

# Hành trình tới 44 usd

Xây dựng Team Research – bài toán kinh tế (thực tế)

# Bước đầu

Của research (07-2024)

- Team size: 2 (Alan Rex)
- GPU – Chủ yếu là 4090 (!!??)
- GPU không train được LLM
- → Mua GPU (😂)
- Lí do: GPUs ở Đài Loan quá rẻ



# Quá trình ban đầu

## Của research

- Team size: 3 (add Bách)
- Train 1 model multi-modal
- Ban đầu là llama3-s
- Training → Đọc paper nhé



# Sai lầm #1 : Bad saving

Của research

- A6000 – Training time 33 tiếng (Gap lớn)

- H100 – 8 Tiếng

- $30 \times 8 = 240$

- $10 \times 33 = 330$  (Lỗi rồi – lỗi cả time)

## Benchmark - A100, H100

Sound instruct - homebrew dataset

4x setup

H100x4 - SXM: ~15hours

A6000x4 - PCIe NVL: ~150hours (cpu offloading)

8x setup

H100x8 - SXM: ~8hours

A100x8 - PCIe: ~25hours

A100x8 - SXM: ~17hours

H100x8 - PCIe NVL: ~16hours

10x setup

A6000x10 - PCIe NVL: ~33 hours (decrease batch to 2 from 3) (edited)

# Sai lầm #1 : Bad saving

Của research

- A6000 – Training time 33 tiếng (Gap lớn)
- H100 – 8 Tiếng
- $30 \times 8 = 240$
- $10 \times 33 = 330$  (Lỗi rồi – lỗi cả time)



# Sai lầm #2 : xây tháp Babel

Của research

- Cố gắng xây tháp cao tới trời
- Xây hồi
- Cãi lộn tùm lum
- Ai về nhà nấy



Ichigo?

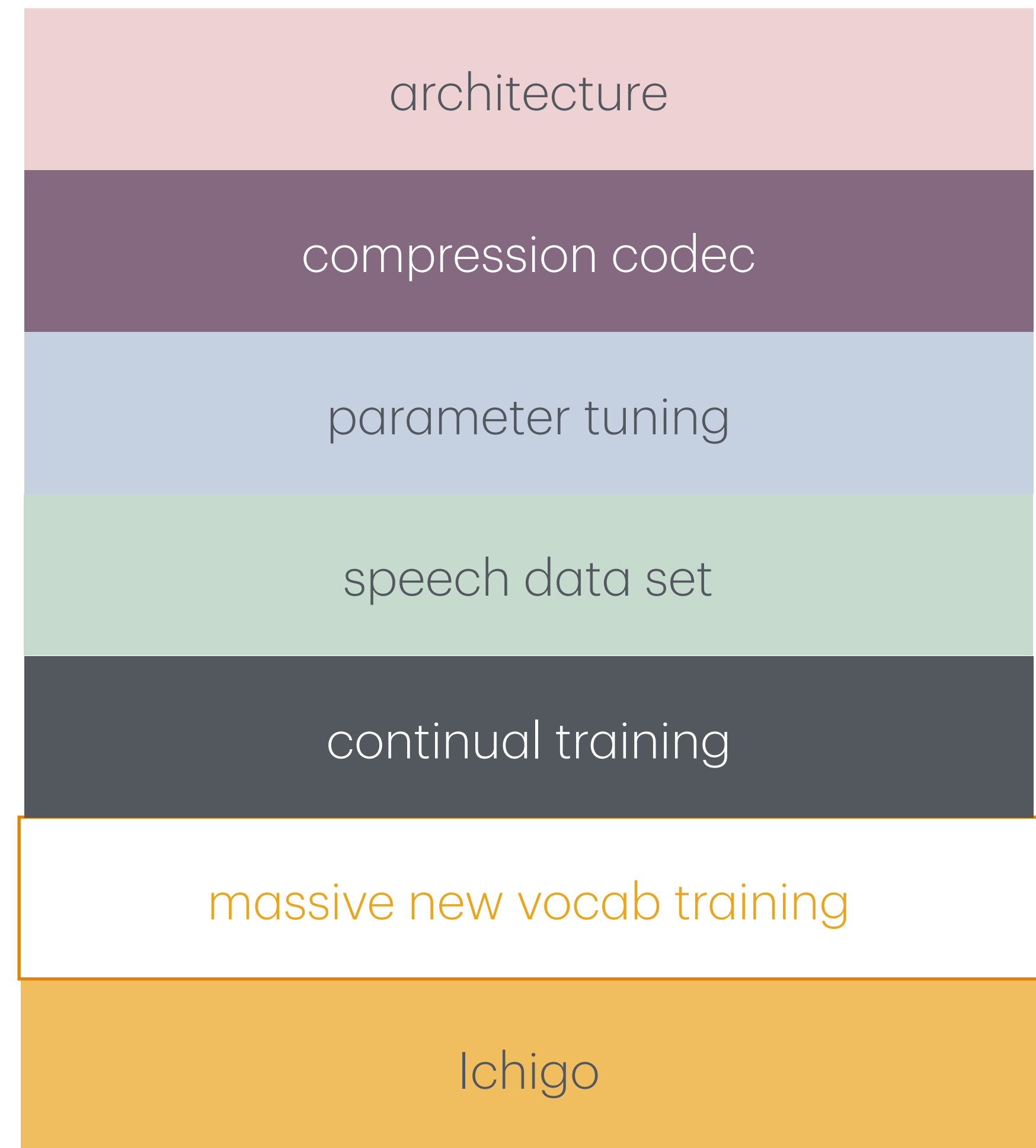




# Sai lầm #2 : xây tháp Babel

Của research

- 240\$ per run
- $240 \times 15-16$  lần = **4k USD**



# Sai lầm #2 : xây tháp Babel

Của research

- 240\$ per run
- $240 \times 15-16$  lần = **4k USD**

- 





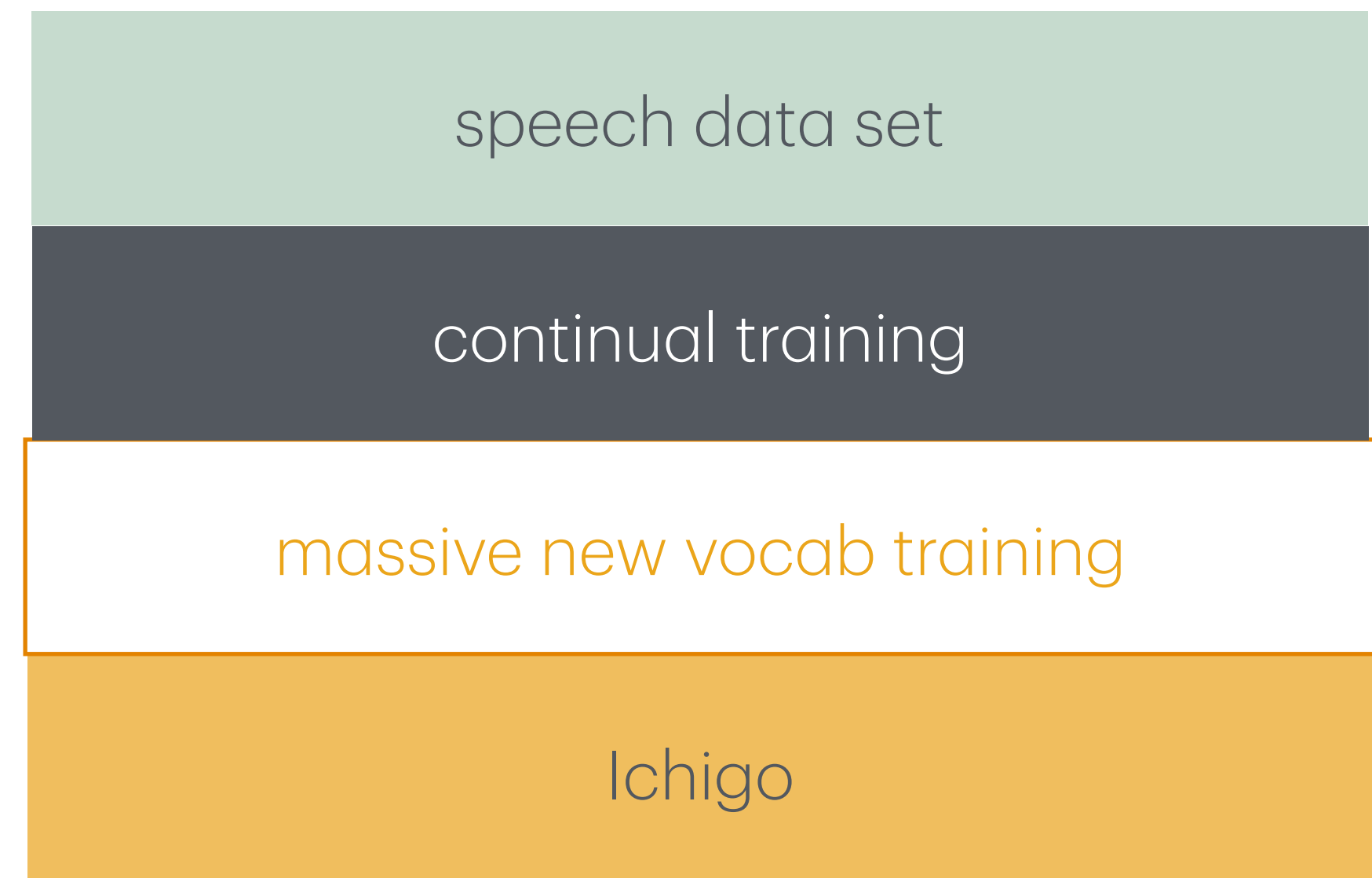
FIX

# Sửa sai : xây tháp Babel

Của research

- 240\$ per run
- $240 \times 15-16$  lần = 4k USD

- 



# Sửa sai : Đập tháp

Của research

- 24\$ per run
- Mỗi run độc lập
- Cô lập vấn đề tốt hơn

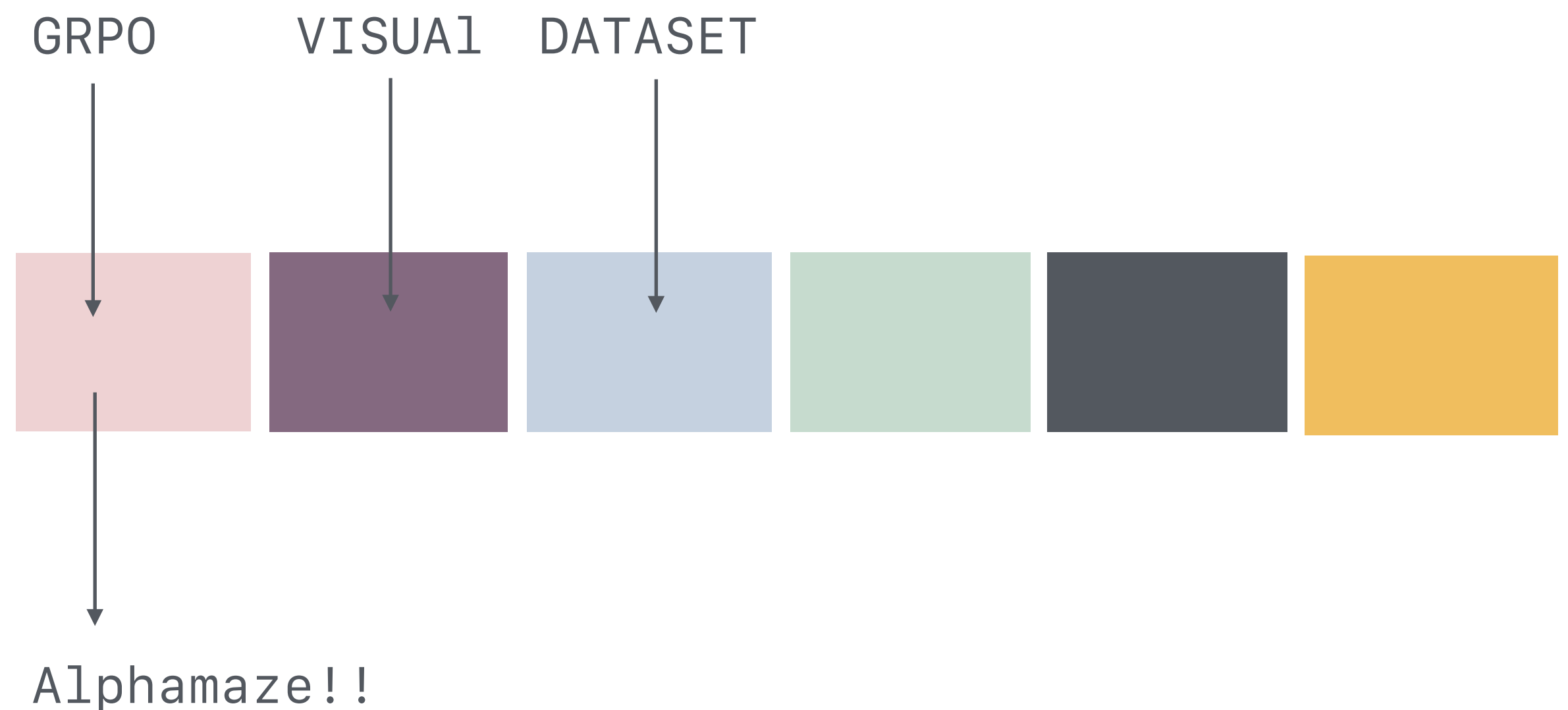




# Sửa sai : Đập tháp

Của research

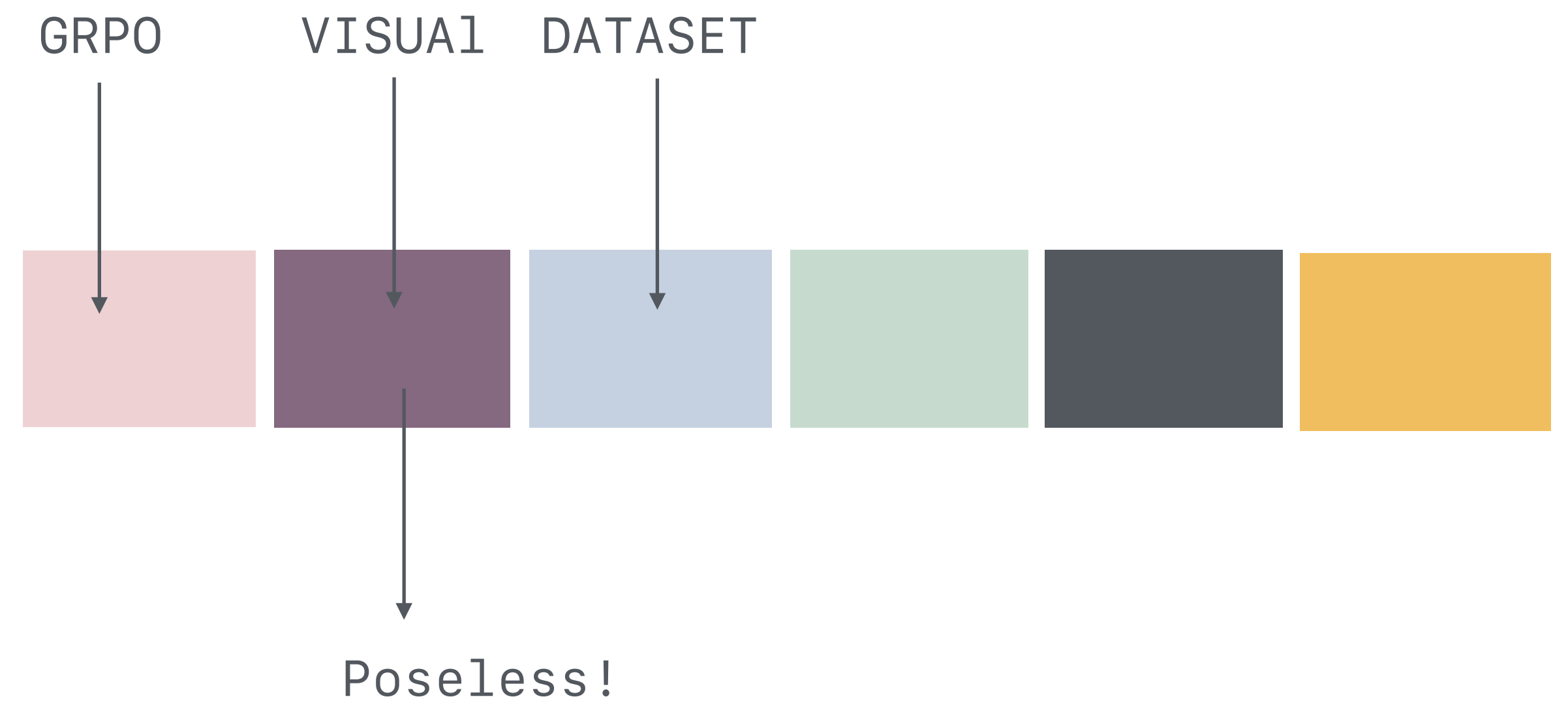
- 24\$ per run
- Mỗi run độc lập
- Cô lập vấn đề tốt hơn



# Sửa sai : Đập tháp

Của research

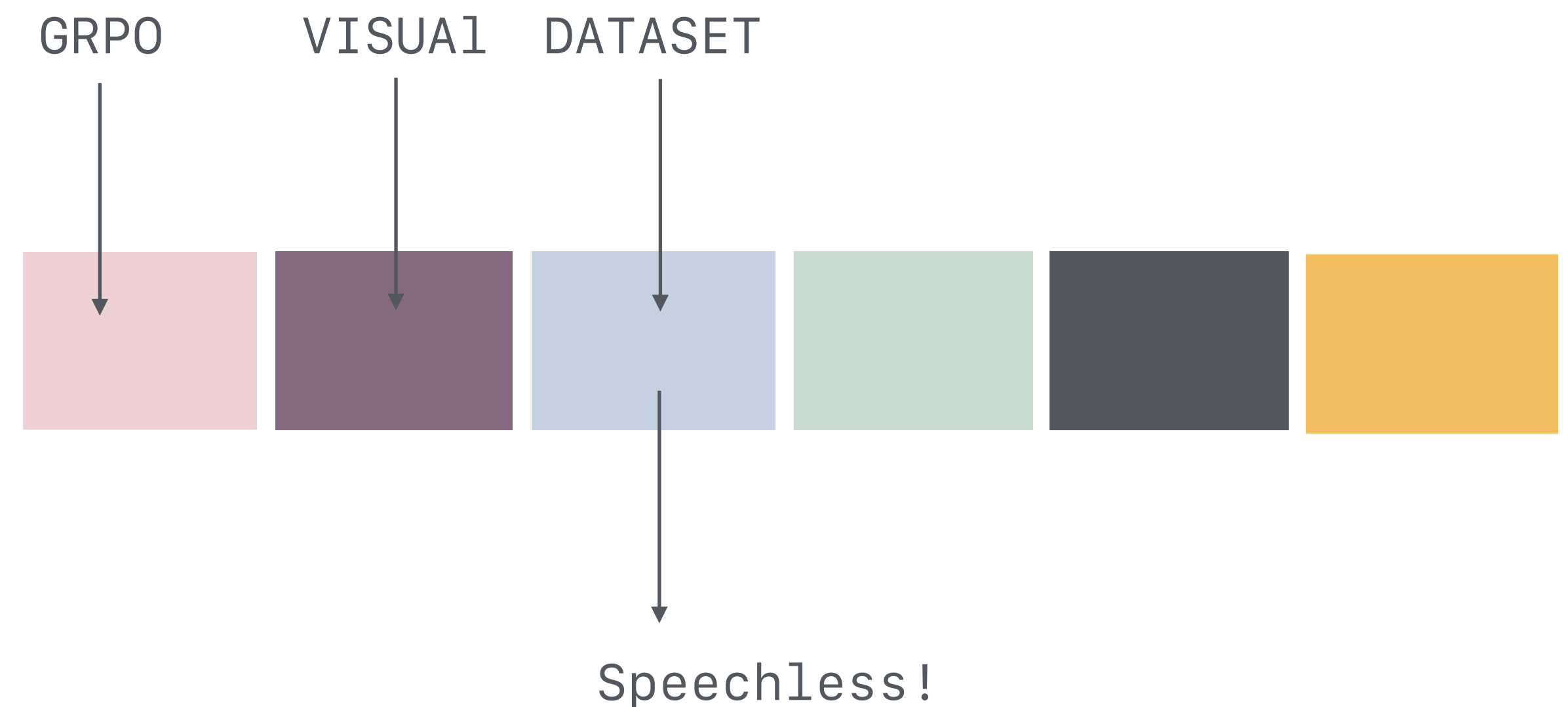
- 24\$ per run
- Mỗi run độc lập
- Cô lập vấn đề tốt hơn



# Sửa sai : Đập tháp

Của research

- 24\$ per run
- Mỗi run độc lập
- Cô lập vấn đề tốt hơn

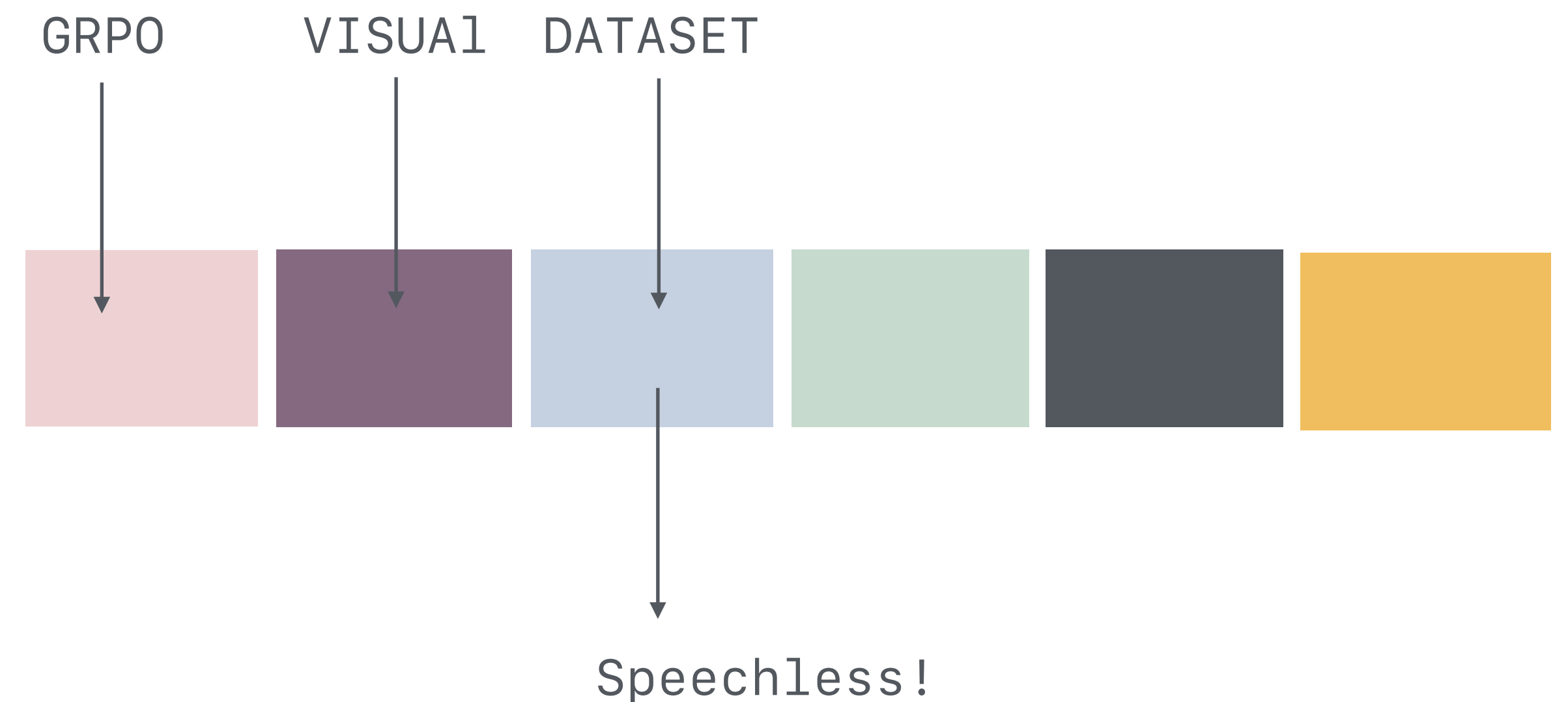




# Sửa sai : Đập tháp

Của research

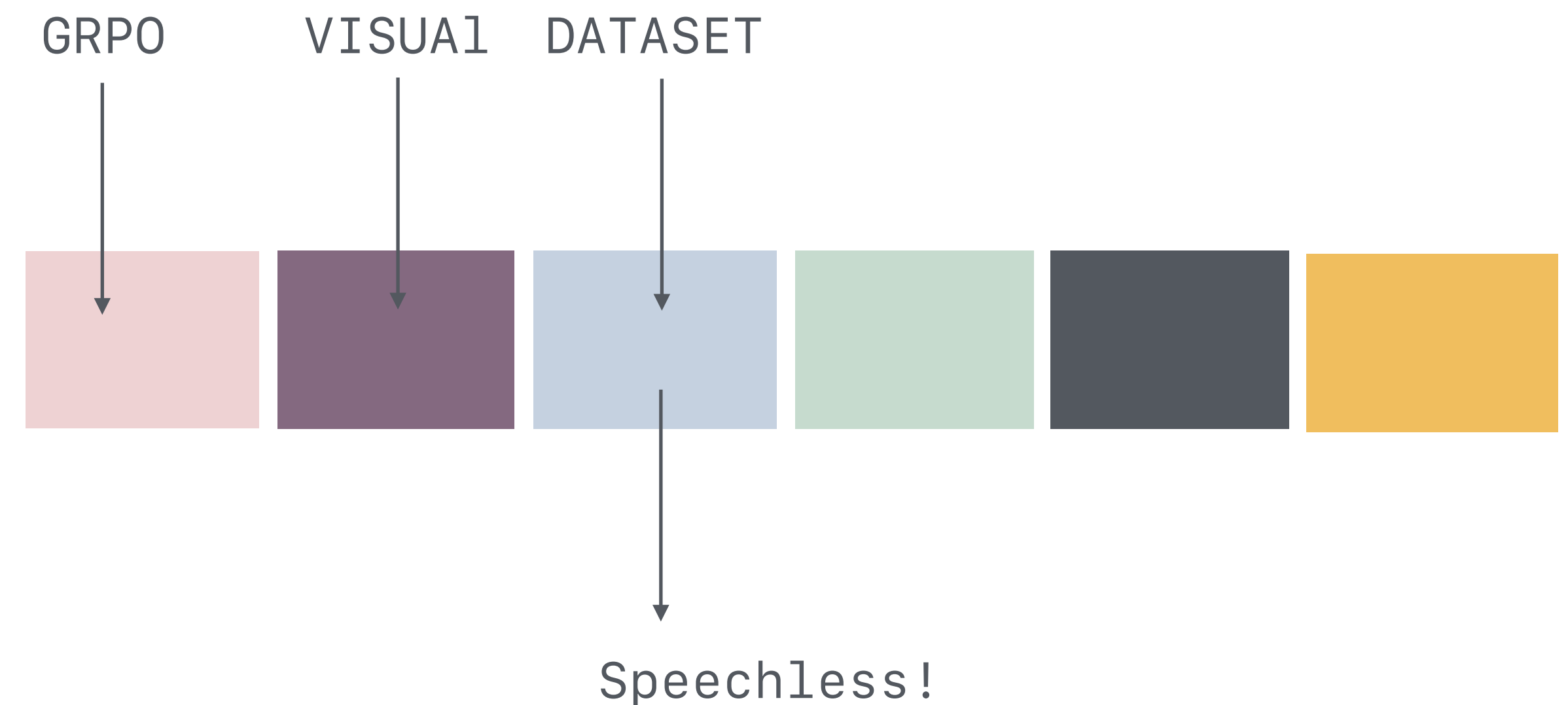
- 24\$ per run
- Mỗi run độc lập
- Cô lập vấn đề tốt hơn
- Đối xử research Investment



# Sửa sai : Đập tháp

Của research

- 24\$ per run
- Mỗi run độc lập
- Cô lập vấn đề tốt hơn
- Đối xử research Investment



# Sửa sai : Đập tháp

Của research

- **44\$**

- < 1.5 tiếng training H200

- Runpod credit

- > 300k total impressions

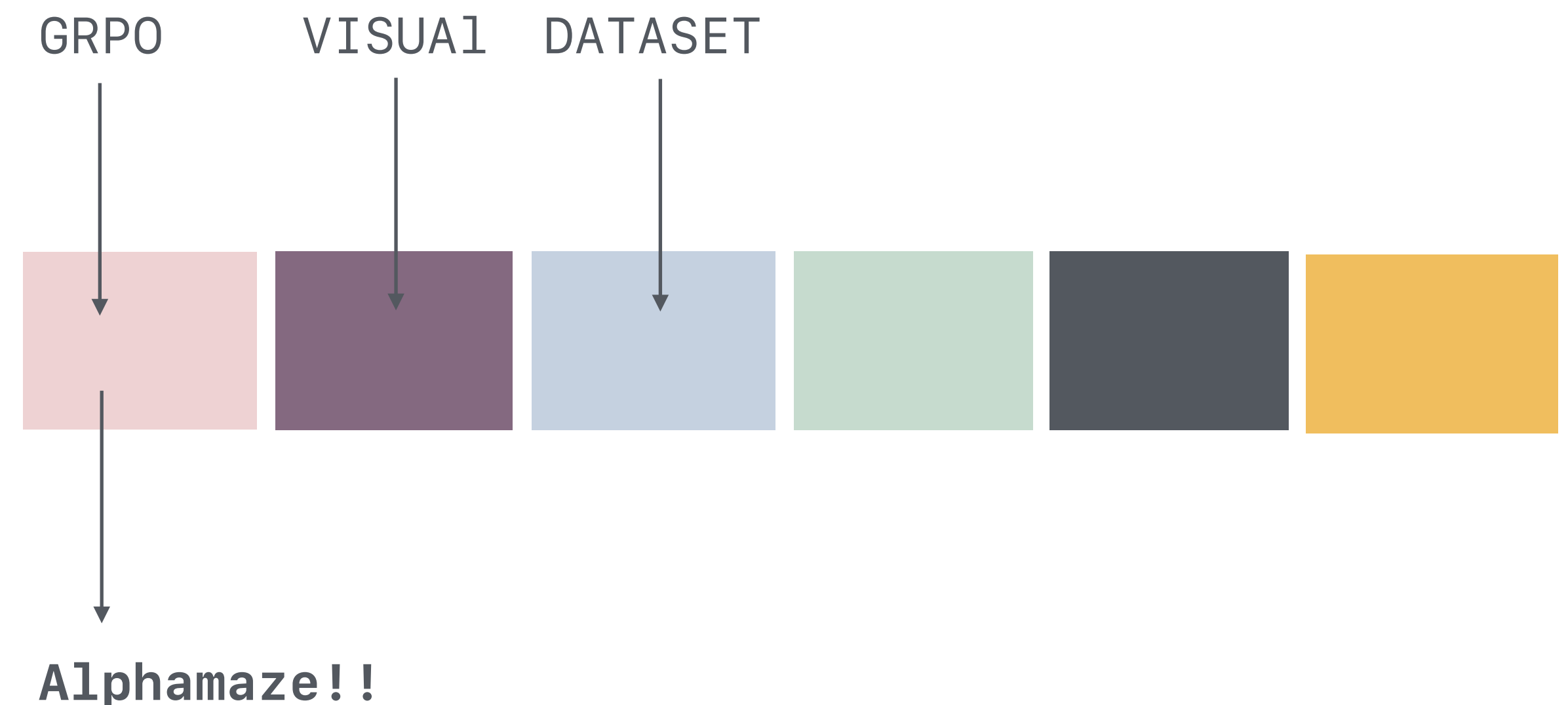
- > 100k plays trên demo

- Featured trên Kaggle

- Indirect traffic > 300k

- **4000\$ Ichigo**

- **Limited success**





# Sửa sai : Tối ưu Của research

Chi phí = Thời gian + Tiền

Thời gian = tiền (tiền lương 🥲)

Thời gian = turnover rate của research

# Sửa sai : Tối ưu Của research



Chi phí = Thời gian + Tiền

Thời gian = tiền (tiền lương 😭)

Thời gian = turnover rate của research

# Hiện tại

## Của research

- Team: 5 người (bài học về nhân sự, maybe kì tới)
- Có gpu internal
- 1 conference paper sau 1 năm
- move fast break things



# Join team ALAN

Ngày hôm nay!

CLICK HERE

[llama.cpp Contributor](#)  
Engineering

Remote  
Remote

[AI Research Scientist](#)  
Research

Remote  
Remote

[Robotic Controls Engineer](#)  
Robots

Remote  
Remote

[Robotics Software Engineer \(C++\)](#)  
Robots

Remote  
Remote

MENLO



# Join team Robotics

Ngay hôm nay!



Kiểm giùm  
C++ Alan ơ

[llama.cpp Contributor](#)  
Engineering

note  
Remote

CLICK HERE

[AI Research Scientist](#)  
Research

Remote  
Remote

[Robotic Controls Engineer](#)  
Robots

Remote  
Remote

[Robotics Software Engineer \(C++\)](#)  
Robots

Remote  
Remote

MENLO

