Contextual Bandits

- Treat each movie as an arm, can be extremely large
- For each round t:

  observe a context x_t for the user arrived

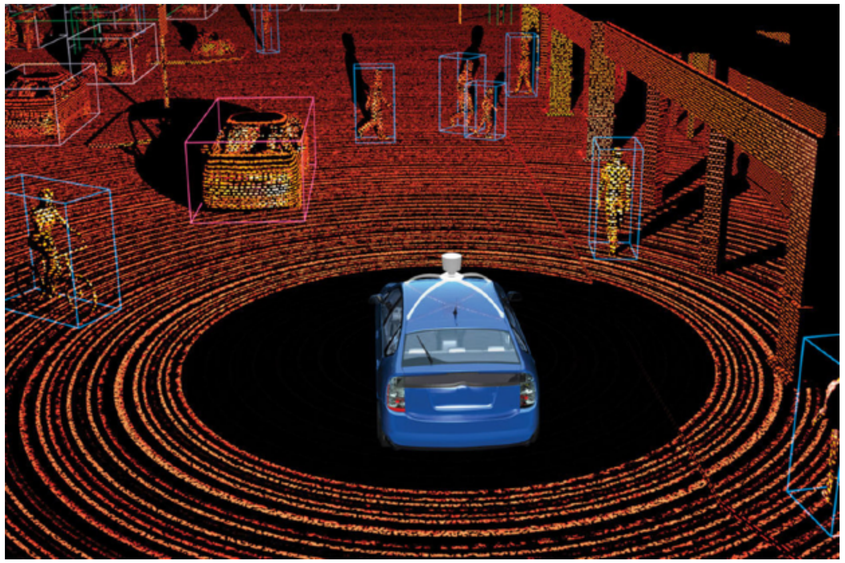
  pick an arm a_t i.e. recommend a movie

  receive a reward $r_t(x_t, a_t)$ which is higher when a user likes a movie

Goal: Given any sequence of context $\{x_1, x_2 \ldots, x_T\}$ or a sequence of users, find an algorithm which maximises the total reward $\sum_{t=1}^{T} r_t(x_t, a_t)$

Reinforcement learning: An example

- Suppose you want to learn how to drive a car such that it follows a lane and minimises the number of collisions



- You observe the state s of the surroundings: which can consist of very high dimensional data like LIDAR, images, etc.
- You take an action a = [steering angle, acceleration]
- Observe the next state $s' \sim \mathbb{P}(\cdot \,|\, s, a)$ and reward $= \mathbb{I}\{\text{on the lane}\} - \#\text{no of collisions}$
- Objective: Design a policy, which is a mapping from state to action e.g. a neural network, such that the reward over T rounds is maximized