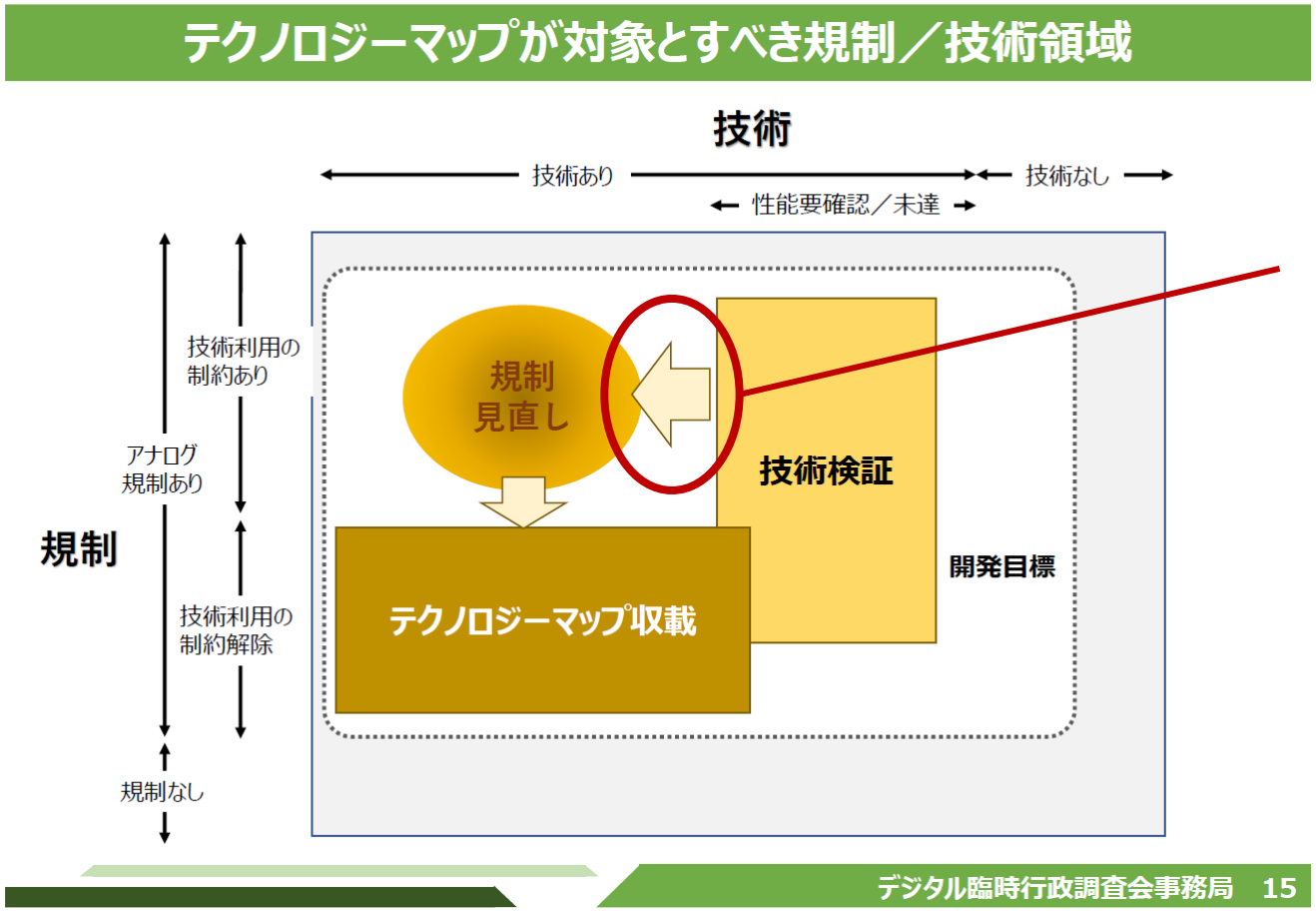
大学共同利用機関法人 情報・システム研究機構
国立情報学研究所
National Institute of Informatics

AI品質へのアプローチ紹介

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

文脈の理解



トラスト確保のアプローチは？？

- 対象技術のトラストが十分かどう評価する？

- 十分あるとしたときに，対象技術に対する個々の適用事例におけるトラスト確保のため必要な施策（新しい規制項目など）は何か？

AIプロダクト・サービス個々に対する
トラスト評価・確保のための
既存ガイドラインを紹介

プロダクト・サービス品質評価の典型的な構造

枠組み・ガイドライン

品質特性（観点） → 評価指標　（具体化）

システムの重要さや要求の強さの区分

- 可用性 → 稼働率，障害復旧水準，…
- 性能 → 時間性能 → 応答時間，…
  → 空間性能 → 最大受容データ量，…
- …

- 企業内利用なら「社会的影響が殆どないシステム」
- …
- 国民生活・社会経済のインフラなら「社会的影響が極めて大きいシステム」

求める基準　（かけ合わせて設定）

「社会的影響が極めて大きい」ならば，
- 「稼働率」は「1年間で数分程度の停止まで許容」
- 「応答時間」のサービスレベルを規定すること
- …

各システムの契約・仕様・評価基準

個別の適用

今回は・・・　「社会的影響が極めて大きいシステム」

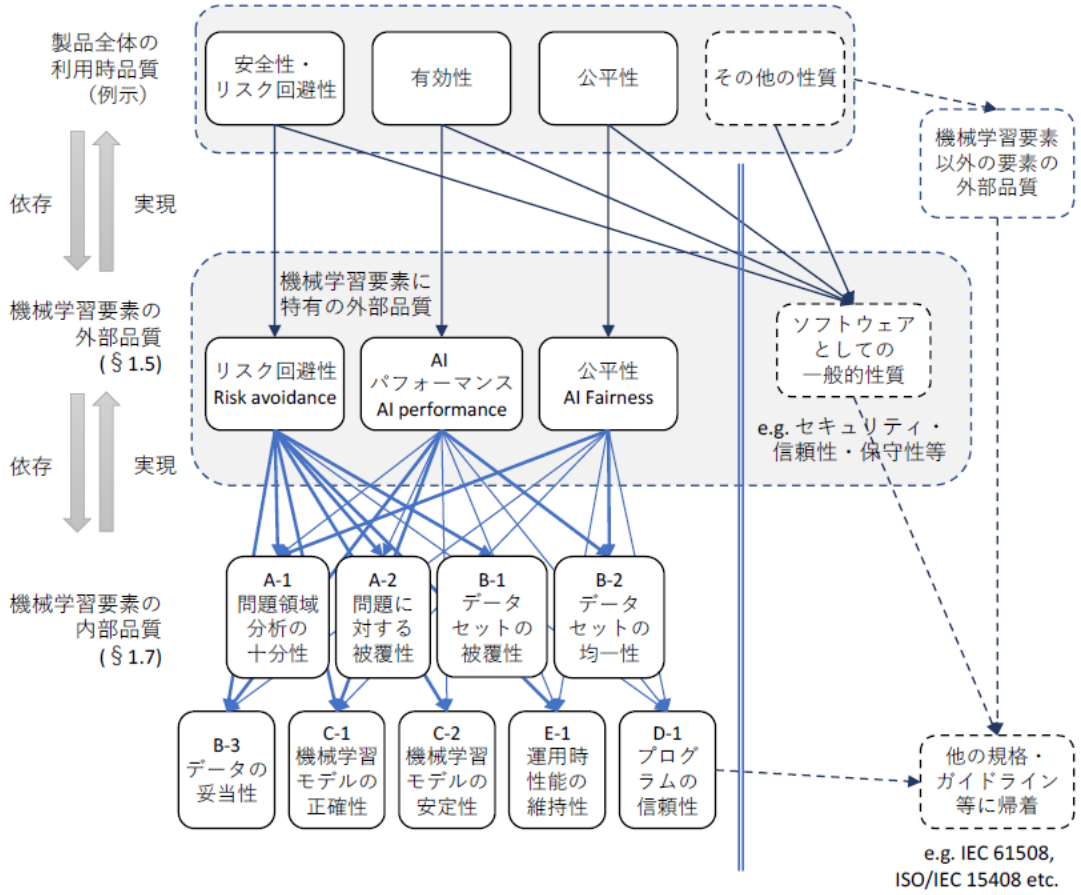
- 「稼働率」は「1年間で数分程度の停止まで許容」
- 「応答時間」は「平均が3秒以内」と規定
- …

機械学習型AIプロダクト・サービスの品質評価

- 対象：データを用いた訓練によりシステム機能を構築
- 国内では2つのガイドライン（2019/2020～）
  - AIQM（経産省系・産総研を中心に）
  - QA4AI（ボランティア）
- 抽象度が高いのでそれぞれ具体化
  - プラント保安，音声インターフェース，文字読み取りなど
    個別領域ごとにサブガイドライン
- 他のガイドラインも大きく異なるものはないはず
  - EUのガイドラインは，倫理・人権を中心とした言い回し

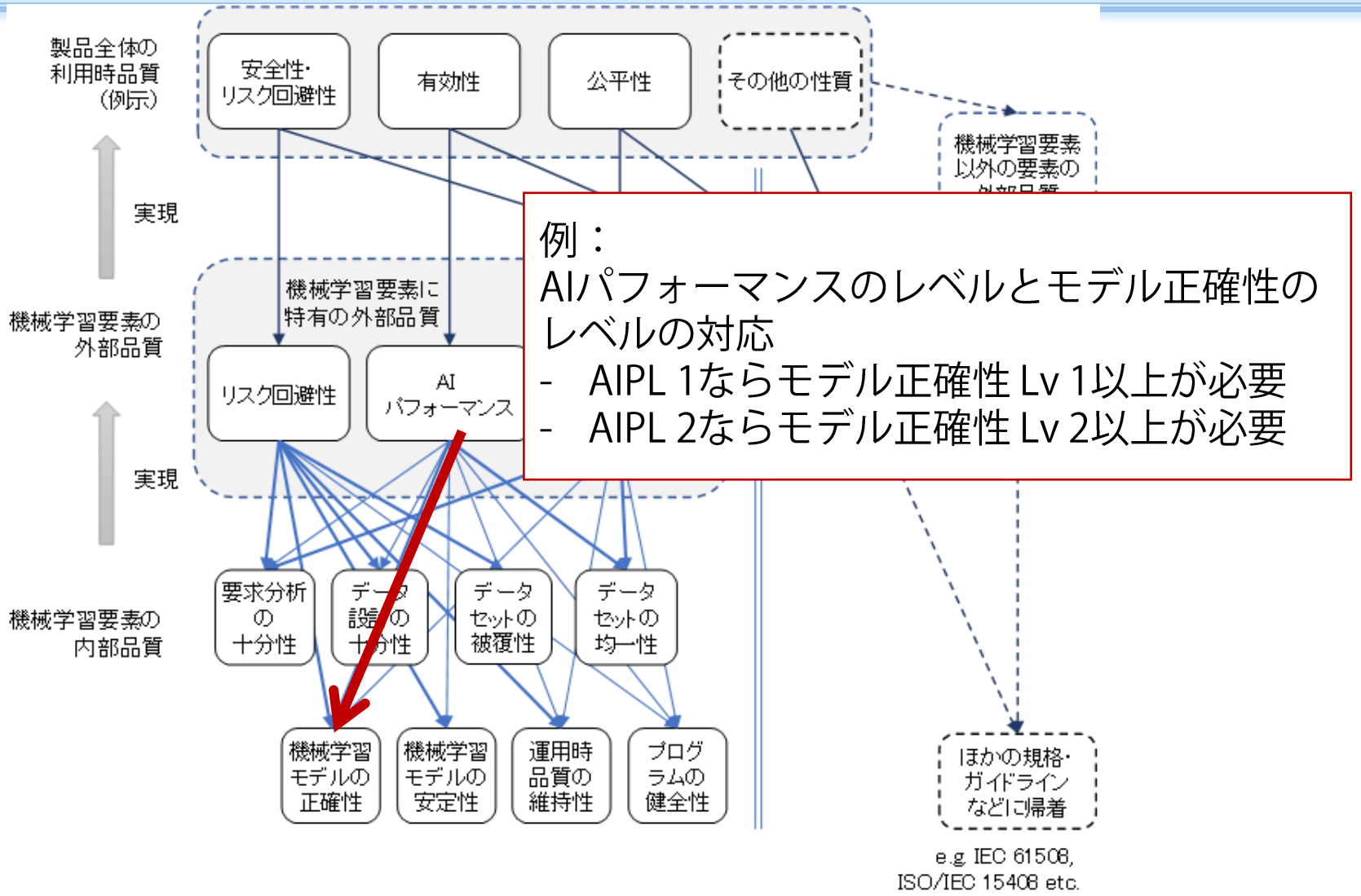
AIQMガイドライン（2020～）

- 標準化を見据えて外部品質特性・内部品質特性を定義
  - 前述の品質モデルの形式
  - 機械学習固有の点を
    外部品質レベルでは
    3特性に集約



[https://www.cpsec.aist.go.jp/achievements/aiqm/]

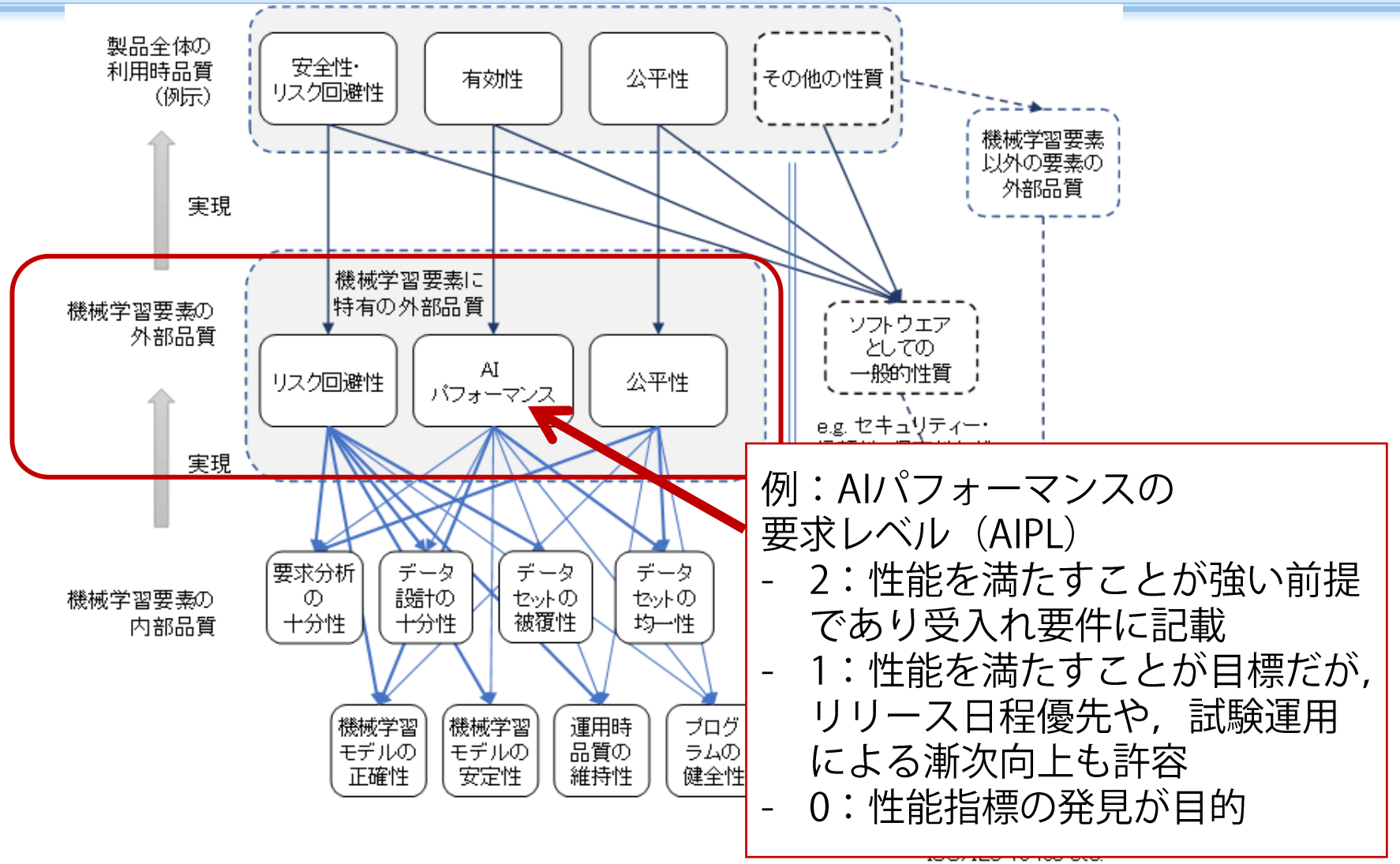
AIQMガイドライン：内部品質レベル分け



例：
AIパフォーマンスのレベルとモデル正確性のレベルの対応
- AIPL 1ならモデル正確性 Lv 1以上が必要
- AIPL 2ならモデル正確性 Lv 2以上が必要

[https://www.aist.go.jp/aist_j/press_release/pr2020/pr20200630_2/pr20200630_2.html]

AIQMガイドライン：外部品質レベル分け

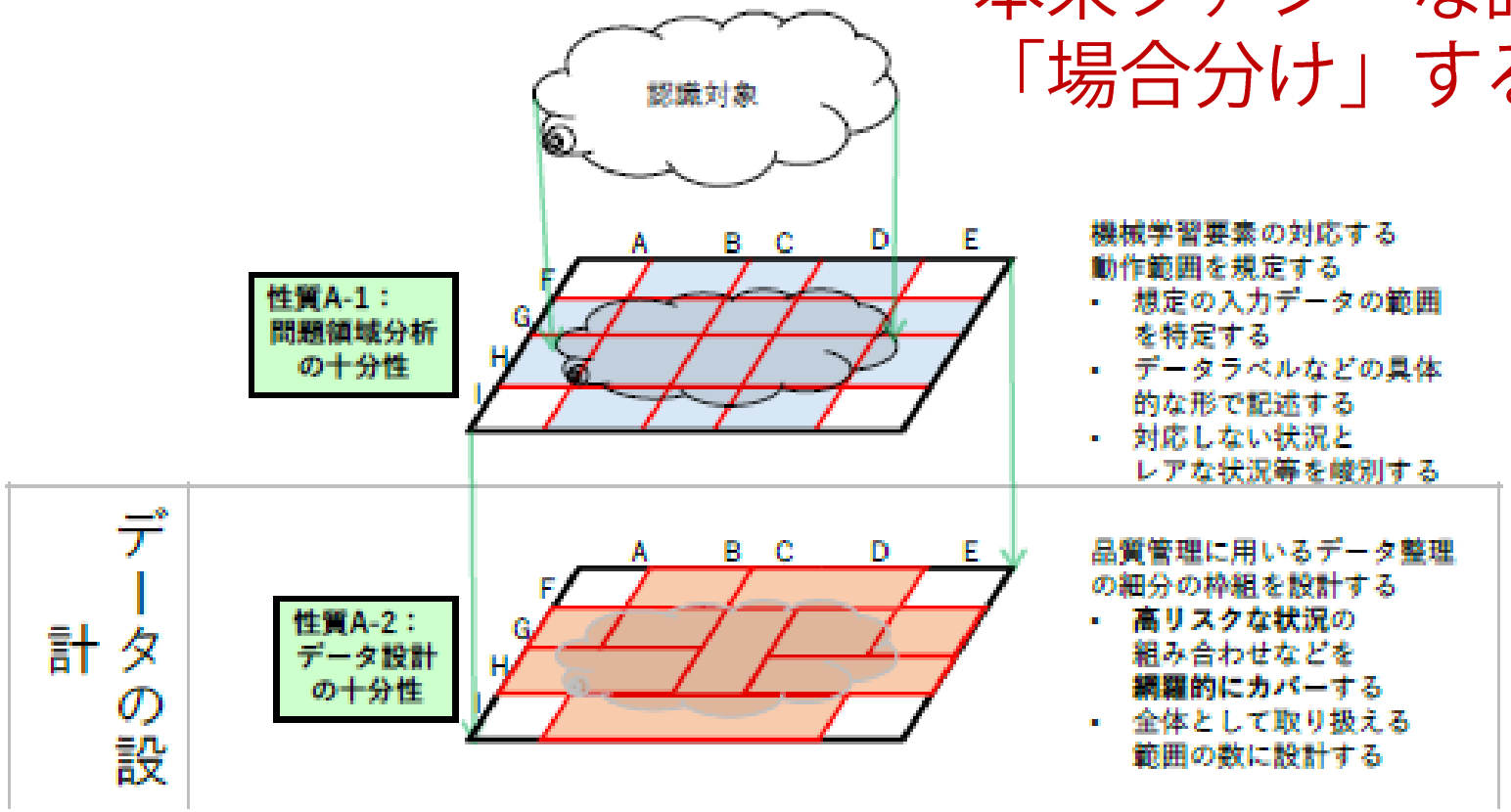


例：AIパフォーマンスの
要求レベル（AIPL）
- 2：性能を満たすことが強い前提であり受入れ要件に記載
- 1：性能を満たすことが目標だが，リリース日程優先や，試験運用による漸次向上も許容
- 0：性能指標の発見が目的

[https://www.aist.go.jp/aist_j/press_release/pr2020/pr20200630_2/pr20200630_2.html]

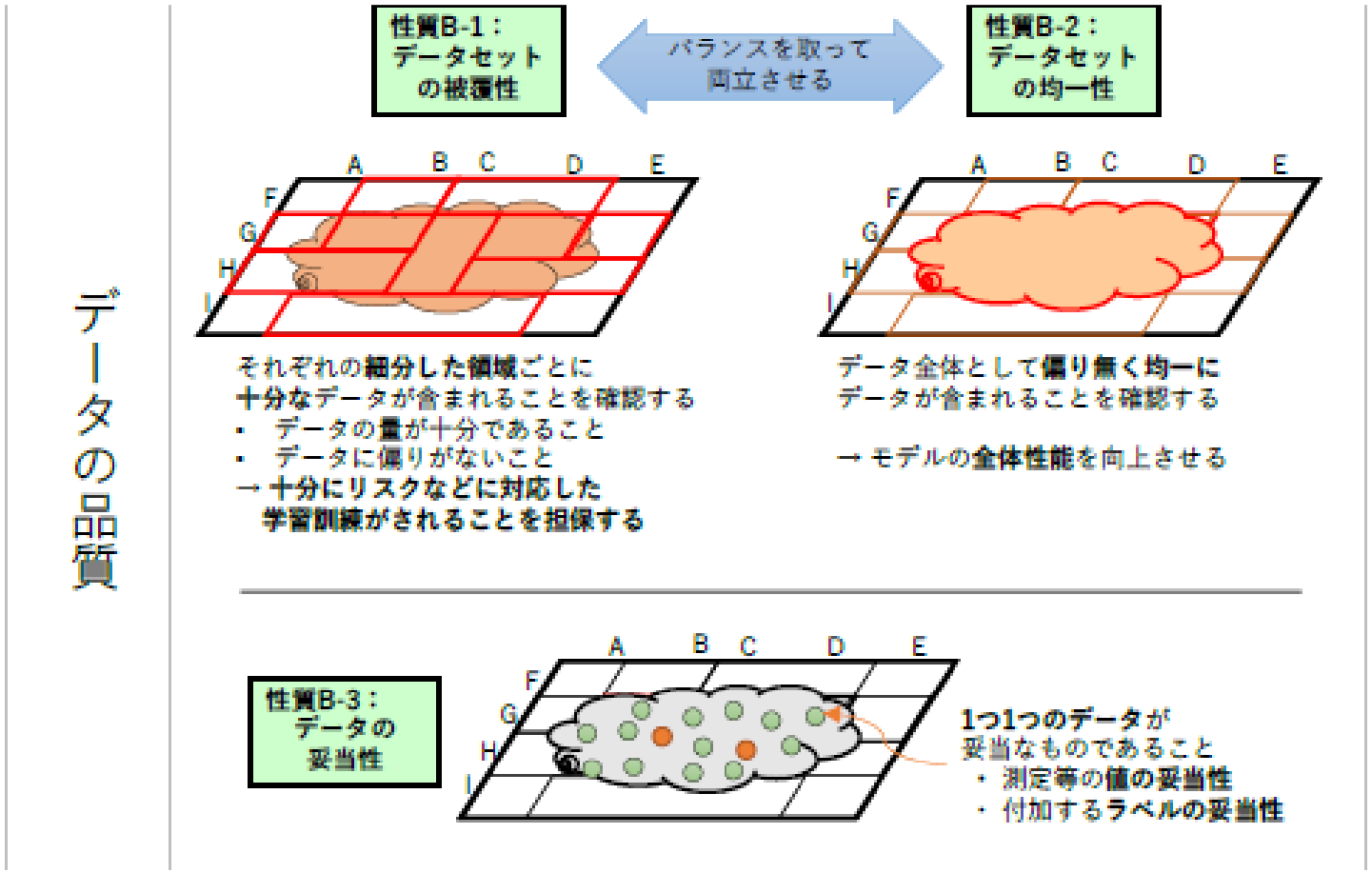
【付録】 AIQMガイドライン：内部品質 (1)

本来ファジーな認識対象を
「場合分け」することで評価軸を検討



[https://www.cpsec.aist.go.jp/achievements/aiqm/]

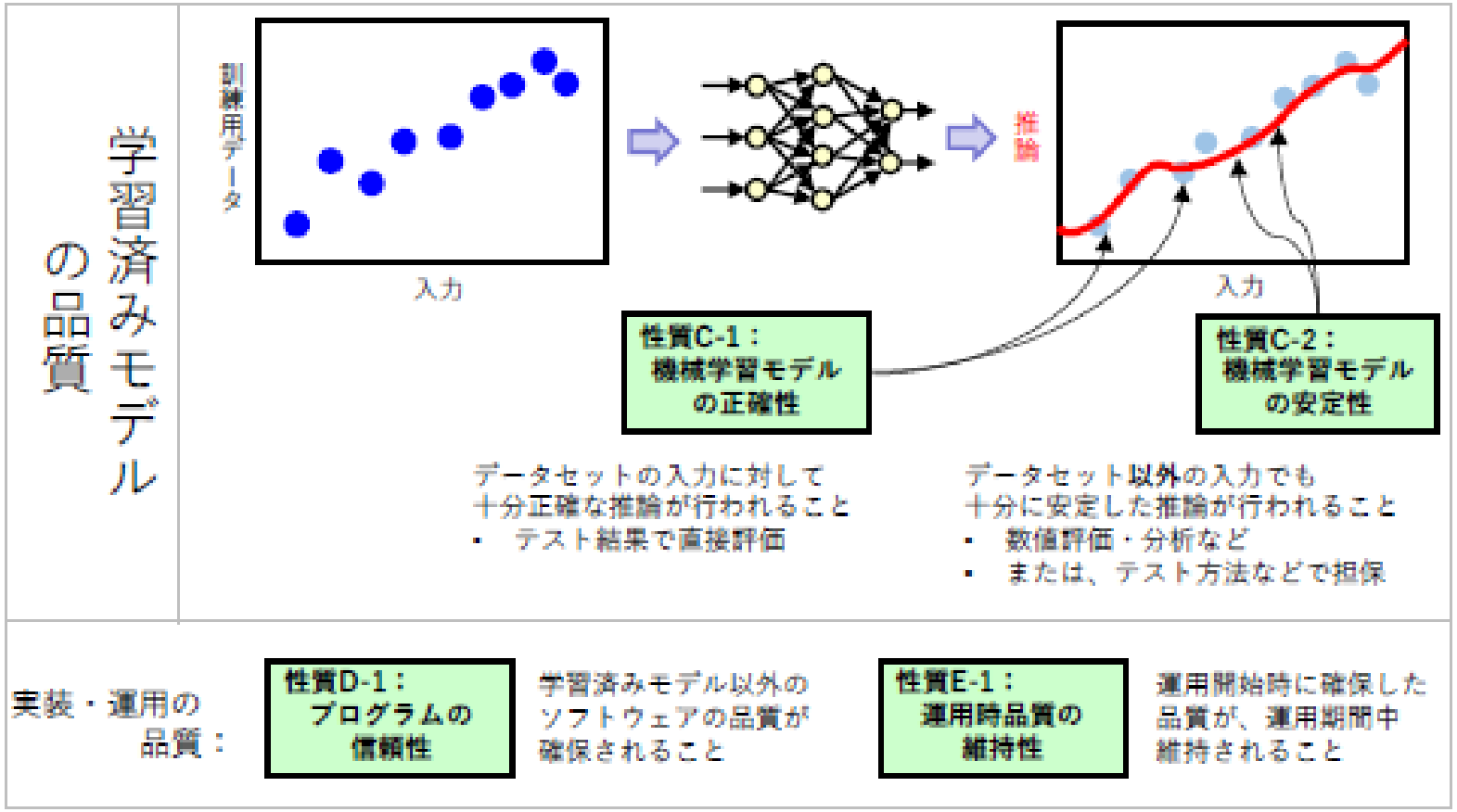
【付録】 AIQMガイドライン：内部品質 (2)



レアケースも充実
vs. 実際の分布に忠実

[https://www.cpsec.aist.go.jp/achievements/aiqm/]

【付録】 AIQMガイドライン：内部品質 (3)



安定性は現状技術での保証は難しいが
今後に期待

[https://www.cpsec.aist.go.jp/achievements/aiqm/]

QA4AIガイドライン（2019～）

- 品質を考えるべき軸の定義

成果物（訓練・評価のためのデータ，予測モデル，システム全体）

＋ プロセス（試行錯誤し改善を継続反復できるagility）

⟷ 顧客の期待とのバランス

- 5ドメインでの踏み込んだ分析
  - 自動運転，産業プロセス，スマートスピーカー，画像や動画の生成，AI-OCR

[http://www.qa4ai.jp/]

QA4AIガイドライン：品質の軸 (1)

- Data Integrity： データに関する品質
  - 量，ラベル妥当性，データドメインの妥当性（期待する運用領域と合致するかなど），外れ値や欠損値の扱い，学習用データとテスト用データの独立性，・・・
- Model Robustness： モデルに関する品質
  - 性能（正解率等），汎化性能，アルゴリズムやハイパーパラメータの妥当性，ノイズに対する頑健性，
    テストに用いたデータの妥当性，・・・

QA4AIガイドライン：品質の軸 (2)

- **System Quality**： システム全体の品質
  - システムが提供する価値の高さ，発生しうる事故のリスク（致命度および発生しやすさ），機械学習部品への依存度，説明可能性・納得性，・・・

- **Process Agility**： プロセスの迅速さ
  - 反復型開発やモデル・システムの品質向上の周期の短さ，運用状況の継続的なフィードバックの頻繁さ，品質向上が期待できる程度，開発・探索・検証・リリースなどの自動化度合い，開発チーム内外において専門家の貢献が期待できる度合い，・・・

QA4AIガイドライン：品質の軸 (3)

- Customer Expectation： 顧客による期待の高さ
  - 期待の高さ，確率的振る舞いの受容度合い，
    リスクの理解度合い，
    継続的な探索・向上に関する理解度合い，
    法の遵守や社会的受容の必要度合い，
    合理的説明の必要度合い，
    想定外への対応や外挿を求める度合い，・・・

Ethics Guidelines for Trustworthy AI (EU, 2019)

1. Human agency and oversight
    1. Fundamental rights
    2. Human agency
    3. Human oversight
2. Technical robustness and safety
    1. Resilience to attack and security
    2. Fallback plan and general safety
    3. Accuracy and reproducibility
3. Privacy and data governance
    1. Respect for privacy and data protection
    2. Quality and integrity of data
    3. Access to data
4. Transparency
    1. Traceability
    2. Explainability
    3. Communication
5. Diversity, non-discrimination and fairness
    1. Unfair bias avoidance
    2. Accessibility and universal design
    3. Stakeholder participation
6. Societal and environmental well-being
    1. Sustainable and environmentally friendly AI
    2. Social impact
    3. Society and democracy
7. Accountability
    1. Auditability
    2. Minimising and reporting negative impact
    3. Documenting trade-offs
    4. Ability to redress

[https://ec.europa.eu/digital-single-market/en/news/ethics-guidelines-trustworthy-ai]

Excellence and Trust in Artificial Intelligence (EU, 2021)

- 高リスクなAIシステムの意識
  - 交通などのインフラ，教育，雇用，法律，入国・移民
- 市場に出す前の責務
  - 適切なリスクの見積もりと低減
  - 高品質なデータセット
  - トレーサビリティのためのロギング
  - 詳細な文書化
  - ユーザへの明確かつ適切な説明
  - 高い頑健性・セキュリティ・正確さ

[https://ec.europa.eu/commission/presscorner/detail/en/ip_21_1682]

まとめ

- AI品質・トラストへのアプローチ
  - 適用先・ステークホルダーからの要請の強さ・ポイントの明確化
    - 例：高リスクな失敗があるのか？平均性能がよければよいのか？
    - 例：倫理・コンプライアンス上の要請があるのか？
  - 要請の強さ・ポイントに対応した成果物の評価
    （AIの場合，データ・モデル・システム全体の評価）
    - 例：失敗が致命的な領域の場合，重要なケースに対応する
      データが十分にあるか？そのケースでの予測性能が十分に高いか？
    - 例：人間がAIの失敗を把握でき，補えるか？
  - 評価はできるが，不完全・不確かという前提で進めることが重要