Maximum Entropy Markov Models (MEMM) for Information Extraction and Segmentation

Roman Firstein

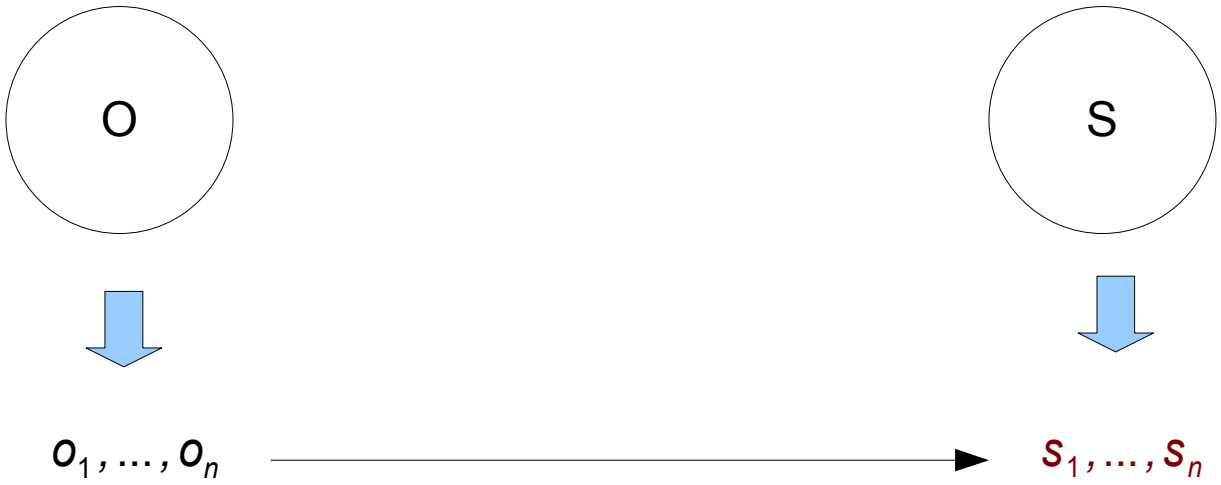
Vortrag

- Vorwort
- Modell
- (Maximale) Entropie am Beispiel der Urne
- Binäres Feature
- Bedingte Entropie
- Maximale Entropie für unseres Modell

Vorwort (1)

Worum geht es überhaupt?!

Gegeben:
- Menge O und S von Objekten (disjunkt)
- W´keitsraum als Teilmenge der Relationen $(s_1, \ldots, s_j, o_1, \ldots, o_j) \forall\, j$
- Stichprobe

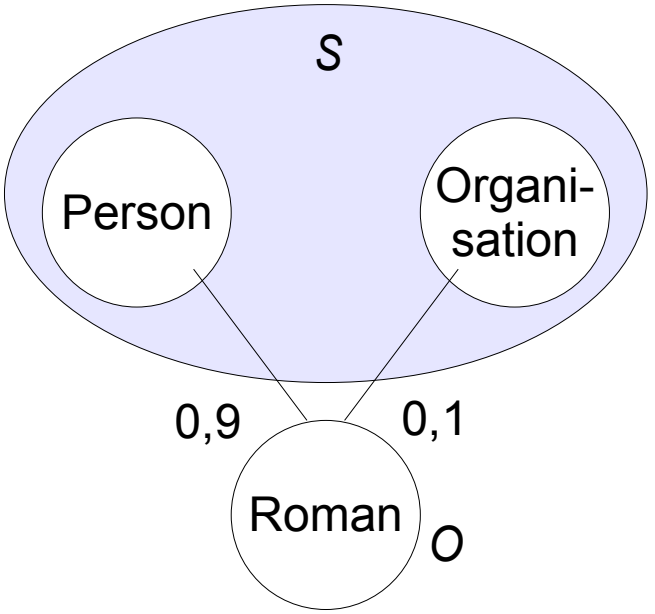


Gesucht:
- Wahrscheinlichkeitsfunktion $P(s_i, o_i)$, $1 \leqslant i \leqslant j$, so dass $(s_1, \ldots, s_j, o_1, \ldots, o_j)$ maximal ist.

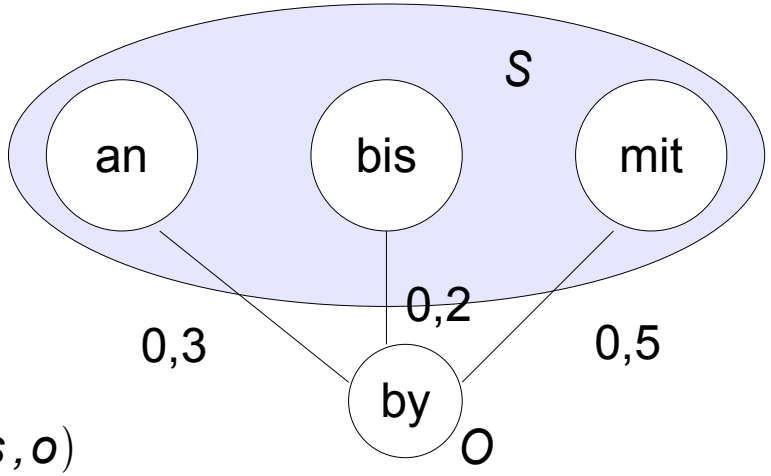
Vorwort (2)

Beispiele

1. Rolle Erkennung:

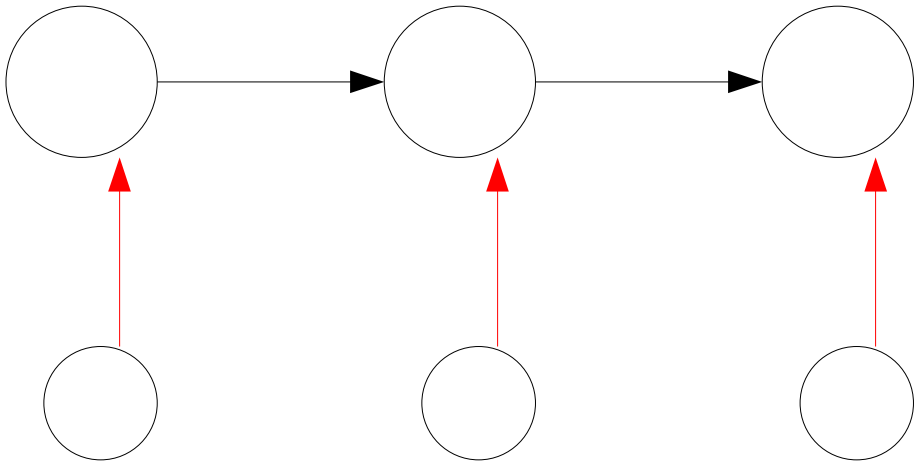


2. Übersetzung aus dem englischen:



Zahlen sind die Wahscheinlickeiten $P(s,o)$

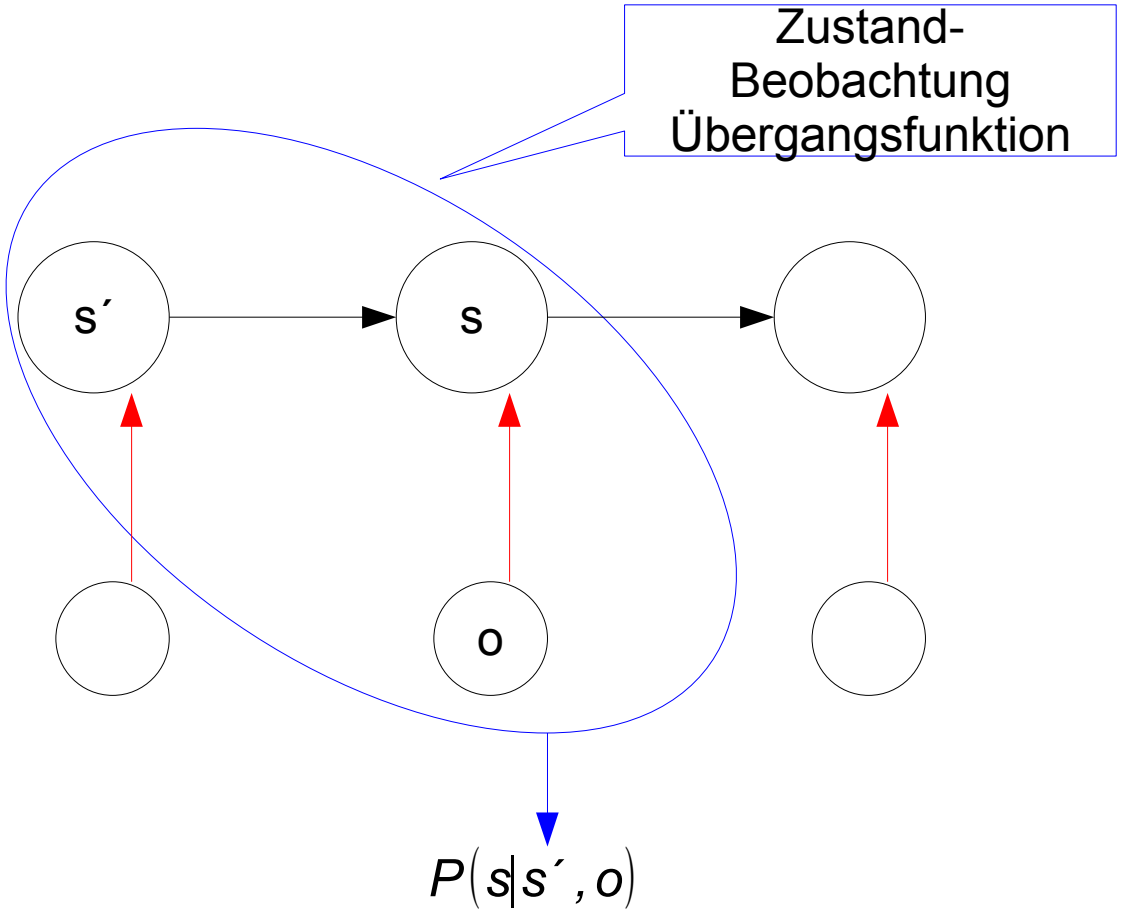
Das NEUE Modell (MEMM)



Zustand hängt jetzt von der Beobachtung und von dem vorherigen Zustand ab

Vorteil:
bei der Übersetzung kann man auch gucken wie das vorherige Wort übersetzt wurde.

Das NEUE Modell (MEMM)



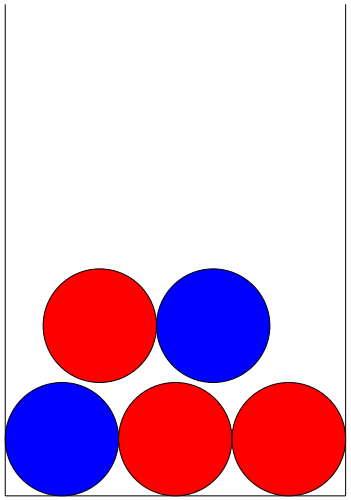
$P(s|s´,o)$ unterteilen wir in $|S|$ Funktionen $P_{s´}(s|o)$.

Entropie

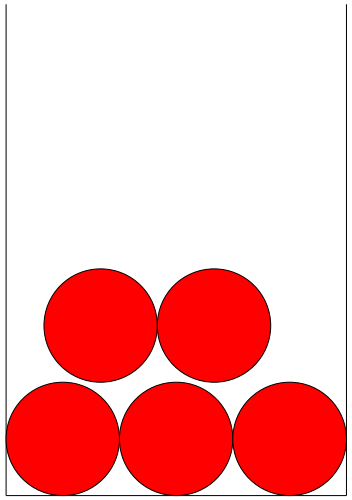
Frage: Wenn man die Verteilungen in der Urne nicht kennt, bei welcher der beiden Urnen wird die Überraschung größer?

Entropie H einer Wahrscheinlichkeitsverteilung X: Maß für die Unvorhersagbarkeit eines Ereignisses.

$$H(X) = -\sum_{i=1}^{|\{rot, blau\}|} p_i \cdot \log_2(p_i)$$



$$H(X) = -(0{,}6\log_2(0{,}6) + 0{,}4\log_2(0{,}4)) \approx 0{,}97$$



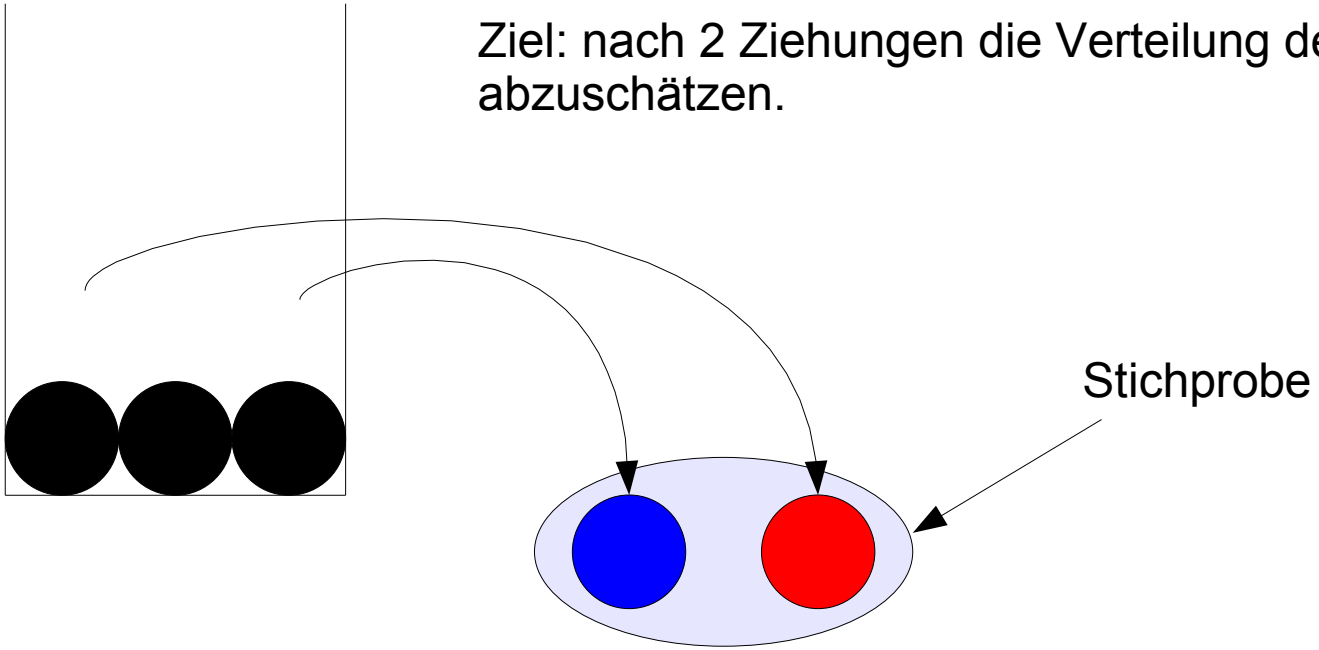
$$H(X) = -(\log_2(1)) = 0$$

Bei der (möglichst) Gleichverteilung wird die maximale Entropie erreicht !!!

Maximale Entropie

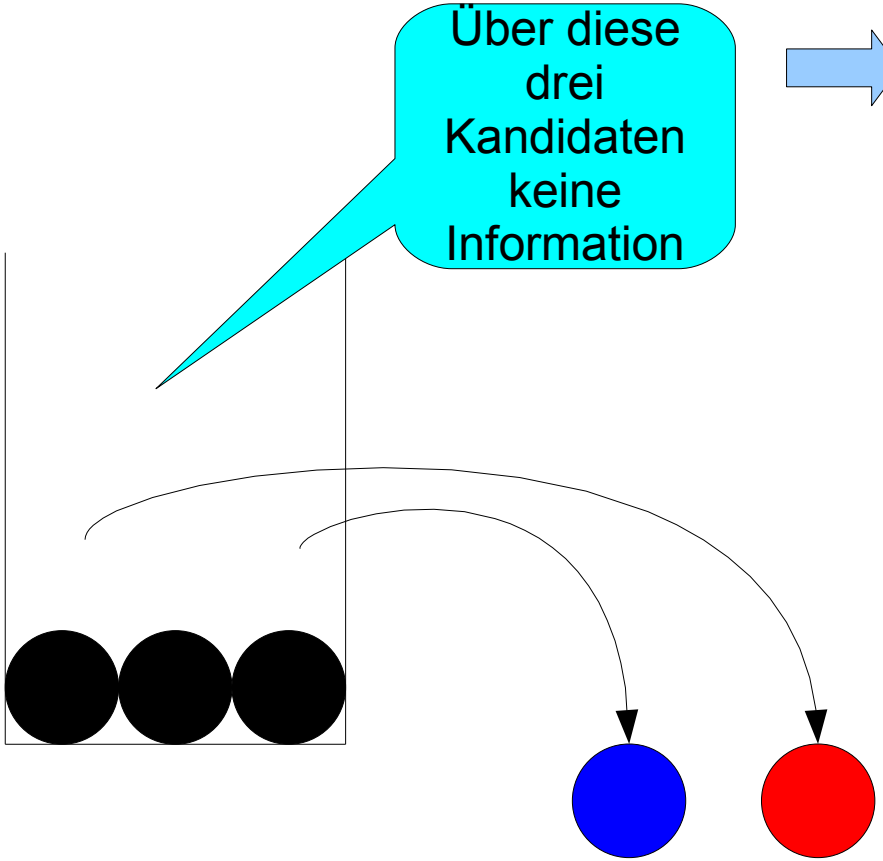
Gegeben: Die Urne enthält 5 Kugeln (rote und blaue). Die Anzahl der Kugeln sei bekannt, aber der jeweiligen Farben nicht.

Ziel: nach 2 Ziehungen die Verteilung der Kugeln abzuschätzen.

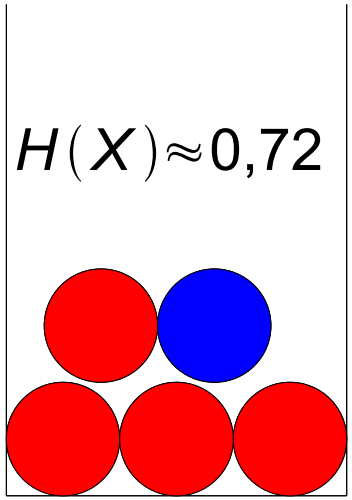


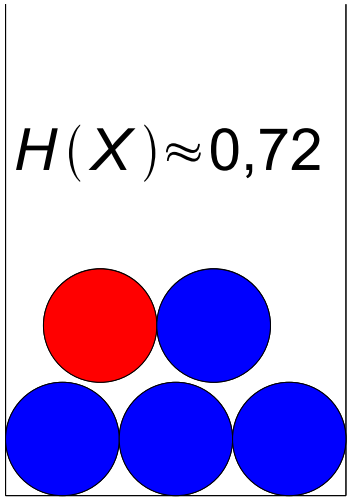
Beobachtung: die Urne enthält mindestens eine rote und eine blaue Kugel.

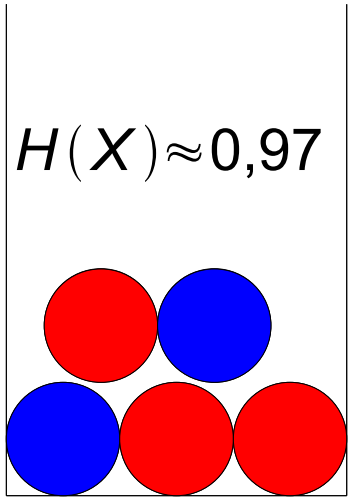
Maximale Entropie

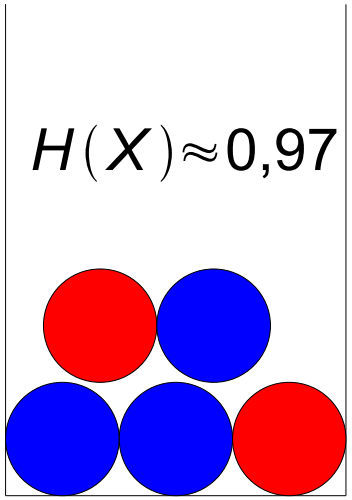


Verteilungen, die mit der Stichprobe vereinbart sind



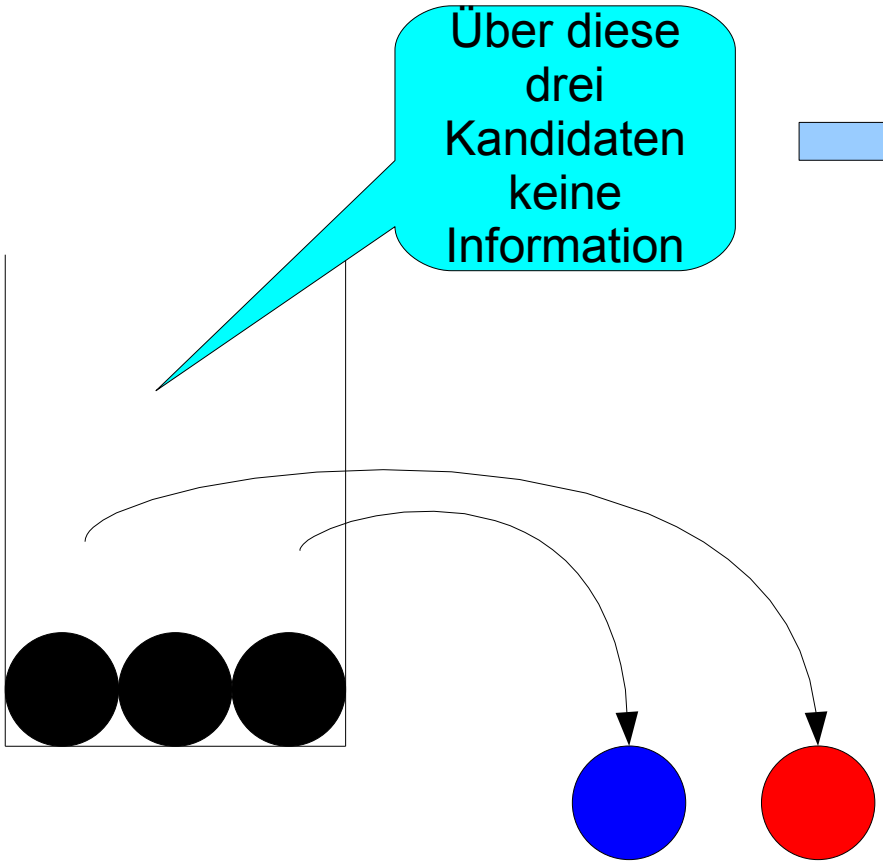






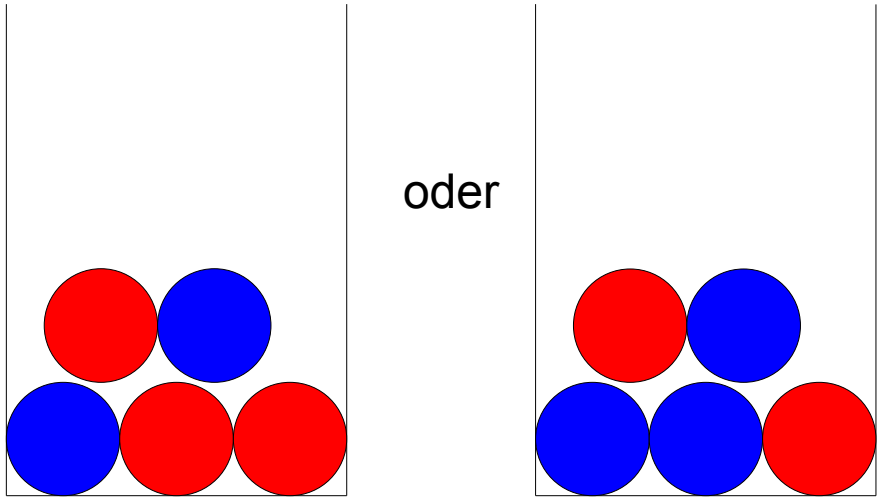
$$H(X) = -\sum_{i=1}^{|\{rot, blau\}|} p_i \cdot \log_2(p_i)$$

Maximale Entropie



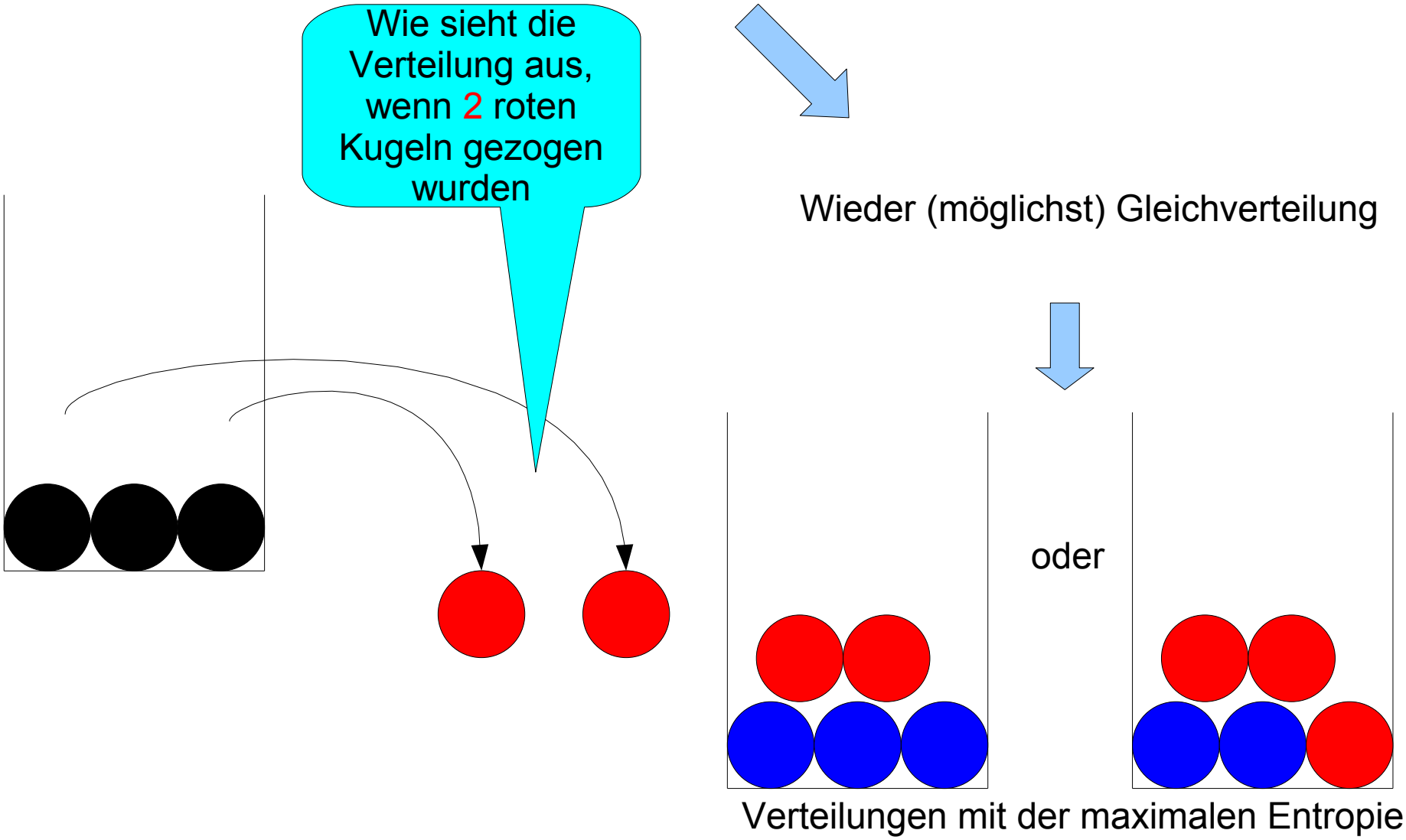
Schätzen wir die Verteilung in der Urne so ab, dass die Entropie H maximal wird.

(möglichst) Gleichverteilung

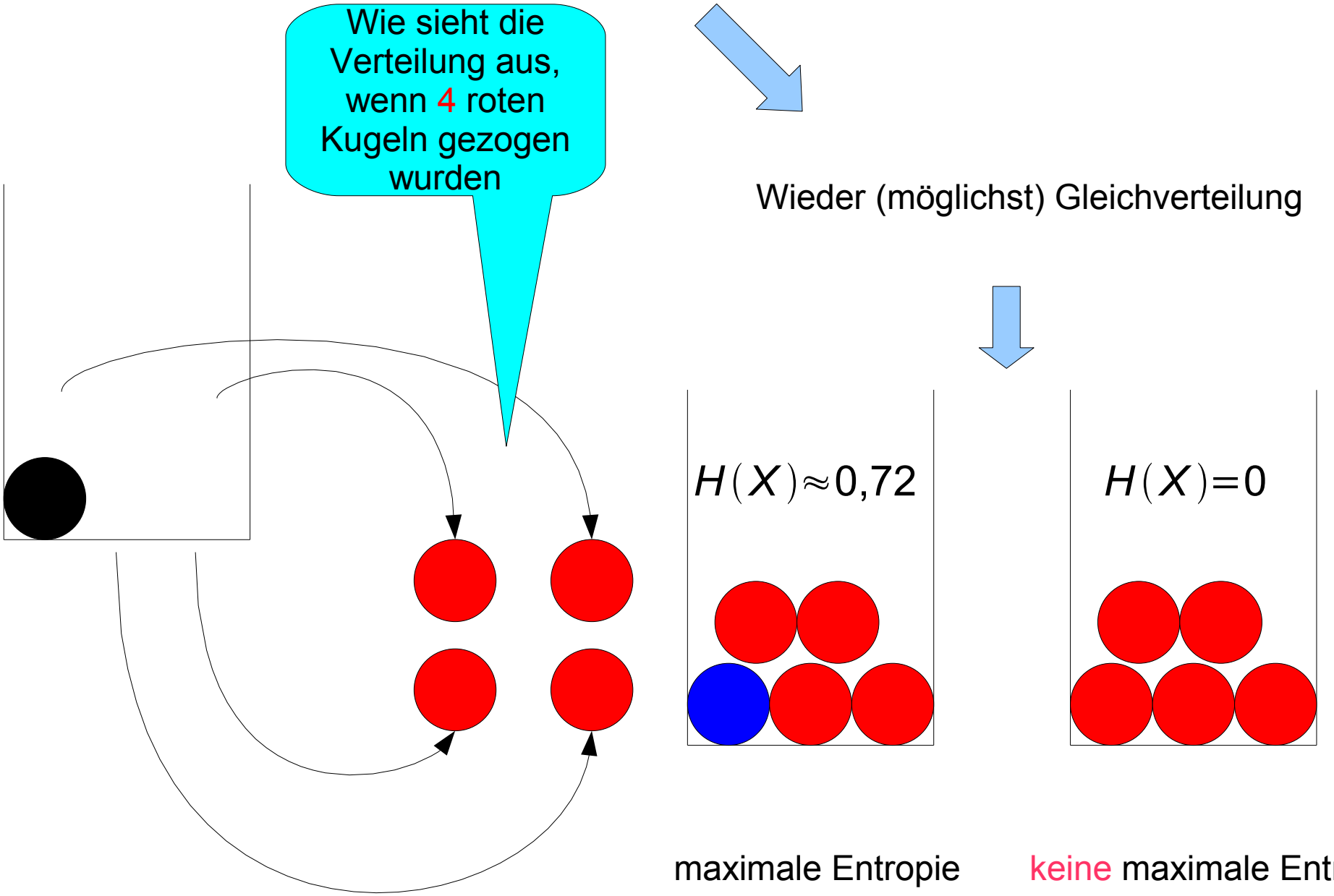


Verteilungen mit der maximalen Entropie

Maximale Entropie



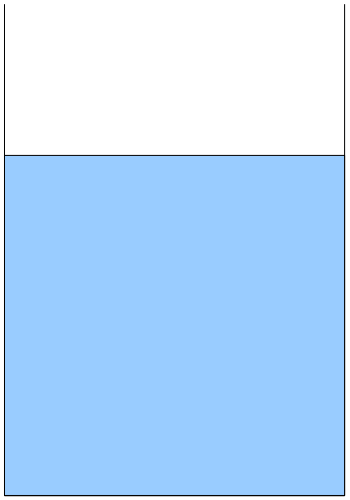
Maximale Entropie



Analogie mit der Urne und zurück zum MEMM

Die Urne enthält Relationen der Form $(s^j , o^j) \forall\, j$

STICHPROBE



Zuordnung per Hand ausgeführt

$(s_1, \ldots, s_n \;,\; o_1, \ldots, o_n)$

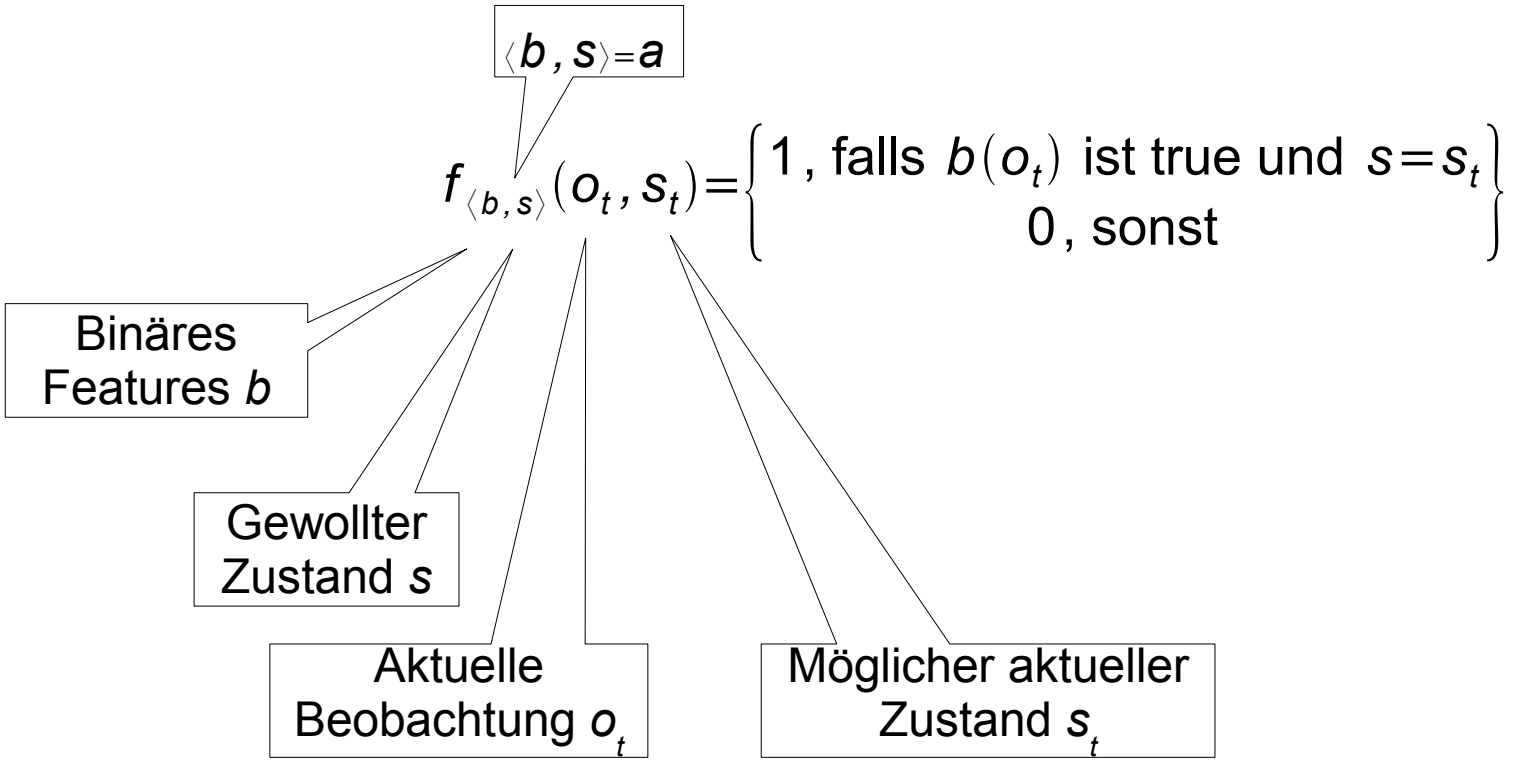
$(s_1, \ldots, s_m \;,\; o_1, \ldots, o_m)$

$(s_1, \ldots, s_k \;,\; o_1, \ldots, o_k)$

da man hier eine s-Folge und eine o-Folge hat, geht es hier um eine bedingte Entropie

Das binäre „Feature“

Stichprobe durch Charakteristische Funktion *f* kodieren,
die einer Beobachtung „1“ oder „0“ zuordnet:



Beispiel

$O=\{she, is, by, me\}$ und $S=\{sie, ist, bei, mit, mir\}$

Stichprobe:

...she is by me... → ...sie ist *mit* mir...
...she is by me... → ...sie ist *bei* mir...
...she is by me... → ...sie ist *bei* mir...
...she is by me... → ...sie ist *mit* mir...
...she is by me... → ...sie ist *bei* mir...

$$f_{\langle b=\text{das Wort ist by}, s=mit\rangle}(o_t, s_t)=\begin{cases}1, \text{ falls } o_t=by \text{ und die Übersetzung von } s_t \text{ ist } mit\\ 0, \text{ sonst}\end{cases}$$

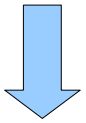
$$f_{\langle b=\text{das Wort ist by}, s=bei\rangle}(o_t, s_t)=\begin{cases}1, \text{ falls } o_t=by \text{ und die Übersetzung von } s_t \text{ ist } bei\\ 0, \text{ sonst}\end{cases}$$

$$f_{\langle b=\text{das Wort ist she}, s=sie\rangle}(o_t, s_t)=\begin{cases}1, \text{ falls } o_t=she \text{ und die Übersetzung von } s_t \text{ ist } sie\\ 0, \text{ sonst}\end{cases}$$

...

Bedingte Entropie

$$H_{s'}(X) = -\sum_{S} P_{s'}(s|o) \cdot \log_2(P_{s'}(s|o)) \quad \text{bei gegebenem } o$$



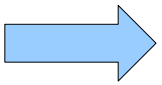
Ziel: Maximierung dieser Funktion
unter den Nebenbedienungen:

$$1 = \sum_{S} P_{s'}(s|o) \quad \forall s', \text{ bei gegebenem } o$$

$$f_a(o, s_k) = \sum_{S} f_a(o, s) \cdot P_{s'}(s|o) \quad \forall s', s_k, a \text{ bei gegebenem } o$$

Vollständig bzgl. P

Vollständig bzgl. f_a



LAGRANGE

Lagrangesche Multiplikatorenmethode zur Maximierung einer Funktion unter Nebenbedingungen

$$L_{s'}(\lambda, P_{s'}) = -\sum_{S} P_{s'}(s|o)\cdot\log_2(P_{s'}(s|o)) + \sum_{a}\lambda_a\left(\sum_{S} f_a(o,s)\cdot P_{s'}(s|o) - f_a(o,s_k)\right)$$

$H_{s'}$ (under the first term) $\qquad 0$ (under the second term)

➡ Erste Ableitung nach *P* bilden und gleich Null setzen:

$$\frac{\partial L_{s'}(\lambda, P_{s'})}{\partial P_{s'}(s|o)} = \left(-1 - \log_2 P_{s'}(s|o) + \sum_{S}\lambda_a\cdot f_a(o,s)\cdot P_{s'}(s|o)\right) = 0$$

Ohne Zwischenschritte

➡

$$P_{s'}(s|o) = \frac{1}{\sum_{s}\exp\left(\sum_{a}\lambda_a f_a(o,s)\right)}\exp\left(\sum_{a}\lambda_a f_a(o,s)\right)$$

Verteilung mit der maximalen Entropie (Ohne Beweis):

$$P_{s'}(s|o) = \frac{1}{Z(o,s')} \exp\left(\sum_a \lambda_a f_a(o,s)\right)$$

Exponentielle Darstellung der W´keit

Normalisierung, damit die Summe gleich 1 wird

Lagrange - Multiplikator, der mit dem GIS Verfahren berechnet wird

$$Z(o,s') = \sum_s \exp\left(\sum_a \lambda_a f_a(o,s)\right)$$

Die Abschätzung der Zustandsfolge aus der Beobachtungsfolge(1)

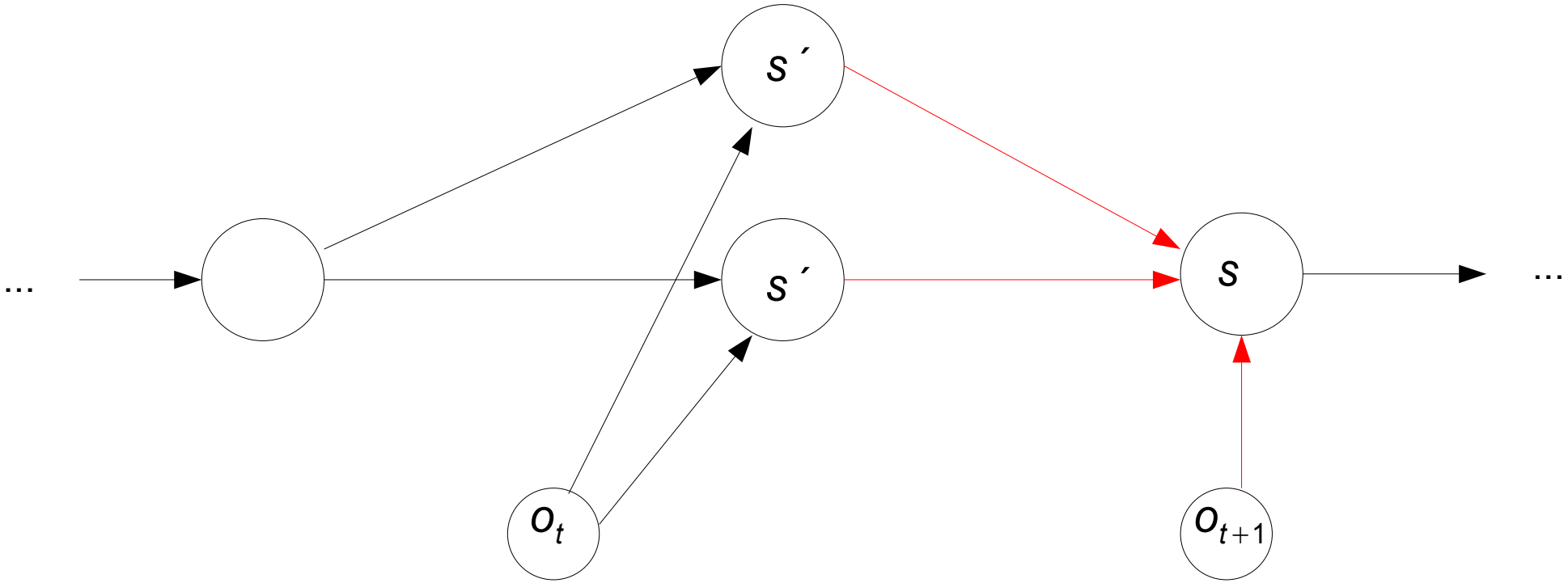
Läuft analog wie bei einem HMM:

$$\delta_{t+1}(s) = \max_{s' \in S} \left\{ \delta_t(s') \cdot P_{s'}(s | o_{t+1}) \right\} \qquad \text{wobei}$$

$\delta_t(s')$: Wahrscheinlichste Zustandsfolge bis zum o_t endet im s'

Die Abschätzung der Zustandsfolge aus der Beobachtungsfolge(2)

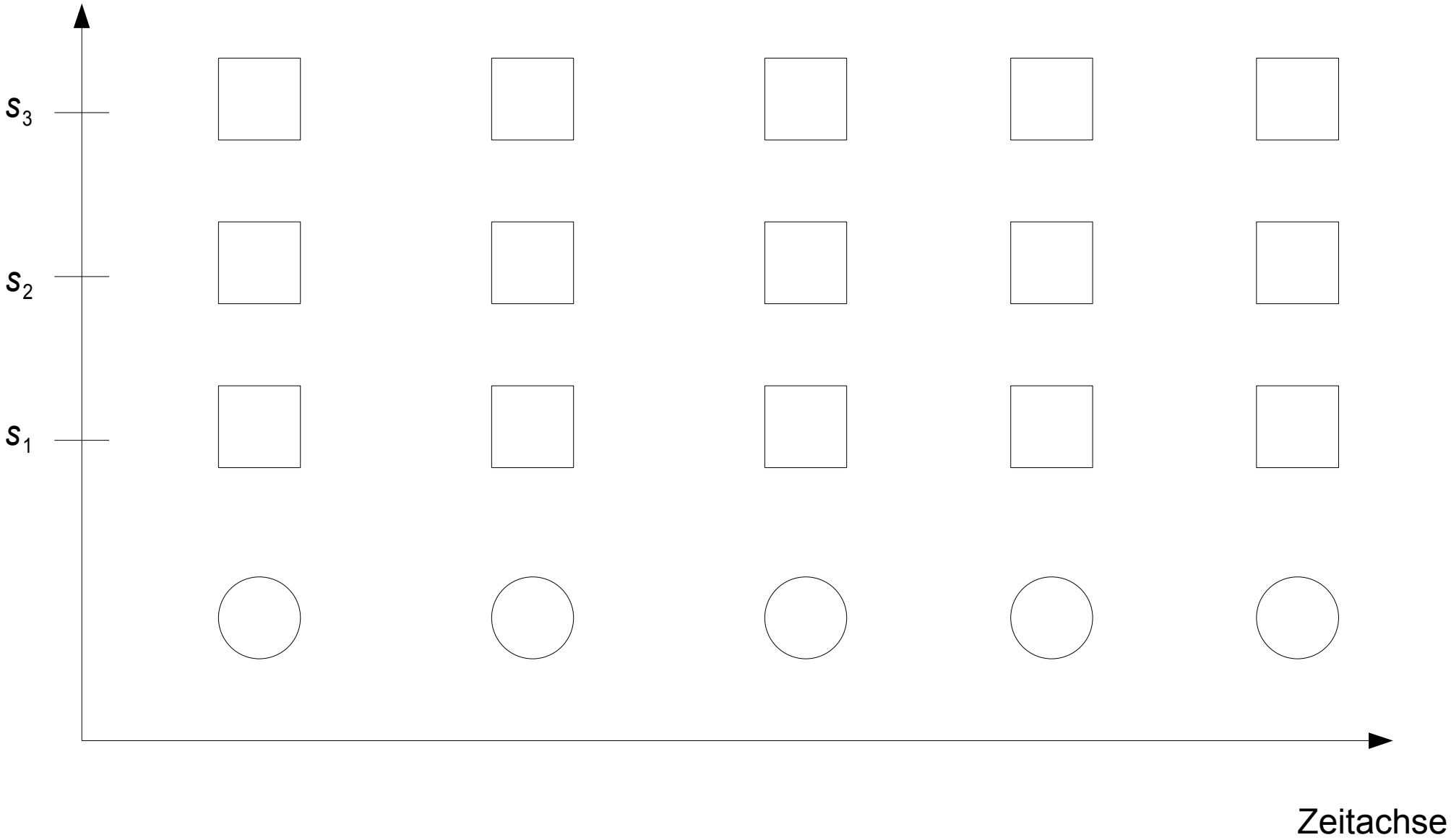
$$\delta_{t+1}(s) = \max_{s' \in S} \left\{ \delta_t(s') \cdot P_{s'}(s | o_{t+1}) \right\}$$



Viterbi Algorithmus für die Abschätzung der wahrscheinlichsten Zustandsfolge aus der Beobachtungsfolge.

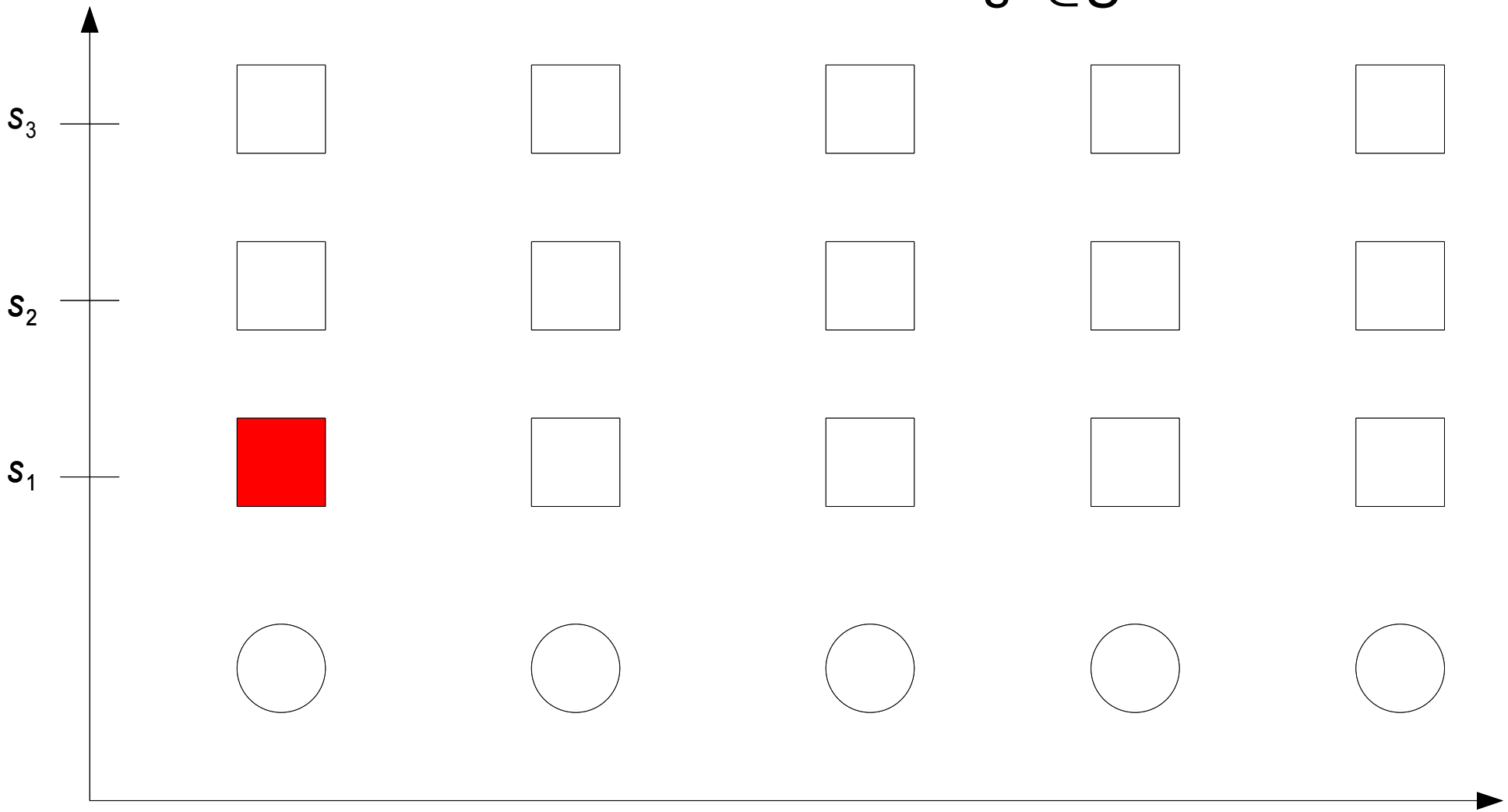
Viterbi Algorithmus

$$\delta_{t+1}(s) = \max_{s' \in S} \left\{ \delta_t(s') \cdot P_{s'}(s|o_{t+1}) \right\}$$



Viterbi Algorithmus

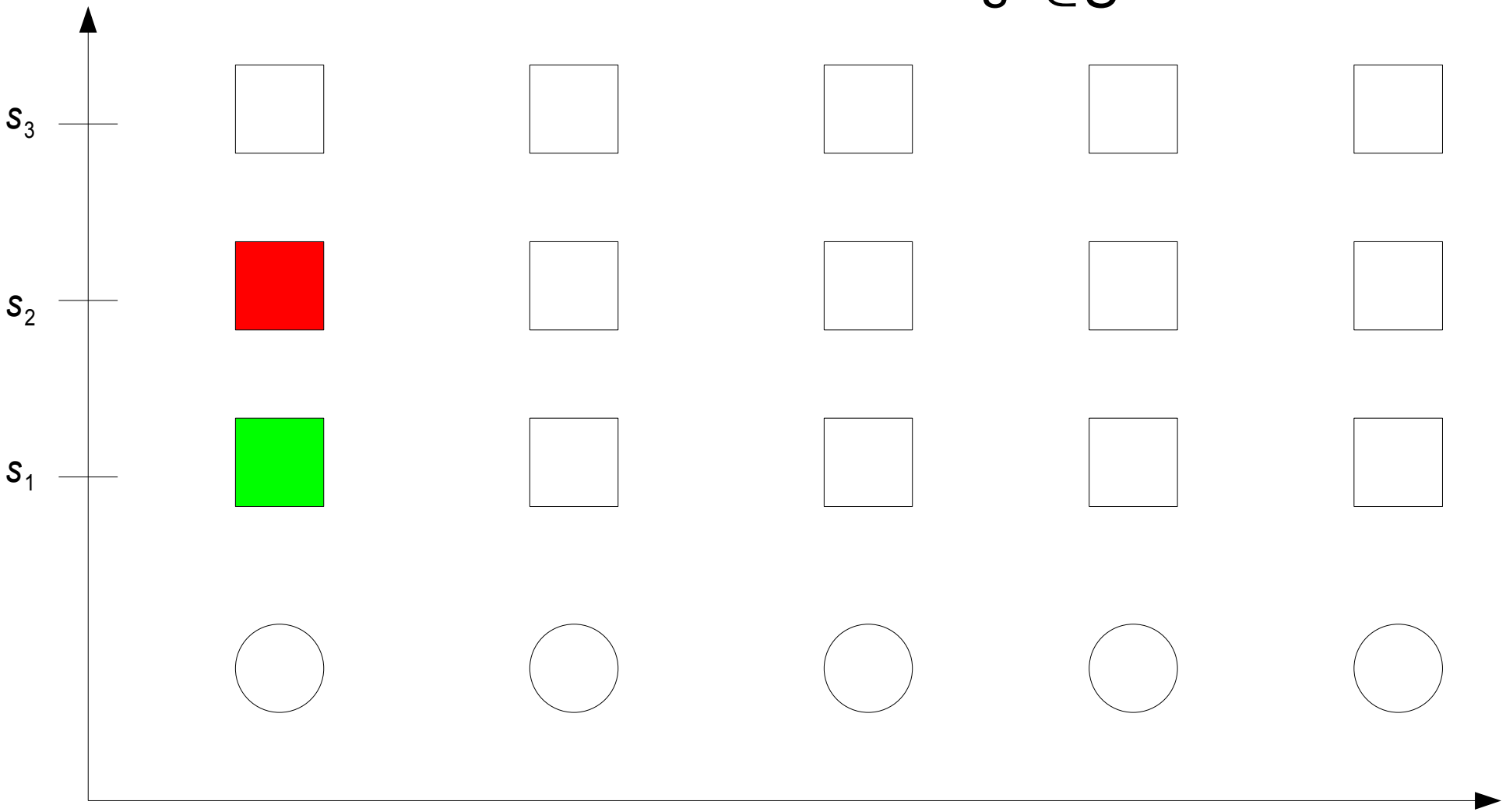
$$\delta_{t+1}(s) = \max_{s' \in S} \left\{ \delta_t(s') \cdot P_{s'}(s|o_{t+1}) \right\}$$



$$\delta_1(s_1) = \max_{s' \in S} \left\{ \delta_0(s') \cdot P_{s'}(s_1|o_1) \right\}$$

Viterbi Algorithmus

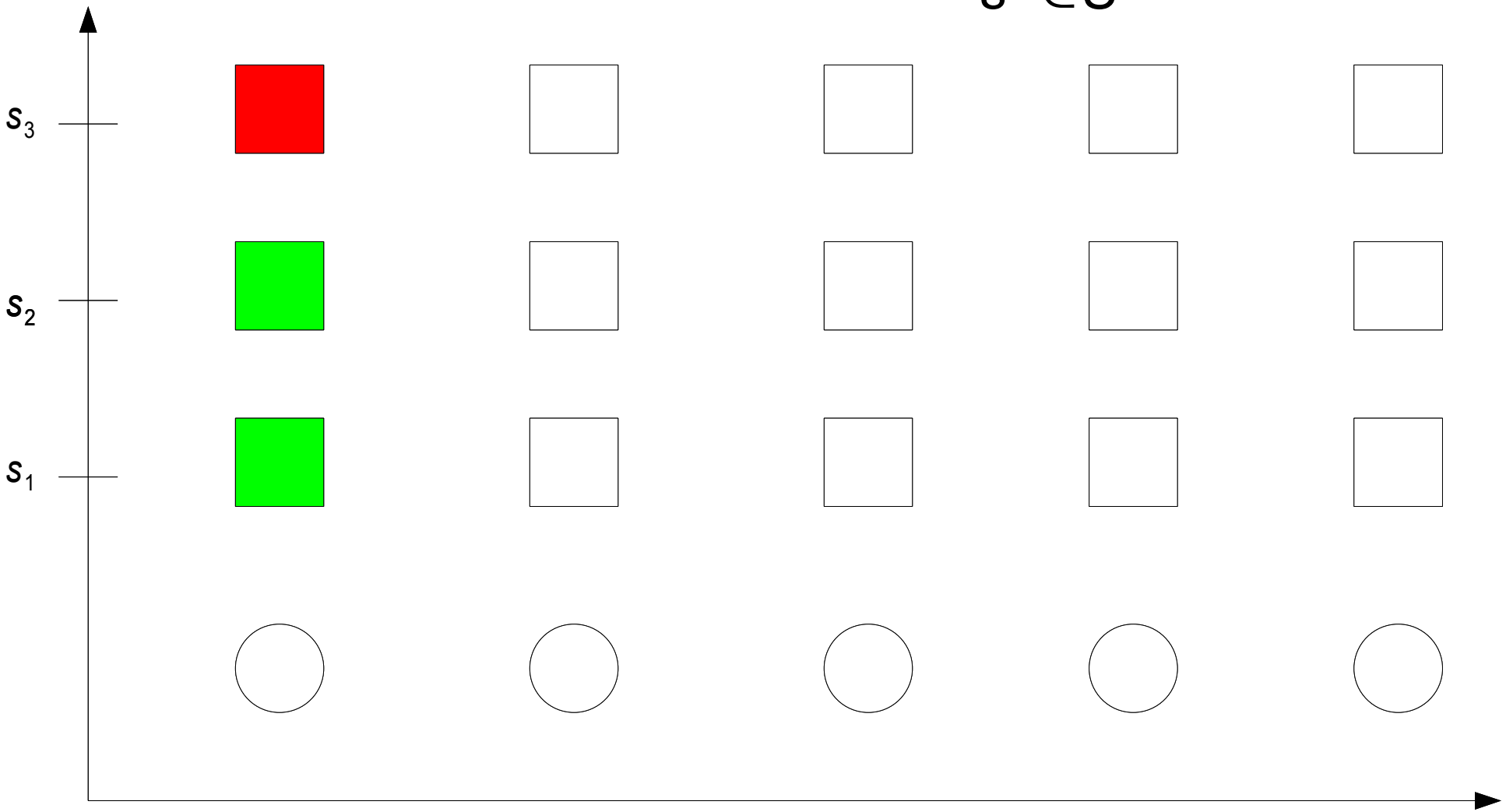
$$\delta_{t+1}(s) = \max_{s' \in S} \left\{ \delta_t(s') \cdot P_{s'}(s|o_{t+1}) \right\}$$



$$\delta_1(s_2) = \max_{s' \in S} \left\{ \delta_0(s') \cdot P_{s'}(s_2|o_1) \right\}$$

Viterbi Algorithmus

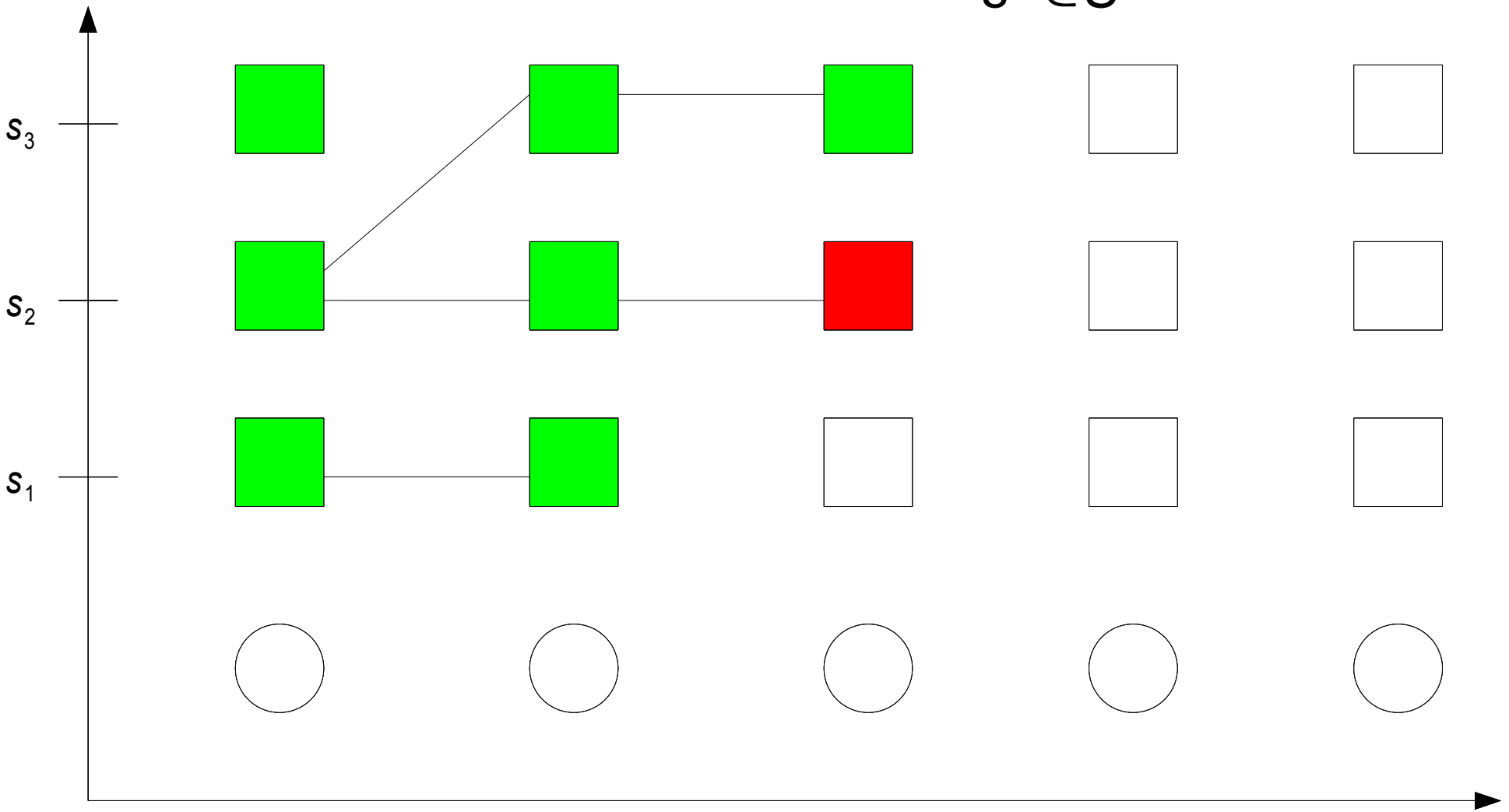
$$\delta_{t+1}(s) = \max_{s' \in S} \left\{ \delta_t(s') \cdot P_{s'}(s|o_{t+1}) \right\}$$



$$\delta_1(s_3) = \max_{s' \in S} \left\{ \delta_0(s') \cdot P_{s'}(s_3|o_1) \right\}$$

Viterbi Algorithmus

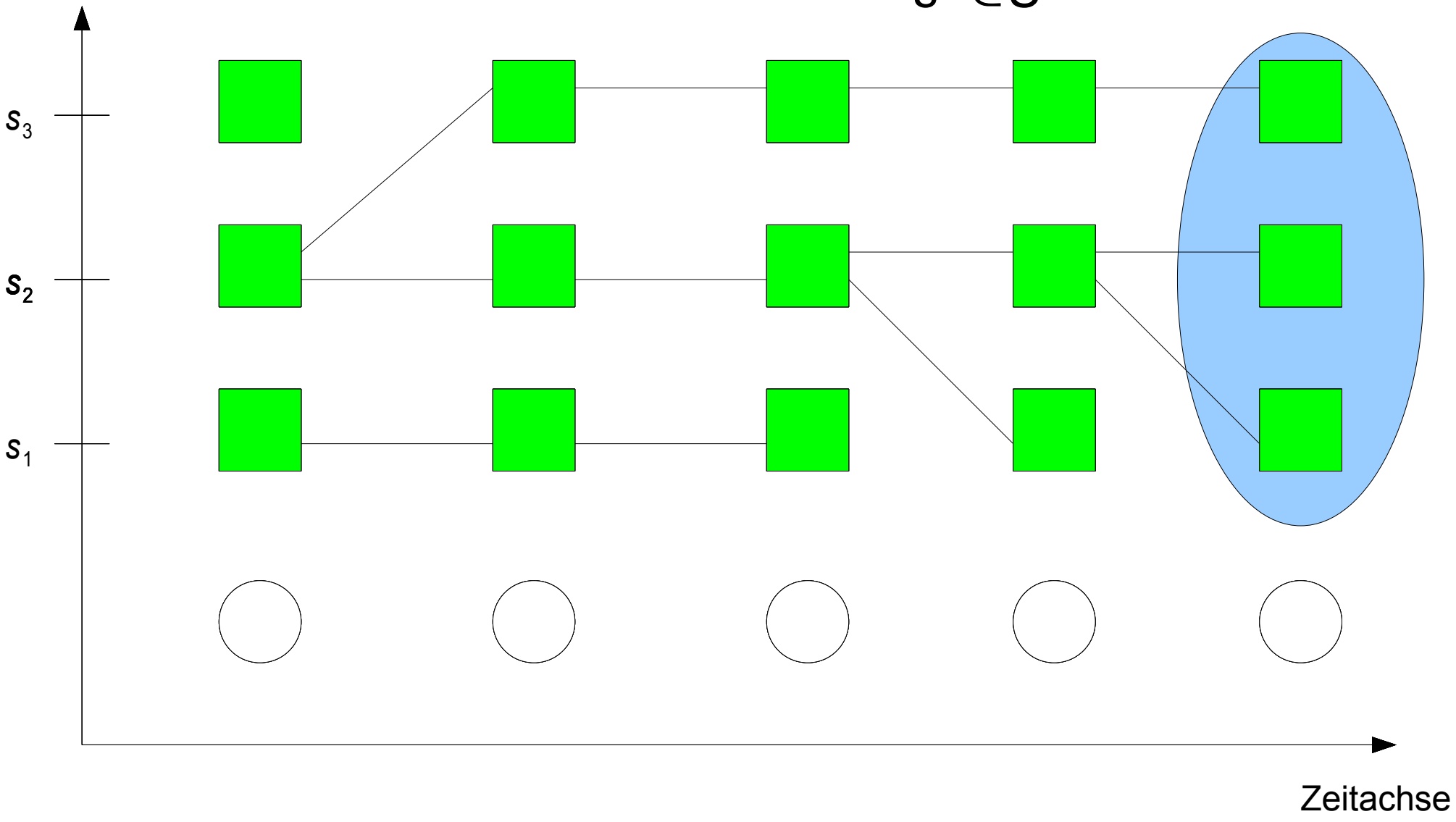
$$\delta_{t+1}(s) = \max_{s' \in S} \left\{ \delta_t(s') \cdot P_{s'}(s|o_{t+1}) \right\}$$



$$\delta_3(s_2) = \max_{s' \in S} \left\{ \delta_2(s') \cdot P_{s'}(s_2|o_3) \right\}$$

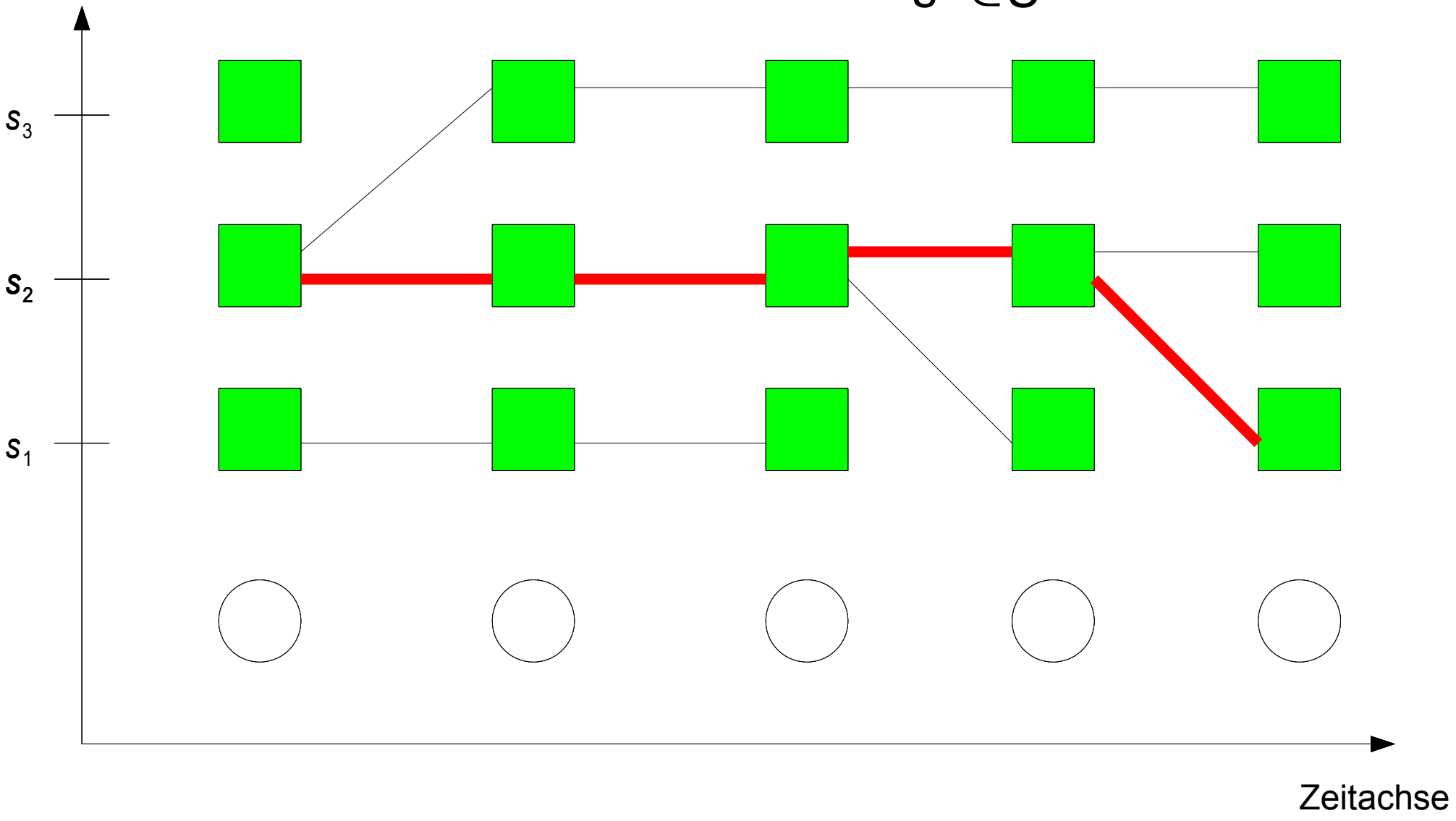
Viterbi Algorithmus

$$\delta_{t+1}(s) = \max_{s' \in S} \left\{ \delta_t(s') \cdot P_{s'}(s | o_{t+1}) \right\}$$



Viterbi Algorithmus

$$\delta_{t+1}(s) = \max_{s' \in S} \left\{ \delta_t(s') \cdot P_{s'}(s | o_{t+1}) \right\}$$

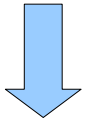


Wie gut ist MEMM?

Ist schon viel besser als das HMM.

Mögliche Verbesserung vom MEMM:

- jedes Element in der Zustandsfolge sollte von der kompletten Beobachtungsfolge abhängen.



Conditional Random Fields (nächster Vortrag)