SPEAKER

Stefan Krawczyk

Co-creator of Hamilton

CEO & Co-Founder

DAGWORKS

STITCH FIX · IDIBON · IBM · VICTORIA UNIVERSITY OF WELLINGTON TE HERENGA WAKA

H2O.ai

H2O OPEN SOURCE GenAI WORLD CONFERENCE • SAN FRANCISCO

LLMOps: Match report from the top of the 5th

Stefan Krawczyk
CEO & Co-founder

DAGWORKS

Why: [Dev|ML|LLM]Ops?

*Ops == Leverage

Leverage ⇒ 📈 ROI

Talk Overview

- 🤔 MLOps vs LLMOps
- 🎙️ Top of the 5th
- 🔮 Forecast
- 🏠 Take home

MLOps vs LLMOps

vs

MLOps vs LLMOps

vs

MLOps ~4 years

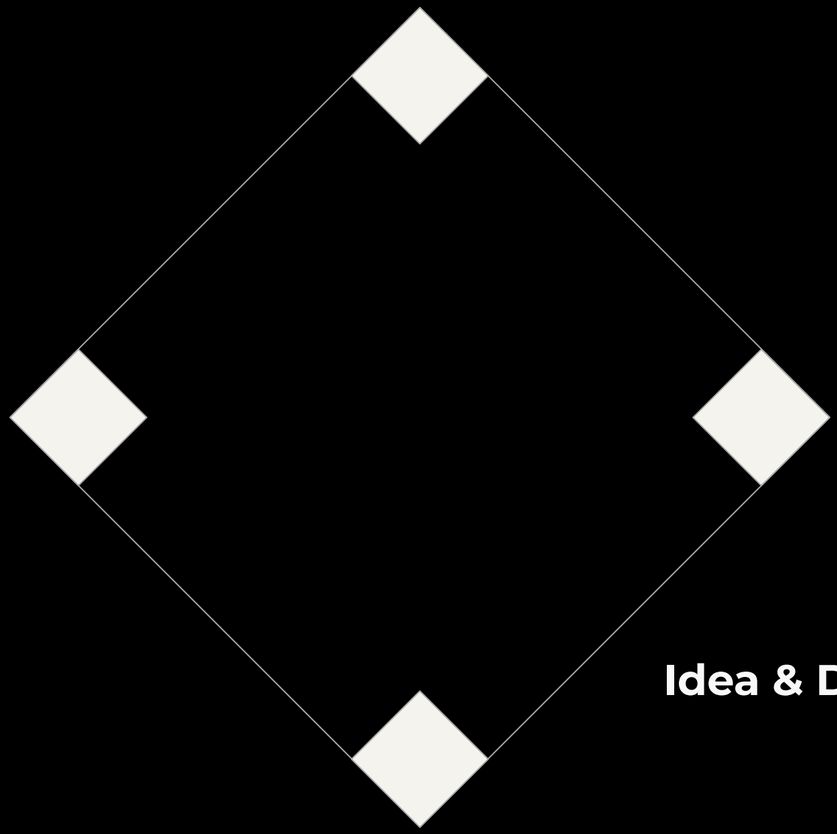
LLMOps < 1 year

MLOps Recap

vs

MLOps Recap

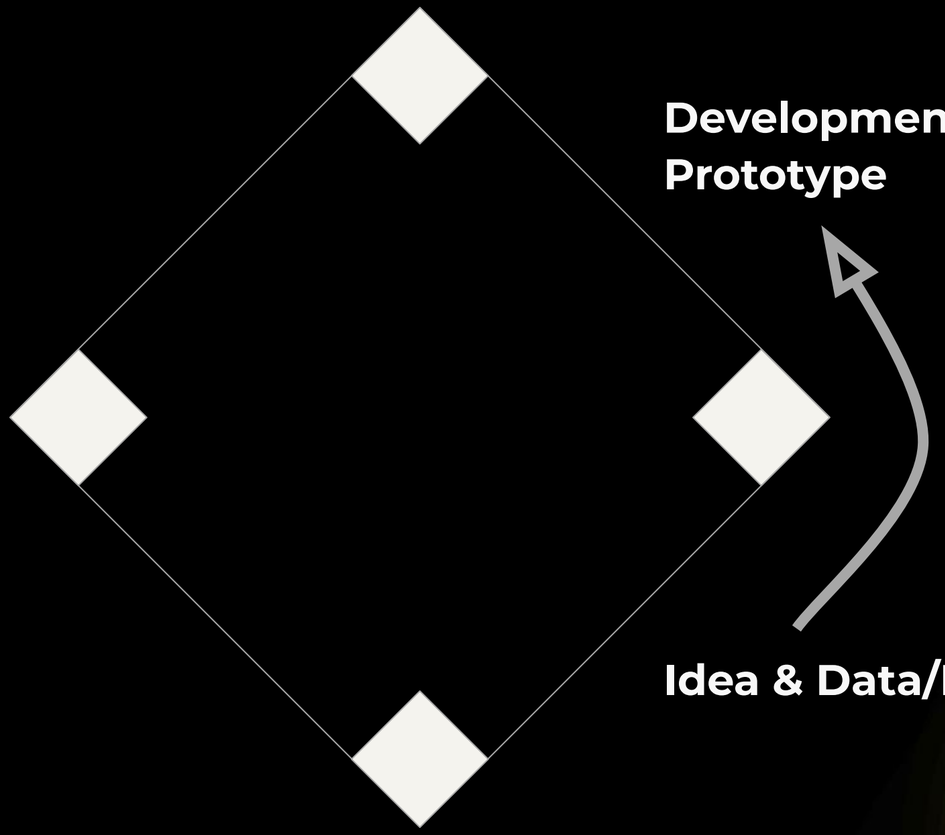
vs



Idea & Data/Resources

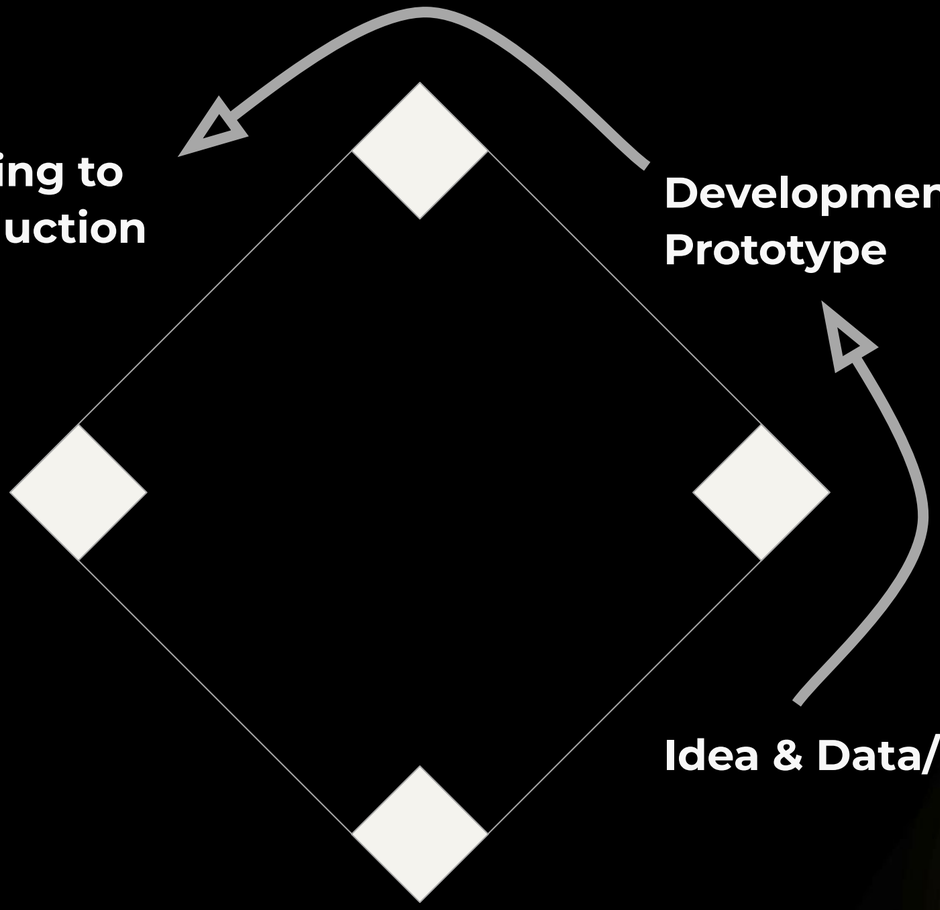
MLOps Recap

vs



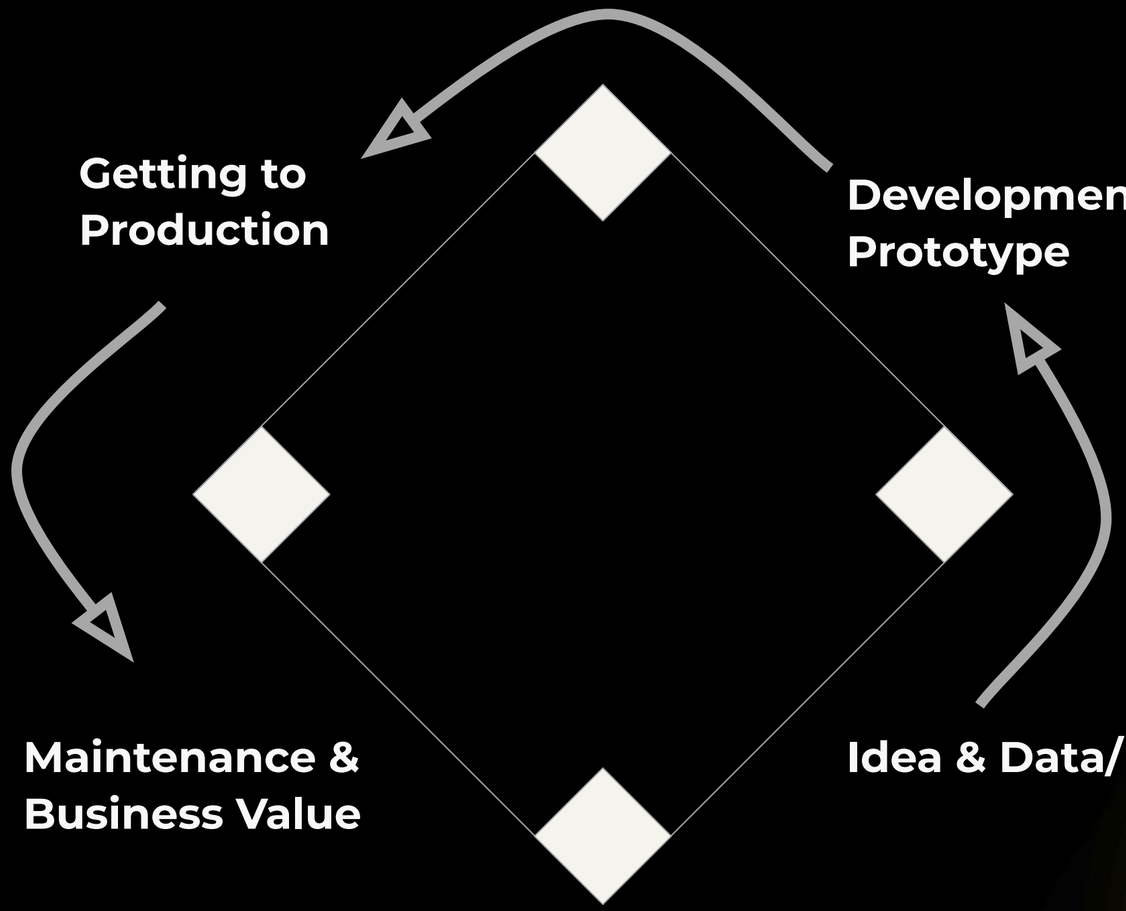
MLOps Recap

vs



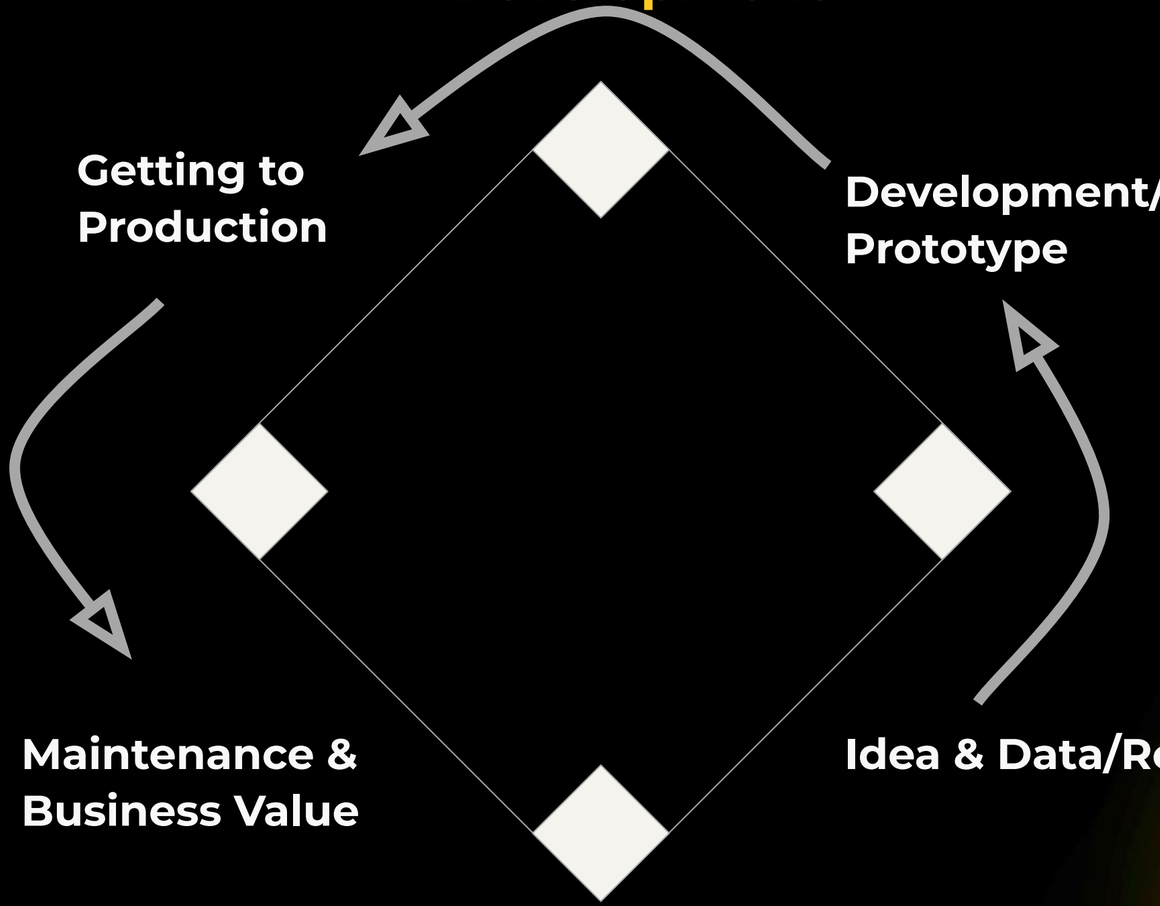
MLOps Recap

VS



MLOps Recap

VS



MLOps Recap

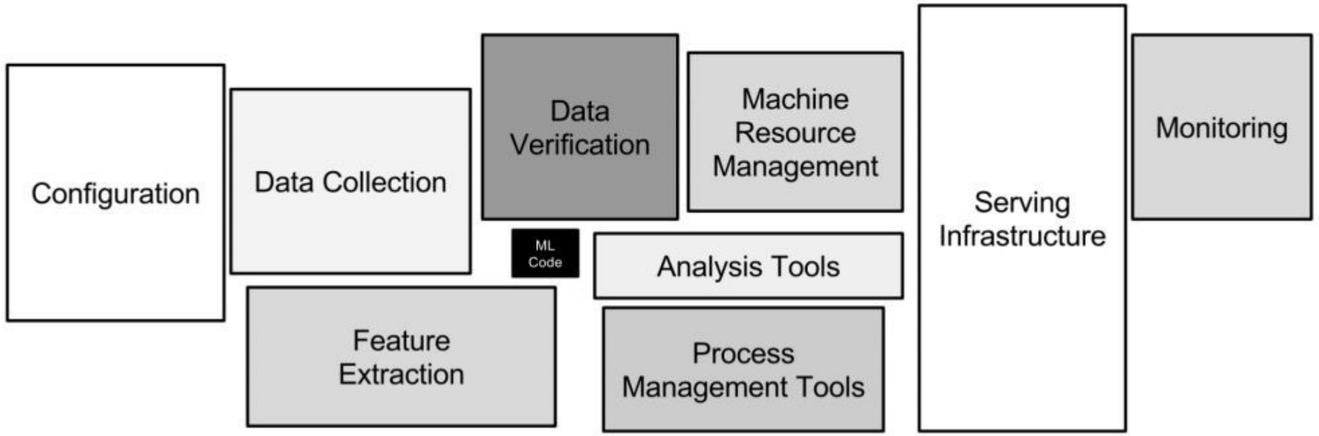
vs

H2O OPEN SOURCE GenAI WORLD

H2O.ai

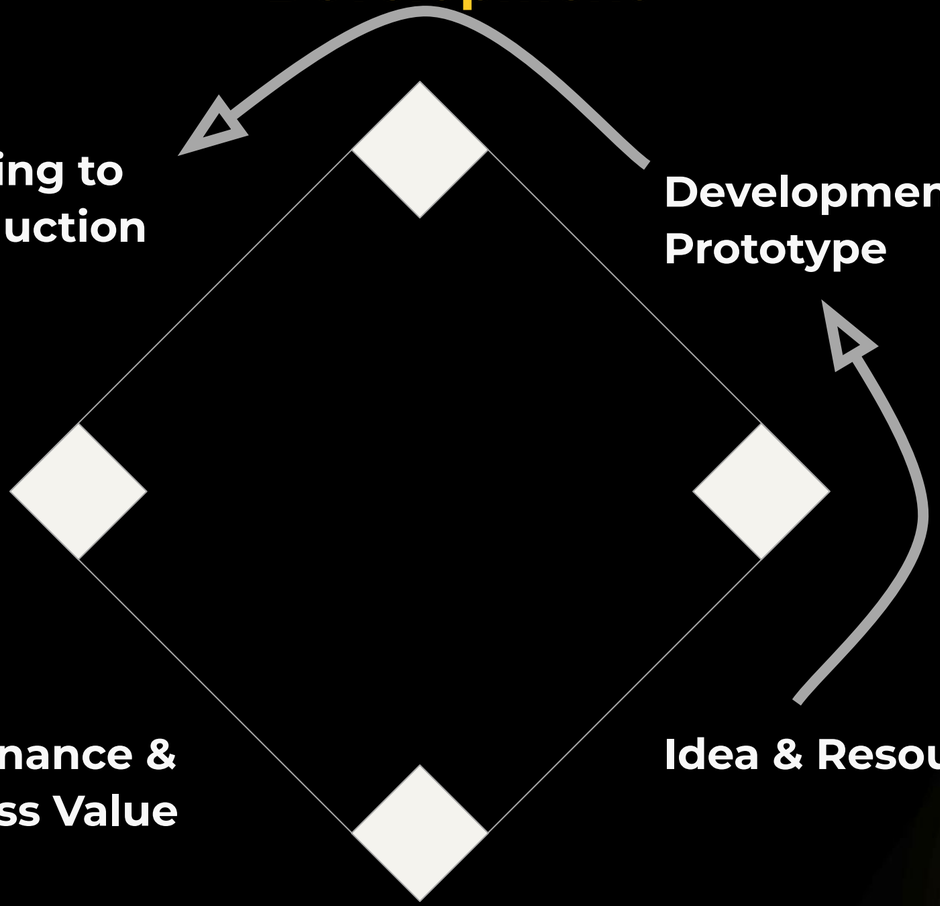
Hidden Technical Debt in Machine Learning Systems

D. Sculley, Gary Holt, Daniel Golovin, Eugene Davydov, Todd Phillips
{dsculley,gholt,dgg,edavydov,toddphillips}@google.com
Google, Inc.



LLMOps ⊆ MLOps

vs

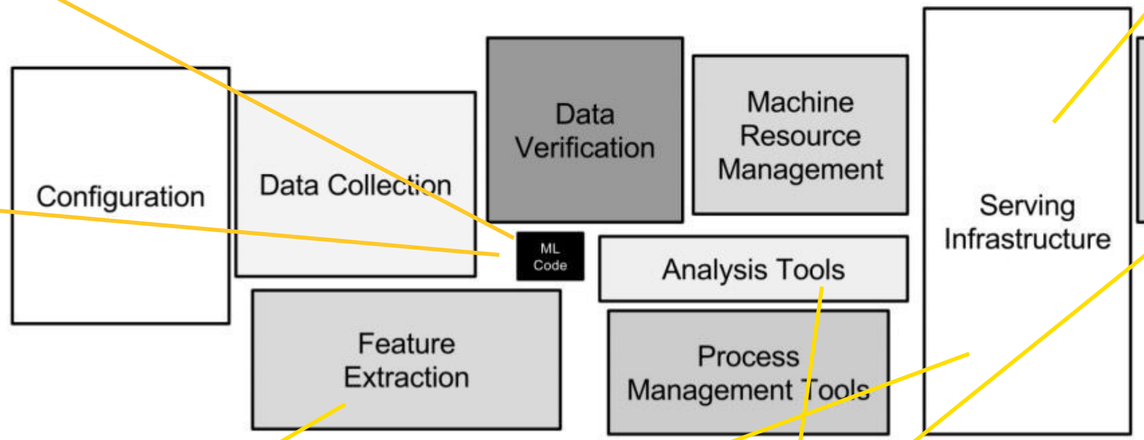


LLMOps ⊆ MLOps

vs

Hidden Technical Debt in Machine Learning Systems

D. Sculley, Gary Holt, Daniel Golovin, Eugene Davydov, Todd Phillips
{dsculley,gholt,dgg,edavydov,toddphillips}@google.com
Google, Inc.



- Prompts
- Fine Tuning
- Foundation Model Platform / Provider
- Safety Filters
- Embeddings
- Vector Databases
- Human Evaluation
- Factuality Assurance

LLMOps ⊆ MLOps

vs

High level: LLMOps vs MLOps

- Same general shape of problems

LLMOps ⊆ MLOps

vs

High level: LLMOps vs MLOps

- Same general shape of problems

Low level: LLMOps vs MLOps

- GPUs required
- Application integration pace
- More “models” in a single application
- Evaluation is fuzzier

Top of the 5th

Top of the 5th

Two teams:

- Proprietary vs Open source

Innings:

- around the 5th generation of models

Fans:

- Privacy vs Cost vs Controllability

Top of the 5th

LLMOps Space:

- Point solutions:
  - prompts
  - tracing
  - evaluation
  - hosting
  - embeddings
  - cost tracking
  - governance
  - vector DBs
  - etc

MLOps Platforms → LLMOps Platforms

A challenging play:
GPT3.5 -> GPT4

A challenging play: GPT3.5 -> GPT4

```
Summarize this text from a scientific article.
Extract any key points with reasoning.

Content:
```

A challenging play: GPT3.5 -> GPT4

```
Summarize this text from a scientific article.
Extract any key points with reasoning.

Content:
```

Same prompt different output!

A challenging play:
GPT3.5 -> GPT4

Other things to think about...

- context windows
- tokens & cost
- latency

A challenging play:
GPT3.5 -> GPT4

Other things to think about...

- context windows
- tokens & cost
- latency

Q: is it worth it?

A challenging play:
GPT3.5 -> GPT4

Other things to think about...

- context windows
- tokens & cost
- latency

Q: is it worth it?

⇒ 🦙 LLMOps processes

⇒ □ Evaluation is your backbone

Some curveballs seen

1. **Not controlling model updates***
2. **New foundational model provider features**
   e.g. plugins, privacy, multi-modal, PDF upload

Forecast

Forecast

Computation

- Cost curves 📉

Forecast

Computation

- Cost curves 📉

Foundational Models

- Proprietary vs Open Source
- Improvements:
  - Context windows
  - Simpler prompts
  - Multi-modal

Forecast

Computation

- Cost curves 📉

Foundational Models

- Proprietary vs Open Source
- Improvements:
  - Context windows
  - Simpler prompts
  - Multi-modal

Organizationally

- Data is your moat
- Evaluations are 🔑
- FE & BE Devs learning LLMOps
- Shift to fine-tuning
- Still need DS & ML

Forecast

Computation

- Cost curves 📉

Foundational Models

- Proprietary vs Open Source
- Improvements:
  - Context windows
  - Simpler prompts
  - Multi-modal

Organizationally

- Data is your moat
- Evaluations are 🔑
- FE & BE Devs learning LLMOps
- Shift to fine-tuning
- Still need DS & ML

LLMOps Space

- Reduction in “wrappers”
- Focus on fine tuning & evaluation

To bring it *home*:

1. 🚀 Plan for rapid evolution
2. ⚾ Strong practices to play

What I'm building:

Standardizing code to enable simpler *Ops.

Hamilton
open source

+

DAGWORKS

THANK YOU!

Questions?

Contact

Stefan Krawczyk
CEO & Co-Founder DAGWorks Inc.
@ stefan@dagworks.io
in/skrawczyk

Yes: all images were generated by DALL-E3