1. The RL problem
2. Formalization
   (a) Discrete, deterministic
   (b) Markov Decision process
   (c) POMDP
3. Solution criteria
   (a) maximum average reward: often mathematically painful
      i. average reward is independent of state.
      ii. average reward is (generall) independent of a one step policy iteration. $\Rightarrow$ measure "transients"
   (b) maximum γ-discounted reward: sometimes hides things.
4. Useful quantities
   (a) State distribution at time $t = d_t(s)$.
   (b) Asymptotic state distribution $d(s)$.
   (c) disounted state distribution $d(s|\gamma, \pi, d_0) = (1-\gamma)\sum_{\gamma=0}^{\infty} \gamma^t d_t(s)$
   (d) Average reward $\eta = E_{a,s\sim\pi(a|s)d(s)} R(s,a)$.
   (e) Discounted reward $\eta_\gamma = E_{a,s\sim\pi(a|s)d(s|\gamma,\pi,d_0)} R(s,a)$
   (f) Expected reward given a state, $J(s) = E_{a,s\sim\pi(a|s)d(s|\gamma,\pi,\delta(s))} R(s,a)$
   (g) Expected reward given a state action, $Q(s,a) = R(s,a) + \gamma E_{a,s\sim\pi(a|s)d(s|\gamma,\pi,\delta(s,a))} R(s,a)$
   (h) Advantage: $A(s,a) = Q(s,a) - J(s)$
5. Deterministic results
   (a) Exploration required
   (b) All states reachable from all others $\Rightarrow$ exploration $= \Theta(VE)$ for exact solution. (http://www.cc.gatech.edu/fac/Sven.Koenig/papers/aaai93.ps.gz)
6. MDP results
   (a) Exploration required
   (b) "mixing time" $= n$ (really state visitation frequency of n) $\Rightarrow \exists c$: $O(n^c)$ algorithm, E^3, for near optimal solution. (http://www.cis.upenn.edu/~mkearns/papers/reinforcement.ps)
      i. Lower bound mixing
      ii. $\frac{1}{1-\gamma} \geq n \Rightarrow$ solutions are similar (http://www.gatsby.ucl.ac.uk/~sham/papers/discounting.ps.gz)
7. POMDP results: strong intractibility
8. Exact RL algorithms

   (a) Dyamic programming
   (b) Q-learning: Learn optimal policy independent of exploration policy

9. Previous Approximate RL algorithms
   (a) Greedy Dynamic Programming: Not guaranteed to improve.
   (b) Policy Gradient: Guaranteed to improve, but very slow.

10. Conservative Policy iteration
   (a) RL reduced to Supervised learning + exploration
   (b) Guaranteed to improve by finite steps $\Rightarrow$ guaranteed to terminate
   (c) "broad" $d_0(s) \Rightarrow$ near optimal