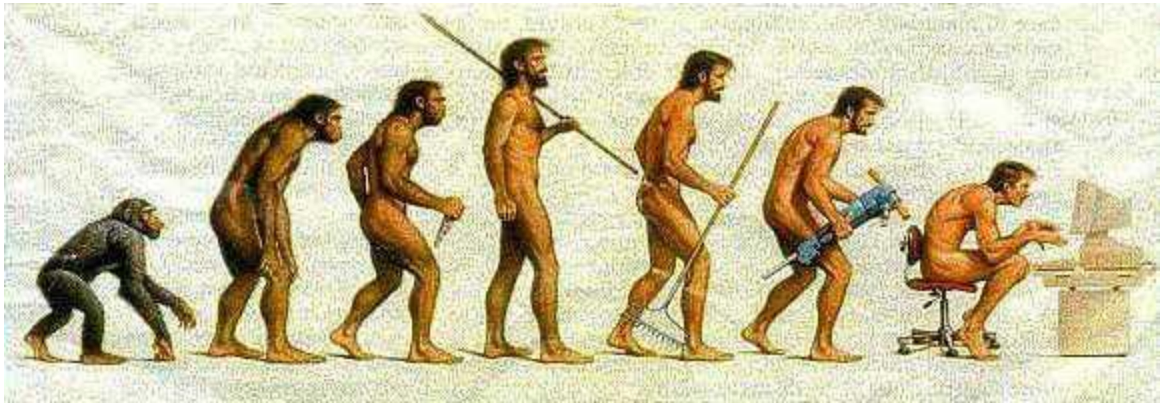
HAMLET

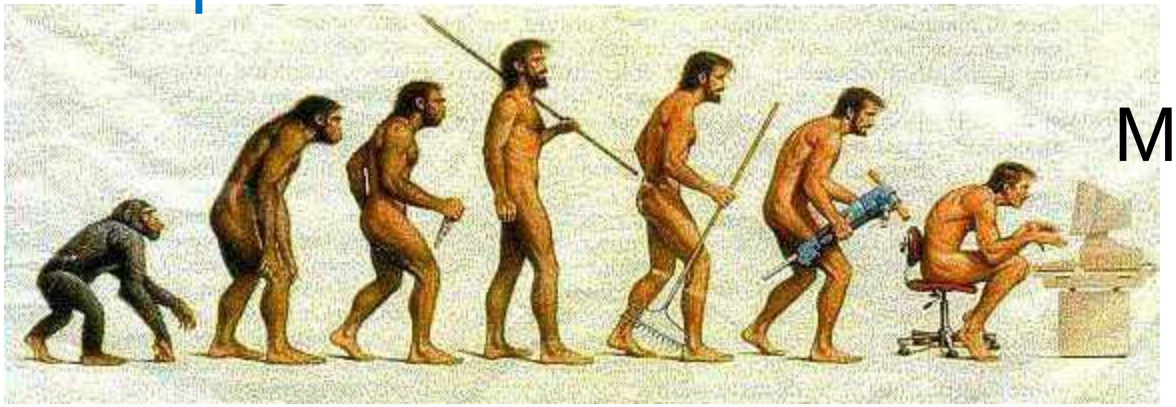
JERRY ZHU
UNIVERSITY OF WISCONSIN

Collaborators:
Rui Castro, Michael Coen, Ricki Colman,
Charles Kalish, Joseph Kemnitz, Robert Nowak,
Ruichen Qian, Shelley Prudom, Timothy Rogers



Somewhere, something went terribly wrong.

Learning: improve with experience



Machines

Animals

Humans

Theory: common mathematical principles

Experiments: behavioral study, computer simulation

Machine Learning + Cognition

Three new case studies of common learning principles in humans, animals and machines:

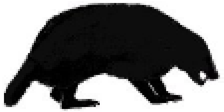
1. Human semi-supervised learning
2. Human active learning
3. Monkey online learning

HAMLET example #1

Human Semi-Supervised Learning

The first work that quantitatively studied human's ability to utilize both labeled and unlabeled data in concept forming.

A Camping Story



badger

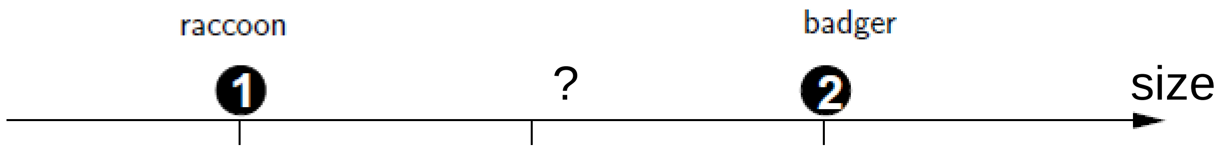
A Camping Story

raccoon

A Camping Story

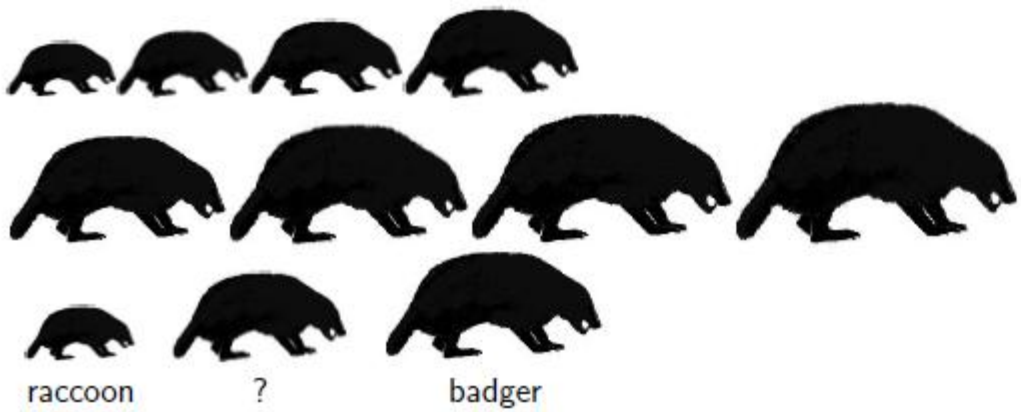
?

Supervised Learning

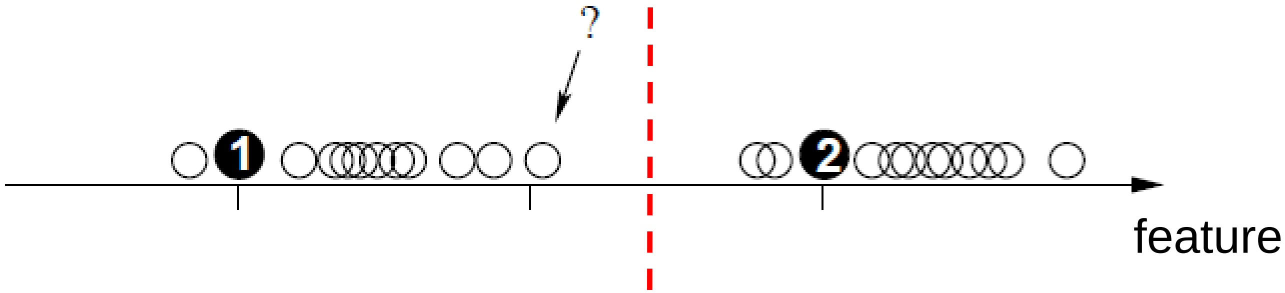


- $x \in \Re^D$: Input item = stimulus = feature vector
- $y \in \{1, 2\}$: class label = category
- Supervised learning: given labeled training examples $(x_1,y_1)\ldots(x_n, y_n)$, learn a classifier $f: X \rightarrow Y$
- In this example, decision boundary is in the middle

Back to the Camp

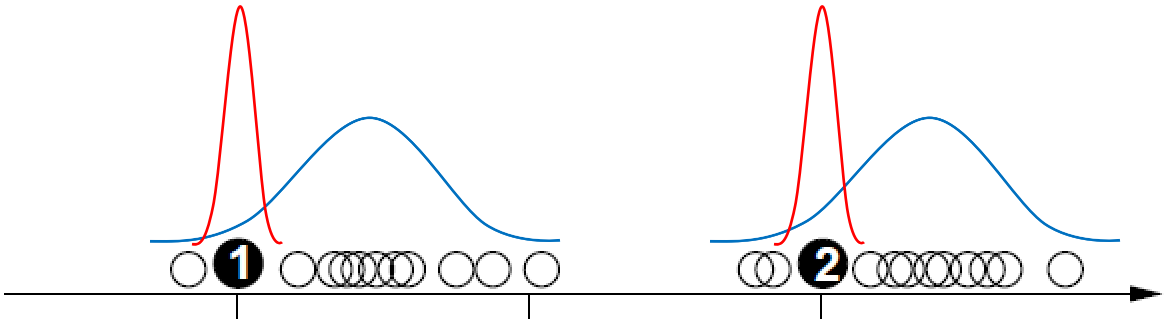


Semi-Supervised Learning



- Semi-supervised learning (SSL): given labeled examples $(x_1,y_1)\ldots(x_n, y_n)$ and unlabeled examples $x_{n+1}\ldots x_{n+m}$ learn a better classifier f: X→Y
- The cluster assumption (one of many assumptions)
- SSL well-studied in machine learning
  - IBM: Vikas Sindhwani

SSL with Gaussian Mixtures



- p(x) is a Gaussian mixtur $w_1 N(\mu_1, \sigma_1^2) + w_2 N(\mu_2, \sigma_2^2)$
- Parameters $\theta = \{w_1, \mu_1, \sigma_1^2, w_2, \mu_2, \sigma_2^2\}$
- p(y|x) from Bayes rule $\frac{w_y \mathrm{N}(x;\mu_y,\sigma_y^2)}{\sum_{k=1,2} w_k \mathrm{N}(x;\mu_k,\sigma_k^2)}$
- Parameter estimation over labeled data (easy)
- Parameter estimation over both labeled and unlabeled data (EM algorithm)

SSL with Gaussian Mixtures

- Prior on parameters:

$$w_k \sim \mathrm{Uniform}[0,1], \mu_k \sim \mathrm{N}(0,\infty), \sigma_k^2 \sim \mathrm{Inv}-\chi^2(\nu, s^2), k = 1,2$$

- Maximize objective

$$\log p(\theta) + \sum_{i=1}^{l} \log p(x_i, y_i|\theta) + \lambda \sum_{i=l+1}^{n} \log p(x_i|\theta)$$

E-step

$$q_i(k) \propto w_k \mathrm{N}(x_i; \mu_k, \sigma_k^2), \quad i = l+1, \ldots, n; k = 1,2$$

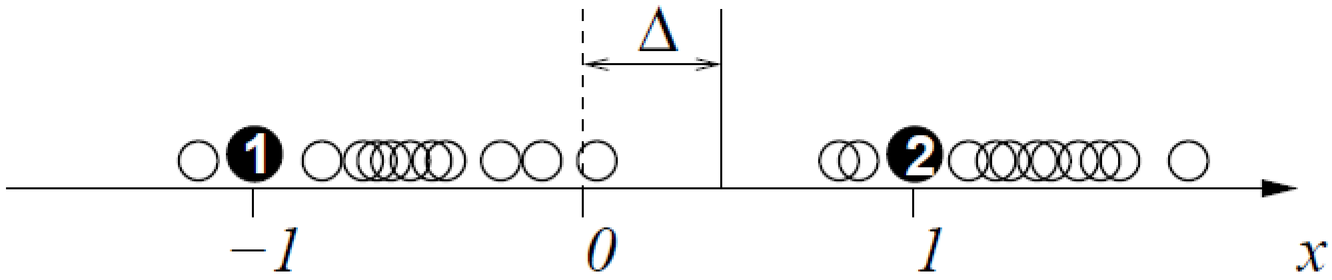
M-step

$$\mu_k = \frac{\sum_{i=1}^{l} \delta(y_i,k) x_i + \lambda \sum_{i=l+1}^{n} q_i(k) x_i}{\sum_{i=1}^{l} \delta(y_i,k) + \lambda \sum_{i=l+1}^{n} q_i(k)}$$

$$\sigma_k^2 = \frac{\nu s^2 + \sum_{i=1}^{l} \delta(y_i,k) e_{ik} + \lambda \sum_{i=l+1}^{n} q_i(k) e_{ik}}{\nu + 2 + \sum_{i=1}^{l} \delta(y_i,k) + \lambda \sum_{i=l+1}^{n} q_i(k)}$$

$$w_k = \frac{\sum_{i=1}^{l} \delta(y_i,k) + \lambda \sum_{i=l+1}^{n} q_i(k)}{l + \lambda(n-l)}$$

Human Semi-Supervised Learning



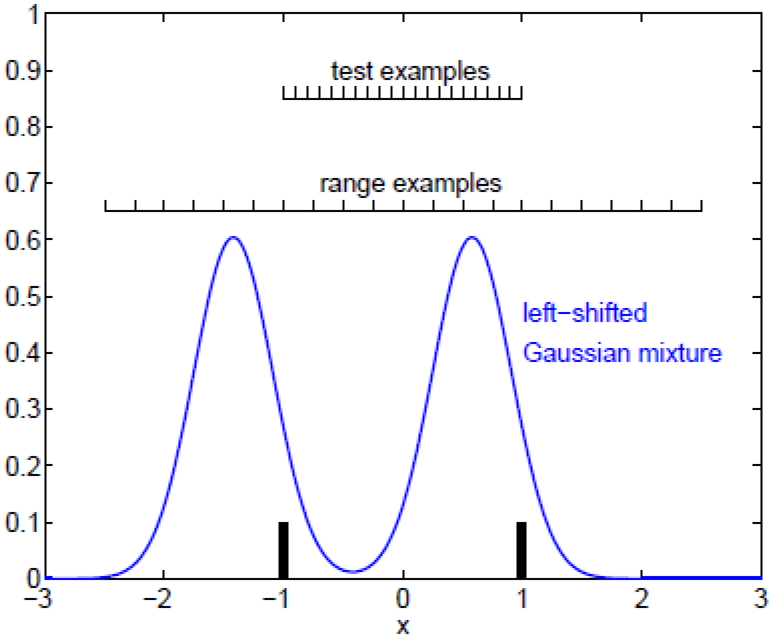
- Machine learning predicts decision boundary shift
- Do humans “do” semi-supervised learning?
  - we are immersed in unlabeled data in supervised tasks (e.g., deciding luggage/bomb)

Materials and Subject

- Stimuli x parameterized in 1D, displayed on screen one at a time

- Label y: 2-way forced choice.
  - Labeled data: audio feedback.
  - Unlabeled data: no audio feedback.
- 22 subjects, two conditions: L and R

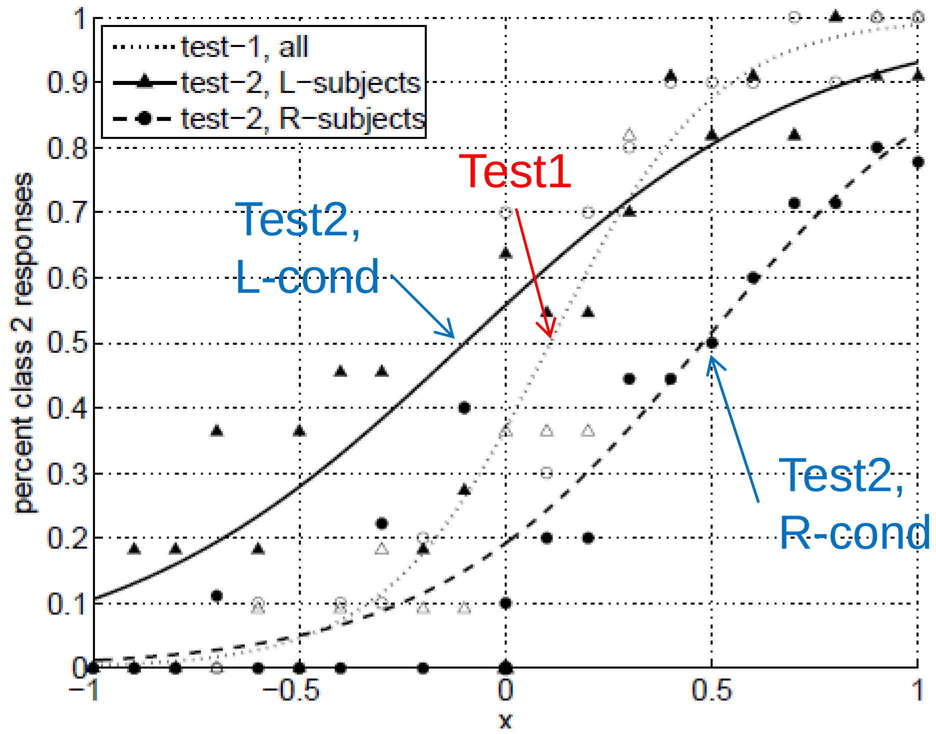
Procedure



1. 20 labeled instances
   10 each: (-1,-), (1,+),
   random order (ditto)
2. Test1: x=-1, -0.9, …
   0.9, 1
3. 690 unlabeled instances sampled from the blue bi-modal distribution, Left- or Right-shifted. Also range examples.
4. Test2: x=-1, -0.9, …, 0.9, 1

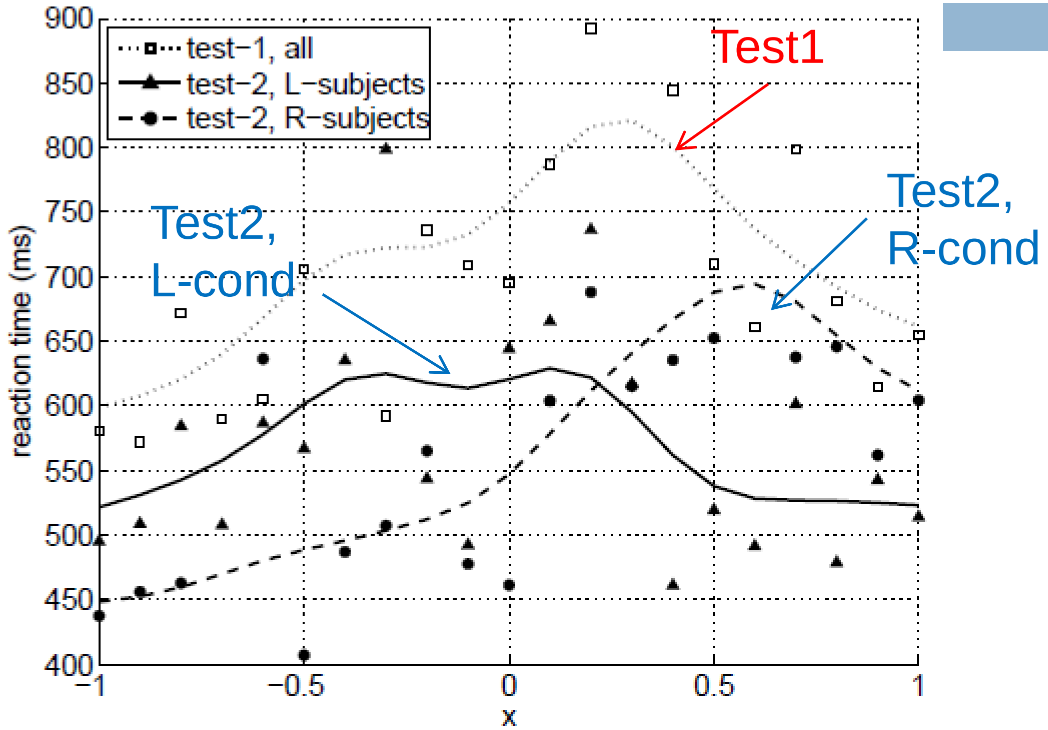
Results: Decision Boundaries

Prob(y=+|x)



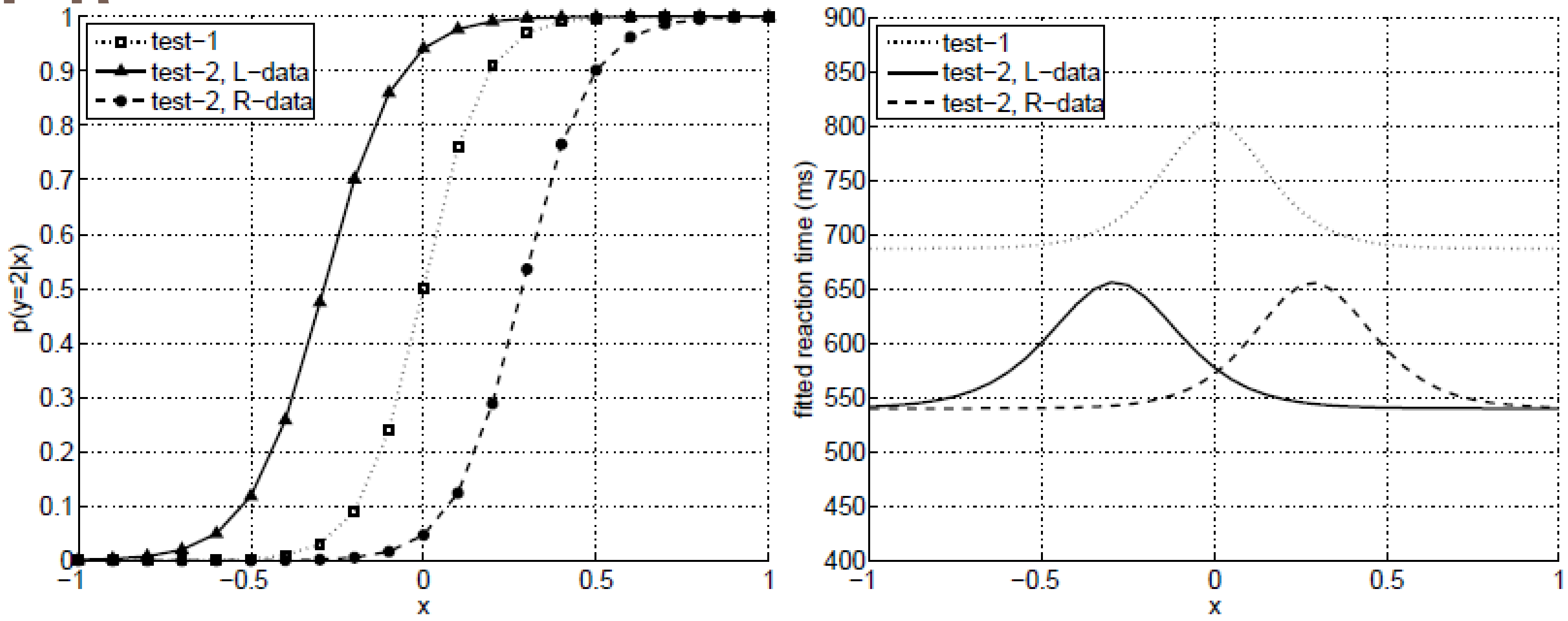
- Human decision boundaries shift after seeing unlabeled data.

Results: Reaction Time



- Peak of reaction time shifts accordingly

SSL Machine Learning Model Fit

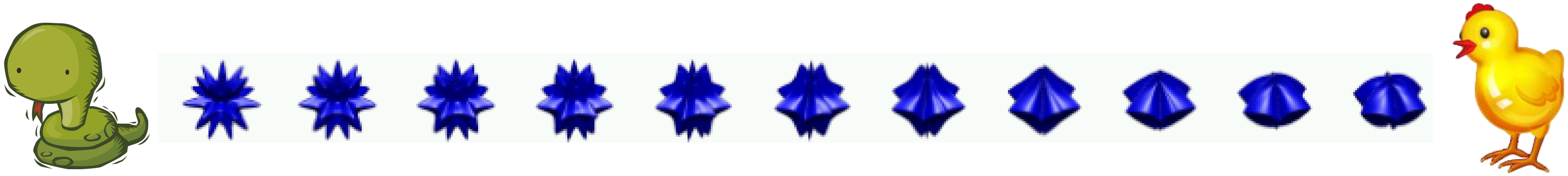


- Prediction of the Gaussian Mixture Model
- The same labeled and unlabeled input, parameters learned with the EM algorithm
- Reaction time modeled as RT = a * Entropy(p(y|x)) + b

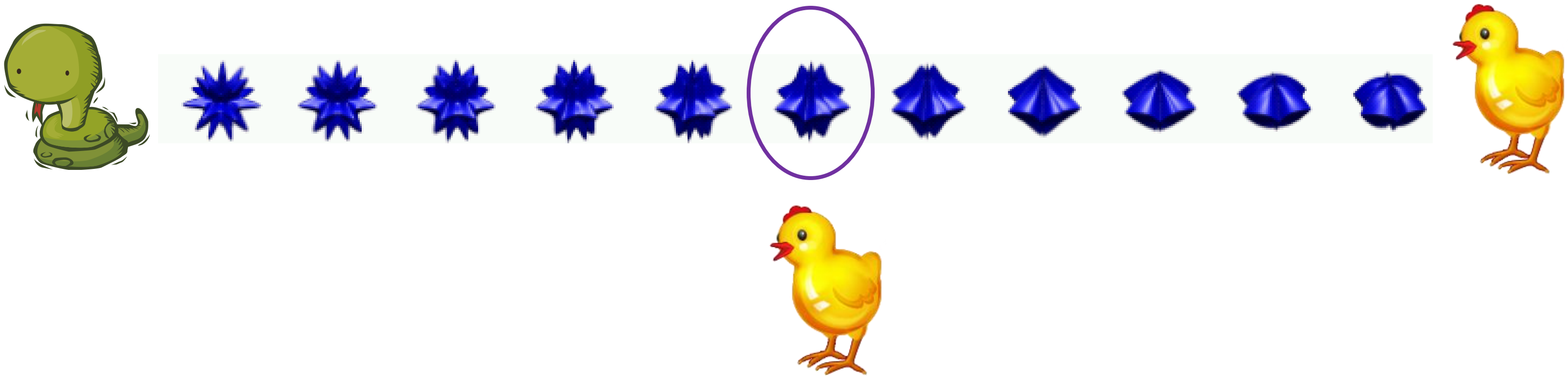
Human Active Learning

The first work that quantitatively studied human's ability to actively select good queries in category learning.

Alien Eggs

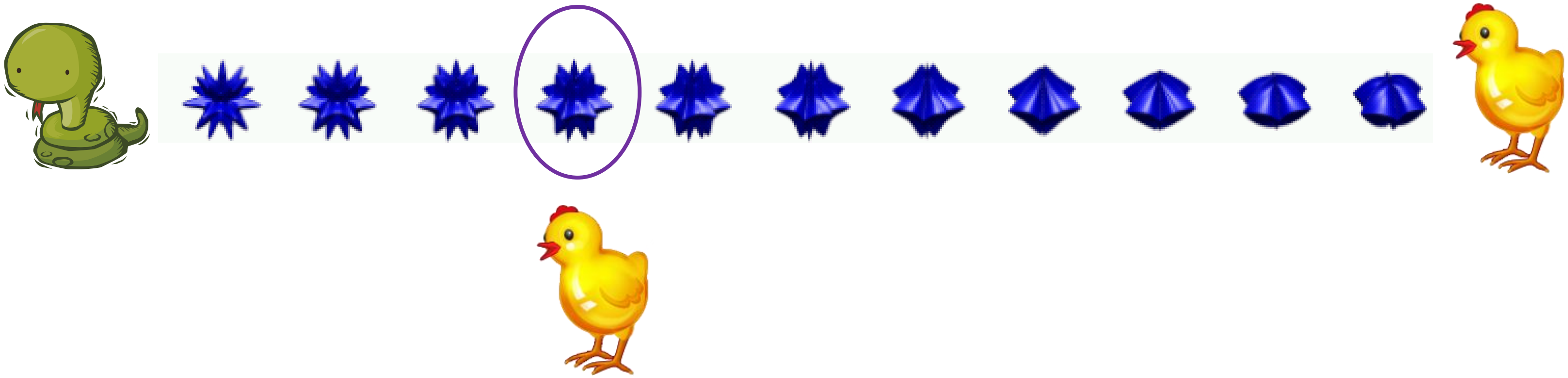


Alien Eggs



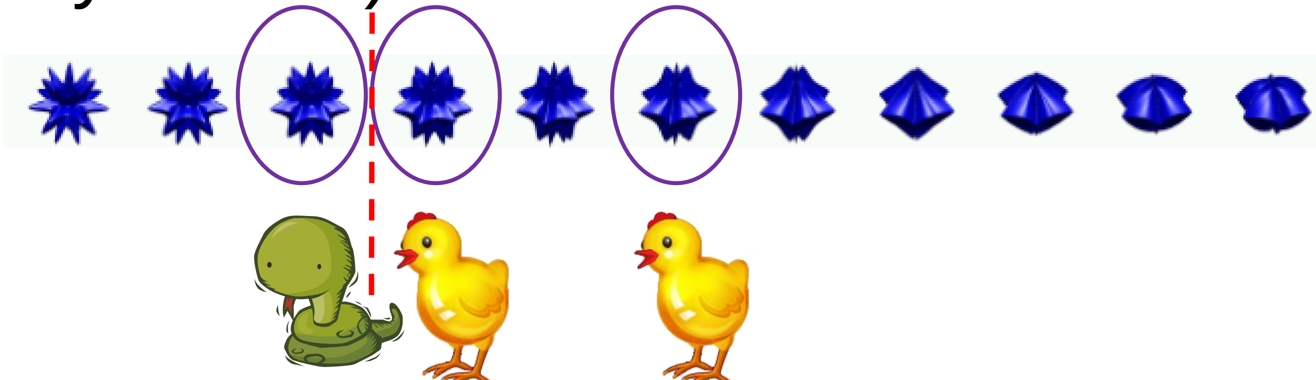
Alien Eggs

Alien Eggs

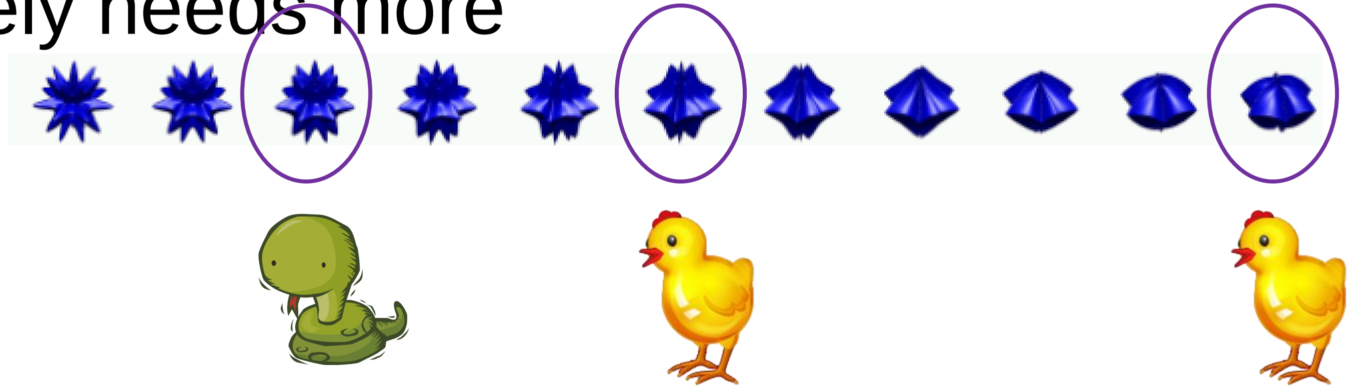


Alien Eggs

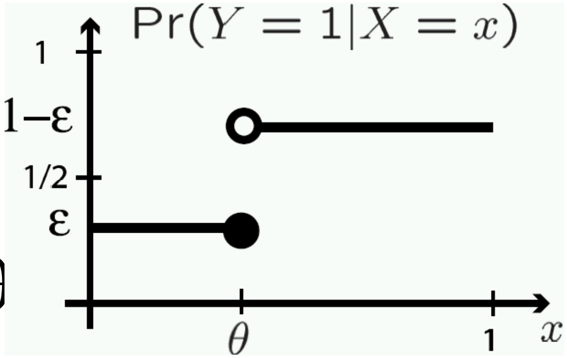
- Active learning required 3 queries (in this case binary search)



- Passive learning with i.i.d. training examples likely needs more



The Learning Task



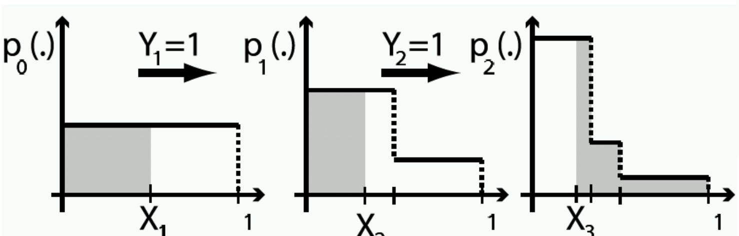
- 1D feature x
- Two classes y
- Unknown but fixed boundary θ
- Label noise ε (no more binary search!)
- Goal: learn θ from training data $(x_1,y_1)\ldots(x_n, y_n)$
- Major difference in how $x_1\ldots x_n$ are chosen
  - Passive learning: x i.i.d. (in this case from uniform[0,1])
  - Active learning: at iteration i, learner selects x_i

Learning-Theoretic Error Bounds

- Passive learning: with n random training examples, the minimax lower bound for boundary estimation error decreases polyn

$$\inf_{\hat{\theta}_n} \sup_{\theta\in[0,1]} \mathbb{E}[|\hat{\theta}_n - \theta|] \geq \frac{1}{4}\left(\frac{1+2\epsilon}{1-2\epsilon}\right)^{2\epsilon} \frac{1}{n+1}$$

- Active learning: there is a probabilistic bisecting algorithm for which the boundary



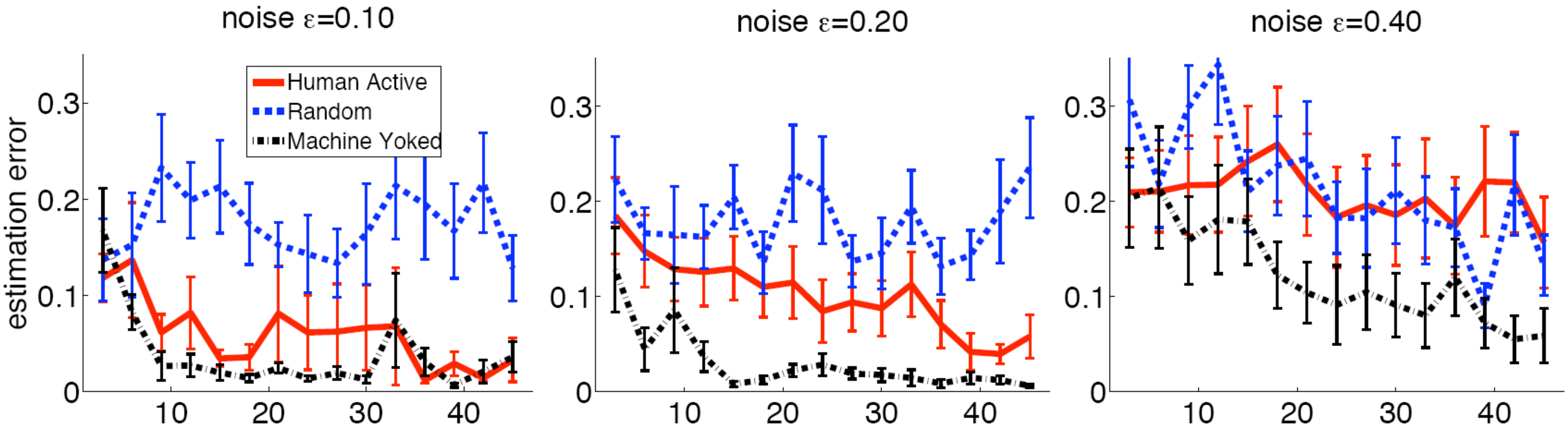
$$\sup_{\theta\in[0,1]} \mathbb{E}[|\hat{\theta}_n - \theta|] \leq 2\left(\sqrt{\frac{1}{2} + \sqrt{\epsilon(1-\epsilon)}}\right)^n$$

Human Active Learning

- 33 subjects randomly divided into three conditions
  - Random (passive): subject receives i.i.d. (x,y) examples
  - Active: subject use mouse scroll to choose x, receives y
  - Yoked: subject receives x chosen by machine active learning algorithm, and its y, as if the machine is teaching the human.
- 5 sessions of 45 iterations, with different θ, ε
- Report boundary guess every 3 iterations.

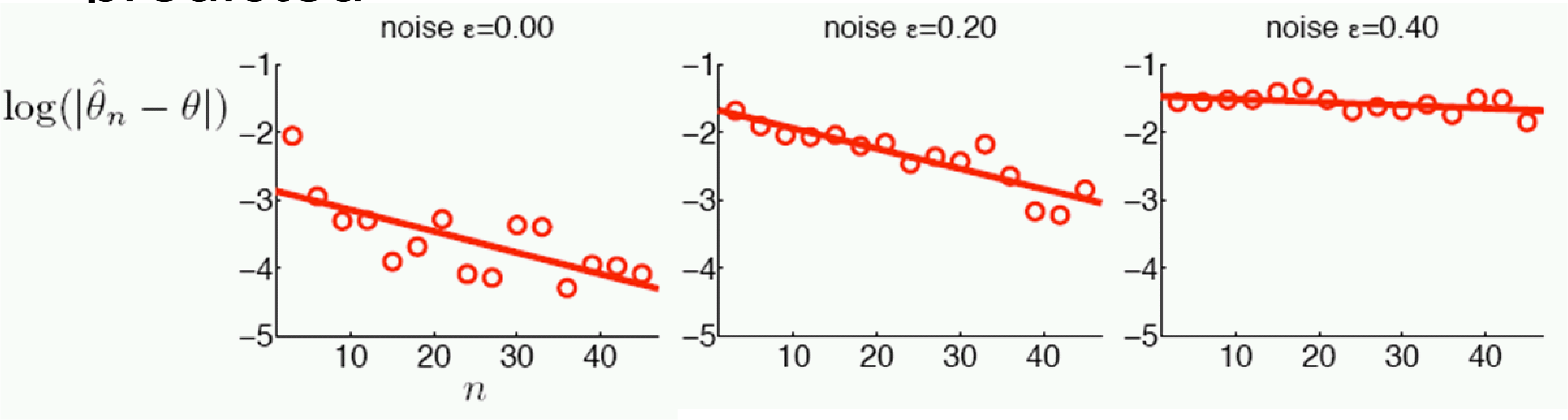
Results

- Human active learning better than passive
- Noise makes human learning difficult



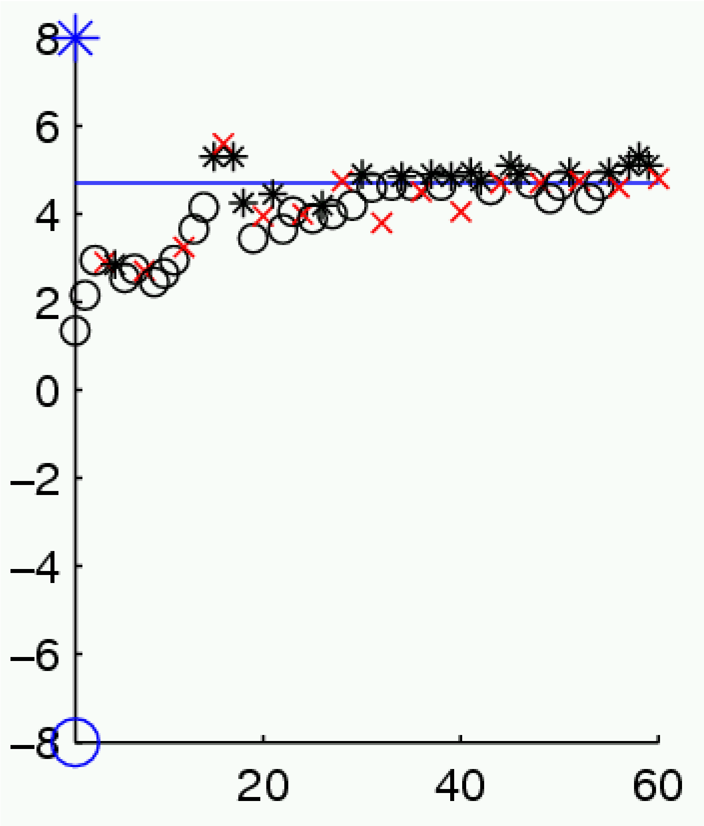
Results

- Human active learning decreases error exponentially, as learning theory predicts
- However, the decay constant is smaller than predicted

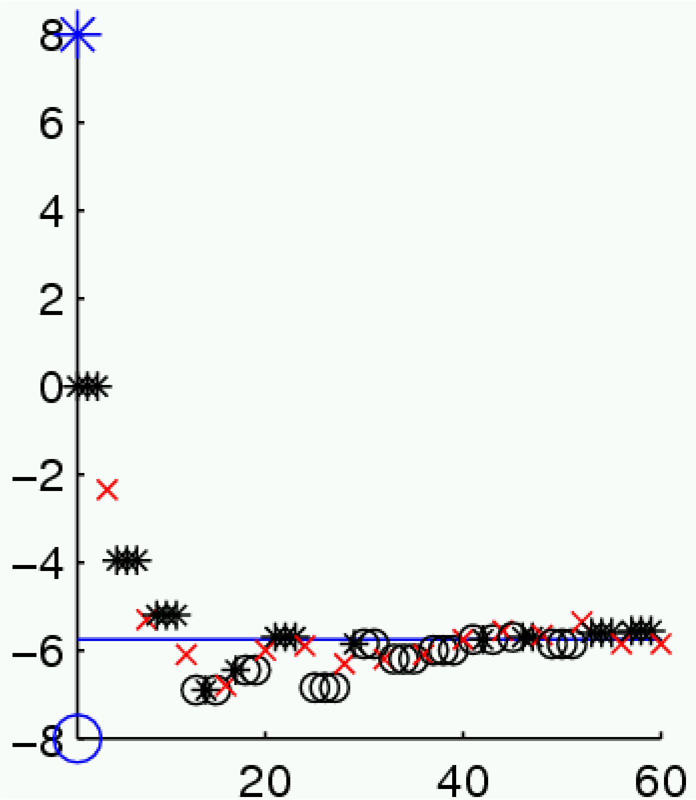


Human Active Strategies

"nudge"



"just to be sure"

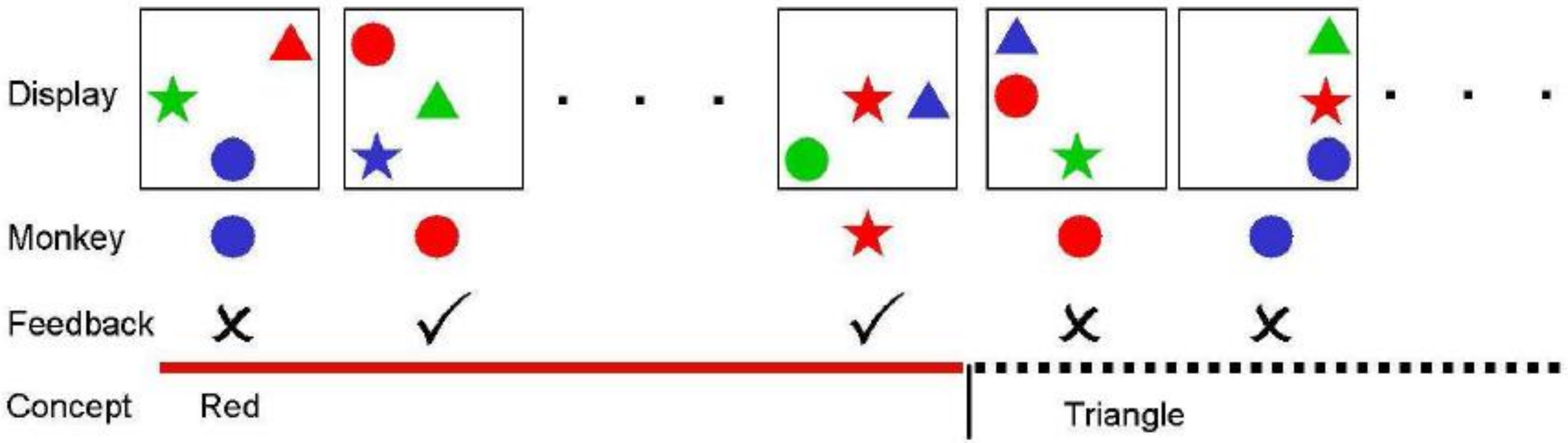


HAMLET example #3

Monkey Online Learning

Faced with an adversary, why do monkeys behave so differently than an online learning algorithm?

Wisconsin Card Sort Task (WCST)



- Three shapes, three colors on each screen
- Initial target concept: “red”, shape irrelevant
- After 10 consecutive correct trials, concept drifts to “triangle” (later to “Blue”, and “Star”)
- How should a learner adjust?

Online Learning Against an Adversary

- Each object *x* has *d*=6 Boolean features (R,G,B,C,S,T).
- Repeat
  - Adversary presents *3* objects, each with two features on (e.g., Red Circle)
  - Adversary can change the taget concept before seeing learner's pick
  - learner picks one, adversary says yes/no
- Want: the number of mistakes not too larger than the number of concept drifts.

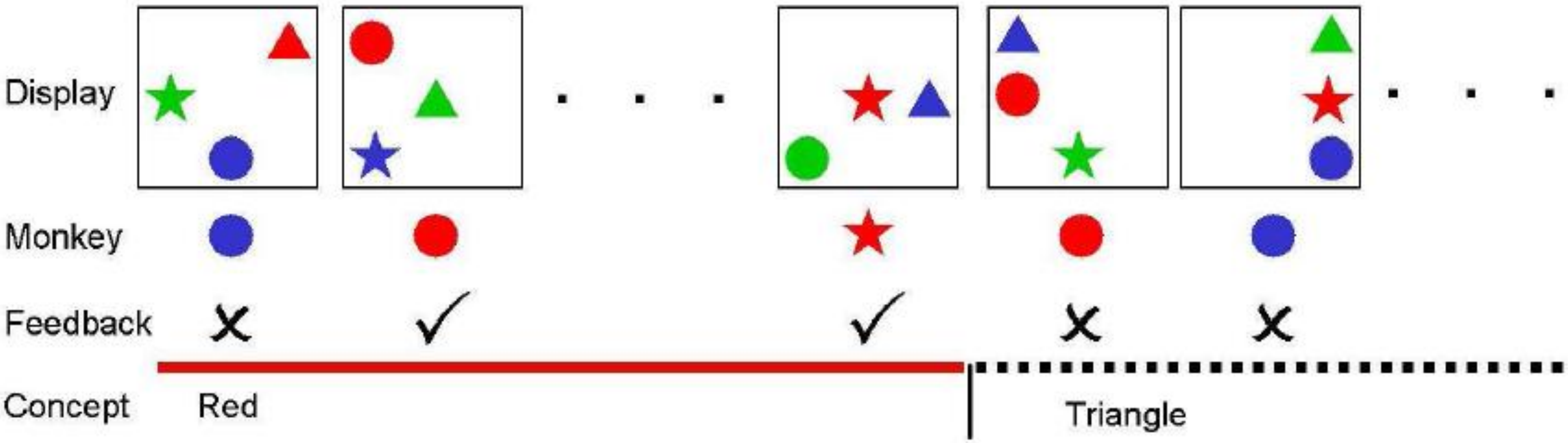
An Online Learning Algorithm

1. Initially $h = (1 \dots 1)$ (d ones). Repeat 2–4:
2. Randomly pick $x \in \{x_1 \dots x_{d/2}\}$ for which $h \wedge x \neq 0$
3. If x is correct, $h = h \wedge x$.
4. If x is wrong, $h = h \wedge \neg x$. If $h = 0$, reset $h = (1 \dots 1)$.

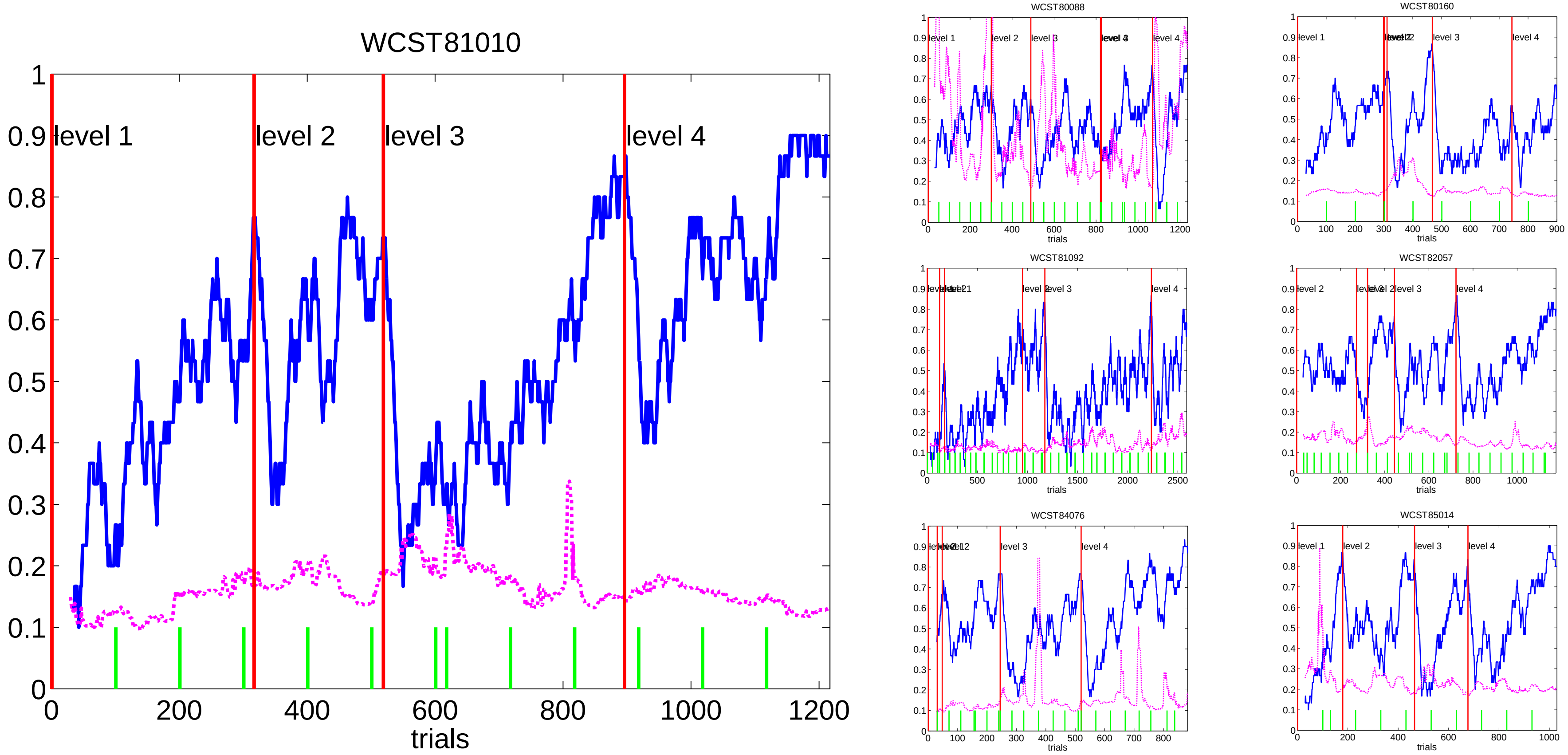
- **Theorem:** For any input sequence with *m* concept drifts, the algorithm makes at most (2*m* + 1)(*d* − 1) mistakes.
- Specifically, the bound is 35 (*m*=3, *d*=6).
- In practice, only 2 to 4 errors per concept drift.

Monkeys Play WCST

- 7 Rhesus monkeys on diet
- Touch screen
- Food pellet reward for touching target concept



Results



Accuracy (30-trial average)

Reaction time (x10 seconds)

Session

Results

	trials	errors	persv
Red	425	242	-
Triangle	249	113	89
Blue	437	247	186
Star	279	132	94

- Monkeys adapt to concept drifts slowly: ~300 trials
- Perservarative error (what would be correct under the previous concept) dominates at 75%
- No “slow down” after concept drifts: do they realize the change?

A Few Lessons Learned

(warning: highly subjective and speculative)

Lessons for Machine Learning

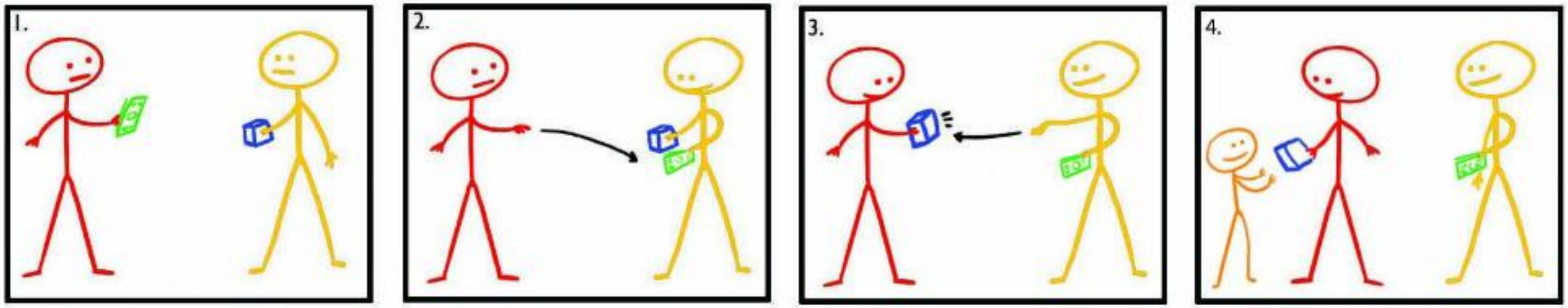
1. Difficulty: Monkeys > Undergrads > Computers
2. There is no train/test split. People always learn and adapt, even on “test data”.
3. Strong sparsity. People focus on one feature.
4. Motivation. Non-diet monkeys refuse to learn.
5. Making existing ML algorithms dumber to explain natural learning is not very interesting.
6. ML should look for things it currently cannot

References

1. Xiaojin Zhu and Andrew Goldberg. **Introduction to Semi-Supervised Learning**. Morgan-Claypool, 2009 (to appear).
2. Xiaojin Zhu, Timothy Rogers, Ruichen Qian, and Chuck Kalish. **Humans perform semi-supervised classification too.** In *Twenty-Second AAAI Conference on Artificial Intelligence (AAAI-07)*, 2007.
3. Xiaojin Zhu. **Semi-supervised learning literature survey**. Technical Report 1530, Department of Computer Sciences, University of Wisconsin, Madison, 2005.
4. Rui Castro, Charles Kalish, Robert Nowak, Ruichen Qian, Timothy Rogers, and Xiaojin Zhu. **Human active learning.** In *Advances in Neural Information Processing Systems (NIPS) 22*, 2008.
5. Xiaojin Zhu, Michael Coen, Shelley Prudom, Ricki Colman, and Joseph Kemnitz. **Online learning in monkeys.** In *Twenty-Third AAAI Conference on Artificial Intelligence (AAAI-08)*, 2008.

Some Other Work

- Multi-manifold, online semi-supervised learning
- Learning bigram LM from unigram bag-of-words
- New year's wishes



Conclusion

Machine learning and cognitive science have much to offer to each other.

Thank you

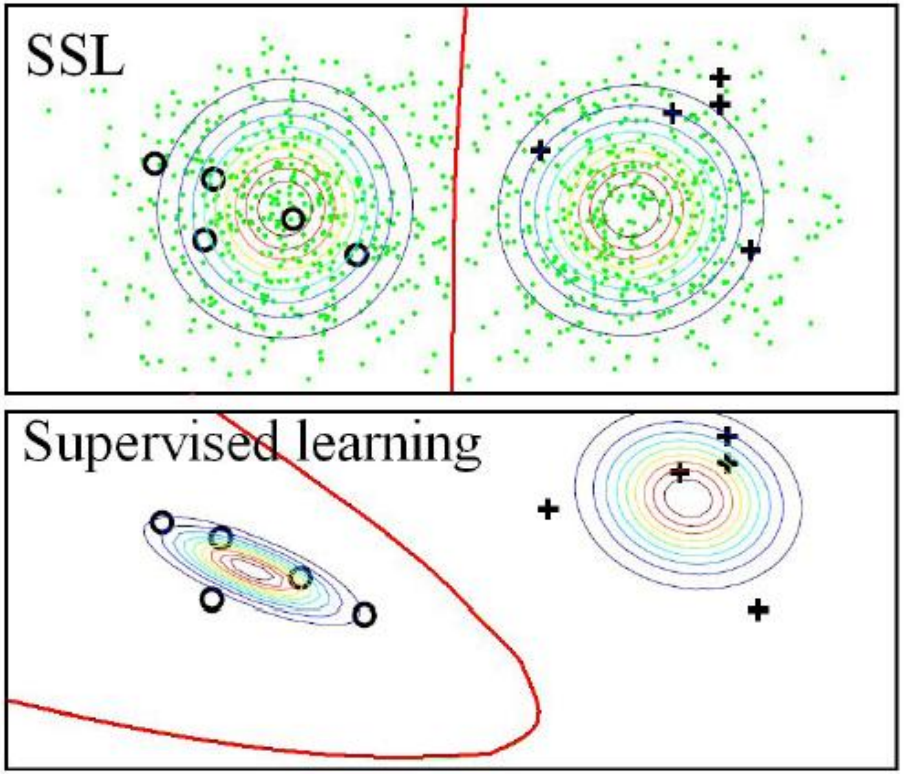
What’s in a Name

- A feature → a dimension
- Instance x (feature vector, point in feature space)→ a stimulus (continuous in this talk; discrete possible)
- Label y → a category (two categories in this talk; multiple categories, or a continuous prediction possible)
- Classification → concept/category learning
- Labeled data → supervised experience (e.g., explicit instructions) from a teacher
- Unlabeled data → passive experiences (including, but not limited to, test instances – be careful)

Learning Paradigms

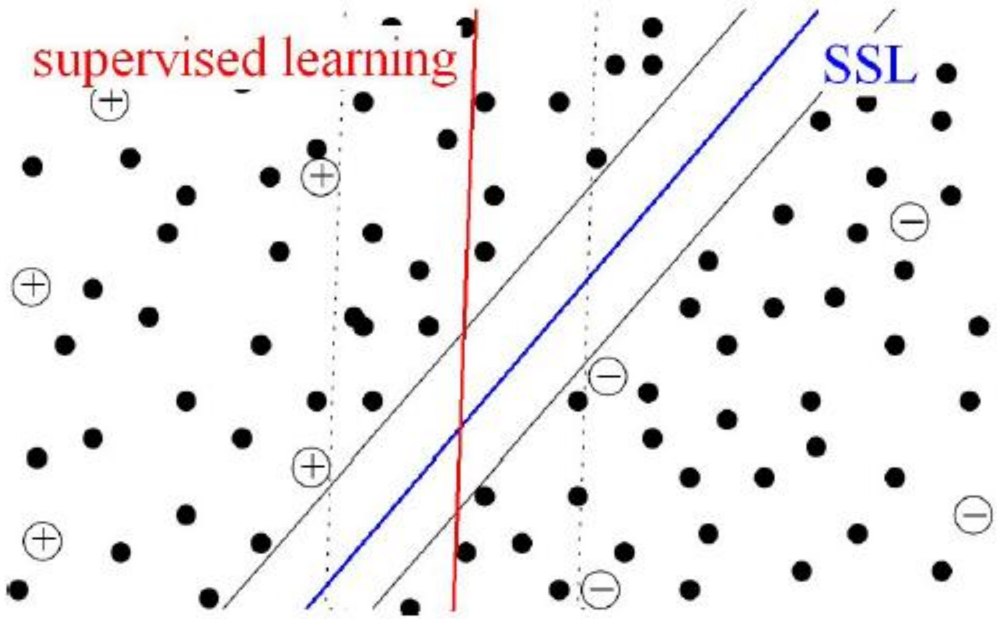
- Unsupervised learning: given $x_1 \ldots x_n$, do clustering, outlier detection etc.
- Supervised learning: given $(x_1, y_1) \ldots (x_n, y_n)$, learn a predictor $f: X \rightarrow Y$
- Semi-supervised learning (SSL): given $(x_1, y_1) \ldots (x_n, y_n)$, $x_{n+1} \ldots x_{n+m}$, learn a better predictor $f: X \rightarrow Y$

SSL Model 1: Mixtures



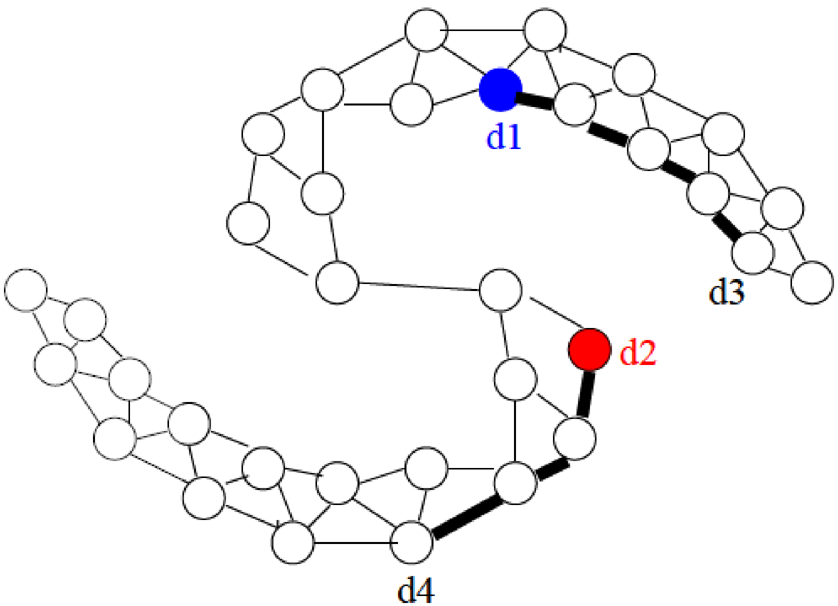
- Gaussian Mixture Models, Multinomial (bag-of-word) mixture
- Assumption: each class y has a specific parametric conditional distribution p(x|y) for its items (e.g. Gaussian).

SSL Model 2: Large Margin



- Transductive Support Vector Machines, Gaussian Processes
- Assumption: instances from different classes are separated by a large gap (the margin).

SSL Model 3: Graph



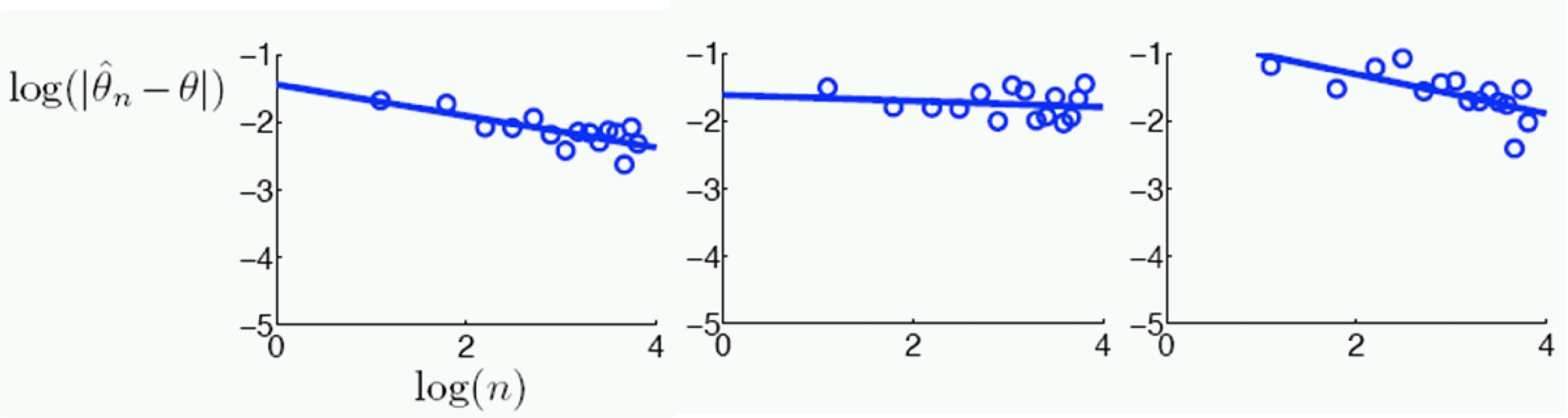
- Graph cut, label propagation, manifold regularization, SSL on tree structure
- Assumption: two instances connected by a strong edge have similar labels.

When does SSL help?

- SSL helps, if the assumption fits the link between:
  - p(x): what unlabeled can tell us, and
  - p(y|x): what the true classification should be
- Warning: wrong SSL assumption can actually lead to worse learning!
  - but even this can be interesting

Results

- Human passive learning even slower than 1/n polynomially.



- Yoked: humans learn to rely on computer.

Monkey Algorithm?

1. Initially $h = (1 \ldots 1)$ (d ones). Repeat 2–4:
2. Randomly pick $x \in \{x_1 \ldots x_{d/2}\}$ for which $h \wedge x \neq 0$
3. If x is correct, $h = h \wedge x$.
4. If x is wrong, $h = h \wedge \neg x$. If $h = 0$, ~~reset $h = (1 \ldots 1)$.~~

- Slow learner: skip step 3, 4 with probability α
- Stubborn: when h=0, retain the incorrect h with probability β
- With α=0.93 and β=0.96, algorithm makes 563 errors, in which 67% perservarative.