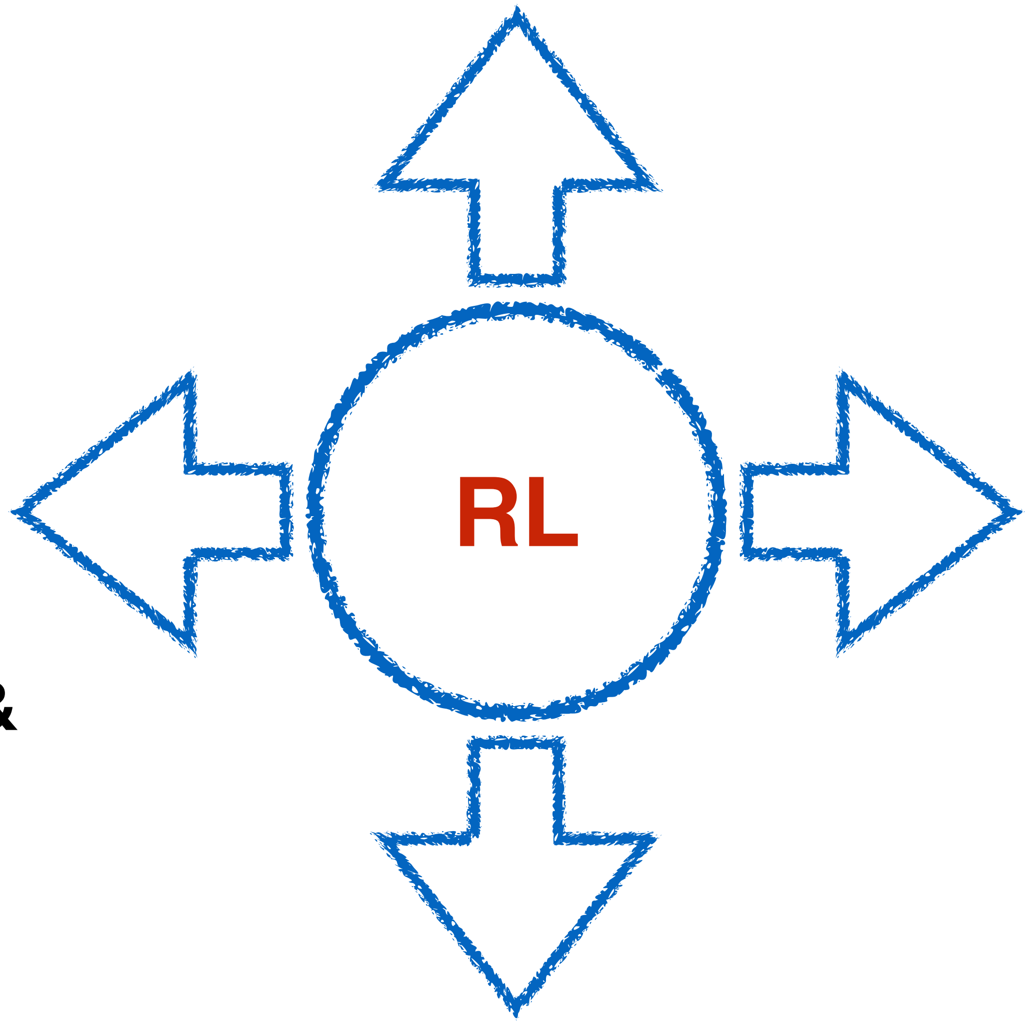
1. Policy Evaluation
(Online Bellman Residual)

[Sun & Bagnell, 15, UAI (Best Student Paper)]

Function Approximation



2. RL via Imitation
(Imitation Learning)

[Sun et.al 17, ICML; 18, ICLR]

Function Approximation &
Imitation

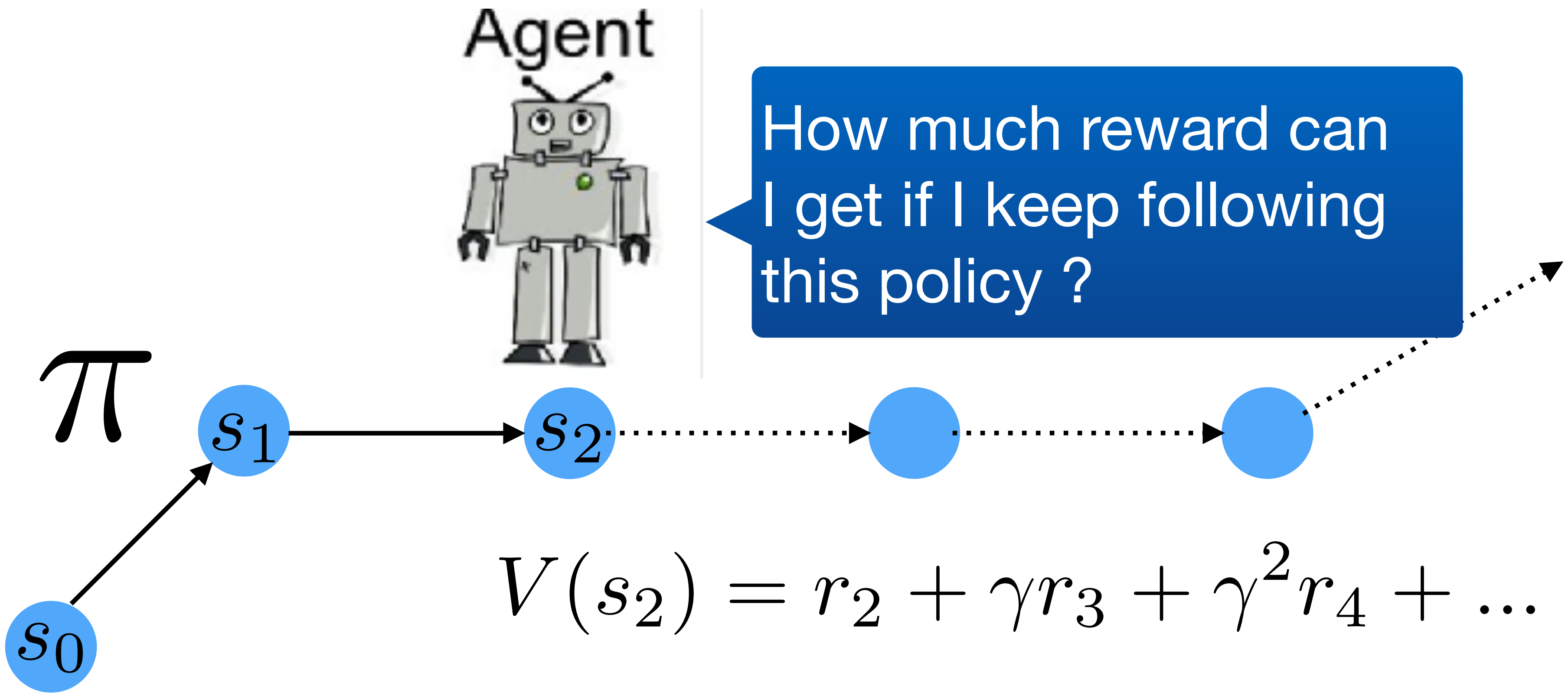
3. RL via Indirect Imitation
(Dual Policy Iteration)

[Sun et.al, 18, submitted to ICML]

Function Approximation
Optimal Control

4. Proposed Work:
Temporal Difference Learning &
Apprenticeship Learning

Policy Evaluation



γ: discount factor

s: state (feature)

r: one-step reward

$$V(s_2) = r_2 + \gamma r_3 + \gamma^2 r_4 + \ldots$$

Predict **Reward-to-go** $\sum_t \gamma^t r_t$

$$f(s) \approx \sum_t \gamma^t r_t$$

Temporal Difference (TD) [Sutton, 1988]
Residual Gradient (RG) [Baird, 1995]

Sequential Online Learning Setting

[Schapire & Warmuth 96, Li 2008]

Environment

$$e_0 = f_0(s_0) - V(s_0)$$

$$V(s_0) = r_0 + \gamma r_1 + \ldots$$

$$e_1 = f_1(s_1) - V(s_1)$$

$$V(s_1) = r_1 + \gamma r_2 + \ldots$$

CHALLENGE #1: Vs are hidden

s_0

$f_0(s_0)$

r_0, s_1

$f_1(s_1)$

r_1, s_2

$f_2(s_2)$

Online Learner

$f_0 \in \mathcal{F}$

Update $f_1 \in \mathcal{F}$
(e.g, TD, RG)

Update $f_2 \in \mathcal{F}$
(e.g, TD, RG)

CHALLENGE #2: No statistical assumption
(e.g., Non-Markovian)

Goal

Goal: minimize the Online **Prediction Error (PE)**:

$$\sum_t e_t^2 = \sum_t (f_t(s_t) - V(s_t))^2$$

Batch **PE:**

$$\sum_t {e_t^*}^2 = \sum_t (f^*(s_t) - V(s_t))^2 \quad f^* \in \mathcal{F}$$

Best Hypothesis in hindsight

Average Online PE

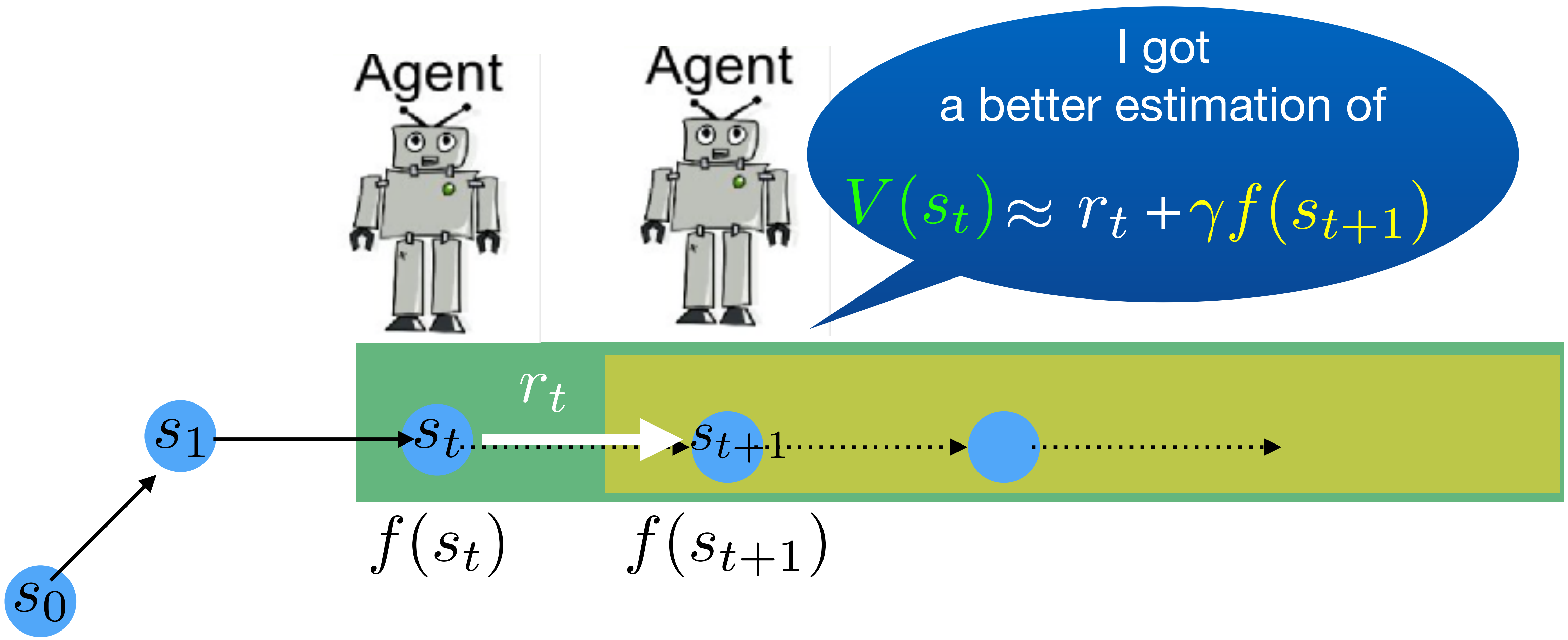
$$\frac{1}{T}\sum_t e_t^2 \le c\frac{1}{T}\sum_t {e_t^*}^2, \quad T \to \infty$$

Smallest possible Batch PE

TD* [Schapire & Warmuth 1996] and RG [Li 2008]

Bellman Loss

Bellman Loss: $l_t(f) = (f(s_t) - r_t - \gamma f(s_{t+1}))^2$



$$f(s_t) - V(s_t) \approx f(s_t) - \boxed{(r_t + \gamma f(s_{t+1}))}$$

Bootstrap

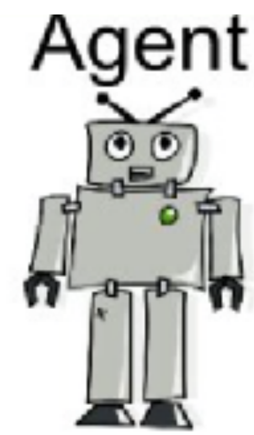
No-regret Online Learning

[Gordon, 99, COLT; Zinkevich, 03, ICML; Shalve-Schwartz, 12]

Hypothesis f_t

Environment
(e.g., Adversary)

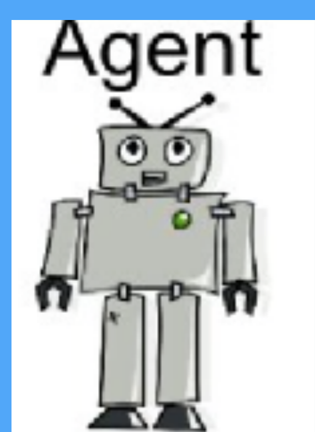
Loss $l_t(f_t)$



$f_1, f_2, ..., f_t$

$l_1, l_2, ..., l_t$

Hypothesis f_{t+1}



$\approx f^*$

Online Stability:

$$\lim_{T\to\infty} \frac{1}{T} \sum_t \|f_{t+1} - f_t\|^2 = 0$$

[Ross & Bagnell 2011, Saha,2012]

Reduction to No-Regret and Stable Online Learning

Recall Bellman Loss at time step t

$$l_t(f) = (f(s_t) - r_t - \gamma f(s_{t+1}))^2$$

No-Regret & **Stable**

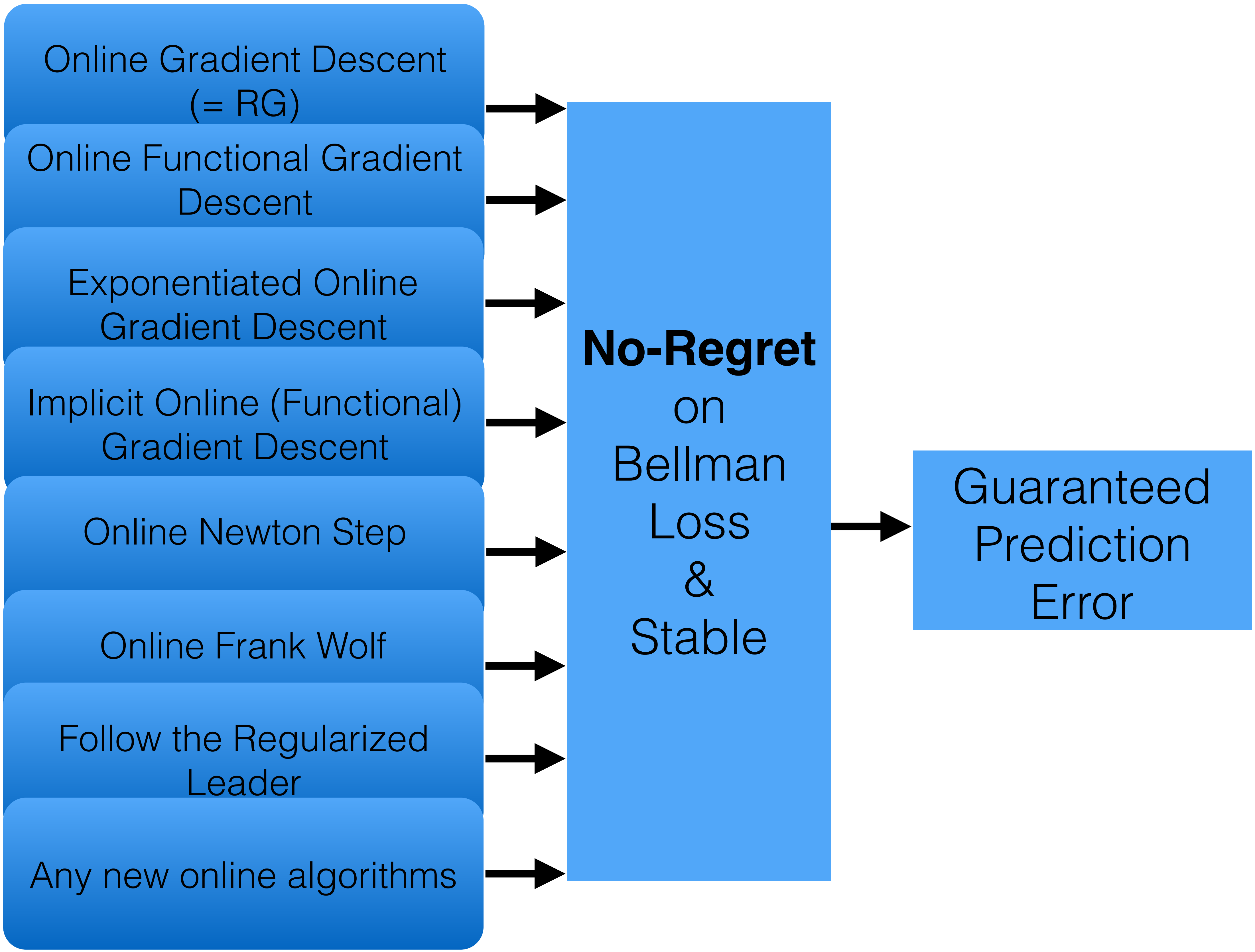
$$l_1(f), l_2(f), ..., l_T(f)$$

Lead to

$$\frac{1}{T}\sum_t e_t^2 \leq \frac{1}{(1-\gamma)^2}\frac{1}{T}\sum_t {e_t^*}^2, \quad T \to \infty$$

[Sun & Bagnell, 15, UAI (Best Student Paper)]

Reduction Leads to a Set of Algorithms



Summary

Message #1:
Agnostic Performance Guarantee with function approximation

Message #2:
Generalization and Efficiency of Policy Evaluation via Reduction to No-Regret Online Learning