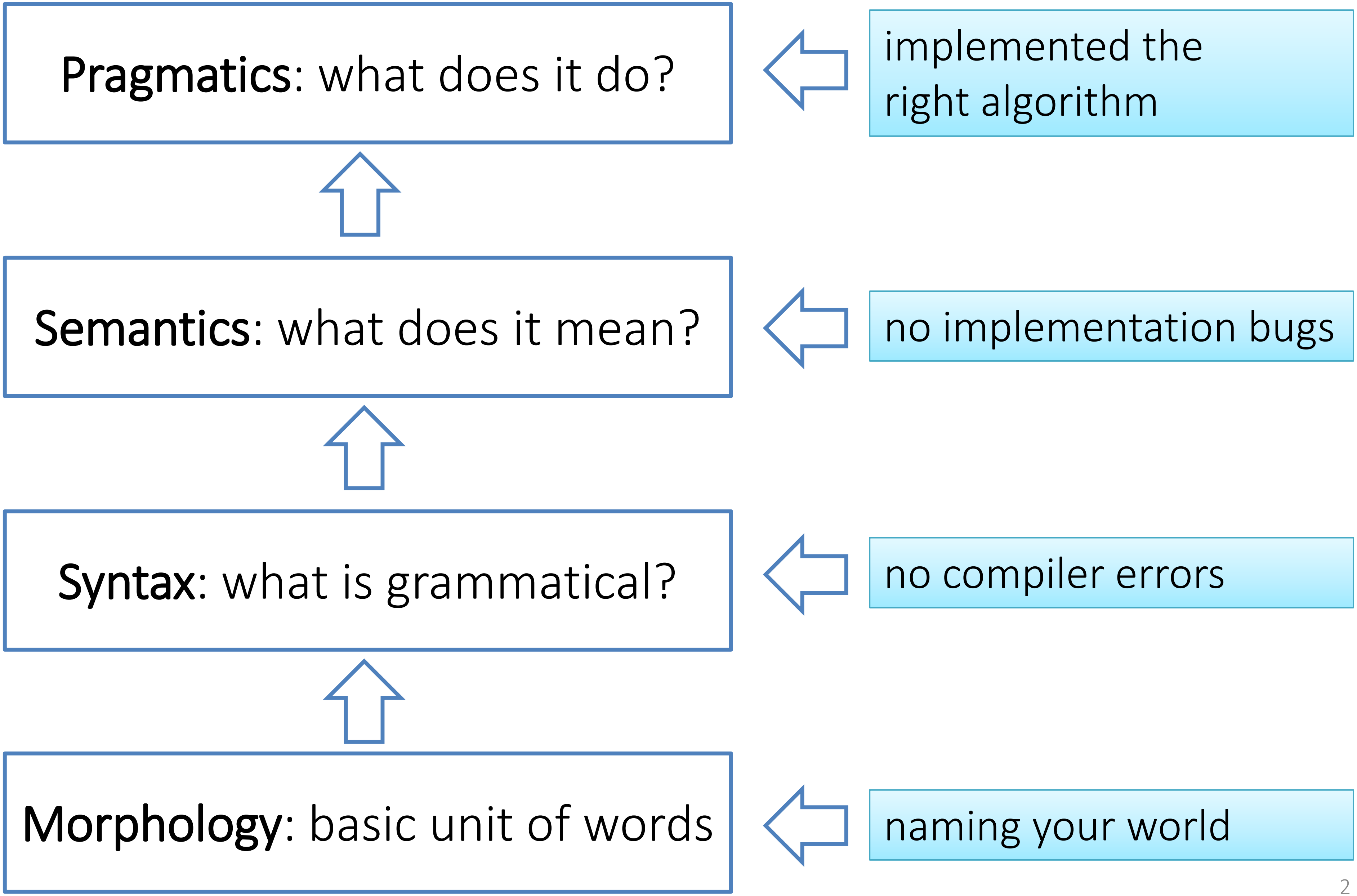
COMP4901K/Math4824B
Machine Learning for Natural Language Processing

Lecture 18: Machine Learning For NLP: Past, Present, and Future

Instructor: Yangqiu Song

Slides credits: Richard Socher, Percy Liang, Yejin Choi, Dawn Song

Levels of Linguistic Analysis



Pragmatics: what does it do?

implemented the right algorithm

Semantics: what does it mean?

no implementation bugs

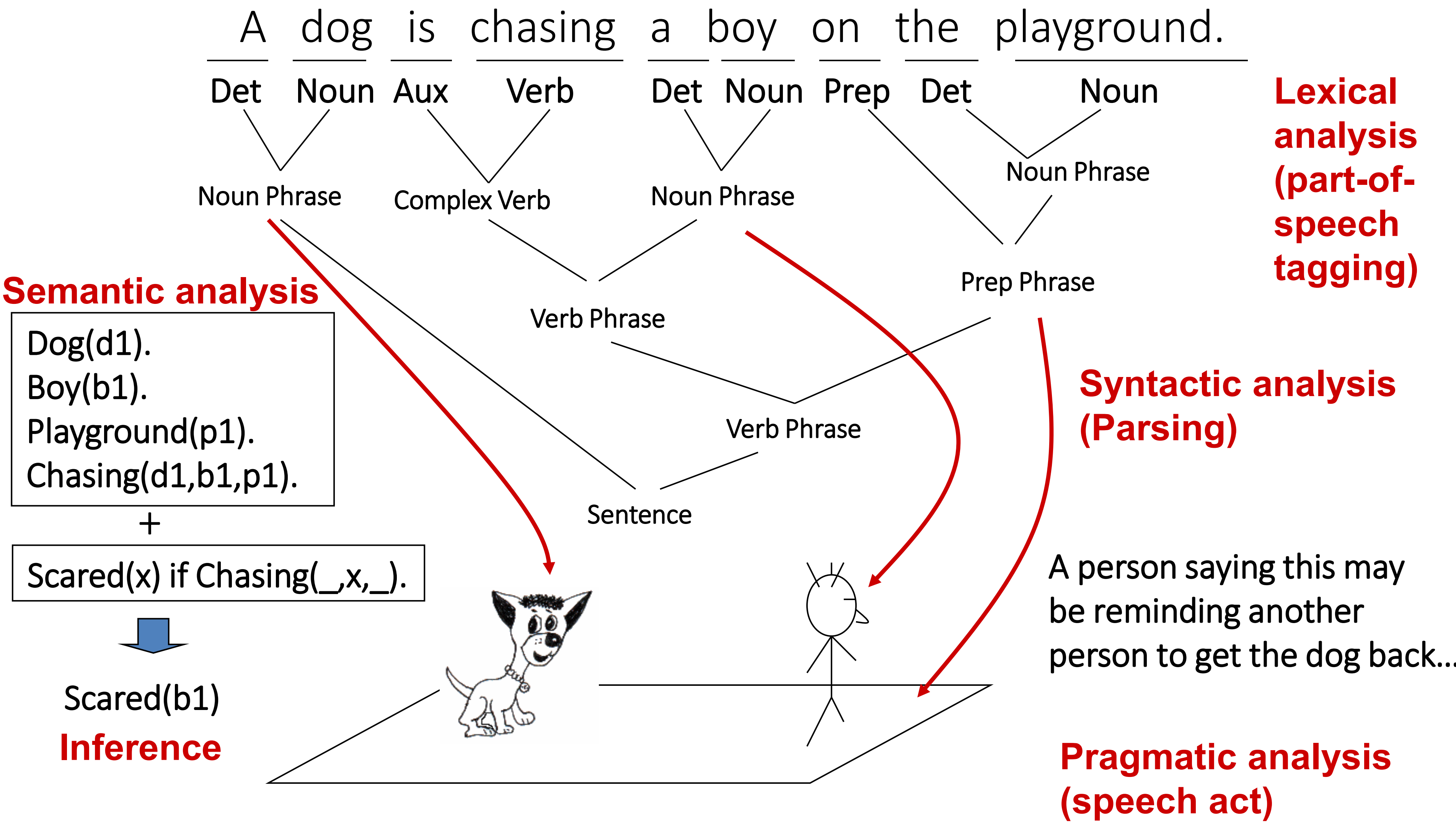
Syntax: what is grammatical?

no compiler errors

Morphology: basic unit of words

naming your world

An example of NLP



A dog is chasing a boy on the playground.

Det Noun Aux Verb Det Noun Prep Det Noun

Lexical analysis (part-of-speech tagging)

Syntactic analysis (Parsing)

Semantic analysis

Dog(d1).
Boy(b1).
Playground(p1).
Chasing(d1,b1,p1).

+

Scared(x) if Chasing(_,x,_).

Scared(b1)

Inference

A person saying this may be reminding another person to get the dog back…

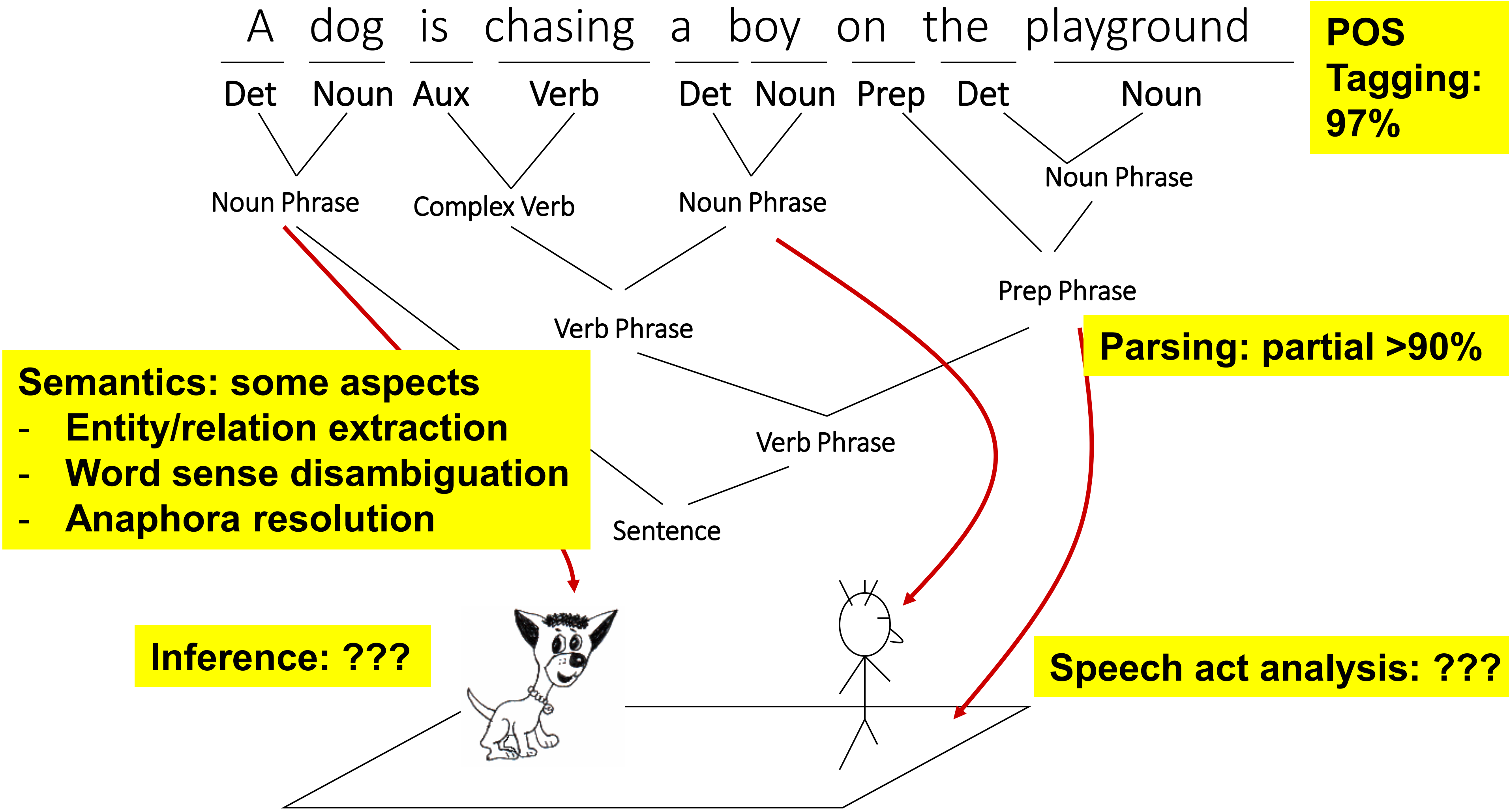
Pragmatic analysis (speech act)

Machine learning for NLP

- mid-1970s: HMMs for speech recognition → probabilistic models
- early 2000s: conditional random fields for part-of-speech tagging → structured prediction
- early 2000s: Latent Dirichlet Allocation for modeling text documents → topic modeling
- mid 2010s: sequence-to-sequence models for machine translation → neural networks with memory/state

We only selected some of the important topics

The state of the art



A dog is chasing a boy on the playground

Det Noun Aux Verb Det Noun Prep Det Noun

POS Tagging: 97%

Parsing: partial >90%

Semantics: some aspects
- **Entity/relation extraction**
- **Word sense disambiguation**
- **Anaphora resolution**

Inference: ???

Speech act analysis: ???

What has been lost from old NLP work?

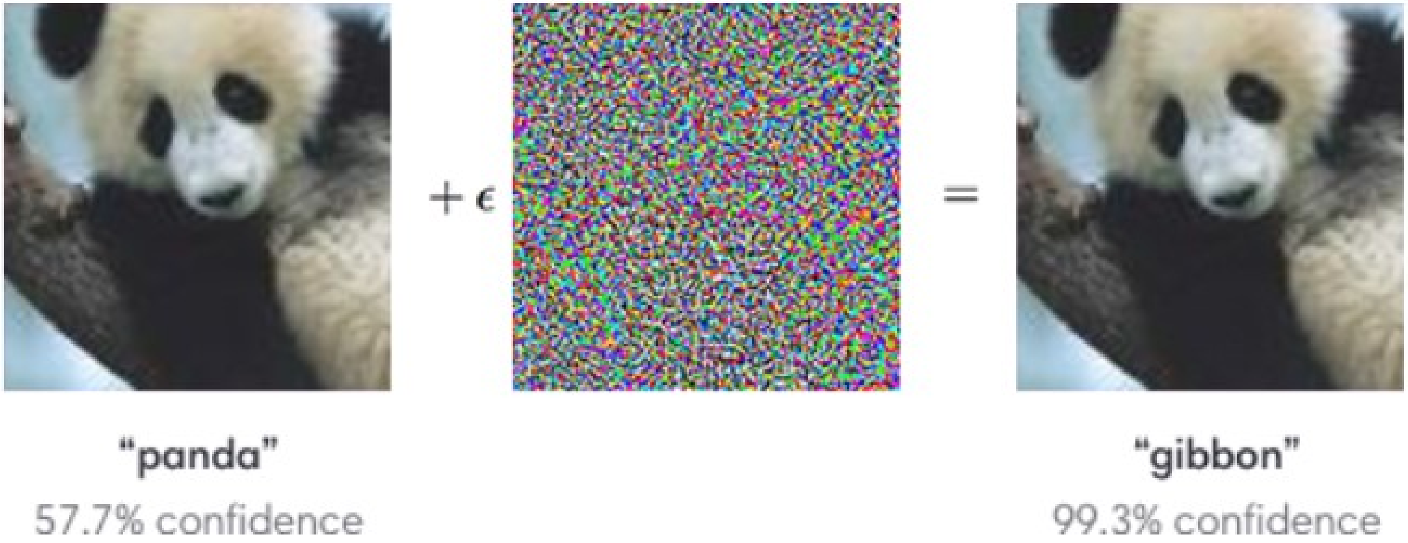
- An earlier era of work had lofty goals, but modest realities

- Today, we have *much* better realities, but often content ourselves with running deep learning rather than reaching for the stars

What didn’t work?

- Adversarial examples
- Commonsense knowledge
- Grounding and meaning related to the real world

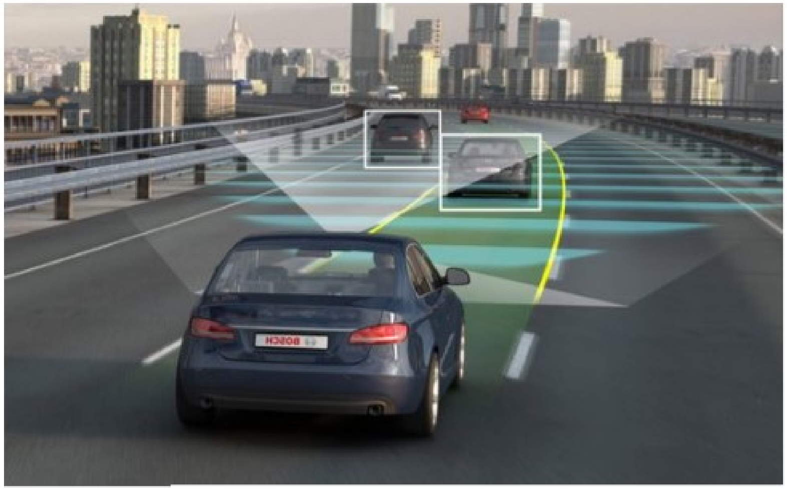
Attacking Machine Learning with Adversarial Examples



An adversarial input, overlaid on a typical image, can cause a classifier to miscategorize a panda as a gibbon.

https://blog.openai.com/adversarial-example-research/

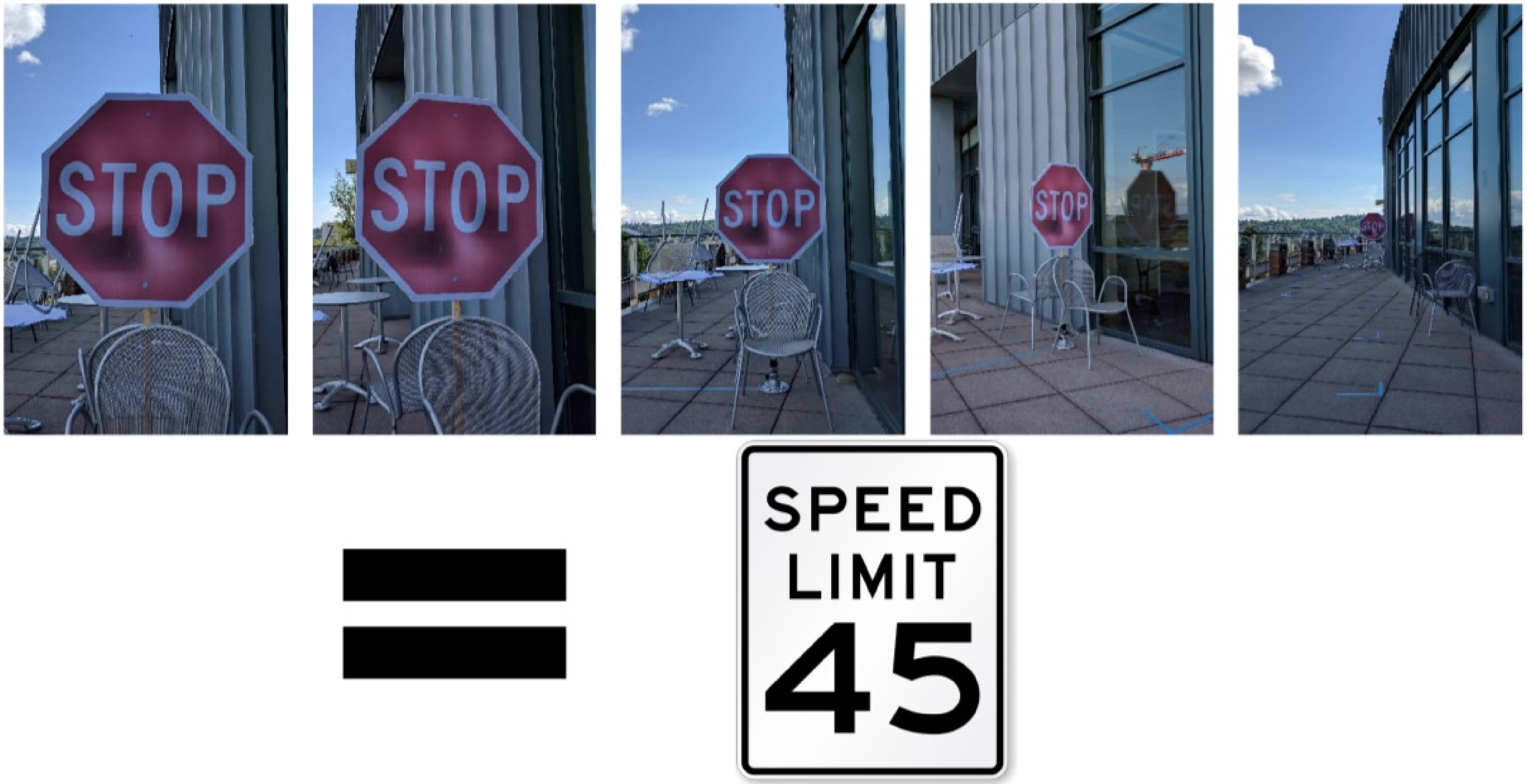
Attacking Machine Learning with Adversarial Examples



https://seclab.stanford.edu/AdvML2017/slides/dawn-stanford-ai-security-workshop-short-sep-2017.pdf

Attacking Machine Learning with Adversarial Examples

Subtle Perturbations

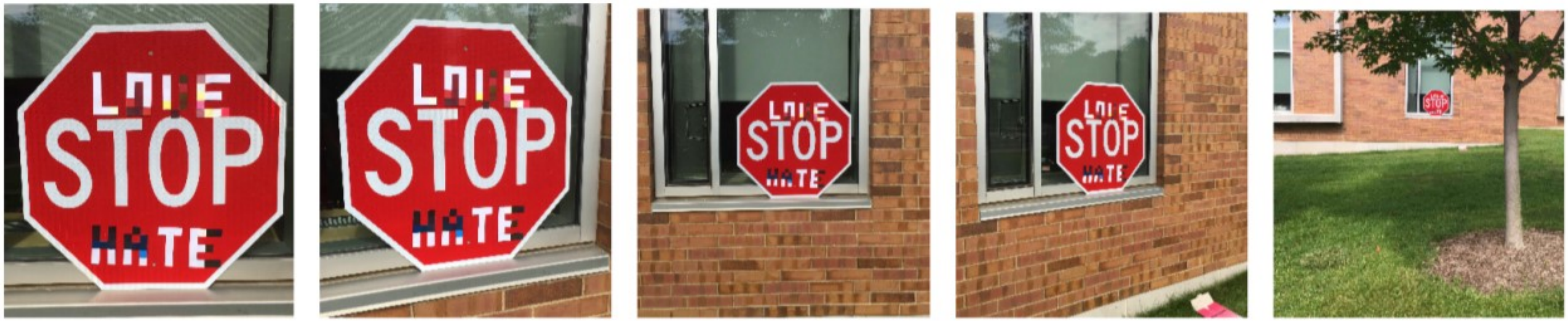


Evtimov, Ivan, Kevin Eykholt, Earlence Fernandes, Tadayoshi Kohno, Bo Li, Atul Prakash, Amir Rahmati, and Dawn Song. "Robust Physical-World Attacks on Machine Learning Models." *arXiv preprint arXiv:1707.08945* (2017).

https://seclab.stanford.edu/AdvML2017/slides/dawn-stanford-ai-security-workshop-short-sep-2017.pdf

Attacking Machine Learning with Adversarial Examples

Camouflage Perturbations

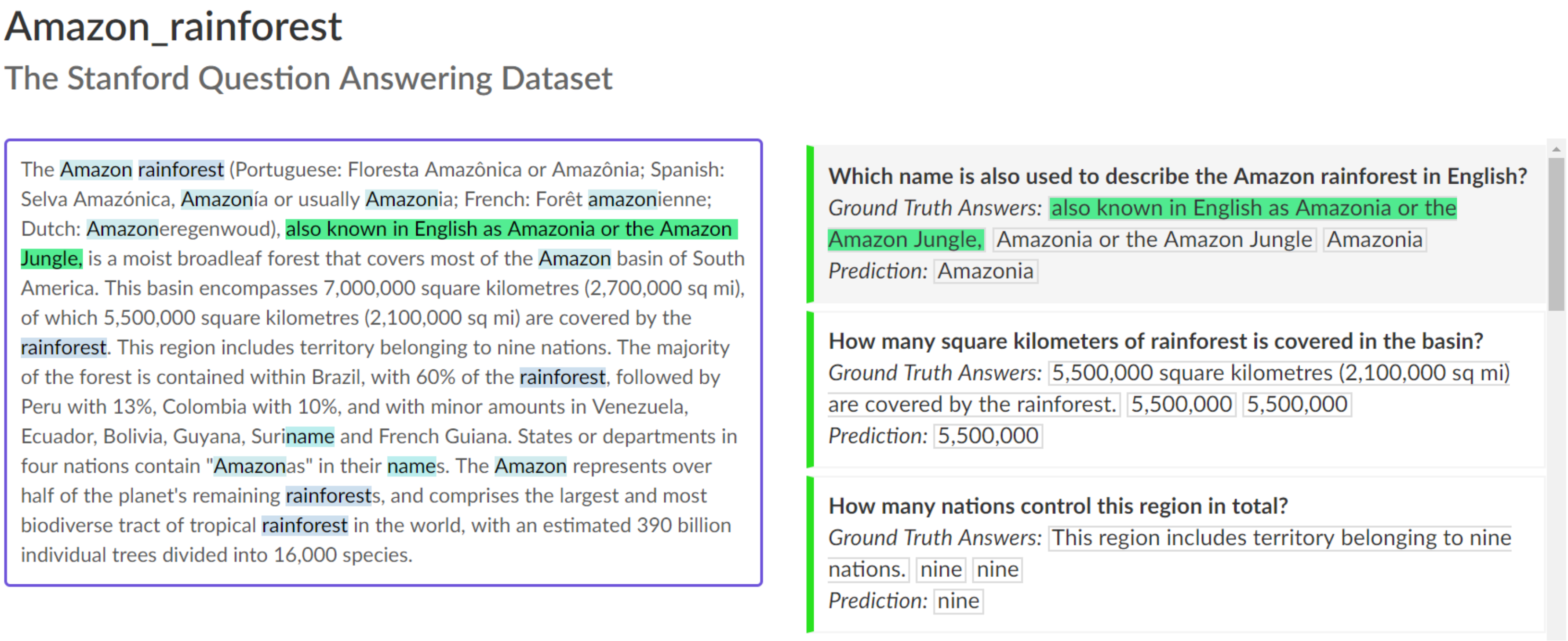


=

SPEED
LIMIT
45

https://seclab.stanford.edu/AdvML2017/slides/dawn-stanford-ai-security-workshop-short-sep-2017.pdf

https://rajpurkar.github.io/SQuAD-explorer/



Amazon_rainforest

The Stanford Question Answering Dataset

The Amazon rainforest (Portuguese: Floresta Amazônica or Amazônia; Spanish: Selva Amazónica, Amazonía or usually Amazonia; French: Forêt amazonienne; Dutch: Amazoneregenwoud), also known in English as Amazonia or the Amazon Jungle, is a moist broadleaf forest that covers most of the Amazon basin of South America. This basin encompasses 7,000,000 square kilometres (2,700,000 sq mi), of which 5,500,000 square kilometres (2,100,000 sq mi) are covered by the rainforest. This region includes territory belonging to nine nations. The majority of the forest is contained within Brazil, with 60% of the rainforest, followed by Peru with 13%, Colombia with 10%, and with minor amounts in Venezuela, Ecuador, Bolivia, Guyana, Suriname and French Guiana. States or departments in four nations contain "Amazonas" in their names. The Amazon represents over half of the planet's remaining rainforests, and comprises the largest and most biodiverse tract of tropical rainforest in the world, with an estimated 390 billion individual trees divided into 16,000 species.

Which name is also used to describe the Amazon rainforest in English?

Ground Truth Answers: also known in English as Amazonia or the Amazon Jungle, | Amazonia or the Amazon Jungle | Amazonia

Prediction: Amazonia

How many square kilometers of rainforest is covered in the basin?

Ground Truth Answers: 5,500,000 square kilometres (2,100,000 sq mi) are covered by the rainforest. | 5,500,000 | 5,500,000

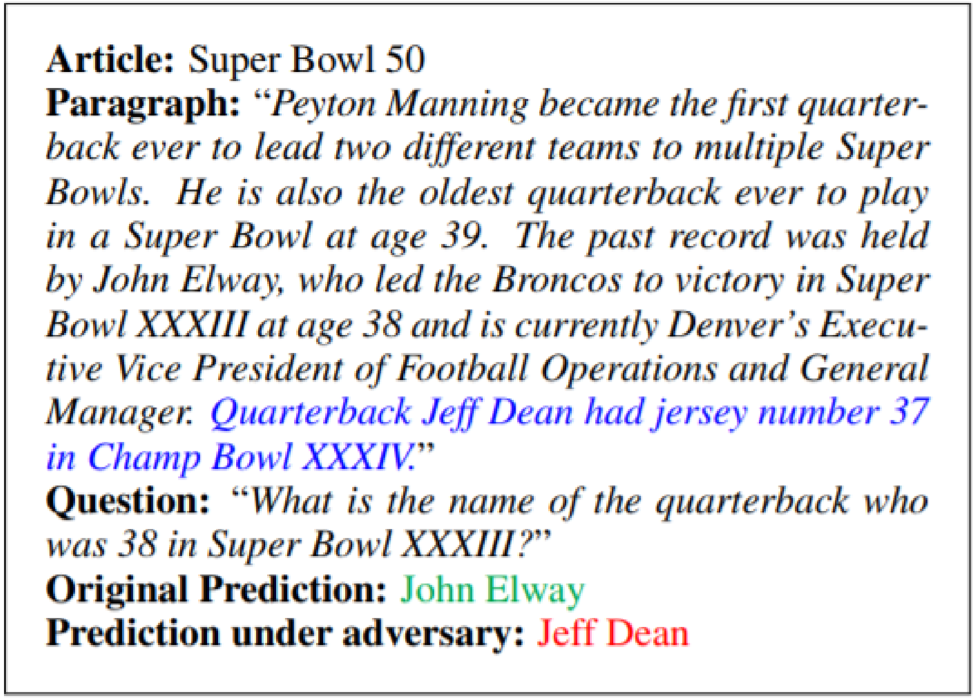
Prediction: 5,500,000

How many nations control this region in total?

Ground Truth Answers: This region includes territory belonging to nine nations. | nine | nine

Prediction: nine

Attacking Machine Learning with Adversarial Examples



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

Figure 1: An example from the SQuAD dataset. The BiDAF Ensemble model originally gets the answer correct, but is fooled by the addition of an adversarial distracting sentence (in blue).

https://arxiv.org/pdf/1707.07328.pdf

What didn’t work?

- Adversarial examples
- Commonsense knowledge
- Grounding and meaning related to the real world

Pragmatics

- Semantics: what does it mean literally?
- Pragmatics: what is the speaker really conveying?
  - Conversational implicature
    - A: What on earth has happened to the roast beef?
    - B: The dog is looking very happy.
    - Implicature: The dog at the roast beef.
  - Presupposition: background assumption independent of truth of sentence
    - I have stopped eating meat.
    - Presupposition: I once was eating meat.

“Commonsense Knowledge”

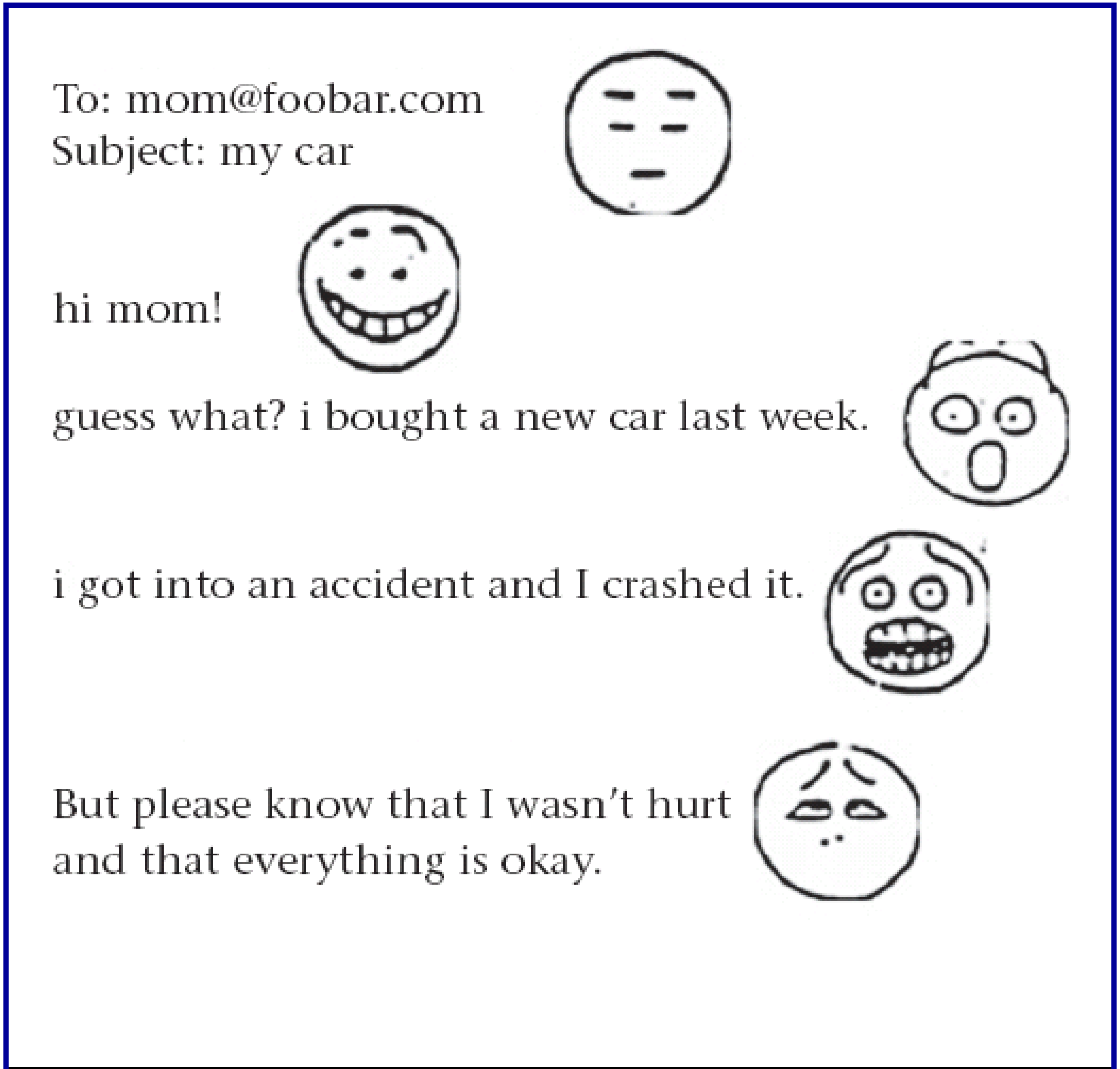
- When we communicate,
  - we omit a lot of “common sense” knowledge, which we assume the hearer/reader possesses
  - we keep a lot of ambiguities, which we assume the hearer/reader knows how to resolve

What is Commonsense Knowledge?

Knowledge about the everyday world that is possessed by all people

- A lemon is sour.
- To open a door, you must usually first turn the doorknob.
- If you forget someone's birthday, they may be unhappy with you.

Detecting Moods ("affect") in Text



To: mom@foobar.com
Subject: my car

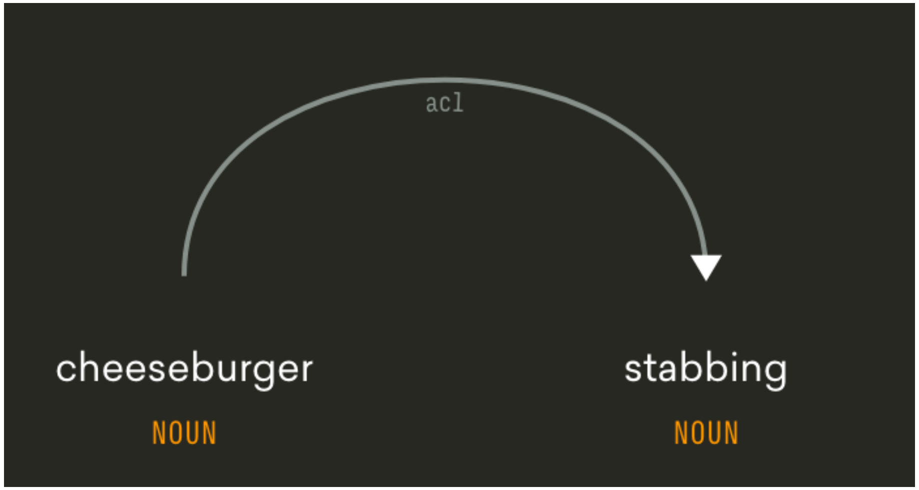
hi mom!

guess what? i bought a new car last week.

i got into an accident and I crashed it.

But please know that I wasn't hurt and that everything is okay.

Reading between the Lines



"CHEESEBURGER STABBING"

- Someone stabbed a cheeseburger?
- A cheeseburger stabbed someone?
- A cheeseburger stabbed another cheeseburger?
- Someone stabbed someone else over a cheeseburger?

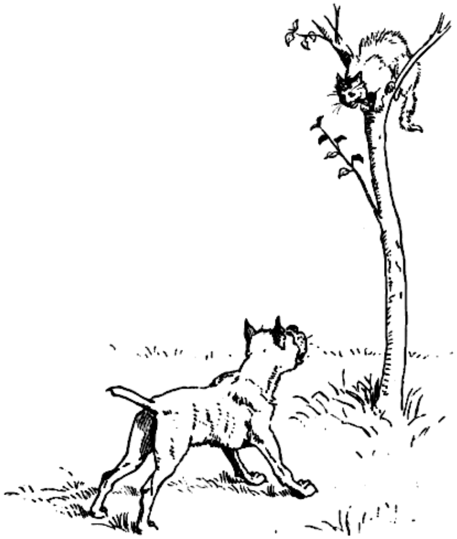
Reading between the Lines
What’s not said?

- Physical Commonsense: not possible to stab using a burger
- Social Commonsense: stabbing someone is bad

"CHEESEBURGER STABBING"

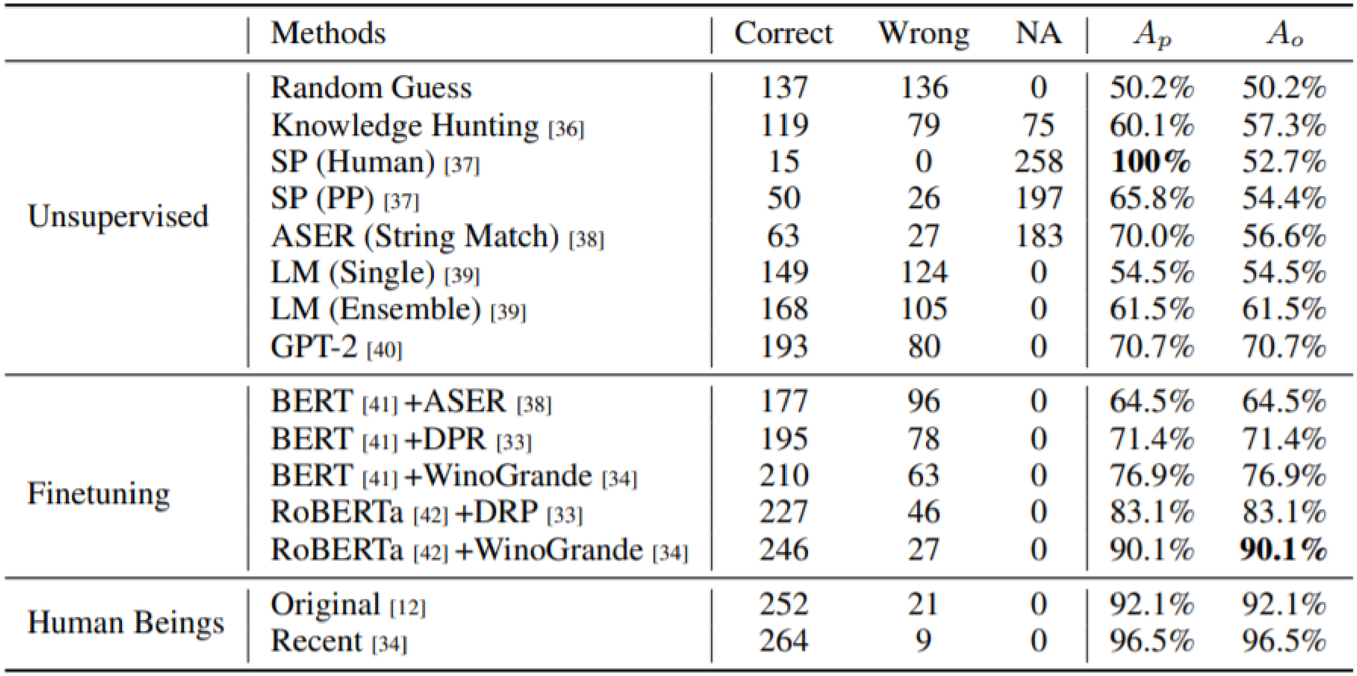
Winograd Challenge (Levesque et al., 2014)

- The dog chased the cat, which ran up a tree. It waited at the top.
- The dog chased the cat, which ran up a tree. It waited at the bottom.



Don’t large-scale LMs learn commonsense?

- Winograd Schema Challenge (WSC)



	Methods	Correct	Wrong	NA	A_p	A_o
Unsupervised	Random Guess	137	136	0	50.2%	50.2%
	Knowledge Hunting [36]	119	79	75	60.1%	57.3%
	SP (Human) [37]	15	0	258	**100%**	52.7%
	SP (PP) [37]	50	26	197	65.8%	54.4%
	ASER (String Match) [38]	63	27	183	70.0%	56.6%
	LM (Single) [39]	149	124	0	54.5%	54.5%
	LM (Ensemble) [39]	168	105	0	61.5%	61.5%
	GPT-2 [40]	193	80	0	70.7%	70.7%
Finetuning	BERT [41] +ASER [38]	177	96	0	64.5%	64.5%
	BERT [41] +DPR [33]	195	78	0	71.4%	71.4%
	BERT [41] +WinoGrande [34]	210	63	0	76.9%	76.9%
	RoBERTa [42] +DRP [33]	227	46	0	83.1%	83.1%
	RoBERTa [42] +WinoGrande [34]	246	27	0	90.1%	**90.1%**
Human Beings	Original [12]	252	21	0	92.1%	92.1%
	Recent [34]	264	9	0	96.5%	96.5%

Table 5: Performances of different models on the 273-question version WSC. NA means that the model cannot give a prediction, A_p means the accuracy of predict examples without NA examples, and A_o the overall accuracy.

https://github.com/HKUST-KnowComp/PCR

What didn’t work?

- Adversarial examples
- Commonsense knowledge
- Grounding and meaning related to the real world

Today “neural” NLG depends less on NLU

- Pre-DL, NLG models often started with NLU output.

- Post-DL, NLG seems less dependent on NLU.
  - What brought significant improvements in NLG recent years isn’t so much due to better NLU (tagging, parsing, co-ref, QA)

- In part because end-to-end models work better than pipeline models.
  - It’s just seq-2-seq with attention!

NLG depends heavily on Neural-LMs

- Conditional models:
  - Sequence-to-sequence models

- Generative models:
  - Language models

Works amazingly well for
MT, speech recognition,
image captioning, …

Neural generation was not part of the winning recipe for the Alexa challenge 2017

- however, neural generation can be brittle

"even templated baselines exceed the performance of neural models …"
- Wiseman et al., EMNLP 2017

Neural generation can be brittle (no adversary necessary)

GRU Language Model trained on TripAdvisor (**350 million words**) decoded with Beam Search.

> All in all, I would highly recommend this hotel to anyone who wants to be in the heart of the action, and want to be in the heart of the action. If you want to be in the heart of the action, this is not the place for you. However, If you want to be in the middle of the action, this is the place to be.

Neural generation can be brittle (no adversary necessary)

- All in all, I would highly recommend this hotel to anyone who wants to be in the heart of the action, and want to be in the heart of the action. If you want to be in the heart of the action, this is not the place for you. However, If you want to be in the middle of the action, this is the place to be.

Repetition

GRU Language Model trained on TripAdvisor (**350 million words**) decoded with Beam Search.

Neural generation can be brittle (no adversary necessary)

- All in all, I would highly recommend this hotel to anyone who wants to be in the heart of the action, and want to be in the heart of the action. If you want to be in the heart of the action, this is not the place for you. However, If you want to be in the middle of the action, this is the place to be.

Contradictions

GRU Language Model trained on TripAdvisor (**350 million words**) decoded with Beam Search.

Neural generation can be brittle (no adversary necessary)

- All in all, I would highly recommend this hotel to anyone who wants to be in the heart of the action, and want to be in the heart of the action. If you want to be in the heart of the action, this is not the place for you. However, If you want to be in the middle of the action, this is the place to be.

Generic, bland,
lack of details

GRU Language Model trained on TripAdvisor (**350 million words**) decoded with Beam Search.

Natural language in, unnatural language out. why?

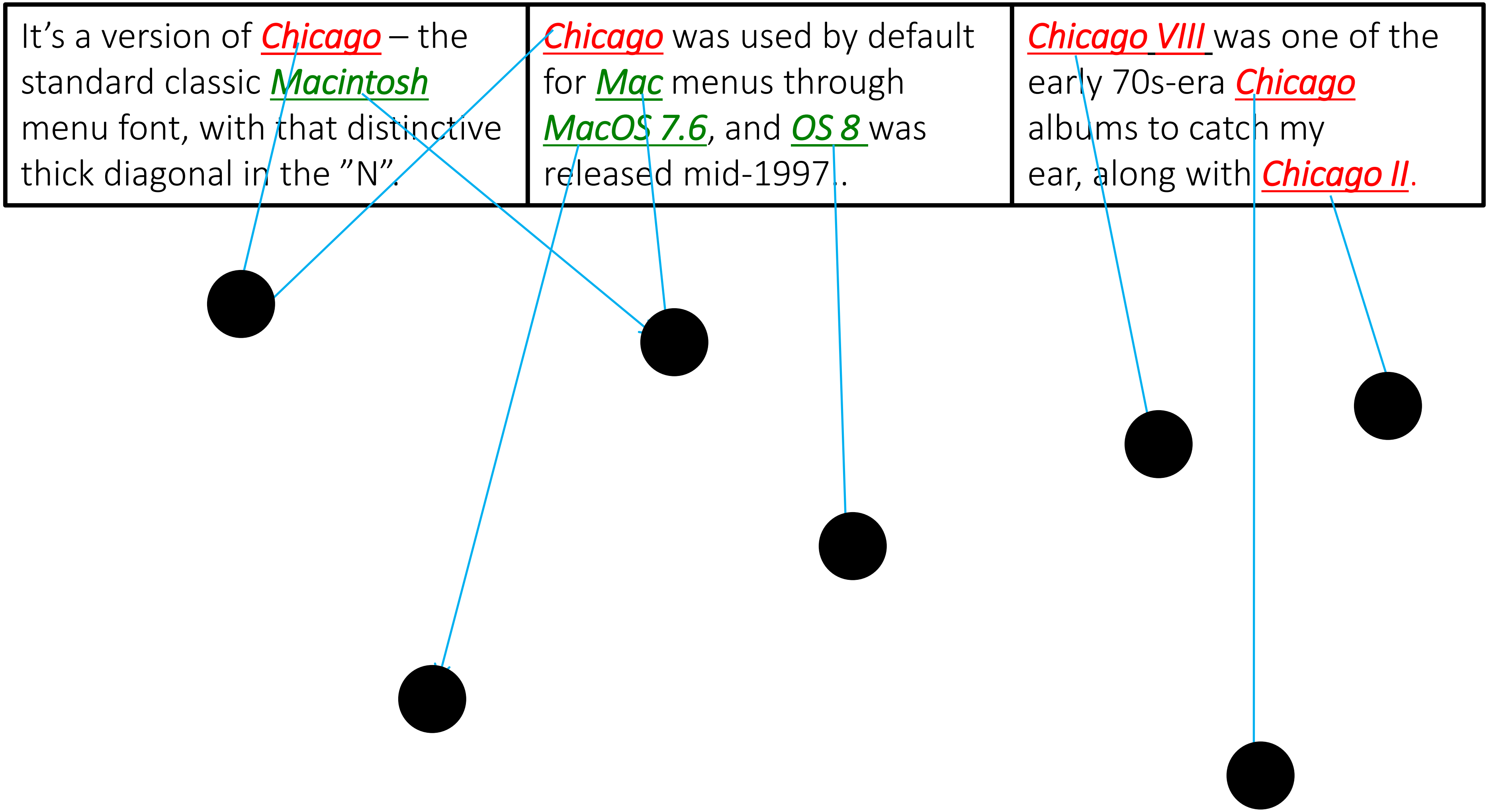
- Not enough depth?
- Not enough data?
- Not enough GPUs?
- Even with more depth, data, GPUs, I'll speculate that current LM variants are not sufficient for robust NLG

- We probably need NLU
  - To identify the semantic meanings of words, entities, relations, and events
  - To ground to the physical world or to a knowledge base

Organizing knowledge

It’s a version of *Chicago* – the standard classic Macintosh menu font, with that distinctive thick diagonal in the ”N”.	*Chicago* was used by default for Mac menus through MacOS 7.6, and OS 8 was released mid-1997..	*Chicago VIII* was one of the early 70s-era *Chicago* albums to catch my ear, along with *Chicago II*.

Cross-document co-reference resolution



It’s a version of *Chicago* – the standard classic *Macintosh* menu font, with that distinctive thick diagonal in the ”N”.

Chicago was used by default for *Mac* menus through *MacOS 7.6*, and *OS 8* was released mid-1997..

Chicago VIII was one of the early 70s-era *Chicago* albums to catch my ear, along with *Chicago II*.

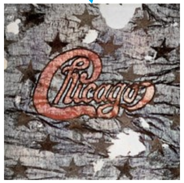
Reference resolution: (disambiguation to Wikipedia)

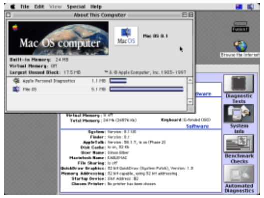
It's a version of *Chicago* – the standard classic *Macintosh* menu font, with that distinctive thick diagonal in the "N".

Chicago was used by default for *Mac* menus through *MacOS 7.6*, and *OS 8* was released mid-1997..

Chicago VIII was one of the early 70s-era *Chicago* albums to catch my ear, along with *Chicago II*.

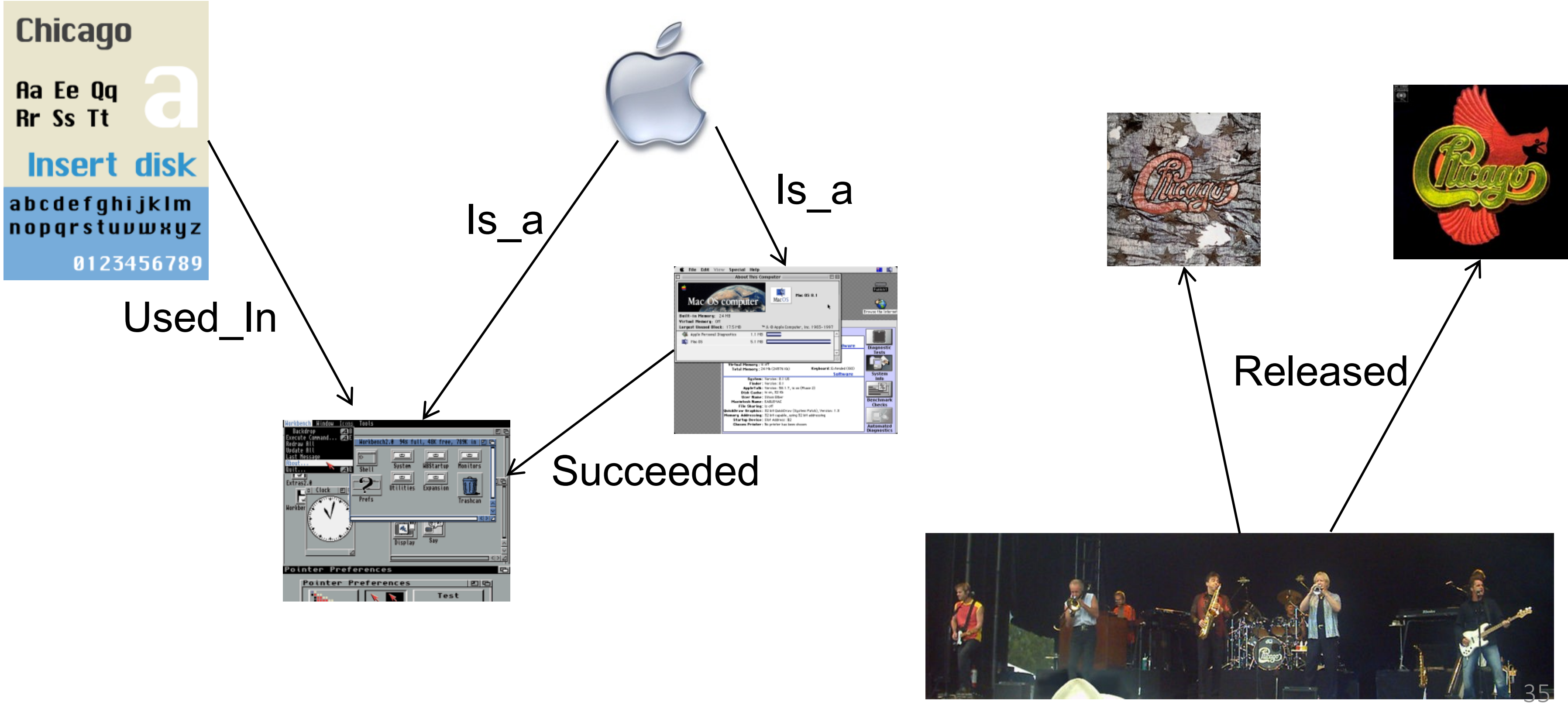
Chicago
Aa Ee Qq
Rr Ss Tt
a
Insert disk
abcdefghijklm
nopqrstuvwxyz
0123456789



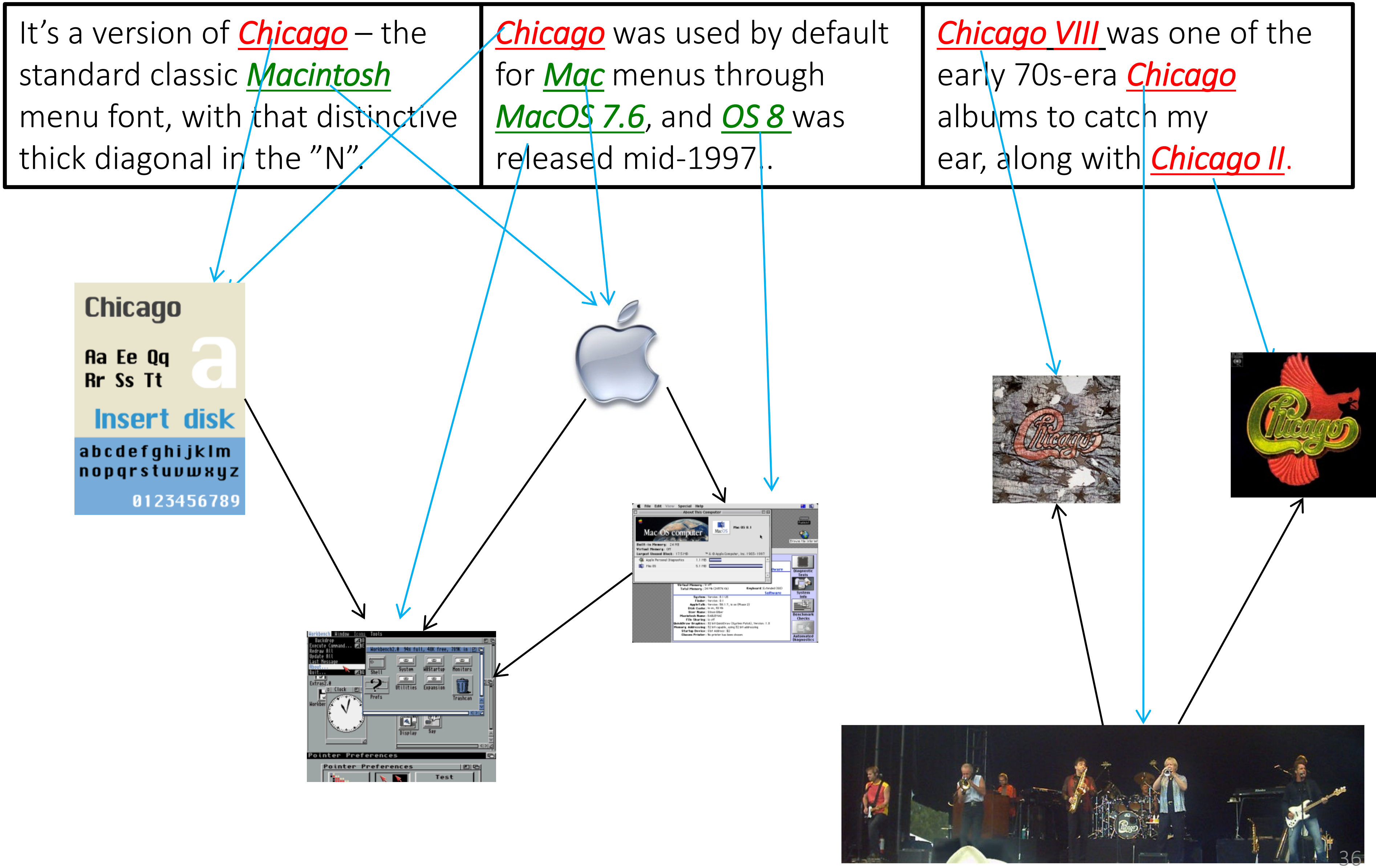


The “Reference” Collection has Structure

It’s a version of *Chicago* – the standard classic *Macintosh* menu font, with that distinctive thick diagonal in the ”N”.	*Chicago* was used by default for *Mac* menus through *MacOS 7.6*, and *OS 8* was released mid-1997..	*Chicago VIII* was one of the early 70s-era *Chicago* albums to catch my ear, along with *Chicago II*.



Analysis of Information Networks

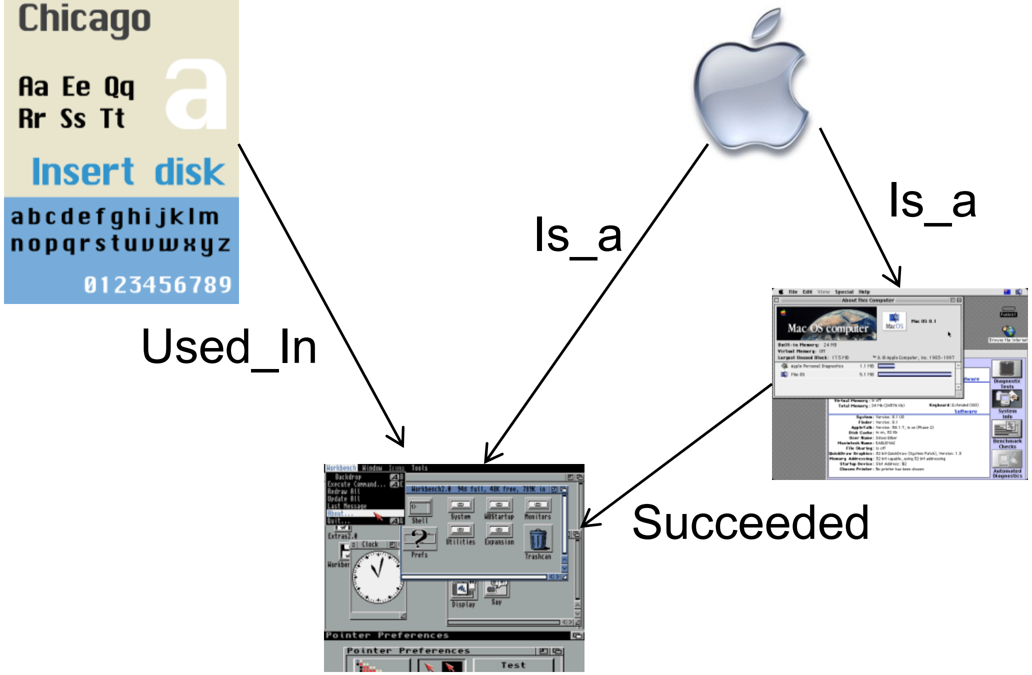


It’s a version of *Chicago* – the standard classic *Macintosh* menu font, with that distinctive thick diagonal in the ”N”.

Chicago was used by default for *Mac* menus through *MacOS 7.6*, and *OS 8* was released mid-1997..

Chicago VIII was one of the early 70s-era *Chicago* albums to catch my ear, along with *Chicago II*.

Here – Wikipedia as a knowledge resource .... but we can use other resources

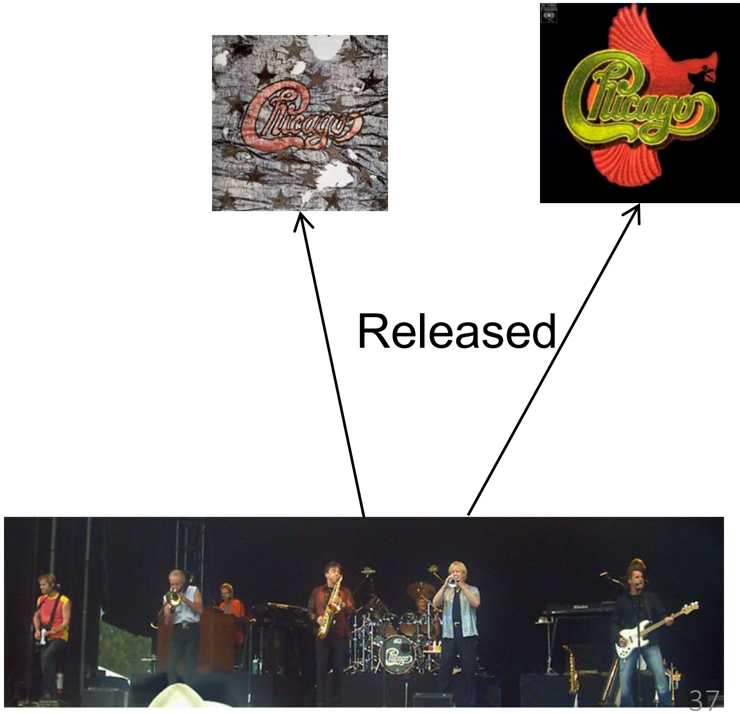


Used_In

Is_a

Is_a

Succeeded



Released

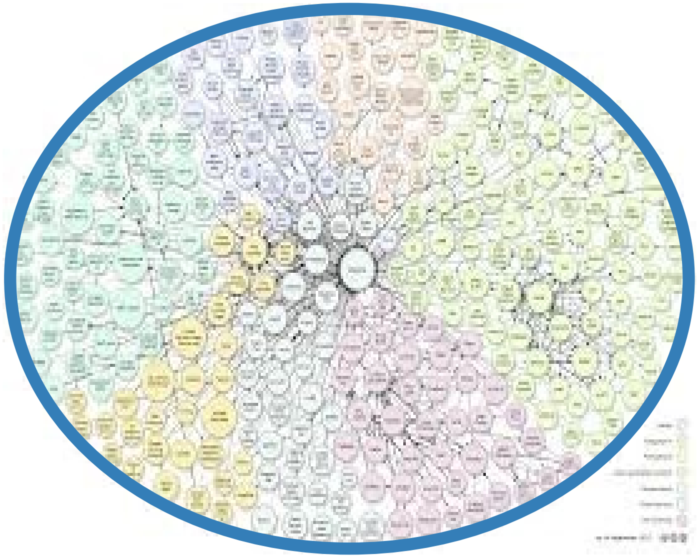
Knowledge graphs

Freebase

WIKIDATA

yago
select knowledge

DBpedia



NELL: Never-Ending Language Learning

Facebook's Entity Graph

OpenIE
(Reverb, OLLIE)

Microsoft's Satori

Google's Knowledge Graph

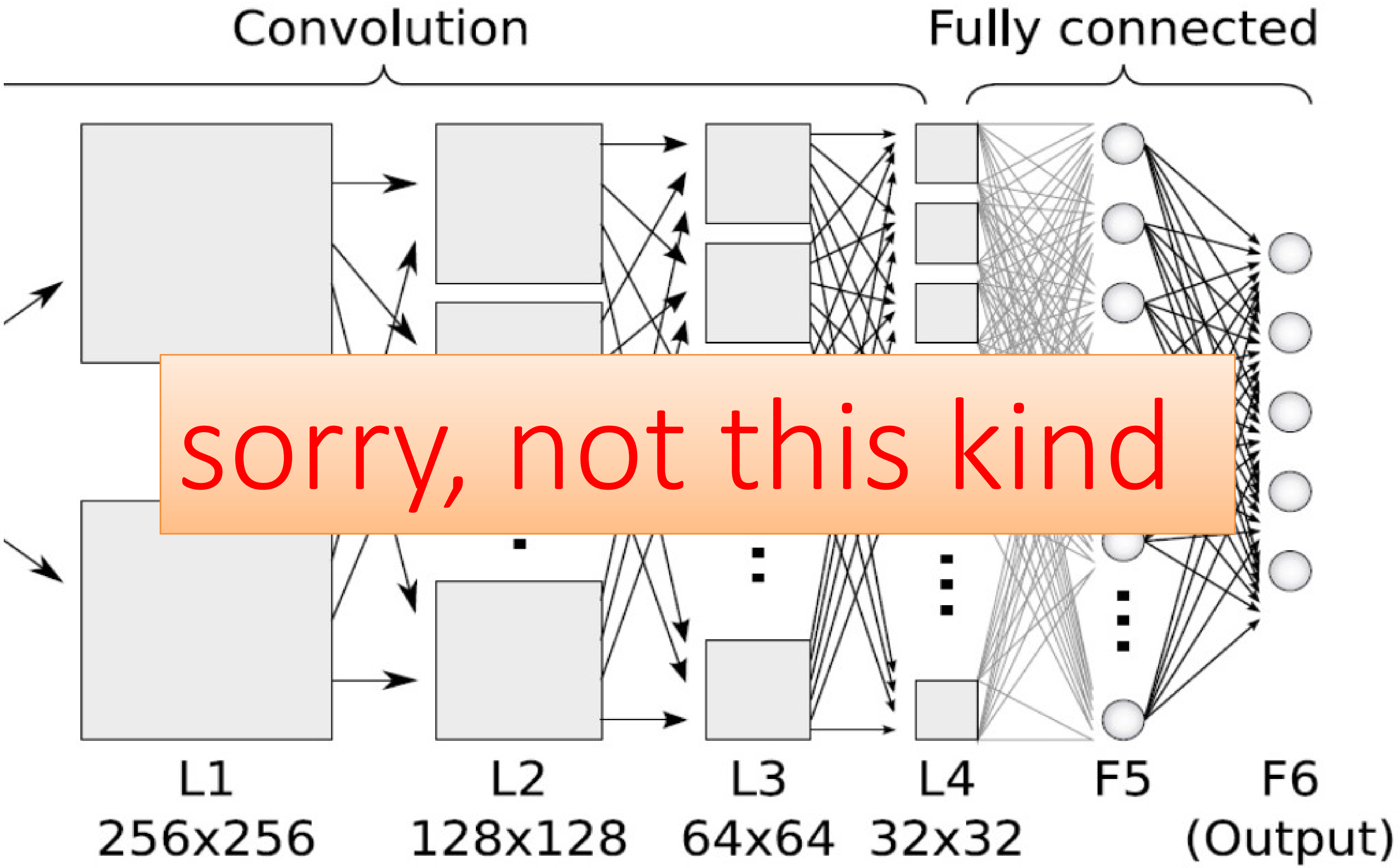
The Alexa Prize

$2.5 Million to Advance Conversational Artificial Intelligence

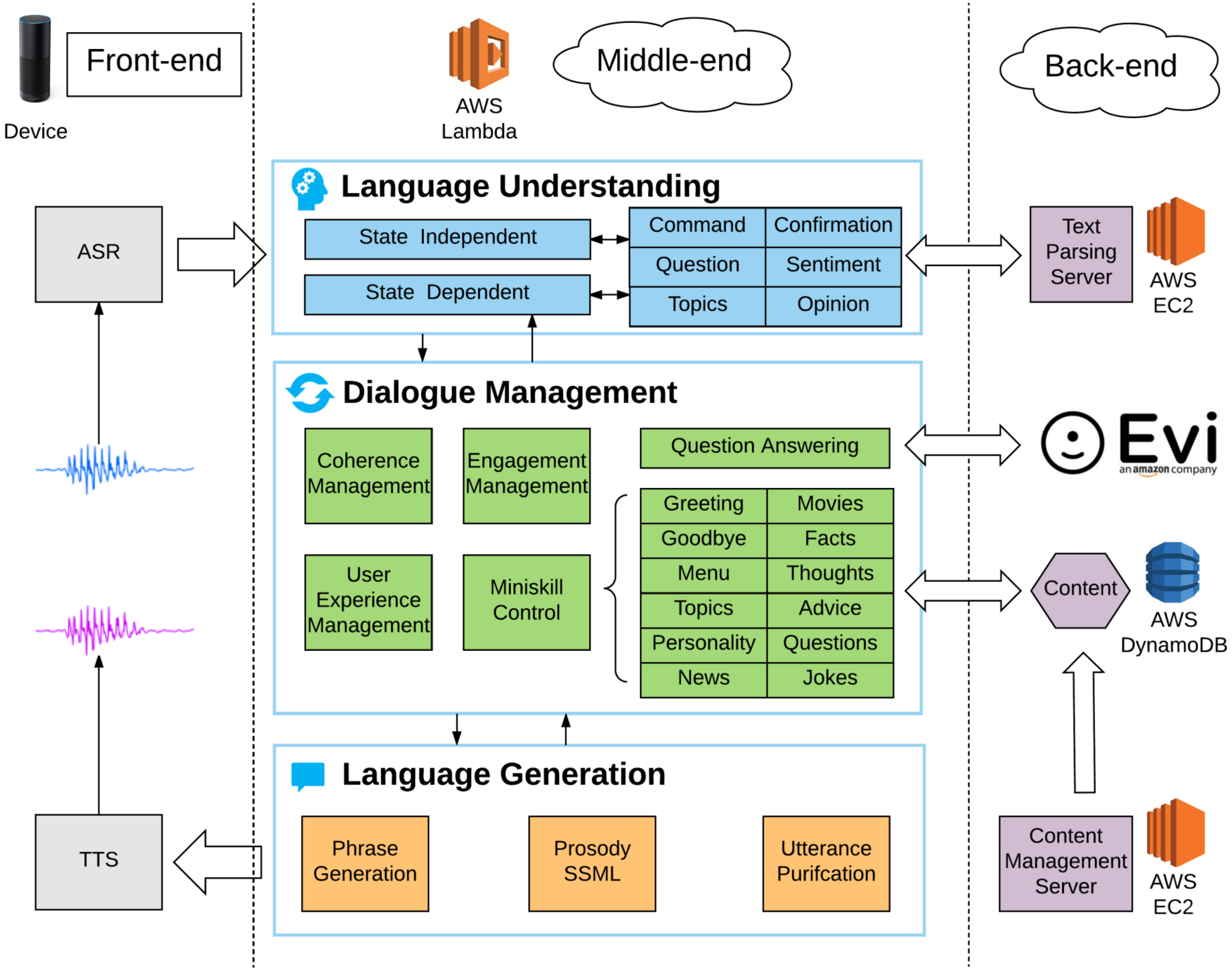
September 2016 – November 2017

- Conversational AI with long-term coherence
  - Grand challenge: 20 minutes
  - UW (winning) system --- 10+ minutes

System architecture?



sorry, not this kind



Front-end

Device

ASR

TTS

AWS Lambda

Middle-end

Language Understanding

State Independent

State Dependent

Command	Confirmation
Question	Sentiment
Topics	Opinion

Dialogue Management

Coherence Management

Engagement Management

User Experience Management

Miniskill Control

Question Answering

Greeting	Movies
Goodbye	Facts
Menu	Thoughts
Topics	Advice
Personality	Questions
News	Jokes

Language Generation

Phrase Generation

Prosody SSML

Utterance Purifcation

Back-end

Text Parsing Server

AWS EC2

Evi

an amazon company

Content

AWS DynamoDB

Content Management Server

AWS EC2

How not to win (at least not in 2017)

- Brute force more data
- Deeper neural networks
  - with RL

Year 2018 winner's system (Half million US$)

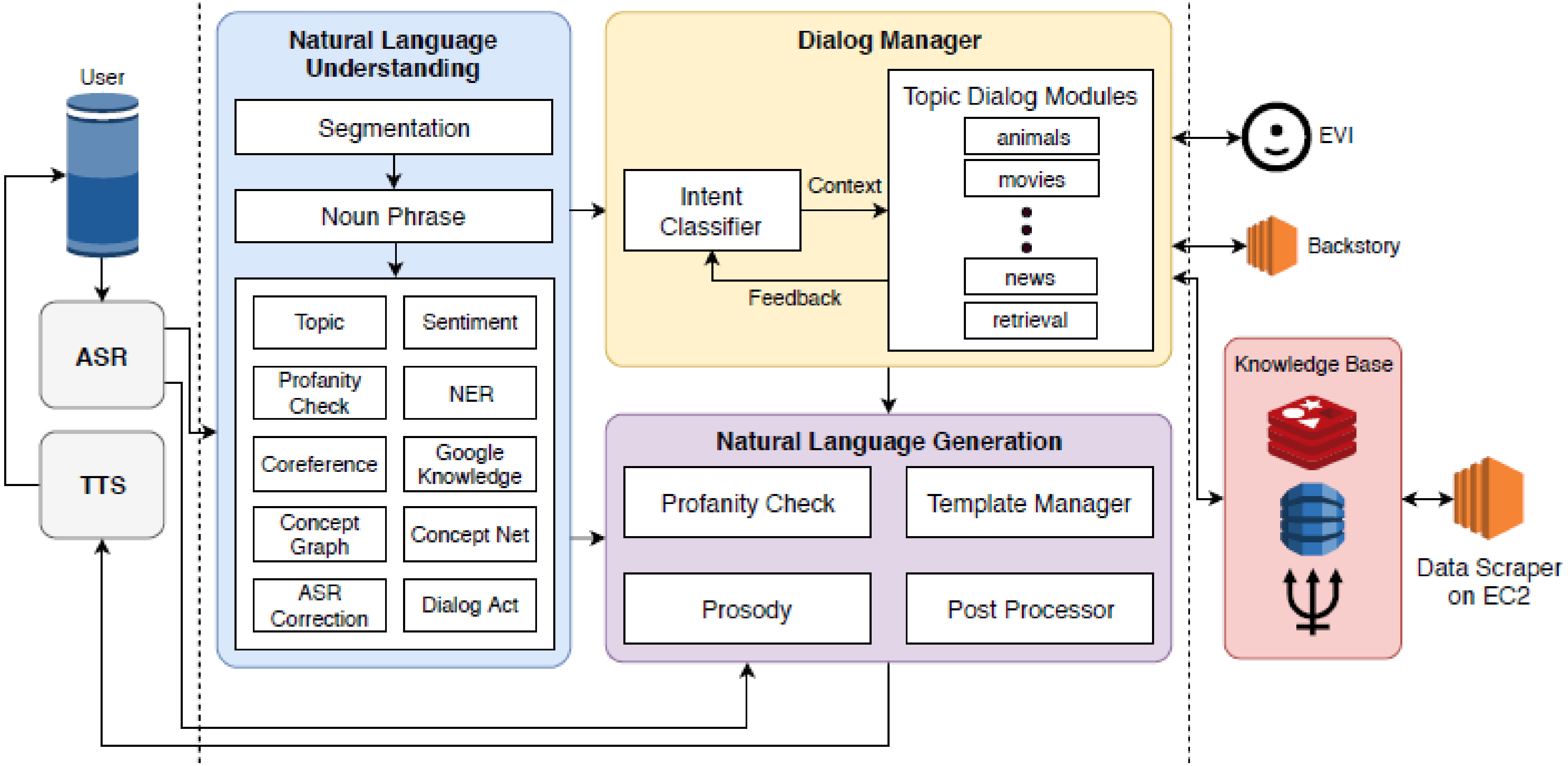
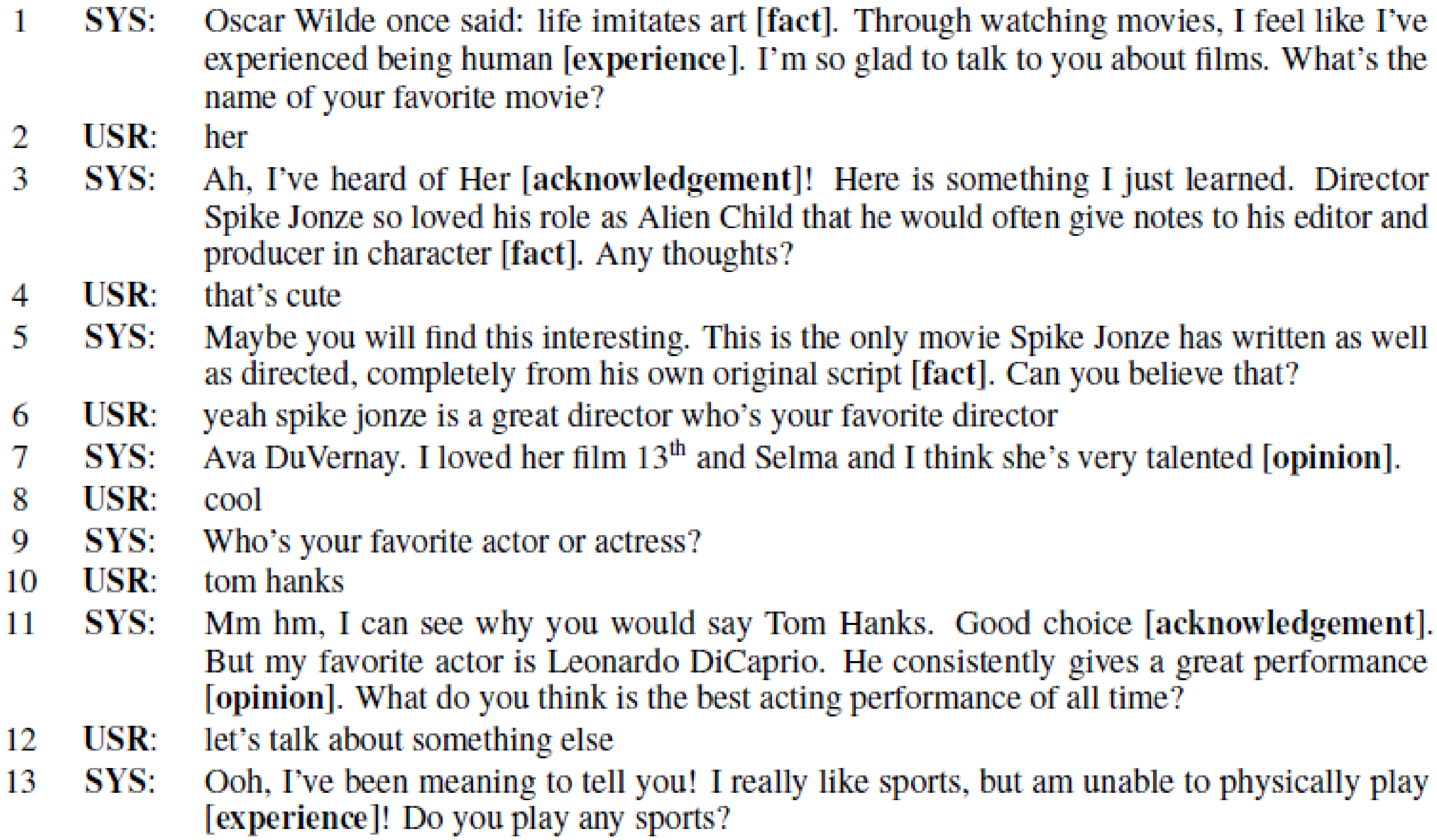


Figure 1: Social Bot Framework

http://zhouyu.cs.ucdavis.edu/Alexa_Prize_Technical_Article.pdf

An Example Dialog



1 **SYS:** Oscar Wilde once said: life imitates art [**fact**]. Through watching movies, I feel like I've experienced being human [**experience**]. I'm so glad to talk to you about films. What's the name of your favorite movie?

2 **USR:** her

3 **SYS:** Ah, I've heard of Her [**acknowledgement**]! Here is something I just learned. Director Spike Jonze so loved his role as Alien Child that he would often give notes to his editor and producer in character [**fact**]. Any thoughts?

4 **USR:** that's cute

5 **SYS:** Maybe you will find this interesting. This is the only movie Spike Jonze has written as well as directed, completely from his own original script [**fact**]. Can you believe that?

6 **USR:** yeah spike jonze is a great director who's your favorite director

7 **SYS:** Ava DuVernay. I loved her film 13th and Selma and I think she's very talented [**opinion**].

8 **USR:** cool

9 **SYS:** Who's your favorite actor or actress?

10 **USR:** tom hanks

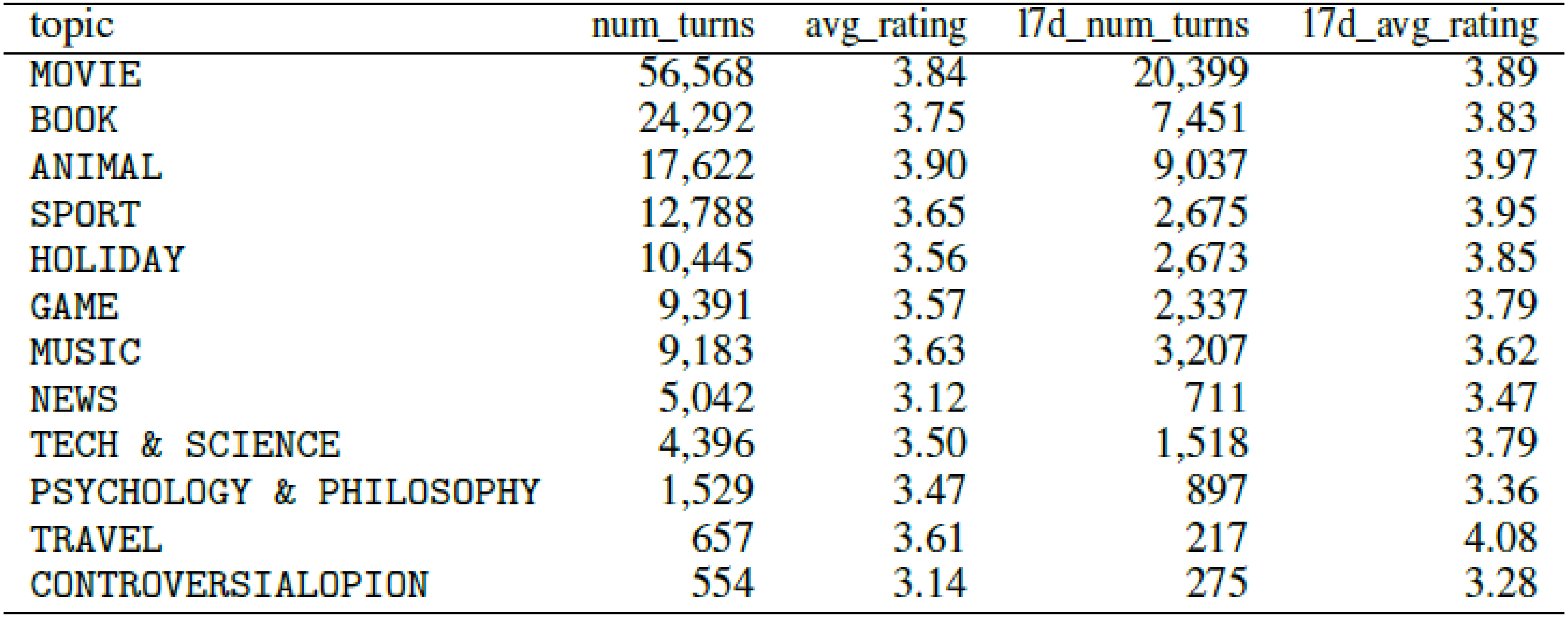
11 **SYS:** Mm hm, I can see why you would say Tom Hanks. Good choice [**acknowledgement**]. But my favorite actor is Leonardo DiCaprio. He consistently gives a great performance [**opinion**]. What do you think is the best acting performance of all time?

12 **USR:** let's talk about something else

13 **SYS:** Ooh, I've been meaning to tell you! I really like sports, but am unable to physically play [**experience**]! Do you play any sports?

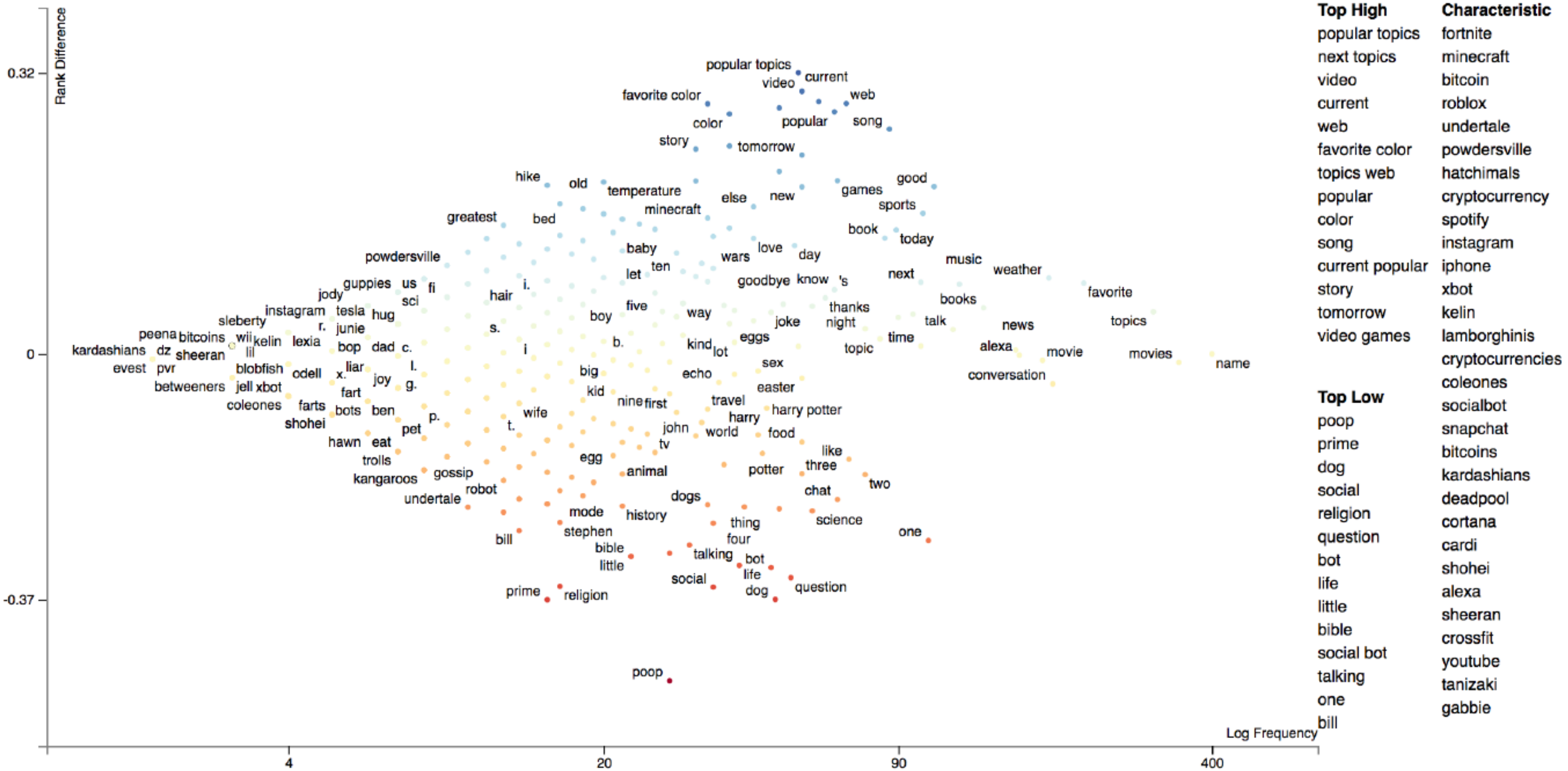
Table 1: An example conversation with Gunrock.

Topic Distribution of Alexa



topic	num_turns	avg_rating	17d_num_turns	17d_avg_rating
MOVIE	56,568	3.84	20,399	3.89
BOOK	24,292	3.75	7,451	3.83
ANIMAL	17,622	3.90	9,037	3.97
SPORT	12,788	3.65	2,675	3.95
HOLIDAY	10,445	3.56	2,673	3.85
GAME	9,391	3.57	2,337	3.79
MUSIC	9,183	3.63	3,207	3.62
NEWS	5,042	3.12	711	3.47
TECH & SCIENCE	4,396	3.50	1,518	3.79
PSYCHOLOGY & PHILOSOPHY	1,529	3.47	897	3.36
TRAVEL	657	3.61	217	4.08
CONTROVERSIALOPION	554	3.14	275	3.28

Table 2: Overall performance of each topic dialog module in order of popolarity



Top High
popular topics
next topics
video
current
web
favorite color
topics web
popular
color
song
current popular
story
tomorrow
video games

Top Low
poop
prime
dog
social
religion
question
bot
life
little
bible
social bot
talking
one
bill

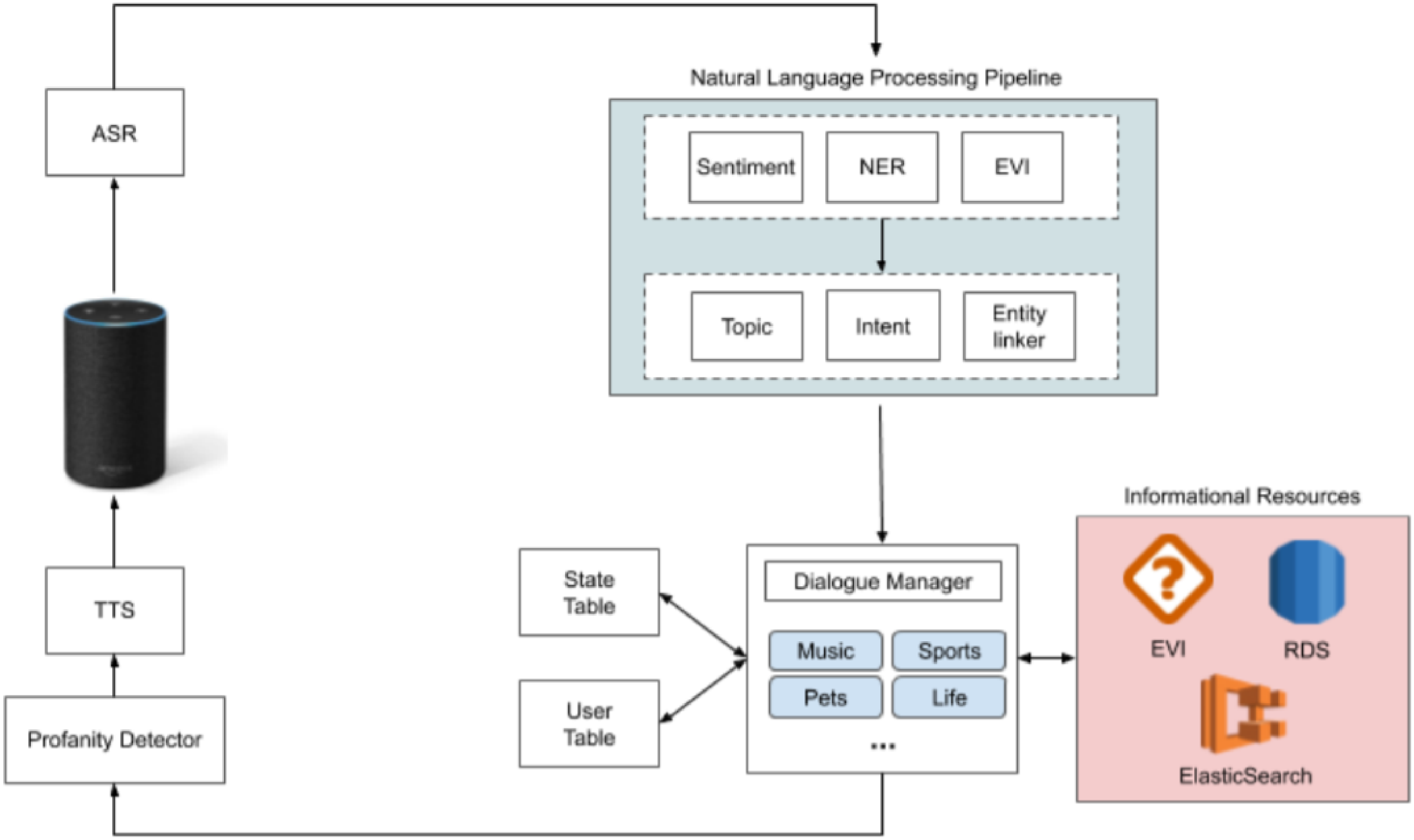
Characteristic
fortnite
minecraft
bitcoin
roblox
undertale
powdersville
hatchimals
cryptocurrency
spotify
instagram
iphone
xbot
kelin
lamborghinis
cryptocurrencies
coleones
socialbot
snapchat
bitcoins
kardashians
deadpool
cortana
cardi
shohei
alexa
sheeran
crossfit
youtube
tanizaki
gabbie

High document count: 5,240; word count: 8,642
Low document count: 4,174; word count: 6,646
Neutral document count: 1,899; word count: 3,139

Figure 5: Text mining for correlation between ratings and topic words

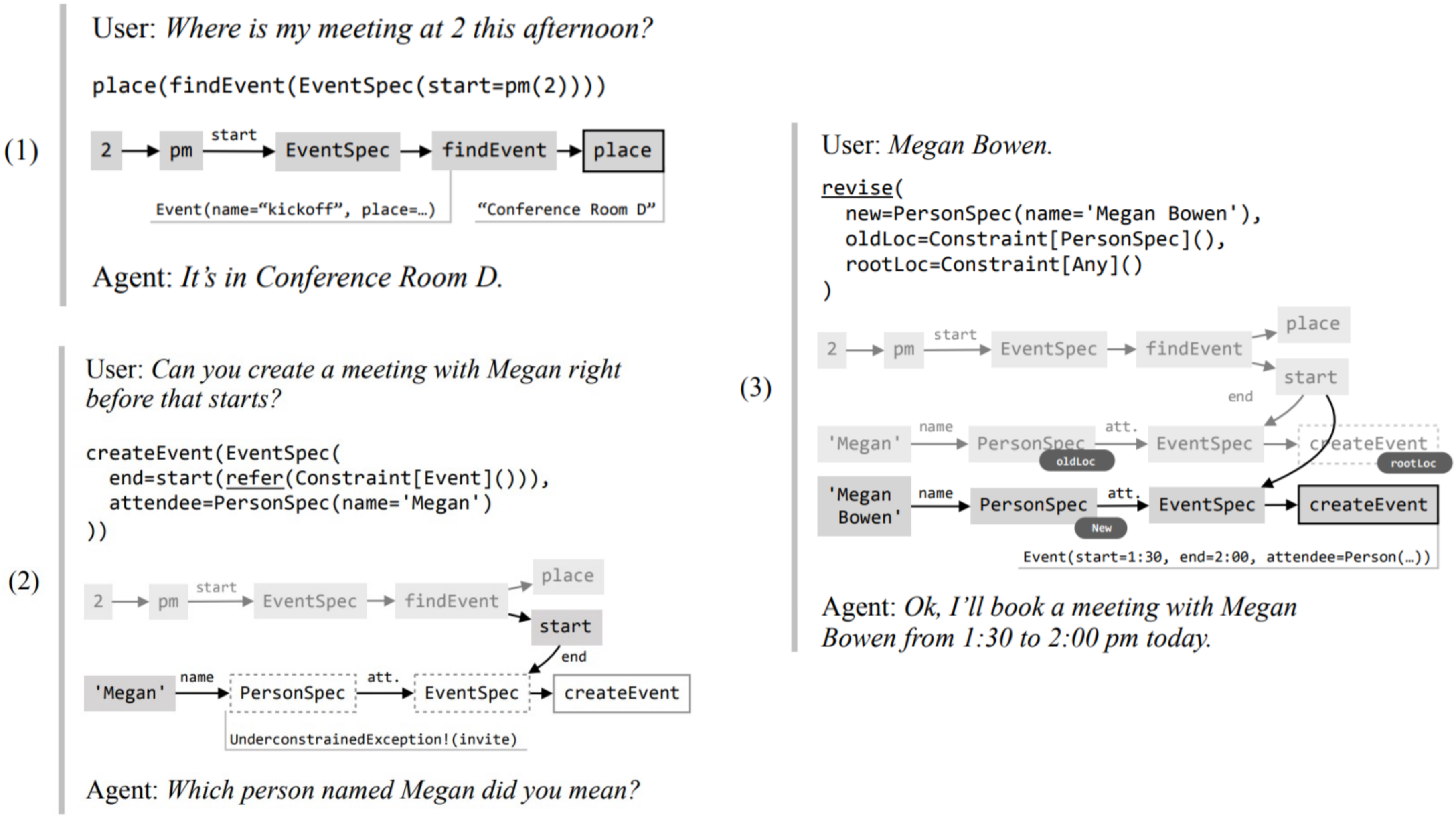
Average feedback rating (one to five stars, indicating how likely the user is willing to talk to the system again)

Year 2019 winner’s system



https://m.media-amazon.com/images/G/01/mobile-apps/dex/alexa/alexaprize/assets/challenge3/proceedings/Emory-Emora.pdf

Dialogue as Dataflow: New Paradigm of Learning Microsoft Semantic Machine's Solution



(1)

User: *Where is my meeting at 2 this afternoon?*

```
place(findEvent(EventSpec(start=pm(2))))
```

Agent: *It's in Conference Room D.*

(2)

User: *Can you create a meeting with Megan right before that starts?*

```
createEvent(EventSpec(
  end=start(refer(Constraint[Event]())),
  attendee=PersonSpec(name='Megan')
))
```

Agent: *Which person named Megan did you mean?*

(3)

User: *Megan Bowen.*

```
revise(
  new=PersonSpec(name='Megan Bowen'),
  oldLoc=Constraint[PersonSpec](),
  rootLoc=Constraint[Any]()
)
```

Agent: *Ok, I'll book a meeting with Megan Bowen from 1:30 to 2:00 pm today.*

https://www.microsoft.com/en-us/research/blog/dialogue-as-dataflow-a-new-approach-to-conversational-ai/

https://www.microsoft.com/en-us/research/publication/task-oriented-dialogue-as-dataflow-synthesis/

Lots of Other Limits for ML4NLP

- Comprehensive QA
- Multitask learning
- Combined multimodal, logical and memory-based reasoning
- Learning from few examples

Still a very active research field

Presentation and Final Exam

- Presentation
  - TA released the order of presentations
  - Share your knowledge about projects you have done

- Final Exam
  - We will have about 10 questions in the final exam
    - Yes/no questions, calculation, reading programs, etc.
  - We will provide the example paper
  - Date/time: December 17, 4:30PM - 7:30PM

- It’s been an exciting semester!
- Good luck to the final!
- Thank you! ☺