高信頼機械学習システムのための Nバージョン構成法

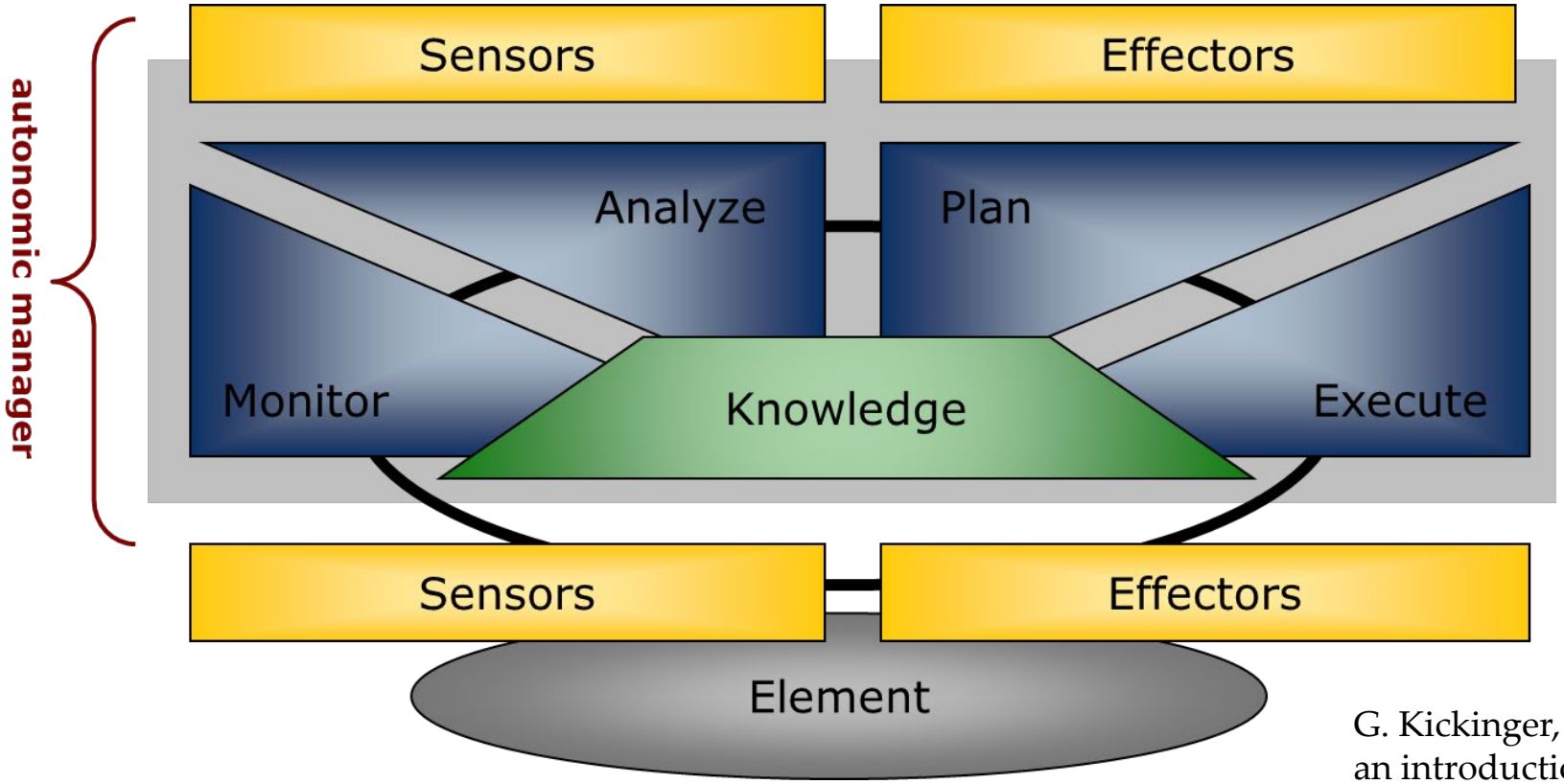
町田 文雄

筑波大学 システム情報系 准教授

電子情報通信学会 情報通信マネジメント研究会2025

20年前

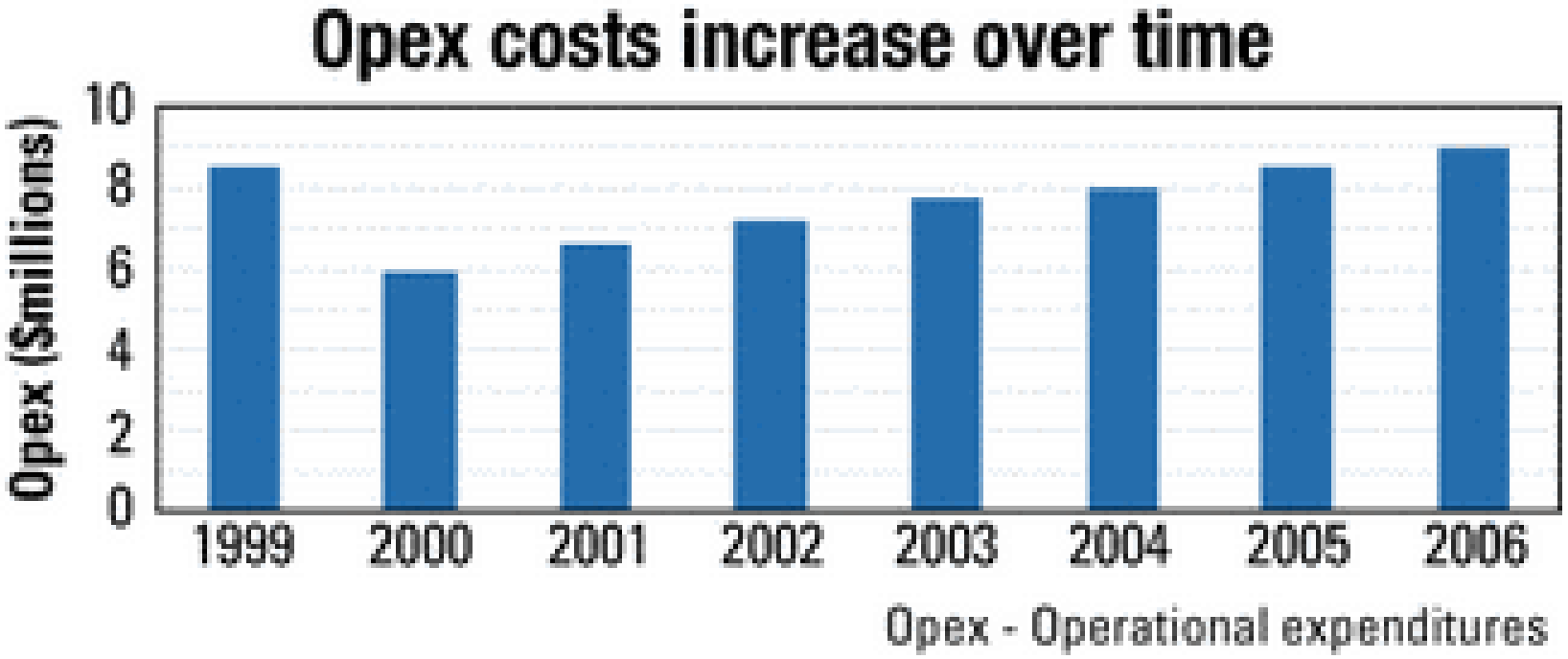
- 自律コンピューティング



G. Kickinger, Autonomic Computing an introduction

自律コンピューティングの目的

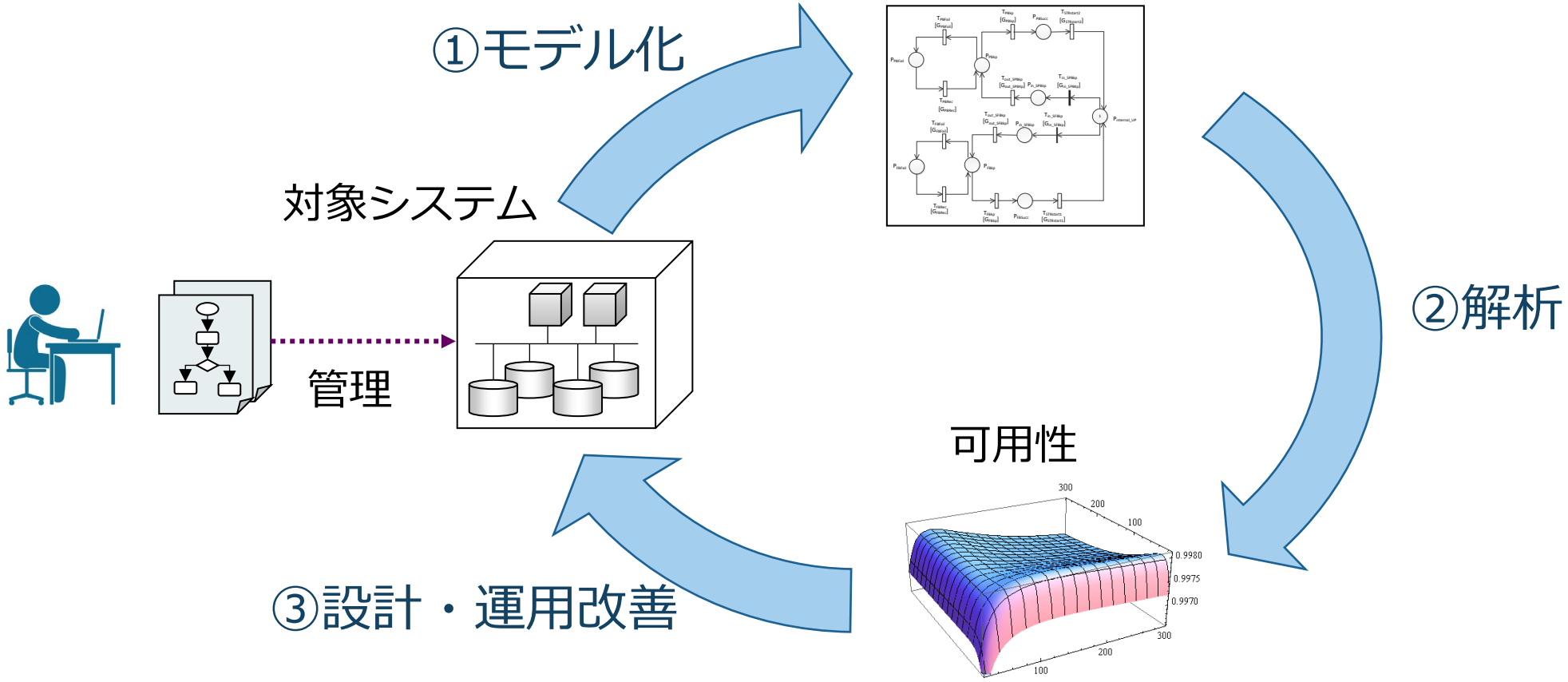
- 増加し続けるシステム運用管理コストの削減



https://www.lightwaveonline.com/network-design/article/16649427/controlling-opex-through-optical-technologies

可用性評価の課題

- 運用自動化の効果を可用性で定量評価



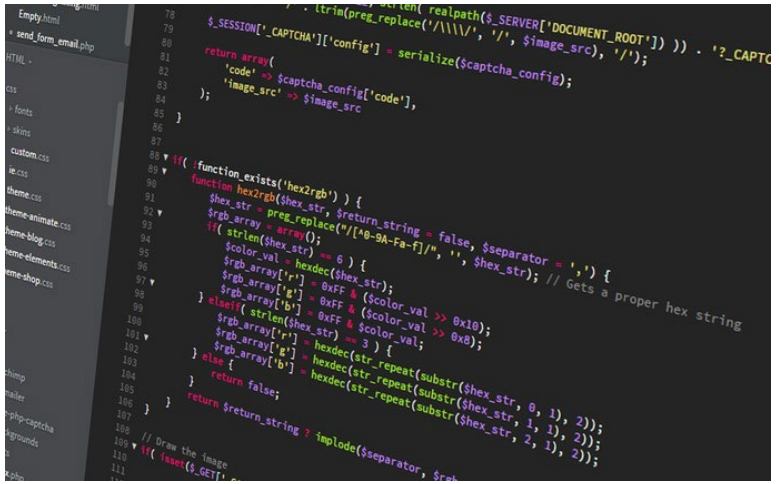
現在：機械学習システムの時代

- 機械学習を使ったシステムの産業応用が広がる

自動運転車

ヘルスケア

ソフトウェア開発



機械学習システムの障害

- 推論エラーが深刻なシステム障害や社会問題を引き起こす

Tesla in self-driving mode causes 8 vehicle crashes

https://bit.ly/3m9kJ8b

Facial recognition technology jailed a man for days

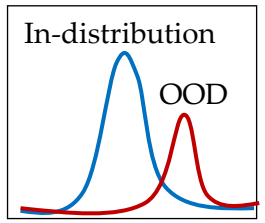
https://shorturl.at/JIob5

GPT-4V often made mistakes when describing the medical image

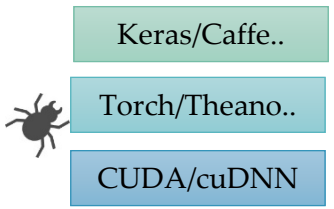
https://shorturl.at/xfqsh

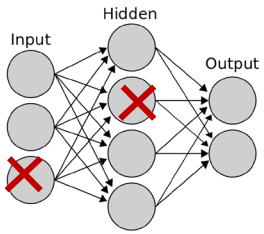
機械学習システムの信頼性リスク

- 機械学習モデルの推論エラー
  - 分布外入力（Out-Of-Distribution）
  - 敵対的サンプル（Adversarial Example）



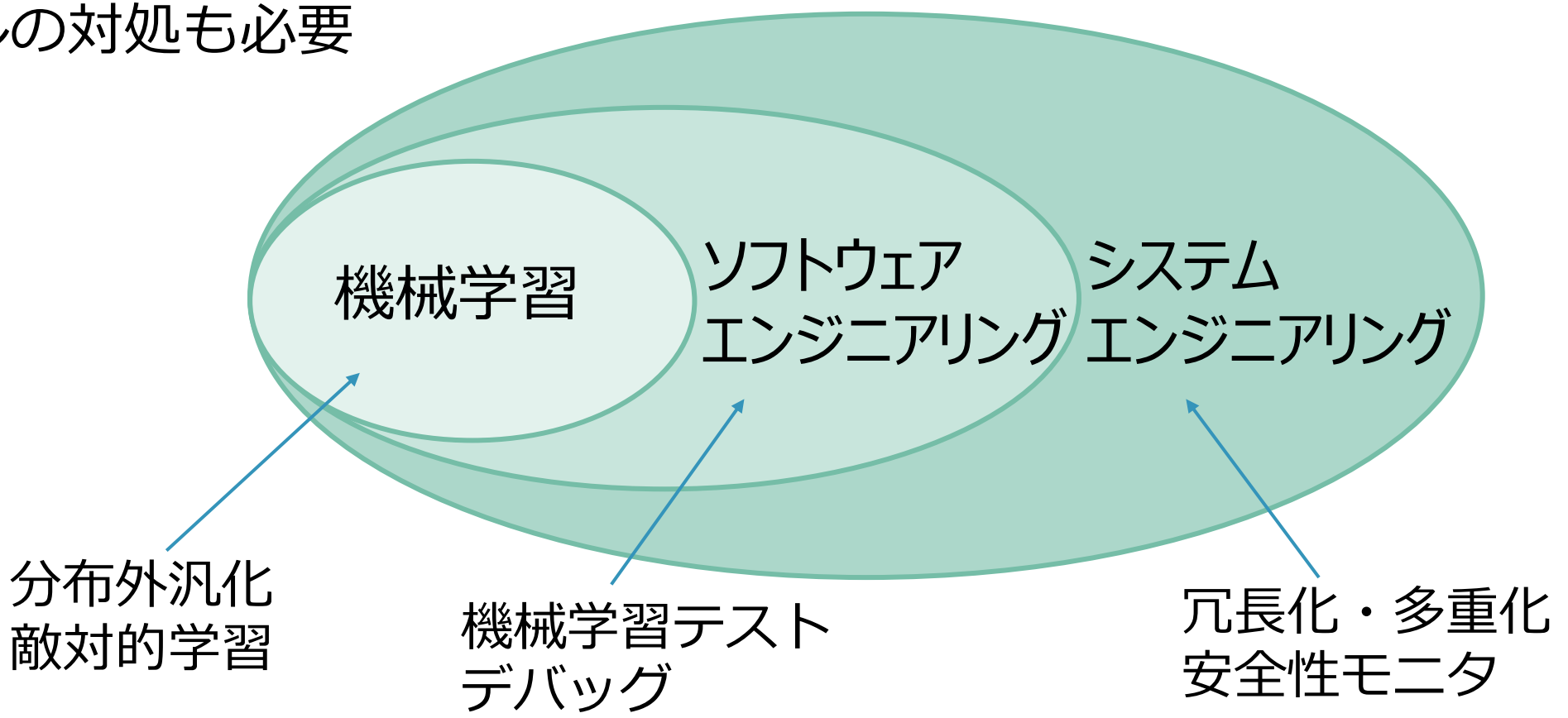
- ソフトウェアやハードウェアの障害
  - ソフトウェアバグ
  - 一時的メモリ障害（Soft Error）





機械学習システム高信頼化のアプローチ

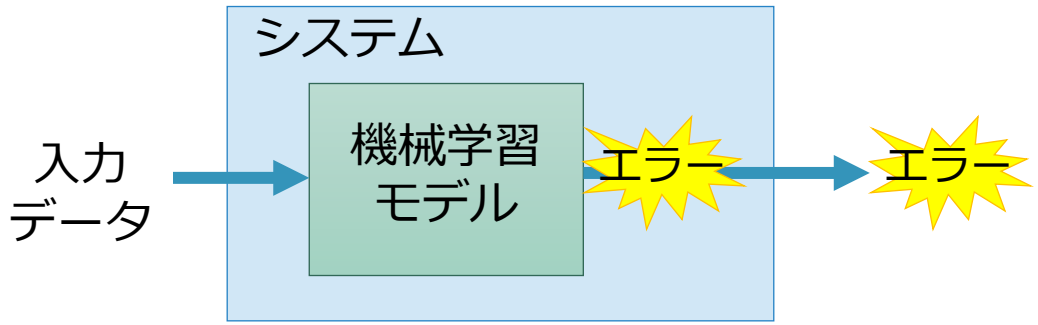
- 機械学習モデルの改良だけでなく、アプリケーションやシステムレベルの対処も必要



Nバージョン機械学習システム

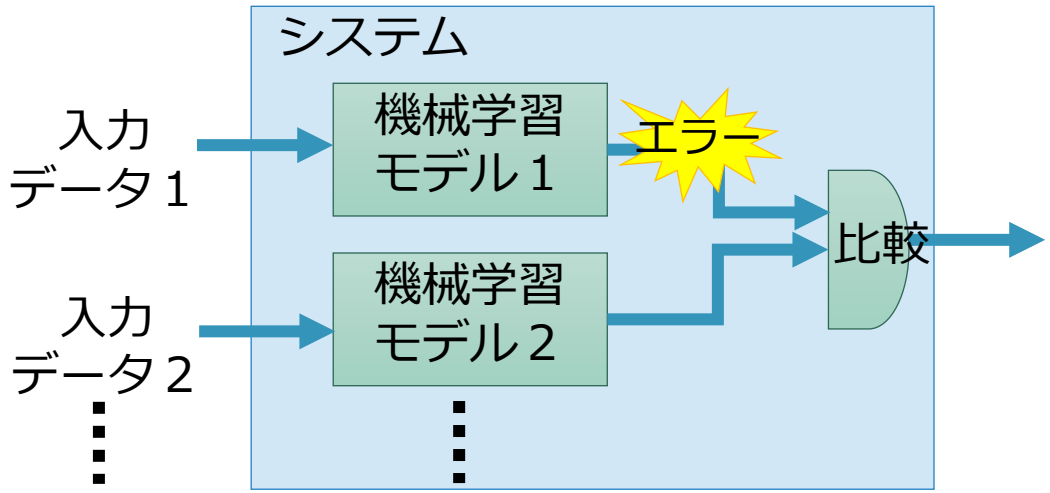
- 機械学習の推論を冗長化してエラー出力を抑える

単一の機械学習モデルを利用する場合



システムの外にエラーがそのまま出てしまう

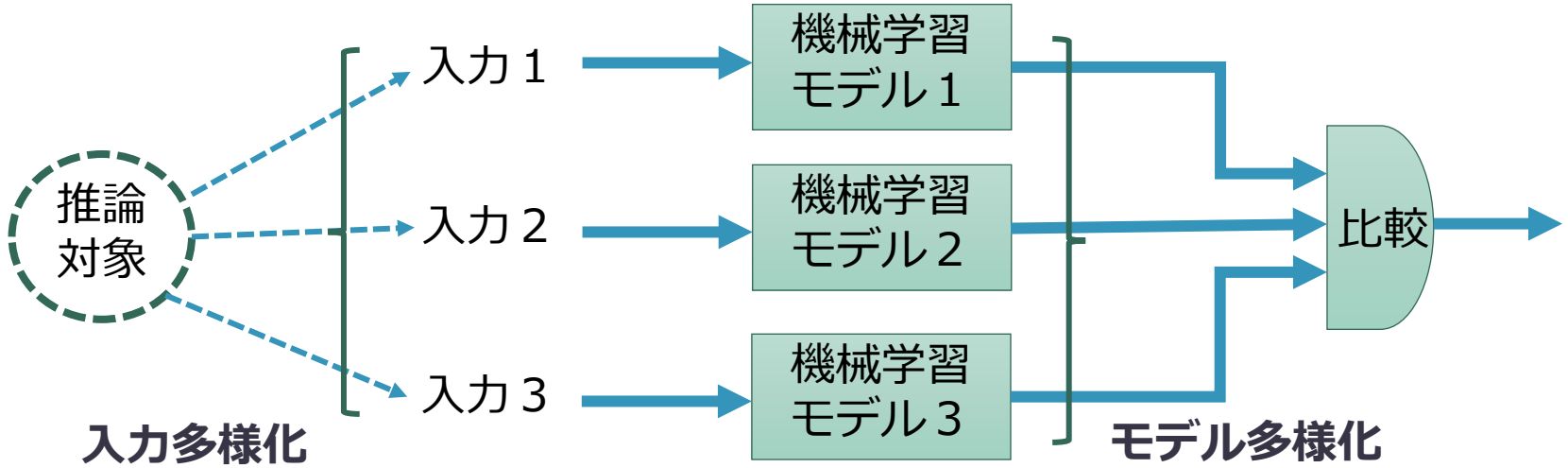
Nバージョン機械学習システムの場合



複数の結果を比較してシステムのエラー出力を抑える

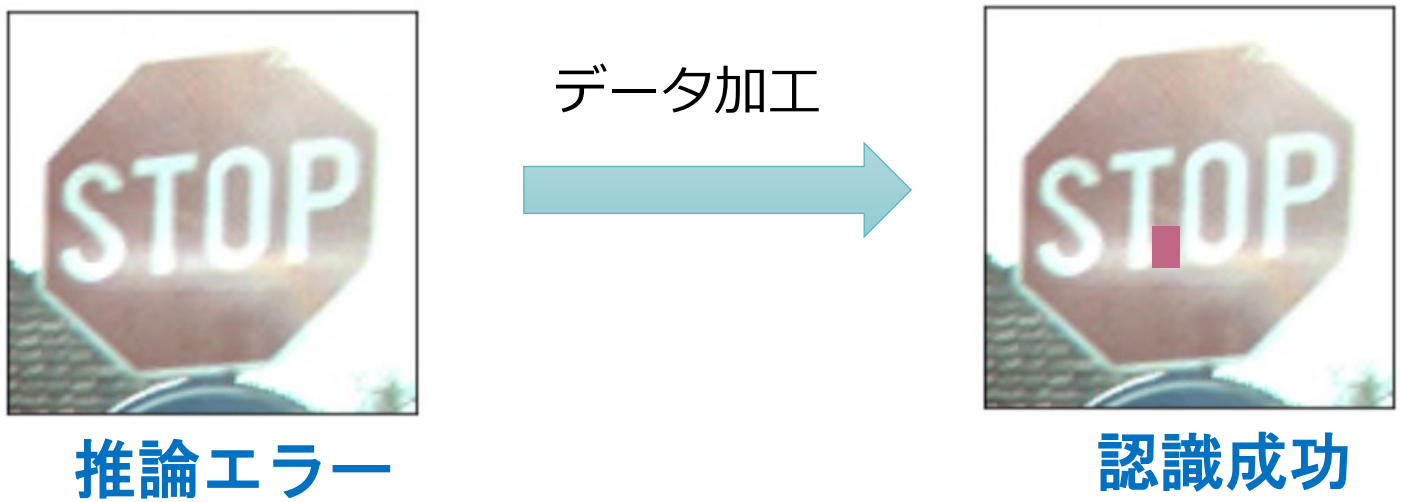
モデルの多様化と入力の多様化

- 複数のモデルが同時にエラーを出力しないように
- モデル多様化
  - 異なる機械学習アルゴリズムや学習データを使ってモデルを作成する
- 入力多様化
  - 同じ推論対象に対する異なる入力データを利用する



入力データ多様化

- 機械学習モデルは入力データの違いに敏感
  - 入力データのわずかな加工で機械学習モデルを騙せる（敵対的サンプル）
    → 逆も起こり得る



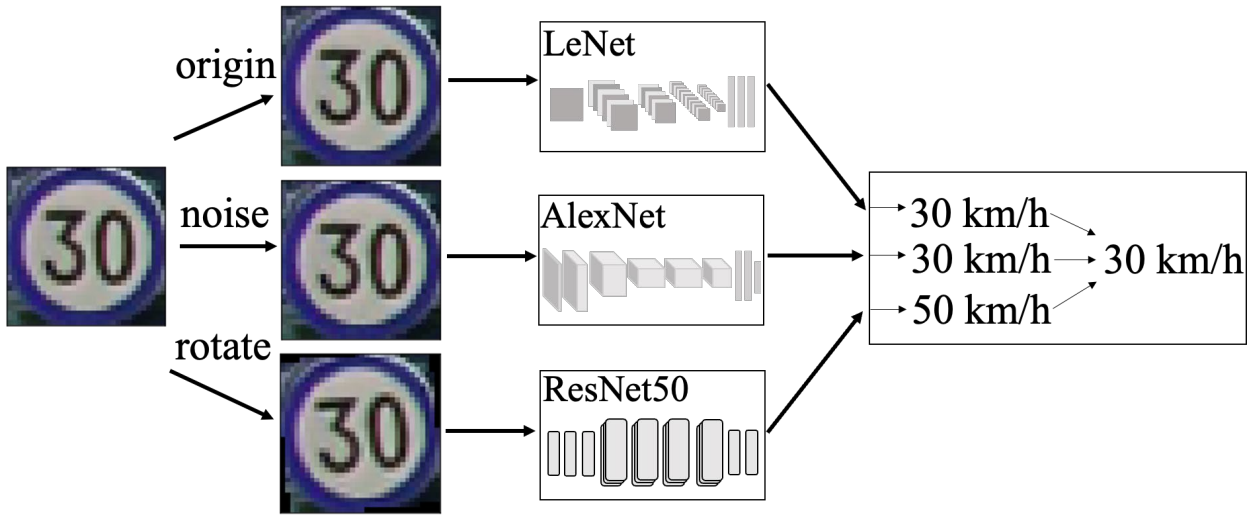
推論エラー　　**認識成功**

Nバージョンプログラミングとの違い

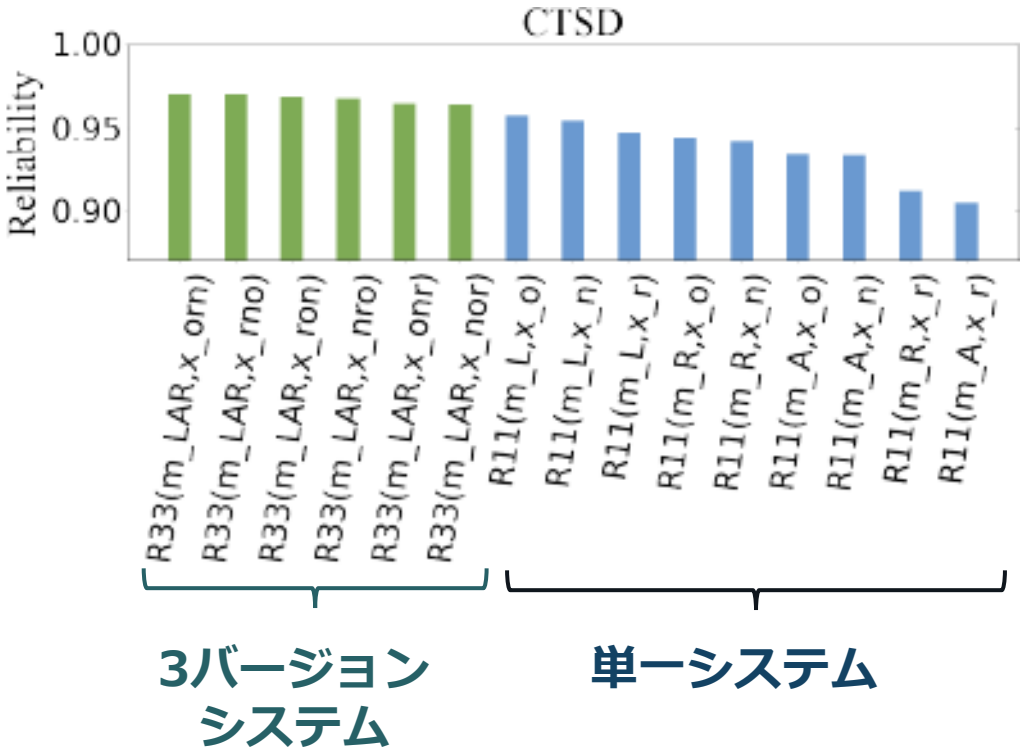
	Nバージョンプログラミング	Nバージョン機械学習システム
対象	プログラム（仕様に基づいて開発される）	機械学習モデル（訓練データから学習される）
対処する問題	ソフトウェアのバグ	誤判断
構成要素	2つ以上の機能的に等価なプログラム	**1つ以上**の同じタスクを実行する機械学習モデル
多様化手法	開発チーム、プログラミング言語、ライブラリ、ツール	学習アルゴリズム、ハイパーパラメータ、学習データ、**入力データ**
導入コスト	高い	**低い**

画像分類システムでの応用

- 3バージョン交通標識分類システム
  - 3つの加工入力データと3つの異なるニューラルネットワークで構成



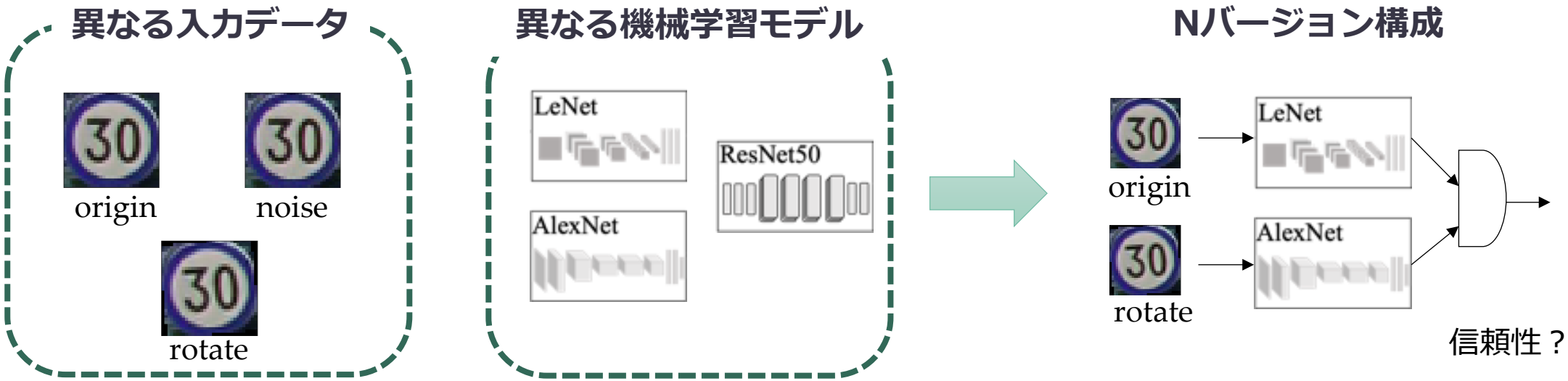
[Q. Wen, et al. ISSRE2023]



信頼性モデルとアーキテクチャの選択

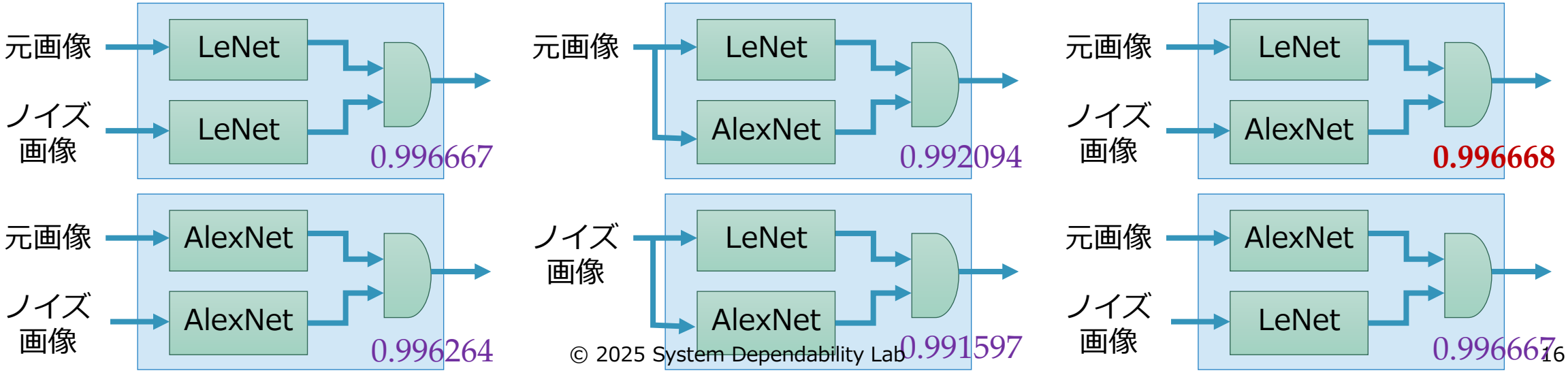
Nバージョン構成選択の問題

- 複数のモデルと異なる入力データが与えられたとき、どのような構成が最も信頼性を向上させるか？
  - どのモデルを使うか？
  - どの入力データをどのモデルに与えるか？



経験則

- 画像分類システムの信頼性はNバージョン構成によって異なる
  - データセット：MNIST（手書き数字0～9）
  - 機械学習モデル：深層ニューラルネットワーク（LeNet, AlexNet）
  - 入力データ多様化：元画像、ノイズ追加画像
  - 比較器：不一致の場合は出力しない



Nバージョン構成の信頼性モデル

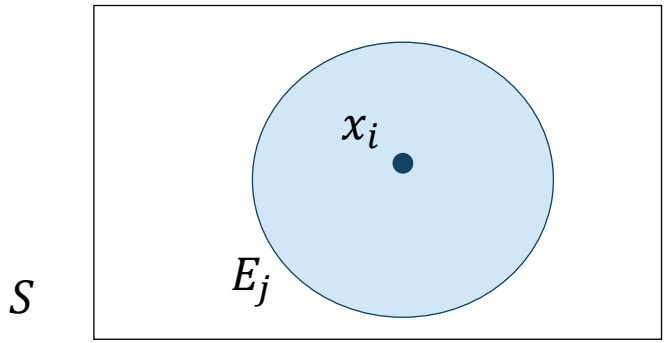
- 異なる入力データと異なるモデルの組み合わせで信頼性が異なる
  → 理論的にどこまで解析できるか？
- 分類システムを対象に信頼性モデルを考える
  - 問題設定

入力データ：同一の対象に対して**2つ**の異なる入力データを利用可能
機械学習モデル：同一の分類タスクを行う**2つ**の分類モデルを利用可能
比較器のルール：出力結果が**一致する場合**にのみその結果を出力する
信頼性：システムが誤った結果を出力しない確率

1バージョン構成の信頼性

- 記法
  - 入力データ：$x_i, i = \{1,2, \dots\}$
  - 機械学習モデル：$m_j, j = \{a, b, \dots\}$
  - 入力データの標本空間：S
  - 機械学習モデルm_jがエラーとなる入力データの集合： $E_j \subset S$
- 機械学習モデルm_jと入力データx_iを組み合わせた場合の信頼性

$$1 - \mathrm{P}[x_i \in E_j]$$



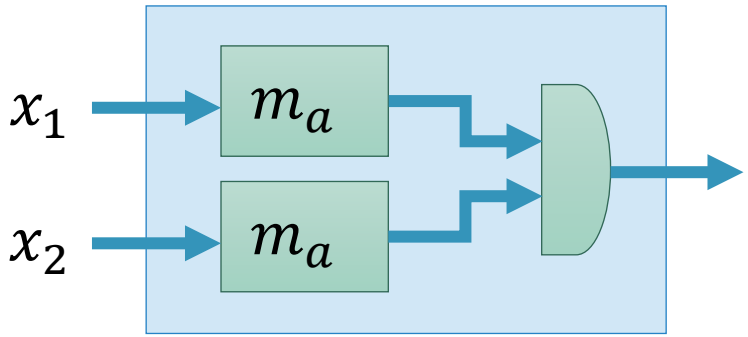
2バージョン構成のアーキテクチャ

6通り

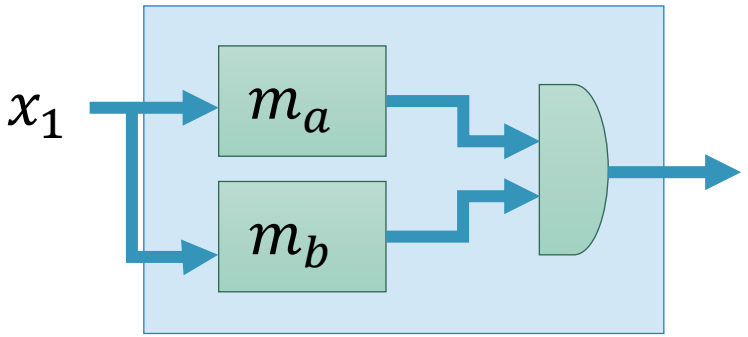
単一モデル二重入力
(Single model double input: SMDI)

二重モデル単一入力
(Double model single input: DMSI)

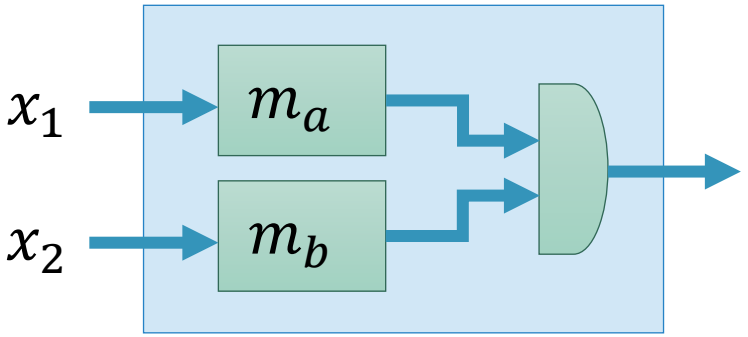
二重モデル二重入力
(Double model double input: DMDI)



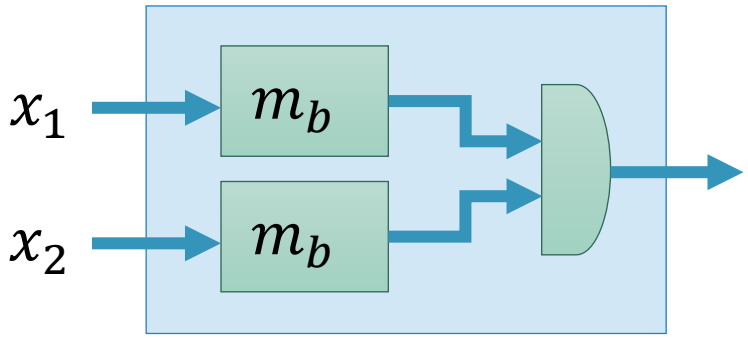
$SMDI(m_a; x_1, x_2)$



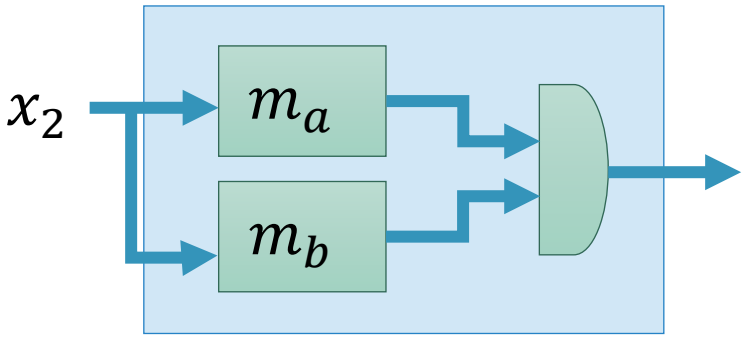
$DMSI(m_a, m_b; x_1)$



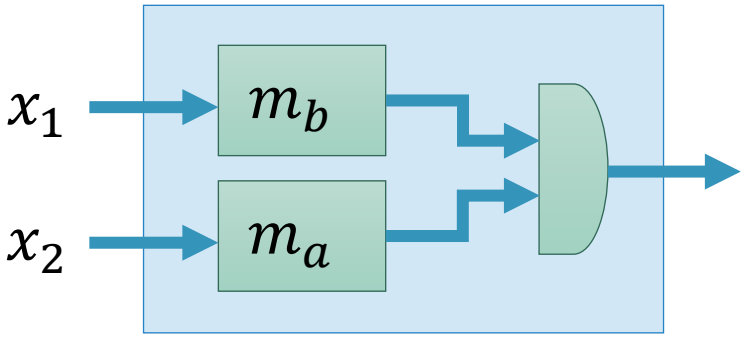
$DMDI(m_a; x_1, m_b; x_2)$



$SMDI(m_b; x_1, x_2)$



$DMSI(m_a, m_b; x_2)$

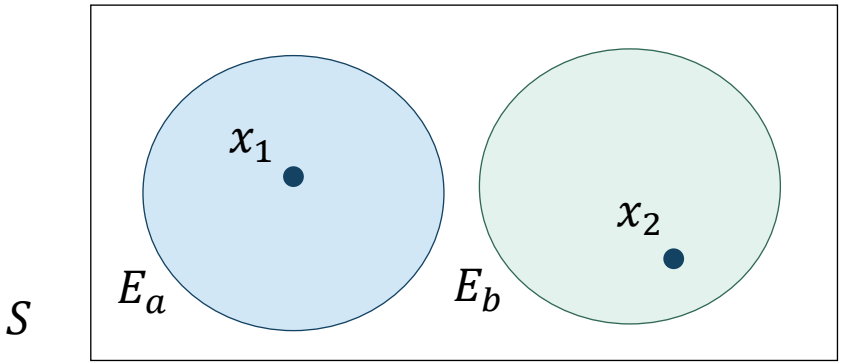


$DMDI(m_a; x_2, m_b; x_1)$

2バージョン構成の信頼性

- エラー確率$\mathrm{P}[x_i \in E_j]$が独立な場合
  - 2バージョン構成のエラー確率は1バージョン構成のエラー確率の積

    $1 - \mathrm{P}[x_1 \in E_a] \cdot \mathrm{P}[x_2 \in E_b]$



- 実際はエラー確率は独立ではない
  - エラー集合E_jは共通部分を持つ可能性がある
  - 入力データx_iの分布は同一とは限らない

多様性指標の導入

- 2つの機械学習モデルを用いる場合、エラー集合に依存関係がある

エラーの共通部分(モデル類似度)

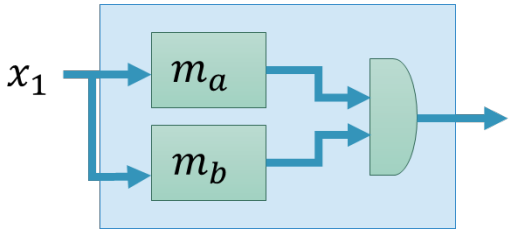
機械学習モデルm_a, m_bが誤出力となる入力データx_iの標本空間Sの部分集合をそれぞれE_a, E_bとし、エラーの共通部分$\alpha_{b|a,i} \in [0,1]$を以下で定義する。

$$\alpha_{b|a,i} = P[x_i \in E_b | x_i \in E_a] = \frac{P[x_i \in E_a \cap E_b]}{P[x_i \in E_a]}.$$

ただし$P[x_i \in E_a] > 0$とする。

- 二重モデル単一入力システム(DMSI)の信頼性

$$R_{DMSI_{a\cap b,1}} = 1 - \alpha_{b|a,1} \cdot P[x_1 \in E_a]$$



多様性指標の導入2

- 2つの入力データを用いる場合、2つのデータ分布は独立ではない

エラーの共起度(入力類似度)

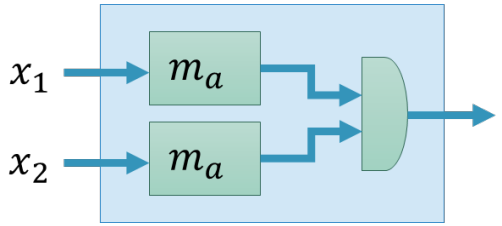
同じ標本空間Sから抽出した機械学習モデルm_j に対する入力データをx_1, x_2とし、エラーの共起度$\beta_{j,2|1} \in [0,1]$を以下のように定義する。

$$\beta_{j,2|1} = Pr\left[x_2 \in E_j | x_1 \in E_j\right] = \frac{P\left[x_1 \in E_j, x_2 \in E_j\right]}{P\left[x_1 \in E_j\right]}.$$

ただし$P\left[x_1 \in E_j\right] > 0$とする。

- 単一モデル二重入力システム(SMDI)の信頼性

$$R_{SMDI_{a,1\cap 2}} = 1 - \beta_{a,2|1} \cdot P[x_1 \in E_a]$$



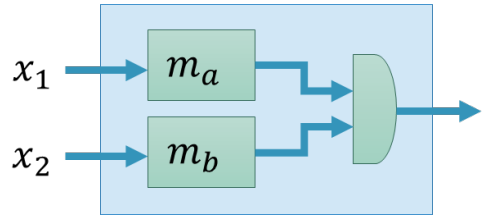
二重モデル二重入力システムの信頼性

- 二重モデル二重入力(DMDI)システムの信頼性はモデルの依存関係と入力データ分布の重なりの両方の影響を受ける
- $DMDI(m_a; x_1, m_b; x_2)$の信頼性

$$R_{DMDI_{a,1\cap b,2}} = 1 - \left[\alpha_{b,2|a,1\cap 2} \cdot \beta_{a,2|1} + \alpha_{b,2|a,1\cap \bar{2}} \cdot \left(1 - \beta_{a,2|1}\right)\right] \cdot P[x_1 \in E_a]$$

$$\alpha_{b,2|a,1\cap 2} = P[x_2 \in E_b | x_2 \in E_a, x_1 \in E_a]$$

$$\alpha_{b,2|a,1\cap \bar{2}} = P[x_2 \in E_b | x_2 \in \overline{E_a}, x_1 \in E_a]$$



入力類似度とモデル類似度に関連したパラメータで特徴づけられる

信頼性モデルから導かれる性質

- モデルの類似度と入力データの類似度が条件付き独立と仮定する場合
  - $\alpha_{b,2|a,1\cap 2} = \alpha_{b|a,2}\ and\ \alpha_{b,2|a,1\cap\overline{2}} = P[x_2 \in E_b | x_2 \in \overline{E_a}]$

性質

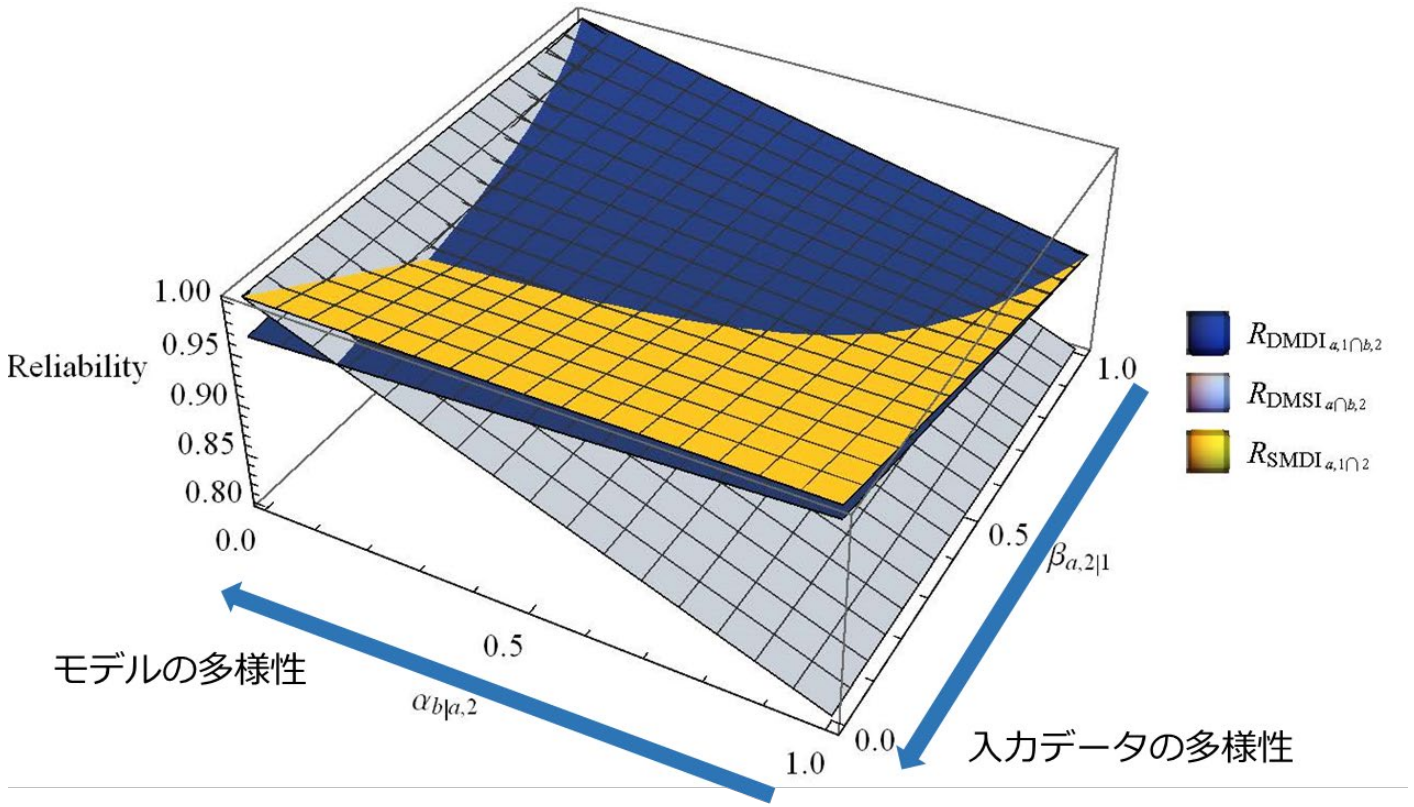
$DMDI_{a,1\cap b,2}$、$DMSI_{a\cap b,2}$ 、 $SMDI_{a,1\cap 2}$の中で最も信頼性が高いアーキテクチャは$\alpha_{b|a,2}$ と $\beta_{a,2|1}$の値から以下で求まる。

$$\begin{cases} DMSI_{a\cap b,2}, & if\ \omega(\alpha_{b|a2'},\beta_{a,2|1}) - \alpha_{b|a2'} \cdot P[x_2 \in E_a] \geq 0\ and \quad \beta_{a,2|1} \geq \alpha_{b|a2'} \cdot \dfrac{P[x_2 \in E_b]}{P[x_1 \in E_a]}, \\ SMDI_{b,1\cap 2}, & if\ \omega(\alpha_{b|a2'},\beta_{a,2|1}) - \beta_{a,2|1} \cdot P[x_1 \in E_a] \geq 0\ and \quad \beta_{a,2|1} \leq \alpha_{b|a2'} \cdot \dfrac{P[x_2 \in E_b]}{P[x_1 \in E_a]}, \\ DMDI_{a,1\cap b,2}, & otherwise. \end{cases}$$

ただし$\omega(\alpha_{b|a,2},\beta_{a,2|1}) = \frac{P[x_1\in E_a]}{1-P[x_2\in E_a]} \cdot [\alpha_{b|a,2} \cdot (\beta_{a,2|1} - P[x_2 \in E_a]) + P[x_2 \in E_b] \cdot (1-\beta_{a,2|1})]$

信頼性モデルから導かれる性質（続き）

- モデルの類似度と入力データの類似度が条件付き独立と仮定する場合



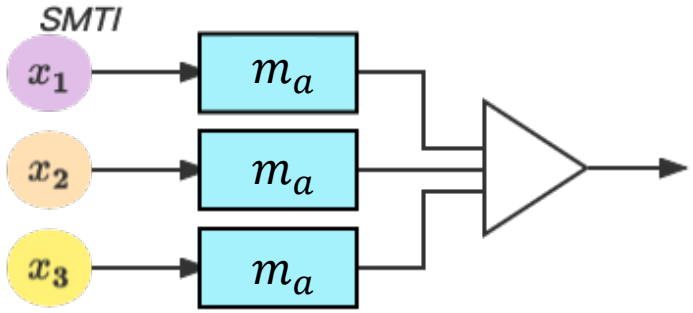
$\alpha_{b|a,2}$ と $\beta_{a,2|1}$のバランスで最も良いアーキテクチャが変わる

現実的なモデルの類似度と入力データの類似度の範囲では $DMDI_{a,1\cap b,2}$が好ましい選択

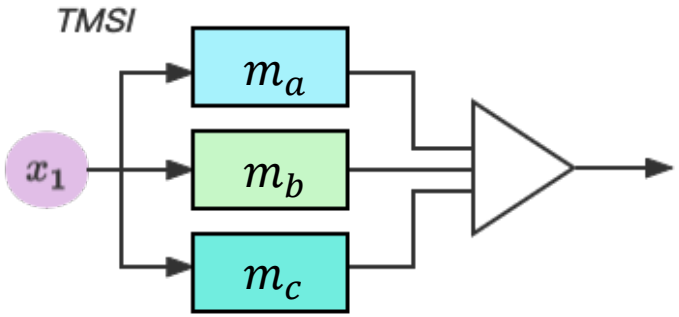
3バージョン構成のアーキテクチャ

- 3つの推論結果の多数決で最終出力を決定する

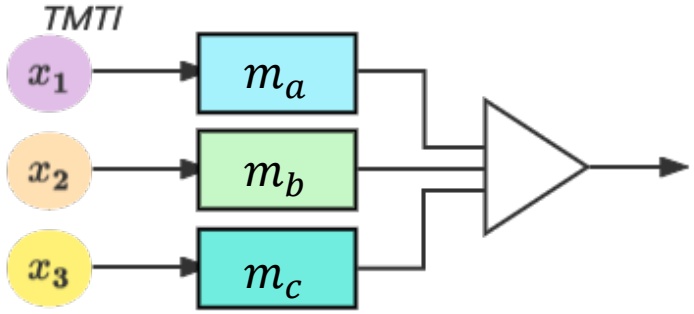
単一モデル三重入力
(Single model triple input: SMTI)



三重モデル単一入力
(Triple model single input: TMSI)

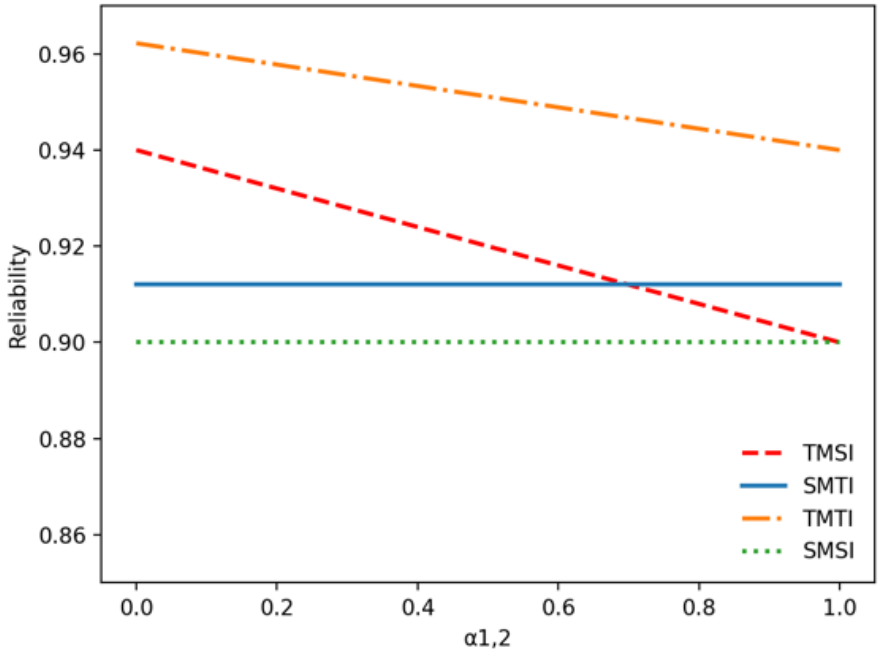


三重モデル三重入力
(Triple model triple input: TMTI)

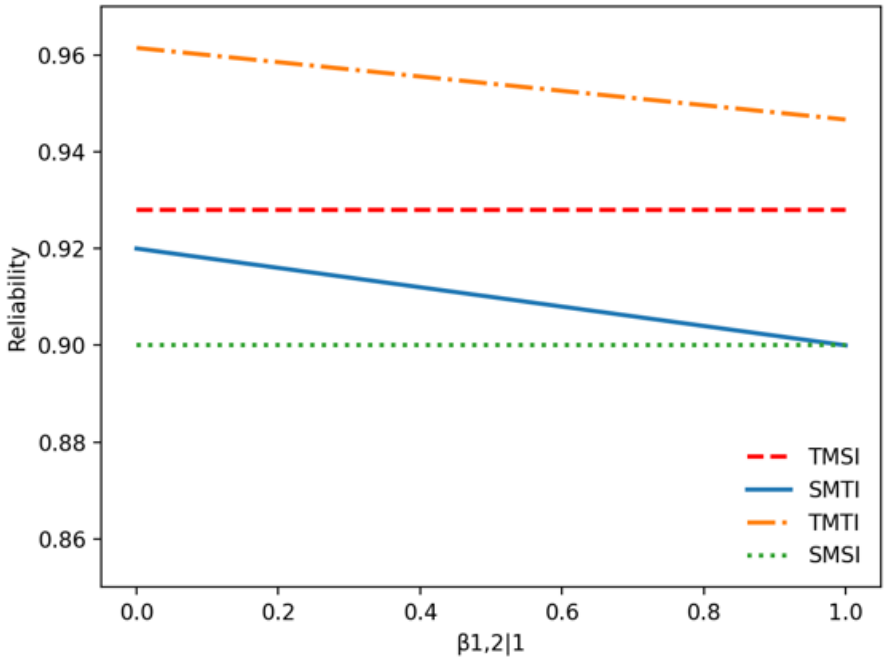


3バージョンアーキテクチャの信頼性比較

モデルの類似度$\alpha_{b|a,1}$の影響を評価



入力の類似度$\beta_{a,2|1}$の影響を評価

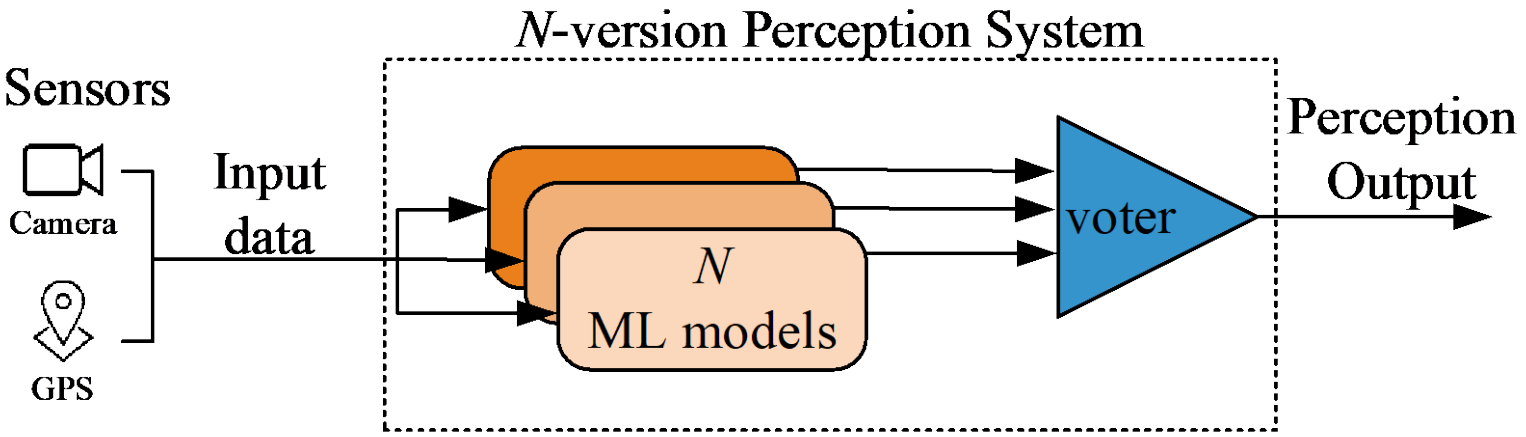


モデルの類似度はTMSIとTMTIの信頼性に影響を与え、
入力の類似度はSMTIとTMTIの信頼性に影響を与える

機械学習システム高信頼化の応用例

自動運転向け物体検出システムへの応用

- 物体検出モデルが障害や攻撃によって精度が低下すると仮定
  - 正常状態　→　劣化状態　→　故障状態
- N個の物体検出モデルを用いて信頼性を向上させる

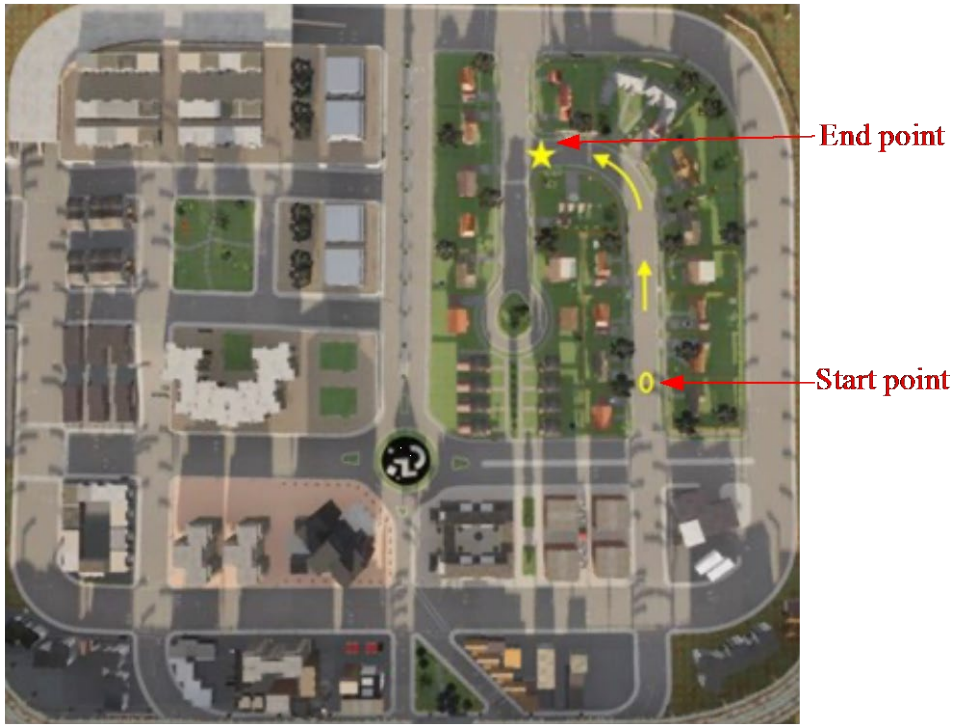


[Q. Wen, et al. AISafety2024]

劣化状態では正しい物体検出ができない

自動運転の安全性評価

- 評価環境
  - 自動運転のシミュレータCarla
  - フレームワークOpenCDA
- オブジェクト検出モデル
  - YOLOv5s6, YOLOv5m6, YOLOv5l6
- 評価指標
  - 衝突率
  - 最初の衝突までのフレーム数

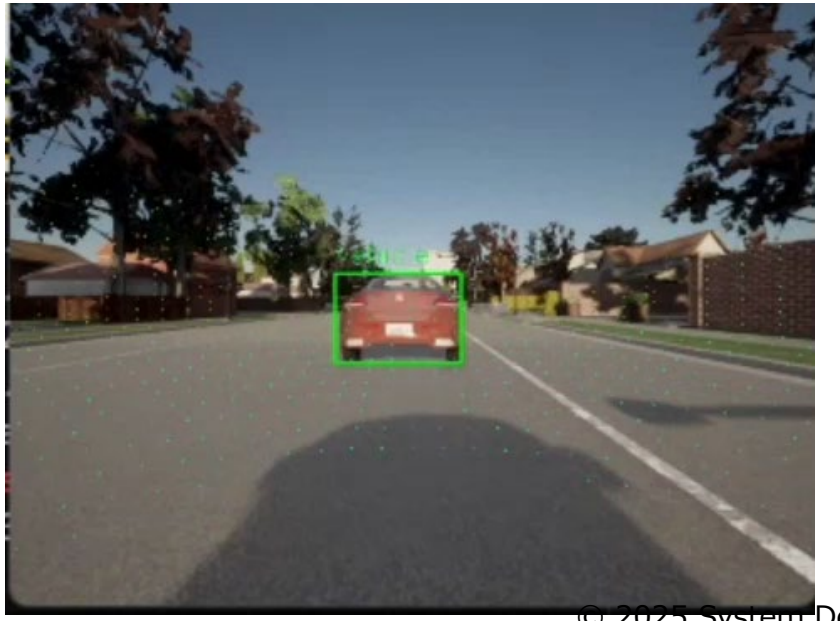


Town03 of the CARLA simulator

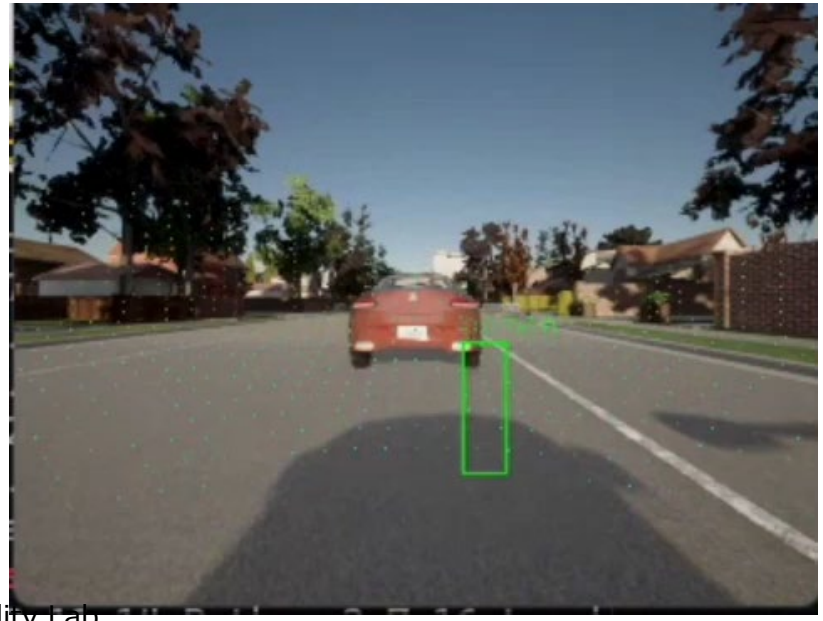
Carlaによる実験

- 故障注入により劣化モデルを生成
  - PyTorchFIでYOLOv5のパラメータをランダムに変更
- 劣化モデルによる自動運転では衝突事故が発生

正常モデルでの運転



劣化モデルでの運転



3バージョン物体検出システムの評価

- 3バージョン構成であれば、1つのモデルが劣化状態になっても自動運転の安全性を維持できる

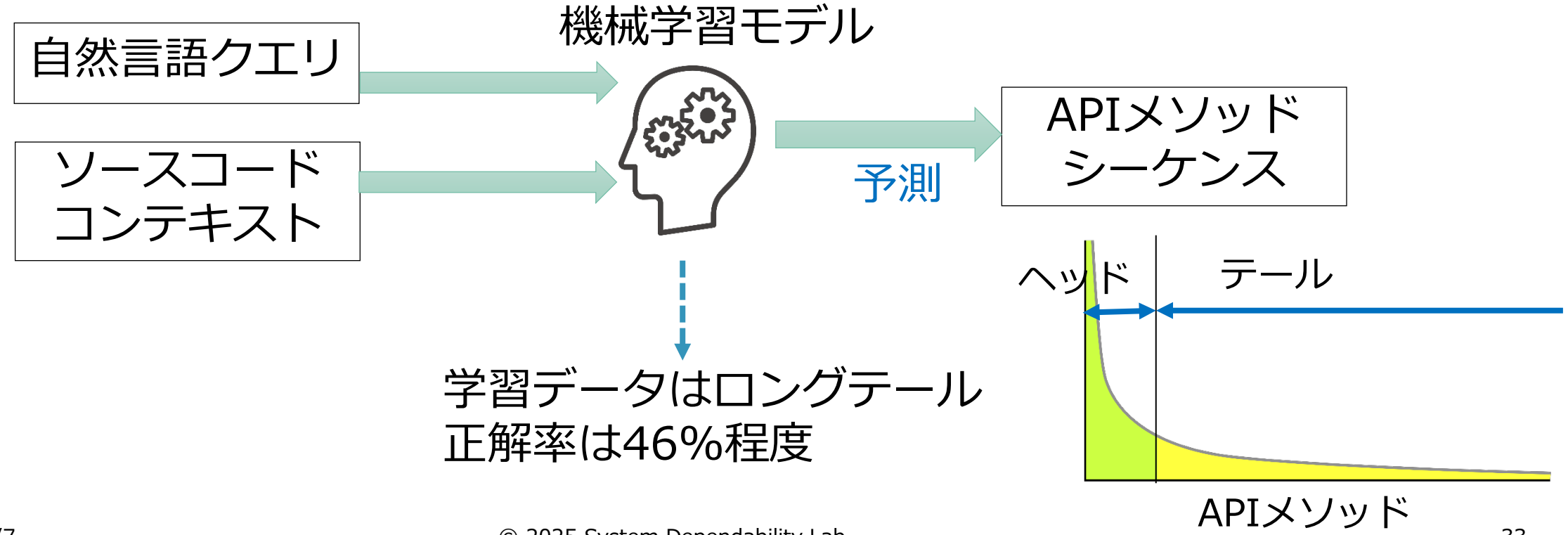
System state	YOLO Model	1st collision frame	Total frames	Collision rate%	# Collisions
Three-version					
(3,0,0)	v5s, v5m, v5l	NA	682	0	0/10
(2,1,0)	v5s, v5m, v5m_FI	NA	693	0	0/10
(2,1,0)	v5s, v5m, v5s_FI	NA	682	0	0/10
(1,2,0)	v5s, v5s_FI, v5m_FI	272	666	28.82	5/10
(1,2,0)	v5m, v5s_FI, v5m_FI	335	654	33.08	7/10
(0,3,0)	v5s_FI, v5m_FI, v5l_FI	187	643	57.00	8/10

劣化したモデルの数

1つのモデルの劣化であれば衝突を防げる

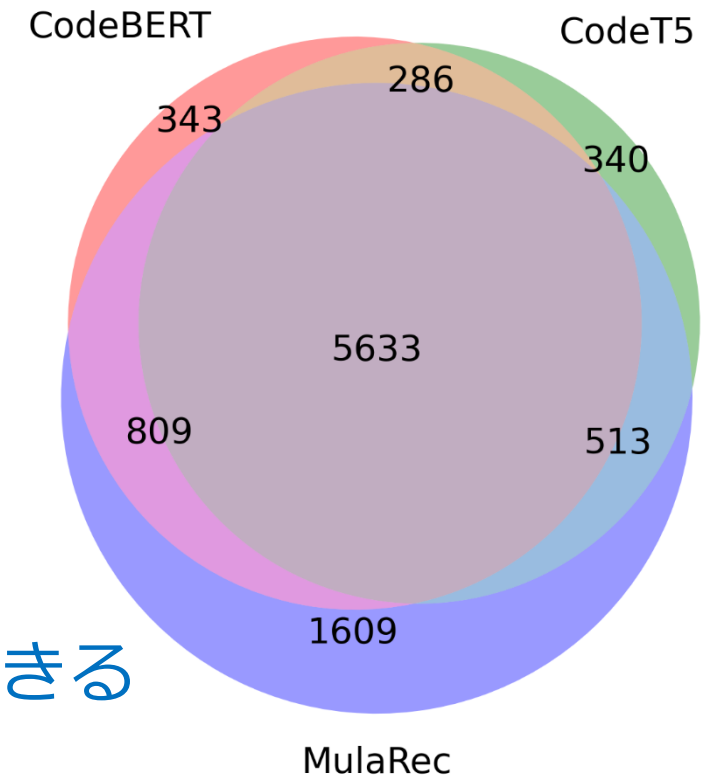
ソフトウェア開発自動化への応用

- APIメソッドシーケンス推薦
  - IDEなどで提供されるプログラム開発支援機能



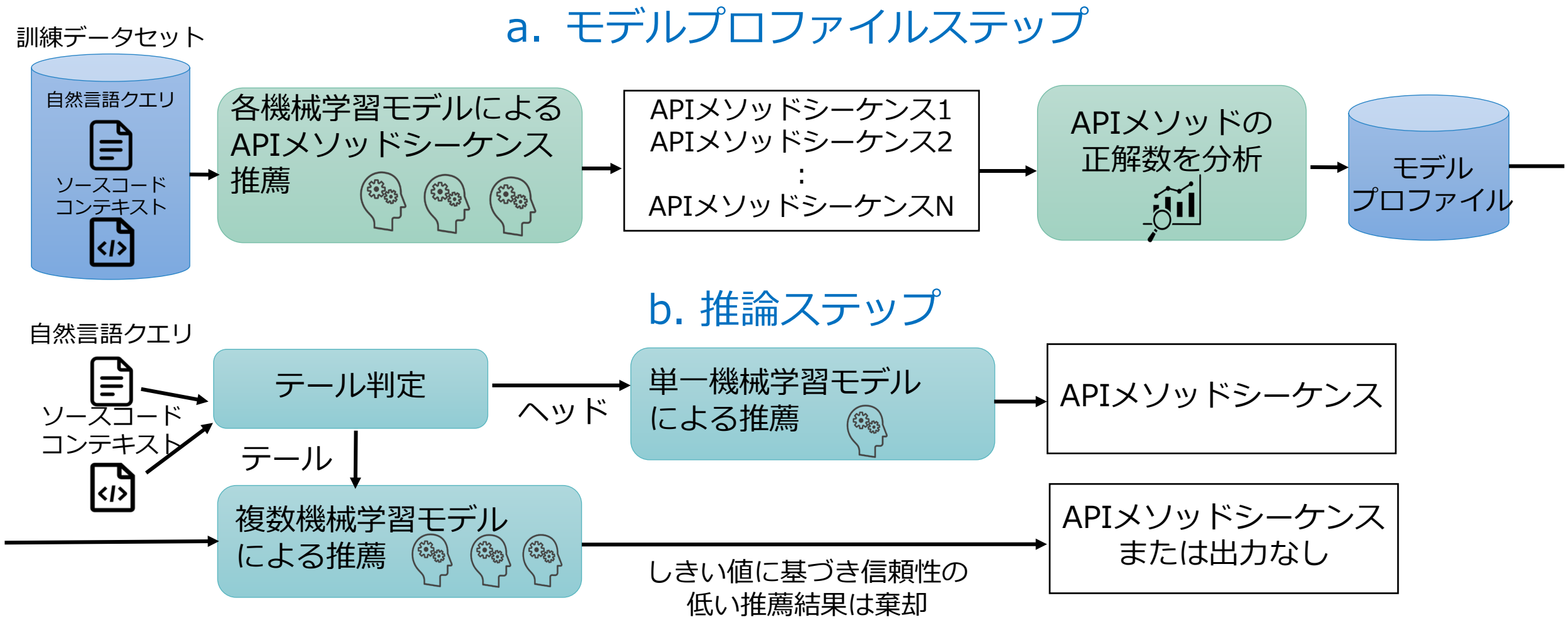
多様な機械学習モデルによるAPI推薦

- 3つの機械学習モデルで同じ推薦タスクを実行
- 正解率の比較
  - CodeBERT　38.2%
  - CodeT5　36.6%
  - MulaRec　46.3%
- 機械学習モデルによって正解しやすいAPIは異なる



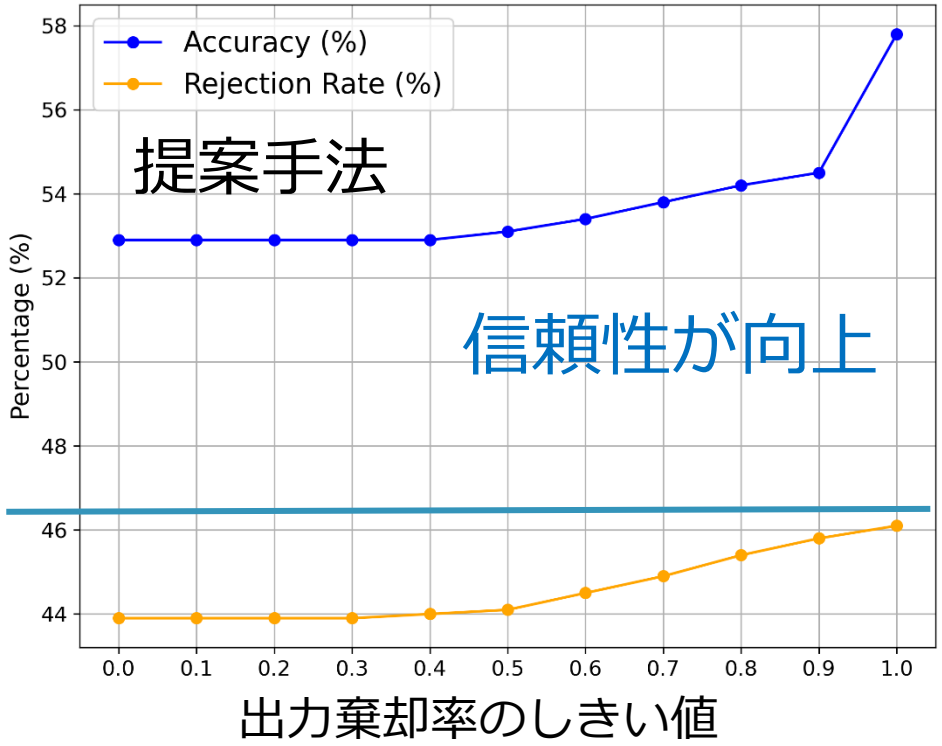
組み合わせることで信頼性向上効果を期待できる

提案手法



評価結果

- APIシーケンス推薦タスクの公開データセット18,500件で評価
  - 出力棄却率のしきい値を上げると正解率と棄却率が増加する



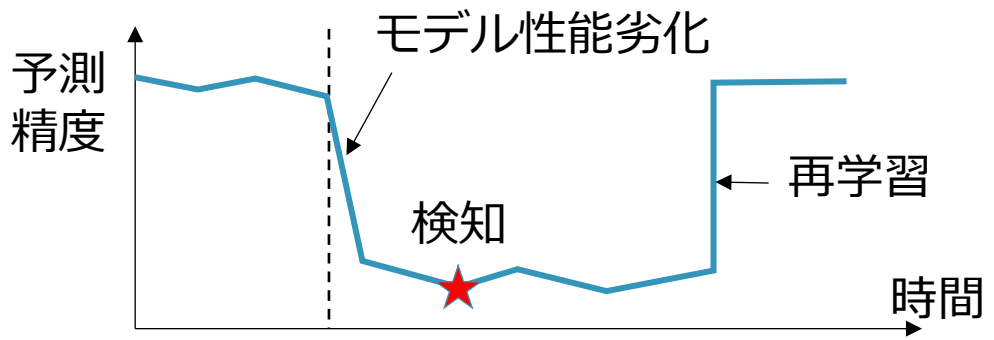
ベースライン　46.3%
（MulaRec正解率）

$$\text{正解率} = \frac{\text{正解数}}{\text{出力数}}$$

$$\text{棄却率} = \frac{\text{出力数}}{\text{入力数}}$$

研究の展望

- システム運用管理タスクへの応用
  - 異常検知システム
    - 異なるセンサーや異なるレイヤの情報源からの推論結果を統合
  - 障害原因特定・修復プラン生成
- 機械学習システムの可用性
  - 長期間稼働する機械学習システムの性能や信頼性を維持
    - モデルの性能劣化検知
    - モデルの再学習



Thanks to collaborators

Qiang Wen
(University of Tsukuba)

Aoi Matsuda
(University of Tsukuba)

Júlio Mendonça
(Tilburg University)

Marcus Völp
(University of Luxembourg)

ご清聴ありがとうございました