Mahindra™ University
Global Thinkers. Engaged Leaders.

Reinforcement Learning and Autonomous Systems (AI 4102)

Lecture 1 (17/08/2023)

Instructor: Gourav Saha

Lecture Content

- What is Reinforcement Learning (RL)?
- Overview of the modules of this course **(through examples)**.
- Course logistics.
- Miscellaneous Topics.

Lecture Content

- **What is Reinforcement Learning (RL)?**
- Overview of the modules of this course (**through examples**).
- Course logistics.
- Miscellaneous Topics.

Motivation

- Reinforcement learning in motivated by how animals (humans are also animals 😂) learn various skills.

- Example 1: Learning to ride a bicycle.
  - If we fall down we get hurt and avoid doing something that lead to falling.
  - If we are about to fall we get scared because of the final outcome and immediately take countermeasures.
  - If we can ride the bike straight for a certain distance, we feel good about ourself and try to redo something similar.

Motivation

- Reinforcement learning in motivated by how animals (humans are also animals 😂) learn various skills.

- Example 2: Teaching dogs a trick/skill.
  - If a dog does the trick, it gets a treat.
  - If it does not then no treat (not evcen just a gooooddddd boy!).

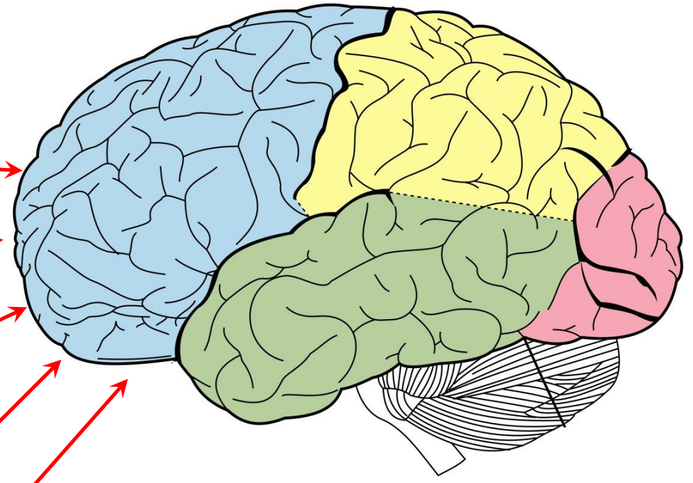
Motivation

Example 1: Learning to ride a bicycle.

- If we fall down we get **hurt** and avoid doing something that lead to falling.
- If we are about to fall we get **scared** because of the final outcome and immediately take countermeasures.
- If we can ride the bike straight for a certain distance, we **feel good** about ourself and try to redo something similar.

Example 2: Teaching dogs a trick/skill.

- If a dog does the trick, it gets a **treat**.
- If it does not then **no treat** (not even just a goooodddd boy!).



The brain converts these experiences (highlighted in red) into some form of REWARD.

Motivation

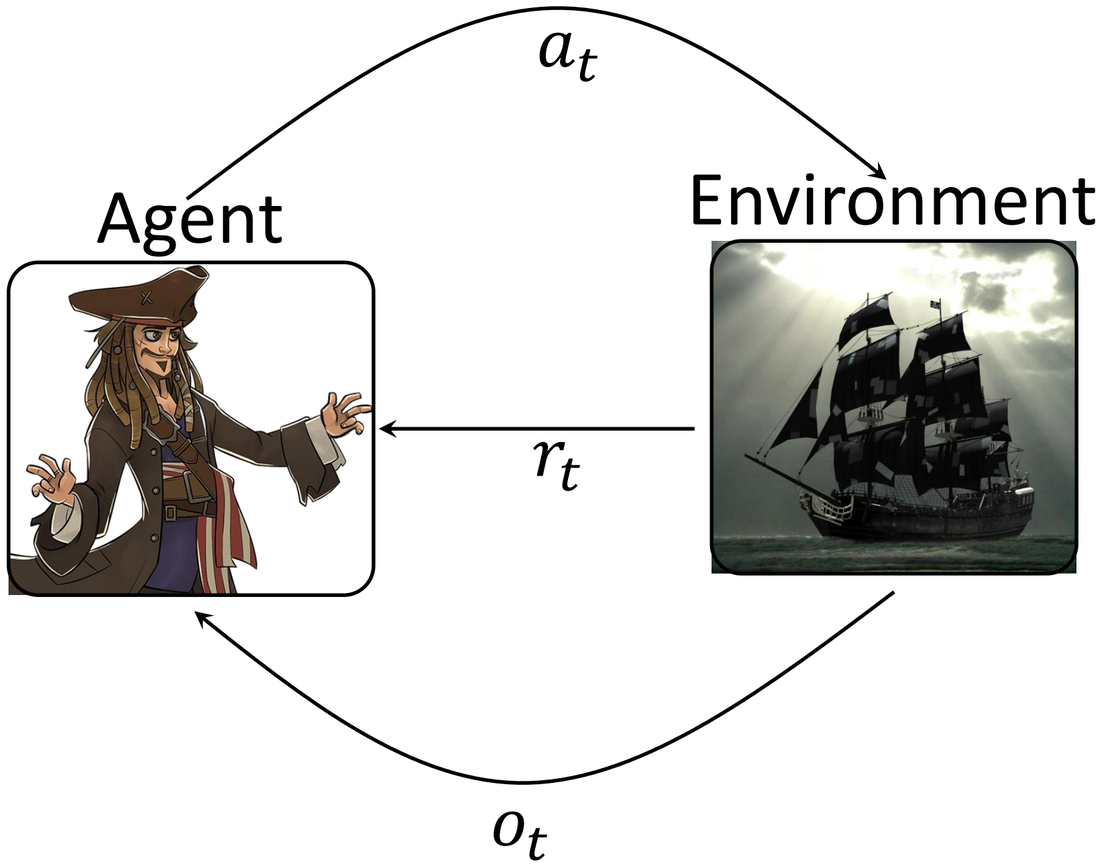
- Reinforcement learning in motivated by how animals (humans are also animals) learn various skills.

- Example 2: Teaching dogs a trick/skill.
  - If a dog does the trick, it gets a treat.
  - If it does not then no treat (not evcen just a goooodddd boy!).

What is Reinforcement Learning?

- Reinforcement learning (RL) deals with:
  - **Sequential decision making**. "Sequential" means over time. This is the **planning** step.
  - **Learning from experience** in order to maximize reward (minimize cost). This is the **learning** step.
- In RL, time is often discretized into **time slots**. This is a default assumption throughout the course.

What is Reinforcement Learning?



- RL setup is often visualized as an **interaction** between an **agent** and an **environment**.

- In time slot t:
  1. The agent makes an observation o_t about the environment.
  2. Based on this observation, it takes an action a_t.
  3. Based on the action a_t:
     a) The agent will get a reward r_t.
     b) The **environment will change** in time slot $t + 1$. Hence, the requirement of the **planning step**.
  4. Based on the reward, the agent will update it's strategy to take actions. **The learning step**.

What is Reinforcement Learning?

VERY IMPORTANT

This leads to one of the two fundamental aspects of RL:

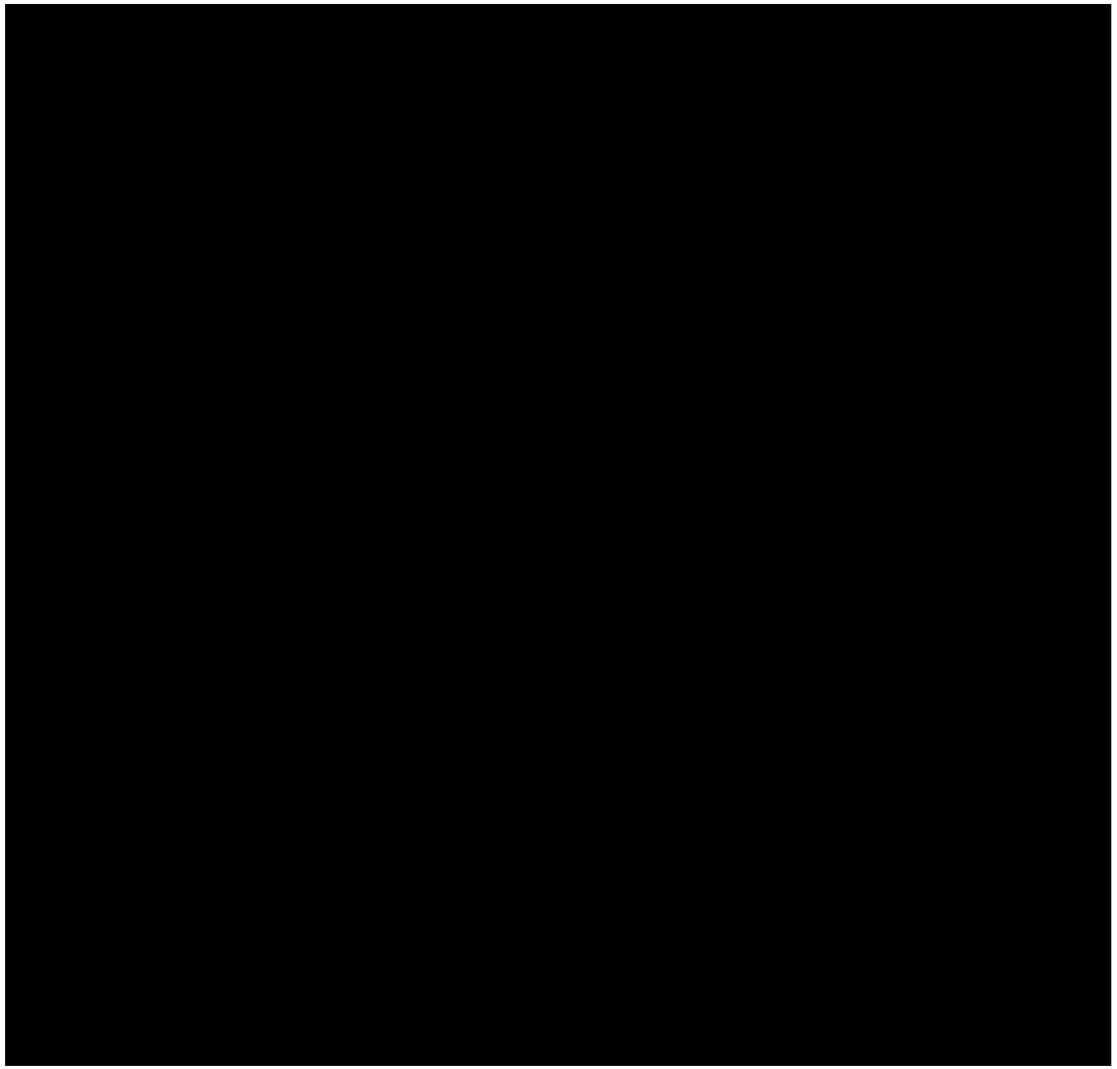
If we take an action to maximize the current reward, then the environment may change such that the future rewards are low.

Example: In chess, if we are greedy to take out opponents pieces, we can put ourself in a bad situation.

- RL setup is often visualized as an **interaction** between an **agent** and an **environment**.

- In time slot t:
  1. The agent makes an observation o_t about the environment.
  2. Based on this observation, it takes an action a_t.
  3. Based on the action a_t:
     a) The agent will get a reward r_t.
     b) The **environment will change** in time slot $t+1$. Hence, the requirement of the **planning step**.
  4. Based on the reward, the agent will update it's strategy to take actions. **The learning step**.

Examples of Reinforcement Learning

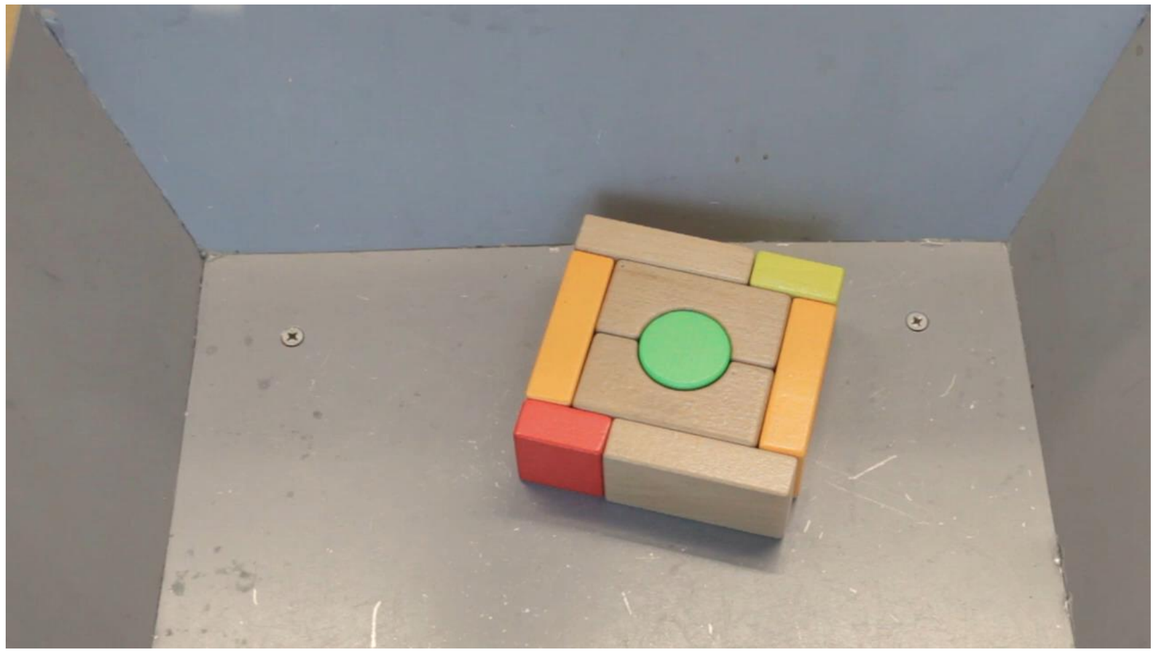


[1] Youtube channel: “Two Minute Papers” [Link]

Example 1: DeepMind Atari

- *Observations*: The RGB image.
- *Actions*: To move the tray left or right.
- *Rewards*: The net score.

Examples of Reinforcement Learning



[2] Dmitry Kalashnikov et al, "Scalable Deep Reinforcement Learning for Vision-Based Robotic Manipulation", PMLR. [Video link]

Example 2: Robotic Grippers

- *Observations:*
  - The RGB image of the objects to grip.
  - The configuration of the robotic arm plus gripper arrangement.
- *Actions:* The servo motor speed.
- *Rewards:* Whether it could grip an object or not.

Examples of Reinforcement Learning

Example 3: Millimeter Wave Communication

- The millimeter wave antenna is transmitting to the women in picture.
- Millimeter waves are highly susceptible to blockages by humans, trees, buildings. The millimeter wave antenna doesn't know about the surrounding of these women that may or may not have blockages. **It has to learn it**.
- The objective is to minimize a weighted sum of transmission delay + transmission power cost.

Examples of Reinforcement Learning

Example 3: Millimeter Wave Communication

- *Observations*: The ACK/NACK (acknowledge signal in wireless communication) signals.
- *Actions*: The number of packets to transmit.
- *Cost (negative reward)*: Transmission Delay + Transmission cost.

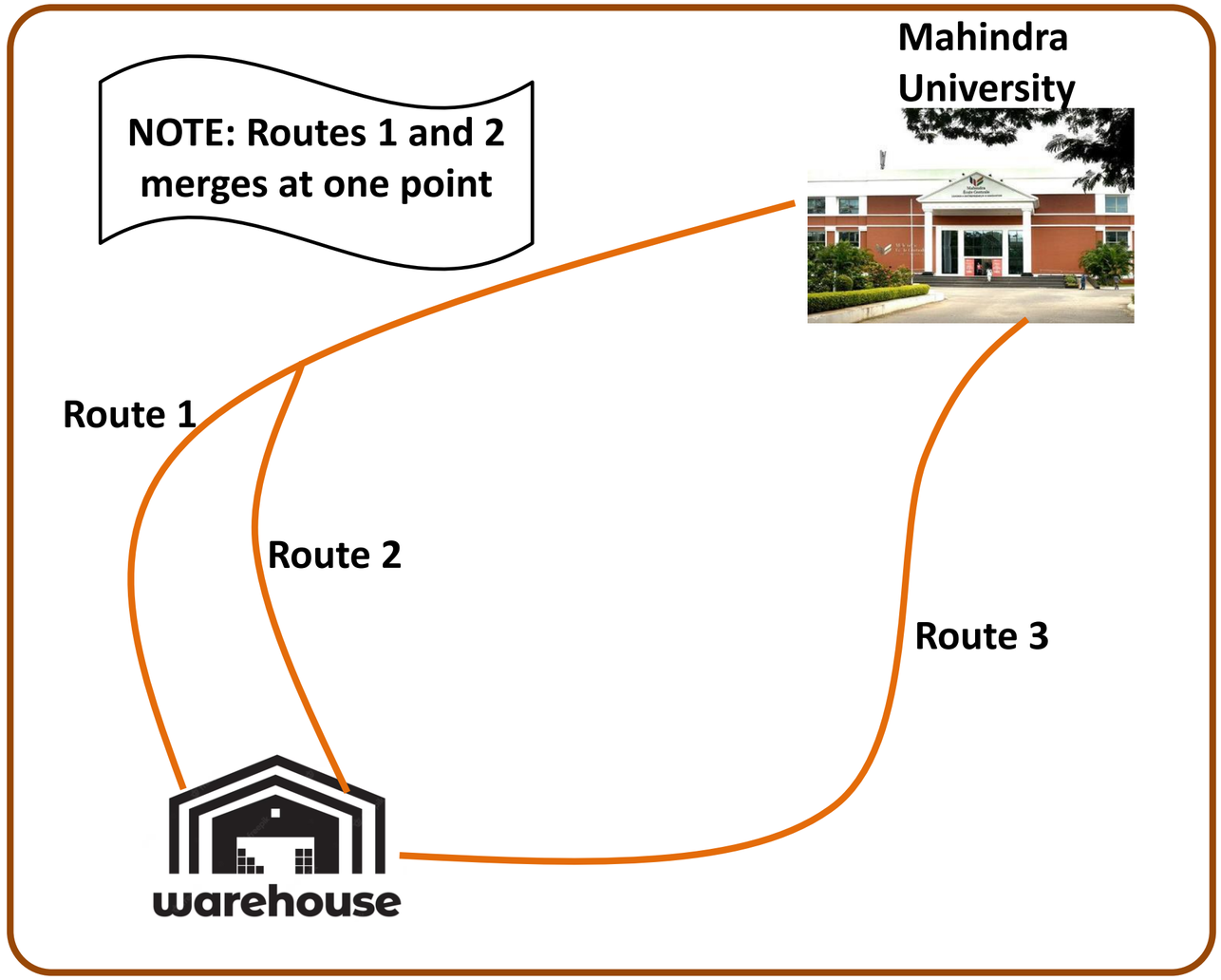
Lecture Content

- What is Reinforcement Learning (RL)?
- **Overview of the modules of this course (through examples).**
- Course logistics.
- Miscellaneous Topics.

Overview of the modules of this course

- Let us get an high level overview of Modules 1 – 4 of this course through a set of similar examples.
- These examples are in one way or the other related to “autonomous system”.
- Module 5 does not deal with Reinforcement Learning (though some of the topics covered in Modules 1 – 4 helps in understanding Module 5) and hence we will discuss it directly towards the end of the course.

Overview of the modules of this course

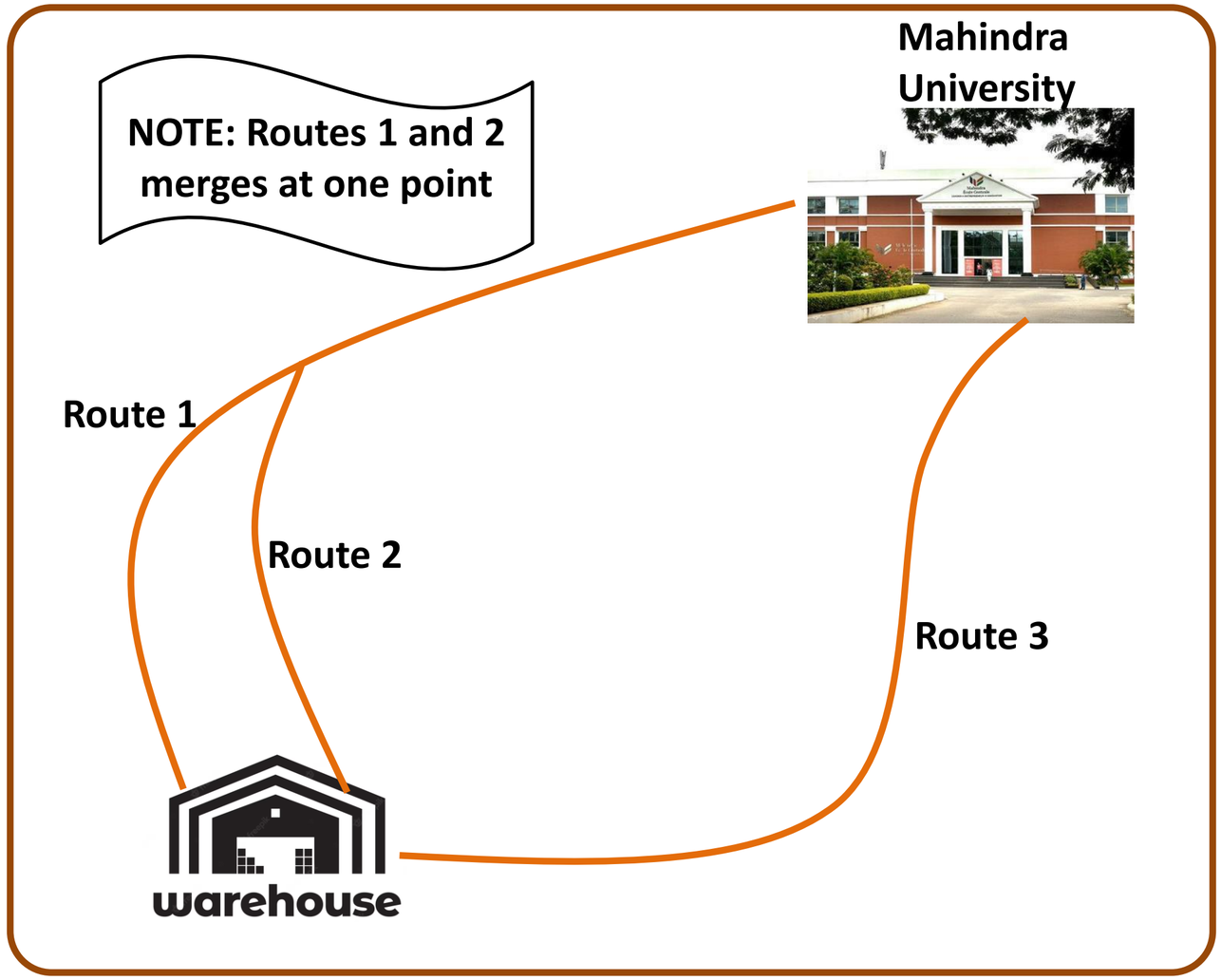


Module 1 (Multi-Armed Bandit)

Example 1:

- The warehouse (Amazon) has to dispatch a lot of autonomous vehicles to Mahindra University.
- There are three routes that the vehicle and take. The route is decided before a vehicle leaves the warehouse.
- Off course, conditions of the routes are uncertain. How should the warehouse decide which route to take in order to:
  1. Minimize fuel cost.
  2. Minimize travel time.

Overview of the modules of this course

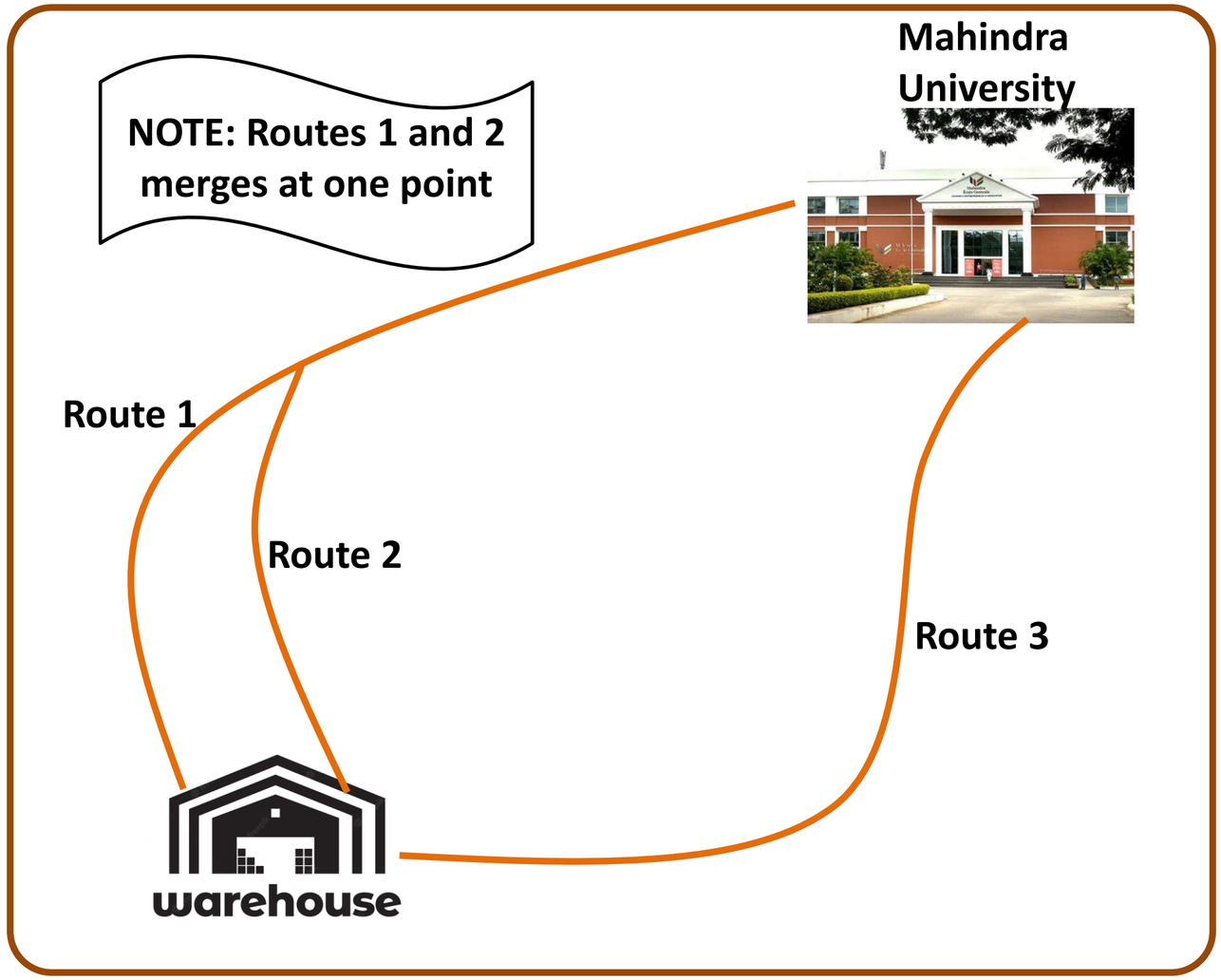


Module 1 (Multi-Armed Bandit)

Example 1:

- The warehouse (Amazon) has to dispatch a lot of autonomous vehicles to Mahindra University.
- There are three routes that the vehicle and take. The route is decided before a vehicle leaves the warehouse.
- IMPORTANT: The warehouse **does not know the probabilistic model of the uncertainty** that decides the fuel cost/travel time.

Overview of the modules of this course

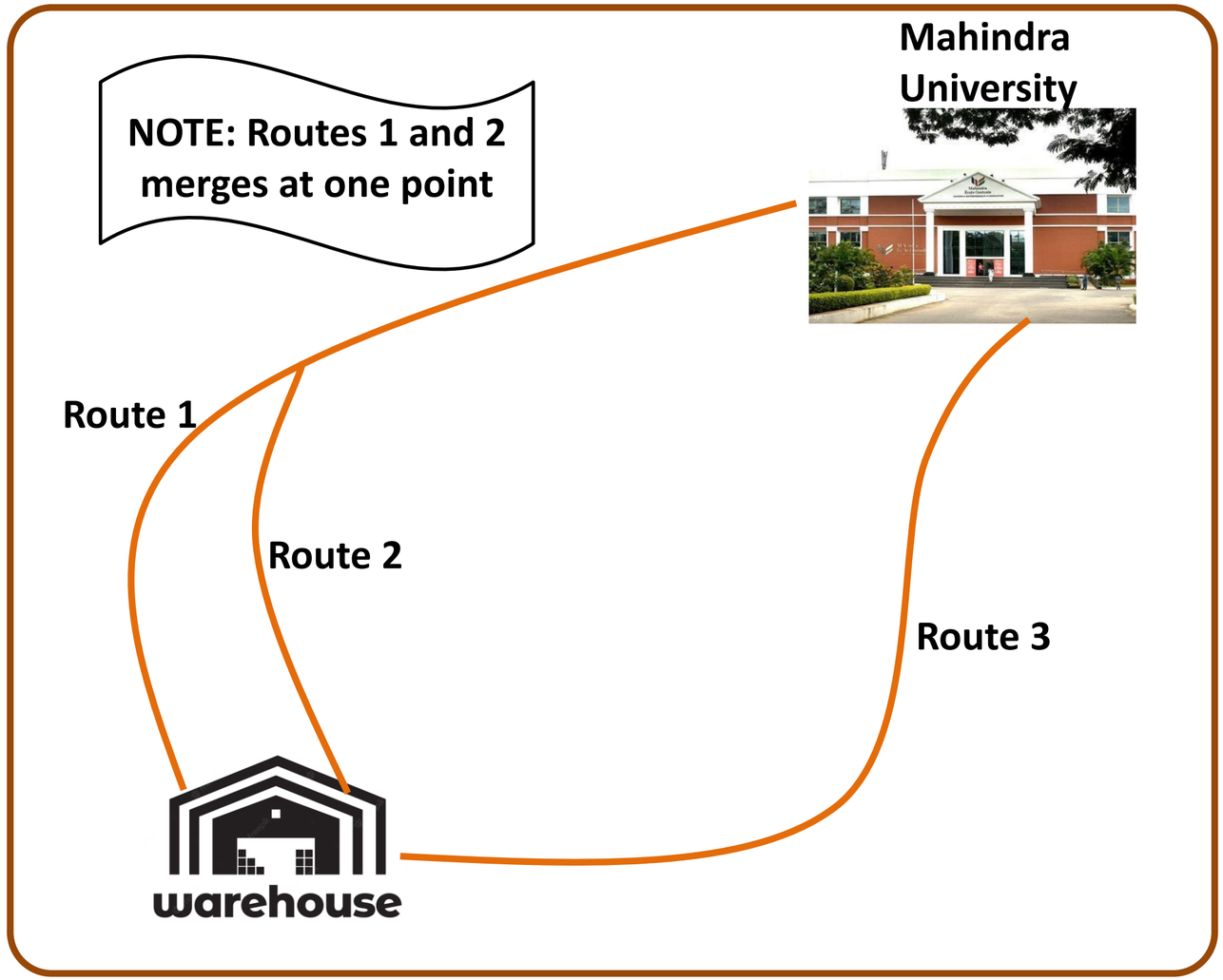


Module 1 (Multi-Armed Bandit)

Broad aspects of the solution to Example 1:

- The warehouse can try different routes and learn from experience.
- Eventually, it will get an idea of which of the three route has the best average fuel cost or travel time.
- IMPORTANT: The warehouse can learn the travel time of a route only if it takes the route. We cannot learn about a route that we are not travelling.

Overview of the modules of this course

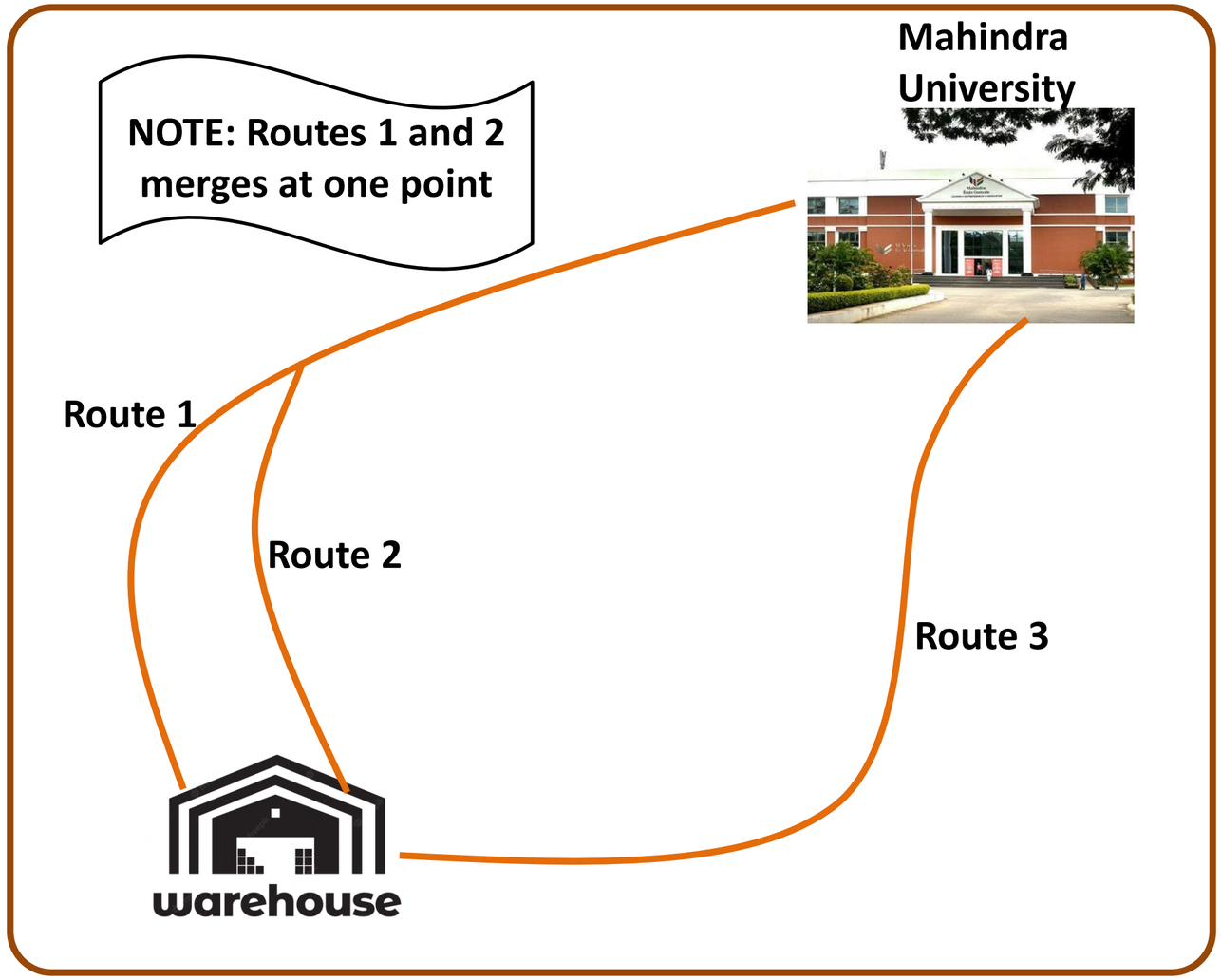


Module 1 (Multi-Armed Bandit)

Broad aspects of the solution to Example 1:

- The warehouse can try different routes and learn from experience.
- Eventually, it will get an idea of which of the three route has the best average fuel cost or travel time.
- SIDE NOTE: The two objectives (fuel cost and travel time) may not lead to the same optimal route.
  - E.x. Route 3 even through it is long (more fuel cost) can lead to the shortest travel time if it has less traffic.

Overview of the modules of this course

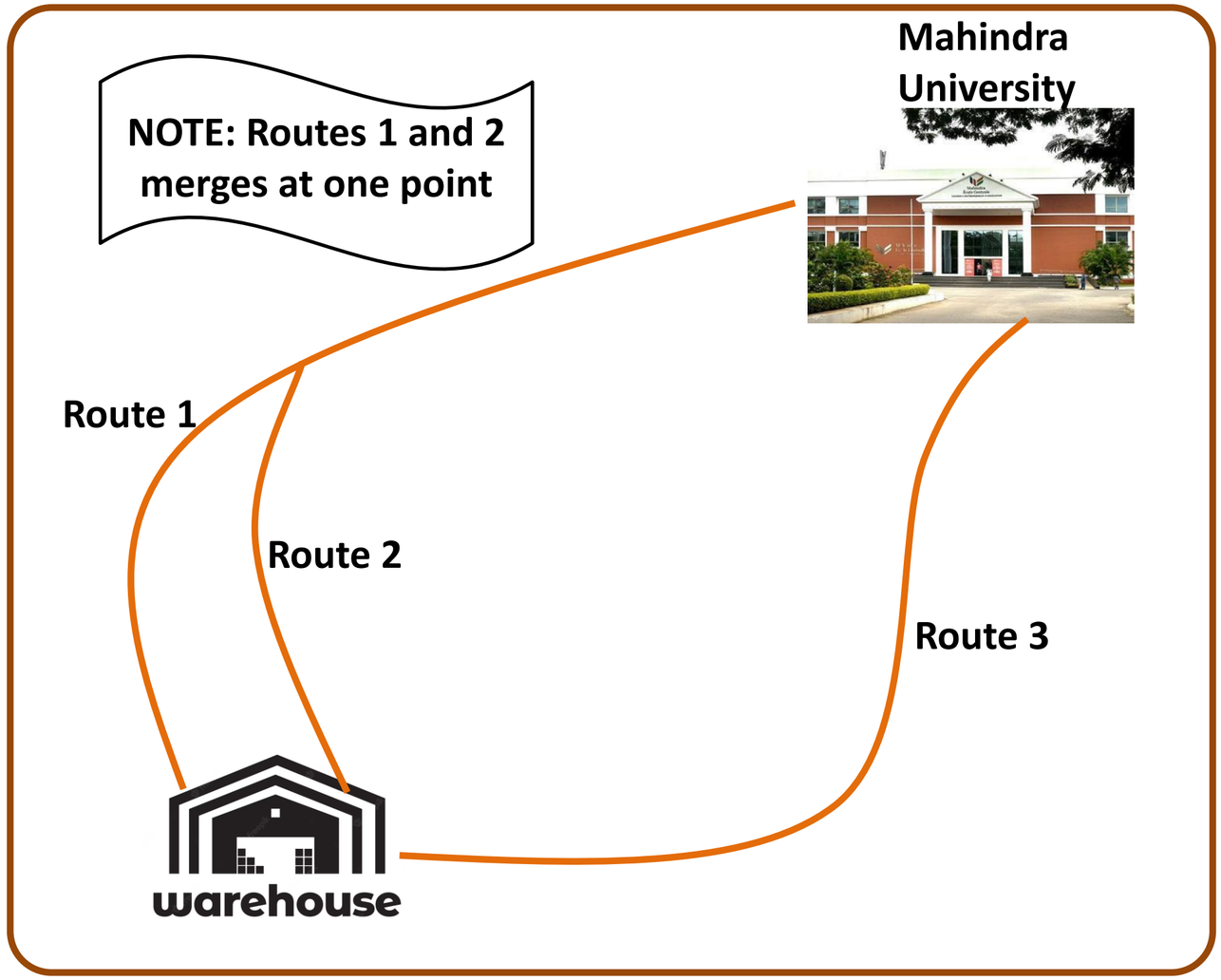


Module 1 (Contextual Multi-Armed Bandit)

Example 2:

- Same as Example 1 with the following modification:
  - The warehouse has two additional information to decide which route to take:
    1. Average travel time of vehicles dispatched in the **previous hour**.
    2. Average travel time of vehicles dispatched in the **current hour of the previous day**.

Overview of the modules of this course

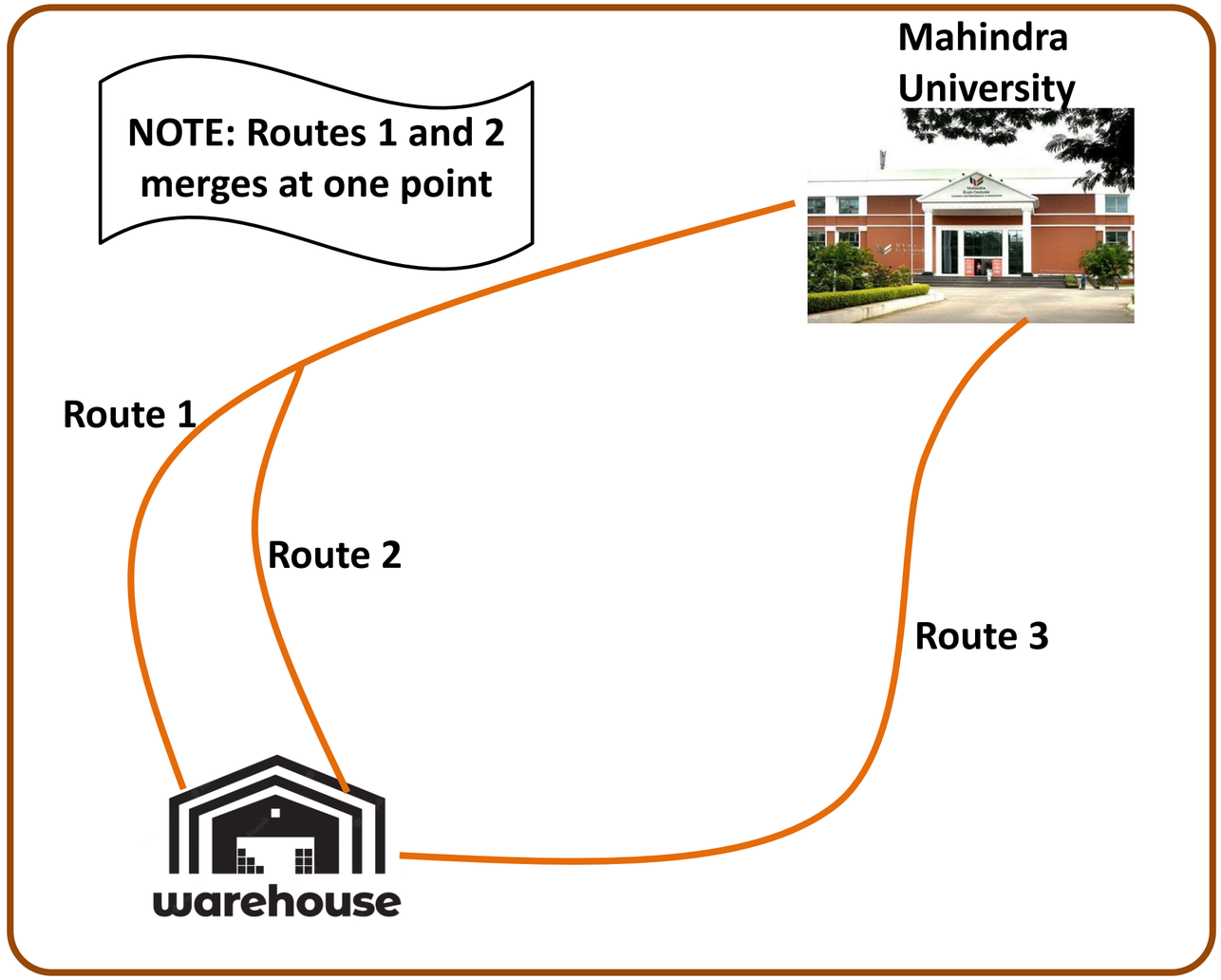


Module 1 (Contextual Multi-Armed Bandit)

Broad aspects of the solution to Example 2:

- The two additional information, **the context**, should intuitively help the warehouse to make better routing decisions.
- However, uncertainty of the three routes still exists in this problem as in Example 1. So, the warehouse still has to learn from experience.

Overview of the modules of this course

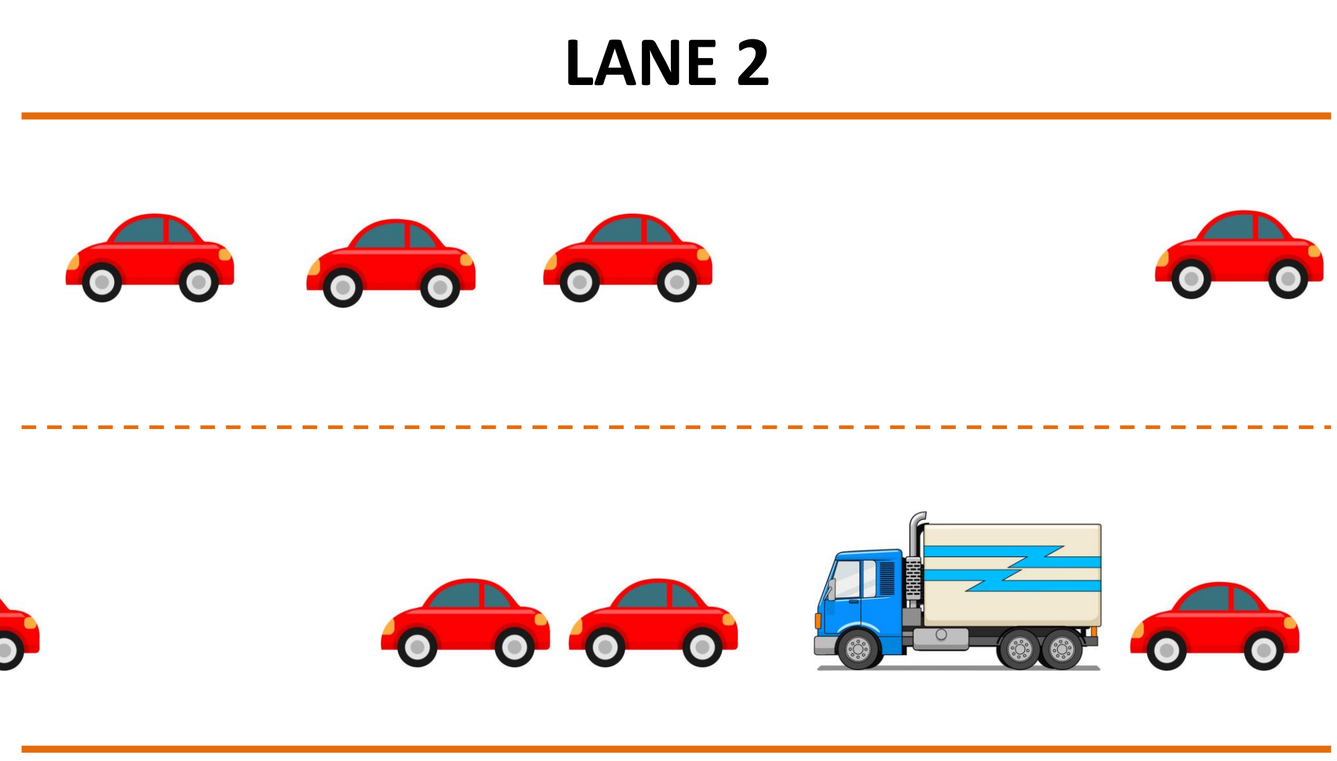


Module 1 (Contextual Multi-Armed Bandit)

Hey! This is just a **forecasting problem**. What is so new about that?

- Yes. But to train a forecasting model, we need data. Here, we have to collect data by traveling a route (hence **online learning**).
- Similar to Example 1, we can collect data about a route only by travelling it.

Overview of the modules of this course

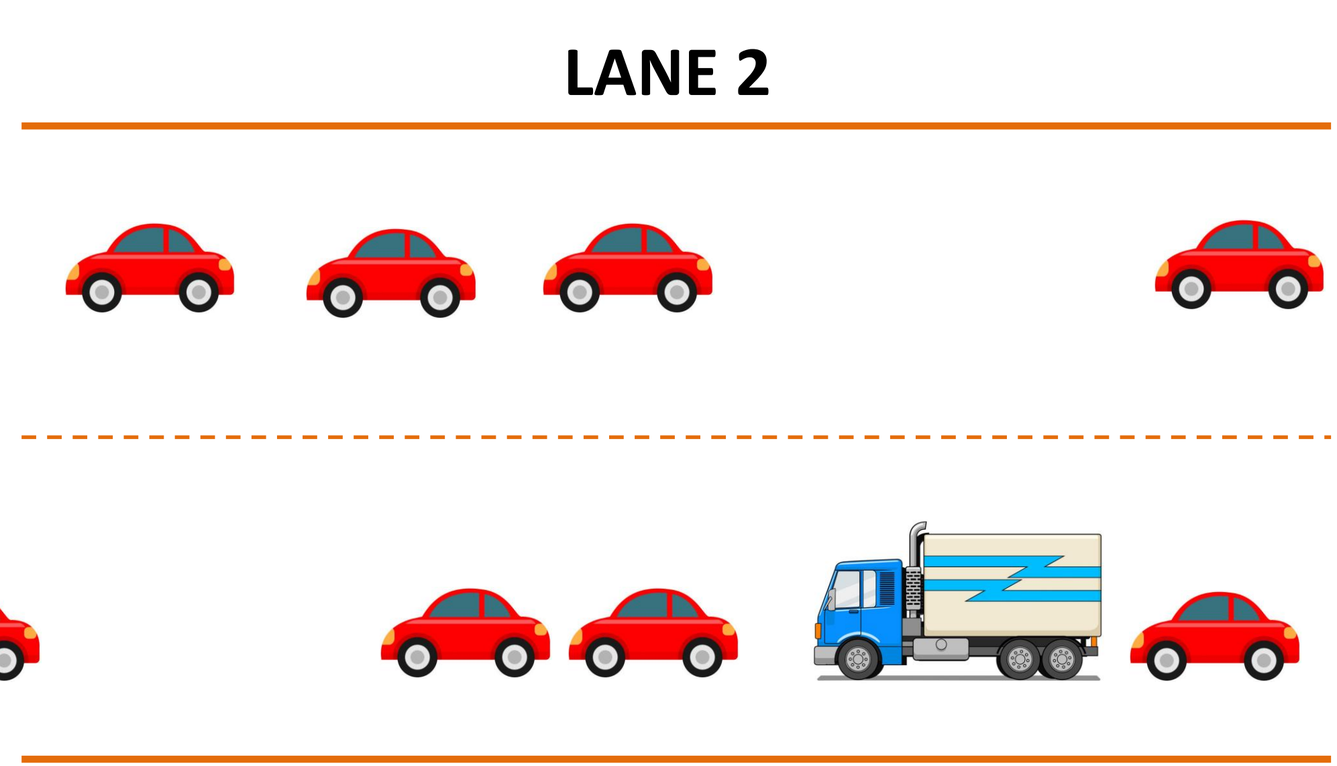


Module 2, 1st half (Markov Decision Process (MDP))

Example 3:

- Suppose the warehouse's vehicle has been dispatched in a given route.
- The job of the autonomous system of the warehouse vehicle is to decide whether to change lane or not.
  - Other aspects of driving like controlling the brakes and accelerator are done by humans.
- Objective is same as Example 1. Either minimize fuel cost or travel time.

Overview of the modules of this course



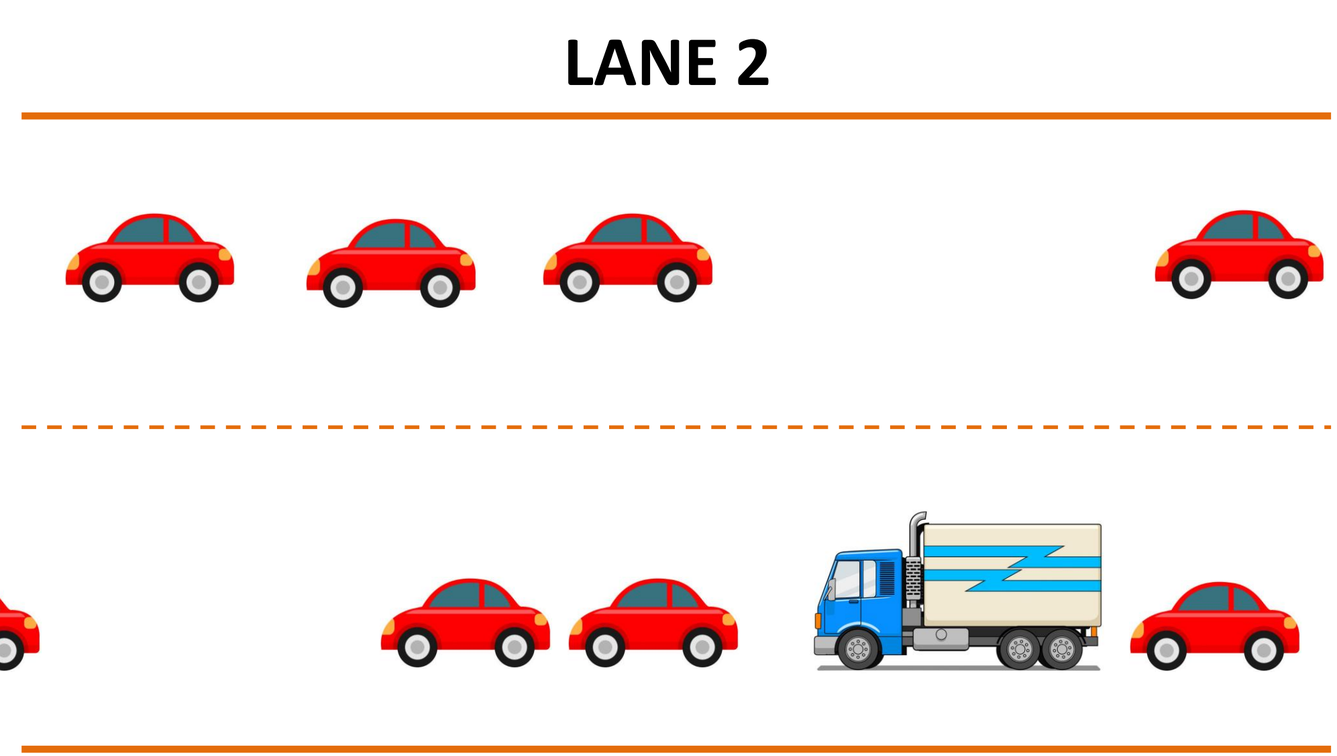
Module 2, 1st half (Markov Decision Process (MDP))

Example 3:

- Uncertainties associated with the system: How the other vehicles will react to lane change. Also, how the human driver in the warehouse vehicle will react by changing its speed.

- IMPORTANT: For MDP, the **autonomous system knows the probabilistic model of the uncertainty**. Hence, a planning problem NOT a learning problem.

Overview of the modules of this course



Module 2, 1st half (Markov Decision Process (MDP))

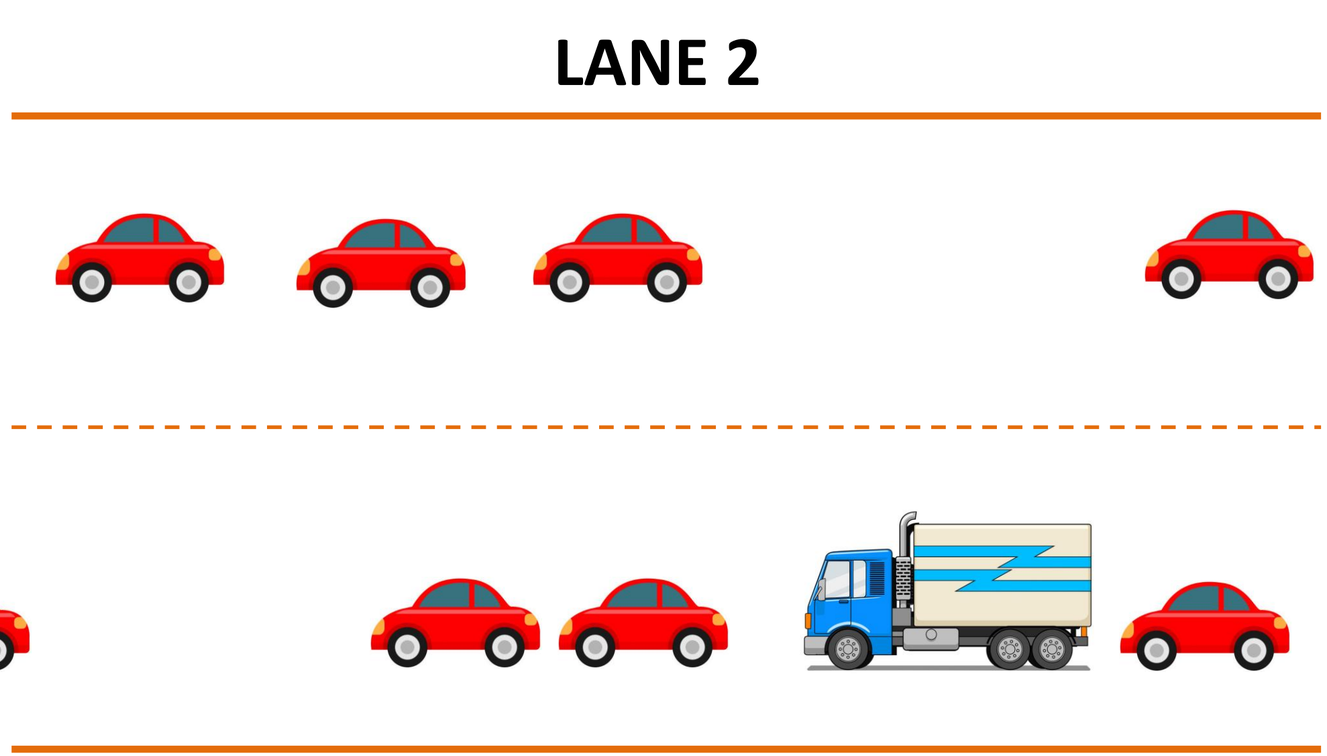
Broad aspects of the solution to Example 3:

- The **state*** of the "environment" are:
  - x - coordinate of all the vehicles.
  - The lane of all the vehicles.
  - The velocity of all the vehicles.

- We assume that these states are available to the warehouse vehicle through GPS and VANET (vehicular adhoc network).

* We will learn more about **states** in the next lecture but for now we can think of states as a set of parameters that sufficient captures how the environment evolves over time.

Overview of the modules of this course



Module 2, 1st half (Markov Decision Process (MDP))

Broad aspects of the solution to Example 3:

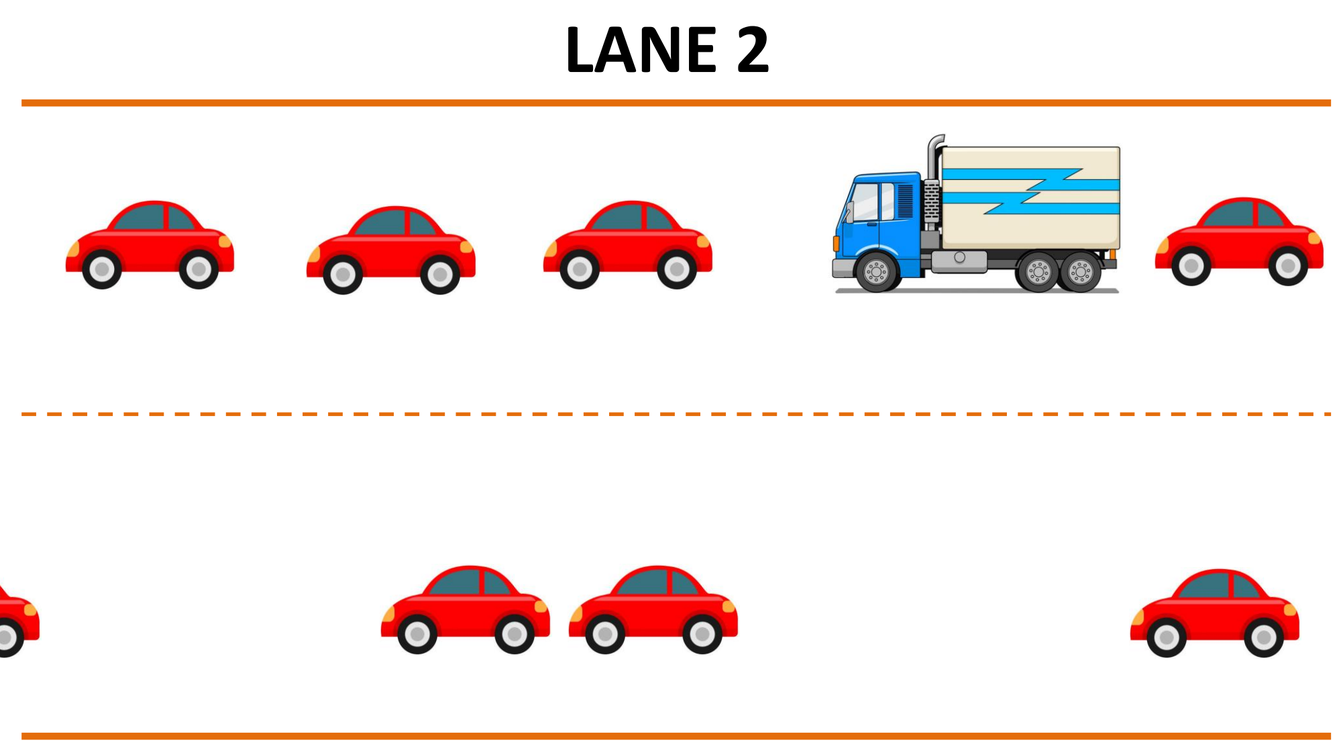
- Just because the:

  1. probabilistic model of uncertainty, and
  2. all the states are available

  it does not mean the deciding the optimal action is an easy task. This is because.....

(PTO)

Overview of the modules of this course

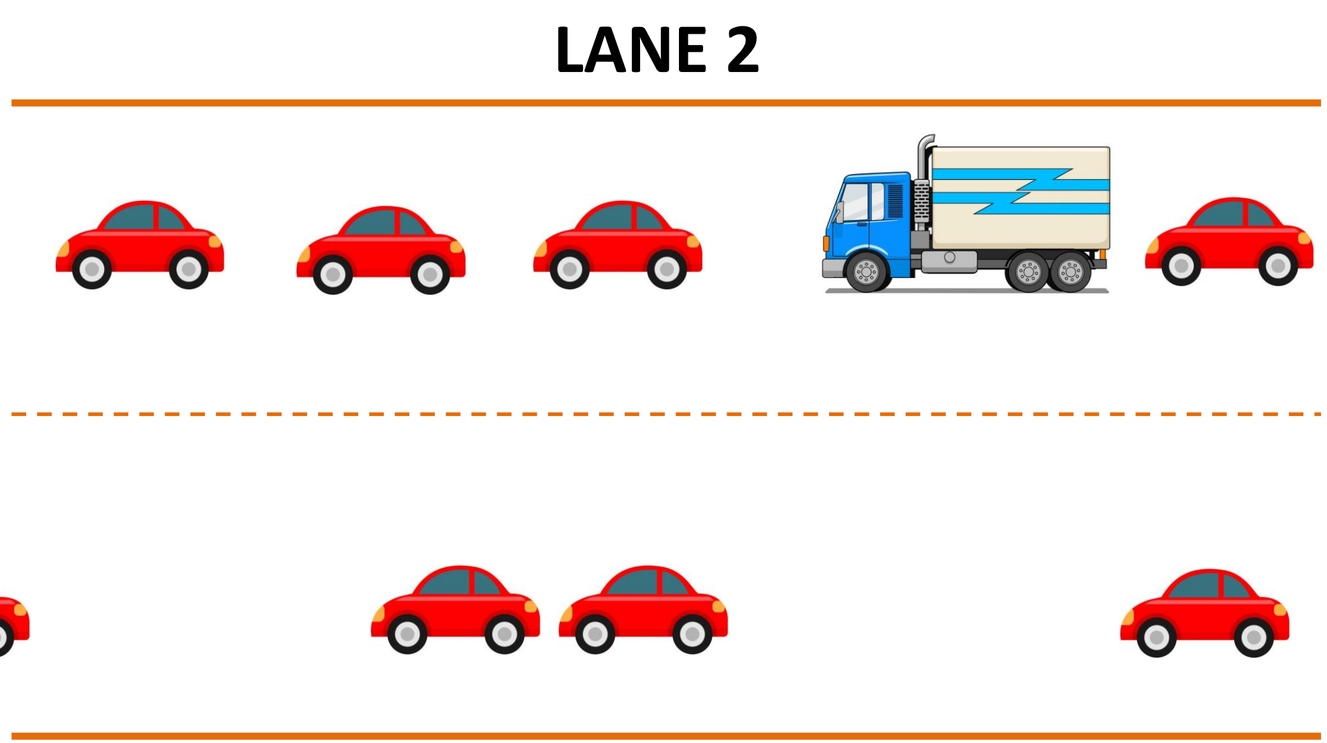


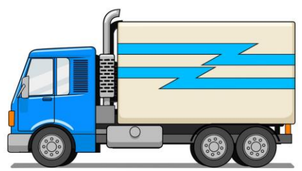
Module 2, 1st half (Markov Decision Process (MDP))

Broad aspects of the solution to Example 3:

- IMPORTANT: Depending on the action of the warehouse vehicle (change lane or not), the states will change. E.x. If the vehicle changes to lane 2:
  - The velocity of the warehouse vehicle may decrease (because a vehicle has to reduce its speed a little during lane change).
  - The speed of **this vehicle** may also reduce.

Overview of the modules of this course





Warehouse vehicle

Other vehicle

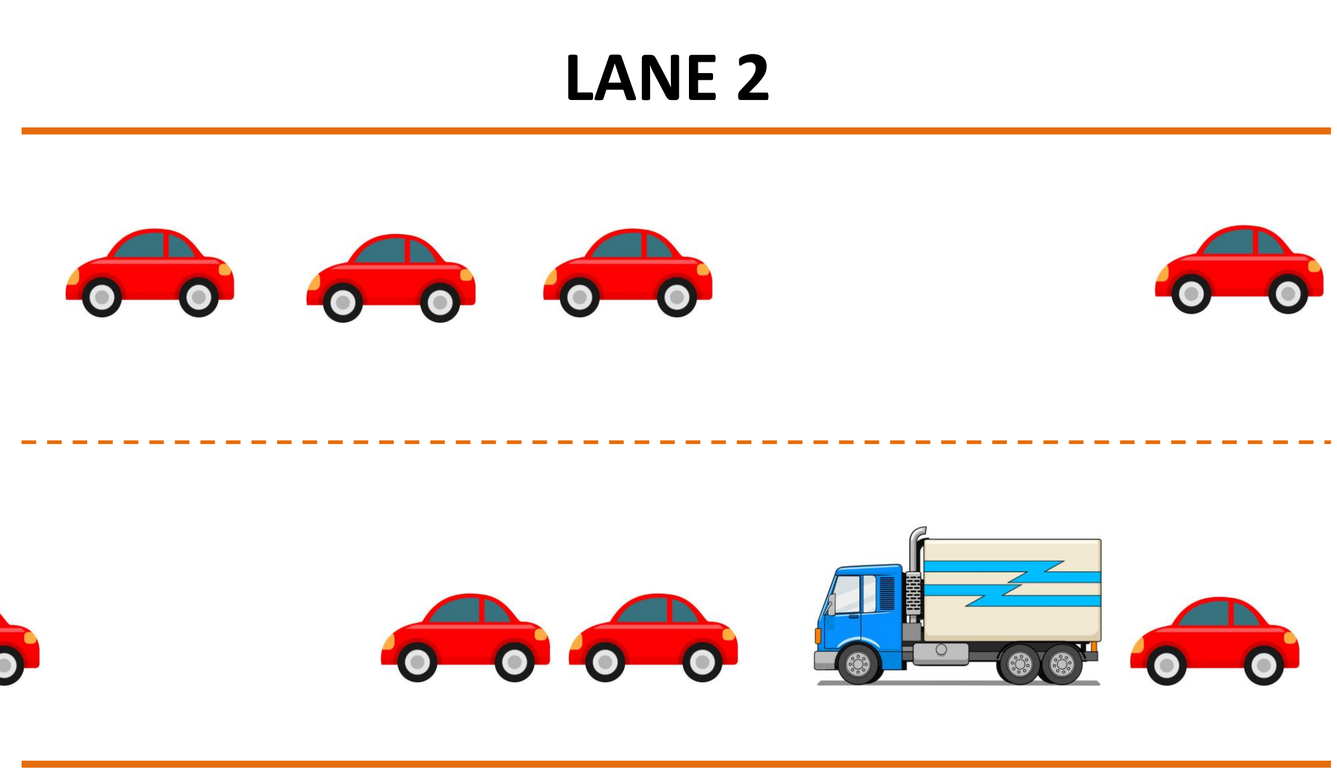
Module 2, 1st half (Markov Decision Process (MDP))

Broad aspects of the solution to Example 3:

- IMPORTANT: Depending on the action of the warehouse vehicle (change lane or not), the states will change. Ex: If the vehicle changes to lane 2:
  - The velocity of the warehouse vehicle may decrease (because a vehicle has to reduce its speed a little during lane change).
  - The speed of this vehicle may also reduce.

This is unique to Example 3 that we didn't see in Examples 1 and 2, i.e. in Examples 1 and 2, the action of the agent didn't change the environment.

Overview of the modules of this course

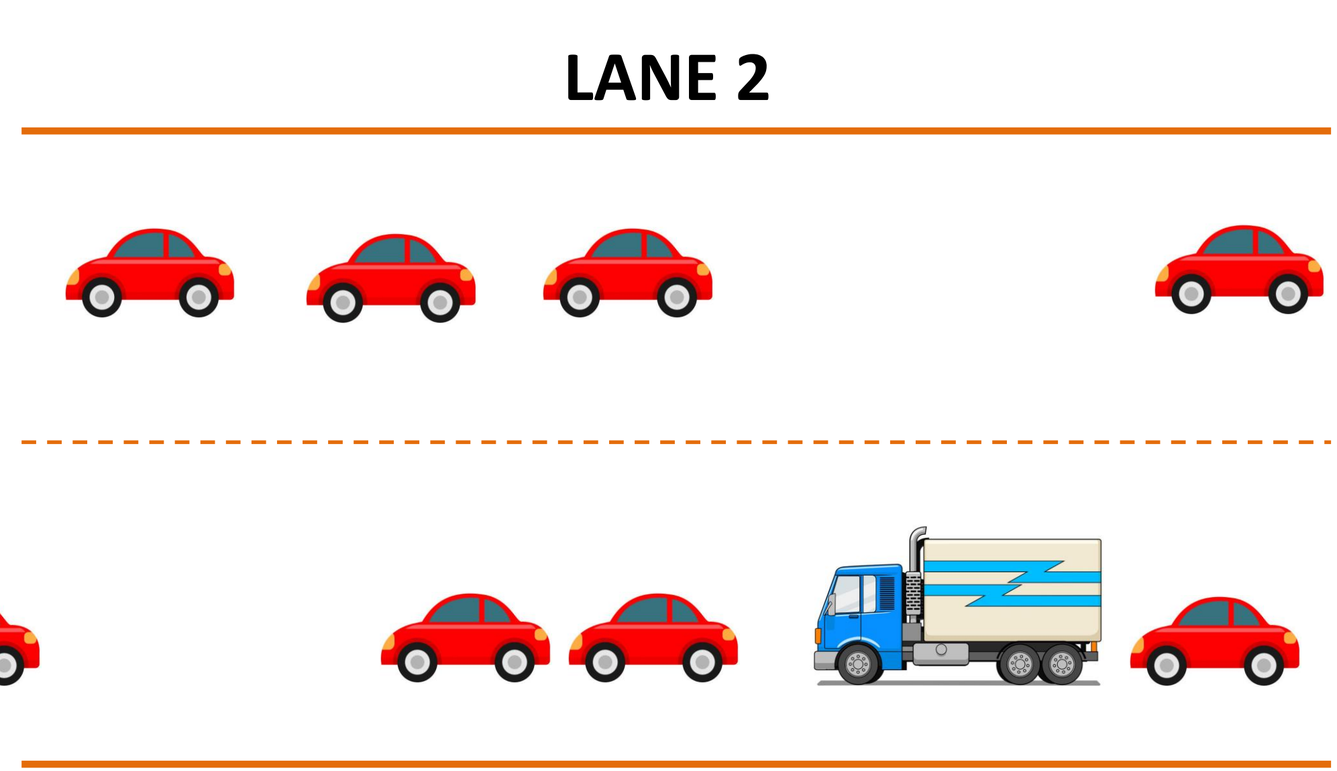


Module 2, 1st half (Markov Decision Process (MDP))

Broad aspects of the solution to Example 3:

- If the environment changes the future rewards (that depends on the environment) also changes.
- Hence, while making a decision, the autonomous system SHOULD NOT employ a "greedy strategy". E.x. in this case even through lane 2 seems less jammed from where the warehouse vehicle is, it is in fact more jammed compared to lane 1 in the long run.

Overview of the modules of this course



Module 2, 2nd half and Module 3 (Reinforcement Learning)

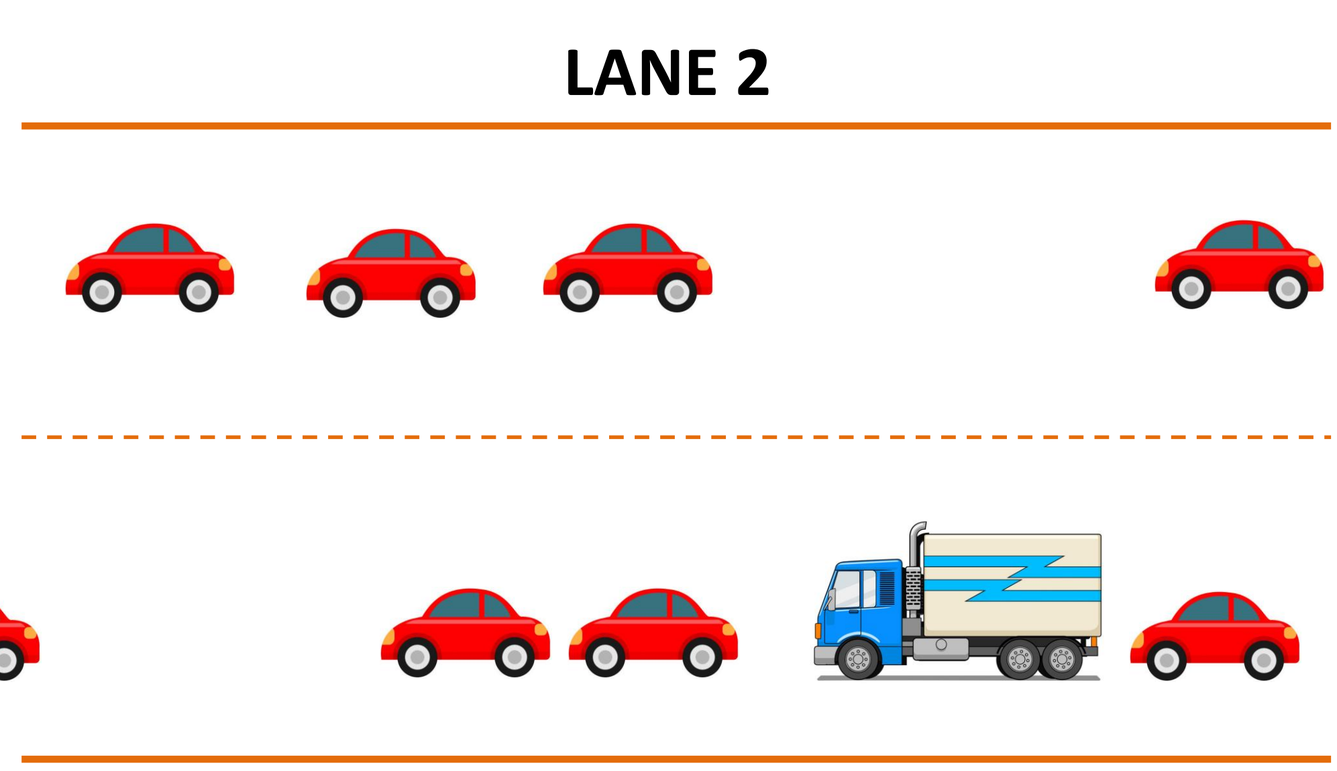
Example 4:

- Same as Example 3 with the following modification:
  - The autonomous system **DOES NOT know the probabilistic model of the uncertainty**. Hence, it is both a planning and a learning problem.

Broad aspects of the solution to Example 4:

- The broad aspects for this example is **same as that of Example 3** with the additional component that the autonomous system has to **learn** the optimal strategy.

Overview of the modules of this course

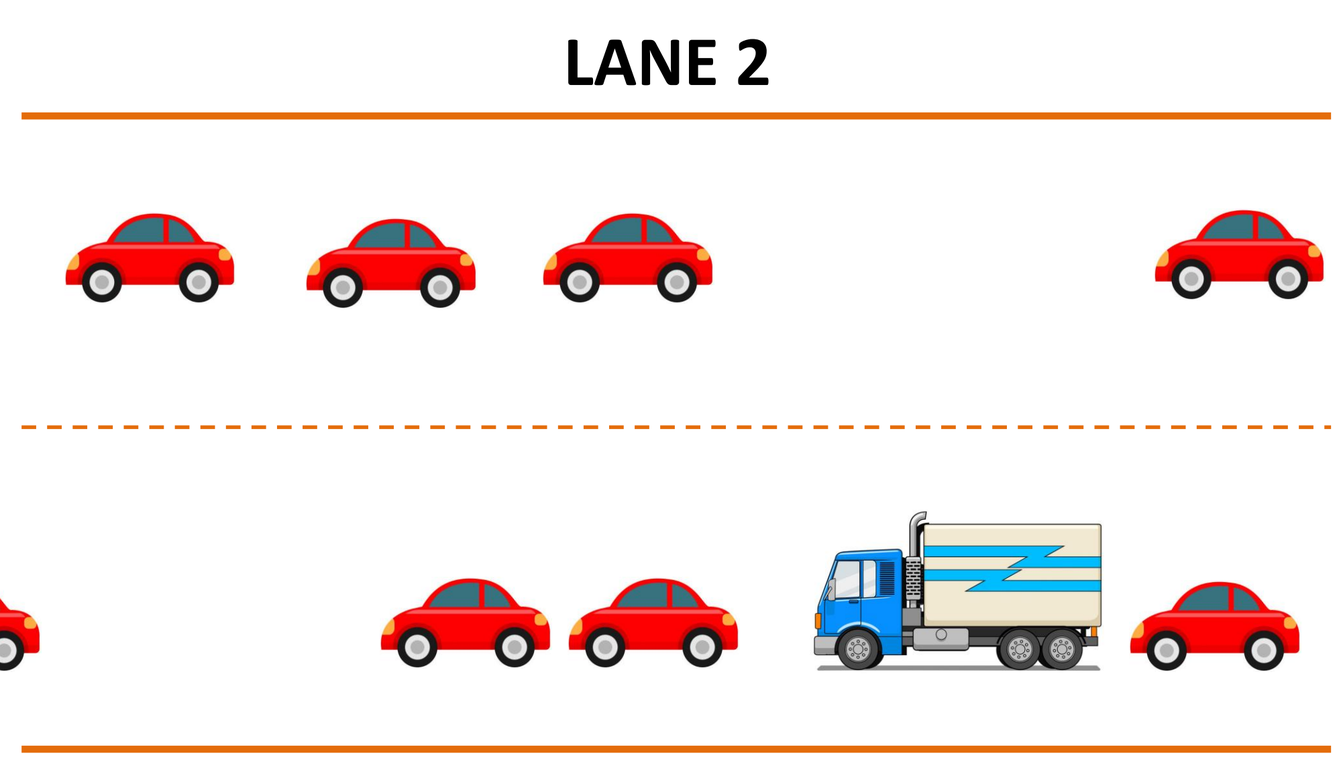


Module 4 (Deep Reinforcement Learning)

Example 4:

- NOTE: We are dealing with Example 4 (same as before and without any change).
- As mentioned while discussing Example 3, the warehouse vehicle knows the states of all the cars in the road through GPA and VANET.

Overview of the modules of this course



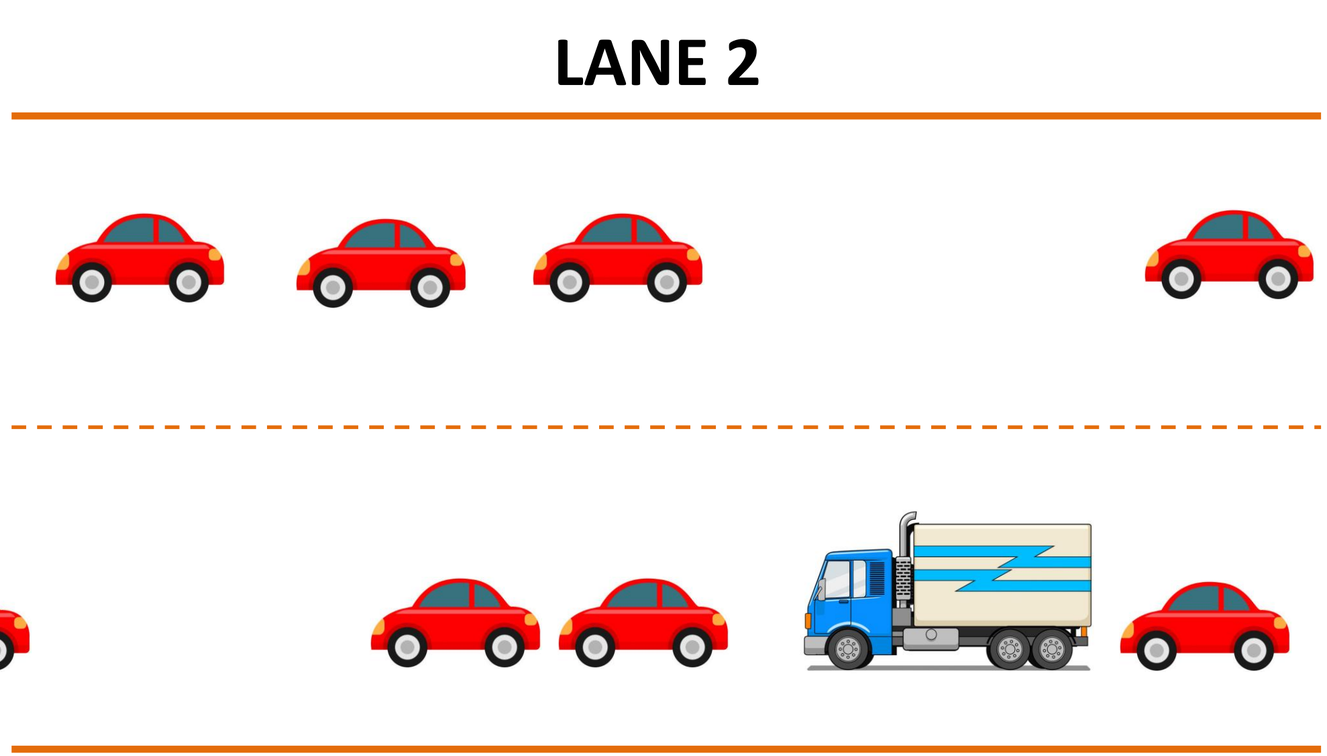
Module 4 (Deep Reinforcement Learning)

Example 4:

- **Question: How many car's states does the warehouse vehicle have to know in order to make a "sufficiently good" decision?**

- This question is important because while the state of a car that is 5 kms ahead may **improve the decision a very little bit**, it will definitely **increase the computational complexity by a significant amount** if the warehouse vehicle has to consider the state of all the cars in 5 kms radius.

Overview of the modules of this course



Module 4 (Deep Reinforcement Learning)

Example 4:

- This is where Deep Reinforcement Learning (DRL) is useful because Neural Networks are very good in doing **automatic feature extraction** for the required task. Here, the features are the "useful states"* and task is to optimize the objective function (fuel cost or travel time).

* Or a function of states that are useful, e.x. the relative distance between cars is a function of two states. While for this example it is easy to accept that relative distance is an useful state, in many situations we may not be able to engineer an useful state (this is same a engineering a new feature in supervised learning). Neural Networks are very good at this task.

Overview of the modules of this course

Module 4 (Deep Reinforcement Learning)

Example 5:

- In this example the observation of the warehouse vehicle is an **unstructured data**; like an image (rather than GPS data through VANET in Example 4).

- The need for automatic feature extraction becomes even more relevant here because if an image is 1028*1028 with 256-bit color, then the state space* is $(1028 \cdot 1028)^{256}$ which is humongous!

* We will learn about state space in Module 2.

Overview of the modules of this course

Module 4 (Deep Reinforcement Learning)

Example 5:

- As we know, CNNs are very good in doing automatic feature extraction for image.

Overview of the modules of this course

Module 4 (Deep Reinforcement Learning)

Example 5:

- As we know, CNNs are very good in doing automatic feature extraction for image.

- A more interesting fact that further justifies the benefit of CNNs (and hence DRL) are the **street signs** because they can give us useful information about flow of traffic. While a human can easily ignore such components of an image while doing "manual" feature extraction, CNNs may not.

Lecture Content

- What is Reinforcement Learning (RL)?
- Overview of the modules of this course (**through examples**).
- **Course logistics.**
  - **Prerequisites**
  - **References**
  - **Marks Distribution**
  - **Office hours**
- Miscellaneous Topics.

Prerequisites

- Probability (strong knowledge).

- Linear algebra (not much)
  - Should at least know how to do matrix multiplication  .
  - Matrix and vector norms (triangle inequality).

- Machine Learning/Deep Learning (required for Module 4).
  - **Who does have any experience with Deep Learning? Email me. I will send you a tutorial.**

References

- Books:

1. Reinforcement Learning: An Introduction, Sutton and Barto, 2nd Edition.
   Available online for free

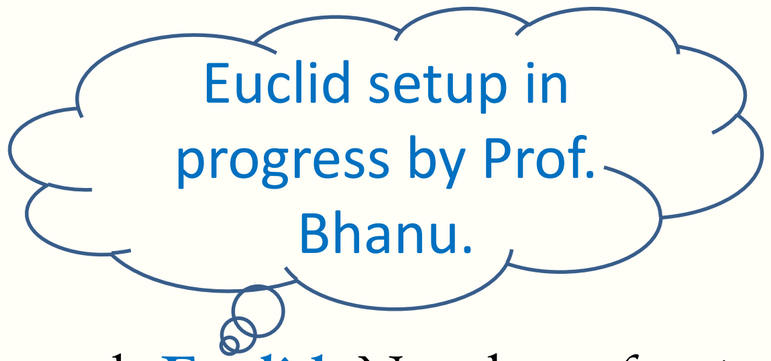
- Video lectures:

1. NPTEL, Indian Institute of Technology Madras, Prof. Balaraman Ravindran.
2. CS 234, Stanford University, Prof. Emma Brunskill.
3. CS 285, University of California Berkeley, Prof. Sergey Levine.
4. CS 885, University of Waterloo, Prof. Pascal Poupart.

- My lecture notes will be available in Dropbox.
  - Will share the link soon after I have included everyone's email id in the group email list.

Marks Distribution

- Programming assignments (40%). **In groups of 3-4. Will send an email about this.**
  - Module 1 (10%).
  - Module 2 (15%).
  - Module 3 (10%).
  - Module 4 (15%).
  - If there is any assignment for Module 5, then it will be **extra marks**.
  - **Academic dishonesty will be heavily penalized. No chance for apologies!**
- Minor 1 (5%).
- Minor 2 (15%).
- Final Exam (30%).



- **Surprise quiz** (10% **extra marks** in total) through **Euclid**. Number of quizzes is unknown!
  - **Please bring your laptop, pen/pencil, paper everyday for the surprise quiz.**

Office Hours

- There will be **one hour of office hours every week**.
  - Will be decided by google form survey.
  - Will conduct the survey as soon as I have everyone's email id in the group email list.

- You can also come to my cubicle by booking **prior appointments** mentioning the **topic of discussion in brief**. (so that I have some time to prepare).

Lecture Content

- What is Reinforcement Learning (RL)?
- Overview of the modules of this course (**through examples**).
- How is RL different from other AI topics?
- Course logistics.
- **Miscellaneous Topics.**
  - **Module 3 by Dr. Neel.**
  - **Mu Email Id.**
  - **Discussion about preliminary survey.**

Module 3 by Dr. Neel

- Module 3 of this course will be handled by **Dr. Gone Neelakantam**.
- His research work is related to the **application of reinforcement learning**, e.x. **smart cities**.
- Any expectations regarding Module 3 should be directly discussed by Dr. Neel.

Discussion about Preliminary Survey

- There are three mentions of something to do with **designing rewards**. **Inverse reinforcement learning** (Module 3) is one way of doing it.

- There is a mention of **Proximal Policy Optimization**. I will cover that while dealing with Module 4.

- There is a mention of **Reinforcement Learning from Human Feedback**.
  - Will try to incorporate it as part of **Programming Assignment for Module 4**. **No Guarantees! Gathering datasets and cleaning them can be a huge issue.**
  - NOT from the perspective of NLP though (I have zero knowledge about NLP).

- Coding in Python and Openai Gym.
  - You can use Tensorflow, Keras, or Pytorch (I don't know Pytorch). But either way, as far as programming assignment for Module 4 is concerned, I will simply run the code and check its performance.

MU Email Id

- I have created a **group email**. If you did not receive any email from me, especially **Masters/PhD students**, please email me your **Roll Number** and **MU Email Id**.

My email: **gourav.saha@mahindrauniversity.edu.in**

Discussion about Preliminary Survey

- There has been a request for **live coding** during the final survey of Control Theory course.
  - **3-4 live coding lectures**.
  - One live coding lecture for every module (maybe two live coding lectures for Module 2).

- There was some comment like:

  "please don't do the "have to build models or learn how to do them without libraries too"... Let's please do it the way one would at a job."

  - I have never worked in corporate  . Don't know what is required in such a setting.
  - By guess is that in an ever changing world, it is more strategical to have a lot of **"transferable skills"** to increase your opportunities.
  - While learning how a library works, you will learn many such transferrable skills by knowing how an algorithm works rather than simply using it. Such knowledge will have applications beyond Reinforcement Learning.
  - So here is the deal:
    - ✓ For Module 1 and 2, you will be coding the algorithm.
    - ✓ For Module 4, you will use pre-existing libraries.


Thank you