FSNet learning in partially-observable stochastic environments

Артем Сорокин, Михаил Бурцев
МФТИ
DeepHackLab

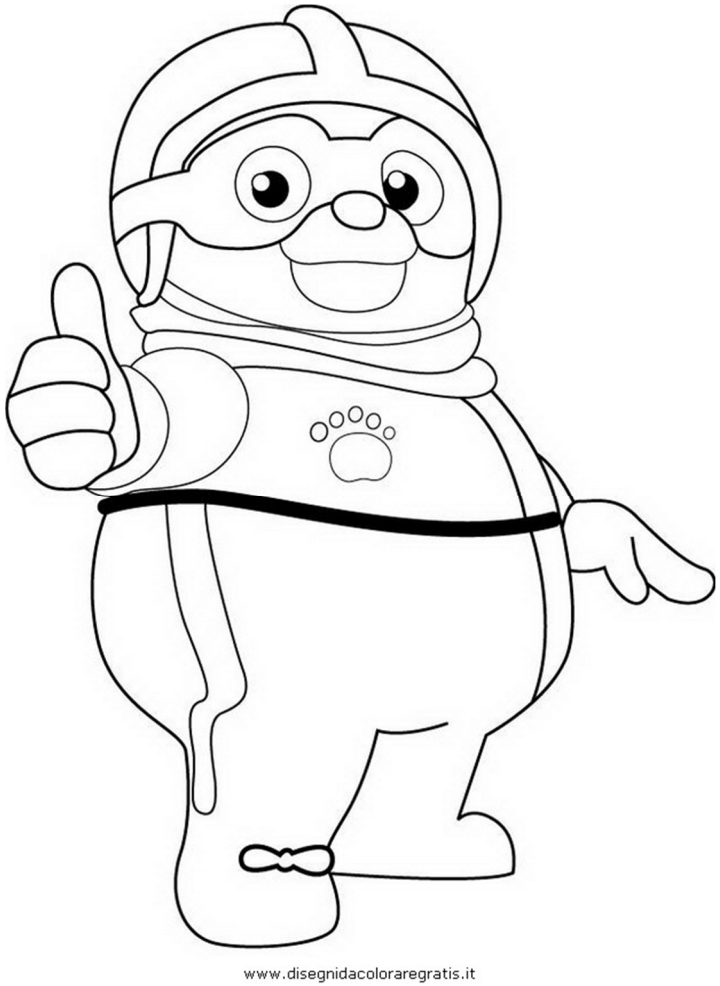
Обучение целенаправленному поведению

Миссия

- Агент должен достичь определенных целей в мире
- Агент
  - воспринимает мир
  - может совершать действия
  - изначально ничего не знает о влиянии действий на состояния мира

Обучение с подкреплением

- Цели задаются, как состояния мира в которых получается награда - *подкрепление*



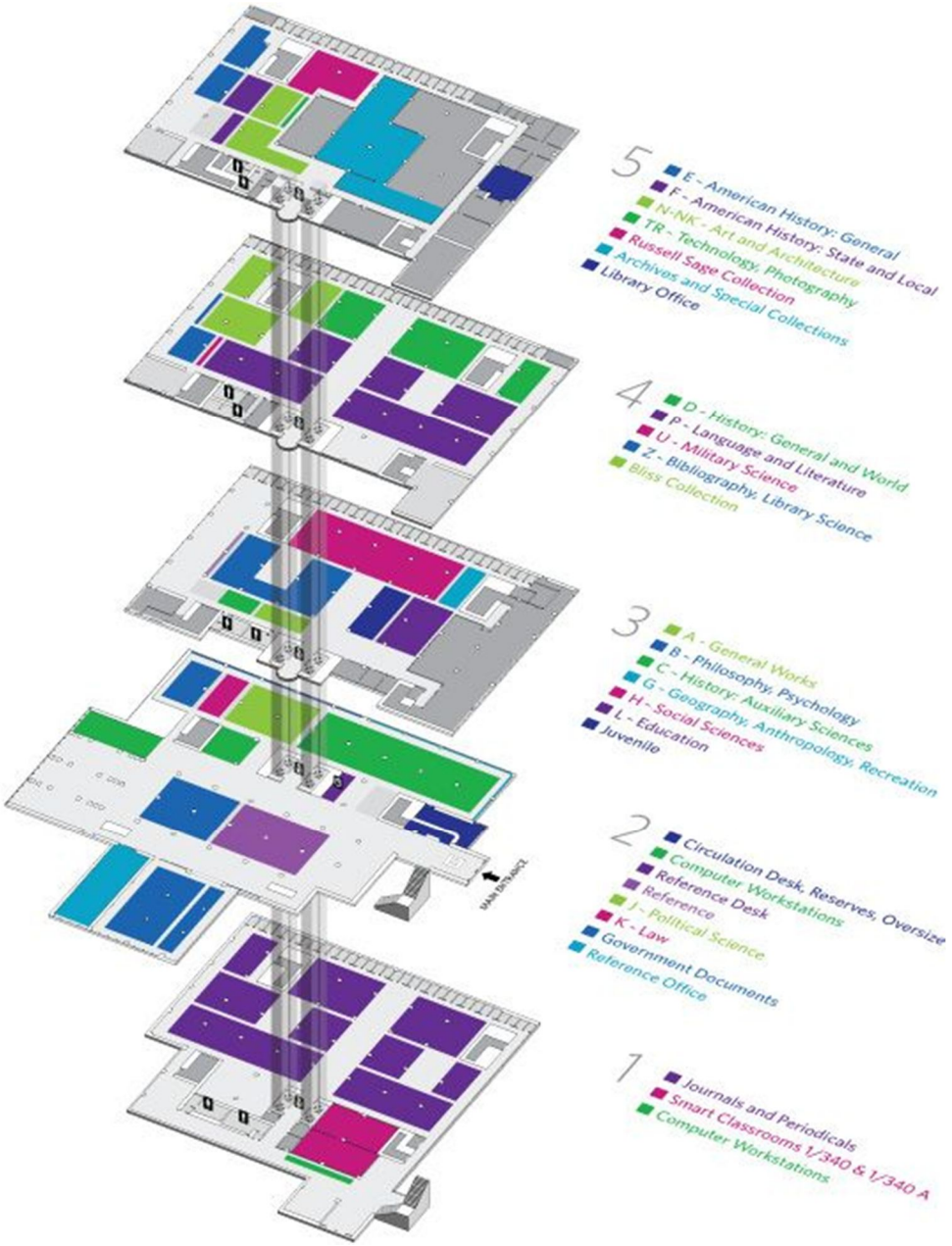
MDP

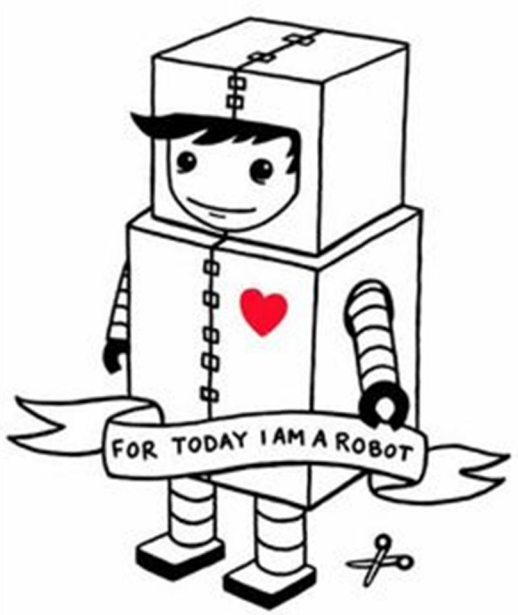
Markov Decision Process (MDP)

$< \mathcal{S}, \mathcal{A}, \mathcal{T}, \mathcal{R} >$, где:

- $\mathcal{S}$ множество состояний мира
- $\mathcal{A}$ множество действий
- $\mathcal{T} : \mathcal{S} \times \mathcal{A} \times \mathcal{S} \to [0, 1]$ вероятности переходов
- $\mathcal{R} : \mathcal{S} \times \mathcal{A} \to \mathbb{R}$ подкрепление

Агент ограничен в восприятии мира





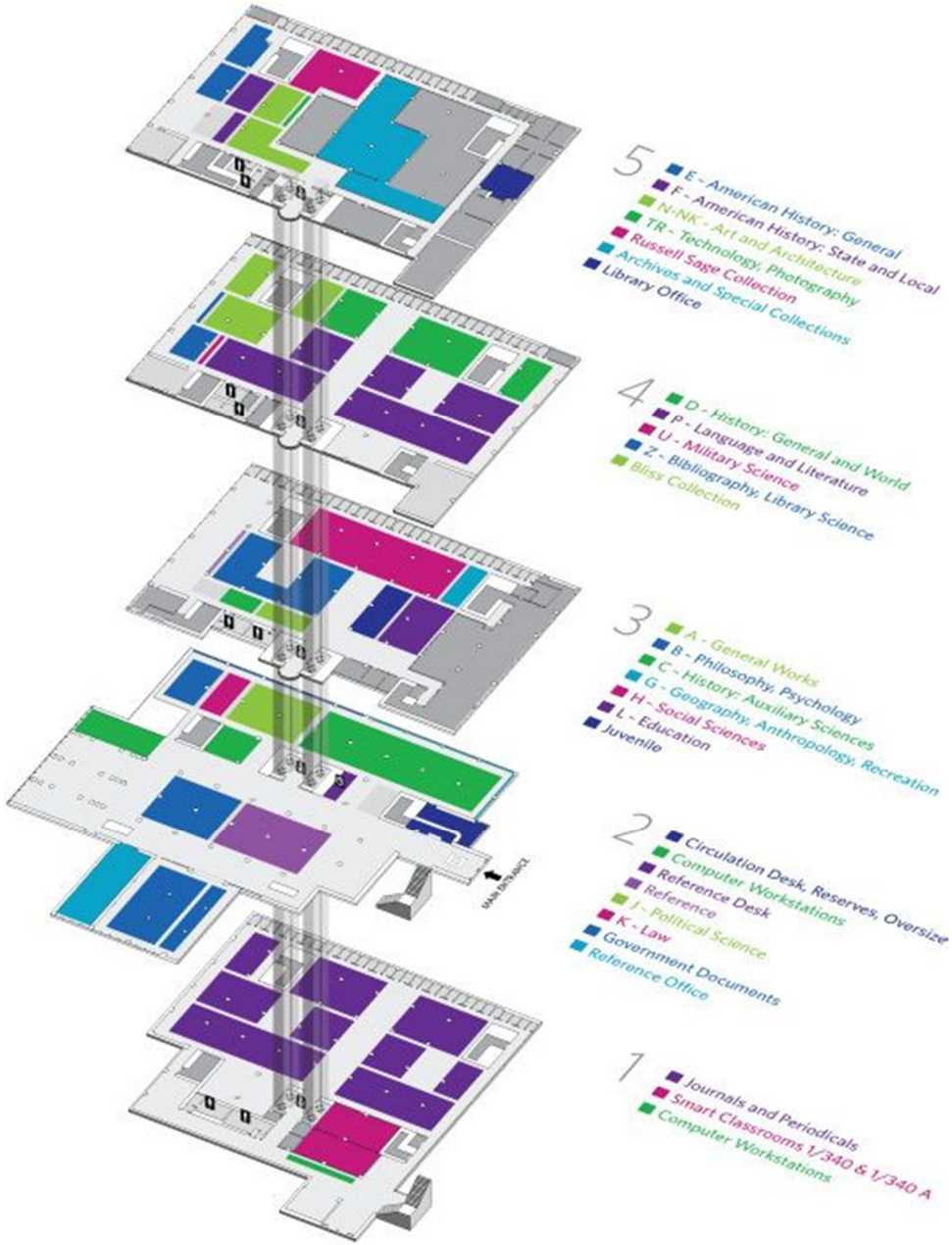
POMDP

Partially Observable Markov Decision Process (POMDP)

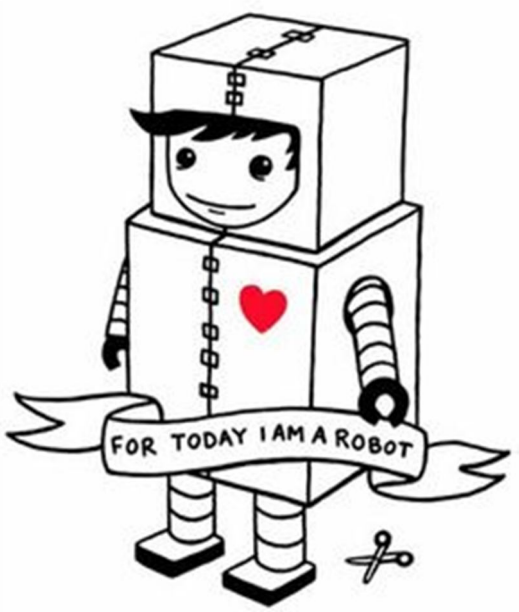
$< \mathcal{S}, \mathcal{A}, \mathcal{T}, \mathcal{R}, \Omega, \mathcal{B} >$, где:

- $\mathcal{S}$ множество состояний мира
- $\mathcal{A}$ множество действий
- $\mathcal{T} : \mathcal{S} \times \mathcal{A} \times \mathcal{S} \to [0, 1]$ вероятности переходов
- $\mathcal{R} : \mathcal{S} \times \mathcal{A} \to \mathbb{R}$ подкрепление
- Ω множество наблюдений
- $\mathcal{B}$ множество убеждений

POMDP


5
E - American History: General
F - American History: State and Local
N-NK - Art and Architecture
TR - Technology, Photography
Russell Sage Collection
Archives and Special Collections
Library Office
4
D - History: General and World
P - Language and Literature
U - Military Science
Z - Bibliography, Library Science
Bliss Collection
3
A - General Works
B - Philosophy, Psychology
C - History: Auxiliary Sciences
G - Geography, Anthropology, Recreation
H - Social Sciences
L - Education
Juvenile
2
Circulation Desk, Reserves, Oversize
Computer Workstations
Reference Desk
Reference
J - Political Science
K - Law
Government Documents
Reference Office
1
Journals and Periodicals
Smart Classrooms 1/340 & 1/340 A
Computer Workstations

$$b$$


FOR TODAY I AM A ROBOT

Основная идея

- На основе предыдущего опыта выбираем действия, которые скорее всего ведут к цели

$$\pi(b, o) = argmax_{a \in \mathcal{A}}(b \cdot Q(o, a))$$

- После каждого действия проверяем в каком «мире» находимся

$$m(b, o, a) \neq o'$$

- Быстро приводим убеждения в соответствии с наблюдениями

$$b_i = 0$$

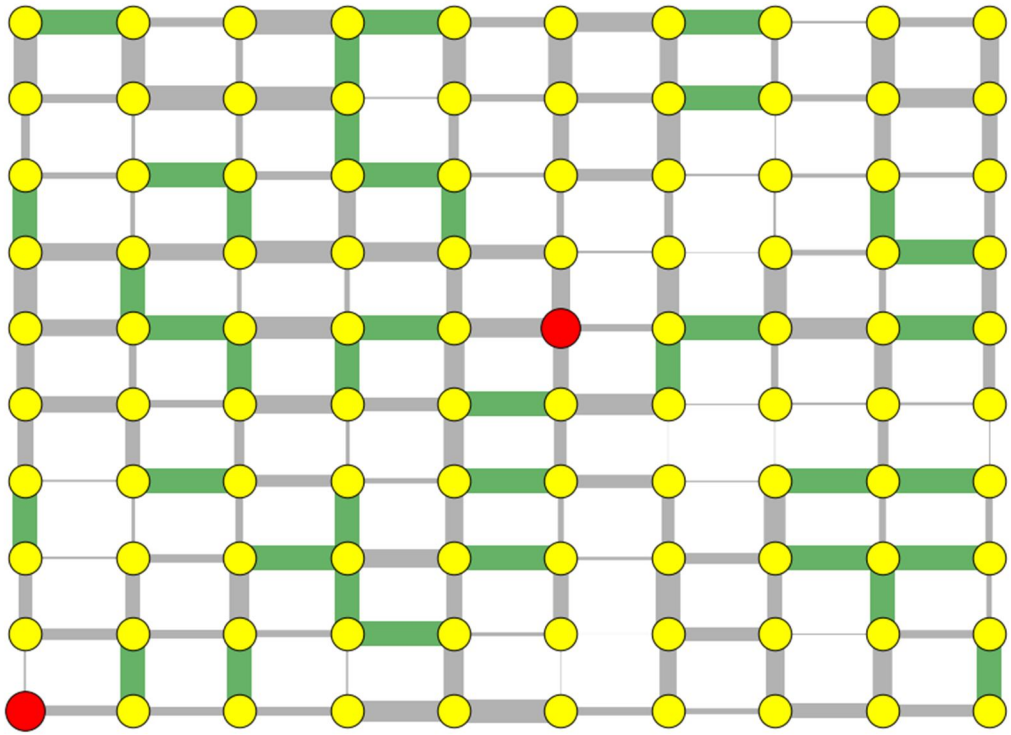
Обучение

- В отличие от табличного RL создаем «сеть» из *FS* элементов аппроксимирующих ценность действия $Q(o,a)$ и модель $m(b,o,a)$
  - если для данного наблюдения не существует *FS* элемента, то выбираем действие случайно
  - если случайное действие привело в состояние для которого существует хотя бы один *FS* элемент, то добавляем в сеть новый *FS* элемент. «запоминающий» действие $Q(o,a)$ и предсказание $m(b,o,a)$
  - после достижения цели обновляем ценности *FS* элементов способом похожим на Q-learning

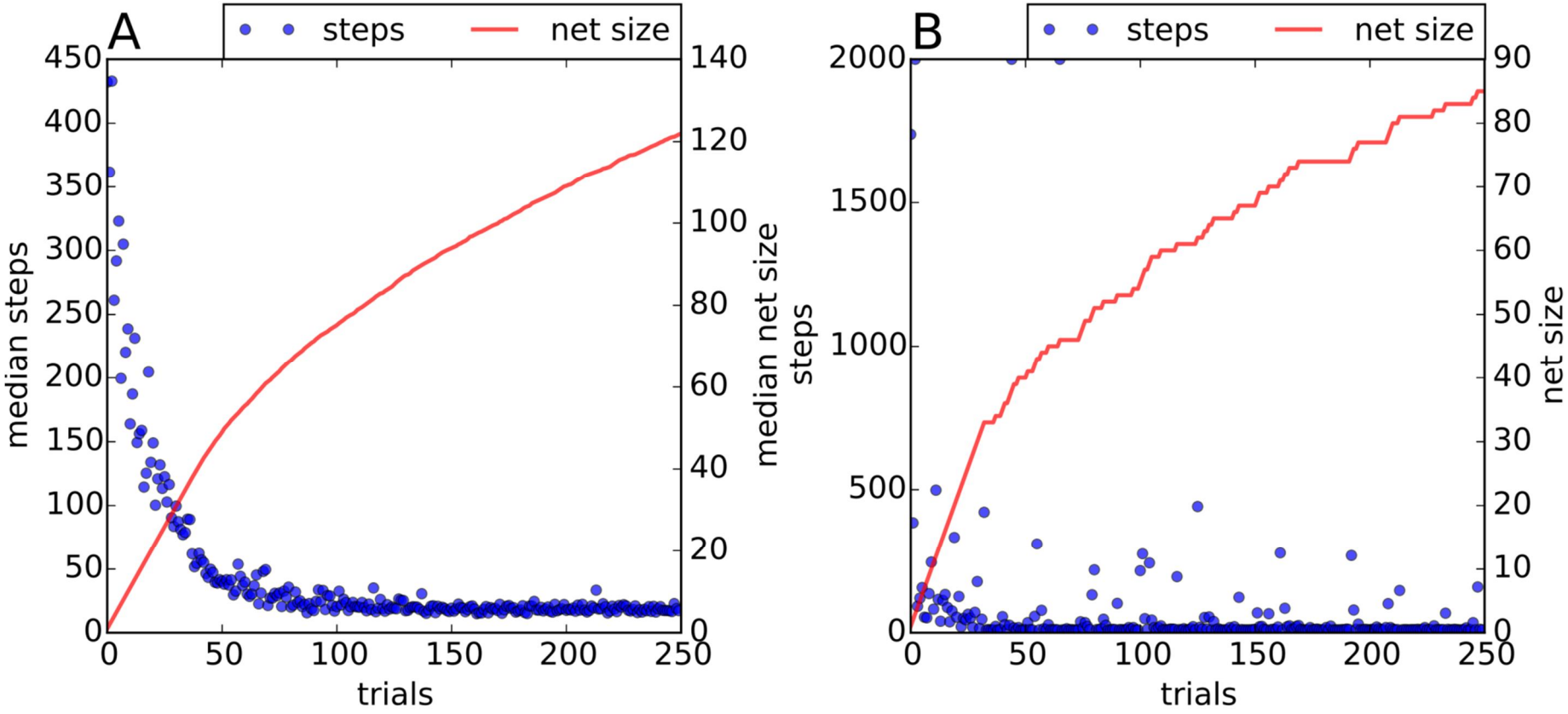
$$Q(o,a) \leftarrow Q(o,a) + \alpha(r + max_{a \in \mathcal{A}}(b' \cdot Q(o',a)) - Q(o,a))$$

Модель среды

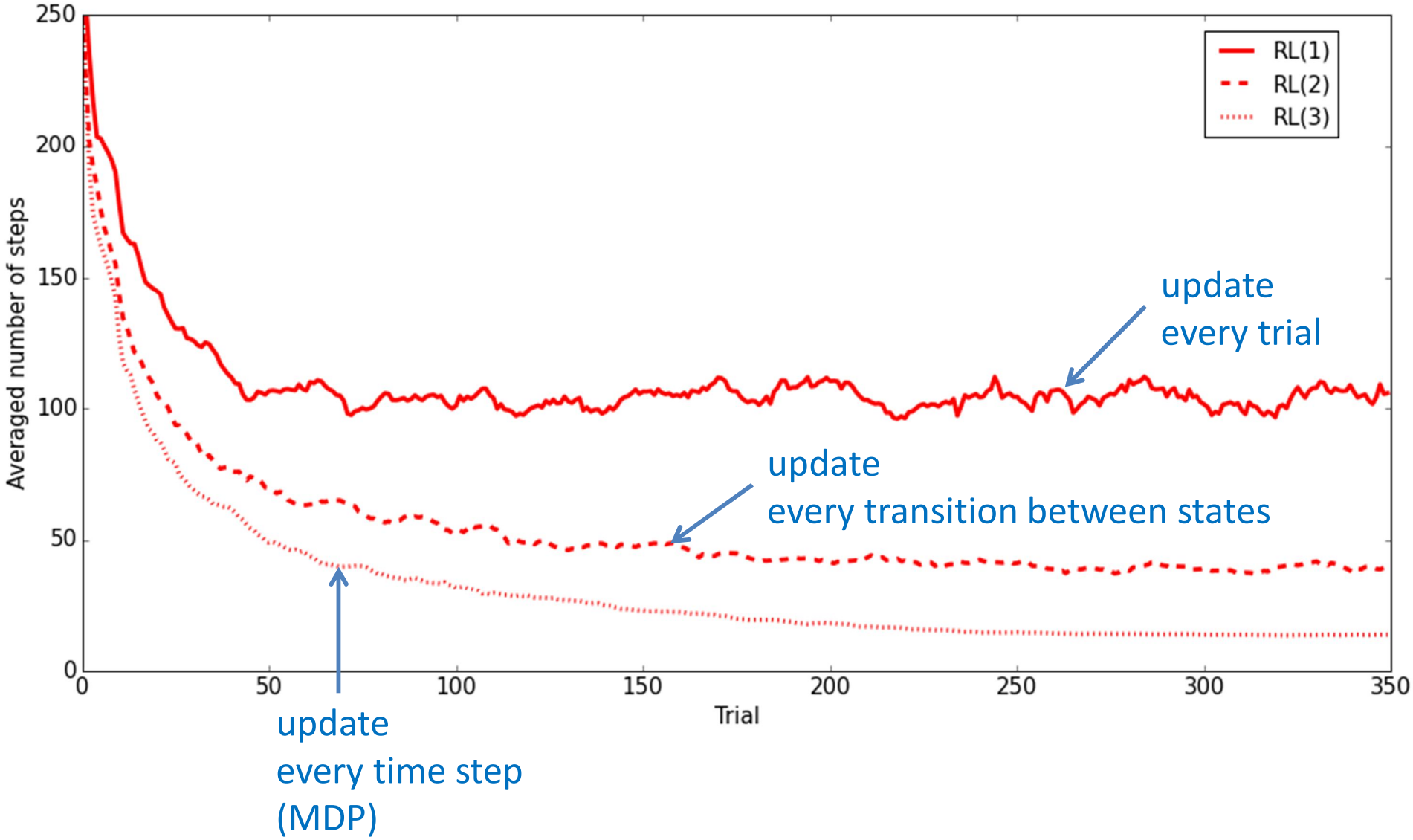
- 2D решетка
  - агент обучается в серии попыток проходить из стартового состояния в целевое
  - 25% переходов имеют вес равный 1
  - оставшимся 75% переходов присваивается вес из равномерного распределения на (0,1)
  - веса задают вероятность сэмплирования перехода:
    1. **в каждой попытке (trial upd)**
    2. при смене состояния (state upd)
    3. после каждого действия (MDP upd)



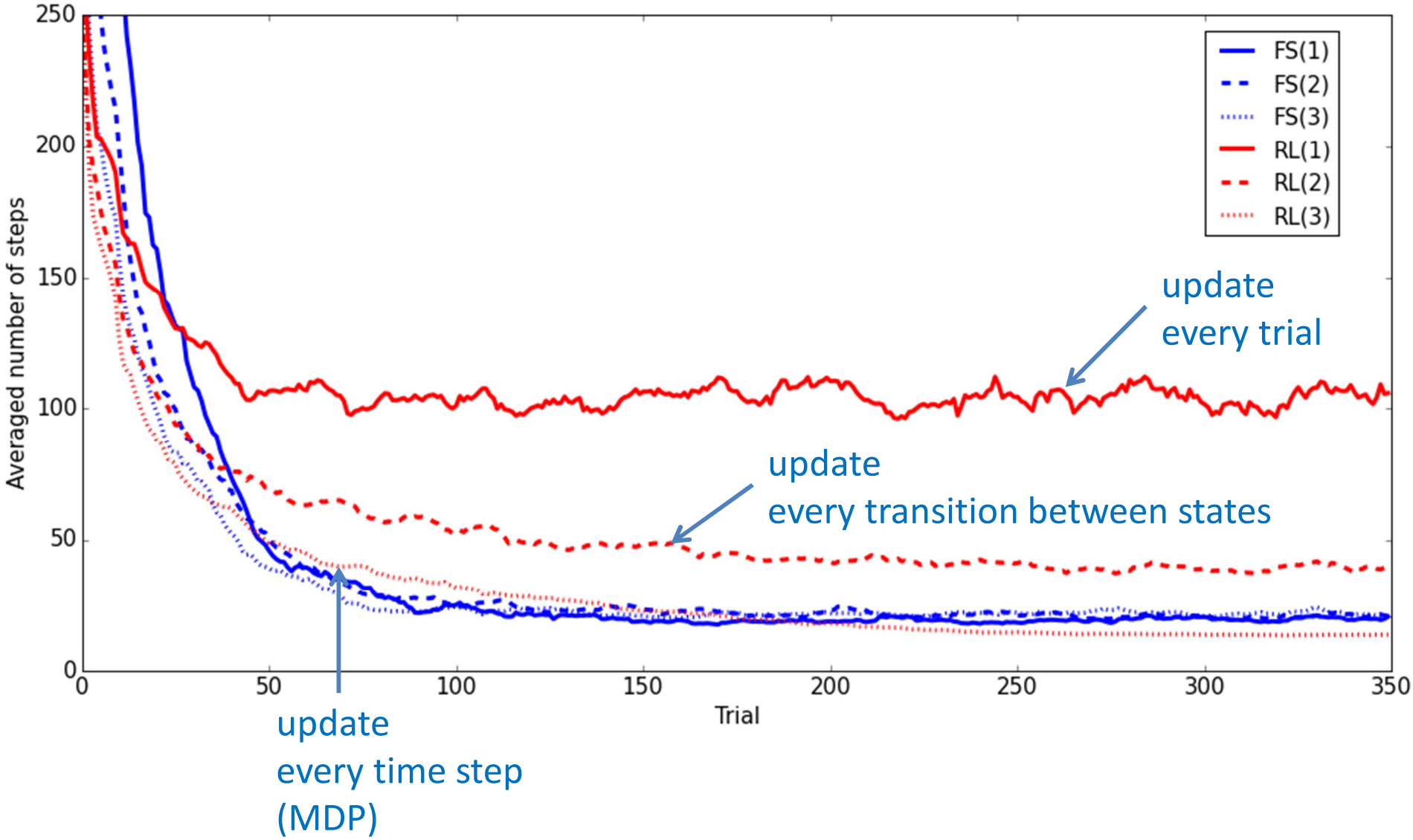
Решетка 10 на 10



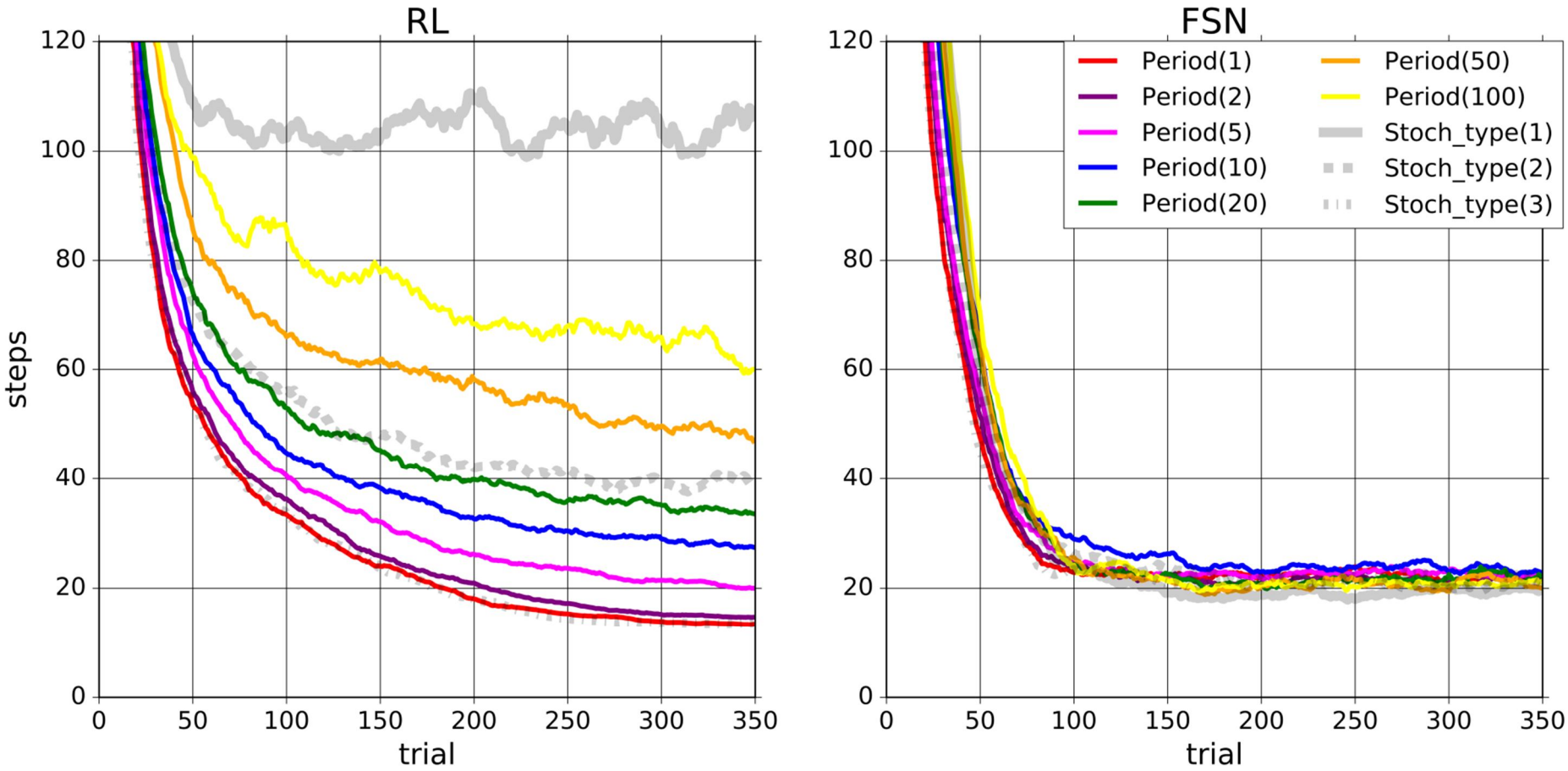
Решетка 10 на 10



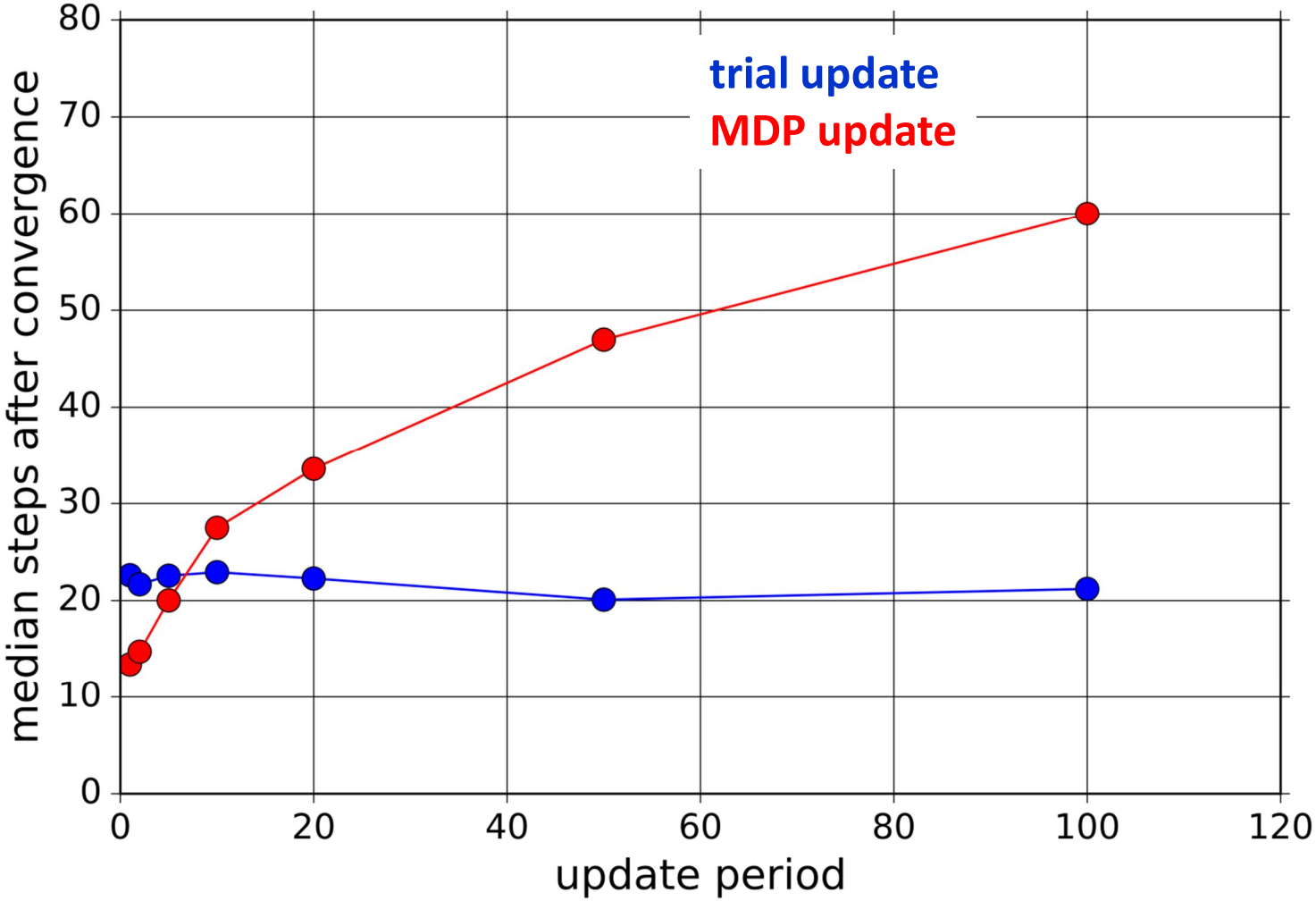
Решетка 10 на 10



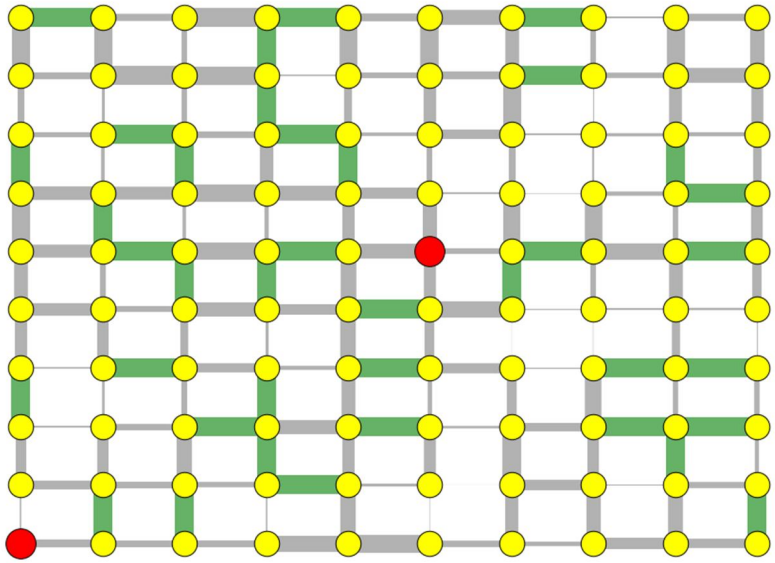
Решетка 10 на 10

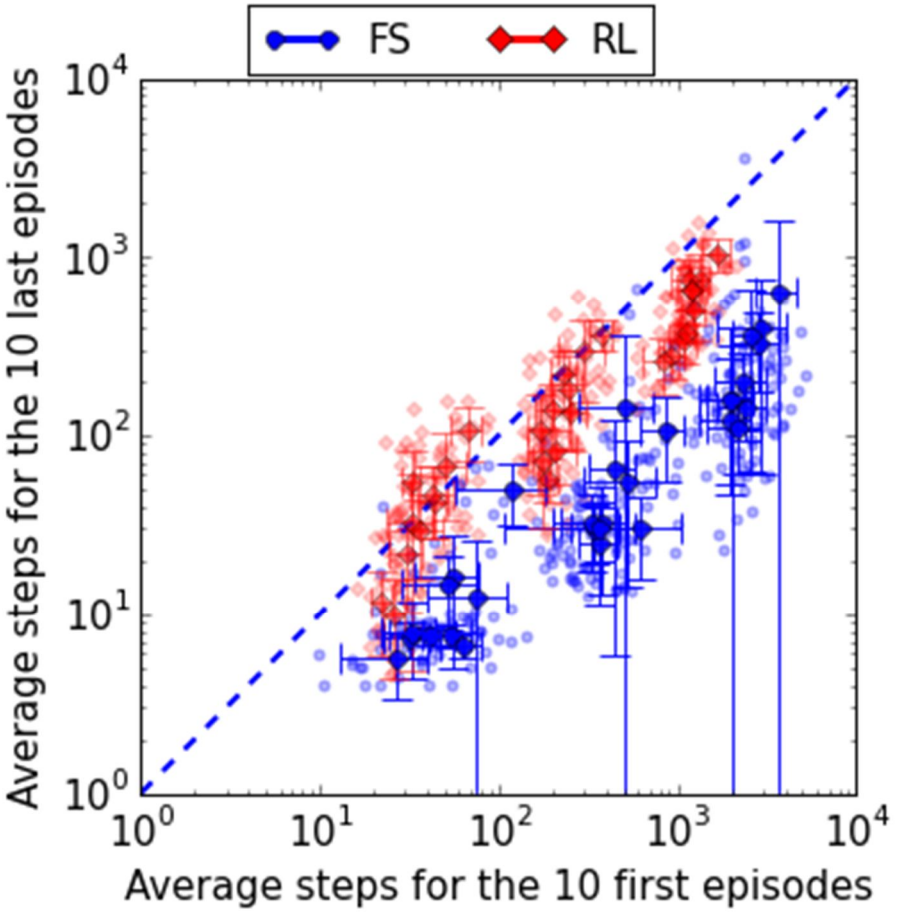


FSNet не зависит от периода обновления «мира»

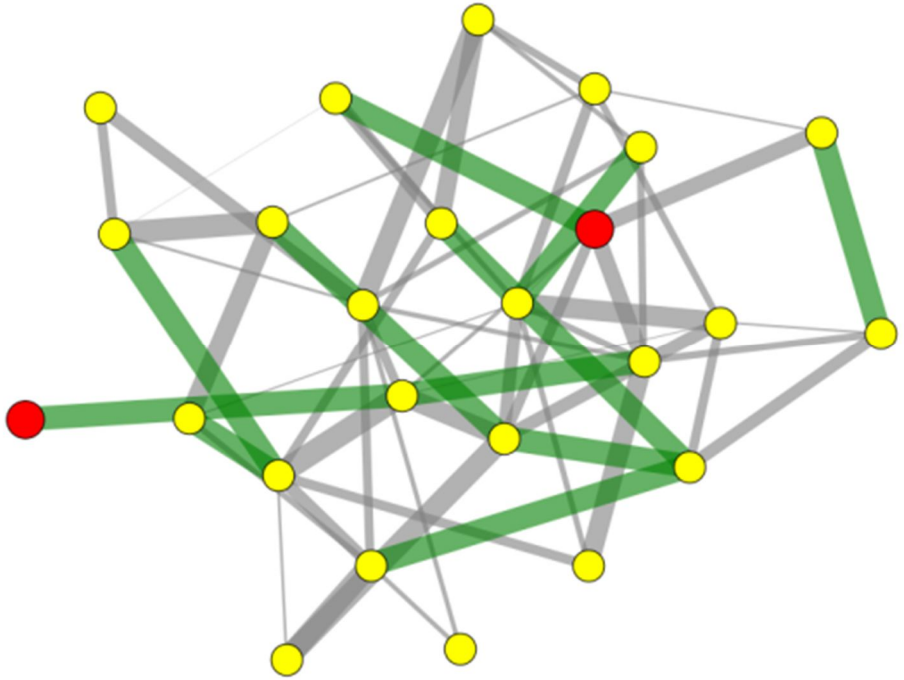


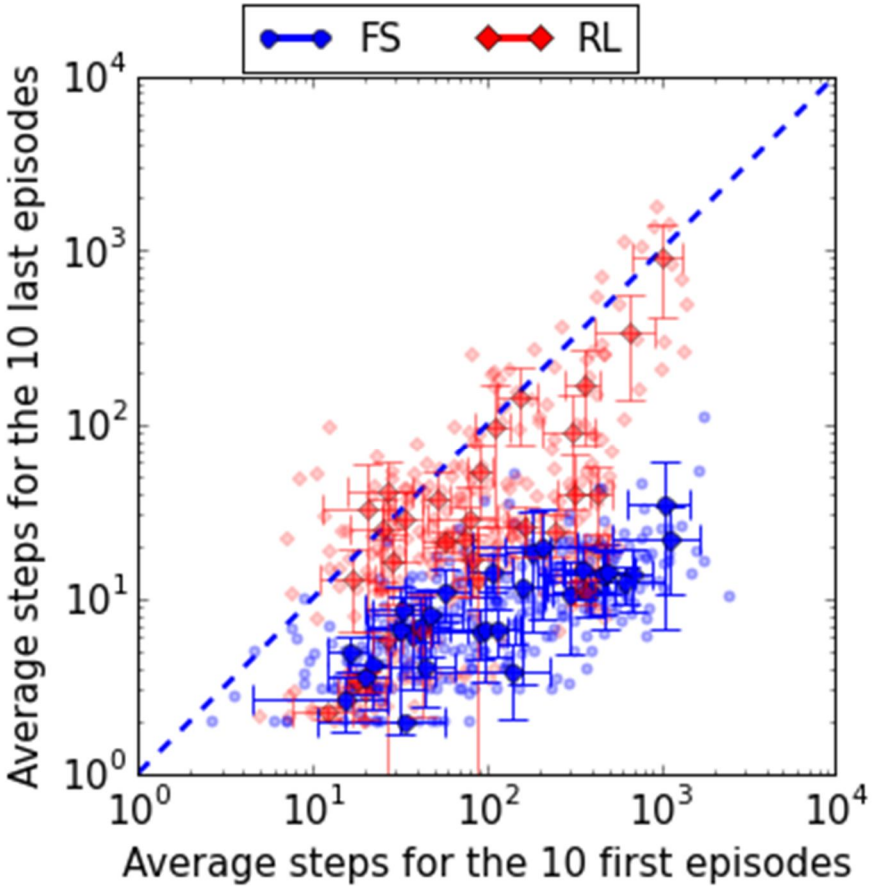
Решетка: 25, 100 и 400 состояний



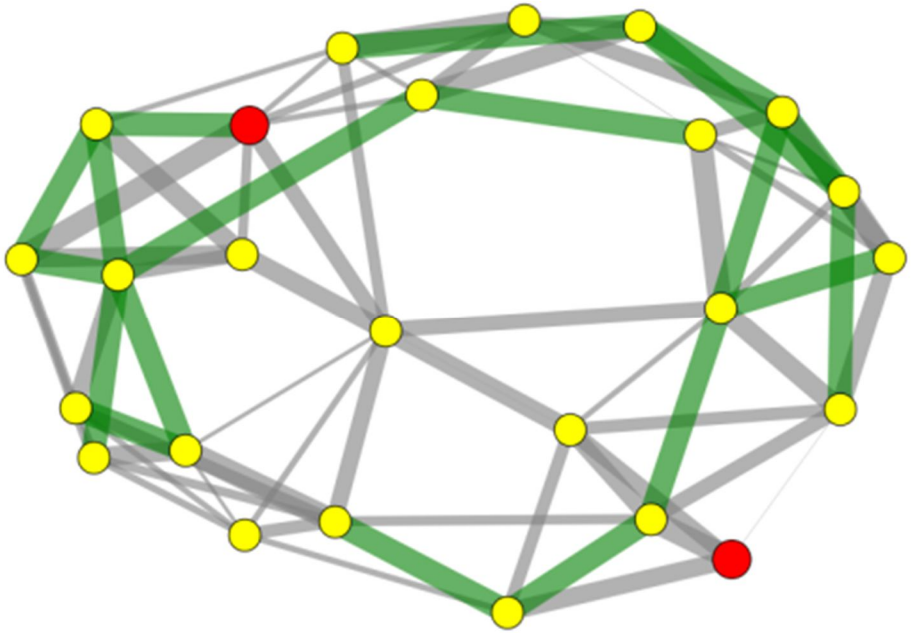


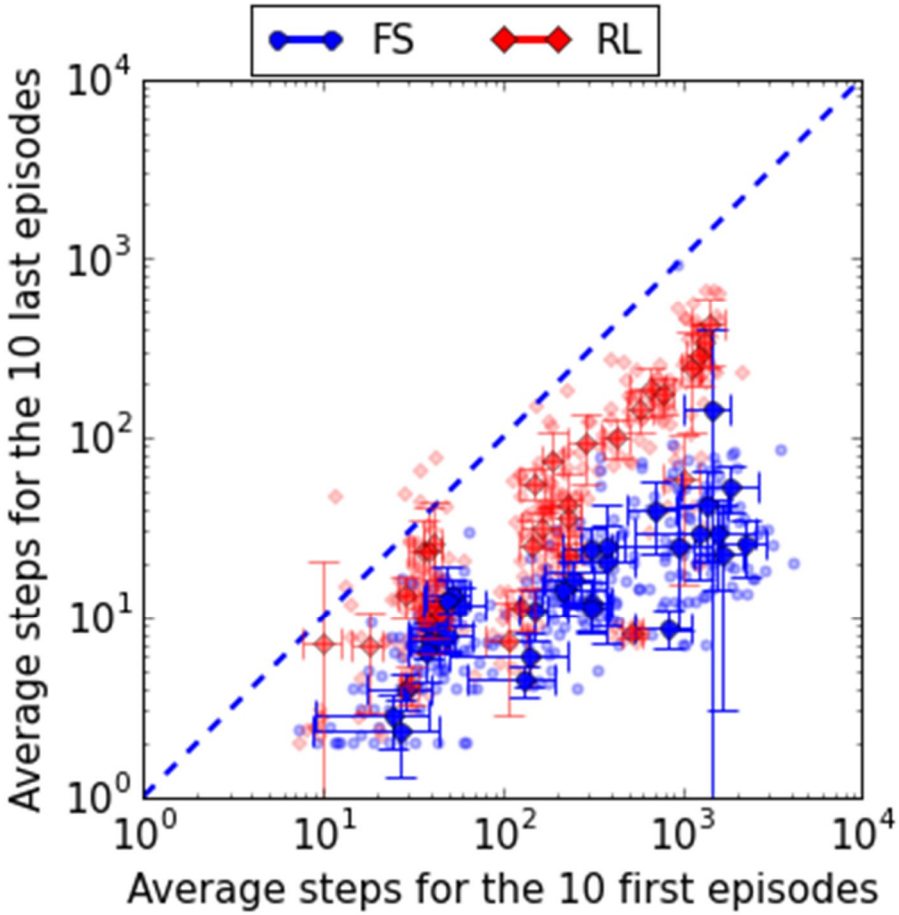
Случайный граф (Erdős–Rényi)





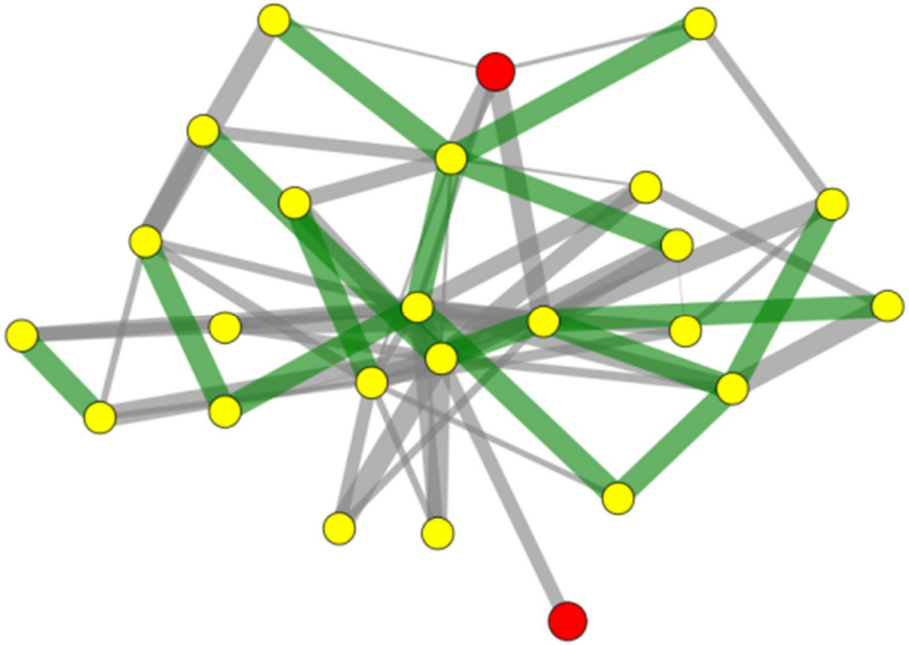
Сеть малого мира (Watts–Strogatz)

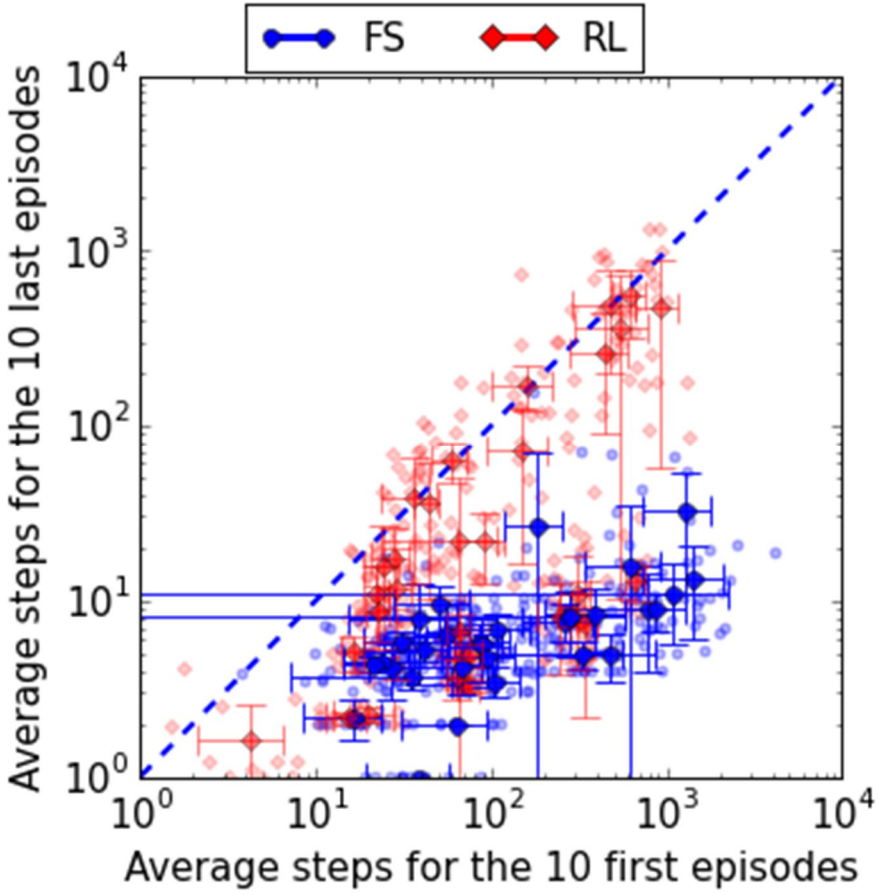




соцсети, трофические цепи, электросети, метаболические сети, структура мозга

Безмасштабная сеть (Barabási–Albert)

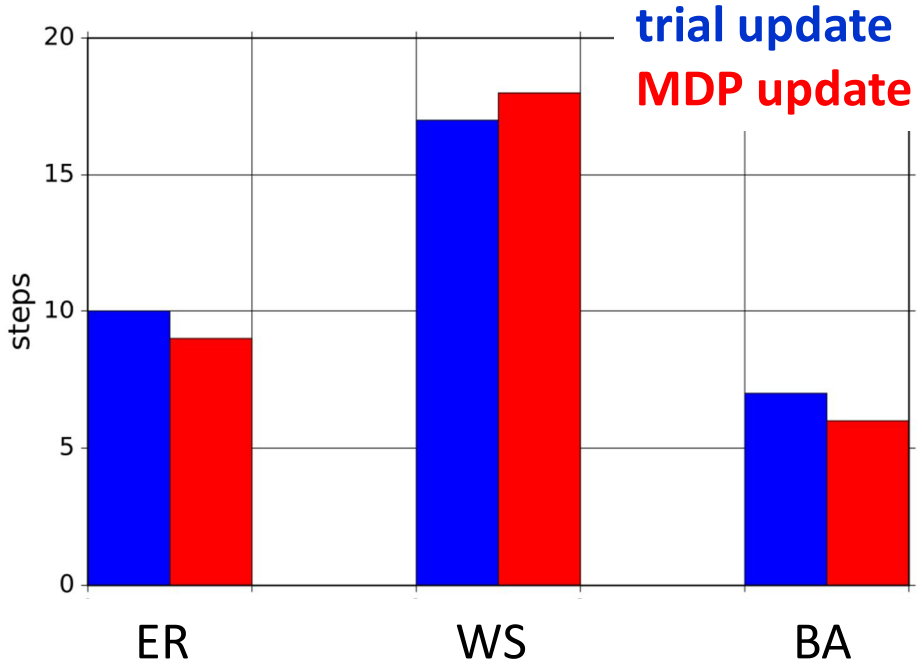




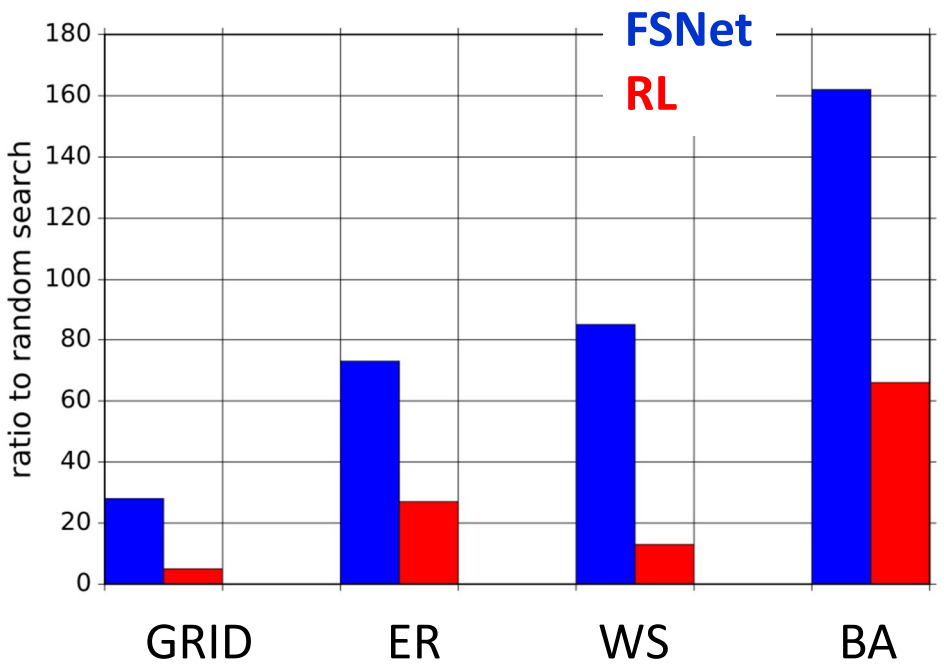
интернет, WWW

Все топологии 400 состояний

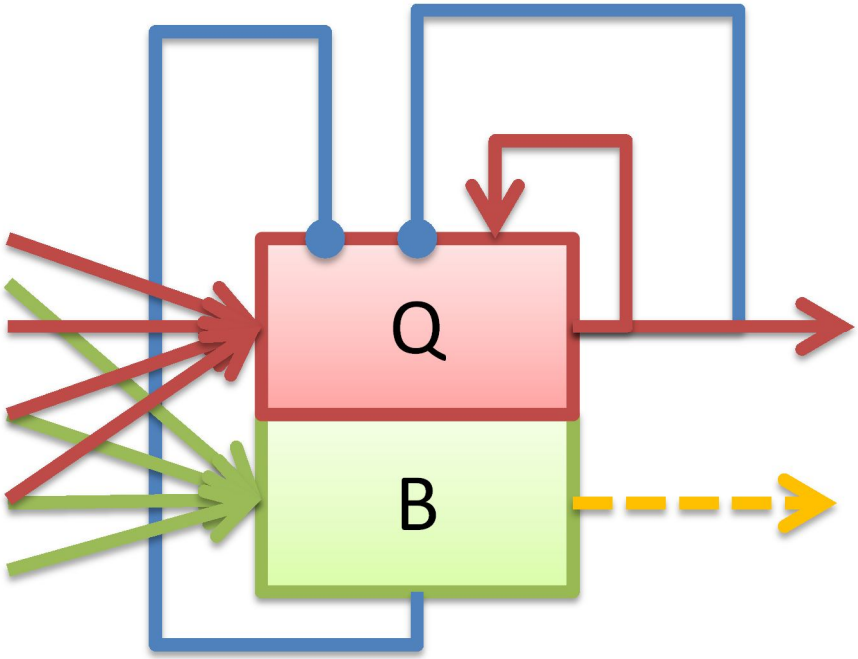
FSNet не зависит от периода изменения среды во всех топологиях

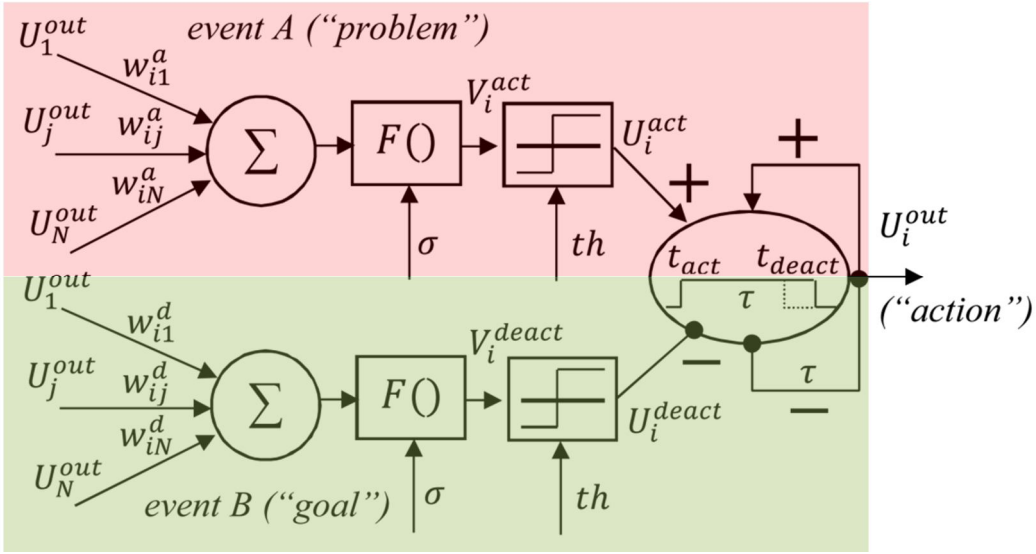


На безмасштабной топологии достигается максимальная разница со случайным блужданием

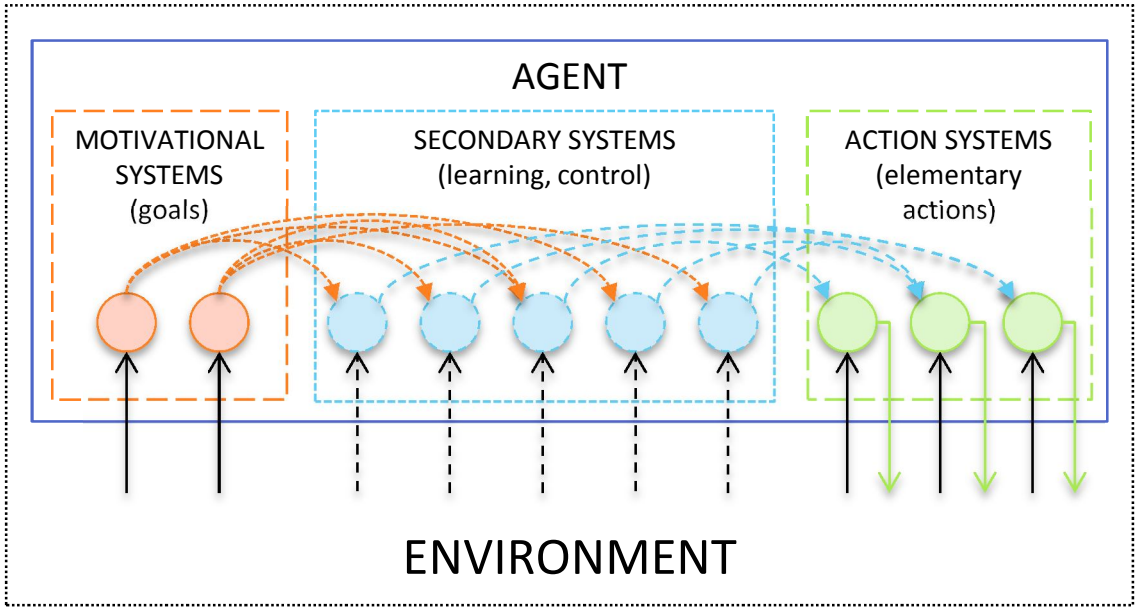


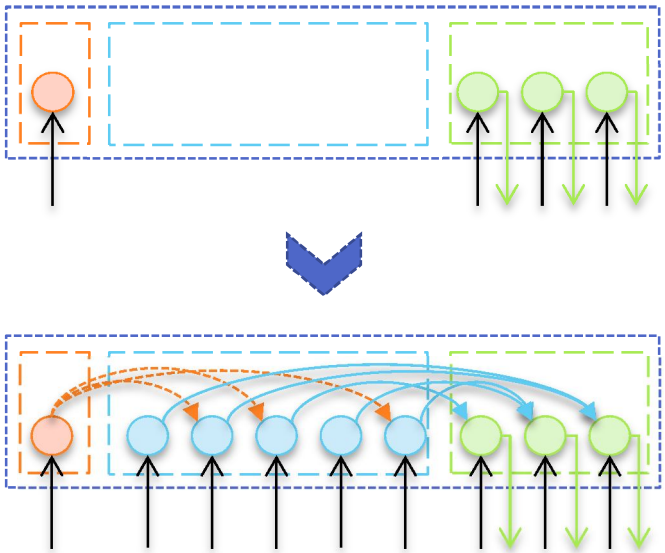
Нейросетевая реализация





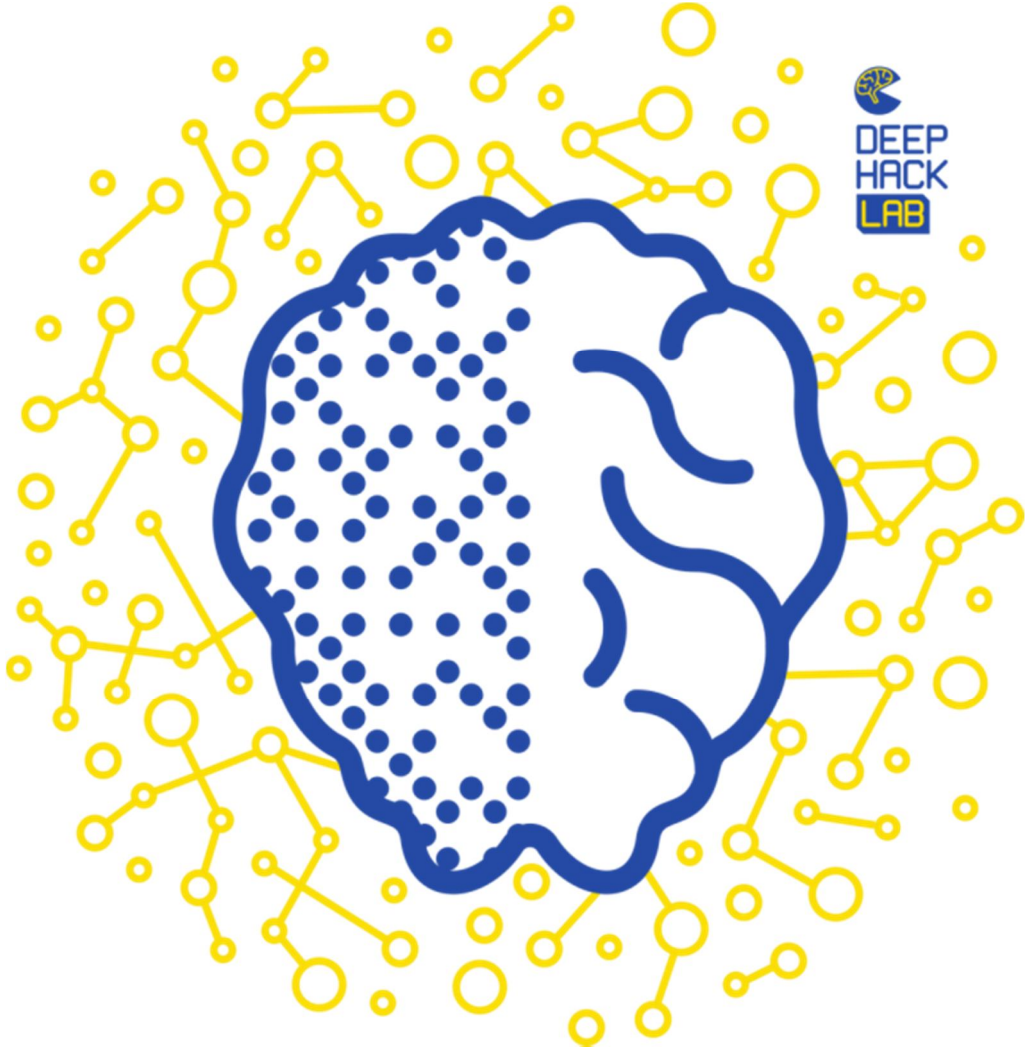
Нейросетевая реализация



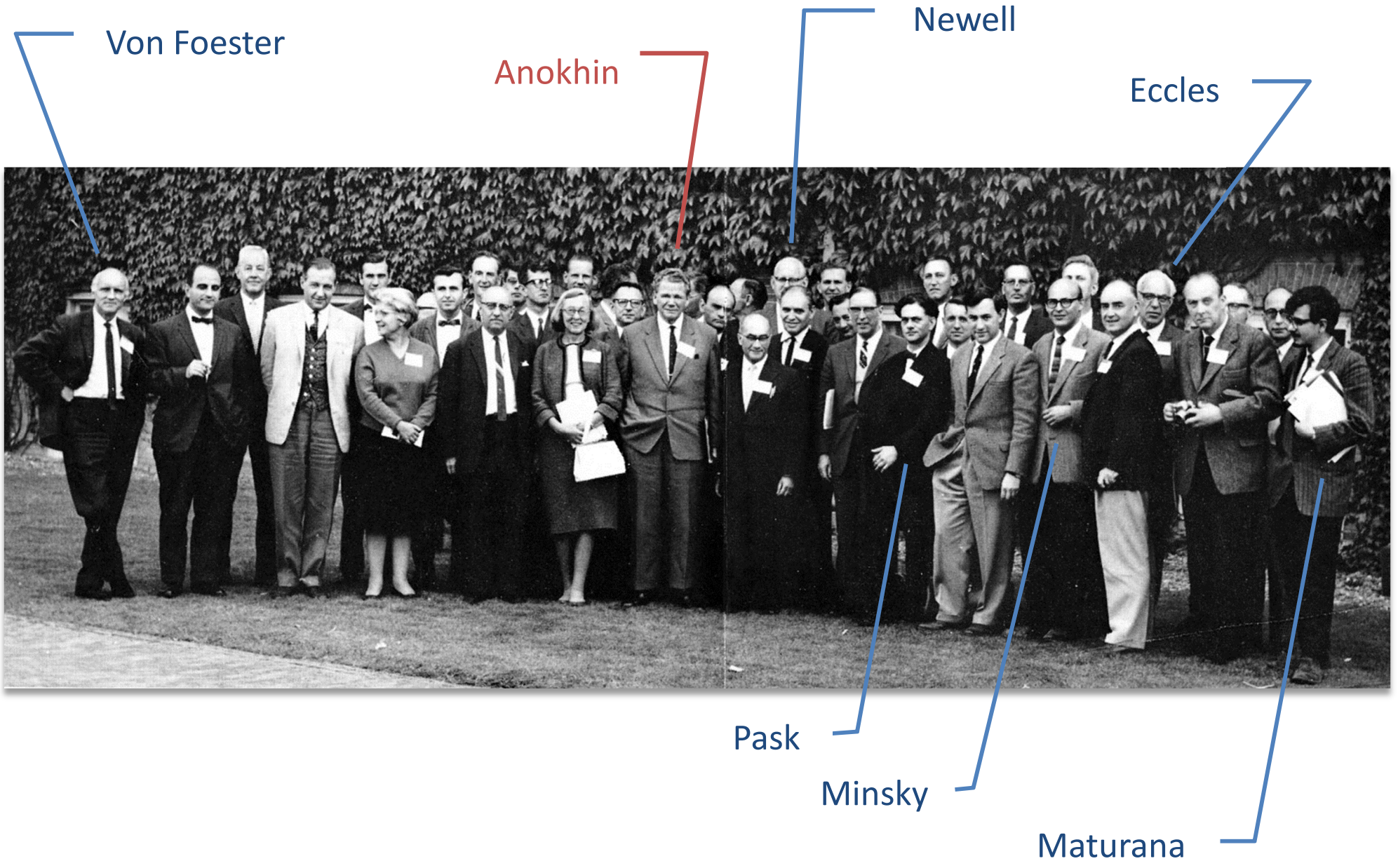


Выводы

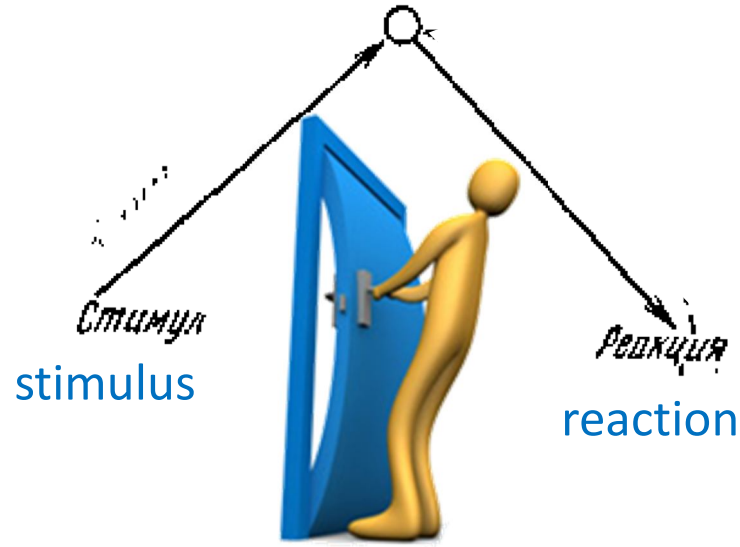
- Введение механизма быстрого переключения между альтернативными действиями, позволяет решать задачу обучения целенаправленному поведению в частично наблюдаемых средах.


DEEP
HACK
LAB

- Neural Information Processing Systems (NIPS), 1987 -
- Information processing in nervous system, Leiden, 1962



Основная идея теории функциональных систем



stimulus

reaction

Open loop

vs

Closed loop

Функциональная система