Active Learning

Knowledge, Learning and Inference 24/02/14

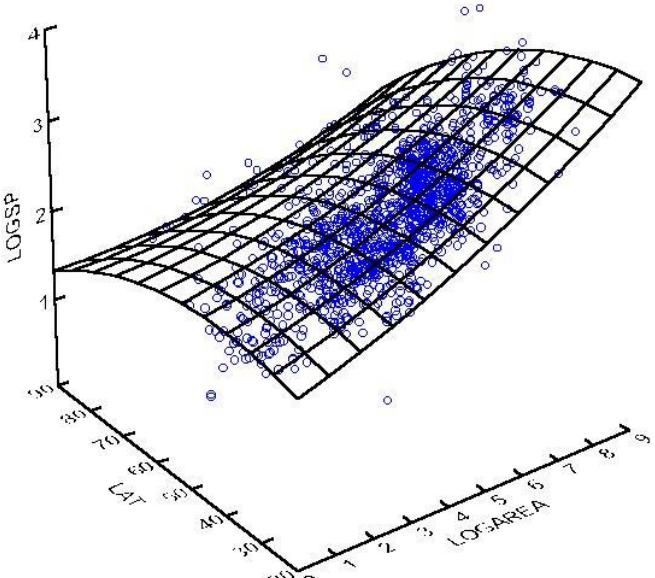
Neil Bramley

Outline

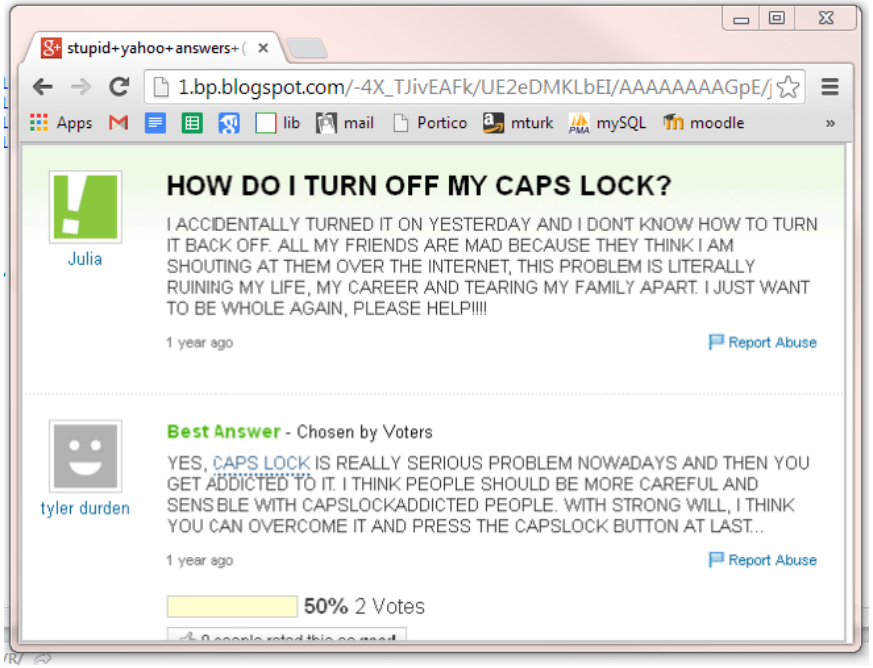
- First hour
  1. What is active learning and why is it important?
  2. What makes a useful query?
  3. Active learning research in cognitive science
- Second hour
  1. Active learning and causal structure
  2. My research

1. What is active learning and why is it important?

Passive learning



vs. Active learning



The ‘banana curve’ of active learning

- Computationally, algorithms that can actively select their own training data during learning exhibit faster learning by avoiding redundant or unhelpful data (Settles, 2009)

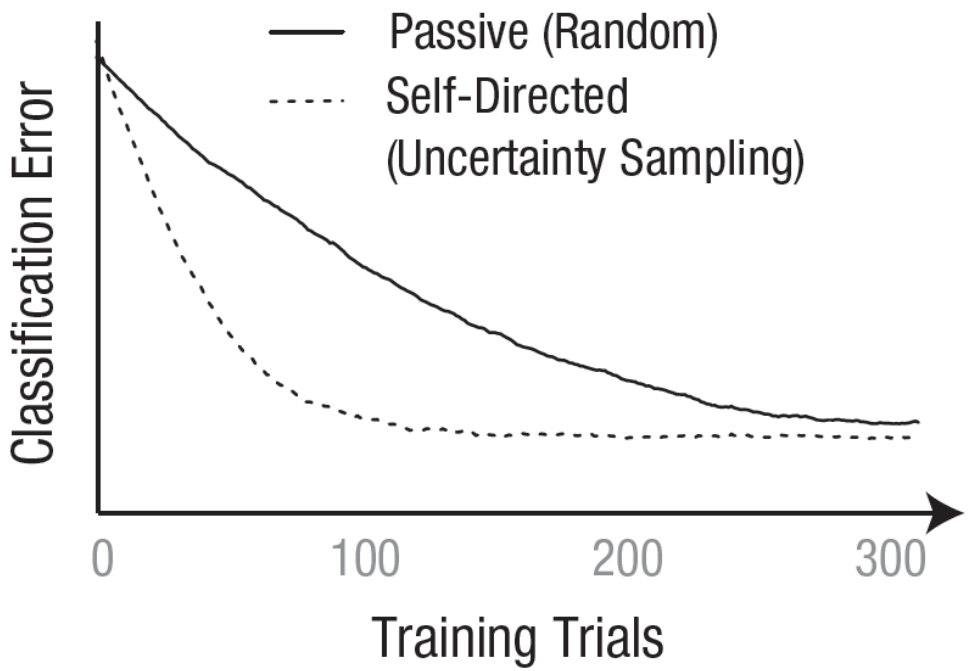


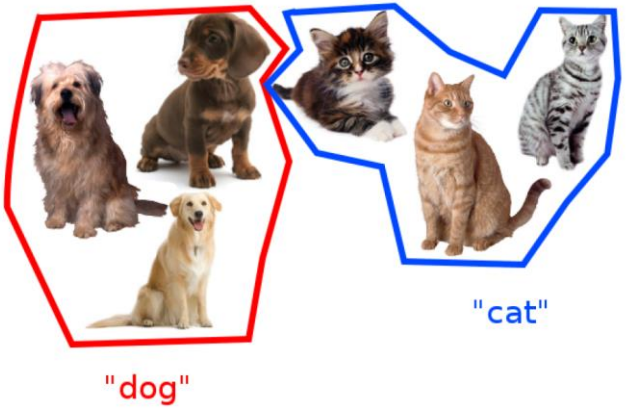
Figure from Gureckis and Markant (2013)

Completing the loop

DOG!

DOG!

CAT! etc



Observation

Generate/update hypotheses

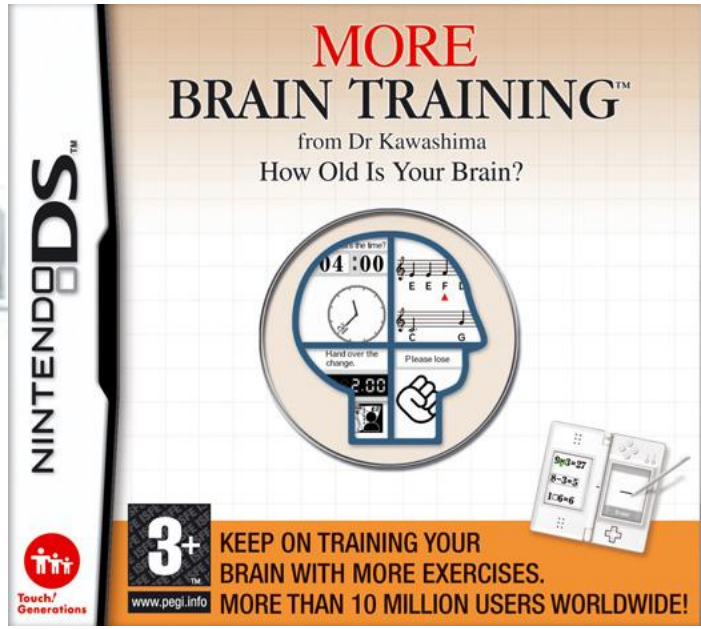
Select next action to test hypotheses

Cat or dog??

Other meanings of active learning…

Nor, ‘actively’ training your brain to be better at learning
e.g. Ball et al 2002 *JAMA*

Not learning while doing exercise
e.g. Hillman et al 2008 *Nat. Reviews Neuro*



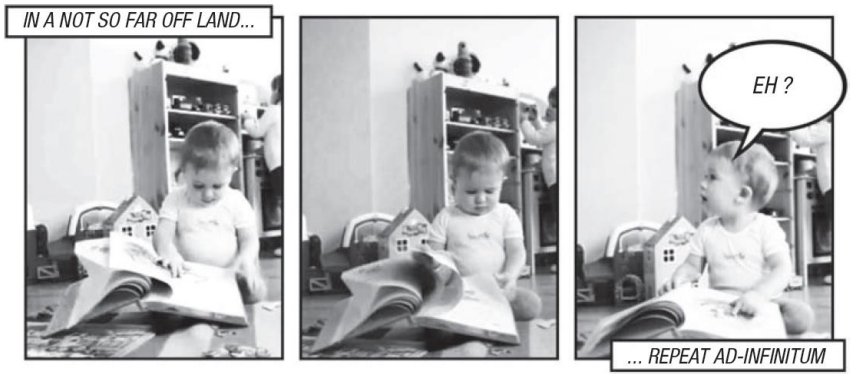
What is active learning and why is it important: Summary

- Much real world learning is active not passive in the sense that learners actively choose what to attend to, query or test in their environment.
- We typically study cognition in highly controlled i.e. passive situations
- These ignore an important part of the dynamics of how we interact with our surroundings
- Active learning choices can tell us more about how beliefs represented and updated by providing another window on learning

2. What makes a query (action, test or experiment) useful?

What is a query?

- Can think of the smallest unit of active learning as a selection from a set of possible *queries*.
- A query can take many forms:
  - Higher level e.g.:
    - Choosing a test to perform(e.g. medical, mechanical)
    - Asking a simple question (e.g. point at an object and ask for its name)
    - Designing an experiment o distinguish between theories
    - Choosing interventions to distinguish between causal structures
  - Lower level
    - Choose where to sacade to next to efficiently resolve what you are looking at.



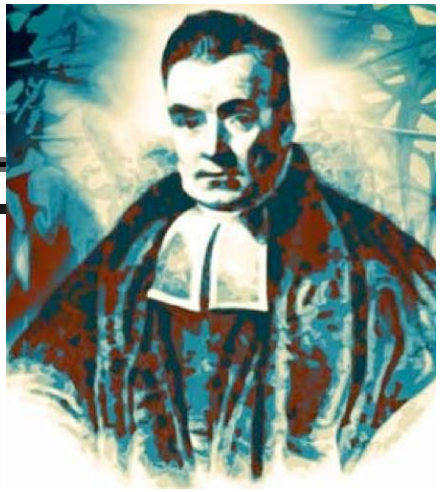
Doctor example

Based on my *passive* examination of the patient, there is a .5 probability this patient has **mumps** and a .5 probability they have **chicken pox** (but definitely has one or other, not both).
But which is it? What test (*i.e. action, query, experiment*) should I perform to help me find out?

Test	**Check potassium**		**Check sodium**	
Result	High	Low	High	Low
P(Result|Mumps)	.8	.2	.95	.05
P(Result|Chicken pox)	.15	.85	.45	.55

What makes a useful query?

- The posterior probability of (**D**)isease being (**m**)umps or (**c**)hicken pox depends on the likelihood of the (**R**)esult of the chosen test (**T**).



- We can calculate each of these with Bayes theorem:

$$P(D = x|R, T) = \frac{P(R|\,T, D = x) \times P(D = x)}{P(R|T, D = m) \times P(D = m) + P(R|T\ D = c) \times P(D = c)}$$

- E.g., $P(D = m|R = high, T = potassium) = \frac{.8\times.5}{.8\times5+.15\times.5} = .84$
- This means that if the doctor uses the blood sample for a potassium test and it comes out *high*, he can be 84% sure that the patient has mumps rather than chicken pox.

What makes a useful query?

Posteriors $\propto$

Likelihoods	Check potassium		Check sodium	
	High	Low	High	Low
P(Result|Mumps)	.8	.2	.95	.05
P(Result|Chicken pox)	.15	.85	.45	.55

$\times$

Prior	
P(Mumps)	.5
P(Chicken pox)	.5

Posteriors	Check potassium		Check sodium	
	High	Low	High	Low
P(Mumps|Result)	.84	.19	.68	.08
P(Chicken pox|Result)	.16	.81	.32	.92

What makes a useful query?

Posteriors	Check potassium		Check sodium	
	High	Low	High	Low
P(Mumps|Result)	.84	.19	.67	.08
P(Chicken pox|Result)	.16	.81	.32	.92

- The potassium test guarantees the doctor a posterior certainty about the true disease of *at least* .81.
- But testing sodium could allow greater certainty (.92 in chicken pox if it comes out *low*), but also be less useful (.67 toward mumps if it comes out *high*).
- Which pair of possible outcomes is more useful in expectancy (i.e. on average)?

What makes a useful query?

- ANSWER: It depends on your motivations.
- We need a measure of the expected utility of each test/query/action to the learner, to compare them:

$$EU(T) = \sum_{Result} U(R,T)P(R|T)$$

- The expected utility of a query is the **sum over possible results/outcomes** of **how useful** each result is, weighted by **how likely** that result is.
- For this we need a measure of the **usefulness** of the various possible posteriors. There are several possibilities for how we can measure this…

Information gain

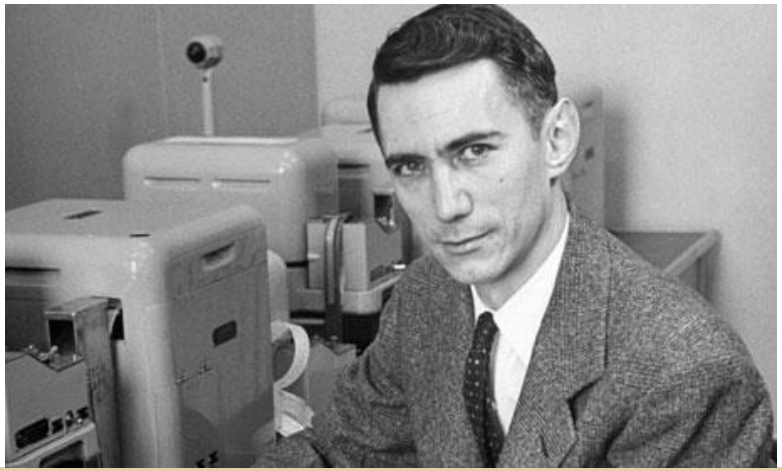
- One possibility: *expected information* gain
- Uncertainty about value of a variable, here (D)isease, is given by:

$$H(D) = -\sum_{d} P(d) \log_2 P(d)$$

(Shannon, 1948)

- *Information gain* is the reduction in one's uncertainty due to seeing some data e.g. a particular test (R)esult or observation

$$I(R) = H(D) - H(D|R)$$



Originally for quantifying how much is transmitted along a telephone line in early days of telecommunications.

Illustration of information gain

- Suppose, for illustrative purposes there are 4 options (1. no disease, 2. mumps, 3. chicken pox, 4. mumps and chicken pox)

$$H(D) = -\sum_{d} P(d) \log P(d)$$

$$= -4 \times \left(\frac{1}{4} \log_2 \frac{1}{4}\right) = 2$$

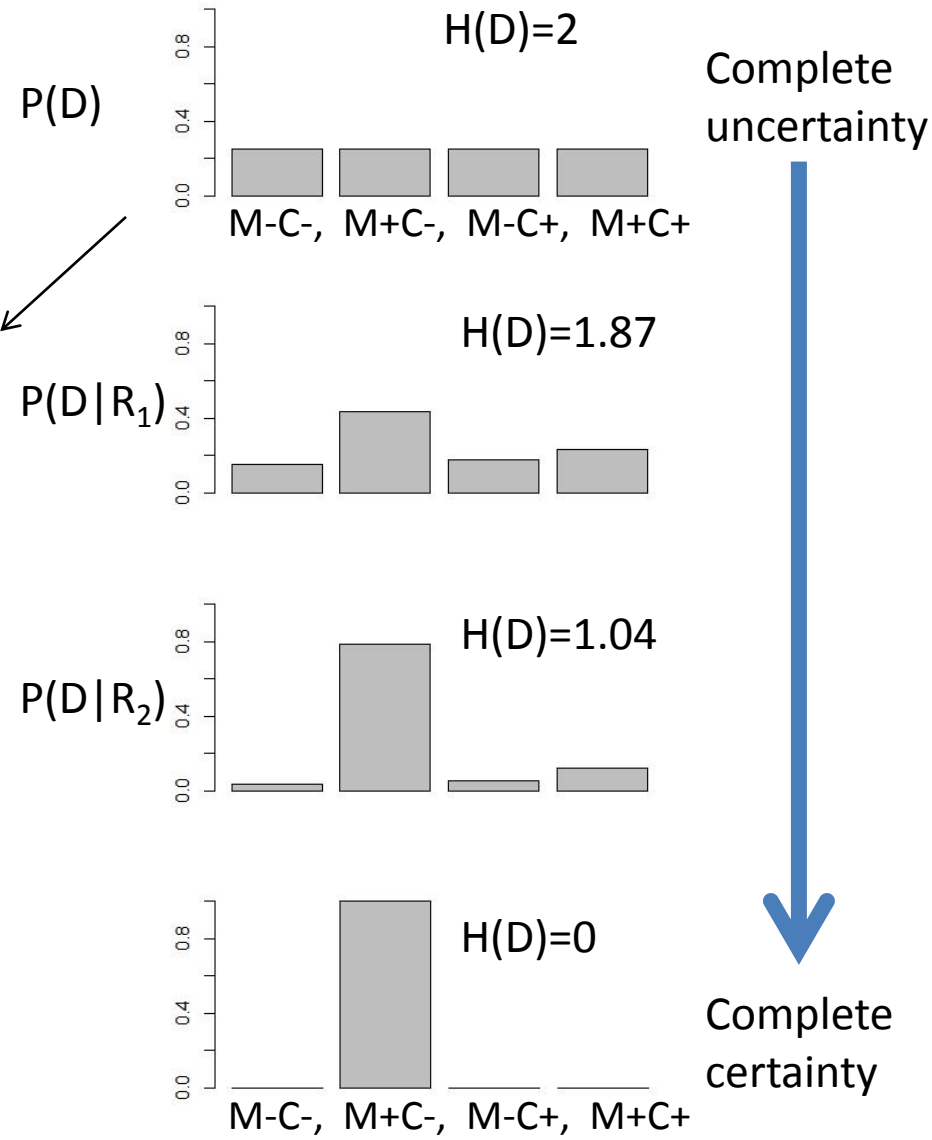
- Result R_1 leads to posterior $P(D|R_1)$ which has information gain:

$$I(R_1) = H(D) - H(D|R_1)$$
$$= 2 - 1.87 = .13\ bits$$

..while

$$I(R_2) = 2 - 1.04 = .96\ bits.$$

So, R_2 is more informative!



Expected Information gain

		Check potassium		Check sodium	
		High	Low	High	Low
Prior uncertainty	H(Disease)		1		
Posterior uncertainty	H(Disease|Result, Test)	.63	.70	.84	.37
Information gain	I(Result)	.37	.30	.16	.63

- *Expected* information gain (*EI)* is the sum of the information gains for each possible outcome, weighted by the probability of getting that outcome.
- In the doctor scenario, the (p)otassium test has *expected information gain:*

$$EI(p) = I(high)P(high|p) + I(low)P(low|p) = .37 \times .475 + .30 \times .525 = .35\,bits$$

- While the sodium test has a lower *expected information gain* of .27 bits.

Expected probability gain

- An alternative conception of a query's usefulness is its' "probability gain" (Baron, 1985)
- How much does a piece of data increase your chance of making a correct guess:

$$\mathrm{p_{gain}(T, R)} = \max P(D|R) - \max P(D)$$

- Again, we get *expected* probability gain by summing over possible test results weighted by their marginal probabilities.

$$\begin{aligned} Ep_{gain}(p) \\ &= p_{gain}(high)P(high|p) \\ &+ p_{gain}(low)P(low|p) \\ &= (.84 - .5) \times .475 + (.81 - .5) \\ &\times .525 = .32 \end{aligned}$$

If I choose the potassium blood test and it comes out *high,* I'll tell the patient she has Mumps. That way I'll have an .84 probability of being right…

Comparing information and probability gain

- Information gain and probability gain sometimes disagree about which action is more useful
- This is the case with this example (again coming from a flat prior P(Mumps)=.5, P(Chicken pox)=.5).

	Check vitamin C		Check vitamin D	
	High	Low	High	Low
P(Result| mumps)	.55	.45	.001	.999
P(Result|chicken pox)	.37	.63	.18	.81

Situation specific utilities

A test is only useful to me if it can result in at least 90% certainty, otherwise I cannot prescribe anything…

	Check potassium		Check sodium	
	High	Low	High	Low
P(Mumps|Result)	.84	.19	.73	.07
P(Chicken pox|Result)	.16	.81	.27	**.92**
P(Result|Test)	.475	.525	.65	.35
U	**0**	**0**	**0**	**1**
EU	**0**		**.35**	

Other measures – Nelson 2005 *Psych Rev.*

- Diagnosticity (or "weight of evidence") - for identifying the category c (0 or 1) diagnosticity of query outcome j is:

$$diagnosticity(q_j) = \max\left(\frac{P(q_j|c_1)}{P(q_j|c_0)}, \frac{P(q_j|c_0)}{P(q_j|c_1)}\right)$$

- Impact, or absolute change in probabilities

$$impact(q_j) = \frac{1}{n}\sum_{c_i} abs[P(c_i|q_j) - P(c_i)],$$

- Impact works relatively similarly to probability gain, diagnosticity ignores the prior so makes quite different (and worse!) predictions about what is a useful test (Nelson 2005).

What makes a useful query? - Summary

- To evaluate how useful a query is, we can think about its possible outcomes and what they would tell us.
- Using Bayes theorem we can compute the posteriors under different potential outcomes of a query.
- With some conception of the utility of these posteriors we can weigh them up to get the expected value of that query.
- Information gain and probability gain, relative to one's prior, are typical utilities to use but others may be better dependent on the situation
- We can do this for all possible tests or queries in order choose which one is the most valuable .

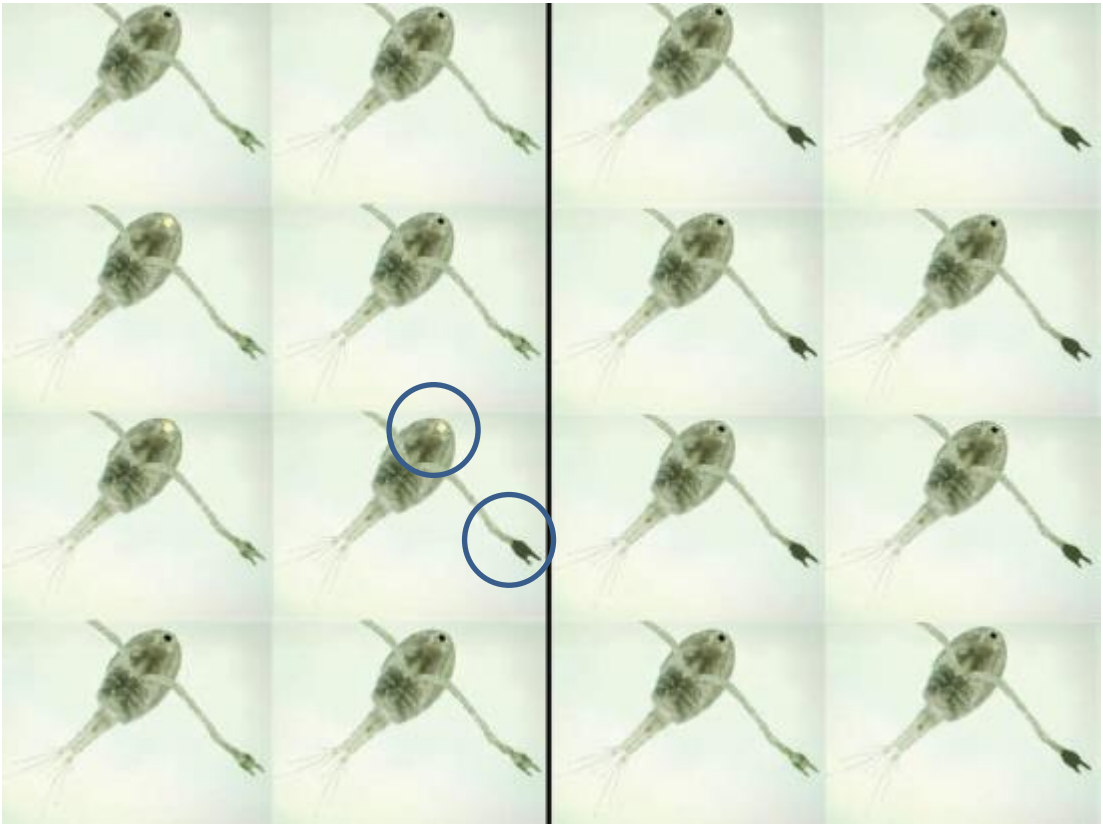
3. Active learning research in cognitive science

Typical research questions

- To what extent are people able to select queries which are useful? Are they more useful than chance selection?
- Are peoples active learning actions better described by information gain, probability gain or some other measure?
- What do people's active learning actions tell us about what hypotheses they are considering?

We will now look at a few examples of active learning research.

Nelson, McKenzie, Cottrel & Sejnowski (2010)



Species A Species B

- Nelson et al, design a classification task which pits probability gain against information gain.
- Features probabilistically related to species. E.g. 7/8 of species A have white claws, but only 1/8 of species B.
- Features initially concealed, participants choose which feature to uncover before classifying each plankton.
- Feature probabilities optimized computationally to maximally separate prediction strength of two models.

Nelson, McKenzie, Cottrel & Sejnowski (2010)

- Nelson et al find, in 3 studies, that participants actions are more consistent with probability gain than information.

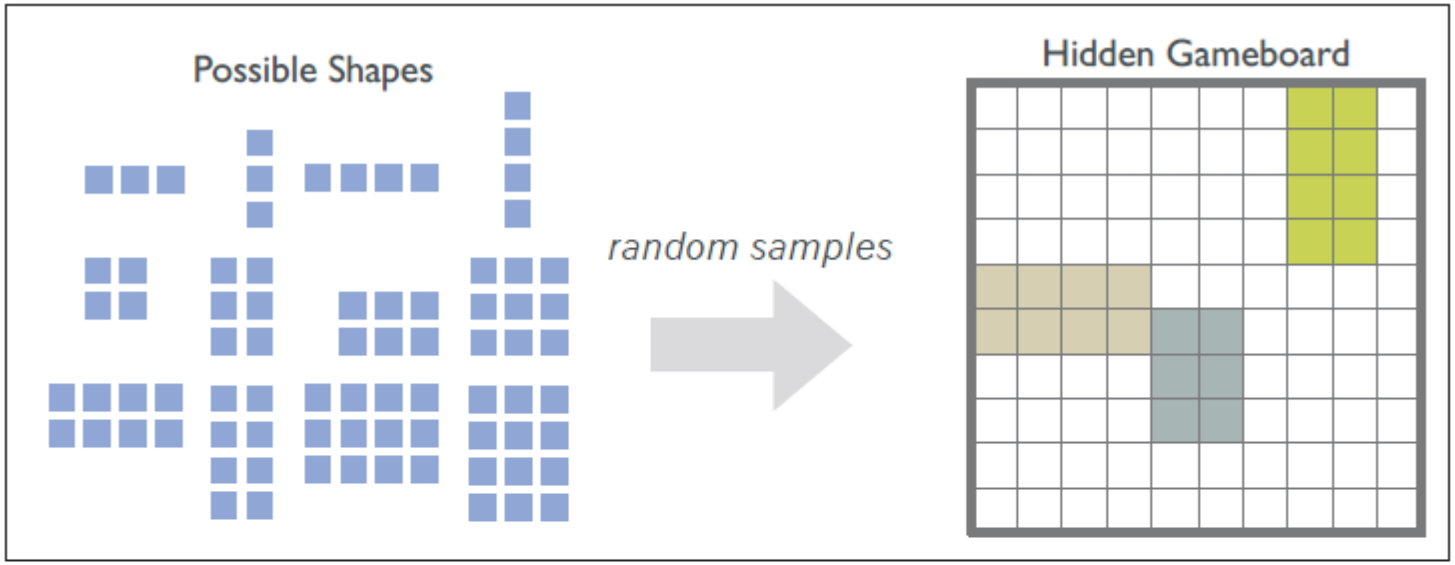
Caveats:

- Limited to scenarios in which active learning is one-shot followed by immediate classification, in this situation maximising probability of being correct is very sensible.
- May not still be true in multi-shot learning.
- May be true of categorisation but not other areas.

“Battleships” - Does the *utility* of information influence sampling behaviour? - Gureckis & Markant, 2012

Does the utility of information influence sampling behaviour?

Experiment 1: Rectangle Search



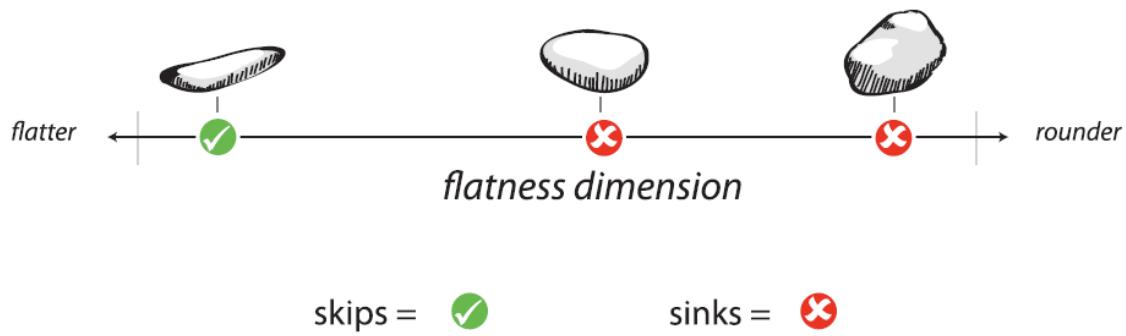
- Participants complete a sampling phase in which they test a number of grid locations and a test phase where they colour in where they think the blocks (ships) are
- They pay a cost for each additional sample and for each error in colouring in the ships
- Task pits information acquisition against payoff maximisation in an explore-exploit dynamic

Battleships

- Participants search the grid in a way that is a closer match to an information maximising strategy than a payoff maximising strategy
- Authors conclude that people are concerned with reducing their uncertainty not maximising their earnings when active learning, suggesting active learning is somewhat "*cost insensitive*"
- Authors argue that maximising information is a sensible general purpose strategy as expected utility is often undefined or intractable to calculate while information will help achieve any future goals, known or unknown

Is it better to select or to receive - Markant and Gureckis, 2013

The result of some stone-skipping tests

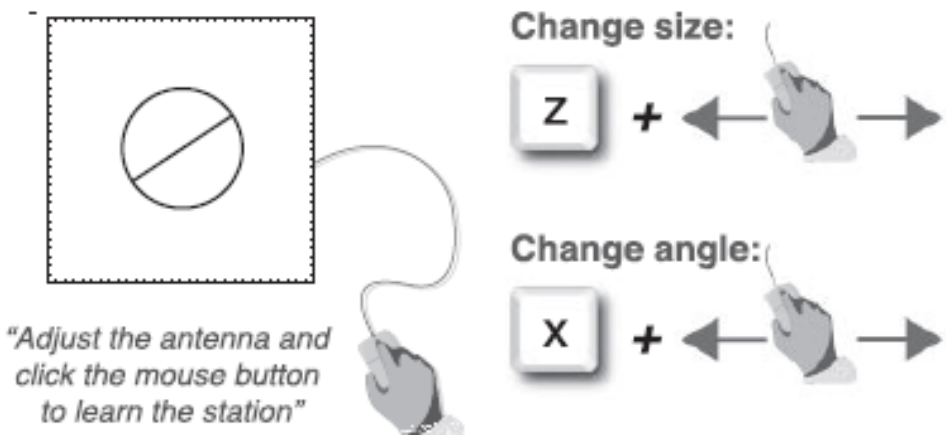


Two different "hypotheses" about which stones will skip that are consistent with the data

Possible selections that might be considered on the next trial

Is it better to select or to receive?

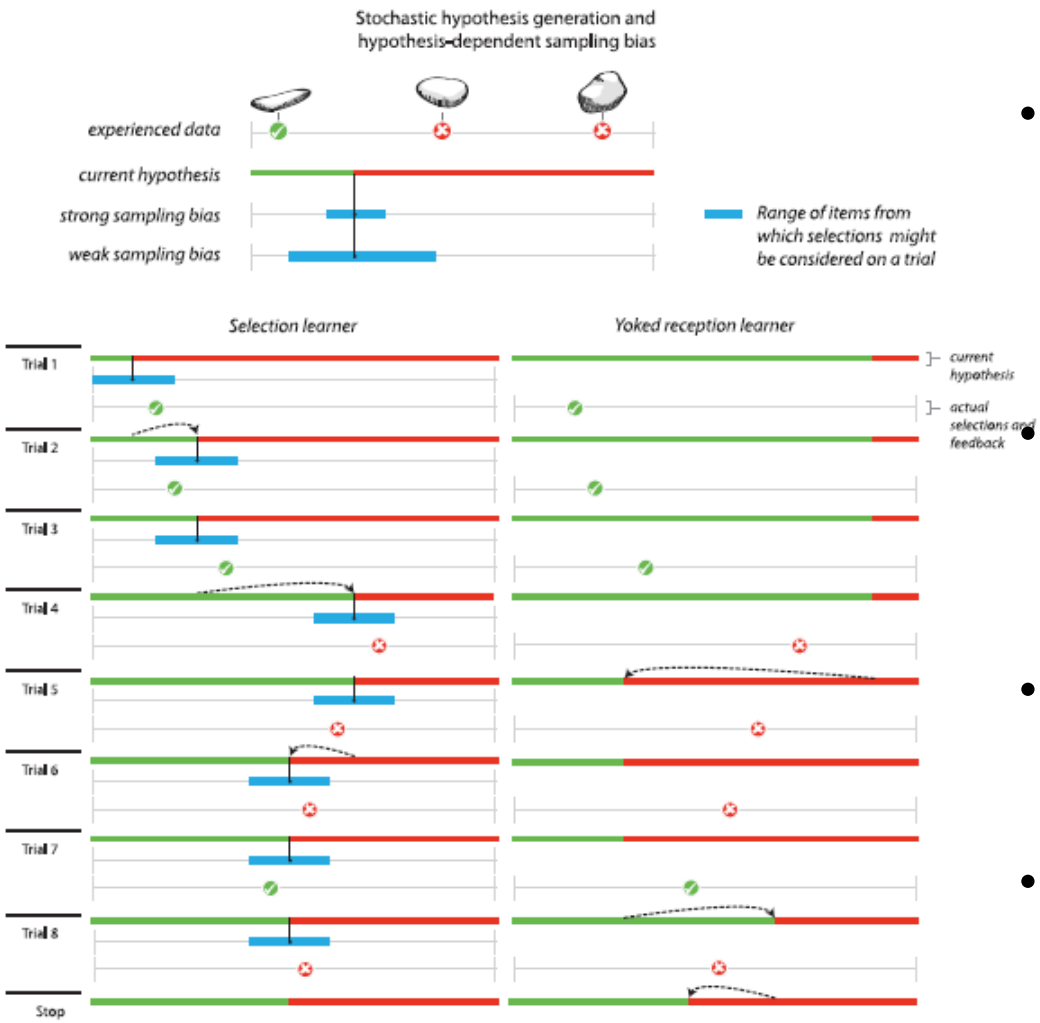
- A “yoking” design
- 240 NYU undergrads:
  - A) 60 design their own stimuli (TV aerials) and ask for the label (which station does it pick up)
  - B) 60 given randomly selected stimuli and their labels
  - C) 60 given the stimuli selected by group 1 with no explanation about origin
  - D) 60 given the stimuli selected by group 1 and told they were designed by someone trying to identify the categories



Is it better to select or to receive?

- Participants who designed their own samples sampled closer to the category boundaries than the randomly generated samples (i.e. generated more useful samples)
- Participants who designed their own samples were more accurate than those who received random samples *and* those that were yoked.
- Yoked participants do no better than those given random samples.
- Authors argue for a *hypothesis-dependent-sampling-bias…*

Hypothesis dependent sampling bias



- Heuristic model: Assumes we maintain a single hypothesis of the category boundary and shift it when we see disconfirmatory evidence
- The active learner selects cases about which they are unsure i.e. near their current category boundary.
- But this may not help the yoked learner who may have a different category boundary.
- Therefore yoked participants do not do as well as the interveners because the data does not address their personal uncertainty

A problem with yoked designs

- Very mixed results. Sobel and Kushnir (2006) and Gureckis and Markant (2013), but not Lagnado and Sloman (2004) or McCormack et al (in press).
- There are many mundane reasons why a yoked participant could perform worse or even better than the active selector e.g. engagement and the time of presentation, cognitive load etc.
- Hard to see how these can be perfectly controlled between self selectors and yokees.

Other research

- Low level: Butko & Movellan (2008). Cast visual search as an optimal information maximising Markov decision process, outperforms other approaches such as saliency maps.
- Other higher level: Oaksford and Chater (1994) do an 'active learning' analysis of the Wason card selection. Show that participants 'biased' responses could be optimally informative queries under minimal assumptions about the make up of the deck.
- Nelson et al (ongoing, but see 2014) explore an example in which greedy information is suboptimal in the long run.
- AND: In causal learning, active query selection plays a special role. We will find out about this after the break!

Half time break!

References so far:

Ball, K., Berch, D. B., Helmers, K. F., Jobe, J. B., Leveck, M. D., Marsiske, M., & ACTIVE Study Group. (2002). Effects of cognitive training interventions with older adults: a randomized controlled trial. *JAMA*, *288*(18), 2271-2281.

Baron J. (1985). Rationality and intelligence. Cambridge University Press.

Butko, N. J., & Movellan, J. R. (2008, August). I-POMDP: An infomax model of eye movement. In *Development and Learning, 2008. ICDL 2008. 7th IEEE International Conference on* (pp. 139-144). IEEE.

Gureckis, T. M., & Markant, D. B. (2012). Self-directed learning a cognitive and computational perspective. *Perspectives on Psychological Science*, *7*(5), 464-481. Nelson, McKenzie, Cottrel & Sejnowski (2010)

Hillman, C. H., Erickson, K. I., & Kramer, A. F. (2008). Be smart, exercise your heart: exercise effects on brain and cognition. *Nature Reviews Neuroscience*,*9*(1), 58-65.

Markant, D. B., & Gureckis, T. M. (2013, March 25). Is It Better to Select or to Receive?

Learning via Active and Passive Hypothesis Testing. Journal of Experimental Psychology: General (in press).

Nelson, J. D. (2005). Finding useful questions: on Bayesian diagnosticity, probability, impact, and information gain. Psychological review, 112(4), 979.

Nelson, J. D., McKenzie, C. R., Cottrell, G. W., & Sejnowski, T. J. (2010). Experience Matters Information Acquisition Optimizes Probability Gain.*Psychological science*, *21*(7), 960-969.

Nelson, J. D., Divjak, B., Gudmundsdottir, G., Martignon, L. F., & Meder, B. (2014). Children's sequential information search is sensitive to environmental probabilities. Cognition, 130(1), 74-80.

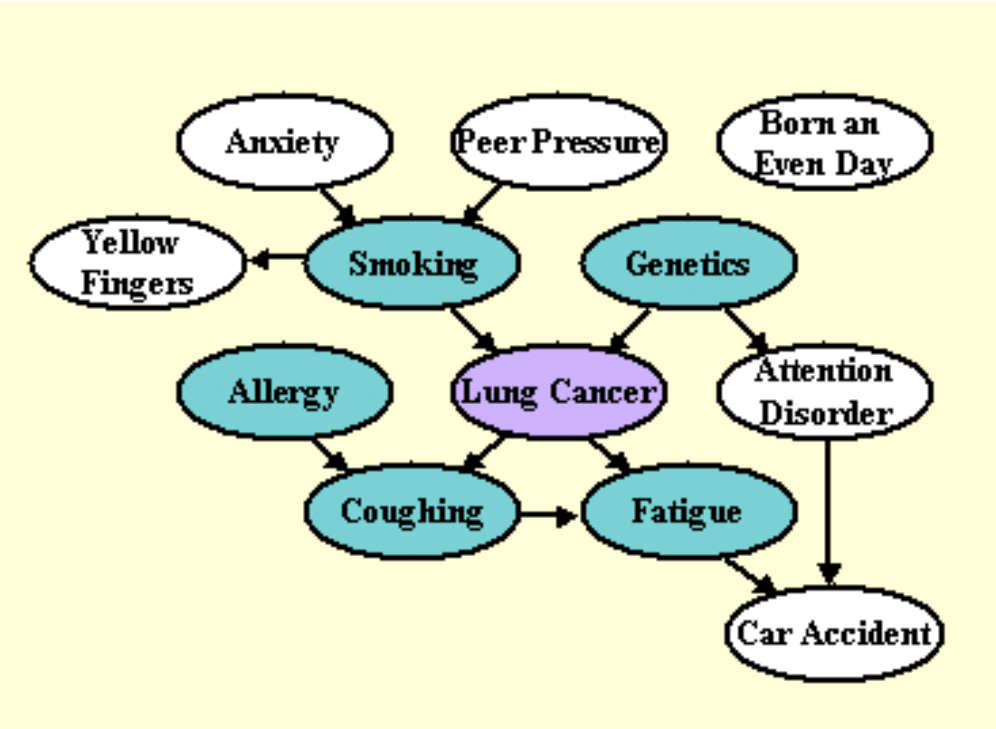
Oaksford, M., & Chater, N. (1994). A rational analysis of the selection task as optimal data selection. *Psychological Review*, *101*(4), 608.

Shannon, C. E. (2001). A mathematical theory of communication. *ACM SIGMOBILE Mobile Computing and Communications Review*, *5*(1), 3-55.

Active learning and causal structure

- Active testing plays a special role in causal learning.
- Without actively manipulating, or “intervening on” a system of interest we cannot, in general, uniquely identify its causal structure (Pearl, 2000).

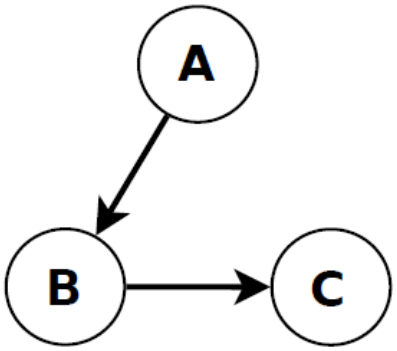
Representing causal structure

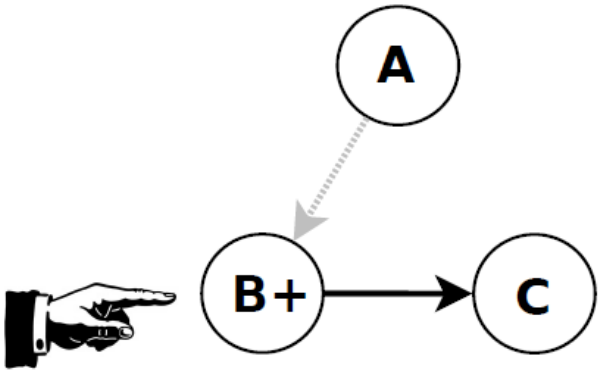


http://www.causality.inf.ethz.ch/challenge.php

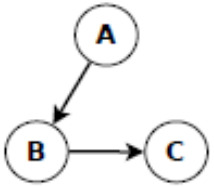
- Can formalise causal knowledge with Directed Acyclic Graphs
- Nodes represent variables of interest
- Arrows (aka 'edges') represent (potentially probabilistic) causal dependencies e.g. "smoking causes lung cancer"
- Structuring knowledge in this way facilitates forward (predictive) and backward (diagnostic) inferences about the world from observations
- Reduces complexity of inference (see next week)

What is an intervention?

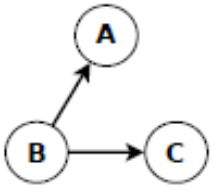




a)



b)



Motivation for current experiments

- Intervention is key to causal learning (Pearl, 2000; Sloman 2005).
- People benefit from ability to intervene during learning (e.g. Lagnado & Sloman 2002; 2004; 2006, Steyvers et al 2003).
- But how are interventions chosen and their outcomes interpreted during learning?

Models of intervention choice

- Scholar model – Selects interventions with high expected information gain(Murphy, 2001; Steyvers et al, 2003)

$$-\sum_{k \in G} p(g_k)\ log_2\ p(g_k)$$

- Gambler model - Selects interventions to increase expected probability of correctly identifying the graph (Kruschke, 2001; Nelson et al, 2010)

$$\max_{k \in G} p(g_k)$$

- Utilitarian model– Selects interventions to increase expected payment (Gureckis & Markant; 2009)

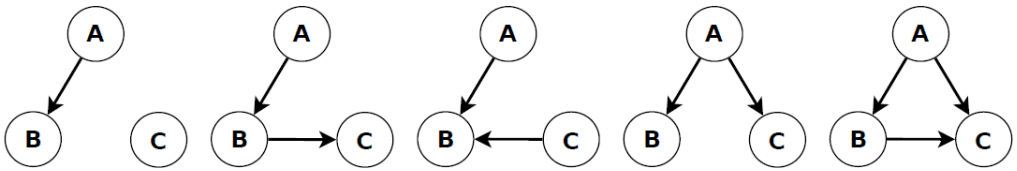
$$max_{i \in G} \sum_{j \in G} L(g_i, g_j) p(g_j)$$

Farsighted active learning?

- When learning is multi-shot, ‘myopic’ maximisation of a function of immediate posterior can be suboptimal.
- One *should* calculate many-step-ahead expected values, (assuming maximisations on later trials), to active learn optimally.
- But this becomes implausibly intractable even by Bayesian standards

The experiment

- 79 participants from Mturk learned probabilistic causal structures over multiple freely selected interventions.
- 5 test problems per participant.



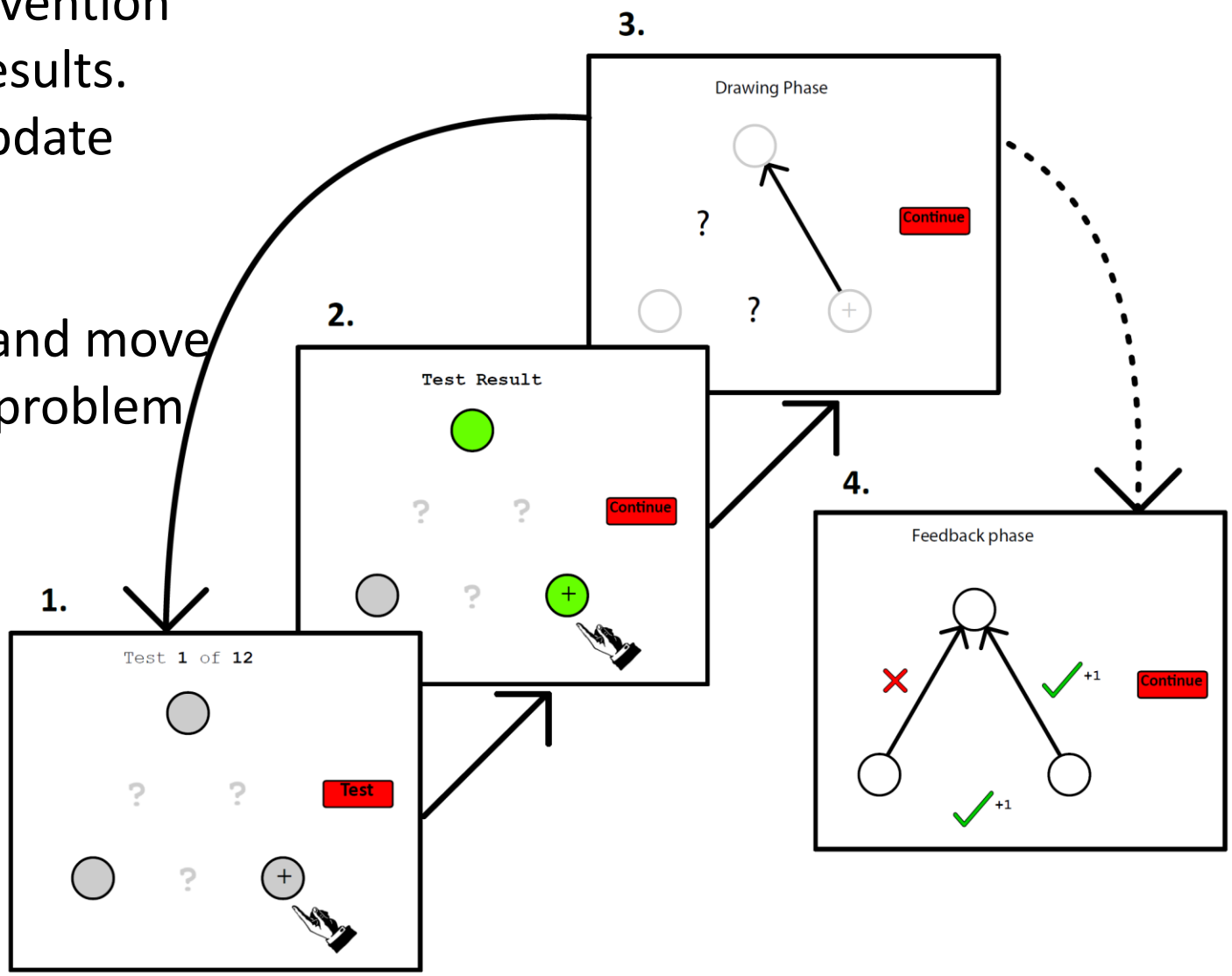
- 12 interventions per problem.
- 27 permissible interventions and 27 possible causal structures.
- Probabilistic environment (spontaneous activations=.1, causal power=.8, combination function = noisy-OR). Participants trained on these.
- Rewarded $1 + $.20 per causal link correctly identified (/3 per problem, /15 overall).

Procedure

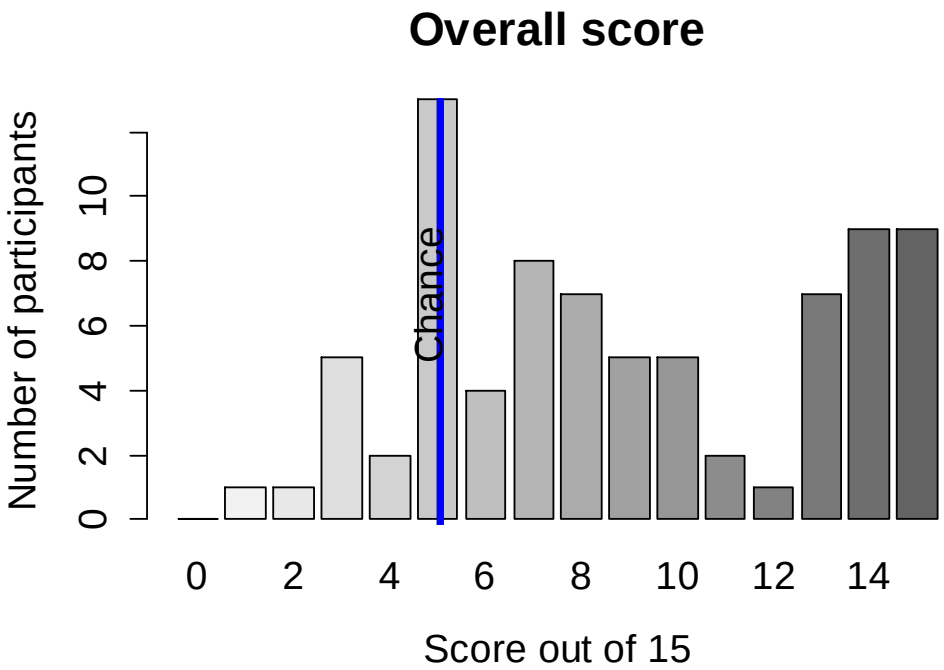
1. Select an intervention
2. Observe the results.
3. (Optionally) update marked links.

After 12 trials:

4. Get feedback and move onto the next problem

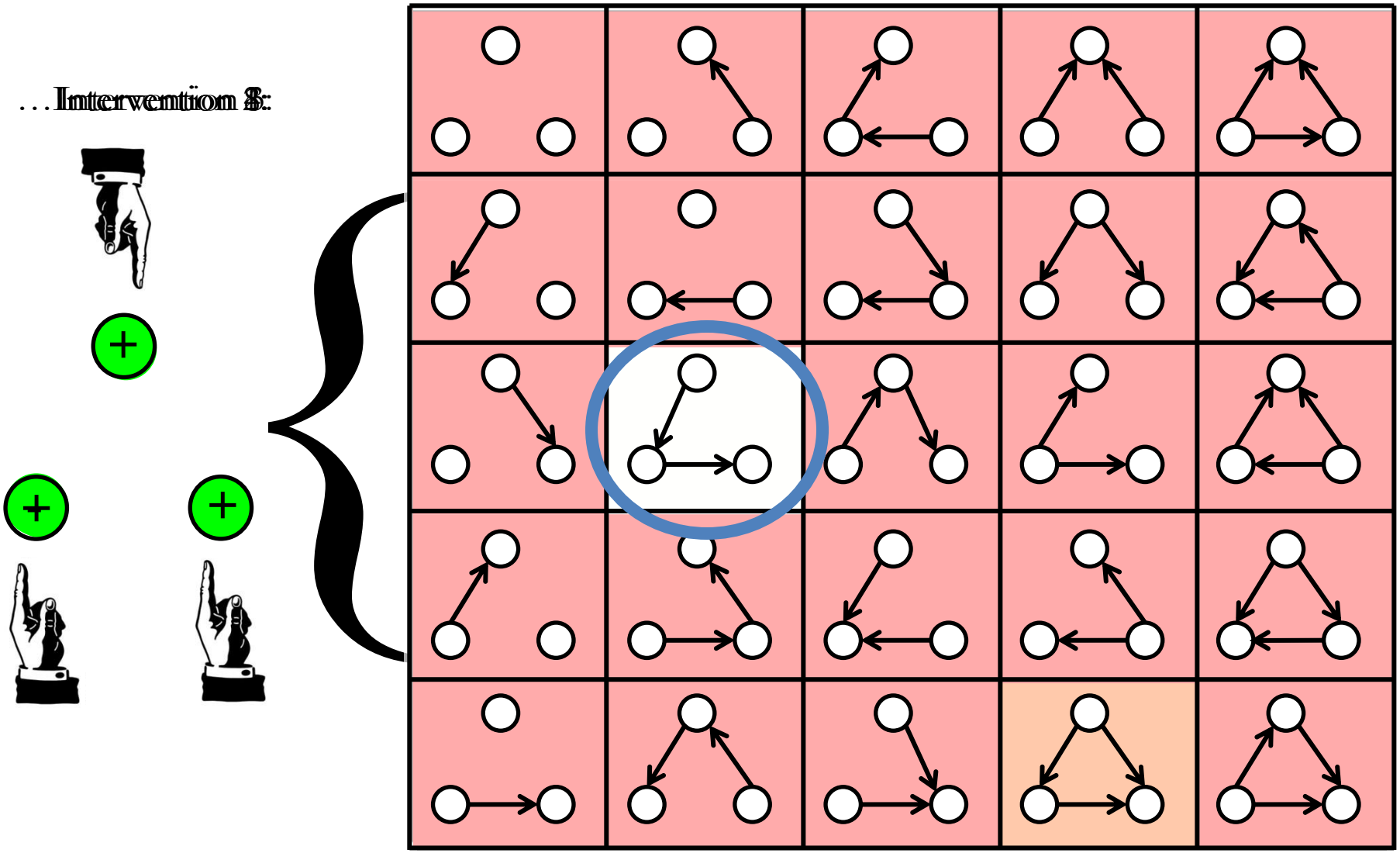


Performance – identifying the causal links



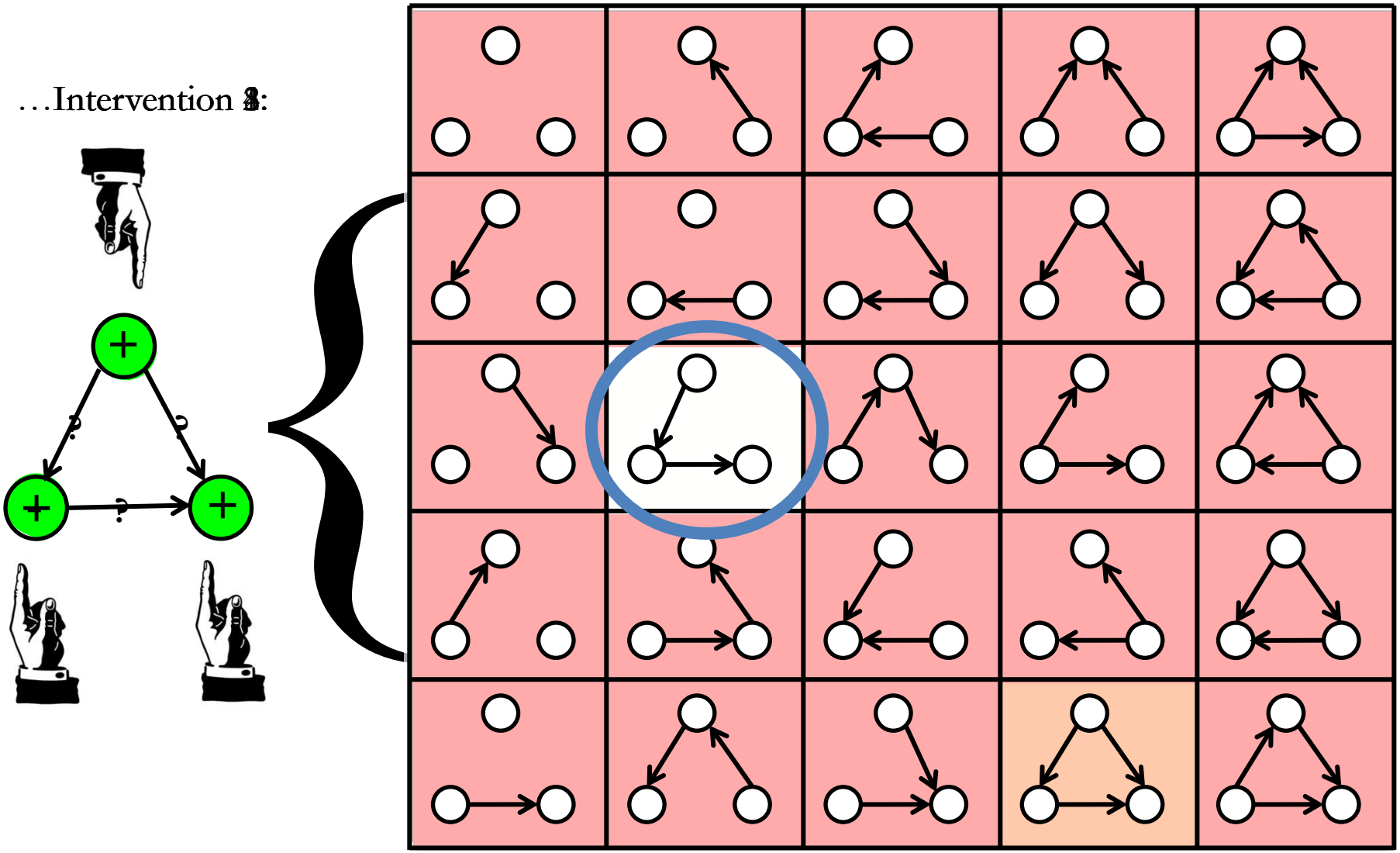
- Performances overall far above chance ($t(78) = 8.60$; $p < .0001$), average 8.97/15 links identified, chance 5.
- No difference in difficulty between problems. No improvement rate over task.
- Frequent attribution error: Mistake chain for fully connected structure 18/79 (vs. 20/79 correct).
- Single interventions most frequent (e.g. A^+). N double 'controlling' interventions (e.g. A^+, B^-) selected strongly predicts good performance ($F_{1,77} = 8.053$, $p = .0058$; +.22 points per double intervention).

Participant 5, problem 2 example:

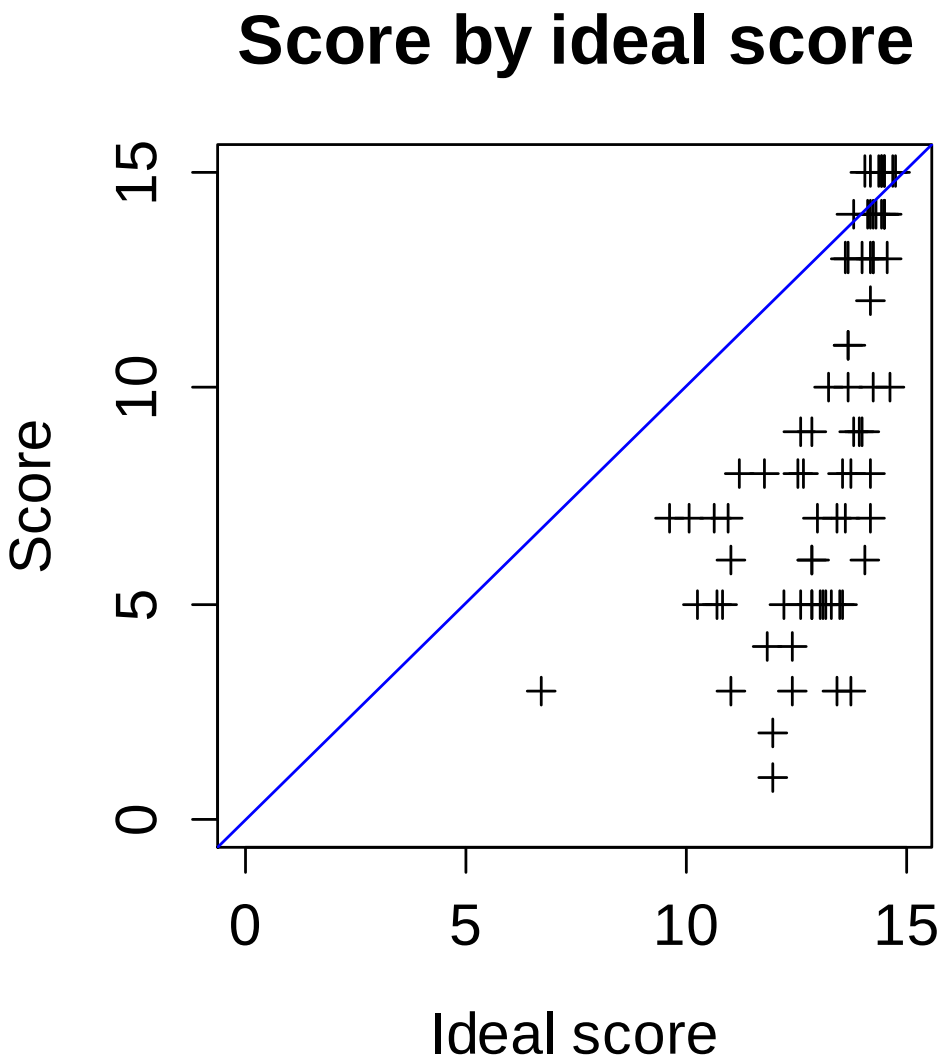


Participant 5, problem 2 example:

…Intervention 2:



Were people effective active learners?



- Intervention choices clearly above chance, average ideal score 13.15, chance 11.55 ($t(78) = 9.56$, $p<.0001$)
- Interventions average 2.77 times as informative as chance, .50 as informative as ideal learner (2.31 and .33 utility increase).

Fitting the models to participants

- Scholar, gambler and utiltarian models define the expected values $v_{1..n}$ of interventions as the *expected increase* in information/probability/utility given current objective posterior.

- Model soft maximises over values to degree α:

$$p(\mathrm{Int}_t = i) = \frac{e^{\alpha\langle v_i\rangle_t}}{\sum_{k=1}^{n} e^{\alpha\langle v_k\rangle_t}}$$

- And, where causal links are marked during learning, they are soft maximisation over posterior *p(G)* to degree β:

$$p(\text{Stated-beliefs}_t = j) = \frac{\sum_{l\epsilon j} e^{\beta q(g=l)_t}}{\sum_{k=1}^{m} e^{\beta q(g=k)_t}}$$

Initial model fits

Model	Steps-ahead	BIC	α	β	$\sharp$ par
Scholar	**1**	**45047**	**5.28**	**3.54**	**68**
Gambler	1	47917	9.09	3.54	2
Utilitarian	1	47464	5.36	3.54	1
Scholar	2	45525	3.48	3.54	6
Gambler	2	46877	11.7	3.54	2
Utilitarian	2	46778	7.69	3.54	0

- An ideal scholar's performance did not improve by looking two steps ahead in simulations, rather all three models converged to the performance of the one-step-ahead scholar model.

Bounded models

- People have limited memory and processing capacities.
- May be better described as intervening and updating under plausible memory constraints

- Can add a forgetting rate

$$P^*(G)^t = (1-\gamma)P(G)^t + \gamma\frac{1}{m}$$

- And a degree of conservatism

$$P^{**}(g_i)^t = \frac{\eta^{I[\delta_i=1]}P^*(g_i)^t}{\sum_{j=1}^{m}\eta^{I[\delta_j=1]}P^*(g_j)^t}$$

Bounded model fits

Model	BIC	α	β	λ	η	$\sharp$ par (/79)	$\sharp$ par$_{>5}$ (/57)
Forgetful Scholar	37840	7.72	23.9	.717		0	0
Forgetful Gambler	40228	40.5	77.4	.946		0	0
Forgetful Utilitarian	39224	12.2	38.1	.894		0	0
Conservative Scholar	40814	6.22	4.76		48.4	0	0
Conservative Gambler	43863	10.4	4.66		18.5	0	0
Conservative Utilitarian	43231	6.83	4.71		32.7	0	0
Conservative Forgetful Scholar	**31145**	**7.87**	**86.1**	**.909**	**2.15**	**31**	**29**
Conservative Forgetful Gambler	31575	61	374	.994	1.35	29	19
Conservative Forgetful Utilitarian	31630	13.4	245	.979	1.59	14	8

- Conservative forgetful scholar is a good overall fit (pseudo-R^2 = .44)
- Forgetting rate strongly predicts overall performance ($p<.001$, +.25 points per each 10% less forgetful.
- More forgetful people are more conservative ($p<.05$)

Heuristic models

- Simple endorsement (Fernbach & Sloman, 2009) – Clamp nodes on (e.g. A^+), one at a time, endorse direct links to any activations
- Disambiguation – Additional step, clamp one node on and all but one of the other nodes off (e.g. $A^+ B^-$), remove endorsed link to unclamped node if it does not activate

Model	BIC	θ	κ	σ	ϱ	$\sharp$ par
Simple endorser	37334	.850		.588	.233	40
Disambiguator	**36938**	.933	.982	.588	.233	39

- Less forgetful, more likely to be disambiguators (p=.016). Trend toward better performance by disambiguators (p=.11)

Conclusions

- Many people are highly effective active causal learners, able to learn fully probabilistic structures over multiple interventions.
- People act to maximise information rather than probability of being correct or utility.
- Most people act like *forgetful conservative scholars* - querying the environment to maximise information but forgetting old evidence as they go along, mitigating this by being conservative about prior conclusions.
- These patterns can also be captured by intervention attribution heuristics.

Continuing work

- Manipulating forgetting – have now run a condition in which participants see a history of past interventions. Comparing with this allows us to identify better to what extent memory is a bottleneck here
- Identifying boundary conditions of effective active causal learning – how many nodes, how much noise before people cannot select useful interventions/cannot learn causal structure?
- Loops and cycles – Bayes nets cannot handle causal structures with feedback but these are pervasive in the world. Currently designing experiments to probe how people interpret data generated by loopy causal systems and how the learn these.

Summing up

- First hour
  - Active query selection is a key part of human learning, and studying it can shed light on representations and processes
  - We can quantify query usefulness using measures including information gain, probability gain
  - Experiments have suggested people benefit from choosing their queries, they seem to maximise probability gain in some contexts (i.e. one shot cateorisation), and be insensitive to costs.
- Second hour
  - Active learning plays a special role in causal learning where is an essential for uniquely identifing structure
  - People appear to optimise information gain in this context, but in a somewhat bounded way

Thanks for listening!

Active causal learning references:

Fernbach, P. M., & Sloman, S. A. (2009). Causal learning with local computations. *Journal of experimental psychology: Learning, memory, and cognition*, *35*(3), 678.

Lagnado, D. A., & Sloman, S. (2004). The advantage of timely intervention. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *30*(4), 856.

Lagnado, D. A., & Sloman, S. A. (2006). Time as a guide to cause. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *32*(3), 451.

Murphy, K. P. (2001). Active learning of causal Bayes net structure.

Pearl, J. (2000). *Causality: models, reasoning and inference* (Vol. 29). Cambridge: MIT press.

Sloman, S. (2009). *Causal models: How people think about the world and its alternatives*. Oxford University Press.

Steyvers, M., Tenenbaum, J. B., Wagenmakers, E. J., & Blum, B. (2003). Inferring causal networks from observations and interventions. *Cognitive science*, *27*(3), 453-489.

Collaborators:

David Lagnado

Maarten Speekenbrink

Jonathan Nelson