

Unveiling the Magic Behind GenAI: From Hallucination to Innovation

Michael Wu, PhD (@mich8elwu)
chief AI strategist @ PROS

2023.12.05



1

Michael Wu, PhD (@mich8elwu)
chief AI strategist @ PROS

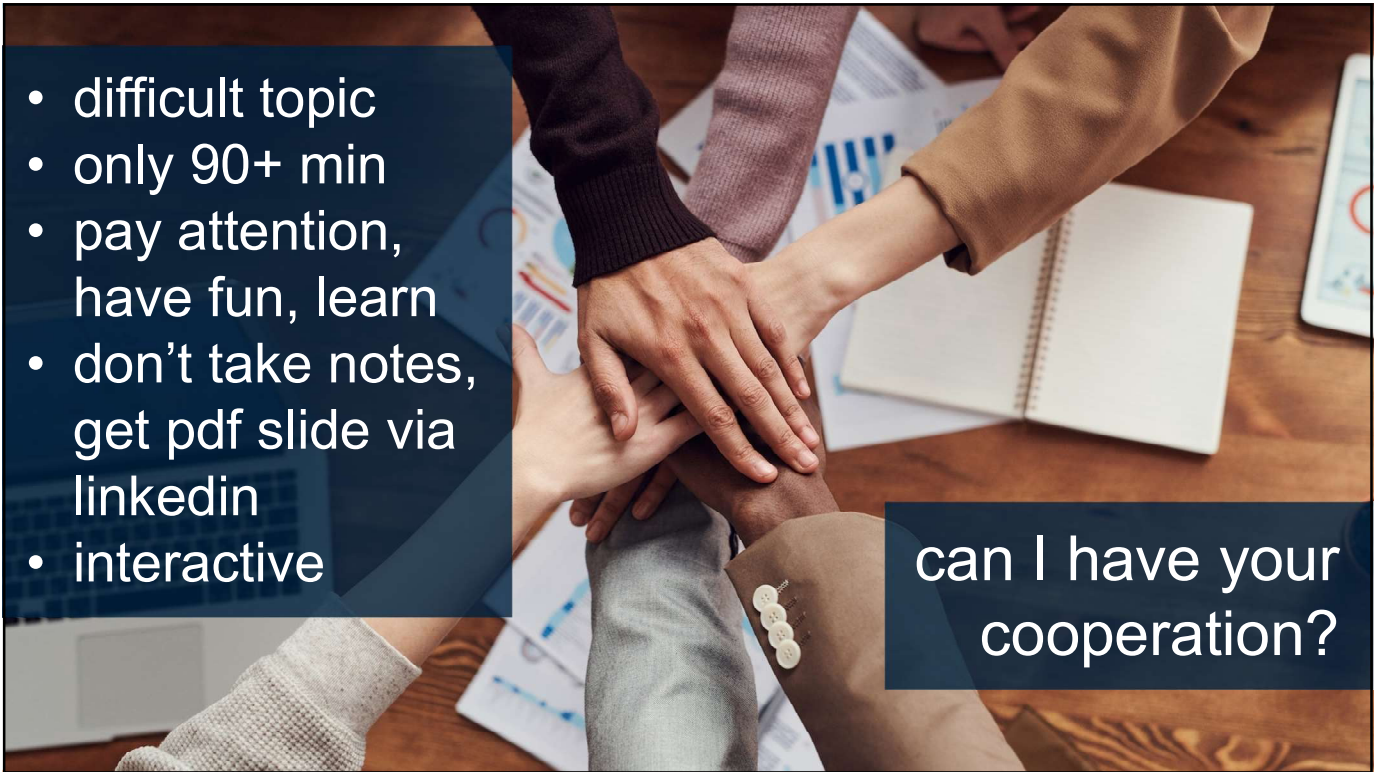
2023.12.05



©2023 PROS, Inc. All rights reserved. Confidential and Proprietary.

page 4

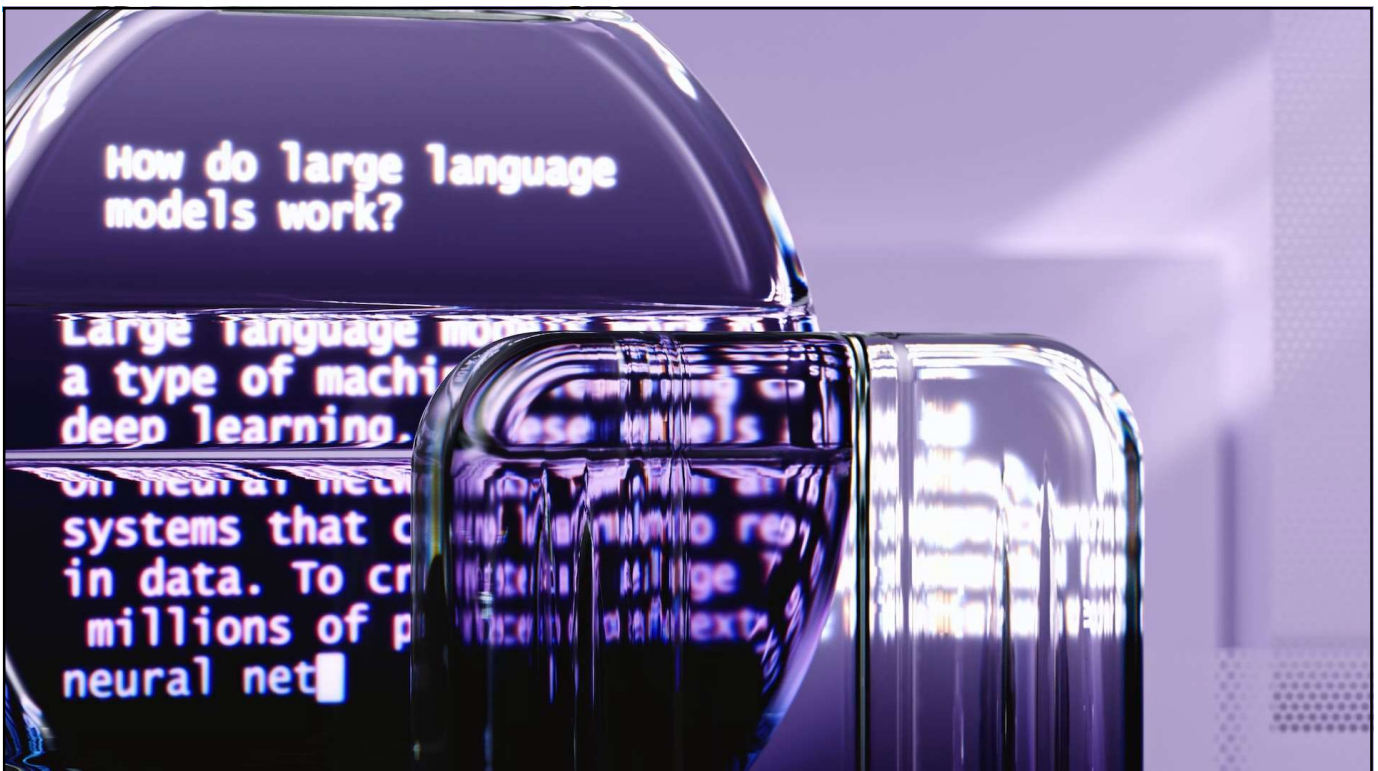
4



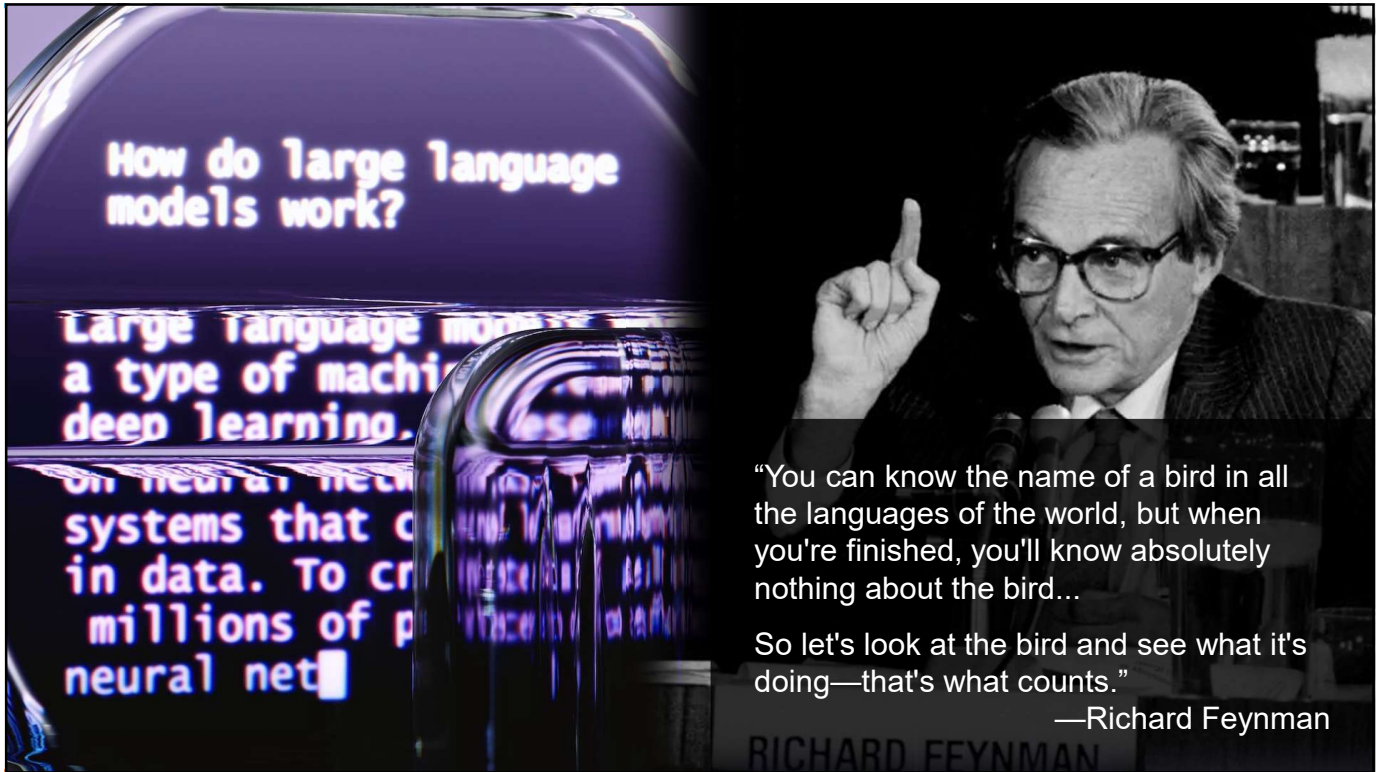
- difficult topic
- only 90+ min
- pay attention, have fun, learn
- don't take notes, get pdf slide via linkedin
- interactive

can I have your cooperation?

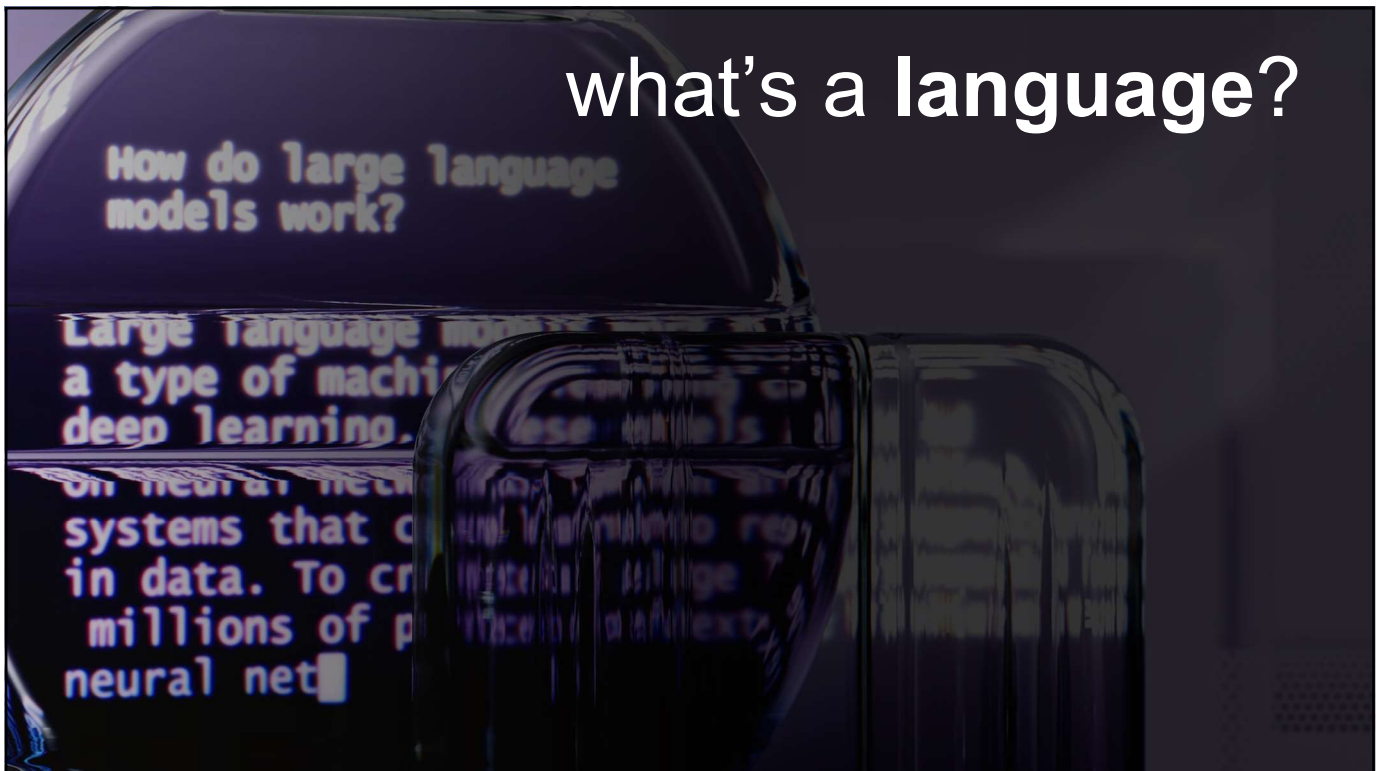
5



6



7



8

How do large language models work?

Large language models are a type of machine learning model based on neural networks. They are systems that learn from large amounts of data. To create these models, researchers use millions of parameters in a neural network.

language:
 vocabulary + grammar
 words + syntaxes
 sequence of word tokens

language model:
 probabilistic model that captures the sequential structure of word tokens

9

How do large language models work?

Large language models are a type of machine learning model based on neural networks. They are systems that learn from large amounts of data. To create these models, researchers use millions of parameters in a neural network.

language model:
 probabilistic model that captures the sequential structure of word tokens

given any sequence s ,
 can we predict the next word w accurately?

$$P(w|s)$$

10

language model:

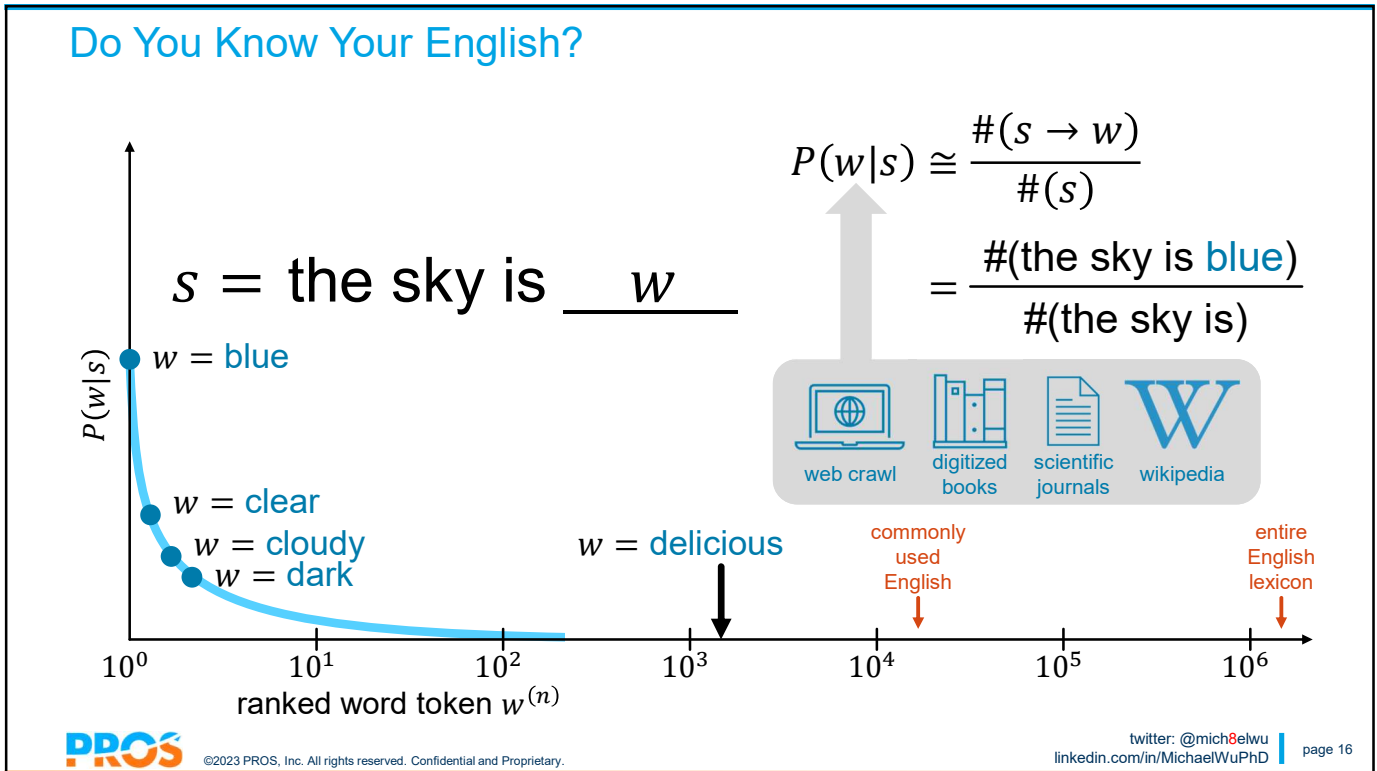
given *any* sequence s ,
can we predict the next
word w accurately?

$$P(w|s)$$

$s = \text{the water is filled with ribulose-bisphosphate-carboxylase-oxygenase, it's very } \underline{\hspace{2cm}}$

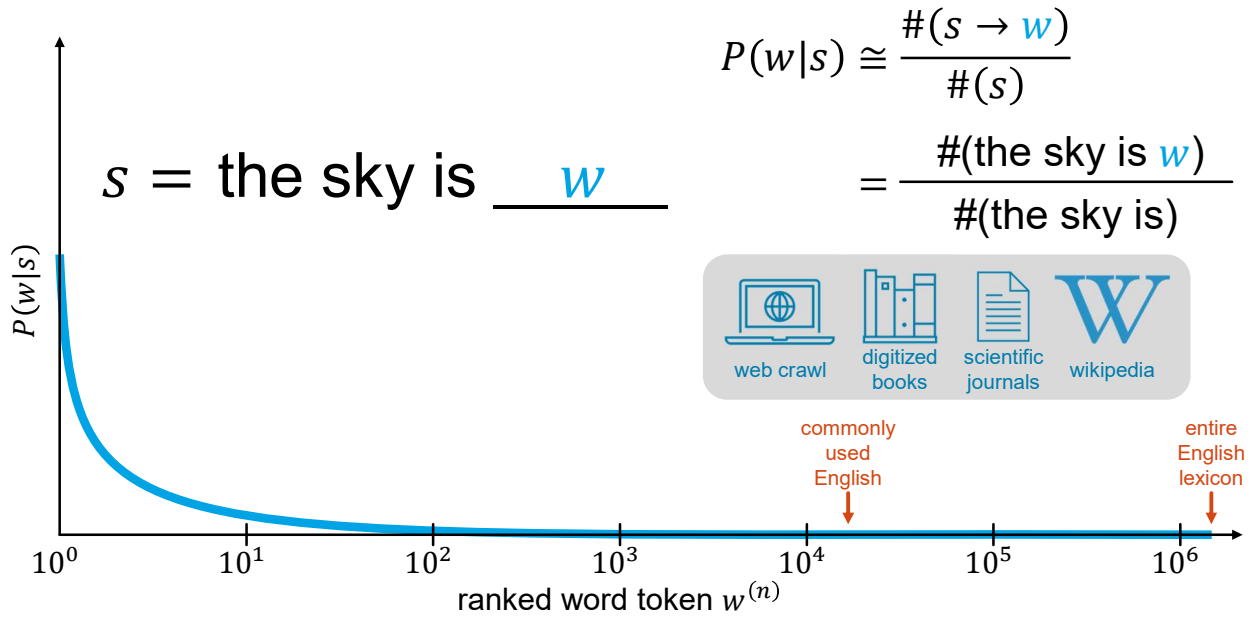
How do large language models work?
Large language models are a type of machine learning model based on deep learning. They use neural networks to process and generate text.

11



16

Do You Know Your English?

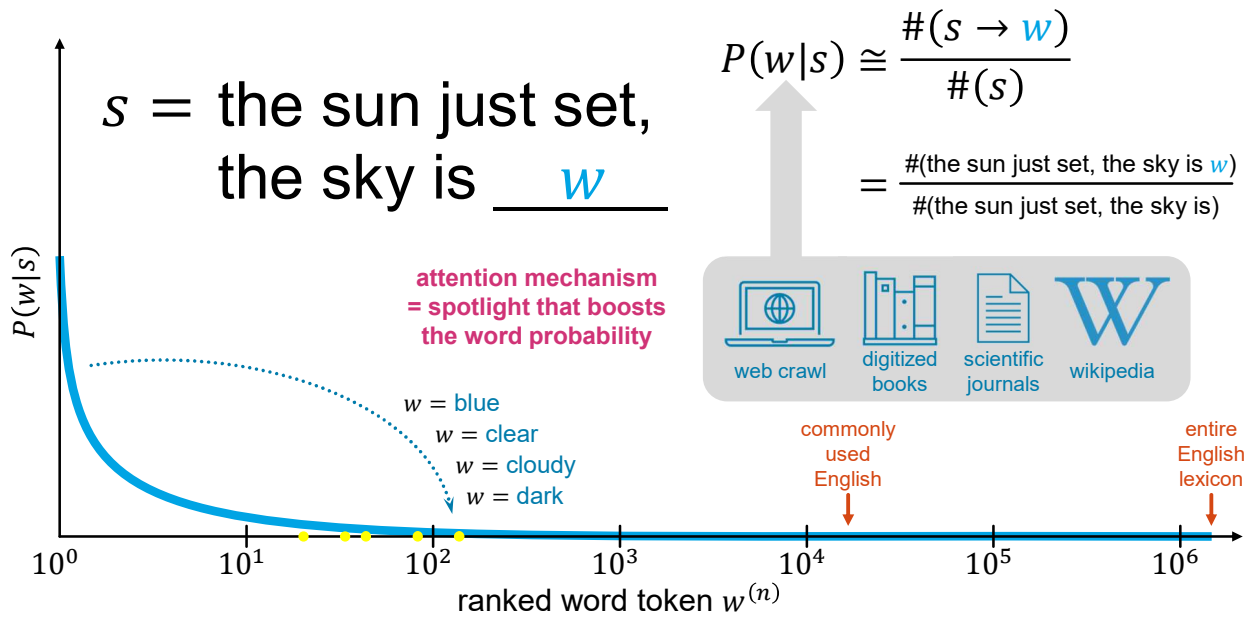


©2023 PROS, Inc. All rights reserved. Confidential and Proprietary.

twitter: @mich8elwu
linkedin.com/in/MichaelWuPhD | page 17

17

Do You Know Your English?

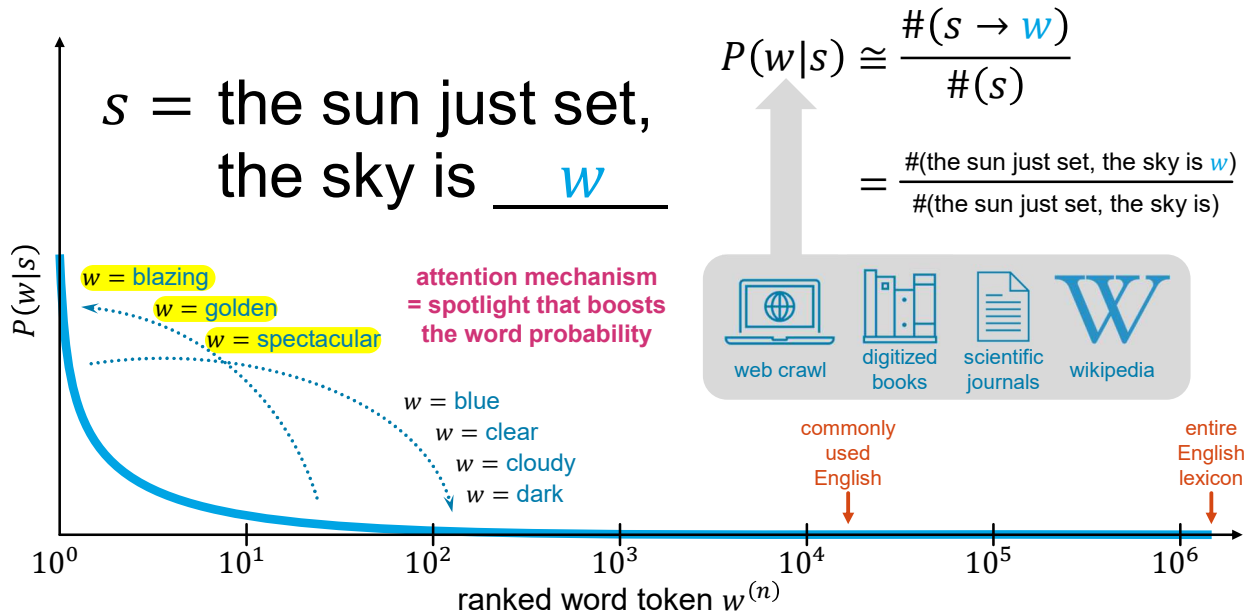


©2023 PROS, Inc. All rights reserved. Confidential and Proprietary.

twitter: @mich8elwu
linkedin.com/in/MichaelWuPhD | page 18

18

Do You Know Your English?

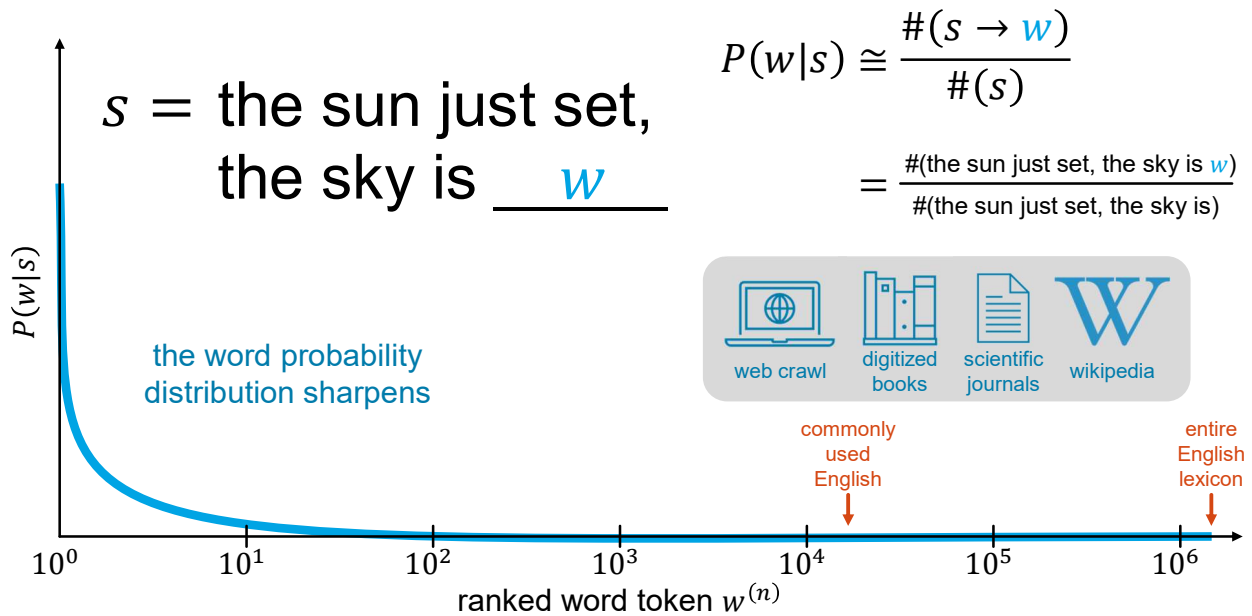


©2023 PROS, Inc. All rights reserved. Confidential and Proprietary.

twitter: @mich8elwu
linkedin.com/in/MichaelWuPhD | page 19

19

Longer Sequence → More Concentrated Word Probability



©2023 PROS, Inc. All rights reserved. Confidential and Proprietary.

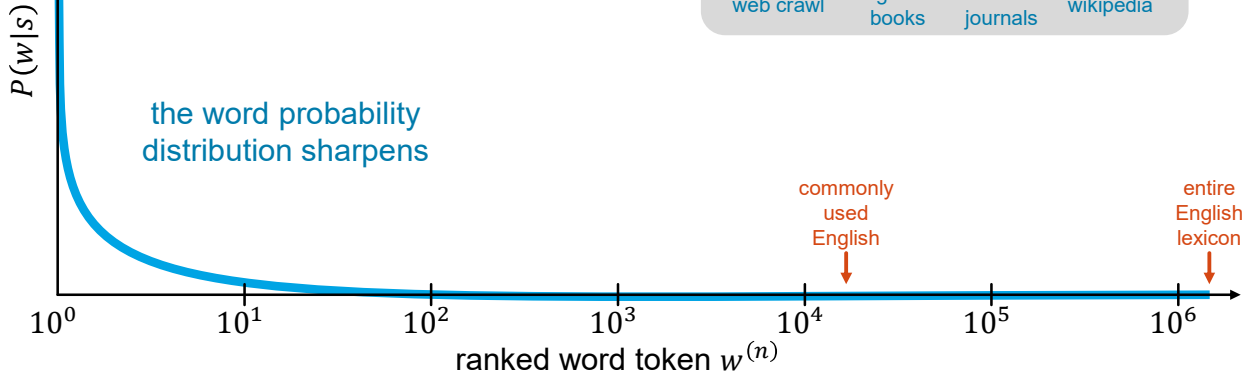
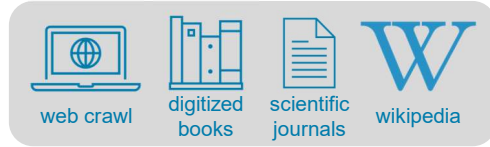
twitter: @mich8elwu
linkedin.com/in/MichaelWuPhD | page 20

20

Longer Sequence → More Concentrated Word Probability

$s =$ it's raining hard,
the sun just set,
the sky is w

$$P(w|s) \cong \frac{\#(s \rightarrow w)}{\#(s)}$$



©2023 PROS, Inc. All rights reserved. Confidential and Proprietary.

twitter: @mich8elwu
linkedin.com/in/MichaelWuPhD | page 21

21

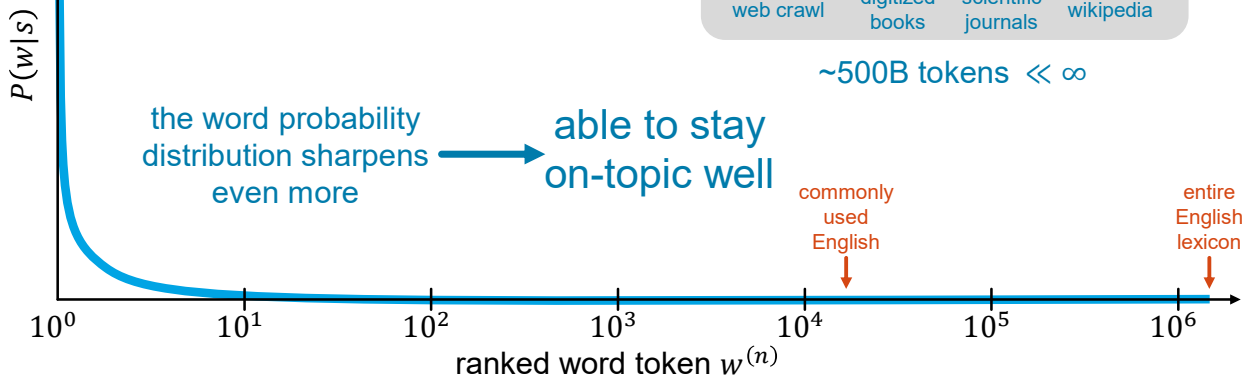
Longer Sequence → More Concentrated Word Probability

$s =$ it's raining hard,
the sun just set,
the sky is w

$$P(w|s) \cong \frac{\#(s \rightarrow w)}{\#(s)}$$



~500B tokens $\ll \infty$



©2023 PROS, Inc. All rights reserved. Confidential and Proprietary.

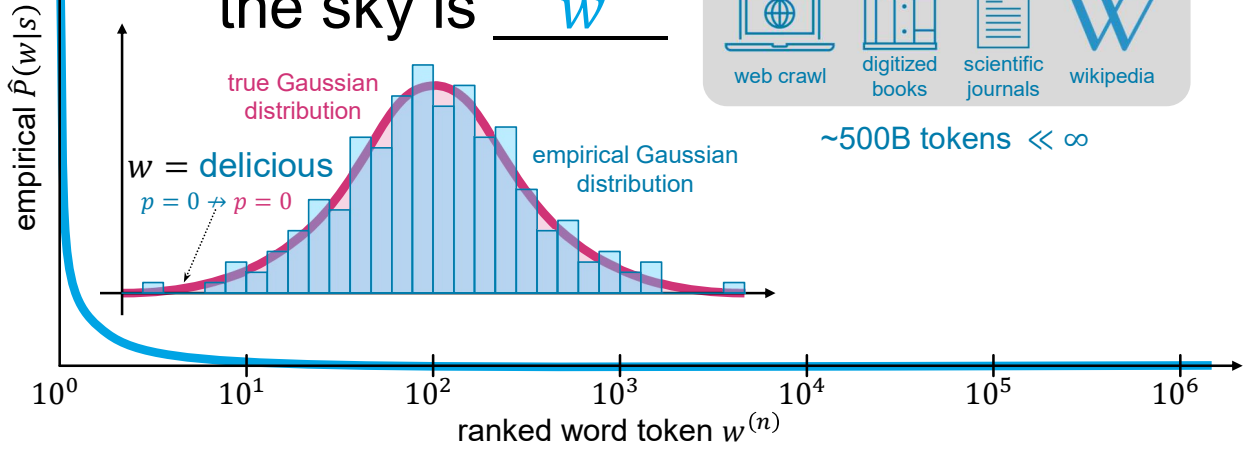
twitter: @mich8elwu
linkedin.com/in/MichaelWuPhD | page 22

22

Empirical Word Probability to Language Model

$s =$ it's raining hard,
the sun just set,
the sky is w

$$P(w|s) \cong \frac{\#(s \rightarrow w)}{\#(s)}$$



©2023 PROS, Inc. All rights reserved. Confidential and Proprietary.

twitter: @mich8elwu
linkedin.com/in/MichaelWuPhD | page 23

23

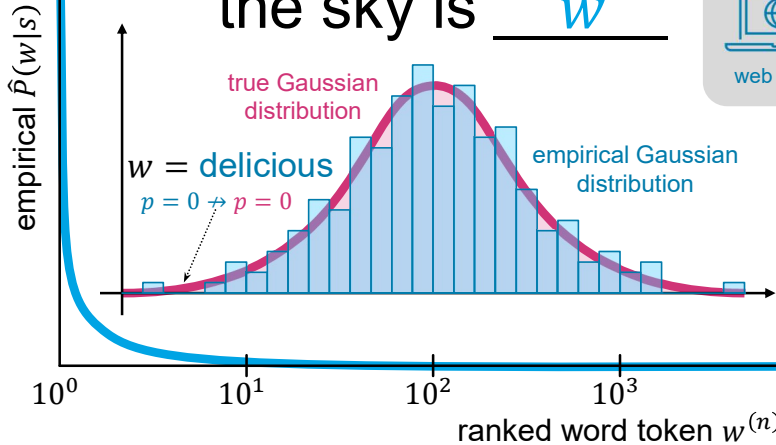


24

Empirical Word Probability to Language Model

$s =$ it's raining hard,
the sun just set,
the sky is w

$$P(w|s) \cong \frac{\#(s \rightarrow w)}{\#(s)}$$



~500B tokens << ∞

$p \neq 0 \rightarrow$ anything is **possible**
under the true distribution

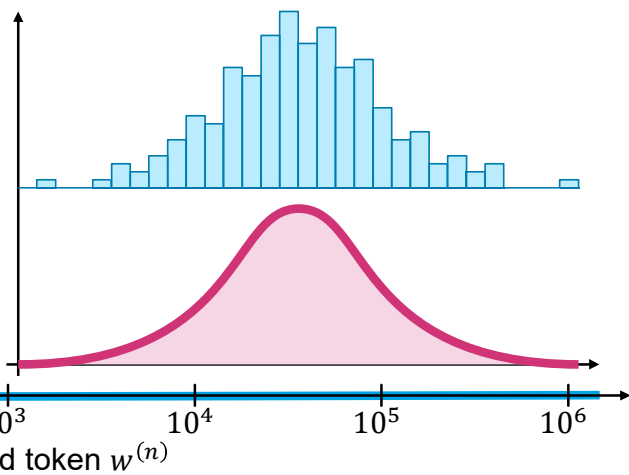
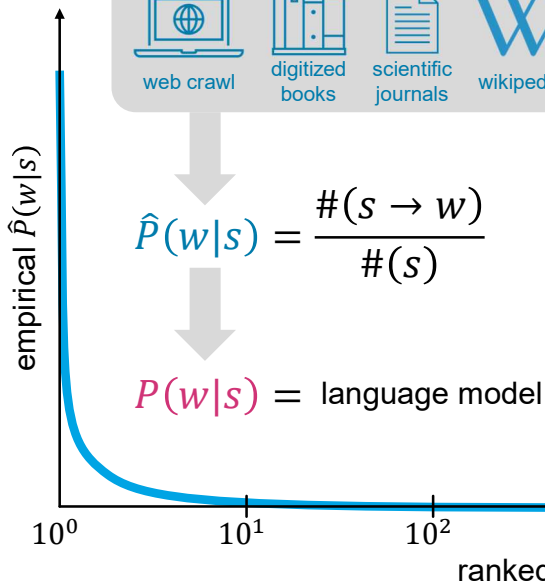
the ability to create novel
content = "generative"



©2023 PROS, Inc. All rights reserved. Confidential and Proprietary.

twitter: @
linkedin.com/in/Michael

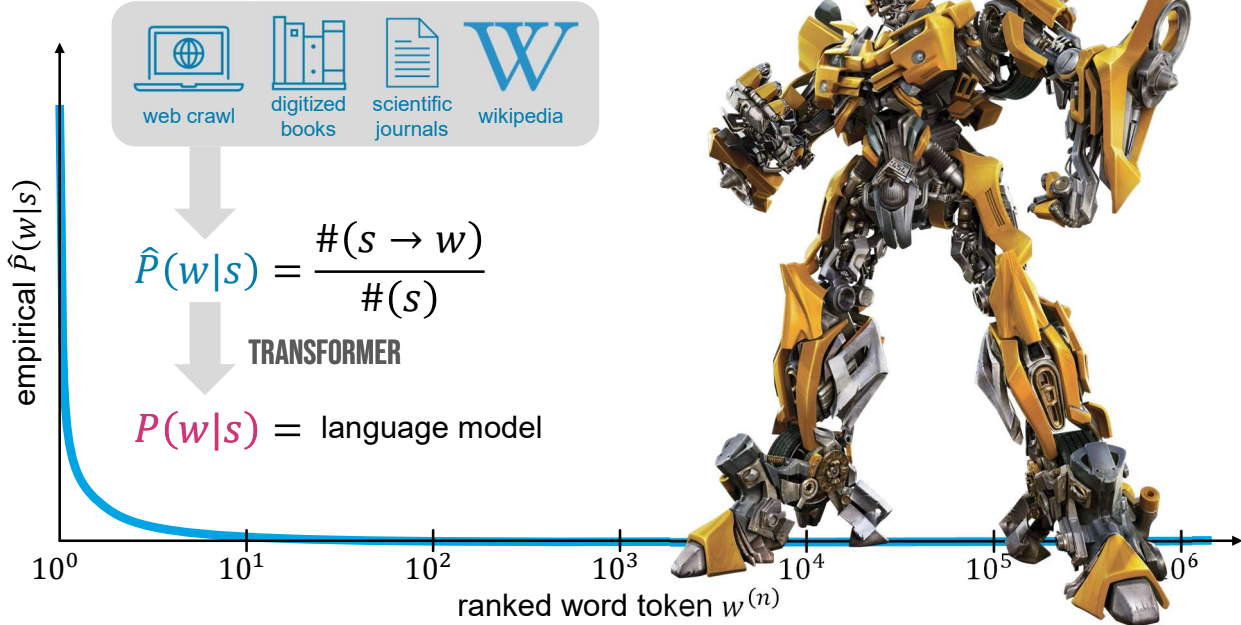
Empirical Word Probability to Language Model



©2023 PROS, Inc. All rights reserved. Confidential and Proprietary.

twitter: @mich8elwu
linkedin.com/in/MichaelWuPhD | page 27

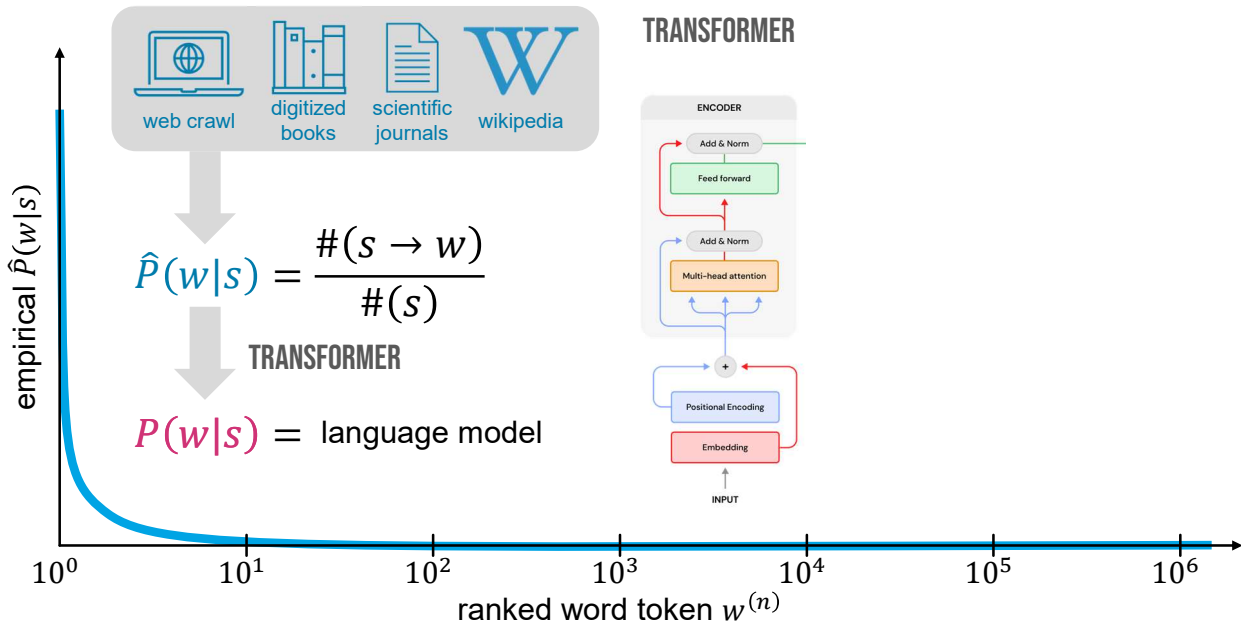
Empirical Word Probability to Language Model



©2023 PROS, Inc. All rights reserved. Confidential and Proprietary.

twitter: @mich8elwu
linkedin.com/in/MichaelWuPhD page 28

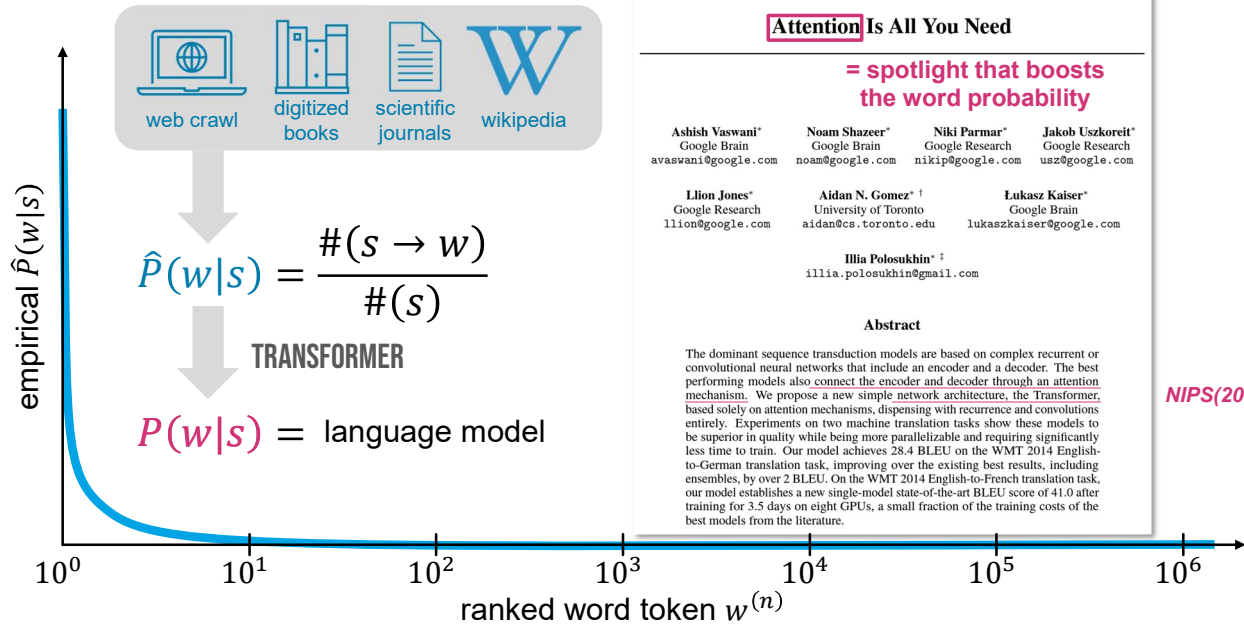
Empirical Word Probability to Language Model



©2023 PROS, Inc. All rights reserved. Confidential and Proprietary.

twitter: @mich8elwu
linkedin.com/in/MichaelWuPhD page 29

Empirical Word Probability to Language Model



Attention Is All You Need

= spotlight that boosts the word probability

Ashish Vaswani*
Google Brain
avaswani@google.com

Noam Shazeer*
Google Brain
noam@google.com

Niki Parmar*
Google Research
nikip@google.com

Jakob Uszkoreit*
Google Research
usz@google.com

Llion Jones*
Google Research
llion@google.com

Aidan N. Gomez* †
University of Toronto
aidan@cs.toronto.edu

Lukasz Kaiser*
Google Brain
lukaszkaizer@google.com

Illia Polosukhin* ‡
illia.polosukhin@gmail.com

Abstract

The dominant sequence transduction models are based on complex recurrent or convolutional neural networks that include an encoder and a decoder. The best performing models also connect the encoder and decoder through an attention mechanism. We propose a new simple network architecture, the Transformer, based solely on attention mechanisms, dispensing with recurrence and convolutions entirely. Experiments on two machine translation tasks show these models to be superior in quality while being more parallelizable and requiring significantly less time to train. Our model achieves 28.4 BLEU on the WMT 2014 English-to-German translation task, improving over the existing best results, including ensembles, by over 2 BLEU. On the WMT 2014 English-to-French translation task, our model establishes a new single-model state-of-the-art BLEU score of 41.0 after training for 3.5 days on eight GPUs, a small fraction of the training costs of the best models from the literature.

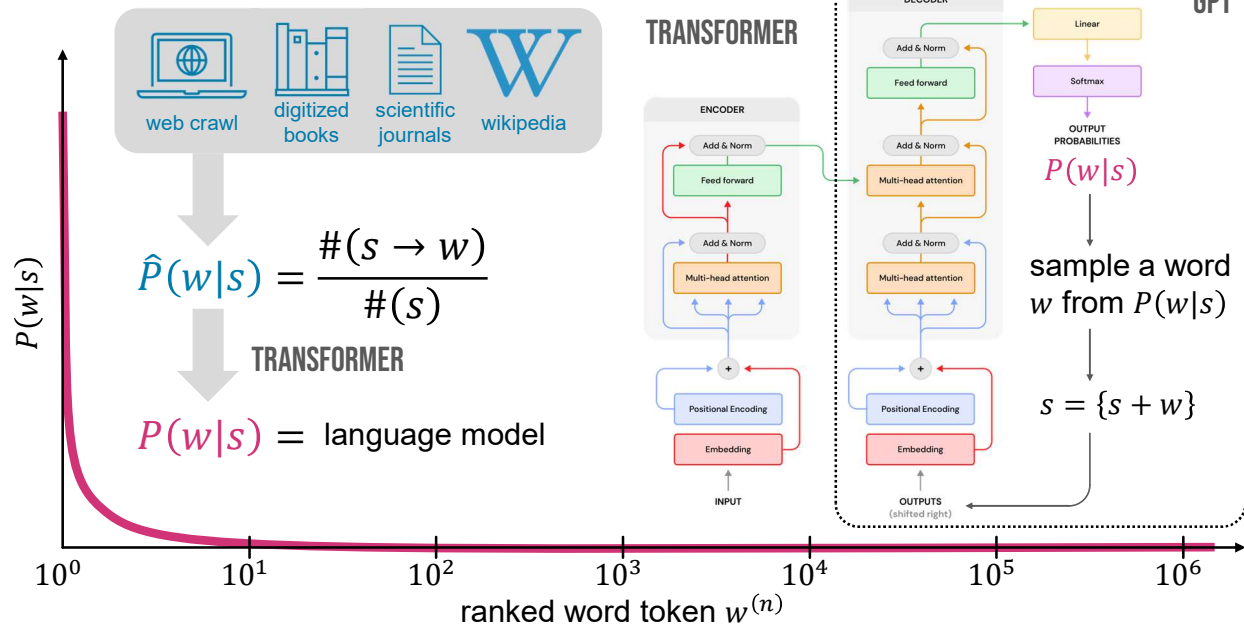
NIPS(2017)



©2023 PROS, Inc. All rights reserved. Confidential and Proprietary.

twitter: @mich8elwu
linkedin.com/in/MichaelWuPhD | page 30

Empirical Word Probability to Language Model



©2023 PROS, Inc. All rights reserved. Confidential and Proprietary.

twitter: @mich8elwu
linkedin.com/in/MichaelWuPhD | page 31

GPT: Generative Pre-trained Transformer is a large language model (LLM)

random number generator
from a distribution over *all* words
given *any* word sequence
trained with human written text
using transformer architecture

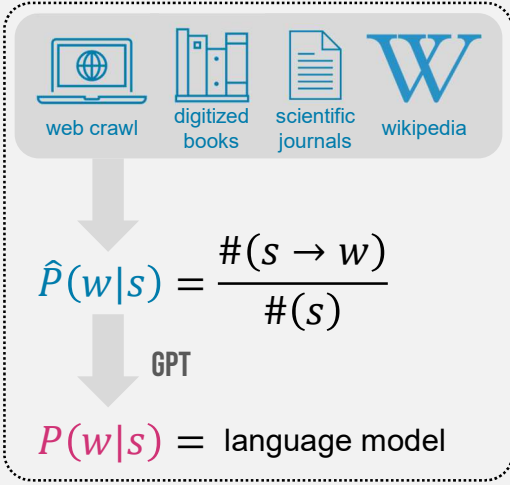
32

doesn't sound
very intelligent

33

From GPT to ChatGPT

pre-trained via supervised/
self-supervised



plausible text continuation
≠ good responses

- supervised *transfer learning* to finetune the model to follow instructions + provide answers

good responses ≠ good dialog

- *reinforcement learning* with human-in-the-loop ranking of good dialogue responses

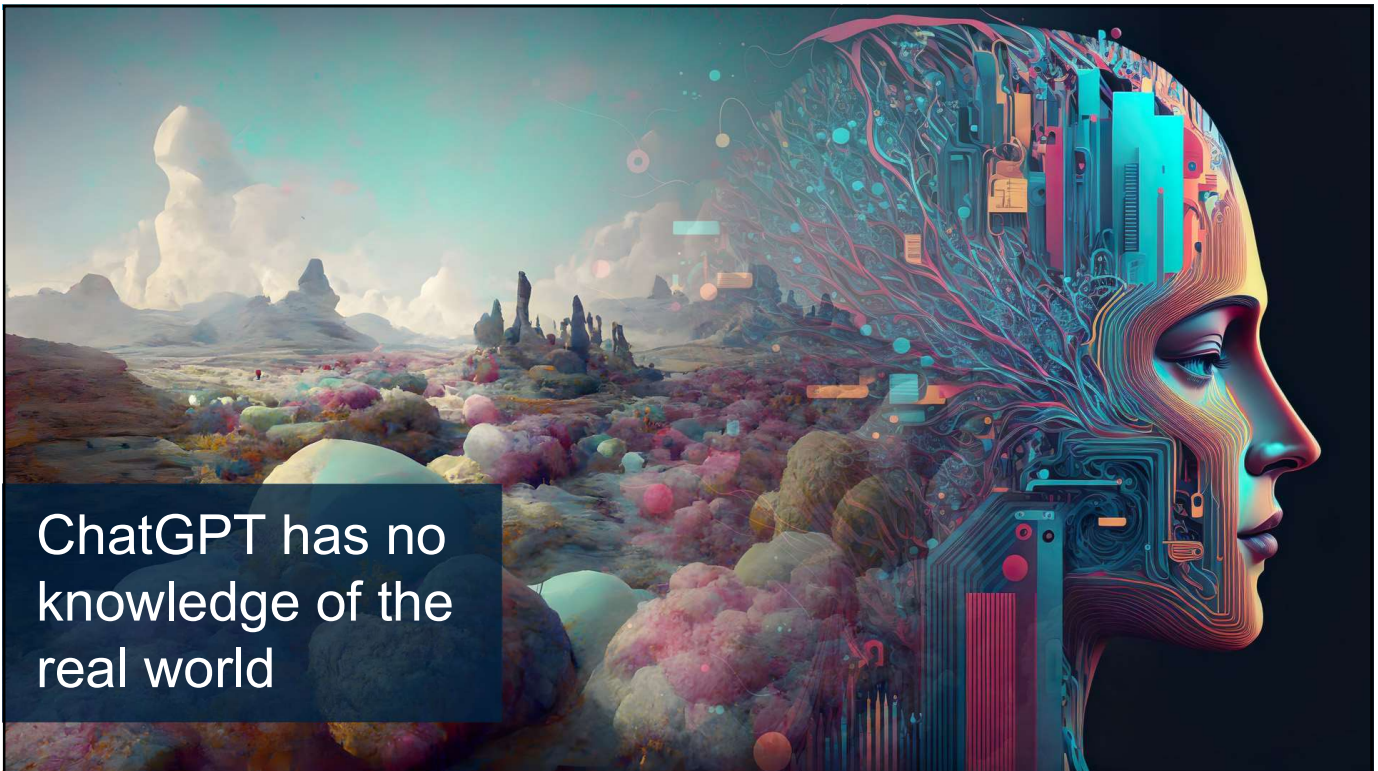
transfer learning

reinforcement learning

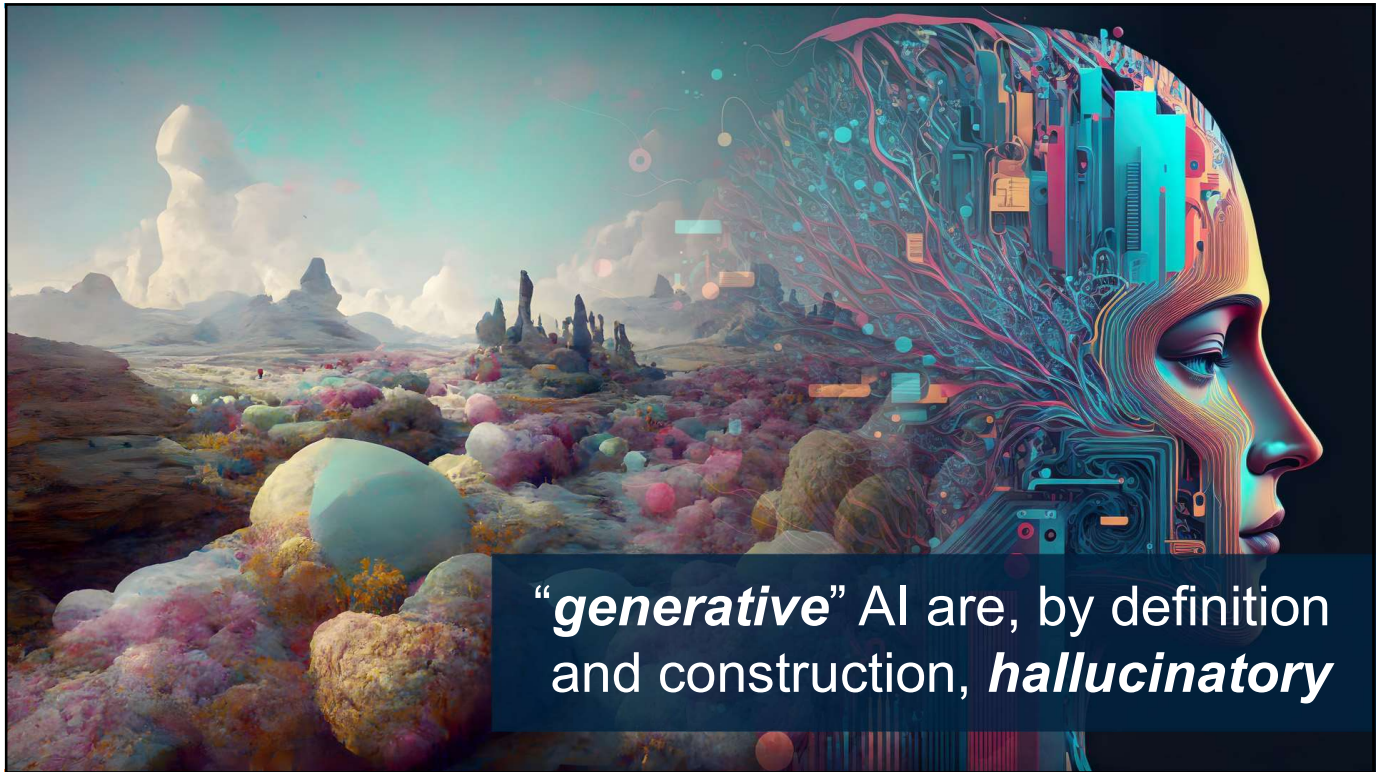


©2023 PROS, Inc. All rights reserved. Confidential and Proprietary.

twitter: @mich8elwu
linkedin.com/in/MichaelWuPhD | page 39



ChatGPT has no knowledge of the real world



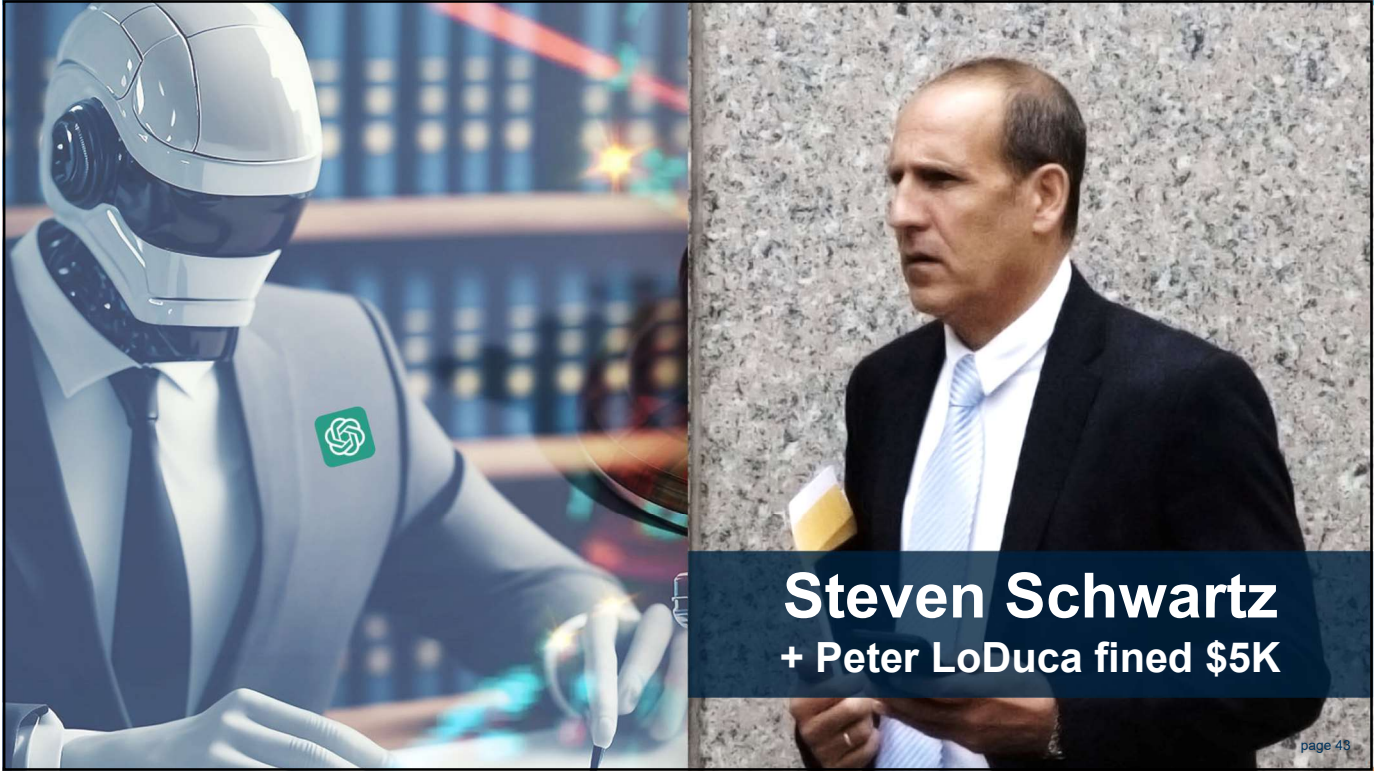
“*generative*” AI are, by definition and construction, *hallucinatory*

41



very convincing legal briefings
citing non-existent court cases

42



43



45



feature: for *design*
+ *creative*
use cases

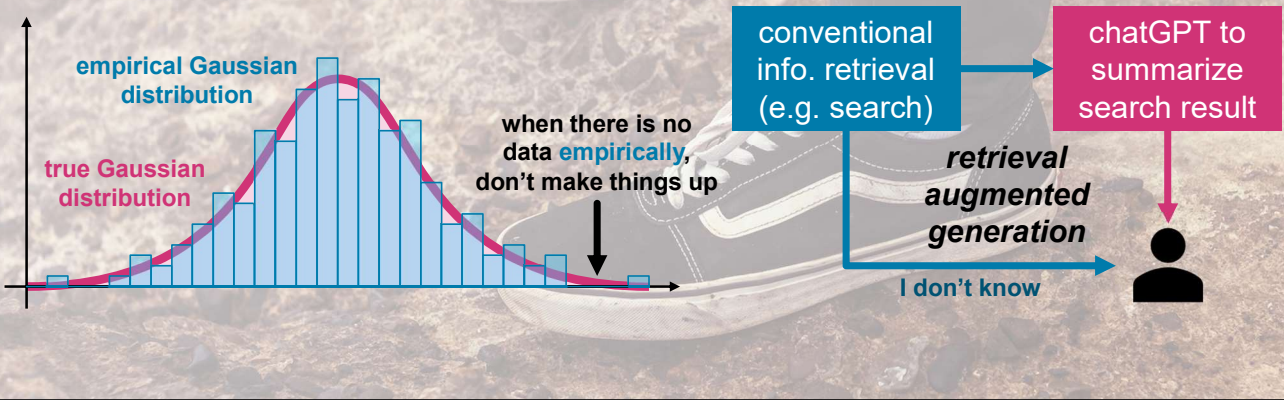
46



bug: for *fact-based*
applications

47

fact-based applications require *grounding*



48



50

2 ways to work with LLM

model fine-tuning

pros

- knowledge encoded into the model parameters
- can teach it anything

cons

- costly: 25,000 × nvidia A100 for ~100 days ~\$63M → GPT4
- must be retrained when there's new data or new LLMs
- hard to iterate, slow time to market

RAG: prompting

pros

- no upfront cost
- no retraining on new data
- easily swap in/out different LLM
- easy to iterate, fast time to market

cons

- limited context length (GPT4: ~8k tokens)
- knowledge accuracy depends on retrieval mechanism (search)

52

Beyond ChatGPT

| | | generic generative AI | | | | specialized generative AI | | | |
|-------------|-------------------------|-----------------------|-------------------|-------------------|-----------------|---------------------------|--------------|--------------------|-------|
| data | textual | | visual | | audio | | game | specialized design | |
| | text | code | image | video | speech | music | 3D model | biotech | other |
| model | BERT | Codex/GPT4 | Dall-E2 | X-Clip | Whisper | Jukebox | DreamFusion | AlphaFold | |
| | GPT | Github copilot | Make-a-Scene | Make-a-Video | data2vec | Riffusion | nvidia Get3D | RoseTTAFold | |
| | Gopher | tabnine | Craiyon | Imagen Video | | dance diffusion | human MDM | | |
| | LaMDA | stability.ai | Midjourney | | | musicLM | | | |
| | PaLM | CodeWhisperer | stable diffusion | | | | | | |
| | BARD | | Imagen | | | | | | |
| | Meta OPT | | nvidia eDiff-I | | | | | | |
| | LLaMA | | | | | | | | |
| application | general writing | code generation | image generation | video generation | voice synthesis | song/music creation | | | |
| | summarize + note taking | documentation | media/advertising | video edit/modify | | | | | |
| | compare/contrast | text to SQL | 2D design | | | | | | |
| | content creation | web app builder | social media | | | | | | |
| | question/answer | | | | | | | | |
| | realtime translation | | | | | | | | |

more models to come

more use cases to come

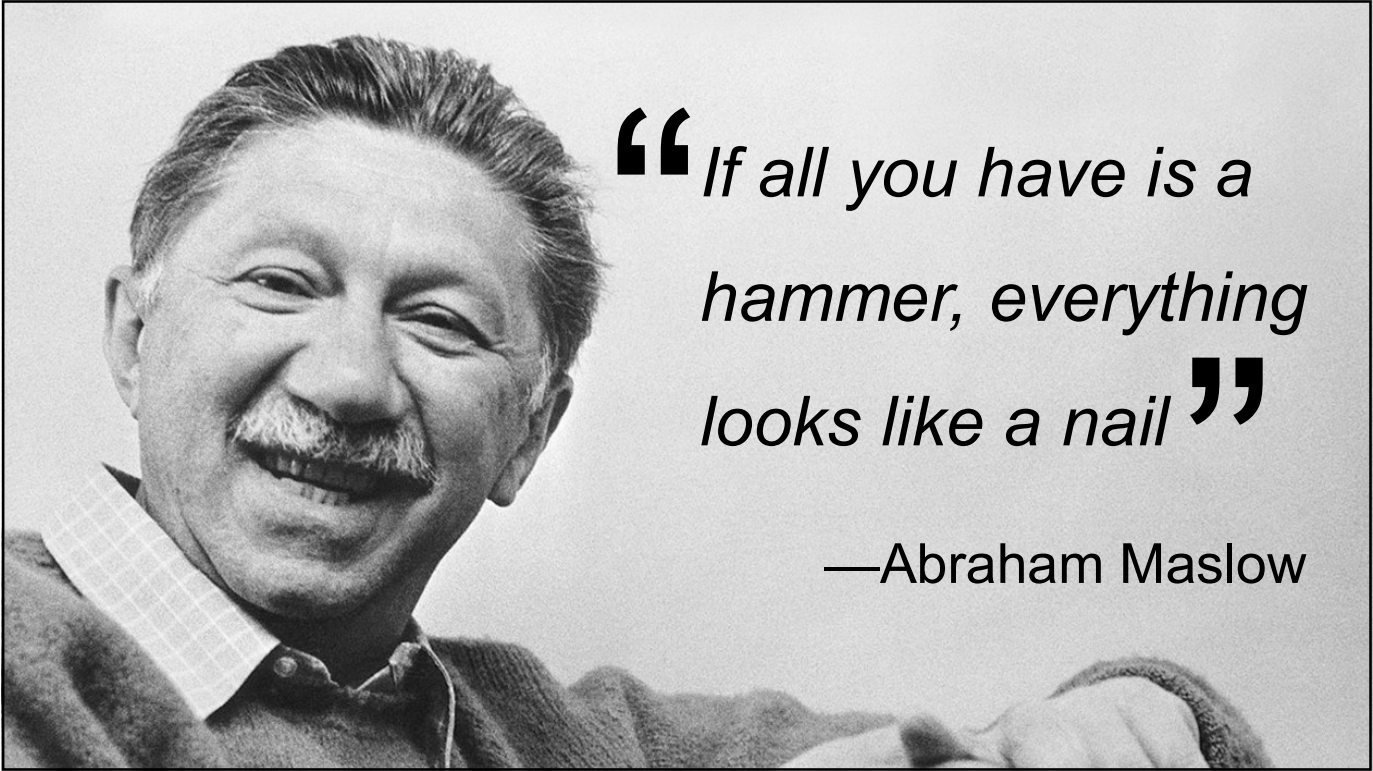
many many more start-ups

©2023 PROS, Inc. All rights reserved. Confidential and Proprietary.

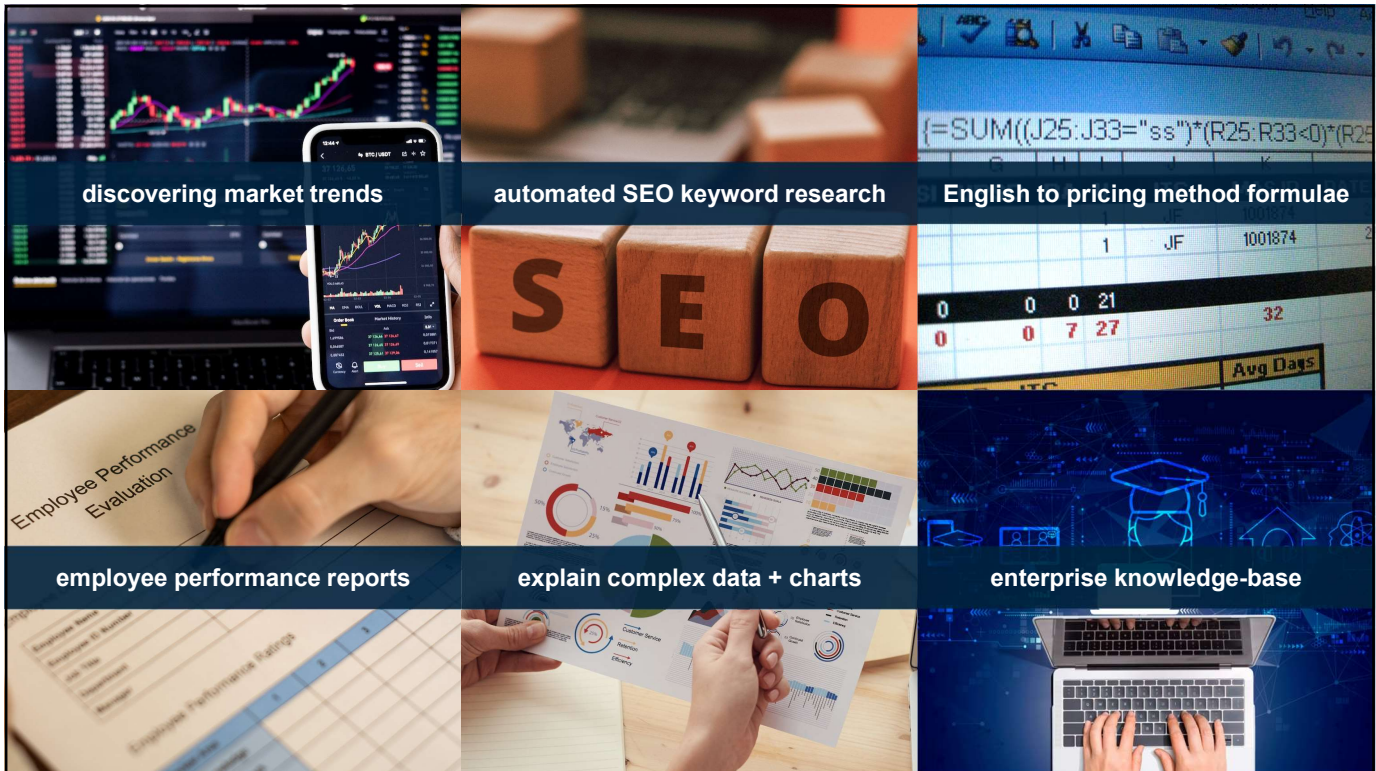
twitter: @mich8elwu
 linkedin.com/in/MichaelWuPhD

page 53

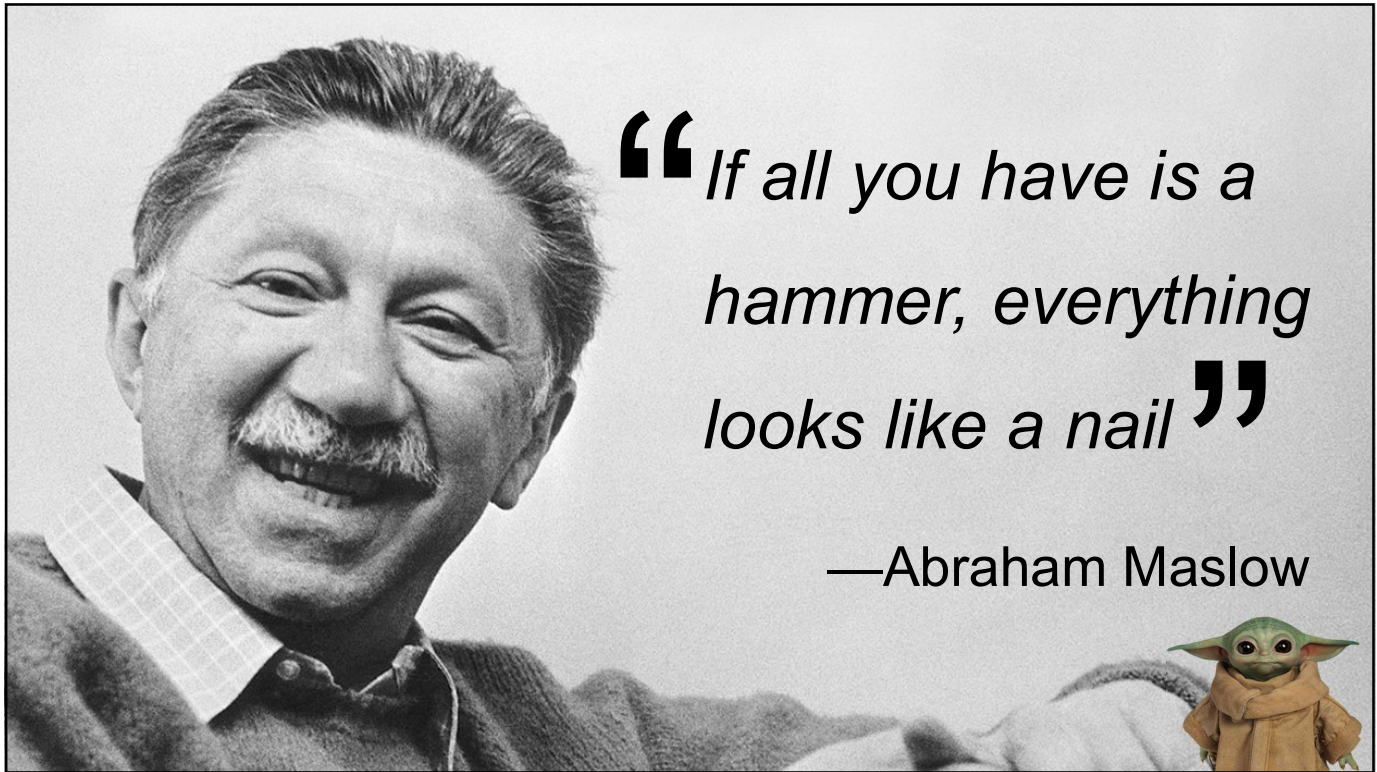
53



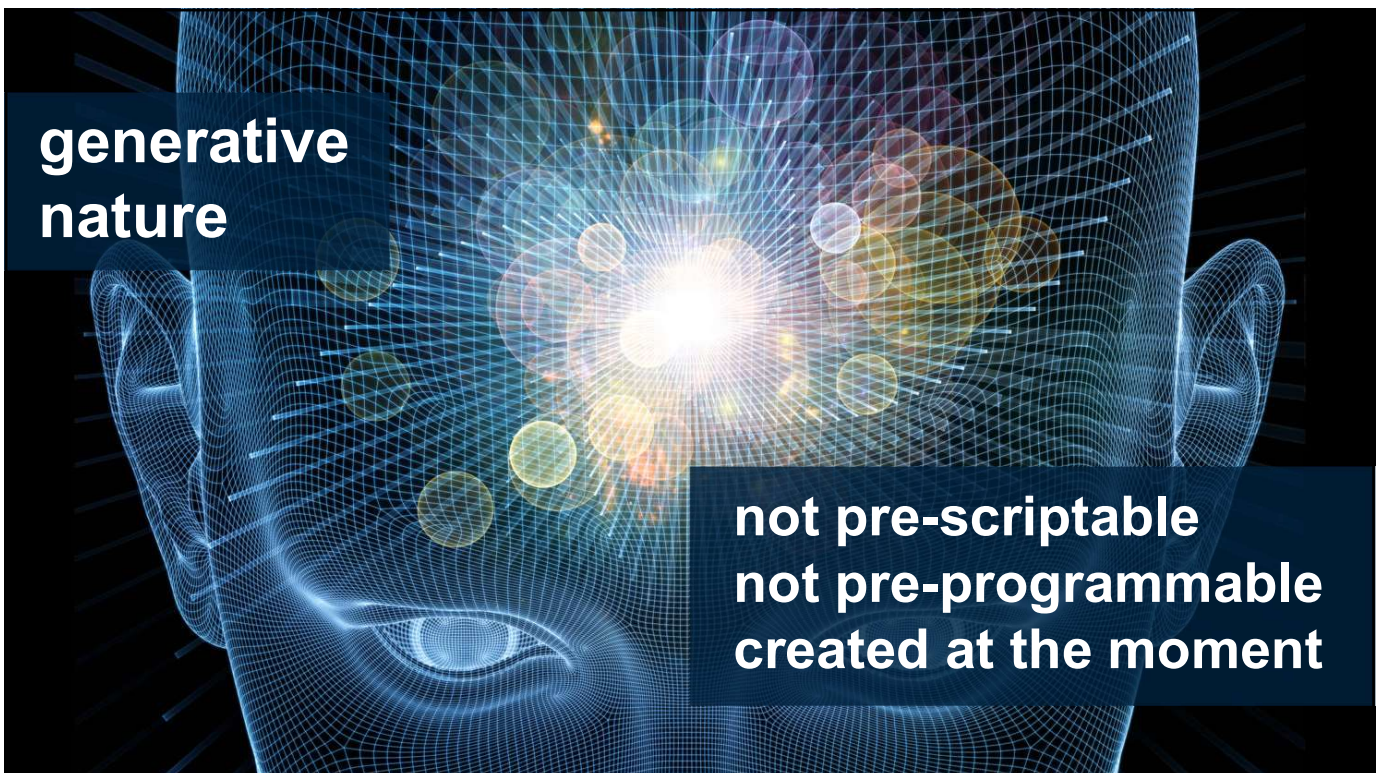
54



55



56



59

Beyond ChatGPT

| | | generic generative AI | | | | | | specialized generative AI | | |
|-------|----------|-----------------------|------------------|--------------|----------|-----------------|--------------|---------------------------|-------|--|
| data | textual | | visual | | audio | | game | specialized design | | |
| | text | code | image | video | speech | music | 3D model | biotech | other | |
| model | BERT | Codex/GPT4 | Dall-E2 | X-Clip | Whisper | Jukebox | DreamFusion | AlphaFold | | |
| | GPT | GitHub copilot | Make-a-Scene | Make-a-Video | data2vec | Riffusion | nvidia Get3D | RoseTTAFold | | |
| | Gopher | tabnine | Crayon | Imagen Video | | dance diffusion | human MDM | | | |
| | LaMDA | stability.ai | Midjourney | | | musicLM | | | | |
| | PaLM | CodeWhisperer | stable diffusion | | | | | | | |
| | BARD | | Imagen | | | | | | | |
| | Meta OPT | | nvidia eDiff-I | | | | | | | |
| | LLaMA | | | | | | | | | |
| | | | | | | | | | | |
| | | | | | | | | | | |

more models to come

| application | general writing | code generation | image generation | video generation | voice synthesis | song/music creation |
|----------------------|-------------------------|-----------------|-------------------|-------------------|-----------------|---------------------|
| | summarize + note taking | documentation | media/advertising | video edit/modify | | |
| compare/contrast | text to SQL | 2D design | | | | |
| content creation | web app builder | social media | | | | |
| question/answer | | | | | | |
| realtime translation | | | | | | |

more use cases to come

many many more start-ups

PROS ©2023 PROS, Inc. All rights reserved. Confidential and Proprietary. twitter: @mich3elwu linkedin.com/in/MichaelWuPhD page 60

60



61



Kristina Kashtanova

62

Good Use Cases

Github copilot/Codex

use case ask *any* natural language question → SQL generation for a known DB schema

prompt DB schema, data dictionary, column definitions, etc.

example "what is the total margin lift for my French customers last quarter?" → **€5.78M**

guardrail

- read-only
- respect access permission

should ALWAYS have some guardrails b/c so much is unknown with GenAI

value/adoption

- explain aggregated results
- human languages are imprecise
- step through calculations
- without trust there is no value

63

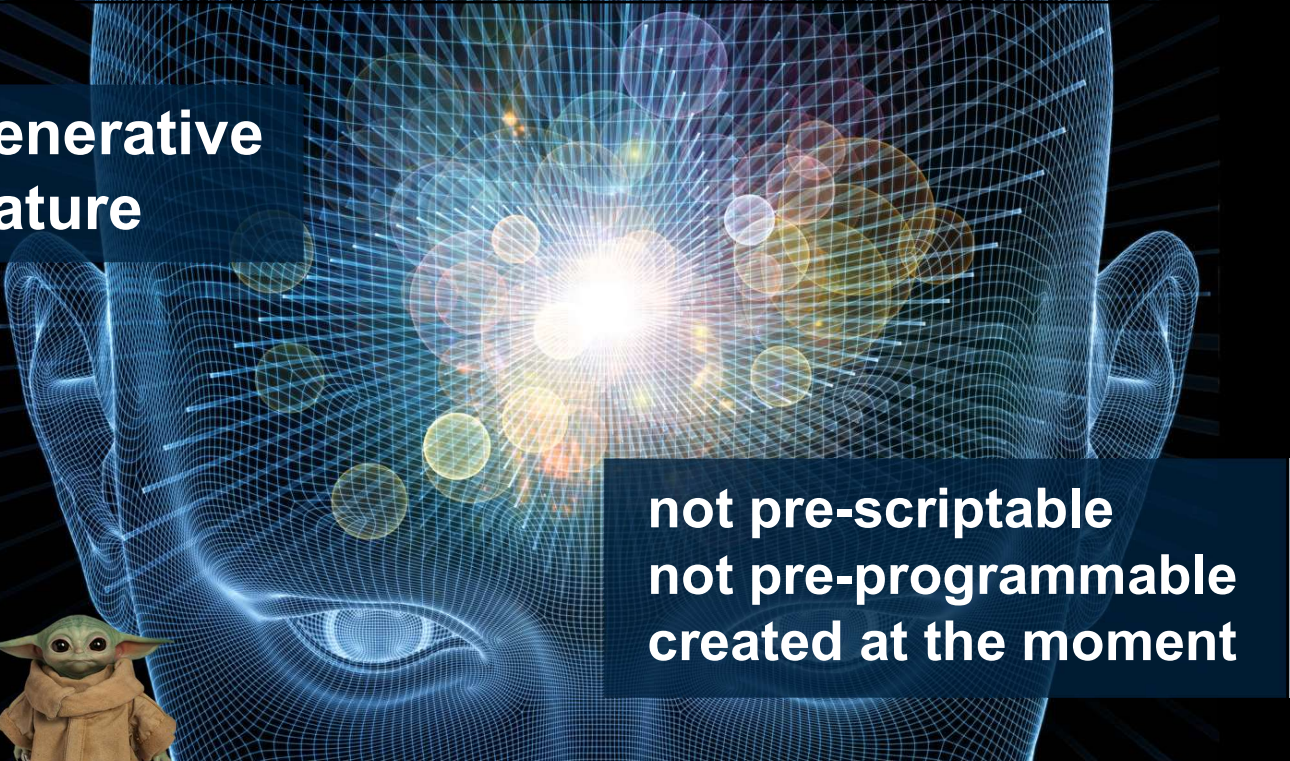


CHATGPT CODE INTERPRETER




if you can translate English to code,
you can pretty much do anything

64

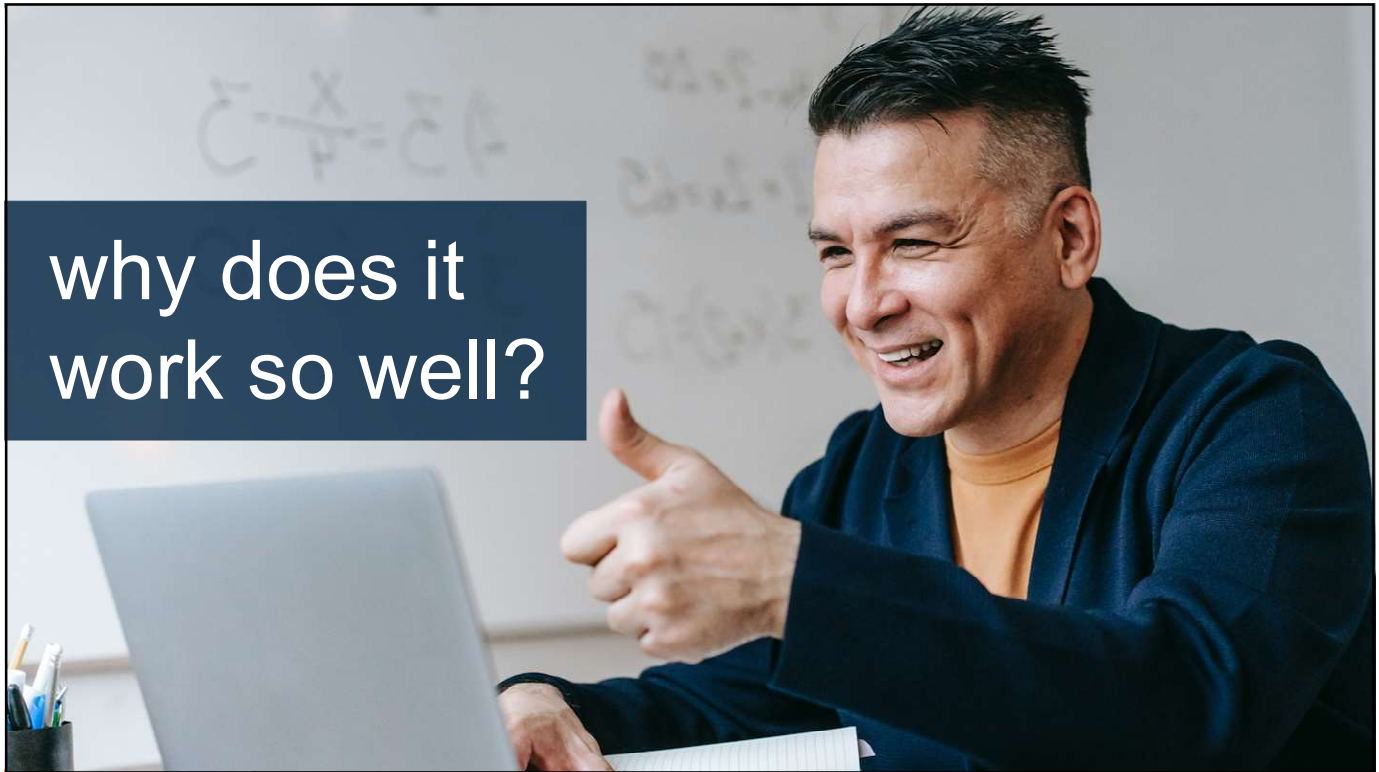


generative
nature

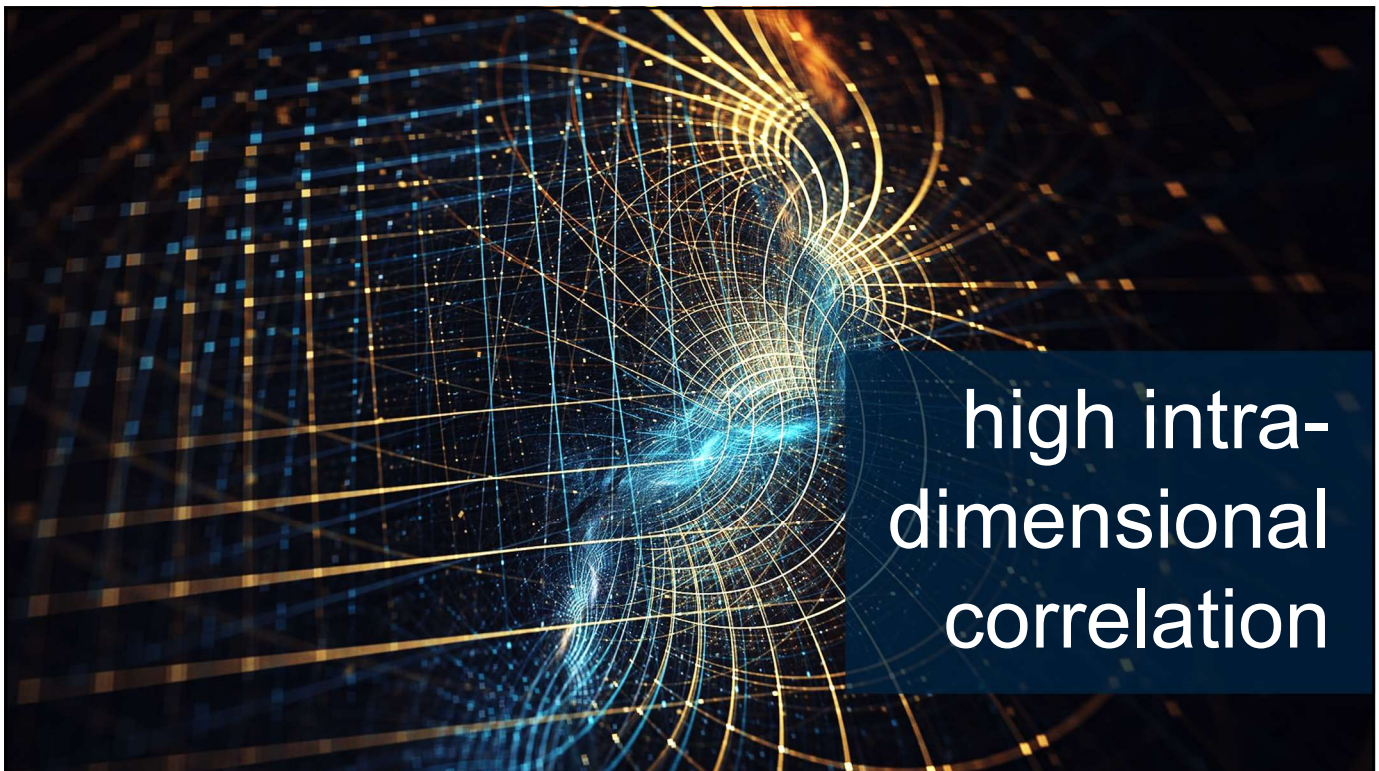
not pre-scriptable
not pre-programmable
created at the moment



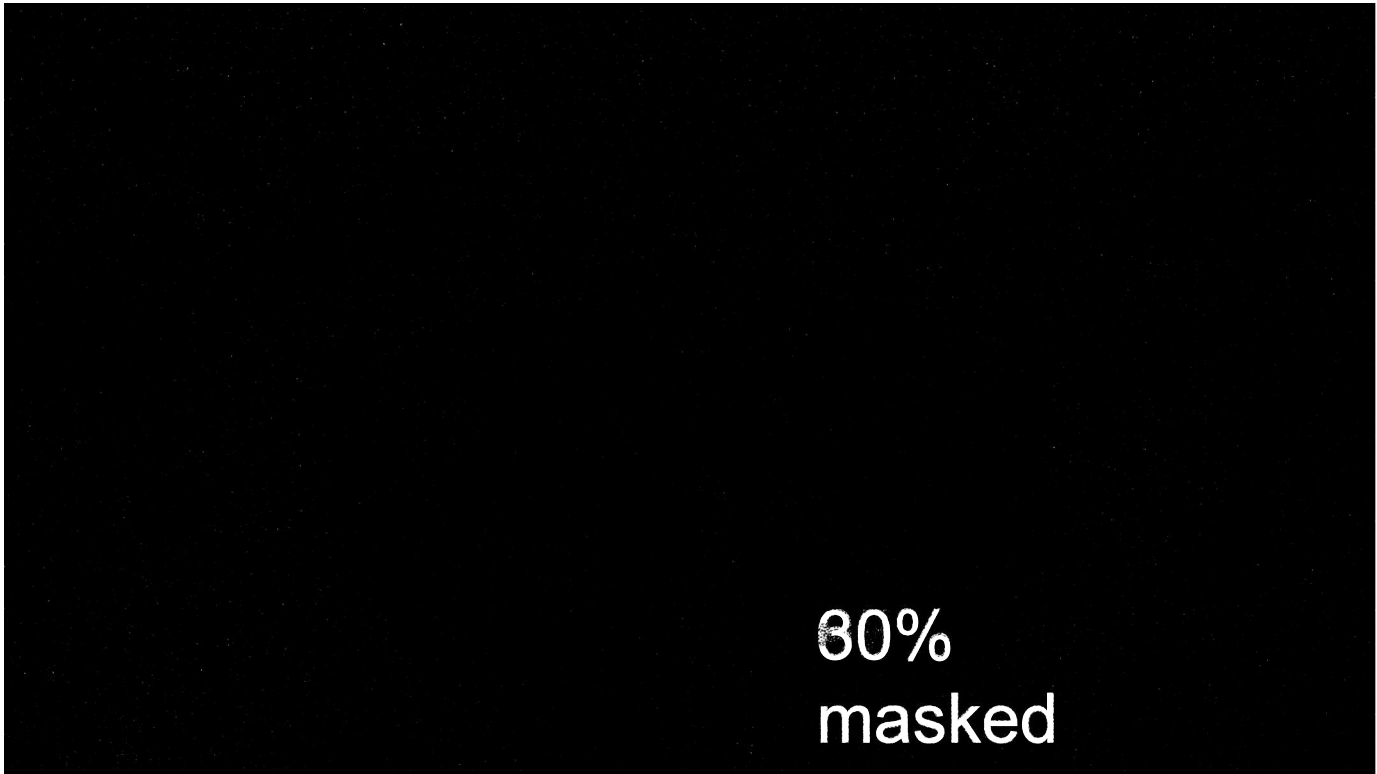
69



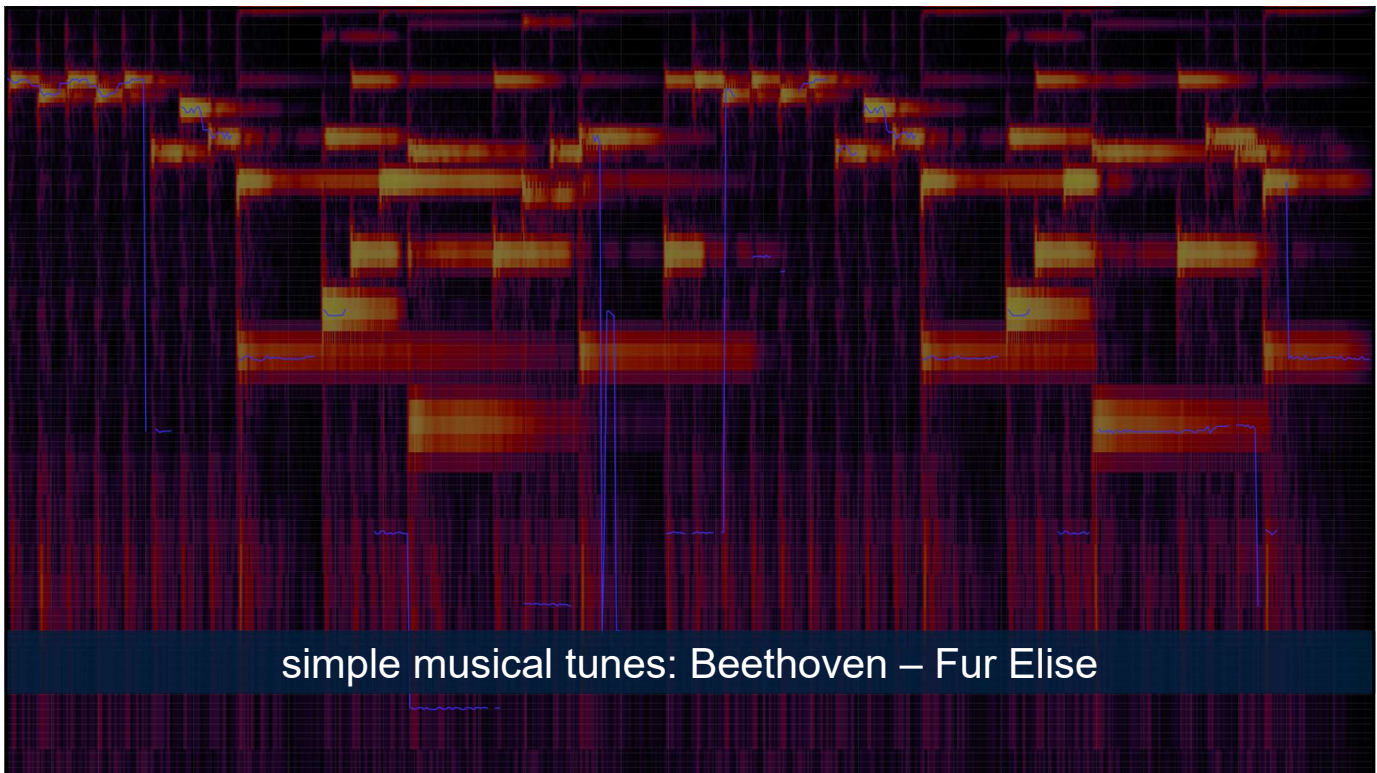
70



71



72



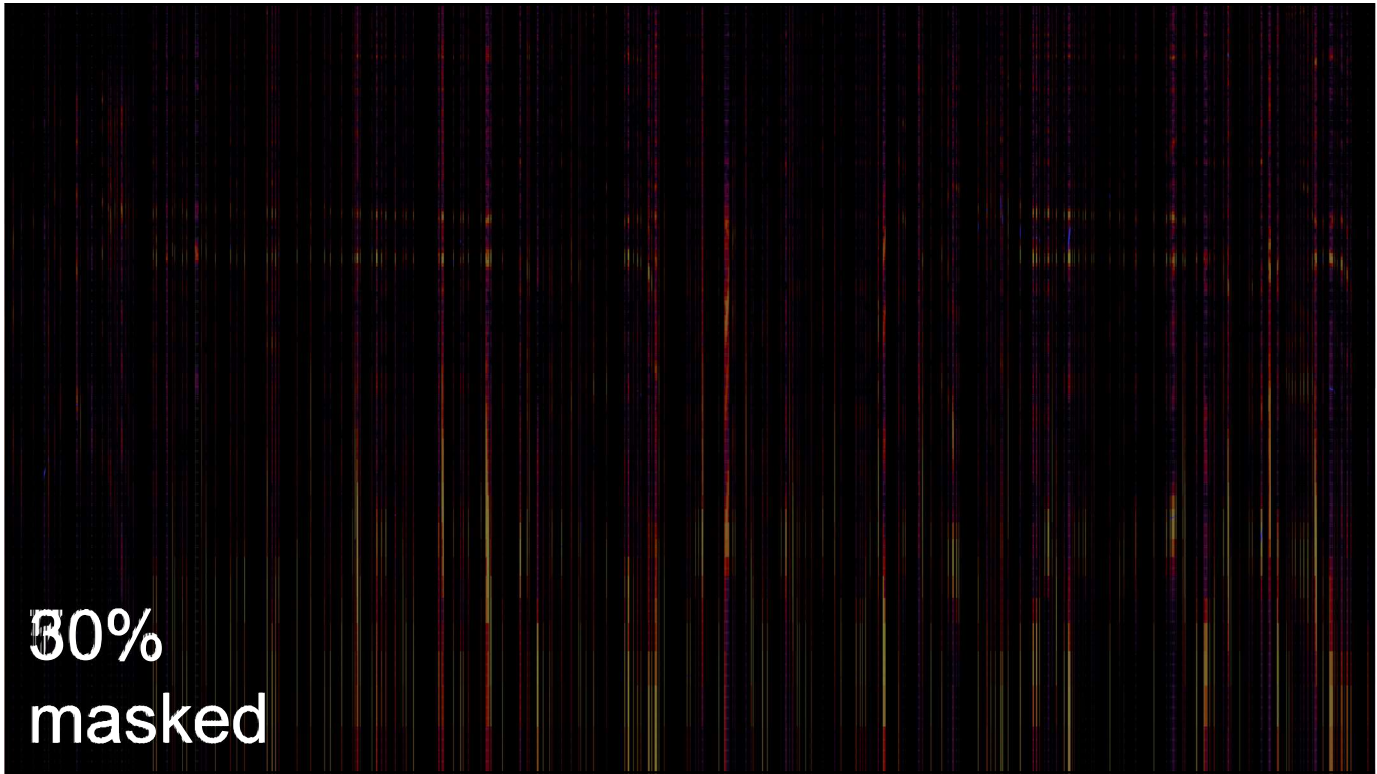
73



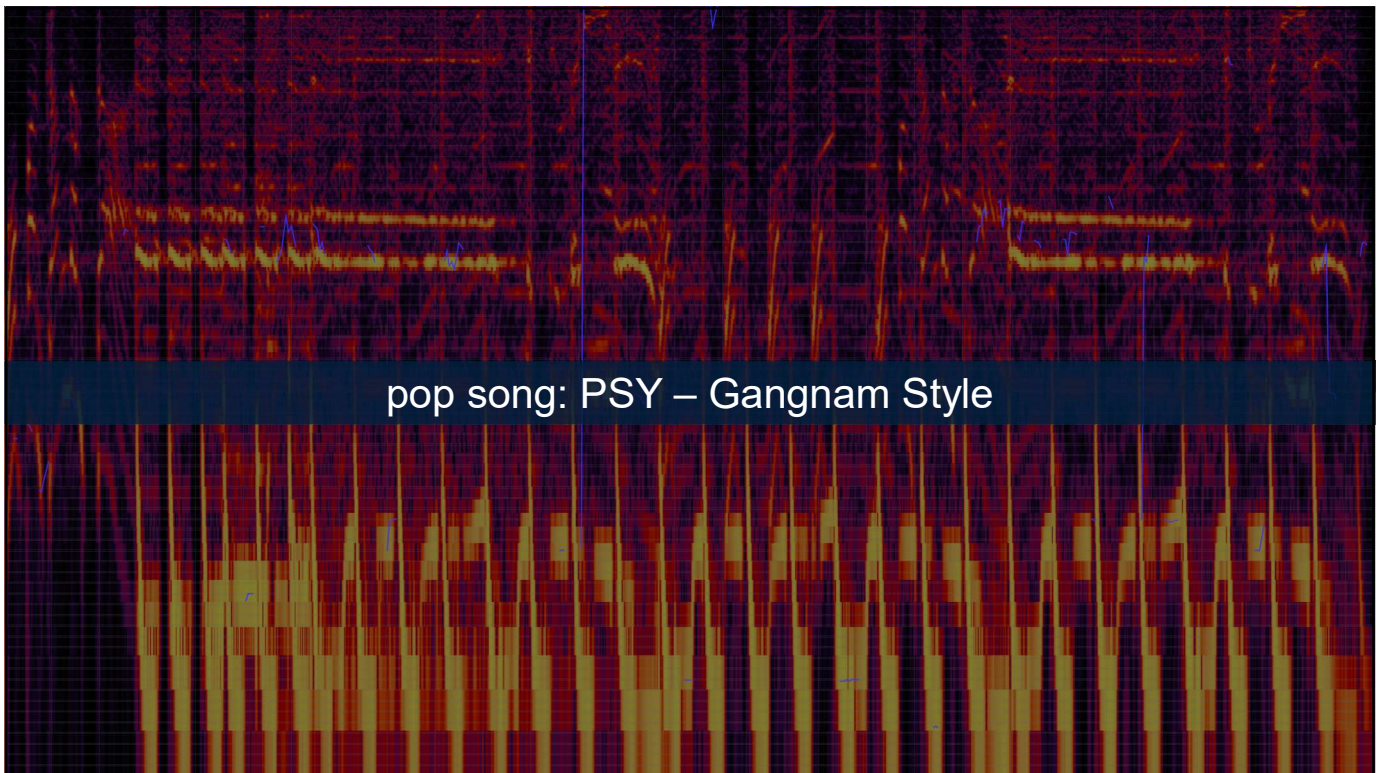
74



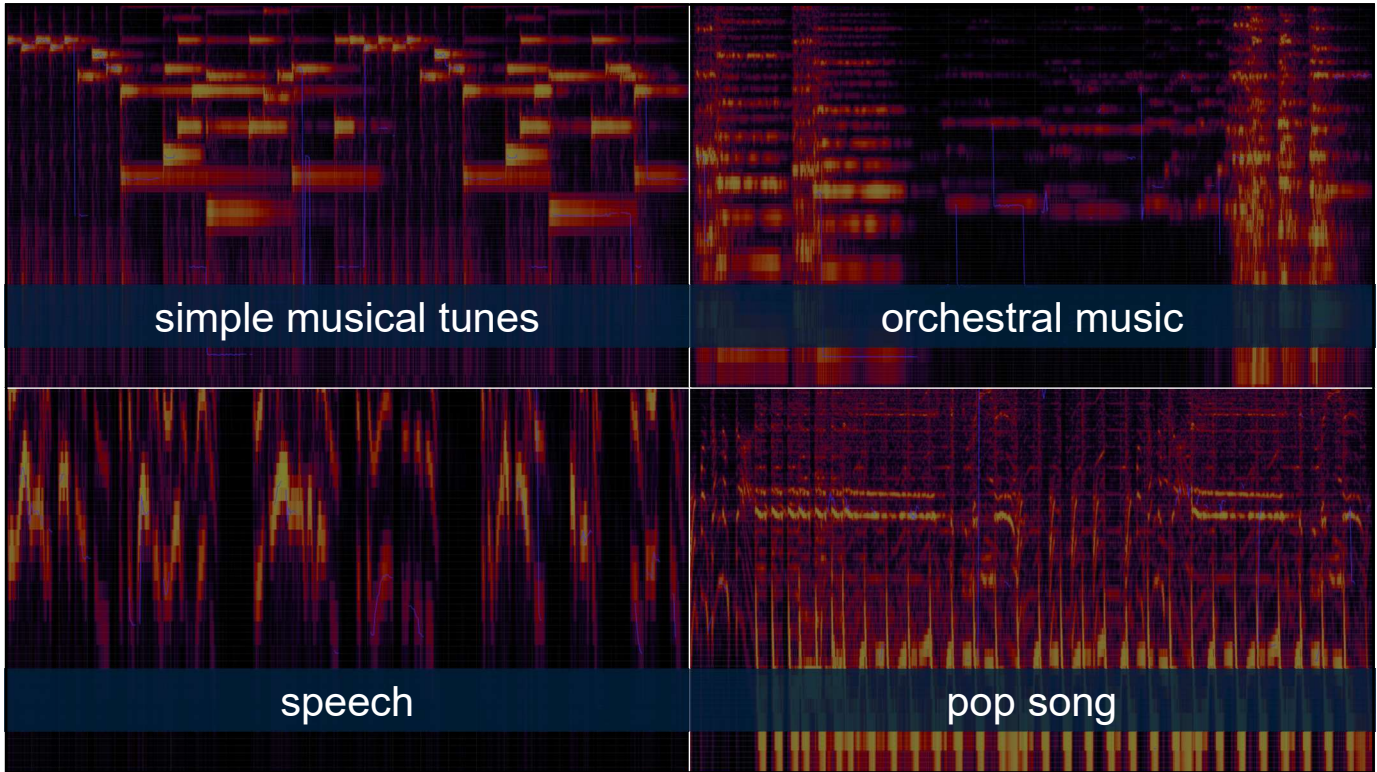
75



76



77



78

ChatGPT and Beyond

50th b-day poem from ChatGPT

Happy birthday, Dr. Michael Wu
 Today we celebrate all that you do
 Your wisdom and insight we hold so dear
 Your guidance and leadership, always so clear

In the field of revenue management
 You are a true expert and phenomenon
 Your knowledge and skills have paved the way
 For success and growth, day by day

Your dedication and hard work
 Are an inspiration, no need to lurk
 Your passion for learning and innovation
 Have brought us all great motivation

So on this special day, we honor you
 For all that you have accomplished, it's true
 We wish you joy, happiness, and health
 May your dreams and goals be realized with wealth

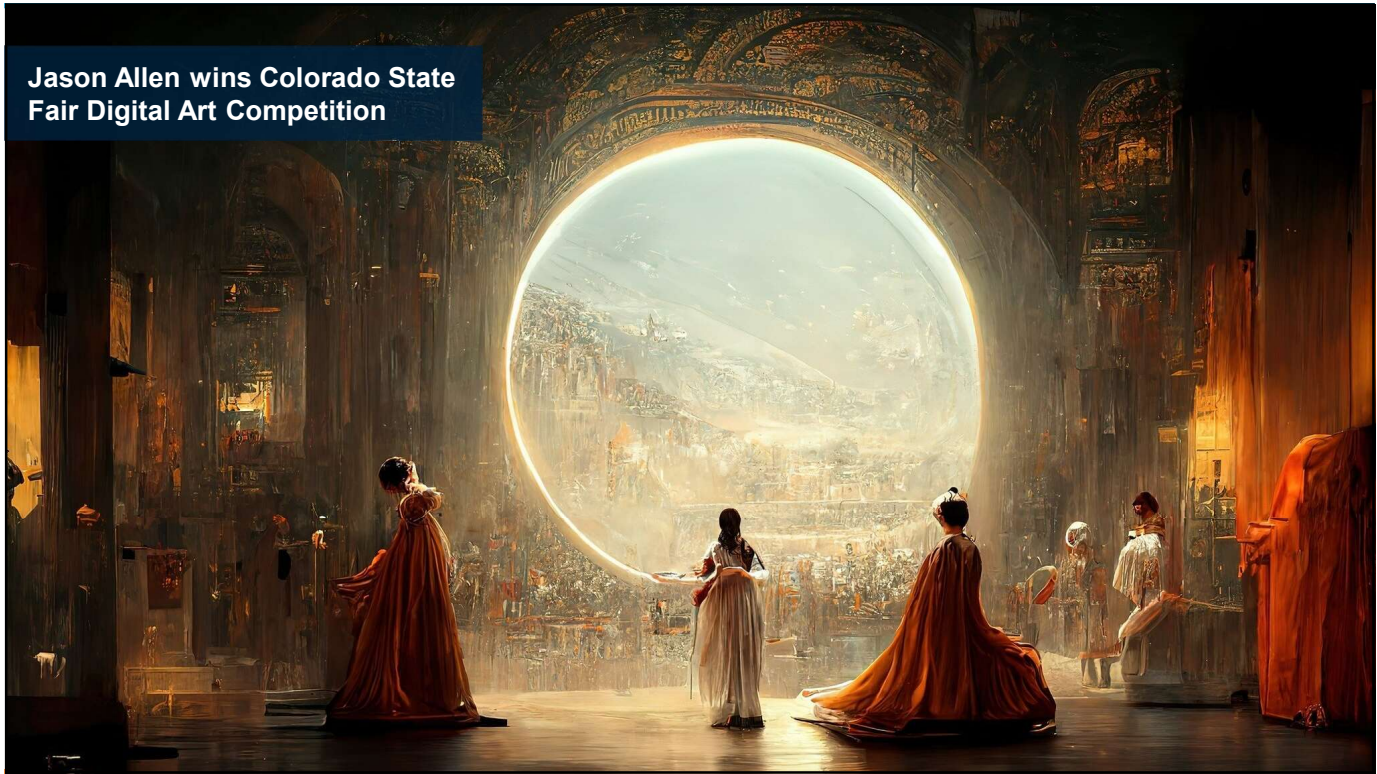
Happy birthday, Dr. Michael Wu
 We are grateful for all that you do
 May your year ahead be filled with delight
 And your future be forever bright



Boris Eldagsen wins Sony World Photography Award



79



80

Forbes

FORBES > INNOVATION

BREAKING

AI-Generative Music Here—‘Siri’ Like Drake Grande—Very Illegal

Arianna Johnson Forbes Senior Editor
I cover the latest trends in sports and healthcare.

0

Drake performs at the 2017 Grammy Awards

WIRED

AMANDA HOOPER CULTURE

AI-Generated Streaming

There are already... And as tunes become...

f t e

PHOTOGRAPH: OBNITOSTOCK

THE HILL

OPINION > TECHNOLOGY

THE VIEWS EXPRESSED BY CONTRIBUTORS ARE THEIR OWN

AI-generated music: how will it change the industry?

BY FERNANDO GARIBAY, KASHYAP KOMPELLA, CONTRIBUTORS - 04/26/23 4:00 PM ET

MUSICBUSINESS WORLDWIDE

MUSIC MADE BY ARTIFICIAL INTELLIGENCE HAS STARTED WINNING AWARDS. GRAMMYS NEXT?

935 SHARES

f t in e

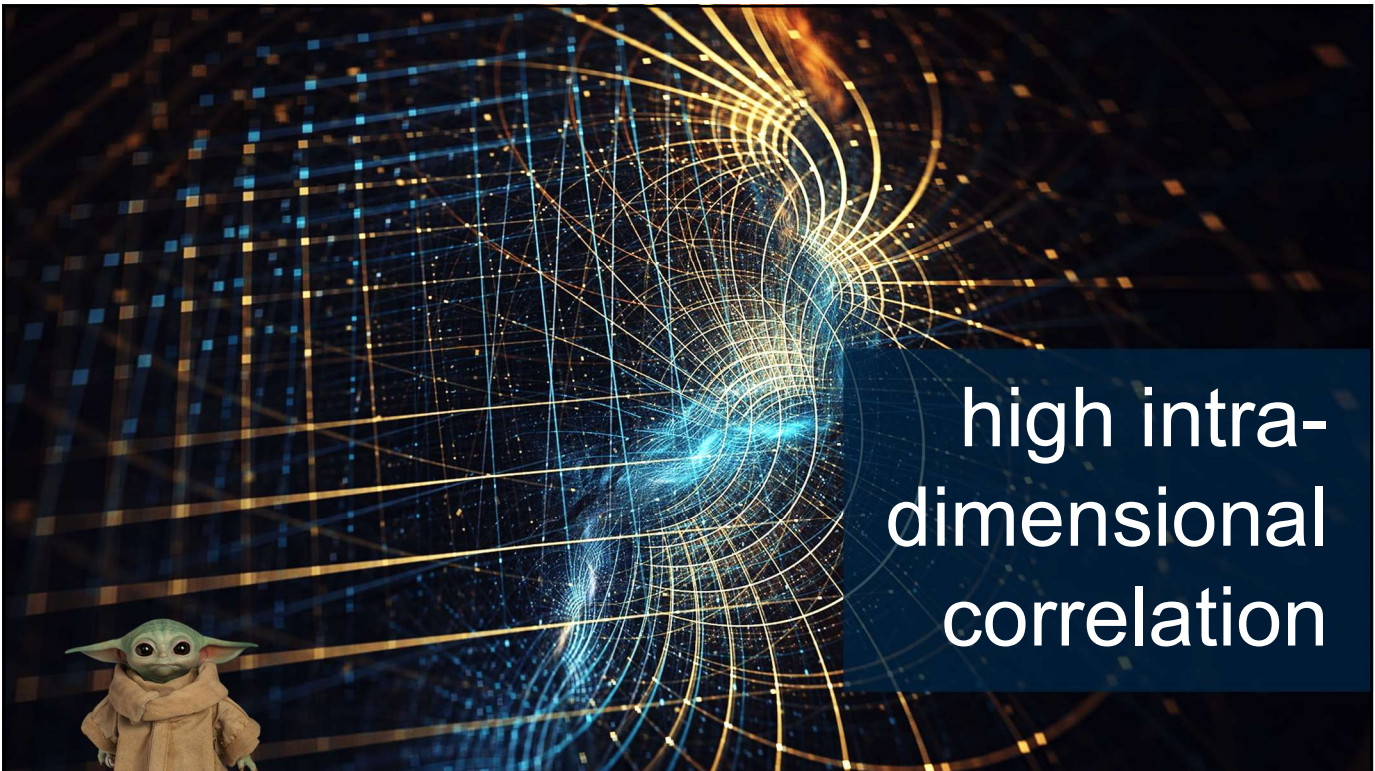
MAY 14, 2020

BY TIM INGHAM

81



82

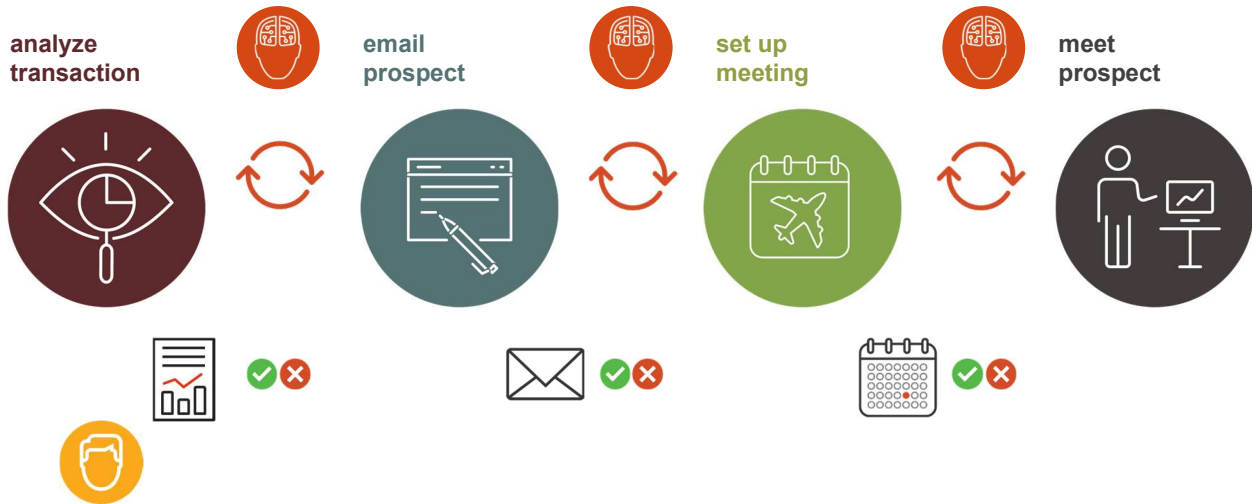


83

AI WILL CATALYZE THE 4TH INDUSTRIAL REVOLUTION



AI Shifting Human's Job (e.g. Sales Rep.)





WE ARE EXCHANGING DATA FOR AUTOMATION

100

AI Shifting Human's Job (e.g. Sales Rep.)

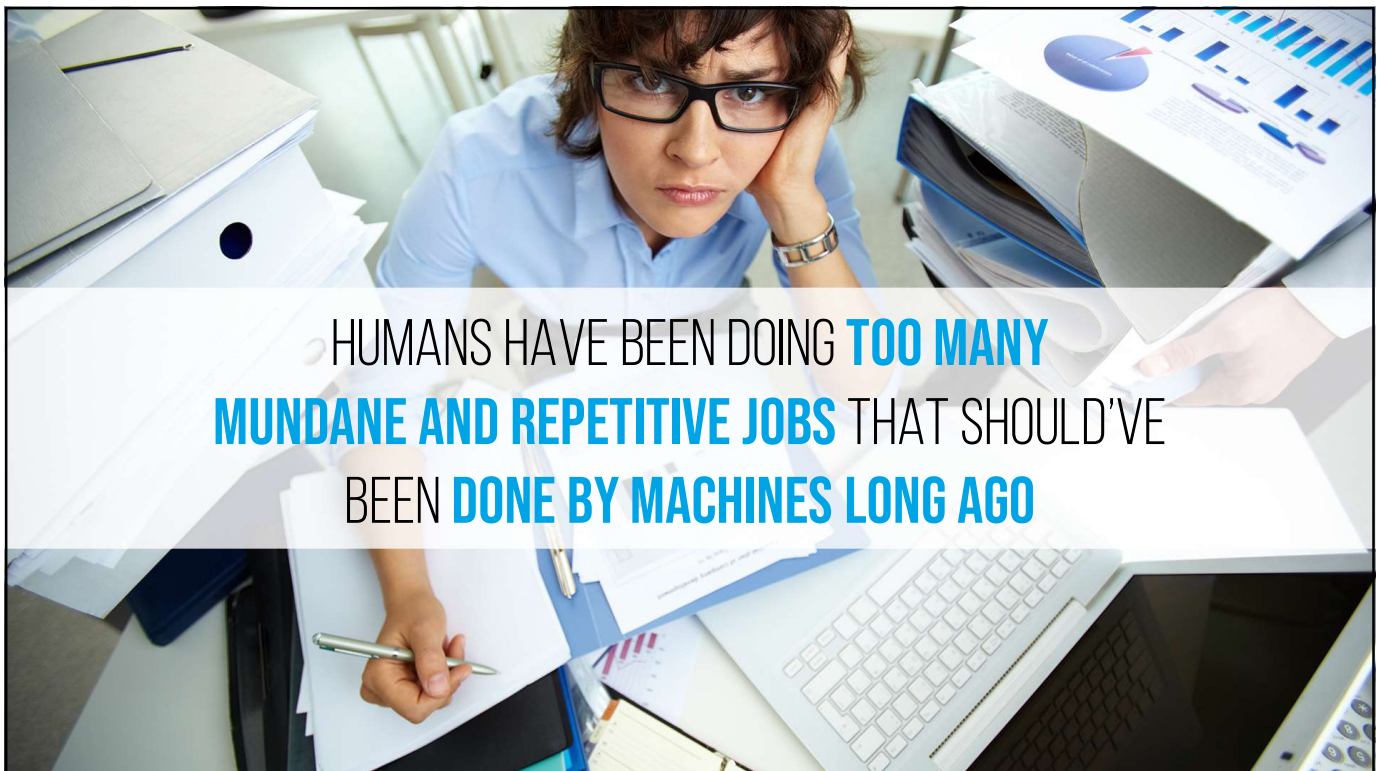


101



AI IS YOUR BEST **COPILOT**, HELPING YOU DO **MORE** AND DO **BETTER**

109



HUMANS HAVE BEEN DOING **TOO MANY MUNDANE AND REPETITIVE JOBS** THAT SHOULD'VE BEEN **DONE BY MACHINES LONG AGO**

110

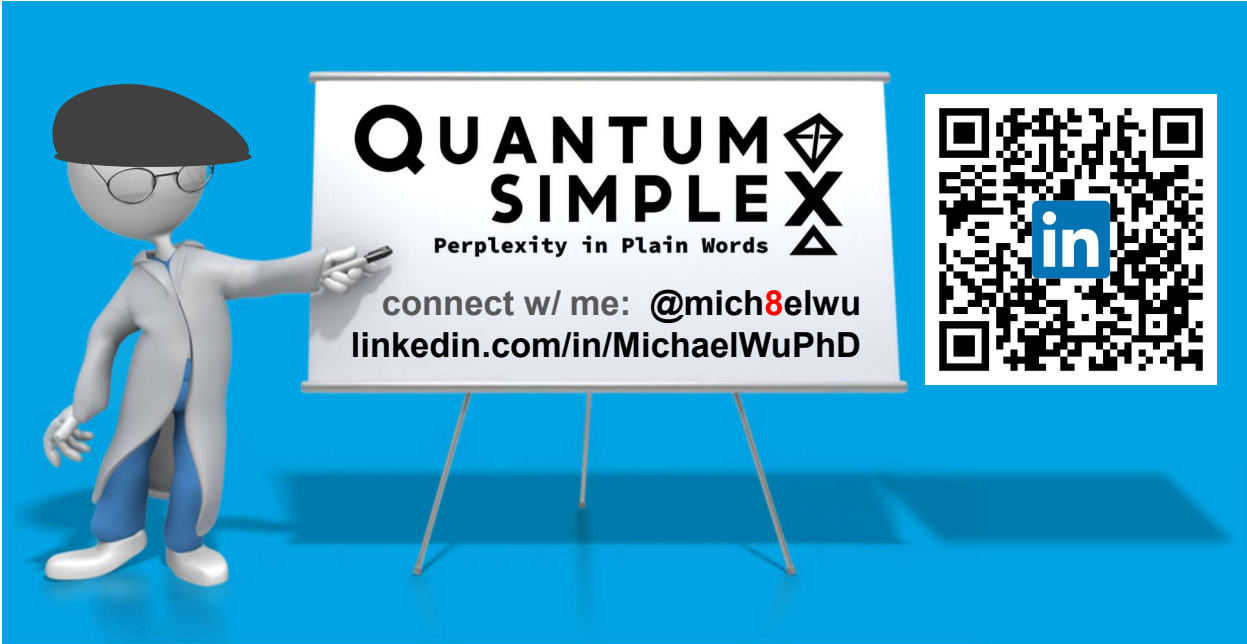


AI AUTOMATION WILL ALLOW US, HUMANS,
TO DO **WHAT WE LOVE AND WHAT WE DO BEST.**

111



113



QUANTUM SIMPLEX
Perplexity in Plain Words

connect w/ me: @mich8elwu
linkedin.com/in/MichaelWuPhD

PROS ©2023 PROS, Inc. All rights reserved. Confidential and Proprietary.

twitter: @mich8elwu
linkedin.com/in/MichaelWuPhD