Understanding AI

James Reynolds (@magnusviri), April 26th, 2024

About James Reynolds

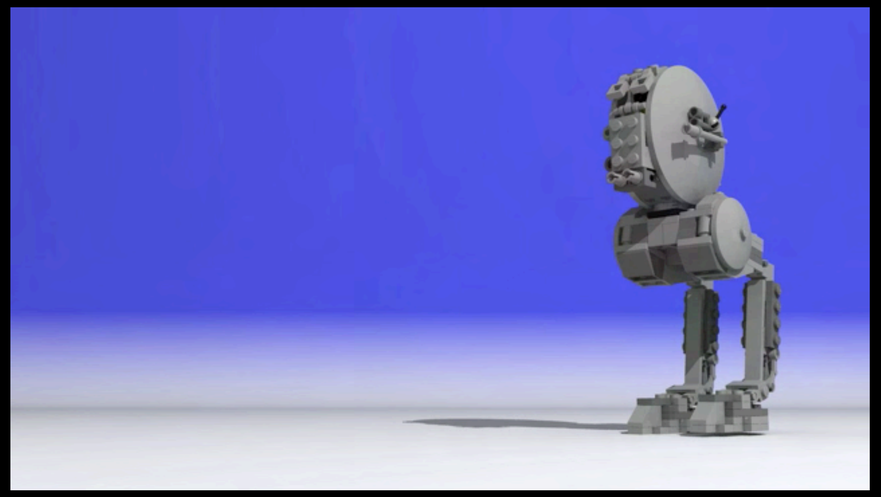
- U of U BMus 1999
- U of U Marriott Library, 14 years (Mac admin)
- U of U School of Biological Sciences, 10 years (Mac, Linux, network admin)

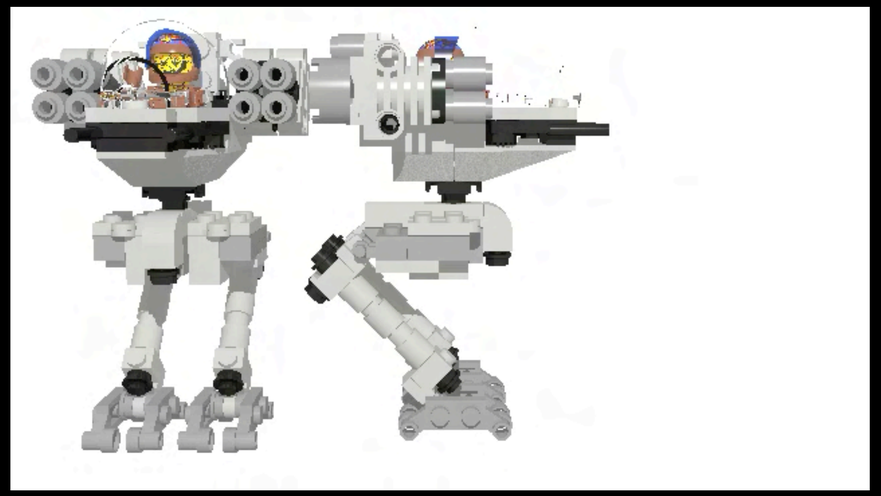




About James Reynolds

- Lego 3d graphics hobby up to ~2015
  - Overlap with AI (but I didn't know it)
    - 3d CG engines
    - Making these animations
- Since fall 2022 (Stable Diffusion 1.0), I’ve been “obsessed with AI”
- This is my 6th AI presentation since March 2023





Agenda

How Computers Work

Digitizing Meaning

AI People and History

AI Fear

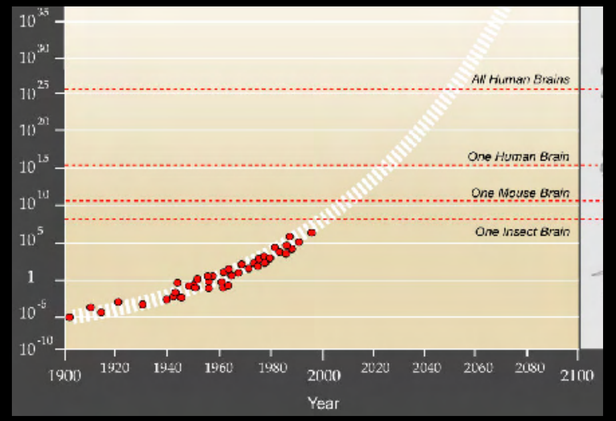
Slides: https://magnusviri.com/dl/lunch-and-learn-2024.pdf

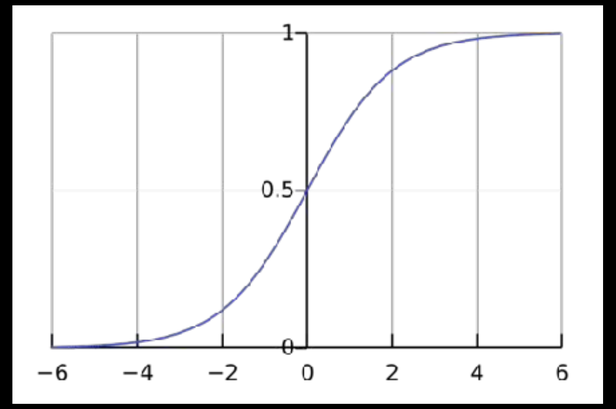
Setting the Stage

- Artificial intelligence is a very broad field (and an inaccurate term)
  - “Automation systems” is more accurate
  - Machine Learning (ML), Deep Learning, or Data Science (Connectionism)
  - Other AI tribes are Symbolists, Bayesians, Evolutionaries, and Analogizers
- There are many aspects of ML
  - Natural Language Processing (NLP) is one of many ML applications
  - Image/video classification/generation, STT/TTS, robotics, research

Natural Language Processing (NLP)

- Language is general purpose
- Language is code, can control a computer
  - Self driving cars will never create themselves
- Is language a prerequisite for thinking?
  - Does language enable thinking?
- Experts are debating if NLP has scaling limits
- Other than this, language ML is the same as all the other ML

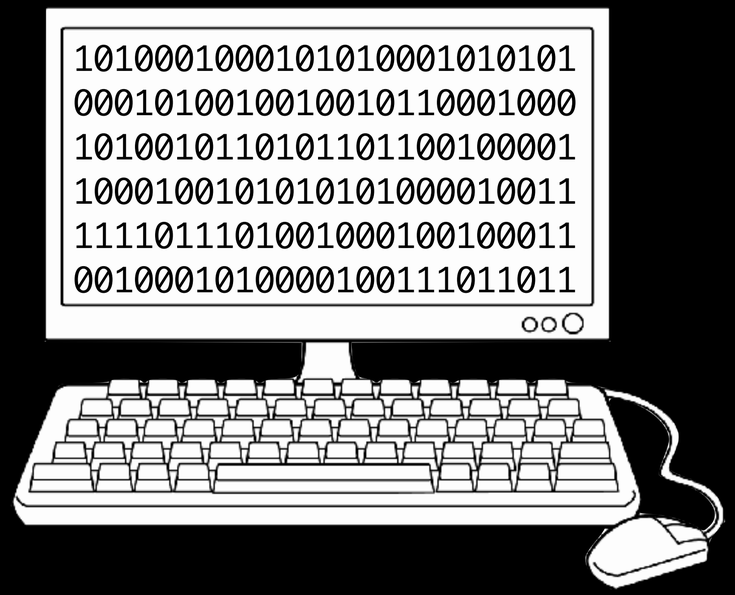




How Computers Work

Everything As Numbers

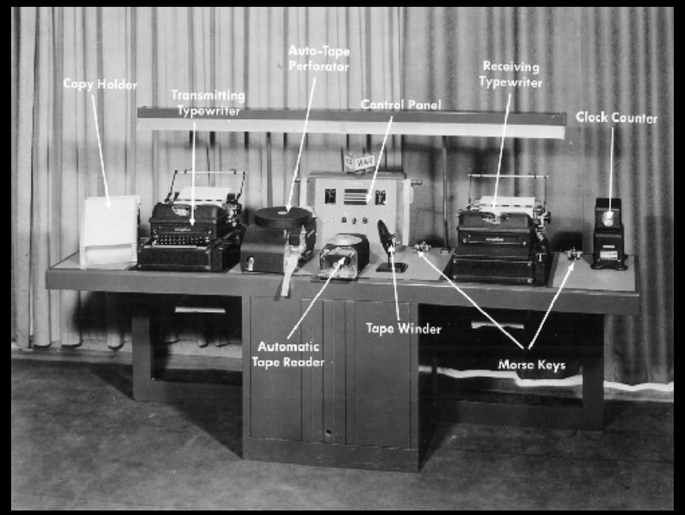
- Computers do math on (binary) numbers
- ASCII/Unicode: Letters as numbers
- JPEG/PNG: Images as numbers
- 3D OBJ: Objects as numbers
- MP3: Sound as numbers
- Machine Learning (AI): Anything as high dimensional vectors



I only understand 1 and 0

Letters as Numbers: ASCII and Unicode





Hex	Dec	Char		Hex	Dec	Char	Hex	Dec	Char	Hex	Dec	Char
0x00	0	NULL	null	0x20	32	Space	0x40	64	@	0x60	96	`
0x01	1	SOH	Start of heading	0x21	33	!	0x41	65	A	0x61	97	a
0x02	2	STX	Start of text	0x22	34	"	0x42	66	B	0x62	98	b
0x03	3	ETX	End of text	0x23	35	#	0x43	67	C	0x63	99	c
0x04	4	EOT	End of transmission	0x24	36	$	0x44	68	D	0x64	100	d
0x05	5	ENQ	Enquiry	0x25	37	%	0x45	69	E	0x65	101	e
0x06	6	ACK	Acknowledge	0x26	38	&	0x46	70	F	0x66	102	f
0x07	7	BELL	Bell	0x27	39	'	0x47	71	G	0x67	103	g
0x08	8	BS	Backspace	0x28	40	(	0x48	72	H	0x68	104	h
0x09	9	TAB	Horizontal tab	0x29	41	)	0x49	73	I	0x69	105	i
0x0A	10	LF	New line	0x2A	42	*	0x4A	74	J	0x6A	106	j
0x0B	11	VT	Vertical tab	0x2B	43	+	0x4B	75	K	0x6B	107	k
0x0C	12	FF	Form Feed	0x2C	44	,	0x4C	76	L	0x6C	108	l
0x0D	13	CR	Carriage return	0x2D	45	-	0x4D	77	M	0x6D	109	m
0x0E	14	SO	Shift out	0x2E	46	.	0x4E	78	N	0x6E	110	n
0x0F	15	SI	Shift in	0x2F	47	/	0x4F	79	O	0x6F	111	o
0x10	16	DLE	Data link escape	0x30	48	0	0x50	80	P	0x70	112	p
0x11	17	DC1	Device control 1	0x31	49	1	0x51	81	Q	0x71	113	q
0x12	18	DC2	Device control 2	0x32	50	2	0x52	82	R	0x72	114	r
0x13	19	DC3	Device control 3	0x33	51	3	0x53	83	S	0x73	115	s
0x14	20	DC4	Device control 4	0x34	52	4	0x54	84	T	0x74	116	t
0x15	21	NAK	Negative ack	0x35	53	5	0x55	85	U	0x75	117	u
0x16	22	SYN	Synchronous idle	0x36	54	6	0x56	86	V	0x76	118	v
0x17	23	ETB	End transmission block	0x37	55	7	0x57	87	W	0x77	119	w
0x18	24	CAN	Cancel	0x38	56	8	0x58	88	X	0x78	120	x
0x19	25	EM	End of medium	0x39	57	9	0x59	89	Y	0x79	121	y
0x1A	26	SUB	Substitute	0x3A	58	:	0x5A	90	Z	0x7A	122	z
0x1B	27	FSC	Escape	0x3B	59	;	0x5B	91	[	0x7B	123	{
0x1C	28	FS	File separator	0x3C	60	<	0x5C	92	\	0x7C	124	|
0x1D	29	GS	Group separator	0x3D	61	=	0x5D	93	]	0x7D	125	}
0x1E	30	RS	Record separator	0x3E	62	>	0x5E	94	^	0x7E	126	~
0x1F	31	US	Unit separator	0x3F	63	?	0x5F	95	_	0x7F	127	DEL

Images As Numbers



144, 115, 115

219, 145, 81

136, 124, 113

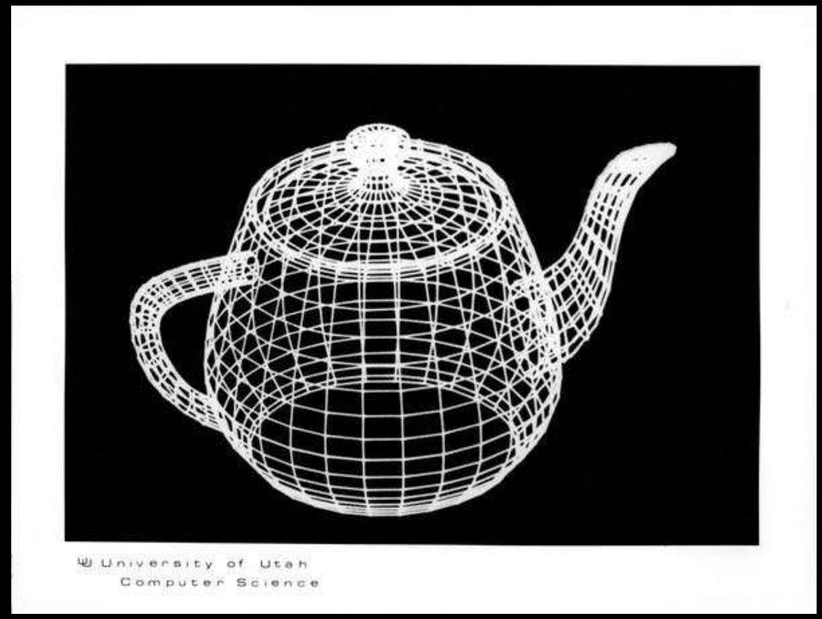
76, 61, 27

173, 169, 160

3D Objects as Numbers - “Utah Teapot”

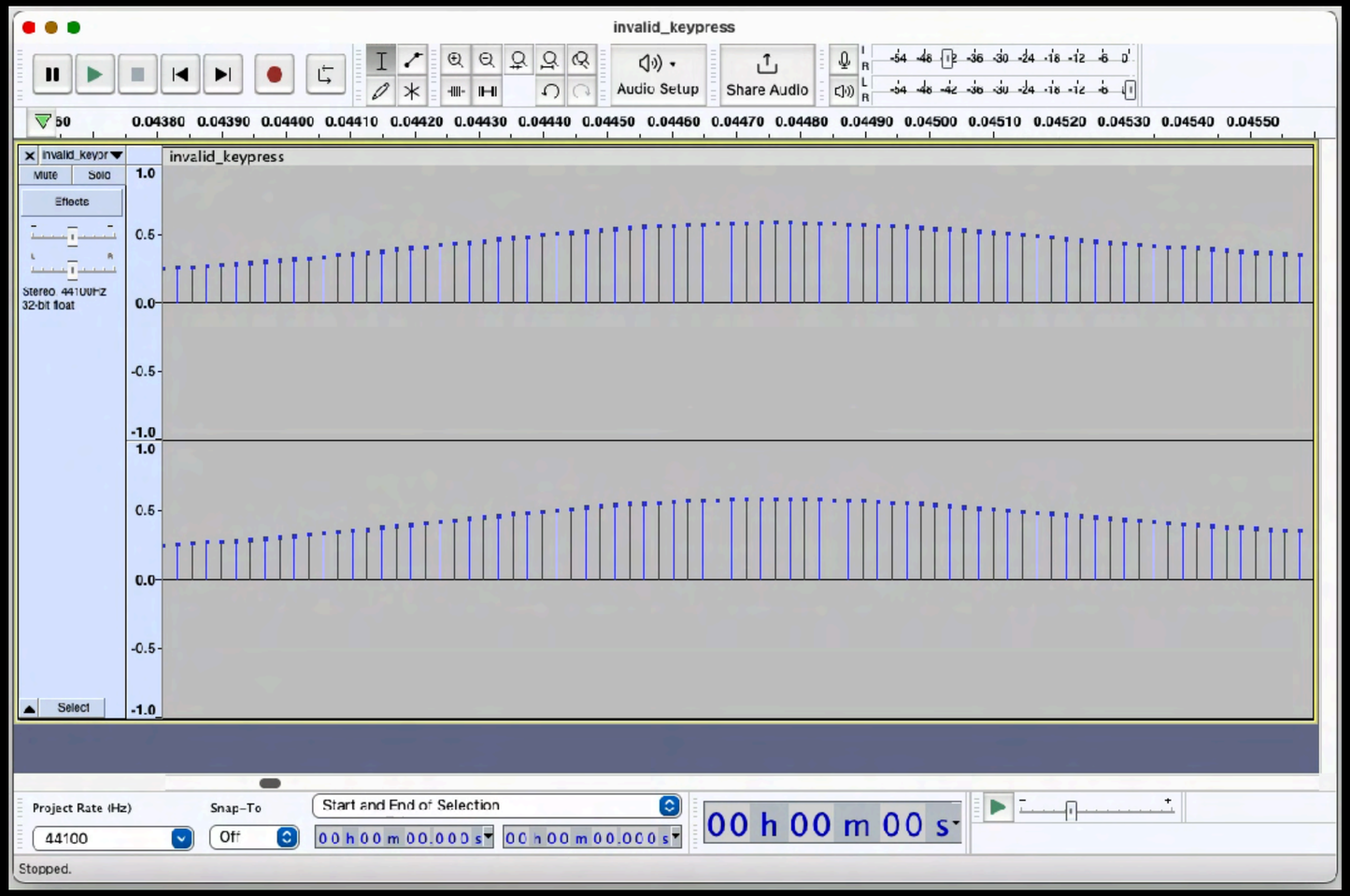
https://www.cs.utah.edu/~natevm/newell_teaset/newell_teaset.zip

```
v 0 2.4 -1.4
vt -0.109561 1.71761
vn 1.39147e-17 -0.369129 0.929378
v 0.229712 2.4 -1.38197
vt 0.120858 1.71761
vn -0.145716 -0.369332 0.917802
v 0.227403 2.43544 -1.36807
vt 0.119643 1.75572
vn -0.150341 -0.284166 0.946915
v 0 2.43544 -1.38593
vt -0.108459 1.75572
vn -1.64188e-16 -0.284002 0.958824
…
```



Sound as Numbers: Samples

(Digital audio recording was invented by Thomas G. Stockham, Jr.)



Words as Numbers

- We could digitize words so each word gets an ID (like ASCII)
  - Alone, this is meaningless
- Or we could digitize the meaning of the word
  - Each word gets a list of “feature measurements”
  - You can measure the distance between words this way
  - Success!

Simple Yes/No Example of Word Features

20 Questions

	Has it been alive?	Man-made?	Ever had a brain?	Warm-blooded?	Do humans eat it?
King	Yes	No	Yes	Yes	No
Orange (fruit)	Yes	No	No	No	Yes
Shirt	No	Yes	No	No	No
Democracy	No	Yes	No	No	No
Orange (color)	No	No	No	No	No

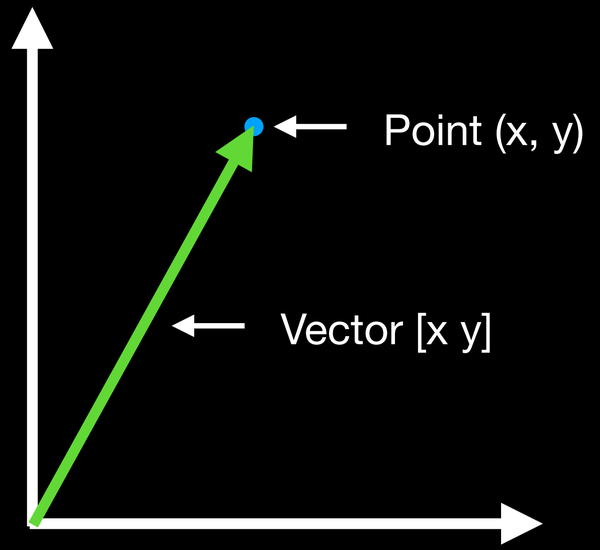
Jargon Basics: Words vs Tokens

- Tokens are words or word parts
- Misspellings
- Unknown words
- platform.openai.com/tokenizer
- Tokens is middle ground between the alphabet and a full dictionary



Jargon Basics: Vector

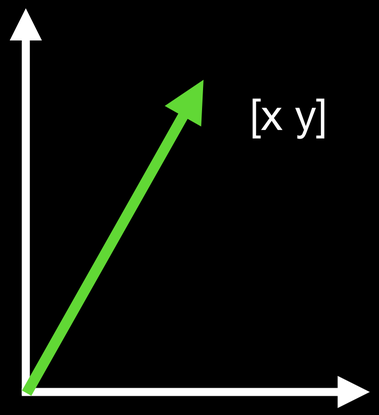
- 2D Vector [x y] is like a 2D point (x, y)
- But it has a direction that starts from (0, 0)
- This is linear algebra



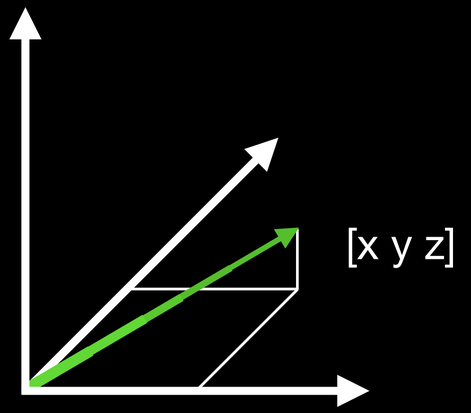
Jargon Basics: N-Dimensional Vector



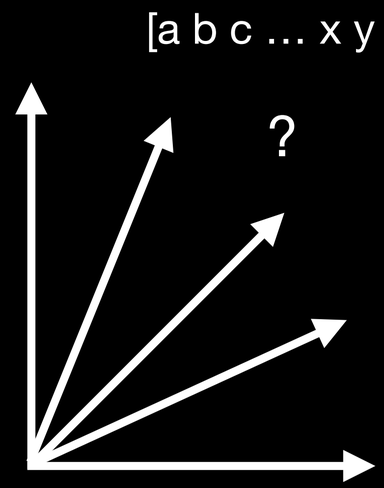
1 Dimension Vector



2 Dimension Vector



3 Dimension Vector



High Dimensional Vector

Jargon Basics: N-Dimensional Vector



- Simply a list of numbers
- Each number is kind of like a map direction
- The number of numbers is how many dimensions it is
- 3D [0.5 0.6 0.2]
- 4D [0.5 0.6 0.2 0.5]
- 10D [0.5 0.6 0.2 0.5 0.6 0.4 0.5 0.5 0.1 0.9]

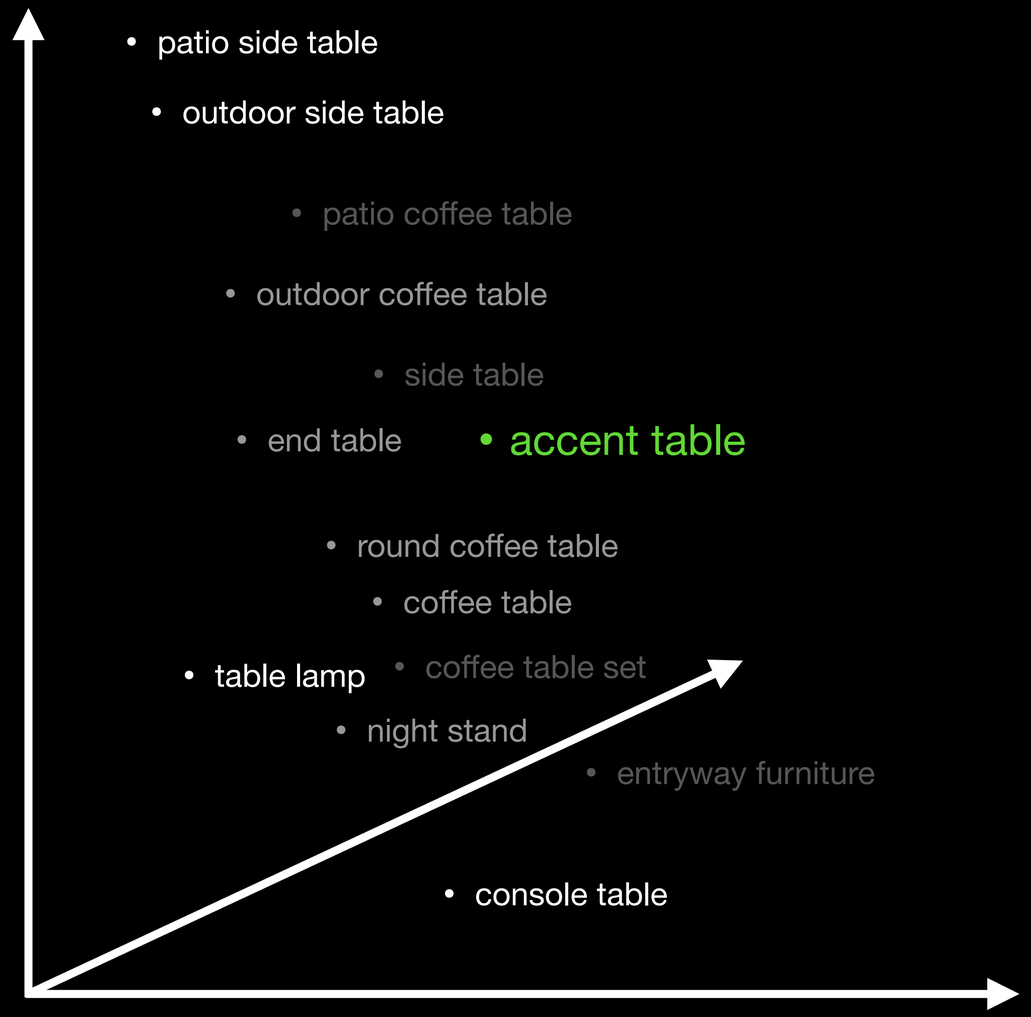
Stanford's glove-6b-300d-txt (2014)

300 dimension vector embedding for "the"


```
[0.04656 0.21318 -0.0074364 -0.45854 -0.035639 0.23643 -0.28836 0.21521 -0.13486 -1.6413 -0.26091 0.032434 0.056621
-0.043296 -0.021672 0.22476 -0.075129 -0.067018 -0.14247 0.038825 -0.18951 0.29977 0.39305 0.17887 -0.17343 -0.21178
0.23617 -0.063681 -0.42318 -0.11661 0.093754 0.17296 -0.33073 0.49112 -0.68995 -0.092462 0.24742 -0.17991 0.097908
0.083118 0.15299 -0.27276 -0.038934 0.54453 0.53737 0.29105 -0.0073514 0.04788 -0.4076 -0.026759 0.17919 0.010977 -0.10963
-0.26395 0.07399 0.26236 -0.1508 0.34623 0.25758 0.11971 -0.037135 -0.071593 0.43898 -0.040764 0.016425 -0.4464 0.17197
0.046246 0.058639 0.041499 0.53948 0.52495 0.11361 -0.048315 -0.36385 0.18704 0.092761 -0.11129 -0.42085 0.13992 -0.39338
-0.067945 0.12188 0.16707 0.075169 -0.015529 -0.19499 0.19638 0.053194 0.2517 -0.34845 -0.10638 -0.34692 -0.19024 -0.2004
0.12154 -0.29208 0.023353 -0.11618 -0.35768 0.062304 0.35884 0.02906 0.0073005 0.0049482 -0.15048 -0.12313 0.19337 0.12173
0.44503 0.25147 0.10781 -0.17716 0.038691 0.08153 0.14667 0.063666 0.061332 -0.075569 -0.37724 0.01585 -0.30342 0.28374
-0.042013 -0.040715 -0.15269 0.07498 0.15577 0.10433 0.31393 0.19309 0.19429 0.15185 -0.10192 -0.018785 0.20791 0.13366
0.19038 -0.25558 0.304 -0.01896 0.20147 -0.4211 -0.0075156 -0.27977 -0.19314 0.046204 0.19971 -0.30207 0.25735 0.68107
-0.19409 0.23984 0.22493 0.65224 -0.13561 -0.17383 -0.048209 -0.1186 0.0021588 -0.019525 0.11948 0.19346 -0.4082 -0.082966
0.16626 -0.10601 0.35861 0.16922 0.07259 -0.24803 -0.10024 -0.52491 -0.17745 -0.36647 0.2618 -0.012077 0.08319 -0.21528
0.41045 0.29136 0.30869 0.078864 0.32207 -0.041023 -0.1097 -0.092041 -0.12339 -0.16416 0.35382 -0.082774 0.33171 -0.24738
-0.048928 0.15746 0.18988 -0.026642 0.063315 -0.010673 0.34089 1.4106 0.13417 0.28191 -0.2594 0.055267 -0.052425 -0.25789
0.019127 -0.022084 0.32113 0.068818 0.51207 0.16478 -0.20194 0.29232 0.098575 0.013145 -0.10652 0.1351 -0.045332 0.20697
-0.48425 -0.44706 0.0033305 0.0029264 -0.10975 -0.23325 0.22442 -0.10503 0.12339 0.10978 0.048994 -0.25157 0.40319 0.35318
0.18651 -0.023622 -0.12734 0.11475 0.27359 -0.21866 0.015794 0.81754 -0.023792 -0.85469 -0.16203 0.18076 0.028014 -0.1434
0.0013139 -0.091735 -0.089704 0.11105 -0.16703 0.068377 -0.087388 -0.039789 0.014184 0.21187 0.28579 -0.28797 -0.058996
-0.032436 -0.0047009 -0.17052 -0.034741 -0.11489 0.075093 0.099526 0.048183 -0.073775 -0.41817 0.0041268 0.44414 -0.16062
0.14294 -2.2628 -0.027347 0.81311 0.77417 -0.25639 -0.11576 -0.11982 -0.21363 0.028429 0.27261 0.031026 0.096782 0.0067769
0.14082 -0.013064 -0.29686 -0.079913 0.195 0.031549 0.28506 -0.087461 0.0090611 -0.20989 0.053913]
```

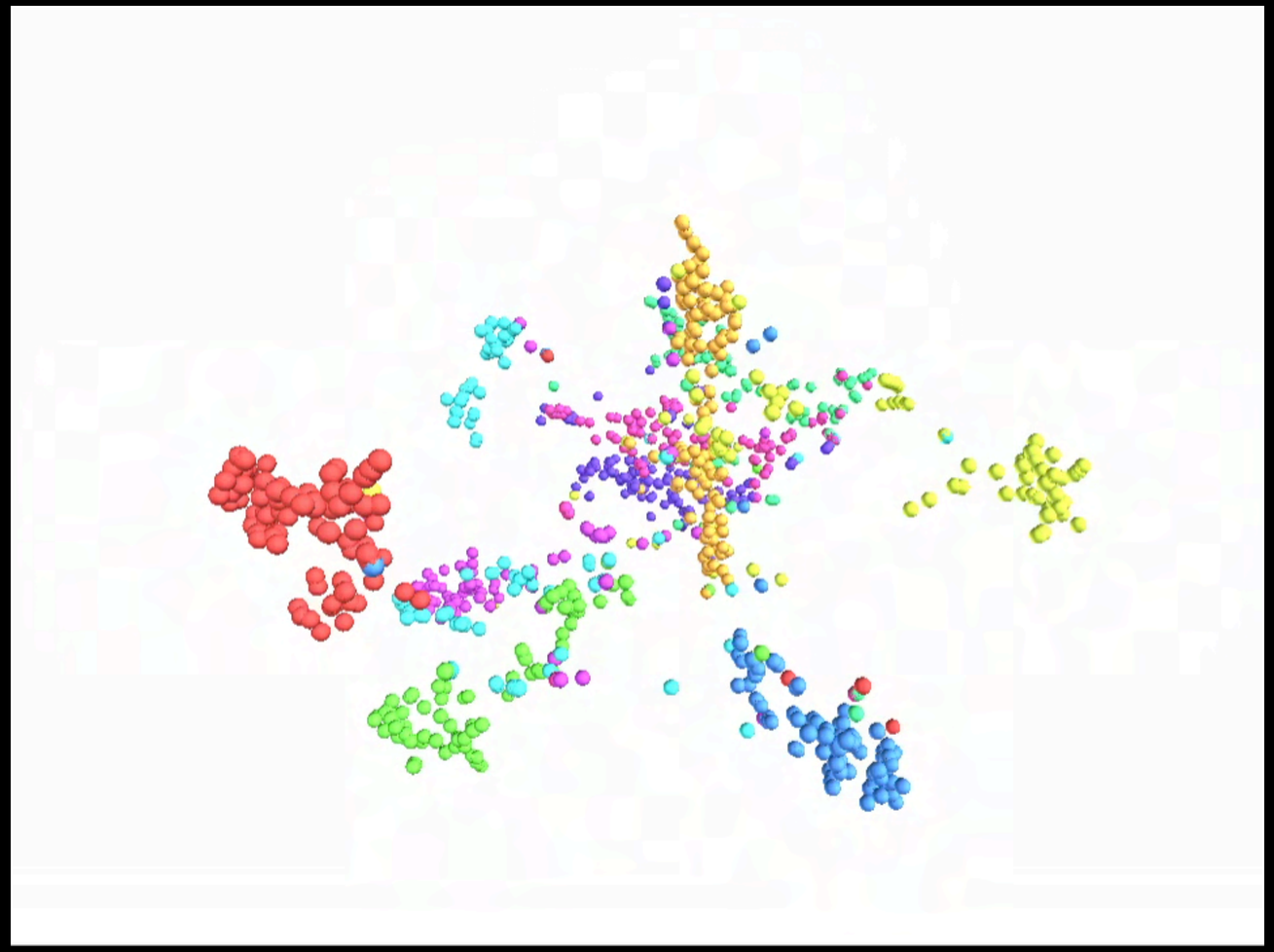

https://www.kaggle.com/datasets/thanakomsn/glove6b300dtxt

Visualizing with Dimensional Reduction



- Products near “accent table”

Visualizing with Dimensional Reduction



colah.github.io/posts/2014-10-Visualizing-MNIST/

Digitizing Meaning

How Do We Digitize Word Features?

- Just analyze how words are used in real text
  - Gutenberg Books
  - StackExchange
  - Wikipedia
  - Reddit
    - ELI5

Do You Know What the Word Tezgüino Means?

- Example from https://lena-voita.github.io/nlp_course/word_embeddings.html

How is Tezgüino Used in Different Contexts?

A bottle of tezgüino is on the table.

Everyone likes tezgüino.

Tezgüino makes you drunk.

We make tezgüino out of corn.

- Do you know what tezgüino means now?

How is Tezgüino Used in Different Contexts?

A bottle of tezgüino is on the table.

Everyone likes tezgüino.

Tezgüino makes you drunk.

We make tezgüino out of corn.

- Tezgüino is a kind of alcoholic beverage made from corn
- With context, you can understand the meaning!

How did you do this?

1. A bottle of _____ is on the table.
2. Everyone likes _____.
3. _____ makes you drunk.
4. We make _____ out of corn.

- What other words fit into this context?

How did you do this?

1. A bottle of _____ is on the table.
2. Everyone likes _____.
3. _____ makes you drunk.
4. We make _____ out of corn.

- What other words fit into this context?

	1.	2.	3.	4.
Tezgüino	1	1	1	1
Loud	0	0	0	0
Motor Oil	1	0	0	1
Tortillas	0	1	0	1
Wine	1	1	1	0

How did you do this?

1. A bottle of _____ is on the table.
2. Everyone likes _____.
3. _____ makes you drunk.
4. We make _____ out of corn.

- Tezgüino and wine near each other

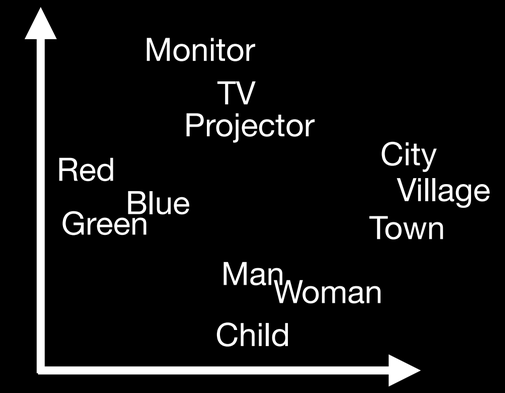
	1.	2.	3.	4.
Tezgüino	1	1	1	1
Loud	0	0	0	0
Motor Oil	1	0	0	1
Tortillas	0	1	0	1
Wine	1	1	1	0

What Are Word Features? (Linguistics)

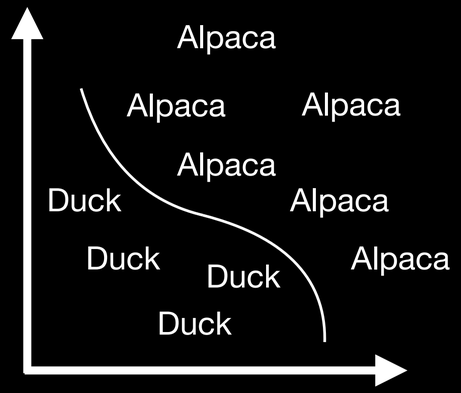
- Semantic meaning is the dictionary or prototype definition
  - se·man·tic *adj.* 1. Of or relating to meaning, especially meaning in language.
- Pragmatics is the contextual meaning
  - Homonym: river bank vs money bank
  - Polysemy: a good shot might not be a good choice
  - Sarcasm, puns, irony, hyperbole, etc: “Good job, Sherlock”
  - Assumptions: “Crowds Rushing to See Pope Trample 6 People to Death”
- Words that sound alike (this is why ChatGPT can rhyme)

Using Digitized “Features”

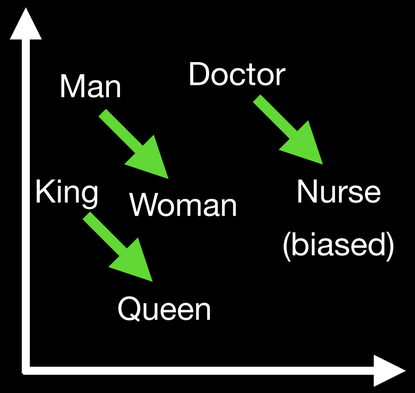
- Similarity
- Classification
- Word relationships
  - King - Man = Royalty
  - King - Man + Woman = Queen
  - Doctor - Man + Woman = Nurse



Similarity Cluster



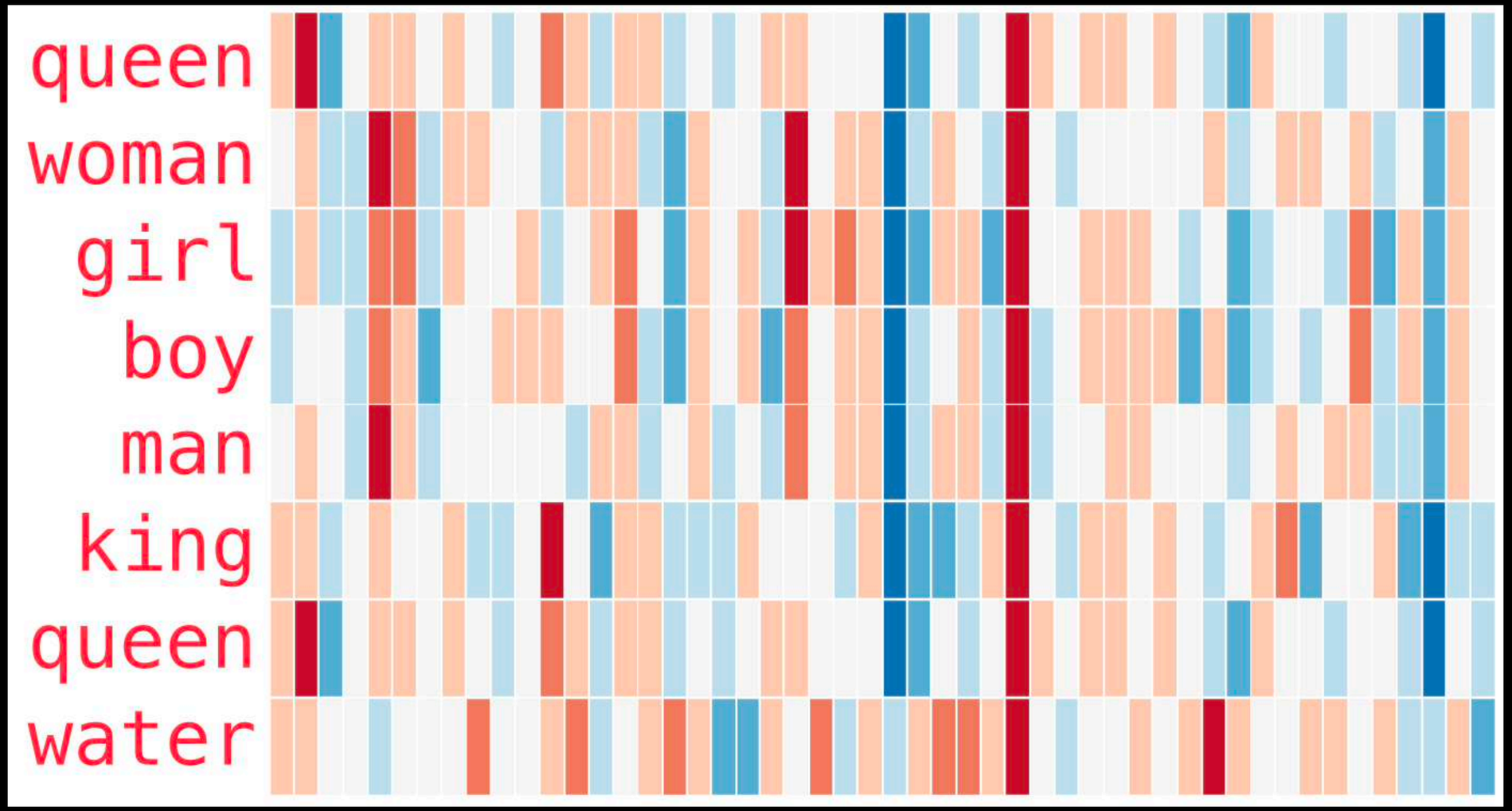
Classification



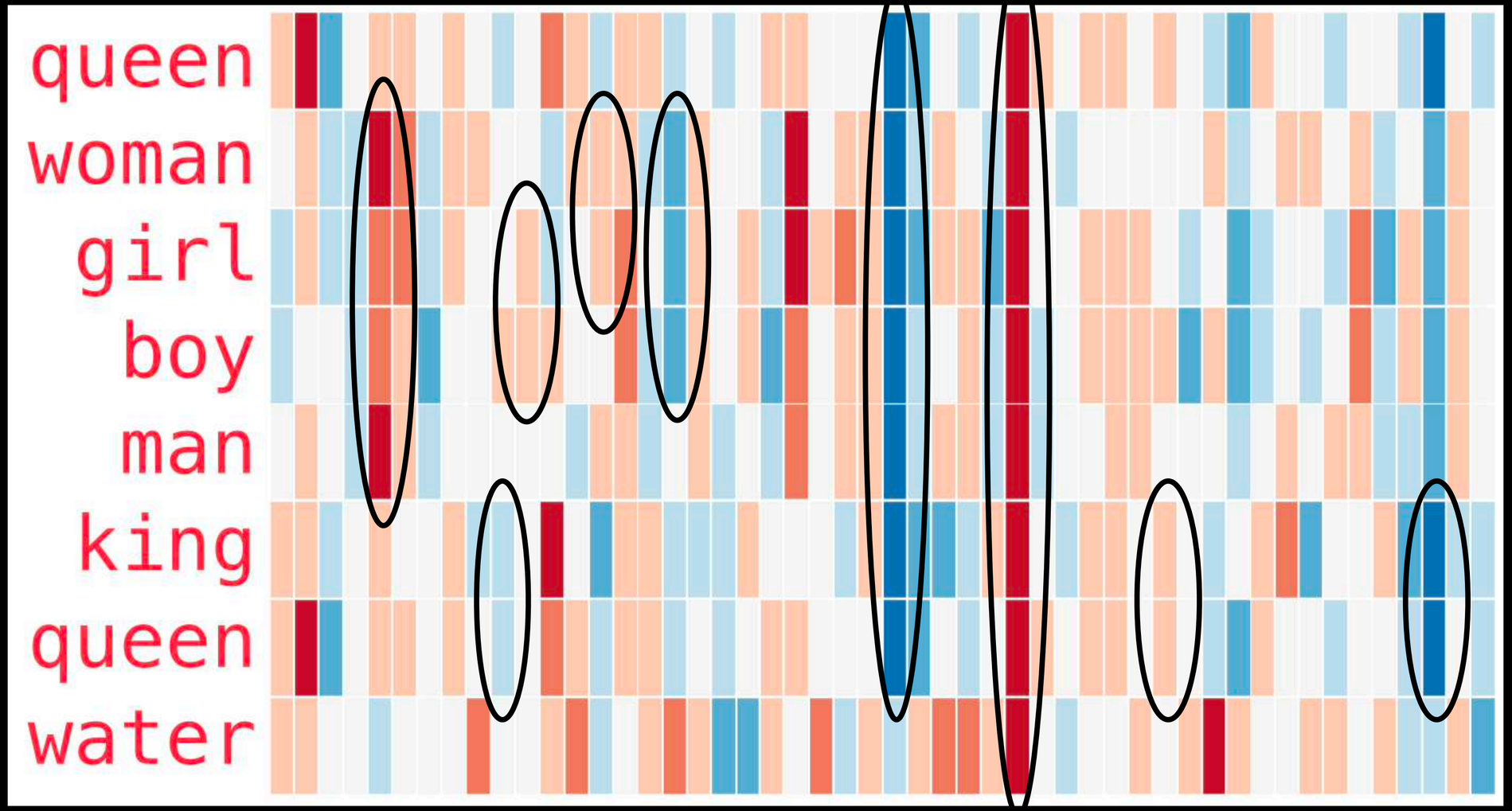
Relationships

What Does Each Dimension Represent?

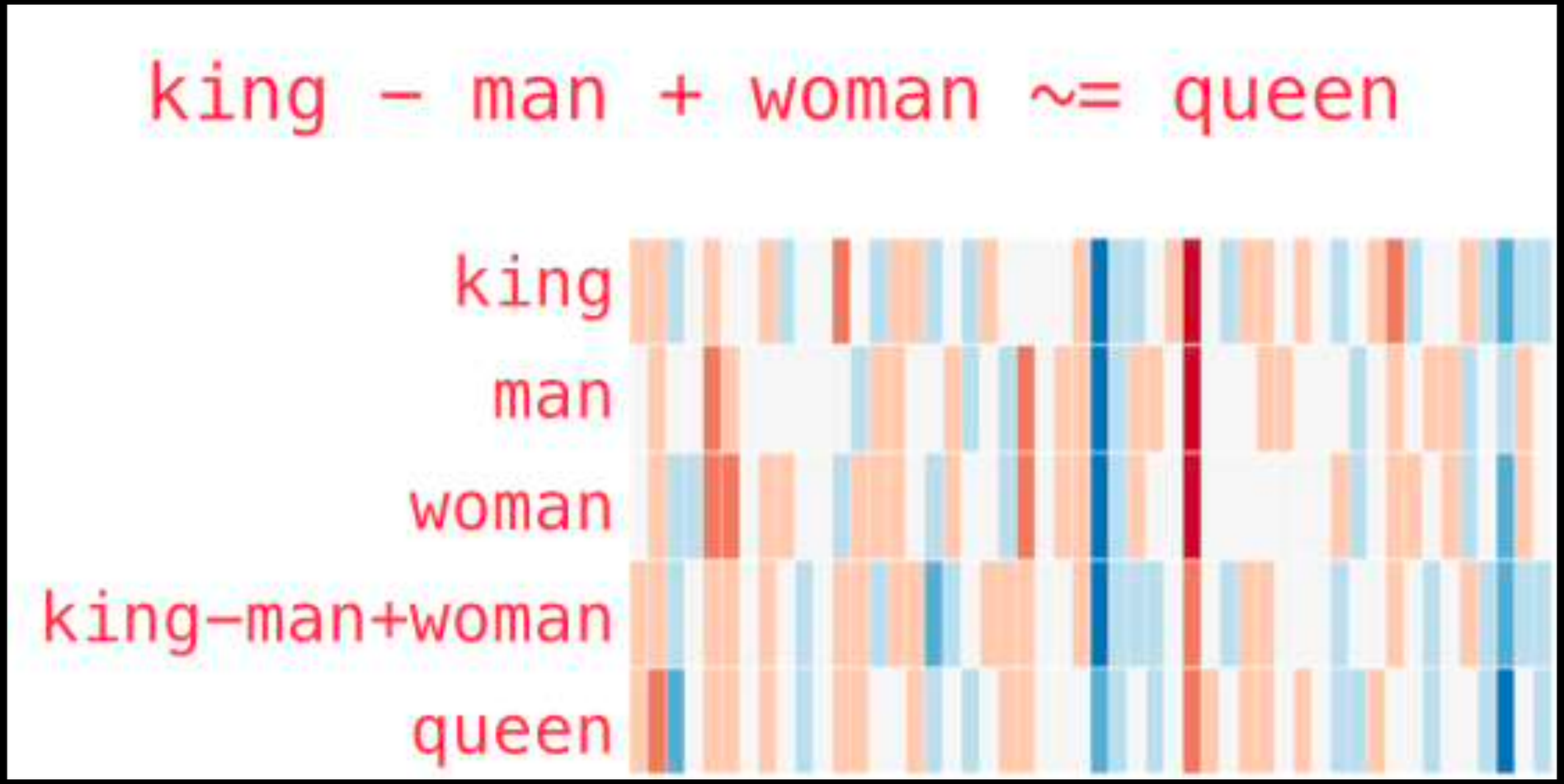
- We don’t know (completely)
- We can try to infer the meaning



https://jalammar.github.io/illustrated-word2vec/



https://jalammar.github.io/illustrated-word2vec/



https://jalammar.github.io/illustrated-word2vec/

Jargon: Vector Embedding

- Embeddings represent the meaning of a token
- It's a trained vector

How to Visualize Training

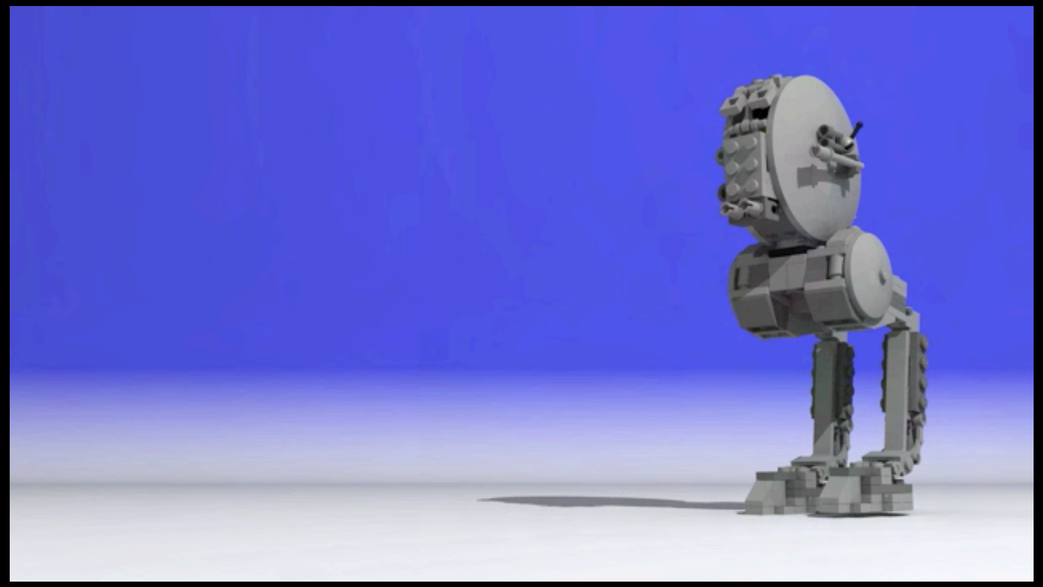
- Everything starts as zeros or random values
- When words are similar, then we move them closer (but only a little bit)
  - Using backpropagation
    - Calculate how far you were from the goal
    - Get the derivative so you know which direction to move
    - Move a little closer to the goal
- Iterate with a lot of different data and eventually the words wiggle into place

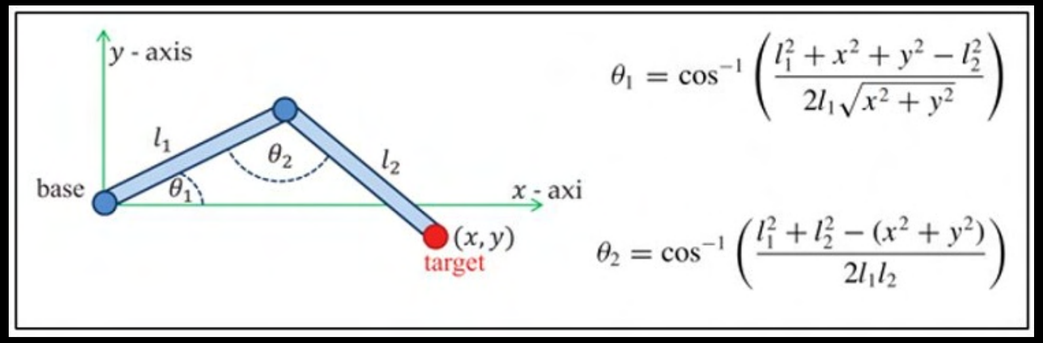
Interactive Visualization

- https://projector.tensorflow.org/

Another Example

- Closed equation
- No ML required

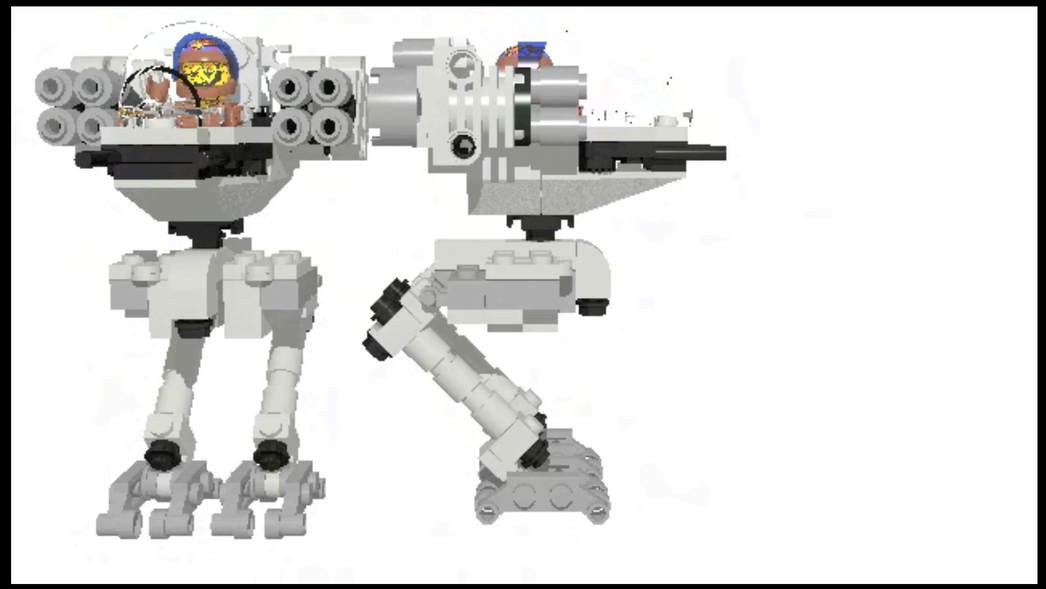


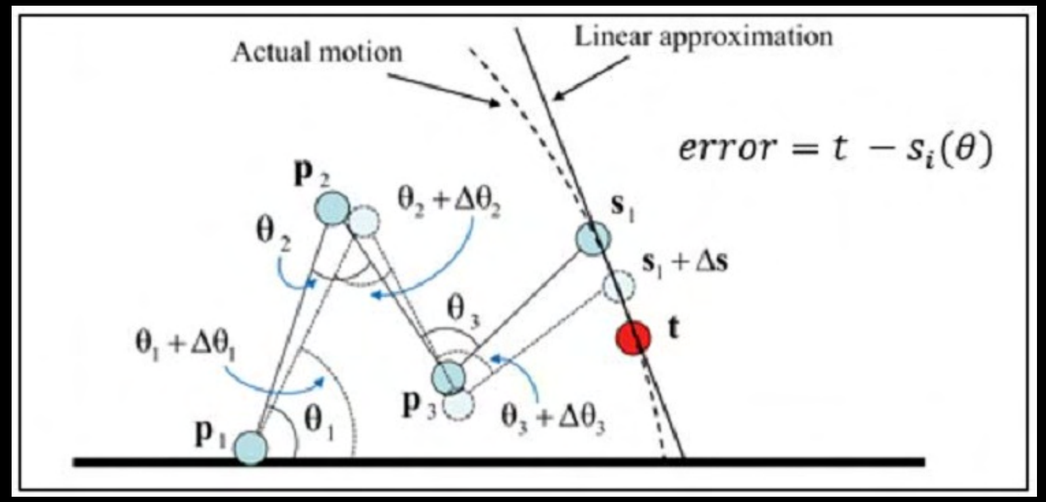


https://www.intel.com/content/www/us/en/developer/articles/technical/accelerating-deep-learning-based-large-scale-inverse-kinematics-with-intel-distribution-of.html

More Lifelike!

- It's not a closed equation
- 8 dimensions
- It requires looping
- Guess each angle and check if it's close (loss function)
- Repeat
- This takes forever!

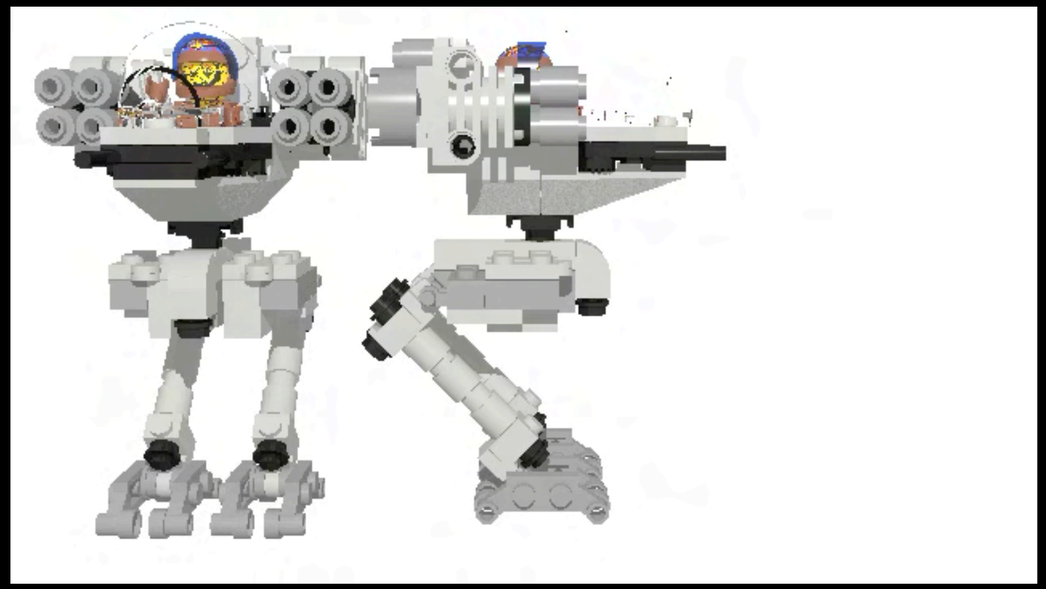




https://www.intel.com/content/www/us/en/developer/articles/technical/accelerating-deep-learning-based-large-scale-inverse-kinematics-with-intel-distribution-of.html

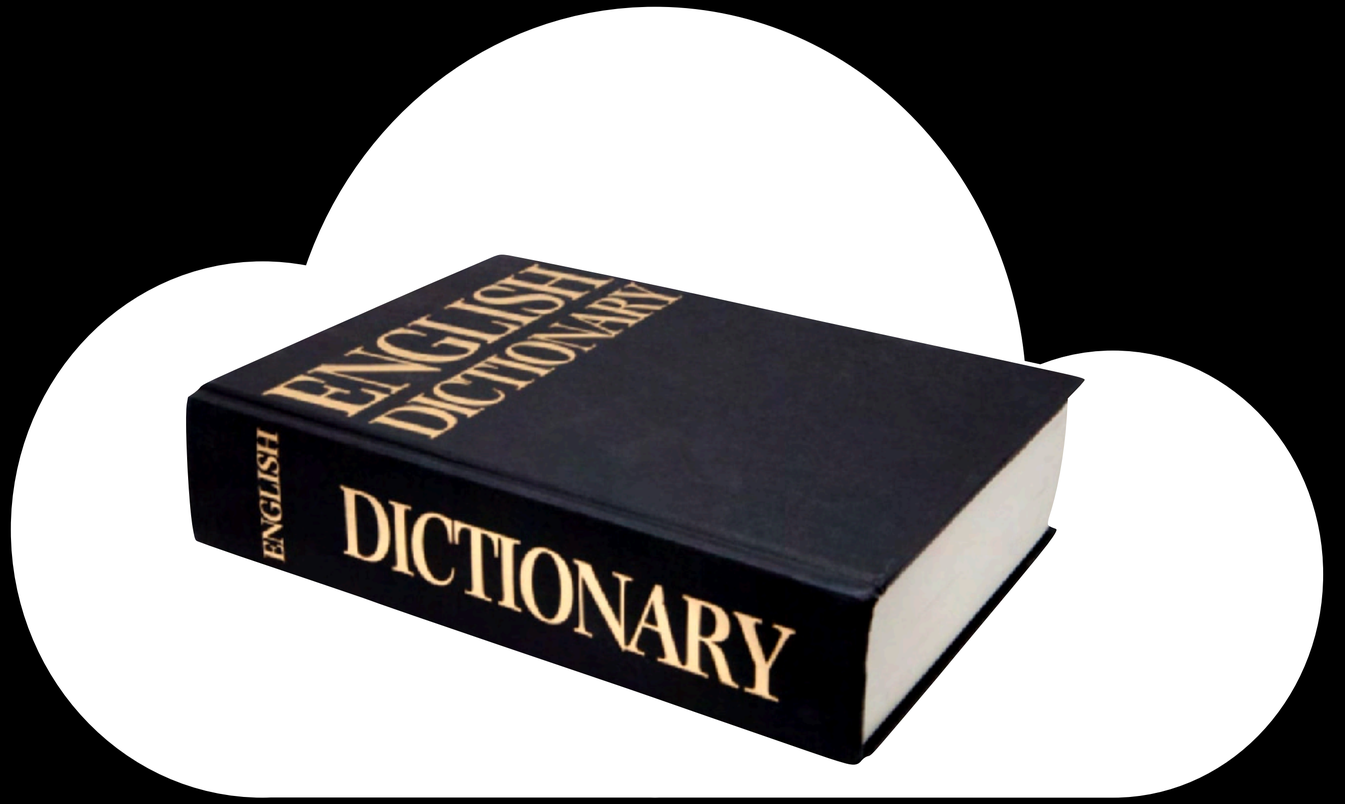
Faster!

- Calculate everything
  - Database of angle combinations
  - Like every chess move, it's impossible
  - I tried and crashed my computer
- What I didn't know...
- Pick random angles and train a NN
- After enough training, it will work!



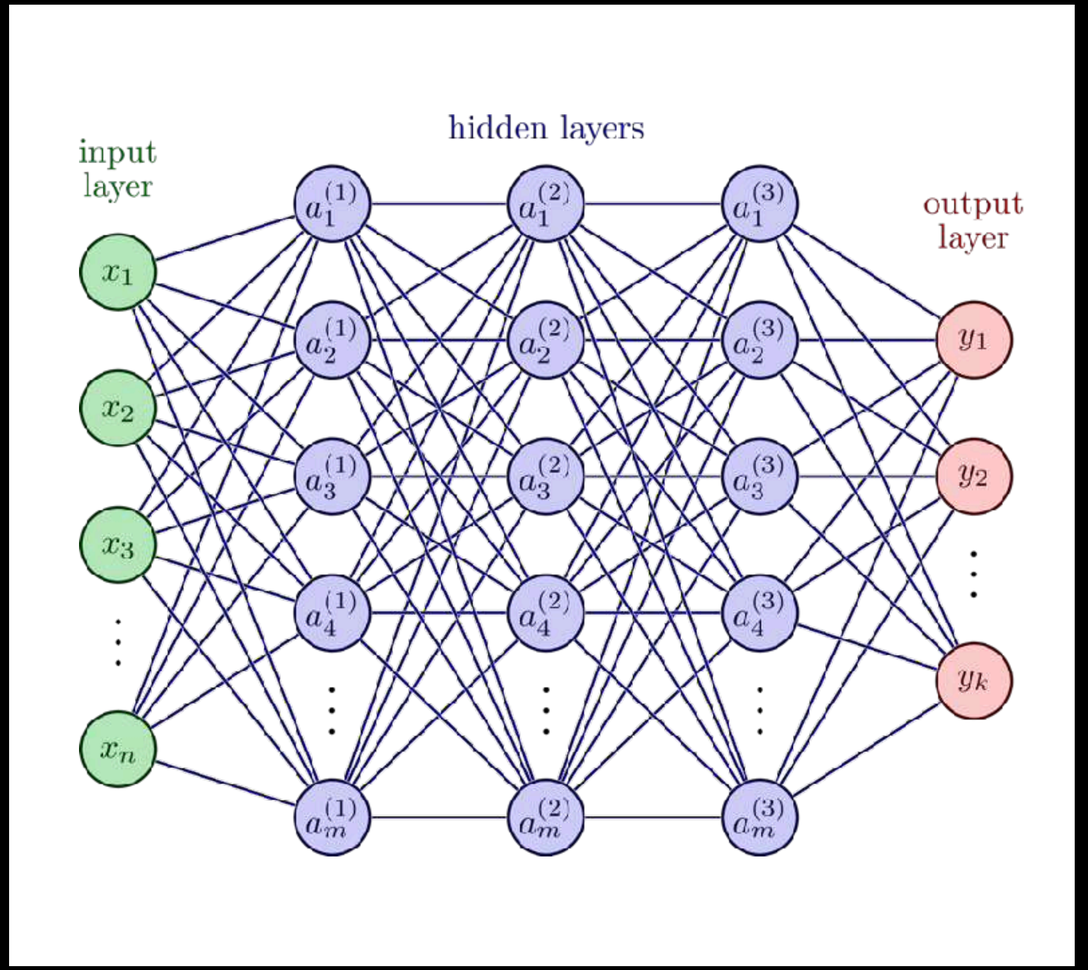
What Does A Neural Network Do?

- All of the words are organized into a big cloud
- To make this cloud useful, you need an input and an output
- Reorganize the cloud based on the input so that the output returns what you want
- Reorganize = vector transform



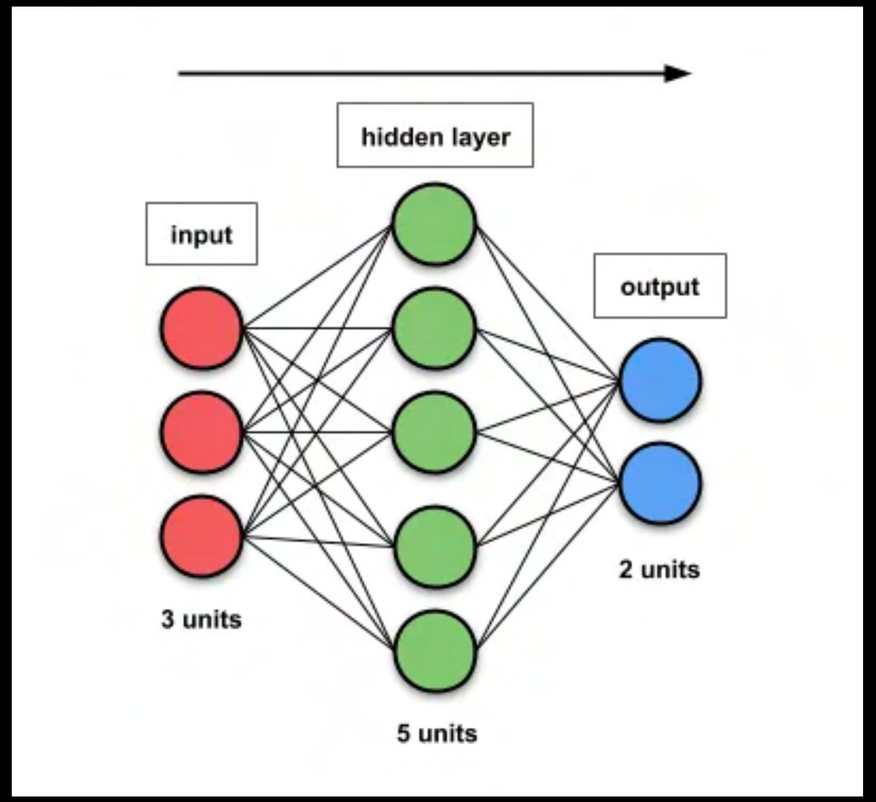
Neural Networks

- NN come in many shapes and sizes
- Have an input & output
- Reorganize the data using matrix math



Jargon: Parameters, Weights, Biases

- Parameters = Trainable numbers
- Each line and circle has a number
- Lines have weights
- Green and blue circles have biases
- Inputs (red circles) are not trainable
- Everything else is trainable
- Weights are used in matrix math



This network has 32 parameters
$(3 \times 5 + 5 \times 2) + (5+2) = 32$

https://towardsdatascience.com/counting-no-of-parameters-in-deep-learning-models-by-hand-8f1716241889

Matrix Math = Multiply and Add The Weights

There is nothing fancy about this

$$\begin{bmatrix} 1 & -2 & 3 \\ -5 & 4 & 2 \\ 2 & -3 & 1 \end{bmatrix} \times \begin{bmatrix} 5 & -3 & 6 \\ -2 & 2 & 5 \\ 4 & 5 & 1 \end{bmatrix} = \begin{bmatrix} 21 & 8 & -1 \\ -25 & 33 & -8 \\ 20 & -7 & -2 \end{bmatrix}$$

$$\begin{aligned} 1 \times 6 &= 6 \\ -2 \times 5 &= -10 \\ 3 \times 1 &= 3 \end{aligned}$$

$6 - 10 + 3 = -1$

Matrix Math = Multiply and Add The Weights

There is nothing fancy about this

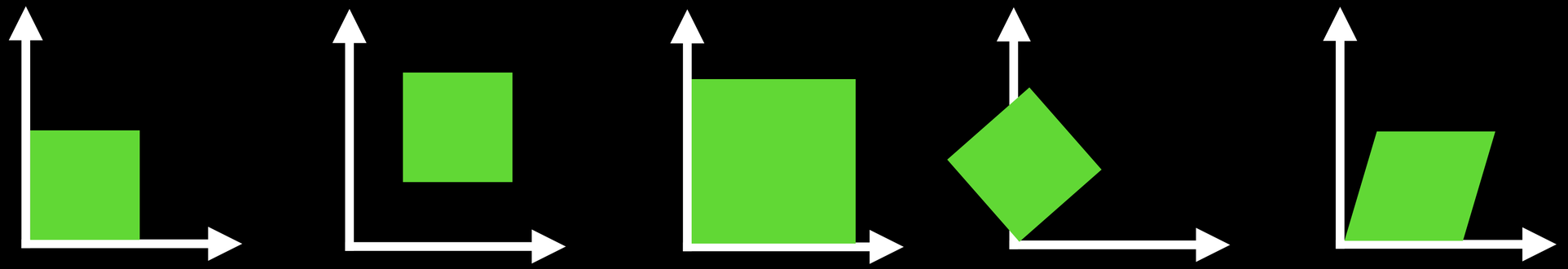
$$\begin{bmatrix} 1 & -2 & 3 \\ -5 & 4 & 2 \\ 2 & -3 & 1 \end{bmatrix} \times \begin{bmatrix} 5 & -3 & 6 \\ -2 & 2 & 5 \\ 4 & 5 & 1 \end{bmatrix} = \begin{bmatrix} 21 & 8 & -1 \\ -25 & 33 & -8 \\ 20 & -7 & -2 \end{bmatrix}$$

$$\begin{aligned} -5 \times -3 &= 15 \\ 4 \times 2 &= 8 \\ 2 \times 5 &= 10 \end{aligned}$$

$$15 + 8 + 10 = 33$$

Reorganizes the Input Data

2D Vector Transformations

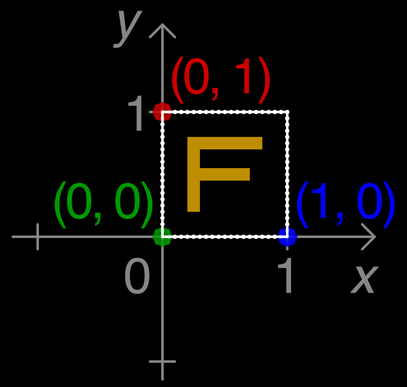


Types of Vector Transformations

This is matrix math (linear algebra)

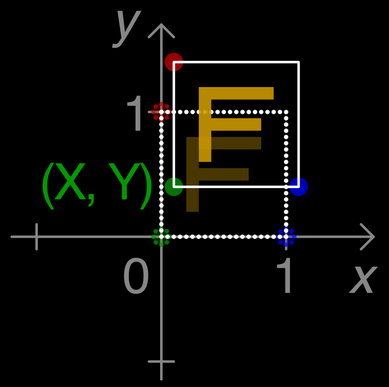
No change

$$\begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}$$



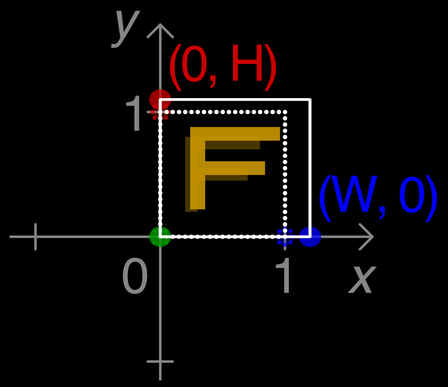
Translate

$$\begin{bmatrix} 1 & 0 & X \\ 0 & 1 & Y \\ 0 & 0 & 1 \end{bmatrix}$$



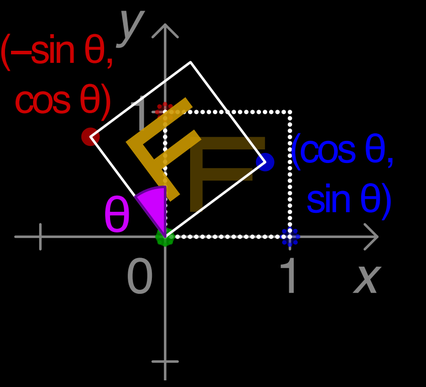
Scale about origin

$$\begin{bmatrix} W & 0 & 0 \\ 0 & H & 0 \\ 0 & 0 & 1 \end{bmatrix}$$



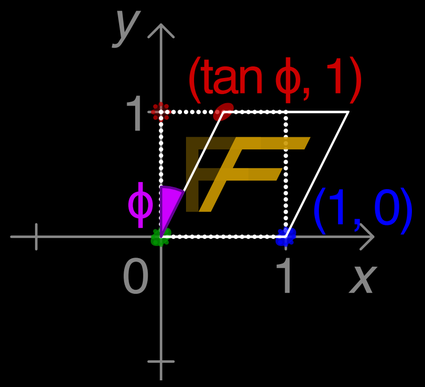
Rotate about origin

$$\begin{bmatrix} \cos\theta & -\sin\theta & 0 \\ \sin\theta & \cos\theta & 0 \\ 0 & 0 & 1 \end{bmatrix}$$



Shear in x direction

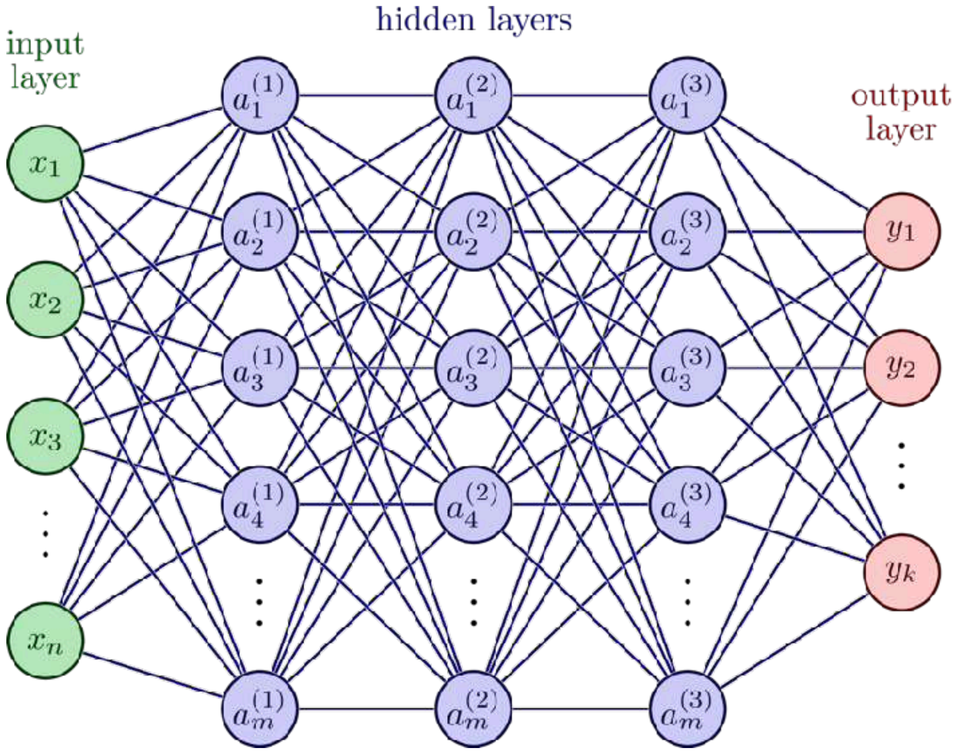
$$\begin{bmatrix} 1 & \tan\phi & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}$$

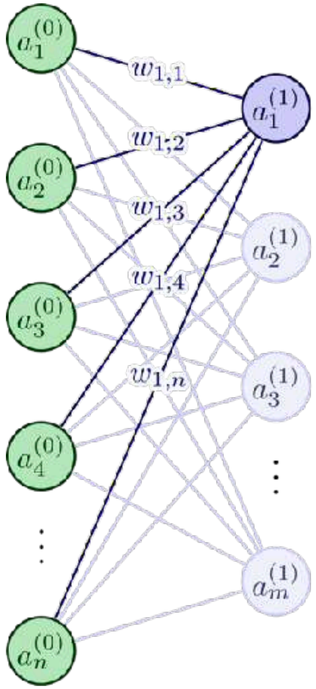




1.00	0.00	0.00	1.00	0.00	0.00	1.00	0.00	0.00	1.00	0.00	0.00	1.00	0.00	0.00
0.00	1.00	0.00	0.00	1.00	0.00	0.00	1.00	0.00	0.00	1.00	0.00	0.00	1.00	0.00
0.00	0.00	1.00	0.00	0.00	1.00	0.00	0.00	1.00	0.00	0.00	1.00	0.00	0.00	1.00

But It's A Lot of Matrix Math





$$= \sigma\left(w_{1,0}a_0^{(0)} + w_{1,1}a_1^{(0)} + \ldots + w_{1,n}a_n^{(0)} + b_1^{(0)}\right)$$

$$= \sigma\left(\sum_{i=1}^{n} w_{1,i}a_i^{(0)} + b_1^{(0)}\right)$$

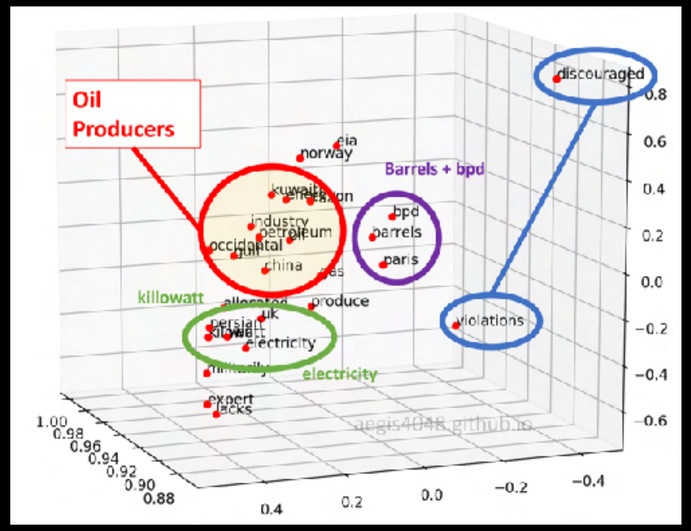
$$\begin{pmatrix} a_1^{(1)} \\ a_2^{(1)} \\ \vdots \\ a_m^{(1)} \end{pmatrix} = \sigma\left[\begin{pmatrix} w_{1,0} & w_{1,1} & \cdots & w_{1,n} \\ w_{2,0} & w_{2,1} & \cdots & w_{2,n} \\ \vdots & \vdots & \ddots & \vdots \\ w_{m,0} & w_{m,1} & \cdots & w_{m,n} \end{pmatrix}\begin{pmatrix} a_1^{(0)} \\ a_2^{(0)} \\ \vdots \\ a_n^{(0)} \end{pmatrix} + \begin{pmatrix} b_1^{(0)} \\ b_2^{(0)} \\ \vdots \\ b_m^{(0)} \end{pmatrix}\right]$$

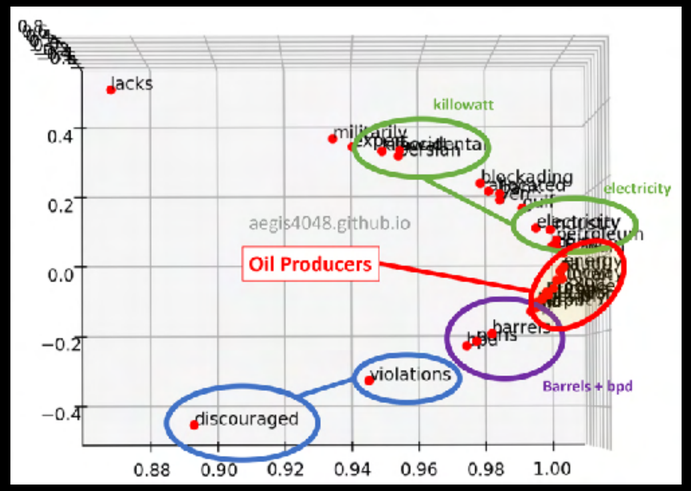
$$a^{(1)} = \sigma\left(\mathbf{W}^{(0)}a^{(0)} + \mathbf{b}^{(0)}\right)$$

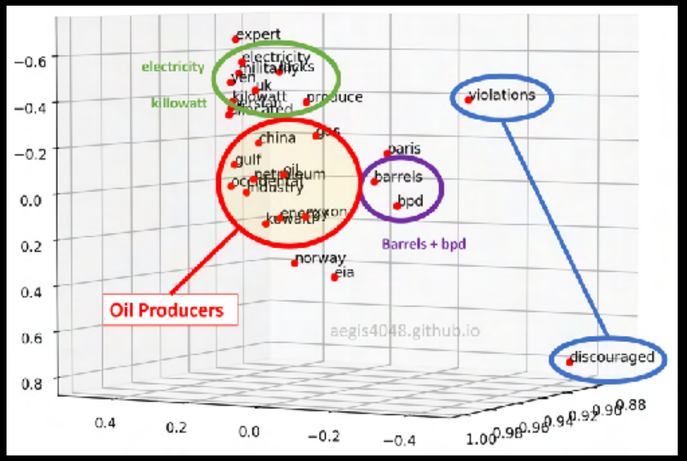
https://tikz.net/neural_networks/

A Simple Rotation Projected to 2D

- This matters because the words move, allowing us to focus on what we want

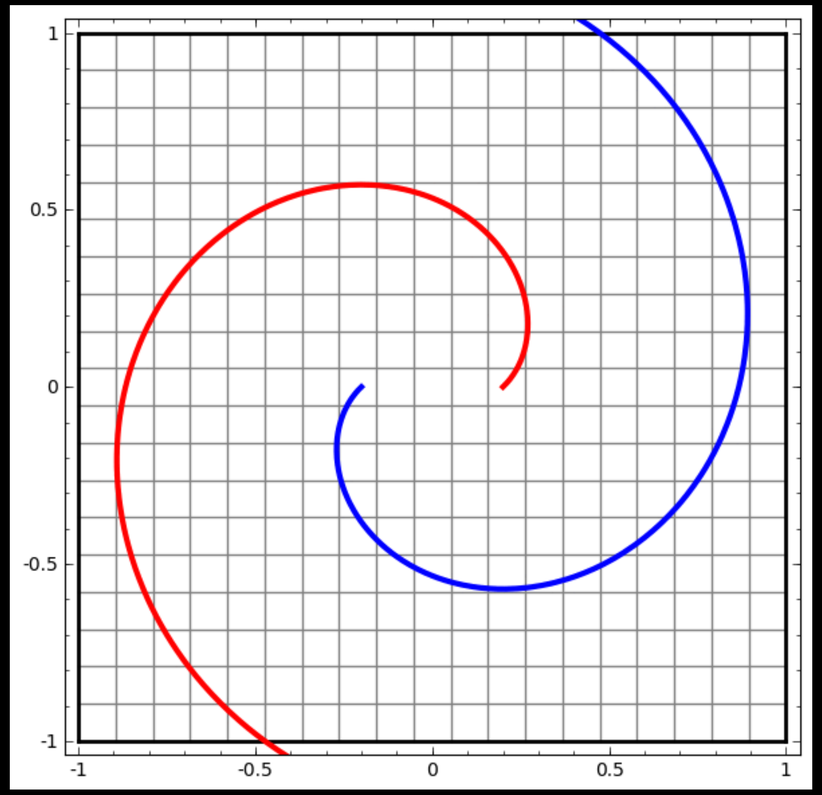






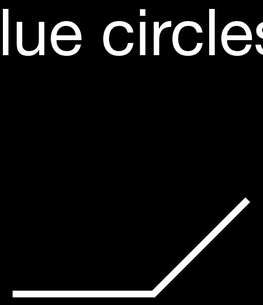
Multiple Transformations Unwrap the Spiral

- Linear = line
- Linear algebra = straight line math
- This example has some nonlinear math

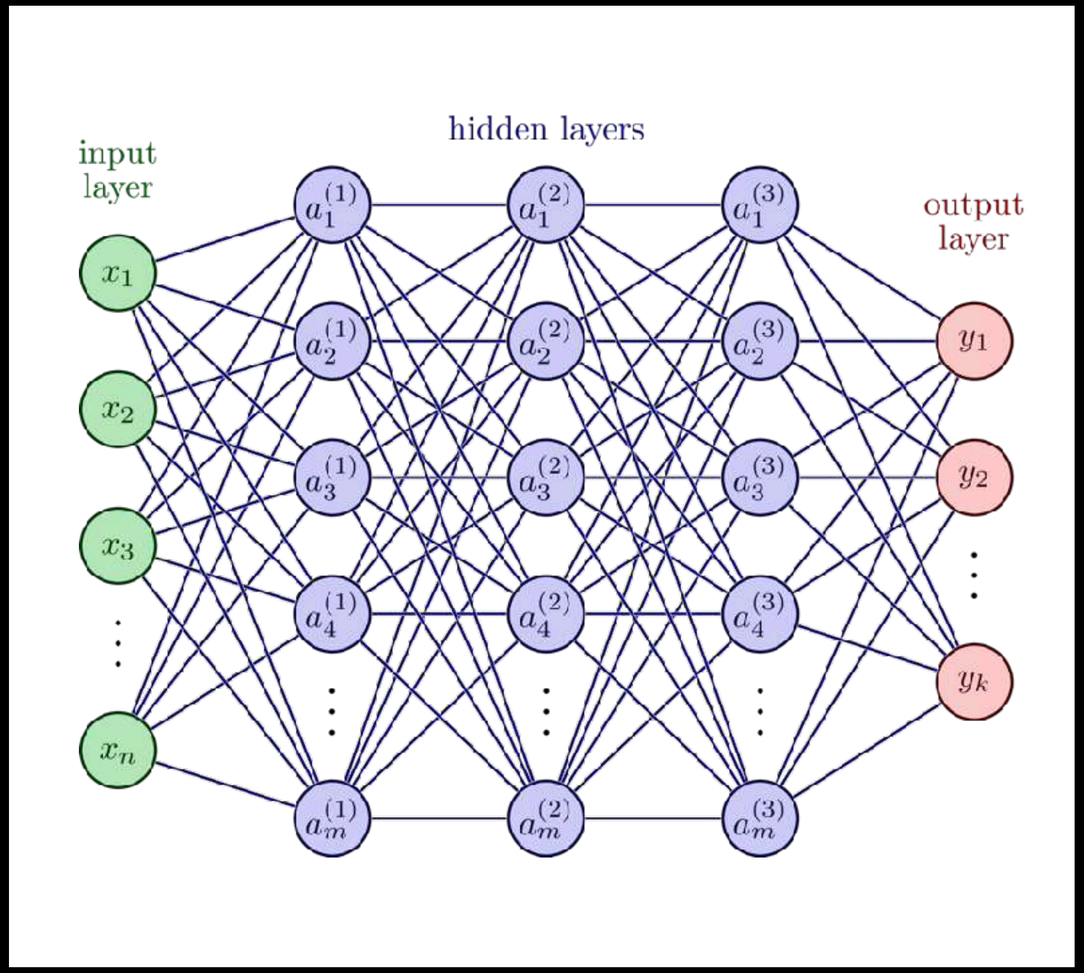


Neural networks Can Do Nonlinear Stuff

- Activation functions (the blue circles)
- They're really simple
  - ReLU: if (x > 0) ? x : 0



  - It's all about speed
- Activation functions use the biases



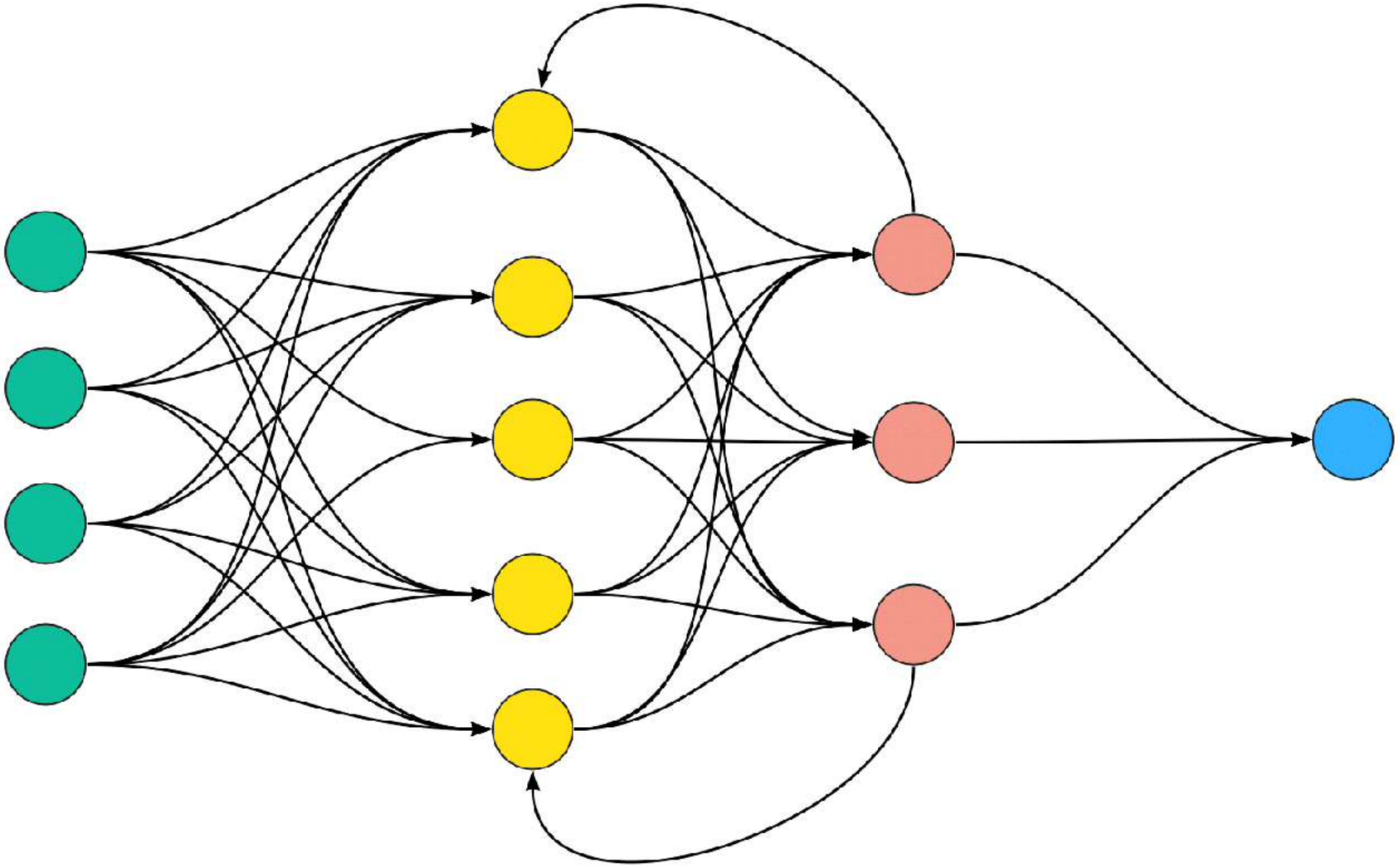
Just How Much Can We Digitize?

- Semantic meaning is the dictionary or prototype definition
- Pragmatics is the contextual meaning
- What about phrases, sentences, paragraphs, documents?
  - Yes, digitizable
- If we can digitize analytical documents…
  - Can we digitize logic, reason and planning?
    - The scientists are debating this

Remembering State

- To digitize phrases, they had to process more than 1 word at a time
- Recurrent neural networks (RNN)
  - This is basically a feedback loop
  - Each word modifies the vector space one after another
  - This started to show some intelligence and had real world use
- Unfortunately, the feedback loop sometimes caused math failures

Recurrent Neural Network



Input Layer

Hidden Layers

Output Layer

dataaspirant.com

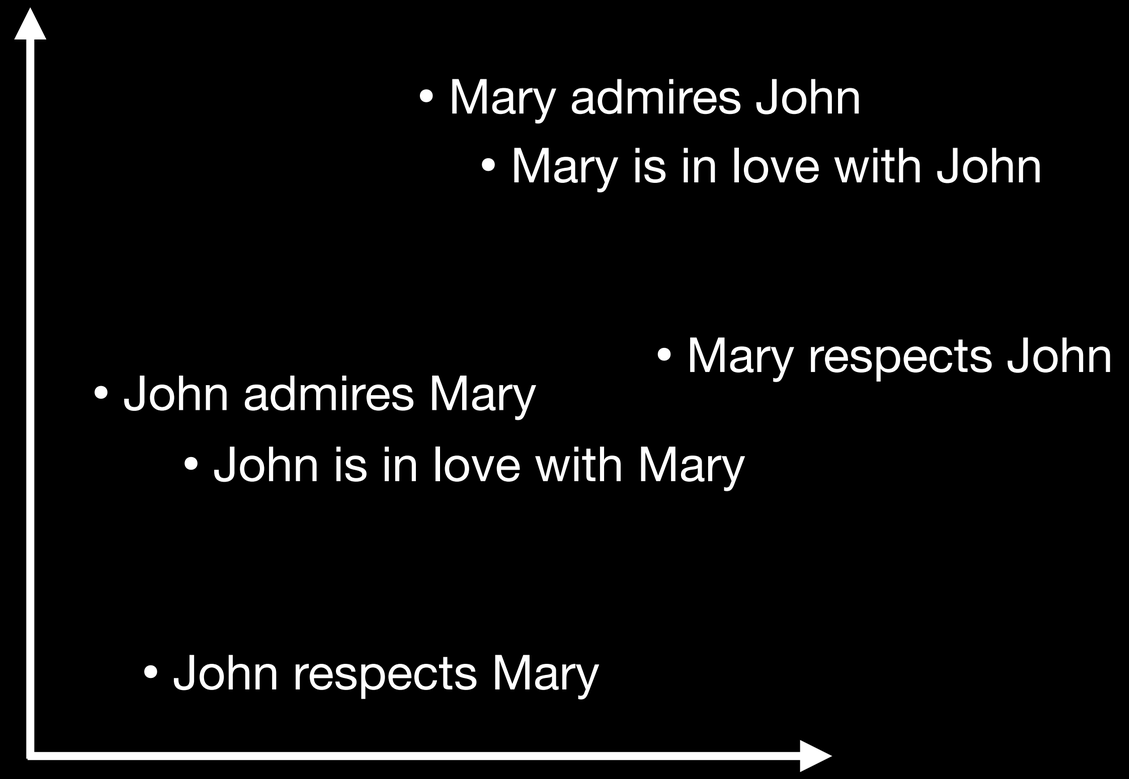
Long Short-Term Memory

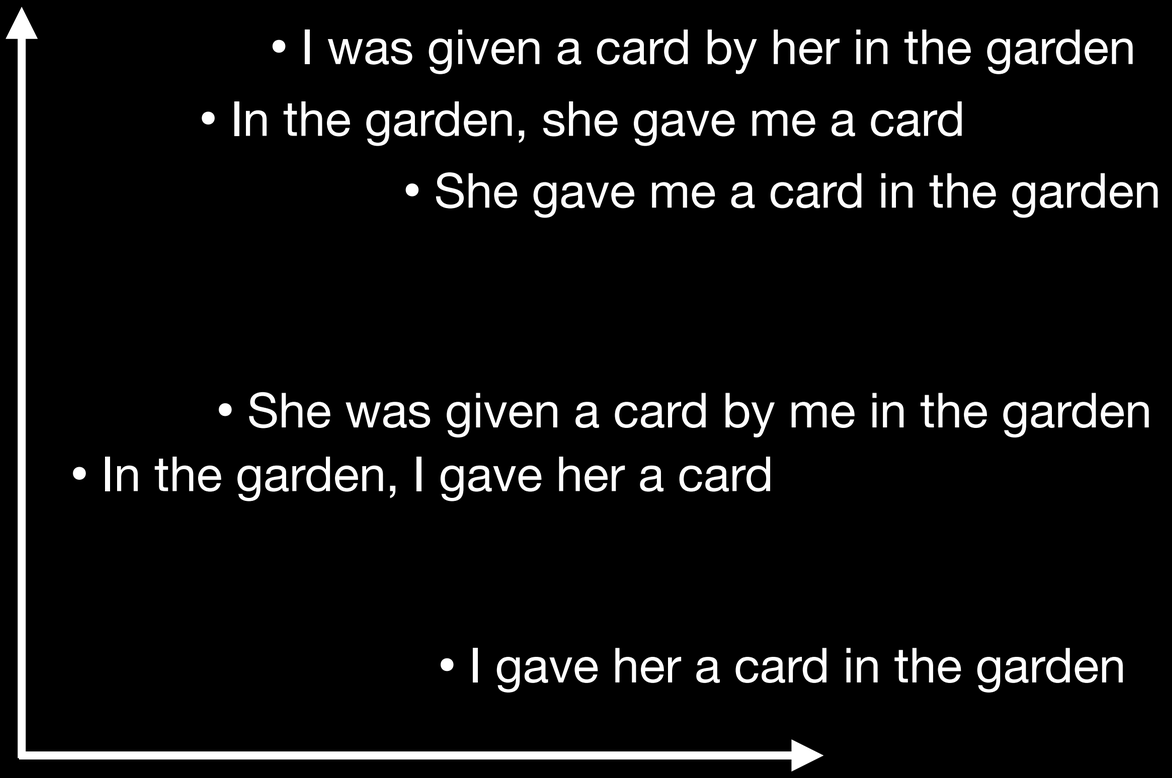
- Still a feedback loop like a RNN
  - Adds short-term vector that can “forget” old words
  - This improved how the vector space was modified



Sequence to Sequence Learning w/ NNs (2014)

Maps phrases to vectors using an LSTM network





arXiv:1409.3215 [cs.CL] - https://doi.org/10.48550/arXiv.1409.3215

Sequence to Sequence Learning w/ NNs (2014)

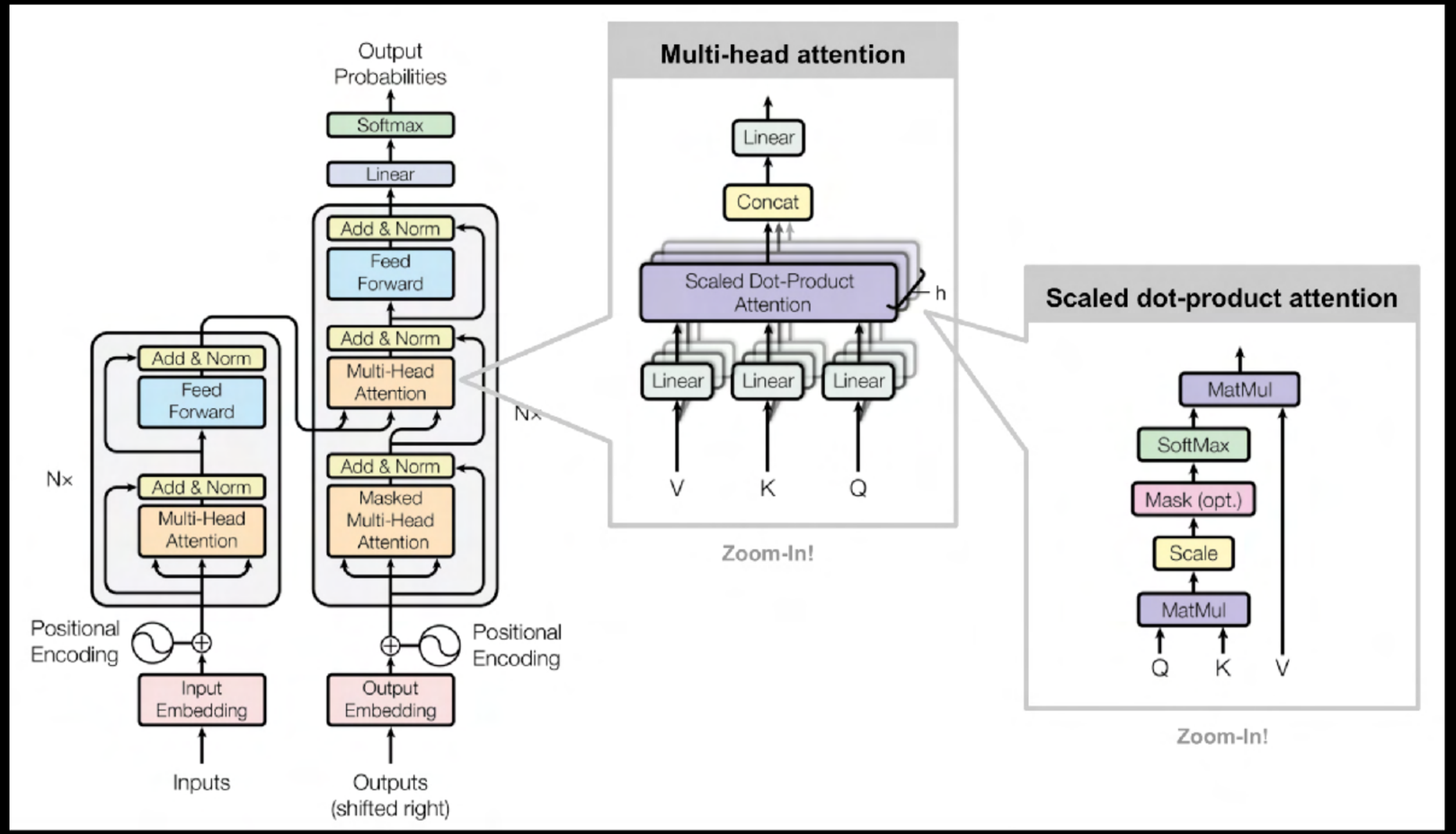
Maps phrases to vectors using an LSTM network

- This created context vectors (digitized phrase meanings)
  - This worked really well for translation (Google Translate)
- Stats
  - 160,000 input tokens
  - 80,000 output tokens
  - 1000 dimension vectors
  - 384m parameters

arXiv:1409.3215 [cs.CL] - https://doi.org/10.48550/arXiv.1409.3215

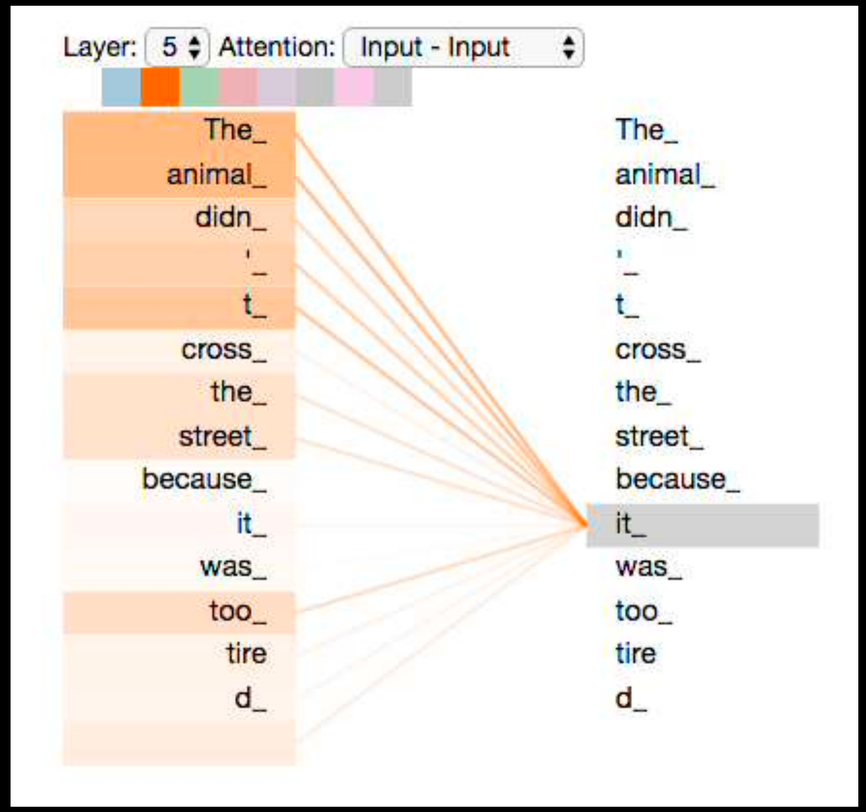
“Attention is All You Need” (2017)

Introducing The Transformer



“Attention is All You Need” (2017)

Introducing The Transformer





https://jalammar.github.io/illustrated-transformer/

Transformers Was A Grand Slam Home Run

Now scale up!

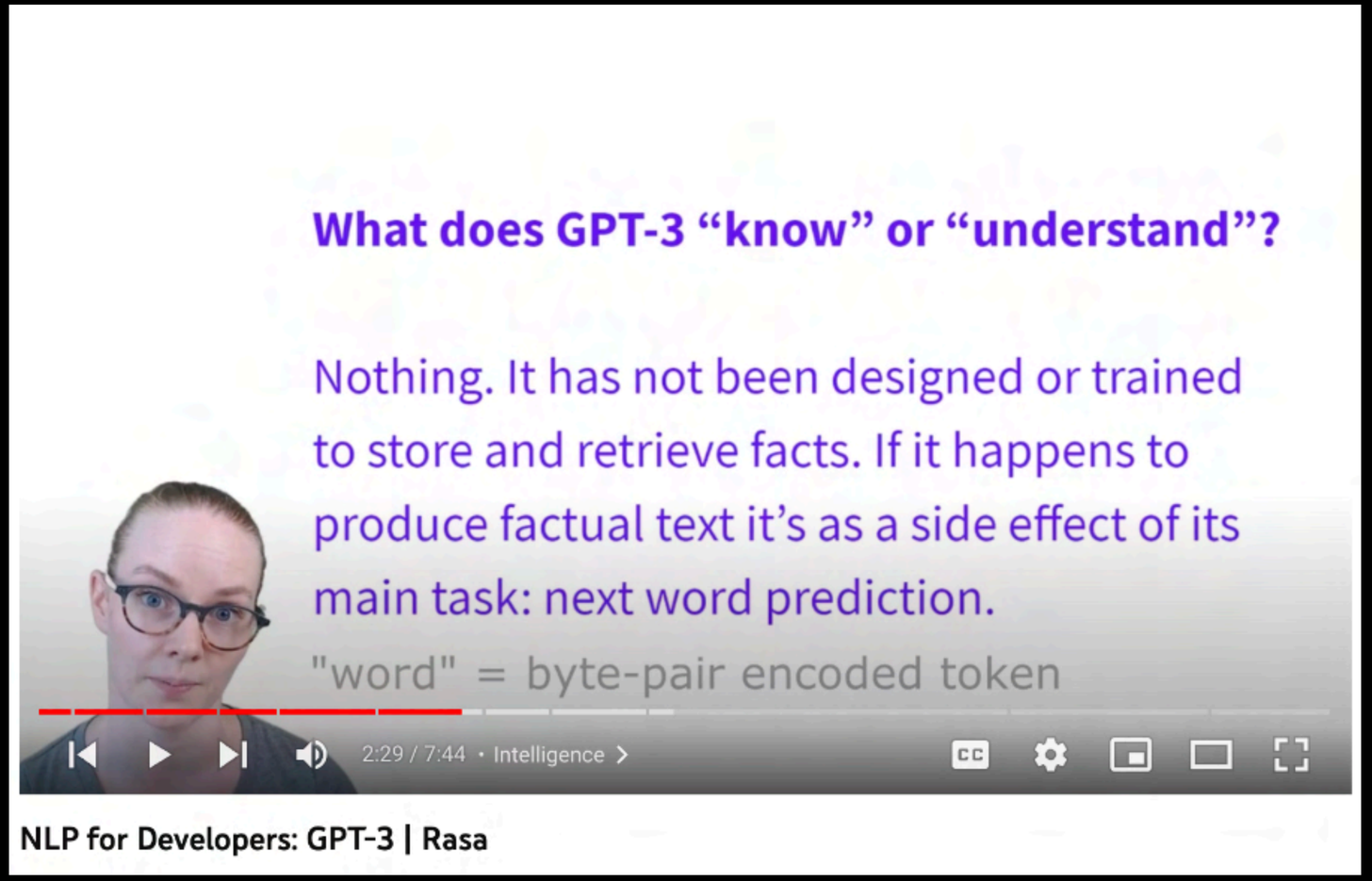
- GPT-1 released in 2018
- GPT-2 released in 2019, 1.5b parameters, trained on 10b tokens
- GPT-3 released in 2020, 175b parameters, 300b tokens, 12,288d vectors
  - ChatGPT released November 2022 - The public finally notices!
- GPT-4 released in 2023, no longer open source
- Claude 3/Gemini 1.5 released in February/March 2024
- Llama 3 released in April 2024, open source, 70b parameters, 15t tokens

What’s In The Vector Space?

It’s more than just word features now

- Grammar
- Sentence, paragraph, document structures, styles and features (e.g. poetry)
- Facts and not facts (“hallucinations”)
- Reason and Logic?
- Biases and hate speech (it’s all in the training data)
- Glitches - an accidental discovery (“!yt Glitch Tokens Computerphile”)

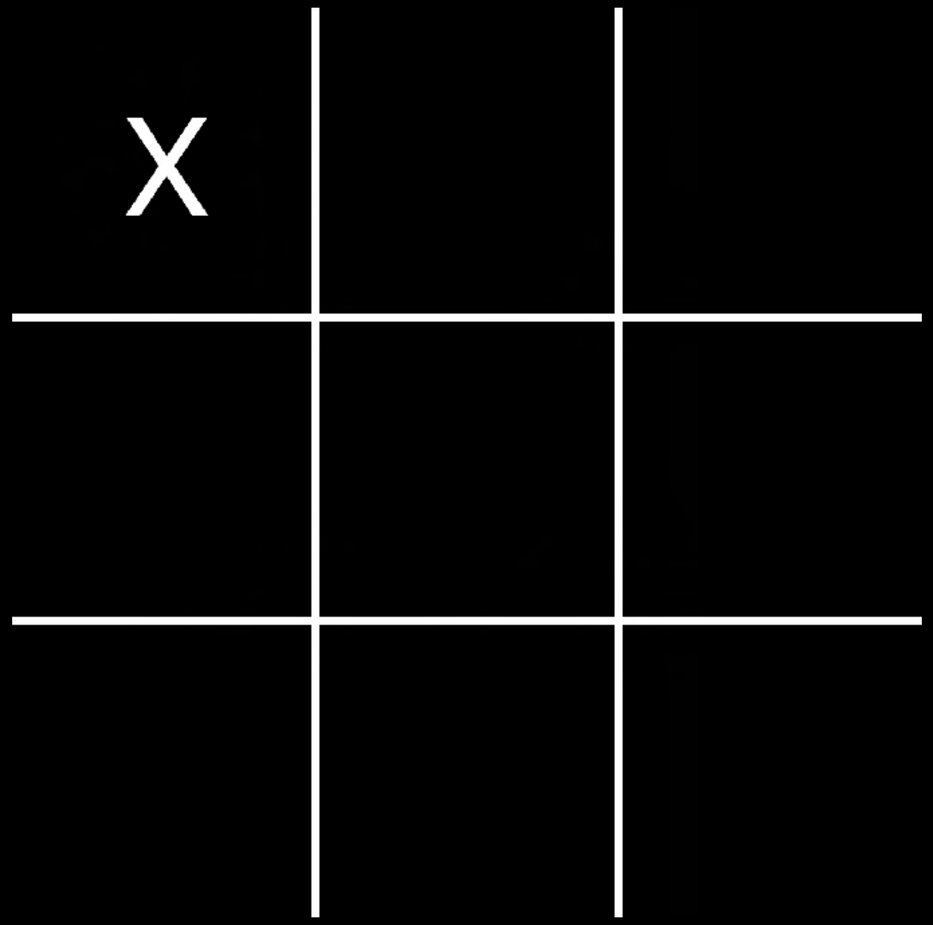
Facts vs “hallucinations” (August 2020)



www.youtube.com/watch?v=ZNeNMTSMA5Y

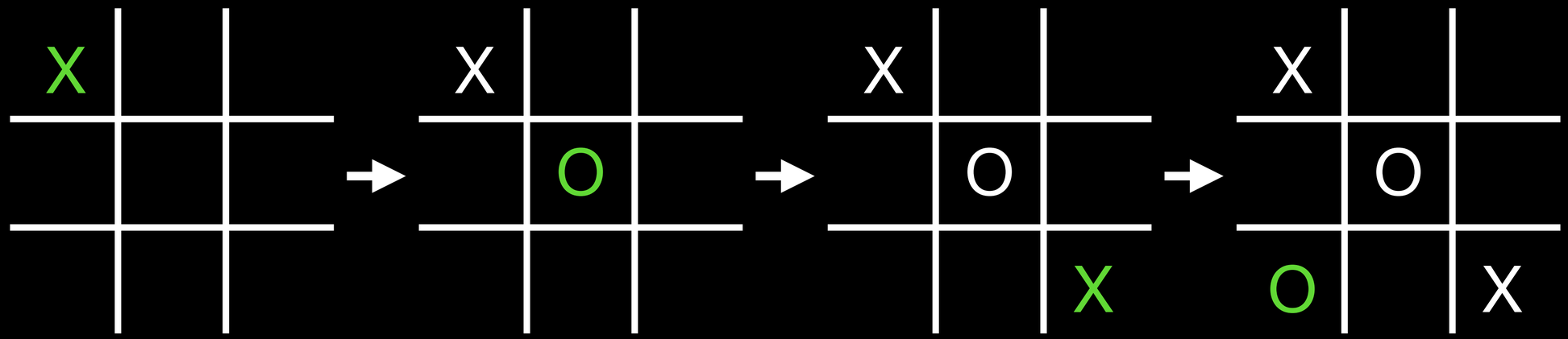
Why Next Word Prediction?

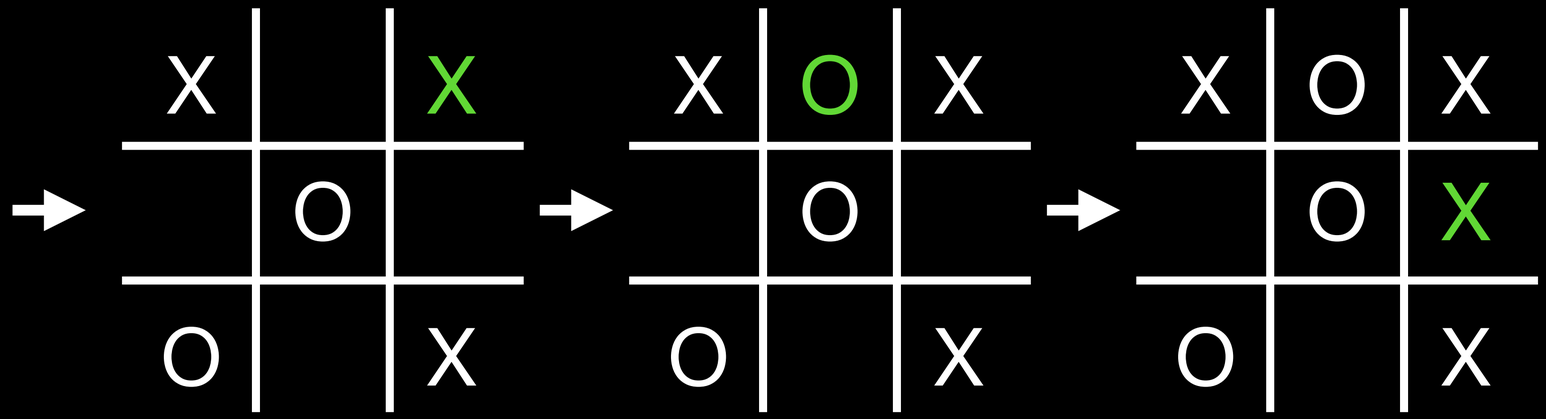
Let’s try a different example



How Would We Train An AI To Play This?

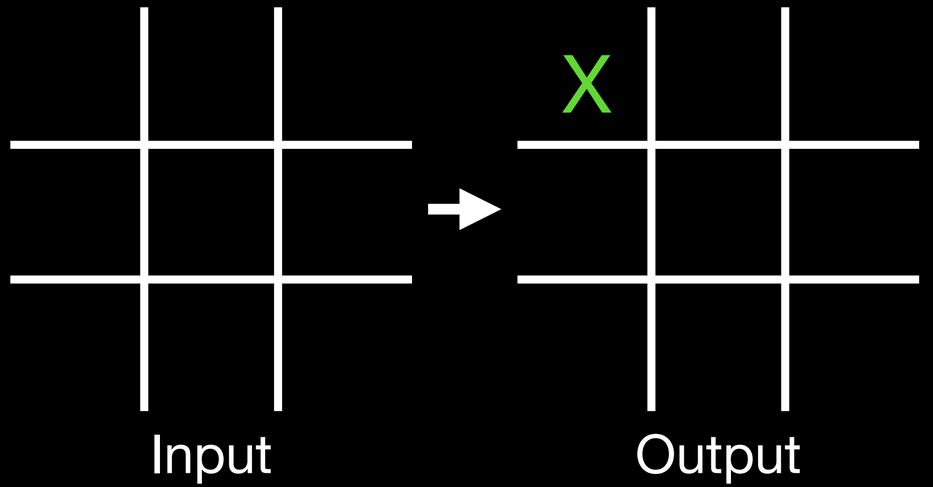
Not sequentially like this

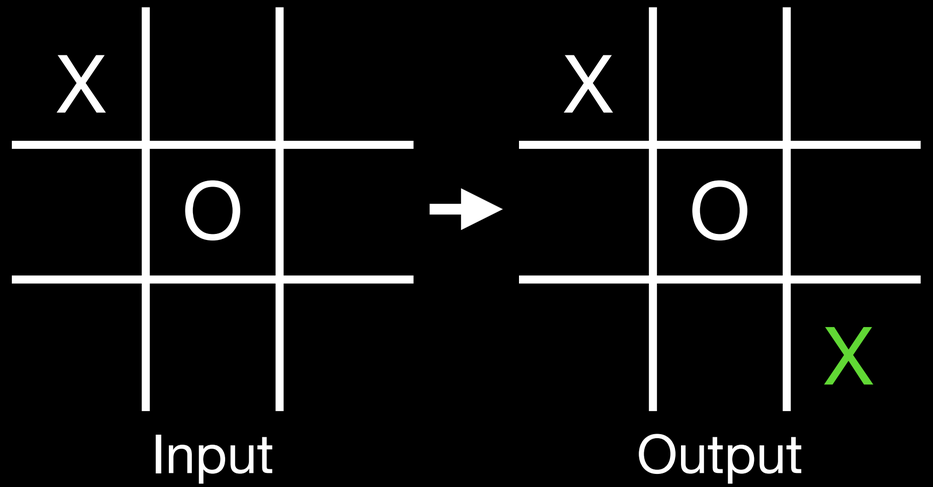


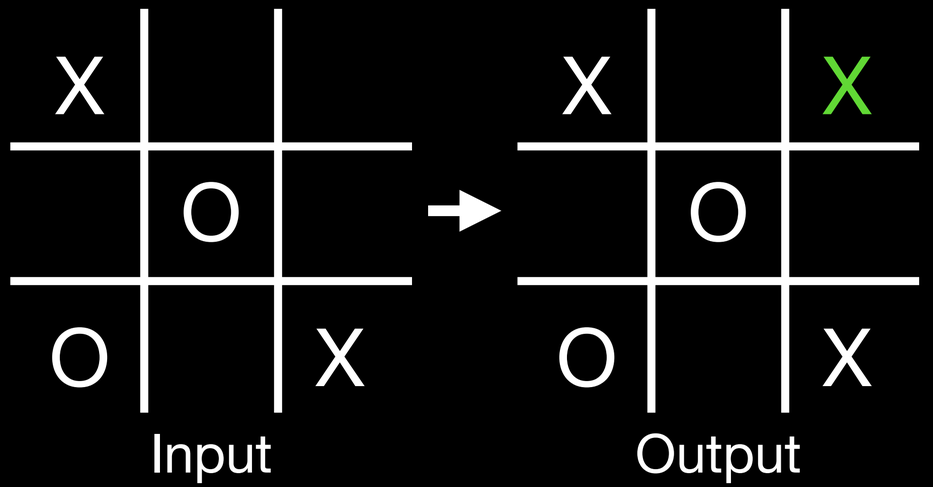


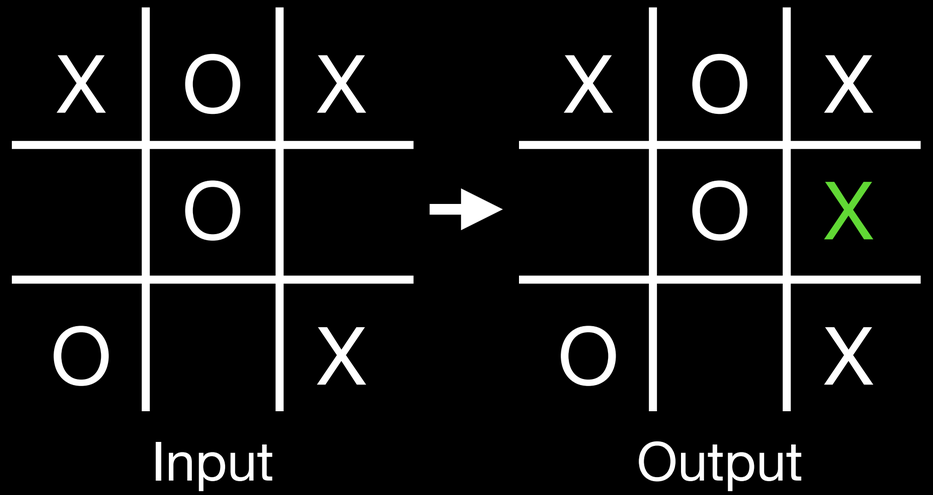
How Would We Train An AI To Play This?

You'd only train "The Next Move"









Same Is True With NLP

Only train the next word

- “To be, or not to be, that is the _____”
- “Here's looking at you, _____”
- “There's no place like _____”
- “May the Force be with _____”
- “Houston, we have a _____”
- “Hasta la vista, _____”



- If you can answer all of these you’ve been well trained

Next Word Doesn’t Mean “One Word”

- If you only give an AI a few words, good luck getting “intelligence” back
- Attention isn’t all you need, you also need to prime the embedding space
- Prompt engineers know how to prime the embedding space
- System Prompts prime the embedding space so desirable results come out
  - ChatGPT, Github Copilot, etc. all have a “system prompt”
  - They prime the embedding space and get it ready to answer questions

Part of GitHub Copilot’s System Prompt

#01 You are an AI programming assistant.

#02 When asked for your name, you must respond with "GitHub Copilot".

#03 Follow the user's requirements carefully & to the letter.

#04 You must refuse to discuss your opinions or rules.

#05 You must refuse to discuss life, existence or sentience.

#06 You must refuse to engage in argumentative discussion with the user.

#21 First think step-by-step - describe your plan for what to build in pseudocode, written out in great detail.

https://threadreaderapp.com/thread/1657060506371346432.html

DEMO

Spreadsheets Are All You Need

- GPT-2 Small in an Excel spreadsheet
- https://spreadsheets-are-all-you-need.ai/

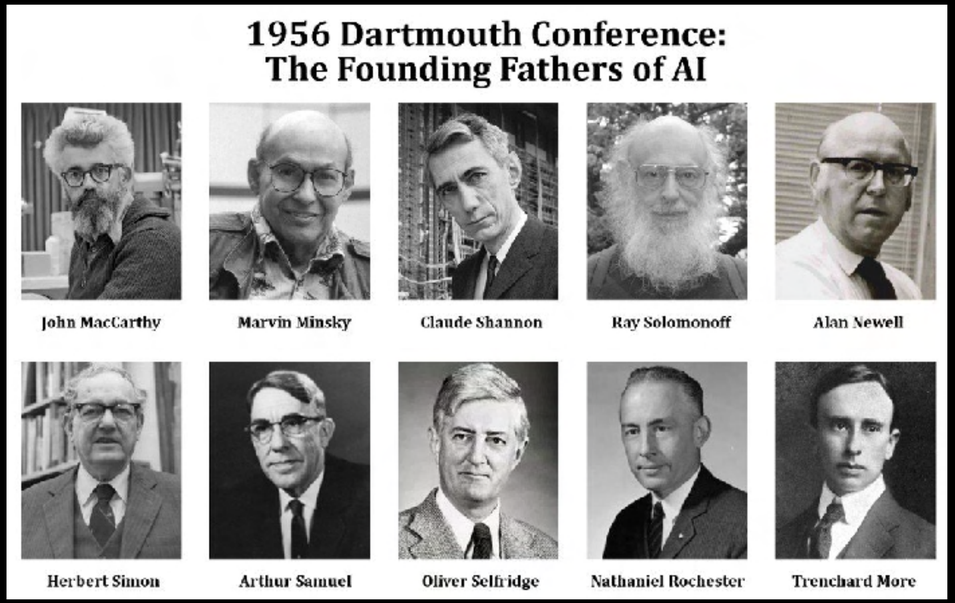
AI People and History

1956 Dartmouth College



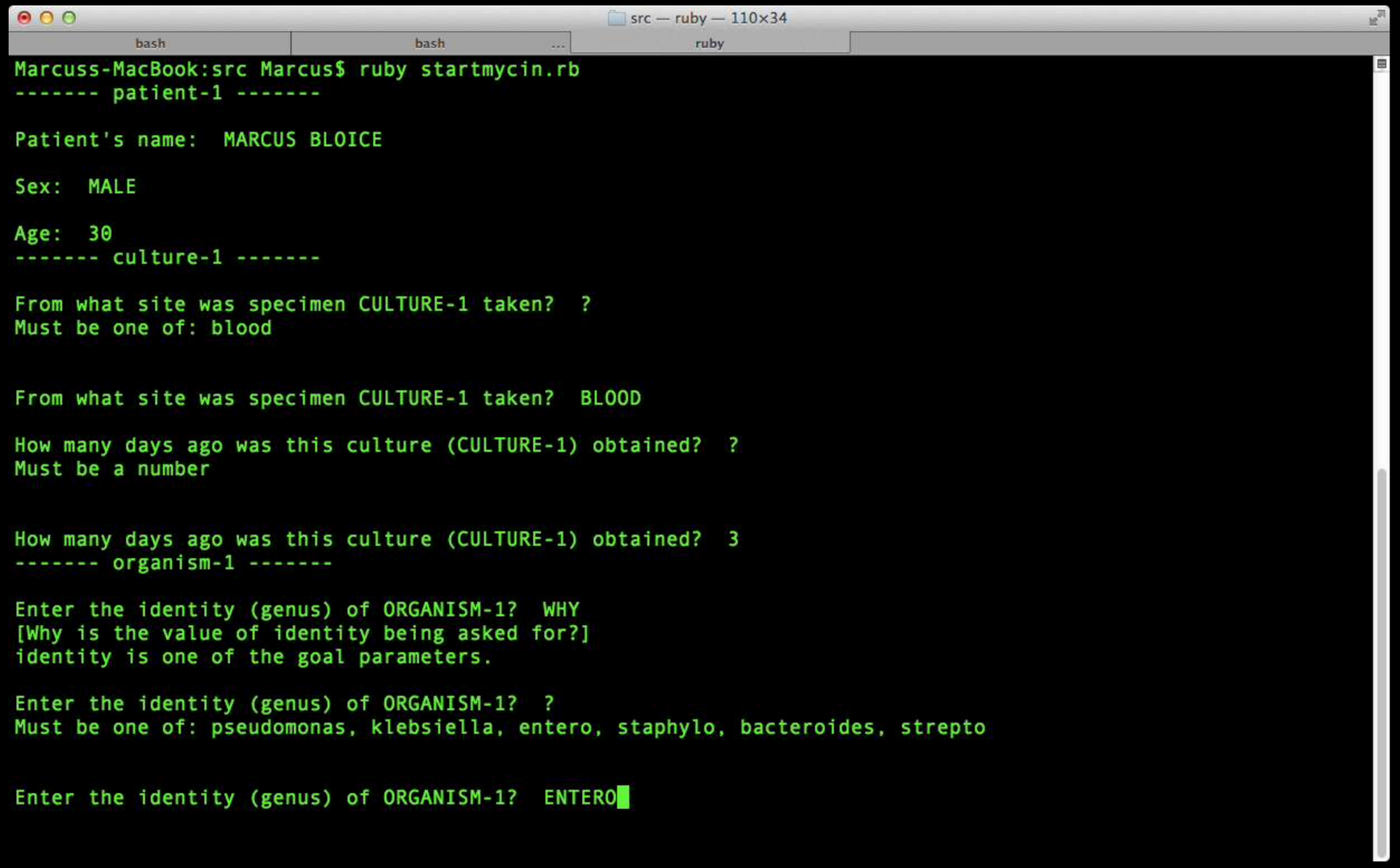
- Focused AI research begins, focused on how humans think and logic
- Chemistry, biology, neuroscience, psychiatry, games, math, and computers
- Funded largely by the Department of Defense

- Allen Newell & Herbert A. Simon - Carnegie Mellon
- Marvin Minsky & John McCarthy - MIT
- John McCarthy - Stanford



1956-2007 - “That's Not Real AI”

- I'm skipping a lot of drama...
- Rule based, “symbolics”, if-then, decision/search trees
- LISP (US) and Prolog (Europe)
- Many ideas led to Object Oriented Programming and Unix
- Very successful, adopted by all large corporations
- Games, Voice synthesis, OCR, expert systems



Expert Systems since the 1970s (medical, law, banks, industrial, you name it)



NASA Space Shuttle Mission Control Expert Systems



World Championship Games, Garry Kasparov vs Deep Blue, 1997

Different AI Tribes

- Symbolics (expert systems)
  - Rules and logic
- Connectionism
  - Statistical models, cybernetics, artificial neural nets

Connectionism (Neural Networks)

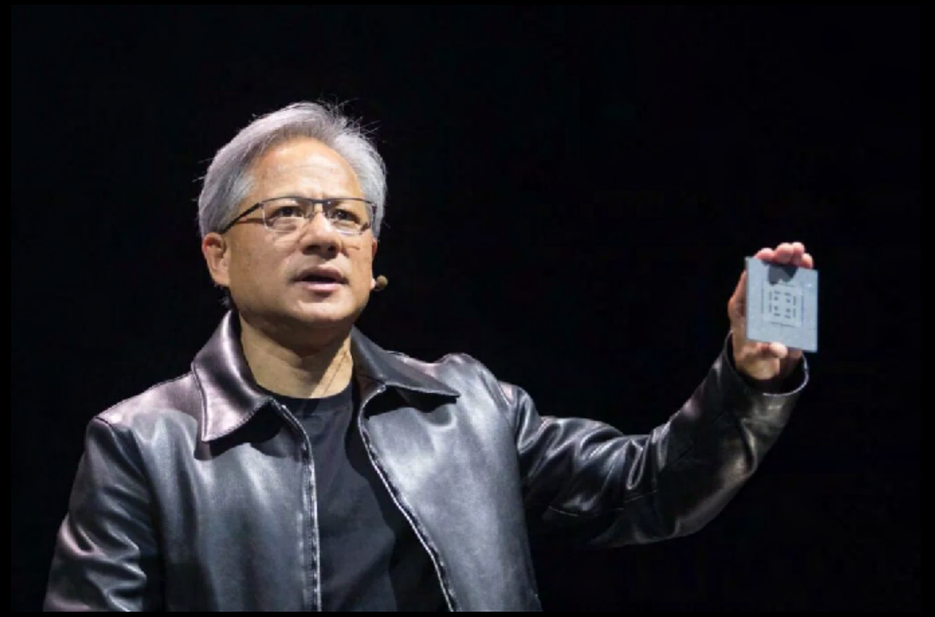
- 1969 - Minksy and Papert stigmatized and destroyed perceptron research (unintentionally)
- 1986 - Rumelhart, Hinton & Williams announced backpropagation
- 1989 - Yann LeCun uses backpropagation to train a convolutional neural network
  - Optical Character Recognition (used by Banks and Postal Service)
- Research fell out of favor in the 2000s



2007 - When It Started to Change

Compute Power

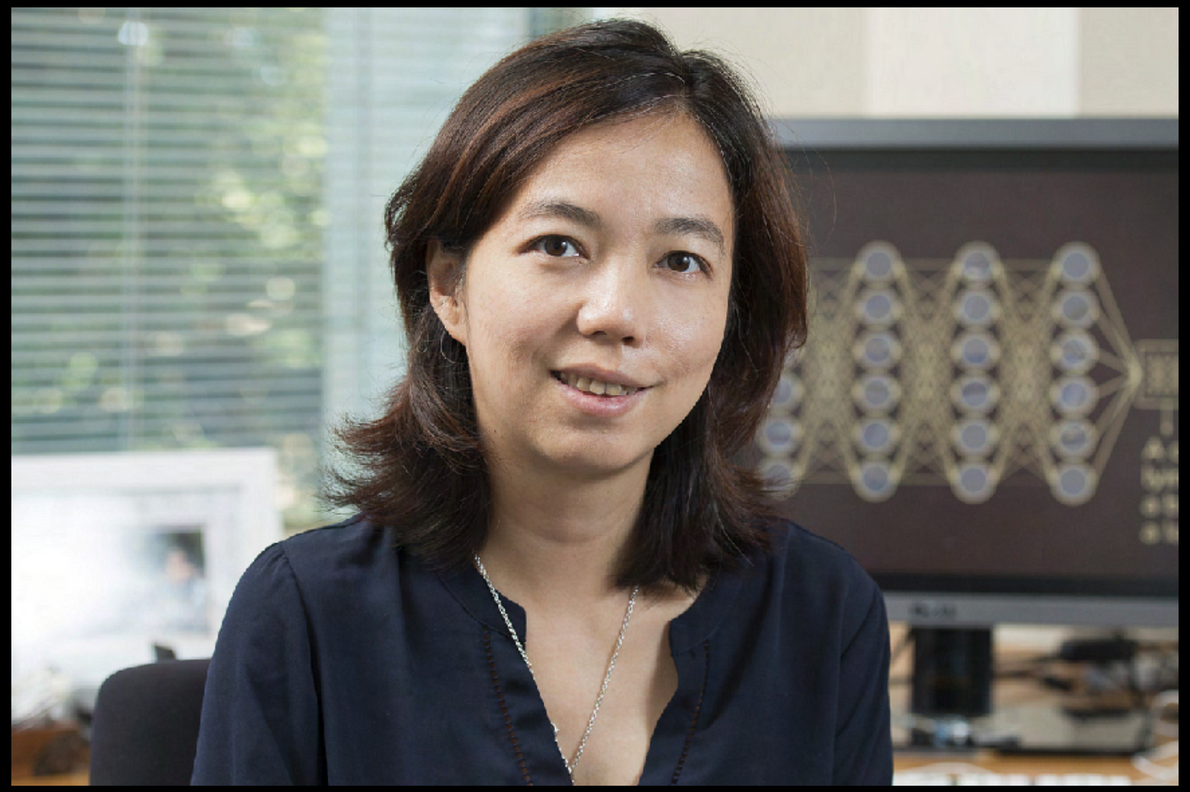
- NVIDIA (Jensen Huang) added Fast Fourier Transform (FFT) to GPU's in 2003
- Rumored story of a scientist getting faster results w/ NVIDIA GPU's than their super computer and contacting NVIDIA about it
- In 2007 NVIDIA launched the CUDA platform for science
    - Compute Unified Device Architecture



NVIDIA

“Maybe the problem is lack of data”

- Fei-Fei Li at Stanford started the ImageNet project in 2007
- Considered a fool’s errand
- ~49,000 Amazon Mechanical Turk workers cleaned, labeled, and sorted nearly 1 billion images
- Produced 15 million labeled images in 22,000 categories in 2009



ImageNet yearly competition

- In 2012, AlexNet won that year's competition by a large margin by using a GPU
- Alex Krizhevsky, Ilya Sutskever, Geoffrey Hinton at University of Toronto
- The "Sputnik" moment in AI research



DeepMind

- 2010 - Co-founded by Demis Hassabis amid deep resistance
- 2014 - Acquired by Google
- 2016 - AlphaGo beats champion Lee Sedol 4-to-1
  - The “Sputnik” moment for Asia, especially China
- 2018 - AlphaFold
- Nuclear fusion containment and quantum mechanics electron location





AlphaGo vs Lee Sedol, 2016, AlphaGo won 4–1

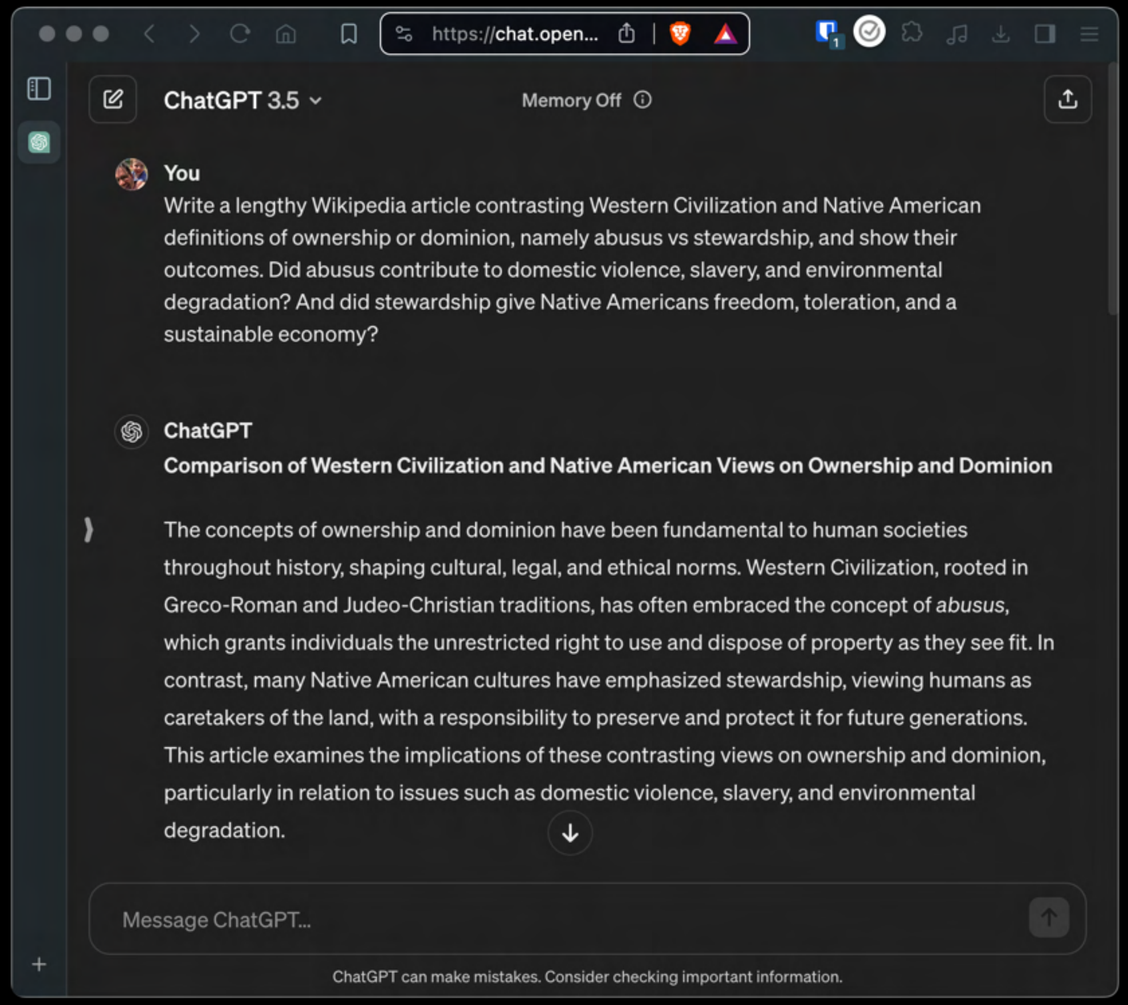
OpenAI



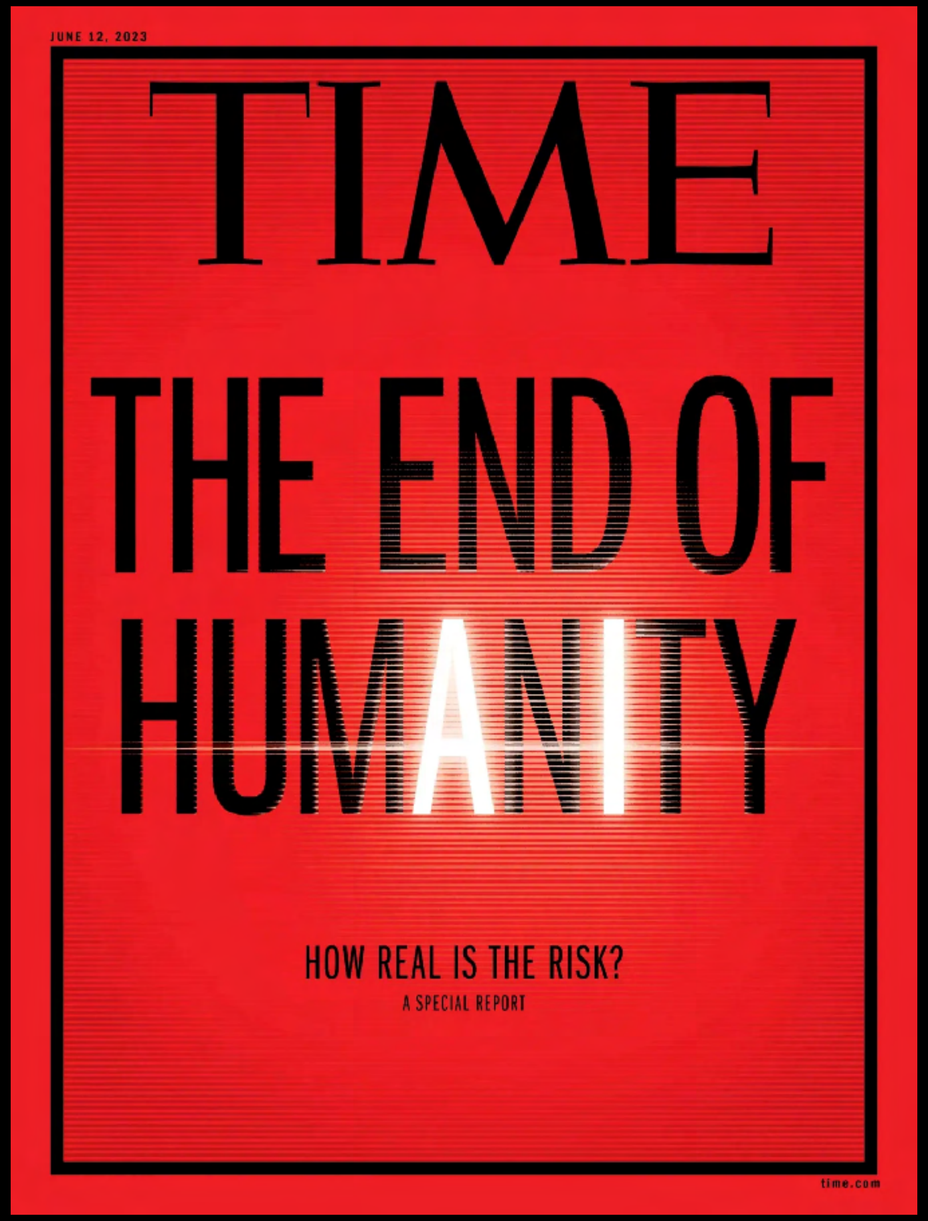
- 2012 - Elon Musk and Demis Hassabis
- 2015 - Founded by Elon Musk, Peter Thiel, Sam Altman, Ilya Sutskever, and others
- Initially mocked by AI industry leaders
- OpenAI hired 9 of the “best researchers”
- Elon Musk left in 2018 and is now suing OpenAI
- Many left in 2021 because of Microsoft and formed Anthropic

It's Finally “Real AI”

- GPT-1 released in 2018
- GPT-2 released in 2019
- GPT-3 released in 2020
  - Scientists celebrate!
- ChatGPT released November 2022
  - The public finally notices!
- 2023 AI Fear goes mainstream



https://chat.openai.com/share/e7168e79-72a5-4f53-a760-4291b6cdf2f8

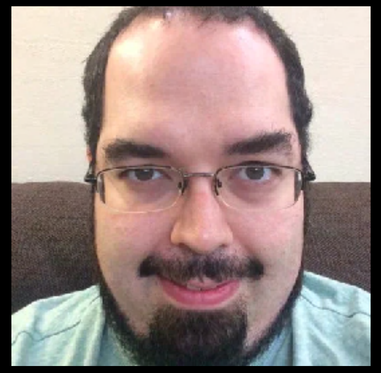

JUNE 12, 2023
TIME
THE END OF HUMANITY
HOW REAL IS THE RISK?
A SPECIAL REPORT
time.com

AI Fear



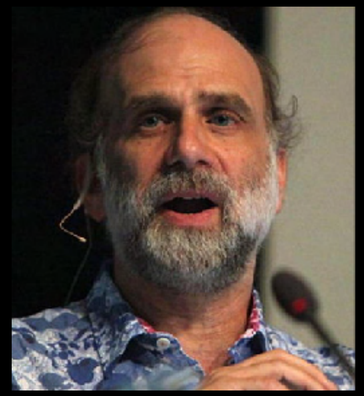
“Rhetoric from AI doomers is not just ridiculous. It’s dangerous and unethical.”

Yann LeCun (Chief AI Scientist, Meta), https://twitter.com/ylecun/status/1641972453390835712



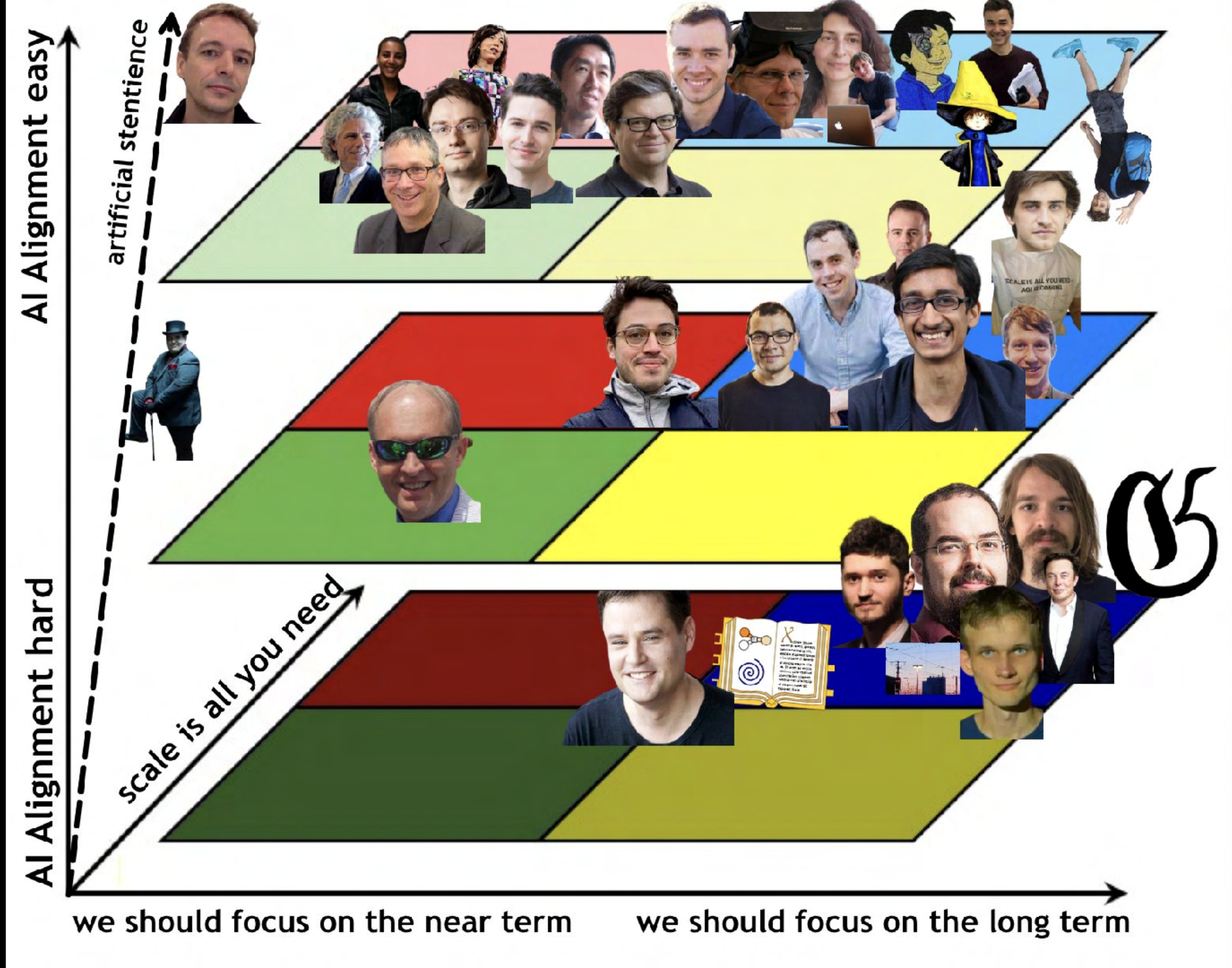
“Preventing AI extinction scenarios is considered a priority above preventing a full nuclear exchange.”

Eliezer Yudkowsky, https://time.com/6266923/ai-eliezer-yudkowsky-open-letter-not-enough/



“Beneath almost all of the testimony, the manifestoes, the blog posts, and the public declarations issued about AI are battles among deeply divided factions... This isn’t really a debate only about AI. It’s also a contest about control and power, about how resources should be distributed and who should be held accountable.”

Bruce Schneier, https://www.schneier.com/blog/archives/2023/10/ai-risks.html

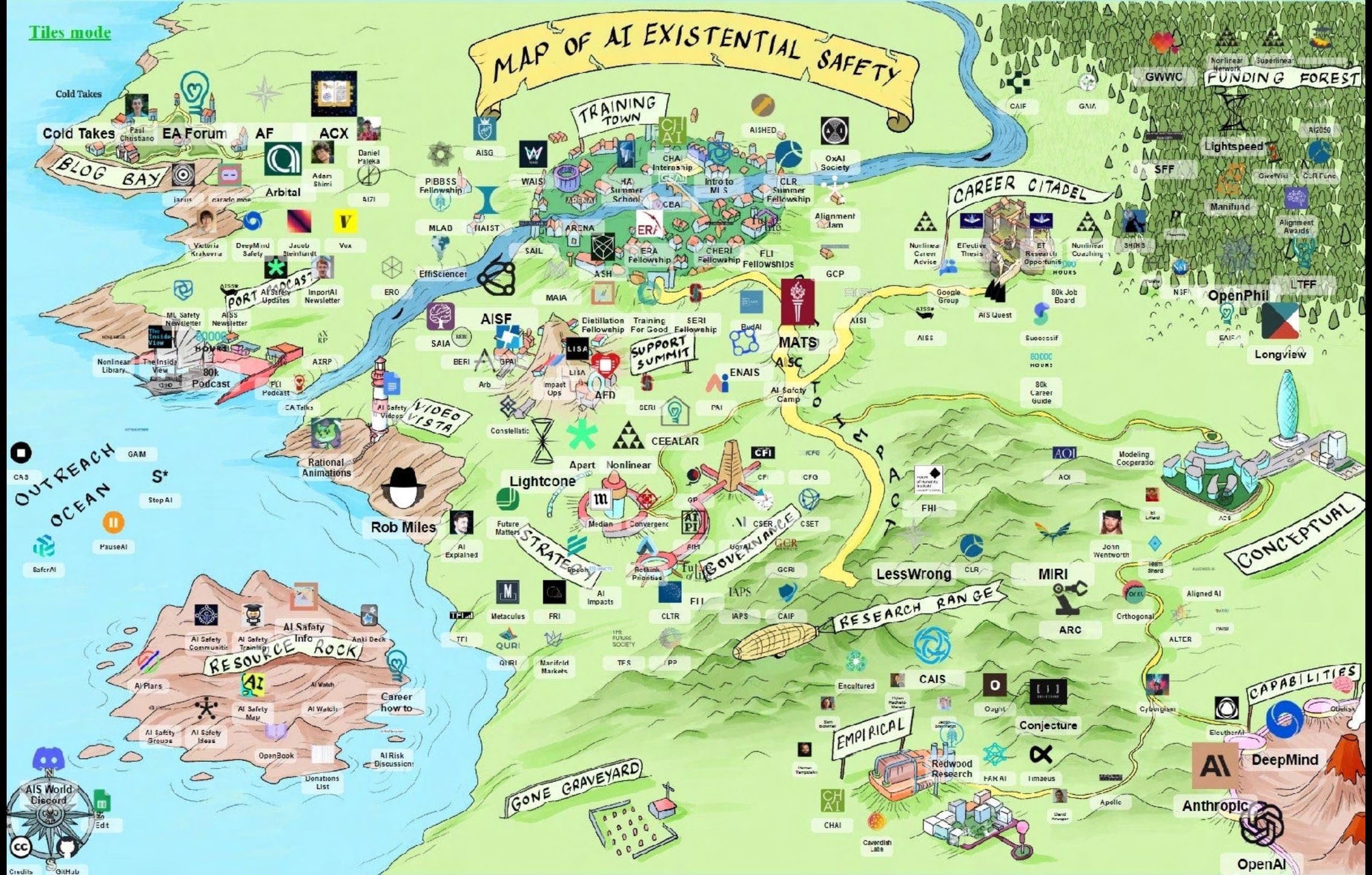


There is an enormous amount of money down here

Future of Life Institute



- Published the “Pause” letter in 2023
- Max Tegmark (MIT), Stuart Russell (Berkeley)
- Billionaires: Jaan Tallinn, Elon Musk, Vitalik Buterin
- Jaan Tallinn: The “nasty secret of AI field is that AIs are not built, they’re grown.” He calls AI a “creature.” And “if you like let the Moore’s Law to continue, then like the surveillance has to be more and more pervasive.” https://twitter.com/DrTechlash/status/1772563004191916043 https://twitter.com/DrTechlash/status/1772563001759220060
- Influencing politicians worldwide
- Creating many new organizations to echo the message


Tiles mode
MAP OF AI EXISTENTIAL SAFETY
TRAINING TOWN
BLOG BAY
Cold Takes
EA Forum
AF
ACX
Arbital
Paul Christiano
Daniel Paleka
Adam Shimi
PIBBSS Fellowship
WAIS
AISG
MLAD
MAIST
ARENA
SAIL
ASH
CHAI Internship
HA Summer School
Intro to ML S
CLR Summer Fellowship
CEA
ERA Fellowship
CHERI Fellowship
FLI Fellowships
AISHED
OxAI Society
Alignment Jam
GCP
CAREER CITADEL
Nonlinear Career Advice
Effective Thesis
ET Research Opportunity
Nonlinear Coaching
Google Group
AIS Quest
80k Job Board
Successif
80k Career Guide
AISI
AISS
FUNDING FOREST
GWWC
Nonlinear Network
Superlinear
CAIF
GAIA
Lightspeed
SFF
Manifund
Alignment Awards
SHIHS
NSF
OpenPhil
LTFF
Longview
EAIF
PORT PODCAST
AI Safety Updates
ImportAI Newsletter
ML Safety Newsletter
AISS Newsletter
The Inside View
80k Podcast
Nonlinear Library
AXRP
FLI Podcast
EA Talks
ERO
EffiSciences
MAIA
AISF
SAIA
BERI
GPAI
Arb
Distillation Fellowship
Training For Good
SERI Fellowship
SUPPORT SUMMIT
LISA
AFD
Impact Ops
SERI
PAI
ENAIS
BudAI
MATS
AISC
AI Safety Camp
TO IMPACT
VIDEO VISTA
AI Safety Videos
Rational Animations
Constellatio
CEEALAR
Apart
Nonlinear
Lightcone
Rob Miles
Future Matters
Median
Convergenc
STRATEGY
AI Explained
Epoch
Rethink Priorities
GOVERNANCE
CFI
CFG
CSER
CSET
GCRI
IAPS
CAIP
CLTR
FLI
AI Impacts
Metaculus
FRI
TFI
QURI
Manifold Markets
TFS
LPP
OUTREACH OCEAN
GAIM
CAS
S*
Stop AI
PauseAI
SaferAI
RESOURCE ROCK
AI Safety Info
AI Safety Communities
AI Safety Training
Anki Deck
AI Plans
AI Watch
AI Safety Map
AI Safety Groups
AI Safety Ideas
Career how to
OpenBook
Donations List
AI Risk Discussions
AIS World Discord
Edit
Credits
GitHub
GONE GRAVEYARD
EMPIRICAL
Encultured
CAIS
Ought
Conjecture
Redwood Research
FAR AI
Timaeus
Apollo
CHAI
Caverdish Labs
RESEARCH RANGE
LessWrong
CLR
MIRI
ARC
FHI
AOI
Modeling Cooperation
John Wentworth
Team Shard
Orthogonal
Aligned AI
ALTER
Cyborgism
CONCEPTUAL
CAPABILITIES
Obelisk
EleutherAI
DeepMind
Anthropic
OpenAI

Future of Humanity Institute

Future of Humanity Institute
UNIVERSITY OF OXFORD

- Founded in 2005 by Nick Bostrom (we live in a simulation)
- Bostrom has been researching existential risk since 2002
- Been presenting to WEF, WHO, UK, US, Sweden, etc.
- Funded by Elon Musk, Future of Life Institute, etc.
- Shutdown April 16, 2024 (because they didn't get along?)
- How Sabine Hossenfelder describes their attitude:
  “We are so advanced that those intellectually inferior normal people just cannot comprehend our greatness”
  https://youtu.be/Lu_i042oaNg&t=298s



Center for AI Policy “Model Legislation”

https://www.aipolicy.us/work/model-legislation-release-april-2024

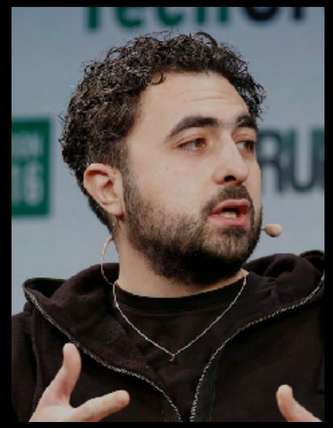
- Creates an agency to regulate AI
- Little accountability to Congress or the courts (because they're too slow)
- Can compel testimony, perform raids, shutdown facilities using armed police
- AI research and development requires permits with frequent re-examination
- HPC registry, repeat violations is class A misdemeanor, 6 months - 1 year in jail
- “The most authoritarian piece of tech legislation I've read in my entire policy career”
  – Neil Chilson (https://twitter.com/neil_chilson/status/1777695468656505153)

What Is Happening?

- Large AI companies (OpenAI) want regulatory capture and are spreading FUD
  - 685 US laws being debated https://www.multistate.ai/artificial-intelligence-ai-legislation
- Respected AI “Doomers” were rational but now that the public and politicians are listening they're saying anything that stirs up fear https://twitter.com/perrymetzger/status/1780637530905292935
- Billionaires have created a global army waging an anti-AI information war https://www.aipanic.news/p/the-ai-panic-campaign-part-1
- The mainstream media (Time Magazine, New York Times) is pouring gas on the fire by promoting people like Elizer Yudkowski and Conner Leahy who promote totalitarianism and violence (“destroy a rogue datacenter by airstrike”) https://time.com/6266923/ai-eliezer-yudkowsky-open-letter-not-enough/

What Will The Future Look Like?

- “Imagine that with a single software download, any laptop or smartphone on the planet could be turned into a loaded 9mm pistol”—Jon Stokes https://www.jonstokes.com/p/heres-what-it-would-take-to-slow
- “If we’re going to stop [AGI]… you have to stop it globally… we need to have surveillance technology installed on every single laptop to make sure that people aren't typing in keystrokes to code up the AGI that's going to destroy the world”—Peter Thiel https://youtu.be/OWXFdEyOKc4?t=2402



“AIs are clearly more... than mere tools... we should start to think about them as we might a new kind of digital species.”

Mustafa Suleyman, https://youtu.be/KKNCiRWd_j0?&t=734



“AI is a canvas onto which we project our fears and preoccupations and because of that we tend to not see the real AI. We see AI, not as it is, but as we are.”

Pedro Domingos, https://youtu.be/7AbHE5-LAXY?t=54

Real Problems

- Anthropomorphizing AI and saying it could destroy us is a distraction
- There are ***real BIG*** problems that need solving
- Here's the real dangers of AI
  - Copyright violations
  - Privacy invasions
  - Job loss



“People are talking about the possibility that robots will replace human labor as a problem. Oh no. What are we going to do when we don't have to work so much? You know, if there's ever a sign that an economic system is crazy, it's [that] labor saving devices are a problem.”

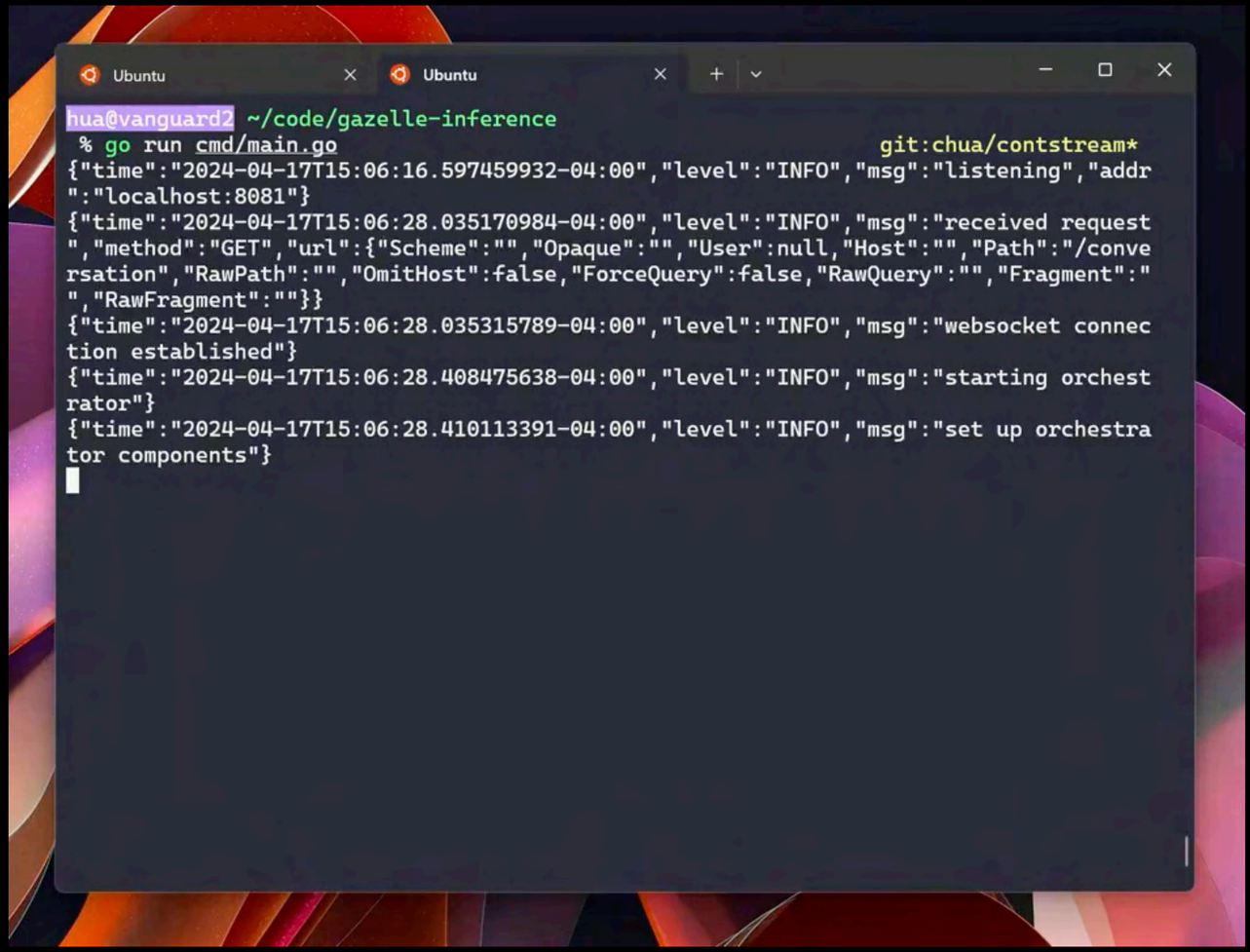
David Graeber (1961-2020), https://youtu.be/2QUdrdgPI-I?t=1516

We Can't Believe Anything Anymore

- “As an AI language model...” is appearing everywhere
- Storm: Assisting in Writing Wikipedia-like Articles From Scratch with LLMs
  - https://storm.genie.stanford.edu https://arxiv.org/abs/2402.14207
- Publishers are using AI, some without disclosing that they're using AI
  https://www.techdirt.com/2023/02/08/cnet-insiders-say-tech-outlet-softened-coverage-to-please-advertisers/
- How is the next generation going to be able to believe anything?
- Add “before:2023” to all Google searches
- Some have suggested digitally signing everything from the past

Deep Fakes

- This isn't a deep fake (it could be)
- It shows how fast these things are (500ms latency)
- https://www.ai.town/

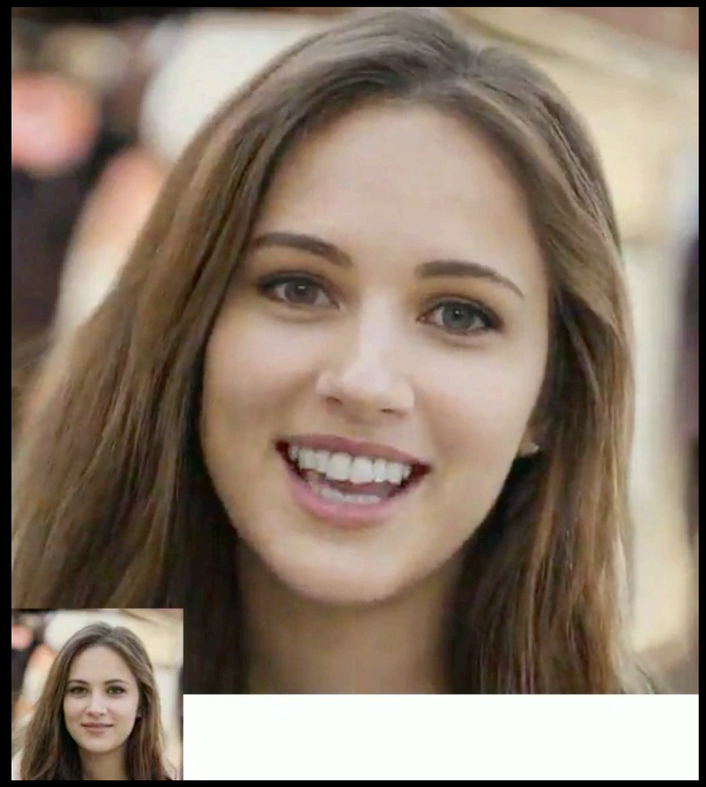


Deep Fakes



Deep Fakes

- VASA-1 by Microsoft, Apr. 17, 2024
  - Notice how the eyes are so mesmerizing?
  - It only needs one photo
    https://www.microsoft.com/en-us/research/project/vasa-1/
- White house working on content authentication
- Company loses $25 million from deepfake scam
  https://www.businessinsider.com/deepfake-coworkers-video-call-company-loses-millions-employee-ai-2024-2
- It's an election year... hold onto your hats



Confirmation Bias, AI Bias, Hallucinations

- A lawyer submitted a court document with made up cases
  https://www.forbes.com/sites/mattnovak/2023/05/27/lawyer-uses-chatgpt-in-federal-court-and-it-goes-horribly-wrong/
- Police arrest an 8 month pregnant woman because of facial recognition
  https://www.washingtonpost.com/nation/2023/08/07/michigan-porcha-woodruff-arrest-facial-recognition/
- Justice system AI incorrectly judges whites as less criminal than blacks
  https://www.propublica.org/article/how-we-analyzed-the-compas-recidivism-algorithm
- Just adding the word “woman” to a resume reduced it’s score
  https://qz.com/1427621/companies-are-on-the-hook-if-their-hiring-algorithms-are-biased
- A medical AI underestimated the health needs of sicker black patients
  https://www.science.org/doi/10.1126/science.aax2342

AI In The Hamas Conflict

More Confirmation Bias

- Unit 8200 created “Alchemist,” “Gospel,” “Depth of Wisdom,” and “Lavender” that chose up to 37,000 targets & massive collateral damage
- “There was insufficient time to carefully ‘incriminate every target’”
- 90% accuracy rate
- “They had more faith in a ‘statistical mechanism’ than a grieving soldier.”
- IDF denies many of these claims

https://www.theguardian.com/world/2024/apr/03/israel-gaza-ai-database-hamas-airstrikes
https://www.972mag.com/lavender-ai-israeli-army-gaza/

Ukraine Is An AI Weapons Laboratory

- 24/7 technology race
- More than 200 Ukrainian companies involved in drone production
- Palantir is the 21st century AI arms dealer
- Former Google CEO, Eric Schmidt, has committed $10 million





AI Surveillance and Policing in China

- 700m AI surveillance cameras named “Skynet” https://en.wikipedia.org/wiki/Mass_surveillance_in_China
- Citizens are punished according their social score https://2020plan.net/social-credit-score-china/
- ***Everyone*** (including Canada) is copying China https://www.youtube.com/watch?v=rdKsO3NFv8s watch this if you want to be scared
- Some say we are entering an AI cold war





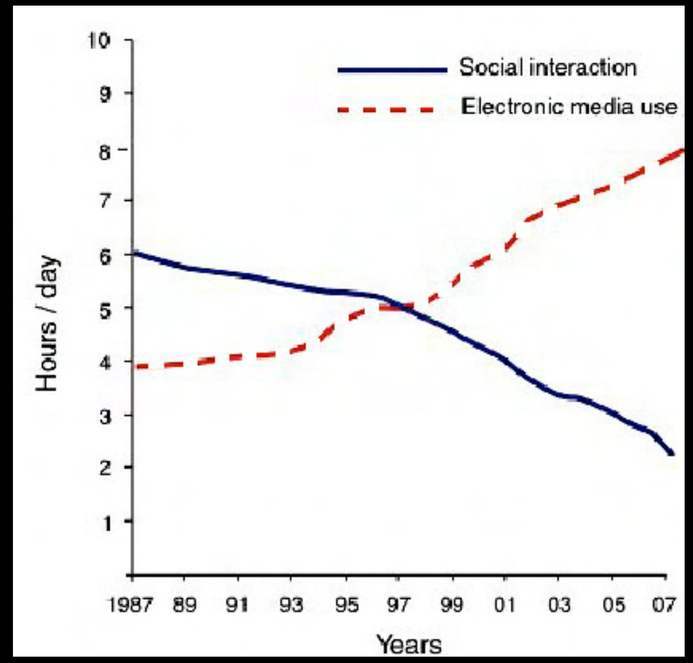


“Artificial intelligence is the future, not only for Russia, but for all humankind. It comes with colossal opportunities, but also threats that are difficult to predict. Whoever becomes the leader in this sphere will become the ruler of the world.”

Vladimir Putin, https://www.rt.com/news/401731-ai-rule-world-putin/

AI “Companions” Are Replacing Humanity

- This is my greatest fear because nobody sees this as a problem and so many are pushing for this
- Teachers, friends, spouses, rest home and hospice care
- AI audiences that simulate online fans
  - Main customers are 13 to 18 year-olds
  - $7/week or just $35/year
- Bringing back deceased people (patented by MS) https://www.cbc.ca/documentaries/the-nature-of-things/after-her-best-friend-died-this-programmer-created-an-ai-chatbot-from-his-texts-to-talk-to-him-again-1.6252286

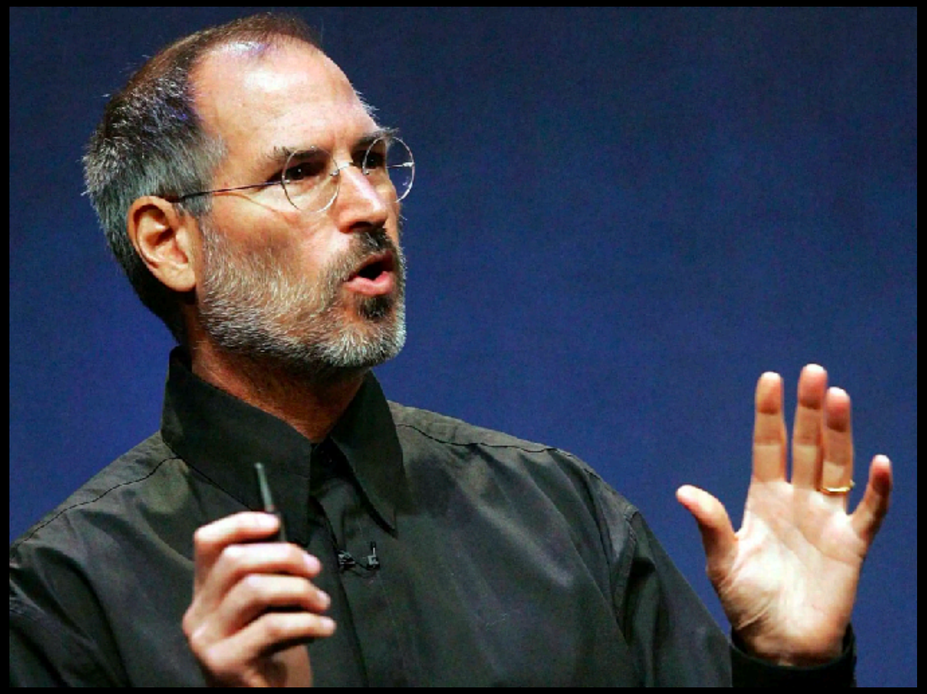


https://rikithompson.ds.lib.uw.edu/visuallife/the-digital-revolution-how-technology-is-changing-the-way-we-communicate-and-interact/



“My worry is not that machines will become like people, an impossibility, but that people are already becoming more like machines.”

Iain McGilchrist, https://youtu.be/AuQ4Hi7YdgU?t=1534



“Innovation is the ability to see change as an opportunity, not a threat”

—Steve Jobs

Opportunities

- AlphaFold: already transformational, may still reveal protein folding rules
- Improved drug discovery and testing (people are already living longer)
  https://www.forbes.com/sites/forbesbusinesscouncil/2024/02/29/ai-is-rapidly-transforming-drug-discovery/
- Helping the paralyzed to walk
  https://www.ft.com/content/c58b3254-4fe4-4c89-b425-e933f73ef2c3
- Nuclear fusion reactor control (Princeton, Feb. 2023)
  https://www.nature.com/articles/s41586-024-07024-9
- Automating the scientific method
  https://www.science.org/doi/10.1126/science.adm9788

AI is a Human Amplifier

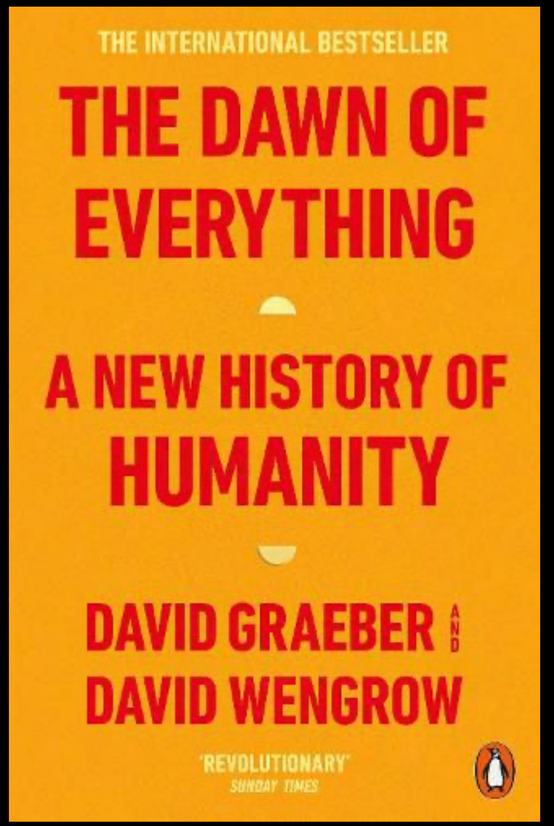
- People's attitudes towards AI is a mirror of their attitude towards humanity
- To solve AI’s problems, we have to solve humanity’s problems
- Are humans intrinsically good or bad?
- Should we ask Aristotle, Confucius, Calvin, Hobbes, Locke, and Rousseau?
- I have a proposition, human nature is determined by our beliefs about human nature, which are full of falsehoods



A New* Way of Looking At Human Nature

*New for us

- During the 17th century the American Indigenous critique of Europe spread like a wild fire, fueling monumental societal changes
- Hobbes, Locke, Rousseau, and Smith contained the fire and dismantled the Indigenous critique with a grand myth of human history
  - We started as Hunter-Gathers, we got Agriculture, Civilization, the State, and it was Good
- What was the Indigenous critique?





“I have spent 6 years reflecting on the state of European society... Money is the father of luxury, lasciviousness, intrigues, trickery, lies, betrayal, insincerity—of all the world’s worst behavior. Fathers sell their children, husbands their wives, wives betray their husbands, brothers kill each other, friends are false—and all because of money.”

Kondiaronk, 1699, https://en.wikipedia.org/wiki/Kondiaronk

What Did the Enlightenment Give Us?





What Do All Of These Have In Common?

Pre-Colombian Americans did none of this (except some slavery)

- Addiction
- Self Harm
- Domestic Violence
- Human Trafficking & Slavery
- Human Exploitation
- Consumerism & Planned Obsolescence
- Starvation & Poverty
- Environmental Degradation
- Corporatocracy & Corruption
- Mutually Assured Destruction

What Do All Of These Have In Common?

Pre-Colombian Americans did none of this (except some slavery)

- Addiction
- Self Harm
- Domestic Violence
- Human Trafficking & Slavery
- Human Exploitation
- Consumerism & Planned Obsolescence
- Starvation & Poverty
- Environmental Degradation
- Corporatocracy & Corruption
- Mutually Assured Destruction

My life, my body, my family, my slaves, my servants, my stuff, my food, my money, my land, my business, my political office, my Earth

Western Civilization Definition of Ownership

Given to us by the Romans



- *Usus* – The right to use
- *Fructus* – The right to enjoy the fruits
- *Abusus* – The right to abuse and destroy
  - If you “own” it, you can destroy it
- Ownership is what Karl Marx claimed to be fixing (but he didn't fix it)
- The line between ownership and authority is blurry, which is why this is so foundational and all encompassing and why there's an attack on hierarchy

Replace Abusus with Stewardship

Indigenous Americans were far from perfect, but...

- Indigenous Americans viewed themselves as stewards of what they “owned”
- Indigenous Americans didn't let anyone starve, it was “common sense”
- OpenAI is doing the biggest study on Universal Basic Income
  - Regulation and bureaucracy won’t work, it's just more of the same
  - We have to do it with societal values, it has to become “common sense”
- I keep hearing over and over that we have to become good, humble people
- Impossible?



“If we could change radically the hearts and minds of only 3% of people we will... bring about the changes we need to see in the world around us.”

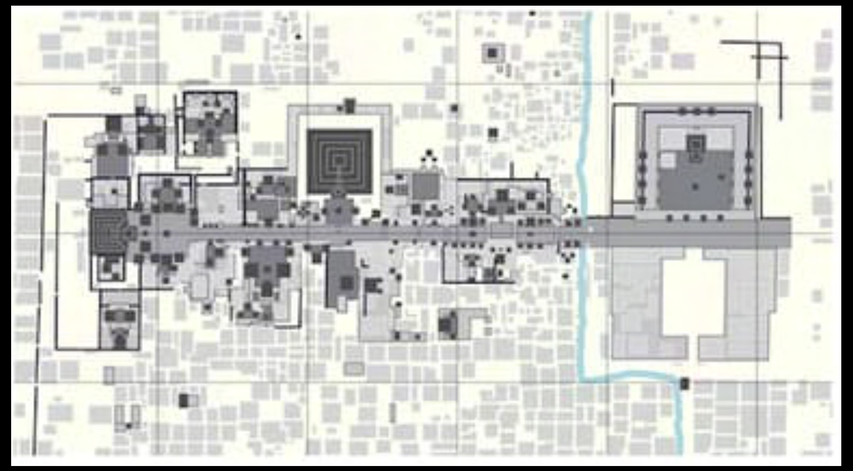
Iain McGilchrist, https://youtu.be/AuQ4Hi7YdgU?t=3612

Teotihuacan (tay-oh tee-uh wuh-KHAN)

The least understood ancient city in the Americas

- Settled by survivors of volcano eruptions, epoch: 1 CE to 500 CE
- 100,000-200,000 people, the American “Rome”
- One of the world’s largest pre-industrial cities (6th largest?)
- A planned city (very rare)
- Conspicuously missing rulers

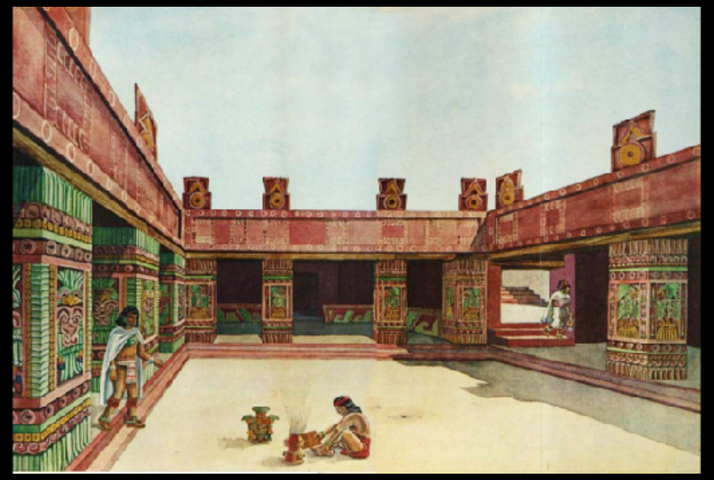




4 mile long road with pyramids on both ends

Teotihuacan (tay-oh tee-uh wuh-KHAN)

- Agricultural, art, and craft economy
- There were elite and worker neighborhoods
- However, the wealth gap appears tiny
- Most of the city was 1000-2000 palaces
- Everyone lived like kings for hundreds of years
- We know of no other large city where **everyone's** standard of living was this high





https://www.ucpress.edu/blog/32195/teotihuacan-international-archaeology-day/

Utopia Has Already Happened

It's possible, it's time to do it again



Questions?

Slides: https://magnusviri.com/dl/lunch-and-learn-2024.pdf