



# Quicksilver

*Training an ML system to generate draft  
Wikipedia articles and Wikidata entries  
simultaneously*

John Bohannon and Vedant Dharnidharka

∴ PRIMER



# Precision vs. Recall



WIKIPEDIA  
The Free Encyclopedia

Main page  
Contents  
Featured content  
Current events  
Random article  
Donate to Wikipedia  
Wikipedia store

Interaction  
Help  
About Wikipedia  
Community portal  
Recent changes  
Contact page

Tools  
What links here  
Related changes  
Upload file  
Special pages  
Permanent link  
Page information  
Wikidata item

Not logged in Talk Contributions Create account Log in

Article Talk

Read Edit View history

Search Wikipedia

## Aleksandr Kogan

From Wikipedia, the free encyclopedia

*This article is about the data scientist involved with Cambridge Analytica. For the Russian artist, see [Aleksandr Kogan \(artist\)](#).*

**Aleksandr Kogan** (born 1985/86), who has also briefly used the name **Dr Spectre**,<sup>[3]</sup> is a Moldovan-born data scientist, who is known for having developed the app that allowed [Cambridge Analytica](#) to collect personal details of 80 million Facebook users. He works as a [research associate](#) at the [University of Cambridge](#).<sup>[4]</sup>

**Contents** [hide]

- Early life
- Career
- Personal life
- References

### Early life [edit]

Kogan was born in what was then the [Moldavian SSR](#) in the [USSR](#) (now [Moldova](#)).<sup>[5][3]</sup> His father is Jewish.<sup>[6]</sup> He lived in Moscow before moving to the United States aged seven.<sup>[3]</sup> He earned a bachelor's degree from the [University of California, Berkeley](#) in 2008, and a PhD from the [University of Hong Kong](#) in 2011.<sup>[3][7]</sup>

### Career [edit]

Kogan had a post-doctoral fellowship at the [University of Toronto](#) before moving to Cambridge.<sup>[7]</sup> He has been a lecturer in the Department of Psychology at

<b>Aleksandr Kogan</b>	
<b>Born</b>	1985/1986 (age 32–33) <sup>[1]</sup> <div>Moldavian SSR, Soviet Union (now Moldova)</div>
<b>Nationality</b>	American <sup>[2]</sup>
<b>Education</b>	University of California, Berkeley (BA) University of Hong Kong (MA, PhD)
<b>Occupation</b>	CEO of Philometrics, psychologist and data scientist, University of Cambridge



# Precision vs. Recall

## References [\[ edit \]](#)

- <sup>1</sup> <sup>a</sup> <sup>b</sup> <sup>c</sup> O'Sullivan, Donie. "Scientist at center of data controversy says Facebook is making him a scapegoat" [↗](#). *cnn.com*. Retrieved 21 March 2018.
- <sup>2</sup> <sup>a</sup> Horwitz, Josh. "Cambridge Analytica's key researcher says his work's impact on elections is "not a real worry" " [↗](#). *Quartz*. Retrieved 26 April 2018.
- <sup>3</sup> <sup>a</sup> <sup>b</sup> <sup>c</sup> <sup>d</sup> <sup>e</sup> <sup>f</sup> "Who is Dr Aleksandr Kogan, the Cambridge academic accused of misusing Facebook data?" [↗](#). *varsity.co.uk*. Retrieved 21 March 2018.
- <sup>4</sup> <sup>a</sup> <sup>b</sup> "Dr Aleksandr Kogan :: Cambridge Neuroscience" [↗](#). *www.neuroscience.cam.ac.uk*. Retrieved 21 March 2018.
- <sup>5</sup> <sup>a</sup> <sup>b</sup> <sup>c</sup> <sup>d</sup> Weaver, Matthew (21 March 2018). "Facebook scandal: I am being used as scapegoat – academic who mined data" [↗](#). *the Guardian*. Retrieved 21 March 2018.
- <sup>6</sup> <sup>a</sup> <sup>b</sup> Aleksandr Kogan speaks! [↗](#). News From Underground. March 21, 2018.
- <sup>7</sup> <sup>a</sup> <sup>b</sup> Adams, Richard (20 March 2018). "Cambridge University asks Facebook for evidence about role of academic" [↗](#). *the Guardian*. Retrieved 21 March 2018.
- <sup>8</sup> <sup>a</sup> "Statement from the University of Cambridge about Dr Aleksandr Kogan" [↗](#). University of Cambridge. 23 March 2018.
- <sup>9</sup> <sup>a</sup> "Philometrics" [↗](#). Retrieved 14 April 2018.
- <sup>10</sup> <sup>a</sup> "Aleksandr Kogan: 5 Fast Facts You Need to Know" [↗](#). Heavy.com. Retrieved 14 April 2018.
- <sup>11</sup> <sup>a</sup> "Scientist: Didn't know data used to target voters" [↗](#). CNN.
- <sup>12</sup> <sup>a</sup> "Aleksandr Kogan: The link between Cambridge Analytica and Facebook" [↗](#). *CBSnews*. Retrieved 25 April 2018.
- <sup>13</sup> <sup>a</sup> Cohen, Marshall (21 March 2018). "Cambridge Analytica researcher touted data-mining in Russia speech" [↗](#). *money.cnn.com*. Retrieved 22 March 2018.
- <sup>14</sup> <sup>a</sup> Cadwalladr, Carole (17 March 2018). "Cambridge Analytica: links to Moscow oil firm and St Petersburg university" [↗](#). *theguardian.com*. Retrieved 22 March 2018.



# Precision vs. Recall

Name	Occupation	Employer	Gender	Wikipedia Word Count ↑	News Mentions	News Claims	Events
<a href="#">Karen R. Lips</a>	Biologist	the University of Maryland	Female	0	141	19	2
<a href="#">Maya Koronyo-Hamaoui</a>	Scientist	the Cedars-Sinai Medical Center	Female	0	137	29	2
<a href="#">Yuka Sasaki</a>	Researcher	Brown University	Female	0	212	11	2
<a href="#">Jane Mendle</a>	Writer	Cornell University	Female	0	163	13	2
<a href="#">Barbara A Cohn</a>	Researcher	the Public Health Institute	Female	0	137	19	2
<a href="#">Jenny C. A. Read</a>	Vision Scientist	Newcastle University	Female	0	160	11	3

# **An Unsupervised Approach to Biography Production using Wikipedia**

Fadi Biadisy, Julia Hirschberg, Elena Filatova

Columbia University

2008

We describe an unsupervised approach to multi-document sentence-extraction based summarization for the task of producing biographies. We utilize Wikipedia to automatically construct a corpus of biographical sentences and TDT4 to construct a corpus of non-biographical sentences. We build a biographical-sentence classifier from these corpora and an SVM regression model for sentence ordering from the Wikipedia corpus. We evaluate our work on the DUC2004 evaluation data and with human judges. Overall, our system significantly outperforms all systems that participated in DUC2004, according to the ROUGE-L metric, and is preferred by human subjects.

# **Generating Wikipedia by Summarizing Long Sequences**

Fadi

Peter J. Liu, Mohammad Saleh, Etienne Pot, Ben Goodrich, Ryan Sepassi, Lukasz Kaiser, Noam Shazeer

Google Brain

2018

We show that generating English Wikipedia articles can be approached as a multi-document summarization of source documents. We use extractive summarization to coarsely identify salient information and a neural abstractive model to generate the article. For the abstractive model, we introduce a decoder-only architecture that can scalably attend to very long sequences, much longer than typical encoder-decoder architectures used in sequence transduction. We show that this model can generate fluent, coherent multi-sentence paragraphs and even whole Wikipedia articles. When given reference documents, we show it can extract relevant factual information as reflected in perplexity, ROUGE scores and human evaluations.



# Google Brain's output...

## Wings Over Kansas

---

From Wikipedia, the free encyclopedia

**Wings Over Kansas.com** is an [aviation website](#) founded in 1998 by [Carl Chance](#) owned by [Chance Communications, Inc.](#)<sup>[a]</sup> to provide information and entertainment to aviation enthusiasts and professionals worldwide. The web site is based in [Wichita, Kansas](#), known as the "Air Capital of the World" due to the many aircraft manufacturers located there. In 2003, the site was upgraded to a data-based web site to better serve the needs of its members. "Wings Over Kansas" has grown steadily and as of 2009 draws over a quarter of a million visitors yearly from over 125 countries.<sup>[1]</sup>

### OUTPUT:

== Wings Over Kansas

==wings over kansas is the best aviation history website i have encountered on the world wide web . it is informative , entertaining , provides ever changing content , and is populated with the true voices of the mainstream aviation community . there is no better place to see where aviation has been and where it is going . centered in the midst of the greatest producer of aircraft in the world , wings over kansas reflects that aviation community to the rest of the world .

# Karen R. Lips

Summary Events Claims

## Summary

**Karen R. Lips** is a biologist at the University of Maryland.

In the winter of 1993, Lips was a PhD student working on her dissertation in Costa Rica when she first observed the then-unknown disease wreak havoc on local frog populations. In some of her studies, she focuses not on skin secretions, but on the genes involved in the frog immune system. Lips' work more than 20 years ago eventually contributed to the identification of the chytrid fungus as the primary cause of frog decline worldwide.

### Associated With

[Vance T. Vredenburg](#)  
[Joyce E Longcore](#)  
[Matthew C Fisher](#)  
[An Martel](#)  
[Samuel T. Turvey](#)

## References

Title ↓	Source	Date
<a href="#">To save salamanders from killer fungus, scientists hit the ground running</a>	Mother Nature Network	04-19-2017
<a href="#">Scientific Journal Launches "Conservation From The Front Lines" To Highlight Human Side Of Research</a>	IFLScience	02-11-2018
<a href="#">How the pet trade is killing frogs — and the genetic sleuthing that uncovered it</a>	The Mercury News	05-09-2018
<a href="#">Exotic pet trade linked to invasive fungus that's killing frogs globally</a>	Washington Post	05-09-2018
<a href="#">A Few Species of Frogs That Vanished May Be on the Rebound</a>	New York Times	03-28-2018

## Events

Title	Summary	Date	Score	Relevance ↓
<a href="#">How the pet trade is killing frogs — and the genetic sleuthing that uncovered it</a>	Genetic sleuthing has revealed the source the deadly fungus that is killing frogs and other amphibians around the world: the Korean peninsula. Scientists attempt to isolate the chytrid fungus <i>Batrachochytrium dendrobatidis</i> from amphibians which have died from chytridiomycosis in the Pyrenees. [Credit: Dirk Schmeller] While there had been a growing number of reports pointing towards Asia as the source of these amphibian pathogens, only recently has it been possible to do the genetic work needed to more clearly identify the source, said Karen Lips of the University of Maryland, a world leader in tropical biology, amphibian declines and conservation policy making. She worries that we might see new lineages of the fungus introduced into the U.S., because so many different genetic lineages are circulating through the international amphibian trade.	05-09-2018	14.27	1.62
<a href="#">A Few Species of Frogs That Vanished May Be on the Rebound</a>	To test that possibility, she and her colleagues collected skin secretions from captive frogs in the Maryland Zoo. To determine how much good skin secretions do, Dr. Lips said, it would be necessary to infect frogs and see whether stronger skin secretions actually keep more frogs alive. In some of her studies, she focuses not on skin secretions, but on the genes involved in the frog immune system. "Their genes are going crazy, but it doesn't matter," Dr. Lips said. It's possible that the immune system of frogs will turn out to be a key to the rebound of some species, or their skin secretions — or both.	03-28-2018	13.60	1.49



# We need a knowledge base!

## Neural Text Generation from Structured Data with Application to the Biography Domain

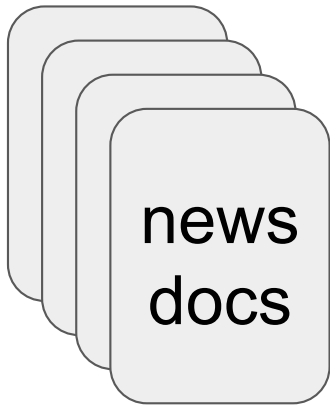
Remi Lebret, David Grangier, Michael Auli  
Facebook AI

2016

This paper introduces a neural model for concept-to-text generation that scales to large, rich domains. We experiment with a new dataset of biographies from Wikipedia that is an order of magnitude larger than existing resources with over 700k samples. The dataset is also vastly more diverse with a 400k vocabulary, compared to a few hundred words for Weathervg or Robocup. Our model builds upon recent work on conditional neural language model for text generation. To deal with the large vocabulary, we extend these models to mix a fixed vocabulary with copy actions that transfer sample-specific words from the input database to the generated output sentence. Our neural model significantly out-performs a classical Kneser-Ney language model adapted to this task by nearly 15 BLEU.



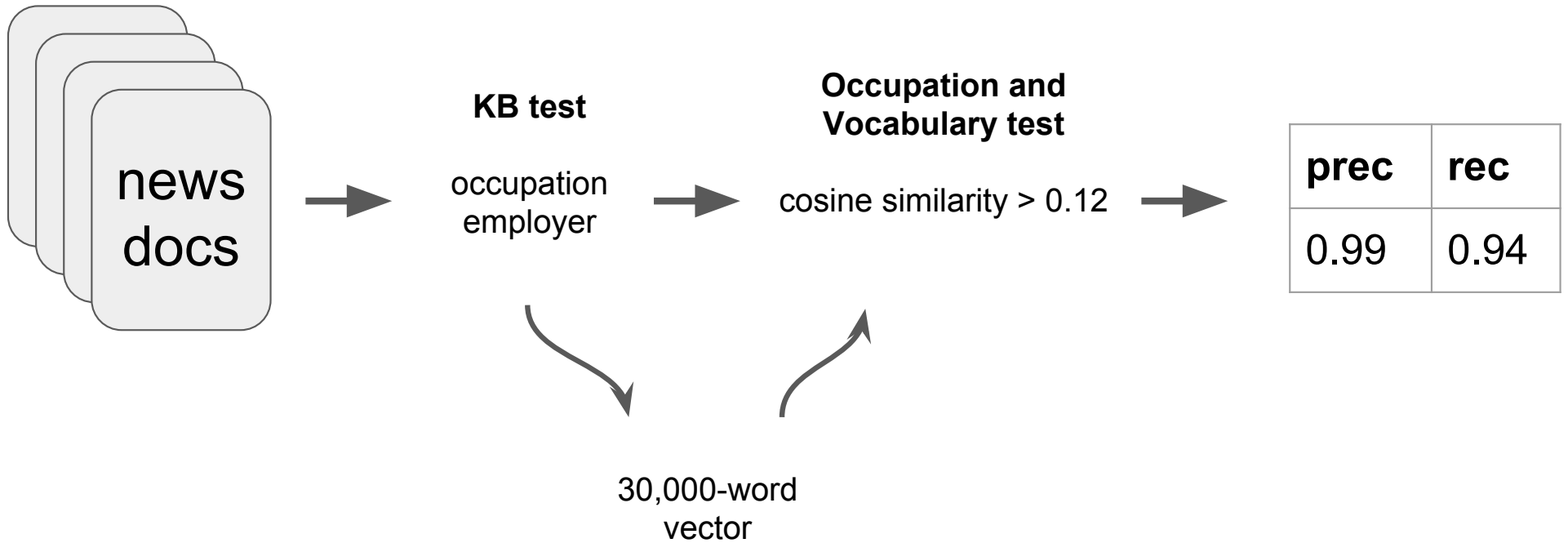
# from news to **Mentions**...



- An international team led by **Janet Kelso**, a bioinformaticist at the Max Planck Institute for Evolutionary Anthropology, has now sequenced the genome of the brown kiwi.
- **Kelso** and her team at the Max Planck Institute for Evolutionary Anthropology in Leipzig, Germany, looked for Neanderthal and Denisovan genetic ancestry.
- **Janet Kelso** passed away on August 31, 2008 at the age of 70 in Indianapolis, Indiana.

← REALLY?

# Entity-linking



# from mentions to **Observations** ...

An international team led by **Janet Kelso**, a **bioinformaticist** at the **Max Planck Institute for Evolutionary Anthropology**, has now sequenced the genome of the brown kiwi.



- **Janet Kelso** — *occupation* → **bioinformaticist**
- **Janet Kelso** — *employer* → **Max Planck Institute for Evolutionary Anthropology**

# from observations to **Claims**

...

- **Janet Kelso** — *has occupation* → **bioinformaticist** (0.92)
- **Janet Kelso** — *has occupation* → **geneticist** (0.45)
- **Janet Kelso** — *has occupation* → **scientist** (0.18)
- **Janet Kelso** — *has occupation* → **researcher** (0.15)
- **Janet Kelso** — *has occupation* → **author** (0.12)

TRUTHINESS



# from claims to new **Beliefs**

- **Janet Kelso** — *has occupation* → **bioinformaticist** (0.92)
- **Stephen Altschul** — *has occupation* → **bioinformaticist** (0.89)
- **Pavel Pevzner** — *has occupation* → **bioinformaticist** (0.90)
- **Emmanuel Yera** — *has occupation* → **bioinformaticist** (0.99)
- **Leonard Apeltsin** — *has occupation* → **bioinformaticist** (0.97)



**bioinformaticist** → is instance of **occupation**





# Wikidata to the rescue!

occupation: **physicist**  
employer: **University of Cambridge**

**Stephen W. Hawking**, the **Cambridge University physicist** and best-selling author who roamed the cosmos from a wheelchair, pondering the nature of gravity and the origin of the universe and becoming an emblem of human determination and curiosity, died early Wednesday at his home in Cambridge, England.

→ *New York Times*, 14 March 2018



## B3: the Bayesian Belief Base

70k scientists in the KB



3k mentioned in the news per day (15k docs)



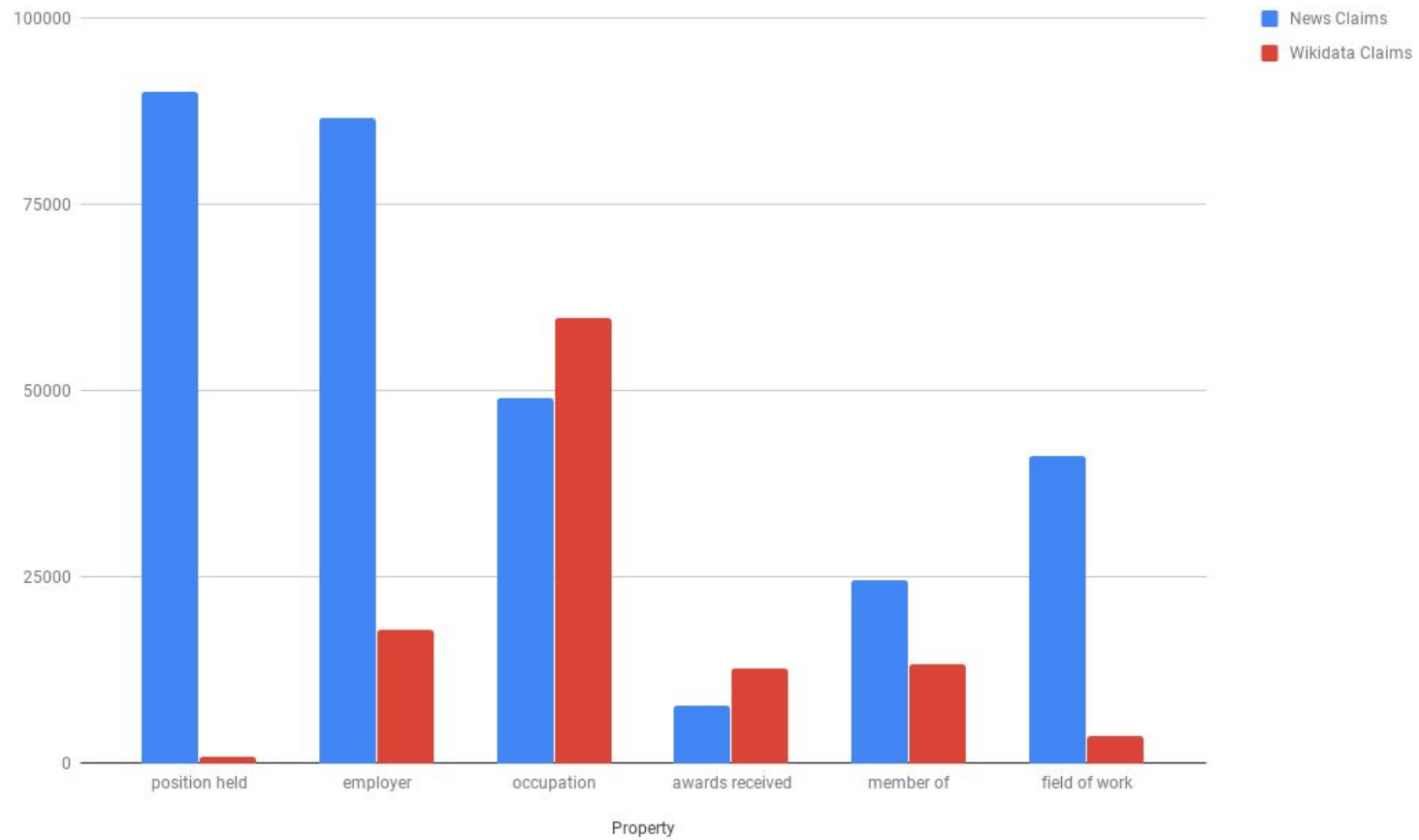
# Quicksilver

<https://quicksilver.primer.ai>

<https://quicksilver-demo.primer.ai>



# Quicksilver vs. Wikidata





# Questions