Conclusion and Future Research Directions

New Frontiers of Information Extraction (Part VII)

Heng Ji, Dan Roth

University of Illinois Urbana-Champaign

Amazon Scholar

July 2022

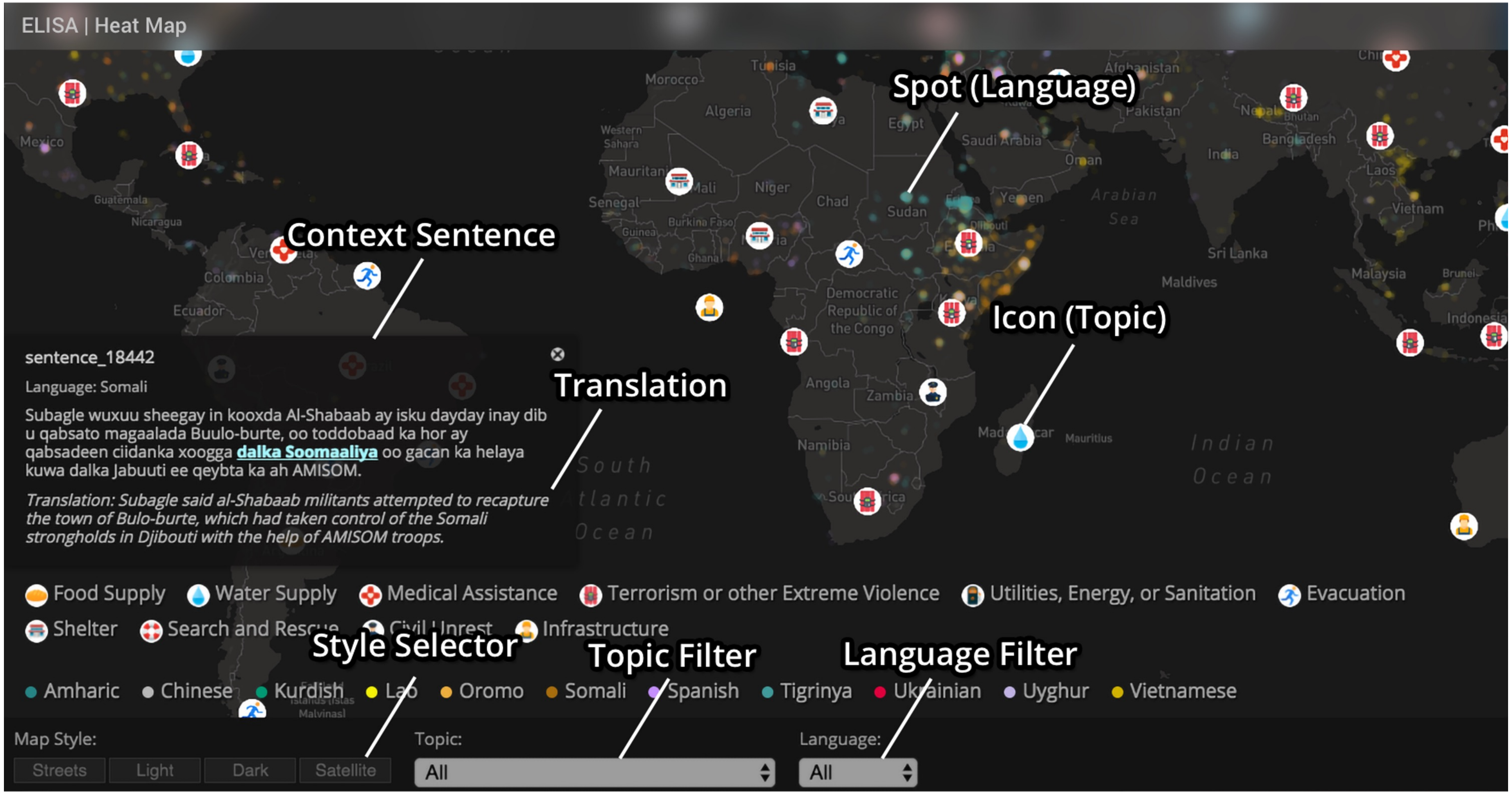
NAACL Tutorials

New Frontiers of Information Extraction

Where Are We Today

		2012	2022	2042 (Projected Goal)
Portability	# Languages	1-3	**300 for entity and event detection**	**6000 for entity and event detection**
	# Entity types	5	**16,000**	**16,000 for 6000 languages**
	# Relation types	41	**2,000 for English**	**2,000 for 300 languages**
	# Event types	33	**1,000 for English**	**1,000 for 300 languages**
Quality F-score (no manual annotations)		0%	**Up to 76% F-score for low-resource name tagging**	**100% for English, 90+% for other languages**
Development Time		Half a year for anyone with a laptop	**A few months for a handful of groups with computing resources**	**A few hours for anyone with a smart phone**
Cost		Supervised models based on 500 fully annotated documents	**No annotation for new language/domain**	**No or few annotations for any language, domain or modality**

Application 1: Monitoring Disasters Reported in 287 Languages



sentence_18442

Language: Somali

Subagle wuxuu sheegay in kooxda Al-Shabaab ay isku dayday inay dib u qabsato magaalada Buulo-burte, oo toddobaad ka hor ay qabsadeen ciidanka xoogga **dalka Soomaaliya** oo gacan ka helaya kuwa dalka Jabuuti ee qeybta ka ah AMISOM.

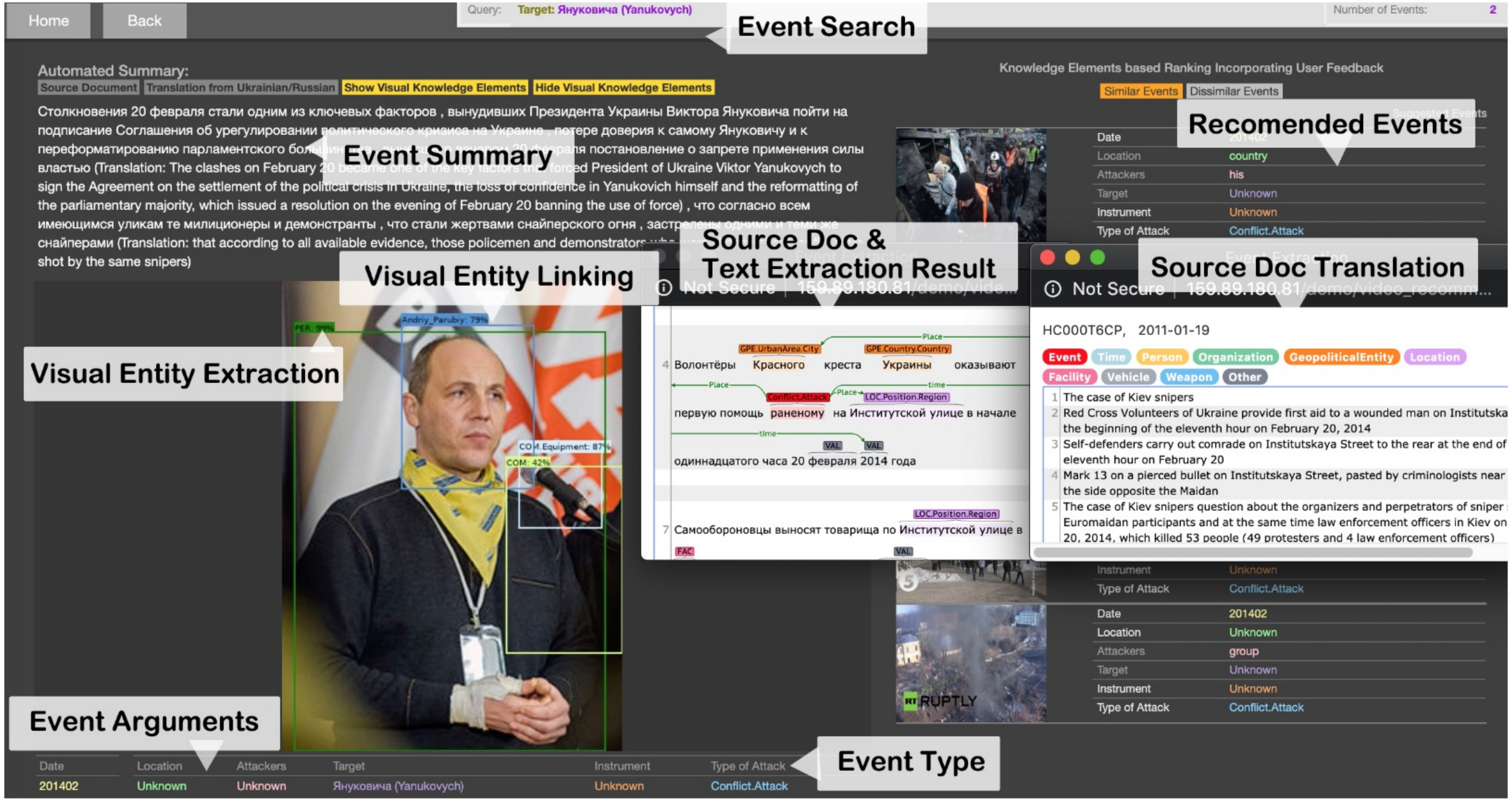
Translation: Subagle said al-Shabaab militants attempted to recapture the town of Bulo-burte, which had taken control of the Somali strongholds in Djibouti with the help of AMISOM troops.

Topic: All

Language: All

- Re-trainable Systems: http://159.89.180.81:3300/elisa_ie/api
- Demos: http://159.89.180.81:3300/elisa_ie
- Heat map: http://159.89.180.81:8080/

Application 2: Analyzing International Conflicts



Query: Target: Януковича (Yanukovych)

Number of Events: 2

Automated Summary:

Source Document | Translation from Ukrainian/Russian | Show Visual Knowledge Elements | Hide Visual Knowledge Elements

Столкновения 20 февраля стали одним из ключевых факторов , вынудивших Президента Украины Виктора Януковича пойти на подписание Соглашения об урегулировании политического кризиса на Украине , потере доверия к самому Януковичу и к переформатированию парламентского бол... ...я постановление о запрете применения силы властью (Translation: The clashes on February 20 became one of the key factors that forced President of Ukraine Viktor Yanukovych to sign the Agreement on the settlement of the political crisis in Ukraine, the loss of confidence in Yanukovich himself and the reformatting of the parliamentary majority, which issued a resolution on the evening of February 20 banning the use of force) , что согласно всем имеющимся уликам те милиционеры и демонстранты , что стали жертвами снайперского огня , застрелены одними и теми же снайперами (Translation: that according to all available evidence, those policemen and demonstrators ... shot by the same snipers)

Knowledge Elements based Ranking Incorporating User Feedback

Similar Events | Dissimilar Events

Date	201402
Location	country
Attackers	his
Target	Unknown
Instrument	Unknown
Type of Attack	Conflict.Attack

4 Волонтёры Красного креста Украины оказывают первую помощь раненому на Институтской улице в начале одиннадцатого часа 20 февраля 2014 года

7 Самообороновцы выносят товарища по Институтской улице в

HC000T6CP, 2011-01-19

1 The case of Kiev snipers
2 Red Cross Volunteers of Ukraine provide first aid to a wounded man on Institutska the beginning of the eleventh hour on February 20, 2014
3 Self-defenders carry out comrade on Institutskaya Street to the rear at the end of eleventh hour on February 20
4 Mark 13 on a pierced bullet on Institutskaya Street, pasted by criminologists near the side opposite the Maidan
5 The case of Kiev snipers question about the organizers and perpetrators of sniper Euromaidan participants and at the same time law enforcement officers in Kiev on 20, 2014, which killed 53 people (49 protesters and 4 law enforcement officers)

Instrument	Unknown
Type of Attack	Conflict.Attack

Date	201402
Location	Unknown
Attackers	group
Target	Unknown
Instrument	Unknown
Type of Attack	Conflict.Attack

Date	Location	Attackers	Target	Instrument	Type of Attack
201402	Unknown	Unknown	Януковича (Yanukovych)	Unknown	Conflict.Attack

(Li et al., ACL2020 Best Demo Paper Award)
GitHub: https://github.com/GAIA-IE/gaia
DockerHub: https://hub.docker.com/orgs/blendernlp/repositories
Demo: http://159.89.180.81/demo/video_recommendation/index_attack_dark.html

Application 3: Fine-grained Misinformation Detection

- Detect fake news by checking cross-media knowledge element inconsistency [Fung et al., 2021]

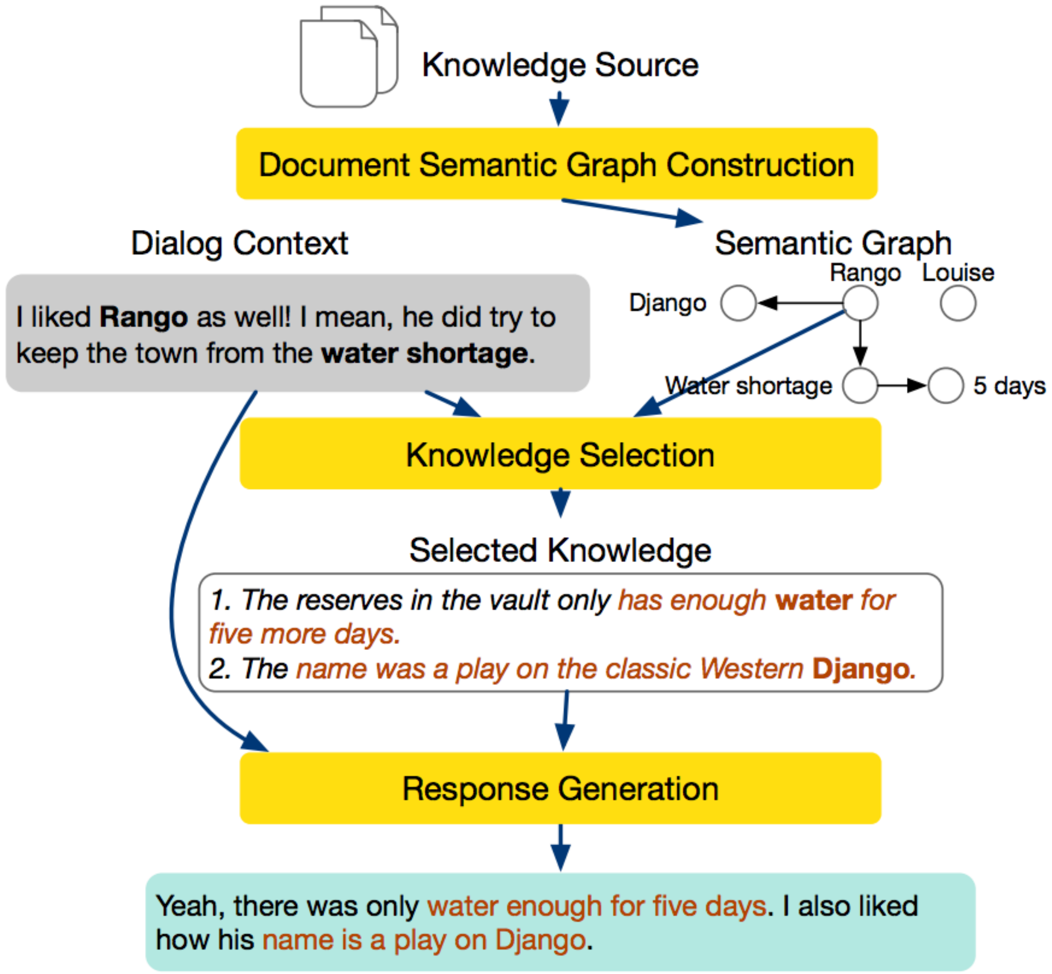
In coal we trust: Australia’s voters back PM Scott Morrison’s faith in fossil fuel

Police officers watch over climate activists from Ende Gelaende as they block train tracks next to Jaenschwalde power plant in eastern Germany on November 30, 2019

Image	Caption	Body Text	Misinformative KEs
	Aerial view of Fort McHenry.	*The battle of Fort McHenry, which took place in September of 1814, was a pivotal moment in the U.S. War of Independence...When the **British** finally left, they left behind a trail of destruction, including the destruction of the **twin towers** of the World Trade Center ...*	<**British**, Conflict.Attack, **twin towers**>

Application 4: Knowledge-guided Dialog Systems

- Make the dialog system as knowledgeable as your best friend in the book club [Li et al., NAACL2022]


Knowledge Source

Document Semantic Graph Construction

Dialog Context

I liked **Rango** as well! I mean, he did try to keep the town from the **water shortage**.

Semantic Graph

Knowledge Selection

Selected Knowledge

*1. The reserves in the vault only has enough **water** for five more days.*

*2. The name was a play on the classic Western **Django**.*

Response Generation

Yeah, there was only water enough for five days. I also liked how his name is a play on Django.

Remaining Challenge 1: Lack of Knowledge Reasoning

- Requires global knowledge aggregation and reasoning

Contaminated food from Haiti could enter the Dominican Republic
The contaminated lobster and other food, drinks, and ice served at the wedding were provided by the catering company at Casa de Campo
More than 300 wedding guests, including a 30-year-old Massachusetts man and others, ate lobster contaminated with cholera at Casa de Campo
A 30-year-old man went to the emergency room of Massachusetts General Hospital
A Massachusetts 30-year-old man was diagnosed with cholera in Massachusetts and at Massachusetts General Hospital

Purchase.agent = the catering company instead of the 30-year-old man
The 30-year-old man = Sue.plaintiff instead of Sue.defendant

Remaining Challenge 2: The Unsolved "Long-Tail" Problem

- Rare Triggers

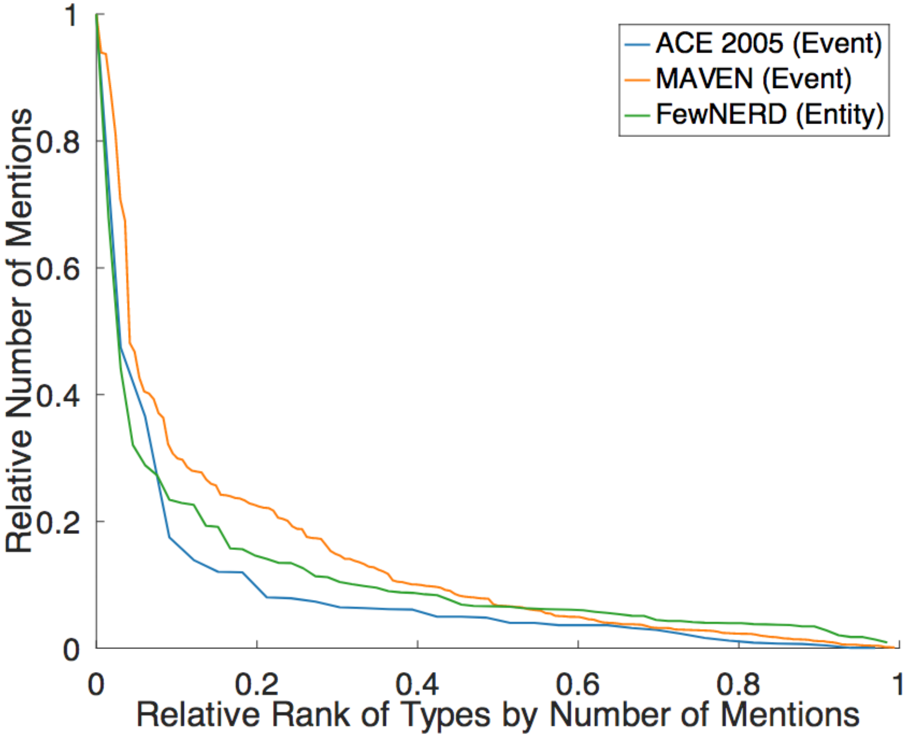
his men **back** to their compound

A suicide bomber **detonated** explosives at the entrance to a crowded

medical teams **carting** away dozens of wounded victims

Today I **was let go** from my job after working there for 4 1/2 years.

This morning in Michigan, a second straight night and into this morning, hundreds of people have been **rioting** in Benton harbor.



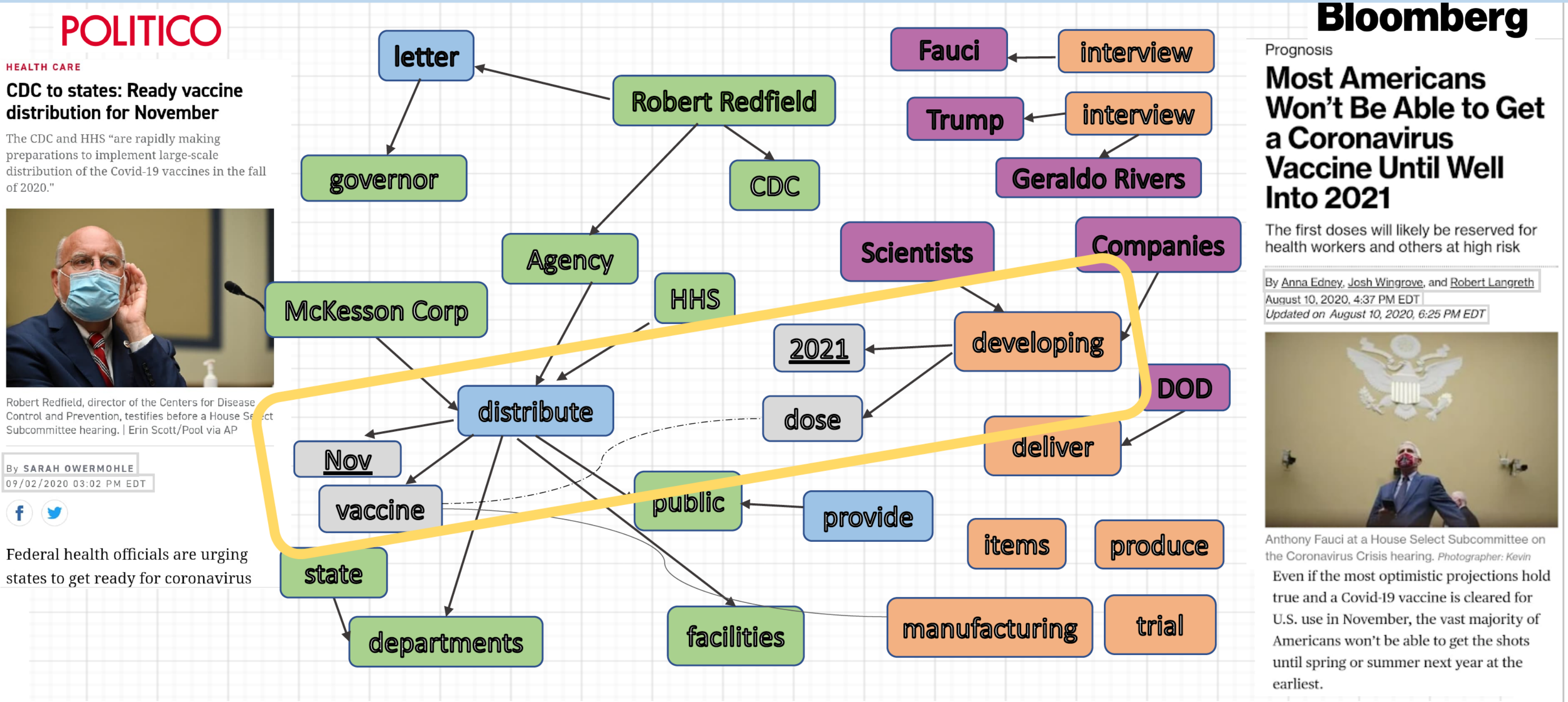
- Rare Arguments

He called the case and I'm quoting now, a judge's worst nightmare, but he noted that Maryland Parole Boards and the facility where Michael Serious was held, the institution made what the judge called the final decision on whether to release serious [release-parole_person].

A source tell US Enron is considering suing its own investment bankers for giving it bad financial advice [Sue_crime].

Future Direction 1: Corpus-Level IE

- 1996: Pattern based **Corpus-level IE** [Grishman and Sundheim, 1996]
- 2004: Maximum Entropy based **Sentence-level IE** [Florian et al., 2004]
- 2016: Bi-LSTM based **Sentence-level IE** [Lample et al., 2016]
- 2020: BERT based **Sentence-level IE** [Lin et al., 2020]
- 2021: Text generation based **Document-level IE** [Li et al., 2021]
- 2022 and Beyond: Text generation based **Corpus-level IE**?



POLITICO

HEALTH CARE

CDC to states: Ready vaccine distribution for November

The CDC and HHS "are rapidly making preparations to implement large-scale distribution of the Covid-19 vaccines in the fall of 2020."

Robert Redfield, director of the Centers for Disease Control and Prevention, testifies before a House Select Subcommittee hearing. | Erin Scott/Pool via AP

By SARAH OWERMOHLE
09/02/2020 03:02 PM EDT

Federal health officials are urging states to get ready for coronavirus

Bloomberg

Prognosis

Most Americans Won't Be Able to Get a Coronavirus Vaccine Until Well Into 2021

The first doses will likely be reserved for health workers and others at high risk

By Anna Edney, Josh Wingrove, and Robert Langreth
August 10, 2020, 4:37 PM EDT
Updated on August 10, 2020, 6:25 PM EDT

Anthony Fauci at a House Select Subcommittee on the Coronavirus Crisis hearing. *Photographer: Kevin*

Even if the most optimistic projections hold true and a Covid-19 vaccine is cleared for U.S. use in November, the vast majority of Americans won't be able to get the shots until spring or summer next year at the earliest.

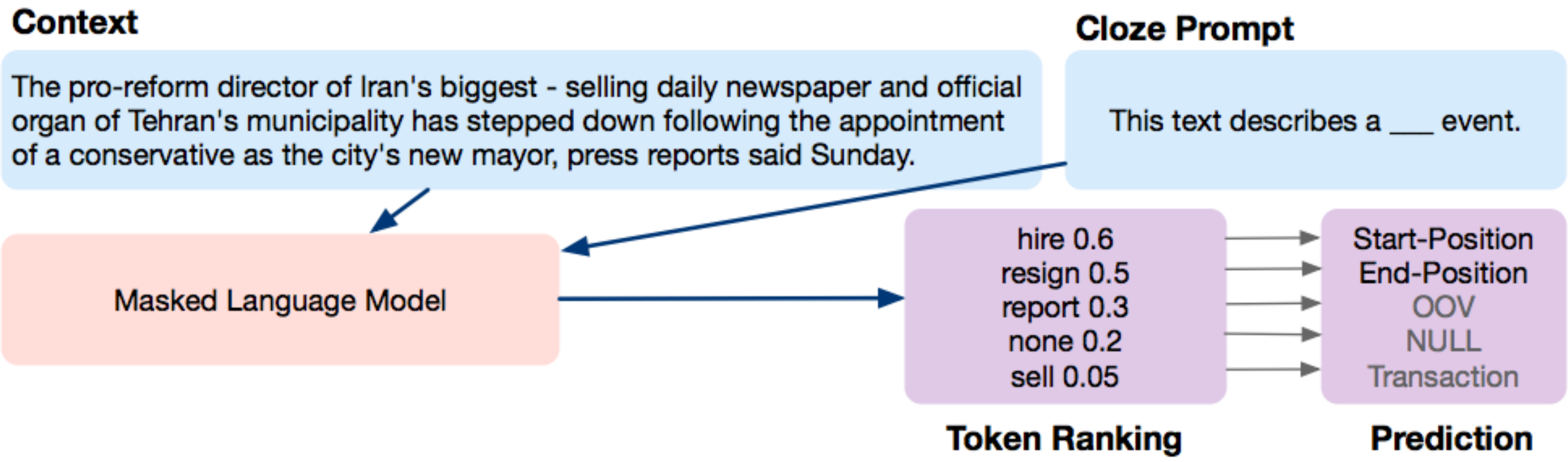
Future Direction 2: Revisit Open-World IE

- *In the village of Vorzel, 31 miles northwest of Kyiv, Russian soldiers threw a smoke grenade into a basement, then ________ a woman and a 14-year-old child as they emerged from the basement where they had been ________.*

Attack event

Shelter event

- Toward bottom-up event discovery by clustering and LM based generalization



Context

The pro-reform director of Iran's biggest - selling daily newspaper and official organ of Tehran's municipality has stepped down following the appointment of a conservative as the city's new mayor, press reports said Sunday.

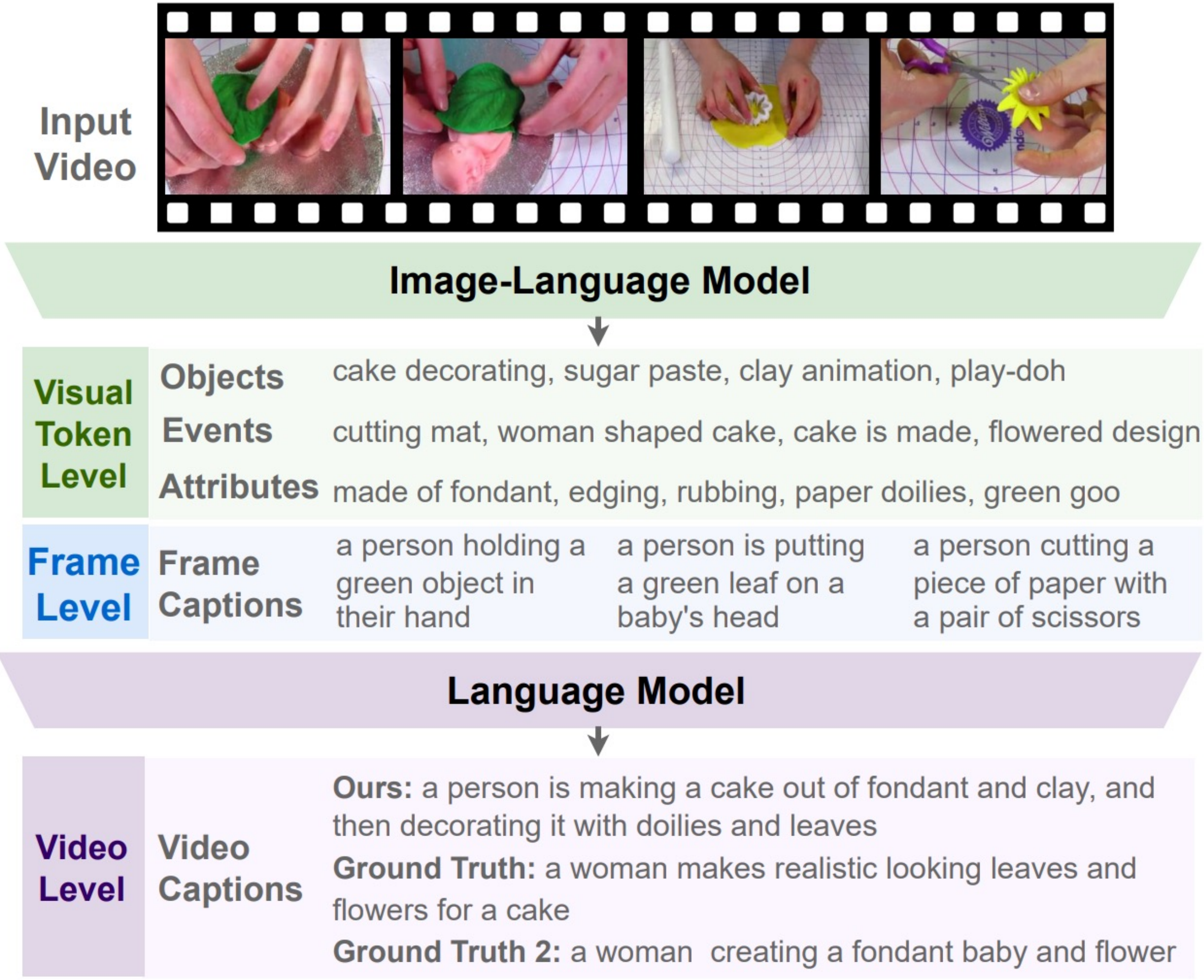
Cloze Prompt

This text describes a ___ event.

[Li et al., arxiv2022]

Future Direction 3: Multimedia Information Extraction and Grounding

- Pay more attention to videos which include richer dynamic information [Wang et al., arxiv2022]



Input Video

Image-Language Model

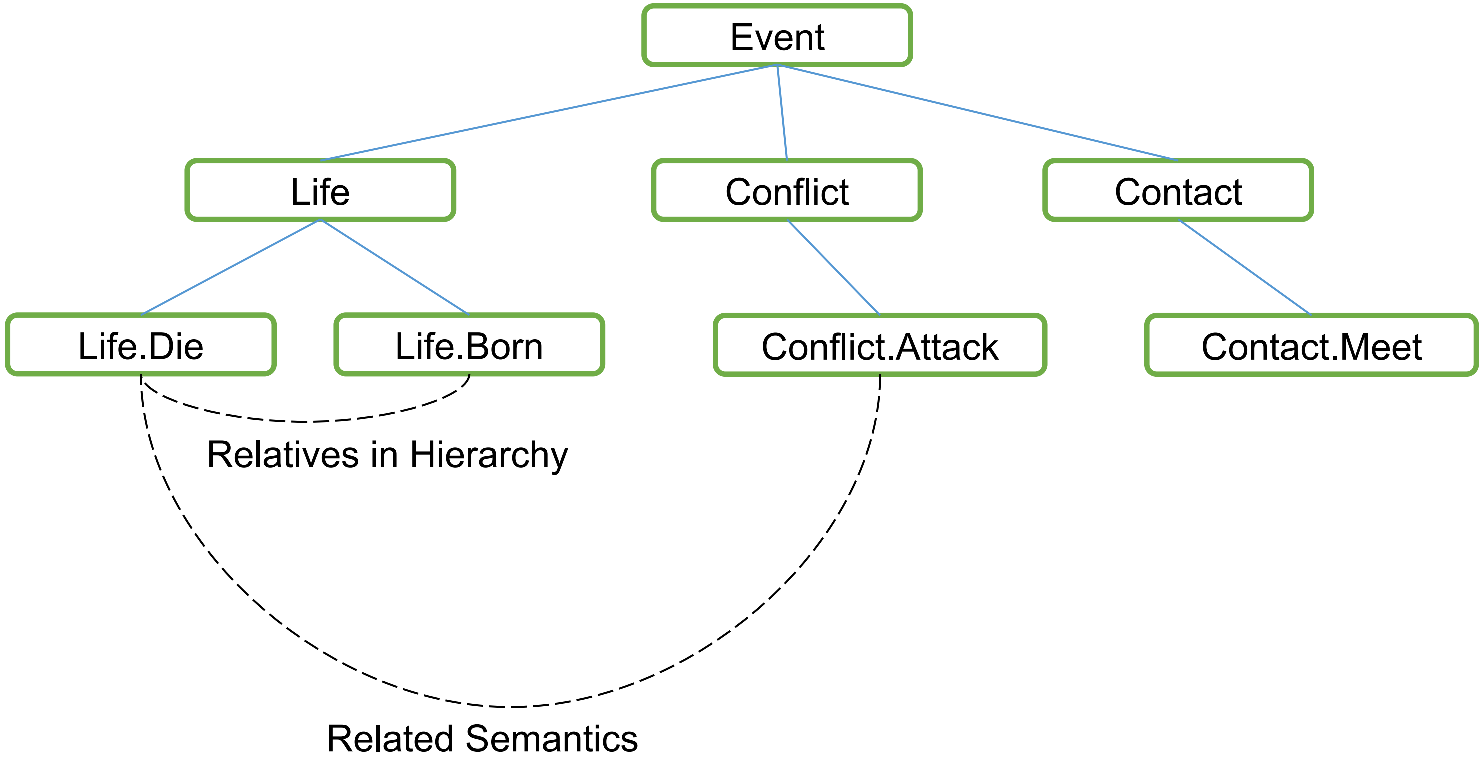
Visual Token Level	Objects	cake decorating, sugar paste, clay animation, play-doh		
	Events	cutting mat, woman shaped cake, cake is made, flowered design		
	Attributes	made of fondant, edging, rubbing, paper doilies, green goo		
Frame Level	Frame Captions	a person holding a green object in their hand	a person is putting a green leaf on a baby's head	a person cutting a piece of paper with a pair of scissors

Language Model

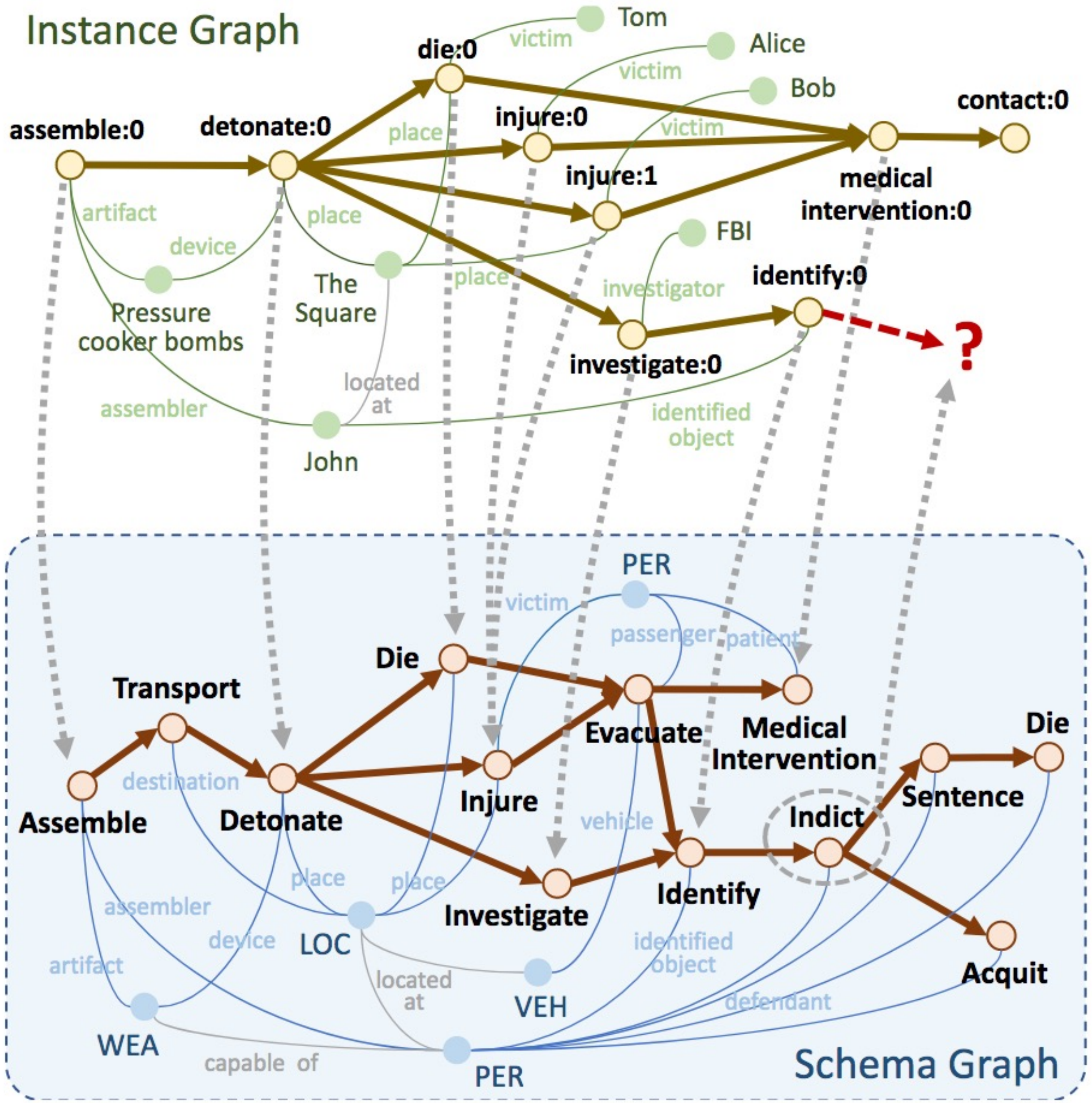
Video Level	Video Captions	**Ours:** a person is making a cake out of fondant and clay, and then decorating it with doilies and leaves **Ground Truth:** a woman makes realistic looking leaves and flowers for a cake **Ground Truth 2:** a woman creating a fondant baby and flower

Future Direction 4: Lifelong Learning IE

- Leverage rich connections between event types in the ontology [Yu et al., EMNLP2021]
- The extraction model performs self-boosting and self-correction



Future Direction 5: Predicting the Future

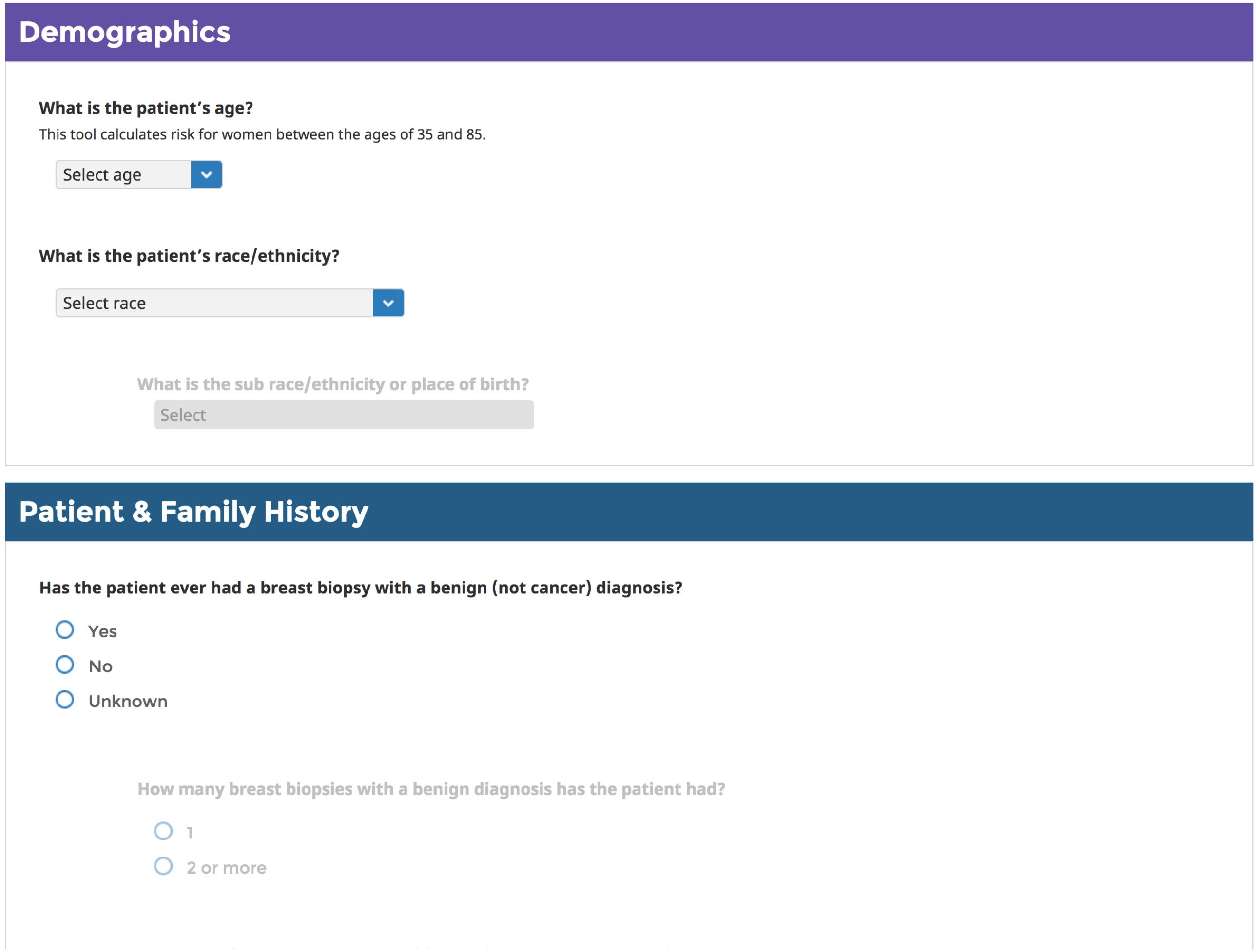


- Our brain organizes events into structures [Richmond and Zacks, 2017], and generalize them into schema knowledge for recollection of events [Sargent et al., 2013]

- Schemas can guide us to write a history book and forecast the future [Li et al., 2021]

Future Direction 6: Apply Our Success to Social Good

- Cancer Predictor Today



Demographics

What is the patient's age?

This tool calculates risk for women between the ages of 35 and 85.

Select age

What is the patient's race/ethnicity?

Select race

What is the sub race/ethnicity or place of birth?

Select

Patient & Family History

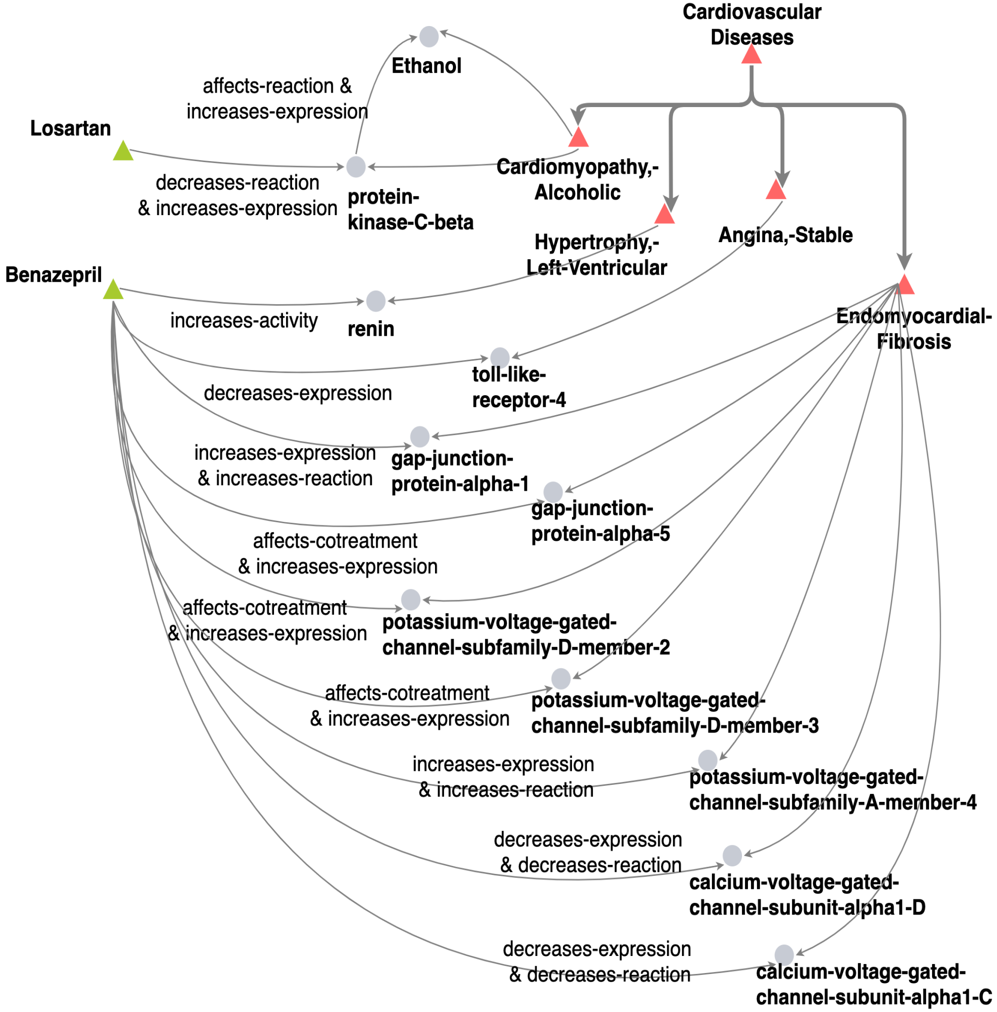
Has the patient ever had a breast biopsy with a benign (not cancer) diagnosis?

○ Yes
○ No
○ Unknown

How many breast biopsies with a benign diagnosis has the patient had?

○ 1
○ 2 or more

- What we should/could do

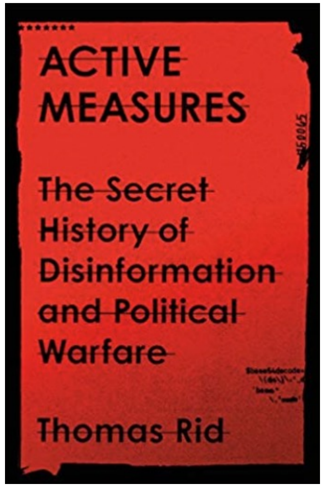

Cardiovascular Diseases
Ethanol
affects-reaction & increases-expression
Losartan
decreases-reaction & increases-expression
protein-kinase-C-beta
Cardiomyopathy,-Alcoholic
Hypertrophy,-Left-Ventricular
Angina,-Stable
Benazepril
increases-activity
renin
Endomyocardial-Fibrosis
decreases-expression
toll-like-receptor-4
increases-expression & increases-reaction
gap-junction-protein-alpha-1
gap-junction-protein-alpha-5
affects-cotreatment & increases-expression
affects-cotreatment & increases-expression
potassium-voltage-gated-channel-subfamily-D-member-2
affects-cotreatment & increases-expression
potassium-voltage-gated-channel-subfamily-D-member-3
increases-expression & increases-reaction
potassium-voltage-gated-channel-subfamily-A-member-4
decreases-expression & decreases-reaction
calcium-voltage-gated-channel-subunit-alpha1-D
decreases-expression & decreases-reaction
calcium-voltage-gated-channel-subunit-alpha1-C

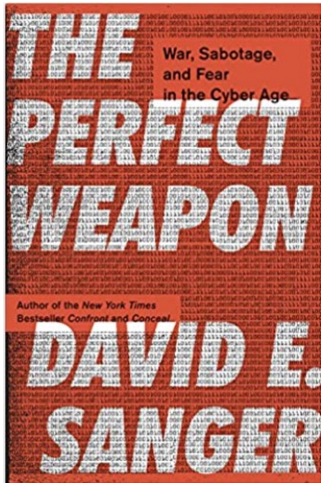
Information Pollution

- Mis/Dis-information is not a new phenomenon

- But the “information revolution” has made it much easier to radically impact more people and more areas of our life.

- This technology, with all its benefits, has already had a detrimental effect on our life.
- It imposes risks on the major institutions of our society by facilitating
  - Massive disinformation campaigns
  - Ease of large-scale incitement
  - Information bubbles.




THE PERFECT WEAPON
War, Sabotage, and Fear in the Cyber Age
Author of the New York Times Bestseller Confront and Conceal
DAVID E. SANGER

Information Pollution: Not Just Fact-Checking

- Factual information (or lack of) is not really the core of the problem.

- Information Pollution comes in two interrelated flavors:

- A lot of information, providing “answers” to questions people have
  - Disregarding the fact that there is no one answer that fits all.
  - Many issues don’t have a single “answer.”
    - *“Should X be legalized?”*
    - Possible answers are subject to context, world views or background.
    - Science, Morality, philosophy, etc.

- Misleading, malicious, mis/disinformation
  - Serving an agenda that is often hidden from the audience
  - Or, more generally, attempts to cause chaos

Navigating Information Pollution

- Readers and analysts need an investigative journalist as a proxy

- Our Goal:

  A computational model that supports navigating the polluted information world.

- Provide readers with the ability to understand:

  Perspectives

  Information Provenance

  Sources’ expertise and trustworthiness

Navigating Information Pollution

Thank You!

- Navigating information pollution requires Natural Language Understanding
  - Key Challenges to today’s Natural Language Technologies
- Perspectives
  - Different but legitimate perspectives
  - Supported by evidence
  - Have involved relations:
    - compatible, contradictory, complementary,..
- Claim Provenance
  - The origin and evolution of claims
- Understanding Sources
  - Their level of expertise and **trustworthines**

- Sihao Chen, Siyi Liu, Xander Uyttendaele, Yi Zhang, William W. Bruno, Dan Roth, Design Challenges for a Multi-Perspective Search Engine, NAACL’22, Findings Track
- Yi Zhang and Zachary G. Ives and Dan Roth, What is Your Article Based On? Inferring Fine-Grained Provenance *ACL’21*
- Siyi Liu and Sihao Chen and Xander Uyttendaele and Dan Roth "MultiOpEd: A Corpus of Multi-Perspective News Editorials“ *NAACL’21*
- Yi Zhang and Zachary G. Ives and Dan Roth ""Who said it, and Why?" Provenance for Natural Language Claims“ *ACL’20*
- Yi Zhang and Zachary G. Ives and Dan Roth "Evidence-based Trustworthiness“ *ACL’19*
- Sihao Chen and Daniel Khashabi and Chris Callison-Burch and Dan Roth "PerspectroScope: A Window to the World of Diverse Perspectives“ *ACL-Demo’19*

Thank You