

Playing with Words

Building Products with NLP

Hello, everyone!



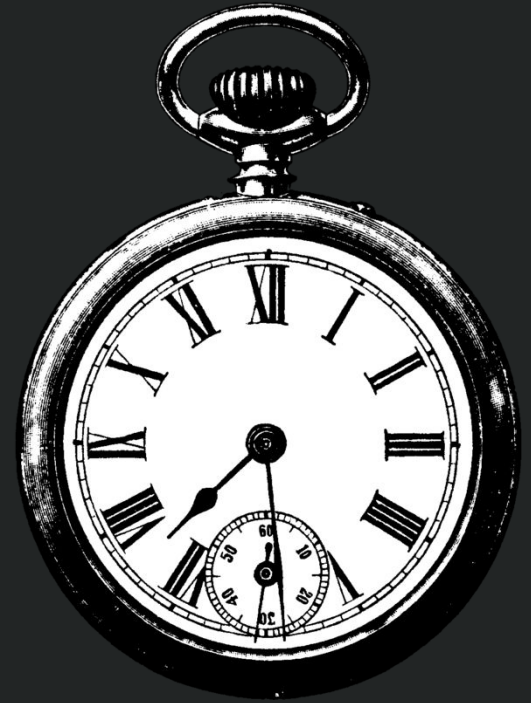
I'm Hilary Mason, co-founder and CEO of Hidden Door.

Twitter: @hmason

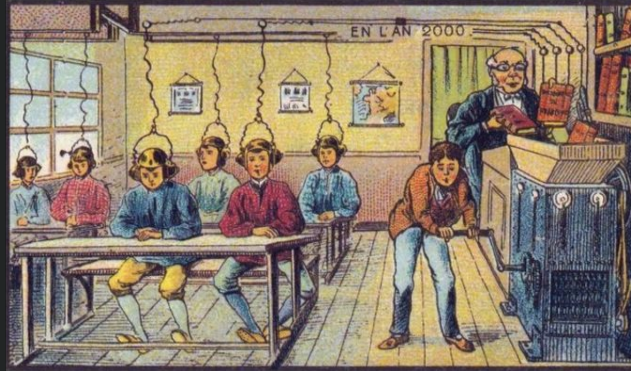
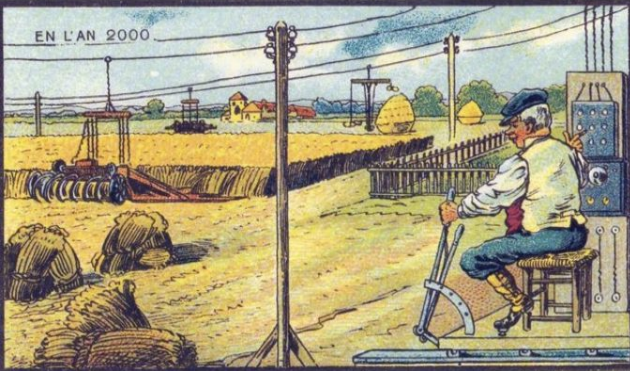
E-mail: hilary@hiddendoor.co

What We'll Talk About Today

...can machines play?



Let's predict the future...



(Postcards from France in 1900 predicting the year 2000.)

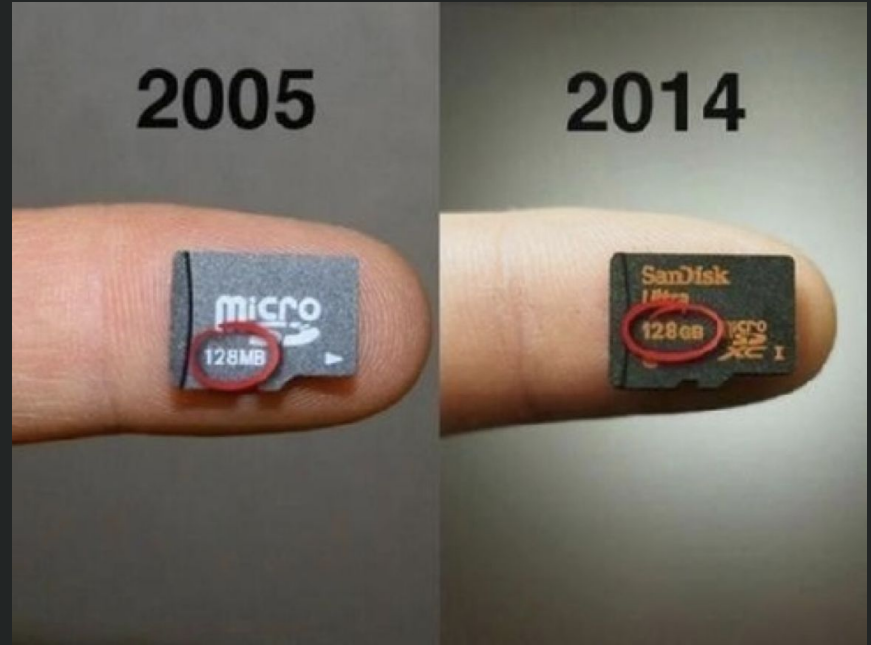
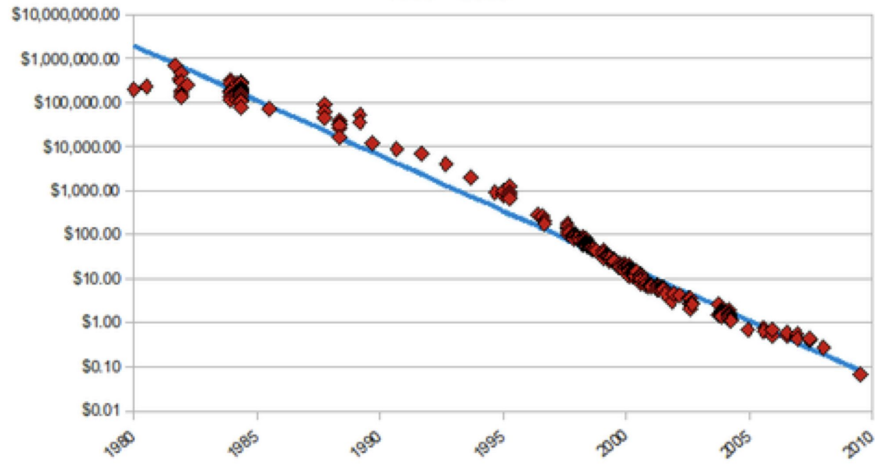
Two ways: Gradually, then suddenly.

Rather, let's predict the present.

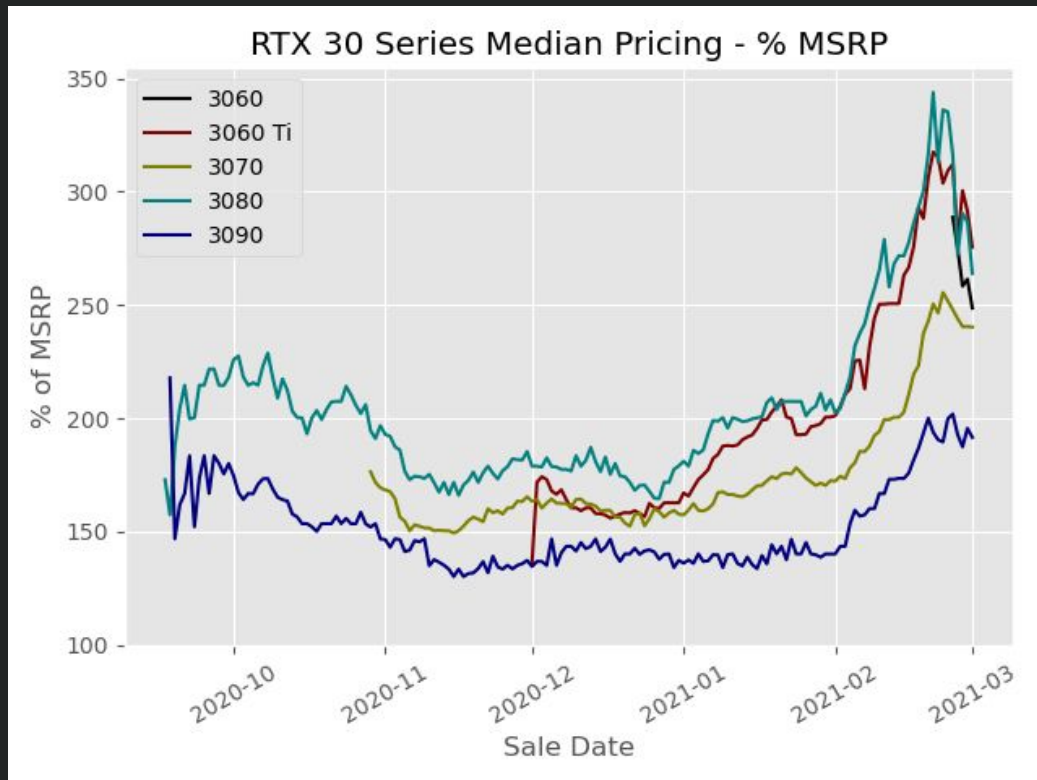
It's linear, or nonlinear.

Linear Changes

Hard Drive Cost per Gigabyte
1980 - 2009

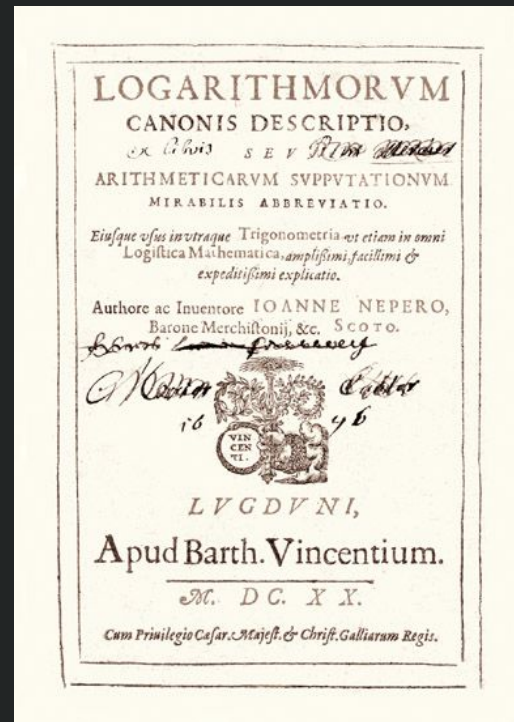
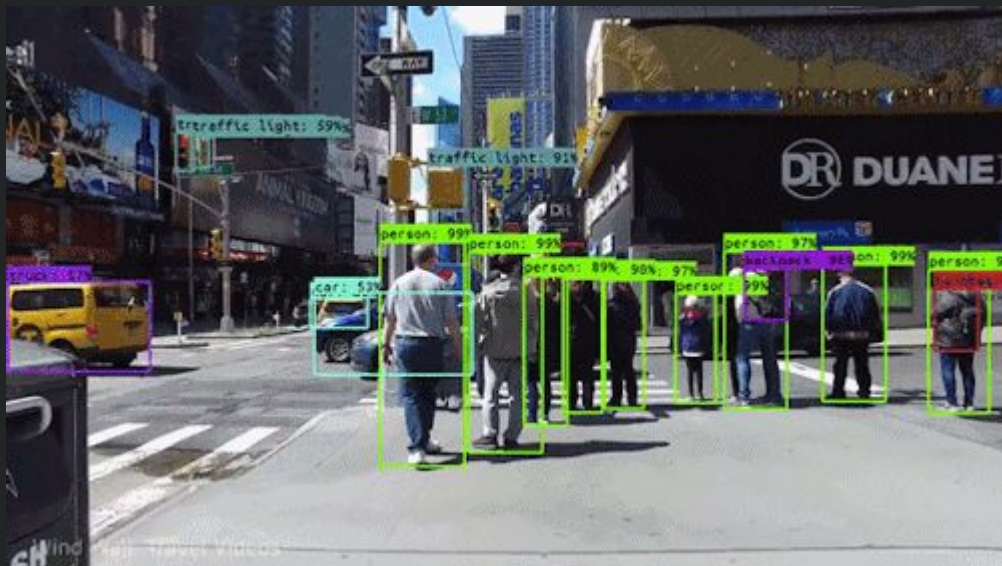


(okay, there are sometimes blips)

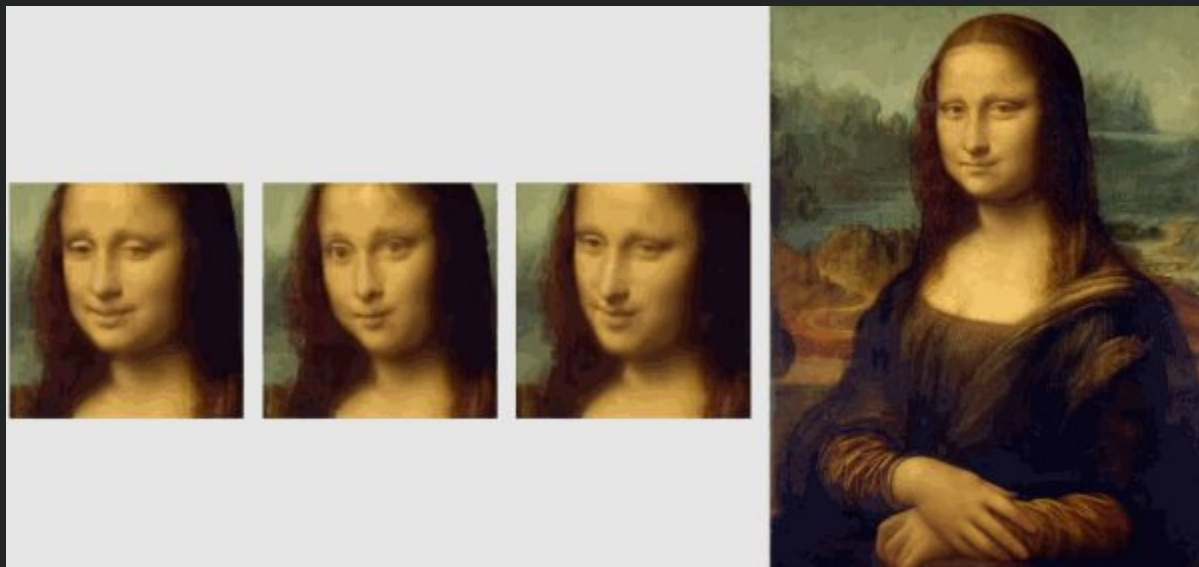


NVIDIA RTX card prices

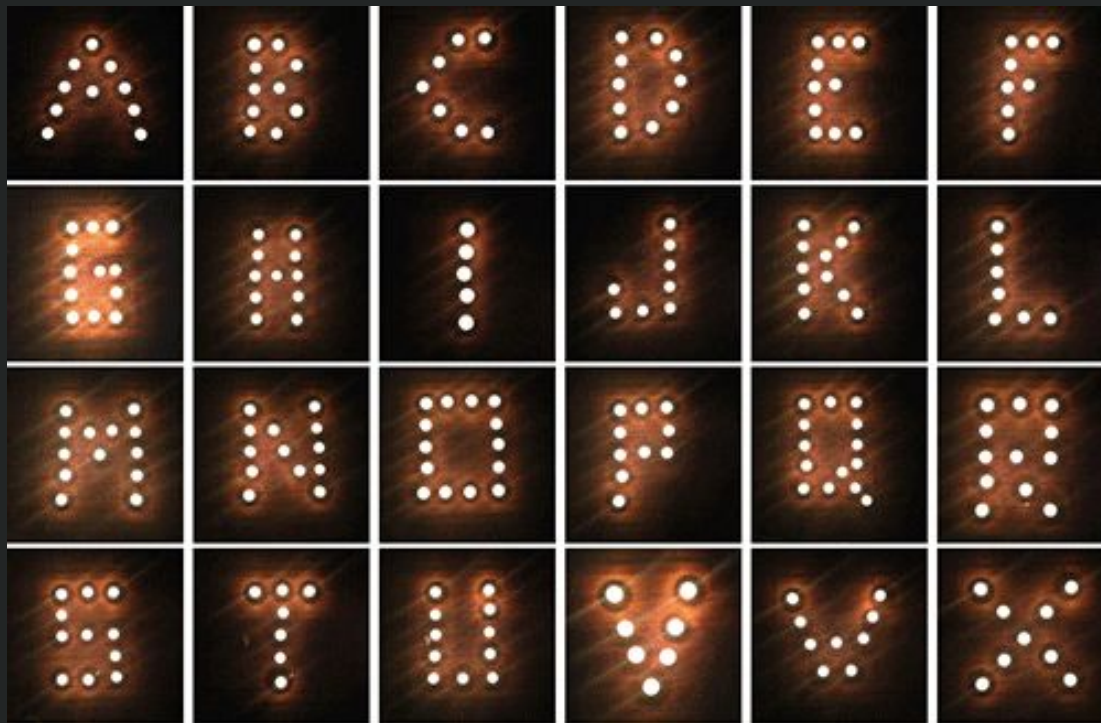
Step-Function Changes



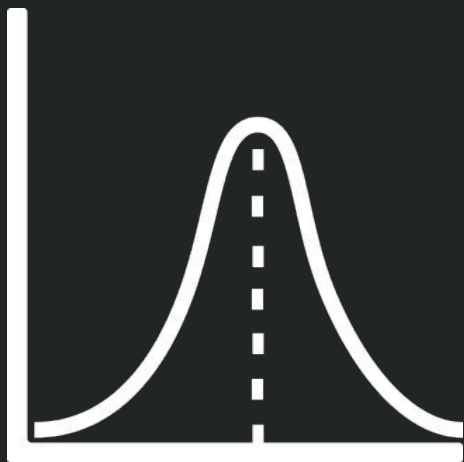
Many examples of this ...



NLP is where the action is today!



Linear growth forces include...



Data



Algorithms
& Software



Compute

Data!



WIKIPEDIA
The Free Encyclopedia

[Main page](#)
[Contents](#)
[Current events](#)
[Random article](#)
[About Wikipedia](#)
[Contact us](#)
[Donate](#)

[Contribute](#)

[Help](#)
[Learn to edit](#)
[Community portal](#)
[Recent changes](#)
[Upload file](#)

[Tools](#)

[What links here](#)
[Related changes](#)
[Special pages](#)
[Permanent link](#)
[Page information](#)
[Cite this page](#)
[Wikidata item](#)

Not logged in [Talk](#) [Contributions](#) [Create account](#) [Log in](#)

Article [Talk](#)

Read [Edit](#) [View history](#)

Language model

From Wikipedia, the free encyclopedia

A statistical **language model** is a [probability distribution](#) over sequences of words. Given such a sequence, say of length *m*, it assigns a probability $P(w_1, \dots, w_m)$ to the whole sequence.

The language model provides [context](#) to distinguish between words and phrases that sound similar. For example, in [American English](#), the phrases "recognize speech" and "wreck a nice beach" sound similar, but mean different things.

Data sparsity is a major problem in building language models. Most possible word sequences are not observed in training. One solution is to make the assumption that the probability of a word only depends on the previous *n* words. This is known as an [n-gram](#) model or unigram model when *n* = 1. The unigram model is also known as the [bag of words model](#).

Estimating the [relative likelihood](#) of different phrases is useful in many [natural language processing](#) applications, especially those that generate text as an output. Language modeling is used in [speech recognition](#),^[1] [machine translation](#),^[2] [part-of-speech tagging](#), [parsing](#),^[2] [Optical Character Recognition](#), [handwriting recognition](#),^[3] [information retrieval](#) and other applications.

In speech recognition, sounds are matched with word sequences. Ambiguities are easier to resolve when evidence from the language model is integrated with a pronunciation model and an [acoustic model](#).

Language models are used in information retrieval in the [query likelihood model](#). There, a separate language model is associated with each [document](#) in a collection. Documents are ranked based on the probability of the query *Q* in the document's language model M_d : $P(Q | M_d)$. Commonly, the [unigram](#) language model is used for this purpose.

Contents [hide]

- 1 [Model types](#)
 - 1.1 [Unigram](#)
 - 1.2 [n-gram](#)
 - 1.2.1 [Bidirectional](#)
 - 1.2.2 [Example](#)
 - 1.3 [Exponential](#)

Commoditized capabilities

... in software, models, and APIs.

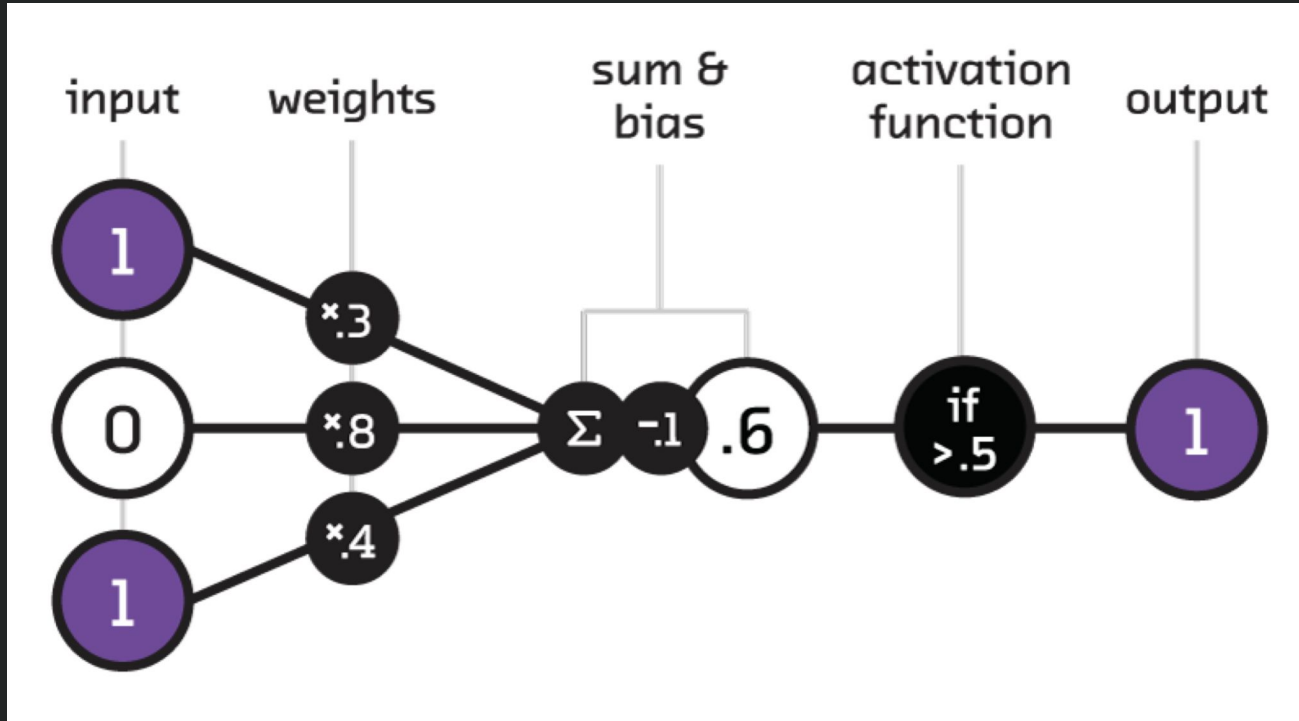


OpenAI

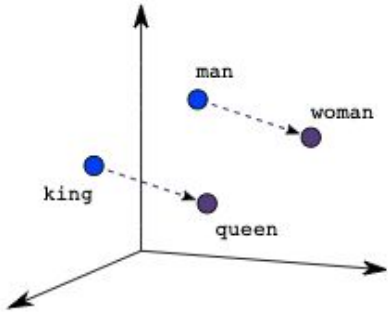
The Cloud



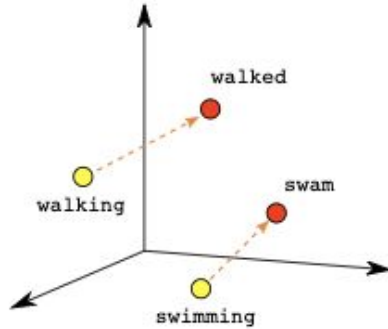
Step functions starting with deep learning...



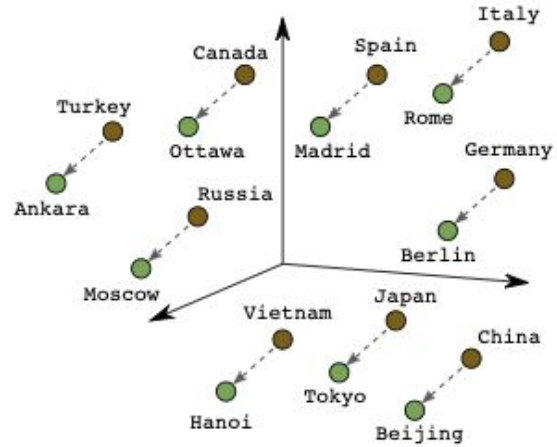
Embeddings



Male-Female



Verb Tense



Country-Capital

Figure 4. Embeddings can produce remarkable analogies.

DATA

5 tensors found

Word2Vec 10K

Label by **word** Color by **No color map**

Edit by **word** Tag selection as

Load Publish Download Label

Sphेरize data

Checkpoint: Demo datasets

Metadata: oss_data/word2vec_10000_200d_labels.tsv

UMAP T-SNE **PCA** CUSTOM

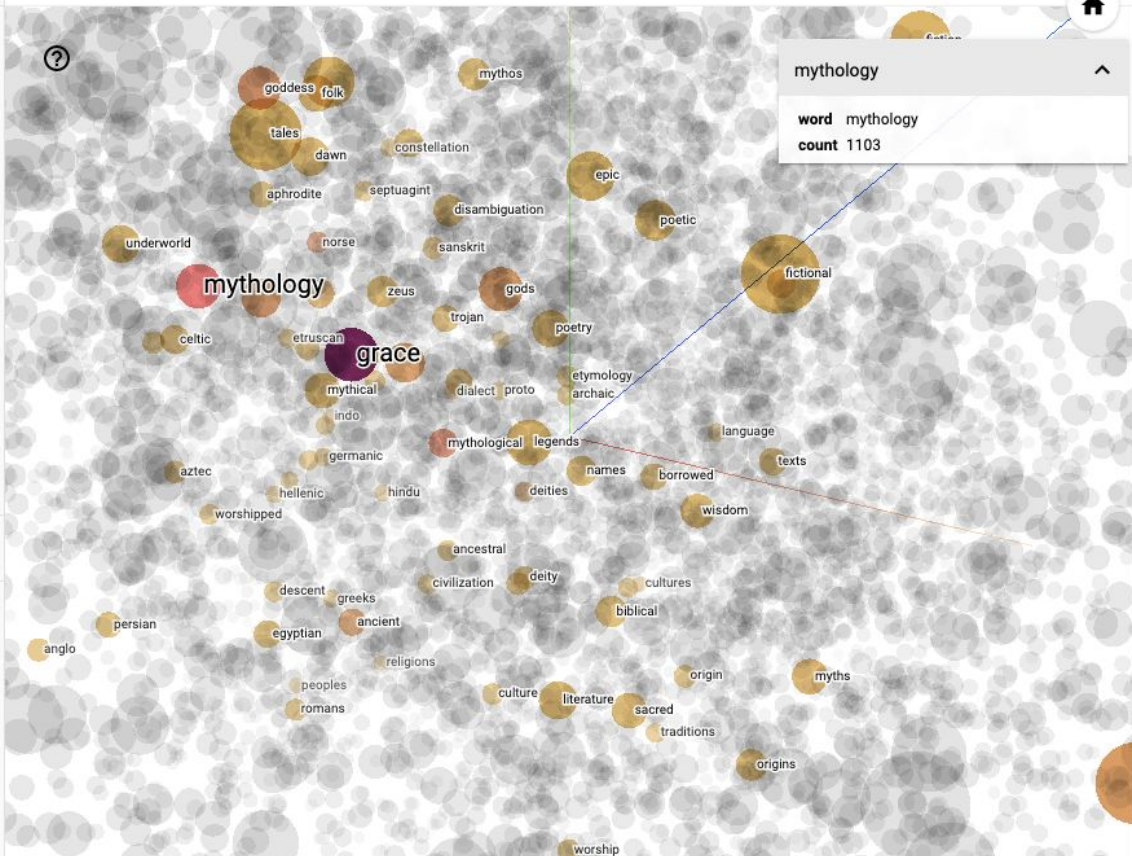
X **Component #1** Y **Component #2**

Z **Component #3**

PCA is approximate.

Total variance described: 8.5%.

Points: 10000 | Dimension: 200 | Selected 101 points



Show All Data Isolate 101 points Clear selection

Search by **word**

neighbors 100

distance COSINE EUCLIDEAN

Nearest points in the original space:

mythological	0.403
norse	0.440
goddess	0.471
greek	0.491
gods	0.500
folklore	0.508
myth	0.508
ancient	0.513
deities	0.537
goddesses	0.564
myths	0.584
mythical	0.595
religion	0.604
constellation	0.604
deity	0.610

BOOKMARKS (0)

Seq2seq and Transformers

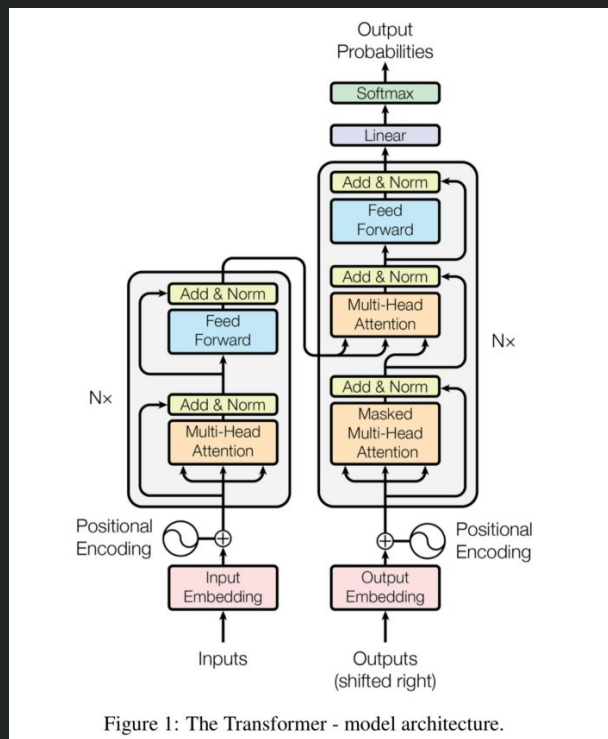
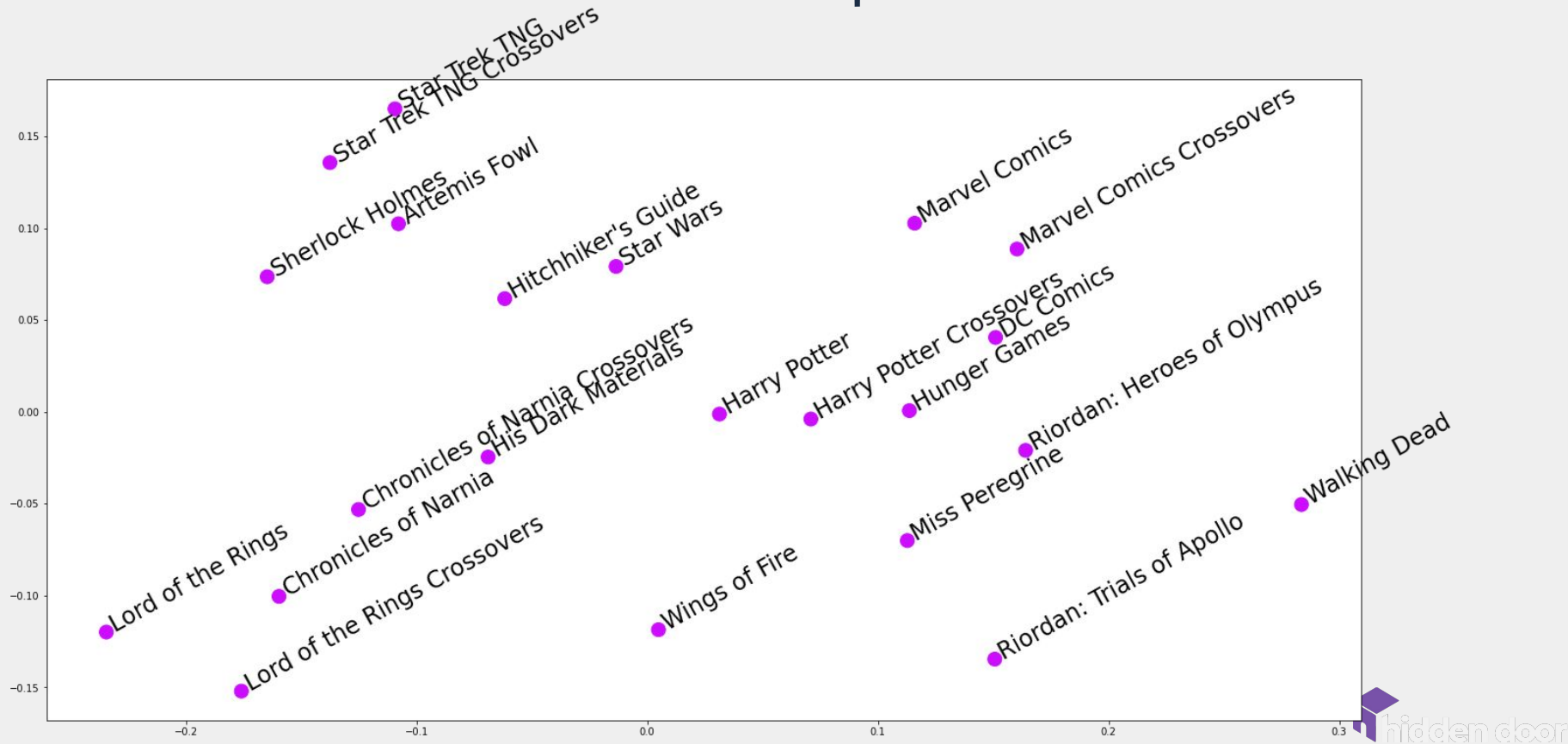


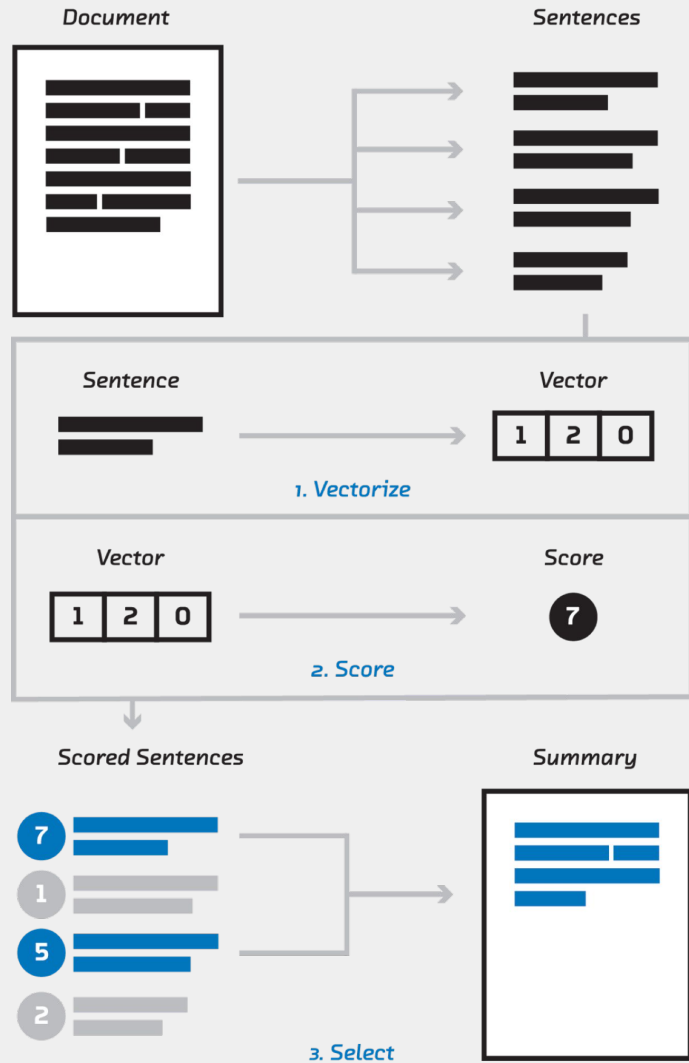
Figure 1: The Transformer - model architecture.

With encoders and decoders and attention!

What can we do with this? (a preview)



We can Summarize





Mark Harris [Follow](#)

Mostly British, mostly tech journalist reporting on the US from mostly rainy Seattle

Jun 6 · 16 min read

Say Goodbye To Your Highly Skilled Job. It's Now a "Human Intelligence Task."

Digital crowdworkers don't only do menial tasks like data entry. They're smart, capable, and hungrier than any algorithm. And they work for cheap.



♥ 216 💬 6



[Next story](#)
[iMcLaren: A Luxury Wearable for ...](#)

Harry K. sits at his desk in Vancouver, Canada, scanning sepia-tinted swirls, loops and blobs on his computer screen. Every second or so, he jabs at his mouse and adds a fluorescent dot to the image. After a minute, a new image pops up in front of him.

Harry is tagging images of cells removed from breast cancers. It's a painstaking job but not a difficult one, he says: "It's like playing Etch A Sketch or a video game where you color in certain dots."

Harry found the gig on [Crowdfunder](#), a crowdworking platform. Usually that cell-tagging task would be the job of pathologists, who typically start their careers with annual salaries of around \$200,000—an hourly wage of about \$80. Harry, on the other hand, earns just four cents for annotating a batch of five images, which takes him between two to eight minutes. His hourly wage is about 60 cents.

Granted, Harry can't perform most of the tasks in a pathologist's repertoire. But in 2016—11 years after the launch of the ur-platform, [Amazon Mechanical Turk](#)—crowdworking (sometimes also called crowdsourcing) is eating into increasingly high-skilled jobs. The engineers who are developing this model of labor have a bold ambition to atomize entire careers into micro-tasks that almost anyone, anywhere in the world, can carry out online. They're banking on the idea that any technology that can make a complex process 100 times cheaper, as in Harry's case, will spread like wildfire.

Top Highlights

But in 2016 -- 11 years after the launch of the ur-platform, Amazon Mechanical Turk -- crowdworking (sometimes also called crowdsourcing) is eating into increasingly high-skilled jobs.

Score: 28 · Scroll: 14%

The engineers who are developing this model of labor have a bold ambition to atomize entire careers into micro-tasks that almost anyone, anywhere in the world, can carry out online.

Score: 30 · Scroll: 15%

They're banking on the idea that any technology that can make a complex process 100 times cheaper, as in Harry's case, will spread like wildfire.

Score: 28 · Scroll: 15%

But as the tech conversation has fixated on how artificial intelligence will affect the job market, crowdwork

Harry K. sits at his desk in Vancouver, Canada, scanning sepia-tinted swirls, loops and blobs on his computer screen. Every second or so, he jabs at his mouse and adds a fluorescent dot to the image. After a minute, a new image pops up in front of him.

Harry is tagging images of cells removed from breast cancers. It's a painstaking job but not a difficult one, he says: "It's like playing Etch A Sketch or a video game where you color in certain dots."

Harry found the gig on Crowdfunder, a crowdworking platform. Usually that cell-tagging task would be the job of pathologists, who typically start their careers with annual salaries of around \$200,000—an hourly wage of about \$80. Harry, on the other hand, earns just four cents for annotating a batch of five images, which takes him between two to eight minutes. His hourly wage is about 60 cents.

Granted, Harry can't perform most of the tasks in a pathologist's repertoire. But in 2016—11 years after the launch of the ur-platform, Amazon Mechanical Turk—crowdworking (sometimes also called crowdsourcing) is eating into increasingly high-skilled jobs. The engineers who are developing this model of labor have a bold ambition to atomize entire careers into micro-tasks that almost anyone, anywhere in the world, can carry out online. They're banking on the idea that any technology that can make a complex process 100 times cheaper, as in Harry's case, will spread like wildfire.

Perhaps it's inevitable that in a few years, software will swallow up these jobs, too. But as the tech conversation has fixated on how artificial intelligence will affect the job market, crowdwork

Merial Frontline Plus Flea and Tick Control for 5-22 Pound Dogs and Puppies, 3-Pack

Review Topics

products product pet company china pets read information review website

“Nothing illegal.”

“Products produced for US distribution are EPA approved.”

“Products for International markets, are not EPA approved.”

away went took got time didn day night thought right

“I was standing maybe 2 feet away from her and she was whining and crying as I was calling her to come to me.”

“It was absolutely terrifying to watch her.”

“We know there are fleas where we walk but I laugh as they attempt to jump on my dogs which like little death ninjas have but to secrete this oil and BAM THEY DIE LIKE THE VERMIN”

price amazon shipping cost buy free store pet cheaper pay

“Amazon also had the cheapest price plus free shipping.”

“Amazon had the lowest price and free shipping.”

“You can't beat prime shipping either.”

vet life years results did said months year ago old

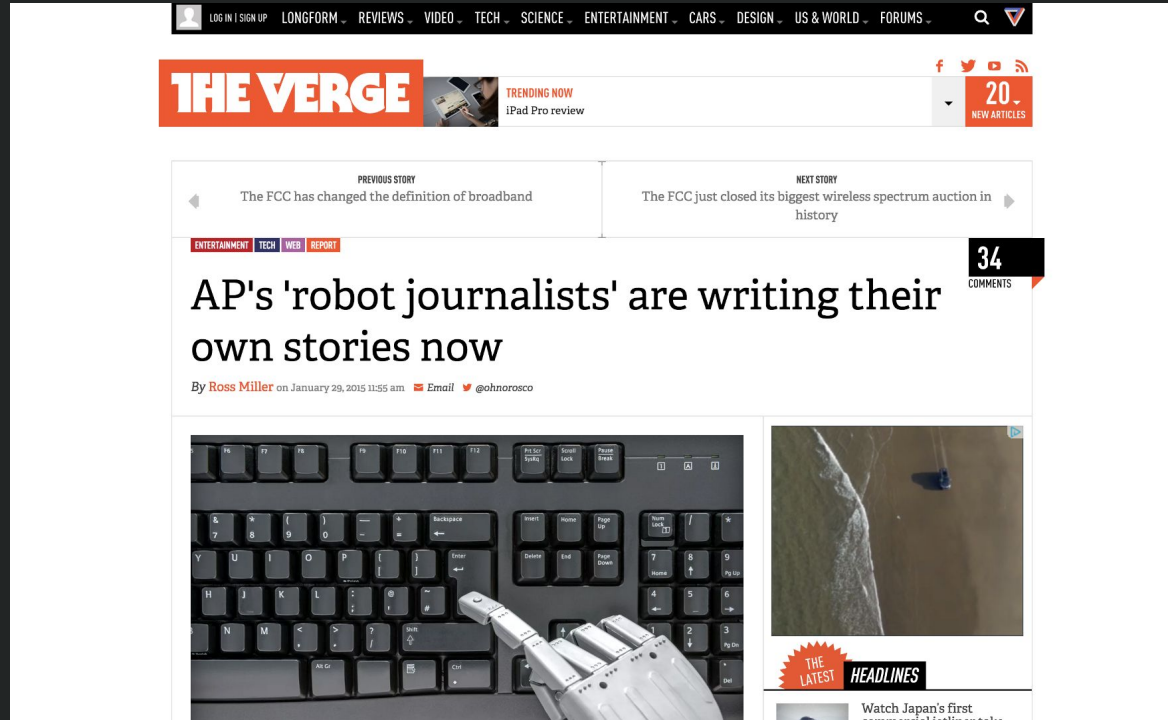
“IV's for dehydration from vomiting.”

“Then he began vomiting and had explosive diarrhea.”

Summaries for Traders, and more!

- Writing headlines for pop culture articles
- Sorting news articles for traders
- Tracking ideas from social media to mainstream news
- Supporting customer support systems

We can generate...



The screenshot shows the top navigation bar of The Verge website with links for LOG IN | SIGN UP, LONGFORM, REVIEWS, VIDEO, TECH, SCIENCE, ENTERTAINMENT, CARS, DESIGN, US & WORLD, and FORUMS. The main header features the THE VERGE logo, a trending article for iPad Pro review, and a badge for 20 NEW ARTICLES. The article title is "AP's 'robot journalists' are writing their own stories now" by Ross Miller, dated January 29, 2015. The main image shows a robotic hand typing on a keyboard. A sidebar on the right shows 34 COMMENTS and a "THE LATEST HEADLINES" section with a video thumbnail.

LOG IN | SIGN UP LONGFORM REVIEWS VIDEO TECH SCIENCE ENTERTAINMENT CARS DESIGN US & WORLD FORUMS

THE VERGE TRENDING NOW iPad Pro review 20 NEW ARTICLES

PREVIOUS STORY The FCC has changed the definition of broadband NEXT STORY The FCC just closed its biggest wireless spectrum auction in history

ENTERTAINMENT TECH WEB REPORT

AP's 'robot journalists' are writing their own stories now

By Ross Miller on January 29, 2015 11:55 am Email @ohnorosco

34 COMMENTS

THE LATEST HEADLINES Watch Japan's first commercial jetliner take

REAL ESTATE LISTINGS GENERATOR

Generate listings for a **2** bedroom **1** bathroom apartment in **Upper East Side** with the following amenities: **Laundry in Building**

Storage Available **Roof Deck** **Gym** **Balcony** **Doorman** **Pets Allowed** **Terrace**

Run for **20** seconds.

Generate

TOP SCORED LISTINGS

from the **1350** listings generated

Additional central **laundry** room, cold storage, pet grooming room and bike storage are also available. This luxuriously-appointed **two** bedroom **one** bathroom apartment in the sleek Milan condominium in Midtown East exudes contemporary elegance with double exposures (north and west). Additionally, the building has **storage** units available, bike room, luggage room, and new **laundry** room in the basement, plus a beautiful garden/patio area in the back. Situated in the coveted Sutton area of **Ues**, The Milan condominium tower built in 2005 with its 2-story Jerusalem stone, Bubinga & Jatoba wood lobby and exquisite residences is striking in design

LISTING SCORE

9

Unique **two** Bedroom and **one** Bath with Chef's Eat-in-Kitchen. Adjacent to the kitchen is a full size **laundry** room, and bedroom with full bath currently configured as a gym. Butler's pantry features miles of **storage** and presently has a **washer** (Washer /Dryers are permitted). This 18th floor apartment, with eastern and southern exposures, has direct **Ues** River views.

LISTING SCORE

9

Desirable **Two** Bedroom Prewar Coop. There is basement **storage** available, and while there is central **laundry**, too, residents are allowed to install a Washer/Dryer in their apartment with board approval. Stately 2,021 SF three bedroom, **one** and one half bath double corner residence at 301 East 50th Street. Located in **ues** and close to fine restaurants, MOMA, shopping and all public transportation; this much sought after location has numerous amenities.

LISTING SCORE

8

OpenAI's GPT-3

YOW Data is a conference about the impact of big data on the future of the enterprise. It will be taking place this year in NYC. For more information and to get your early bird tickets, please visit <http://www.cvent.com/events/big-data-conference-nyc/event-summary-0f38d8>

YOW Data is a conference about big data. It's not the kind of big data where there are too many numbers to count. It's the kind of big data where there are too many important things to count.

It's about 'data science' in the same way that 'computer science' is about computers. It's not about Excel spreadsheets or the database thingy with the green table. It's about finding interesting things out about the world using data.

This year's event will take place at the University of Oxford, on 24-27 September.

YOW Data is a conference about data. We're a small team who've been running data-related events, with a conference in London since 2013, and another one in New York since 2014. We're based in London, and we focus on bringing together data practitioners from across the business, IT and data science worlds.

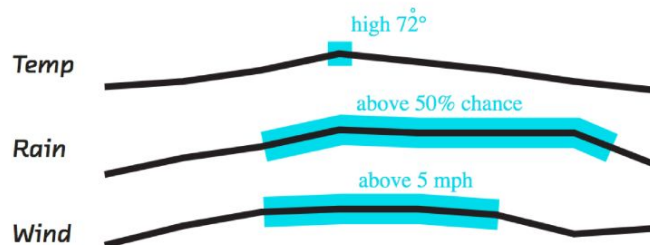
We organised the first EVERYOW conference in 2013, and we've been doing it every year since. We're London-based, and we focus on bringing together data practitioners from across the business, IT and data science worlds. We run three events a year: EVERYOW in London, EVERYOW New York and EVERYOW Silicon Valley.

Generating tasks include

1. Raw Data

	Temperature	Rain %	Wind Speed
Time of Day	62°	30	4
	70°	40	7
	72°	70	8

2. Document Planning



3. Micro Planning

- high 72° → a warm day, with a high of 72°
- over 50% chance of rain → grab an umbrella, it's going to be wet out there
- over 5 mph wind → hold onto your hat, it's going to be a very windy day

4. Realization

A warm day, with a high of 72°. Grab an umbrella and hold onto your hat. It's going to be a wet and very windy day.

What are people building?

DIGITAL HEALTH

One Doctor's Prescription for the Health Tech Industry



FORTUNE

LEADERSHIP

What to Do When Your Coworker Sees You As a Threat



FORTUNE

DONALD TRUMP

Trump's FCC Chairman Disagrees About the Role of the Media



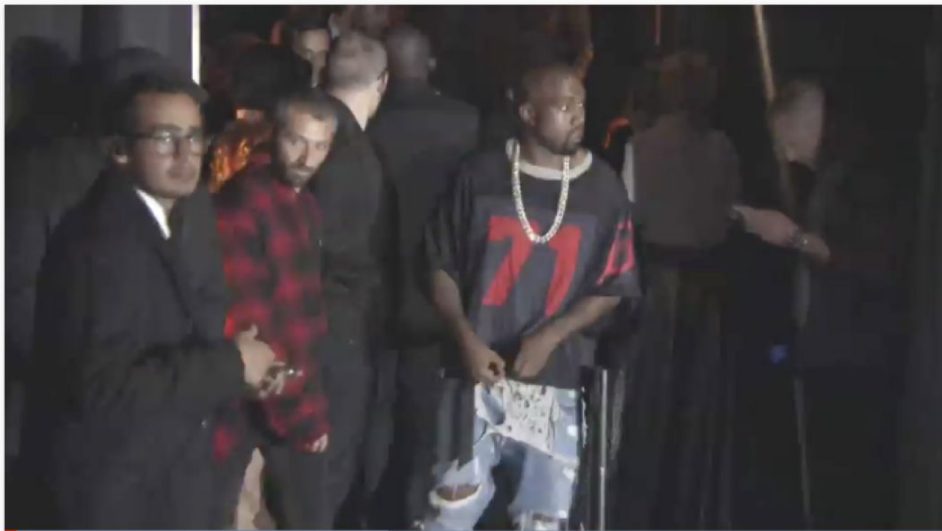
FORTUNE

FOOD TECH

The Future of Food Production Will Look A Lot Like Brewing Beer



FORTUNE



KANYE WEST

Teenager Teaches A.I. to Rap Like Kanye

David Z. Morris
Mar 19, 2017



A 17-year-old from West Virginia has used an archive of Kanye West lyrics to train a neural network to write rhymes on its own. The results are braggadocious, intermittently obscene, frequently incomprehensible, and laced with occasional profanity.

RELATED CONTENT

GRAMMYS 2017

What to Watch for and How to Watch the Grammys This Weekend



FORTUNE

KANYE WEST

Here's How Much Kanye West Stands to Lose for Canceling His Saint Pablo Tour



FORTUNE

KANYE WEST

Kanye West Cuts Concert Short After Ranting About Jay Z, Beyoncé and the Media



We believe great stories can change the world.

**Invent new characters for
your detective potboiler.**

Writing can be lonely, but
Sudowrite is always up for a
brainstorming sesh.

Characters for a detective story that
takes place in New Orleans:

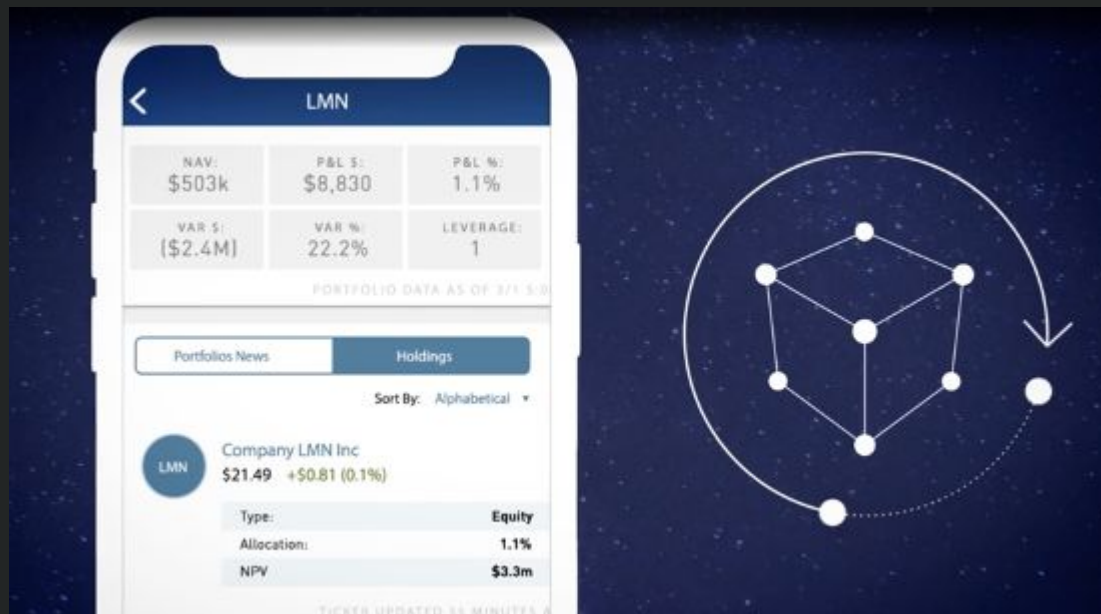
He owns a private detective agency in
New Orleans.

Billie Stone. A 37-year-old woman who
lives with her sister and two nieces
in a trailer on the side of a highway
in Louisiana. She owns a car wash
with her sister, named Wonder Wipe.
They don't really do business and

Janice is a librarian who helps a

Discover surprising story

Summaries for Traders



CAUTION

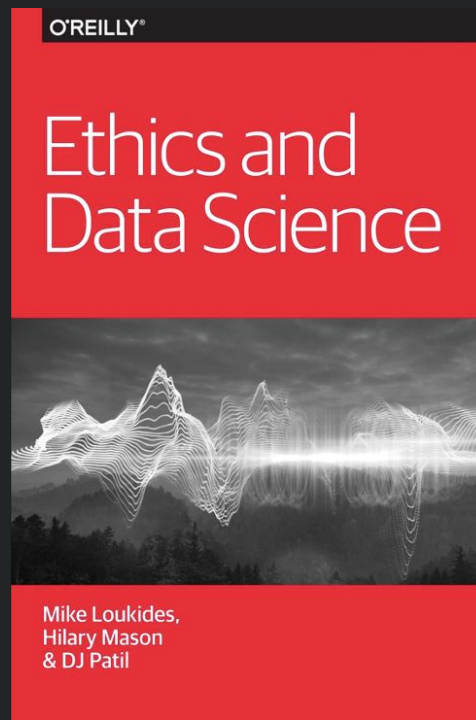


THIS IS SPARTA

Language Models aren't 'safe'!

...they aren't safe for *anyone*.

Building ethical products and businesses requires too much invention (and more intervention).



We Need New UX Metaphors

Search for movie...

! Women Art Revolution

#1

#Horror

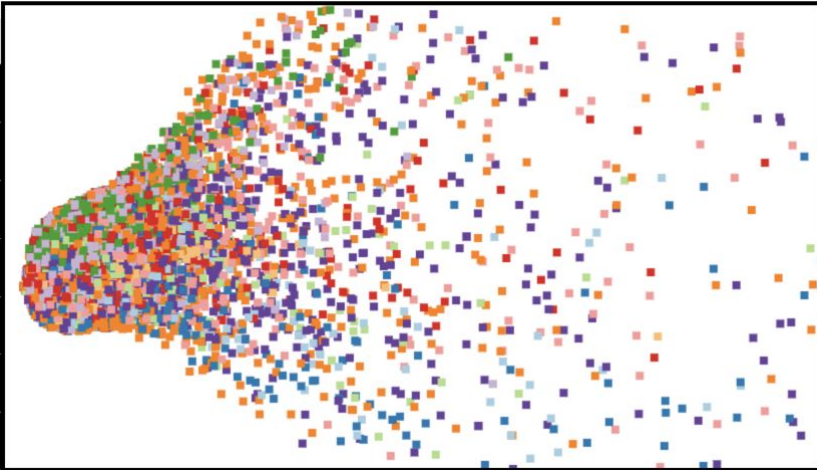
\$

\$9.99

'71

'Doc'

'E'

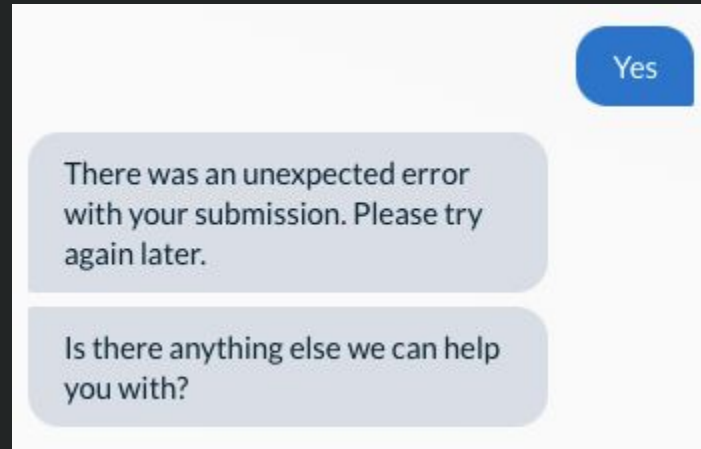


Action Adventure Animation Anthology Avant-Garde Biography Comedy Commercial Crime Cult Documentary Drama Erotica

Fantasy Family Film noir Gay & Lesbian History Horror Music Musical Music Video Mystery Romance Sci-Fi Short

Silent Sport Thriller TV Mini-series TV Movie War Western

(chatbots aren't it)



How does the product get built?

Data scientist? Data engineer? Machine learning engineer? Machine learning researcher?

How do we all work together?

How is the capability managed?



Move from efficiencies into new opportunities

Harvard Business Review

Latest Magazine Popular Topics Podcasts Video Store The Big Idea Visual Library Reading


DATA

How to Decide Which Data Science Projects to Pursue

by Hilary Mason

OCTOBER 17, 2018

Summary Save Share Comment 0 Text Size Print \$8.95 Buy Copies



We need more ambition!

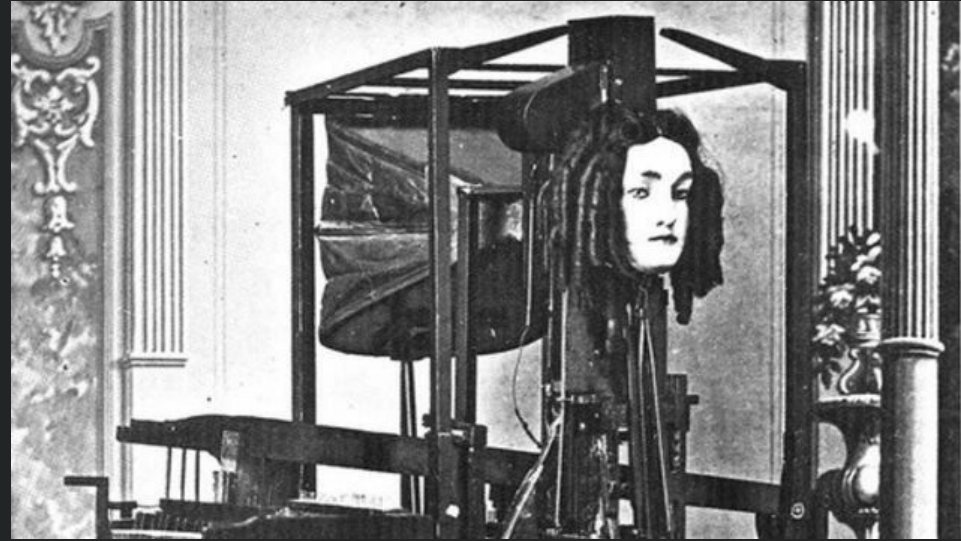
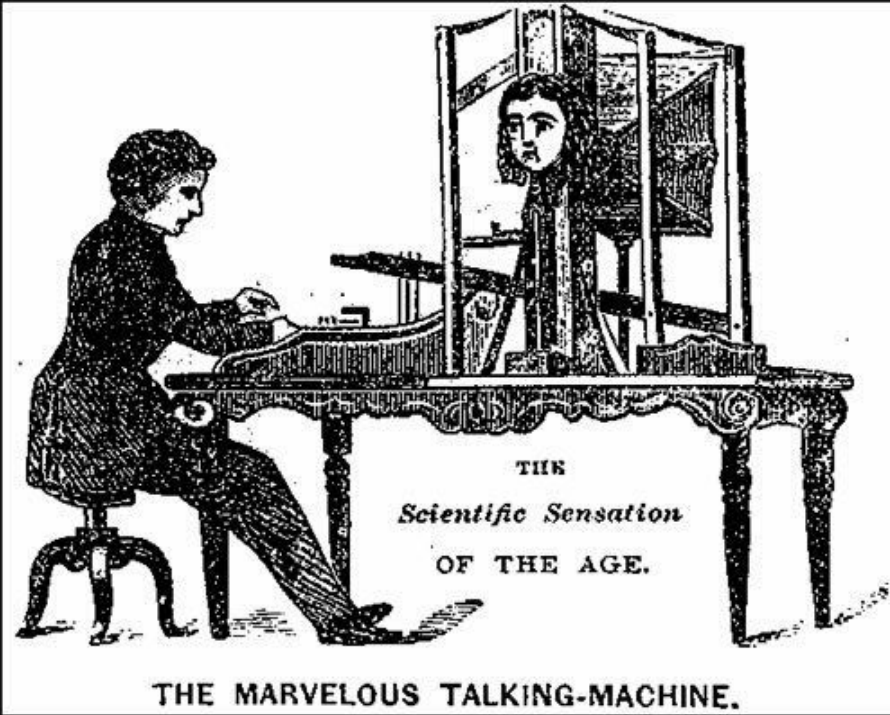
I'm excited about Augmented Creativity



People are creative, not machines.



The Euphonia, 1846 Speaking Robot

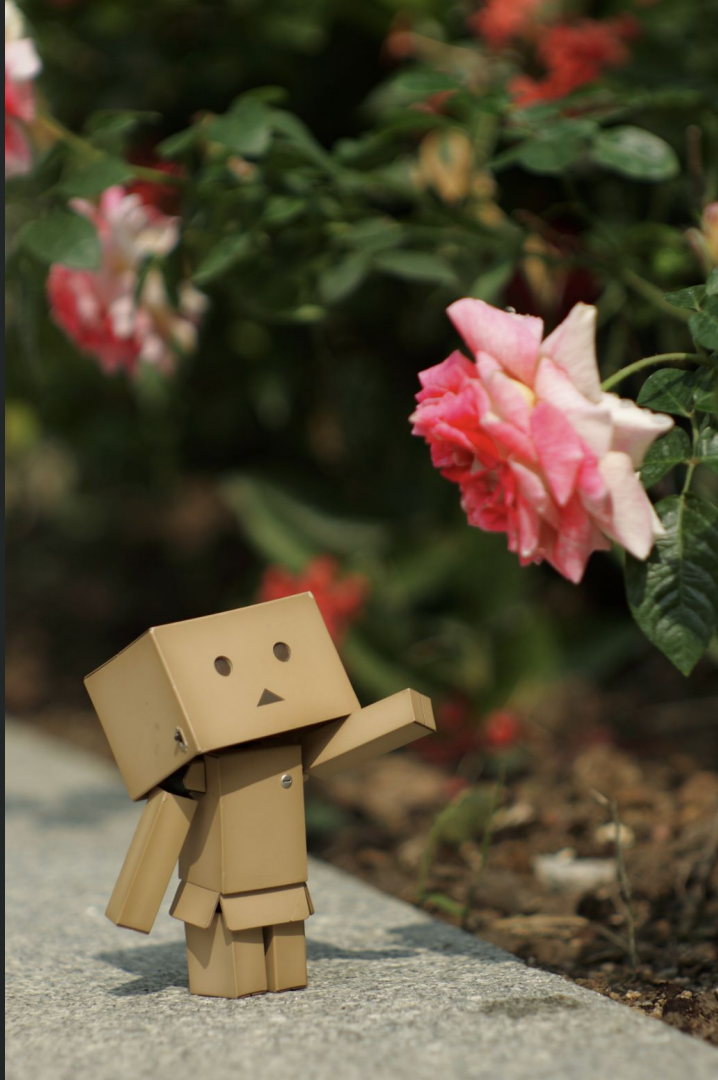


Tools to express the things you can imagine, but haven't yet created.

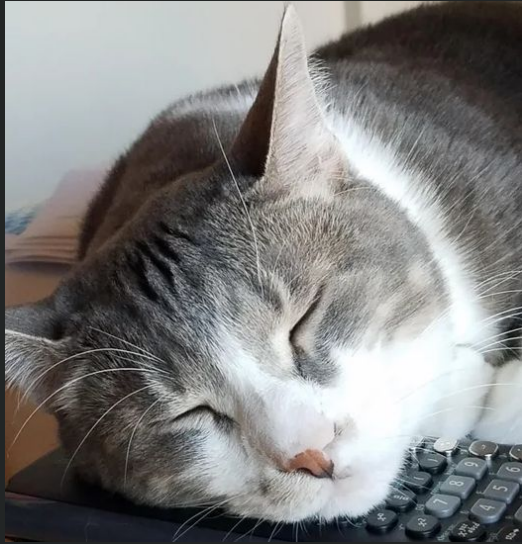




NLP is awesome!



Thank you so much, YOW Data 2021!



I'm Hilary Mason, co-founder and CEO of Hidden Door.

Q&A on Slack!

Twitter: @hmason

E-mail: hilary@hiddendoor.co