

マルチステップ強化学習の話

～Q学習、Alpha-Go、モンテカルロ木探索などを一般化しよう～

東京大学 北村俊徳

北村 俊徳 (Toshinori Kitamura)



@syuntoku14

所属

東大松尾研究室 (D1)

研究内容

強化学習の理論 (と応用)

インターン:

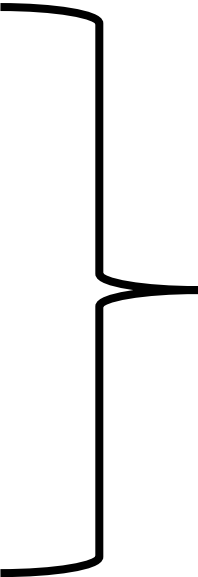
MiraRobotics, AIST,
OMRON SINIC X, Integral AI



今日の内容

色々な逐次意思決定アルゴリズムを一般化する枠組みについて紹介

- Q学習
- モデル予測制御
- A^*
- Alpha-Go
- モンテカルロ木探索
- ...



実は全部
「マルチステップ強化学習」で
一般化できます！

論文：[Beyond the One Step Greedy Approach in Reinforcement Learning](#)

もっと数式スライド：https://github.com/syuntoku14/Shumi-Note/blob/main/notebooks/Multi_step_RL.pdf

こんな人にとっては面白い or 役に立つかも

- 強化学習をさわったことがある
- これから強化学習を勉強する
- 次のアルゴリズムの名前を二つ以上聞いたことがある
 - Q学習
 - モデル予測制御
 - A*
 - Alpha-Go
 - モンテカルロ木探索
 - ...

目次

1. 逐次意思決定問題って？迷路の例

2. ヒントを使ったアルゴリズム

- ・ ヒントの使用頻度とトレードオフ

3. ヒントの学習

- ・ ヒントの学習頻度とトレードオフ

逐次意思決定問題とは

行動を選択→次の状態に遷移→行動を選択→...を繰り返す環境の
最適な行動の選択ルールを考える問題（身の回りにもたくさん！）

将棋・囲碁



行動：駒を動かす
状態の遷移：盤面の変化

Alpha-Goとか

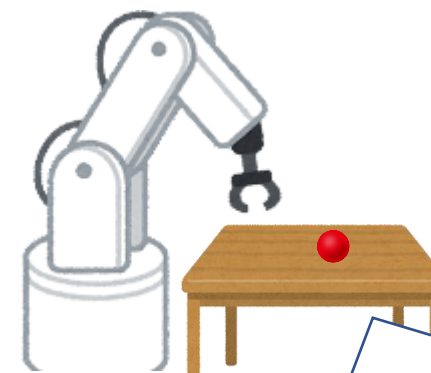
迷路



行動：進む方向を決める
状態の遷移：隣のマスに移動

A*とか

ロボット



行動：関節を動かすトルクを決める
状態の遷移：ロボットの関節角が変化

モデル予測制御とか

逐次意思決定問題とは

行動を選択→次の状態に遷移→行動を選択→...を繰り返す環境の
最適な行動の選択ルールを考える問題（身の回りにもたくさん！）

将棋・囲碁



行動：駒を動かす
状態の遷移：盤面の変化

迷路



行動：進む方向を決める
状態の遷移：隣のマスへ移動

ロボット

逐次意思決定問題を
解くアルゴリズムは実は全部似ている

体験してみよう！（次ページから）

行動：関節を動かすトルクを決める
状態の遷移：ロボットの関節角が変化

Alpha-Goとか

A*とか

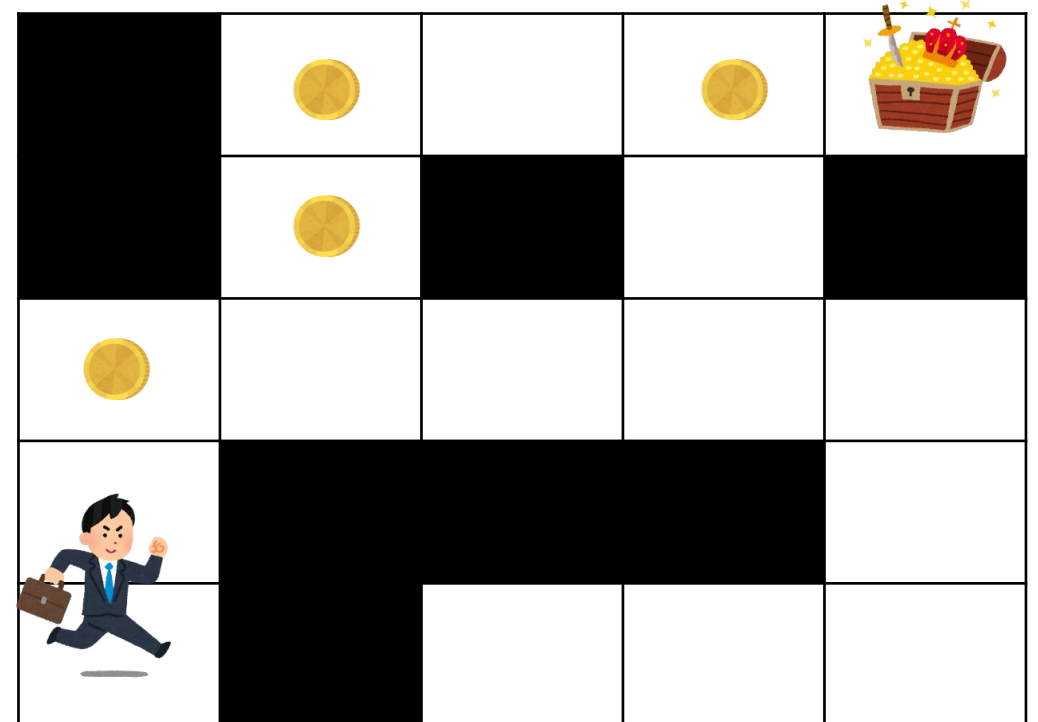
モデル予測制御とか

迷路を使ってアルゴリズムを体験しよう

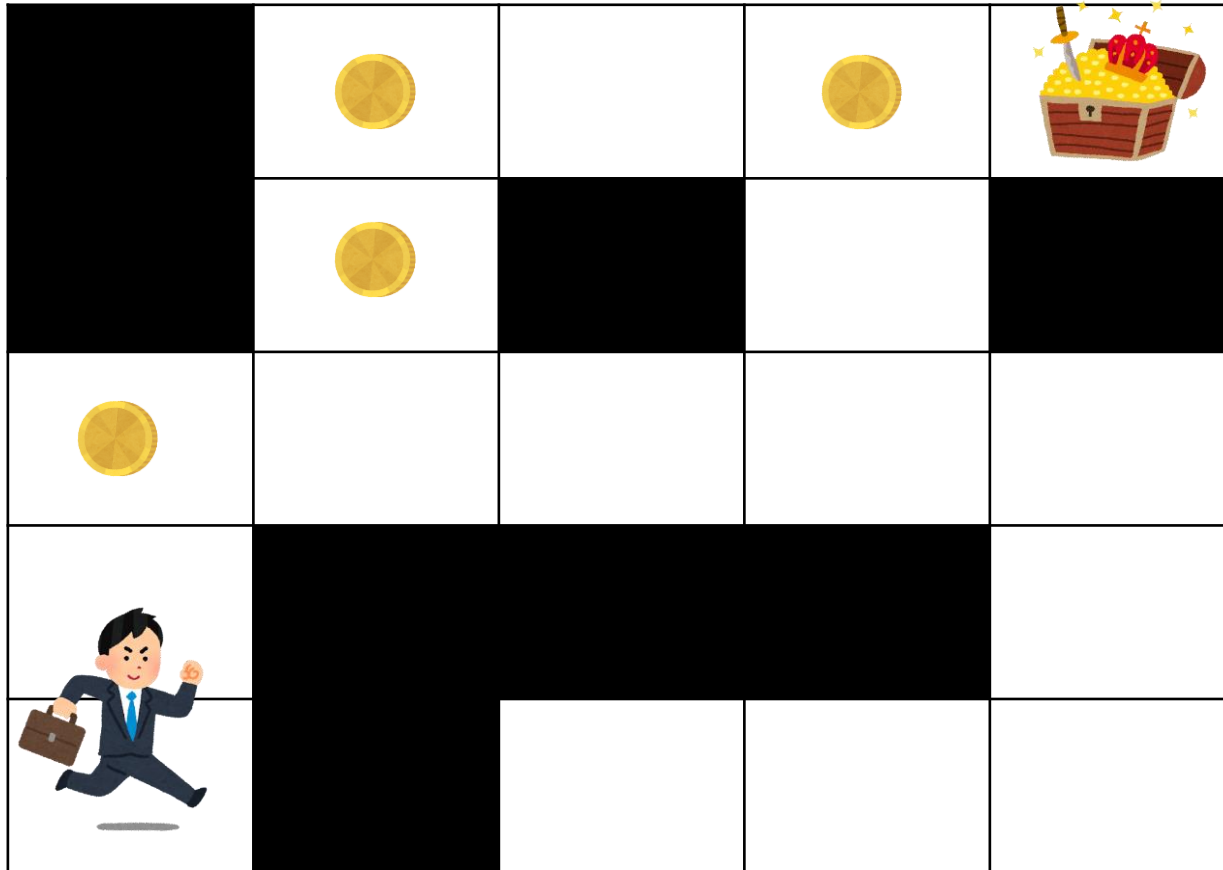
- 次のスライドから迷路を見ていきます
- 🪙を一番多く拾い、🏆までの最短経路を頭の中で描いてみてください

迷路の中身：

- 移動のたびマイナス 🪙
- 黒マスは壁
- 白マスは歩道
- 来た道は戻れない



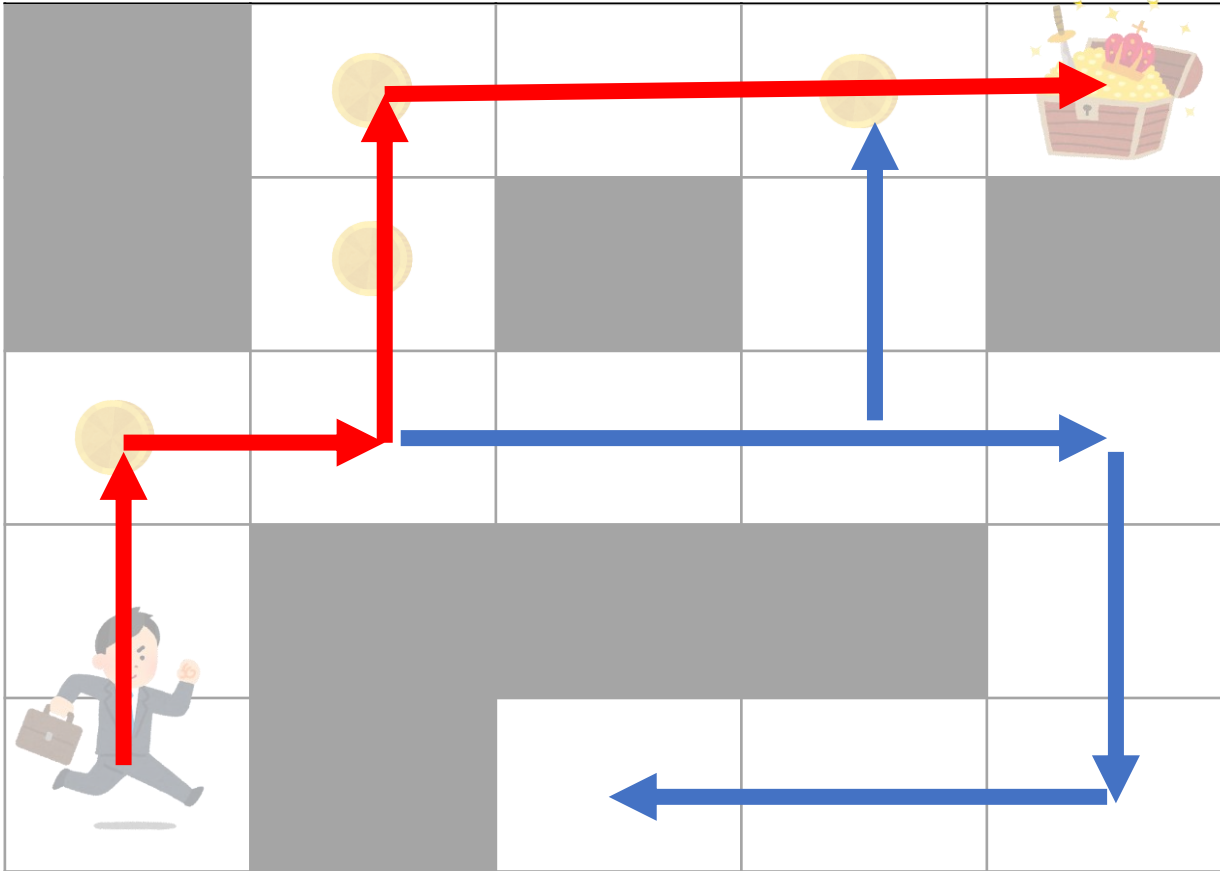
全部の選択肢を列挙しよう 1



●が一番拾える
最短経路は？

- 移動のたびマイナス ●
- 黒マスは壁
- 白マスは歩道

全部の選択肢を列挙しよう 1

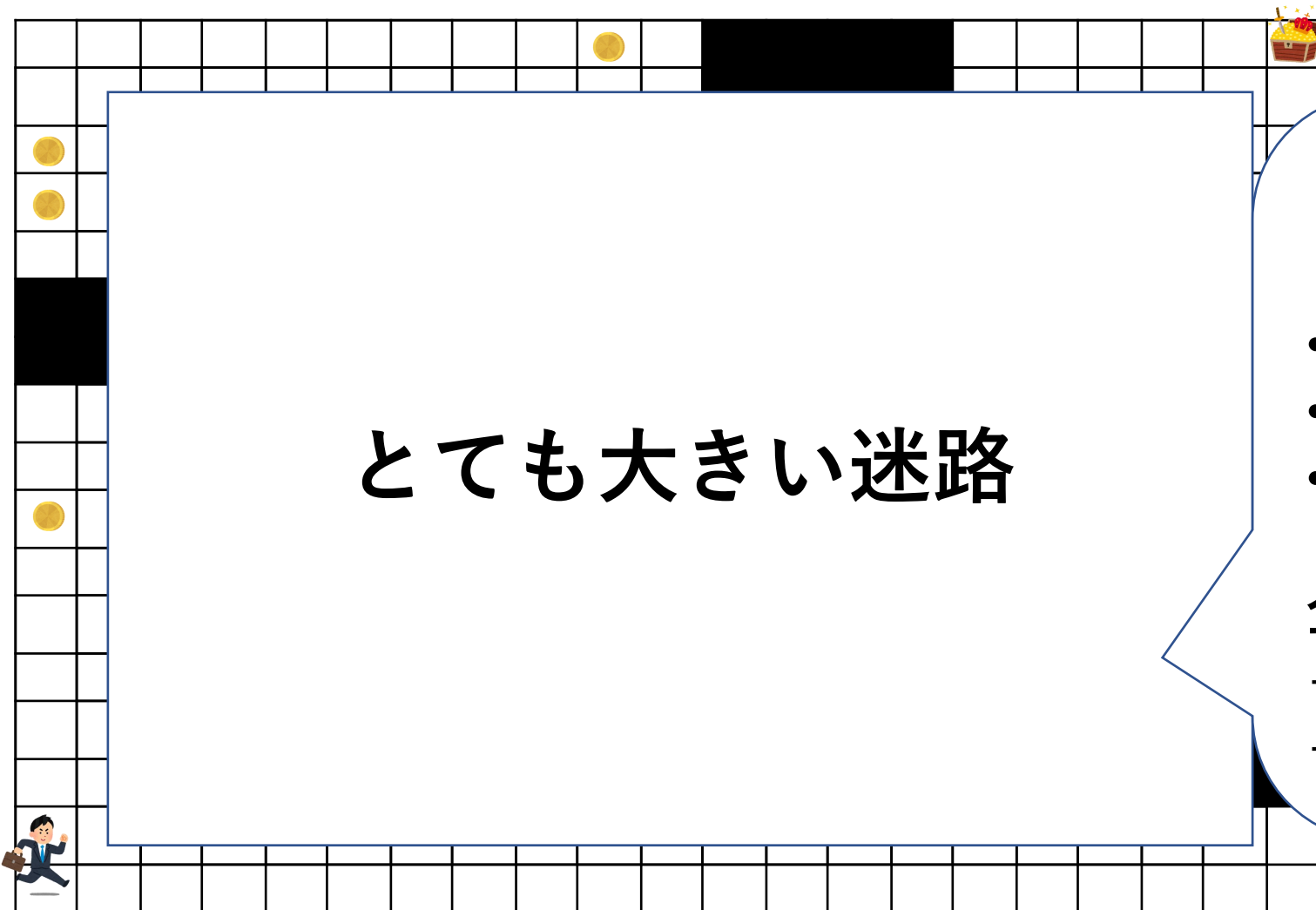


「すぐ解ける迷路」

- 小さい
- 次の遷移先が確定
- ゴール場所が既知

→ 全列挙で十分

難しい迷路 | 大きい迷路



「大きい迷路」

- 大きい
- 次の遷移先が確定
- ゴールの場所が既知

全列挙は大変

→ 効率よく解くためには？

→ A*アルゴリズムとか

目次

1. 逐次意思決定問題って？迷路の例

2. ヒントの使い方

- ・ ヒントの使用頻度とトレードオフ

3. ヒントの学習

- ・ ヒントの学習頻度とトレードオフ

追加ルール：ヒントのおじさん

これから先は
● 15個くらい拾ってから
ゴールだよ

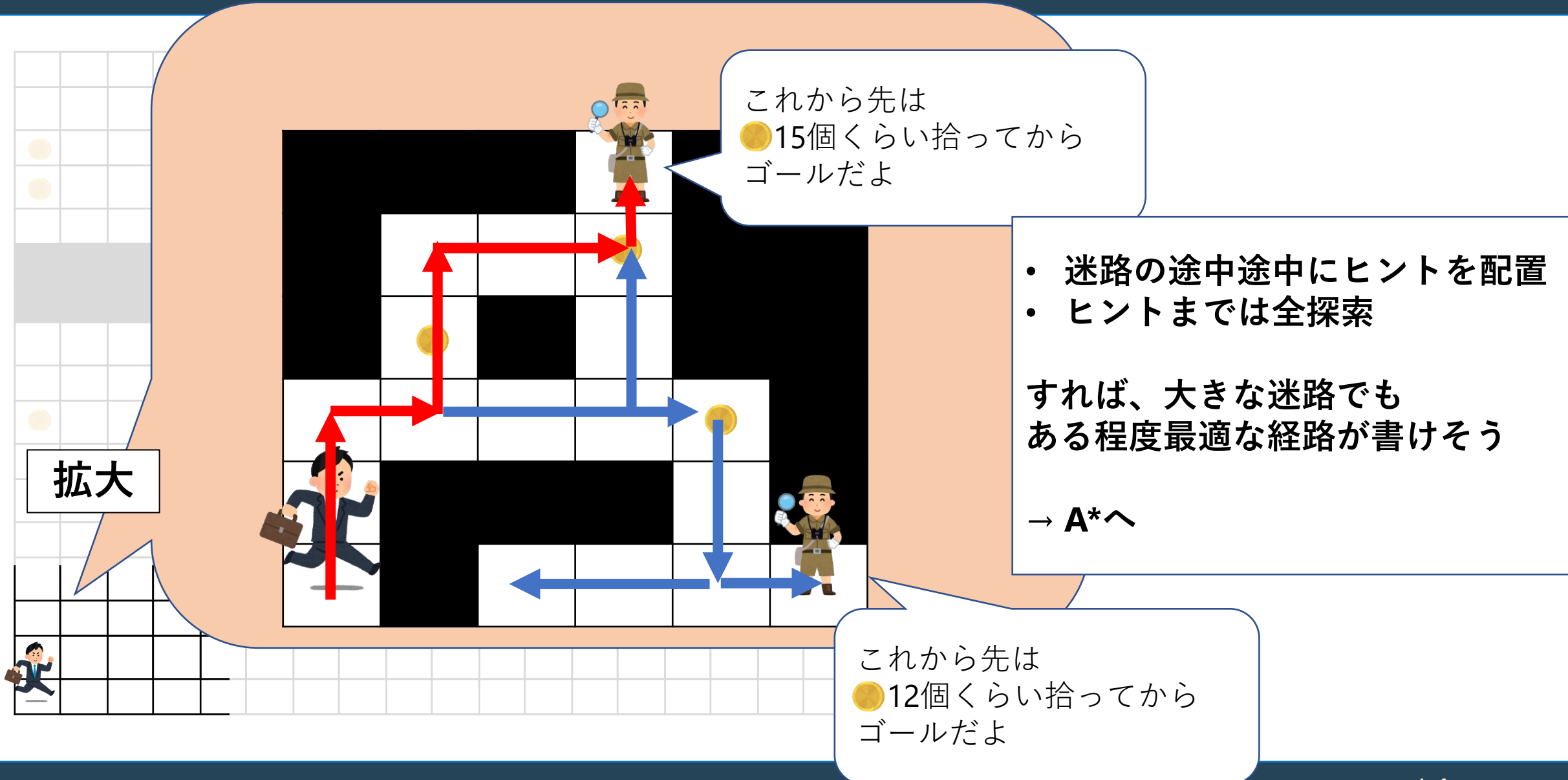
拡大

これから先は
● 12個くらい拾ってから
ゴールだよ

追加ルール：ヒントのおじさん

- ヒントのおじさんは、これから先のマスについて **90%正しい情報を返します**
- ヒントのおじさんは好きな箇所に配置できます

追加ルール：ヒントのおじさん

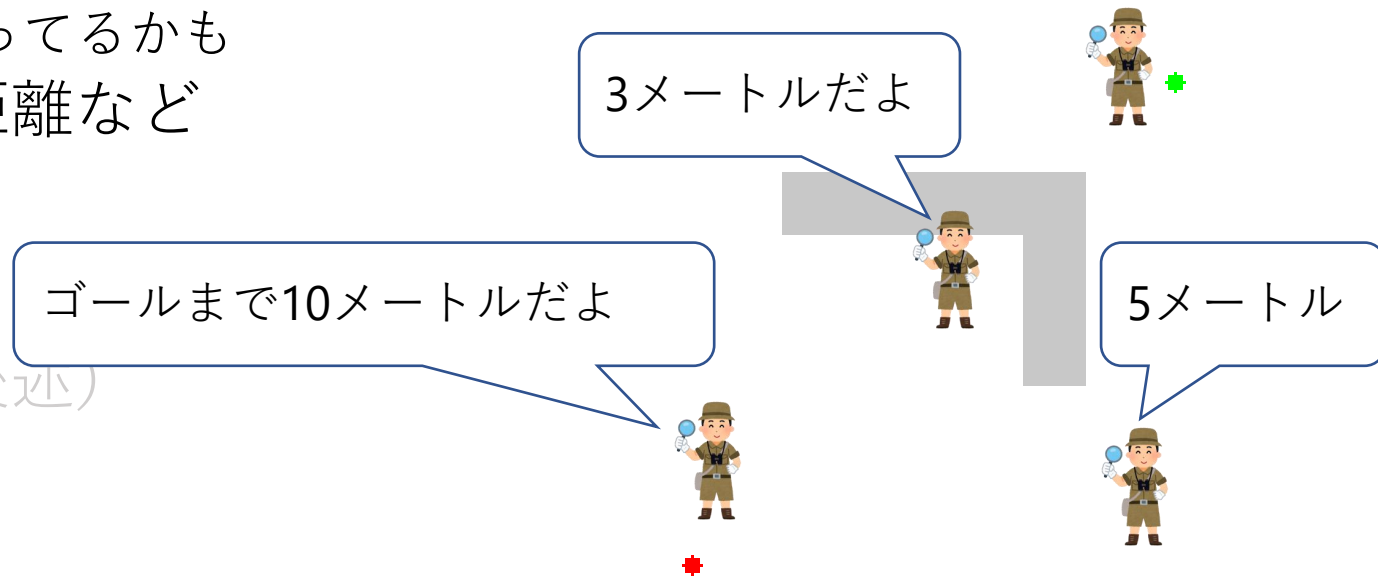


基本のアルゴリズム

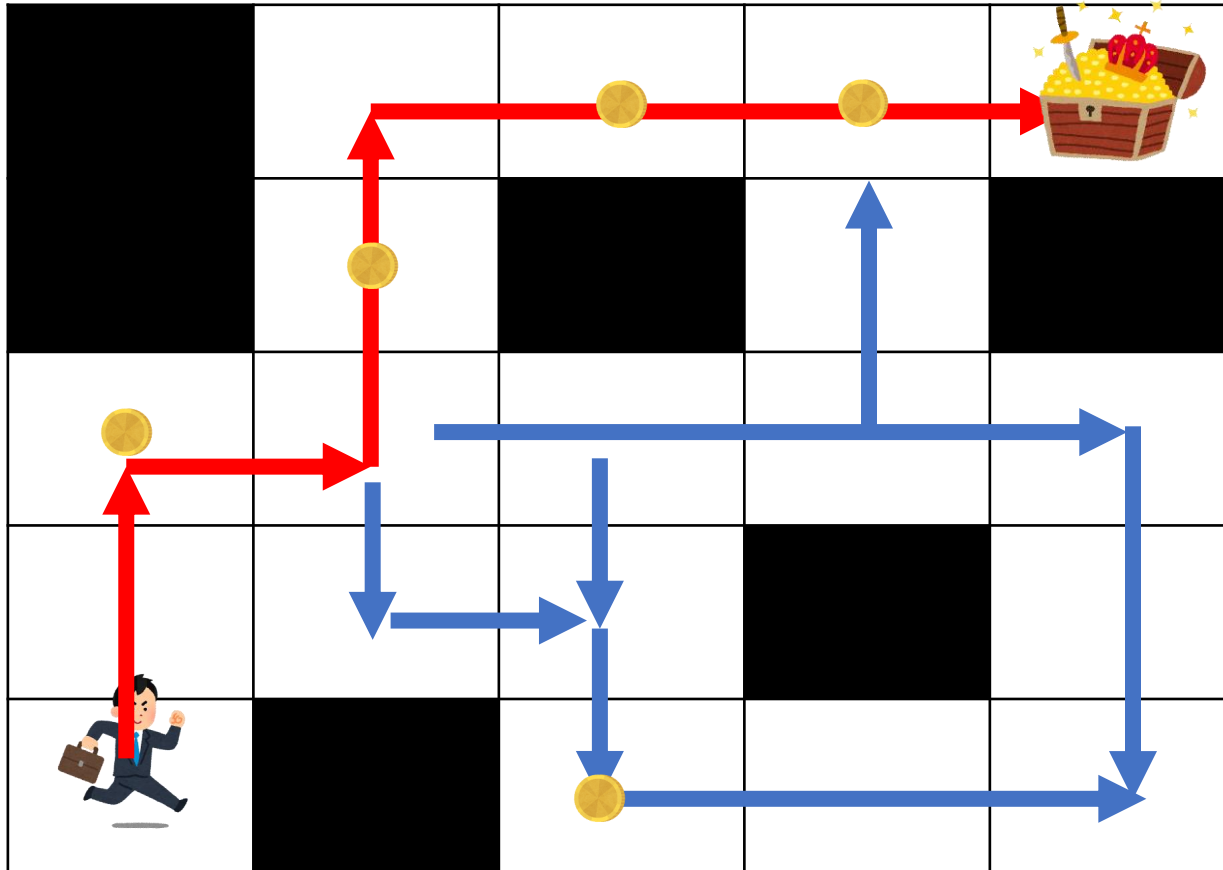
- 全列挙：ヒントなし
- A*：ヒントあり
 - ヒントは**人間が設計する関数**（ヒューリスティック関数）
 - なので、ヒントの情報は間違ってるかも
 - ヒントの例：ゴールまでの距離など

- 強化学習：ヒントあり
 - ヒントは試行錯誤で学習（後述）

次：ヒントのおじさんの使い方
アルゴリズムが変わるよ



ヒントの使用頻度とトレードオフ

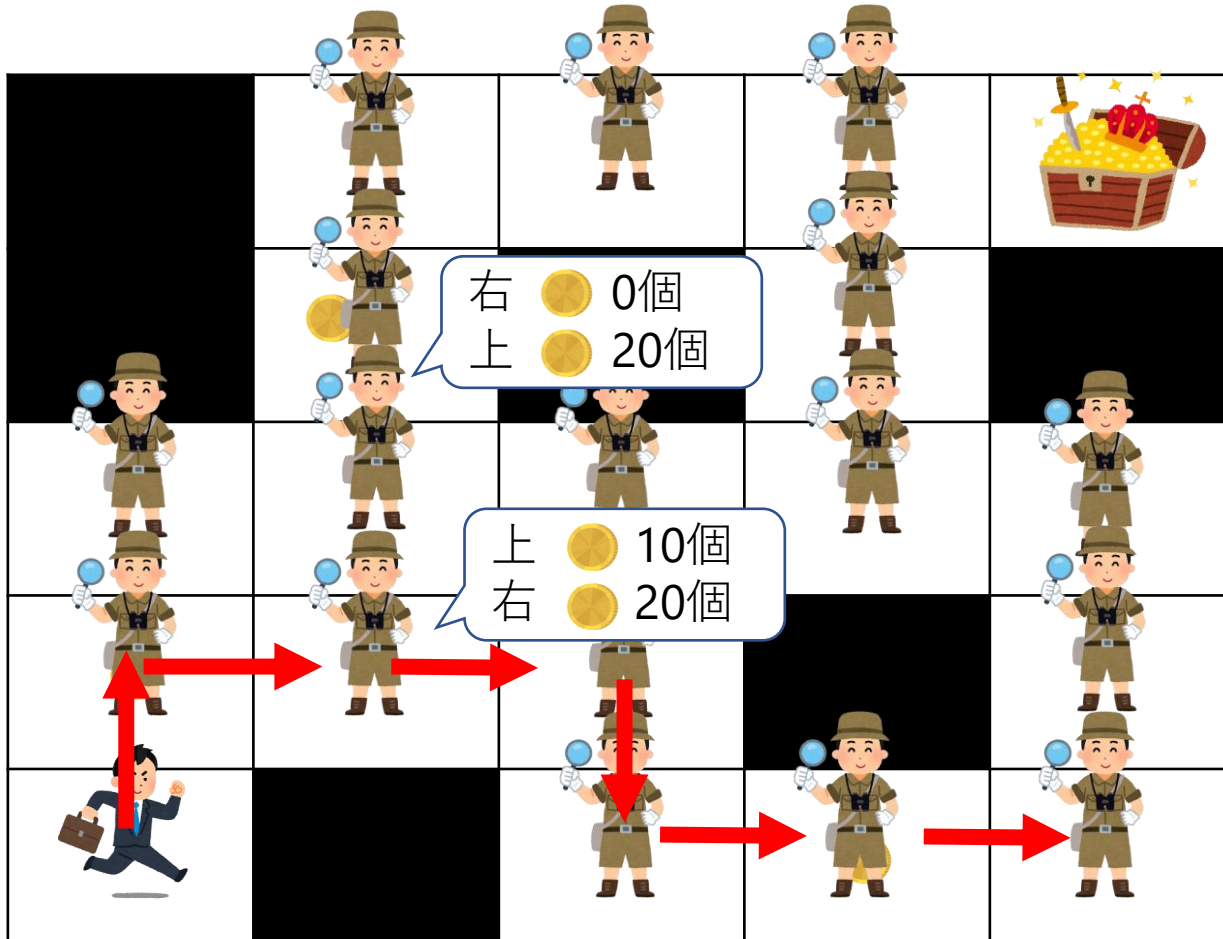


ヒントのおじさん

- ヒントのおじさんは、これから先のマスについて **90%正しい情報を返します**

- 一回もヒントのおじさんに従わない場合...
- 最適な経路は見つかるけど
列挙のコストがかかる...

ヒントの使用頻度とトレードオフ

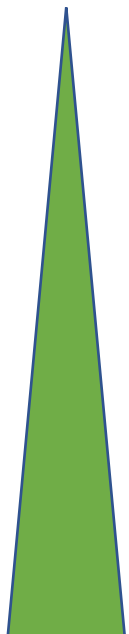


ヒントのおじさん

- ヒントのおじさんは、これから先のマスについて **90%正しい情報を返します**
- 全てのマスでヒントのおじさんに従えば 列挙の必要はないけど...
- 毎回**10%の確率**で間違えるので、最適ではない経路に誘導されちゃうかも...

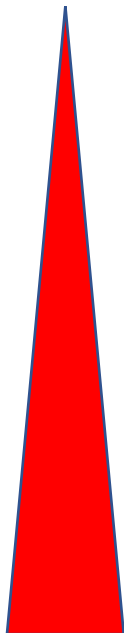
ヒントの使用頻度とトレードオフ

最適精度
低



最適精度
高

列挙コスト
低

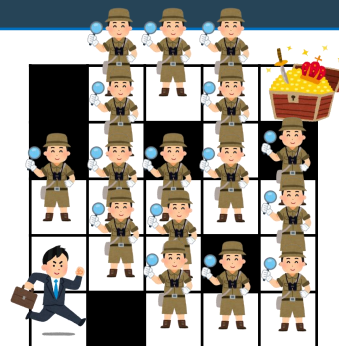


列挙コスト
高

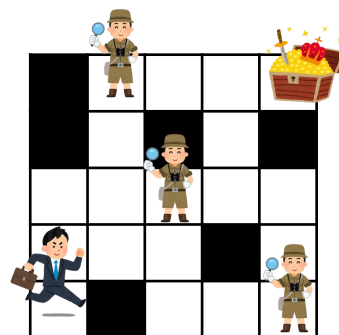
使用頻度
高



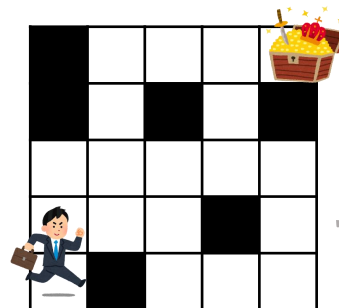
使用頻度
低



Q学習



A*
Alpha-Go
モンテカルロ
木探索



全列挙
モデル予測制御

アルゴリズムの違いは
ヒントの使い方！

次ページから：
ヒントの設計の仕方にも
違いがあるよ

目次

1. 逐次意思決定問題って？迷路の例

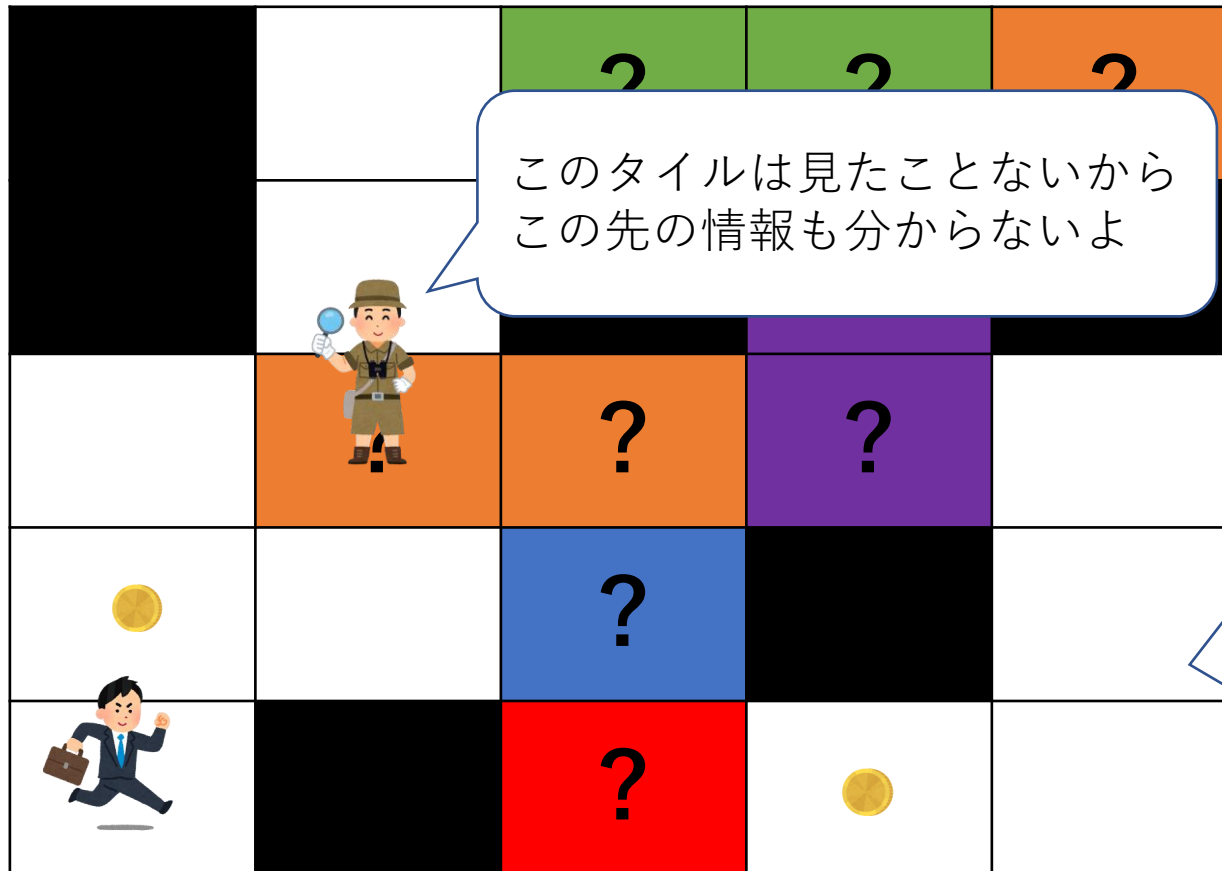
2. ヒントの使い方

- ・ ヒントの使用頻度とトレードオフ

3. ヒントの学習

- ・ ヒントの学習頻度とトレードオフ

難しい迷路 | 何が起こるかわからない迷路



「何が起こるかわからない迷路」

- ゴール位置不明
- 色付きタイルは何が起こるか不明
 - ランダムに宝を10個失うかも
 - スタートに飛ばされるかも
 - ゴールが隠れてるかも

→ 事前にヒントのおじさんが設計できない...

難しい迷路 | 何が起こるかわからない迷路

おっけー！

右に行くと+0だよ

下に行くと+100だよ

タイルの情報が分からないので、実際に動き回って試す
& ヒントを学習させよう

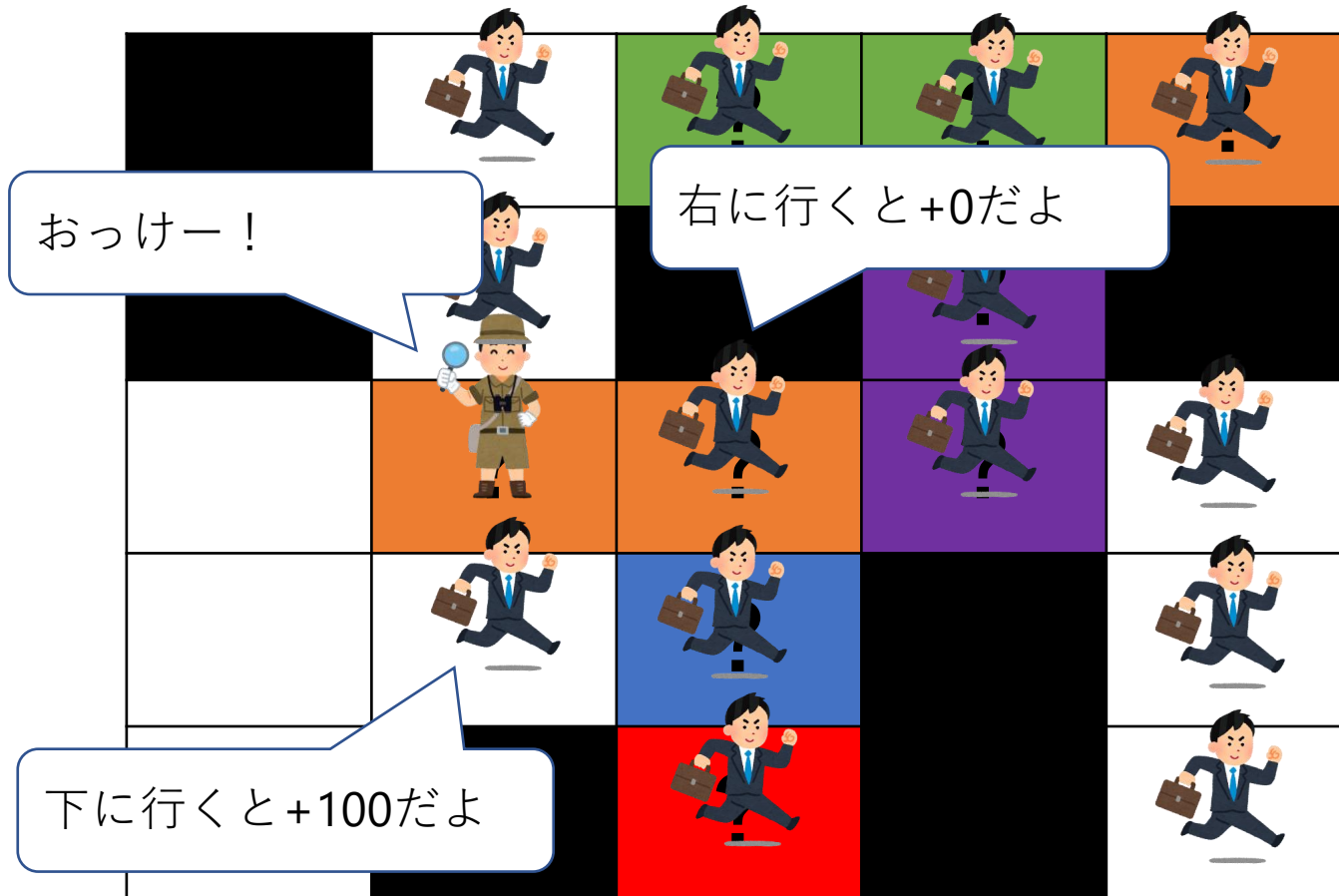
→ ヒントの学習させ方で
アルゴリズムが変化

基本のアルゴリズム

- モデル予測制御：ヒントなし
 - ただし**タイルの情報が全部既知**の場合の全列挙
- Q学習、モンテカルト木探索、Alpha-Go：ヒントあり
 - タイルの情報が未知の場合を対処
 - ヒントは**試行錯誤で学習**（ヒントはQ値と呼ばれる）

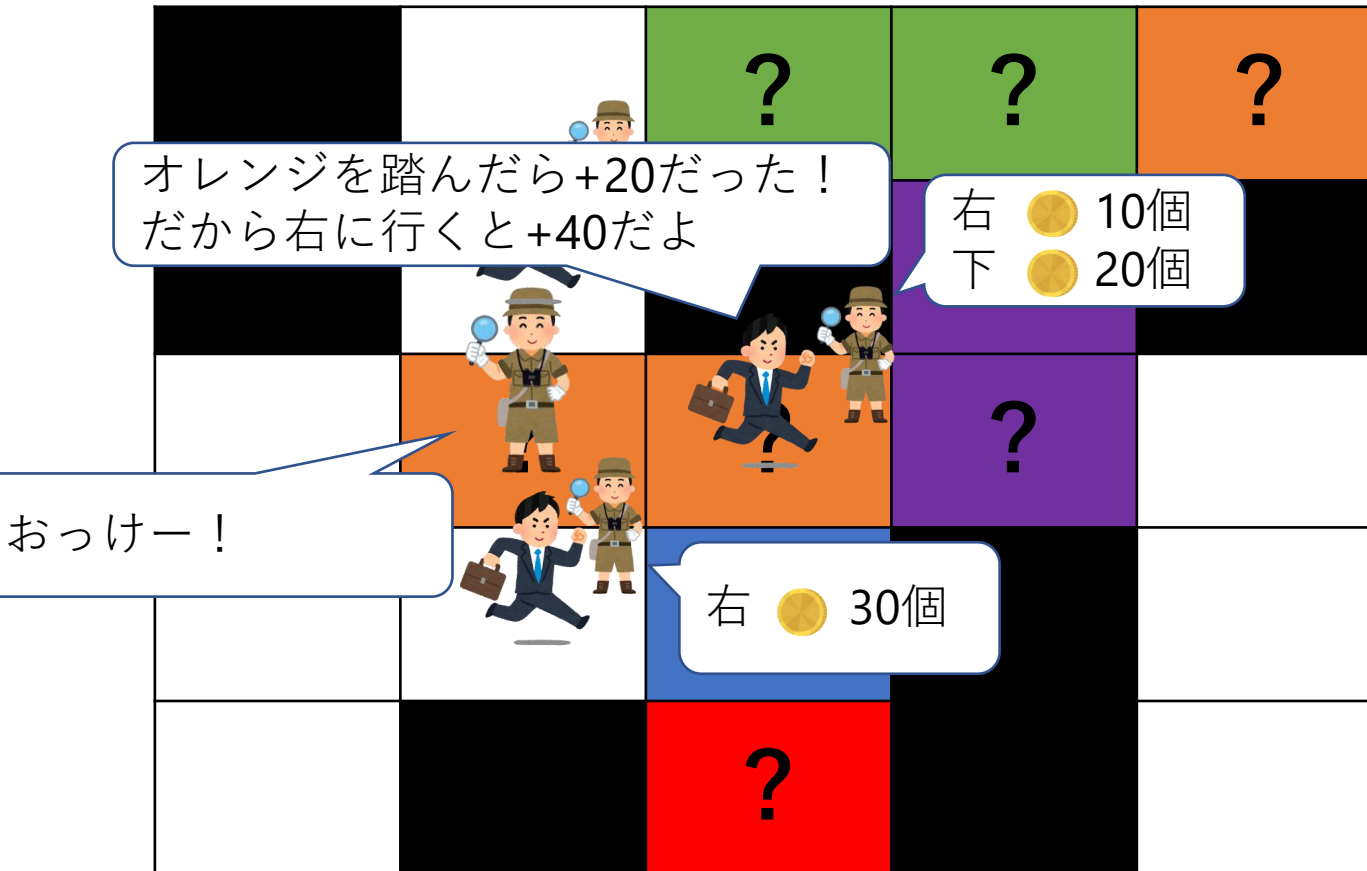
次：ヒントのおじさんはどうやって学習させればいい？

ヒントの学習頻度とトレードオフ



- ヒントを学習させたい場所以降の可能性を全列挙すれば、ヒントおじさんの精度は上がるけど...
- 列挙にコストがかかっちゃう...

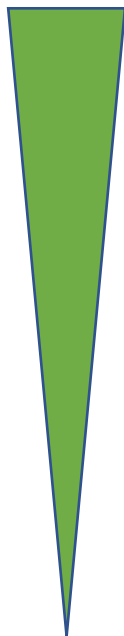
ヒントの学習頻度とトレードオフ



- 1ステップだけ進んで判明した情報を、まだ未熟なヒントの情報と合体させよう
- 列挙のコストは低いけど、ヒントの精度は悪い...

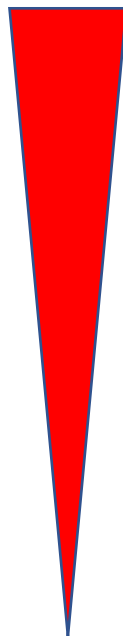
ヒントの学習頻度とトレードオフ

学習精度
高



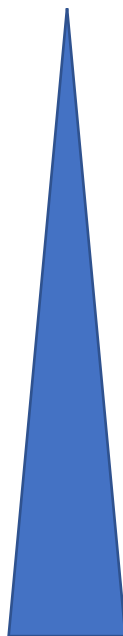
学習精度
低

学習コスト
高



学習コスト
低

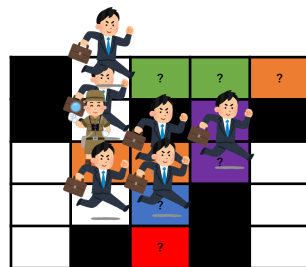
学習頻度
低



学習頻度
高



モンテカルロ
木探索



??
(Rainbow)



Q学習

アルゴリズムの違いは
ヒントの学習頻度！

まとめ

- 「ヒントの使用頻度」と「ヒントの学習頻度」の違いがアルゴリズムの違いに繋がるよ
 - **マルチステップ強化学習**と呼ばれる枠組みだよ
 - 専門用語：「使用頻度→方策更新」、「学習頻度→方策評価」に対応
- 実は**まだ見つかっていないアルゴリズムが結構あるよ**
 - 特に使用頻度側の発展は微妙。もっと良いアルゴリズムがあるかも。