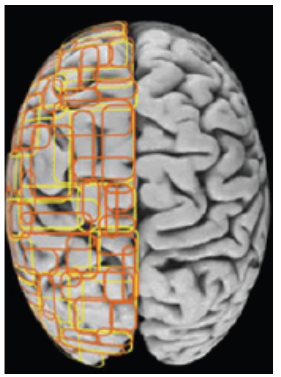
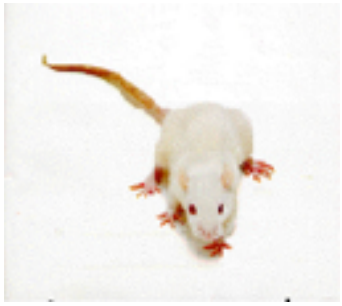


Cuando el perro de Pavlov se robotizó: aprendizaje por refuerzos en psicología, robótica, neurociencias y juegos de Atari

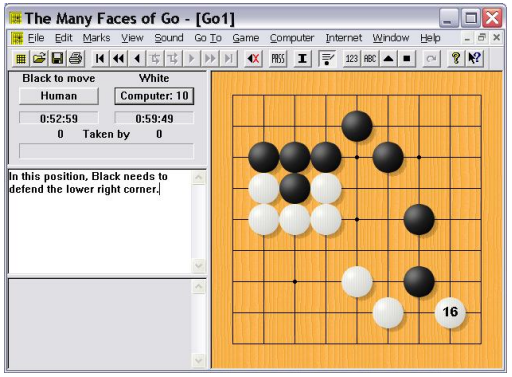
Carlos Greg Diuk

Princeton Neuroscience Institute

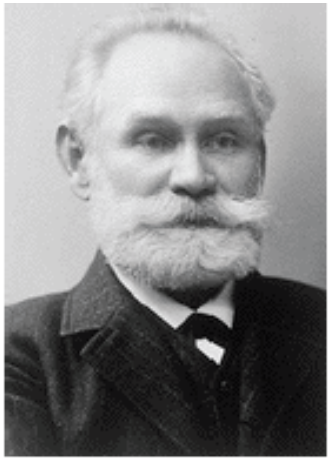
Charla de borrachos – 15 de junio de 2012



Aprendizaje y toma de decisiones



Aprendizaje (y refuerzos)



Ivan Pavlov
(1849-1936)

- Primeras teorías “modernas”:
  - Condicionamiento clásico o Pavloviano.

Visión ultra-limitada del aprendizaje, pero funciona.
Demuestra que los animales pueden aprender relaciones *arbitrarias* entre estímulo->respuesta.

Dibujitos: Yael Niv

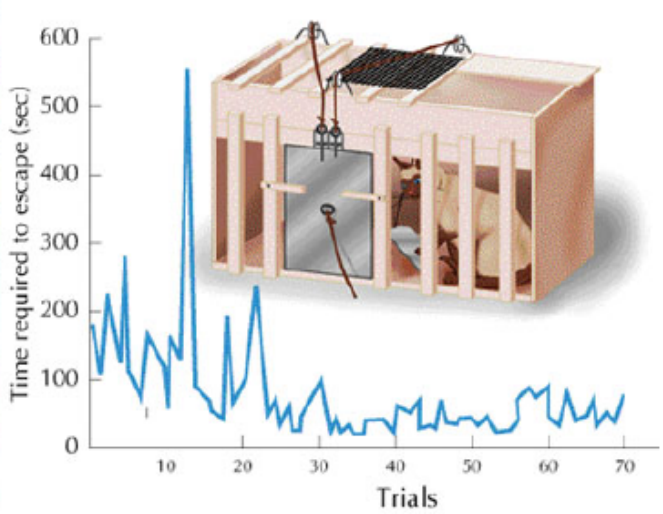
Pero el perro de Pavlov no tomó decisiones

- Hasta ahora no tuvo que hacer nada, no tomó decisiones, sólo salivó porque la campana le recordó el churrasco.
- Agreguemos “control”: no solamente estímulo-respuesta
- Las acciones que tomamos tienen consecuencias.

Condicionamiento Instrumental/ Operacional

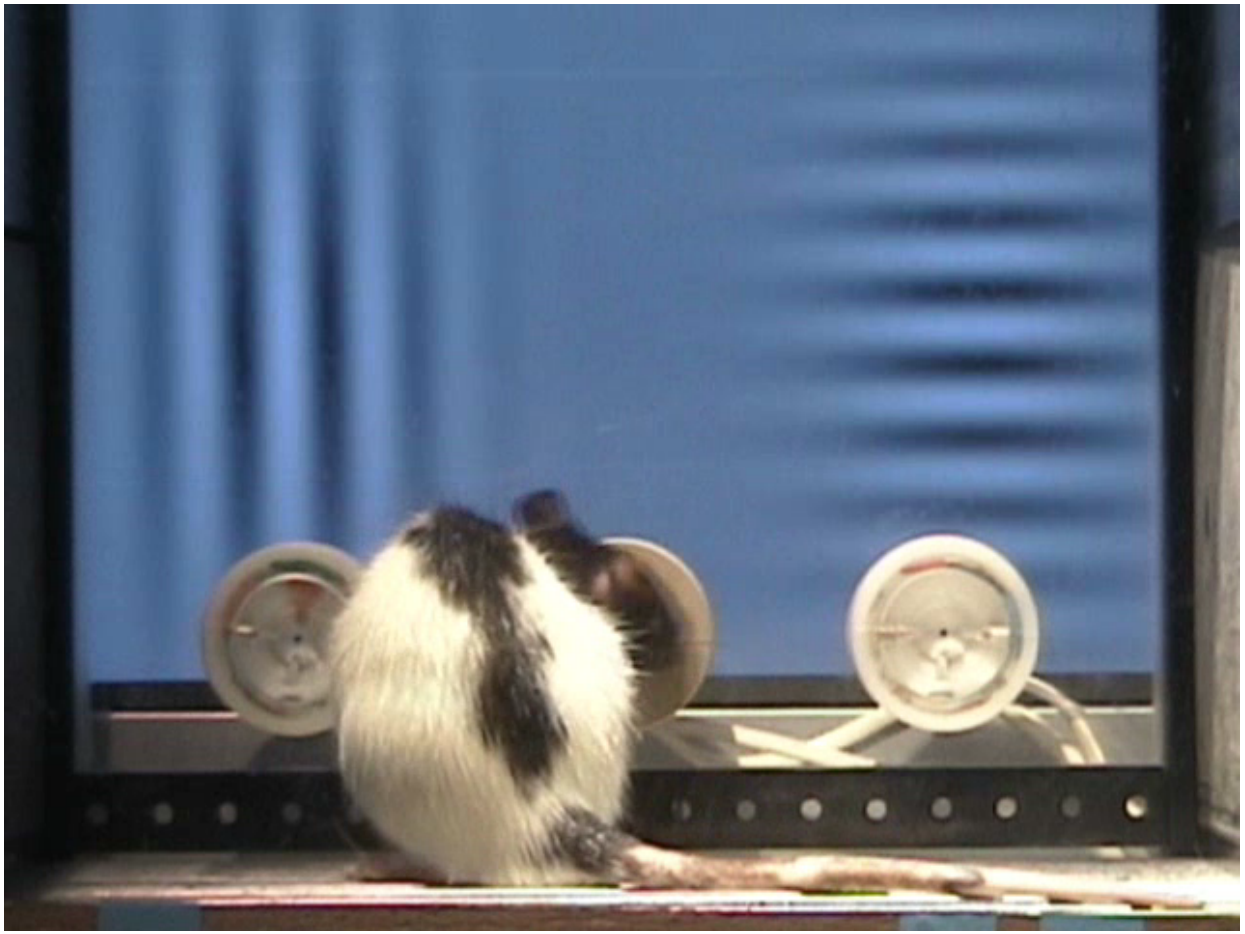
- Thorndike experimentó con gatos hambrientos tratando de escapar de una jaula.
- Midió el "tiempo hasta escapar" como métrica de aprendizaje.
- "Curva de aprendizaje"

Edward Thorndike
(1874-1949)



Diapositiva: Yael Niv

Aprendizaje por Condicionamiento



Cuál es la regla?

Crédito: Reinagel lab, UCSD

Condicionamiento Instrumental

- La lección importante a extraer:

Los animales no sólo pueden aprender relaciones estímulo-respuesta arbitrarias, sino también **comportamientos** arbitrarios en base a dichos estímulos.

Crédito: Björn Brembs, FU Berlin

Aprendiendo con refuerzos

- Idea clave de Rich Sutton (1980s):

  Para desarrollar un sistema *inteligente*, el sistema tiene que *desear* algo.

- Sutton y Barto desarrollan AR computacional

Sistemas que aprenden a fuerza de desear lograr un objetivo

Aprendizaje por refuerzos

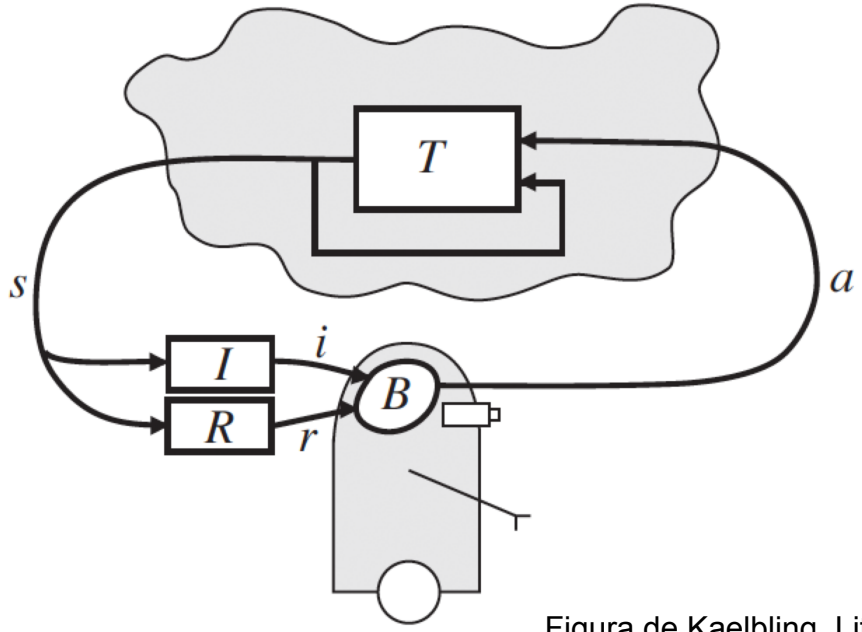
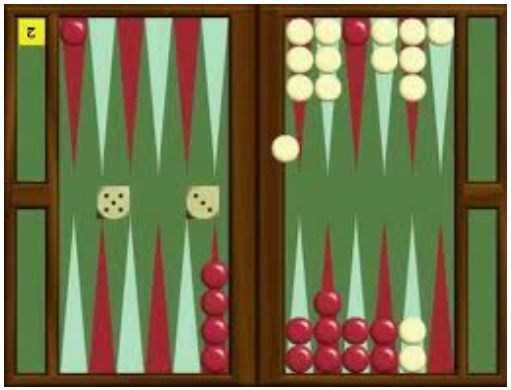


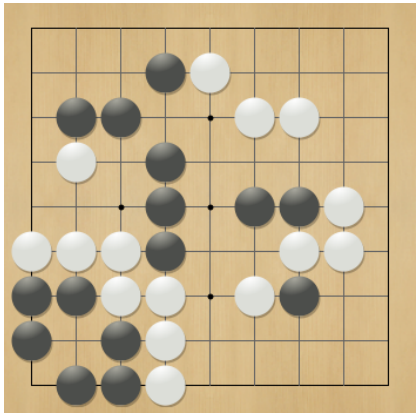
Figura de Kaelbling, Littman & Moore 96

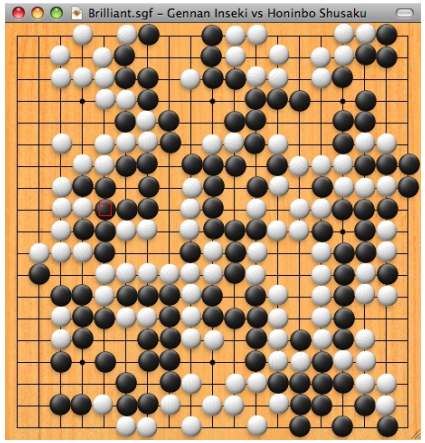
- Problema: transiciones entre estados y refuerzos son desconocidos.

Algunos éxitos

- 1995: jugador de backgammon basado en AR les empata a campeones mundiales.
- 2006-2011: jugadores de Go basado en AR alcanzan nivel de Grand Master en 9x9 y le ganan a otros programas en 19x19.







Aplicaciones

- Muchas aplicaciones en robótica:

- Juegos para X-Box de Microsoft Research que se adaptan al usuario.

El problema de qué comer

V(Locro)

$V \leftarrow V + \alpha (R - V)$

Prediction Error (>0)

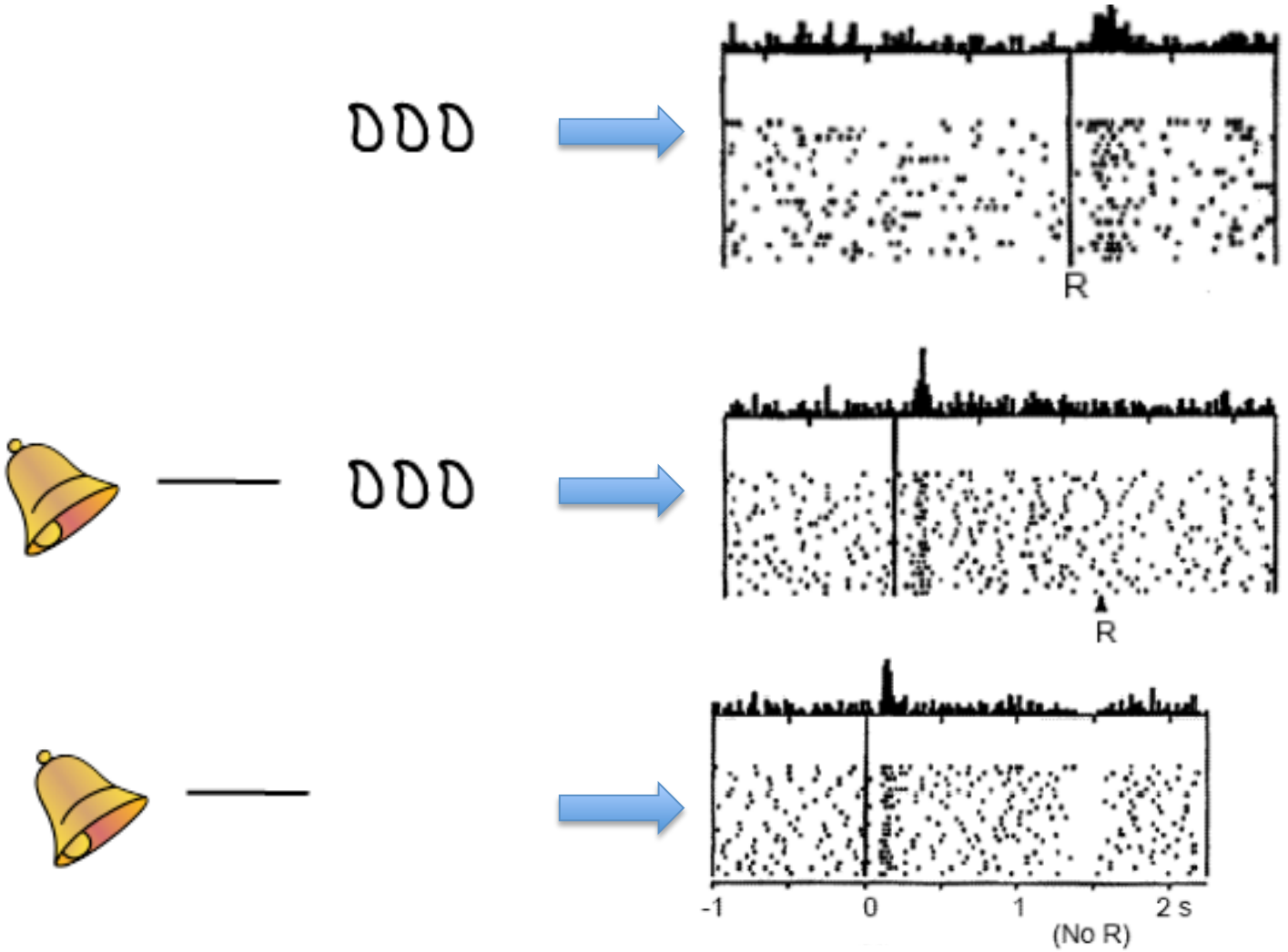
V(Humita)

$V \leftarrow V + \alpha (R - V)$

Prediction Error (<0)

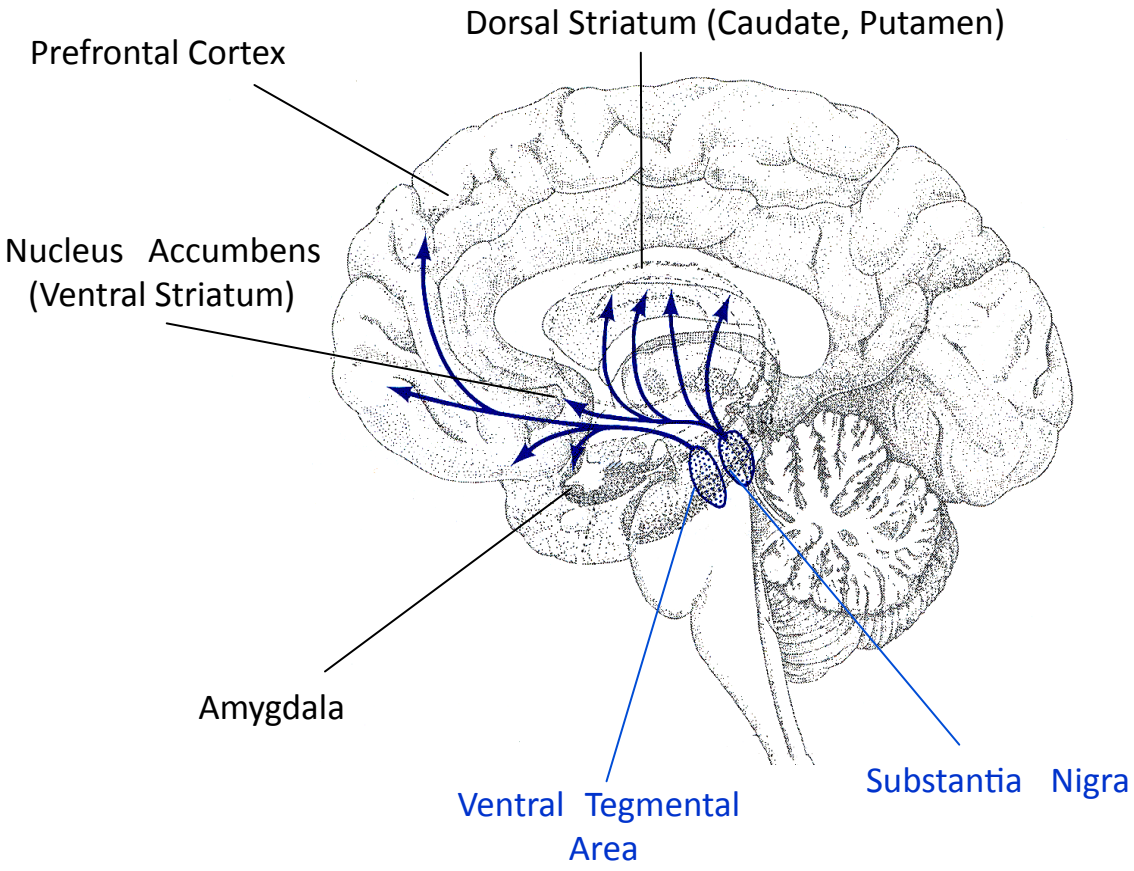
Algoritmo: Temporal-difference (TD) learning

Prediction errors en el cerebro

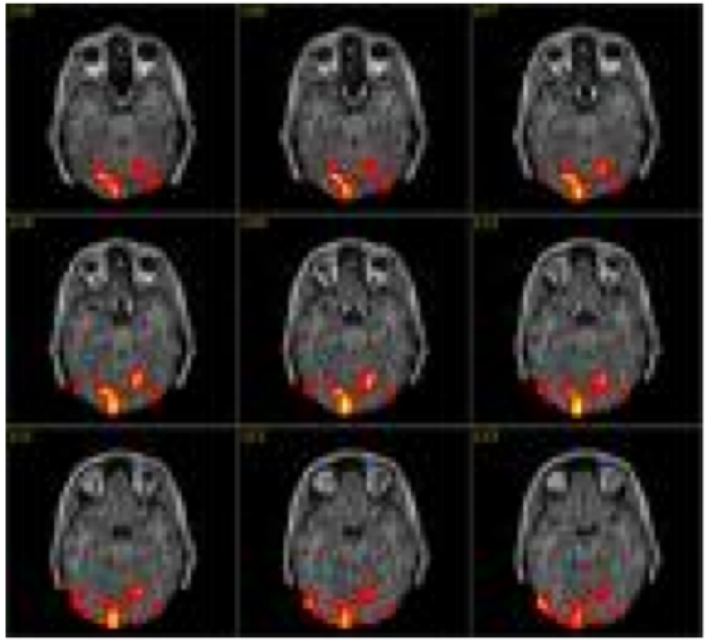


Schultz, Dayan & Montague, *Science*, 1997

AR en el cerebro: la dopamina

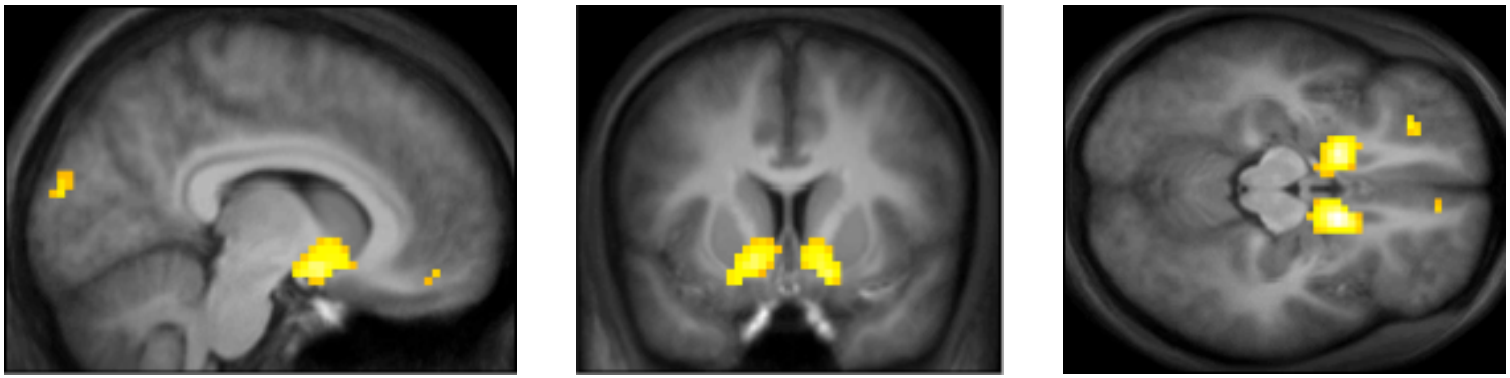


Estudiando AR en humanos…



Resonancia magnética funcional (fMRI)

Estudiando AR en humanos…



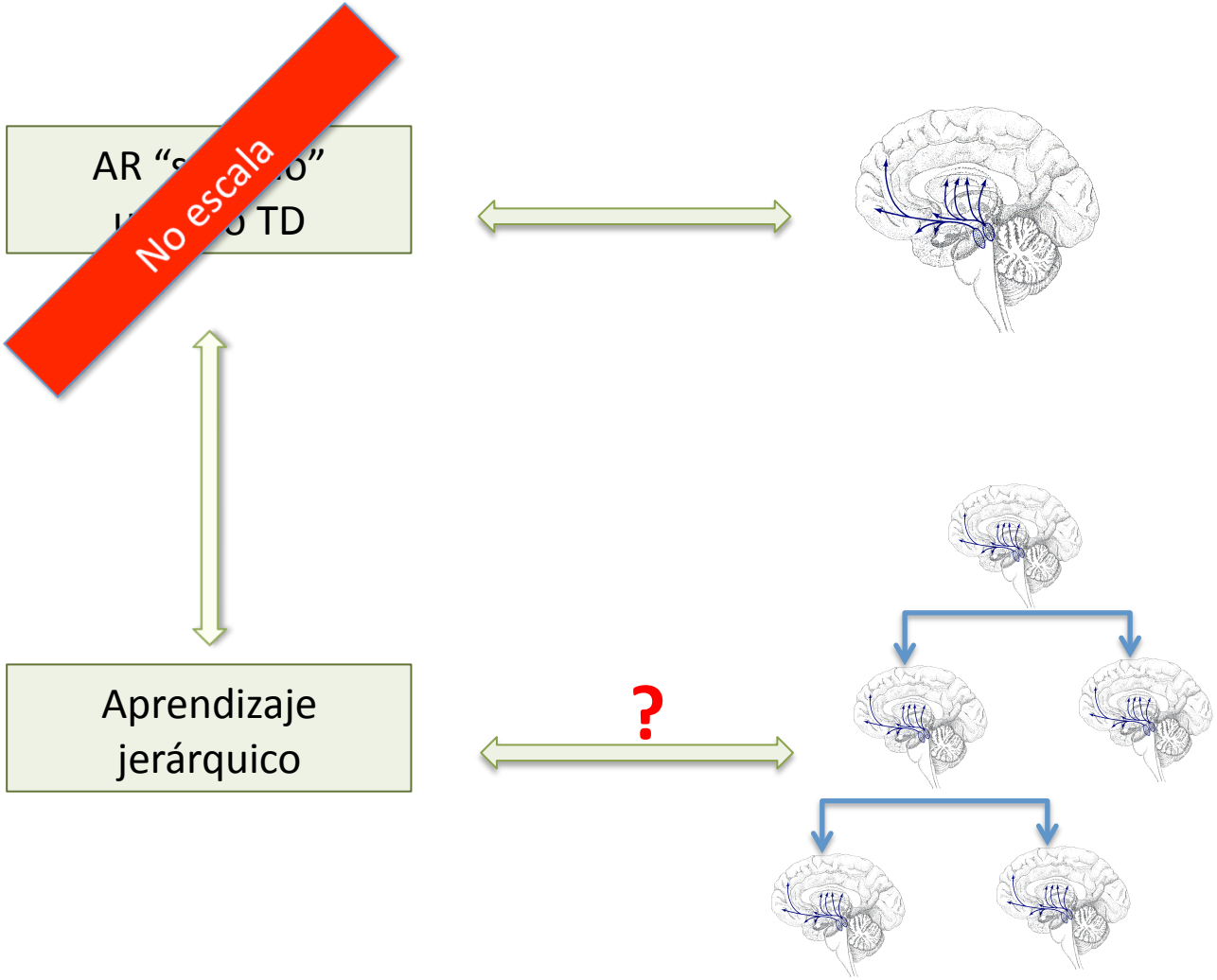
18 Participantes jugaron al juego de “qué comer” (tenían 8 opciones)

Análisis basado en un modelo:

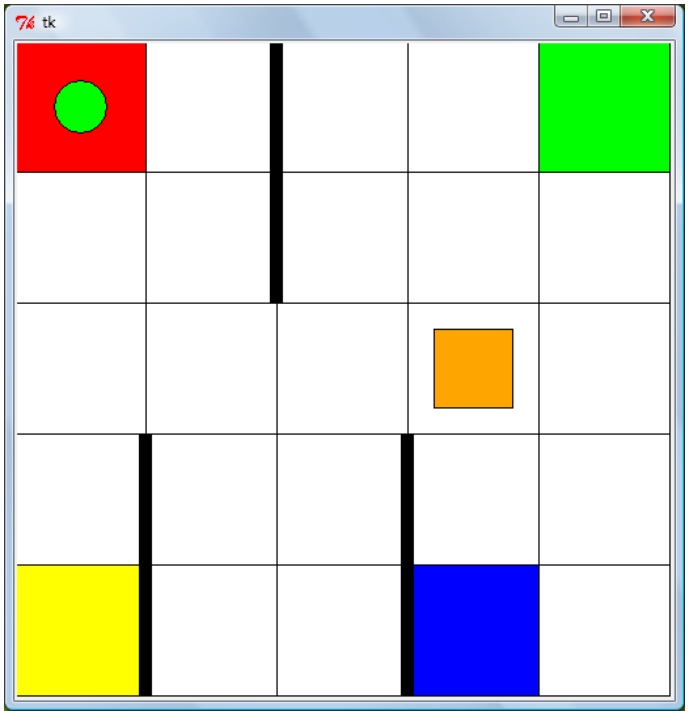
Postulamos que el algoritmo de aprendizaje es TD.

Buscamos áreas del cerebro cuyo patrón de activación correlaciona con Prediction Errors que genera TD.

Problema con TD

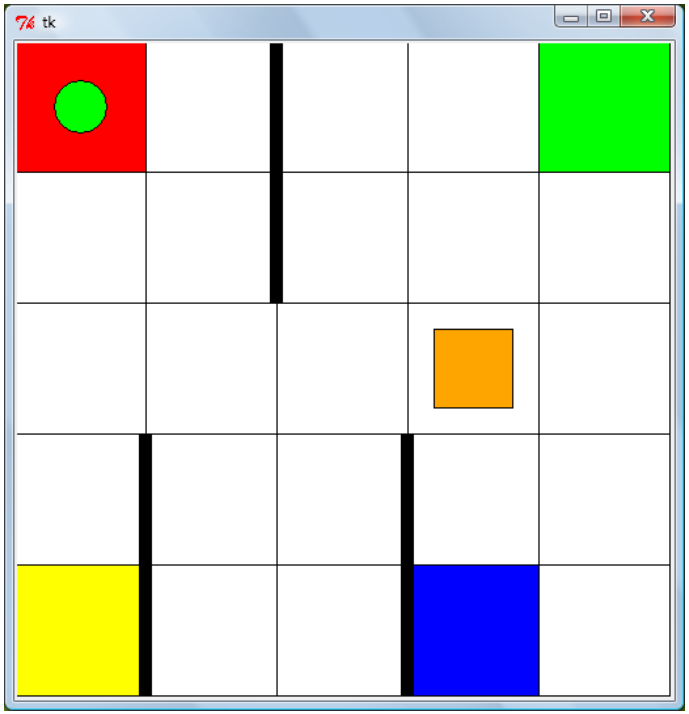


Juguemos un juego



- Pueden elegir entre 6 acciones.
- Cuando el juego se cierre, es porque ganaron.

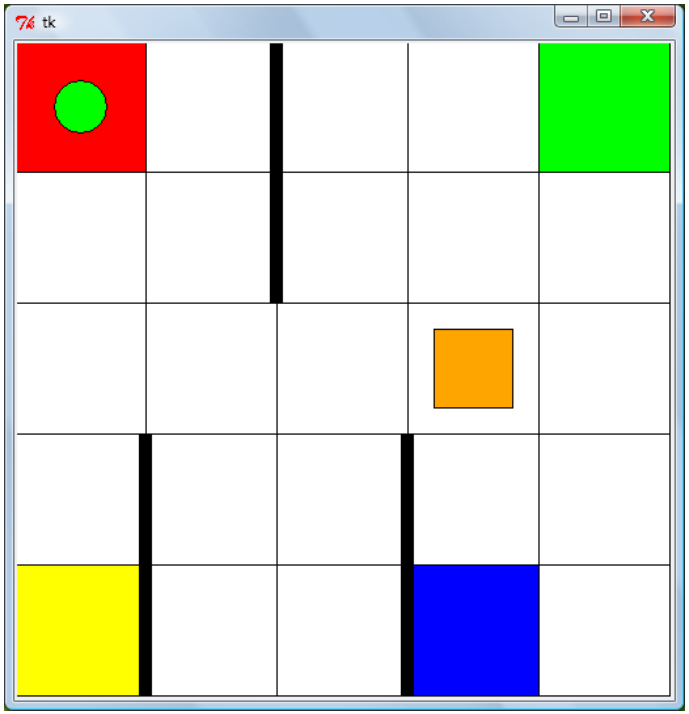
La maldición de la dimensionalidad



El Taxi Problem (Dietterich, 1999)

Problem variables	# of states
Taxi location (25)	25

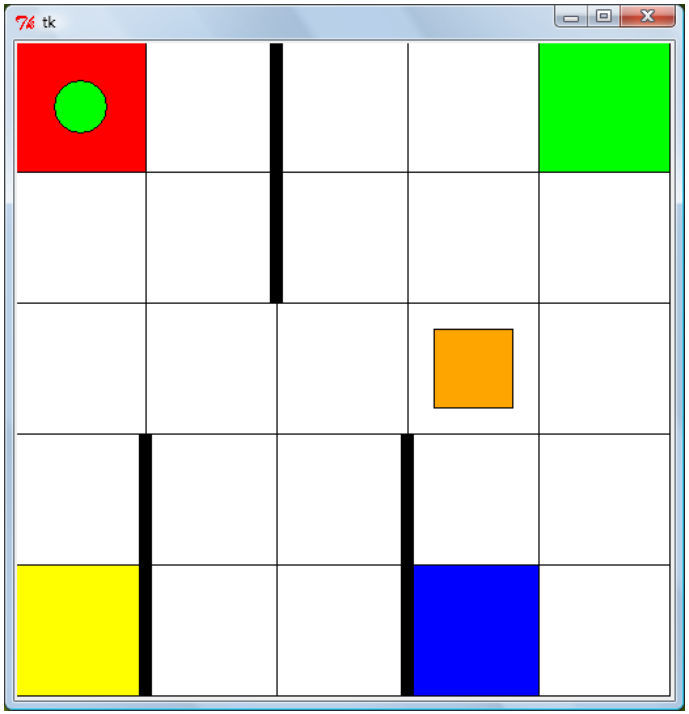
La maldición de la dimensionalidad



The Taxi Problem (Dietterich, 1999)

Problem variables	# of states
Taxi location (25)	25
Passenger location (5)	25 x 5 = 125

La maldición de la dimensionalidad



The Taxi Problem (Dietterich, 1999)

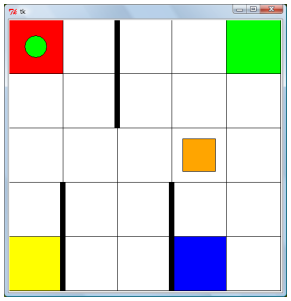
Problem variables	# of states
Taxi location (25)	25
Passenger location (5)	25 x 5 = 125
Destination (4)	125 x 4 = 500

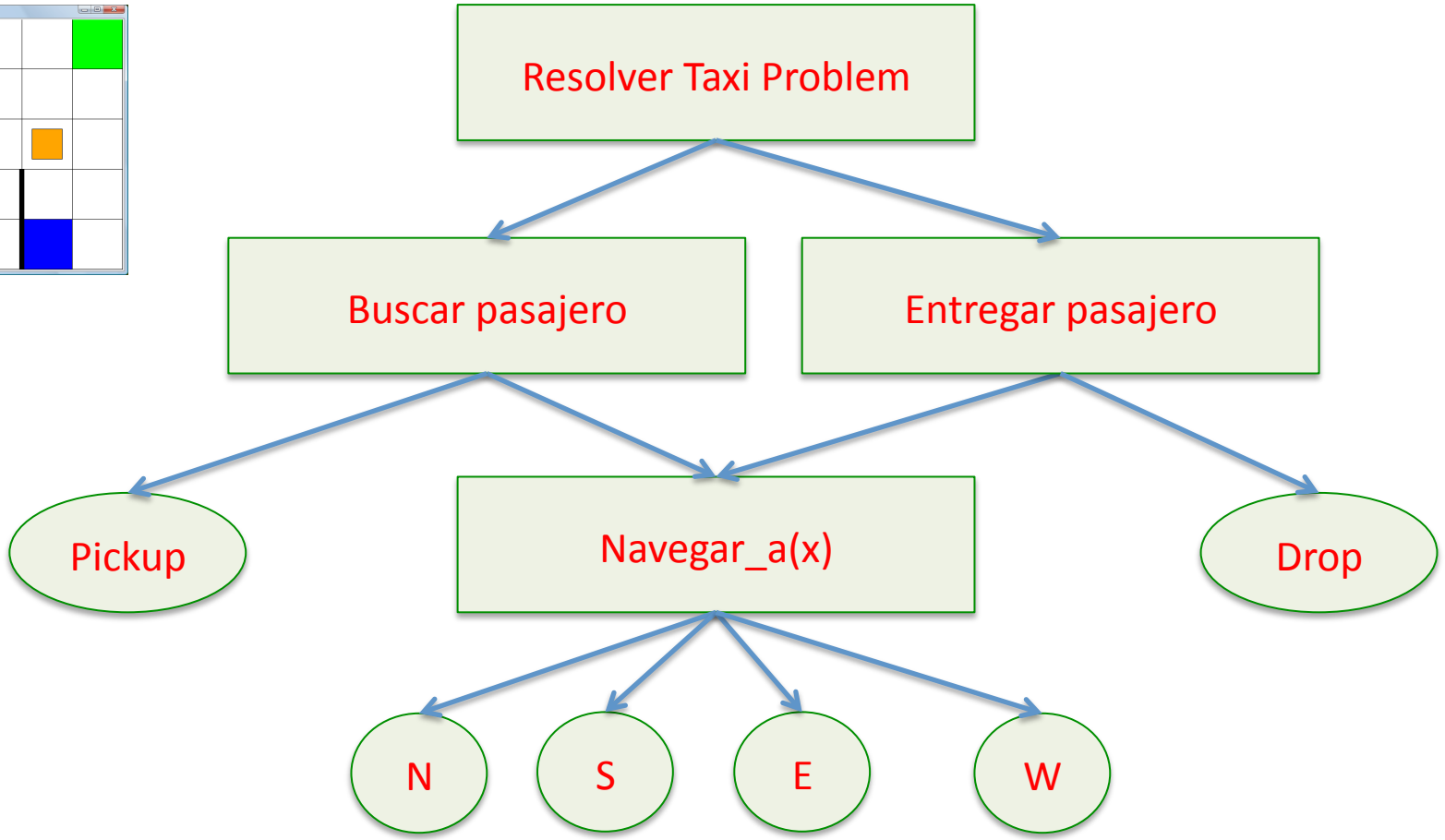
Pickup N S E W Drop

Costos en experiencia: Taxi

Método de aprendizaje	# pasos hasta comportamiento cuasi-óptimo
Temporal-difference simple (Q-learning) [Watkins & Dayan, 1992]	47157

Descompongamos la tarea





Costos en experiencia: Taxi

Método de aprendizaje	# pasos hasta comportamiento cuasi-óptimo
Temporal-difference simple (Q-learning) [Watkins & Dayan, 1992]	47157
Agregamos jerarquía (MaxQ) [Dietterich 1999, 2000]	6298

Costos en experiencia: Taxi

Método de aprendizaje	# pasos hasta comportamiento cuasi-óptimo
Temporal-difference simple (Q-learning) [Watkins & Dayan, 1992]	47157
Agregamos jerarquía (MaxQ) [Dietterich 1999, 2000]	6298
Usando un planner (R-Max) [Braffman & Tennenholtz 2002]	4151

Costos en experiencia: Taxi

Método de aprendizaje	# pasos hasta comportamiento cuasi-óptimo
Temporal-difference simple (Q-learning) [Watkins & Dayan, 1992]	47157
Agregamos jerarquía (MaxQ) [Dietterich 1999, 2000]	6298
Usando un planner (R-Max) [Braffman & Tennenholtz 2002]	4151
Usando un planner y aprendiendo a ignorar variables (Met R-Max) [Diuk et al., 2008]	2246

Costos en experiencia: Taxi

Método de aprendizaje	# pasos hasta comportamiento cuasi-óptimo
Temporal-difference simple (Q-learning) [Watkins & Dayan, 1992]	47157
Agregamos jerarquía (MaxQ) [Dietterich 1999, 2000]	6298
Usando un planner (R-Max) [Braffman & Tennenholtz 2002]	4151
Usando un planner y aprendiendo a ignorar variables (Met R-Max) [Diuk et al., 2008]	2246
Lo mismo, pero con una jerarquía [Diuk et al., 2006]	**319**

Las jerarquías ayudan!
La bendición de la abstracción

Jerarquías en el comportamiento humano

Karl Lashley
(1890-1958)

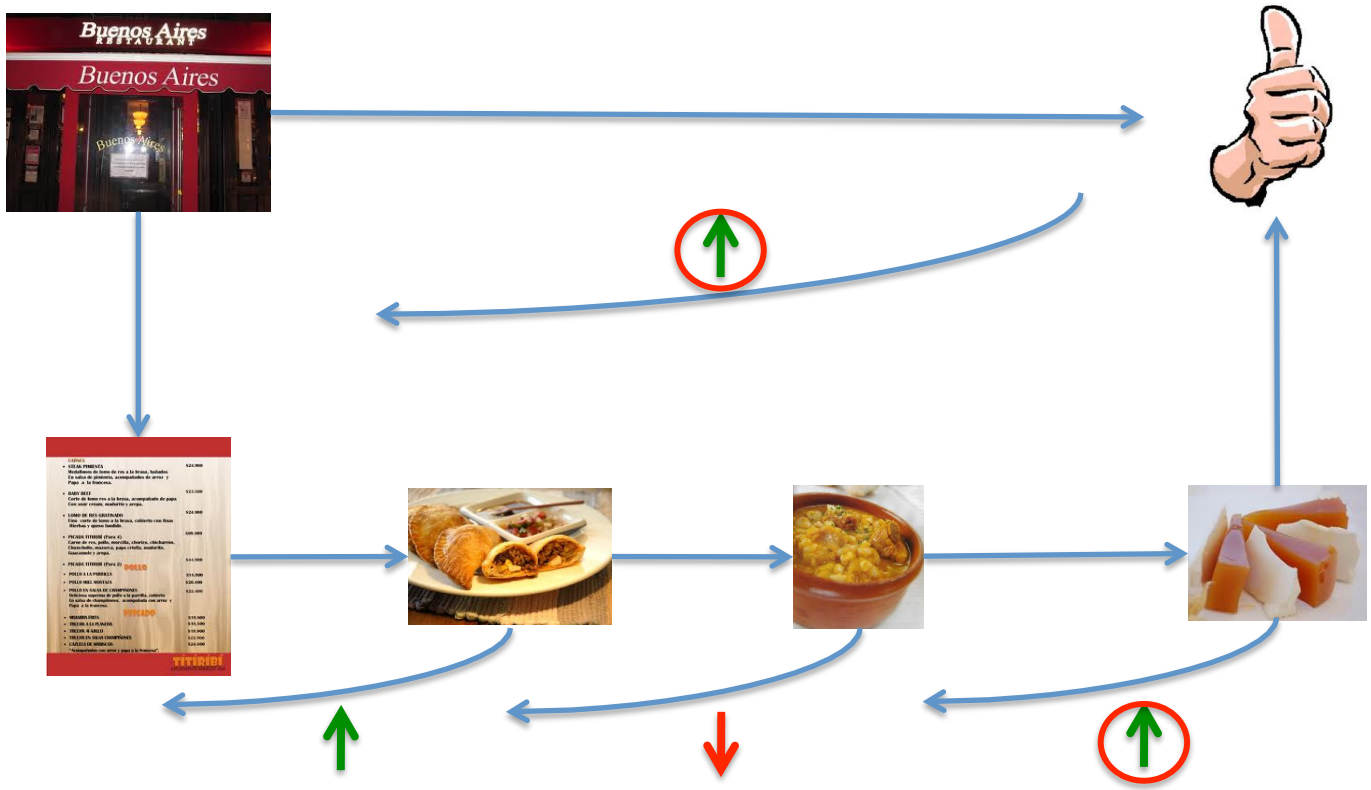
"El comportamiento secuencial no es una cadena de estímulo-respuestas, sino una jerarquía de subrutinas anidadas."

Parte del comienzo de la "revolución cognitiva".

El problema de qué comer

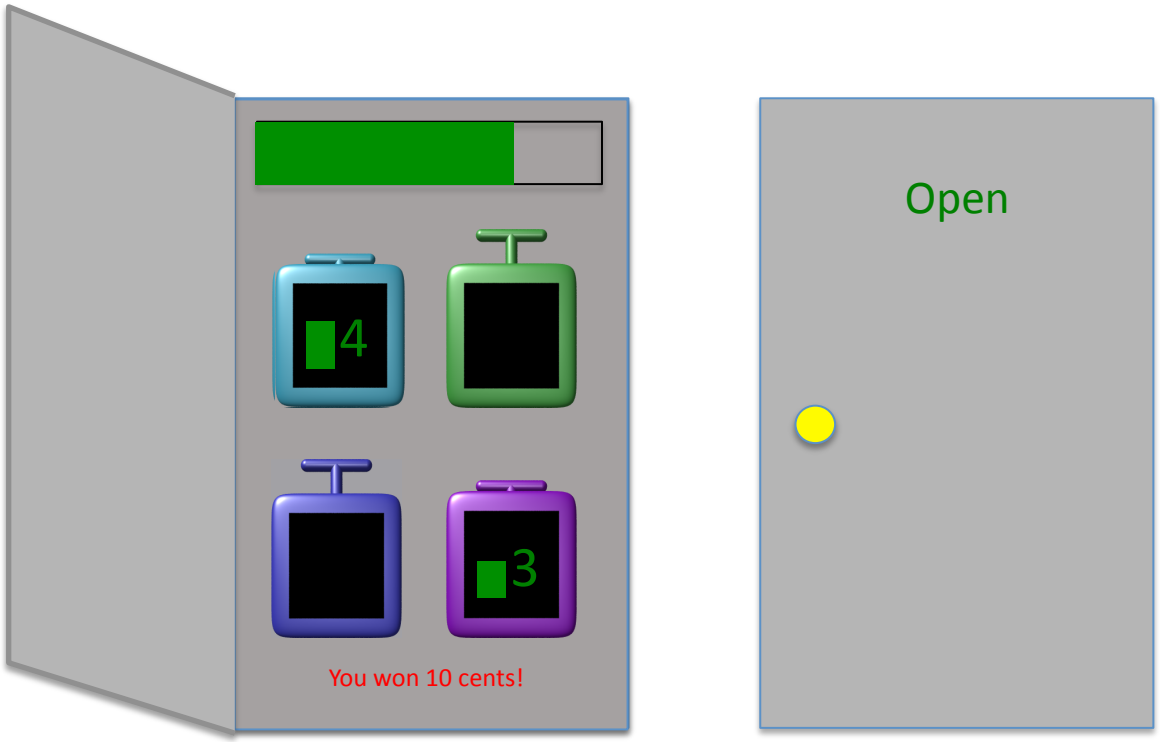
El problema de **dónde** comer

Método jerárquico basado en TD

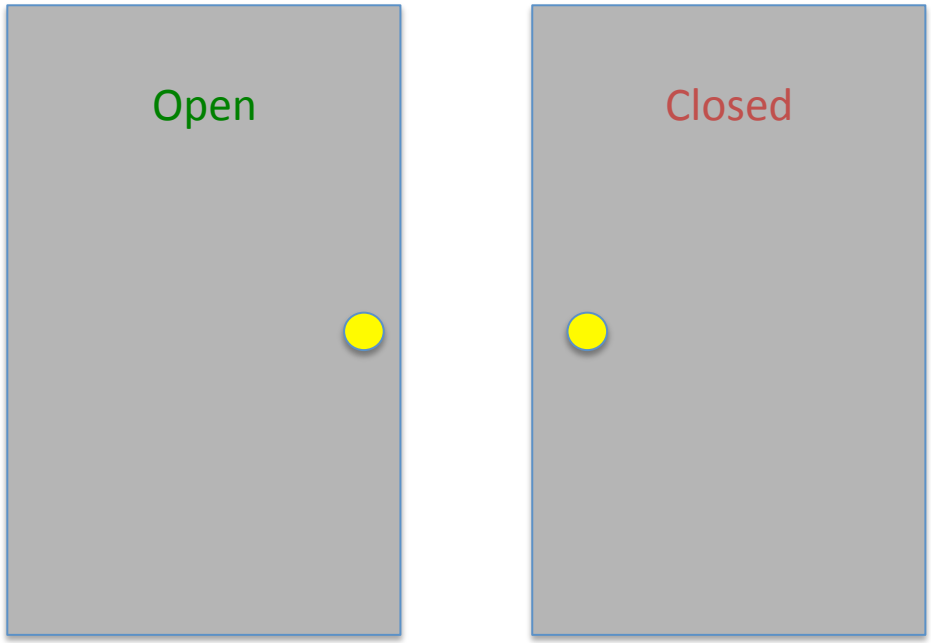


Algoritmo: Options (Sutton, Precup & Singh, 1999)

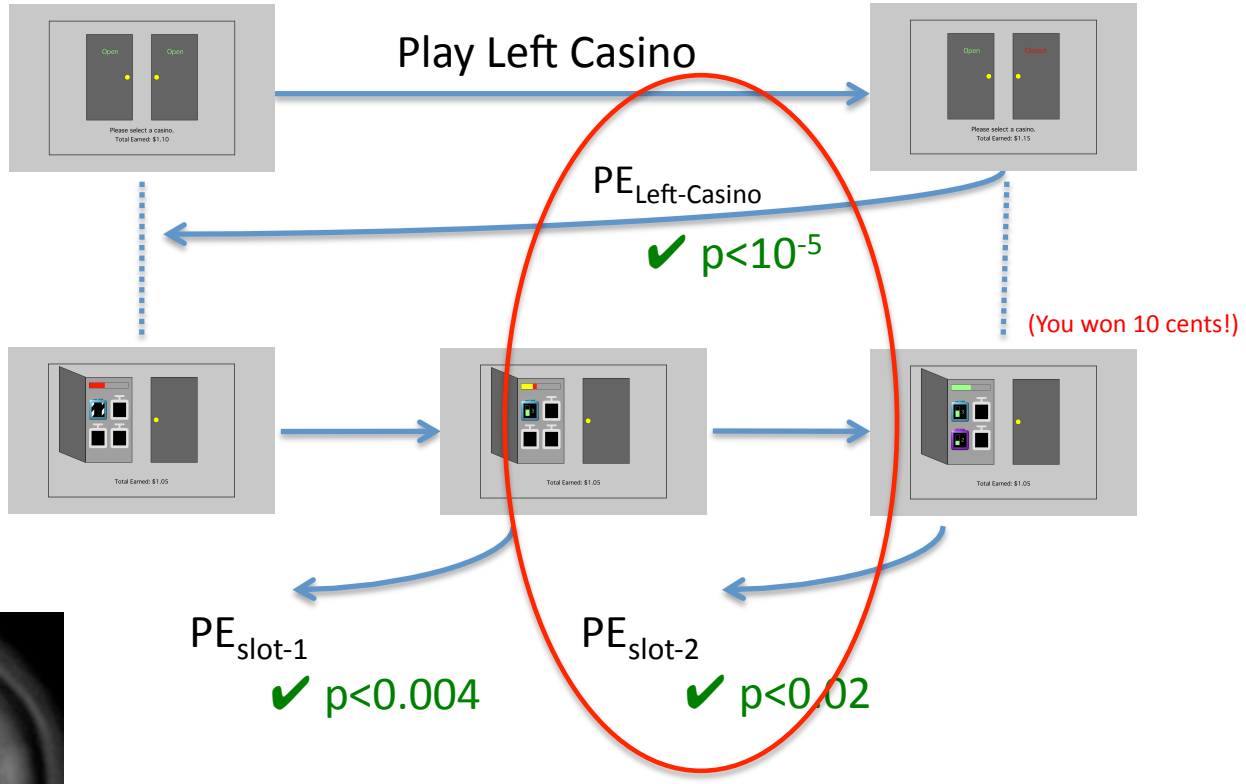
El problema del Casino

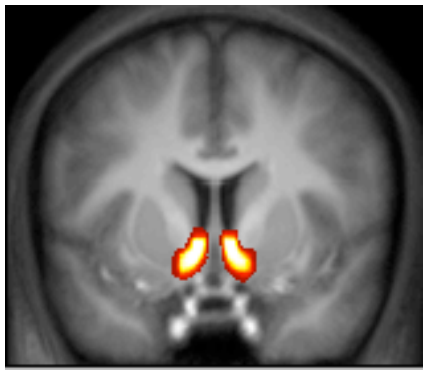


The Casino Task



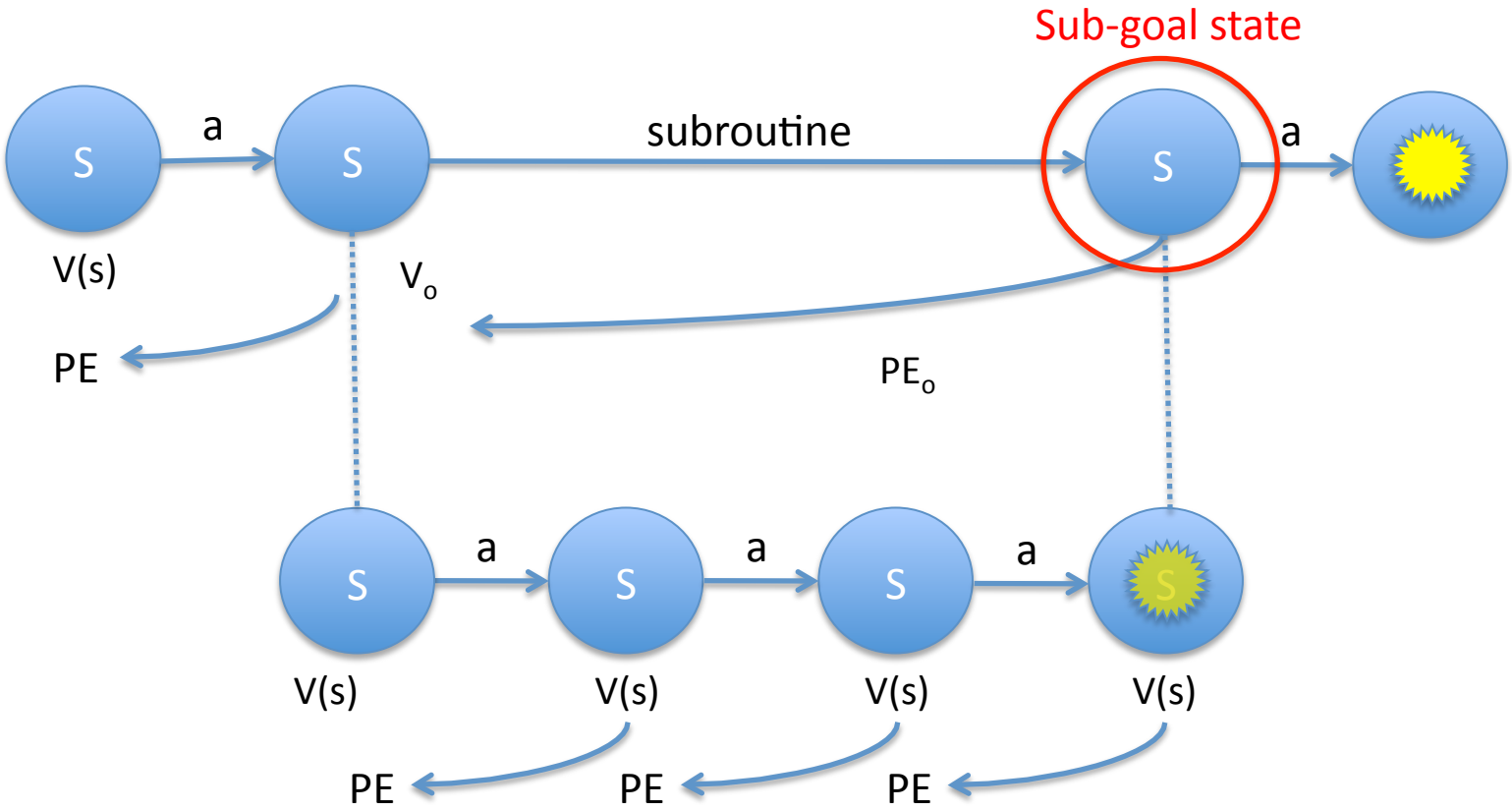
Resultados Casino



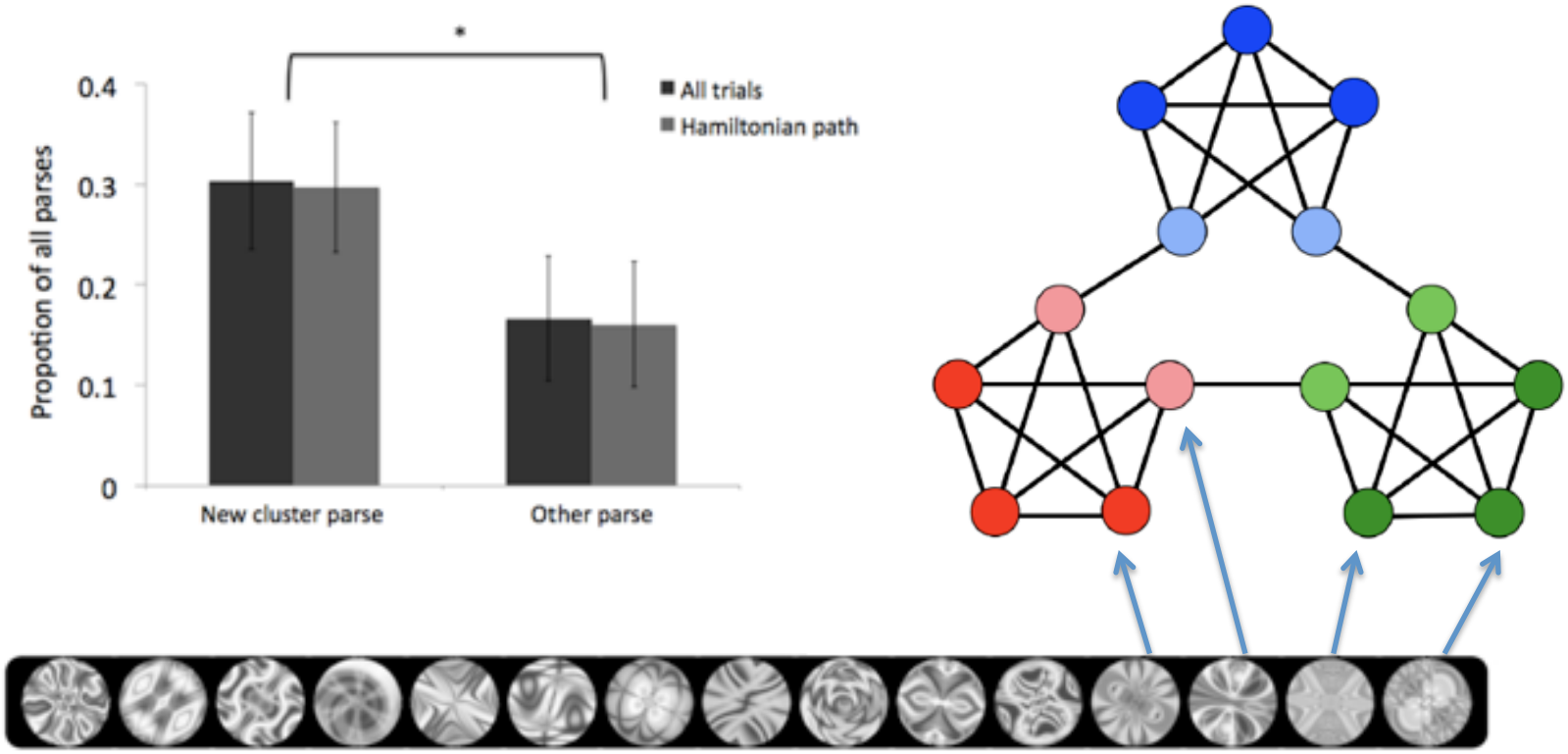


El costo de la abstracción

- Problema complejo y profundo: de dónde salen los sub-objetivos?



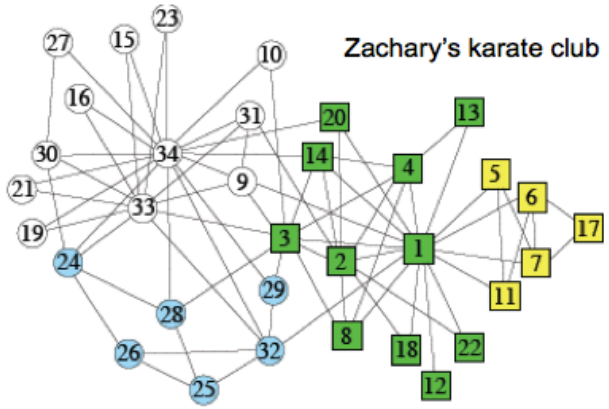
“Carving nature at its joints”

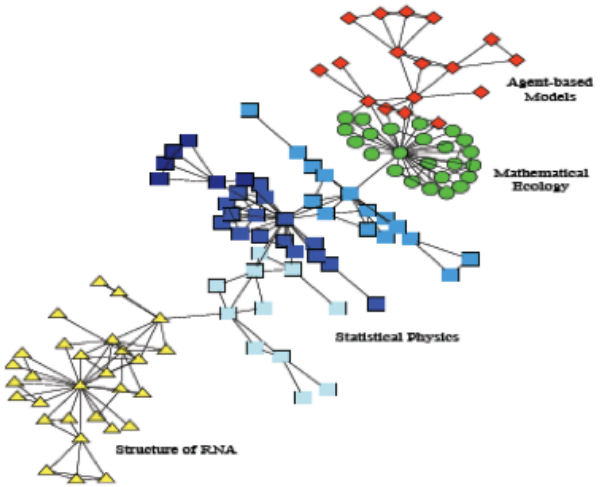


Schapiro et al., *in preparation*

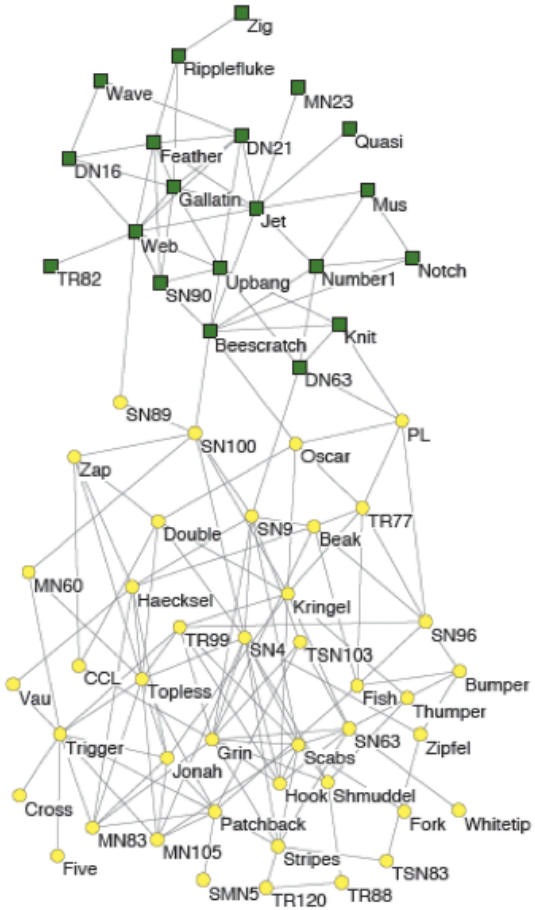
“Carving nature at its joints”

Node/vertex betweenness
Max-flow / Min-cut
Spectral clustering





Santa Fe Institute collaborations



Lusseau's bottleneck dolphins

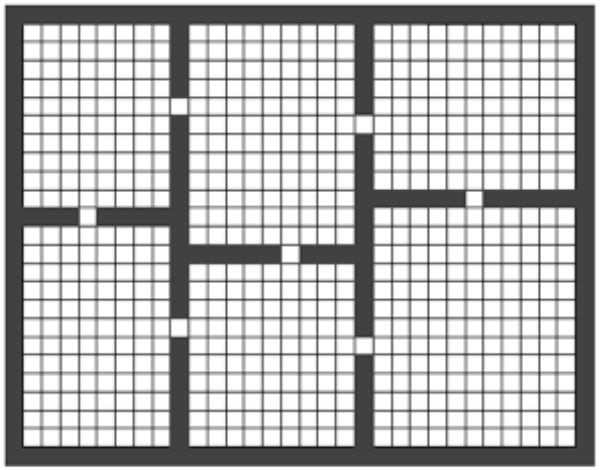
From Fortunato, *Physics Reports*, 2010

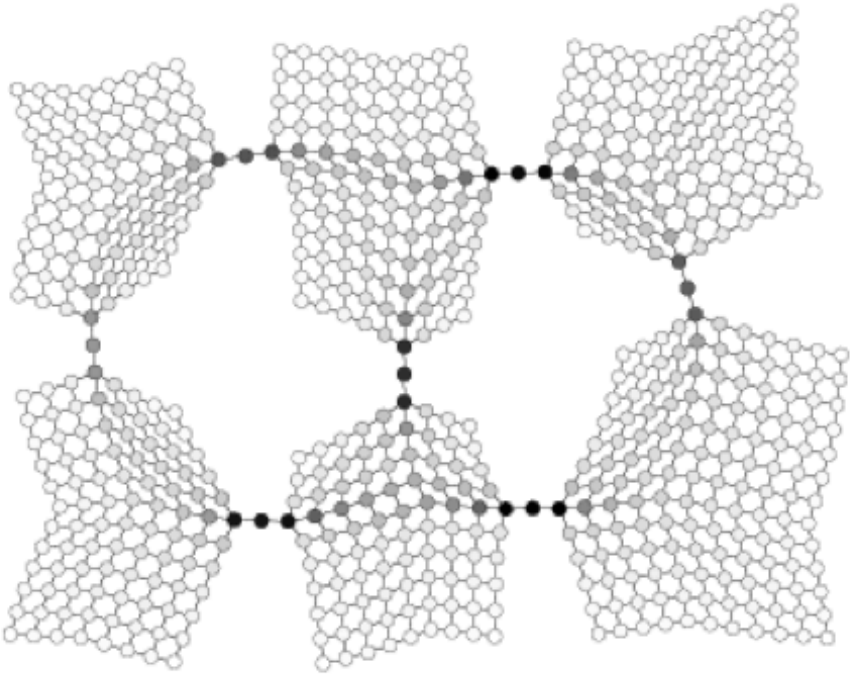
Betweenness

- Betweenness de v es la cantidad de caminos más cortos entre *s* y *t* que pasan por v, sobre la cantidad de caminos más cortos totales.

$$c_s(v) = \sum_{\forall t \neq v \neq s} \frac{\sigma_{st}(v)}{\sigma_{st}}.$$

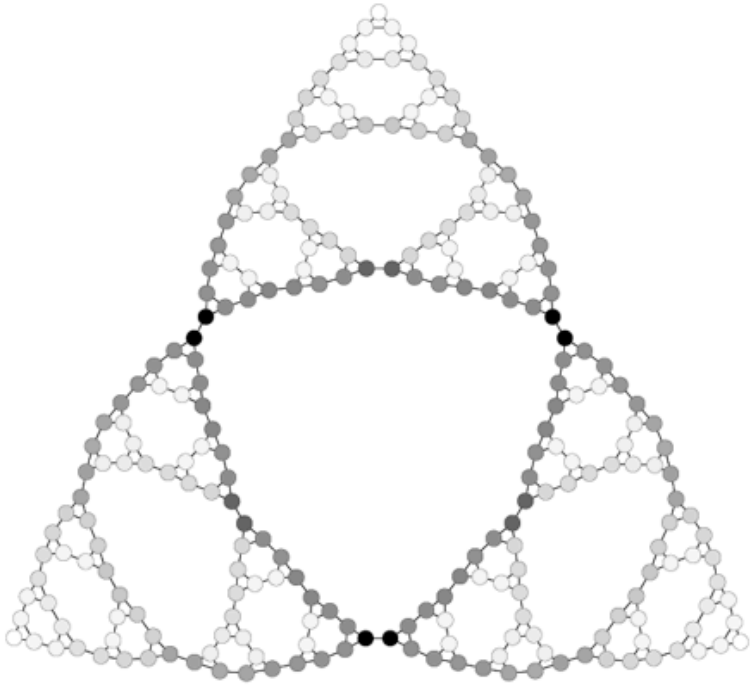
Bottlenecks en AR





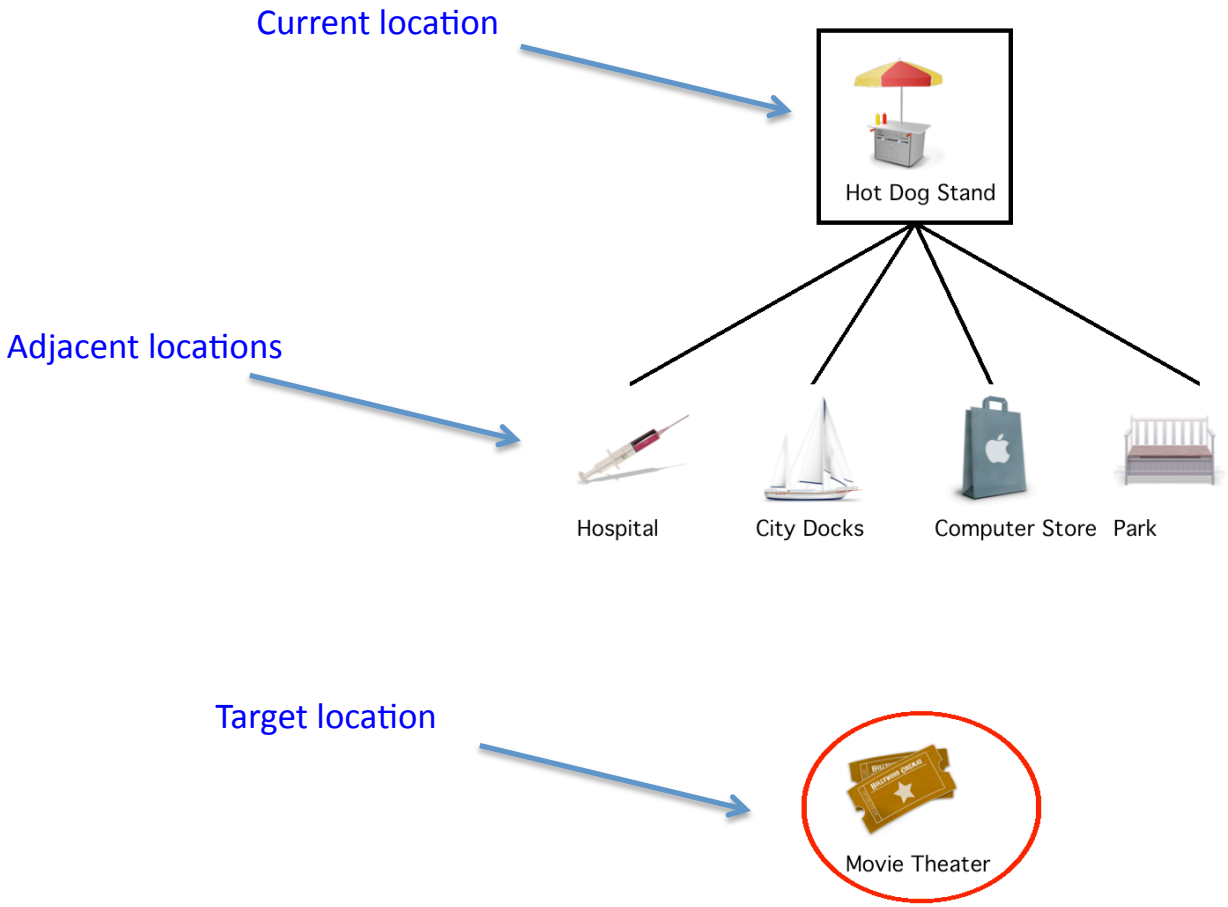
Simsek, 2008

Bottlenecks en RL

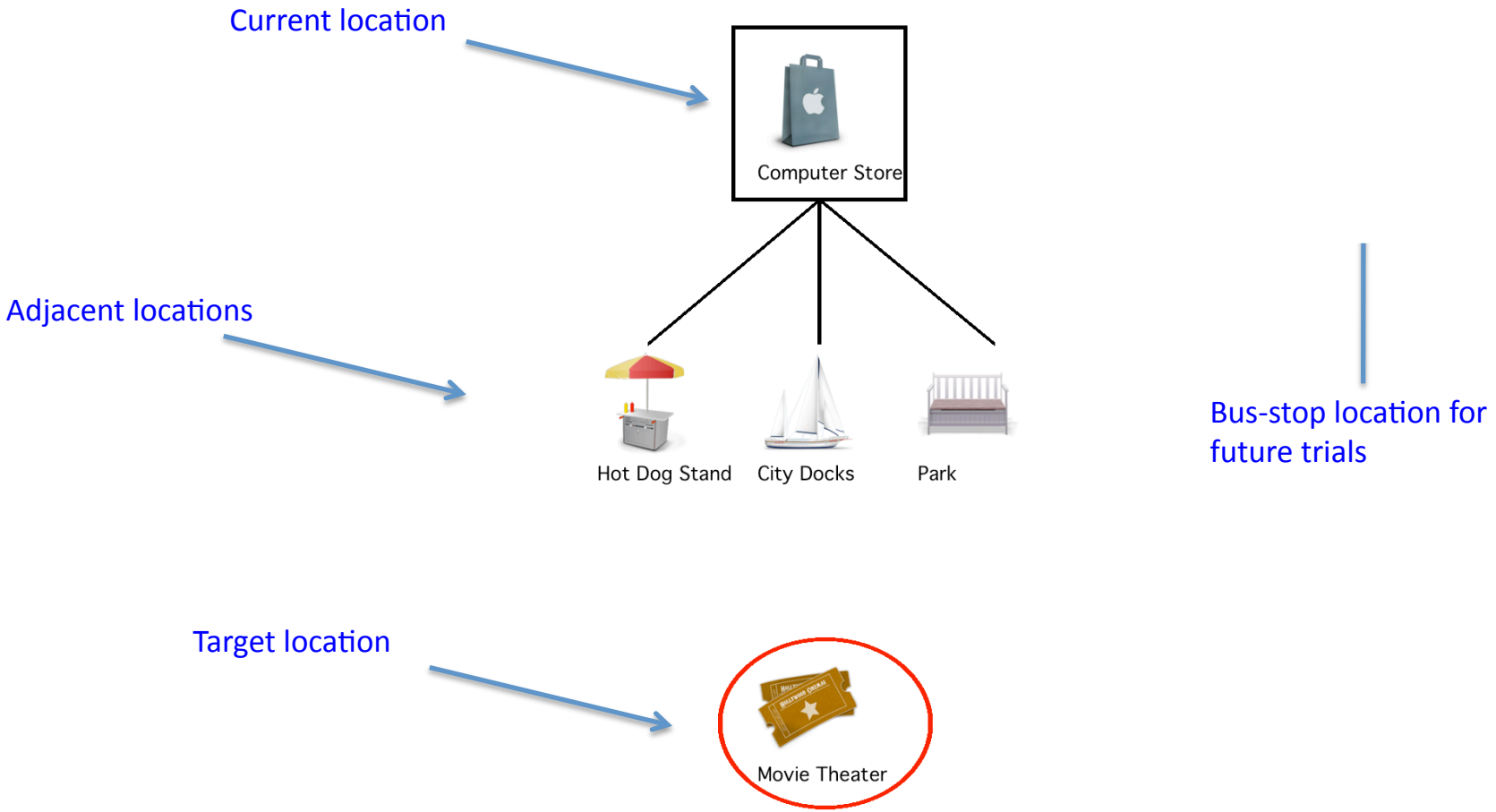


Simsek, 2008

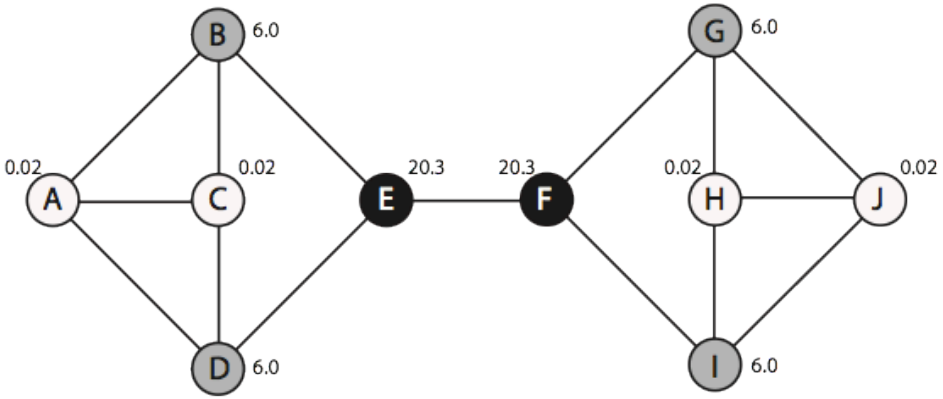
Identificando sub-objetivos

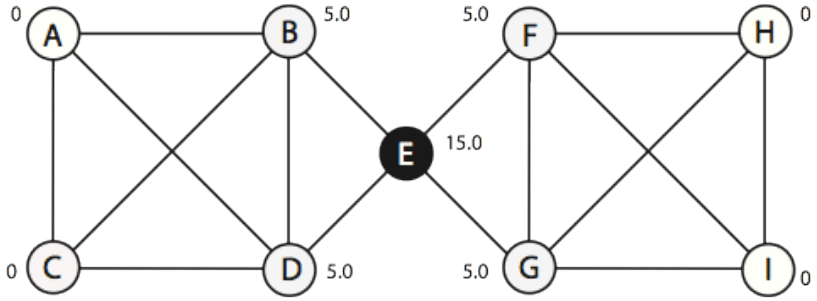


Identificando sub-objetivos



Mapas

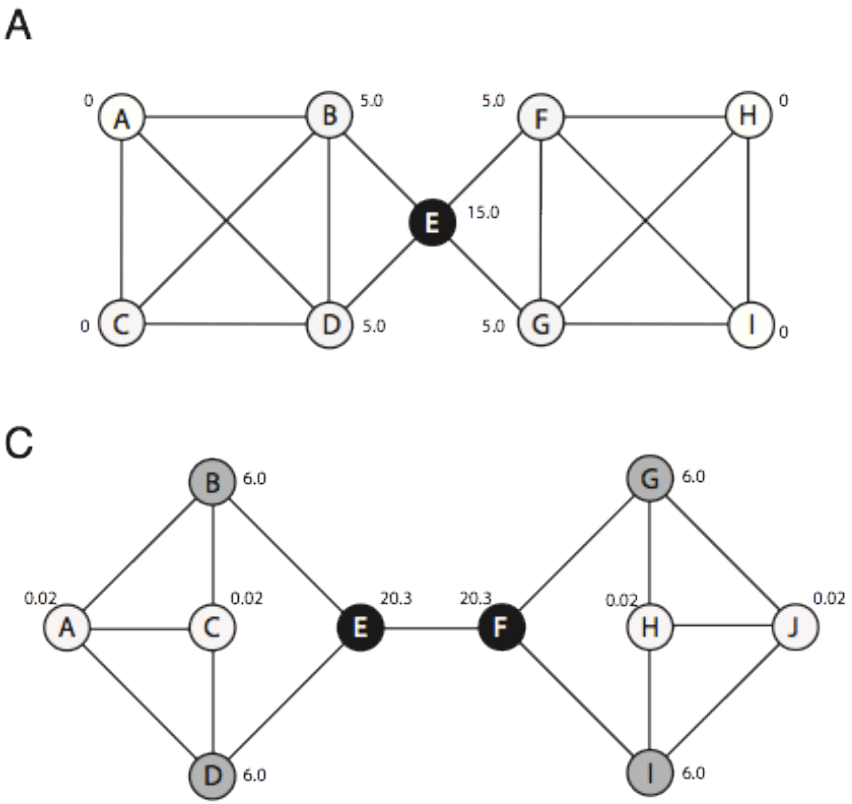




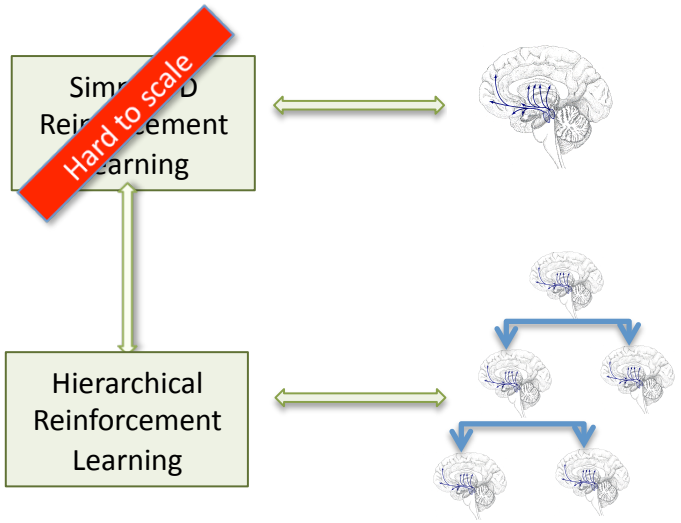
Recuerden:

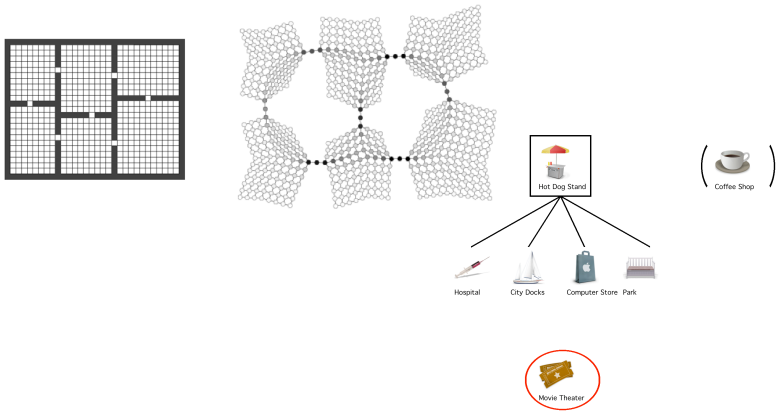
Los participantes no tienen acceso al mapa de la ciudad, sólo la navegan.

Elección de paradas de bondi



Resúmen





- Modelos de AR basados en TD explican comportamientos simples de AR en el cerebro.
- TD no escala a situaciones más complejas.
- Las jerarquías parecen ayudar.
- Creemos que los humanos piensan jerárquicamente.
- Modelamos tareas usando AR jerárquico, y verificamos algunas predicciones en el cerebro.
- Pregunta profunda: de dónde salen las jerarquías? Exploramos ideas de network analysis.

Cómo jugarías este juego?

Objetos en RL

Todavía falta…

- Creo que representaciones basadas en objetos y relaciones nos van a servir mucho.
- La percepción humana está especialmente preparada para identificar objetos.
- Sabemos muy poco sobre mecanismos de aprendizaje en el cerebro con estas representaciones.

Salud!