Simulation-based optimisation

Problem: what to do if no prior system or data are available? (esp. new application domains)

Frame: MDP Learning Problem

Method and Results

Method:
1. Build a simulated learning environment from WOZ data
2. Train and test policy by simulated interaction
3. Compare against supervised baseline policy
4. Test policy with real users
5. Post-evaluate: compare results from 3 corpora

Results:
- RL significantly outperforms SL
- 50-times more reward when tested in simulation
- 18-times more reward with real users
- Users rate the RL-based policy 10% higher.

References: