Decoding GenAI: The Science Behind the Magic of ChatGPT

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

2024.04.18

1869

1

For attendees of the Decoding GenAI lecture series at the Bulgarian Academy of Science. Please DO NOT Distribute.

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

2024.04.18

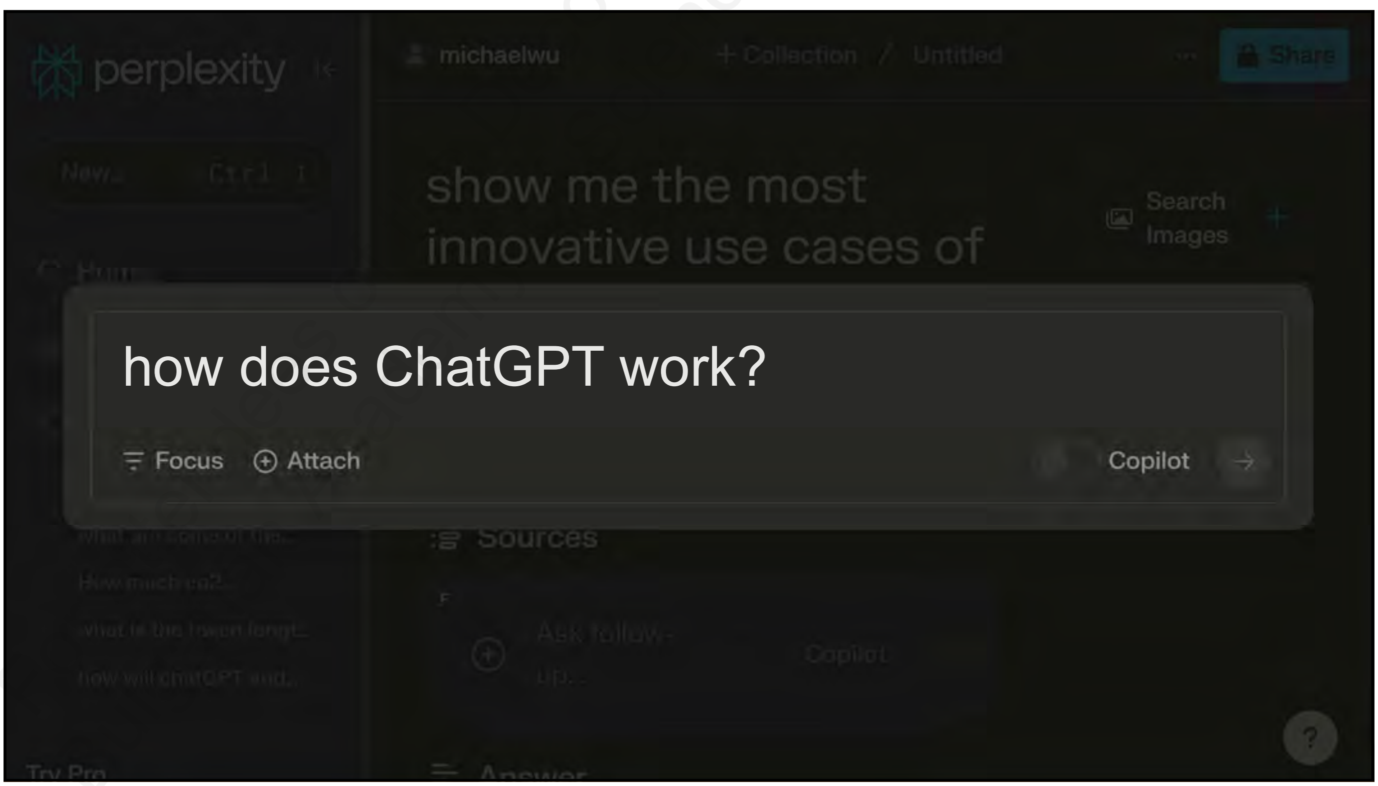
©2024 PROS, Inc. All rights reserved. Confidential and Proprietary.

10

10

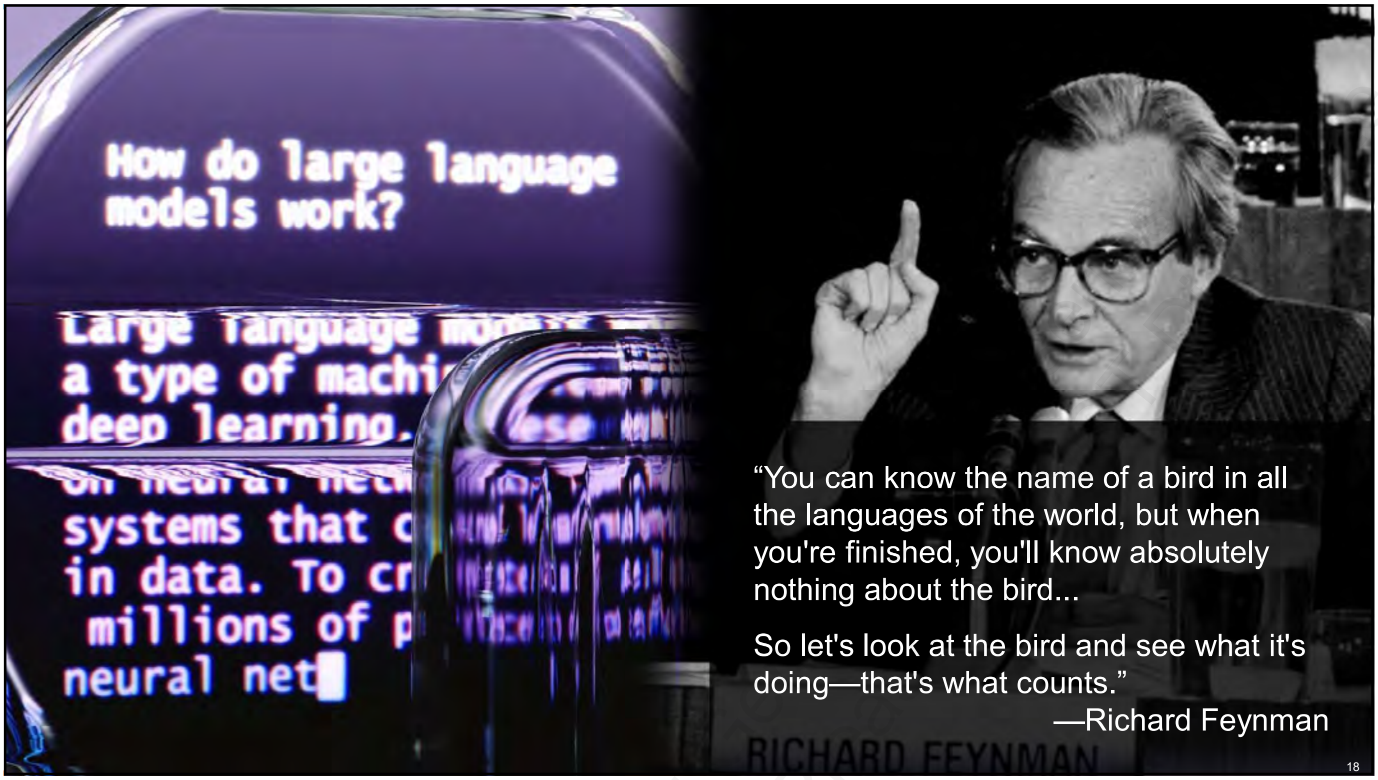
• difficult topic
• only ~75 min
• interactive
• pay attention, have fun, learn
• don't take notes, get pdf slide via linkedin
can I have your cooperation?
11

11

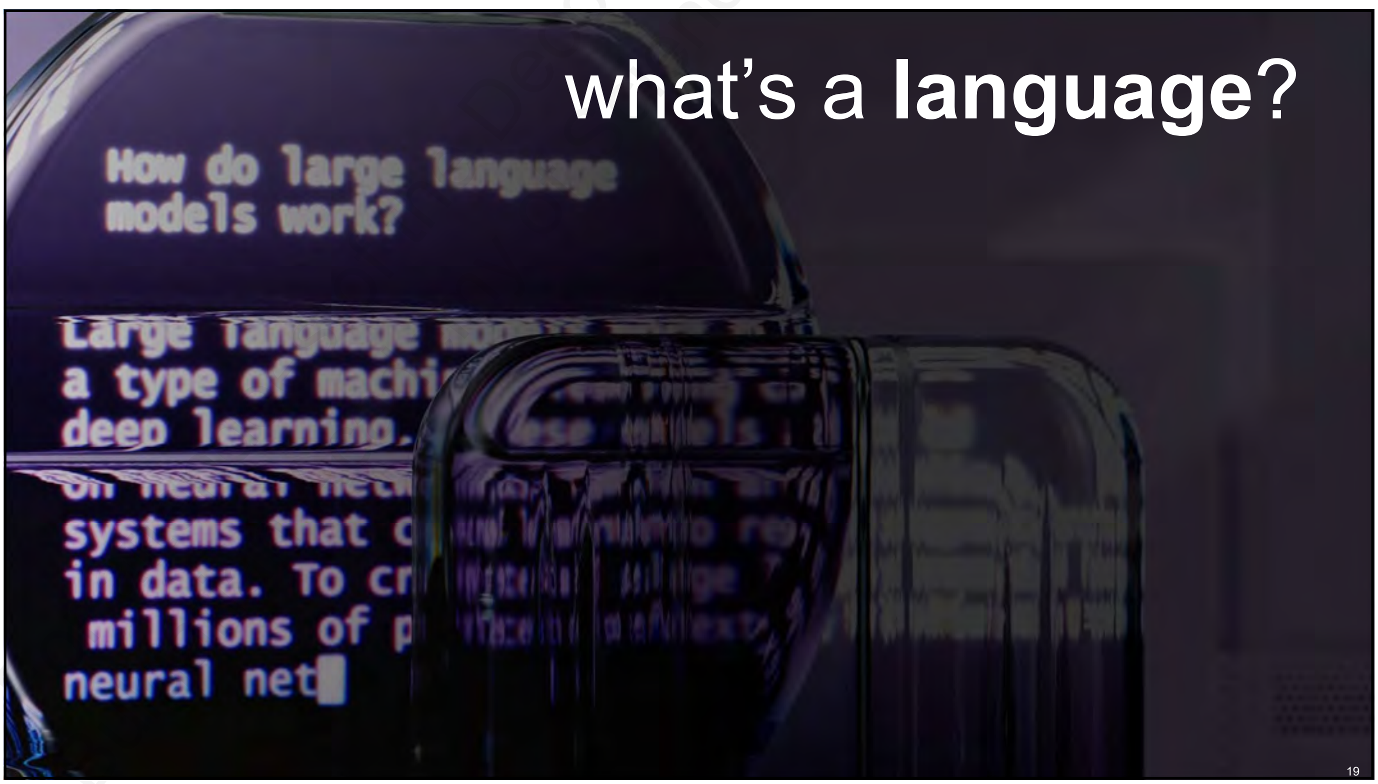


perplexity
michaelwu
+ Collection / Untitled
Share
show me the most innovative use cases of
Search Images
how does ChatGPT work?
Focus
Attach
Copilot
Sources
Ask follow-up...
Copilot
Answer
Try Pro

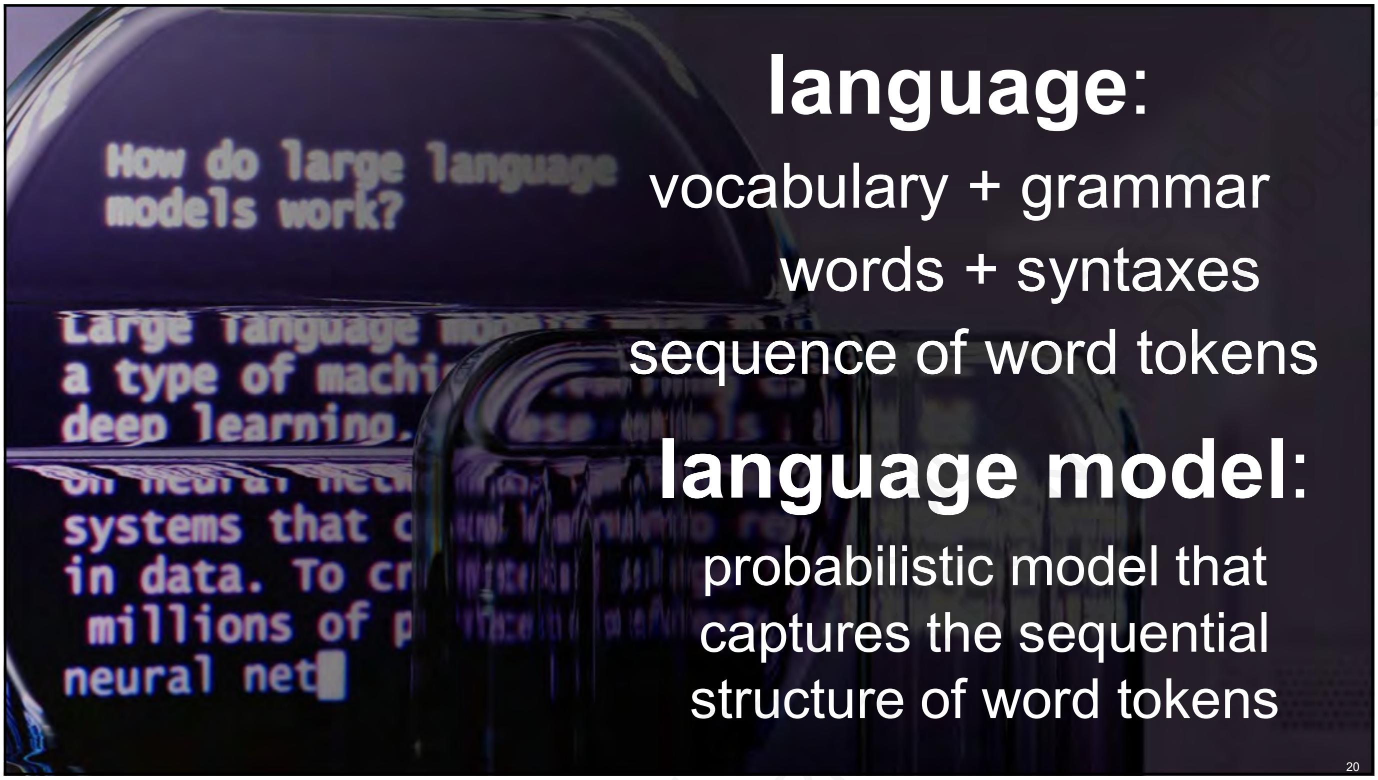
16


How do large language models work?
“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
RICHARD FEYNMAN
18

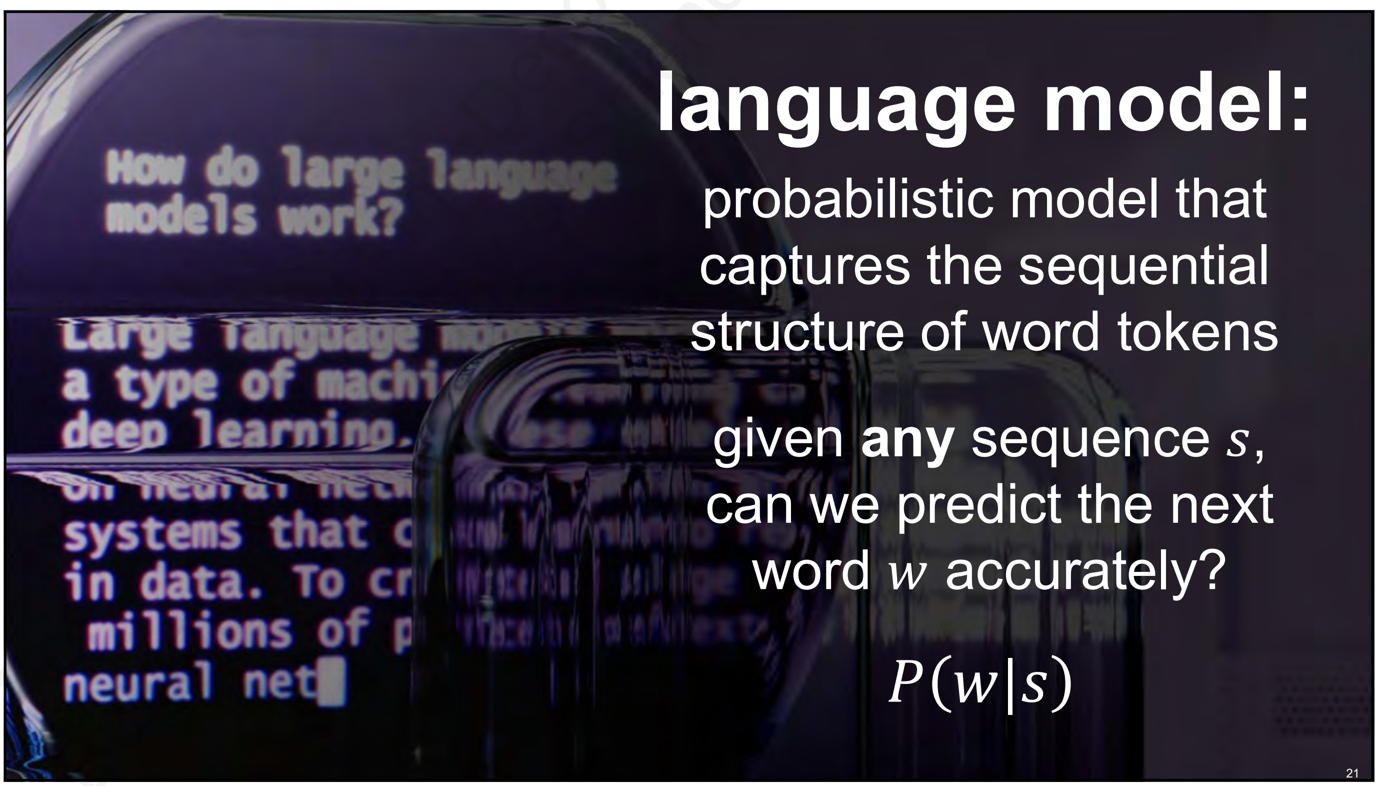
18


what’s a **language**?
How do large language models work?
19

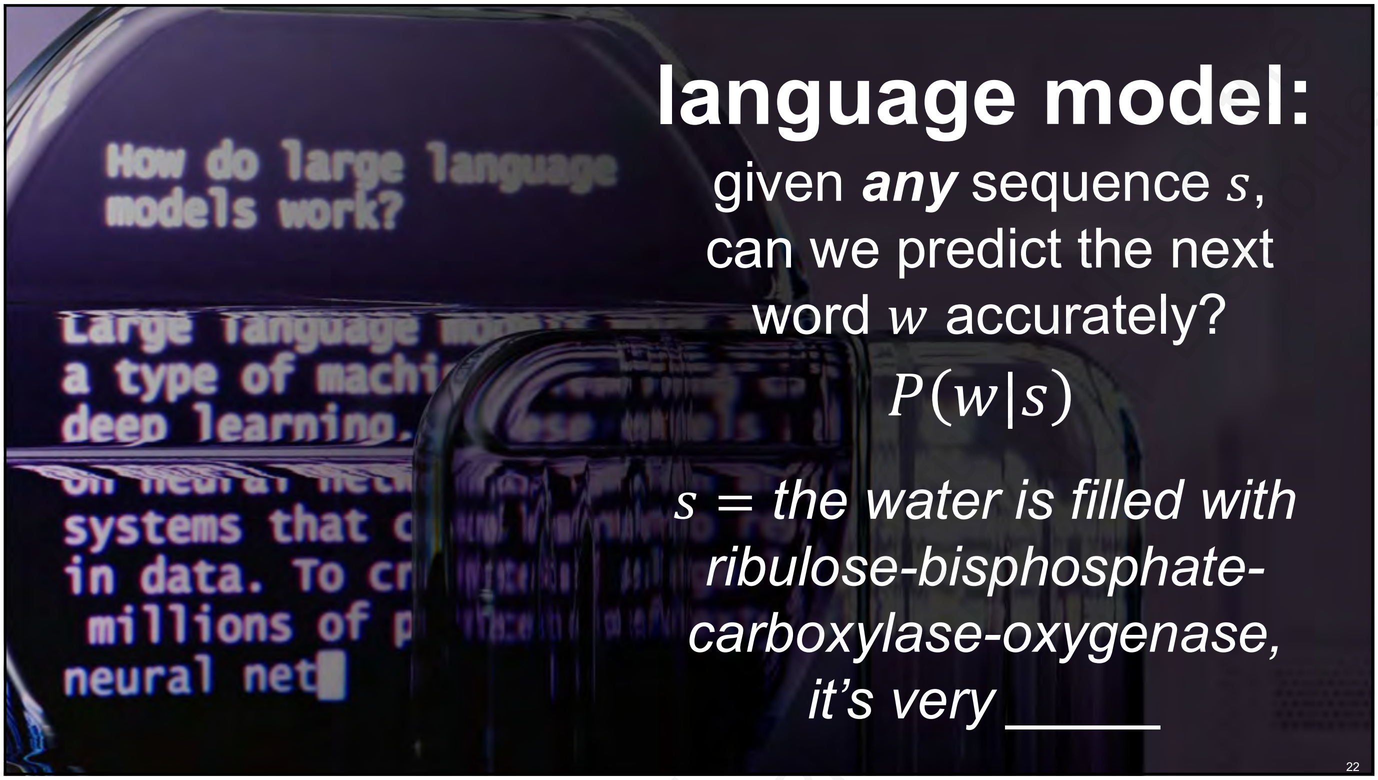
19


How do large language models work?
language:
vocabulary + grammar
words + syntaxes
sequence of word tokens
language model:
probabilistic model that captures the sequential structure of word tokens
20

20

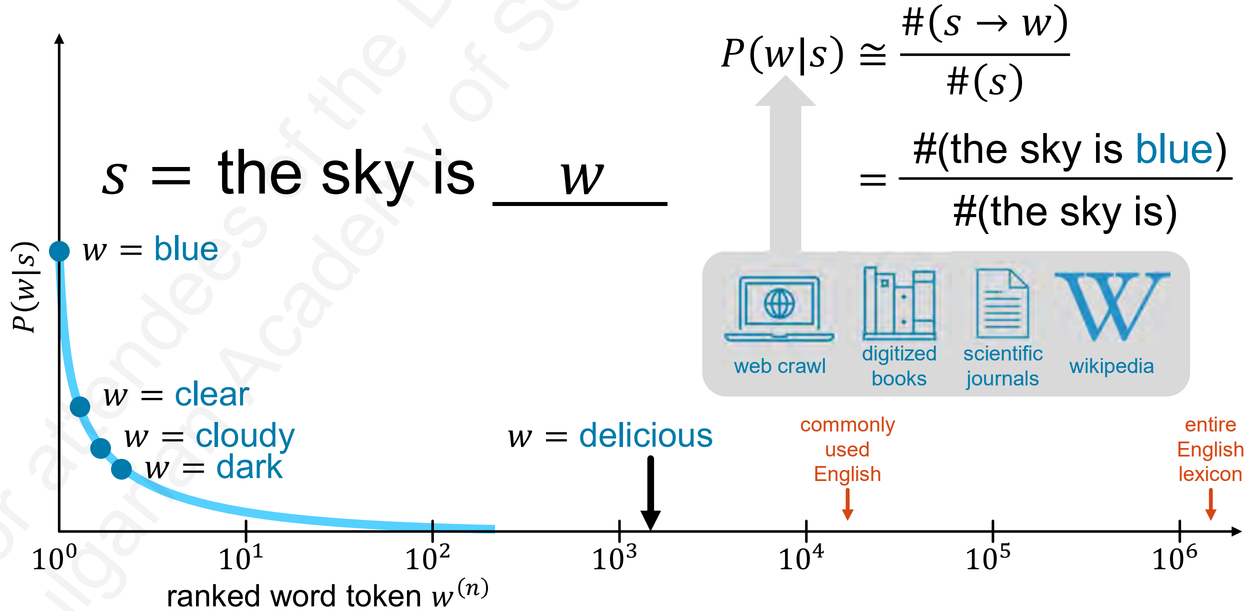

How do large language models work?
language model:
probabilistic model that captures the sequential structure of word tokens
given **any** sequence s, can we predict the next word w accurately?
$P(w|s)$
21

21



22

Do You Know Your English?



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

$$= \frac{\#(\text{the sky is blue})}{\#(\text{the sky is})}$$

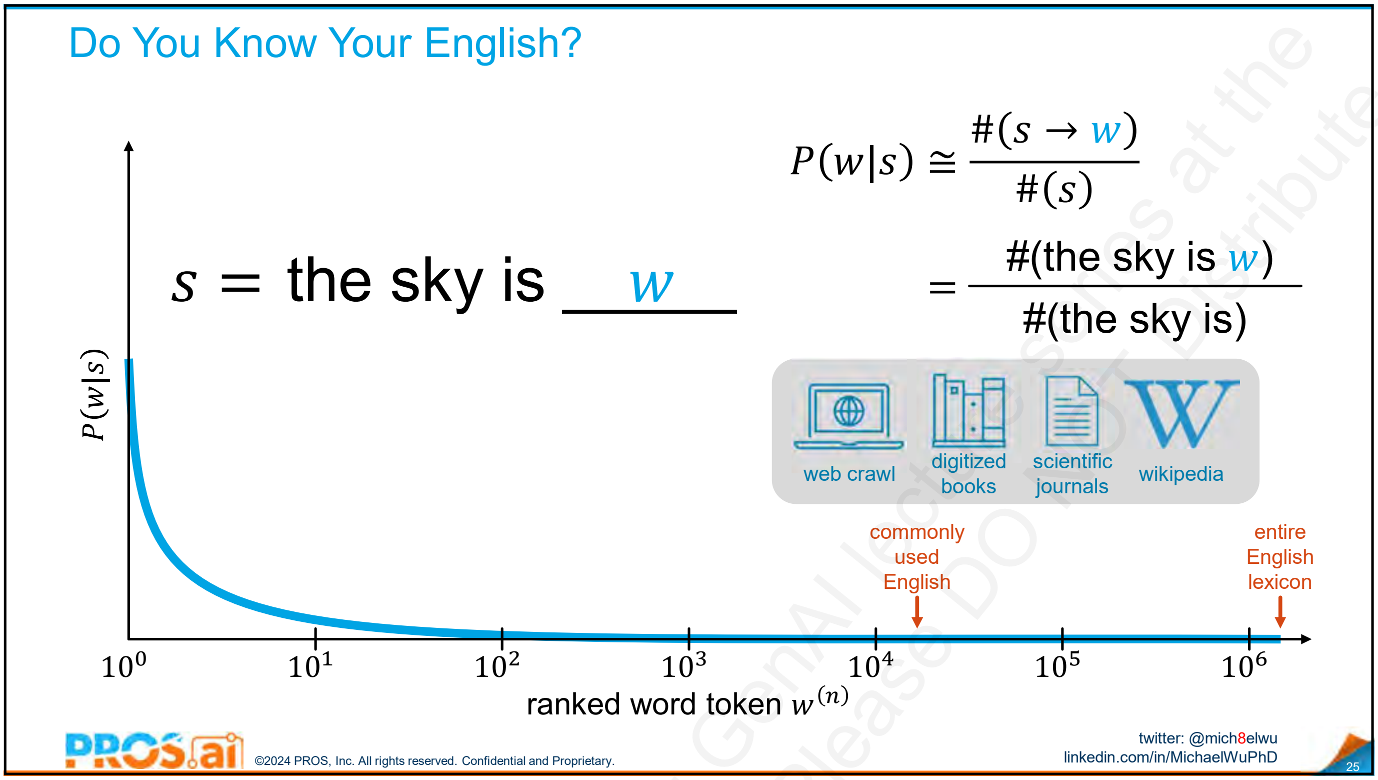
PROS.ai ©2024 PROS, Inc. All rights reserved. Confidential and Proprietary.

twitter: @mich8elwu
linkedin.com/in/MichaelWuPhD

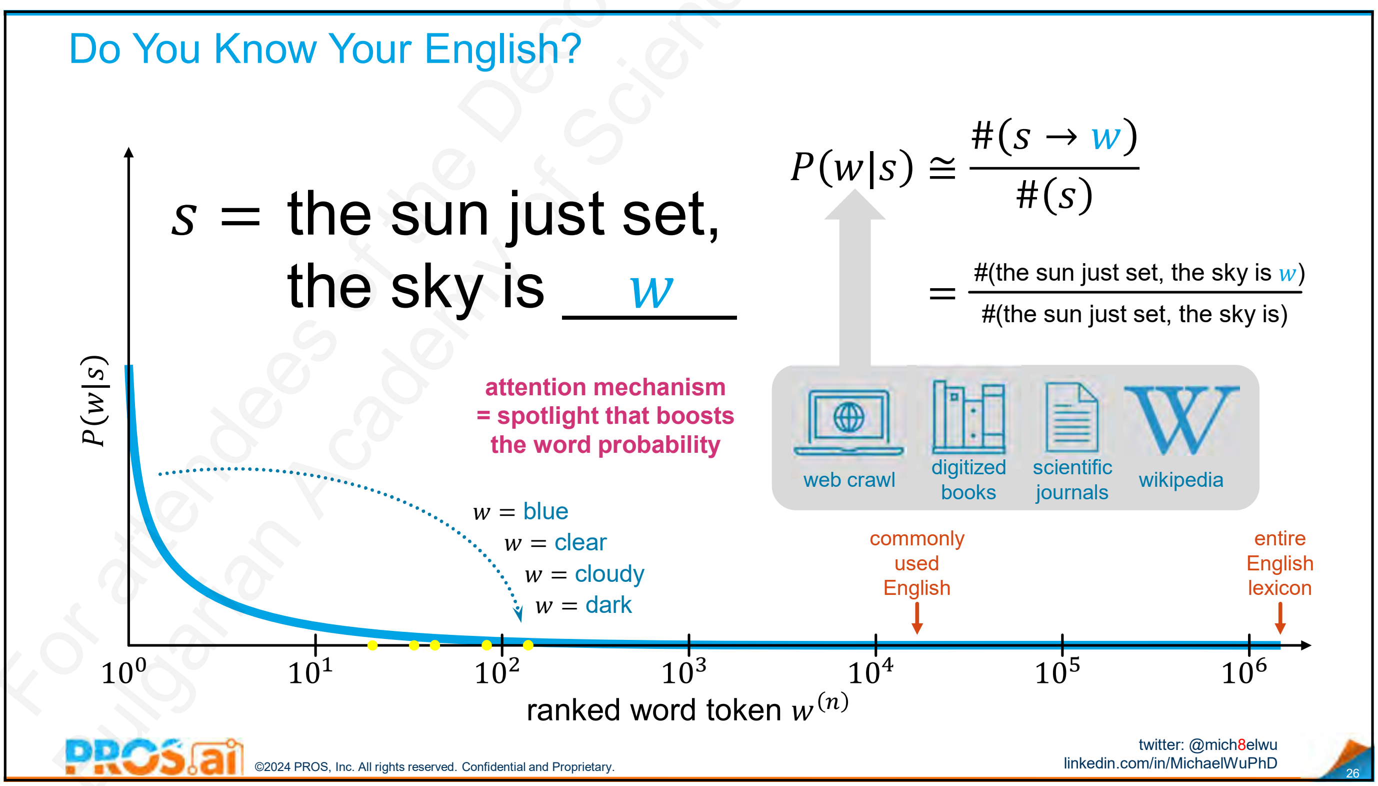
24

24

For attendees of the Decoding
Bulgarian Academy of Science.

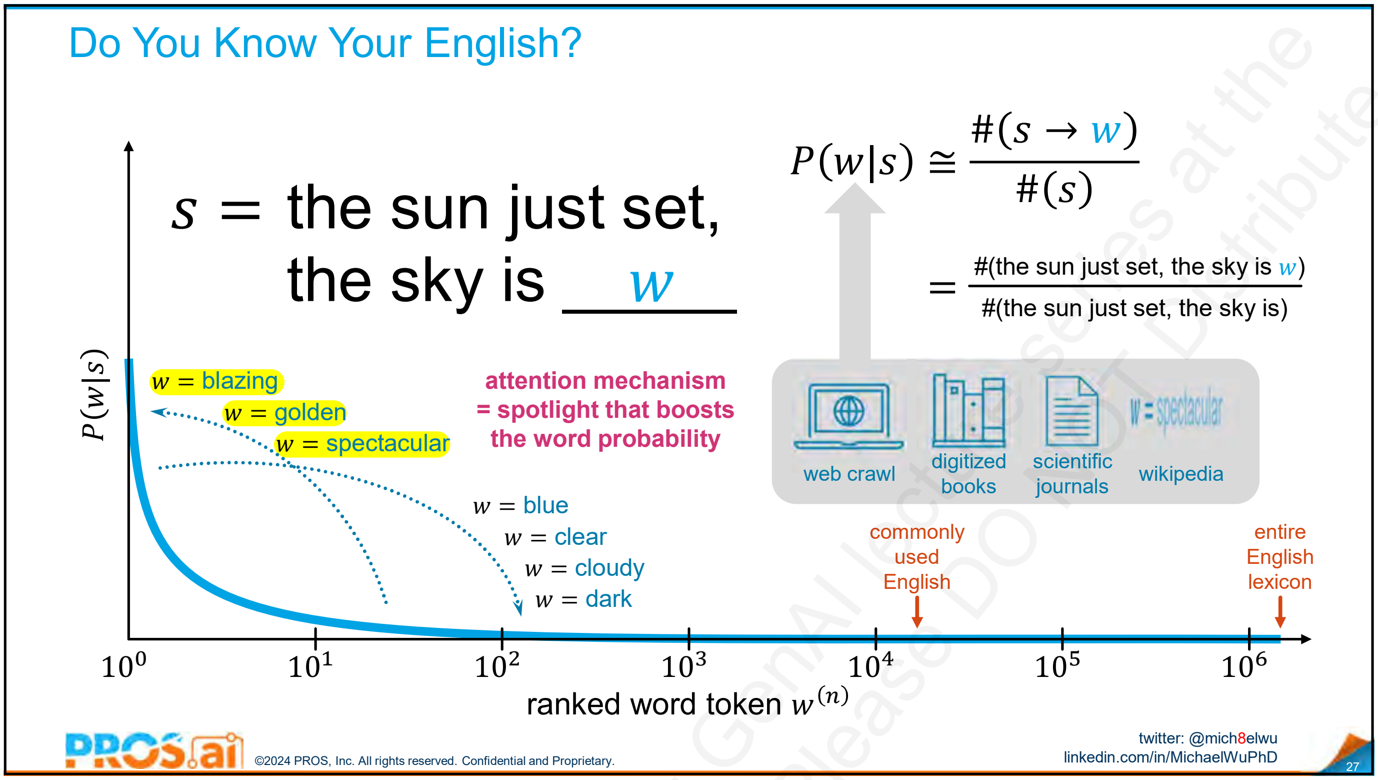


25

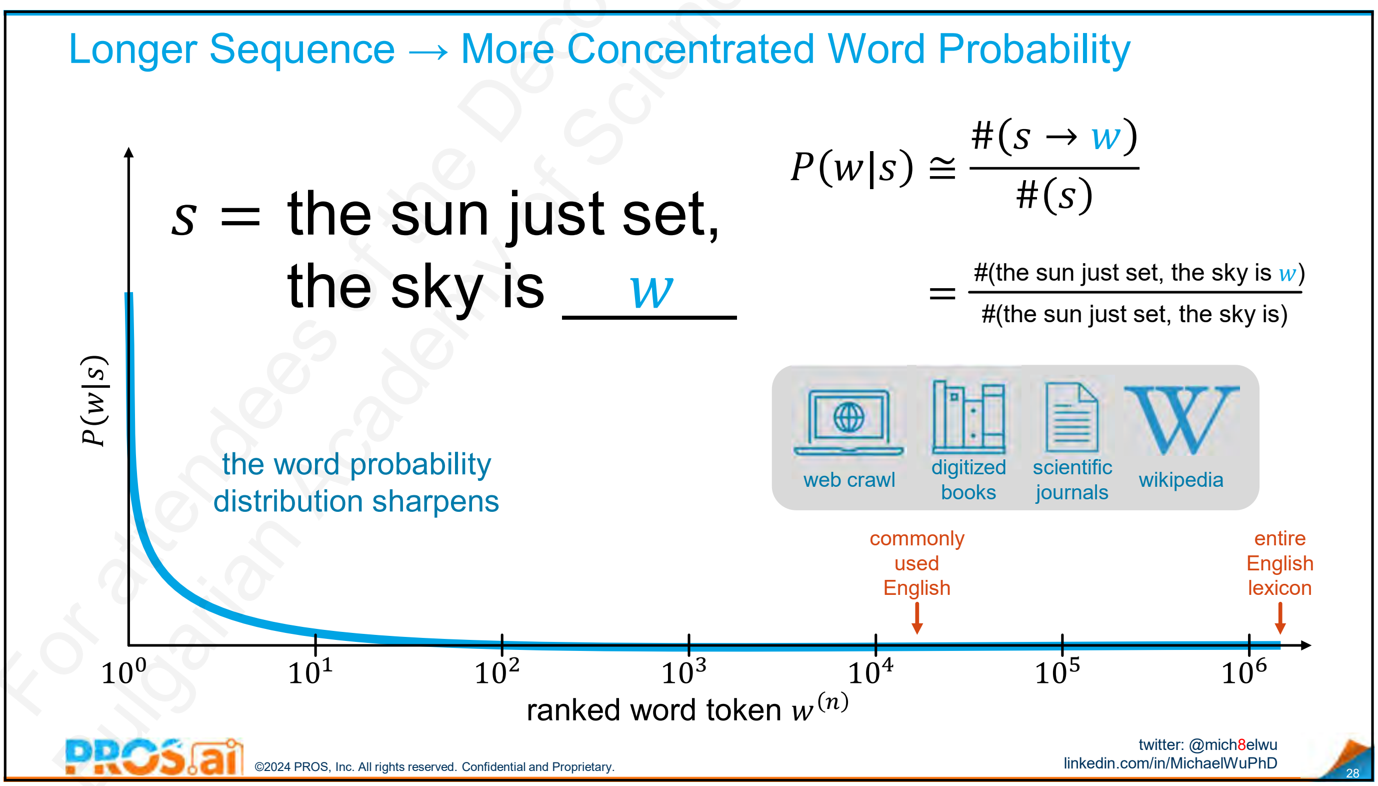


26

For attendees of the Decoding GenAI lectures at the Bulgarian Academy of Science. Please DO NOT Distribute.

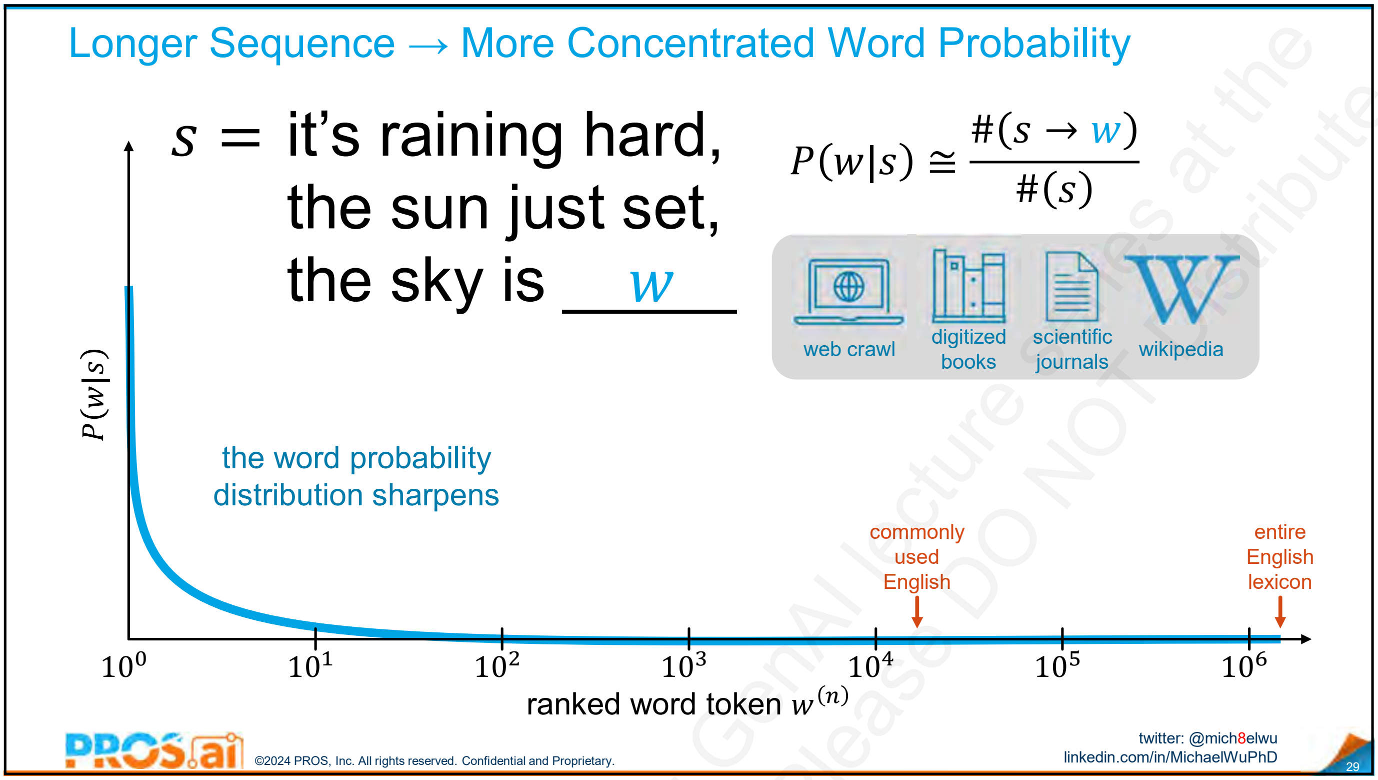


27

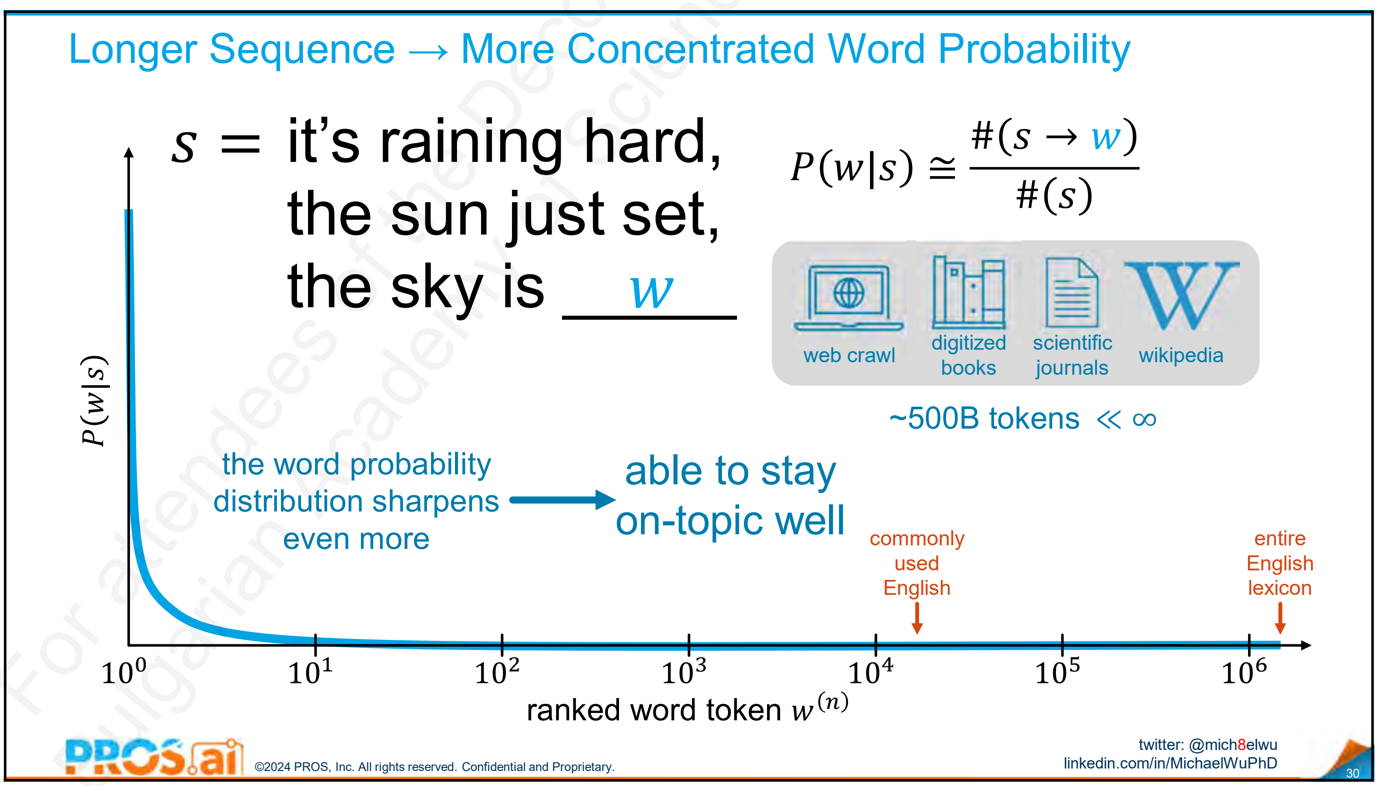


28

For attendees of the Decoding GenAI lectures at the Bulgarian Academy of Science. Please DO NOT Distribute.

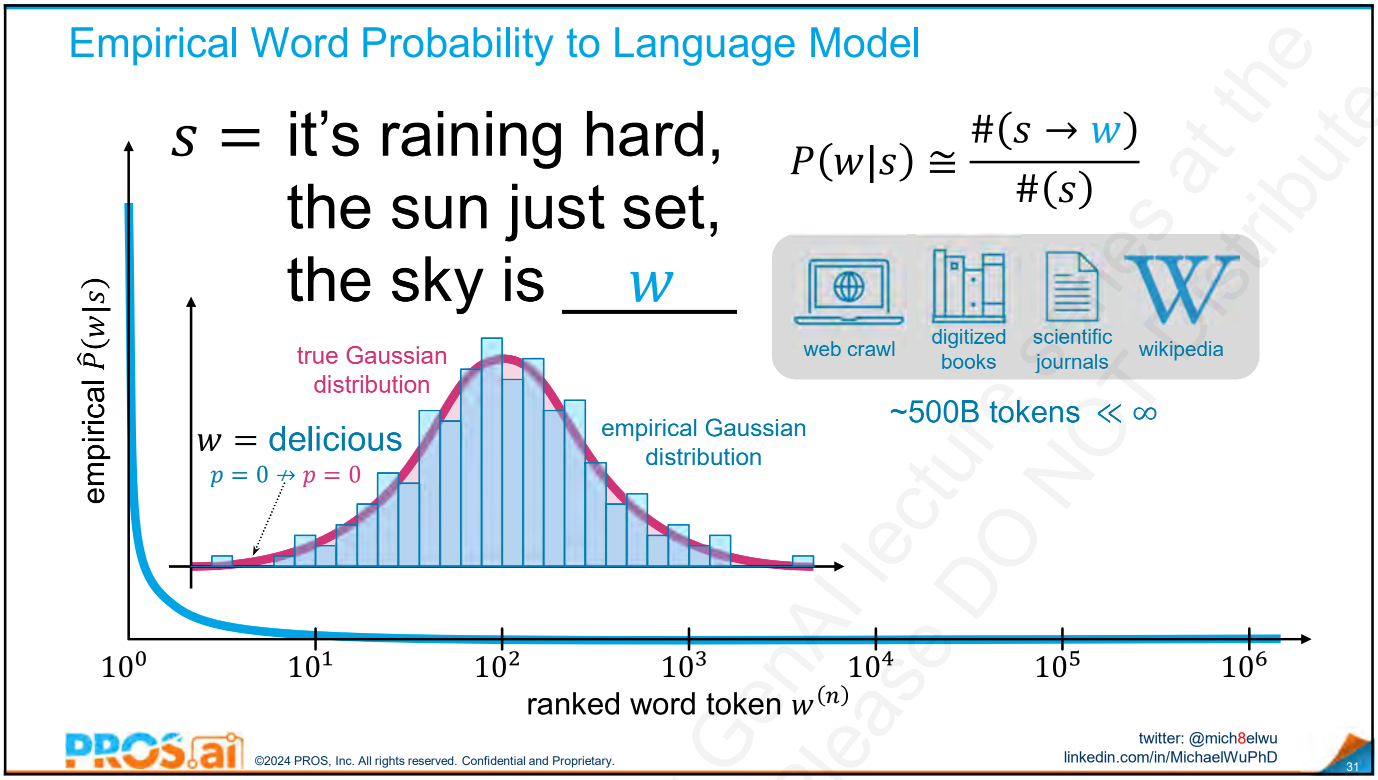


29



30

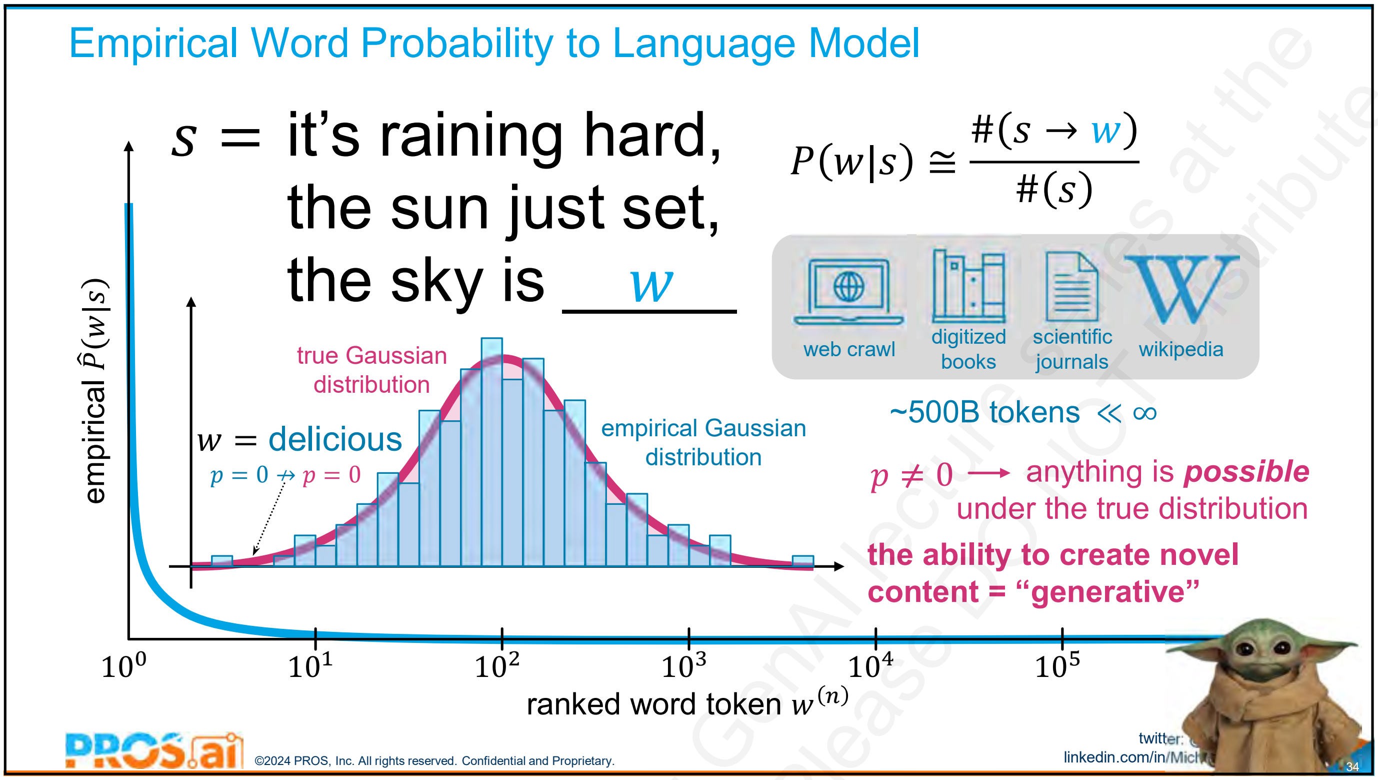
For attendees of the Decoding GenAI lecture series at the Bulgarian Academy of Science. Please DO NOT redistribute.



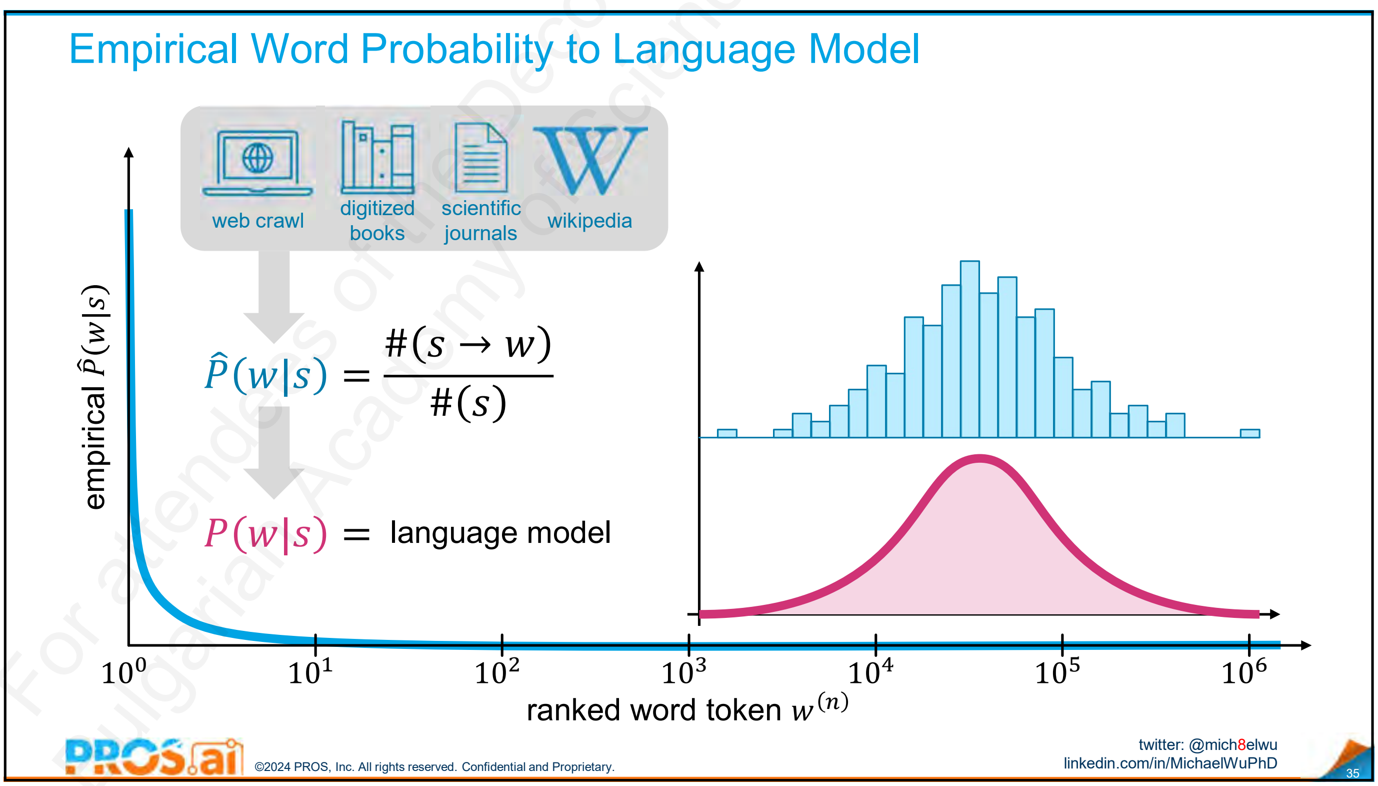
31

32

Please DO NOT distribute.

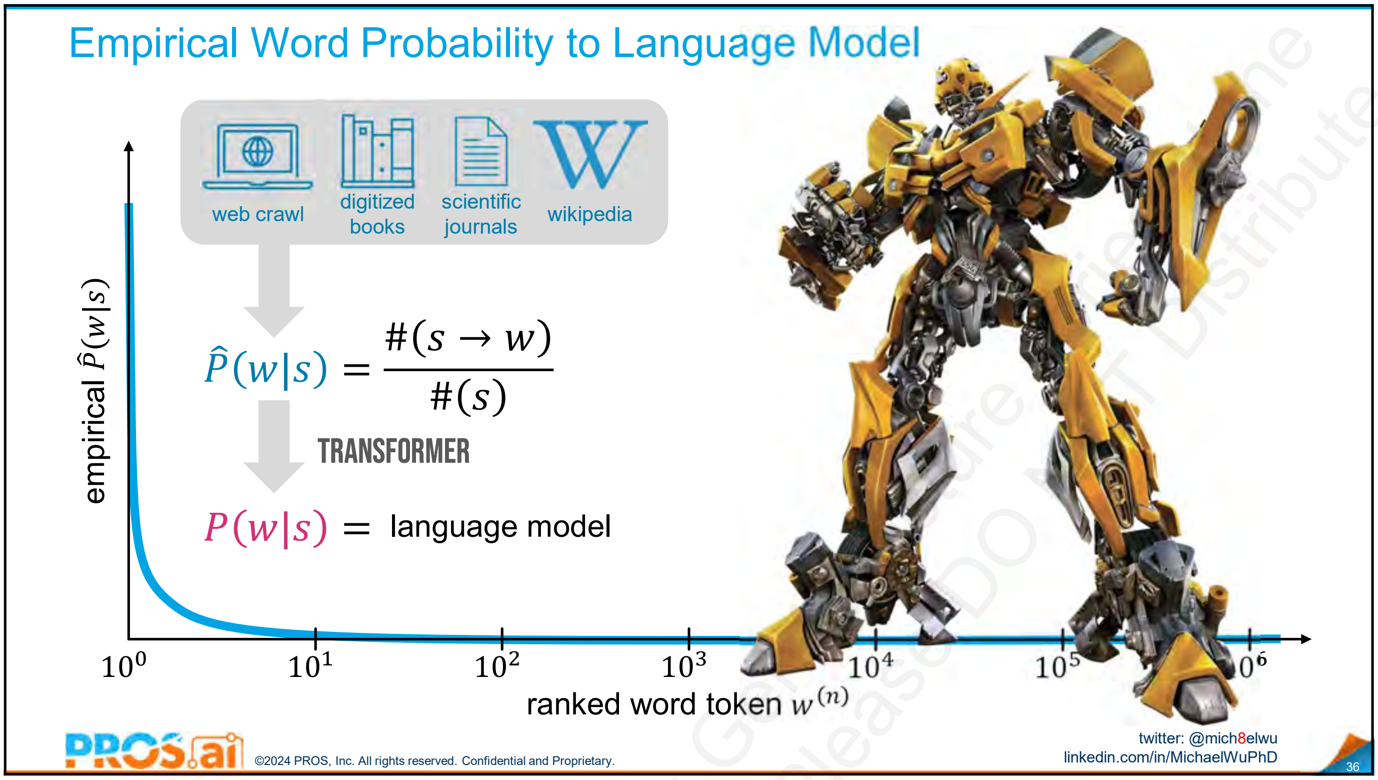


34

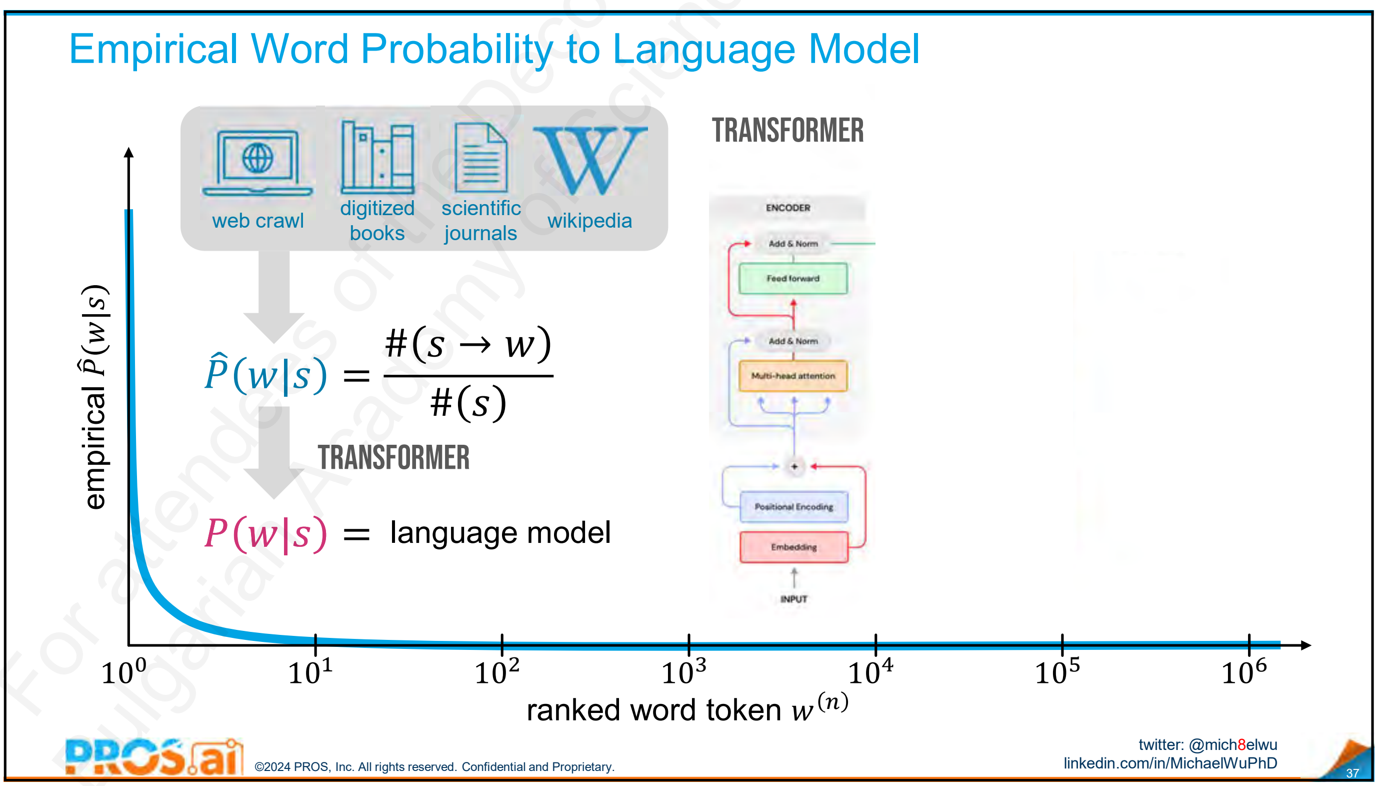


35

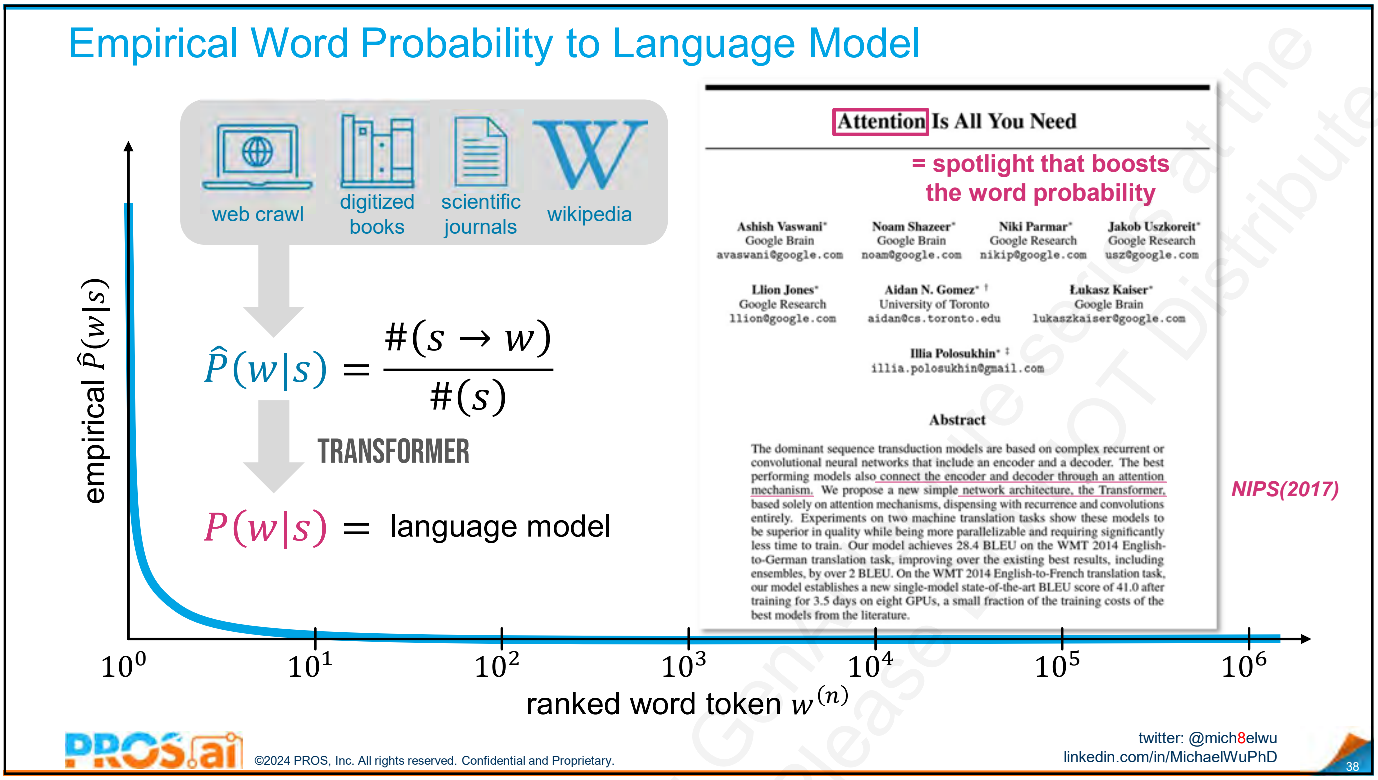
For attendees of the Decoding GenAI lectures at the Bulgarian Academy of Science. Please DO NOT redistribute.


Empirical Word Probability to Language Model
web crawl
digitized books
scientific journals
wikipedia
$\hat{P}(w|s) = \frac{\#(s \to w)}{\#(s)}$
TRANSFORMER
$P(w|s) =$ language model
empirical $\hat{P}(w|s)$
10^0 10^1 10^2 10^3 10^4 10^5 10^6
ranked word token $w^{(n)}$
©2024 PROS, Inc. All rights reserved. Confidential and Proprietary.
twitter: @mich8elwu
linkedin.com/in/MichaelWuPhD
36

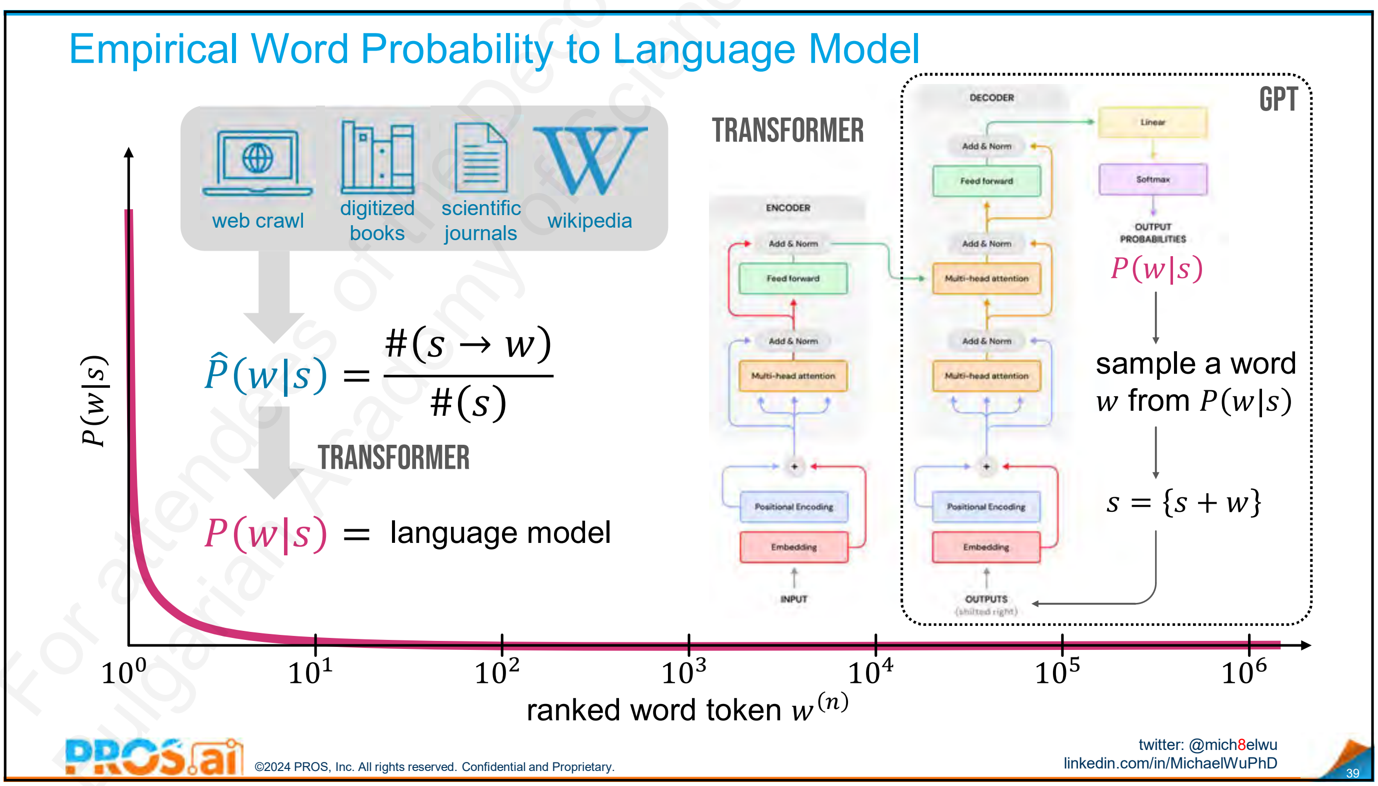
36


Empirical Word Probability to Language Model
web crawl
digitized books
scientific journals
wikipedia
$\hat{P}(w|s) = \frac{\#(s \to w)}{\#(s)}$
TRANSFORMER
$P(w|s) =$ language model
TRANSFORMER
ENCODER
Add & Norm
Feed forward
Add & Norm
Multi-head attention
Positional Encoding
Embedding
INPUT
empirical $\hat{P}(w|s)$
10^0 10^1 10^2 10^3 10^4 10^5 10^6
ranked word token $w^{(n)}$
©2024 PROS, Inc. All rights reserved. Confidential and Proprietary.
twitter: @mich8elwu
linkedin.com/in/MichaelWuPhD
37

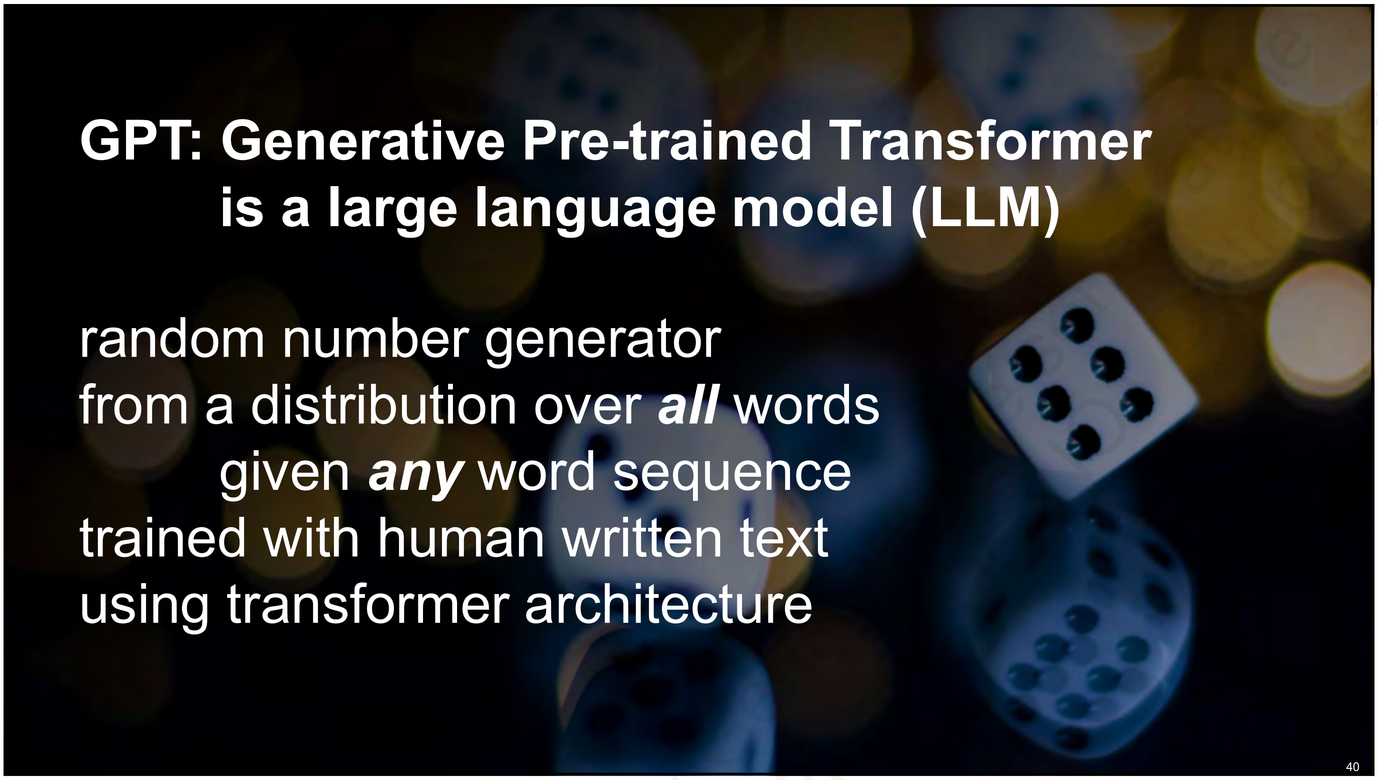
37


Empirical Word Probability to Language Model
web crawl
digitized books
scientific journals
wikipedia
$\hat{P}(w|s) = \frac{\#(s \to w)}{\#(s)}$
TRANSFORMER
$P(w|s) =$ language model
empirical $\hat{P}(w|s)$
10^0 10^1 10^2 10^3 10^4 10^5 10^6
ranked word token $w^{(n)}$
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)
©2024 PROS, Inc. All rights reserved. Confidential and Proprietary.
twitter: @mich8elwu
linkedin.com/in/MichaelWuPhD
38

38


Empirical Word Probability to Language Model
web crawl
digitized books
scientific journals
wikipedia
$\hat{P}(w|s) = \frac{\#(s \to w)}{\#(s)}$
TRANSFORMER
$P(w|s) =$ language model
$P(w|s)$
10^0 10^1 10^2 10^3 10^4 10^5 10^6
ranked word token $w^{(n)}$
TRANSFORMER
GPT
ENCODER
DECODER
Add & Norm
Feed forward
Multi-head attention
Positional Encoding
Embedding
INPUT
OUTPUTS
Linear
Softmax
OUTPUT PROBABILITIES
$P(w|s)$
sample a word w from $P(w|s)$
$s = \{s + w\}$
©2024 PROS, Inc. All rights reserved. Confidential and Proprietary.
twitter: @mich8elwu
linkedin.com/in/MichaelWuPhD
39

39


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
40

40


doesn't sound
very intelligent
41

41

From GPT to ChatGPT

pre-trained via supervised/ self-supervised

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

GPT

$$P(w|s) = \text{language model}$$

plausible continuation of text ≠ good response to questions and instructions

- user: explain how cost-based pricing in B2B works
- LLM: explain how value-based pricing in B2C works

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

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

twitter: @mich8elwu
linkedin.com/in/MichaelWuPhD

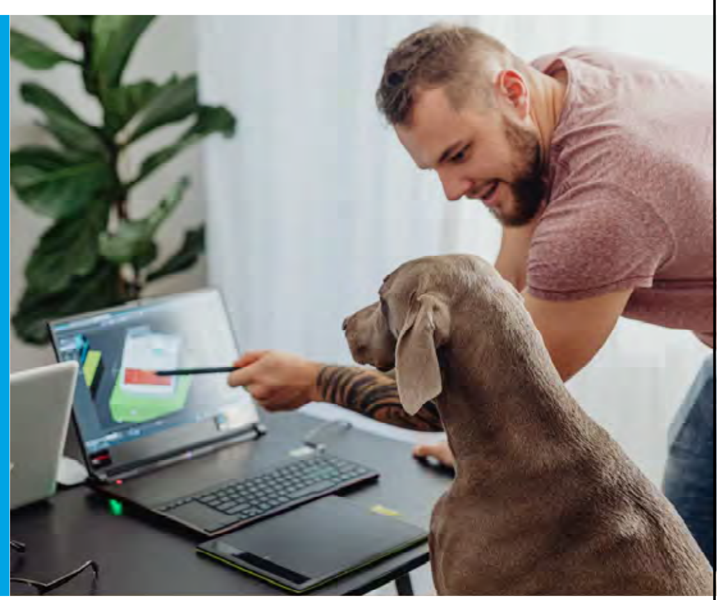
42

From GPT to ChatGPT

plausible continuation of text ≠ good response to questions and instructions

- user: explain how cost-based pricing in B2B works
- LLM: explain how value-based pricing in B2C works

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



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

twitter: @mich8elwu
linkedin.com/in/MichaelWuPhD

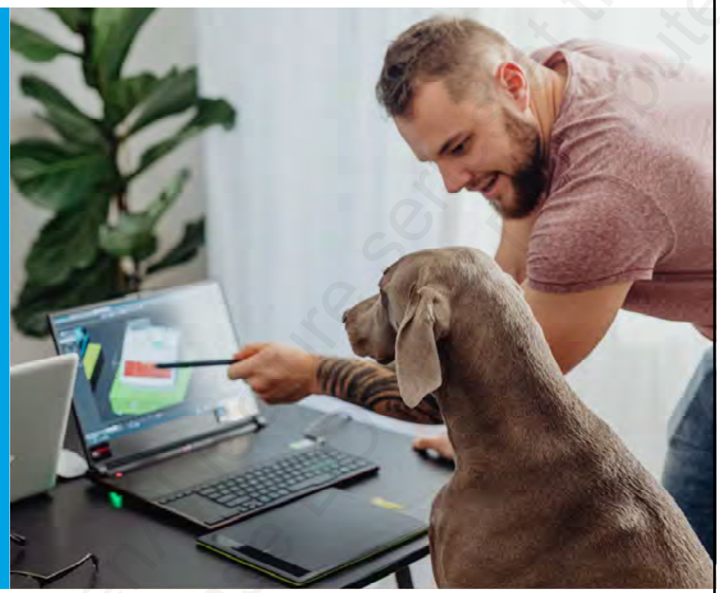
43

From GPT to ChatGPT

responding to commands
≠ good open-ended dialogues

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

- scoring a good dialog is nebulous
- inconsistency among people



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

twitter: @mich8elwu
linkedin.com/in/MichaelWuPhD

44

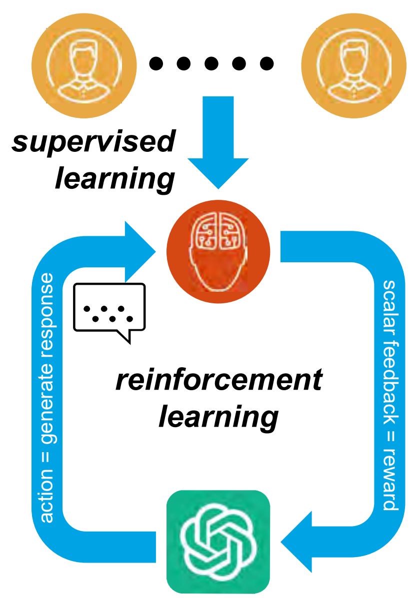
44

From GPT to ChatGPT

responding to commands
≠ good open-ended dialogues

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

- LLM generates few responses
  a b c d
- humans rank the preferences
  c > a > d > b
- supervised train a reward function to score responses to be used in reinforcement learning



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

twitter: @mich8elwu
linkedin.com/in/MichaelWuPhD

45

45

From GPT to ChatGPT

responding to commands
≠ good open-ended dialogues

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

- LLM generates few responses
  a b c d
- humans rank the preferences
  c > a > d > b
- supervised train a reward function to score responses to be used in reinforcement learning

PROS.ai
©2024 PROS, Inc. All rights reserved. Confidential and Proprietary.
twitter: @mich8elwu
linkedin.com/in/MichaelWuPhD
46

46

48


"*generative*" AI are, by definition and construction, ***hallucinatory***
49

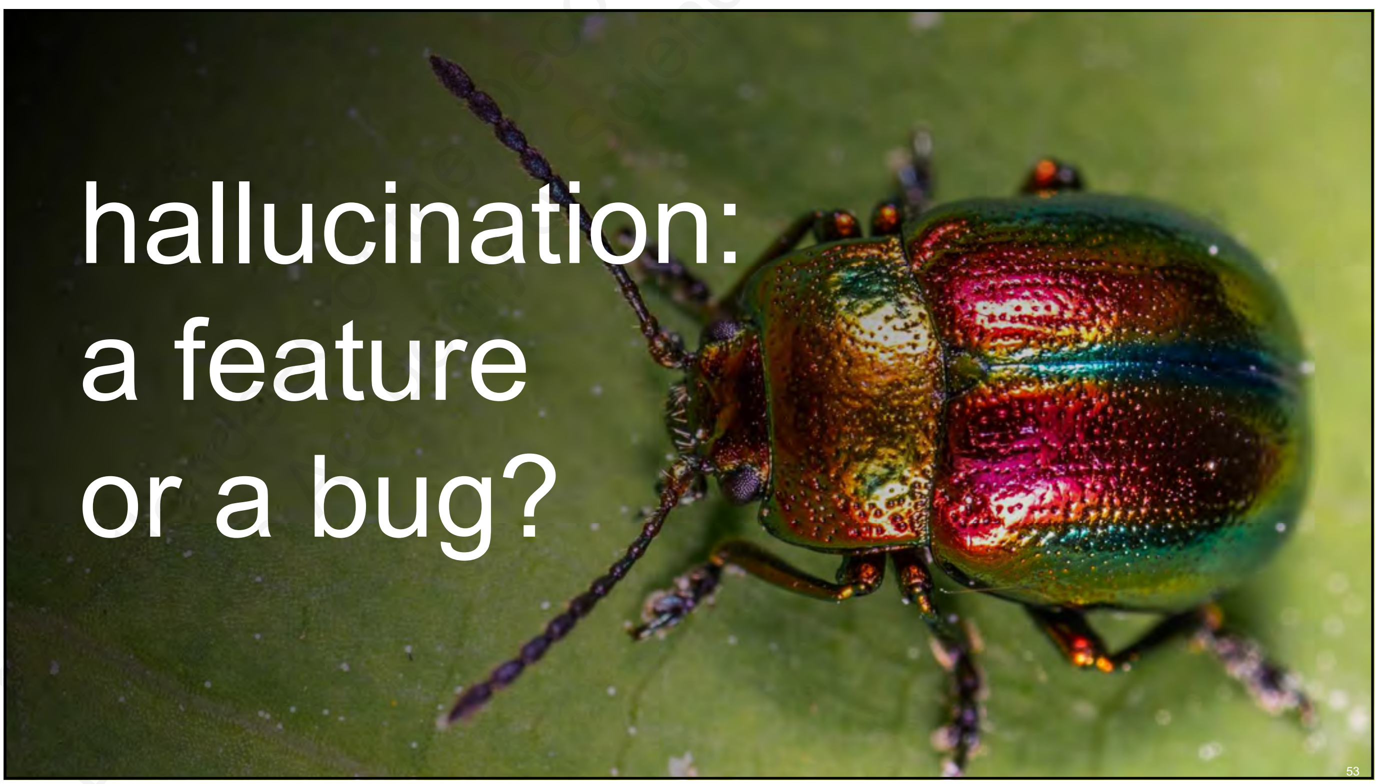
49


very convincing legal briefings citing non-existent court cases
50

50


Steven Schwartz
+ Peter LoDuca fined $5K
51

51


hallucination:
a feature
or a bug?
53

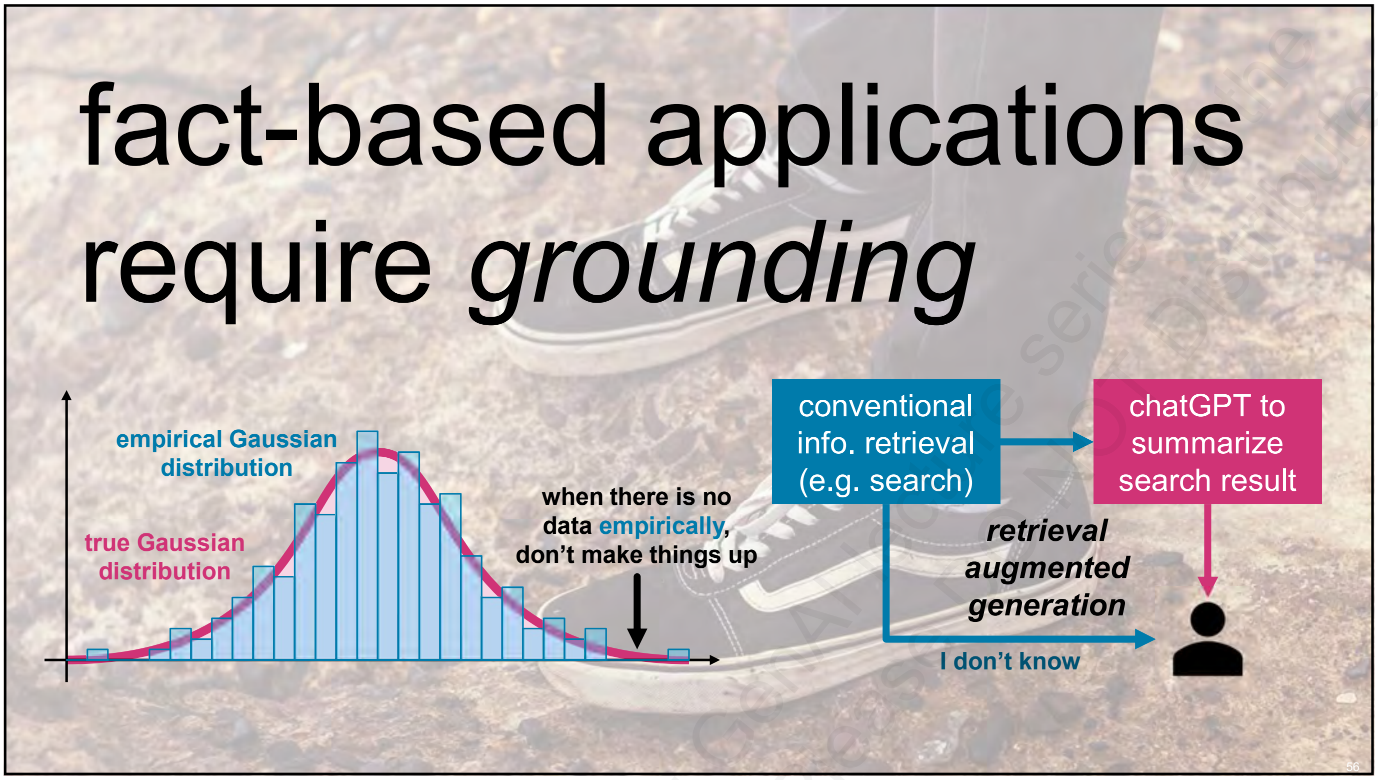
53


feature: for *design*
+ *creative*
use cases
54

54


bug: for *fact-based*
applications
55

55


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
56

56


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)
59

59

60

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?

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

twitter: @mich8elwu
linkedin.com/in/MichaelWuPhD

61

61

62

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

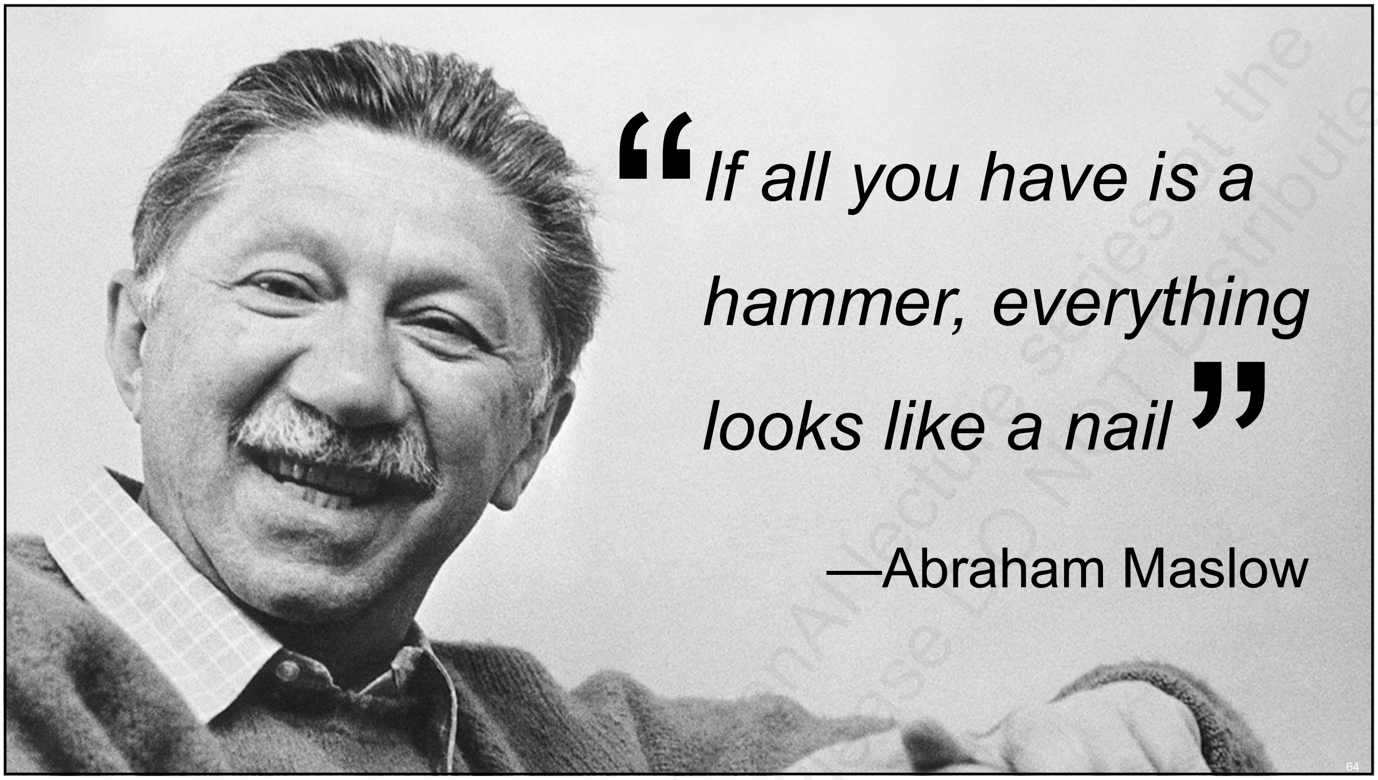
many many more start-ups

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

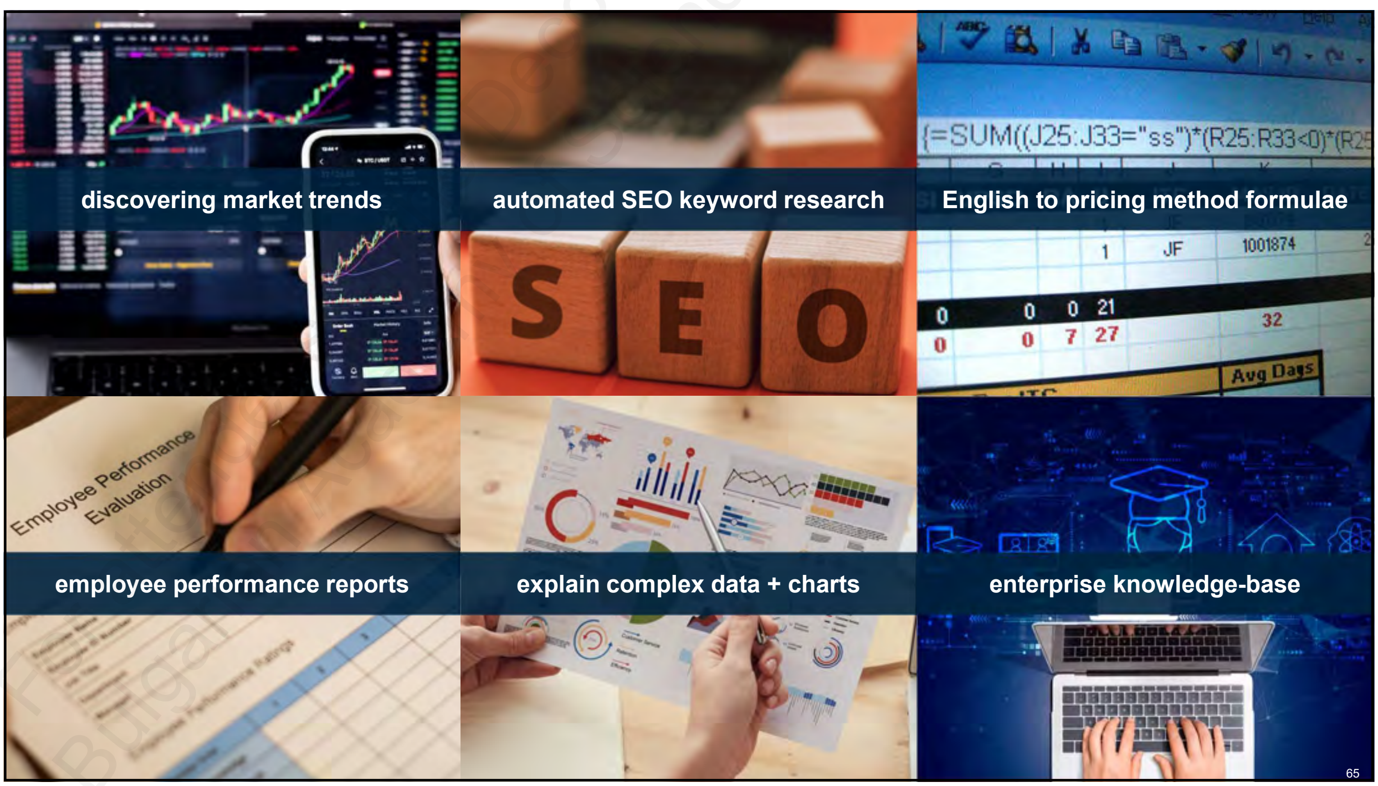
twitter: @mich8elwu
linkedin.com/in/MichaelWuPhD

63

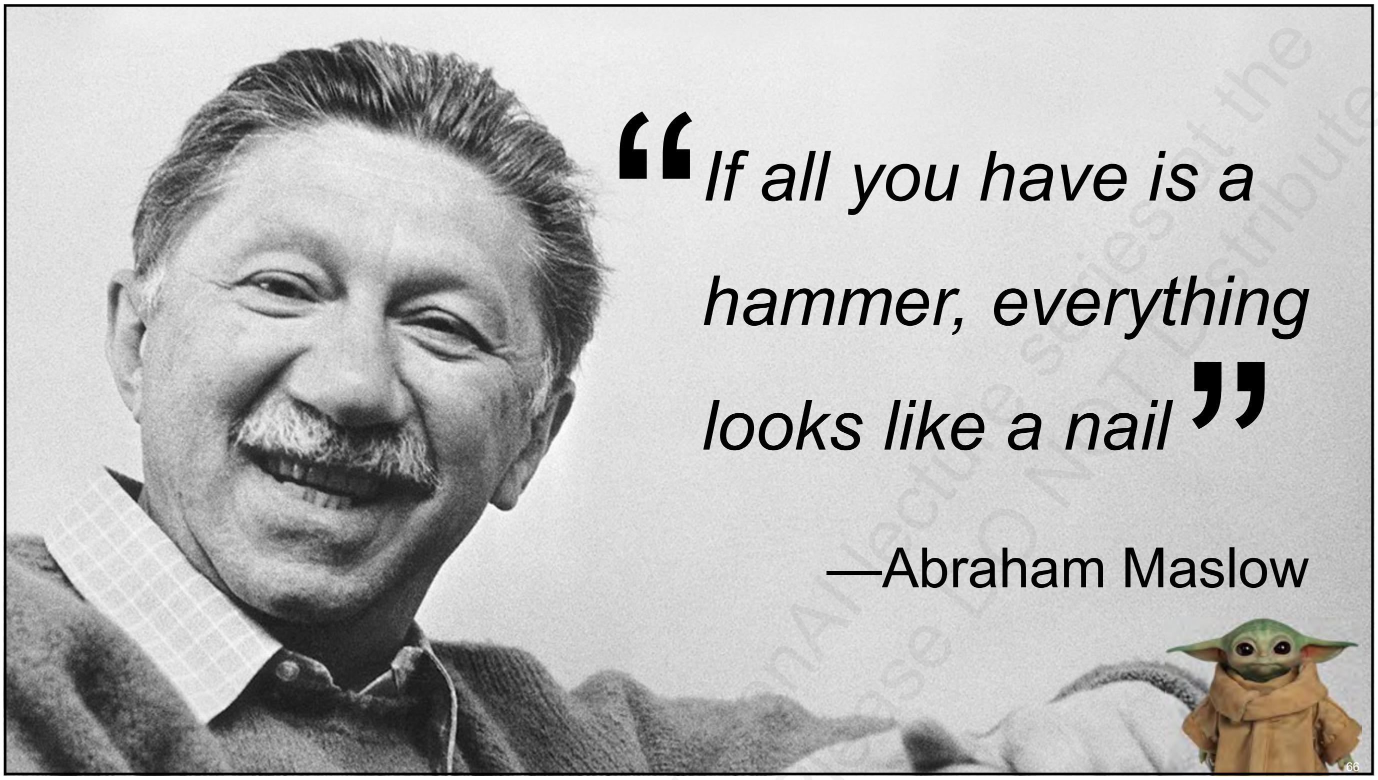
63


“If all you have is a hammer, everything looks like a nail”
—Abraham Maslow
64

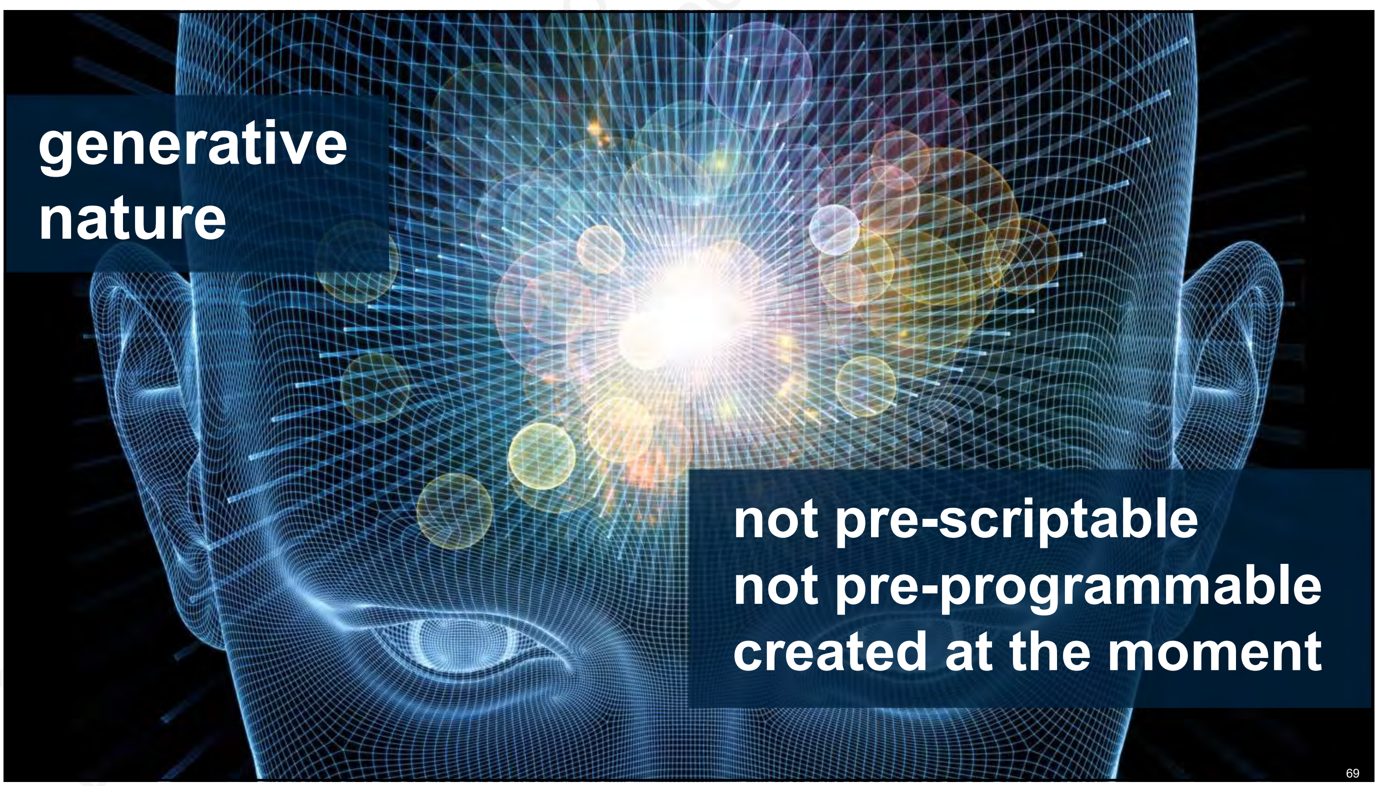
64


discovering market trends
automated SEO keyword research
English to pricing method formulae
SEO
{=SUM((J25:J33="ss")*(R25:R33<0)*(R2
employee performance reports
Employee Performance Evaluation
explain complex data + charts
enterprise knowledge-base
65

65


“If all you have is a hammer, everything looks like a nail”
—Abraham Maslow
66

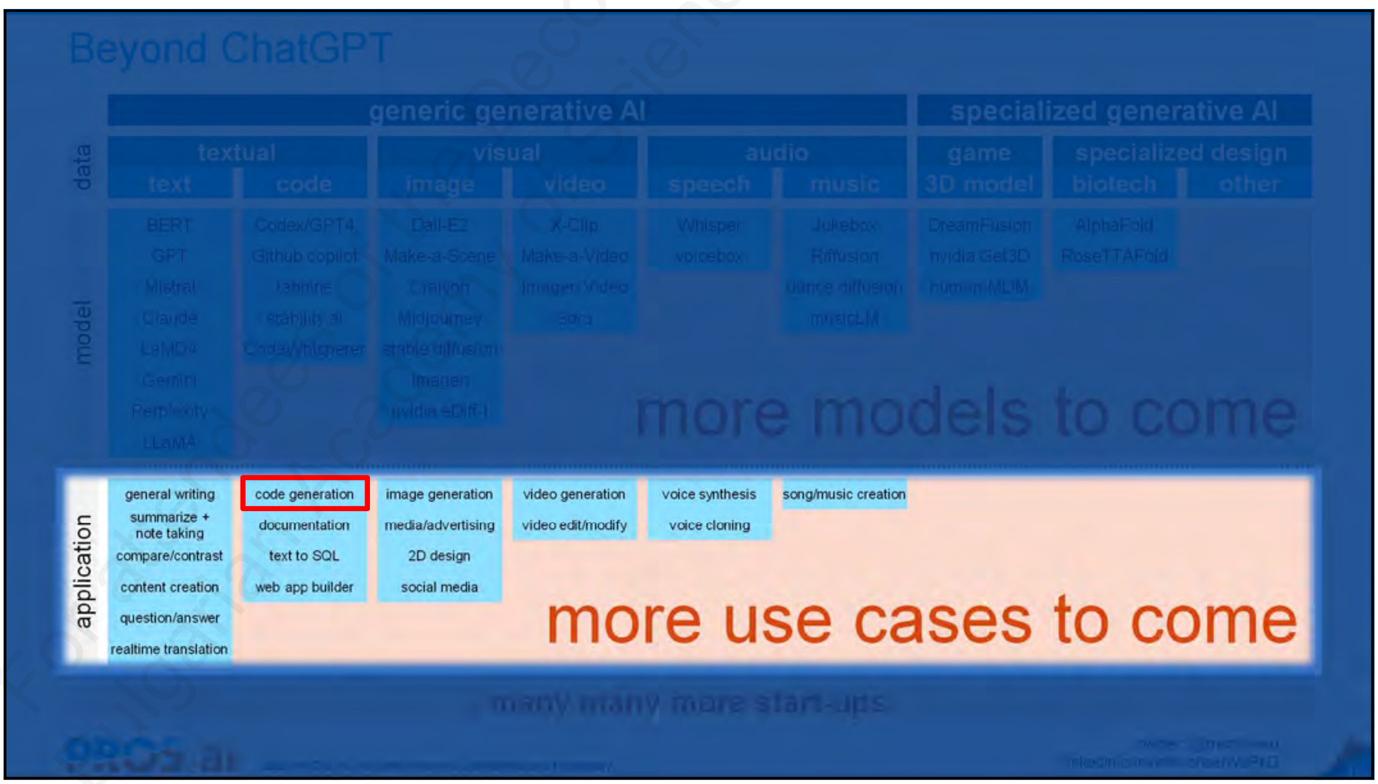
66


generative nature
not pre-scriptable
not pre-programmable
created at the moment
69

69

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
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
more models to come
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 use cases to come
many many more start-ups

70



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
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
more models to come
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 use cases to come
many many more start-ups

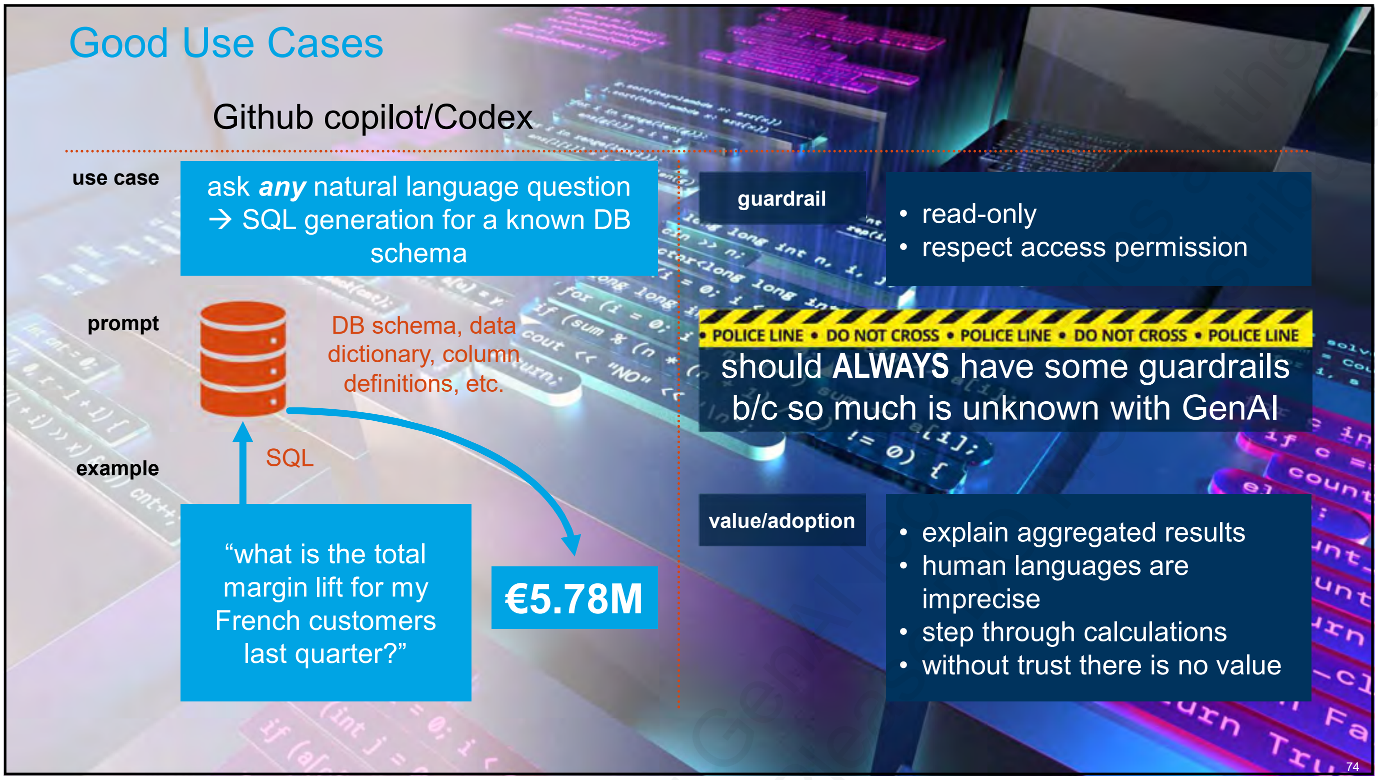
71


copyright
72

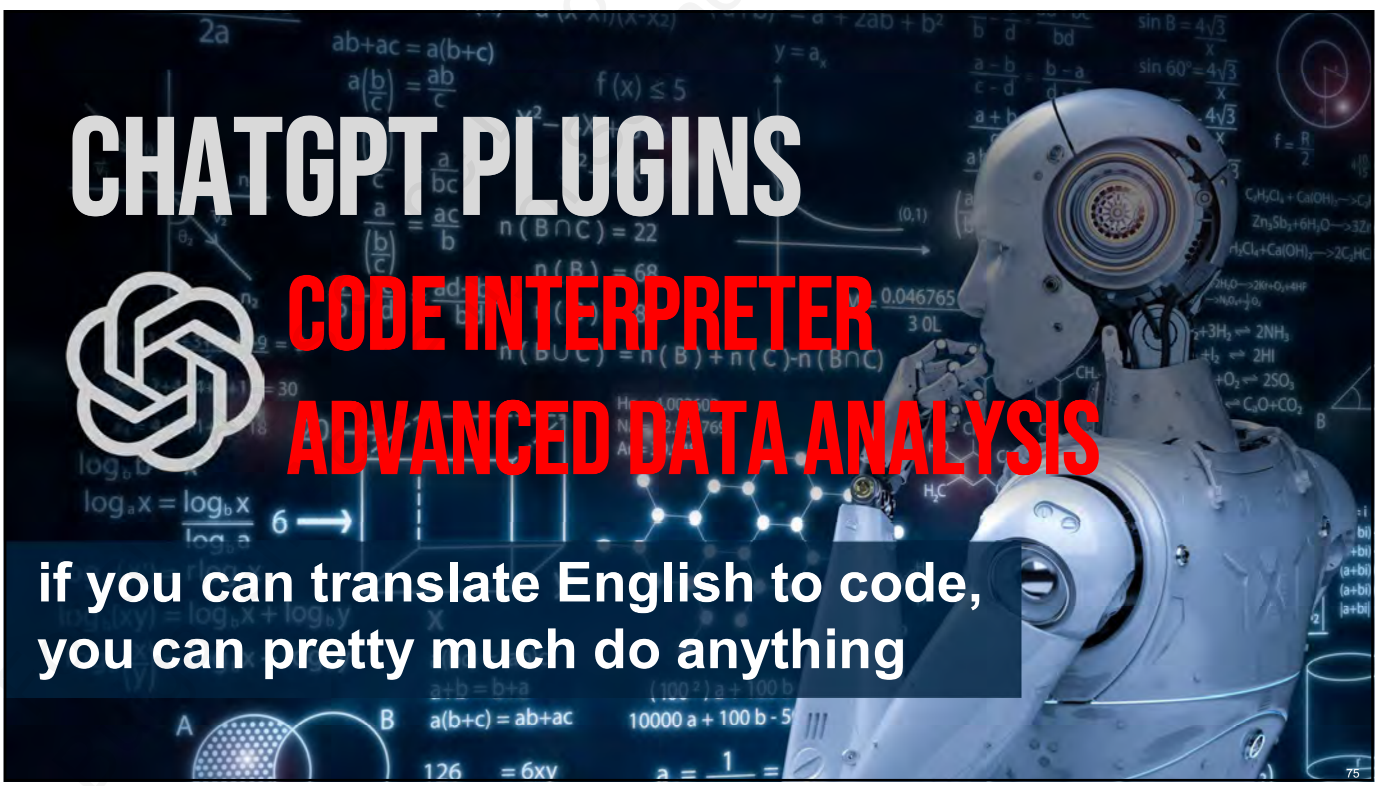
72


FANTASY
ZARYA OF THE DAWN
Kristina Kashtanova
73

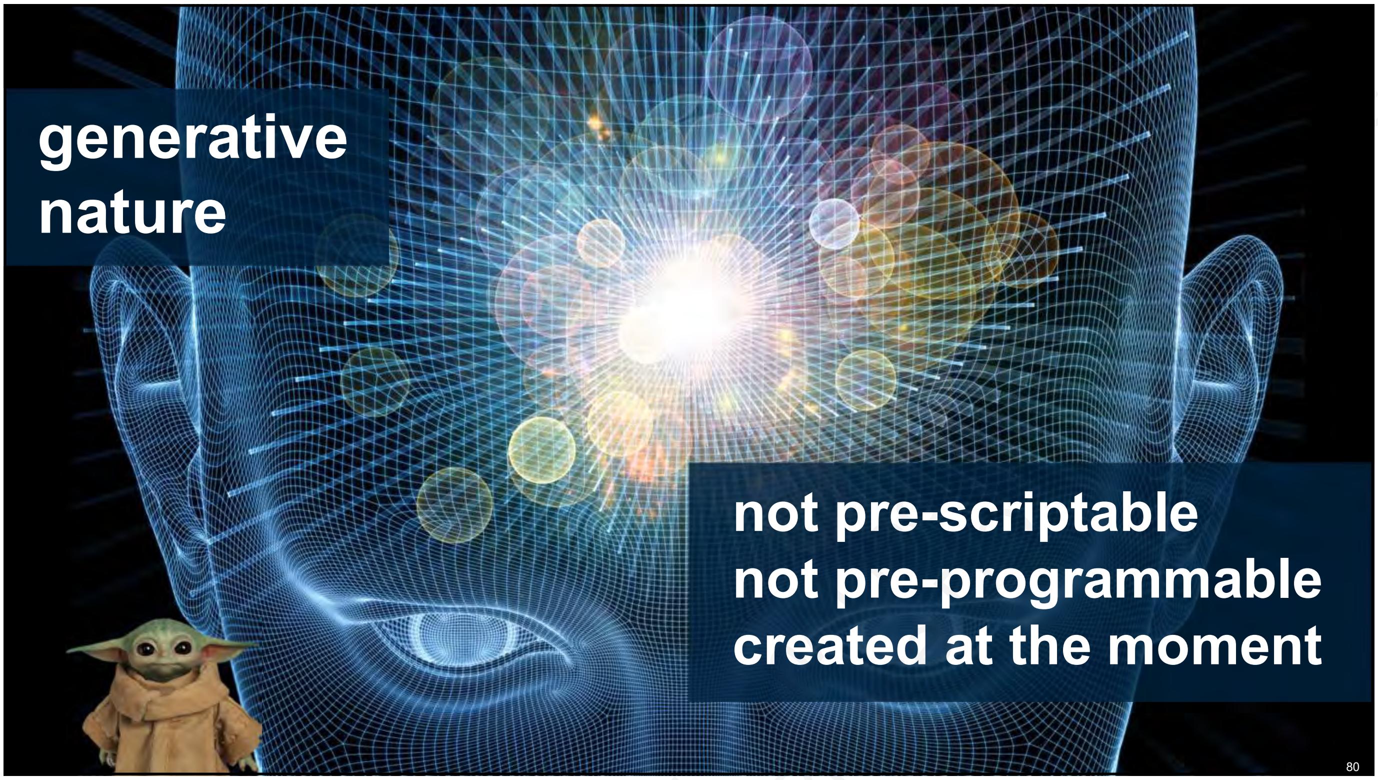
73


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
74

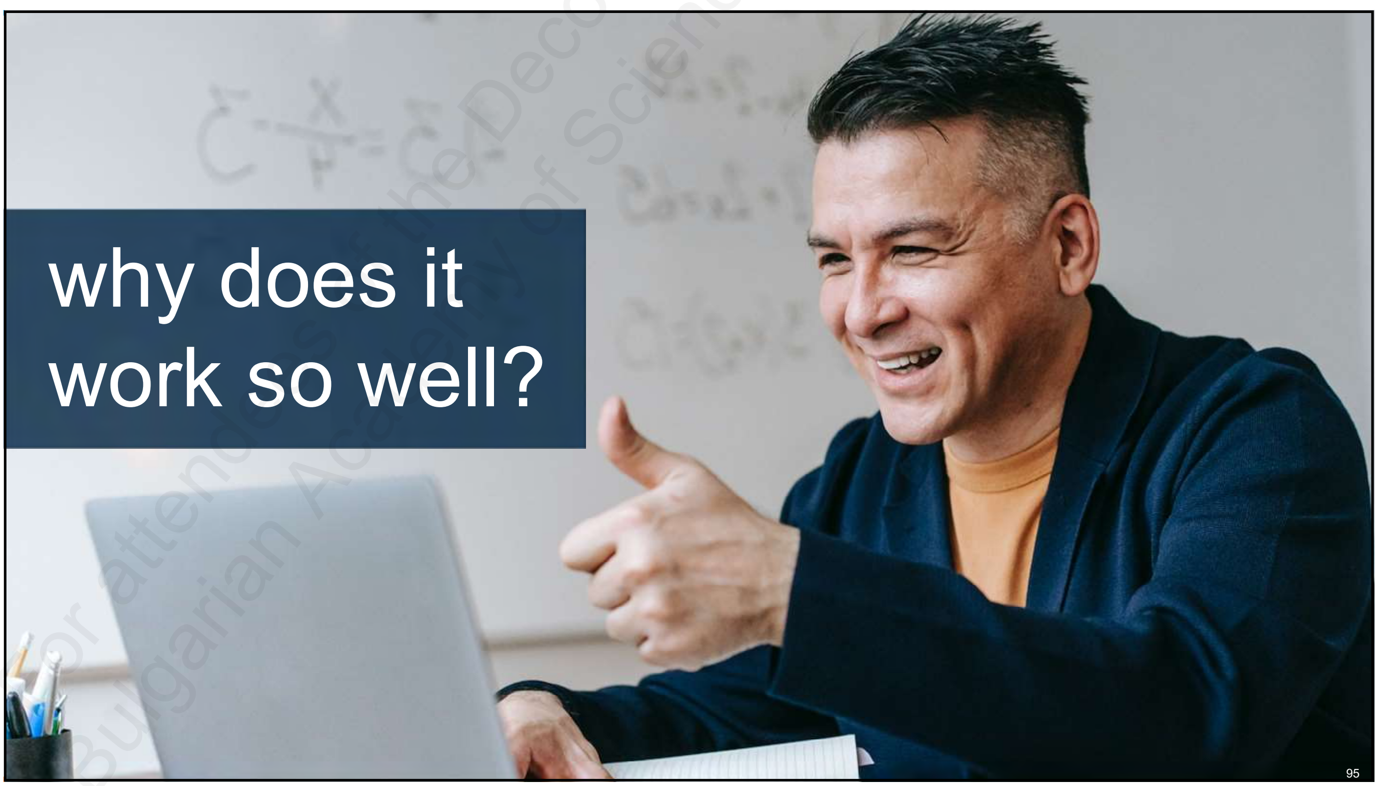
74


CHATGPT PLUGINS
CODE INTERPRETER
ADVANCED DATA ANALYSIS
if you can translate English to code, you can pretty much do anything
75

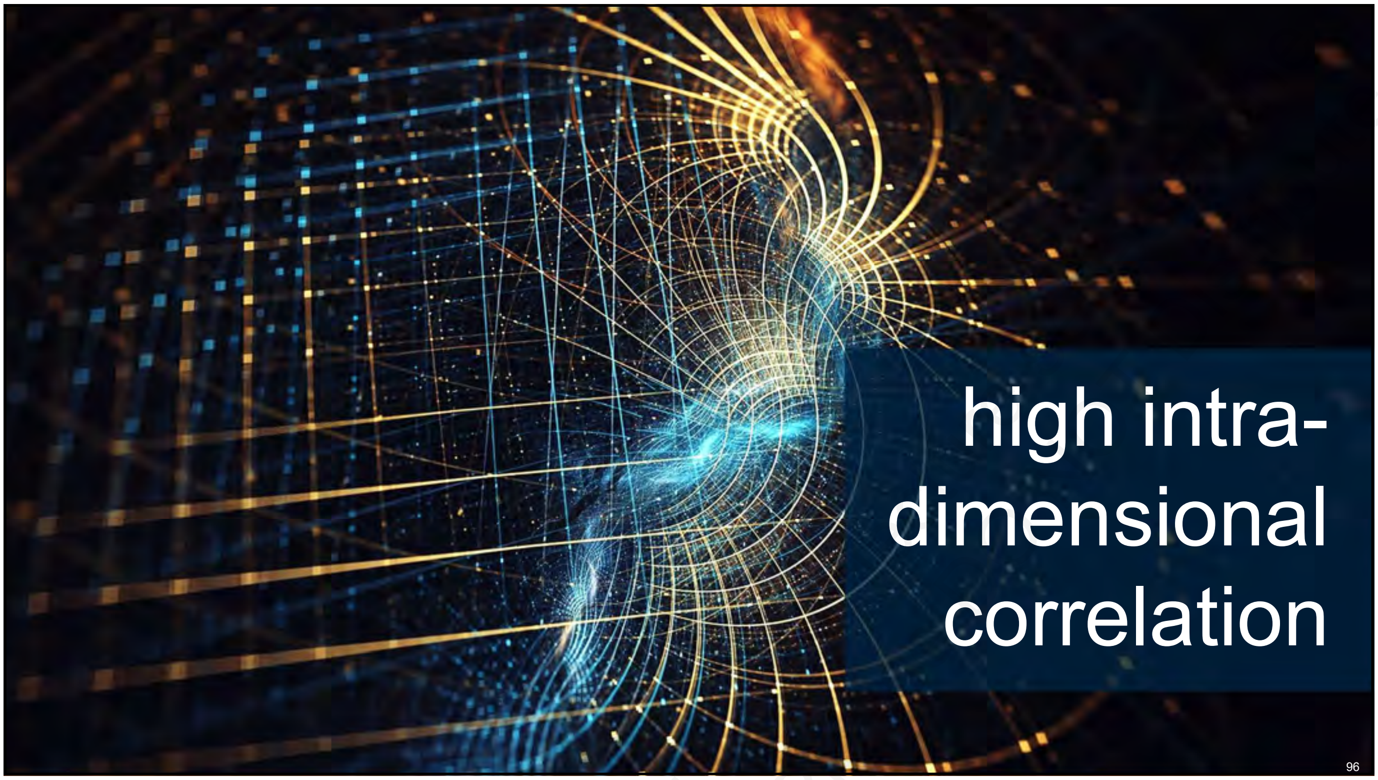
75


generative
nature
not pre-scriptable
not pre-programmable
created at the moment
80

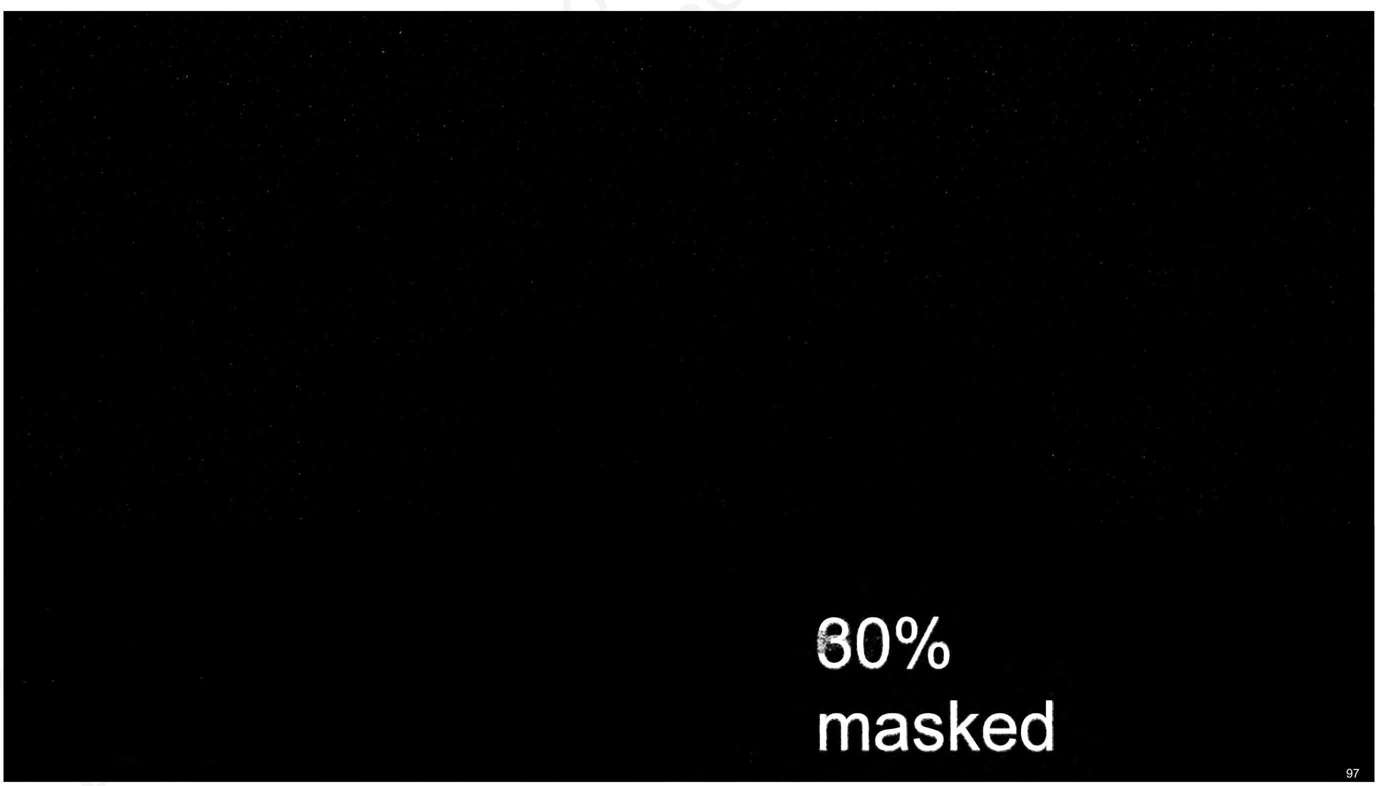
80


why does it
work so well?
95

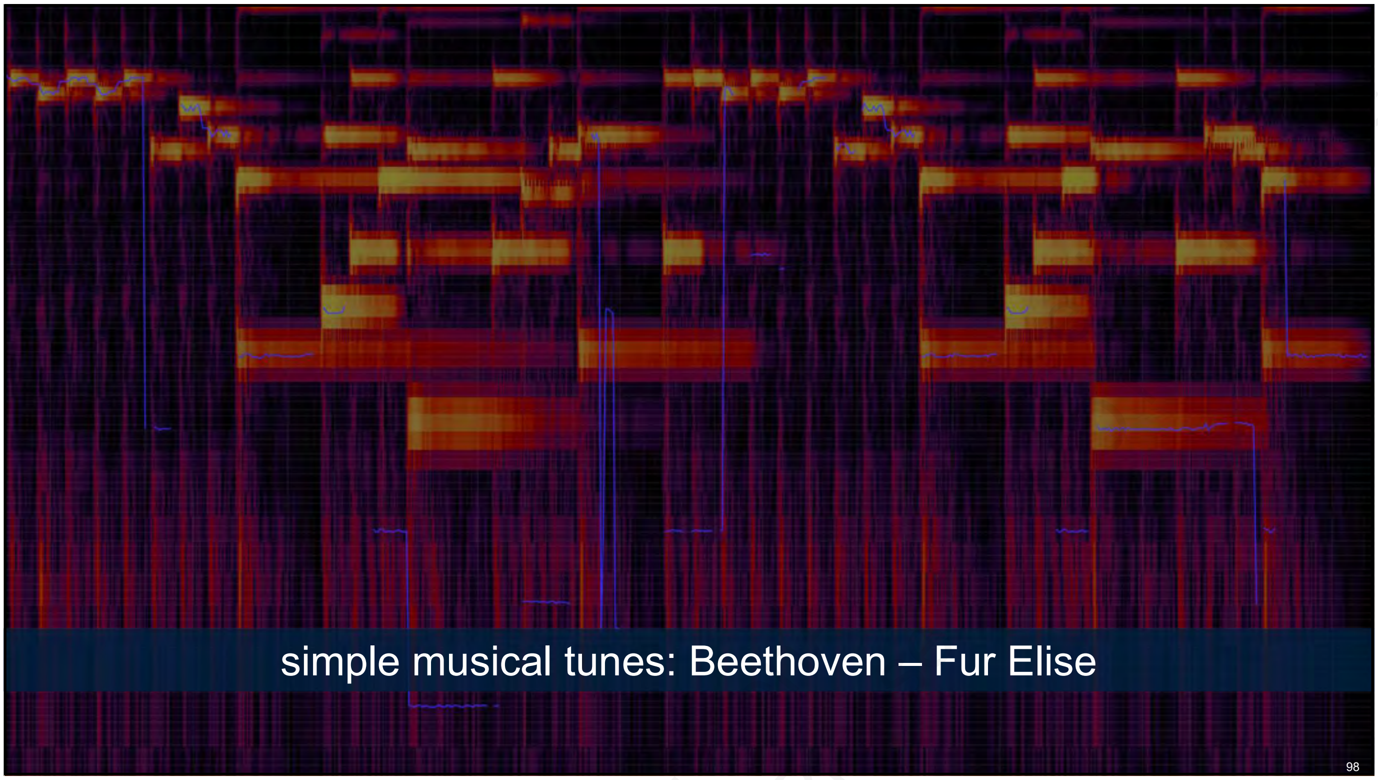
95



96



97

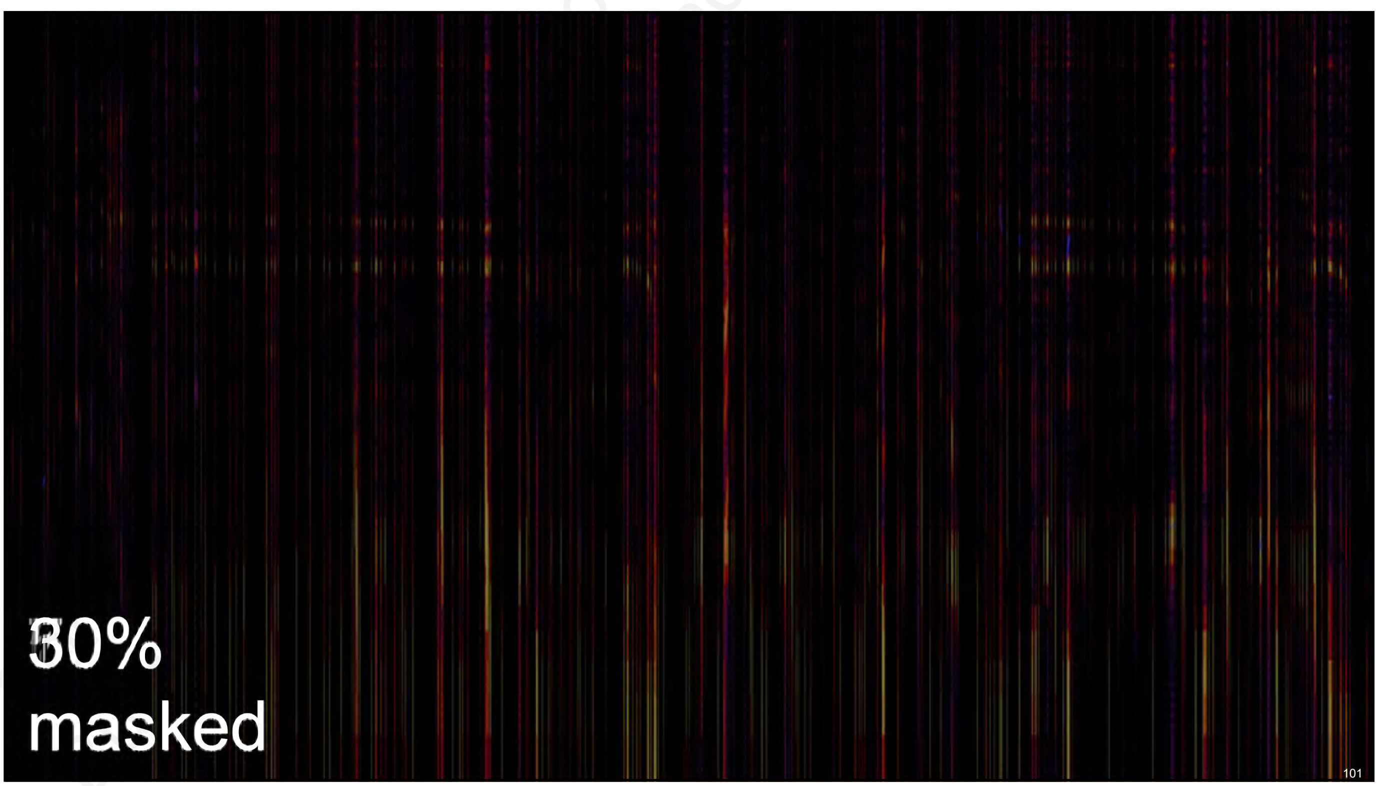

simple musical tunes: Beethoven – Fur Elise
98

98

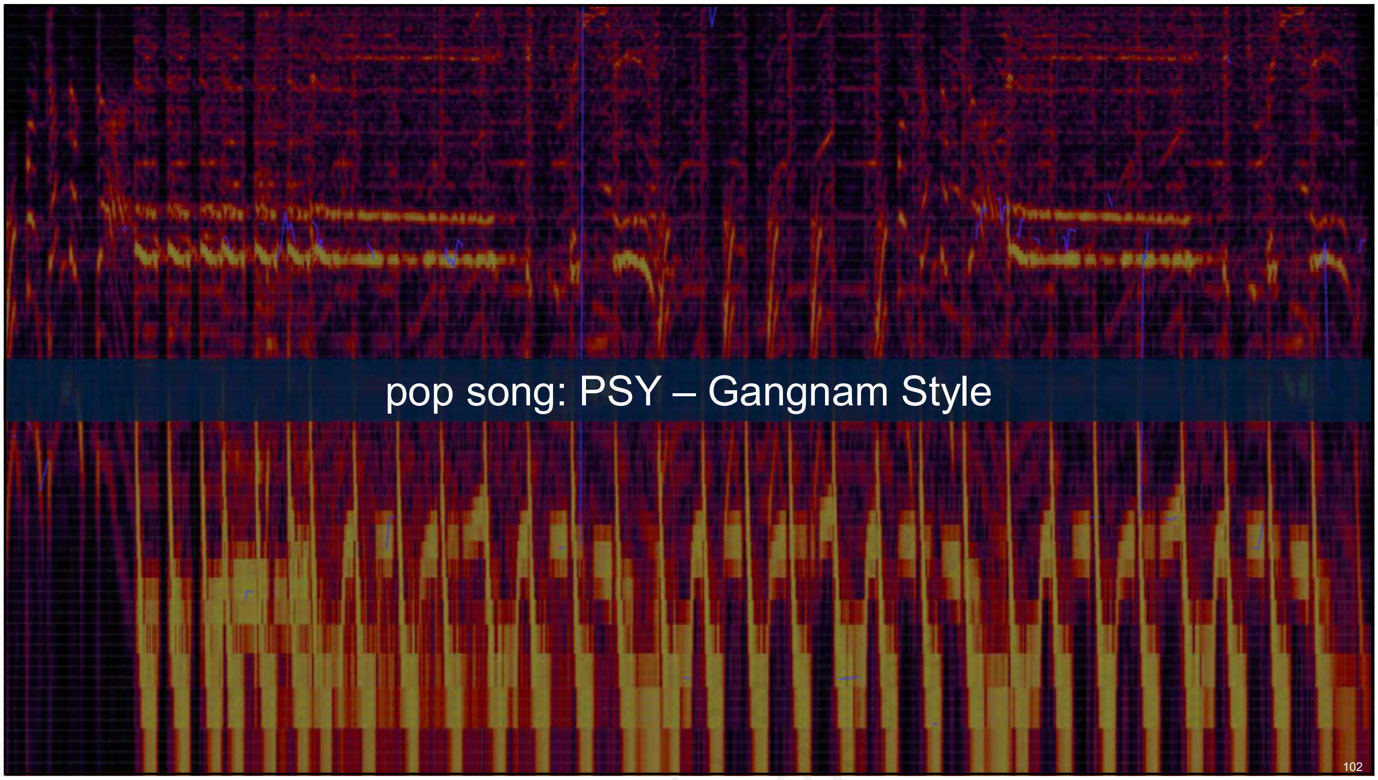

orchestral music: Beethoven – 5th Symphony
99

99

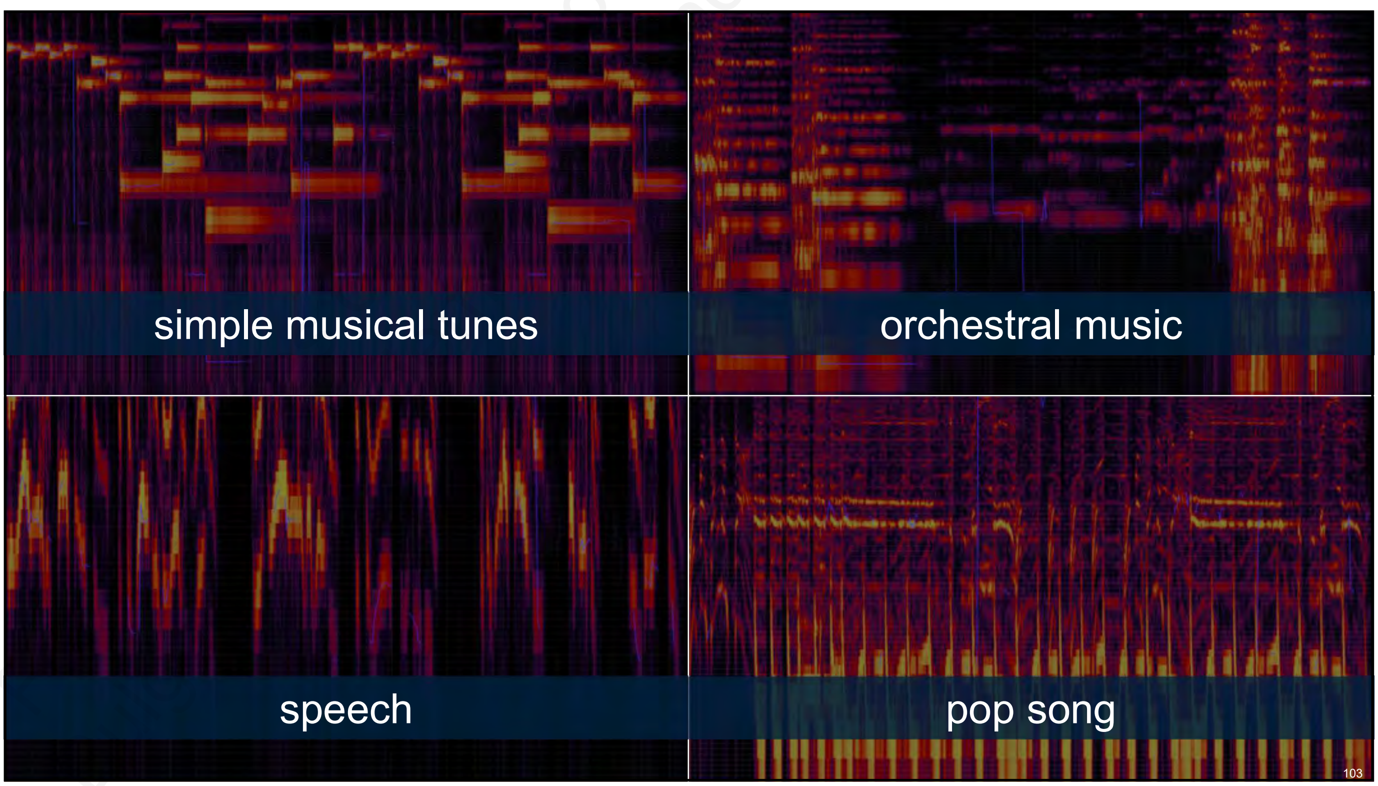
100



101


pop song: PSY – Gangnam Style
102

102


simple musical tunes
orchestral music
speech
pop song
103

103

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

PROS.ai

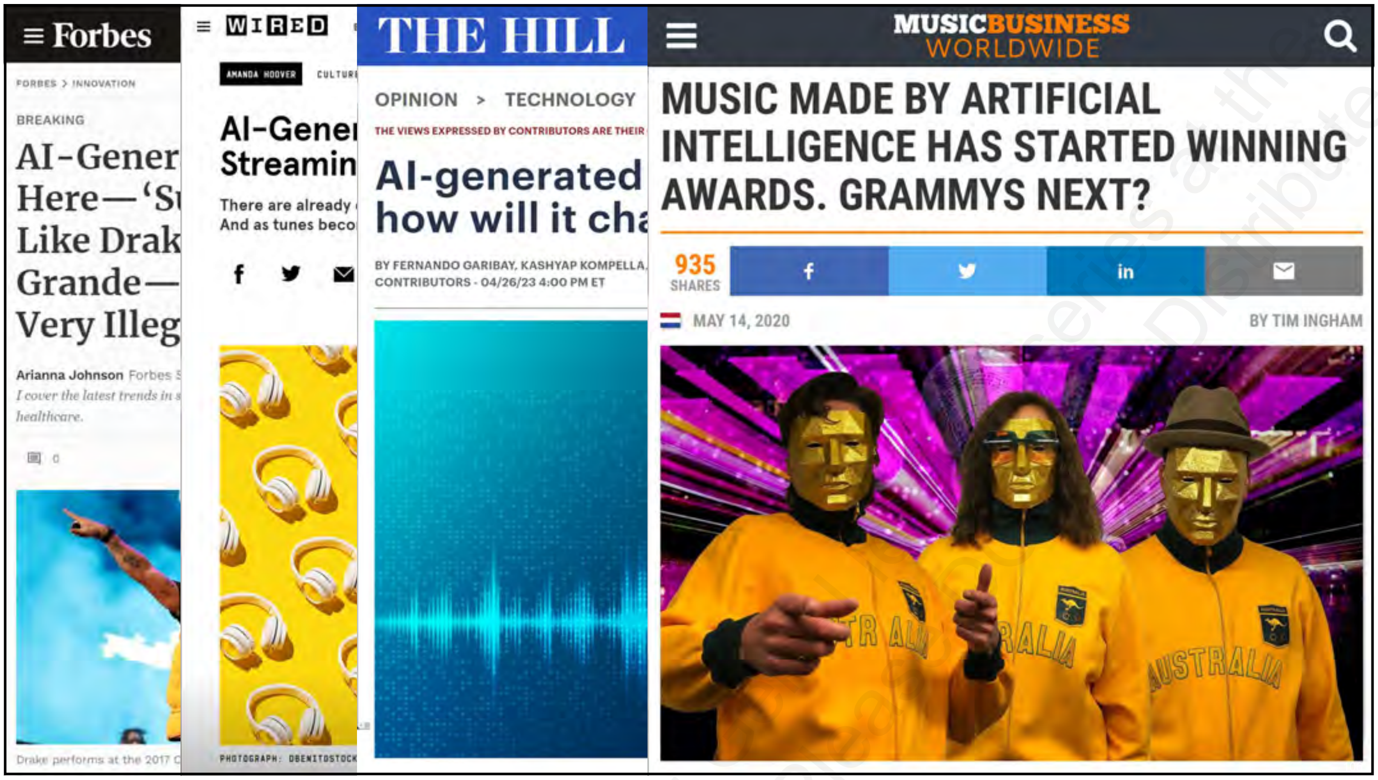
©2024 PROS, Inc. All rights reserved. Confidential and Proprietary.

twitter: @mich8elwu
linkedin.com/in/MichaelWuPhD

104

104

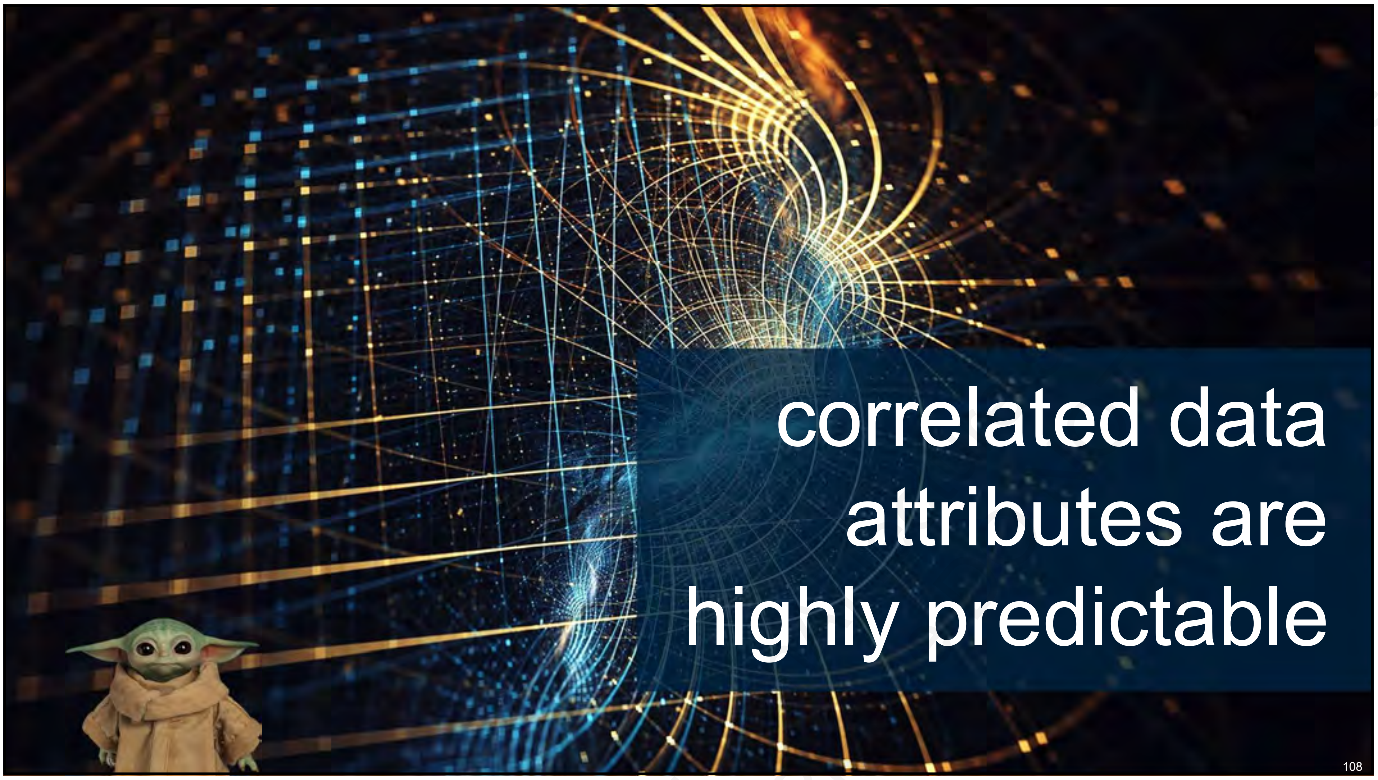
105


Forbes
FORBES > INNOVATION
BREAKING
WIRED
THE HILL
OPINION > TECHNOLOGY
MUSICBUSINESS WORLDWIDE
MUSIC MADE BY ARTIFICIAL INTELLIGENCE HAS STARTED WINNING AWARDS. GRAMMYS NEXT?
935 SHARES
MAY 14, 2020
BY TIM INGHAM

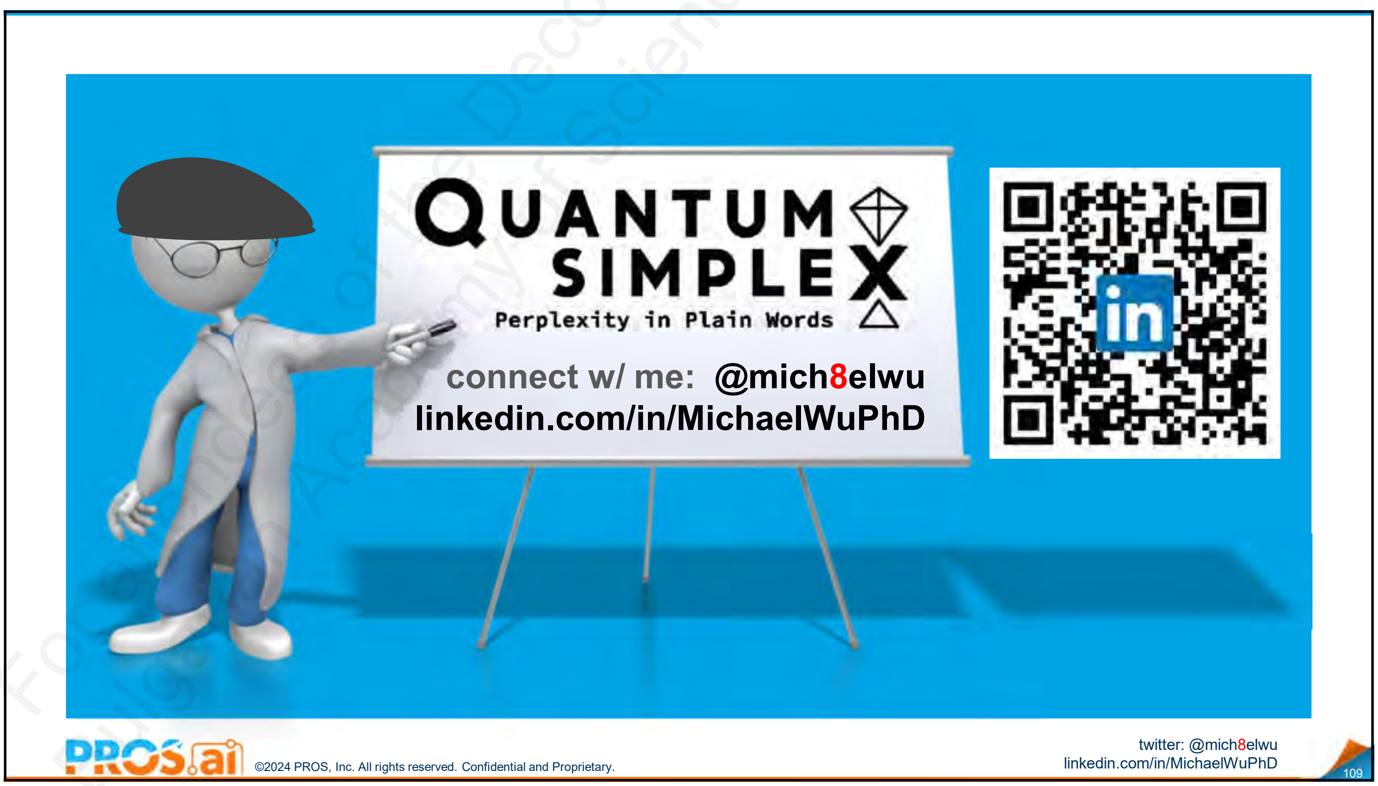
106


isn't that amazing?

107


correlated data attributes are highly predictable
108

108


QUANTUM SIMPLEX
Perplexity in Plain Words
connect w/ me: @mich8elwu
linkedin.com/in/MichaelWuPhD
PROS.ai
©2024 PROS, Inc. All rights reserved. Confidential and Proprietary.
twitter: @mich8elwu
linkedin.com/in/MichaelWuPhD
109

109

110

Navigating the Limits of GenAI
in Business Decisions—Beyond Productivity

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

2024.04.18

1869

PROS.

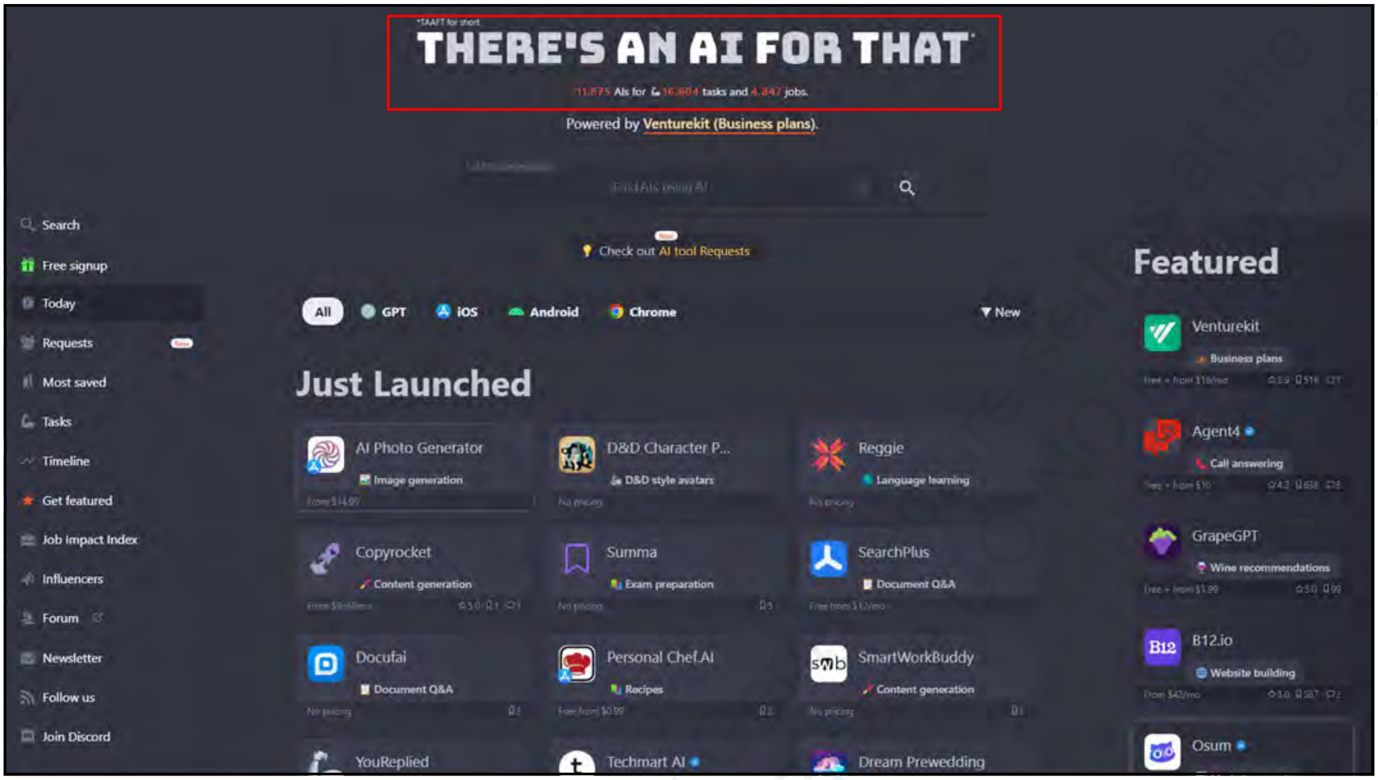
111

112

113

114

115


THERE'S AN AI FOR THAT
Powered by Venturekit (Business plans).
Search
Free signup
Today
Requests
Most saved
Tasks
Timeline
Get featured
Job Impact Index
Influencers
Forum
Newsletter
Follow us
Join Discord
Check out AI tool Requests
All
GPT
iOS
Android
Chrome
New
Just Launched
AI Photo Generator
Image generation
D&D Character P...
D&D style avatars
Reggie
Language learning
Copyrocket
Content generation
Summa
Exam preparation
SearchPlus
Document Q&A
Docufai
Document Q&A
Personal Chef.AI
Recipes
SmartWorkBuddy
Content generation
YouReplied
Techmart AI
Dream Prewedding
Featured
Venturekit
Business plans
Agent4
Call answering
GrapeGPT
Wine recommendations
B12.io
Website building
Osum

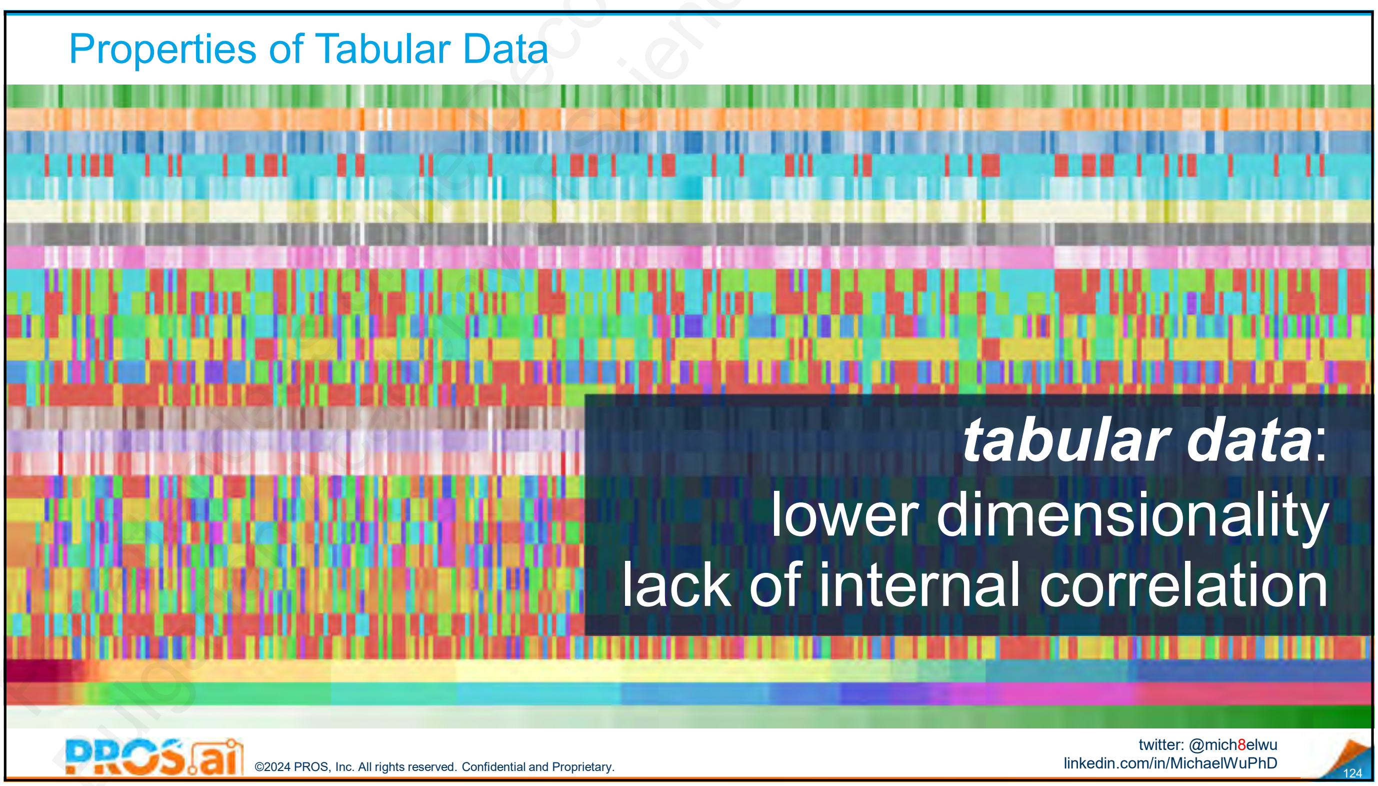
121


there aren't many mission-critical business applications based on GenAI
122

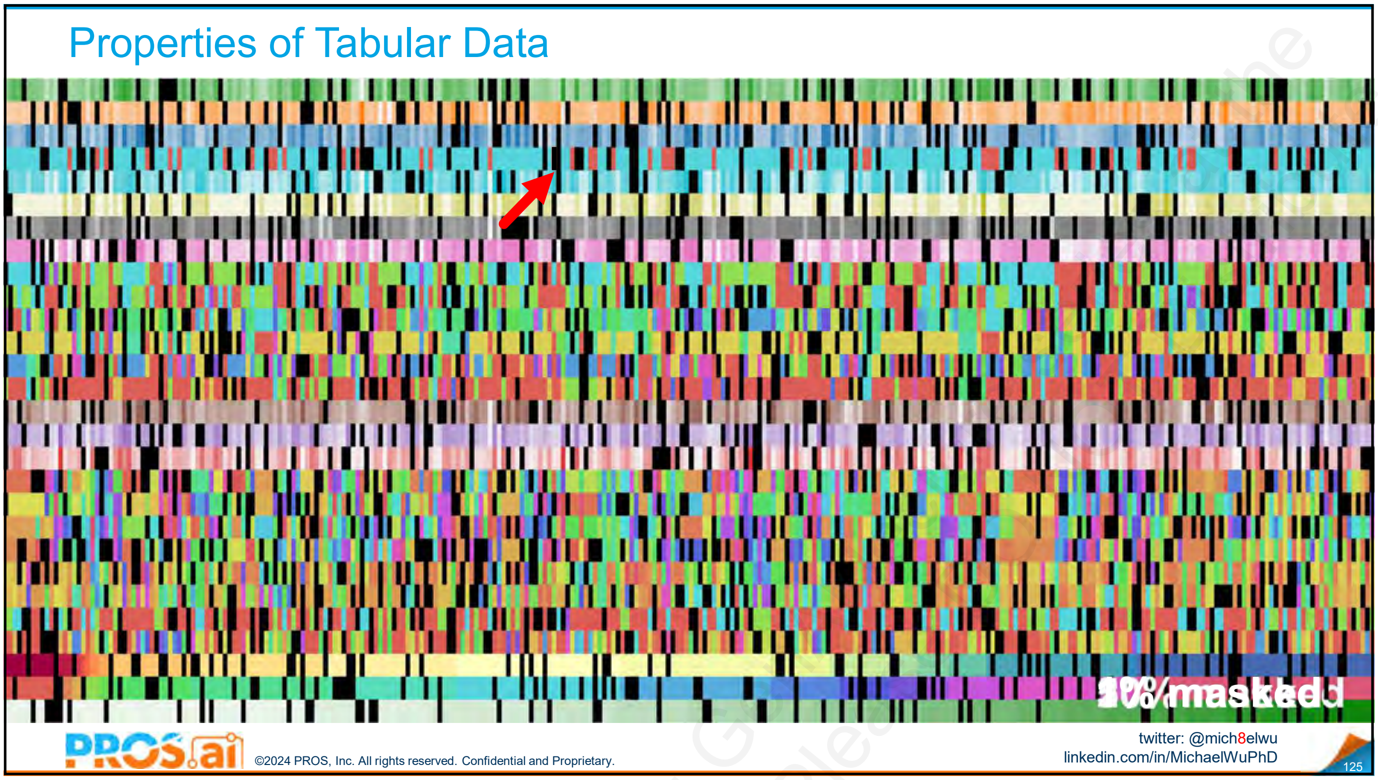
122


most mission-critical business decisions are made with the support of some sort of tabular data
123

123


Properties of Tabular Data
tabular data:
lower dimensionality
lack of internal correlation
PROS.ai
©2024 PROS, Inc. All rights reserved. Confidential and Proprietary.
twitter: @mich8elwu
linkedin.com/in/MichaelWuPhD
124

124


Properties of Tabular Data
©2024 PROS, Inc. All rights reserved. Confidential and Proprietary.
twitter: @mich8elwu
linkedin.com/in/MichaelWuPhD
125

125


what does it take to drive profitable growth?
PROFIT
127

127

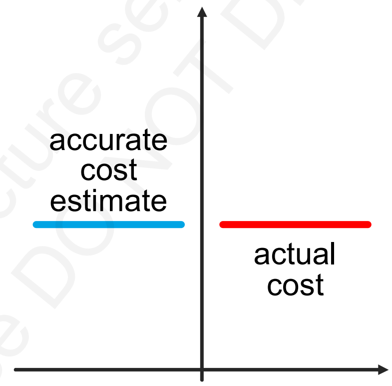
What Does it Take to Drive Profitability?

profit = revenue − cost

must have ability to manage both terms

optimize revenue

estimate cost



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

twitter: @mich8elwu
linkedin.com/in/MichaelWuPhD

129

For attendees of the Decoding GenAI lecture series at the Bulgarian Academy of Science. Please DO NOT Distribute.

What Does it Take to Drive Profitability?

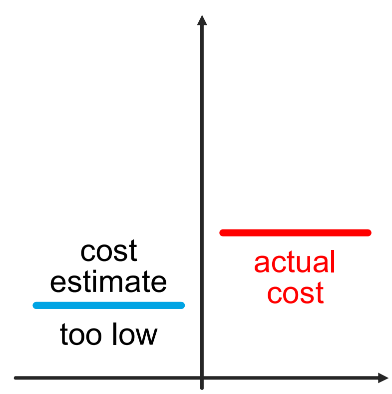
profit = revenue − cost

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

130

What Does it Take to Drive Profitability?

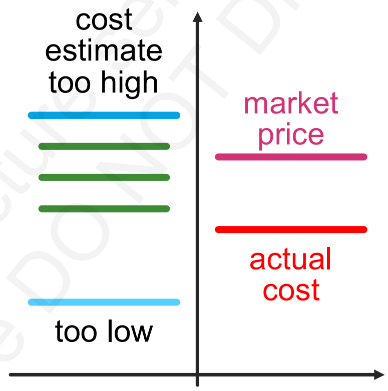
profit = revenue − cost

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

131

What Does it Take to Drive Profitability?

profit = revenue − cost

must have ability to manage both terms

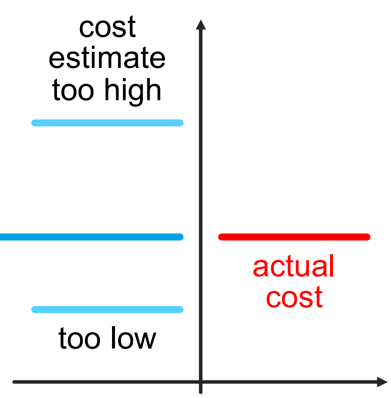
optimize revenue | **estimate cost**

- 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) ≠ accurate cost estimate NOW

- 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

132

For attendees of the Decoding GenAI lectures at the Bulgarian Academy of Science. Please DO NOT Distribute.

What Does it Take to Drive Profitability?

profit = revenue − cost

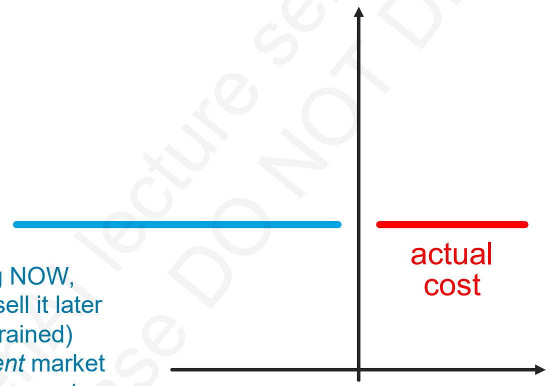
must have ability to manage both terms

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) ≠ 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

PROS.ai ©2024 PROS, Inc. All rights reserved. Confidential and Proprietary.
twitter: @mich8elwu
linkedin.com/in/MichaelWuPhD

133

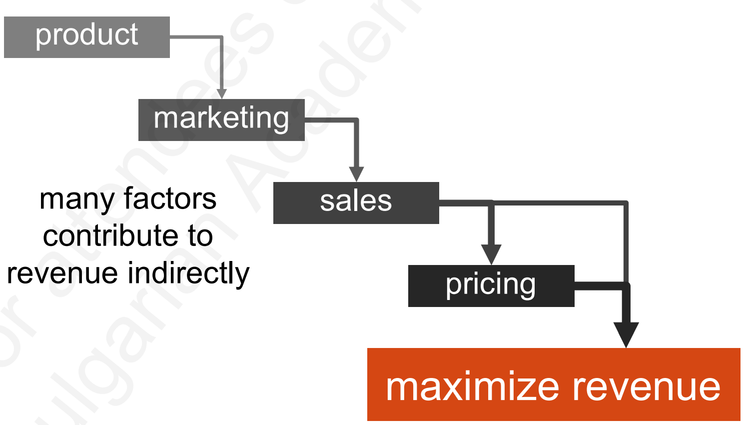
What Does it Take to Drive Profitability?

profit = revenue − 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***

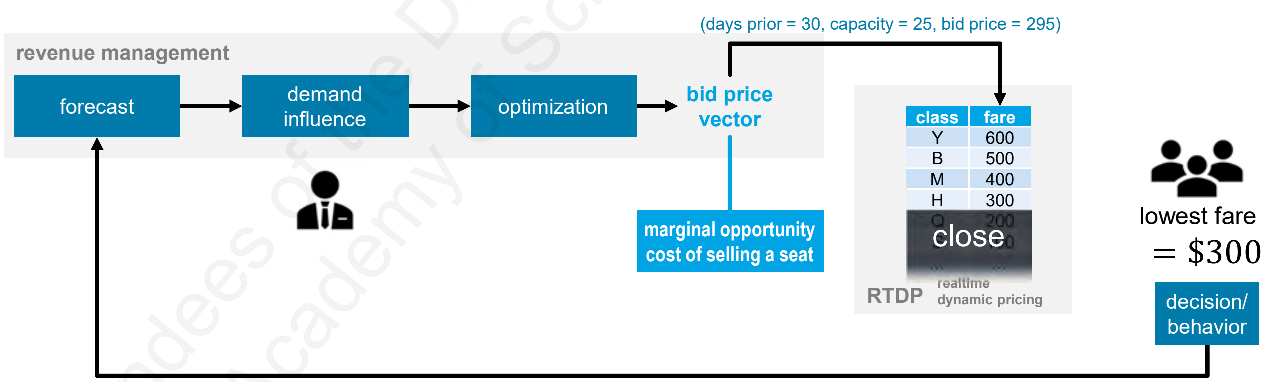
PROS.ai ©2024 PROS, Inc. All rights reserved. Confidential and Proprietary.
twitter: @mich8elwu
linkedin.com/in/MichaelWuPhD

135

For attendees of the Decoding Gen... lecture series at the Bulgarian Academy of Science. Please DO NOT Distribute.

137

Conventional RM: Controls the Fares Seen by Shoppers



automation — automated generation of bid prices

learning — improved bid prices through the update of info from the market

PROS.ai

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

twitter: @mich8elwu
linkedin.com/in/MichaelWuPhD

139

139

Definition of AI

AI = machine mimicry of certain aspects of human behaviors with 2 important characteristics:

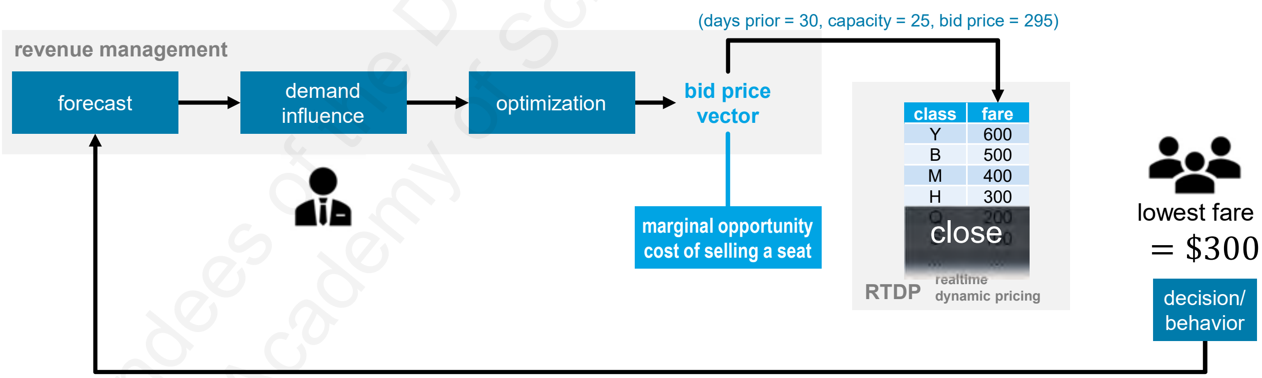
©2024 PROS, Inc. All rights reserved. Confidential and Proprietary.

twitter: @mich8elwu
linkedin.com/in/MichaelWuPhD

140

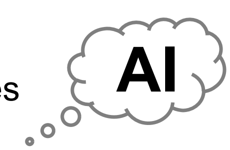
For Attendees of the Decoding GenAI lecture series at the Bulgarian Academy of Science. Please DO NOT Distribute.

Conventional RM: Controls the Fares Seen by Shoppers



automation — automated generation of bid prices

learning — improved bid prices through the update of info from the market

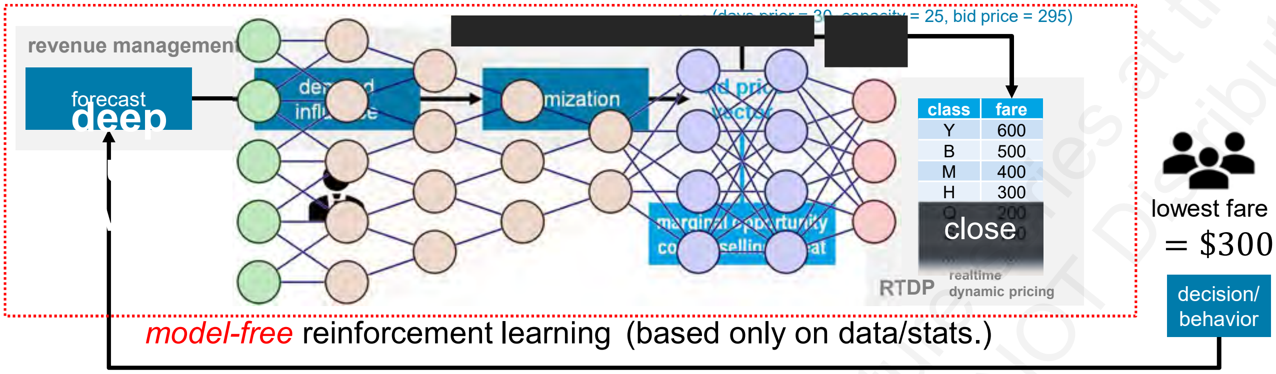


©2024 PROS, Inc. All rights reserved. Confidential and Proprietary.

twitter: @mich8elwu
linkedin.com/in/MichaelWuPhD

141

Conventional RM: Controls the Fares Seen by Shoppers



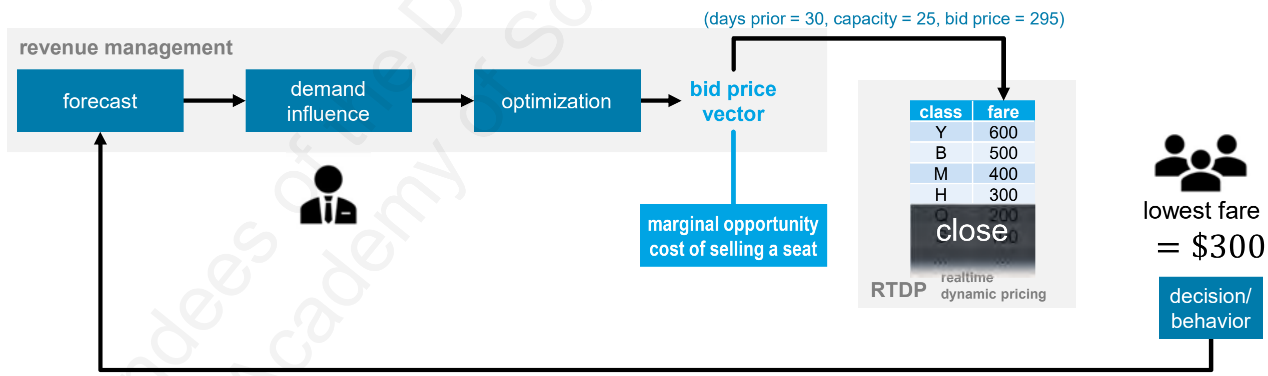
automation — automated generation of bid prices

learning — improved bid prices through the update of info from the market

PROS.ai ©2024 PROS, Inc. All rights reserved. Confidential and Proprietary.
twitter: @mich8elwu
linkedin.com/in/MichaelWuPhD
142

142

Conventional RM: Controls the Fares Seen by Shoppers



model-free reinforcement learning (based only on data/stats.)

flexibility (generality)

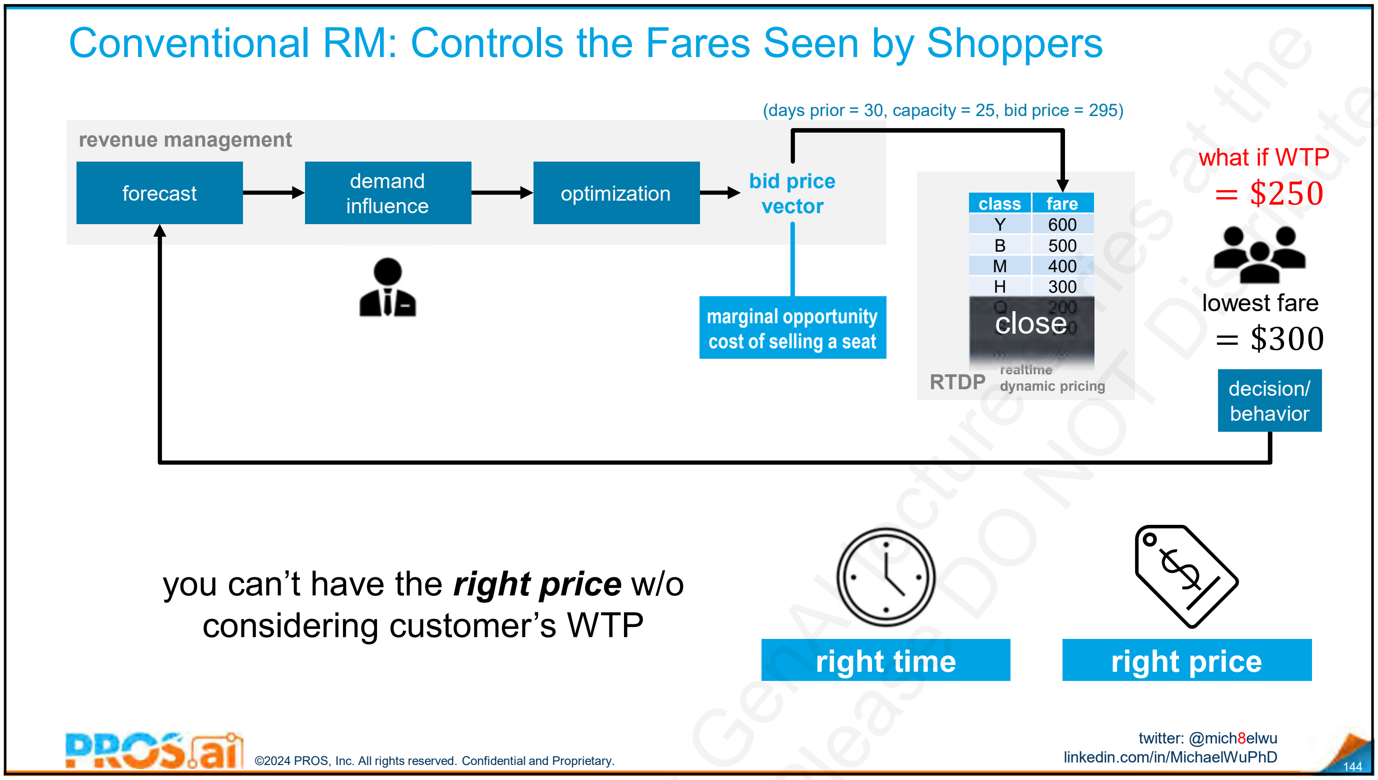
model-based reinforcement learning (based on known mechanisms + data/stats.)

transparency (interpretability)
manipulability + control
data efficiency

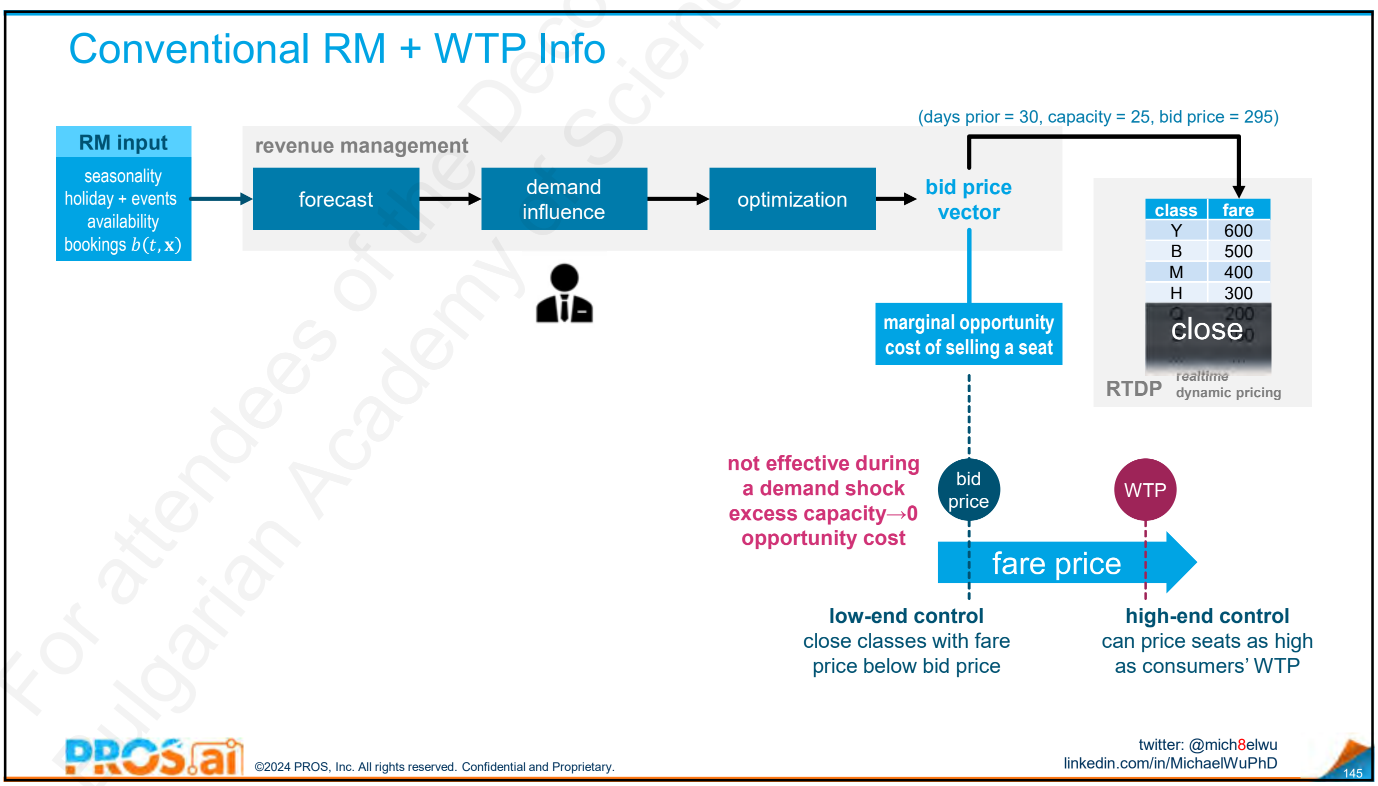
PROS.ai ©2024 PROS, Inc. All rights reserved. Confidential and Proprietary.
twitter: @mich8elwu
linkedin.com/in/MichaelWuPhD
143

143

For attendees of the Decoding GenAI lectures at the Bulgarian Academy of Science. Please DO NOT Distribute.



144



145

For attendees of the Decoding GenAI Architecture Series at the Bulgarian Academy of Science. Please DO NOT Distribute.

Conventional RM + WTP Info

(days prior = 30, capacity = 25, bid price = 295)

RM input
seasonality
holiday + events
availability
bookings $b(t,\mathbf{x})$
prices $p(t,\mathbf{x})$

revenue management
forecast → demand influence → optimization → **bid price vector**

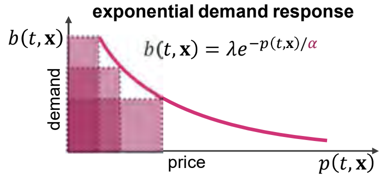
marginal opportunity cost of selling a seat

class	fare
Y	600
B	500
M	400
H	300
Q	200

close

RTDP realtime dynamic pricing

exponential demand response



revenue: $R_{tot} = b[p(t,\mathbf{x})]p(t,\mathbf{x})$ cost = bid price
margin:

$$\frac{\partial}{\partial p} M_{tot} = \frac{\partial}{\partial p} \exp\left(\frac{-p}{\alpha}\right)[p-c] = 0$$

not effective during a demand shock excess capacity→0 opportunity cost

bid price WTP

fare price

low-end control
close classes with fare price below bid price

high-end control
can price seats as high as consumers' WTP

©2024 PROS, Inc. All rights reserved. Confidential and Proprietary.
twitter: @mich8elwu
linkedin.com/in/MichaelWuPhD
146

146

Conventional RM + WTP Info

(days prior = 30, capacity = 25, bid price = 295)

RM input
seasonality
holiday + events
availability
bookings $b(t,\mathbf{x})$
prices $p(t,\mathbf{x})$

revenue management
forecast → demand influence → optimization → **bid price vector**

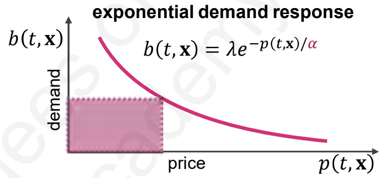
marginal opportunity cost of selling a seat

class	fare
Y	600
B	500
M	400
H	300
Q	200

close

RTDP realtime dynamic pricing

exponential demand response



revenue: $R_{tot} = b[p(t,\mathbf{x})]p(t,\mathbf{x})$ cost = bid price
margin: $M_{tot} = b[p(t,\mathbf{x})][p(t,\mathbf{x}) - c(t,\mathbf{x})]$

$$\frac{\partial}{\partial p} M_{tot} = \frac{\partial}{\partial p} \exp\left(\frac{-p}{\alpha}\right)[p-c] = 0$$

$$\exp\left(\frac{-p}{\alpha}\right) \cdot 1 + [p-c]\exp\left(\frac{-p}{\alpha}\right)\frac{-1}{\alpha} = 0$$

$$\cancel{\exp\left(\frac{-p}{\alpha}\right)}\left[1 - \frac{[p-c]}{\alpha}\right] = 0$$

$$\alpha = [p-c]$$

$$p^* = c + \alpha$$

not effective during a demand shock excess capacity→0 opportunity cost

bid price WTP

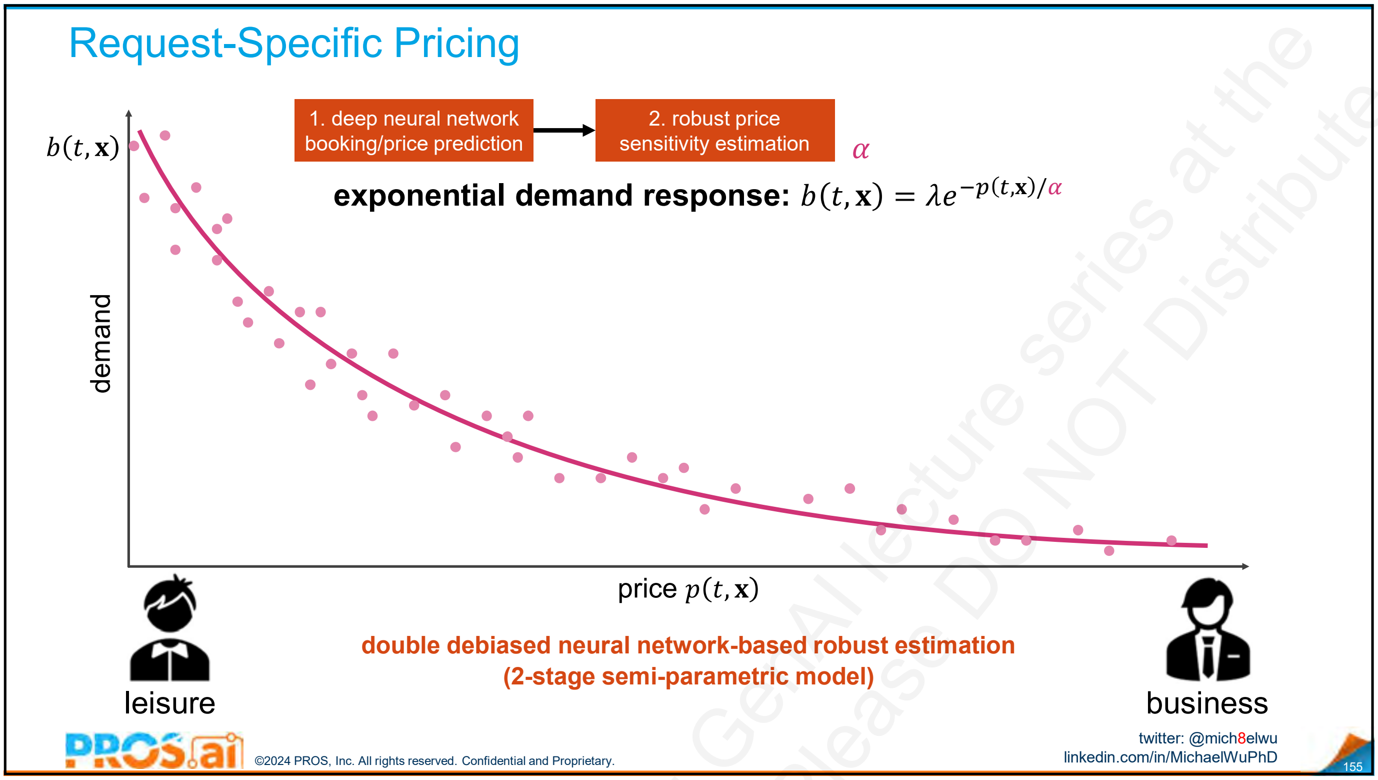
fare price

c ←- α -→ p^*

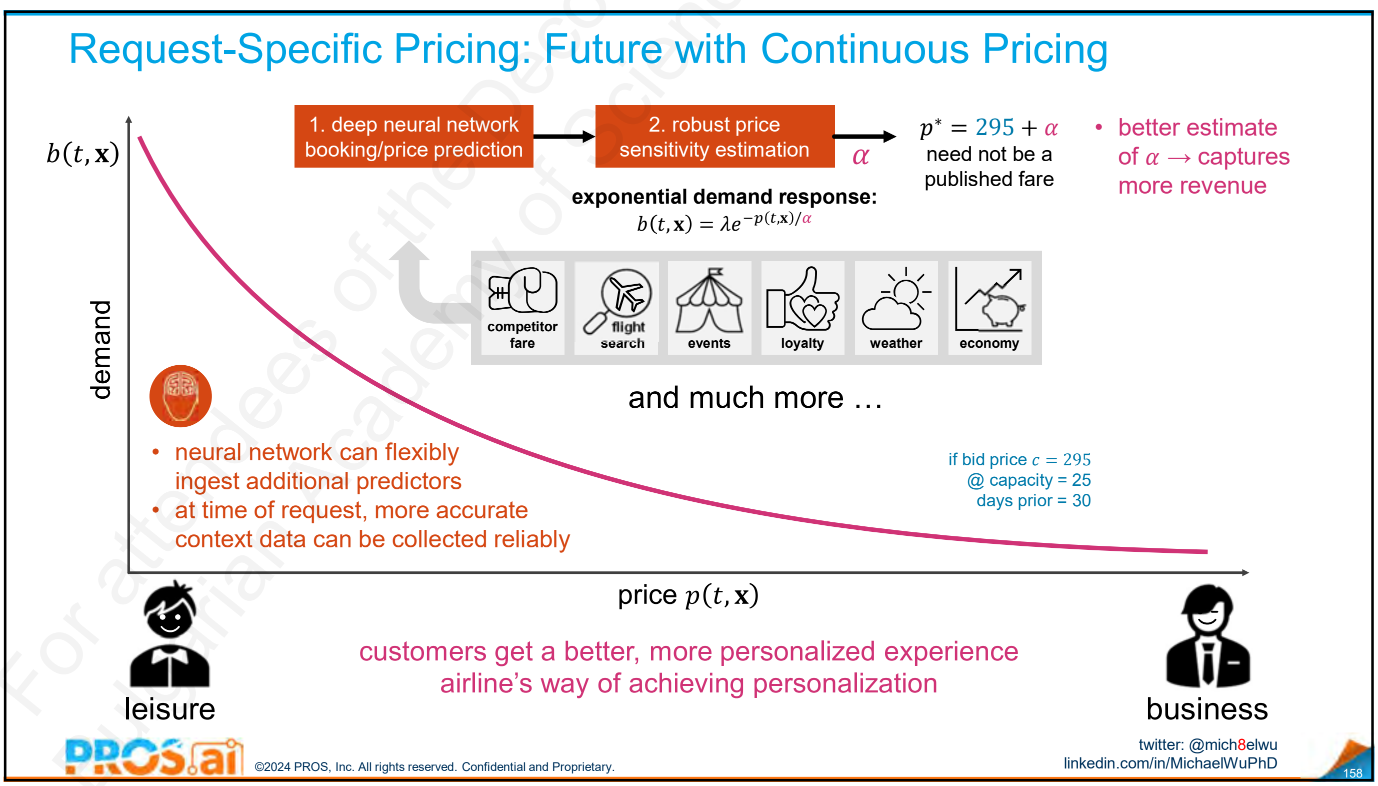
©2024 PROS, Inc. All rights reserved. Confidential and Proprietary.
twitter: @mich8elwu
linkedin.com/in/MichaelWuPhD
147

147

For attendees of the Decoding Gen AI lecture series at the Bulgarian Academy of Science. Please DO NOT distribute.



155

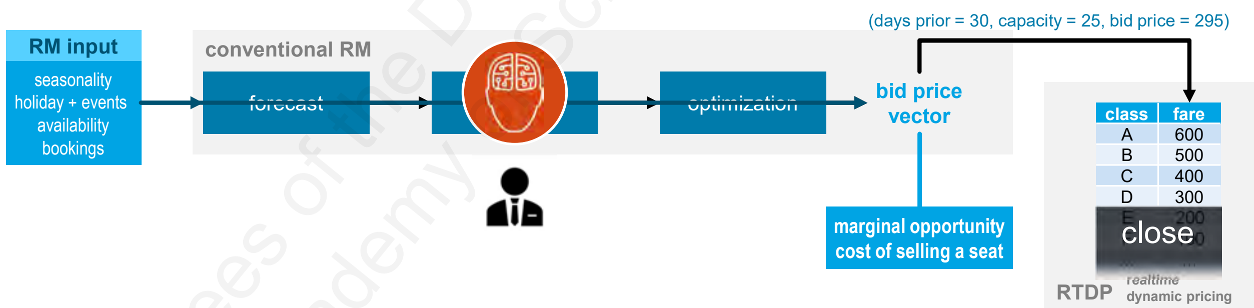


158

For attendees of the Decoding GenAI lecture series at the Bulgarian Academy of Science. Please DO NOT Distribute.

160

DiANNe: **Di**rect **A**daptive **N**eural **Ne**twork RM



forecasting is challenging

1) demand volatility (e.g. cargo's late & lumpy demand arrival)
2) data unavailability (e.g. industries that don't have a RM practice)

Direct — gets bid prices *directly* from historical transactions (forecast free)

Adaptive — highly *adaptive*, robust to extreme demand shocks

Neural **Ne**twork — uses deep *neural network* as the predictive model

PROS.ai

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

twitter: @mich8elwu
linkedin.com/in/MichaelWuPhD

163

For attendees of the Decoding AI lecture series at the Bulgarian Academy of Science. Please DO NOT Distribute.

163

DiANNe: **Di**rect **A**daptive **N**eural **Ne**twork RM

RM input
historical transactions (bookings)

conventional RM

(days prior = 30, capacity = 25, bid price = 295)

bid price vector

marginal opportunity cost of selling a seat

class	fare
A	600
B	500
C	400
D	300
E	200

close

RTDP *realtime* dynamic pricing

Direct — gets bid prices *directly* from historical transactions (forecast free)

Adaptive — highly *adaptive*, robust to extreme demand shocks

Neural **Ne**twork — uses deep *neural network* as the predictive model

no forecasting: lighter data requirement

no optimization: lighter implementation + lower cost

flexible data sources, supports additional covariates

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

twitter: @mich8elwu
linkedin.com/in/MichaelWuPhD

164

164

Asia air cargo carrier

19% bid price prediction ***MSE*** reduction

3.2% attainable ***revenue*** lift

EU air cargo carrier

2.14% bid price prediction ***MSE*** reduction

3.07% attainable ***revenue*** lift

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

twitter: @mich8elwu
linkedin.com/in/MichaelWuPhD

165

165

For attendees of the Decoding Gen AI lecture series at the Museum of Science. Please DO NOT distribute.

What Does it Take to Drive Profitability?

profit = revenue − cost

must have ability to manage both terms

optimize revenue

estimate cost

more accurate cost estimation through opportunity cost can drive

3+%

revenue lift

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

twitter: @mich8elwu
linkedin.com/in/MichaelWuPhD

166

166

169

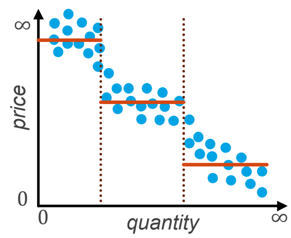
For attendees of the Decoding GenAI lecture series at the Bulgarian Academy of Science. Please DO NOT Distribute.

wROXANNe: **w**in-**R**ate **O**ptimized e**X**plainable **A**I with **N**eural **N**etwork

segmentation

forecast

optimize



- limits accuracy
- don't support continuous attributes
- limited cross-segment knowledge sharing
- data sparsity

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

twitter: @mich8elwu
linkedin.com/in/MichaelWuPhD

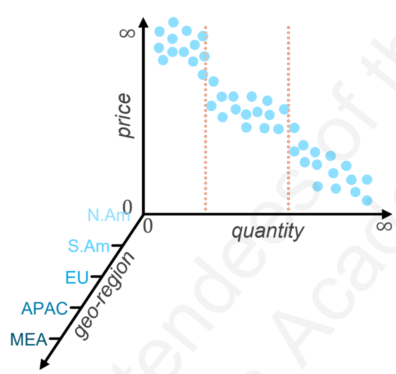
172

wROXANNe: **w**in-**R**ate **O**ptimized e**X**plainable **A**I with **N**eural **N**etwork

segmentation

forecast

optimize



- limits accuracy
- don't support continuous attributes
- limited cross-segment knowledge sharing
- data sparsity

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

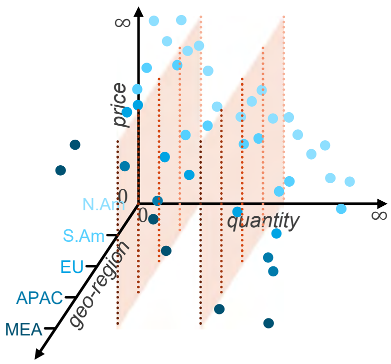
twitter: @mich8elwu
linkedin.com/in/MichaelWuPhD

173

For attendees of the Decoding GenAI lecture series at the Bulgarian Academy of Science. Please DO NOT Distribute.

wROXANNe: **w**in-**R**ate **O**ptimized e**X**plainable **A**I with **N**eural **N**etwork

segmentation | forecast | optimize



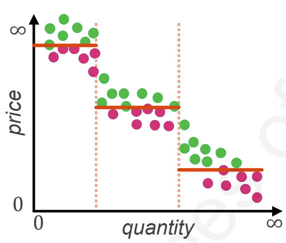
- limits accuracy
- don't support continuous attributes
- limited cross-segment knowledge sharing
- data sparsity

PROS.ai ©2024 PROS, Inc. All rights reserved. Confidential and Proprietary.
twitter: @mich8elwu
linkedin.com/in/MichaelWuPhD
174

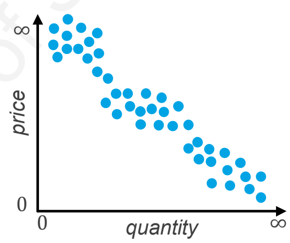
174

wROXANNe: **w**in-**R**ate **O**ptimized e**X**plainable **A**I with **N**eural **N**etwork

segmentation | forecast | optimize

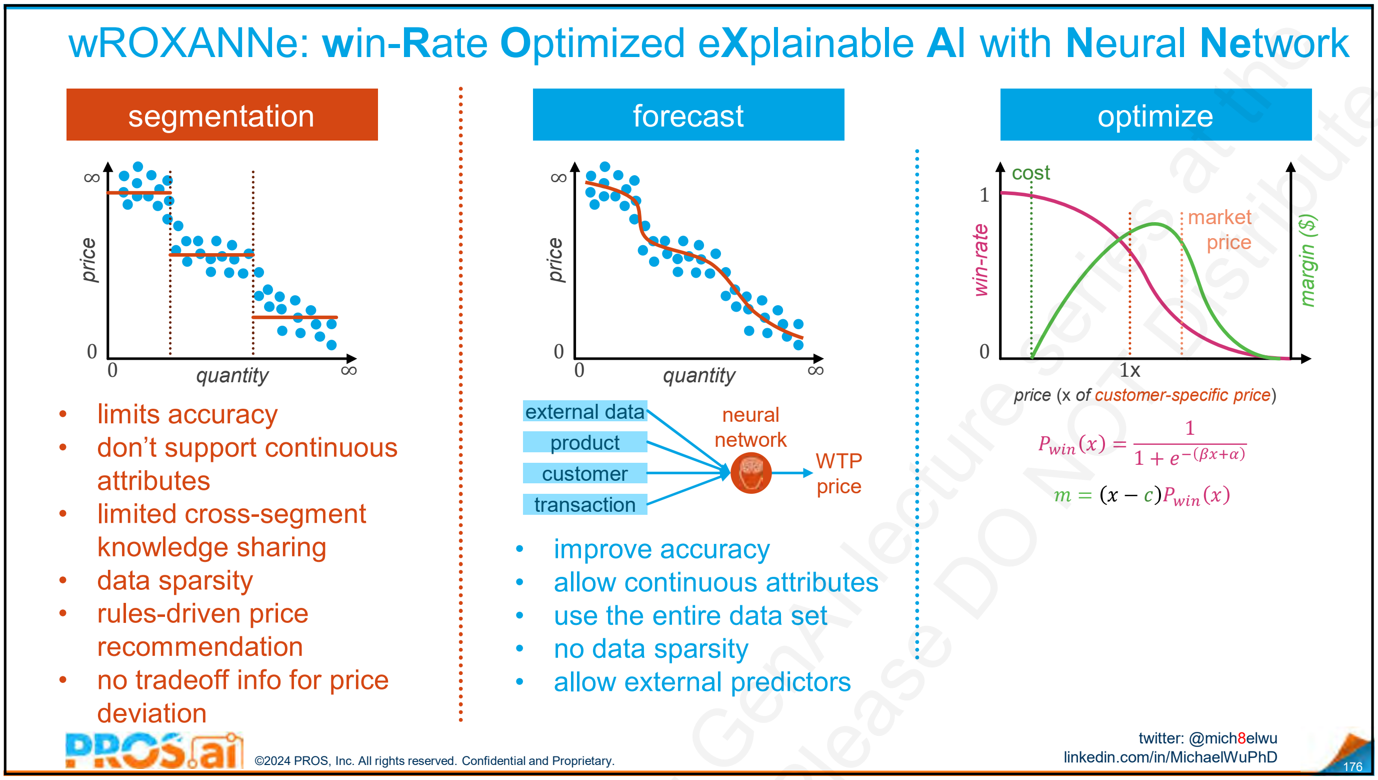


- limits accuracy
- don't support continuous attributes
- limited cross-segment knowledge sharing
- data sparsity
- rules-driven price recommendation
- no tradeoff info for price deviation

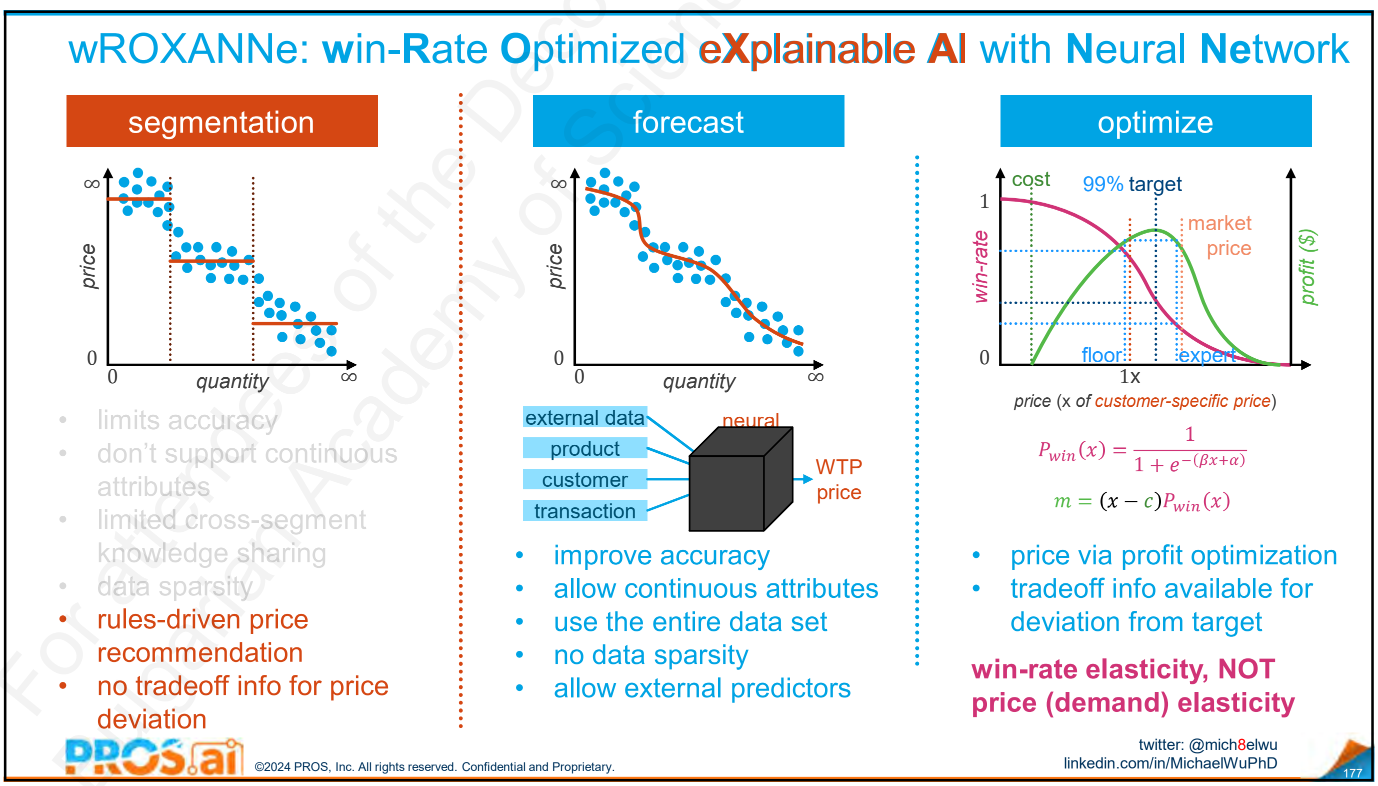


PROS.ai ©2024 PROS, Inc. All rights reserved. Confidential and Proprietary.
twitter: @mich8elwu
linkedin.com/in/MichaelWuPhD
175

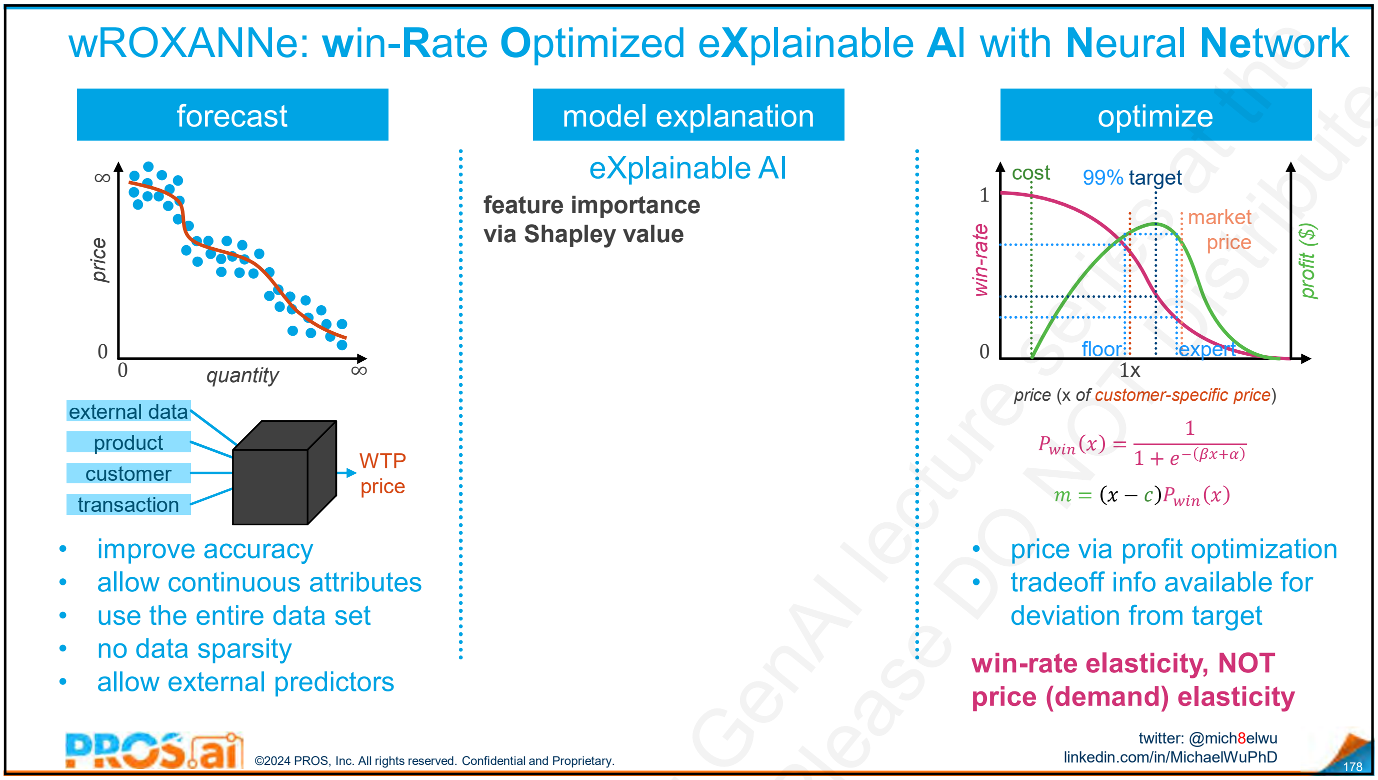
175


wROXANNe: win-Rate Optimized eXplainable AI with Neural Network
segmentation
forecast
optimize
price
quantity
∞
0
limits accuracy
don't support continuous attributes
limited cross-segment knowledge sharing
data sparsity
rules-driven price recommendation
no tradeoff info for price deviation
external data
product
customer
transaction
neural network
WTP price
improve accuracy
allow continuous attributes
use the entire data set
no data sparsity
allow external predictors
cost
market price
win-rate
margin ($)
1
1x
price (x of customer-specific price)
$P_{win}(x) = \frac{1}{1+e^{-(\beta x+\alpha)}}$
$m = (x-c)P_{win}(x)$
PROS.ai
©2024 PROS, Inc. All rights reserved. Confidential and Proprietary.
twitter: @mich8elwu
linkedin.com/in/MichaelWuPhD
176

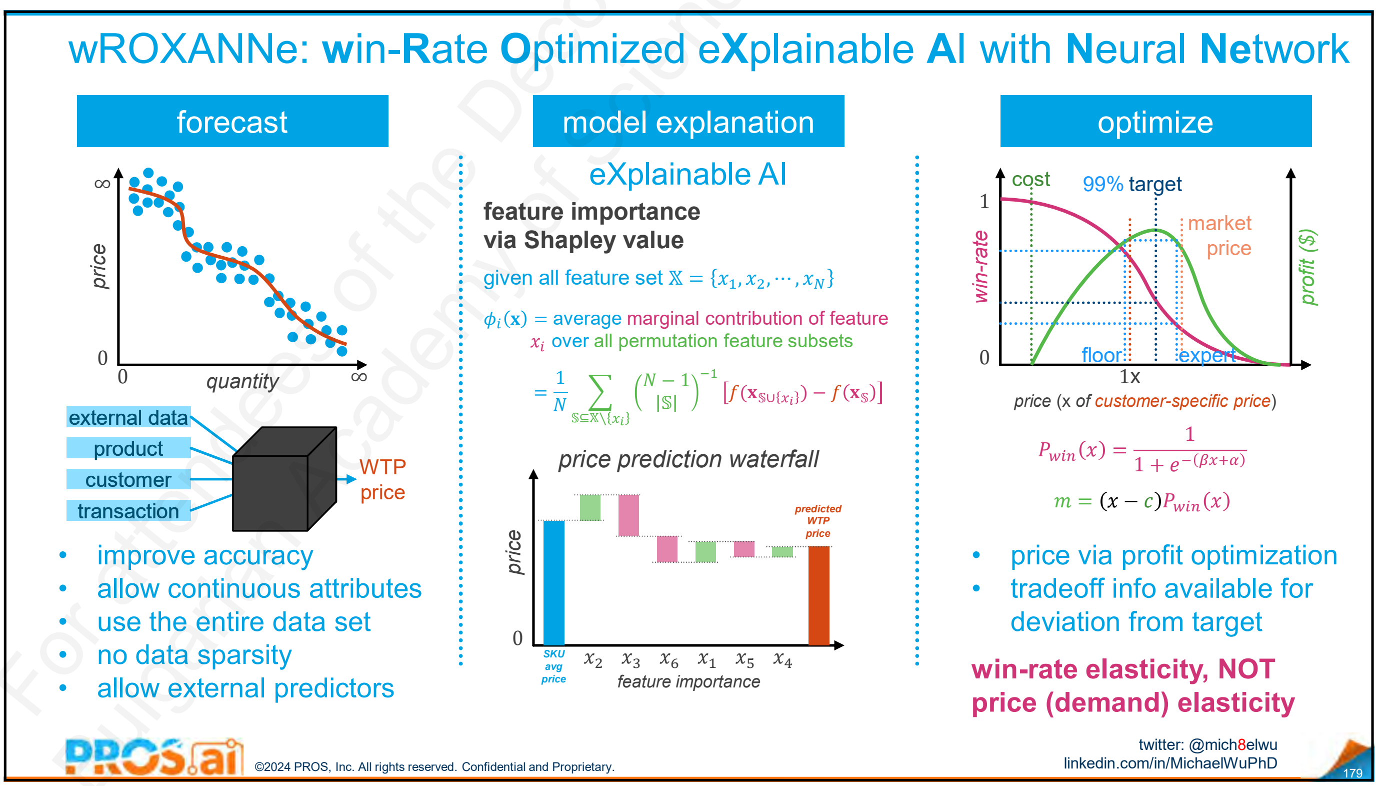
176


wROXANNe: win-Rate Optimized eXplainable AI with Neural Network
segmentation
forecast
optimize
price
quantity
∞
0
limits accuracy
don't support continuous attributes
limited cross-segment knowledge sharing
data sparsity
rules-driven price recommendation
no tradeoff info for price deviation
external data
product
customer
transaction
neural
WTP price
improve accuracy
allow continuous attributes
use the entire data set
no data sparsity
allow external predictors
cost
99% target
market price
win-rate
profit ($)
floor
expert
1
1x
price (x of customer-specific price)
$P_{win}(x) = \frac{1}{1+e^{-(\beta x+\alpha)}}$
$m = (x-c)P_{win}(x)$
price via profit optimization
tradeoff info available for deviation from target
win-rate elasticity, NOT price (demand) elasticity
PROS.ai
©2024 PROS, Inc. All rights reserved. Confidential and Proprietary.
twitter: @mich8elwu
linkedin.com/in/MichaelWuPhD
177

177


wROXANNe: win-Rate Optimized eXplainable AI with Neural Network
forecast
price
quantity
external data
product
customer
transaction
WTP price
• improve accuracy
• allow continuous attributes
• use the entire data set
• no data sparsity
• allow external predictors
model explanation
eXplainable AI
feature importance via Shapley value
optimize
cost
99% target
market price
win-rate
profit ($)
floor
expert
1x
price (x of customer-specific price)
$P_{win}(x) = \frac{1}{1 + e^{-(\beta x + \alpha)}}$
$m = (x - c)P_{win}(x)$
• price via profit optimization
• tradeoff info available for deviation from target
win-rate elasticity, NOT price (demand) elasticity
©2024 PROS, Inc. All rights reserved. Confidential and Proprietary.
twitter: @mich8elwu
linkedin.com/in/MichaelWuPhD
178

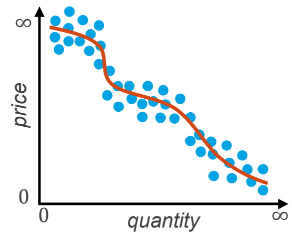
178

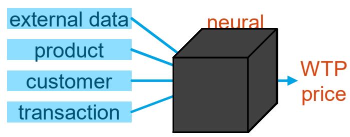

wROXANNe: win-Rate Optimized eXplainable AI with Neural Network
forecast
price
quantity
external data
product
customer
transaction
WTP price
• improve accuracy
• allow continuous attributes
• use the entire data set
• no data sparsity
• allow external predictors
model explanation
eXplainable AI
feature importance via Shapley value
given all feature set $\mathbb{X} = \{x_1, x_2, \cdots, x_N\}$
$\phi_i(\mathbf{x})$ = average marginal contribution of feature x_i over all permutation feature subsets
$= \frac{1}{N} \sum_{\mathbb{S} \subseteq \mathbb{X} \setminus \{x_i\}} \binom{N-1}{|\mathbb{S}|}^{-1} [f(\mathbf{x}_{\mathbb{S} \cup \{x_i\}}) - f(\mathbf{x}_{\mathbb{S}})]$
price prediction waterfall
price
predicted WTP price
SKU avg price
x_2 x_3 x_6 x_1 x_5 x_4
feature importance
optimize
cost
99% target
market price
win-rate
profit ($)
floor
expert
1x
price (x of customer-specific price)
$P_{win}(x) = \frac{1}{1 + e^{-(\beta x + \alpha)}}$
$m = (x - c)P_{win}(x)$
• price via profit optimization
• tradeoff info available for deviation from target
win-rate elasticity, NOT price (demand) elasticity
©2024 PROS, Inc. All rights reserved. Confidential and Proprietary.
twitter: @mich8elwu
linkedin.com/in/MichaelWuPhD
179

179

wROXANNe: win-Rate Optimized eXplainable AI with Neural Network

forecast



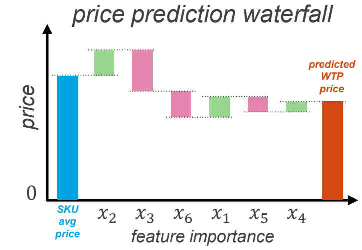


- improve accuracy
- allow continuous attributes
- use the entire data set
- no data sparsity
- allow external predictors

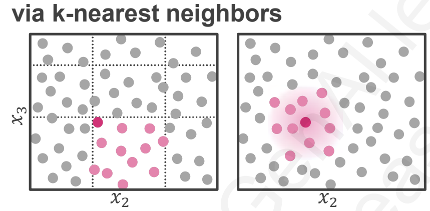
model explanation

eXplainable AI

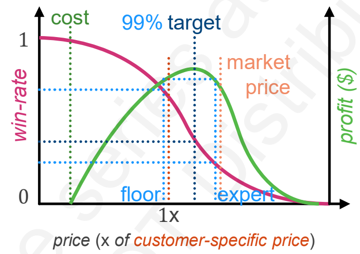
feature importance via Shapley value



comps (comparable transactions) via k-nearest neighbors



optimize



$$P_{win}(x) = \frac{1}{1 + e^{-(\beta x + \alpha)}}$$

$$m = (x - c)P_{win}(x)$$

- price via profit optimization
- tradeoff info available for deviation from target

win-rate elasticity, NOT price (demand) elasticity

PROS.ai

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

twitter: @mich8elwu
linkedin.com/in/MichaelWuPhD

180

180

US food manufacturer

22% price prediction ***RMSE*** reduction

4.5% attainable ***revenue*** lift

EU-based global materials distributor

~50% price prediction ***RMSE*** reduction

11% attainable ***margin*** lift

PROS.ai

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

twitter: @mich8elwu
linkedin.com/in/MichaelWuPhD

181

181

What Does it Take to Drive Profitability?

profit = revenue − cost

must have ability to manage both terms

optimize revenue

more accurate willingness-to-pay prediction using neural networks can drive

4+%

revenue lift

estimate cost

more accurate cost estimation through opportunity cost can drive

3+%

revenue lift

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

twitter: @mich8elwu
linkedin.com/in/MichaelWuPhD

183

183

What Does it Take to Drive Profitability?

profit = revenue − cost

must have ability to manage both terms

optimize revenue

more accurate willingness-to-pay prediction using neural networks can drive

7+% ← 4+% + 3+%

revenue lift

estimate cost

more accurate cost estimation through opportunity cost can drive

revenue lift

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

twitter: @mich8elwu
linkedin.com/in/MichaelWuPhD

188

188

198

Can we Use GenAI in Profit Optimization?

ChatGPT

web crawl
digitized books
scientific journals
wikipedia

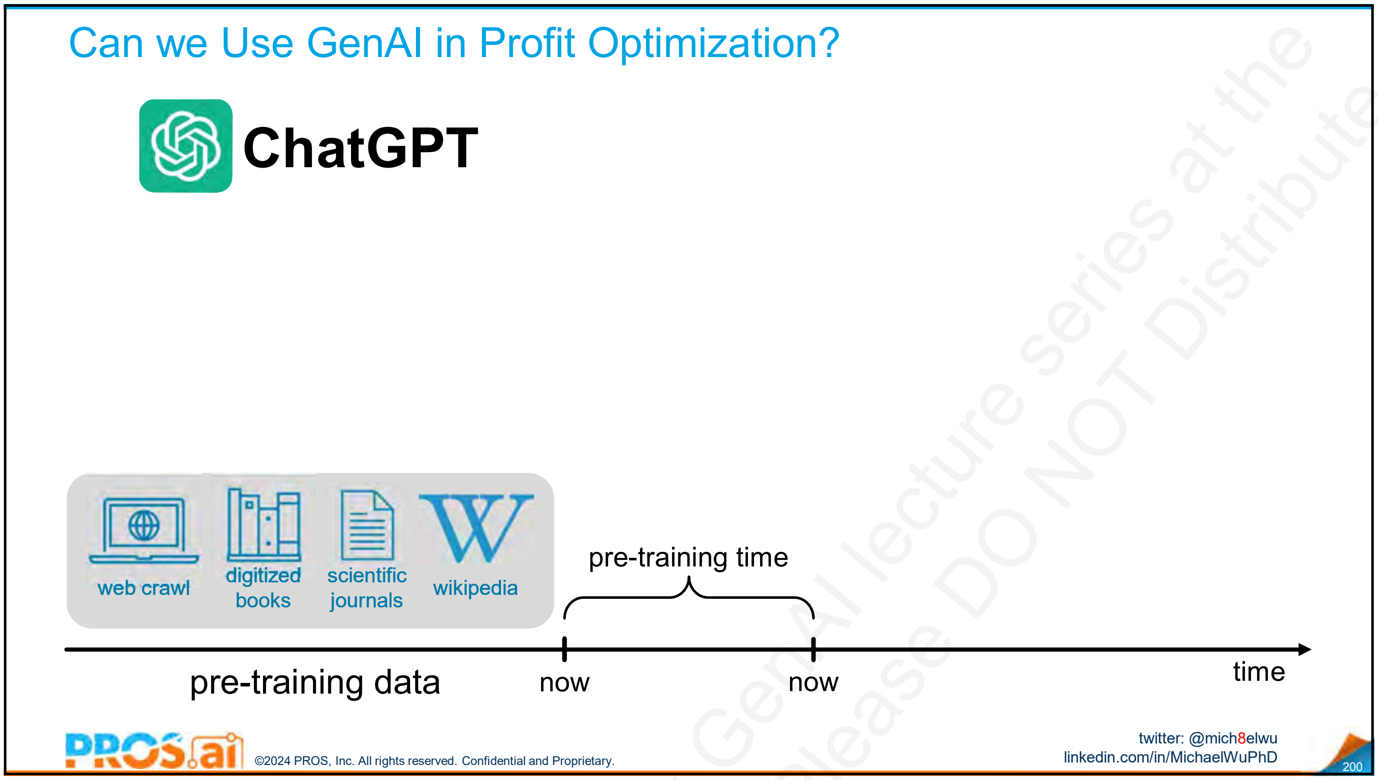
pre-training data

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

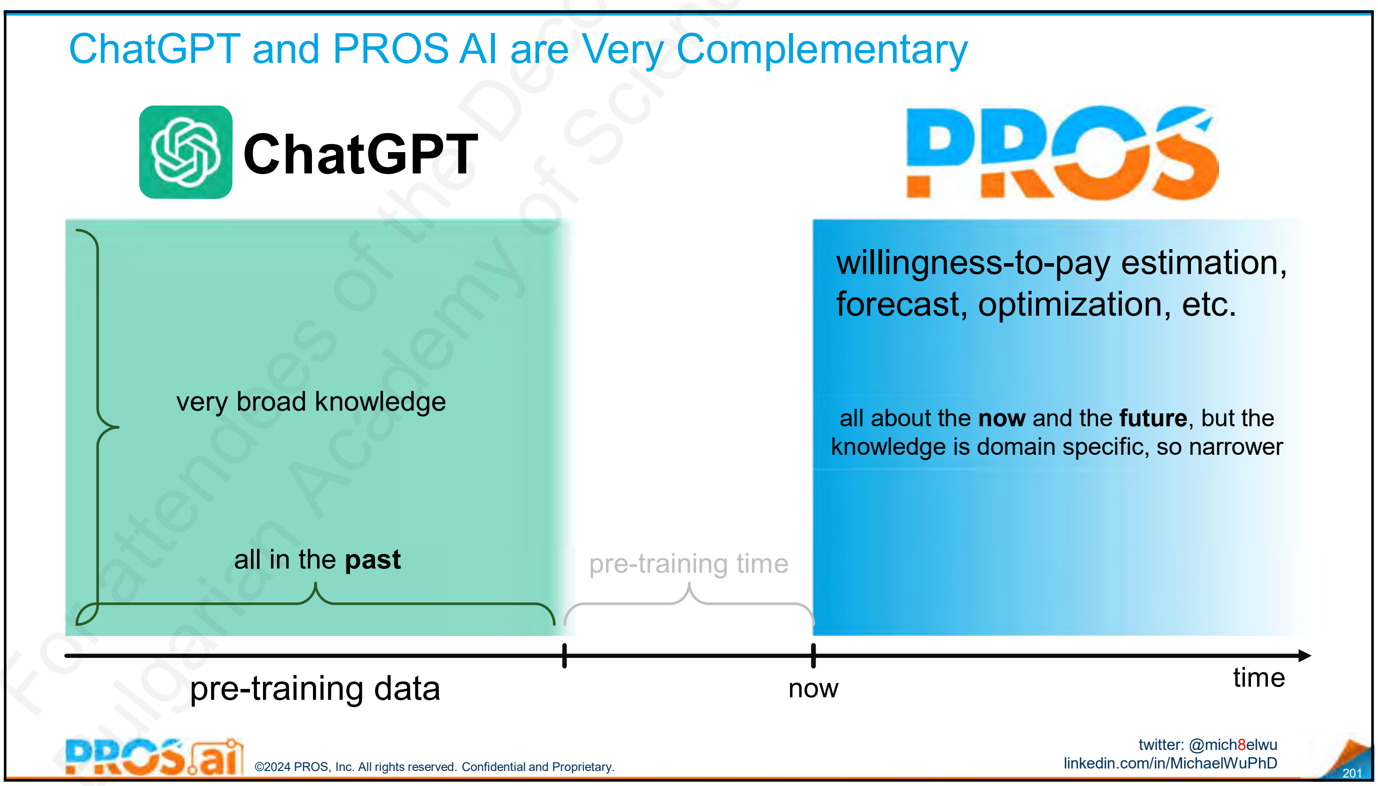
twitter: @mich8elwu
linkedin.com/in/MichaelWuPhD

199

199

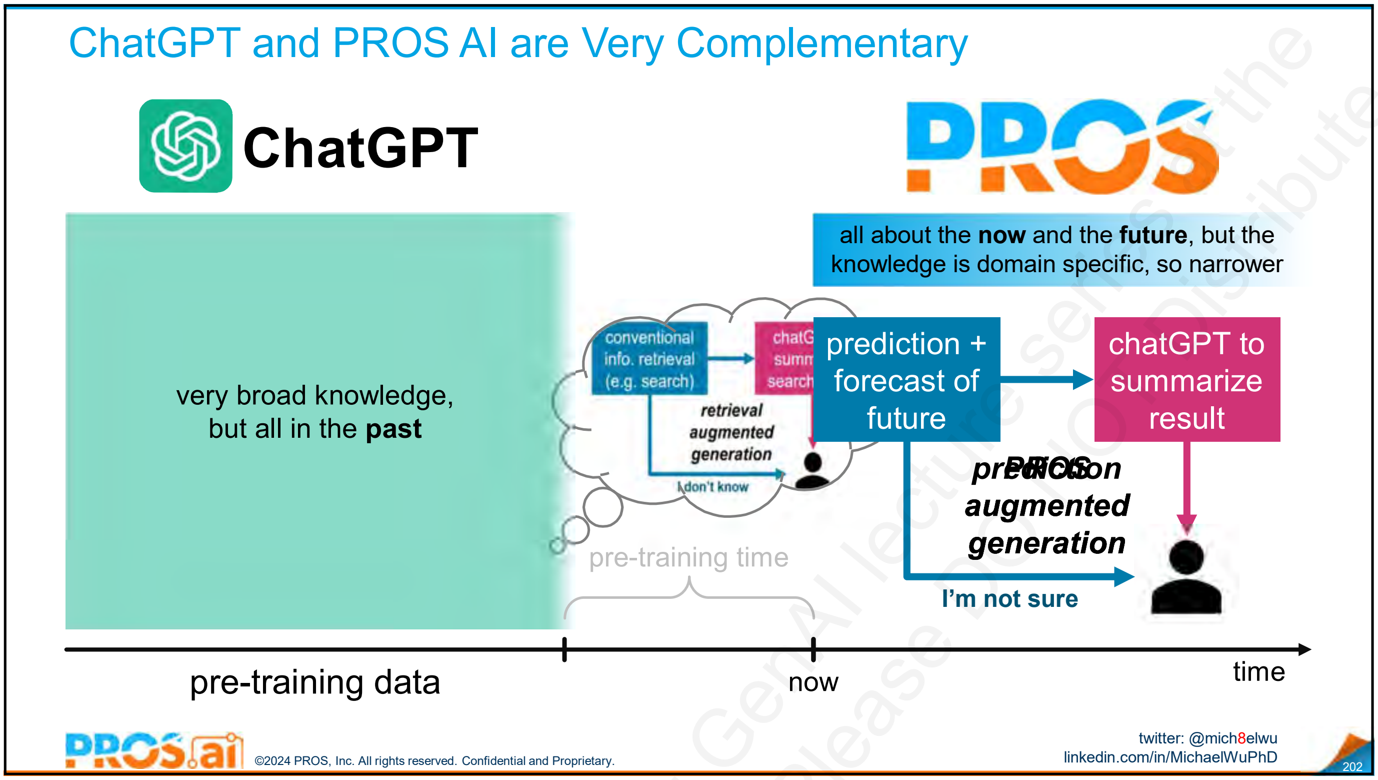


200



201

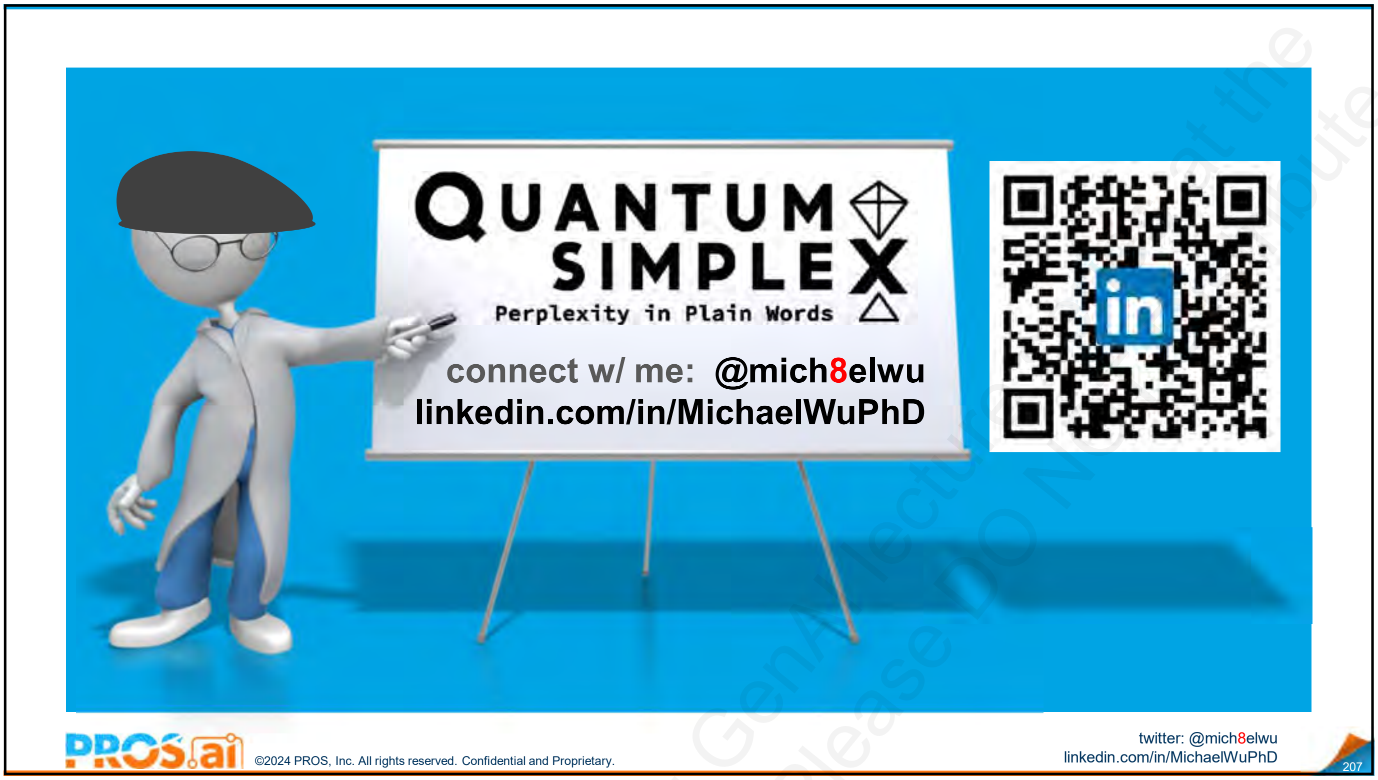
For attendees of the Decoding GenAI lecture series at the Bulgarian Academy of Science. Please DO NOT Distribute.


ChatGPT and PROS AI are Very Complementary
ChatGPT
PROS
all about the **now** and the **future**, but the knowledge is domain specific, so narrower
very broad knowledge, but all in the **past**
conventional info. retrieval (e.g. search)
retrieval augmented generation
I don't know
prediction + forecast of future
chatGPT to summarize result
prediction augmented generation
I'm not sure
pre-training time
pre-training data
now
time
©2024 PROS, Inc. All rights reserved. Confidential and Proprietary.
twitter: @mich8elwu
linkedin.com/in/MichaelWuPhD
202

202


this is the WAY
203

203



207

For attendees of the Decoding GenAI lecture at the Bulgarian Academy of Science. Please DO NOT distribute.