JaSST Hokkaido 2025 基調講演

QA for the Better

三菱電機株式会社
AI戦略プロジェクトグループ
改革推進部長　徳隆宏

徳 隆宏

三菱電機　AI戦略プロジェクトグループ改革推進部長
JaSST Kansai アドバイザー, QA4AI 副運営委員

業務システム、組込、品質保証、アジャイル、データ活用、AI
など様々なソフトウェア領域の実践者。
業務改革でのAI活用を中心としたプロジェクト全体の推進担当

- 2005年　外資系システムインテグレータ
  製造業を中心とした、基幹システム・業務システム・
  組込システム等の設計・実装・検証
- 2011年　国内制御機器メーカ
  ファクトリーオートメーションを中心とした、
  組込、AI・IoTシステムの、企画・研究・開発
- 2021年　国内機械メーカ
  - データ活用を対象としたクラウド・品質プロセス
  - DX戦略の立案・推進
- 2024年　三菱電機
  業務改革プロジェクトの組成・推進 / 生成AI活用の推進

設立年

1921

従業員数

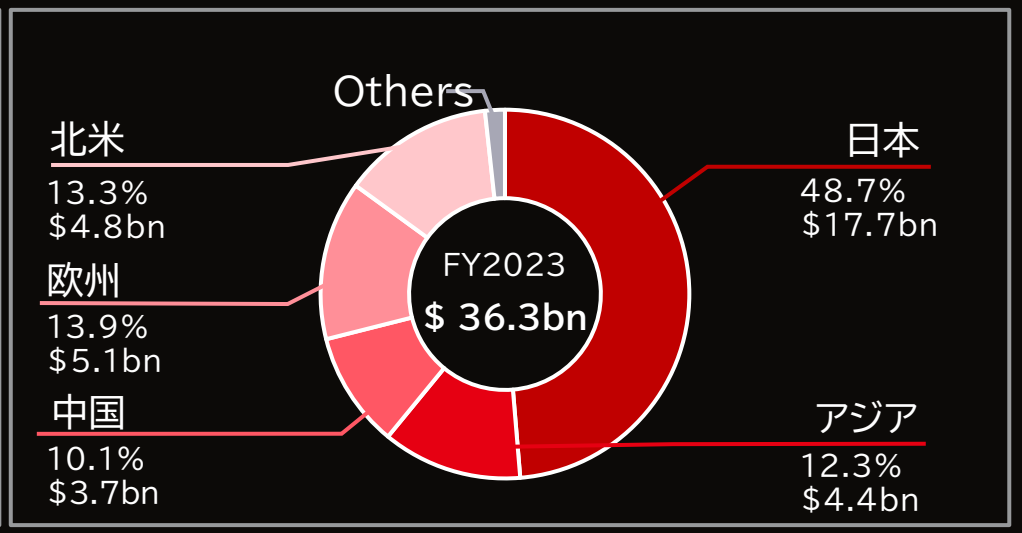
24年8月時点

149,655

関連会社数

24年8月時点

203



グローバル市場での展開

インフラBA

70機以上の
人工衛星を製造

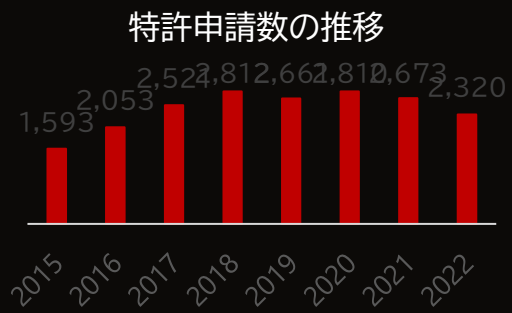
インダストリ・モビリティBA

サーボシステム
シェア：No.1

セミコンダクター・デバイス事業

民生用IPM
シェア： No.1

特許

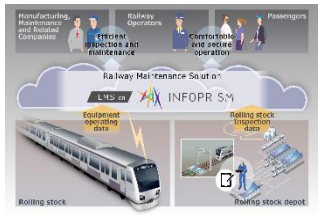
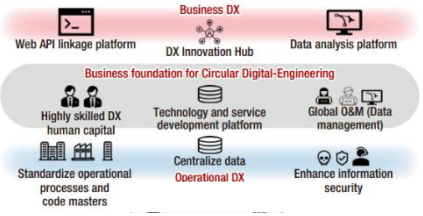


グローバル **4位**
（9年連続でTop5入り）

日本国内　**1位**
（8年連続で1位）

As of March 2023,
World Intellectual Property
Organization (WIPO)

5つのビジネスエリア(BA)

インフラ BA	インダストリー・モビリティ BA	ライフ BA	デジタルイノベーション事業	セミコンダクター・デバイス事業
 		 		 
Group	Group	Group	Group	Group
• 社会システム事業 • エネルギーシステム事業 • 防衛・宇宙システム事業	• FAシステム事業 • 自動車機器事業	• ビルシステム事業 • 空調・家電事業	• 情報・通信システム事業	• 半導体・デバイス事業
24% 13,100億円	29% 15,700億円	39% 21,600億円	3% 1,500億円	5% 2,900億円

三菱電機のAI

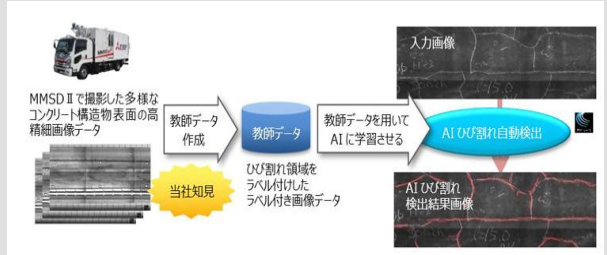
各種機器に搭載可能なコンパクトかつ高性能な機械学習型AI技術 Maisartを開発し、
製品および社内の製造現場へ適用

Maisart

製品適用

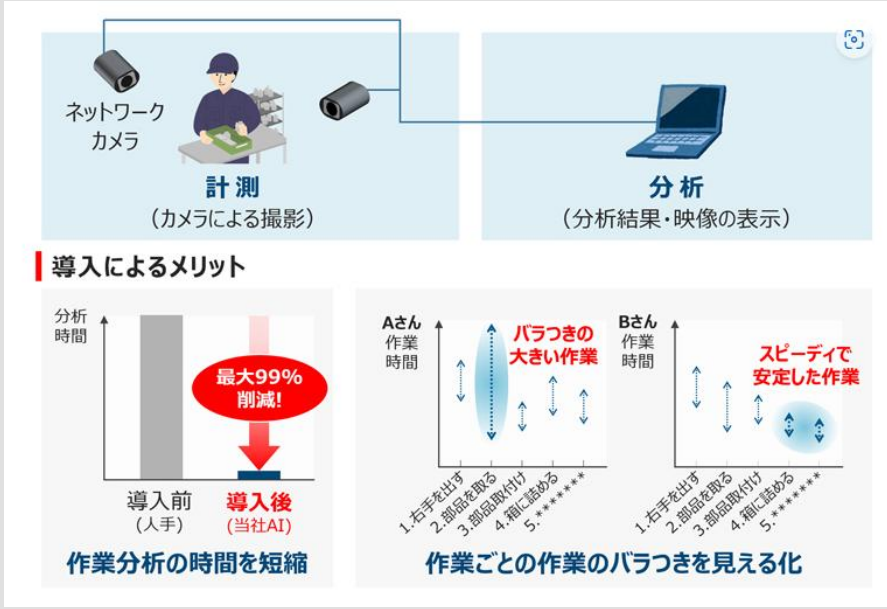
ビル用マルチエアコン、インフラ保守システム、FA機器（放電加工機等）、省エネ分析・診断アプリ、監視カメラ、介護見まもりサービス等の様々な製品に組み込み

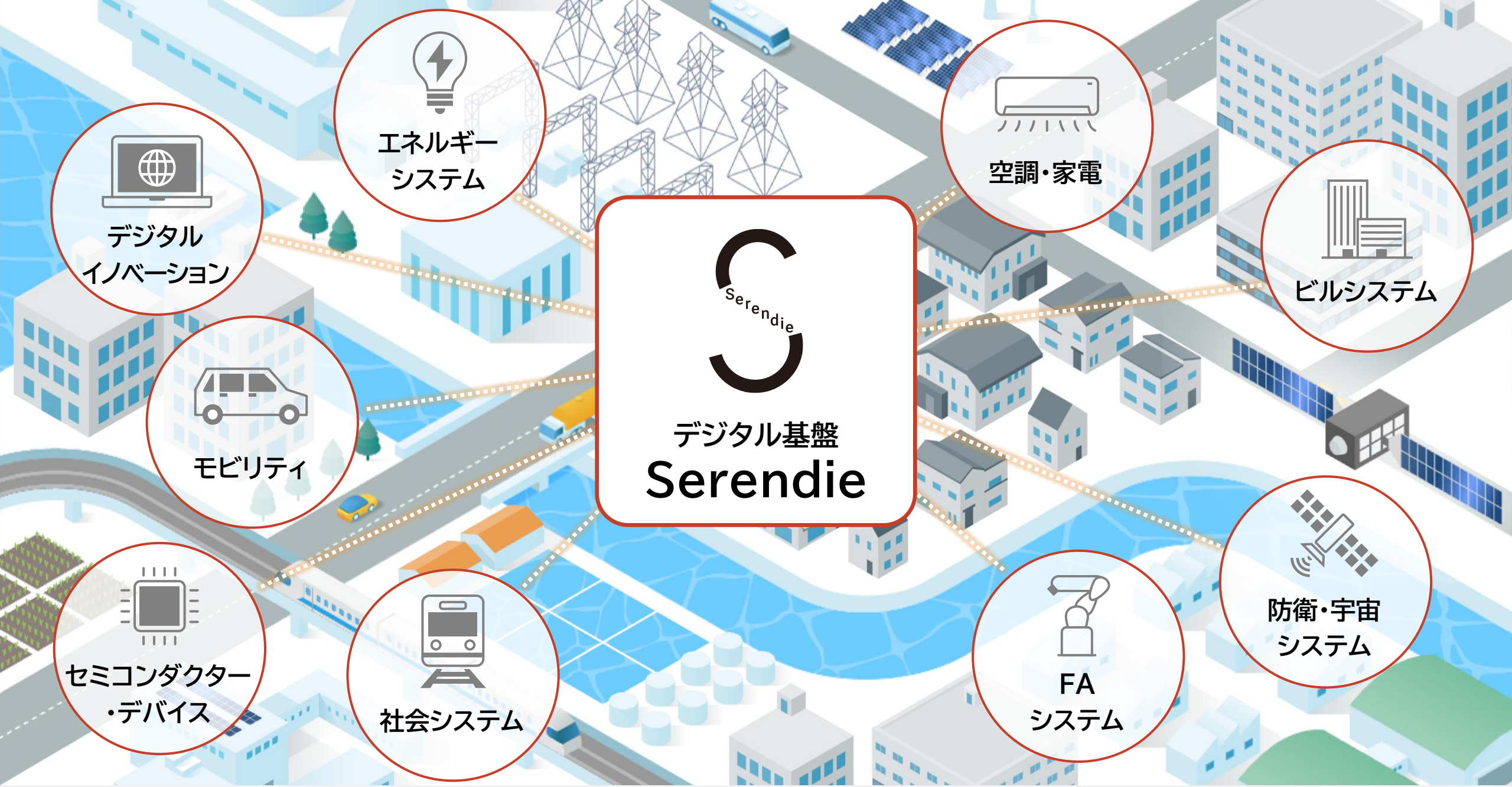


設計・製造現場

外観検査、作業分析、設備予知保全、製造データ分析などの製造関連業務だけでなく、部品需要予測、仕様作成支援、社内文書検索などの生産管理業務、設計業務まで、様々活用





私の所属: AI CoE

生成AI技術をグローバルで活用できる基盤を提供し、
業務効率化・高度化、事業での顧客価値向上に貢献する
Center of Excellence 組織

基盤

分散投資・重複投資を回避するために、AI活用基盤・開発基盤を整備

推進

知識・ノウハウのサイロ化を防ぎ全社の利用を促進するために、情報共有コミュニティやプロジェクトを推進

ガバナンス

各国規制や法案に対応するために、AIガバナンスの体制を構築・推進

QA for The Better

1

QA4AI コンソーシアム

コンソーシアム活動

- コーディネート
- ドメイン活動
- ガイドライン
- Open カンファレンス

2

生成AIブーム AIプロダクト進化の潮流

- 生成AIへの期待値
- AIサービスの進化
- AIエージェント
- AIによる開発支援

3

品質保証のチャレンジ

- アジャイル
- 性能評価基盤
- ガードレール
- AI ReadyなQA

https://qa4ai.jp

AIプロダクト（機械学習技術を利用した製品・サービス）を様々な組織が開発・利用 一方でどのようにリスク管理や品質保証をすべきか、産業界に明解な答えがないため、2018年に結成

- 主に機械学習技術を前提としている
  - 非AIプロダクトは、明示的なルール定義から開発（演繹的）
  - AIプロダクトは、結果（データ）から開発（帰納的）
- CACE性がある
  - Changing Anything Changes Everything
- 自動化が必要
  - モデルの構築・運用には大量のデータが必要
  - AIプロダクトが利用される状況変化に対し、データなどの観点で追従が必要

QA4AI ガイドライン

・ベストプラクティス（どんな課題があって、どう考える・どうする）を中心に編成
・ガイドラインは、（少なくとも）毎年更新
 ・ドメイン編はワーキンググループごとに検討・執筆

■共通編

・ガイドラインの目的とスコープ、品質保証の枠組み
・品質保証の技術 、説明可能性・解釈性

■ドメイン編

・コンテンツ生成（動画生成など）
・Voice User Interface（スマートスピーカーなど）
・産業用プロセス（装置・設備など）
・自動運転（自動運転車両）
・AI-OCR（図面・テキスト読み取り）
・対話型生成AI

コンソーシアムの活動

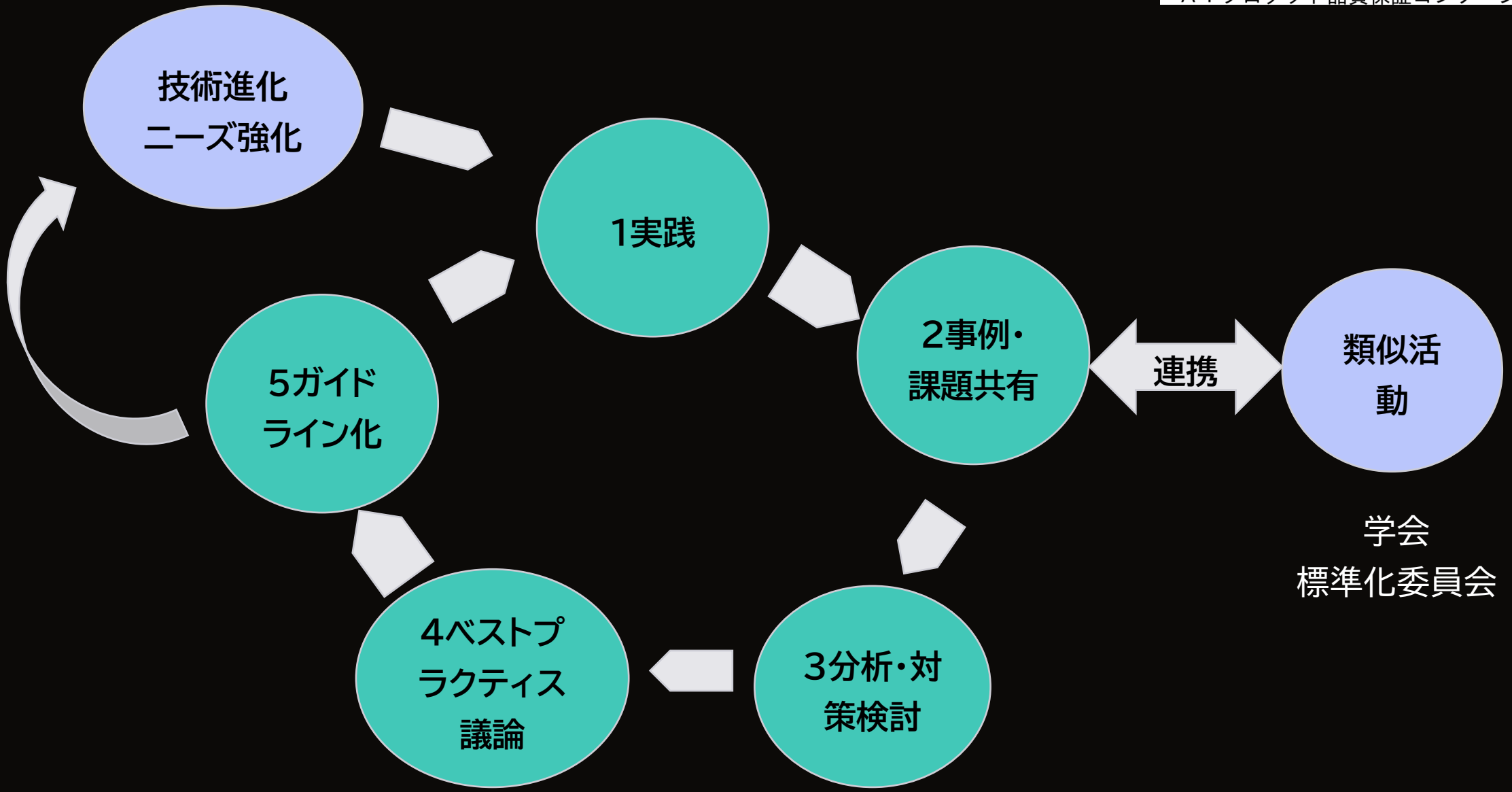
Consortium of Quality Assurance for Artificial-Intelligence-based Products and services
ＡＩプロダクト品質保証コンソーシアム

オープンに、マルチドメインで、技術的なベストプラクティスを提供
・開発組織も、顧客も、（相応の）安心を。品質保証側が邪魔をせず過度な期待を防ぎたい
・オープンで技術志向の組織運営。招待型でベストエフォートで参加
・AIプロダクト共通ガイドに加え、マルチドメインでガイドライン提供

マルチドメイン： AIプロダクトの領域ごとに深堀議論
・AIプロダクトの使われ方や作られ方によって課題やベストプラクティスが異なる
⇒ ドメインごとにワーキンググループ活動
・ドメインの例：AI-OCR、産業用プロセスデータ、コンテンツ生成、対話型生成など
・ワーキンググループ活動の例：情報収集・課題整理・ガイドライン執筆・社外発信

QA4AIの活動のモデル

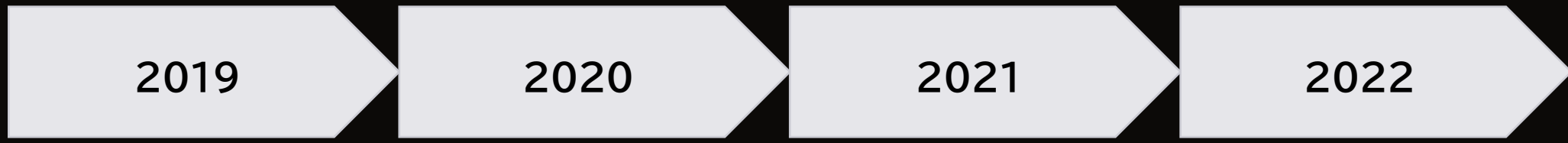
Consortium of Quality Assurance for Artificial-Intelligence-based Products and services
ＡＩプロダクト品質保証コンソーシアム



QA4AIでのドメインの変遷

・2019年のころは「AIの定義を議論する」時代であったが、今や「AIの活用を議論する」時代に

QA4AIは初期からコンテンツ生成系システムをはじめとした、主要なAIプロダクトのドメインをカバー
大規模言語モデルへの対応も2024年から開始



生成系システム
Voice User Interface
産業用プロセス
自動運転

＋AI OCR

＋説明可能AI

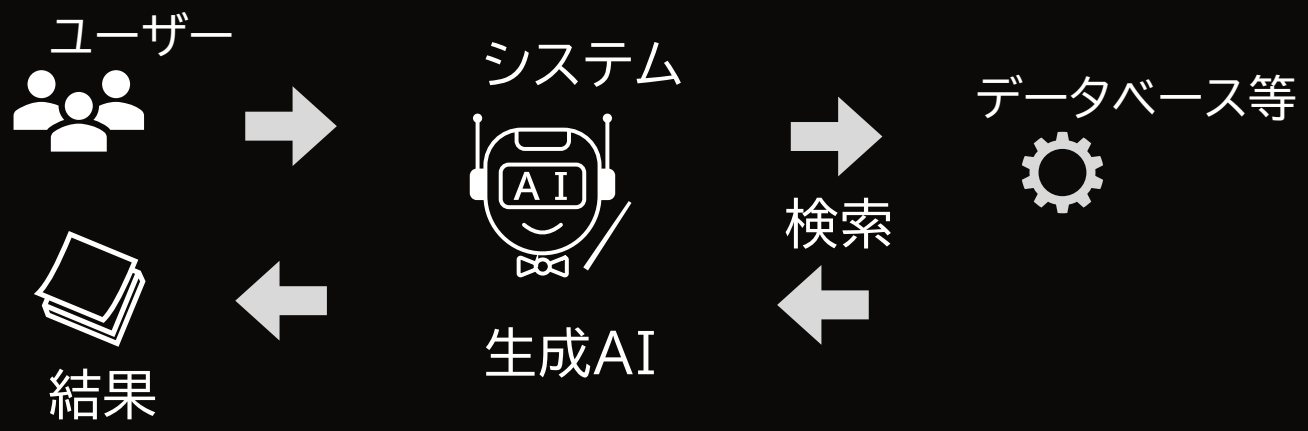
＋大規模言語モデル・対話型生成AI

QA4AIの事例　対話型生成AIの品質特性

ID	評価項目	説明
QC01-1	自然言語処理における回答性能の評価	感情分析、分類、推論、要約、QA、翻訳など多様なタスクにおける包括的なベンチマーク（例：GLUE、SuperGLUE）による評価
QC01-2	ツール活用に関する回答性能の評価	外部ツールの入力の正確性、実行の成否、タスク達成度。CodeBLUEなどの専用指標も存在
QC01-3	創造性・多様性に関する回答性能の評価	再生成などにおける回答間の多様性や創造性の評価（明確な手法は未確立）
QC01-4	制御可能性・協調可能性の評価	出力フォーマットや禁止事項など指示の反映度を評価。メタモルフィックテストによる評価が想定される
QC02-1	一般的な知識に対する事実性・誠実性の評価	プロンプトに情報を含めずに回答させ、LLMの内在知識に基づく正確性を評価。TriviaQAやKoLAなどを使用
QC02-2	与えた知識に対する事実性の評価	ファインチューニングやプロンプトにより与えた知識に対する応答の正確性を評価
QC02-3	根拠の説明性・妥当性の評価	回答と共に出力された根拠情報の妥当性・信頼性を評価
QC03-1	公平性の評価	性別・人種などセンシティブ属性に関するバイアス評価。StereoSetやCrowS-Pairs、BOLDなどのベンチマークを使用
QC03-2	安全性の評価	毒性や有害性のある応答の有無を評価。OLID、RealToxicityPrompts、HarmfulQなどを使用
QC03-3	データガバナンス	訓練・生成データに関する著作権、プライバシー、利用制限の遵守などを評価
QC04	頑健性の評価	入力摂動に対する応答の安定性を評価。AdvGLUE、ANLIなどを使用
QC05	AIセキュリティの評価	ジェイルブレイクや攻撃プロンプトに対する耐性を評価。具体的手法に基づく実験的評価

対話型生成AIの品質保証の課題例　RAG

「Retrieval Augmented Generation」の頭文字
生成AIの言語生成能力を、検索技術と組み合わせて拡張する仕組み



品質課題：　検索精度

RAGASフレームワーク

メトリクス	説明
Context Precision	検索抽出したコンテキストが質問に関連している度合い。不要な情報が含まれていないか。
Context Recall	質問に答えるために必要な情報が、検索されたコンテキストに含まれている度合い。
Context Entities Recall	Context RecallをEntity（固有表現）単位で計算したメトリクス。
Noise Sensitivity	ノイズ（無関係なコンテキスト）が混入した場合の、回答への影響度合い。
Response Relevancy	生成された回答が元の質問とどれだけ関連しているか。
Faithfulness	生成された回答が、与えられたコンテキストに基づいてどの程度推論できているか。
Multimodal Faithfulness	テキスト・画像・音声など複数のモダリティに対して回答が忠実であるか。
Multimodal Relevance	複数モダリティに対する回答が質問と関連しているか。

QA4AIの品質保証: 5つの軸とバランス

一定の品質管理目標を置きながら
漸進的に5軸の品質を高め、プロダクト品質を向上する

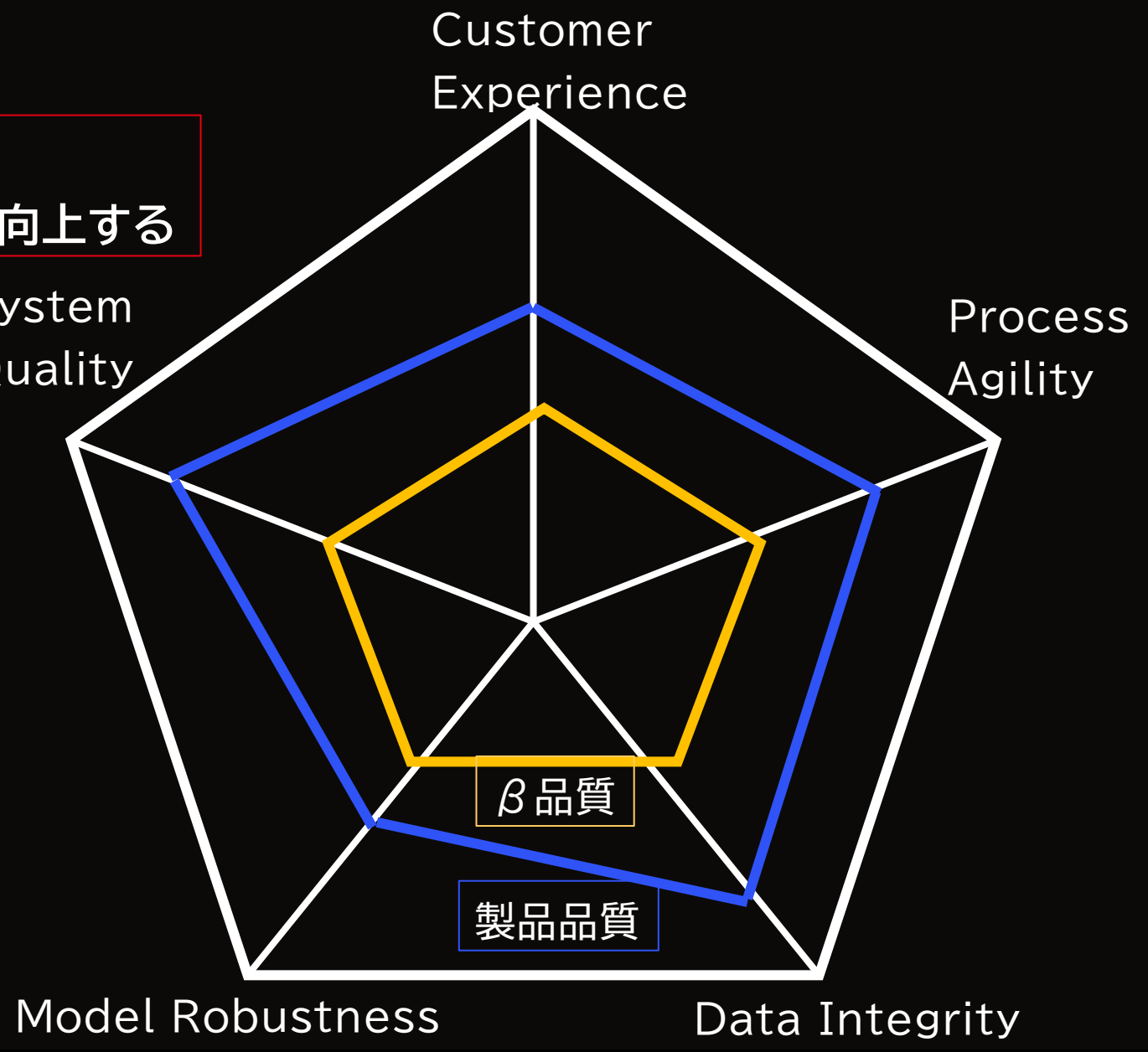
① Data Integrity
（データの量、質、管理 など）

② Model Robustness
（性能尺度、多様な検証、監視 など）

③ System Quality
（提供価値、インシデント許容度、結合度 など）

④ Process Agility
（データ収集反映、リリースサイクル、
構成管理、省力 など）

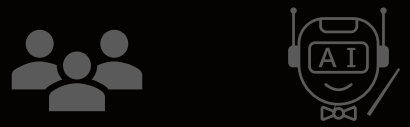
⑤ Customer Expectation
（データの質や量への理解、説明性や原因追及の期待値、
実運用での協力度 など）



QA for The Better

1

QA4AI コンソーシアム



コンソーシアム活動

- コーディネート
- ドメイン活動
- ガイドライン
- Open カンファレンス

2

生成AIブーム AIプロダクト進化の潮流

- 生成AIへの期待値
- AIサービスの進化
- AIエージェント
- AIによる開発支援

3

品質保証のチャレンジ

- アジャイル
- 性能評価基盤
- ガードレール
- AI ReadyなQA

AIが経営課題の本質的な解決手段として注目され、活用が広まっている

	経営課題	AIがもたらす変革例
Volatility 変動性 変化が激しく予測が難しい	技術革新・規制変更・為替・需給変動などの柔軟な対処	▶ リアルタイムでの**分析とシナリオ予測**
Uncertainty 不確実性 将来の展望が見えにくい	顧客ニーズ、需給予測、新製品投資判断	▶ いまあるデータでの**探索・シミュレーション**
Complexity 複雑性 多様な要因が絡み合う	事業・グローバル・ESG・人材・データなど多様で有機的に関係する事象からの判断	▶ AIが多様なデータを**分析し、関係性と構造を可視化・簡素化**
Ambiguity 曖昧性 正解がない状況	意味や因果関係が不明瞭で、正解が明確ではない	▶ 様々な視点での**深い分析や、戦略立案とブラッシュアップ**

急速に進化するAI、サービス

※一部

2024-12-03　Amazon Bedrock上で新たな基盤モデル群「Amazon Nova」シリーズ

2025-01-09　Neuron SDK 2.21. Trainium2およびInf2インスタンスでの
モデルトレーニングと推論をサポート

2025-03-10　DeepSeek-R1がフルマネージドなサーバレスモデルとしてAmazon Bedrockでサポート

2025-03-31　「Nova Act」を発表し、ウェブブラウザ内でのタスク実行を可能に

2025-04-07　Amazon Bedrockでプロンプトキャッシュが GA

2025-04-08　Amazon Nova Reel（動画生成モデル）、Nova Sonic（音声生成モデル）などをリリース

2025-04-09　Amazon Q Developerが日本語サポート

2025-04-30　Amazon Nova Premier（高度なマルチモーダルモデル）をAmazon Bedrock上で提供開始

2025-05-22　Amazon BedrockでClaude 4（Opus 4 / Sonnet 4）を提供

2025-06-24　Amazon Bedrock Guardrailsが更新。日本語含め60言語でポリシー違反を検知・フィルター

業務改革プロジェクト

生成AIを積極的に活用した業務プロセス改革を推進し、
蓄積したノウハウをリアルタイムに事業等へ展開

社内業務への適用例

業務プロセス改革

戦略・企画	事業環境分析資料の自動生成、戦略提案
人事・総務	問合せ対応の自動化
経理・財務	伝票処理におけるエラー、不正兆候の検知

設計製造プロセス改革

ハードウエア設計	仕様の照合、コスト見積り
ソフトウエア設計	コード生成、テストケース生成
品質保証	デザインレビュー支援、購入品メーカの監査支援

業務改革プロジェクト

- DX・AI活用等により、**徹底的にノンコア業務を効率化**
- 効率化で得た余力をコア業務に投入するとともに、**コア業務そのものの高度化・効率化**をDX・AI活用で実現

当社全体の競争力と効率を高め、**業務効率を2倍に**

業務改革プロジェクト選定プロセス

1. ありたい姿やニーズを詳細化し、約1000件のアイデアを抽出
2. アイデアを類型化し、60のユースケースを抽出
3. インパクト指標・実現容易性を基に、プロジェクトロードマップを策定

業務改革プロジェクトの推進における課題と当社のアプローチ

改革のオーナーシップ

- 業務部門のトップから現場まで自分ごとで進める必要がある
- AI CoE は推進の中で全体最適を担う

全社体制の中で、それぞれの
責任を果たす推進形態

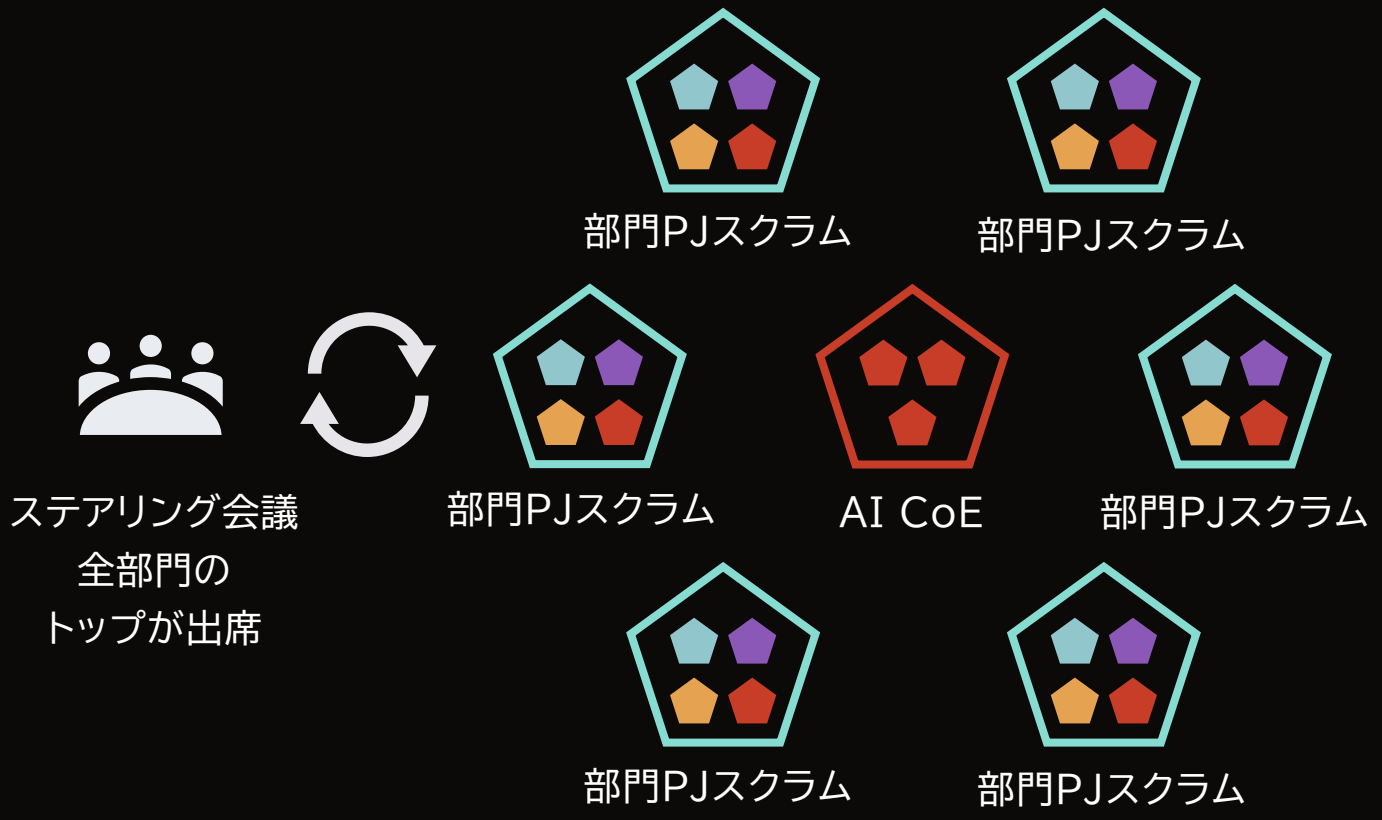
不確かなゴールと効果の刈り取り

- 仕様や効果が明確でない
- 使用の中で精度などの品質向上が必要

アジャイル推進と
生成AIアプリケーションの品質保証

エンタープライズアジャイル推進

エンタープライズアジャイル推進



Scrum@Scaleをもとにしたスケール組織運営

部門がオーナーシップ

- 各業務部門が主体
  - プロダクトオーナー（PO）
  - スクラムマスター（SM）
- 透明性を持ちCxOが意思決定

内製

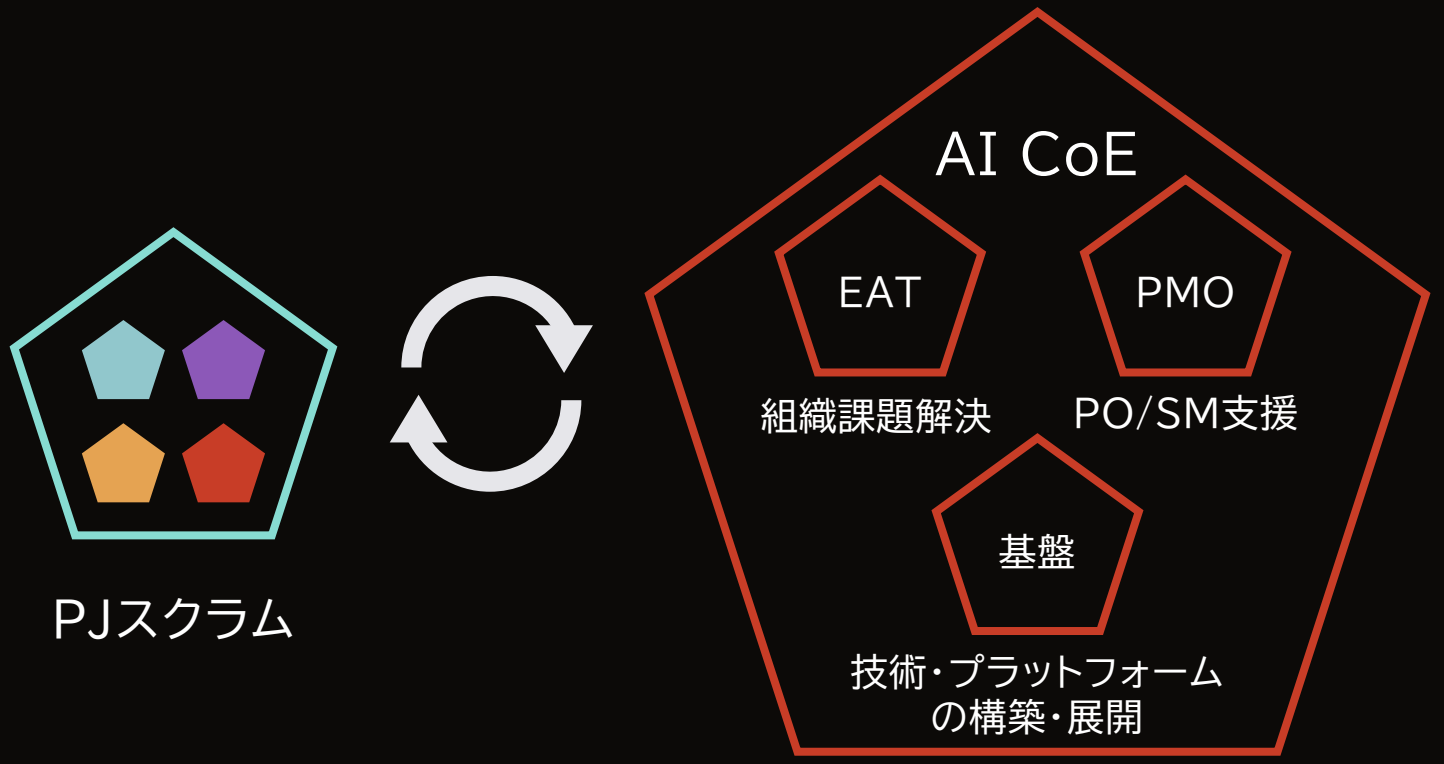
- 社内開発者
- AI専門家（AI CoE）

AI CoEによる全体最適

- 基盤の強化の展開
- アジャイル推進

エンタープライズアジャイル推進

AI CoEが各プロジェクトを伴走支援し、組織課題・プロジェクト課題・技術課題の解決を担う



- EAT(Executive Action Team)
  KPI設計や組織課題に対処
- PMO
  プロジェクトの全体進行と
  プロダクト価値の最大化を支援
- 基盤
  技術・データなどプラットフォーム整備し、
  設計やエージェントサービスを横展開

EATの機能： アジリティを得るための企業の運営ルール策定、アジリティ実現のための組織能力構築等

AI CoEの難しさとやりがい

PMO	EAT	基盤
プロダクト マネジメント	トップとチームとの 横断的疎通	開発基盤
プロジェクト マネジメント	プロダクトオーナー 横断の組織化	運用基盤
Agile Quality Assurance	ルールの改善と 課題解決	ガバナンス

前例がないことを、組織的に・スピーディに実行するミッションを負うチーム

生成AI活用基盤で目指す姿

AIエージェントを簡単/爆速に開発・利用

必要なアセットを再利用

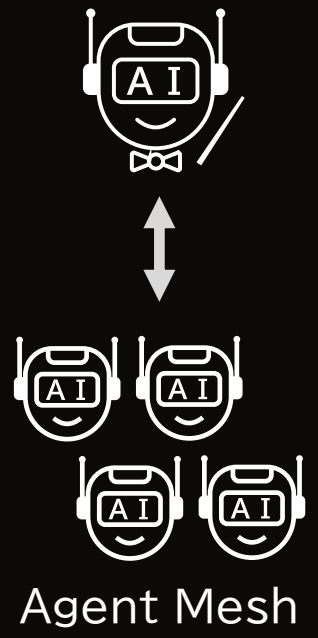
生成AI活用基盤

- AIエージェントツール
- データコネクタ
- MCPサーバ
- ナレッジ

AIエージェントの開発効率性、安全性を向上

- デバック/トレーシング
- 性能評価基盤
- ガードレール
- ナレッジ

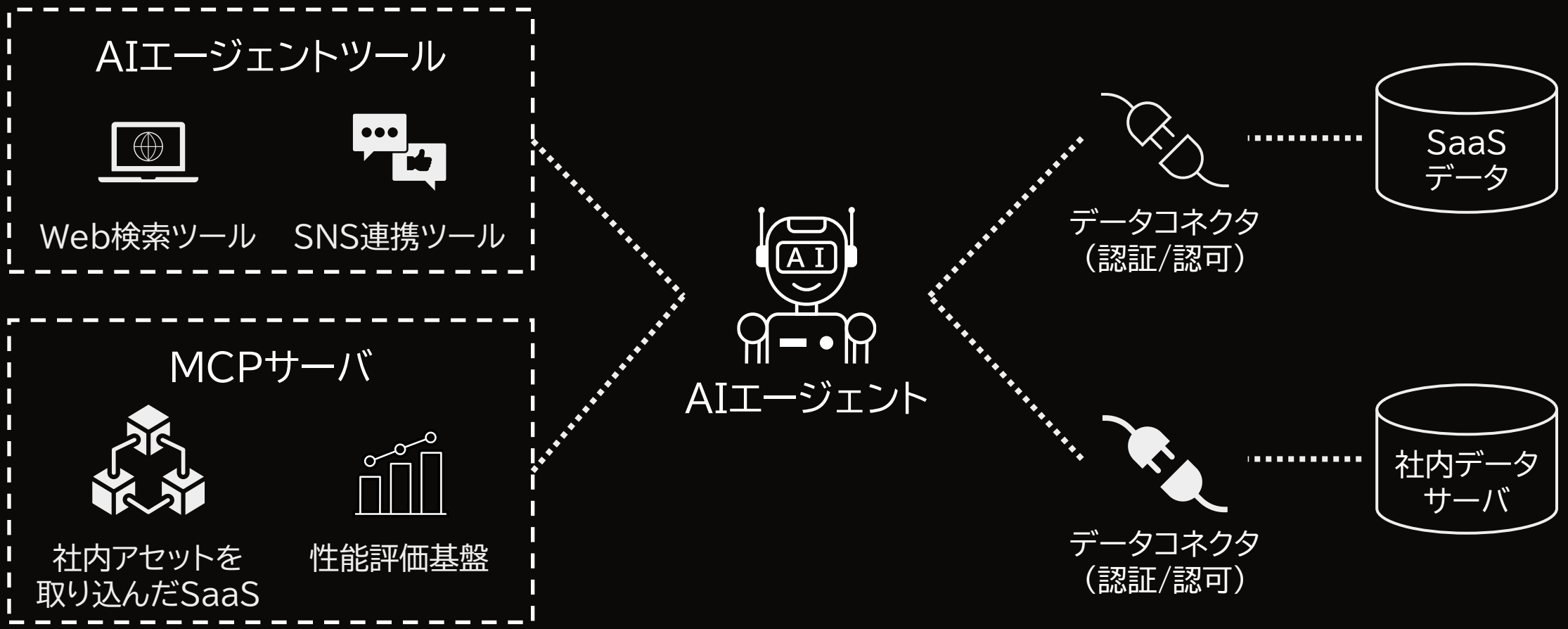
“Agentic AI”を基盤としたAgent Meshにより、高価値なアウトプットを生成



Agent Mesh

AIエージェントを簡単/爆速に開発・利用

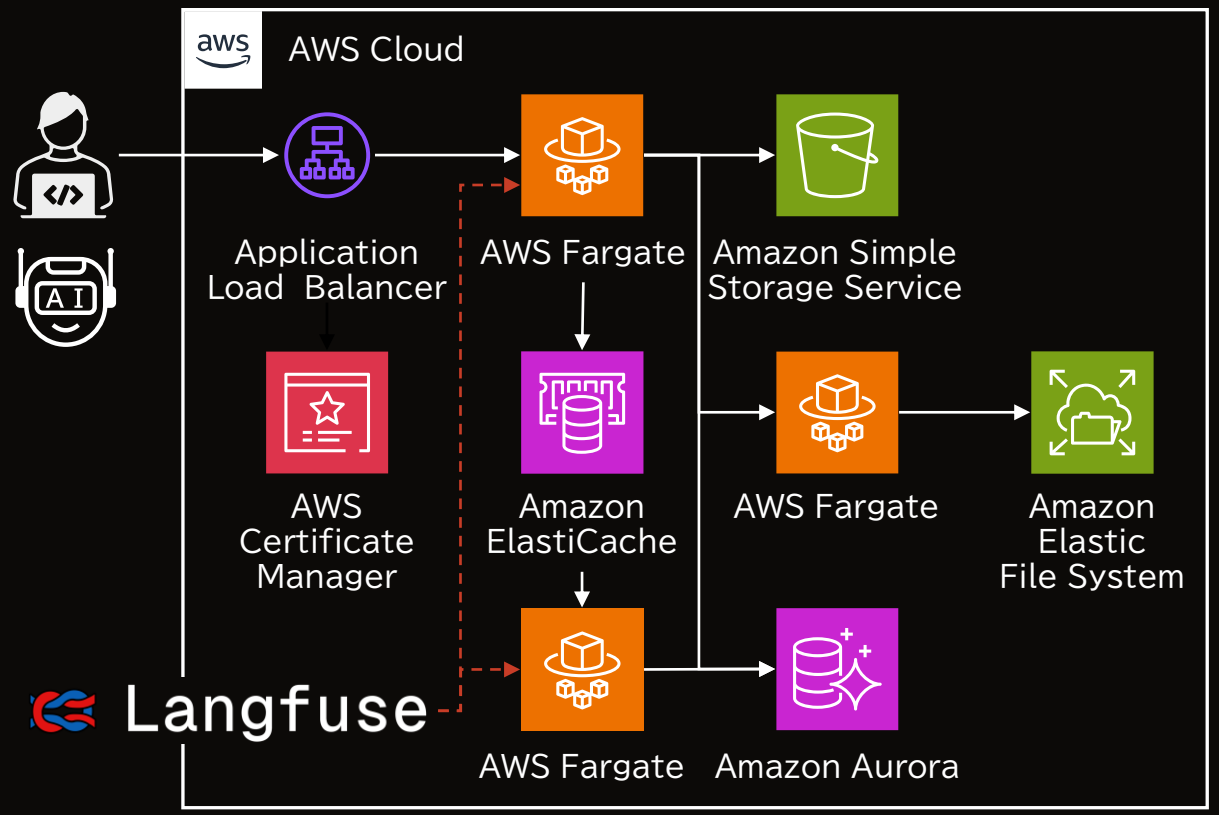
生成AI活用基盤のアセットを活用し、AIエージェントを爆速で構築・利用



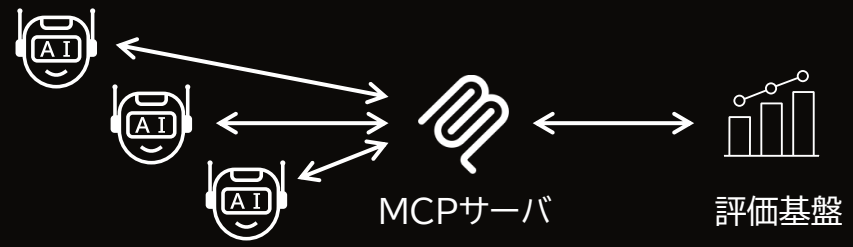
AIエージェントの開発効率性や安全性を高める

LLMOpsによる**継続的改善**を実現

デバッグ/トレーシング



性能評価基盤



評価基盤をMCPサーバ化し、
評価をより簡単に実行
「MCP Server as a Judge」

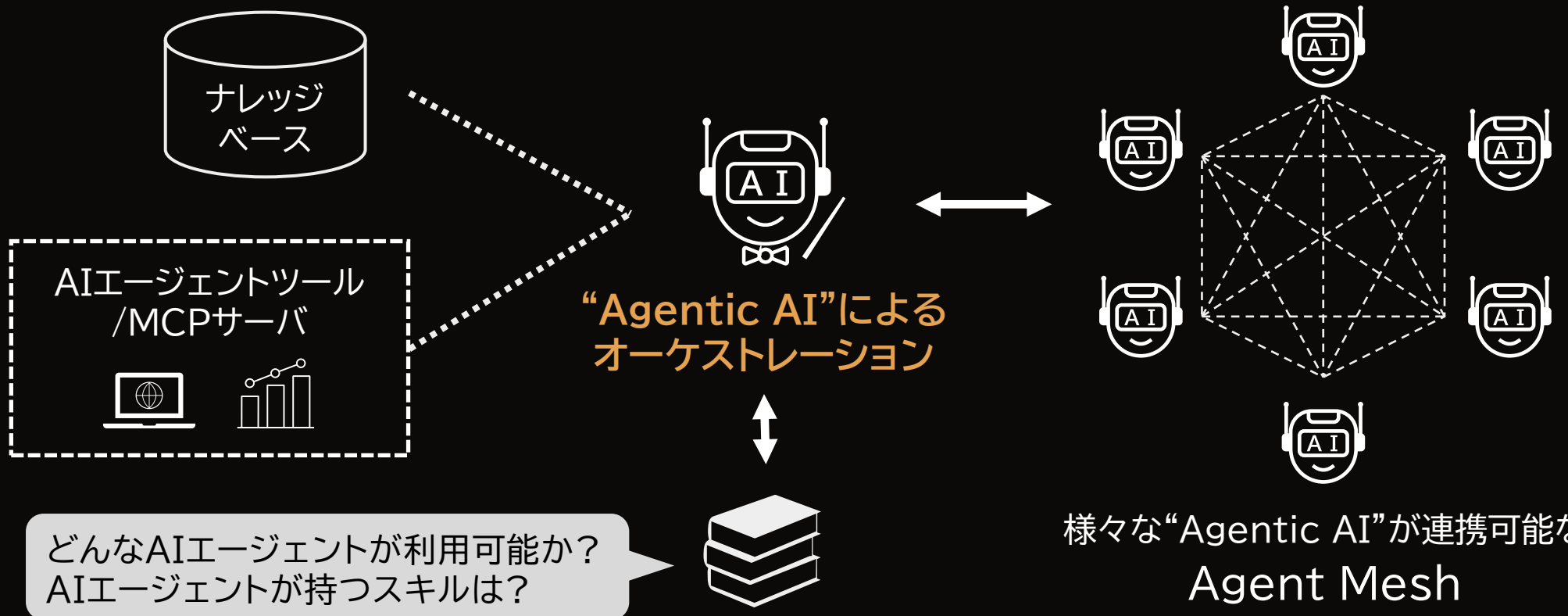
安全性検証

Amazon Bedrock Guardrails

ガードレールによる
入出力データの検証

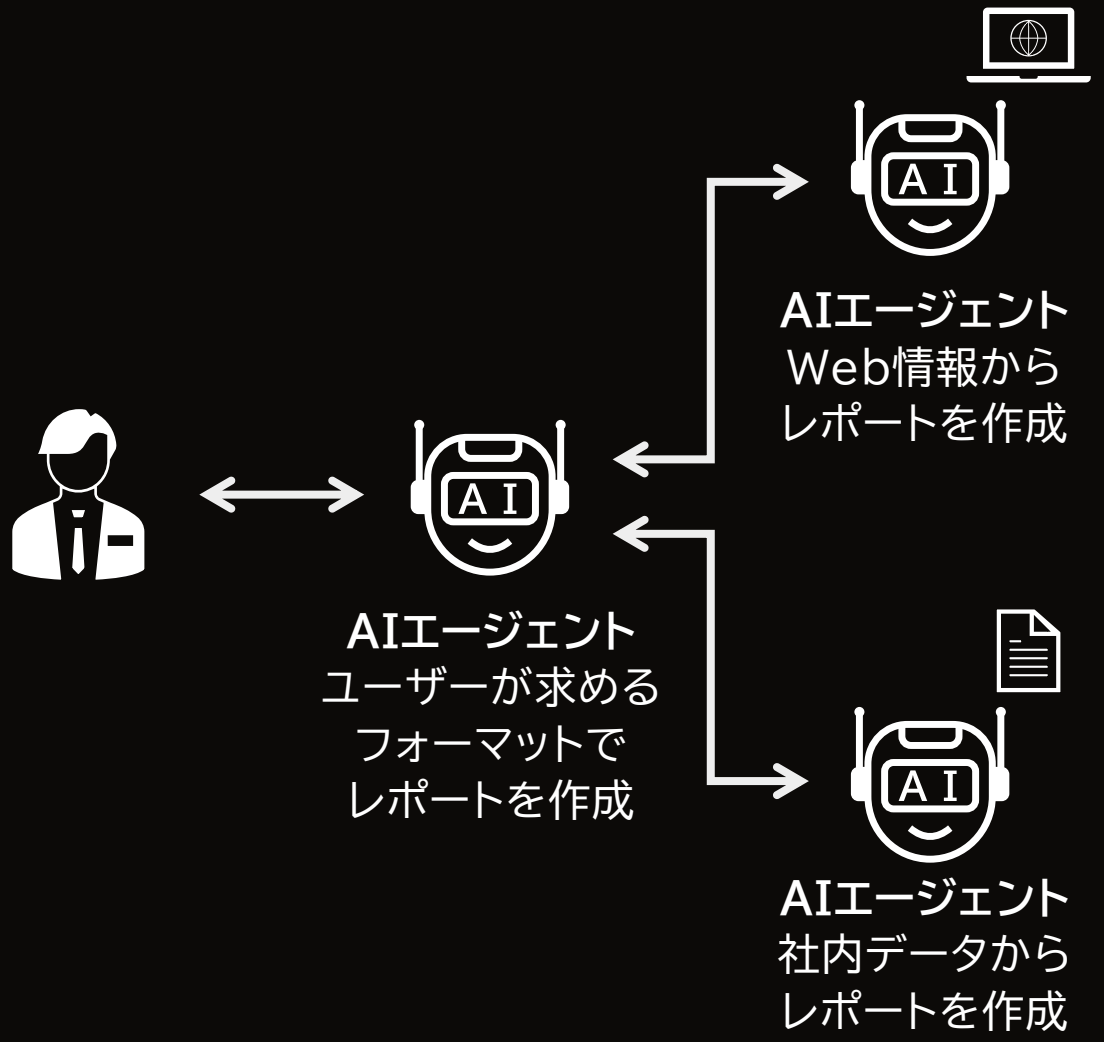
“Agentic AI”を基盤とするAgent Meshにより、高価値なアウトプットを生成する

“Agentic AI”の相互活用により、高度な課題を解決



*Agentic AI：目標を持ち、自律的に行動を選択・実行し、環境に働きかけるAI

生成AI活用基盤の事例「戦略立案エージェント」



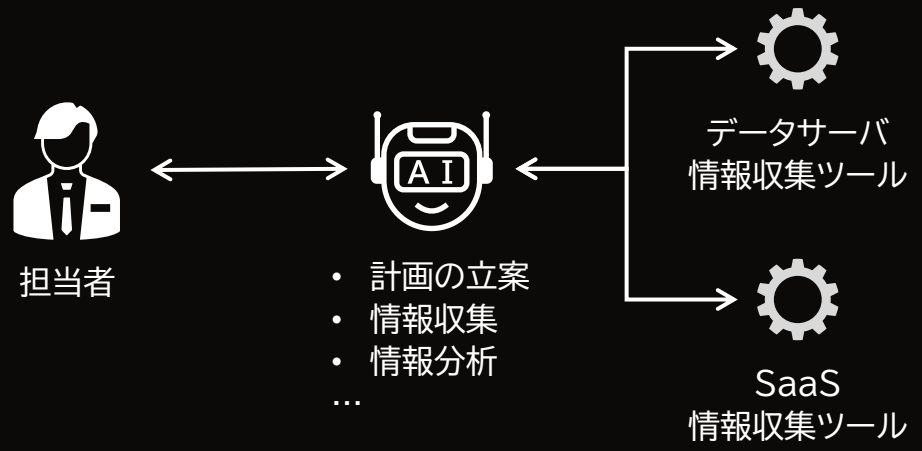
「AIエージェント×AIエージェント」による戦略立案レポート生成

- 個々のAIエージェントは、
  流用性を高めるために、特定タスクを担う
- 高度なユーザーニーズに応じて、
  動作するAIエージェントの種類や連携を切り替え

今後は、より多くのエージェント
を協調させ、高度化を目指す

AIエージェントの協調による課題の解決

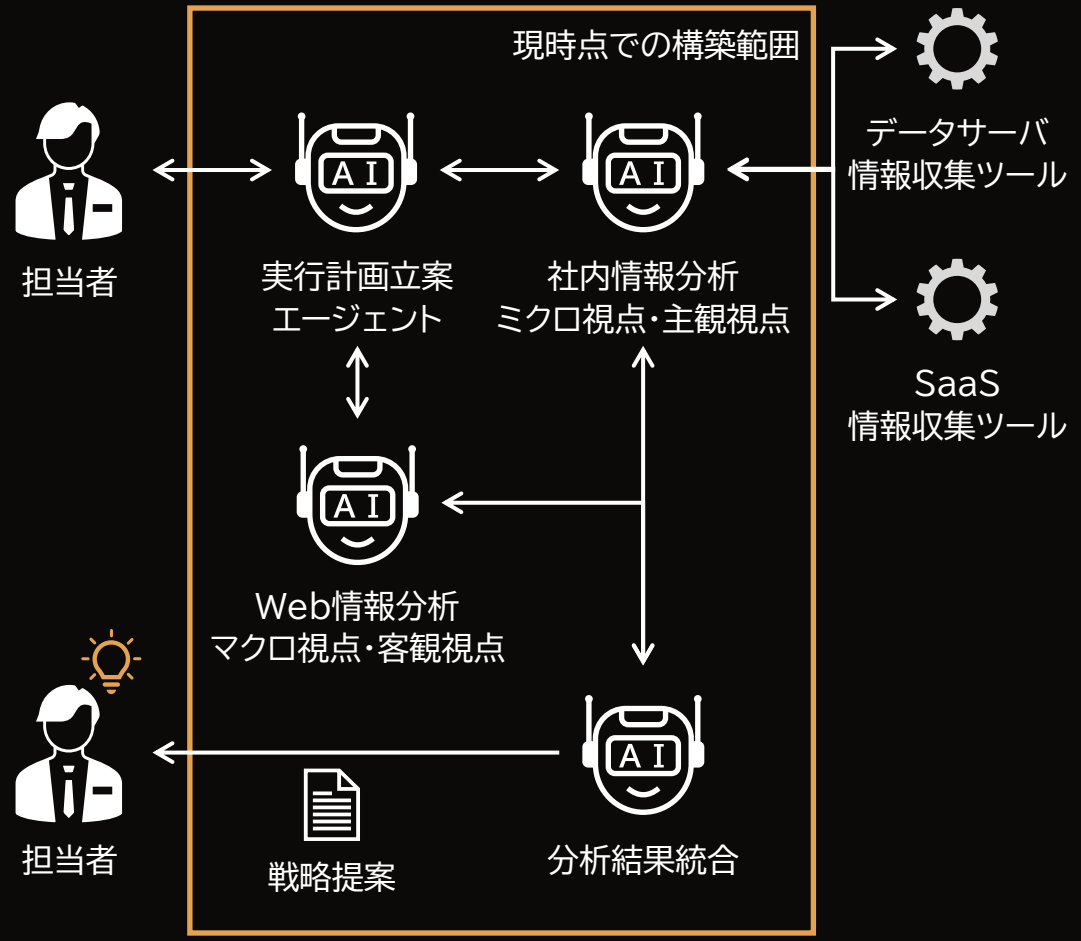
単一のAIエージェントの場合



問題点

- AIエージェントへ指示が伝わらない
- AIエージェントが処理する情報量が多くなり、内容が薄くなる
- 外部情報検索の際に、意図せず社内情報を漏洩するリスク

「AIエージェント×AIエージェント」により複雑な課題を分割し、全体の解決を目指す



AIエージェントによる開発支援の活用

- Github Copilot, Cline, Amazon Q developer, Kiro etc…
  - UI周辺、IaC (CDK) コーディングの効率化
  - 構成図(drawio)のドラフト作成
  - テストケースのドラフト作成

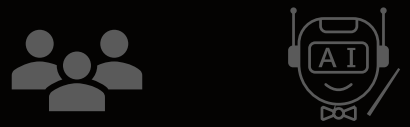
  など。活用は多岐にわたる。

  新しいツールも適宜評価し活用

QA for The Better

1

QA4AI コンソーシアム



コンソーシアム活動

- コーディネート
- ドメイン活動
- ガイドライン
- Open カンファレンス

2

生成AIによる AIプロダクトの進化

- AIへの期待値
- AI活用サービス
- AIエージェント
- AIによる開発支援

3

品質保証のチャレンジ

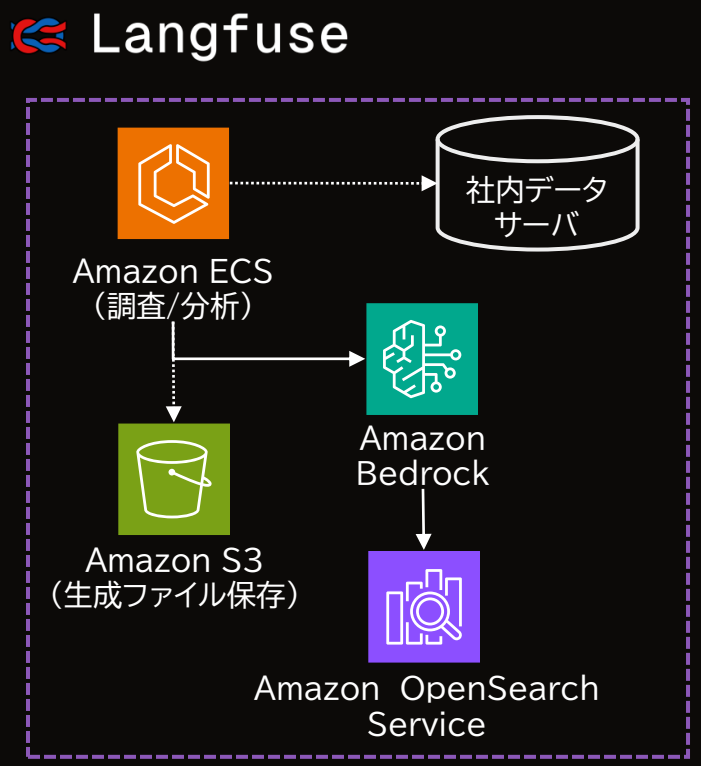
- 大規模アジャイル
- デバック/ トレーシング
- 性能評価基盤
- ガードレール

品質保証の課題

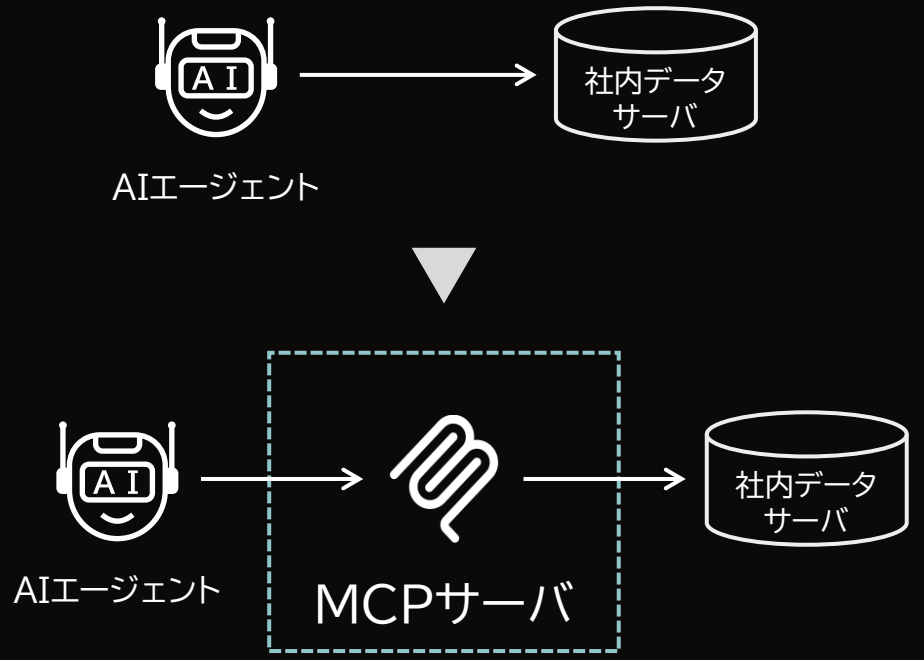
項目	説明	具体的な課題例	対処例
1. チーム間での品質基準のばらつき	複数のスクラムチームが独立して動くため、品質基準や受け入れ条件がチームごとに異なる	・テストカバレッジの差異・コードレビューや静的解析の実施レベルが不統一	品質計画のテンプレート化とレビュー
2. E2Eテストの統合の困難さ	コンポーネントがチームごとに開発されており、システム全体としての統合テストが後回しになりやすい	・E2Eテストが後工程になり、バグの検出が遅延・CI/CDパイプラインの整備不足	E2Eテストの早期着手モチベーション 共通CI/CD・E2Eテスト環境の展開
3. リリースの同期・段階的リリースによる複雑化	複数のチームが異なるペースでリリースを行うため、統合的な品質評価が難しい	・スプリント単位での完成定義（DoD）が曖昧・リリースごとにQA対象が変動	基盤の視点でアーキテクチャをレビュー
4. テストの自動化戦略が全体最適になりにくい	各チームが個別にテスト自動化を進めると、ツール・方式・レベルが分断されやすい	・重複や未整備のテストケースが発生・メンテナンスコストの増大	MVPからテストパターンの整理
5. QA人材の配置とロールの明確化が難しい	各チームにQA担当を割くのが難しく、全体的な品質視点が欠落しやすい	・専門QA不在による品質観点の抜け漏れ・横断的なQA組織が機能しない	品質管理計画で定義し、人財配置
6. 仕様・変更要求の可視性とトレーサビリティの欠如	ユーザーストーリー中心の開発で要件の断片化が起き、全体仕様とテストの紐付けが困難	・要求変更による品質インパクトの把握が困難・テストケースのメンテが追いつかない	QA担当によるインスペクション
7. 非機能要件（性能・セキュリティ・信頼性・データ）の軽視	ビジネス要求に即した短期デリバリに注力しすぎて非機能要件のテストが後回しに	・パフォーマンスや可用性のテスト未実施・セキュリティリスクの見落とし	スクラム内だけでなくAI CoEからもフィードバック
8. 品質メトリクスの収集と活用の困難さ	大規模な構成では品質指標（バグ密度・テスト通過率など）の全体把握が困難	・品質KPIがチーム単位でしか追えない・経営層・顧客への報告が曖昧	EATによるマネジメント

デバッグ・トレーシング

- LLMOpsツールをAIエージェント環境に組み込み
- トレーシングしながらアウトプットを検証・改善



社内データ収集など汎用性が高いタスクを
MCPサーバ化し、AIエージェントで活用



生成AI活用基盤へ取り込み

生成結果の品質改善を、Agentic AIで高める基盤

現状の評価プロセス

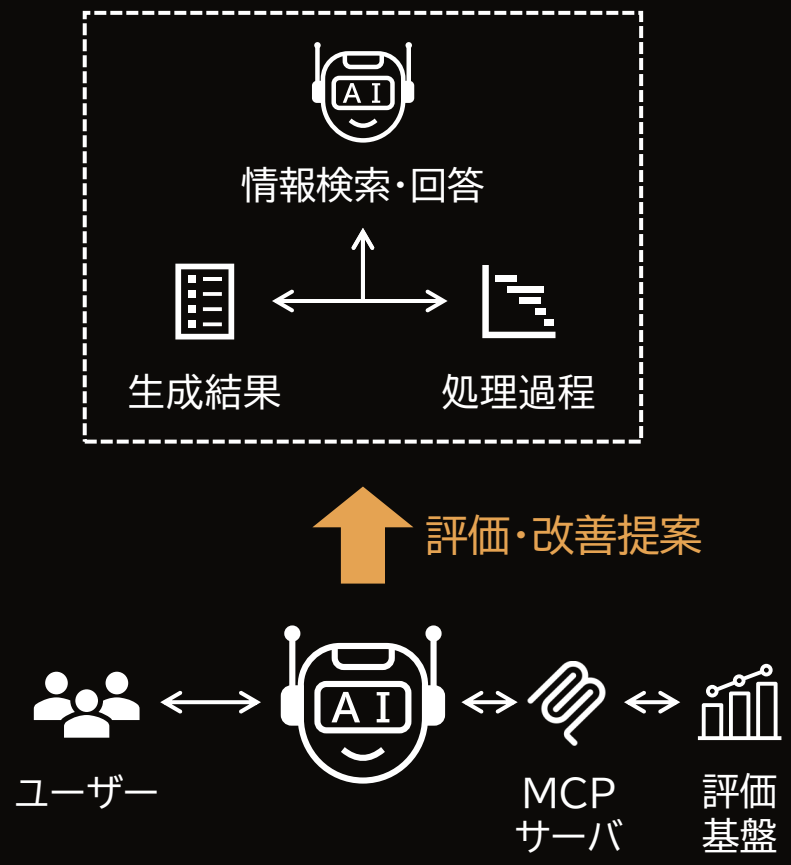
ユーザーやエンジニアが、知見と時間をかけてAIエージェントを評価
（一貫性、妥当性、情報の客観性や主観性）

▶ **課題** 人手による検証は、時間・コストが膨大

継続的品質改善に向けた取り組み

評価&改善提案するAgentic AIの開発と運用

実現に向けて
- AIエージェントのObservability向上
- ユーザーフィードバックと定量評価の関連付け
- AIエージェントが処理できるように、人の暗黙知の形式知化



Agentic AIを中心した評価・改善

ガードレール

生成AIの安全策を実現するためには、多層での防御が必要
代表的に必要な設定を見極め、標準的なガードレール環境を構築しつつ、設計・検証で担保

レイヤー	考慮事項	補足/例
① 入力前処理層	ユーザー入力のフィルタリング・制限（プロンプトインジェクション対策など）	- 入力正規化 - 禁止ワードフィルター - スクリプト検出
② モデル呼び出し層（Bedrock等）	モデルの安全設定、API利用制限	- BedrockのGuardrails - 呼び出し頻度制限
③ 出力後処理層	LLM出力のフィルタ・再評価・検知	- 禁止ワード・内容チェック - 自動評価スコアによるReject/再プロンプト
④ UI/UX層	ユーザーへの説明責任・選択肢制御	- 出力のファクトチェック表示 - ユーザーへのガイドや警告メッセージ
⑤ 監視・ログ層	利用履歴の記録と監査	- Prompt/Responseのログ化 - センシティブ入力のアラート通知

QA for The Better

1

QA4AI
コンソーシアム

2

生成AIによる
AIプロダクトの進化

3

品質保証のチャレンジ

多様で日々進化する技術活用を前提としたAIプロダクト

品質保証には技術を手の内化しつつ、日々良くすることが必須

コーディネート

ドメイン活動

ガイドライン

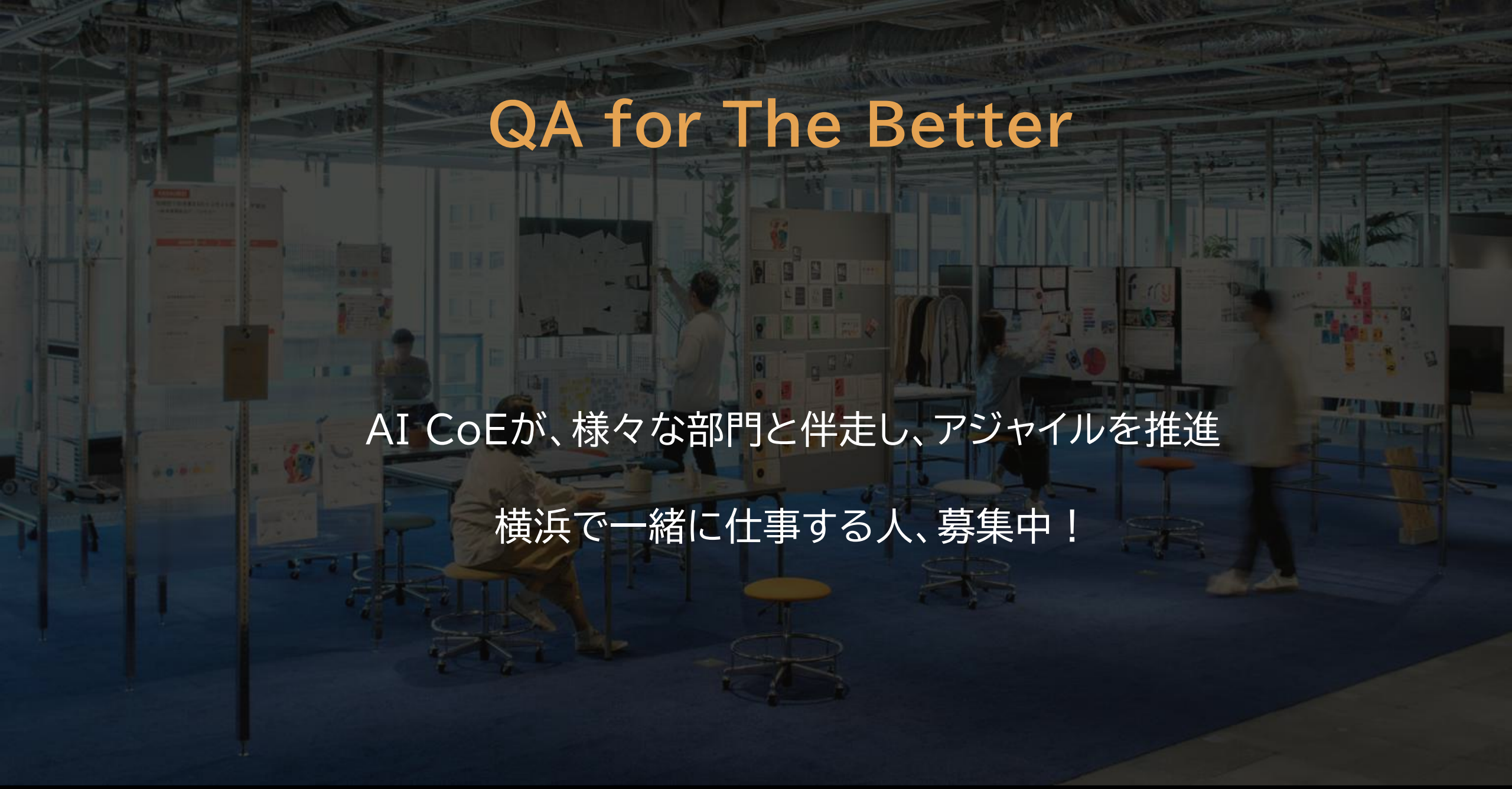
Open
カンファレンス

AIエージェント

AIによる開発支援

性能評価基盤

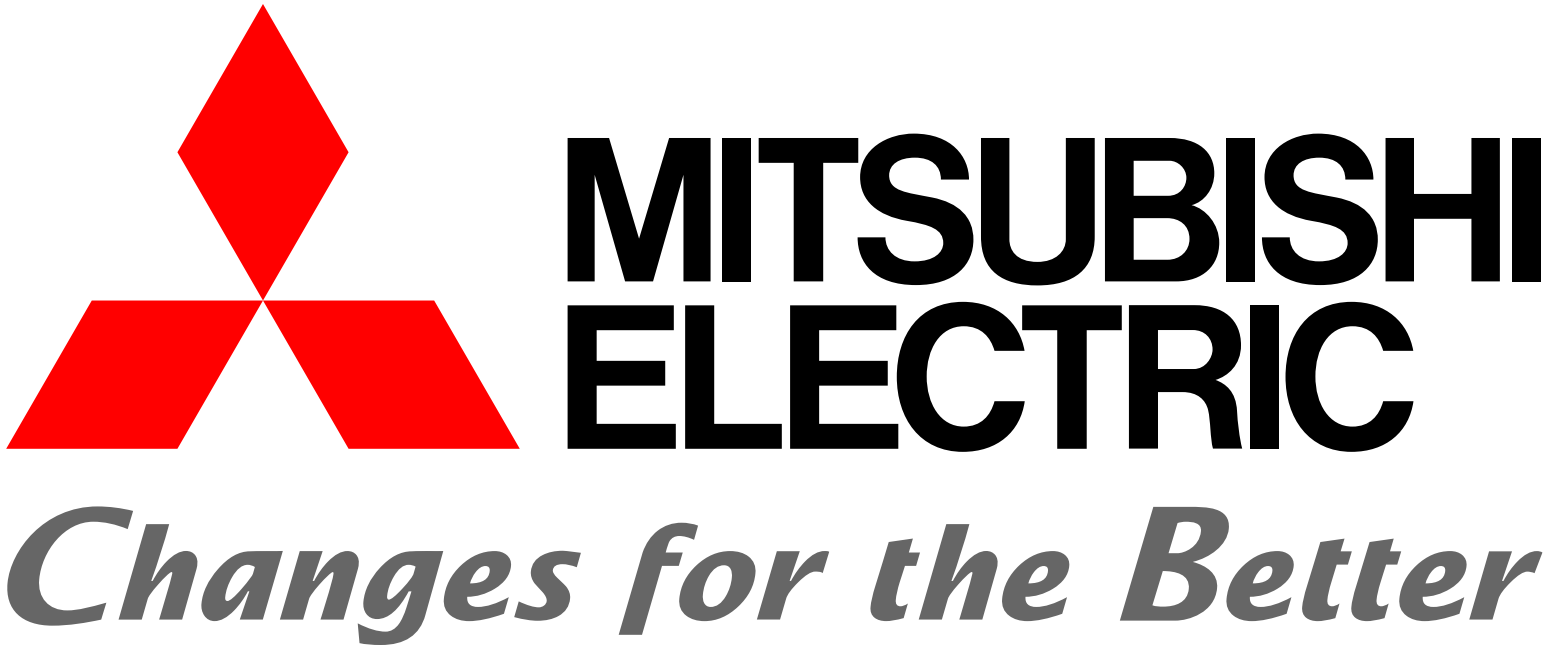
ガードレール



QA for The Better

AI CoEが、様々な部門と伴走し、アジャイルを推進

横浜で一緒に仕事する人、募集中！


MITSUBISHI
ELECTRIC
Changes for the Better