人工知能プログラミング演習

第 6 回

株式会社ディー・エヌ・エー／東京電機大学 講師

甲野 佑

東京電機大学鳩山キャンパス 2018/May/21

Google drive

授業ドライブ：https://goo.gl/QSppHs

前々回まで見ていた google drive は見れなくなります
開けない人は副手に連絡

今回の内容

強化学習概要

↓

A/B テスト

↓

速さと正確さのトレードオフ

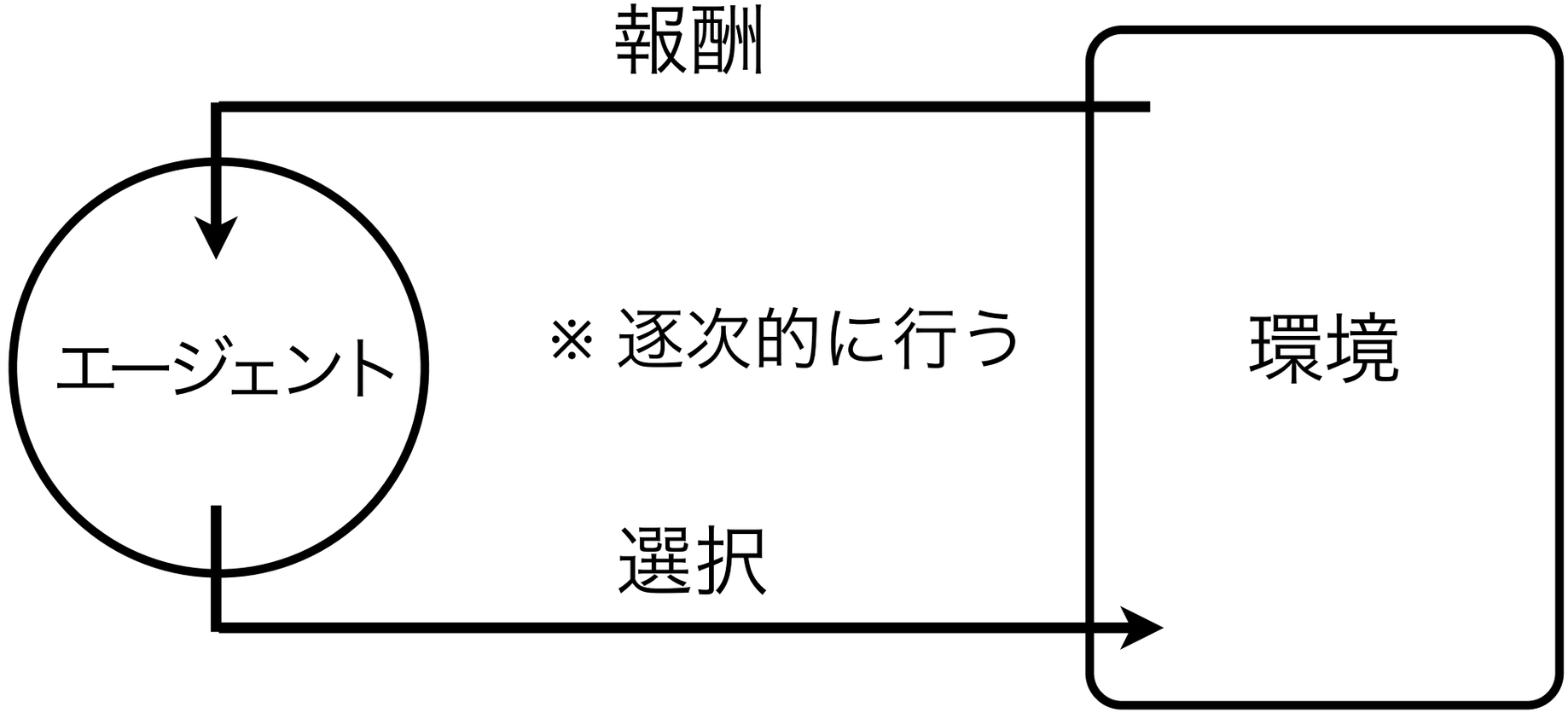
↓

多本腕バンディット問題

↓

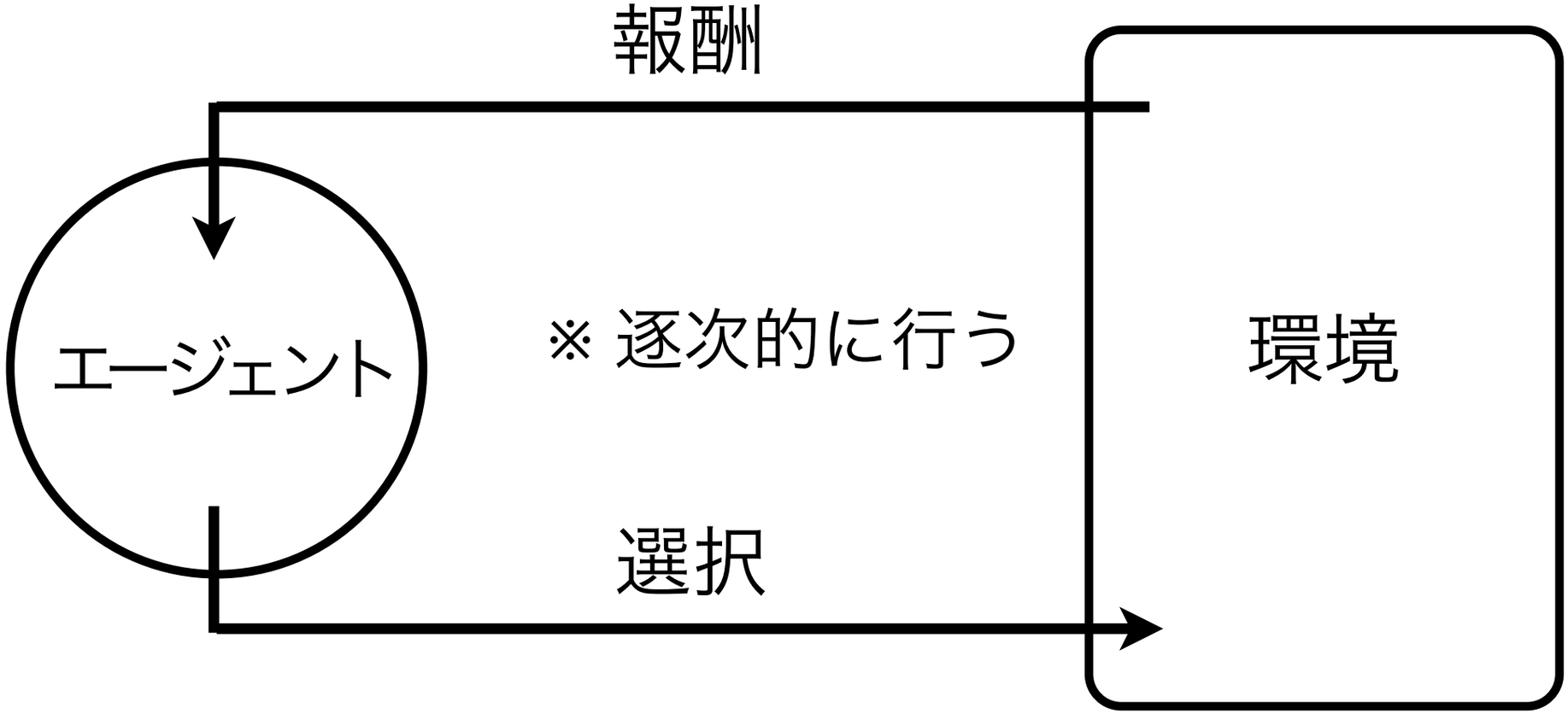
メタバンディットアルゴリズム

強化学習



エージェントが実際に環境に**働きかけ**て
試行錯誤により多数の選択肢から最良の選択肢を**見つけ**て
獲得報酬を最大化する事が目的

強化学習

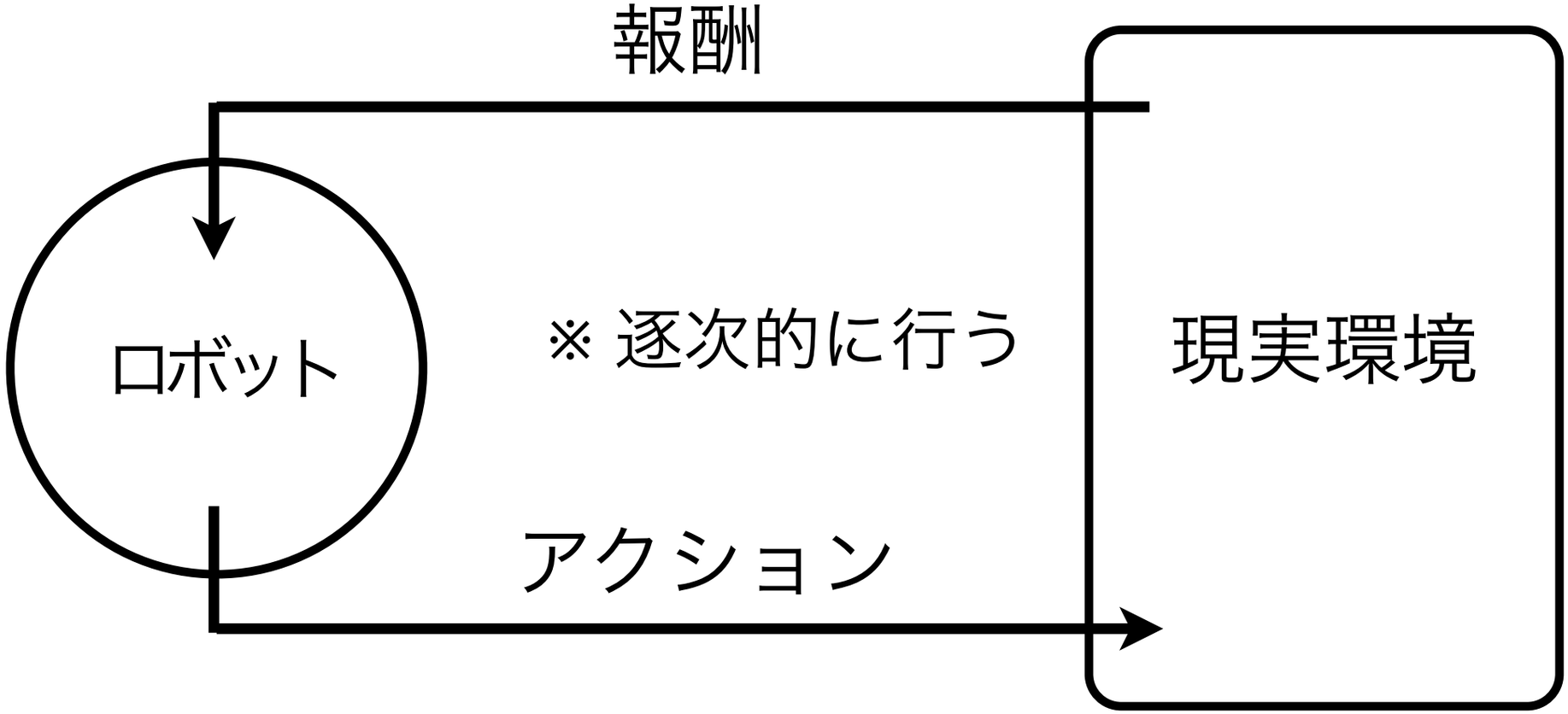


強化学習には様々な側面があるが

試行錯誤が必要な

未知の環境での**逐次的**な**意思決定**として扱った

強化学習



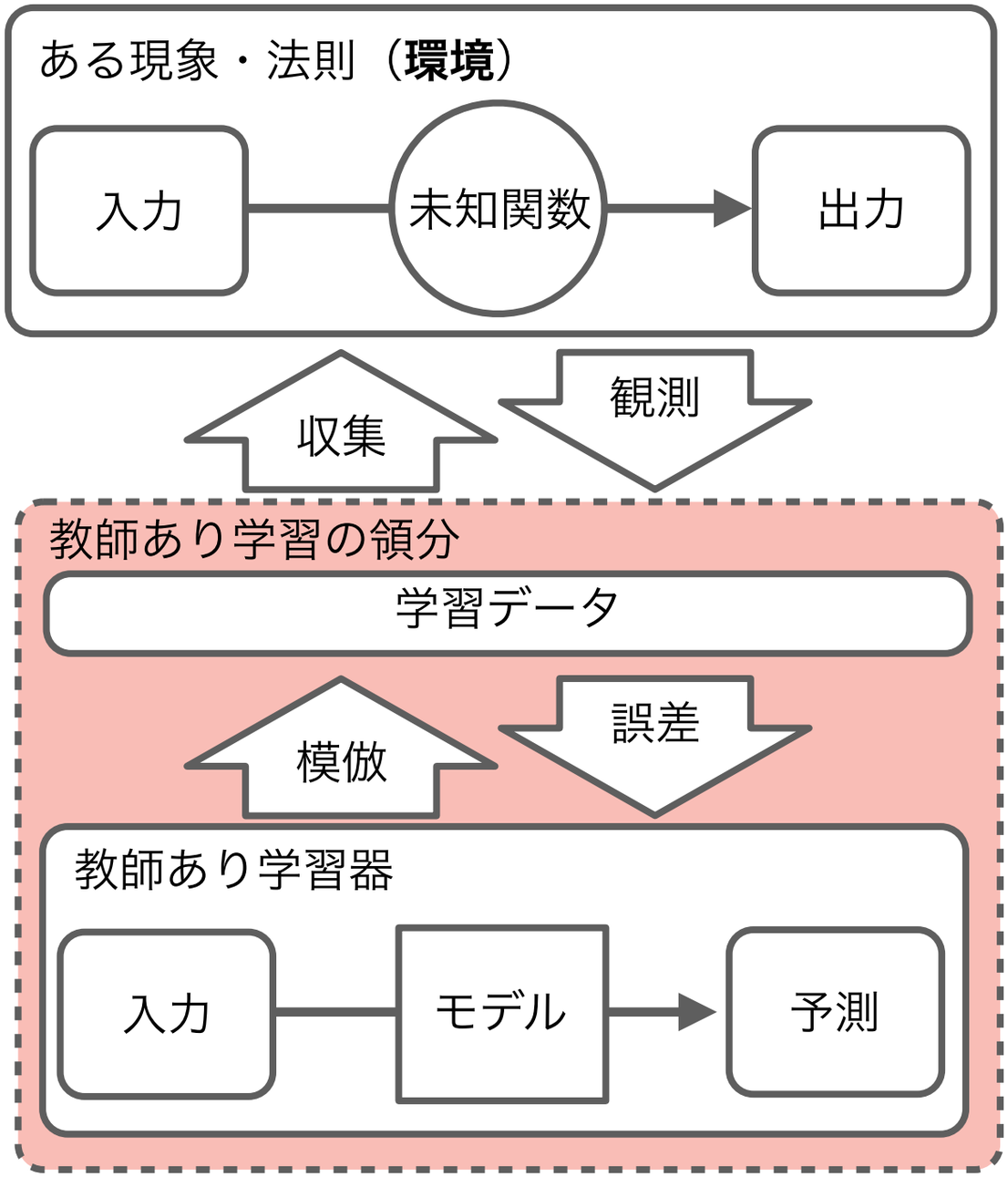
その理由は

【**ゼロ**から学習し**新たな状況**にも対処できる人工知能】

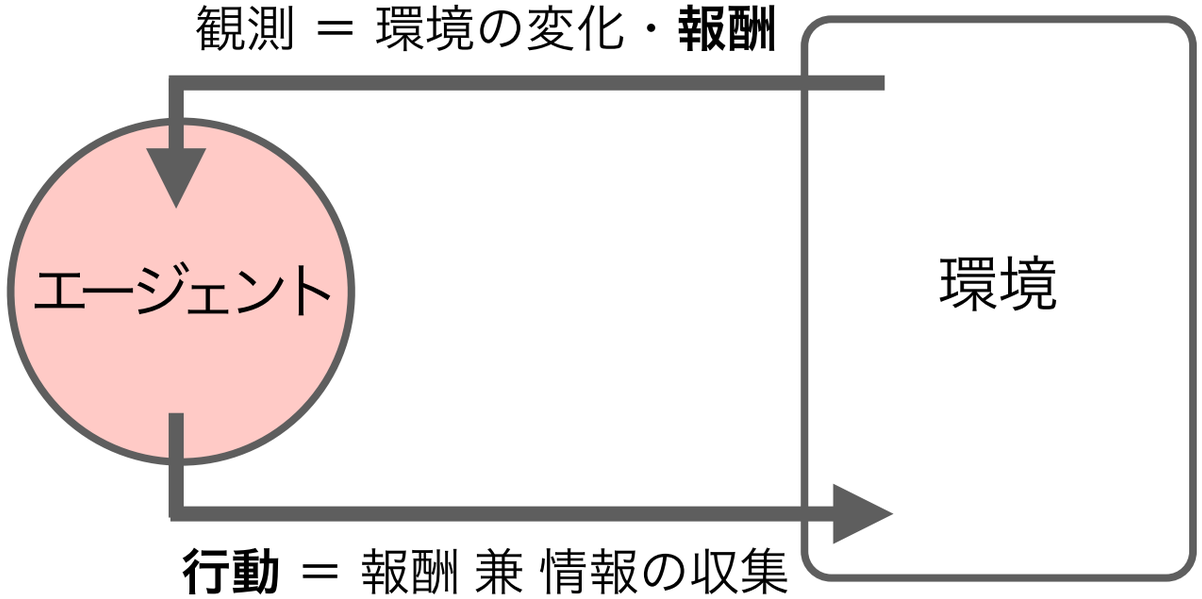
への寄与を目指しているから

強化学習とは

【教師あり学習】



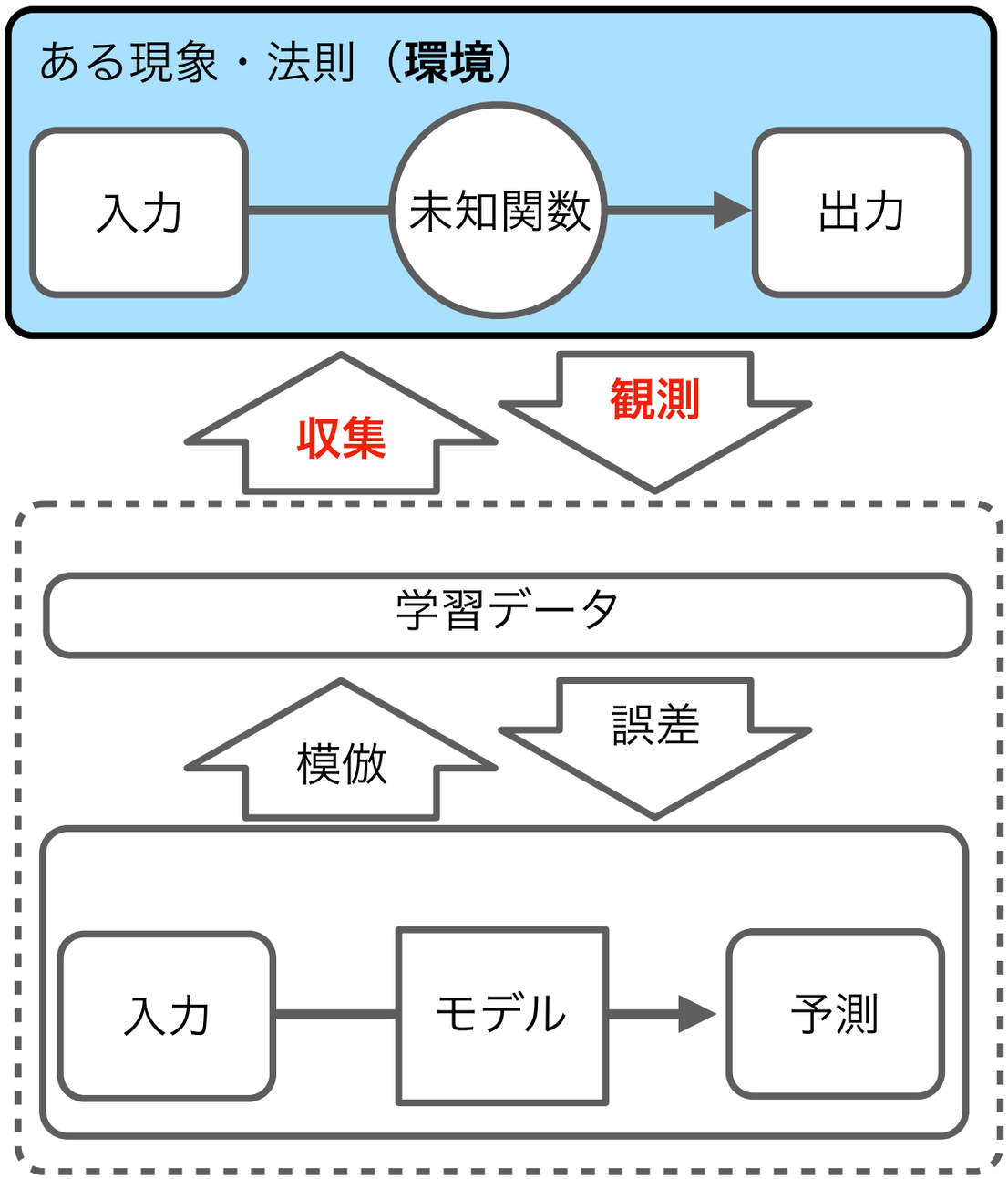
【強化学習】



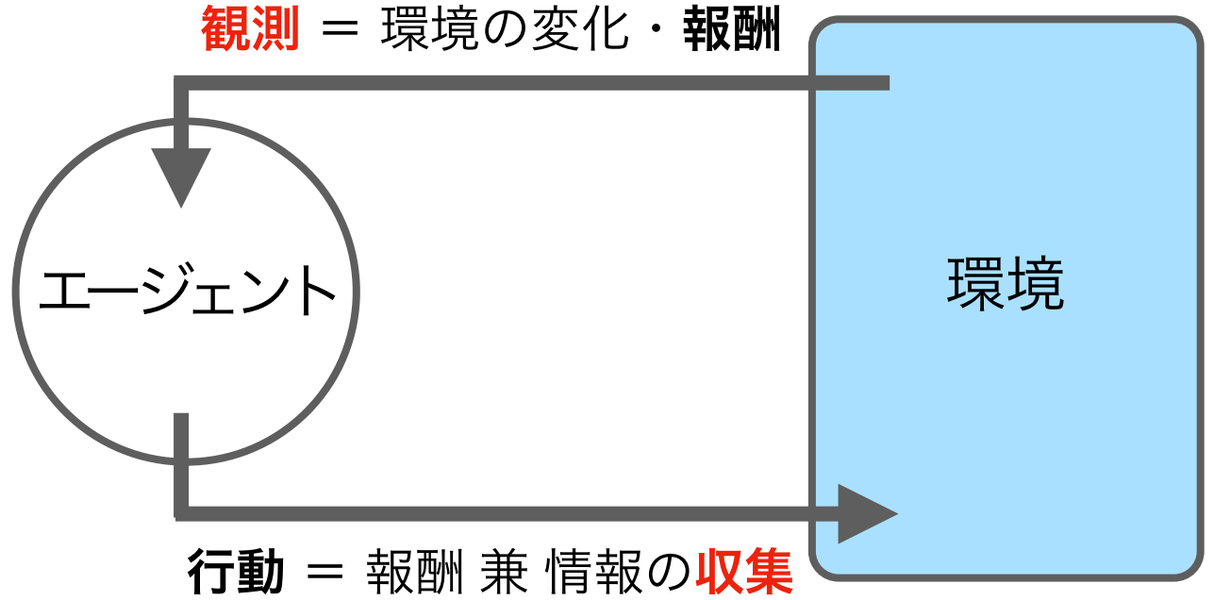
- **学習データ**から未知関数を学習モデルに模倣させるのが**教師あり学習**
- 環境 (母集団) から学習データの収集は教師あり学習では**与えられる前提**

強化学習とは

【教師あり学習】

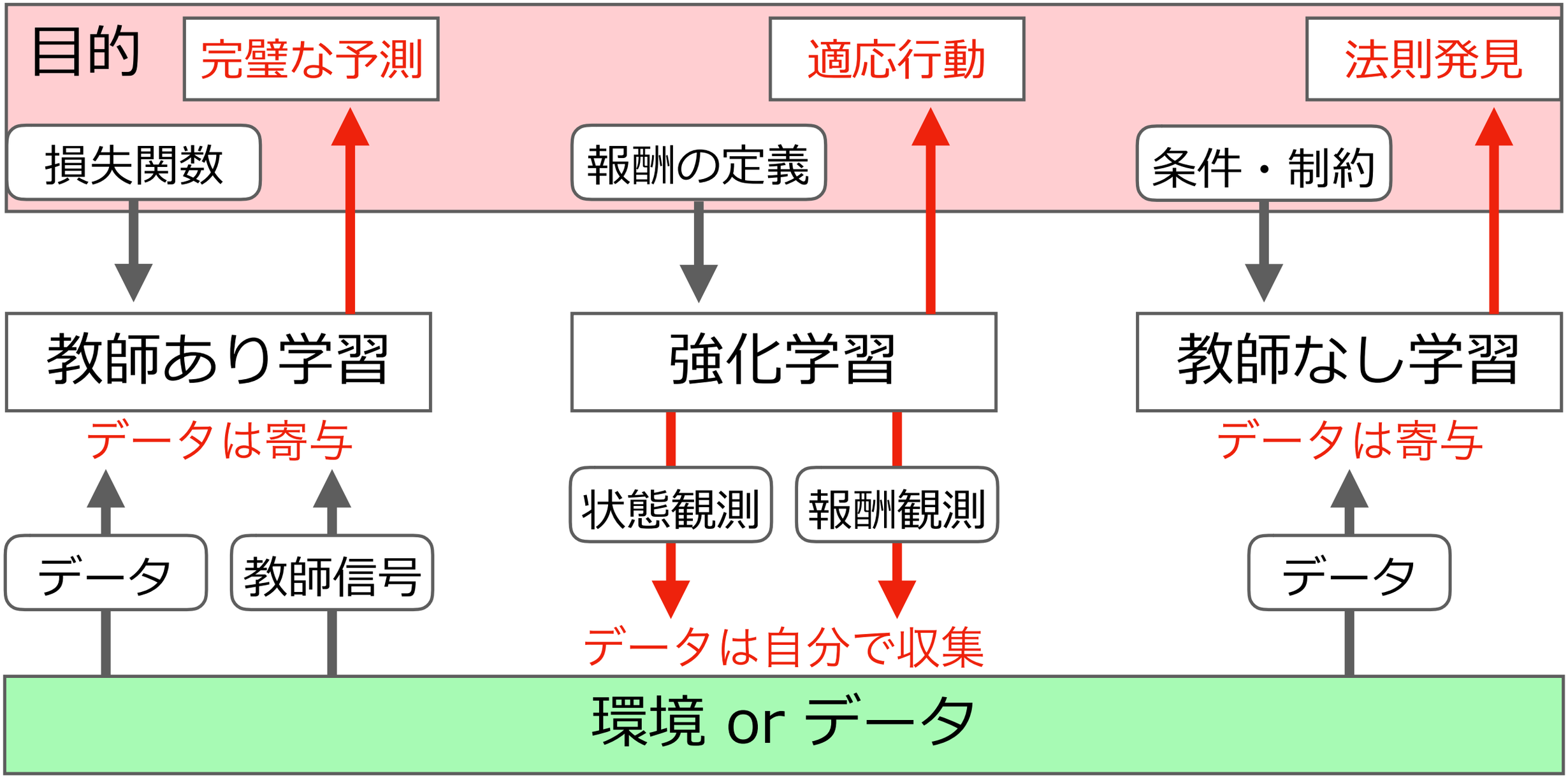


【強化学習】



- 環境からデータを環境から主体的に取得して来ることが**前提**
- 環境の遷移 (＝どんなデータが手に入るか) はエージェント自身の選択次第

機械学習



強化学習は“半教師あり学習”と呼ばれるがそれは間違い

環境を探索して主体的にデータを獲得しつつ最適化

単純な強化学習課題の例

【A/Bテスト】が有名

※ 2008年大統領選挙でオバマ陣営が宣伝サイトデザインで使用して40%ほどコンバージョン(広告成果)を向上させた

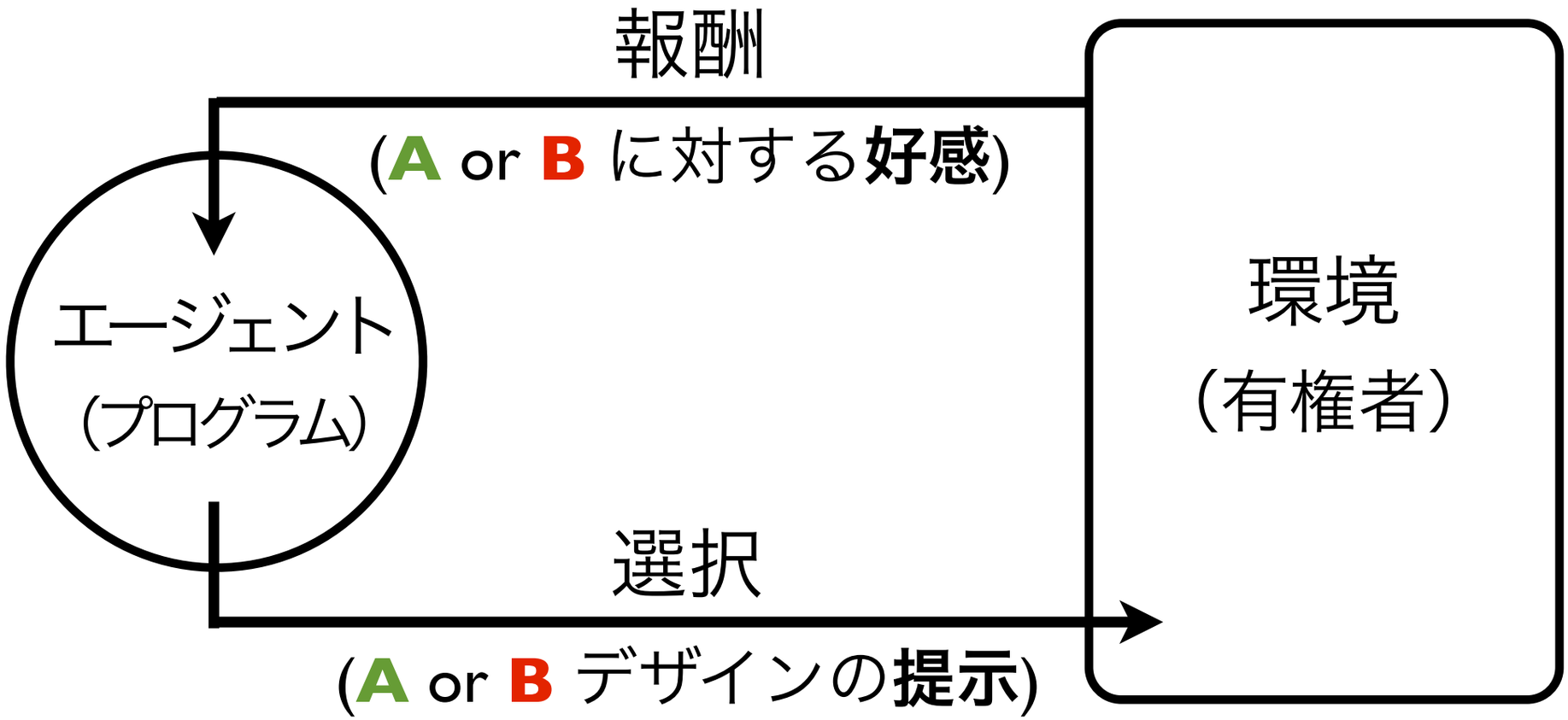
単純な強化学習課題の例

【A/Bテスト】が有名

逐次的に **A** or **B** のサイトデザインを選択させて

人間の好みのデザインに最適化させる

A/B テストと強化学習



【A/Bテスト】が有名

選択 ＝ A or B のサイトデザインの提示

報酬 ＝ 提示されたデザインへの好感

A/B テストと強化学習

【A/Bテスト】が有名

利点としては**デザインの改善**が

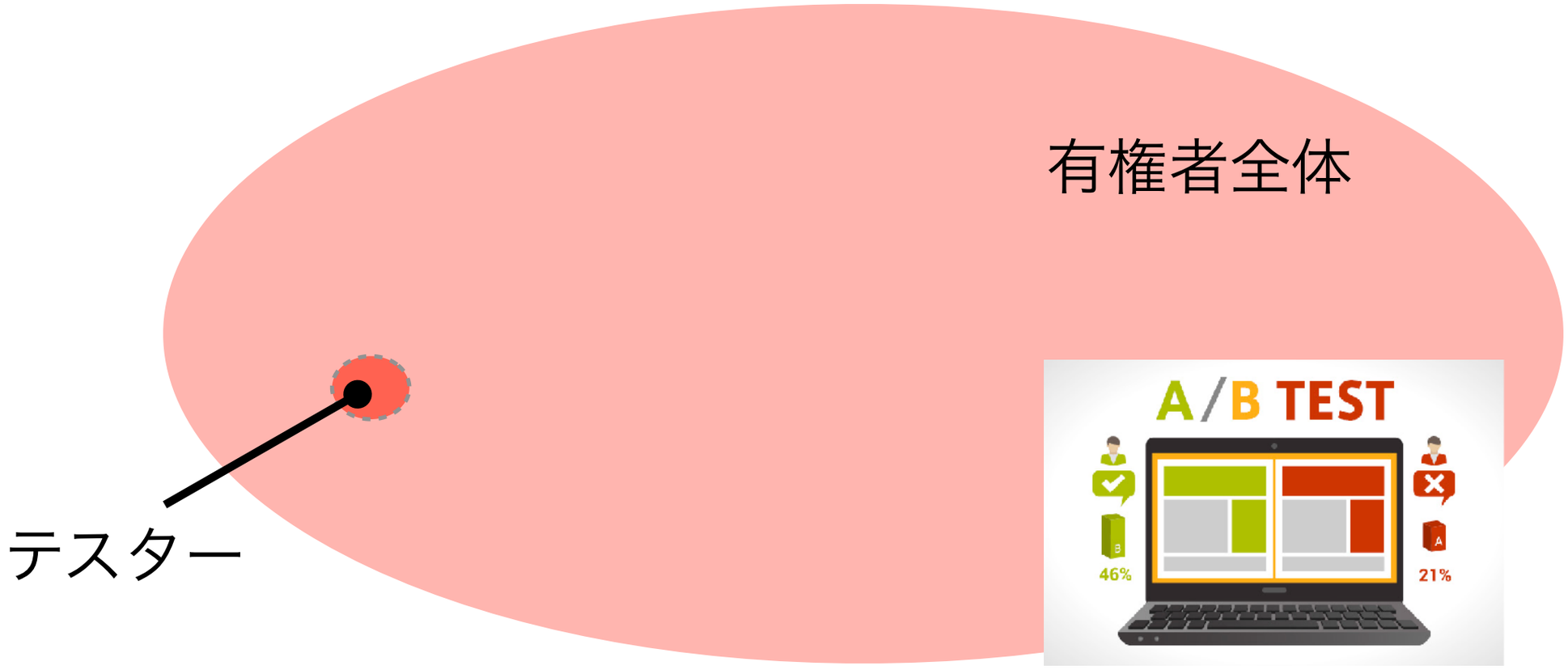
掲載期間中に**自律的**に行える事が挙げられる

A/B テストと強化学習

【A/Bテスト】が有名

有権者全体とテスターの好みの**偏り**を考慮しなくて良い

A/B テストと強化学習



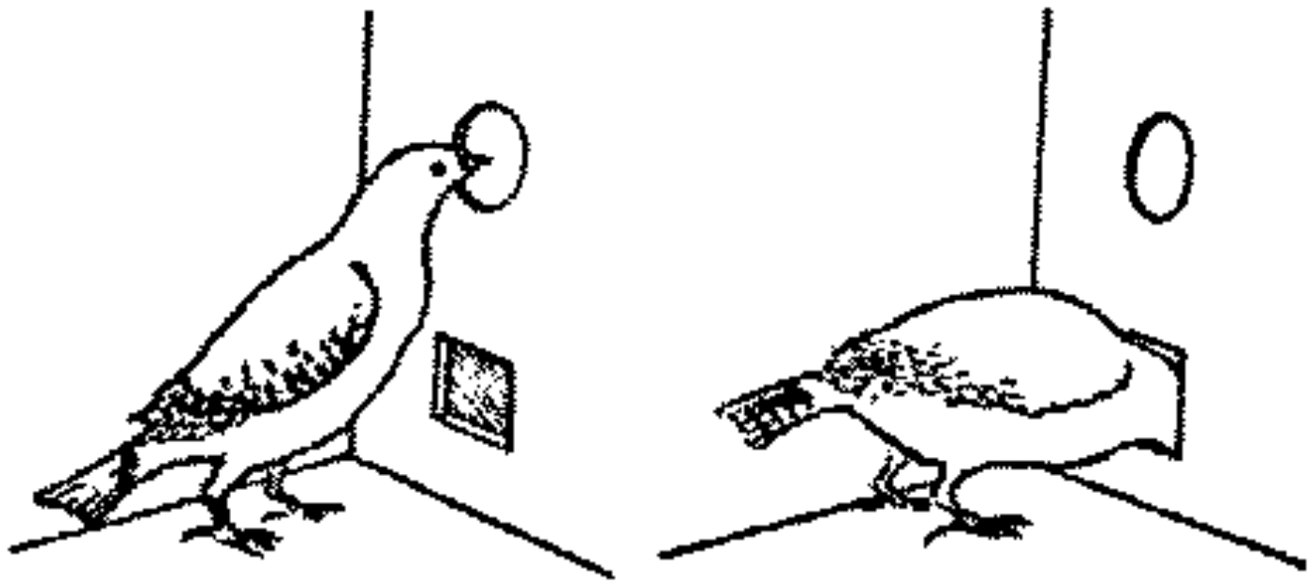
【A/Bテスト】が有名

有権者全体とテスターの好みの**偏り**を考慮しなくて良い

掲載期間中の**流行の変化**にも対応できる

強化学習と基本

スキナー箱等に代表される動物の行動学習が
アルゴリズムの原点のひとつ

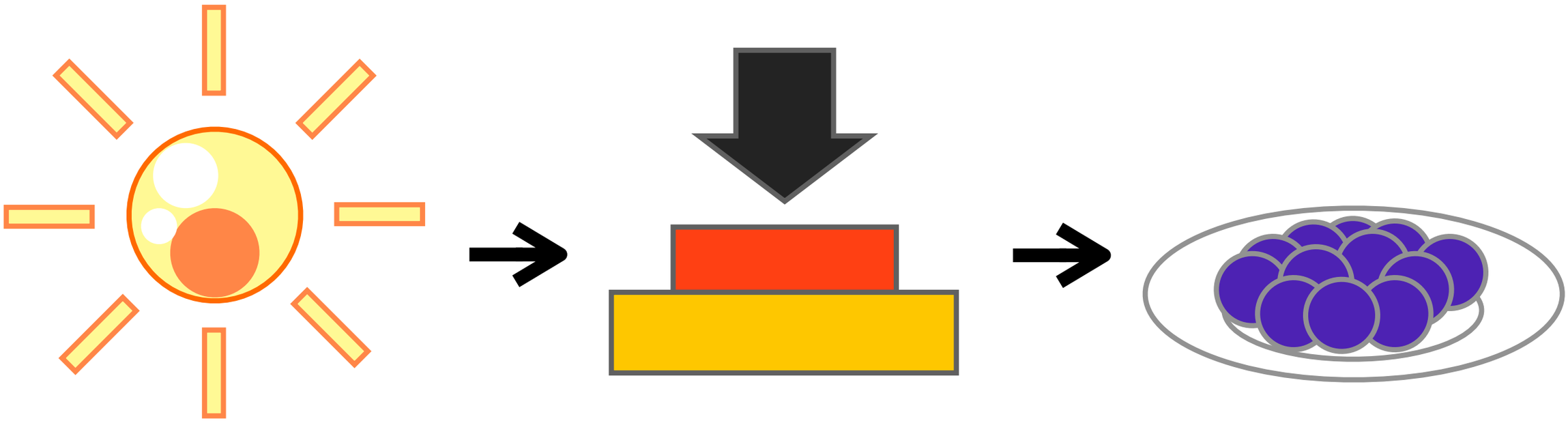


※ ボタンを押すと餌が出る

餌＝**強化子**(ABテストにおける有権者の評価)による
行動(特定のデザイン)の**強化**が基本

強化学習とは

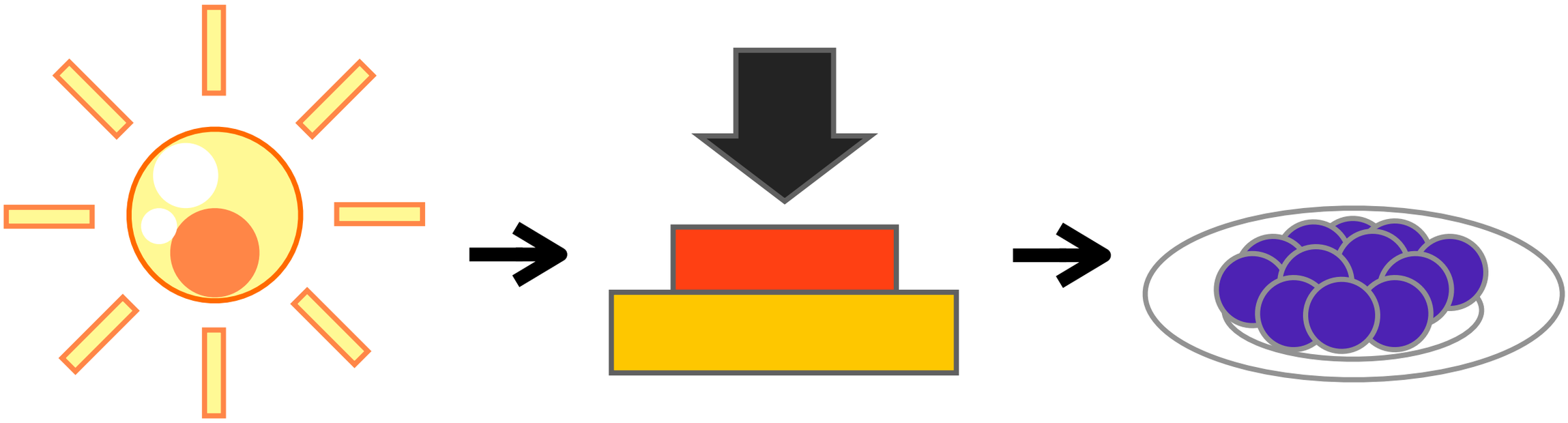
基本的には鳩の条件付け実験 ”スキナー箱” そのもの



状態：ランプ点灯　　行動：ボタン押下　　報酬：エサ獲得

強化学習とは

基本的には鳩の条件付け実験 ”スキナー箱” そのもの

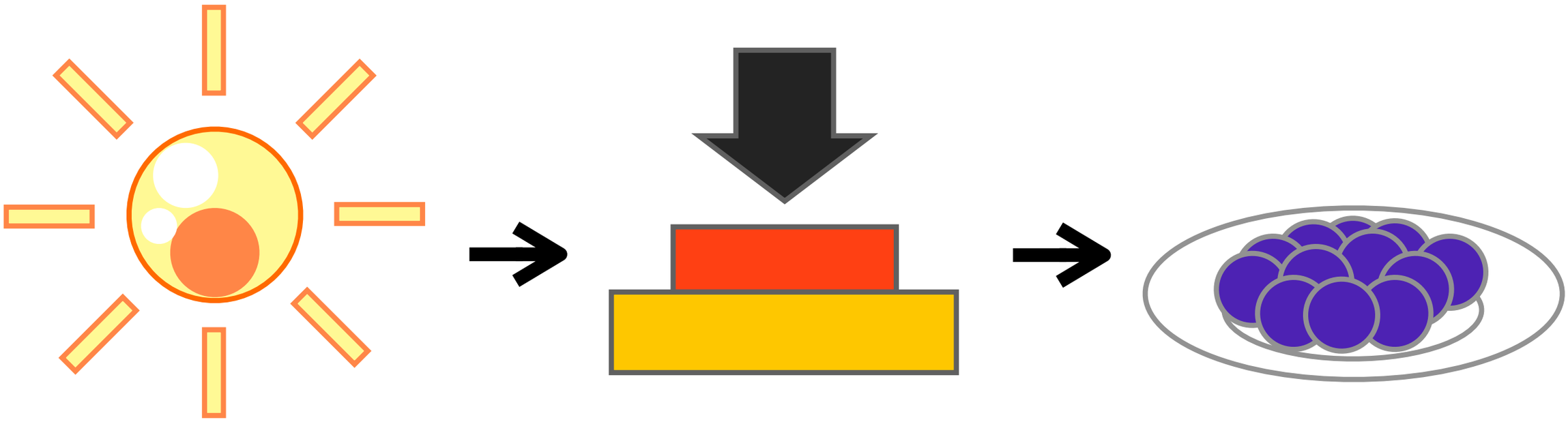


状態：ランプ点灯　　行動：ボタン押下　　報酬：エサ獲得

強化

強化学習とは

基本的には鳩の条件付け実験 ”スキナー箱” そのもの



状態：ランプ点灯　　行動：ボタン押下　　報酬：エサ獲得

強化

理屈は犬に芸を覚えさせるのとほぼ同じ (繰り返し)

→ 基本の理屈は単純

強化学習課題の例

【例】

・web広告のデザインや提示バナー選択

強化学習課題の例

【例】

・web広告のデザインや提示バナー選択

・新薬，新農法の臨床・実験

・リクエストへのサーバーのリソース割り当て

強化学習課題の例

【例】

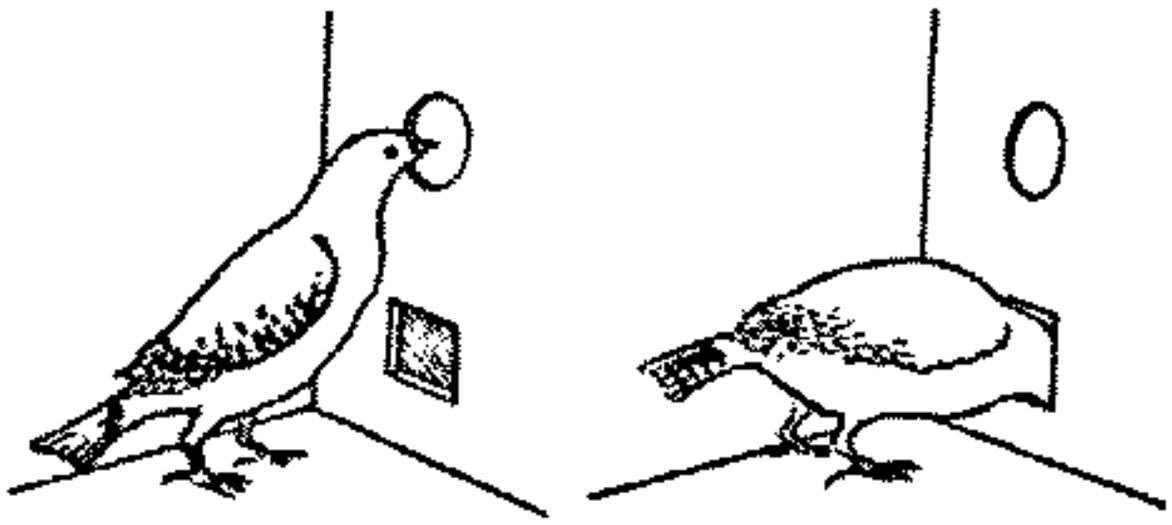
・web広告のデザインや提示バナー選択

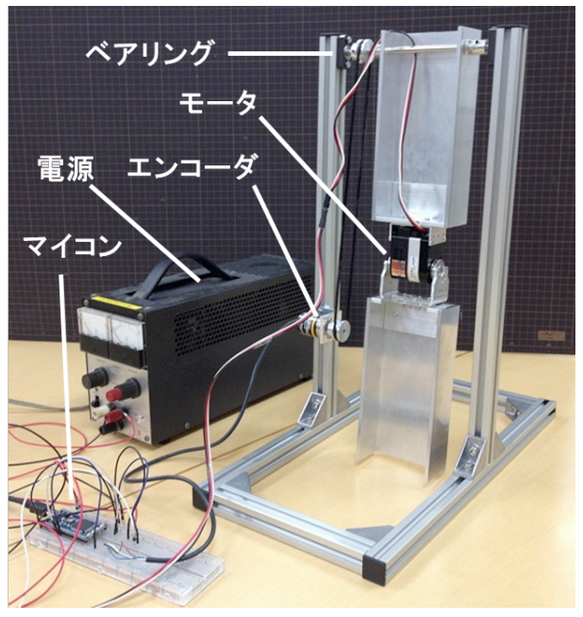
・新薬，新農法の臨床・実験

・リクエストへのサーバーのリソース割り当て

・動物の行動学習モデル(原点の一つ)

・ロボットの自律的な行動学習





強化学習課題の例

【例】

・web広告のデザインや提示バナー選択

・新薬，新農法の臨床・実験

・リクエストへのサーバーのリソース割り当て

・動物の行動学習モデル(原点の一つ)

・ロボットの自律的な行動学習

限られた**サンプリング数**での動的な**意思決定**に用いられる

強化学習課題の例

【例】

・web広告のデザインや提示バナー選択

・新薬，新農法の臨床実験

・リクエストへのサーバーのリソース割り当て

・動物の行動学習モデル(原点の一つ)

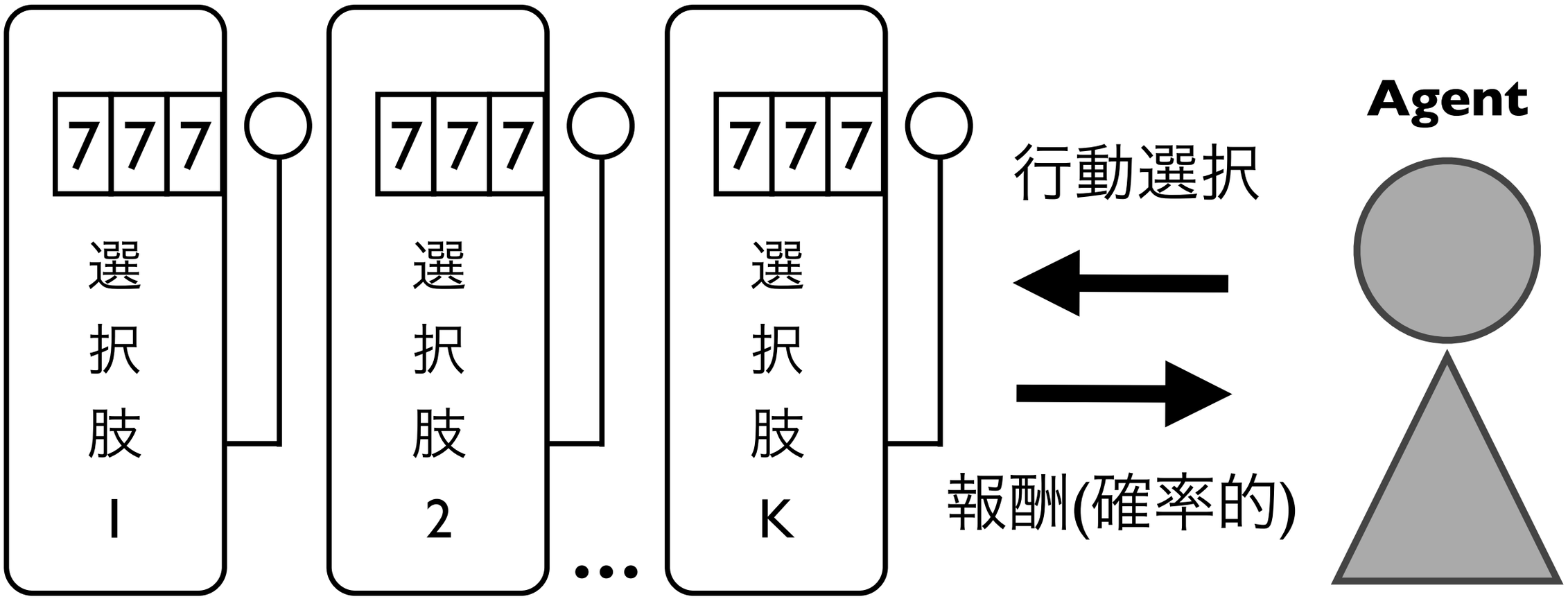
・ロボットの自律的な行動学習

限られた**サンプリング数**での動的な**意思決定**に用いられる

本授業ではこれらを**抽象化**した**意思決定課題**の一種である

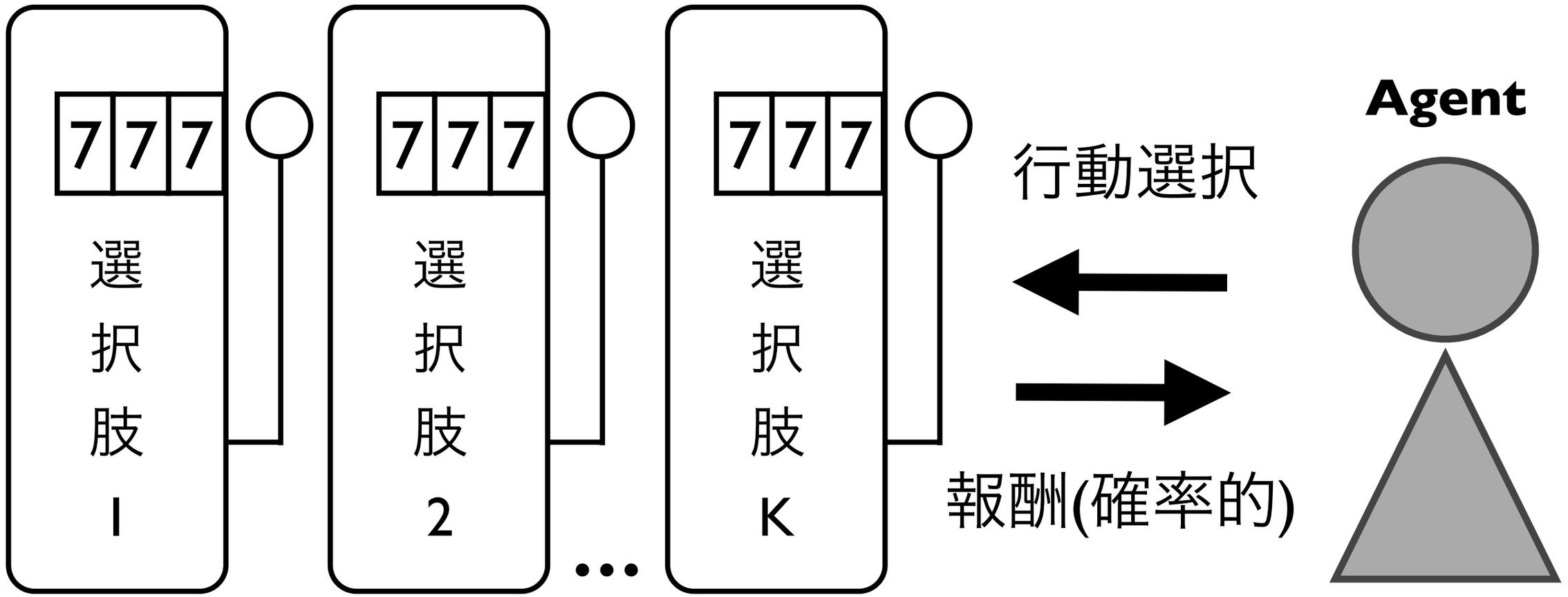
K本腕バンディット問題を扱う

多腕バンディット問題



エージェントにとって**未知**の**割合**で発生する**報酬**を
試行錯誤により**良い選択肢**を**なるべく早く**発見する事が目的

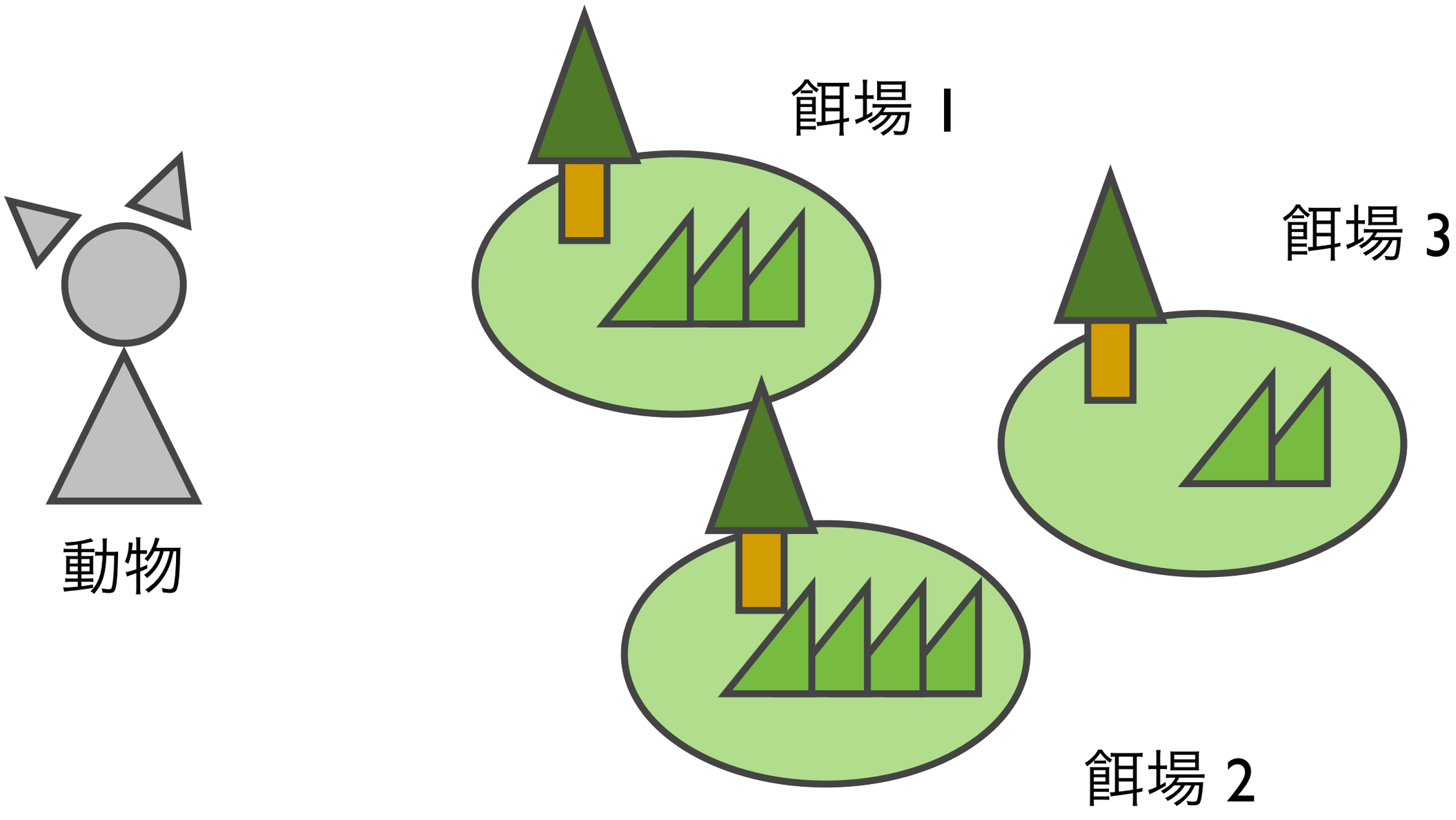
多腕バンディット問題



正直，強化学習の**ビジネス応用**としては**一番有効**

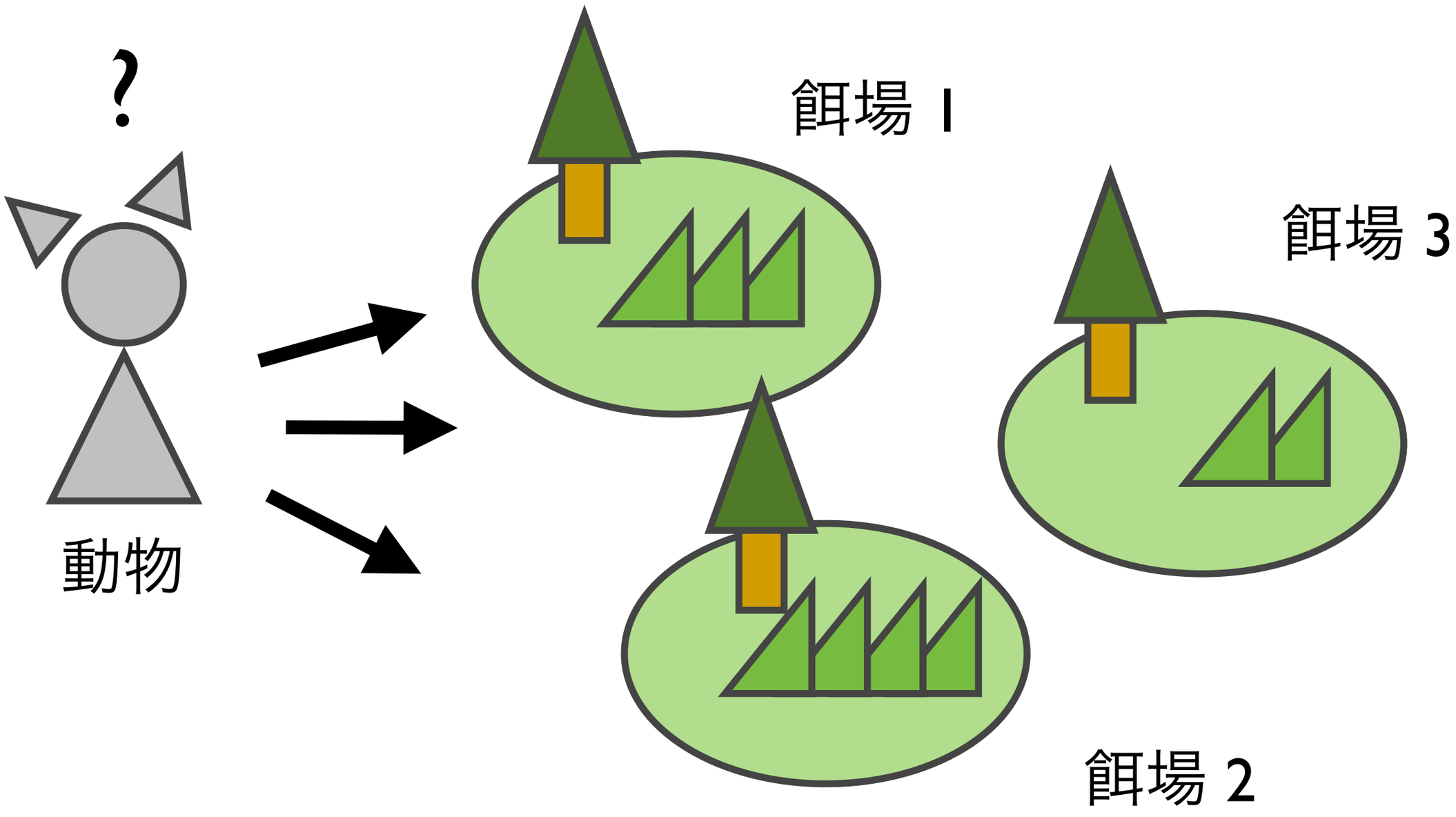
バンディット(問題の)アルゴリズムによる**広告配信**の改善

多腕バンディット問題



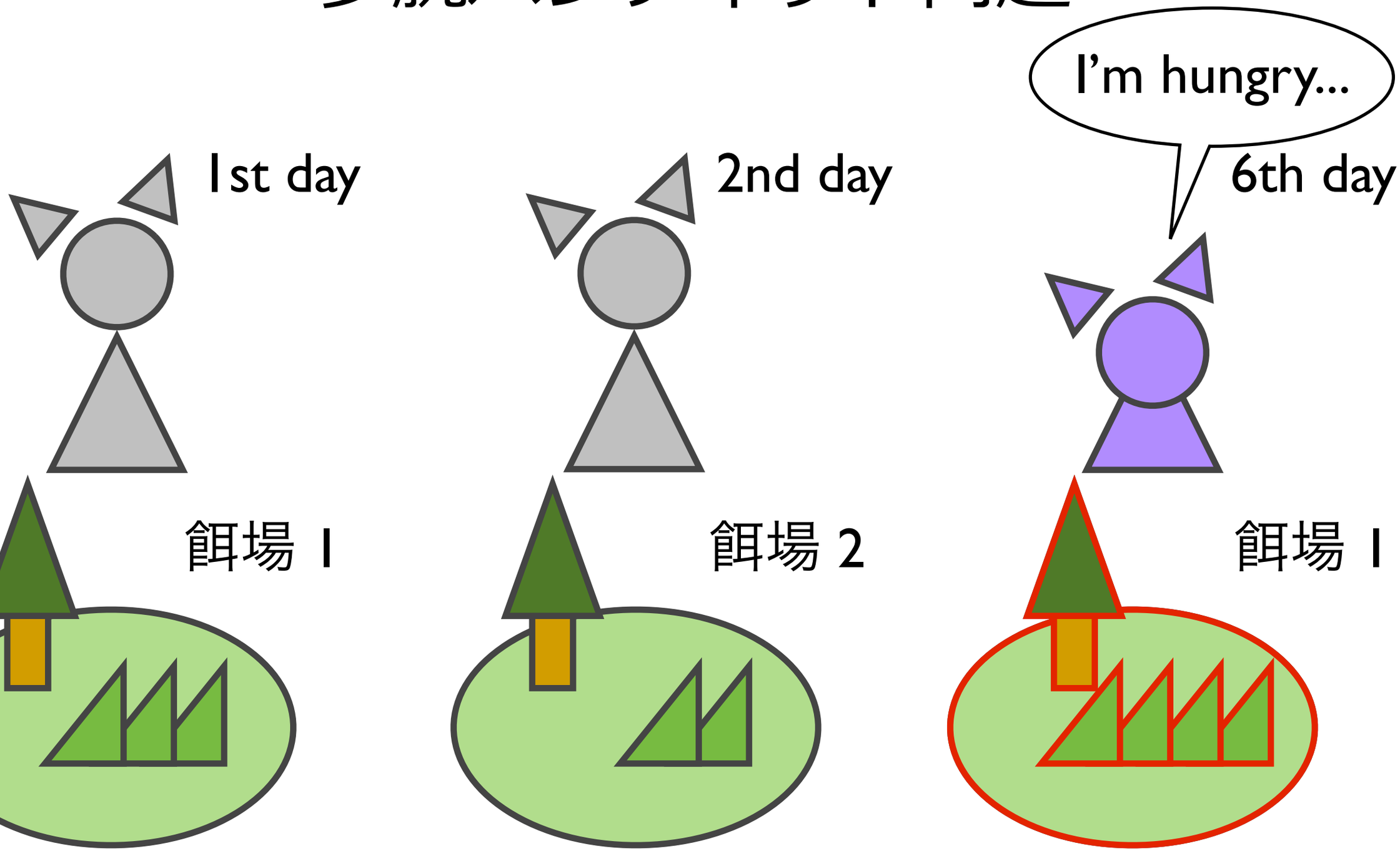
バンディット問題は動物の採餌行動に例えられる

多腕バンディット問題



未知の環境ではどこが最良の餌場かわからない

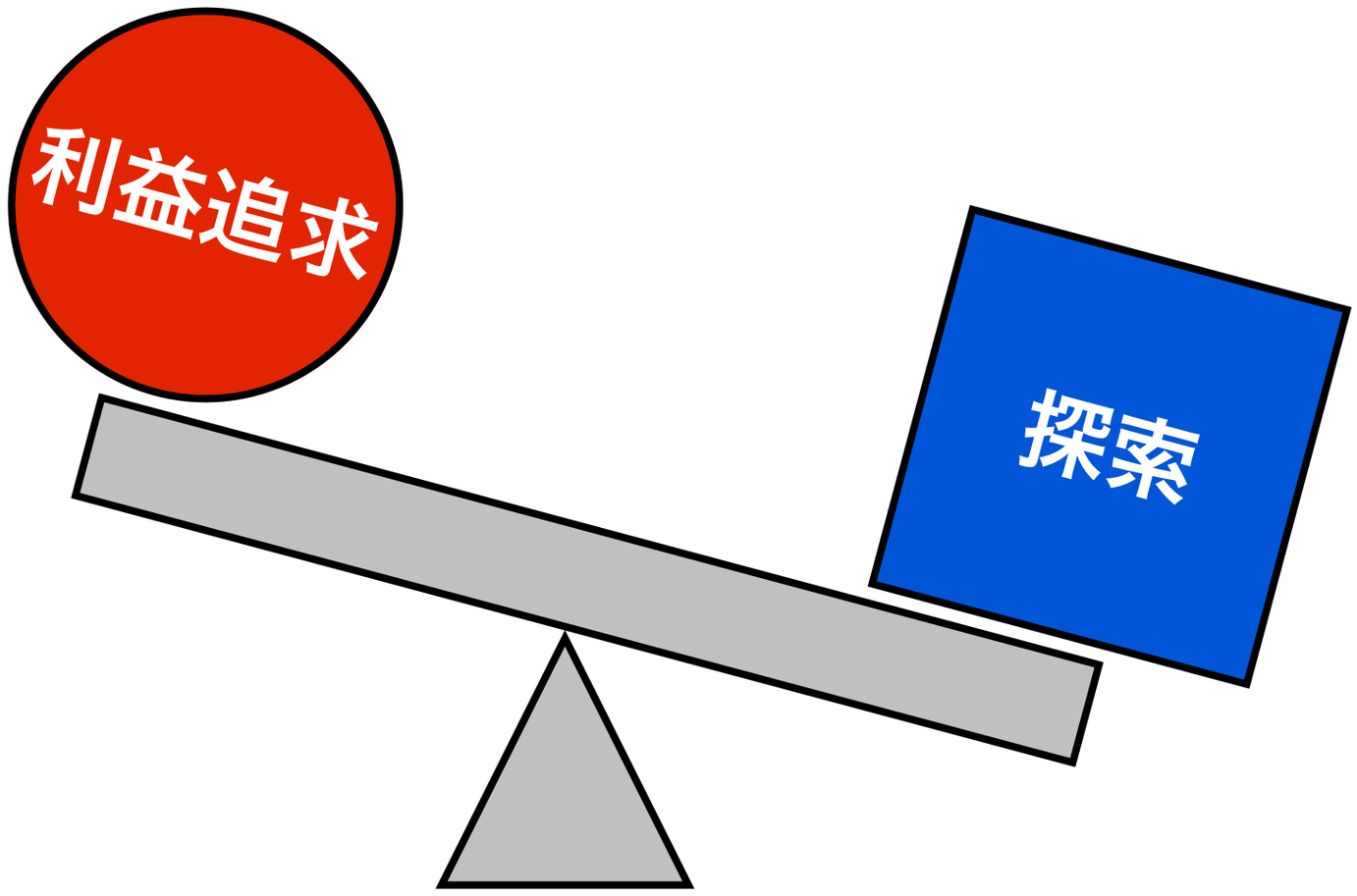
多腕バンディット問題



最良の餌場を見つけるためには**探索**をしなければならないが

探索をしすぎると獲得できる**餌**が**少ない**ため**飢えて**しまう

速さと正確さのトレードオフ



選択肢が**複数**ある**未知環境**下では**同時に**

情報収集のための探索と情報を利用した利益追求は両立不可

→ 限られた時間内でのバランシングが大切

速さと正確さのトレードオフ

教師あり学習：

強化学習　　：

強化学習は他の学習課題と異なり

【学習期間】と**【実践期間】**を分離せず行う

強化学習の**利点**であり**困難**さでもある

速さと正確さのトレードオフ

教師あり学習：

強化学習　　：

(研修はあるけど)社会人

強化学習は他の学習課題と異なり

【学習期間】と**【実践期間】**を分離せず行う

強化学習の**利点**であり**困難**さでもある

速さと正確さのトレードオフ

UCB1-tuned ： **特殊な価値関数で探索と利益追求を制御**

※ どれも強化学習の方策(戦略)

強化学習課題上は分離していないが**方策(戦略)**によって

【学習期間=探索】と**【実践期間=利益追求】**を分離する事も

→ 方策の評価の方法は？

速さと正確さのトレードオフ

バンディット問題の評価は**現在までの選択の報酬期待値**と
最良の選択肢を初めから選択し続けた期待値との**差**
で表される**後悔**の度合いが**低い**ほど良い

【後悔の度合い】

試行回数×最良の報酬確率−Σ(各選択肢の試行回数×報酬確率)

なるべく**早く正確に**最良の選択肢を見つけ出す必要がある

速さと正確さのトレードオフ

実際に対応すべき**環境は非定常**である場合が多く

探索しても**過去のサンプル**が無意味になる可能性がある

柔軟に**再探索**をして**環境の変化**に追従するべき

探索アルゴリズム

【 ε -greedy】

確率 ε でランダム探索をして,

確率 1-ε でもっとも高い観測期待値 (獲得報酬の平均値) が高い選択肢を選択する

探索確率 ε をステップ数 O(log t) で減少させる

最も基本的な探索アルゴリズム

探索アルゴリズム

【softmax】

exp (Ei/τ) / Σ exp (Ek/τ)

観測期待値 (Ei) の大きさに応じて選択確率を決める

温度パラメータ τ が高いほどランダム性が高くなる

パラメータ τ を減少させていくがチューニングが難しい

探索アルゴリズム

【KL-UCB】

※ベルヌーイバンディット用

後に紹介する UCB 系アルゴリズムで
パラメータを KL-divergence の逆関数から求めたもの

(数年前の知識として)
最も低い後悔の度合いが保証されている

探索アルゴリズム

【Thompson Sampling】

※ベルヌーイバンディット用

ベータ分布のハイパーパラメータとして報酬の獲得/不獲得
を転用して評価値をサンプリングする

(数年前の知識として)

最も低い後悔の度合いが保証されている

非定常環境

A/Bテストで扱っている**人間**とは**非定常**なもの
エージェントが完全に変化を捉えきれるとは**限らない**

※ 人間との協力作業は常に非定常

今後の人間がいる環境での**相互作用**を考慮すると
非定常環境への対応は**不可欠**である

非定常環境での意思決定

環境が変化したなら**過去の記憶は足枷**になる

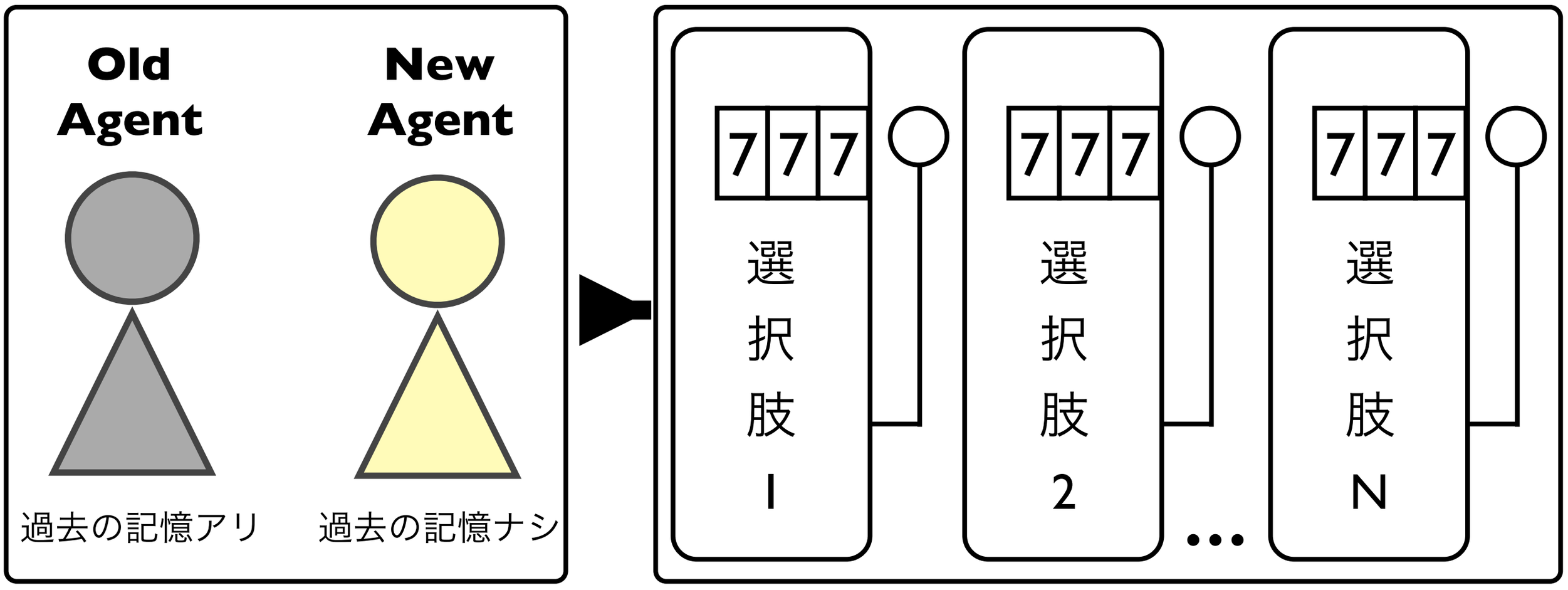
→ 過去の記憶は消したほうが良い

メタバンディットアルゴリズム

非定常環境に対応するために考案されたアルゴリズム

記憶を初期化したエージェントと従来のエージェントの間で

一定期間**二択**のバンディット課題を行う



エージェントを選択　してから　**選択肢**を選択する

メタバンディットアルゴリズム

選択肢の価値関数には UCB 系アルゴリズムを用いる

UCB1-tuned :

$$P(E|A_i) + \sqrt{\frac{\ln n}{n_i} min\left(1/4, V_i(n_i)\right)}$$

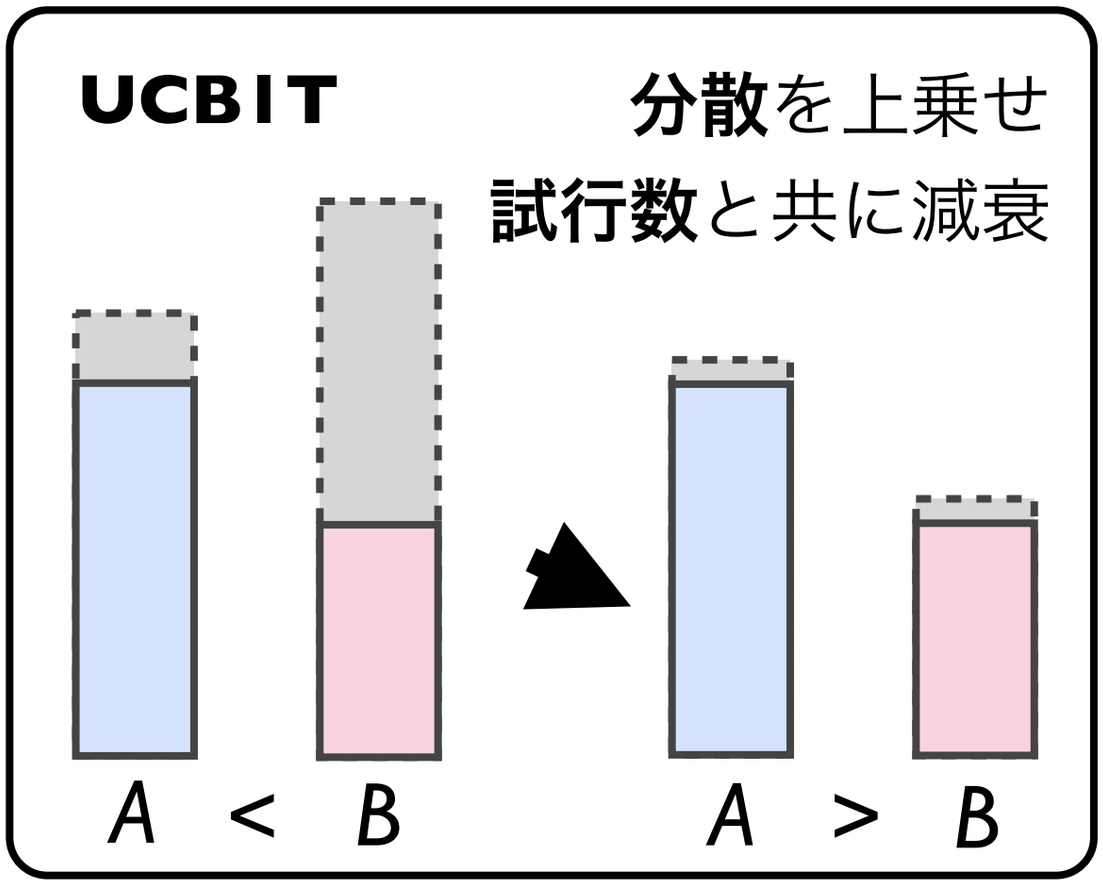
$$V_i(s) = \left(\frac{1}{n_i}\sum_{k=1}^{n_i} r_{k,i}\right) - P(E|A_i) + \sqrt{\frac{\ln n}{n_i}}$$

統計的な背景から導出された評価関数

理想的な選択との損失(後悔の度合い)の上限を保証

メタバンディットアルゴリズム

選択肢の価値関数には UCB 系アルゴリズムを用いる



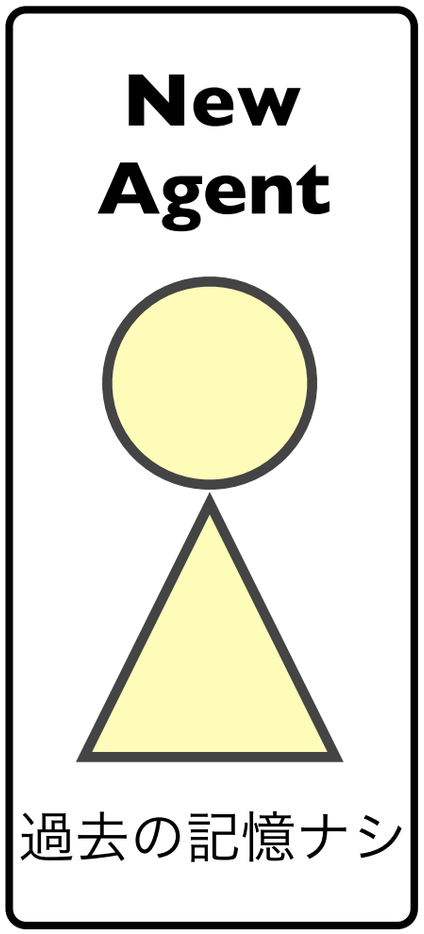
統計的な背景から導出された評価関数

理想的な選択との損失(後悔の度合い)の上限を保証

メタバンディットの強み

環境が変化したなら**過去の記憶は足枷**になる

→ 過去の記憶は消したほうが良い



メタバンディットの強み

環境が変化したなら**過去の記憶は足枷**になる

→ 過去の記憶は消したほうが良い

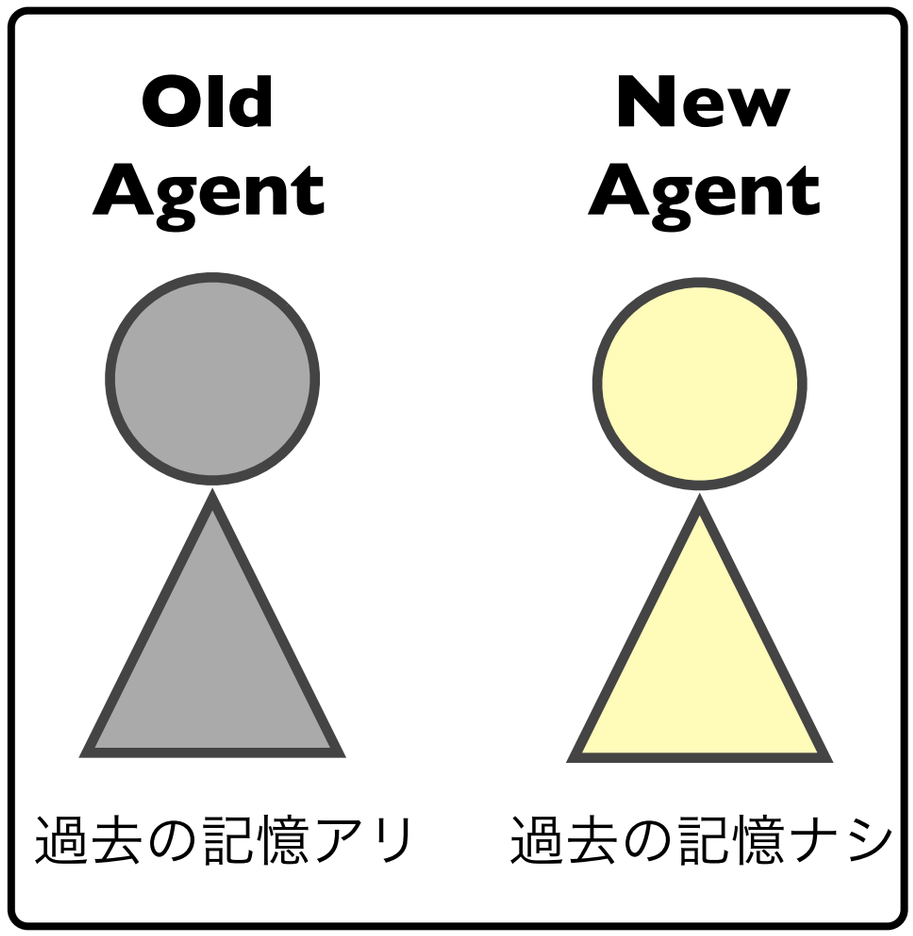
しかし環境の変化の検出が**誤りであったら**無駄

また大きな変化ではないのに記憶を**全て消す必要は無い**

メタバンディットの強み

記憶を **消す**／**残す** かを一定期間比較する事で

誤検出による早まった忘却と**不必要な忘却**を回避する



※記憶を **消した**／**残した** エージェントで一定期間比較してから選択

メタバンディットの強み

1. 現時点でベストな選択肢の報酬確率が下降した場合

のみ

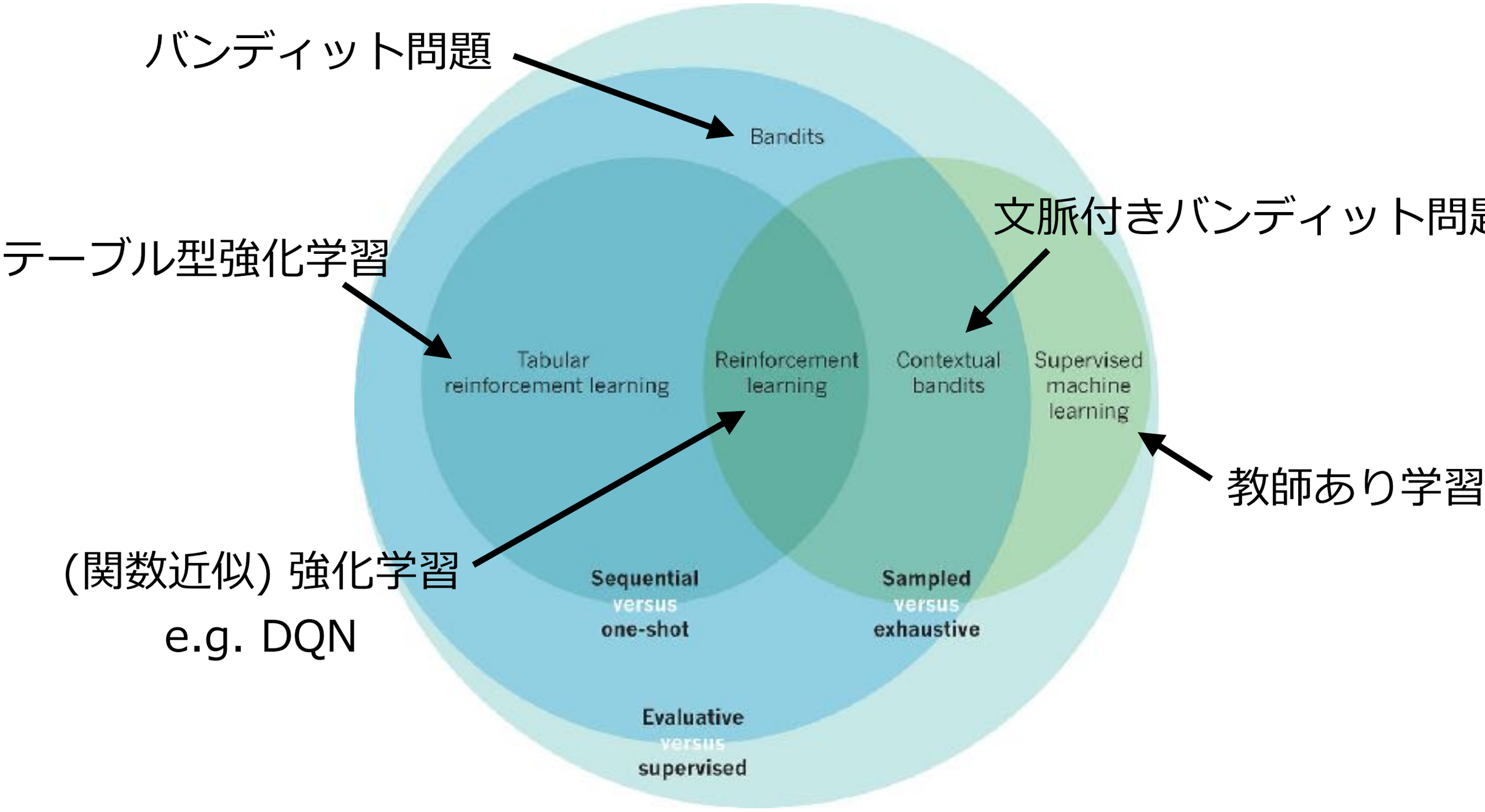
非定常環境に対処できる

2. 非ベストな選択肢の報酬確率が上昇した場合
3. 下降と上昇が同時に起こった場合

には**対処できない**

現時点で最適化は不可能

強化学習の区分

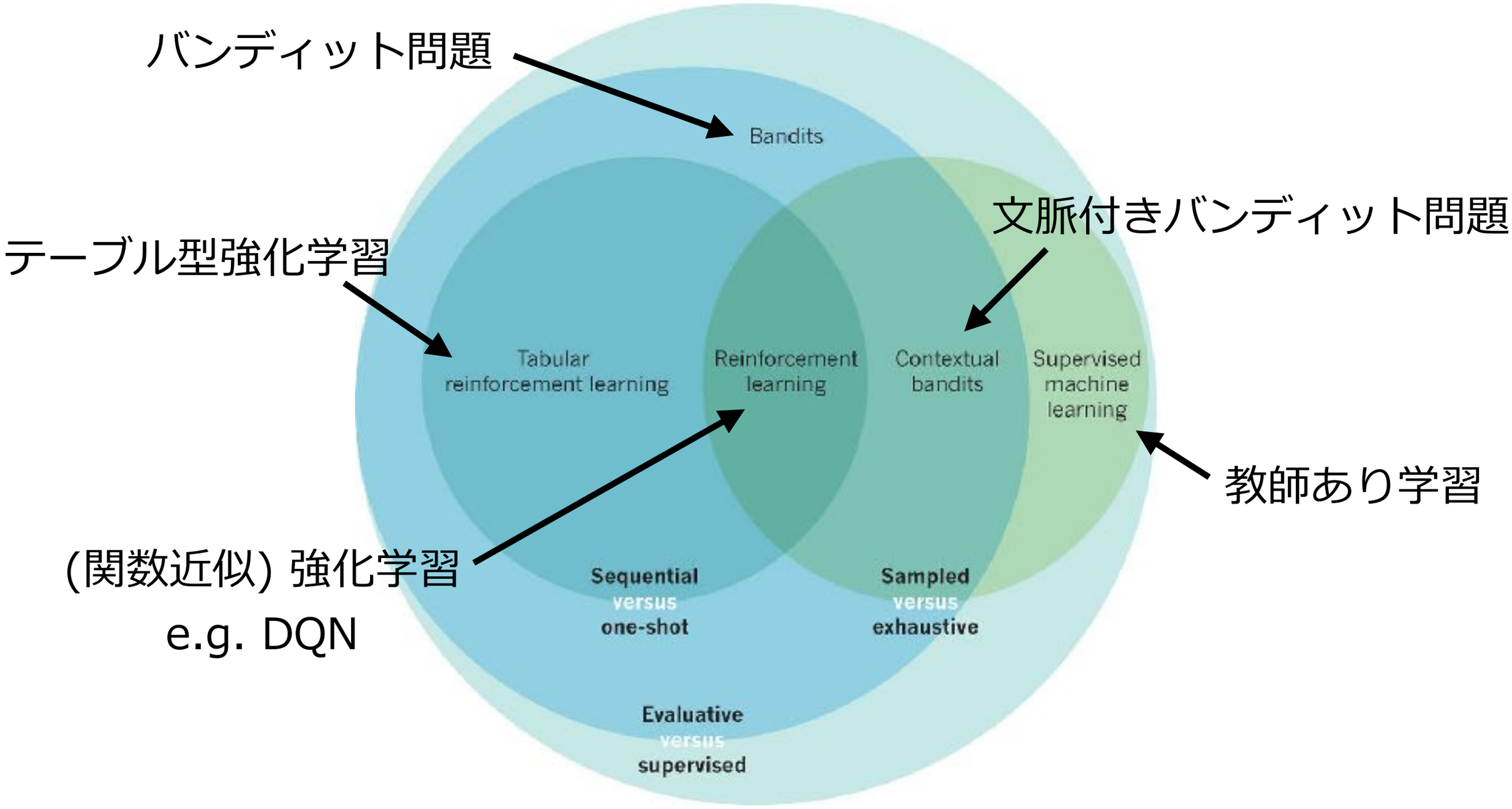


実用向き　　：バンディット問題・文脈付きバンディット問題

応用事例は限定的　：テーブル型強化学習

高度な応用に向けて：(関数近似) 強化学習

強化学習の区分



次回: バンディット問題の実装課題

次々回: 高度な強化学習概要

今回のまとめ

強化学習は環境からのデータ収集を前提とした分野

↓

試行錯誤配分と報酬の最大化のバランスが困難

↓

確率論の知識により最適化可能

↓

非定常環境 (現実環境) を想定すると更に困難