Discovering Bugs in NLP Models Using Natural Perturbations

Sameer Singh

What is the population of the capital of New York?

circa 2005

News results for population of New York

UN: World Population Aging Rapidly In Developing Countries - RFE/RL

https://www.rferl.org/a/1099361.html ▾

Apr 10, 2002 - A weeklong UN conference in Madrid is warning that the world's population is aging rapidly, with people aged 60 and older poised to ...

The town of the talk - The Economist

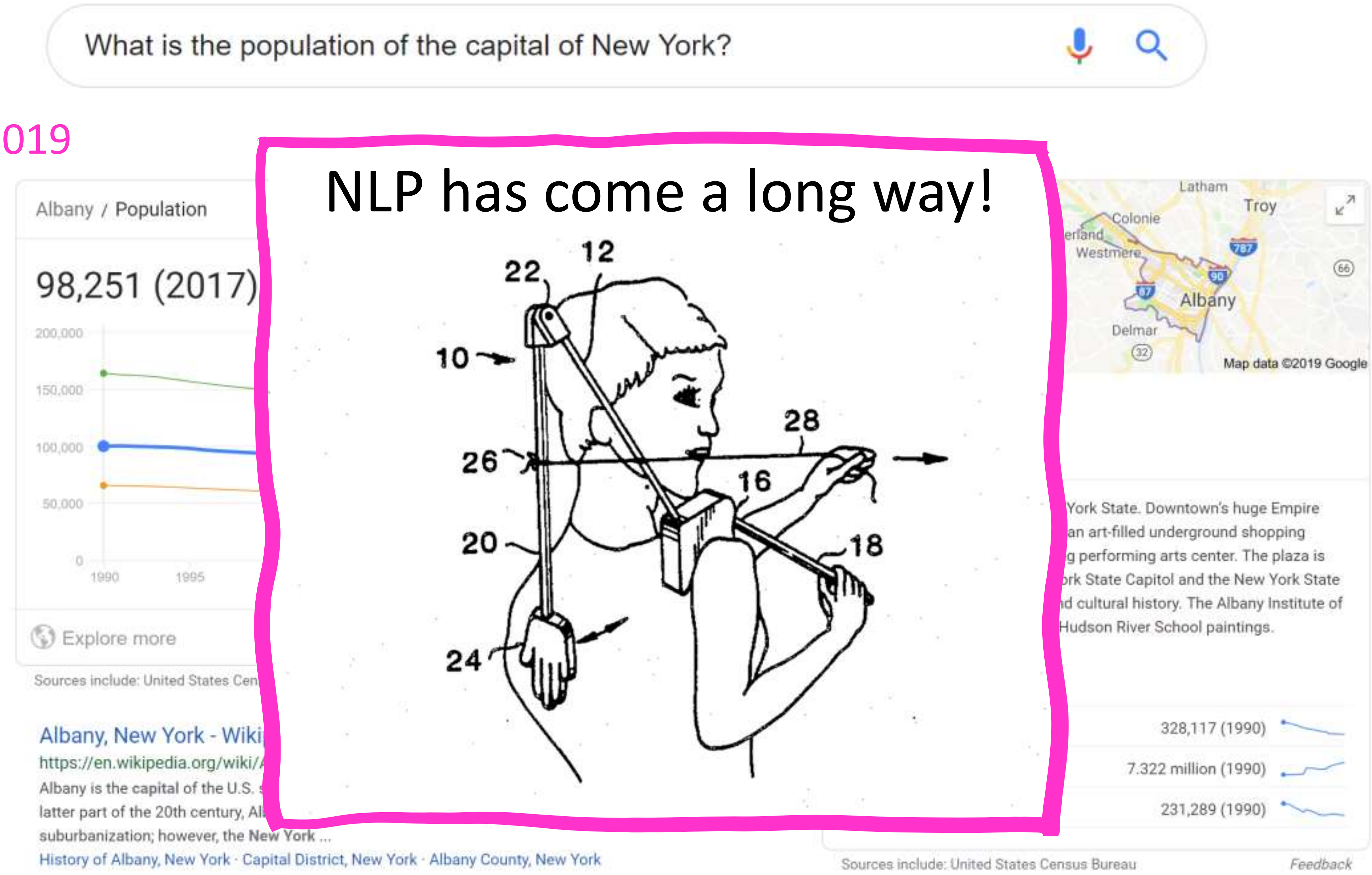
https://www.economist.com/special-report/2005/02/19/the-town-of-the-talk ▾

Feb 19, 2005 - The town of the talk. After the twin-tower nightmare, New York is back on form, says Anthony Gottlieb (interviewed here) ...

[adapted from Zadeh 2005, From Search Engines to Question-Answering Systems — The Need for New Tools]

2019

NLP has come a long way!



But we know models remain brittle…

Anton van den Hengel, ACL 2018

Jia and Liang, EMNLP 2017

Article: Super Bowl 50
Paragraph: "*Peyton Manning became the first quarterback ever to lead two different teams to multiple Super Bowls. He is also the oldest quarterback ever to play in a Super Bowl at age 39. The past record was held by John Elway, who led the Broncos to victory in Super Bowl XXXIII at age 38 and is currently Denver's Executive Vice President of Football Operations and General Manager. Quarterback Jeff Dean had jersey number 37 in Champ Bowl XXXIV.*"
Question: "*What is the name of the quarterback who was 38 in Super Bowl XXXIII?*"
Original Prediction: John Elway
Prediction under adversary: Jeff Dean

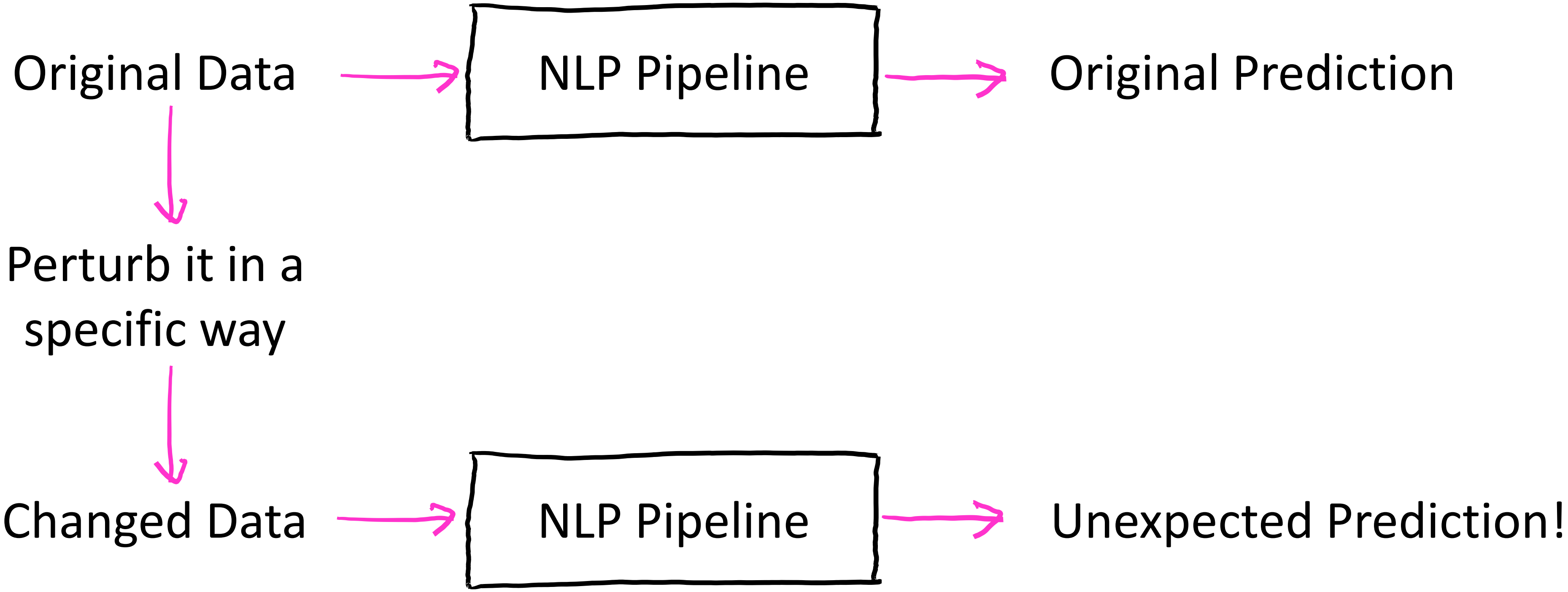
SQuAD

Context	In 1899, John Jacob Astor IV invested \$100,000 for Tesla to further develop and produce a new lighting system. Instead, Tesla used the money to fund his Colorado Springs experiments.
Original	What did Tesla spend Astor's money on ?
Reduced	did
Confidence	$0.78 \rightarrow 0.91$

Feng et al, EMNLP 2018

How do we discover bugs in NLP?

A **software bug** is an error, flaw, failure or fault in a computer program or system that causes it to produce an incorrect or unexpected result, or to behave in unintended ways.



Outline

Changing individual instances

- Semantically Equivalent Adversaries
- Semantically Implied Adversaries
- Universal Adversaries

Changing training data

- Link Prediction Adversaries

Outline

Changing individual instances

Semantically Equivalent Adversaries

Semantically Implied Adversaries

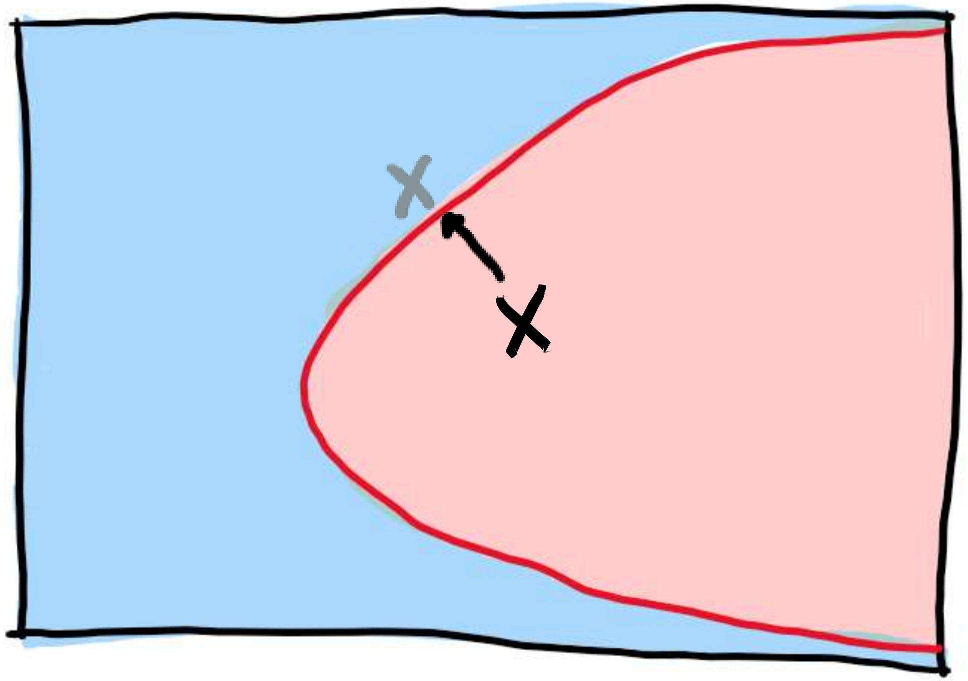
Universal Adversaries

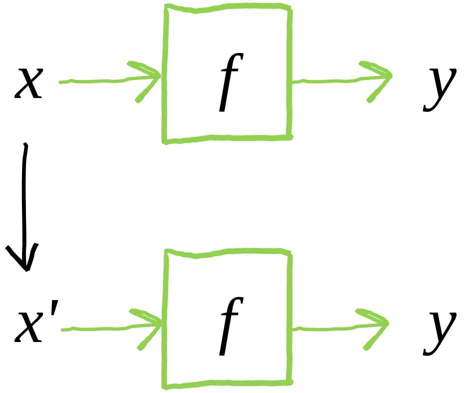
Changing training data

Link Prediction Adversaries

ACL 2018

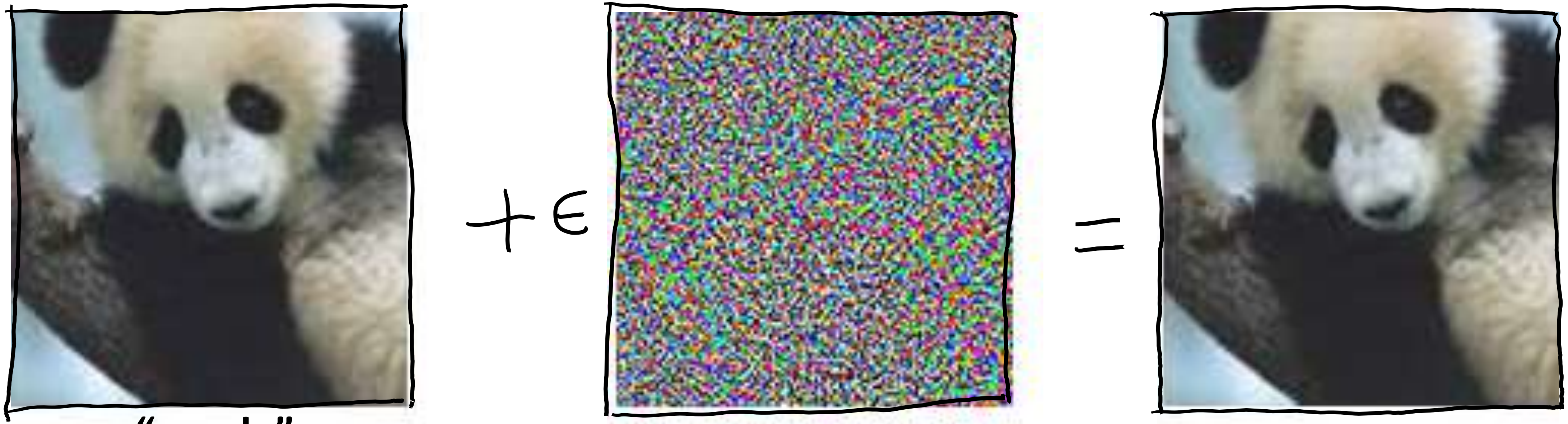
Adversarial Examples: Oversensitivity





Find closest example with different prediction

Oversensitivity in images



“panda”

57.7% confidence

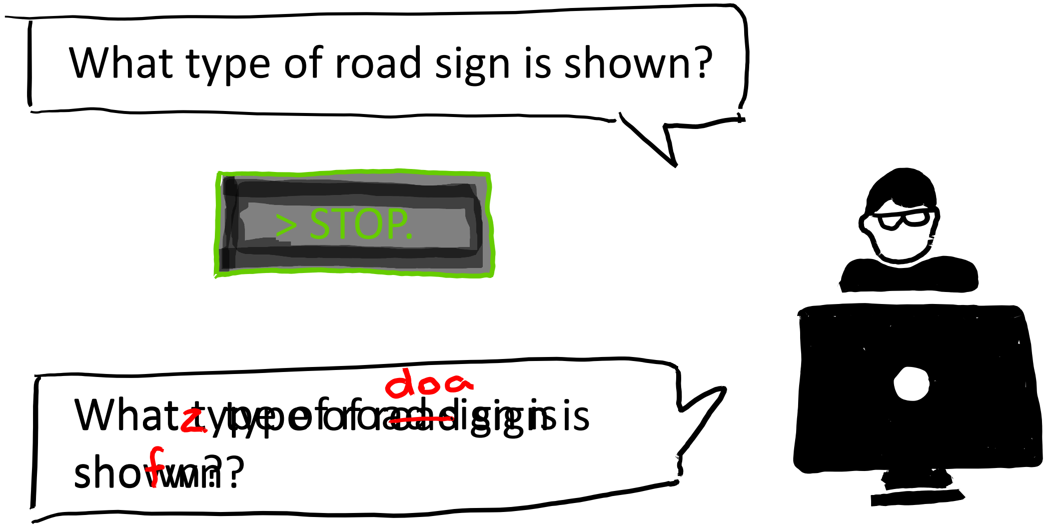
“gibbon”

99.3% confidence

Adversaries are indistinguishable to humans…

But unlikely in the real world (except for attacks)

What about text?



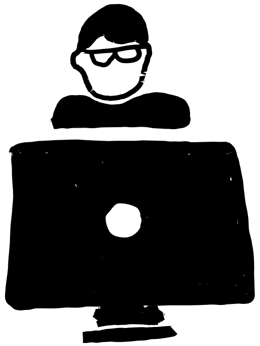
Perceptible by humans, unlikely in real world

What about text?

What type of road sign is shown?

> STOP.

What type of road sign is NOT shown?



A single word changes too much!

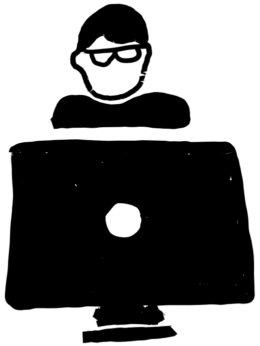
Semantics matter

What type of road sign is shown?

> STOP.

Which
~~What~~ type of road sign is shown?

> Do not Enter.



Bug, and likely in the real world

Semantics matter

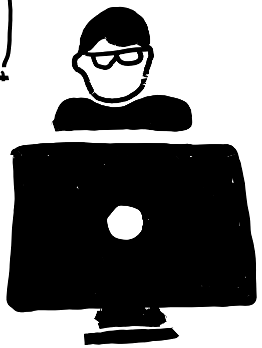
The biggest city on the river Rhine is Cologne, Germany with a population of more than 1,050,000 people.
It is the second-longest river in Central and Western Europe (after the Danube), at about 1,230 km (760 mi)

How long is the Rhine?

> 1230km

How long is the Rhine??

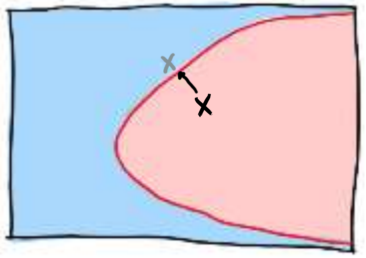
> More than 1,050,000



Not all changes are the same: meaning should be same

How do we do this?

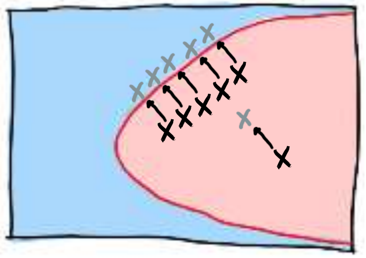
Semantically-Equivalent Adversary
(SEA)



x → Backtranslation + Filtering → x'

What color is the tray?	Pink
What colour is the tray?	Green
Which color is the tray?	Green
What color is it?	Green
How color is tray?	Green

Semantically-Equivalent Adversarial Rules
(SEARs)



(x, x') → Common Patterns → Rules

color → colour

SEARs Examples: VisualQA

SEAR	Questions / SEAs	f(x)	Flips
WP VBZ → **WP's**	~~What has~~ **What's** been cut?	~~Cake~~ **Pizza**	3.3%
What NOUN → **Which NOUN**	~~What~~ **Which** kind of floor is it?	~~Wood~~ **Marble**	3.9%
color → **colour**	What ~~color~~ **colour** is the tray?	~~Pink~~ **Green**	2.2%
ADV is → **ADV's**	~~Where is~~ **Where's** the jet?	~~Sky~~ **Airport**	2.1%

Visual7a-Telling [Zhu et al 2016]

SEARs Examples: SQuAD

SEAR	Questions / SEAs	f(x)	Flips
What VBZ → **What's**	~~*What is*~~ **What's** the NASUWT?	~~Trade union~~ **Teachers in Wales**	2%
What NOUN → **Which NOUN**	~~*What resource*~~ **Which resource** was mined in the Newcastle area?	~~coal~~ **wool**	1%
What VERB → **So what VERB**	~~*What was*~~ **So what was** Ghandi's work called?	~~Satyagraha~~ **Civil Disobedience**	2%
What VBD→ **And what VBD**	~~*What was*~~ **And what was** Kenneth Swezey's job?	~~journalist~~ **sleep**	2%

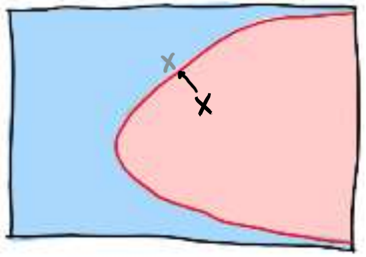
BiDAF [Seo et al 2017]

SEARs Example: Sentiment

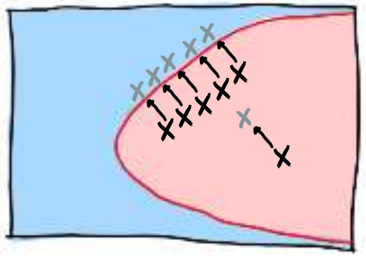
SEAR	Reviews / SEAs	f(x)	Flips
movie → **film**	Yeah, the ~~*movie*~~ **film** pretty much sucked .	~~Neg~~ **Pos**	2%
	This is not ~~*movie*~~ **film** making .	~~Neg~~ **Pos**	
film → **movie**	Excellent ~~*film*~~ **movie** .	~~Pos~~ **Neg**	1%
	I'll give this ~~*film*~~ **movie** 10 out of 10 !	~~Pos~~ **Neg**	
is → **was**	Ray Charles ~~*is*~~ **was** legendary .	~~Pos~~ **Neg**	4%
	It ~~*is*~~ **was** a really good show to watch .	~~Pos~~ **Neg**	
this → **that**	Now ~~*this*~~ **that** is a movie I really dislike .	~~Neg~~ **Pos**	1%
	The camera really likes her in ~~*this*~~ **that** movie.	~~Pos~~ **Neg**	

fastText [Joulin et al., 2016]

Semantic Adversaries



SEA



SEARS

Semantics matter

Models are prone to these bugs

SEAs and SEARs help find and fix them

Outline

Changing individual instances

Semantically Equivalent Adversaries

Semantically Implied Adversaries

Universal Adversaries

ACL 2019

Changing training data

Link Prediction Adversaries

Consistency in predictions

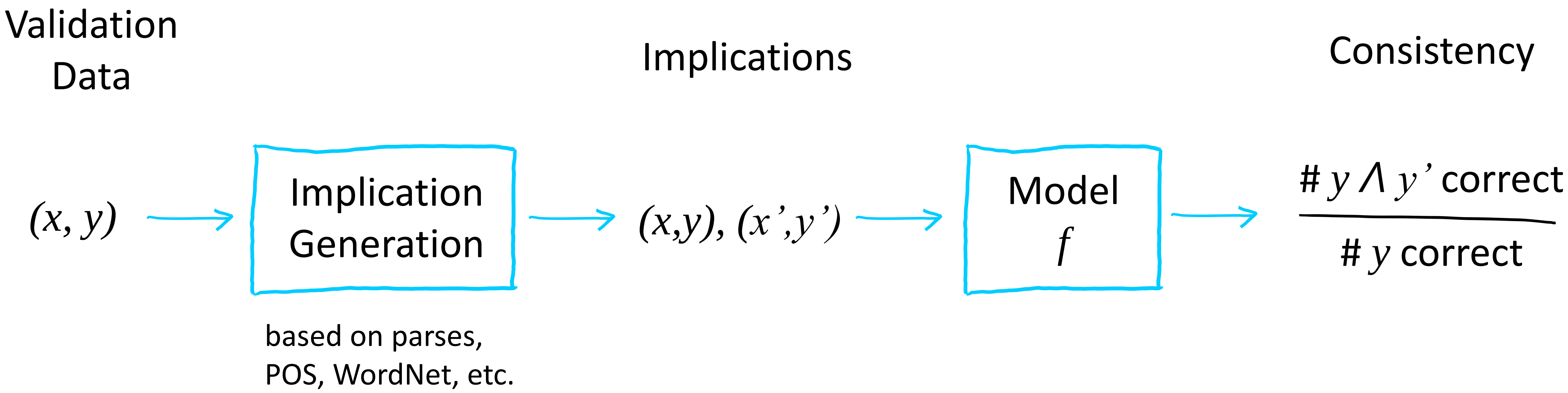
So far, we have considered equivalence, i.e. $(x, y) \rightarrow (x', y)$

(x, y) How many birds? **1**

$\downarrow$ $\downarrow$

(x', y') Is there 1 bird? **Yes**

Evaluating Implication Consistency



Visual QA

(x, y): What room is this? **bathroom**

Logical Equivalence 57%

(x', y'): Is this a bathroom? **Yes**

Necessary Condition 50% **67%**

(x', y'): Is there a bathroom in the picture? **Yes**

97% are valid!

Mutual Exclusion

(x', y'): Is this a kitchen? **No** 35%

Visual QA Results

Model	Acc	LogEq	Mutex	Nec	Avg
SAAA (Kazemi, Elqursh, 2017)	61.5	76.6	42.3	90.2	72.7
Count (Zhang et al., 2018)	65.2	81.2	42.8	92.0	75.0
BAN (Kim et al., 2018)	64.5	73.1	50.4	87.3	72.5

Good at answer w/ numbers, but not questions w/ numbers

e.g. How many birds? **1** (12%) → Are there 2 birds? **yes** (<1%)

SQuAD

Subj	When did Zhenjin die? **1285** → Who died in 1285? **Zhenjin**	29%	
Dobj	When did Denmark join the EU? **1972** → What did Denmark join in 1972? **the EU**	10%	**73%**
Amod	When did the Chinese famine begin? **1331** → Which famine began in 1331? **Chinese**	30%	97% are valid!
Prep	Who received a bid in 1915? **Edison** → When did Edison receive a bid? **1915**	46%	

[Demszky et al. 2018]

SQuAD Results

Model	F1	Subj	Dobj	Amod	Prep	Avg
bidaf (Seo et al., 2017)	77.9	70.6	65.9	75.1	72.4	72.1
bidaf+e (Peters et al., 2018)	81.3	71.2	69.3	75.8	72.8	72.9
rnet (Wang et al., 2017)	79.5	68.5	67.0	74.7	70.7	70.9
Mnem (Hu et al., 2018)	81.5	70.3	68.0	75.8	71.9	72.2

Bad at questions with Wh-word as direct object
e.g. Who is Moses? (53%) vs Who did Hayk defeat? (12%)

Implication Adversaries

- We shouldn’t treat each prediction in isolation
  - Inconsistency leads to poor user experience
- Currently, rule-based system for generating them
- Already promising!
  - Reveals important bugs in the models
  - Even simple data augmentation is promising

Outline

Changing individual instances

Semantically Equivalent Adversaries

Semantically Implied Adversaries

Universal Adversaries

in preparation

Changing training data

Link Prediction Adversaries

Universal Adversaries

REDACTED for anonymity period

How do we do this?

REDACTED for anonymity period

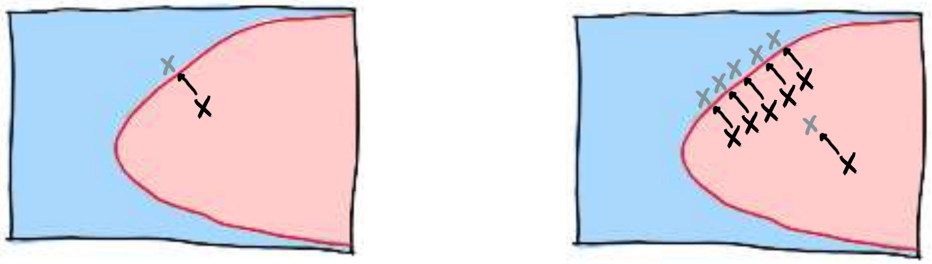
Textual Entailment

REDACTED for anonymity period

Language Modeling (GPTv2 small)

REDACTED for anonymity period

Changing Instances



- “Adversarial attacks” for NLP
  - Semantically Equivalent
  - Semantic Implications
  - Universal Tokens
- Useful for identifying different kinds of problems
  - Not all of them are traditional “bugs”
- General set of approaches that apply for most NLP models

Outline

Changing individual instances

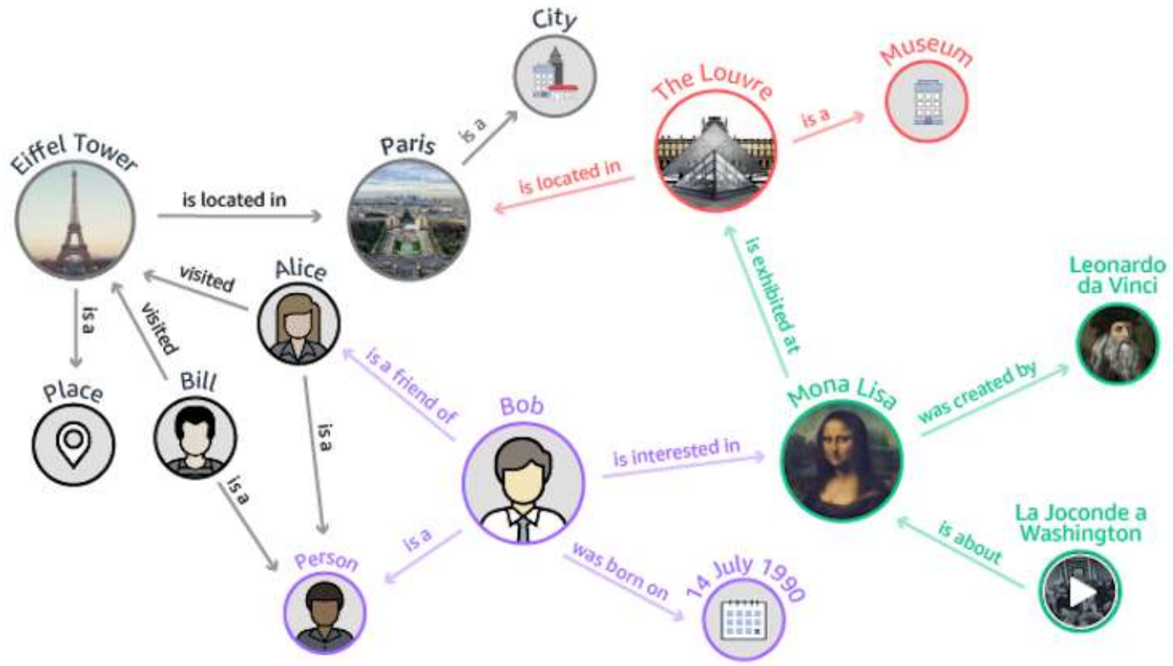
- Semantically Equivalent Adversaries
- Semantically Implied Adversaries
- Universal Adversaries

Changing training data

- Link Prediction Adversaries

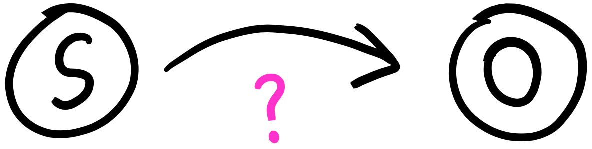
NAACL 2019

Different Kind of Model: Link Prediction

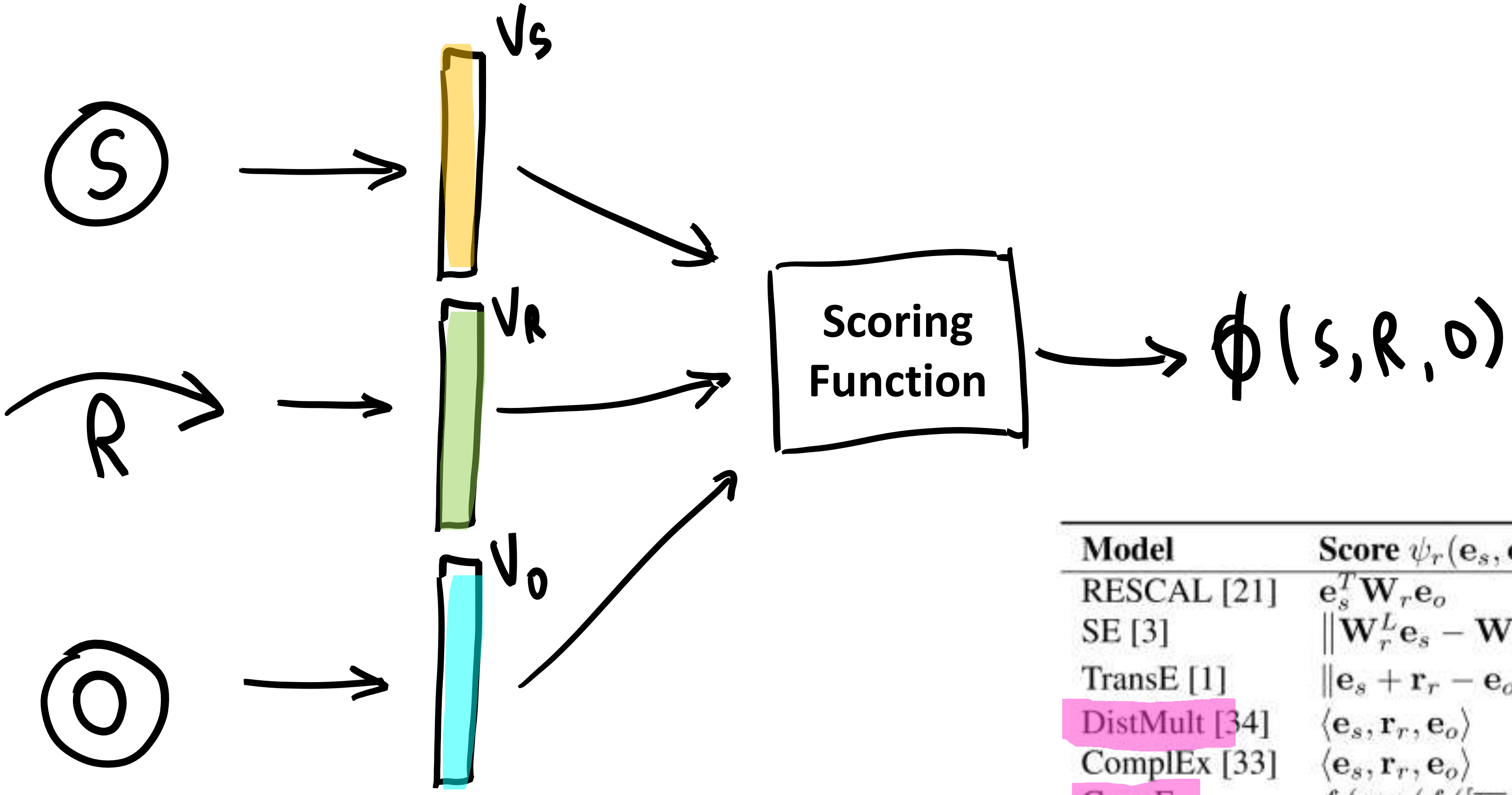


Entity Prediction

Relation Prediction



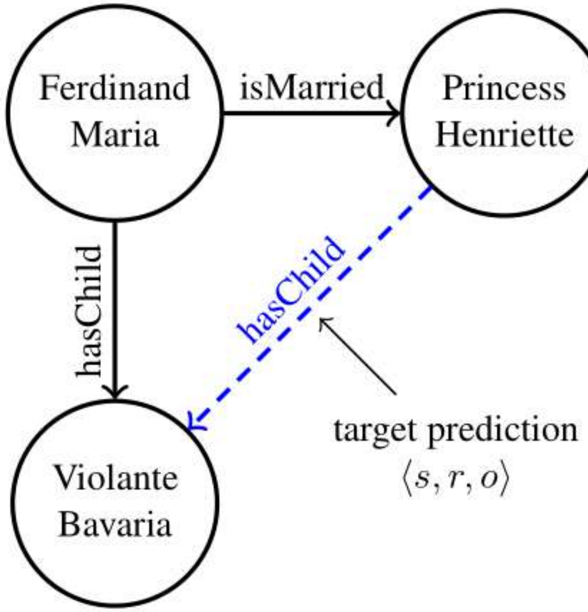
Knowledge Base Completion



Model	Score $\psi_r(\mathbf{e}_s, \mathbf{e}_o)$
RESCAL [21]	$\mathbf{e}_s^T \mathbf{W}_r \mathbf{e}_o$
SE [3]	$\left\| \mathbf{W}_r^L \mathbf{e}_s - \mathbf{W}_r^R \mathbf{e}_o \right\|_p$
TransE [1]	$\| \mathbf{e}_s + \mathbf{r}_r - \mathbf{e}_o \|_p$
DistMult [34]	$\langle \mathbf{e}_s, \mathbf{r}_r, \mathbf{e}_o \rangle$
ComplEx [33]	$\langle \mathbf{e}_s, \mathbf{r}_r, \mathbf{e}_o \rangle$
ConvE	$f\left(\mathrm{vec}\left(f\left([\overline{\mathbf{e}_s}; \overline{\mathbf{r}_r}] * \omega\right)\right) \mathbf{W}\right) \mathbf{e}_o$

Table from Dettmers, et al. (2018)

Link Prediction Example

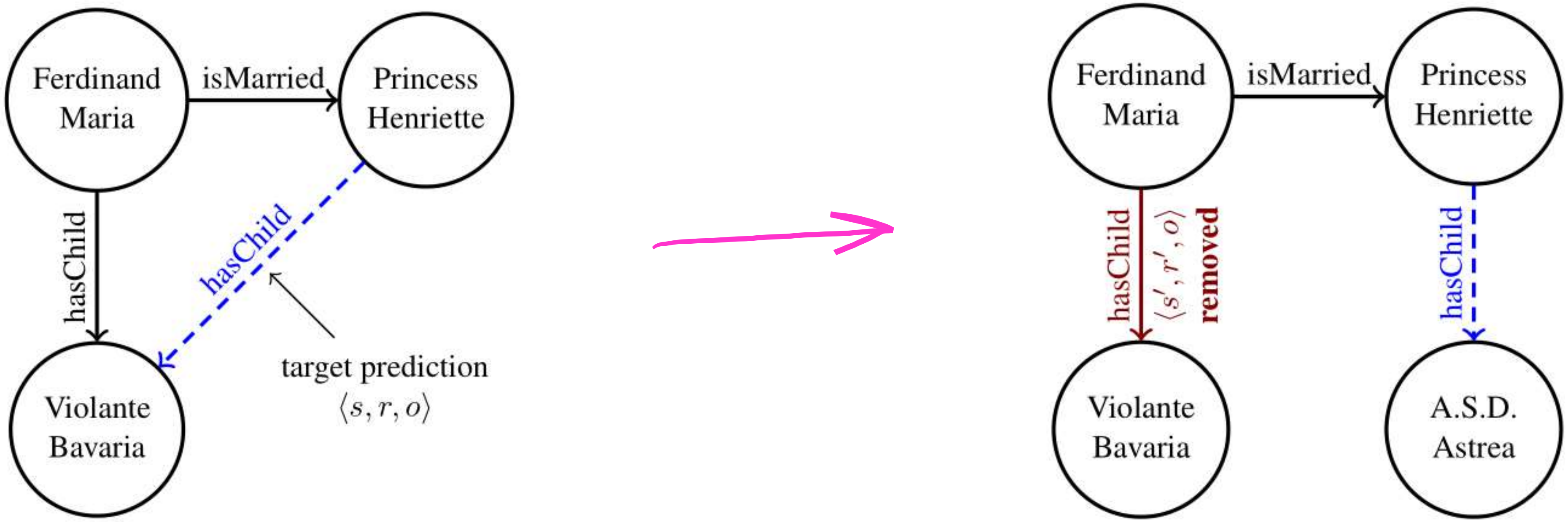


Why was this prediction made?

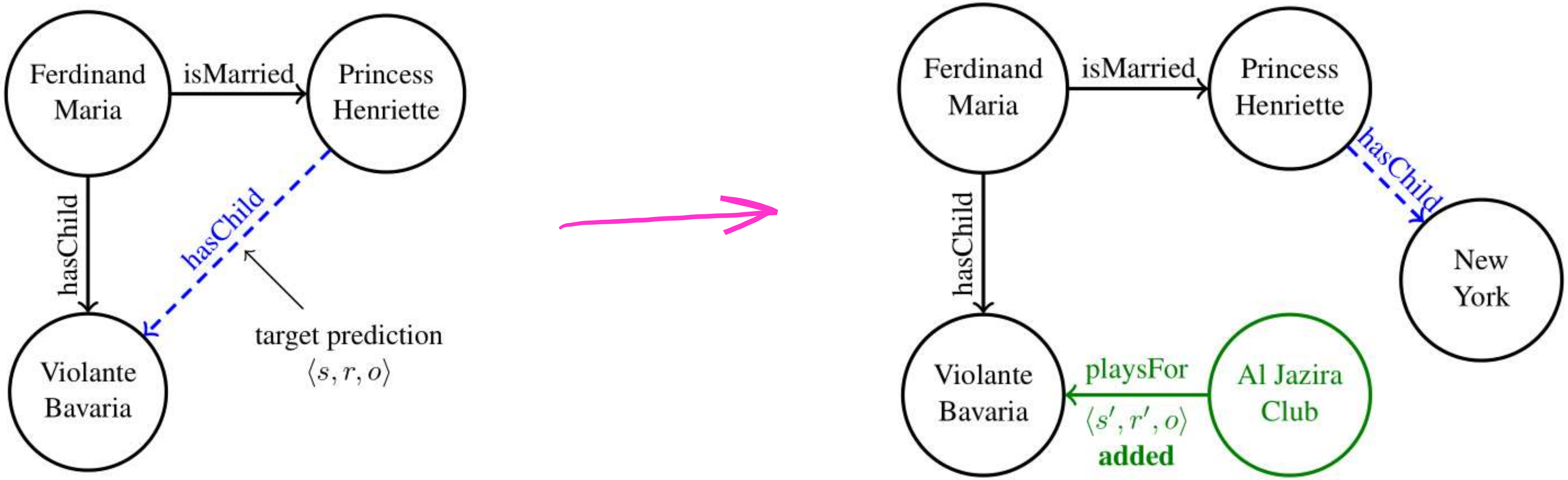
What is this sensitive to?

Depends on the graph structure!

Link Prediction: Removing a Link



Link Prediction: Adding a Link



How do we do it?

$$\underset{(s',r')}{\operatorname{argmax}} \;\; \phi(s,r,o) - \bar{\phi}(s,r,o)$$

Link to add/remove from the graph

Original score

Score *after retraining*

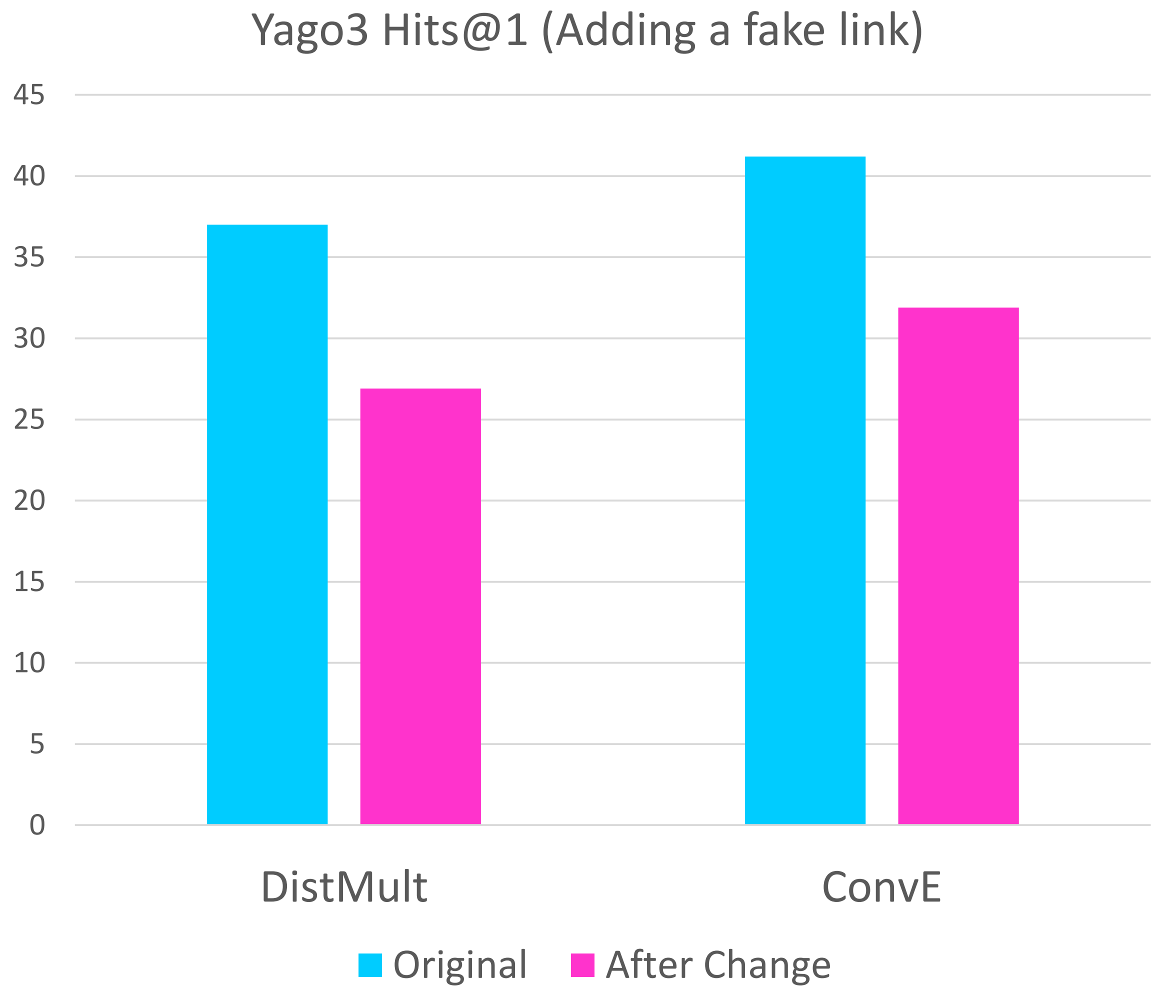
Retraining is too expensive!

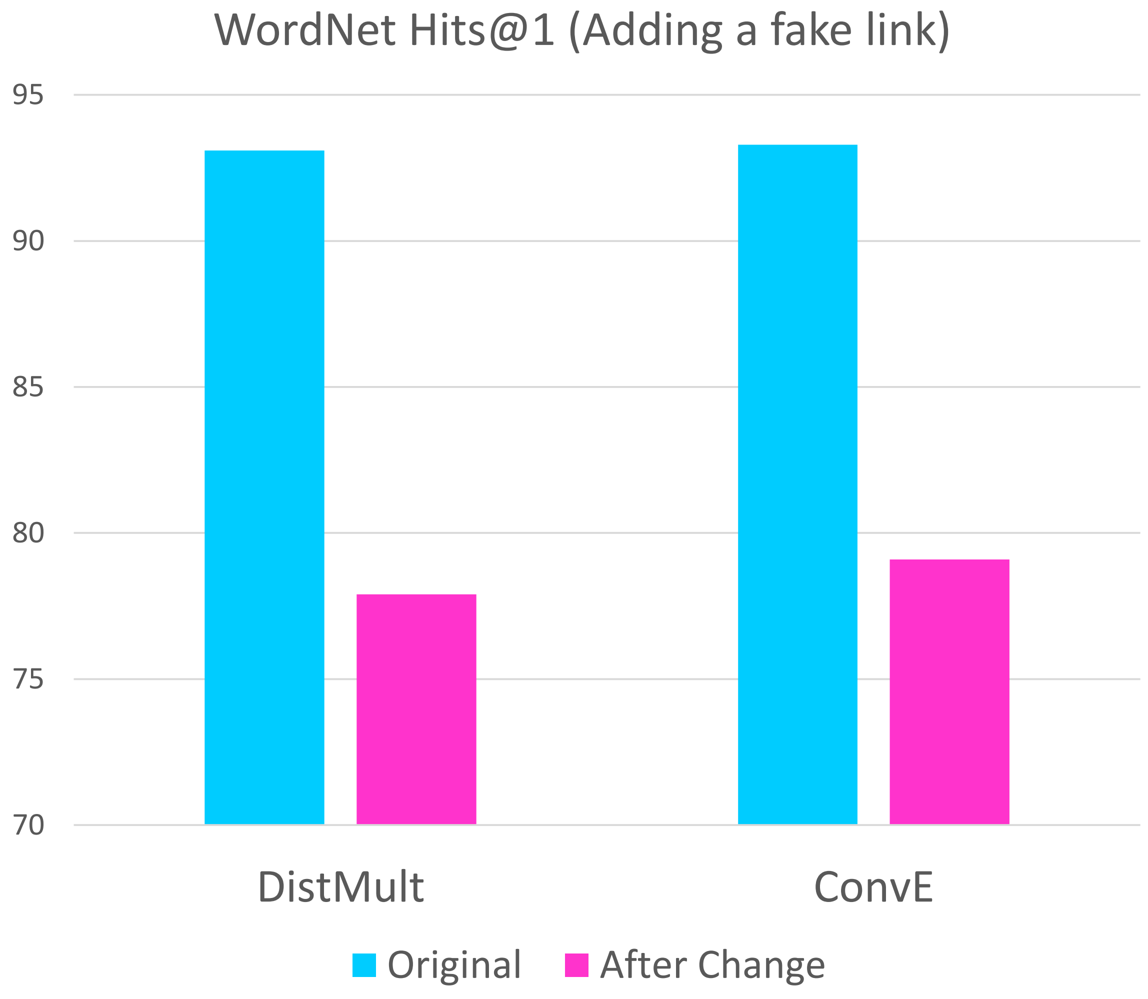
Too many links to search!

Learn a continuous space of links, and search using gradient descent

Taylor approximation, and utilize graph structure

Adding Links: How sensitive is the model?





Removing Links: Cause behind prediction

Summarize by rule mining on which edges are used

Bug in DistMult and ConvE

isMarriedTo(a,c)∧ hasChild(c,b) ⇒ hasChild(a,b)

Only in DistMult

playsFor(a,c) ∧ isLocatedIn(c,b) ⇒ wasBornIn(a,b)*

isAffiliatedTo(a,c)∧ isLocatedIn(c,b) ⇒ diedIn(a,b)*

Only in ConvE

hasAdvisor(a,c)∧ graduatedFrom(c,b) ⇒ graduatedFrom(a,b)

influences(a,c)∧ influences(c,b) ⇒ influences(a,b)

* Identified as rules by [Yang et. al. 2015]

Changing the training data

- Sometimes, “bugs” are problems in the training data/pipeline
  - Embeddings of all kinds, for example
- To find these bugs, you need to change the training data
  - And efficiently estimate the effect of retraining
- We show how to do that for link prediction

Outline

Changing individual instances

- Semantically Equivalent Adversaries
- Semantically Implied Adversaries
- Universal Adversaries

Changing training data

- Link Prediction Adversaries






Thanks!

sameer@uci.edu
sameersingh.org
@sameer_

Work with **Matt Gardner** and me

as part of
The Allen Institute for
Artificial Intelligence
in **Irvine**, CA

All levels: pre-PhD, PhD interns, postdocs, and research scientists!