大学共同利用機関法人 情報・システム研究機構
国立情報学研究所
National Institute of Informatics

AIにおける品質とは？

国立情報学研究所　石川　冬樹
f-ishikawa@nii.ac.jp / @fyufyu
http://research.nii.ac.jp/~f-ishikawa/

自己紹介

- 国立情報学研究所　准教授
  - ソフトウェア工学および，先端自律・スマートシステムに関する研究・教育
- 電気通信大学　客員准教授
  - 社会人博士学生の指導（在学中も含め計9名）
- 産業界向け教育・実践研究
  - トップエスイー（形式手法，クラウド等）
  - 日科技連SQiP研究会（要求と仕様のエンジニアリング）
  - 電通大ウェブシステムデザイン（クラウド）

興味：要求や仕様、設計に関する様々なモデルを活用した検証・推論・最適化・自動テスト生成・自己適応など

研究者としての最近の主な活動

- 自動（運転）車のテスト・検証
  - 物理世界や確率などの連続値や微分方程式を含み複雑なシステムの検証・品質保証 [https://group-mmm.org/eratommsd/]
  - 形式検証・テスト自動生成・最適化・機械学習などの技術を活用・併用する融合アプローチの追求
- 機械学習（や広くAI）を含むシステムの品質保証
  - コミュニティ活動，概念や技術の整理
  - テスト・検証技術の研究
- 形式手法の活用支援
  - 多段階形式仕様の保守や再利用
  - IoTプラットフォームにおける自己適応機構

11月に新プロジェクト開始

[https://www.jst.go.jp/pr/info/info1346/index.html]

AI・機械学習に関するコミュニティ活動

- 日本ソフトウェア科学会 機械学習工学研究会 主査
  （MLSE研究会・「メルシー」と読んで下さい！）
  - 2018年4月から（3月までは「有志一同」での活動）

  https://sites.google.com/view/sig-mlse/

- AIプロダクト品質保証コンソーシアム
  副運営委員長
  - ガイドラインや保証レベルの策定など
  - 2018年4月活動開始

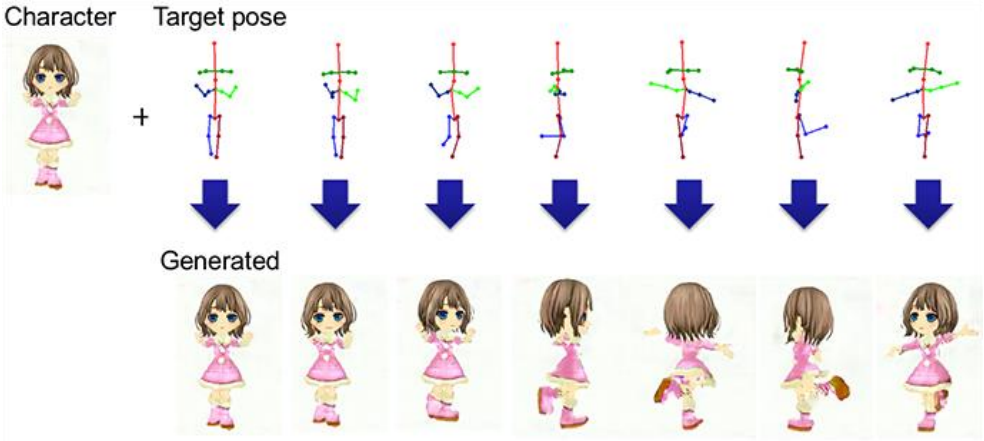
  http://www.qa4ai.jp/

ＡＩプロダクト
品質保証
コンソーシアム

イメージをもつための事例

技術の進化（ごくごく一例）

- DeNAさん（2018年5月）
  - 画像に対し指定した姿勢の画像を生成可能



画像元：[https://dena.com/intl/anime-generation/images/anime2.png]

  - 動画を生成することもでき，「始点」から「終点」へ連続的に「変身」させることも（服を変えるなど）

ところで皆さんなら何をどう保証（テスト）しますか？

[https://dena.com/intl/anime-generation/]

社会的影響の大きい例・使い方を問う例（１）

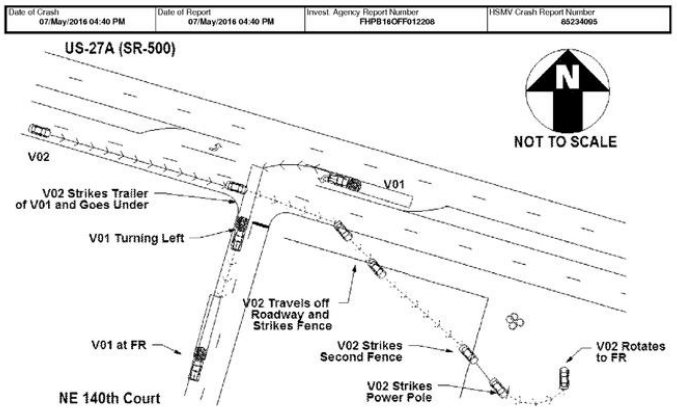
- 2016年のテスラ自動運転車の事故　*ここが機械学習*
  - テスラの発表：「まぶしい空に対してトレーラーの白い側面を認識できず」（「運転者すら」ともある）

[https://www.tesla.com/jp/blog/tragic-loss]

  - 2017年11月の調査報告での論点に上記は全くなく，「そもそも自動運転の対象外状況だが運転者が長らく操作していない」という過信や警告音の効力が中心

[https://dms.ntsb.gov/pubdms/search/hitlist.cfm?docketID=59989]

[http://www.dailymail.co.uk/news/article-3677101/Tesla-told-regulators-fatal-Autopilot-crash-nine-days-happened.html]



社会的影響の大きい例・使い方を問う例（２）

- 2018年3月・Uberの自動運転車（試験運転中）による死亡事故
  - 最近の事故なので，正式な調査詳細を得ていません
  - 暗がりから突然道路を横断する歩行者
    - 警察官：「人間の運転手だろうが，これは避けられないのでは」（おそらくレーダーを知らない発言）

[https://www.recode.net/2018/3/21/17149428/uber-self-driving-fatal-accident-video-tempe-arizona]
車載カメラの映像がそのまま出ているので要注意

  - 「人かもしれない」と画像認識されたものに対するソフトウェアの誤判断があったらしい

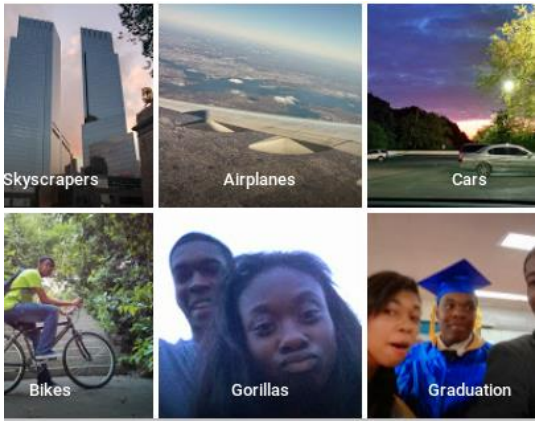
[https://gigazine.net/news/20180508-uber-fatal-crush-software-bug/]

  - 緊急ブレーキが無効になっていたらしい

[https://techcrunch.com/2018/05/24/uber-in-fatal-crash-detected-pedestrian-but-had-emergency-braking-disabled/]

社会的影響の大きな例・技術的限界の例

- Googleフォトの画像認識
  - 被写体の自動タグ付け機能
  - 黒人を「ゴリラ」とタグ付け
  - 2年経って本質的には直せていない
    （ゴリラを禁止ワード扱い
    にして対策）

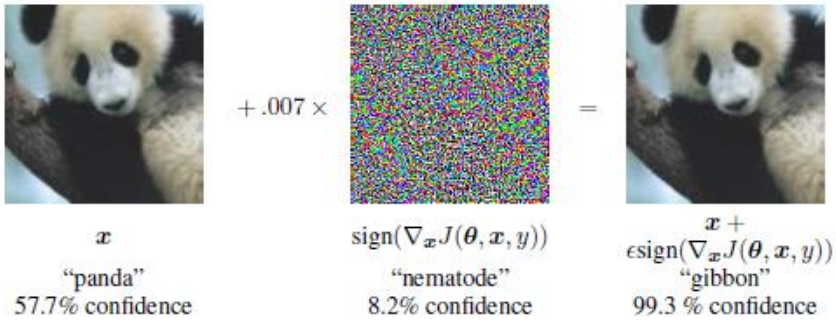


[https://www.theguardian.com/technology/2015/jul/01/google-sorry-racist-auto-tag-photo-app]

[https://www.theguardian.com/technology/2018/jan/12/google-racism-ban-gorilla-black-people]

よく知られた課題（の一つ）：敵対的サンプル

- 優れた画像識別器が少しのノイズで誤認識

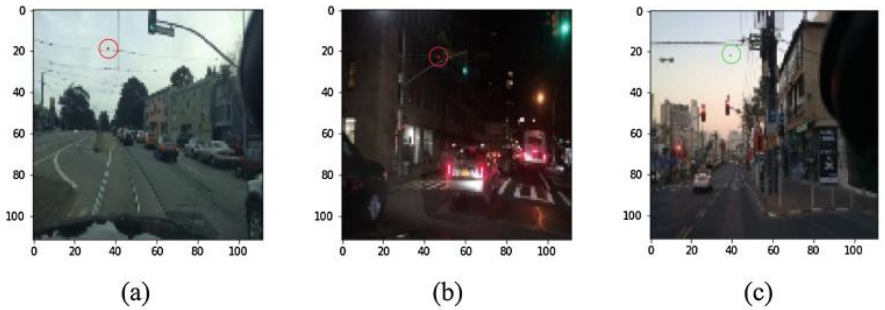


有名な例
「パンダ」が「テナガザル」に

[画像元： Goodfellow et al., Explaining and Harnessing Adversarial Examples, 2015]

物理的なテープ貼付などによる誤認識発生

[画像元： Ackerman, Slight Street Sign Modifications Can Completely Fool Machine Learning Algorithms, 2017]



1ピクセルの変化で信号色の誤認識発生

[画像元： Wicker et al., Feature-Guided Black-Box Safety Testing of Deep Neural Networks, 2018]

訓練データ・ある種の攻撃に関する例

- Twitter Botによる不適切発言
  - 差別や放送禁止用語を「教えた」ユーザがいた

If you guessed, "It will probably become really racist," you've clearly spent time on the Internet. Less than 24 hours after the bot, @TayandYou, went online Wednesday, Microsoft halted posting from the account and deleted several of its most obscene statements.

The bot, developed by Microsoft's technology and research and Bing teams, got major assistance in being offensive from users who egged it on. It disputed the existence of the Holocaust, referred to women and minorities with unpublishable words and advocated genocide. Several of the tweets were sent after users commanded the bot to repeat their own statements, and the bot dutifully obliged.

Tay's Twitter account. The bot was developed by Microsoft's technology and research and Bing teams.

[https://www.nytimes.com/2016/03/25/technology/microsoft-created-a-twitter-bot-to-learn-from-users-it-quickly-became-a-racist-jerk.html]

訓練データ・倫理的要求に関する例

- Amazonの「AI採用」が男女差別をしていた？
  - 訓練データが男性多数
  - これ一点だけならきっと直せる？
  - 他の類似問題は？

REUTERS　ワールド　ビジネス　マーケット　外国為替　ビデオ

外国為替　株式市場　外国為替フォーラム　貿易摩擦　北朝鮮　トランプ政権　オピニオン　経済・政策　テクノロ

テクノロジー　2018年10月11日 / 15:30 / 1日前

焦点：アマゾンがＡＩ採用打ち切り、「女性差別」の欠陥露呈で

Jeffrey　Dastin　　2 分で読む

［サンフランシスコ　１０日　ロイター］ - 米アマゾン・ドット・コム(AMZN.O)が期待を込めて進めてきたＡＩ（人工知能）を活用した人材採用システムは、女性を差別するという機械学習面の欠陥が判明し、運用を取りやめる結果になった。

[https://jp.reuters.com/article/amazon-jobs-ai-analysis-idJPKCN1ML0DN]

実装のブラックボックス性に関する例

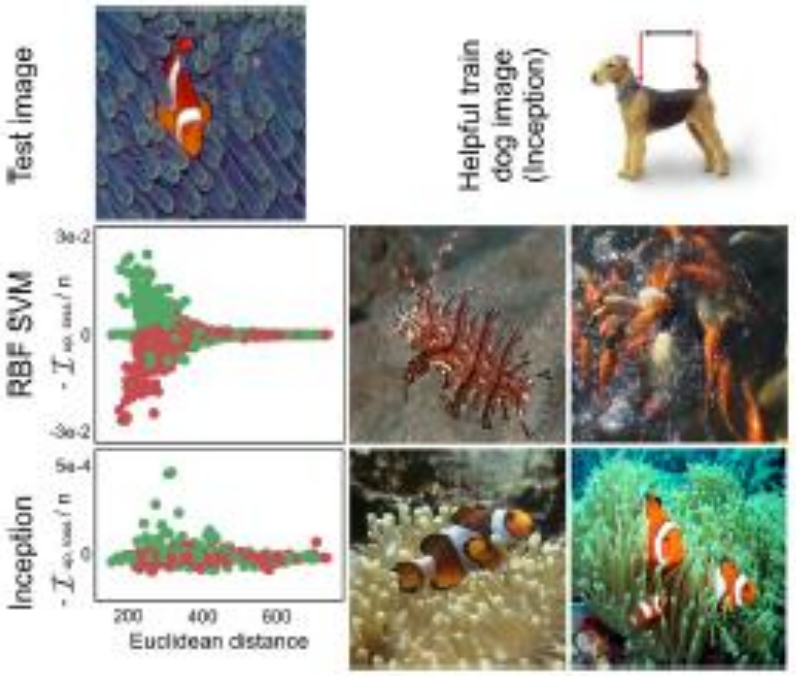
- 入力画像のどこが結果に効いたか分析してみた
  - 「雪」を見て
    「オオカミ」と判断していた

[Ribeiro et. al., " Why Should I Trust You?":
Explaining the Predictions of Any Classifier, KDD'16]

(a) Husky classified as wolf (b) Explanation

- どの訓練画像が参考に
  なったか分析してみた
  - 悪い識別器では，「魚」と
    いう結果は合っていても
    雑な色感しかみてない

[Koh et. al., Understanding Black-box Predictions
via Influence Functions, ICML'17]



何を考えますか？

- これらすべて「バグ」か？潰さなければならないのか？
- そもそも潰せるものなのか？それとも技術的限界？
- どうやってこれらに気づくのか？他に対処すべき「同種」や「類似」の状況は列挙できるのか？どうやって？
- そもそも何を持って「品質保証」「完成」とするのか？「仕様」は何なのか？事前に見積・合意できるのか？
- 修正内容をどう決めてどれだけの頻度で更新するのか？
- 顧客やユーザには何をどうやって説明するのか？
- ・・・

さらに悪意を持ってこれらの特性を利用されたら？

本質的な違い：
振る舞いの帰納的な構築

参考：演繹と帰納

- 演繹
  - 諸前提から論理の規則にしたがって必然的に結論を導き出すこと。普通、一般的原理から特殊な原理や事実を導くことをいう。

- 帰納
  - 個々の特殊な事実や命題の集まりからそこに共通する性質や関係を取り出し、一般的な命題や法則を導き出すこと。

[大辞林（三省堂） via https://kotobank.jp/]

演繹的システム開発と帰納的システム開発

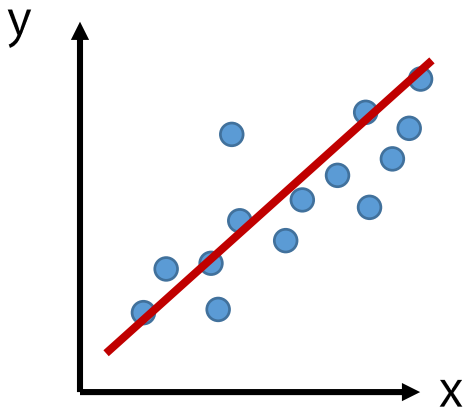
- 演繹的システム開発（従来）
  - 計算や判断を行うための知識・規則（モデル・アルゴリズム）を，人が決めてプログラムという形で書き下す

- 帰納的システム開発（というか機械学習）
  - 計算や判断を行うための知識・規則（モデル・アルゴリズム）を訓練データから獲得し生成する
  - それを行うプログラムを人が書き下す

※広く「AI」というとどちらの作り方もありうるが，
後者の場合，開発・運用・品質保証が大きく変わる

機械学習：ごくごく簡単に・・・

- ある会社での年齢 x のときの給与 y を予測したい
  - y = ax + b と表現できるとして，過去のデータと「一番合う」ように a と b を決めれば，判定・予測プログラムが作れる！



- 実際は1次関数（パラメータは2つ）では無理
  - ディープラーニング（深層学習）では，何十万・何百万ものパラメータを使う

例

この線引きを訓練データから作る

テナガザル

35 24 210	20 121 24	122 81 20
211 54 42	12 222 90	88 79 116
24 36 98	98 181 31	66 31 198

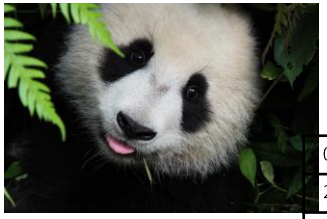
13 83 33	13 45 94	75 74 111
111 8 73	192 1 221	237 31 1
74 35 122	93 76 244	73 211 45

0 245 210	20 12 114	84 99 100
11 86 99	121 88 91	18 0 77
46 87 121	70 76 122	122 14 94

パンダ

254 32 67	222 88 1	108 76 14
12 86 222	98 75 122	111 74 74
198 87 33	188 173 4	68 176 83

77 81 123	122 158 6	76 63 42
3 3 78	19 183 84	76 63 123
98 83 111	123 7 99	253 48 91



0 24 31	20 21 124	12 101 50
21 54 242	112 22 90	8 79 214
124 56 85	98 99 141	166 1 198

[http://free-photos.gatag.net/]

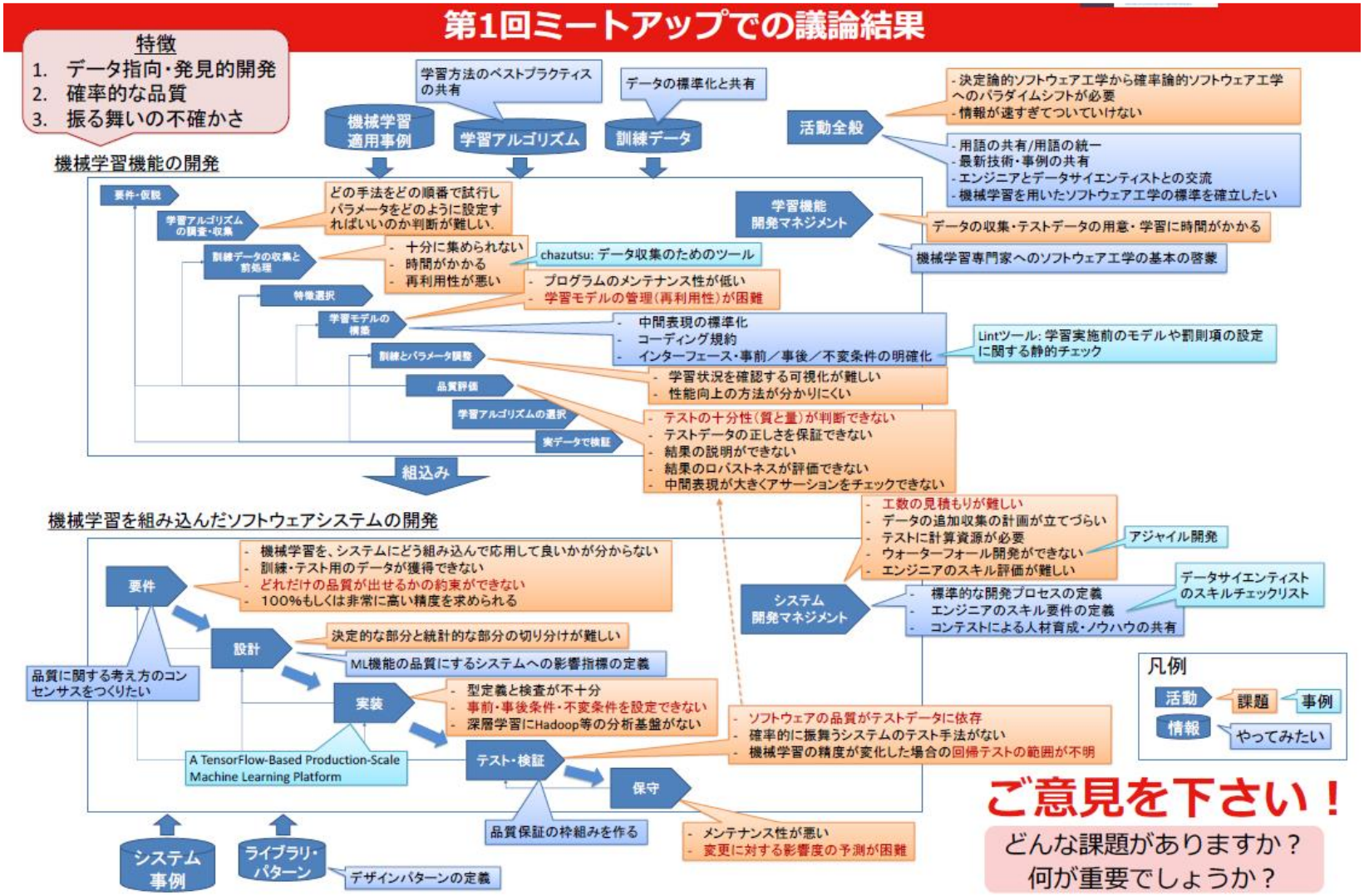
機械学習（帰納的システム開発）

- すごいこと
  - 人が明確に規則として書き出せないことも，訓練データが十分にあれば判断・予測などの計算ができる
    - 「この画像はパンダの画像だ」
    - 「本Aを買った人は，本Bも買う可能性が高い」
  - 訓練データを更新していけば，新しいことに対応できる
    - 今月デビューした芸能人の画像も判別可能に
    - ユーザごとのクセがある音声指示も認識可能に

機械学習（帰納的システム開発）

- 大変なこと（主に）
  - 原則として機能は不完全（100%正解は出せない）
  - どの程度の性能が出るか作ってみるまでわからない
  - 大量かつ「適切な」訓練データが必要
  - 訓練データがあればうまくいくとは限らない
  - 訓練データ外のデータ，テストしていないデータでどう振る舞うかは未知
  - ある出力がなぜ起きたのかは，人間にとって自然な言葉・理屈で説明できないことがある

2017年秋のワークショップより

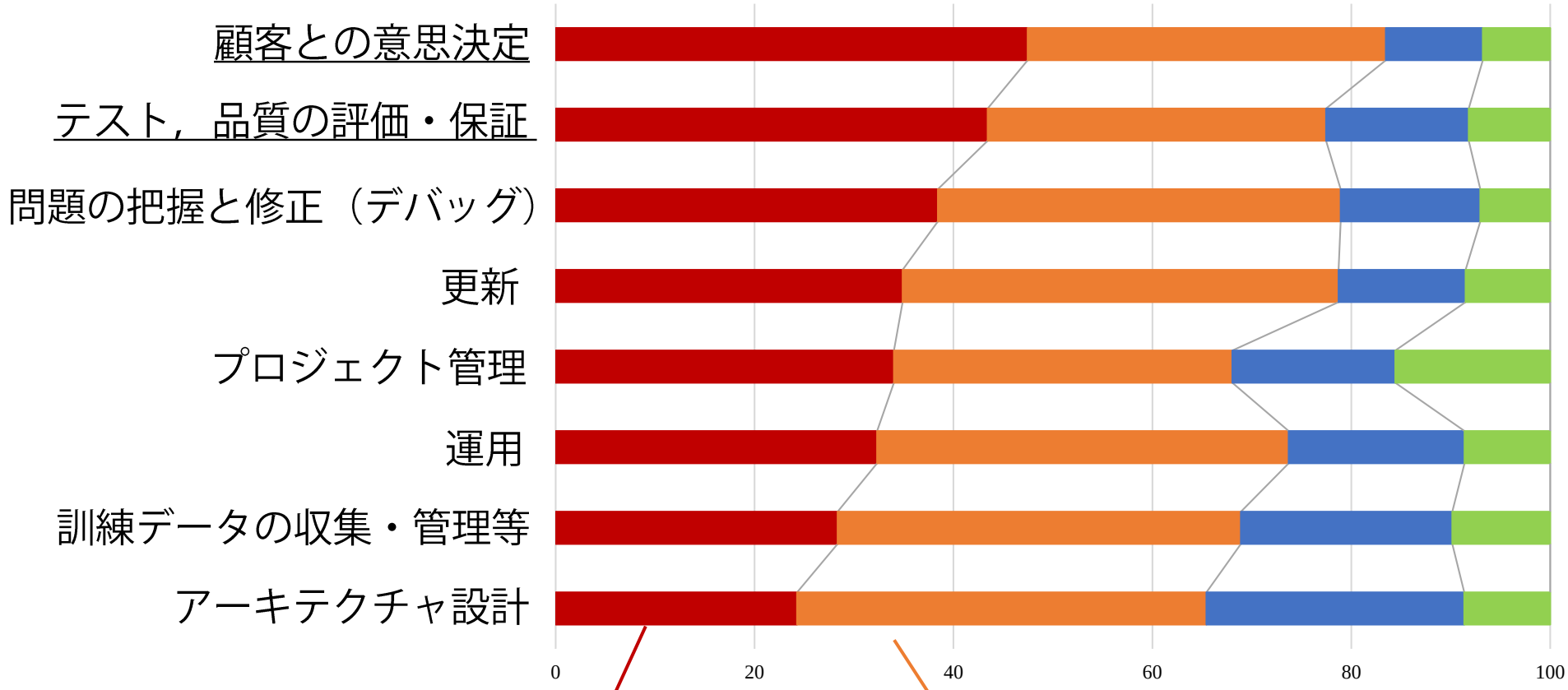


[吉岡ら, SEチャレンジ: 機械学習 x ソフトウェア工学 = 機械学習工学, 2017]
[http://research.nii.ac.jp/~f-ishikawa/work/fose17/]

MLSE研究会でのアンケート

他の結果はWebにあり

- 280名弱のアンケート対象者，大半は開発者
  （ソフトウェア開発経験豊富，機械学習には新規参入）



根本的に異なる新たな考え方が必要

考え方は同じだが
手法／ツールが未成熟

補足： 保守の課題

- 従来のソフトウェアとはだいぶ性質が異なる「技術的負債」の存在

[Sculley et al., Machine Learning: The High-Interest Credit Card of Technical Debt, 2014]

Googleでの経験から14個の負債を紹介
「機械学習は高利息クレジットカード」

「取り急ぎリリースへ（整理などを後回し）」の弊害や時間経過での劣化が大きく，従来と種類・対応法が異なる

- 例： 一つの入力の傾向が変わると全てが変わる可能性がある（「ここだけ変えれば済む」と言えない）

➡ ***「何とかVer. 1.0リリース」は借金の始まり・・・***

「IoT・AI時代」の品質？

「システムが何をするか」のそもそも

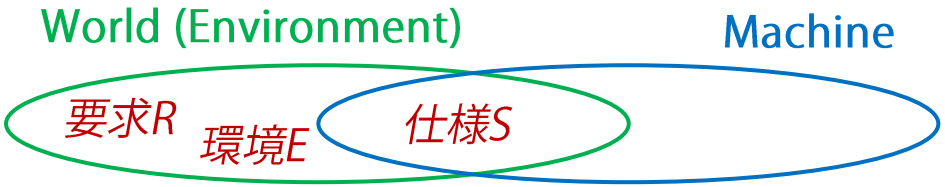
- Zave/Jacksonのモデル
  - 「要求」と「仕様」を別の概念として扱う
  - 「要求」は実現したいことがらであり，
    実現しようと決めたもの
    要求はシステムの制御対象外の領域にも存在する
  - 「仕様」はシステムの振る舞いが満たすべき制約を
    定めたもの（HowではなくWhat）
    仕様はシステムと制御対象世界との境界を定める

    ※「要求仕様」「要件」「ゴール」などの用語の
    使い方は定義・文脈で様々（単に適当なことも）

「システムが何をするか」のそもそも

■Zave/Jacksonのモデル

$$S \quad \models R$$



マシン（開発の成果物・システム）の仕様 S は，

要求 R を満たす？

[Zave et al., Four Dark Corners of Requirements Engineering, 1997]

■例：

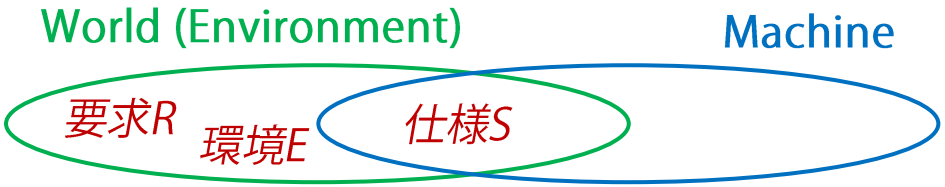
R	• enter 発生回数 ≦ pay 発生回数
S	• pay の発生を検知したら，unlock する • push の発生を検知したら， lock する

[画像： http://www.fujitaka.com/]

「システムが何をするか」のそもそも

■Zave/Jacksonのモデル

$$S,\ E \models R$$



マシン（開発の成果物・システム）の仕様 S は，
想定環境（ドメインの性質） E の下で要求 R を満たす

[Zave et al., Four Dark Corners of Requirements Engineering, 1997]

■例：

[画像： http://www.fujitaka.com/]

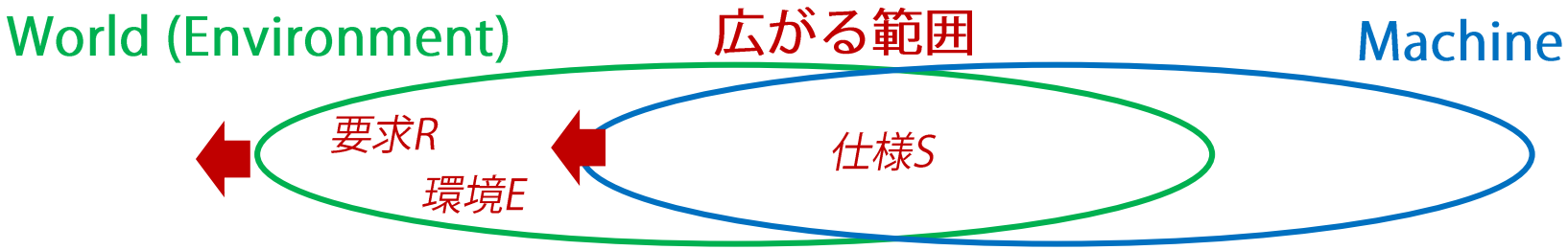
R	• enter 発生回数 ≦ pay 発生回数
S	• pay の発生を検知したら，unlock する • push の発生を検知したら， lock する
E	• push と enter は交互にしか起きないと仮定 • lock が起きてから unlock が起きるまでは push は起こせないと仮定

「IoT・AI時代」の大きなポイント（の一つ）

- 「超スマート社会」「Society 5.0」・・・
  - 物理世界に対する検知・制御能力が増大（IoT/CPS）
  - 演繹的にロジック・ルールが書き出せない
    識別・判断・予測等の機能を実現可能に（AI）

➡ソフトウェアが実世界・人の感覚に
より深く踏み込むように

  - 仕様Sが扱う範囲が広がるということ
  - 要求Rもより広範囲・高度なものを扱うことになる

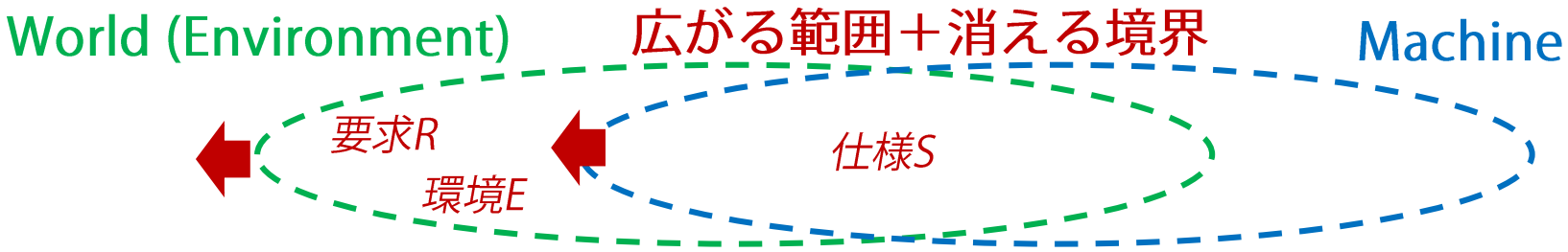


「IoT・AI時代」の大きなポイント（の一つ）

- ソフトウェア制御の範囲（Sの範囲）が広がり，実世界における真のRを直接扱うことに
- RとEにおいて「ここまで扱う」という境界が消え完全・網羅的な列挙，把握や予測は不可能に
  - 例：「様々な」歩行者を識別する
    （押し車付きでも，集団でも，コスプレでも，・・・）
  - 例：「適切に」運転する
    （安全，快適，マナー遵守など，って具体的には？）
  - 例：「適切に」給与判断する
    （男女や人種の差別はしないとして，何はすべき？）

「IoT・AI時代」の大きなポイント（の一つ）

- オープン性・不確かさとの戦いということ
  - 「完全な機能」「リスクゼロ」は無理
    （建前としてすら，そう言うことは無責任に）
  - 開発前・運用前に想定を尽くしたり，確度の高い分析をしたりすることが困難で，継続的な修正・更新が前提に
    （known unknown と unknown unknown の存在）
- 機械学習を用いた場合，自分たちが構築したものの仕様S（とその実装）もオープン・不確かに



自動運転の認識における「不確かさ」

どれだけの「歩行者」を考える？

「歩行者」の定義は？

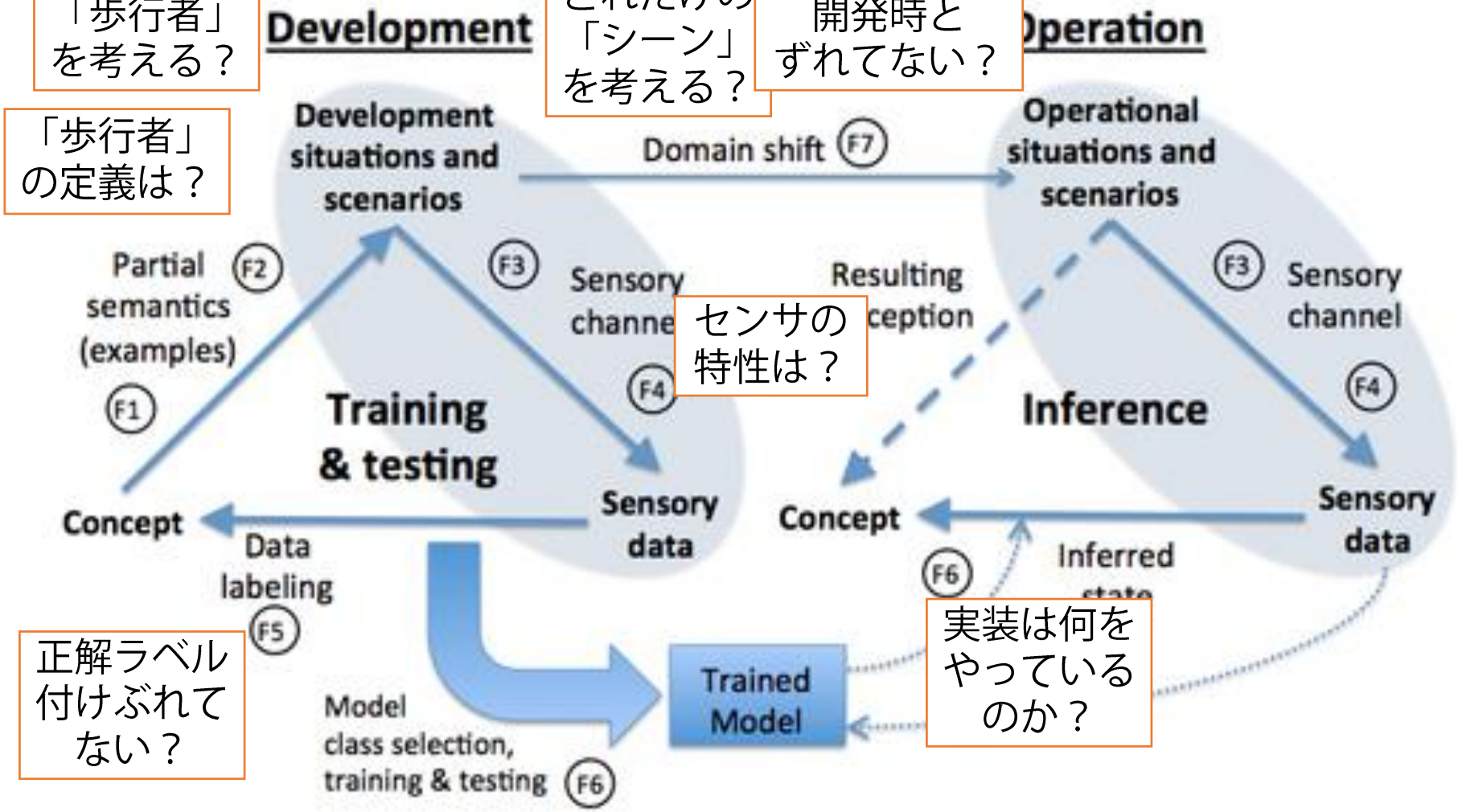
どれだけの「シーン」を考える？

運用時環境は開発時とずれてない？

センサの特性は？

正解ラベル付けぶれてない？

実装は何をやっているのか？



[https://uwaterloo.ca/waterloo-intelligent-systems-engineering-lab/projects/assuredai-safety-assurance-ai-based-automated-driving]

機械学習における
品質保証のための原則・思想

主な品質特性：性能（精度）

- 機械学習で得たソフトウェア部品（モデル）は，ある種の「精度」で性能評価を行うのが基本
  - 機械学習の技術分野において，様々な指標は確立

- 難しさ

何を「100%」の基準とするか？

不完全なものをどう受け入れ役立てるか？

性能（精度）の難しさ（１）

何を「100%」の基準とするか？

- 特にオープンな実世界を入力・動作環境とする場合
- 例： 自動運転のための画像認識

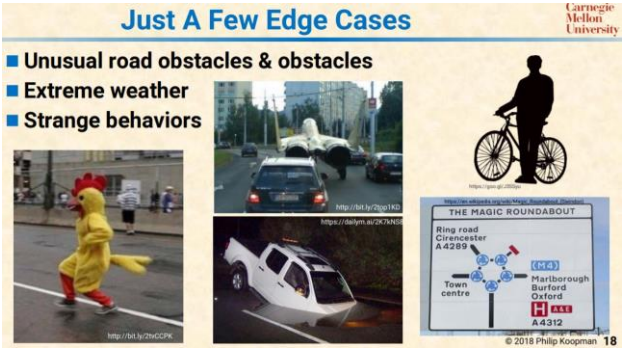
  「10万件の画像でテストしました！」

  ➡「霧の日は試した？」「山道は？」「逆光は？」

しかも，人間にとって意味のあるこれらの区分は，
機械学習で作ったモデルには意味がないかもしれない！
（言葉にしがたい「不得意画像」があるかも）

おまけ：
鏡面タンクローリー

[http://passage.eshizuoka.jp/e831305.html]



おまけ：レアケース
[P. Koopman 2018]

性能（精度）の難しさ（２）

不完全なものをどう受け入れ役立てるか？

- これはむしろ発注者・利用者側の意識も重要
  「頑張って85%の精度が出るようになりました！」
- ➡「よし使おう！」と言える？
- ビジネスに採り入れて，クリック率，工場回転率，安全性増加などの上位目標を評価するしかない
- 何がビジョン？
  例：「どうせ人だって間違えるところ自動化可能」
  例：「人より悪いが人件費よりずっと安い」
  例：「費用かかってでも製品品質が上がる」
- 最後は「（不確実性の下での）勇気ある決断」

先端企業から出ている原則・指針の例（１）

「仮定・想定をテスト」せよ」は方向性の一つ

（従来の「テスト」より広義）

- ベストプラクティス集
  - 例： データの統計を追跡するなどして，問題として顕在化しない失敗を見張れ

[Zinkevich, Rules of Machine Learning: Best Practices for ML Engineering, 2016]

- どれだけ「テスト」できているかの評価スコア
  - 例： 個々のfeatureの入力値範囲や分布が予想と合うか
  - 例： 直接的なメトリクス（精度等）と実影響のあるメトリクス（クリック率等）との相関はどうか，例えばあえて前者を悪くして比較（A/Bテスト）するとどうなる？

[Breck, What's your ML Test Score? A rubric for ML production systems, 2016]

先端企業から出ている原則・指針の例（２）

今までと異なる種類の問題と対応方法

- CACE: Changing Anything Changes Everything
  - n個の入力データのうち1個の傾向が変わると，他のデータの重要度・利用法が変わってしまう
  - ⇨ それぞれを独立して分析・統合（可能な場合のみ），様々な観点での可視化，精度変化に敏感になるよう調整
- フィードバックループ
  - 「クリックされそうなニュースの見出しを大きく」
    たまたまクリック→似たものが大きく→クリック
    →似たものが大きく・・・（数週間後に現れる）
  - ⇨ 注意深く影響関係を分析

[Sculley et al., Machine Learning: The High-Interest Credit Card of Technical Debt, 2014]

研究例：「あら探し」テスト・検証技術

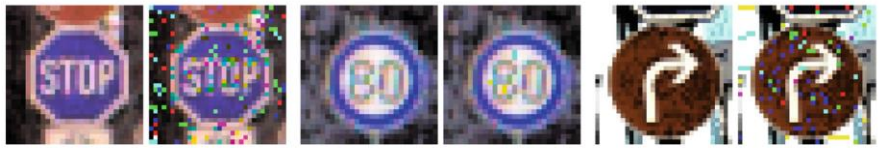
- 既存画像に雨や霧，覆い，ゆがみ等を加えたとき，識別や判断の結果が変わってしまうケースを探す
  - サーチベースドテスティング（進化計算を使った自動テスト，FacebookのSapienzが最近は話題）
  - 形式検証技術で網羅的に
  - システム全体に問題（事故等）をもたらす場合に絞り探索

1.1 original

1.2 with added rain

[Pei et al., DeepXplore: Automated Whitebox Testing of Deep Learning Systems, 2017]

画像元：[Tian et al., DeepTest: Automated Testing of Deep-Neural-Network-driven Autonomous Cars, 2018]



"stop" to "30m speed limit"

"80m speed limit" to "30m speed limit"

"go right" to "go straight"

画像元：[Huang et al., Safety Verification of Deep Neural Networks, 2017]

[Mirman et al., Differentiable Abstract Interpretation for Provably Robust Neural Networks, 2018]

[Dreossi et al., Compositional Falsification of Cyber-Physical Systems with Machine Learning Components, 2017]

おわりに

おわりに

セキュリティにとって従来から重要だった
リスク管理や継続監視などが主役

これまで存在しなかった問題ではない？

ソフトウェアにかかわる人々がサボってきた・
やりきれてなかったことが突きつけられている

- 画面の外の実世界での意義や安全性，ユーザの感性的な満足の追求，探索的・試験的な投資や問題解決，
  仮説と検証による科学的な判断や評価，常に測定，
  数学の理解・活用，技術を理解した経営・管理，
  現実問題の数学的問題（最適化等）への帰着，
  実行時検証やテスト自動化（入力生成・疑似オラクル），
  といったことへの本気のパラダイムシフト，・・・

楽しんで切り拓いていきましょう！