GEORGE MASON UNIVERSITY Department of Computer Science

CS/SWE 795 Seminar

Preventing Undesirable Behavior of Intelligent Machines

Authors: Philip Thomas, Bruno Silva, Andrew Barto, Stephen Giguere
Published in 2019

Seyed Mohammadreza Noei
Department of Computer Science
George Mason University

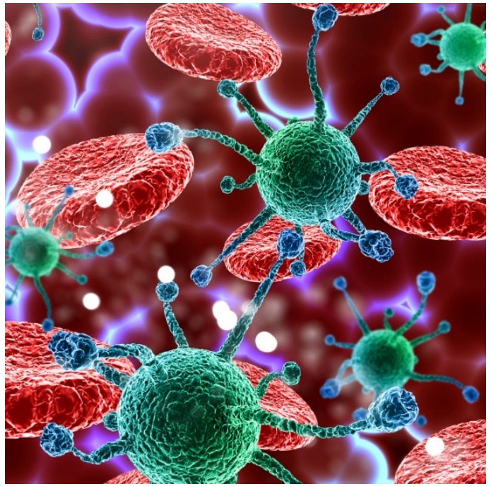
23th Feb 2023

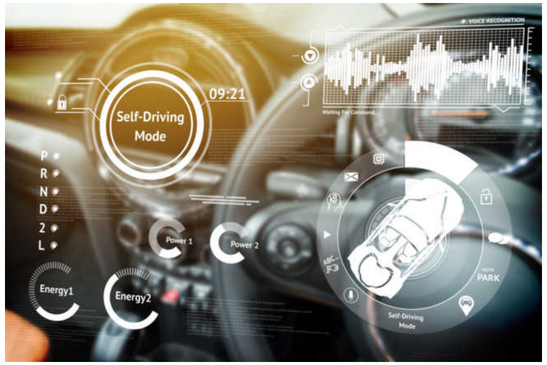
Outline

- Problem Statement
- Methodology
- Results
- Conclusion

Problem Statement

- Biology,
- Geology,
- Autonomous Vehicles,





Problem Statement

□ Standard ML algorithms

- Solution:

  θ ,

- Objective function:

  f: $\Theta \rightarrow \mathbb{R}$,

$$\underset{\theta \in \Theta}{\arg\max} \, f(\theta)$$

□ Proposed Framework

(Seldonian Optimization)

- Objective function:

  f: $A \rightarrow \mathbb{R}$,

$$\underset{a \in \mathcal{A}}{\arg\max} \, f(a)$$

$$\text{s.t. } \forall i \in \{1, \ldots, n\}, \Pr\Big(g_i(a(D)) \leq 0\Big) \geq 1 - \delta_i$$

Proposed Method

3 steps for designing the framework:

1. Define the goal

2. Define the interface

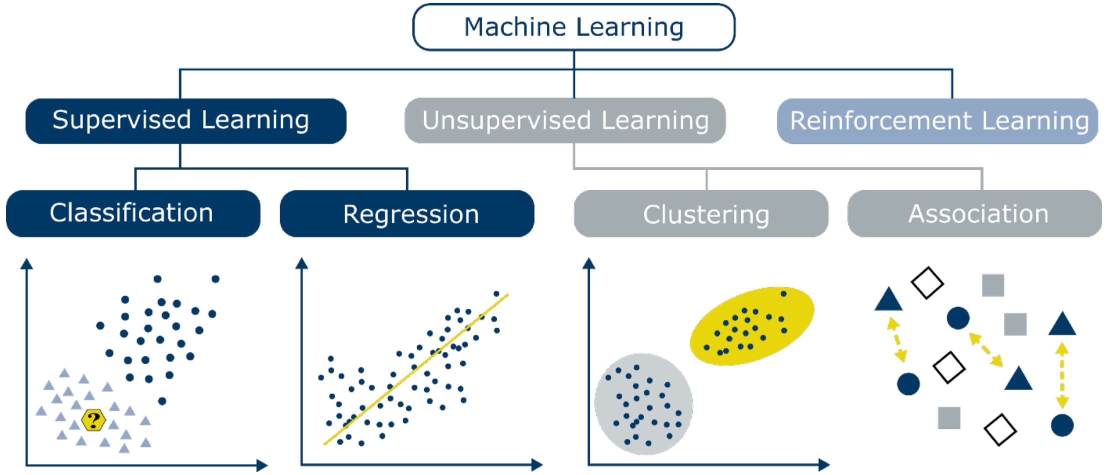
3. Create the algorithm

Statistics Models

1. Regression

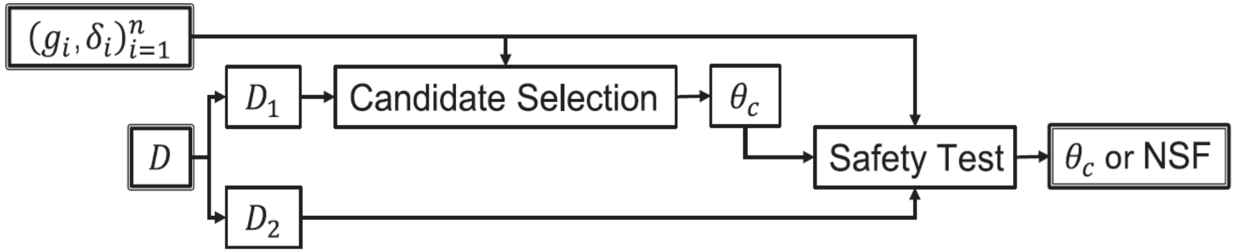
2. Classification

3. Reinforcement Learning



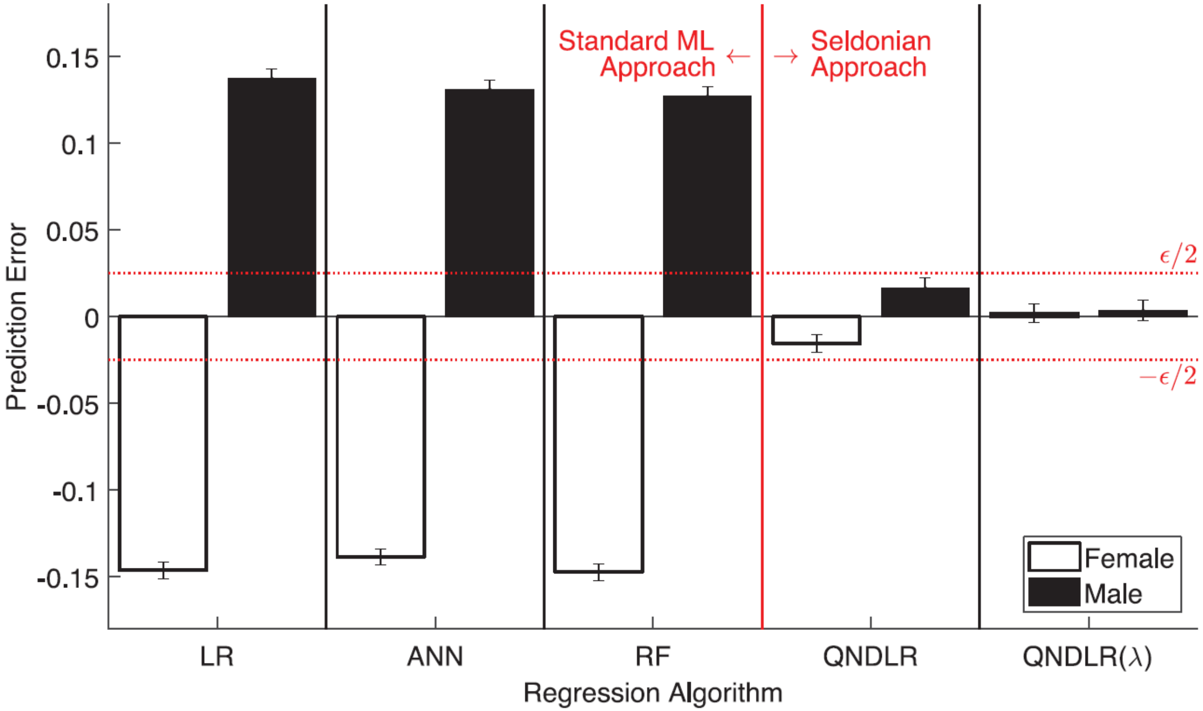
Statistics Models

- Schema of Seldonian regression algorithms



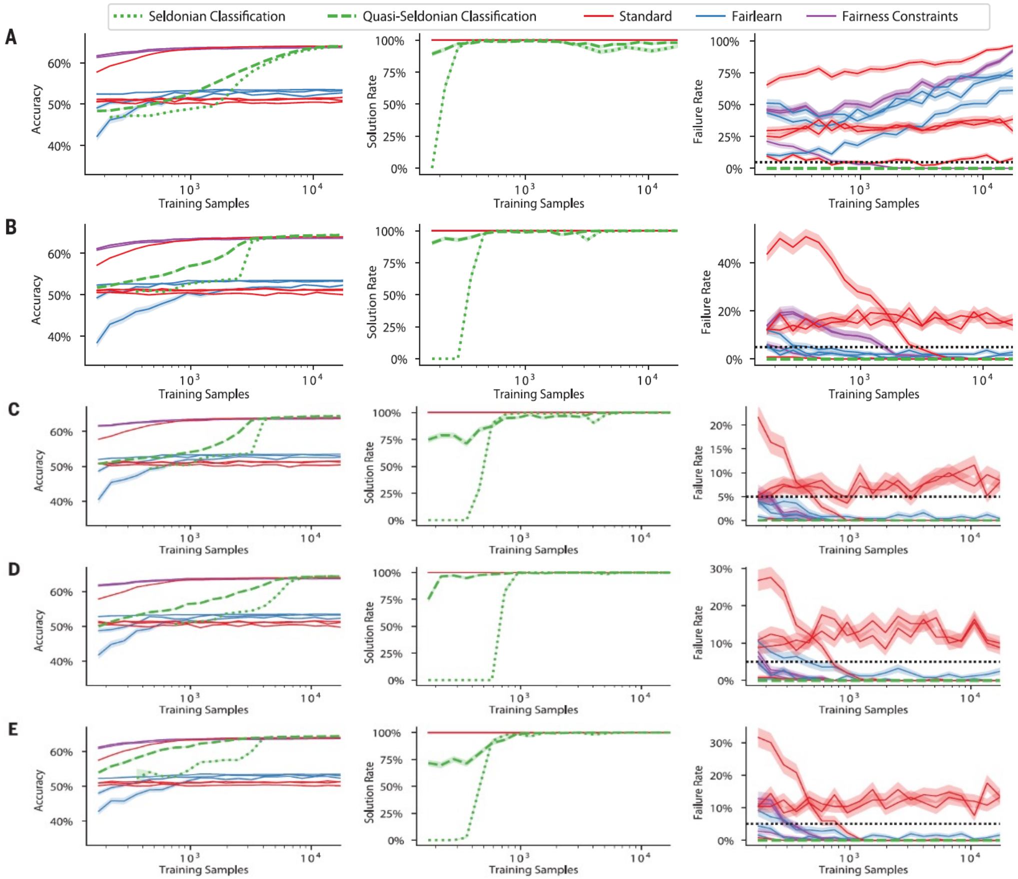
Statistics Models

- Result of Regression



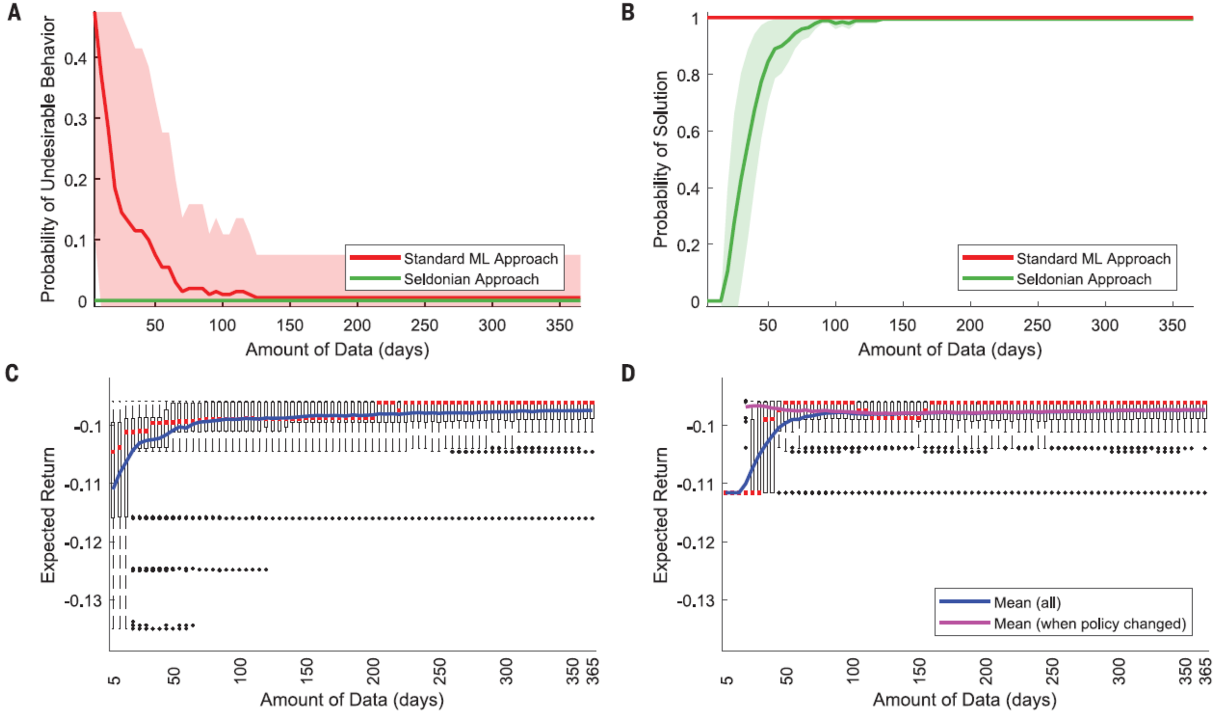
Statistics Models

- Result of Classification



Statistics Models

- Result of Reinforcement Learning



Discussion points

- Solving environmental problems and ethically and socially responsible
- Importance of human control in AI systems
- Need for continued research and development in this area

Q
A