
AWS
re:Invent

MDS206

Empowering users with natural language search

Ramin Beheshti
Chief Product & Technology Officer
Dow Jones

Sheri Moran
Principal Solutions Architect
AWS

AWS re:Invent

© 2020, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Agenda

Enterprise Search—still a challenge today

Rethinking Enterprise Search with Amazon Kendra

The WSJ Talk2020

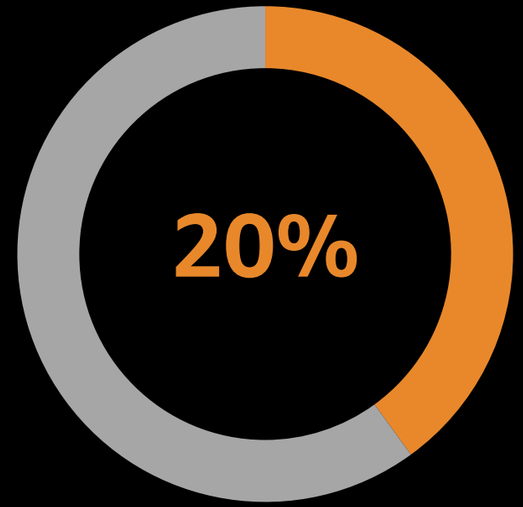
Call to action



© 2020, Amazon Web Services, Inc. or its affiliates. All rights reserved.

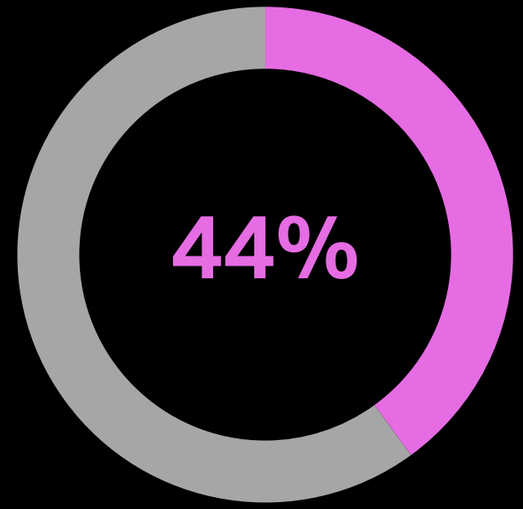
Employees spend **20%** of their time looking for information

—McKinsey



44% of the time, they cannot find the information they need to do their job

—IDC



Key challenges

Low accuracy

- 80% of data is unstructured
- Keyword engines

Complexity

- Scattered data silos
- Difficult to set up
- Extensive manual maintenance

Impacts

Lower employee productivity

Increased risk & liability

Duplication of work

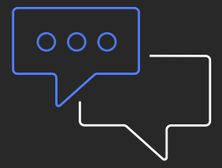
Created a negative customer experience

Our mission at AWS

Put machine learning in the
hands of every developer

Amazon Kendra

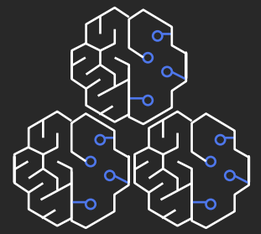
HIGHLY ACCURATE AND EASY-TO-USE INTELLIGENT SEARCH SERVICE POWERED BY MACHINE LEARNING



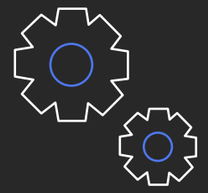
Natural language queries



NLU and Machine Learning core



Domain expertise



Continuous improvement



Native connectors



Simple API and console experiences code samples

Intuitively search unstructured data using natural language search

The AWS ML Stack

BROADEST AND MOST COMPLETE SET OF MACHINE LEARNING CAPABILITIES

AI SERVICES

VISION	SPEECH		TEXT			SEARCH	CHATBOTS	PERSONALIZATION	FORECASTING	FRAUD	DEVELOPMENT	CONTACT CENTERS
Amazon Rekognition	Amazon Polly	Amazon Transcribe *+Medical*	Amazon Comprehend *+Medical*	Amazon Translate	Amazon Textract	Amazon Kendra	Amazon Lex	Amazon Personalize	Amazon Forecast	Amazon Fraud Detector	Amazon CodeGuru	Contact Lens *For Amazon Connect*

ML SERVICES

Amazon SageMaker	Ground Truth	AWS Marketplace for ML	SageMaker Studio IDE								Neo	Augmented AI
			Built-in algorithms	Notebooks	Experiments	Processing	Model training and tuning	Debugger	Autopilot	Model hosting	Model Monitor	

ML FRAMEWORKS & INFRASTRUCTURE

TensorFlow, mxnet, PYTORCH	GLUON, Keras, DeepGraphLibrary	Deep Learning AMIs & Containers	GPUs & CPUs	Elastic Inference	Inferentia	FPGA

AWS Machine Learning Services for Search

AMAZON PROVIDES CUSTOMERS WITH CHOICE WHEN IMPLEMENTING INTELLIGENT SEARCH SOLUTIONS FOR THEIR BUSINESS



Amazon Kendra

- Makes it quick and easy for developers and builders to implement AI and improve business outcomes
- No ML experience required
- Fully managed and pre-trained AI service that automates adding intelligence to your applications



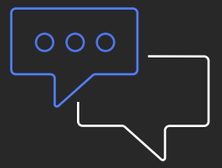
Amazon SageMaker

- Build, train, and deploy your own ML models quickly using a visual IDE
- All tools required for ML in a single service
- Prebuilt models, algorithms, and solutions plus AutoML
- Fully managed for scalability, performance, and cost-effectiveness

Quickly and easily finding accurate information

AMAZON KENDRA

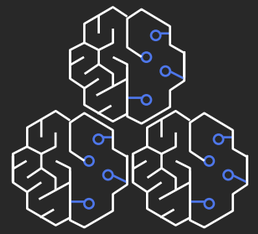
EASY TO FIND WHAT YOU ARE LOOKING FOR



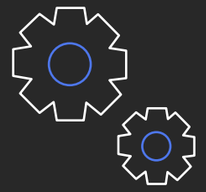
Natural language queries



NLU and Machine Learning core



Domain expertise



Continuous improvement

SIMPLE AND QUICK TO SET UP



Native connectors



Simple API and console experiences code samples

Low Accuracy

Complexity

Amazon Kendra

NATURAL LANGUAGE QUERIES



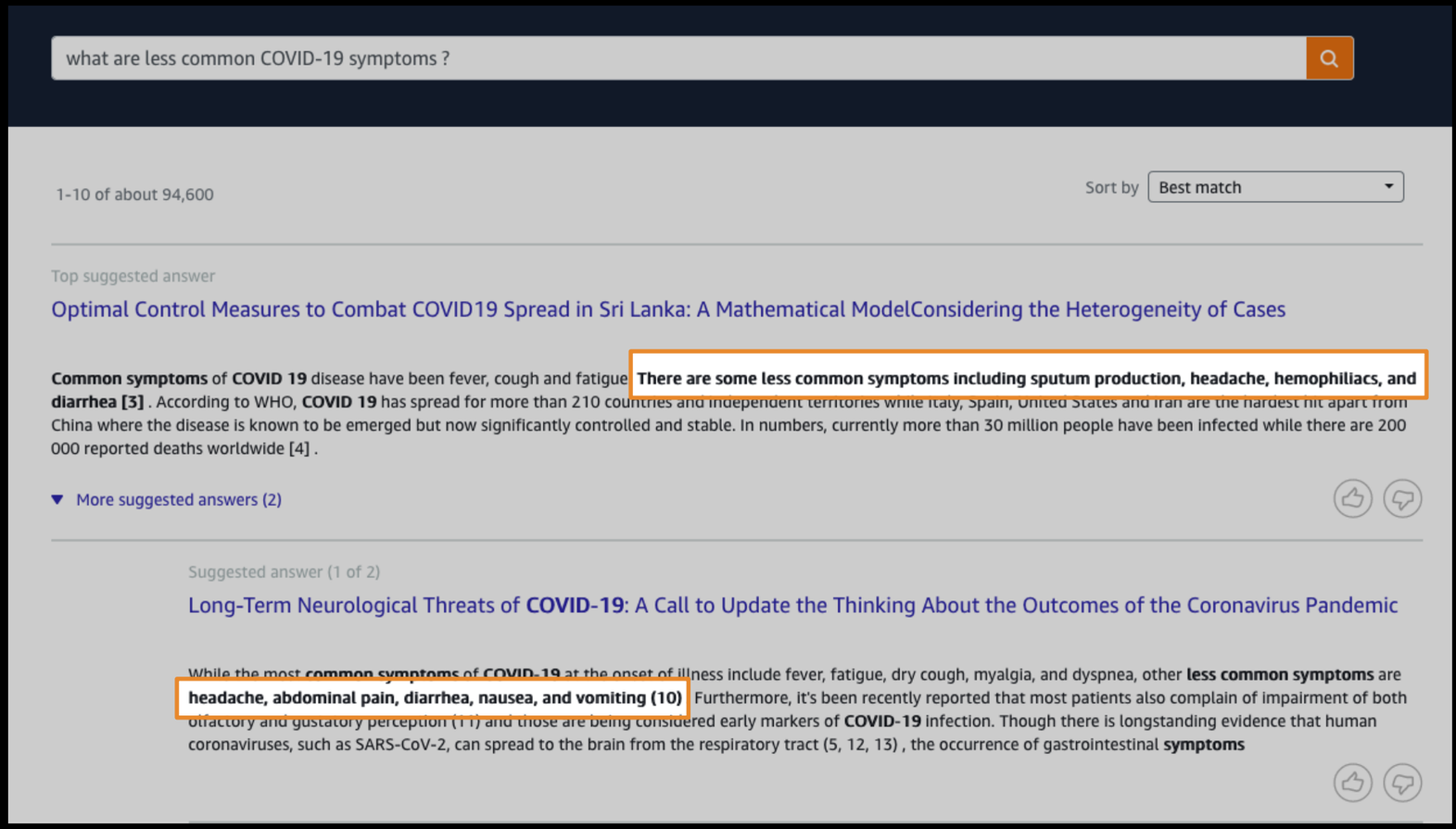
Natural language queries

Natural language queries

"Where is the IT support desk in Kumo?" → "1st floor"

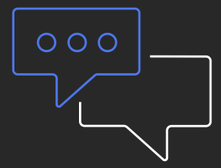
Keyword queries

"Health benefits" or "parking allowance"



Amazon Kendra

NLU AND ML CORE



Natural language Queries



NLU and Machine Learning core

Better answers

Reading Comprehension

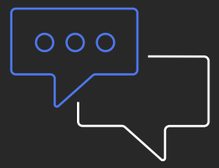
FAQ matching

Document ranking



Amazon Kendra

DOMAIN EXPERTISE



Natural language Queries



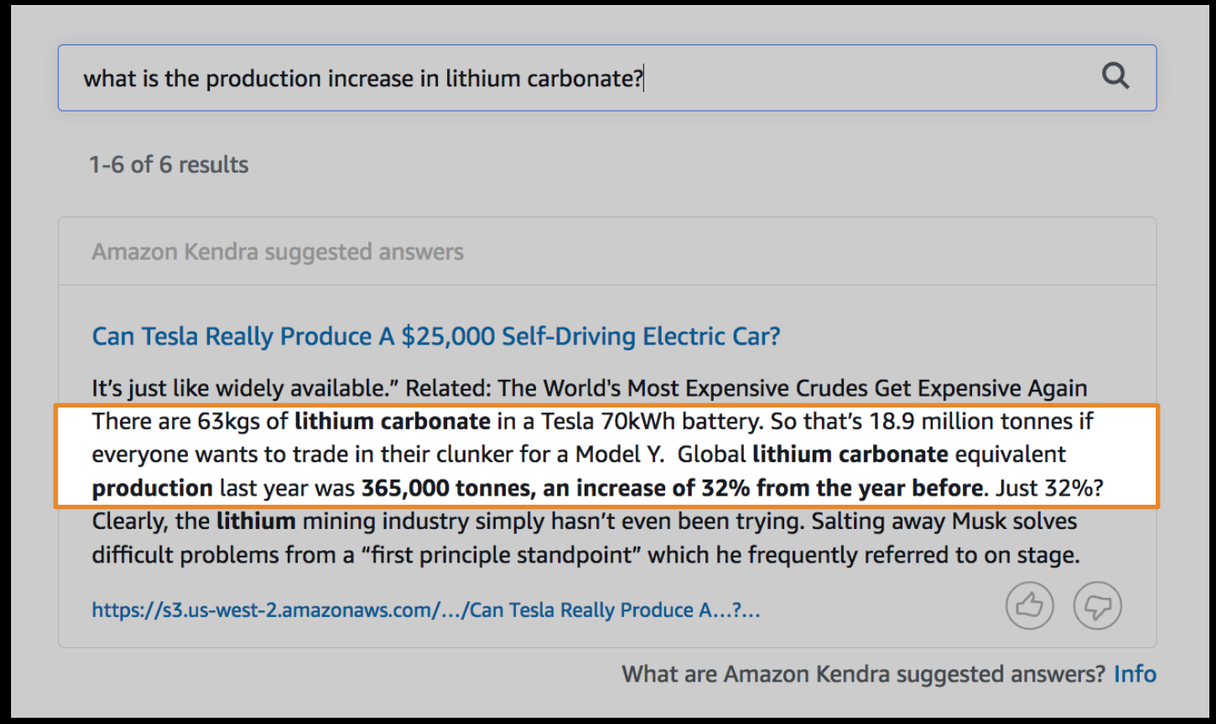
NLU and Machine Learning core



Domain expertise

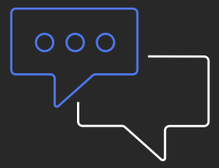
Optimized for 14 major domains

IT, **FINANCIAL SERVICES**, Insurance, Pharma, **INDUSTRIAL,** Energy, Legal, Media and Entertainment, Travel and Hospitality, Health, HR, News, Telecommunications, and Automotive



Amazon Kendra

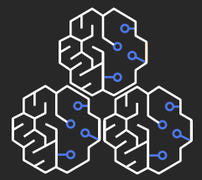
INCREMENTAL LEARNING



Natural language Queries



NLU and Machine Learning core



Domain expertise

COMING SOON!

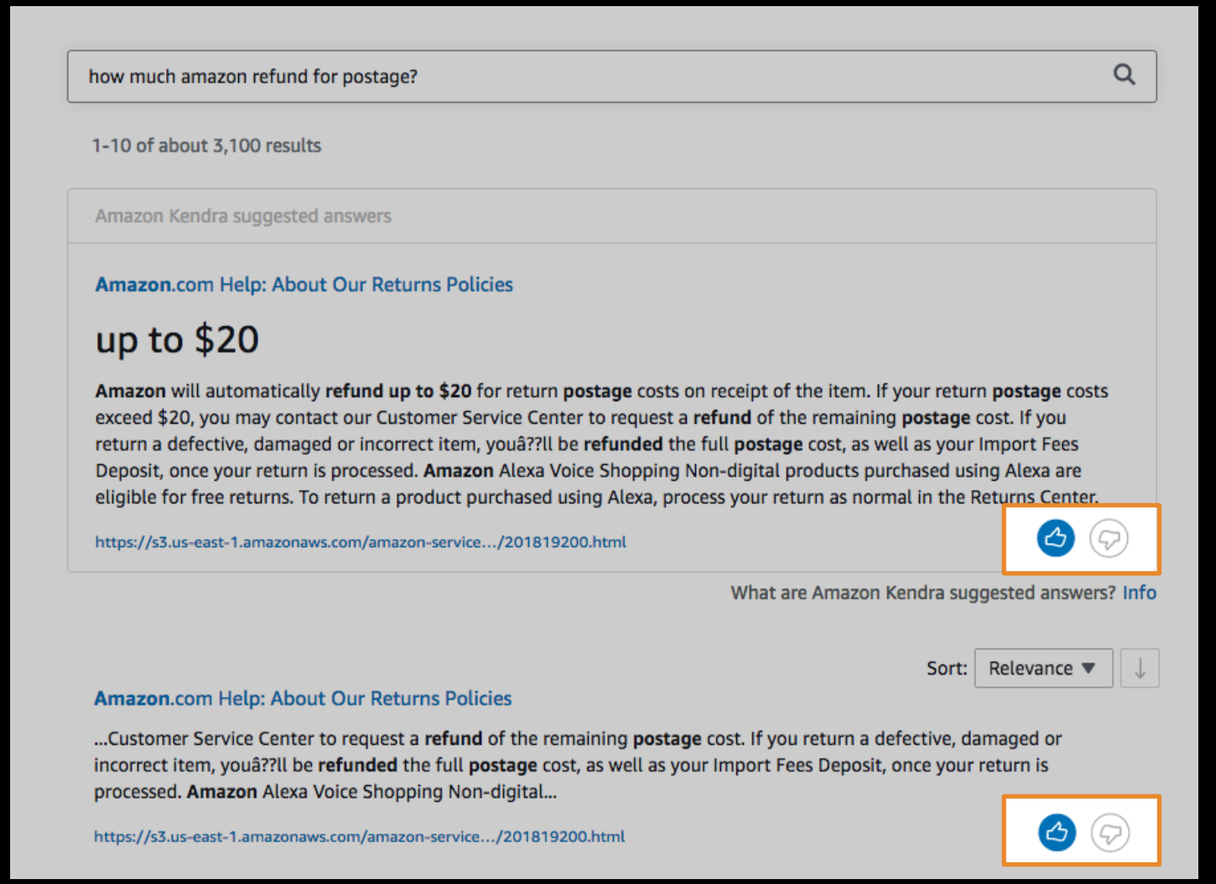


Continuous Improvement

Amazon Kendra improves automatically over time

Captures click-through and user feedback

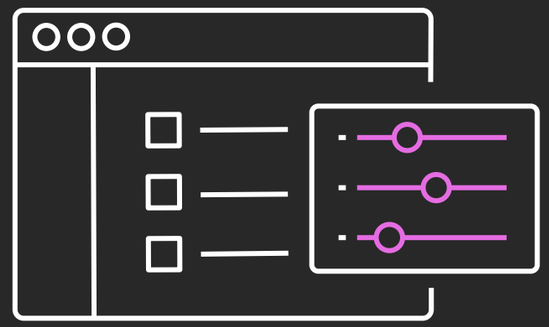
Retrains models periodically



Fine-tune accuracy of Amazon Kendra with relevance tuning

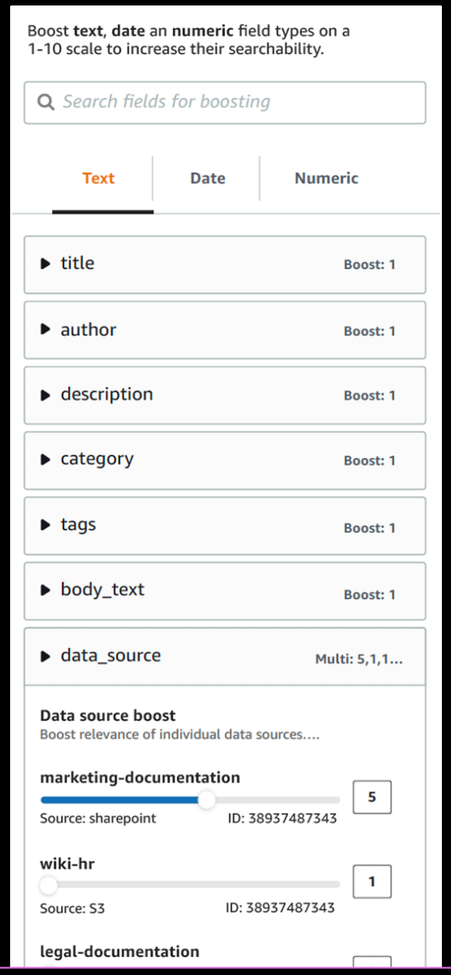
Prioritize your data

- Authoritative data sources
- Document freshness
- Authors or departments
- Upvotes and popularity

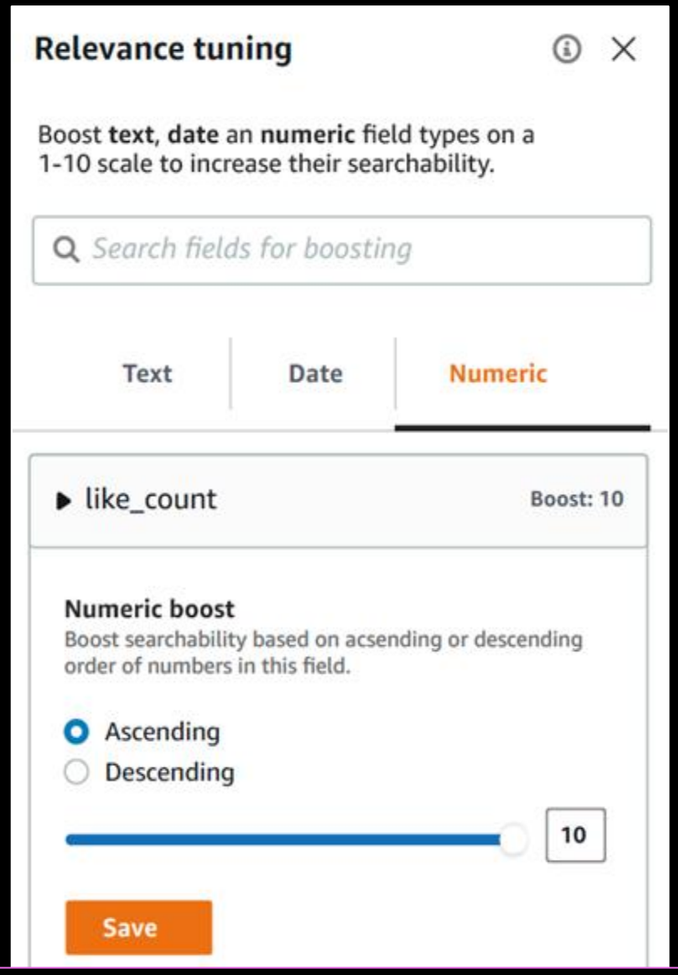


→ Fine-tune Amazon Kendra's accuracy

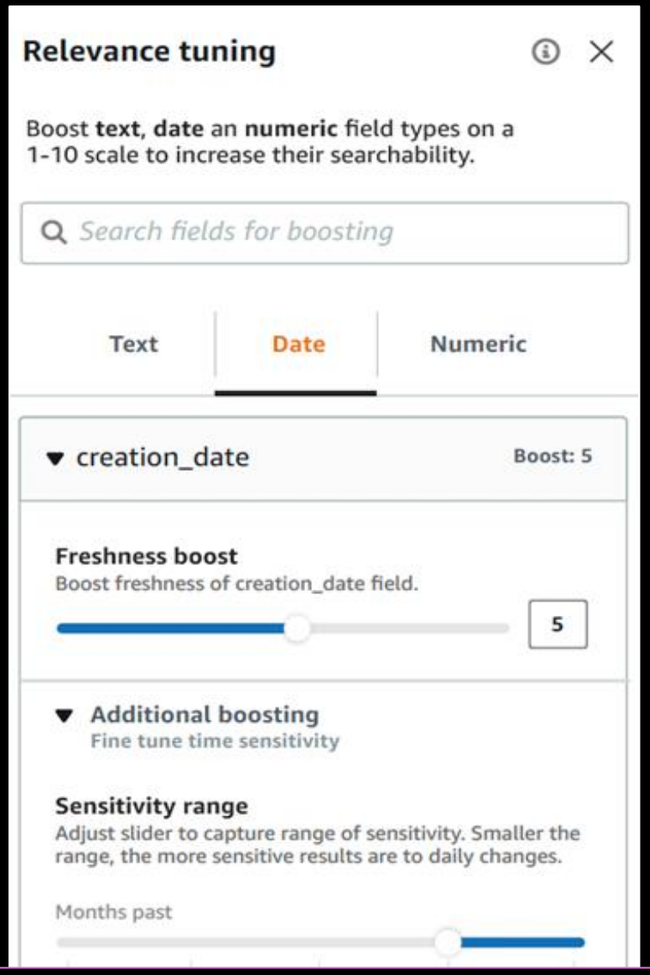
Boosting results



Data sources



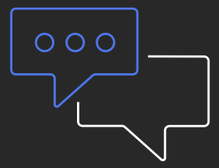
View/like counts



Document freshness

Amazon Kendra

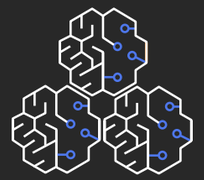
NATIVE CONNECTORS



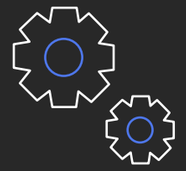
Natural language Queries



NLU and Machine Learning core



Domain expertise



Continuous improvement

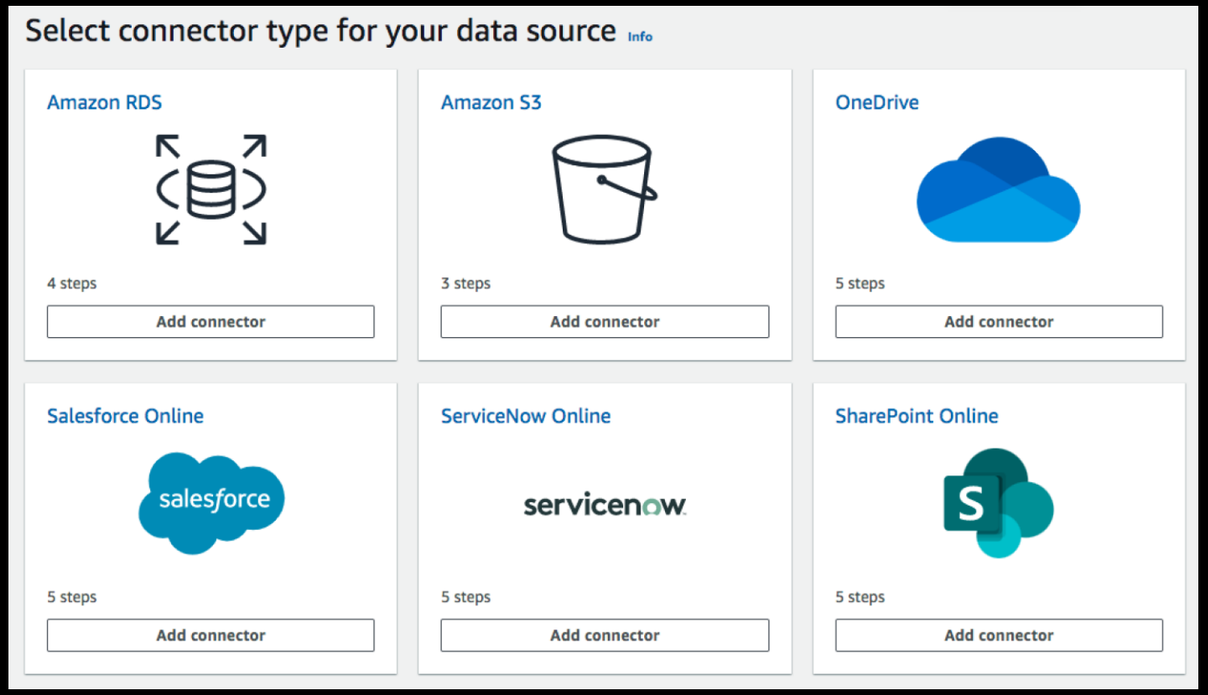


Native connectors

The most popular and counting

e.g., S3, file systems (SMB), Web crawler, Databases, SharePoint

Add credentials, point and shoot



Quickly and easily find accurate information



Increase employee productivity
by equipping employees with the data they need to do their jobs (e.g., HR, Ops)



Accelerate research and development
by finding relevant information in highly technical research and engineering documents



Enhance customer experiences
by easily finding answers to questions on websites, in call centers and via chatbots (without bot training)



Minimize regulatory and compliance risks
enhancing regulatory intelligence, correspondence and access to up-to-date compliance data



Easily build more intelligent applications
enhancing applications with a semantic search engine for content management

Customer spotlight

AMAZON KENDRA AND WSJ TALK2020



AWS re:Invent

© 2020, Amazon Web Services, Inc. or its affiliates. All rights reserved.

The beginning: Talk2020 for reporters

BACKGROUND

Internal tool made from reporter need

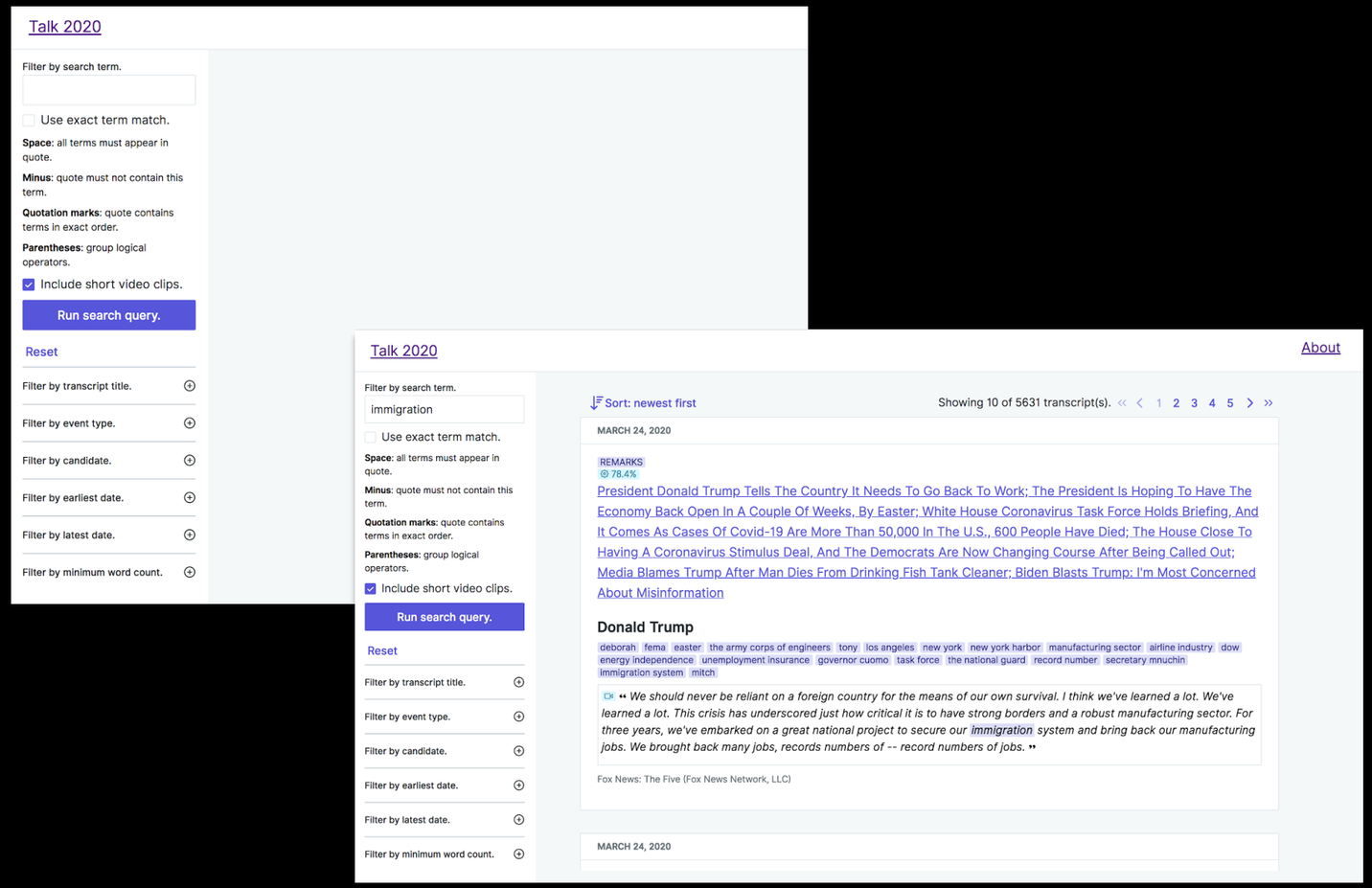
Idea was generated from the Wall Street Journal's DC Bureau so they can perform research for analytical stories. It started with an internal developer and then became a partnership with our newsroom's R&D team.

Experience tailored for Power Users

It was just controls and an empty UI, utilized an internal design system for quick delivery but with minimal UX.

Laid the groundwork

Went through a lot of development phases from a simple Django-powered app to something more robust and real-time with internally developed ML models.



The Challenge: Take it public

BACKGROUND

Can it answer a user's need?

We know that readers want facts, and that they want to be able to quickly identify and share them.

Search vs. browse

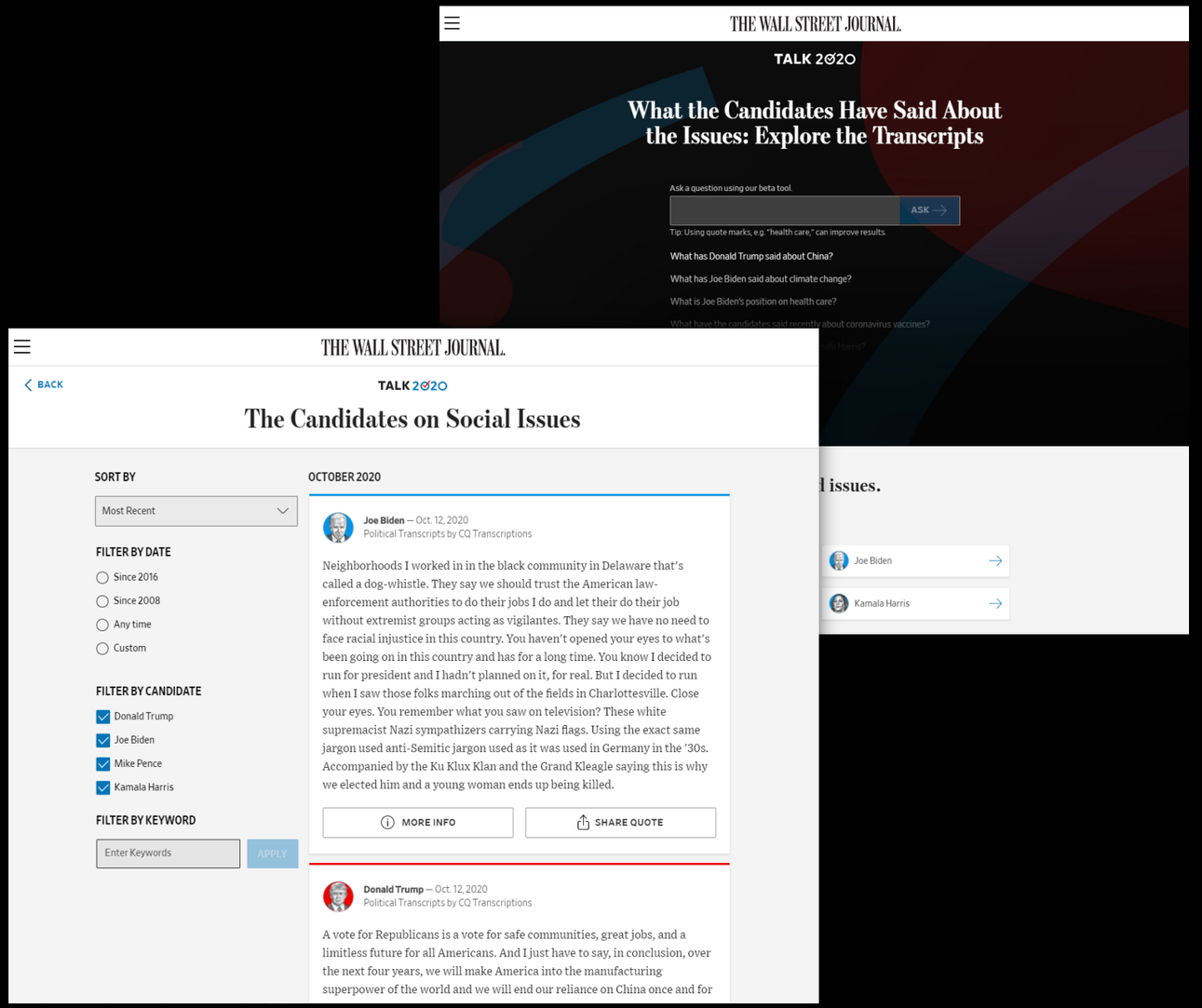
Allow users to observe a general topic or candidate stances as well as utilizing direct search.

Can we make smart search scale?

Making sure that a free product released to the public wouldn't incur performance problems was a big concentration.

Make it part of the WSJ universe

Search and reader-driven utility as an entry way to other WSJ content.



Architecting a tool to move with the news

LESSONS LEARNED

Real-Time Ingestion Pipeline

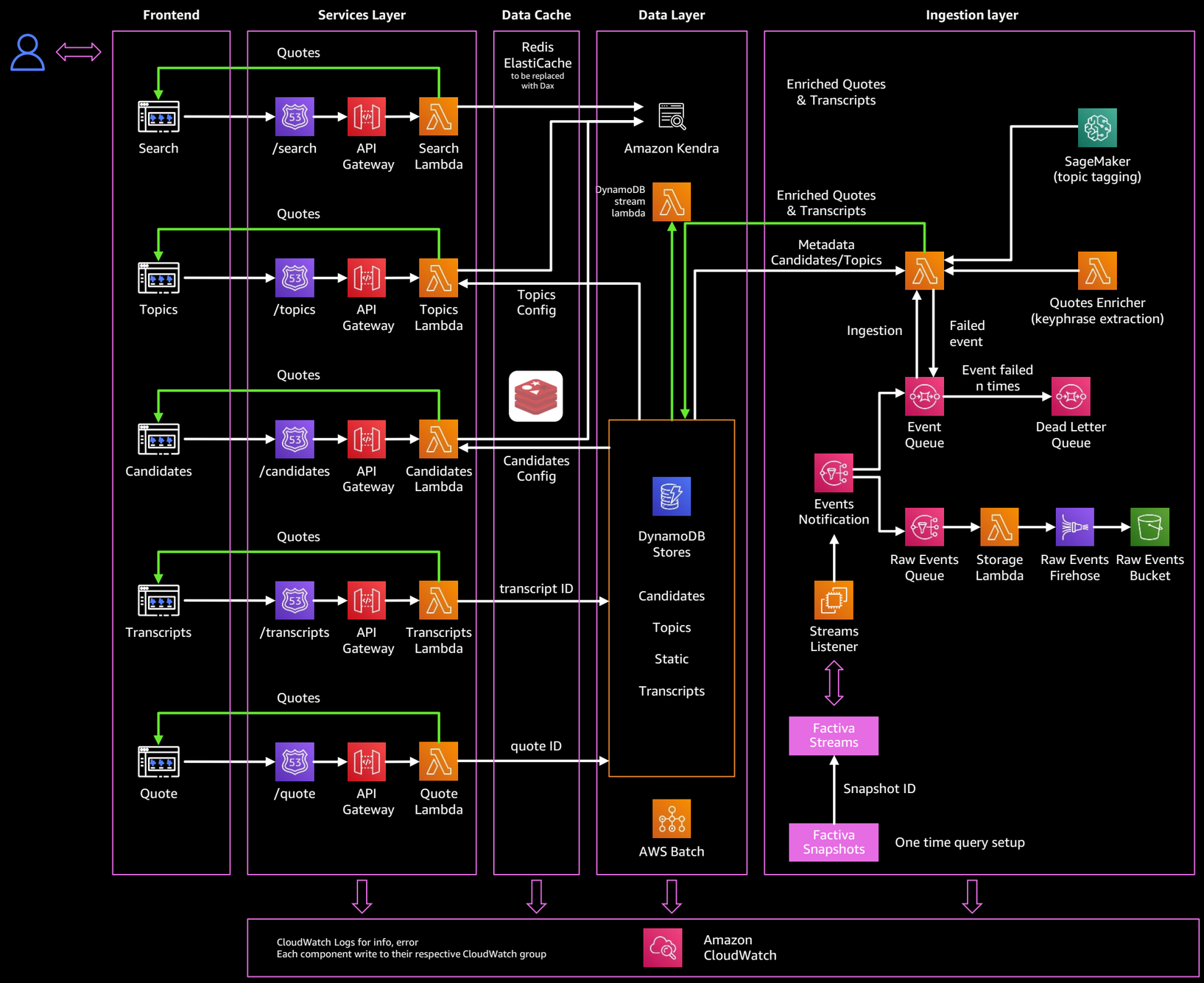
Transcripts are coming in on a constant basis, building for regular ETL and ingestion into Amazon Kendra was essential.

Configurable/flexible lists of candidates/topics

We built to be able to easily add entire candidates or topics into the tool and run data ingestion quickly.

Data caching for search performance

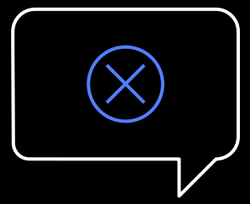
We kept searches cached for approx. an hour for repeat queries, but still had them update regularly as new documents came in.



Defining quality in our quotes

LESSONS LEARNED

Beyond keywords, the team identified more characteristics that helped us qualify results



Boost vector 1: is it a full quote?

We wanted to make sure that the quotes themselves contained as much of the full thought or stance as possible, instead of half transcribed crosstalk.

WALLACE: I I didn't ask sir, you're

TRUMP: Chris, that was the worst part

WALLACE: debating him, not me

TRUMP: of Obama

WALLACE: Let me ask my question.

Defining quality in our quotes

LESSONS LEARNED

Beyond keywords, the team identified more characteristics that helped us qualify results



Boost vector 2: is it timely? Original?

We played with several date ranges for de-duplication efforts to make sure we didn't clog the results.

ROMANS: President Trump bringing up his son, Barron, by name during that rally while attacking presidential candidate Elizabeth Warren.

KRISTEN WELKER : The White
most shameful political episod
impeachment abuse of power
his part, the President tweetin
be his 2020 rallying cry, the st
they're not after me, they're af
hours on stage, the longest spe
invoking his thirteen-year-old son to slam Elizabeth Warren.

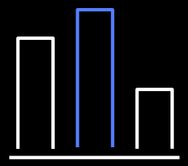
TRUMP: Pocahontas goes to the middle of Central Park or she's in Manhattan. You got me. And, people, I mean, I would I would have Barron go into Central Park and he would get a crowd just as large.

PRESIDENT DONALD TRUMP : Pocahontas goes to the middle of Central Park or whatever park, she's in Manhattan, you got me. And people, I mean, I could have I could have Barron Trump go into Central Park and he'd get a crowd that would be just as good. He's thirteen. He'd get a bigger crowd.

Defining quality in our quotes

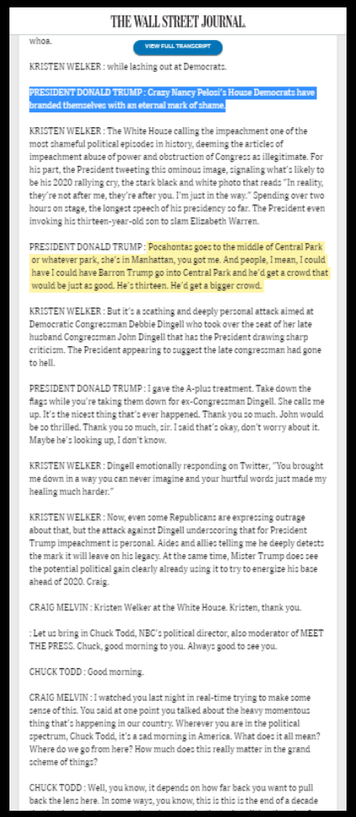
LESSONS LEARNED

Beyond keywords, the team identified more characteristics that helped us qualify results



Boost Vector 3: is it part of a full speech or appearance?

Transcripts that contained higher percentages of candidate quotes helped narrow down media appearances.



THE WALL STREET JOURNAL.

whoa.

VIEW FULL TRANSCRIPT

KRISTEN WELKER : while lashing out at Democrats.

PRESIDENT DONALD TRUMP : Crazy Nancy Pelosi's House Democrats have branded themselves with an eternal mark of shame.

KRISTEN WELKER : The White House calling the impeachment one of the most shameful political episodes in history, deeming the articles of impeachment abuse of power and obstruction of Congress as illegitimate. For his part, the President tweeting this ominous image, signaling what's likely to be his 2020 rallying cry, the stark black and white photo that reads "In reality, they're not after me, they're after you. I'm just in the way." Spending over two hours on stage, the longest speech of his presidency so far. The President even invoking his thirteen-year-old son to slam Elizabeth Warren.

PRESIDENT DONALD TRUMP : Pocahontas goes to the middle of Central Park or whatever park, she's in Manhattan, you got me. And people, I mean, I could have I could have Barron Trump go into Central Park and he'd get a crowd that would be just as good. He's thirteen. He'd get a bigger crowd.

KRISTEN WELKER : But it's a scathing and deeply personal attack aimed at Democratic Congressman Debbie Dingell who took over the seat of her late husband Congressman John Dingell that has the President drawing sharp criticism. The President appearing to suggest the late congressman had gone to hell.

PRESIDENT DONALD TRUMP : I gave the A-plus treatment. Take down the flags while you're taking them down for ex-Congressman Dingell. She calls me up. It's the nicest thing that's ever happened. Thank you so much. John would be so thrilled. Thank you so much, sir. I said that's okay, don't worry about it. Maybe he's looking up, I don't know.

KRISTEN WELKER : Dingell emotionally responding on Twitter, "You brought me down in a way you can never imagine and your hurtful words just made my healing much harder."

KRISTEN WELKER : Now, even some Republicans are expressing outrage about that, but the attack against Dingell underscoring that for President Trump impeachment is personal. Aides and allies telling me he deeply detests the mark it will leave on his legacy. At the same time, Mister Trump does see the potential political gain clearly already using it to try to energize his base ahead of 2020. Craig.

CRAIG MELVIN : Kristen Welker at the White House. Kristen, thank you.

: Let us bring in Chuck Todd, NBC's political director, also moderator of MEET THE PRESS. Chuck, good morning to you. Always good to see you.

CHUCK TODD : Good morning.

CRAIG MELVIN : I watched you last night in real-time trying to make some sense of this. You said at one point you talked about the heavy momentous thing that's happening in our country. Wherever you are in the political spectrum, Chuck Todd, it's a sad morning in America. What does it all mean? Where do we go from here? How much does this really matter in the grand scheme of things?

CHUCK TODD : Well, you know, it depends on how far back you want to pull back the lens here. In some ways, you know, this is this is the end of a decade



Taxonomy for Search & Browse

LESSONS LEARNED

Developing categories that speak the user's language

The great corpus provided by Factiva had their own taxonomy applied to the documents; but we had to take an archival service and put it in the language of the casual user.

2017-2020 view						2019/2020
Highlevel Topic	Subtopic	Relevance	Frequency	Importance	Popularity Rank within Highlevel Topic	Subtopic Keywords
Black Lives Matter	Black Lives Matter	0.295	506	1.000	0	['protester', 'violent anti - police hostility', 'civil right movement', 'racial justice', 'antifa', 'black caucus']
Business	Work-Life Balance	0.266	359	0.106	0	#N/A
Business	Workers Pay	0.289	30	0.098	1	['wage', 'bonus pay', 'american workers', 'payroll']
Business	Corporate Mega Mergers	0.292	192	0.076	2	['general motors']
Business	Capital Gains Tax	0.312	123	0.074	3	['capital tax']
Business	Small Business Lending	0.355	55	0.072	4	['small business loan', 'loan', 'sba']
Catastrophe	Accidents/Man-Made Disasters	0.295	182	0.287	0	['unrelated accident', 'natural disaster', 'fault']
Catastrophe	Natural Disasters/Catastrophes	0.284	470	0.235	1	['natural disaster', 'major emergency disaster declaration', 'fema']
Catastrophe	Geological Disasters	0.264	3	0.132	2	#N/A
Catastrophe	Building Fires/Explosions/Collapses	0.252	41	0.128	3	['building']
Catastrophe	Famine	0.265	39	0.109	4	['hunger']
Coronavirus	Coronavirus/Covid-19	0.277	88	0.500	0	['contagious virus', 'disease', 'new cure']
Corporate	Corporate Subsidies	0.305	428	1.000	0	['private company']
Crime	Torture	0.267	927	0.047	0	['waterboarding', 'enhanced interrogation', 'guantanamo bay']
Crime	Corruption	0.315	508	0.044	1	['corruption', 'corrupt broken political establishment', 'biden', 'people', 'vice president biden', 'ukraine']
Crime	Gangs	0.337	520	0.044	2	['gang', 'ms-13 gang member', 'horrible gang', 'ms-13']
Crime	Criminal Justice Reform	0.316	353	0.043	3	['criminal justice reform', 'juvenile justice', 'crime', 'law enforcement']
Crime	Fraud	0.288	362	0.031	4	['fraud', 'fraudulent crime', 'scam']
Discrimination	Racism	0.345	364	0.353	0	#N/A
Discrimination	Safe Spaces	0.254	170	0.136	1	#N/A
Discrimination	Religious Discrimination	0.262	74	0.124	2	['discriminatory policy', 'religious freedom', 'muslim ban']
Discrimination	Discrimination	0.299	100	0.117	3	['discrimination', 'lgbtq issue']
Discrimination	Age Discrimination	0.306	8	0.076	4	['age']
Disease	Pandemic	0.278	483	0.178	0	['pandemic', 'outbreak', 'ebola', 'infectious disease', 'quarantine', 'cdc people']
Disease	H1N1 Flu/Influenza	0.309	99	0.062	1	['swine flu', 'bad flu', 'bad pandemic', 'vaccine', 'people']
Disease	Animal Influenza	0.342	64	0.040	2	['flu', 'flu season', 'bad pandemic']
Disease	Outbreaks/Epidemics	0.260	141	0.037	3	#N/A
Disease	Aids/Hiv	0.268	27	0.036	4	['aids epidemic', 'aids', 'disease']
Domestic Policy	Impeachment	0.325	5510	0.174	0	['impeach trump', 'impeachment nonsense', 'president', 'mueller', 'people', 'democrats']
Domestic Policy	Congressional Term Limits	0.305	984	0.083	1	['congress people', 'senate', 'democrats', 'nancy pelosi']

The outcomes

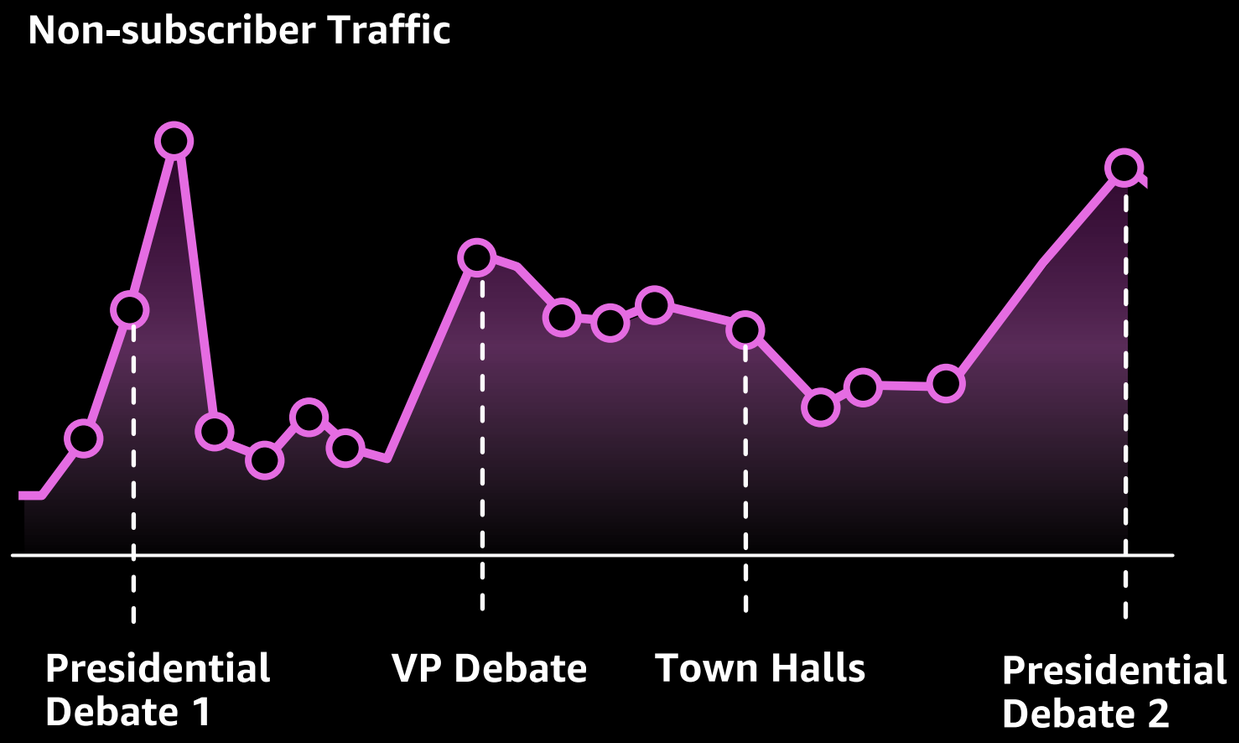
RESULTS AND LESSONS LEARNED

How did they use it?

1) WHEN
A decent second-screen experience during the debates, but the traffic really surged the morning after the first debate, when most of the analysis was published.

Other, more straight-forward debates let this shine as a second screen, so perhaps the chaotic nature of the first one.

Next-day users were also more likely to do more searches.



The outcomes

RESULTS AND LESSONS LEARNED

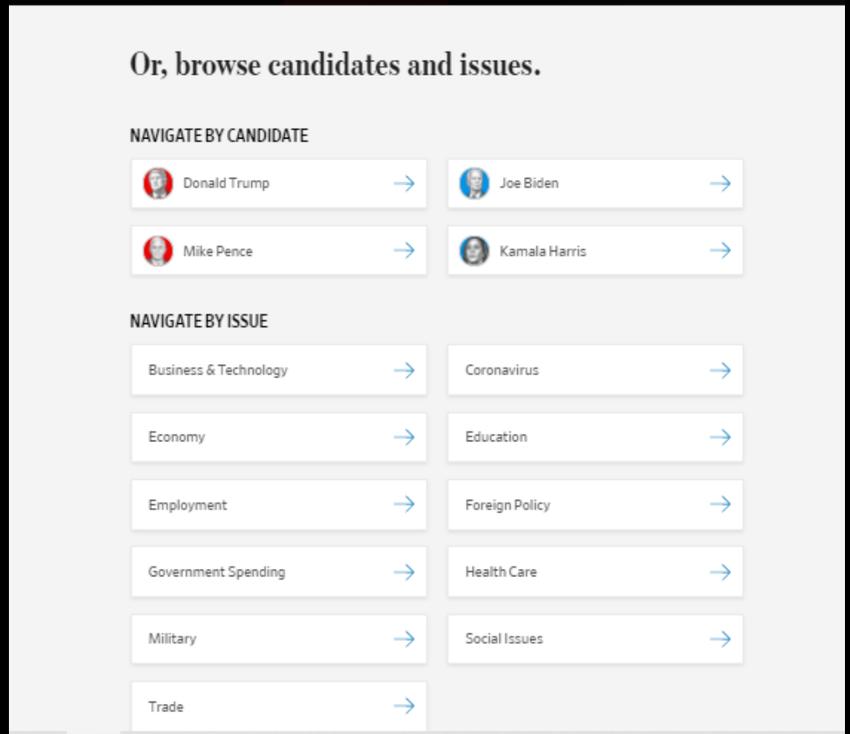
How did they use it?

2) WHICH PATH DID THEY CHOOSE?

Of the users that landed on our search tools, 70% did by direct search, 30% did by browsing by candidate or topic.

This was a bit by design: the first whole page load was the “Ask” button and associated features. We may seek to better balance these features out in the future.





The outcomes

RESULTS AND LESSONS LEARNED

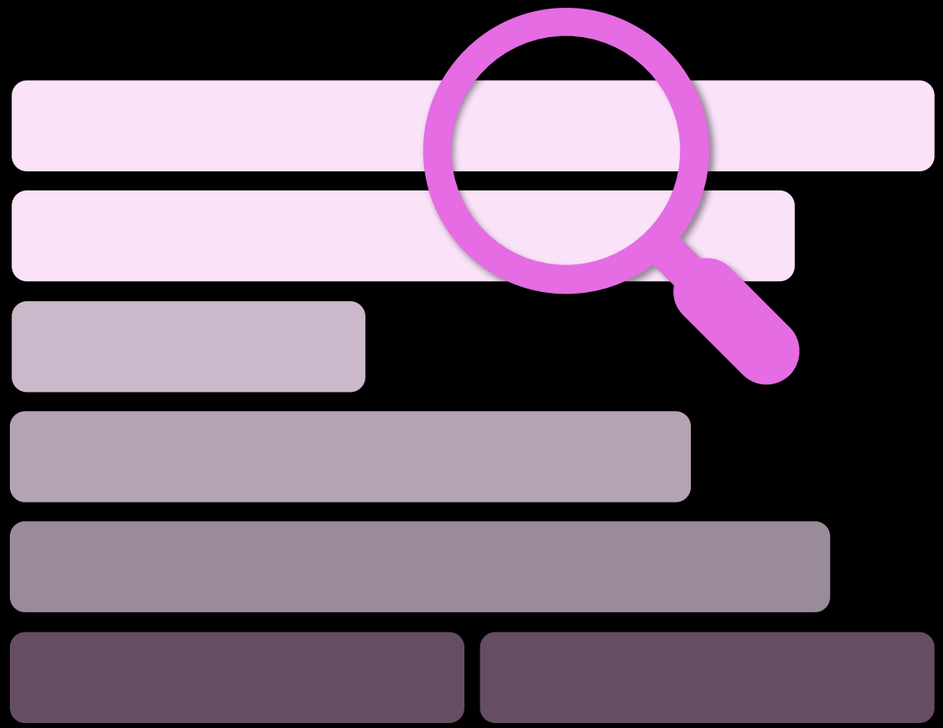
How did they use it?

3) WHAT DID THEY SEARCH FOR?

Users performed almost 15,000 different kinds of queries.

The main natural-language queries that were entered were based on the sample questions.

Still a mix of single keyword/keyphrase queries, generally pegged to the debate. Keywords were generally topic-based



The next steps: Where do we go from here?

RESULTS AND LESSONS LEARNED

Where do we see potential?

1) MORE USE CASES

As we have now built a knowledge base-type application for WSJ readers based on natural language query technology, there have been a lot of other potential uses that could be worth exploring.

2) WHERE WE WANT TO IMPROVE

There are some things we know we would like to go back and improve if this were to be the kind of tool we would do separate flavors for.

a) Better keyphrase detection vs. keyphrases

b) More direct control over how deep of a pool of results to return to the person performing the search (vs. the current top 100)

c) Better tools for raw data analysis and preparation

Call to action

Amazon Kendra Home Page:

https://aws.amazon.com/kendra/

Amazon Kendra Blogs:

https://aws.amazon.com/blogs/machine-learning/category/artificial-intelligence/amazon-kendra/

eGuides to help business case development & getting started with Amazon Kendra POC:

https://pages.awscloud.com/NAMER-devadopt-DL-Amazon-Kendra-eBook-7-Reasons-Why-2020-learn.html

https://pages.awscloud.com/NAMER-devadopt-DL-Amazon-Kendra-eBook-how-to-build-a-POC-2020-learn.html

Thank you!

Ramin Beheshti
Chief Product & Technology Officer
Dow Jones

Sheri Moran
Principal Solutions Architect
AWS

AWS re:Invent

© 2020, Amazon Web Services, Inc. or its affiliates. All rights reserved.



Please complete the session survey

AWS re:Invent

© 2020, Amazon Web Services, Inc. or its affiliates. All rights reserved.