Learning Representations by Humans, for Humans

発表者：楊明哲

2022年4月18日 @読み会

論文情報と選んだ理由

選択理由

- 人間に対して情報技術で作用（介入）について知るため
- Human-in-the-loop系の論文

Sophie Hilgard [1 *] **Nir Rosenfeld** [2 *] **Mahzarin Banaji** [3] **Jack Cao** [3] **David C. Parkes** [1]

[*]Equal contribution [1]School of Engineering and Applied Science, Harvard University, Cambridge, MA, USA [2]Department of Computer Science, Technion - Israel Institute of Technology [3]Department of Psychology, Harvard University, Cambridge, MA, USA. Correspondence to: Sophie Hilgard <ash798@g.harvard.edu>, Nir Rosenfeld <nirr@cs.technion.ac.il>.

貢献：人間を介して人間に有益な表現を獲得可能に

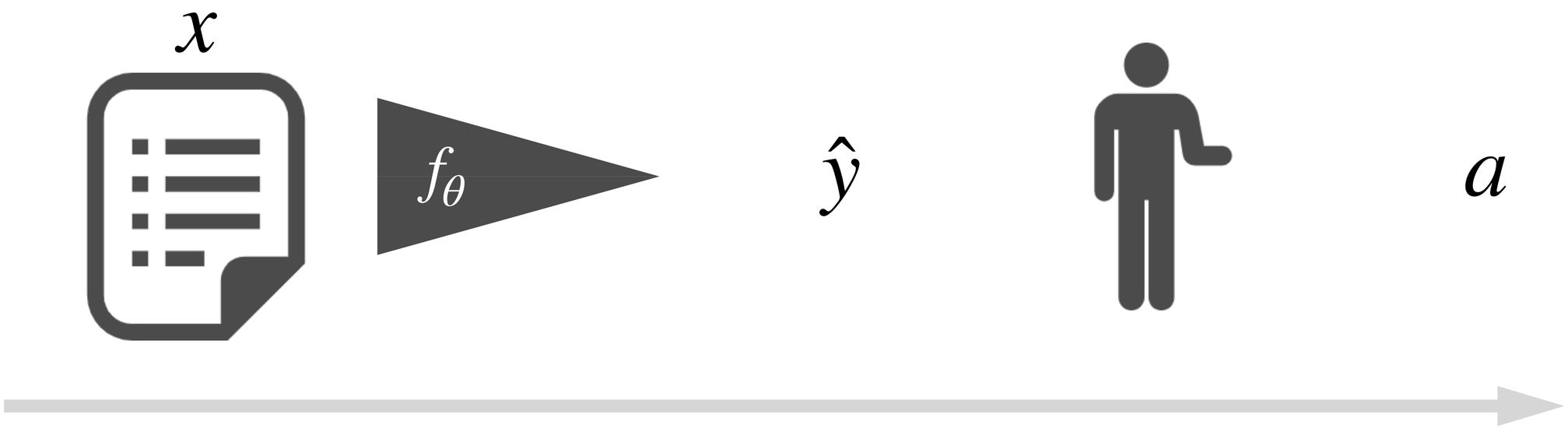
背景：機械学習が発展してきて，意思決定に使われきている

問題：安全性，公平性などを考慮する際に鵜呑みにしずらい

提案：人間が理解可能な方法で情報を提示し，
人間の最終決定を支援するフレームワークMoMを提案

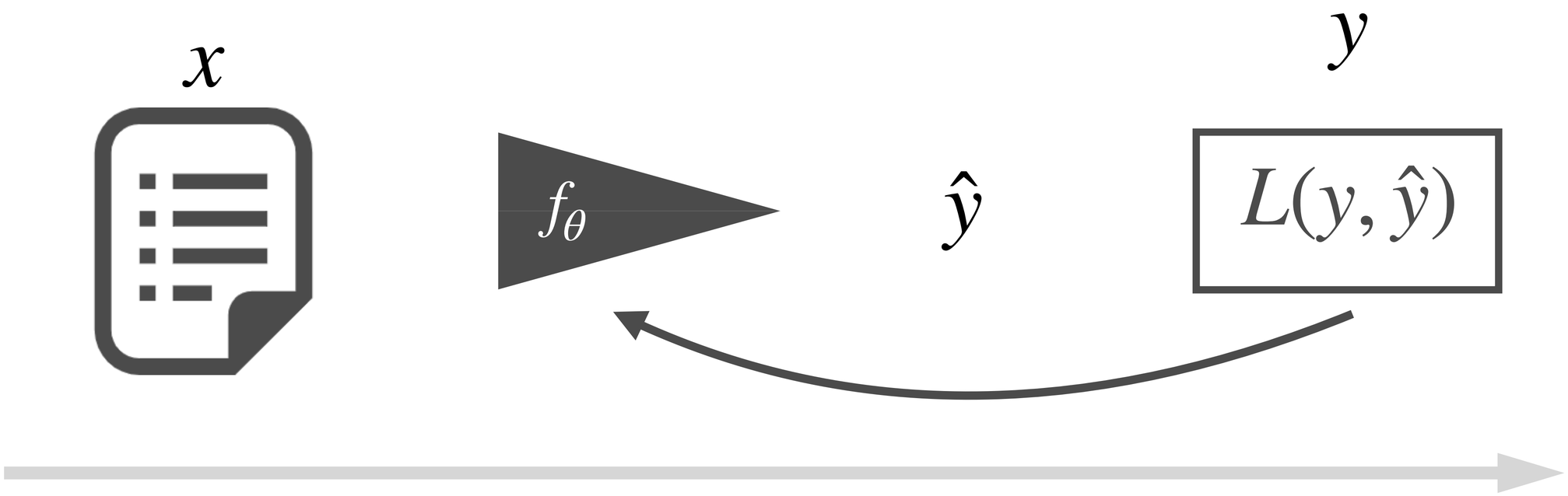
結果：3つのタスクを実施.
人間の意思決定に対して向上させることを示した

いままでの意思決定支援：出力を鵜呑みしない



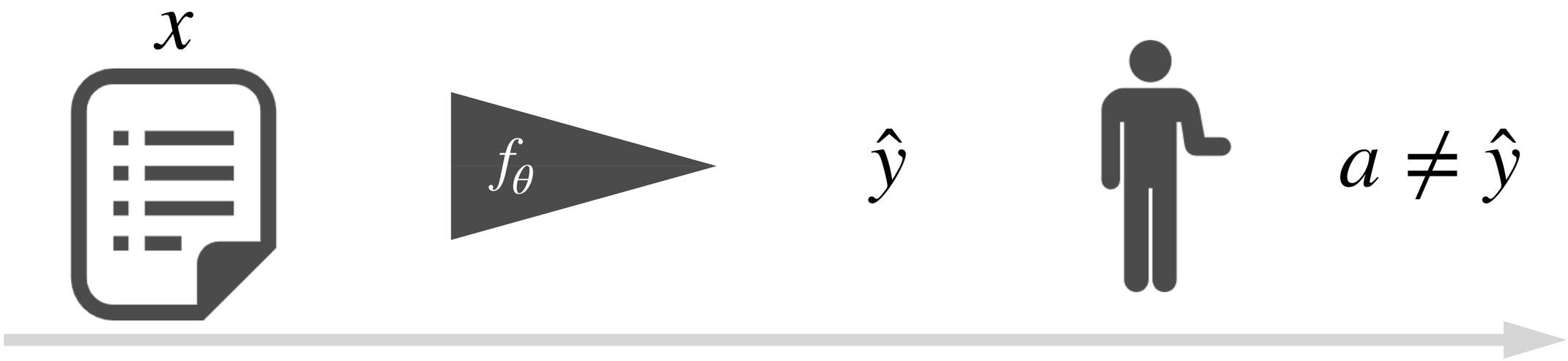
機械の出力 $\hat{y}$ を人間が受け取り，行動 a をとる

いままでの手法：よい出力を学習（訓練時）



教師あり学習で良い出力になるように学習

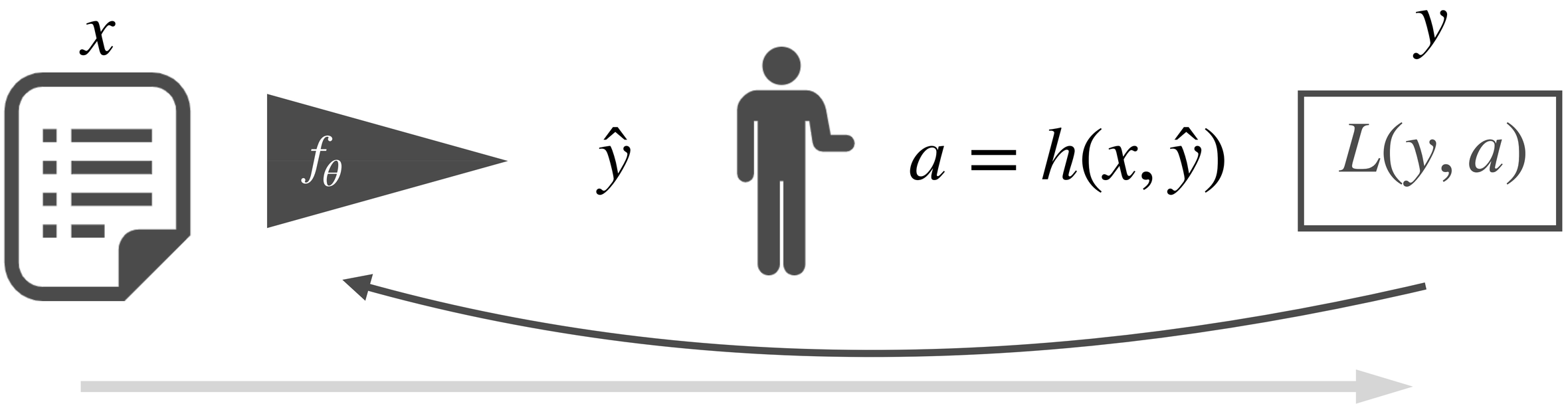
現実：推論結果と行動は一致しない



人間が間に入るので $\hat{y} \neq a$ になることがある

意思決定の状況によって機械と人間の意思決定が一致しない

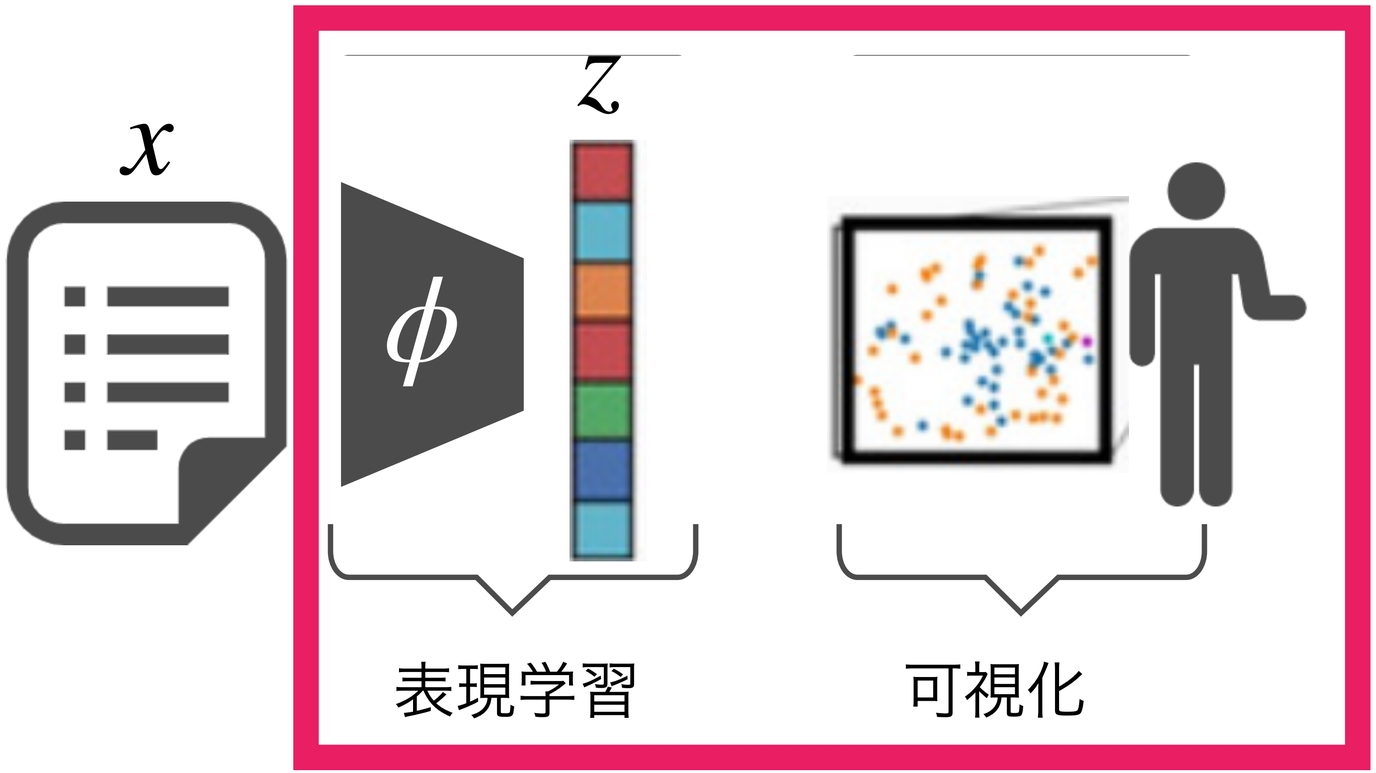
理想：人間の意思決定を最適化



人間の行動を含めて最適化することが理想

$\hat{y}$ がただの出力で人間にとって理解し難い

提案手法：人間でも理解できる方法で支援したい



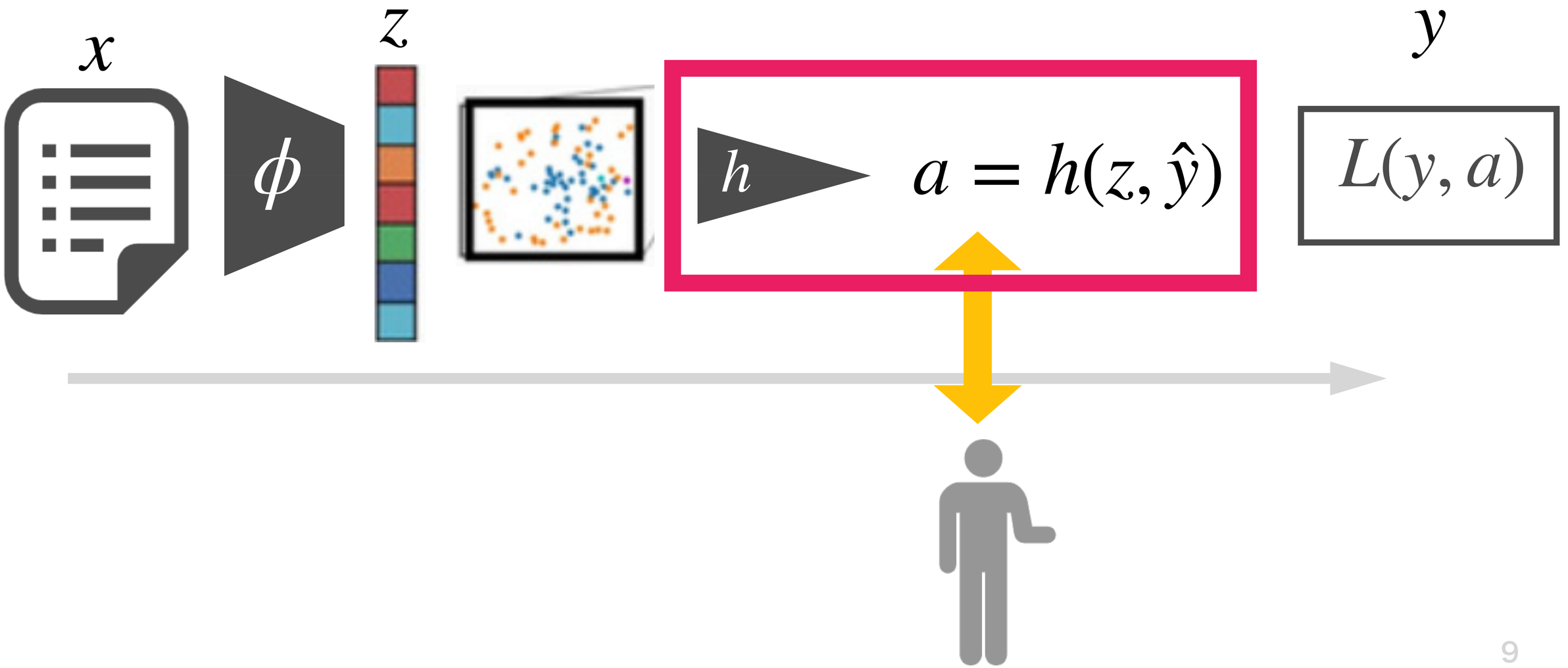
$$a = h(z, \hat{y})$$

y

$$L(y, a)$$

人間がいるため，勾配を伝播できない

人間の代理モデルを設定して，勾配を伝播可能に！



表現を使った適切な可視化は難しい

表現 z から人間が理解できるようにするために,

可視化の操作が必要

- 棒グラフや，画像，ハイライトとか…

意思決定に対して，良い介入が行えるような可視化はそれだけで研究になるくらい難しい

本研究では，良さそうな可視化を選んで利用する

実験：3つのタスクで有効性を示す

タスク1：高次元データを低次元に圧縮したときに
有用な表現を獲得できるか？

タスク2：実際のタスクでも
獲得した表現による支援は有効か？

タスク3：機械が知り得ない追加情報も
表現として獲得できるか？

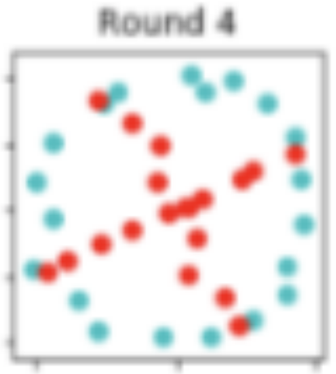
タスク1：良い表現を獲得できるか？

次元圧縮をしてデータの可視化が可能になる

- 既存の次元圧縮は統計的な最適化しか考えていない

高次元なデータを人工的に作り，分類問題に取り組む

- 直交射影した時に，”X”と”O”の形に並ぶように作成



タスク1：表現モデルと代理モデル

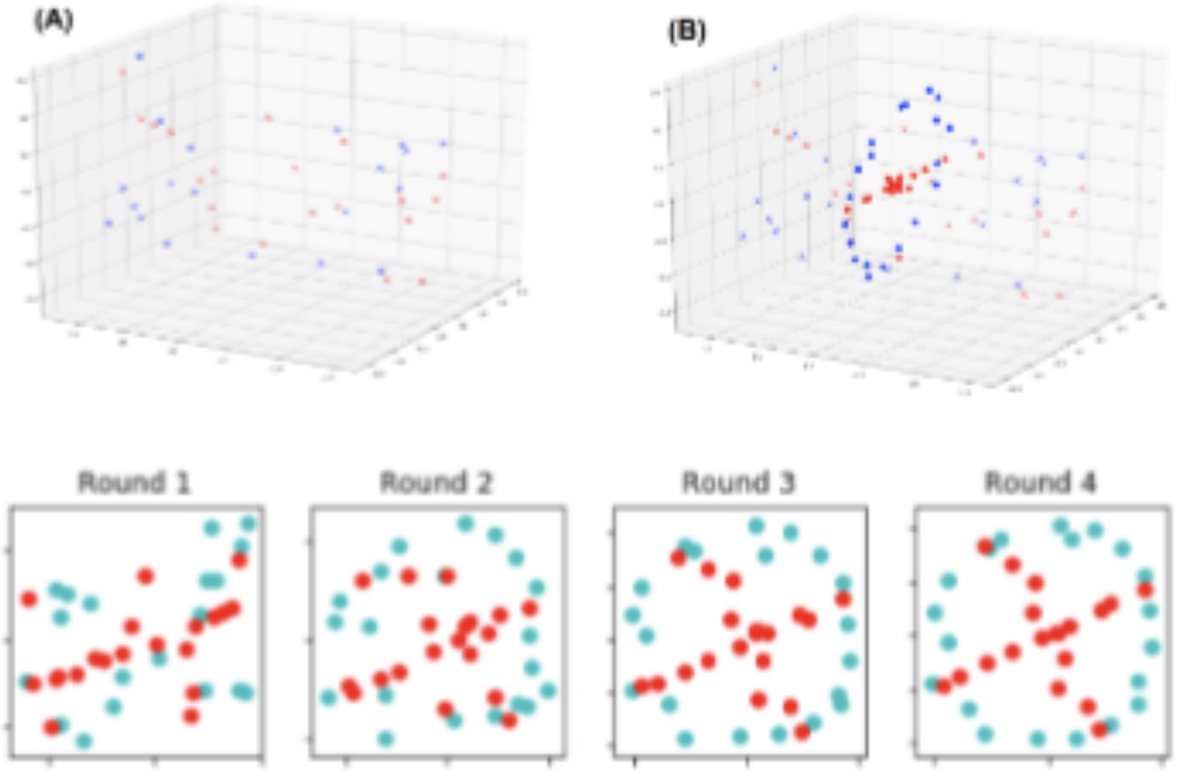
表現モデル

- 3x2の線型写像行列

代理モデル

- 一層で3x3の畳み込みネットワーク
- 人間の視覚をおおまかに再現する

タスク1：結果



Accuracy 68% → 91%

75%の実験参加者が精度100%を達成

タスク2：リアルなデータでも支援できるか

ローン審査を題材にしたタスク

- $y = 1$: 完済した，$y = 0$: 滞納，$a = 1$: 承認，$a = 0$: 拒否
- 損失関数：$l(y, a) = \mathbf{1}_{y \neq a}$

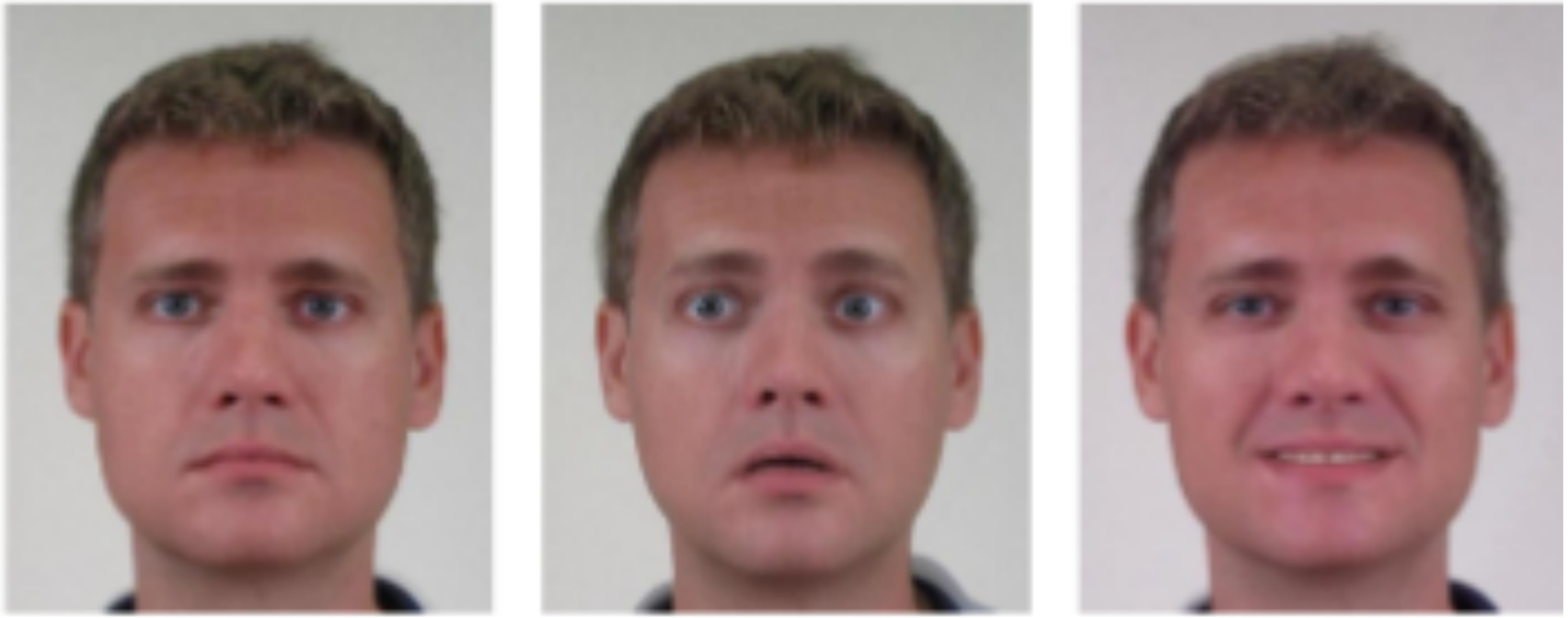
MTurkで参加者を集め計5000回答を集めた

目標：表現によるアドバイスで
意思決定の支援が可能かどうか

タスク2：顔表情によって表現を可視化

Facial avatarを使って，顔表情で可視化を行う

- Chernoff faces: 顔パーツにそれぞれ数値が対応して変化



タスク2：表現モデルと代理モデル

表現モデル

- 全結合の25ユニットを1層

代理モデル

- 全結合の20ユニットを2層

タスク2：直接的なアドバイスと同等

No adviceより11.5%向上, 予測アドバイスと同等

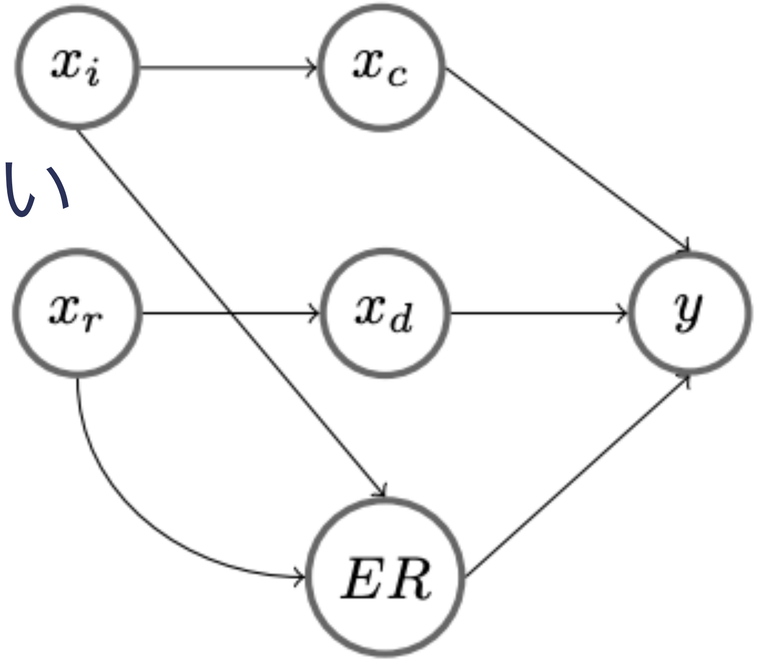
タスク3：人間しか知り得ない情報も獲得できるか

人間しか知り得ない情報（知識，常識とか）を表現として獲得できるかをみたい

医療診断のタスクを設計

- 追加情報 s を設定する → 人間しか見れない
- この追加情報は出力に影響を及ぼす設定

データセット自体は人工的に生成



タスク3：表現モデルと代理モデル

Always: The human always fully incorporates the side information,

$$h(x, w, s) = w^\intercal x + s$$

Never: The human never incorporates the side information,

$$h(x, w, s) = w^\intercal x$$

Or: The human becomes less likely to incorporate side information as weight is put on x_i, x_r,

$$h(x, w, s) = w^\intercal x + \sigma(1/\max(\max(x_i, x_r), .0001) - 2). \cdot s$$

Note that max(.0001) is required to prevent numerical overflow, and -2 recenters the sigmoid to allow for values $< .5$.

Coarse: The human incorporates s as in Or, but uses a coarse, noisy version of s, $s' = 2 \cdot \mathbb{1}\{s \geq 2\}$

$$h(x, w, s) = w^\intercal x + \sigma(1/\max(\max(x_i, x_r), .0001) - 2). \cdot s'$$

表現モデル

- 重みと特徴量の線形和

代理モデル

- 重みと特徴量の線形和+追加情報の線形和

タスク3：

	M∘M	h(Machine)
Or	1.0	.894
Coarse Or	.951	.891
Never	.891	.891
Always	1.0	.674

h(Machine)

- 純粋に訓練を行う場合

MoM

- 表現も一緒に訓練する

MoM以外では，追加情報によって過学習が起きやすい

まとめと感想

まとめ

人間から人間のために意思決定を支援する
フレームワークを提案

人間の最終決定を支援できるように，表現を学習し可視化

感想

モデル自体を厳密に定義しないからシンプルだけど
実利用を考えた時うまくいかない気がしている

最後の実験はしっくりしないが，人間しか知り得ないものを
抽出できたらいいなにはすごく同意