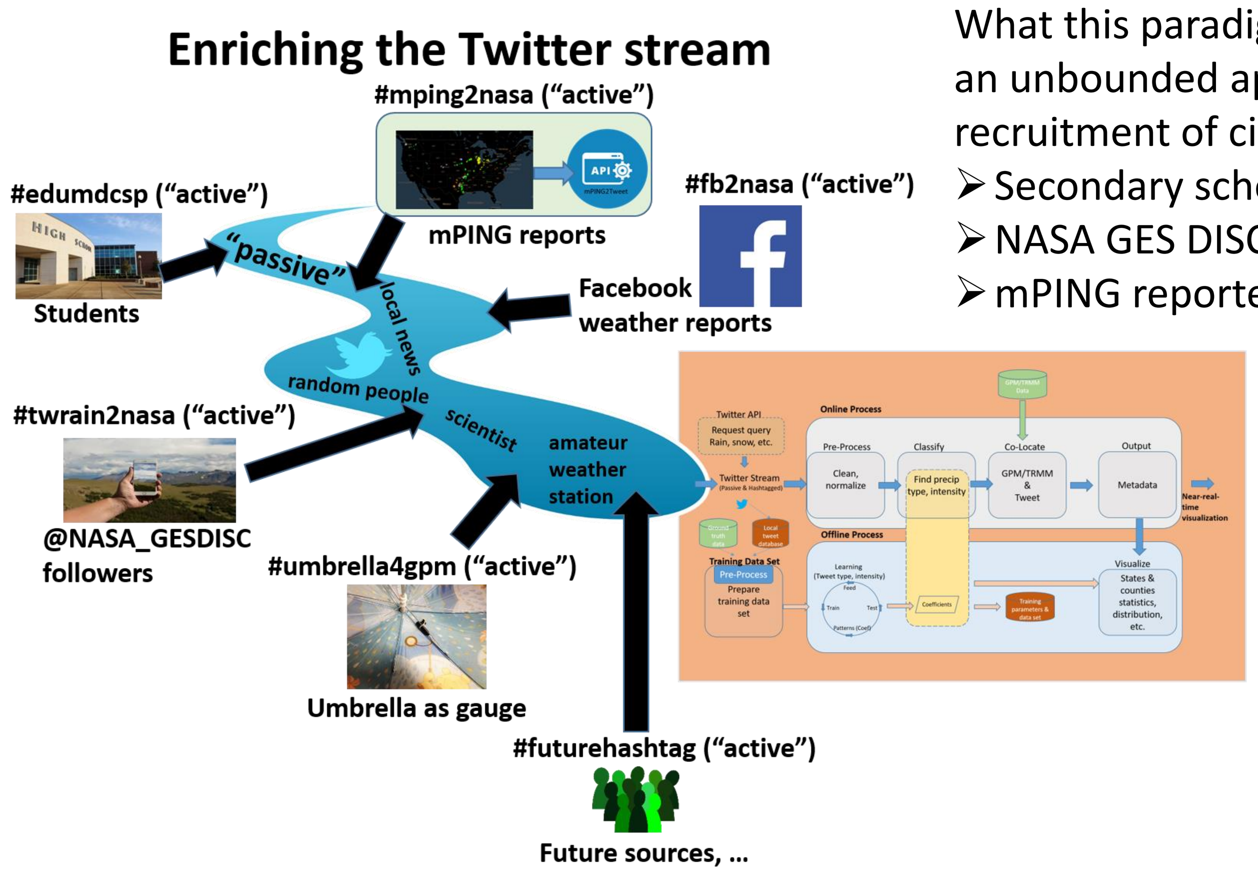
[1]NASA Goddard Space Flight Center; [2]ADNET Systems, Inc.; [3]Cornell University; [4]University of Michigan Ann Arbor; [5]CUNY Queens College

*Emails: Arif.Albayrak@nasa.gov; William.L.Teng@nasa.gov*

## A Paradigm Shift

- Social media data streams, such as Twitter, 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).
- Determinant of output tweet quality from our tweet processing infrastructure is the quality of the tweets retrieved from the Twitter stream.
- Twitter provides a large source of citizen scientists for crowdsourcing. These contributors of "precipitation tweets" do so either knowingly ("active") or not ("passive"). "Active" and "passive" tweets are complementary; "active" tweets serve to "enrich the Twitter stream."
- Concept of enriching the Twitter stream is a paradigm shift from the traditional focus on recruiting citizen scientists.

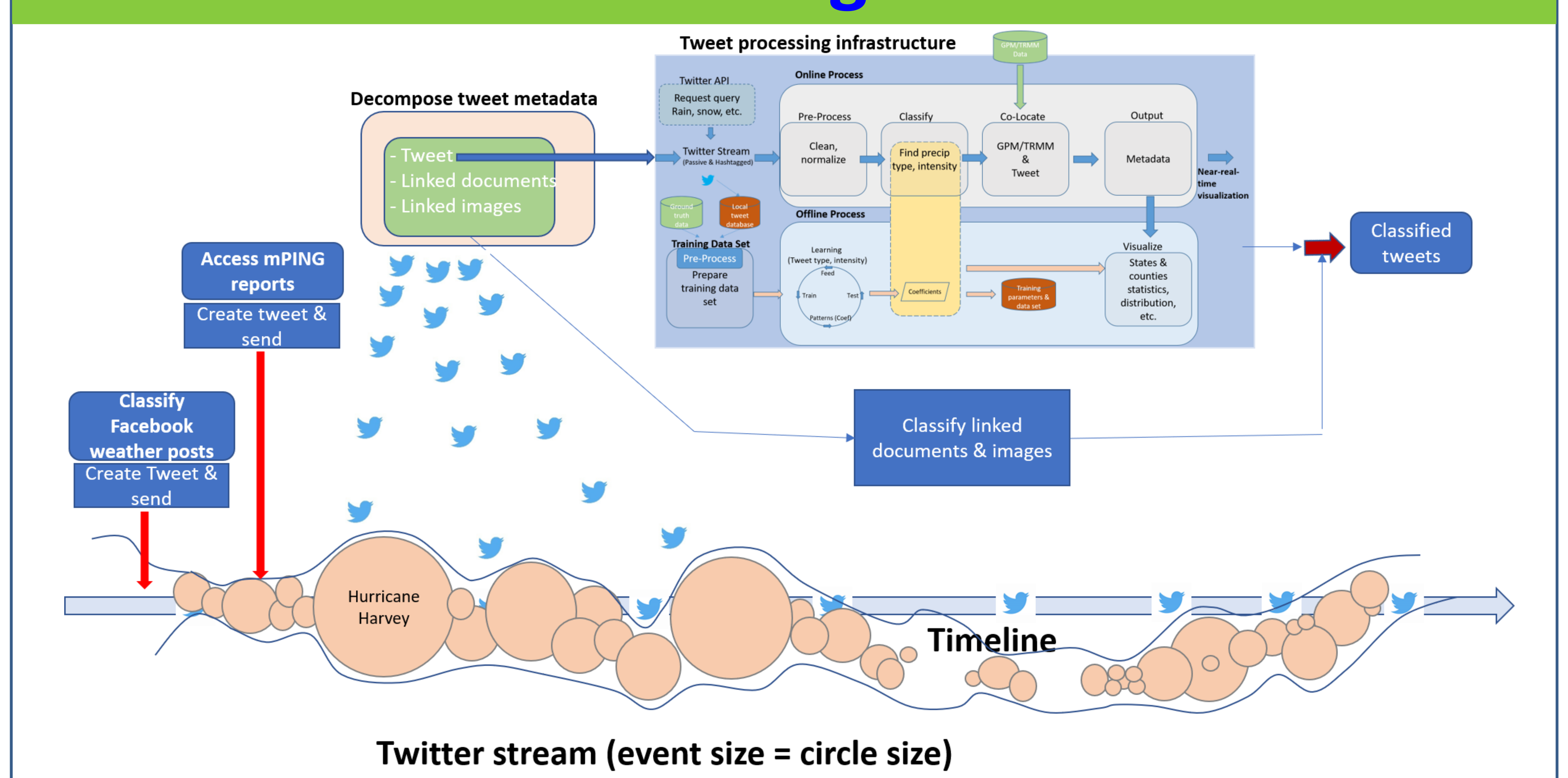
### Passive and Active Tweets




What this paradigm shift enables is an unbounded approach to recruitment of citizen scientists, e.g.,

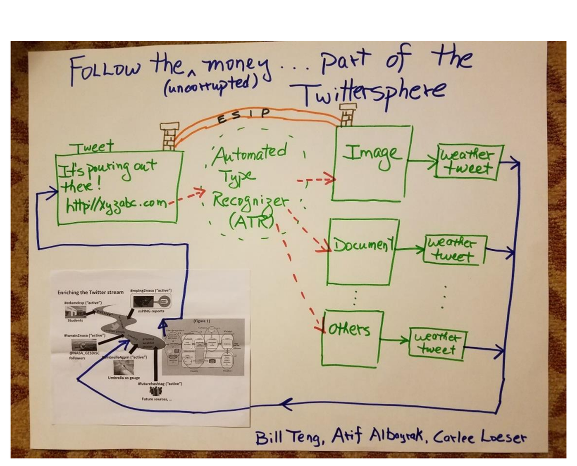
- Secondary school students
- NASA GES DISC Twitter followers
- mPING reporters
- Rain-sensing umbrellas
- Other social media weather posts
- …

## Machine Learning Architecture




### Ongoing work to Improve Quality of Tweets (details in next three sections)

- Classify documents and images that are endpoints of links in tweets, to extract additional information relevant to the tweets.
- Classify Facebook weather posts; convert to "active tweets"
- Automated Type Recognizer of either document or image in endpoint of link in tweet. On the right is our proposal for an informal funding opportunity at the ESIP 2018 Summer Meeting.

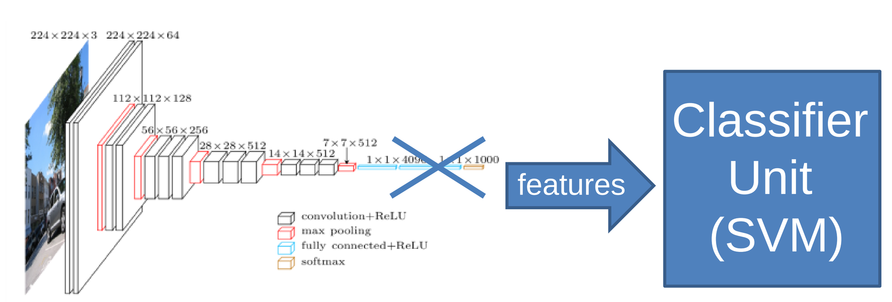



## Classifying Tweet-linked Images (Corcoran)

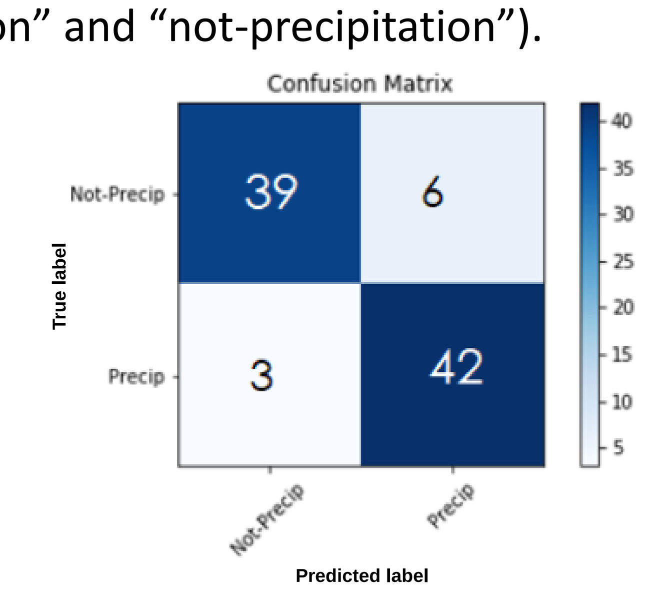
Construct classifier to analyze images for precipitation-related information (e.g., is there rain in the image? is it a forecast map?)

### A Transfer-Learning Approach

- Deep learning models, particularly Convolutional Neural Networks (CNN), have been shown to be very effective for large-scale image recognition and classification.
- Because a large number of labeled images is required to develop CNN, doing so from scratch would be very costly in compute and time resources.
- **Transfer Learning** takes advantage of pre-trained models as a starting point, thus mitigating the cost of model development for the current task. These reused models are, in effect, feature extractors, the outputs from which then become inputs for training smaller, more manageable classifiers.
- For the current task, we used VGG-16[1] as the feature extractor, by removing the final fully connected layers, and trained a linear support vector machine (SVM) to output the final classification (i.e., "precipitation" and "not-precipitation").




Basic structure of our model: VGG-16 architecture[2] (left); fully connected layers removed ("X"); extracted features with which to train classifier (for current task, SVM).




Confusion matrix of classification on a 90-image test set, split into two target categories, "precipitation" and "not-precipitation."

## Classifying Tweet-linked Documents (Wang)

Use Hierarchical Attention Network (HAN)-based model to classify tweet documents, i.e., **precipitation occurrence**, **type**, and **intensity,** at given locations and times.

### Architecture: Hierarchical Attention Networks[3]

*Word/Sentence Encoder* – Embed words to vectors through embedding matrix; apply bidirectional GRU[4] to obtain **representative, contextual** hidden annotations of words/sentences.

*Word/Sentence Attention* – Combine **learned measure of importance** with contextual word/sentence annotations; more relevant words have greater weighting, and vice versa.

*Classification* – Softmax*, categorical cross-entropy#

*Regression* – Softmax, mean squared error

Word Encoder -> Word Attention -> Sentence Encoder -> Sentence Attention -> Classification/Regression

*Softmax – reduces influence of extreme values by constraining data into the range 0-1 before classification.

#Categorical cross-entropy – used as final layer in classifying data to predefined classes.

### Results / Alternate Model Comparisons

| Data Type | Accuracy |
|---|---|
| Precipitation Occurrence (Yes/No) | 83.6% |
| Precipitation Type (Rain/Sleet/Snow/Null) | 77.4% |
| Precipitation Intensity (mm/hr) | 74.2% (±1.5 mm/hr) |


### Model Attention Weight Visualizations

Model places greater importance on darker words during classification / regression.

**Example 1:** At 9:04 AM, Deerfield [Franklin Co, MA] HAM RADIO reports SNOW of 3.00 INCH #BOX https://t.co/123

**Word-level Attention Weights Visualization:**

at <time> , deerfield [ franklin co , ma ] <allcaps> ham radio </allcaps> reports <allcaps> snow </allcaps> of <number> <allcaps> inch </allcaps> <hashtag> box </hashtag> <url>

**Example 2:** It's chilly outside but so beautiful! Enjoy the snow day ❄ #staywarm #snowmageddon… https://t.co/123

**Sentence-level Attention Weights Visualization:**

it ' s chilly outside but so beautiful !
enjoy the snow day ❄ stay warm

**Word-level Attention Weights Visualization:**

it ' s chilly outside but so beautiful !
enjoy the snow day ❄ <hashtag> stay warm </hashtag> <hashtag> snowmageddon </hashtag> … <url>

**Example 3:** All this snow has us wishing for summer! Sipping on our beachy 🌴 Coco Rooibos Chai this morning… https://t.co/123

**Sentence-level Attention Weights Visualization:**

all this snow has us wishing for summer !
sipping on our beachy 🌴 coco rooibos chai this

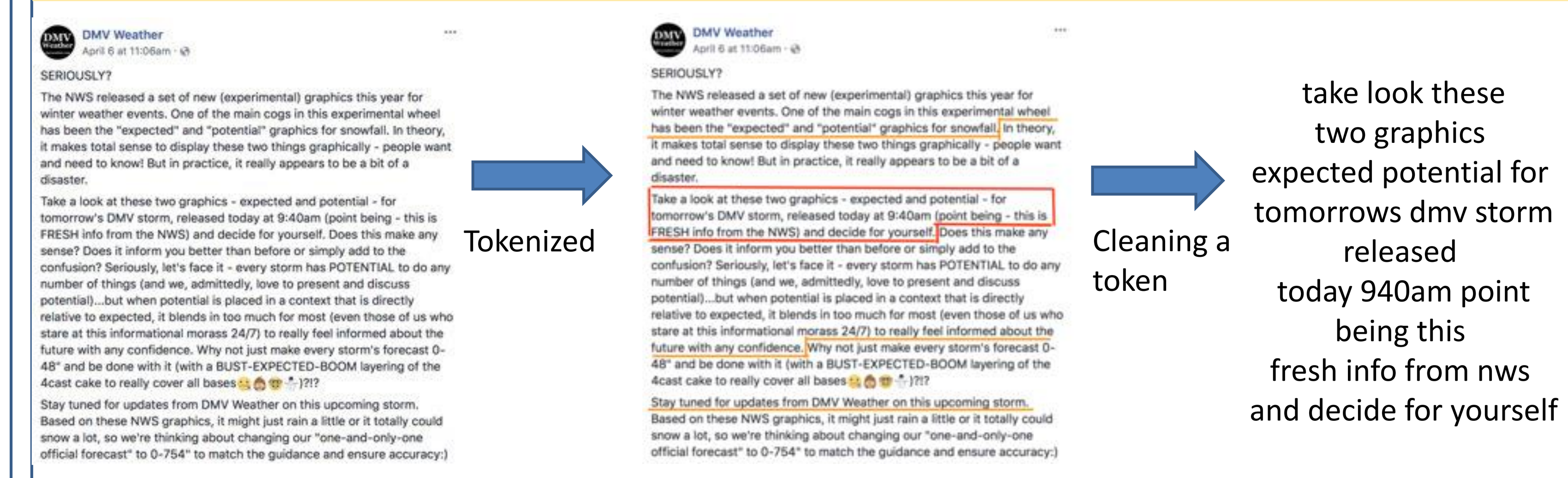
**Word-level Attention Weights Visualization:**

all this snow has us wishing for summer !
sipping on our beachy 🌴 coco rooibos chai this morning … <url>

## Classifying Facebook Posts (Maksumov)

Construct classifiers to label posts on Facebook weather pages as precipitation related (e.g., is this post suggesting that it's snowing right now?)

### Data Preprocessing: Cleaning Facebook Posts




- Original Facebook posts scraped using Python program by minimaxir[5]

### Creating Feature Vectors: TF-IDF (Term Frequency-Inverse Document Frequency)

For each word in a token, multiply together:

- # times the word appears in the token
- Log (# tokens / # tokens with the word)

Converted, messy text data -> numerical data to analyze

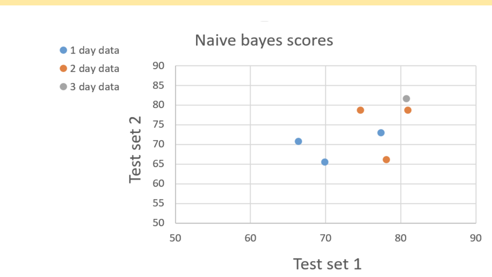
### Creating a Classifier: Naive Bayes Algorithm

Best label maximizes chances of finding label in training set and finding token's words in label

$$c_{NB} = \underset{c \in C}{\operatorname{argmax}}\, P(c_j) \prod_{x \in X} P(x \mid c)$$

### Results: Classifying the Token's Words

- More data in training set -> greater label accuracy
- 300+ training data -> about 80% correct classification




## Summary

- Paradigm shift from a focus on recruiting citizen scientists to enriching the Twitter stream enables an unbounded approach to recruitment.
- Output tweet quality and quantity from our tweet processing infrastructure is increased by complementing "passive" tweets from the Twitter stream with tweets from "active" participants.
- Ongoing work to improve quality of tweets include (1) classifying documents and images that are endpoints of links in tweets, to extract additional information relevant to the tweets and (2) classifying Facebook weather posts and converting them to "active" tweets.

### References

[1]VGG-16 (Visual Geometry Group-16; also OxfordNet), https://gist.github.com/baraldilorenzo/07d7802847aaad0a35d3

[2]Image of VGG-16 architecture, https://blog.heuritech.com/2016/02/29/a-brief-report-of-the-heuritech-deep-learning-meetup-5/

[3]Yang et al., 2016, Hierarchical attention networks for document classification, *in* Proc. 2016 Conf. NAACL-HLT, San Diego, 1480-1489.

[4]Bahadanau et al., 2014, Neural machine translation by jointly learning to align and translate, arXiv:1409.0473 [cs.CL].

[5]Woolf, M. (minimaxir), 2017, Facebook page post scraper, https://github.com/minimaxir/facebook-page-post-scraper.

**Acknowledgments: This work was supported, in part, by**

- NASA Citizen Science for Earth Systems Program, for the project, "Mining Twitter data to augment NASA GPM validation."
- NASA Goddard's Office of Education, for the internship of Daniel Maksumov.
- NASA GES DISC, for the internship of Sky Wang.
- ADNET Systems, Inc., for the internship of John Corcoran.