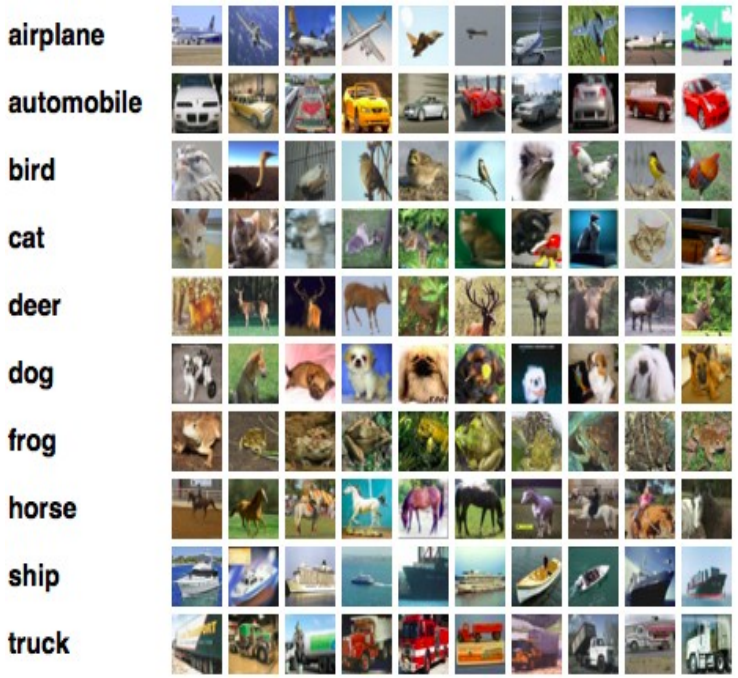
CS6700:
Reinforcement Learning

B. Ravindran

J Slot: W: 2 – 3:15, Th 3:30 – 4:45
Tutorial: M: 5 – 5:50

Machine Learning

Learn functions from input to output from data



Learning to Control

- Familiar models of machine learning
  - Learning from data.

- How did you learn to cycle?
  - Trial and error!
  - Falling down hurts!
  - Evaluation, not instruction
  - Reinforcement Learning

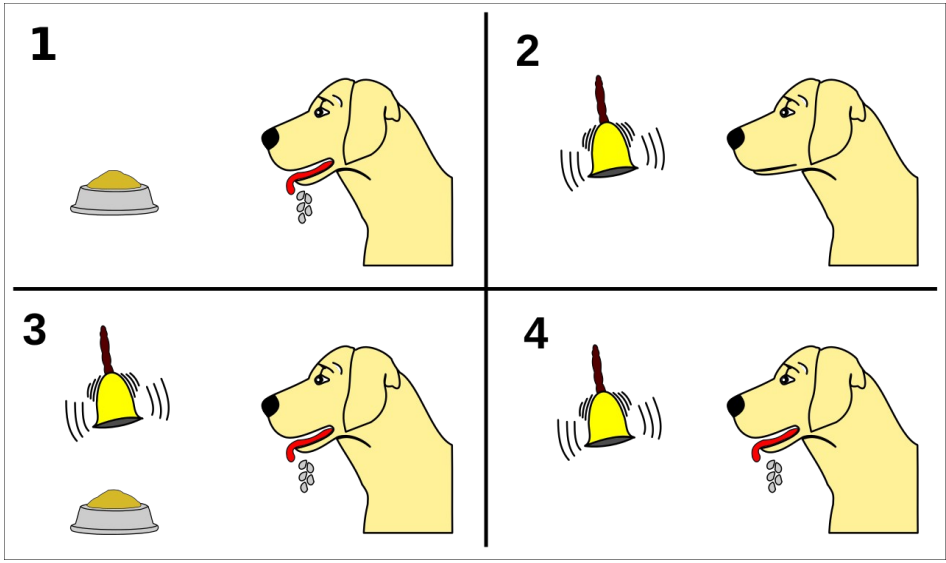
- Walk, Talk, etc.

Reinforcement Learning

- A trial-and-error learning paradigm
  - Rewards and Punishments
- Learn about a system through interaction
- Inspired by behavioural psychology!
  - Pavlov's dog

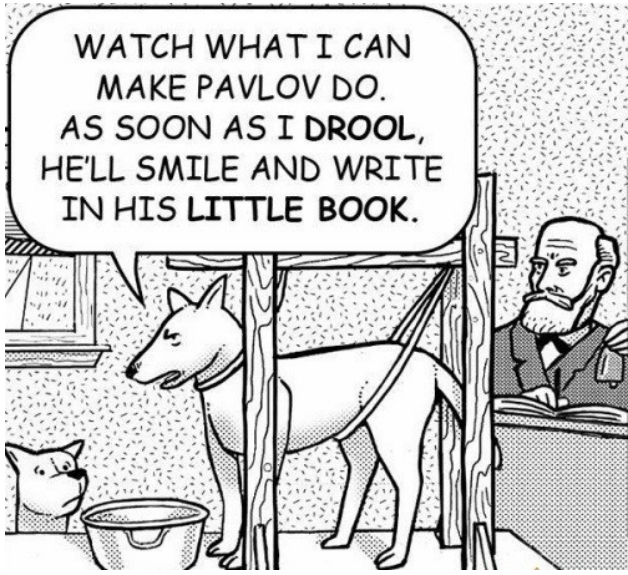
Reinforcement Learning

- A trial-and-error learning paradigm
  - Rewards and Punishments
- Learn about a system through interaction
- Inspired by behavioural psychology!
  - Pavlov's dog

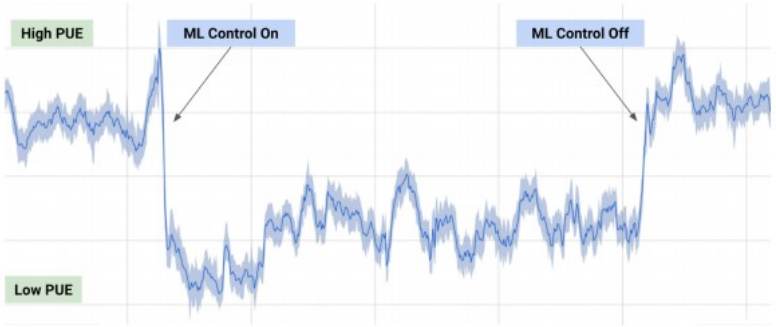


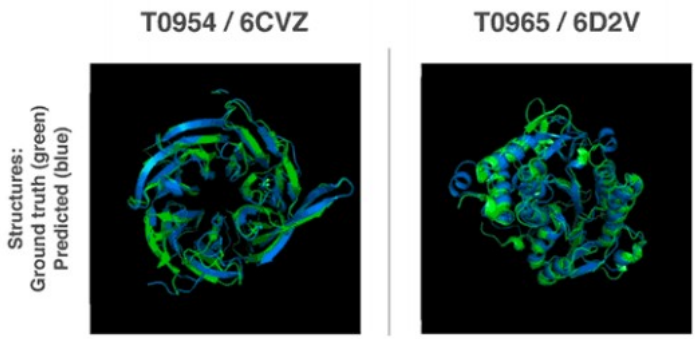
Reinforcement Learning

- A trial-and-error learning paradigm
  - Rewards and Punishments
- Learn about a system through interaction
- Inspired by behavioural psychology!
  - Pavlov's dog



Reinforcement Learning Works!

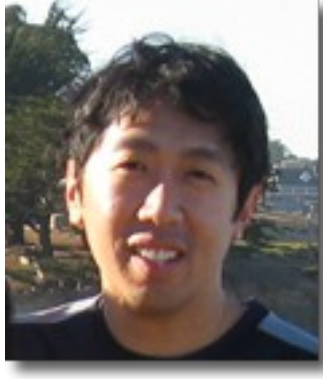




Why RL

- Complex Dynamics
  - Helicopter control

Helicopter Control

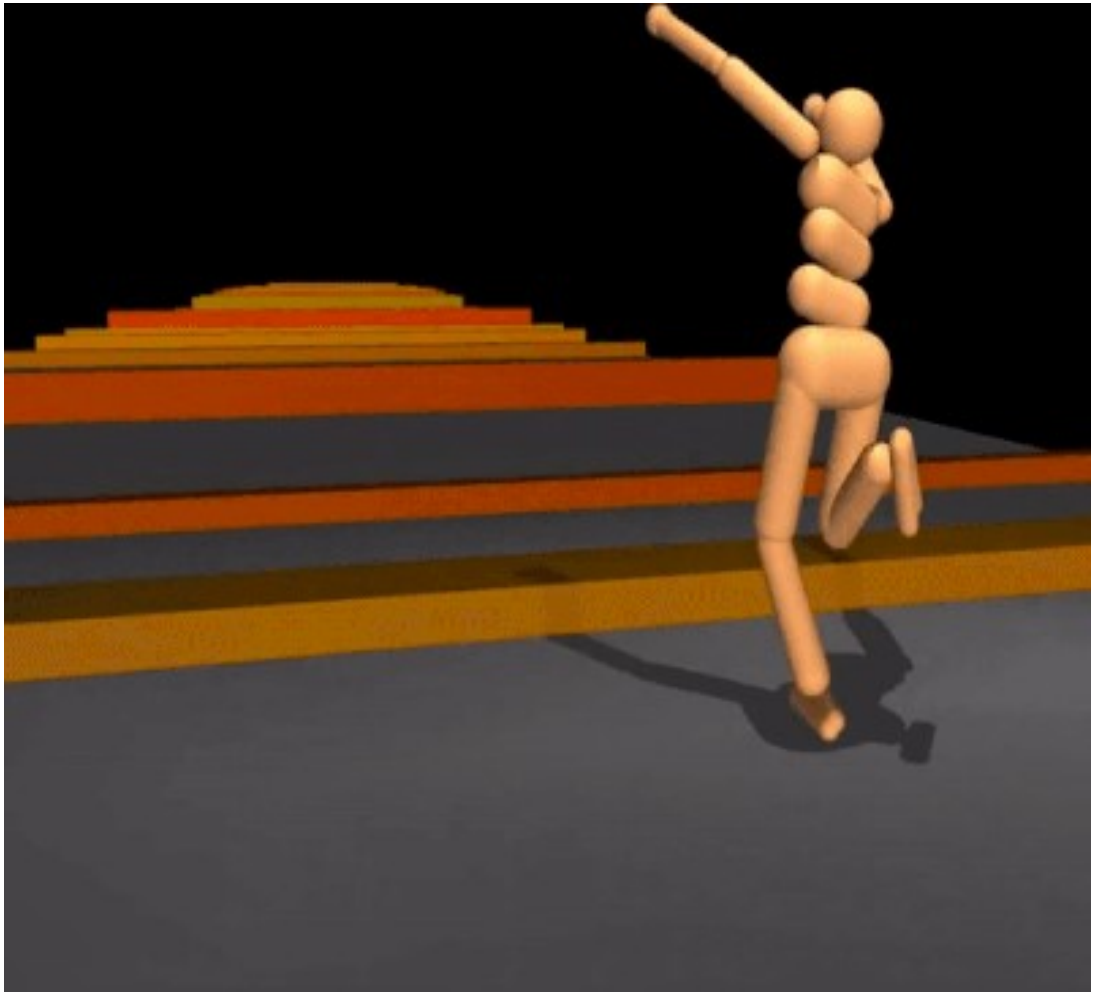


Why RL

- Complex Dynamics
  - Helicopter control
  - Humanoid control

Humanoid Control

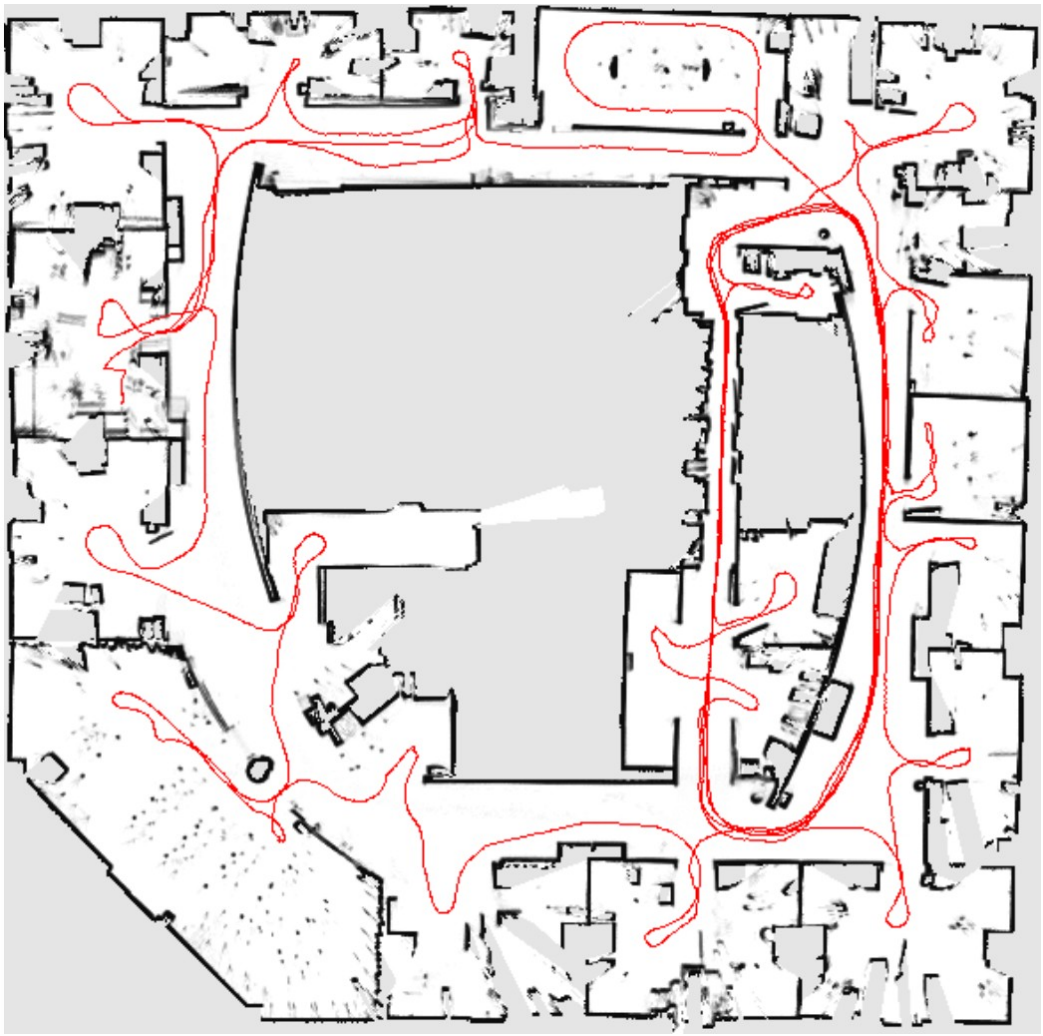
Humanoid Running!



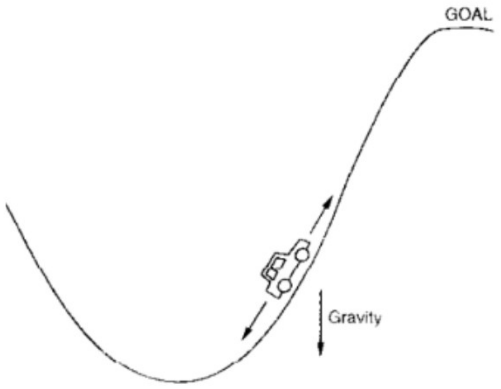
Why RL

- Complex Dynamics
- Complex Workspace

Path Planning



Path Planning

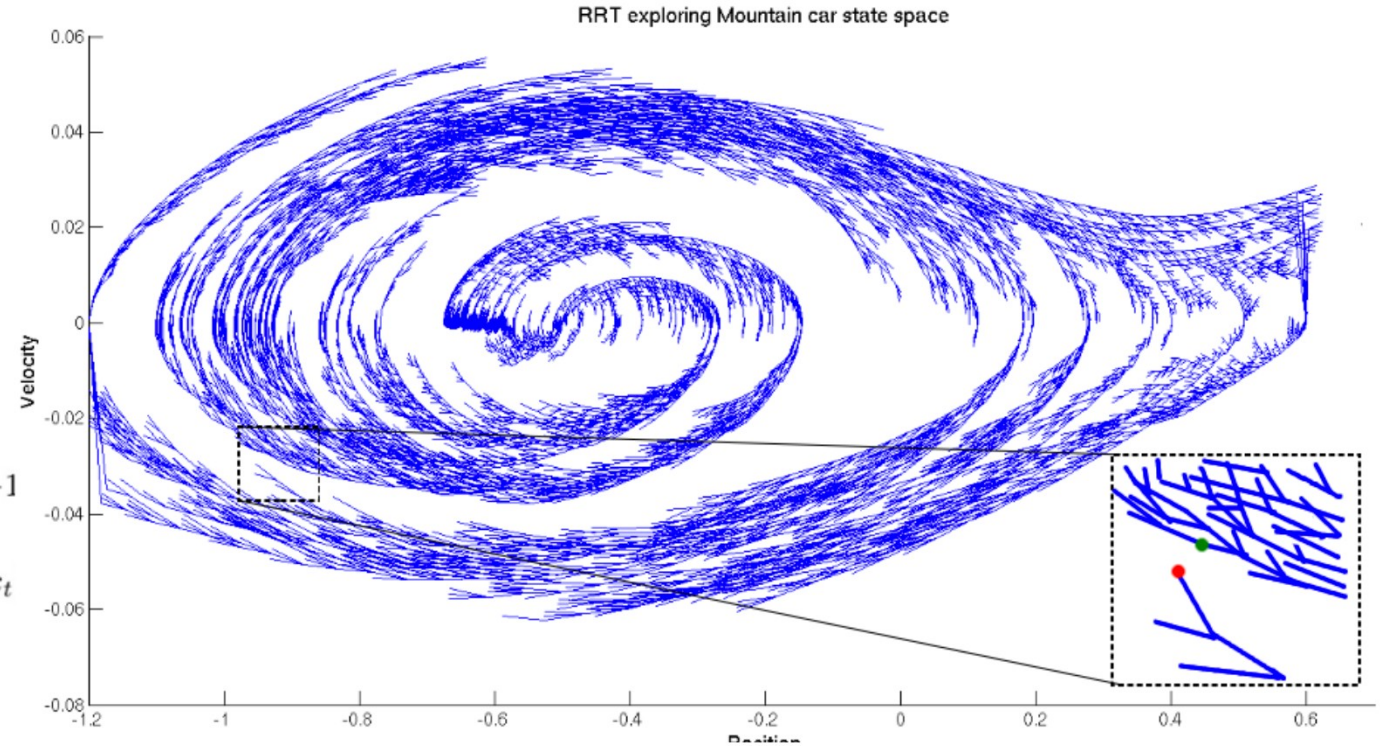


State :– (x, v), Control :– $\{+1, 0, -1$

$x_{t+1} = x_t + v_t$

$v_{t+1} = v_t + 0.001u_t - 0.0025\cos(3x_t$

-1 reward till goal



- Learn distance function ICRA 2014

Cluttered Workspace

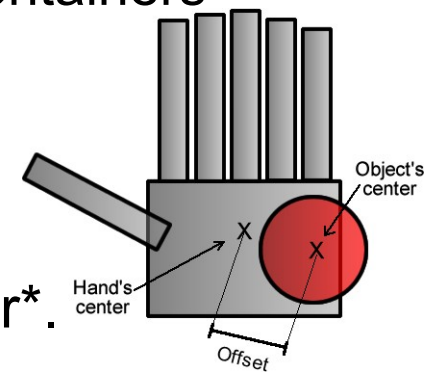
Why RL

- Complex Dynamics
- Complex Workspace
- Stochastic sensing and actuation

Visual Attention ICRA 2012

Manipulation Task:

- Goal: To put as many objects as possible inside the containers
- Uncertainty: Location of objects and containers
  - Duration: 5 minutes each trial
  - We implement our model using the iCub simulator*.



*Metta el al, "The iCub humanoid robot: An open platform for research in embodied cognition," *Proc. ACM Perf. Metrics for Inf. Sys.* 2008, 50-56.

Why RL

- Complex Dynamics
- Complex Workspace
- Stochastic sensing and actuation
- Human-in-the-loop Learning
  - Instructions
  - Imitation

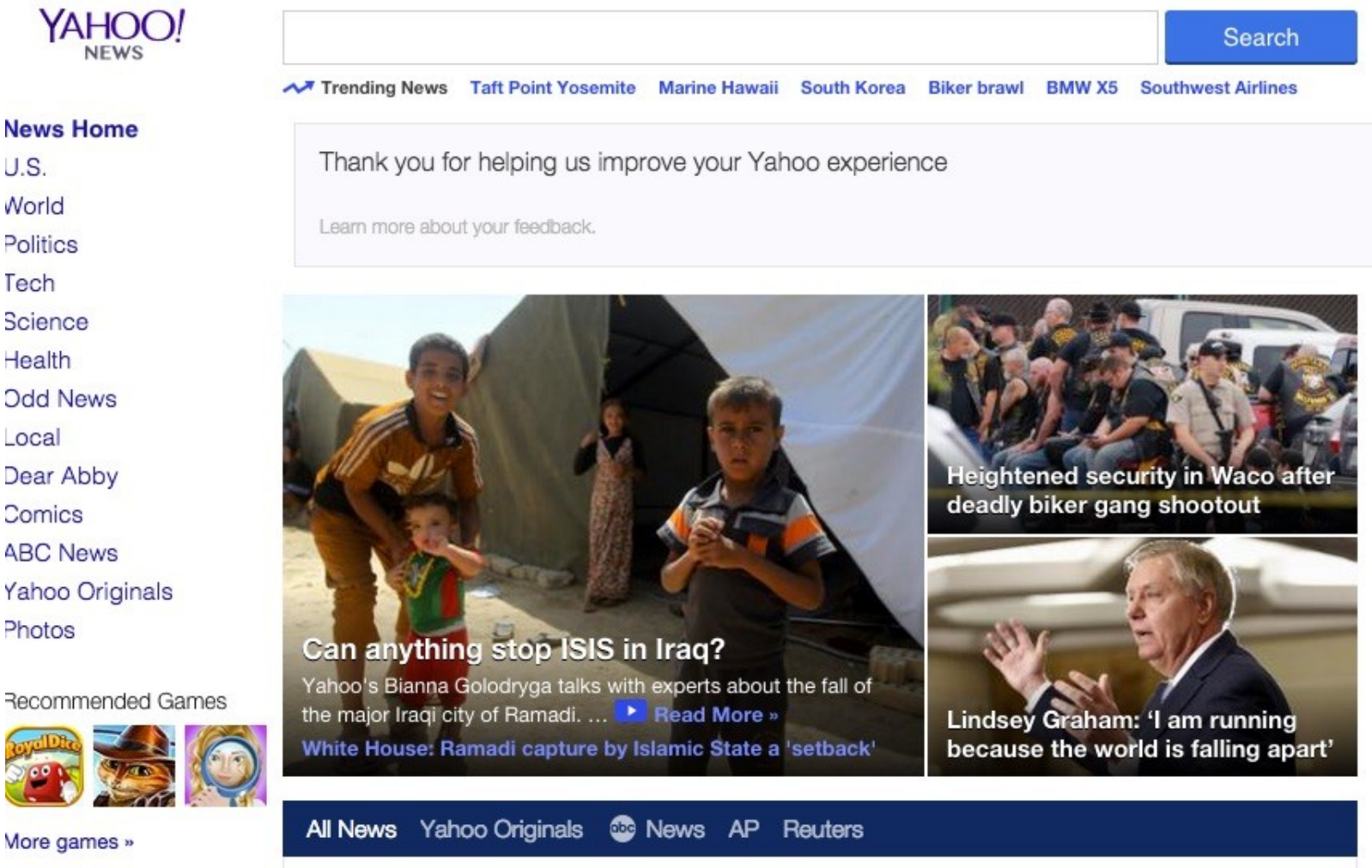
Interpreting Instructions FLAIRS 2012

- Show Video

Why RL

- Complex Dynamics
- Complex Workspace
- Stochastic sensing and actuation
- Human-in-the-loop Learning
- Cognitively motivated Learning
- Customization/Personalization

Customization



Ad Selection

Shop for Florists in Chennai on Google

Sponsored

Online Flowers Delivery
₹1,749
Ferns N Petals

Message In A Bottle with teg
₹349
Ferns N Petals

Classic Bunch - online flower ...
₹499
FlowerAura
Special offer

Online Flower Delivery
₹599
Ferns N Petals

Relish Of Heavenly Treat
₹1,399
Ferns N Petals

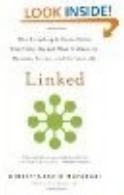
Florists In Chennai - Same Day Delivery Within 4 Hrs - floweraura.com
Ad www.floweraura.com/Online-Florist/Chennai
Online **Flowers** & Gifts Delivery @ Rs 399. Best Price, 100% Smile Guaranteed.
Delivery in 4 Hrs · Mid-Night Delivery · No Hidden Cost · Free Shipping · Flowers Starting @ Rs 399
Types: Cakes, Flowers, Gifts, Chocolate

Flowers Delivery in Chennai - Expess Delivery in 2-3 Hrs
Ad www.flowersnfruits.com/Flower_Delivery/Chennai 099300 06747
Order **Flowers** Now For Express Delivery within 2-3 hrs Anywhere in **Chennai**.

Recommendation

Really?

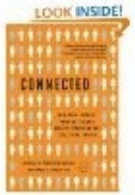
Customers Who Bought This Item Also Bought



Linked: How Everything Is Connected to... by Albert-Laszlo Barabasi

★★★★☆ (116)

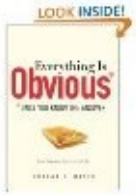
$10.38



Connected: The Surprising Power of Our... by Nicholas A. Christakis

★★★★☆ (42)

$10.87

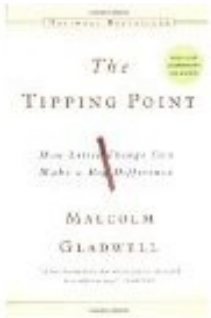


Everything Is Obvious: *Once You Know the Answer by Duncan J. Watts

★★★★☆ (54)

$17.16

Your Recent History Also Bought



The Tipping Point: How Little...

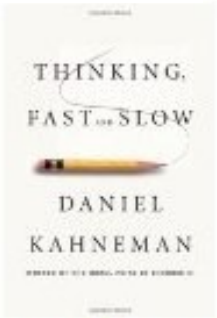
› Malcolm Gladwell

★★★★☆ (1,262)

Paperback

$9.59

Fix this recommendation



Thinking, Fast and Slow

› Daniel Kahneman

★★★★☆ (91)

Hardcover

$16.65

Fix this recommendation

Adapt: Why Success Always Starts...

› Tim Harford

★★★★½ (31)

Hardcover

$17.82

Fix this recommendation

Comment Recommendation

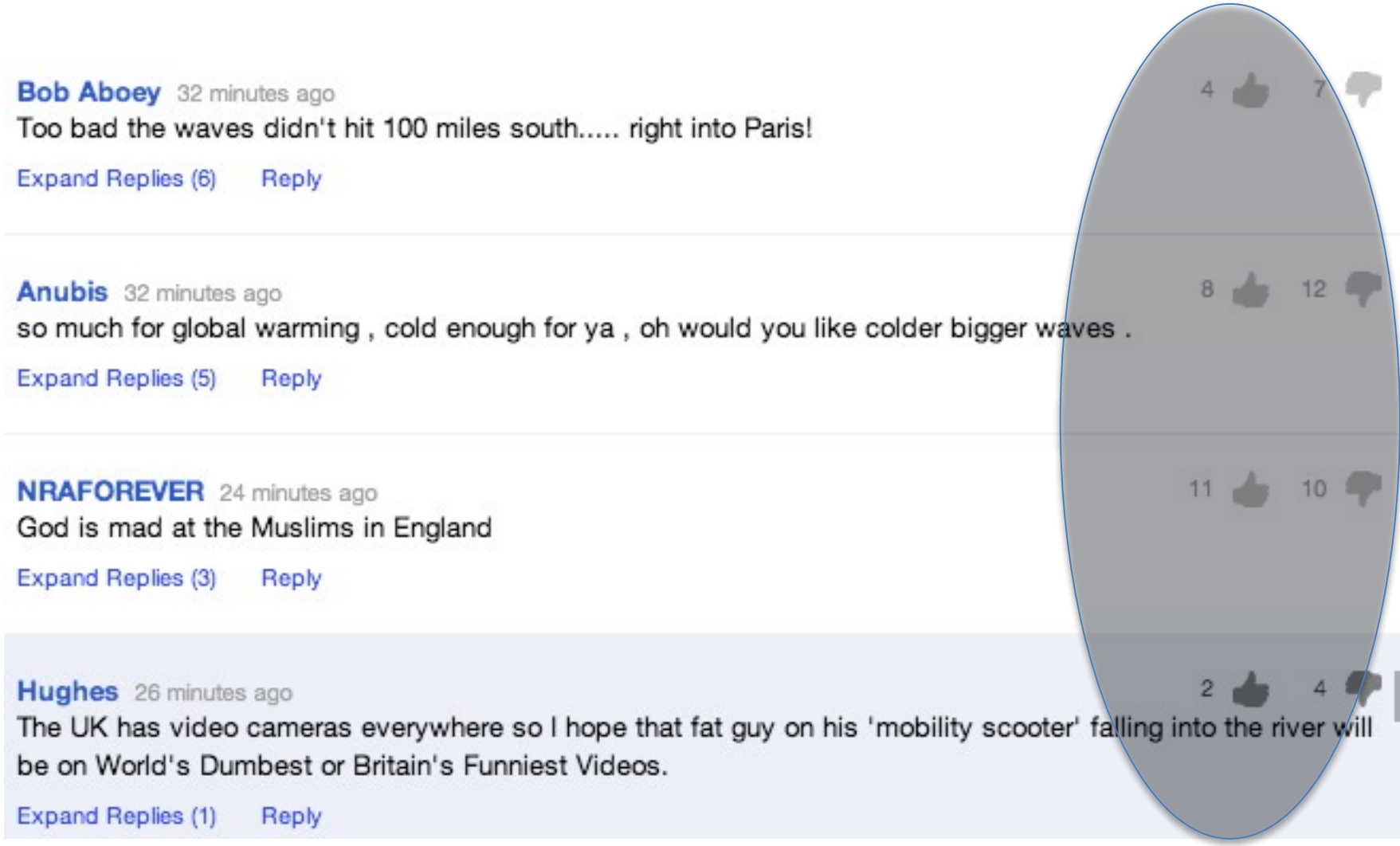
Bob Aboey 32 minutes ago 4 7
Too bad the waves didn't hit 100 miles south..... right into Paris!
Expand Replies (6) Reply

Anubis 32 minutes ago 8 12
so much for global warming , cold enough for ya , oh would you like colder bigger waves .
Expand Replies (5) Reply

NRAFOREVER 24 minutes ago 11 10
God is mad at the Muslims in England
Expand Replies (3) Reply

Hughes 26 minutes ago 2 4
The UK has video cameras everywhere so I hope that fat guy on his 'mobility scooter' falling into the river will be on World's Dumbest or Britain's Funniest Videos.
Expand Replies (1) Reply

Comment Recommendation



Bob Aboey 32 minutes ago
Too bad the waves didn't hit 100 miles south..... right into Paris!
Expand Replies (6) Reply
4 7

Anubis 32 minutes ago
so much for global warming , cold enough for ya , oh would you like colder bigger waves .
Expand Replies (5) Reply
8 12

NRAFOREVER 24 minutes ago
God is mad at the Muslims in England
Expand Replies (3) Reply
11 10

Hughes 26 minutes ago
The UK has video cameras everywhere so I hope that fat guy on his 'mobility scooter' falling into the river will be on World's Dumbest or Britain's Funniest Videos.
Expand Replies (1) Reply
2 4

Why RL

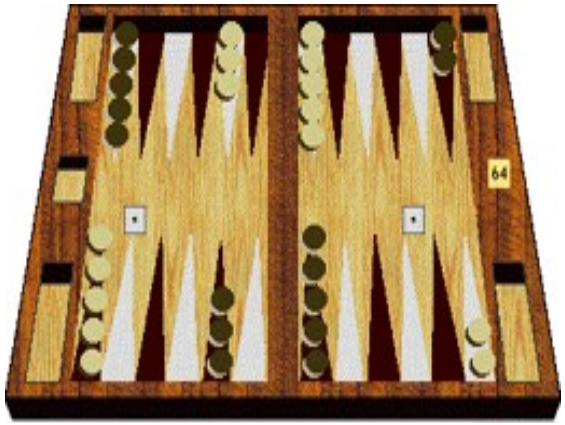
- Complex Dynamics
- Complex Workspace
- Stochastic sensing and actuation
- Human-in-the-loop Learning
- Cognitively motivated Learning
- Customization/Personalization
- Going beyond human knowledge
  - Learn through Self-Play

Human Level Backgammon player

TD-Gammon (Tesauro 92, 94, 95)

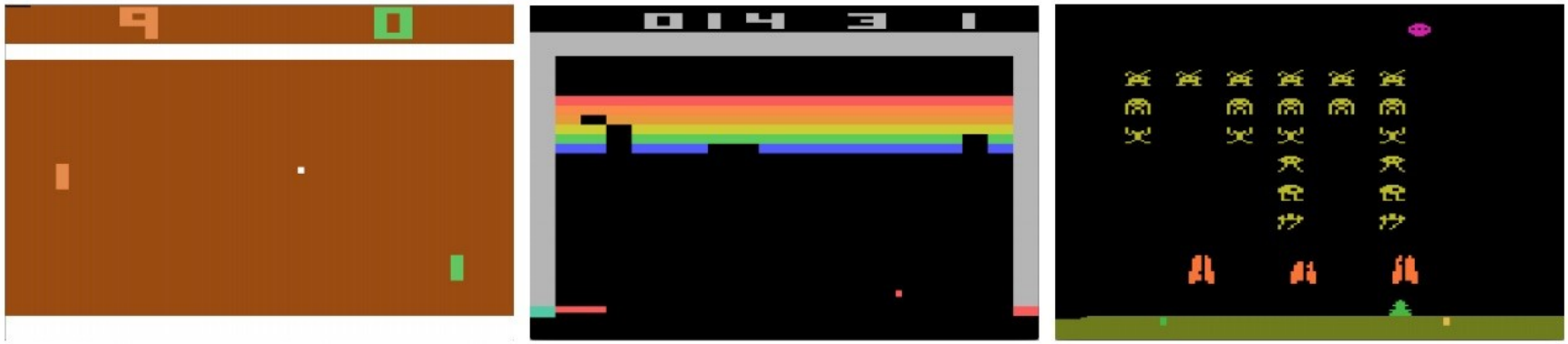
Beat the best human player in 1995

Learnt completely by *self play*



New moves not recorded by humans in centuries of play

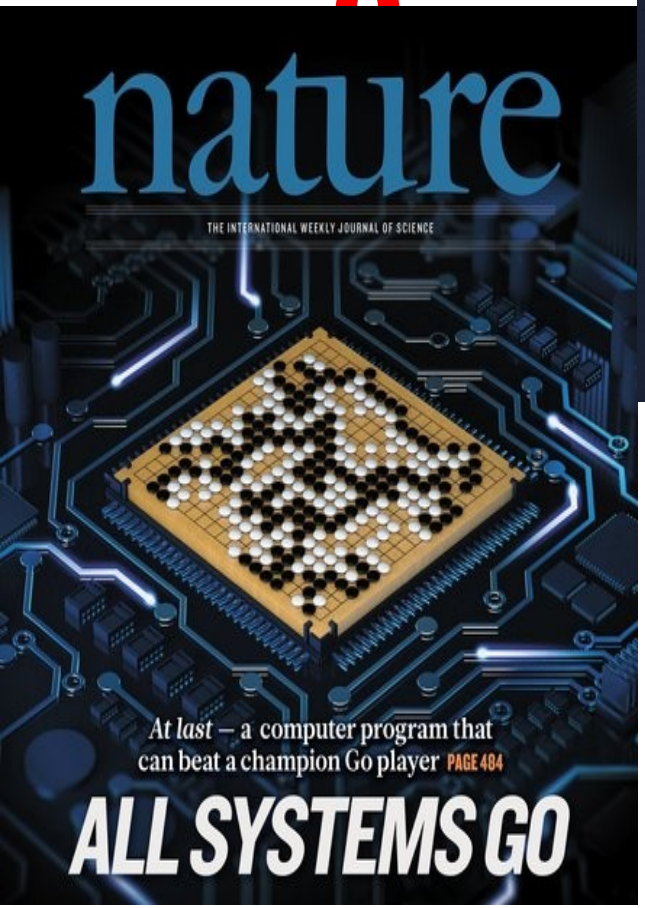
Game Playing – Arcade Games

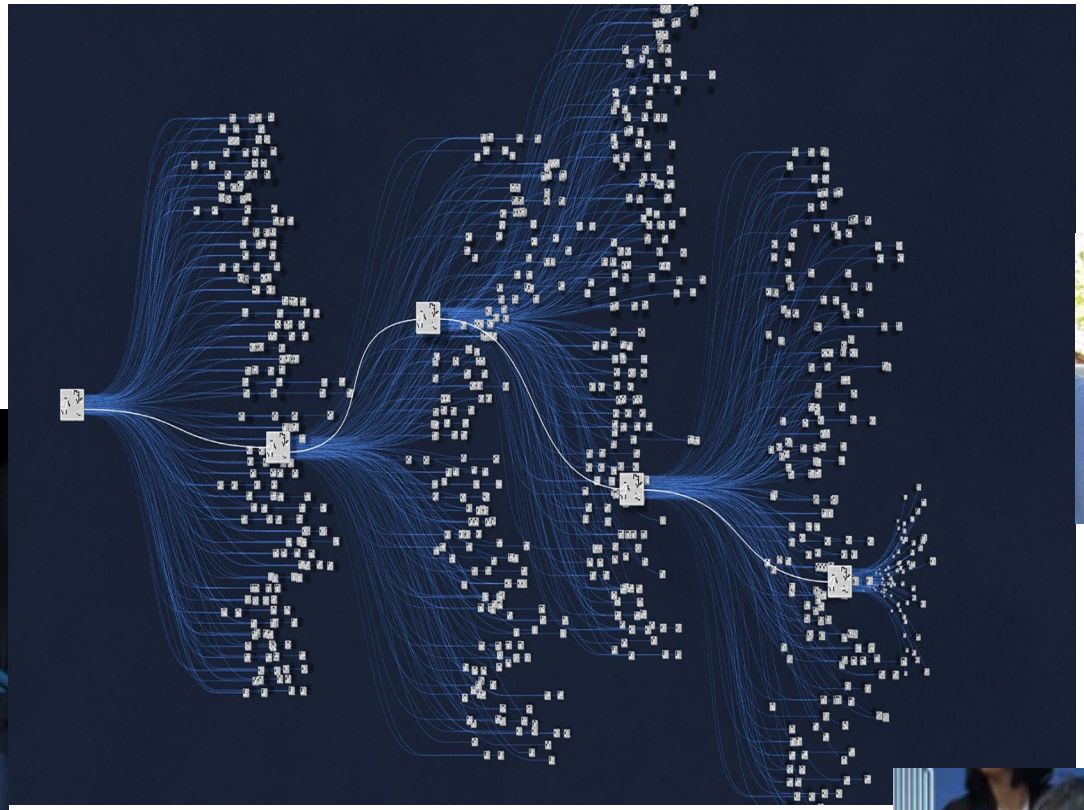


- Learnt to play from video input!
- Learnt from scratch
- Used a complex *neural network!*
- Considered one of the hardest learning problems solved by a computer!

Deepmind's Atari Player

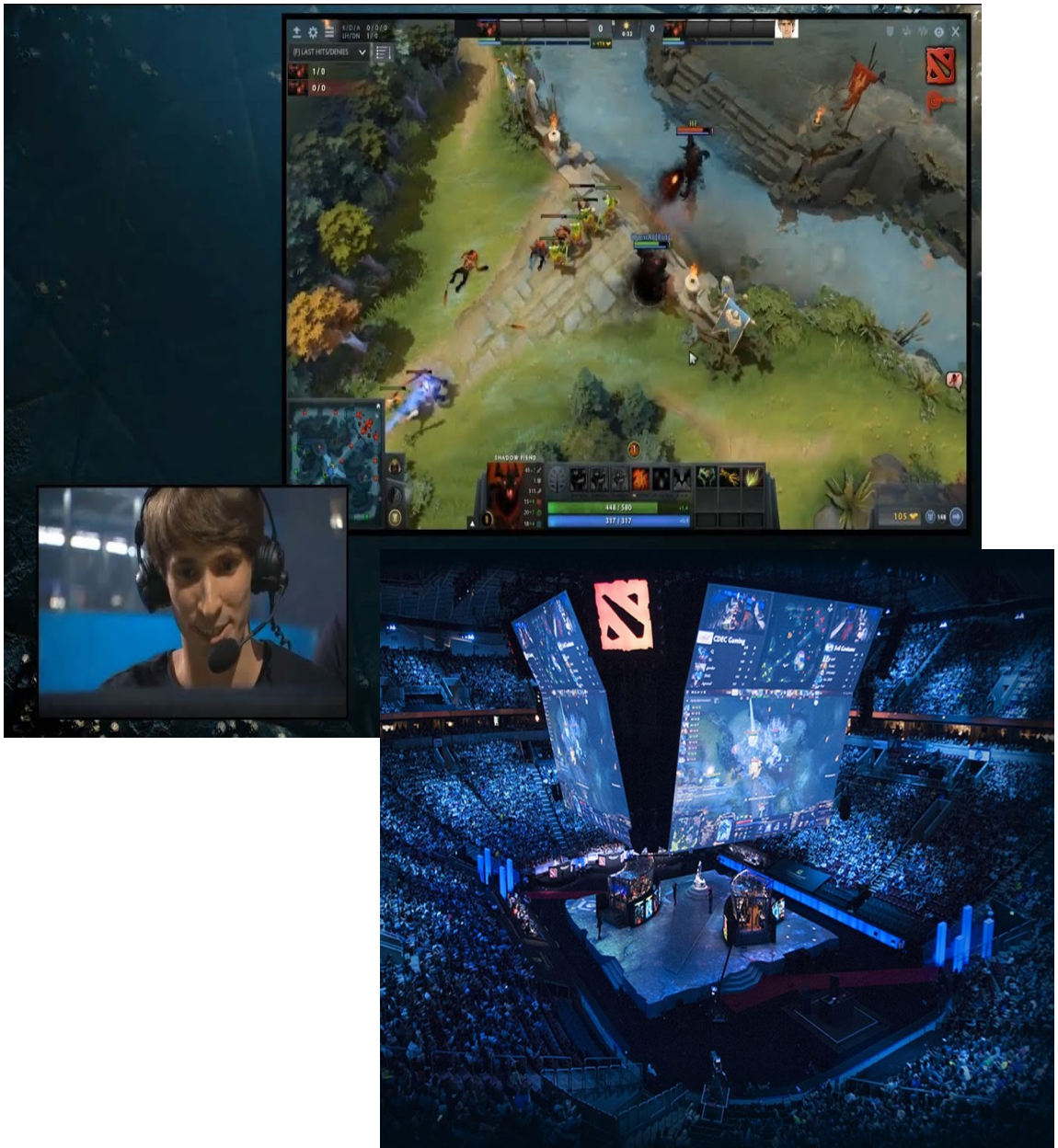
AlphaG
o





- Branching factor : Go 250 vs 35 Chess
- AlphaGo Master defeated the 18-time World Champion* 4-1.

Defence of the Ancients 2 (DoTA 2)



- The AI bot won 1v1 matches against top players in the world at the International DoTA Championships.
- 100 different heroes. 100 different items. Many different tactics. **Much more complex than board games.**
- Trained for 2 weeks by using just self-play.

- The full game is played 5v5. Multi-agent coordination required. Still being developed.

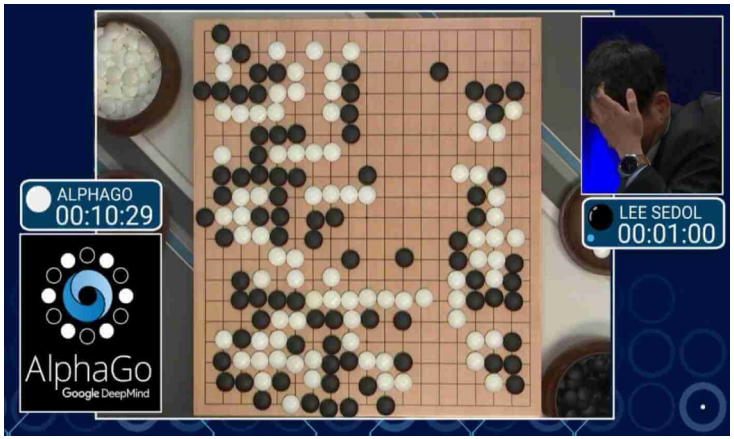
AlphaZero: A general AI Algorithm

- A general AI agent; Not limited to Go. Superhuman Performance on Chess, Shogi and Go.
- *No human data:* Trained from Scratch RL by playing against itself.
- *No human features:* Only the raw positions from the board are provided to the agent.
- *Simpler Search:* No randomized Monte Carlo Rollouts. Use a Neural Network to evaluate.
- Beat AlphaGo Lee by **100 – 0**
- Beat Stockfish and Elmo on Chess and Shogi.

AlphaGo

Move 37

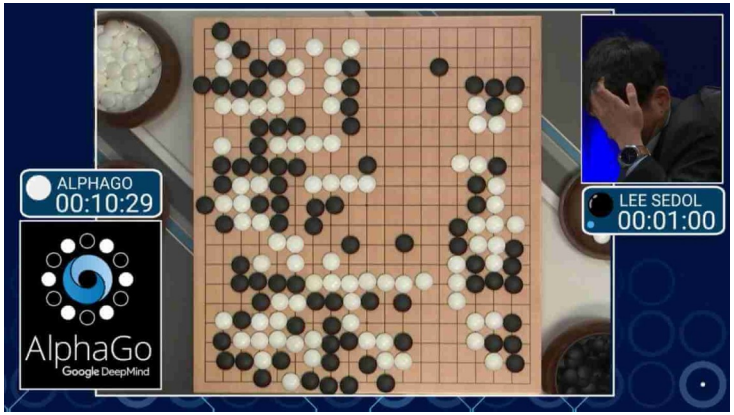
With the 37th move in the match's second game, AlphaGo landed a surprise on the right-hand side of the 19-by-19 board that flummoxed even the world's best Go players, including Lee Sedol. "That's a very strange move," said one commentator, himself a nine dan Go player, the highest rank there is. "I thought it was a mistake," said the other. Lee Sedol, after leaving the match room, took nearly fifteen minutes to formulate a response. Fan Gui—the three-time



AlphaGo

Move 37

With the 37th move in the match's second game, AlphaGo landed a surprise on the right-hand side of the 19-by-19 board that flummoxed even the world's best Go players, including Lee Sedol. "That's a very strange move," said one commentator, himself a nine dan Go player, the highest rank there is. "I thought it was a mistake," said the other. Lee Sedol, after leaving the match room, took nearly fifteen minutes to formulate a response. Fan Gui—the three-time

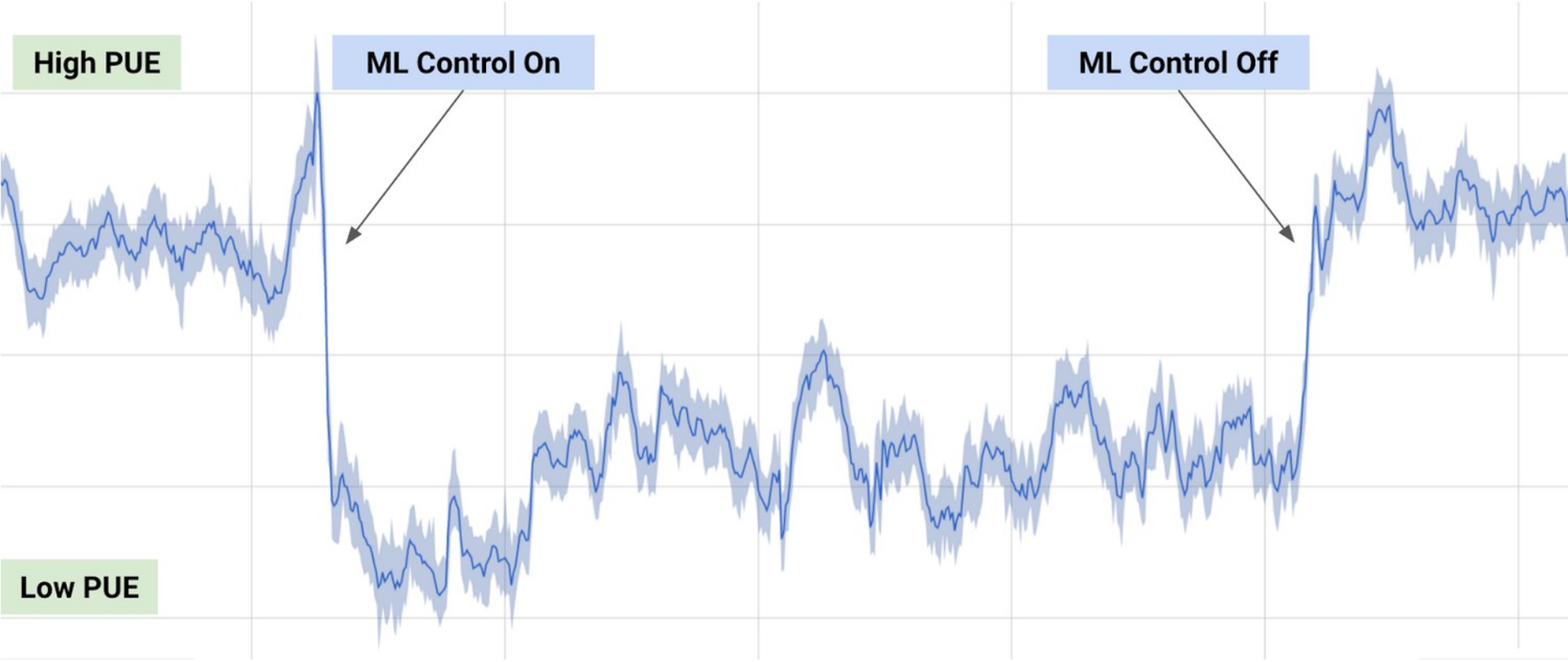


Why RL

- Complex Dynamics
- Complex Workspace
- Stochastic sensing and actuation
- Human-in-the-loop Learning
- Cognitively motivated Learning
- Customization/Personalization
- Going beyond human knowledge
  - Improving heuristics

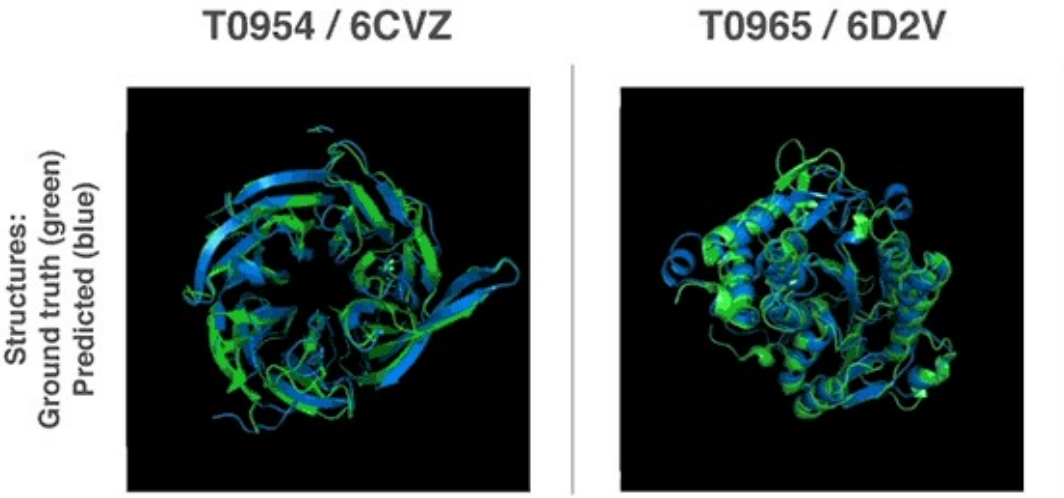
Power Management

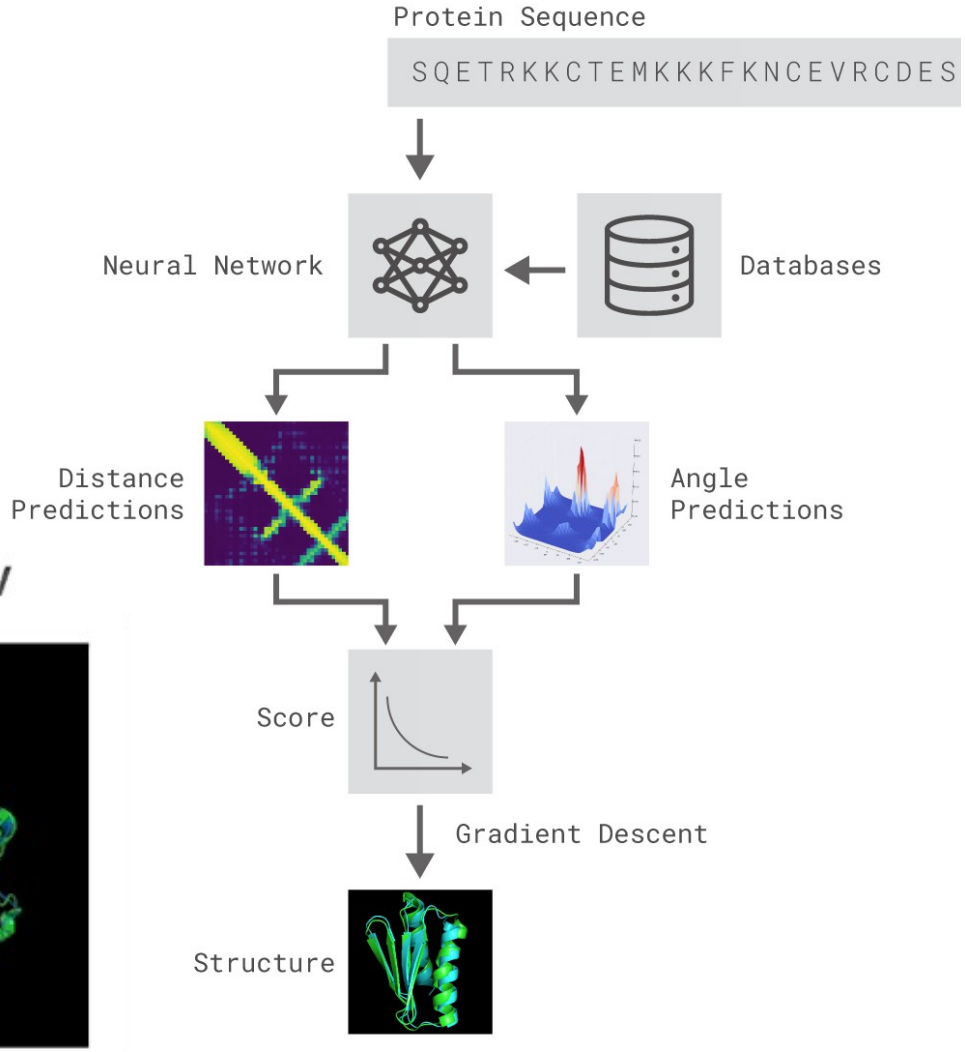
- Reduced Google Data Centre cooling bill by 40%



AlphaFold

- Protein Folding: One of the hardest problems in biology
- An AI agent achieved 25% improvement over best human effort

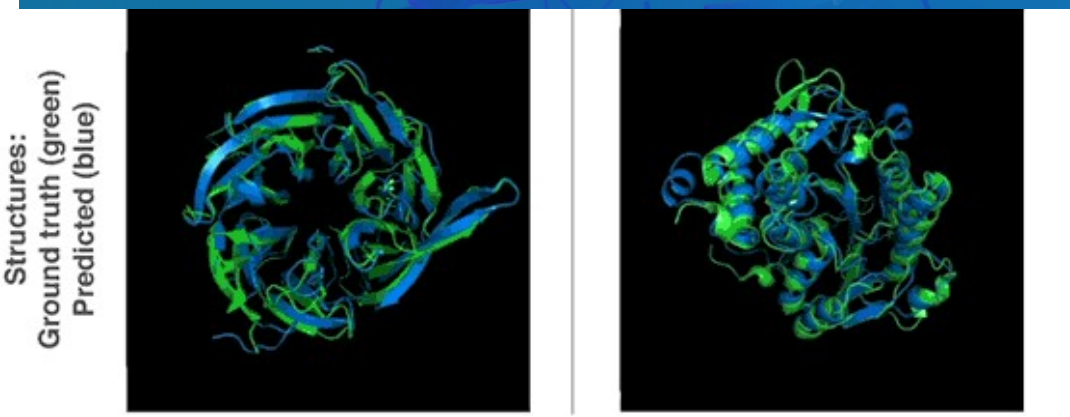


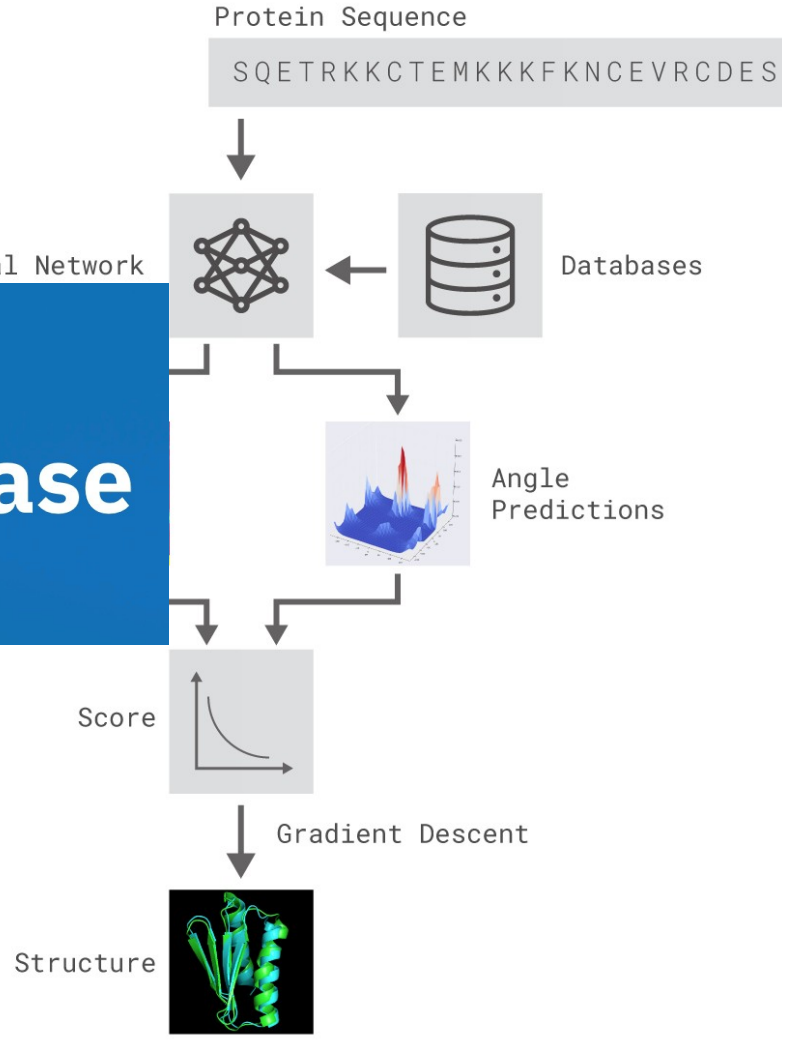


AlphaFold 2.0

- Protein Folding: One of the hardest problems in biology

SOLVED?

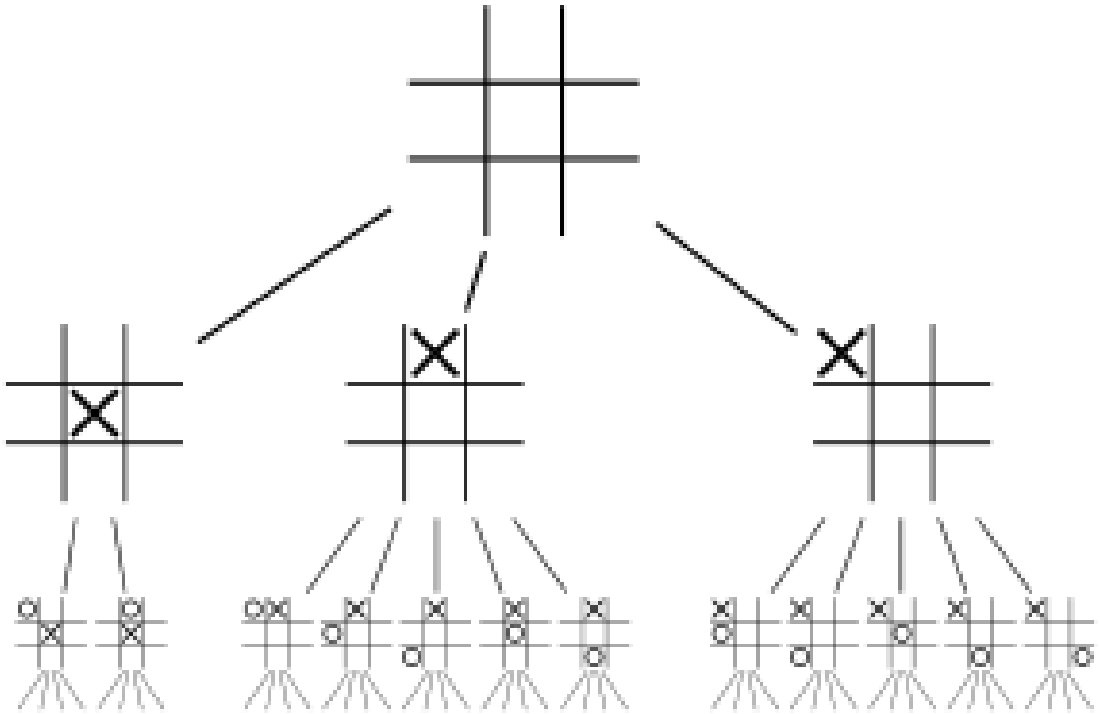




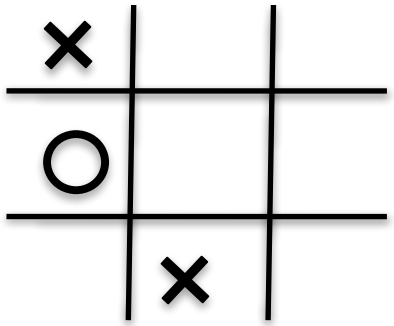
Other Applications

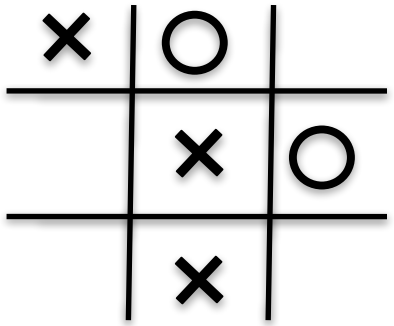
- Optimal Control
  - Robot Navigation
  - Chemical Plants
- Combinatorial Optimization
  - Elevator Dispatching
  - VLSI placement and routing
  - Job-shop scheduling
  - Routing algorithms
  - Call admission control
- More
  - Intelligent Tutoring Systems
- Computational Neuroscience
  - Primary mechanism of learning
- Psychology
  - Behavioral and operant conditioning
  - Decision making
- Operations Research
  - Approximate Dynamic Programming
- More
  - Dialogue systems

Tic-Tac-Toe

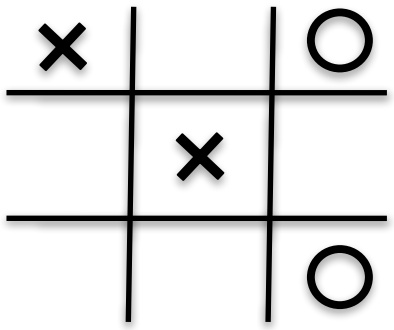


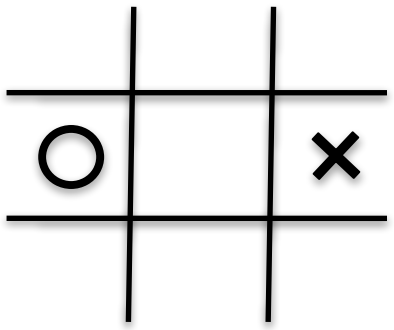
Supervised Learning



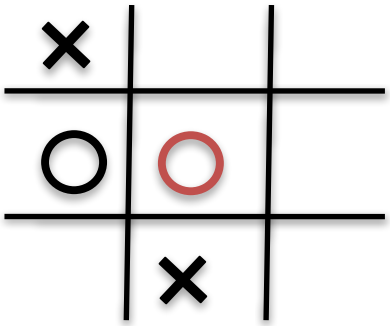


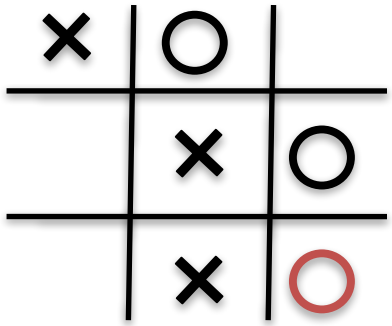
Current Positions



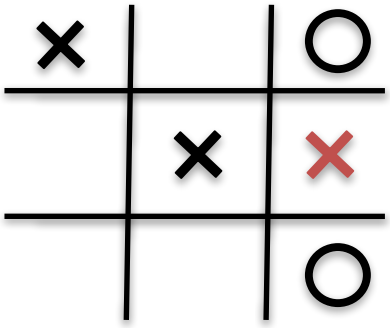


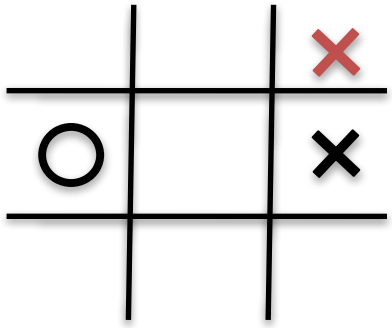
Supervised Learning





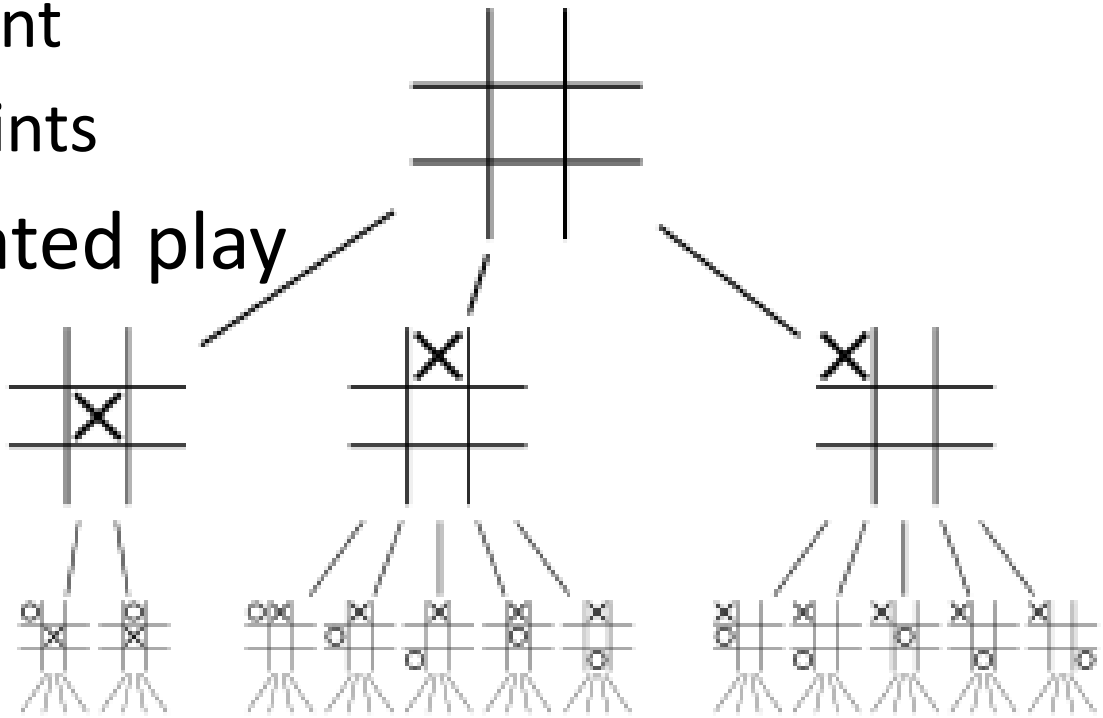
Expert Moves





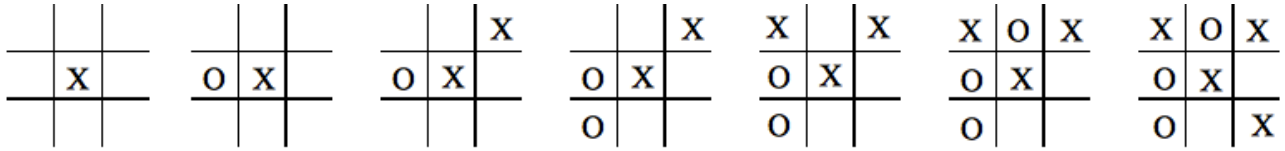
Reinforcement Learning

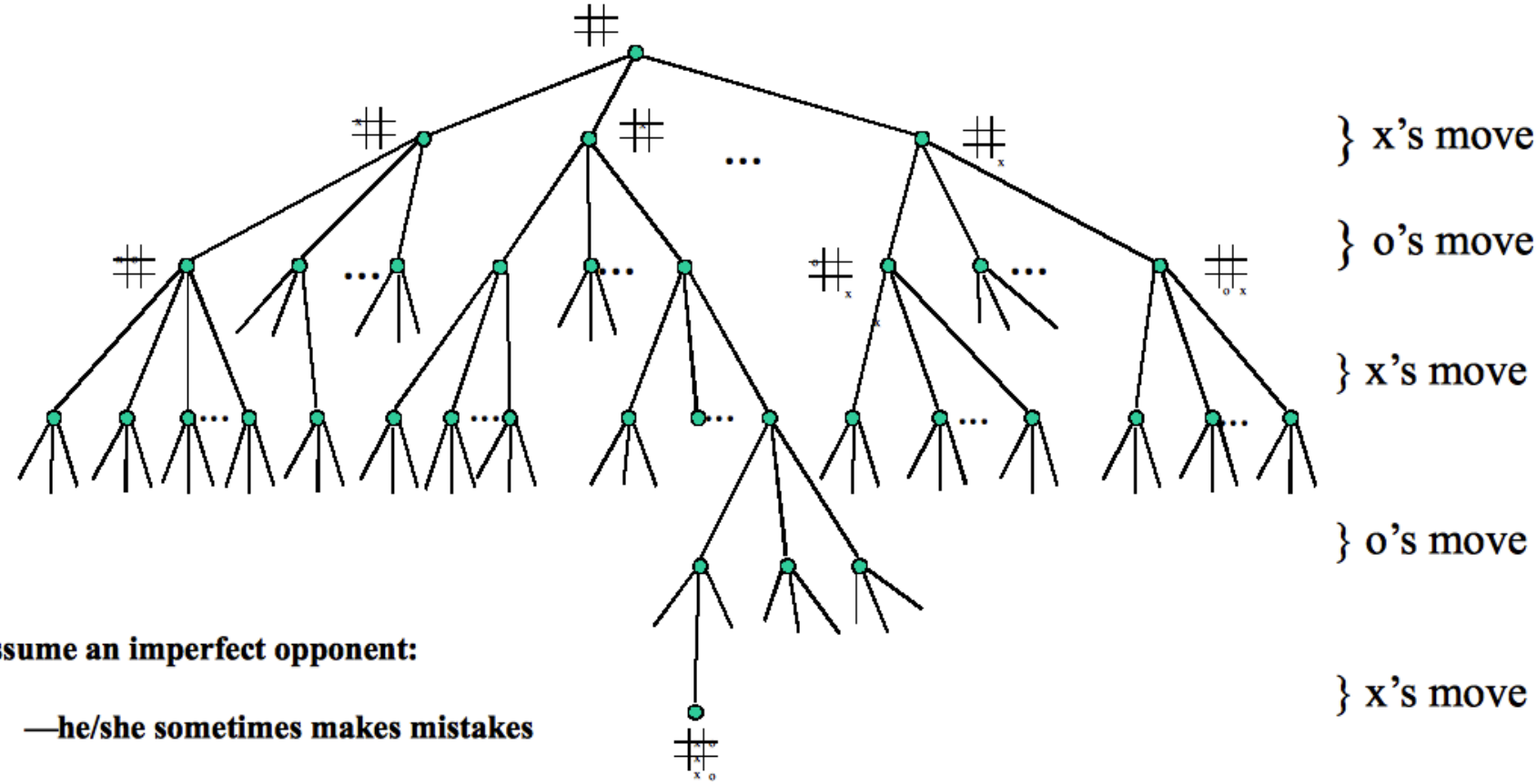
- Learn from evaluation
  - Win gives 1 point
  - Loss gives -1 point
  - Draw gives 0 points
- Learn from repeated play



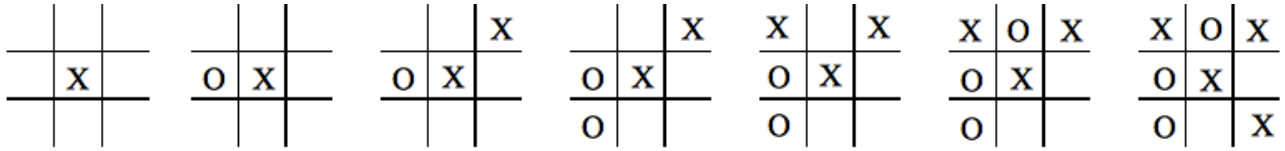
MENACE Michie and Chambers ’60

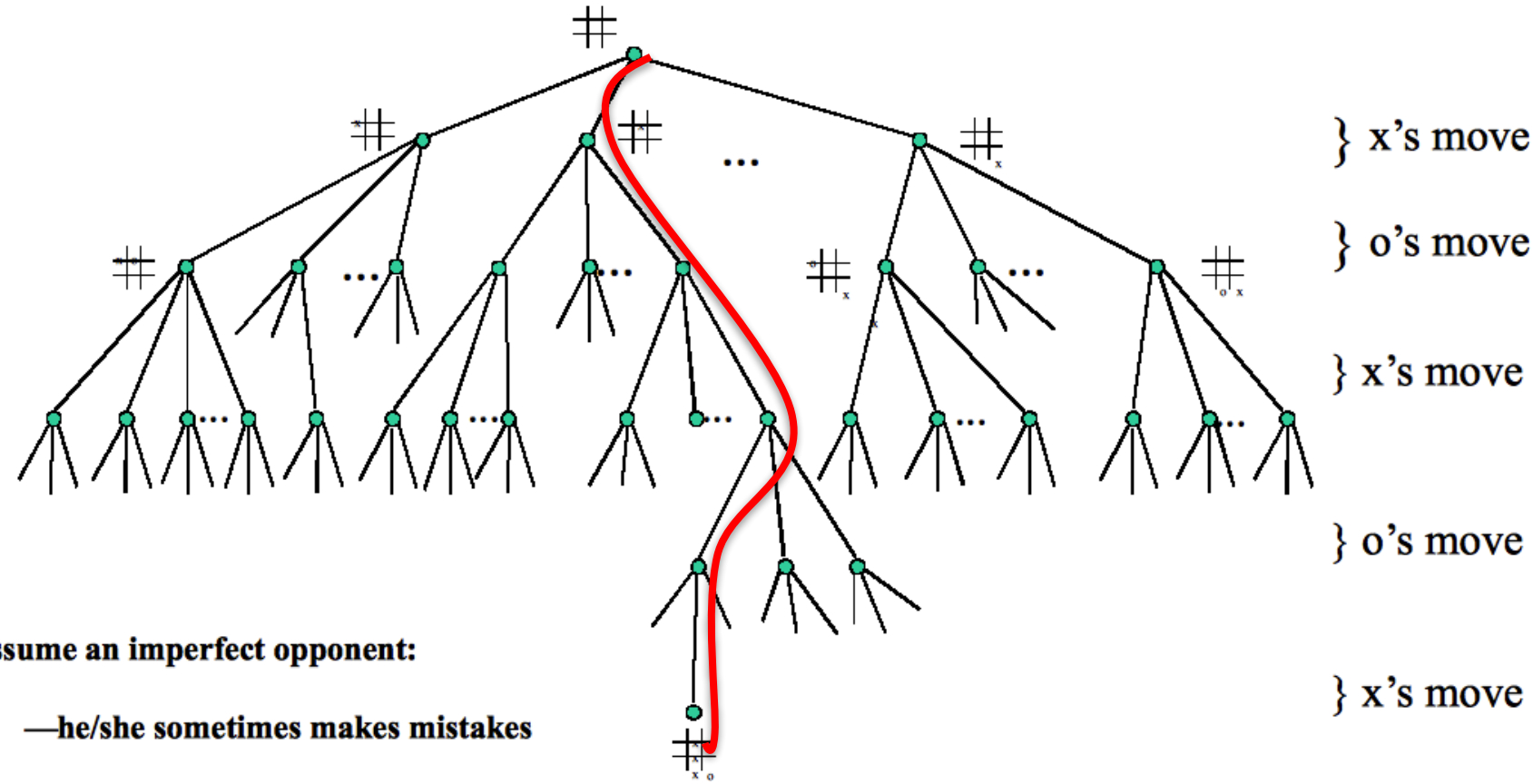
More Tic-Tac-Toe





More Tic-Tac-Toe

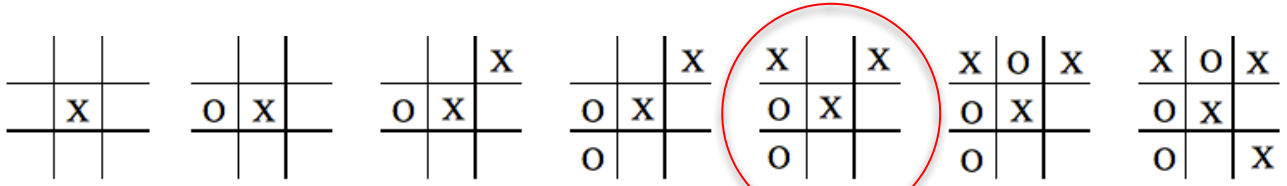


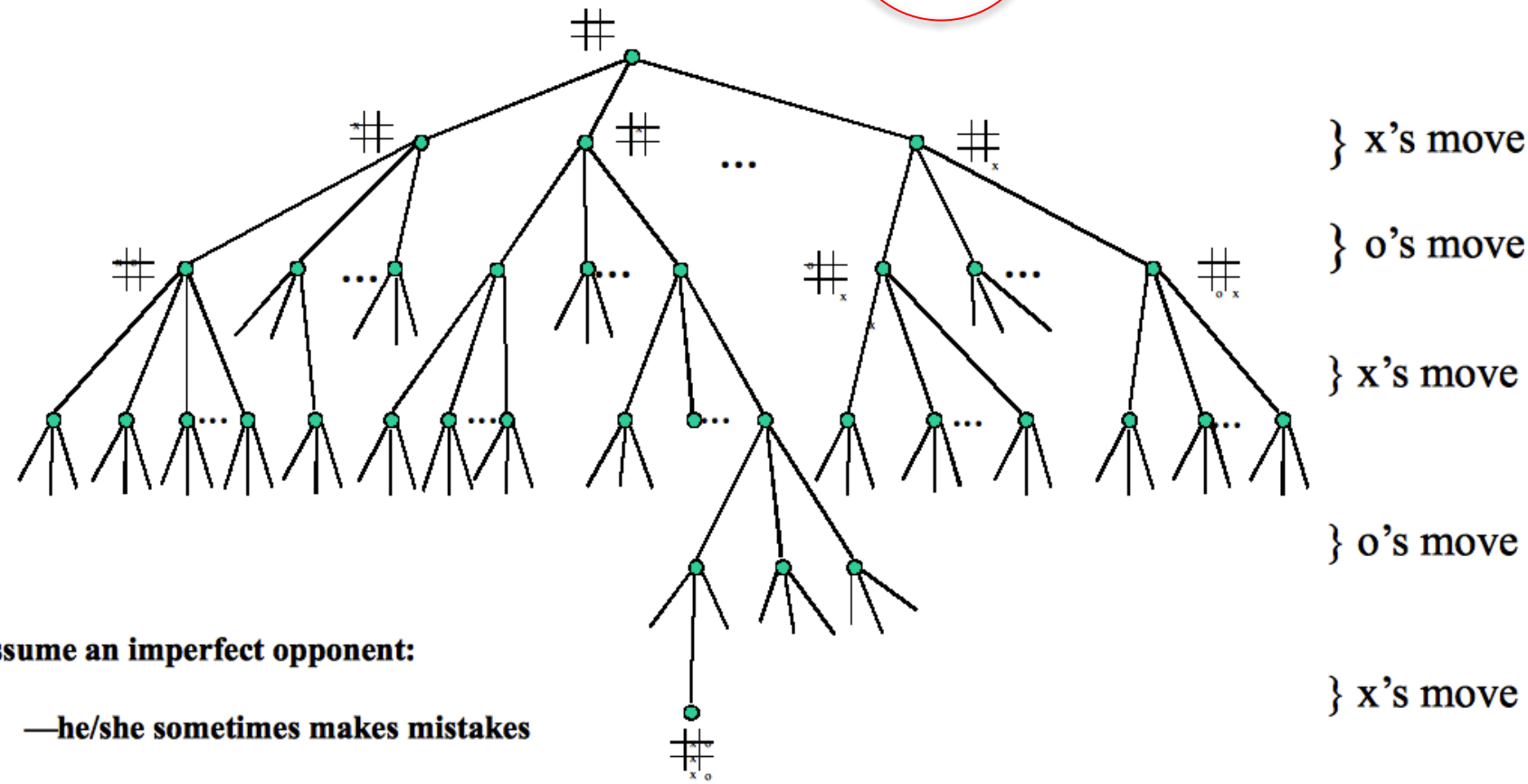


Assume an imperfect opponent:

—he/she sometimes makes mistakes

More Tic-Tac-Toe





Assume an imperfect opponent:

—he/she sometimes makes mistakes

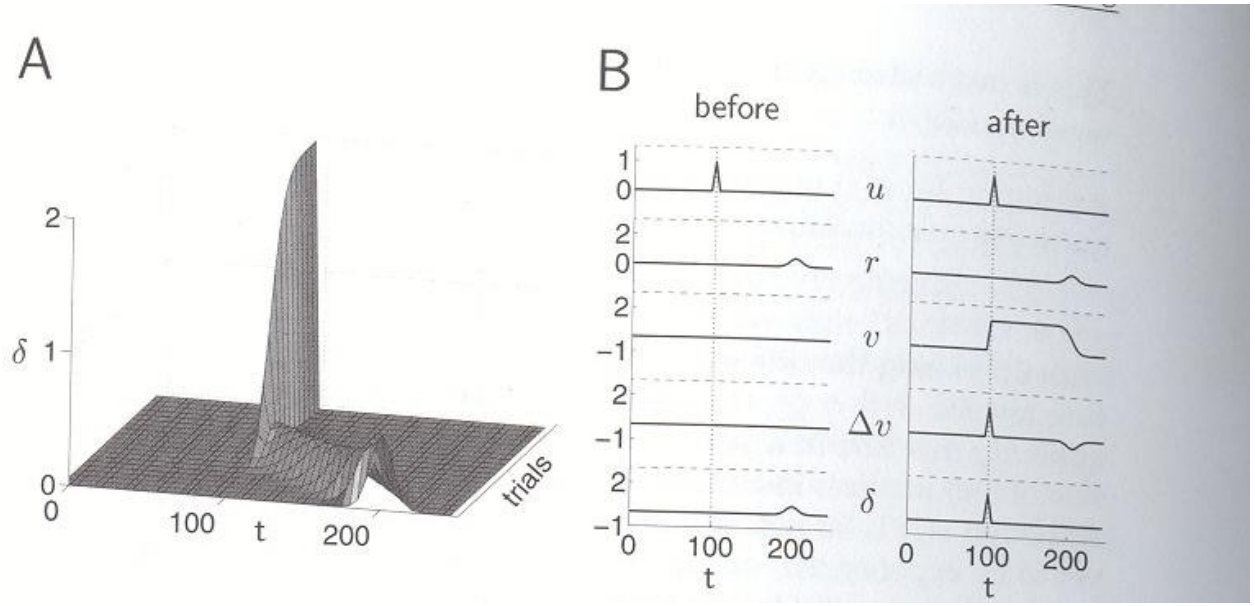
Temporal Difference

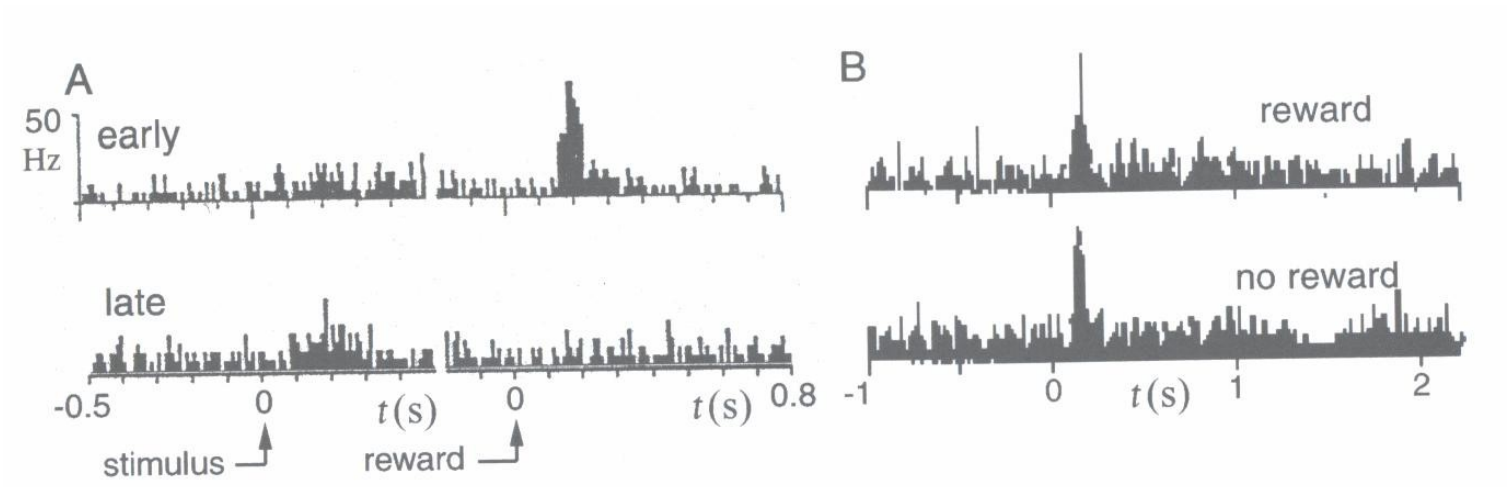
Barto, Sutton, Anderson '83

- Simple rule to explain complex behaviors
- Intuition: Prediction of outcome at time $t+1$ is better than the prediction at time t. Hence use the later prediction to adjust the earlier prediction.
- Has also had profound impact in behavioral psychology and neuroscience!



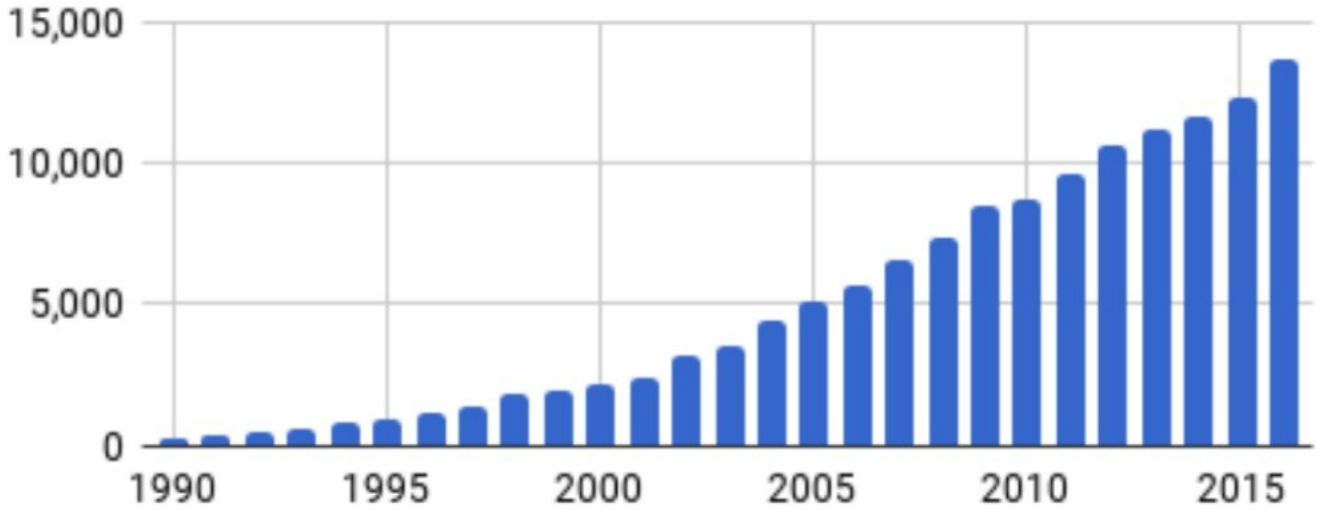
TD in the Brain





What's Next

- Deep Reinforcement Learning has revived excitement in the community



From: Henderson et al. "Deep Reinforcement Learning that Matters". Arxiv 2017. https://arxiv.org/pdf/1709.06560.pdf

What's Next

- Deep Reinforcement Learning has revived excitement in the community
- But many fundamental questions still to be addressed
- Goal: Omnivorous learning – consume any information to learn
  - Closer to how humans learn

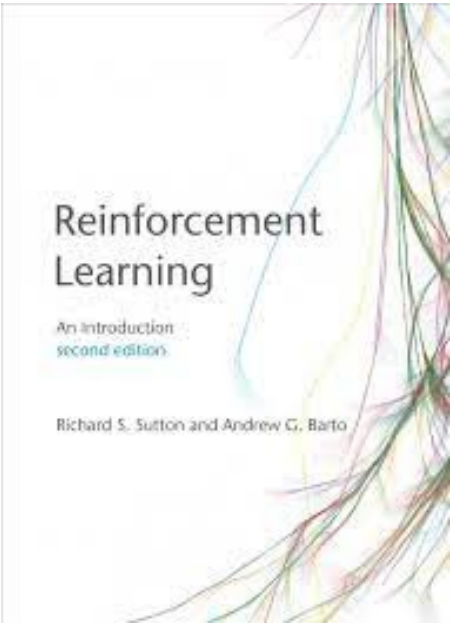
Administrivia

Textbook: Sutton, R. S., and Barto, A. G. "Reinforcement Learning: An Introduction", 2nd Edition. MIT Press.

Get Moodle Access

Marking Scheme

Written Assignments	24%
Programming Assignments	36%
Written Exam 1	14%
Written Exam 2	14%
Tutorials	12%
Total	100%



TAs:

Vihaan Akshaay Rajendiran <me17b171@smail.iitm.ac.in>

G. Prashant <bs17b011@smail.iitm.ac.in>