Learning for the Long Tail

Train AI 2018

Marti Hearst
UC Berkeley

With Philippe Laban

How Many Distinct People Were Mentioned In Accounts of the Cambridge Analytical Scandal?

Audience Exercise!

How Many Distinct People Were Mentioned In Accounts of the Cambridge Analytical Scandal?

We find > 700

Why Handle The Long Tail?

How to Handle Rare Cases?

Why Handle Rare Cases?

Avoid Damaging Errors

Don’t erroneously say someone died!

Who Is John Allen?

In 2014, John Allen died in a plane downed by a Russian-made missile.

John Allen was also a retired U.S. General and Presidential Envoy in 2014.

Confusing these people would have been problematic!

John Allens are Plentiful!

Results **1 – 20** of **204**

John L. **Allen**, Jr. (Q1254163)
American journalist
22 statements, 10 sitelinks - 20:15, 8 May 2018

John R. **Allen** (Q615356)
United States Marine Corps general
25 statements, 14 sitelinks - 20:15, 8 May 2018

John S. **Allen** (Q6012096)
American university president, professor of astronomy
19 statements, 2 sitelinks - 20:15, 8 May 2018

John J. **Allen**, Jr. (Q1700644)
American politician
35 statements, 2 sitelinks - 20:15, 8 May 2018

John P. **Allen** (Q16197280)
Canadian country/rock/bluegrass fiddler
10 statements, 1 sitelink - 20:15, 8 May 2018

John Hensleigh **Allen** (Q15072629)
politician
23 statements, 1 sitelink - 20:15, 8 May 2018

John James **Allen** (Q198304)
American lawyer and judge
26 statements, 4 sitelinks - 20:15, 8 May 2018

John Mills **Allen** (Q1701104)
American politician
24 statements, 2 sitelinks - 20:15, 8 May 2018

Why Do We Care?

newsLens

newsLens Search...

STORY CATEGORIES

- World
- Business
- People

World

Syria Talks 1	Trump Lawyer Michael C.. 2	Israeli Police 4	Salisbury Nerve Agent A.. 5
Rohingya Crisis	FBI Director	Nelson Mandela's Ex-Wife..	Bill Cosby Trial
Saudi-Led Coalition	Obamacare Repeal	North Korea	Junior Hockey Team's Bus..
Americas Summit ' Failure	Brexit Talks	Trump's EPA Budget Prop..	Catalan Independence
Same-Sex Marriage	EU Air Traffic Control Ag..	Deputy National Security ..	Solo Alps Ski Trip
Putin's Oligarch Allies	Greek Fighter Jet Crashes	Former Brazil President L..	Missouri Governor's Bid

Read the story

Business

Cambridge Analytica Sca.. 3	Migrant Crisis	Co-Pilot In Algeria's Wors..	US STOCKS-Wall St
Trump Trade War	VW Scandal	Its Self-Driving Cars	

Read the story

Cambridge Analytica Scandal

Apr. 2018

Apr. 10 2018 — Facebook CEO Mark Zuckerberg is about to testify to Congress
Open details

Apr. 06 2018 — EU says Facebook confirmed data of 2.7 million Europeans 'improperly shared'
Open details

Apr. 04 2018 — Zuckerberg will testify before Congress on April 11
Open details

Mar. 27 2018 — Cambridge Analytica whistleblower: Vote Leave 'cheating' may have swayed Brexit referendum - video
Open details

Mar. 27 2018 — Facebook CEO Mark Zuckerberg will reportedly testify before Congress
Open details

Mar. 27 2018 — Zuckerberg will not appear before UK parliament committee
Open details

Mar. 26 2018 — FTC Confirms Facebook Probe Over Privacy Practices
Open details

Cambridge Analytica Scandal

Actors

Search for an actor...

37 director | 36 president | 29 senator | 24 professor

Mark Zuckerberg • 2138 mentions
ceo | executive | chief executive | founder

Aleksandr Kogan • 773 mentions
researcher | professor | scientist | cambridge

Donald Trump • 741 mentions
candidate | president | presidential candidate | us president

Alexander Nix • 550 mentions
ceo | executive | chief executive | chief officer

Christopher Wylie • 494 mentions
employee | former employee | whistleblower | former cambridge analytica employee

Shahmir Sanni • 242 mentions
volunteer | whistleblower | blower | - blower

Damian Collins • 235 mentions
chair | chairman | mp | head

< 1/126 >

Apr. 10 2018 — 3
Facebook CEO Mark Zuckerberg is about to testify to Congress
Open details

Apr. 06 2018 — 3
EU says Facebook confirmed data of 2.7 million Europeans 'improperly shared'
Open details

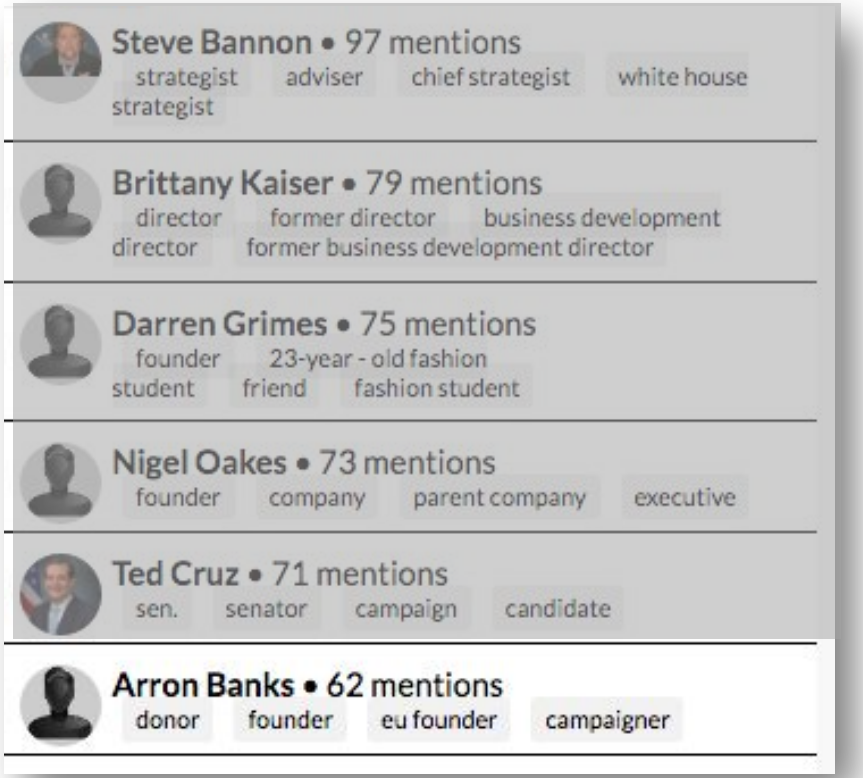
Apr. 04 2018 — 7
Zuckerberg will testify before Congress on April 11
Open details

Mar. 27 2018 — 4
Cambridge Analytica whistleblower: Vote Leave 'cheating' may have swayed Brexit referendum - video
Open details

Mar. 27 2018 — 5
Facebook CEO Mark Zuckerberg will reportedly testify before Congress
Open details

Mar. 27 2018 — 4
Zuckerberg will not appear before UK parliament committee
Open details

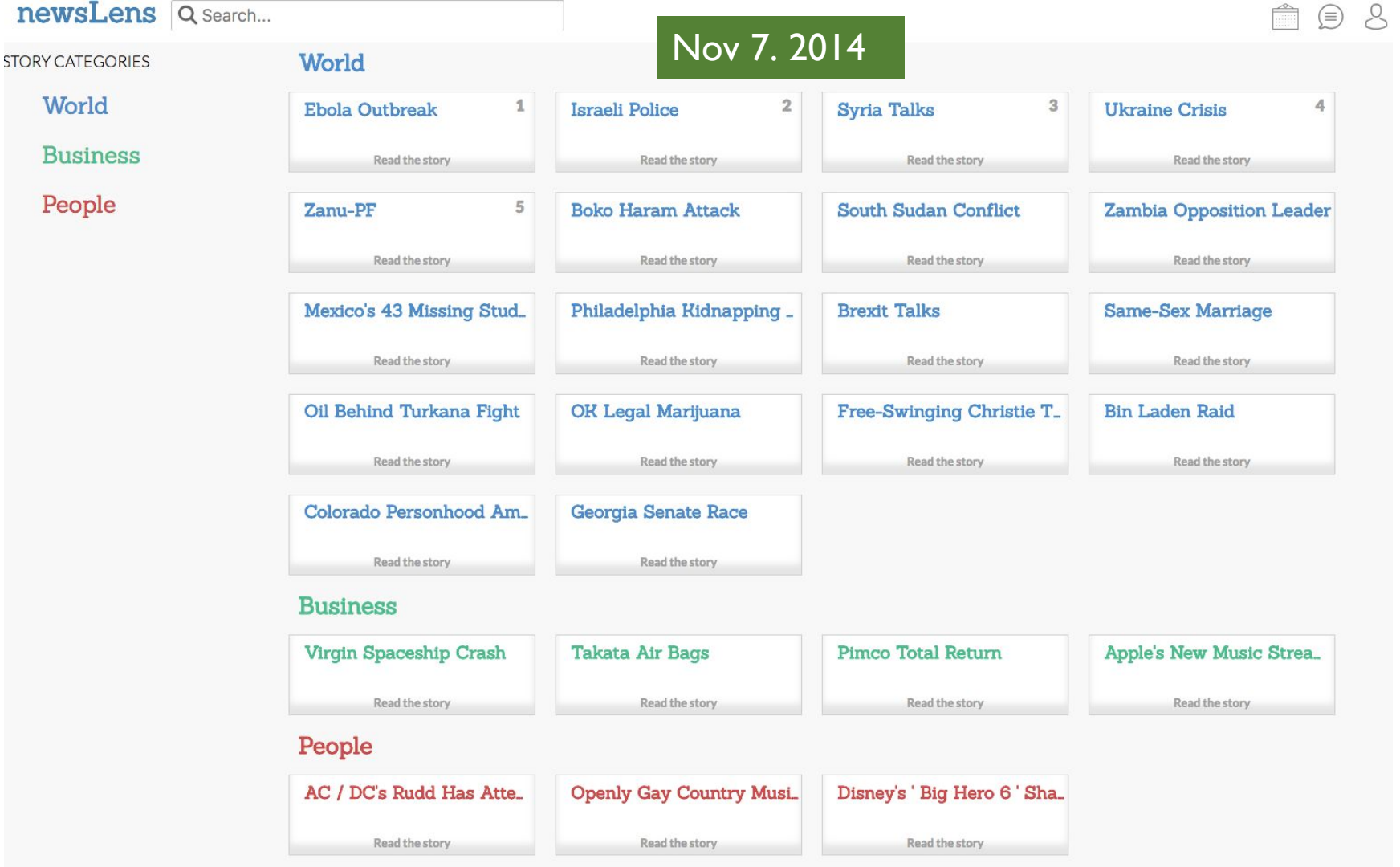
Mar. 26 2018 — 7
FTC Confirms Facebook Probe Over Privacy Practices
Open details


Steve Bannon • 97 mentions
strategist adviser chief strategist white house strategist
Brittany Kaiser • 79 mentions
director former director business development director former business development director
Darren Grimes • 75 mentions
founder 23-year - old fashion student friend fashion student
Nigel Oakes • 73 mentions
founder company parent company executive
Ted Cruz • 71 mentions
sen. senator campaign candidate
Arron Banks • 62 mentions
donor founder eu founder campaigner


Arron Banks
businessman
62 mentions
Open newsLens profile
Open Wikidata entry
X
Introductions
funder
its
a key financier in the Brexit campaign
who has funded the Leave
British businessman
A prominent campaigner to leave the European Union
EU
the insurance magnate
who campaigned for Britain to leave the European Union in a 2016 referendum
EU founder

NewsLens Shows The Long Tail in News

- Represent disparate articles as stories
- Can go back in time to see past stories
- Shows many points of view at once
- Shows the less famous people in context


newsLens
Search...
Nov 7. 2014
STORY CATEGORIES
World
Business
People
World
Ebola Outbreak 1
Read the story
Israeli Police 2
Read the story
Syria Talks 3
Read the story
Ukraine Crisis 4
Read the story
Zanu-PF 5
Read the story
Boko Haram Attack
Read the story
South Sudan Conflict
Read the story
Zambia Opposition Leader
Read the story
Mexico's 43 Missing Stud..
Read the story
Philadelphia Kidnapping ..
Read the story
Brexit Talks
Read the story
Same-Sex Marriage
Read the story
Oil Behind Turkana Fight
Read the story
OK Legal Marijuana
Read the story
Free-Swinging Christie T..
Read the story
Bin Laden Raid
Read the story
Colorado Personhood Am..
Read the story
Georgia Senate Race
Read the story
Business
Virgin Spaceship Crash
Read the story
Takata Air Bags
Read the story
Pimco Total Return
Read the story
Apple's New Music Strea..
Read the story
People
AC / DC's Rudd Has Atte..
Read the story
Openly Gay Country Musi..
Read the story
Disney's ' Big Hero 6 ' Sha..
Read the story

Hawaii volcano eruption - n

newslens.berkeley.edu/story/20880

120%

Hawaii Volcano Eruption

May 2018

Actors

Search for an actor...

9 resident | 5 geologist | 4 spokeswoman | 3 scientist

Janet Babb • 47 mentions
geologist | spokeswoman | hvo geologist | usgs geologist

Darryl Oliveira • 42 mentions
director | civil defense director | chiefl | hawaii county director

Jeremy Wilson • 17 mentions

Jessica Gauthier • 16 mentions
realtor

David Ige • 14 mentions
governor | # kilauea # volcano-- governor | gov. | d

Jordan Sonner • 12 mentions

Tulsi Gabbard • 12 mentions
rep. | u.s. representative | hawaii | representative

< 1 / 14 >

May 07 2018 (4)
Dramatic footage shows lava engulfing car in Hawaii - video
Open details
- Dramatic footage shows lava engulfing car in Hawaii - video
- Dramatic video shows lava bursting from the ground in a volcano eruption in Hawaii
- 'Fountains of lava': Dramatic video shows scale of Mount Kilauea eruption in Hawaii (PHOTO,VIDEO)
- Aerial footage shows volcanic lava destroying homes in Hawaii - video

May 07 2018 (5)
Hawaii volcano destroys dozens of homes
Open details

May 06 2018 (5)
Thousands evacuated after Hawaii's Kilauea volcano erupts
Open details

Apr. 29 2018

447 days with no activity

Feb. 05 2017

newslens.berkeley.edu:4005/article/5aec711acb2c981d83a89513

Peter Thiel

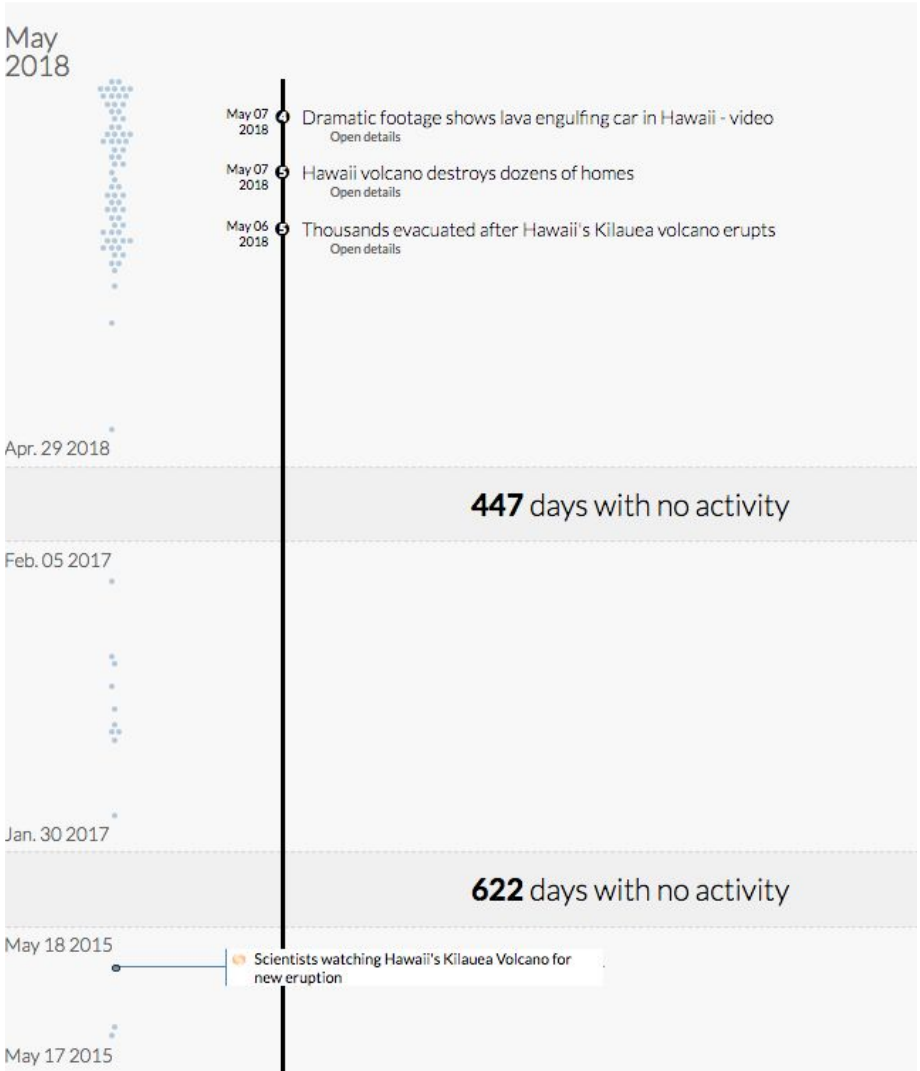
1200 mentions

in 117 stories

Actor

Hulk Hogan's Gawker La.. Read the story	His Silicon Valley Read the story	Republican Convention Read the story
Cambridge Analytica Sca.. Read the story	FBI Director Read the story	Silicon Valley's Famous S.. Read the story
New Zealand Citizenship Read the story	Republican Debate Read the story	Republican Debate Read the story
Donald Trump's Behavior .. Read the story	Silicon Valley's Liberal Bu.. Read the story	Trump Administration Read the story
Trump Tower Meeting Read the story	Google Employee Anti-Di.. Read the story	Up To 2 Dozen Parties Ey.. Read the story
Elon Musk's New Company Read the story	Facebook Shareholder Vote Read the story	California Senate Leader .. Read the story
Obama Administration Read the story	Trump's Immigration Ord.. Read the story	Clayton's SEC Nomination Read the story

Hawaii Volcano Eruption



Hawaii Volcano Eruption

Actors

Search for an actor...

9 resident | 5 geologist | 4 spokeswoman | 3 scientist

Janet Babb • 47 mentions
geologist | spokeswoman | hvo geologist | usgs geologist

Darryl Oliveira • 42 mentions
director | civil defense director | chief | hawaii county director

Jeremy Wilson • 17 mentions

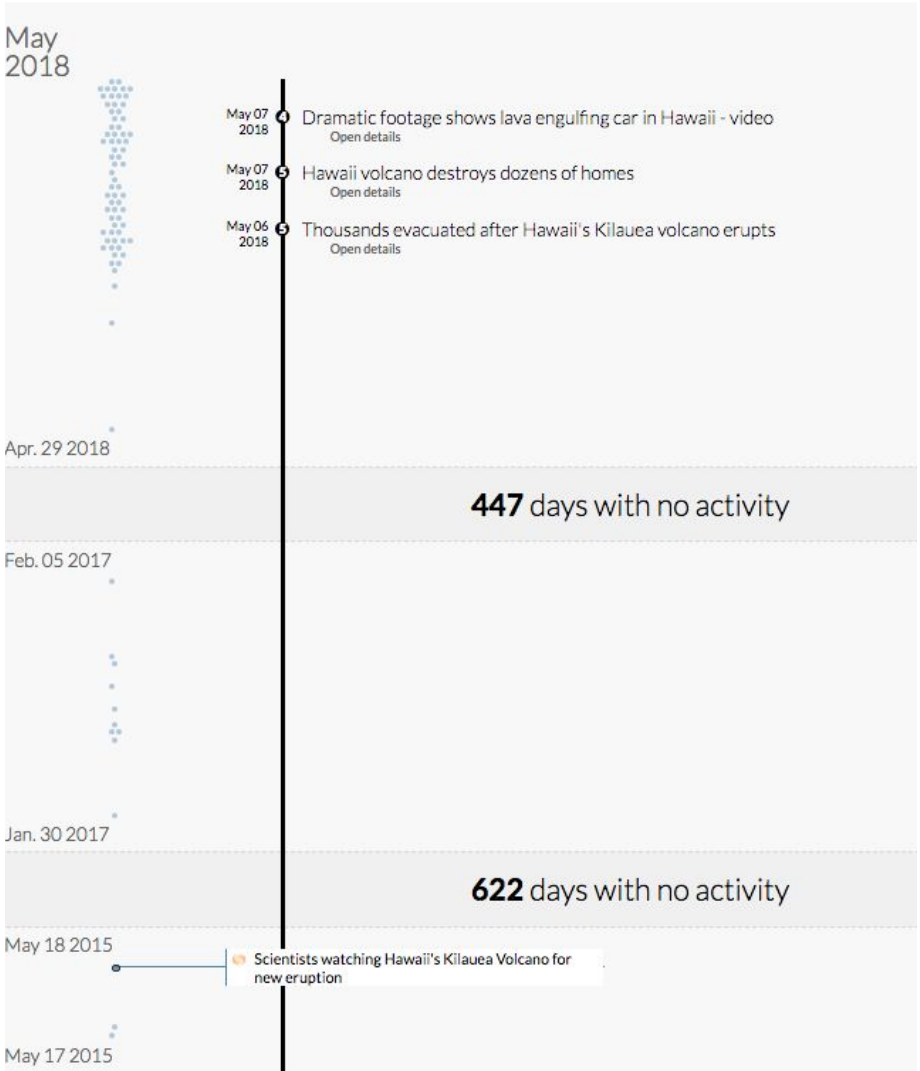
Jessica Gauthier • 16 mentions
realtor

David Ige • 14 mentions
governor | # kilauea # volcano-- governor | gov. | d

Jordan Sonner • 12 mentions

Tulsi Gabbard • 12 mentions
rep. | u.s. representative | hawaii | representative

< 1 / 14 >



Why Handle The Long Tail?

How to Handle Rare Cases?

Goal: Entity Linking People to Knowledge Base Entries

Journalists Write Using "Introductions"

A properly sourced news story states some background for anyone who is mentioned.

"*Vatican analyst* John Allen"

We use an "introduction" as **very precise evidence** for a person's identity.

Alternatives / Redundancy Reduce Errors

The first introduction below suggests someone who might not be known on wikidata.

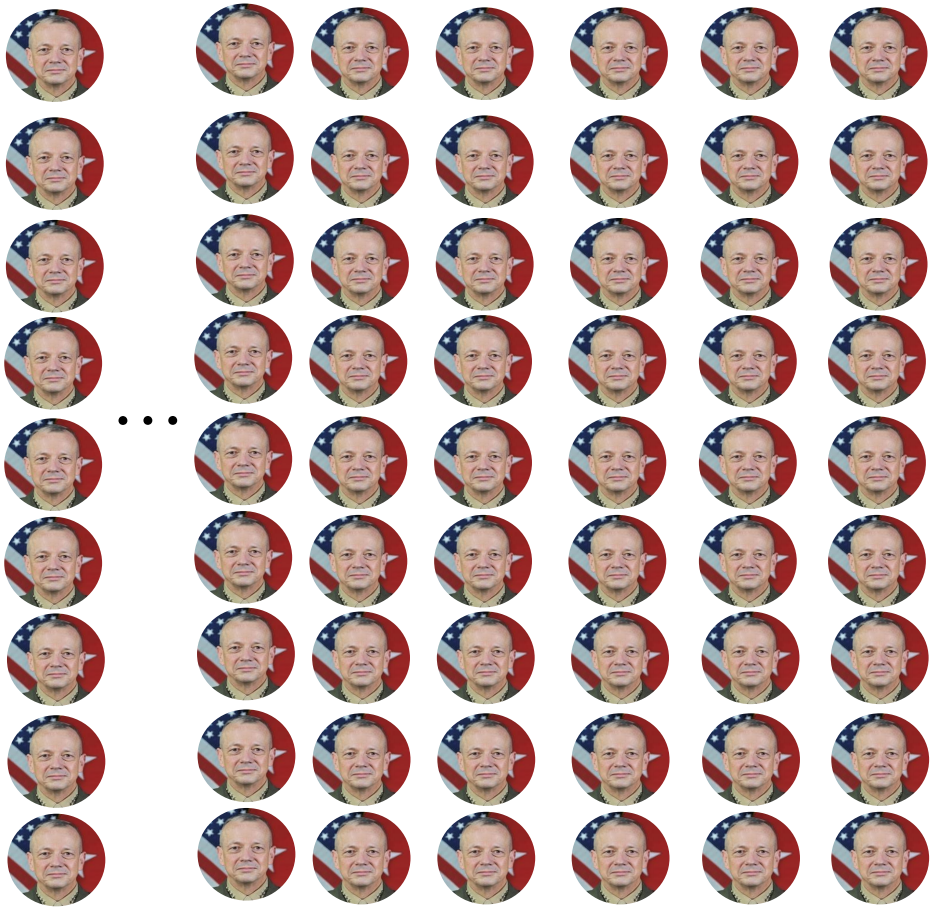
"another victim "
"Oscar winning actress"
"the American actress"

The additional introductions enrich the description.

Marine General

The top U.S. and NATO commander in Afghanistan
Marine Corps Gen.

CNN's Vatican analyst




The Easier Case

South African
Author,
Not on Wikidata

M17 Victim,
not on Wikidata

The Harder Case

?

"our much-loved colleague"

An Amsterdam-based British lawyer

one of the last two of the ten UK victims to be identified

(The Telegraph, July 20, 2017)

?

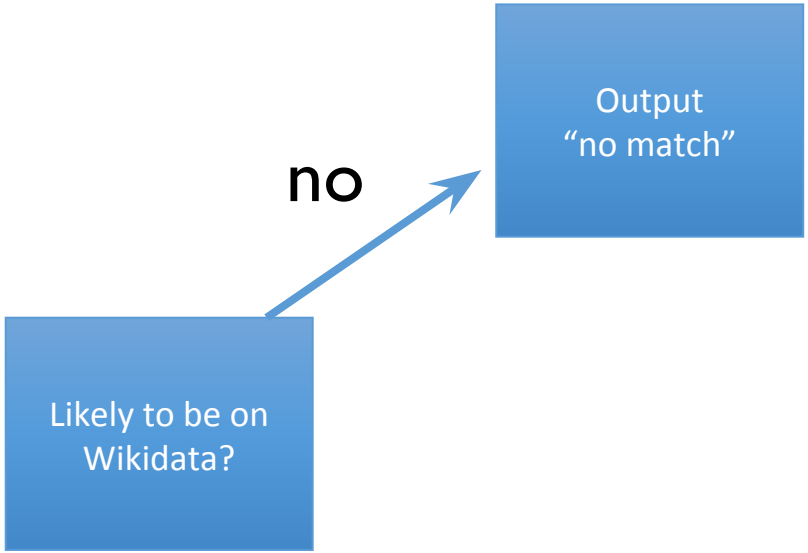
Not likely to link

"our much-loved colleague"

An Amsterdam-based British lawyer

one of the last two of the ten UK victims to be identified

(The Telegraph, July 20, 2017)

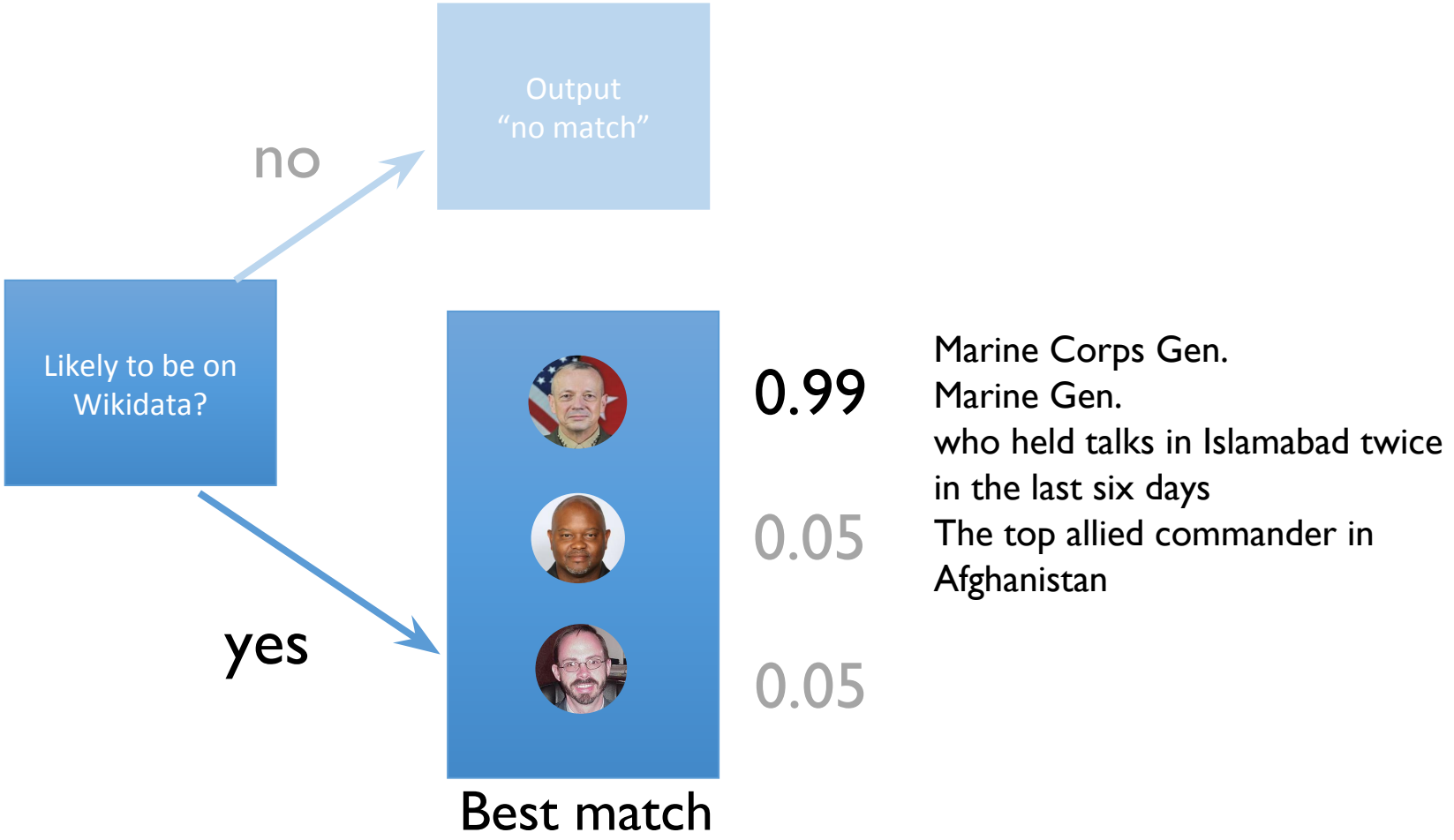


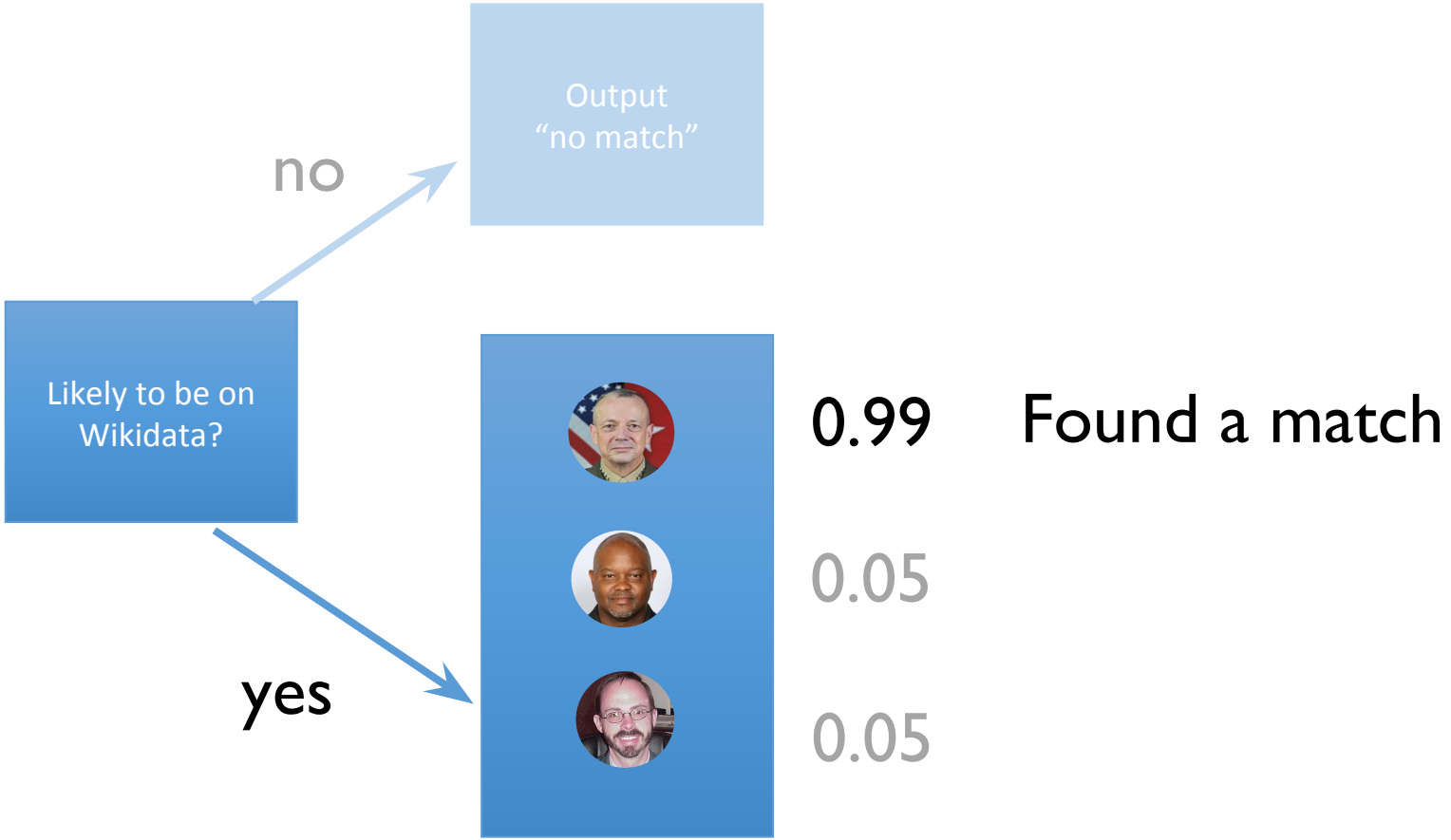
Classifier trained
on introductions

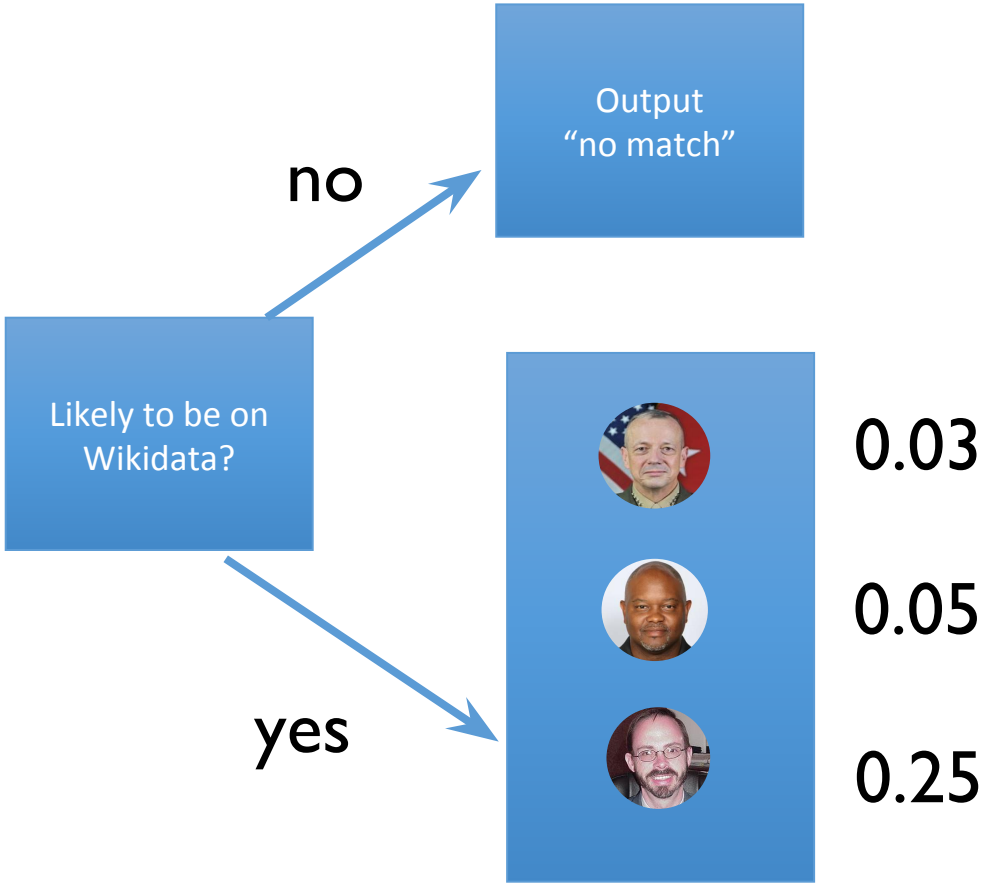
“our much-loved colleague”

?

The top U.S. and NATO commander in Afghanistan
Marine Corps Gen.
Marine Gen.
who held talks in Islamabad twice in the last six days
The top allied commander in Afghanistan


Output
“no match”
no
Likely to be on Wikidata?
yes
0.99
0.05
0.05
Marine Corps Gen.
Marine Gen.
who held talks in Islamabad twice in the last six days
The top allied commander in Afghanistan
Best match


Output
“no match”
no
Likely to be on Wikidata?
yes
0.99
Found a match
0.05
0.05



“author of a biography of Desmond Tutu”

Nothing above threshold, No match

A Challenging Dataset

- Extracted 1000 (article, people) pairs
- Each person has 2+ potential entries in Wikidata
- Manually labeled the pairs for ground truth
- 15% (293) are instances that do not match to WD

RESULTS

Easy cases: When there is a match in Wikidata

84.5% correct (91.5 using position in Wikidata)
(compared to a popular commercial system of 82.1)

Hard cases: When **no match** in Wikidata

50.8% correct
(compared to a popular commercial system of **14.5**)

Summary: How We Handled the Long Tail

- Identified evidence that is:
  - High precision
  - Contextually determined
  - Can often label with just one instance, but also can be improved when additional evidence is present.
- Two-phase classification:
  - First decide yes-or-no for match to KB
  - This greatly improves accuracy on no-match cases.

Thank you!

newslens.berkeley.edu

Marti Hearst and Phillipe Laban
UC Berkeley