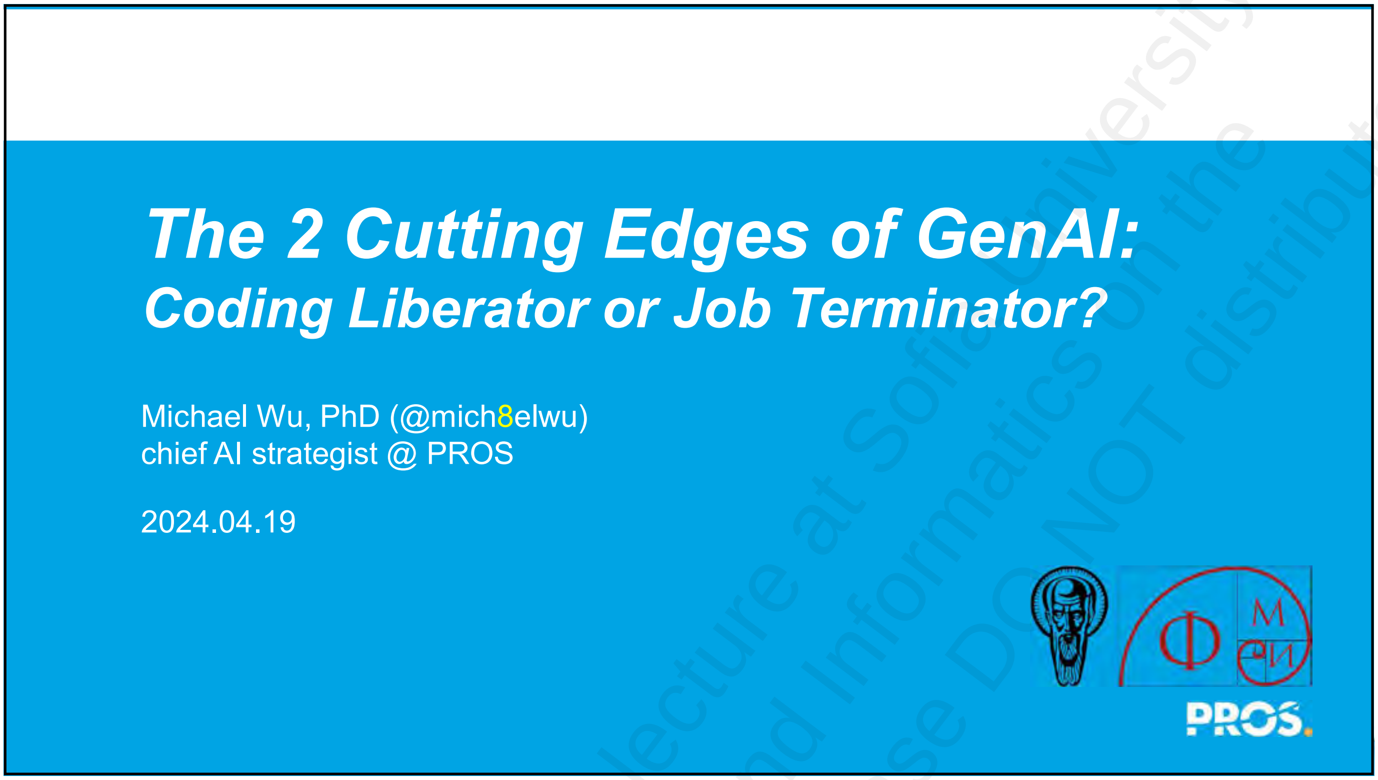


The 2 Cutting Edges of GenAI:
Coding Liberator or Job Terminator?

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

2024.04.19

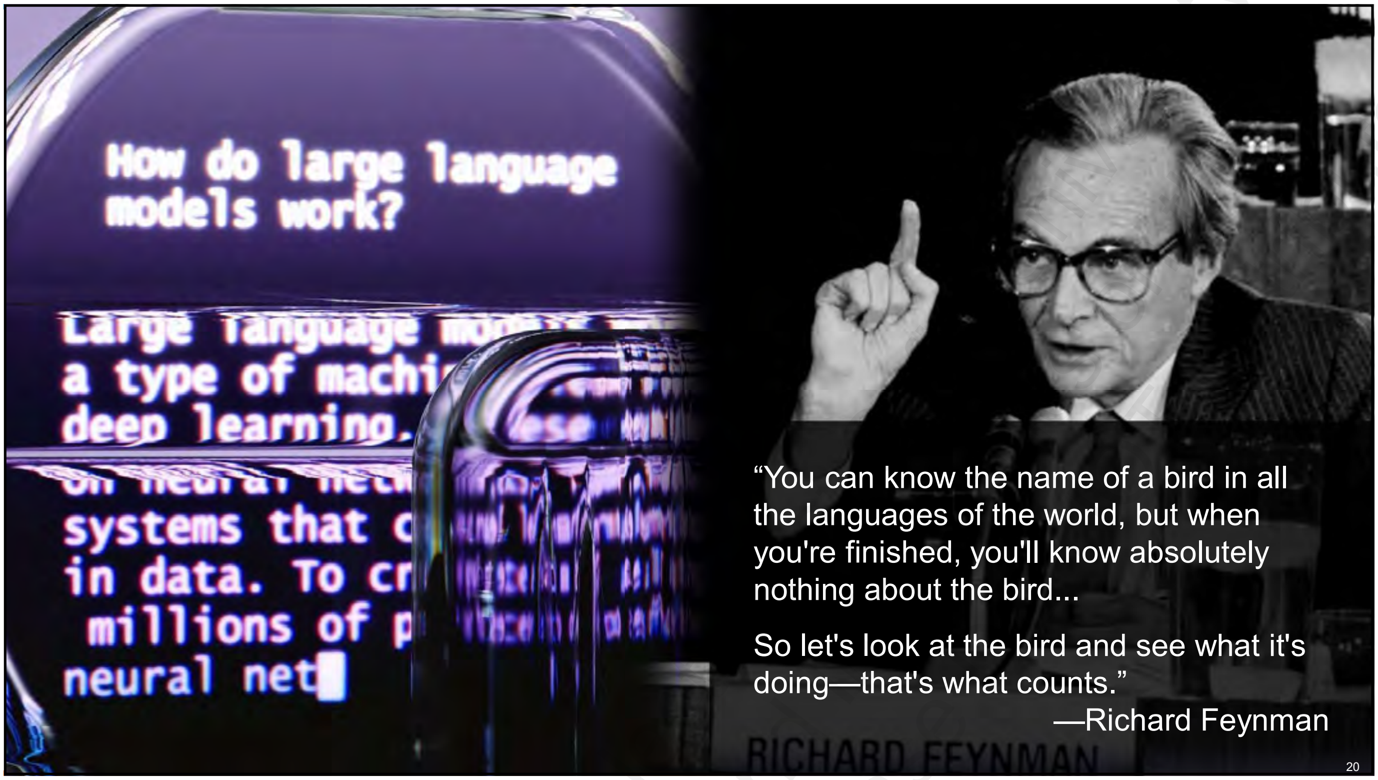
1

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

2024.04.19

12

12



“You can know the name of a bird in all the languages of the world, but when you're finished, you'll know absolutely nothing about the bird...

So let's look at the bird and see what it's doing—that's what counts.”

—Richard Feynman

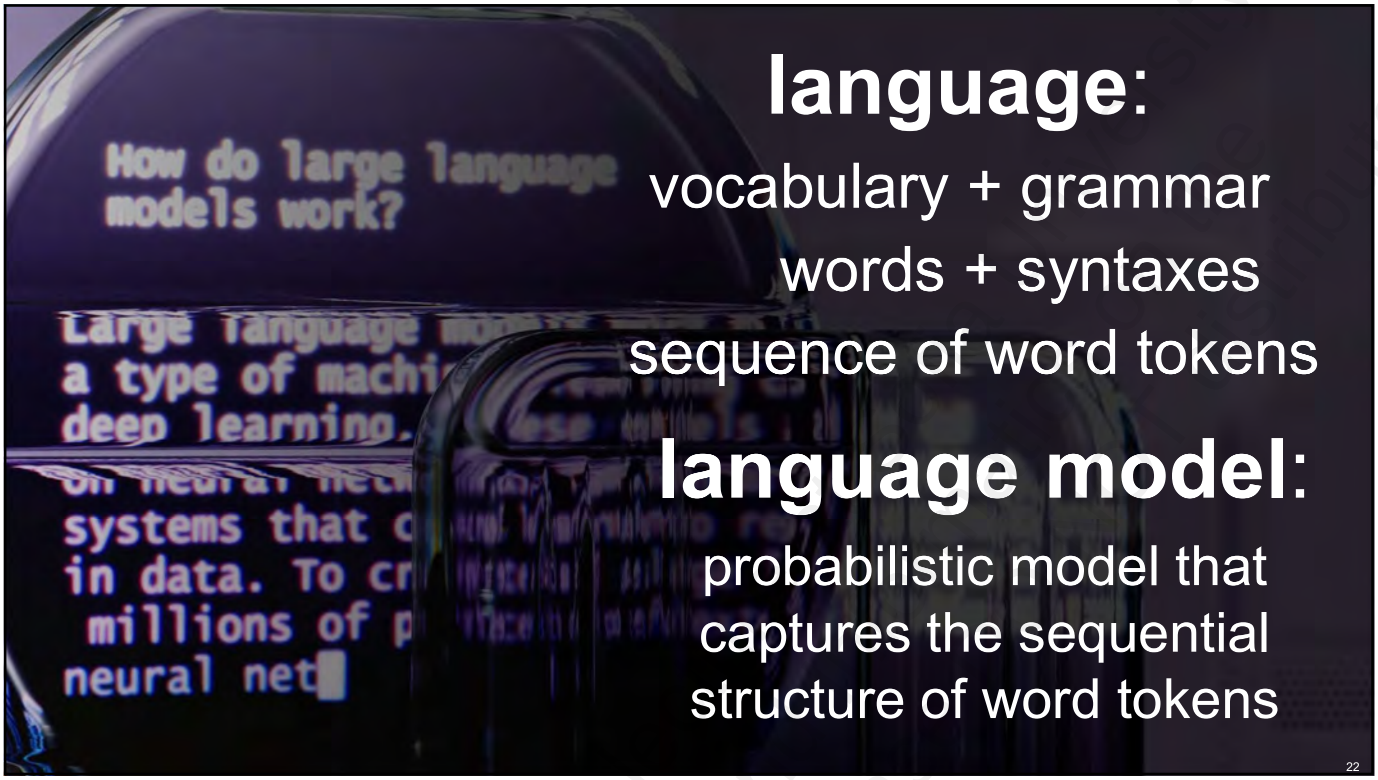
20

20

what’s a **language**?

21

21



language:

vocabulary + grammar

words + syntaxes

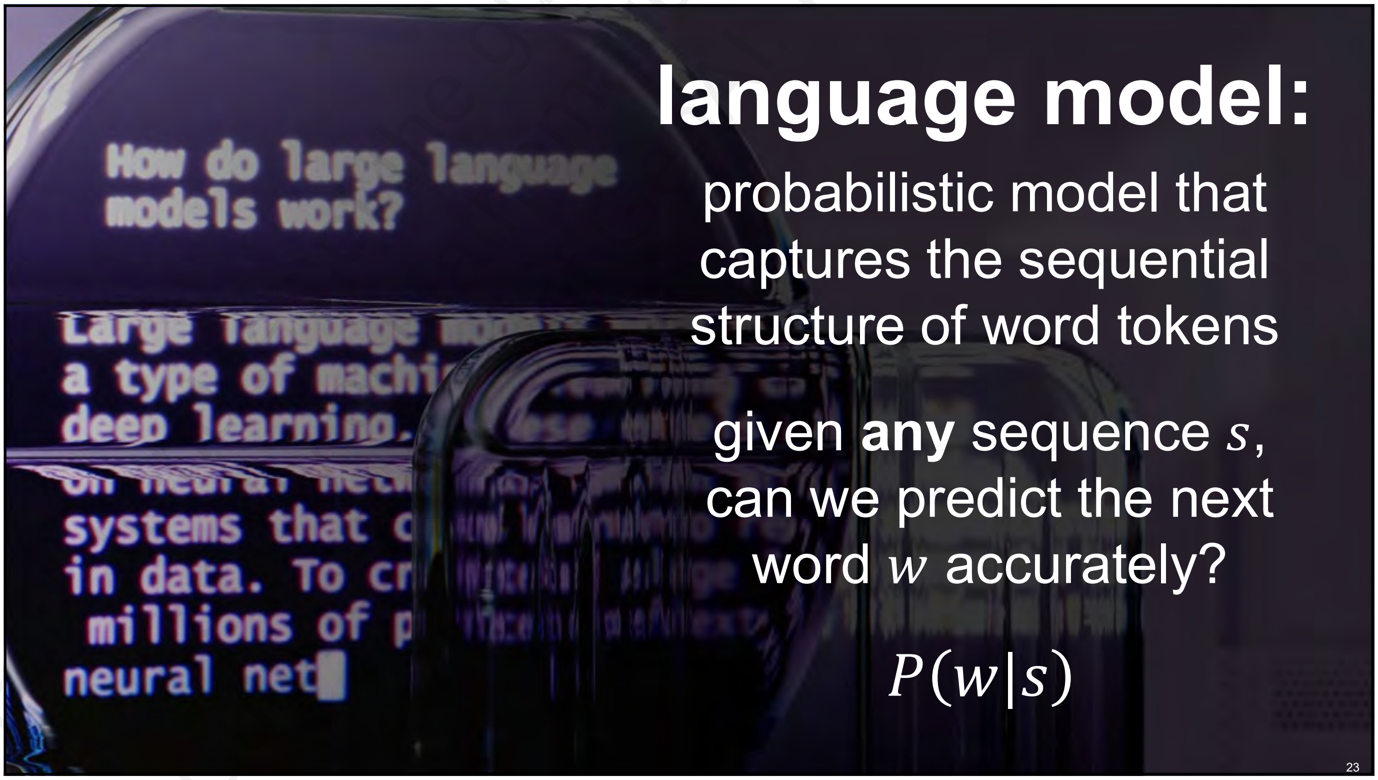
sequence of word tokens

language model:

probabilistic model that captures the sequential structure of word tokens

22

22



language model:

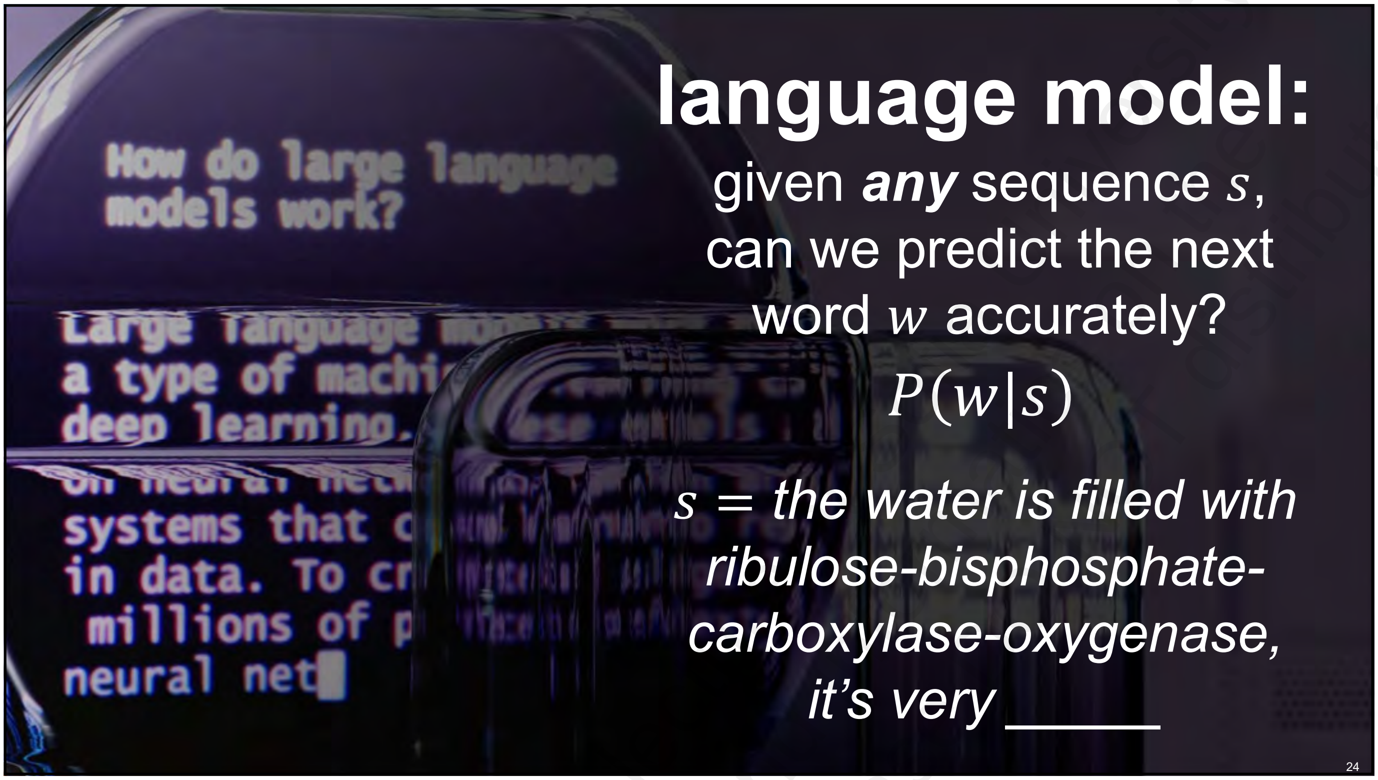
probabilistic model that captures the sequential structure of word tokens

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

$$P(w|s)$$

23

23



language model:

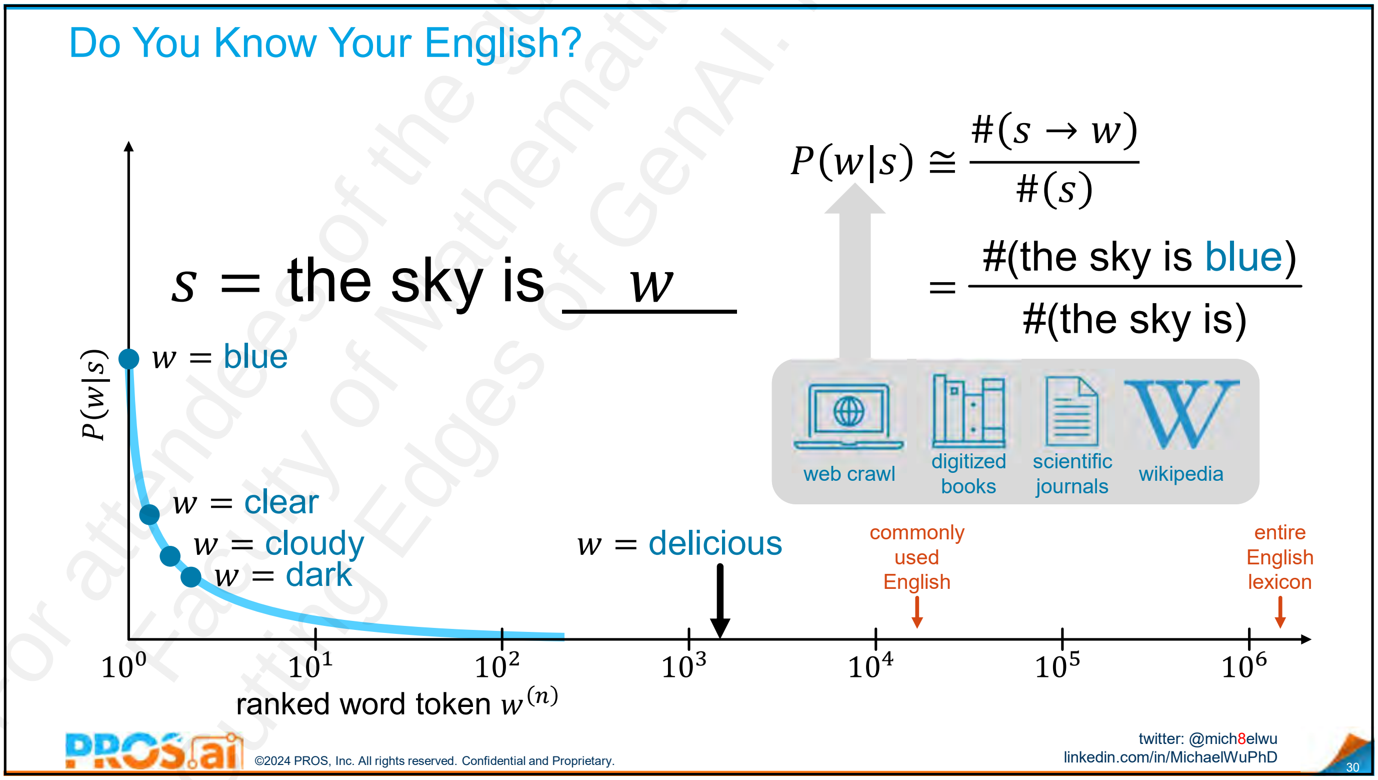
given ***any*** sequence s, can we predict the next word w accurately?

$$P(w|s)$$

s = the water is filled with ribulose-bisphosphate-carboxylase-oxygenase, it's very _____

24

24



Do You Know Your English?

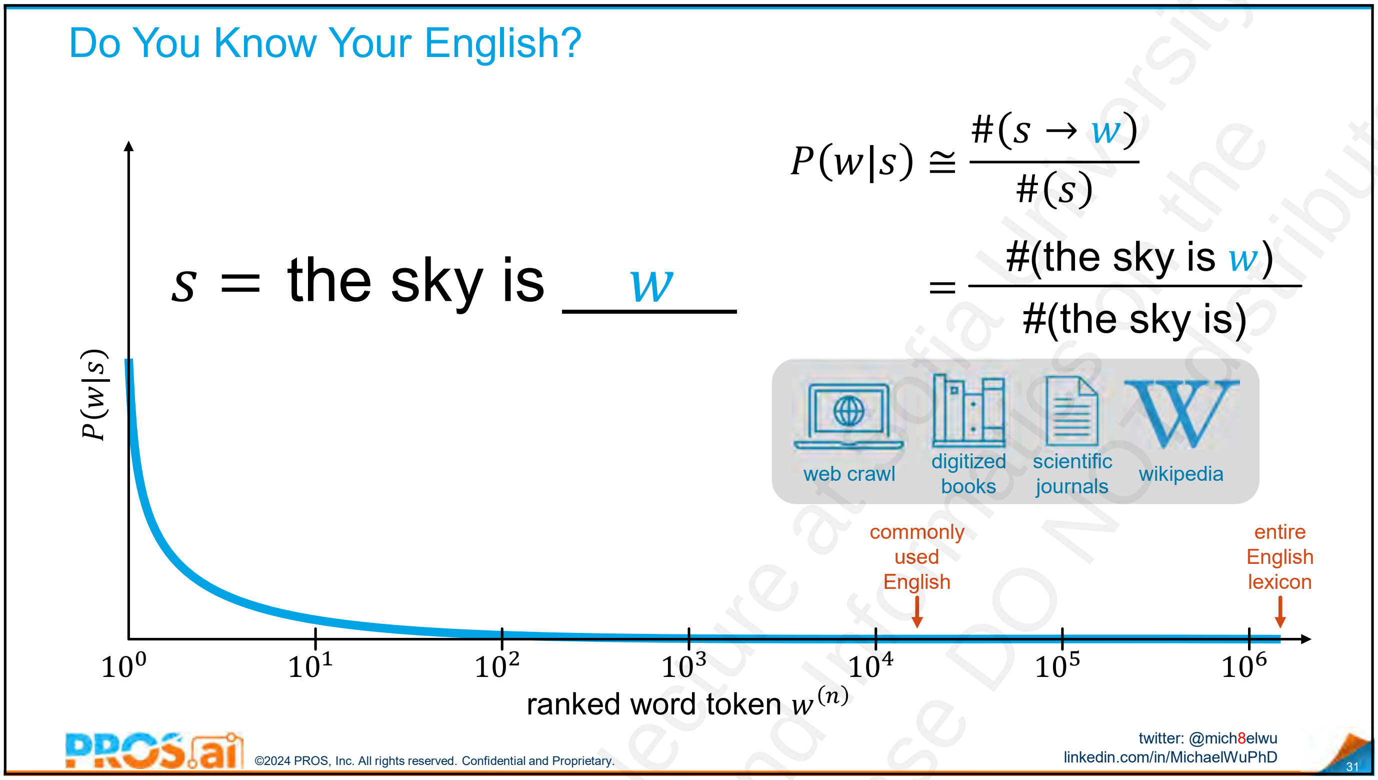
s = the sky is __w__

$$P(w|s) \cong \frac{\#(s \rightarrow w)}{\#(s)} = \frac{\#(\text{the sky is blue})}{\#(\text{the sky is})}$$

twitter: @mich8elwu
linkedin.com/in/MichaelWuPhD

30

30

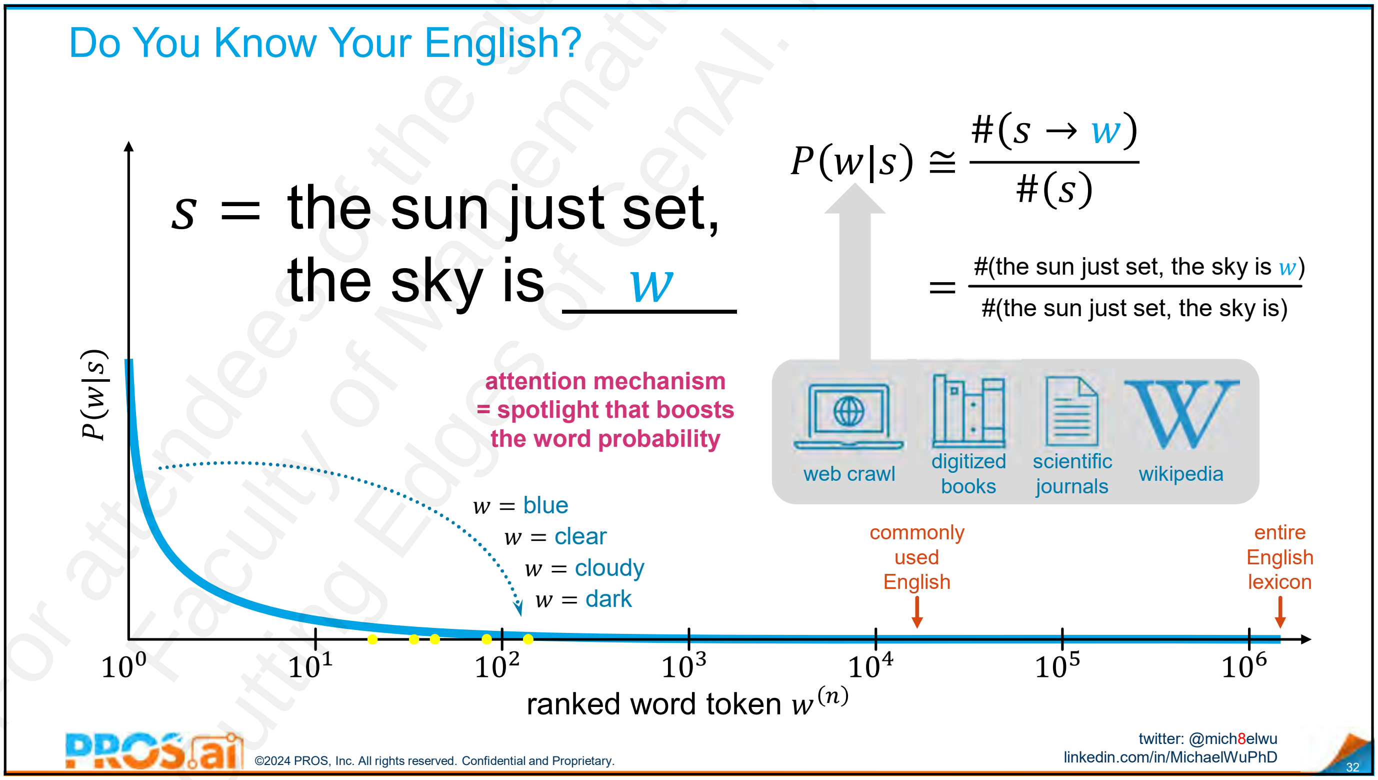


Do You Know Your English?

$s =$ the sky is ___w___

$$P(w|s) \cong \frac{\#(s \to w)}{\#(s)} = \frac{\#(\text{the sky is } w)}{\#(\text{the sky is})}$$

twitter: @mich8elwu
linkedin.com/in/MichaelWuPhD



Do You Know Your English?

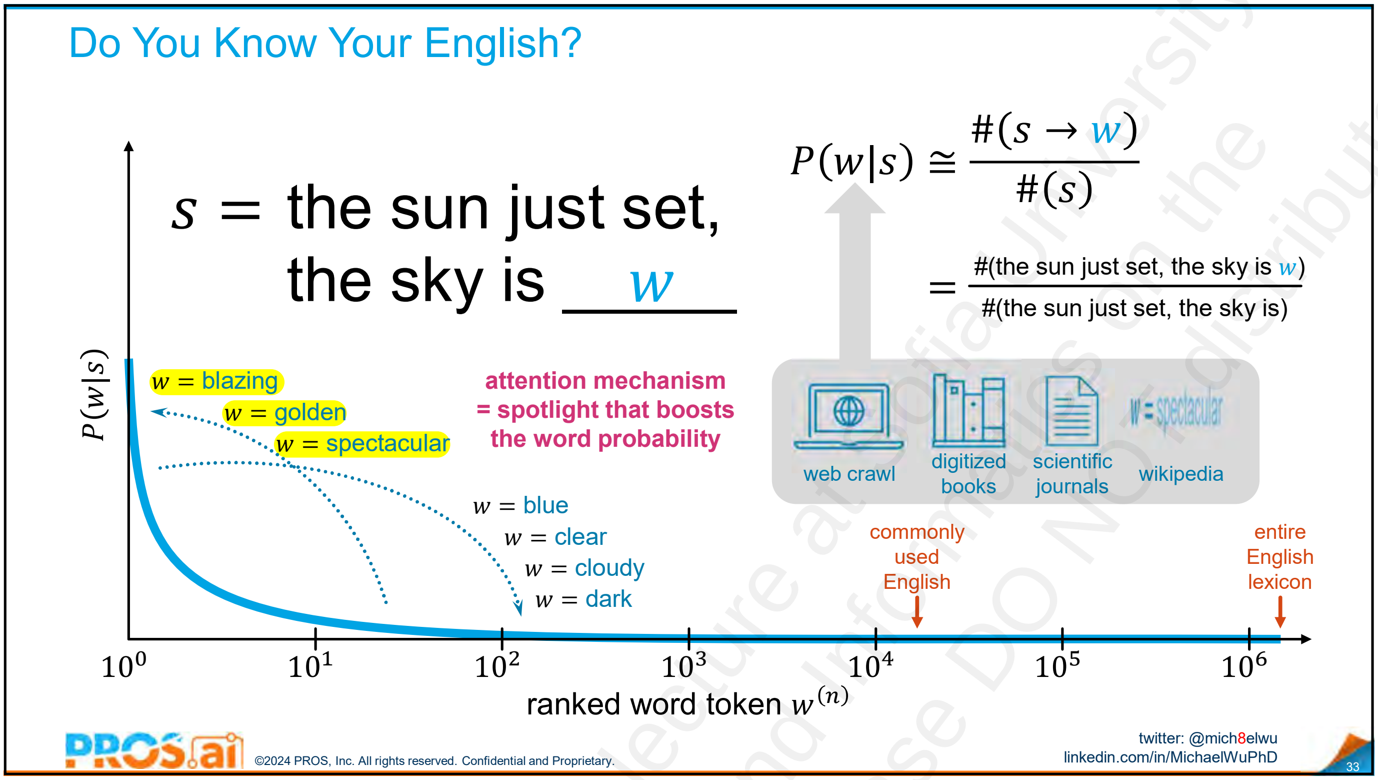
$s =$ the sun just set,
the sky is ___w___

$$P(w|s) \cong \frac{\#(s \to w)}{\#(s)} = \frac{\#(\text{the sun just set, the sky is } w)}{\#(\text{the sun just set, the sky is})}$$

attention mechanism
= spotlight that boosts
the word probability

$w =$ blue
$w =$ clear
$w =$ cloudy
$w =$ dark

twitter: @mich8elwu
linkedin.com/in/MichaelWuPhD



Do You Know Your English?

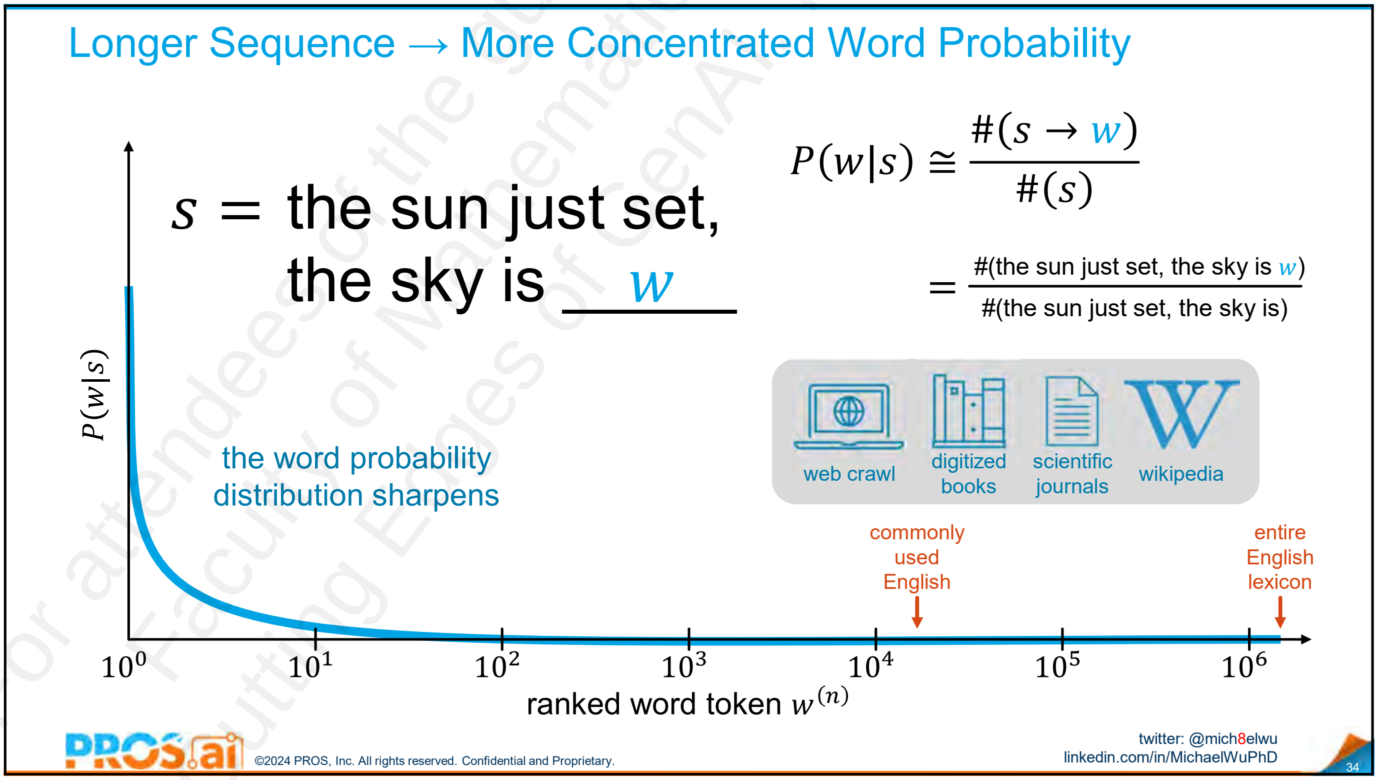
$s =$ the sun just set, the sky is $\underline{\quad w \quad}$

$$P(w|s) \cong \frac{\#(s \to w)}{\#(s)} = \frac{\#(\text{the sun just set, the sky is } w)}{\#(\text{the sun just set, the sky is})}$$

attention mechanism = spotlight that boosts the word probability

twitter: @mich8elwu
linkedin.com/in/MichaelWuPhD

33



Longer Sequence → More Concentrated Word Probability

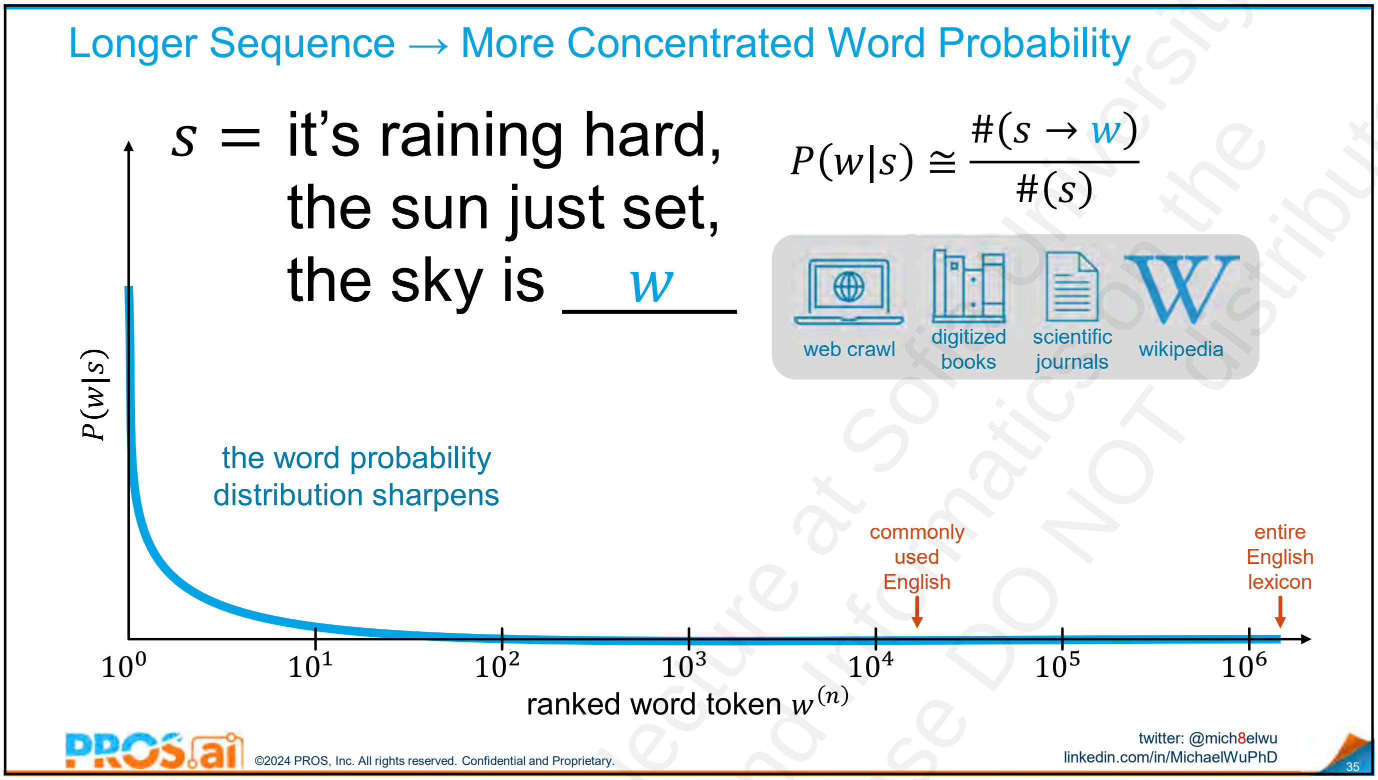
$s =$ the sun just set, the sky is $\underline{\quad w \quad}$

$$P(w|s) \cong \frac{\#(s \to w)}{\#(s)} = \frac{\#(\text{the sun just set, the sky is } w)}{\#(\text{the sun just set, the sky is})}$$

the word probability distribution sharpens

twitter: @mich8elwu
linkedin.com/in/MichaelWuPhD

34

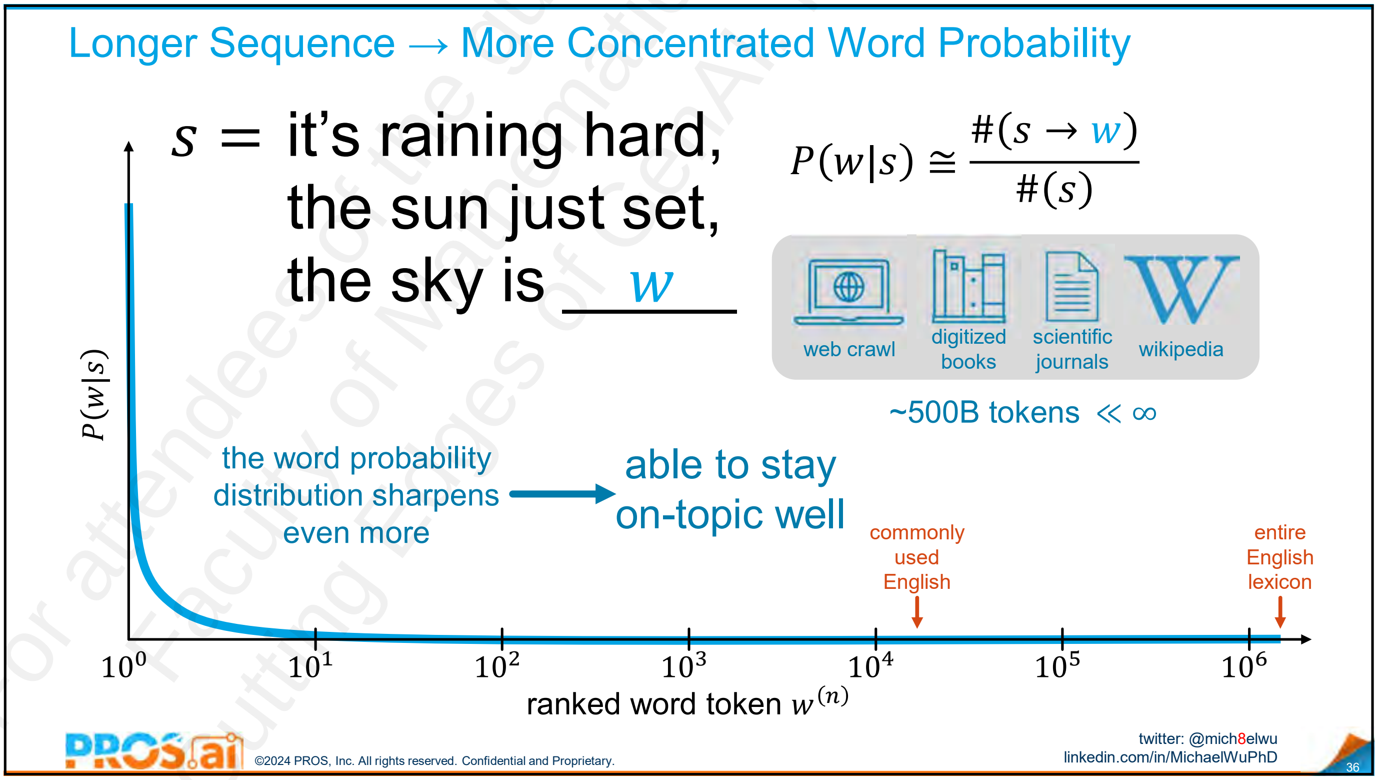


Longer Sequence → More Concentrated Word Probability

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

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

twitter: @mich8elwu
linkedin.com/in/MichaelWuPhD



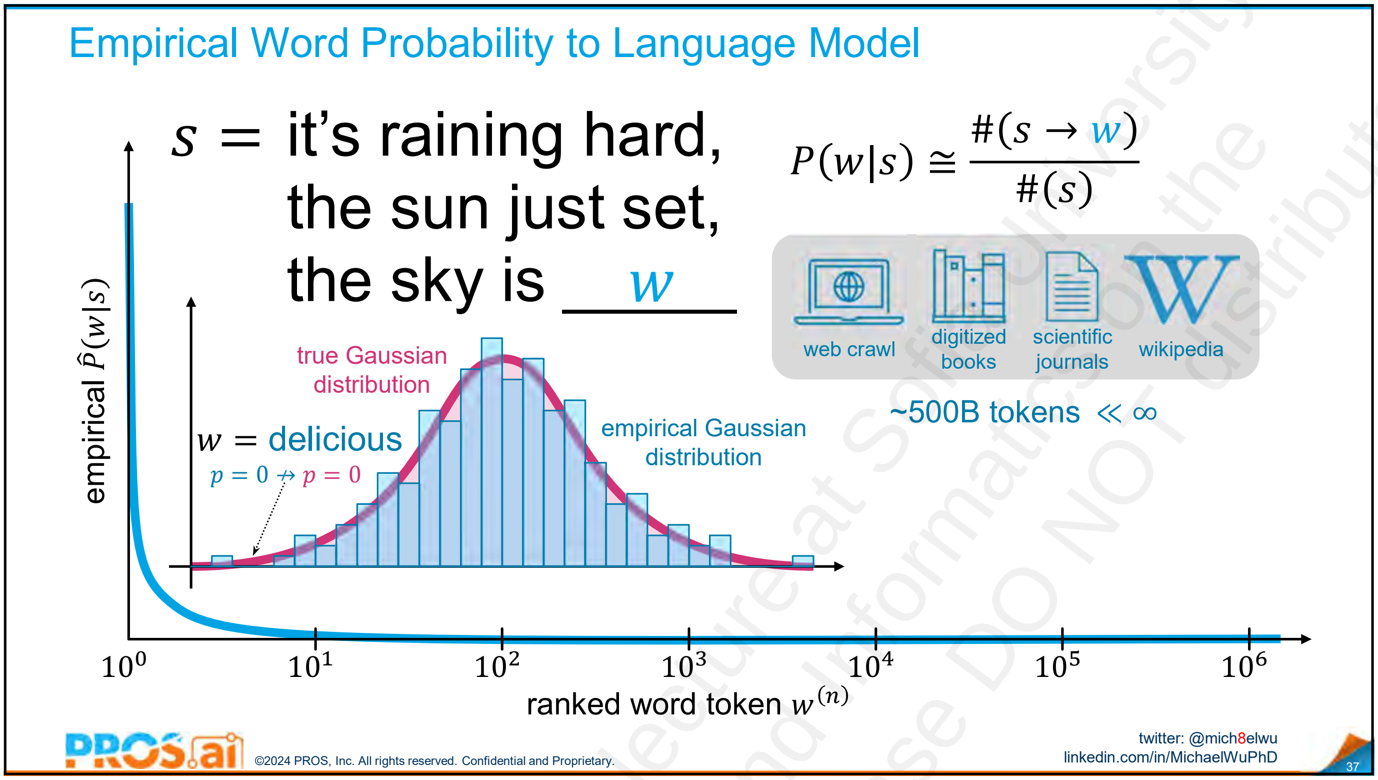
Longer Sequence → More Concentrated Word Probability

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

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

~500B tokens $\ll \infty$

twitter: @mich8elwu
linkedin.com/in/MichaelWuPhD



Empirical Word Probability to Language Model

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

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

web crawl, digitized books, scientific journals, wikipedia

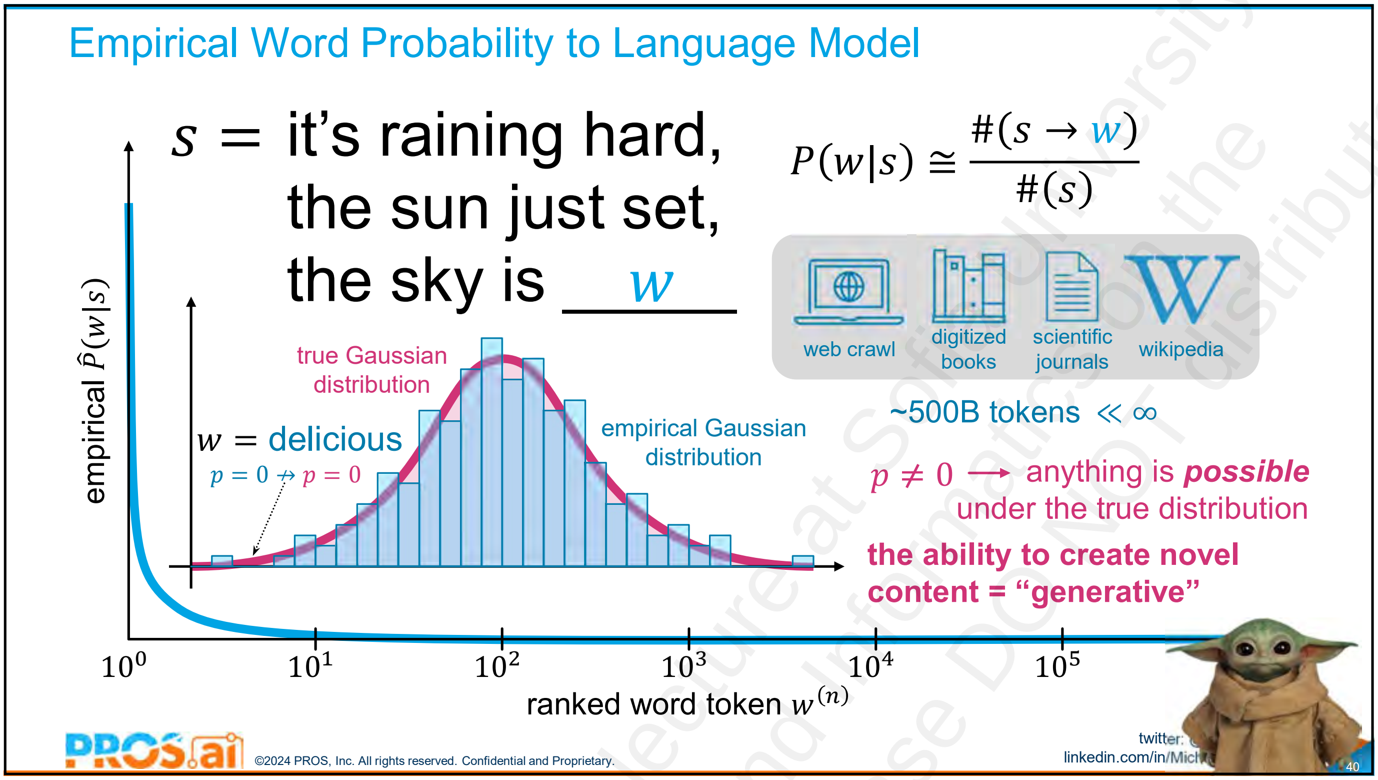
~500B tokens $\ll \infty$

PROS.ai

twitter: @mich8elwu
linkedin.com/in/MichaelWuPhD

37

38



Empirical Word Probability to Language Model

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

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

web crawl · digitized books · scientific journals · wikipedia

~500B tokens $\ll \infty$

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

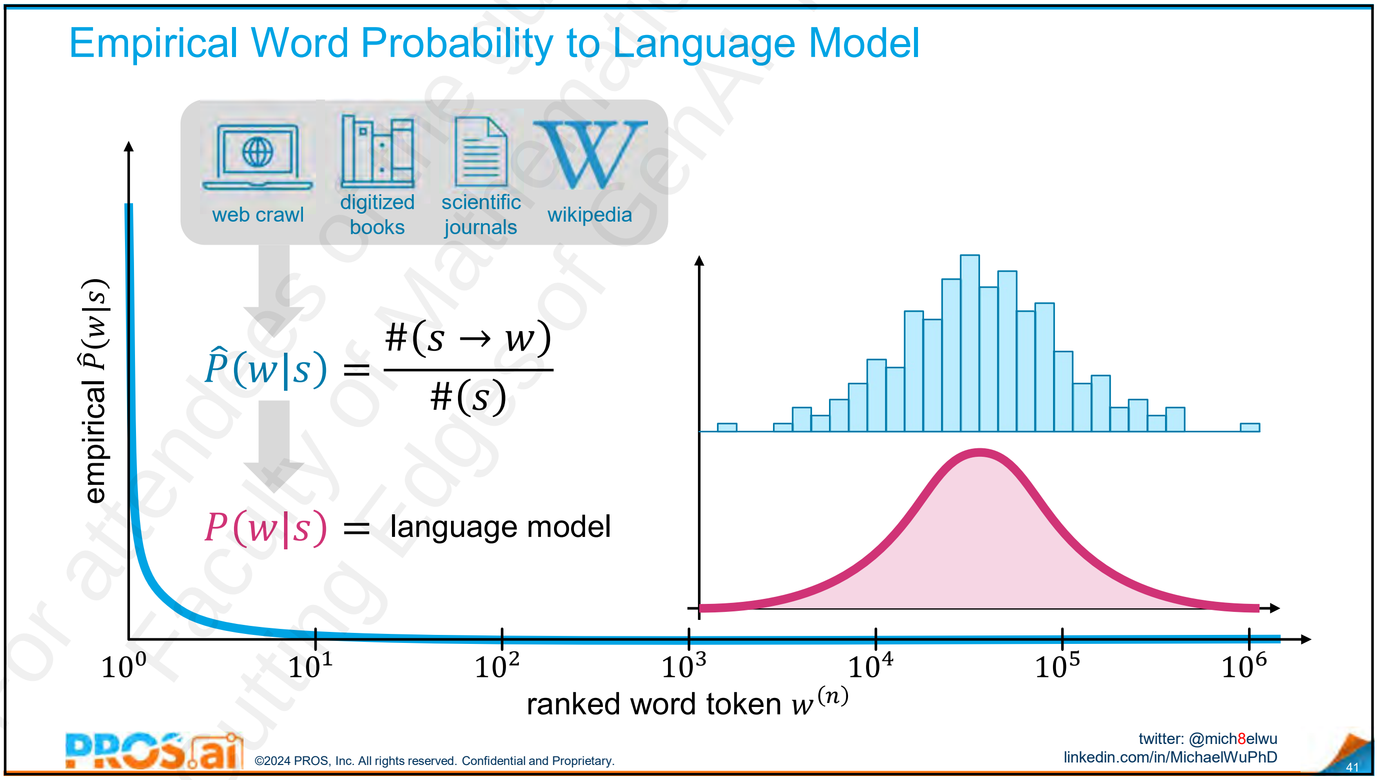
the ability to create novel content = "generative"

PROS.ai ©2024 PROS, Inc. All rights reserved. Confidential and Proprietary.

twitter: @
linkedin.com/in/Mich

40

40



Empirical Word Probability to Language Model

web crawl · digitized books · scientific journals · wikipedia

$$\hat{P}(w|s) = \frac{\#(s \to w)}{\#(s)}$$

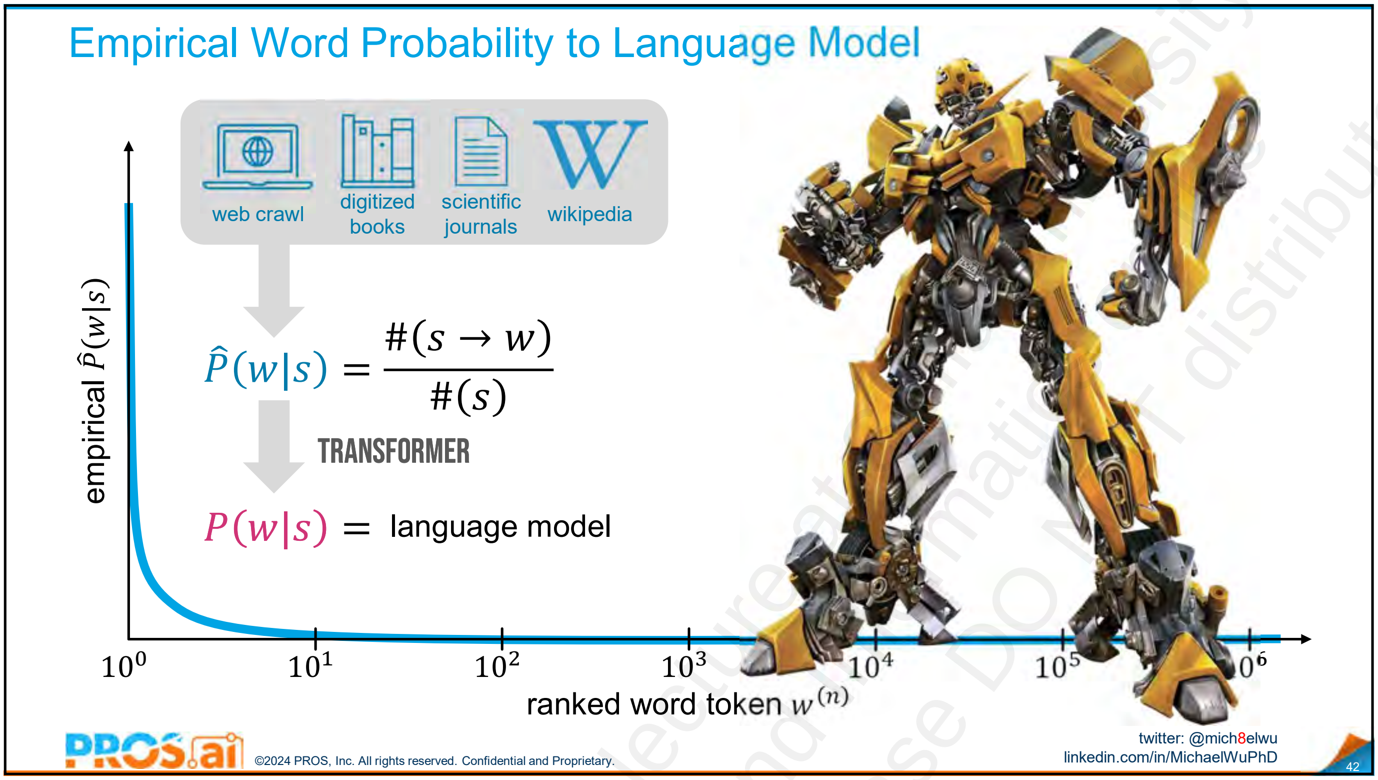
$P(w|s) =$ language model

PROS.ai ©2024 PROS, Inc. All rights reserved. Confidential and Proprietary.

twitter: @mich8elwu
linkedin.com/in/MichaelWuPhD

41

41



Empirical Word Probability to Language Model

web crawl · digitized books · scientific journals · wikipedia

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

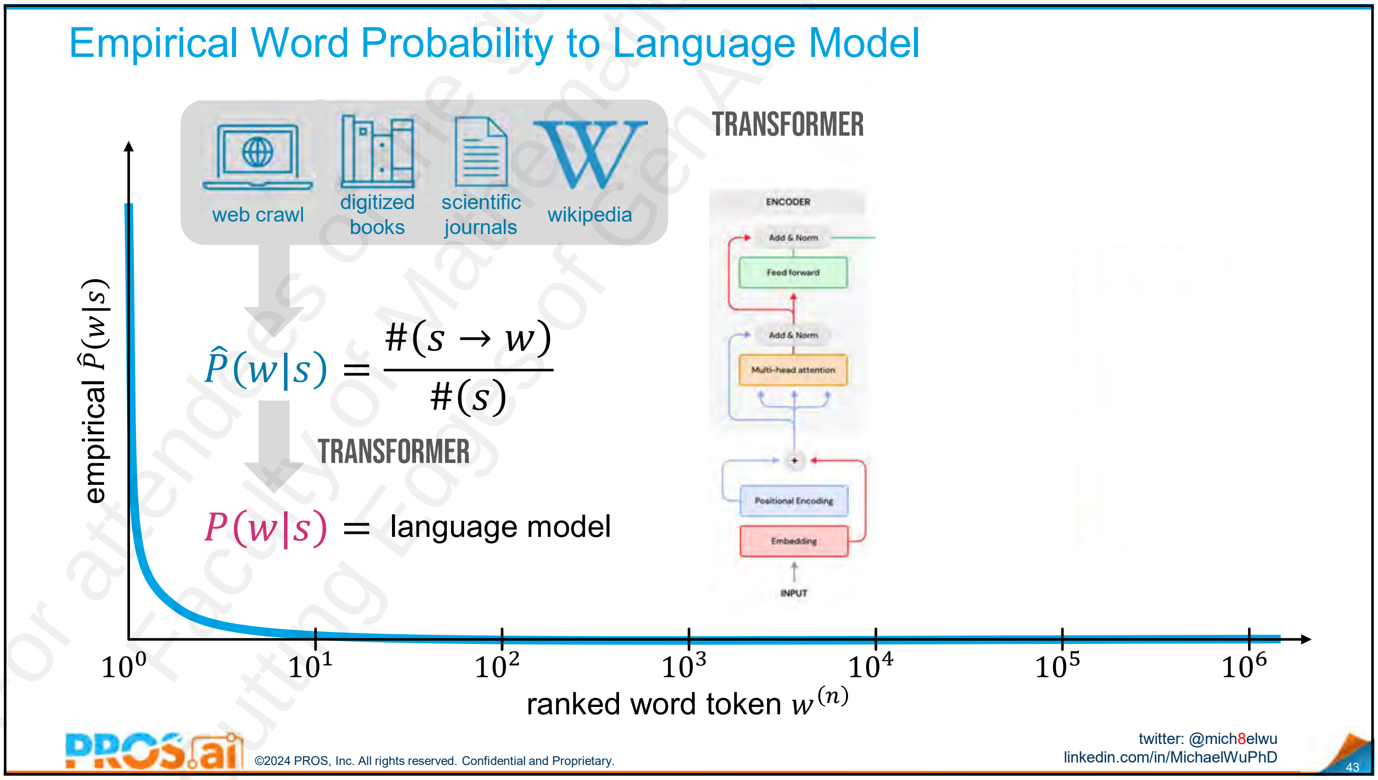
TRANSFORMER

$P(w|s) =$ language model

©2024 PROS, Inc. All rights reserved. Confidential and Proprietary.

twitter: @mich8elwu
linkedin.com/in/MichaelWuPhD

42



Empirical Word Probability to Language Model

web crawl · digitized books · scientific journals · wikipedia

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

TRANSFORMER

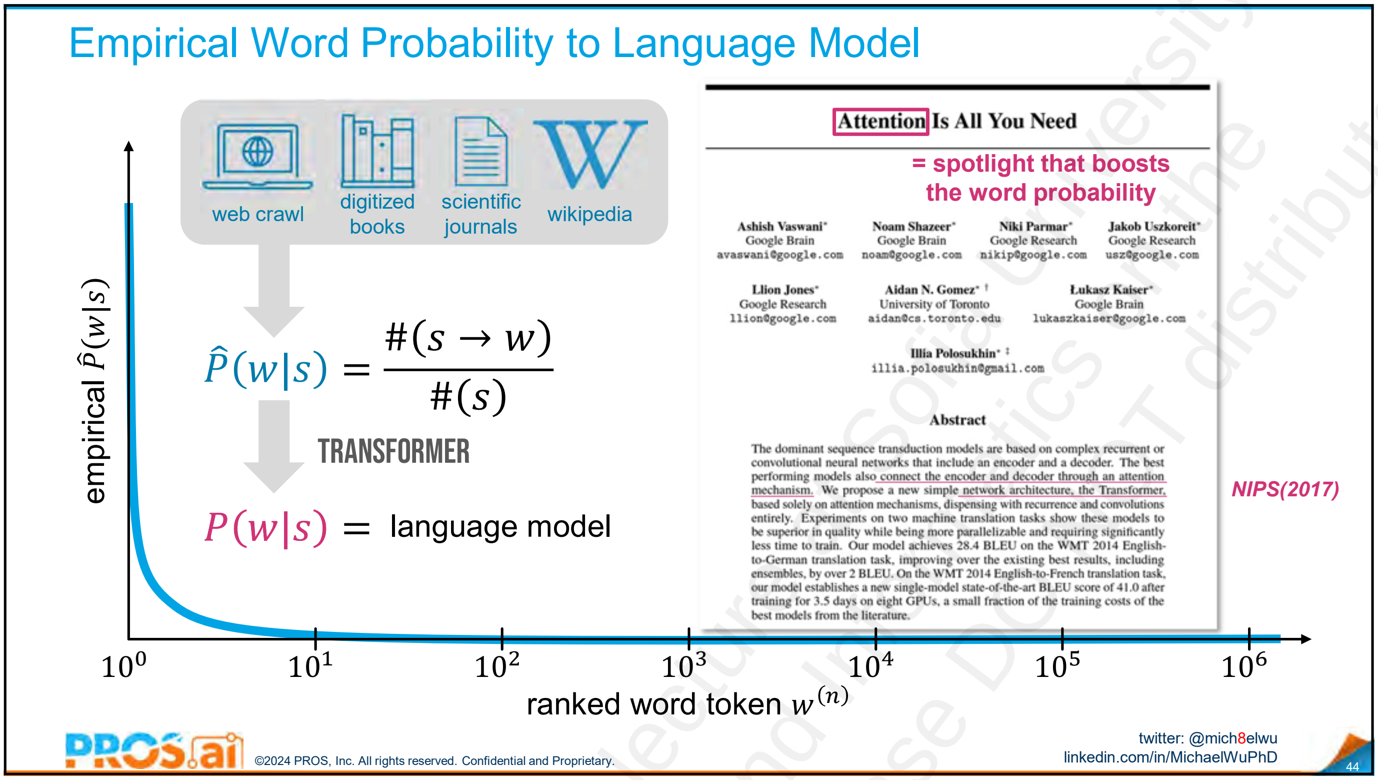
$P(w|s) =$ language model

TRANSFORMER

©2024 PROS, Inc. All rights reserved. Confidential and Proprietary.

twitter: @mich8elwu
linkedin.com/in/MichaelWuPhD

43



Empirical Word Probability to Language Model

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

TRANSFORMER

$P(w|s) =$ 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
Łukasz Kaiser* Google Brain lukaszkaiser@google.com
Illia Polosukhin* ‡ illia.polosukhin@gmail.com

Abstract

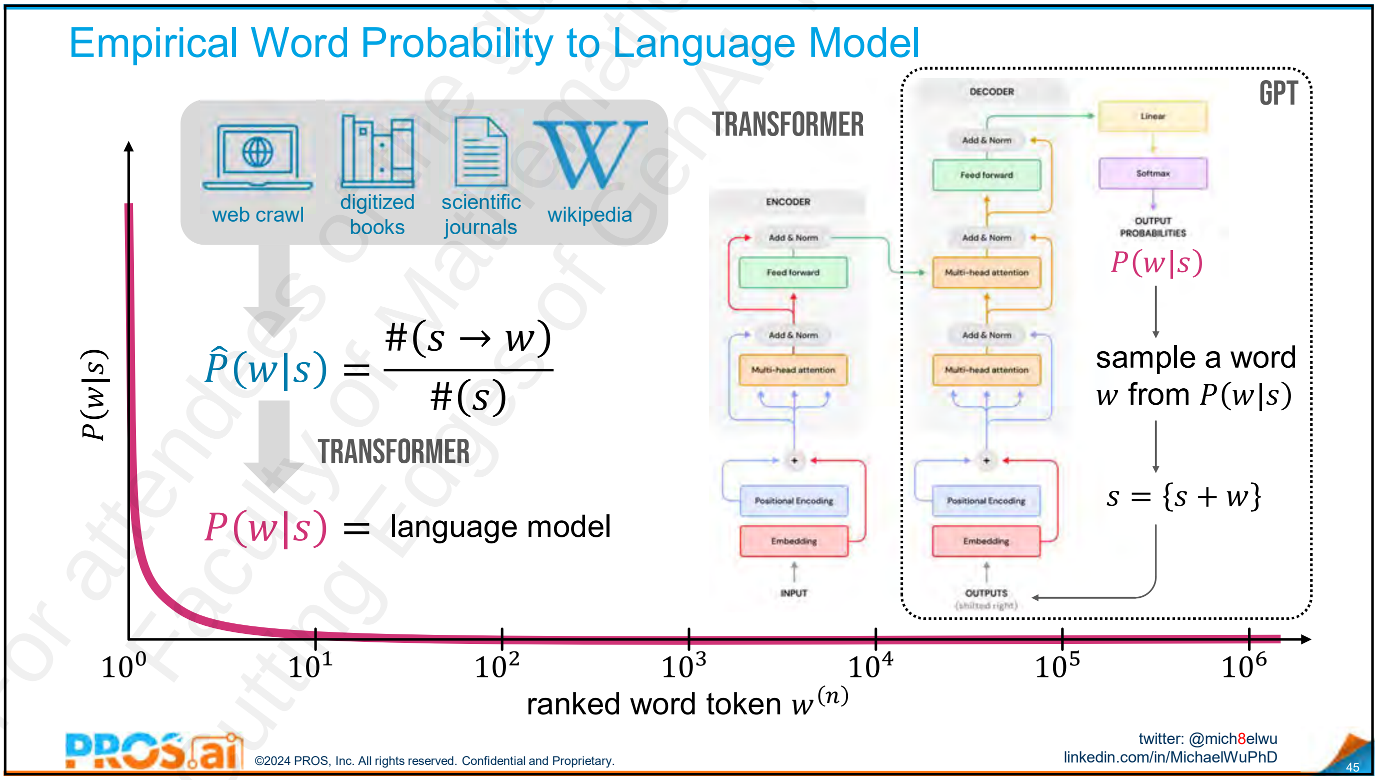
The dominant sequence transduction models are based on complex recurrent or convolutional neural networks that include an encoder and a decoder. The best performing models also connect the encoder and decoder through an attention mechanism. We propose a new simple network architecture, the Transformer, 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)

twitter: @mich8elwu
linkedin.com/in/MichaelWuPhD

44

44



Empirical Word Probability to Language Model

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

TRANSFORMER

$P(w|s) =$ language model

$P(w|s)$

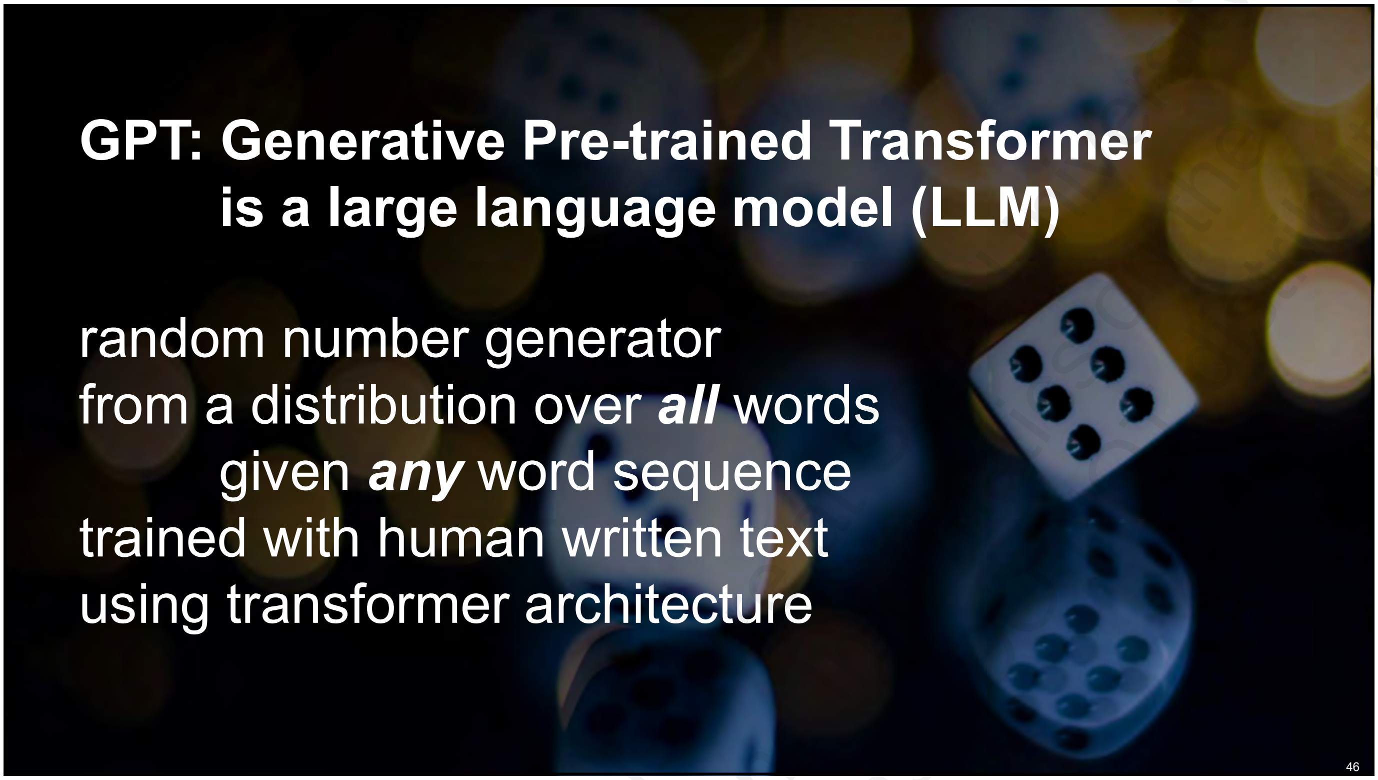
sample a word w from $P(w|s)$

$s = \{s + w\}$

twitter: @mich8elwu
linkedin.com/in/MichaelWuPhD

45

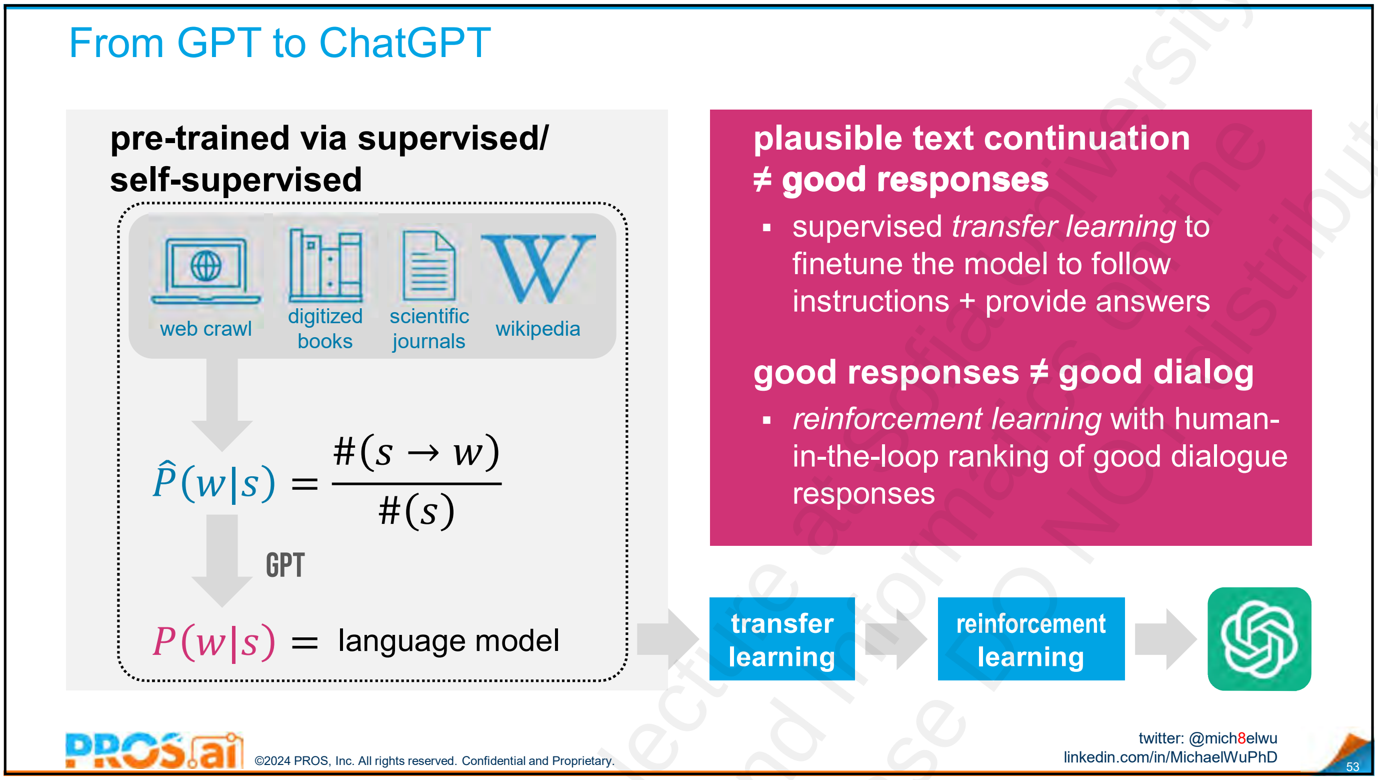
45



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

doesn’t sound
very intelligent



From GPT to ChatGPT

pre-trained via supervised/ self-supervised

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

$P(w|s) =$ language model

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

twitter: @mich8elwu
linkedin.com/in/MichaelWuPhD

53

53

ChatGPT has no knowledge of the real world

54

54

Created with the help of Beautiful.ai

"***generative***" AI are, by definition and construction, ***hallucinatory***

55

55

very convincing legal briefings citing non-existent court cases

56

56

Steven Schwartz

+ Peter LoDuca fined $5K

57

hallucination:
a feature
or a bug?

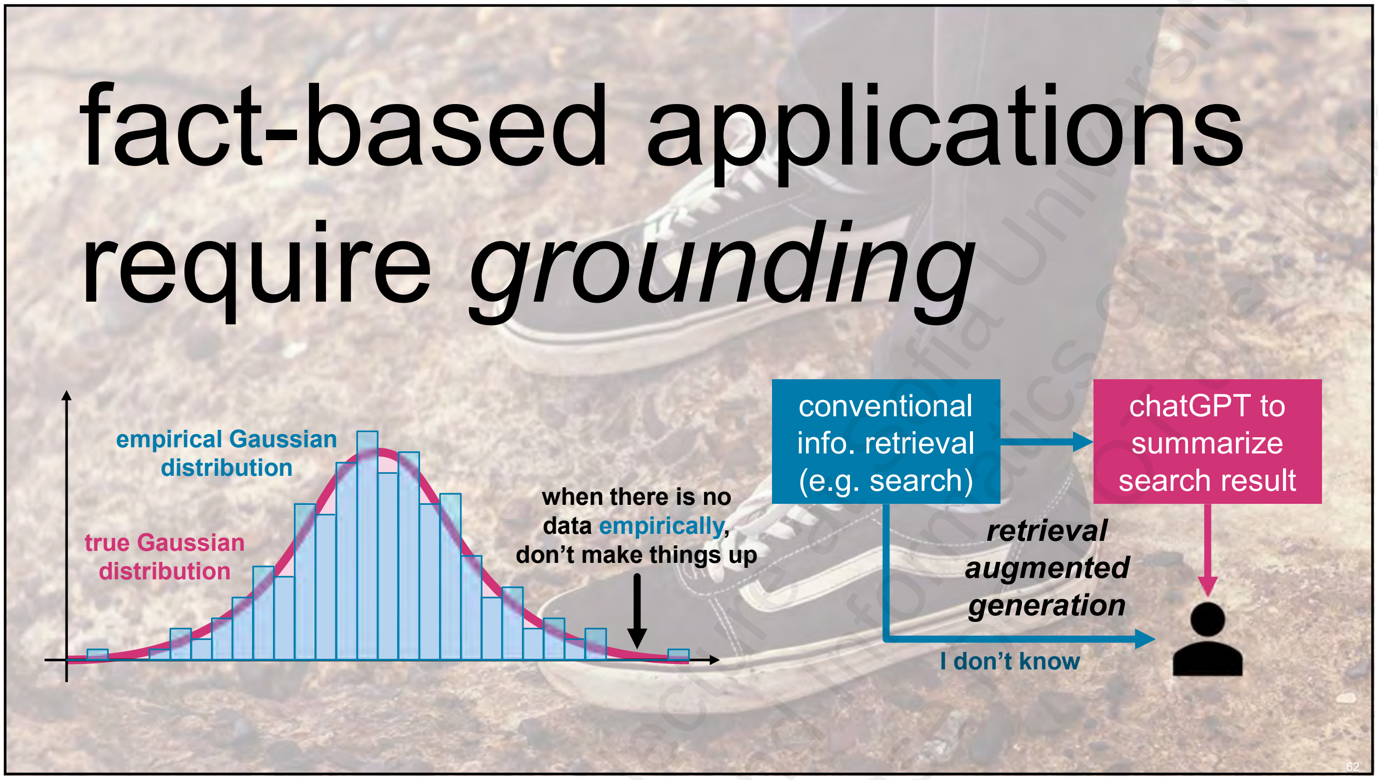
59

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

60

bug: for *fact-based* applications

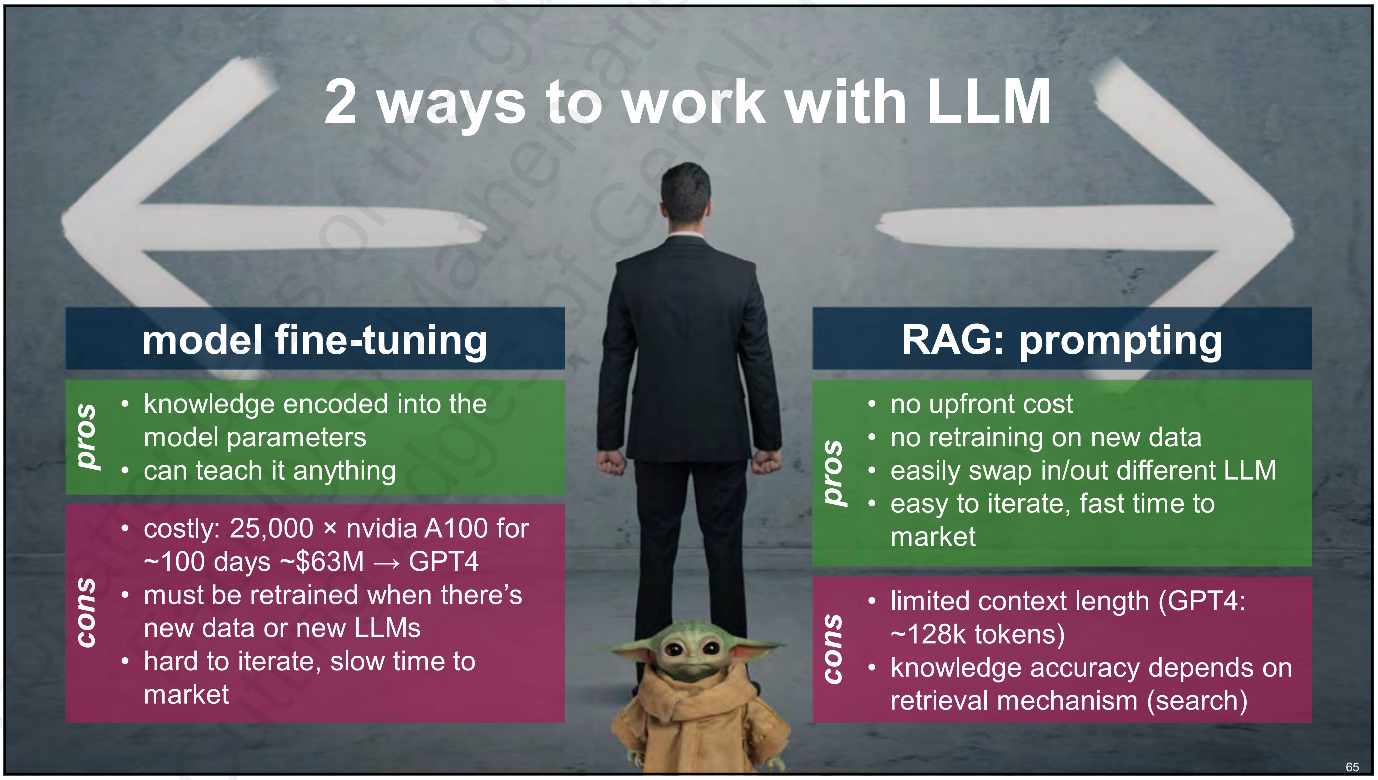
61



fact-based applications require *grounding*

empirical Gaussian distribution
true Gaussian distribution
when there is no data empirically, don't make things up
conventional info. retrieval (e.g. search)
chatGPT to summarize search result
retrieval augmented generation
I don't know
62

62



2 ways to work with LLM

model fine-tuning

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

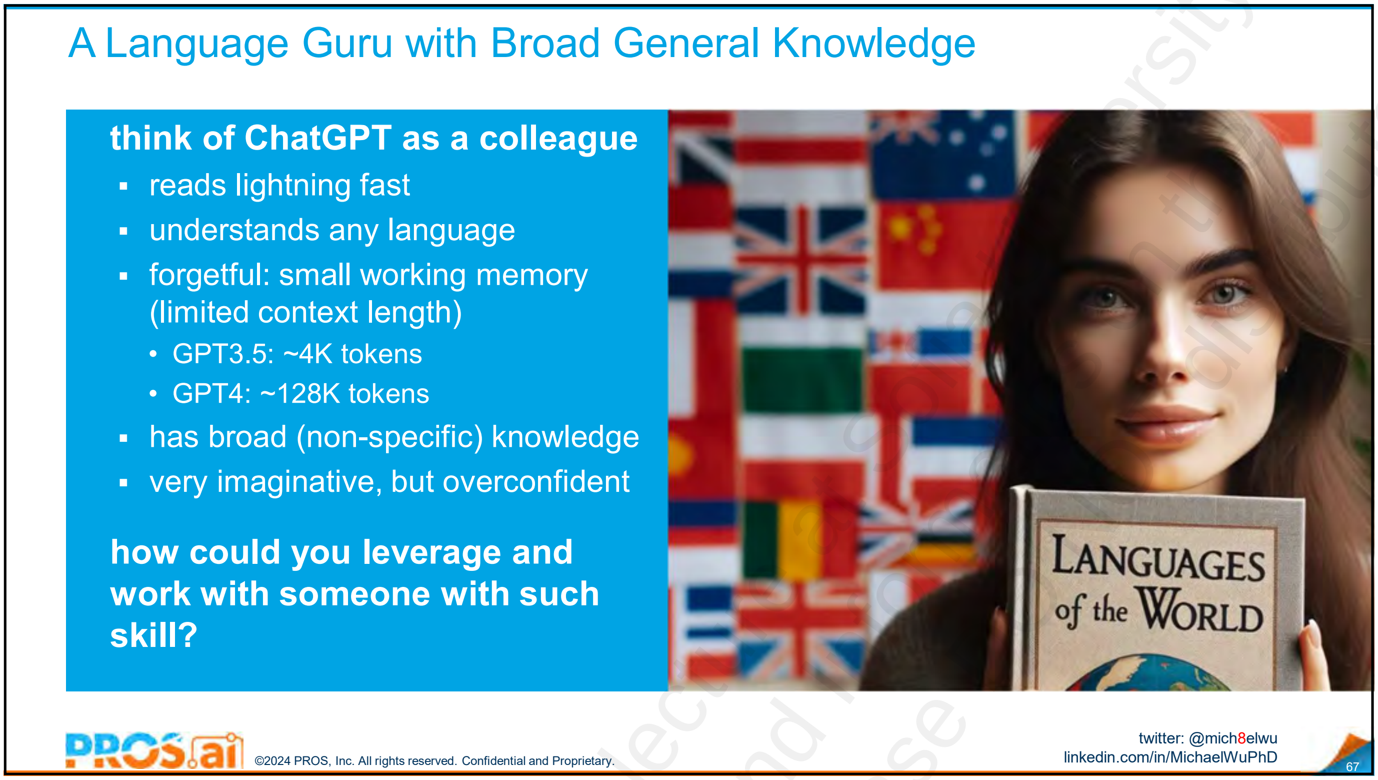
cons
- costly: 25,000 × nvidia A100 for ~100 days ~$63M → GPT4
- must be retrained when there's new data or new LLMs
- hard to iterate, slow time to market

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: ~128k tokens)
- knowledge accuracy depends on retrieval mechanism (search)

65



A Language Guru with Broad General Knowledge

think of ChatGPT as a colleague

- reads lightning fast
- understands any language
- forgetful: small working memory (limited context length)
  - GPT3.5: ~4K tokens
  - GPT4: ~128K tokens
- has broad (non-specific) knowledge
- very imaginative, but overconfident

how could you leverage and work with someone with such skill?

LANGUAGES of the WORLD

PROS.ai

twitter: @mich8elwu
linkedin.com/in/MichaelWuPhD

67

67

68

Beyond ChatGPT

	generic generative AI						specialized generative AI		
	textual		visual		audio		game	specialized design	
data	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	voicebox	Riffusion	nvidia Get3D	RoseTTAFold	
	Mistral	tabnine	Craiyon	Imagen Video		dance diffusion	human MDM		
	Claude	stability.ai	Midjourney	Sora		musicLM			
	LaMDA	CodeWhisperer	stable diffusion						
	Gemini		Imagen						
	Perplexity		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	voice cloning				
	compare/contrast	text to SQL	2D design						
	content creation	web app builder	social media						
	question/answer								
	realtime translation								

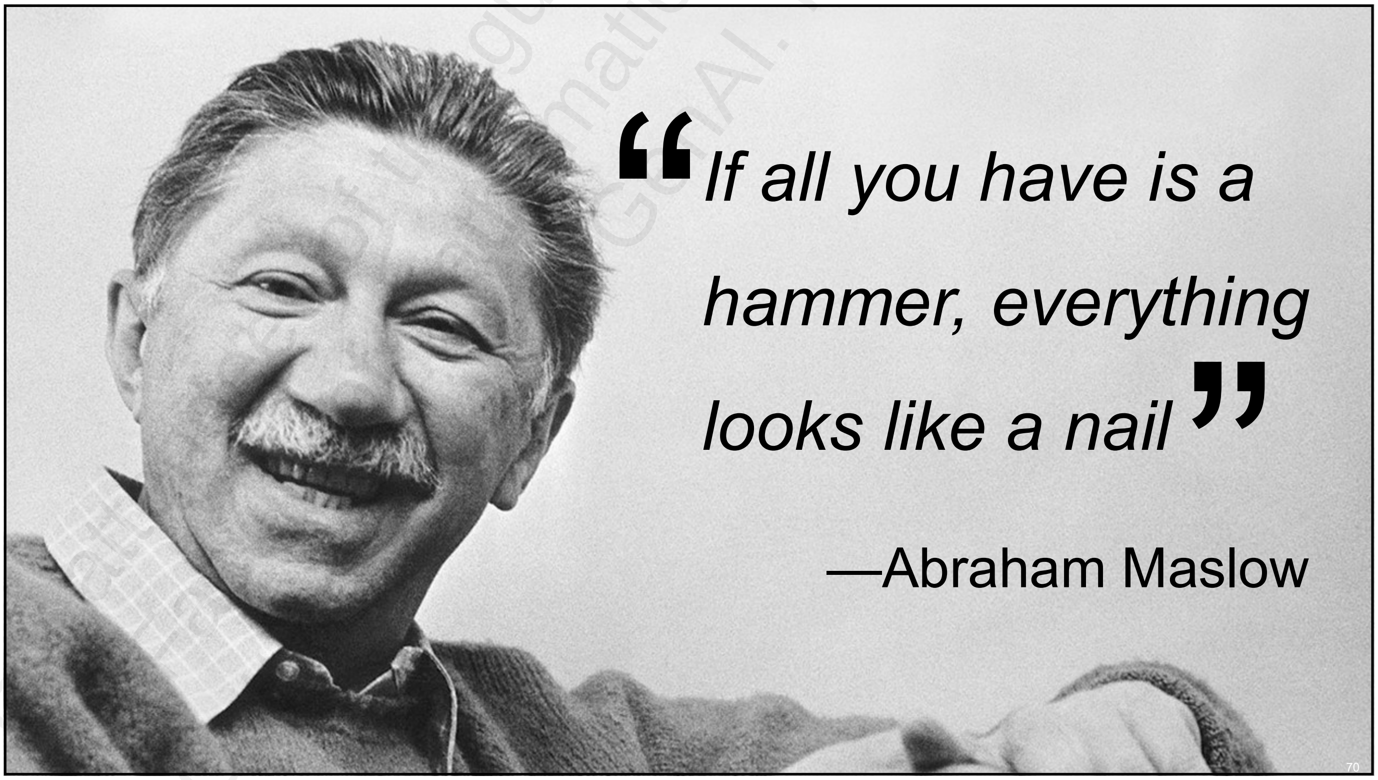
more models to come

more use cases to come

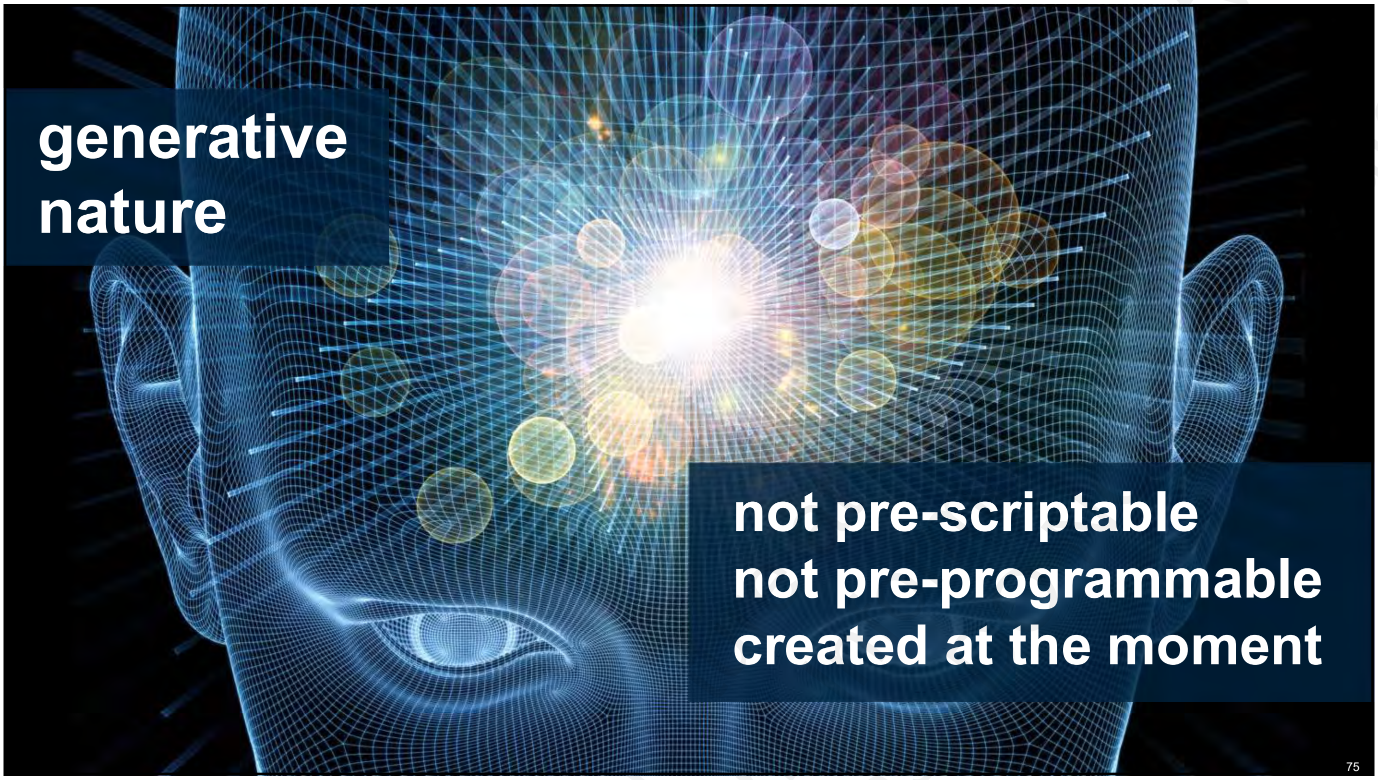
many many more start-ups

PROS.ai

twitter: @mich8elwu
linkedin.com/in/MichaelWuPhD



> “If all you have is a hammer, everything looks like a nail”
>
> —Abraham Maslow

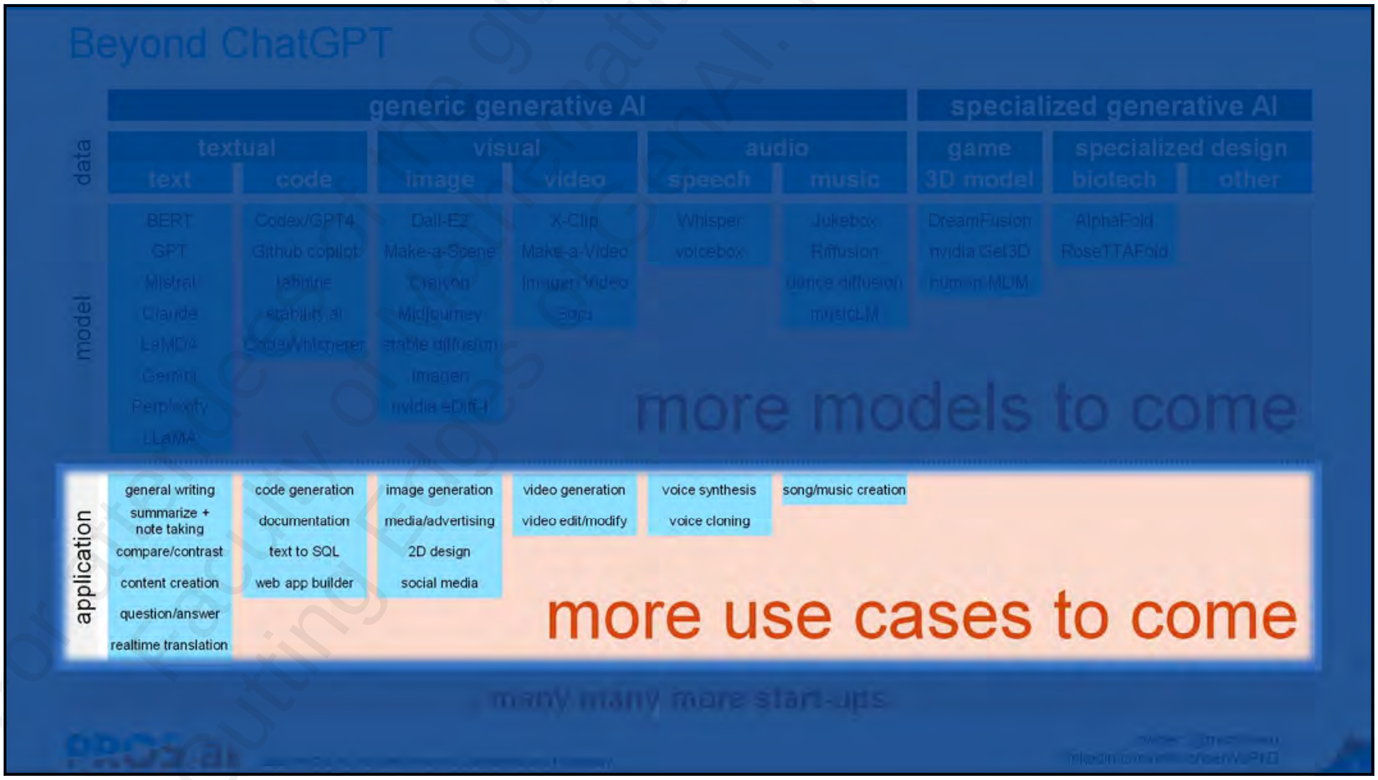


generative nature

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

75

75



Beyond ChatGPT

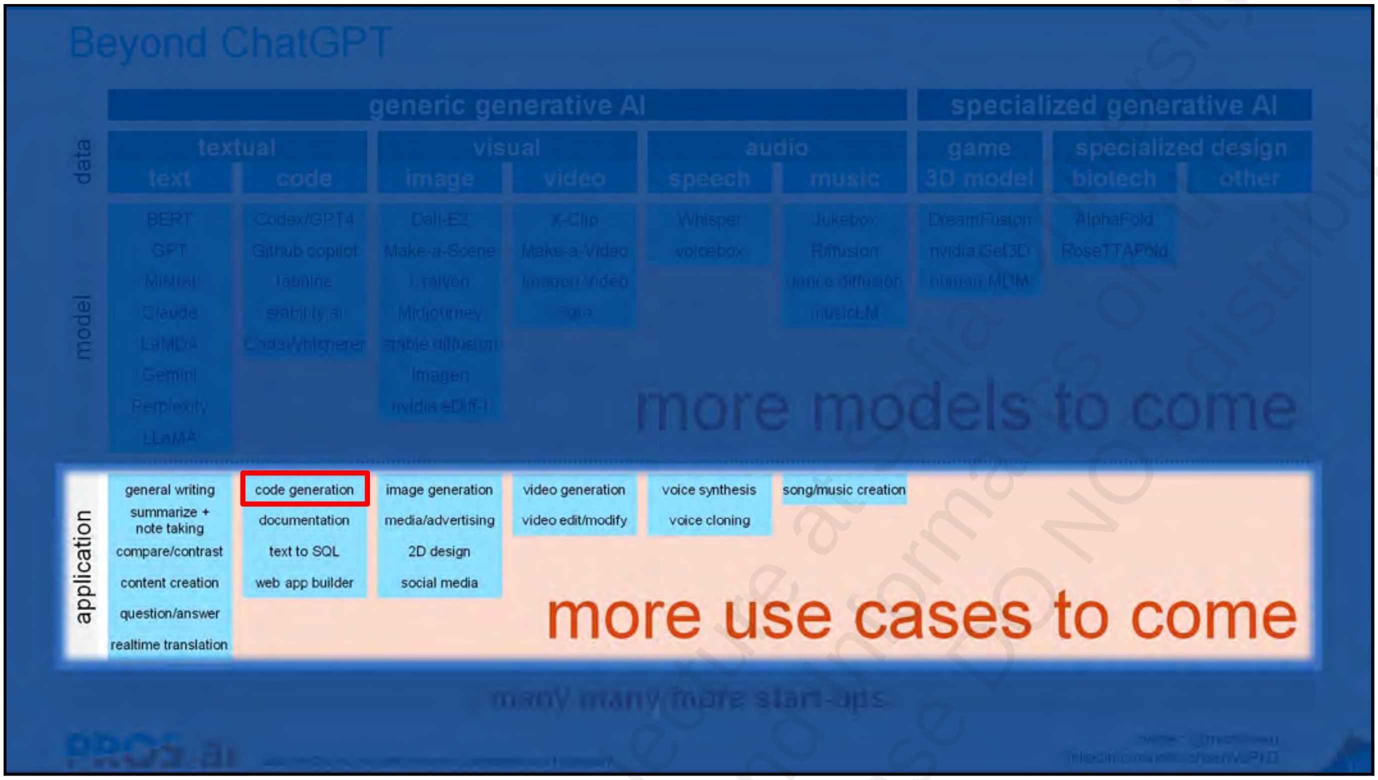
	generic generative AI						specialized generative AI		
	textual		visual		audio		game	specialized design	
data	text	code	image	video	speech	music	3D model	biotech	other
model	BERT GPT Mistral Claude LaMDA Gemini Perplexity LLaMA	Codex/GPT4 Github copilot tabnine stability.ai CodeWhisperer	Dall-E2 Make-a-Scene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

more models to come

more use cases to come

many many more start-ups

76



Beyond ChatGPT

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

more models to come

more use cases to come

many many more start-ups

PROS.ai

77

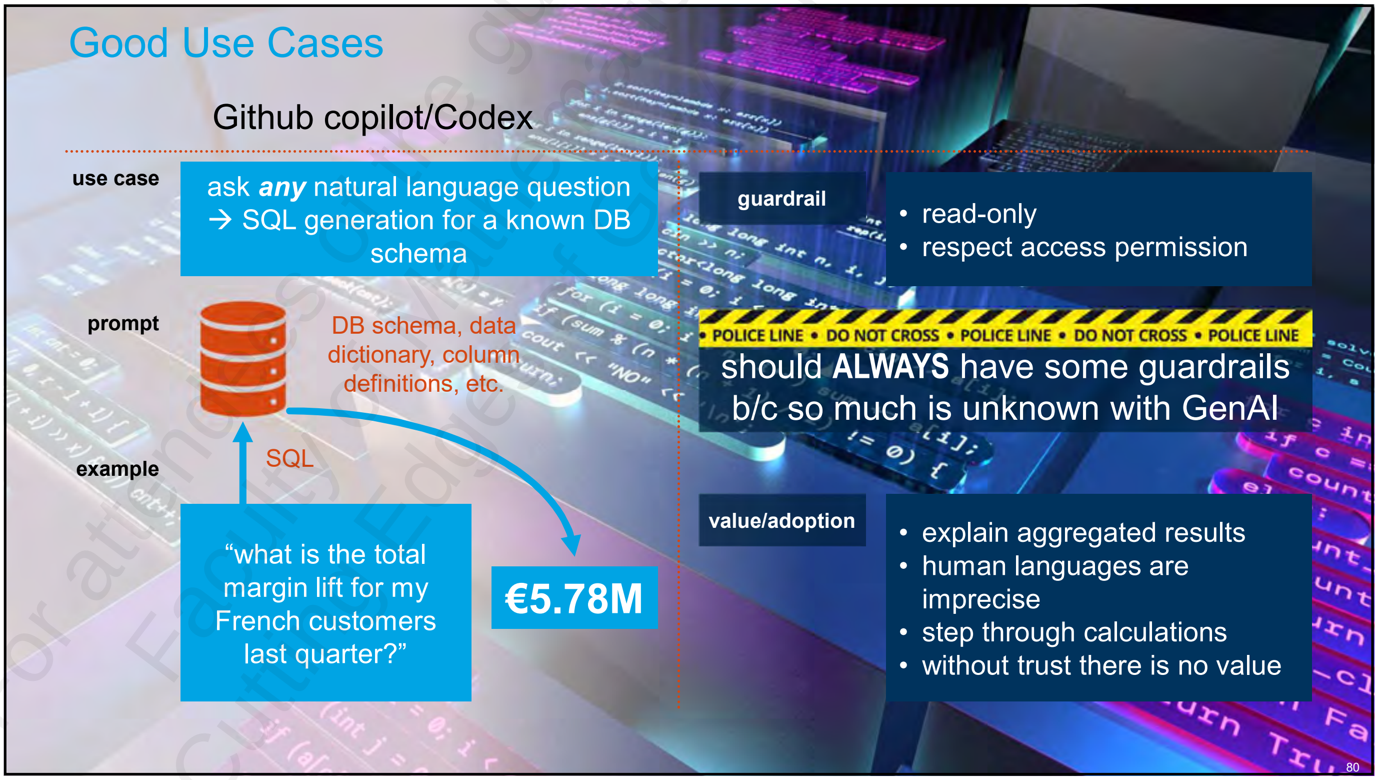
copyright

78

78

Kristina Kashtanova

79



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.

SQL

example

“what is the total margin lift for my French customers last quarter?”

€5.78M

guardrail

- read-only
- respect access permission

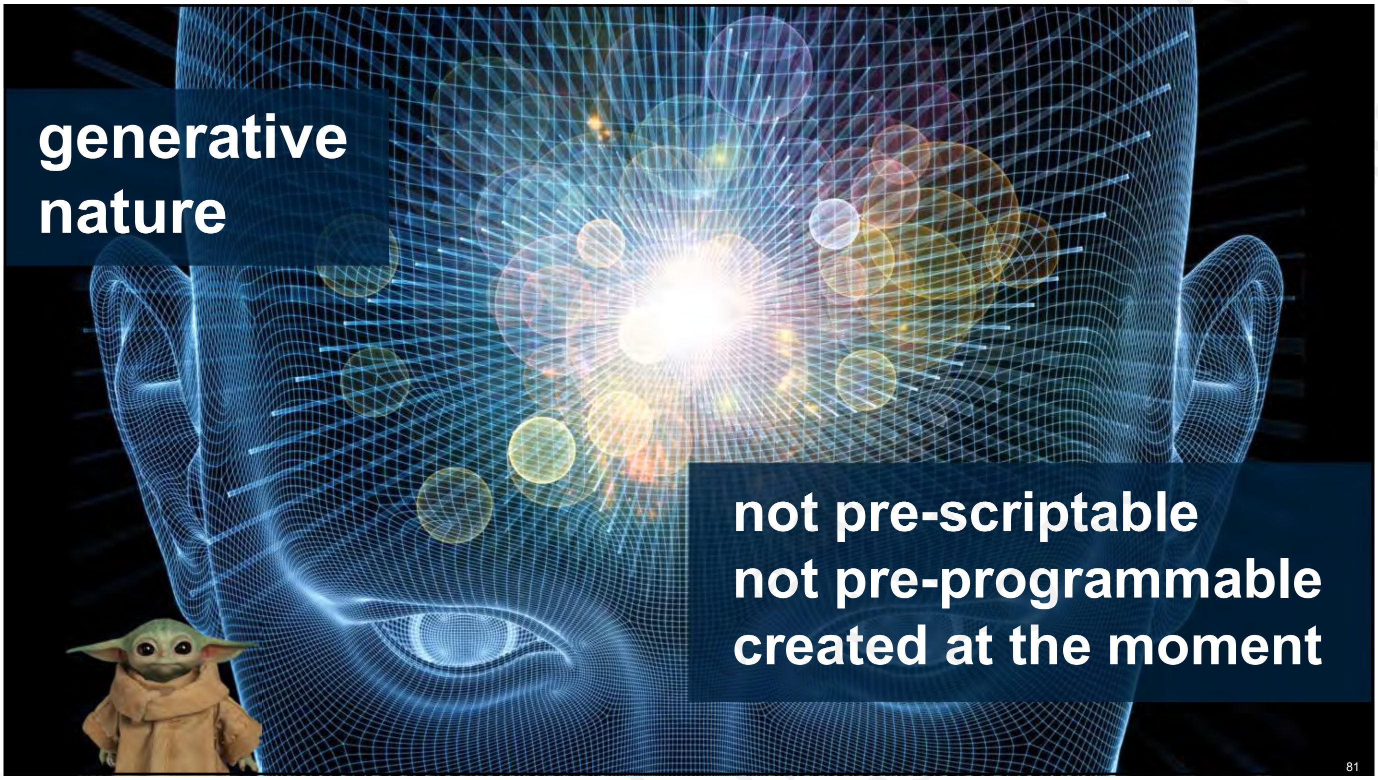
• POLICE LINE • DO NOT CROSS • POLICE LINE • DO NOT CROSS • POLICE LINE

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

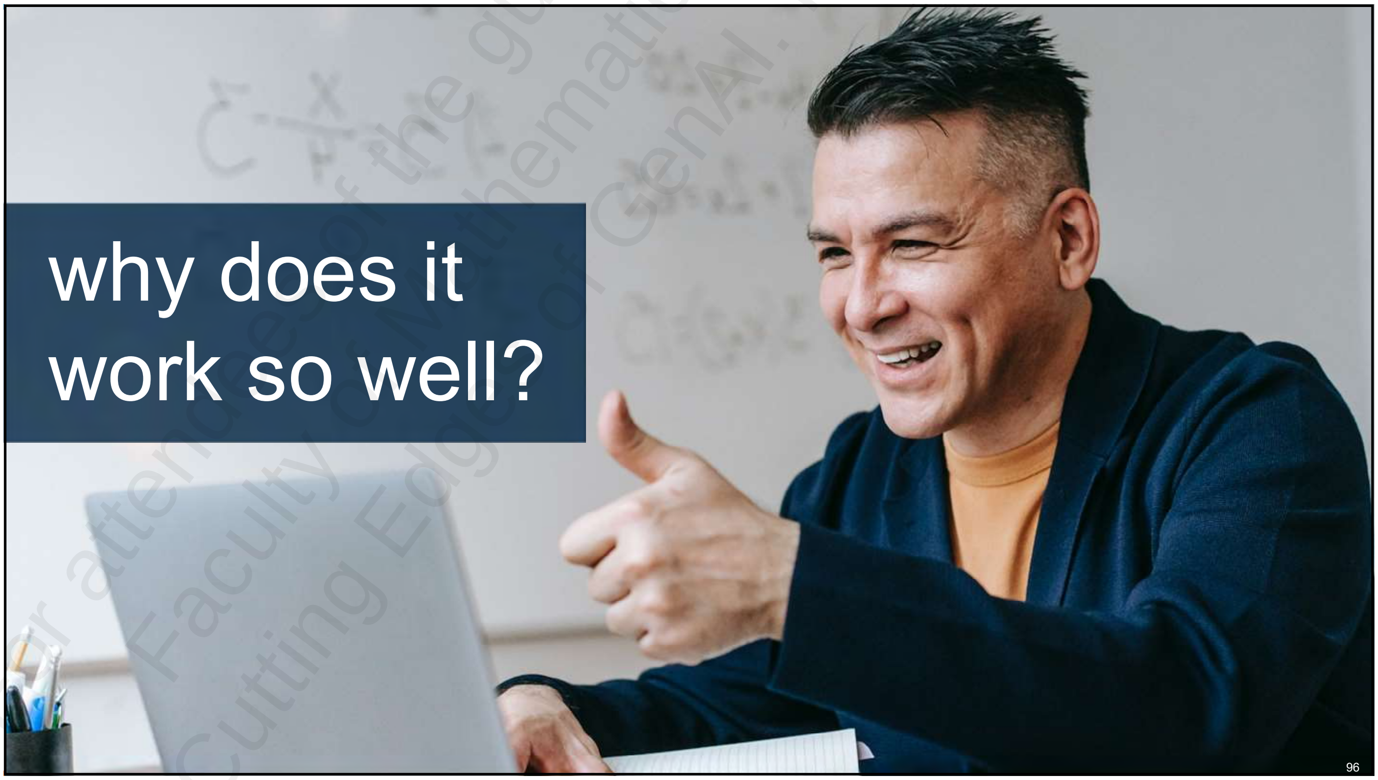
80



generative nature

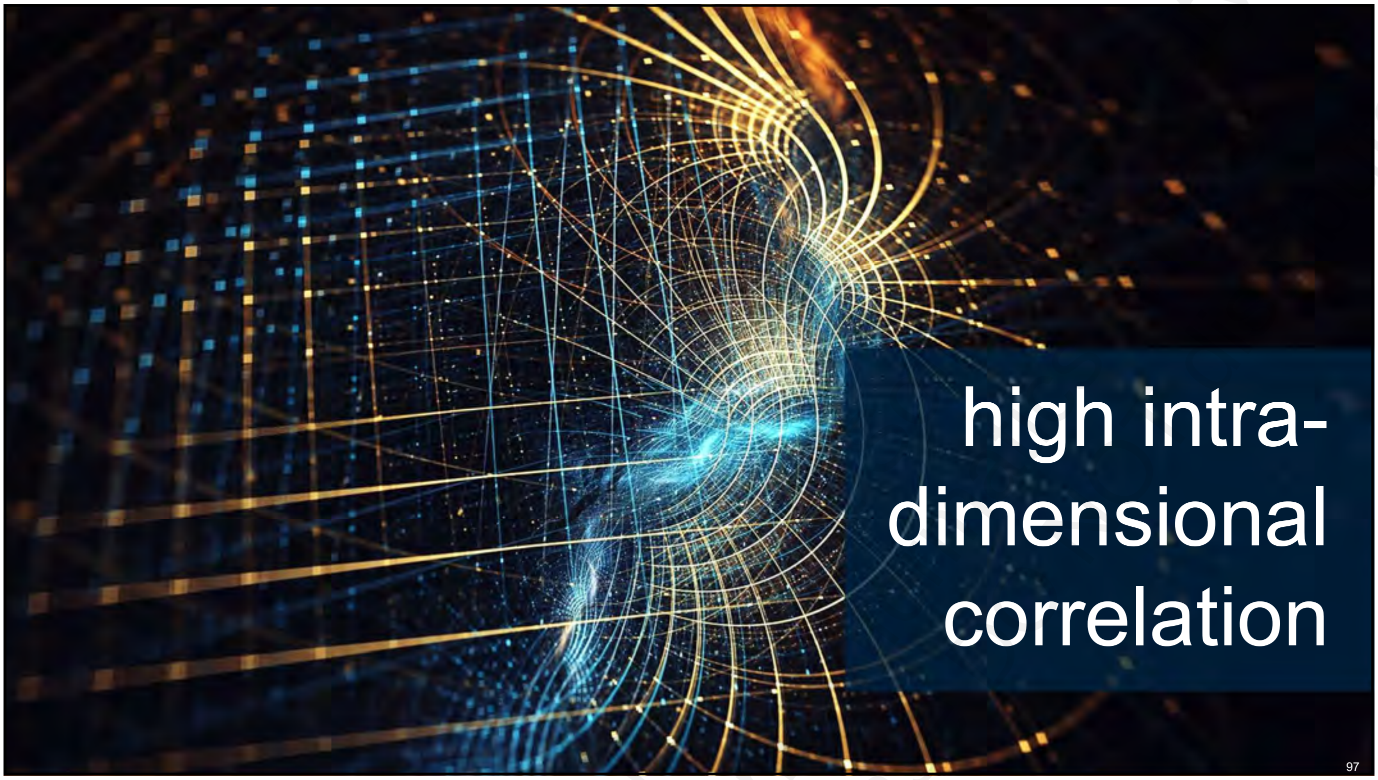
not pre-scriptable
not pre-programmable
created at the moment

81



why does it work so well?

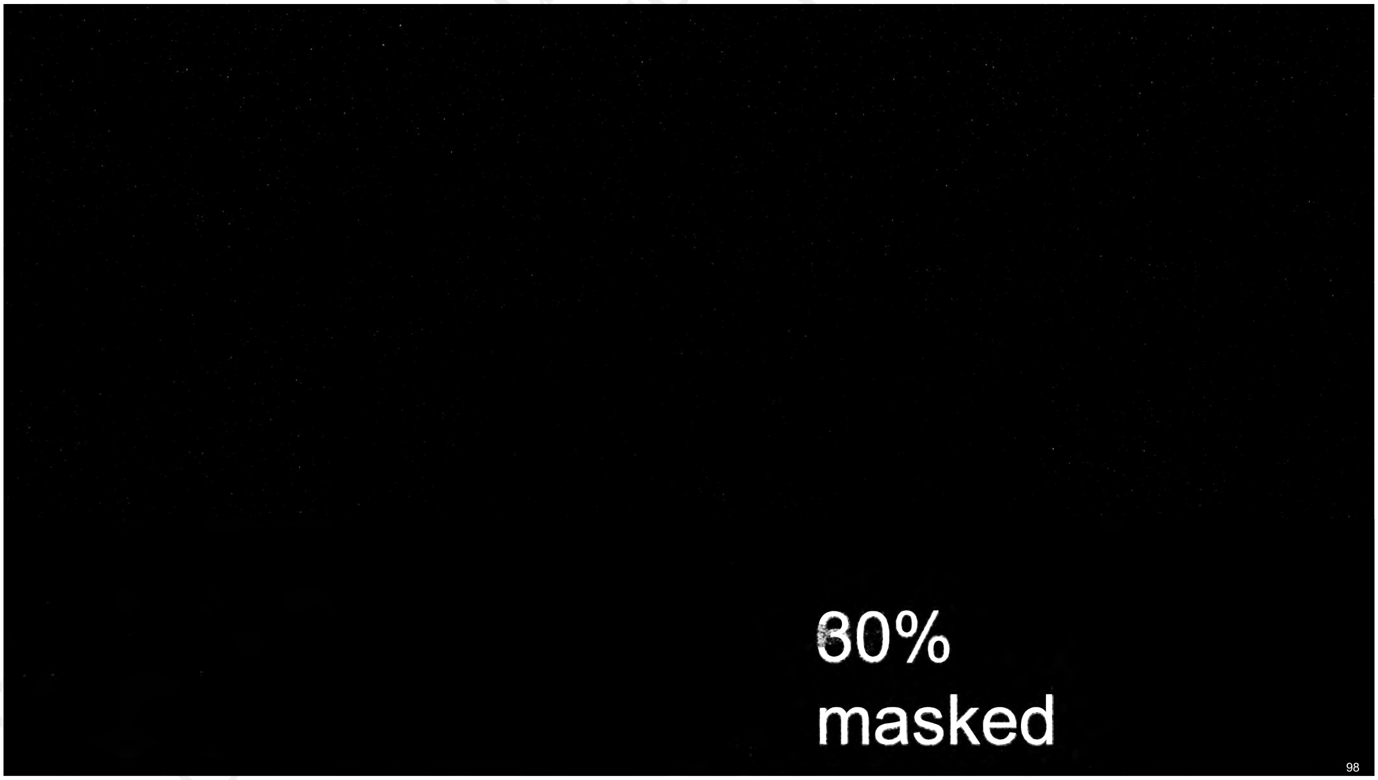
96



high intra-dimensional correlation

97

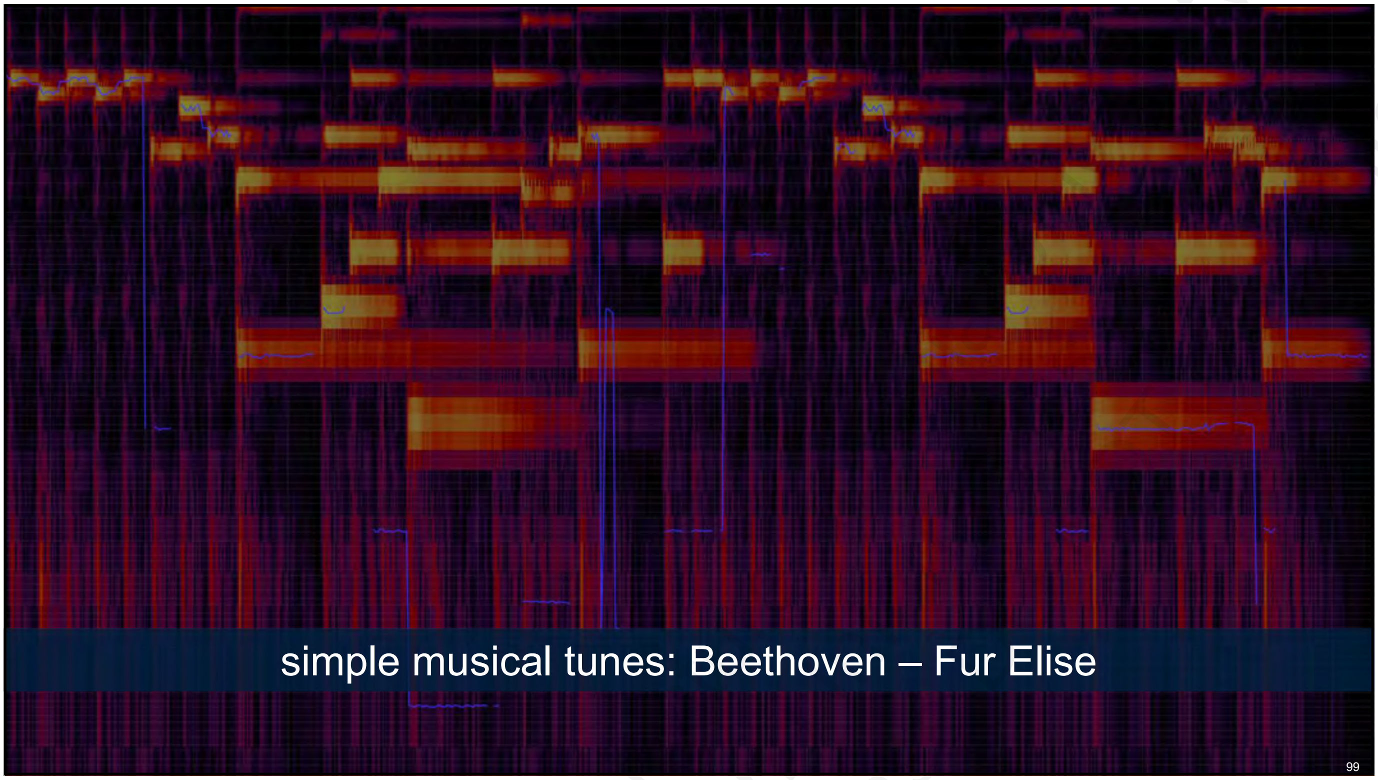
97



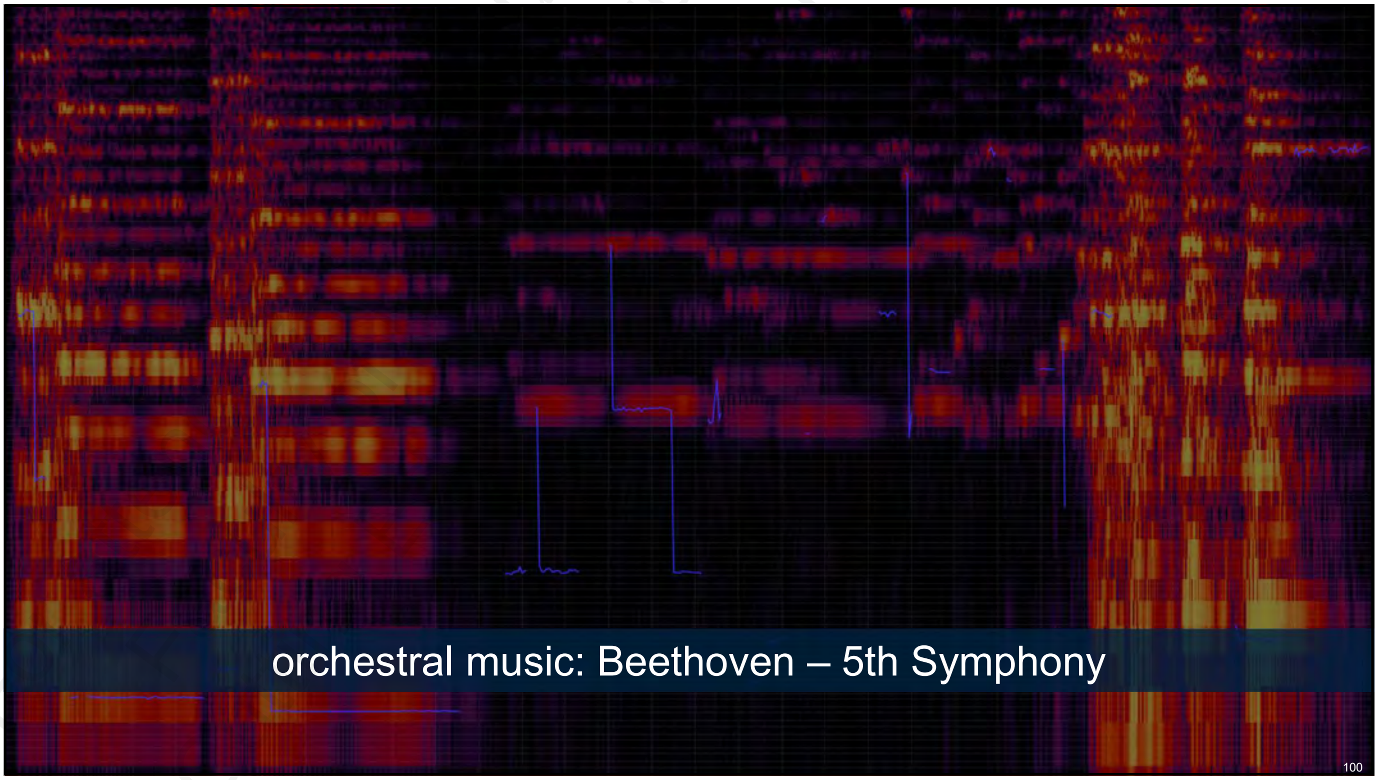
80% masked

98

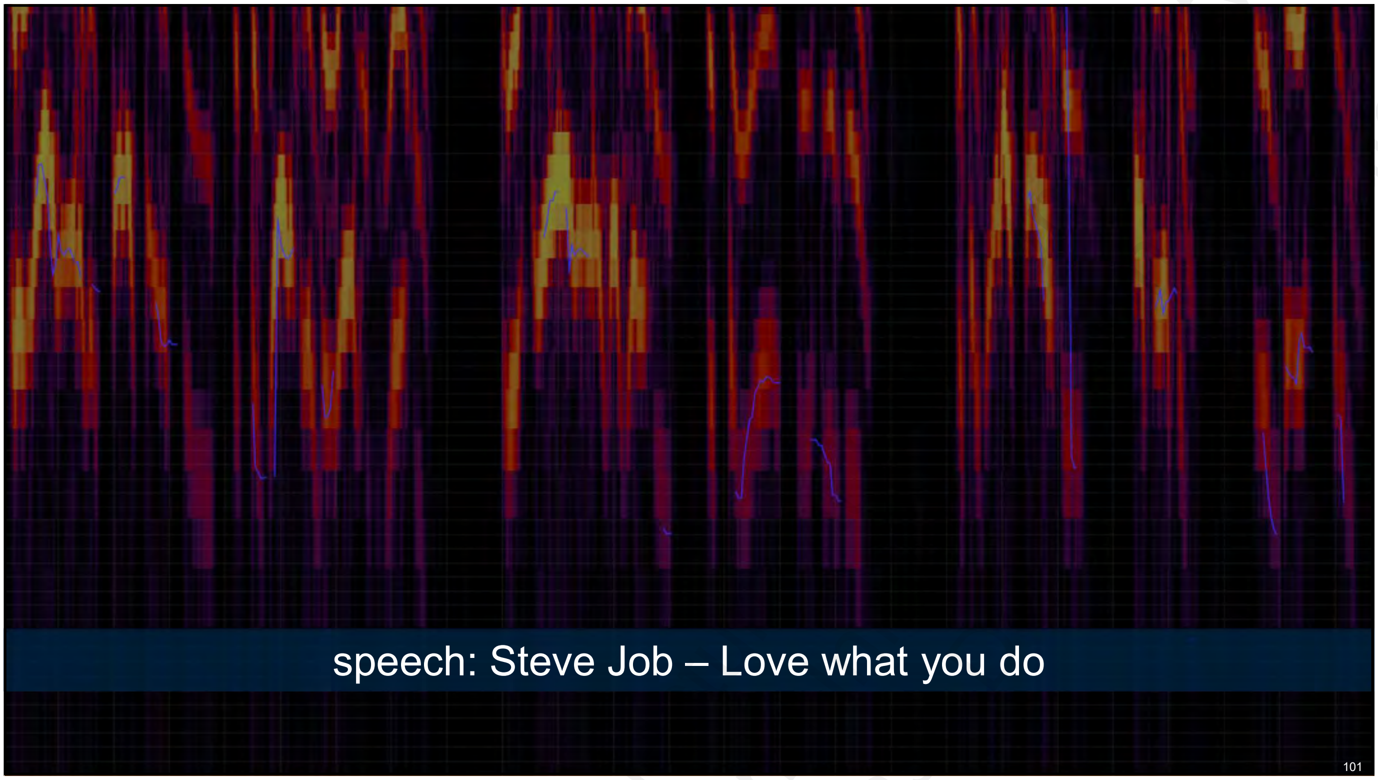
98



simple musical tunes: Beethoven – Fur Elise

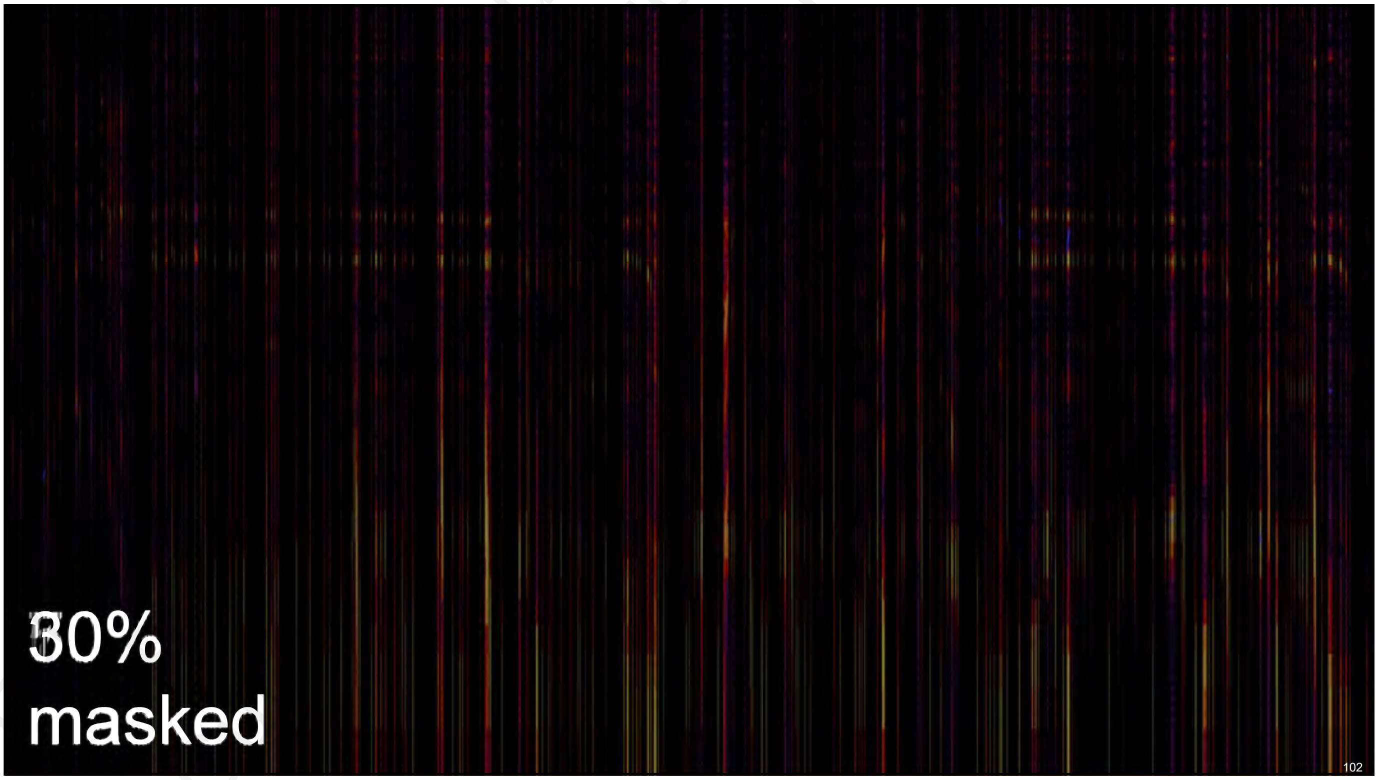


orchestral music: Beethoven – 5th Symphony



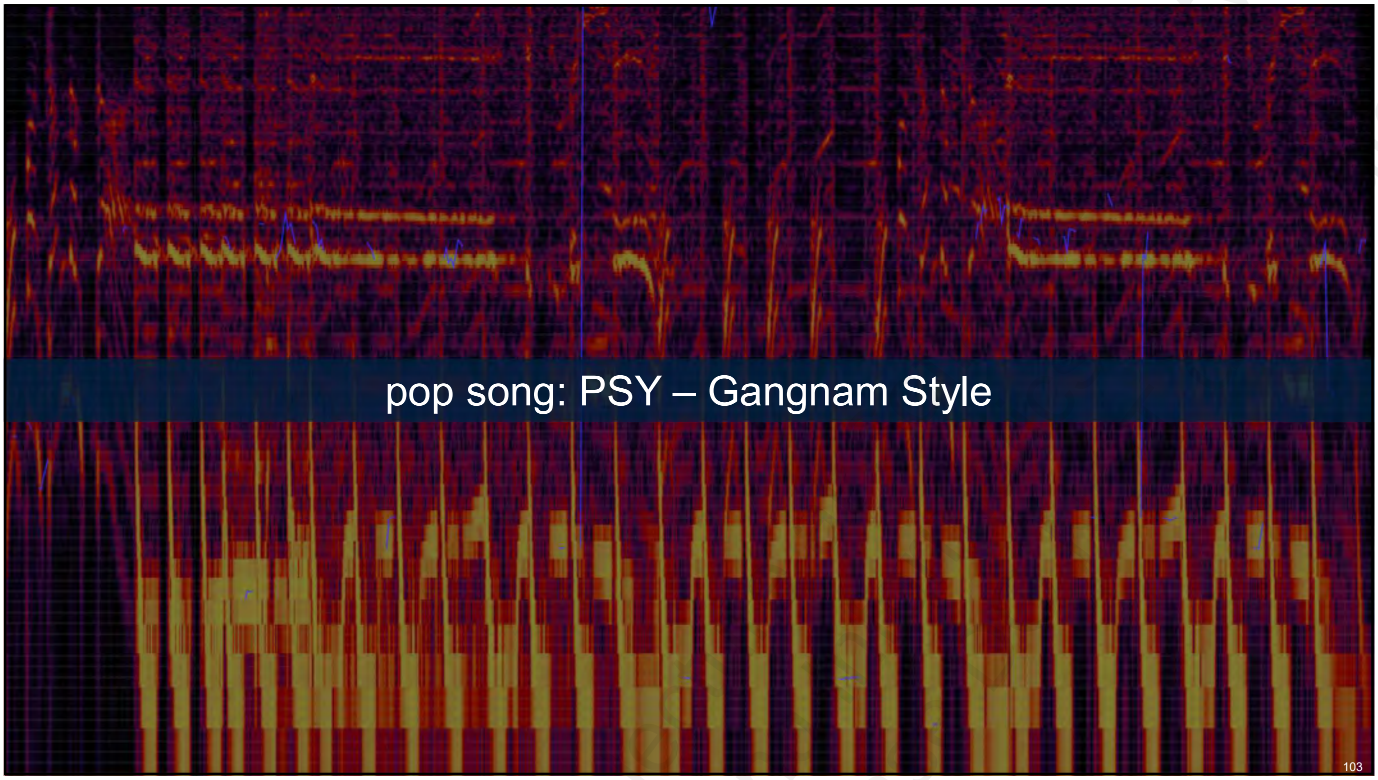
speech: Steve Job – Love what you do

101



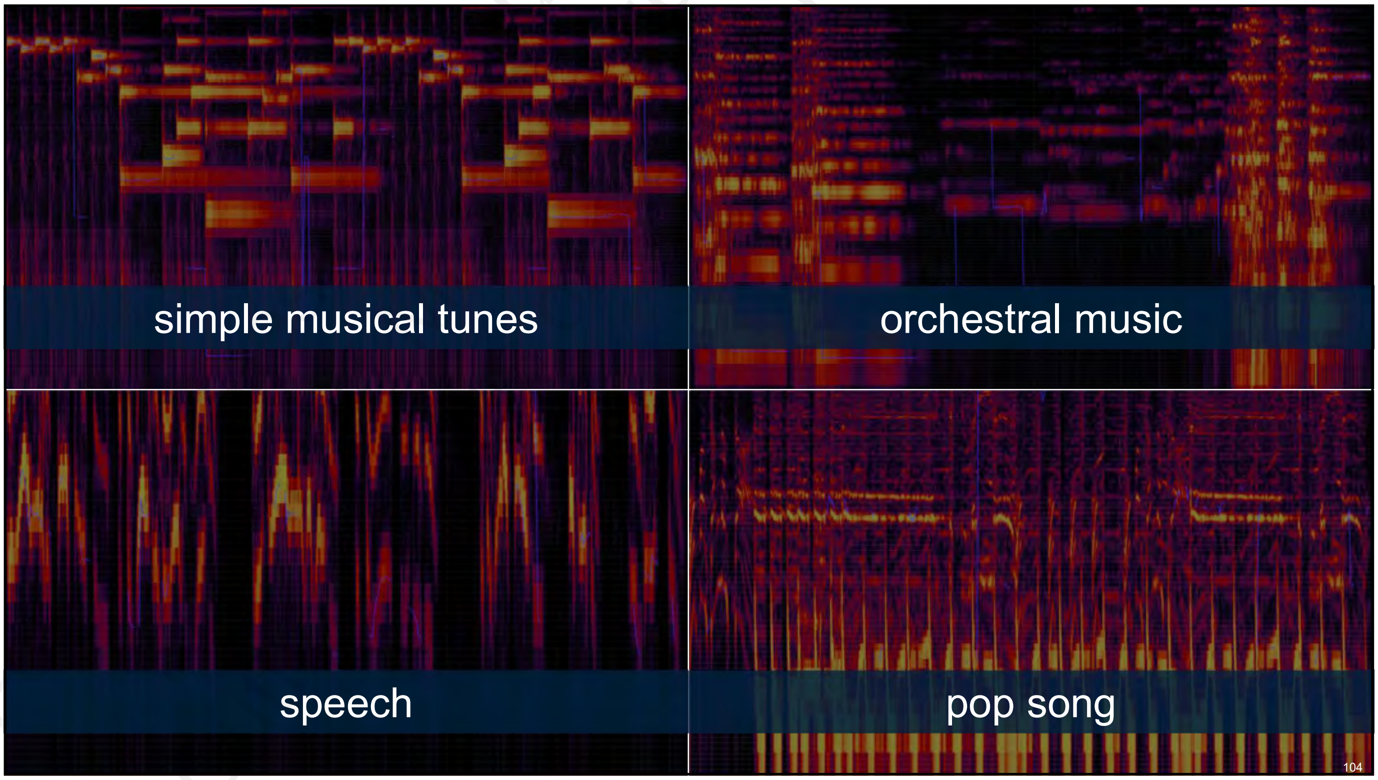
30%
masked

102



pop song: PSY – Gangnam Style

103



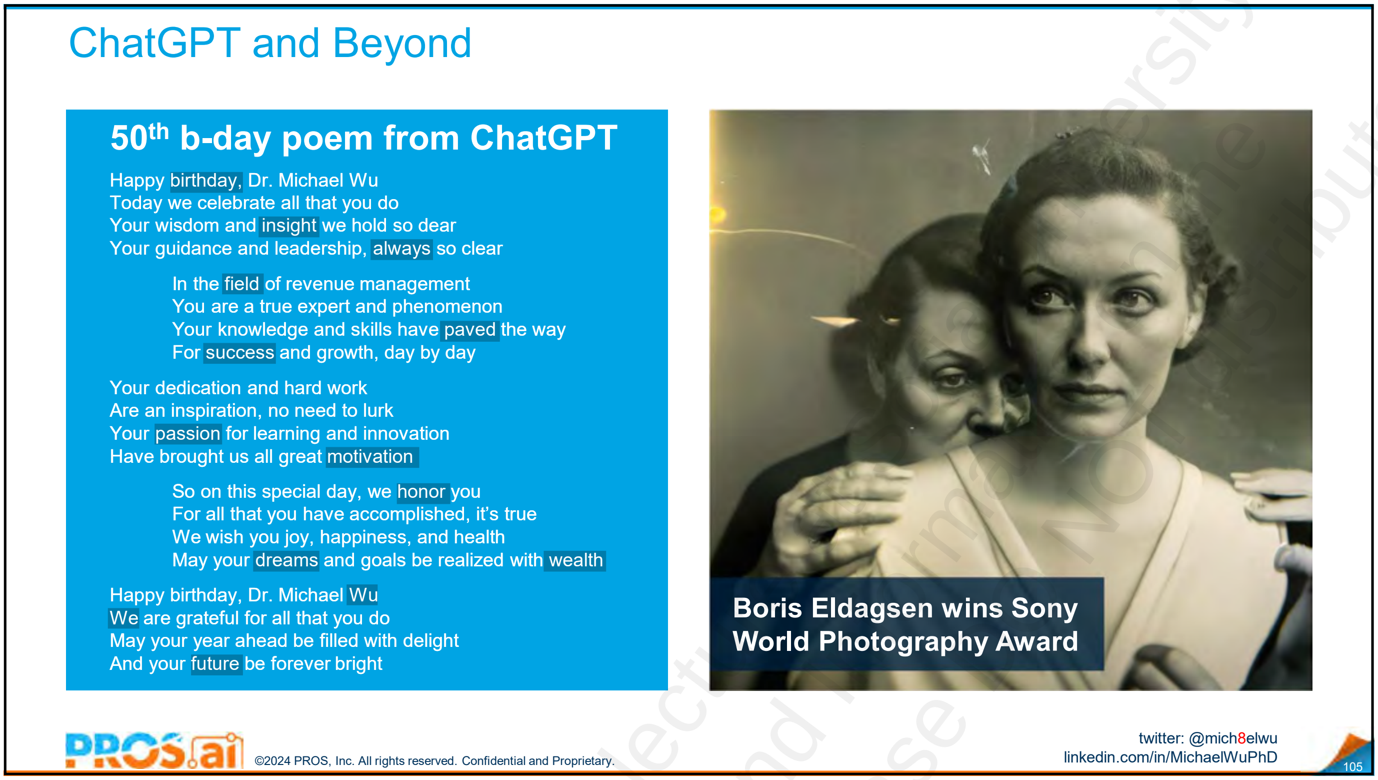
simple musical tunes

orchestral music

speech

pop song

104



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

PROS.ai ©2024 PROS, Inc. All rights reserved. Confidential and Proprietary.

twitter: @mich8elwu
linkedin.com/in/MichaelWuPhD

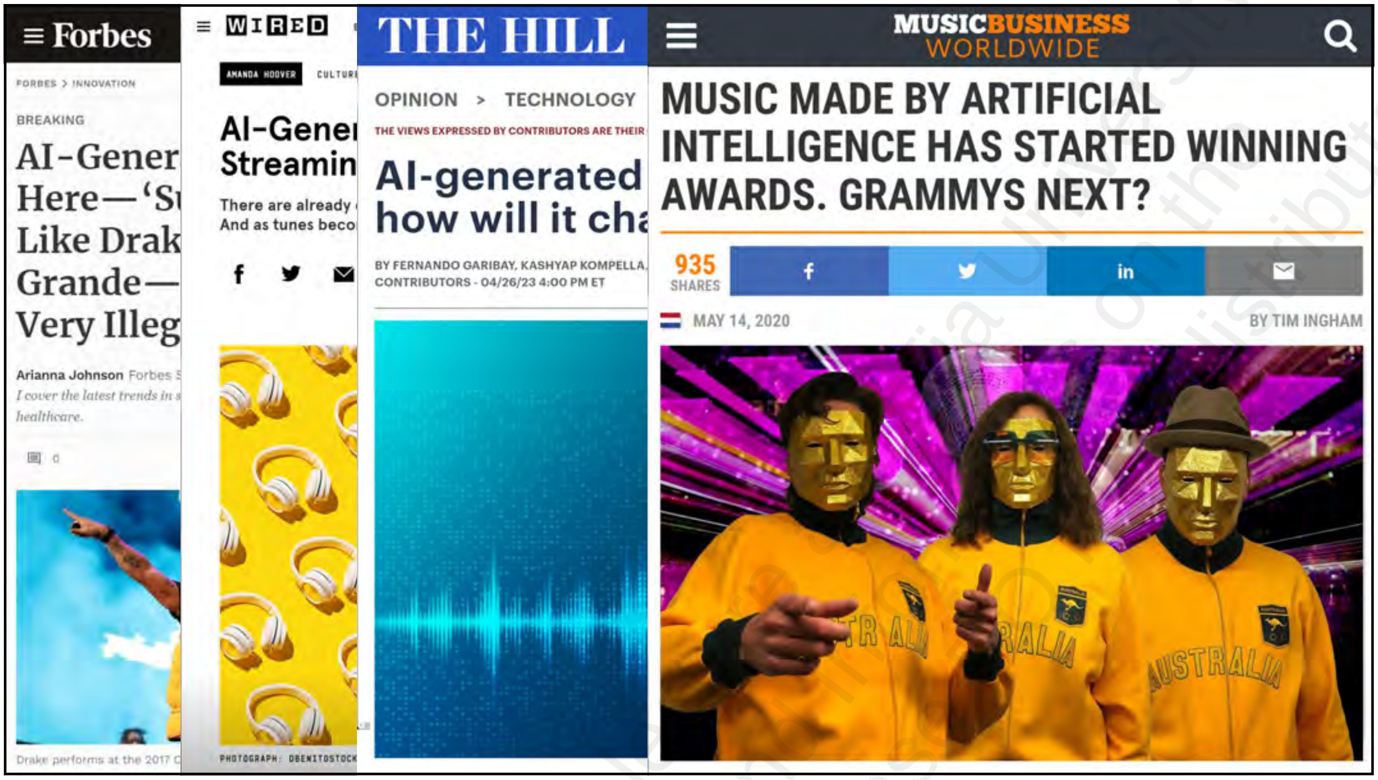
105

105

Jason Allen wins Colorado State Fair Digital Art Competition

106

106



≡ Forbes

FORBES > INNOVATION

BREAKING

AI-Gener
Here—'Su
Like Drak
Grande—
Very Illeg

Arianna Johnson Forbes S
I cover the latest trends in s
healthcare.

Drake performs at the 2017 C

≡ WIRED

AMANDA HOOVER CULTURE

AI-Gener
Streamin

There are already
And as tunes beco

PHOTOGRAPH: DBENITOSTOCK

THE HILL

OPINION > TECHNOLOGY

THE VIEWS EXPRESSED BY CONTRIBUTORS ARE THEIR

AI-generated
how will it cha

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

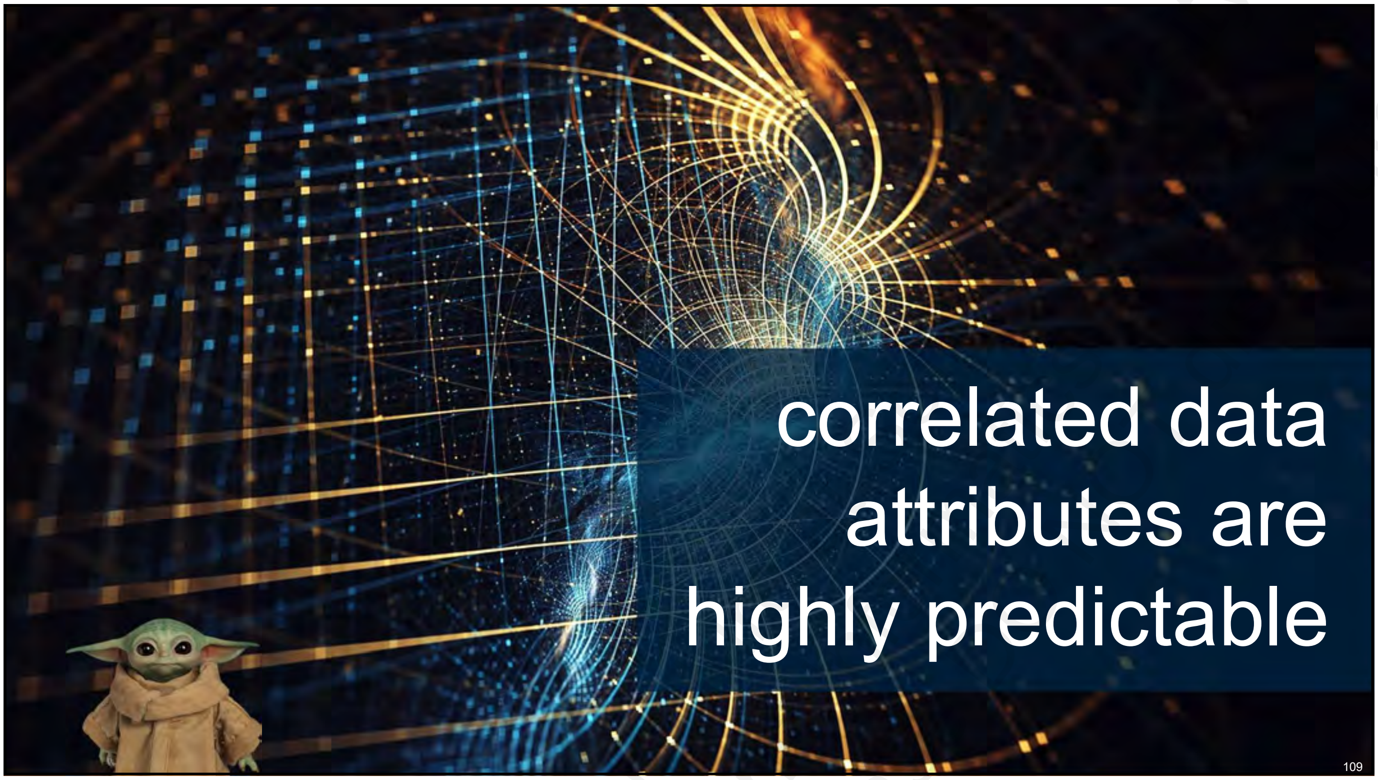
MAY 14, 2020

BY TIM INGHAM

107

isn't that amazing?

108

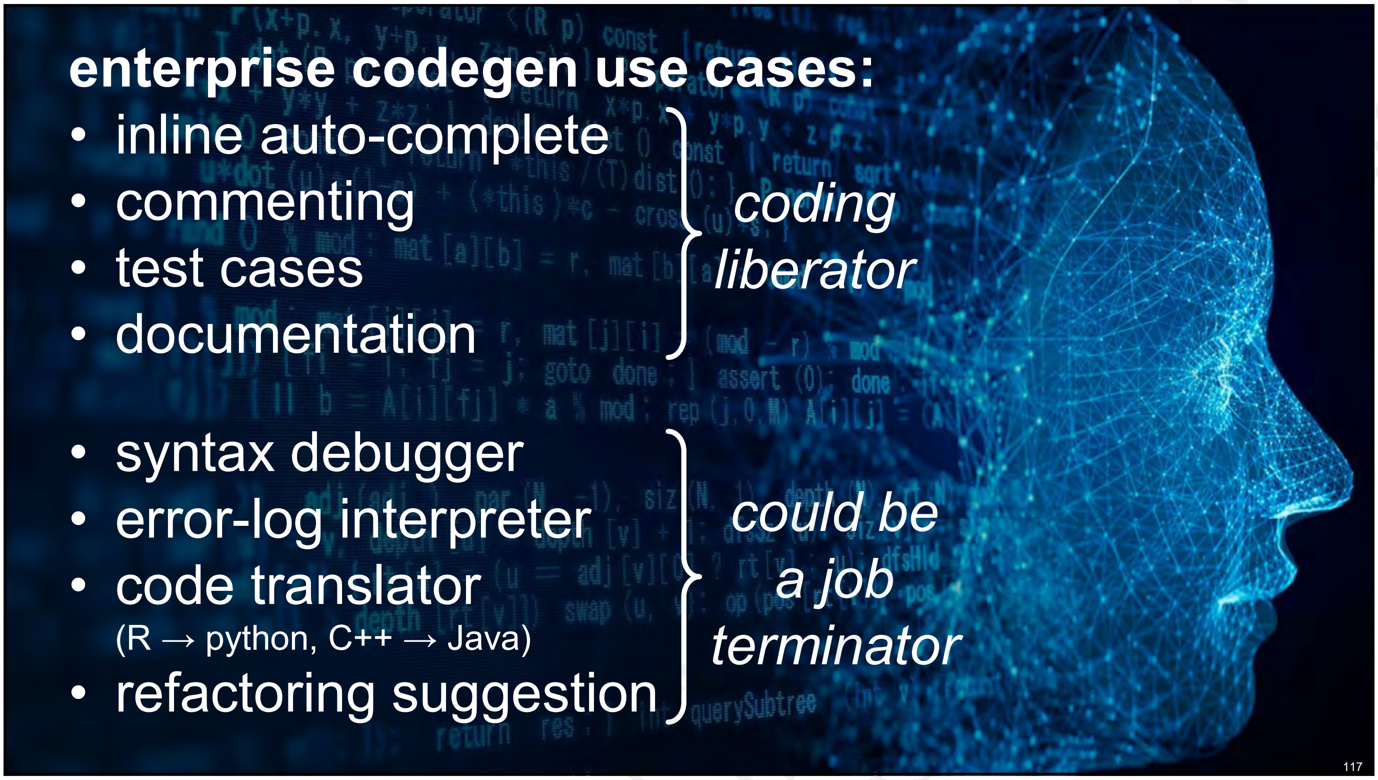


correlated data attributes are highly predictable

109

where does that leave software engineers and data scientists?

116



enterprise codegen use cases:

- inline auto-complete
- commenting
- test cases
- documentation

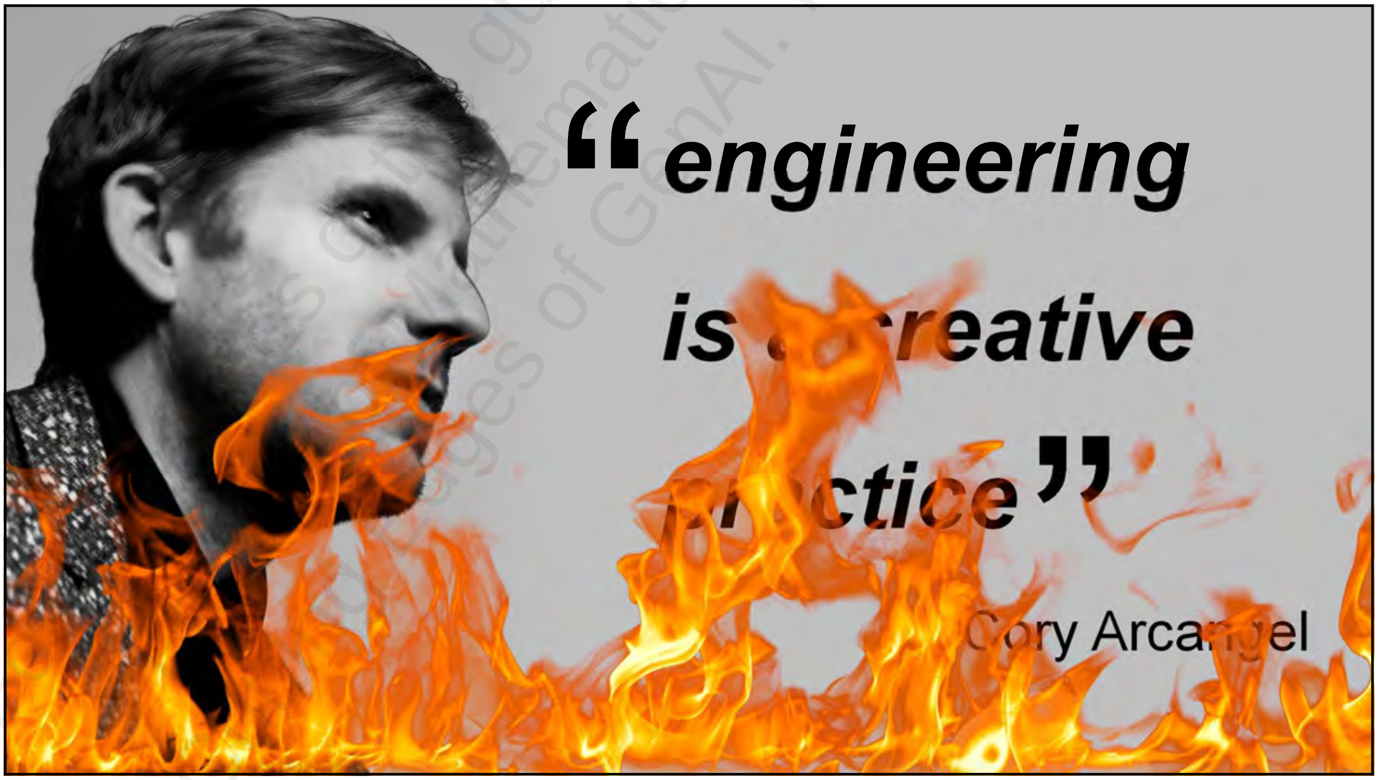
} *coding liberator*

- syntax debugger
- error-log interpreter
- code translator
  (R → python, C++ → Java)
- refactoring suggestion

} *could be a job terminator*

117

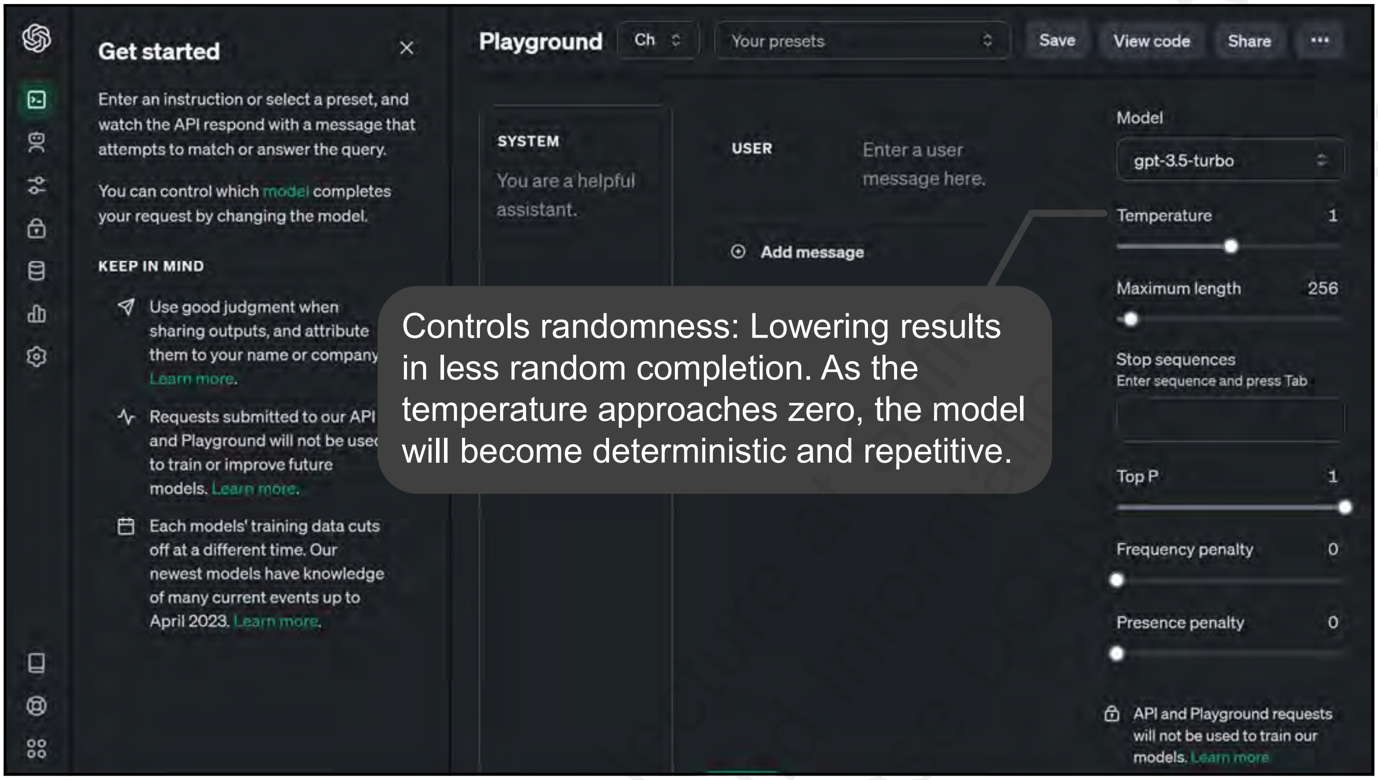
117



“***engineering is a creative practice***”

Cory Arcangel

118



Get started

Enter an instruction or select a preset, and watch the API respond with a message that attempts to match or answer the query.

You can control which model completes your request by changing the model.

KEEP IN MIND

- Use good judgment when sharing outputs, and attribute them to your name or company. Learn more.
- Requests submitted to our API and Playground will not be used to train or improve future models. Learn more.
- Each models' training data cuts off at a different time. Our newest models have knowledge of many current events up to April 2023. Learn more.

SYSTEM

You are a helpful assistant.

USER

Enter a user message here.

Add message

Model

gpt-3.5-turbo

Temperature 1

Maximum length 256

Stop sequences

Enter sequence and press Tab

Top P 1

Frequency penalty 0

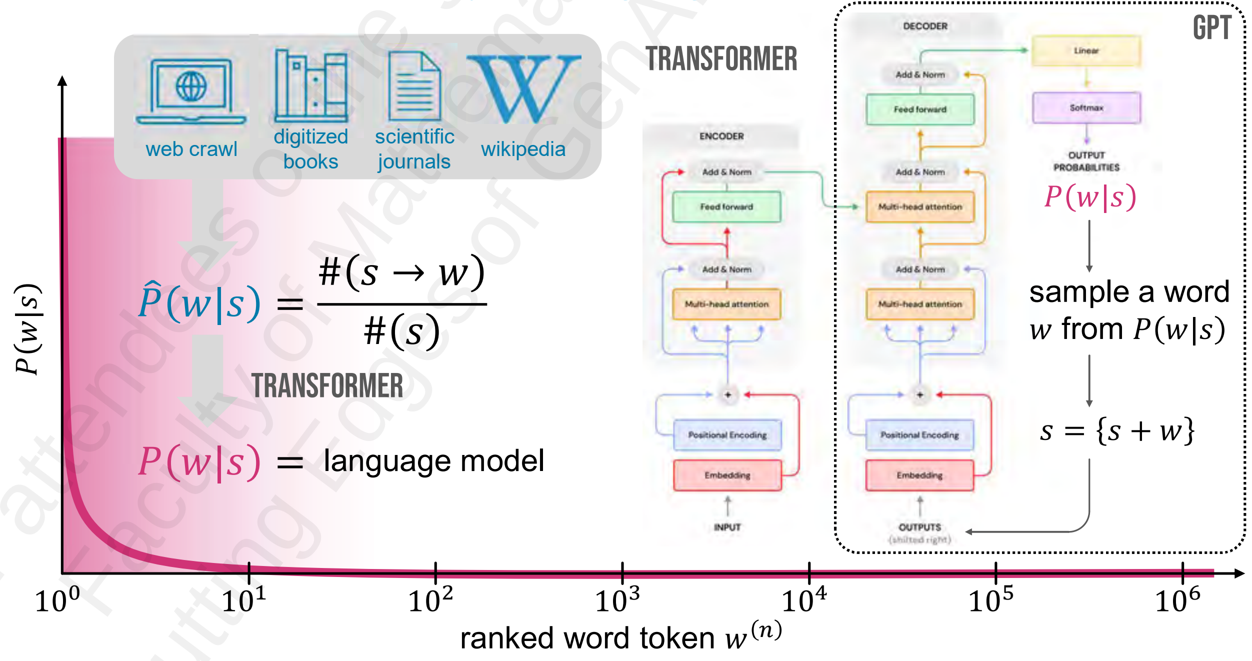
Presence penalty 0

API and Playground requests will not be used to train our models. Learn more

Controls randomness: Lowering results in less random completion. As the temperature approaches zero, the model will become deterministic and repetitive.

119

Empirical Word Probability to Language Model



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

$P(w|s) =$ language model

sample a word w from $P(w|s)$

$s = \{s + w\}$

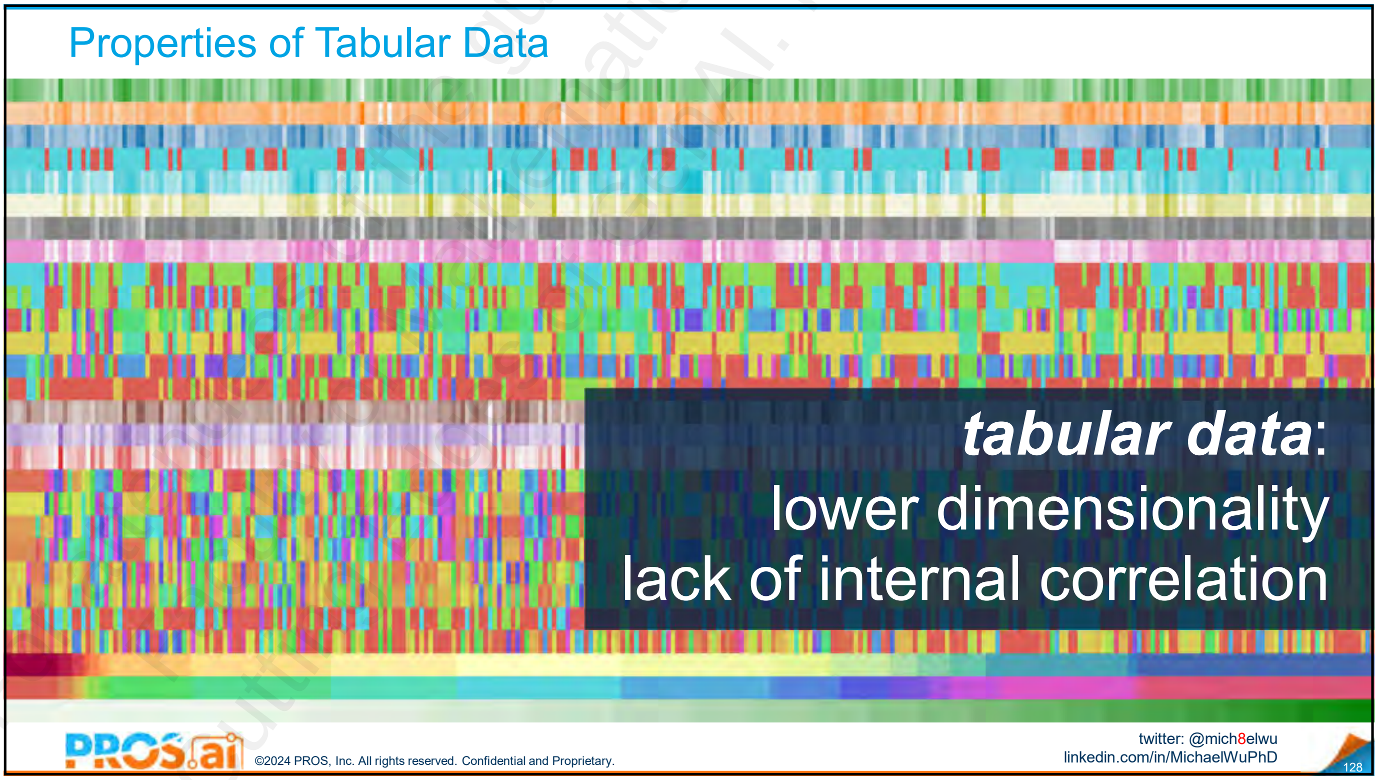
twitter: @mich8elwu
linkedin.com/in/MichaelWuPhD

120

most data scientists in the industry work with some sort of structured tabular data

127

127



Properties of Tabular Data

***tabular data*:**

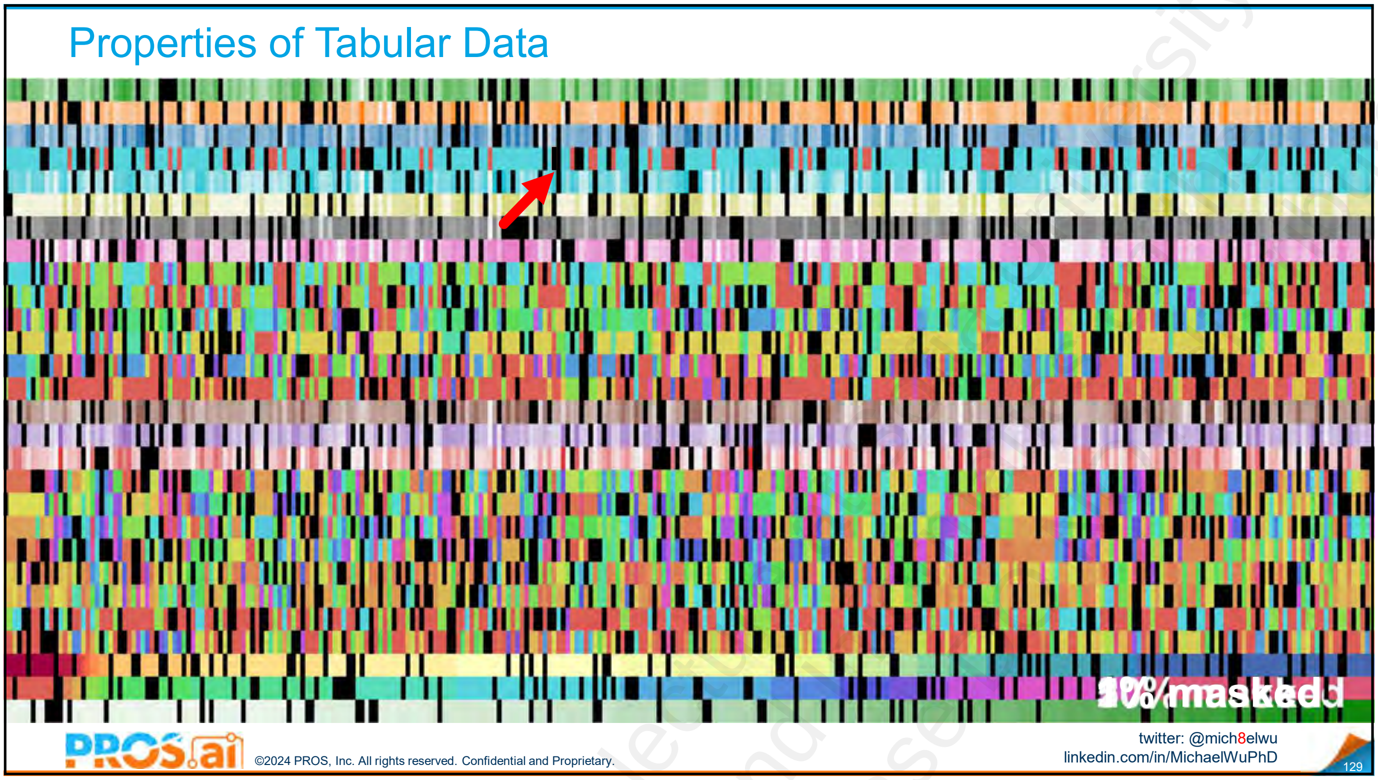
lower dimensionality

lack of internal correlation

twitter: @mich8elwu
linkedin.com/in/MichaelWuPhD

128

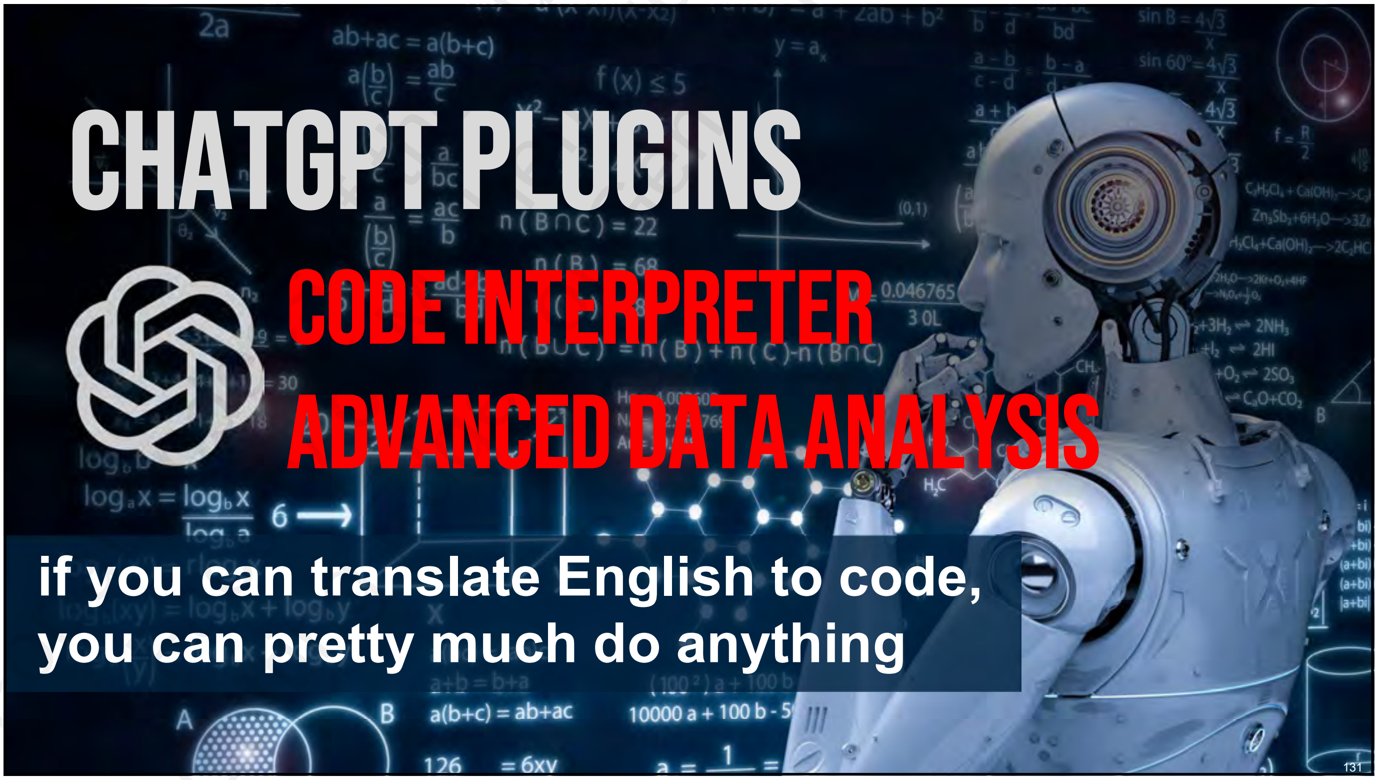
128



Properties of Tabular Data

twitter: @mich8elwu
linkedin.com/in/MichaelWuPhD

129



CHATGPT PLUGINS

CODE INTERPRETER

ADVANCED DATA ANALYSIS

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

131

the future of software development is one where engineers collaborate with different kinds of AI tools

users must have the relevant expertise in discerning and mitigating risk of hallucinations

132

132

AI *will* ***not*** replace you, but someone using AI ***will***

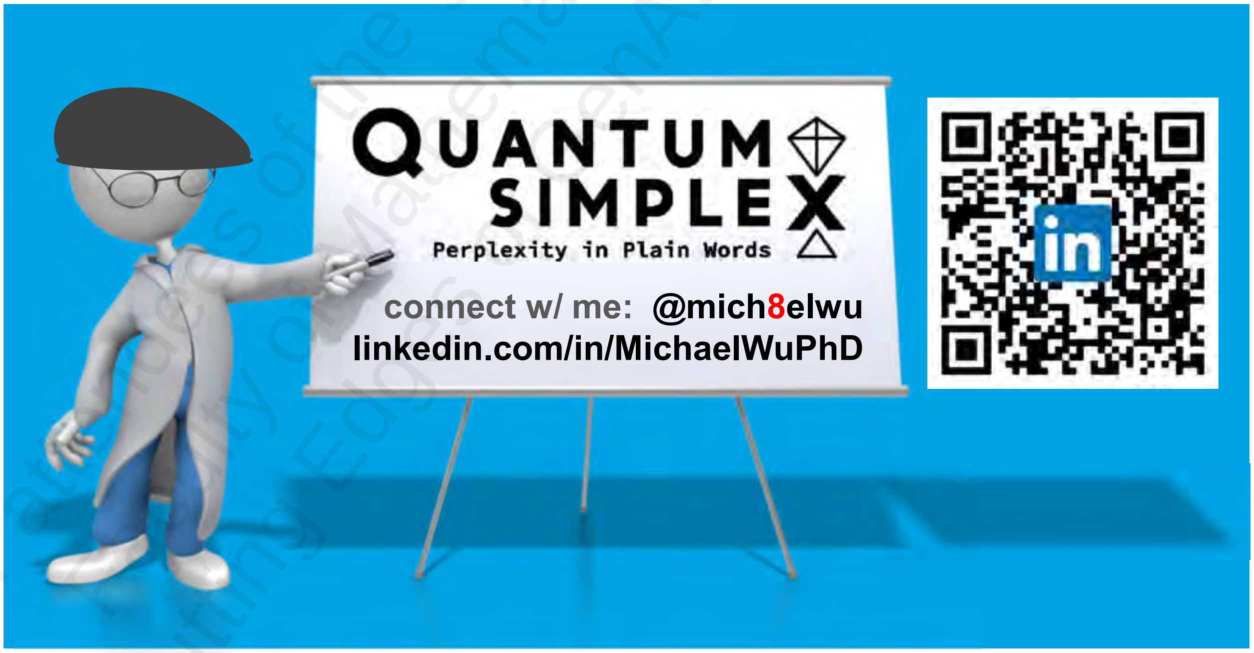
users must have the relevant expertise in discerning and mitigating risk of hallucinations

133

133

145

145



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

twitter: @mich8elwu
linkedin.com/in/MichaelWuPhD

149

149