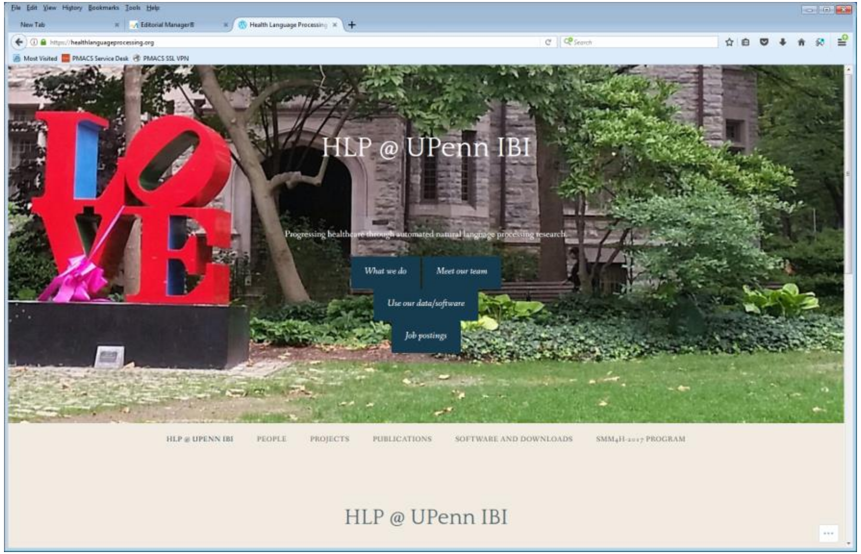
Natural Language Processing for Health (HLP)

September 2018

Tweet @UPennHLP #HLPMeeting



HLP @ UPenn IBI

Progressing healthcare through automated natural language processing research.

HLP @ UPenn IBI

https://healthlanguageprocessing.org

Graciela Gonzalez-Hernandez

Contact: gragon@upenn.edu

Twitter: @gracielagon

Institute for Biomedical Informatics

Perelman School of Medicine
University of Pennsylvania

Program & Speaker List

- **Welcome/Introduction –** Graciela Gonzalez-Hernandez, Ph.D.

- **KEYNOTES *(15 minutes + 15 minutes for questions each)***

Deep neural networks and distant supervision for geographic location mention extraction
Arjun Magee
Dept. of Biomedical Informatics
Arizona State University

Social Media Mining for Pharmacovigilance: challenges and opportunities: Case Control Studies from Twitter?
Graciela Gonzalez-Hernandez, Ph.D
Health Language Processing Lab – Penn IBI
University of Pennsylvania

Deep neural networks and distant supervision for geographic location mention extraction

Arjun Magge [1,2], Davy Weissenbacher [3], Abeed Sarkar [3], Matthew Scotch [1,2], and Graciela Gonzalez [3]

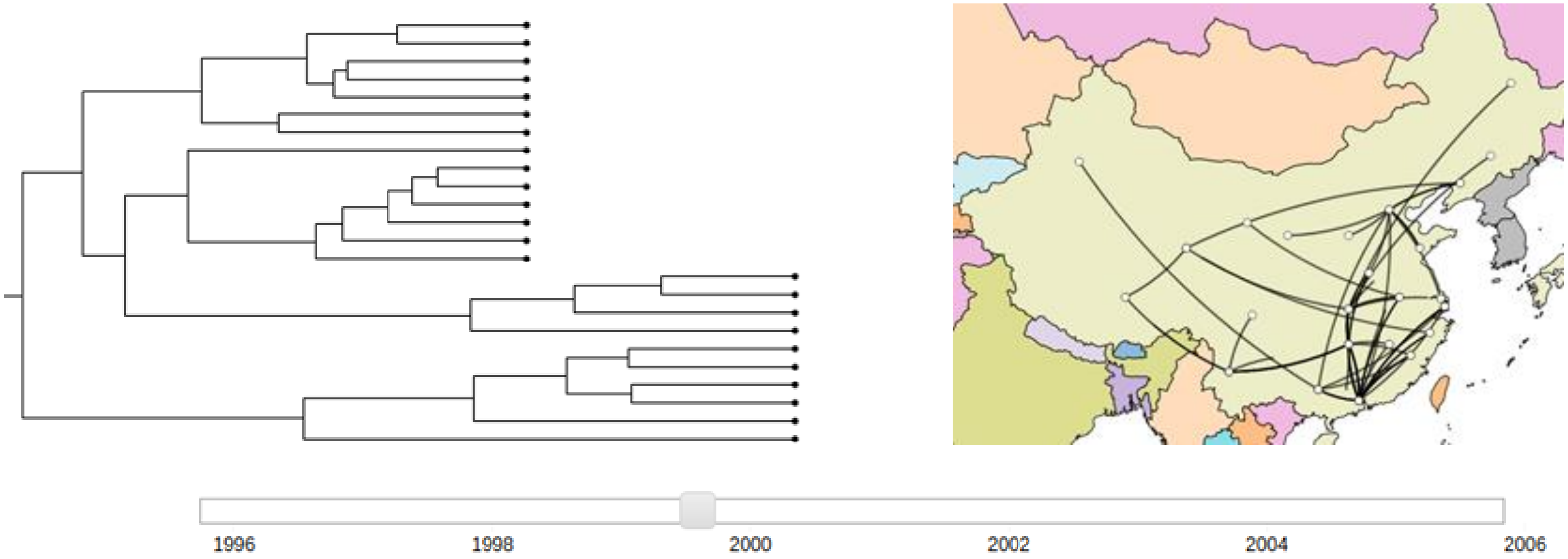
[1] Department of Biomedical Informatics, Arizona State University
[2] Biodesign Center for Environmental Health Engineering, Biodesign Institute, Arizona State University
[3] Department of Biostatistics, Epidemiology and Informatics, The Perelman School of Medicine, University of Pennsylvania

Sept 6, 2018

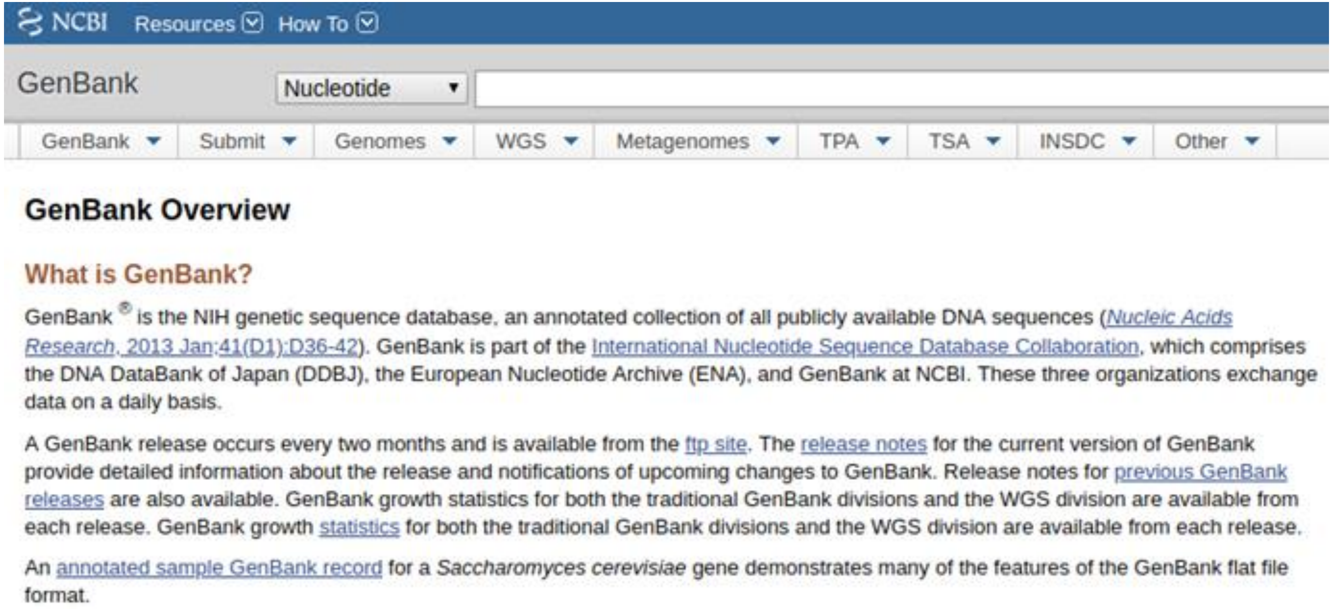
ASU College of Health Solutions
ARIZONA STATE UNIVERSITY

Phylogenetic tree and spread reconstruction

- Virus phylogeography and epidemiology research relies on nucleotide sequence repositories like **GenBank**



GenBank



GenBank Overview

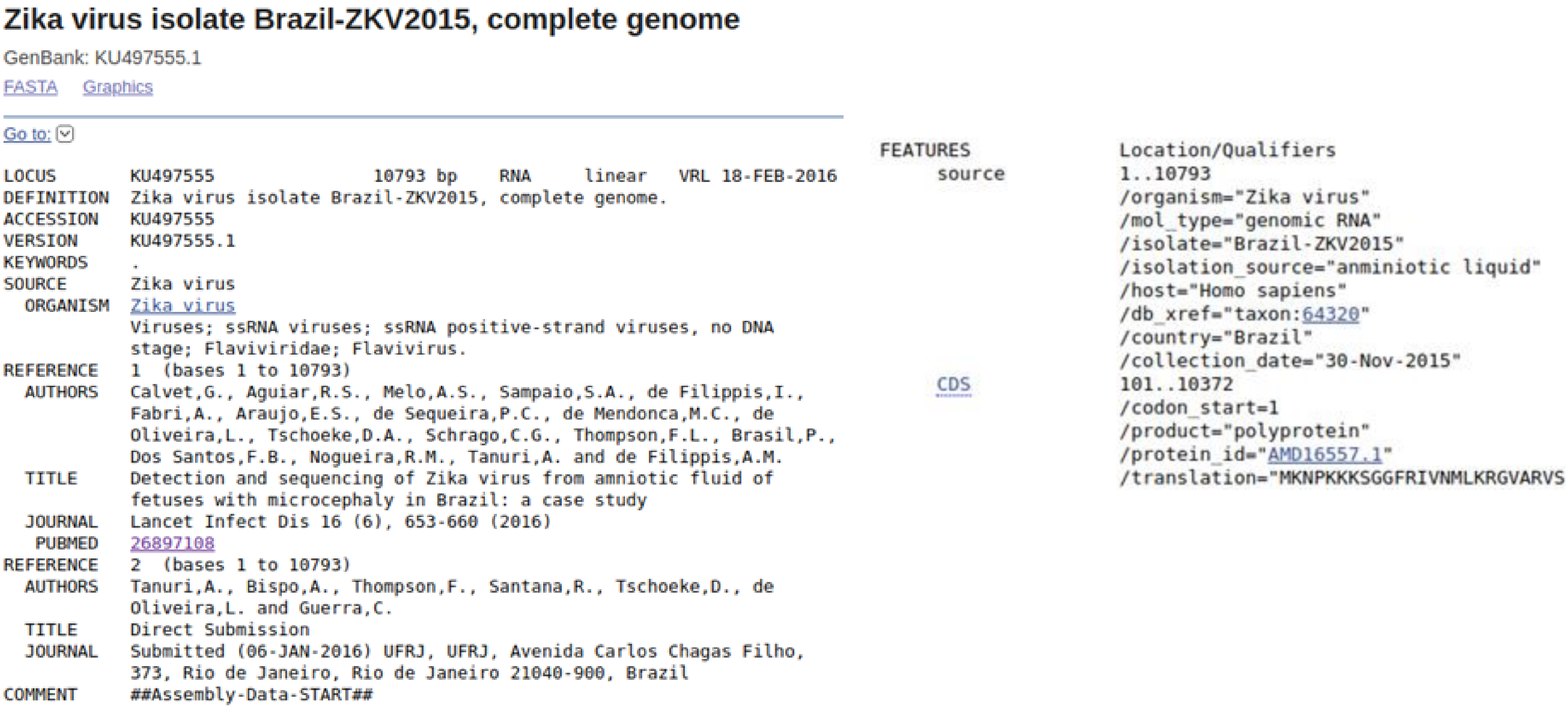
What is GenBank?

GenBank [®] is the NIH genetic sequence database, an annotated collection of all publicly available DNA sequences (*Nucleic Acids Research*, 2013 Jan;41(D1):D36-42). GenBank is part of the International Nucleotide Sequence Database Collaboration, which comprises the DNA DataBank of Japan (DDBJ), the European Nucleotide Archive (ENA), and GenBank at NCBI. These three organizations exchange data on a daily basis.

A GenBank release occurs every two months and is available from the ftp site. The release notes for the current version of GenBank provide detailed information about the release and notifications of upcoming changes to GenBank. Release notes for previous GenBank releases are also available. GenBank growth statistics for both the traditional GenBank divisions and the WGS division are available from each release. GenBank growth statistics for both the traditional GenBank divisions and the WGS division are available from each release.

An annotated sample GenBank record for a *Saccharomyces cerevisiae* gene demonstrates many of the features of the GenBank flat file format.

Genbank Record and Metadata



Zika virus isolate Brazil-ZKV2015, complete genome

GenBank: KU497555.1

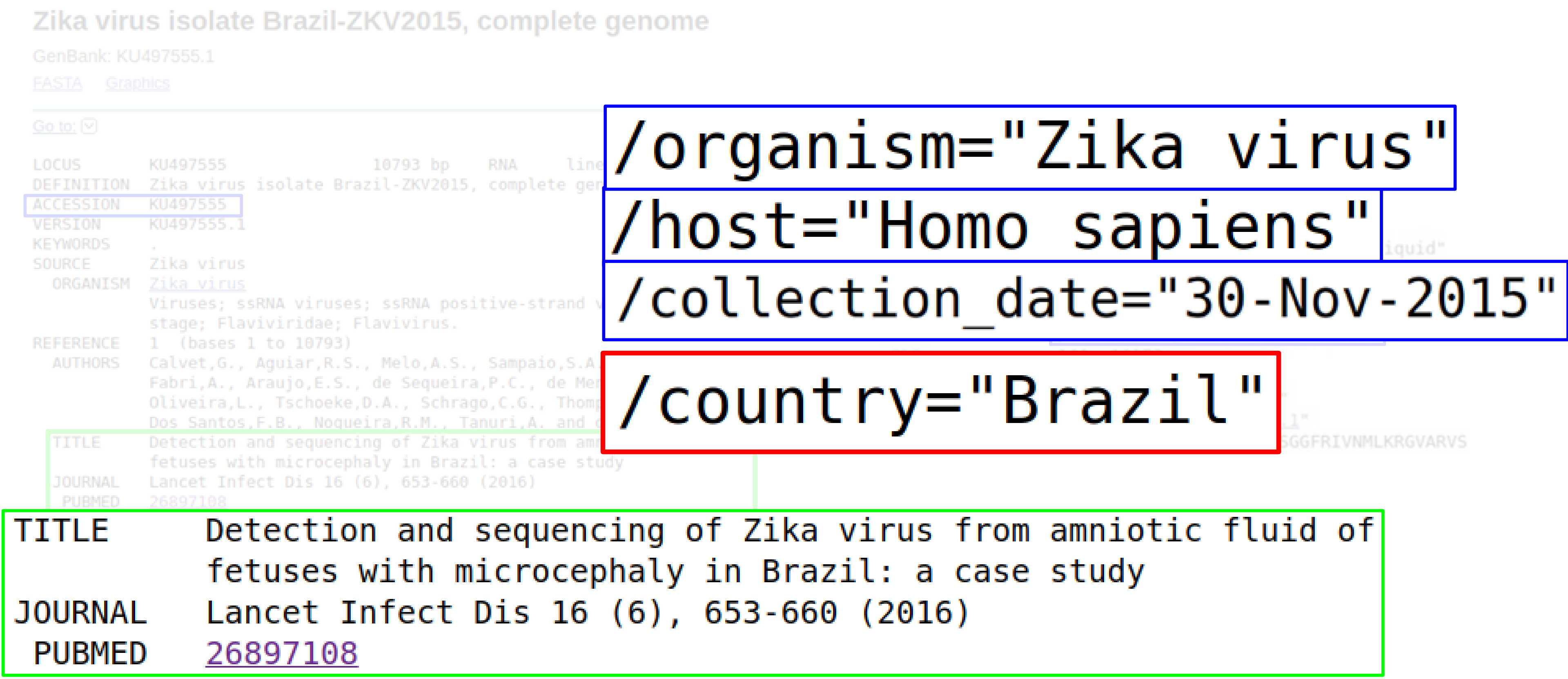
FASTA Graphics

Go to:

```
LOCUS       KU497555               10793 bp    RNA     linear   VRL 18-FEB-2016
DEFINITION  Zika virus isolate Brazil-ZKV2015, complete genome.
ACCESSION   KU497555
VERSION     KU497555.1
KEYWORDS    .
SOURCE      Zika virus
  ORGANISM  Zika virus
            Viruses; ssRNA viruses; ssRNA positive-strand viruses, no DNA
            stage; Flaviviridae; Flavivirus.
REFERENCE   1  (bases 1 to 10793)
  AUTHORS   Calvet,G., Aguiar,R.S., Melo,A.S., Sampaio,S.A., de Filippis,I.,
            Fabri,A., Araujo,E.S., de Sequeira,P.C., de Mendonca,M.C., de
            Oliveira,L., Tschoeke,D.A., Schrago,C.G., Thompson,F.L., Brasil,P.,
            Dos Santos,F.B., Nogueira,R.M., Tanuri,A. and de Filippis,A.M.
  TITLE     Detection and sequencing of Zika virus from amniotic fluid of
            fetuses with microcephaly in Brazil: a case study
  JOURNAL   Lancet Infect Dis 16 (6), 653-660 (2016)
   PUBMED   26897108
REFERENCE   2  (bases 1 to 10793)
  AUTHORS   Tanuri,A., Bispo,A., Thompson,F., Santana,R., Tschoeke,D., de
            Oliveira,L. and Guerra,C.
  TITLE     Direct Submission
  JOURNAL   Submitted (06-JAN-2016) UFRJ, UFRJ, Avenida Carlos Chagas Filho,
            373, Rio de Janeiro, Rio de Janeiro 21040-900, Brazil
COMMENT     ##Assembly-Data-START##
```

```
FEATURES             Location/Qualifiers
     source          1..10793
                     /organism="Zika virus"
                     /mol_type="genomic RNA"
                     /isolate="Brazil-ZKV2015"
                     /isolation_source="anminiotic liquid"
                     /host="Homo sapiens"
                     /db_xref="taxon:64320"
                     /country="Brazil"
                     /collection_date="30-Nov-2015"
     CDS             101..10372
                     /codon_start=1
                     /product="polyprotein"
                     /protein_id="AMD16557.1"
                     /translation="MKNPKKKSGGFRIVNMLKRGVARVS
```

Genbank Record and Metadata



Zika virus isolate Brazil-ZKV2015, complete genome

GenBank: KU497555.1

FASTA Graphics

Go to:

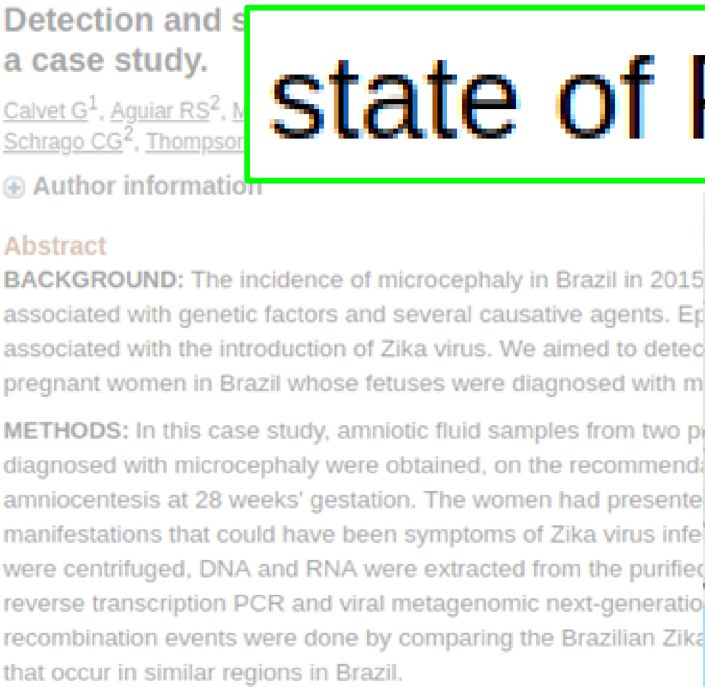
```
LOCUS       KU497555               10793 bp    RNA     line
DEFINITION  Zika virus isolate Brazil-ZKV2015, complete gen
ACCESSION   KU497555
VERSION     KU497555.1
KEYWORDS    .
SOURCE      Zika virus
  ORGANISM  Zika virus
            Viruses; ssRNA viruses; ssRNA positive-strand v
            stage; Flaviviridae; Flavivirus.
REFERENCE   1  (bases 1 to 10793)
  AUTHORS   Calvet,G., Aguiar,R.S., Melo,A.S., Sampaio,S.A.
            Fabri,A., Araujo,E.S., de Sequeira,P.C., de Men
            Oliveira,L., Tschoeke,D.A., Schrago,C.G., Thomp
            Dos Santos,F.B., Nogueira,R.M., Tanuri,A. and d
  TITLE     Detection and sequencing of Zika virus from amn
            fetuses with microcephaly in Brazil: a case study
  JOURNAL   Lancet Infect Dis 16 (6), 653-660 (2016)
   PUBMED   26897108
```

```
iquid"
.1"
SGGFRIVNMLKRGVARVS
```

```
/organism="Zika virus"
/host="Homo sapiens"
/collection_date="30-Nov-2015"
/country="Brazil"
```

```
TITLE     Detection and sequencing of Zika virus from amniotic fluid of
          fetuses with microcephaly in Brazil: a case study
JOURNAL   Lancet Infect Dis 16 (6), 653-660 (2016)
 PUBMED   26897108
```

PubMed Article



Detection and s
a case study.

Calvet G[1], Aguiar RS[2], M
Schrago CG[2], Thompson

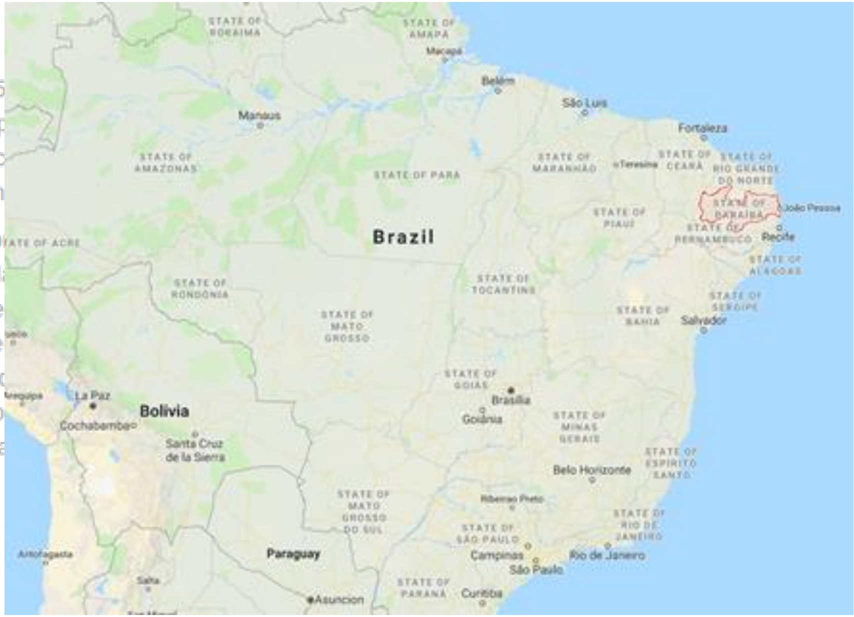
⊕ Author information

Abstract

BACKGROUND: The incidence of microcephaly in Brazil in 2015
associated with genetic factors and several causative agents. Ep
associated with the introduction of Zika virus. We aimed to detec
pregnant women in Brazil whose fetuses were diagnosed with m

METHODS: In this case study, amniotic fluid samples from two p
diagnosed with microcephaly were obtained, on the recommenda
amniocentesis at 28 weeks' gestation. The women had presente
manifestations that could have been symptoms of Zika virus infe
were centrifuged, DNA and RNA were extracted from the purifie
reverse transcription PCR and viral metagenomic next-generatio
recombination events were done by comparing the Brazilian Zika
that occur in similar regions in Brazil.

state of Paraíba in Brazil



Insufficient location information

Problem: Locations in GenBank metadata are not sufficient for Phylogeography research

- Especially for countries like USA, Canada, Russia, China, Brazil

Solution: Enrich location information in GenBank by extracting locations from the associated PubMed article using Natural Language Processing (NLP)

Natural Language Processing

1. Named Entity Recognition
   - identifying words of interest in text (usually nouns)
   - e.g. names, genes, proteins, locations, organizations, time, etc.

1. Concept Resolution
   - perform disambiguation by assigning a unique gazetteer ID
   - e.g. Paris can refer to Paris, Texas, USA or Paris, France

1. Determine Location of Infected Host (LOIH)
   - assign probabilities to all identified locations using heuristics

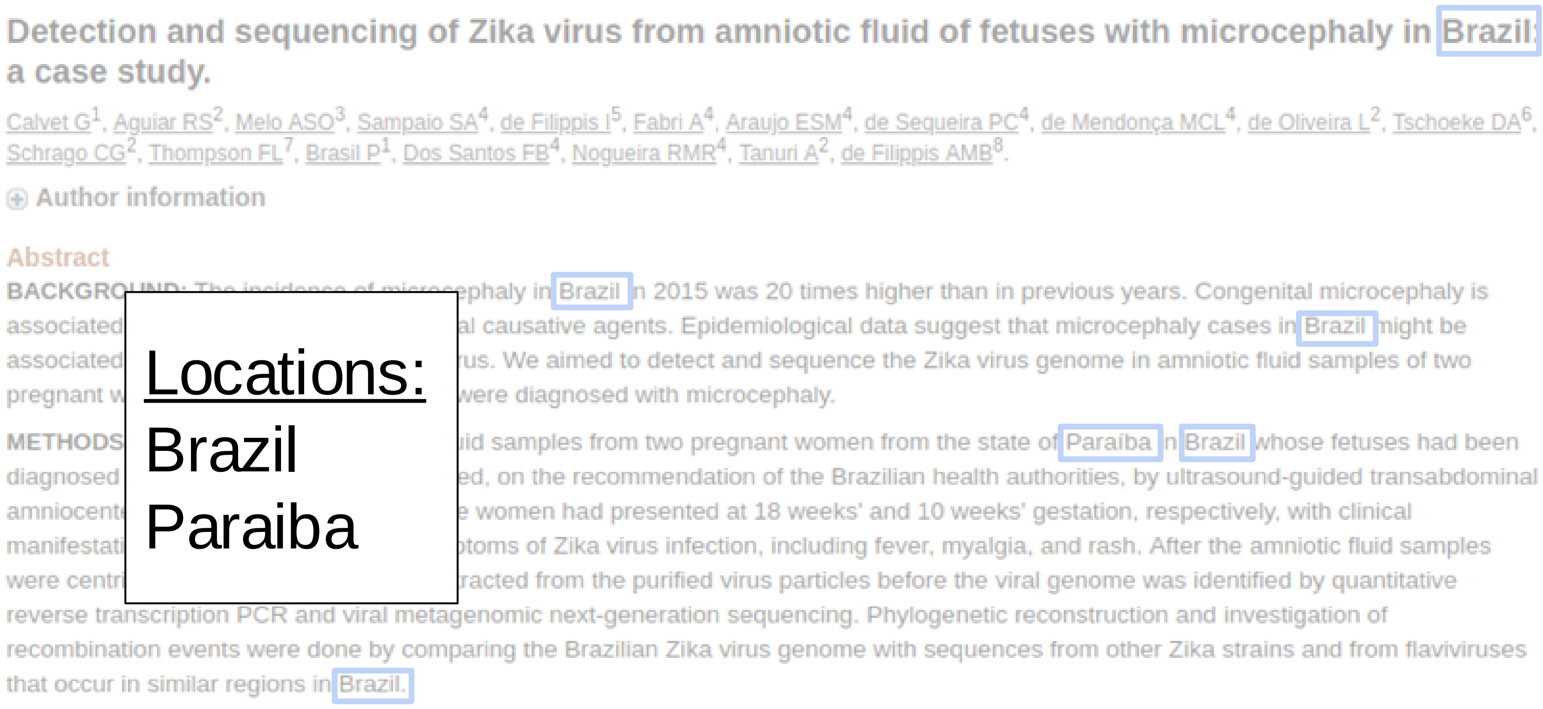
Natural Language Processing

1. Named Entity Recognition
   - identifying words of interest in text (usually nouns)
   - e.g. names, genes, proteins, locations, organizations, time, etc.

1. Concept Resolution
   - perform disambiguation by assigning a unique gazetteer ID
   - e.g. Paris can refer to Paris, Texas, USA or Paris, France

1. Determine Location of Infected Host (LOIH)
   - assign probabilities to all identified locations using heuristics

PubMed Article : NER



Detection and sequencing of Zika virus from amniotic fluid of fetuses with microcephaly in Brazil: a case study.

Calvet G[1], Aguiar RS[2], Melo ASO[3], Sampaio SA[4], de Filippis I[5], Fabri A[4], Araujo ESM[4], de Sequeira PC[4], de Mendonça MCL[4], de Oliveira L[2], Tschoeke DA[6], Schrago CG[2], Thompson FL[7], Brasil P[1], Dos Santos FB[4], Nogueira RMR[4], Tanuri A[2], de Filippis AMB[8].

⊕ Author information

Abstract

BACKGROUND: The incidence of microcephaly in Brazil in 2015 was 20 times higher than in previous years. Congenital microcephaly is
associated
al causative agents. Epidemiological data suggest that microcephaly cases in Brazil might be
associated
rus. We aimed to detect and sequence the Zika virus genome in amniotic fluid samples of two
pregnant w
were diagnosed with microcephaly.

METHODS
uid samples from two pregnant women from the state of Paraíba in Brazil whose fetuses had been
diagnosed
ed, on the recommendation of the Brazilian health authorities, by ultrasound-guided transabdominal
amniocente
e women had presented at 18 weeks' and 10 weeks' gestation, respectively, with clinical
manifestati
ptoms of Zika virus infection, including fever, myalgia, and rash. After the amniotic fluid samples
were centri
racted from the purified virus particles before the viral genome was identified by quantitative
reverse transcription PCR and viral metagenomic next-generation sequencing. Phylogenetic reconstruction and investigation of recombination events were done by comparing the Brazilian Zika virus genome with sequences from other Zika strains and from flaviviruses that occur in similar regions in Brazil.

<u>Locations:</u>

Brazil

Paraiba

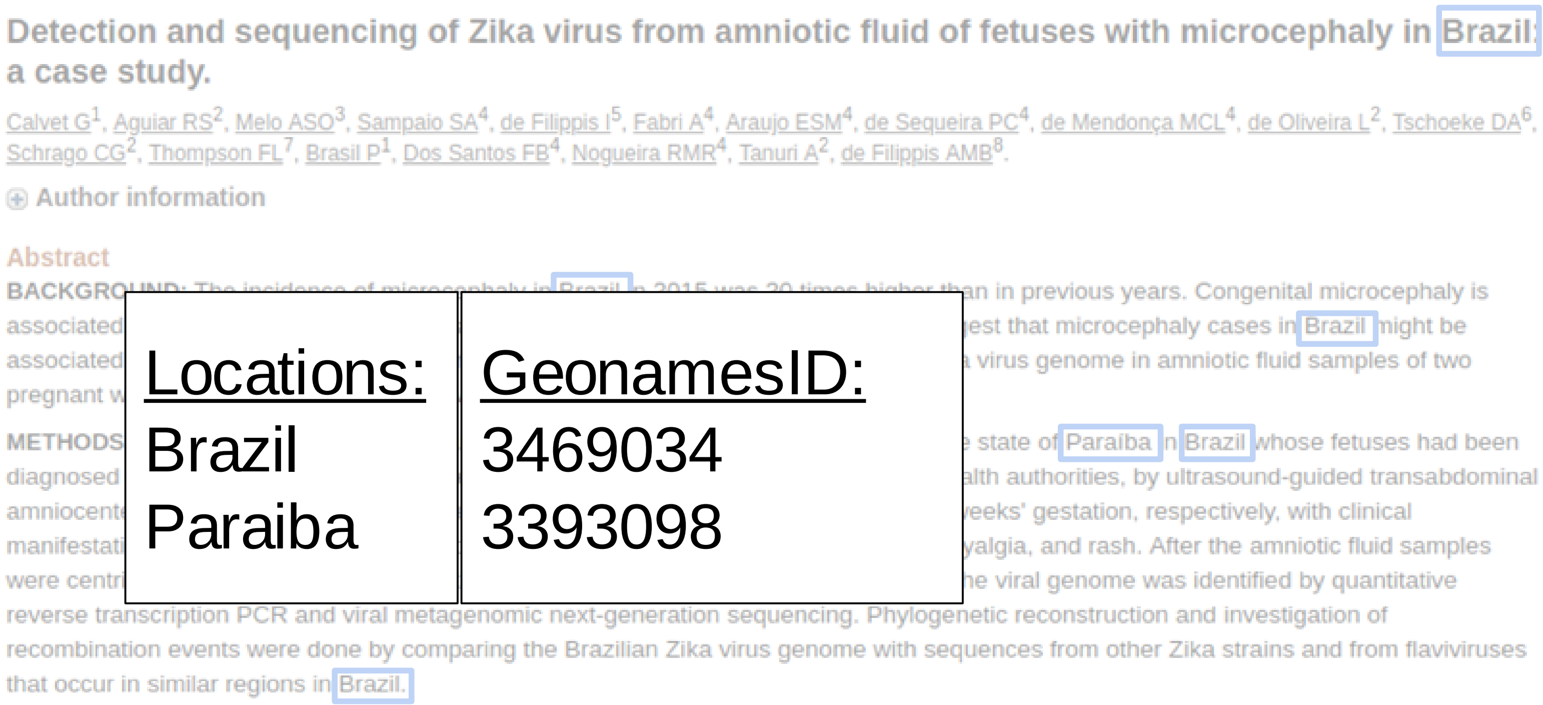
Natural Language Processing

1. Named Entity Recognition
   - identifying words of interest in text (usually nouns)
   - e.g. names, genes, proteins, locations, organizations, time, etc.

1. Concept Resolution
   - perform disambiguation by assigning a unique gazetteer ID
   - e.g. Paris can refer to Paris, Texas, USA or Paris, France

1. Determine Location of Infected Host (LOIH)
   - assign probabilities to all identified locations using heuristics

PubMed Article : NER



Detection and sequencing of Zika virus from amniotic fluid of fetuses with microcephaly in Brazil: a case study.

Calvet G[1], Aguiar RS[2], Melo ASO[3], Sampaio SA[4], de Filippis I[5], Fabri A[4], Araujo ESM[4], de Sequeira PC[4], de Mendonça MCL[4], de Oliveira L[2], Tschoeke DA[6], Schrago CG[2], Thompson FL[7], Brasil P[1], Dos Santos FB[4], Nogueira RMR[4], Tanuri A[2], de Filippis AMB[8].

⊕ Author information

Abstract

BACKGROUND: The incidence of microcephaly in Brazil in 2015 was 20 times higher than in previous years. Congenital microcephaly is
associated
gest that microcephaly cases in Brazil might be
associated
a virus genome in amniotic fluid samples of two
pregnant w

METHODS
e state of Paraiba in Brazil whose fetuses had been
diagnosed
alth authorities, by ultrasound-guided transabdominal
amniocent
eeks' gestation, respectively, with clinical
manifestati
yalgia, and rash. After the amniotic fluid samples
were centri
he viral genome was identified by quantitative
reverse transcription PCR and viral metagenomic next-generation sequencing. Phylogenetic reconstruction and investigation of recombination events were done by comparing the Brazilian Zika virus genome with sequences from other Zika strains and from flaviviruses that occur in similar regions in Brazil.

Locations:	GeonamesID:
Brazil	3469034
Paraiba	3393098

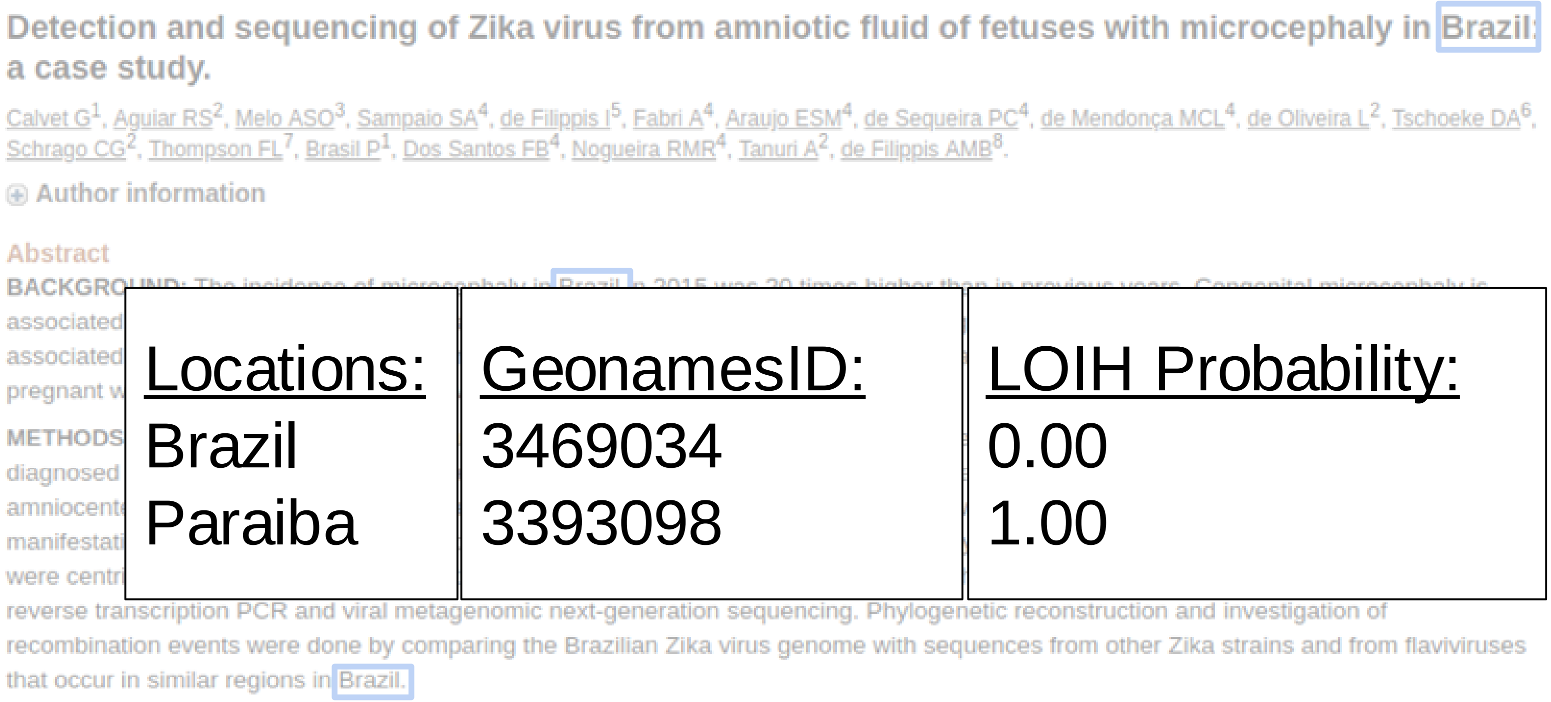
Natural Language Processing

1. Named Entity Recognition
   - identifying words of interest in text (usually nouns)
   - e.g. names, genes, proteins, locations, organizations, time, etc.

1. Concept Resolution
   - perform disambiguation by assigning a unique gazetteer ID
   - e.g. Paris can refer to Paris, Texas, USA or Paris, France

1. Determine Location of Infected Host (LOIH)
   - assign probabilities to all identified locations using heuristics

PubMed Article : NER



Detection and sequencing of Zika virus from amniotic fluid of fetuses with microcephaly in Brazil: a case study.

Calvet G[1], Aguiar RS[2], Melo ASO[3], Sampaio SA[4], de Filippis I[5], Fabri A[4], Araujo ESM[4], de Sequeira PC[4], de Mendonça MCL[4], de Oliveira L[2], Tschoeke DA[6], Schrago CG[2], Thompson FL[7], Brasil P[1], Dos Santos FB[4], Nogueira RMR[4], Tanuri A[2], de Filippis AMB[8].

⊕ Author information

Abstract

BACKGROUND: The incidence of microcephaly in Brazil in 2015 was 20 times higher than in previous years. Congenital microcephaly is

associated

associated

pregnant w

METHODS

diagnosed

amniocente

manifestati

were centri

reverse transcription PCR and viral metagenomic next-generation sequencing. Phylogenetic reconstruction and investigation of recombination events were done by comparing the Brazilian Zika virus genome with sequences from other Zika strains and from flaviviruses that occur in similar regions in Brazil.

Locations:	GeonamesID:	LOIH Probability:
Brazil	3469034	0.00
Paraiba	3393098	1.00

Natural Language Processing

1. Named Entity Recognition
   - identifying words of interest in text (usually nouns)
   - e.g. names, genes, proteins, locations, organizations, time, etc.

1. Concept Resolution
   - perform disambiguation by assigning a unique gazetteer ID
   - e.g. Paris can refer to Paris, Texas, USA or Paris, France

1. Determine Location of Infected Host (LOIH)
   - assign probabilities to all identified locations using heuristics

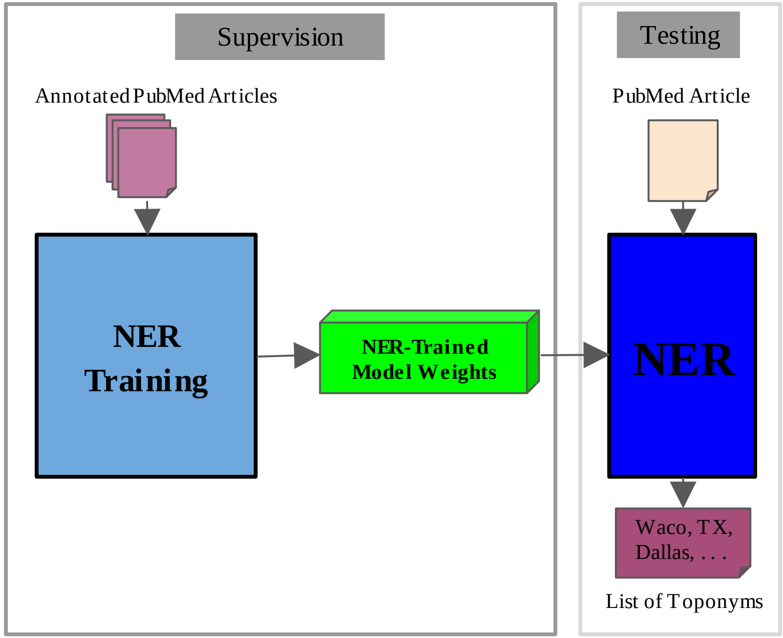
Ambiguity in Natural Language

*in **May, Russia** in 2010.*	✓
*found in **May** 2013.*	X
*pigs, **turkey** and quail*	X
*University of **Las Vegas**.*	X

Why deep neural nets?

- Rule based systems
- Machine Learning and Deep learning
  - Better performance with more annotated data

- Most times, you can only annotate a few articles.
  - So, distant supervision?

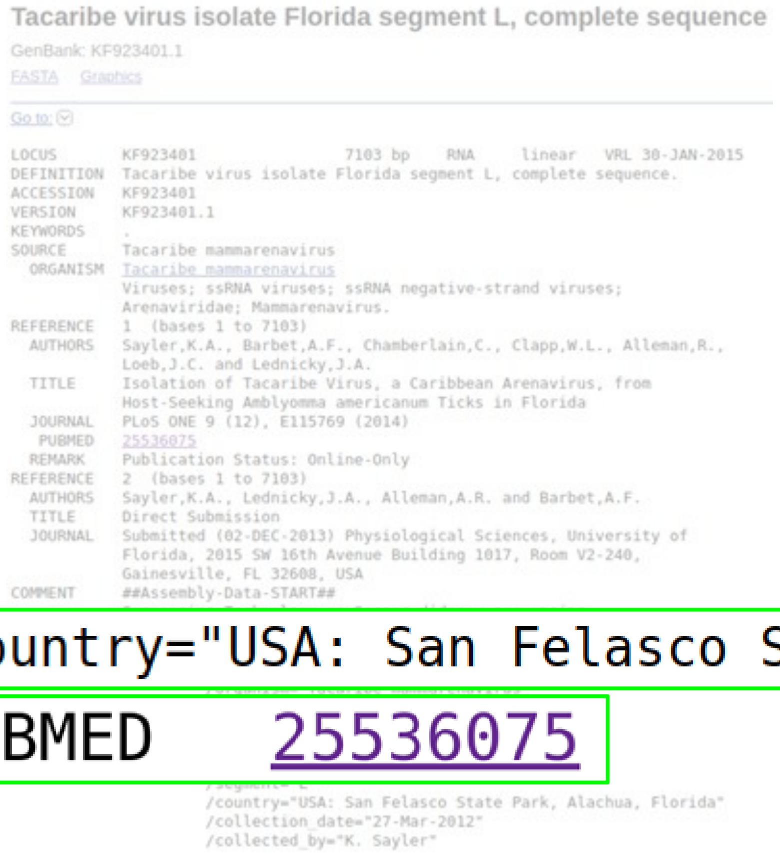


Dataset

- A set of 60 full-text articles (~300,000 words) from Pubmed containing 1881 location annotations
  - 48 for training and 12 for testing

- Distant supervision
  - Use GenBank articles where locations are known
  - Generate positive and negative examples based on rules
  - They are noisy! But, that’s okay.
  - We use them to generate ~8 million training instances(words)

Collecting Distant Supervision Samples



Tacaribe virus isolate Florida segment L, complete sequence

GenBank: KF923401.1

FASTA Graphics

Go to:

```
LOCUS       KF923401                7103 bp    RNA     linear   VRL 30-JAN-2015
DEFINITION  Tacaribe virus isolate Florida segment L, complete sequence.
ACCESSION   KF923401
VERSION     KF923401.1
KEYWORDS    .
SOURCE      Tacaribe mammarenavirus
  ORGANISM  Tacaribe mammarenavirus
            Viruses; ssRNA viruses; ssRNA negative-strand viruses;
            Arenaviridae; Mammarenavirus.
REFERENCE   1  (bases 1 to 7103)
  AUTHORS   Sayler,K.A., Barbet,A.F., Chamberlain,C., Clapp,W.L., Alleman,R.,
            Loeb,J.C. and Lednicky,J.A.
  TITLE     Isolation of Tacaribe Virus, a Caribbean Arenavirus, from
            Host-Seeking Amblyomma americanum Ticks in Florida
  JOURNAL   PLoS ONE 9 (12), E115769 (2014)
   PUBMED   25536075
  REMARK    Publication Status: Online-Only
REFERENCE   2  (bases 1 to 7103)
  AUTHORS   Sayler,K.A., Lednicky,J.A., Alleman,A.R. and Barbet,A.F.
  TITLE     Direct Submission
  JOURNAL   Submitted (02-DEC-2013) Physiological Sciences, University of
            Florida, 2015 SW 16th Avenue Building 1017, Room V2-240,
            Gainesville, FL 32608, USA
COMMENT     ##Assembly-Data-START##
                     /country="USA: San Felasco State Park, Alachua, Florida"
                     /collection_date="27-Mar-2012"
                     /collected_by="K. Sayler"
```

Process records which have fine-grained locations

/country="USA: San Felasco State Park, Alachua, Florida"

PUBMED 25536075

Collecting Positive Samples

Isolation of Tacaribe Virus, a Caribbean Arenavirus, from Host-Seeking Amblyomma americanum Ticks in Florida. We report the re-isolation of the virus from a pool of 100 host-seeking Amblyomma americanum (lone star ticks) collected in a Florida state park in 2012.

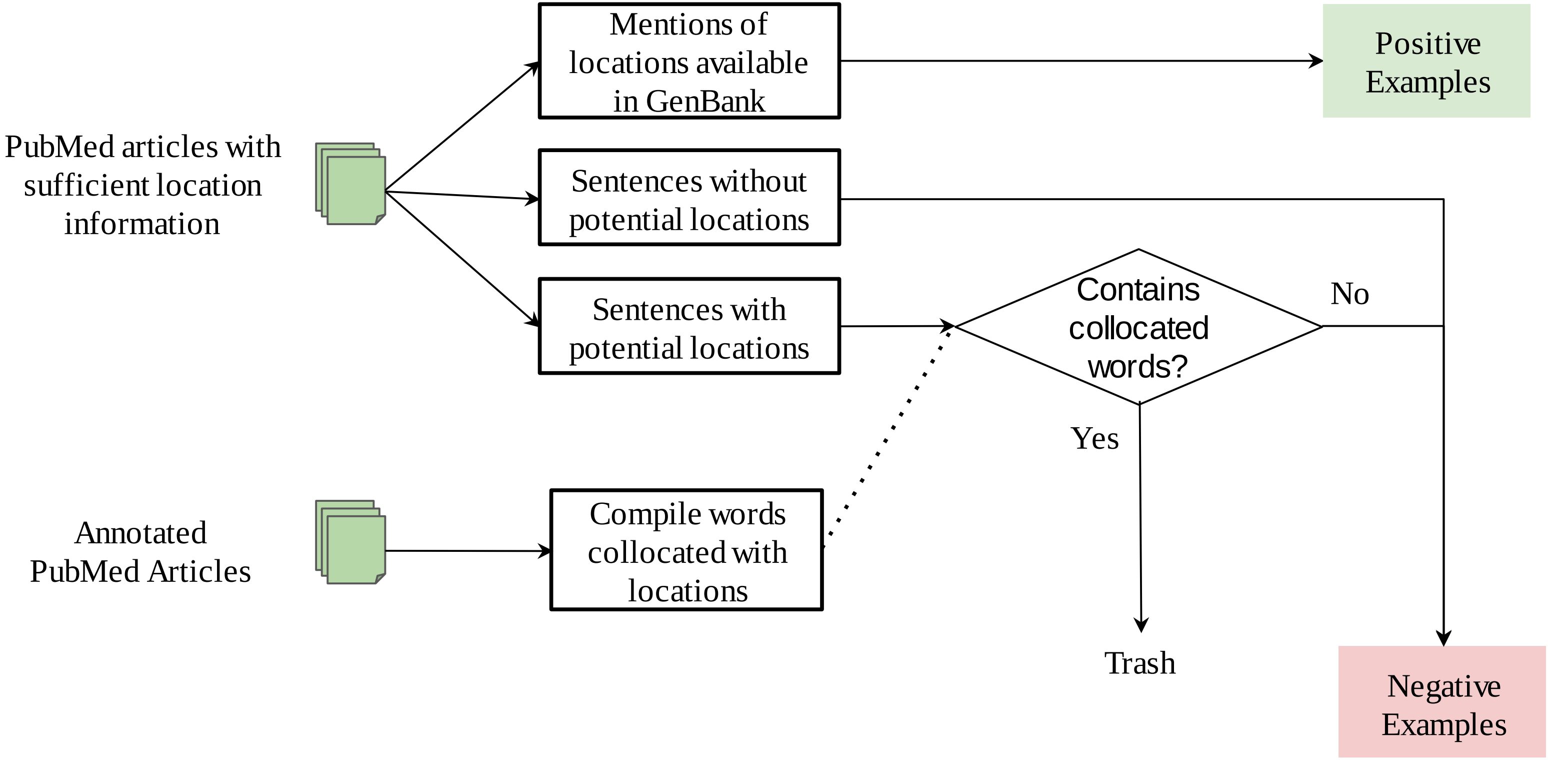
At least ten of these viruses are associated with human disease in many parts of the world including western Africa, Argentina, Bolivia, Venezuela and Brazil [2].

All tick trapping was performed in accordance with the Florida Department of Environmental Protection Research and Collection Permit #05231210.

Only host-seeking tick species common in Florida were collected because these species are most likely to attach to a person and take a blood meal.
In 2013, ticks were collected using the same methods as 2012 from two additional Florida state parks: Manatee Springs State Park in Chiefland, Florida (29° 29′47.401″ N, 82°58′4.429″ W) and O'Leno State Park in High Springs, Florida (29°55′11.863″ N and 82° 35′15.427″ W), to determine if the virus could be detected in other locations in North Central Florida (Fig.1).

A total of 500 host-seeking ticks were collected from three state parks located in North Central Florida, including the original field site where ticks were trapped for virus isolation attempts (Fig. 1).

Collecting Negative Samples



PubMed articles with sufficient location information

Mentions of locations available in GenBank

Positive Examples

Sentences without potential locations

Sentences with potential locations

Contains collocated words?

No

Yes

Trash

Negative Examples

Annotated PubMed Articles

Compile words collocated with locations

Filter them based on some guidelines

Positive Examples

Ticks in Florida <PAD> <PAD>

in a Florida state park

~~with the Florida Department of~~

common in Florida were collected

two additional Florida state parks

Chiefland, Florida (29

Springs , Florida (29

North Central Florida (Fig

North Central Florida, including

~~University of Florida Interdisciplinary Center~~

in Central Florida , USA

Negative Examples

<PAD> <PAD> Gene UL111A encodes

<PAD> Gene UL111A encodes viral

Gene UL111A encodes viral interleukin

UL111A encodes viral interleukin -

encodes viral interleukin - 10

viral interleukin - 10 (

interleukin - 10 (Lockridge

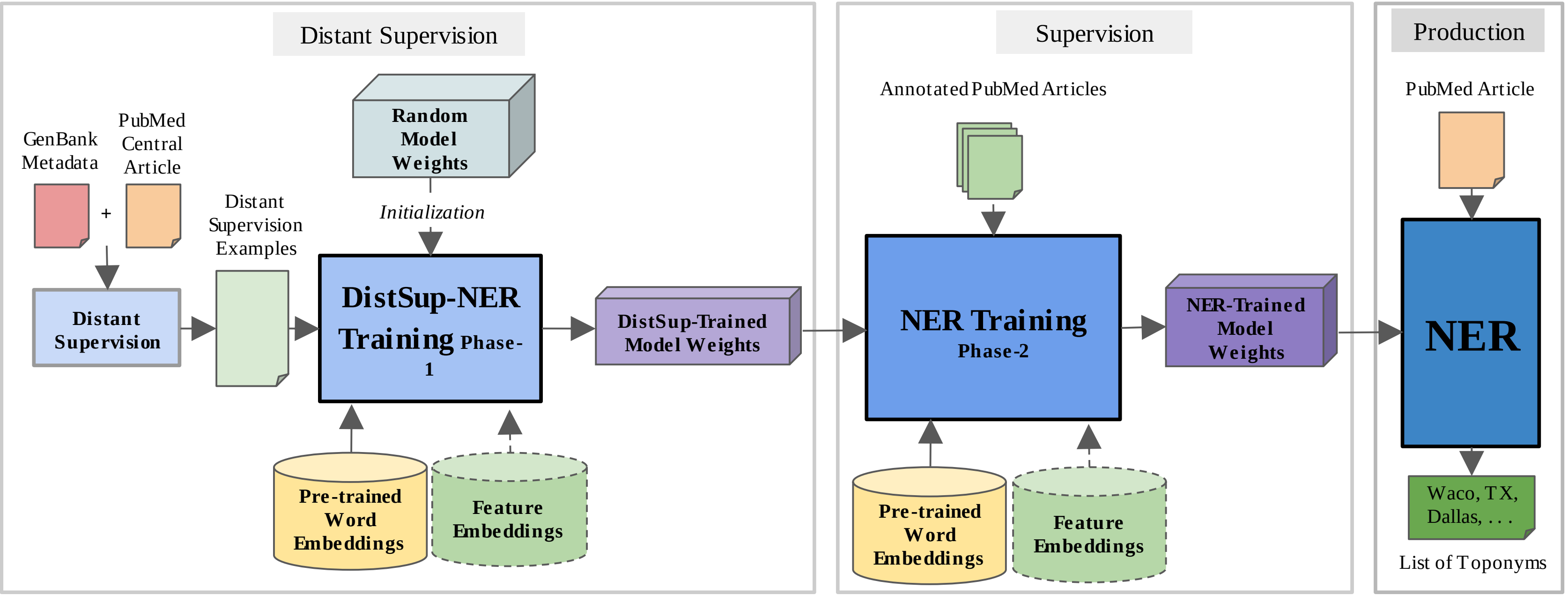
- 10 (Lockridge et

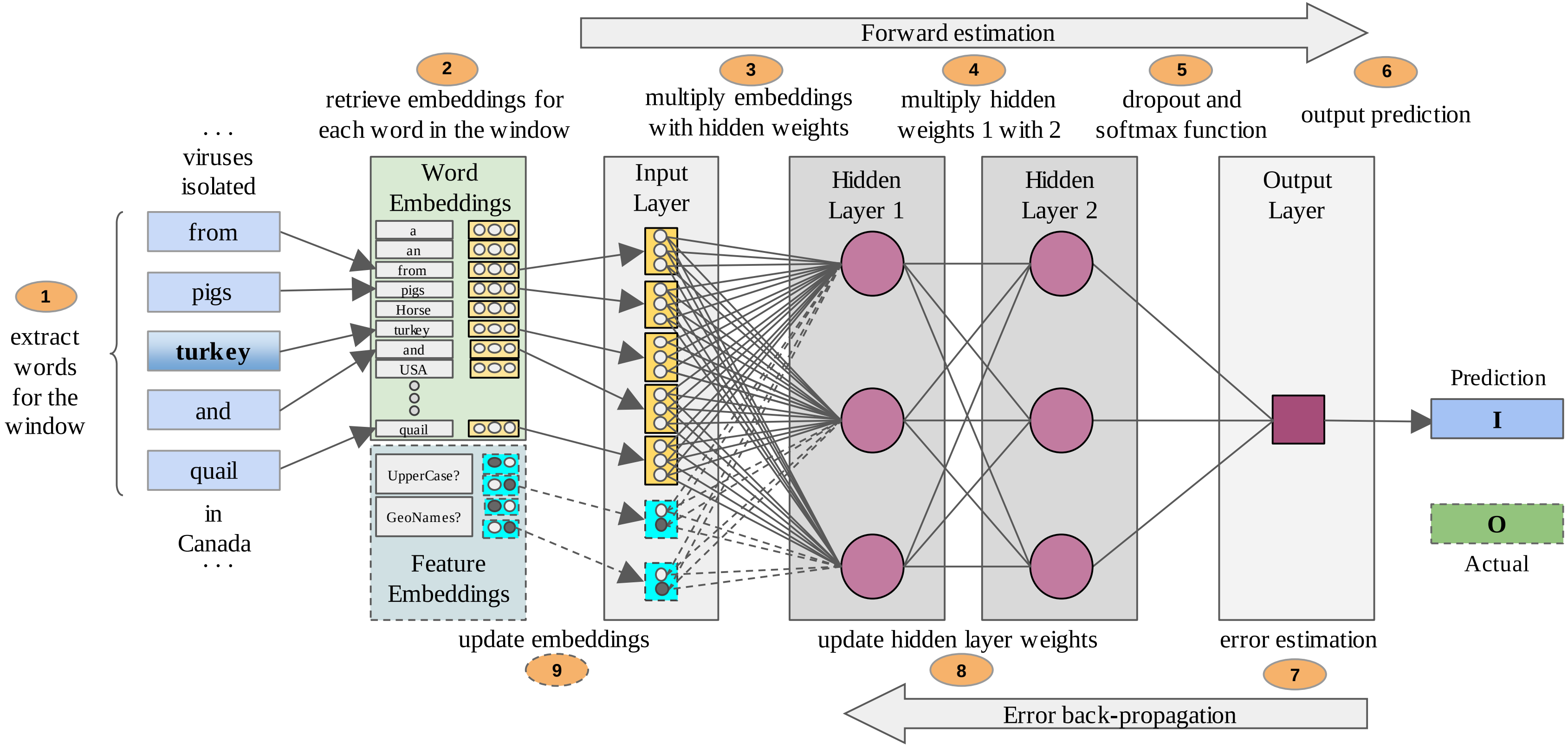
10 (Lockridge et al

(Lockridge et al .

Lockridge et al . ,

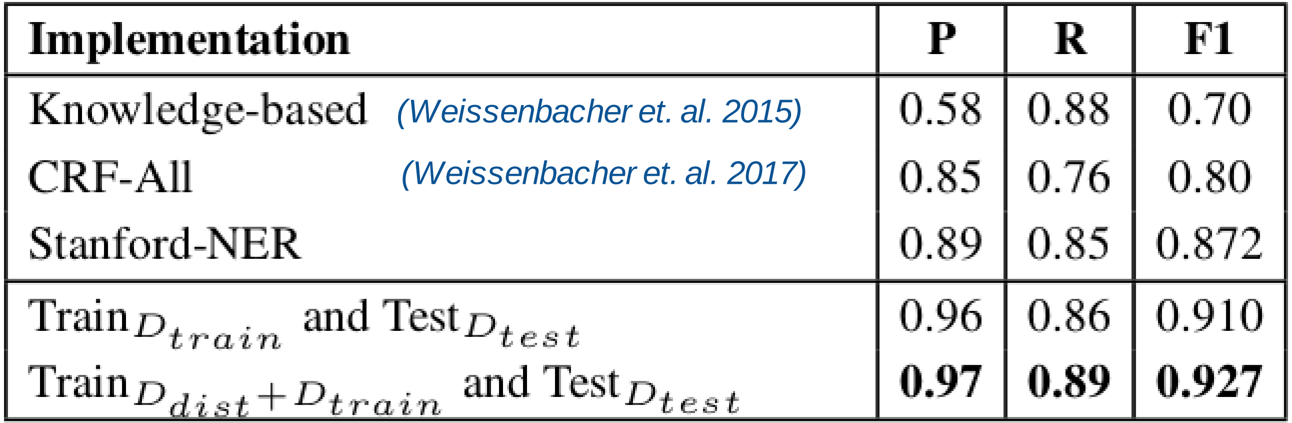
Layered training





Collobert, Ronan, et al. "Natural language processing (almost) from scratch." *Journal of Machine Learning Research* 12.Aug (2011): 2493-2537.

Improved performance



Implementation	**P**	**R**	**F1**
Knowledge-based *(Weissenbacher et. al. 2015)*	0.58	0.88	0.70
CRF-All *(Weissenbacher et. al. 2017)*	0.85	0.76	0.80
Stanford-NER	0.89	0.85	0.872
Train D_{train} and Test D_{test}	0.96	0.86	0.910
Train $D_{dist}+D_{train}$ and Test D_{test}	**0.97**	**0.89**	**0.927**

92.7% on tokenwise evaluation

91.5% on strict evaluation

Limitations and Future Work

- Potential for improving performance
  - Deal with table data
  - Second layer of supervision for trying advanced recurrent models like Bi-LSTM-CRFs
- What is the improvement to resolution/normalization?
- Any improvements to the phylogeographic models?
- Distant supervision *and* Supervision - a systematic analysis of how much data for both is sufficient
- Validate with other entities like hosts, virus, genes etc.

One last thing . . . since we are at ISMB

Tell your peers:

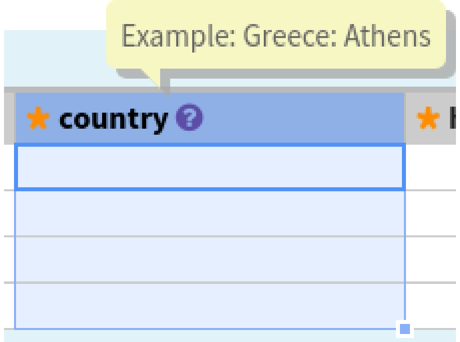
Even though the field says *country*

please add in addition to *country* information:

- state
- county (if available)
- city (if available)

U.S. National Library of Medicine
National Center for Biotechnology Information

Submission Portal



Softwares and Applications

NER (source code) : https://github.com/amagge/ner-topo-ff

GeoBoost v1 : https://tinyurl.com/geoboost (Tahsin *et. al. 2017*)

ZoDo (under development): https://zodo.asu.edu/zodo

ZooPhy (under development): https://zodo.asu.edu/zoophy

Acknowledgments

Bio-Ontologies

For the travel grant to present at ISMB

Funding and Acknowledgments

Principal Investigators:

Dr. Matthew Scotch

Dr. Graciela Gonzalez

Collaborators:

Dr. Davy Weissenbacher

Dr. Abeed Sarker

Annotators:

Karen O'Connor

Megan Rorison

Briana Trevino

NIH National Institute of Allergy and Infectious Diseases

ASU Biodesign Institute Arizona State University

ASU Department of Biomedical Informatics Arizona State University

Penn

DEPARTMENT of BIOSTATISTICS EPIDEMIOLOGY & INFORMATICS

NVIDIA.

Thank you!

Questions?

Social Media Mining for Pharmacovigilance: challenges and opportunities

Case-control studies from Twitter???

Health Language Processing Lab – Penn IBI

Graciela Gonzalez-Hernandez, PhD

email: gragon@pennmedicine.upenn.edu

@gracielagon

Funded by NLM/NIH Grant R01LM011176

Perelman School of Medicine
UNIVERSITY *of* PENNSYLVANIA

SM data for pharmacovigilance studies

- **There are about 38,220 tweets / minute about the user's current medical conditions**[1,2,3]
- **Patient reporting brings different perspective**, more detail, info on severity and impact of ADRs in daily life. (34 studies - PMID 27558545).
- **Abundant adverse event reports in SM, with a** higher frequency of adverse events, particularly for 'mild' adverse events. (51 studies = PMID 26271492).

amazon
DailyStrength
Drugs.com

[1] http://www.pewinternet.org/fact-sheets/health-fact-sheet/
[2] http://www.statista.com/statistics/282087/number-of-monthly-active-twitter-users/
[3] http://www.internetlive**stats**.com/twitter-**statistics**/

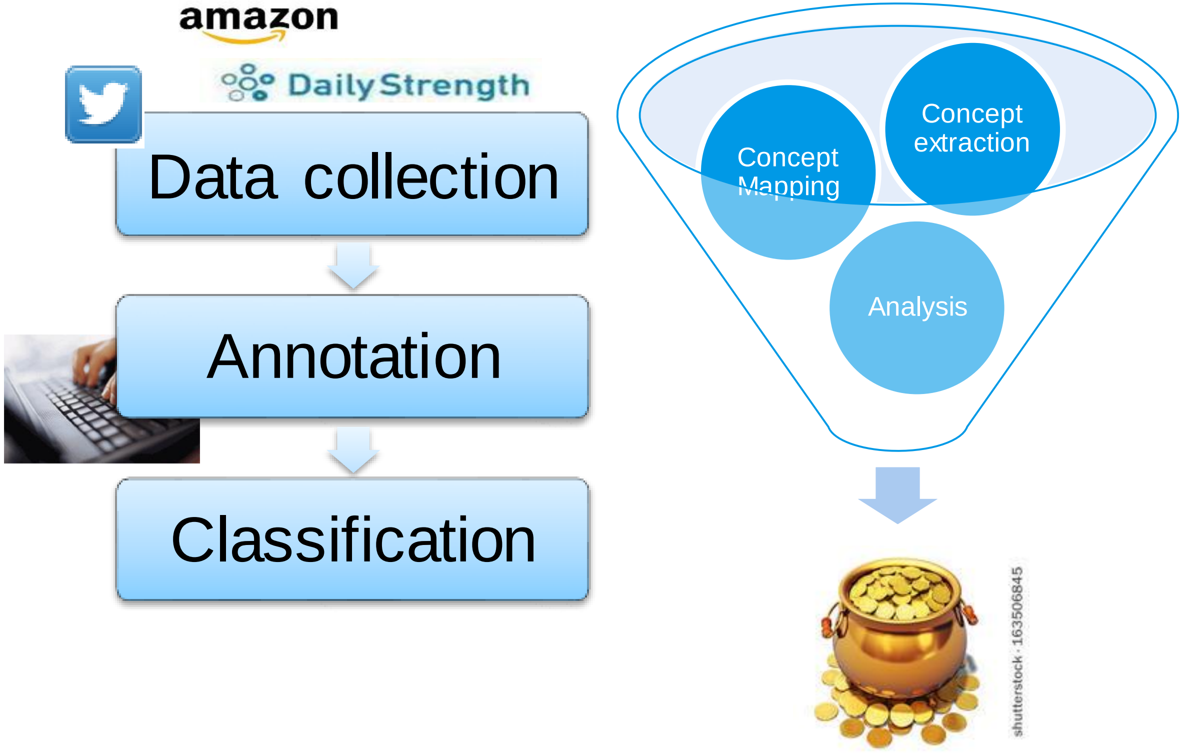
Work during first funding cycle

- Our prior work addressed the challenges of automatically collecting and processing SM reports on medication side effects.
- It resulted in over 16 publications, numerous annotated datasets, and novel automatic language processing (NLP) methods for side effect mention extraction and normalization to a standardized vocabulary (the UMLS/MedDRA).

Overview

- **Develop novel NLP methods to leverage SM data for specific pharmacovigilance efforts that are hindered by known drawbacks of SRSs.**
- **We focus on methods to facilitate the use of SM data for exploring**
  - (a) factors affecting medication adherence and persistence among the general population (Aim 1), and
  - (b) possible associations between medications taken during pregnancy and pregnancy outcomes (Aim 2).
- **These are areas of significant impact for which SM data could meaningfully complement current PV efforts**

Social Media Mining pipeline



Data collection

Annotation

Classification

Concept extraction

Concept Mapping

Analysis

The Aims

- ***Develop and evaluate NLP methods to identify non-adherence and non-persistence and related information from Twitter data.***

- ***Develop and evaluate NLP methods to identify medication use during pregnancy and pregnancy outcomes from Twitter data.***

- ***Develop and evaluate methods for automatic selection of control groups to address the challenge faced when information from SM is to be used for epidemiological studies.***

Aim 1

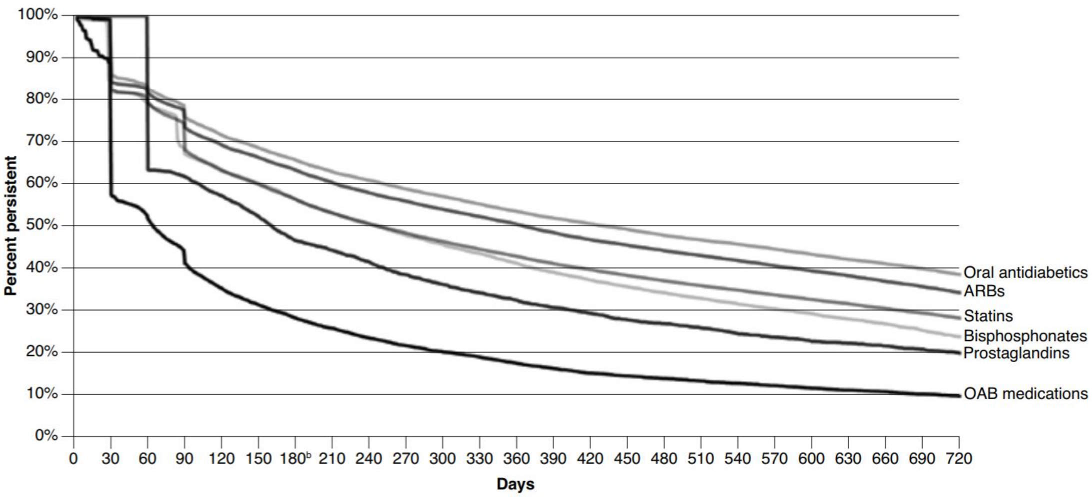
- ***Develop and evaluate NLP methods to identify non-adherence and non-persistence and related information from Twitter data. The methods will***
  - dynamically collect a cohort of SM users that stopped taking or switched medications, did not fill a prescription, or altered their treatment,
  - extract information from the user's *timeline* (publicly available postings over time) and *conversation threads* (postings by the user and others in reply to a posting of interest) relevant to
    - (a) an expressed reason for these actions,
    - (b) dosage/duration of treatment,
    - (c) concomitant treatments, and
    - (d) diagnosed health conditions.

Adherence/persistence studies from SM

- **Social media may be particularly useful for identifying sources of intolerability that lead to non-adherence/non-persistence**
- **These are often not reported by physicians or patients through standard means because are considered “mild”, “not serious” or are unexpected**
- **Significant problem, given that, on average:**
  - 30% of treated patients have a beneficial response
  - 30% do not respond
  - 10% have only side effects
  - 35%-70% are non-adherent / non-persistent, often due to side-effects or perceived/real non-response

6-month persistence rate

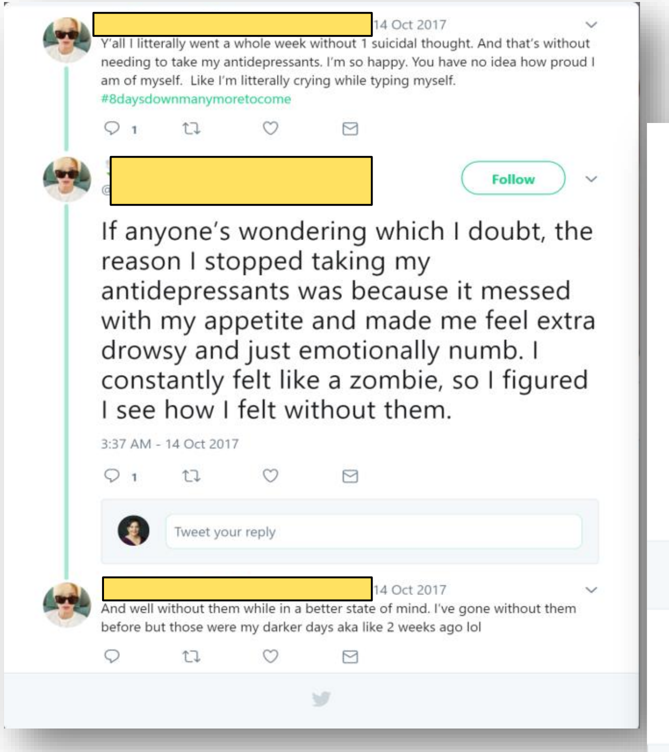
FIGURE 2 Time to Discontinuation[a] of 6 Chronic Therapy Classes, Allowing for 60-Day Treatment Gap



- prostaglandin analogs 47%
- statins 56%
- bisphosphonates 56%
- oral antidiabetics 66%
- angiotensin II receptor blocker 63%
- overactive bladder medications 28%

Comparing Adherence and Persistence Across 6 Chronic Medication Classes
Jason Yeaw, J Manag Care Spec Pharm, 2009 Nov;15(9):728-740.

“I stopped taking” & “made me”



[redacted] 14 Oct 2017

Y'all I litterally went a whole week without 1 suicidal thought. And that's without needing to take my antidepressants. I'm so happy. You have no idea how proud I am of myself. Like I'm litterally crying while typing myself. #8daysdownmanymoretocome

1

[redacted]

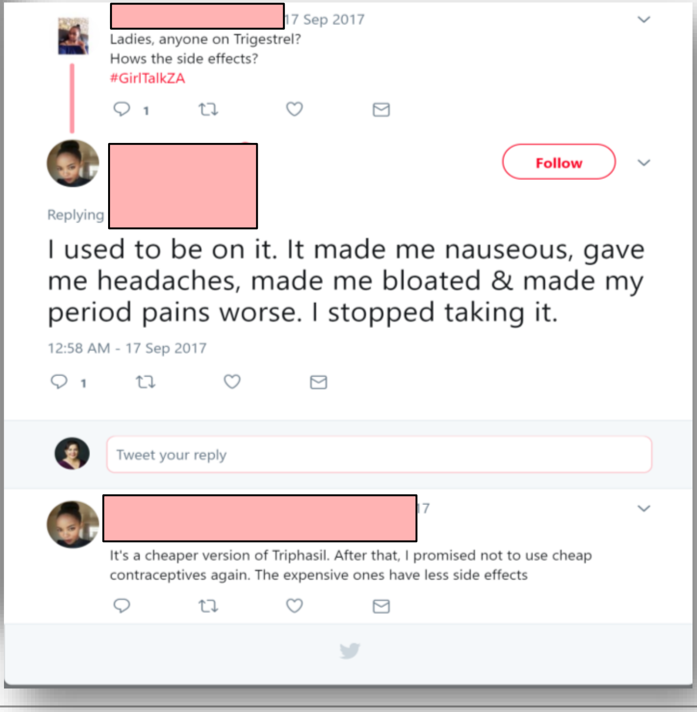
If anyone's wondering which I doubt, the reason I stopped taking my antidepressants was because it messed with my appetite and made me feel extra drowsy and just emotionally numb. I constantly felt like a zombie, so I figured I see how I felt without them.

3:37 AM - 14 Oct 2017

1

[redacted] 14 Oct 2017

And well without them while in a better state of mind. I've gone without them before but those were my darker days aka like 2 weeks ago lol



[redacted] 17 Sep 2017

Ladies, anyone on Trigestrel?
Hows the side effects?
#GirlTalkZA

1

[redacted]

Replying [redacted]

I used to be on it. It made me nauseous, gave me headaches, made me bloated & made my period pains worse. I stopped taking it.

12:58 AM - 17 Sep 2017

1

[redacted] 7

It's a cheaper version of Triphasil. After that, I promised not to use cheap contraceptives again. The expensive ones have less side effects

Aim 2

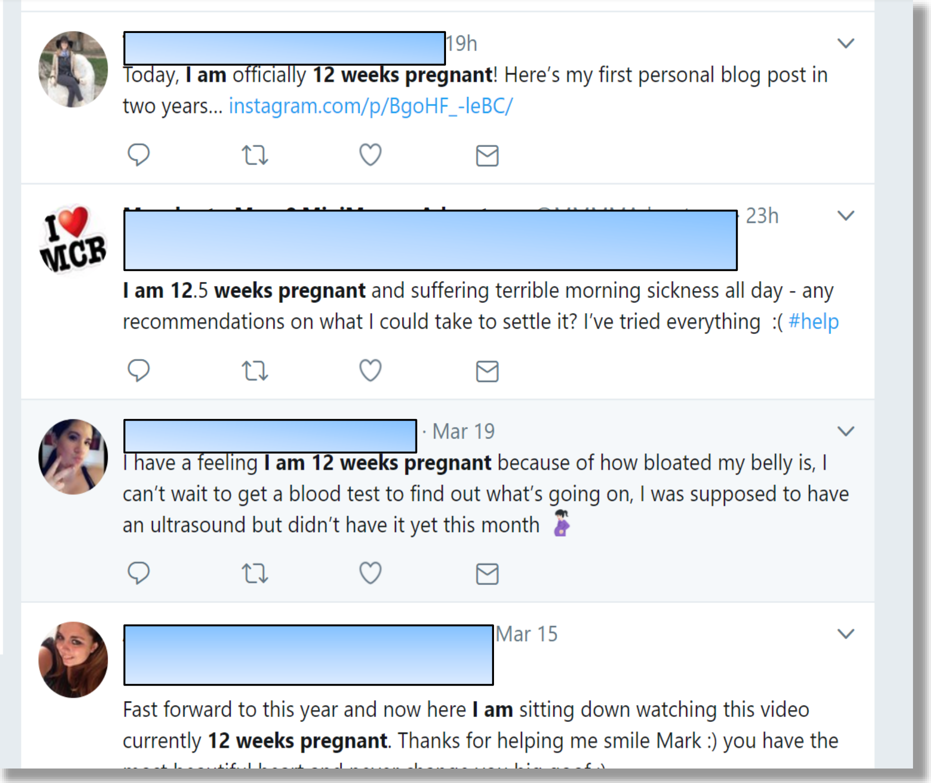
- ***Develop and evaluate NLP methods to identify medication use during pregnancy and pregnancy outcomes from Twitter data.***
  - Development and evaluation of NLP methods to dynamically collect a cohort of SM users who report a pregnancy, and
  - Methods to extract information from the user’s timeline to
    - (a) distinguish when mention of a medication indicates possible intake of it,
    - (b) infer the estimated pregnancy timeframe (beginning and end of pregnancy), and
    - (c) extract or infer pregnancy outcomes from those postings (including at least live birth, fetal death, hemorrhage, miscarriage, low-birth weight, pre-term birth, and reported congenital malformations)

Case-control study with SM data?

- **Select cohort of pregnant women from SM**[1]
  - About 120 thousand, 700 million tweets
- **Within that, find cases of interest**
  - *"Women who gave birth to a child with a birth defect and whose public tweets include tweets during pregnancy"*
- **Annotate (100% of the data found)**
- **Find matching (control) subjects**
  - *"Women pregnant around the same time, for whom there is no evidence that their child was born with a birth defect"*

[1] Sarker *et al* Discovering cohorts of pregnant women .. J Med Internet Res. 2018

From Twitter, “I am 12 weeks pregnant”



[redacted] 19h

Today, **I am** officially **12 weeks pregnant**! Here’s my first personal blog post in two years... instagram.com/p/BgoHF_-leBC/

[redacted] 23h

I am 12.5 **weeks pregnant** and suffering terrible morning sickness all day - any recommendations on what I could take to settle it? I’ve tried everything :(#help

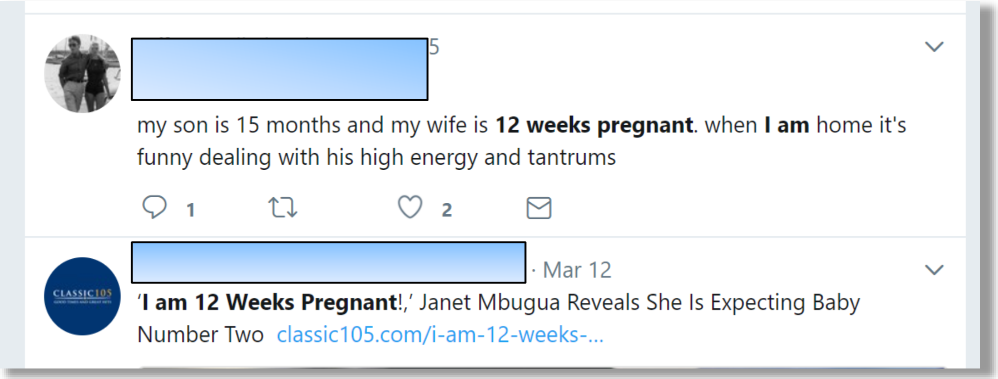
[redacted] · Mar 19

I have a feeling **I am 12 weeks pregnant** because of how bloated my belly is, I can’t wait to get a blood test to find out what’s going on, I was supposed to have an ultrasound but didn’t have it yet this month 🤰

[redacted] Mar 15

Fast forward to this year and now here **I am** sitting down watching this video currently **12 weeks pregnant**. Thanks for helping me smile Mark :) you have the most beautiful heart and never change you big goof :)

From Twitter, noise



[redacted]

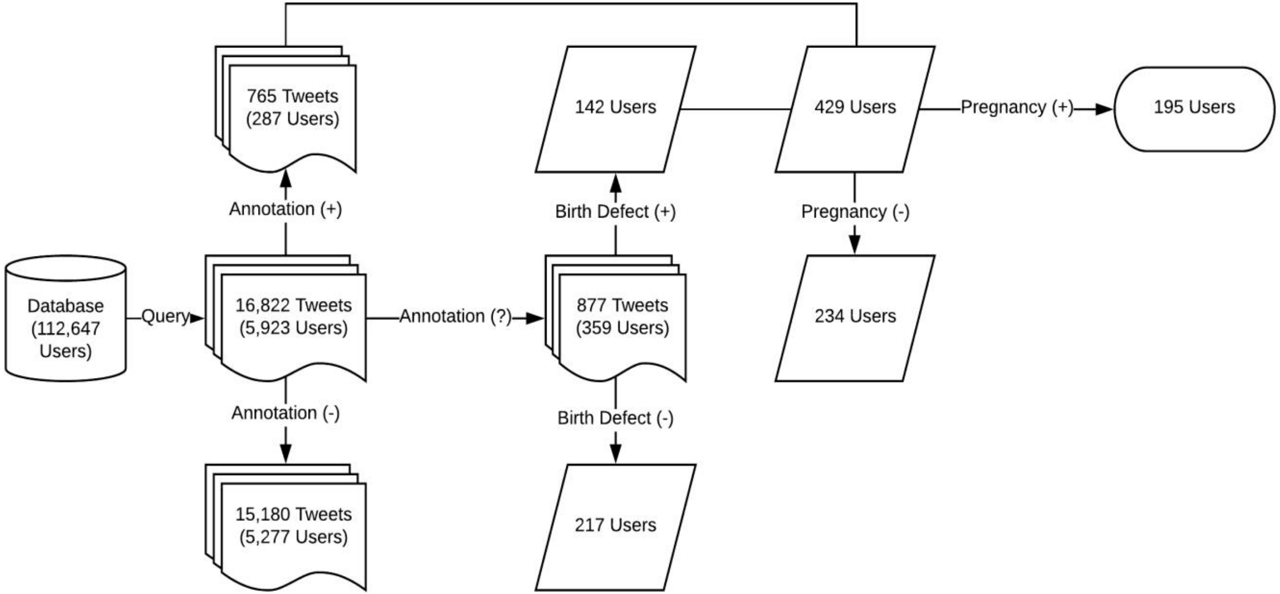
my son is 15 months and my wife is **12 weeks pregnant**. when **I am** home it's funny dealing with his high energy and tantrums

1 2

[redacted] · Mar 12

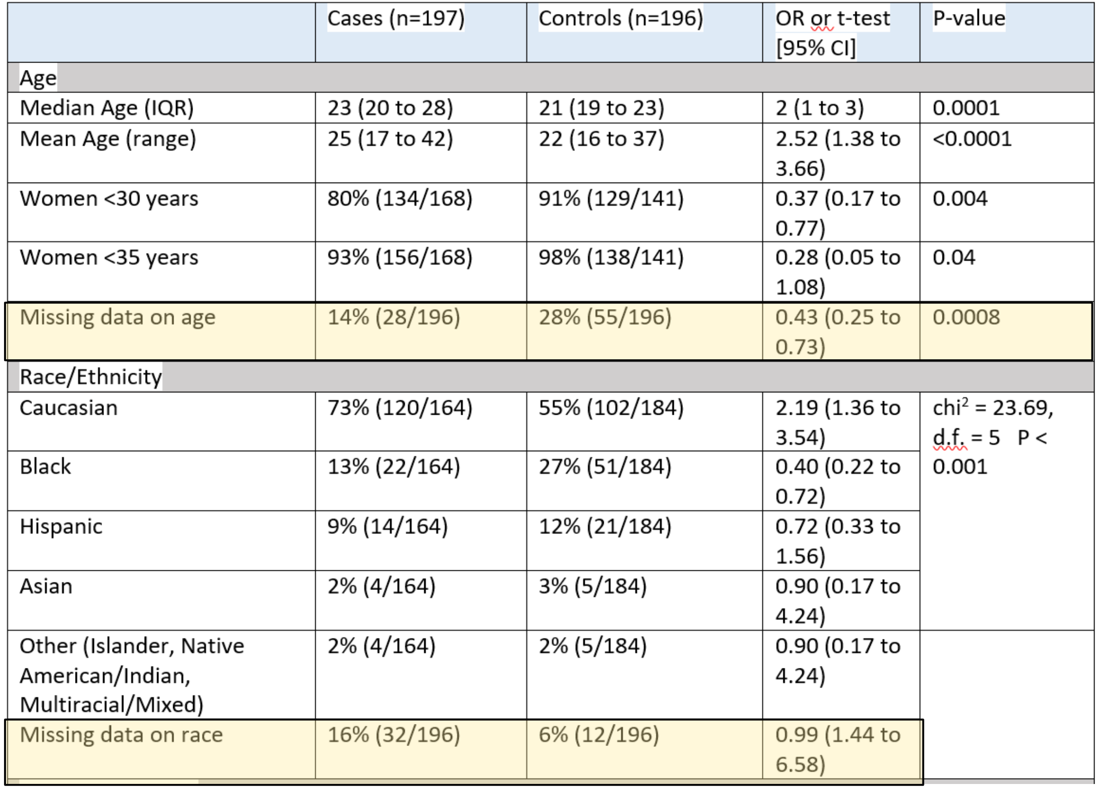
'**I am 12 Weeks Pregnant**!,' Janet Mbugua Reveals She Is Expecting Baby Number Two classic105.com/i-am-12-weeks-...

Finding cases – birth defects cohort



Klein et al, 2018 (in preparation)

Birth defects data from Social Media



	Cases (n=197)	Controls (n=196)	OR or t-test [95% CI]	P-value
Age				
Median Age (IQR)	23 (20 to 28)	21 (19 to 23)	2 (1 to 3)	0.0001
Mean Age (range)	25 (17 to 42)	22 (16 to 37)	2.52 (1.38 to 3.66)	<0.0001
Women <30 years	80% (134/168)	91% (129/141)	0.37 (0.17 to 0.77)	0.004
Women <35 years	93% (156/168)	98% (138/141)	0.28 (0.05 to 1.08)	0.04
Missing data on age	14% (28/196)	28% (55/196)	0.43 (0.25 to 0.73)	0.0008
Race/Ethnicity				
Caucasian	73% (120/164)	55% (102/184)	2.19 (1.36 to 3.54)	chi^2 = 23.69, d.f. = 5 P < 0.001
Black	13% (22/164)	27% (51/184)	0.40 (0.22 to 0.72)	
Hispanic	9% (14/164)	12% (21/184)	0.72 (0.33 to 1.56)	
Asian	2% (4/164)	3% (5/184)	0.90 (0.17 to 4.24)	
Other (Islander, Native American/Indian, Multiracial/Mixed)	2% (4/164)	2% (5/184)	0.90 (0.17 to 4.24)	
Missing data on race	16% (32/196)	6% (12/196)	0.99 (1.44 to 6.58)	

Klein et al, 2018 (in preparation)

Thank you!

gragon@pennmedicine.upenn.edu
Twitter: @gracielagon
HLP lab (datasets and software available):
https://healthlanguageprocessing.org