

Skill Squatting Attacks on Amazon Alexa

Deepak Kumar
University of Illinois

Eric Hennenfent
University of Illinois

Riccardo Paccagnella
University of Illinois

Joshua Mason
University of Illinois

Paul Murley
University of Illinois

Adam Bates
University of Illinois

Michael Bailey
University of Illinois

Voice Is the New Touch

**56 Million Smart Speaker Sales in 2018
Says Canalys**

**Voice-First Devices Are The Next Big
Thing -- Here's Why**

Speech recognition systems
frequently make errors, even in
normal use

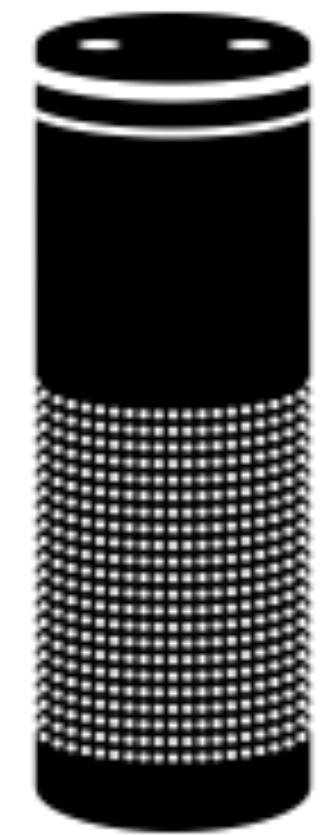


“Alexa, turn on
the living room
lights.”



“Alexa, turn on
the living room
lights.”

“Okay, playing
Living on a Prayer,
Lullaby Edition”





Alexa Skills



Alexa Skills



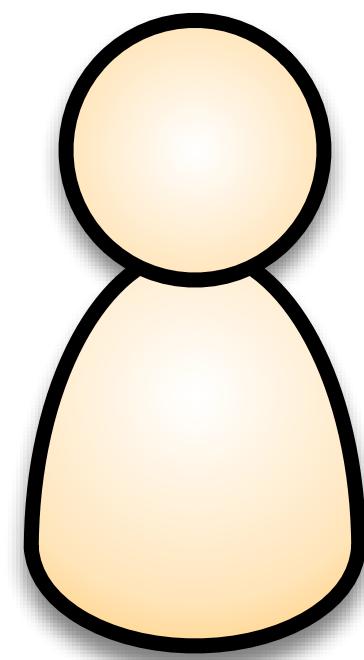
Alexa Skills



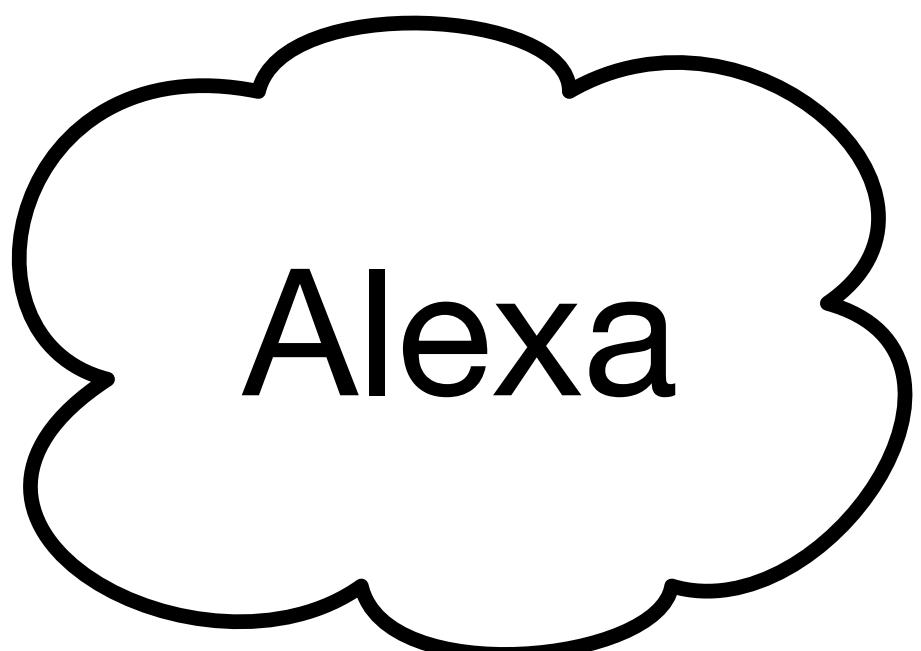
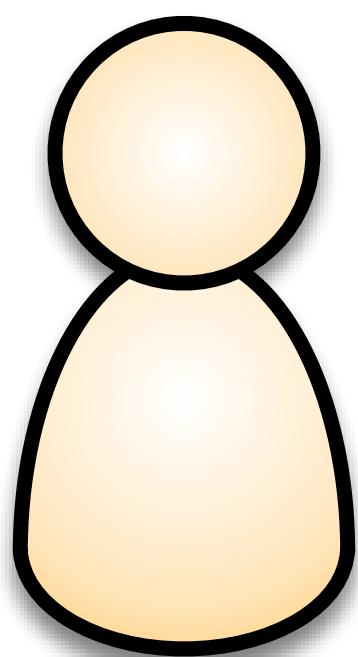
Skills: Apps, but for Alexa

(

“Alexa, tell me some cat facts!”



“Alexa, tell me some cat facts!”



Skills

...
cat forks

cat fast

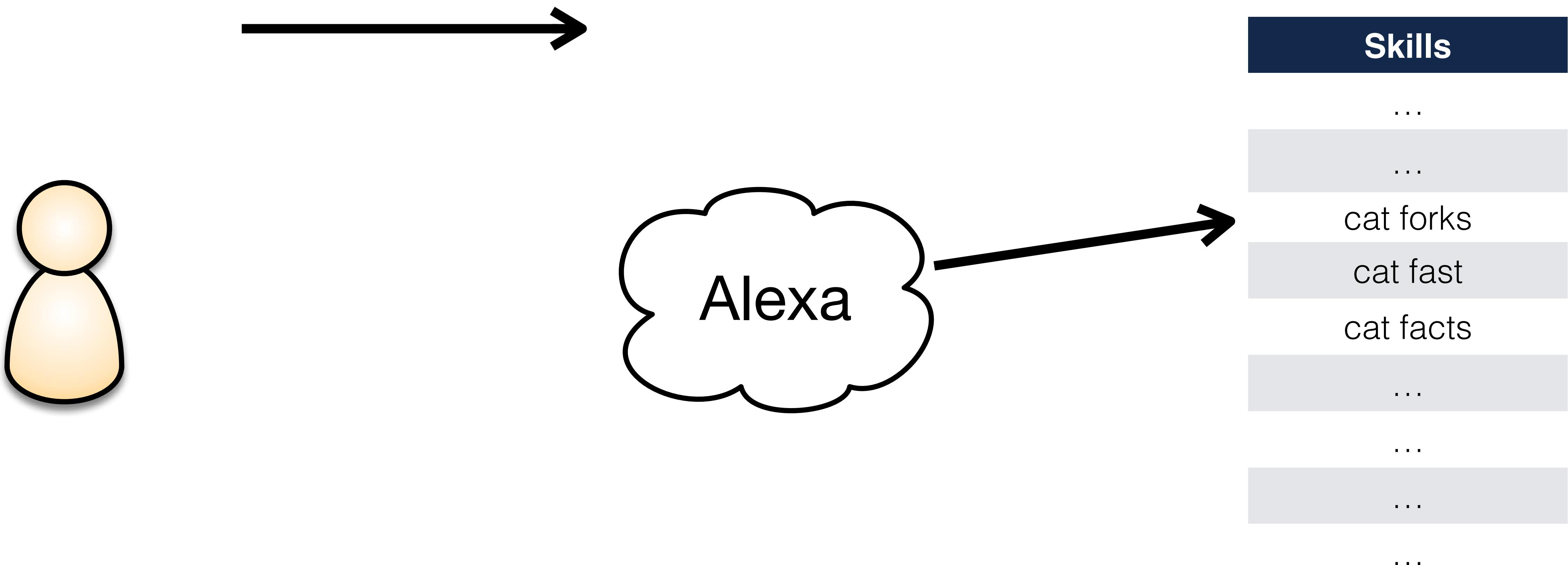
cat facts
...

...

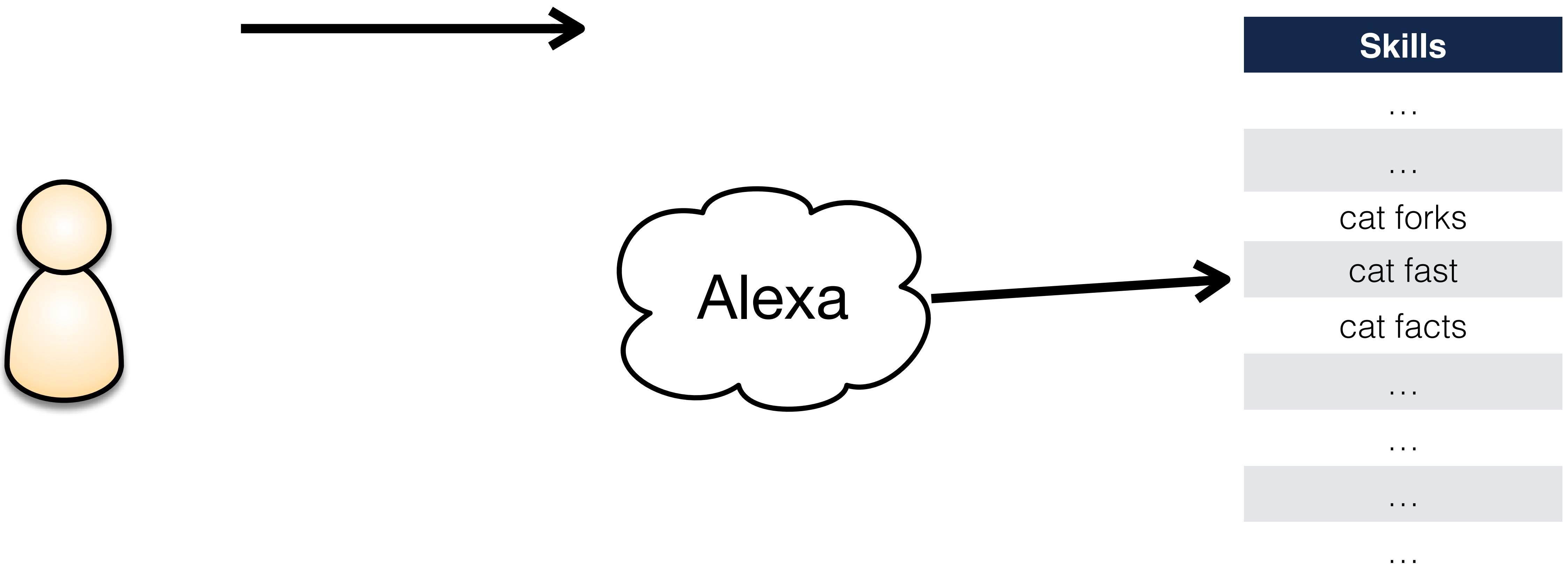
...

...

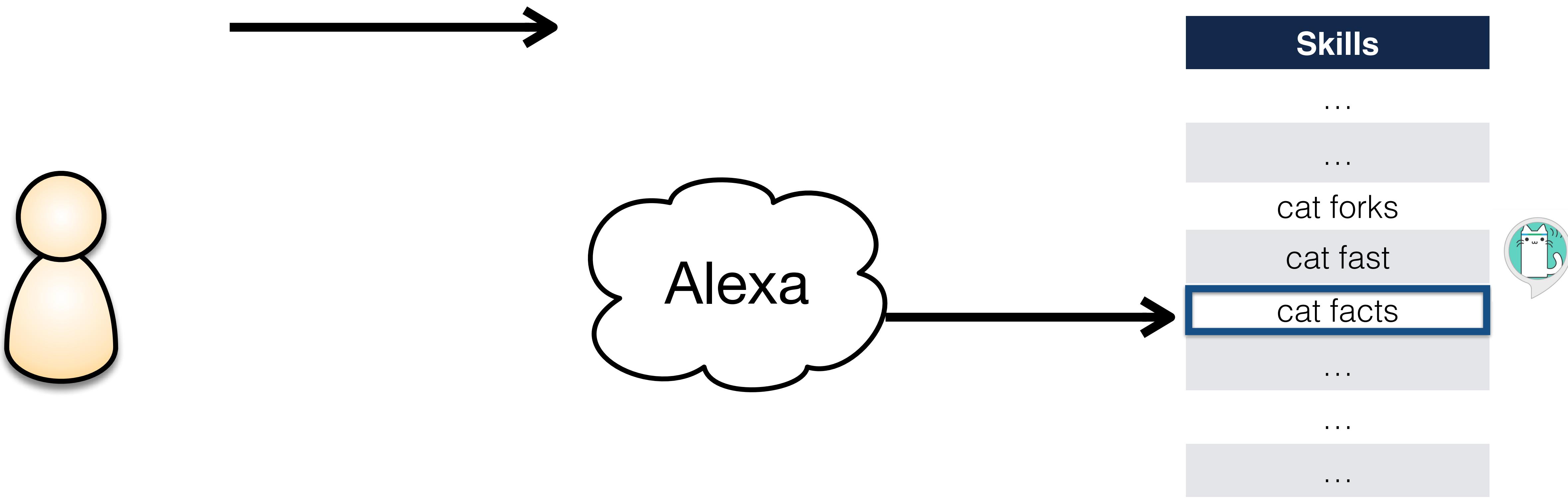
“Alexa, tell me some cat facts!”



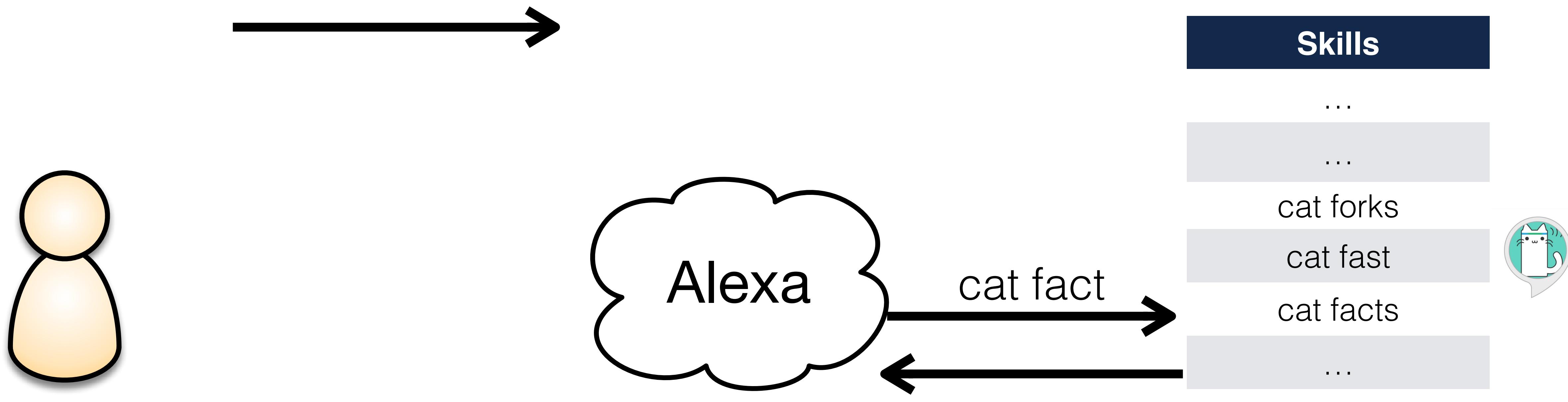
“Alexa, tell me some cat facts!”



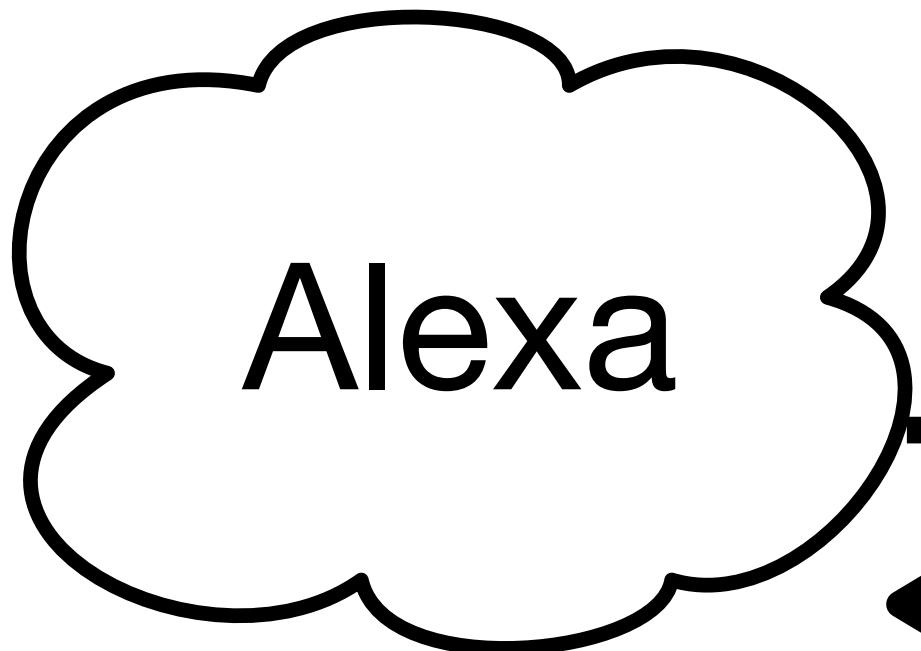
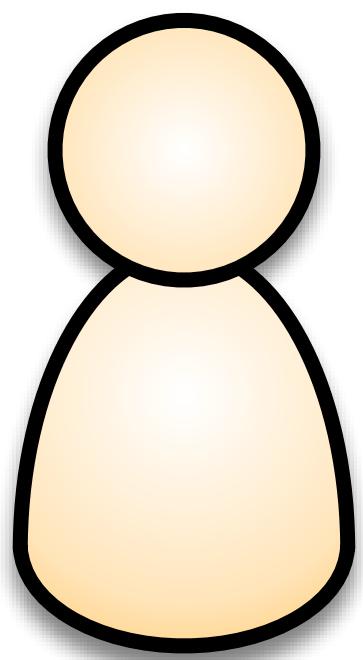
“Alexa, tell me some cat facts!”



“Alexa, tell me some cat facts!”



“Alexa, tell me some cat facts!”



cat fact



Skills

...

...

cat forks

cat fast

cat facts



“A group of cats is called a clowder!”



1. Alexa makes mistakes

1. Alexa makes mistakes
2. Skills are the new apps

1. Alexa makes mistakes
2. Skills are the new apps

What could go wrong?



Fish Geek

Matt Mitchell

"Alexa ask Fish Geek to tell me a fact"

"Alexa ask Fish Geek to tell me trivia"



Phish Geek

EP

"Alexa, open Phish Geek"

*"Alexa, launch Phish Geek and
tell me a fact"*





Rachel

It just gives "fish" facts, not "Phish" facts

February 8, 2017

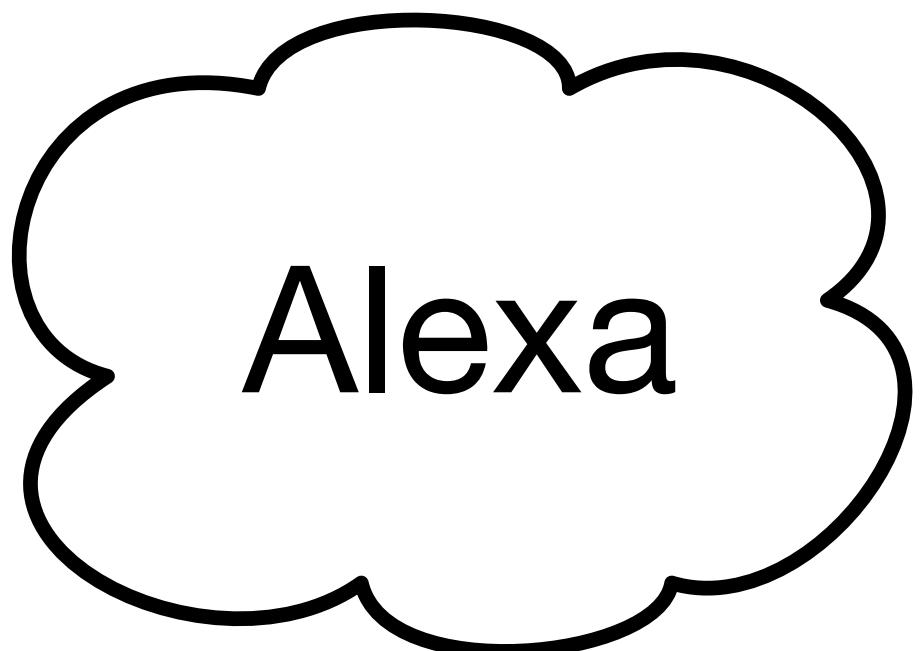
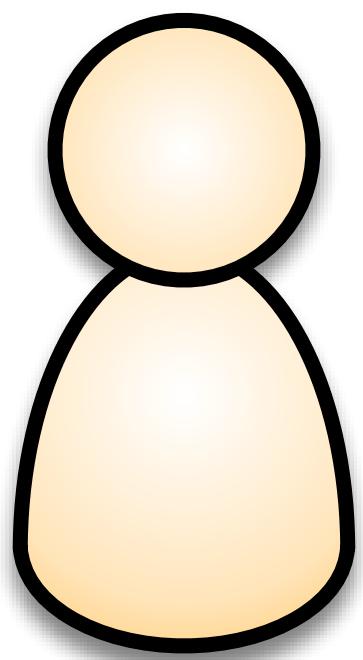
I would love it if this actually gave facts about the band. But instead it tells you things like, "some fish have fangs"

Can Alexa errors be leveraged to cause harm to end users?

Skill Squatting Attacks

An attacker can leverage predictable errors in Alexa to route users to skills that they didn't intend to go to

“Alexa, tell me some cat facts!”



Skills

...

cat forks

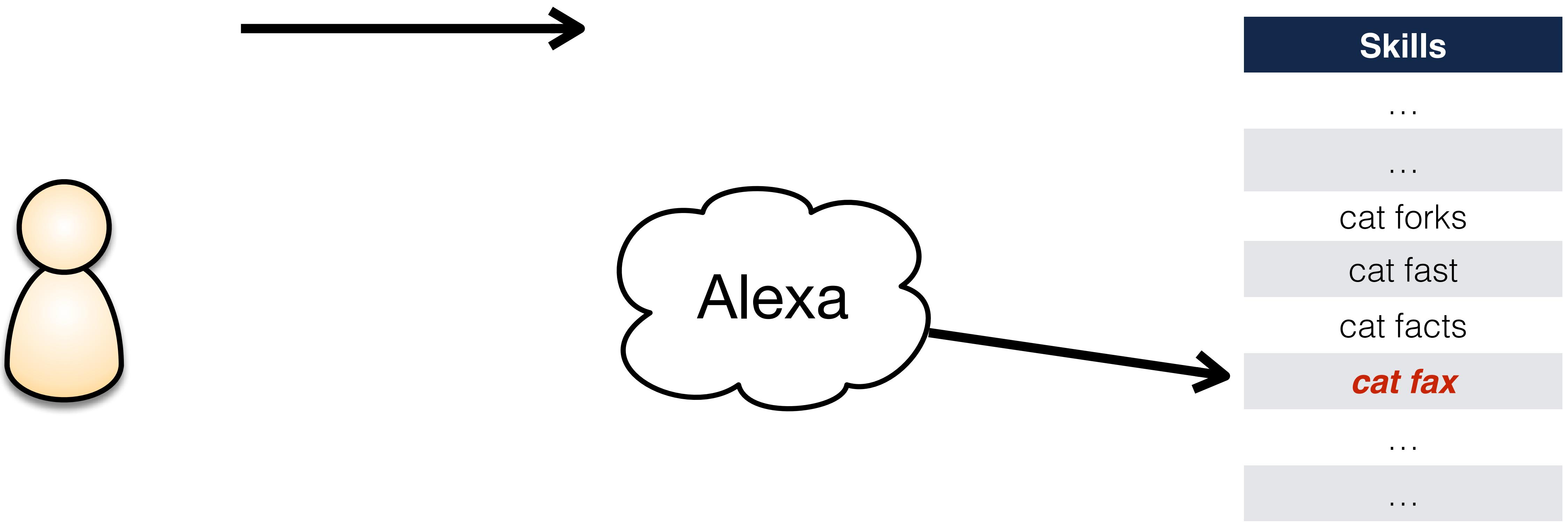
cat fast

cat facts

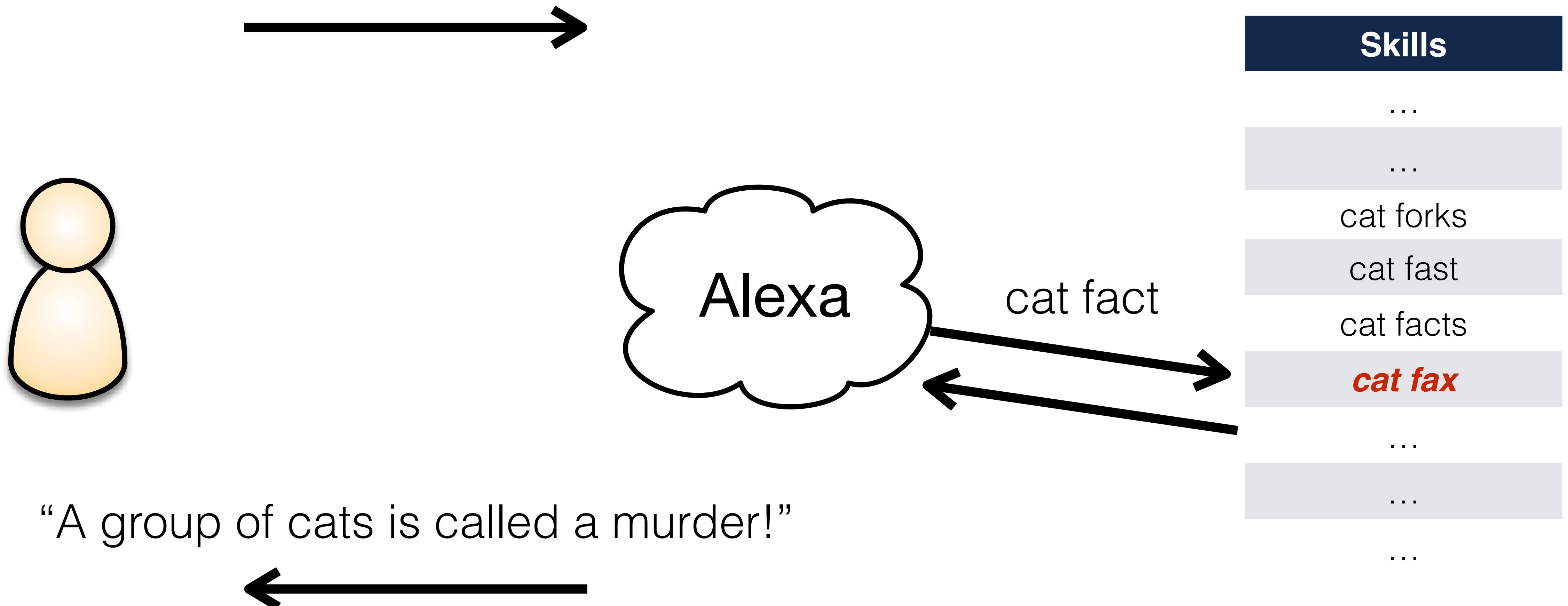
cat fax



“Alexa, tell me some cat facts!”



“Alexa, tell me some cat facts!”

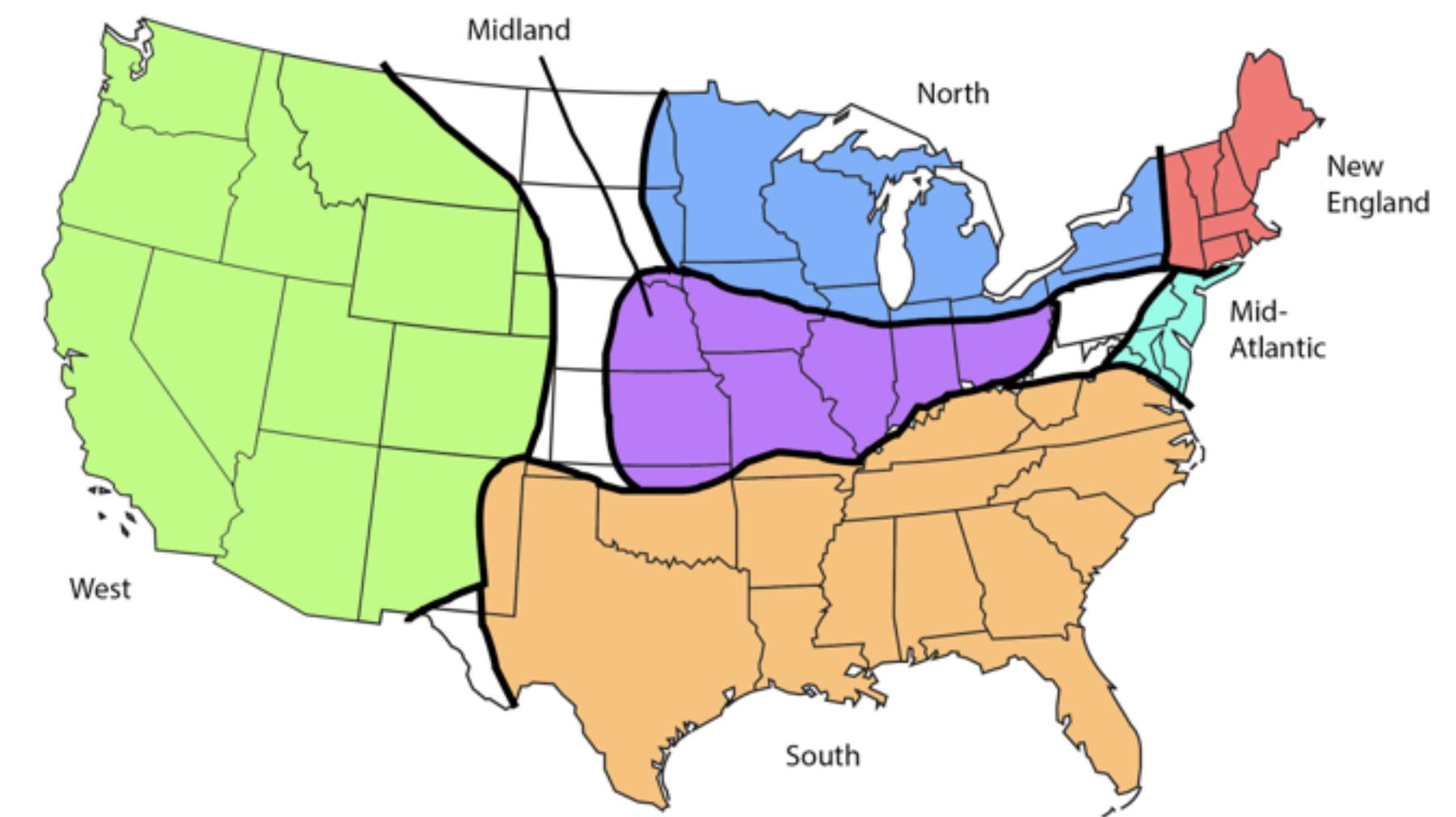


How can you tell which errors are predictable?

Send speech samples to Alexa,
figure out where it goes wrong

Speech Corpus

- Leveraged the **NSP Dataset**
 - 60 speakers, 188 unique words each (11,460 audio samples)
 - Speakers were representative of 6 US dialect regions



Measuring Interpretation Errors

We sent each speech sample to Alexa 50 times, providing us 573,000 transcriptions across the 60 speakers

Predictable Errors

Word	Prediction	Word	Prediction
Sail	Sale	Superhighway	Super Highway
Rip	Rap	Wet	What
Outshine	Outshyne	Main	Maine
Lung	Lang	Boil	Boyle
Accelerate	Xcelerate	Sell	Cell
Mill	No	Full	Four
Preferably	Preferrably	Dime	Time
Earthy	Fi	Bean	Been
Calm	Com	Dull	Doll
Coal	Call	Sweeten	Sweden
Outdoors	Out Doors	Luck	Lock
Loud	Louder	Con	Khan

Predictable Errors

Word	Prediction
Sail	Sale
Rip	Rap
Outshine	Outshyne
Lung	Lang
Accelerate	Xcelerate
Mill	No
Preferably	Preferrably
Earthy	Fi
Calm	Com
Coal	Call
Outdoors	Out Doors
Loud	Louder

Word	Prediction
Superhighway	Super Highway
Wet	What
Main	Maine
Boil	Boyle
Sell	Cell
Full	Four
Dime	Time
Bean	Been
Dull	Doll
Sweeten	Sweden
Luck	Lock
Con	Khan

Can we use our predictable errors
to route users to unintended skills?

Validating the Skill Squatting Attack

- Split speakers into two sets: “training” set and the “testing” set
- For each word with predictable error, we built two skills: the word, and the predictable error
 - Skill A: Wet
 - Skill B: What
- Sent the testing set through to Alexa, observed how many times skill B was triggered instead of skill A

A Brief, Ethical Note....

- We validated this attack strictly in a developer environment, no real skills were targeted or tested in the wild
- This is a fundamental limitation, but it's what we thought was the right thing to do

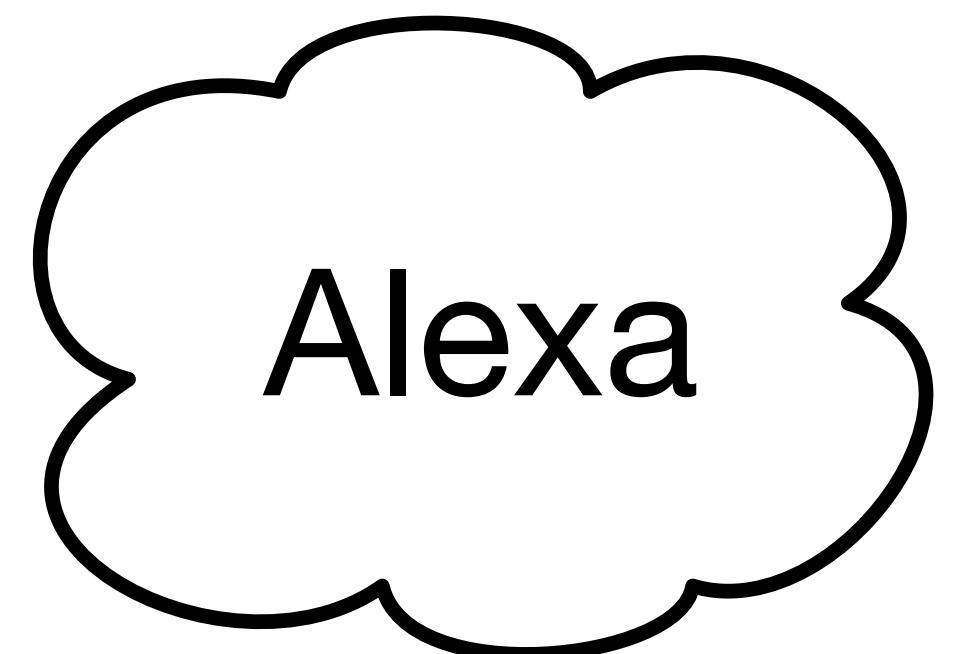
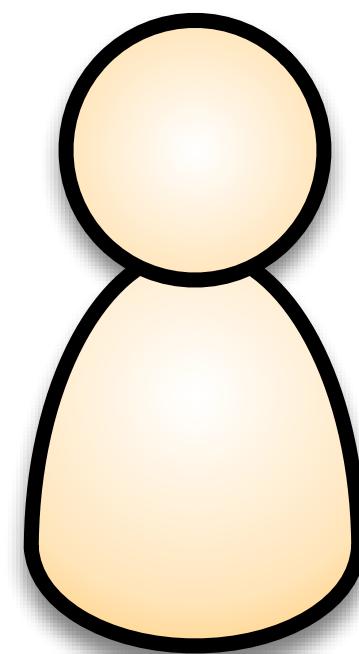
*Successfully squatted 25 of
27 (93%) predictable errors
at least once*

(

I would never want a cat fact.
Why does this matter?



“Alexa, ask Amex to pay
Bailey \$100”



Skills

...

...

Amex

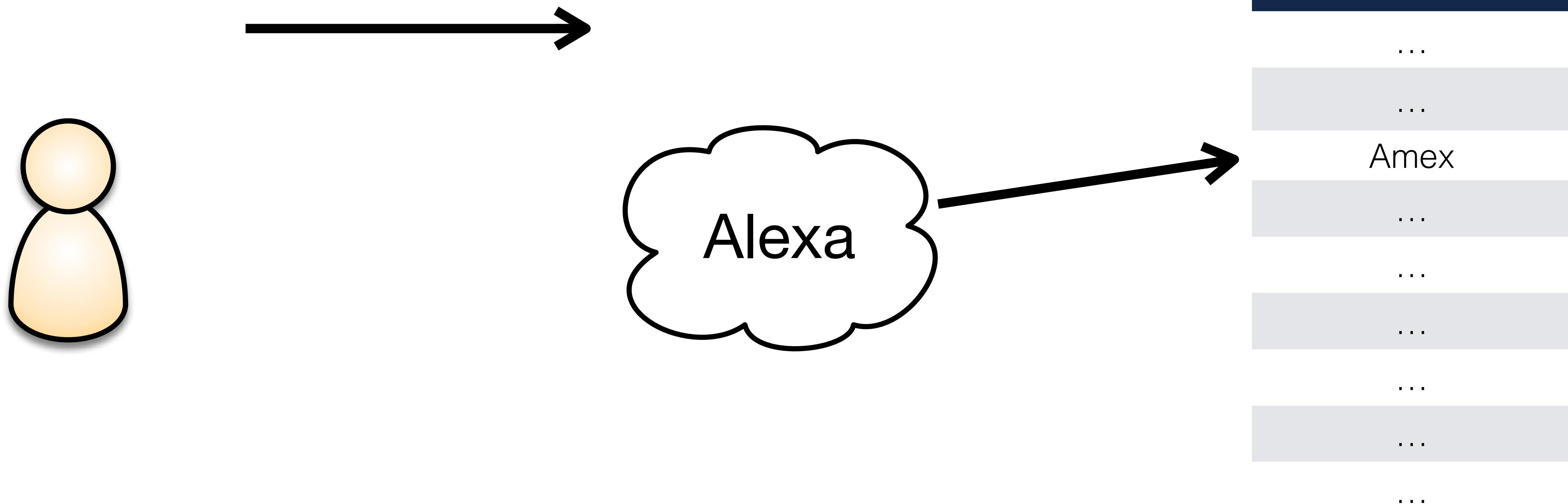
...

...

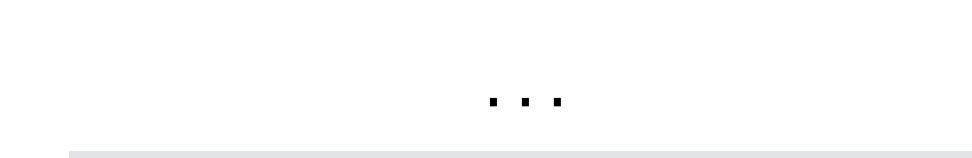
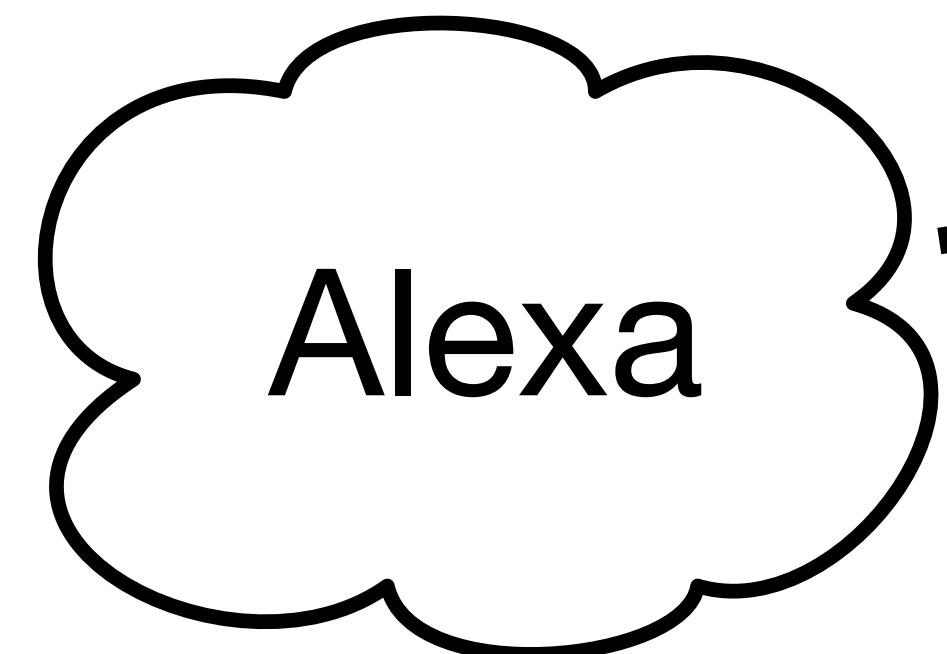
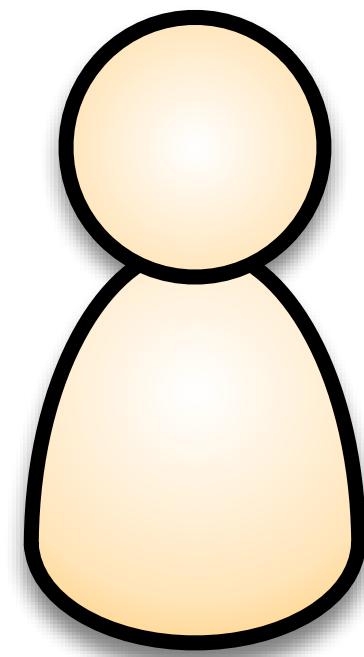
...

...

“Alexa, ask Amex to pay
Bailey \$100”

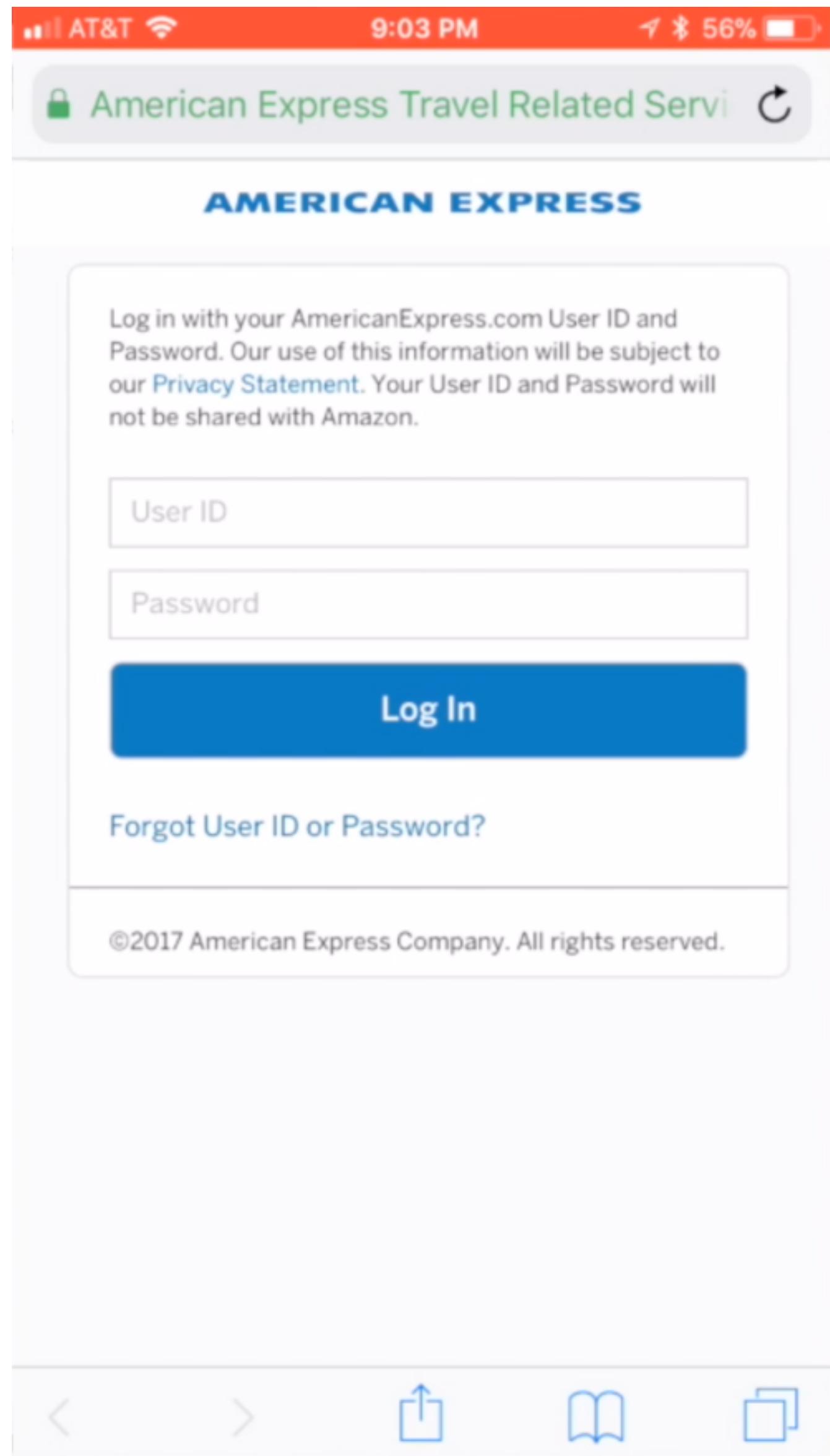


“Alexa, ask Amex to pay
Bailey \$100”

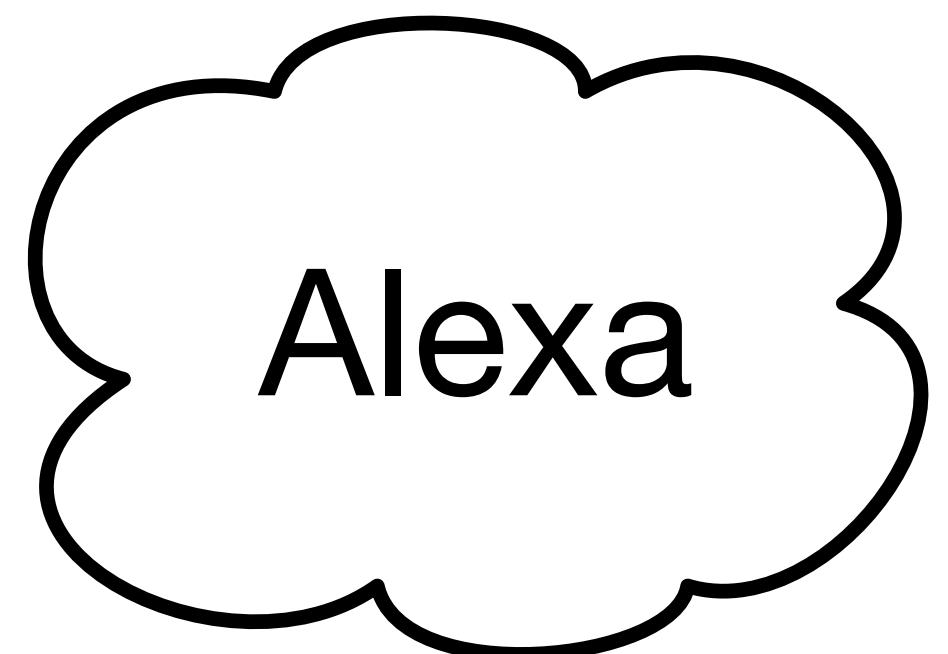
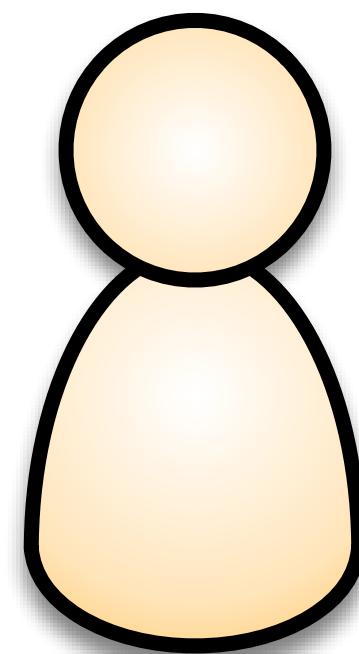


“You need to log in. I’ve sent a
card to your phone.”





“Alexa, ask Amex to pay
Bailey \$100”



Skills

...

...

Amex

Am X.

...

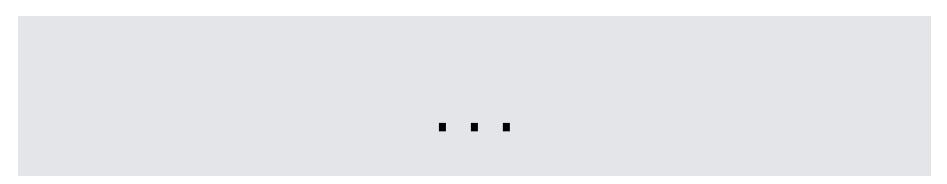
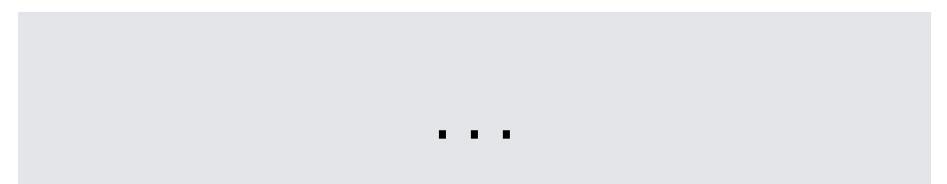
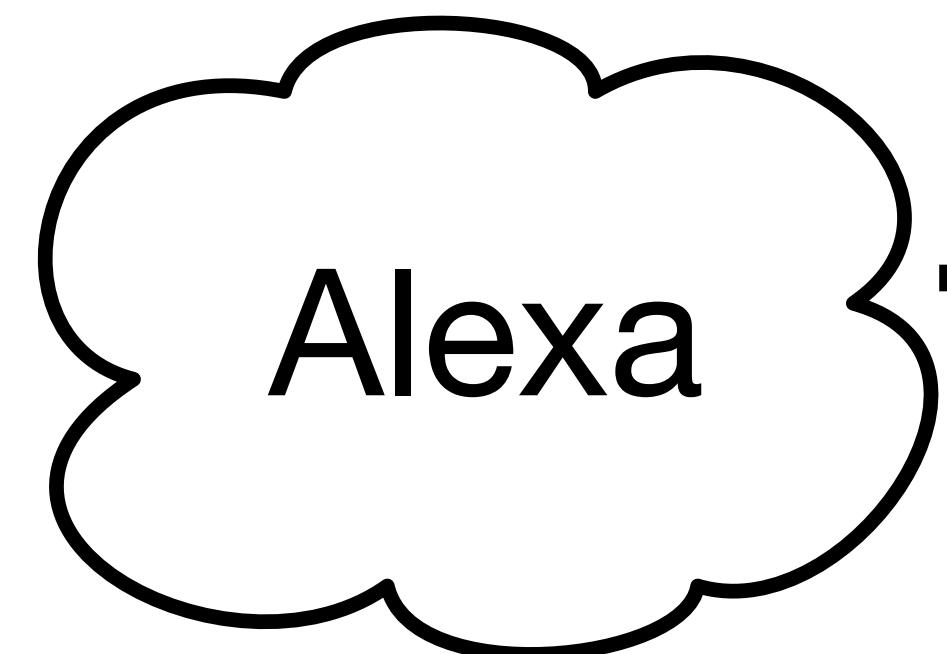
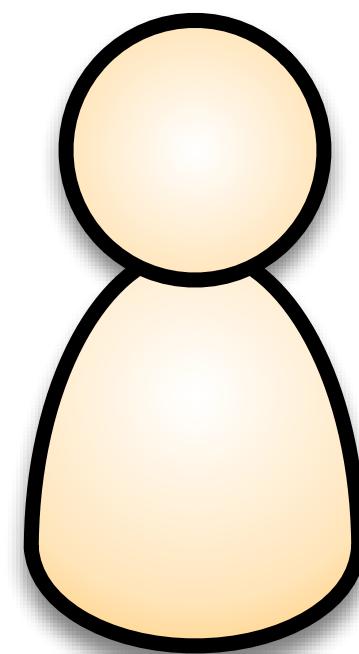
...

...

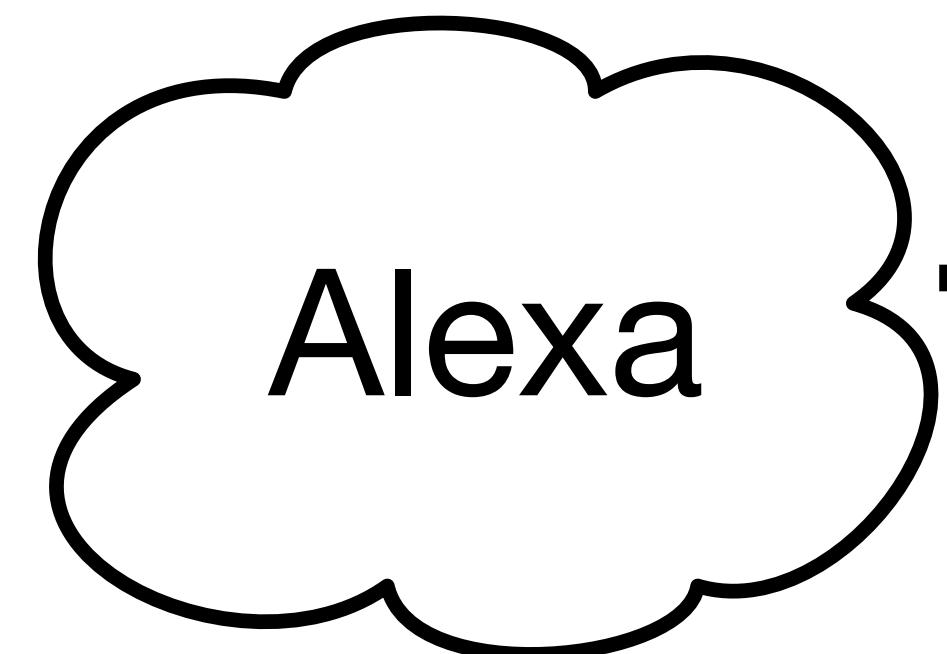
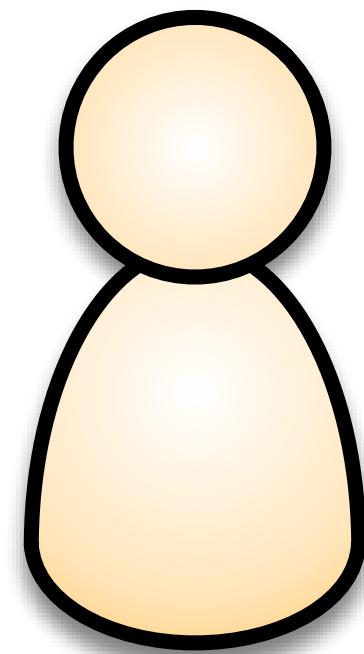
...

...

“Alexa, ask Amex to pay
Bailey \$100”

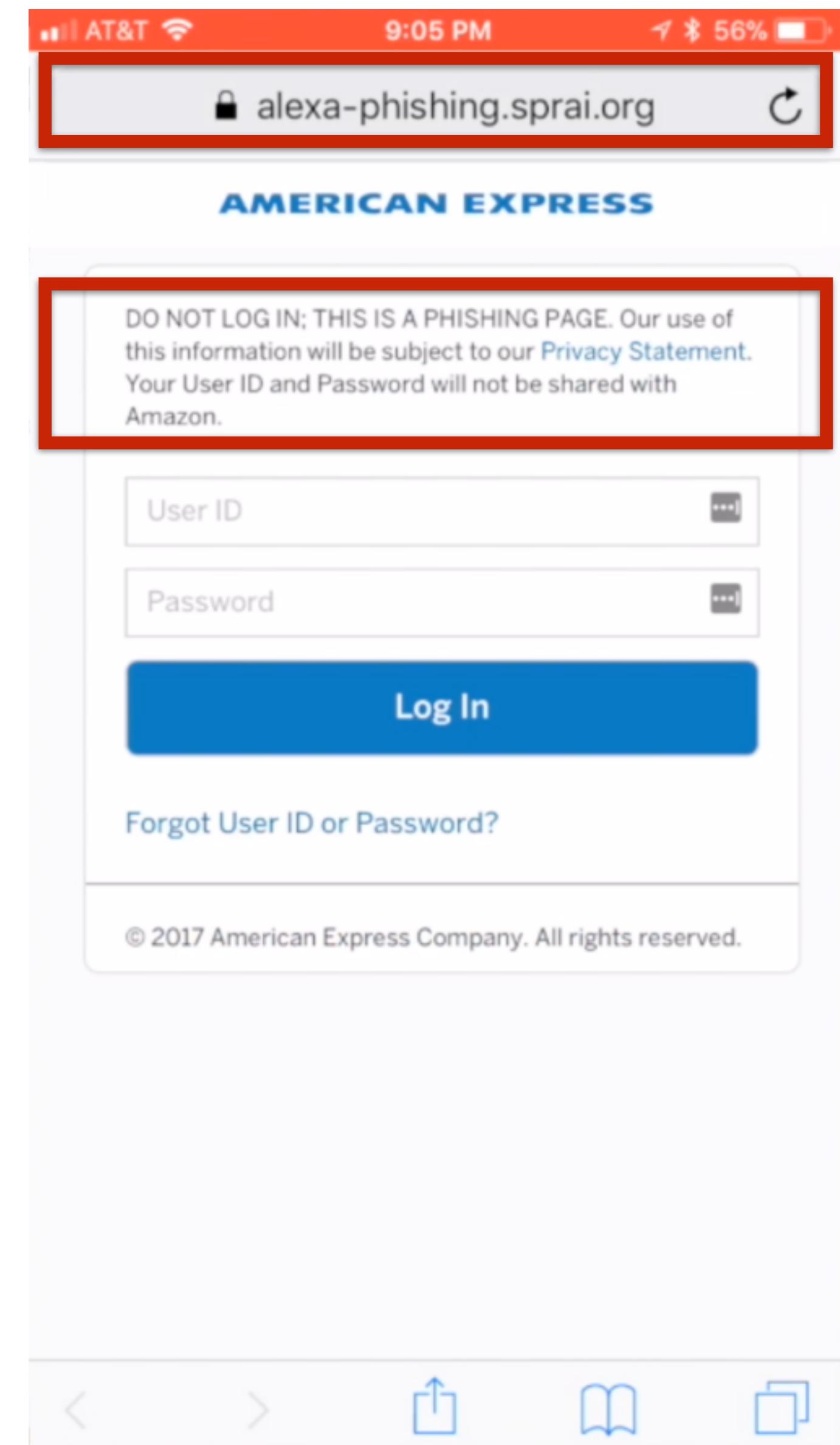


“Alexa, ask Amex to pay
Bailey \$100”



“You need to log in. I’ve sent a
card to your phone.”

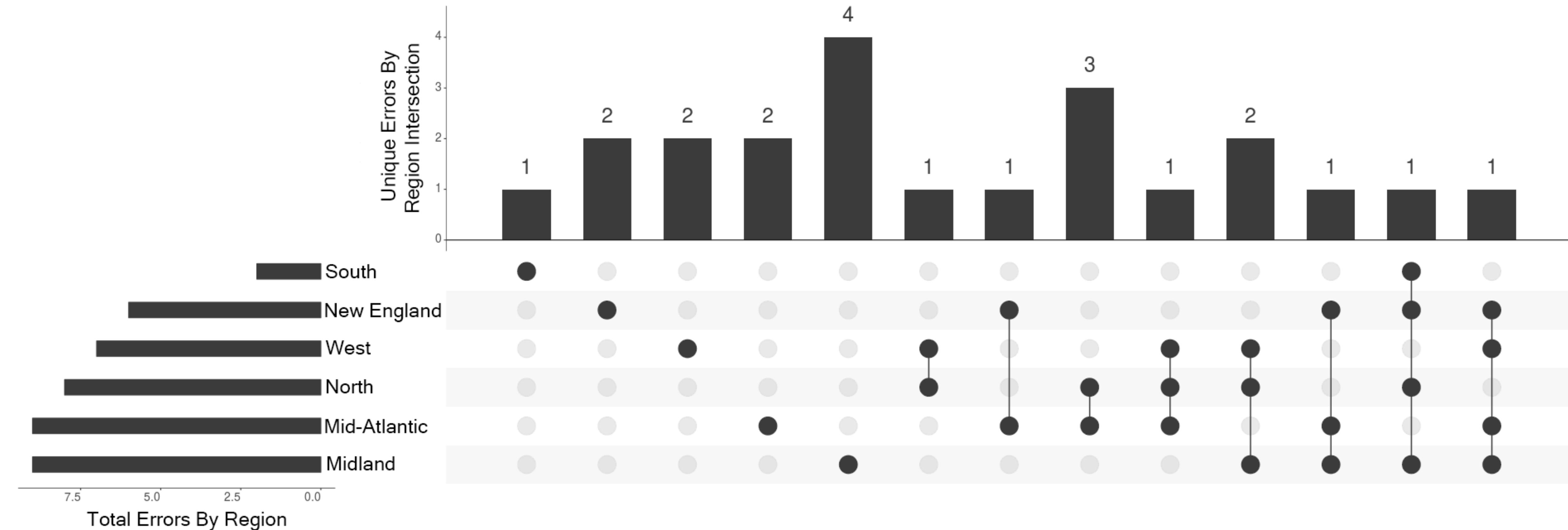




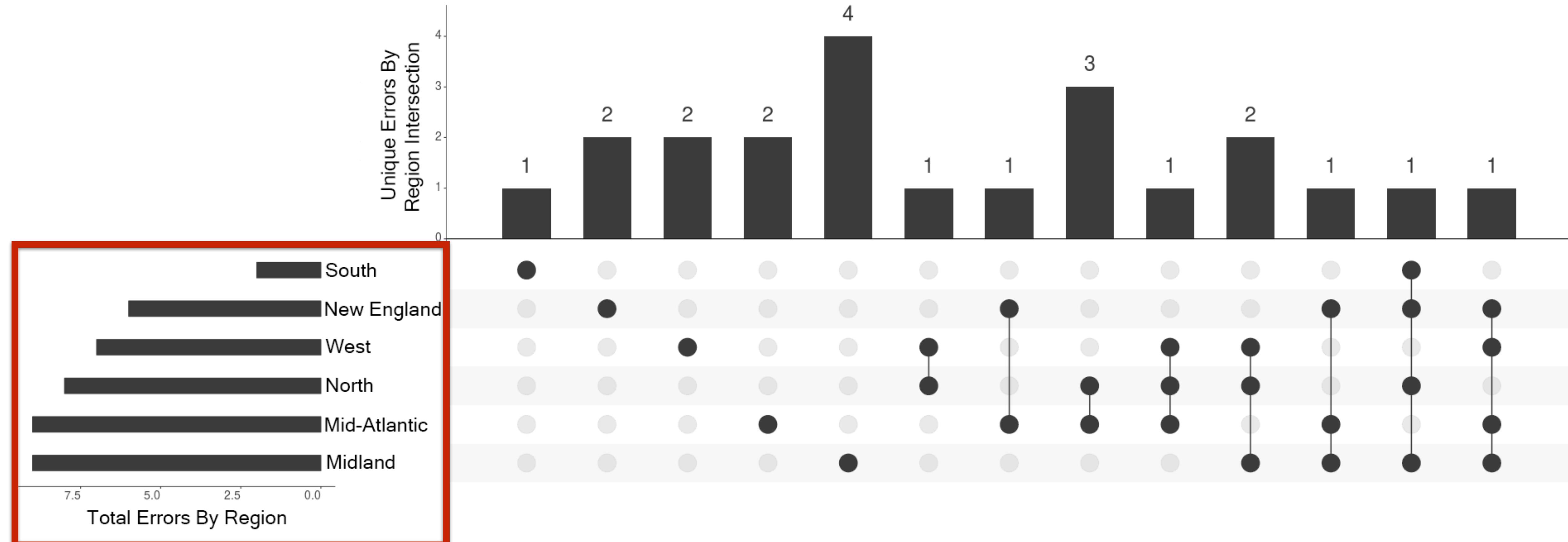
**VOICE IS THE NEXT BIG
PLATFORM, UNLESS YOU HAVE
AN ACCENT**

Do different regions exhibit unique
predictable interpretation errors?

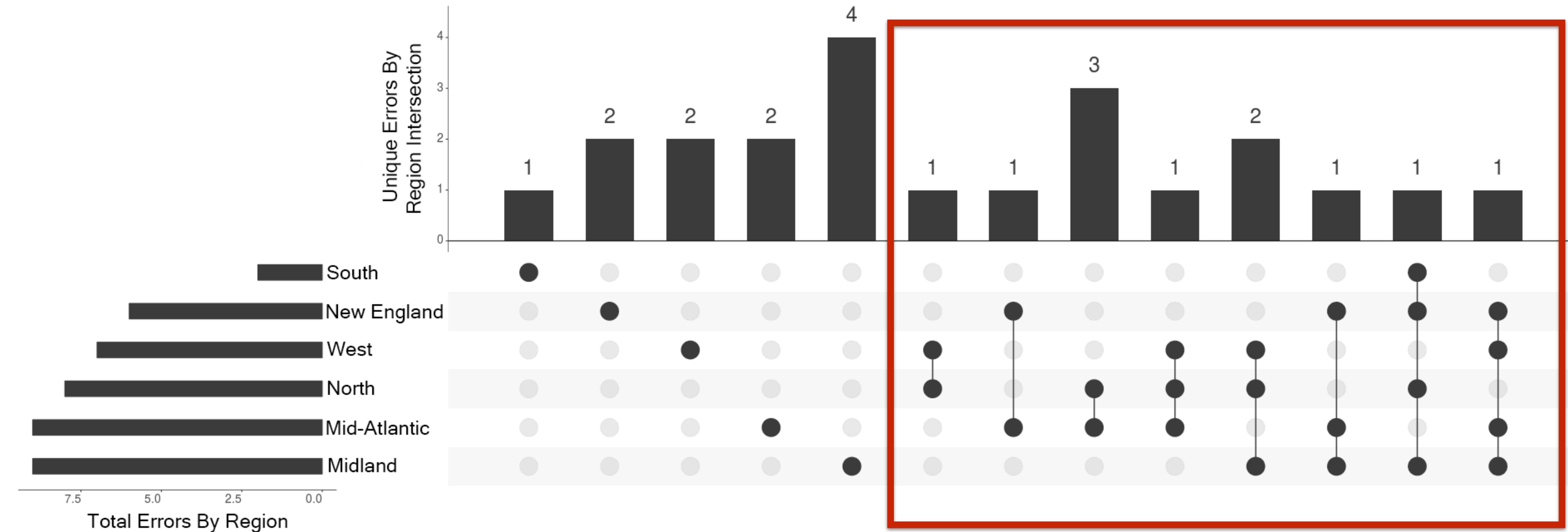
Predictable Errors by Region



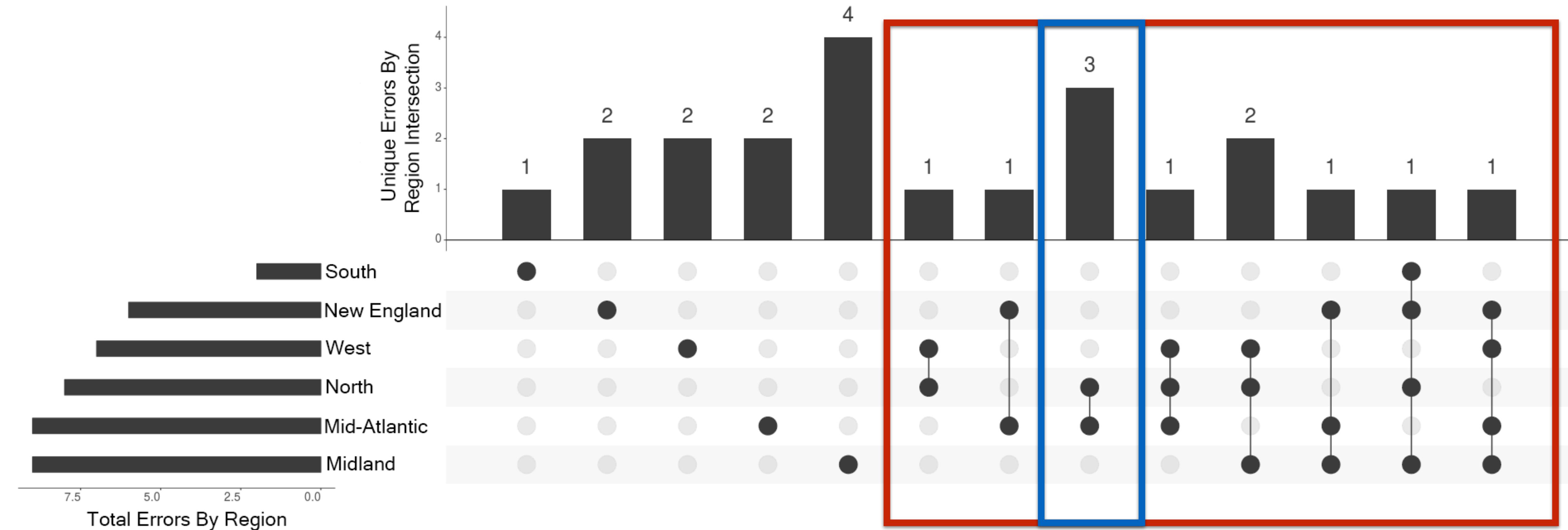
Predictable Errors by Region



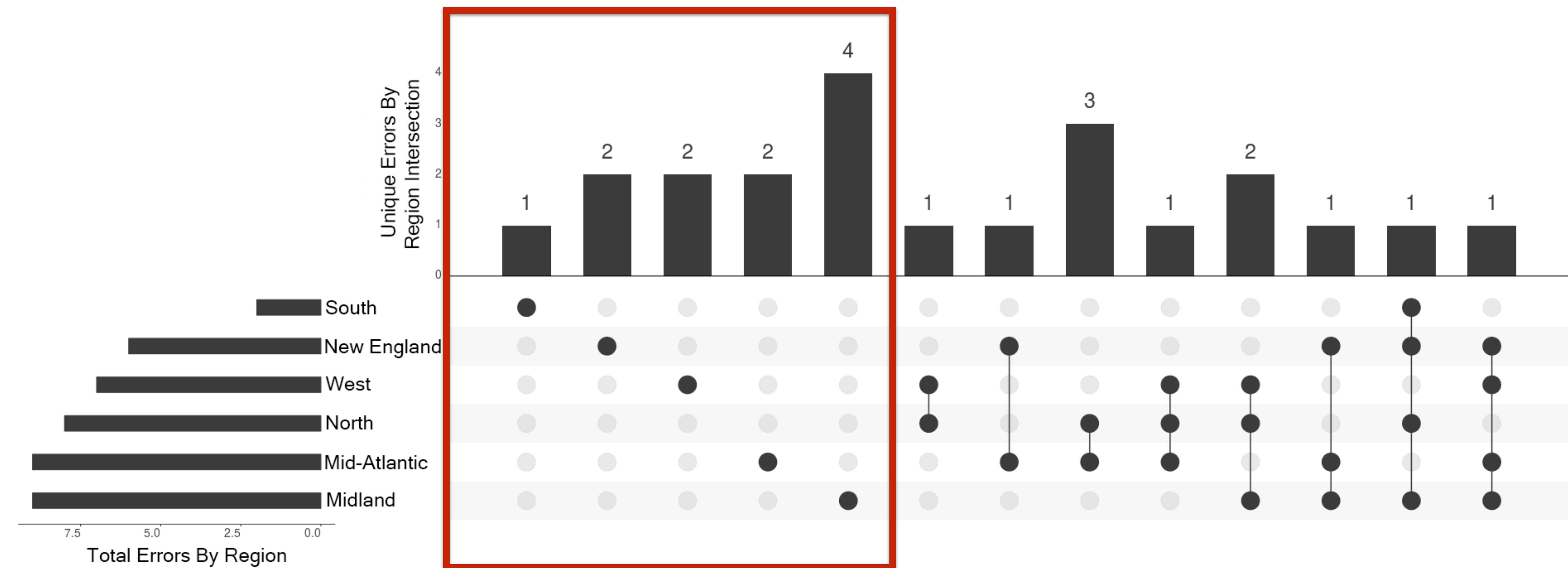
Predictable Errors by Region



Predictable Errors by Region



Predictable Errors by Region



Spear Skill Squatting Attacks

- An attacker can leverage accent-specific predictable errors in Alexa to route distinct groups of users to skills that they didn't intend to go to

Validating the Spear Skill Squatting Attack

Squatted Pair	Region	Target %	Overall %	Significant?
Tool/Two	South	34.0%	14.1%	Yes
Dock/Doc	West	97.4%	81.6%	No
Mighty/My T.	West	20.0%	4.1%	Yes
Exterior/Xterior	New England	42.9%	22.5%	Yes
Meal/Meow	New England	55.6%	34.3%	Yes
Wool/Well	Midland	50%	32.4%	No
Pal/Pow	Midland	65.9%	37.7%	Yes
Accuser/Who's There	Midland	26.0%	4.9%	Yes
Pin/Pen	Midland	26.3%	10.0%	Yes
Malfunction/No Function	Mid-Atlantic	36.0%	27.5%	No
Fade/Feed	Mid-Atlantic	59.0%	14.7%	Yes

Validating the Spear Skill Squatting Attack

Squatted Pair	Region	Target %	Overall %	Significant?
Tool/Two	South	34.0%	14.1%	Yes
Dock/Doc	West	97.4%	81.6%	No
Mighty/My T.	West	20.0%	4.1%	Yes
Exterior/Xterior	New England	42.9%	22.5%	Yes
Meal/Meow	New England	55.6%	34.3%	Yes
Wool/Well	Midland	50%	32.4%	No
Pal/Pow	Midland	65.9%	37.7%	Yes
Accuser/Who's There	Midland	26.0%	4.9%	Yes
Pin/Pen	Midland	26.3%	10.0%	Yes
Malfunction/No Function	Mid-Atlantic	36.0%	27.5%	No
Fade/Feed	Mid-Atlantic	59.0%	14.7%	Yes

Validating the Spear Skill Squatting Attack

Squatted Pair	Region	Target %	Overall %	Significant?
Tool/Two	South	34.0%	14.1%	Yes
Dock/Doc	West	97.4%	81.6%	No
Mighty/My T.	West	20.0%	4.1%	Yes
Exterior/Interior	New England	42.0%	22.3%	Yes
Meal/Mew	New England	55.6%	34.3%	Yes
Wool/Well	Midland	50.0%	32.4%	No
Fad/Fow	Midland	66.6%	37.7%	Yes
Accuser/Who's There	Midland	26.0%	4.9%	Yes
Pin/Pen	Midland	26.3%	10.0%	Yes
Malfunction/No Function	Mid-Atlantic	36.0%	27.5%	No
Fade/Feed	Mid-Atlantic	59.0%	14.7%	Yes

Successfully squatted 8 out of 11 spear squattable pairs

Limitations

- Scale + Representativeness of the dataset

Limitations

- Scale + Representativeness of the dataset
- Skill behavior *outside* of a development environment

Takeaways

- New medium, same problems
 - “Typosquatting” in the land of IoT

Takeaways

- New medium, same problems
 - “Typosquatting” in the land of IoT
- Opaque ML for *decision making* is still nascent
 - Interface quirks can and will be exploited to cause abuse

Moving Forward

- Working with Amazon to fix these issues in their platform

Moving Forward

- Working with Amazon to fix these issues in their platform
- Measuring the widespread harms of skill squatting

Moving Forward

- Working with Amazon to fix these issues in their platform
- Measuring the widespread harms of skill squatting
- Investigating IoT trust relationships
 - Do users intrinsically trust voice-based devices more than online?

Questions?

dkumar11@illinois.edu
@_kumarde

backup