PolyAI

Conversational AI Platform for Customer Service

Nikola Mrkšić, Cofounder & CEO

PolyAI

Founded in

2017

UNIVERSITY OF CAMBRIDGE

Raised

$15 million

POINT72 VENTURES

Amadeus Capital Partners

Published

100+

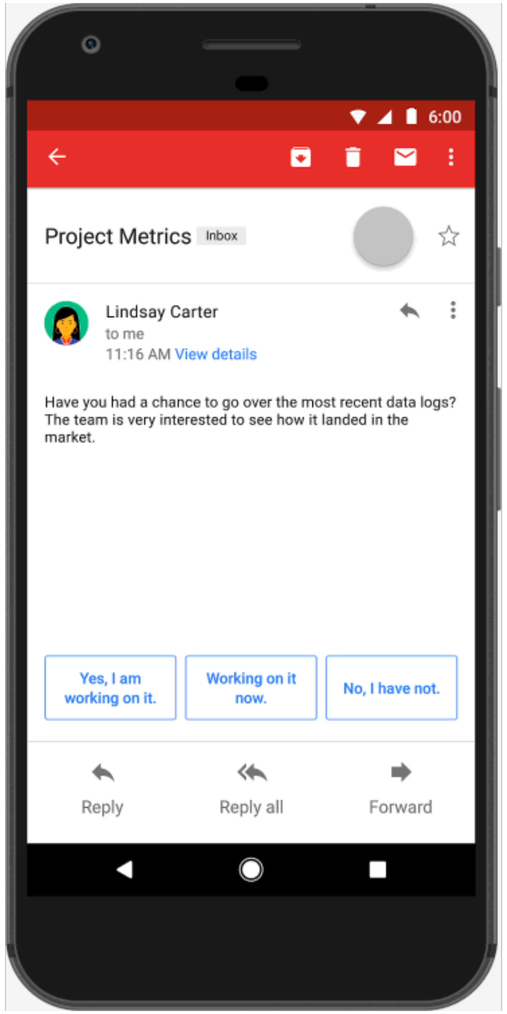
Research papers, with over 4,000 citations between us

Association for Computational Linguistics

IEEE Advancing Technology for Humanity

UNIVERSITY OF CAMBRIDGE

Previously, we've worked on...



We Work With

Latest Progress of Conversational AI

Audible now offers live customer service through Alexa devices

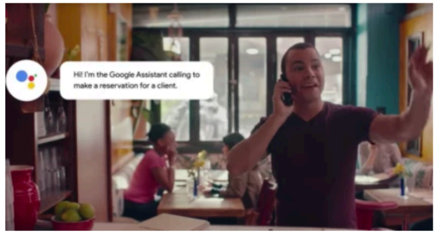
2 months ago **Sarah Perez**

Alexa devices just got a new use case: live customer service help. This morning, Amazon's e-book company Audible announced the first live customer service experience on Alexa devices, activat...

Google's Duplex calls still frequently require human intervention

4 weeks ago **Brian Heater**

When Google launched Duplex with a demo at I/O last year, the audience was left wondering how much of the call was staged. The AI-based reservation booking service seemed almost too impressive to b...

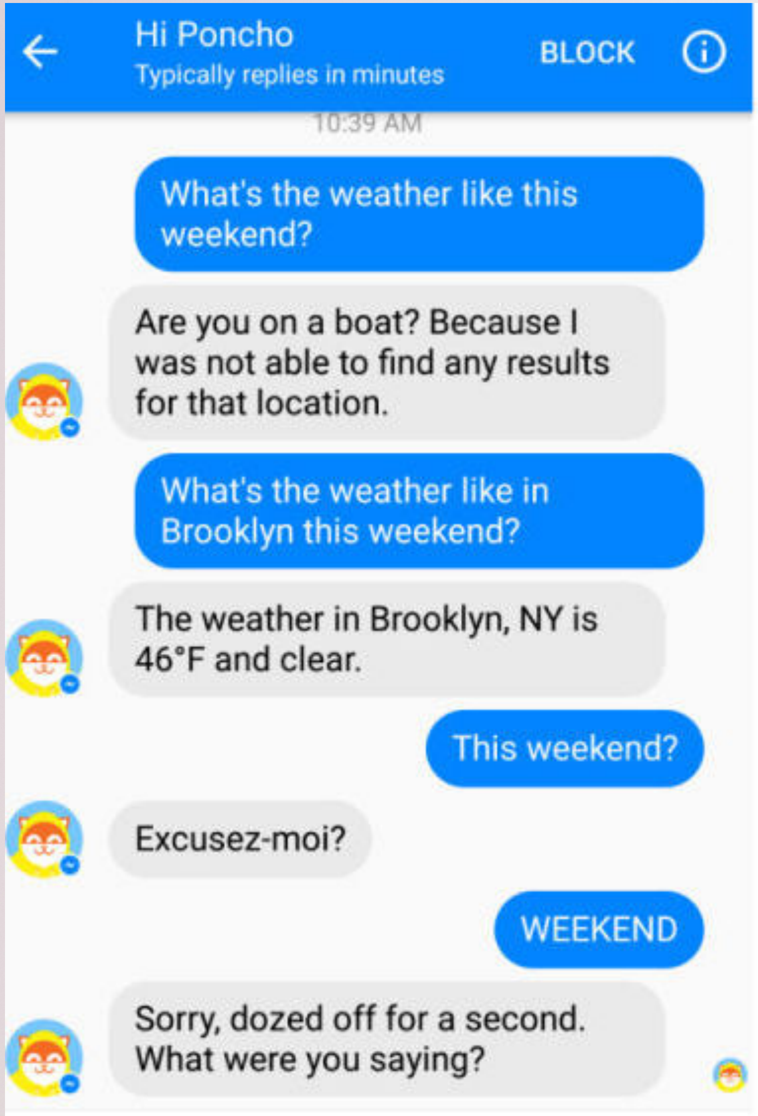


Twilio launches Autopilot to help developers build better bots

8 months ago **Frederic Lardinois**

Bots went through the hype cycle faster than a speeding roller coaster, as the promise of chatting with a computer quickly turned sour. Now, Twilio wants to take another stab at this market with th...

Conversational AI Today



Conversational AI - Where Are We?

RASA

Bot Framework

amazon web services Lex

IBM WATSON

clinc

twilio

Dialogflow

interactions

KISS FM
"Alexa, open KISS FM"
11

Question of the Day
"Alexa, ask Question of the Day."
684

4AFart
"Alexa, Ask For A Fart"
61

Jurassic Bark
"Alexa, ask Jurassic Bark to bark"
13

Xiaoice

Poncho

Swelly

AVA (Autodesk)

What Are the Challenges?

Developers have to anticipate all phrases that could be used to invoke a command

- An alarm for 8.30am.
- Set an alarm for half eight.
- Turn on my 8.30am alarm.
- Wake me up in six hours.
- Alarm at 8.30am, please.

…

- Extra hot.
- Very spice.
- Burning hot.
- The hottest one.
- Extra spicy please.

…

What Are the Challenges?

Developers have to anticipate all phrases that could be used to invoke a command

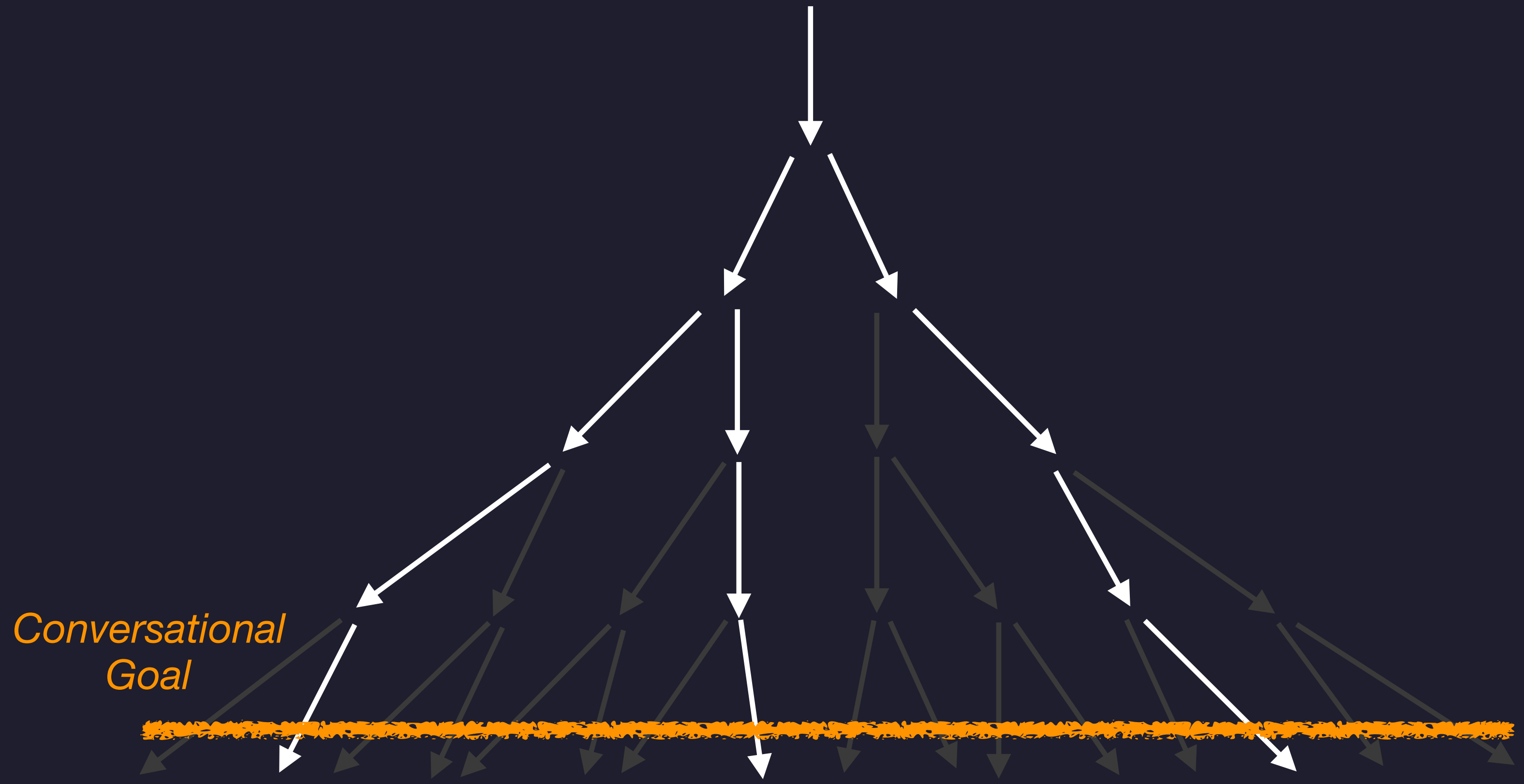
- An alarm for 8.30am.
- Set an alarm for half eight.
- Turn on my 8.30am alarm.
- Wake me up in six hours.
- Alarm at 8.30am, please.

...

- Extra hot.
- Very spice.
- Burning hot.
- The hottest one.
- Extra spicy please.

...

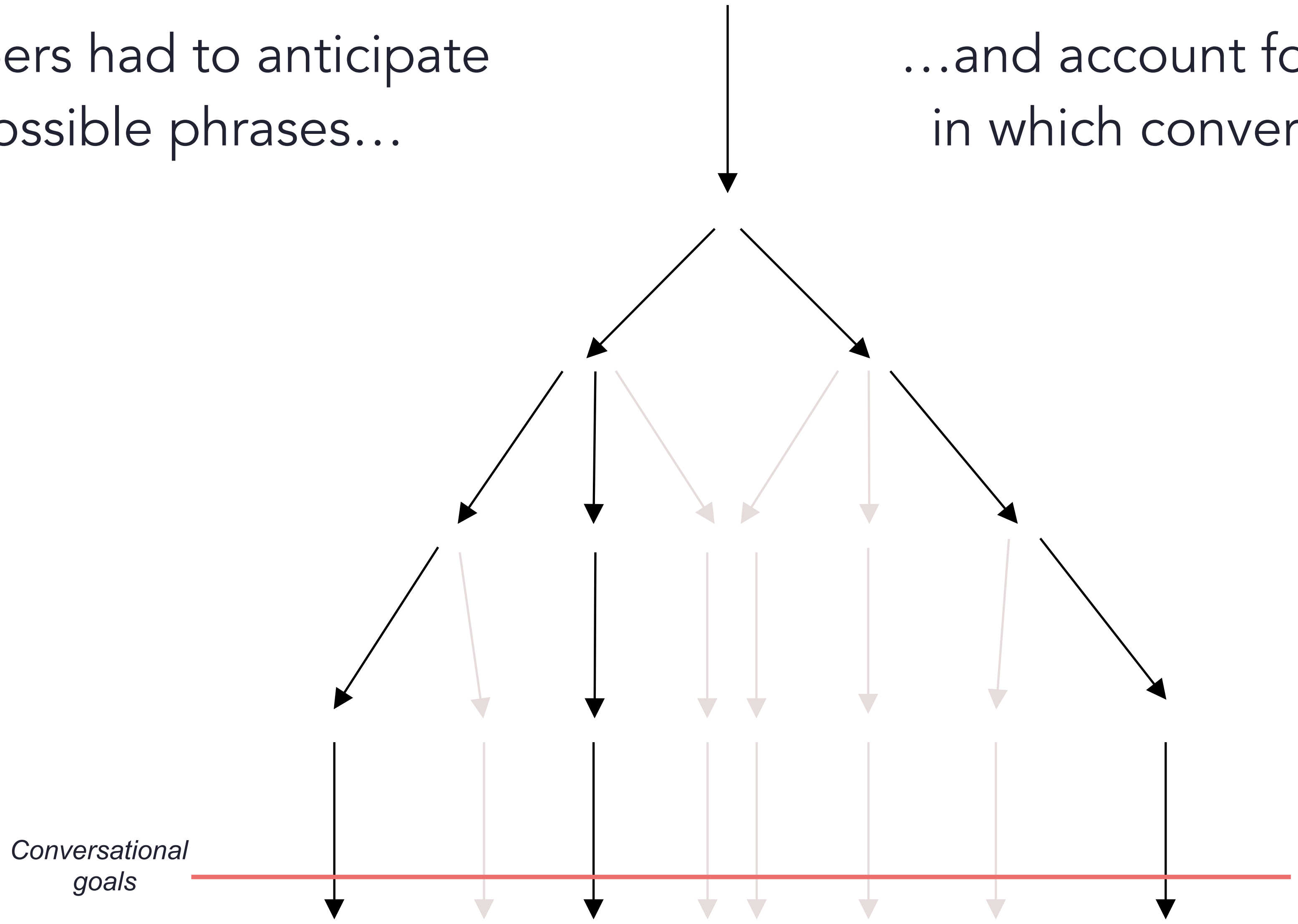
... and they have to account for all the conversational scenarios that their users might try to follow



Non AI Automation is Linear

Developers had to anticipate all possible phrases…

…and account for all the directions in which conversation could flow



We Imagine Customer Scenarios to be Simple (Linear)

What time would you like your booking for?

5pm

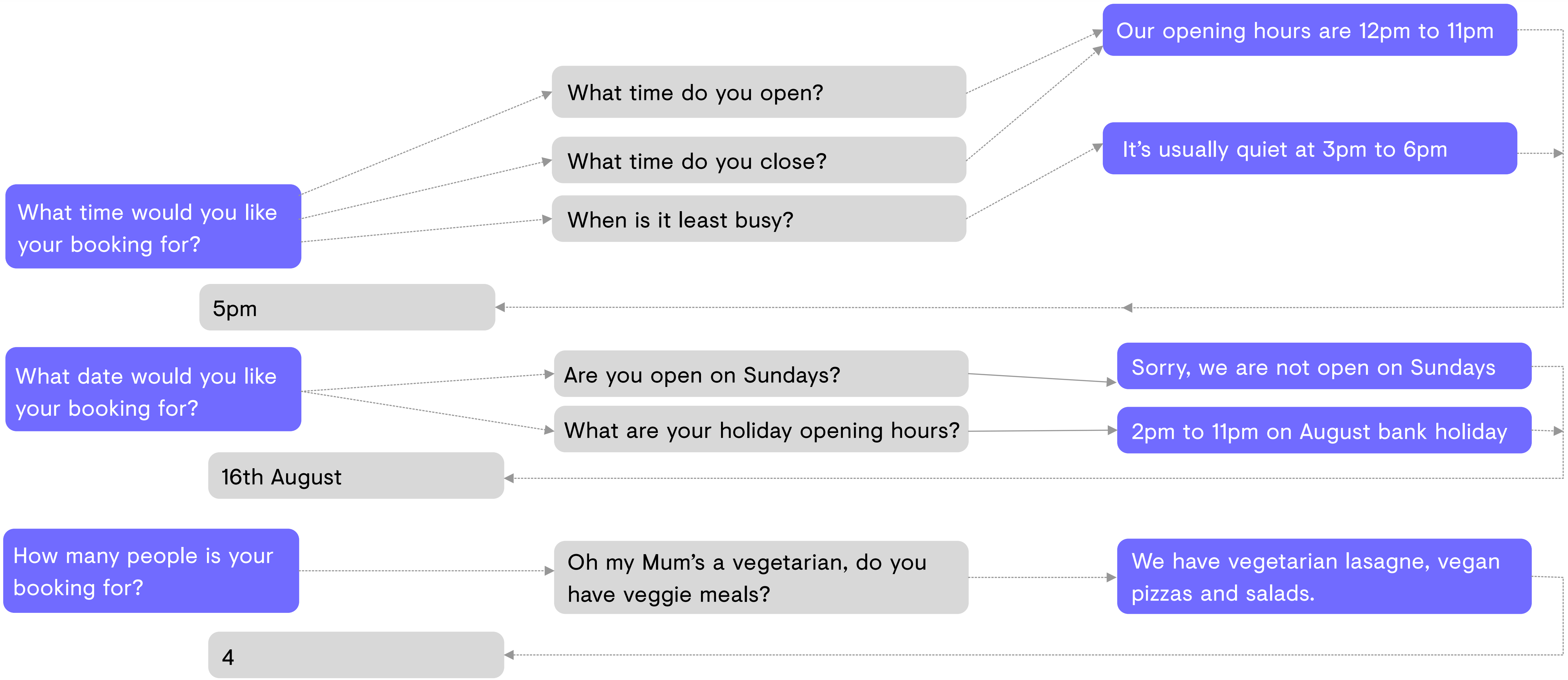
What date would you like your booking for?

16th August

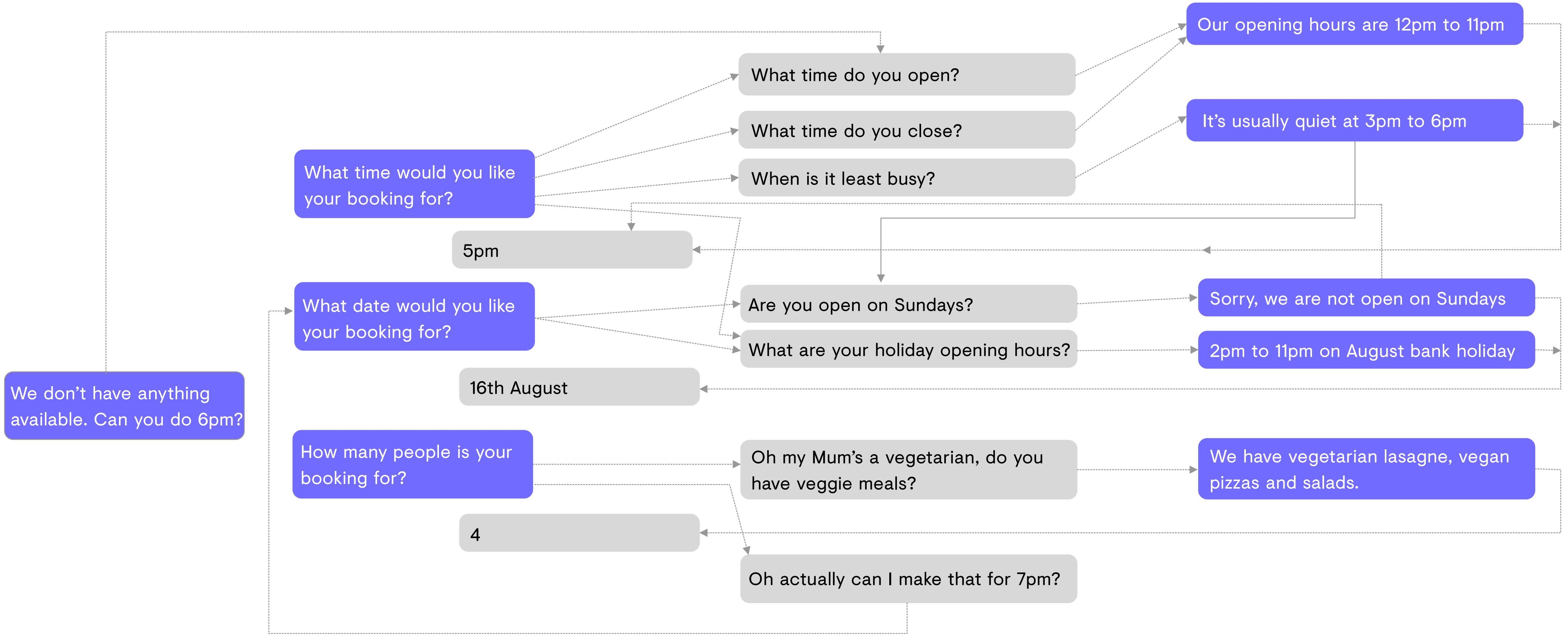
How many people is your booking for?

4

Customers Would Like to Ask Questions

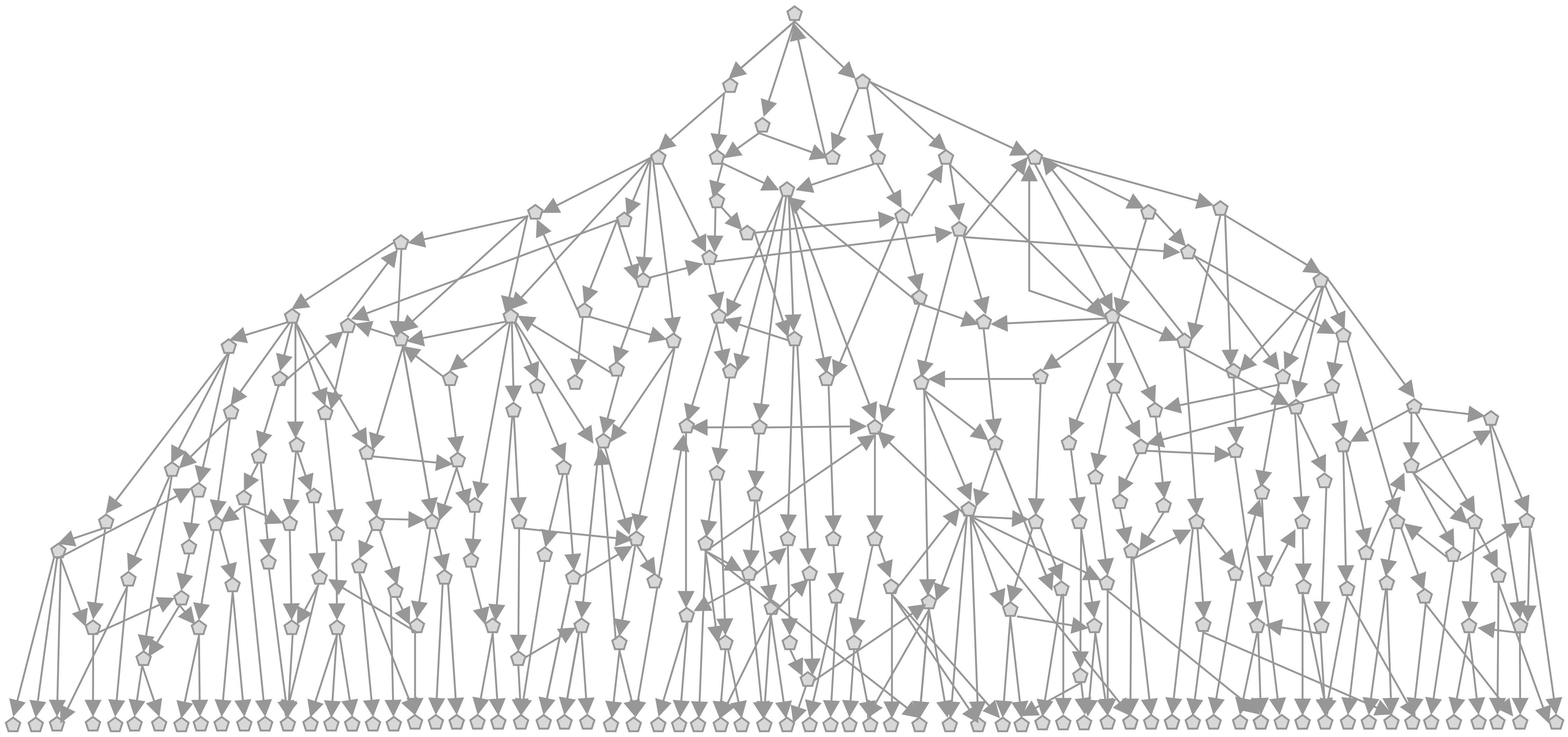


Conversations Move Back and Forth



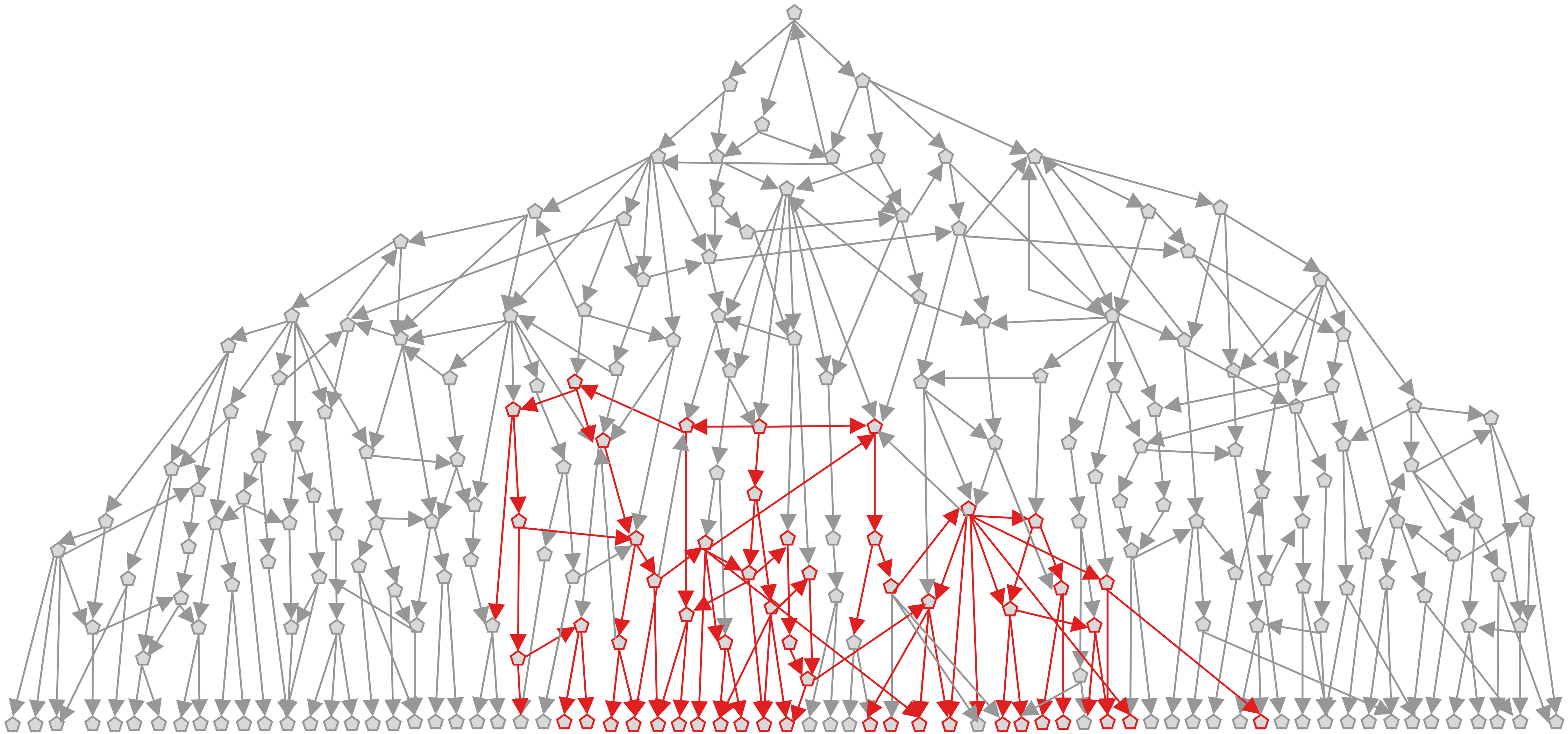
Linear Conversations Get Complicated

But in real life, customers change their mind or give information in a different order



So Updating Them is Over-Complicated and Expensive

But in real life, customers change their mind or give information in a different order



PolyAI Content Programming

Logic trees (linear) = not scalable

Content Programming (non-linear) = scalable

Creating Task-based Dialogue Systems

Convincing Application

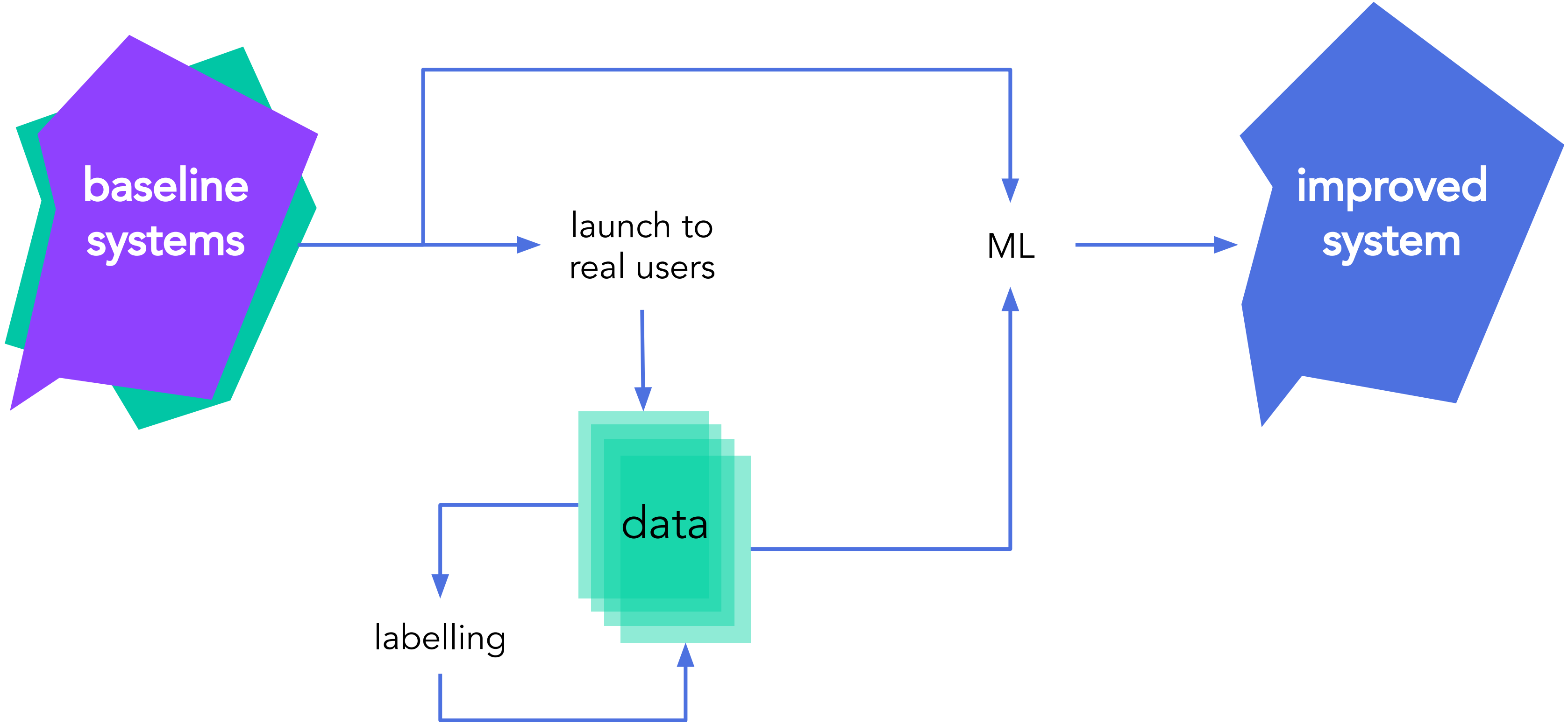
solves a real problem

Meaningful Evaluation

can measure progress

Annotated Data

is machine-learnable


baseline systems
launch to real users
ML
improved system
data
labelling

how do we get a baseline system?

how can we minimise reliance on annotated data?

how can we scale better?
(skills, domains, languages…)

by using large pre-trained models that encapsulate knowledge of conversational response

Pre-training in NLP

- recent trend to pre-train large models of language, then fine-tune

  BERT, ELMo, GPT etc.

- uses unlabelled text + unsupervised objective

  same idea as cbow, skip gram, skip thought etc.

- learns general representations of text, useful for downstream tasks

PolyAI Conversational Datasets

Reddit	OpenSubtitles	AmazonQA
3.7 billion comments from online discussions on many topics	over 400 million lines of subtitles from movies and TV	over 3.6 million product question-answer pairs
↓	↓	↓
727 million examples	316 million examples	3.6 million examples

github.com/PolyAI-LDN/conversational-datasets

Public Conversational Datasets

	~ Turns	Annotations
DSTC 2&3	10^4	response, ASR, SLU
MultiWoz	10^5	response, NLU
DSTC7 Reddit	10^6	response, entities
DSTC7 Ubuntu	10^6	response
PolyAI AmazonQA	10^6	product, response
PolyAI OpenSubtitles	10^8	'response'
PolyAI Reddit	$\mathbf{10^9}$	**response**

Next word prediction

The launch of India's second lunar mission has been → **halted**

apple
called
halted
celebrate
passport
...

Masked word prediction

The launch of ■ 's second lunar mission has been ???
less than an hour before the scheduled blast- ■ ,
due to a ■ problem.

apple
called
halted
celebrate
passport
...

Response Selection

Any recommendations for short trips from Singapore?

→

It doesn't feel like July.
That type of music isn't really my cup of tea.
Bintan is just a quick ferry trip away.
You have to try the vegetarian Haggis!
I'd do a short trip to Paris.
...

Response Selection

- large conversational datasets

Language Modelling

- large text datasets

Response Selection

- large conversational datasets
- representations encode conversational cues

Language Modelling

- large text datasets
- representations encode word/phrase/sentence cues

Response Selection

- large conversational datasets
- representations encode conversational cues
- encodes full sentences

Language Modelling

- large text datasets
- representations encode word/phrase/sentence cues
- encodes words contextually

Response Selection

- large conversational datasets
- representations encode conversational cues
- encodes full sentences
- directly applicable to retrieval-based dialogue

Language Modelling

- large text datasets
- representations encode word/phrase/sentence cues
- encodes words contextually
- maybe applicable to generation/scoring

a lot of the power of neural techniques is finding good embeddings / encodings

- so learn encoder model on large conversational data

- then use various tricks and small models on the learned vector space for domain specific tasks

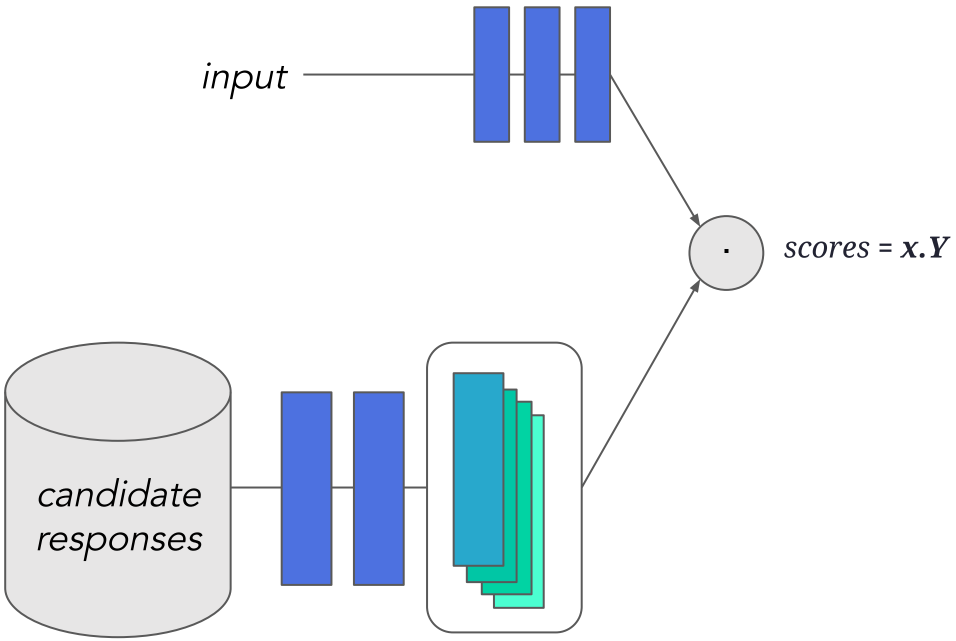
Dual Encoders for Response Selection

dual encoder dot product model

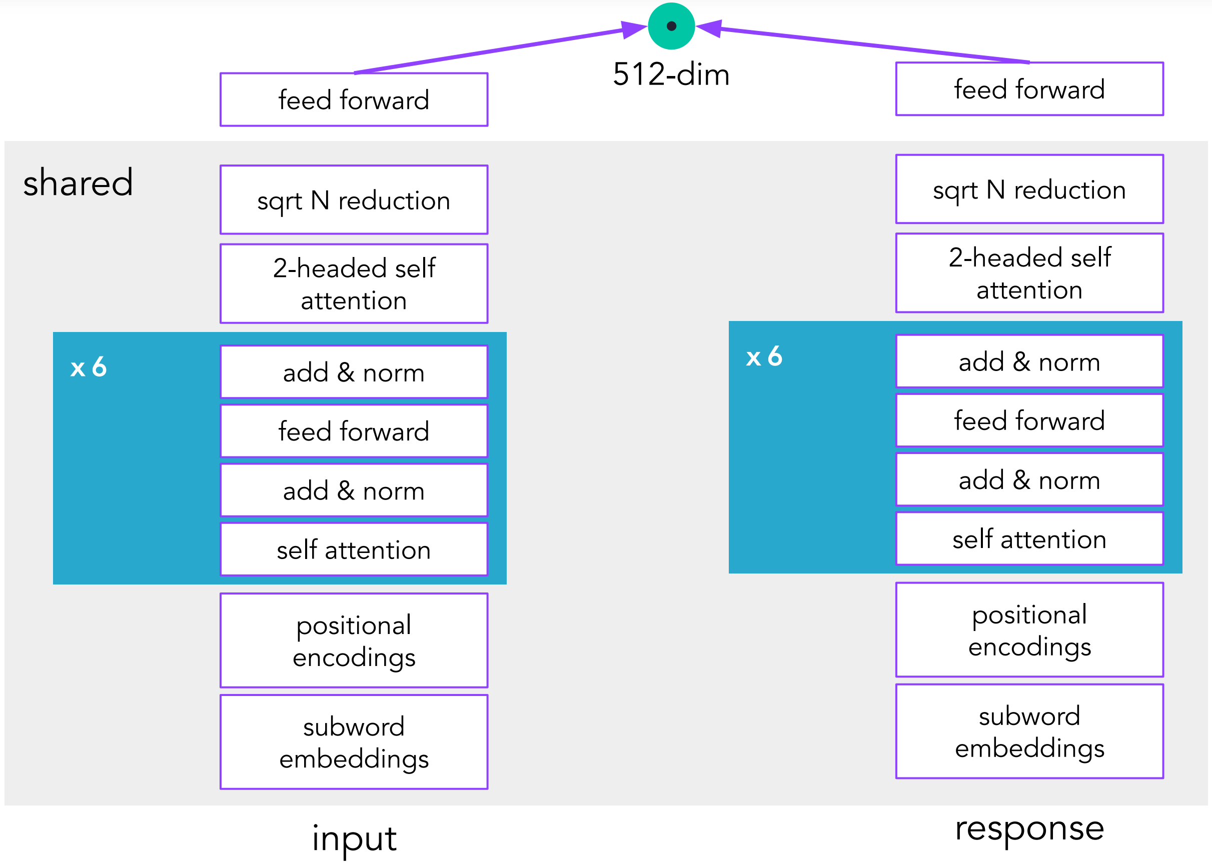
- gmail smart reply
- universal sentence encoder

trained to give a high score for the response found in the data, low score for random responses

final score of an input and response is a dot-product of two vectors



PolyAI Encoder



network encodes a batch of inputs to vectors:

$$x_1 \quad x_2 \quad \dots \quad x_N$$

and responses to vectors:

$$y_1 \quad y_2 \quad \dots \quad y_N$$

$x_1.y_1$	$x_1.y_2$	$x_1.y_3$	$x_1.y_4$	$x_1.y_5$
$x_2.y_1$	$x_2.y_2$	$x_2.y_3$	$x_2.y_4$	$x_2.y_5$
$x_3.y_1$	$x_3.y_2$	$x_3.y_3$	$x_3.y_4$	$x_3.y_5$
$x_4.y_1$	$x_4.y_2$	$x_4.y_3$	$x_4.y_4$	$x_4.y_5$
$x_5.y_1$	$x_5.y_2$	$x_5.y_3$	$x_5.y_4$	$x_5.y_5$

the *N* x *N* matrix of all scores is a fast matrix product.

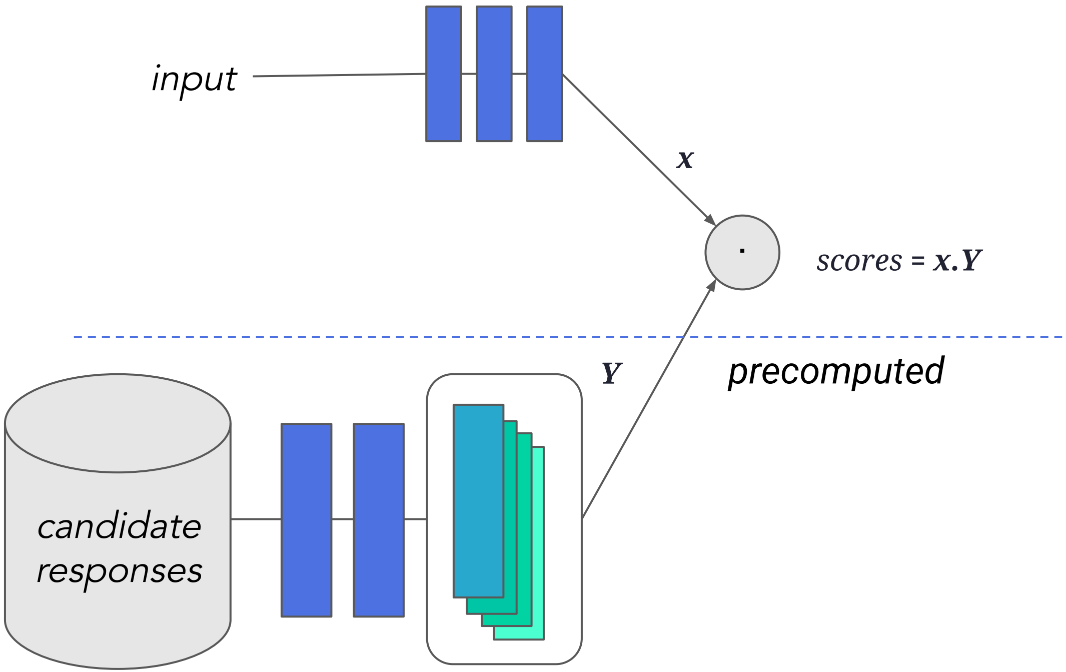
large improvement in 1 of 100 ranking accuracy over binary classification.

$x_1.y_1$	$x_1.y_2$	$x_1.y_3$	$x_1.y_4$	$x_1.y_5$
$x_2.y_1$	$x_2.y_2$	$x_2.y_3$	$x_2.y_4$	$x_2.y_5$
$x_3.y_1$	$x_3.y_2$	$x_3.y_3$	$x_3.y_4$	$x_3.y_5$
$x_4.y_1$	$x_4.y_2$	$x_4.y_3$	$x_4.y_4$	$x_4.y_5$
$x_5.y_1$	$x_5.y_2$	$x_5.y_3$	$x_5.y_4$	$x_5.y_5$

Precomputation for dot product model

the representations of the candidates $\boldsymbol{Y}$ can be precomputed

approximate nearest neighbor search can speed up the top N search



at inference, a user query has N words, there are M responses with N_R words each

- dot product model
  - $O(N)$ to encode input to vector space
  - $O(\log M)$ to find top scoring response with approximate search

at inference, a user query has N words, there are M responses with N_R words each

- dot product model
  - $O(N)$ to encode input to vector space
  - $O(\log M)$ to find top scoring response with approximate search
- general sequence model (e.g. BERT next sentence scoring)
  - $O(M(N + N_R))$ to score all responses
  - $O(M)$ to find top response

1-of-100 accuracy

how often the correct response is ranked top vs 99 random

PolyAI Encoder

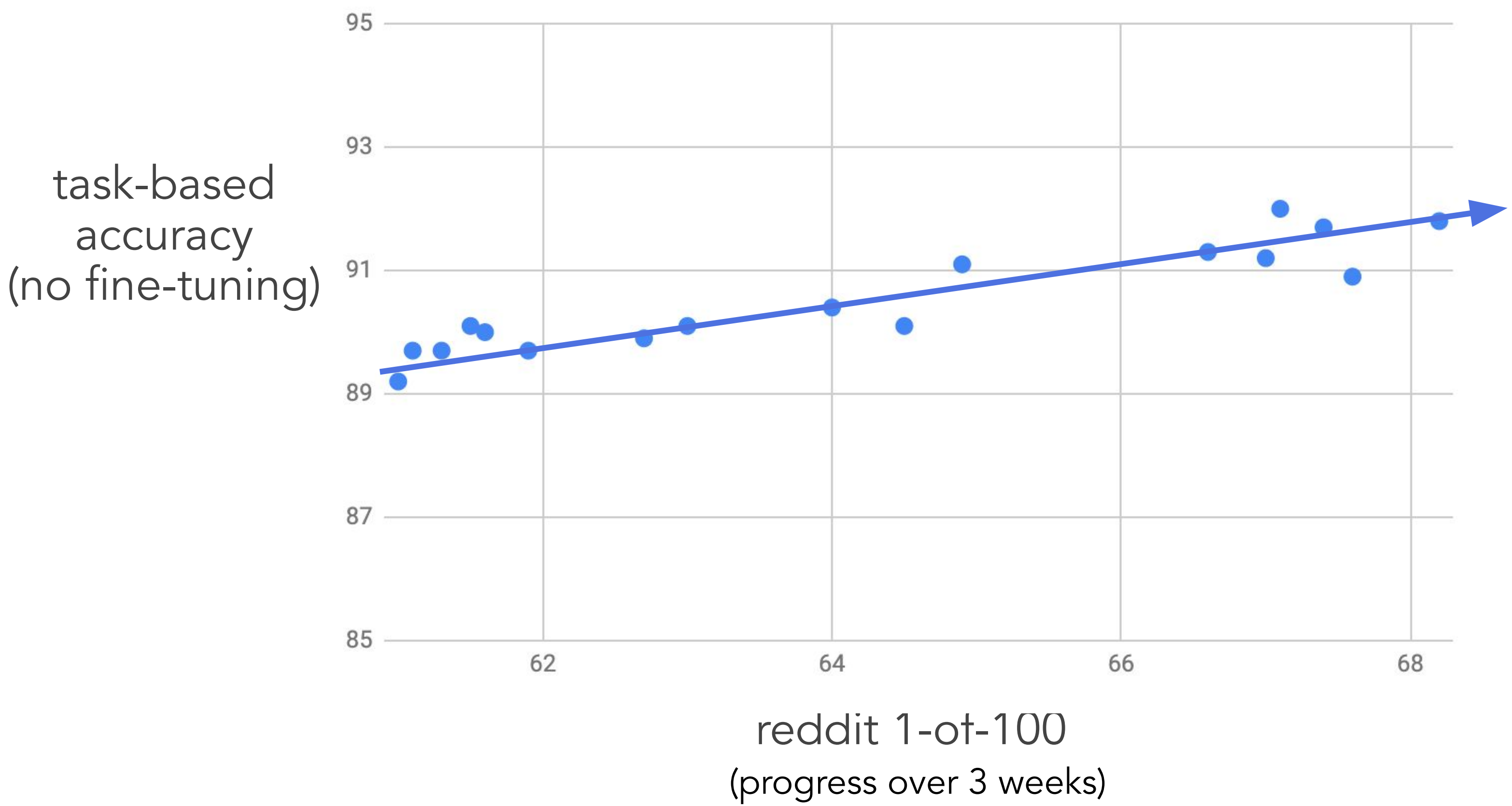
		reddit 1-of-100 accuracy
keyword-based	TF-IDF	26.7%
	BM25	27.6%
MAP dot product models	ELMo	19.3%
	BERT	24.5%
	USE	40.8%
	USE_QA	46.3%
	BERT dot-product model	55.0%
PolyAI Encoders	n-grams	61.3%
	subwords	**68.2%**

PolyAI Encoder

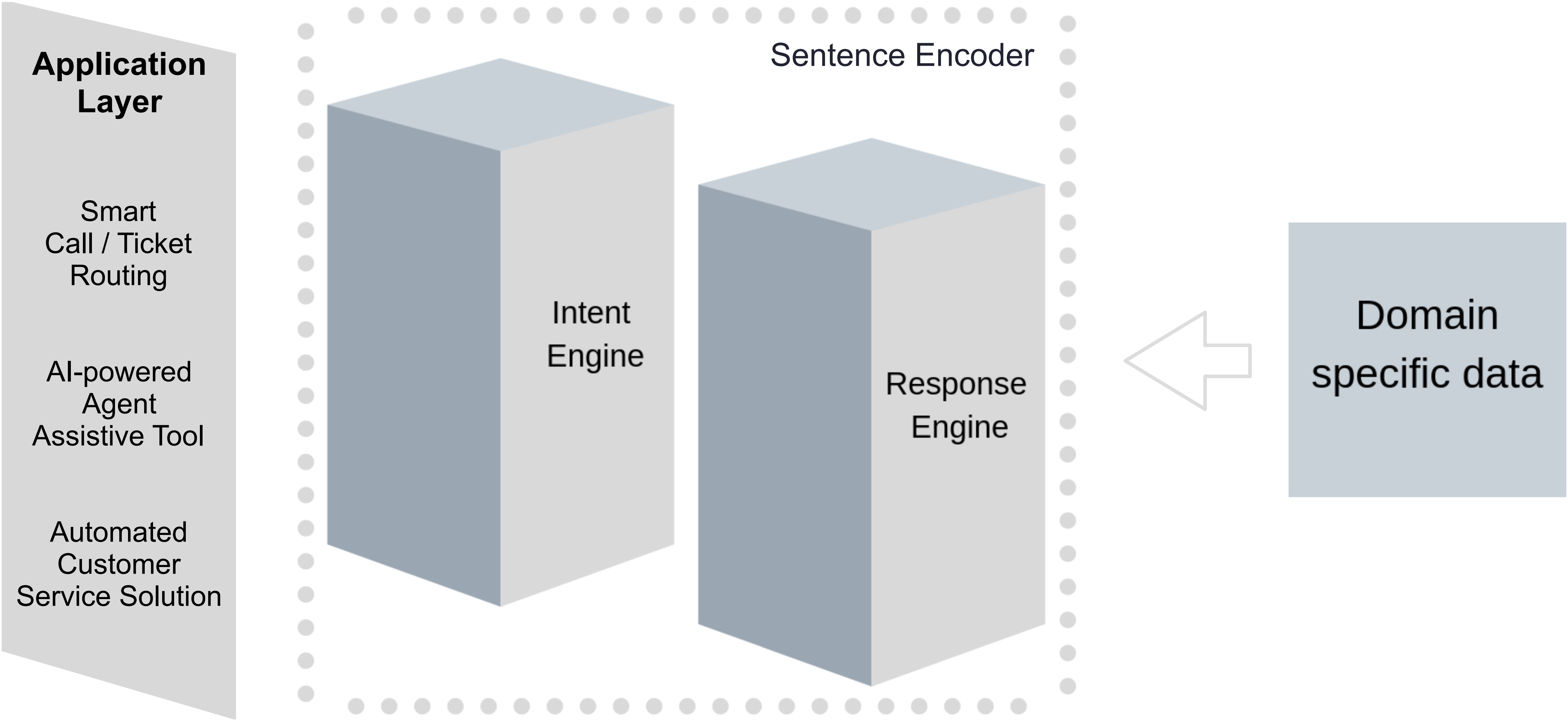
resource-constrained optimization:
pick the best model after training 18 hours on 12 GPUs

- fast ML engineering cycle, rapid progress
- we own the whole training pipeline

- training costs under $100
- model runs fine on CPU
- final model is 40MB


task-based
accuracy
(no fine-tuning)
95
93
91
89
87
85
62
64
66
68
reddit 1-of-100
(progress over 3 weeks)

Smart, Accurate Customer Service Solutions



intent classification

Intent Classification

initiate-booking

can i make a booking
can i reserve a table
okay i want to book a table for tonight

cancel-booking

cancel it
i don't want the table anymore

restart

let's start over
forget this

Intent Classification

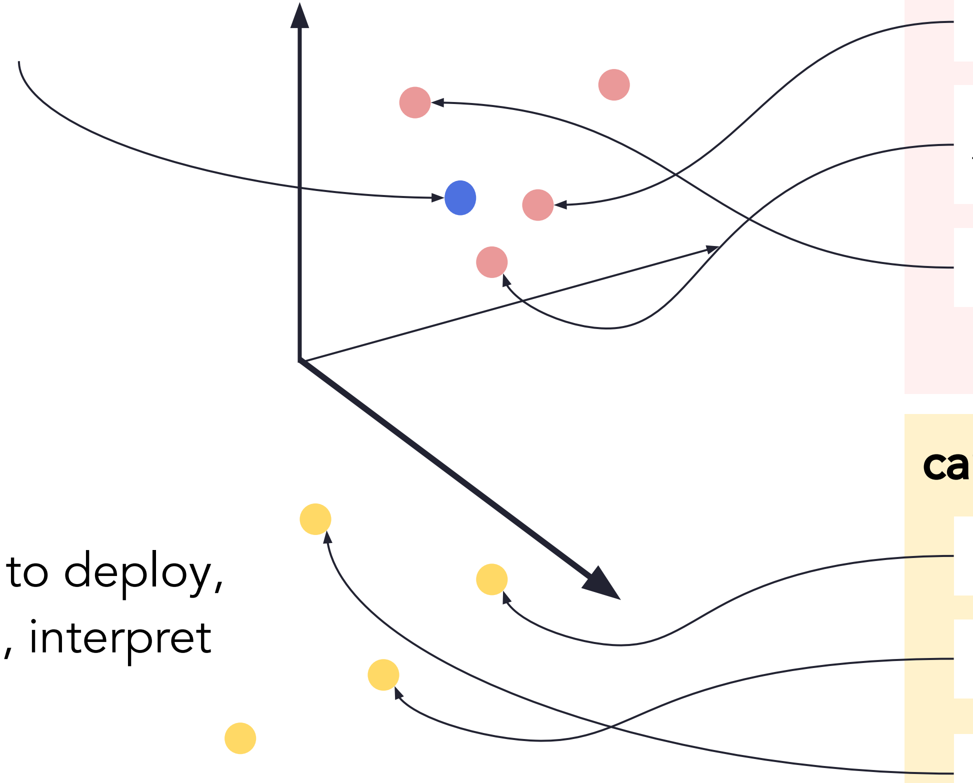
- can train an MLP on top of encoding representation
- can jointly fine-tune the encoding parameters

- can treat similarity in encoding space as as a kernel
    - SVM (more interpretable, encoding-agnostic)

Why is Out-of-the-Box Performance Important?

$(\text{Intent accuracy\%})^{\wedge\text{number of turns}}$ = success rate of automation

$(80\%)^{\wedge 2\text{ turns}}$ = 64% success rate can i make a reservation



initiate-booking

- can i make a booking
- i want to reserve a table for tonight
- okay let's book
- ...

cancel-booking

- actually forget the booking
- i don't want the table anymore
- ok actually i don't want the table
- ...

\+ simple to deploy, control, interpret

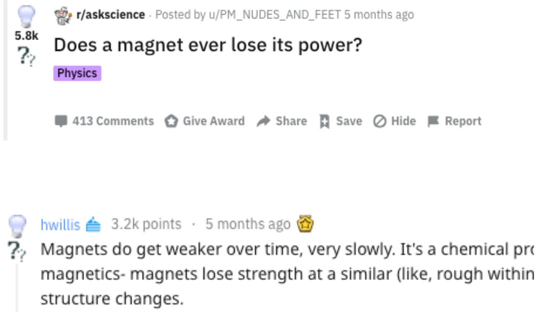
Encoder Has Been Trained on Billions of Conversations

Jessep: You want answers?!

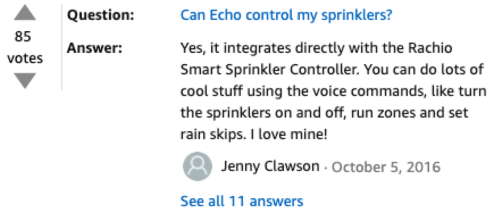
Kaffee: I want the truth!

Jessep: You can't handle the truth!

40 million lines of subtitles from film and TV

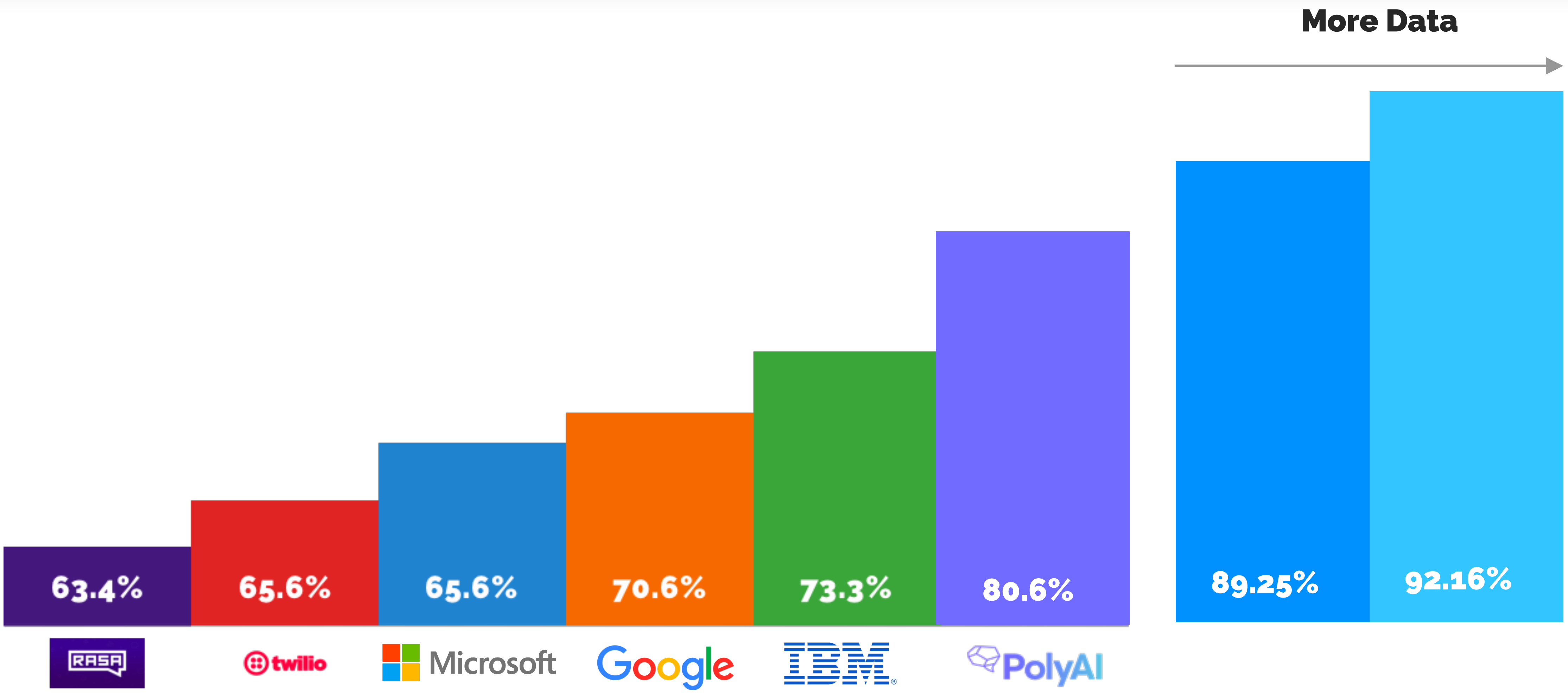


3.7 billion comments from online forums



3.6 million question/answer pairs from FAQ

Poly AI Encoder: Better Performance Out-of-the-Box

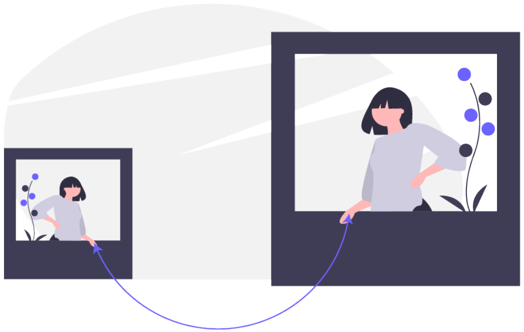


https://poly-ai.com/blog/were-building-the-most-accurate-intent-detector-on-the-market-6

The PolyAI Encoder: Understanding as a Service

Needs less data
for deployment

Empower customers
to speak naturally



Lightweight:
MBs instead of GBs

Intent Engine w/ Shared Representations

- I want an extra spicy one.
- Extra spicy chicken please.
- Extra hot, please.
- A pungent one, please.

...

- Quiero uno extra picante.
- Pollo extra picante por favor.
- Extra caliente, por favor.
- Una picante, por favor.

...

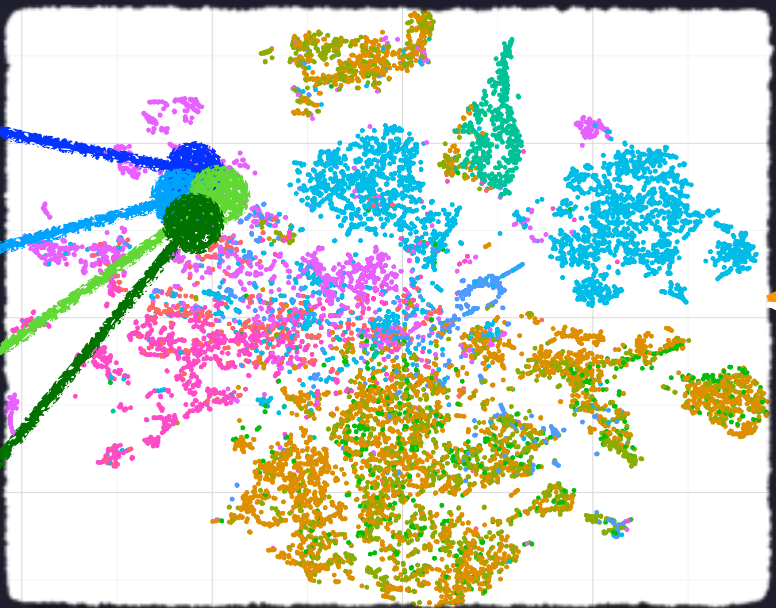
- 나는 여분의 매운 것을 원한다
- 여분의 매운 닭고기주세요
- 아주 더워주세요
- 매운 것, 제발

...

- 我想要特辣的
- 请给我一份特辣烤鸡
- 我吃很辣的
- 来一份大辣

...

Canvas of Semantics



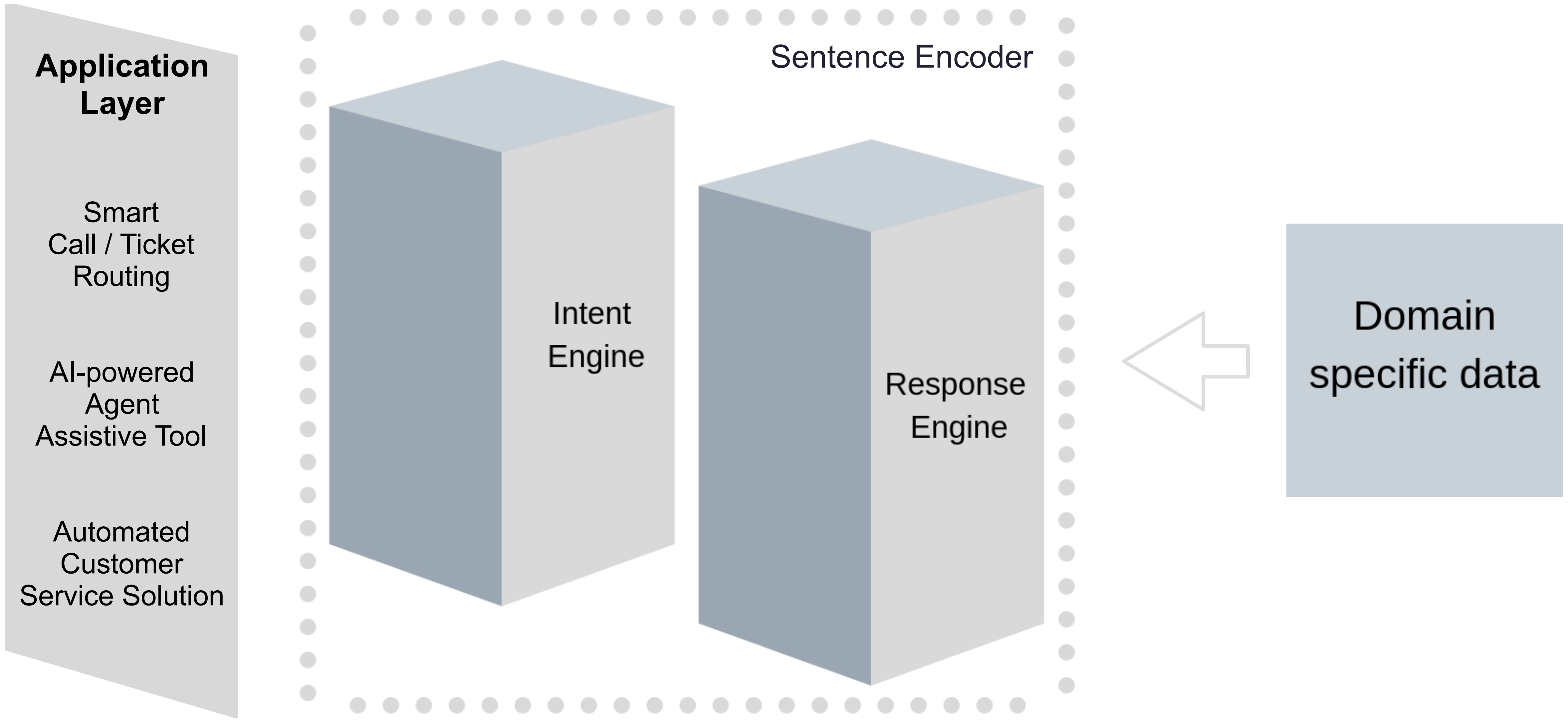
Ultra Spicy

Spicy

Mild

- Leverage public large datasets for generalization
- One model, multiple intents, many languages
- State-of-the-art performance comparing to other solutions

Smart, Accurate Customer Service Solutions



Giving the Best Answer at Every Turn

Conversation between IVA and customer

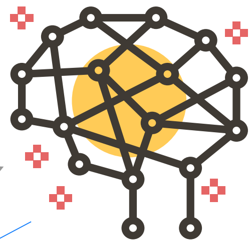
The pump is broken

Okay, let’s look into that for you. Which pump is it?

It’s pump 3

When was the filter last changed?

PolyAI Model



Content

Record pump number

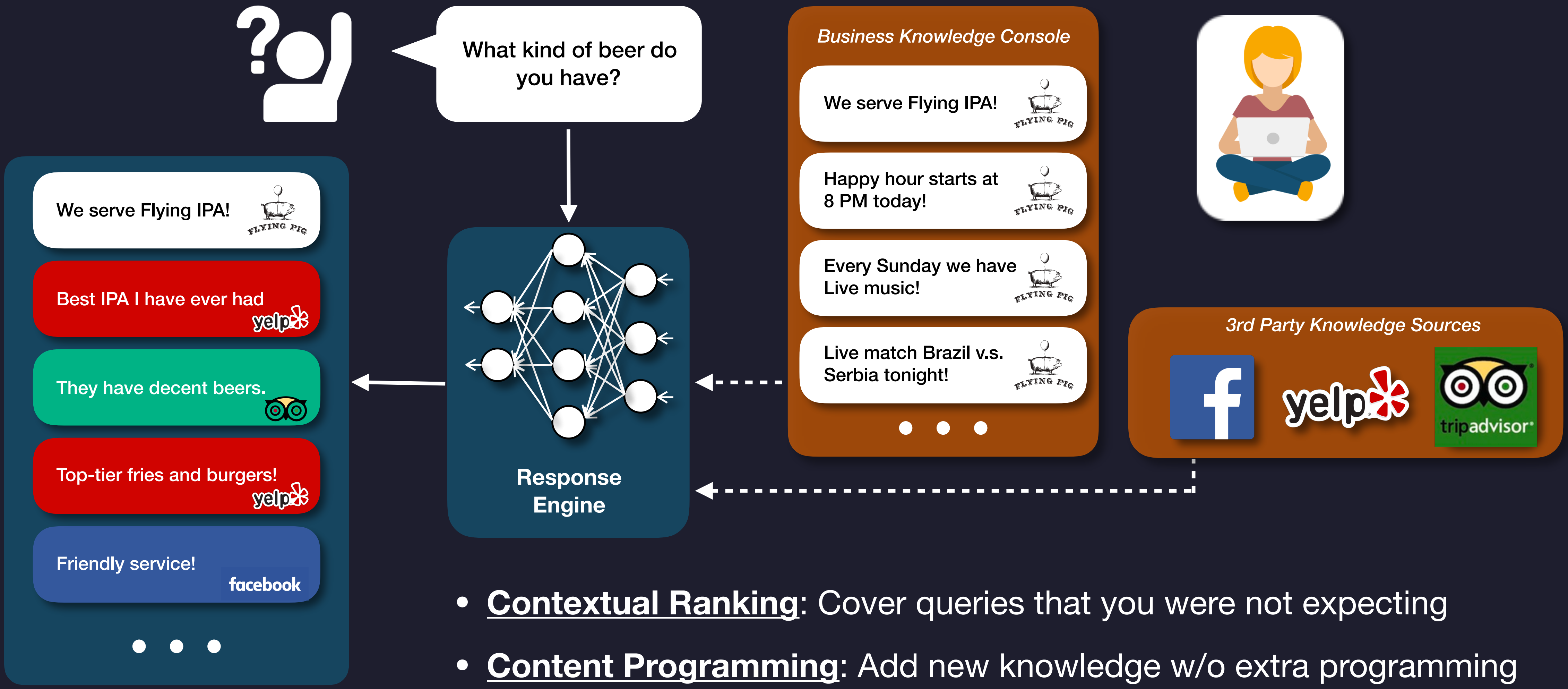
Station has 5 pumps

SVB: 0123456

Check last filter date

Engineer available 16th August

Response Engine



- **Contextual Ranking**: Cover queries that you were not expecting
- **Content Programming**: Add new knowledge w/o extra programming
- **Business Logic Engine**: Build precise business logics on top

restaurant search

DSTC 2 & 3

hello I am looking for a cheap place in the east
> inform(pricerange=cheap, area=east)

sure, what type of food?
> request(food)

i want gastropub food
> inform(food=gastropub)

there are no cheap places serving gastropub in the east.
> inform(name=none, area=east, pricerange=cheap)

how about any pricerange? and i need to know if they have wifi.
> inform(pricerange=dontcare) request(has_wifi)

The King's Arms is a nice place in the east of town serving gastropub food. It has wifi.
> offer(name="The King's Arms", area=east, food=gastropub, has_wifi=true)

DSTC 2 & 3

- explicit semantics forces unnaturally constrained dialogues

    users need to know the ontology

- requires special annotated data, one specialised model per 'slot'

DSTC 2 & 3

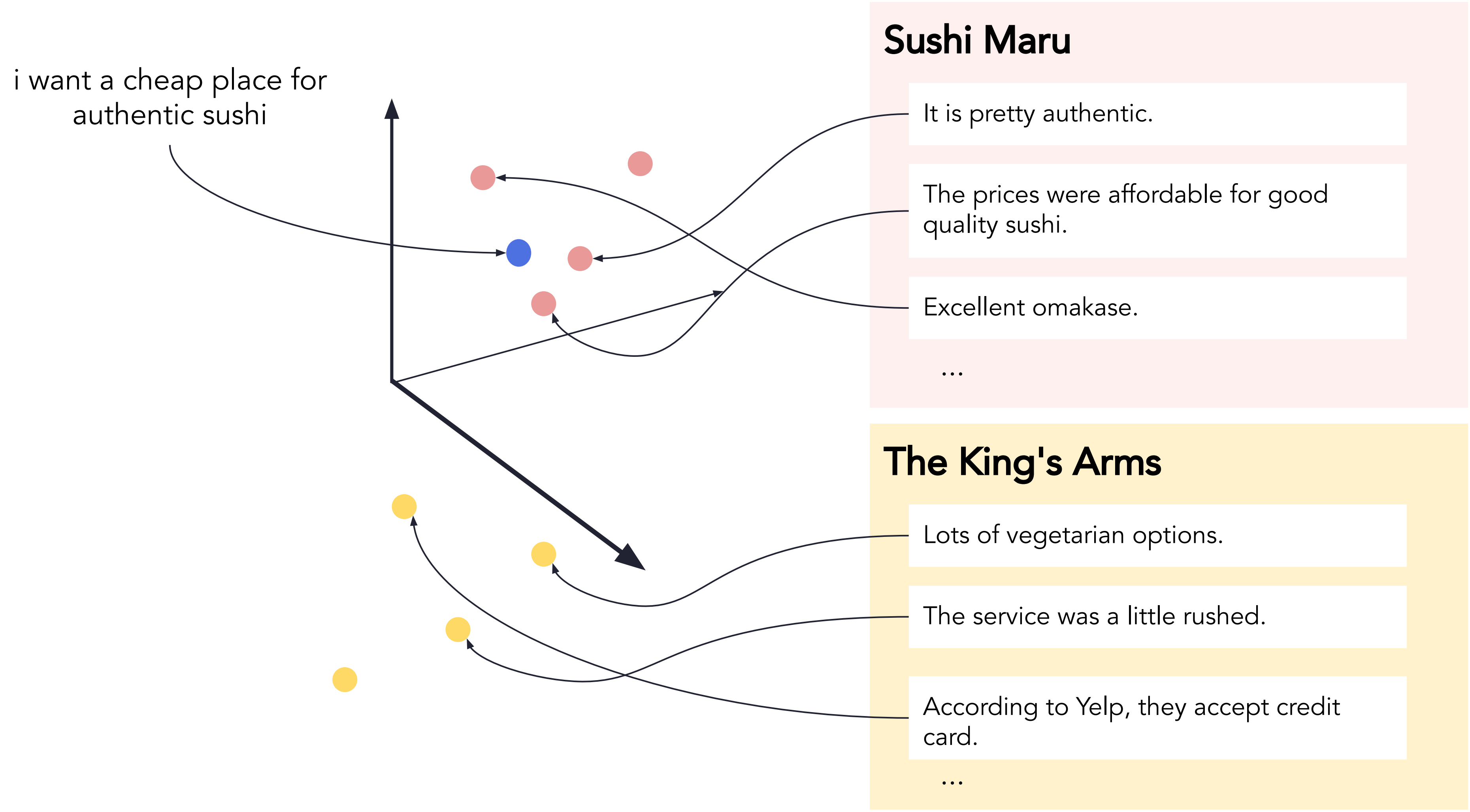
- some slots are necessary

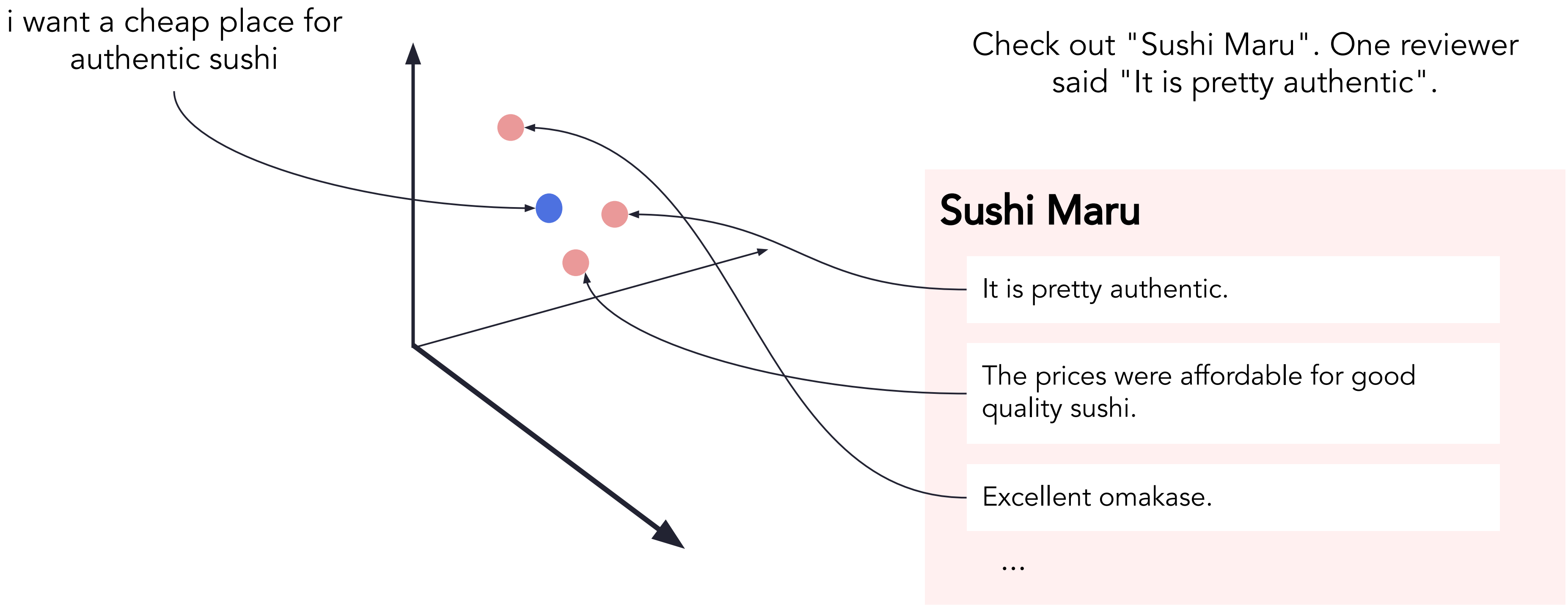
  number of people, booking time, name

- some might not be

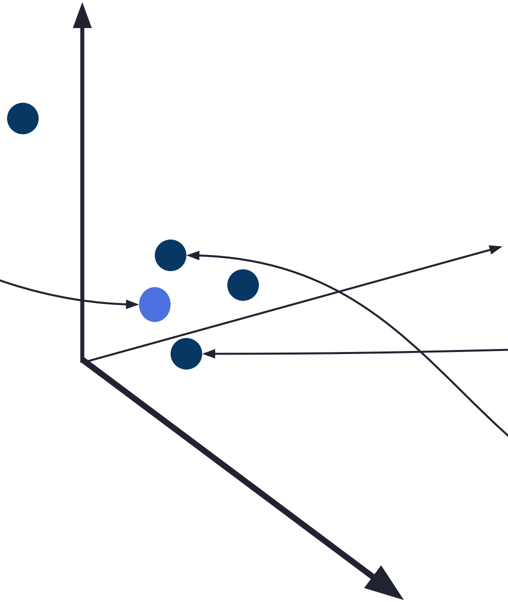
  food, price range, has wifi, has vegetarian, has vegan, serves cocktails….

- use all sentences in all reviews of all restaurants in a city
- treat dialogue as an iterative search
- perform search in implicit vector space learned by encoders


i want a cheap place for authentic sushi
Sushi Maru
It is pretty authentic.
The prices were affordable for good quality sushi.
Excellent omakase.
...
The King's Arms
Lots of vegetarian options.
The service was a little rushed.
According to Yelp, they accept credit card.
...


i want a cheap place for authentic sushi
Check out "Sushi Maru". One reviewer said "It is pretty authentic".
Sushi Maru
It is pretty authentic.
The prices were affordable for good quality sushi.
Excellent omakase.
...

does it have a view of
the ocean?



Sotto il Mare

our table had an excellent view

…

The Elephant House

Restaurant in Old Town. *Brasseries, Coffee & Tea, Food, British, Sandwiches, Cafes, and Scottish*

where did JK rowling write Harry Potter

“We got to see the spot where it is said JK Rowling wrote Harry Potter.”

“So apparently JK Rowling often wrote Harry Potter here.”

“After all, if JK Rowling was inspired maybe it can rub off?”

The Elephant House

Restaurant in Old Town. *Brasseries, Coffee & Tea, Food, British, Sandwiches, Cafes, and Scottish*

can i book a table for 2

What date and time?

Restaurant search

- entirely powered by a single model, trained on hundreds of millions of examples

- bootstrapped using only raw text representations- restaurants + reviews + facts

- allows more natural search, not bottlenecked by explicit semantics / ontology

Value Extraction

- limit slots to obvious values that the system needs to extract

    booking time & date, your name, number of people

- value extraction can benefit from pre-trained representations

- see our blog post on *Neural language understanding of people's names*

Response Selection for Bootstrapping Dialogue

efficient task tailored to dialogue

smaller cheaper faster models

robust performance on downstream tasks

competitive intent classification

driven by paraphrase collection

powers conversational search

efficient search

reduced dependency on strict ontology

Live Demo


PolyAI
www.polyai.com