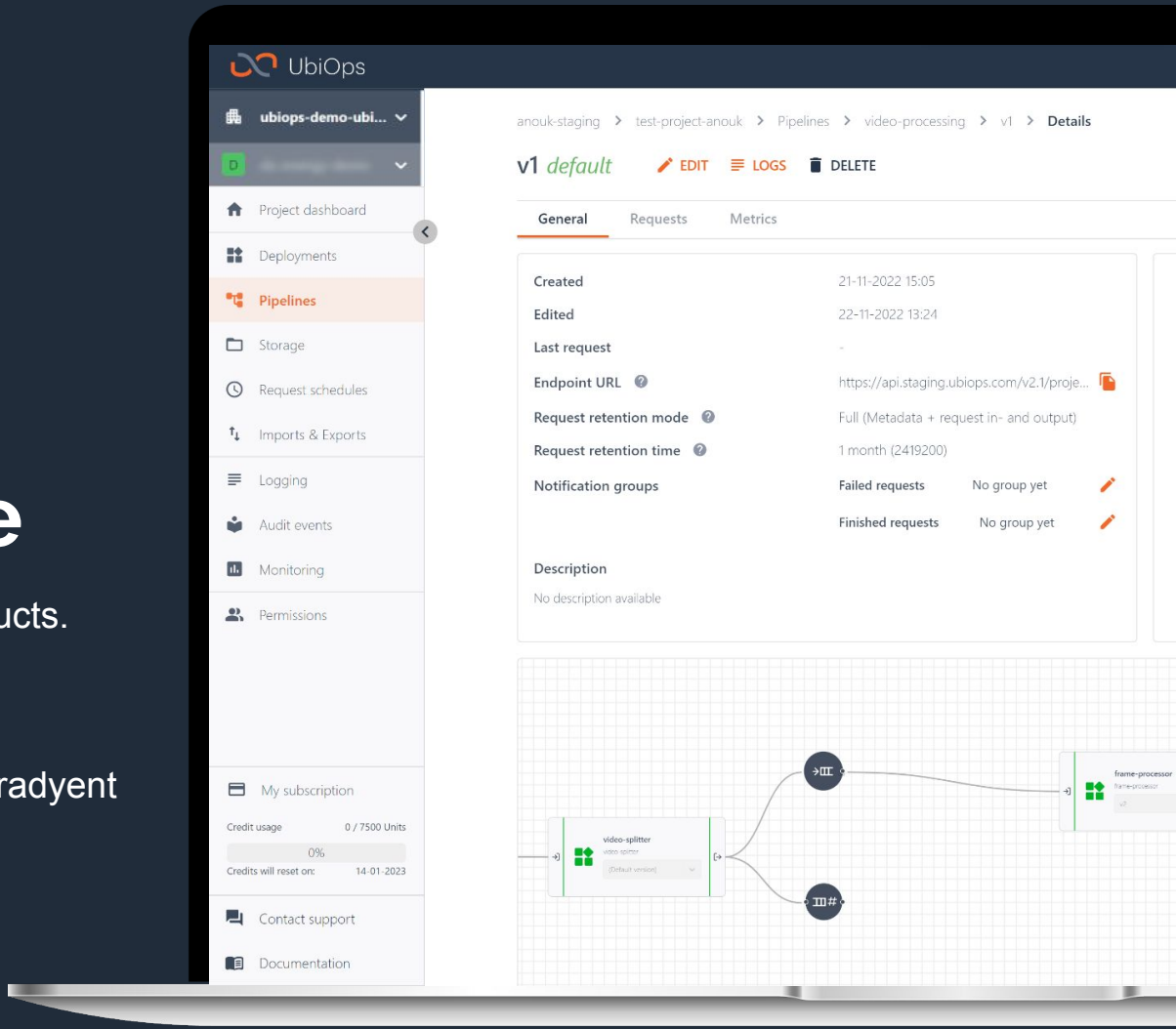
UbiOps

WSAI Montréal 2023

Deploying AI models at scale

Building production-grade AI / ML products.

Victor Pereboom - CTO @ UbiOps
Tanja Crijns - Software Engineer @ Gradyent



UbiOps

/ There is a huge growth in AI services. Many businesses want to turn **ML & AI models into** products**.**

UbiOps

Productionizing AI & ML: The next big thing?

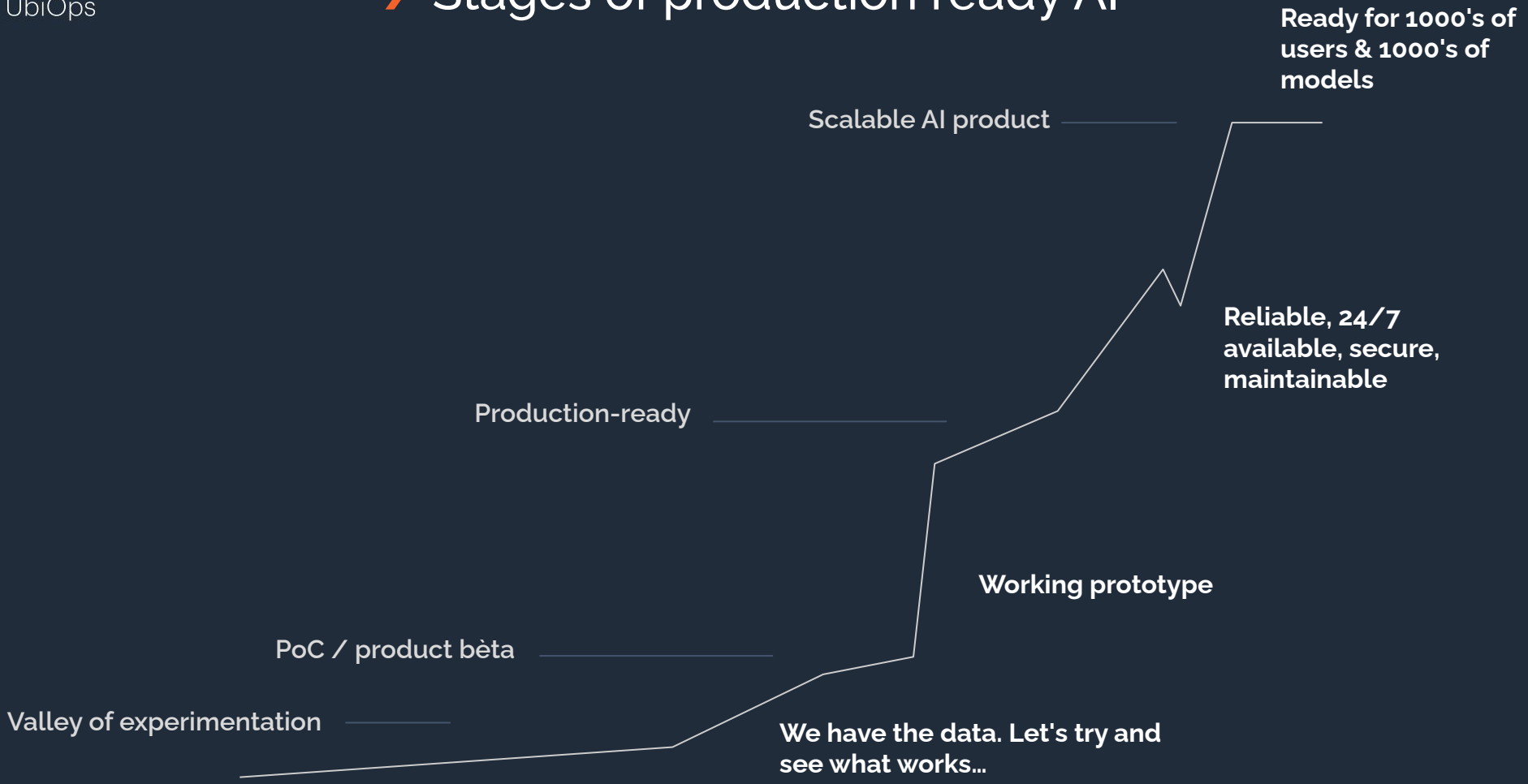
UbiOps

/ **Inference** is becoming more important than **training.**

We're entering a new era of AI applications

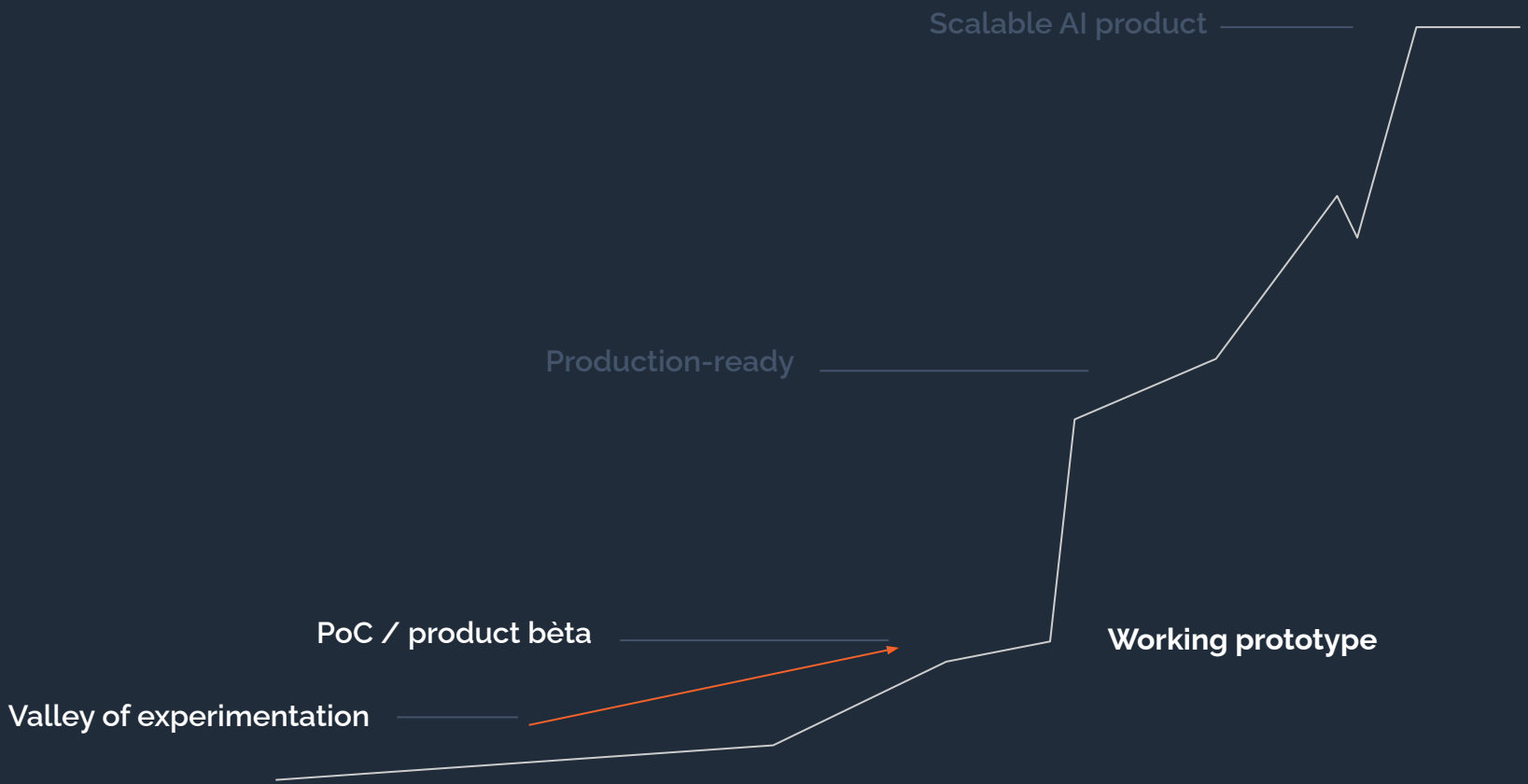
UbiOps

/ Stages of production ready AI



UbiOps

Stages of production ready AI



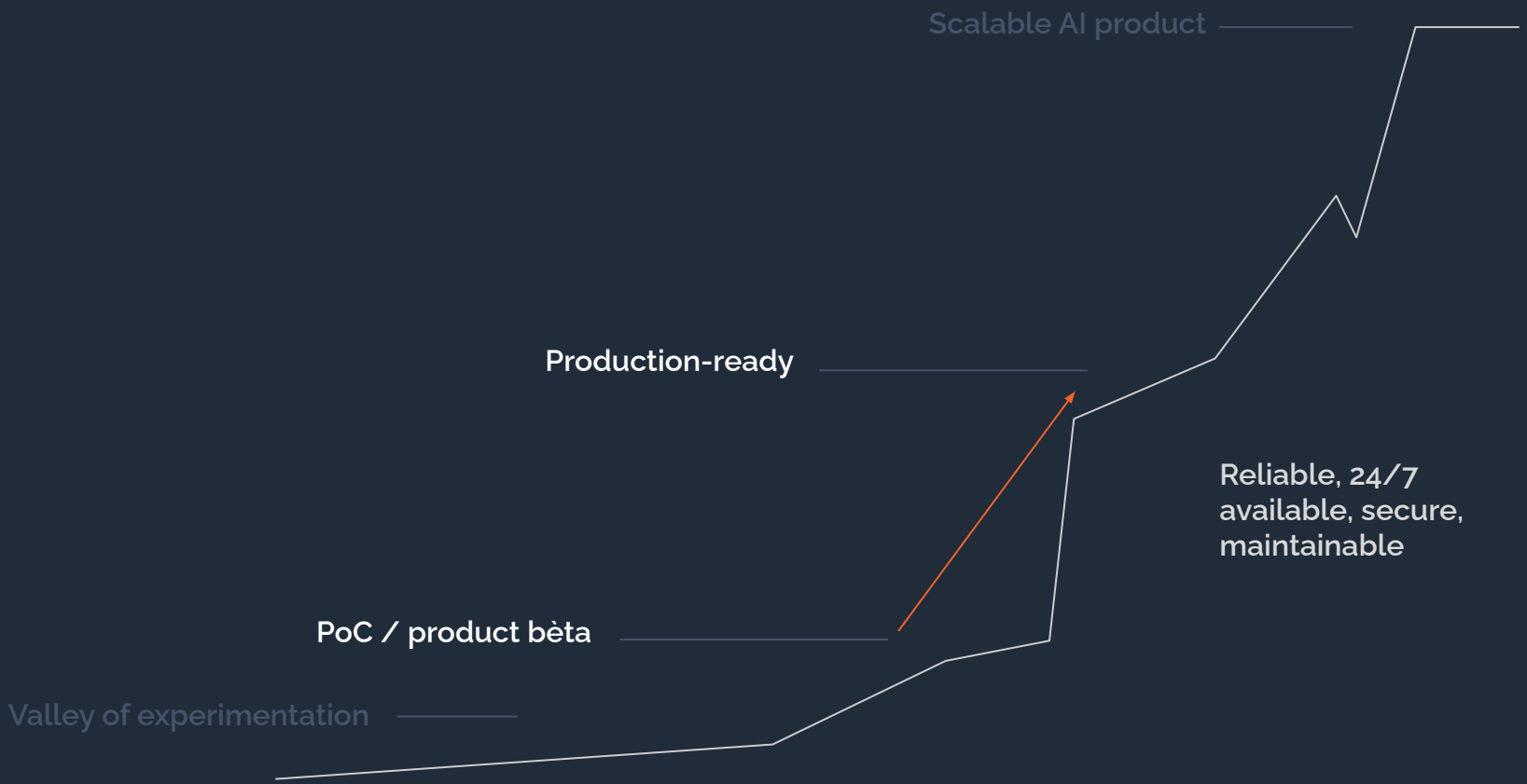
UbiOps

/ **Foundation models** will change how we approach AI system development. But there are also **thousands** of business cases that need a more traditional data science approach.

"*Not everyone is Google*"

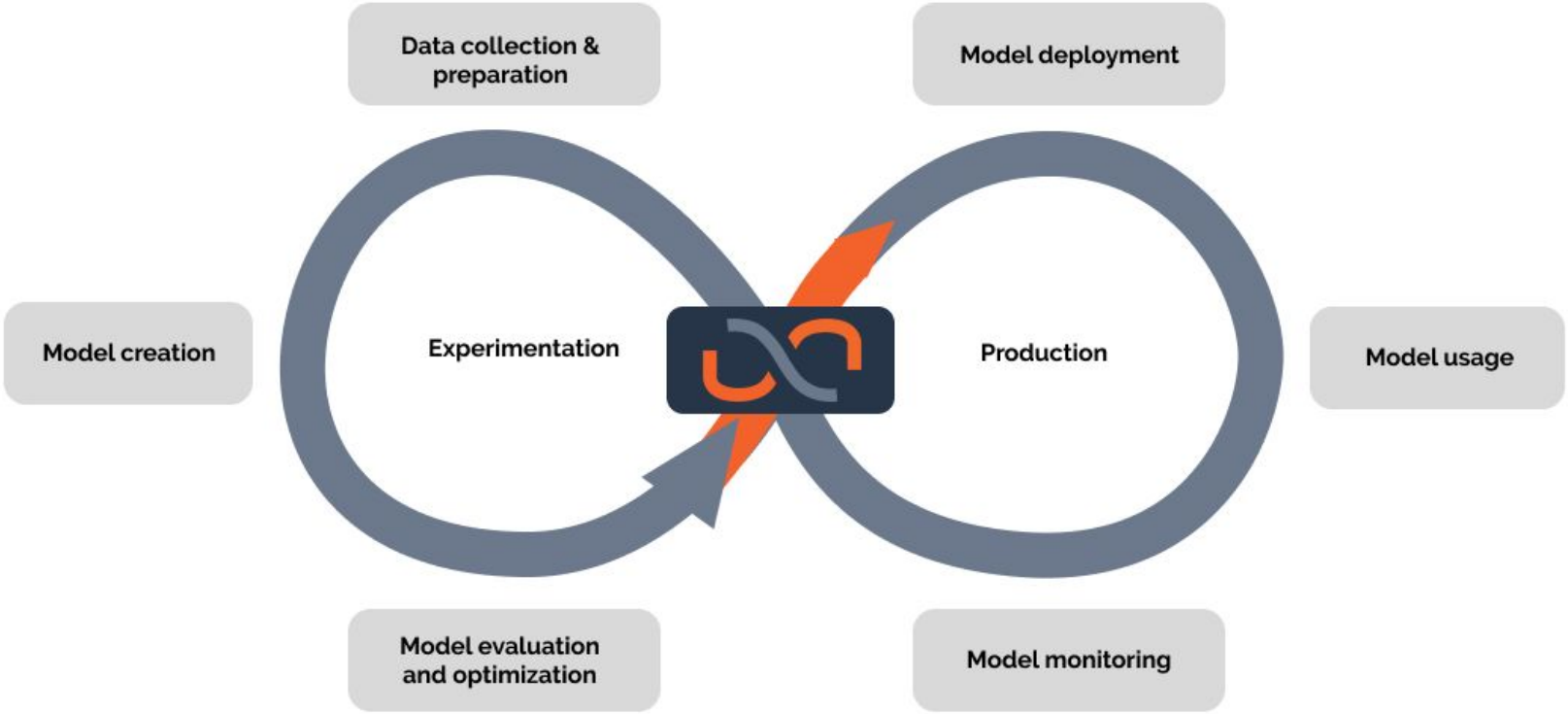
UbiOps

From PoC to a model in production



UbiOps

/ What does **productionization** even mean?



UbiOps

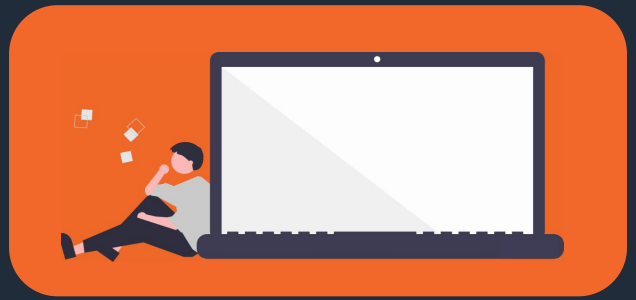
In the end, with AI in production, teams suddenly need to care about **reliability, uptime** and **security** of their solutions

UbiOps

/ What does **productionization** even mean?

New, non AI / ML related things to deal with

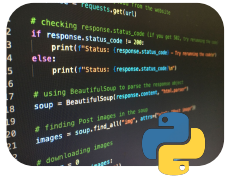
- Reproducibility *("It worked on my machine this morning …")*
- Maintainability *("Was it 'churn_v5_thisone.py' or 'churn_v5_final.py'…?")*
- Robustness *("Let's hope they don't kill our spot instance during the demo… ")*
- Security & compliance *("Do we have someone named Cody in our team??")*



UbiOps

/ The **AI model** stays the same, the complexity of the solution explodes

AI Model



+

IT infra

User & permission management

Auto scaling

Security

?!

Integrations

kubernetes

docker

Server configuration

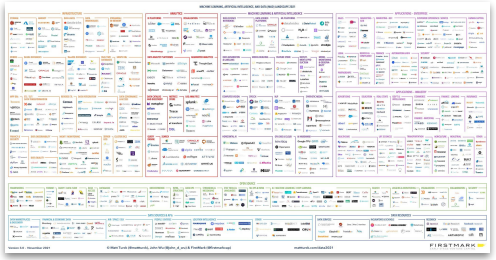
Logging & monitoring

Code dependencies & libraries

Version control

UbiOps

/ Solving the model deployment challenge



MLOps/ModelOps 2

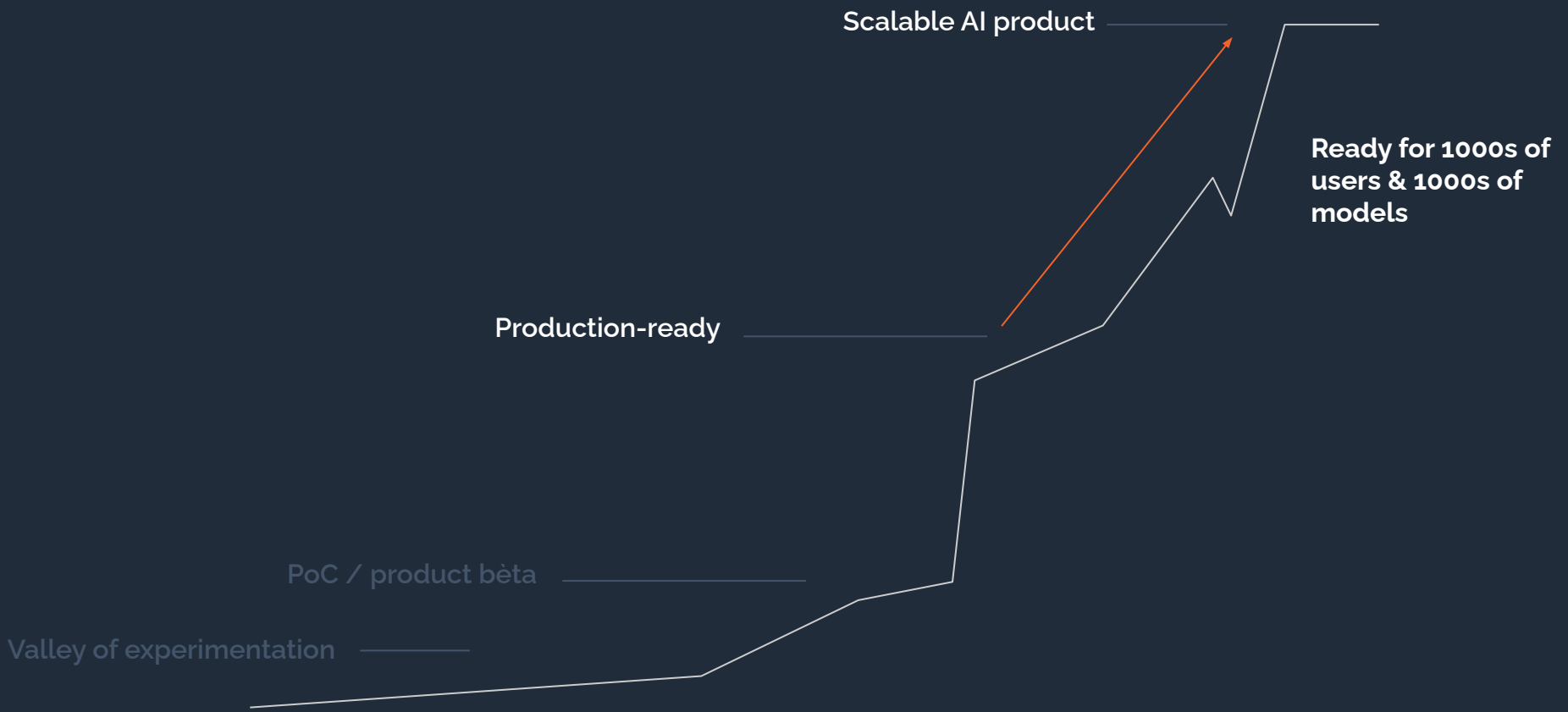
Tools, technologies, and practices to deploy, monitor, and govern ML/AI algorithms and other analytical models in production-grade applications.

1 **Data Science**

3 **Infrastructure**

UbiOps

Scaling it up (and down)



UbiOps

/ Scaling **up**

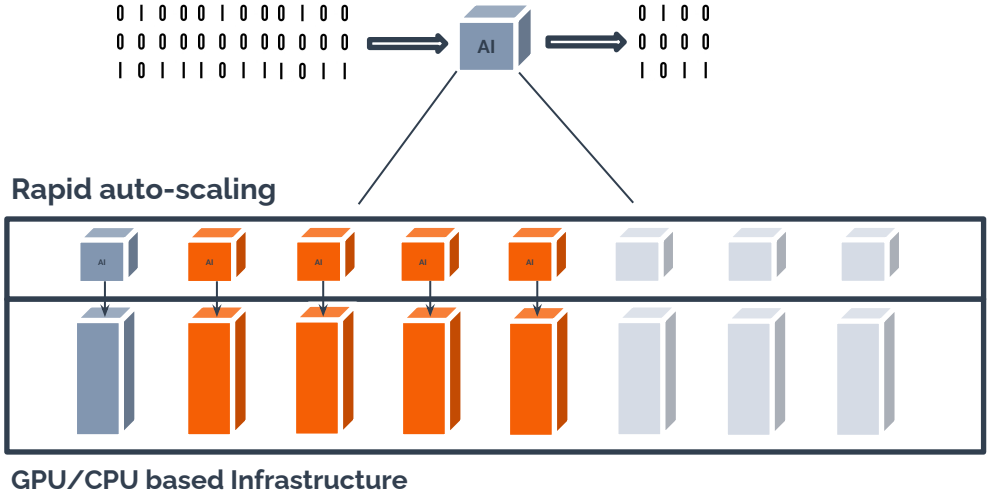
More non AI / ML related things to deal with

- Automatic scaling *("We need to be ready for all those new users")*
- Cost efficiency *("What's wrong? Ah, he saw the cloud bill…")*
- Resource availability *("What do you mean, they ran out of GPUs ?!")*
- Hybrid cloud *("This data can't move to the cloud…sorry")*

UbiOps

/ Scaling up...

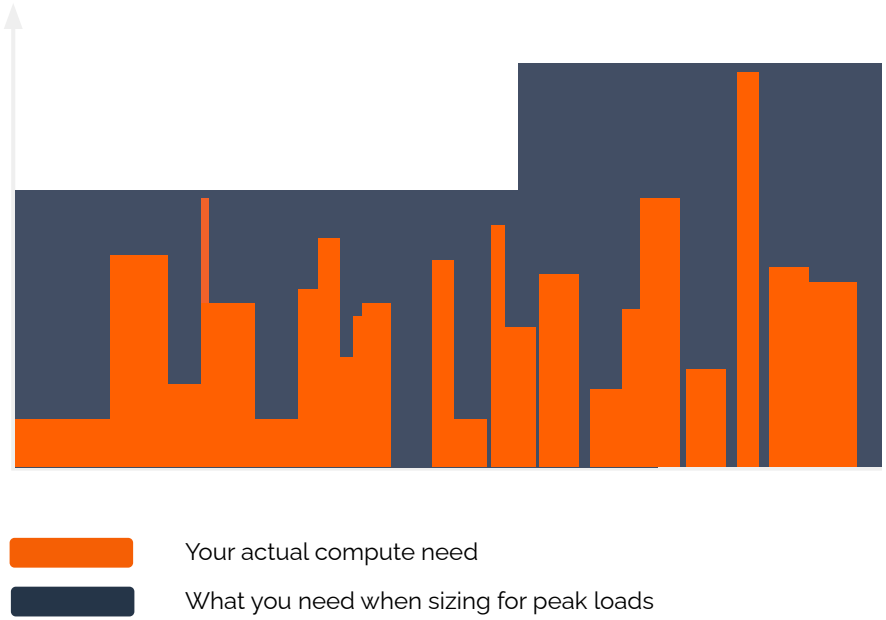
Making sure the AI service scales with demand. Ensuring it can deal with spiky usage without the whole solution going down



UbiOps

/ ...and scaling back down

Scaling up and down (back to zero) rapidly ensures you follow your compute demand curve as close as possible, optimizing resource usage

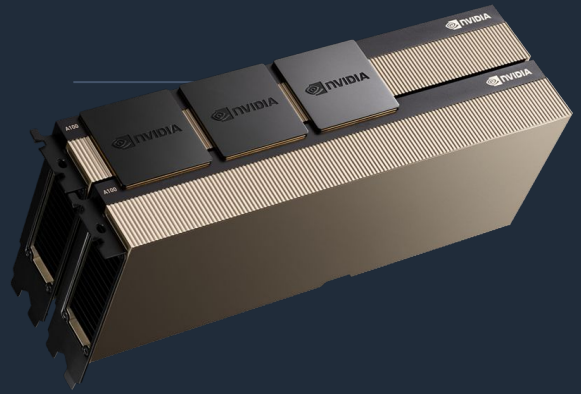


UbiOps

Resource availability

Challenges with GPUs

- Necessary for acceleration of AI models (and training)
- They're expensive
- Sharing a GPU between containers is not easy
- Issues with GPU availability, driven by supply chain issues and AI models that get heavier and heavier

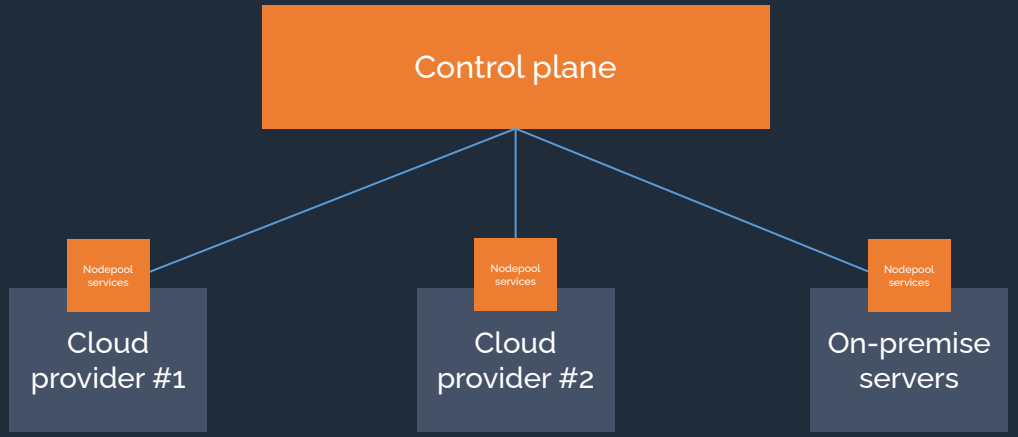


UbiOps

/ Run across environments: Hybrid & multi-cloud

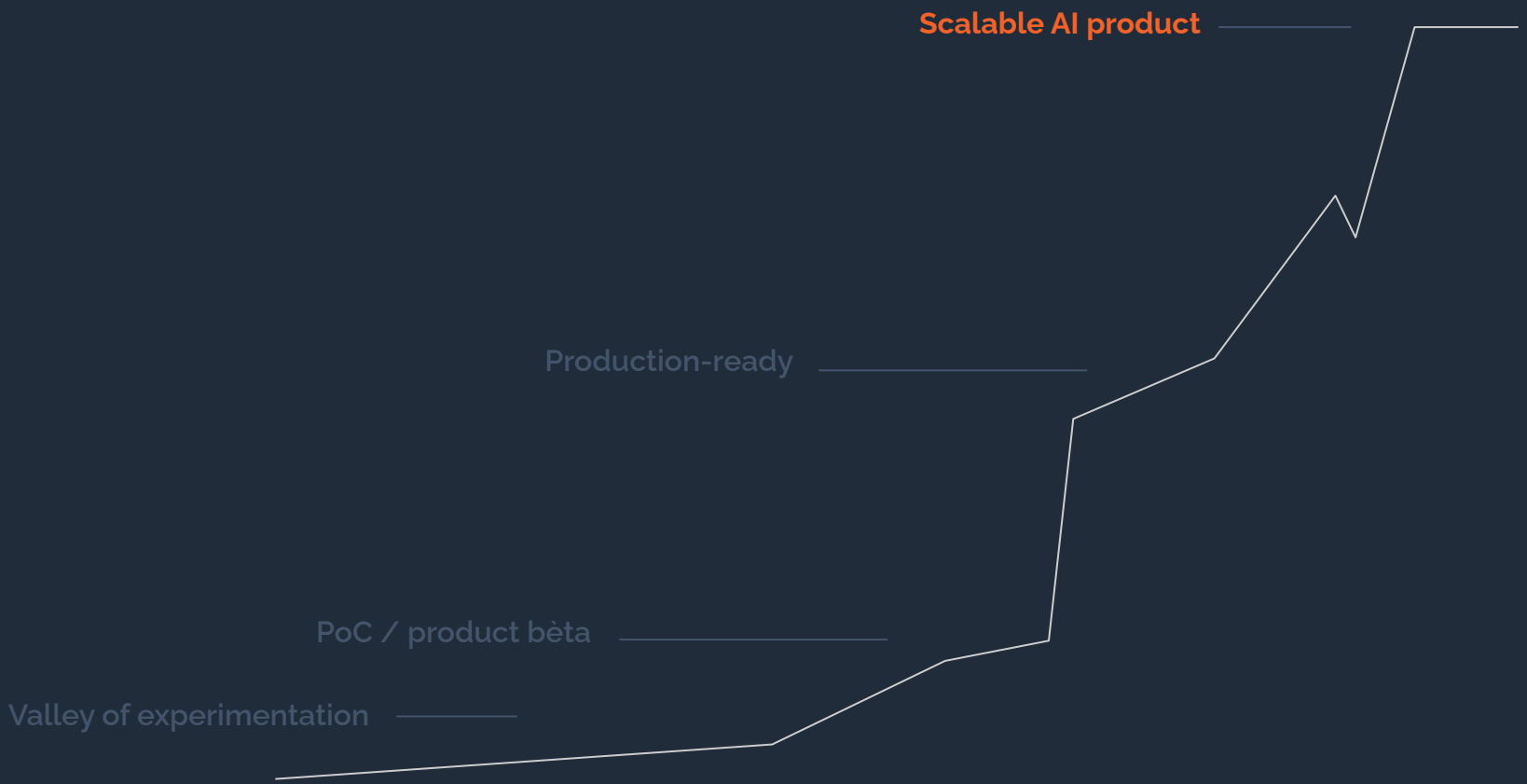
Enabling to lift and shift workloads dynamically across different environments.

- Higher availability of resources (and GPUs)
- Bring compute to where the data is.
- Solve compliance and security blockers



UbiOps

/ Reaching the summit

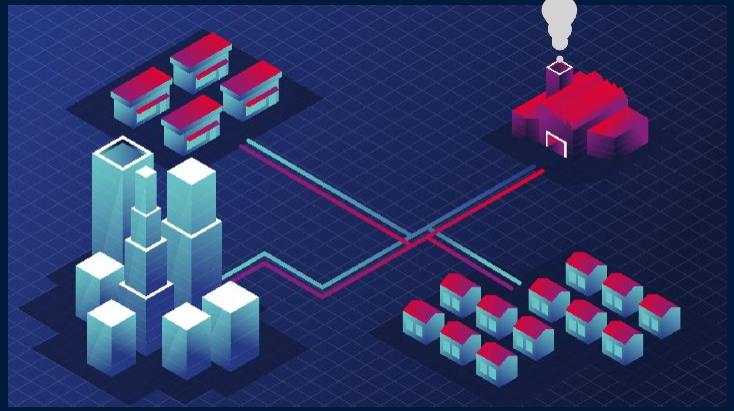


Gradyent

Usecase
Gradyent Digital Twin for heating grids

Tanja Crijns

District and industrial heating networks need to transform



High temperature & losses

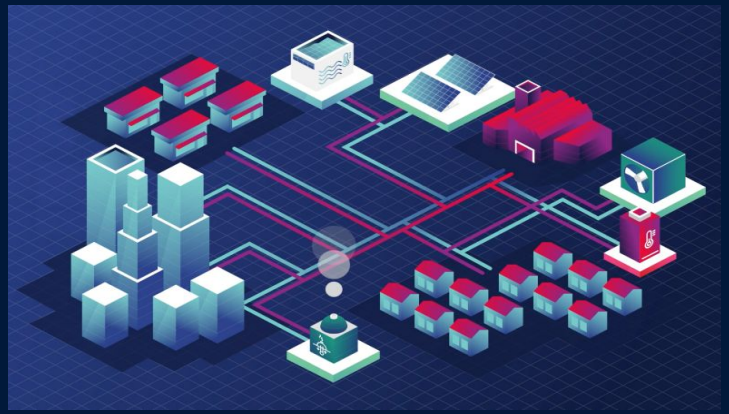
Single / few heat sources

Stand alone

Optimize

Decarbonize

Grow



Lower temperatures

Multiple, different types of heat sources

Integrated with industry, electricity & cooling

Gradyent

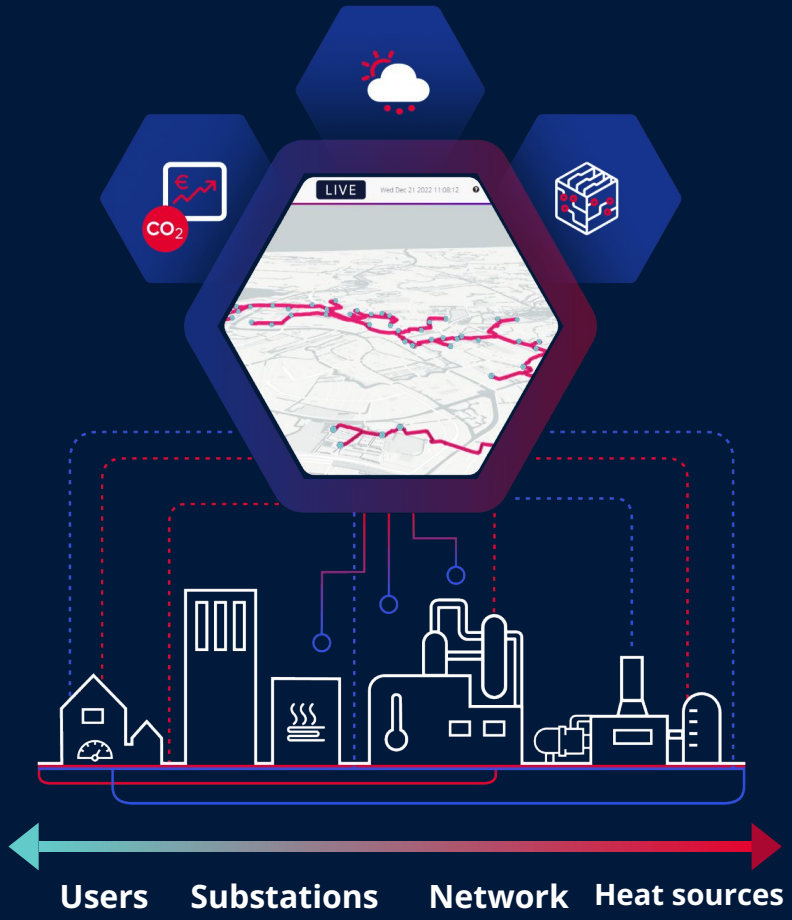
The Gradyent Digital Twin brings a new level of optimization & design by covering the entire system in real time

Real time operational optimization

By optimizing source scheduling, temperatures and pressure, debottlenecking, demand response, issue resolution

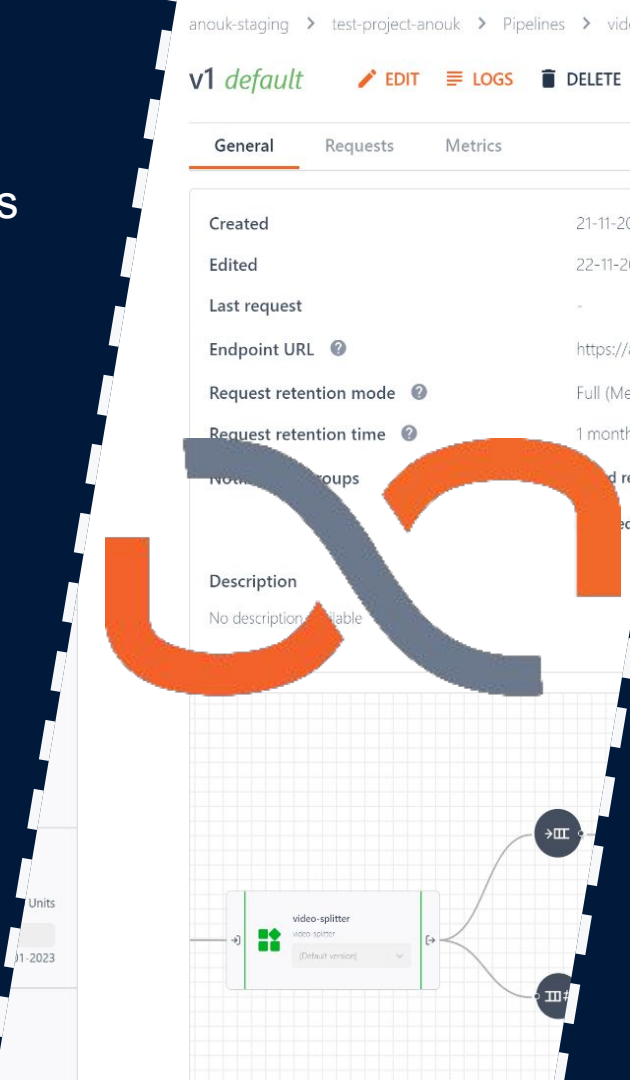
Design & simulation of future proof networks

Enhancing decision making around low carbon sources, system changes and extensions



Gradyent

UbiOps provides an abstraction layer between our engineers and the cloud



Gradyent

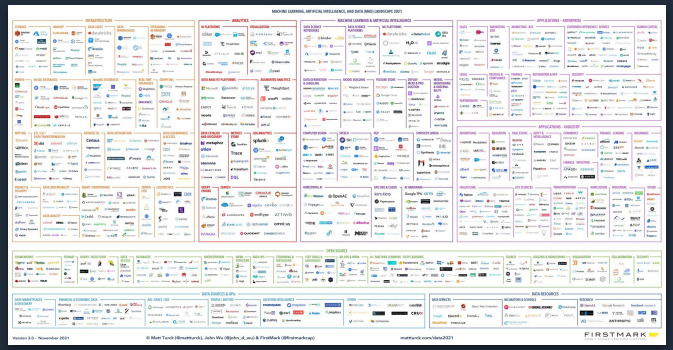
Gradyent - Five key benefits

- Ease of use
- Fast paced development
- Production grade stability
- Security and compliance
- Clear insight in compute usage

Gradyent

UbiOps

Fast track to productized AI



Scalable AI product

Production-ready

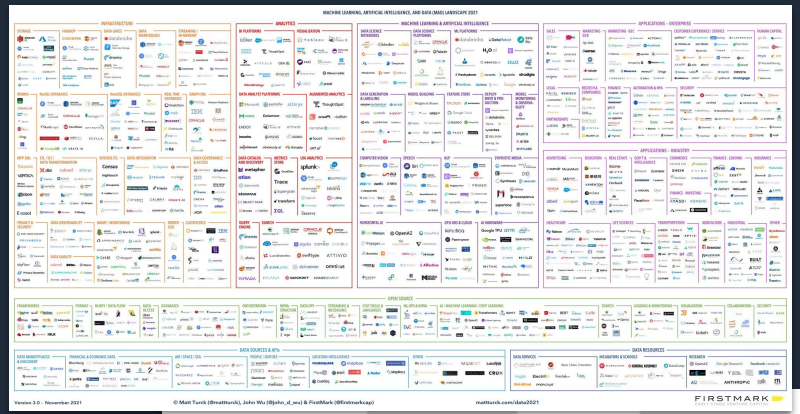
PoC / product bèta

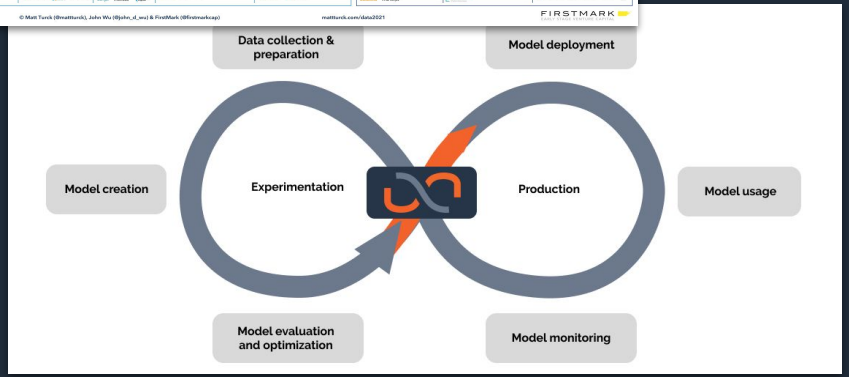
Valley of experimentation

- Short time to market
- Lower costs
- Increased reliability and maintainability of a solution

UbiOps

/ The future for AI products





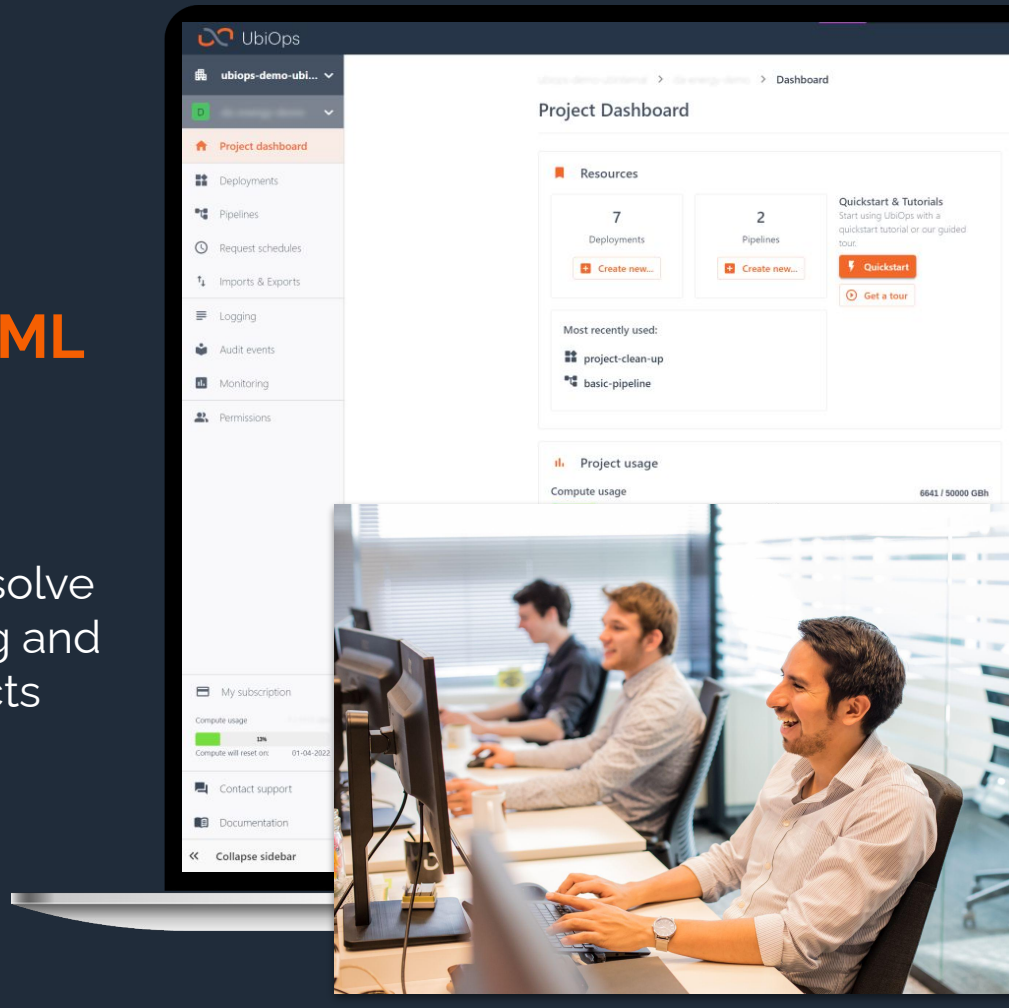
- A **new reality** where implementation of AI has a significant impact on competitiveness, market share and profitability.
- The MLOps ecosystem is expanding rapidly
- Teams will start picking their own preferred stack of tools
- Inference will become increasingly important
- Foundation models will drive product creation, but also hardware and accelerator needs

UbiOps

UbiOps

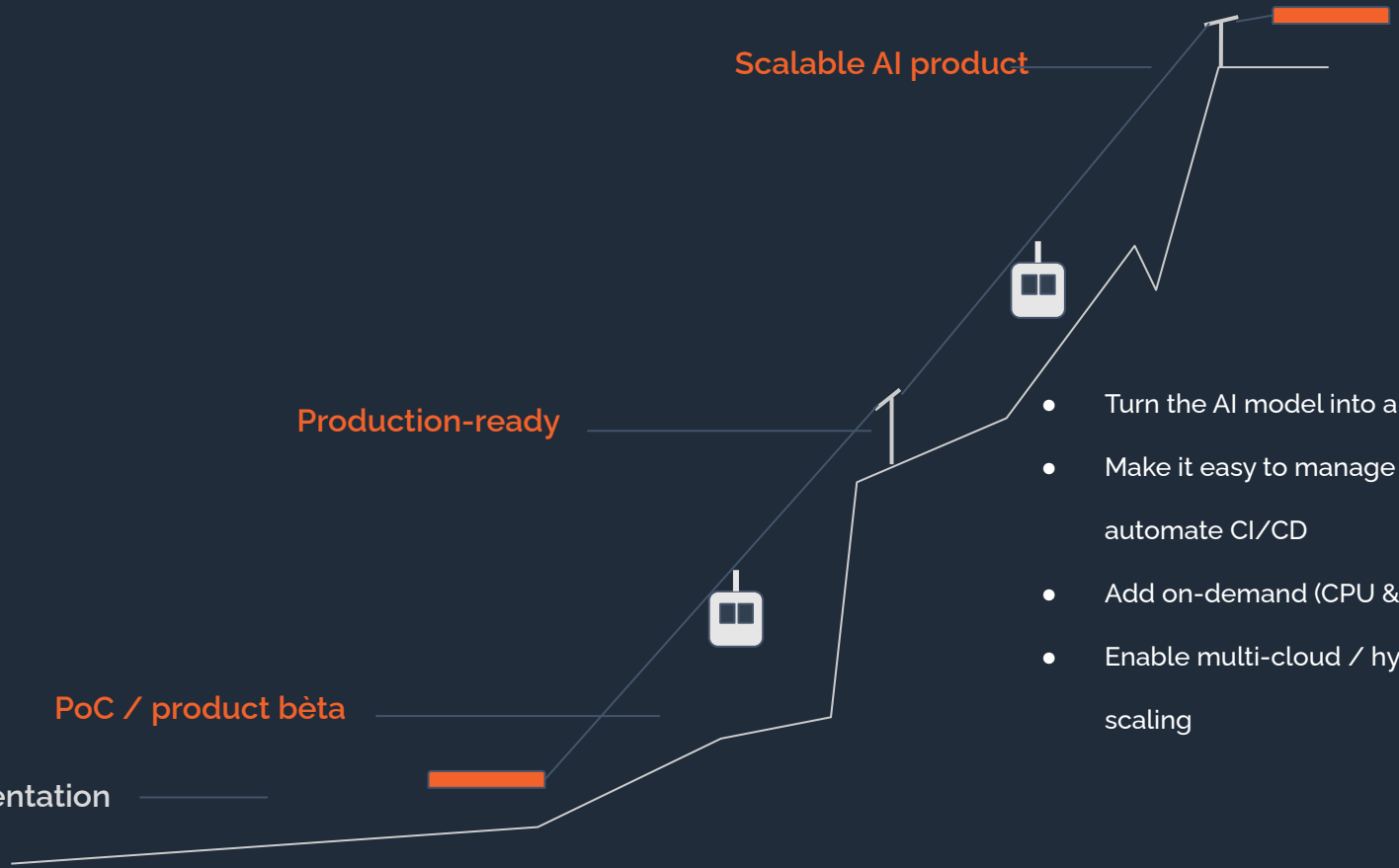
Build production-grade AI / ML products.

Our mission is to help teams of all sizes solve the technical challenges around building and running the next generation of AI products and services



UbiOps

/ Fast-track to production ready AI



- Turn the AI model into a microservice
- Make it easy to manage its lifecycle and automate CI/CD
- Add on-demand (CPU & GPU) compute
- Enable multi-cloud / hybrid-cloud scaling

UbiOps

WSAI Montréal 2023

Thank you!

Victor Pereboom - CTO @ UbiOps
victor@ubiops.com | visit **ubiops.com**

Tanja Crijns - Software Engineer @ Gradyent
tanja@gradyent.ai

Meet us at booth A70