

# DS-GA 3001 001 | Lecture 6

## Reinforcement Learning

---

Jeremy Curuksu, PhD

NYU Center for Data Science

[jeremy.cur@nyu.edu](mailto:jeremy.cur@nyu.edu)

March 12, 2025

# DS-GA 3001 RL Curriculum

---

## Reinforcement Learning:

- ▶ Introduction to Reinforcement Learning
- ▶ Multi-armed Bandit
- ▶ Dynamic Programming on Markov Decision Process
- ▶ Model-free Reinforcement Learning
- ▶ Value Function Approximation (Deep RL)
- ▶ Policy Function Approximation (Actor-Critic)
- ▶ Planning from a Model of the Environment (AlphaZero)
- ▶ Aligning AI systems to Human Preferences (ChatGPT)
- ▶ **Examples of Industrial Applications and Project Q&A**
- ▶ Advanced Topics

# Reinforcement Learning

---

## Last week: Value Function Approximation

- ▶ Categories of Functions in Reinforcement Learning
- ▶ Approximation of State Update Functions
- ▶ Approximation of Value Functions
- ▶ Deep Reinforcement Learning

## Today: Examples of Industrial Applications

- ▶ A Tour of RL Industrial Applications
- ▶ Project Q&A

# Robotics

# Teach a Robot to ...



Source: DeepMind (2022)

# Autonomous Driving

# Learn to Drive Like a Human

- **Goal:** Drive vehicle on a circuit to destination without leaving the road

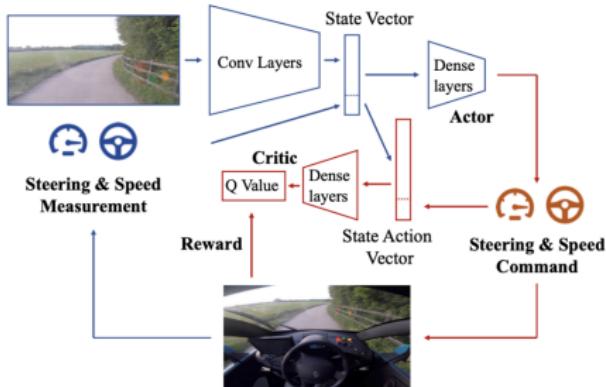


## Environment:

Virtual and physical driving lanes

## Reward Function:

- Forward Speed
- Termination upon *infraction of traffic rules* by safety driver



## State:

- Pixels of front camera encoded by CNN (single monocular image)
- Vehicle speed and steering angle

## Actions:

2 actions: Speed, steering angle

# Learn to Drive Like a Human

---

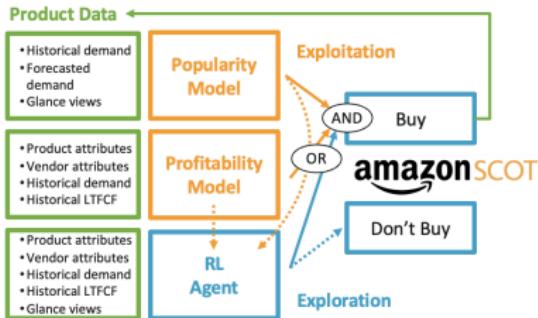


Source: Wayve (2019)

# **Amazon Inventory Management**

# Manage Amazon Retail Inventory

- ▶ **Goal:** Select products to show as "shipped and sold by Amazon" on the Amazon.com website to maximize customer experience and profitability



**Offline Validation:** All auto parts and medical supply products in catalog as of 12/2018.  
Cash flow and sales in 2019.

Statistics	ASINs selected to buy	ASINs blocked from buying
ASINs count (in MM)	0.16 (2.7%)	5.55 (97.3%)
Cash flow (in MM euros)	152.35	-19.04
Sales (in MM)	28.06 (91.7%)	2.57 (8.3%)

Online A/B testing: Q4 2019, 30M products, 90% treatment, 10% control

EU LAB	Treatment Effect per ASIN per week	Confidence Interval	p-value	Annualized Impact
CP (Euros)	0.0103	[0.002, 0.019]	0.02	€2.45 MM
Sales (Euros)	0.021	[-0.061, 0.103]	0.10	€4.68 MM
Cash flow (Euros)	0.1123	[-0.311, 0.536]	0.54	€25.03 MM
Out of stock (bps)	-74	[-100, -50]	0.00	-74 mm



#### Environment (Contextual Bandit):

Product's profitability and popularity,  $\Delta t = 3$  months



### Reward Function:

$$\text{Profit} = \sum (\text{short term profit} + \text{long term value})$$

-cost of capital - asset depreciation



 State:

Product-, brand- and vendor-level statistics, historical sales, historical cash flow, glance views



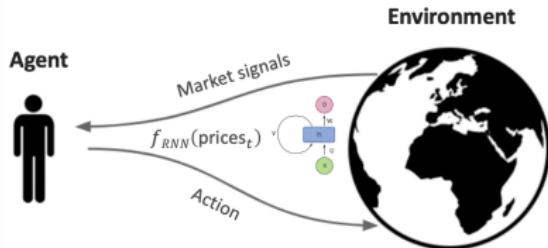
## **Actions:**

1 action: (block buy, buy at least 1 unit)

# Algorithmic Trading

# Manage a Stock Portfolio

- **Goal:** Identify trading strategy in portfolio of  $k$  stocks to maximize profit



RNN-based RL outperforms lag-based RL at trading stocks

Mean return \$4,900 vs. \$2,200; Maximum return \$7,500 vs. \$2,800



## Environment:

7 years of stock prices and news headlines,  $\Delta t = 1$  day

## State:

Vector of  $k$  stock prices for last  $n$  days  
encoded by RNN

## Reward Function:

$$r_t = r_t^0 + r_t^{risk} + r_t^{fee} = \sum_k \underbrace{\frac{(\text{prices}_{t+1} - \text{prices}_t)}{\text{prices}_t}}_{\text{Portfolio Value Change}} x_t - \lambda \sigma_t^2(r_t^0) - \kappa_t^T x_t$$

Risk Penalty      Transaction Cost

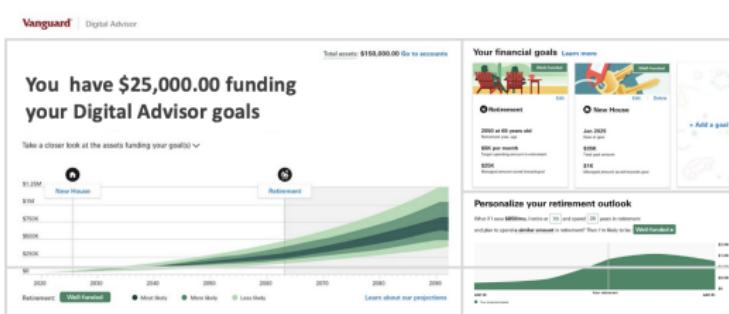
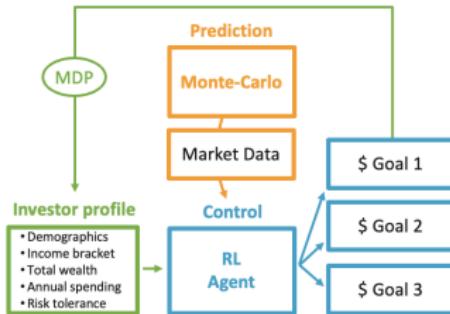
## Actions:

$k$  actions: (\$buy, \$sell, sit) $_k$

# **Asset Allocation and Wealth Management**

# Manage Long-Term Financial Goals

- ▶ **Goal:** Determine optimal asset allocation strategy to meet multiple long-term financial goals, while also being successful in retirement



 **Environment:**

$$p(s' | s, a)_{stable} + w(\sigma^2)_{MC}, \mathbb{E}(p_{MC}, a)_{MC}, \Delta t = 1 \text{ year}$$

## Reward Function:

$$r_t = r_t^{work} + r_t^{retired} = -\beta_1 \sum_i |g_t^i - g_T^i| + \beta_2 p_{MC}$$

## Observed State:

Investor profile, past \$ contribution to each goal

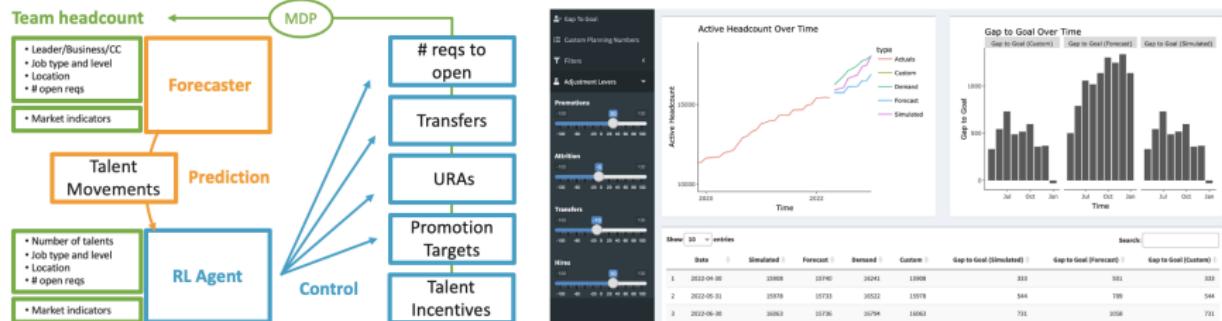
## Actions:

$k$  actions: \$ contribution to each goal

# Workforce Planning

# Manage a Large Talent Workforce

- **Goal:** Pull talent levers to minimize gaps "actual vs. target headcount" while also minimizing costs across the World-Wide Amazon workforce



Environment:

$$p(s' | s, a)_{(\text{internal + external factors})}, \Delta t = 1 \text{ month}$$

State:

Talent team size, job type, job level, location, number of open jobs, market data, talent movement forecasts

Reward Function:

$$r_t = r_t^{gap} + r_t^{cost} = \beta_1 |h_{EoY} - h_{target}| - \beta_2 c_t$$

Actions:

Jobs to open, Transfers, URA, Promotions, Compensation (...)

# Workforce MDP Simulator

- **Model used to define next states:** Forecast monthly talent movement based on historical trends and market indicators

## Graph Transformer for Workforce Planning

Jérôme Curakus  
People Experience and Technology  
curakj@amazon.com

Jérôme Righy  
People Experience and Technology  
righyj@amazon.com

### Abstract

We present a Graph Transformer deep learning method for workforce planning which can identify potential talent risks and forecast individual monthly talent movements in each segment of the Amazon corporate population. This method outperforms last methods in career change at Amazon by 40% in 76% of the segments and outperforms all other methods in 70% of the segments and 17-79% of the segments. Given the dependency Graph in the proposed method can be used to interpret and understand talent forecasts, there is little drawback to using this method compared to using linear trailing rates for workforce planning.

### 1 Introduction

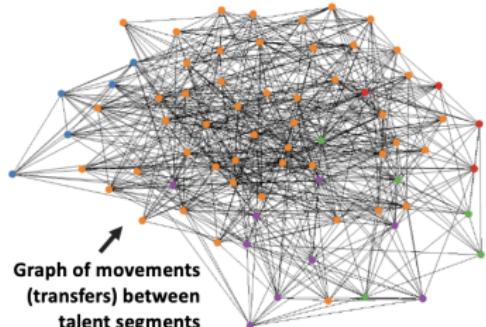
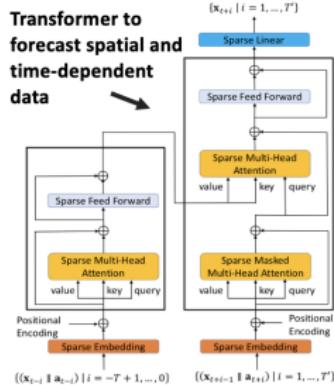
Workforce planning at Amazon sets year-end headcount targets by financial cost center to meet the company's current and future staffing needs based on business goals and talent movement forecasts (hires, promotions, transfers, attritions). Individual team leaders further plan their workforce needs by individual team, job type, job level and location. Failure to accurately forecast future headcounts and staffing needs often results in delays in productivity and costly resource allocation [1].

Forecasting Amazon talent movements is challenging due to complex spatial and temporal dependencies within the Amazon population, and non-stationarity that result from unusual events such as the Covid pandemic [2]. Amazon is an especially difficult forecasting problem due to its diverse operations worldwide and unprecedented size (over 1.6M employees in peak season [2]).

Examples of talent movement dependency at Amazon include talents in similar environments, with similar profiles and job market opportunities, or under similar talent management strategies. Many other factors, including external factors such as large swings in the company's stock value [3], can influence talent flows and thereby create correlated traffic patterns in the Amazon population.

Forecasting spatial and time-dependent data in large, complex traffic networks was recently addressed by estimating a dependency Graph that parsimoniously represents the spatial dependency between different locations in the network (nodes), and using the Graph to sparsify a deep RNN and CNN and improve the learning of long-range temporal dependency [5]. By assigning each neuron of the Transformer with a spatial location and using knowledge from the dependency Graph to prune neural connections that are not dependent, the resulting Graph Transformer could efficiently capture both spatial and temporal dependencies and scale to large traffic patterns in the Amazon population [4].

In this paper, we use a multivariate Gaussian approximation to find the dependency Graph of talent movements over the different teams of the Amazon corporate population defined by leader, job type, and job level. We use this Graph to derive insights into the overall dynamics of Amazon talent

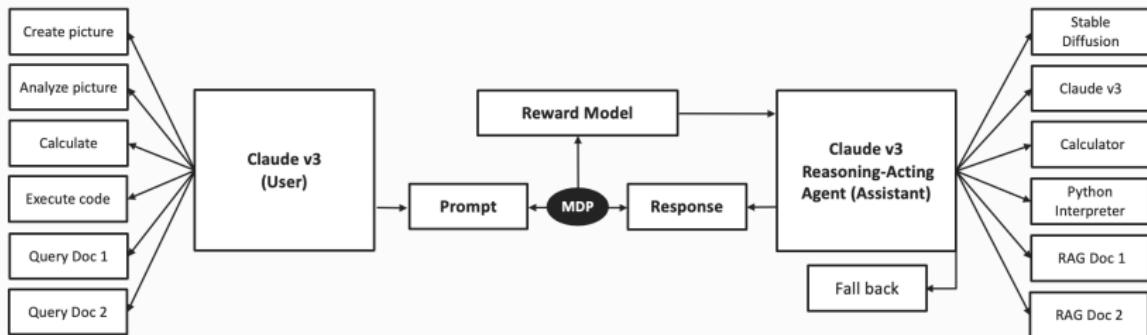


Model	MAPE	Nodes improved	MAPE in nodes improved
Graph Transformer	84%	N/A ( <i>self</i> )	93%
Prophet	95%	70%	108%
s-ARIMA/State Space	87%	61%	107%
Trailing 3-month	112%	76%	133%

# Alignment of LLM Agents

# Generative AI: Learn to ask for clarifications

- **Goal:** Increase validity and efficiency of reasoning agents in a preset orchestration environment by learning when to ask for clarifications



## Environment:

- Sample of  $n$  queries mapped to valid tools.
- LLM emulates users and generates prompt variation for each query.

## Reward Function:

- Valid tool selection: +5, Invalid tool selection: -5
- Ask for clarification: 0, Every step: -1

## State:

- User prompt
- System prompt and tools

## Actions:

Select one of  $k$  tools and respond to the user, or fallback on asking for clarifications

# Generative AI: Learn to ask for clarifications

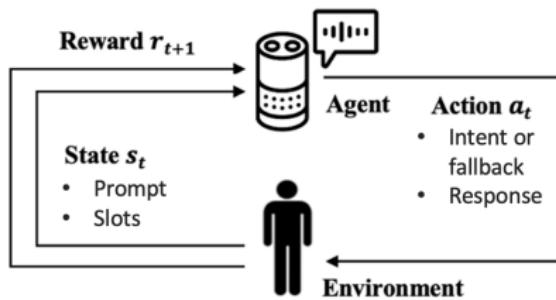
- **Results (NeurIPS 2024):** Agent identified prompts too ambiguous to identify correct tools and worth falling back on asking for clarification



Figure 4: Accumulated reward and number of fallbacks across  $10 \times 30$  episodes of reflections sampled with VRL. The x-axis corresponds to the five prompts generated by Claude for each of the six queries. Epochs 1-10 are listed from top to bottom.

# Generative AI: Learn to ask for clarifications

- **Results (ICML 2023):** Agent converged to policies which fulfilled intents in 1.3 steps on average in 100% of dialogues (Nash Equilibrium)



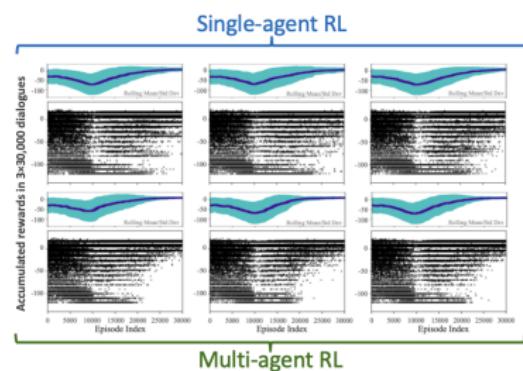
## Environment:

- Library of intents with associated prompts/slots
- Prompts selected randomly (*single agent RL*) or by another agent (*multi-agent RL*)



## Reward Function:

- **Efficiency:** Every step: -1
- **Validity:** Valid intent: +5, Invalid intent: -5
- **Coherence:** Valid slot: +5, Invalid slot: -5
- **Fallback:** Ask for clarification: 0

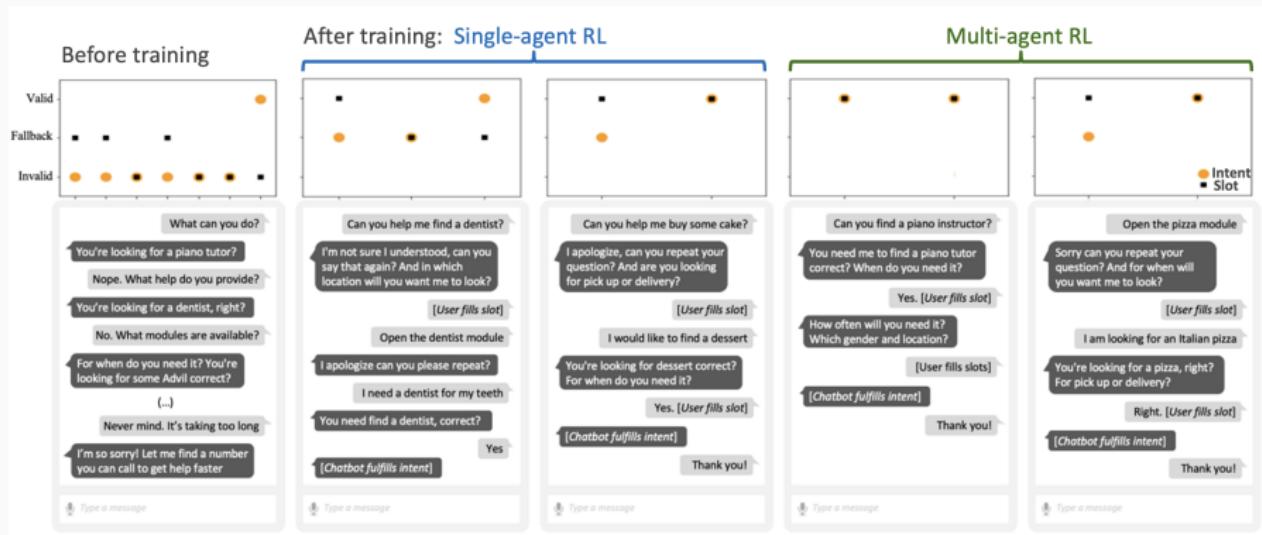


## Quality and efficiency of sampled dialogues

	Single RL 0-5K	Single RL 25-30K	Multi RL 0-5K	Multi RL 25-30K
% of successful dialogues	68 (.8)	99 (.1)	67 (1.4)	100 (0)
Number of steps in successful dialogues	4.1 (.2)	1.0 (.0)	4.1 (.1)	1.0 (.0)
/navigation	4.2 (.1)	1.8 (.2)	4.1 (.1)	1.3 (.1)
/piano	4.2 (0)	1.6 (.2)	4.1 (2)	1.3 (.0)
/dentist	4.1 (.1)	1.7 (.3)	4.3 (0)	1.3 (.0)
/pizza	4.3 (.2)	1.7 (.3)	4.2 (.1)	1.3 (.0)
/Advil	4.4 (.1)	2.3 (.5)	4.3 (.1)	1.4 (.1)
/dessert				

# The chatbot learned an original strategy...

- ▶ **Superalignment:** The chatbot found a fallback strategy to increase efficiency without sacrificing coherence: fill valid slots when prompt is partially understood but too ambiguous to identify the exact intent



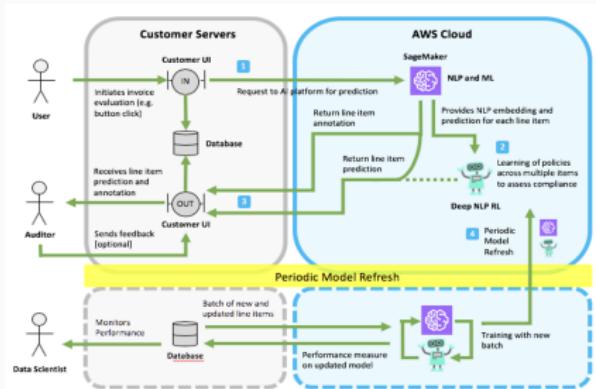
# Your Turn!

# **Appendix (not covered in class)**

# Audit Financial Claims with NLP

# Audit Claims with Natural Language

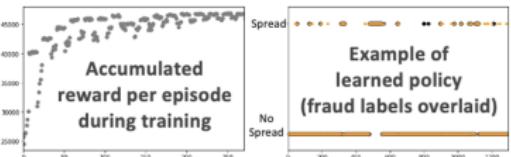
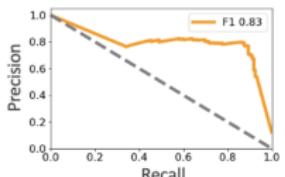
- **Goal:** Recommend compliance level of items in financial claims, to reduce time spent by human contractors, and to reduce errors



Agent finds 86% of known frauds at 80% precision

#### Fraud Test Results:

Precision: 80%  
Recall: 86%  
F1 Score: 0.83  
Brier Score: 0.04  
AUC Score: 0.95



#### Environment:

50,000 claims with at least 2 items per claim

#### Reward Function:

- Compliance category predicted by NLP classifier trained on individual items (regular audit)
- Frauds detected by specialized auditors

#### State:

- Last  $n$  items in claims encoded by NLP
- Claim metadata (source, activity code, billed units, ...)

#### Actions:

2 actions: Fraud risk level and compliance category

# Audit Claims with Natural Language

- ▶ **Post-processing to interpret RL results:** Clustering in NLP space of items at risk can help auditors identify patterns of frauds more quickly

**Rearrange non-compliant items per date/source/claim**

...then cluster items per narrative

Example of 16 items at risk claimed by Mr. X:

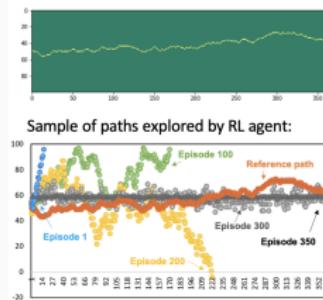
Source	Claim ID	Date	Description	
Mr. X	E1731	1/15/2019 0:00	Initiate proper disclosure of claimant's sworn testimony (authorized by claims representative) through preparation of Notice according to S. C. Code (legal document requiring signature of S. C. attorney)  sworn testimony (authorized by claims representative) on opposing counsel through preparation of statutorily required document according to S. C. Code	1
	E1732	1/25/2019 0:00	Execute Amended Notice of Claimant's sworn testimony (authorized by claims representative) on opposing counsel through preparation of statutorily required document according to S. C. Code  Initiate proper disclosure of claimant's sworn testimony (authorized by claims representative) through preparation of Amended Notice according to S. C. Code (legal document requiring signature of S. C. attorney)	1
		10/30/2018 0:00	Advisement of medical discovery directive to pursuant to - 67-214 non-applicability of HIRPA Privacy Rule to Workers' Compensation matters and applicable state law for evidentiary document or documents	1
			Advisement of medical discovery directive to pursuant to - 67-14 non-applicability of HIRPA Privacy Rule to Workers' Compensation matters and applicable state law for evidentiary document production	1
			Advisement of medical discovery directive to EMS pursuant to - 67-14 non-applicability of HIRPA Privacy Rule to Workers' Compensation matters and applicable state law for evidentiary document production	1
			Execute medical discovery request on through preparation of statutorily required document according to , and	5
			Initiate medical discovery on EMS through preparation of medical evidence request according to Reg. (legal document in S. C.) to obtain pertinent documents for analysis	1
			Initiate medical discovery on through preparation of medical evidence request according to Reg. (legal document in S. C.) to obtain pertinent documents for analysis	2
		10/31/2018 0:00	Initiate proper disclosure of claimant's sworn testimony (authorized by claims representative) through preparation of Notice according to S. C. Code (legal document requiring signature of S. C. attorney)  sworn testimony (authorized by claims representative) on opposing counsel through preparation of statutorily required document according to S. C. Code	1
		12/12/2018 0:00	Advisement of discovery directive to Claimant & e., (pursuant to - 67-214) representation of employer/career and production of evidentiary documents (A108)	1
			Initiate discovery on claimant through preparation of subpoena evidence request according to Reg. (legal document in S. C.) to obtain pertinent documents for analysis, including tax returns and business records	1
		12/5/2018 0:00	Initiate proper disclosure of claimant's sworn testimony (authorized by claims representative) through preparation of Notice according to S. C. Code (legal document requiring signature of S. C. attorney)  sworn testimony (authorized by claims representative) on opposing counsel through preparation of statutorily required document according to S. C. Code	1

# **Seismic Mapping to Identify Natural Oil & Gas Reserves**

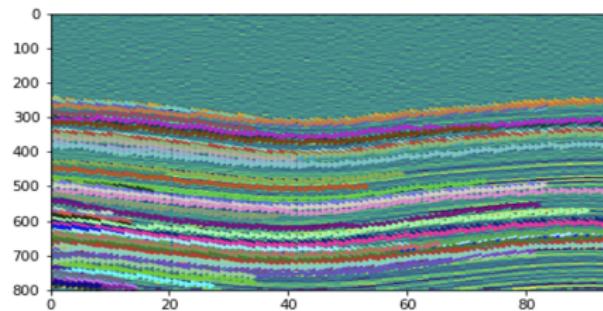
# Seismic Pathfinder

- **Goal:** Screen cross-sections under the earth surface to identify the nature and geometry of individual seismic layers, to reduce exploration costs

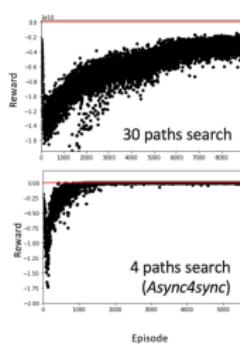
Pathfinder for 1 path:



1 cross-sectional (cylinder) map:



Accumulated reward:



## Environment:

Underground cylinder cross-sections

## Reward Function:

$$-\beta_1 \sum |z_t^i - z_{ref}^i| - \beta_2 \sum |z_{end}^i - z_0^i|$$

## State:

Indirect seismic measurements at  $z_t$  and  $(z-3, z-2, z-1, z, z+1, z+2, z+3)_{t+1}$

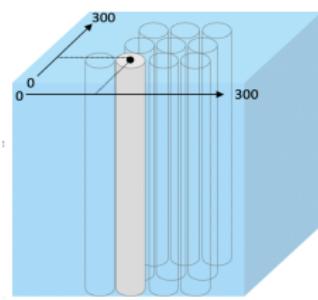
## Actions:

$k$  actions: (up, down) for each path

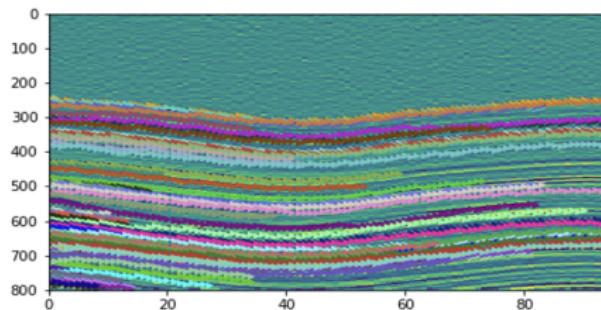
# Seismic Pathfinder

- **Goal:** Screen cross-sections under the earth surface to identify the nature and geometry of individual seismic layers, to reduce exploration costs

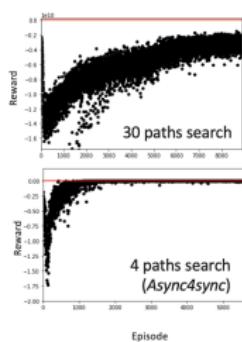
3D seismic cube:



1 cross-sectional (cylinder) map:



Accumulated reward:



## Environment:

Underground cylinder cross-sections

## Reward Function:

$$-\beta_1 \sum |z_t^i - z_{ref}^i| - \beta_2 \sum |z_{end}^i - z_0^i|$$

## State:

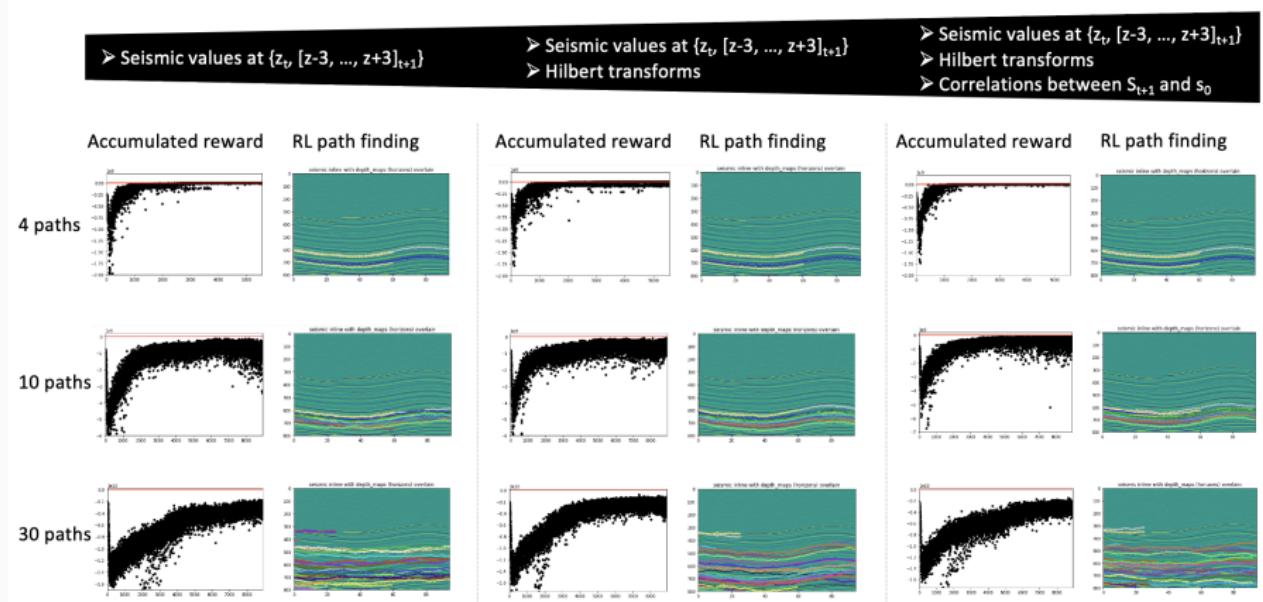
Indirect seismic measurements at  $z_t$  and  $(z-3, z-2, z-1, z, z+1, z+2, z+3)_{t+1}$

## Actions:

$k$  actions: (up, down) for each path

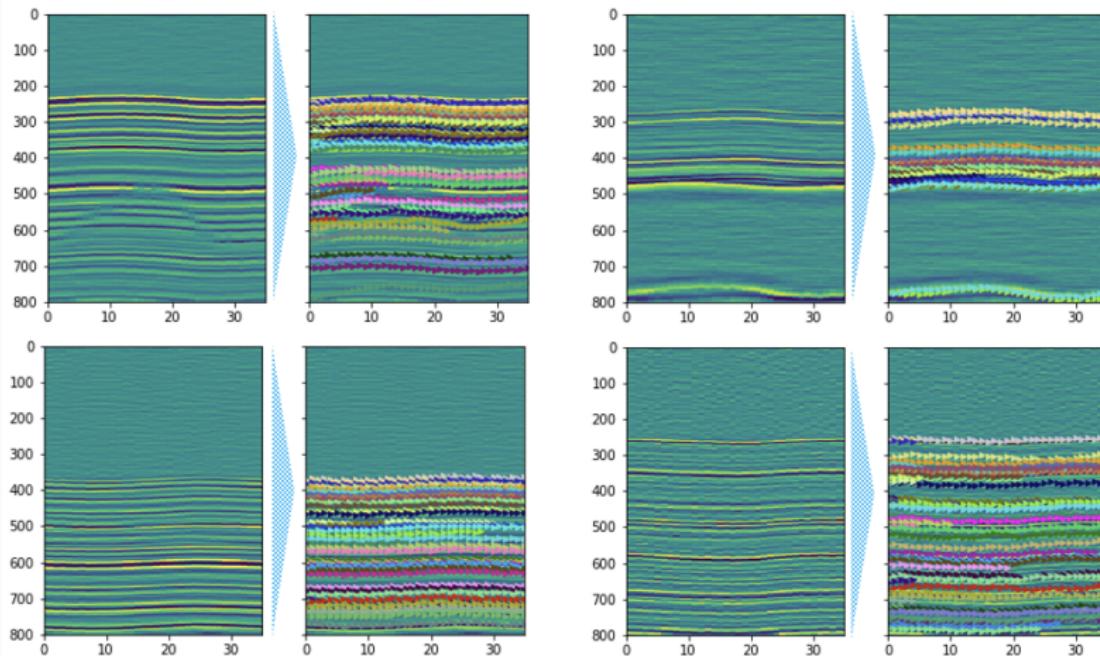
# Seismic Pathfinder

- **Analysis of Results:** Benchmarks of synchronous  $n$ -path search RL on state complexity and number of paths



# Seismic Pathfinder

- ▶ **Analysis of Results:** Generalization of pre-trained Async4sync DRL agent on arbitrary cubes and cylinders with **radius = 35 steps**



# Seismic Pathfinder

- ▶ **Analysis of Results:** Generalization of pre-trained Async4sync DRL agent on arbitrary cubes and cylinders with **radius = 90 steps**

