Petite SRE：

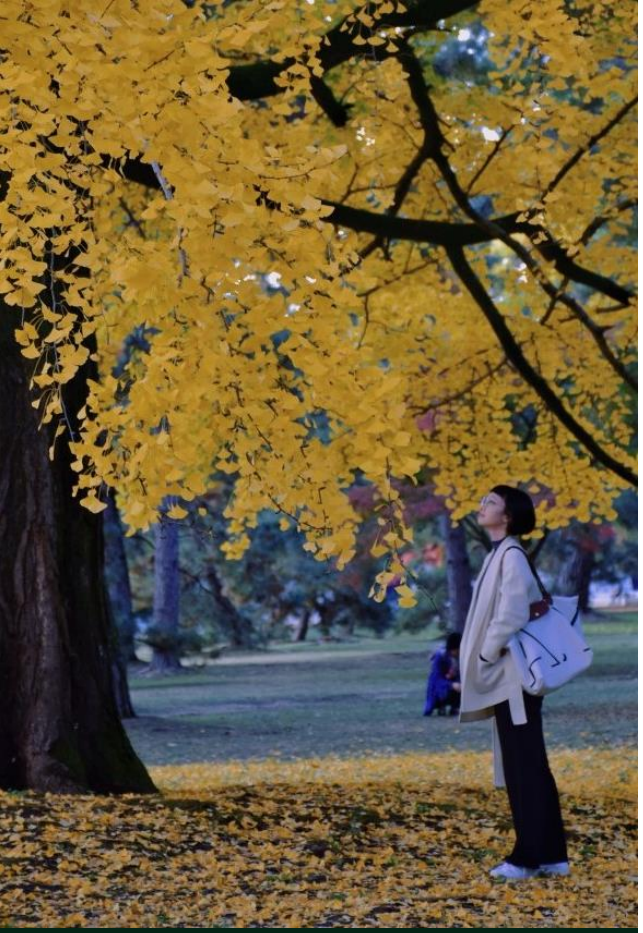
GenAI Eraにおけるインフラのあり方観察

キンドリルジャパン株式会社

叶　奕池（ICHICHI）

Infrastructure/Cloud Architect

kyndryl



自己紹介

よう　えきち
叶 奕池（ICHICHI）

キンドリルジャパン株式会社　Infrastructure/Cloud Architect

Japan AWS Junior.Champion 2023

略歴：

2024年6月ー現在　大手カード会社様　**生成AI基盤開発案件　Lead Architect, Tech Lead**

- GPU-Basedコンテナ基盤、AI/MLデータベース、自動化設計・開発
- Full Stackスクラム開発支援

2024年6月ー現在　大手自動車産業のお客様　**DX基盤開発案件　Associate Architect**

- Over 2000＋サーバを有する大規模AWS基盤設計・構築
- スクラム開発推進

2022年4月ー2024年6月　大手保険会社様　**アウトソーシング案件　AWS Team Lead, Architect**

- Over 1000＋サーバーを有する大規模ハイブリッドクラウド基盤構築・運用管理自動化設計
- 先端ソリューション開発：Outposts、オンプレミスへのAWSエージェント統合、など
- 日米協同クラウド活用推進

2021年4月ー2022年3月　IBM Cloud Advanced Customer Support　セキュリティサポートエンジニア

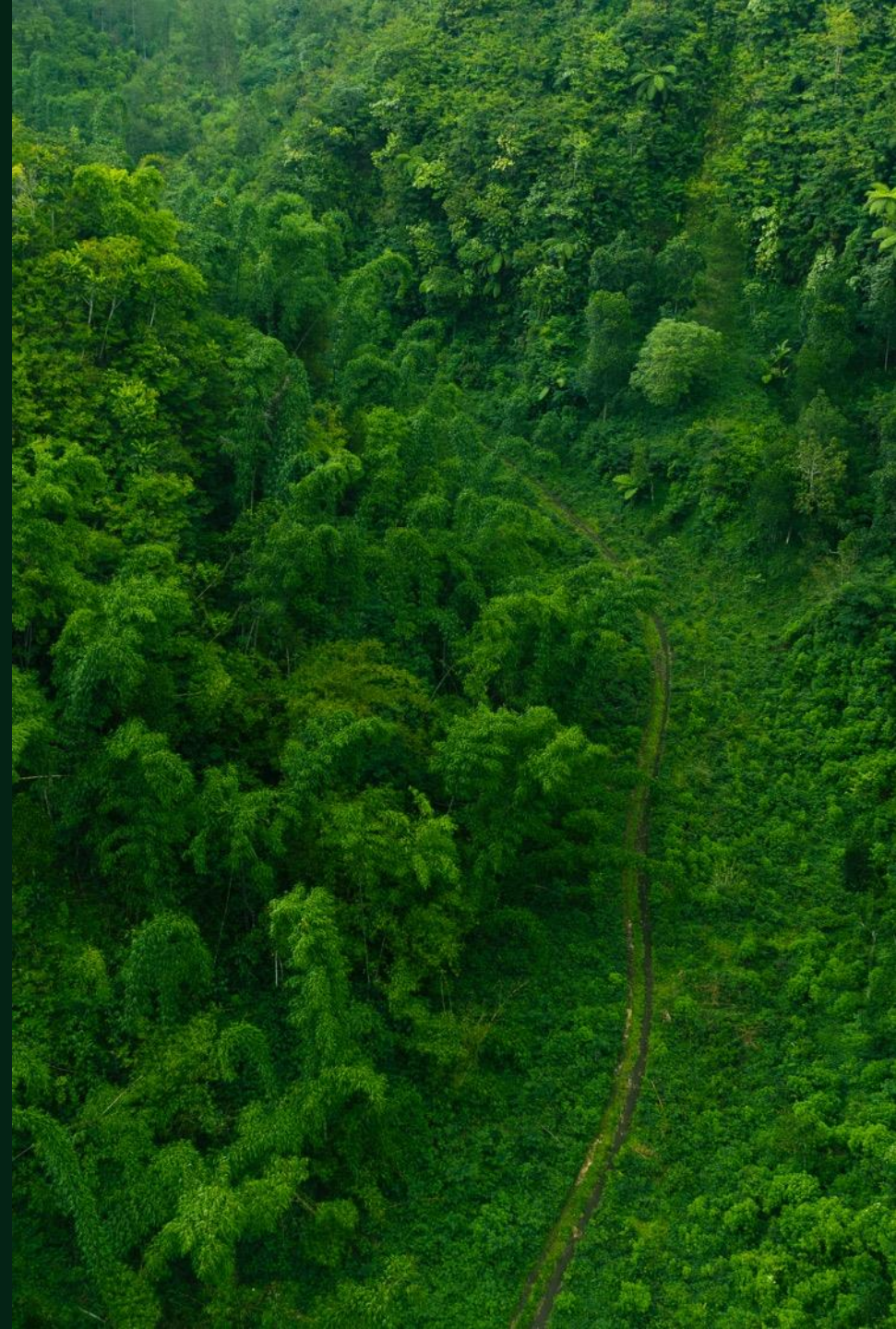
2021年4月　新卒入社

kyndryl

Agenda

1. The Edge of GenAI Research and Report
2. The Review of AWS AI Day
3. Business Case-based Modeling
    - Enterprise Solution Design/Delivery
    - Agile Product Design/Delivery
    - Team/Technical Agility
    - Continuous Input/Output
    - Lean Portfolio Management
4. Conclusion

kyndryl



1. The Edge of GenAI Research and Report

❶注目すべき機械学習モデルの開発状況：2023年

- 産業界が51モデルを開発
- 学術界が15モデルを開発
- 産学連携から21モデルが誕生

「機械学習モデルの開発は依然として、産業界が主導権を握る」

❷基盤モデル開発の加速

- 2023年の開発数は2022年の2倍以上
- 65.7%がOSSとして公開

「開発の民主化が進んでいる」

❸技術性能

- 画像分類や言語理解で人間を凌駕
- 競技数学や視覚的常識判断では未だ課題

❹責任のあるAI

- AI関連インシデント： 2023年に123件報告（前年比32.2%増）
- 社会的影響への懸念拡大

「技術革新と責任ある開発の両立が急務」

kyndryl

❶Agentic AI（エージェント型AI）の台頭

- 次世代の強力な基盤モデルへの進化
- 自律的な意思決定が可能なデジタルエージェント
- 高度な推論能力を持つチャットボット・コパイロット

❷インダストリーにおけるGenAIの適用

- 消費者向けと企業向けのタスク処理は継続、主な適用分野

❸アクセス可能なAIツール

- より安価、安全なAIツールは日常生活の一部として浸透

❹AIを統合したソフトウェア開発

- AIは開発を支援する役割から、ソフトウェア自体の一部に変更

2. The Review of AWS AI Day

AI Rethink Framework

❶Process: 具体的なニーズの基にAIを活用
- 実際の業務課題やニーズからAI活用を検討
- 明確な目的意識を持ったAI導入

❷Formulation: AIで解決すべき問題の明確化
- 課題の具体的な定義づけ
- 解決したい問題の詳細を記述

❸Tools and Technologies: AIの提供する可能性を評価
- 利用可能なAIツールの機能を評価
- 技術的な実現可能性を検討

❹Data: 適切な情報に基づくAIの活用
- 必要なデータの特定と収集
- データの品質と適合性の確認

❺Context: 実践におけるAI活用評価基準の形成
- 実際の業務環境での適用検討
- メリットとリスクを軸に評価

Infrastructure and SRE in GenAI Era

❶アプリケーション開発と統合
- 開発パターンと関連する機能構造を理解
- 階層化・構造化されたリソース割り当て戦略を策定

❷インフラストラクチャ設計/デリバリーモデル
- プロダクト/レイヤーベースではなく、機能/サービスベースのRACI策定
- 変化するサービスメッシュの要件を対応可能に

❸SRE（Site Reliability Engineering）の適用
- 自動化されたデプロイメントと改善パイプラインを確立
- 早期段階からのObservabilityを実装

GenAI Eraのインフラ/クラウド設計/デリバリ現場において、SREが中心的な役割を果たす

kyndryl

3. Business Case-based Modeling：Process View

Enterprise Solution Design/Delivery

❶Complete Cloud Native
- クラウド環境に最適化された新しいアーキテクチャへの完全移行

❷Semi-Cloud Native
- クラウド上にあるシステムと密接に連携するものが対象
- オンプレミスに残しつつ、AWS AgentやAWS Outpostsを導入

❸On-premises Remains
- 現時点移行不可なシステム

Agile Product Design/Delivery

❶Individual Service-based Modules
- サービスの単体デプロイまたは導入検証のために利用
- 例：環境定義スクリプトが組み込んだEC2起動テンプレート

❷Technical Solution-based Modules
- 日常的なシステム運用から生まれたソリューションモジュール
- 例：障害対応訓練のためのAWS Fault Injection Simulatorテンプレート

❸Industry Business Case-based Modules
- 特定のインダストリーの特徴を考慮したビジネスケースモジュール
- 例：BSEA for FSI

❹Observability Modules
- 上記のモジュールとセットにする監視モジュール

Team/Technical Agility

❶タスク管理の高度化
- 各タスクの目標、成果物、期限を明確に定義し共有
- パイプライン式スケジューリング

❷多層的なチーム戦略
- L1チーム（基本デリバリ）
- L2チーム（高度デリバリ）
- L3チーム（先端デリバリ）

❸AI統合サービスの活用

Continuous Input/Output

Lean Portfolio Management

kyndryl

3. Business Case-based Modeling：Stack View

Lean Portfolio Management

Continuous Input/Output
- Automation
- Agile
- Observability

Agile Product Design/Delivery
- Observability Modules

Agile Product Design/Delivery
- Industry Business Case-based Modules

Fast responses to customer inquires

Social Media Monitoring of customer sentiments

Chatbots for user queries and products

Customer retention and user experience

L1 / L2 Help Desk Support

Faster Root Cause Analysis of IT Incidents

Triaging Service request

IT Incident Management

Team/Technical Agility

LangChain

OpenAI

Agile Product Design/Delivery
- Individual Service-based Modules
- Technical Solution-based Modules

kyndryl

Conclusion

Key Takeaways

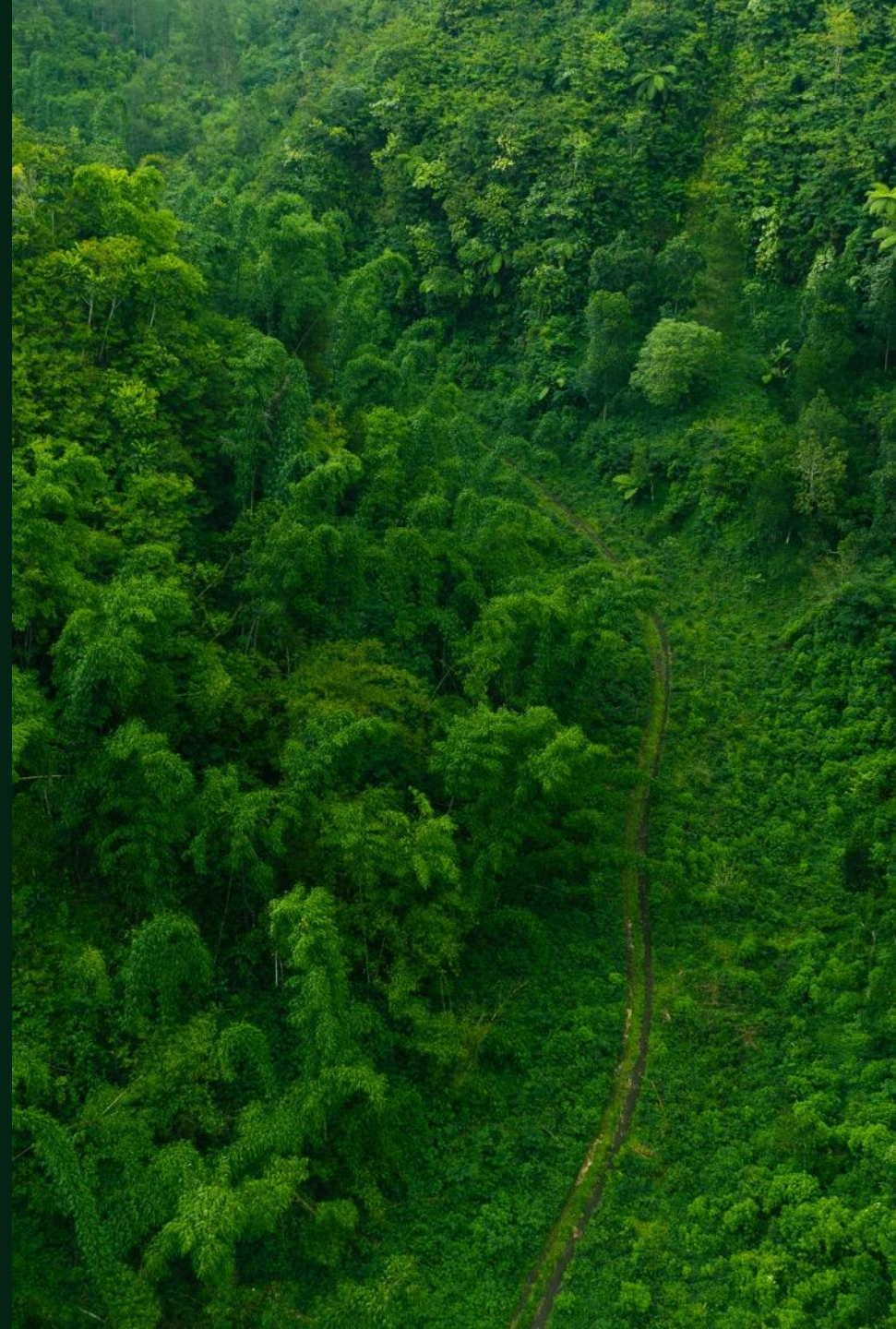
1.GenAI時代におけるアドバンスドインフラデザイン/デリバリの成功要因：

- SREの原則と組み合わせたバランスの取れたアプローチ
- ビジネス実践から生まれた明確で実用性のあるモデリングフレームワーク
- 継続的な改善サイクル

2.実用的なAIインフラストラクチャは、まだ発展途上の段階にあり、

現行のクラウドインフラストラクチャとの主な相違点：

- 創発的な知能能力
- 自己組織化システム
- 最小限の人的介入
- 高度なメタ学習フレームワーク



kyndryl

Conclusion

Future Tasks

1.AI駆動の自動化：

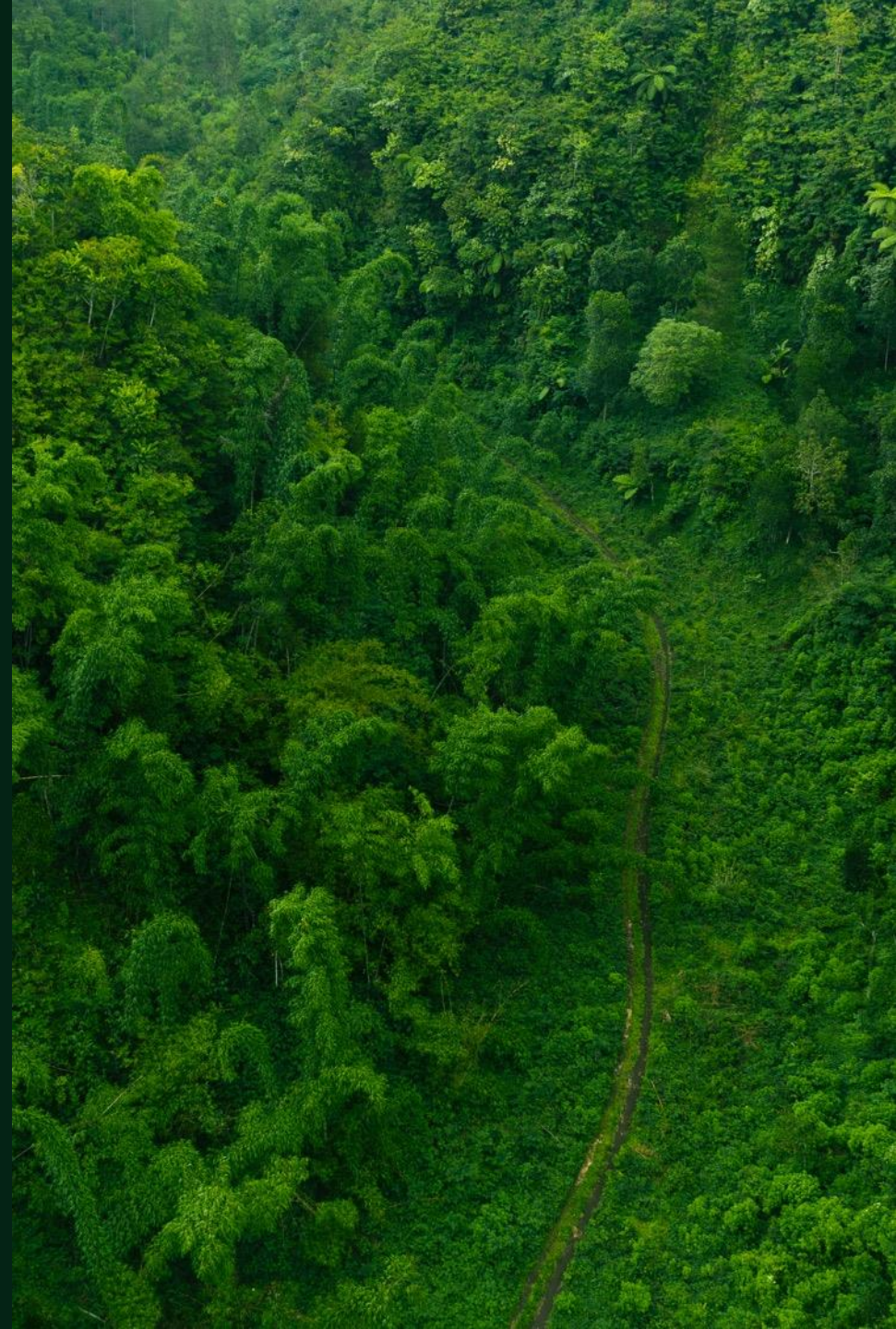
- ルーチン作業の自動化から複雑な意思決定の自動化へ
- より知的な自動化システムの開発

2.高度な可観測性の実装：

- メトリクス、ログ、トレースの統合的な監視
- AIを活用した異常検知と分析

3.自己回復可能なインフラストラクチャー：

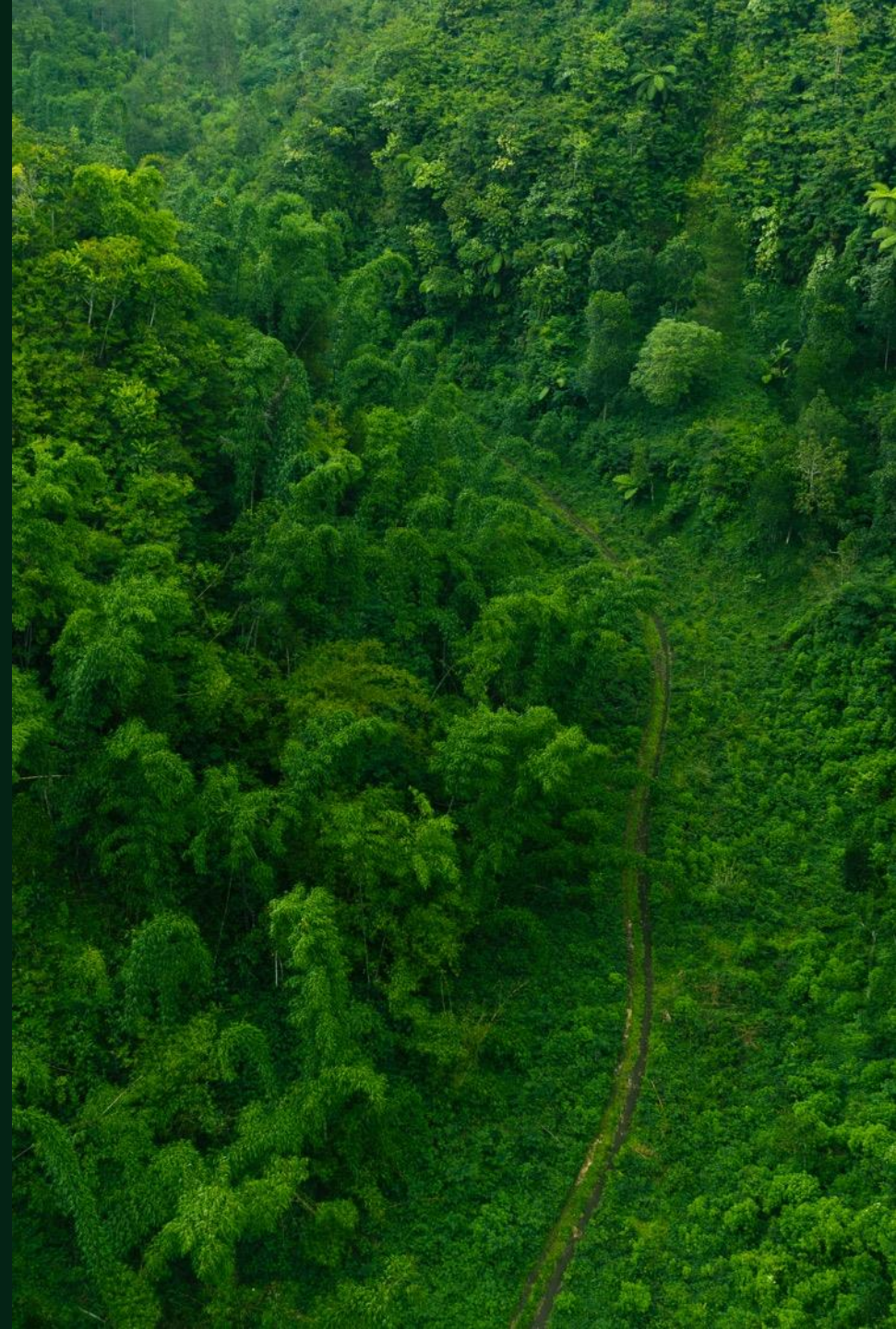
- 予測的なメンテナンス
- 自動的な障害回復メカニズム
- システムの自己最適化機能



kyndryl

Appendix

1. Agentic AI: Omnipresent Intelligence, BofA Global Research, 2024
2. Agent AI: Surveying The Horizons of Multimodal Interaction, Stanford University, Microsoft Research, Redmond, 2024
3. What's next in AI? A Framework for thinking About Inference Compute, Barclays, 2024
4. Artificial Intelligence Index Report 2024, Stanford University, 2024
5. AI Thinking: A Framework for Rethinking Artificial Intelligence in Practice, Denis Newman-Griffis, 2024
6. The Process Matters, Joel Brockner, 2016
7. A Brief History of Intelligence: Why the Evolution of the Brain Holds the Key to th Future of AI, Max S. Bennett, 2023
8. Nexus: A Brief History of Information Networks from the Stone Age to AI, Yuval Noah Harari, 2024
9. Scaled Agile Framework 6.0: Core Competencies



kyndryl

ありがとうございました

Re:Inventで再会しましょう―

本日発表の日本語文面は、Claude 3.5 Sonnet、Chat GPT 4.0およびMicrosoft Copilotの協賛でお送りいたしました

キンドリルジャパン株式会社

叶　奕池（ICHICHI）

Infrastructure/Cloud Architect

kyndryl