RUB

INSTITUT
FÜR
NEUROINFORMATIK

Reinforcement Learning

Abdullah Sahin

27.11.2018

„Maschinelles Lernen“ bis jetzt

INSTITUT FÜR NEUROINFORMATIK

- **Supervised Learning**
  - erfordert gelabelte Trainingsdaten
- *Unsupervised Learning*
  - erfordert Interpretation der Ausgabe

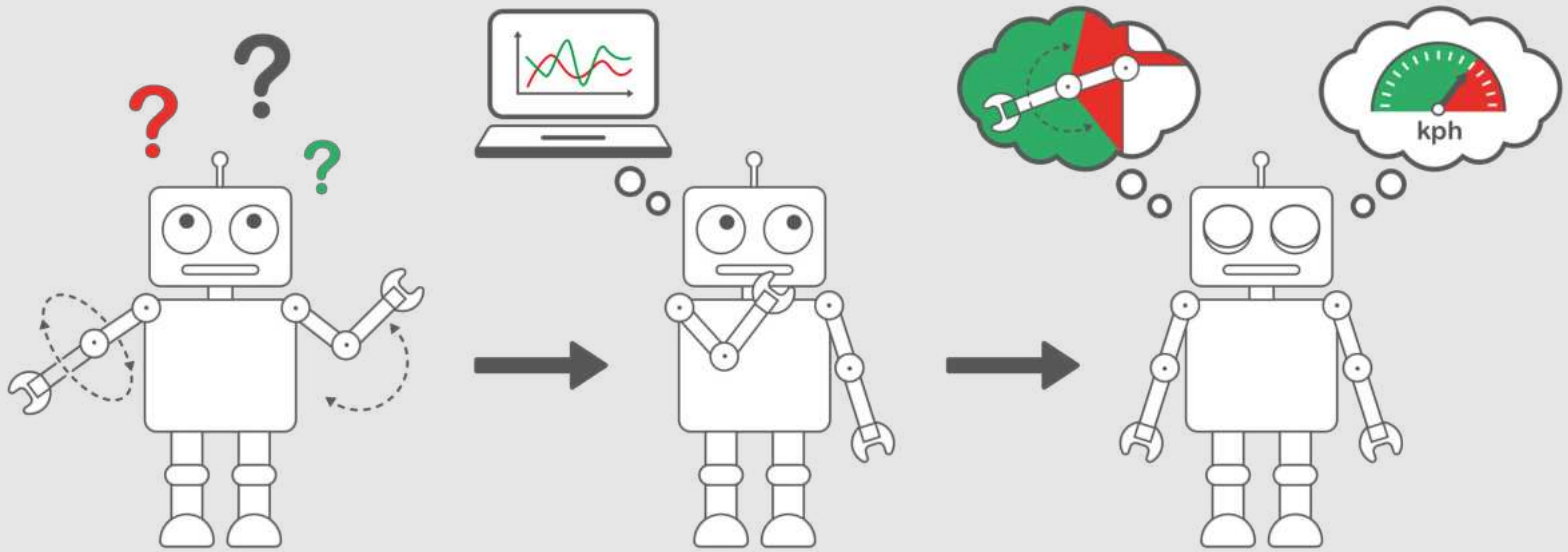
„Maschinelles Lernen“ bis jetzt

INSTITUT FÜR NEUROINFORMATIK

- **Supervised Learning**
  - erfordert gelabelte Trainingsdaten

beide Verfahren erfordern manuelle Arbeit

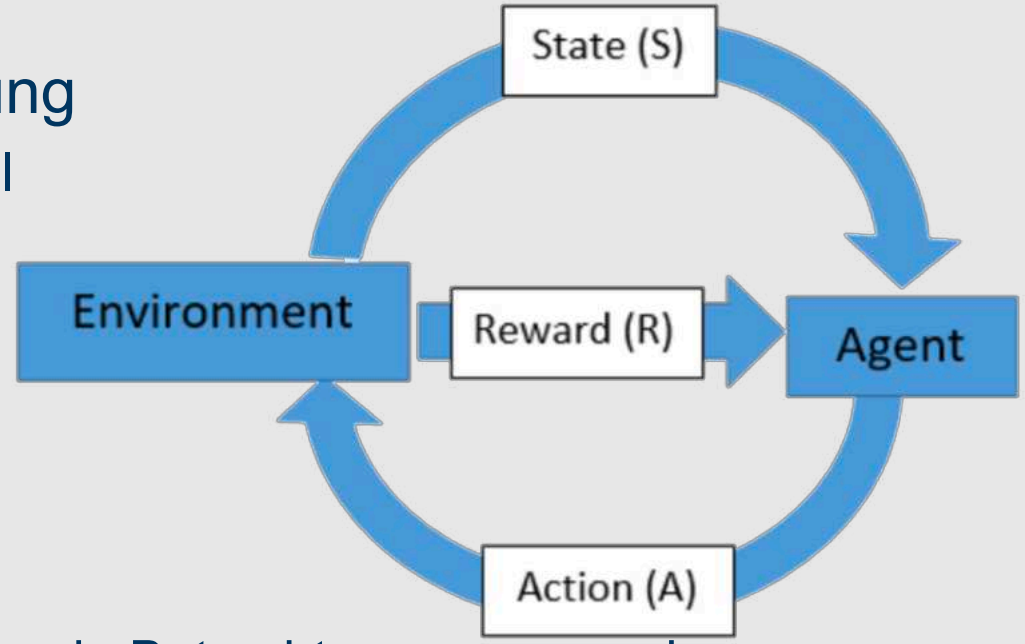
- *Unsupervised Learning*
  - erfordert Interpretation der Ausgabe



Reinforcement Learning

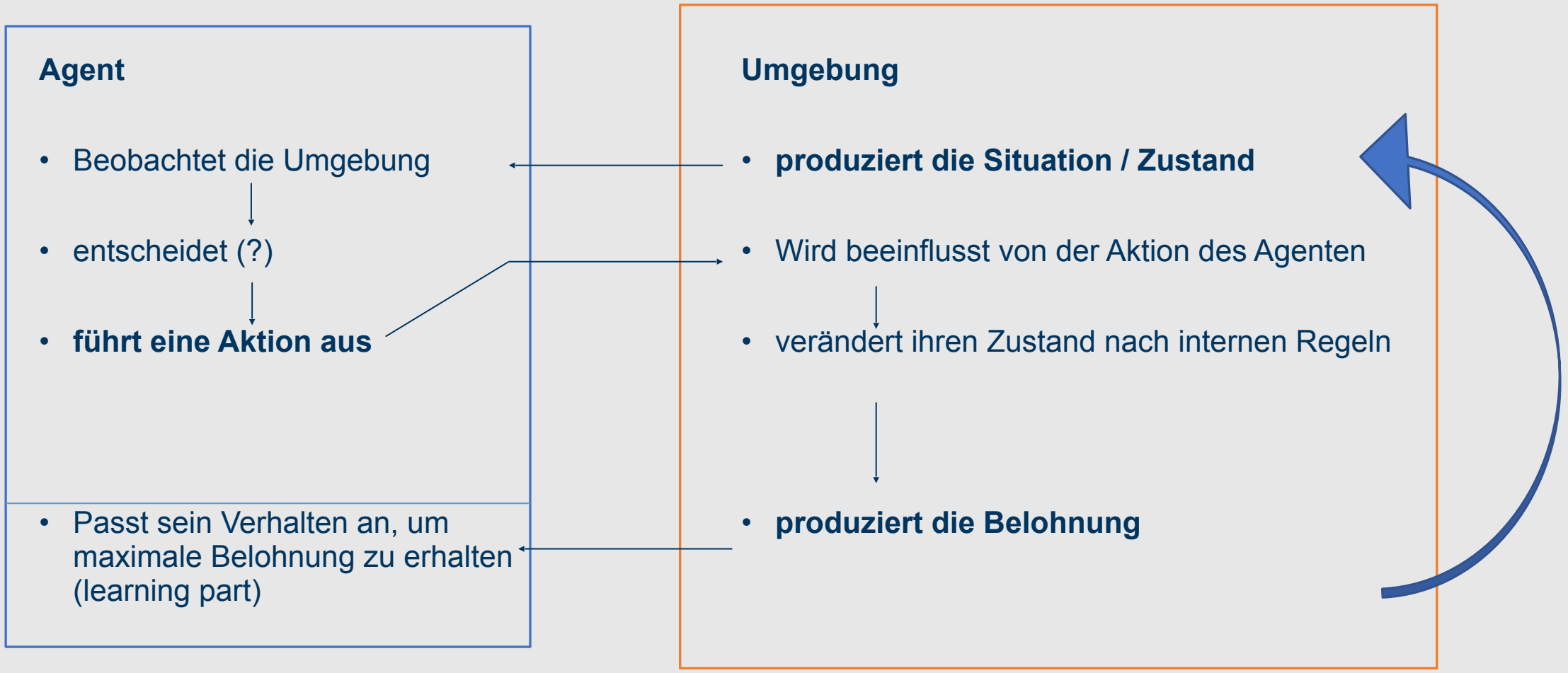
INSTITUT FÜR NEUROINFORMATIK

- am Lernverhalten des Menschen orientiert
- Lernen erfolgt durch Interaktion in einer Umgebung
  - dem Agent ist unbekannt welche Aktion er ausführen soll
  - Lernen durch „trial and error“
  - Neues ausprobieren und Bekanntes verbessern



- Ziel ist es die Gesamtbelohnung zu maximieren
  - Bei Auswahl von Aktion müssen zukünftige Konsequenzen in Betracht gezogen werden

Konzept


Agent
• Beobachtet die Umgebung
• entscheidet (?)
• führt eine Aktion aus
• Passt sein Verhalten an, um maximale Belohnung zu erhalten (learning part)
Umgebung
• produziert die Situation / Zustand
• Wird beeinflusst von der Aktion des Agenten
• verändert ihren Zustand nach internen Regeln
• produziert die Belohnung

Unterschiede

INSTITUT FÜR NEUROINFORMATIK

- **Supervised Learning**
  - lernt von gelabelten Trainingsdaten
  - kann maximal so gut sein wie die Trainingsdaten
  - Wissen ist statisch
  - bildet unbekanntes Input in **bekannte** Outputs ab

- **Reinforcement Learning**
  - lernt eigenständig durch Entscheidungen
  - kann übermenschliche Leistungen erbringen
  - Wissen ist dynamisch
  - bildet Zustände auf Aktionen ab (Strategie)

Unterschiede

INSTITUT FÜR NEUROINFORMATIK

- **Supervised Learning**
  - lernt von gelabelten Trainingsdaten
  - kann maximal so gut sein wie die Trainingsdaten
  - Wissen ist statisch
  - bildet unbekanntes Input in **bekannte** Outputs ab
- **Unsupervised Learning**
  - versucht in einer Datenmenge unbekannte Strukturen zu finden
  - Ausgabe muss interpretiert werden

- **Reinforcement Learning**
  - lernt eigenständig durch Entscheidungen
  - kann übermenschliche Leistungen erbringen
  - Wissen ist dynamisch
  - bildet Zustände auf Aktionen ab (Strategie)
  - versucht die Belohnung zu maximieren
  - Ausgabe wird von der Umgebung konsumiert

Unterschiede

INSTITUT FÜR NEUROINFORMATIK

- **Supervised Learning**
  - lernt von gelabelten Trainingsdaten
  - kann maximal so gut sein wie die Trainingsdaten
  - Wissen ist statisch
  - bildet unbekanntes Input in **bekannte** Outputs ab
- **Unsupervised Learning**
  - versucht in einer Datenmenge unbekannte Strukturen zu finden
  - Ausgabe muss interpretiert werden

- **Reinforcement Learning**
  - lernt eigenständig durch Entscheidungen
  - kann **übermenschliche Leistungen** erbringen
  - Wissen ist dynamisch
  - bildet Zustände auf Aktionen ab (Strategie)
  - versucht die Belohnung zu maximieren
  - Ausgabe wird von der Umgebung konsumiert

Ausgezeichnete Leistungen

IBM‘s Watson

INSTITUT FÜR NEUROINFORMATIK

- **Watson**
  - Supercomputer kombiniert mit künstlicher Intelligenz
  - Entwickelt um Antworten auf Fragen (Suchmaschine)
  - Kommuniziert in natürlicher Sprache

IBM‘s Watson

INSTITUT FÜR NEUROINFORMATIK

- **Watson**
  - Supercomputer kombiniert mit künstlicher Intelligenz
  - Entwickelt um Antworten auf Fragen (Suchmaschine)
  - Kommuniziert in natürlicher Sprache

- Februar 2011: **gewann** gegen zwei Champions

IBM‘s Watson

INSTITUT FÜR NEUROINFORMATIK

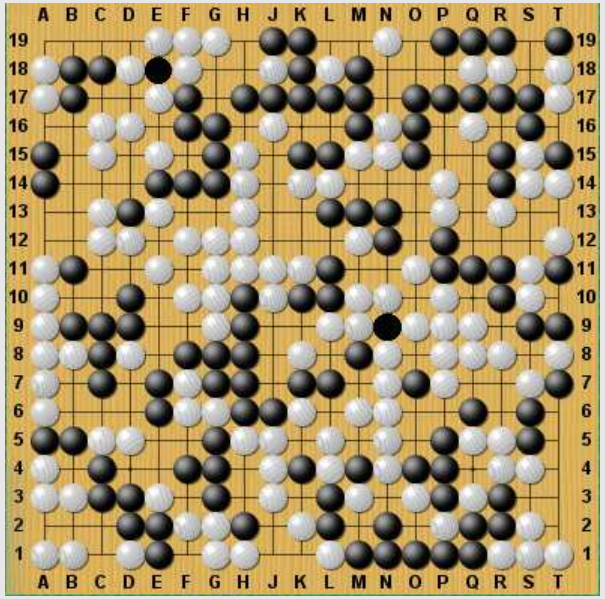
- **Watson**
  - Supercomputer kombiniert mit künstlicher Intelligenz
  - Entwickelt um Antworten auf Fragen (Suchmaschine)
  - Kommuniziert in natürlicher Sprache

- **Jeopardy!**
  - Ein Fernseh-Quizshow aus den USA
  - erfordert Schnelligkeit und Fachwissen

- Februar 2011: **gewann** gegen zwei Champions

AlphaGo

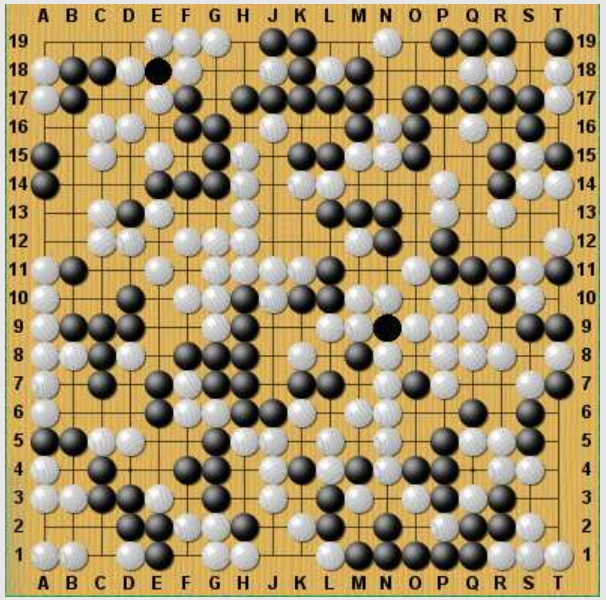
- **AlphaGo**
  - entwickelt von Google‘s Tochterunternehmen **DeepMind**
  - soll ausschließlich **Go** spielen



AlphaGo

- **AlphaGo**
  - entwickelt von Google‘s Tochterunternehmen **DeepMind**
  - soll ausschließlich **Go** spielen

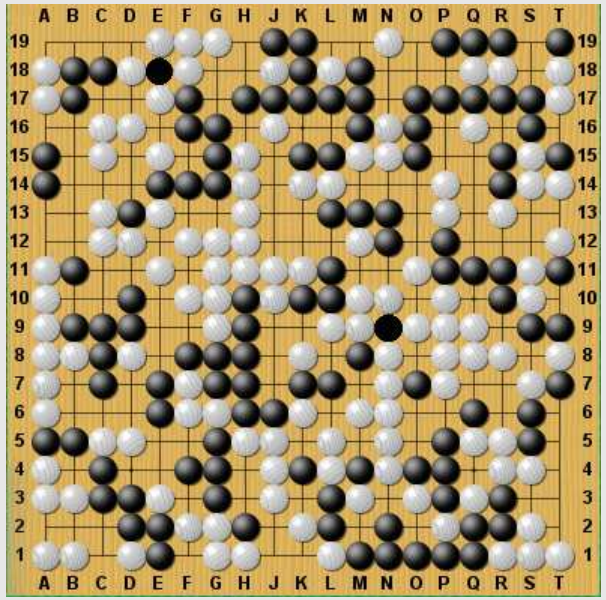
- **Go**
  - Ein strategisches Brettspiel
  - Bei 19x19 Brettgröße ca. 10^{171} gültige Spielpositionen



AlphaGo

- **AlphaGo**
  - entwickelt von Google‘s Tochterunternehmen **DeepMind**
  - soll ausschließlich **Go** spielen

- **Go**
  - Ein strategisches Brettspiel
  - Bei 19x19 Brettgröße ca. 10^{171} gültige Spielpositionen



- Mai 2016: **4:1** gegen einer der besten Spieler (Lee Sedol)
- Mai 2017: **3:0** gegen Weltranglistenerste (Ke Jie)

AlphaZero

INSTITUT FÜR NEUROINFORMATIK

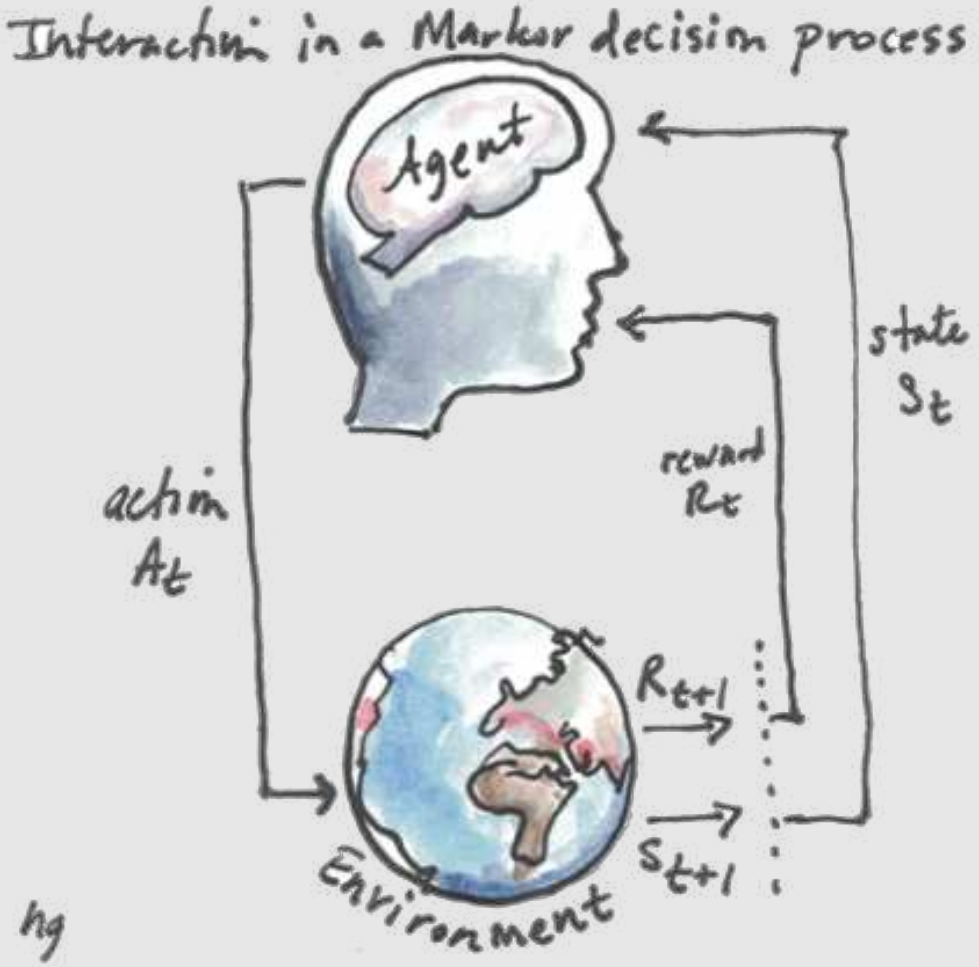
- **AlphaZero** kann **Schach**, **Shogi** und **Go** auf übermenschlichem Niveau spielen
- nach **nur 24 Stunden** trainieren / lernen mit sich selbst
- verwendet die gleichen Hyperparameter für alle Spiele

Spiel	Brettgröße Felderanzahl	Zustandsraum-Komplexität (als dekadischer Logarithmus $\log_{10}$)	Spielbaum-Komplexität ($\log_{10}$)
Schach	8×8=64	50[6]	123[6]
Shōgi	9×9=81	71[8]	226[8]
Go	19×19=361	171[10]	360[11]

Markov Decision Process

Basiskomponente

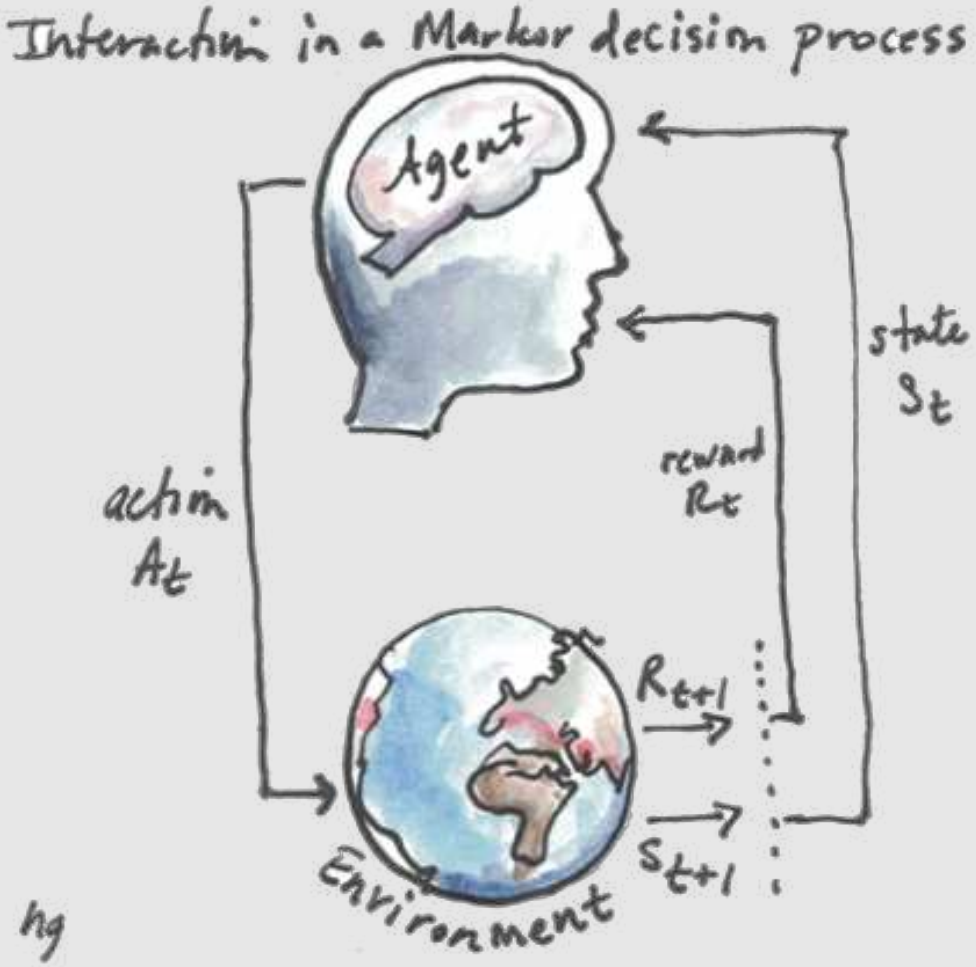
- **Strategie π**
  - Verhalten des Agenten
- **Belohnungssignal**
  - Wie gut hat er sich verhalten



Basiskomponente

INSTITUT FÜR NEUROINFORMATIK

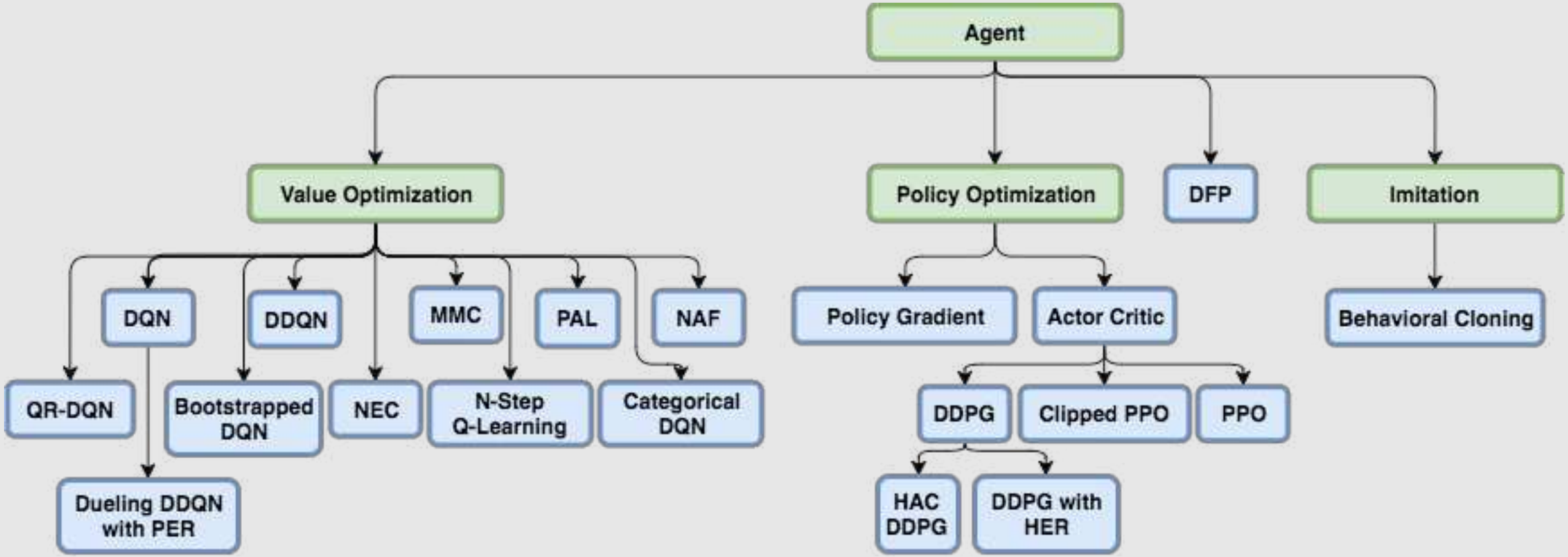
- **Strategie π**
  - Verhalten des Agenten
- **Belohnungssignal**
  - Wie gut hat er sich verhalten
- **Nutzenfunktion**
  - **V(s)** : erwartete Belohnung bei Startzustand **s**
  - **q(s,a)** : erwartete Belohnung bei Aktion **a** in **s**
- **Return G_t**
  - G_t : Gesamtbelohnung zur Zeit **t**
- **Modell** *(optional)*
  - $\mathscr{P}$**: Transaktionsmatrix**
  - $\mathscr{R}$**: Belohnungsfunktion**



Lernmethoden

RL Algorithmen *unvollständig*

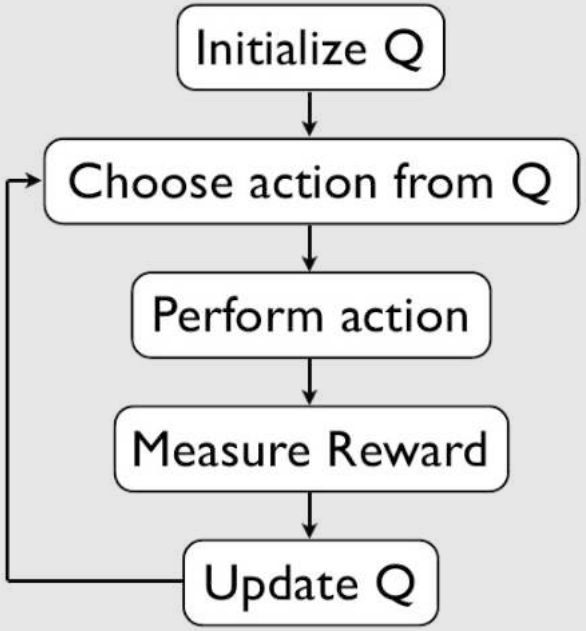
INSTITUT FÜR NEUROINFORMATIK



Q-Learning

INSTITUT FÜR NEUROINFORMATIK

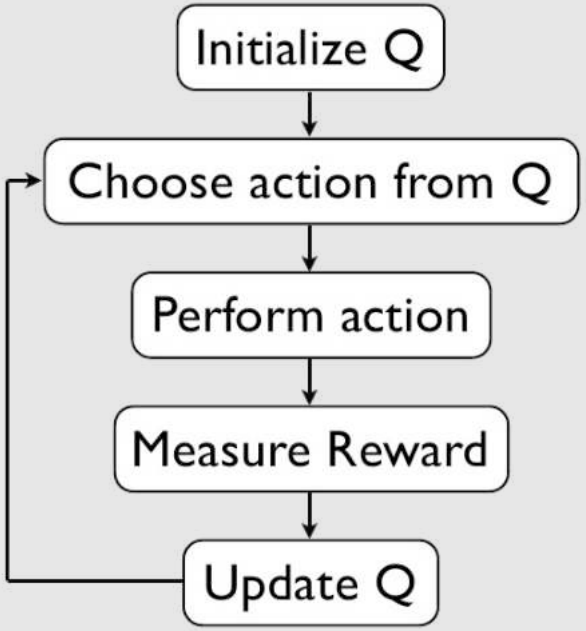
- **Einfaches Lernalgorithmus**
  - lernt langfristig ein optimales Verhalten
  - Ungeeignet für große Zustand-Aktion-Räume
- **Modell frei**
  - erfordert keine Kenntnisse über die Dynamik der Umgebung



Q-Learning

INSTITUT FÜR NEUROINFORMATIK

- **Einfaches Lernalgorithmus**
  - lernt langfristig ein optimales Verhalten
  - Ungeeignet für große Zustand-Aktion-Räume
- **Modell frei**
  - erfordert keine Kenntnisse über die Dynamik der Umgebung

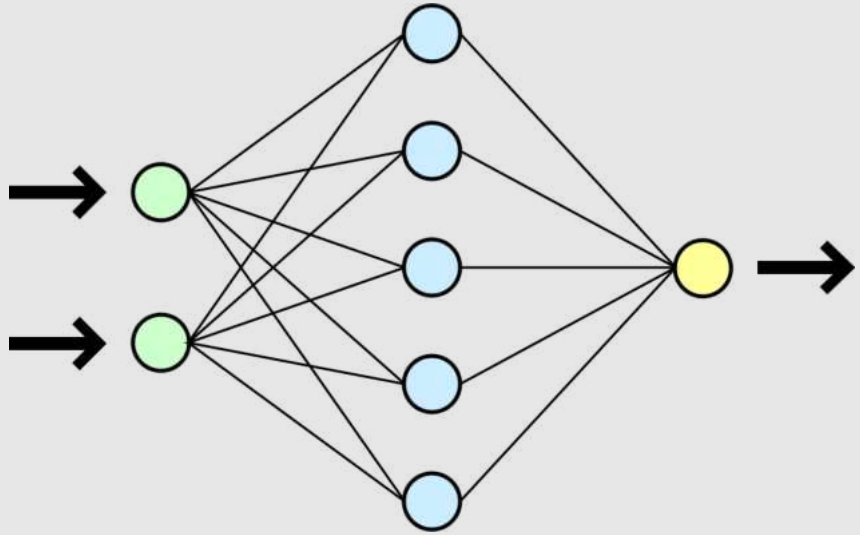


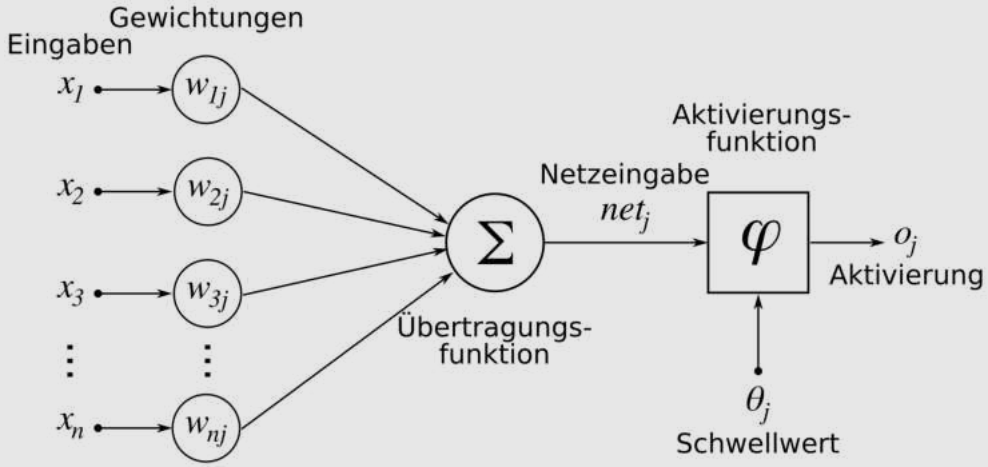
$$Q^{new}(s_t, a_t) \leftarrow (1-\alpha) \cdot \underbrace{Q(s_t, a_t)}_{\text{old value}} + \underbrace{\alpha}_{\text{learning rate}} \cdot \overbrace{\left(\underbrace{r_t}_{\text{reward}} + \underbrace{\gamma}_{\text{discount factor}} \cdot \underbrace{\max_a Q(s_{t+1}, a)}_{\text{estimate of optimal future value}} \right)}^{\text{learned value}}$$

Anwendungen

Hyperparameterwahl bei KNN

- Anzahl der Hidden-Layer
- Initialwerte der Gewichte
- Aktivierungsfunktionen
- Lernrate
- Anzahl der Epochen



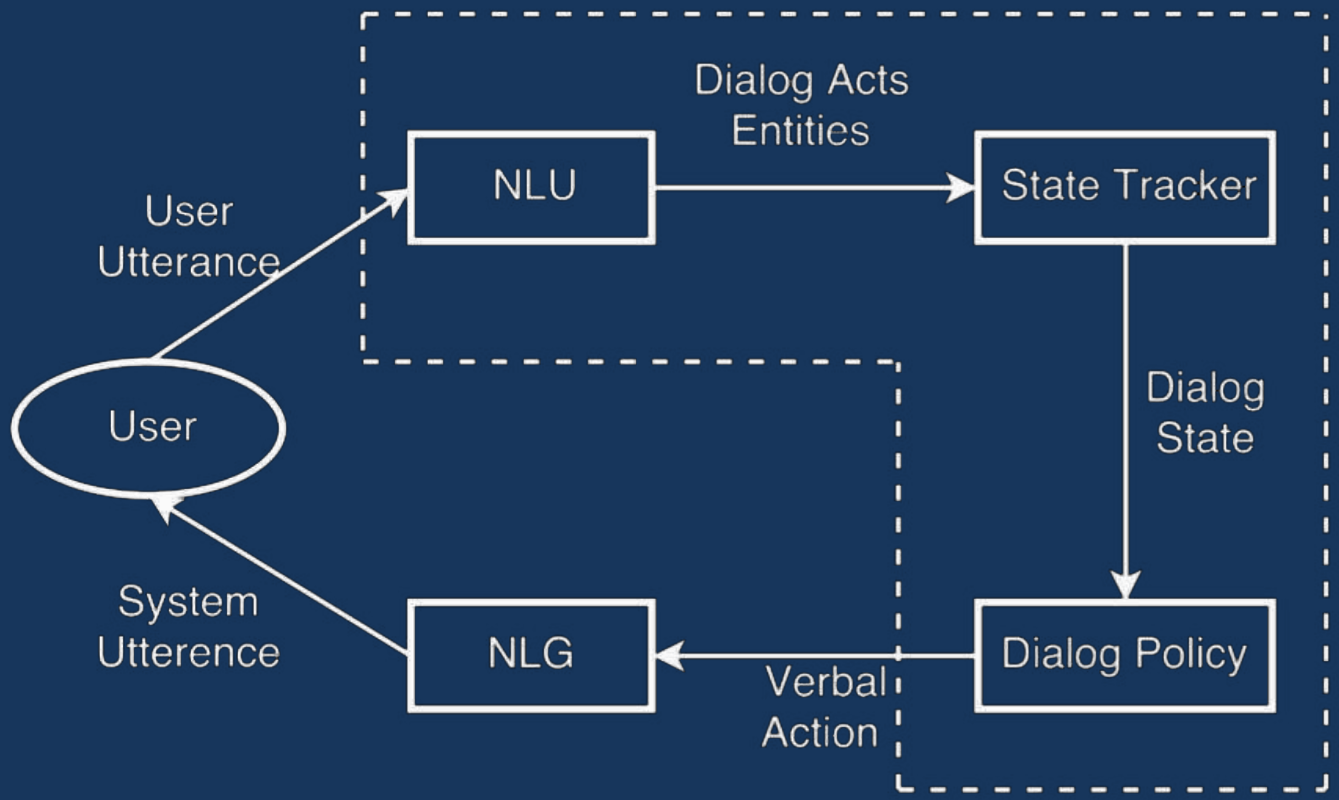


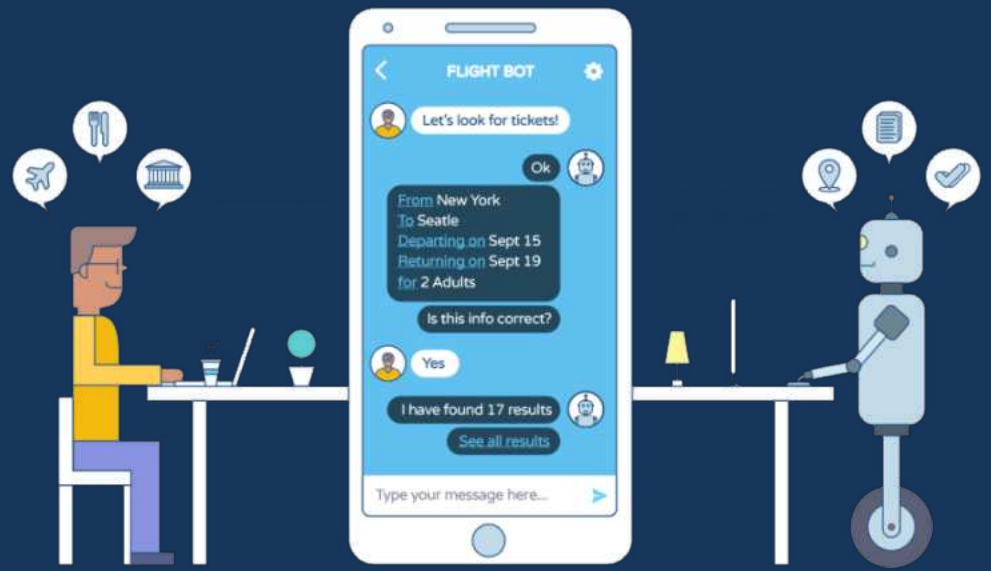
Intelligent Transportation Systems

E-Commerce


Page Request
Ranking Action
Search Engine
Page 1
click
back
Page Request
Ranking Action
Search Engine
Page 2
click
back
Page Request
Ranking Action
Search Engine
Page n
buy
leave
End

Dialog Systems, chatbots

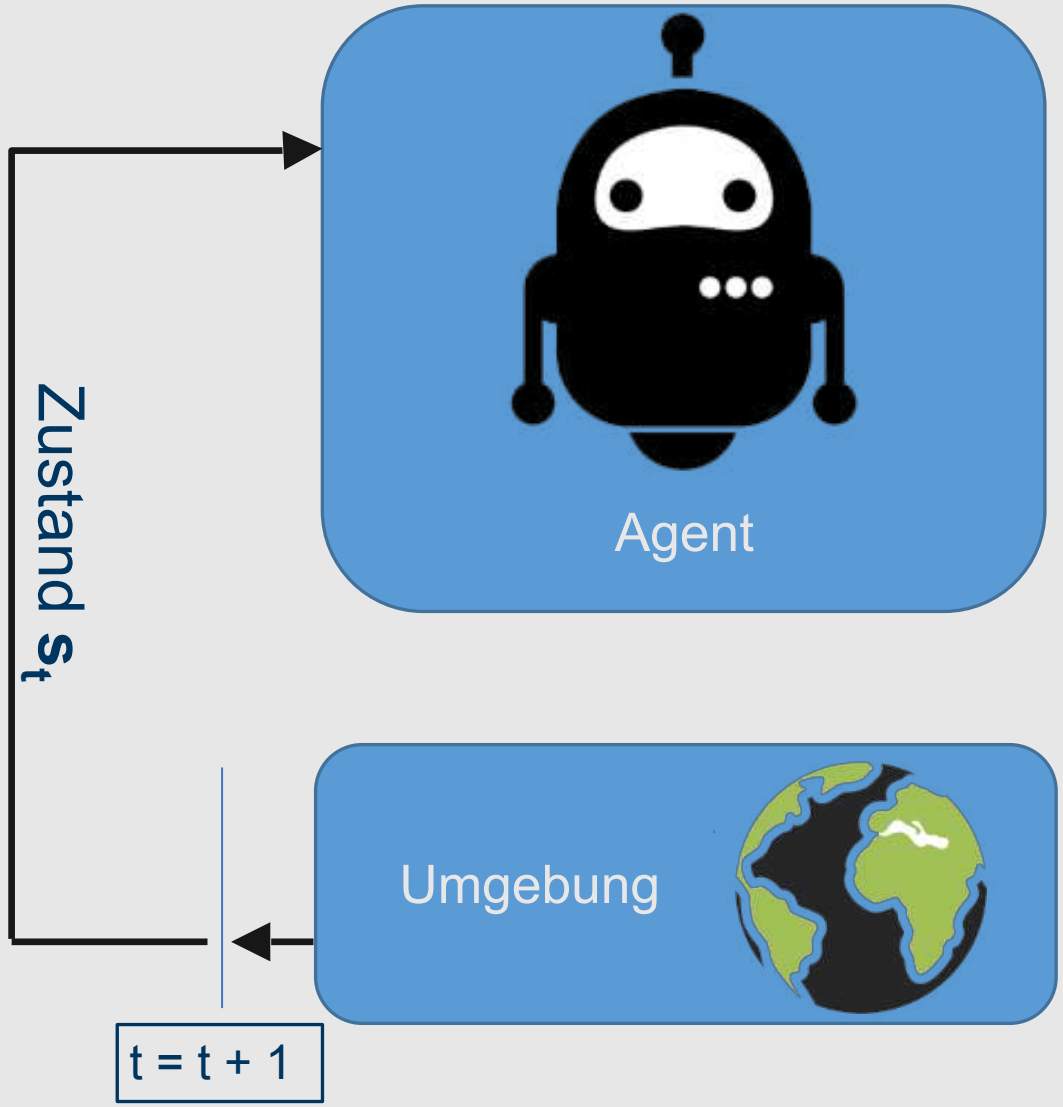




NLU: natural language understanding
NLG: natural language generation

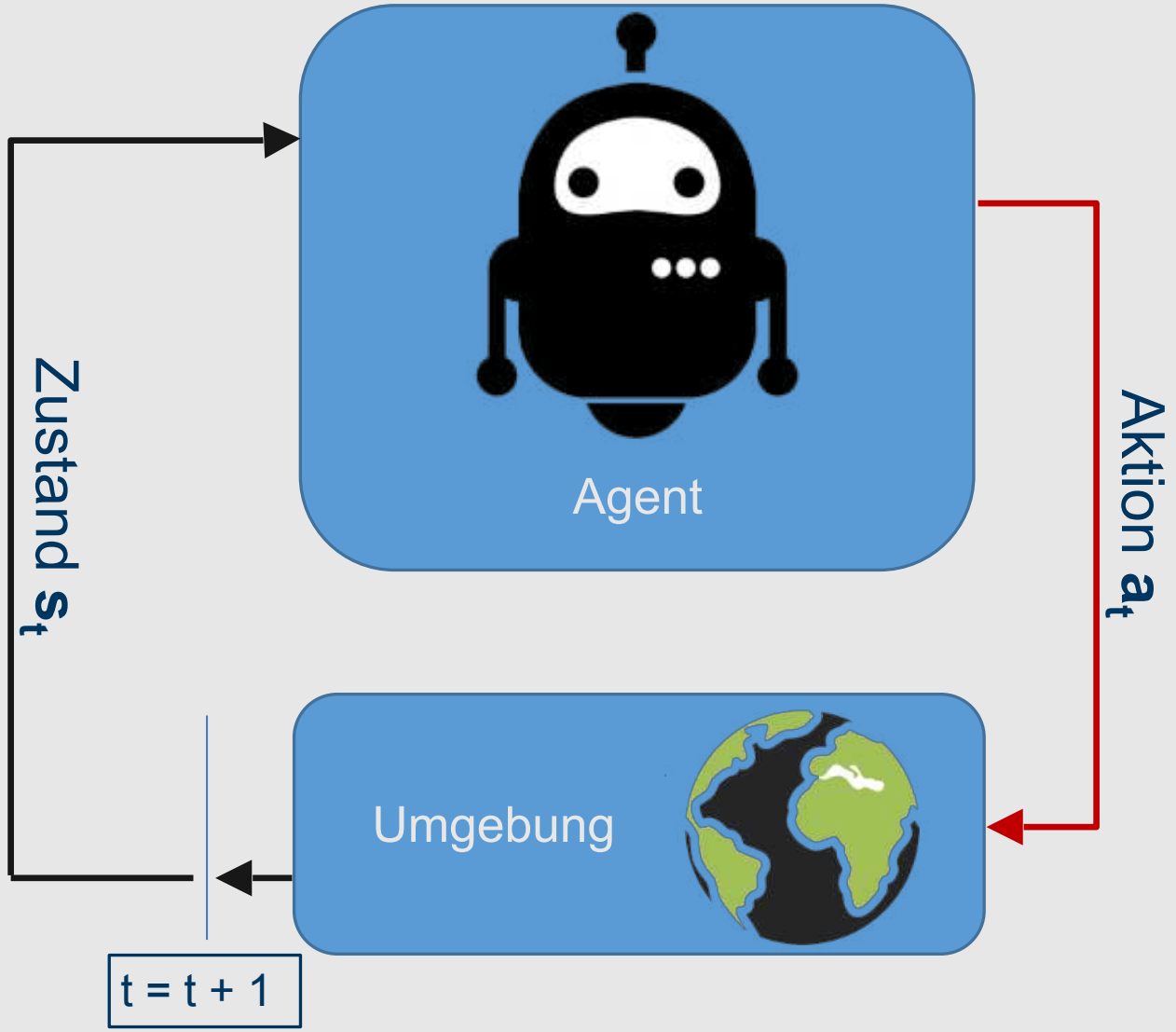
Abschluss

Abschluss



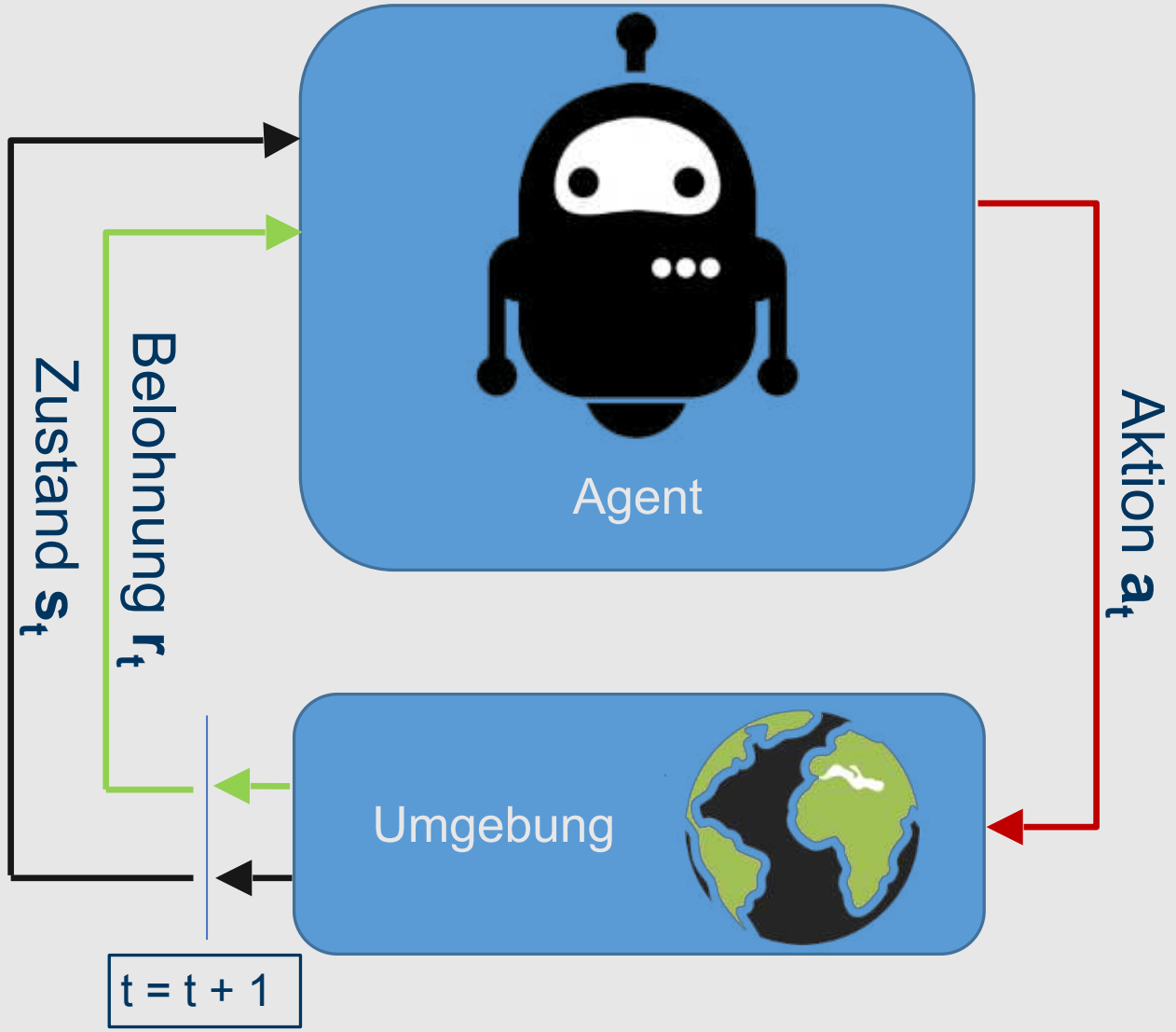
- hat Zustände

Abschluss



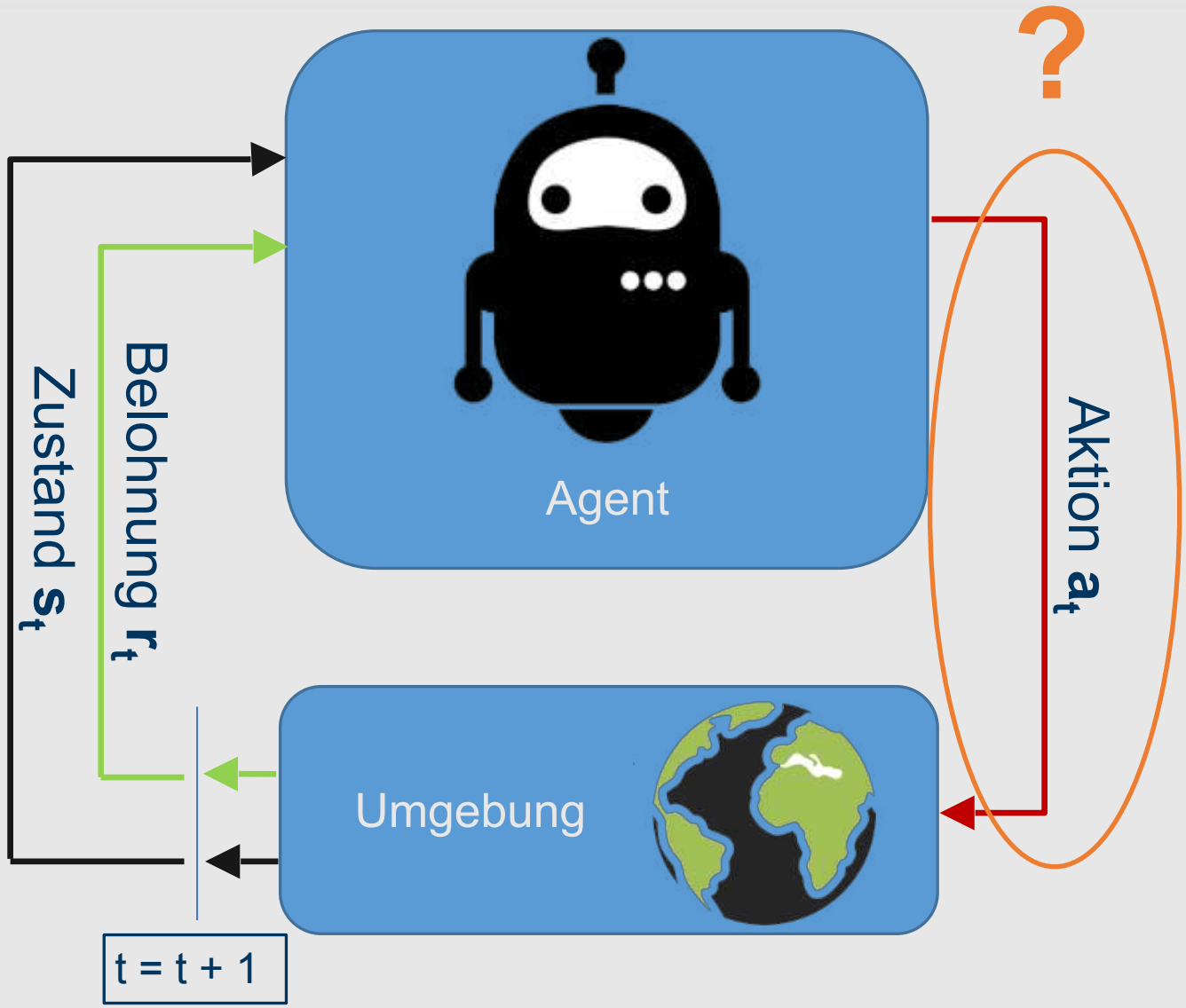
- hat Zustände
- hat Aktionen

Abschluss



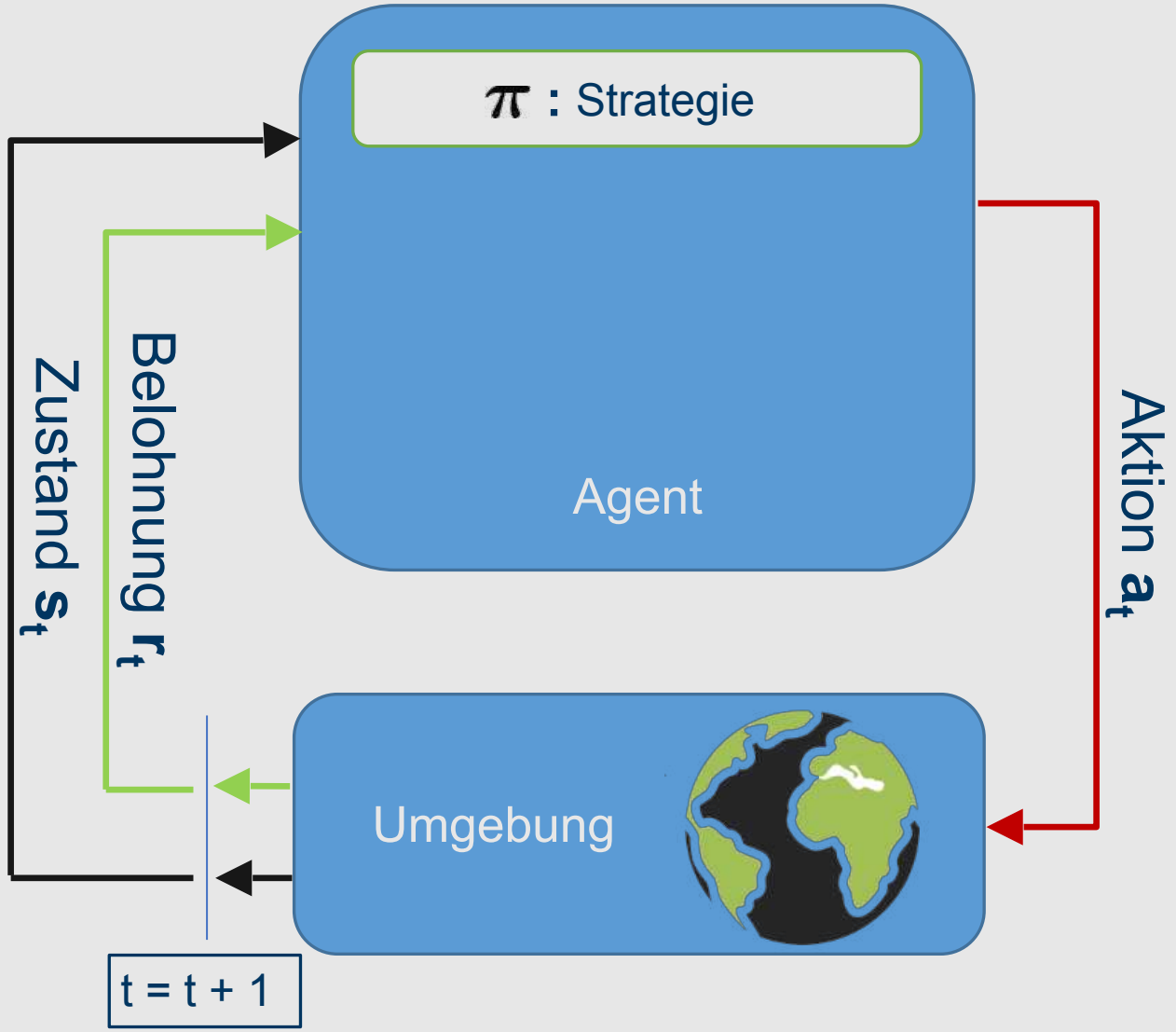
- hat Zustände
- hat Aktionen
- erhält Belohnung von der Umgebung

Abschluss



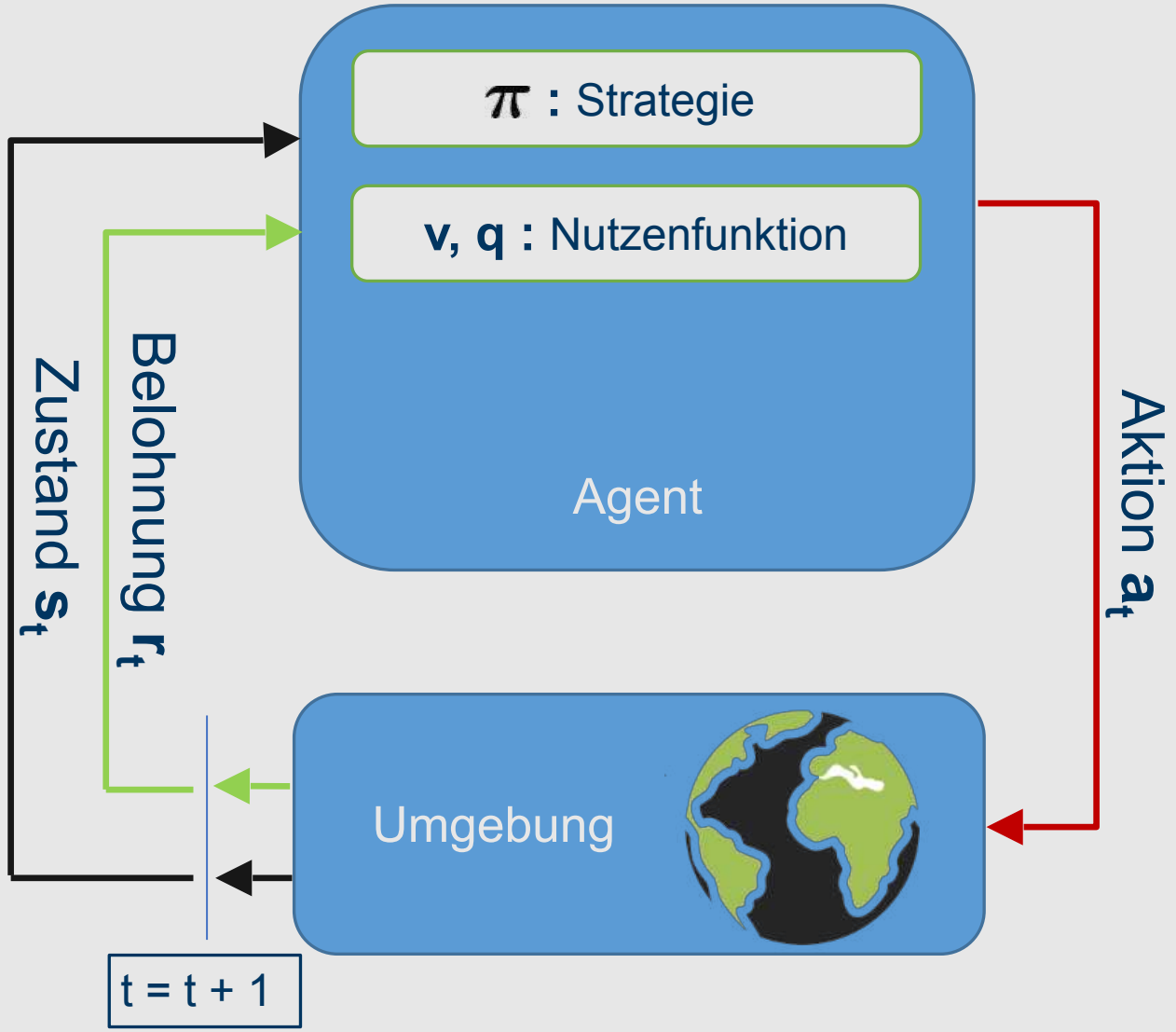
- hat Zustände
- hat Aktionen
- erhält Belohnung von der Umgebung

Abschluss



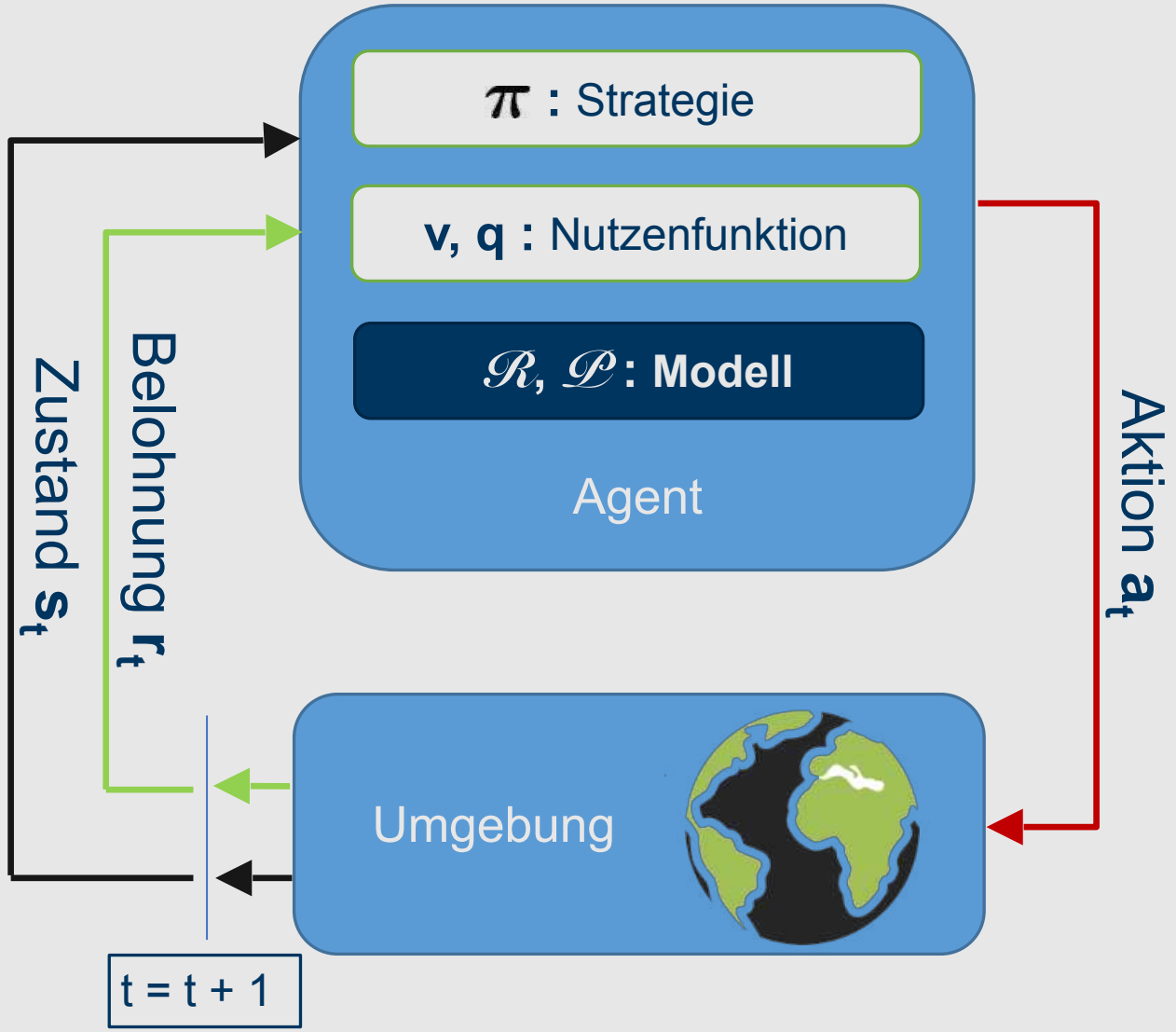
- hat Zustände
- hat Aktionen
- erhält Belohnung von der Umgebung
- verhält sich nach $\pi(s)$

Abschluss



- hat Zustände
- hat Aktionen
- erhält Belohnung von der Umgebung
- verhält sich nach $\pi(s)$
- **v(s):** Nutzen von Zustand **s**
- **q(s, a):** Nutzen von Aktion **a** in **s**

Abschluss



- hat Zustände
- hat Aktionen
- erhält Belohnung von der Umgebung
- verhält sich nach **π(s)**
- **v(s):** Nutzen von Zustand **s**
- **q(s, a):** Nutzen von Aktion **a** in **s**
- **ℛ: Belohnungsfunktion**
- **𝒫: sagt nächsten Zustand s‘ voraus**

Fragen ?