Will Graphs Lead to the Next Breakthrough of Conversational AI?

Yu Su
The Ohio State University
Microsoft Semantic Machines

The Ohio State University

Microsoft

A Bit About Me

- Been working on knowledge graphs since 2014
  - Construction, querying, and reasoning
- Been working on conversational AI since 2015
  - Semantic parsing, task-oriented dialogue, question answering, embodied instruction following
- Been working at Microsoft Semantic Machines since 2018 and helped develop a new conversational interface for Outlook
- Won the 3rd place in the inaugural Amazon Alexa Prize Taskbot Challenge

Natural Language is the Universal Interface

Easier Said Than Done

yahoo!

US airlines are bracing customers for what will probably be another bumpy holiday weekend as the industry struggles to manage a surge in travel demand that probably exceeds its current capacity.

Yu Su, a computer science professor at Ohio State University, was stranded last Saturday night in Charlotte, North Carolina after his connecting flight home never left.

The airline didn't cancel the 8:30 pm flight until around midnight after numerous delays that created "the delusion of hope," said Yu, who never got a clear explanation for the problem.

https://www.yahoo.com/now/flight-trouble-strained-us-airlines-014234130.html

Easier Said Than Done

When is my flight to Seattle?

Tomorrow at 5:00 pm.

How long will it take to get to the airport?

It will take 20 minutes according to Google Maps.

Book a Uber 1.5 hours before that.

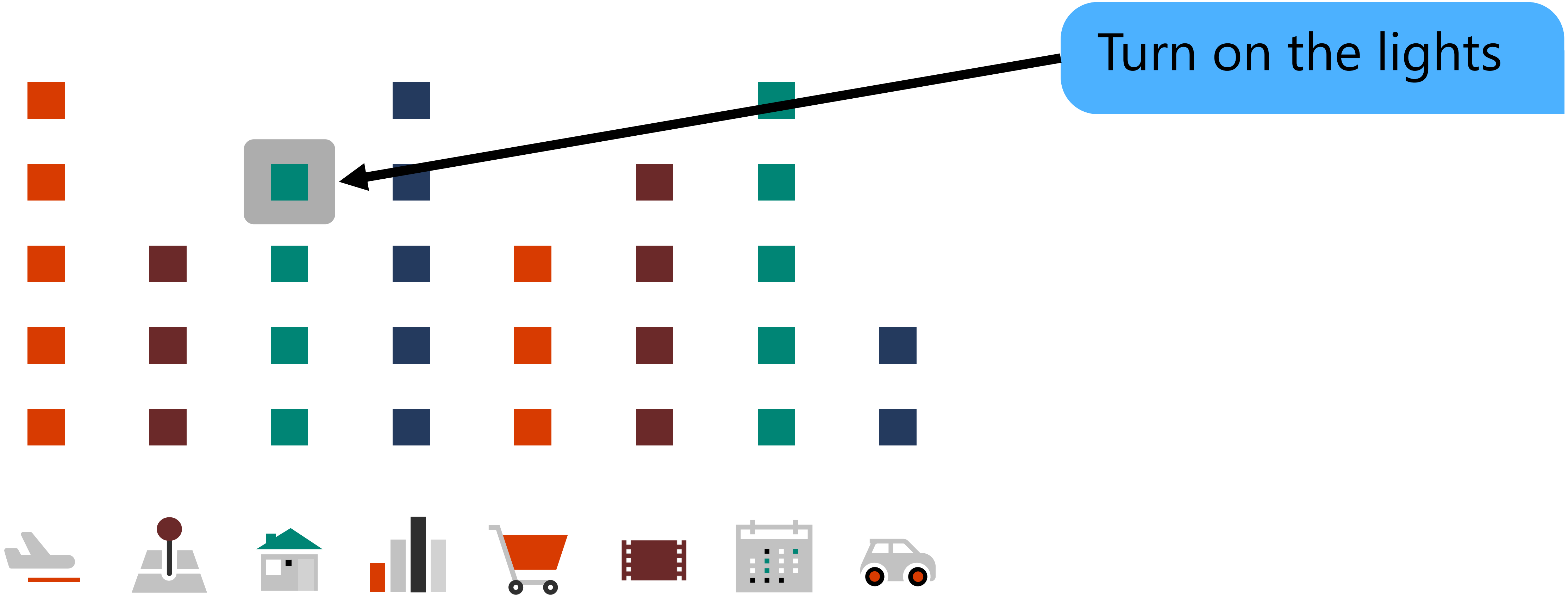
Sure. Booked a Uber for 3:30 pm tomorrow to the Columbus airport.

Any good Chinese restaurants close to my hotel?

According to Yelp, Haidilao has 4.5 stars and is 2-min walk from Hyatt.

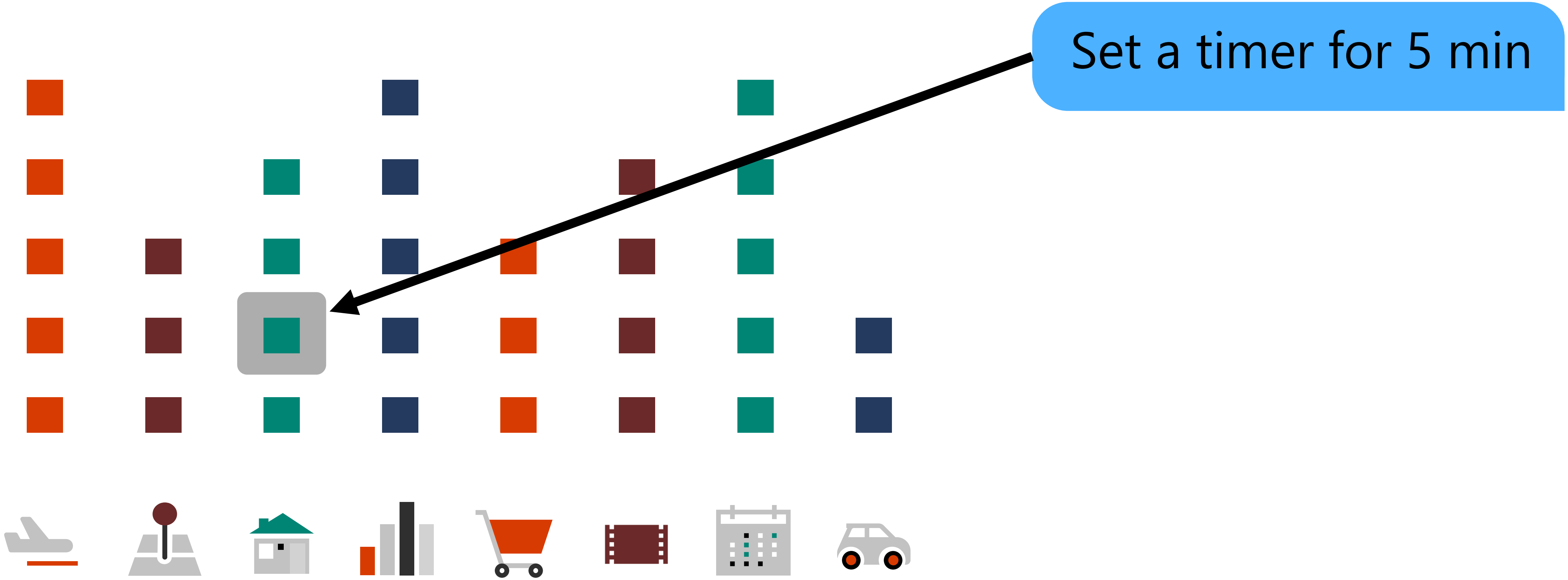
Standard Approaches to Language Interfaces

Standard Approach: Intents and Slots



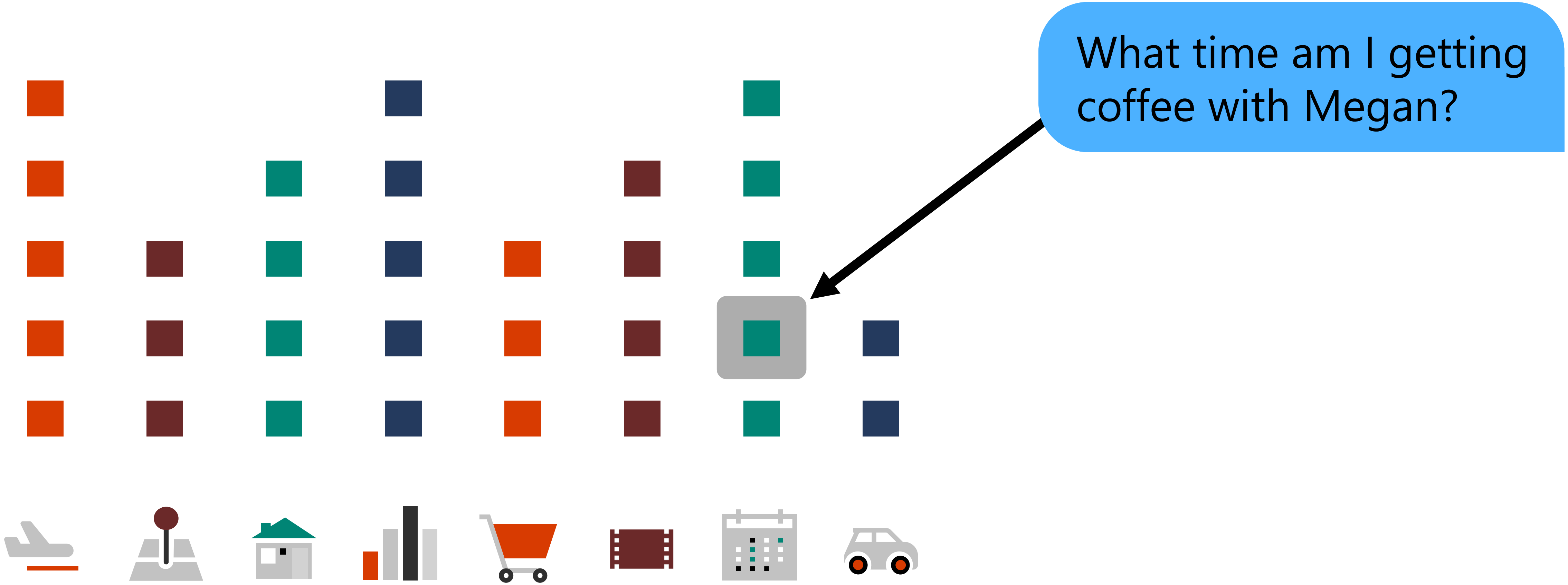
[(Bobrow et al., 1977), (Zue et al., 1991), (Henderson et al., 2015)]

Standard Approach: Intents and Slots



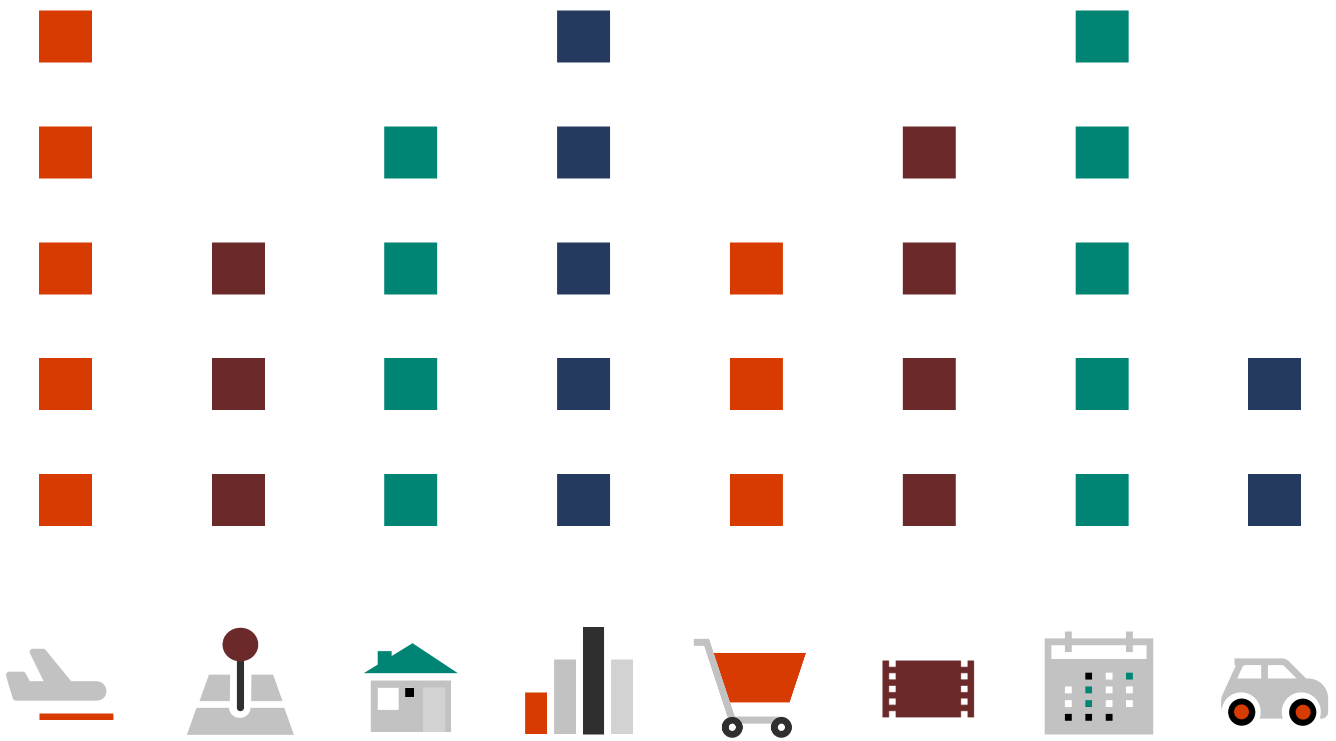
[(Bobrow et al., 1977), (Zue et al., 1991), (Henderson et al., 2015)]

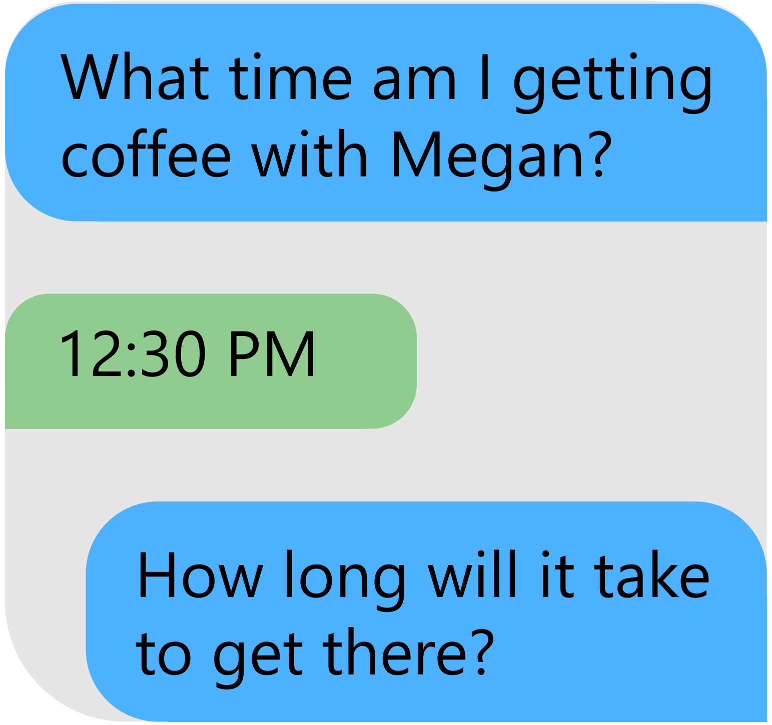
Standard Approach: Intents and Slots



[(Bobrow et al., 1977), (Zue et al., 1991), (Henderson et al., 2015)]

Standard Approach: Intents and Slots





[(Bobrow et al., 1977), (Zue et al., 1991), (Henderson et al., 2015)]

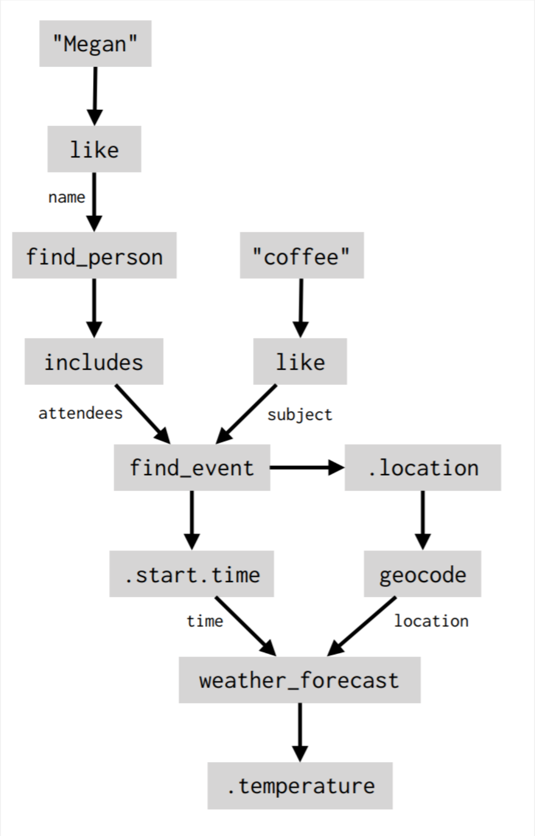
Proposition:

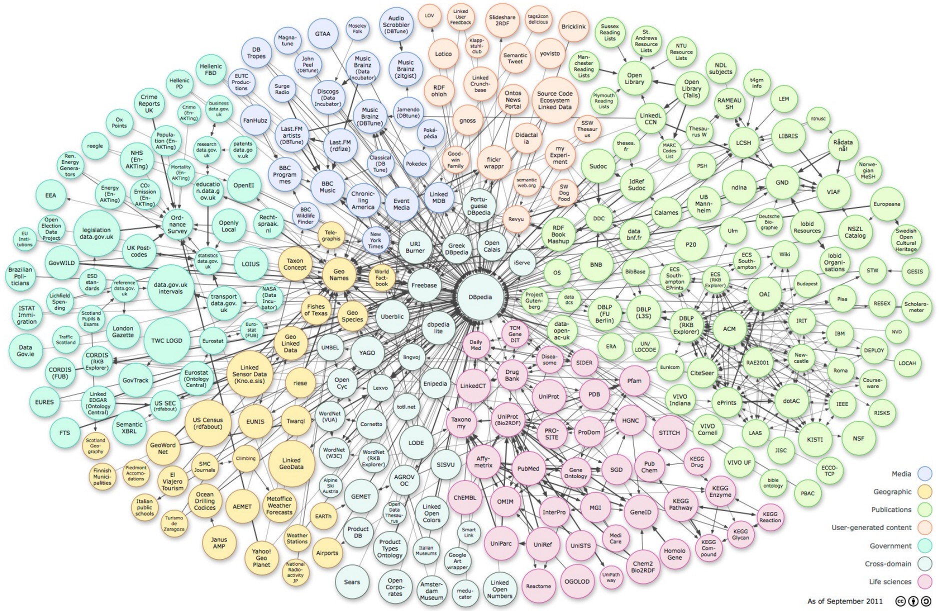
The world where conversational AI agents are grounded is inherently structured and interconnected, so graphs should be an integral part of conversational AI.

How Graphs May Help?

Dialogue as Dataflow Graph	Broad-Coverage Meaning Representation

What's the temperature going to be for my coffee with Megan?





https://lod-cloud.net/

Microsoft

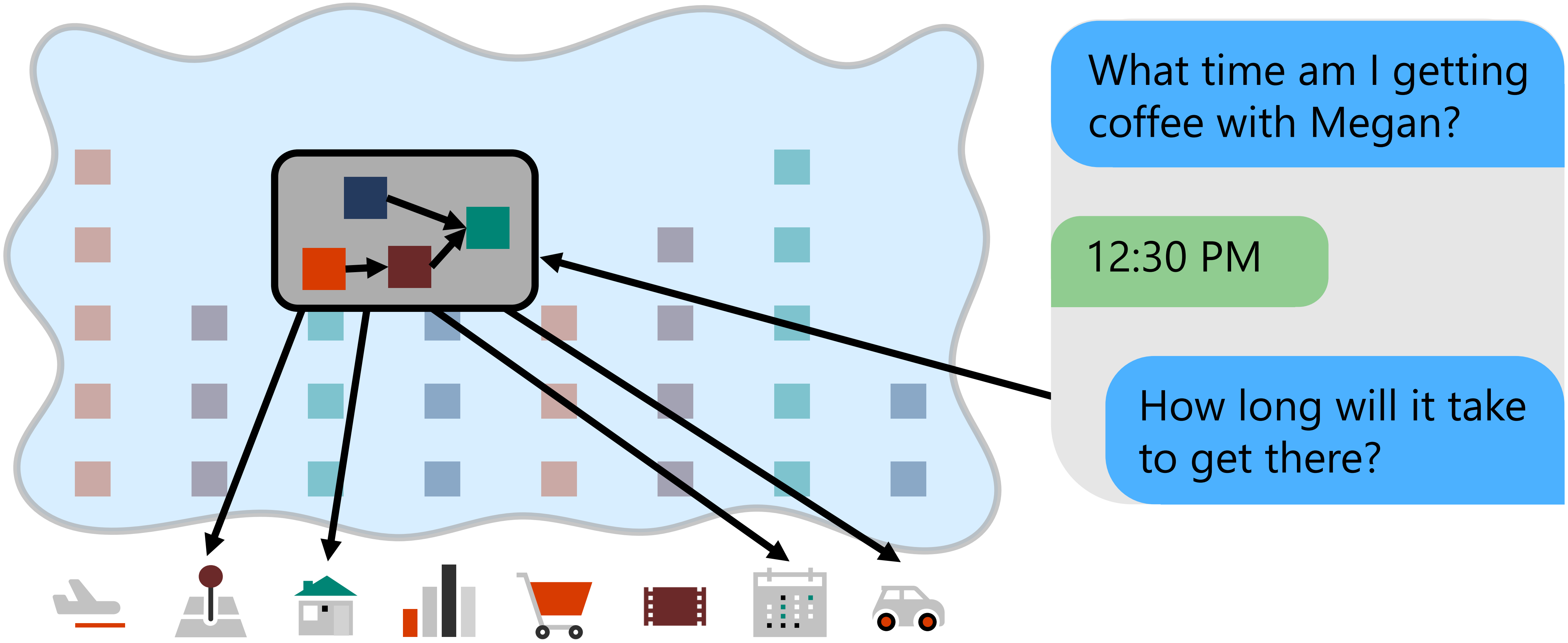
semanticmachines

Task-Oriented Dialogue as Dataflow Synthesis

Semantic Machines, Jacob Andreas, John Bufe, David Burkett, Charles Chen, Josh Clausman, Jean Crawford, Kate Crim, Jordan DeLoach, Leah Dorner, Jason Eisner, Hao Fang, Alan Guo, David Hall, Kristin Hayes, Kellie Hill, Diana Ho, Wendy Iwaszuk, Smriti Jha, Dan Klein, Jayant Krishnamurthy, Theo Lanman, Percy Liang, Christopher H Lin, Ilya Lintsbakh, Andy McGovern, Aleksandr Nisnevich, Adam Pauls, Dmitrij Petters, Brent Read, Dan Roth, Subhro Roy, Jesse Rusak, Beth Short, Div Slomin, Ben Snyder, Stephon Striplin, Yu Su, Zachary Tellman, Sam Thomson, Andrei Vorobev, Izabela Witoszko, Jason Wolfe, Abby Wray, Yuchen Zhang, Alexander Zotov

[Transactions of ACL 2020]

SM View: Contextual Program Synthesis



Five Key Ideas

1. Dialogs are programs (program synthesis)
2. Complex tasks are built from simpler ones (compositionality)
3. Meanings depend on context (metacomputation)
4. Things will go wrong (exception handling)
5. Systems should tell the truth (dynamic grounded generation)

Idea 1: Dialog as Program Synthesis

Dialogue as Program Synthesis: Intents and Slots

Turn on the lights

```
turn_on_lights()
```

Set a timer for 5 minutes

```
set_timer(time: “5 minutes”)
```

What time am I getting coffee with Megan?

```
time_of_calendar_event_by_title_and_attendee(
    “coffee”, “Megan”)
```

Dialogue as Program Synthesis: General Programs

What time am I getting coffee with Megan?

```
[p] = find_person(name: like(“Megan”))

[e] = find_event(
    subject: like(“coffee”),
    attendees: includes([p])))

describe([e].start.time)
```

[c.f. (Zelle 1995, Dong & Lapata 2016)]

Dialogue as Program Synthesis: General Programs

What time am I getting coffee with Megan?

```
[p] = find_person(name: like(“Megan”))

[e] = find_event(
    subject: like(“coffee”),
    attendees: includes([p]))

describe([e].start.time)
```

[c.f. (Zelle 1995, Dong & Lapata 2016)]

Idea 2: Complex Tasks via Compositionality

Compositionality: Within a Turn

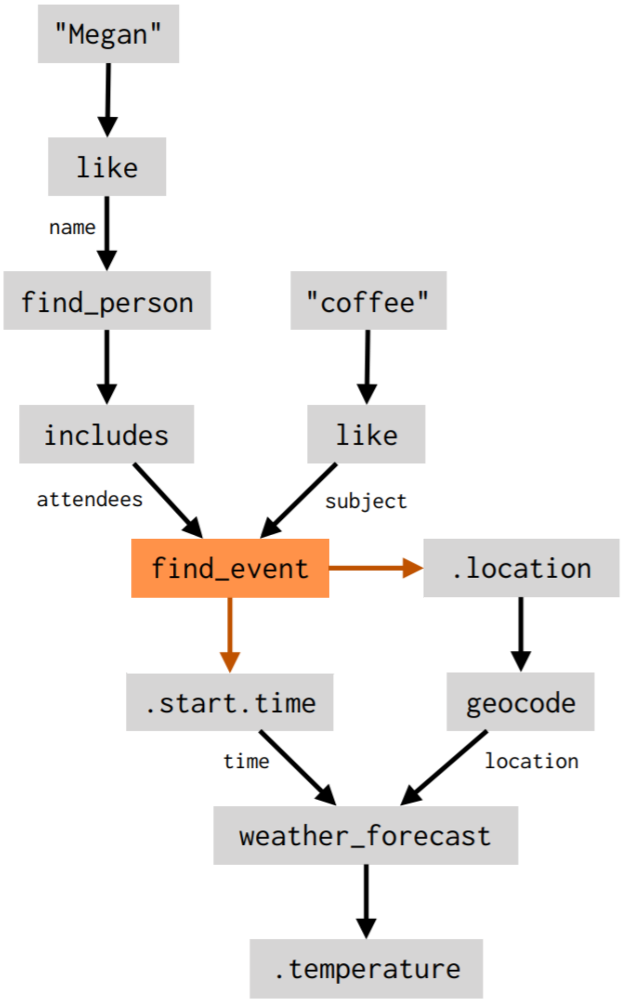
What's the temperature going to be for my coffee with Megan?

74° F

```
[p] = find_person(name: like("Megan"))
[e] = find_event(
    subject: like("coffee"),
    attendees: includes([p])))

[l] = geocode([e].location)
[w] = weather_forecast(
    location: [l]
    time: [e].start.time)
describe([w].temperature)
```

Dataflow

```
[p] = find_person(name: like("Megan"))
[e] = find_event(
    subject: like("coffee"),
    attendees: includes([p])))

[l] = geocode([e].location)
[w] = weather_forecast(
    location: [l]
    time: [e].start.time)
describe([w].temperature)
```



Compositionality: Multi-Turn

What time am I getting coffee with Megan?

12:30 PM

What's the temperature going to be?

74° F

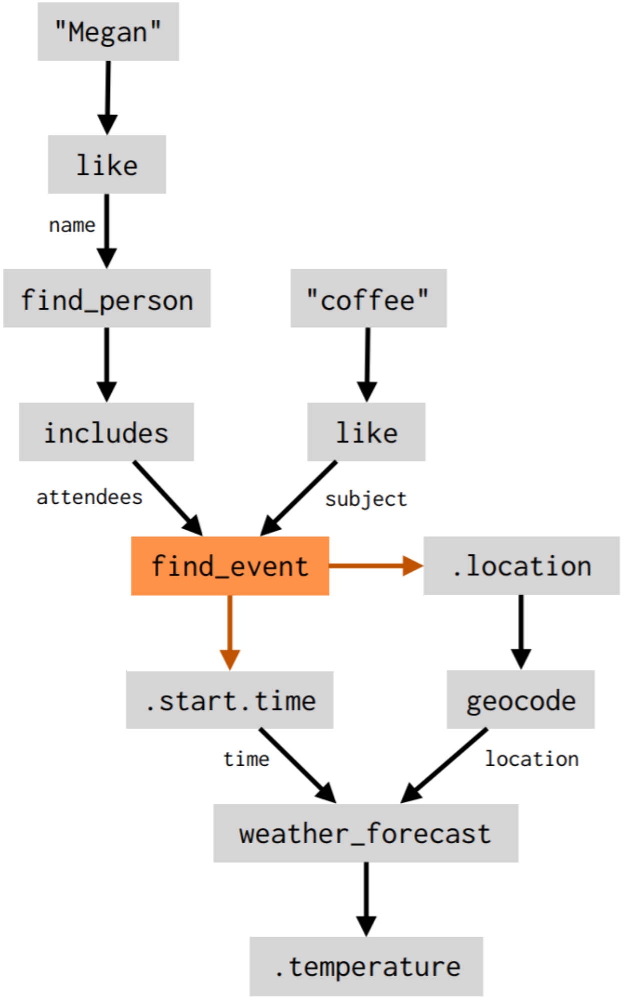
```
[p] = find_person(name: like("Megan"))

[e] = find_event(
    subject: like("coffee"),
    attendees: includes([p])))

describe([e].start.time)
```

```
[l] = geocode([e].location)

[w] = weather_forecast(
    location: [l]
    time: [e].start.time)

describe([w].temperature)
```

Dataflow

```
[p] = find_person(name: like(“Megan”))

[e] = find_event(
    subject: like(“coffee”),
    attendees: includes([p])))

describe([e].start.time)
```

```
[l] = geocode([e].location)

[w] = weather_forecast(
    location: [l]
    time: [e].start.time)

describe([w].temperature)
```



Idea 3: Meanings Depend on Context

Context through Metacomputation: Reference

What will the weather be like?

```
weather_forecast(
    location: here(),
    time: now()
)
```

When's my coffee with Megan?

12:30 PM

What will the weather be like?

```
weather_forecast(
    location: [e].location,
    time: [e].time
)
```

[cf. (Lappin & Leass 1994), (Zettlemoyer & Collins 2009), (Suhr et al. 2018)]

Context through Metacomputation: Reference

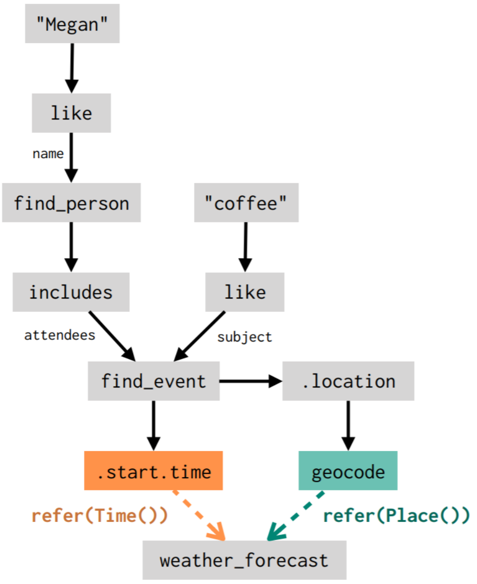
What will the weather be like?

When's my coffee with Megan?

12:30 PM

What will the weather be like?

```
weather_forecast(
    location:
        refer(Constraint[Place]()),
    time:
        refer(Constraint[Time]()))
```



[cf. (Lappin & Leass 1994), (Zettlemoyer & Collins 2009), (Suhr et al. 2018)]

Context through Metacomputation: Revision

Do I have any meetings today?

No

What about tomorrow?

```
[d] = today()
[e] = find_event(start.date: [d])
describe([e])
```

```
[d2] = tomorrow()
[e2] = find_event(start.date: [d2])
describe([e2])
```

What's the weather today?

Sunny

What about tomorrow?

What's the weather going to be like during my third meeting today?

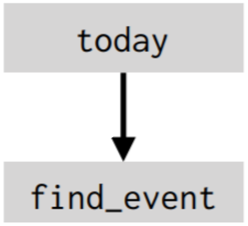
Cloudy

What about tomorrow?

Context through Metacomputation: Revision

Do I have any meetings today?

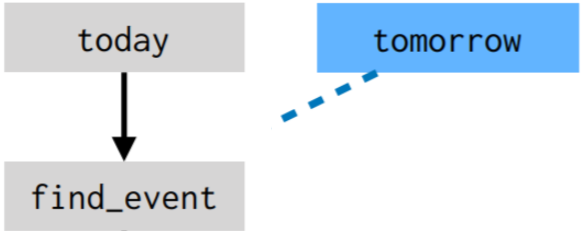
```
[d] = today()
[e] = find_event(
        start.date:[d]
      )
describe([e])
```



No

What about tomorrow?

```
[f] = revise(
  rootLoc:
    RoleConstraint(output),
  oldLoc: Constraint[Date](),
  new: tomorrow()
)

describe([f])
```



Yes

Program transformation (Visser 2001)

Idea 4: Things Will Go Wrong

Exception Handling

Disambiguation, Confirmation, Missing Slots, API Errors, …

Set up a meeting with Megan

```
[p] = find_person(
        name: like(“Megan”))

[e] = create_event(
        attendees: includes([p]))
```

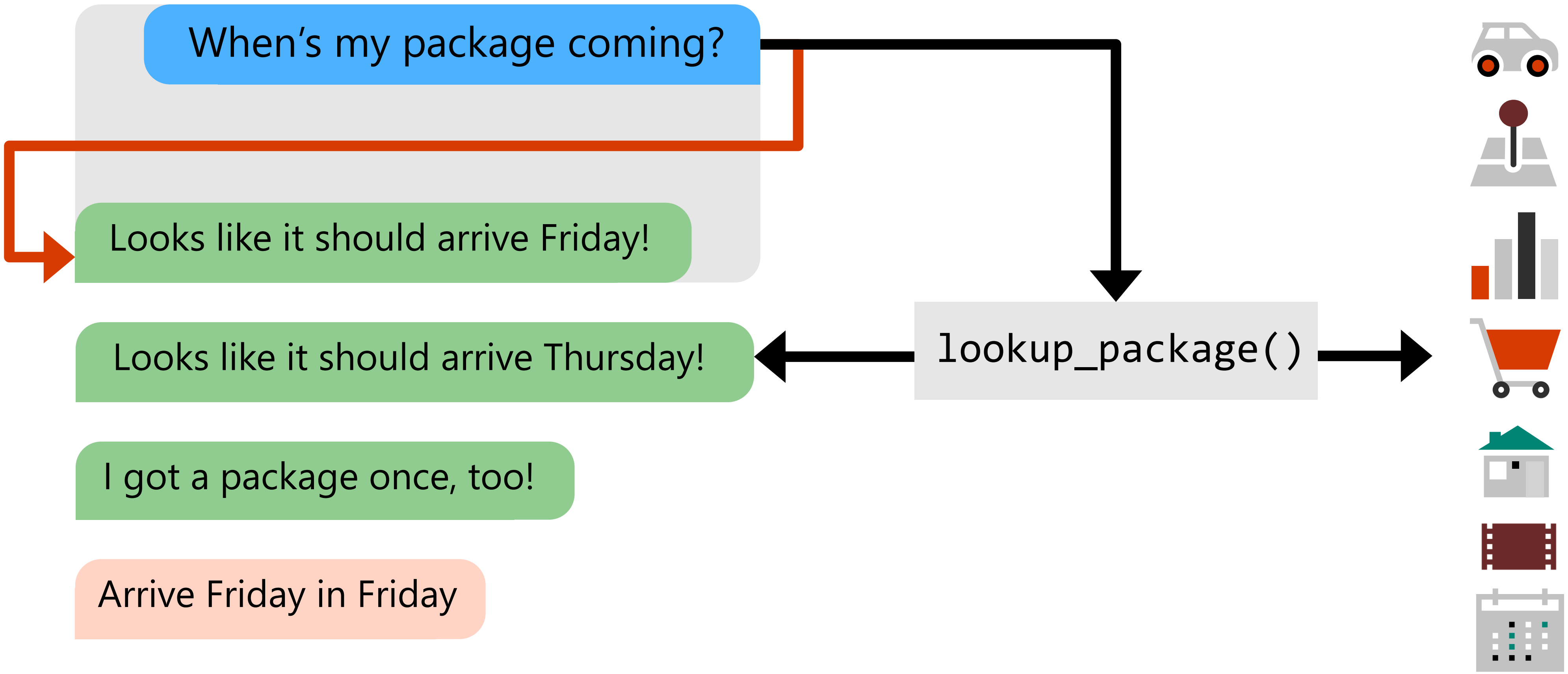
```
Except: MultipleMatchesFound([p])
Except: NoMatchFound([p])
Except: APIError([e])
Except: MissingSlot([e])
```

```
Except: MultipleMatchesFound([p])
```

Did you mean Megan Bowen?

Idea 5: Truthful Generation

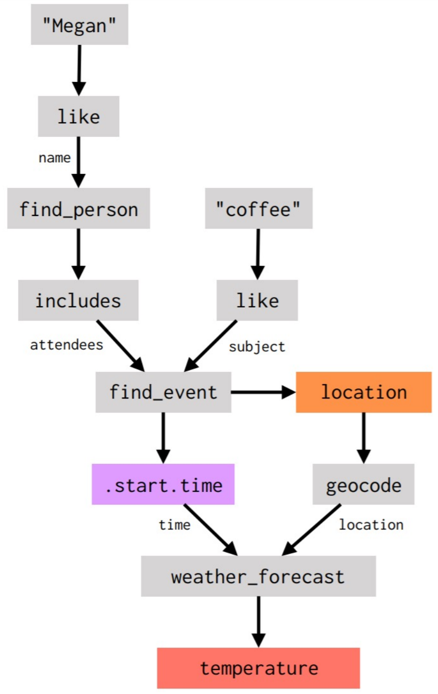
Grounded vs. Ungrounded Generation



Dynamic Grounded Generation

What’s the temperature going to be at my coffee with Megan?

```
...
describe([temperature])
```



It will be 74° F tomorrow at noon at Rosie’s Café.

Demo

May

S	M	T	W	T	F	S
2	3	4	5	6	7	8

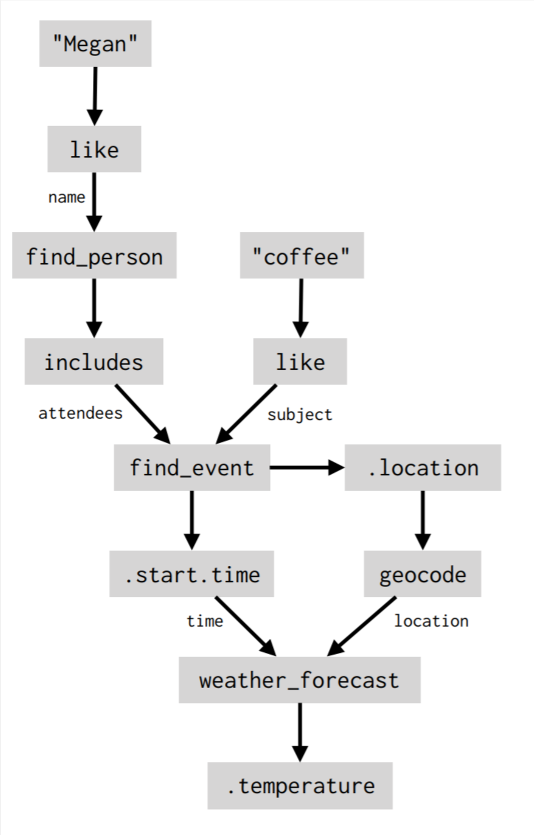
9 AM
10 AM
11 AM
12 PM
1 PM
2 PM
3 PM
4 PM
5 PM
6 PM
7:05 PM

5

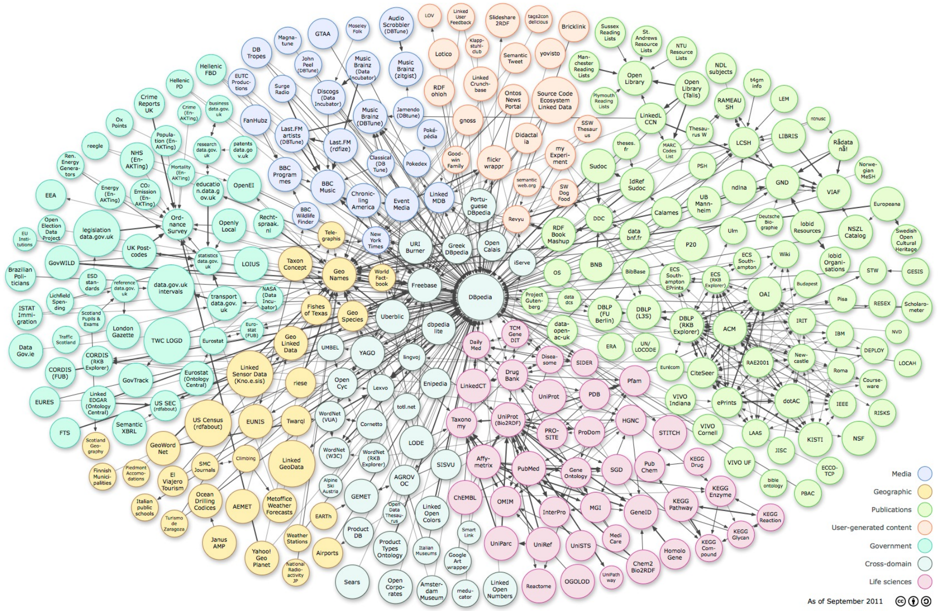
How Graphs May Help?

Dialogue as Dataflow Graph

What's the temperature going to be for my coffee with Megan?



Broad-Coverage Meaning Representation



https://lod-cloud.net/

Beyond I.I.D.: Three Levels of Generalization for Question Answering on Knowledge Bases

Yu Gu

Brian Sadler

Percy Liang

Xifeng Yan

w/ Sue Kase and Michelle Vanni

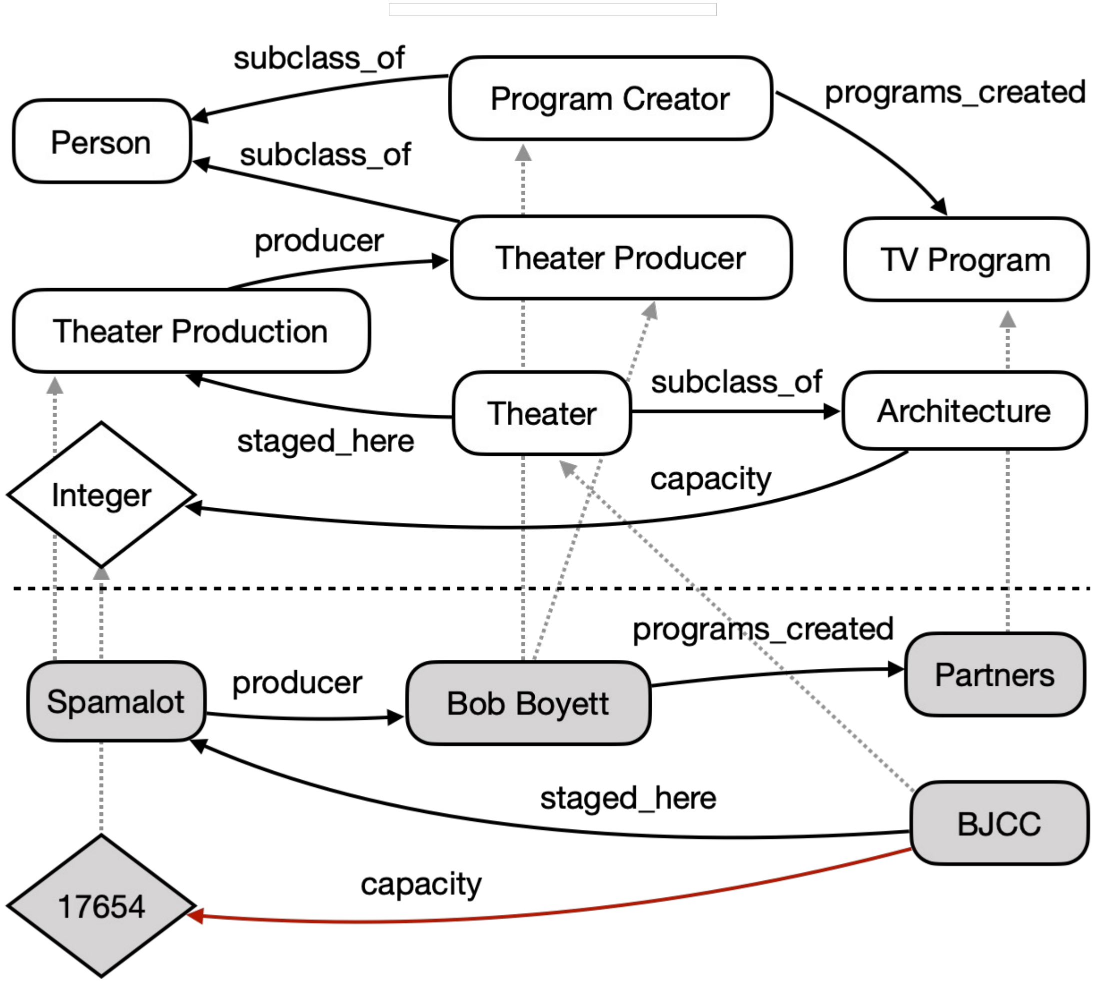
[The Web Conference 2021]

Universal Interface Needs
Broad-Coverage Meaning Representation

Large-Scale Knowledge Graphs as a Testbed

Modern knowledge graphs/bases (KGs/KBs) are extremely **large** and **broad**

- **Freebase**: 100 domains, 19,000 relations, 45 million entities, and 3 billion facts
- **Google Knowledge Graph**: 5 billion entities and 500 billion facts (as of 2020)[1]



[1] https://blog.google/products/search/about-knowledge-graph-and-knowledge-panels/

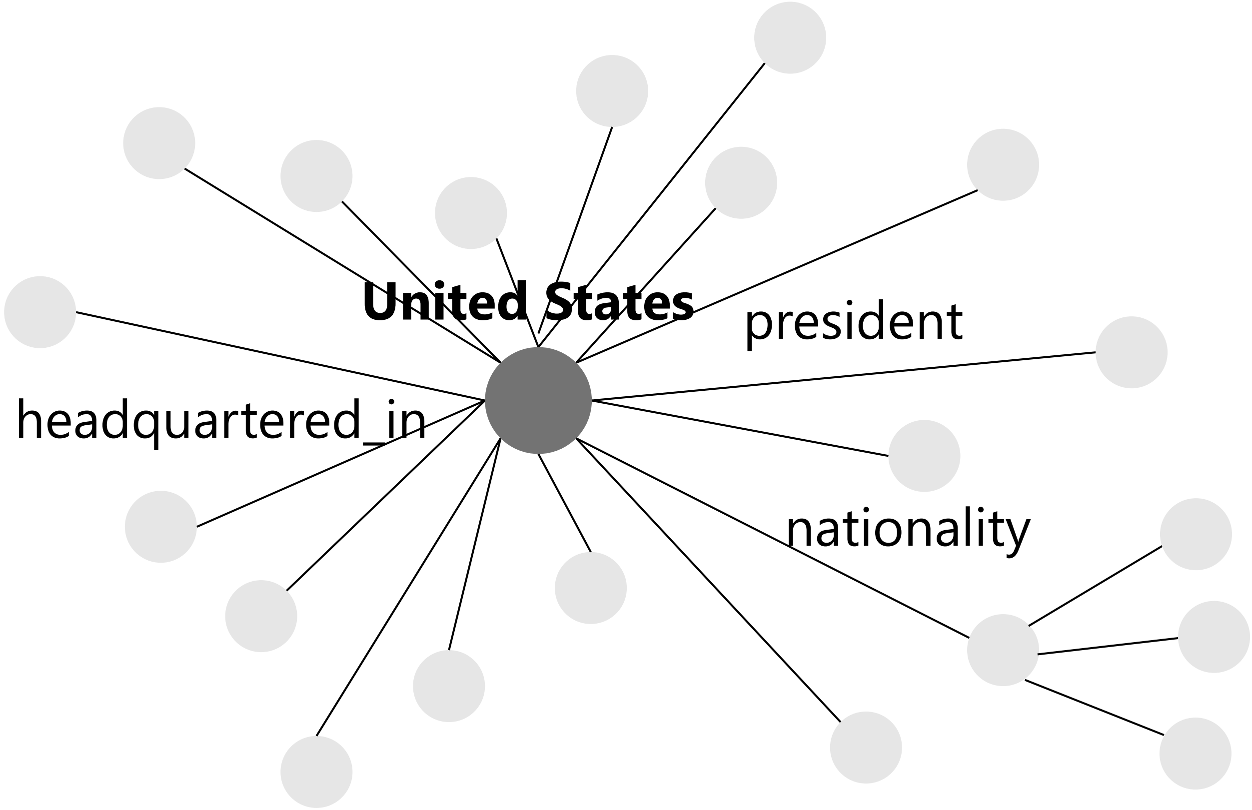
New Challenges for Broad-Coverage Conversational AI

Question answering on large-scale KGs reveals (new) challenges for developing broad-coverage conversational AI

Large Search Space	Non-I.I.D. Generalization	Semantic Ambiguities

Large Search Space

Which American actor has got the most Oscars nominations?



Q: How many **1-hop neighbors** does United States have in Freebase?

A: 1,092,532

Q: How about **2-hop**?

A: 82,130,962

Combinatorial explosion! 😱

Non-I.I.D. Generalization

Practical KBQA models should be **built with non-i.i.d. generalizability**

Training Data

- *Who is the producer of Spamalot?*

  (AND Theater_Producer (JOIN (R producer) Spamalot))

- *How many plays has Bob Boyett produced?*

  (COUNT (AND Theater_Production
  (JOIN producer Bob_Boyett))

- *Find plays that were staged in large theaters that could hold at least 20,000 people.*

  (AND Theater_Production
  (JOIN (R staged_here) (JOIN (GE capacity 20000)))

Non-I.I.D. Generalization

Practical KBQA models should be **built with non-i.i.d. generalizability**

Training Data

- *Who is the producer of Spamalot?*
  (AND Theater_Producer (JOIN (R producer) Spamalot))
- *How many plays has Bob Boyett produced?*
  (COUNT (AND Theater_Production
  (JOIN producer Bob_Boyett))
- *Find plays that were staged in large theaters that could hold at least 20,000 people.*
  (AND Theater_Production
  (JOIN (R staged_here) (JOIN (GE capacity 20000)))

I.I.D. Generalization

- *How many theater productions has Oprah produced?*
  (COUNT (AND Theater_Production (JOIN producer Oprah_Winfrey))

Non-I.I.D. Generalization

Practical KBQA models should be **built with non-i.i.d. generalizability**

Training Data

- *Who is the producer of Spamalot?*
  (AND Theater_Producer (JOIN (R producer) Spamalot))
- *How many plays has Bob Boyett produced?*
  (COUNT (AND Theater_Production
  (JOIN producer Bob_Boyett))
- *Find plays that were staged in large theaters that could hold at least 20,000 people.*
  (AND Theater_Production
  (JOIN (R staged_here) (JOIN (GE capacity 20000)))

I.I.D. Generalization

- *How many theater productions has Oprah produced?*
  (COUNT (AND Theater_Production (JOIN producer Oprah_Winfrey))

Compositional Generalization

- *Bob Boyett's production was housed in what theater capable of holding at least 10,000 people?*
  (AND Theater (AND (GE capacity 10000)
  (JOIN staged_here (JOIN producer Bob_Boyett))))

Non-I.I.D. Generalization

Practical KBQA models should be **built with non-i.i.d. generalizability**

Training Data

- *Who is the producer of Spamalot?*
  (AND Theater_Producer (JOIN (R producer) Spamalot))
- *How many plays has Bob Boyett produced?*
  (COUNT (AND Theater_Production
  (JOIN producer Bob_Boyett))
- *Find plays that were staged in large theaters that could hold at least 20,000 people.*
  (AND Theater_Production
  (JOIN (R staged_here) (JOIN (GE capacity 20000)))

I.I.D. Generalization

- *How many theater productions has Oprah produced?*
  (COUNT (AND Theater_Production (JOIN producer Oprah_Winfrey))

Compositional Generalization

- *Bob Boyett's production was housed in what theater capable of holding at least 10,000 people?*
  (AND Theater (AND (GE capacity 10000)
  (JOIN staged_here (JOIN producer Bob_Boyett))))

Zero-Shot Generalization

- *How many TV programs has Bob Boyett created?*
  (COUNT (AND TV_Program (JOIN (R program_created) Bob_Boyett))

Semantic Ambiguities

Flaminia Brasini was the designer of what game?

Q: This question is referring to the relation of …

a. computer.computer.key_designers

b. sports.golf_course.designer

c. games.game.designer

Prior Work on KBQA

EMNLP'13

Semantic Parsing on Freebase from Question-Answer Pairs

Jonathan Berant Andrew Chou Roy Frostig Percy Liang

EMNLP'16

- Mostly focusing on i.i.d.
- Small scale (*w.r.t.* size, coverage, or diversity)

The Web as a

Alon Talmor
Tel-Aviv University
alontalmor@mail.tau.ac.il

Abstract

ISWC'19

LC-QuAD 2.0: A Large Dataset for Complex Question Answering over Wikidata and DBpedia

Mohnish Dubey[1,2], Debayan Banerjee[1,2], Abdelrahman Abdelkawi[2,3], and Jens Lehmann[1,2]

[1] Smart Data Analytics Group (SDA), University of Bonn, Germany

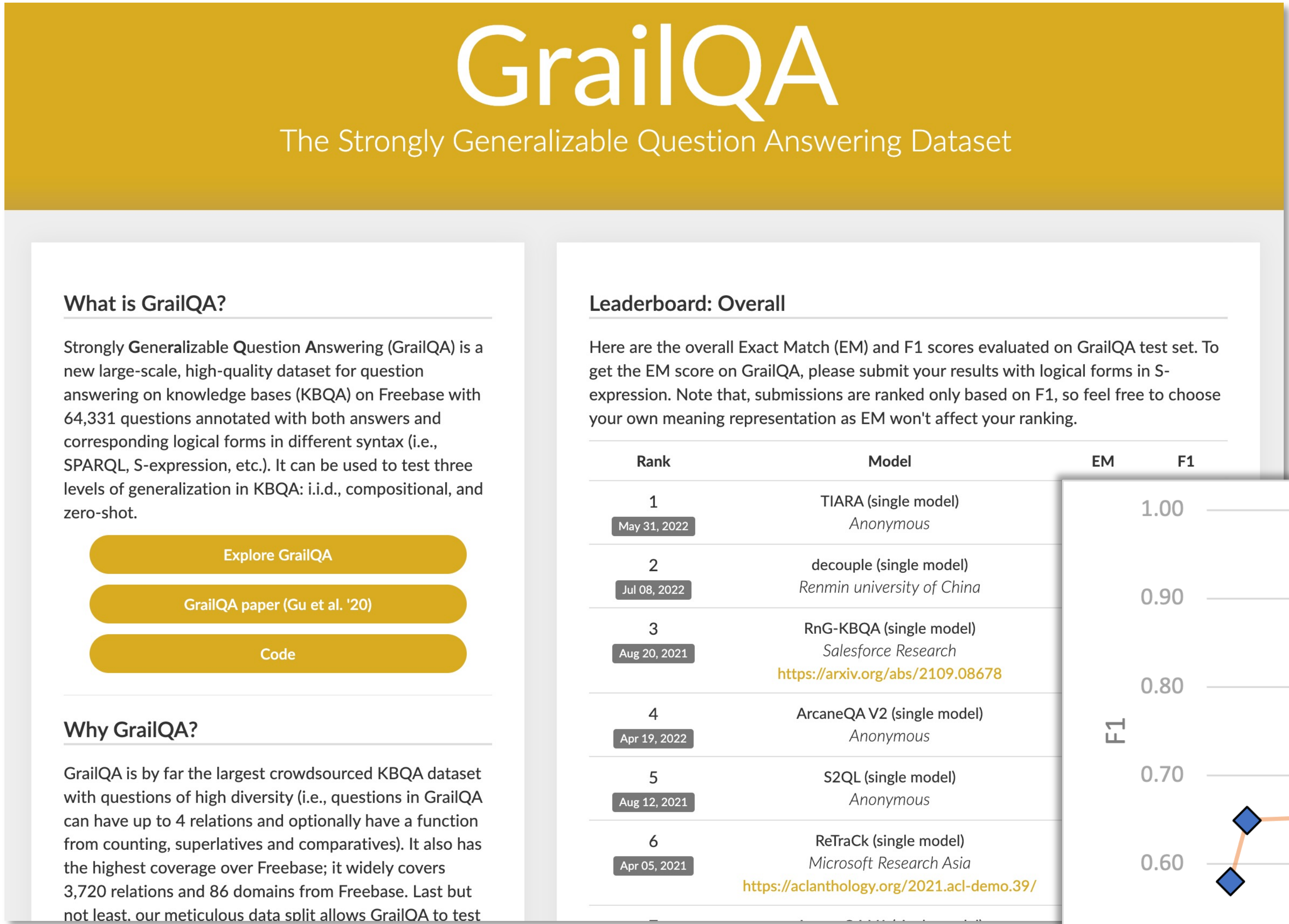
GrailQA: First Dataset for Strongly Generalizable KBQA

64,495 questions, **86** domains, **3,720** relations, **32,585** entities

Support three levels of generalization: **i.i.d.**, **compositional**, and **zero-shot**

Question	Domain	Answer	# of Relations	Function
Beats of Rage is a series of games made for which platform?	Computer Video Game	DOS	1	none
Which tropical cyclone has affected Palau and part of Hong Kong?	Location, Meteorology	Typhoon Sanba	3	none
Marc Bulger had the most yards rushing in what season?	Sports, American Football	2008 NFL Season	3	superlative
How many titles from Netflix have the same genre as The Big Hustle?	Media Common	20,104	2	count
What bipropellant rocket engine has less than 3 chambers?	Spaceflight	RD-114 RD-112, ...	1	comparative

Gu et al. Beyond I.I.D.: Three Levels of Generalization for Question Answering on Knowledge Bases. The Web Conference, 2021



GrailQA

The Strongly Generalizable Question Answering Dataset

What is GrailQA?

Strongly **G**ene**ral**izable **Q**uestion **A**nswering (GrailQA) is a new large-scale, high-quality dataset for question answering on knowledge bases (KBQA) on Freebase with 64,331 questions annotated with both answers and corresponding logical forms in different syntax (i.e., SPARQL, S-expression, etc.). It can be used to test three levels of generalization in KBQA: i.i.d., compositional, and zero-shot.

Explore GrailQA

GrailQA paper (Gu et al. '20)

Code

Why GrailQA?

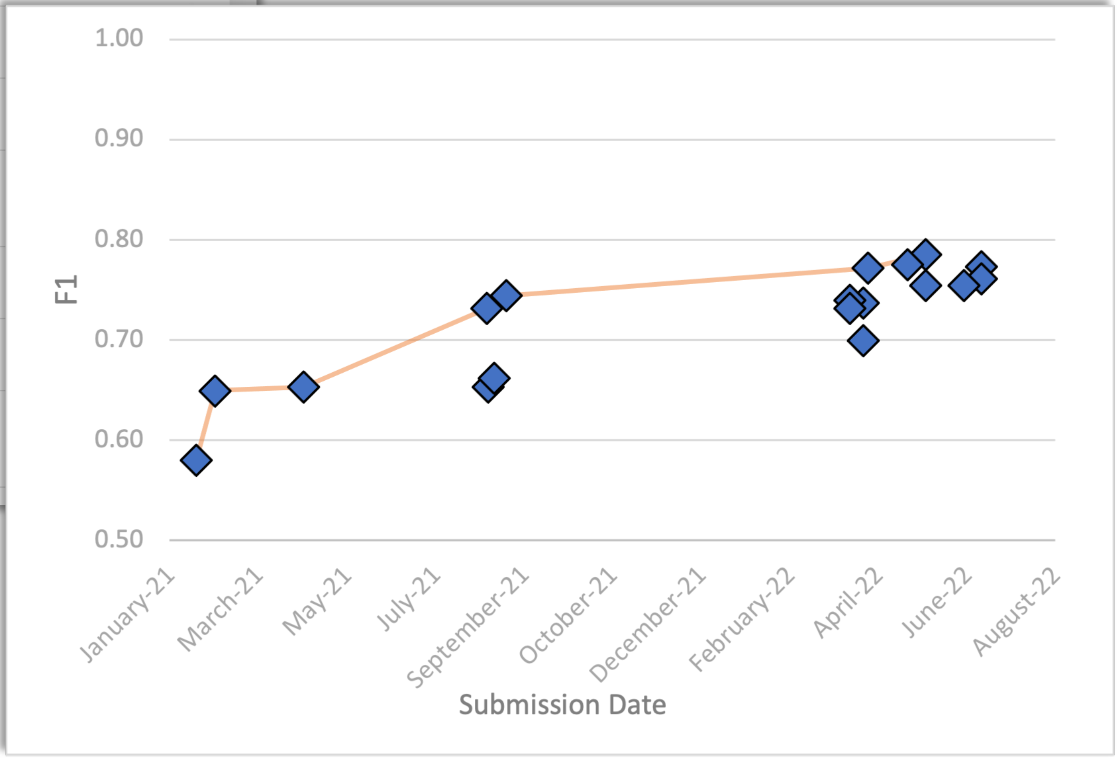
GrailQA is by far the largest crowdsourced KBQA dataset with questions of high diversity (i.e., questions in GrailQA can have up to 4 relations and optionally have a function from counting, superlatives and comparatives). It also has the highest coverage over Freebase; it widely covers 3,720 relations and 86 domains from Freebase. Last but not least, our meticulous data split allows GrailQA to test

Leaderboard: Overall

Here are the overall Exact Match (EM) and F1 scores evaluated on GrailQA test set. To get the EM score on GrailQA, please submit your results with logical forms in S-expression. Note that, submissions are ranked only based on F1, so feel free to choose your own meaning representation as EM won't affect your ranking.

Rank	Model	EM	F1
1 May 31, 2022	TIARA (single model) *Anonymous*		
2 Jul 08, 2022	decouple (single model) *Renmin university of China*		
3 Aug 20, 2021	RnG-KBQA (single model) *Salesforce Research* https://arxiv.org/abs/2109.08678		
4 Apr 19, 2022	ArcaneQA V2 (single model) *Anonymous*		
5 Aug 12, 2021	S2QL (single model) *Anonymous*		
6 Apr 05, 2021	ReTraCk (single model) *Microsoft Research Asia* https://aclanthology.org/2021.acl-demo.39/		

>1,600 downloads
20 submissions

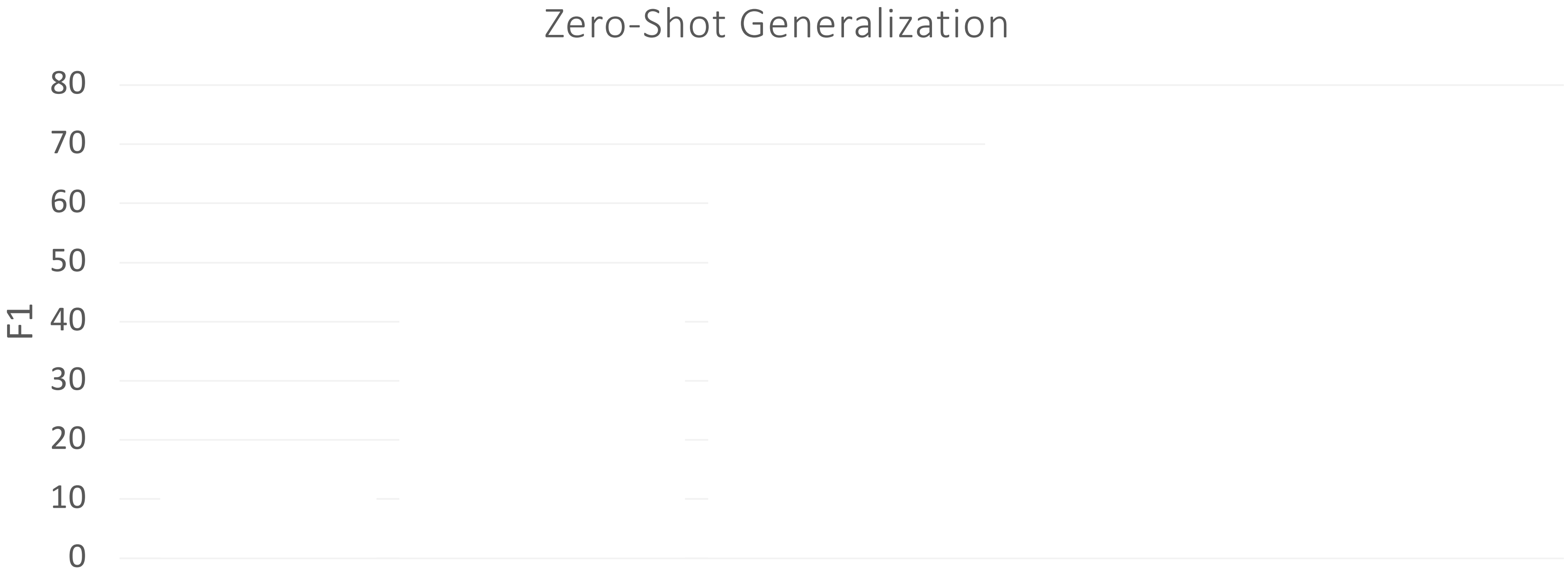


Homepage: https://dki-lab.github.io/GrailQA/

Code: https://github.com/dki-lab/GrailQA

[(Gu et al., 2021), (Ye et al., 2022), (Gu and Su, 2022)]

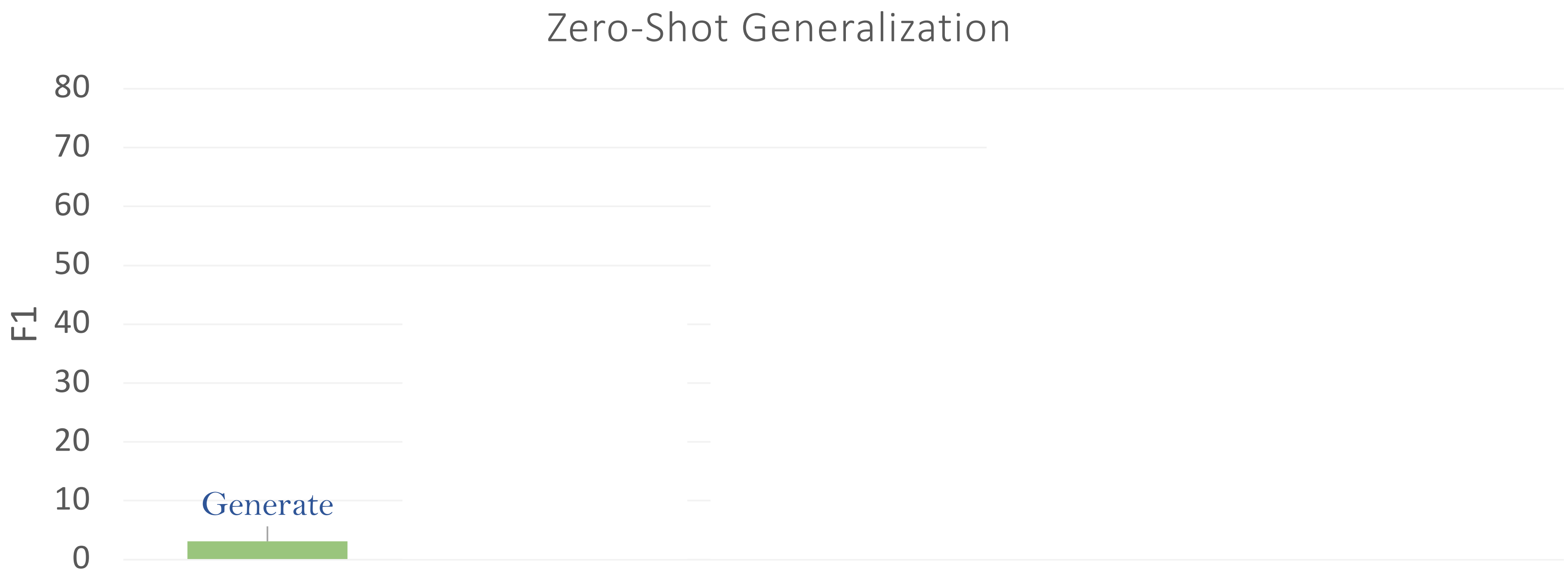
Lessons Learned So Far



[(Gu et al., 2021), (Ye et al., 2022), (Gu and Su, 2022)]

Lessons Learned So Far

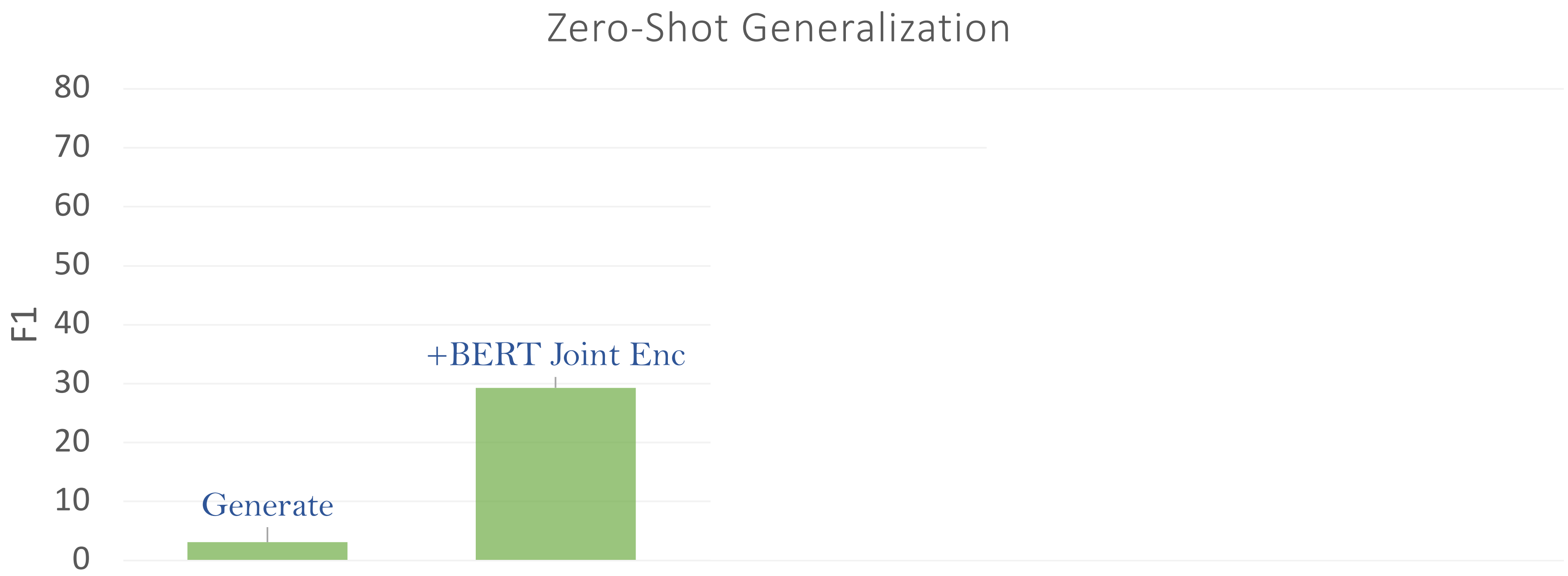
- Simple transductive parsing (Seq2Seq or Seq2Graph) is unlikely to suffice



[(Gu et al., 2021), (Ye et al., 2022), (Gu and Su, 2022)]

Lessons Learned So Far

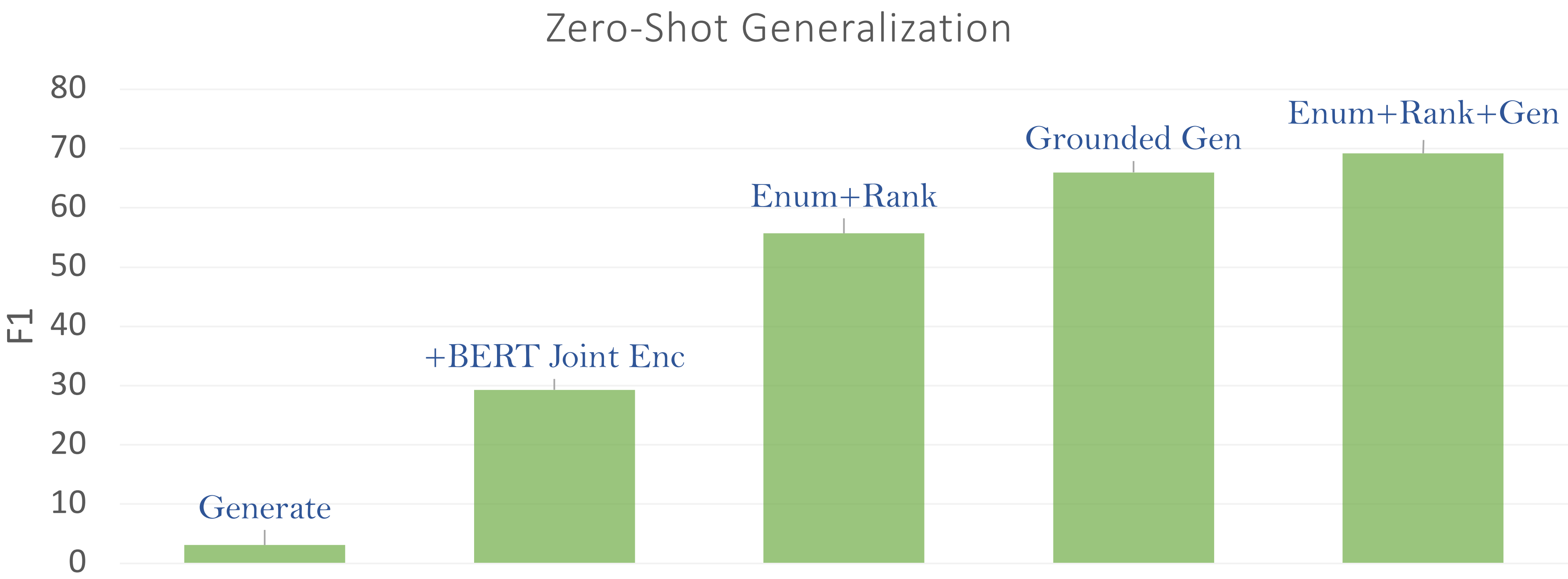
- Simple transductive parsing (Seq2Seq or Seq2Graph) is unlikely to suffice
- Joint utterance-schema encoding is critical for schema linking



[(Gu et al., 2021), (Ye et al., 2022), (Gu and Su, 2022)]

Lessons Learned So Far

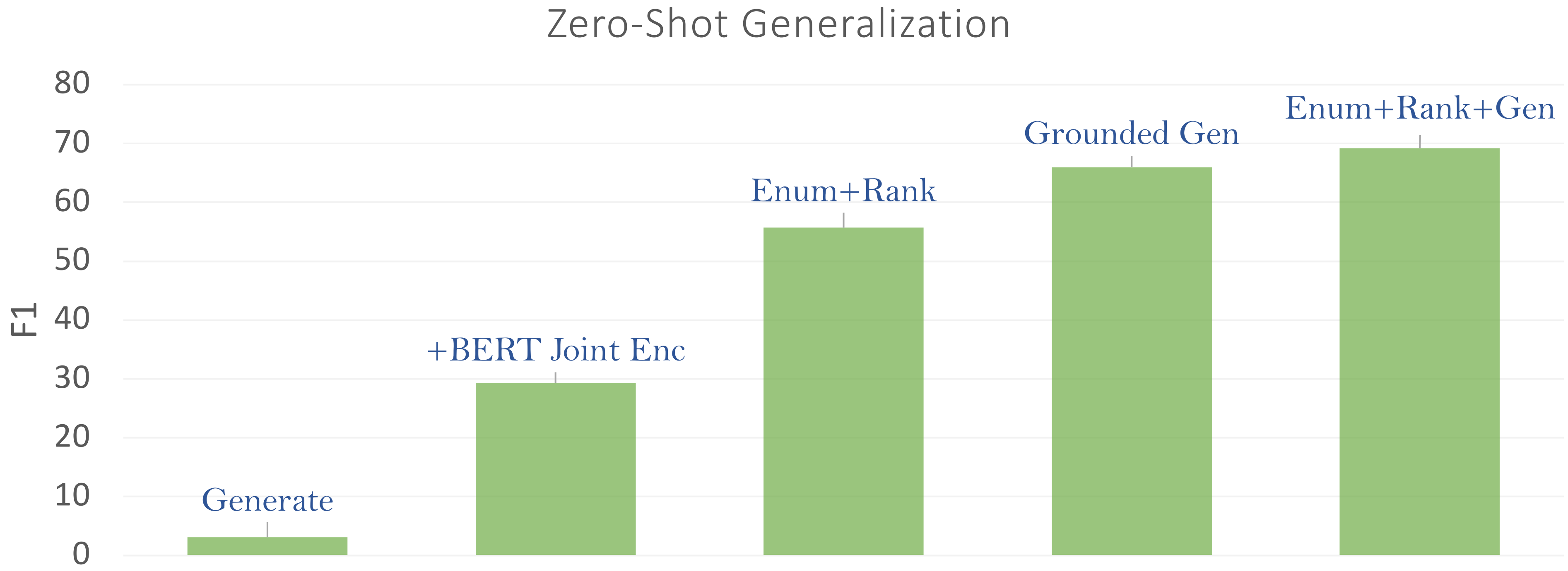
- Simple transductive parsing (Seq2Seq or Seq2Graph) is unlikely to suffice
- Joint utterance-schema encoding is critical for schema linking
- Grounded generation/search can go a long way



[(Gu et al., 2021), (Ye et al., 2022), (Gu and Su, 2022)]

Lessons Learned So Far

- Simple transductive parsing (Seq2Seq or Seq2Graph) is unlikely to suffice
- Joint utterance-schema encoding is critical for schema linking
- Grounded generation/search can go a long way
- **Zero-shot generalization is possible!**

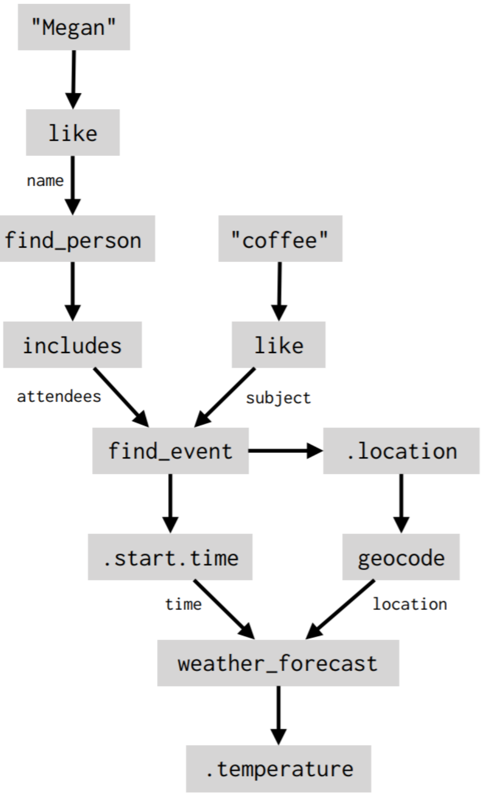


Recap

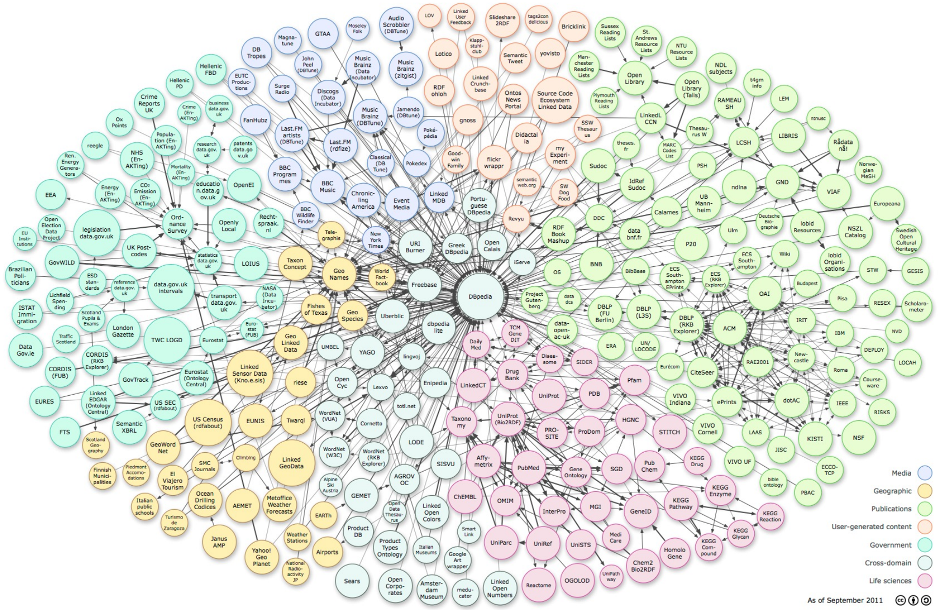
The world where conversational AI agents are grounded is inherently structured and interconnected, so graphs should be an integral part of conversational AI.

Dialogue as Dataflow Graph

What's the temperature going to be for my coffee with Megan?



Broad-Coverage Meaning Representation



https://lod-cloud.net/

Interesting Future Directions

- Reconciling dataflow graphs and knowledge graphs
- Graph neural networks for context modeling
- Data collection, sample efficiency, learning from use
- Foundation models for graph-based conversational AI

Thanks &