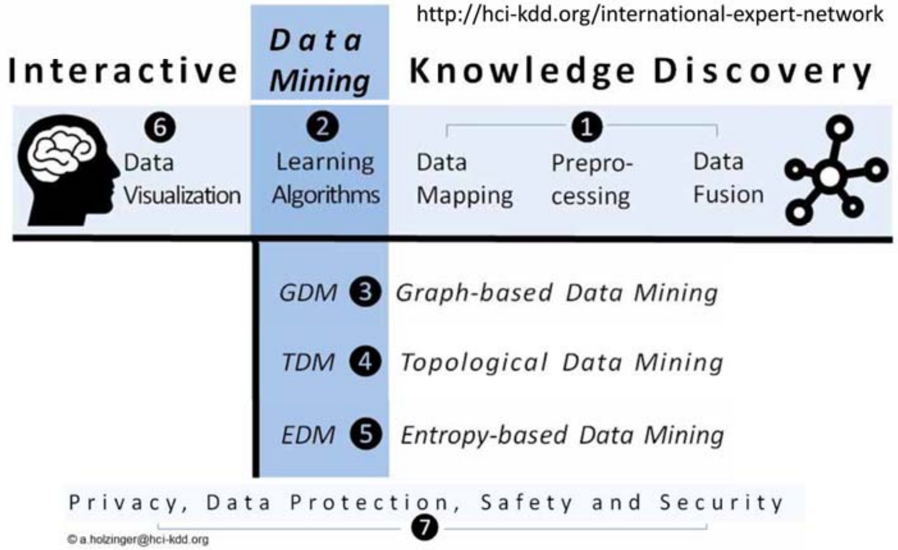
Andreas Holzinger
340.300 Principles of Interaction
Summer Term 2017

Selected Topics of interactive Machine Learning (iML): Interaction with Agents Part 3: Reinforcement Learning

a.holzinger@hci-kdd.org
http://hci-kdd.org/interactive-machine-learning

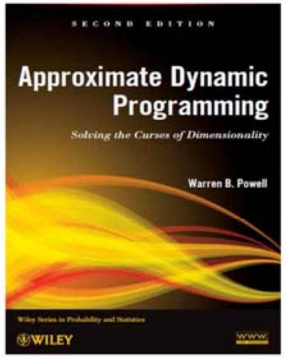
iML needs a concerted effort fostering integrated research

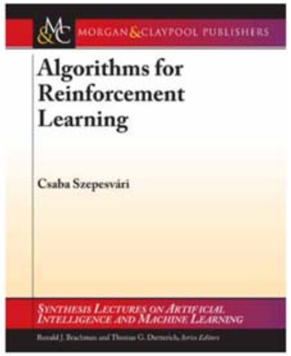
http://hci-kdd.org/international-expert-network



Holzinger, A. 2014. Trends in Interactive Knowledge Discovery for Personalized Medicine: **Cognitive Science meets Machine Learning.** IEEE Intelligent Informatics Bulletin, 15, (1), 6-14.

Standard Textbooks for RL





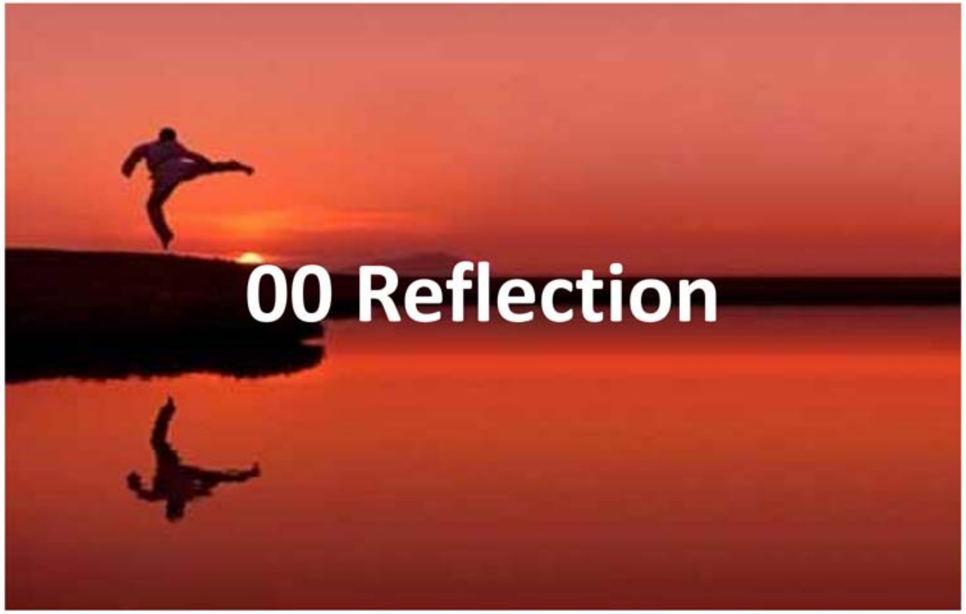
Sutton, R. S. & Barto, A. G. 1998. *Reinforcement learning: An introduction*, Cambridge, MIT press, http://incompleteideas.net/sutton/book/the-book-1st.html.

Powell, W. B. 2007. *Approximate Dynamic Programming: Solving the curses of dimensionality*, John Wiley & Sons, http://adp.princeton.edu/.

Szepesvári, C. 2010. Algorithms for reinforcement learning. Synthesis lectures on artificial intelligence and machine learning, 4, (1), 1-103.

Red thread through this lecture

- 00 Reflection
- 01 What is RL? Why is it interesting?
- 02 Decision Making under uncertainty
- 03 Roots of RL
- 04 Cognitive Science of RL
- 05 The Anatomy of an RL agent
- 06 Example: Multi-Armed Bandits
- 07 RL-Applications in health
- 08 Future Outlook



Problem of a Human-in-the-loop ?

- Humans are irrational, inconsistent, lacking robustness, error-prone, adaptive, subjective, …
- Problem: Preferences often are biased, subjective, constructed on the fly, or even do not exist …
- (Daniel Kahnemann, Nobel-Prize 2002)

Kahneman, D. 2011. Thinking, fast and slow, New York, Macmillan.

Quiz (Supervised S, Unsupervised U, Reinforcement R)

1) Given x, y; find f that map a new $x \mapsto y$ (S/U/R?)
2) Finding similar points in high-dim X (S/U/R)?
3) Learning from interaction to achieve a goal (S/U/R)?
4) Human expert provides examples (S/U/R)?
5) Automatic learning by interaction with environment (S/U/R)?
6) An agent gets a scalar reward from the environment (S/U/R)?

01 What is RL? Why is it interesting?

"I want to understand intelligence and how minds work. My tools are computer science, statistics, mathematics, and plenty of thinking" Nando de Freitas, Univ. Oxford and Google."

In press at *Behavioral and Brain Sciences.*

Building Machines That Learn and Think Like People

Brenden M. Lake,[1] Tomer D. Ullman,[2,4] Joshua B. Tenenbaum,[2,4] and Samuel J. Gershman[3,4]
[1]Center for Data Science, New York University
[2]Department of Brain and Cognitive Sciences, MIT
[3]Department of Psychology and Center for Brain Science, Harvard University
[4]Center for Brains Minds and Machines

Abstract

Recent progress in artificial intelligence (AI) has renewed interest in building systems that learn and think like people. Many advances have come from using deep neural networks trained end-to-end in tasks such as object recognition, video games, and board games, achieving performance that equals or even beats humans in some respects. Despite their biological inspiration and performance achievements, these systems differ from human intelligence in crucial ways. We review progress in cognitive science suggesting that truly human-like learning and thinking machines will have to reach beyond current engineering trends in both what they learn, and how they learn it. Specifically, we argue that these machines should (a) build causal models of the world that support explanation and understanding, rather than merely solving pattern recognition problems; (b) ground learning in intuitive theories of physics and psychology, to support and enrich the knowledge that is learned; and (c) harness compositionality and learning-to-learn to rapidly acquire and generalize knowledge to new tasks and situations. We suggest concrete challenges and promising routes towards these goals that can combine the strengths of recent neural network advances with more structured cognitive models.

JKU Remember: Three main types of Machine Learning

1-S; 2-U; 3-R; 4-S; 5-R; 6-R

- **I) Supervised learning (classification)**
  - $y = f(x)$
  - Given x, y pairs; find a f that map a new x to a proper y
  - Regression, logistic regression, classification
  - Expert provides examples e.g. classification of clinical images
  - Disadvantage: Supervision can be expensive
- **II) Unsupervised learning (clustering)**
  - $f(x)$
  - Given x (features only), find f that gives you a description of x
  - Find similar points in high-dim X
  - E.g. clustering of medical images based on their content
  - Disadvantage: Not necessarily task relevant
- **III) Reinforcement learning**
  - $y = f(x)$
  - more general than supervised/unsupervised learning
  - learn from interaction to achieve a goal
  - Learning by direct interaction with environment (automatic ML)
  - Disadvantage: broad difficult approach, problem with high-dim data

JKU Why is RL interesting?

- Reinforcement Learning is the **oldest approach,** with the longest history and can provide insight into understanding human learning [1]
- RL is the **"AI problem in the microcosm"** [2]
- Future opportunities are in Multi-Agent RL (MARL), Multi-Task Learning (MTL), Generalization and **Transfer-Learning** [3], [4].

[1] Turing, A. M. 1950. Computing machinery and intelligence. Mind, 59, (236), 433-460.

[2] Littman, M. L. 2015. Reinforcement learning improves behaviour from evaluative feedback. Nature, 521, (7553), 445-451, doi:10.1038/nature14540.

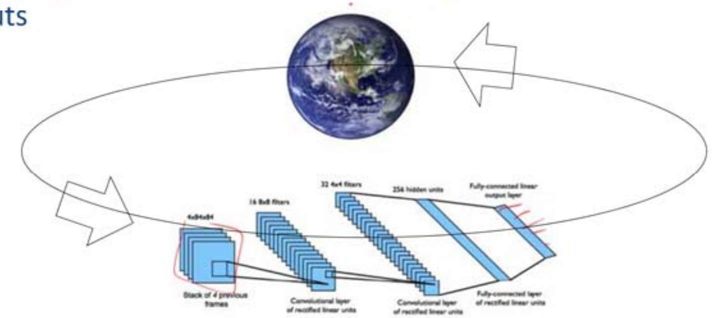
[3] Taylor, M. E. & Stone, P. 2009. Transfer learning for reinforcement learning domains: A survey. The Journal of Machine Learning Research, 10, 1633-1685.
[4] Pan, S. J. & Yang, Q. A. 2010. A Survey on Transfer Learning. IEEE Transactions on Knowledge and Data Engineering, 22, (10), 1345-1359, doi:10.1109/tkde.2009.191.

JKU RL is key for ML according to Demis Hassabis

https://www.youtube.com/watch?v=XAbLn66iHcQ&index=14&list=PL2ovtN0KdWZiomydY2yWhh9-QOn0GvrCR
Go to time 1:33:00

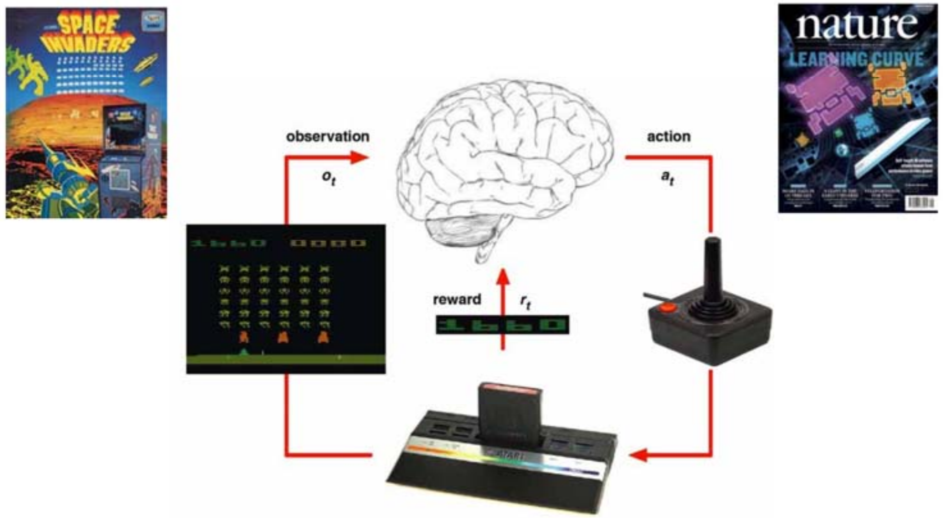
A very recent approach is combining RL with DL

- Combination of deep neural networks with reinforcement learning = Deep Reinforcement Learning
- Weakness of classical RL is that it is not good with high-dimensional sensory inputs
- Advantage of DRL: Learn to act from high-dimensional sensory inputs



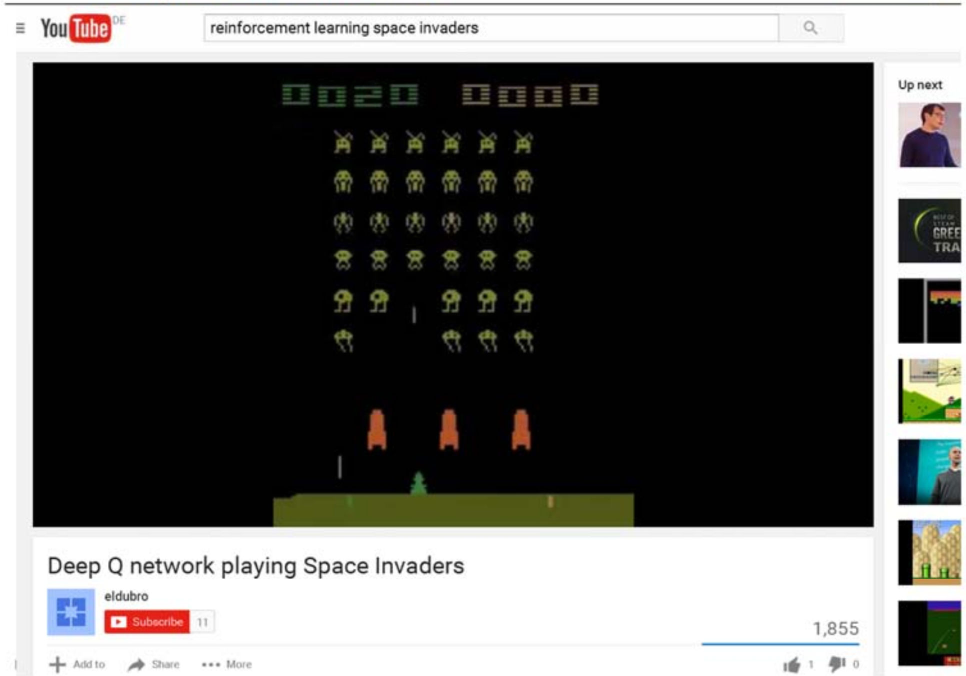
Volodymyr Mnih et al (2015), https://sites.google.com/a/deepmind.com/dqn/
https://www.youtube.com/watch?v=iqXKQf2BOSE

Learning to play an Atari Game

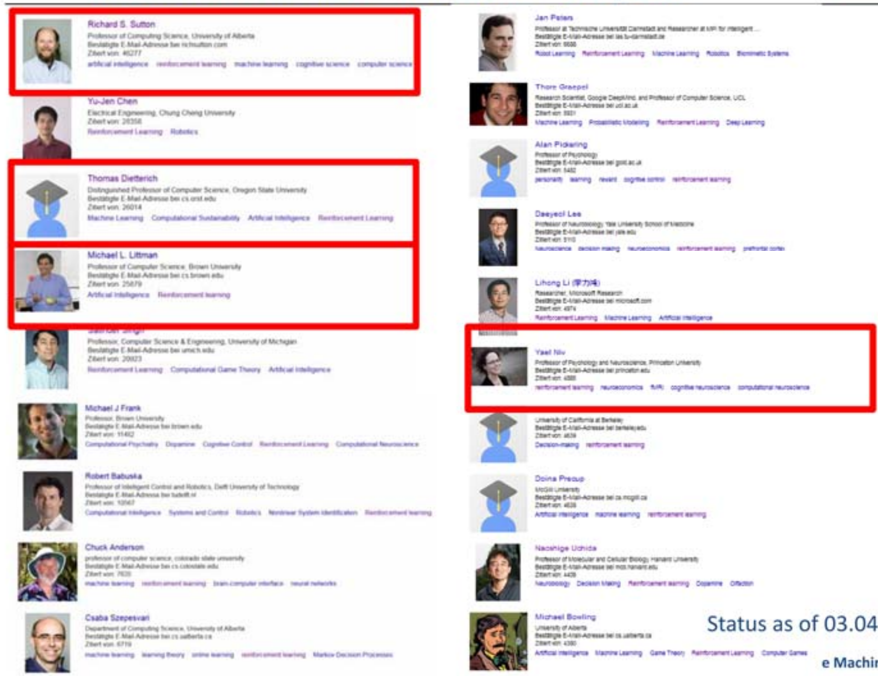


Mnih, V., Kavukcuoglu, K., Silver, D., Rusu, A. A., Veness, J., Bellemare, M. G., Graves, A., Riedmiller, M., Fidjeland, A. K., Ostrovski, G., Petersen, S., Beattie, C., Sadik, A., Antonoglou, I., King, H., Kumaran, D., Wierstra, D., Legg, S. & Hassabis, D. 2015. Human-level control through deep reinforcement learning. Nature, 518, (7540), 529-533, doi:10.1038/nature14236

Example Video Atari Game



Scientists in this area - selection - incomplete!



Status as of 03.04.2016

JKU Always keep in mind: Learning and Inference

HCI-KDD

04

d … $data$

h … $hypotheses$

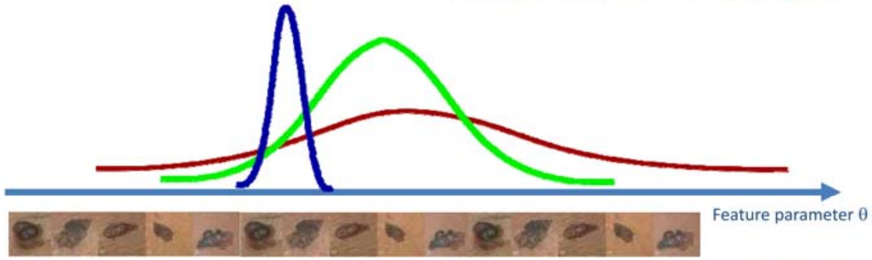
$\mathcal{H}$ … $\{H_1, H_2, \dots, H_n\}$ $\forall\, h, d$ …

Likelihood Prior Probability

$$p(h|d) = \frac{p(d|h) * p(h)}{\sum_{h \in H} p(d|h')\, p(h')}$$

Posterior Probability

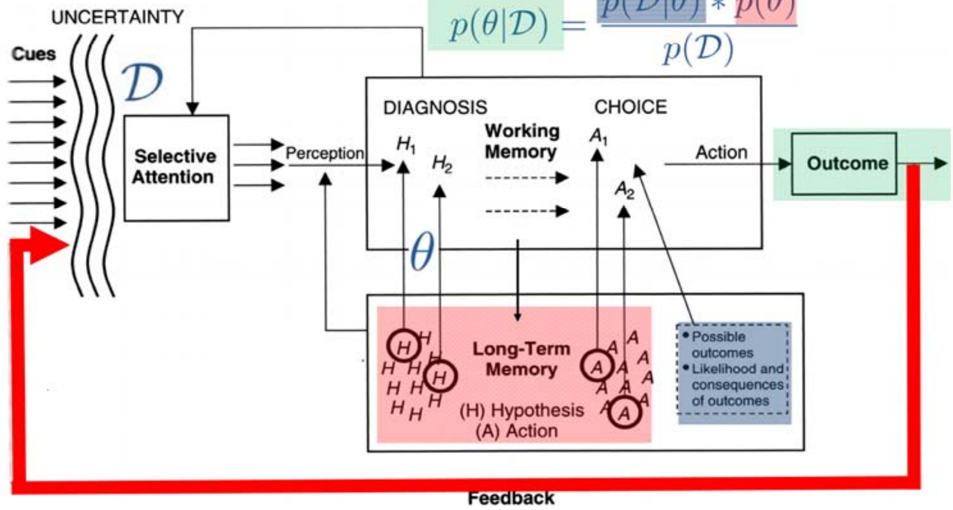
Problem in $\mathbb{R}^n \rightarrow$ complex



JKU Human Decision Making: probabilistic reasoning

HCI-KDD

$$p(\theta|\mathcal{D}) = \frac{p(\mathcal{D}|\theta) * p(\theta)}{p(\mathcal{D})}$$



Wickens, C. D. (1984) *Engineering psychology and human performance. Columbus (OH), Charles Merrill, Altered by Holzinger, A. (2017)*

JKU Bayesian Learning from data → Generalize

HCI-KDD

09

$$\mathcal{D} = x_{1:n} = \{x_1, x_2, \dots, x_n\}$$

$p(\mathcal{D}|\theta)$

$$p(\theta|\mathcal{D}) = \frac{p(\mathcal{D}|\theta) * p(\theta)}{p(\mathcal{D})}$$

$$posterior = \frac{likelihood * prior}{evidence}$$

The inverse probability allows to learn from data, infer unknowns, and make predictions

JKU Goal: Select actions to maximize total future reward

HCI-KDD

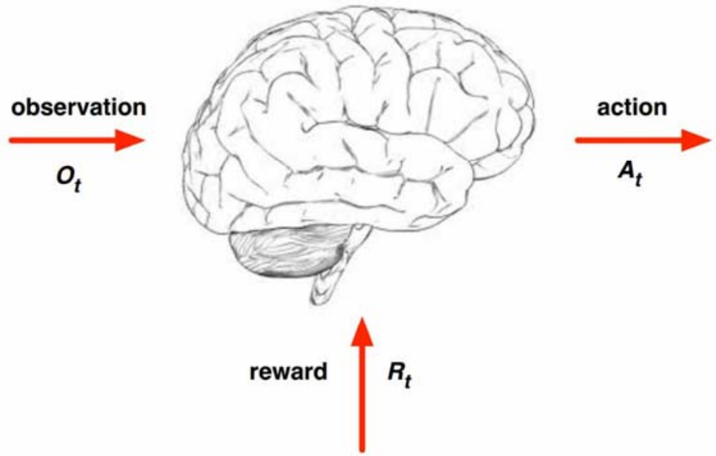
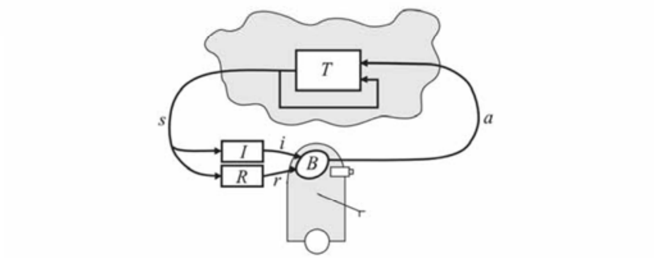


Image credit to David Silver, UCL

Standard RL-Agent Model goes back to Cybernetics 1950



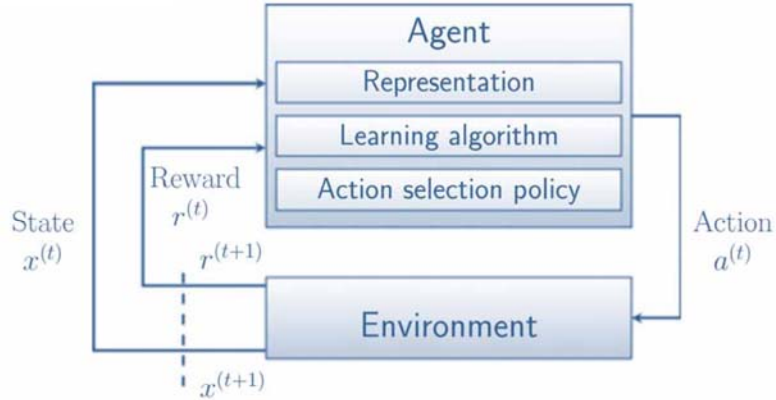

```
initialize V(s) arbitrarily
loop until policy good enough
    loop for s ∈ S
        loop for a ∈ A
            Q(s, a) := R(s, a) + γ Σ_{s'∈S} T(s, a, s') V(s')
        V(s) := max_a Q(s, a)
    end loop
end loop
```

Kaelbling, L. P., Littman, M. L. & Moore, A. W. 1996. Reinforcement learning: A survey. Journal of Artificial Intelligence Research, 4, 237-285.

RL-Agent seeks to maximize rewards

for $t = 1, \ldots, n$ **do**
The agent perceives state s_t
The agent performs action a_t
The environment evolves to s_{t+1}
The agent receives reward r_t
end for

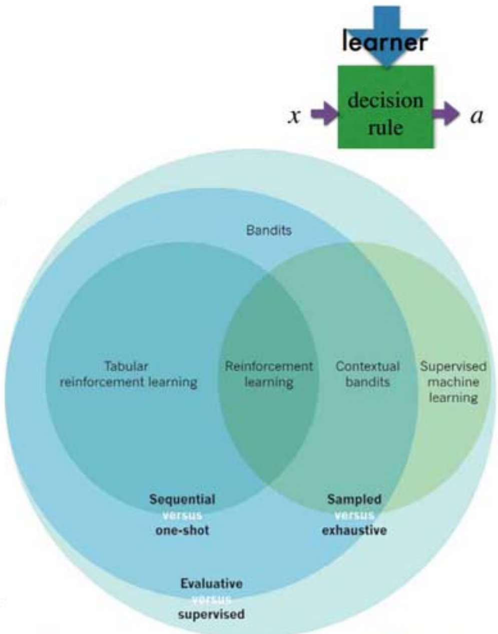
Intelligent behavior arises from the actions of an individual seeking to **maximize its received reward** signals in a **complex and changing world**



Sutton, R. S. & Barto, A. G. 1998. Reinforcement learning: An introduction, Cambridge MIT press

RL – Types of Feedback (crucial!)

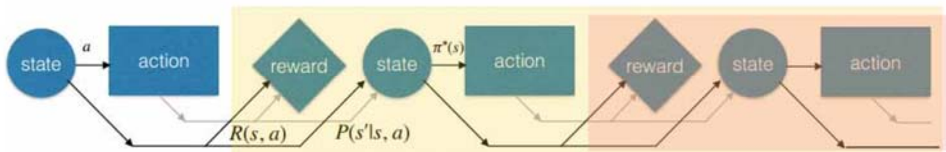
- Supervised: Learner told best a
- Exhaustive: Learner shown every possible x
- One-shot: Current x independent of past a



Littman, M. L. 2015. Reinforcement learning improves behaviour from evaluative feedback. Nature, 521, (7553), 445-451.

Problem Formulation in a MDP

- Markov decision processes specify setting and tasks
- Planning methods use knowledge of P and R to compute a good policy π
- Markov decision process model captures both sequential feedback and the more specific one-shot feedback (when $P(s'|s, a)$ is independent of both s and a



$$Q^*(s, a) = R(s, a) + \gamma \Sigma P(s'|s, a) \max_{a'} Q^*(s', a')$$

Littman, M. L. 2015. Reinforcement learning improves behaviour from evaluative feedback. Nature, 521, (7553), 445-451.

Agent observes environmental state at each step t

- 1) Overserves
- 2) Executes
- 3) Receives Reward
- Executes action A_t:
- $O_t = sa_t = se_t$
- Agent state = environment state = information state
- Markov decision process (MDP)

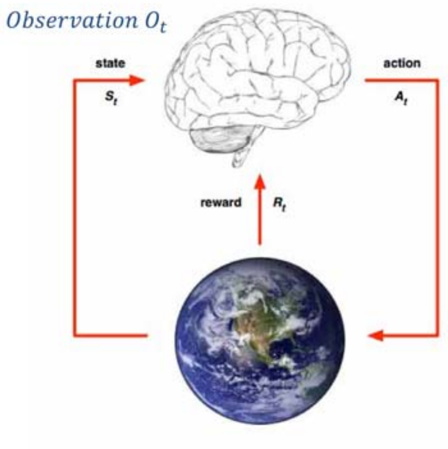
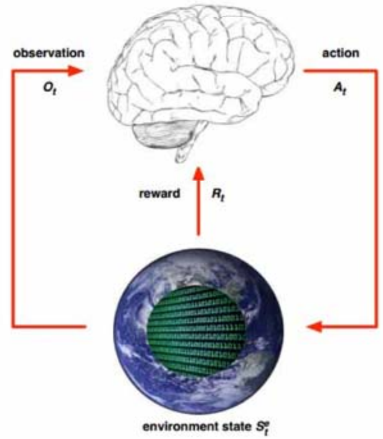


Image credit to David Silver, UCL

Environmental State is the current representation

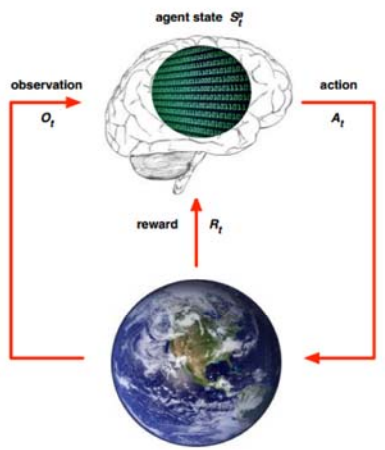
- i.e. whatever data the environment uses to pick the next observation/reward
- The environment state is not usually visible to the agent
- Even if S is visible, it may contain irrelevant information
- A State S_t is Markov iff:

$$\mathbb{P}[S_{t+1}|S_t] = \mathbb{P}[S_{t+1}|S_1, \ldots, S_t]$$



Agent State is the agents internal representation

- i.e. whatever information the agent uses to pick the next action
- it is the information used by reinforcement learning algorithms
- It can be any function of history:
- $S = f(H)$



$$H_t = O_1, R_1, A_1, \ldots, A_{t-1}, O_t, R_t$$

Components of RL Agents and Policy of Agents

- RL agent components:
  - Policy: agent's behaviour function
  - Value function: how good is each state and/or action
  - Model: agent's representation of the environment
- Policy as the agent's behaviour
  - is a map from state to action, e.g.
  - Deterministic policy: a = (s)
  - Stochastic policy: (ajs) = P[At = ajS t = s
- Value function is prediction of future reward:

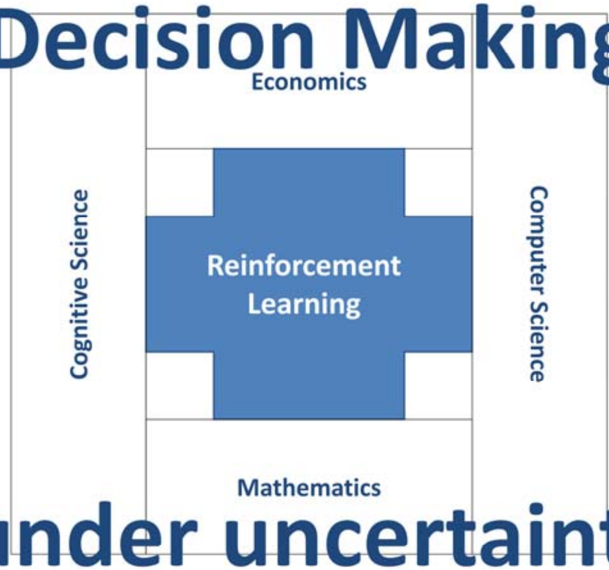
$$v_\pi(s) = \mathbb{E}_\pi\left[R_{t+1} + \gamma R_{t+2} + \gamma^2 R_{t+3} + \ldots \mid S_t = s\right]$$

What if the environment is only partially observable?

- Partial observability: when agent only indirectly observes environment (robot which is not aware of its current location; good example: Poker play: only public cards are observable for the agent):
- Formally this is a partially observable Markov decision process (POMDP):
  - Agent must construct its own state representation S, for example:
- Complete history: $S_t^a = H_t$
- Beliefs of environment state: $S_t^a = (\mathbb{P}[S_t^e = s^1], \ldots, \mathbb{P}[S_t^e = s^n])$
- Recurrent neural network: $S_t^a = \sigma(S_{t-1}^a W_s + O_t W_o)$

RL is multi-disciplinary and a bridge within ML

Decision Making



under uncertainty

02 Decision Making under uncertainty

Decision Making is central in Health Informatics

Source: Cisco (2008). Cisco Health Presence Trial at Aberdeen Royal Infirmary in Scotland

Reasoning Foundations of Medical Diagnosis

3 July 1959, Volume 130, Number 3366

SCIENCE

Reasoning Foundations of Medical Diagnosis

Symbolic logic, probability, and value theory aid our understanding of how physicians reason.

Robert S. Ledley and Lee B. Lusted

The purpose of this article is to analyze the complicated reasoning processes inherent in medical diagnosis. The importance of this problem has received recent emphasis by the increasing interest in the use of electronic computers as an aid to medical diagnostic processes

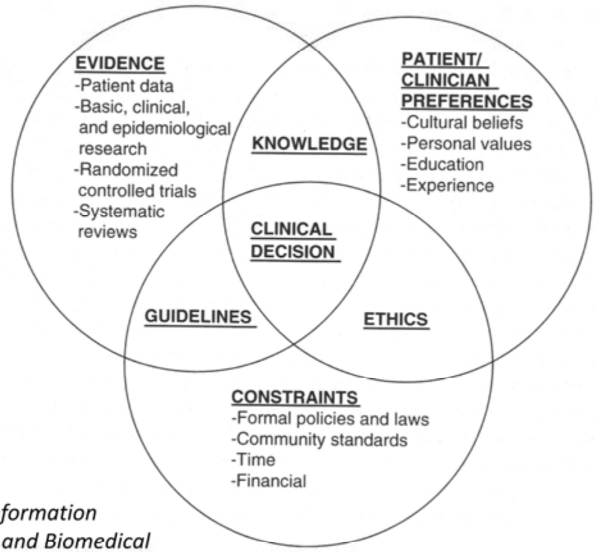
fitted into a definite disease category, or that it may be one of several possible diseases, or else that its exact nature cannot be determined." This, obviously, is a greatly simplified explanation of the process of diagnosis, for the physician might also comment that after seeing a

ance are the ones who do remember and consider the most possibilities."

Computers are especially suited to help the physician collect and process clinical information and remind him of diagnoses which he may have overlooked. In many cases computers may be as simple as a set of hand-sorted cards, whereas in other cases the use of a large-scale digital electronic computer may be indicated. There are other ways in which computers may serve the physician, and some of these are suggested in this paper. For example, medical students might find the computer an important aid in learning the methods of differential diagnosis. But to use the computer thus we must understand how the physician makes a medical diagnosis. This, then, brings us to the subject of our investigation: the reasoning foundations of medical diagnosis and treatment.

Medical diagnosis involves processes that can be systematically analyzed, as well as those characterized as "intangible." For instance, the reasoning foundations of medical diagnostic procedures

Clinical Medicine is Decision Making!



Hersh, W. (2010) *Information Retrieval: A Health and Biomedical Perspective. New York, Springer.*

History of DSS is a history of artificial intelligence

E. Feigenbaum, J. Lederberg, B. Buchanan, E. Shortliffe

Stanford Heuristic Programming Project
Memo HPP-78-1
February 1978

Computer Science Department
Report No. STAN-CS-78-649

DENDRAL AND META-DENDRAL:
THEIR APPLICATIONS DIMENSION

by

Bruce G. Buchanan and Edward A. Feigenbaum

COMPUTER SCIENCE DEPARTMENT
School of Humanities and Sciences
STANFORD UNIVERSITY

Rheingold, H. (1985) *Tools for thought: the history and future of mind-expanding technology. New York, Simon & Schuster.*

Buchanan, B. G. & Feigenbaum, E. A. (1978) DENDRAL and META-DENDRAL: their applications domain. *Artificial Intelligence, 11, 1978, 5-24.*

03 Roots of RL

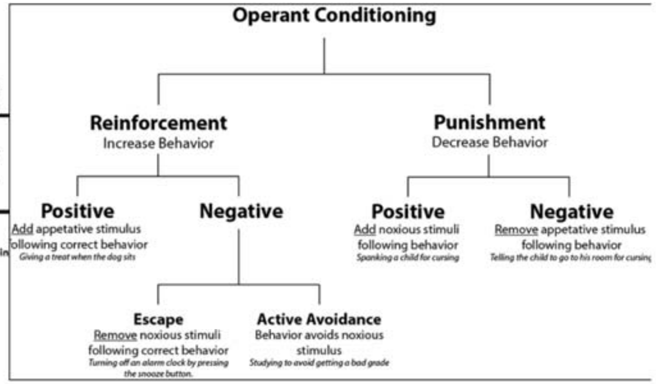
Pre-Historical Issues of RL

Ivan P. Pavlov (1849-1936)
1904 Nobel Prize
Physiology/Medicine

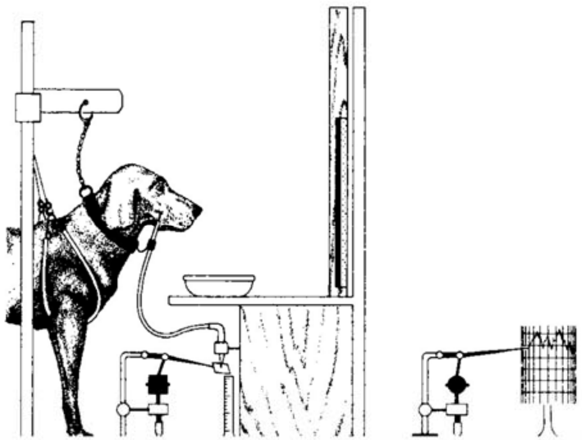
Edward L. Thorndike
(1874-1949)
1911 Law of Effect

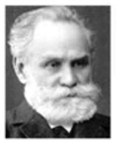
Burrhus F. Skinner
(1904-1990)
1938 Operant Conditioning

	Punishment (decreasing behavior)	Reinforcement (increasing behavior)
Positive (adding)	adding something to decrease behavior	adding something to increase behavior
Negative (subtracting)	subtracting something to decrease behavior	subtracting somethin to increase behavior



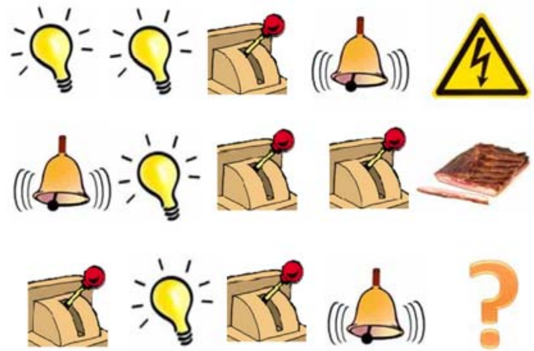
Classical Experiment with Pavlov's Dog





- *Classical (human and) animal conditioning*: "the magnitude and timing of the conditioned response changes as a result of the contingency between the conditioned stimulus and the unconditioned stimulus" [Pavlov, 1927].

Back to the rats ... roots ☺



- What if agent state = last 3 items in sequence?
- What if agent state = counts for lights, bells and levers?
- What if agent state = complete sequence?

JKU Historical Issues of RL in Computer Science

https://webdocs.cs.ualberta.ca/~sutton/book/the-book.html

Turing, A. M. 1950. Computing machinery and intelligence. Mind, 59, (236), 433-460.

Richard Bellman 1961. Adaptive control processes: a guided tour. Princeton.

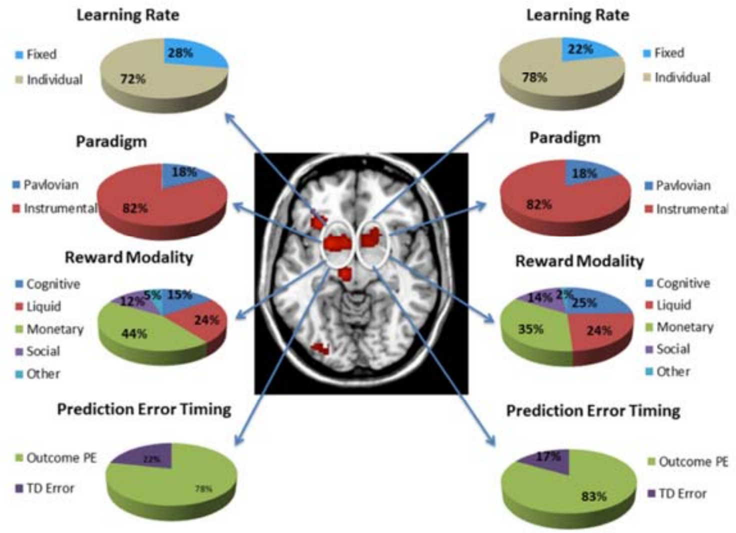
Watkins, C. J. & Dayan, P. 1992. Q-learning. Machine learning, 8, (3-4), 279-292.

Sutton, R. S. & Barto, A. G. 1998. Reinforcement learning: An introduction, Cambridge, MIT press.

Littman, M. L. 2015. Reinforcement learning improves behaviour from evaluative feedback. Nature, 521, (7553), 445-451.

Excellent Review Paper:
Kaelbling, L. P., Littman, M. L. & Moore, A. W. 1996. Reinforcement learning: A survey. Journal of Artificial Intelligence Research, 4, 237-285

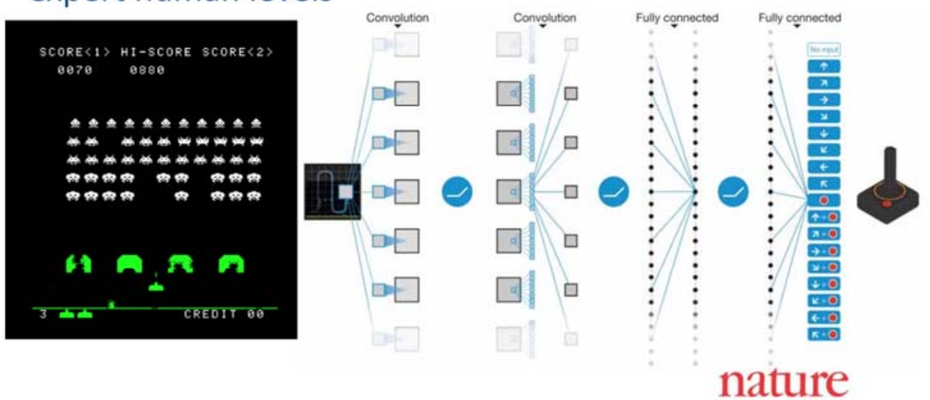
JKU This is still state-of-the-art in 2015



Chase, H. W., Kumar, P., Eickhoff, S. B. & Dombrovski, A. Y. 2015. Reinforcement learning models and their neural correlates: An activation likelihood estimation meta-analysis. Cognitive, Affective & Behavioral Neuroscience, 15, (2), 435-459, doi:10.3758/s13415-015-0338-7.

JKU 2015 – the year of reinforcement learning ☺

Deep Q-networks (Q-Learning is a model-free RL approach) have successfully played Atari 2600 games at expert human levels

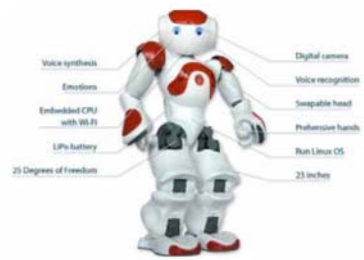


Mnih, V., Kavukcuoglu, K., Silver, D., Rusu, A. A., Veness, J., Bellemare, M. G., Graves, A., Riedmiller, M., Fidjeland, A. K., Ostrovski, G., Petersen, S., Beattie, C., Sadik, A., Antonoglou, I., King, H., Kumaran, D., Wierstra, D., Legg, S. & Hassabis, D. 2015. Human-level control through deep reinforcement learning. Nature, 518, (7540), 529-533, doi:10.1038/nature14236

JKU Typical Reinforcement Learning Applications: aML

httpimages.computerhistory.orgtimelinetimeline_ai.robotics_1939_elektro.jpg

1985



http://cyberneticzoo.com/robot-time-line/

http://www.neurotechnology.com/res/Robot2.jpg

https://royalsociety.org/events/2015/05/breakthrough-science-technologies-machine-learning

Kober, J., Bagnell, J. A. & Peters, J. 2013. Reinforcement Learning in Robotics: A Survey. The International Journal of Robotics Research.

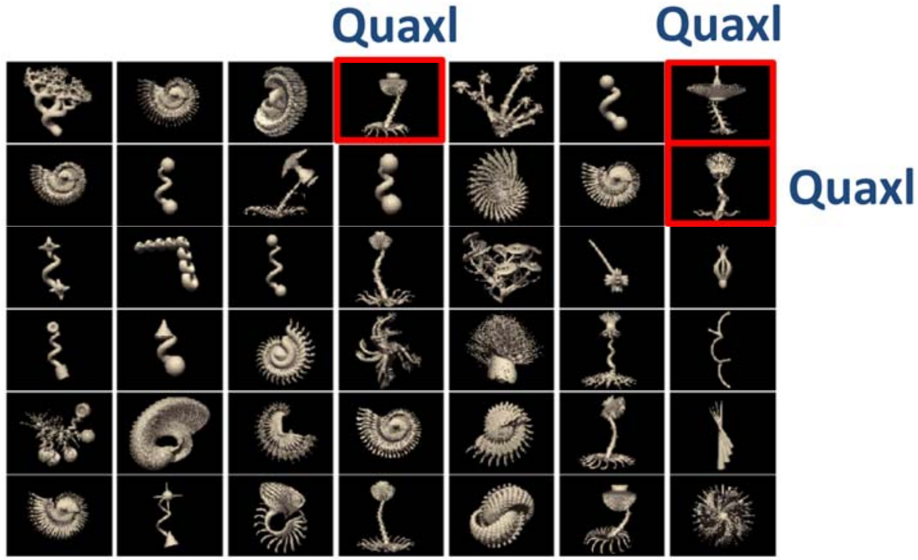
Nogrady, B. 2015. Q&A: Declan Murphy. Nature, 528, (7582), S132-S133, doi:10.1038/528S132a.

04 Cognitive Science of R-Learning: Human Information Processing

How does our mind get so much out of it …

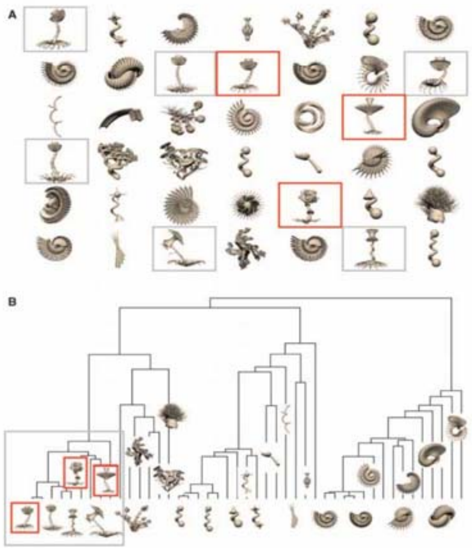
Salakhutdinov, R., Tenenbaum, J. & Torralba, A. 2012. One-shot learning with a hierarchical nonparametric Bayesian model. Journal of Machine Learning Research, 27, 195-207.

Learning words for objects – concepts from examples

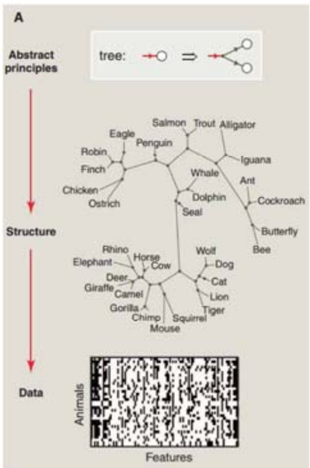


Salakhutdinov, R., Tenenbaum, J. & Torralba, A. 2012. One-shot learning with a hierarchical nonparametric Bayesian model. Journal of Machine Learning Research, 27, 195-207.

How do we understand our world …



$$P(h|d) = \frac{P(d|h)P(h)}{\sum_{h' \in H} P(d|h')P(h')} \propto P(d|h)P(h)$$



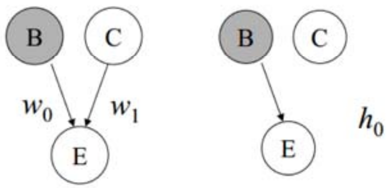
Tenenbaum, J. B., Kemp, C., Griffiths, T. L. & Goodman, N. D. 2011. How to grow a mind: Statistics, structure, and abstraction. Science, 331, (6022), 1279-1285.

One of the unsolved problems in human concept learning

- which is highly relevant for ML research, concerns the factors that determine the subjective difficulty of concepts:
- Why are some concepts psychologically extremely simple and easy to learn,
- while others seem to be extremely difficult, complex, or even incoherent?
- These questions have been studied since the 1960s but are still unanswered …

Feldman, J. 2000. Minimization of Boolean complexity in human concept learning. Nature, 407, (6804), 630-633, doi:10.1038/35036586.

A few certainties

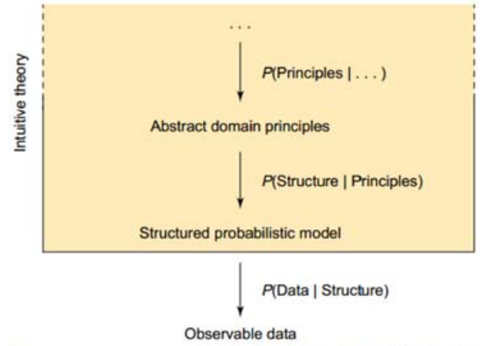


- Cognition as probabilistic inference
  - Visual perception, language acquisition, motor learning, associative learning, memory, attention, categorization, reasoning, causal inference, decision making, theory of mind
- Learning concepts from examples
- Learning and applying intuitive theories (balancing complexity vs. fit)

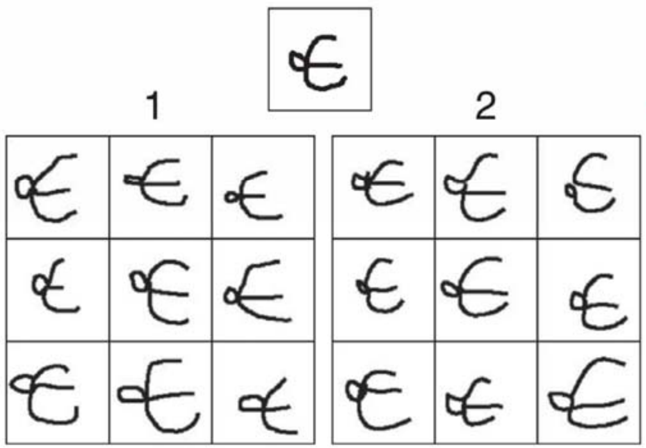
Modeling basic cognitive capacities as intuitive Bayes

- Similarity
- Representativeness and evidential support
- Causal judgement
- Coincidences and causal discovery
- Diagnostic inference
- Predicting the future

Tenenbaum, J. B., Griffiths, T. L. & Kemp, C. 2006. Theory-based Bayesian models of inductive learning and reasoning. Trends in cognitive sciences, 10, (7), 309-318.



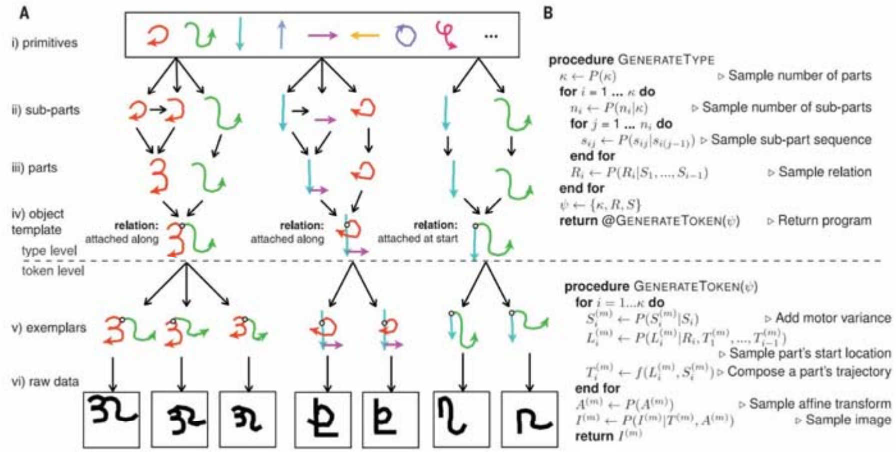
Drawn by Human or Machine Learning Algorithm?



Lake, B. M., Salakhutdinov, R. & Tenenbaum, J. B. 2015. Human-level concept learning through probabilistic program induction. Science, 350, (6266), 1332-1338, doi:10.1126/science.aab3050.

Human-Level concept learning – probabilistic induction

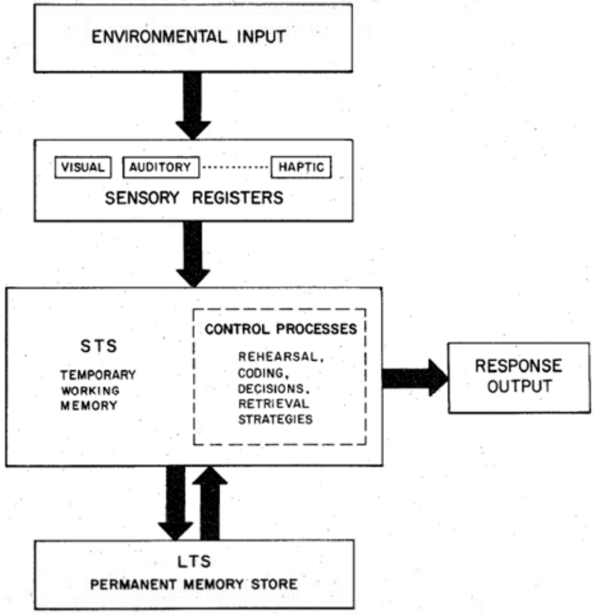
A Bayesian program learning (BPL) framework, capable of learning a large class of visual concepts from just a single example and generalizing in ways that are mostly indistinguishable from people



Lake, B. M., Salakhutdinov, R. & Tenenbaum, J. B. 2015. Human-level concept learning through probabilistic program induction. Science, 350, (6266), 1332-1338, doi:10.1126/science.aab3050.

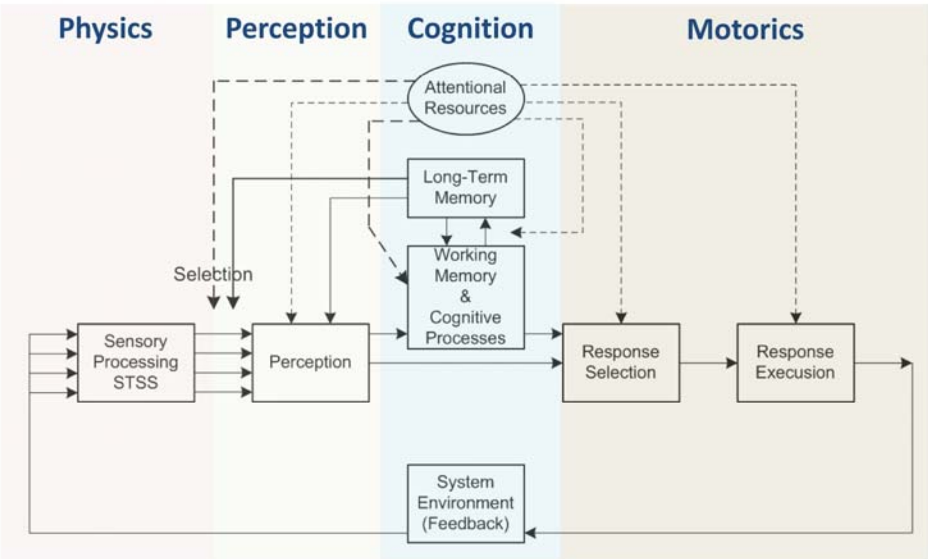
How does our mind get so much out of so little?

Human Information Processing Model (A&S)



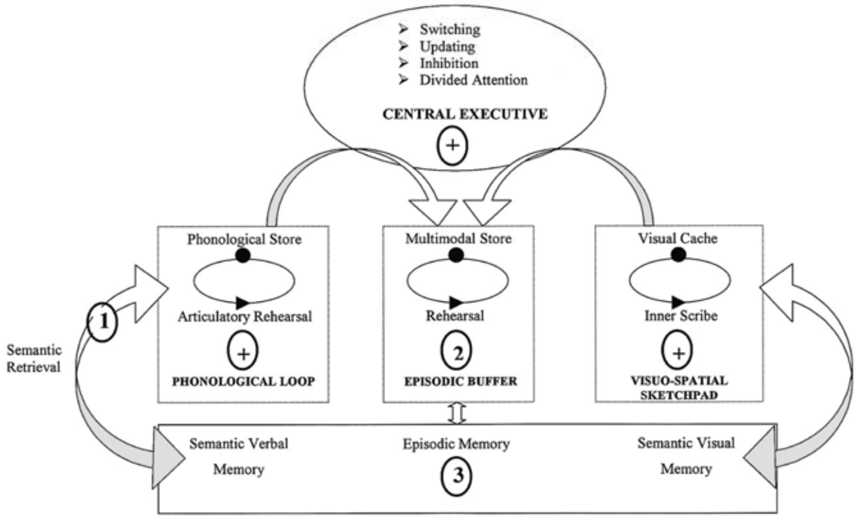
Atkinson, R. C. & Shiffrin, R. M. (1971) *The control processes of short-term memory (Technical Report 173, April 19, 1971). Stanford, Institute for Mathematical Studies in the Social Sciences, Stanford University.*

General Model of Human Information Processing



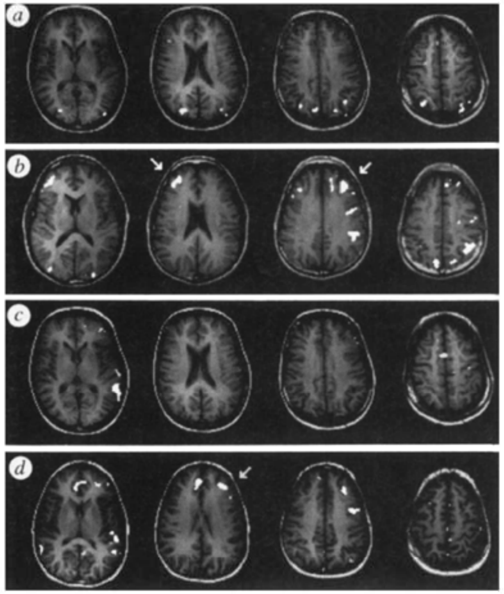
Wickens, C., Lee, J., Liu, Y. & Gordon-Becker, S. (2004) *Introduction to Human Factors Engineering: Second Edition. Upper Saddle River (NJ), Prentice-Hall.*

Alternative Model: Baddeley - Central Executive



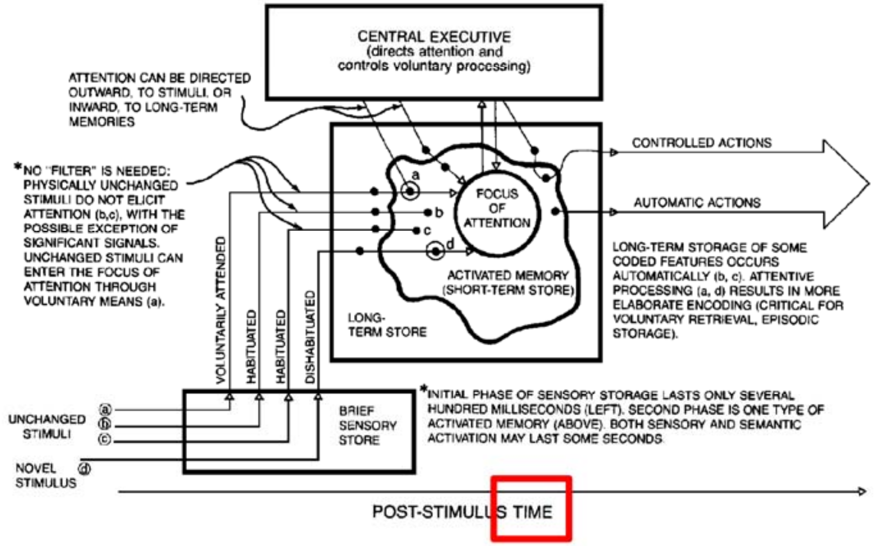
Quinette, P., Guillery, B., Desgranges, B., de la Sayette, V., Viader, F. & Eustache, F. (2003) Working memory and executive functions in transient global amnesia. *Brain, 126, 9, 1917-1934.*

Neural Basis for the "Central Executive System"



D'Esposito, M., Detre, J. A., Alsop, D. C., Shin, R. K., Atlas, S. & Grossman, M. (1995) The neural basis of the central executive system of working memory. *Nature, 378, 6554, 279-281.*

Slide 7-14 Central Executive – Selected Attention



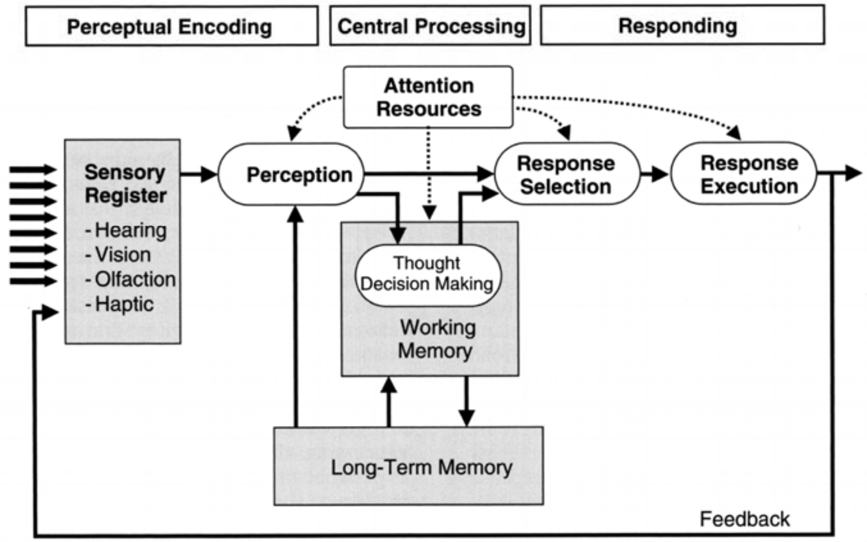
Cowan, N. (1988) Evolving conceptions of memory storage, selective attention, and their mutual constraints within the human information-processing system. *Psychological Bulletin, 104, 2, 163.*

Selective Attention Test

Note: The Test does NOT properly work if you know it in advance or if you do not concentrate on counting

Simons, D. J. & Chabris, C. F. 1999. Gorillas in our midst: sustained inattentional blindness for dynamic events. Perception, 28, (9), 1059-1074.

Human Attention is central for decision making

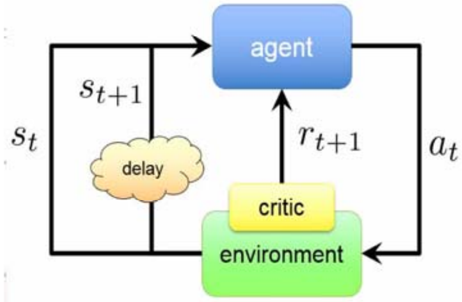


Wickens, C. D. (1984) *Engineering psychology and human performance. Columbus (OH), Charles Merrill.*

05 The Anatomy of an R-Learning Agent

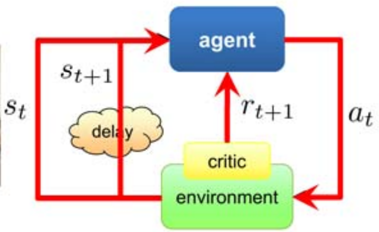
Why is this relevant for health informatics?

- Decision-making under uncertainty
- Limited knowledge of the domain environment
- Unknown outcome – unknown reward
- Partial or unreliable access to "databases of interaction"



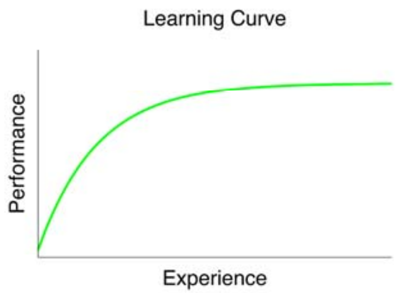
Russell, S. J. & Norvig, P. 2009. Artificial intelligence: a modern approach (3rd edition), Prentice Hall, Chapter 16, 17: Making Simple Decisions and **Making Complex Decisions**

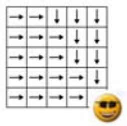
Decision Making under uncertainty



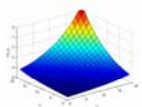
$$\langle s, a, s', r\rangle_1$$
$$\langle s, a, s', r\rangle_2$$
$$\vdots$$
$$\langle s, a, s', r\rangle_n$$

<position, speed> *<carpet, chair>* *<new position, new speed>, advancement*





Control Policy



Value Function

Image credit to Allessandro Lazaric

Taxonomy of RL agents 1/2: A Components

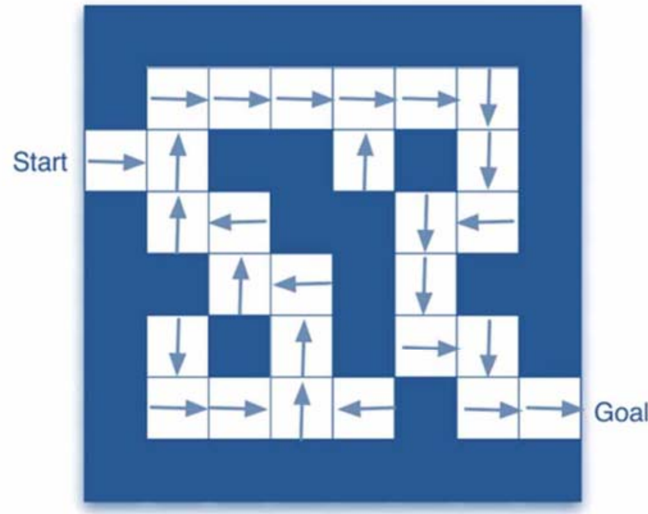
- **Policy:** agent's behaviour function
  e.g. stochastic policy $\pi(a|s) = \mathbb{P}[A_t = a | S_t = s]$
- **Value function:** how good is each state and/or action
  e.g. $v_\pi(s) = \mathbb{E}_\pi \left[R_{t+1} + \gamma R_{t+2} + \gamma^2 R_{t+3} + \ldots \mid S_t = s\right]$
- **Model:** agent's representation of the environment
  $\mathcal{P}$ predicts the next state; $\mathcal{R}$ the next reward

$$\mathcal{P}^a_{ss'} = \mathbb{P}[S_{t+1} = s' \mid S_t = s, A_t = a]$$
$$\mathcal{R}^a_s = \mathbb{E}\left[R_{t+1} \mid S_t = s, A_t = a\right]$$

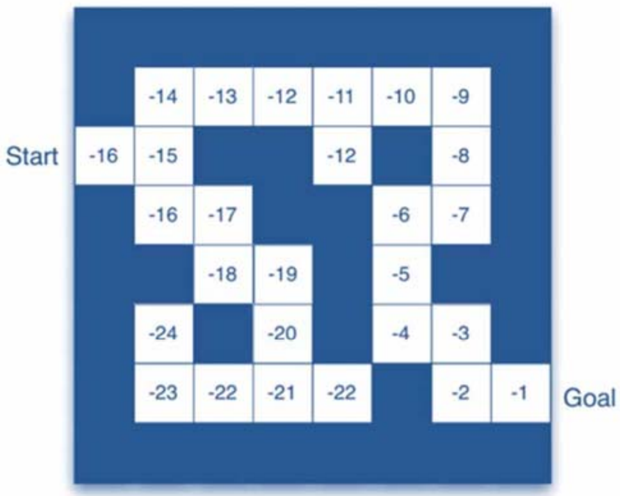
Taxonomy of agents 2/2 B Categorization

- 1) Value-Based
  (no policy, only value function)
- 2) Policy-Based
  (no value function, only policy)
- 3) Actor-Critic
  (both)
- 4) Model free
  (and/or) – but no model
- 5) Model-based
  (and/or – and model)

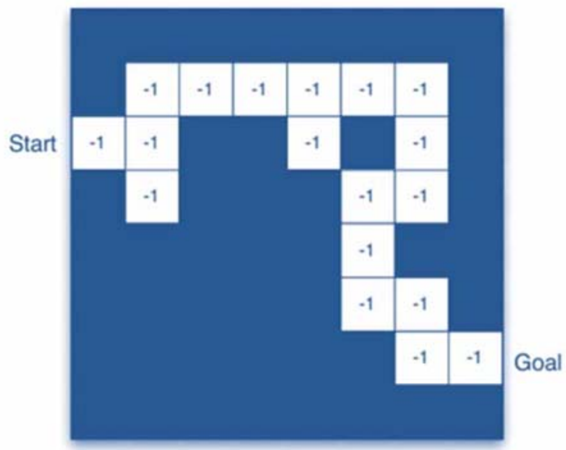
Maze Example: Policy



Maze Example: Value Function



Maze Example: Model



- Grid layout represents transition model $\mathcal{P}^a_{ss'}$
- Numbers represent immediate reward $\mathcal{R}^a_s$ from each state s (same for all a)

Principle of a RL algorithm is simple

Time steps $t_1, t_2, \ldots, t_n$

- Observe the state x_t
- Take an action a_t (problem of **exploration** and **exploitation**)
- Observe next state and earn reward x_{t+1}, r_t
- Update the policy and the value function π_t, Q_t

$$Q(x_t, a_t) = Q(x_t, a_t) + \alpha\big(r_t + \gamma \max_a Q(x_{t+1}, a) - Q(x_t, a_t)\big)$$

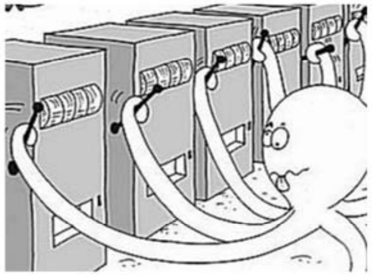
$$\pi(x) = \arg\max_a Q(x, a)$$

Example RL Algorithms

- Temporal difference learning (1988)
- Q-learning (1998)
- BayesRL (2002)
- RMAX (2002)
- CBPI (2002)
- PEGASUS (2002)
- Least-Squares Policy Iteration (2003)
- Fitted Q-Iteration (2005)
- GTD (2009)
- UCRL (2010)
- REPS (2010)
- DQN (2014)

06 Example: Multi-Armed Bandits (MAB)

Principle of the Multi-Armed Bandits problem (1/2)



- There are n slot-machines (“einarmige Banditen”)
- Each machine i returns a reward $y \approx P(y;\ \Theta_i)$
- Challenge: The machine parameter Θ_i is unknown
- Which arm of which slot machine should a gambler pull to **maximize** his cumulative reward over a sequence of trials? (stochastic setting or adversarial setting)

Image credit and more information: http://research.microsoft.com/en-us/projects/bandits

Principle of the Multi-Armed Bandits problem (2/2)

- Let $a_t \in \{1, \dots, n\}$ be the choice of a machine at time t
- Let $y_t \in \mathbb{R}$ be the outcome with a mean of $\langle y_{a t} \rangle$
- Now, the given policy maps all history to a new choice:

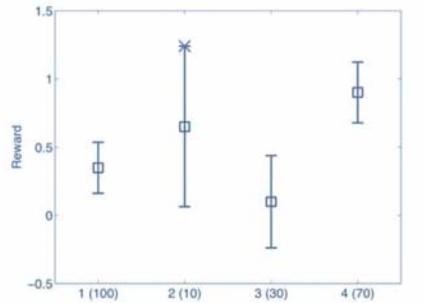
$$\pi : [(a_1, y_1), (a_2, y_2), \dots, (a_{t\text{-}1}, y_{t\text{-}1})] \mapsto a_t$$

- The problem: Find a policy π that $\max\langle y_T \rangle$
- Now, two effects appear when choosing such machine:
  - You collect more data about the machine (=knowledge)
  - You collect reward
- Exploration and Exploitation
  - **Exploration:** Choose the next action a_t to $min\langle H(b_t)\rangle$
  - **Exploitation:** Choose the next action a_t to $max\langle y_t \rangle$
- models an agent that simultaneously attempts to acquire new knowledge (called "exploration") and optimize his or her decisions based on existing knowledge (called "exploitation"). The agent attempts to balance these competing tasks in order to maximize total value over the period of time considered.

More information: http://research.microsoft.com/en-us/projects/bandits

MAP-Principle: "Optimism in the face of uncertainty"

$$a_t = \max_{a \in \mathcal{A}} \left(\widehat{r}_t(a) + \sqrt{\frac{\log(1/\delta)}{T_t(a)}} \right)$$



$$a_t = \max_{a \in \mathcal{A}} \left(\text{rew}_t(a) + \text{uncert}_t(a) \right)$$

Exploitation
the higher the (estimated) reward the higher the chance to select the action

Exploration
the higher the (theoretical) uncertainty the higher the chance to select the action

Auer, P., Cesa-Bianchi, N. & Fischer, P. 2002. Finite-time analysis of the multiarmed bandit problem. Machine learning, 47, (2-3), 235-256.

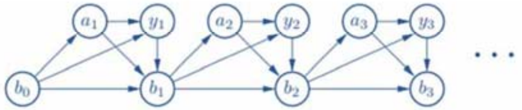
Knowledge Representation in MAB

- Knowledge can be represented in two ways:
- 1) as full history $h_t = [(a_1, y_1), (a_2, y_2), \dots, (a_{t\text{-}1}, y_{t\text{-}1})]$

  or
- 2) as belief $b_t(\theta) = P(\theta | h_t)$

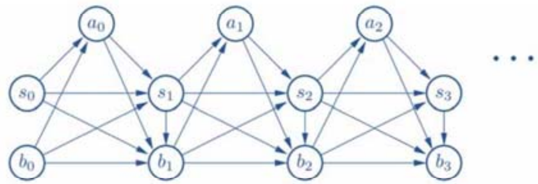
where Θ are the unknown parameters of all machines

The process can be modelled as belief MDP:



$$P(b' | y, a, b) = \begin{cases} 1 & \text{if } b' = b'_{[b,a,y]} \\ 0 & \text{otherwise} \end{cases}, \quad P(y|a, b) = \int_{\theta_a} b(\theta_a)\, P(y|\theta_a)$$

The optimal policies can be modelled as belief MDP



$$P(b' | s', s, a, b) = \begin{cases} 1 & \text{if } b' = b[s', s, a] \\ 0 & \text{otherwise} \end{cases}, \quad P(s'|s, a, b) = \int_\theta b(\theta)\, P(s'|s, a, \theta)$$

$$V(b, s) = \max_a \left[\mathsf{E}(r|s, a, b) + \textstyle\sum_{s'} P(s'|a, s, b)\, V(s', b') \right]$$

Poupart, P., Vlassis, N., Hoey, J. & Regan, K. An analytic solution to discrete Bayesian reinforcement learning. Proceedings of the 23rd international conference on Machine learning, 2006. ACM, 697-704.

Applications of MABs

- Clinical trials: potential treatments for a disease to select from new patients or patient category at each round, see:

W. Thompson. On the likelihood that one unknown probability exceeds another in view of the evidence of two samples. Bulletin of the American Mathematics Society, vol. 25, pp. 285–294, 1933.

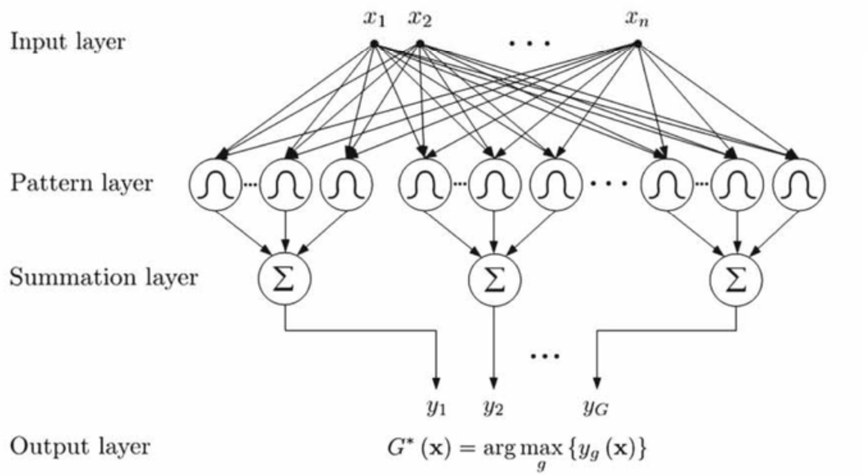
- Games: Different moves at each round, e.g. GO
- Adaptive routing: finding alternative paths, also finding alternative roads for driving from A to B
- Advertisement placements: selection of an ad to display at the Webpage out of a finite set which can vary over time, for each new Web page visitor

07 Applications in Health

Example for Health

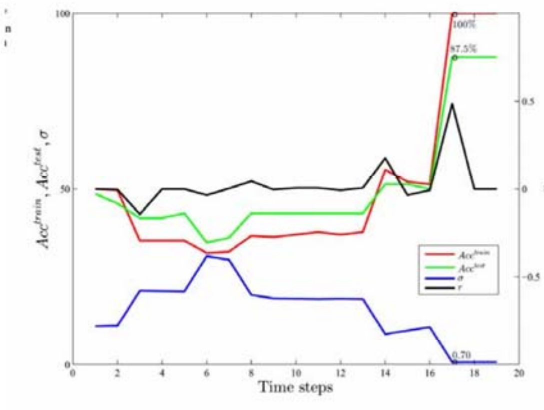
https://www.youtube.com/watch?v=20sj7rRfzm4

Example: Q-Learning



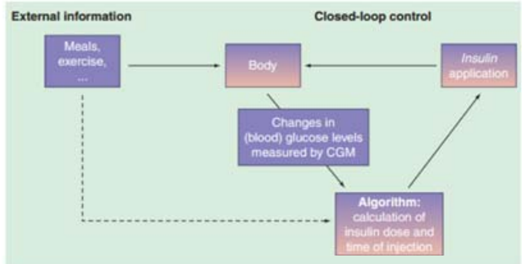
Kusy, M. & Zajdel, R. 2014. Probabilistic neural network training procedure based on Q(0)-learning algorithm in medical data classification. *Applied Intelligence,* 41, (3), 837-854, doi:10.1007/s10489-014-0562-9.

Sample Data Sets



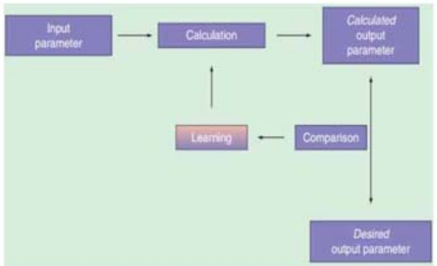
– Wisconsin breast cancer database [24] that consists of 683 instances with 9 attributes. The data is divided into two groups: 444 benign cases and 239 malignant cases.
Pima Indians diabetes data set [36] that includes 768 cases having 8 features. Two classes of data are considered: samples tested negative (500 records) and samples tested positive (268 records).
Haberman's survival data [21] that contains 306 patients who underwent surgery for breast cancer. For each instance, 3 variables are measured. The 5-year survival status establishes two input classes: patients who survived 5 years or longer (225 records) and patients who died within 5 years (81 records).
Cardiotocography data set [3] that comprises 2126 measurements of fetal heart rate and uterine contraction features on 22 attribute cardiotocograms classified by expert obstetricians. The classes are coded into three states: normal (1655 cases), suspect (295 cases) and pathological (176 cases).
Dermatology data [13] that includes 358 instances each of 34 features. Six data classes are considered: psoriasis (111 cases), lichen planus (71 cases), seborrheic dermatitis (60 cases), cronic dermatitis (48 cases), pityriasis rosea (48 cases) and pityriasis rubra pilaris (20 cases).

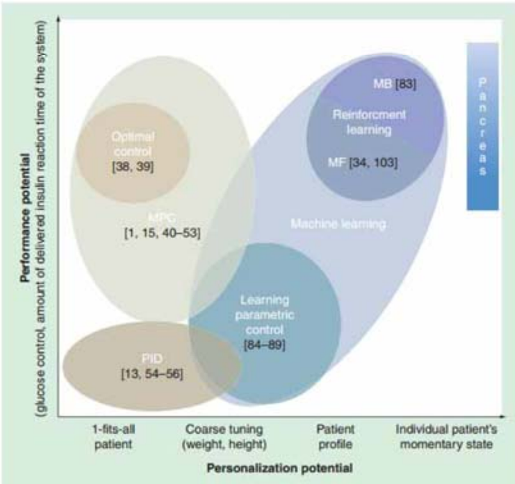
Example (1/3)



Bothe, M. K., Dickens, L., Reichel, K., Tellmann, A., Ellger, B., Westphal, M. & Faisal, A. A. 2013. The use of reinforcement learning algorithms to meet the challenges of an artificial pancreas. Expert Review of Medical Devices, 10, (5), 661-673, doi:10.1586/17434440.2013.827515.

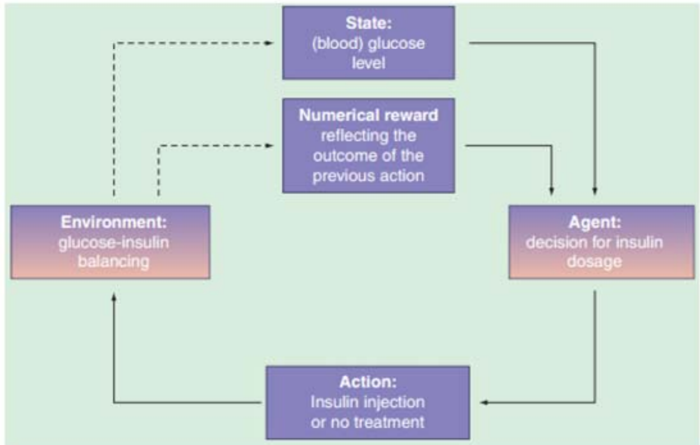
Example (2/3)





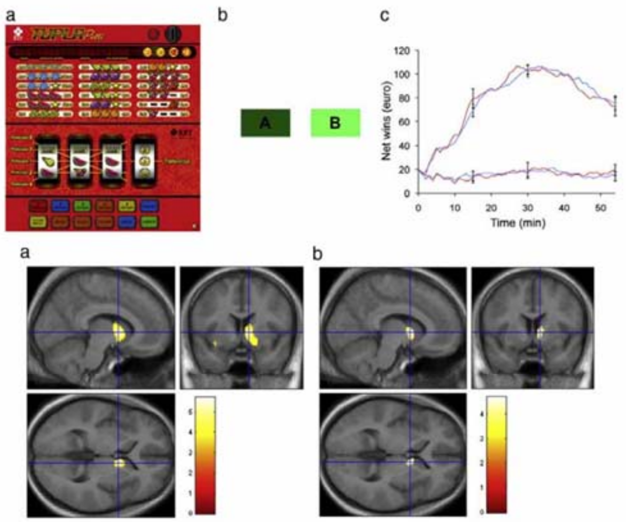
Bothe, M. K., Dickens, L., Reichel, K., Tellmann, A., Ellger, B., Westphal, M. & Faisal, A. A. 2013. The use of reinforcement learning algorithms to meet the challenges of an artificial pancreas. Expert Review of Medical Devices, 10, (5), 661-673, doi:10.1586/17434440.2013.827515.

Example (3/3)



Bothe, M. K., Dickens, L., Reichel, K., Tellmann, A., Ellger, B., Westphal, M. & Faisal, A. A. 2013. The use of reinforcement learning algorithms to meet the challenges of an artificial pancreas. Expert Review of Medical Devices, 10, (5), 661-673, doi:10.1586/17434440.2013.827515.

Example of a work with MABs



Joutsa et al. (2012) Mesolimbic dopamine release is linked to symptom severity in pathological gambling. *NeuroImage*, 60, (4), 1992-1999, doi.org/10.1016/j.neuroimage.2012.02.006.

Thank you!

Questions

Sample Questions

- Why is RL - for us in health informatics - interesting?
- What is a medical doctor in daily clinical routine doing most of the time?
- Please explain the human decision making process on the basis of the model by Wickens (1984) !
- What is the underlying principle of DQN?
- What is probabilistic inference? Give an example!
- Why is selective attention so important?
- Please describe the “anatomy” of a RL-agent!
- What does policy-based RL-agent mean? Give an example!
- What is the underlying principle of a MAB? Why is it interesting for health informatics?

Keywords

- Reinforcement Learning
- Trial-and-Error Learning
- Markov-Decision-Process
- Utility-based agent
- Q-Learning
- Passive reinforcement learning
- Adaptive dynamic programming
- Temporal-difference learning
- Active reinforcement learning
- Bandit problems

Advance Organizer (1)

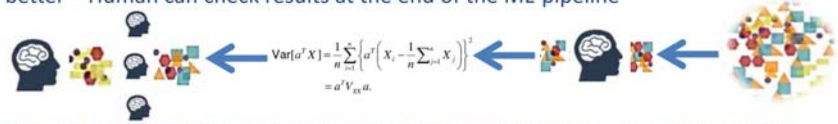
- RL:= general problem, inspired by behaviorist psychology; how software agents learn to make decisions from success and failure, from reward and punishment in an environment – aiming to maximize cumulative reward.
- RL is studied in game theory, control theory, operations research, information theory, simulation-based optimization, multi-agent systems, swarm intelligence, genetic algorithms.
- Aka: approximate dynamic programming.
- The problem has been studied in the theory of optimal control, though most studies are concerned with the existence of optimal solutions and their characterization, and not with the learning or approximation aspects. In economics and game theory, reinforcement learning may be used to explain how equilibrium may arise under bounded rationality.

Appendix

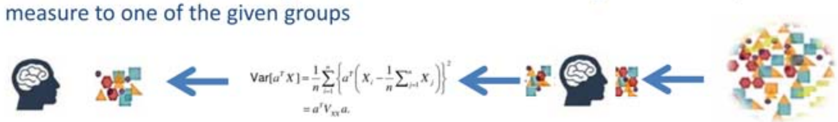
Unsupervised – Supervised – Semi-supervised

A) Unsupervised ML: Algorithm is applied on the raw data and learns fully automatic – Human can check results at the end of the ML-pipeline

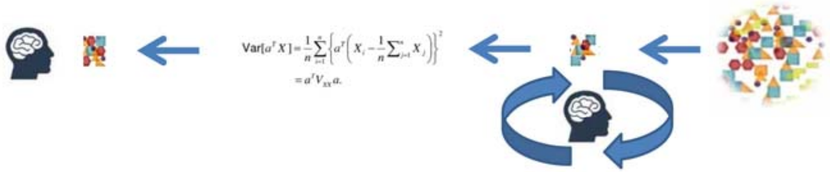
B) Supervised ML: Humans are providing the labels for the training data and/or select features to feed the algorithm to learn – the more samples the better – Human can check results at the end of the ML-pipeline



C) Semi-Supervised Machine Learning: A mixture of A and B – mixing labeled and unlabeled data so that the algorithm can find labels according to a similarity measure to one of the given groups

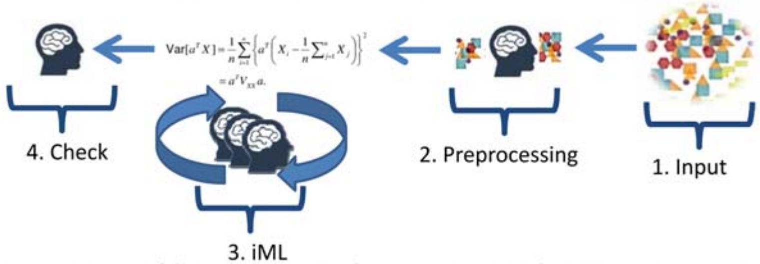


D) Reinforcement Learning: Algorithm is continually trained by human input, and can be automated once maximally accurate



- Advantage: non-greedy nature
- Disadvantage: must learn model of environment

E) **Interactive Machine Learning:** Human is seen as an agent involved in the actual learning phase, step-by-step influencing measures such as distance, cost functions …



Constraints of humans: Robustness, subjectivity, transfer?

Open Questions: Evaluation, replicability, …

Holzinger, A., Plass, M., Holzinger, K., Crisan, G., Pintea, C. & Palade, V. 2016. Towards interactive Machine Learning (iML): Applying Ant Colony Algorithms to solve the Traveling Salesman Problem with the Human-in-the-Loop approach. Springer Lecture Notes in Computer Science LNCS 9817. Heidelberg, Berlin, New York: Springer, pp. 81-95, doi:10.1007/978-3-319-45507-56.