Introduction to Reinforcement Learning

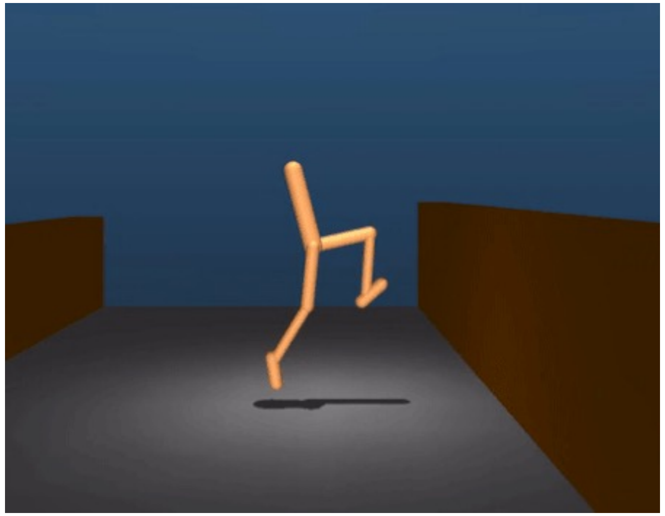
David I. Inouye

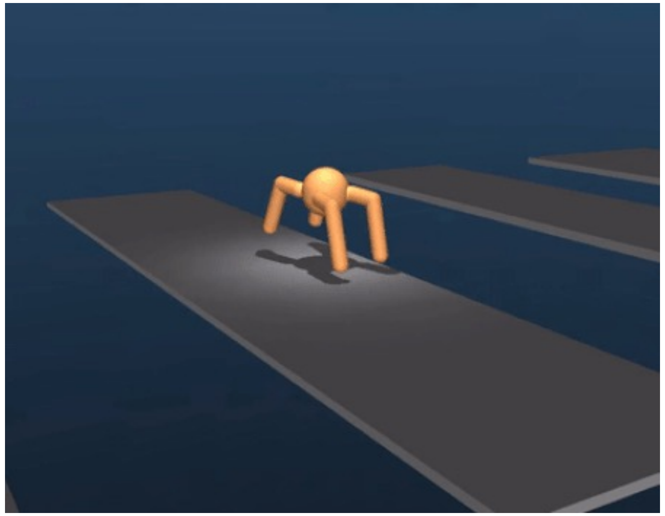
PURDUE UNIVERSITY®

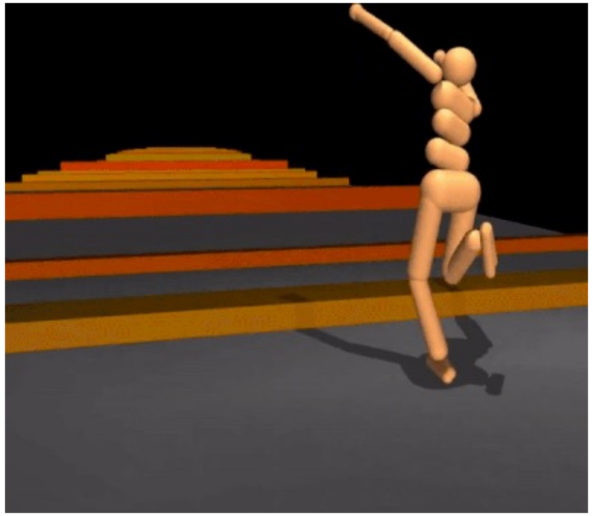
Elmore Family School of Electrical and Computer Engineering

How can we formalize the problem of "learning to walk"?

- Babies seem to learn to walk by just trying it multiple times until they learn to control their bodies.
- Given only some sensors like touch (pressure sensor) and the goal of moving forward, can an agent learn to walk?







Heess, N., TB, D., Sriram, S., Lemmon, J., Merel, J., Wayne, G., ... & Silver, D. (2017). Emergence of locomotion behaviours in rich environments. *arXiv preprint arXiv:1707.02286*.
https://arxiv.org/abs/1707.02286

How can we formalize the problem of “learning optimal oil refinery configuration”?

- How do we configure each component to produce the most refined oil?
- Especially in dynamic situations where operating conditions change

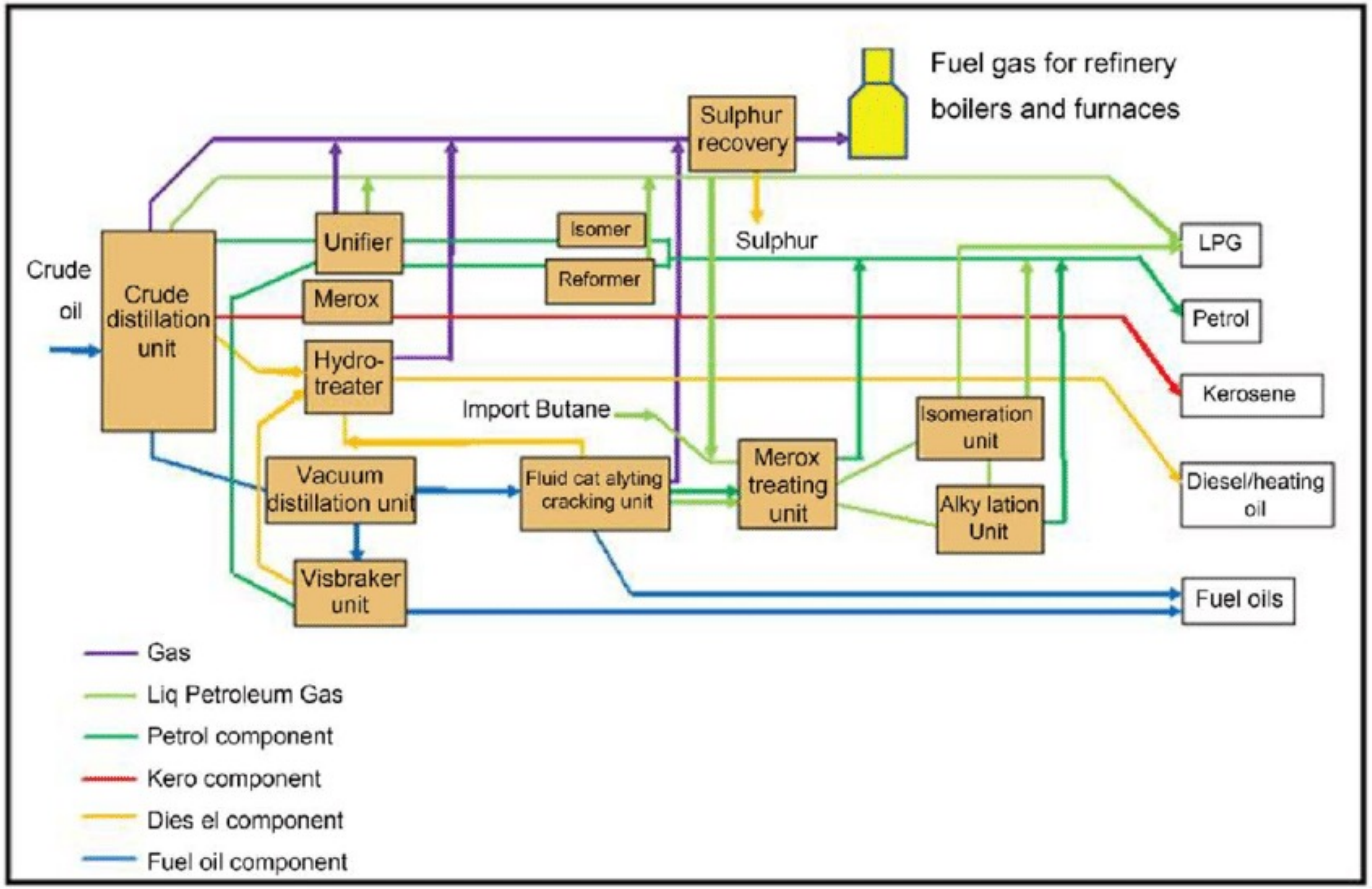
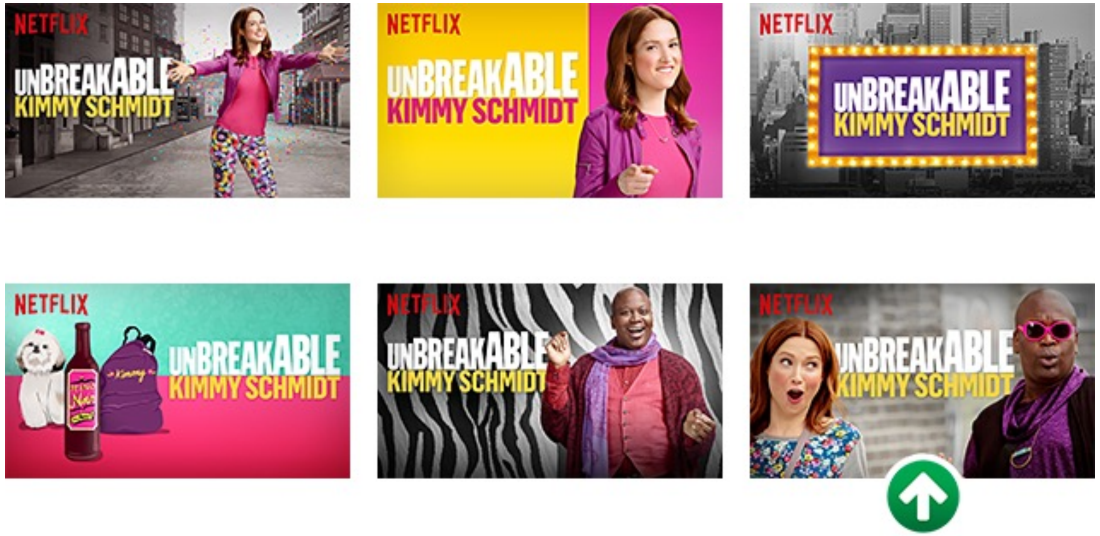
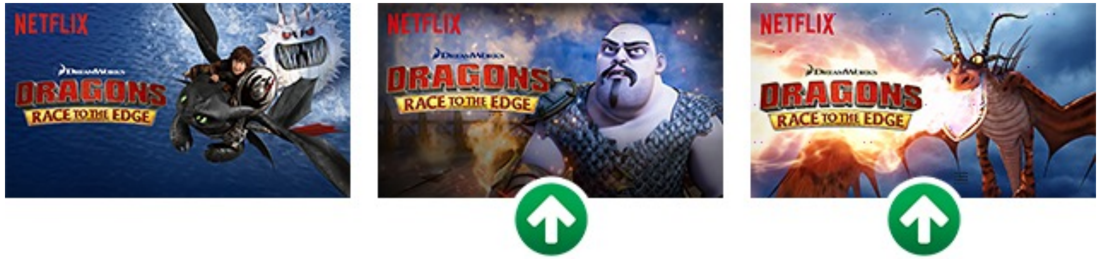
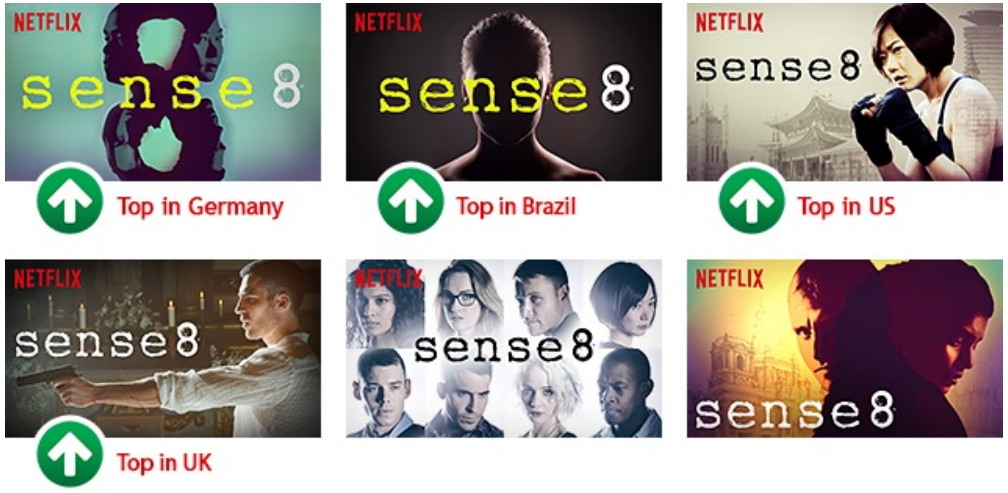


Figure from: Alanbari, M., Rahman, I., Al-Ansari, N., & Knutsson, S. (2016). Comparison of potential environmental impacts on the production of gasoline and kerosene, Al-Daura refinery, Baghdad, Iraq. *Engineering*, *8*(11), 767-776.

How can we formalize the problem of “learning optimal Netflix images”?







Figures from https://medium.com/@guy_leb anon/image-optimization-at-netflix-287dcc74df17

https://netflixtechblog.com/art work-personalization-c589f074ad76

Reinforcement learning provides the formal framework to analyze each of these problems

- What is reinforcement learning?
  - "Goal-directed learning from interaction" (Sutton & Barto, 2018)
- Learning
  - The system should adapt to new situations and learn from the past
- Goal-directed
  - The learning process has an ultimate desired state
- **Inter**action
  - The system can act to affect something outside itself
  - The system receives information and feedback

“Reinforcement learning” can refer to a problem, a solution, or a field of study

- Problem
  - How do we formalize the problem of RL?
  - How do we analyze the bounds of this?
  - How can we map a concrete application to this abstract problem?
- Solutions
  - Once the (abstract or concrete) problem has been defined, how do we solve it with algorithms?
  - How can we approximate the solution?
- Field of study
  - Everything surrounding RL including problem, solutions, preprocessing, etc.

Trial-and-error search and delayed reward are two key aspects of RL

- **Trial-and-error** search (or optimization)
  - Loop of action + feedback and repeat
  - Cannot use gradient descent (at least directly)
  - Cannot enumerate all possible solutions
- **Delayed** reward
  - The ultimate value of an action may not be known **immediately**
  - **Greedy** approaches generally will not work
  - The concept of **planning ahead** can be critical
  - Ultimately, delayed reward is related to the idea of a "goal"

1. Learning to walk
2. Learning optimal oil refinery configuration
3. Learning optimal Netflix images

What are the differences between **supervised learning** and RL?

- Supervised learning
    1. Correct actions (i.e., labels) are given **a priori** for each training sample
        - Fixed set of (hopefully) representative examples
    2. Training and testing/inference phases are usually separate
    3. No real interaction with an environment / passive involvement (e.g., predictions don't directly change the environment)
    4. Focused on prediction / Goal is **generalization**

- Reinforcement learning
    1. Actions (or labels) are evaluated during deployment
        - New examples can be gathered
    2. Training (can) happen at the same time as testing/deployment
    3. Direct interaction with environment
    4. Focused on **causal** action / active involvement
    5. In practice, much more general but also much harder

1. Learning to walk

2. Learning optimal oil refinery configuration

3. Learning optimal Netflix images

What are the differences between **unsupervised learning** and RL?

- Unsupervised learning
  - Aimed at discovering hidden structure
  - Training and inference/sampling phases are usually separate
  - No real interaction with an environment / passive involvement (e.g., predictions don't directly change the environment)

- Reinforcement learning
  - Aimed at maximizing reward
    - Unsupervised learning could be subtask
  - Training (can) happen at the same time as inference/sampling
    - The agent can collect new samples as it learns
  - Direct interaction with environment

Exploitation-exploration tradeoff and specification of goal are also differences

- **Exploitation-exploration** tradeoff
  - Another consequence of the **interaction** part of RL

- RL focuses on the **whole** real-world problem rather than just subproblem
  - Supervised and unsupervised learning can be used as subproblems within RL

1. Learning to walk
2. Learning optimal oil refinery configuration
3. Learning optimal Netflix images

An “agent” and its “environment” are the core components of the RL framework

- The **agent**:
  - **Senses** its environment / Observe the **state** of its environment
    - The sensors could be physical (e.g., pressure sensor) or virtual (e.g., get Twitter trending topics)
  - **Takes actions** (even no action is an “action”)
- The **environment**:
  - **Provides feedback** based on an agent’s actions
  - *Can* change over time
  - *Can* be affected by external events
  - *Can* be affected by agent’s actions
  - *Can* have other agents inside of it

1. Learning to walk
2. Learning optimal oil refinery configuration
3. Learning optimal Netflix images

<u>Policy</u>, **<u>reward</u>**, **<u>value</u>**, and **<u>model</u>** are subelements of the RL framework

1. **<u>Policy</u>**
   - Maps from environment state to action (perhaps stochastically)
2. **<u>Reward</u>**
   - Encodes long-term goal via short-term sensations/rewards
   - Easy to define and observe/estimate
3. **<u>Value</u>**
   - Represents **long-term** value of an environment state or action
   - Hard to define and estimate
4. **<u>Model</u>** of the environment (optional)
   - Enables the agent to hypothesize about future states of the environment (e.g., planning)
   - Could be seen as part of the policy itself (my take)
   - Could be physics simulation environment or an ML prediction model
   - NOTE: Generally, different than what we have called "model", which in ML is usually the function approximator.

1. Learning to walk

2. Learning optimal oil refinery configuration

3. Learning optimal Netflix images

Summary: RL is a more general framework than ML that enables interactive learning towards a desired state

- **Interaction** is the key difference from other learning paradigms because actions can **both**
  - Change the environment (e.g., causal effects)
  - Determine which data is collected during learning
- **Long-term** goals are encoded through **short-term** rewards
  - Analogous to supervised learning
    - Formally defining what is a cat or what is a proper English sentence is **very hard**
    - Giving examples of cat images or proper English sentences is **easy**
  - In RL
    - Defining all the tasks/subtasks needed to achieve a complex goal is **very hard**
    - Defining reward functions is **easy**
- However, RL is much harder than supervised or unsupervised learning
  - If you can solve the problem without RL, you should.

Reference

- Based on the excellent RL book by Sutton and Barto
  - http://incompleteideas.net/book/the-book-2nd.html