AI搭載ソフトウェアの開発プロジェクトにおける品質確保

2019.9 SQiP2019

株式会社 日立製作所
システム＆サービスビジネス統轄本部
品質保証統括本部

〇今谷恵理 eri.imatani.no@hitachi.com
中川純貴 junki.nakagawa.gs@hitachi.com

自己紹介、前提

HITACHI
Inspire the Next

・ソフト開発ベンダー、SIerのQAの立場からの経験を発表します。

・QAとしての経歴は10年。うちAI関係は直近２年。
＋学生時代にMLを専攻していました。

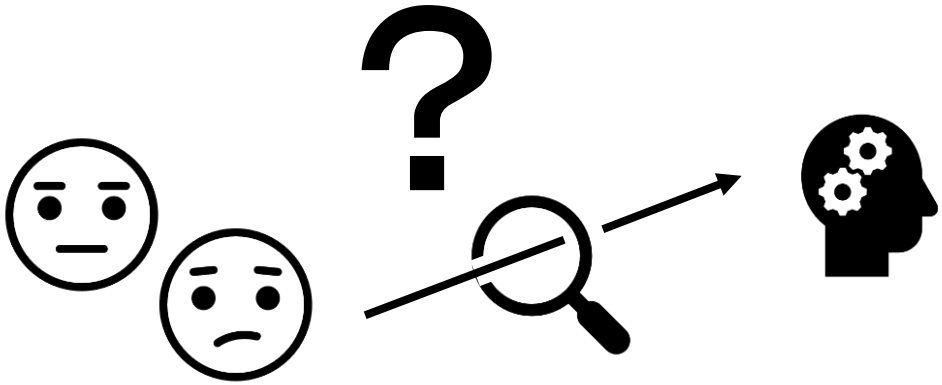
・前提：教師あり学習　※教師なし学習、強化学習は対象外です。

・用語の省略
CL：チェックリストのこと
ML：機械学習のこと
精度：汎化性能のこと

HITACHI
Inspire the Next

Contents

AIの検査がなぜ難しいのか



- **データドリブン**

  入力に対する出力は訓練データ次第。出力が導出された理由を演繹的に説明できない。だから**検査CLを作りにくい**。

- **確率的**

  出力に誤りがあっても、低頻度ならOK。**バグか判断しにくい。**

  ※バグ＝対策が必要な不良

- **検査技法**

  どの技法も限定的か提唱段階。**セオリーと言える検査技法はまだ無い**。

じゃあどうする

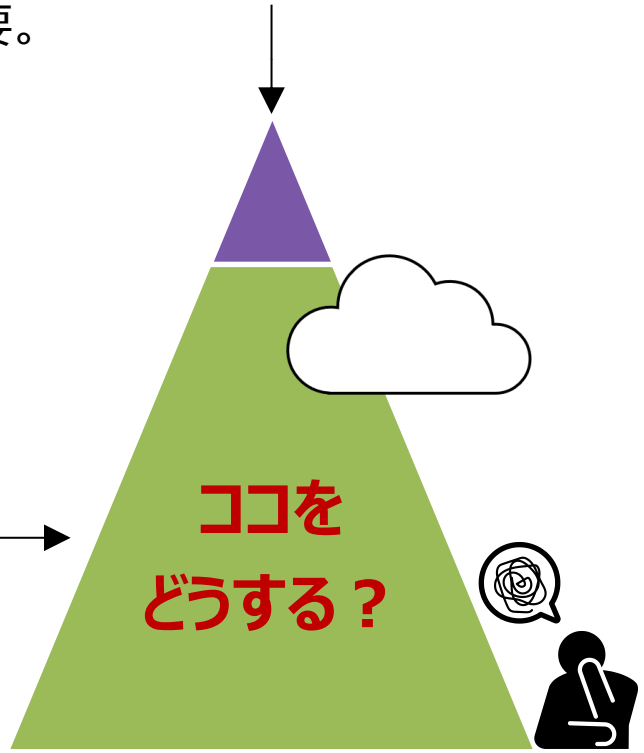
AI搭載ソフトも様々。できるところからやろう。

- 高い汎化性能、安全性、ロバスト性を強く求められるソフト

  e.g. 自動運転など。数は限られる。厳しいテストが必要。

- 多少粗くてもOKなソフト

  人の手間を軽減してくれればいいソフト。
  厳しいテストは不要(顧客要件と予算次第)、
  だが最低限レベルは担保したい！



考え方

品質＝テストデータでの汎化性能（精度）　と覚悟を決める。

例：　ナスの出荷判定をする画像識別AIに
　　パンダやサルの画像が入力される心配はしなくていい！

汎化性能の実現方法に問題が無いか、
プロセスをチェックすることで品質を間接的に確認する。

- 前提条件の整理
- データの選定／管理
- 精度の測定方法 等

ハイパーパラメータ？　訓練データ？

過学習？　交差検証？　オンライン学習？

汎化性能の劣化？再学習？前提条件？

え、なに？

案件増加で、
AI分野に新しく挑戦する
開発者／SE／QAも多い。
→ プロセスチェックリストが必要！

プロセスチェックの位置付けを
別の角度から表すと…（次ページ）

AI搭載ソフトの中身とスコープ

AI搭載ソフトの中身とスコープ

HITACHI
Inspire the Next

広義のAIはSE作業の塊。作業プロセスを検査する！

精度測定の方法および、それに至るまでの学習過程が妥当かどうかを、
プロセスや訓練データから判断する。

AI搭載ソフ

広義のAI

「学習」の
試行錯誤（作業）
と、その結果の実装

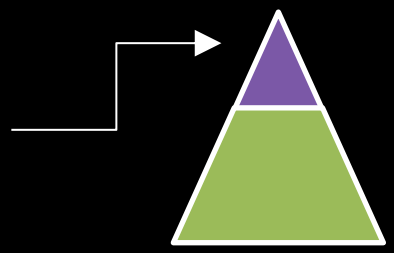
- 特徴量の選定／加工
  - ⑩クレンジング／補完
  - ⑩補正／マージ／次元圧縮
- ハイパーパラメタ調整
- 精度測定

AI

狭義のAI

- 内部パラメタ（モデル）

モデルの振る舞いを直接的にテストする必要が
あるのはコッチ（自動運転等）。今回は対象外



HITACHI
Inspire the Next

ＭＬプロセスチェックリストを作りました。

ＭＬプロセスチェックリストを作りました

開発工程で失敗しそうなポイントをチェックリスト（以下CLと呼ぶ）にまとめました

インプット

- 教師あり学習において一般的に気を付けるべきこと（文献、記事、シンポジウム等から）
- 社内の失敗／ヒヤリ事例
- 内閣府から提示されている「人間中心のＡＩ社会原則検討会議」の動向
  - フェールセーフの考慮
  - AIが人種・性別で差別をしないこと

ベース

- ウォータフォール型のWBS
- AI搭載システムの開発プロセスフロー（CRISP-DMの日立版）

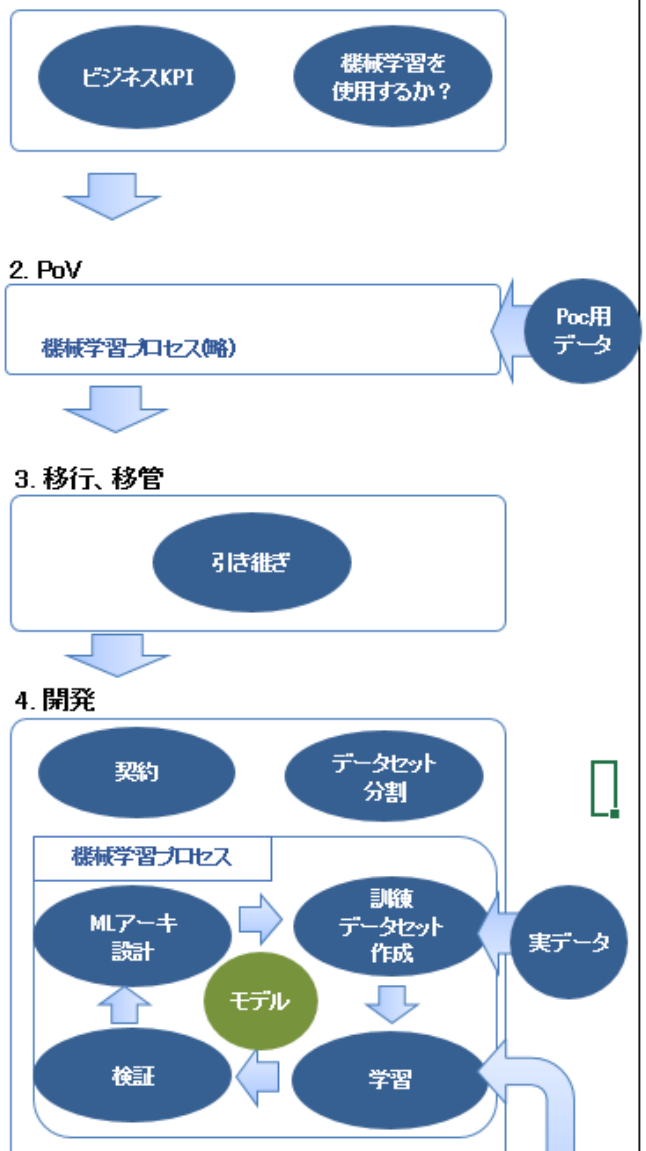
人材

- QAメンバにて作成
- 社内有識者にレビュー依頼

MLプロセスCL

HITACHI
Inspire the Next

AIシステムの開発プロセス



MLプロセスCL

1. 企画

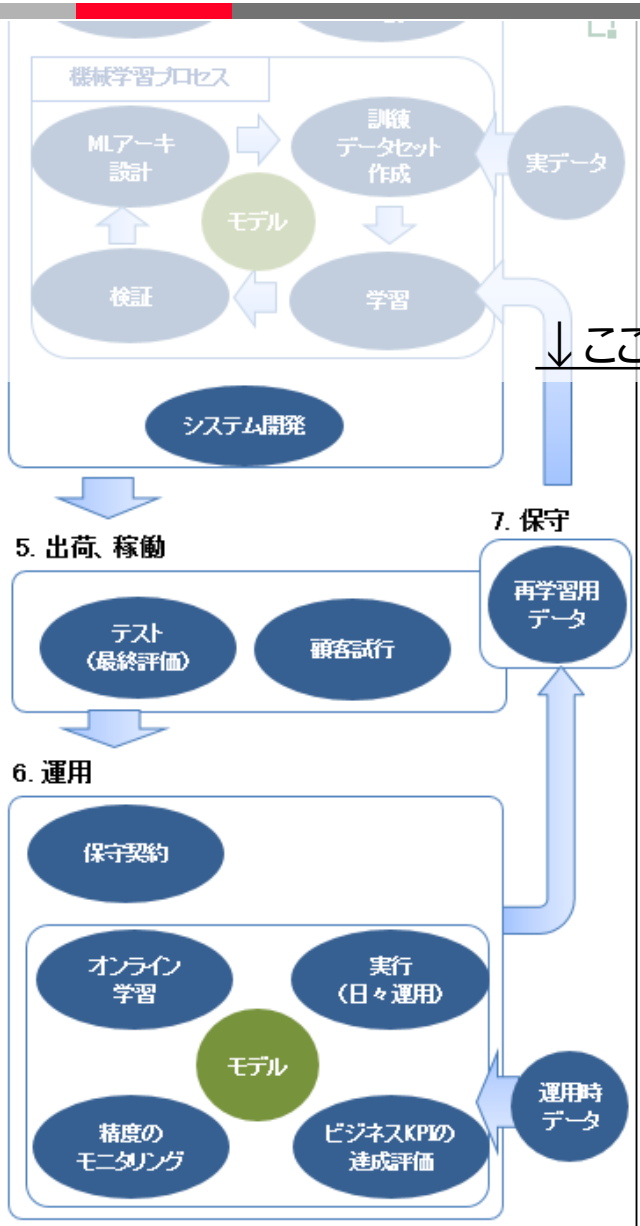
			チェック欄 日付
1-1	ビジネスKPI	顧客が抱える問題がKPIに落とし込まれており、それが明確であること。	
1-2			
1-3	機械学習を使用するか？	機械学習を本当に使うべきなのかが検討されていること。	
1-4		機械学習で解ける見込みのある問題であること。	

2&4. PoV /開発

				PoV	Development
2-1	契約	契約など	契約書やサービス仕様書のしかるべきレビューが実施されていること。	■該当	■
2-2	データセット分割	データセット分割	最後に汎化性能を測定する為に、データセットからあらかじめテストデータセット(最終評価用のデータセット)を分離確保していること。	■	■
2-3	アーキ設計	アルゴリズムの選定	データの特性を把握した上でアルゴリズムを選定していること。	■	■
2-4		システム設計	予測結果をどういう形で利用するのか、予測誤りをどこで吸収するのかを検討・設計されていること。(顧客とともに検討すること)		■
2-5			モデルのバックアップ方法も検討・設計されていること。		■
2-6	訓練データセット作成	訓練データ・正解ラベル	機械学習の入力データが、予測に必要な特徴量を含んでいるか、業務有識者を交えて検討していること。		■
2-7			訓練データに含まれる偏見などに起因して、AIの出力による不当な差別が生じないように配慮していること。		■
2-8			質の高い訓練データと正解ラベルを準備していること。	▲少し該当	■
2-9		データの前処理	精度の実現のため、訓練データに対して適切な処理をしていること ※AIアルゴリズムによって必要な処理は異なる。	▲	■
2-10			データ前処理の内容を顧客に説明できる形で記録していること。 (保守の為に、設計資料としても残すこと)	■	■
2-11	学習	学習・パラメータチューニング	パラメータチューニングの手順・方式を事前検討し、学習過程の記録を残すこと。	▲	■
2-12			過学習を防ぐための施策を検討すること。 ※さらに多くの訓練データを集めるなど。詳細(解説)のシート参照。	▲	■
2-13	検証	検証	モデル構築後、検証データセットを利用して精度(汎化性能)を評価していること。 精度が悪い場合はデータの前処理方法やハイパーパラメータの見直しを行っていること。(そのようなプロセスを踏んでいること)	▲	■
2-14			過学習が発生していないか検証すること。		
2-15			データの状況や学習の状況を踏まえたうえで、到達すべき精度について顧客との同意形成を行っていること。 (実現の見込みがある精度で同意形成していること)		■
2-16	システム開発	AI以外の要素	AI以外のシステムと同様に、エラー処理や信頼性、性能、セキュリティ、保守性などの設計と実装も行うこと。		■

MLプロセスCL（つづき）

HITACHI
Inspire the Next



↓ここまでは前ページと同じ（つづき）

				日付
2-11	学習	学習・パラメータチューニング	パラメータチューニングの手順・方式を事前検討し、学習過程の記録を残すこと。	▲ ■
2-12			過学習を防ぐための施策を検討すること。 ※さらに多くの訓練データを集めるなど。詳細（解説）のシート参照。	▲ ■
2-13	検証	検証	モデル構築後、検証データセットを利用して精度（汎化性能）を評価していること。 精度が悪い場合はデータの前処理方法やハイパーパラメータの見直しを行っていること。（そのようなプロセスを踏んでいること）	▲ ■
2-14			過学習が発生していないか検証すること。	
2-15			データの状況や学習の状況を踏まえたうえで、到達すべき精度について顧客との同意形成を行っていること。（実現の見込みがある精度で同意形成していること）	■
2-16	システム開発	AI以外の要素	AI以外のシステム同様に、エラー処理や信頼性、性能、セキュリティ、保守性などの設計と実装を行うこと。	■

3. 移行、移管

				日付
3-1	引継ぎ	担当者が変わった時	POVで何をしたかが分かるドキュメントがあること。 （担当者が変わった場合には特に確認すること）	
3-2		プロジェクト状況の確認	POV時からビジネス環境が変わってないこと。	

5. 出荷、稼働

				日付
5-1	テスト 最終評価	プロセス確認	ここまでの 各工程で、上述のCLが消化され、懸案等が解消されていること。	
5-2		精度の確認	事前に確保したテストデータセットを用いて精度（汎化性能）を最終確認していること。 関連項番：2-2	
5-3	顧客試行	試行	必要に応じて、顧客内で試行運用する期間を設けていること。（その検討がされていること）	

6. 運用

				日付
6-1	ビジネスKPIの達成評価	ビジネスKPIの達成評価	企画段階で設定したビジネスKPIが稼働時点で達成できているか確かめ、開発したシステムの価値を評価していること。	
6-2	精度のモニタリング	モニタリング	ビジネス環境の変化などによるデータ傾向の変動により、精度は悪化していくことを織り込んだ運用になっていること。	
6-4	オンライン学習	モデルのバックアップ	悪い学習をしてしまった際にモデルを初期化をする方法を準備していること。	

7. 保守

				日付
7-1	再学習用データ	ー	再学習で必要になるデータセットを、あらかじめ蓄積しておく仕組みがあること。	

チェック項目例 1

HITACHI
Inspire the Next

訓練データの与え方が適切か？

- **十分な量／種類／質のデータや正解ラベルを使っているか？**
  機械学習で解決できる問題／状況か？
  無責任な／いい加減なラベリングをしていないか？

- **差別を招くようなデータを使っていないか？**
  性別、人種、住所、思想　がデータにある時は要注意

- **特徴量の加工法が検討・実装されているか？**
  特徴量選択、次元圧縮、クレンジング、補正、正規化、サンプリング　は適切か？

MLプロセスCLは、全ての基礎である訓練データに強くフォーカス。

チェック項目例２

HITACHI
Inspire the Next

汎化性能の測定方法が適切か

- **データ、ハイパーパラメータ、モデルの管理**

  汎化性能測定用のテストデータに、訓練データが混ざっていないか？

  チャンピオンデータで精度測定していないか？

  データ、モデル、パラメタのセットが構成管理されているか？（再現性）

- **過学習**　　過学習していないか確認したか？

- **評価軸の検討／合意**

  例：偽陽性、偽陰性（適合率、再現率）

  A　ガンの可能性が少しでもあればアラートする

  B　確実にガンと判断できた場合だけアラートする

  →どっちを軸に精度（汎化性能）を評価するか、顧客と合意したか？

MLプロセスCLは、精度測定にもフォーカス。

HITACHI
Inspire the Next

開発フェーズ以外

企画・契約

- AI導入の成果をどう評価するか、KPIに落とし込まれているか？
- 顧客受入における精度測定の前提に、Sier⇔顧客で認識差がないか？
  - 受け入れテストで使うデータセットの中身は、訓練データと同質か？

運用

- AIの出力が誤っていても、顧客に大きな被害が出ない運用か？（フェールセーフ）
- 稼働後の精度劣化（データの傾向変化）を監視する運用になっているか？
  - 精度劣化の可能性を顧客に伝えているか？　劣化が障害扱いされないか？

保守

- 再学習に必要なデータを確保する運用／仕組みになっているか？
- オンライン学習の場合、悪い学習をしてしまった際に元に戻せるか？

MLプロセスCLは、企画～開発～保守をトータルでカバー。

HITACHI
Inspire the Next

使ってみた。

適用プロジェクトその１

PJ-需要予測

- 需要発生予測システム（XGBoost ）　※XGBoost ＝ 勾配ブースティング決定木の一種
- 受注案件（顧客がいるプロジェクト）

過去データから、各地域での
注文発生量を予測・シミュレーション

供給計画の参考にする

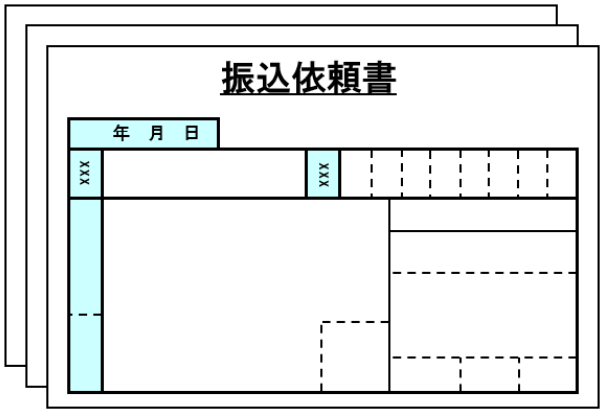
地図は国土地理院の電子国土Webから掲載

適用プロジェクトその２

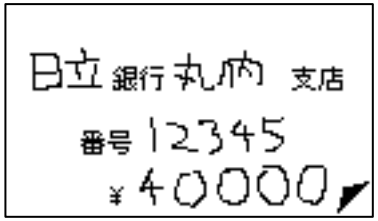
HITACHI
Inspire the Next

PJ-帳票認識

- 帳票認識ソフトウェア（DNN）　※DNN ＝　ディープニューラルネットワーク
- 製品開発
- プレ学習した状態で提供（販売後、顧客データで追加学習）



手書き帳票



手書き文字

銀行名：　日立銀行
支店名：　丸ノ内支店
口座番号：12345
振込金額：40,000円

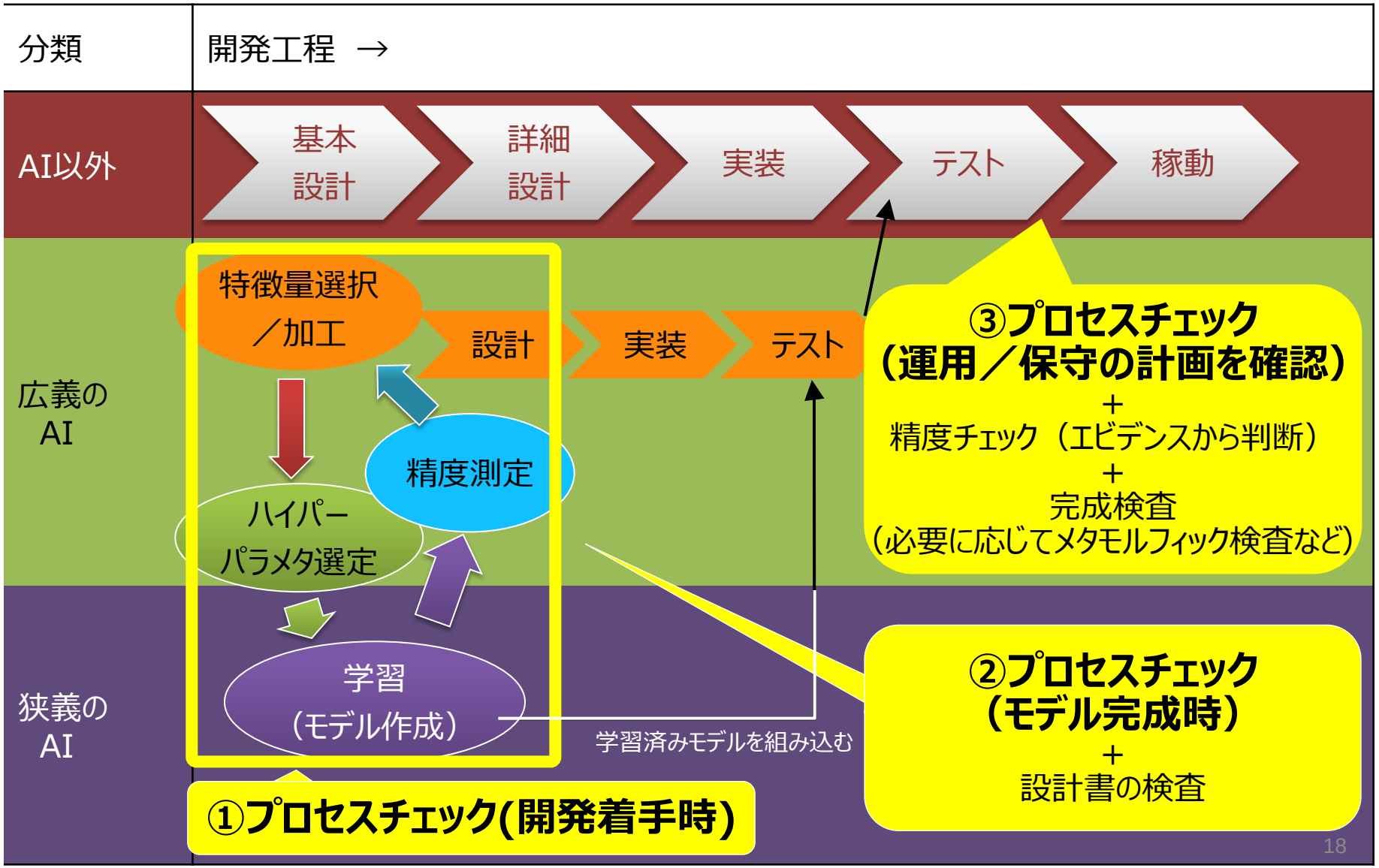
OCRでデータ化

【OCRの課題】
癖字／崩し字　が苦手

↓

ディープラーニングと組み合わせて認識精度を向上

MLプロセスCLの適用タイミング



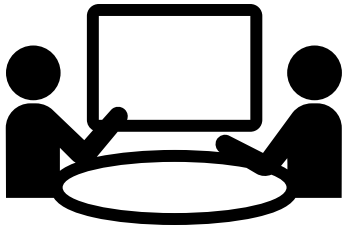
やりかた

HITACHI
Inspire the Next

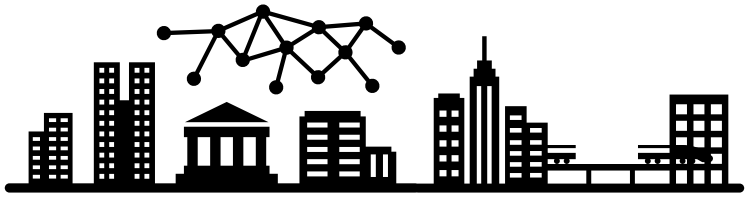
- レビューの形でヒアリング、CLを一緒に埋めていく

  参加者

  開発者（AI部分の実装・学習の担当者）
  SE（プロジェクトリーダ）
  QA



- 一部、エビデンスを提示してもらう



適用結果＜PJ-需要予測＞

HITACHI
Inspire the Next

以下の問題が摘出された

開発

- 訓練データの作成方法や学習手順がドキュメントとして整備されていない
  （不在やメンテナンス不足）

【MLプロセスCLで回避した事象／リスク】
Revできない、開発した担当者以外には再現できない。
保守の再学習時に「前はなぜこうしたんだっけ？」を振り返れない。

運用保守

- 稼働後の精度をどう確保していくか、顧客との合意が足りていなかった
  → 半年ごとに精度測定し、規定値を下回ったら再学習する方針に

【MLプロセスCLで回避した事象／リスク】
保守費を頂いているのに精度劣化に気付かない。
劣化が酷くなった時にお客様から叱られる事態に。

適用結果＜PJ-帳票認識＞

以下の問題が摘出された

開発

- 過学習を実際に回避したか確認する手段を検討できていなかった

運用保守

- 顧客受け入れテストの前提データセットが曖昧

  → （販売後の、顧客データでの追加学習サービスにおいて）受入テストとしての精度計測に使用するデータセットを、**事前に顧客と取り決める手順とした。**

【MLプロセスCLで回避した事象／リスク】

ゐ・ゑ・ヰ・ヱ

顧客 「とっておきの崩し字で精度を測定するぞ。
毛筆の草書体で書いた歴史的な仮名文字だ！」

SIer 「硬筆、楷書、現代仮名が暗黙の前提だったはずでは・・・？」

→ 工期遅延、コスト追加のリスク

- 再学習のサービスメニューが無い

総括

- 漏れがちな検討項目を救う効果があった

  運用・保守など、開発中にあまり意識しない部分の問題を抽出できた

- 教師あり学習としての基礎的なインシデントは発見されなかった

  基礎的なインシデントの例;　- テストデータが訓練データを混同
  - 過学習の考慮が全くない

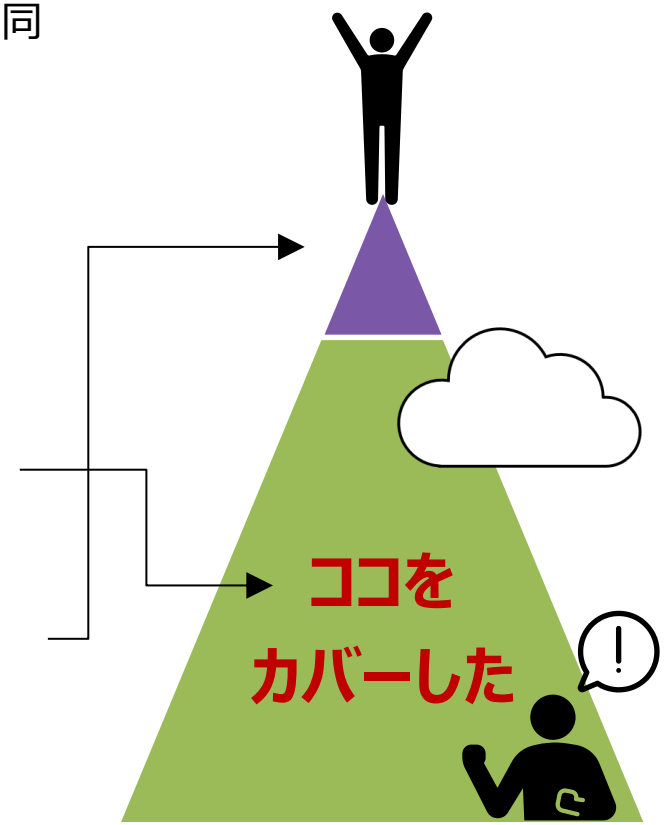
  考察;　開発案件が少ない今はまだML人材が足りている。
  今後人材が足りなくなると、基礎的なインシデントが発見されることも増える。

- 多少粗くてもOKなソフト

  必要最低限な品質はMLプロセスCLだけで担保できた！

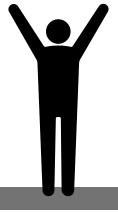
- 高い精度、厳しいテストが求められるソフト

  基礎としてMLプロセスCL　は使える！
  ※AIモデルに対する直接的なテスト も必要



HITACHI
Inspire the Next

END



AI搭載ソフトウェアの
開発プロジェクトにおける品質確保

2019.9　SQiP2019

株式会社 日立製作所
システム＆サービスビジネス統轄本部
品質保証統括本部

今谷恵理、中川純貴


HITACHI
Inspire the Next