HITACHI
Inspire the Next

JaSST'18 Tokyo

AI搭載システムの品質保証

株式会社日立製作所 研究開発グループ
システムイノベーションセンタ システム生産性研究部

明神智之

HITACHI
Inspire the Next

本発表は・・・

Test/QA for AI(Machine Learning) ← こちらです

AI(ML) for Test/QA

HITACHI
Inspire the Next

1. 背景
2. アプローチ
3. AI搭載システム検証
4. AI搭載システム品質アセスメント
5. 社会的コンセンサスの形成
6. 結論・今後の課題

HITACHI
Inspire the Next

1. 背景
2. アプローチ
3. AI搭載システム検証
4. AI搭載システム品質アセスメント
5. 社会的コンセンサスの形成
6. 結論

1-1. AIに関する社会的動向

HITACHI
Inspire the Next

AI(ML)に対する信頼性確保が各方面で議論され始める。また、次回ICSTのキーノートは「Safety Verification for Deep Neural Networks」

- **社会動向**
  - テスラ社の自動運転による初の死亡事故。米国家運輸安全委員会は製造業者に対して「自動運転機能を設計された条件に制限するセーフガードシステムの導入」を勧告
  - 「AIネットワーク社会推進フォーラム」（国際シンポジウム）にて、AIシステムの連携、透明性、制御可能性、安全性、プライバシー、セキュリティ、公正に関する意見が提唱
  - FLI（Future of Life Institute）にて「アシロマAI原則」を公表。23の原則に「AIシステムの安全性とセキュリティ」「AIシステムの事故時の透明性」等、品質保証関連項目が提唱

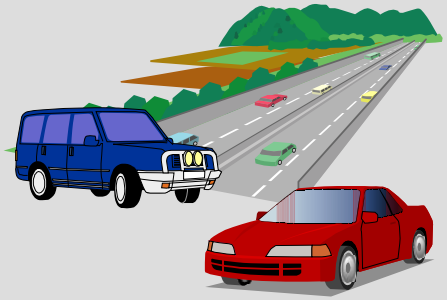
- **学会動向**
  - ICST2018のキーノートに「Safety Verification for Deep Neural Networks」
  - ソフトウェアエンジニアリングシンポジウム 2017。基調講演にて「設計から学習へ：かわりゆく人工知能ソフトウェア」。パネル討論 にて「機械学習とソフトウェア工学」
  - AIの主要学会の一つであるAAAI-18 で品質に関するセッションが新たに登場
    - W8: Engineering Dependable and Secure Machine Learning Systems
    - W16: Statistical Modeling of Natural Software Corpora

1-2. AI搭載システム開発の動向

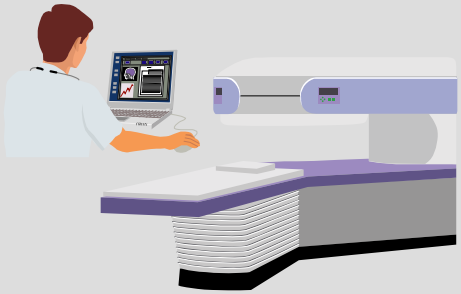
HITACHI
Inspire the Next

高品質が要求される様々な事業へのAI(ML)技術の適用が推進中

自動運転



医療・医用



経営判断

物流

1-3. AI活用における社会課題

HITACHI
Inspire the Next

左図は何で、右図は何か？

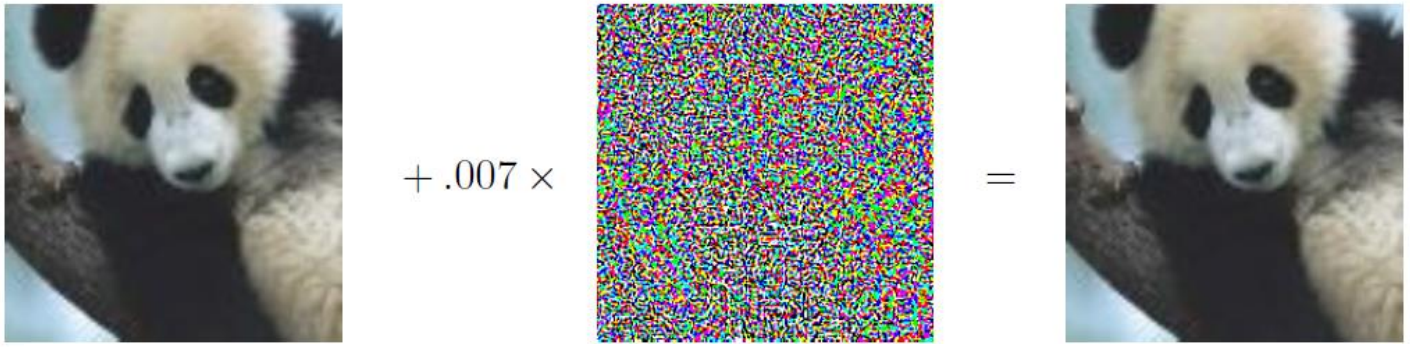


Figure 1: A demonstration of fast adversarial example generation applied to GoogLeNet (Szegedy et al., 2014a) on ImageNet. By adding an imperceptibly small vector whose elements are equal to the sign of the elements of the gradient of the cost function with respect to the input, we can change GoogLeNet's classification of the image. Here our ϵ of .007 corresponds to the magnitude of the smallest bit of an 8 bit image encoding after GoogLeNet's conversion to real numbers.

引用元：Explaining and Harnessing Adversarial Examples: Ian Goodfellow, Jonathon Shlens, Christian Szegedy, ICLR2015

HITACHI
Inspire the Next

左図は何で、右図は何か？

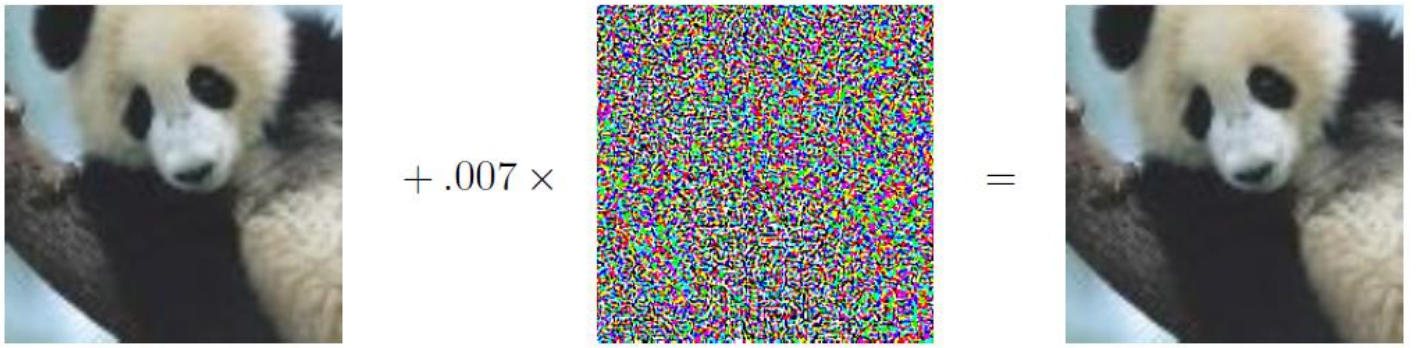


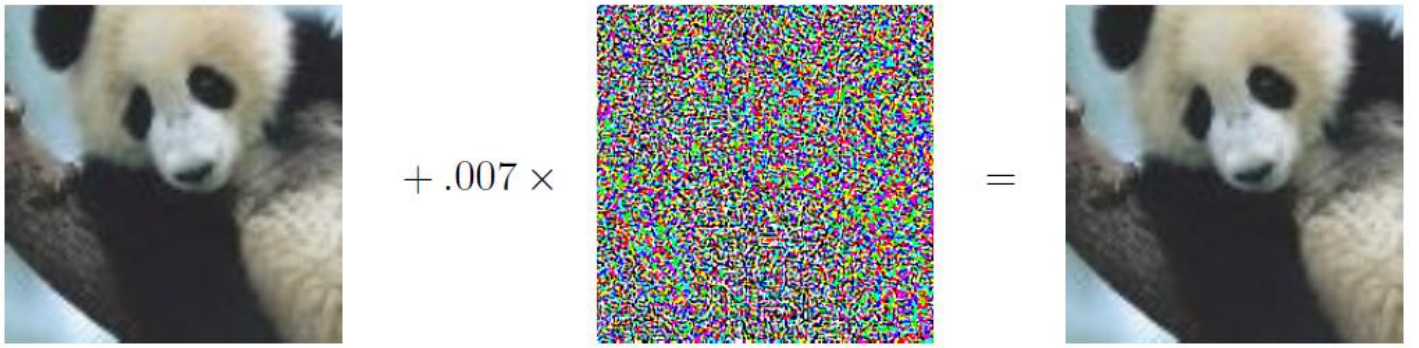
Figure 1: A demonstration of fast adversarial example generation applied to GoogLeNet (Szegedy et al., 2014a) on ImageNet. By adding an imperceptibly small vector whose elements are equal to the sign of the elements of the gradient of the cost function with respect to the input, we can change GoogLeNet's classification of the image. Here our ϵ of .007 corresponds to the magnitude of the smallest bit of an 8 bit image encoding after GoogLeNet's conversion to real numbers.

引用元：Explaining and Harnessing Adversarial Examples: Ian Goodfellow, Jonathon Shlens, Christian Szegedy, ICLR2015

1-3. AI活用における社会課題

HITACHI
Inspire the Next

AIシステムの振る舞いは(一般的に)開発者さえも理解できない
利用者にはAIシステムに対する不安がつきまとう



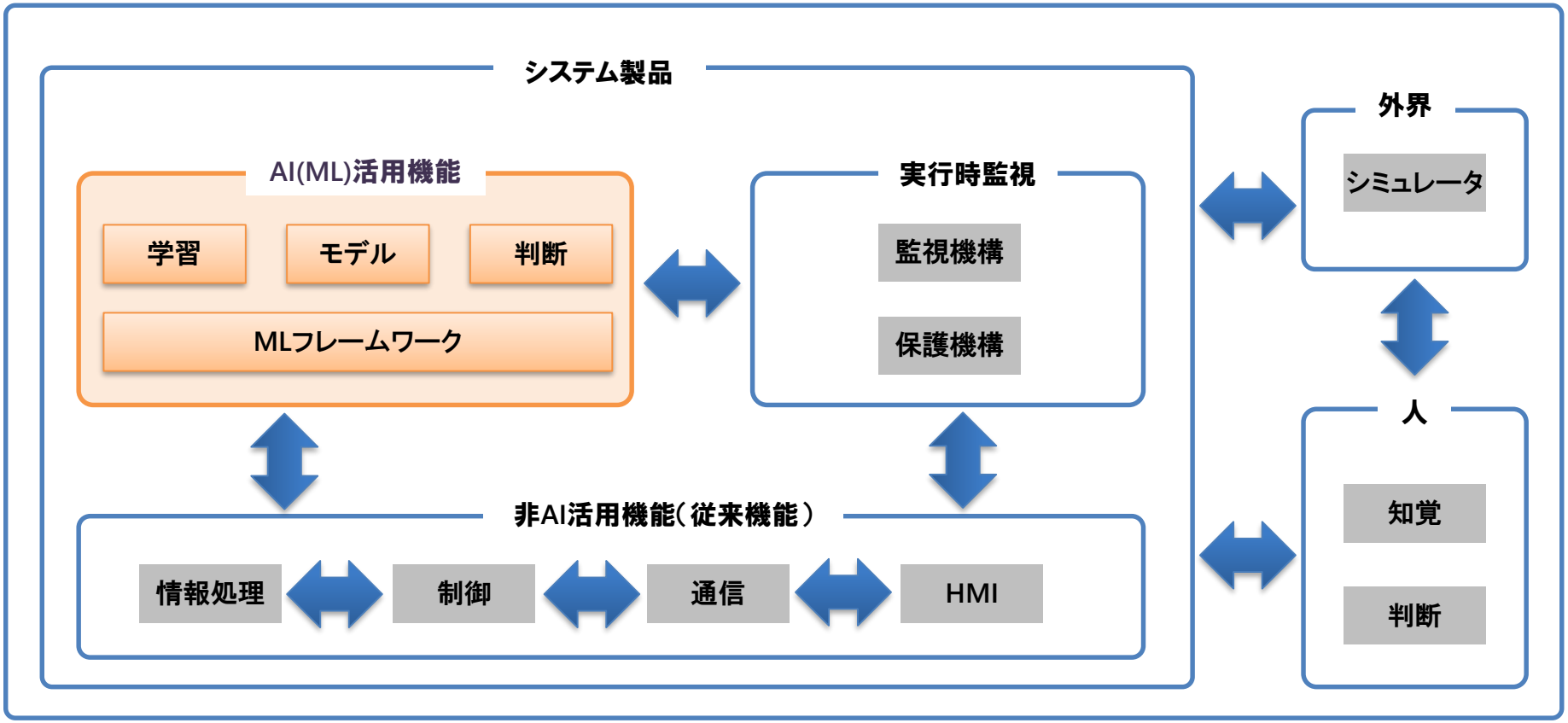
引用元:Explaining and Harnessing Adversarial Examples: Ian Goodfellow, Jonathon Shlens, Christian Szegedy, ICLR2015

高品質が要求される事業へのAI技術の適用が進む中、
今後、AI搭載システムの品質保証が問題視されると予想される

1-4. 本発表のフォーカス

HITACHI
Inspire the Next

**AI搭載システムを取り巻く環境において，
AI(ML)活用機能の品質保証にフォーカス**



HITACHI
Inspire the Next

1. 背景
2. アプローチ
3. AI搭載システム検証
4. AI搭載システム品質アセスメント
5. 社会的コンセンサスの形成
6. 結論

2-1. AI搭載システムにおける品質保証とは

HITACHI
Inspire the Next

AI(ML)の品質保証とは、AI(ML)を搭載したシステム全体のリスクが社会で許容可能なリスク以下であることを示すこと

品質保証のゴール：社会が許容可能なリスク上限 ≧ システム全体のリスク

- 従来から許容しているリスク
  - AI活用の追加利益により新たに許容可能なリスク
- 従来のシステムのリスク
  - AI活用により新たに発生するリスク

品質保証のゴールを達成するために、
「(i) AI活用により新たに発生するリスクの低減」するとともに
「(ii) AI活用の追加利益によって拡大可能なリスク基準の策定と社会的な合意形成」を実現

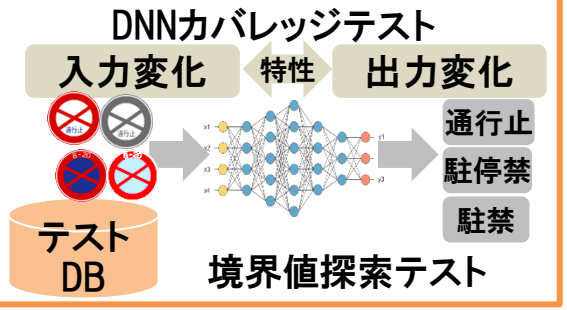
2-2. AI搭載システム品質保証アプローチの提案

HITACHI
Inspire the Next

①AIシステム検証技術、②AIシステム品質アセスメント手法
③社会的コンセンサスの形成により、AI搭載システムの品質保証を実現

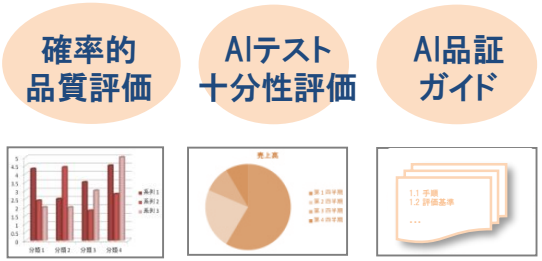
(i) AI起因リスクの低減

①AIシステム検証技術



(ii) AI活用の追加利益によって拡大可能なリスク基準の策定と社会的な合意形成

②AIシステム品質アセスメント手法



③社会的コンセンサスの形成

HITACHI
Inspire the Next

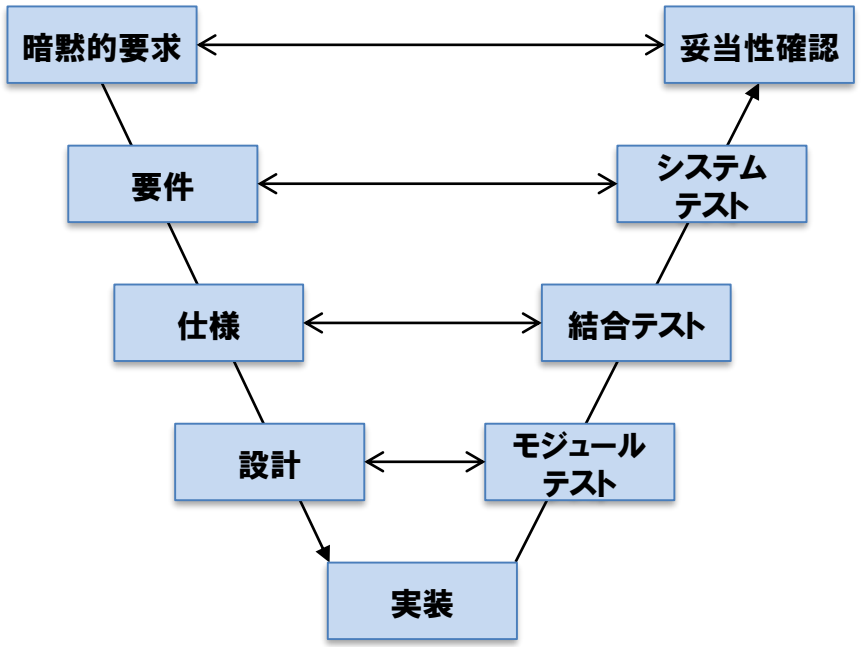
1. 背景
2. アプローチ
3. AI搭載システム検証技術
4. AI搭載システム品質アセスメント手法
5. 社会的コンセンサスの形成
6. 結論

3-1. AI搭載システム検証の課題

HITACHI
Inspire the Next

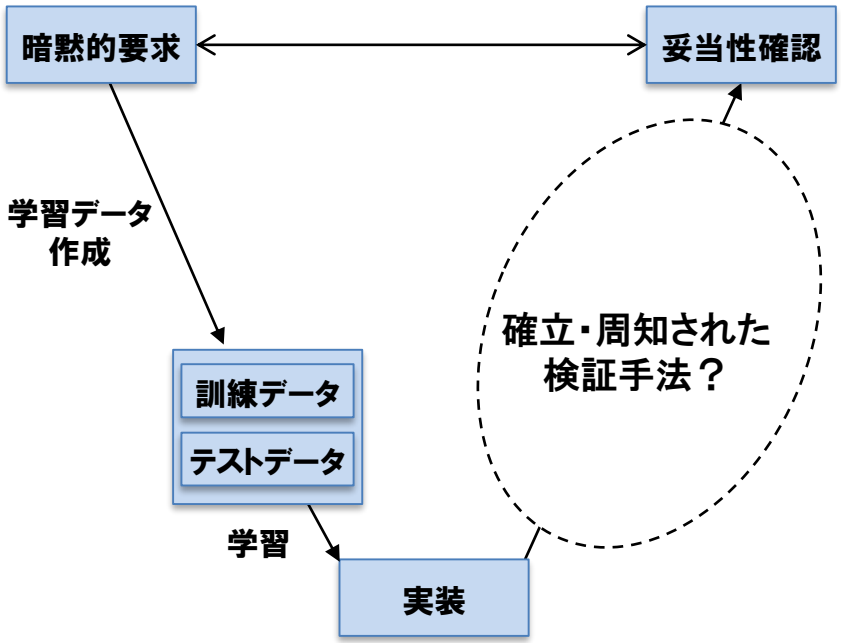
AI搭載システム開発(帰納)では、リスク低減のための工程が妥当性確認に集中しているため、現実的な時間内で求められる品質の達成が困難

従来システム開発(演繹型開発)



各工程で異なる観点で検証することにより
様々な不具合を網羅的に摘出可能

AI搭載システム開発(帰納型開発)



妥当性確認だけで摘出困難な不具合は
埋めこめられたまま出荷

3-2. AI搭載システムの検証観点

HITACHI
Inspire the Next

自動車向け機能安全規格ISO26262記載のテスト手法をヒントにして，AI搭載システムに必要な検証観点，検証アプローチを整理

#	検証観点	検証アプローチ
1	結果が同一となる入力のパターンを満遍なくカバーして検査する	入力と出力の関係を表す性質を仕様から取得し，それらの性質ごとに入力値を作成
2	入力のパターンの境界面に誤りがないか検査する	入力と出力の関係を表す性質を仕様から取得し，それらの性質の境界面となる入力値を作成
3	実行パターンを満遍なくカバーして検査する	実行パターンを制御パスと捉え，異なる制御パスを通る入力を与える（カバレッジテスト）
4	入出力関係の規則性が崩れるなど，特徴の異なるケースを検査する	特徴の異なる入出力を，仕様やプログラムロジックから逆算
5	異常値が出力されるケースを検査する	異常出力を得る入力を，仕様やプログラムロジックから逆算
6	ユースケースの観点から想定外の入力について検査する	ユースケースの観点から想定外となり得る入力を任意に作成
7	ユースケースシナリオに従って検査する	ユースケースシナリオに従う入力を任意に作成

⋮

3-2. AI搭載システムの検証観点

HITACHI
Inspire the Next

自動車向け機能安全規格ISO26262記載のテスト手法をヒントにして，AI搭載システムに必要な検証観点，検証アプローチを整理

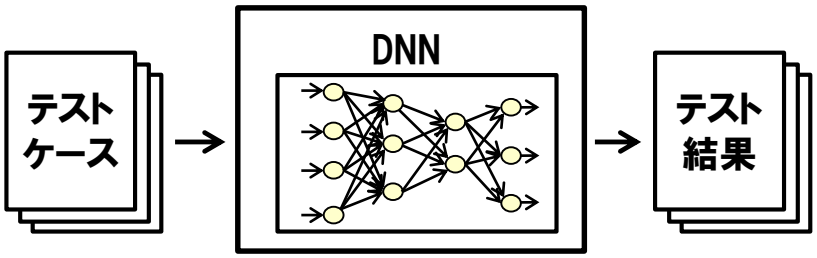
#	検証観点	検証アプローチ
1	結果が同一となる入力のパターンを満遍なくカバーして検査する	入力と出力の関係を表す性質を仕様から取得し，それらの性質ごとに入力値を作成
2	入力のパターンの境界面に誤りがないか検査する	入力と出力の関係を表す性質を仕様から取得し，それらの性質の境界面となる入力値を作成
3	実行パターンを満遍なくカバーして検査する	実行パターンを制御パスと捉え，異なる制御パスを通る入力を与える（カバレッジテスト）
4	入出力関係の規則 特徴の異なるケー	AI(DNN)向け検証アプローチ： DNNカバレッジテスト 学習済みネットワーク内のニューロンの活性化を網羅するようにテストデータを生成・テスト
5	異常値が出力されるケースを検査する	異常出力を得る入力を，仕様やプログラムロジックから逆算
6	ユースケースの観点から想定外の入力について検査する	ユースケースの観点から想定外となり得る入力を任意に作成
7	ユースケースシナリオに従って検査する	ユースケースシナリオに従う入力を任意に作成

3-3. 開発技術：DNNカバレッジテスト手法

HITACHI
Inspire the Next

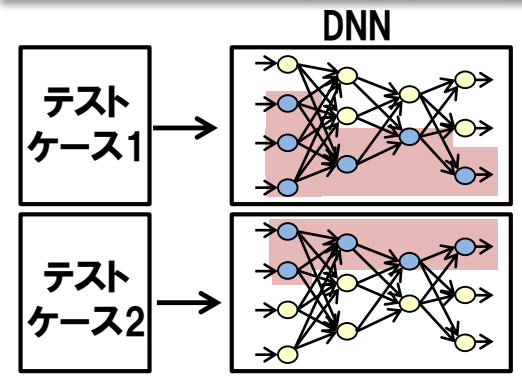
DNN内のニューロン活性化状態を基準とした、テストケース評価技術および活性化網羅率を上げるテストケース生成技術の開発

AI搭載システムのテストの課題



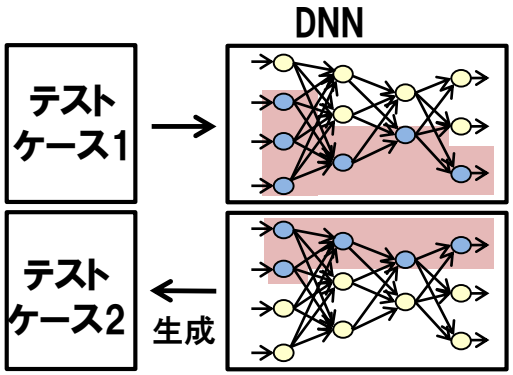
テストの定量的な充足判定が困難

DNNテストケース評価技術



DNNのノード活性化を基準として、活性化網羅率でテストの充足性を判定（道路標識認識を用いて検討）

DNNテストケース生成技術

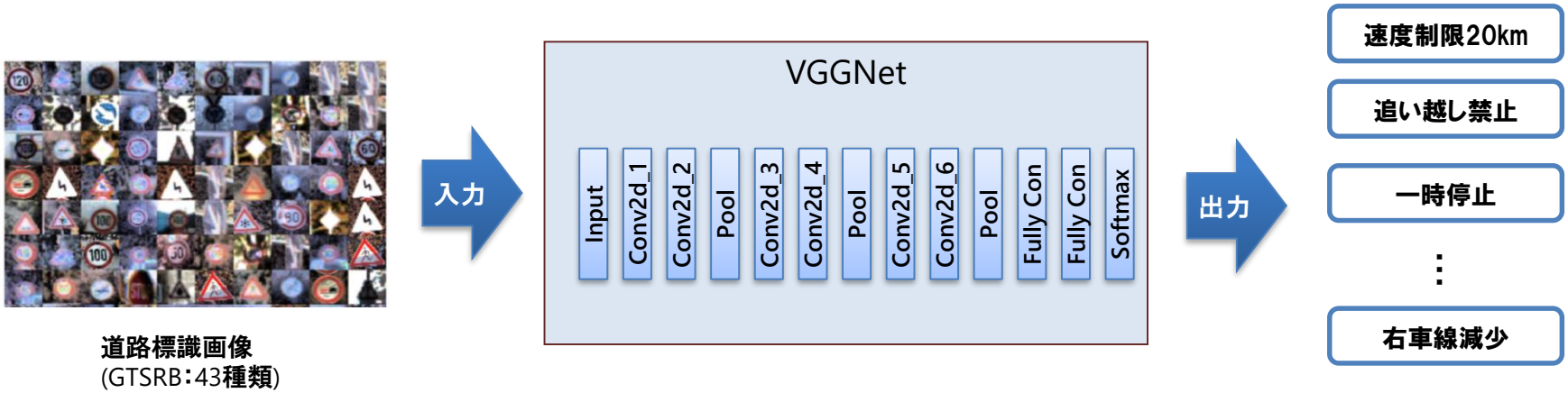


活性化網羅率を向上するようにテストパタンを選別・生成

3-4. 実験対象システム

HITACHI
Inspire the Next

- 道路標識画像の教師あり分類
  - ドイツ道路標識画像(GTSRB*)43種類を分類
  - 学習データ　26640枚
  - テストデータ 12630枚
- 学習済みモデルを対象にカバレッジの評価実験を実施
  - 学習そのものは対象外



*J. Stallkamp, M. Schlipsing, J. Salmen, and C. Igel. The German Traffic Sign Recognition Benchmark: A multi-class classification competition. In Proceedings of the IEEE International Joint Conference on Neural Networks, pages 1453-1460. 2011.

3-5. DNNカバレッジテストの定義

HITACHI
Inspire the Next

DNNカバレッジによるテストケースの充足性判定手法を試作・可視化

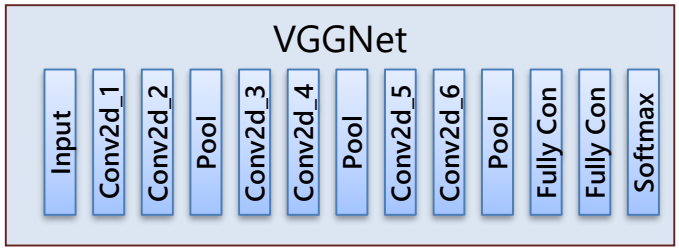
- 仮説：テストケースが充足 ⇒ ニューラルネットワークをより多く活性化
  - （充足とは，例えば誤判定を引き起こす画像が十分な種類あること）
- 充足判定式の定義（カバレッジ）：
  - テストケースが活性化したニューロンを評価
  - 一つのテストケースで活性化したニューロンが少ないほどニューロン一つ当たりの活性化度合いが高いと評価する（先行研究*との差異）

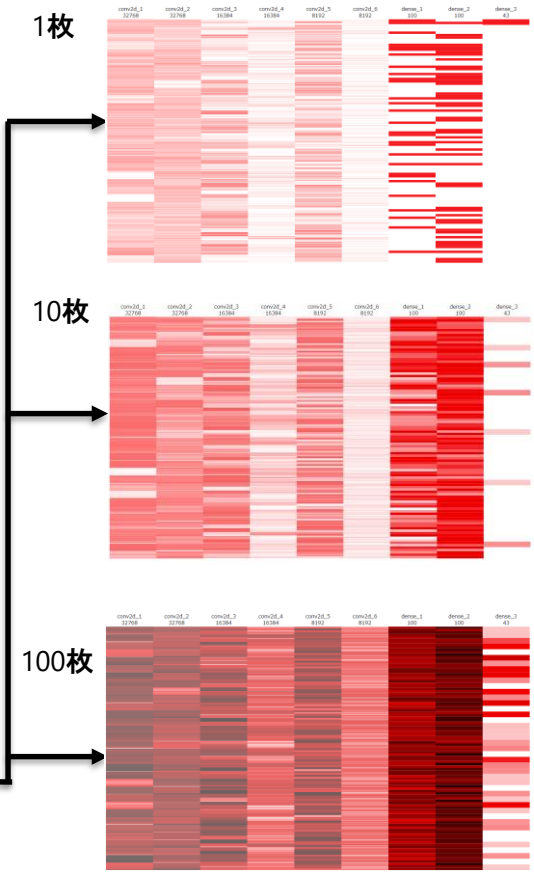
$$E(n,T) = \sum_{t}^{T} \frac{1}{|\{n|\forall n \in N, act(n,x) > \mathrm{t}\}|}$$

N：ニューラルネットワーク内のニューロン$n0, n1, \ldots, nk$ の集合，
T：テストケース $x0,x1,\ldots,xk$ の集合，$out(n, x)$：活性化関数，t：閾値
E：ネットワークの活性化度

- 複数テストケースを入力した結果，活性化したニューロンが多いほどテストが充足していると評価する

道路標識画像
(GTSRB：43種類)





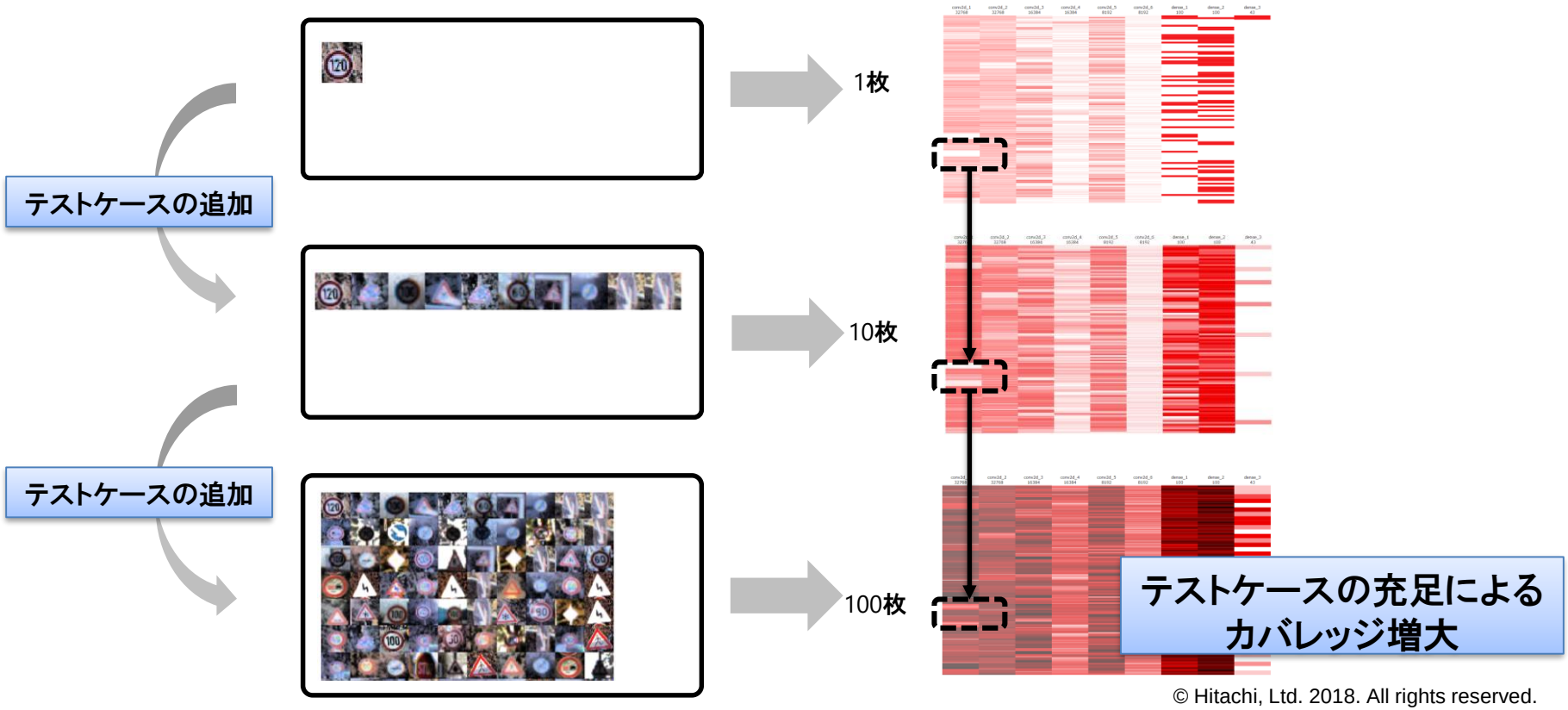
*DeepXplore: Automated Whitebox Testing of Deep Learning Systems, Kexin Pei, Yinzhi Cao, Junfeng Yang, Suman Jana, SOSP' 17

3-6. DNNカバレッジテストの結果

HITACHI
Inspire the Next

誤判定ケースについて，中間層の充足と画像のバリエーションの関係を評価

- 仮説検証のため，誤判定を引き起こす画像とカバレッジを調査
  - 画像増加によるカバレッジ増大を確認
  - テストケースとして適切にバリエーションを増やせているかを確認



HITACHI
Inspire the Next

1. 背景
2. アプローチ
3. AI搭載システム検証
4. **AI搭載システム品質アセスメント**
5. 社会的コンセンサスの形成
6. 結論

4-1. ②AI搭載システム品質目標の決定

HITACHI
Inspire the Next

AIによる誤り発生の影響の大きさとAI活用の追加利益に基づいて、AIによる誤り発生確率の上限を規定し、システムの品質目標として定義

#	誤り発生の影響の大きさ	AI活用による追加利益
1	多数の人命に関わる	小
2		大
3	少数の人命に関わる	小
4		大
5	資産の損失に繋がる	小
6		大
7	資産の損失に繋がるが回復可能（保険など）	小
8		大
9	人命や資産への影響なし	小
10		大

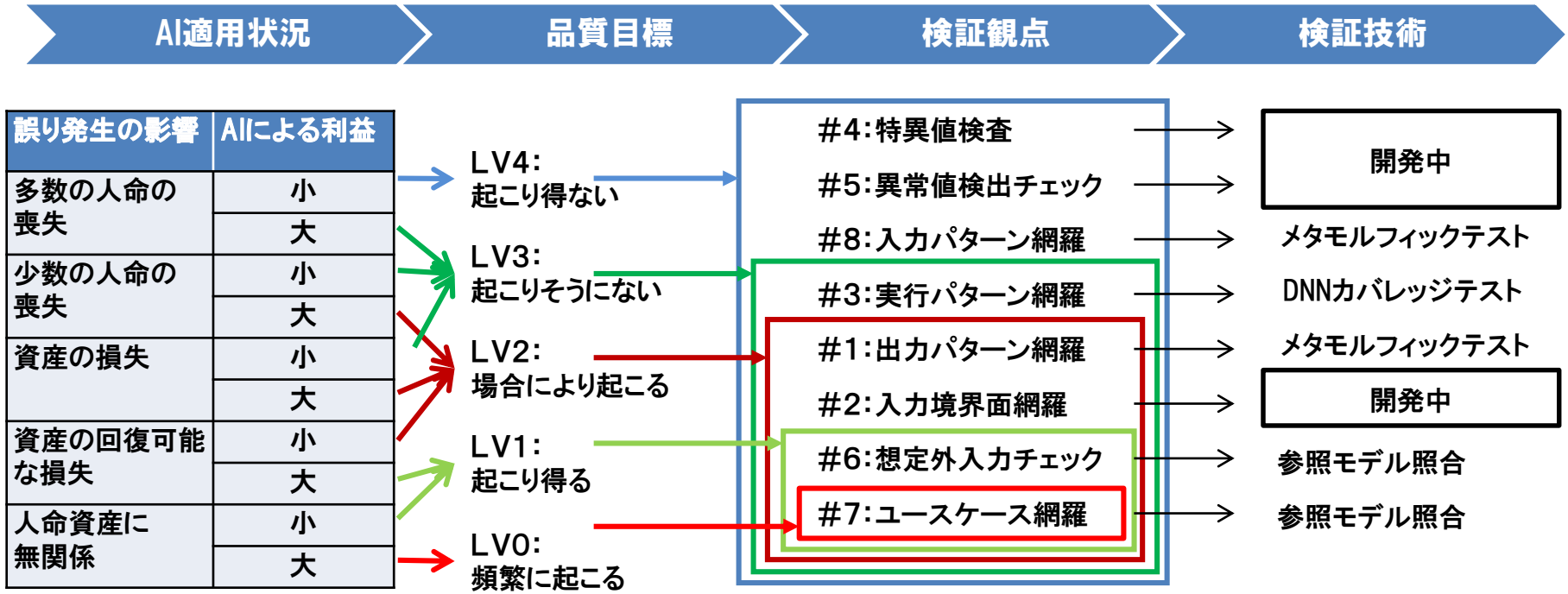
AIによる誤り発生確率の上限	品質目標（レベル）
起こり得ない（Improbable）	LV4
起こりそうにない（Remote）	LV3
起こりそうにない（Remote）	LV3
場合により起こる（Occasional）	LV2
起こりそうにない（Remote）	LV3
場合により起こる（Occasional）	LV2
場合により起こる（Occasional）	LV2
起こりうる（Probable）	LV1
起こりうる（Probable）	LV1
頻繁に起こる（Frequent）	LV0

4-2. ②AI搭載システム品質アセスメント技術の確立

HITACHI
Inspire the Next

定義したAI搭載システムの品質目標が、適用した検証技術によって達成されていることを確認する、「品質アセスメント手法」を確立

品質アセスメント手法の基本アイデア



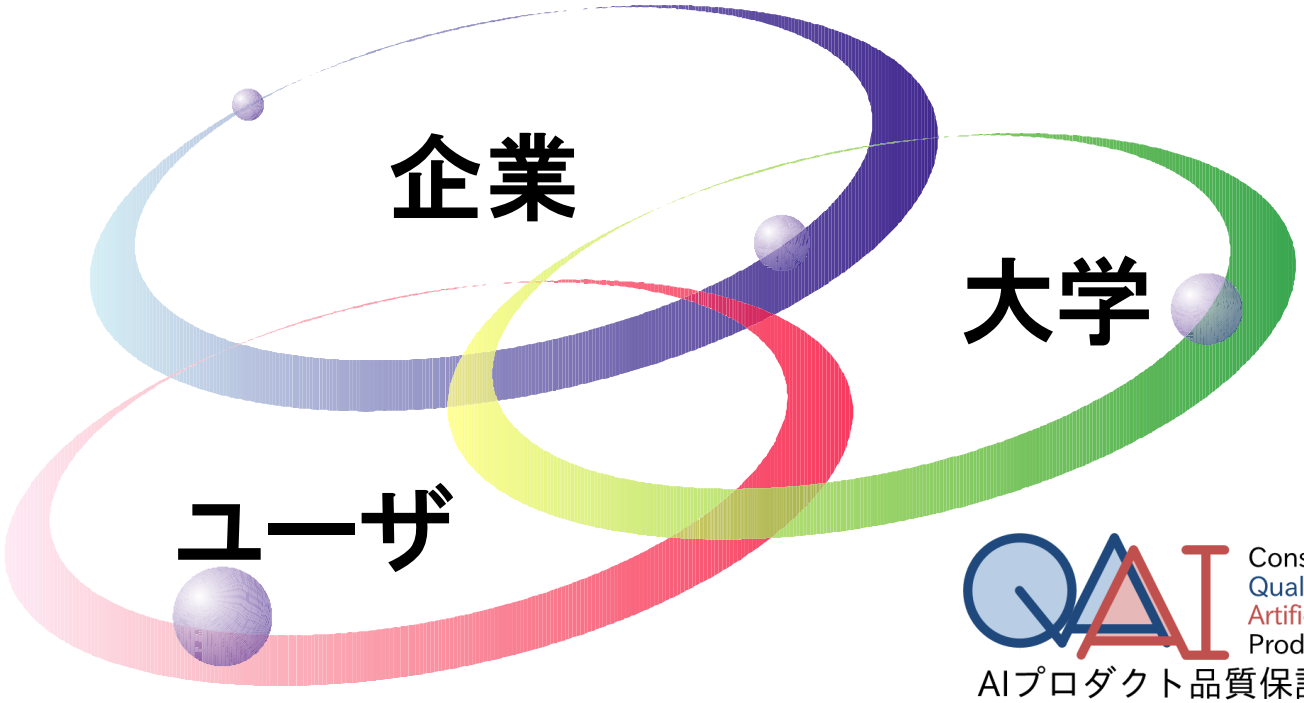
HITACHI
Inspire the Next

1. 背景
2. アプローチ
3. AI搭載システム検証
4. AI搭載システム品質アセスメント
5. **社会的コンセンサスの形成**
6. 結論

5-1. ③社会的コンセンサスの形成

HITACHI
Inspire the Next

AI搭載システムの品質保証に関する調査・研究・体系化、および品質に対する適切な期待を社会に啓発するための連携組織が必要



QAI
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

AI搭載システムと社会との安心できる共生の実現を社会全体でめざす

HITACHI
Inspire the Next

1. 背景
2. アプローチ
3. AI搭載システム検証
4. AI搭載システム品質アセスメント
5. 社会的コンセンサスの形成
6. **結論**

6. 結論・今後の課題

HITACHI
Inspire the Next

- 結論
  - 品質保証のゴール設定
  - 社会が許容可能なリスク上限 ≧ システム全体のリスク

- AI(機械学習)搭載システムの品質保証に関するアプローチを制定
  1. AIシステム検証技術
  2. AIシステム品質アセスメント手法
  3. 社会的コンセンサスの形成

- 今後の課題
  - 検証技術の拡充
  - 実案件での評価
  - 社会的合意の形成に向けた組織作り

HITACHI

Inspire the Next