



AIプロダクトに対する品質保証の 基本的考え方

2019/3/27 (水)

ソフトウェアテストシンポジウム東京 2019

石川 冬樹 (国立情報学研究所 / QA4AIコンソーシアム)

小川 秀人 (日立製作所 / QA4AIコンソーシアム)

西 康晴 (電気通信大学 / QA4AIコンソーシアム)



本日の流れ

AGENDA

「AIプロダクトに対する品質保証の基本的考え方」の解説
QA4AIコンソーシアムの2018年度成果物の α 版を基に解説します

会場からのQ&Aのパネル
石川、小川、にしの3名が会場からの質問にお答えします



石川 冬樹

国立情報学研究所 アーキテクチャ科学研究系 准教授 および 先端ソフトウェア工学・国際研究センター 副センター長。ソフトウェア工学および自律・スマートシステムに関する研究・教育に従事。電気通信大学 大学院情報理工学研究科 客員准教授。日本ソフトウェア科学会 機械学習工学研究会 生査。AIプロダクト品質保証コンソーシアム 運営副委員長。日本ソフトウェア科学会理事。博士(情報理工学)。



小川 秀人

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ー 生管研究員 兼 ソフトウェアモダライゼーションラボ ラボ長。AIプロダクト品質保証コンソーシアム運営副委員長。ソフトウェア工学の研究およびソフトウェア開発プロジェクトへの技術適用に従事。博士(情報科学)。



西 康晴

電気通信大学 大学院理工学研究科 講師。AIプロダクト品質保証コンソーシアム(QA4AI)運営委員長。NPO法人ソフトウェアテスト技術振興協会(ASTER)理事。ISO/IEC JTC1/SC7/WG26 (ISO/IEC/IEEE29119など)国内委員会生査。日本科学技術連盟ソフトウェア品質委員会(SQiP)副委員長。ソフトウェアのテストやQA、サービス学などに関する研究・教育・コンサルティングに従事。博士(工学)。

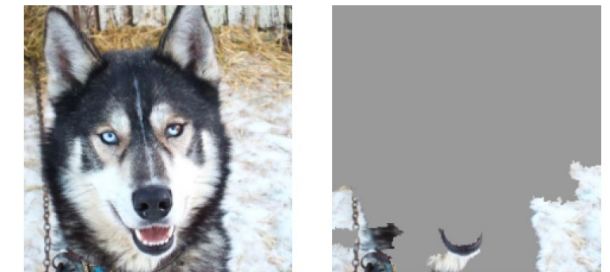
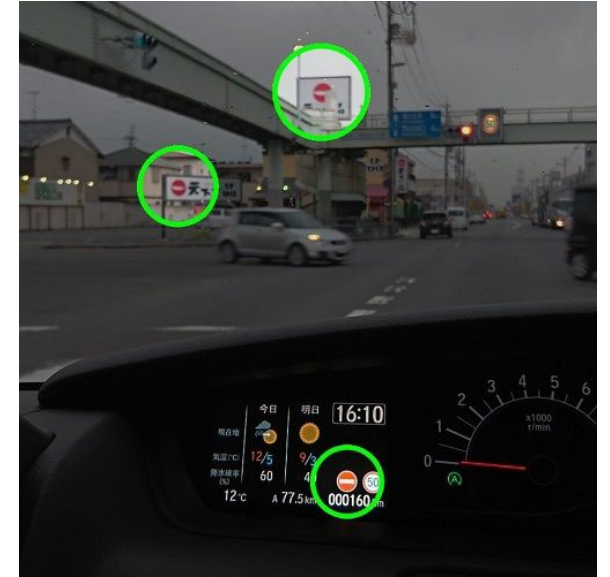
背景

- 各企業でAIプロダクト（主に機械学習、特にDNNを使った製品やサービス）を開発・利用するようになり、社会的・経済的影響が大きく、競争力の源泉として破壊的影響を持つようになってきた
- しかしどのように「品質保証」を行っていけばいいのか、についての知見は誰も持っていないように思われ、困っている組織は非常に多く見られる
- そこで産官学でオープンなコンソーシアムを設立し、AIプロダクトの品質保証について産業界に寄与する必要がある



AIプロダクトの特徴

- 主にML技術を前提としている
 - 帰納的開発である
- 確率的ふるまいをする
- CACE性がある / 非線形である
- 自動化が必要
- 詳しくは以下を参照するとよいでしょう
 - IoT/AI時代のテストング・検証技術の最前線
(石川冬樹 / SQiP研究会特別講義 2019/1/11)
 - <https://www.slideshare.net/Fuyukilshikawa/iotai-128126704>



A woman with long dark hair, wearing a dark jacket and a necklace, is shown from the chest up. She has a thoughtful or questioning expression, with her hand near her face. The background is a solid light blue color, overlaid with several large, semi-transparent black question marks of varying sizes. The text is centered in the lower half of the image.

はたして
品質保証（QA）は
どうやればいいんだろう？

QA4AIコンソーシアムとは

- 「AIプロダクト品質保証コンソーシアム」
 - <http://qa4ai.jp/>
 - 産官学連携のためのオープンなコンソーシアム
 - 様々な類似の活動をハーモナイズしたい
 - 現在40名弱のメンバと3組織から構成されている
 - 2018年の春に設立
 - 世話人：西康晴（電気通信大学）、石川冬樹（国立情報学研究所）、小川秀人（日立製作所）
- QA4AIコンソーシアムの特徴
 - オープンなコラボレーションの場
 - 議論し（て成果物を作り）たい人だけが集まる場
 - 月1回の会合とslackによる議論・情報交換
 - 会費などは不要である
 - マルチドメイン
 - 現在スマートスピーカー、産業用プロセス、自動運転、生成系の4ドメインがある



QA4AIコンソーシアムのコンセプト

- 開発組織にも顧客にも（相応の）安心をして欲しい
- QAがジャマをしてはいけない
- オープンで技術指向の組織運営
- マルチドメイン
- 類似の活動を行っている全ての組織とハーモナイズする
 - MLSE / 産総研-NII / JAXA-自工会 / 名古屋大サポイン / QAMLなど
 - それぞれの主要メンバがQA4AIコンソーシアムに名を連ねている

QA4AIコンソーシアムの活動

1. AI プロダクトの品質保証技術の調査・体系化
2. AI プロダクトの品質保証技術の適用支援・応用例の収集
3. AI プロダクトの品質保証技術の研究開発の促進
4. AI プロダクトの品質保証レベルの策定と、各レベルに必要とされる技術の対応付け
5. AI プロダクトの開発・提供組織による4)の自己宣言の方法の策定
6. 社会への4)、5)の提示と啓発

ガイドラインの
作成

タイムライン

- α 版：3月頭に内部リリース
 - AIプロダクトの品質保証において、
どういった方向性で考えればよいか、をまとめた段階
 - 具体的にどのようにQAを行えばよいかの基準ではない
- β 版：4月頭に内部リリース
 - α 版+各ドメイングループのガイドライン
- 正式版リリース：5月中旬
 - 5/17（金）午後にイベントを行う
 - 予定：QA4AI、産総研-NII、JAXA（-自技会）、名古屋大サポイン、QAML

ガイドラインの構造

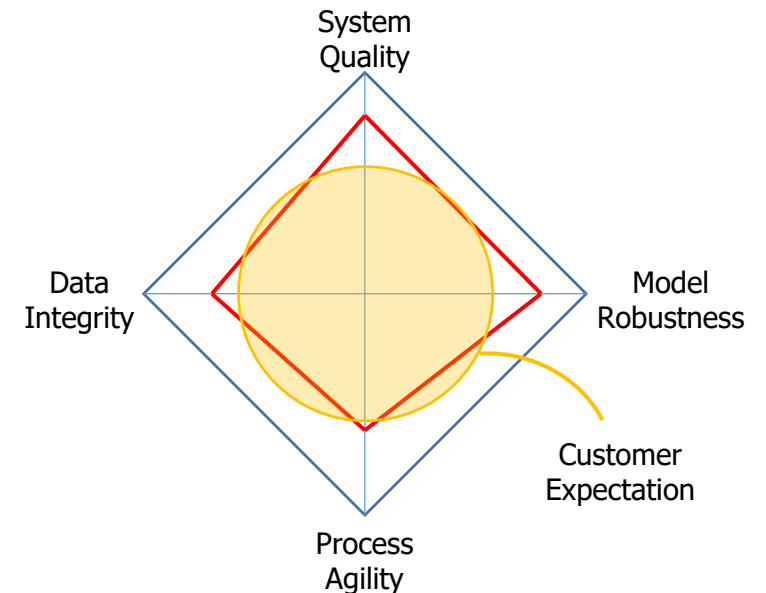
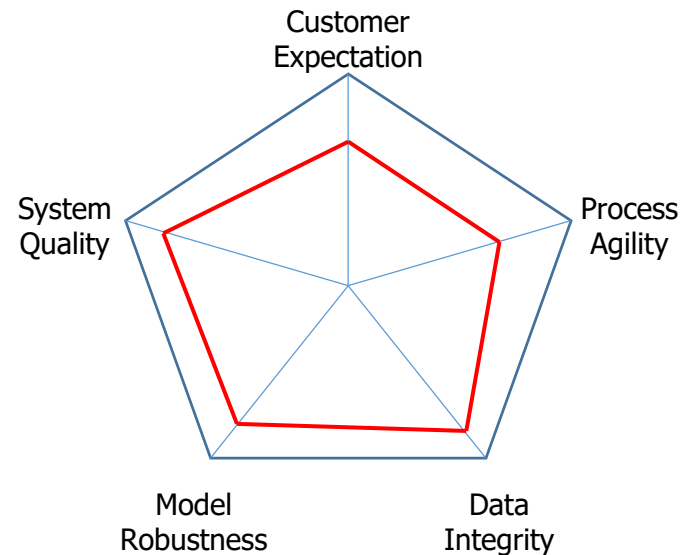
- コンソーシアムの趣旨・目的
- 品質保証活動の5つの軸とバランス
 - 類似の活動を行っている組織も似た方向で考えているらしい
- 品質保証技術のカタログ
 - SQUBOK V3と同期させる - 重複投資はなるべく避ける
- 各ドメインでの成果物
 - スマートスピーカー、産業用プロセス、自動運転、生成系の4ドメインによるガイドライン

品質保証の枠組み

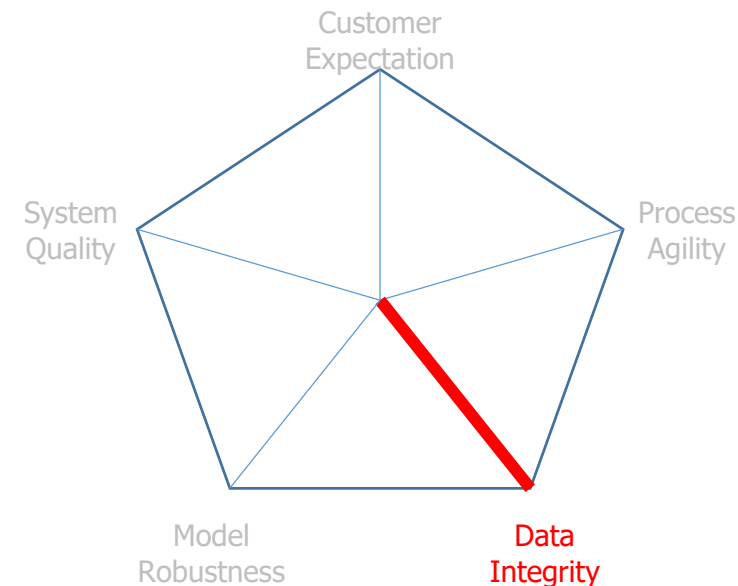
- 品質保証活動の5つの軸とバランス
- 品質保証技術のカタログ

品質保証活動の5つの軸とバランス

- 5つの軸とそのバランスでQAの活動を評価する
 - Data Integrity
 - Model Robustness
 - System Quality
 - Process Agility
 - Customer Expectation



Data Integrity



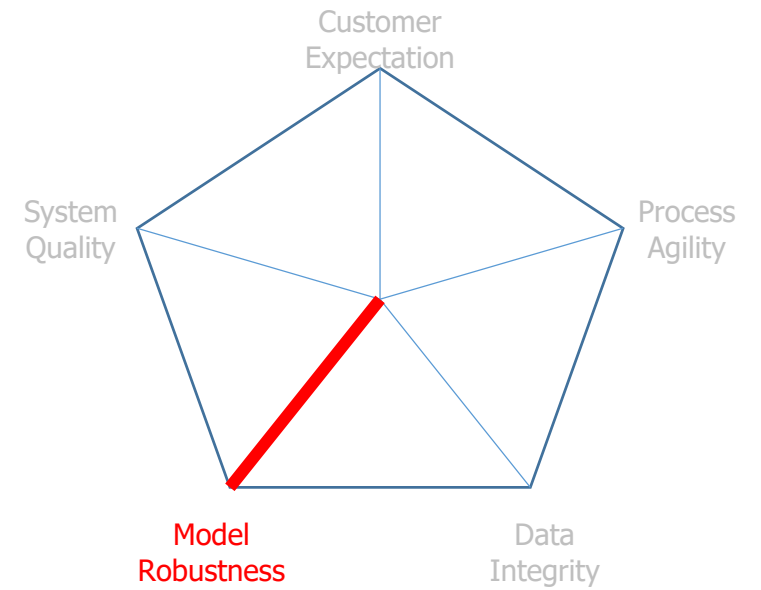
- データがきちんとしているか

- 量は充分か、コストは適正か
 - 意味のある量か
 - 「かさ増し」しても大丈夫か
- ラベルは妥当か
- 求める母集団のサンプルか、実データか、 unnecessary データが含まれていないか、データに関する要求事項を満たしているか、データに関する制約に反していないことを監視しているか
- 必要な要素を適切に含んだサンプルか
- 偏りやバイアス、汚染は無い
 - 自分たちが考えている「偏りを発生させるもの」だけでよいか
- 複雑すぎないか、単純すぎないか
- それぞれのデータは常識的な値か、外れ値は本当に外れているデータか、欠損に意味はないか、外れ値や欠損値の扱いは適切か
- データ内の性質（多重共線性など）は適切に考慮されているか
- 含むべきでない母集団のデータと混ざっていないか
- 学習用データと検証用データは独立しているか
- 学習時のふるまいと提供時のふるまいに齟齬はないか
- 所有権や著作権・知的財産権、機密性、プライバシーは適切に考慮されているか
- オンライン学習を行う場合、その影響を適切に考慮しているか
- 学習用プログラムやデータ生成プログラムの不具合によってデータの意味が毀損されないか
- など

Model Robustness

- モデルがきちんとしているか

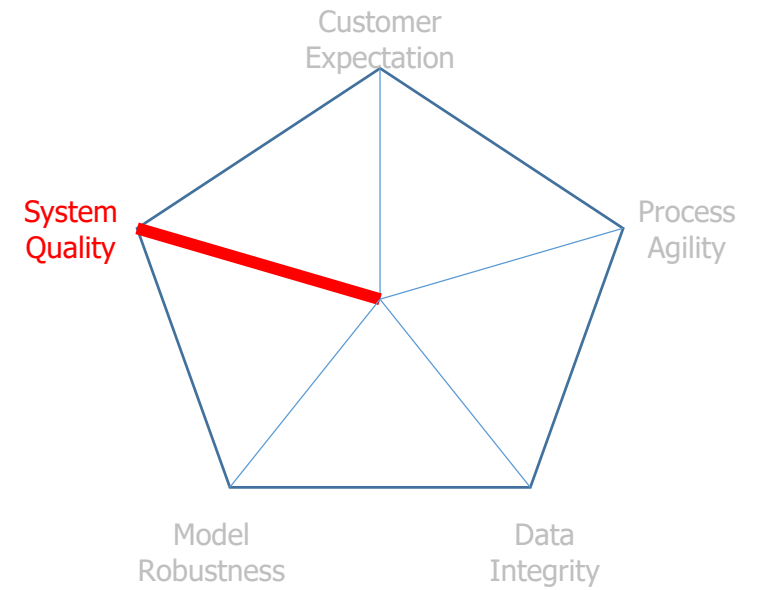
- 正答率、適合率、再現率、F値
- 汎化性能
- AUROC (Area under Receiver Operating Characteristic) など
- 学習は適切に進行したか
- 局所最適に陥っていないか
- 適切なアルゴリズムかどうかの検討は行ったか
- 適切なハイパーパラメータかどうかの検討は行ったか
- 十分に交叉検証などを行ったか
- ノイズに対して頑健か
- 十分に多様なデータで検証を行ったか
 - 数理的多様性、意味的多様性、社会的文化的多様性
- デグレードは許容可能な範囲か、デグレードの影響範囲を把握できているか、学習は再現可能か
- モデルが陳腐化していないか
- 実データに対する予測品質が劣化していないか
- 目標メトリクスの計測が難しい場合、計測できるメトリクスとの関連は妥当か
- など



「精度が高ければ品質がよい」
ではない！

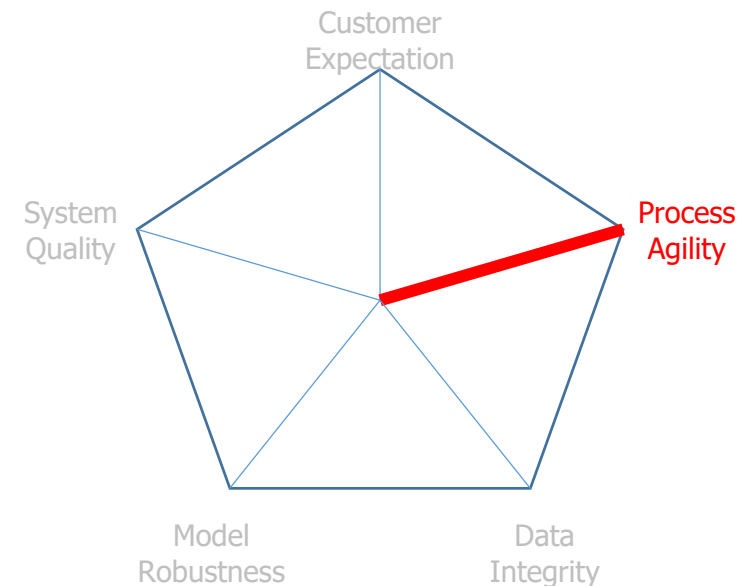
System Quality

- システム全体として価値は高いか、何か起きててもなんとかなるか
 - システムを全体として、および意味のある単位で評価を行ったか
 - 価値は適切に提供されているか
 - 発生しうる品質事故の致命度は許容できる程度に低く抑えられているか
 - 品質事故の致命度はドメインによって異なる
 - 身体や生命への危害、経済的ダメージ、社会や環境への影響、不快感、魅力の低さ、意味のなさ、反倫理
 - システムの事故到達度・安全機能・耐攻撃性は充分か
 - 防護機構のよさ・多さ / 回避性や制御性の高さ / 自己修復性
 - 品質事故を引き起こしうる事象の発生頻度は低いと見積もることができるか
 - 事象の発生頻度、検証した事象の網羅性
 - 環境統制性（意図、偶発、攻撃）
 - 保証性、説明可能性、納得性は充分か
 - AIの寄与度を抑えられているか
 - 性能などシステムのふるまいが劣化していないか
 - システムが依存する他の（AIの、もしくは非AIの）システムの変更は迅速かつ適切に反映できるか、不具合の影響を充分低く抑えられるかなど



※Quality、Reliability、
Dependability、
Safety、Securityなどを
ここでは厳密に区別していない

Process Agility

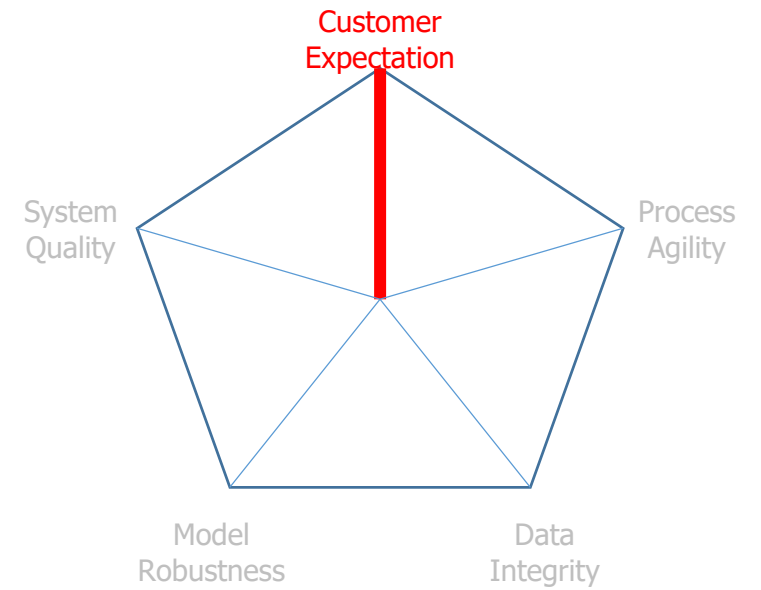


- プロセスが機動的か

- データ収集の速度とスケーラビリティは充分か
- 充分短い反復単位で反復型開発は行っているか、モデル・システムの品質向上の周期は充分短いか
- 運用状況の継続的なフィードバックは頻繁か
- リリースロールバックは簡便で迅速に行えるか
- よりよくなっていく見込みはあるか
 - 新しい特徴量を迅速に追加できるか、モデルを迅速に改善できるか / 学習や推論のデバッグを行う手段を有しているか
- 段階的リリースやカナリアリリースの度合は適切か、リリース直前にシステム全体やモデルの評価を行っているか
- 開発・探索・検証・リリースなどの自動化は充分か
- データ、モデル、環境、コード、出力などについて適切に構成管理が行われているか
- 開発者やチームは技術的に充分納得し共感しているか
 - 自分たちがいま何をどう作っているか / 何かが起こったらどこまで戻せばよいか / どのようなリスクがあるか / それにどう対処すればよいか概ね分かっているか、分からないことが分かっているか
- 開発チームは適切な能力を持った人財を備えているか
 - 機械学習やデータサイエンスの「専門家」やドメインの専門家は含まれているか？
- 経験を技術に反映させられているか
- 開発チーム外のステークホルダーは充分納得しているか
- など

Customer Expectation

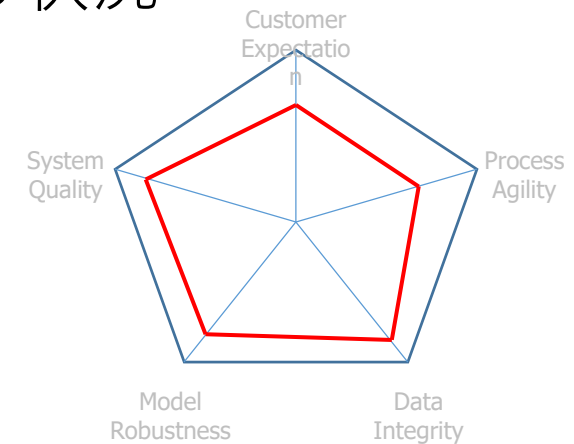
- (良くも悪くも) 顧客の期待は高いか
 - 顧客側期待の高さ
 - 確率的動作という考え方の非受容
 - 顧客側でのリスク・副作用の無理解や非受容
 - 継続的実運用への近さ
 - PoCや β リリースという概念を理解していない度合
 - データの量や質に対する認識の甘さ
 - 狙っているのが「人間並み」なのかどうか
 - 法規制、著作権や知財、プライバシー、コンプライアンス、社会的受容が必要な度合
 - “合理的”説明を求める度合、“外挿”や“予測”をしたがる度合
 - “原因”や“責任(者)”を求めたがる度合
 - 納得感を共感する風土や雰囲気、仕事の進め方の少なさ
 - 顧客担当者・チームで意思決定できる権限の少なさ・範囲の狭さ
 - など



バランスがよい・バランスが悪いとは？

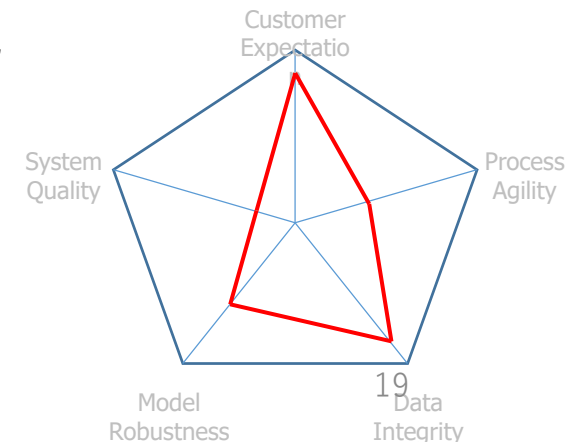
- バランスがよい：
顧客の期待に沿っていると保証できることが期待できる状況

- DI/MR/SQ/PAに不足がない状況
 - DI/MR/SQ/PAのどれか一つが欠けていても品質が確保できたとは言えない
- (DI/MR/SQ/PAに対して) 顧客の期待が適切である状況
 - 顧客の期待を適切に把握し、可能ならば期待を適切な程度に収め、DI/MR/SQ/PAが総合的に期待を満たすようにする
 - 顧客の期待を適切な程度に収め保つ（期待値コントロール）も重要な活動になる



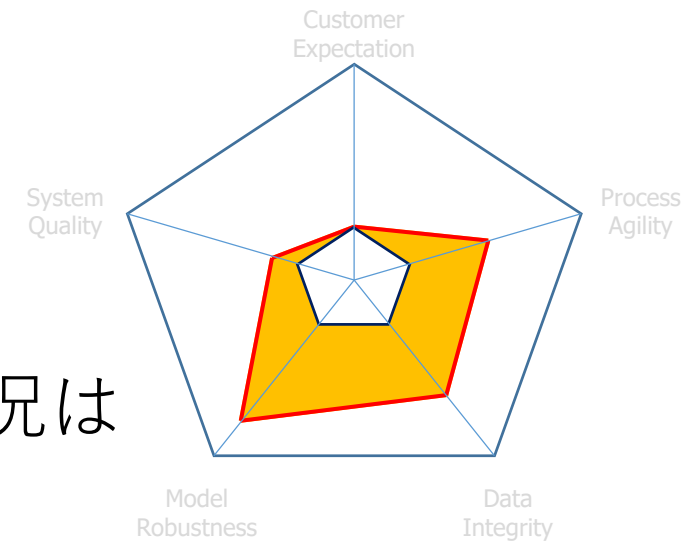
- バランスが悪い：
顧客の期待に対してDI/MR/SQ/PAが不足している状況

- DI/MR/SQ/PAのいずれかが不足している状況
 - 品質は「最弱リンク」で作用することが多いので、最も劣っているものが効いてくる
- 顧客の期待が高すぎる状況
 - 可能ならば期待値コントロールが重要である
 - 短期的で個別的な期待値コントロールと、中長期的で社会的な期待値コントロールの両方が本来必要である



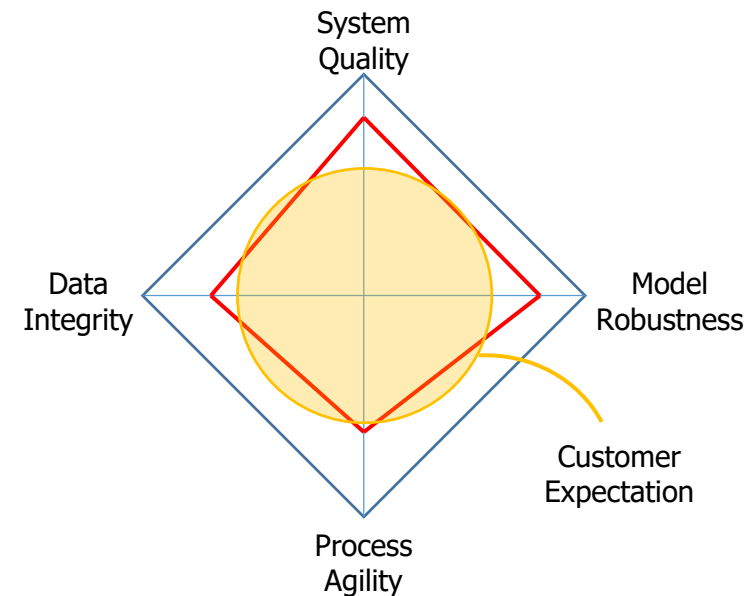
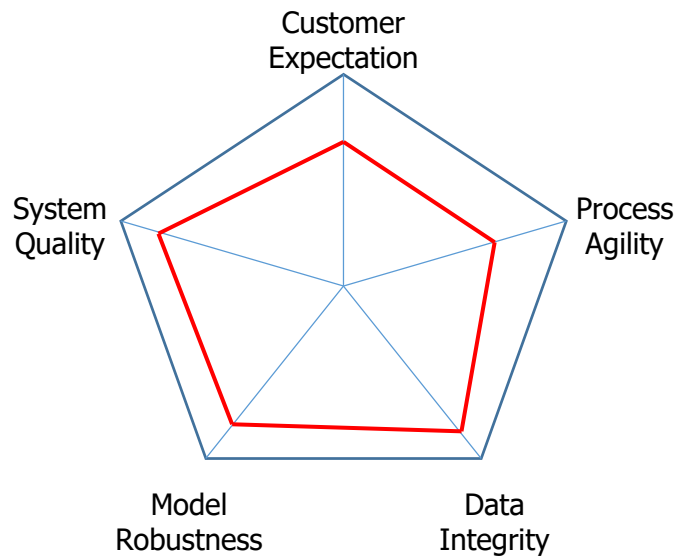
余力と過剰品質

- DI/MR/SQ/PAが顧客の期待を上回っている状況は「余力」であって過剰品質ではない
 - 将来的に開発段階が進んだり運用範囲が広がると顧客の期待が高めるため、その時に備えた活動を事前に行っていると捉える
 - 一つのプロジェクトに余力があると、他のプロジェクトの品質向上の基盤にもなりうるため、「過剰」ではない
 - 「過剰品質」はむしろ、プロジェクトやQAが顧客の期待やビジネスの方向性、自分たちの開発の状況や技術力を見失い、一体自分たちが何をやっているのかを分からずにただ技術的手数だけを増やしている状況を指す



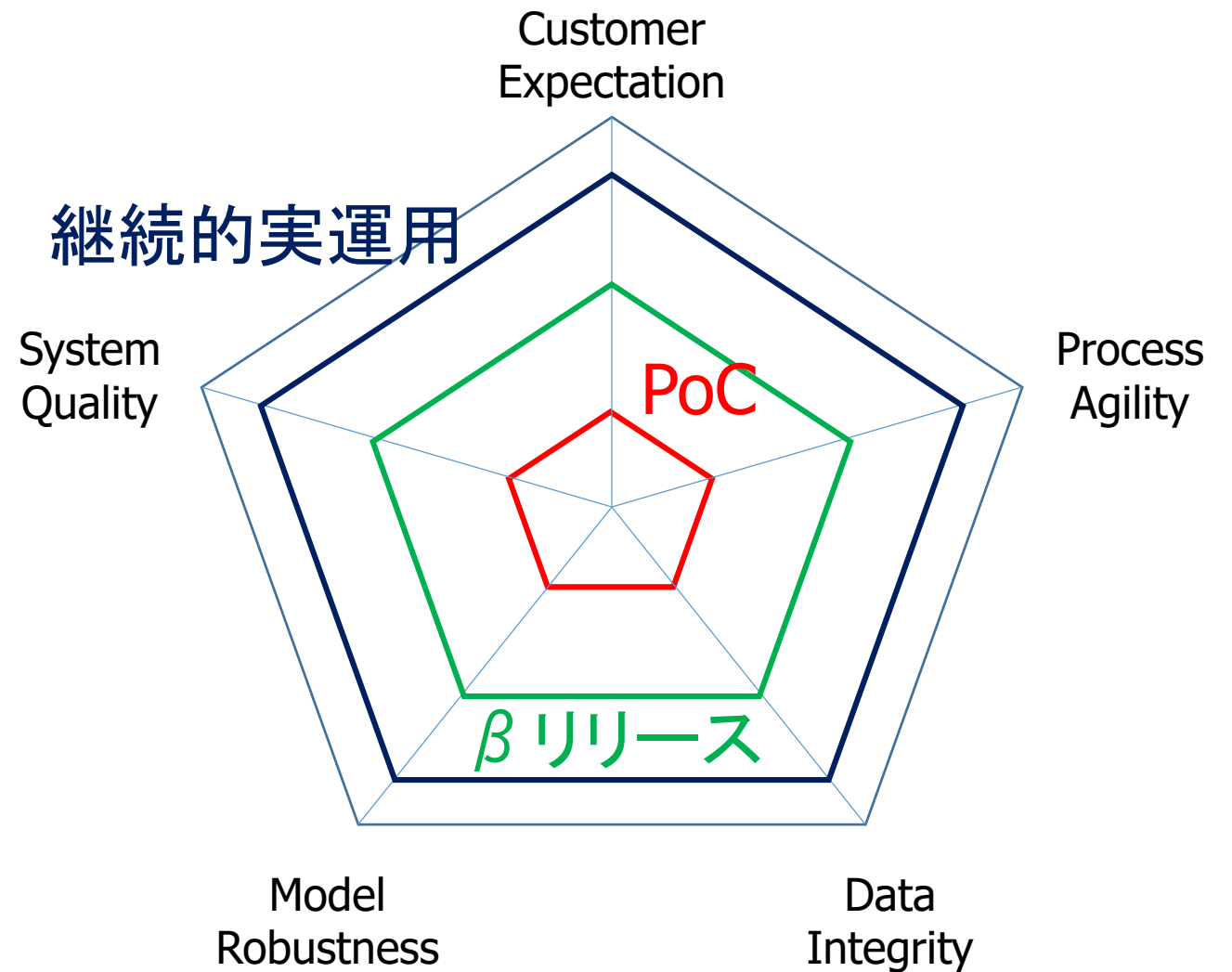
品質保証活動のバランスの可視化

- DI/MR/SQ/PAおよびCEの5つの軸のバランスを可視化するとよい
 - 可視化の方法は色々考えられるので、組織ごとで構わない
 - 例えば以下の2つの可視化がありうる
 - 1) 5つの軸を並立させる可視化
 - 2) CEとDI/MR/SQ/PAとを対比させる可視化



開発の段階と面積

- 開発の段階
(PoC、 β リリース、
継続的実運用など) ごとに
面積は異なるが
バランスは取れている
必要がある



品質保証技術のカatalog

- MLプロダクトには、
主要なソフトウェア品質保証技術が
無力化ないし限定的になってしまう
 - プロセス
 - 構成管理
 - メトリクス
 - レビュー
 - テスト
 - 品質の保証
 - MLSE技術

品質保証技術のカatalog

• プロセス

- 「こう作ればうまく作れる」「こう作ると失敗する」という演繹的因果関係が不明なため、作業手順を定義し遵守しても品質は上がらない
- ML技術とプロセスをより密接に結びつける必要があるが、従来型のソフトウェア開発であってもできている組織は少ない
- 学習データと検証用データを厳密に管理したりプライバシーを保護するなどのプロセスは必要となる
 - が、それはAIプロダクトの品質保証そのものの話とは異なる

• 構成管理

- プログラムコードを中心とした構成管理は排他的バージョン管理程度の意味しかなく、実務的に役に立たない
- 学習データ、ハイパーパラメータ、モデル、プログラム、検証データなどが一体化した「探索環境」が必要となる

品質保証技術のカatalog

• メトリクス

- プロセスメトリクス
 - プロセスが無力であればプロセスメトリクスも無力である
- プロダクトメトリクス
 - モデルの情報の定量的集約ができるほどDNNに対する知見があるわけではない
 - ある程度であれば線形統計のメトリクスが有用な場合がある
 - データの質に関するメトリクスを測定し評価するのは可能であるが、従来のソフトウェアメトリクスとは大きく異なる
 - SQuaRE (ISO/IEC 25Kシリーズ・旧9126) を読み替えるという試行はなされている

• レビュー

- 「こう作ればうまく作れる」「こう作ると失敗する」という演繹的因果関係が不明なため、中間成果物（学習中モデル）に対する静的な検証はできない
- データの（量や）質に関するレビューは可能であるが、従来のソフトウェアレビューとは大きく異なる
- モデル構築に関する様々な知見の共有という意味でのレビューは重要である

品質保証技術のカタログ

• テスト

- プログラム構造がほとんど品質に寄与しないため
従来のホワイトボックステストは意味をなさない
 - DNNの構造に基づいてホワイトボックステストをすればいい...?
 - ニューロンカバレッジ
 - DNN内のニューロンが全て活性化するようにテストケースを生成する
 - 効果があるという論文はあるが、
本当に効果的かどうか広く実証されているわけではない

品質保証技術のカatalog

• テスト

- CACE性によって同値分割が意味をなさないため、テストケースを減らすことができない
 - テスト設計やテスト実行の自動化が必須となる
 - モデルベースドテストやサーチベースドテスト
 - テスト観点や最適化目標の「テスト要求分析」がキーになる
 - CI環境が構築されていないと非常に厳しい
 - オラクル（期待結果）が自動生成できないとテスト自動化は非常に困難になる
 - MLに限った話ではない
 - メタモルフィックテスト
 - 「入力を変えると出力はこう変わるはず」という関係からテストケースを生成する
 - 形式検証技術による頑健性検証
 - 画像認識に対して特定の画像操作を行い、認識結果が変わらないことを網羅的に検証する

品質保証技術のカatalog

• テスト

- AIにはAIでテストすればいいじゃないか！
 - サーチベースドテストは広い意味のAIである
 - FacebookではSBTを応用して「自動でテストして自動でデバッグ」している？
 - Sapienz / SapFix
 - 日本にも同様の技術のスタートアップが出てきている
 - RPA(Robotic Process Automation)のアプローチで、AIを用いてテストを自動化する取り組みが盛んになっている
 - AI「で」テストするものであって、AI「を」テストするものではないが、テストを自動化するという考え方はそれなりに使えるのかもしれない
 - 日本にも同様の技術のスタートアップが出てきている
 - GAN(Generative Adversarial Network)などでテストケースを生成する
 - 「AIを誤解させる技術」を使ってピンポイントにテストを行う
 - GANなどで「ヤバイ」テストケースを見つけることは可能だが、それが実世界で問題になるかどうかまではあまり議論されていない

品質保証技術のカタログ

• 品質の保証？

- 「品質の保証」とは一体何を意味するのか、をもう一度問い直さなければならない
- 品質とは何か？
 - 結果としての品質
 - モデルの誤判別の防止？ロバスト性？事故の予防？機能性？非機能特性？価値？社会的問題解決？
 - 「致命的」ってなに？経済性？倫理？社会的受容？
 - 組織能力としての品質
 - 日本的品質管理では組織能力の高さも「品質」と捉える
- 保証とは何か？
 - そもそも我々は「保証する」ことをきちんと理解できているのか？
 - プロセスの遵守？文書の監査？指標の達成？チェックリスト？（感覚的）網羅テスト？
 - 「品質が保証されたと感じる」ために何が必要なのか？
 - eXplainable AI技術として盛んに研究・議論されている

品質保証技術のカatalog

• MLSE（機械学習工学）技術

- そもそも品質向上・確保は下流や品質保証組織だけで取り組むものではない
 - 開発上流から取り組む必要がある = MLSE（機械学習工学技術）
- 様々なMLSE技術への取り組みがなされる必要がある
 - 適切な安全機能を持つようなアーキテクチャは開発可能なのか？
 - 緊急停止ボタンモジュールをMLでつくることは可能か？
 - MLモデルモジュールを確率動作するコンポーネントと捉え、システムの故障率を計算することは果たして妥当なのか？
 - 演繹的開発を前提としたソフトウェアに対する機能安全規格の考え方は適用できるのか？
 - 安全要求やハザード解析からトレーサビリティを確保しても確率動作してしまう
 - プロセスの遵守は意味をなさないし適用技術でレベル分けできるほど成熟していない
 - 「ある条件で必ず誤動作するが全体として確率動作し、故障率を実績主張できないコンポーネント」は、ハードウェアに対する機能安全の枠組みに合致するのか？
- MLSE技術全体を、我が国全体で発展させていく必要がある

QA4AIコンソーシアムのこれから

- 5/17（金）にガイドラインの正式版を公開する
 - 東京でイベントも行います / 告知はQA4AIのWebサイトで行います (<http://qa4ai.jp/>)
- 参画企業でガイドラインの試行を行う
 - 試行したい組織はぜひQA4AIコンソーシアムにご参画ください
- 年次程度にガイドラインのアップデートを行う
 - 技術的にはSQuBOK V3と同期する
- 新たなドメインでの議論を始める
 - AIOCRのドメインは手を挙げている
- 類似した活動のハーモナイズを行う
 - 産総研-NII / JAXA-自工会 / 名古屋大サポイン / QAMLとハーモナイズするために各組織からQA4AIコンソーシアムの運営委員会に参画し、より密な連携を図る

本日の流れ

AGENDA

「AIプロダクトに対する品質保証の基本的考え方」の解説
QA4AIコンソーシアムの2018年度成果物の α 版を基に解説します

会場からのQ&Aのパネル
石川、小川、にしの3名が会場からの質問にお答えします

時間の許す限り質問に答えます



石川 冬樹 / 小川 秀人 / 西 康晴