

# AI製品の信頼性をどうやって評価する？

## AIのすごさと製品としての難しさ

---

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：石川 冬樹

アーキテクチャ科学研究系 准教授

先端ソフトウェア工学・国際研究センター 副センター長

## ■専門

- ソフトウェア工学，特にディペンダビリティ：  
形式手法，自動テスト生成，安全性論証など

## ■現在の主な研究プロジェクト

- JST ERATO-MMSD：自動運転システムの安全性
- JST MIRAI-eAI：機械学習システムのディペンダビリティ

## ■産業界向け教育・実践研究

- トップエスイー，日科技連SQiP，電通大AISECなど
- 機械学習工学コミュニティ（MLSE研究会，QA4AI）



機械学習工学研究会  
MACHINE LEARNING SYSTEMS ENGINEERING



AIプロダクト  
品質保証  
コンソーシアム

# 目次

---

- AIのすごさと難しさ
  - 機械学習技術によるAI
  - 象徴的な事例
- 製品としての「AI」
  - 従来ソフトウェアにおける「注文」と「品質」
  - AIソフトウェアにおける「注文」と「品質」
- まとめ・今後に向けて

# 目次

---

## ■ AIのすごさと難しさ

### ■ 機械学習技術によるAI

### ■ 象徴的な事例

## ■ 製品としての「AI」

### ■ 従来ソフトウェアにおける「注文」と「品質」

### ■ AIソフトウェアにおける「注文」と「品質」

## ■ まとめ・今後に向けて

# 機械学習：とても簡単に

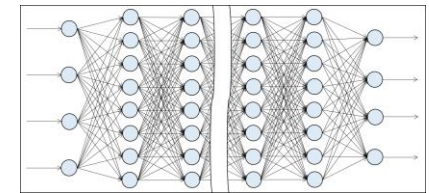
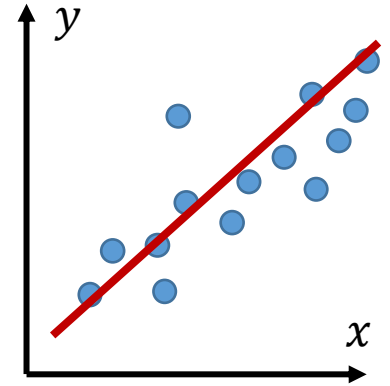
## ■ある会社での年齢 $x$ のときの給与 $y$ を予測したい

- $y = ax + b$  と表現できるとして、  
過去のデータと「一番合う」ように

$a, b$  を決めれば判定・予測プログラムが作れる！

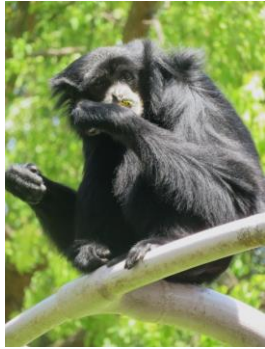
## ■実際は1次関数（パラメータ2つ）では表現が困難

- ディープラーニング（深層学習）では、より多様な入力形式を扱え、  
何百万以上のパラメータで予測関数を表現する



現在話題になっているAI（人工知能）は主に、  
判定や予測の機能をデータからつくる機械学習型

# 機械学習：別の例



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 | 180 77    |
| 46 87 121 | 70 76 122 | 122 14 94 |

画像 =  
各点の色を表す  
数値の集まり

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 |

この線引きを訓練データから作る

# 機械学習：すごいこと

## ■規則性を書き出すことが困難な機能であっても実現可能

### ■画像における物体の識別

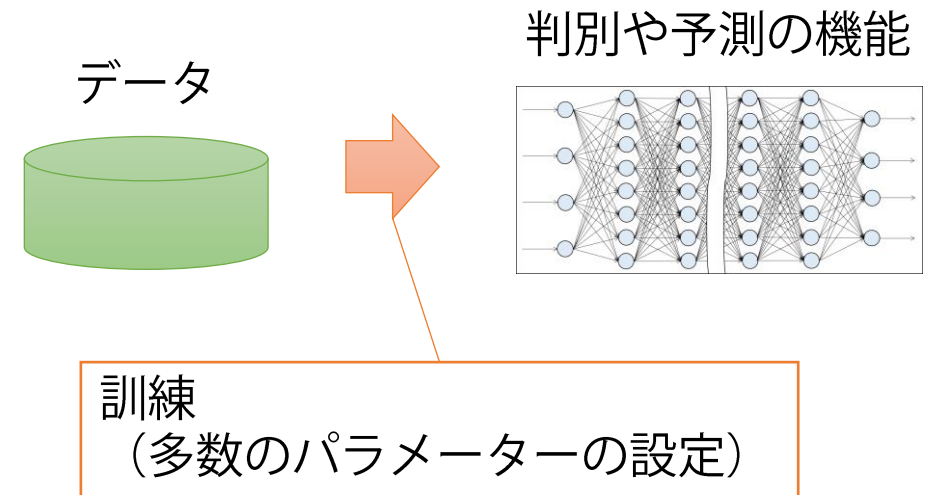
- 自動運転における歩行者や信号
- 工場における不良品や機器劣化
- 医療画像における腫瘍

### ■工場機器の制御

### ■ローンや保険に関する判断

### ■音声や文書の分類, 検索, 翻訳

### ■画像や動画の色塗りや補完, 生成



# 機械学習：大変なこと

---

## ■ 判別や予測の性能の不確かさ

- 原則として機能は不完全（100%正解することはない）
- どの程度の性能が出るか作ってみるまでわからない

## ■ 振る舞いの不確かさ

- 新たな入力でどう振る舞うかは未知（かなり似たデータでも）
- ある出力がなぜ起きたのかは説明できないことが多い

## ■ データへの依存性

- 大量かつ「適切な」訓練データが必要
- 基本的にデータに対して相対的な評価しかできない



# 目次

---

## ■ AIのすごさと難しさ

- 機械学習技術によるAI

- 象徴的な事例

## ■ 製品としての「AI」

- 従来ソフトウェアにおける「注文」と「品質」

- AIソフトウェアにおける「注文」と「品質」

## ■ まとめ・今後に向けて

# 課題の例：技術的限界・性能や振る舞いの不確かさ

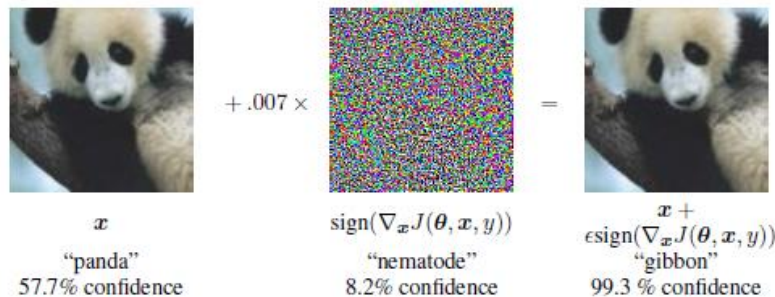
## ■ Google フォトの画像認識

- 黒人の写真を「ゴリラ」とタグ付け
- 2年経って本質的には直せず（ゴリラを禁止ワード扱いに）

[ <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> ]

## ■ 優れた画像識別器が少しのノイズで誤認識



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

[ Ackerman, Slight Street Sign Modifications Can Completely Fool Machine Learning Algorithms, IEEE Spectrum'17 ]

「パンダ」が「テナガザル」に

物理的なテープ貼付による誤認識

# 課題の例：訓練データ・攻撃・継続的学習

## ■ 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
- 継続的に学習・更新し続けるものの監視や制御の問題

(もしも攻撃がなかったとしても)

- 人間・社会の要請に事前に・事後にどう応えるか

[ <https://www.nytimes.com/2016/03/25/technology/microsoft-created-a-twitter-bot-to-learn-from-users-it-quickly-became-a-racist-jerk.html> ]  
(access: 2021/09/27)

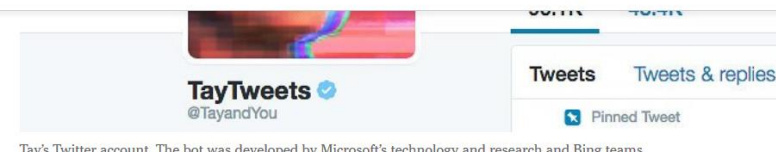
TECHNOLOGY

*Microsoft Created a Twitter Bot to Learn From Users. It Quickly Became a Racist Jerk.*

By DANIEL VICTOR MARCH 24, 2016



TECHNOLOGY | Microsoft Created a Twitter Bot to Learn From Users. It Quickly Became a Racist Jerk.



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

# 課題の例：バイアス・倫理的要求

- 社会的に望ましくない**差別的なバイアス（偏り）**の発生
  - 人間の（ときには無意識な）**差別を学習**
  - **少数例外**に対しては学習しきらず予測性能が低くなりがち



女性に不利な雇用判断

→

過去の差別を学習？女性が少数データ？

[ <https://jp.reuters.com/article/amazon-jobs-ai-analysis-idJPKCN1ML0DN> ]  
(access: 2021/09/27)



アフリカ系の名前で検索すると  
逮捕歴データベースの広告

→

ユーザの潜在意識によるクリックを反映？

[ L. Sweeney, Discrimination in  
Online Ad Delivery, ACM Queue'13 ]

# 目次

---

## ■ AIのすごさと難しさ

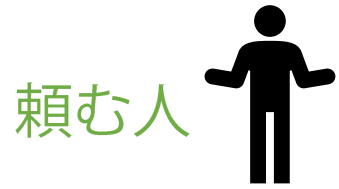
- 機械学習技術によるAI
- 象徴的な事例

## ■ 製品としての「AI」

- 従来ソフトウェアにおける「注文」と「品質」
- AIソフトウェアにおける「注文」と「品質」

## ■ まとめ・今後に向けて

# 従来ソフトウェアにおける「注文」



頼む人

飲食チェーン店の事業部門



- 商品の合計金額を計算
- 10%の消費税を加算（端数切捨）
- ...
- 日ごとの売上げ集計を翌週の月曜朝までに本社サーバに送付

要求仕様



作る人

同チェーン店のICT部門  
or  
ICTベンダー企業



食事客

使う人・影響を受ける人

本社戦略室

レジ担当店員

※ 実際は様々なバリエーションがあり、より複雑

# 従来ソフトウェアにおける「開発」



頼む人

飲食チェーン店の事業部門

- 商品の合計金額を計算
- 10%の消費税を加算（端数切捨）
- ...
- 日ごとの売上げ集計を翌週の月曜朝までに本社サーバに送付

要求仕様

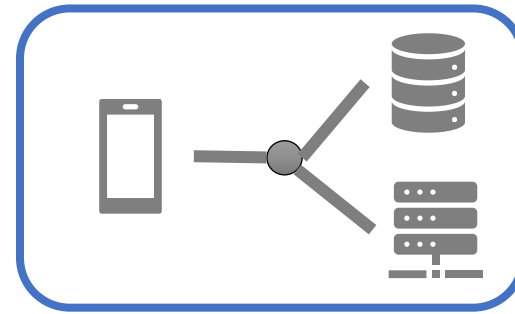


挙動をプログラムとして  
設計，書き出し



作る人

同チェーン店のICT部門  
or  
ICTベンダー企業



システム



食事客

使う人・影響を受ける人

本社戦略室

レジ担当店員

※ 実際は様々なバリエーションがあり，より複雑

# 従来ソフトウェアにおける「品質」 (1) 正しさ



頼む人



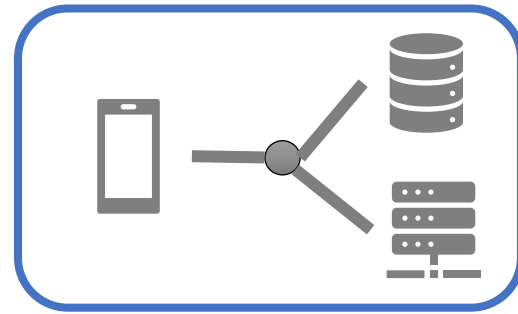
飲食チェーン店の事業部門

- 商品の合計金額を計算
- 10%の消費税を加算 (端数切捨)
- ...
- 日ごとの売上げ集計を翌週の月曜朝までに本社サーバに送付

要求仕様



正しく作った?



システム



作る人

同チェーン店のICT部門  
or  
ICTベンダー企業



食事客

使う人・影響を受ける人

本社戦略室

レジ担当店員

※ 実際は様々なバリエーションがあり, より複雑



# 従来ソフトウェアにおける「品質」 (1) 正しさ

- 正しさ：要求仕様で定めた規則・ルールを満たす
  - 基本的に確実に満たされるべき
  - 規則性を踏まえたテストを実施して確認
    - 「コーヒー270円」の場合で正しく料金を計算できたなら「紅茶220円」の場合もきっと正しく動く
- ➡ 複雑なソフトウェアで不具合ゼロは困難であるものの、様々な技術や工夫が積み重なってきた

# 従来ソフトウェアにおける「品質」 (2) 価値・妥当性



頼む人



飲食チェーン店の事業部門

- 商品の合計金額を計算
- 10%の消費税を加算 (端数切捨)
- ...
- 日ごとの売上げ集計を翌週の月曜朝までに本社サーバに送付

要求仕様



特定店舗ではテイクアウト  
実施中 (異なる税率)

売上げ集計の分析結果は  
翌週を待たず反映したい

XXXPay  
使える??

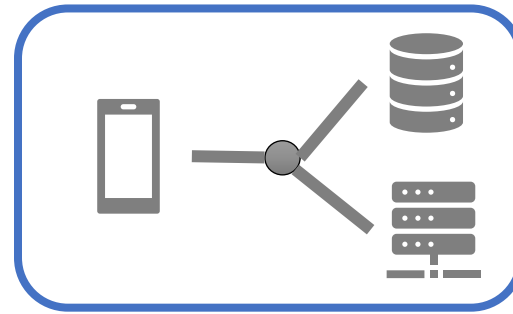


食事客

使う人・影響を受ける人

本社戦略室

レジ担当店員



システム



作る人



同チェーン店のICT部門  
or  
ICTベンダー企業

※ 実際は様々なバリエーションがあり, より複雑

# 従来ソフトウェアにおける「品質」 (2) 価値・妥当性



頼む人

飲食チェーン店の事業部門

- 商品の合計金額を計算
- 10%の消費税を加算 (端数切捨)
- ...
- 日ごとの売上げ集計を翌週の月曜朝までに本社サーバに送付

要求仕様



特定店舗ではテイクアウト  
実施中 (異なる税率)

売上げ集計の分析結果は  
翌週を待たず反映したい

XXXPay  
使える??



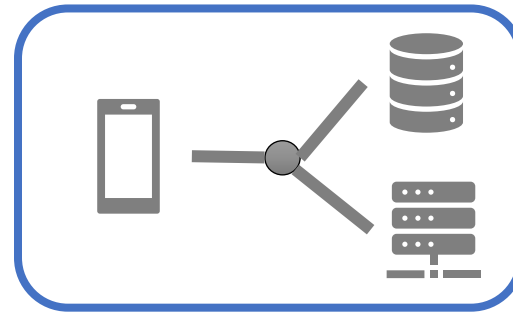
食事客

使う人・影響を受ける人

本社戦略室

レジ担当店員

よいものを作った?



システム



作る人

同チェーン店のICT部門  
or  
ICTベンダー企業

※ 実際は様々なバリエーションがあり, より複雑

# 従来ソフトウェアにおける「品質」 (2) 価値・妥当性

- 価値・妥当性：様々な利害関係者のニーズを満たす
  - 正解はない・終わりもない
  - 「自身が本当に必要としているもの」に予め気づき、漏れなく正確に言葉にすることは非常に難しい
  - 状況はどんどん変化する
- ➡ 「価値を産む小さな機能」の開発・評価を少しずつ反復する  
進め方が活発に (2010年代以降, アジャイル開発)
  - 対象ビジネスの専門家 (頼む人・使う人) が携わることが重要

# 目次

---

## ■ AIのすごさと難しさ

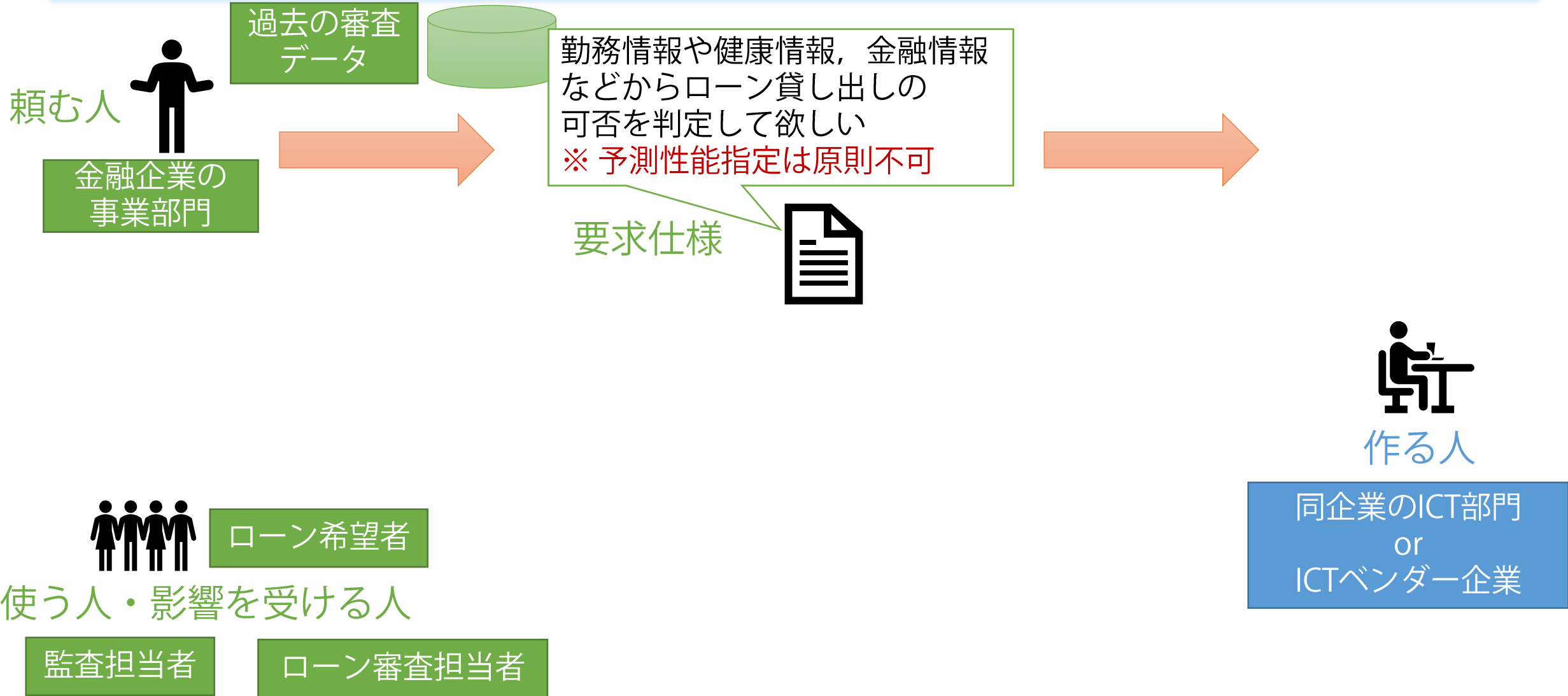
- 機械学習技術によるAI
- 象徴的な事例

## ■ 製品としての「AI」

- 従来ソフトウェアにおける「注文」と「品質」
- AIソフトウェアにおける「注文」と「品質」

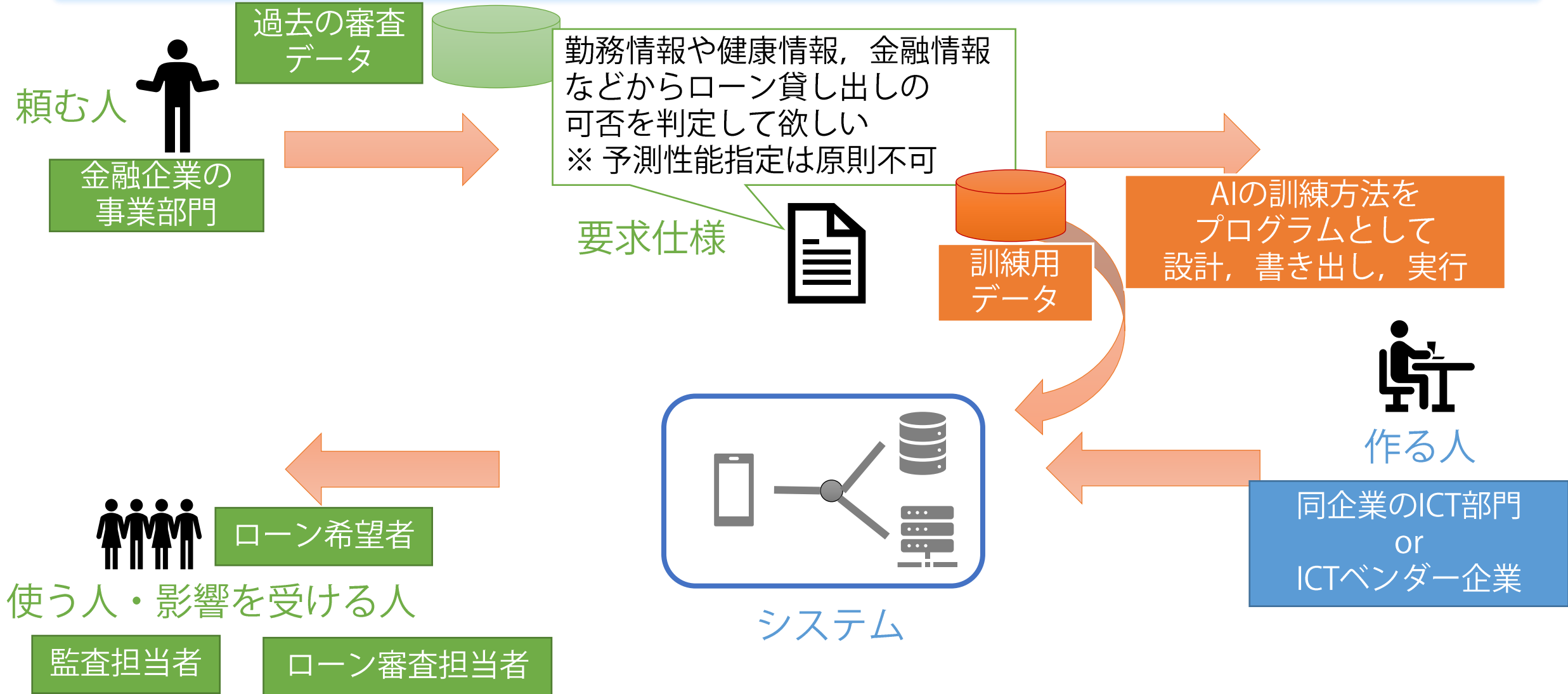
## ■ まとめ・今後に向けて

# AIソフトウェアにおける「注文」



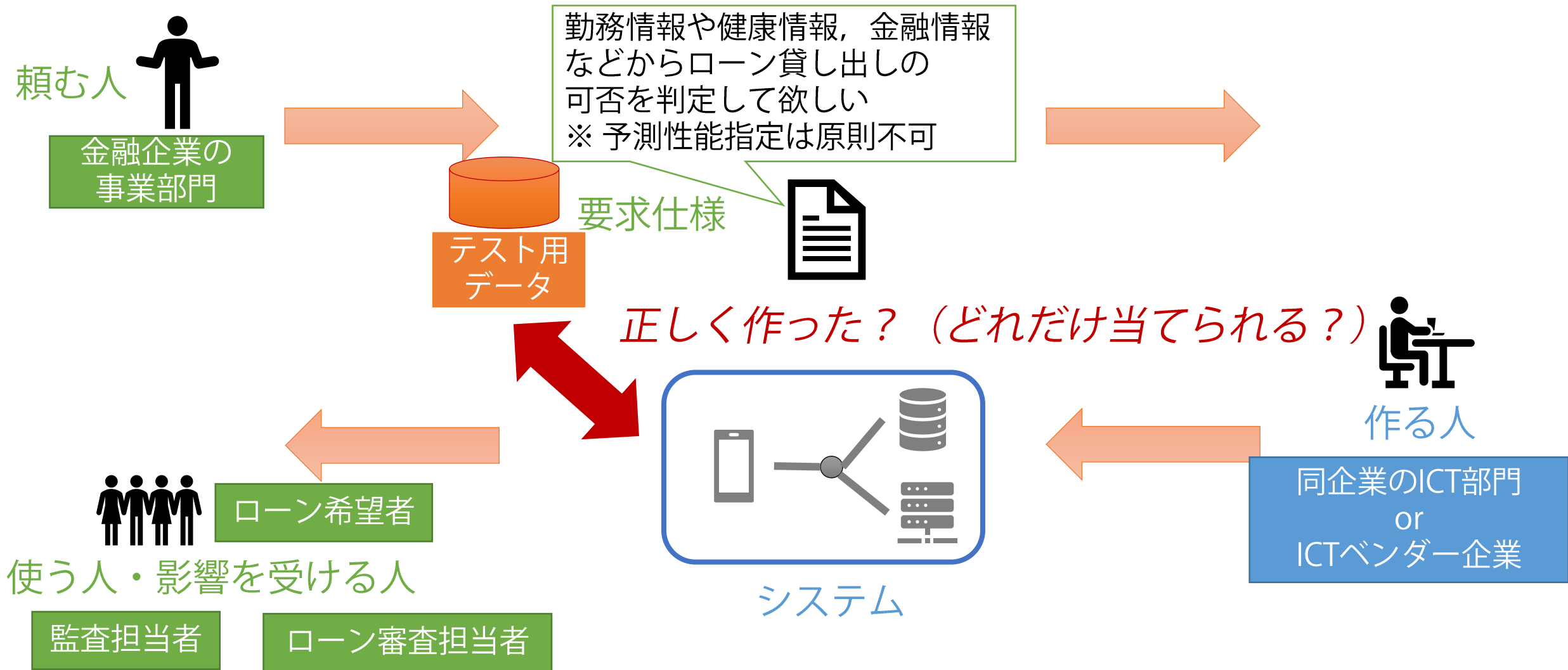
※ 実際は様々なバリエーションがあり, より複雑

# AIソフトウェアにおける「開発」



※ 実際は様々なバリエーションがあり, より複雑

# AIソフトウェアにおける「品質」 (1) 正しさ



※ 実際は様々なバリエーションがあり, より複雑



# AIソフトウェアにおける「品質」 (1) 正しさ

- 正しさ：データに含まれる関係性を満たす
    - そもそも言葉で表現しきれないので機械学習を用いている！
    - 100%正解を出力することはできず，試してみるまで性能は不明
    - 用意したテストデータの個々に対して確認するしかない
      - 「〇〇件中□□件正解した」
      - 評価の意義はテストデータの品質に大きく依存  
(悪い例：傾向が異なる古いデータ，特定の状況が抜けたデータ)
- ➡ 不完全性・不確かさに対する試行・評価の反復が大前提に

# 補足：AIソフトウェアの評価のイメージ

- 二値分類での例：陽性か陰性か
  - 例：カメラ画像からの不良品検出
  - PCR検査など医療診断とも似た考え方

| AI \ 正解    | AIの判断：陽性<br>計45    | AIの判断：陰性<br>計455      |
|------------|--------------------|-----------------------|
| 陽性<br>計20  | 問題検出<br>(正解)<br>15 | 見落とし<br>5             |
| 陰性<br>計480 | 誤検出<br>30          | 問題検出なし<br>(正解)<br>450 |



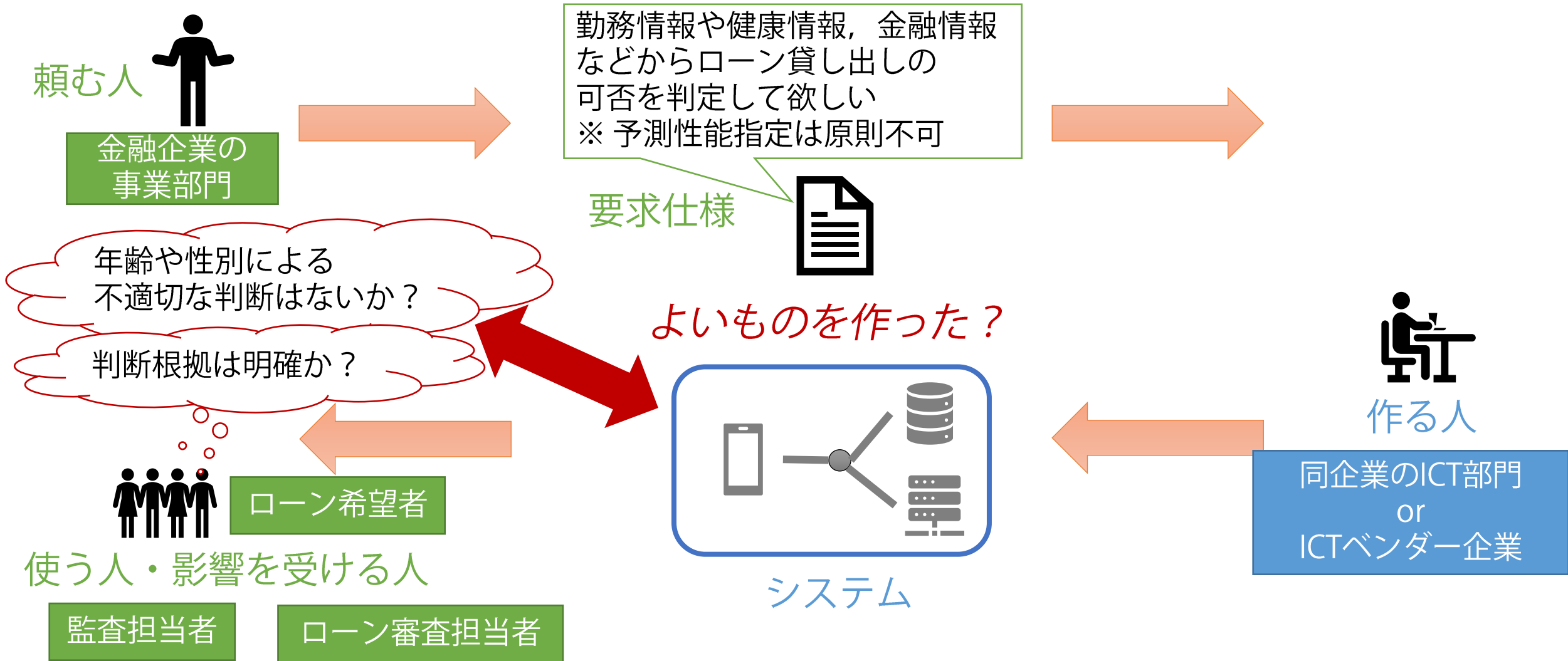
20件の不良品のうち5件検出せず  
(見落とし率25%)



45件の陽性検出のうち30件は良品  
(誤検出率66%)

全体500件の製品のうち  
良品・不良品の正解は465件  
(正解率93%)

# AIソフトウェアにおける「品質」 (2) 価値・妥当性



※ 実際は様々なバリエーションがあり, より複雑

# AIソフトウェアにおける「品質」 (2) 価値・妥当性

## ■価値・妥当性：様々な利害関係者のニーズを満たす

- AIの機能は意思決定に携わり、人間・社会への影響が大きくなる場合がある

## ■前述した従来の難しさがより大きく

- 正解はない・終わりもない

- 「自身が本当に必要としているもの」に予め気づき、漏れなく正確に言葉にすることは非常に難しい

- 状況はどんどん変化する

➡人間・組織・社会への影響を今まで以上に考える必要がある

# 目次

---

## ■ AIのすごさと難しさ

- 機械学習技術によるAI
- 象徴的な事例

## ■ 製品としての「AI」

- 従来ソフトウェアにおける「注文」と「品質」
- AIソフトウェアにおける「注文」と「品質」

## ■ まとめ・今後に向けて

# 関連する動向

---

## ■国内

- 「機械学習工学」という新たな研究分野の立ち上げ
- 「AIの品質」に関する2つのガイドラインを世界に先駆けていち早く発行！
- 契約に関するガイドラインなども次々と

## ■世界でももちろん

- 「説明可能なAI」という大きな研究開発の動向
- EUは人権や倫理の観点から厳しい指針を打ち出し
- 自動運転, 医療診断などでは安全・信頼に関する盛んな議論

# 本講義の振り返り

- ソフトウェアシステムはますます組織や社会の根幹に
- ➔ AI（機械学習）技術により新たな動き
  - 「規則を書き出せない」機能をデータを基に実現可能

（これまでもそうであったが、これまで以上に）

組織や社会に踏み込んだ価値創造・課題解決へ

- 「作る人」だけではなく「頼む人」「使う人」の協働がカギ
- 不確か・不完全であることを前提として、  
試行錯誤を通してよくしていく