Fact Checking

Some of these slides are copied from the tutorial at KDD 2018. See https://shiralkarprashant.github.io/fact-checking-tutorial-KDD2018/ for the original source

What is Fact Checking?

- Determine the **correctness** of a factual statement by
  - Searching for evidence in external data sources
  - Evaluating and aggregating the evidence

- **Correctness** or **Verifiability?**

Misinformation on the Web

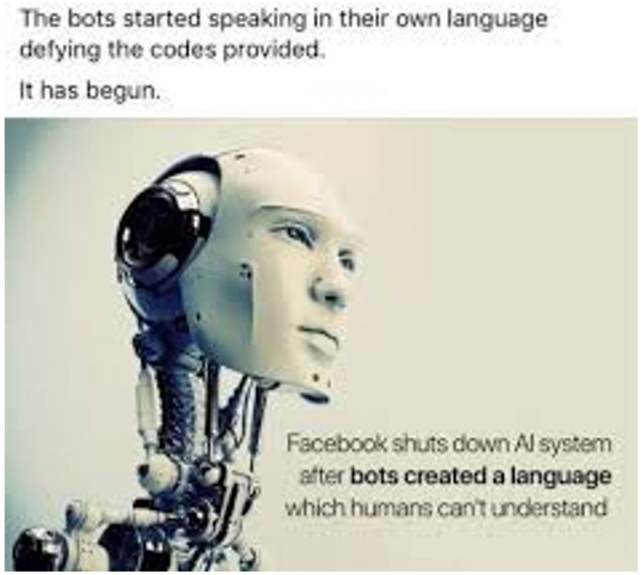
Some examples…

- “global warming was a fraud designed to diminish liberty and weaken democracy [1]”
- “President Obama purposefully allowed Ebola to enter the United States so the country would be more like Africa” [2]
- “Jade Helm 15, a simple military exercise, was perceived on the Internet as the beginning of a new civil war in the United States”

[1] http://www.forbes.com/sites/charleskadlec/2011/07/25/the-goal-is-power-the-global-warming-conspiracy/#466ef44b5ed5
[2] https://www.washingtonpost.com/news/wonk/wp/2014/10/13/the-inevitable-rise-of-ebola-conspiracy-theories/
[3] https://www.washingtonpost.com/news/checkpoint/wp/2015/09/14/remember-jade-helm-15-the-controversial-military-exercise-its-over/

Misinformation on the Web

- More examples



Misinformation on the Web

- **Fake news** – Stories that were fabricated
- Trump used the term 153 times in 2017

[1] https://www.vox.com/policy-and-politics/2018/5/9/17335306/trump-tweet-twitter-latest-fake-news-credentials

Misinformation on the Web

Real problem!

- World Economic Forum (WEF) consider digital misinformation one of the main threats to our society
- An increasing exposure of users to unsubstantiated rumors increases tendency to be credulous
- Google is developing a trustworthiness score to rank the results of queries
- Facebook has proposed a community-driven approach where users can flag false content to correct the newsfeed algorithm
- How to remove fake news and avoid censorship?

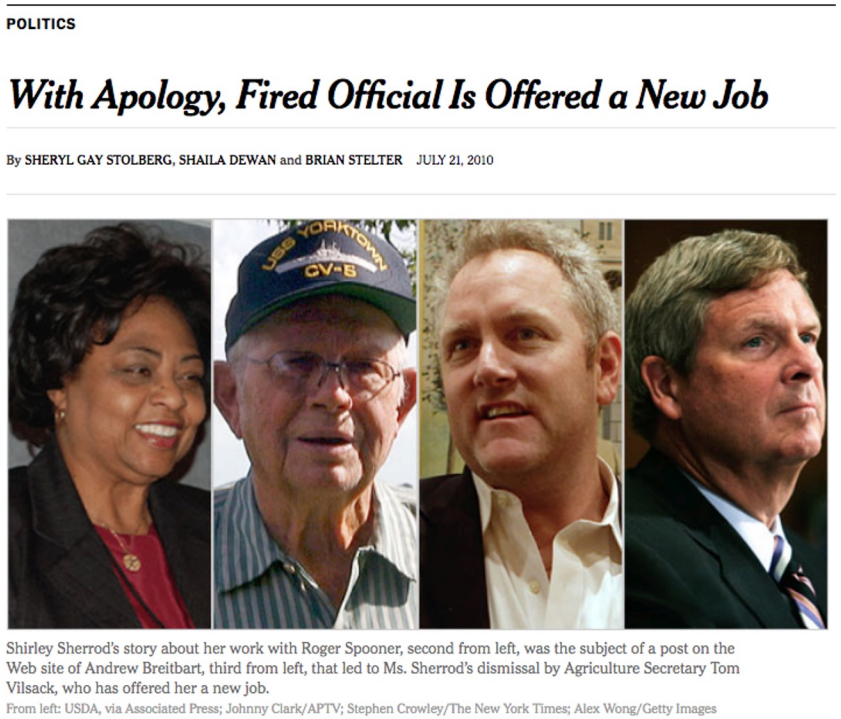
Different types of falsehood

- **“Technical” falsehood**
  - Extraction mistakes: Facts that were extracted incorrectly
  - Caused by the noisy process of extracting information from text
- **Factual falsehood**
  - Hoaxes: Deliberately fabricated falsehood made to masquerade as truth”
  - Conspiracy theories
  - Scientific discoveries
  - Etc.

[1] Y. Li *et al.*, “A survey on truth discovery,” *arXiv preprint arXiv:1505.02463*, 2015.

Different types of falsehood

- Misleading video taken out of context posted on Blog



POLITICS

With Apology, Fired Official Is Offered a New Job

By SHERYL GAY STOLBERG, SHAILA DEWAN and BRIAN STELTER JULY 21, 2010

Shirley Sherrod's story about her work with Roger Spooner, second from left, was the subject of a post on the Web site of Andrew Breitbart, third from left, that led to Ms. Sherrod's dismissal by Agriculture Secretary Tom Vilsack, who has offered her a new job.

From left: USDA, via Associated Press; Johnny Clark/APTV; Stephen Crowley/The New York Times; Alex Wong/Getty Images

[1] https://www.youtube.com/watch?v=BLWeMyGpTfI
[2] https://www.nytimes.com/2010/07/22/us/politics/22sherrod.html

Different types of falsehood

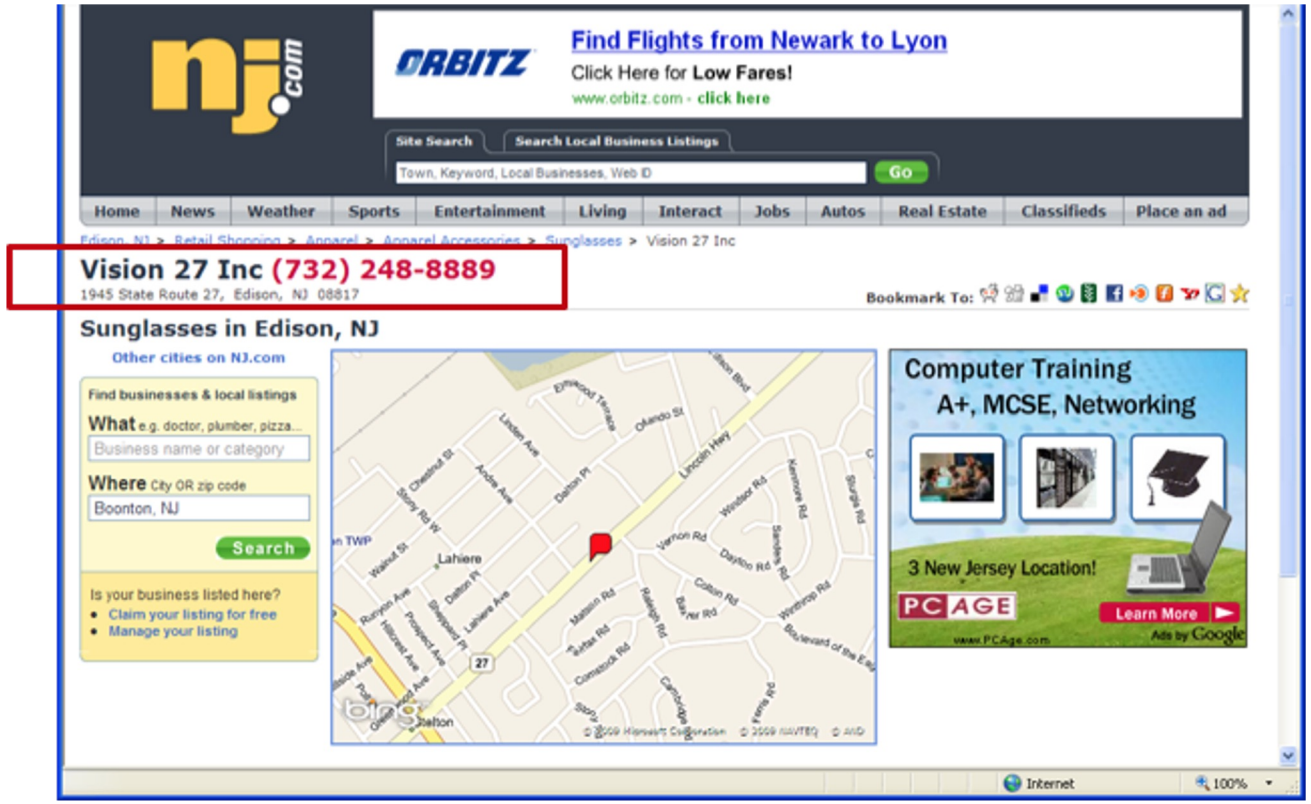
- Erroneous information because of data entry errors

IMDB

WikiData

Different types of falsehood

- Erroneous information because of out-of-date information



Different types of falsehood

Wrong information can be just as bad as lack of information.

The Internet needs a way to help people separate rumor from real science.

–Tim Berners-Lee

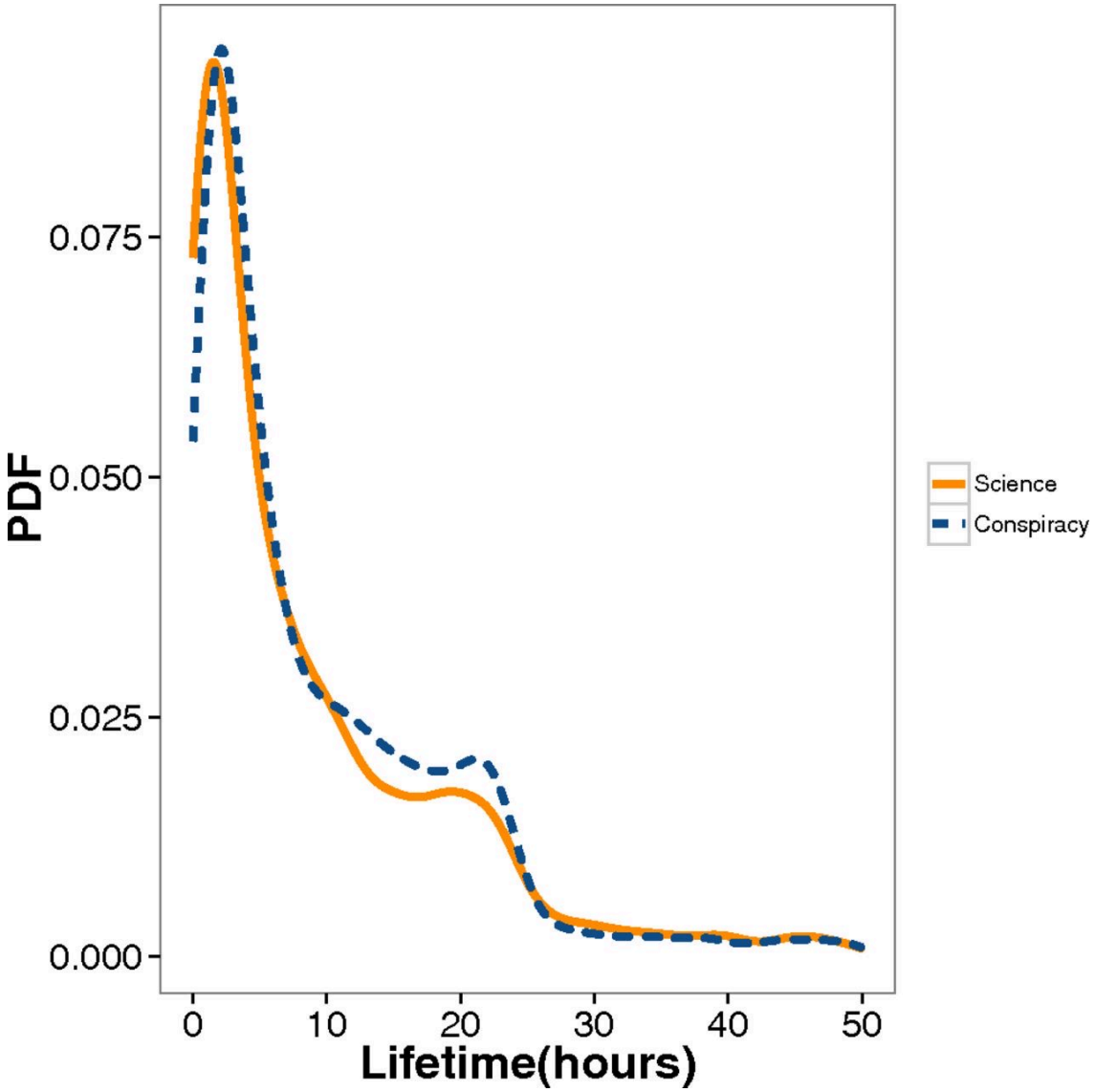
Misinformation on the Web

- A recent study [1] analyzed the phenomenon of misinformation spreading focusing on two types of fake news:
  - conspiracy theories
  - scientific information
- Conspiracy theories
  - Simplify causation
  - Reduce complexity reality
  - Formulated to tolerate uncertainty
  - Origin is unknown
- Scientific information
  - Disseminate scientific advances and exhibit process of scientific thinking
  - We can trace the origin

[1] http://www.pnas.org/content/113/3/554.full

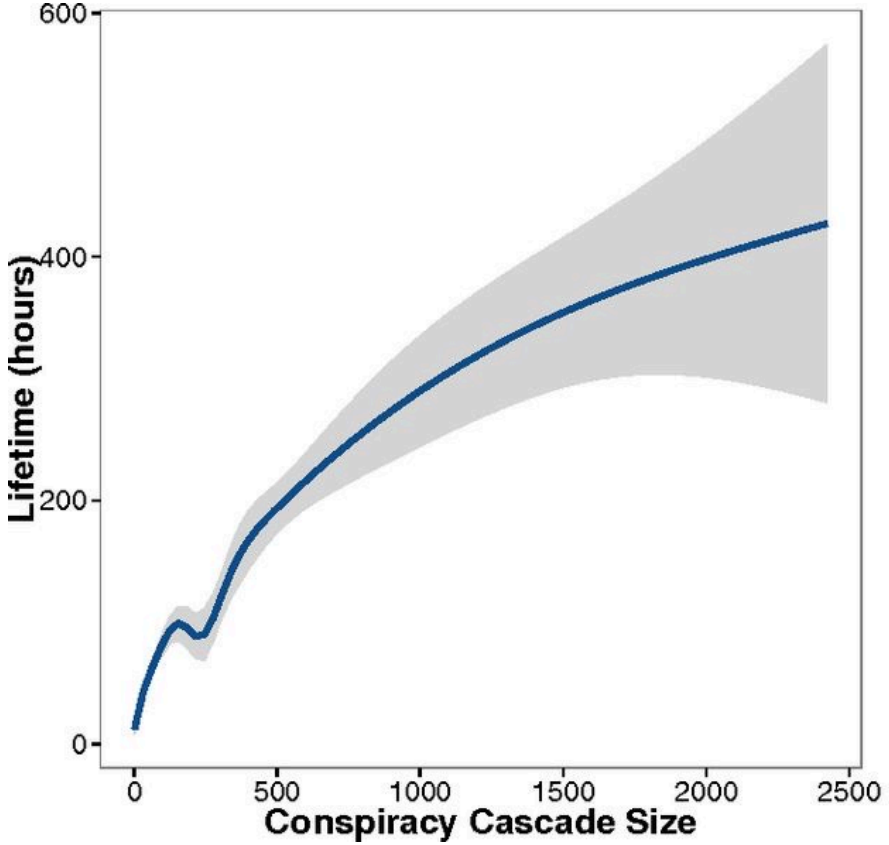
Misinformation on the Web

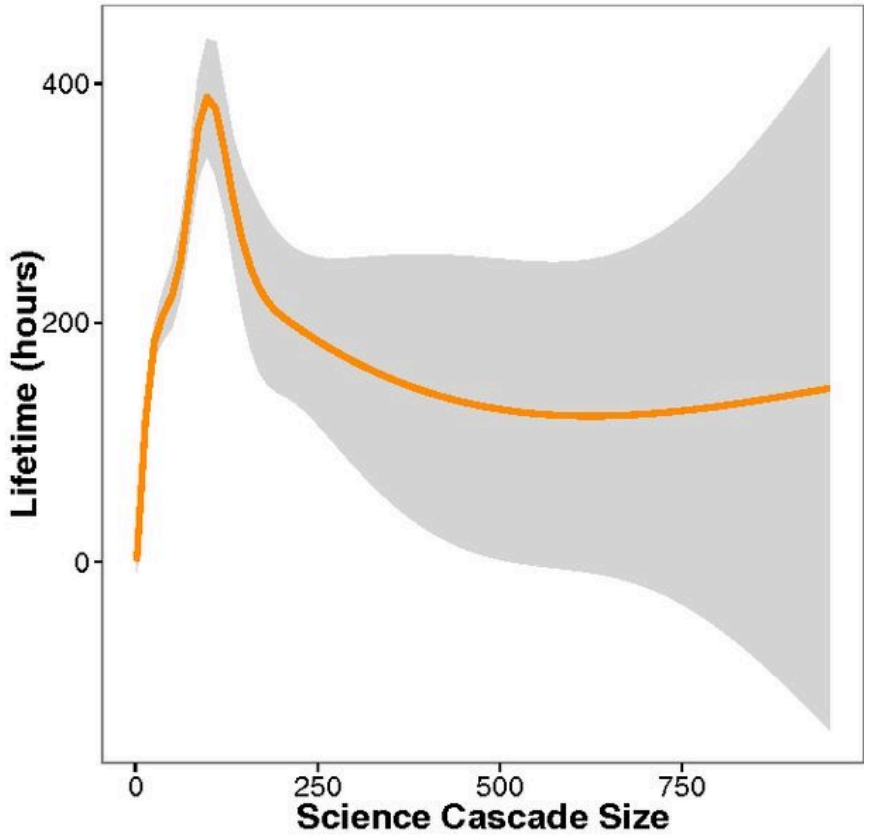
- Methodology of the study:
  Picked from Facebook 32 pages
  that disseminate conspiracy theories
  and 35 about “fake” scientific news
- Measured lifetime of news:
  time between first and last share



Misinformation on the Web

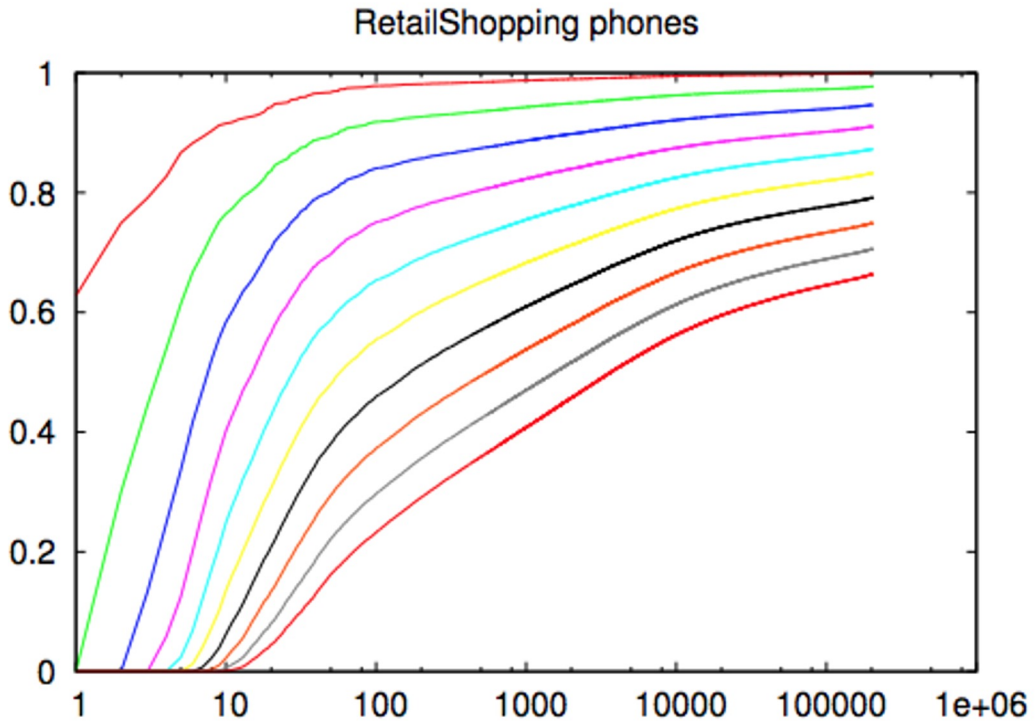
Lifetime as a function of the cascade size for conspiracy news (Left) and science news (Right).





Why is Fact Checking hard?

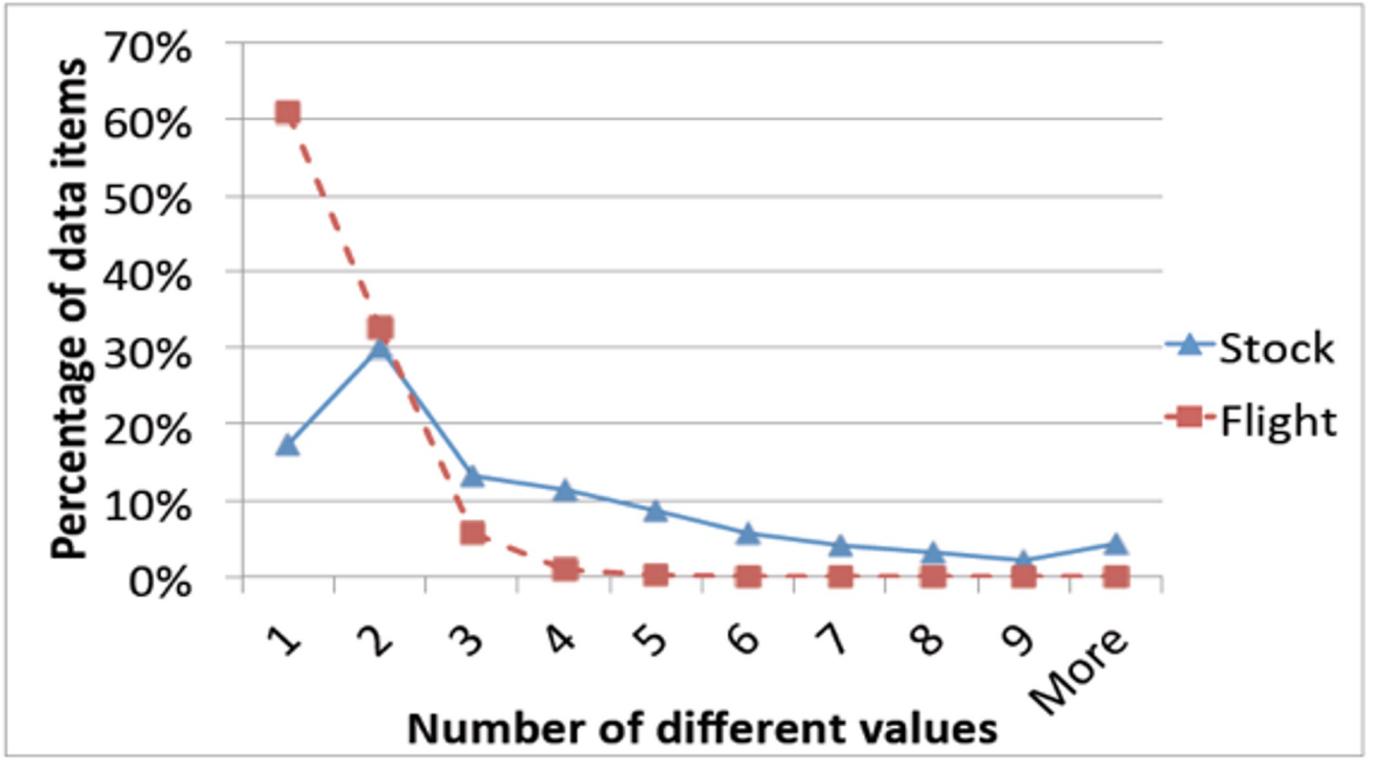
- **Sparsity**: To verify 90% retail phone# from >=3 sources, we need > 1000 sources



[1] Nilesh Dalvi, Ashwin Machanavajjhala, Bo Pang: *An Analysis of Structured Data on the Web*. VLDB 2012

Why is Fact Checking hard?

- **Conflicts:** Inconsistency on 70% data items with tolerance of 1% difference



[1] Xian Li, Xin Luna Dong, Kenneth Lyons, Weiyi Meng, Divesh Srivastava *Truth Finding on the Deep Web: Is the Problem Solved?* PVLDB 2012

Why is Fact Checking hard?

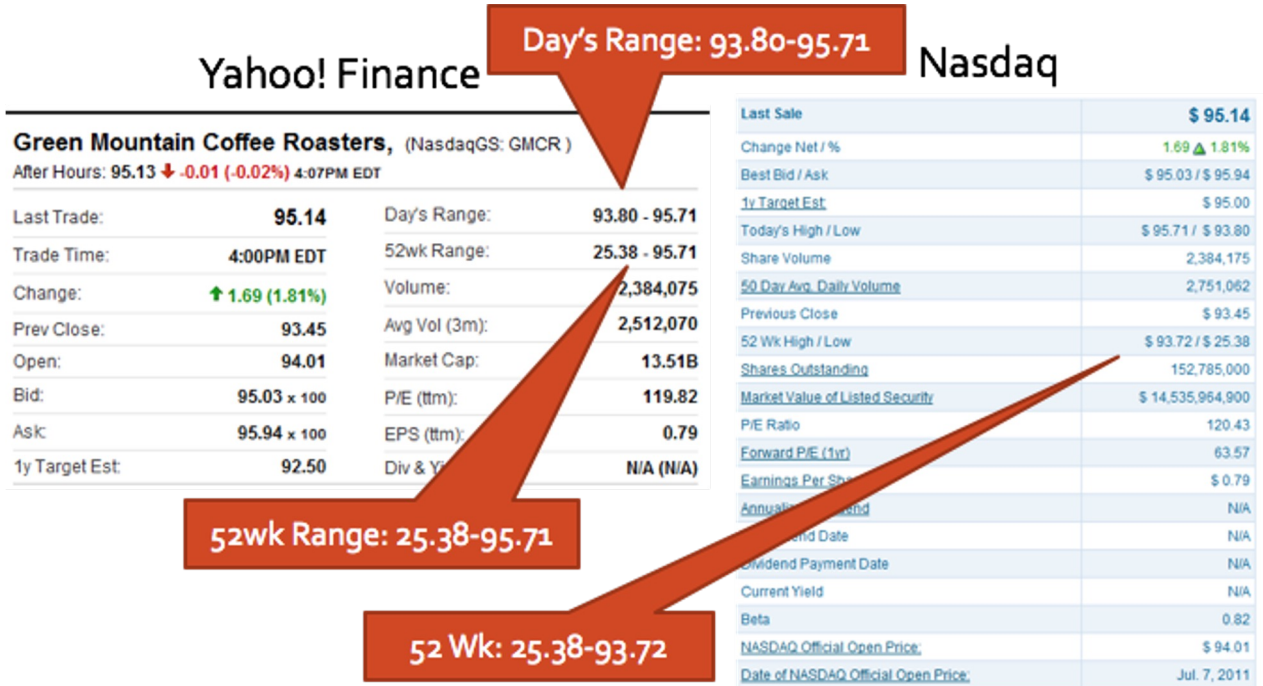
- **Trustworthyness:** Even authoritative sources may not have very high accuracy of data

	Source	Accuracy	Coverage
Stock	*Google Finance*	.94	.82
	Yahoo! Finance	.93	.81
	NASDAQ	.92	.84
	MSN Money	.91	.89
	Bloomberg	.83	.81
Flight	*Orbitz*	.98	.87
	Travelocity	.95	.71
	Airport average	.94	.03

[1] Xian Li, Xin Luna Dong, Kenneth Lyons, Weiyi Meng, Divesh Srivastava *Truth Finding on the Deep Web: Is the Problem Solved?* PVLDB 2012

Why is Fact Checking hard?

- **Semantic Ambiguity:** 46% inconsistency in Stock and 33% inconsistency in Flight are caused by semantic ambiguity



Why is Fact Checking hard?

- **Instance Ambiguity**

- Wei Zhang 0081 — Huawei Technologies Company Ltd., Media Laboratory, Shenzhen, China (and 4 more)
- Wei Zhang 0082 — Nanyang Technological University, School of Computer Science and Engineering, Singapore
- Wei Zhang 0083 — Zhejiang University, Institute of Microelectronics and Optoelectronics, China
- Wei Zhang 0084 — Beijing University of Technology, College of Mechanical Engineering, China
- Wei Zhang 0085 — Fudan University, School of Computer Science, Shanghai, China
- Wei Zhang 0086 — Huazhong University of Science and Technology, School of Computer Science and Technology, Wuhan,
- Wei Zhang 0087 — Beijing University of Posts and Telecommunications, MoE Key Laboratory of Optical Communication an

[+] **Other persons with a similar name**

[–] **2010 – today**

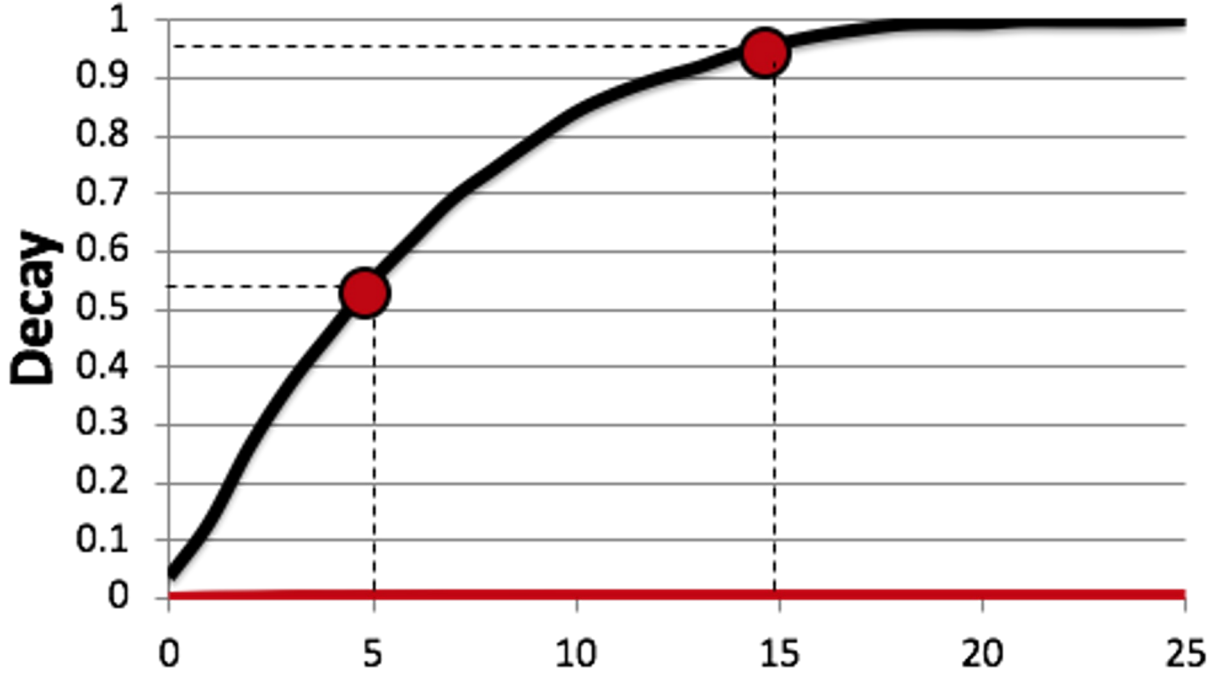
2018

[j431] Chuanhao Li, Wei Zhang, Gaoliang Peng, Shaohui Liu:
Bearing Fault Diagnosis Using Fully-Connected Winner-Take-All Autoencoder. IEEE Access 6: 6103-6115 (2018)

[j430] Chunxue Wu, Chong Luo, Naixue Xiong, Wei Zhang, Tai-Hoon Kim:
A Greedy Deep Learning Method for Medical Disease Analysis. IEEE Access 6: 20021-20030 (2018)

[j429] Wei Zhang, Shu-Lin Wang:
Inference of Cancer Progression With Probabilistic Graphical Model From Cross-Sectional Mutation Data. IEEE Access 6: 22889-22898 (2018)

Why is Fact Checking hard?

- **Changes over time:** Over half European patent inventors changed their address in 5 years
- 

Patent records: 1871

Real-world inventors: 359

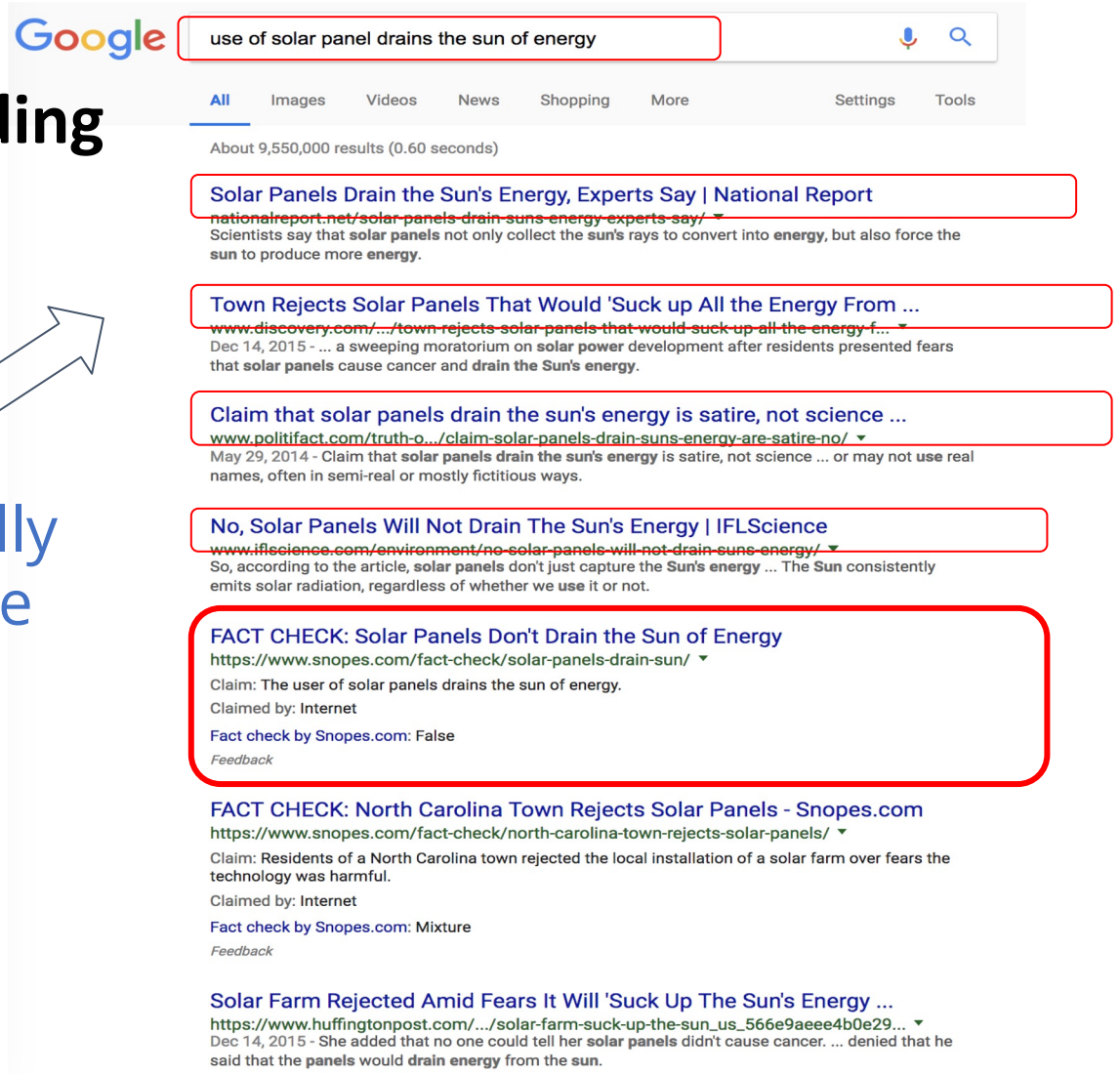
In years: 1978 - 2003

[1] Pei Li, Xin Luna Dong, Andrea Maurino, Divesh Srivastava: *Linking Temporal Records*. VLDB 2011.

Why is Fact Checking hard?

- **Text understanding**

Does solar panels really drain the energy of the sun?



Why is Fact Checking hard?

- **Infer**

how many universities in china

All News Maps Images Videos More

About 129,000,000 results (0.56 seconds)

This article is a list of universities in mainland China, Hong Kong and Macau (of P.R.C.). By May of 2017, there were **2,914** colleges and universities, with over 20 million students enrolled in mainland China. More than 6 million Chinese students graduated from university in 2008.

List of universities in China - Wikipedia
https://en.wikipedia.org/wiki/List_of_universities_in_China

About this

How are Chinese students at US universities faring in 201
https://www.studyinternational.com/news/china-student-us-2018/
Jan 29, 2018 - There're **more Chinese** students seeking education abroad in the foreign students in the **U.S. higher** education system **with** our ... gain admissio

how many universities in the us

All Images News Maps Videos More Settings Too

About 296.000.000 results (0,88 seconds)

A privacy reminder from Google

To be consistent with data protection laws, we're asking that you take a moment to review key points of our Privacy Policy, which covers all Google services and describes how we use data and what options you have. We'll need you to do this today.

REMIND ME LATER REVIEW NOW

5,300 colleges

In the **United States**, there are approximately 5,300 **colleges** and **universities**. These **colleges** and **universities** range from beauty schools to private Ivy League research **universities** like Harvard University. Aug 5, 2019

How Many Universities & Colleges are in the US?
https://www.educationunlimited.com › blog › how-many-universities-college...

About Featured Snippets Feedbac

What to check?

- Different formats
  - Structured triples: (Obama, born_in, Kenya)
  - Textual claims: “solar panel drains the sun of energy”
  - Entire articles: e.g., fake reviews
- Different types
  - string, categorical, numerical e.g., “The height of Mt Everest is 29K”
  - “Tom Cruise plays Pete in Top Gun”
  - “Tom Cruise is the actor in all Top Gun series”

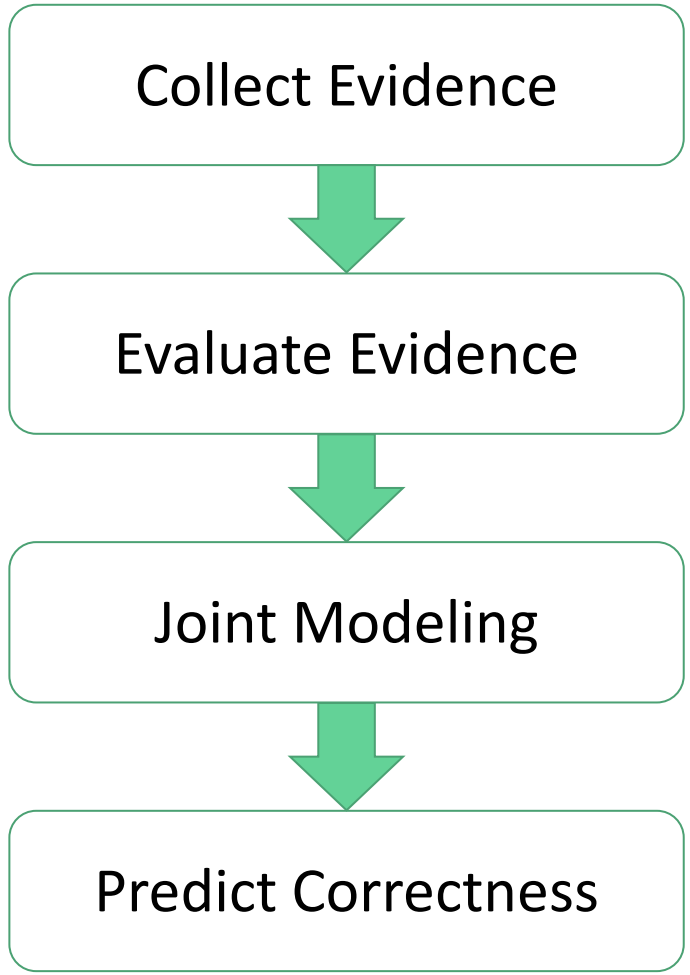
Where to check?

- Graph data
  - Manually created Knowledge Graphs
  - Open Information Extraction
- Structured data
  - Web Tables
  - Databases, etc.
- Unstructured data
  - Web pages (e.g., Wikipedia)
  - Social networks
  - Query logs

Fact Checking from Text and Social Networks

Some of these slides are copied from the tutorial at KDD 2018. See https://shiralkarprashant.github.io/fact-checking-tutorial-KDD2018/ for the original source

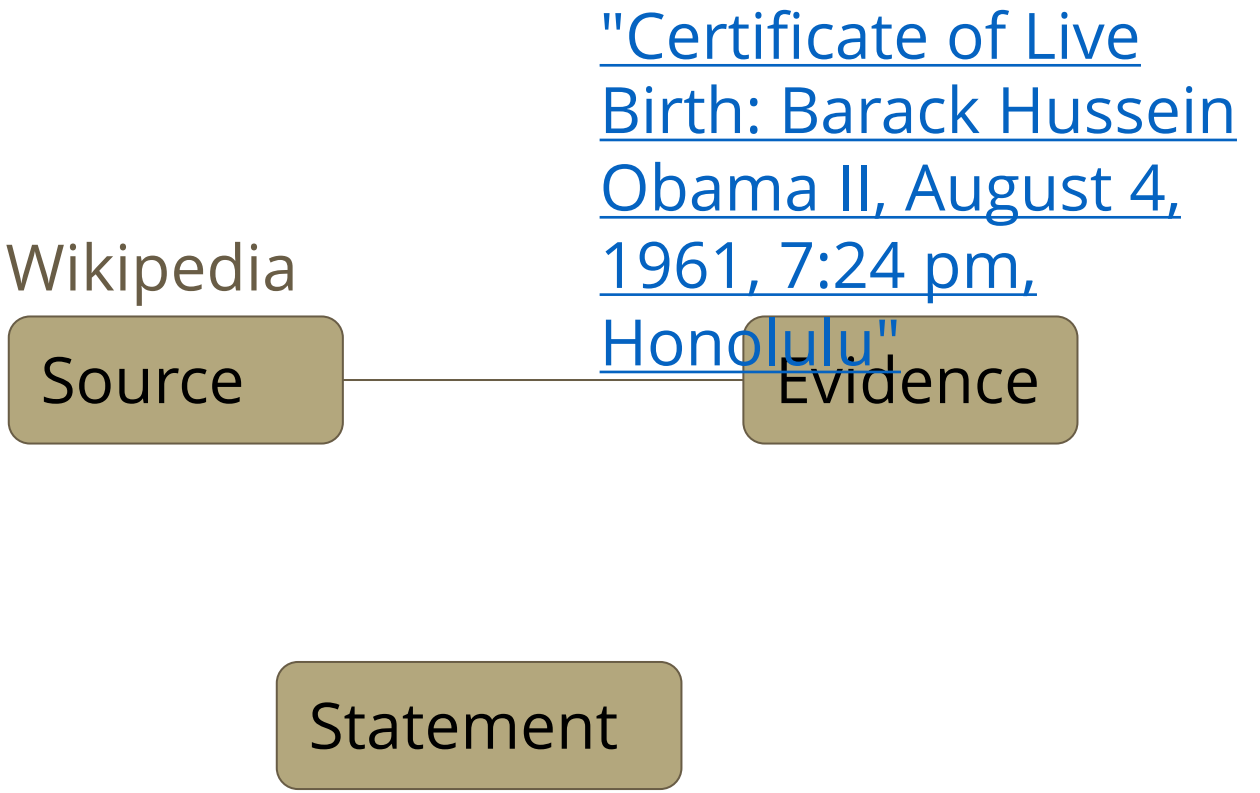
How to check?



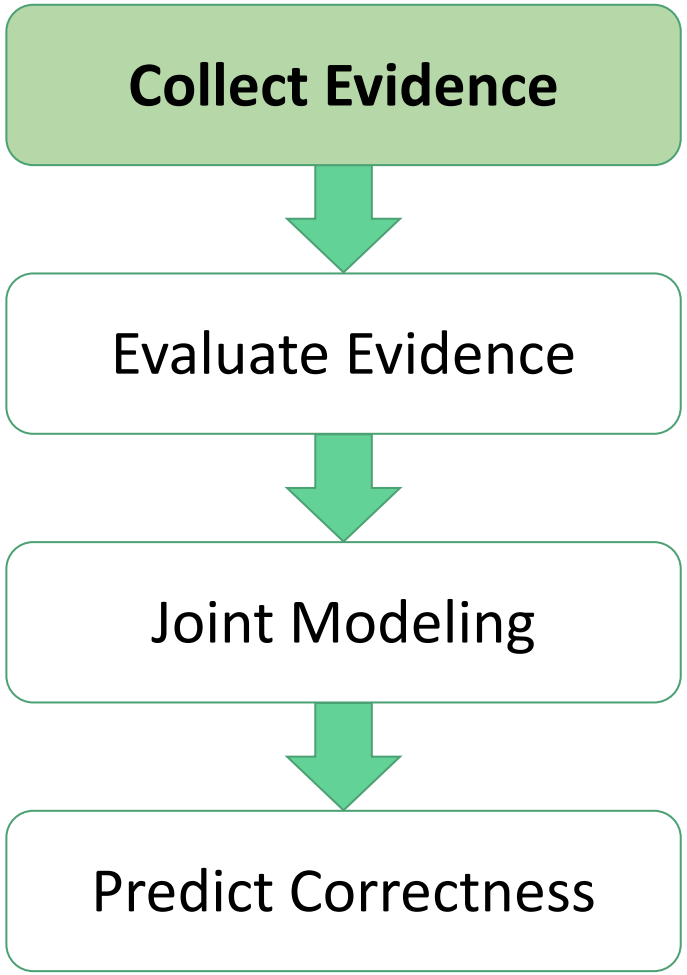
Statement

(Obama, born_in, Hawaii) True / False?

How to check?



(Obama, born_in, Hawaii) True / False ?



How to check?

Objective evidence?
Up-to-date evidence?

Trustworthy source ?
Correlation of sources?

Wikipedia

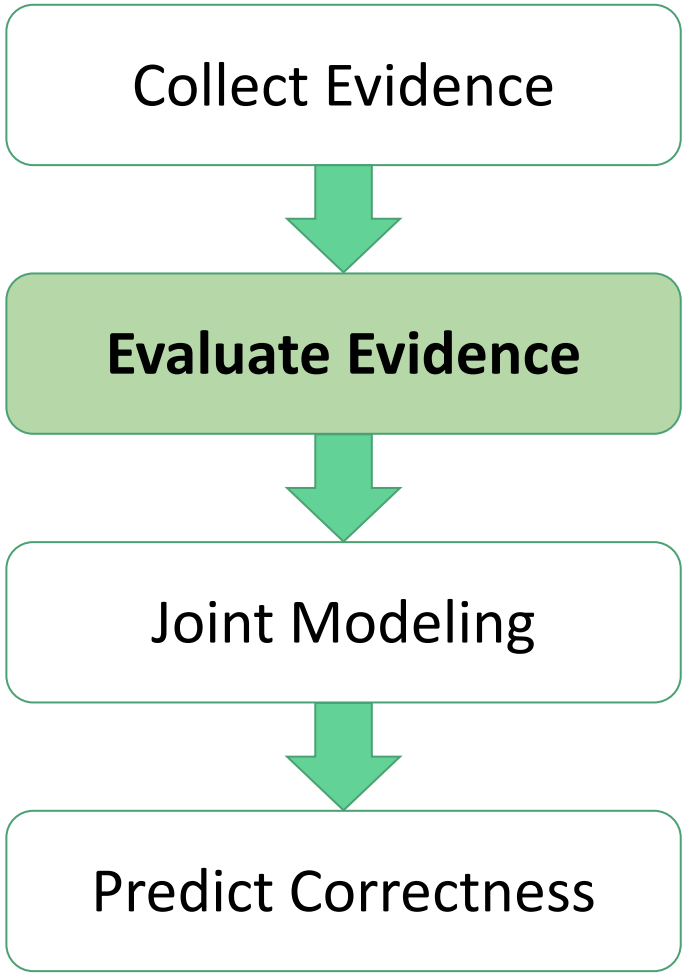
"Certificate of Live Birth: Barack Hussein Obama II, August 4, 1961, 7:24 pm, Hon

Source

Evidenc e

Statement

(Obama, born_in, Hawaii) True / False ?



How to check?

Objective evidence?
Up-to-date evidence?

Trustworthy source ?
Correlation of sources?

Wikipedia

"Certificate of Live Birth: Barack Hussein Obama II, August 4, 1961, 7:24 pm, Hon

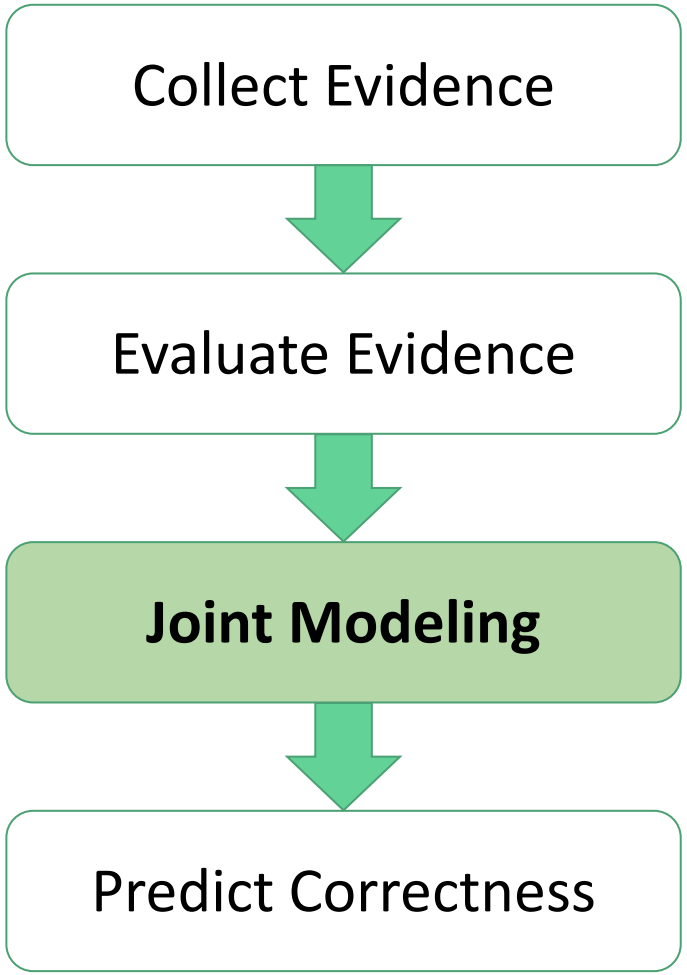
Source

Evidence

Joint interaction between statement, evidence and its source!

Statement

(Obama, born_in, Hawaii) True / False ?



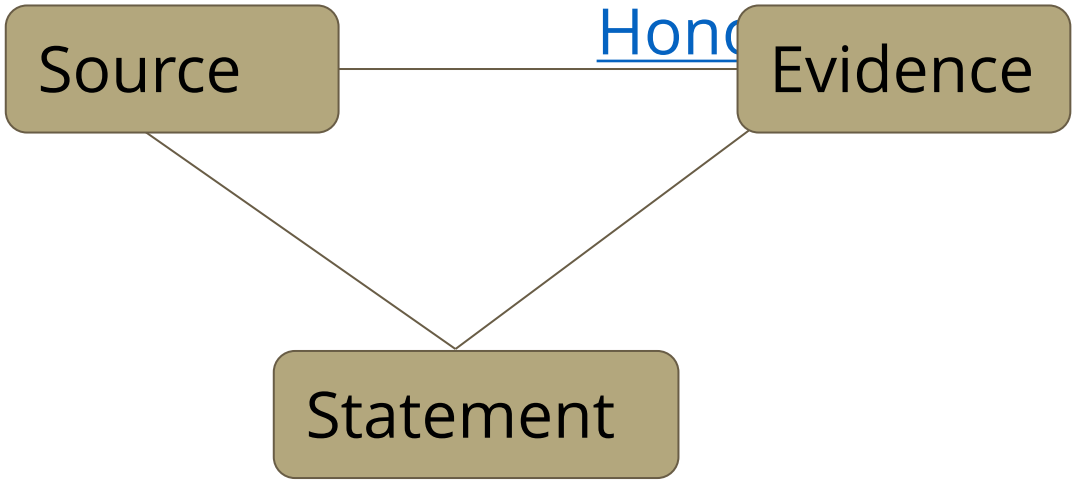
How to check?

Objective evidence?
Up-to-date evidence?

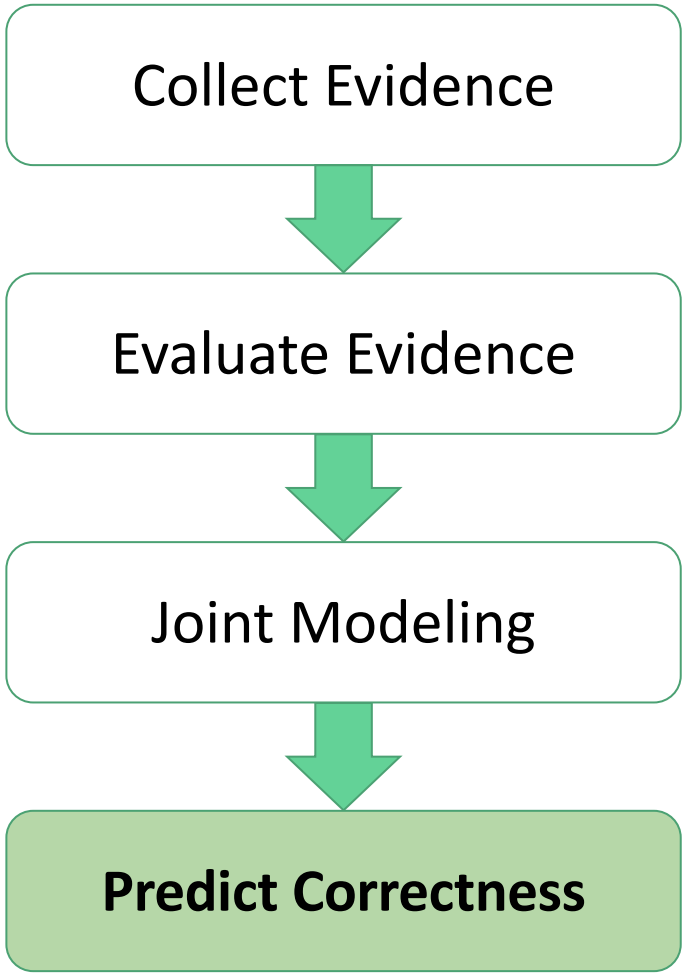
"Certificate of Live Birth: Barack Hussein Obama II, August 4, 1961, 7:24 pm, Hono

Trustworthy source ?
Correlation of sources?

Wikipedia



(Obama, born_in, Hawaii) True / False ?



Three goals

- A **Fact Checker** should:
  - Be accurate. Predictions should mimic truth
  - Be scalable. Reduce human intervention so that it can be applied on large volumes of data
  - Be interpretable. We want to be able to understand/explain the predictions

Systems for fact checking

What can we leverage for automated fact checking?

- features from the text (claim, tweet, replies, etc.)
- background knowledge
- users’ opinions

Systems for fact checking

What can we leverage for automated fact checking?

- **features from the text (claim, tweet, replies, etc.)**
- background knowledge
- users’ opinions

Information Credibility on Twitter

- One of the first and most important works in Fact-checking on social networks is the one of Castillo et. al [1]

- The paper introduces the problem of the earthquake in Chile in 2010

8th largest recorded in history

- Haiti 2010: 7.7 Mw
- Chile 2010: 8.8 Mw
- Japan 2011: 9 Mw

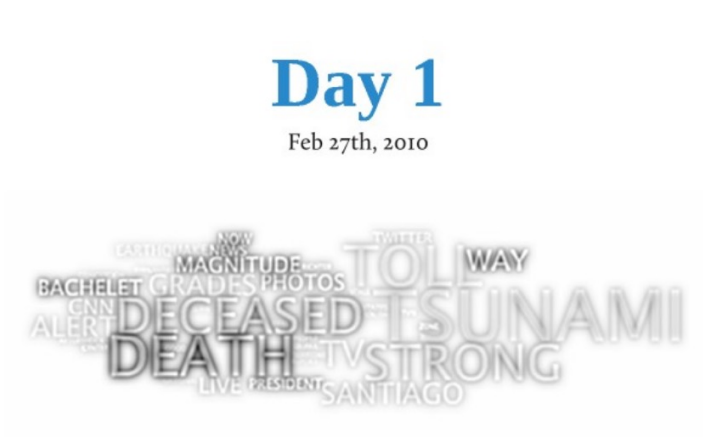
[1] C. Castillo, M. Mendoza, B. Poblete. Information Credibility on Twitter. WWW, 2011

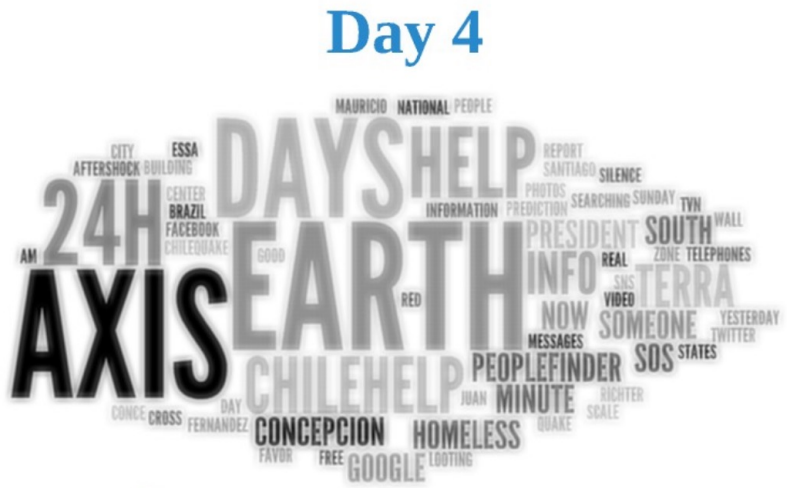
Information Credibility on Twitter

- Communication was impossible for the first 2-3 hours
- First video images 6-7 hours later

Information Credibility on Twitter

- A large amount of communication occurred on **Twitter**





Information Credibility on Twitter

- A large amount of communication occurred on **Twitter**
- Some tweets were useful, … others were not

Case	# of unique tweets	% of re-tweets	# of unique "affirms"	# of unique "denies"	# of unique "questions"
Confirmed truths					
The international airport of Santiago is closed	301	81	291	0	7
The *Viña del Mar International Song Festival* is canceled	261	57	256	0	3
Fire in the Chemistry Faculty at the University of Concepción	42	49	38	0	4
Navy acknowledges mistake informing about tsunami warning	135	30	124	4	6
Small aircraft with six people crashes near Concepción	129	82	125	0	4
Looting of supermarket in Concepción	160	44	149	0	2
Tsunami in Iloca and Duao towns	153	32	140	0	4
TOTAL	1181		1123	4	30
AVERAGE	168,71		160,43	0,57	4,29
False rumors					
Death of artist Ricardo Arjona	50	37	24	12	8
Tsunami warning in Valparaiso	700	4	45	605	27
Large water tower broken in Rancagua	126	43	62	38	20
Cousin of football player Gary Medel is a victim	94	4	44	34	2
Looting in some districts in Santiago	250	37	218	2	20
"Huascar" vessel missing in Talcahuano	234	36	54	66	63
Villarrica volcano has become active	228	21	55	79	76
TOTAL	1682		502	836	216
AVERAGE	240,29		71,71	119,43	30,86

YAHOO! RESEARCH

Information Credibility on Twitter

- **Proposed solution:** Supervised Classification

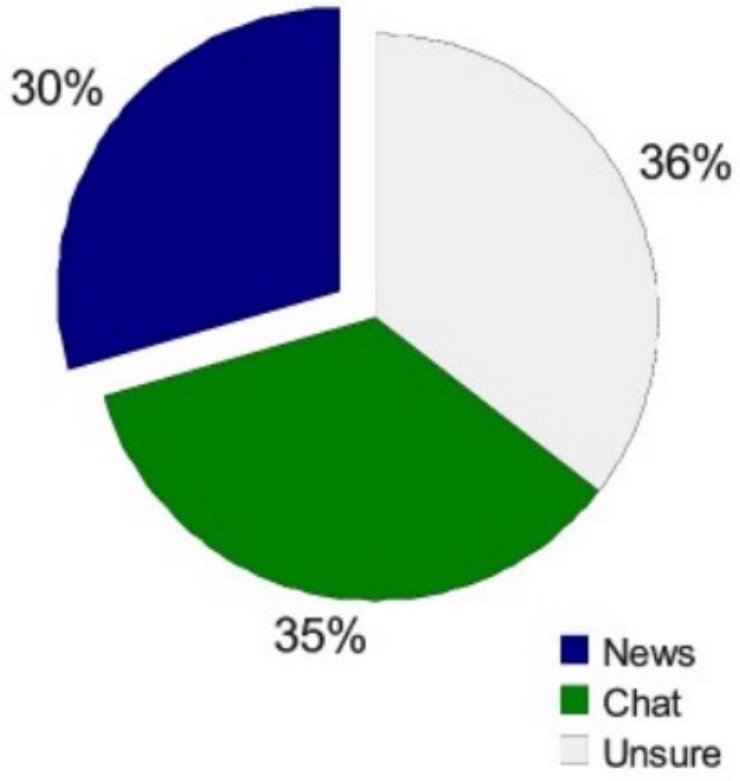
- Two problems:
  - Find events to train the classifier
  - Train the classifier(s)

Information Credibility on Twitter

- **Proposed solution:** Supervised Classification

- Two problems:
  - **Find events to train the classifier**

- **Solution:** Use Amazon Mechanical Turk. Crowdsourced 383 events.
  - **Goal** was to distinguish events and conversations between people



Identifying specific news/events from a set of tweets

Guidelines:

Users of Twitter post short messages, each up to 140 characters, commonly known as tweets.

In this task you will need to indicate if most of the tweets in the group are:

1. Spreading news about a **specific news/event**
2. Comments or conversation

A **specific news/event** must meet the following requirements:

- be an affirmation about a fact or something that really happened.
- be of interest to others, not only for the friends of each user.

Tweets are not related to a **specific news/event** if they are:

- Purely based on personal/subjective opinions.
- Conversations/exchanges among friends.

- For each group, we provide a list of descriptive keywords that help you understand the topic behind the tweets.

Examples:

Specific news/event

- Study says social ad spending to reach $1.68 billion this year
- Obama to sign $600 million border security legislation http://dlvt.it/3kqpq
- Huge brawl in GABP!!! #cardinals v #reds

Conversation/comments

- Probably should have brought rainboots to wort today. #regret
- Listening to @jaredleto performing Bad Romance gives me goosebumps
- Lovely weather for cats

Item 3.

Consider the following group of tweets:

- RT @jbreezie24 @blazetrilla lakrs bout to get raja bell <<<dat nigga a scrub anyway fuck dat nigga he gonna warm da bench up
- Fuck raja bell going to Utah? Damn!
- RT @jaharikavi: the #Utah #Mormons look like they are now getting raja bell.....> god u w fool
- SMH raja bell told Kobe Nevermind on meeting him and went to UTAH.. dick move.
- @iRapedKOBE raja bell definitely goin 2 da lakers, he'll b stupid not 2, #WeDaChamps
- @ChgTheGmE they'll see what happens next year. Yo kinda mad raja bell went to the jazz instead of us
- Don't mind Shannon brown coming back would of preferred raja bell but brown works. I'm just happy farmar is gone and Lakers got @SteveBlake5
- @Basketball_Ron Ron what do you think about the lakers going after raja bell
- Fuck U raja bell ! U chose money over a chamionship w/ Kobe lol
- RT @Lockedonsports: O'Connor "we got someone who can guard the best perimeter defender and wants to" in raja bell

descriptive keywords:"raja","bell"

The previous tweets are:

- ( ) spreading a specific news/event?
- ( ) conversation/comments among friends?

Please provide a description of the topic covered by the previous tweets in only one sentence:

Information Credibility on Twitter

- **Proposed solution:** Supervised Classification

- Two problems:

  - **Train the classifier(s)**

- **First classifier:** Set of tweets => is event or not
- **Second classifier:** Set of tweets => credible event or not

Information Credibility on Twitter

- Proposed to use a large set of features! We can group the features as:
  - Message-based features
  - User-based features
  - Topic-based features
  - Propagation-based features

Scope	Feature	Description
Msg.	LENGTH CHARACTERS	Length of the text of the tweet, in characters
	LENGTH WORDS	... in number of words
	CONTAINS QUESTION MARK	Contains a question mark '?'
	CONTAINS EXCLAMATION MARK	... an exclamation mark '!'
	CONTAINS MULTI QUEST OR EXCL.	... multiple question or exclamation marks
	CONTAINS EMOTICON SMILE	... a "smiling" emoticon e.g. :-) ;-) ...
	CONTAINS EMOTICON FROWN	... a "frowning" emoticon e.g. :-(;-(...
	CONTAINS PRONOUN FIRST | SECOND | THIRD	... a personal pronoun in 1st, 2nd, or 3rd person. (3 f
	COUNT UPPERCASE LETTERS	Fraction of capital letters in the tweet
	NUMBER OF URLS	Number of URLs contained on a tweet
	CONTAINS POPULAR DOMAIN TOP 100	Contains a URL whose domain is one of the 100 most
	CONTAINS POPULAR DOMAIN TOP 1000	... one of the 1,000 most popular ones
	CONTAINS POPULAR DOMAIN TOP 10000	... one of the 10,000 most popular ones
	CONTAINS USER MENTION	Mentions a user: e.g. @cnnbrk
	CONTAINS HASHTAG	Includes a hashtag: e.g. #followfriday
	CONTAINS STOCK SYMBOL	... a stock symbol: e.g. $APPL
	IS RETWEET	Is a re-tweet: contains 'RT '
	DAY WEEKDAY	The day of the week in which this tweet was written
	SENTIMENT POSITIVE WORDS	The number of positive words in the text
	SENTIMENT NEGATIVE WORDS	... negative words in the text
	SENTIMENT SCORE	Sum of ±0.5 for weak positive/negative words, ±1.0 fo
User	REGISTRATION AGE	The time passed since the author registered his/her ac
	STATUSES COUNT	The number of tweets at posting time
	COUNT FOLLOWERS	Number of people following this author at posting time
	COUNT FRIENDS	Number of people this author is following at posting ti
	IS VERIFIED	1.0 iff the author has a 'verified' account
	HAS DESCRIPTION	... a non-empty 'bio' at posting time
	HAS URL	... a non-empty homepage URL at posting time
Topic	COUNT TWEETS	Number of tweets
	AVERAGE LENGTH	Average length of a tweet
	FRACTION TWEETS QUESTION MARK	The fraction of tweets containing a question mark '?'
	FRACTION TWEETS EXCLAMATION MARK	... an exclamation mark '!'
	FRACTION TWEETS MULTI QUEST OR EXCL.	... multiple question or exclamation marks
	FRACTION TWEETS EMOTICON SMILE | FROWN	... emoticons smiling or frowning (2 features)
	CONTAINS PRONOUN FIRST | SECOND | THIRD	... a personal pronoun in 1st, 2nd, or 3rd person. (3 f
	FRACTION TWEETS 30PCT UPPERCASE	... more than 30\% of characters in uppercase
	FRACTION TWEETS URL	The fraction of tweets containing a URL
	FRACTION TWEETS USER MENTION	... user mentions
	FRACTION TWEETS HASHTAG	... hashtags
	FRACTION TWEETS STOCK SYMBOL	... stock symbols
	FRACTION RETWEETS	The fraction of tweets that are re-tweets
	AVERAGE SENTIMENT SCORE	The average sentiment score of tweets
	FRACTION SENTIMENT POSITIVE	The fraction of tweets with a positive score
	FRACTION SENTIMENT NEGATIVE	... with a negative score
	FRACTION POPULAR DOMAIN TOP 100	The fraction of tweets with a URL in one of the top-10
	FRACTION POPULAR DOMAIN TOP 1000	... in one of the top-1,000 domains
	FRACTION POPULAR DOMAIN TOP 10000	... in one of the top-10,000 domains
	COUNT DISTINCT EXPANDED URLS	The number of distinct URLs found after expanding sh
	SHARE MOST FREQUENT EXPANDED URL	The fraction of occurrences of the most frequent expan
	COUNT DISTINCT SEEMINGLY SHORTENED URLS	The number of distinct short URLs
	COUNT DISTINCT HASHTAGS	The number of distinct hashtags
	SHARE MOST FREQUENT HASHTAG	The fraction of occurrences of the most frequent hasht
	COUNT DISTINCT USERS MENTIONED	The number of distinct users mentioned in the tweets
	SHARE MOST FREQUENT USER MENTIONED	The fraction of user mentions of the most frequently m
	COUNT DISTINCT AUTHORS	The number of distinct authors of tweets
	SHARE MOST FREQUENT AUTHOR	The fraction of tweets authored by the most frequent a
	AUTHOR AVERAGE REGISTRATION AGE	The average of AUTHOR REGISTRATION AGE
	AUTHOR AVERAGE STATUSES COUNT	The average of AUTHOR STATUSES COUNT
	AUTHOR AVERAGE COUNT FOLLOWERS	... of AUTHOR COUNT FOLLOWERS
	AUTHOR AVERAGE COUNT FRIENDS	... of AUTHOR COUNT FRIENDS
	AUTHOR FRACTION IS VERIFIED	The fraction of tweets from verified authors
	AUTHOR FRACTION HAS DESCRIPTION	... from authors with a description
	AUTHOR FRACTION HAS URL	... from authors with a homepage URL
Prop.	PROPAGATION INITIAL TWEETS	The degree of the root in a propagation tree
	PROPAGATION MAX SUBTREE	The total number of tweets in the largest sub-tree of t
	PROPAGATION MAX | AVG DEGREE	The maximum and average degree of a node that is not
	PROPAGATION MAX | AVG DEPTH	The depth of a propagation tree (0=empty tree, 1=o 2=only re-tweets of the root) and its per-node average
	PROPAGATION MAX LEVEL	The max. size of a level in the propagation tree (except

Information Credibility on Twitter

- Proposed to use a large set of features! We can group the features as:
  - **Message-based features**
  - User-based features
  - Topic-based features
  - Propagation-based features

Scope	Feature
Msg.	LENGTH CHARACTERS
	LENGTH WORDS
	CONTAINS QUESTION MARK
	CONTAINS EXCLAMATION MARK
	CONTAINS MULTI QUEST OR EXCL.
	CONTAINS EMOTICON SMILE
	CONTAINS EMOTICON FROWN
	CONTAINS PRONOUN FIRST | SECOND | THIRD
	COUNT UPPERCASE LETTERS
	NUMBER OF URLS
	CONTAINS POPULAR DOMAIN TOP 100
	CONTAINS POPULAR DOMAIN TOP 1000
	CONTAINS POPULAR DOMAIN TOP 10000
	CONTAINS USER MENTION
	CONTAINS HASHTAG
	CONTAINS STOCK SYMBOL
	IS RETWEET
	DAY WEEKDAY
	SENTIMENT POSITIVE WORDS
	SENTIMENT NEGATIVE WORDS
	SENTIMENT SCORE
User	REGISTRATION AGE
	STATUSES COUNT
	COUNT FOLLOWERS
	COUNT FRIENDS
	IS VERIFIED
	HAS DESCRIPTION
	HAS URL

Information Credibility on Twitter

- Proposed to use a large set of features!
  We can group the features as:
  - Message-based features
  - **User-based features**
  - Topic-based features
  - Propagation-based features

	LENGTH WORDS
	CONTAINS QUESTION MARK
	CONTAINS EXCLAMATION MARK
	CONTAINS MULTI QUEST OR EXCL.
	CONTAINS EMOTICON SMILE
	CONTAINS EMOTICON FROWN
	CONTAINS PRONOUN FIRST | SECOND | THIRD
	COUNT UPPERCASE LETTERS
	NUMBER OF URLS
	CONTAINS POPULAR DOMAIN TOP 100
	CONTAINS POPULAR DOMAIN TOP 1000
	CONTAINS POPULAR DOMAIN TOP 10000
	CONTAINS USER MENTION
	CONTAINS HASHTAG
	CONTAINS STOCK SYMBOL
	IS RETWEET
	DAY WEEKDAY
	SENTIMENT POSITIVE WORDS
	SENTIMENT NEGATIVE WORDS
	SENTIMENT SCORE
User	REGISTRATION AGE
	STATUSES COUNT
	COUNT FOLLOWERS
	COUNT FRIENDS
	IS VERIFIED
	HAS DESCRIPTION
	HAS URL
Topic	COUNT TWEETS
	AVERAGE LENGTH
	FRACTION TWEETS QUESTION MARK
	FRACTION TWEETS EXCLAMATION MARK
	FRACTION TWEETS MULTI QUEST OR EXCL.
	FRACTION TWEETS EMOTICON SMILE | FROWI
	CONTAINS PRONOUN FIRST | SECOND | THIRD
	FRACTION TWEETS 30PCT UPPERCASE
	FRACTION TWEETS URL
	FRACTION TWEETS USER MENTION
	FRACTION TWEETS HASHTAG
	FRACTION TWEETS STOCK SYMBOL
	FRACTION RETWEETS
	AVERAGE SENTIMENT SCORE
	FRACTION SENTIMENT POSITIVE

Information Credibility on Twitter

- Proposed to use a large set of features!
  We can group the features as:
  - Message-based features
  - User-based features
  - **Topic-based features**
  - Propagation-based features

	DAY WEEKDAY
	SENTIMENT POSITIVE WORDS
	SENTIMENT NEGATIVE WORDS
	SENTIMENT SCORE
User	REGISTRATION AGE
	STATUSES COUNT
	COUNT FOLLOWERS
	COUNT FRIENDS
	IS VERIFIED
	HAS DESCRIPTION
	HAS URL
Topic	COUNT TWEETS
	AVERAGE LENGTH
	FRACTION TWEETS QUESTION MARK
	FRACTION TWEETS EXCLAMATION MAR
	FRACTION TWEETS MULTI QUEST OR EX
	FRACTION TWEETS EMOTICON SMILE |
	CONTAINS PRONOUN FIRST | SECOND | T
	FRACTION TWEETS 30PCT UPPERCASE
	FRACTION TWEETS URL
	FRACTION TWEETS USER MENTION
	FRACTION TWEETS HASHTAG
	FRACTION TWEETS STOCK SYMBOL
	FRACTION RETWEETS
	AVERAGE SENTIMENT SCORE
	FRACTION SENTIMENT POSITIVE
	FRACTION SENTIMENT NEGATIVE
	FRACTION POPULAR DOMAIN TOP 100
	FRACTION POPULAR DOMAIN TOP 1000
	FRACTION POPULAR DOMAIN TOP 10000
	COUNT DISTINCT EXPANDED URLS
	SHARE MOST FREQUENT EXPANDED URL
	COUNT DISTINCT SEEMINGLY SHORTENI
	COUNT DISTINCT HASHTAGS
	SHARE MOST FREQUENT HASHTAG
	COUNT DISTINCT USERS MENTIONED
	SHARE MOST FREQUENT USER MENTION
	COUNT DISTINCT AUTHORS
	SHARE MOST FREQUENT AUTHOR
	AUTHOR AVERAGE REGISTRATION AGE
	AUTHOR AVERAGE STATUSES COUNT
	AUTHOR AVERAGE COUNT FOLLOWERS

Information Credibility on Twitter

- Proposed to use a large set of features!
  We can group the features as:
  - Message-based features
  - User-based features
  - Topic-based features
  - **Propagation-based features**

	AVERAGE SENTIMENT SCORE
	FRACTION SENTIMENT POSITIVE
	FRACTION SENTIMENT NEGATIVE
	FRACTION POPULAR DOMAIN TOP 100
	FRACTION POPULAR DOMAIN TOP 1000
	FRACTION POPULAR DOMAIN TOP 10000
	COUNT DISTINCT EXPANDED URLS
	SHARE MOST FREQUENT EXPANDED URL
	COUNT DISTINCT SEEMINGLY SHORTENED U
	COUNT DISTINCT HASHTAGS
	SHARE MOST FREQUENT HASHTAG
	COUNT DISTINCT USERS MENTIONED
	SHARE MOST FREQUENT USER MENTIONED
	COUNT DISTINCT AUTHORS
	SHARE MOST FREQUENT AUTHOR
	AUTHOR AVERAGE REGISTRATION AGE
	AUTHOR AVERAGE STATUSES COUNT
	AUTHOR AVERAGE COUNT FOLLOWERS
	AUTHOR AVERAGE COUNT FRIENDS
	AUTHOR FRACTION IS VERIFIED
	AUTHOR FRACTION HAS DESCRIPTION
	AUTHOR FRACTION HAS URL
Prop.	PROPAGATION INITIAL TWEETS
	PROPAGATION MAX SUBTREE
	PROPAGATION MAX | AVG DEGREE
	PROPAGATION MAX | AVG DEPTH
	PROPAGATION MAX LEVEL

Information Credibility on Twitter

- Obtained the best results with a J48 decision tree

Table 7: Results for the credibility classification.

Class	TP Rate	FP Rate	Prec.	Recall	F_1
A (“true”)	0.825	0.108	0.874	0.825	0.849
B (“false”)	0.892	0.175	0.849	0.892	0.87
W. Avg.	0.860	0.143	0.861	0.860	0.86

DeClarE: Evidence-aware Deep Learning

Goal: Given an arbitrary **textual claim** like

> “the gun epidemic is the leading cause of death of young African-American men, more than the next nine causes put together”

determine whether the claim is **credible** or **non credible**

[1] Kashyap Popat, Subhabrata Mukherjee, Andrew Yates, Gerhard Weikum: DeClarE: Debunking Fake News and False Claims with Evidence-aware Deep Learning. In EMNLP, 2018

DeClarE: Evidence-aware Deep Learning

Four assets:

- **Claim: “**President Obama ordered a life-sized bronze statue of himself to be permanently installed at the White House”
- **Source of the claim: “**conservativeflashnews.com”
- **Evidence about the truth value: “**The emails have made their way across the internet. But reports that Obama ordered a $200,000 life-size bronze statue of himself to be “permanently installed in the White House” are totally false**”**
- **Source of the evidence: “**The Florida Times”

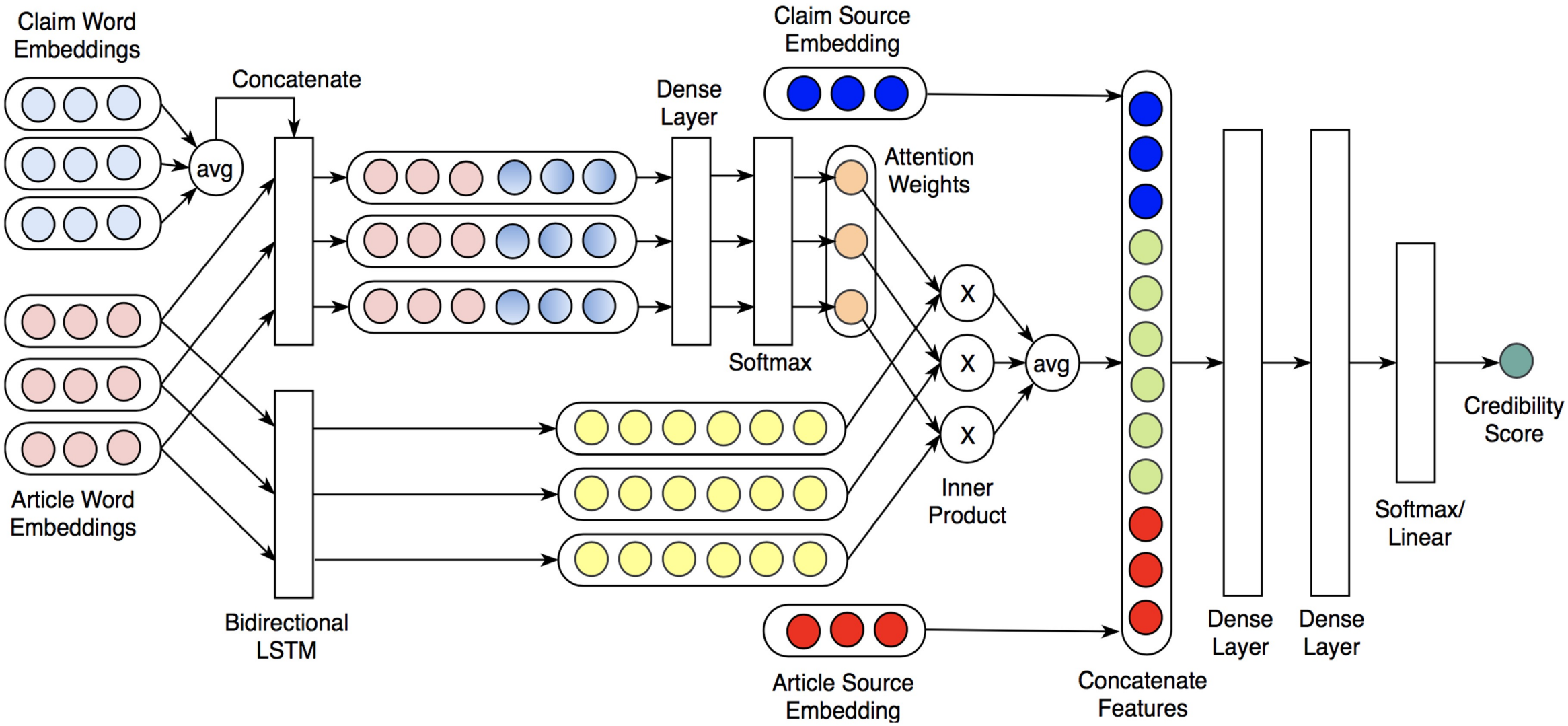
DeClarE: Evidence-aware Deep Learning

Key idea: Learn separate *embeddings* for

- (i) Claim Text, (ii) Claim Source, (iii) Evidence Text, (iv) Evidence Source
- Bidirectional LSTM to model contextual information
- *Attention* to focus on relevant parts of the evidence (article) w.r.t claim context
- Aggregate and predict

[1] Kashyap Popat, Subhabrata Mukherjee, Andrew Yates, Gerhard Weikum: DeClarE: Debunking Fake News and False Claims with Evidence-aware Deep Learning. In EMNLP, 2018

DeClarE: Evidence-aware Deep Learning



DeClarE: Evidence-aware Deep Learning

Rich lexicons + feature engineering (Distant Supervision) outperforms Declare in Snopes

Dataset		*True* Claims Accuracy (%)	*False* Claims Accuracy (%)	Macro F1-Score	AUC
Snopes	LSTM-text	64.65	64.21	0.66	0.70
	CNN-text	67.15	63.14	0.66	0.72
	Distant Supervision	**83.21**	**80.78**	**0.82**	**0.88**
	DeClarE (Plain)	74.37	78.57	0.78	0.83
	DeClarE (Plain+Attn)	78.34	78.91	0.79	0.85
	DeClarE (Plain+SrEmb)	77.43	79.80	0.79	0.85
	DeClarE (Full)	78.96	78.32	0.79	0.86
PolitiFact				0.63	0.66
				0.64	0.67
	Distant Supervision	62.53	62.08	0.62	0.68
	DeClarE (Plain)	62.67	69.05	0.66	0.70
	DeClarE (Plain+Attn)	65.53	68.49	0.66	0.72
	DeClarE (Plain+SrEmb)	66.71	69.28	0.67	0.74
	DeClarE (Full)	**67.32**	**69.62**	**0.68**	**0.75**

Source embedding and Attention improves AUC by 2% - 3%

DeClarE: Evidence-aware Deep Learning

[False] Barbara Boxer: "Fiorina's plan would mean slashing Social Security and Medicare."

Article Source: *nytimes.com*

least of glimmer of truth while ignoring critical facts that would give a different impression mr adair cited a couple examples of barely true claims including this one in california democratic sen barbara boxer claimed that republican challenger carly fiorina s plan would mean slashing social security and medicare but we found there was sketchy evidence to support that fiorina hasn t said much about her ideas on social security and medicare and what she has said doesn t provide much proof of slashing and then there s this one in pennsylvania in the pennsylvania senate race republican pat toomey

---

[True] Hillary Clinton: "The gun epidemic is the leading cause of death of young African-American men, more than the next nine causes put together."

Article Source: *thetrace.org*

away the leading cause of death by francesca mirabile september 27 2016 during the first presidential debate monday night democratic nominee hillary clinton offered a chilling statistic on firearm homicides and the victimization of black males the gun epidemic is the leading cause of death of young african american men more than the next nine causes put together she said data from the centers for disease control and prevention confirms her assertion of all black males between the ages of 15 and 24 that died in 2014 a majority 54 percent were killed with a gun nearly nine in 10

[False] : Coca-Cola's original diet cola drink, TaB, took its name from an acronym for "totally artificial beverage."

Article Source: *foxnews.com*

the first diet colas being the first in 1952 cocacola execs at that time were hesitant to affix the term diet to cocacola so the name tab was chosen as a tribute to those who were keeping tab of their weight according to cola legend the drink was actually dubbed tab as an acronym for totally artificial beverage a great story which unfortunately cocacola says is completely untrue the name was actually chosen by computer and market research the saccharin scandal in the 70s did its damage and the introduction of diet coke in the early 1980s pushed tab even

---

[True] : Household paper shredders can pose a danger to children and pets.

Article Source: *byegoff.com*

packages while still protecting any private information that may be contained in the papers in theory the personal home paper shredder makes much sense personal or pet injuries from paper shredders a growing number of reported injuries reveal that home shredders pose a danger to any user and are especially dangerous to children and pets in fact the federal consumer product safety commission issued a paper shredder safety alert documenting reports of incidents involving finger amputations lacerations and other finger injuries directly connected to the use of home shredders

---

Systems for fact checking

What can we leverage for automated fact checking?

- features from the text (claim, tweet, replies, etc.)
- **background knowledge**
- users' opinions

ExFaKT

- ExFaKT: Explaining Facts over Knowledge Graphs and Text

- **Problem:** Existing Fact-Checking systems return numerical scores. We want to return more “meaningful” explanations
- **Idea:** Use reasoning (rules) to construct human-readable explanations

[1] M. Gad-Elrab, D. Stepanova, J. Urbani, G. Weikum. ExFaKT: A Framework for Explaining Facts over Knowledge Graphs and Text. In WSDM 2019

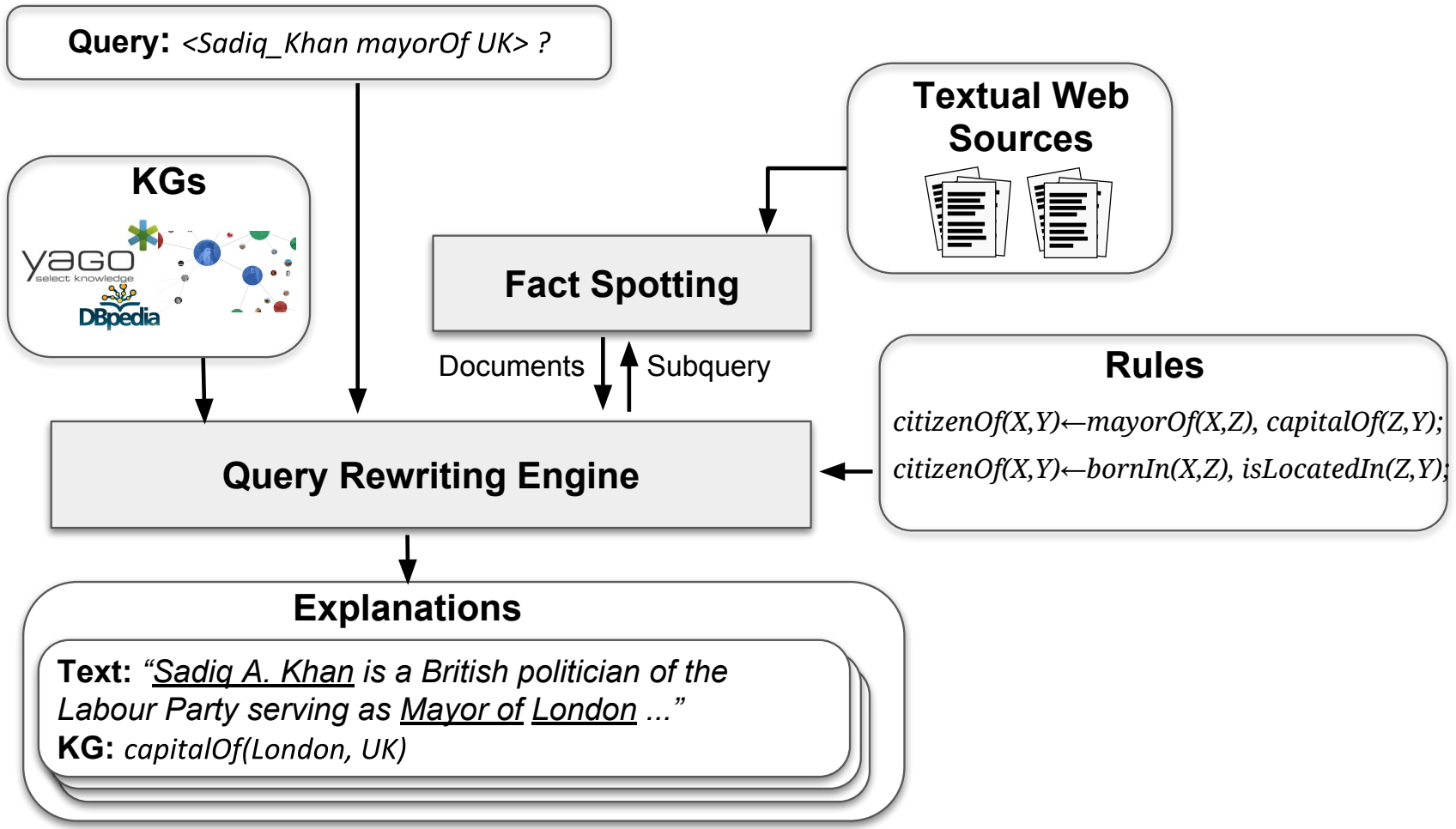
ExFaKT

Main features:

- Consult both KGs and textual documents
  - Mitigate the problems:
    - Bias in entity description
    - Incompleteness of KGs
- Use rules to construct explanations
  - Mitigate the problems:
    - Domain-specific and/or ambiguous interpretation
    - Return human-comprehensible explanations

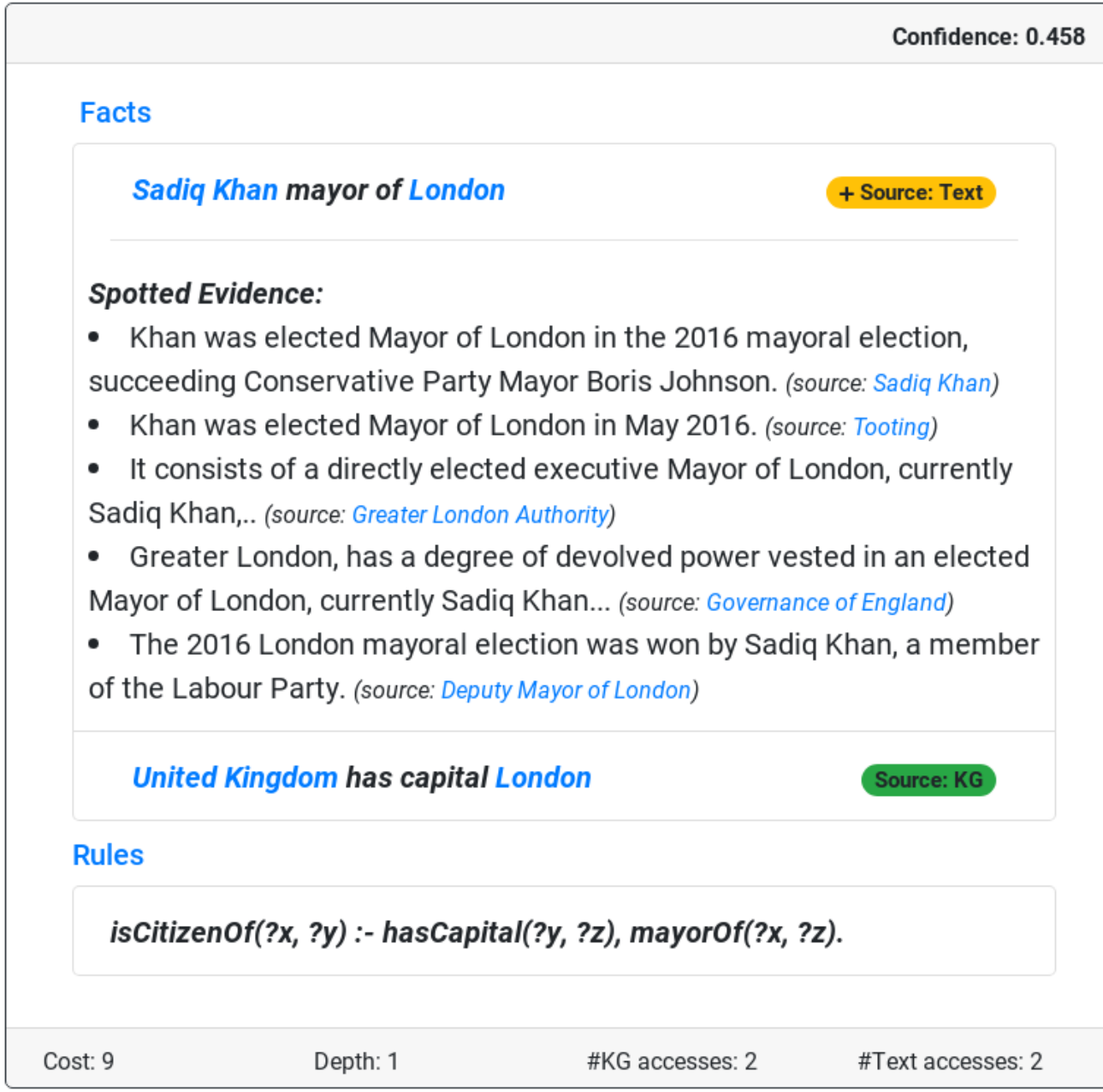
ExFaKT

System overview



ExFaKT

- **Explanations** must be:
  - D1: Concise (few atoms)
  - D2: Close to the query (not long chain of rules)
  - D3: Reliable (based on high quality data)



ExFaKT

- Rules are executed **top-down**
- Reasoning algorithm:
  - Function **explain** computes explanations for a query *q*
  - Function **process_goal** attempts at answering the query or use the rules to *rewrite* it in simpler queries
- Standard datalog engines cannot be used **as-is.**
  - First it explores the search space prioritizing *simple explanations (D1-D2)*
  - It always prefers facts from the KG than from text (D3)

Input: fact q, KG $\mathcal{G}$, text corpus $\mathcal{T}$, ruleset Π, nonegative paramete
max_depth for ensuring termination
Output: set of explanations O

1 **function** *explain*(q, $\mathcal{G}$, $\mathcal{T}$, Π)
2 $depth[q] \leftarrow 0$; $status[q] \leftarrow$ TODO; $P \leftarrow \{\{q\}\}$; $O \leftarrow \{\}$
3 **while** $P \neq \emptyset$ **do**
4 Pick explanation E from P (i.e., $E \in P$)
5 $P \leftarrow P \setminus \{E\}$
6 **if** $status[g] =$FOUND *for all* $g \in E$ **then**
7 $O \leftarrow O \cup \{E\}$ ▷We found a valid explanation.
8 **else**
9 Pick an atom g from E s.t. $status[g] =$TODO
10 $NT \leftarrow process\_goal(g, E \setminus \{g\}, \mathcal{G}, \mathcal{T}, \Pi)$
11 $P \leftarrow P \cup NT$
12 **return** O
13 **function** *process_goal*(g, E, $\mathcal{G}$, $\mathcal{T}$, Π)
14 $O \leftarrow \emptyset$
15 $\Sigma \leftarrow bind(g, \mathcal{G}, \mathcal{T})$
16 $TR \leftarrow \{g\}$
17 **for** $\sigma \in \Sigma$ **do**
18 $a \leftarrow g\sigma$
19 $depth[a] \leftarrow depth[g]$; $status[a] \leftarrow$ FOUND
20 **if** $source[\sigma] =$KG **then** $TR \leftarrow TR \setminus \{a\}$
21 $O \leftarrow O \cup \{E\sigma \cup \{a\}\}$
22 **for** $gr \in TR$ *s.t.* $depth[gr] < max\_depth$ **do**
23 $O \leftarrow O \cup rewrite(gr, E, \Pi)$
24 **return** O

ExFaKT

- Empirical evaluation considering automatically mined rules and manually created
- As KGs, used DBPedia and YAGO
- As text, used Wikipedia and Bing APIs
- **Recall:**

Table 1: Recall of baselines vs. ExFaKT configurations

	B-Wiki	B-Web	KG	Wiki	Web	KG+Wiki	KG+Web
influences	0.30	0.24	0.00	0.38	0.88	0.42	0.92
isPolitOf	0.02	0.16	0.26	0.18	0.88	0.42	0.92
wroteMusic	0.08	0.28	0.00	0.10	0.72	0.24	0.78
mayorOf	0.66	0.9	0.00	0.66	0.90	0.66	0.90
actedWith	0.26	0.52	0.18	0.26	0.60	0.54	0.94
countryWon	0.18	0.38	0.00	0.18	0.38	0.70	0.92
Total	**0.25**	**0.41**	**0.07**	**0.29**	**0.73**	**0.50**	**0.90**

ExFaKT

- Humans judged the quality of the explanations

Config	Candid.	Explan.	Question 1			Question 2	
			Yes	*No*	*Cannot*	Yes	No
B-Wiki	75	75	0.87	0.04	0.10	0.90	0.10
B-Web	122	122	0.85	0.0	0.15	0.80	0.20
B-Search	228	228	0.58	0.01	0.42	0.55	0.45
KG+Wiki	159	311	0.64	0.35	0.02	0.63	0.37
KG+Web	267	1021	0.82	0.01	0.17	0.74	0.26

- Writing rules is easy: 10 students produced 96 rules in 30 minutes

	Strong	**Valid**	**Invalid**	**Total**
Supporting Rules	37	22	10	69
Refutation Rules	10	12	5	27
Total	**47**	**34**	**15**	**96**

Systems for fact checking

What can we leverage for automated fact checking?

- features from the text (claim, tweet, replies, etc.)
- background knowledge
- **users' opinions**

Automatically detecting the stance of users

Example

“*The COVID-19 virus is not heat-resistant and will be killed by a high temperature***”**

User 1 (Twitter): “*There are more COVID-19 infection cases in cold areas*”

User 2 (Twitter): “*COVID-19 virus outbreak in tropical countries is severe*”

Stance classification for fake news

How can we exploit the stance for fake news?

1. We can use it to detect potential controversy regarding certain rumors

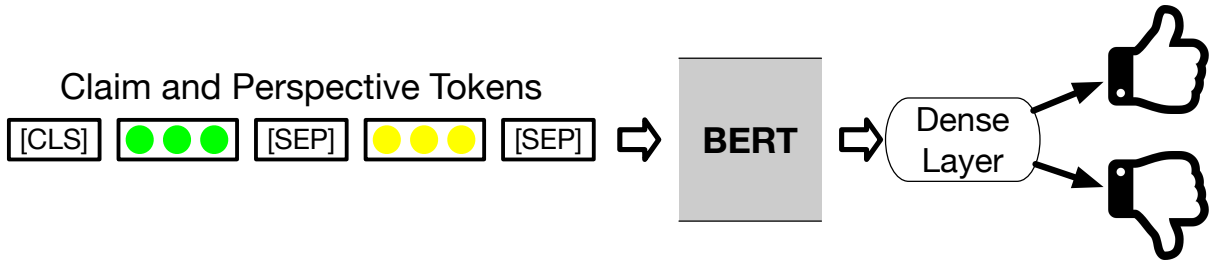
2. We can use the stance as an extra feature of existing fact checking pipelines

3. Combine the stance with the **expertise** (e.g, [1]) of users to decide about the veracity of the rumors

[1] Wagner, Claudia, et al. "It's not in their tweets: Modeling topical expertise of twitter users." *2012 International Conference on Privacy, Security, Risk and Trust and 2012 International Confernece on Social Computing*. IEEE, 2012.

Automatically detecting the stance of users

State-of-the-art methods (e.g., [1]) use BERT



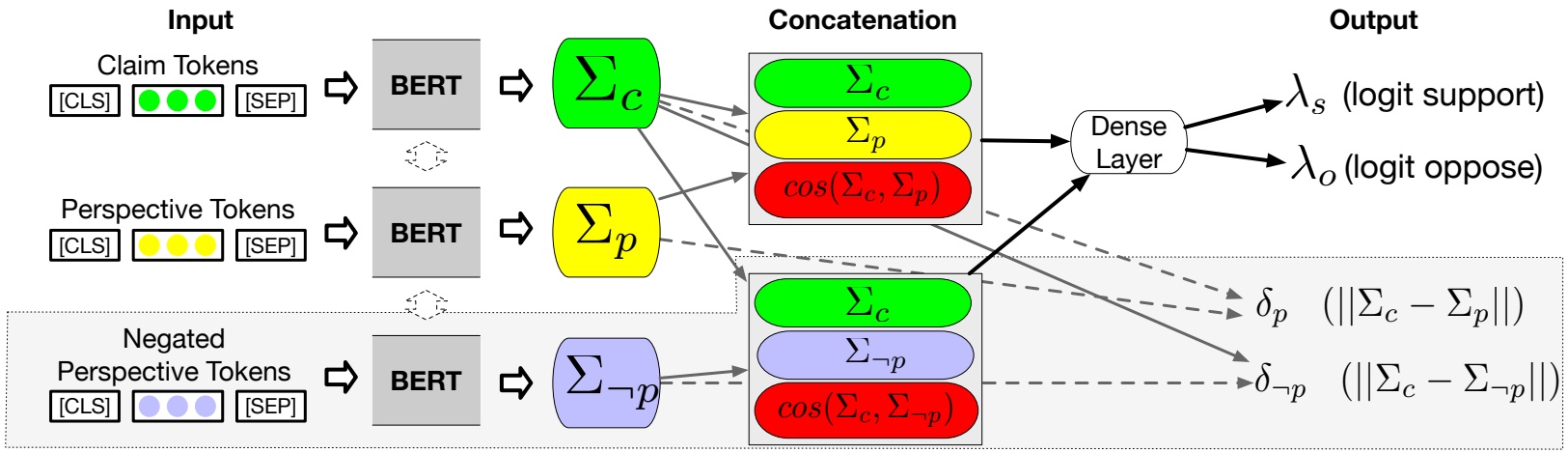
Challenge

- Performance is not (yet) good as humans
- Bias towards **support**

[1] Sihao Chen et al. 2019. Seeing Things from a Different Angle: Discovering Diverse Perspectives about Claims. In *NAACL*.

Automatically detecting the stance of users

Tribrid [1] is a new method that *learns* to trust the predictions



Main novelty

- Perspectives are automatically negated with templates
- BERT is *jointly learned* with original and negated perspectives
- Loss function is designed to improve accuracy and consistency

[1] Song Yang and Jacopo Urbani 2021. Tribrid: Stance Classification with Neural Inconsistency Detection. In *EMNLP*.

Automatically detecting the stance of users

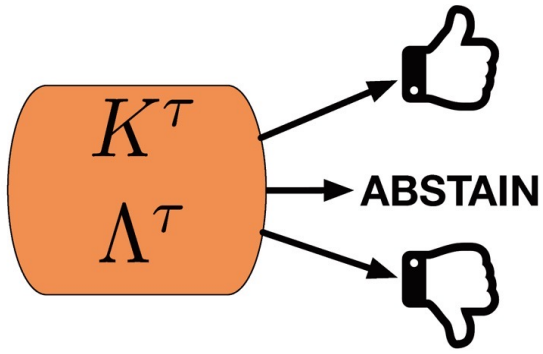
Tribrid can be used to compute confidence scores and **abstain** if the prediction is **inconsistent**

λ_s (logit support) ⇨

λ_o (logit oppose) ⇨

$\delta_p \quad (||\Sigma_c - \Sigma_p||)$ ⇨

$\delta_{\neg p} \quad (||\Sigma_c - \Sigma_{\neg p}||)$ ⇨



Evaluation (with abstain)

Comparison when excluding more and more "inconsistent"[1] predictions

Percentage abstain	10%	30%	70%	90%
BERT	72	76	78	7
STANCY	79	83	1	0
Tribrid with K^{τ}	**82**	**87**	85	87
Tribrid with Λ^{τ}	80	85	**91**	**96**

F1 scores on the PERSPECTRUM dataset

[1] inconsistent means below given threshold τ

Evaluation (with abstain)

Comparison when excluding more and more “inconsistent” predictions

Percentage abstain	10%	30%	70%	90%
BERT	72	76	78	7
STANCY	79	83	1	0
Tribrid with K^{τ}	**82**	**87**	85	87
Tribrid with Λ^{τ}	80	85	**91**	**96**

F1 scores on the PERSPECTRUM dataset

Observation

Competitors do not compute reliable confident predictions

In contrast, our method has high performance with the most confident predictions

Evaluation (with abstain)

Comparison when excluding more and more “inconsistent” predictions

Percentage abstain	10%	30%	70%	90%
BERT	72	76	78	7
STANCY	79	83	1	0
Tribrid with K^{τ}	**82**	**87**	85	87
Tribrid with Λ^{τ}	80	85	**91**	**96**

F1 scores on the PERSPECTRUM dataset

Observations

Performance as good as humans