

要求に応じ信頼できるAIを 仕立て上げるための "Engineerable AI"技術

国立情報学研究所 アーキテクチャ科学研究系
准教授 石川 冬樹

JST未来社会創造事業 “Engineerable AI” プロジェクト

自己紹介

■国立情報学研究所 准教授

- ソフトウェア工学，特にディペンダビリティ：
形式手法，自動テスト生成，安全性論証など



■現在の主な研究プロジェクト

- 機械学習型AIシステムのディペンダビリティ
- 自動運転システムの安全性やサービス品質



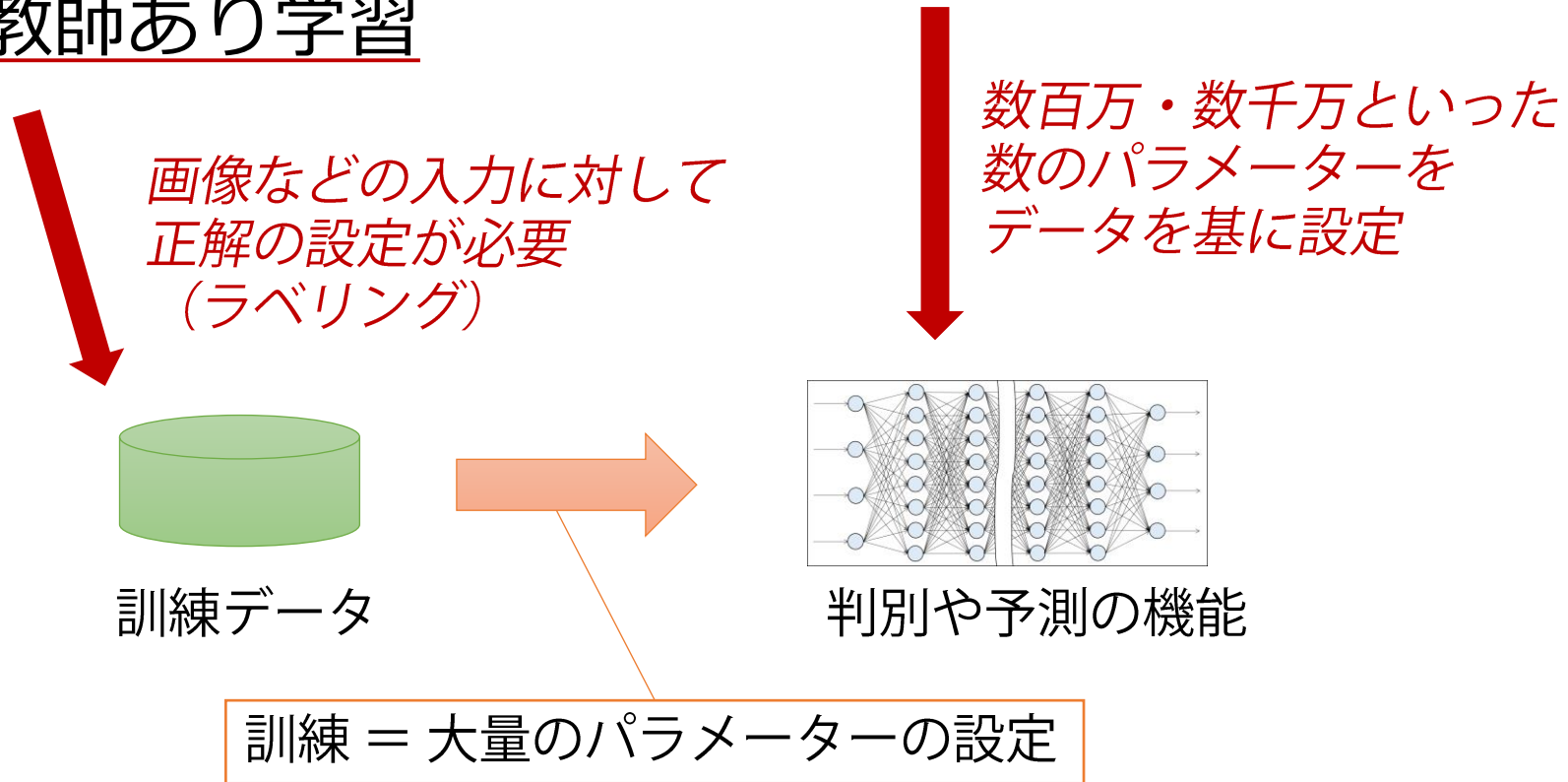
■産業界向け教育・実践研究

- トップエスイー，日科技連SQiP，電通大AISECなど
- 機械学習工学コミュニティ（MLSE研究会，QA4AI）

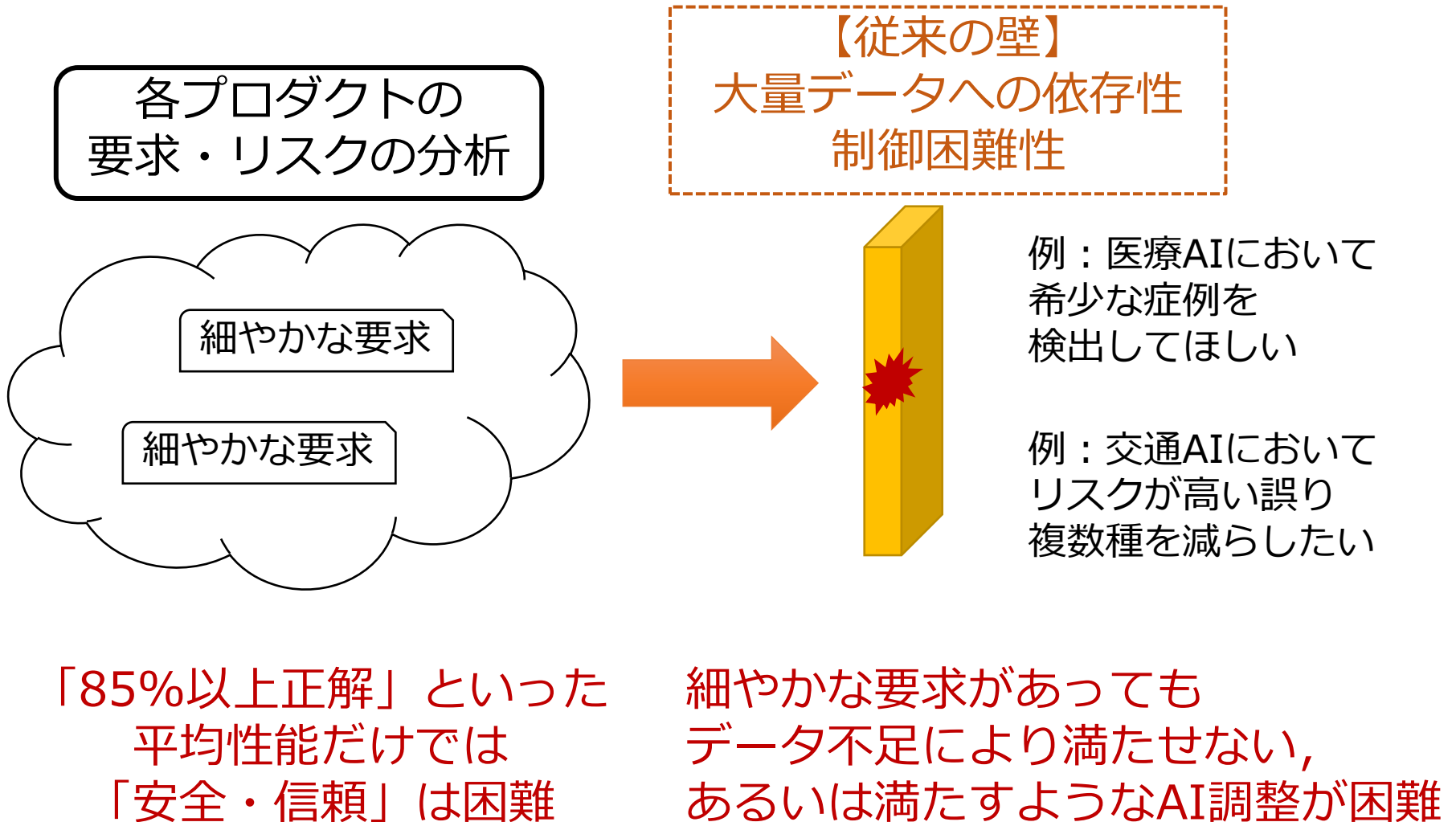


■現在の産業応用における一つの主流：

深層学習（ディープラーニング）技術を用いた 教師あり学習



動機：AIエンジニアリングの難しさ



Engineerable AI (eAI) プロジェクト

細やかなニーズに対しAIを仕立て上げられる
“Engineerable AI” 技術を創出し、
医療・交通の二分野において実証

■参考： eXplainable AI (XAI) 「説明できるAI」

■例：「今回は画像のここを見てガンだと判定した」といった
説明をAIの出力に付加

**安全・信頼の実現・向上・保証のためにエンジニアが
具体的な「アクション」をとっていくための技術に注力**

eAIプロジェクト：技術アプローチ

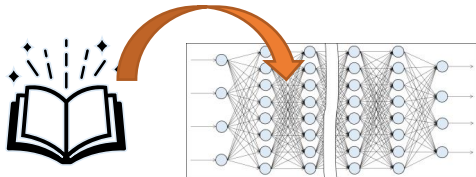
大量のデータが必要

最低でも10万以上のデータを
収集し整備することが必要

データだけでなく
「知識」をAIへ



例：知識・制約を埋め込んだ設計



少数の訓練データでも
要求を踏まえた挙動に

コア技術：AI構築
知識融合型深層学習技術

課題

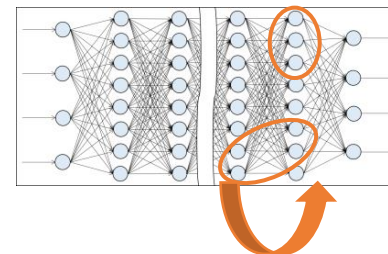
ニーズと対応づけた制御が不可能

特定状況での性能を狙って
上げたり直したりするのは困難

AI挙動の「知識」を得て
細やかな制御を可能に



例：バグ位置推定技術の応用



避けたい誤りに
影響する部分に
絞って修正

コア技術：AI継続改善
深層学習自動デバッグ技術

成果例 (1) スモールデータ・医療AI

■スモールデータ深層学習の発展・展開

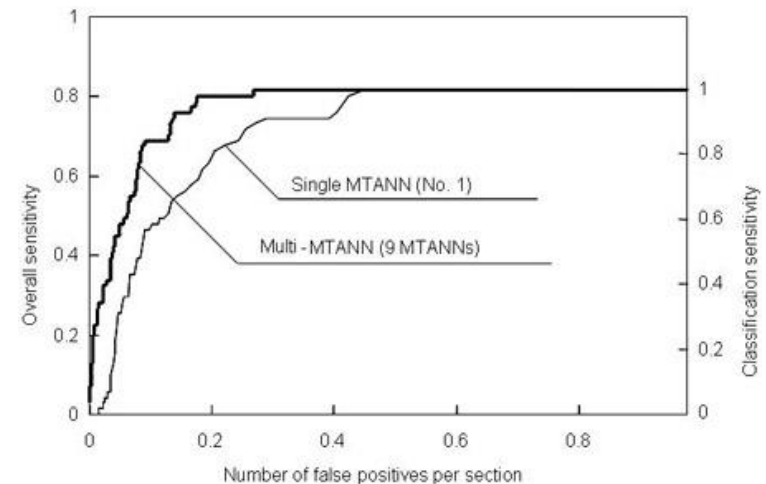
■例：MTANN（東工大 鈴木賢治研究室）

(Massive-Training Artificial Neural Network)

■「症例の検出・分類」ではなく「症例の模範画像への近さ」のようなものを学習

■100件程度の訓練データで高信頼な診断を実現

■症例の形状などの知識を
埋め込むことでさらに高信頼に



成果例 (2) AI修正・交通AI

■AI修正技術の提案・展開（NII石川冬樹研究室）

■例：DistRep



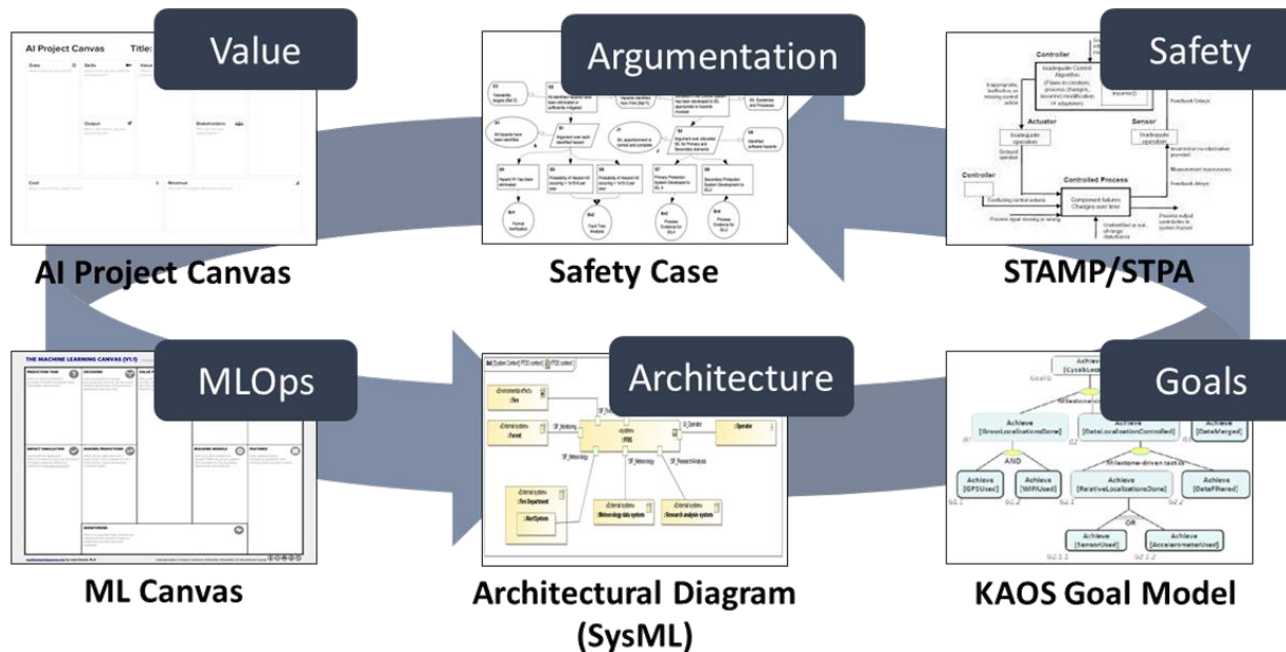
- 複数種別の誤りそれぞれを修正してマージ
- 産業界と定めた障害物検出ベンチマークにおいて、既存手法ではそもそも「狙った形で直す」ことが困難な設定でも総合リスクを下げることを実証

■例：NeuRecover

- 「デグレ」防止：更新により重要なケースで外すように変わってしまわないようにする
- 富士通における社会アプリケーションにおいて実証

成果例 (3) フレームワーク

- 要求・リスクの包括的なモデリングから、
AIの構築・修正技術や運用・監視につないでいく
(早大 鷲崎弘宜研究室)



Engineerable AI

データへの依存性・制御困難性を大きく低減



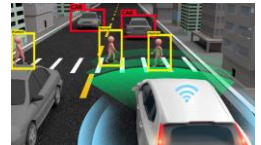
「ニーズに応じAIを工学的に仕立て上げる」技術へ

医療POC：
少数データでの
高信頼な診断



様々な応用領域での
高品質化・安定化へ

交通POC：
安全性向上の
ための改善制御



AI構築技術：
知識融合型深層学習

AI継続改善技術：
深層学習自動デバッグ

一体としてのeAIフレームワーク