機械学習による
不適切なクラウドソーシング
タスクの検出

馬場雪乃、鹿島久嗣（東大）
木下慶、山口豪志、秋好陽介（ランサーズ(株)）

第5回データ工学と情報マネジメントに関するフォーラム
(DEIM2013), 2013年3月4日

クラウドソーシングでは利用規約に違反する「不適切なタスク」が投稿されることがある

不適切なタスクと判断される例：個人情報の入力を要求

生活保護を不正受給している恐れがある方が近辺にいたら、その人の情報を教えてください

氏名 [　　]　住所 [　　]

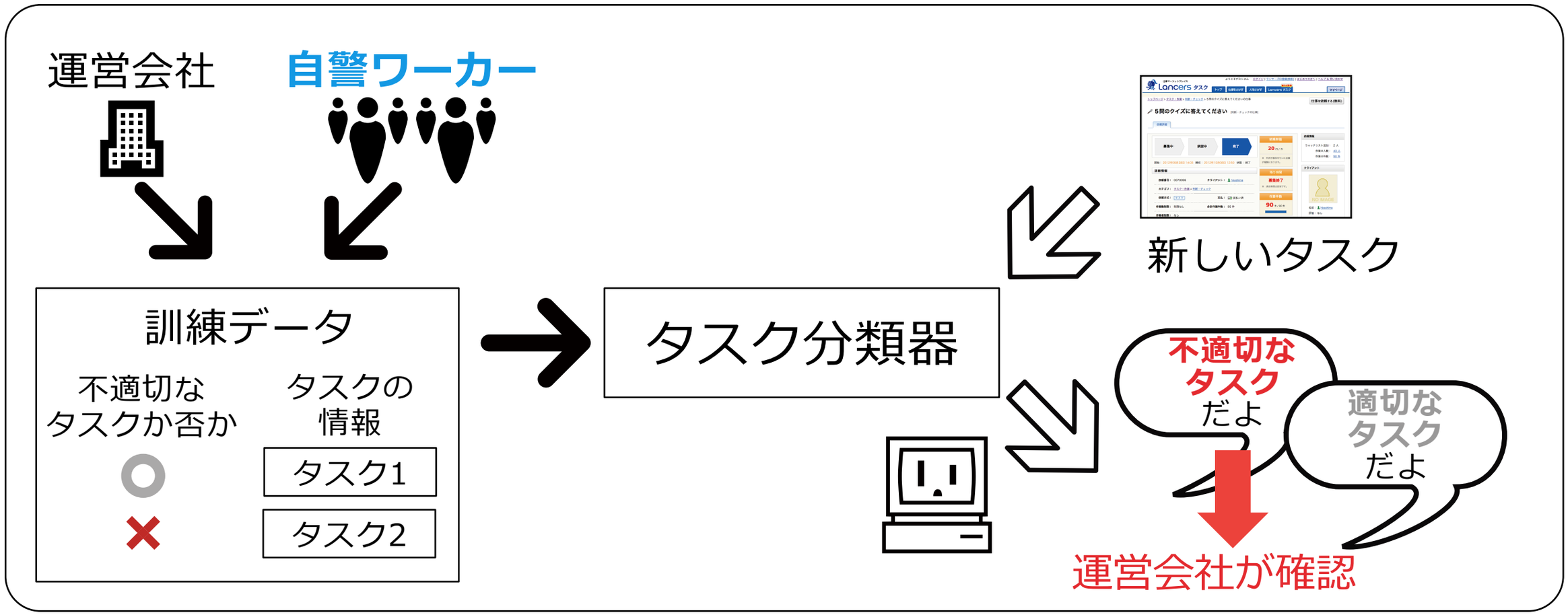
その他コメント

[　　]

運営会社はタスクを監視し不適切タスクを見つけ次第依頼停止

タスク数が増えるにつれ監視コストが問題になる

運営会社の監視作業を支援するために
不適切クラウドソーシングタスクを機械学習で検出



商用クラウドソーシングサービスLancersのデータを使って実験

結果1: **運営会社**の監視結果で訓練することで**高精度**を達成

結果2: **自警ワーカー**の監視結果を組み合わせることで**精度向上**

結果3: 自警ワーカー利用で運営会社の**訓練データ作成コスト削減**

クラウドソーシングでは品質管理が課題

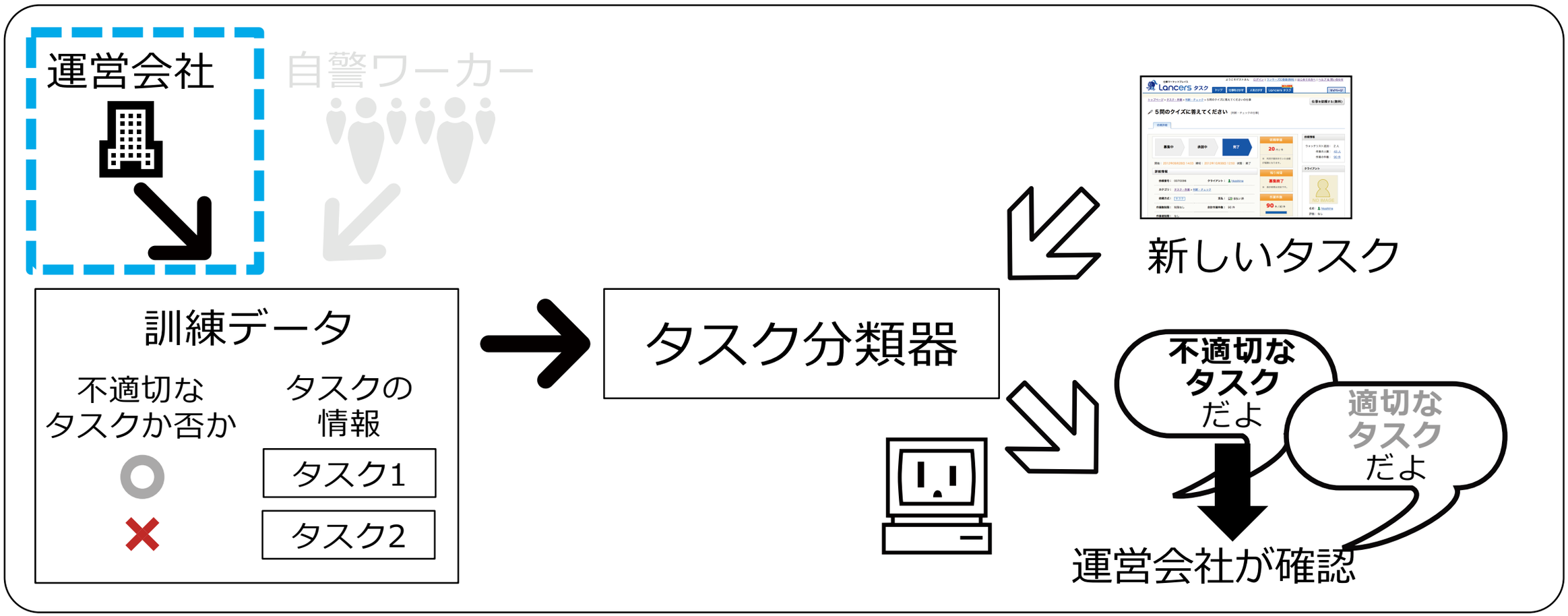
リクエスター・ワーカーともに信頼性が事前にわからない
品質管理が課題

リクエスター・**タスクの品質管理**は
取り扱われていない

ワーカー・作業結果の品質管理は
既存研究で対象とされている

実験 1

運営会社の監視結果だけを用いて分類器を訓練

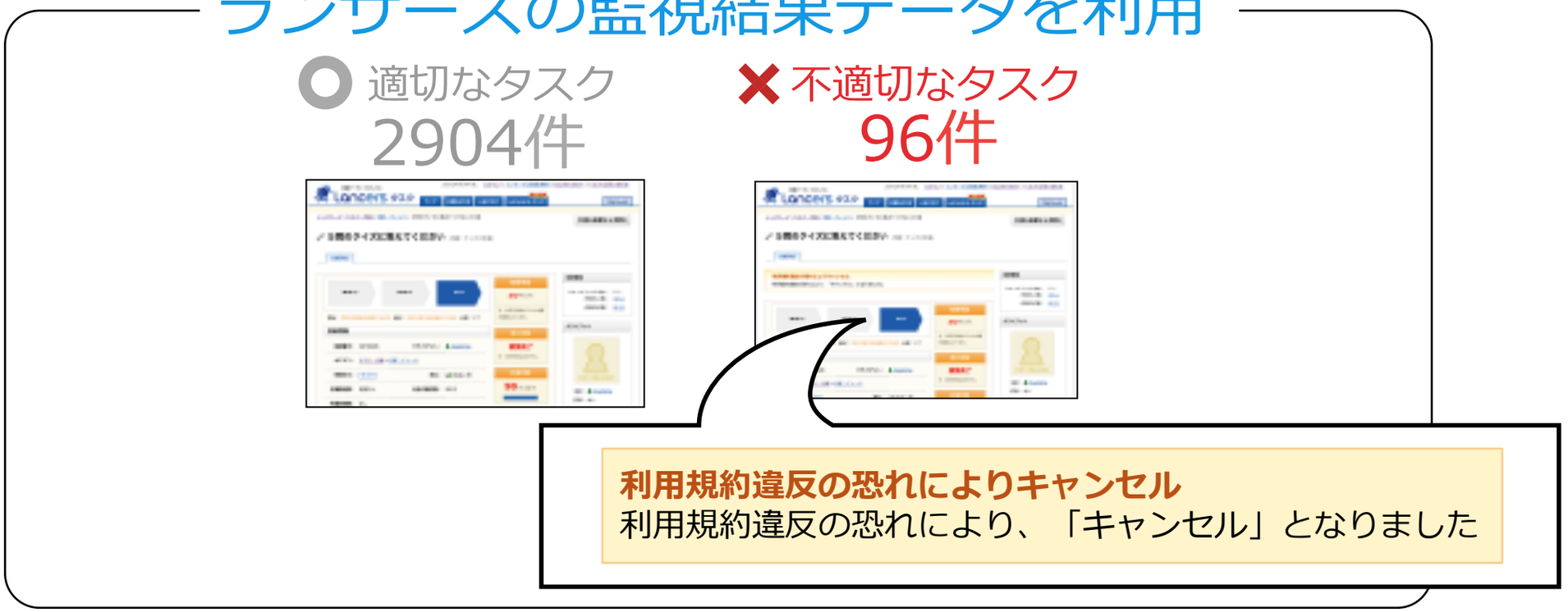


結果1: 運営会社の監視結果で訓練することで高精度を達成

結果2: 自警ワーカーの監視結果を組み合わせることで精度向上

結果3: 自警ワーカー利用で運営会社の訓練データ作成コスト削減

商用クラウドソーシングサービス「Lancers」の運営会社による監視結果データを利用



※ランサーズに投稿されたタスク方式の依頼の一部

分類器構築に用いた特徴量

タスクのテキスト特徴量	タイトル，概要，本文のBag-of-Words
タスクの非テキスト特徴量	依頼人数，作業単価，ワーカー条件の有無など
依頼者ID	依頼者ごとに割り当てられるID
依頼者の非テキスト特徴量	依頼者属性，信頼性，これまでの実績など

SVMを用いて平均AUC 0.950を達成
タスクのテキスト特徴量が特に有効

訓練に用いる特徴量を変化させたときのAUC値
（用いた特徴量を塗りつぶして表示）

タスクの テキスト特徴	タスクの 非テキスト特徴	依頼者ID	依頼者の 非テキスト特徴	AUC （平均と標準偏差）
			■	0.771 ±0.040
		■		0.848 ±0.032
		■	■	0.840 ±0.036
	■			0.734 ±0.042
	■		■	0.835 ±0.036
	■	■		0.890 ±0.030
	■	■	■	0.854 ±0.040
■				0.902 ±0.032
■			■	0.910 ±0.030
■		■		0.904 ±0.031
■		■	■	0.911 ±0.031
■	■			0.946 ±0.016
■	■		■	**0.950** ±0.015
■	■	■		0.947 ±0.015
■	■	■	■	**0.950** ±0.015

タスクのテキスト特徴量が特に有効

平均AUC 0.950と**高精度**

※分類手法はSVMを採用．60%で訓練し残りでテスト．100回繰り返した際のAUC平均値と標準偏差を提示

実験２

「自警ワーカー」のデータも訓練に使用

運営会社

自警ワーカー

クラウドソーシングで監視を依頼

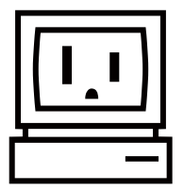
1タスクあたり3人・計97人が作業

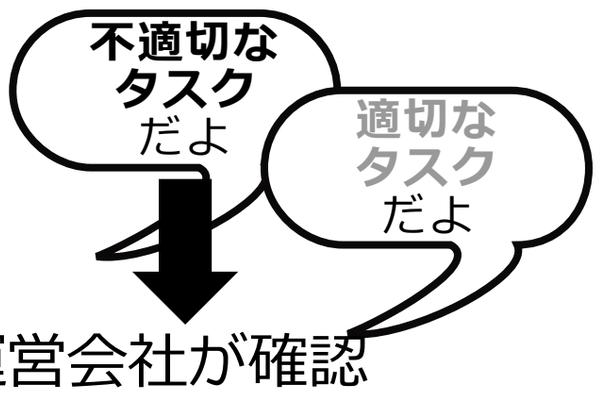
新しいタスク

訓練データ

不適切なタスクか否か	タスクの情報
○	タスク1
×	タスク2

タスク分類器





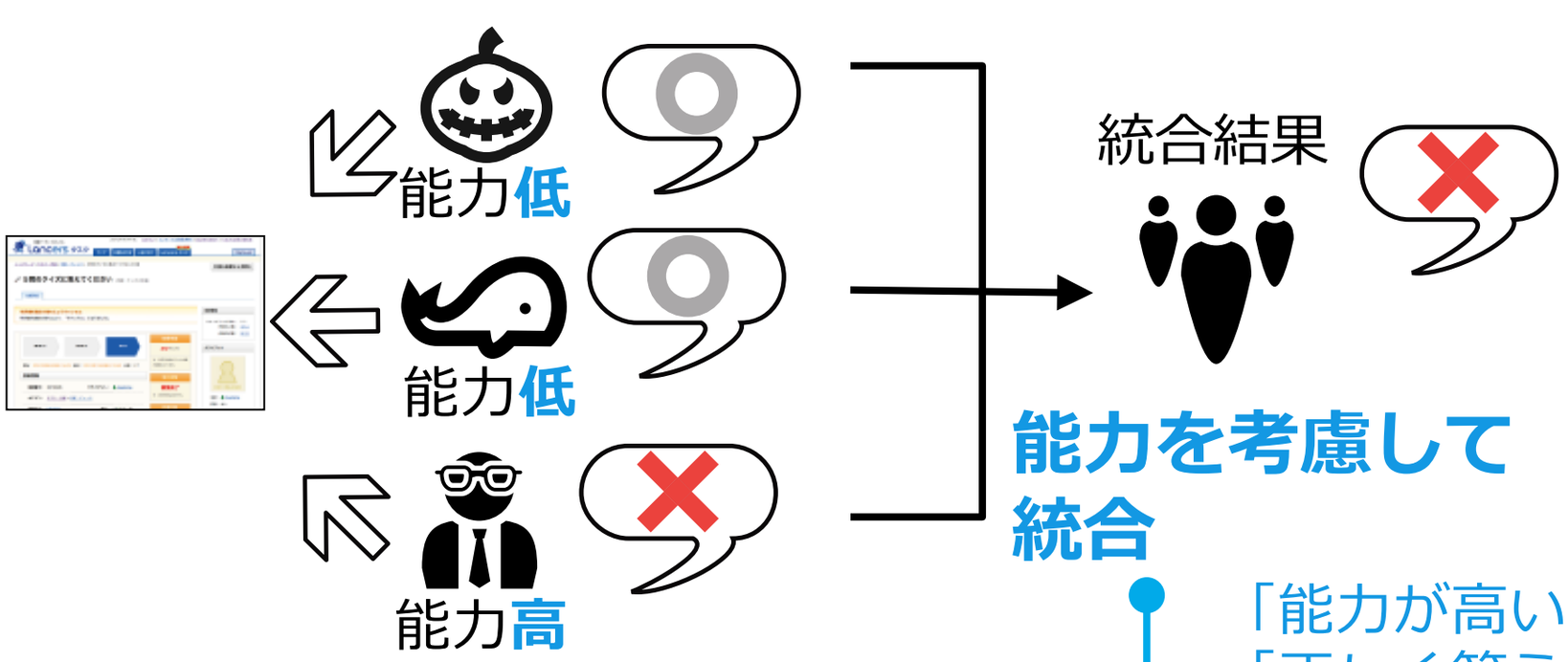
結果1: 運営会社の監視結果で訓練することで高精度を達成

結果2: 自警ワーカーの監視結果を組み合わせることで精度向上

結果3: 自警ワーカー利用で運営会社の訓練データ作成コスト削減

工夫１
能力を考慮した既存の回答統合手法を利用

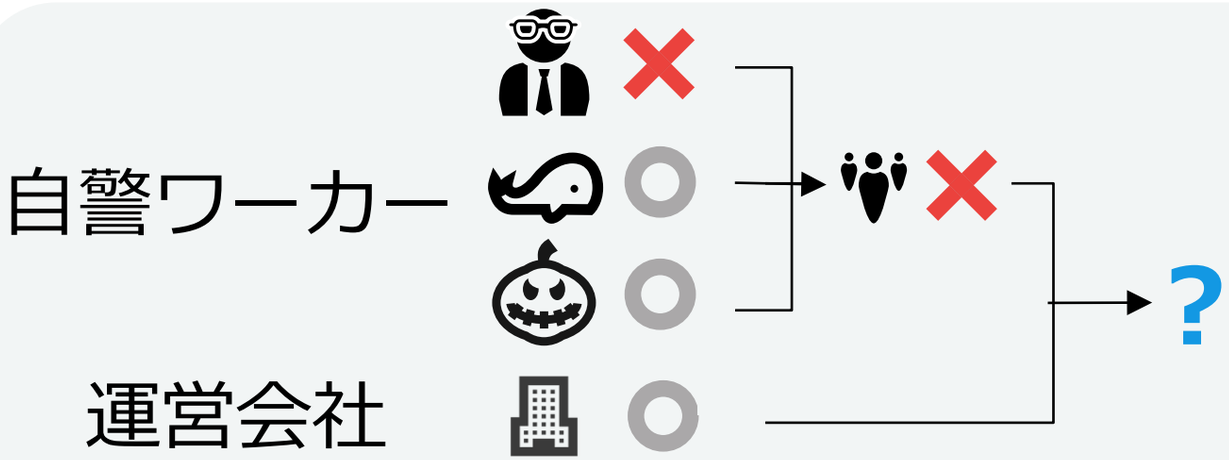
「不適切タスク」の基準を熟知している運営会社と異なり
ワーカーは**能力にばらつき**があり皆が正しく判断できるとはない
→能力を考慮して統計的に回答統合 [Dawid&Skene 79]



「能力が高い人は正しく答える」
「正しく答える人は能力が高い」
を利用して**能力**と**正しい回答**を
交互に推定

[Dawid&Skene 79] A. P. Dawid and A. M. Skene.: Maximum Likelihood Estimation of Observer Error-Rates Using the EM Algorithm, Journal of the Royal Statistical Society, Series C (Applied Statistics) , Vol. 28, No. 1 (1979)

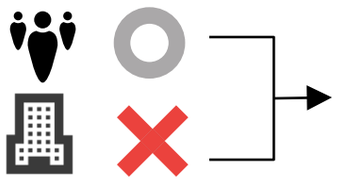
工夫２
運営・自警ワーカーの監視をルールを定めて統合



運営会社と自警ワーカーの
監視結果を
どう統合するか？

(1) 監視結果が同じ場合

(2) 監視結果が異なる場合：３通りのルール候補



❶ 常に「適切○」を採用

❷ 常に「不適切✕」を採用

❸ 当該サンプルを訓練に加えない

○✕と✕○で３×３＝計９通りのルール候補

自警ワーカーの監視結果を組み込んで精度向上することを確認

ルールごとのAUCの平均値と標準偏差

ルール		👣 ✕ 🏢 ○		
		「適切」を採用	「不適切」を採用	訓練に追加しない
👣 ○ 🏢 ✕	「適切」を採用	0.763 ±0.076	0.817 ±0.063	0.816 ±0.075
	「不適切」を採用	0.950 ±0.015	0.951 ±0.017	**0.962** ±0.013
	訓練に追加しない	0.791 ±0.070	0.841 ±0.061	[illegible] ±0.032

運営会社の監視結果だけで訓練した場合（0.950）よりも**高精度**

最も上手くいったルール

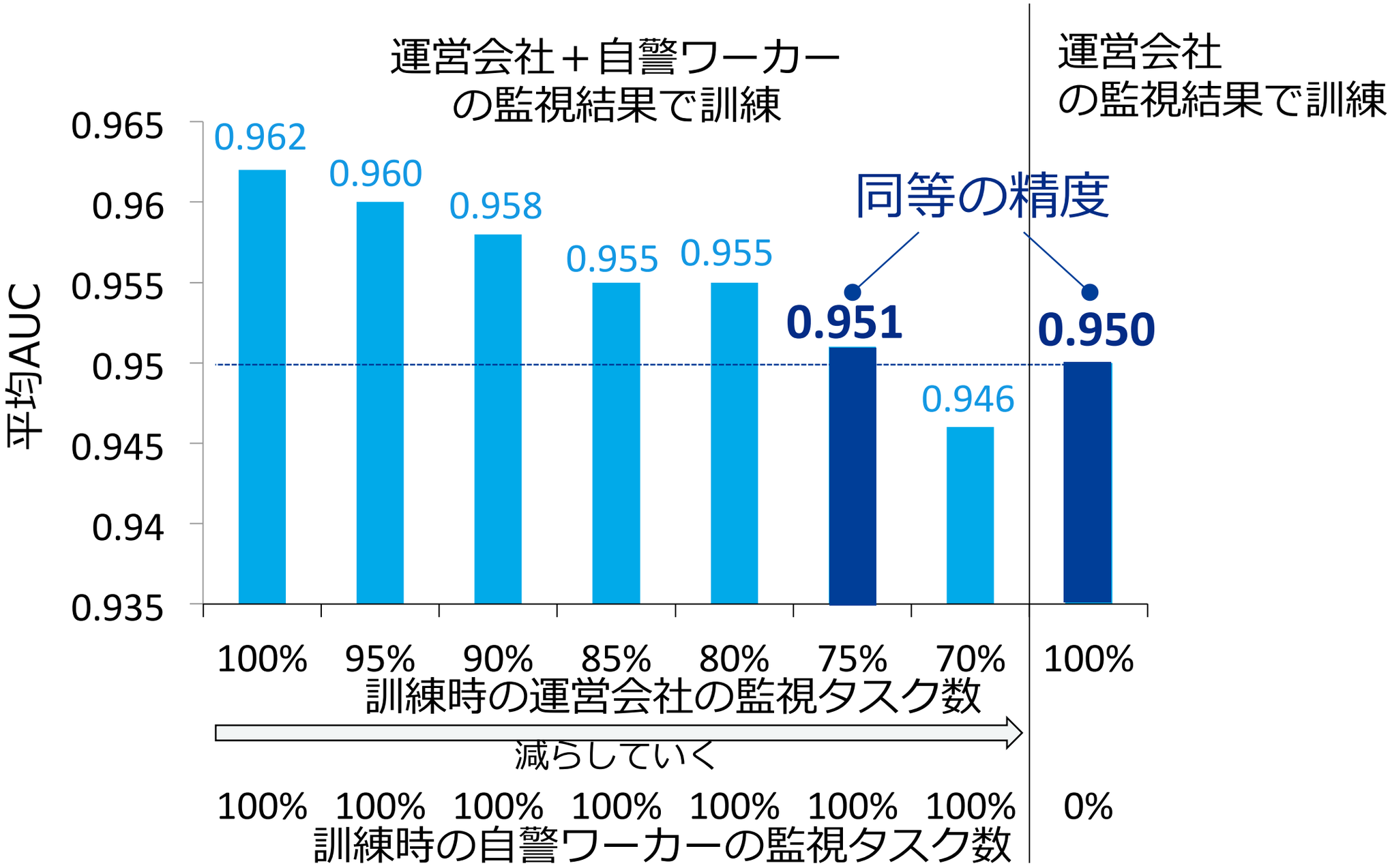
👣 ○ 適切
🏢 ✕ 不適切
→ ✕

👣 ✕ 不適切
🏢 ○ 適切
→ SKIP

🏢 **運営会社は厳しい判断**
「不適切」判定は常に正しい

👣 **自警ワーカーは運営会社よりも甘い判断**
自警ワーカーの「不適切」判定に運営会社が反対するときは訓練に使わない方がいい

自警ワーカーの判断を組み込むと
訓練時の運営会社監視数を25%削減しても精度維持



自警ワーカーに監視依頼することで
運営会社の訓練データ作成コストを削減できる

不適切なクラウドソーシングタスクの検出を自警ワーカーにも監視依頼することで高精度で達成

運営会社のタスク監視作業を支援するため
運営会社と**自警ワーカー**の監視結果を使って
不適切なタスクを検出する分類器を構築

1. 運営会社の監視結果で訓練した分類器が平均AUC 0.950
   不適切タスクを高精度検出
2. 運営会社と自警ワーカーの監視結果で訓練した分類器が
   平均AUC 0.962
   自警ワーカーの監視結果も利用することで精度が向上
3. 運営会社の監視結果を25%削減しても平均AUC 0.951
   自警ワーカーの監視結果の利用は，運営会社の訓練データ作成コスト削減にも効果あり