# Reinforcement Learning

# und

# Vier Gewinnt

von
Robin Christopher Ladiges

HAW-Hamburg
Projekt Lernende Agenten
16.01.2012

# Gliederung

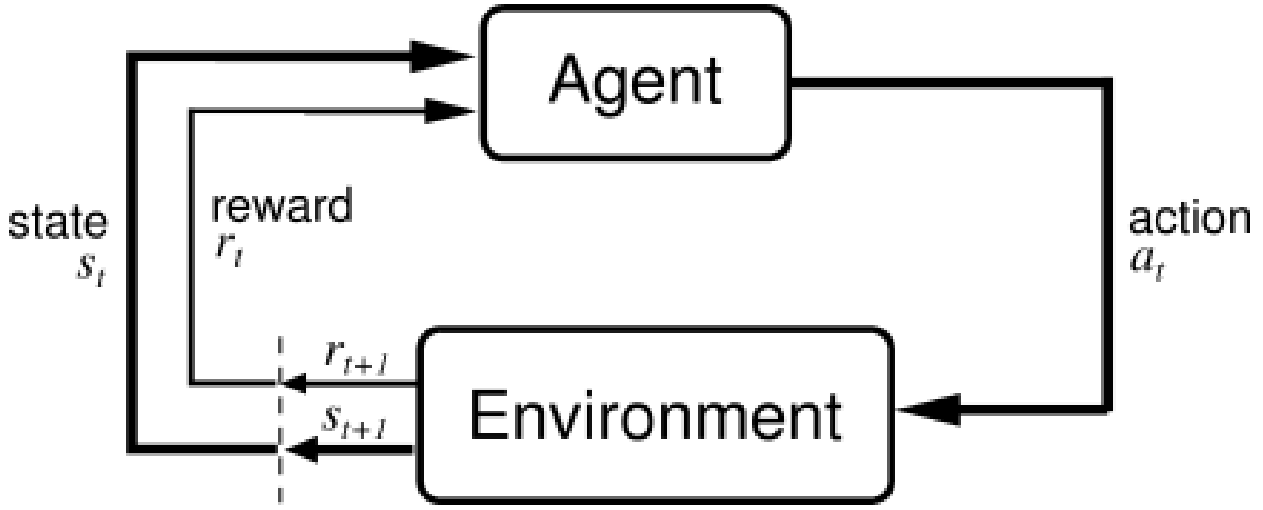
- **Reinforcement Learning**
  - Maschinelles Lernen
  - Reinforcement Learning
  - Bewertungsfunktionen
  - Temporal-difference Learning
- Vier Gewinnt

# Maschinelles Lernen

- Überwachtes Lernen (Supervised Learning)
  - Eingabe- und Zielwerte bekannt
  - Trainingsdaten → Modell
  - Testdaten zur Validierung des Modells
- Unüberwachtes Lernen (Unsupervised Learning)
  - Ähnlichkeit von Eingabedaten → Modell
- Verstärkendes Lernen (Reinforcement Learning)
  - Ausprobieren → eigene Erfahrungen sammeln

# Reinforcement Learning

- Agent lernt selbstständig durch ausprobieren
- Agent nimmt Umwelt mittels Sensoren wahr
- Probiert Aktionen aus
- Erhält Belohnung oder Bestrafung



# Bewertungsfunktionen

- Zustand-Wert-Funktion (V-Werte)

  V(s) : Situation → Bewertung

  Wie gut ist diese Situation?

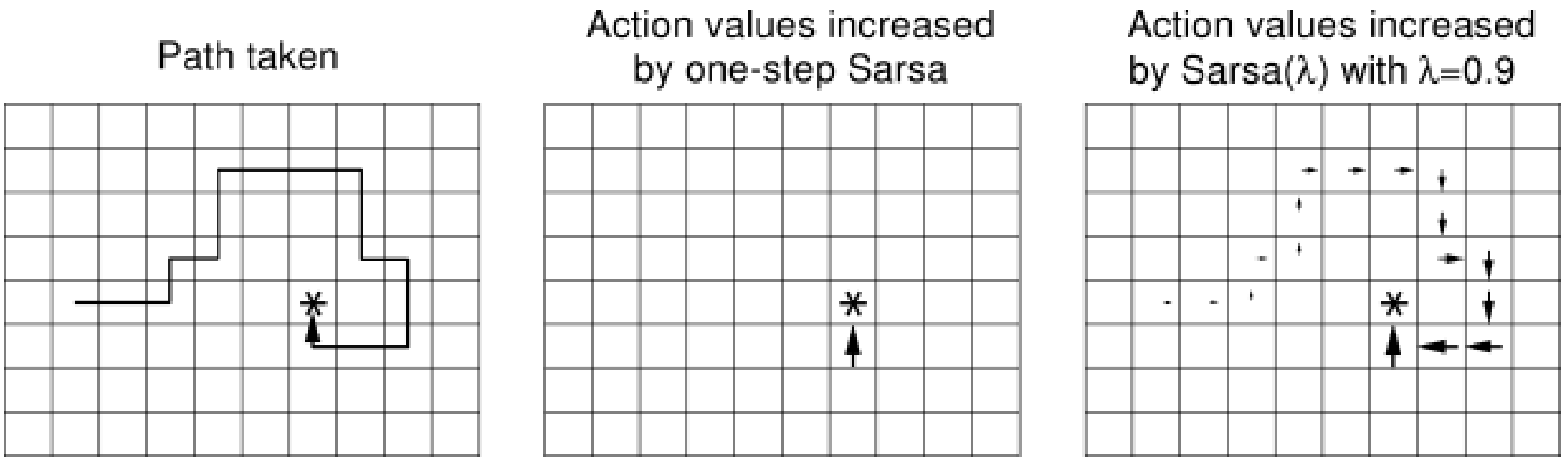
- Aktion-Wert-Funktion (Q-Werte)

  Q(s,a) : Situation x Aktion → Bewertung

  Wie gut ist diese Aktion in dieser Situation?

# Temporal-difference Learning

- Nicht nur die letzte Aktion beeinflusst Ergebnis
- Umso länger die Aktion zurückliegt, desto weniger Einfluss hatte sie auf das Ergebnis



# Gliederung

- Reinforcement Learning
- **Vier Gewinnt**
  - Kommunikation Umwelt & Agent
  - Situations-ID
  - Einsparungen für Wertetabelle
  - Größe der Wertetabelle
  - Live-Demonstration

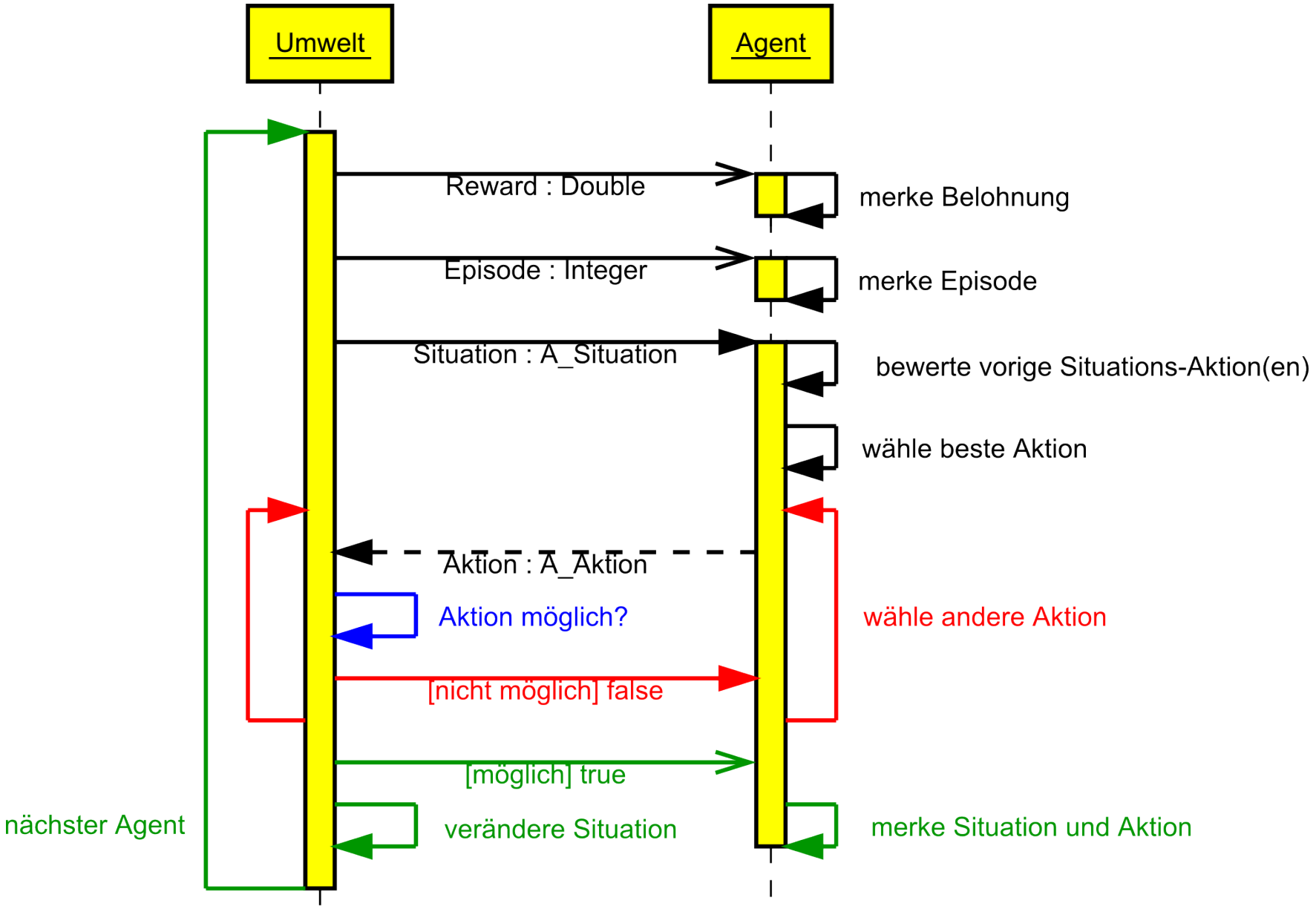
# Vier Gewinnt

RL-Framework von Patrick Boekhoven [1]

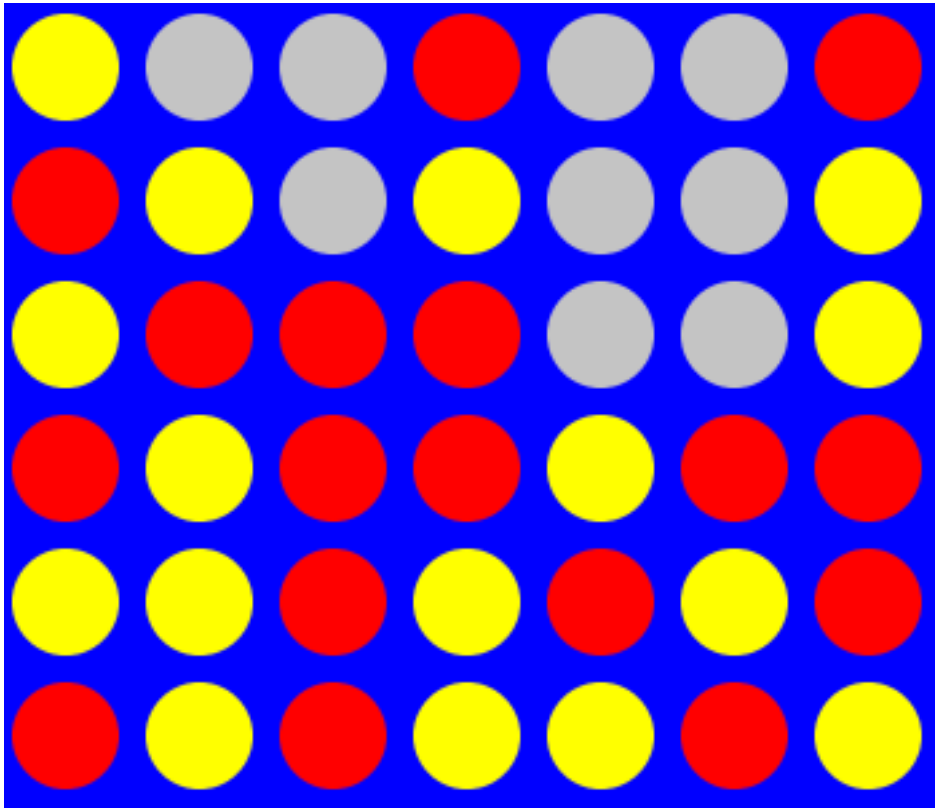
Projektübernahme SS2011 von Daniel Wehring

[1] http://opus.haw-hamburg.de/volltexte/2011/1269/

# Kommunikation Umwelt & Agenten



# Situations-ID



Pro Feld 3 Zustände

7 Reihen, 6 Zeilen

naiv

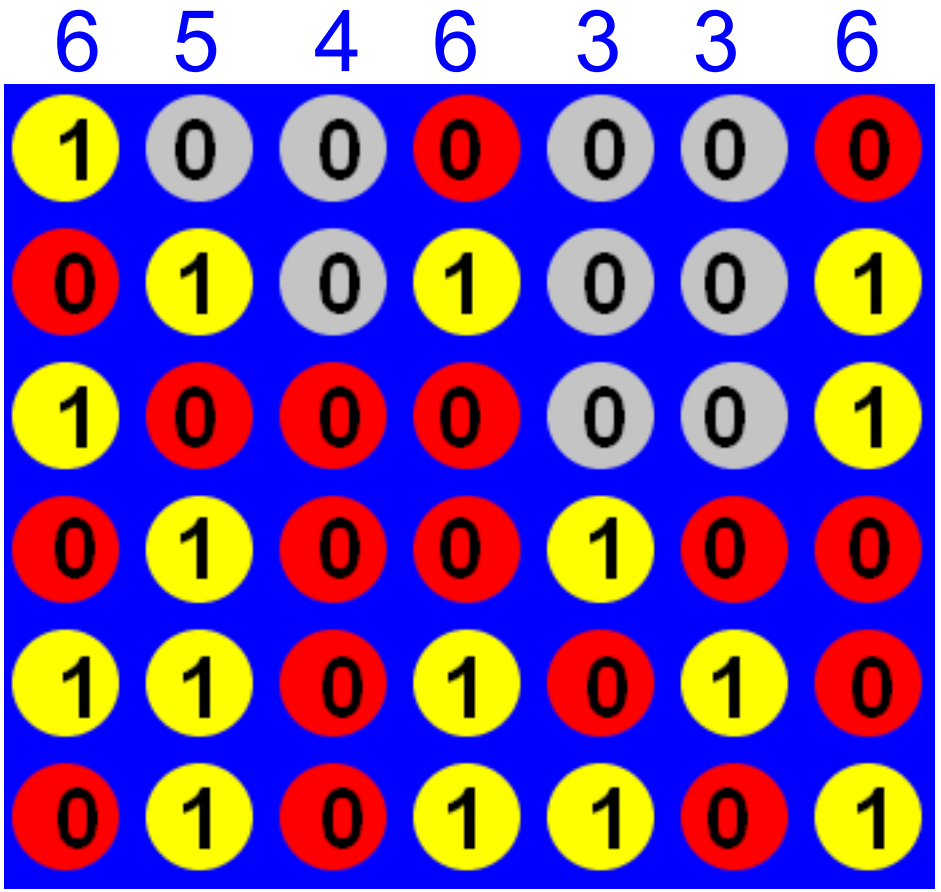
$$3^{6\cdot 7} = 1{,}094 \cdot 10^{20} \rightarrow 67\ Bit$$

optimal [2]

$$4{,}531 \cdot 10^{12} \rightarrow 43\ Bit$$

[2] http://homepages.cwi.nl/~tromp/c4/c4.html

# Situations-ID




Problem:

Long nur 64 Bit

Farbe 1 Bit / Feld

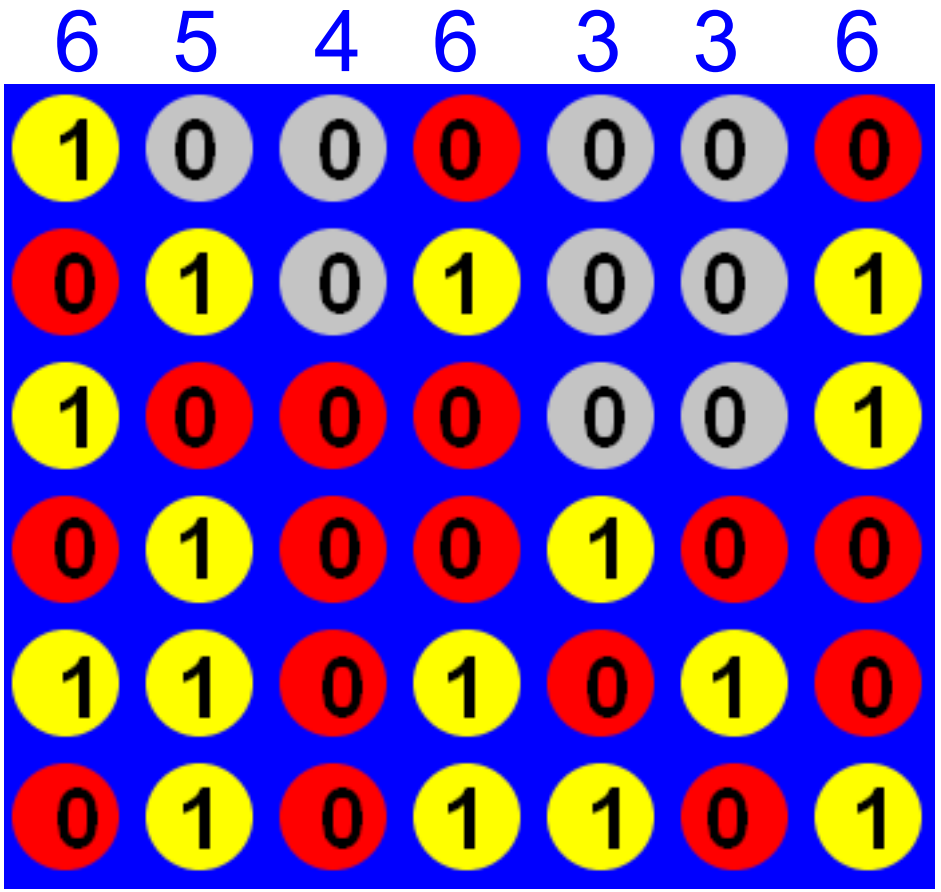
→ 42 Bit

Anzahl Steine / Reihe

→ 7 x 3 Bit = 21 Bit

42 + 21 = 63 Bit

# Situations-ID




**0** (Vorzeichen)

| | | |
|---|---|---|
| **110** (6) | 101010 | (Reihe 0) |
| **101** (5) | 010111 | (Reihe 1) |
| **100** (4) | 000000 | (Reihe 2) |
| **110** (6) | 010011 | (Reihe 3) |
| **011** (3) | 000101 | (Reihe 4) |
| **011** (3) | 000010 | (Reihe 5) |
| **110** (6) | 011001 | (Reihe 6) |

= 7.686.219.650.993.325.465

# Einsparung für Wertetabelle

V-Werte (Situation → Bewertung)

Q-Werte (Situation x Aktion → Bewertung)

Situation x Aktion = Folgesituation

Folgesituation → Bewertung

Einsparung in Wertetabelle von Faktor 7

- RL-Framework arbeitet mit Q-Werten

  Möglichkeit equals & hashCode von Situations-Aktionen zu überschreiben

# Größe der Wertetabelle

Datenbank

- 64 Bit für Situation (Long)
- 32 Bit für Bewertung (Float)
- 4.531.985.219.092 legale Spiel-Situationen [2]

→ 435.070.581.032.832 Bit = ~54 TB

Näherungsverfahren nötig (Neuronales Netz)

[2] http://homepages.cwi.nl/~tromp/c4/c4.html

# Live-Demonstration