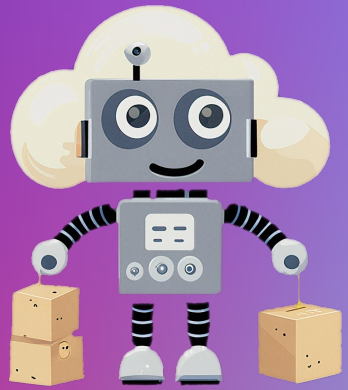
KEY001

Generative AI

What we have learned so far…

Slavik Dimitrovich

Head of AI/ML Specialist
Solutions Architects

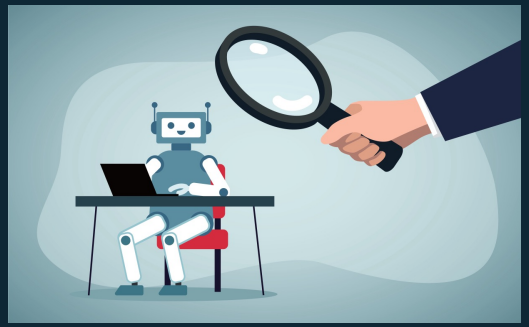


Why we should be skeptical of Generative AI

Why you should be skeptical of Generative AI

Why you should be skeptical of Generative AI

Why you should be skeptical of Generative AI



Emerging risks and challenges with generative AI

Veracity (e.g., hallucinations)

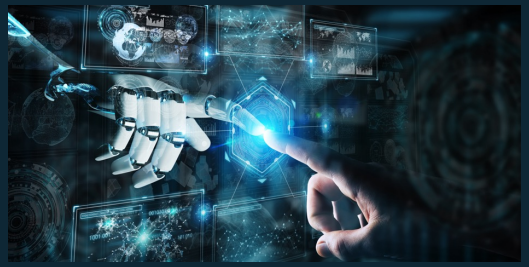
Toxicity & Safety

Intellectual property

Data privacy

Why we should be excited about Generative AI

New Frontiers



New Frontiers

New Frontiers

Generative AI has potential to create significant business value

NEW EXPERIENCES

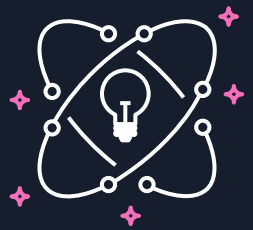
Create new innovative and engaging ways of interacting with your customers and employees

PRODUCTIVITY

Radically improve productivity across all lines of business

INSIGHTS

Extract insights and clear answers from all your corporate information, enabling faster and better decisions



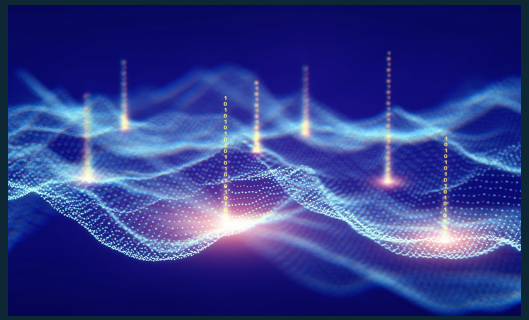
CREATIVITY

Create new content and ideas, including conversations, stories, images, videos, and music

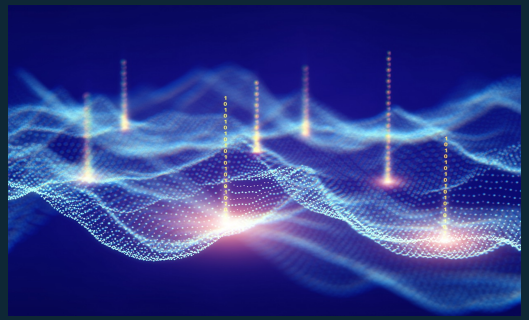
Democratization of expertise & Intelligence augmentation

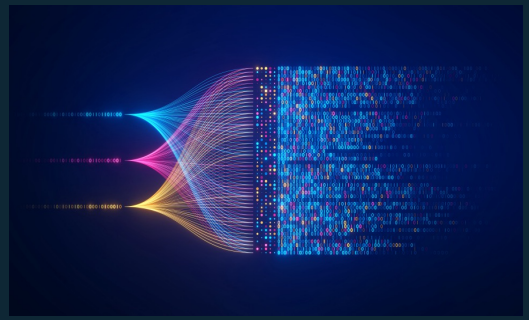
Why you may be skeptical of Generative AI with AWS

Why you may be skeptical of Generative AI with AWS

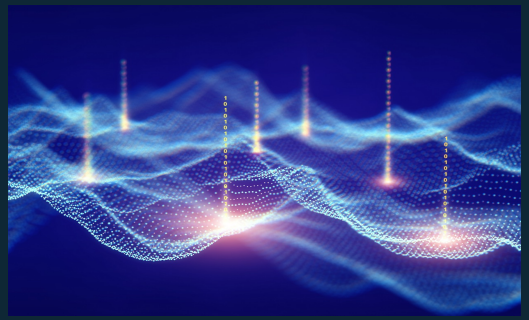


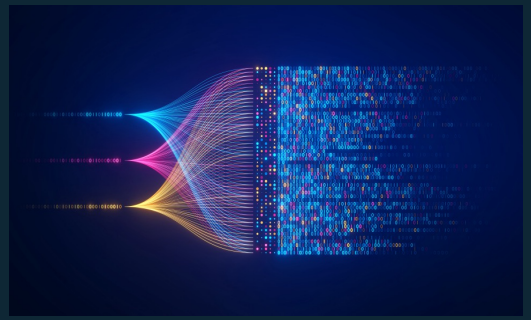
Why you may be skeptical of Generative AI with AWS

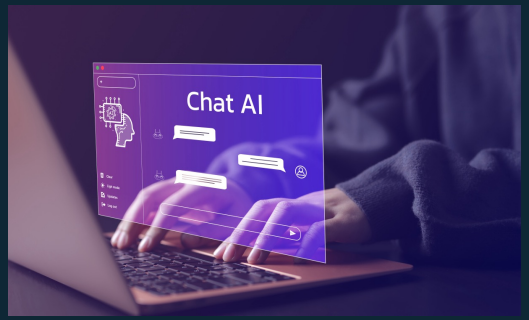




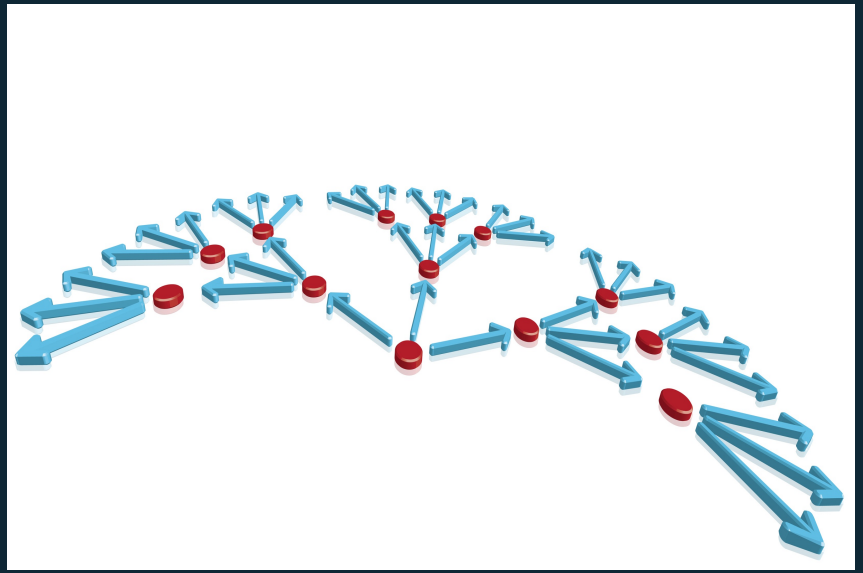
Why you may be skeptical of Generative AI with AWS

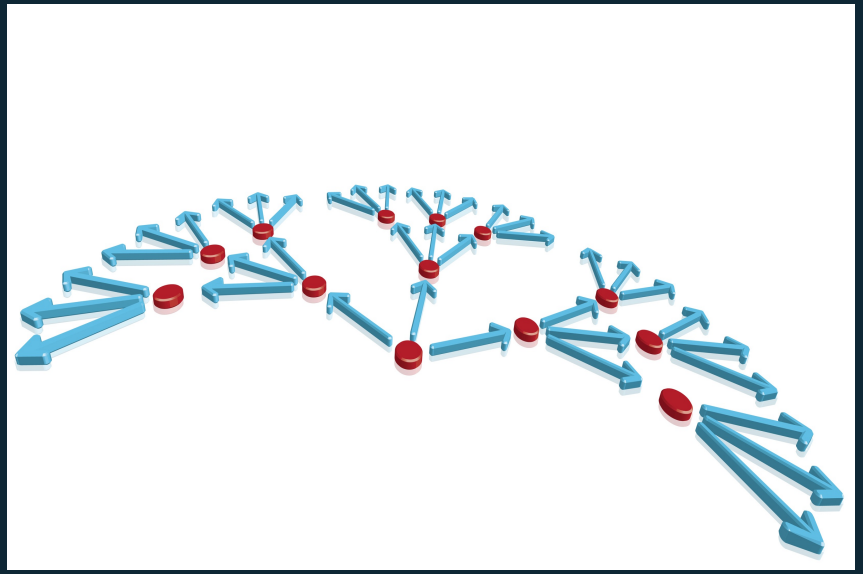




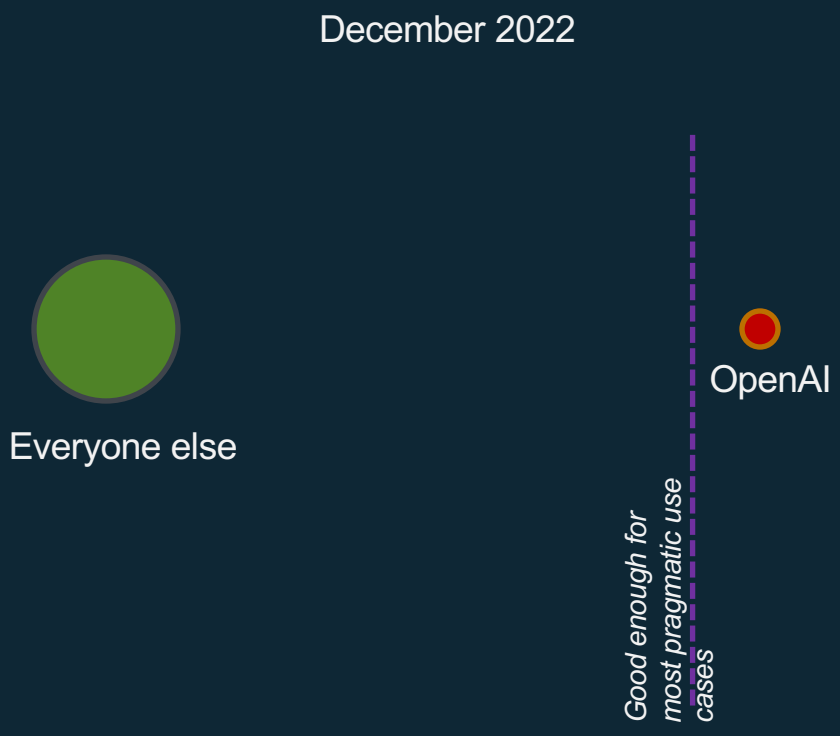


Why this does not matter

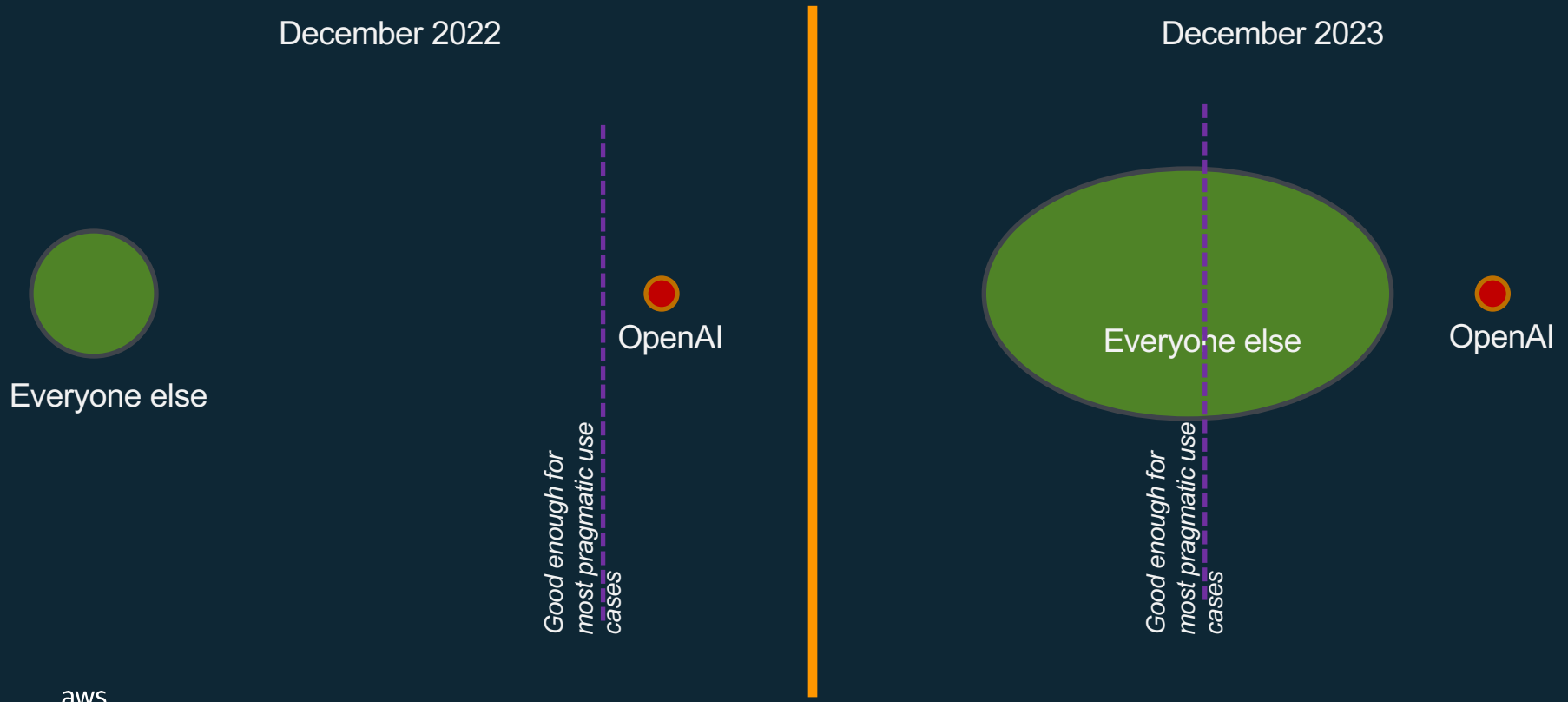




Competitive Landscape in Foundation Models

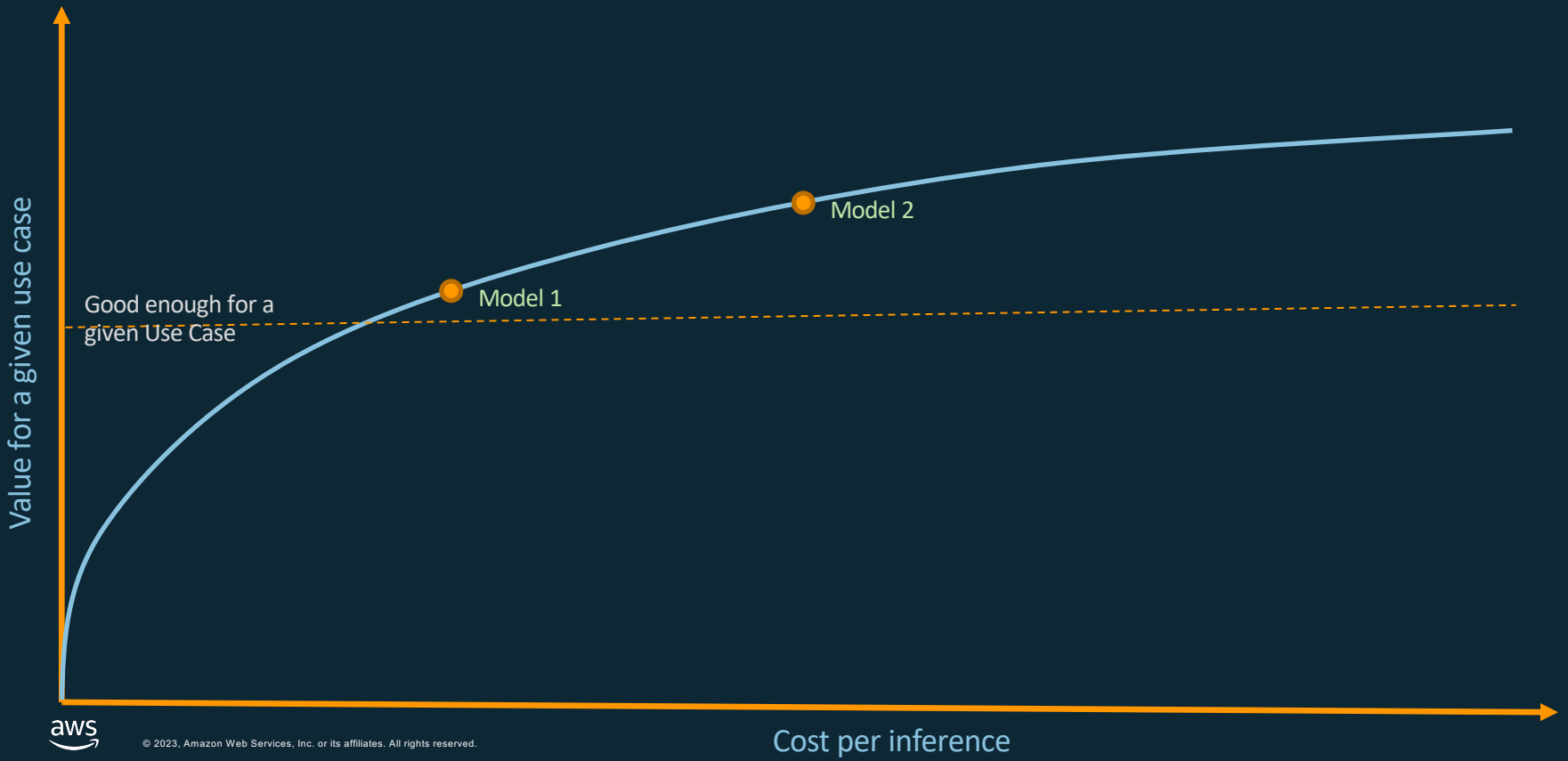


Competitive Landscape in Foundation Models

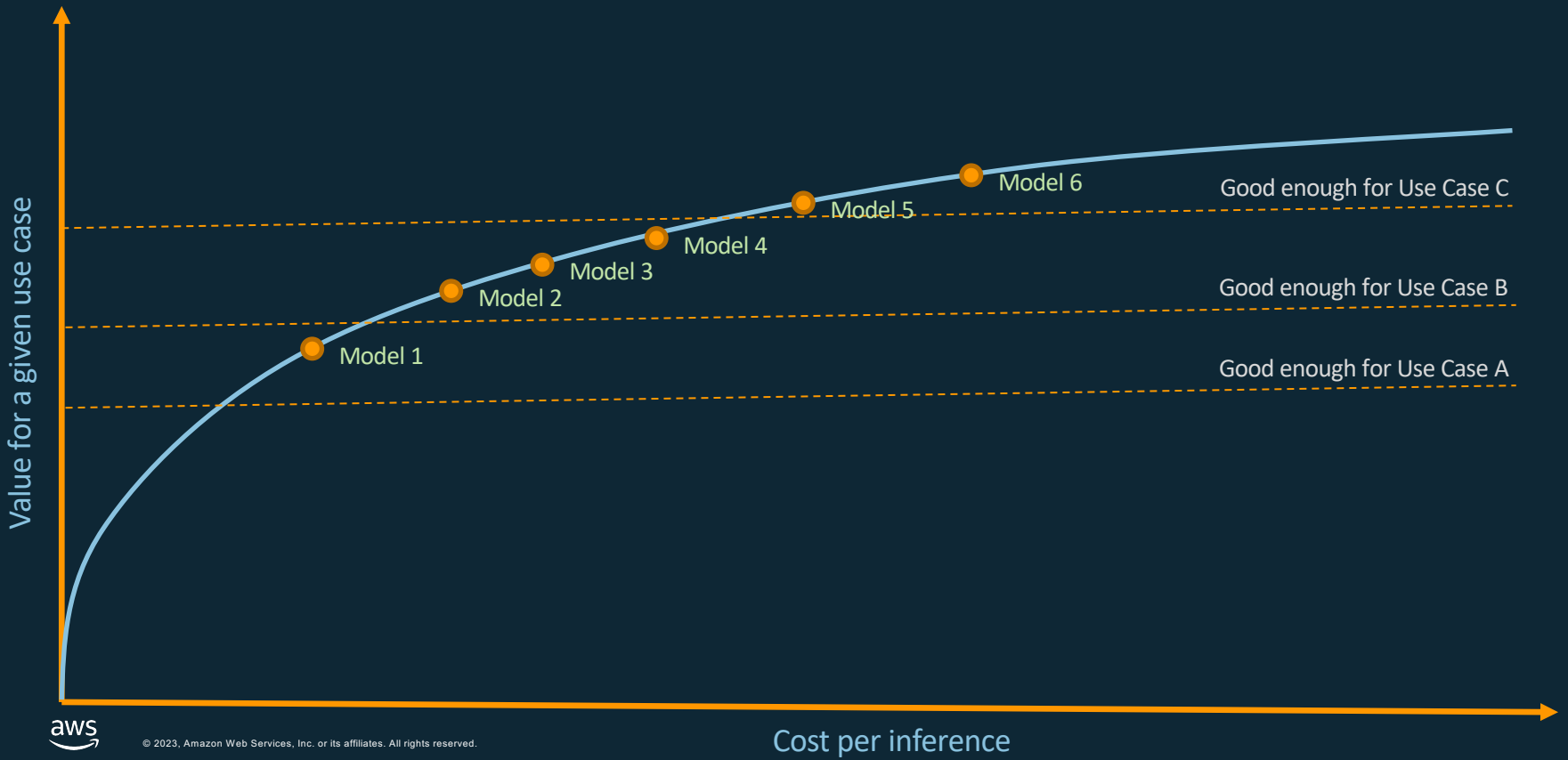


Pay attention to the trend, not a snapshot in time

Goldilocks match between model and use case

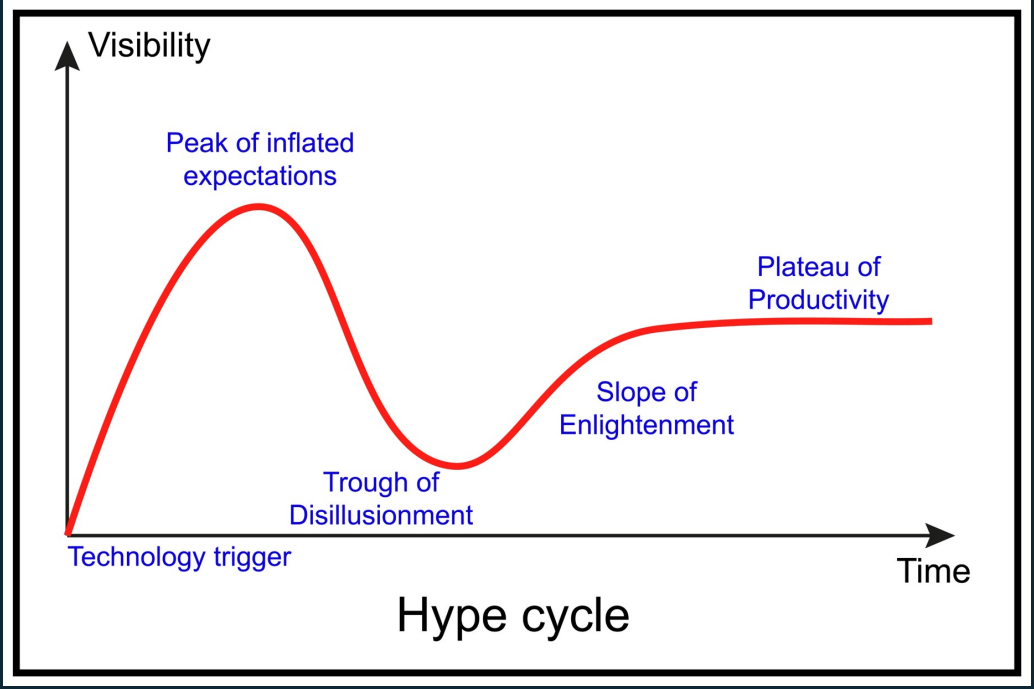


Goldilocks match between model and use case

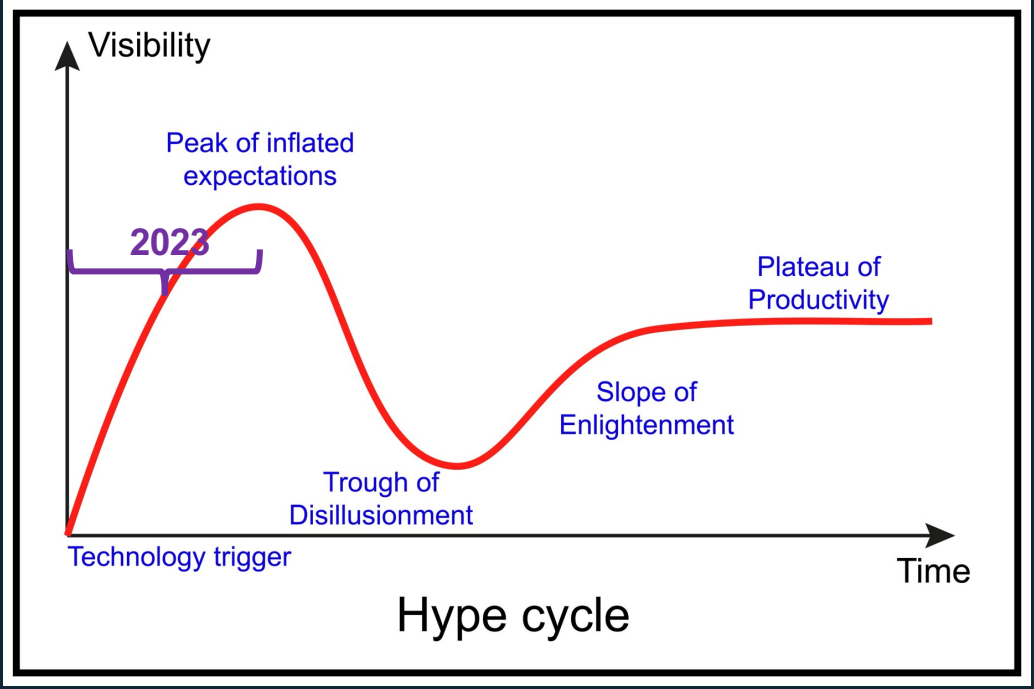


Beyond the point of “good enough” you are likely overpaying for limited additional value

Technology Hype Curve



Generative AI & the Hype Curve



2023 – the Year of POCs

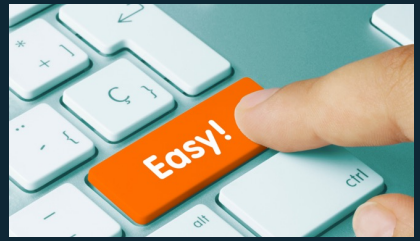
Attributes of the POC Mindset:

Attributes of the POC Mindset:

FM is the Focus

Attributes of the POC Mindset:

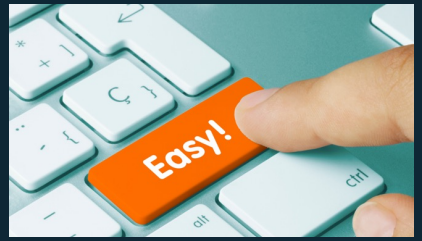
FM is the Focus



Ease of getting Outcome is Very Important

Attributes of the POC Mindset:

FM is the Focus

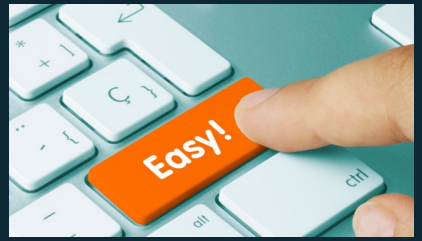


Ease of getting Outcome is Very Important

Exciting Outcome is more important than Business Value

Attributes of the POC Mindset:

FM is the Focus



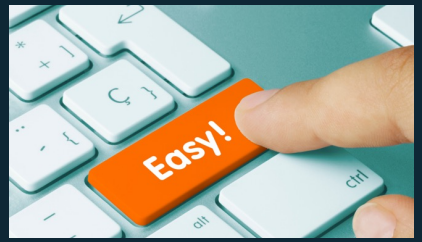
Ease of getting Outcome is Very Important

Exciting Outcome is more important than Business Value

FOMO is an Important Driver

Attributes of the POC Mindset:

FM is the Focus



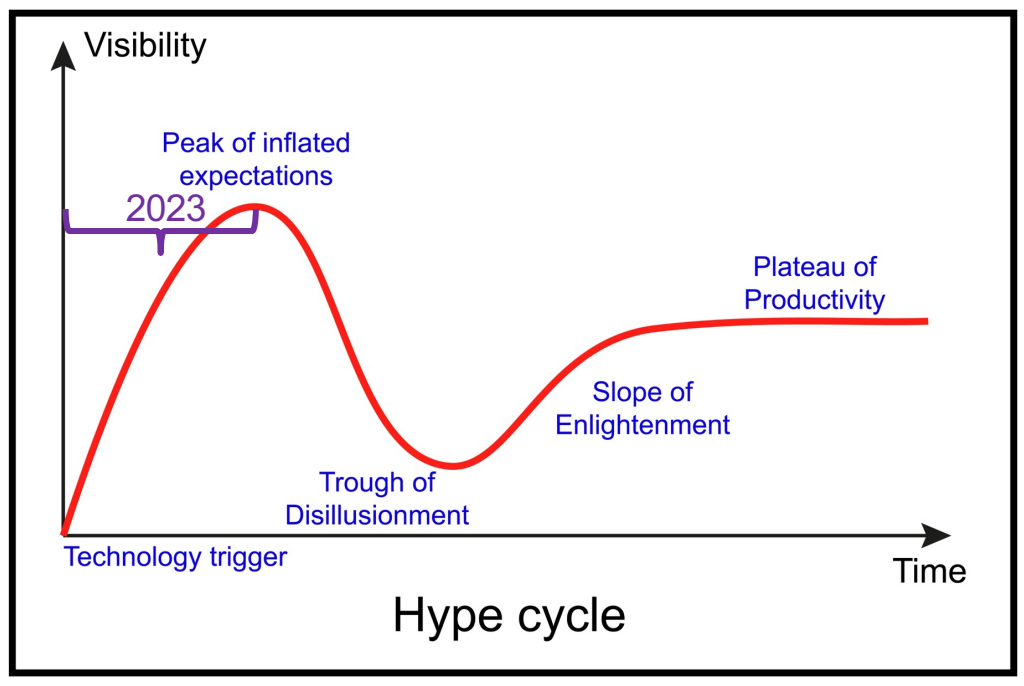
Ease of getting Outcome is Very Important

Exciting Outcome is more important than Business Value

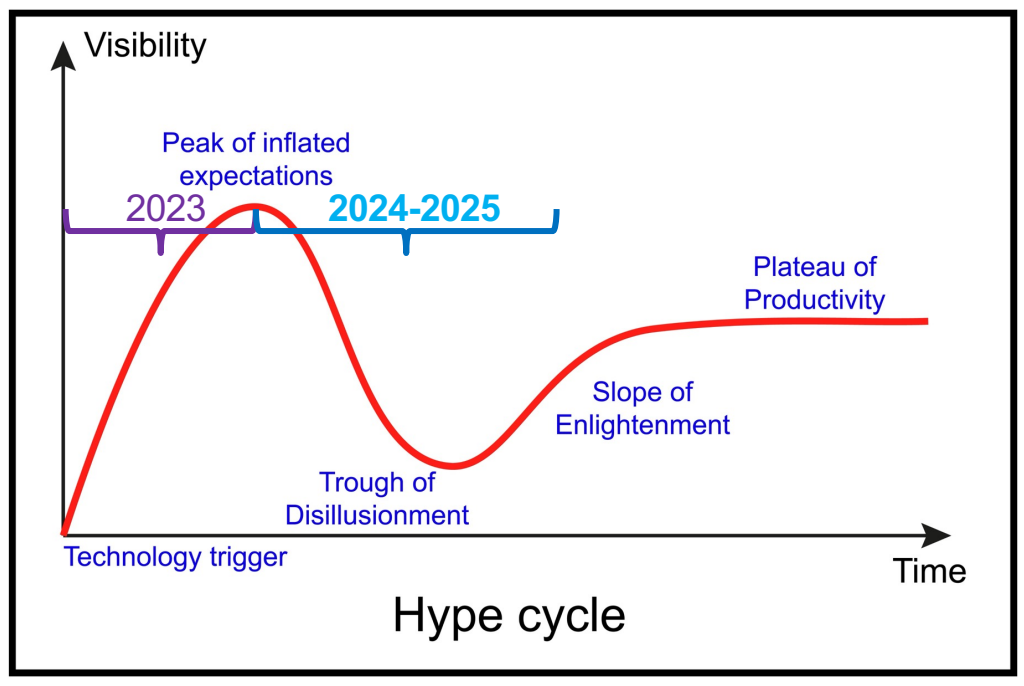
FOMO is an Important Driver

Lack of attention to Risks and Costs

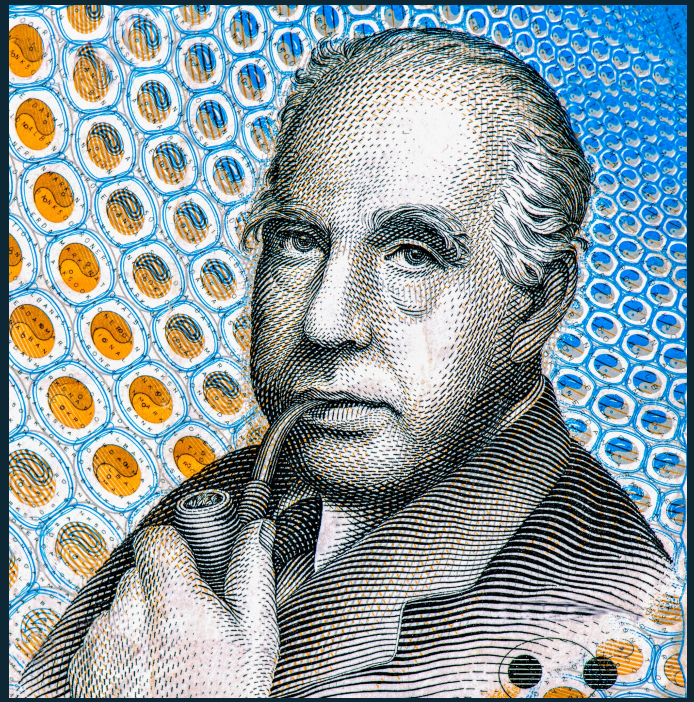
Generative AI Hype Curve



Generative AI Hype Curve



Prediction is very difficult, especially if it’s about the future



Niels Bohr, the Nobel laureate in Physics and father of the atomic model

Trends expected in 2024

Industry- and Domain-specific use cases

Trends expected in 2024

Industry- and Domain-specific use cases

Democratization of tech for building / tuning FMs

Trends expected in 2024

Industry- and Domain-specific use cases

Democratization of tech for building / tuning FMs

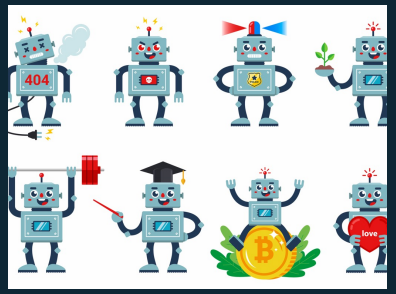
Proprietary valuable data will grow in importance

Trends expected in 2024

Industry- and Domain-specific use cases

Democratization of tech for building / tuning FMs

Proprietary valuable data will grow in importance



Specialized Models, including new modalities

2024 – the year of Production

(for some)

Challenges on the way to production

Challenges on the way to production

Security, Compliance,
Regulation, Brand

Challenges on the way to production

Security, Compliance, Regulation, Brand

Difficulties quantifying business value

Challenges on the way to production

Security, Compliance, Regulation, Brand

Difficulties quantifying business value

Unit economics

Challenges on the way to production

Security, Compliance, Regulation, Brand

Difficulties quantifying business value

Unit economics

Difficulties operationalizing GenAI systems

Challenges on the way to production

Security, Compliance, Regulation, Brand

Difficulties quantifying business value

Unit economics

Difficulties operationalizing GenAI systems

Skills shortage

Attributes of the production mindset

Attributes of the Production Mindset:

FM as part of a larger system

Attributes of the Production Mindset:

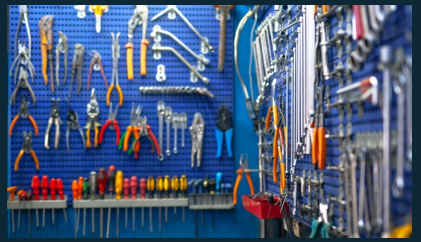
FM as part of a larger system

Cost and Latency become more important

Attributes of the Production Mindset:

FM as part of a larger system

Cost and Latency become more important

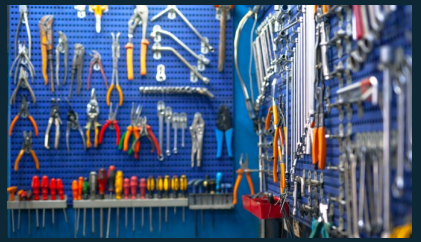


Choose model optimized for a use case

Attributes of the Production Mindset:

FM as part of a larger system

Cost and Latency become more important



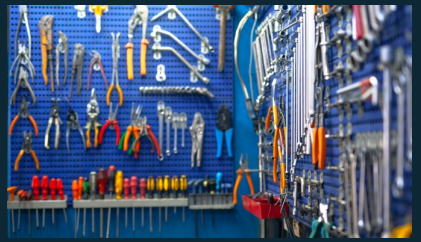
Choose model optimized for a use case

Security, Privacy, Compliance, Responsible / Ethical AI

Attributes of the Production Mindset:

FM as part of a larger system

Cost and Latency become more important



Choice of models optimized for use cases

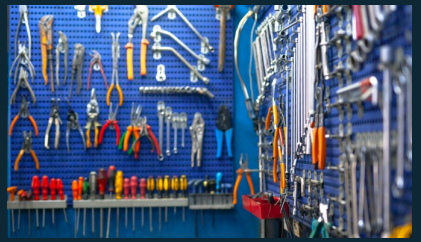
Security, Privacy, Compliance, Responsible / Ethical AI

Tangible business value is the main driver

Attributes of the Production Mindset:

FM as part of a larger system

Cost and Latency become more important



Choice of models optimized for use cases

Security, Privacy, Compliance, Responsible / Ethical AI

Tangible business value is the main driver

Flexibility in consumption models

The power of choice by design

Gen AI-powered Users and Developers

Model Consumers and Tuners

Model Builders and FM Providers

Gen AI-powered Users and Developers

Model Consumers and Tuners

Model Builders and FM Providers

Infrastructure for FM training and inference

 GPUs  Trainium  Inferentia  SageMaker

 UltraClusters  EFA  EC2 Capacity Blocks  Nitro  Neuron

Gen AI-powered Users and Developers

Model Consumers and Tuners

Tools to build with LLMs and other FMs

Amazon Bedrock

Guardrails | Agents | Customization Capabilities

Model Builders and FM Providers

Infrastructure for FM training and inference

GPUs

Trainium

Inferentia

SageMaker

UltraClusters

EFA

EC2 Capacity Blocks

Nitro

Neuron

Gen AI-powered Users and Developers

Applications that transform experiences

Amazon Q | Amazon Q in Amazon QuickSight | Amazon Q in Amazon Connect

Model Consumers and Tuners

Tools to build with LLMs and other FMs

Amazon Bedrock

Guardrails | Agents | Customization Capabilities

Model Builders and FM Providers

Infrastructure for FM training and inference

GPUs | Trainium | Inferentia | SageMaker

UltraClusters | EFA | EC2 Capacity Blocks | Nitro | Neuron

aws

Your data is your differentiator


Generative AI Application

Generative AI Application

Storage
Structured and unstructured data

Operational databases
SQL, NoSQL, document, graph, vector

Analytics and data lakes
Search, streaming, batch, interactive

Data integration
Capture, transformation, streaming

Data governance
Catalog, quality, privacy, access controls

aws

Path to meaningful AI transformation for most enterprises is not through a single magical “killer use case”, but rather by achieving a snowball effect of Generative AI-powered use cases becoming the very fabric of how they operate

Key takeaways & next steps

Key Takeaways

- Now is the time to get tangible business value out of Generative AI

Key Takeaways

- Now is the time to get tangible business value out of Generative AI
- Adopt the Production Mindset

Key Takeaways

- Now is the time to get tangible business value out of Generative AI
- Adopt the Production Mindset
- Start with low-hanging fruit and go for the snowball effect

Key Takeaways

- Now is the time to get tangible business value out of Generative AI
- Adopt the Production Mindset
- Start with low-hanging fruit and go for the snowball effect
- Look at the trendline, not a snapshot in time

Key Takeaways

- Now is the time to get tangible business value out of Generative AI
- Adopt the Production Mindset
- Start with low-hanging fruit and go for the snowball effect
- Look at the trendline, not a snapshot in time
- Partner with technology providers who provide you choice of models and flexibility of engaging with GenAI technology across all levels of the stack

Next steps

- If you haven’t explored this space yet, let’s set up a discovery workshop to help you identify and prioritize potential use cases

Next steps

- Let’s explore together!
- If you have done exploration and looking for the first use cases to take to production, let’s work on a Prototype / Pilot to connect all the pieces together and evaluate cost/benefits at a deeper level

Next steps

- Let’s explore together!
- Let’s test the assumptions!
- If you are ready to take your first use case(s) to production, let’s discuss the delivery strategy: we have rich and growing partner ecosystem as well as our own Professional Services to help with that

Next steps

- Let’s explore together!
- Let’s test the assumptions!
- Let’s launch in productin and realize business value!

Thank you!

Slavik Dimitrovich

slavik@amazon.com

Please complete the session survey.