

# *Navigating the Limits and Potential of GenAI in Mission-Critical Business Decisions*

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

2024.04.23



**PROS.**

1

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

2024.04.23



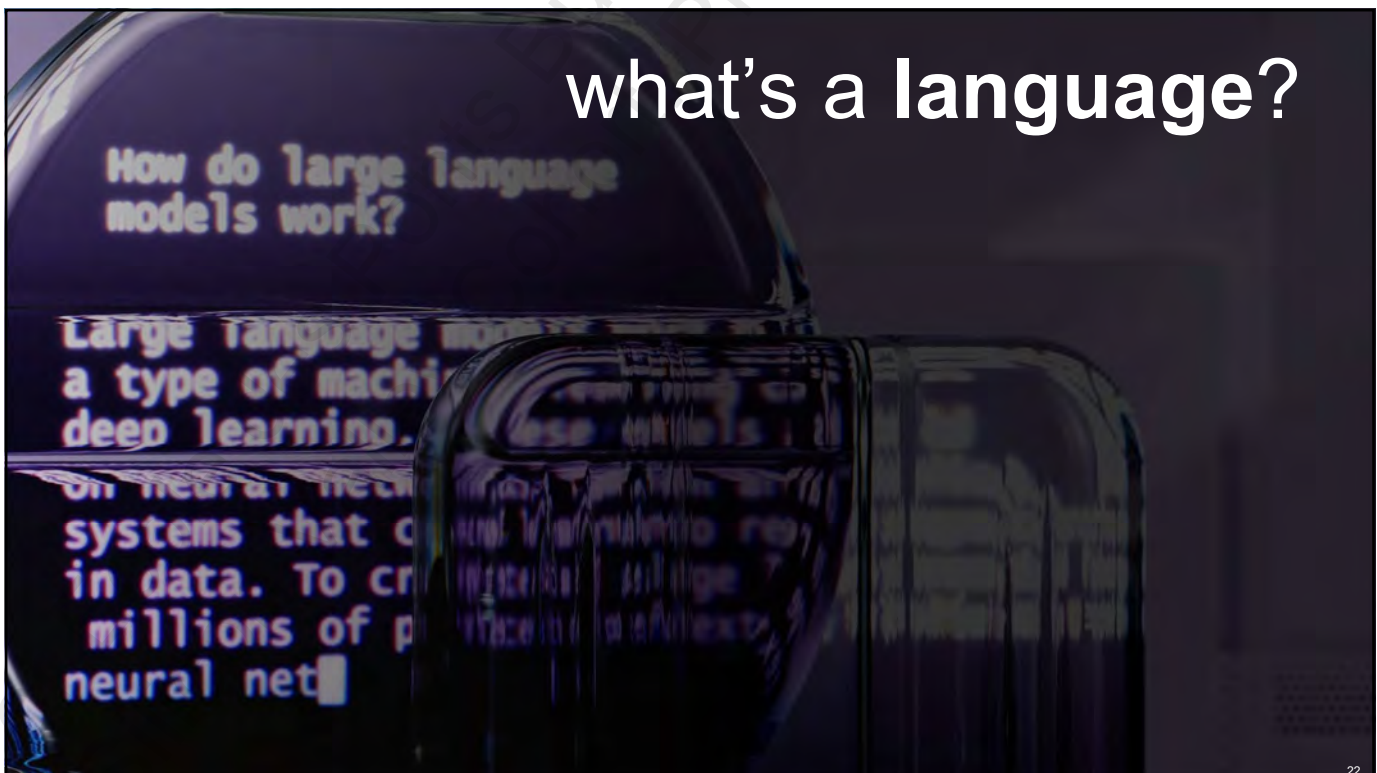
13



- difficult topic
- only ~100 min
- interactive
- pay attention, have fun, learn
- don't take notes, get pdf slide via linkedin

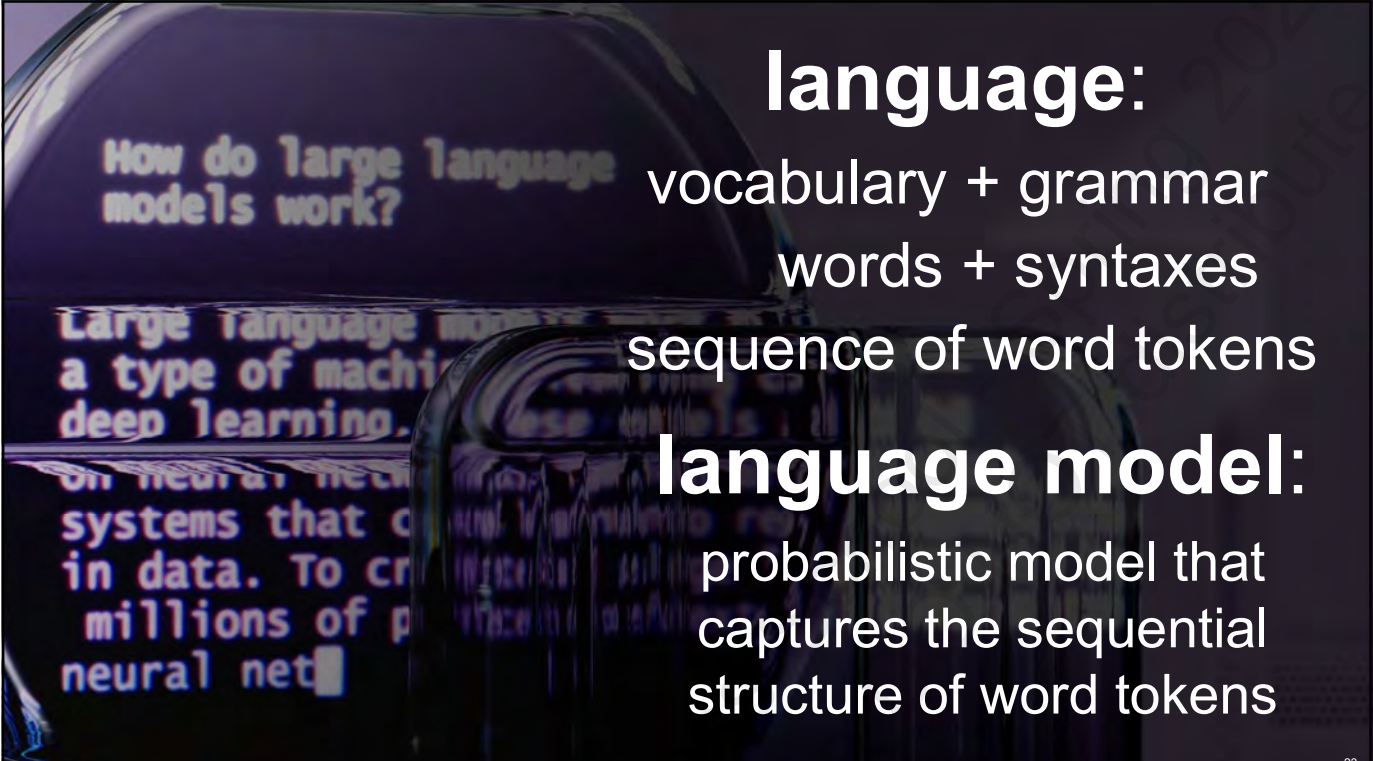
can I have your cooperation?

14



what's a language?

22



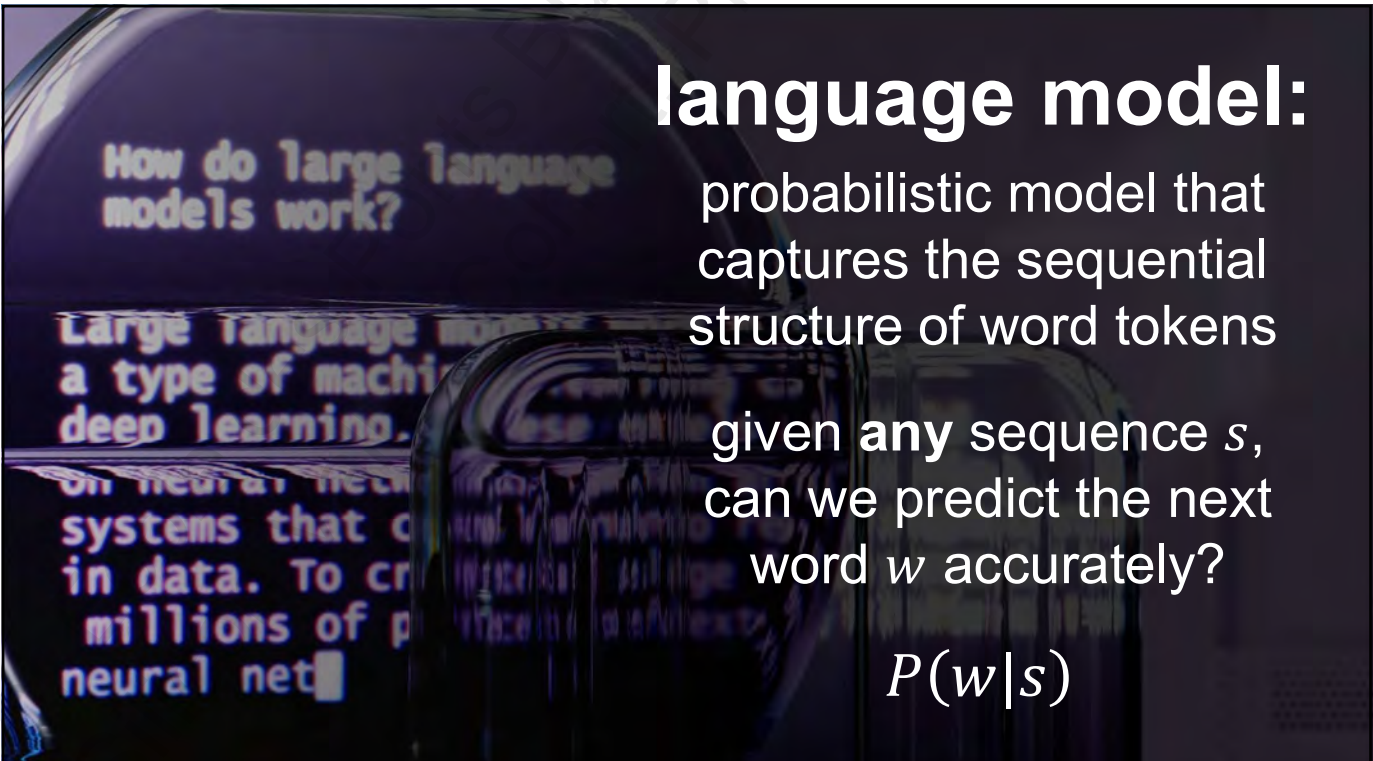
How do large language models work?

Large language models are a type of machine learning model based on neural networks. They are trained on massive amounts of data. To create these models, researchers use deep learning systems that process 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

23



How do large language models work?

Large language models are a type of machine learning model based on neural networks. They are trained on massive amounts of data. To create these models, researchers use deep learning systems that process 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?

$$P(w|s)$$

24



# language model:

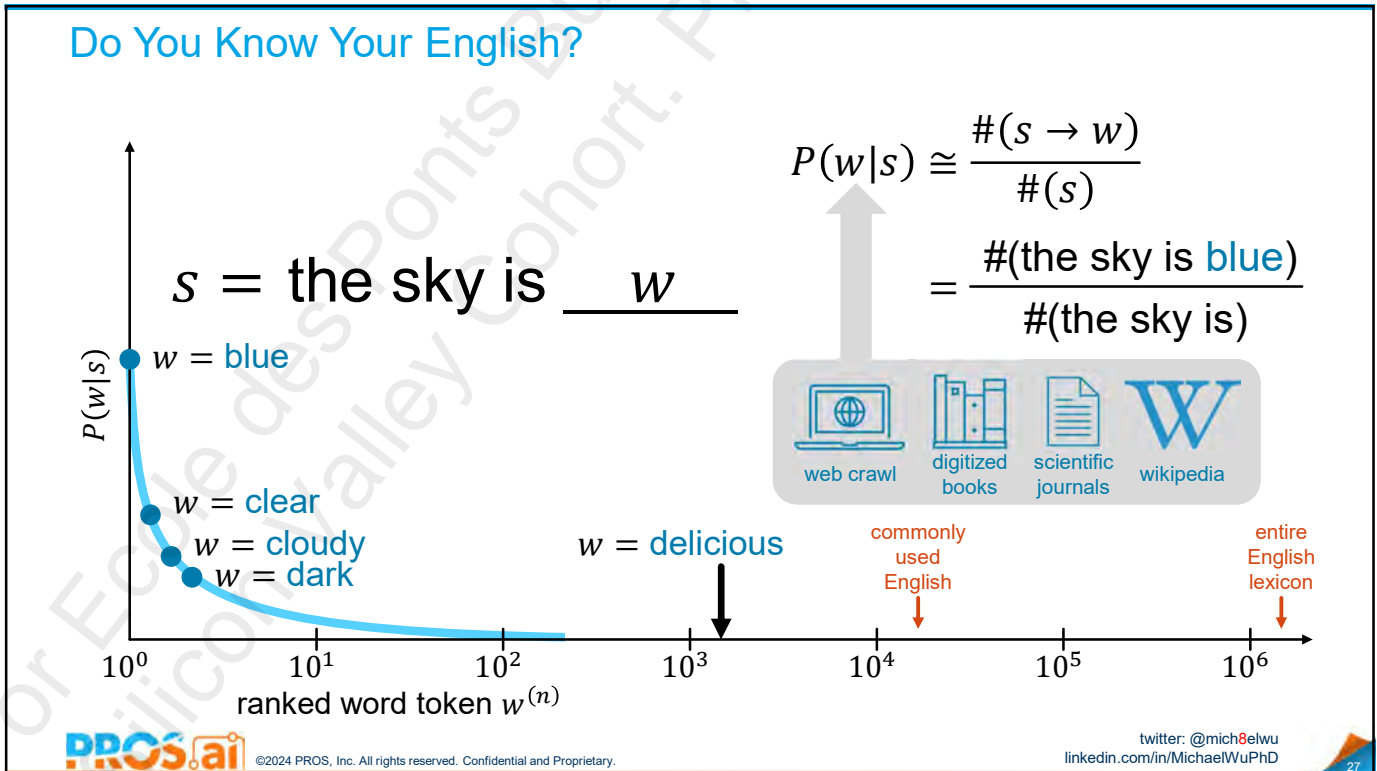
given *any* sequence  $s$ ,  
can we predict the next  
word  $w$  accurately?

$$P(w|s)$$

$s = \textit{the water is filled with ribulose-bisphosphate-carboxylase-oxygenase, it's very _____}$

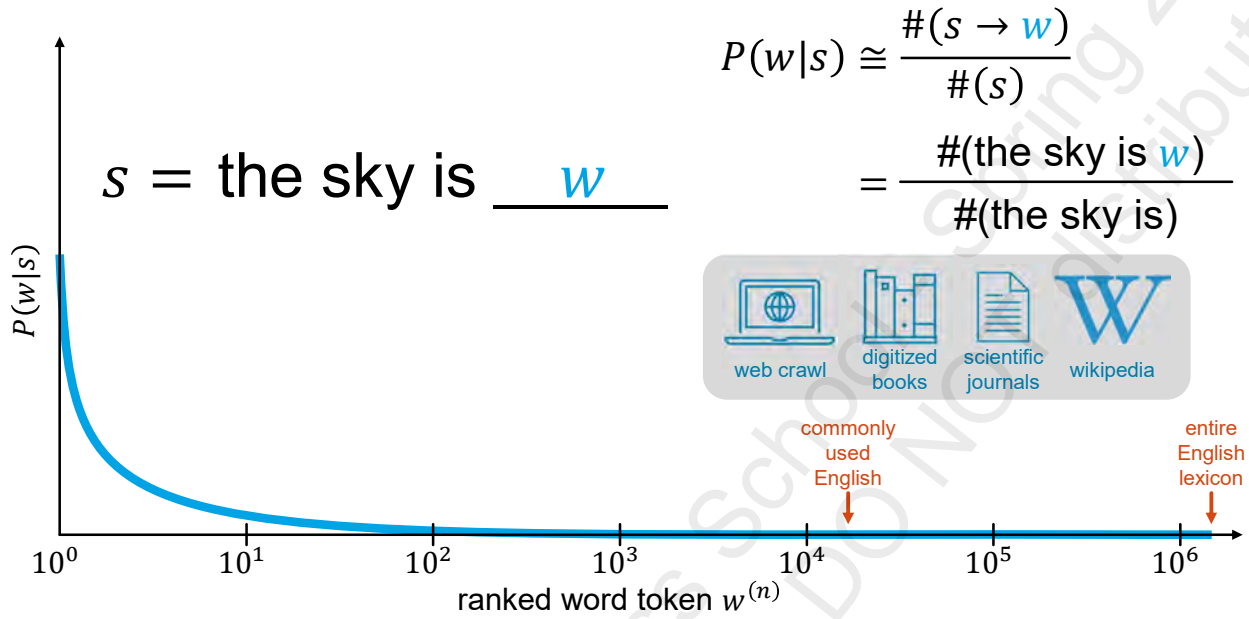
*How do large language models work?*  
*Large language models are a type of machine learning model based on neural networks that process data. To create them, millions of parameters are trained on neural networks.*

25



27

### Do You Know Your English?

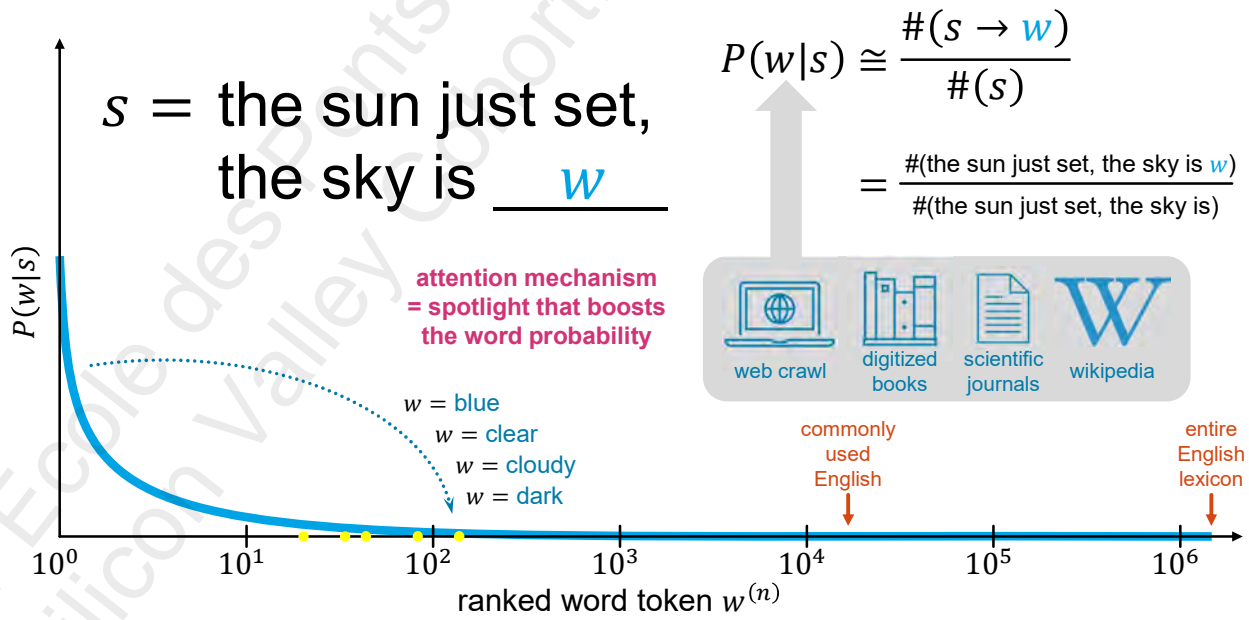


©2024 PROS, Inc. All rights reserved. Confidential and Proprietary.

twitter: @mich8elwu  
linkedin.com/in/MichaelWuPhD

28

### Do You Know Your English?

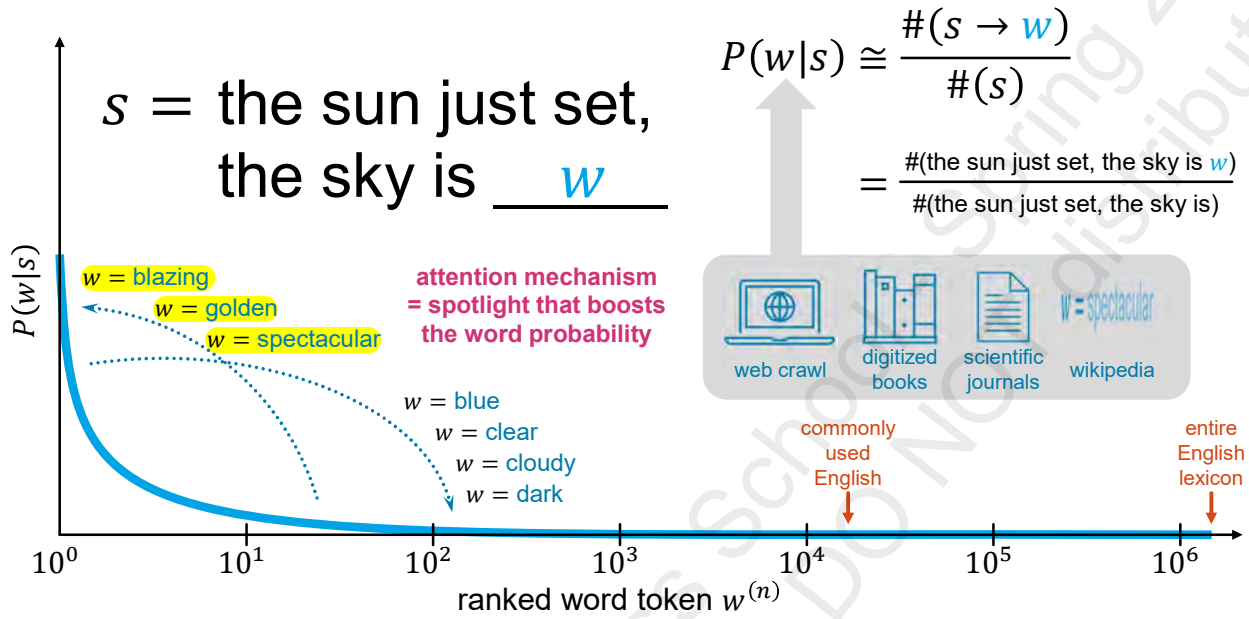


©2024 PROS, Inc. All rights reserved. Confidential and Proprietary.

twitter: @mich8elwu  
linkedin.com/in/MichaelWuPhD

29

### Do You Know Your English?

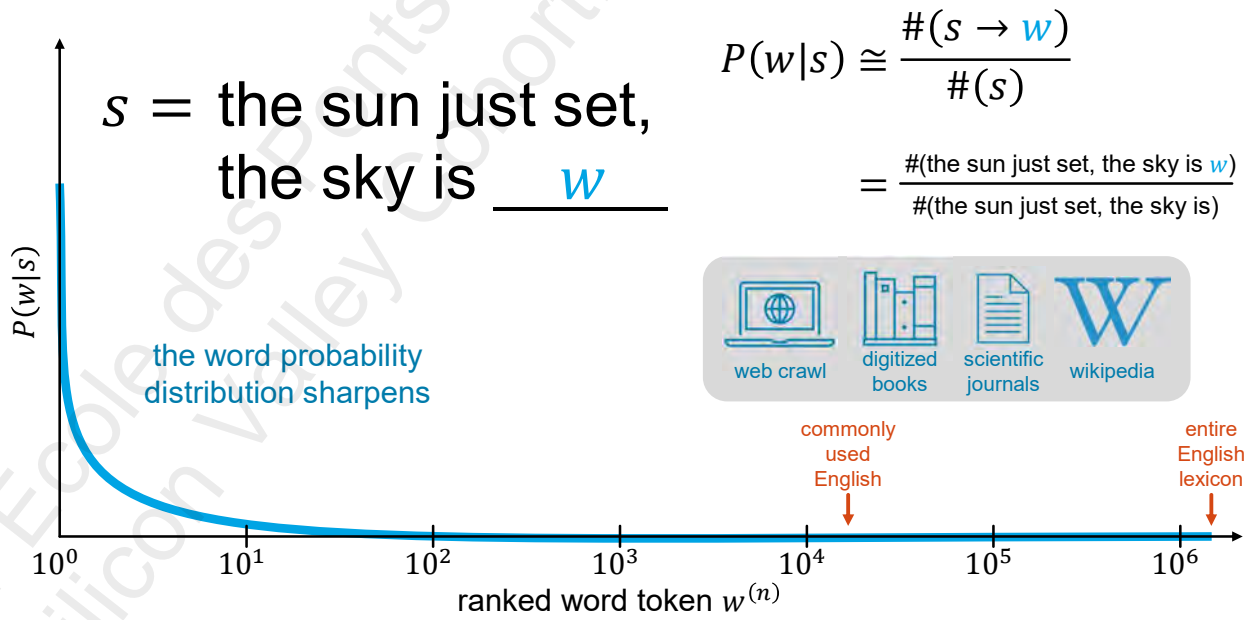


©2024 PROS, Inc. All rights reserved. Confidential and Proprietary.

twitter: @mich8elwu  
linkedin.com/in/MichaelWuPhD

30

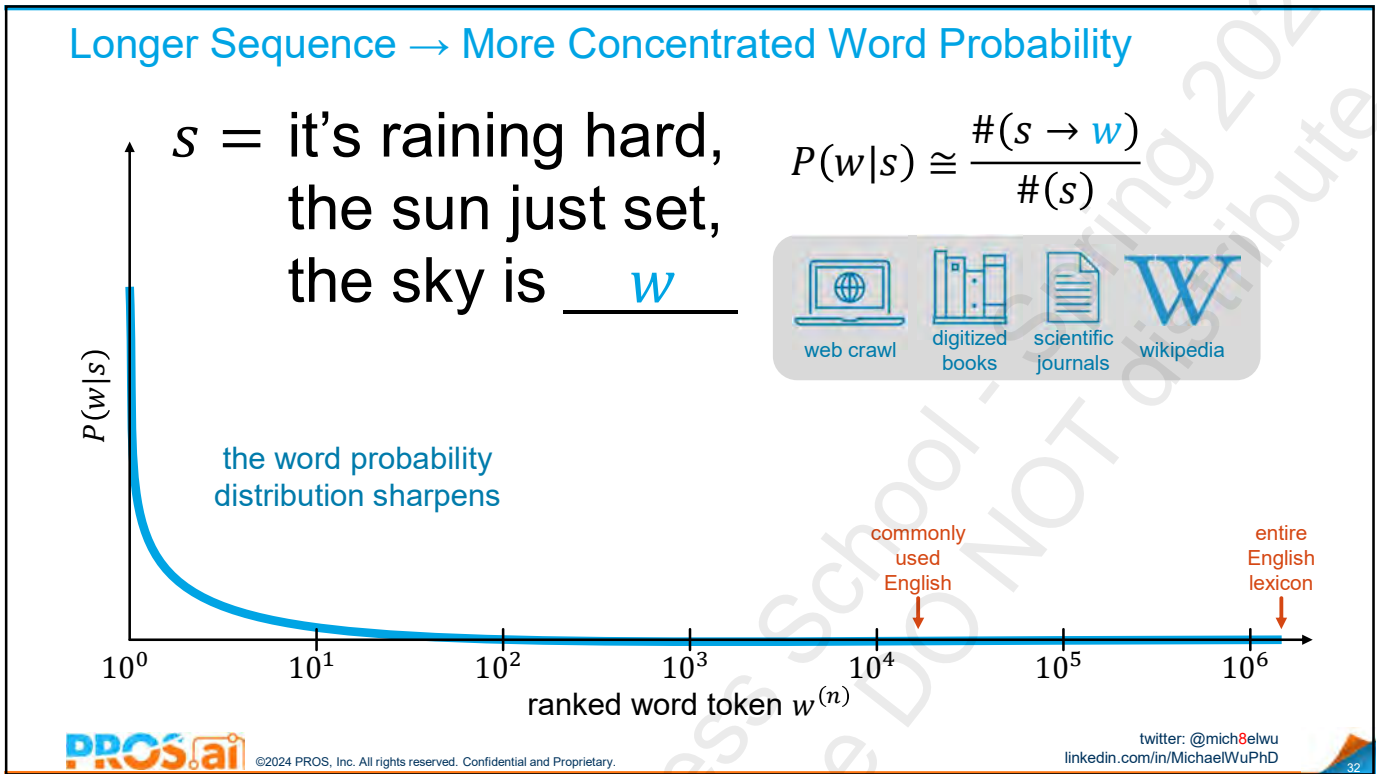
### Longer Sequence → More Concentrated Word Probability



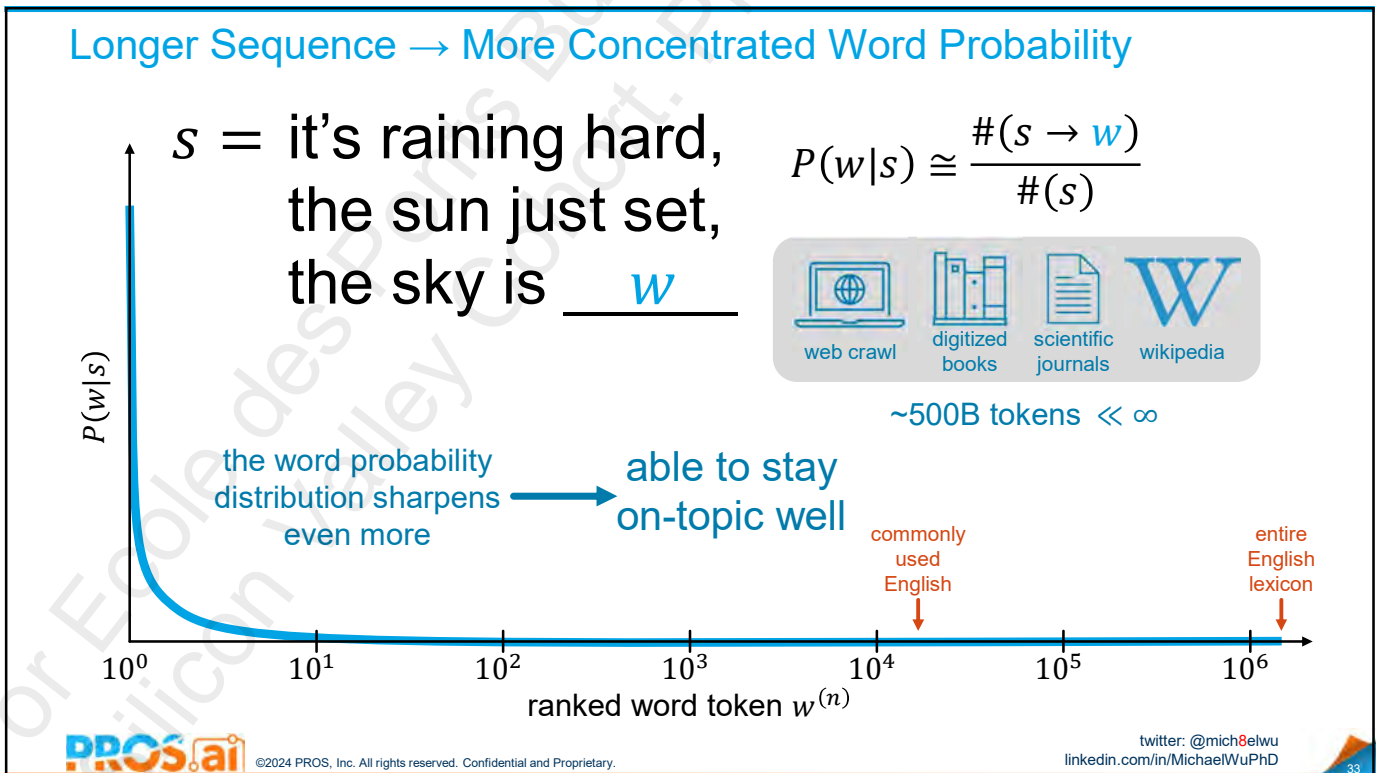
©2024 PROS, Inc. All rights reserved. Confidential and Proprietary.

twitter: @mich8elwu  
linkedin.com/in/MichaelWuPhD

31



32



33



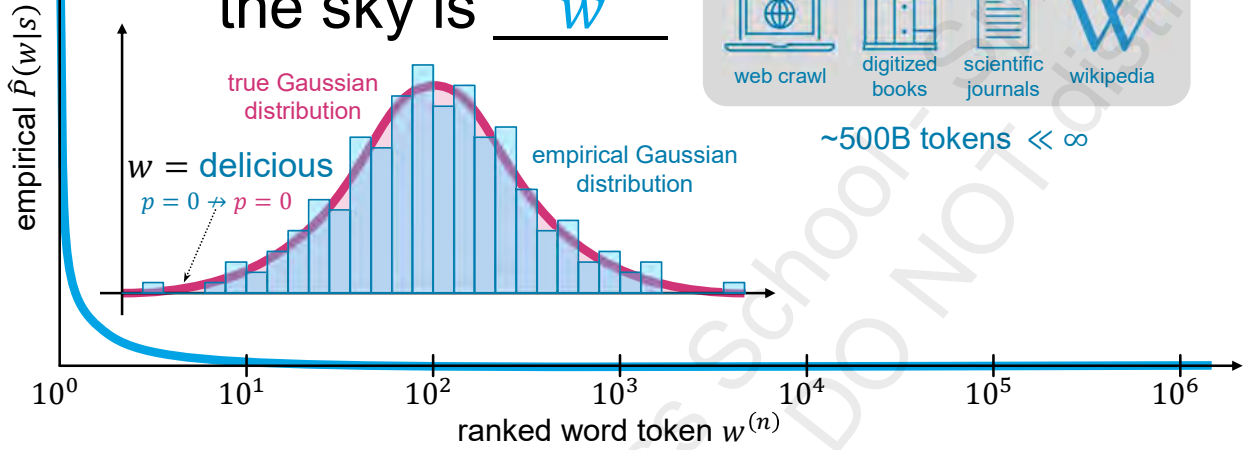
### Empirical Word Probability to Language Model

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

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



~500B tokens  $\ll \infty$



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

twitter: @mich8elwu  
linkedin.com/in/MichaelWuPhD

34



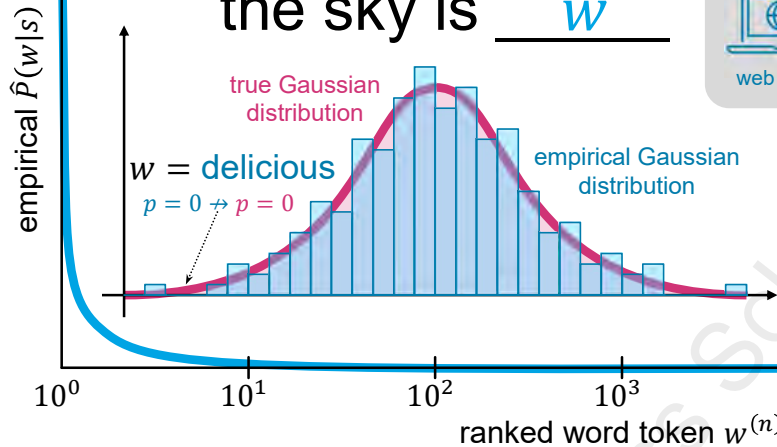
35



## 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  $\ll \infty$

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

the ability to create novel  
content = "generative"

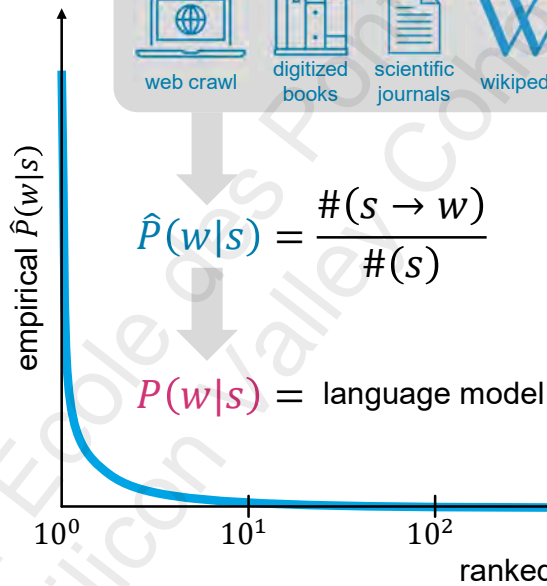


©2024 PROS, Inc. All rights reserved. Confidential and Proprietary.

twitter: @mich8elwu  
linkedin.com/in/MichaelWuPhD

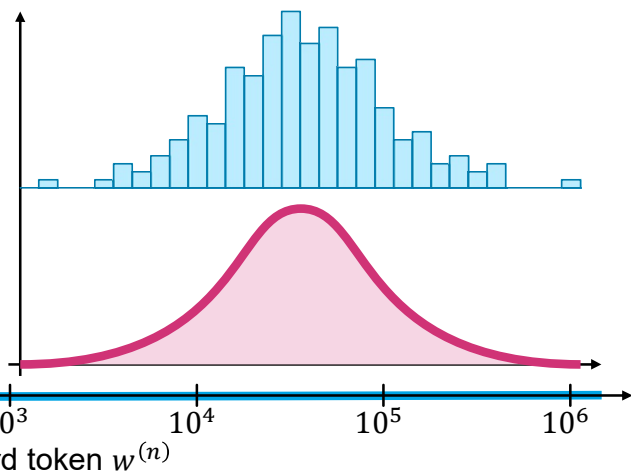
37

## Empirical Word Probability to Language Model



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

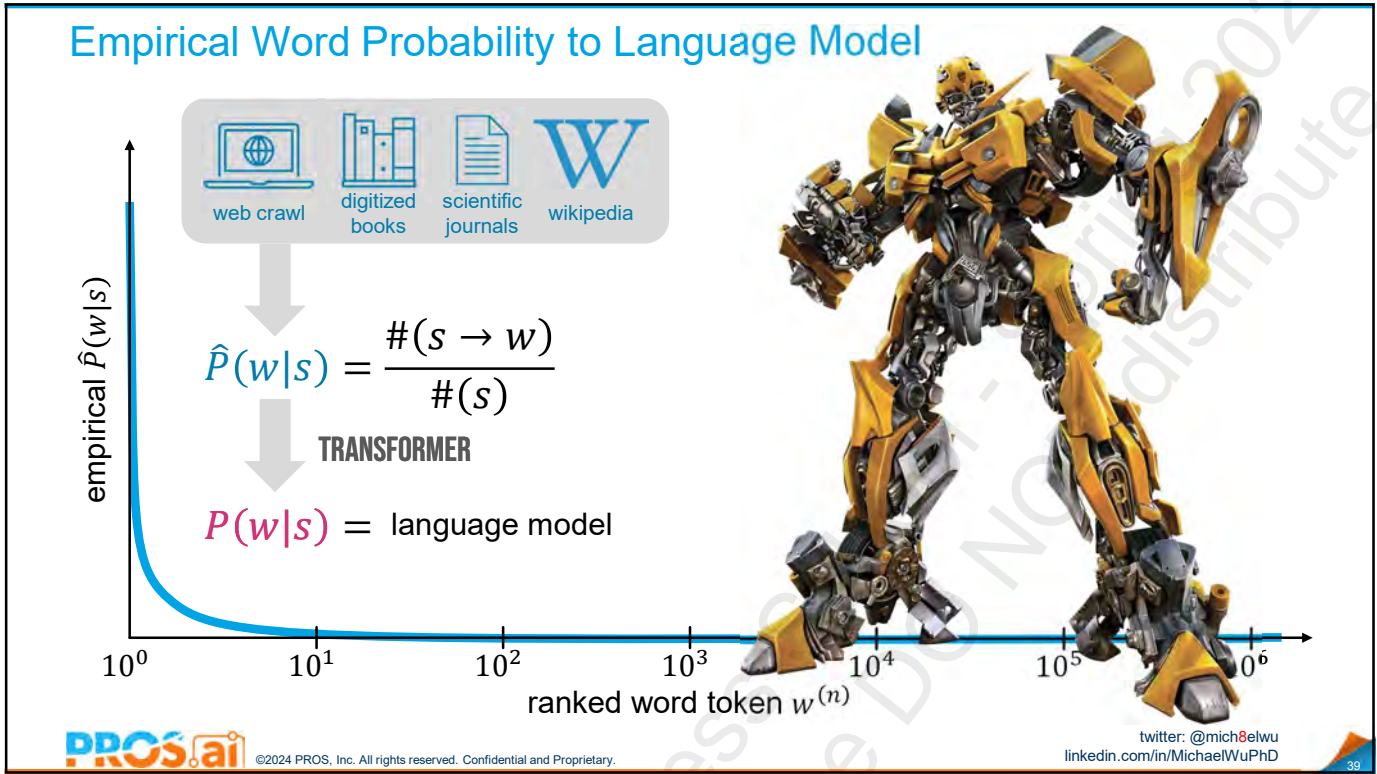
$P(w|s) =$  language model



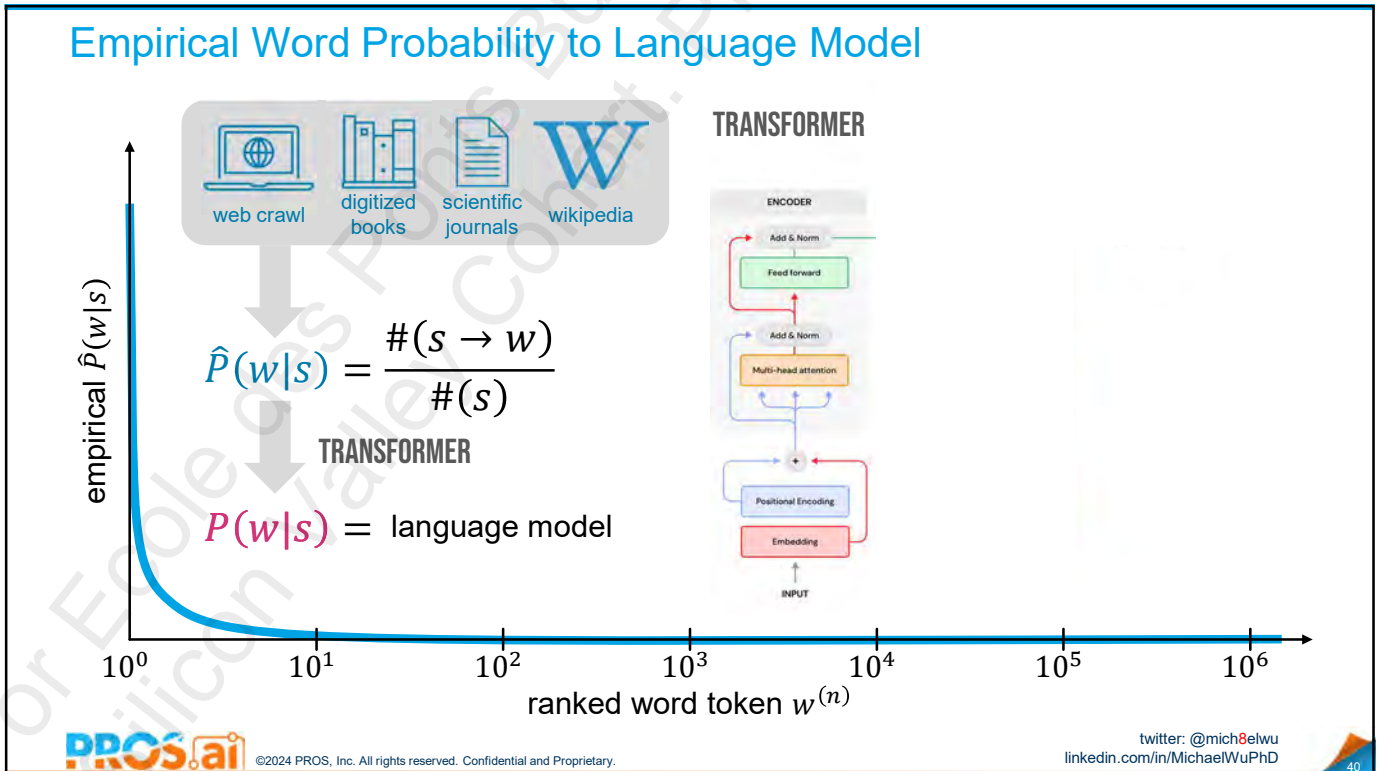
©2024 PROS, Inc. All rights reserved. Confidential and Proprietary.

twitter: @mich8elwu  
linkedin.com/in/MichaelWuPhD

38

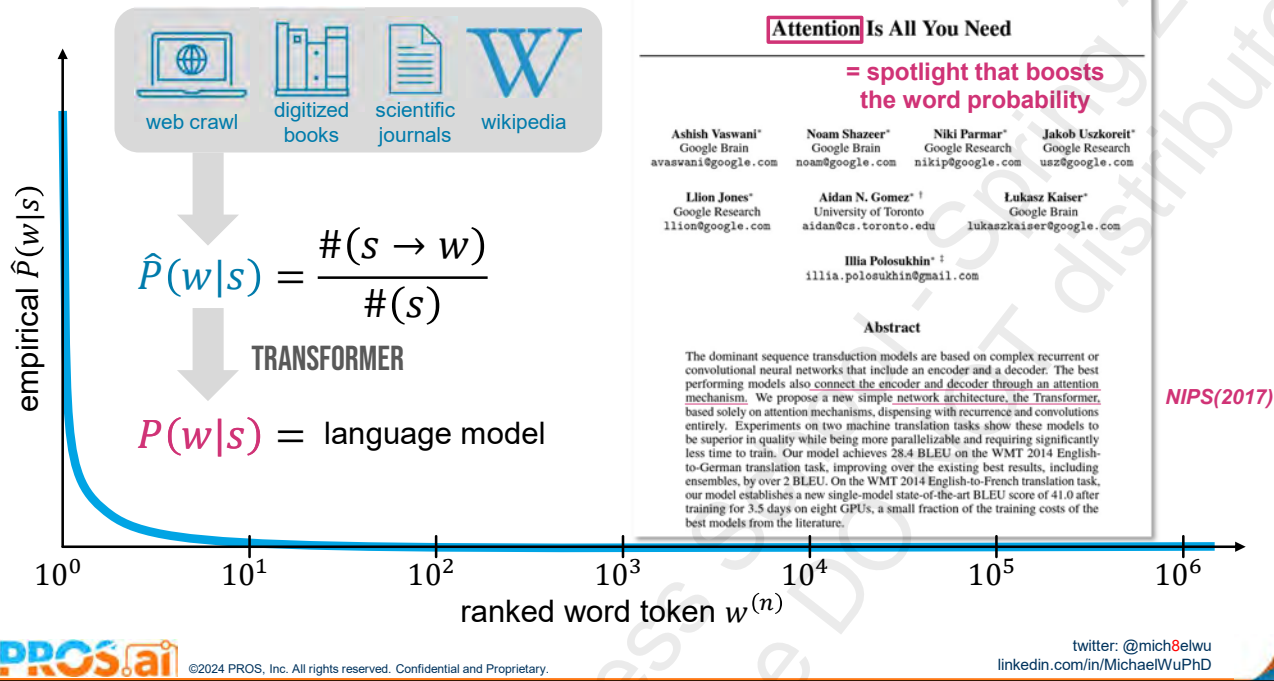


39



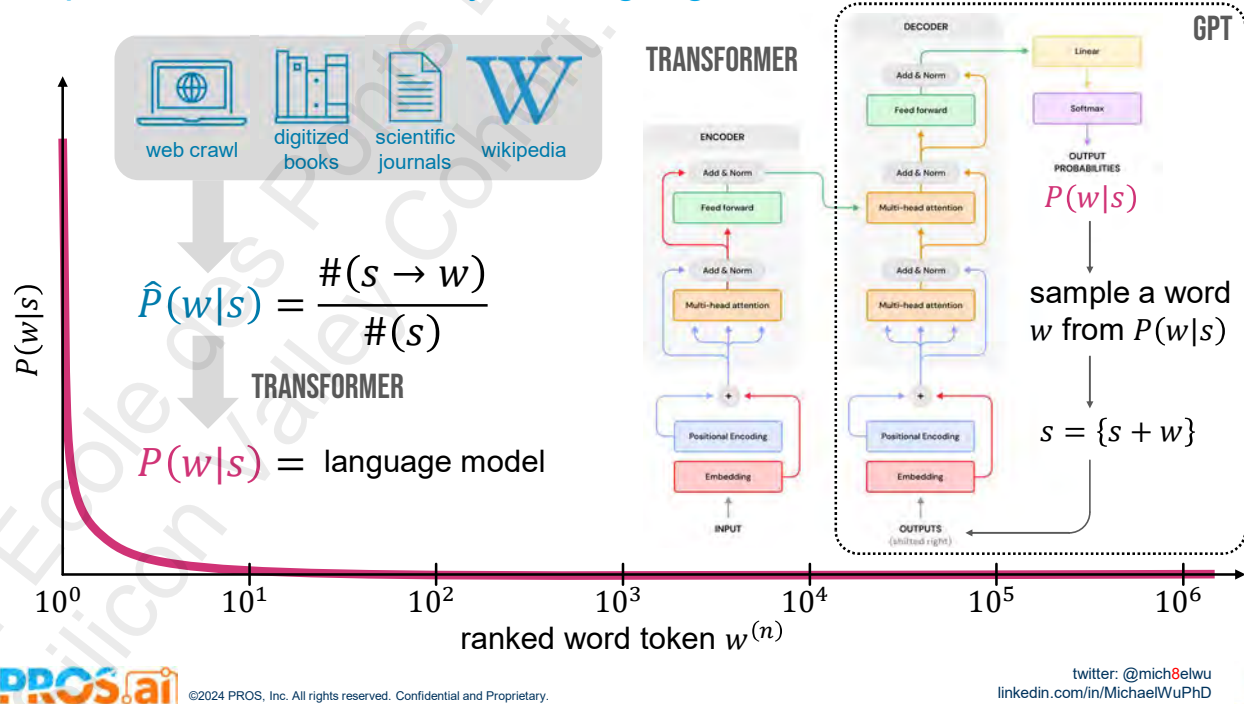
40

## Empirical Word Probability to Language Model



41

## Empirical Word Probability to Language Model



42

## 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

43

doesn't sound  
very intelligent

44





51



52






53



54



hallucination:  
a feature  
or a bug?



56

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



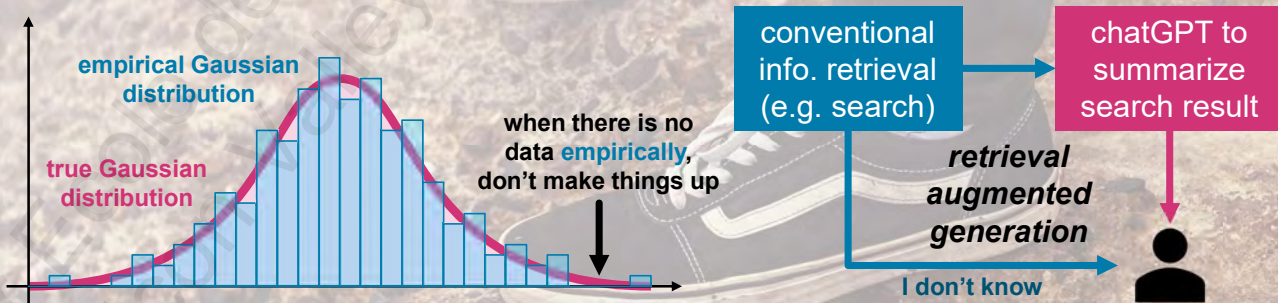
57



# bug: for *fact-based* applications

58

# fact-based applications require *grounding*



59



## 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)

62

## 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?**



64



65

## 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	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

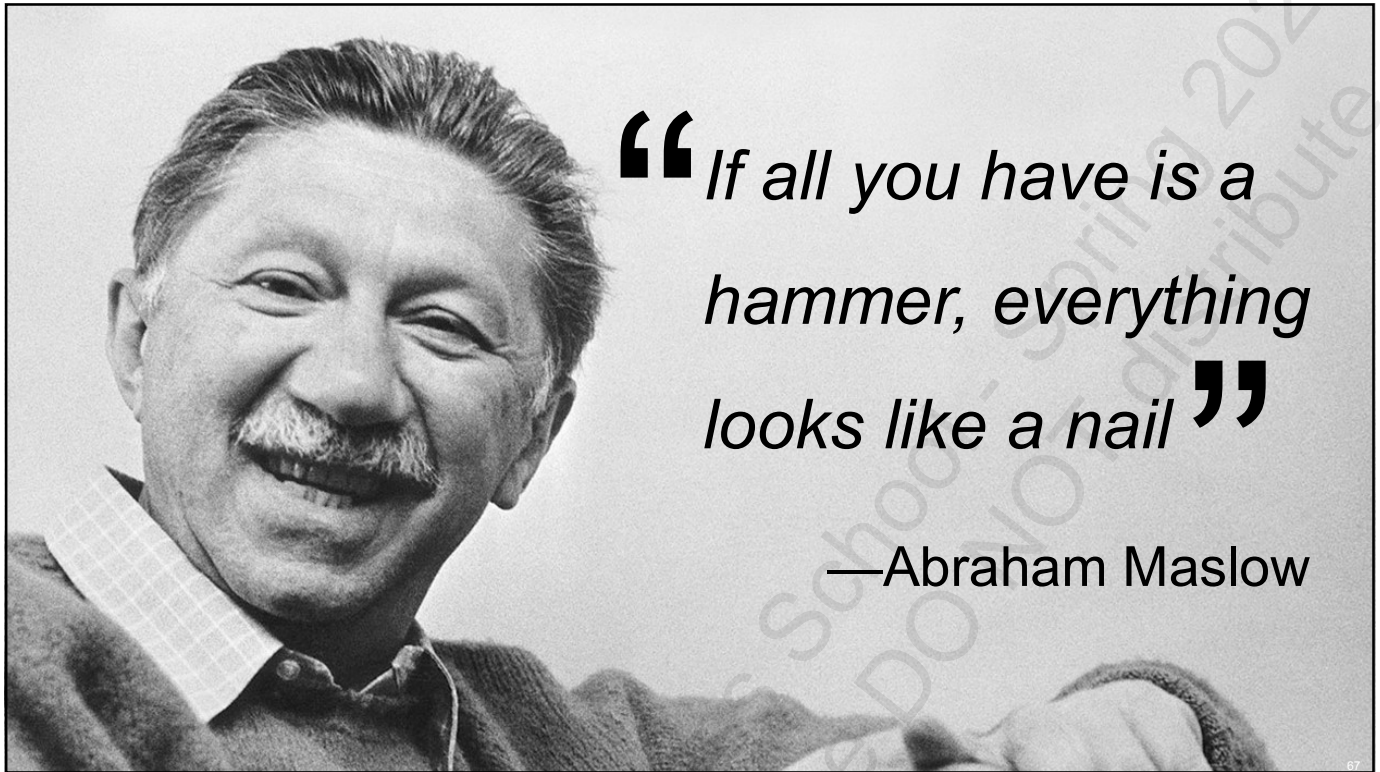


©2024 PROS, Inc. All rights reserved. Confidential and Proprietary.

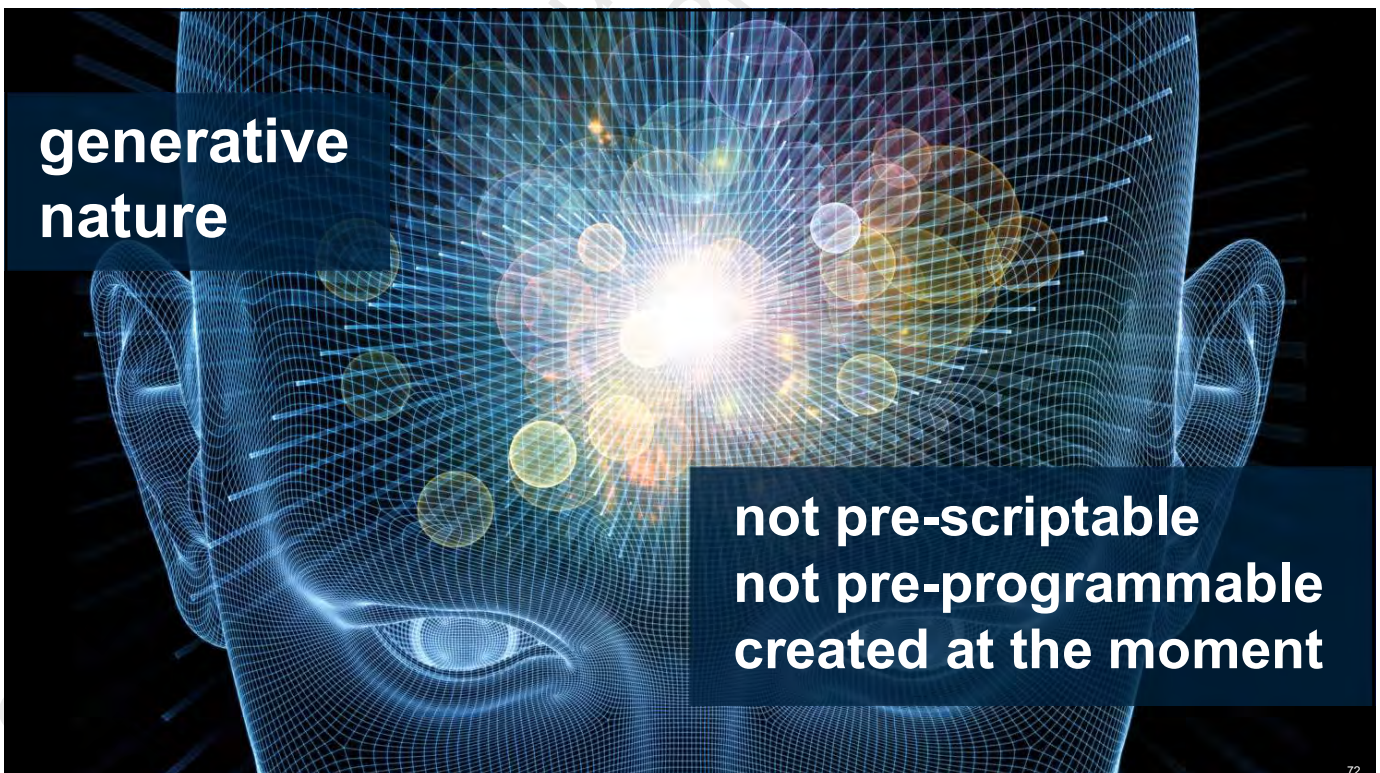
twitter: @mich8elwu  
linkedin.com/in/MichaelWuPhD

66



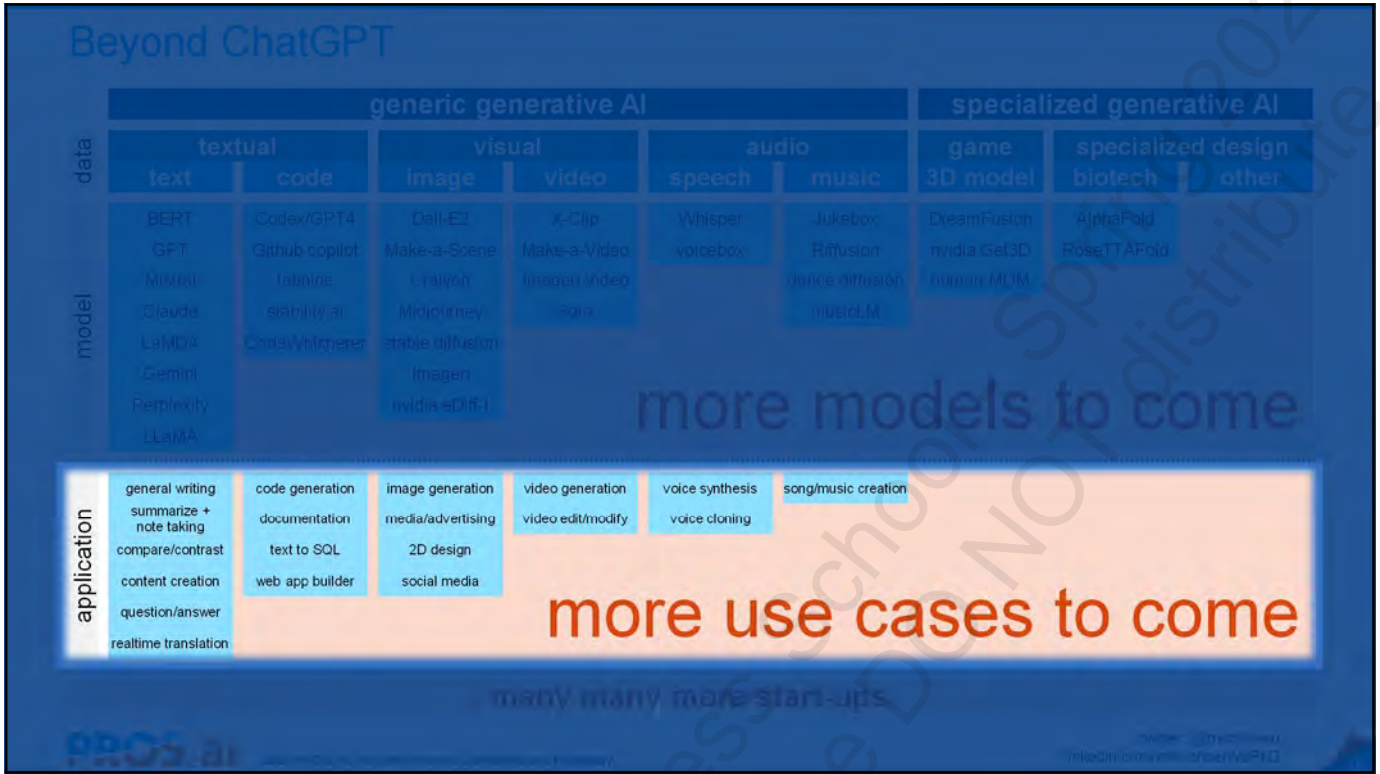


67



72





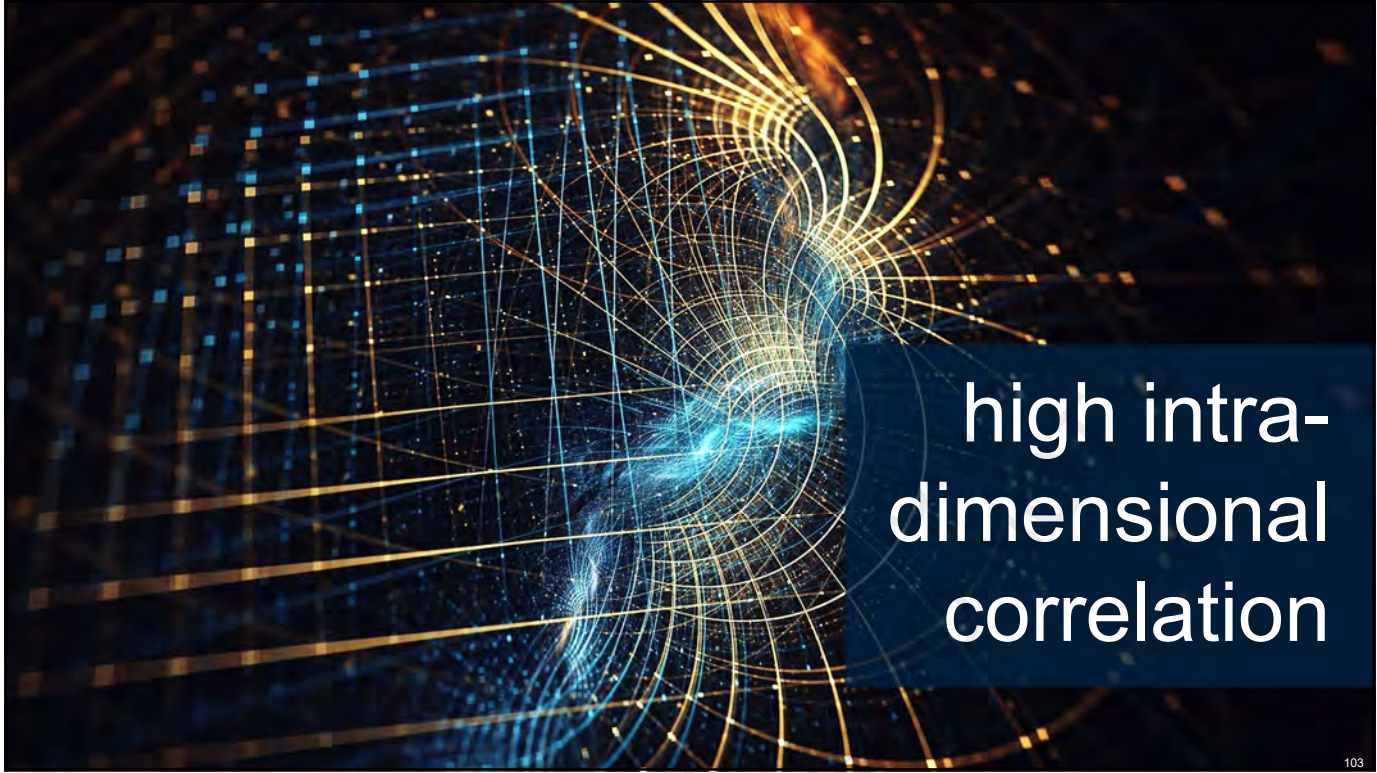
73



102

102

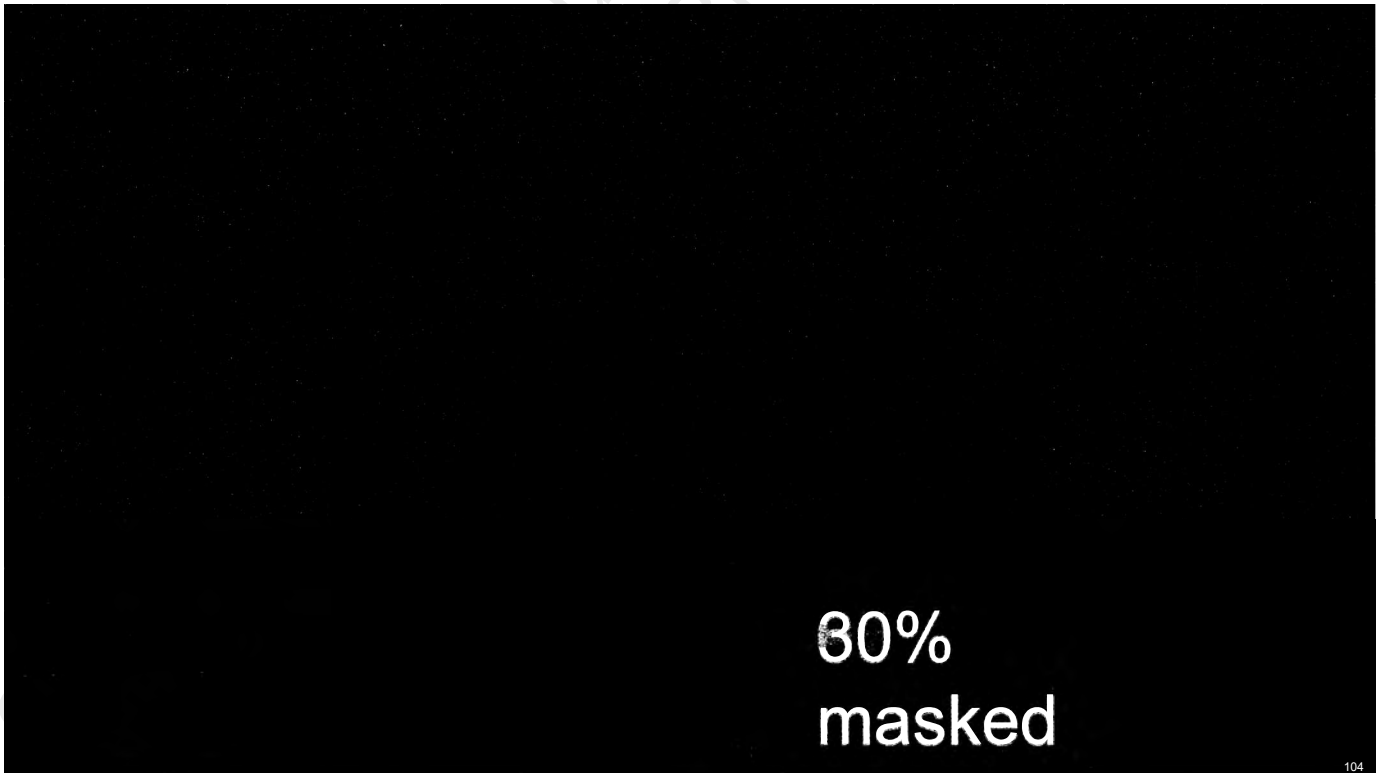




high intra-  
dimensional  
correlation

103

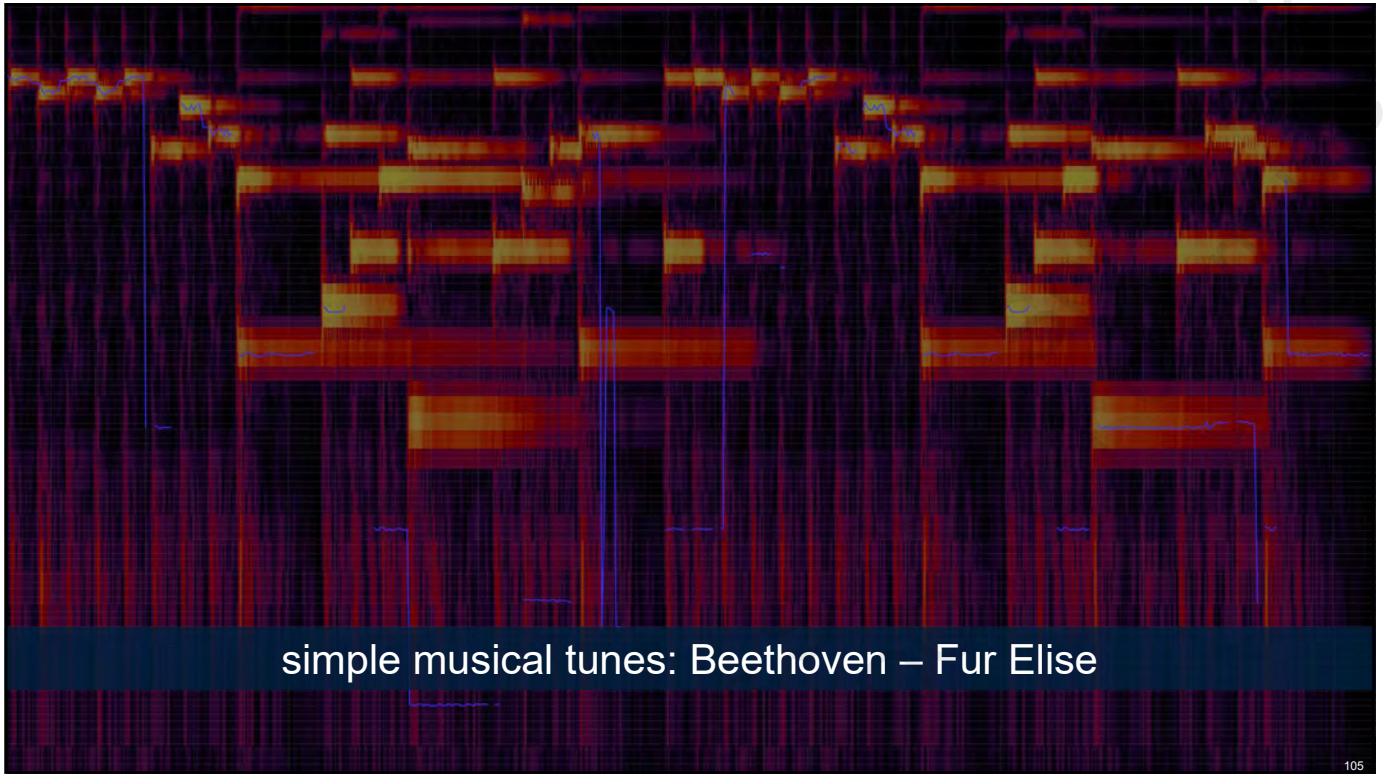
103



60%  
masked

104

104



simple musical tunes: Beethoven – Fur Elise

105



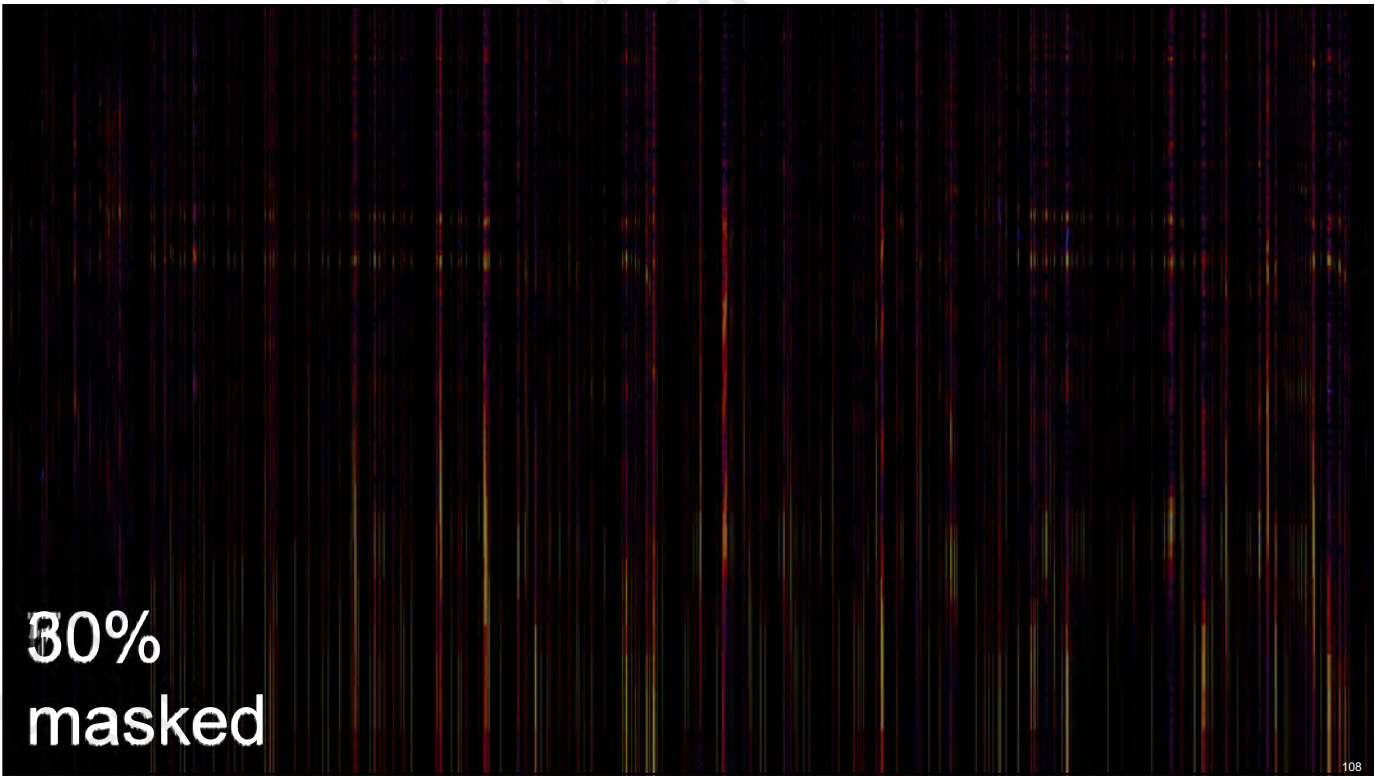
orchestral music: Beethoven – 5th Symphony

106

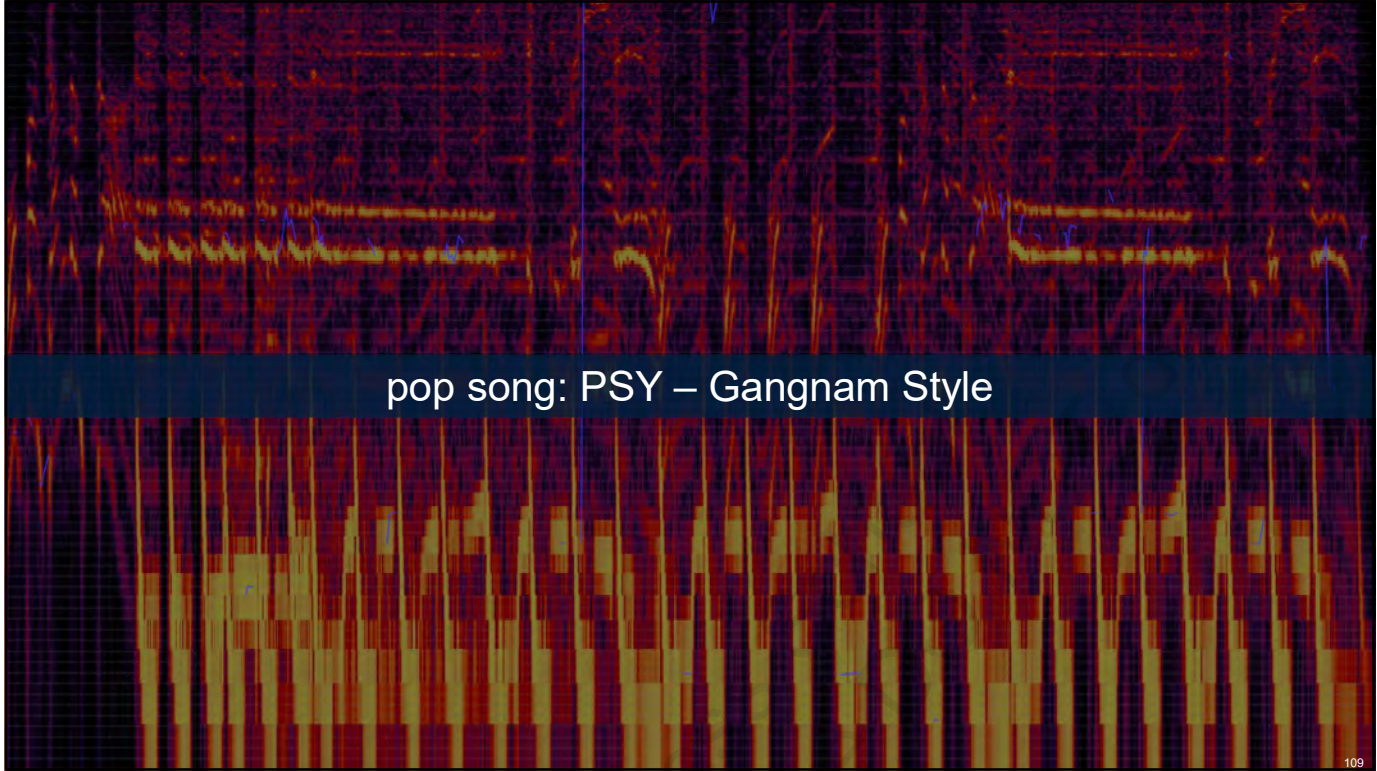




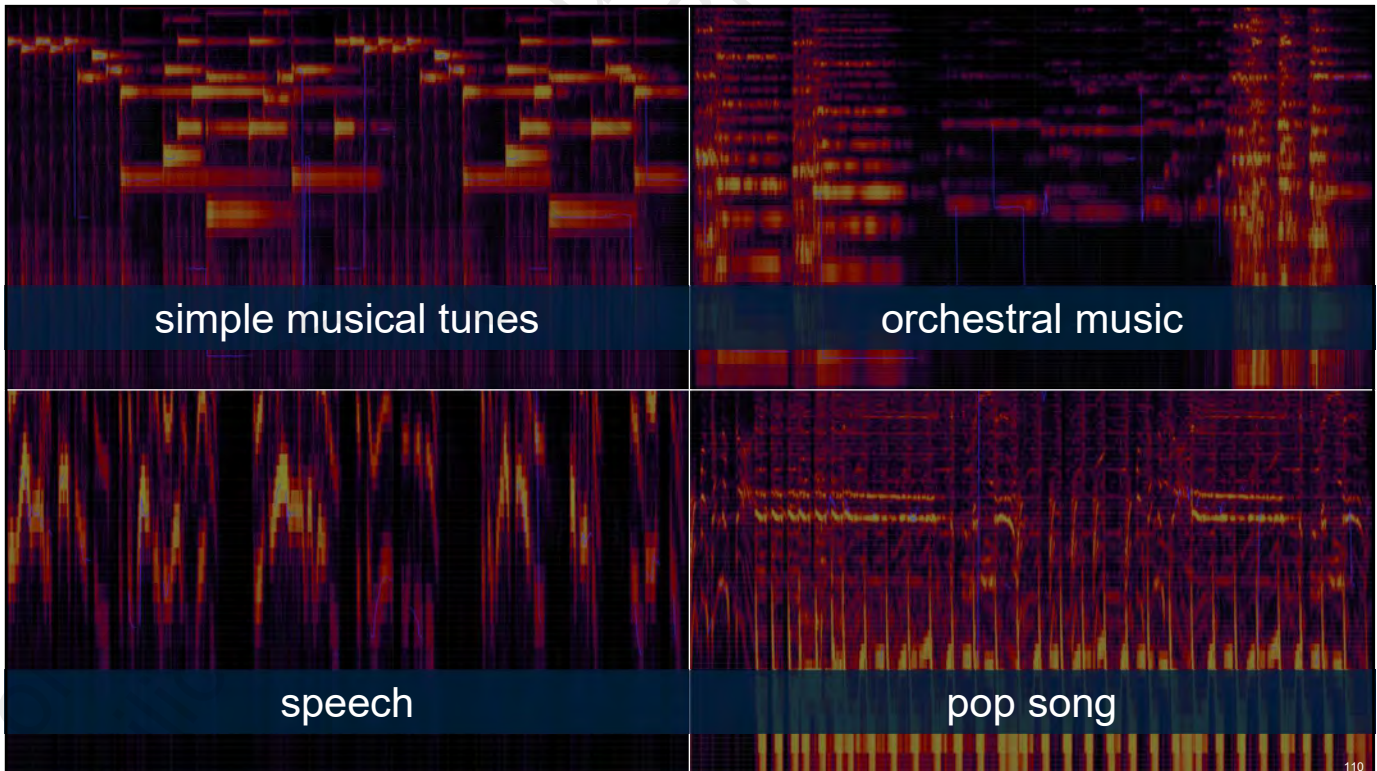
107



108



109



110



# ChatGPT and Beyond

## 50<sup>th</sup> 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**

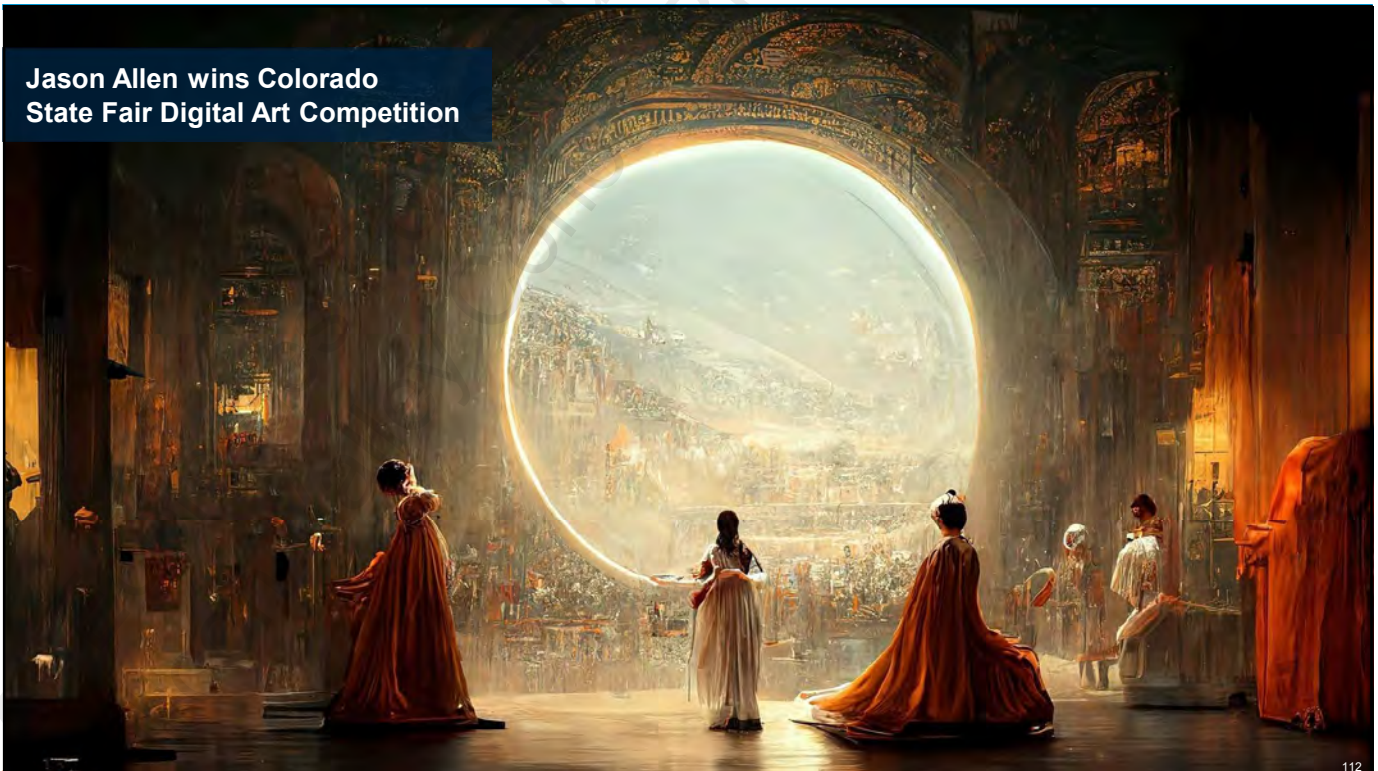


©2024 PROS, Inc. All rights reserved. Confidential and Proprietary.

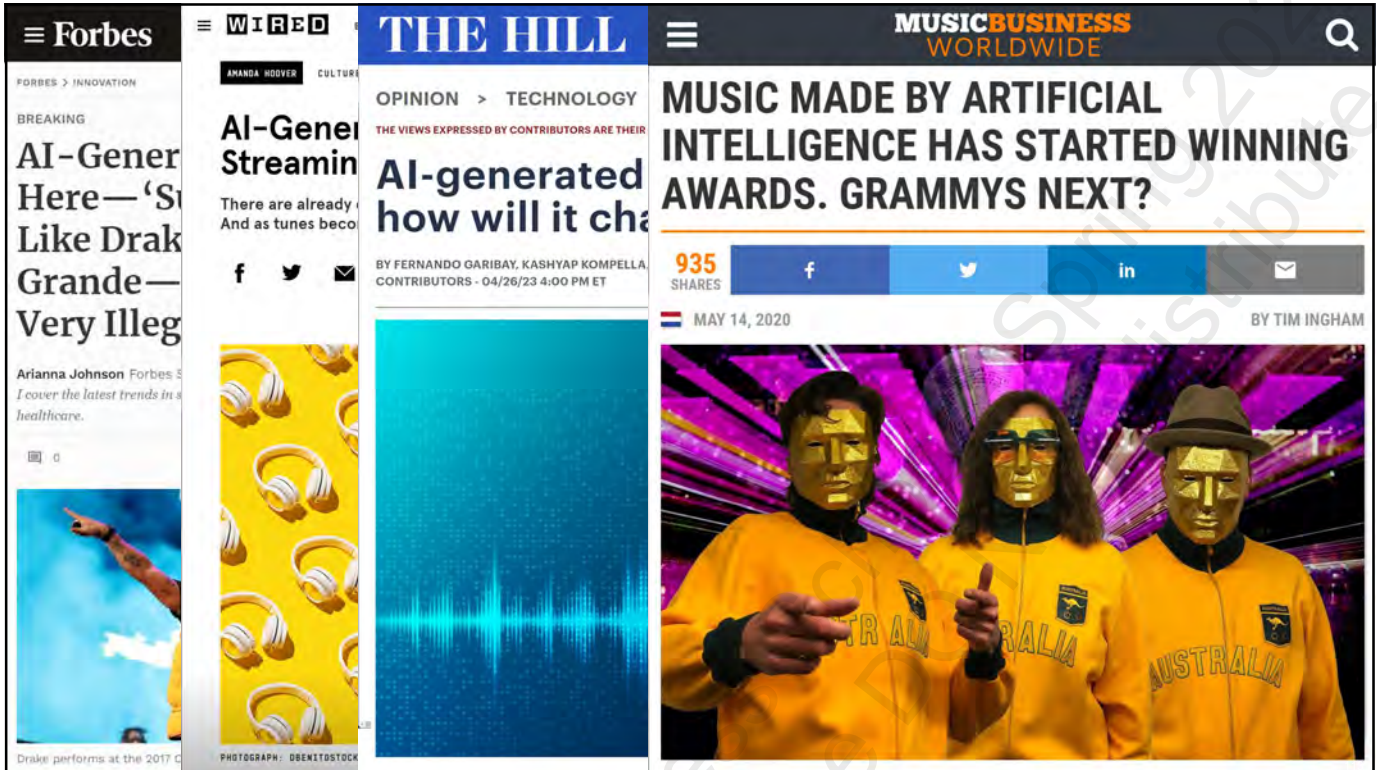
twitter: @mich8elwu  
linkedin.com/in/MichaelWuPhD

111

## Jason Allen wins Colorado State Fair Digital Art Competition



112

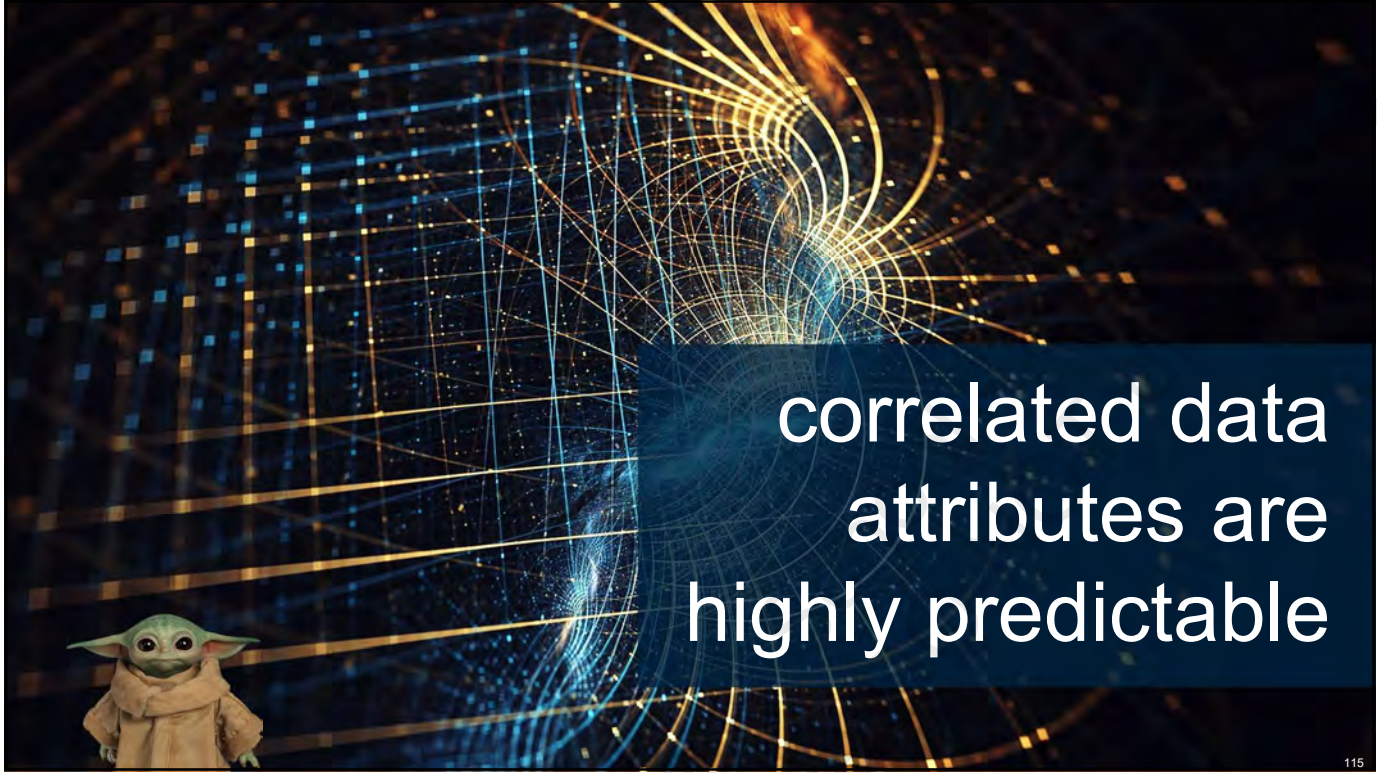


113



114





115



116





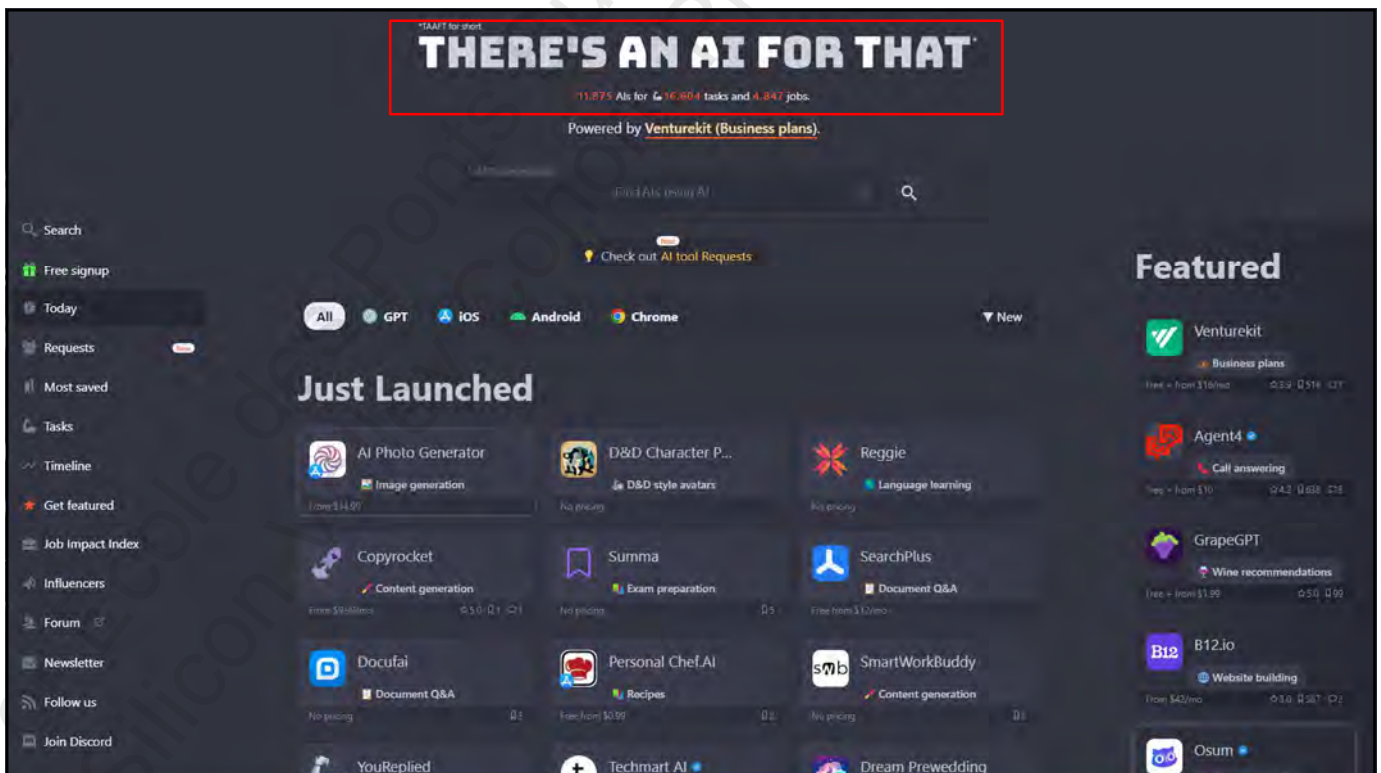
117



118



119



125



A photograph of a business meeting. In the foreground, two people in dark suits are pointing at a large, glowing blue digital display that shows a complex network or data visualization. In the background, another person in a light-colored shirt and tie is visible, looking towards the display. The scene is dimly lit, with the primary light source being the digital display itself, creating a professional and high-tech atmosphere.

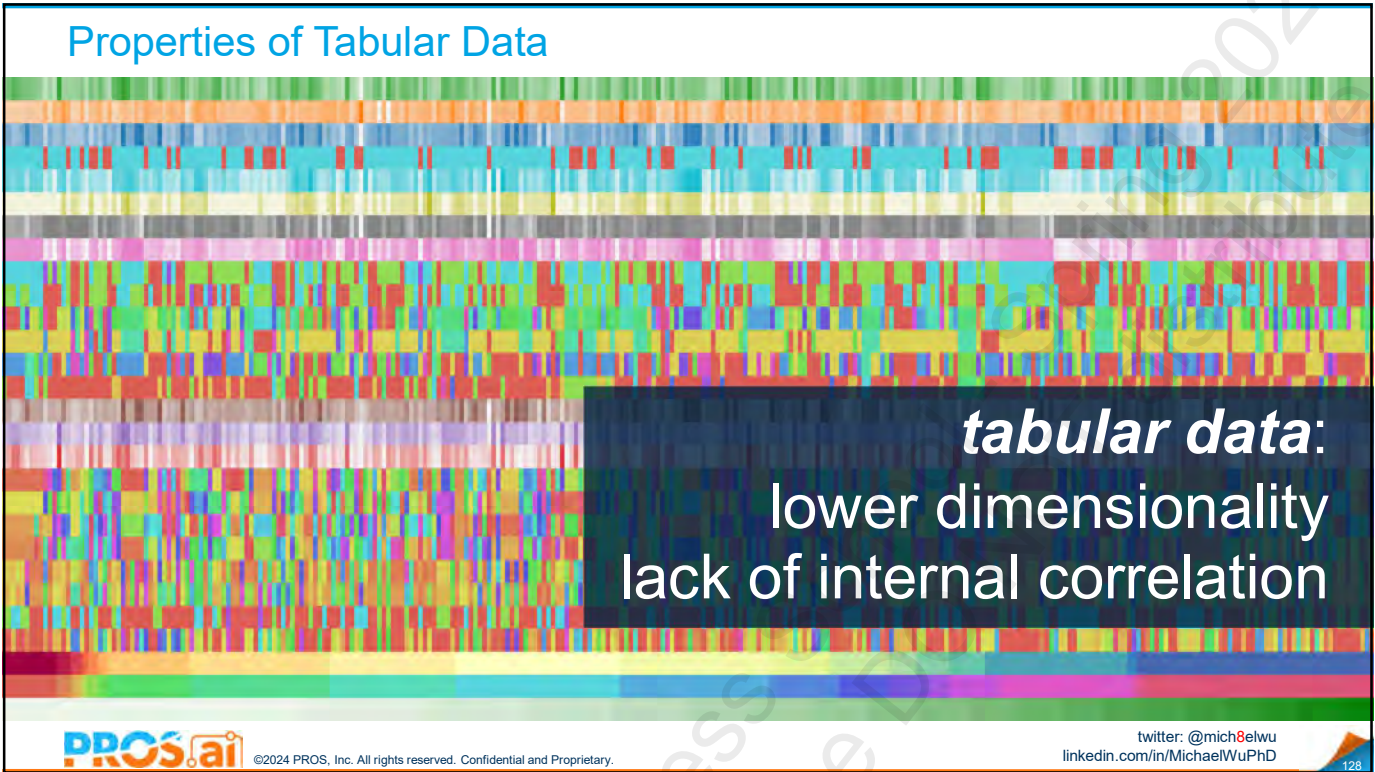
**there aren't many mission-critical business applications based on GenAI**

126

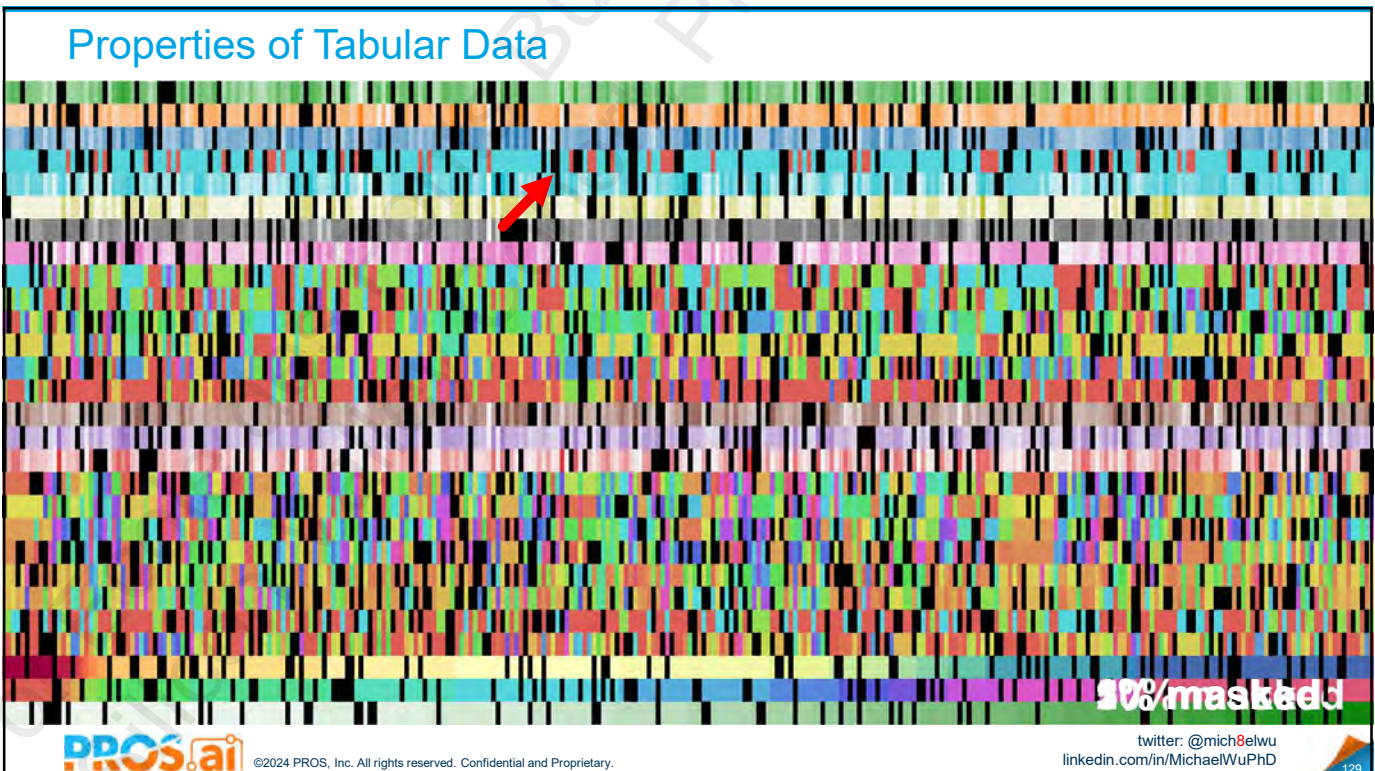
A photograph of a business meeting, identical to the one above. Two people in dark suits are pointing at a large, glowing blue digital display showing a network visualization. A third person in a light shirt and tie is in the background. The lighting is focused on the display, creating a professional and high-tech atmosphere.

**most mission-critical business decisions are made with the support of some sort of tabular data**

127



128



129



what does it takes to  
drive profitable growth?



132

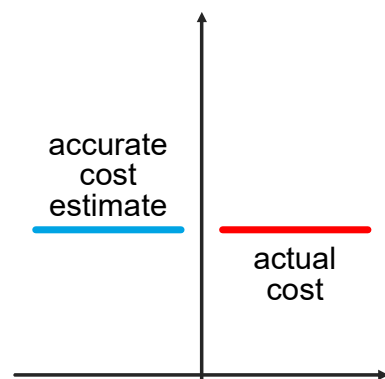
What Does it Take to Drive Profitability?

**profit = revenue – cost**

must have ability to  
manage both terms

optimize revenue

estimate cost



**PROS.ai**

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

twitter: @mich8elwu  
linkedin.com/in/MichaelWuPhD

133

133

### What Does it Take to Drive Profitability?

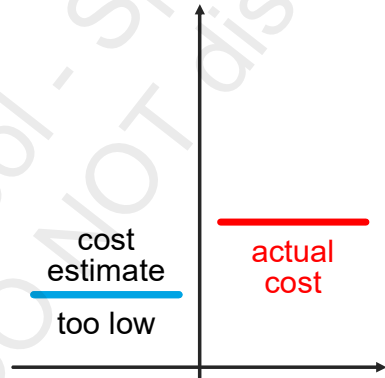
$$\text{profit} = \text{revenue} - \text{cost}$$

must have ability to manage both terms

**optimize revenue**

**estimate cost**

- selling below the **actual cost** → incur losses (negative margin)



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

twitter: @mich8elwu  
linkedin.com/in/MichaelWuPhD

134

### What Does it Take to Drive Profitability?

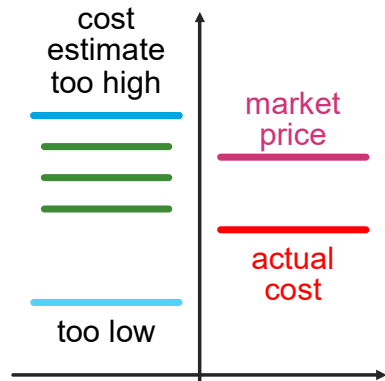
$$\text{profit} = \text{revenue} - \text{cost}$$

must have ability to manage both terms

**optimize revenue**

**estimate cost**

- selling above **market price** → loss of competitiveness
- turn down **deals** that could contribute positive margin → loss of biz opportunities
- selling below the **actual cost** → incur losses (negative margin)



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

twitter: @mich8elwu  
linkedin.com/in/MichaelWuPhD

135



## What Does it Take to Drive Profitability?

# profit = revenue - cost

must have ability to manage both terms

**optimize revenue**

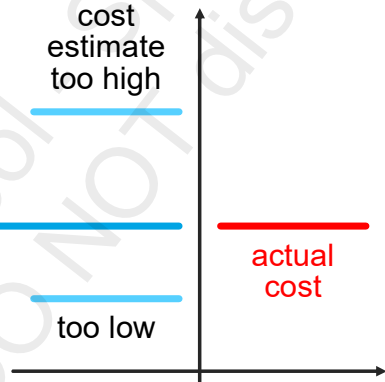
**estimate cost**

- many products/services are not directly linked to production
- industry with supply constrained (e.g. travel, transport, logistics, utilities, perishables goods, etc.)

- selling above **market price** → loss of competitiveness
- turn down **deals** that could contribute positive margin → loss of biz opportunities

material cost (cost of production)  $\neq$  accurate cost estimate NOW

- selling below the **actual cost** → incur losses (negative margin)



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

twitter: @mich8elwu  
linkedin.com/in/MichaelWuPhD

136

136

## What Does it Take to Drive Profitability?

# profit = revenue - cost

must have ability to manage both terms

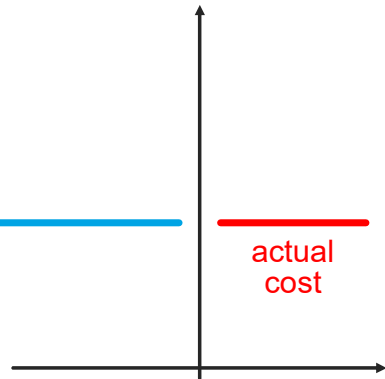
**optimize revenue**

**estimate cost**

- many products/services are not directly linked to production
- industry with supply constrained (e.g. travel, transport, logistics, utilities, perishables goods, etc.)

material cost (cost of production)  $\neq$  accurate cost estimate NOW = opportunity cost

- cost incur by selling NOW, because you can't sell it later (b/c supply is constrained)
- cost based on *current* market demand + biz environment



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

twitter: @mich8elwu  
linkedin.com/in/MichaelWuPhD

137

137

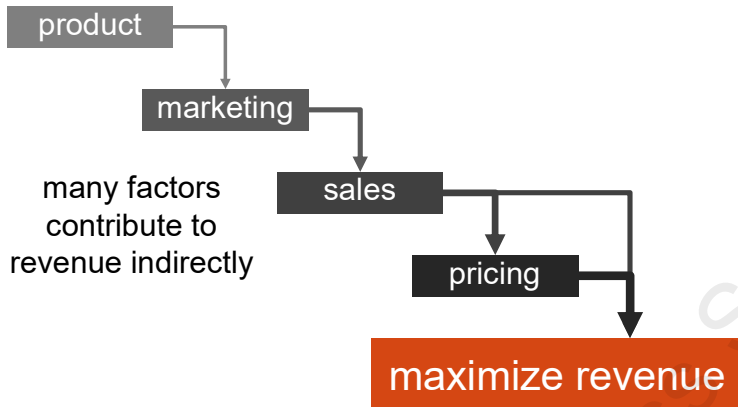
## What Does it Take to Drive Profitability?

$$\text{profit} = \text{revenue} - \text{cost}$$

must have ability to manage both terms

**optimize revenue**

**estimate cost**



- pricing = most direct driver for revenue
- 1% price change
  - ▶ 11% margin improvement
  - ▶ more impact than 1% change in anything else about the business
- best way to maximize revenue = **optimize price**



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

twitter: @mich8elwu  
linkedin.com/in/MichaelWuPhD

139

139



146



# Can we Use GenAI in Profit Optimization?



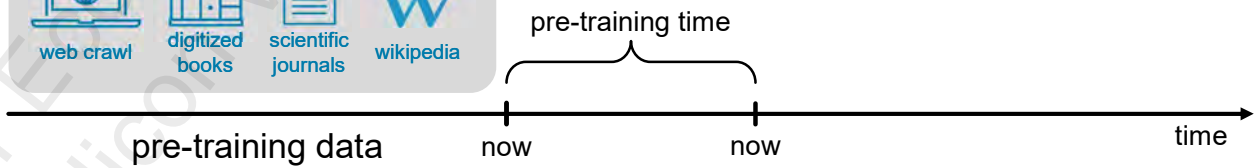
©2024 PROS, Inc. All rights reserved. Confidential and Proprietary.

twitter: @mich8elwu  
linkedin.com/in/MichaelWuPhD

147

147

# Can we Use GenAI in Profit Optimization?



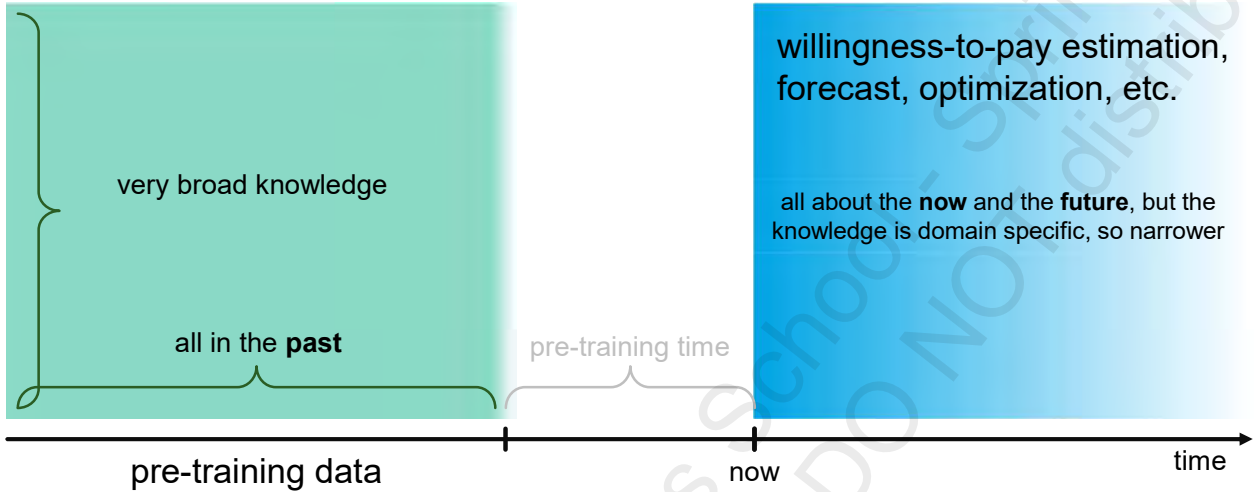
©2024 PROS, Inc. All rights reserved. Confidential and Proprietary.

twitter: @mich8elwu  
linkedin.com/in/MichaelWuPhD

148

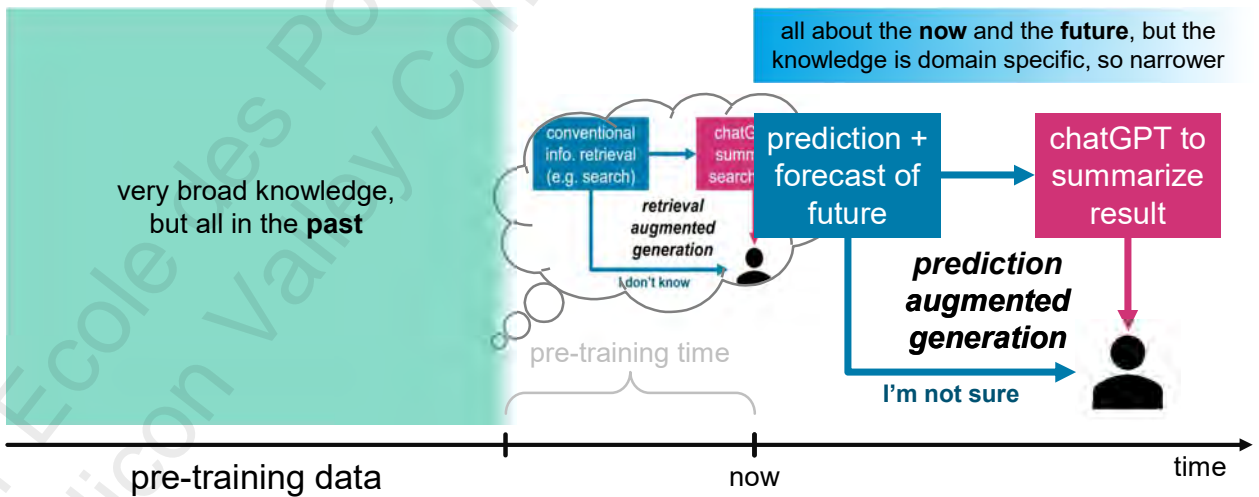
148

### ChatGPT and PROS AI are Very Complementary



149

### ChatGPT and PROS AI are Very Complementary



150





151

A promotional graphic with a blue background. On the left, a 3D rendered character wearing a grey suit, a white shirt, a blue tie, and a black beret points with a white marker towards a white sign on a tripod stand. The sign contains the following text: "QUANTUM SIMPLEX" in large, bold, black letters, with a diamond and triangle symbol to the right. Below this, it says "Perplexity in Plain Words". At the bottom of the sign, it reads "connect w/ me: @mich8elwu" and "linkedin.com/in/MichaelWuPhD". To the right of the sign is a QR code with a blue LinkedIn logo in the center. At the bottom left of the graphic is the "PROS.ai" logo and the text "©2024 PROS, Inc. All rights reserved. Confidential and Proprietary." At the bottom right, it says "twitter: @mich8elwu" and "linkedin.com/in/MichaelWuPhD".

155