



Sharif University of Technology



Data Science and Machine Learning Lab

Life as a Data Scientist

System Analysis and Design

S.M.F. Sani

Data Science and Machine Learning Lab (DML)
Department of Computer Engineering, Sharif University of Technology
Oct 2024

Introduction

Introduction

About Me



VectorStock® VectorStock.com/48387854



Introduction (Cont.)

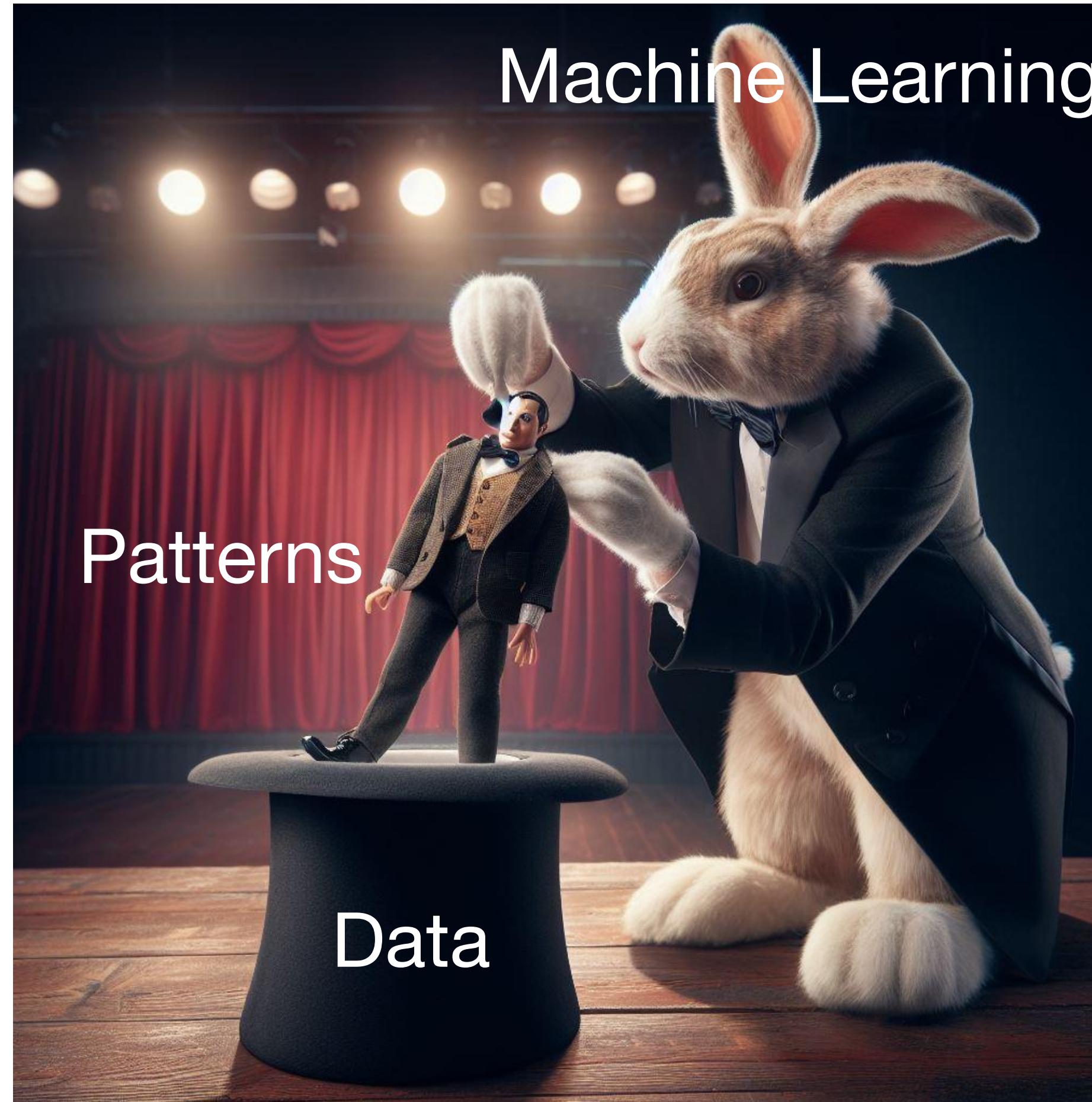
About Me



What is this machine learning thing...

Machine Learning

Magic of our era!



Machine Learning

Magic of our era!



Cyberpunk Style Mercedes Benz in Night City



super realistic portrait of an Iranian model, little smile, rainy day, standing in crossway



A blonde woman with elvish ears waring green dress that has silver buttons puts flower hairpin in her hairs sitting on a wooden chair inside of wooden cottage.

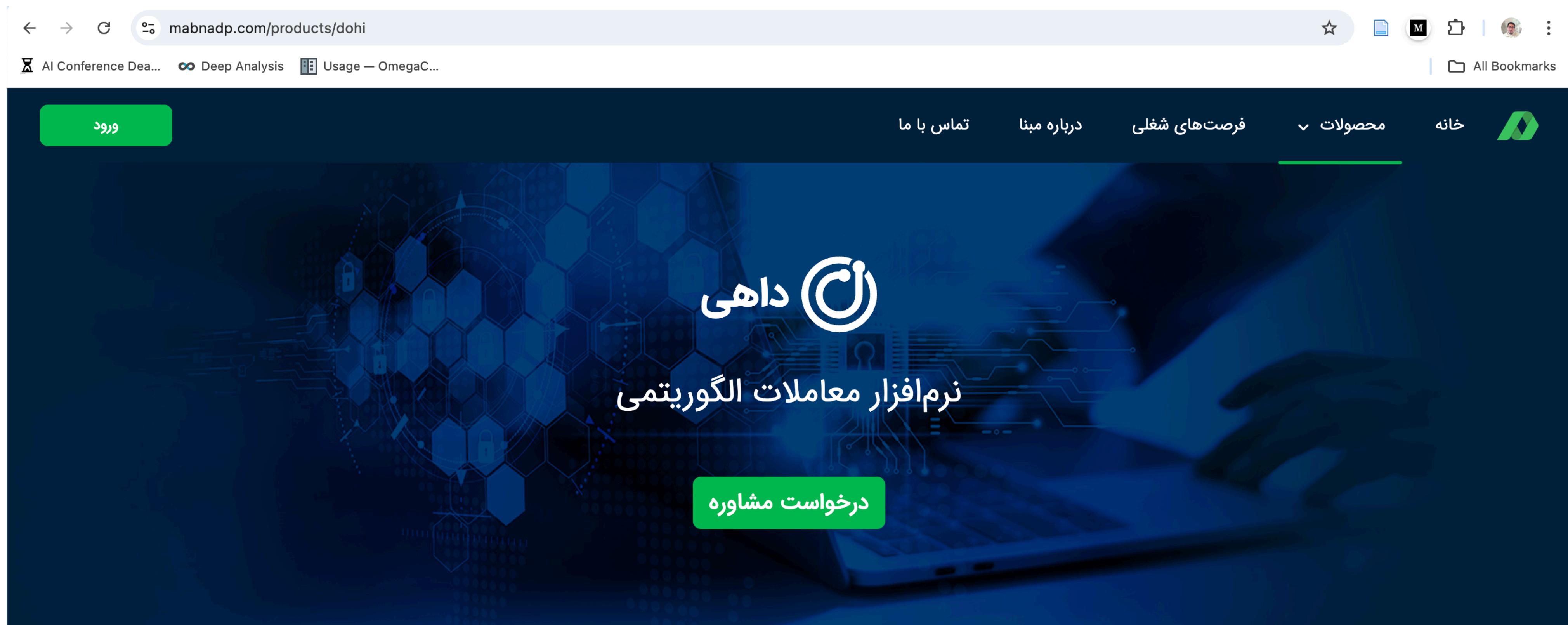
Applications (Iran)

Digital Advertising



Applications (Iran)

Trade



Applications (Iran)

Product Search

The screenshot shows a web browser window with the URL 'mori.style' in the address bar. The page content is in Persian. At the top right, there are navigation icons for back, forward, search, and other browser functions. Below the address bar, there are tabs for 'AI Conference Dea...', 'Deep Analysis', and 'Usage — OmegaC...'. On the far right, there's a 'All Bookmarks' link.

In the center, there's a navigation bar with a dropdown arrow and the text 'ورود به موري' (Log in to mori). To the right of the navigation bar are the 'نمایشگاه پاییزه' (Autumn exhibition) logo and the 'mori' logo.

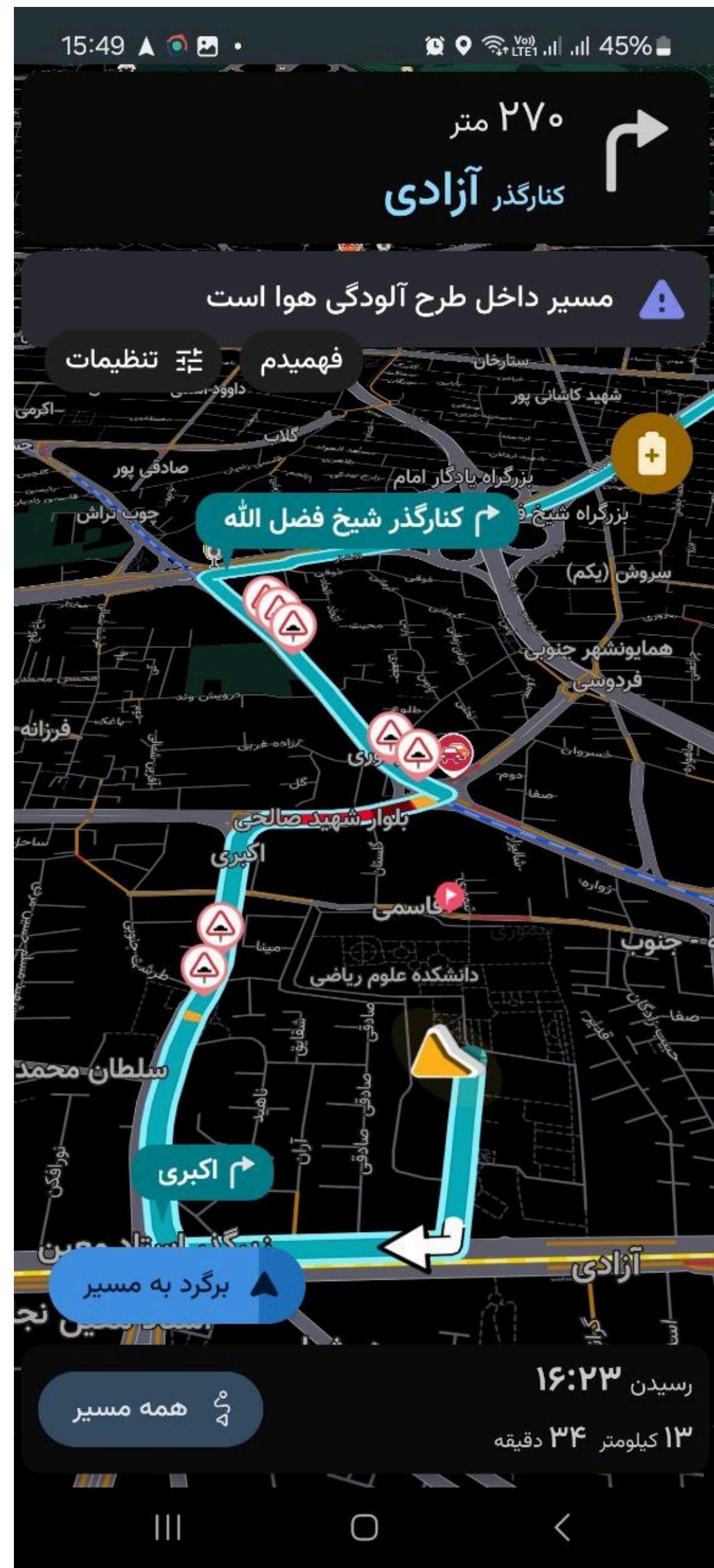
The main content area displays four product images: a person wearing a light-colored hoodie with a floral graphic on the back, a close-up of a hoodie with 'DUANE' printed on it, a person in a green sweatshirt, and a person in a white hoodie. A callout bubble over the second image contains the text 'دورس و یا هودی اورسایز برای پاییز که به هر استایل روزمره‌ای بیاد' (A plus-size hoodie for autumn that goes with every outfit style).

Below the images, a large text block reads 'جستجو بین ۱.۴ میلیون لباس از صدها فروشنده' (Search between 1.4 million clothes from hundreds of sellers).

At the bottom, there's a search bar with the placeholder text 'به دنبال چه جور لباس یا اکسسوری می‌گردی؟ توصیفش کن تا برات پیدا کنیم.' (What kind of clothes or accessories are you looking for? Describe them to find them). The search bar includes a magnifying glass icon and a character count indicator '۰/۱۰۰۰'. There are also back and forward navigation buttons.

Applications (Iran)

Navigation



Applications (Iran)

Speech Processing

The screenshot displays a grid of 16 cards, each representing a different application or service related to speech processing:

- بیوشا**: تلفن گویای هوشمند (Smartphone Voice Assistant) - پاسخگوی هوشمند تلفنی.
- آریانا**: متن به گفتار آریانا (Text-to-Speech) - متن خوان هوشمند فارسی.
- نیسا**: گفتار به نوشته رسمی (Speech-to-Text) - نایپ هوشمند.
- نیسالایف**: گفتار به نویسنده (Speech-to-Text) - نایپ هوشمند.
- روزآقا**: جستجو در گفتار (Search in Speech) - جستجوگر کلمات کلیدی در صوت.
- احراز هویت گفتاری**: احراز هویت صوتی (Voice Biometric Verification) - رمز صوتی.
- تشخیص هویت صوتی**: شناسایی گوینده از روی صدا.
- ربات هوشمند صوتی**: چت‌بات هوشمند فارسی.
- پارسیا**: مترجم هوشمند گفتار به گفتار.
- دادگان هوش مصنوعی**: تهیه و تولید دادگان هوش مصنوعی.
- دستیارهای صوتی هوشمند**: مبتنی بر هوش مصنوعی.
- نویسه‌خوان هوشمند فارسی**: استخراج متن از تصویر.
- پروژه‌ها**: پروژه‌های سفارشی هوش مصنوعی.
- شیوا**: سامانه بیبود گفتار.
- تبديل صدا**: Voice Conversion.

On the right side of the grid, there is a large, stylized illustration of a smartphone displaying various app icons, with a magnifying glass focusing on one of them. Below the phone, three people are shown interacting with it, and a shopping cart is nearby. The background of the page features a light blue gradient.

Applications (Iran)

Chat Bot

گپیفای - Gapify

ابزار چت آنلاین با مشتریان + هوش مصنوعی برای پاسخ‌دهی سریع

گپیفای ترکیبی از امکانات جامع چت آنلاین با مشتریان به اضافهی چتبات هوش مصنوعی (باتیفای) برای خودکارسازی حداکثری پاسخ‌دهی به مشتریان است. با گپیفای هیچ مشتری بی‌پاسخ نمی‌ماند.

نیاز به اطلاعات بیشتری دارم

گپیفای را امتحان می‌کنم

با ما چت کنید!



Academia vs Industry

Academia Vs Industry

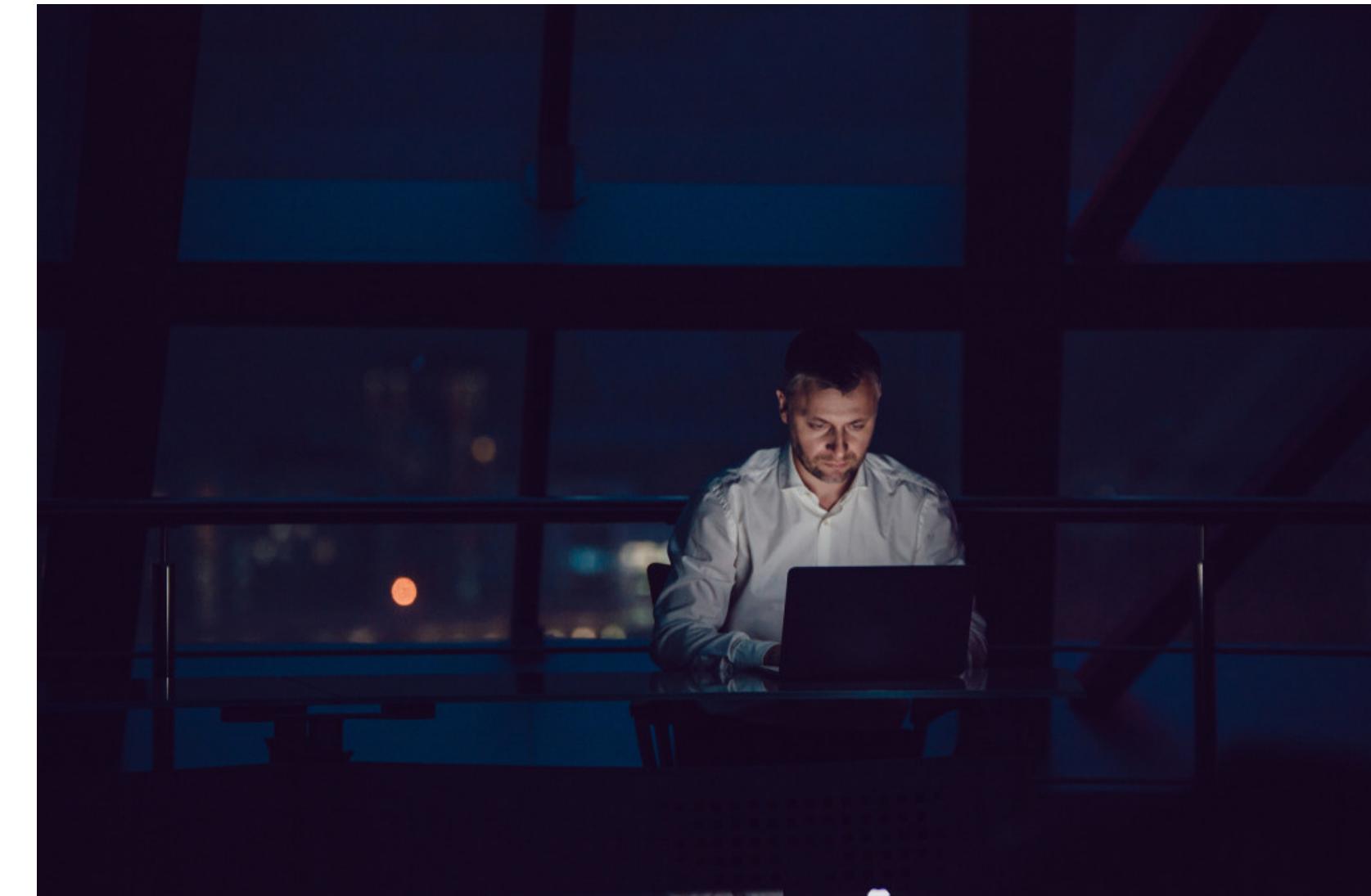
Money



Infrastructure

Academia Vs Industry

Team Work, Communication (Maybe our Problem)



Academia Vs Industry

Priorities, Deadlines

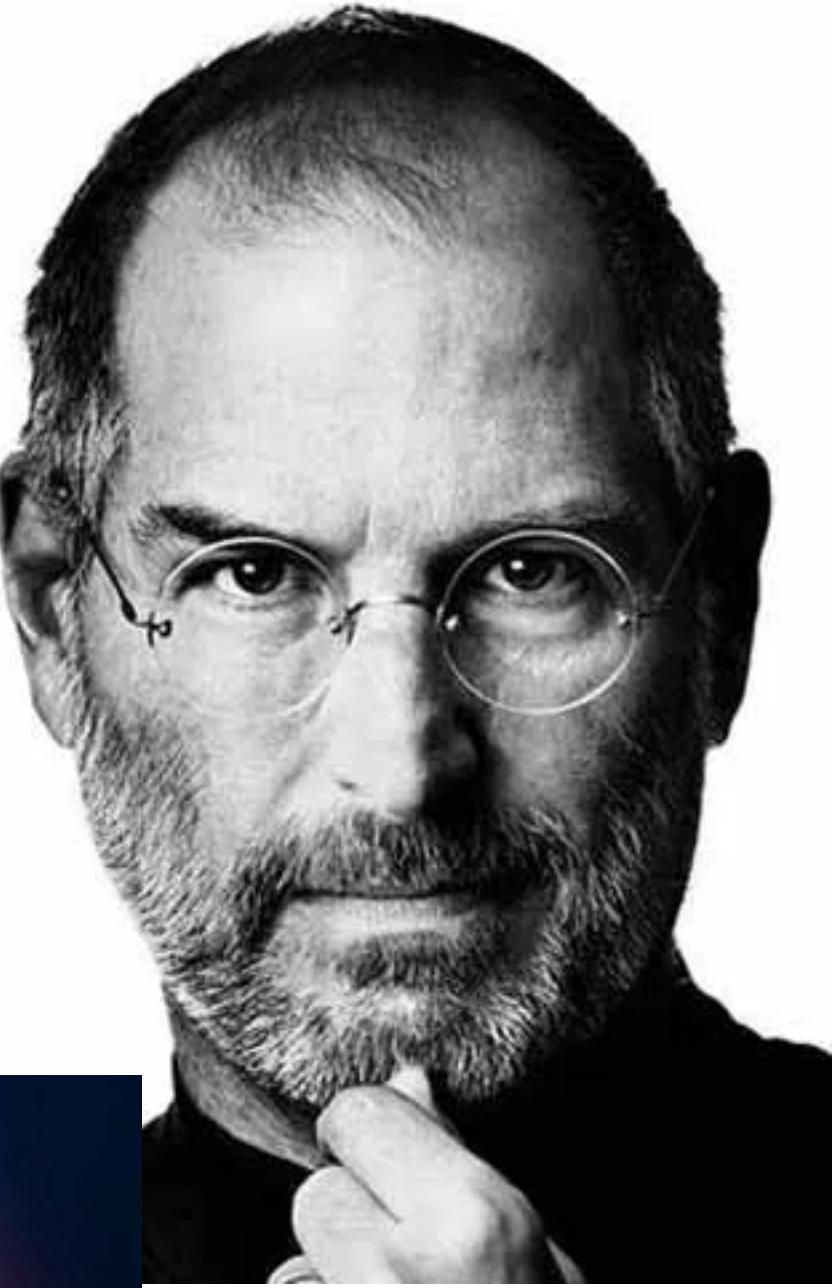


"The only way to do great work is to love what you do. If you haven't found it, keep looking. Don't settle."

- Steve Jobs



SOMEONE IN THEIR 40S WHO'S STILL IN GRAD SCHOOL



Academia Vs Industry

Importance of Data



Facebook and Google Are the New Data Brokers

By Chris Hoofnagle | University of California, Berkeley

They claim not to sell data, but Facebook and Google have paid developers with data. This selling of your personal data is central to platform economics. Could quality signals change the terms of the bargain?

"We do not sell your personal information to anyone"
[<https://safety.google/privacy/ads-and-data/>] – Google, Ads and Data Webpage

"No, we don't sell any of your information to anyone and we never will."
[<https://www.facebook.com/help/152637448140583?helpref=related>] – Facebook,
Does Facebook sell my information? Webpage



157	
158	
159	سلام خیلی خوشحالم ک این وقت شب برای ما شب بیدارها فعال هستید، لطفا در صورت امکان شیرموز پسته شکر نداشته باشد بهتر است. دکتر کاظمی
160	
161	
162	
163	
164	

بهروزسانی ۲: توضیح کافه‌بازار در مورد آسیب‌پذیری امنیتی گزارش شده

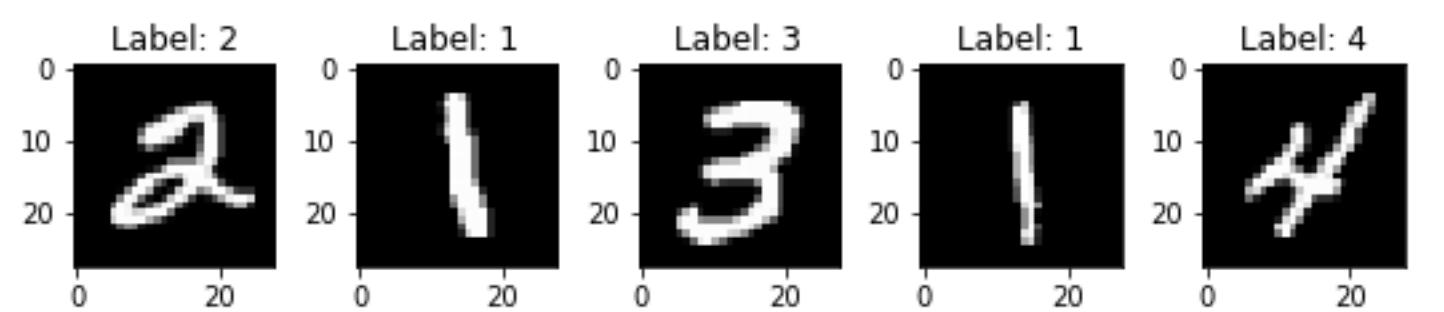
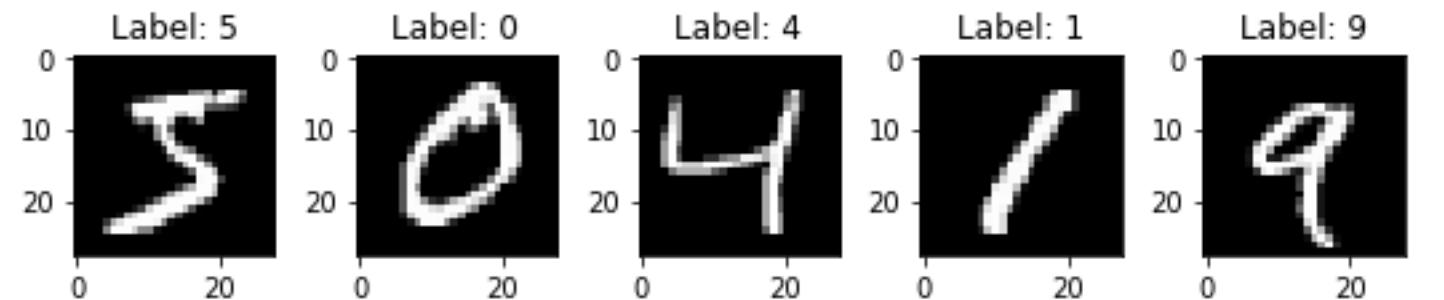
کافه‌بازار پیرو بیانیه‌ی قبلی و در راستای سیاست خود مبنی بر شفافیت، کاربران را در جریان جزئیات جدید نتایج بررسی‌های تیم فنی کافه‌بازار درباره‌ی گزارشی که صبح روز شنبه ۷ اردیبهشت در شبکه‌های اجتماعی منتشر یافت، قرار می‌دهد.

طبق بررسی‌های تیم فنی کافه‌بازار مشخص شده است که منبع‌کد یکی از زیرسیستم‌های وبسایت کافه‌بازار به دست افرادی خارج از مجموعه رسیده است؛ در عین حال، همچنان هیچ‌گونه شواهدی در زمینه نشت اطلاعات حساب‌های کاربری به دست نیامده است.

تیم فنی این مسئله را از ساعات ابتدایی روز شنبه با جدیت پیگیری کرد و اشکال امنیتی را شناسایی و برطرف کرد.

Academia Vs Industry

Data Type, Volume



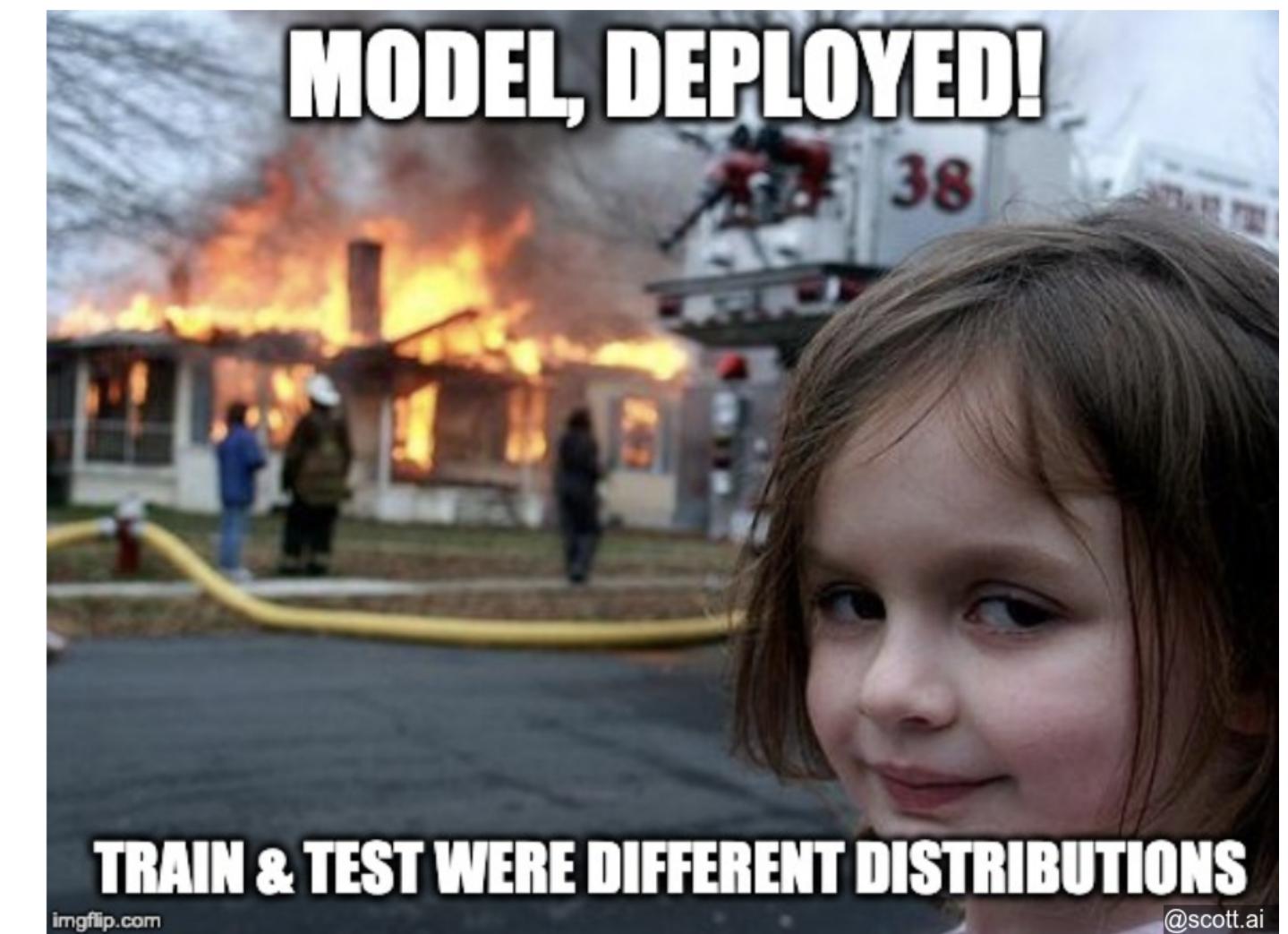
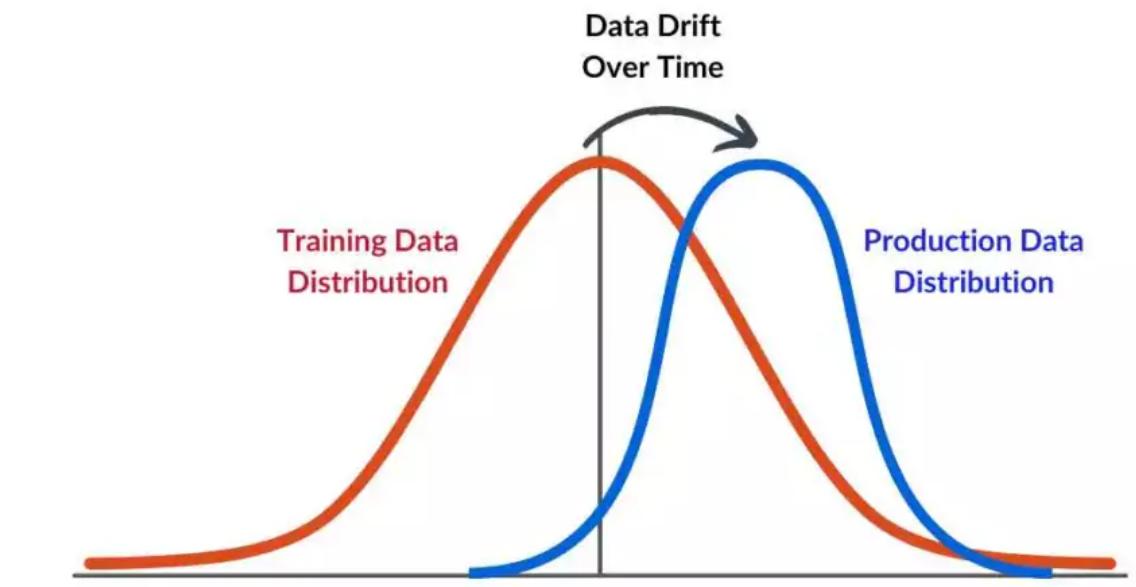
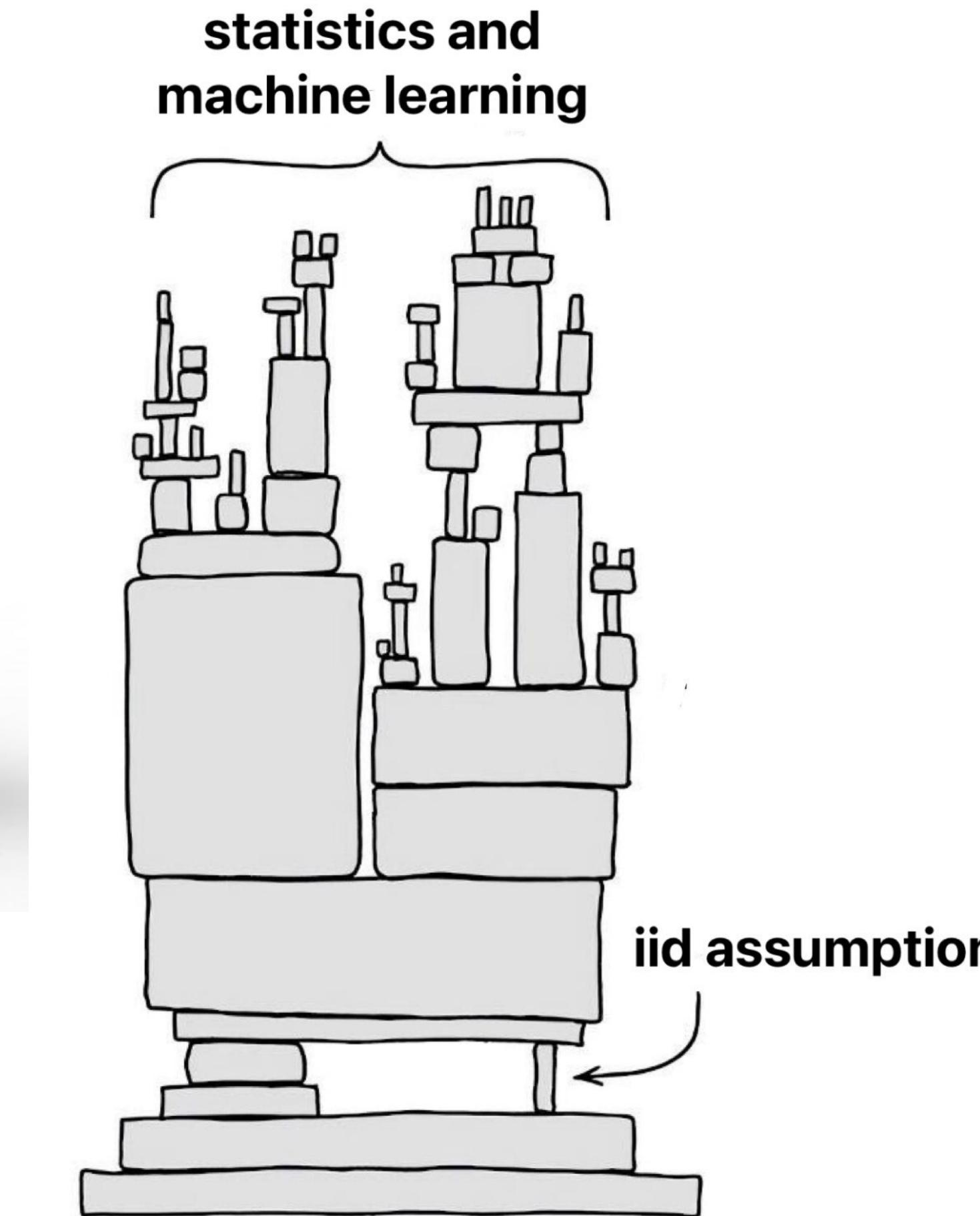
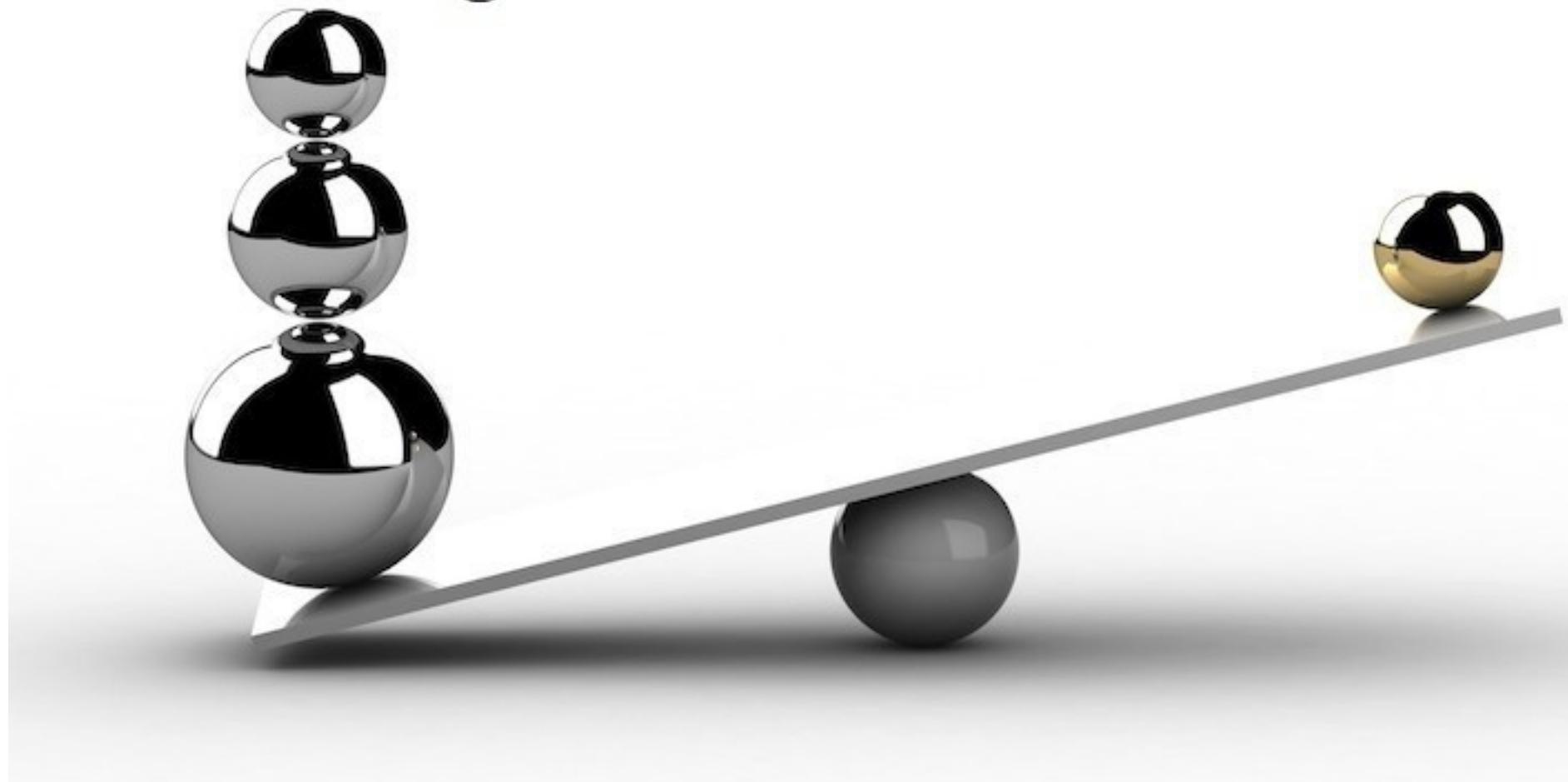
```
.....  
{  
    "sort_timestamp": 1634275259292,  
    "rating": 3.0,  
    "helpful_votes": 0,  
    "title": "Meh",  
    "text": "These were lightweight and soft but much too small for my liking. I would have  
preferred two of these together to make one loc. For that reason I will not be  
repurchasing.",  
    "images": [  
        {  
            "small_image_url": "https://m.media-amazon.com/images/I/81FN4c0VHzL._SL256_.jpg",  
            "medium_image_url": "https://m.media-amazon.com/images/I/81FN4c0VHzL._SL800_.jpg",  
            "large_image_url": "https://m.media-amazon.com/images/I/81FN4c0VHzL._SL1600_.jpg",  
            "attachment_type": "IMAGE"  
        },  
        {  
            "asin": "B088SZDGXG",  
            "verified_purchase": true,  
            "parent_asin": "B08BBQ29N5",  
            "user_id": "AEYORY2AVPMCPDV57CE337YU5LXA"  
        }  
    ]  
}.....
```



Academia Vs Industry

Data Quality

Handling Imbalanced Dataset in ML

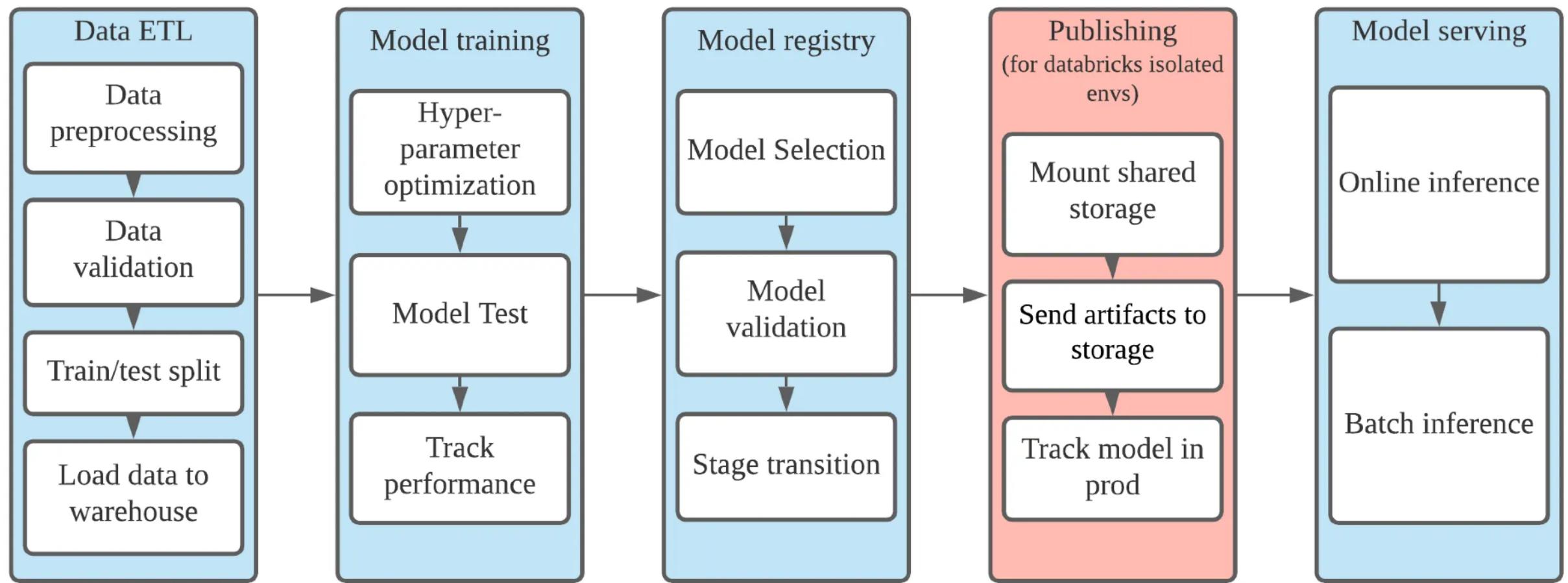


Academia Vs Industry

Infrastructure, Continuous Training

Some key facts about GPT-4:

- **Total parameters** — ~1.8 trillion (over 10x more than GPT-3)
- **Architecture** — Uses a mixture of experts (MoE) model to improve scalability
- **Training compute** — Trained on ~25,000 Nvidia A100 GPUs over 90-100 days
- **Training data** — Trained on a dataset of ~13 trillion tokens
- **Inference compute** — Runs on clusters of 128 A100 GPUs for efficient deployment
- **Context length** — Supports up to 32,000 tokens of context



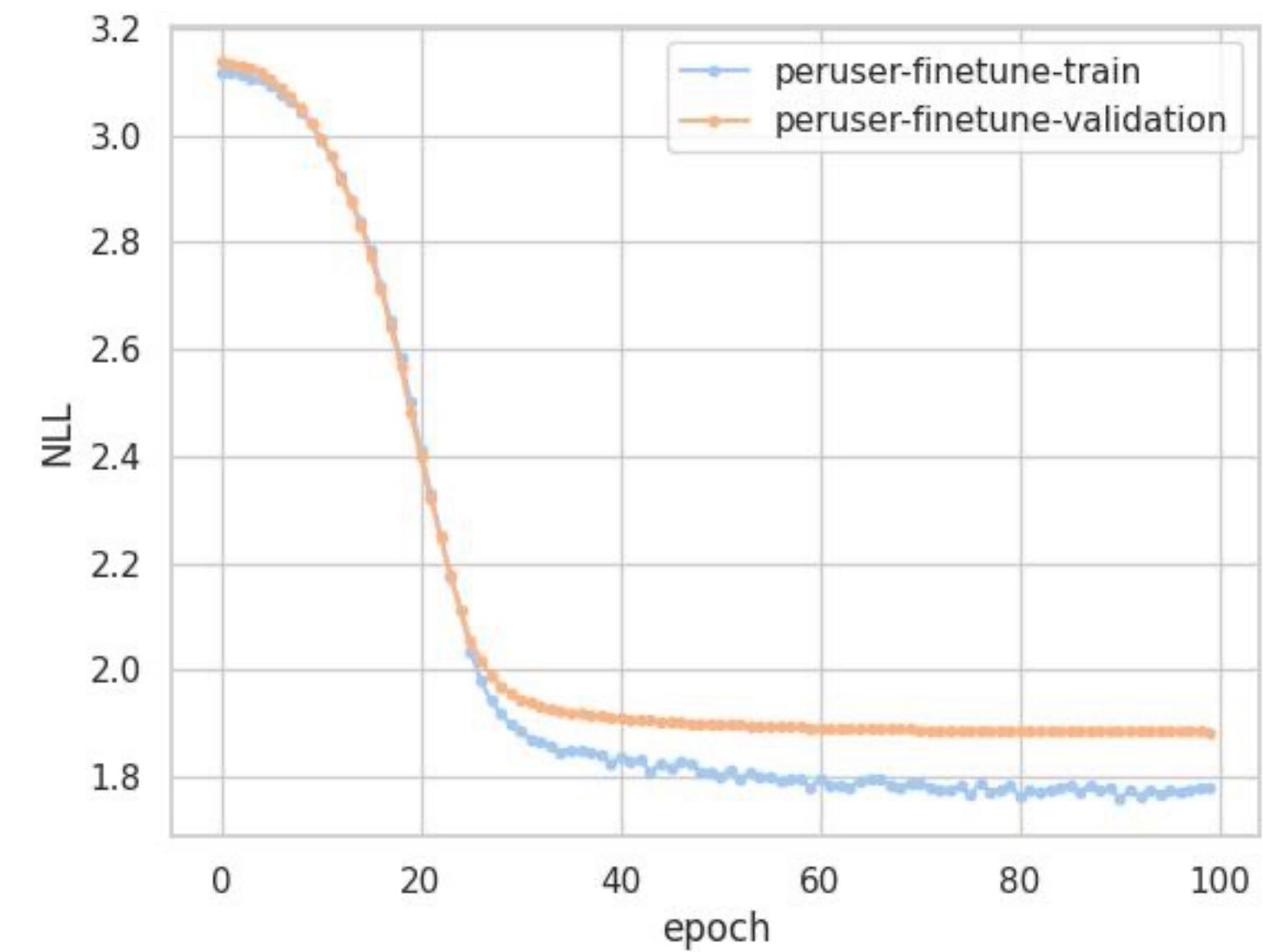
Running Large Language Models (Llama 3) on Apple Silicon with Apple's MLX Framework

Step-by-Step Guide to Implement LLMs like Llama 3 Using Apple's MLX Framework on Apple Silicon (M1, M2, M3, M4)

Manuel · Follow
5 min read · Jun 10, 2024

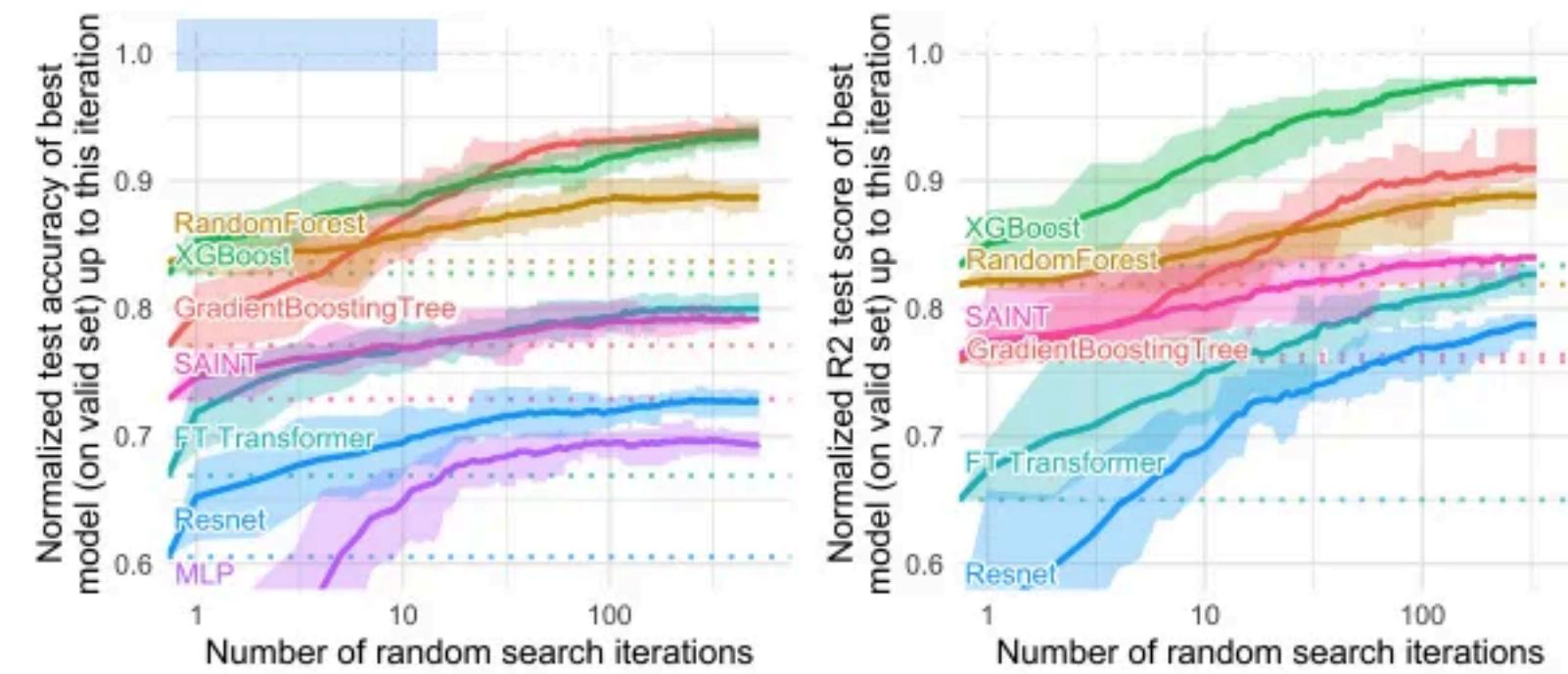
136 Q 1

...

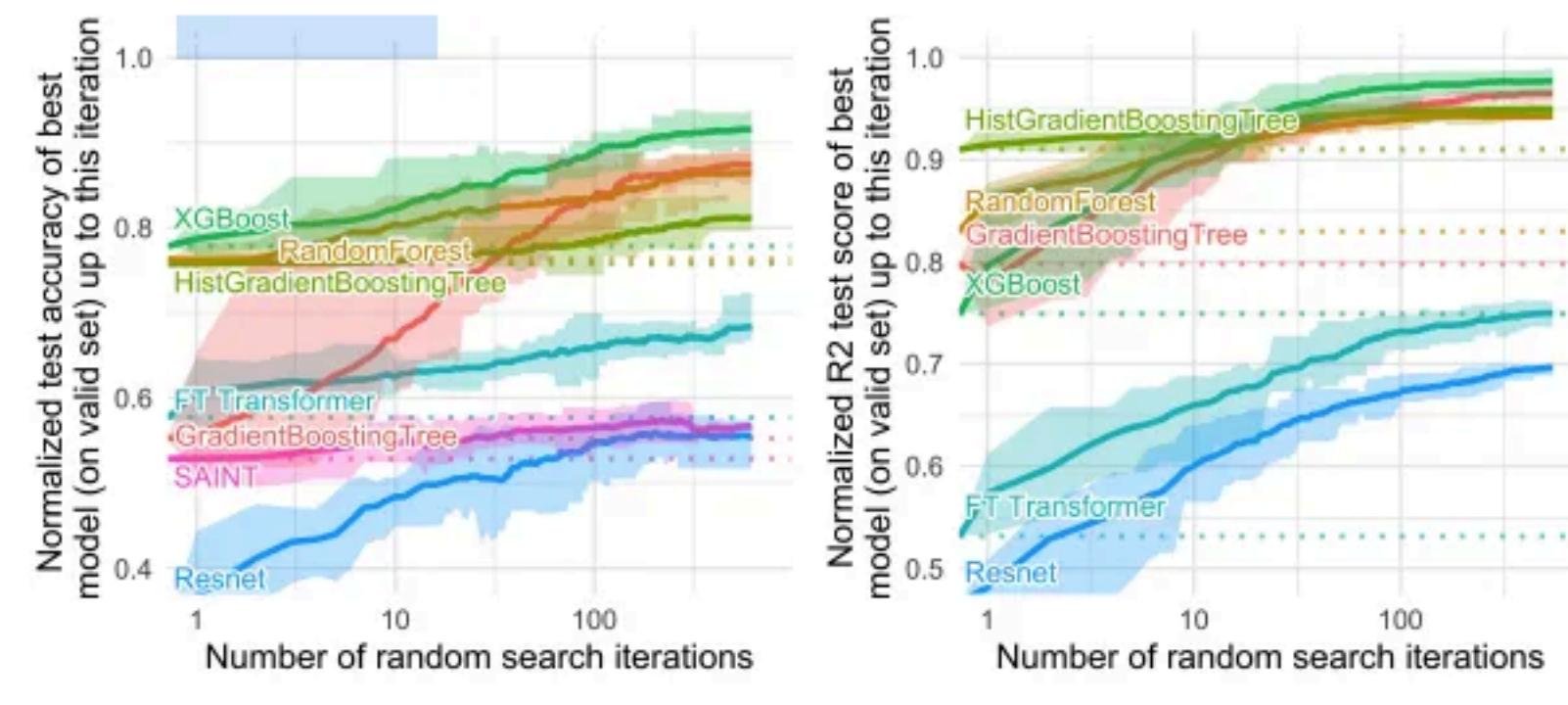


Academia Vs Industry

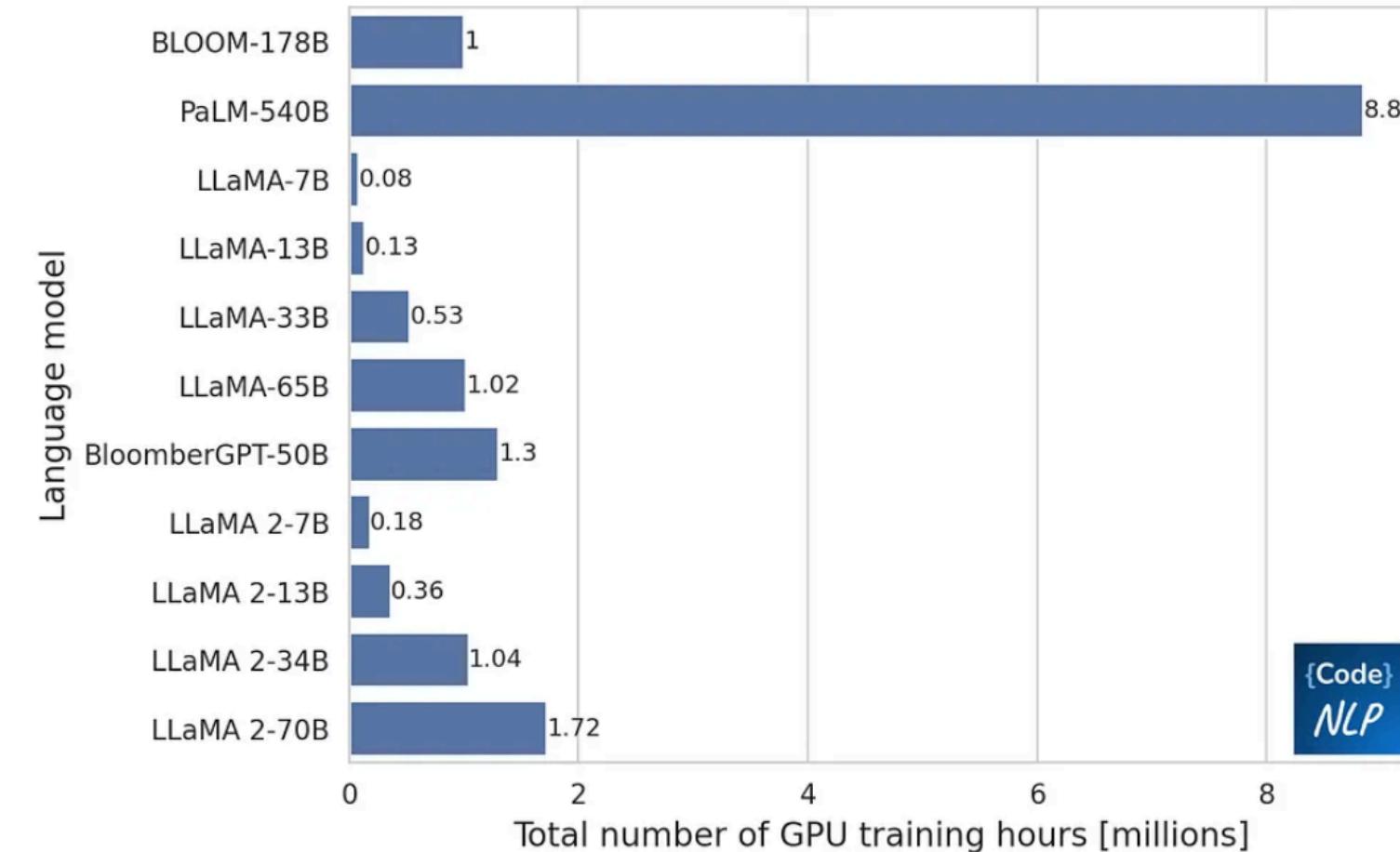
Solution Complexity



Medium-sized datasets, with only numerical features



Medium-sized datasets, with both numerical and categorical features



Model Name	Rossmann	CoverType	Higgs	Gas	Eye	Gesture
XGBoost	490.18 ± 1.19	3.13 ± 0.09	21.62 ± 0.33	2.18 ± 0.20	56.07 ± 0.65	80.64 ± 0.80
NODE	488.59 ± 1.24	4.15 ± 0.13	21.19 ± 0.69	2.17 ± 0.18	68.35 ± 0.66	92.12 ± 0.82
DNF-Net	503.83 ± 1.41	3.96 ± 0.11	23.68 ± 0.83	1.44 ± 0.09	68.38 ± 0.65	86.98 ± 0.74
TabNet	485.12 ± 1.93	3.01 ± 0.08	21.14 ± 0.20	1.92 ± 0.14	67.13 ± 0.69	96.42 ± 0.87
ID-CNN	493.81 ± 2.23	3.51 ± 0.13	22.33 ± 0.73	1.79 ± 0.19	67.9 ± 0.64	97.89 ± 0.82
Simple Ensemble	488.57 ± 2.14	3.19 ± 0.18	22.46 ± 0.38	2.36 ± 0.13	58.72 ± 0.67	89.45 ± 0.89
Deep Ensemble w/o XGBoost	489.94 ± 2.09	3.52 ± 0.10	22.41 ± 0.54	1.98 ± 0.13	69.28 ± 0.62	93.50 ± 0.75
Deep Ensemble w XGBoost	485.33 ± 1.29	2.99 ± 0.08	22.34 ± 0.81	1.69 ± 0.10	59.43 ± 0.60	78.93 ± 0.73

Each column is a dataset. Better performances are highlighted with bold numbers.

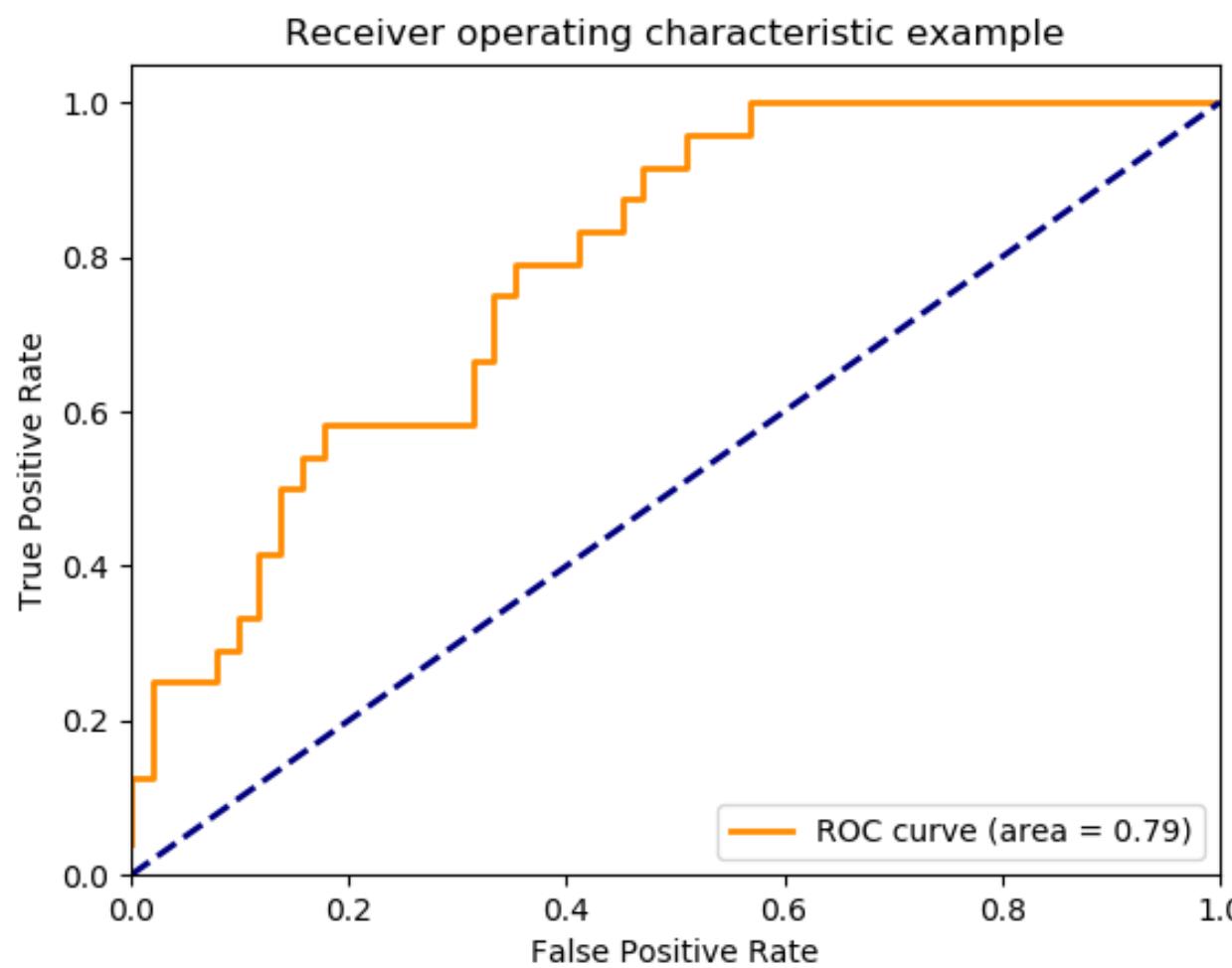
Model Name	YearPrediction	MSLR	Epsilon	Shrutime	Blastchar
XGBoost	77.98 ± 0.11	$55.43 \pm 2e-2$	$11.12 \pm 3e-2$	13.82 ± 0.19	20.39 ± 0.21
NODE	76.39 ± 0.13	$55.72 \pm 3e-2$	$10.39 \pm 1e-2$	14.61 ± 0.10	21.40 ± 0.25
DNF-Net	81.21 ± 0.18	$56.83 \pm 3e-2$	$12.23 \pm 4e-2$	16.8 ± 0.09	27.91 ± 0.17
TabNet	83.19 ± 0.19	$56.04 \pm 1e-2$	$11.92 \pm 3e-2$	14.94 ± 0.13	23.72 ± 0.19
ID-CNN	78.94 ± 0.14	$55.97 \pm 4e-2$	$11.08 \pm 6e-2$	15.31 ± 0.16	24.68 ± 0.22
Simple Ensemble	78.01 ± 0.17	$55.46 \pm 4e-2$	$11.07 \pm 4e-2$	13.61 ± 0.14	21.18 ± 0.17
Deep Ensemble w/o XGBoost	78.99 ± 0.11	$55.59 \pm 3e-2$	$10.95 \pm 1e-2$	14.69 ± 0.11	24.25 ± 0.22
Deep Ensemble w XGBoost	76.19 ± 0.21	$55.38 \pm 1e-2$	$11.18 \pm 1e-2$	13.10 ± 0.15	20.18 ± 0.16

Each column is a dataset. Better performances are highlighted with bold numbers.



Academia Vs Industry

Evaluation Metrics



Doubly Robust

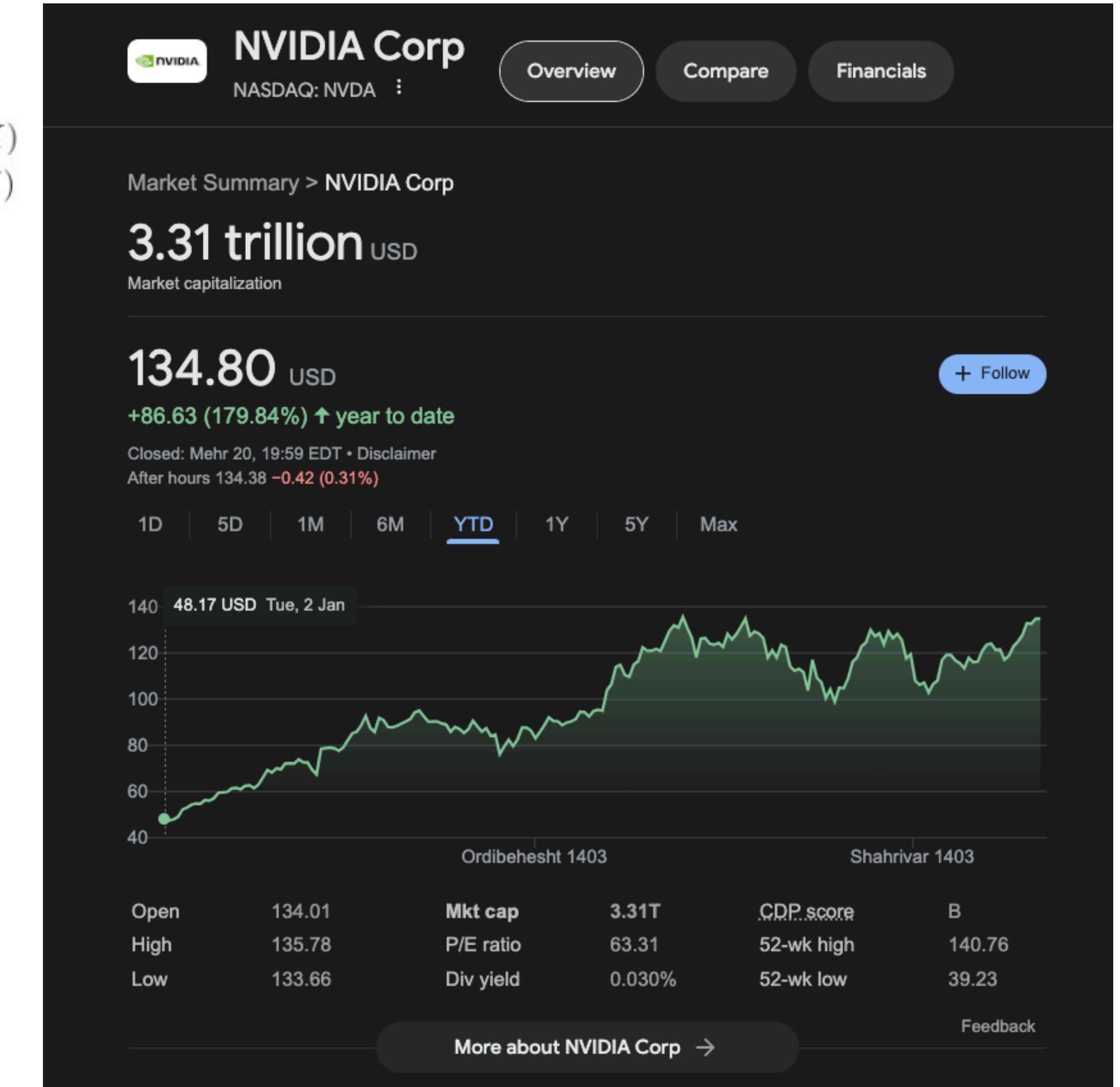
- Estimating ATE with Doubly Robust estimator:

$$\begin{aligned}ATE_{DR} &= \frac{1}{n} \sum_{i=1}^n \left[\frac{T_i Y_i}{\hat{e}(X_i)} - \frac{\{T_i - \hat{e}(X_i)\}}{\hat{e}(X_i)} \hat{m}_1(X_i) \right] \\&- \frac{1}{n} \sum_{i=1}^n \left[\frac{(1 - T_i) Y_i}{1 - \hat{e}(X_i)} + \frac{\{T_i - \hat{e}(X_i)\}}{1 - \hat{e}(X_i)} \hat{m}_0(X_i) \right]\end{aligned}$$

- *Unbiased* if either **propensity score** or **regression model** is correct
- This property is referred to as *double robustness*

		Predicted values		$Recall = Sensitivity = \frac{TP}{TP+FN}$
		True	False	
Actual	True	True Positive (TP)	False Negative (FN)	$Specificity = \frac{TN}{TN+FP}$
	False	False Positive (FP)	True Negative (TN)	
		Precision = $\frac{TP}{TP+FP}$		Accuracy = $\frac{TP+TN}{TP+TN+FP+FN}$
				$F1 = \frac{2 \times Precision \times Recall}{Precision + Recall}$

B. Classification Model Results



Academia Vs Industry

Values, Blue or Red pill



≡ Google Scholar Attention is all you need! [Search]

Articles About 8,700,000 results (0.06 sec)

Any time [PDF] **Attention is all you need** [PDF] hhu.de

Since 2024 A Vaswani - Advances in Neural Information Processing Systems, 2017 - user.phil.hhu.de

Since 2023 Attention is all you need Attention is all you need ...

Since 2020 ☆ Save Cite Cited by 136701 Related articles

Custom range... ...

An image of a smartphone screen showing a news article. The article features a dark background with a red and blue abstract graphic. It includes a small circular icon with a heart rate graph and some Persian text at the top. The main title is "انقلابی در تشخیص سرطان پروستات با هوش مصنوعی ایرانی" (Revolutionary in cancer diagnosis prostate cancer with artificial intelligence). Below the title are smaller Persian text elements like "دھر میں" and "Mr.News".

The Choice is Yours!

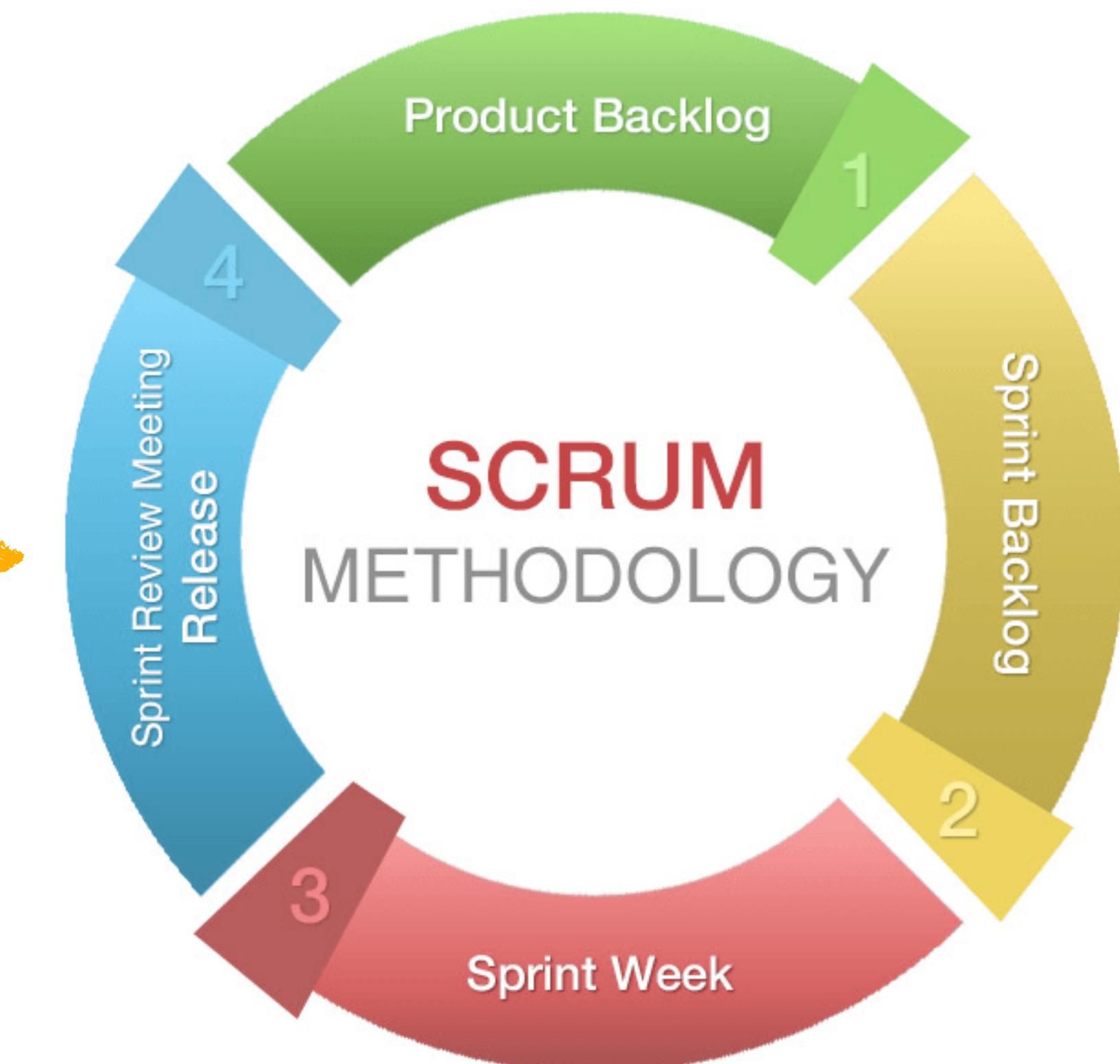
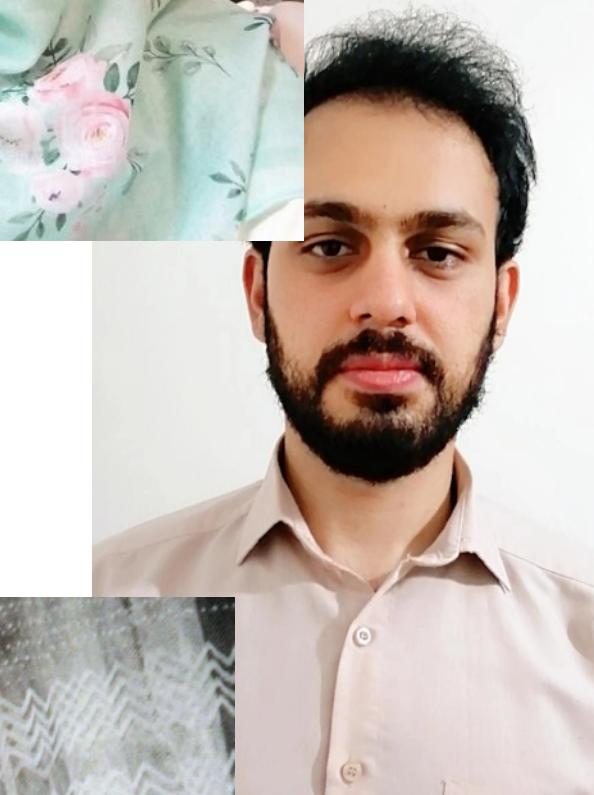
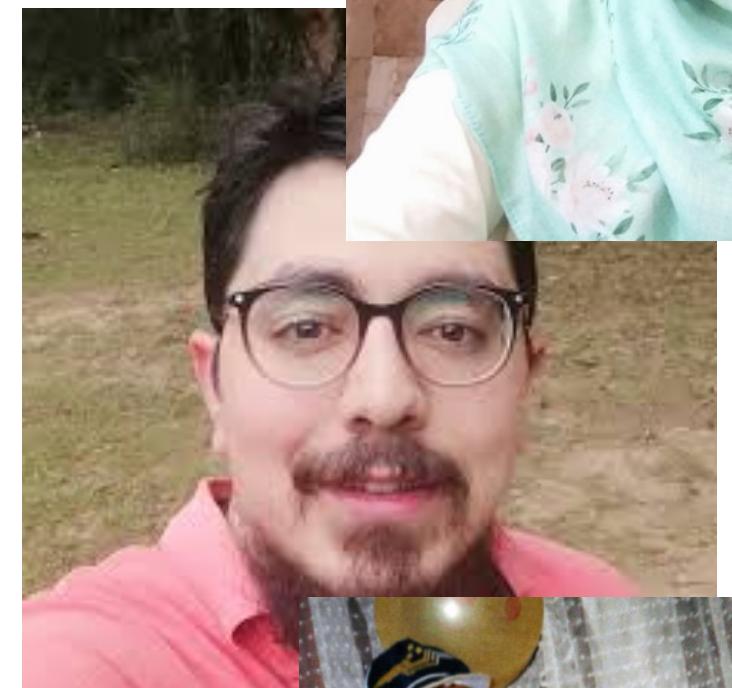


digikala

فیلم‌مو 

Academia Vs Industry

Project Management



Scrum and Machine Learning

Positions

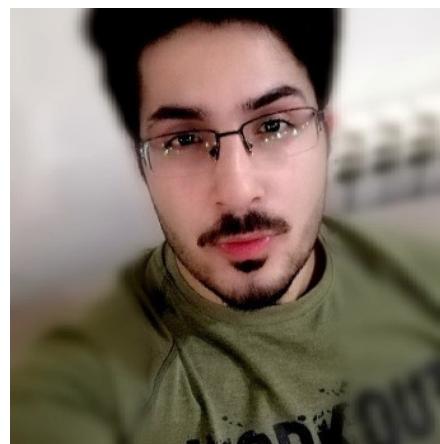
Machine Learning Roles



Machine Learning



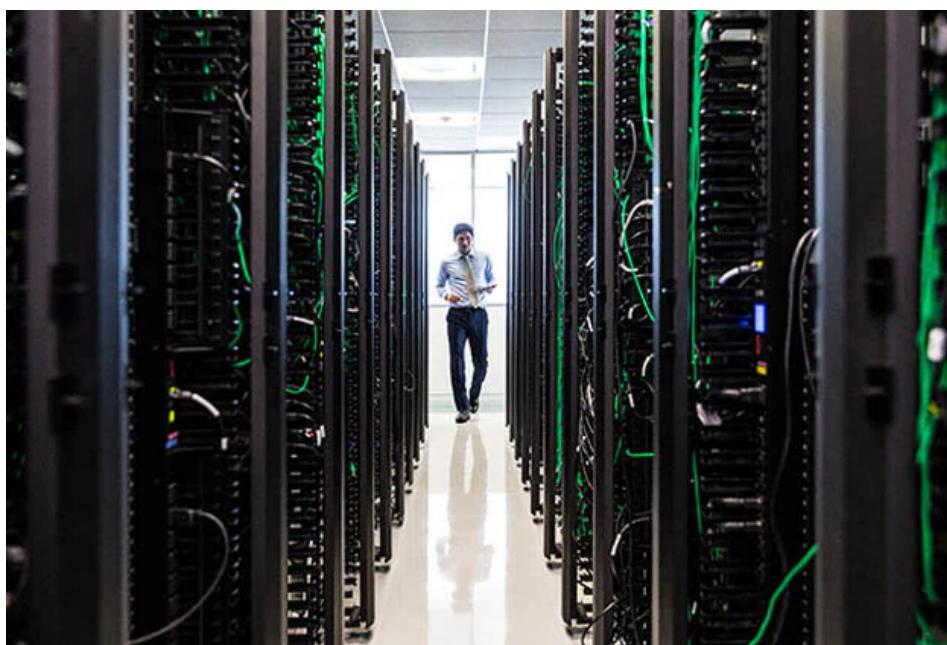
Data Scientist



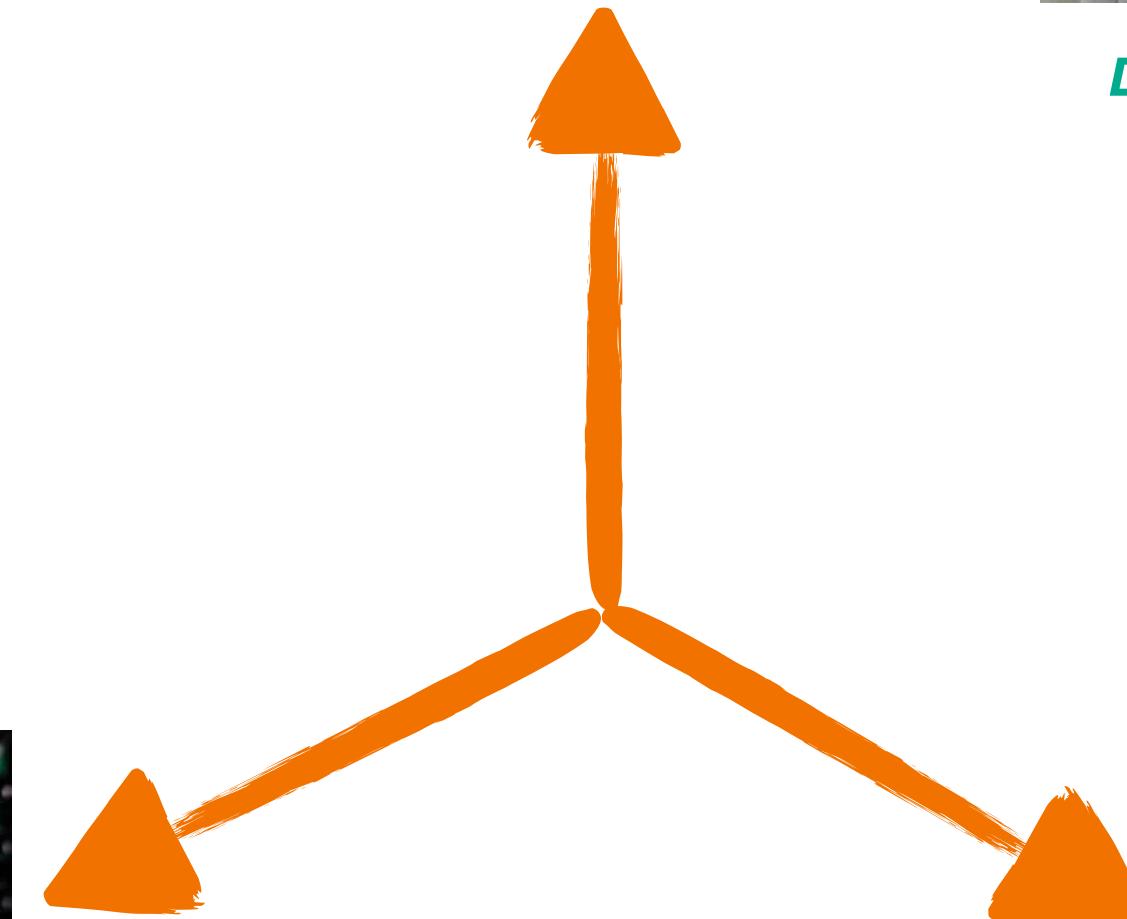
ML Engineer



Data Engineer



Infrastructure



Business



Data Analyst

Scrum and ML

Product Manager is Super Important

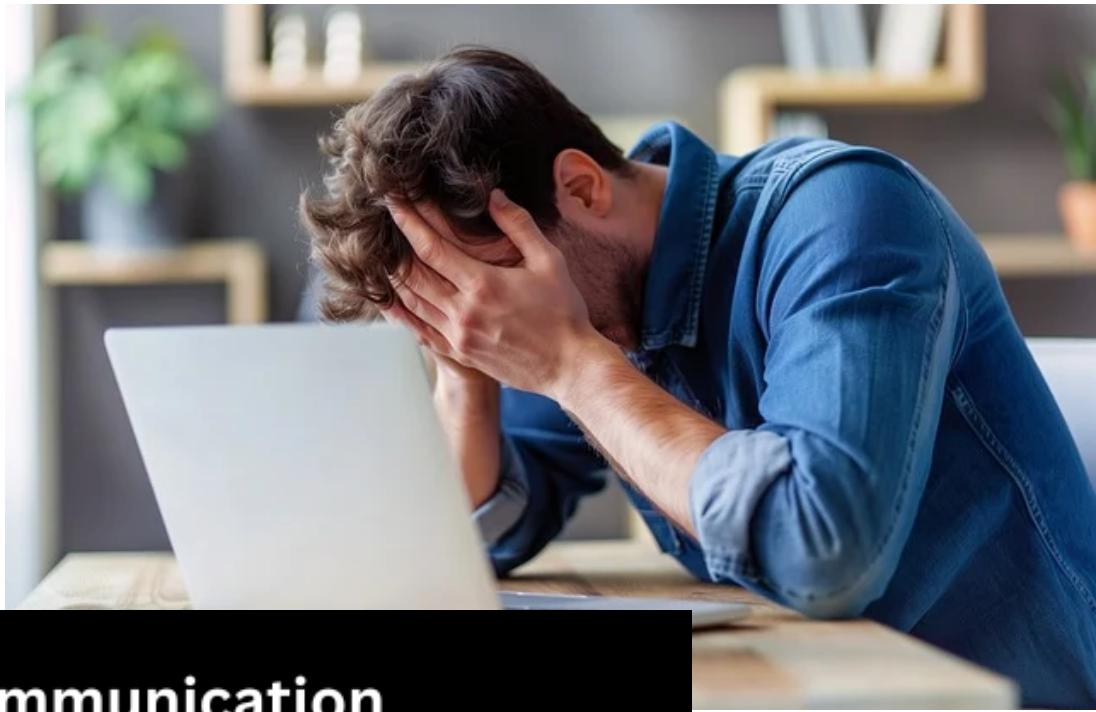
- **View ML system as product is HARD.**
- Maintaining Backlog is a Job.
- Finding Business Opportunities.
- Academic People do not pay attention to business.
- Intra-team Communications.
- Ego in academic people.
- ***Technical Product Manager (TPM).***



Scrum and ML

Standup Meetings

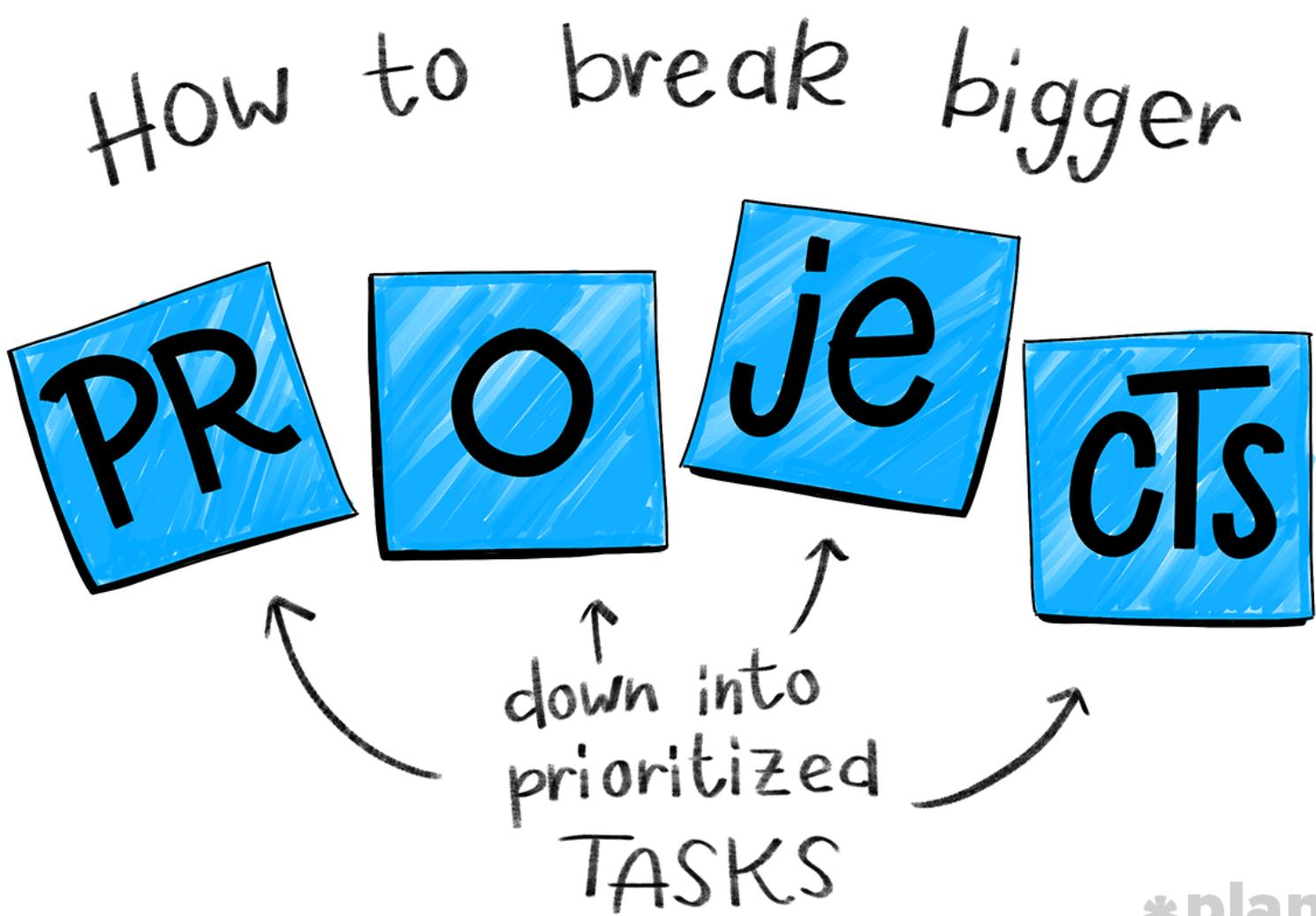
- **Sense of Progress.**
- **Fail-fast Mindset.**
- **Pickup fights.**
- **Speed Control.**
- **Inter-team speed difference.**
- **Irrelevant tasks Issue.**
- **Means Start of The Day.**
- **Easily Could Break the Time Box.**



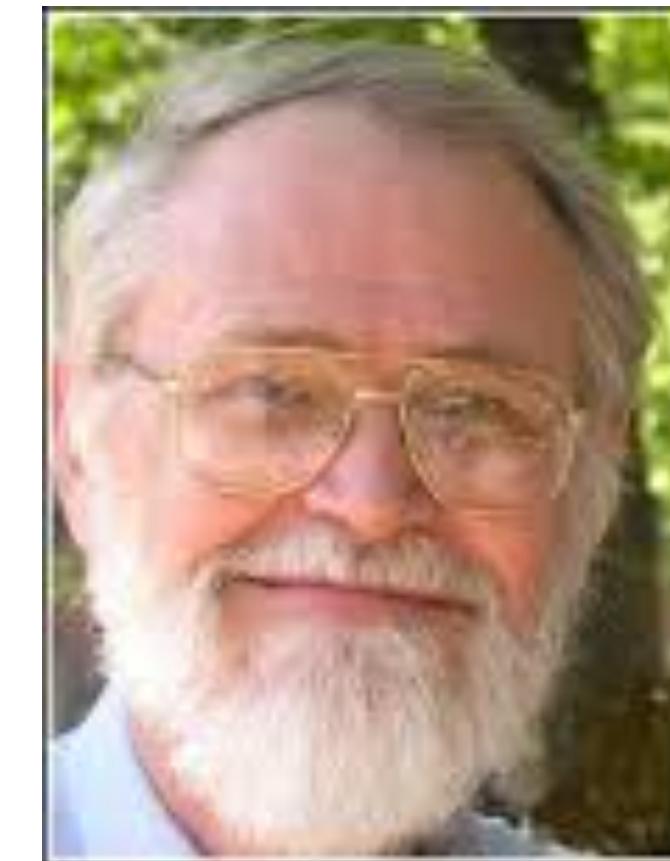
Scrum and ML

Refinement / Vague Backlog

- Define DoDs are challenging in ML Systems.
- Time Boxing Issues. (EDA, Increasing performance x%, debugging, ...)
- Data Driven Scrum Insights.



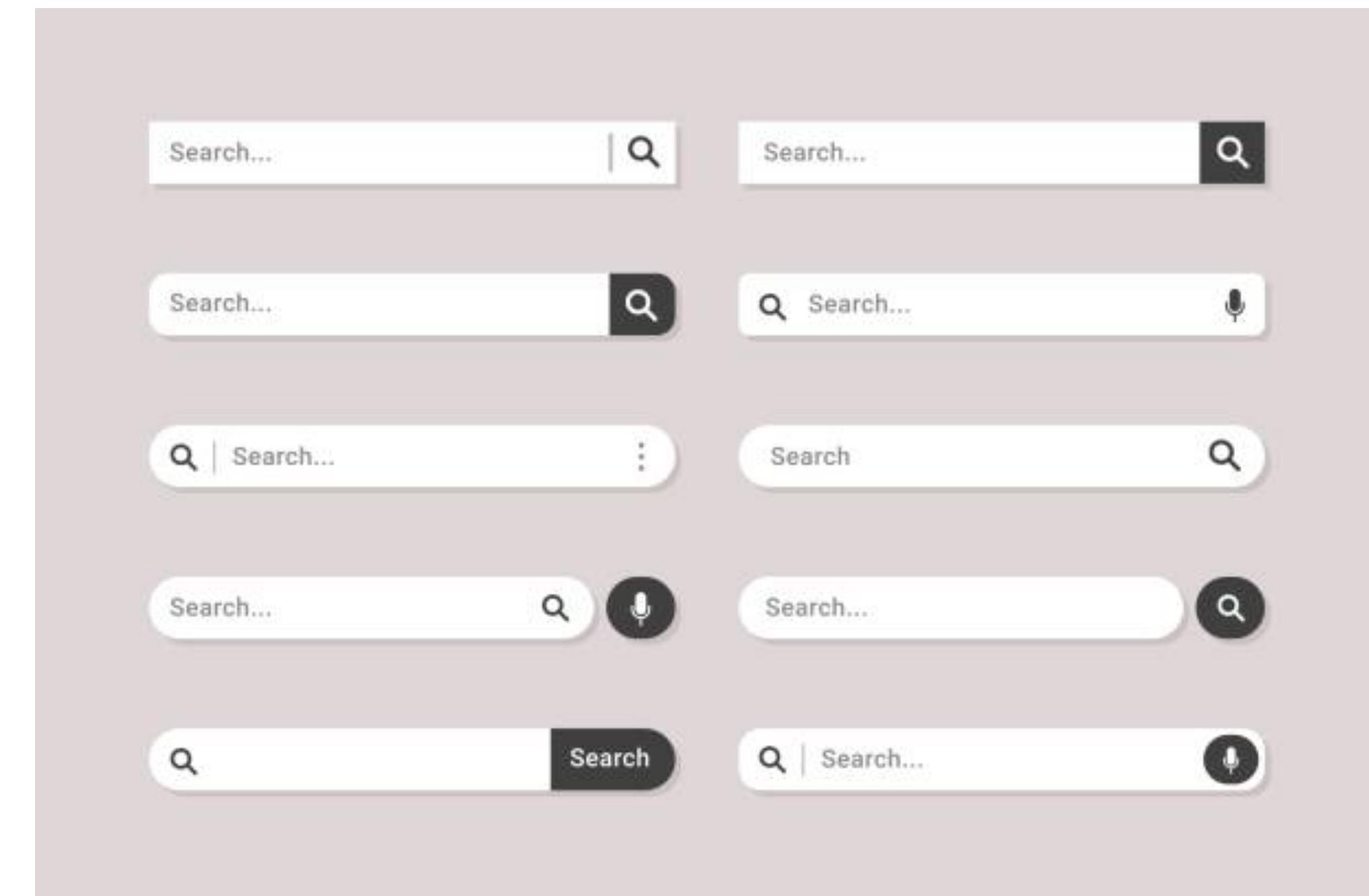
*planio



Debugging is twice as hard as writing the code in the first place.
Therefore, if you write the code as cleverly as possible, you are, by definition, not smart enough to debug it.

— Brian Kernighan —

AZ QUOTES



Scrum and ML

Refinement

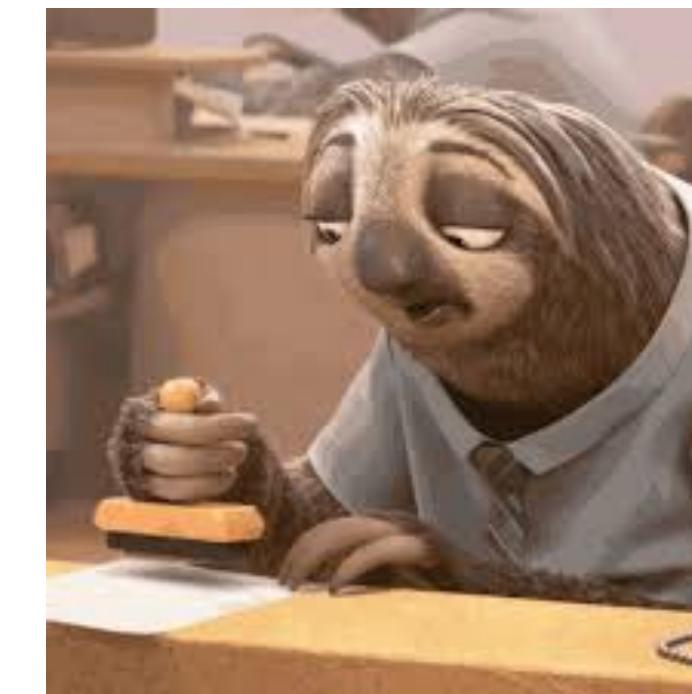
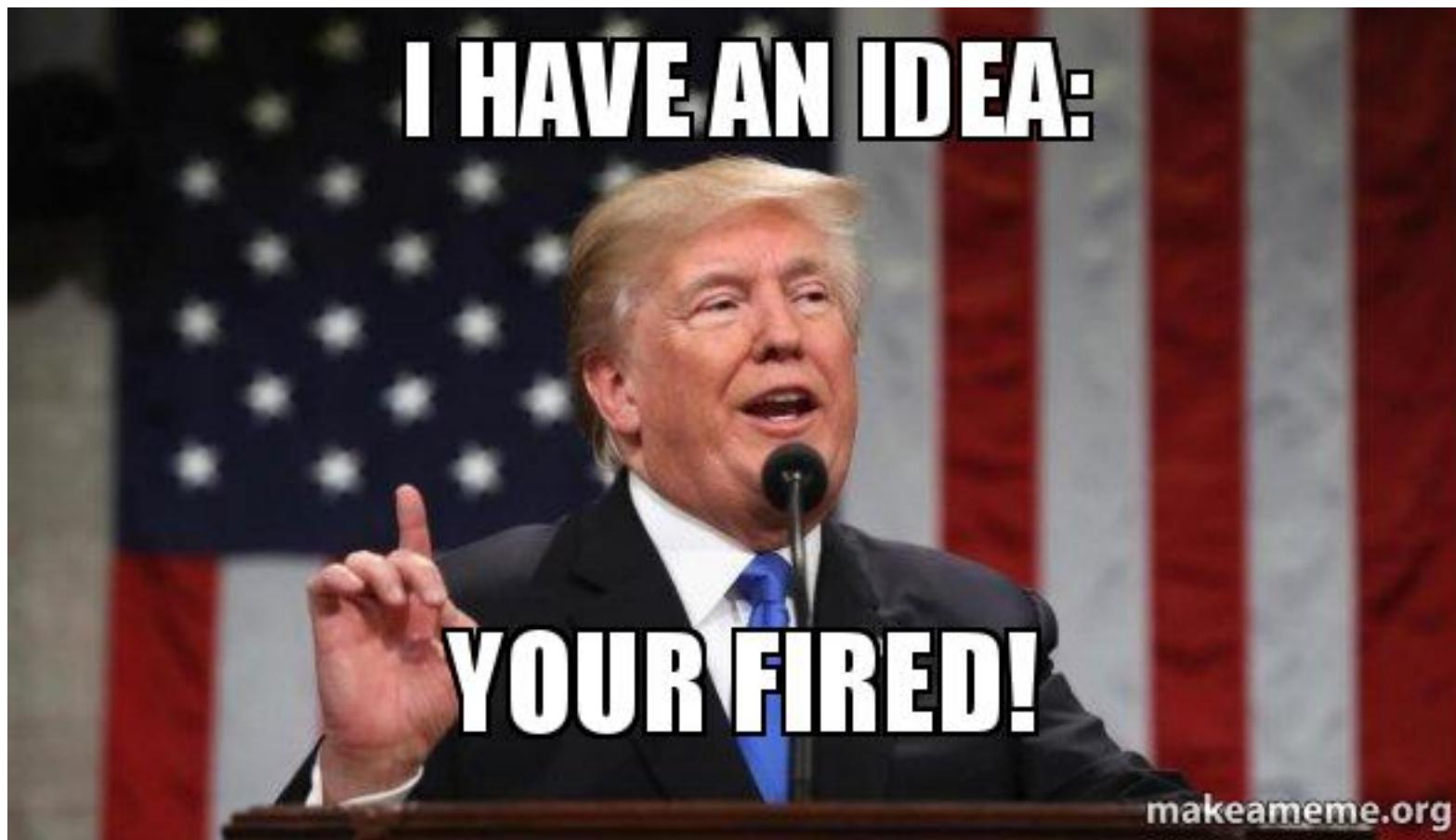
- Curse of Brilliant People.
- Performance Monitoring.
- Maintenance.
- POC in machine learning.
- Double checking core idea/implementation.
- Novel things estimation?
- It's enough.
- Consider Buffer.



Scrum and ML

Review

- Evaluation Metric of Sprint.
- You fire some one here.
- Excuses emerge here.
- Be strict on DoDs.



Scrum and ML

Retro

- If you do not seek team's progress:
 - Go to therapy.
 - Leave.
- Successful Stories.

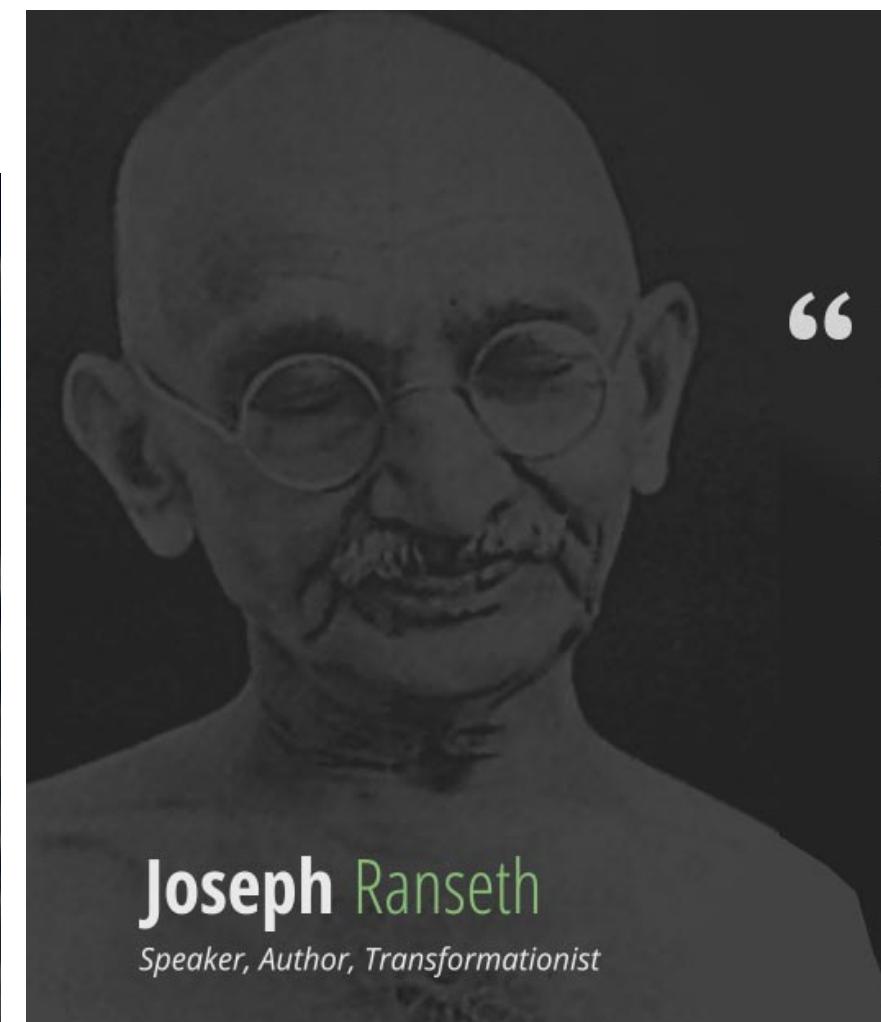
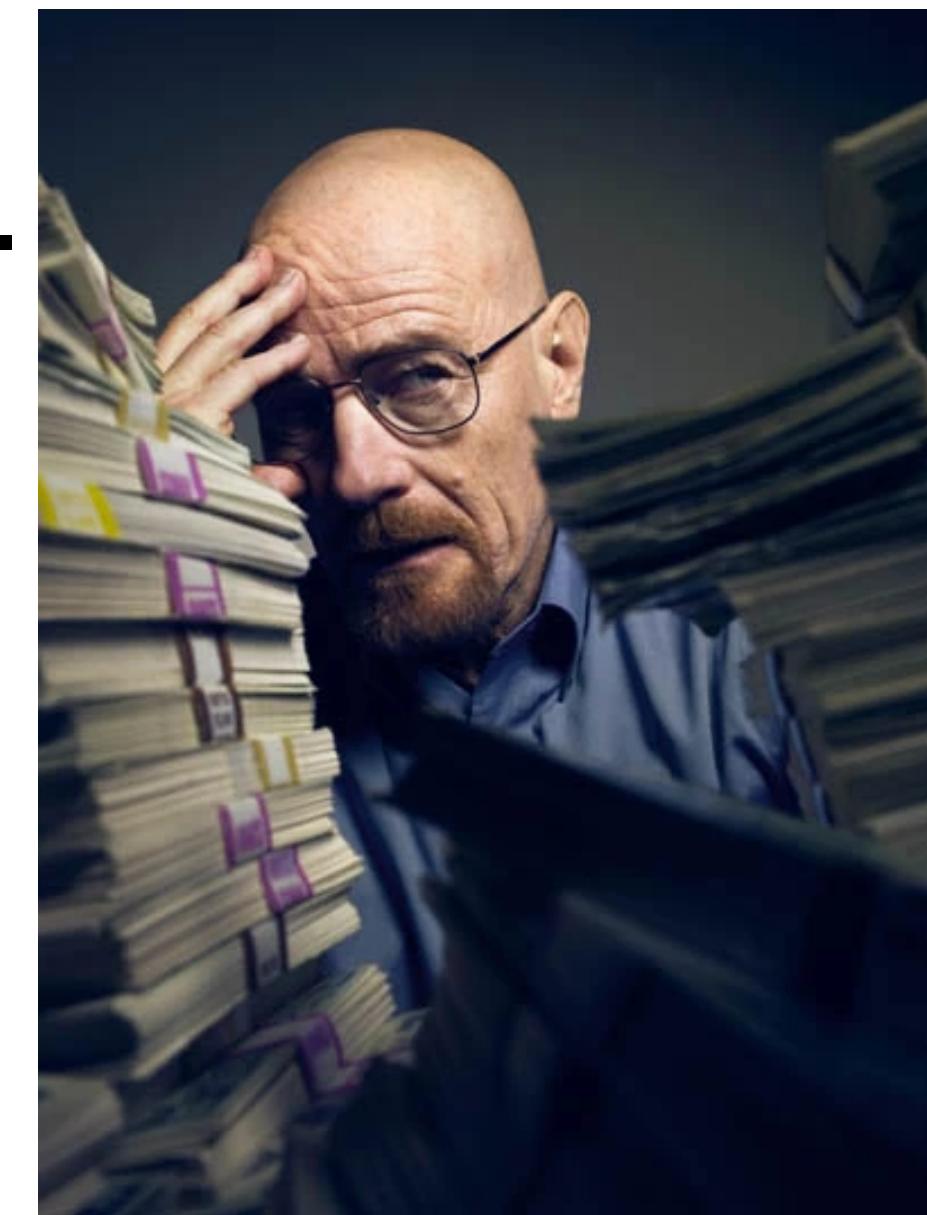


How to Success As a Data Scientist

- Learn fundamentals by your heart.
- Intuitively know what you do.
- Blend in teams.
- Not too aggressive, Not a potato.
- Be ambitious.
- Be Multi-Dimensional.
- **Work Ethics.**



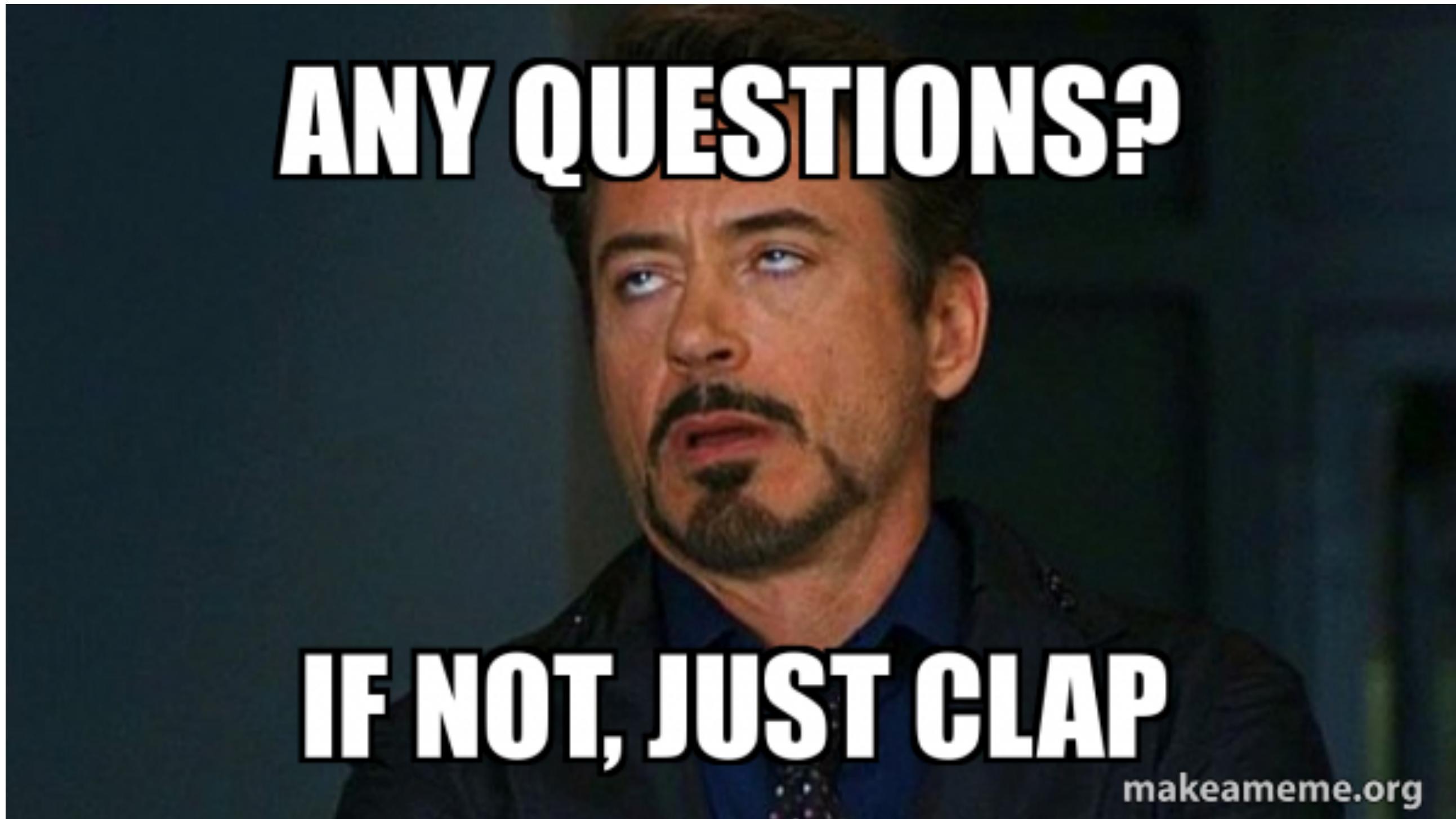
You should become Gholamreza in your field



Joseph Ranseth
Speaker, Author, Transformationist

“ Be the change you want
to see in the world.”

– Not ~~Gandhi~~



ANY QUESTIONS?

IF NOT, JUST CLAP

makeameme.org