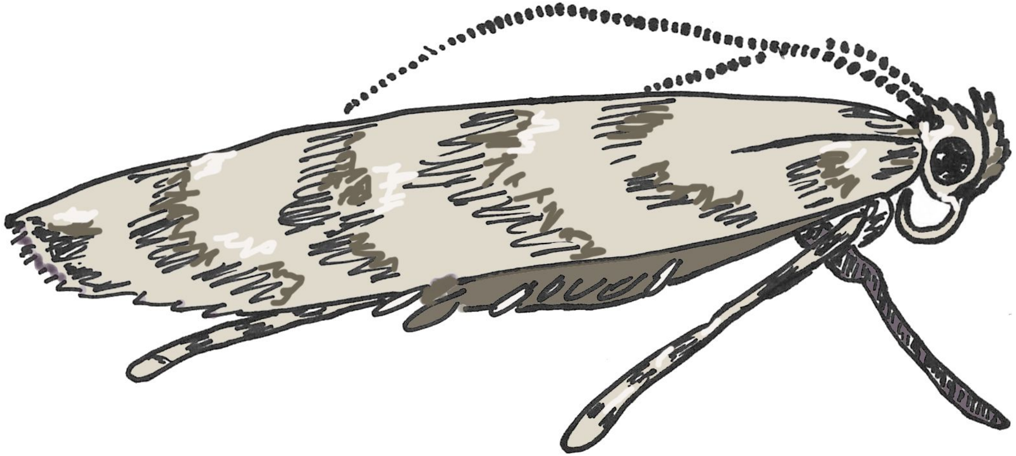
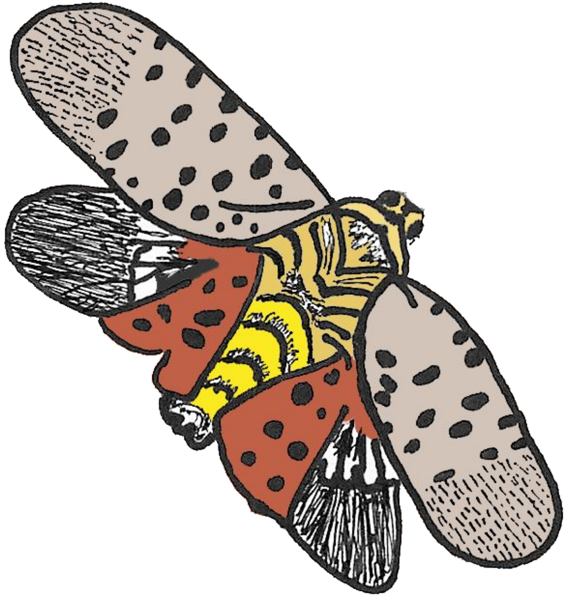
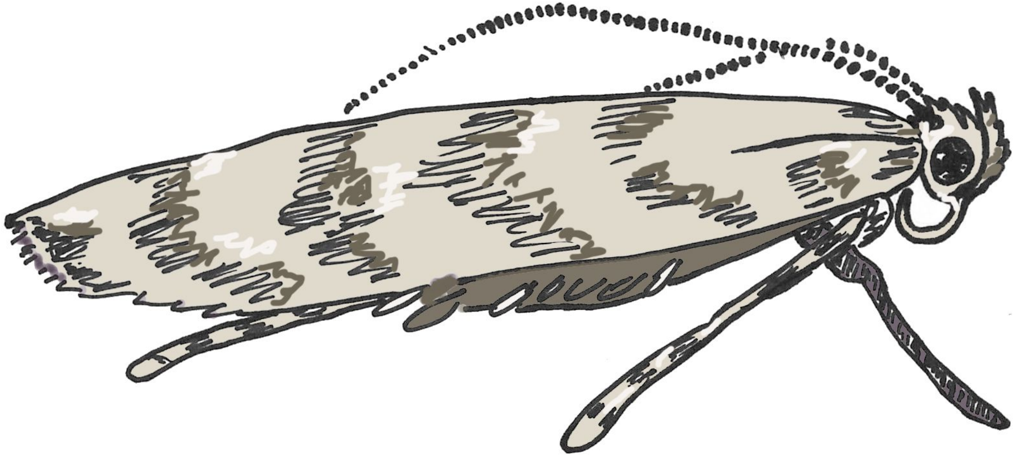


Case Studies in Mining Text for Plant Pests and Pathogens

Laura Tateosian, Ariel Saffer, Makiko Shukunobe, & Chelsey Walden-Schreiner

Center for Geospatial Analytics
North Carolina State University



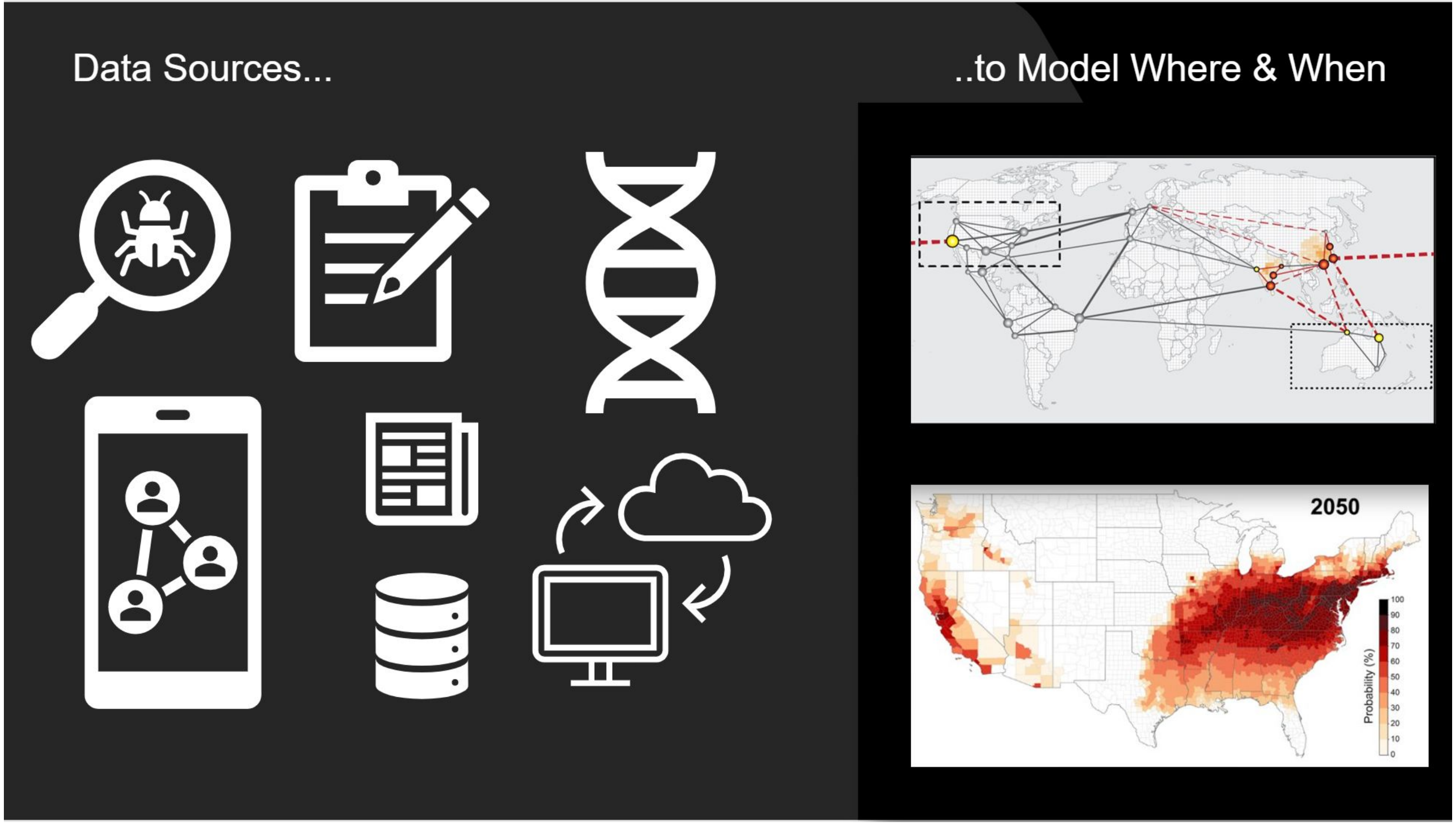


Case Studies in Mining Text for Plant Pests and Pathogens

Laura Tateosian, Ariel Saffer, Makiko Shukunobe, & Chelsey Walden-Schreiner

Center for Geospatial Analytics
North Carolina State University

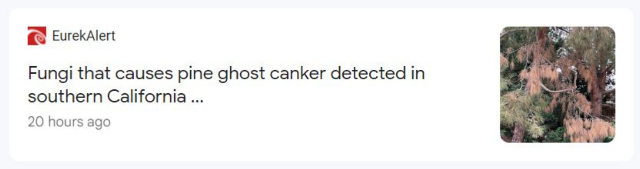
The problem: timely pest records at scale



Opportunities in Data and AI

CBS AUSTIN

The Amazing Acro-Cats are coming to Austin, including the only all-cat band in the world



EurekAlert

Fungi that causes pine ghost canker detected in southern California ...

20 hours ago

The Verge

Meta has its own new AI tech — meet LLaMa

5 hours ago

24 38.5k 🔥 Turtle stretches its webbed feet while sunbathing

submitted 15 hours ago by SinjiOnO to r/NatureIsFuckingLit

541 comments share save hide give award report crosspost

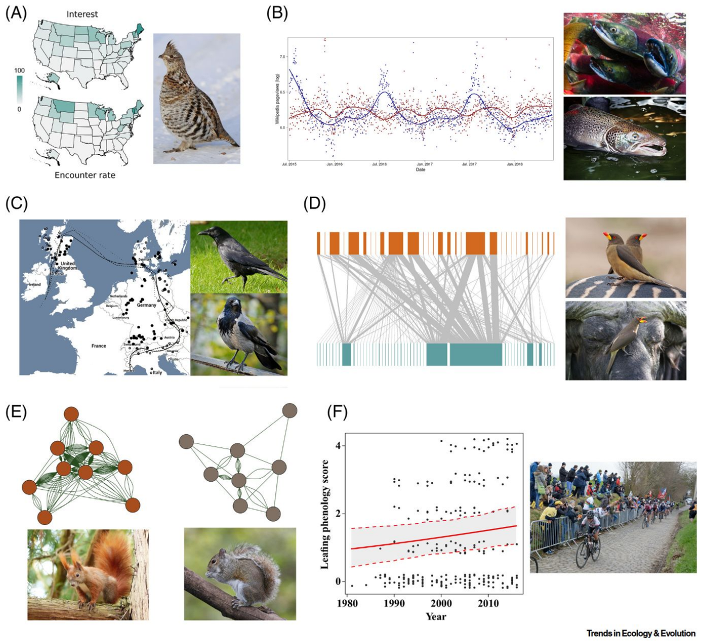
a BigScience initiative

BLOOM

176B params · 59 languages · Open-access

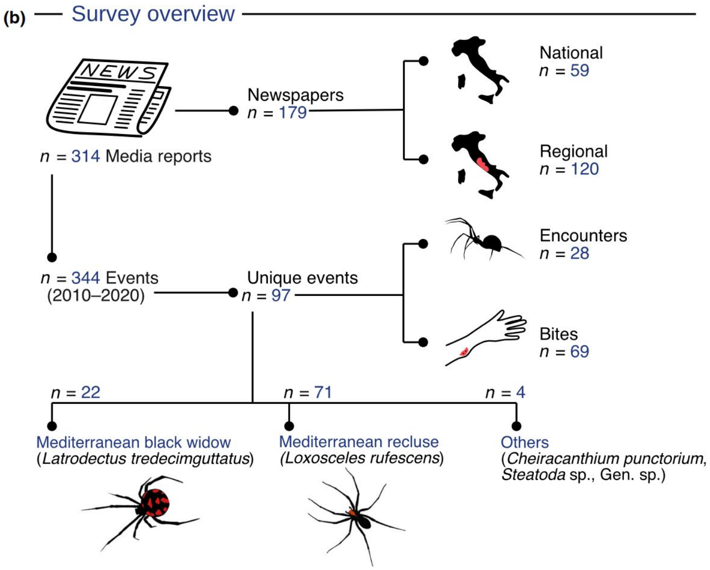
Volume + Velocity + Variety + Veracity + Value

Observing species (including humans) through text mining



Trends in Ecology & Evolution

Jarić, Ivan, et al. "**iEcology**: harnessing large online resources to generate ecological insights." *Trends in Ecology & Evolution* 35.7 (2020): 630-639.



(b) Survey overview

n = 314 Media reports

Newspapers
n = 179

National
n = 59

Regional
n = 120

n = 344 Events
(2010–2020)

Unique events
n = 97

Encounters
n = 28

Bites
n = 69

n = 22

n = 71

n = 4

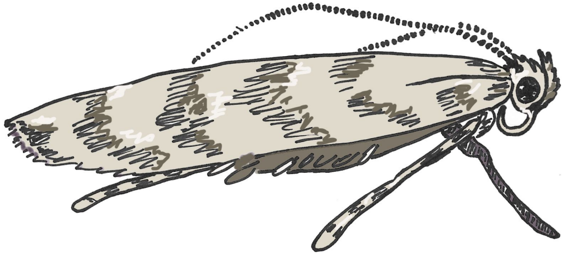
Mediterranean black widow
(*Latrodectus tredecimguttatus*)

Mediterranean recluse
(*Loxosceles rufescens*)

Others
(*Cheiracanthium punctorium*, *Steatoda* sp., Gen. sp.)

Mammola, Stefano, et al. "Media framing of spiders may exacerbate arachnophobic sentiments." *People and Nature* 2.4 (2020): 1145-1157.

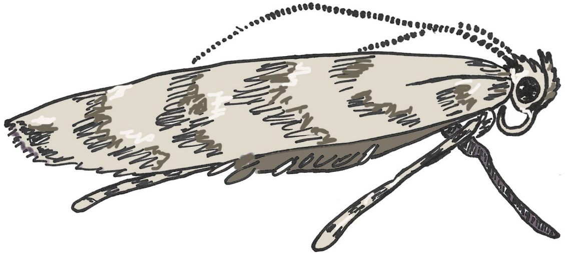
Four case studies



Goals:

1. Describe the potential for multiple **text sources** to provide **content valuable** to **document pest spread**,
2. Help others who seek to track pests through text media **overcome the methodological hurdles** associated with using text data.

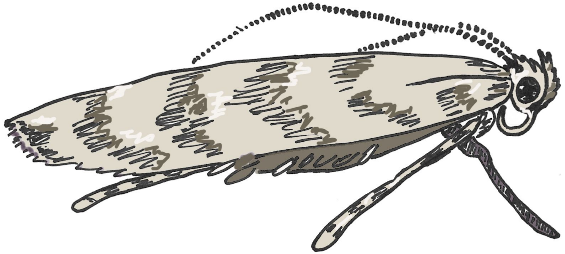
Four case studies



1843-1850

2009-2023

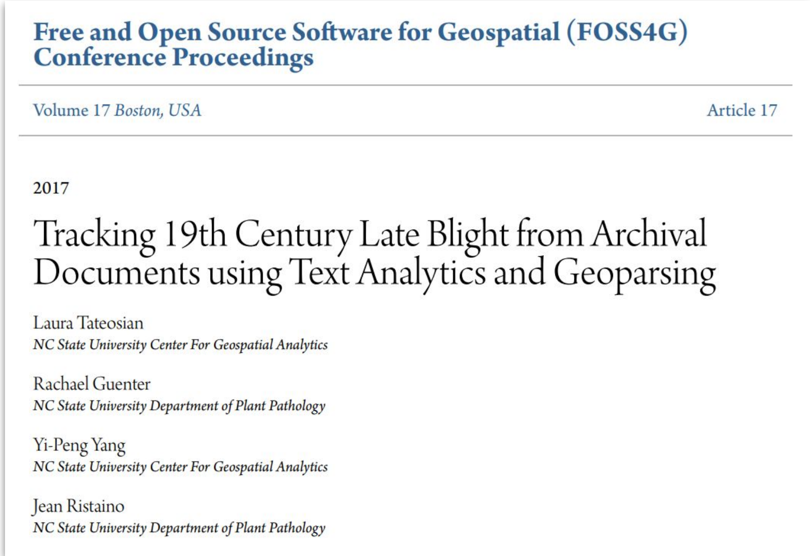
Four case studies



Pennsylvania

Nigeria

Case study 1:   1840’s P. Infestins



Free and Open Source Software for Geospatial (FOSS4G) Conference Proceedings

Volume 17 *Boston, USA* — Article 17

2017

Tracking 19th Century Late Blight from Archival Documents using Text Analytics and Geoparsing

Laura Tateosian
NC State University Center For Geospatial Analytics

Rachael Guenter
NC State University Department of Plant Pathology

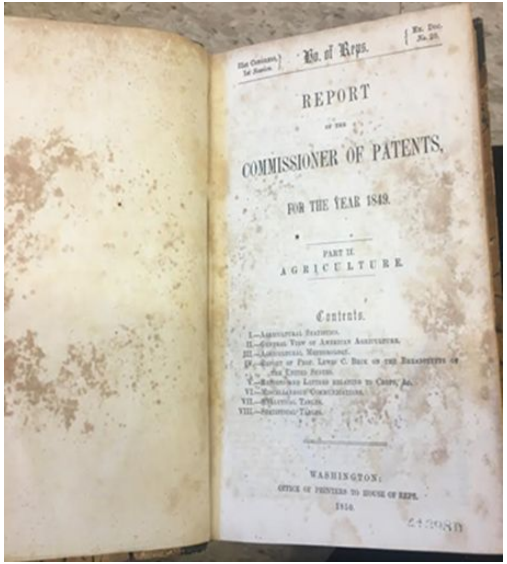
Yi-Peng Yang
NC State University Center For Geospatial Analytics

Jean Ristaino
NC State University Department of Plant Pathology

PATENT OFFICE REPORT
PATENT OFFICE REPORT 1852
PATENT OFFICE REPORT 1852
PATENT OFFICE REPORT
PATENT OFFICE REPORT 1851

US Annual Report of the Commissioner of Patents

1841-1850 (1 doc./year)



31st Congress, 1st Session. — Ho. of Reps. — Ex. Doc. No. 20.

REPORT

OF THE

COMMISSIONER OF PATENTS,

FOR THE YEAR 1849.

PART II.

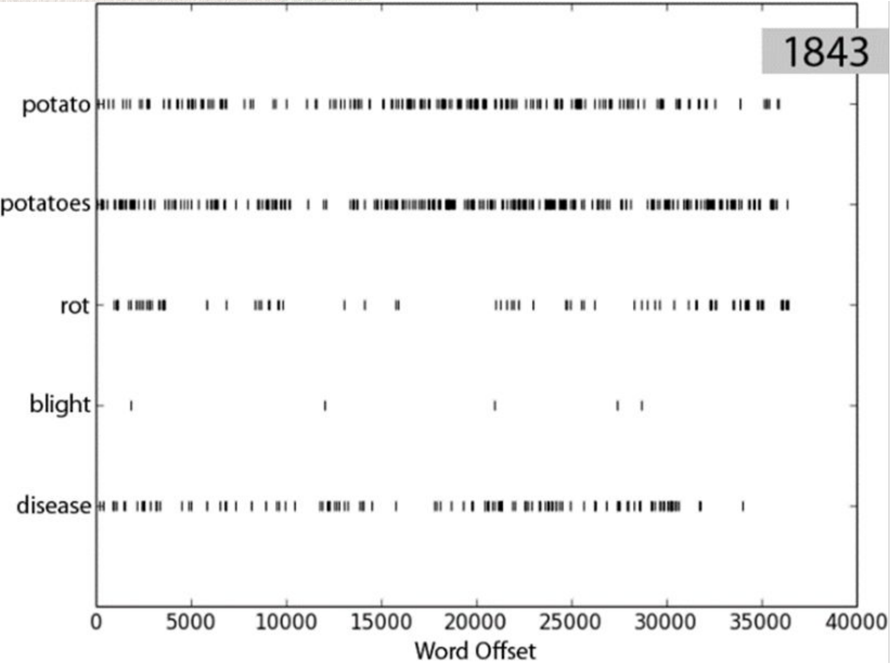
AGRICULTURE.

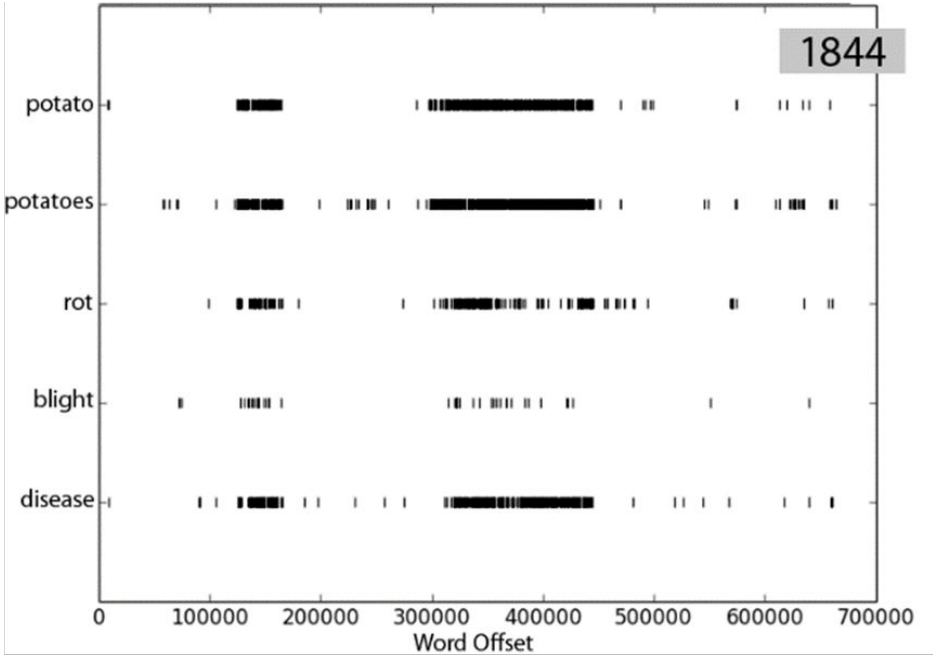
Contents.

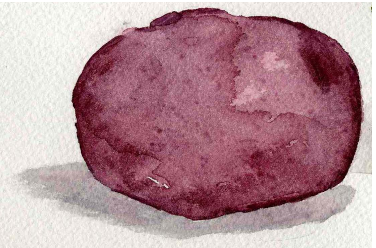
I.—Agricultural Statistics.
II.—General View of American Agriculture.
III.—Agricultural Meteorology.
IV.—Report of Prof. Lewis C. Beck on the Breadstuffs of the United States.
V.—Reports and Letters relating to Crops, &c.
VI.—Miscellaneous Communications.
VII.—Statistical Tables.
VIII.—Statistical Tables.

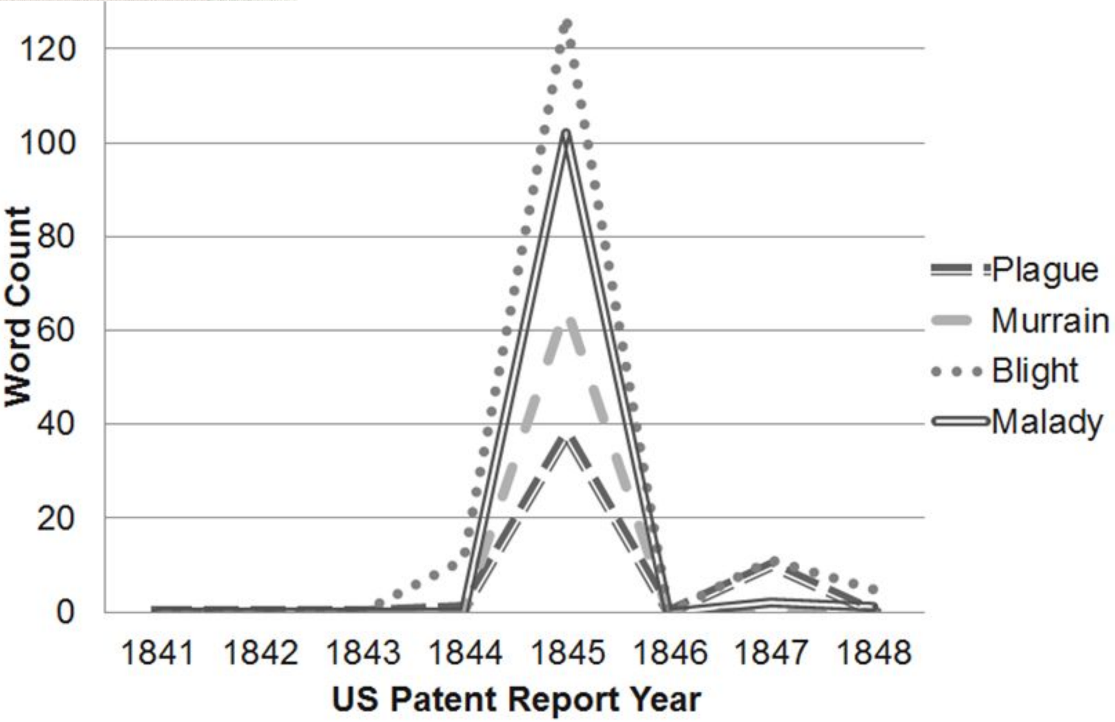
WASHINGTON:
OFFICE OF PRINTERS TO HOUSE OF REPS.
1850.

243980


1843
potato
potatoes
rot
blight
disease
0
5000
10000
15000
20000
25000
30000
35000
40000
Word Offset

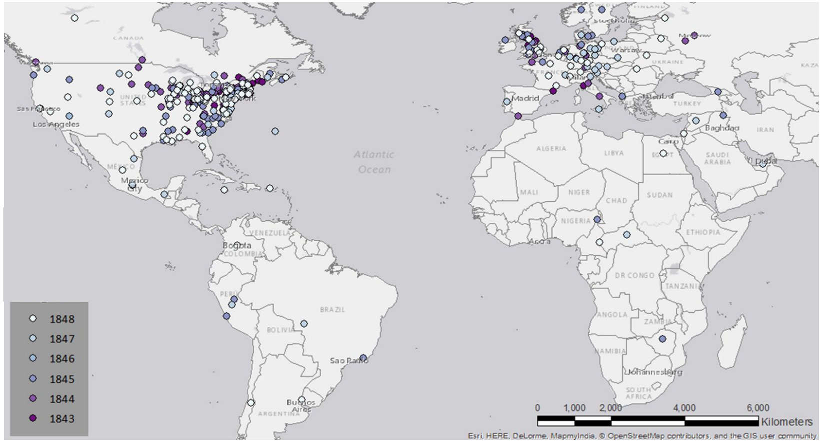

1844
potato
potatoes
rot
blight
disease
0
100000
200000
300000
400000
500000
600000
700000
Word Offset



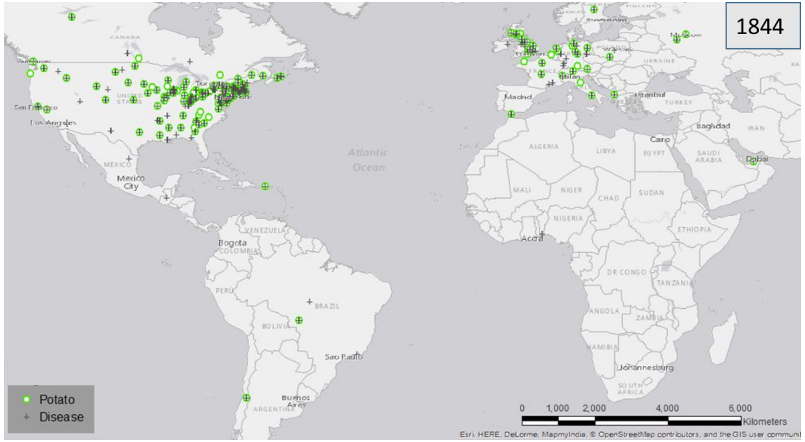


Terminology varies across pest/pathogen, geographically, and over time.

Mapping potato and disease mentions*



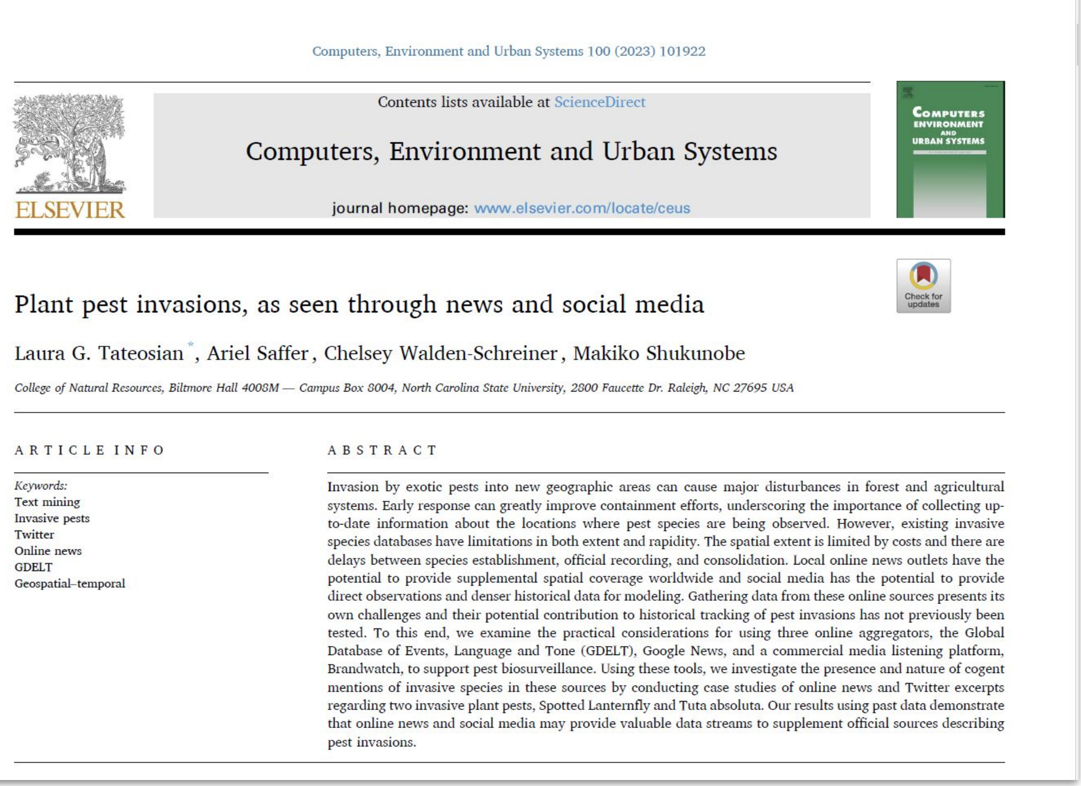
Esri, HERE, DeLorme, MapmyIndia, © OpenStreetMap contributors, and the GIS user community



Esri, HERE, DeLorme, MapmyIndia, © OpenStreetMap contributors, and the GIS user communi

*Assumption: Place names that are mentioned in the text close to topically relevant terms likely to be involved.

Case studies 2 & 3



Computers, Environment and Urban Systems 100 (2023) 101922

Contents lists available at ScienceDirect

Computers, Environment and Urban Systems

journal homepage: www.elsevier.com/locate/ceus

Plant pest invasions, as seen through news and social media

Laura G. Tateosian *, Ariel Saffer, Chelsey Walden-Schreiner, Makiko Shukunobe

College of Natural Resources, Biltmore Hall 4008M — Campus Box 8004, North Carolina State University, 2800 Faucette Dr. Raleigh, NC 27695 USA

ARTICLE INFO

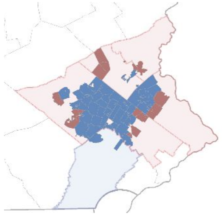
Keywords:
Text mining
Invasive pests
Twitter
Online news
GDELT
Geospatial–temporal

ABSTRACT

Invasion by exotic pests into new geographic areas can cause major disturbances in forest and agricultural systems. Early response can greatly improve containment efforts, underscoring the importance of collecting up-to-date information about the locations where pest species are being observed. However, existing invasive species databases have limitations in both extent and rapidity. The spatial extent is limited by costs and there are delays between species establishment, official recording, and consolidation. Local online news outlets have the potential to provide supplemental spatial coverage worldwide and social media has the potential to provide direct observations and denser historical data for modeling. Gathering data from these online sources presents its own challenges and their potential contribution to historical tracking of pest invasions has not previously been tested. To this end, we examine the practical considerations for using three online aggregators, the Global Database of Events, Language and Tone (GDELT), Google News, and a commercial media listening platform, Brandwatch, to support pest biosurveillance. Using these tools, we investigate the presence and nature of cogent mentions of invasive species in these sources by conducting case studies of online news and Twitter excerpts regarding two invasive plant pests, Spotted Lanternfly and Tuta absoluta. Our results using past data demonstrate that online news and social media may provide valuable data streams to supplement official sources describing pest invasions.

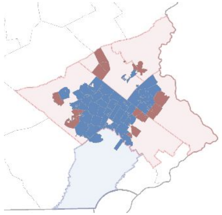
Case study 2: Spotted Lanternfly in the news

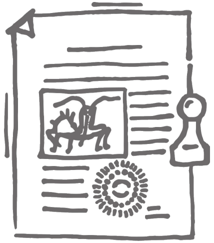
Spotted lanternfly terms + 2017 + PA →



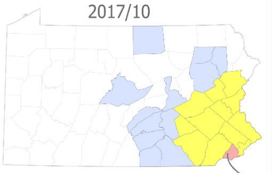
Case study 2:   Spotted Lanternfly in the news

Spotted lanternfly terms + 2017 + PA →



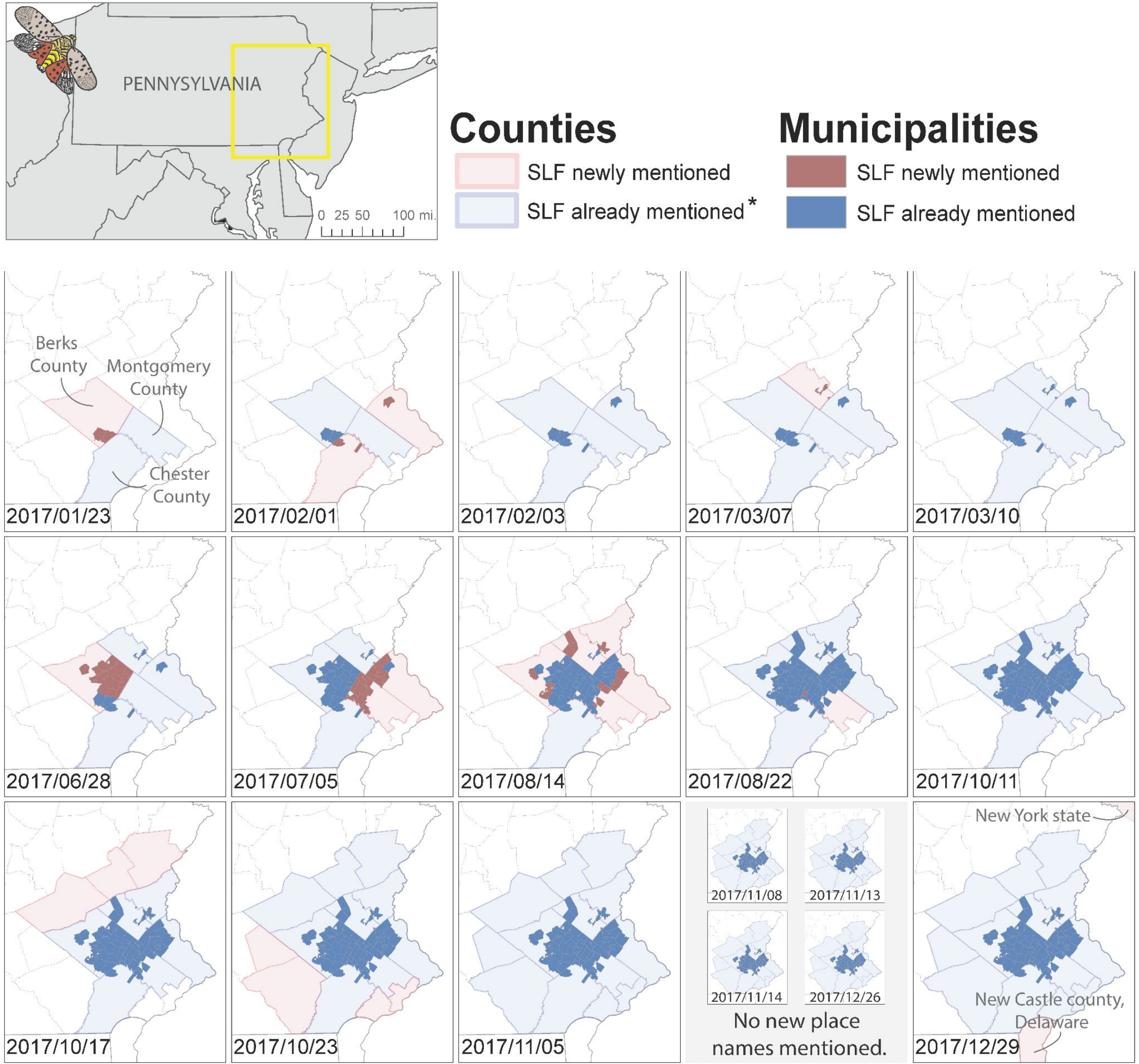


Spotted lanternfly terms + 2014-2017 + PA →

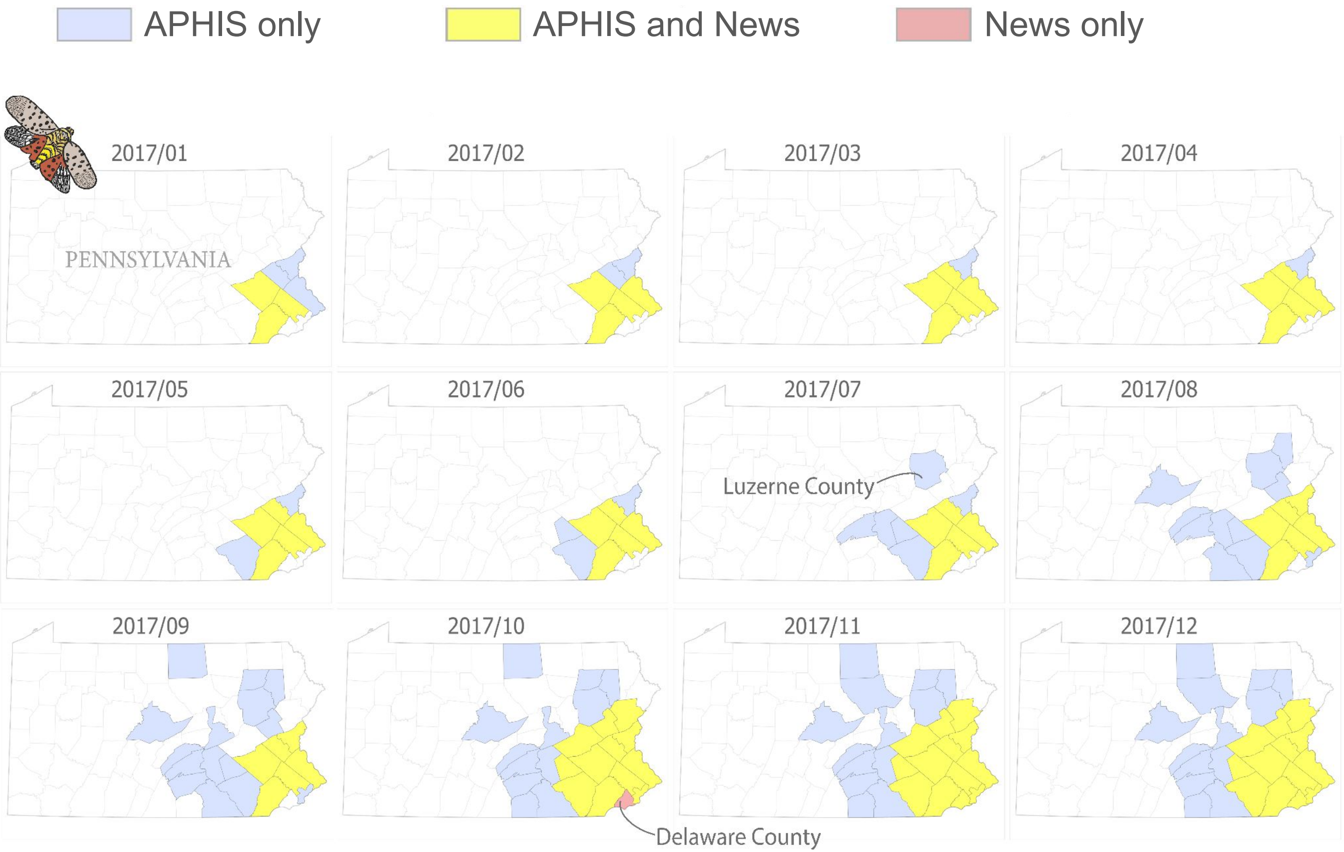


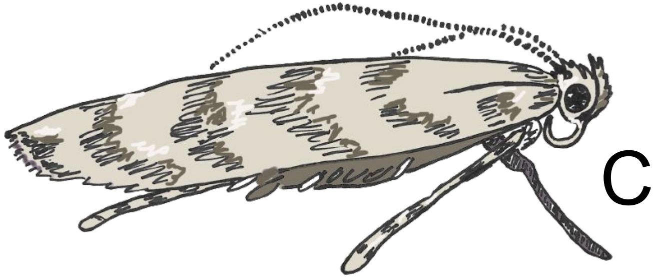
USDA- APHIS records,
aggregated to county

Current quarantine conditions, readily available. *Historical* temporal progression of quarantine, less so.



News shows how county quarantine efforts closely followed the early spread.



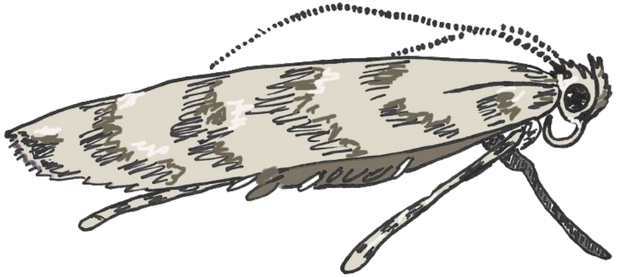


Cast study 4: Tuta absoluta posts

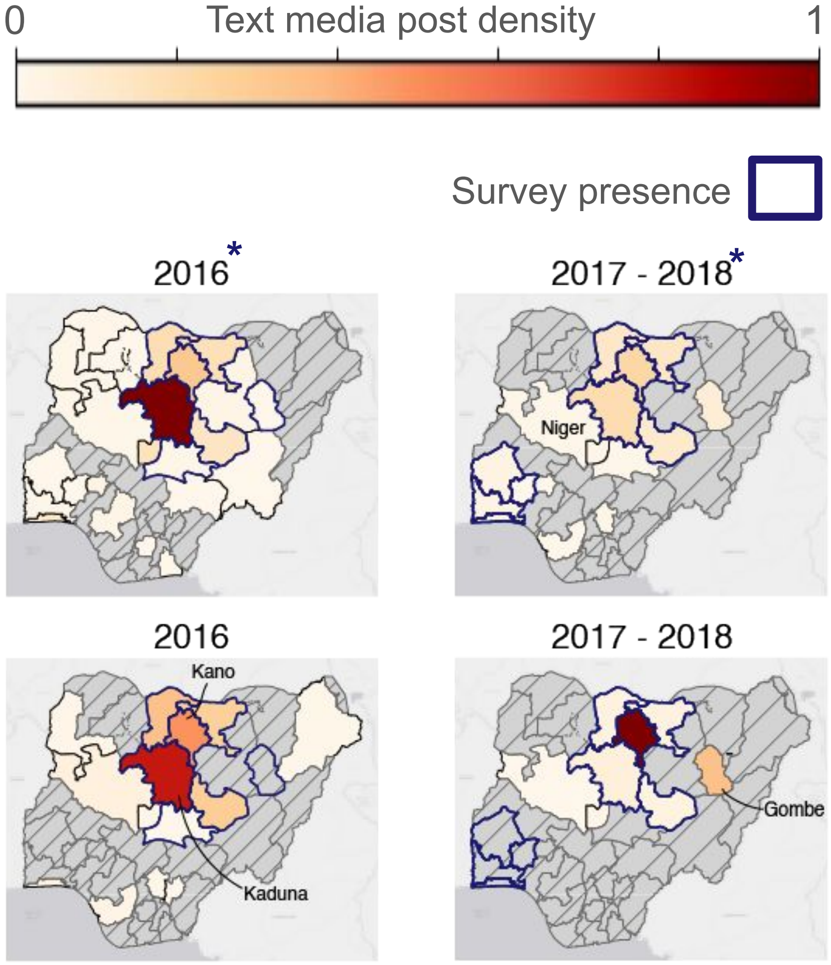
Tuta absoluta + Tweets/News → state mentioned

Tuta absoluta + Nigeria + survey data → presence by state

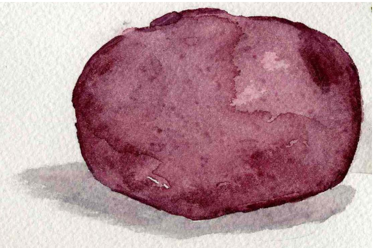
2016 survey (Borisade et al., 2017) & 2017-2018 survey (Aigbedion-Atalor et al., 2019)



News and Twitter posts about a location closely matched known pest locations, and highlight other potential pest locations



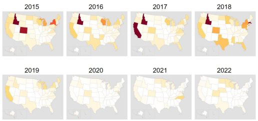
* **2016** data published in ***2017***, **2017 - 2018** data published in ***2019***



Case study 4: P. Infestans now

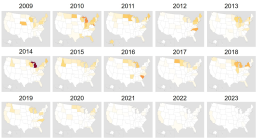
GDELT

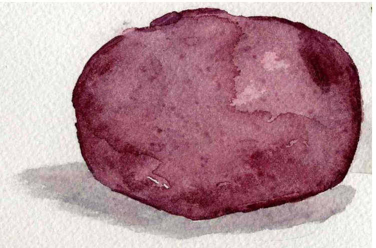
P. Infestans + 2015–2022 + USA →



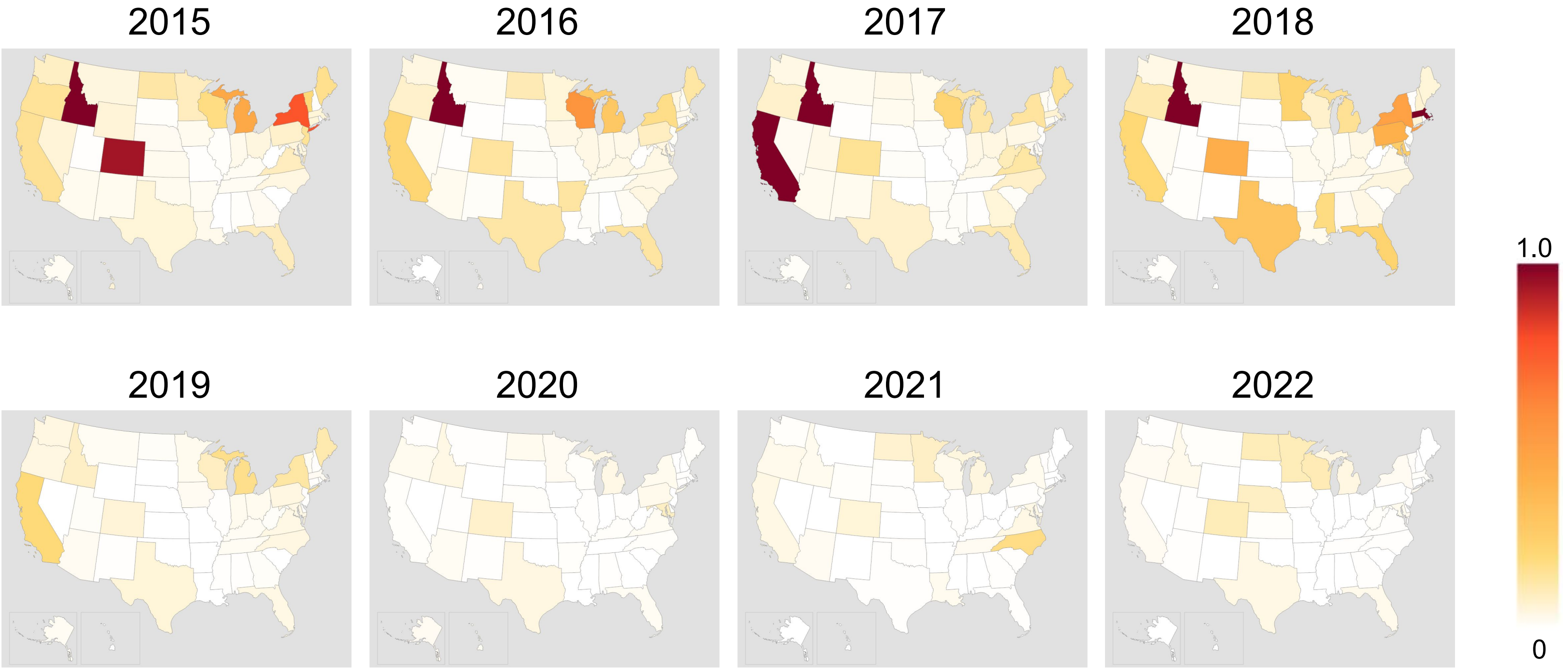
Twitter API

P. Infestans + 2009–2023 + USA →





P. Infestans news from GDELT*

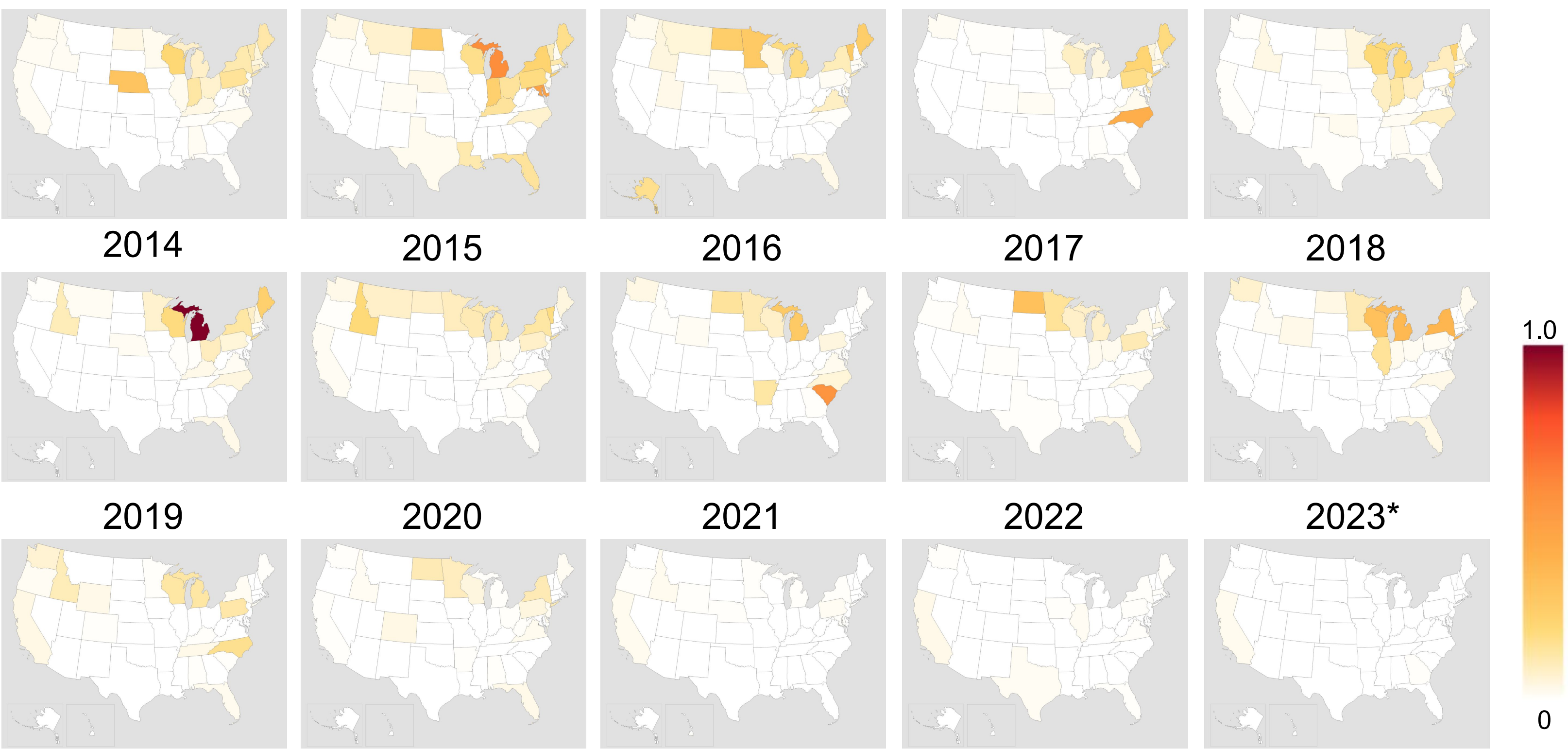


Min: 0
Max: 410

*GDELT Project = Global Database of Events, Language and Tone

P. Infestans Tweets from Twitter API

Min: 0
Max: 257



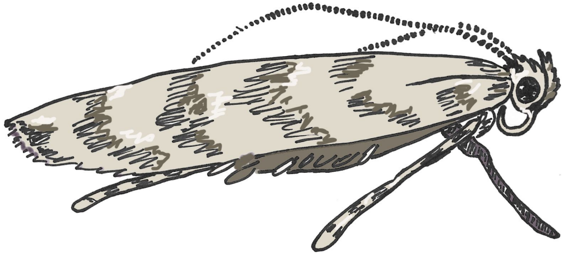
2009 2010 2011 2012 2013

2014 2015 2016 2017 2018

2019 2020 2021 2022 2023*

*2023 only through Feb 1

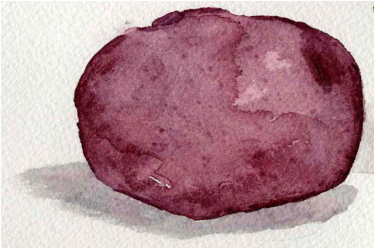
Discussion

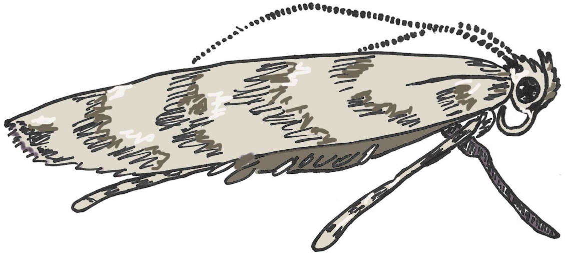


Potential to catch first records, faster

Fill gaps in survey data, with volume as an indicator of **invasion intensity**

Discussion



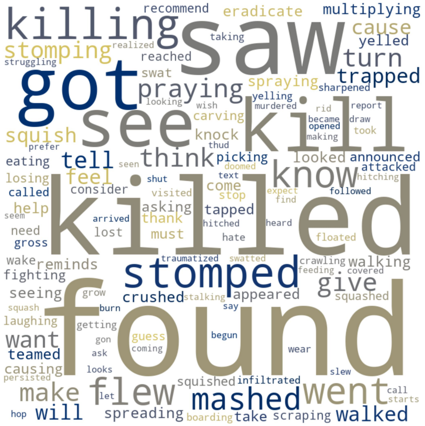


Potential to catch first records, faster

Fill gaps in survey data, with volume as an indicator of **invasion intensity**

Opportunity to extract **continuous, time-sensitive pest information** to get it data faster

Technical challenges for automation: place-name **disambiguation** and **text classification**



Tailoring Named Entity Recognition (NER) to extract pest event data from online news and Tweets

Ariel Saffer, Laura Tateosian, Makiko Shukunobe, Chelsey Walden-Schreiner, Ross Meentemeyer

NC STATE UNIVERSITY

Center for Geospatial Analytics

Obtaining **timely, complete records** is a challenge for global pest biosurveillance

At a global scale, formal observation records used to track and manage spreading pests may be spatially and temporally sparse.

Further, the time from data collection to research publication delays access to this data.

Informal pest observations shared in **news** and **social media** may offer an abundant, low-cost alternative for accessing **real-time and historical spatiotemporal information** about spreading pests.

We explored **web media** as **an alternative data** source, comparing posts to official records

10 years of web news and Twitter posts about 2 emerging pests

Tuta absoluta
Nigeria

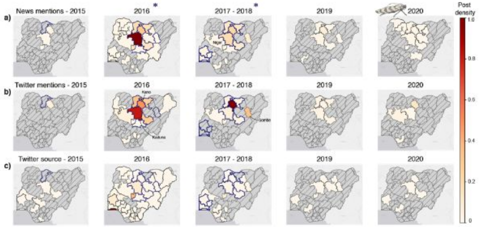
Spotted lanternfly
Pennsylvania, USA

Compared timing, post origin, and places mentioned in posts with scientific pest observations

- Tuta absoluta: data from two published ground surveys in Nigeria
- Spotted lanternfly: point observations collected by USDA APHIS in Pennsylvania

Media provided a **low latency source** of data to **fill spatial and temporal gaps** in pest observations

States mentioned in news and Tweets **match** and provide **earlier access** to Tuta absoluta locations observed in surveys and span beyond survey years.



* 2016 survey data published in 2017, 2017 - 2018 data published in 2019

Results published in: Tateosian, Laura G., Ariel Saffer, Chelsey Walden-Schreiner, and Makiko Shukunobe. "**Plant Pest Invasions, as Seen through News and Social Media.**" Computers, Environment and Urban Systems 100 (March 1, 2023).

This work highlights the potential for **news and Tweets to fill gaps in existing data** and **reduce the latency** of new pest records… but further work is still needed to **reliably and automatically extract pest events** from large volumes of **unstructured text.**

Up next: Automatically extracting pest events from text

ACQUIRE DATA

3 species	Unique posts
 Tuta absoluta *Global, emerging*	12,235 news 7,013 Tweets
 Spotted lanternfly *USA, emerging*	17,667 news 33,253 Tweets
 Phytophthora infestans *Global, widespread*	18,975 news 9,033 Tweets

DEFINE PEST EVENT "ENTITIES"

Named Entity Recognition (NER) is a Natural Language Processing (NLP) approach to classify words or multi-word "**entities**" in text.

Pest events entities include…

Pest (What) · Geo-location (Where) · Date (When)

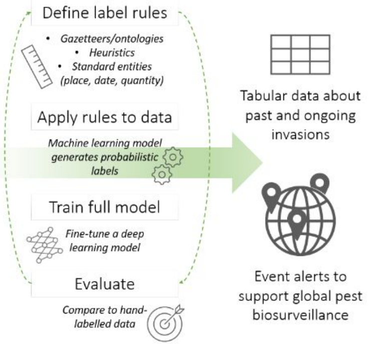
And additional entities like…

Host · Symptoms and damage · Loss and costs

LABEL DATA AND TRAIN MODELS

"UPDATE: 1 of 8 sites in Kent County, Ontario LOC tested positive for Phytophthora infestans PEST , the causal agent of late blight PEST in tomato HOST and potato HOST , this week DATE July-15-18 DATE sampling period."

"The spotted lanternfly PEST has been seen in Greenwich LOC this fall DATE …. first arrived from Asia LOC in 2014 DATE and can be particularly harmful to apples HOST and grapes HOST ."


Define label rules

- *Gazetteers/ontologies*
- *Heuristics*
- *Standard entities (place, date, quantity)*

Apply rules to data

Machine learning model generates probabilistic labels

Train full model

Fine-tune a deep learning model

Evaluate

Compare to hand-labelled data

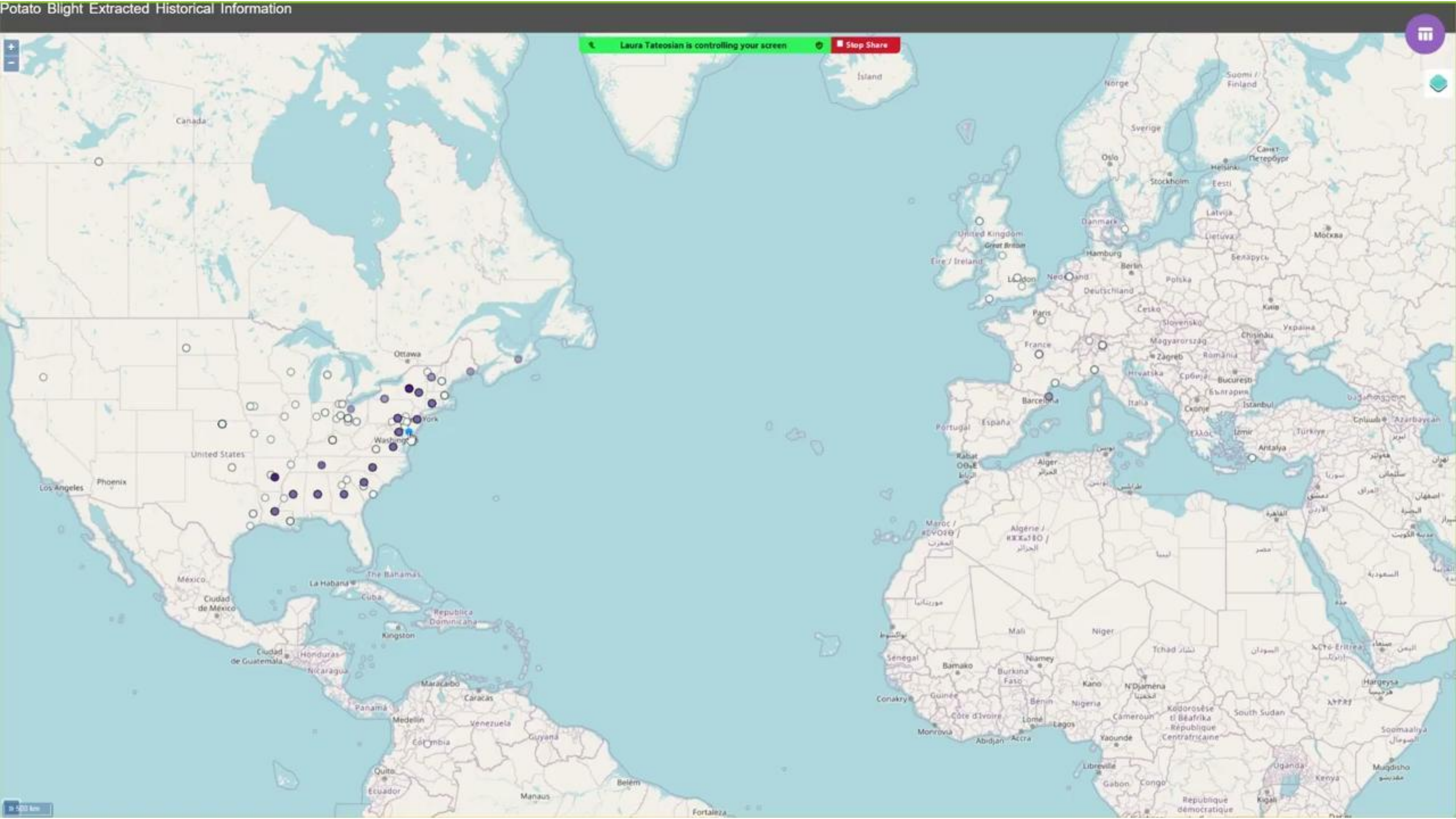
Tabular data about past and ongoing invasions

Event alerts to support global pest biosurveillance

CHALLENGES

- Place name disambiguation (geoparsing)
- Pipelines for underrepresented languages
- Generalization across pests
- Consolidating unique pest events
- Capturing observation uncertainty

…. and more!

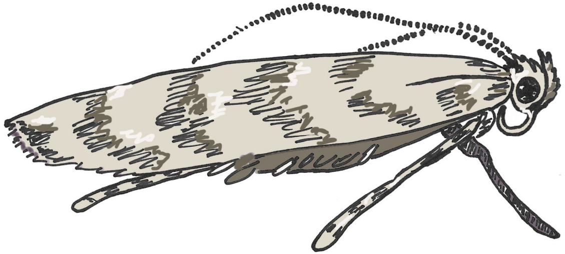


Potato Blight Extracted Historical Information

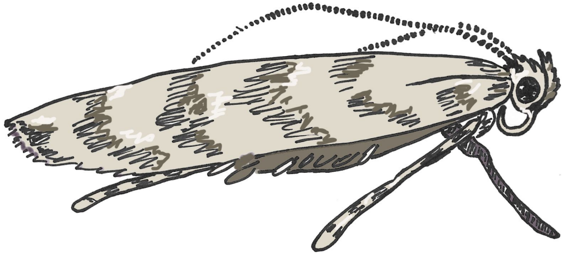
Laura Tateosian is controlling your screen

Stop Share

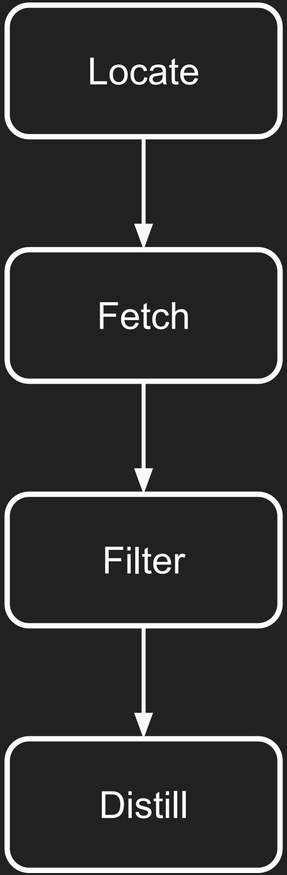
Questions?



Activity



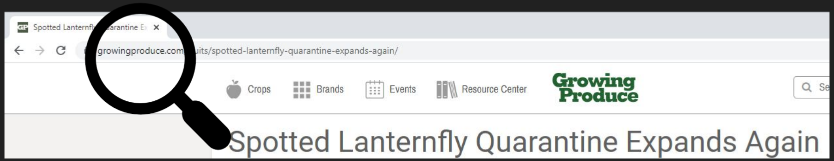
https://go.ncsu.edu/pipp


Locate

Fetch

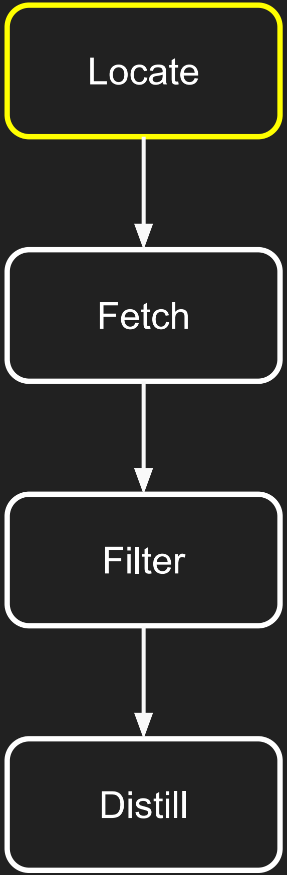
Filter

Distill



growingproduce.com uits/spotted-lanternfly-quarantine-expands-again/

Spotted Lanternfly Quarantine Expands Again

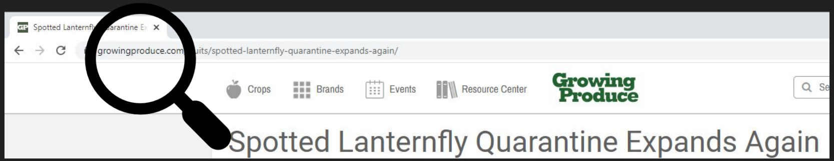


Locate

Fetch

Filter

Distill



Spotted Lanternfl arantine E
growingproduce.com uits/spotted-lanternfly-quarantine-expands-again/
Crops
Brands
Events
Resource Center
Growing Produce
Se

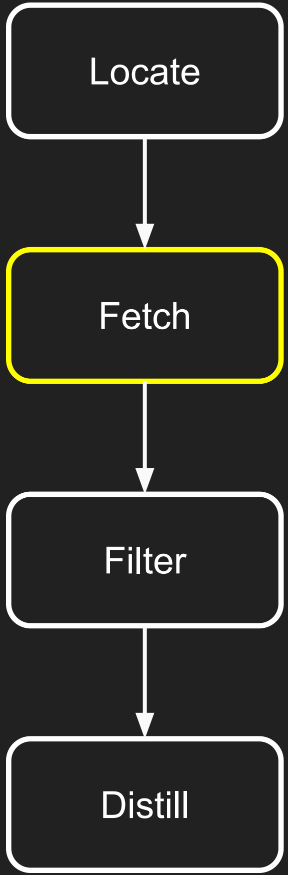
Spotted Lanternfly Quarantine Expands Again

Brandwatch

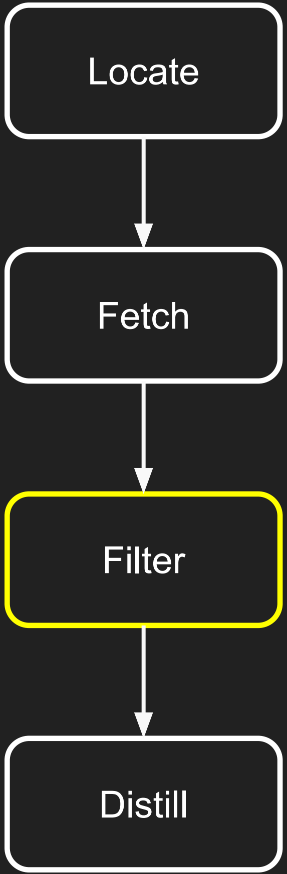
Global Database of Events, Language and Tone (GDELT)

Google News

Twitter API



HTML scraping

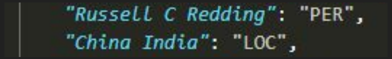


Regex for text matching

spotted*lantern[w]fl* in “https://growingproduce.com/spotted-lanternfly-quarantine-expands-again”?

Natural language processing

Named Entity Recognition & geoparsing



```
    "Russell C Redding": "PER",
    "China India": "LOC",
```

Translation

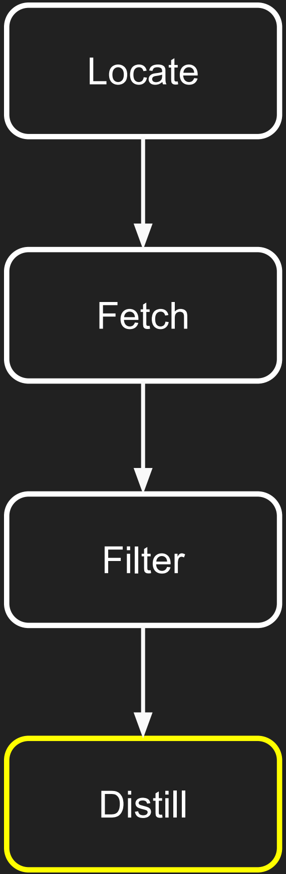
>>> translator.translate('歡迎').text)
welcome

Text classification

tomato blight got my once admired plant. doesn't look like I'll save it : (<—-YES

Seated on a toadstool, the deathflower of the potato blight on her <—-NO

Human-in-the-loop



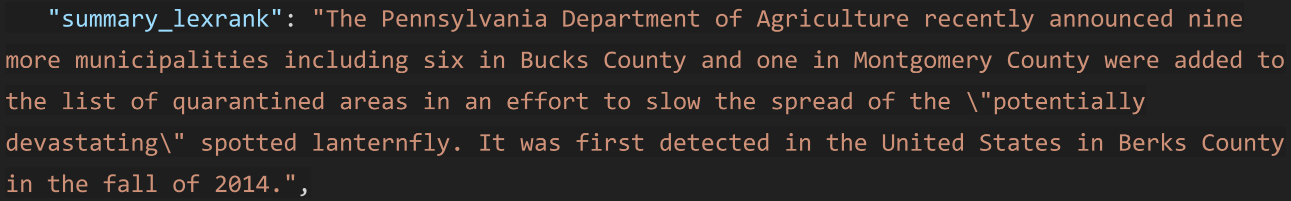
Locate

Fetch

Filter

Distill

Unsupervised summarization



```
  "summary_lexrank": "The Pennsylvania Department of Agriculture recently announced nine
more municipalities including six in Bucks County and one in Montgomery County were added to
the list of quarantined areas in an effort to slow the spread of the \"potentially
devastating\" spotted lanternfly. It was first detected in the United States in Berks County
in the fall of 2014.",
```

Search server

Apache Solr

Text classification

Brandwatch

Visualization