Human and Machine Learning

Tom Mitchell
Machine Learning Department
Carnegie Mellon University

November 20, 2006

How can studies of machine (human) learning inform studies of human (machine) learning?

Learning = improving performance
at some task
through experience

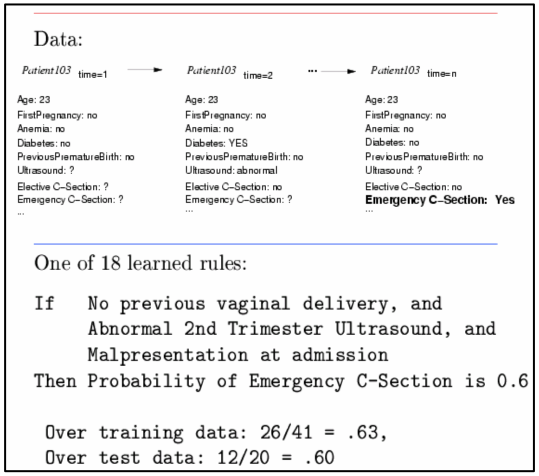
Outline

1. Machine Learning and Human Learning

2. Aligning specific results from ML and HL
   - Learning to predict and achieve rewards
     - TD learning ↔ Dopamine system in the brain
   - Value of redundancy in data inputs
     - Cotraining ↔ Intersensory redundancy hypothesis

3. Core questions and conjectures

Machine Learning - Practice



Data:

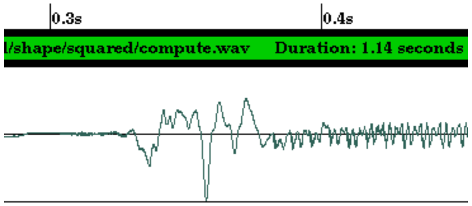
Patient103 time=1 →	Patient103 time=2 ... →	Patient103 time=n
Age: 23	Age: 23	Age: 23
FirstPregnancy: no	FirstPregnancy: no	FirstPregnancy: no
Anemia: no	Anemia: no	Anemia: no
Diabetes: no	Diabetes: YES	Diabetes: no
PreviousPrematureBirth: no	PreviousPrematureBirth: no	PreviousPrematureBirth: no
Ultrasound: ?	Ultrasound: abnormal	Ultrasound: ?
Elective C–Section: ?	Elective C–Section: no	Elective C–Section: no
Emergency C–Section: ?	Emergency C–Section: ?	**Emergency C–Section: Yes**
...	...	...

One of 18 learned rules:

```
If   No previous vaginal delivery, and
     Abnormal 2nd Trimester Ultrasound, and
     Malpresentation at admission
Then Probability of Emergency C-Section is 0.6

  Over training data: 26/41 = .63,
  Over test data: 12/20 = .60
```

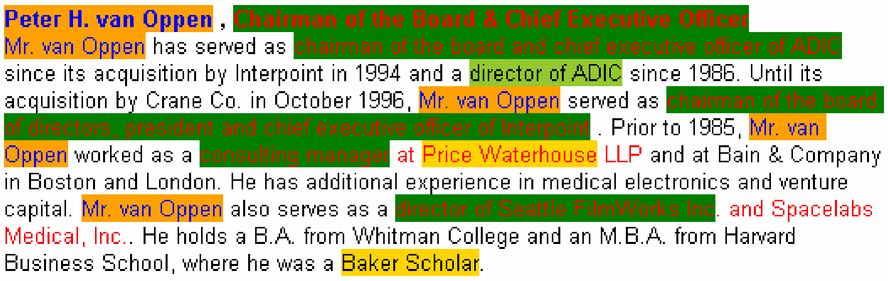
Mining Databases



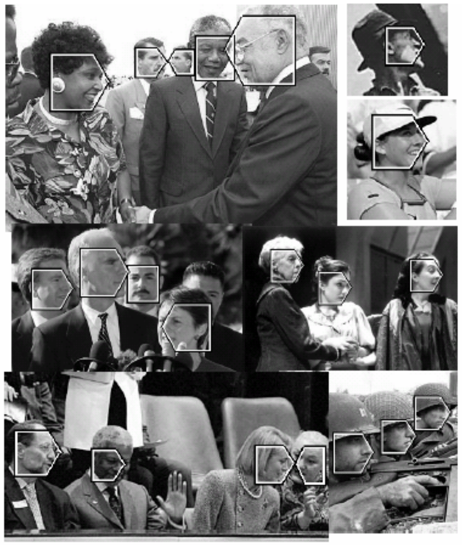
Speech Recognition

Control learning

Text analysis



Peter H. van Oppen , **Chairman of the Board & Chief Executive Officer**
Mr. van Oppen has served as chairman of the board and chief executive officer of ADIC since its acquisition by Interpoint in 1994 and a director of ADIC since 1986. Until its acquisition by Crane Co. in October 1996, Mr. van Oppen served as chairman of the board of directors, president and chief executive officer of Interpoint . Prior to 1985, Mr. van Oppen worked as a consulting manager at Price Waterhouse LLP and at Bain & Company in Boston and London. He has additional experience in medical electronics and venture capital. Mr. van Oppen also serves as a director of Seattle FilmWorks Inc. and Spacelabs Medical, Inc.. He holds a B.A. from Whitman College and an M.B.A. from Harvard Business School, where he was a Baker Scholar.



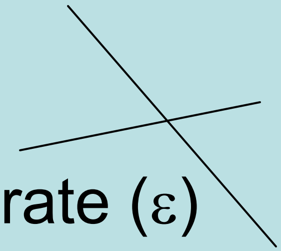
Object recognition

- Reinforcement learning
- Supervised learning
- Bayesian networks
- Hidden Markov models
- Unsupervised clustering
- Explanation-based learning
- ....

Machine Learning - Theory

PAC Learning Theory

(for supervised concept learning)



# examples (*m*)

representational complexity (*H*)

error rate (ε)

failure probability (δ)

$$m \geq \frac{1}{\epsilon}(\ln|H| + \ln(1/\delta))$$

Similar theories for

- Reinforcement skill learning
- Unsupervised learning
- Active student querying
- …

… also relating:

- # of mistakes during learning
- learner's query strategy
- convergence rate
- asymptotic performance
- …

What We Know About ML

- Excellent algorithms for pure induction
  - SVM’s, decision trees, graphical models, neural nets, ...
- Algorithms for dimensionality reduction
  - PCA, ICA, compression algorithms, ...
- Fundamental information theoretic bounds relate data and biases to probability of successful learning
  - PAC learning theory, statistical estimation, grammar induction, ...
- Active learning by querying teacher is much more data-efficient than random observation
- Algorithms to learn from delayed feedback (reinforcement)
  - Temporal difference learning, Q learning, policy iteration, ...
- ...

ML Has Little to Say About

- Learning cumulatively over time
- Learning from instruction (lectures, discussion)
- Role of motivation, forgetting, curiosity, fear, boredom, ...
- Implicit (unconscious) versus explicit (deliberate) learning
- ...

What We Know About HL*

Neural level:

- Hebbian learning: connection between the pre-synaptic and post-synaptic neuron increases if pre-synaptic neuron is repeatedly involved in activating post-synaptic
  - Biochemistry: NMDA channels, Ca^{2+}, AMPA receptors, ...
- Timing matters: strongest effect if pre-synaptic action potential occurs within 0 - 50msec before postsynaptic firing.
- Time constants for synaptic changes are a few minutes.
  - Can be disrupted by protein inhibitors injected after the training experience

* I'm not an expert

What We Know About HL*

System level:

- In addition to single synapse changes, memory formation involves longer term ‘consolidation’ involving multiple parts of the brain

- Time constant for consolidation is hours or days: memory of new experiences can be disrupted by events occurring after the experience (e.g., drug interventions, trauma).
  - E.g., injections in amygdala 24 hours after training can impact recall experience, with no impact on recall within a few hours

- Consolidation thought to involve regions such as amygdala, hippocampus, frontal cortex. Hippocampus might orchestrate consolidation without itself being home of memories

- Dopamine seems to play a role in reward-based learning (and addictions)

* I’m not an expert

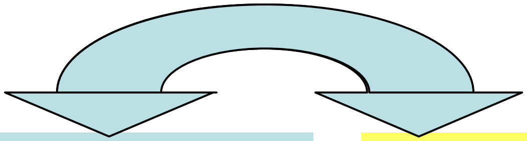
What We Know About HL*

Behavioral level:

- Power law of practice: competence vs. training on log-log plot is a straight line, across many skill types
- Role of reasoning and knowledge compilation in learning
  - chunking, ACT-R, Soar
- Timing: Expanded spacing of stimuli aids memory, ...
- Theories about role of sleep in learning/consolidation
- Implicit and explicit learning. (unaware vs. aware).
- Developmental psychology: knows much about sequence of acquired expertise during childhood
  - Intersensory redundancy hypothesis

* I'm not an expert

Models of Learning Processes



Machine Learning:

- # of examples
- Error rate
- Reinforcement learning
- Explanations

- Learning from examples
- Complexity of learner’s representation
- Probability of success
- Exploitation / exploration
- Prior probabilities
- Loss functions

Human Learning:

- # of examples
- Error rate
- Reinforcement learning
- Explanations

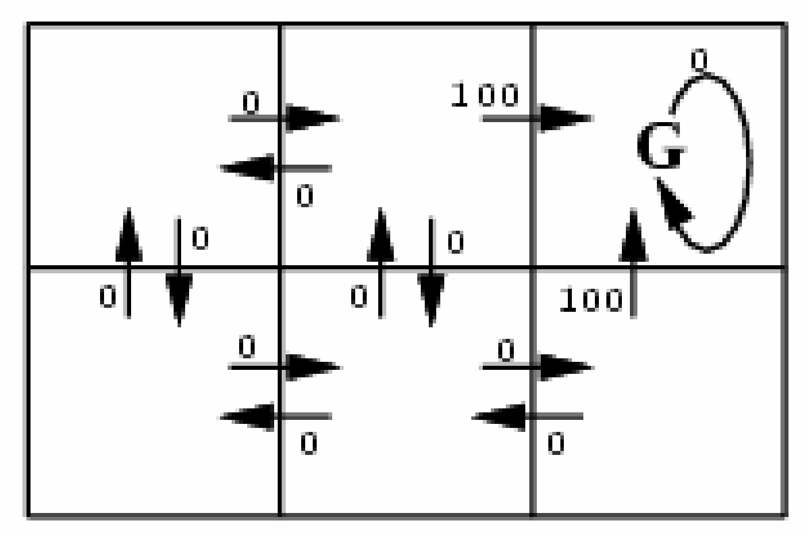
- Human supervision
  - Lectures
  - Question answering
- Attention, motivation
- Skills vs. Principles
- Implicit vs. Explicit learning
- Memory, retention, forgetting

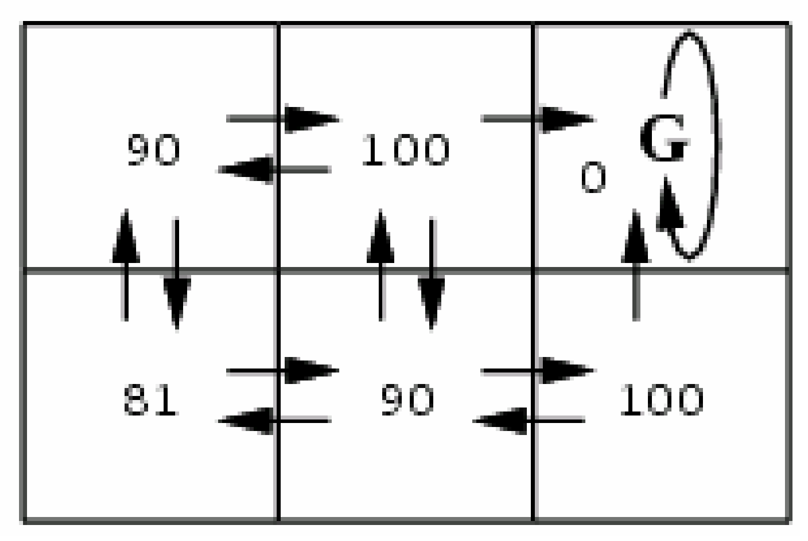
1. Learning to predict and achieve rewards

TD learning ↔ Dopamine in the brain

Reinforcement Learning

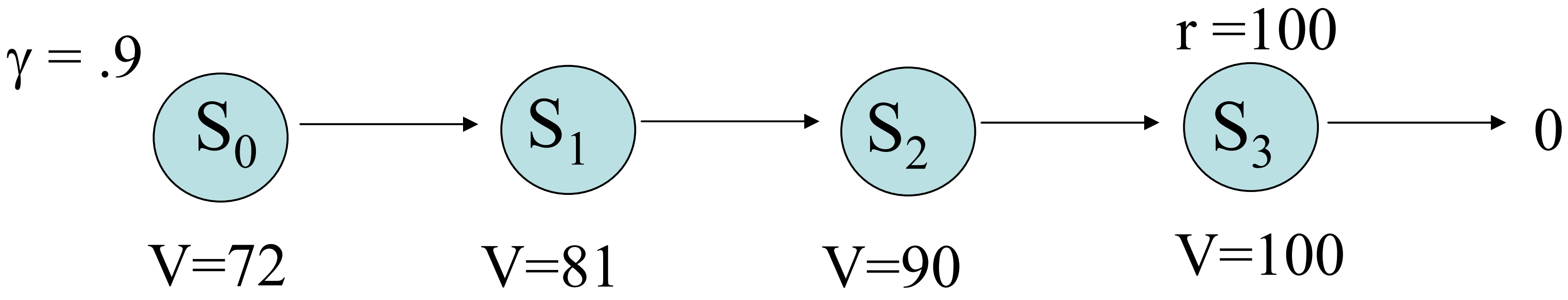
[Sutton and Barto 1981; Samuel 1957]





$$V^*(s) = E[r_t + \gamma\, r_{t+1} + \gamma^2 r_{t+2} + \ldots]$$

Reinforcement Learning in ML



$$V(s_t) = E[r_t + \gamma\, r_{t+1} + \gamma^2 r_{t+2} + \ldots]$$

$$V(s_t) = E[r_t] + \gamma\, V(s_{t+1})$$

To learn V, use each transition to generate a training signal:

$$training\_error_t = r_t + \gamma \hat{V}(s_{t+1}) - \hat{V}(s_t)$$

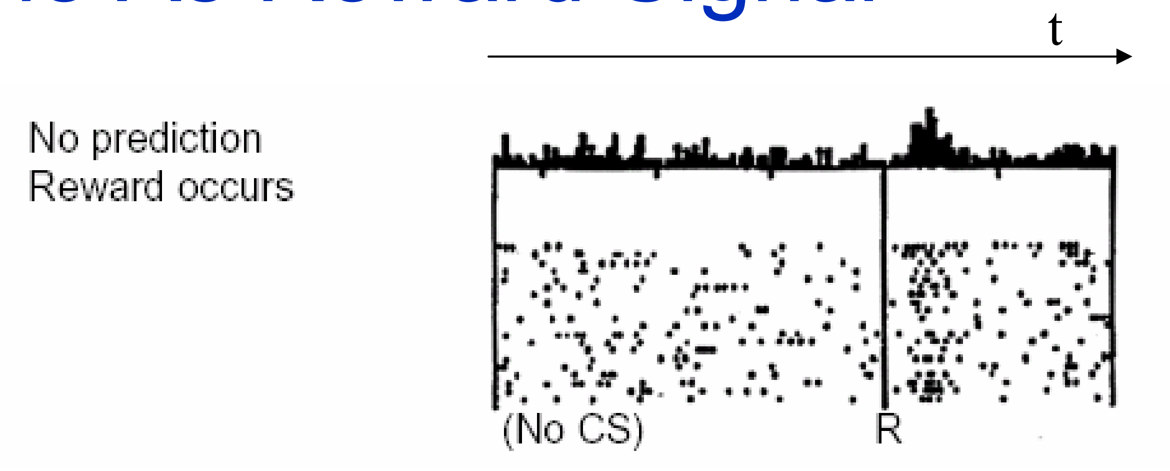
Reinforcement Learning in ML

$$\text{training error} = r_t + \gamma V(s_{t+1}) - V(s_t)$$

- Variants of RL have been used for a variety of practical control learning problems
  - Temporal Difference learning
  - Q learning
  - Learning MDPs, POMDPs

- Theoretical results too
  - Assured convergence to optimal V(s) under certain conditions
  - Assured convergence for Q(s,a) under certain conditions

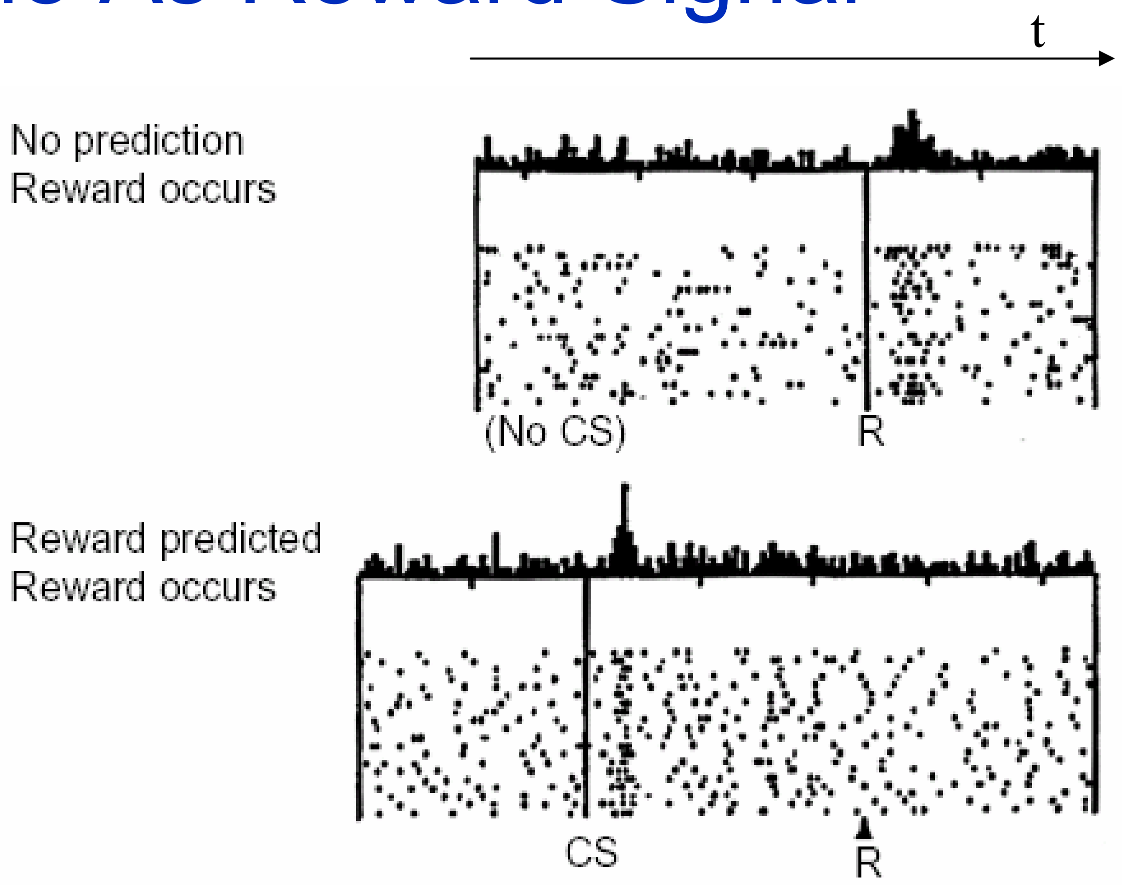
Dopamine As Reward Signal



[Schultz et al.,
Science, 1997]

Dopamine As Reward Signal

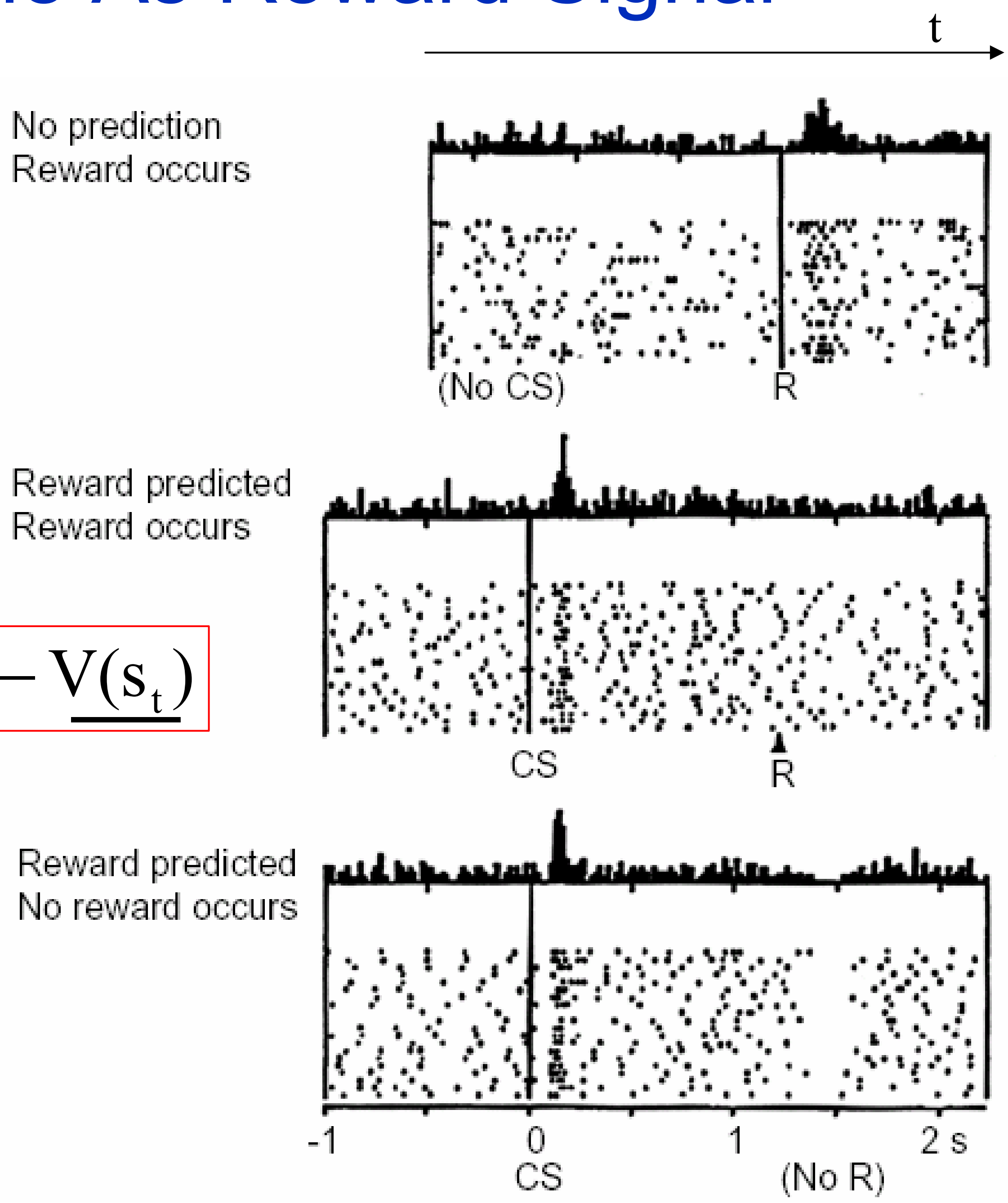
[Schultz et al., *Science*, 1997]



Dopamine As Reward Signal

[Schultz et al., *Science*, 1997]

$$\boxed{\text{error} = \underline{r_t + \gamma\, V(s_{t+1})} - \underline{V(s_t)}}$$



RL Models for Human Learning

[Seymore et al., Nature 2004]

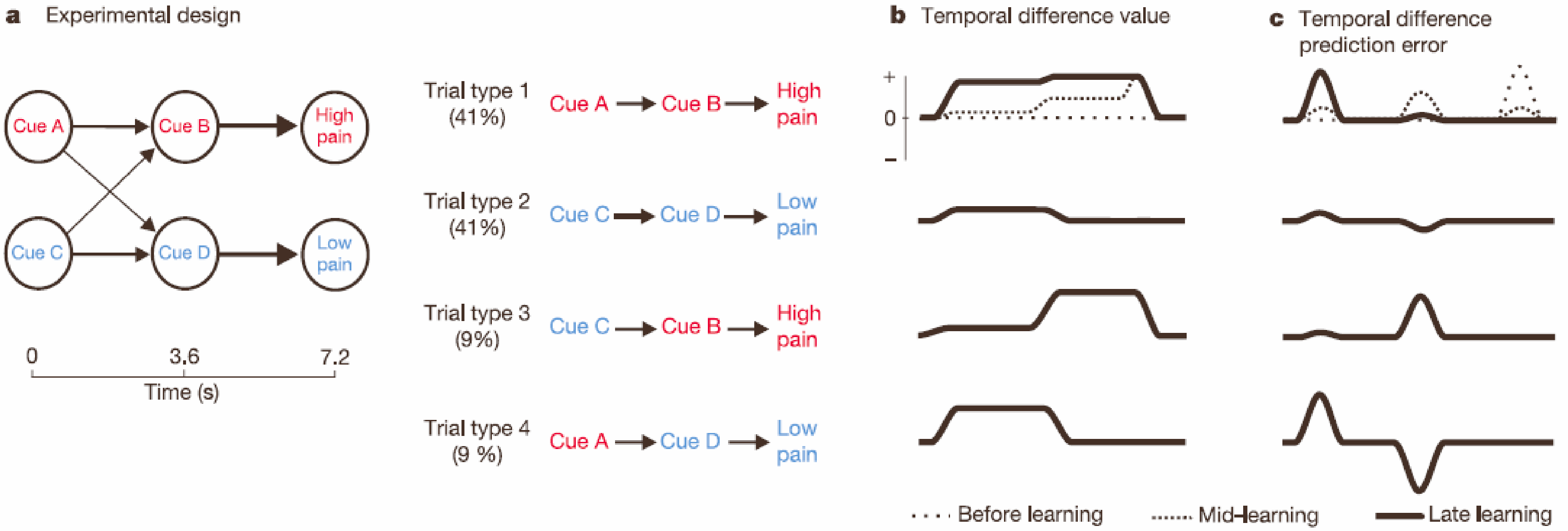
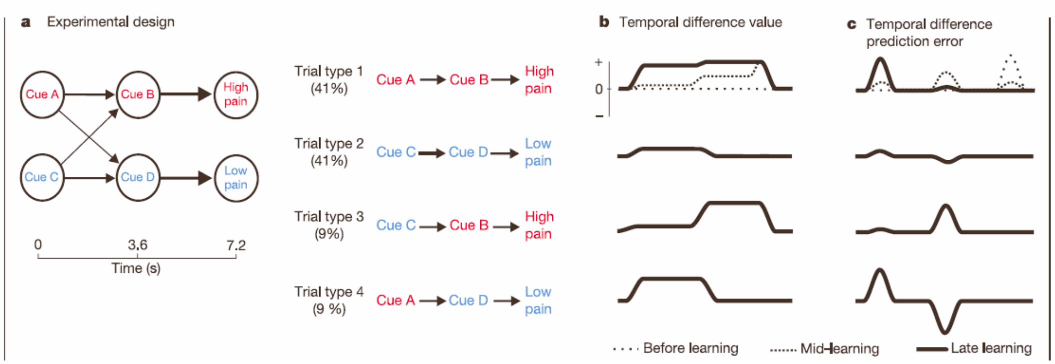


Figure 1 Experimental design and temporal difference model. **a**, The experimental design expressed as a Markov chain, giving four separate trial types. **b**, Temporal difference value. As learning proceeds, earlier cues learn to make accurate value predictions (that is, weighted averages of the final expected pain). **c**, Temporal difference prediction error; during learning the prediction error is transferred to earlier cues as they acquire the ability to make predictions. In trial types 3 and 4, the substantial change in prediction elicits a large positive or negative prediction error. (For clarity, before and mid-learning are shown only for trial type 1.)



[Seymore et al., Nature 2004]

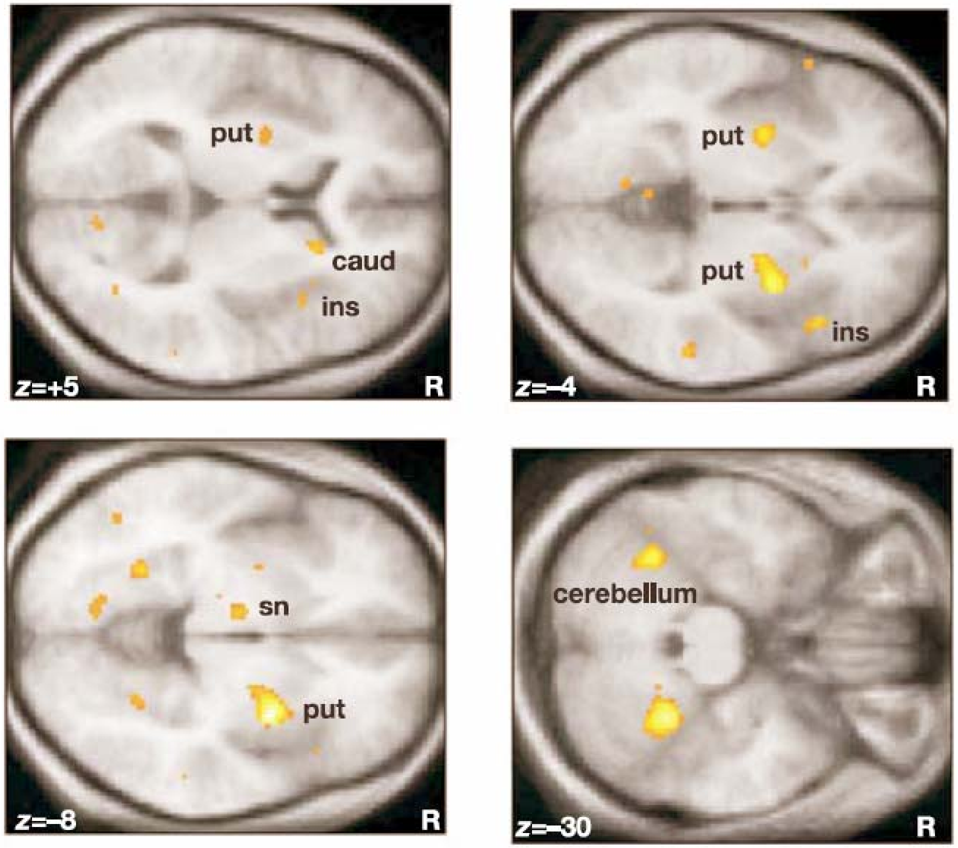


Figure 2 Temporal difference prediction error (statistical parametric maps). Areas coloured yellow/orange show significant correlation with the temporal difference

Human EEG responses to Pos/Neg Reward

from [Nieuwenhuis et al.]

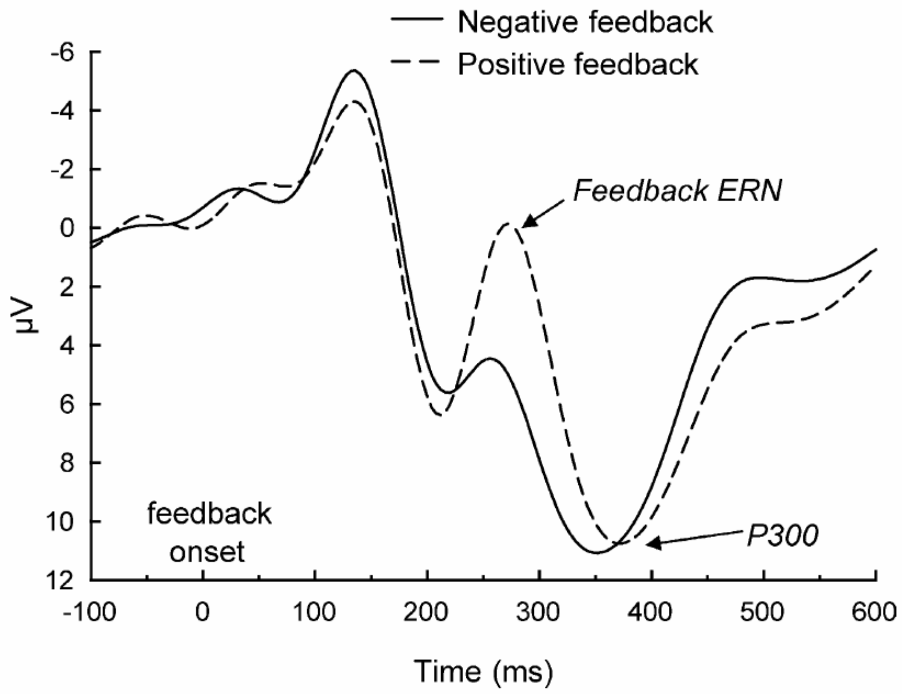


Fig. 1. Typical example of event-related brain potentials associated with negative and positive feedback (adapted from Ref. [25]). Negative is

Response due to feedback on timing task (press button exactly 1 sec after sound).

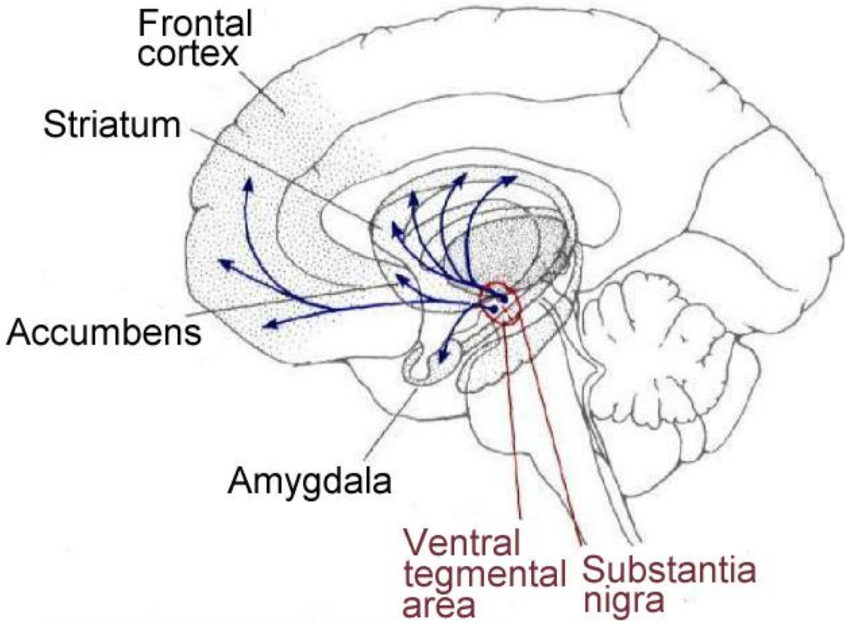
Neural source appears to be in anterior cingulate cortex (ACC)

Response is abnormal in some subjects with OCD

One Theory of RL in the Brain

from [Nieuwenhuis et al.]

- Basal ganglia monitor events, predict future rewards
- When prediction revised upward (downward), causes increase (decrease) in activity of midbrain dopaminergic neurons, influencing ACC
- This dopamine-based activation somehow results in revising the reward prediction function. Possibly through direct influence on Basal ganglia, and via prefrontal cortex



Summary: Temporal Difference ML Model Predicts Dopaminergic Neuron Acitivity during Learning

- Evidence now of neural reward signals from
  - Direct neural recordings in monkeys
  - fMRI in humans (1 mm spatial resolution)
  - EEG in humans (1-10 msec temporal resolution)

- Dopaminergic responses track temporal difference error in RL

- Some differences, and efforts to refine HL model
  - Better information processing model
  - Better localization to different brain regions
  - Study timing (e.g., basal ganglia learns faster than PFC ?)

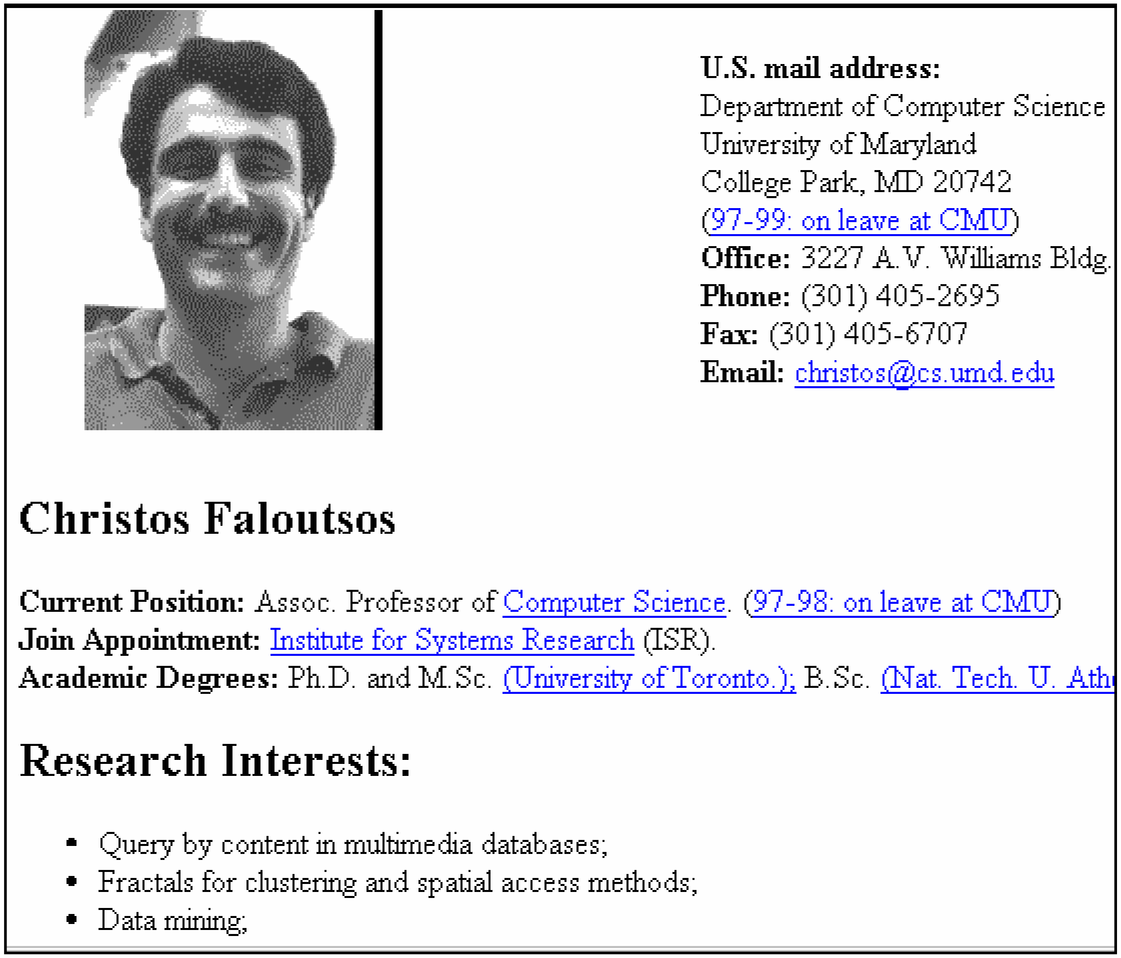
2. The value of unlabeled multi-sensory data for learning classifiers

Cotraining ↔ Intersensory redundancy hypothesis

Redundantly Sufficient Features

Professor Faloutsos

my advisor



U.S. mail address:
Department of Computer Science
University of Maryland
College Park, MD 20742
(97-99: on leave at CMU)
Office: 3227 A.V. Williams Bldg.
Phone: (301) 405-2695
Fax: (301) 405-6707
Email: christos@cs.umd.edu

Christos Faloutsos

Current Position: Assoc. Professor of Computer Science. (97-98: on leave at CMU)
Join Appointment: Institute for Systems Research (ISR).
Academic Degrees: Ph.D. and M.Sc. (University of Toronto.); B.Sc. (Nat. Tech. U. Ath

Research Interests:

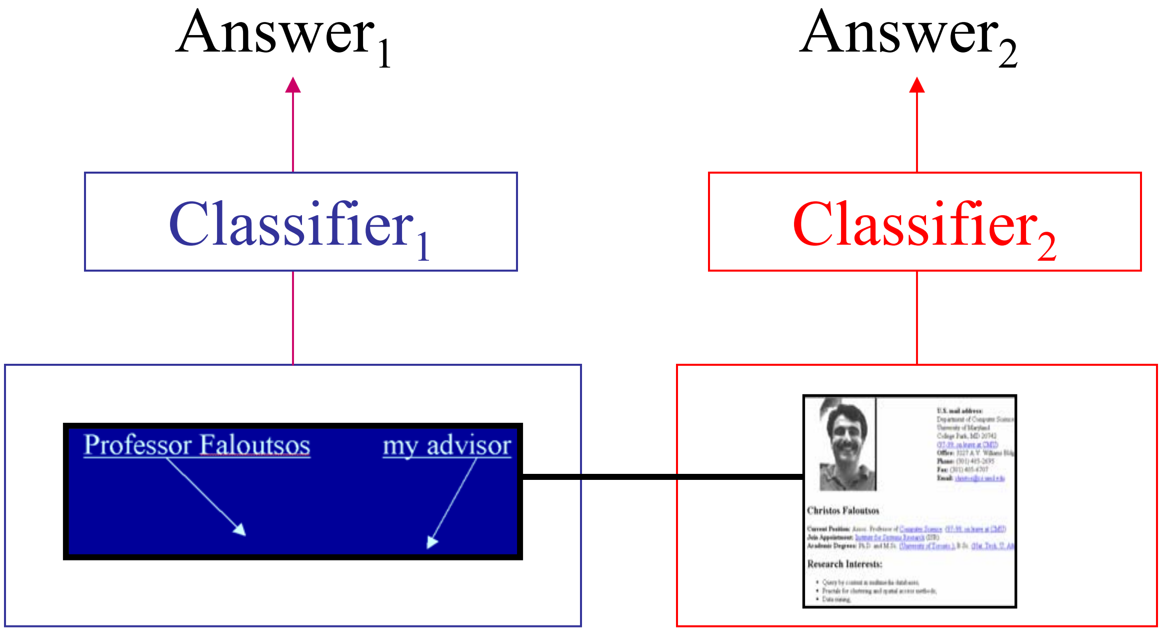
- Query by content in multimedia databases;
- Fractals for clustering and spatial access methods;
- Data mining;

Co-Training

Idea: Train Classifier$_1$ and Classifier$_2$ to:

1. Correctly classify labeled examples

2. Agree on classification of unlabeled



Answer$_1$

Answer$_2$

Classifier$_1$

Classifier$_2$

Professor Faloutsos

my advisor

U.S. mail address:
Department of Computer Science
University of Maryland
College Park, MD 20742
(97-98: on leave at CMU)
Office: 3227 A.V. Williams Bldg.
Phone: (301) 405-2695
Fax: (301) 405-6707
Email: [illegible]

Christos Faloutsos

Current Position: Assoc. Professor of Computer Science. (97-98: on leave at CMU)
Join Appointment: Institute for Systems Research (ISR)
Academic Degrees: Ph.D. and M.Sc. (University of Toronto.), B.Sc. (Nat. Tech. U. Ath

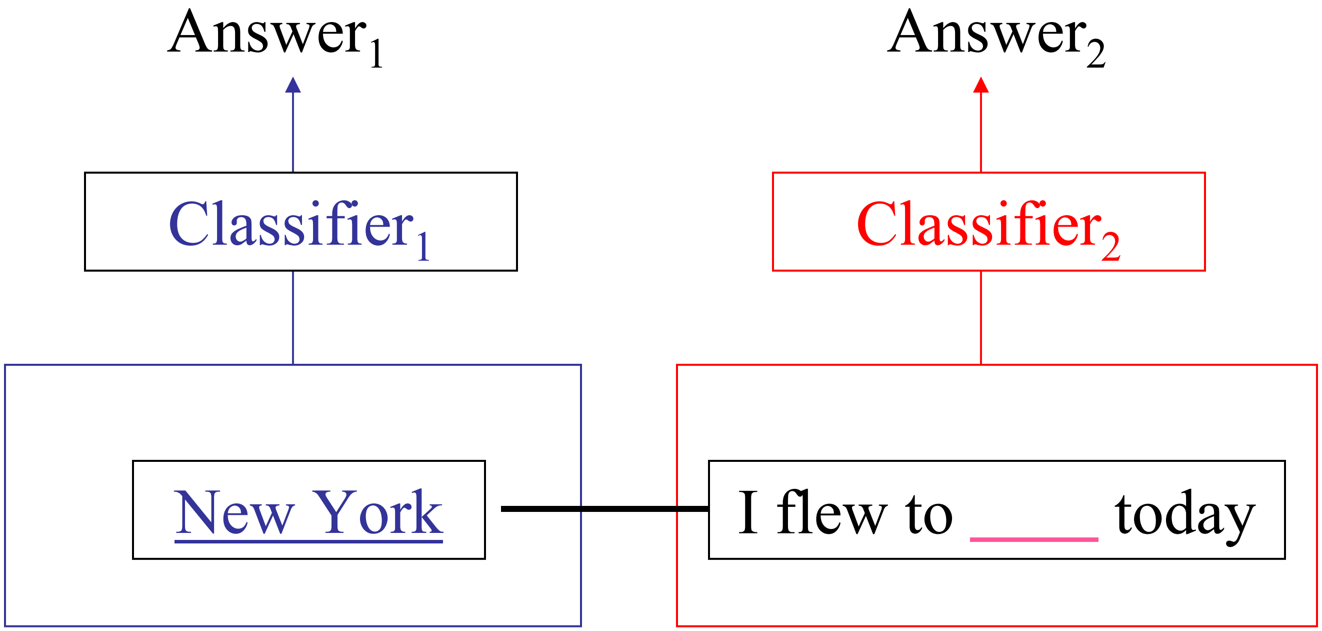
Research Interests:

- Query by content in multimedia databases;
- Fractals for clustering and spatial access methods;
- Data mining.

Co-Training

Where else might this work?

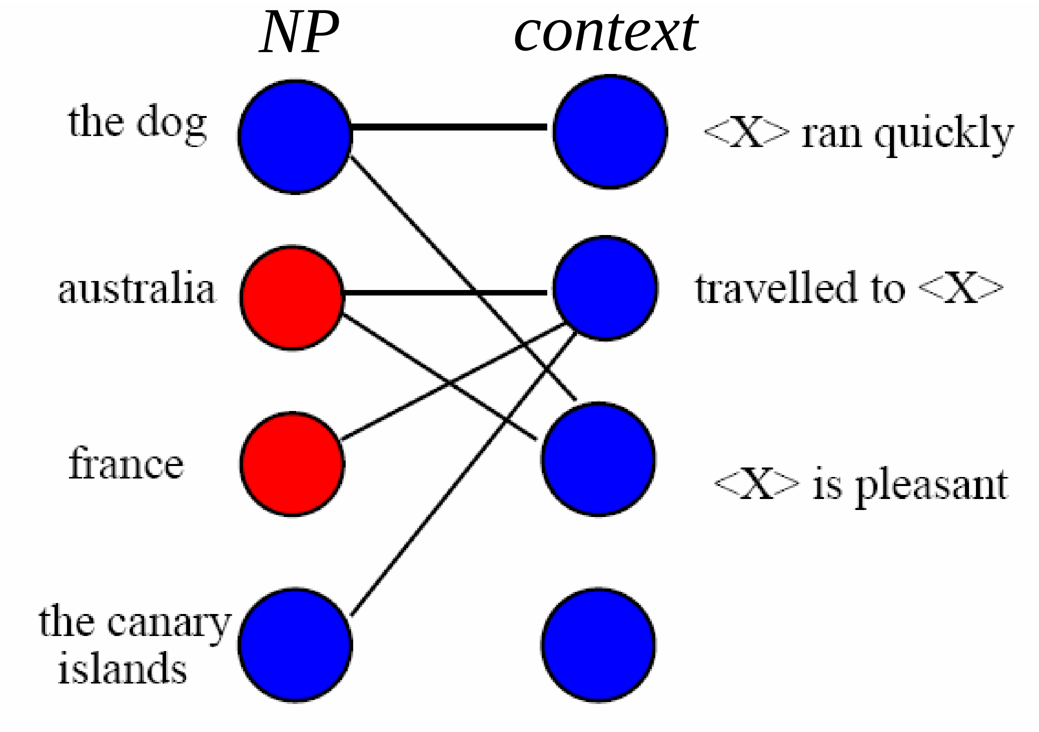
- learning lexicons and named-entity recognizers for people, places, dates, books, … (eg., Riloff&Jones; Collins et al.)



I flew to **New York** today.

CoEM applied to Named Entity Recognition

[Rosie Jones, 2005], [Ghani & Nigam, 2000]

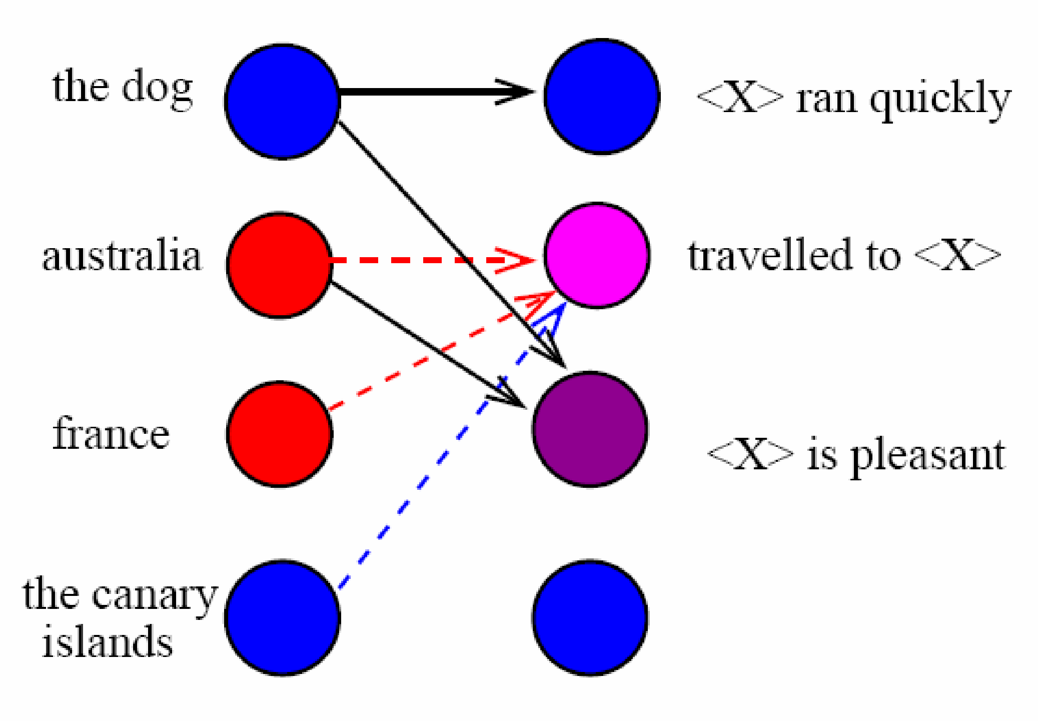


Update rules:

$$
\begin{aligned}
P(class|context_i) &= \sum_j P(class|NP_j)P(NP_j|context_i) \\
P(class|NP_i) &= \sum_j P(class|context_j)P(context_j|NP_i)
\end{aligned}
$$

CoEM applied to Named Entity Recognition

[Rosie Jones, 2005], [Ghani & Nigam, 2000]

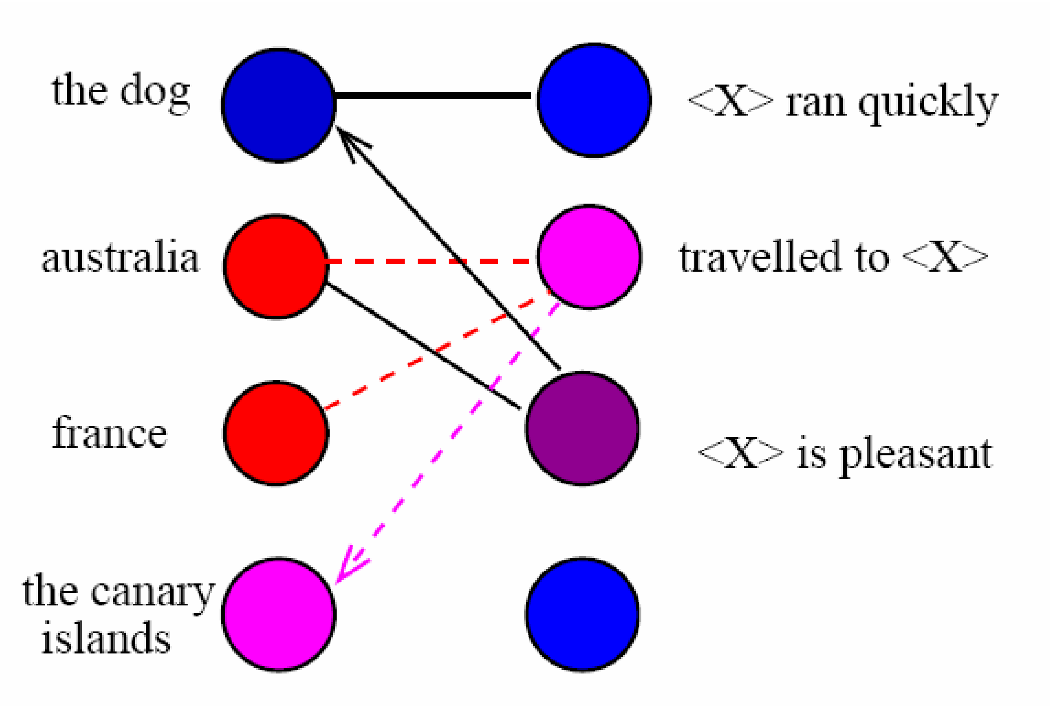


Update rules:

$$\begin{aligned} P(class|context_i) &= \sum_j P(class|NP_j)P(NP_j|context_i) \\ P(class|NP_i) &= \sum_j P(class|context_j)P(context_j|NP_i) \end{aligned}$$

CoEM applied to Named Entity Recognition

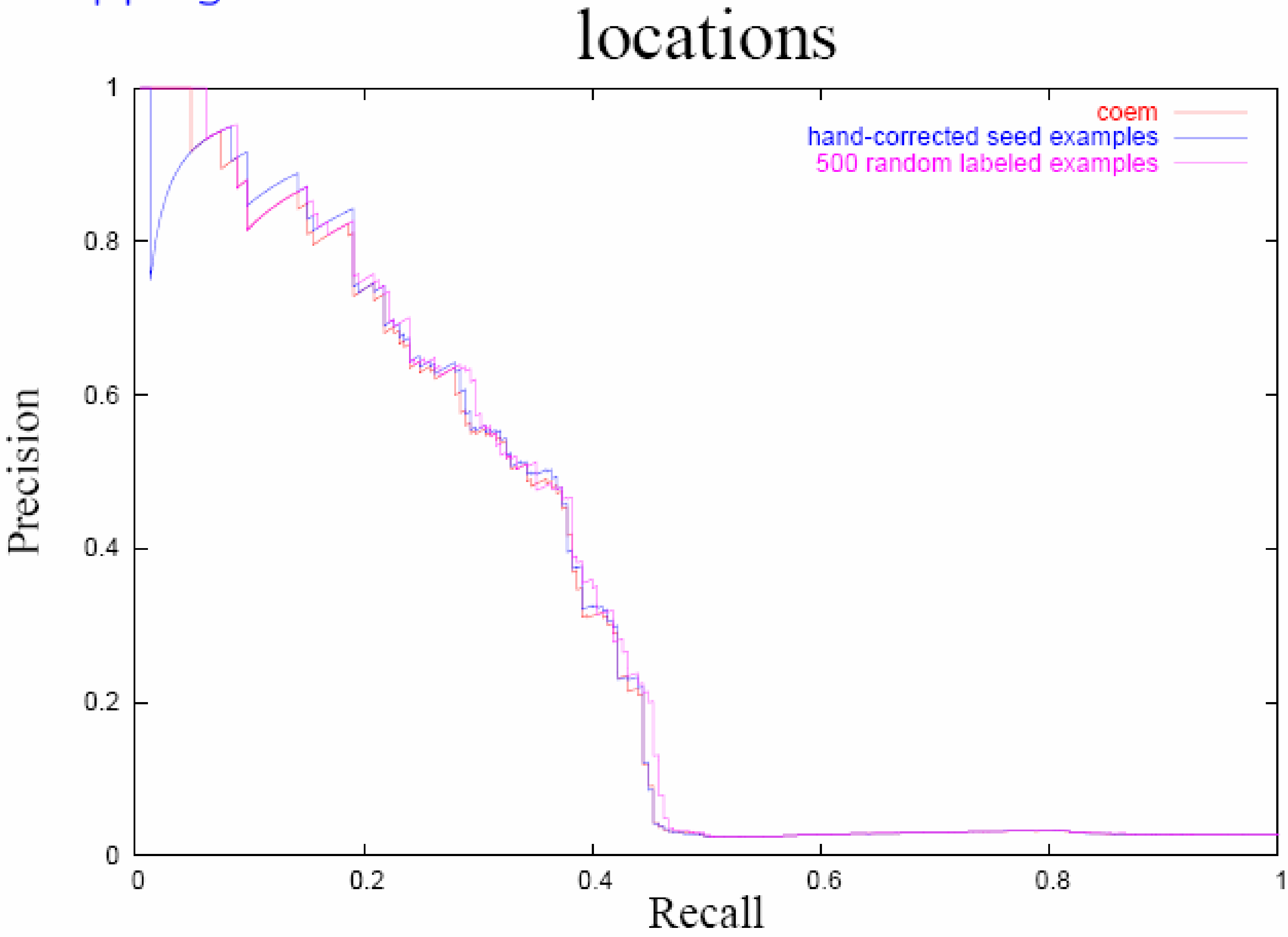
[Rosie Jones, 2005], [Ghani & Nigam, 2000]



Update rules:

$$
\begin{aligned}
P(class|context_i) &= \sum_j P(class|NP_j)P(NP_j|context_i) \\
P(class|NP_i) &= \sum_j P(class|context_j)P(context_j|NP_i)
\end{aligned}
$$

Bootstrapping Results



Co-Training Theory

[Blum&Mitchell 98; Dasgupta 04, ...]

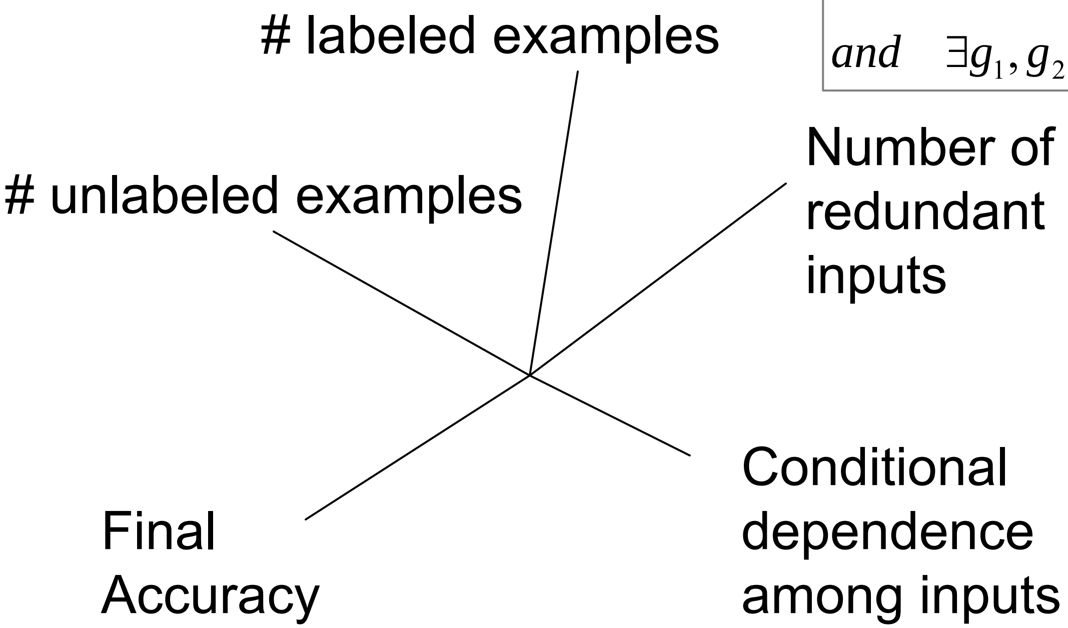
CoTraining setting:

learn $f : X \rightarrow Y$

where $X = X_1 \times X_2$

where x *drawn from unknown distribution*

and $\exists g_1, g_2 \quad (\forall x) g_1(x_1) = g_2(x_2) = f(x)$



# labeled examples

# unlabeled examples

Number of redundant inputs

Final Accuracy

Conditional dependence among inputs

→ want inputs less dependent, increased number of redundant inputs, ...

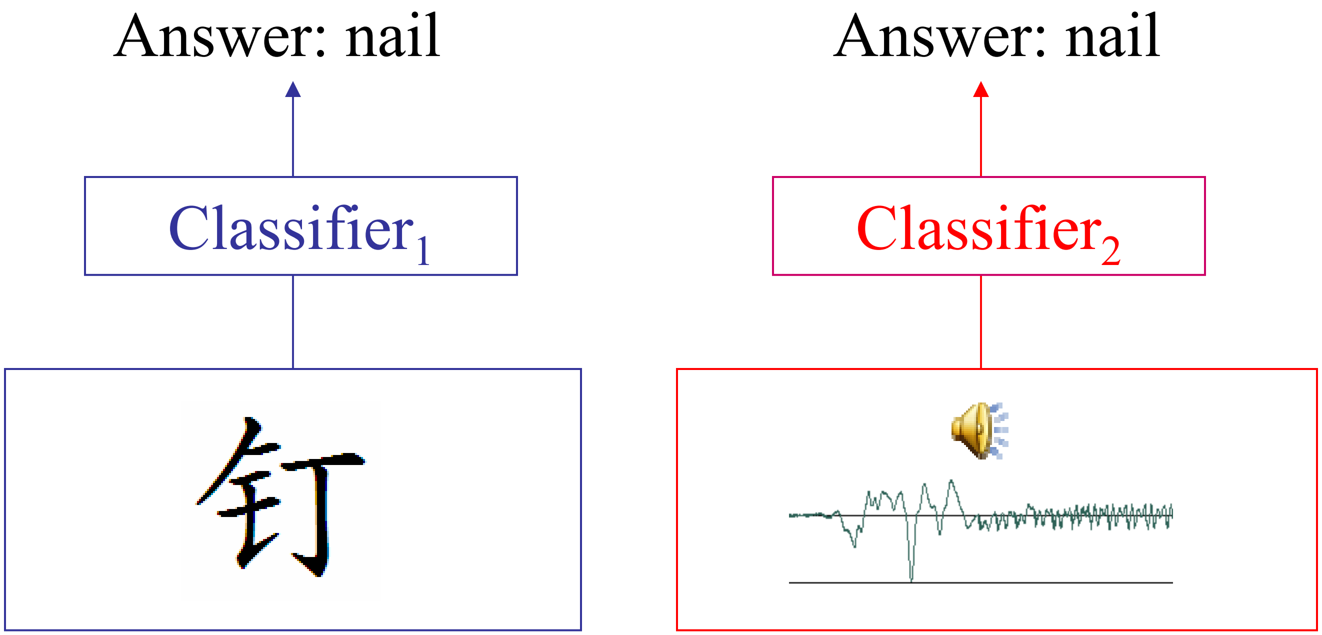
Theoretical Predictions of CoTraining

- Possible to learn from unlabeled examples
- Value of unlabeled data depends on
  - How (conditionally) independent are X_1 and X_2
    - The more the better
  - How many redundant sensory inputs X_i there are
    - Expected error decreases exponentially with this number
- Disagreement on unlabeled data predicts true error

Do these predictions hold for human learners?

Co-Training [joint work with Liu, Perfetti, Zi]

Can it work for humans learning chinese as a second language?



Answer: nail

Answer: nail

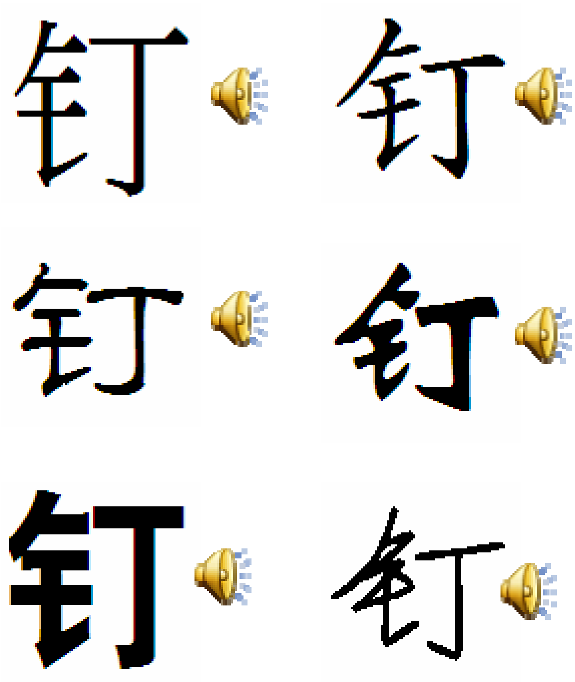
Classifier_1

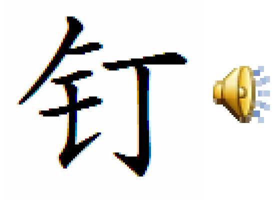
Classifier_2

Examples

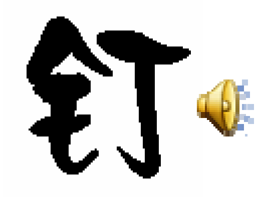
- Training fonts and speakers for “nail”

- Testing fonts and speakers for “nail”





Familiar

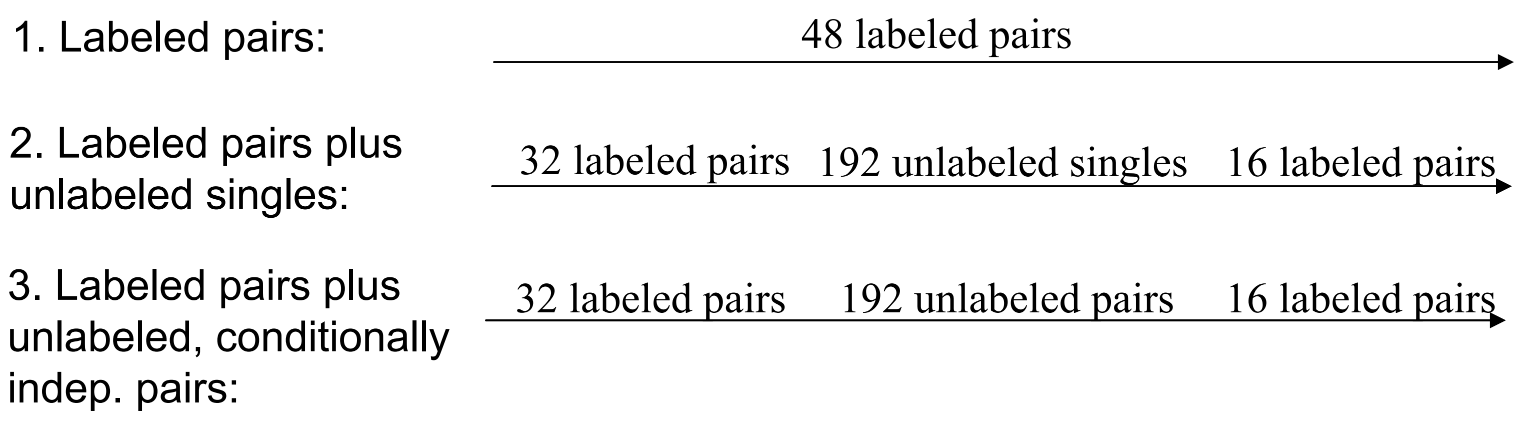


Unfamiliar

Experiment: Cotraining in Human Learning

[with Liu, Perfetti, Zi 2006]

- 44 human subjects learning Chinese as second lanuage
- Target function to be learned:
  - chinese word (spoken / written) → english word
  - 16 distinct words, 6 speakers, 6 writers = 16x6x6 stimulus pairs
- Training conditions:



1. Labeled pairs: 48 labeled pairs

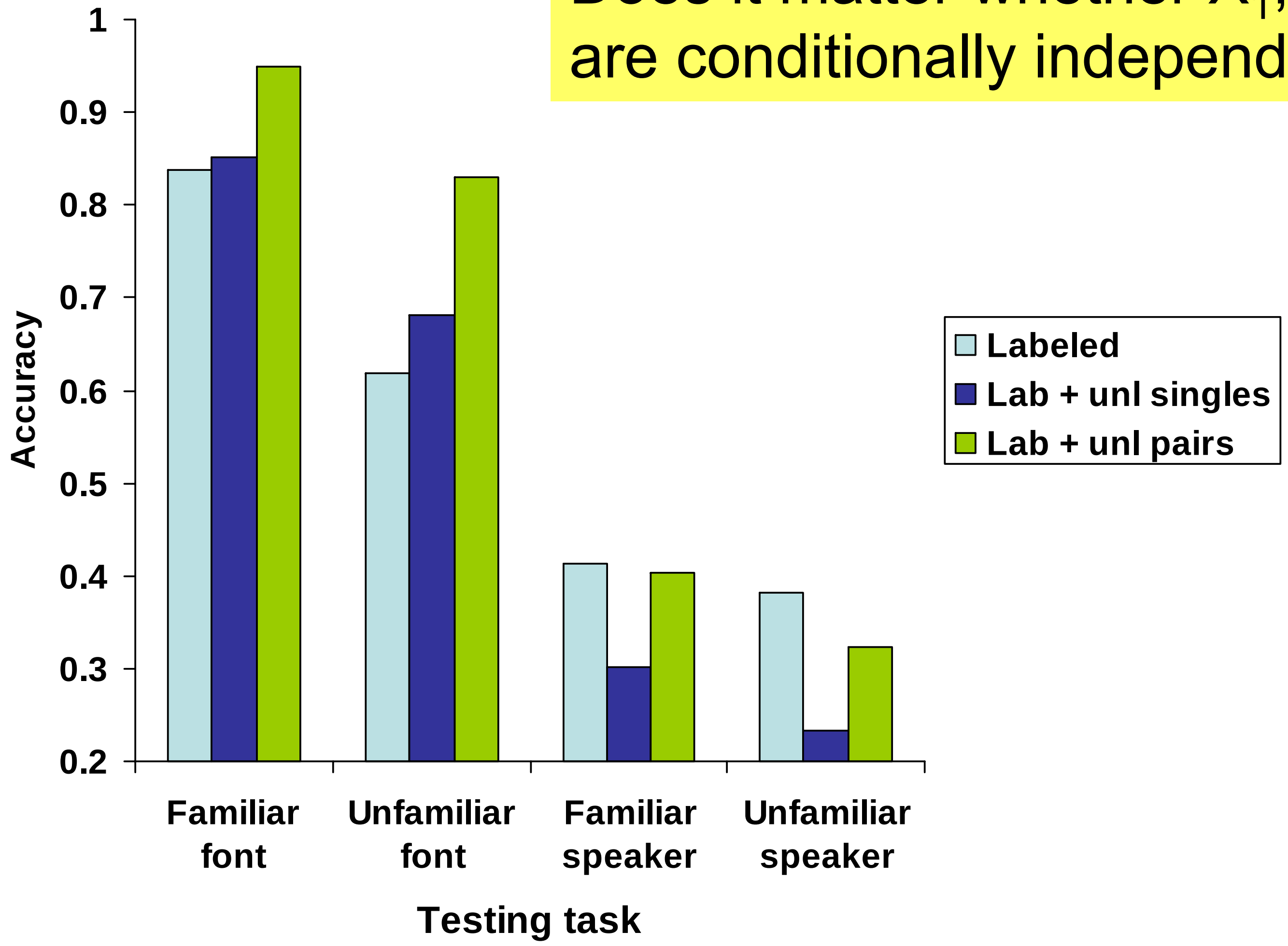
2. Labeled pairs plus unlabeled singles: 32 labeled pairs 192 unlabeled singles 16 labeled pairs

3. Labeled pairs plus unlabeled, conditionally indep. pairs: 32 labeled pairs 192 unlabeled pairs 16 labeled pairs

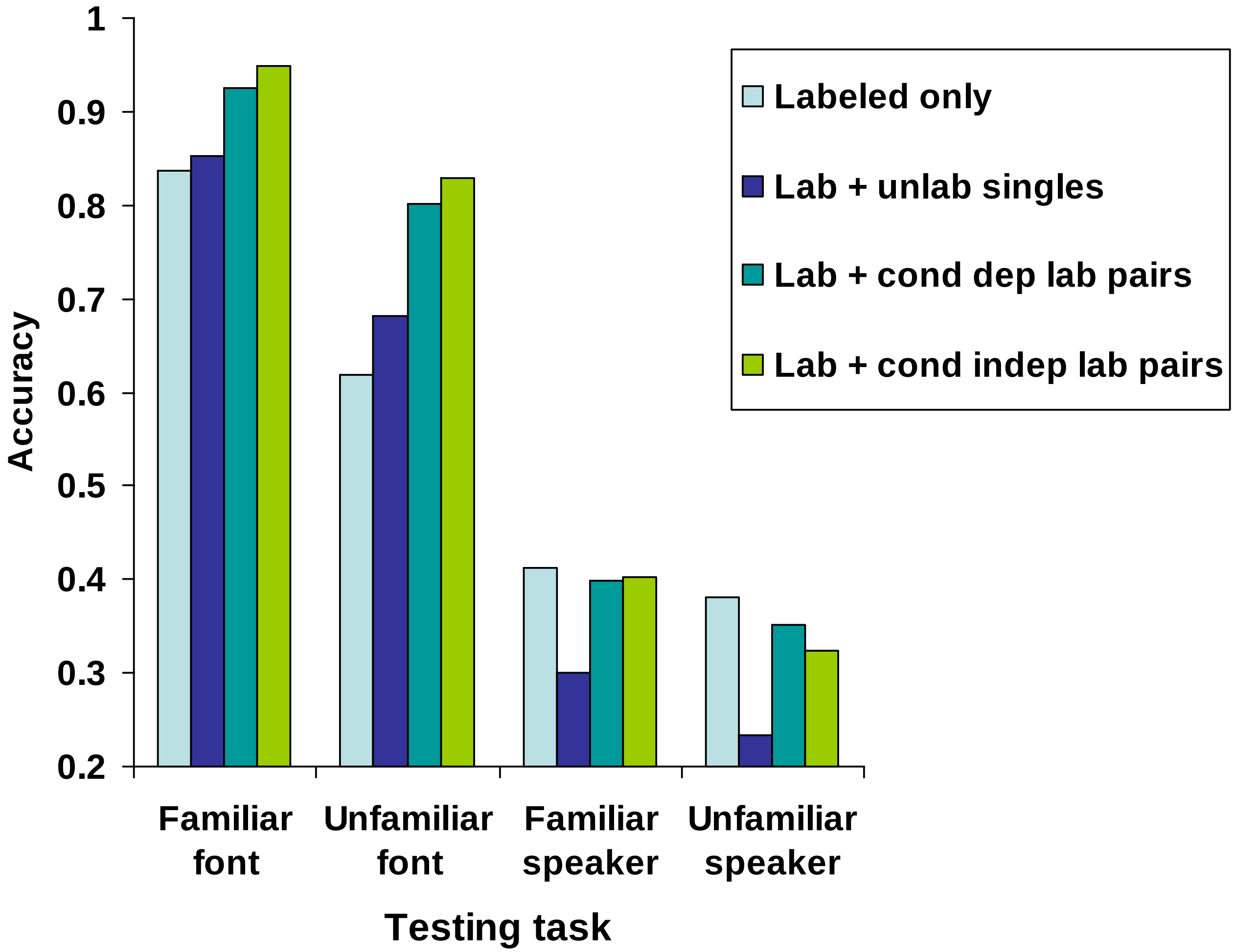
- Test: 16 test words (single chinese stimulus), require english label

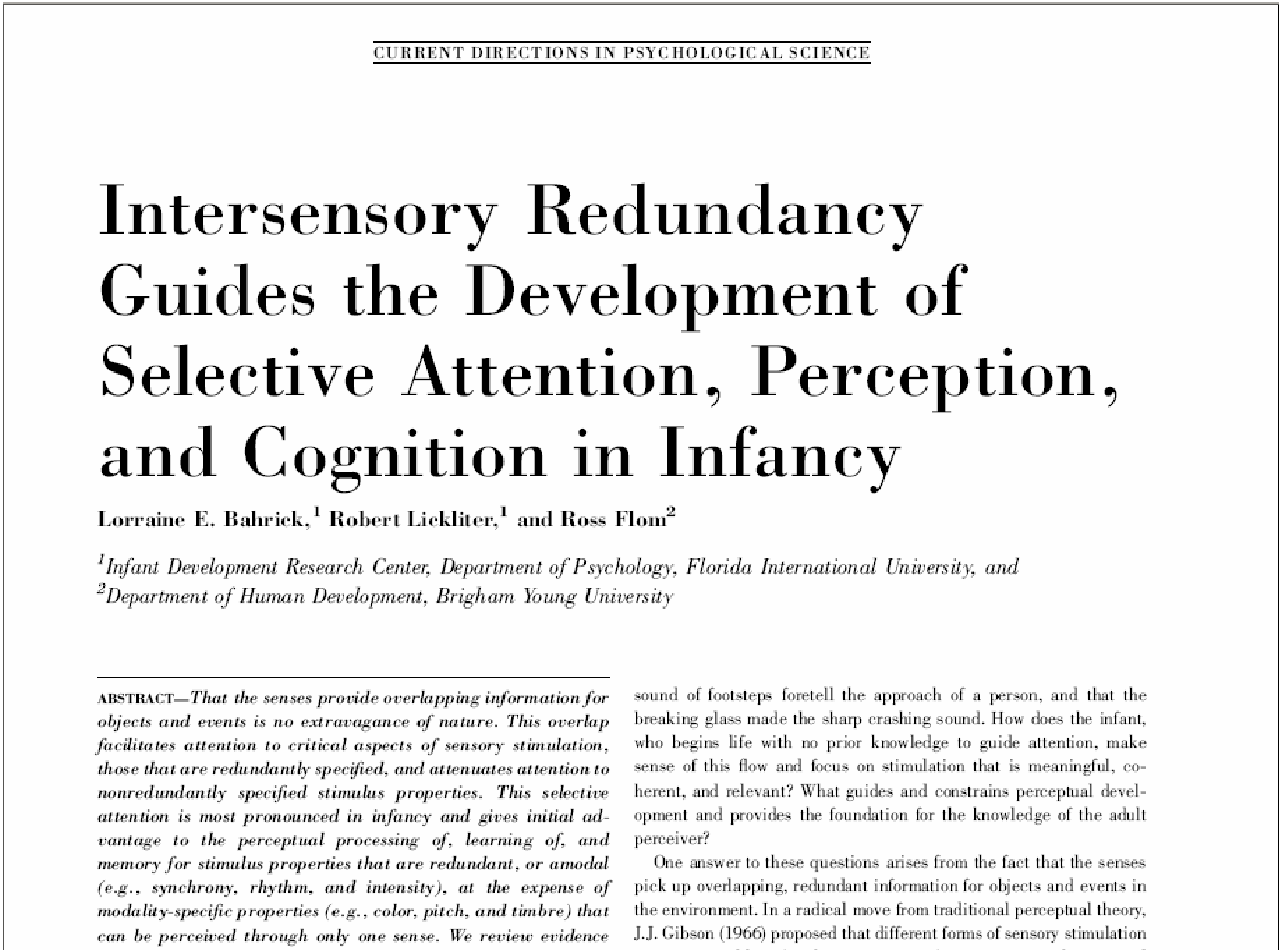
Results

Does it matter whether X_1, X_2 are conditionally independent?



Impact of Conditional Independence in unlabeled pairs





CURRENT DIRECTIONS IN PSYCHOLOGICAL SCIENCE

Intersensory Redundancy Guides the Development of Selective Attention, Perception, and Cognition in Infancy

Lorraine E. Bahrick,[1] Robert Lickliter,[1] and Ross Flom[2]

[1]*Infant Development Research Center, Department of Psychology, Florida International University, and*
[2]*Department of Human Development, Brigham Young University*

ABSTRACT—*That the senses provide overlapping information for objects and events is no extravagance of nature. This overlap facilitates attention to critical aspects of sensory stimulation, those that are redundantly specified, and attenuates attention to nonredundantly specified stimulus properties. This selective attention is most pronounced in infancy and gives initial advantage to the perceptual processing of, learning of, and memory for stimulus properties that are redundant, or amodal (e.g., synchrony, rhythm, and intensity), at the expense of modality-specific properties (e.g., color, pitch, and timbre) that can be perceived through only one sense. We review evidence*

sound of footsteps foretell the approach of a person, and that the breaking glass made the sharp crashing sound. How does the infant, who begins life with no prior knowledge to guide attention, make sense of this flow and focus on stimulation that is meaningful, coherent, and relevant? What guides and constrains perceptual development and provides the foundation for the knowledge of the adult perceiver?

One answer to these questions arises from the fact that the senses pick up overlapping, redundant information for objects and events in the environment. In a radical move from traditional perceptual theory, J.J. Gibson (1966) proposed that different forms of sensory stimulation

Infant Learning and Intersensory Redundancy

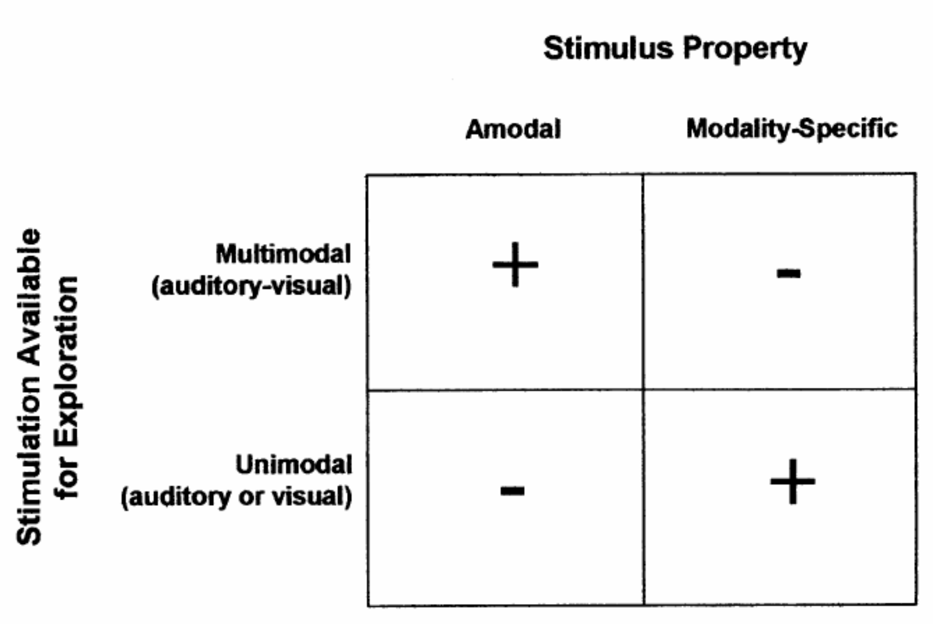
- Infants
  - 3 month olds attend to amodal properties (tempo of hammer) when given multisensory inputs, but not when given single modality input [Bahrick et al., 2002]

- Animals
  - Quail embryos learned an individual maternal call 4x faster when given multisensory data (synchronizing light with rate and rhythm of the sound) [Lickliter et al., 2002]

Intersensory Redundancy and Infant Development

[Bahrick & Lickliter, *Dev. Psy*, 2000]

- Intersensory redundancy: “spatially coordinated and temporally synchronous presentation of the same information across two or more senses”
- Sight & sound of ball bouncing
  Amodal property: tempo



		Stimulus Property	
		Amodal	Modality-Specific
Stimulation Available for Exploration	Multimodal (auditory-visual)	+	-
	Unimodal (auditory or visual)	-	+

Intersensory Redundancy Hypothesis [Bahrick & Lickliter]:

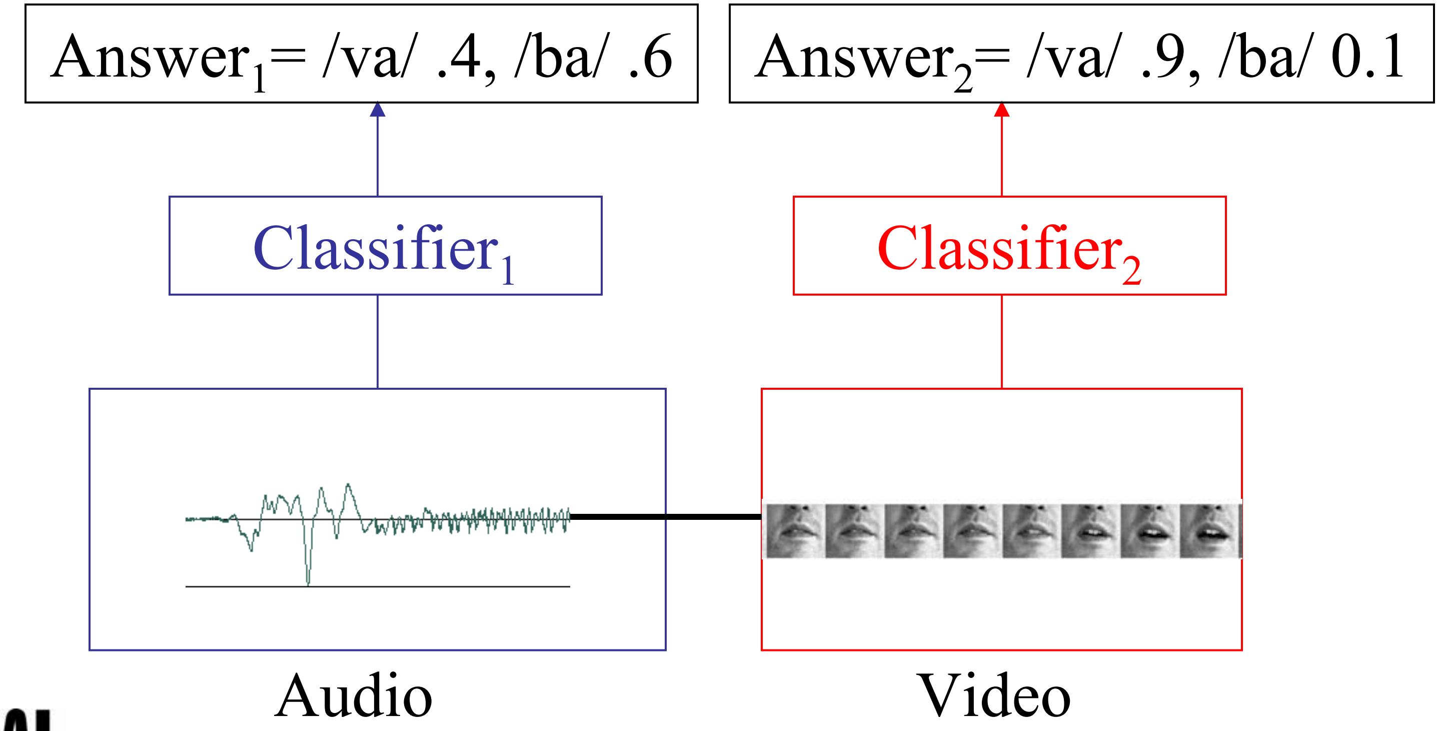
1. Learning of <u>amodal</u> properties is facilitated by multimodal stimulation
2. Learning of <u>modality-specific</u> properties facilitated by unimodal stimulation
3. These effects are most pronounced in early development

Co-Training

Where else might this work?

- learning to recognize phonemes/vowels

[de Sa, 1994; Coen 2006]



Answer$_1$= /va/ .4, /ba/ .6

Answer$_2$= /va/ .9, /ba/ 0.1

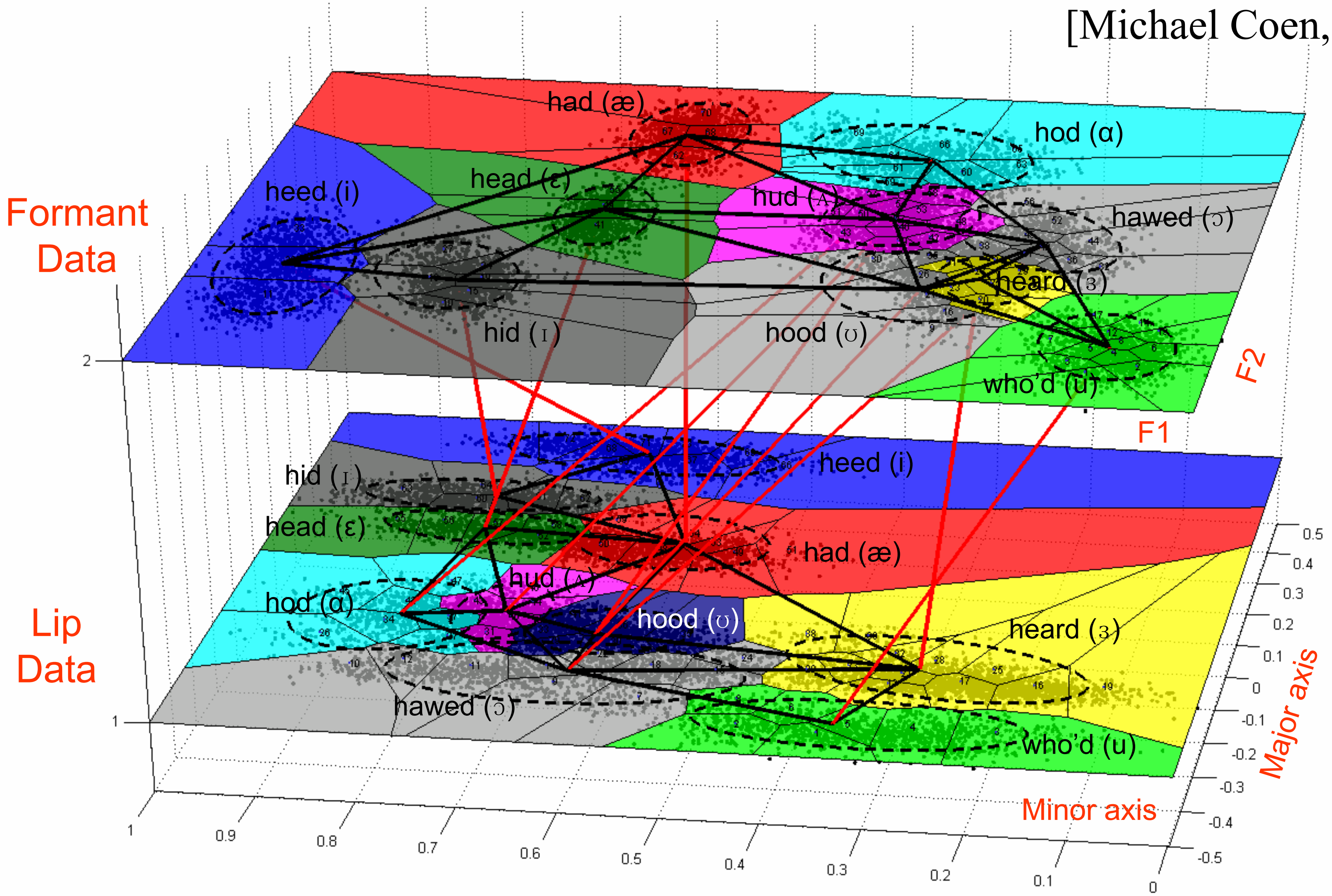
Classifier$_1$

Classifier$_2$

Audio

Video

[Michael Coen, 2006]



Mutual clustering

CoTraining Summary

- Unlabeled data improves supervised learning when example features are redundantly sufficient and only weakly (conditionally) correlated

- Theoretical results
  - If X1,X2 conditionally independent given Y
    - PAC learnable from weak initial classifier plus unlabeled data
    - disagreement between g1(x1) and g2(x2) bounds final classifier error
  - Disagreement between classifiers over unlabeled examples predicts true classification error

- Aligns with developmental psychology claims about importance of multi-sensory input

- Unlabeled conditionally independent pairs improve second language learning in humans
  - But dependent pairs are also helpful !

Human and Machine Learning

Additional overlaps:

- Learning representations for perception
  - Dimensionality reduction methods, low level percepts
  - Lewicky et al.: optimal sparse codes of natural scenes yield gabor filters found in primate visual cortex
- Learning using prior knowledge
  - Explanation-based learning, graphical models, teaching concepts & skills, chunking
  - VanLehn et al: explanation-based learning accounts for some human learning behaviors
- Learning multiple related outputs
  - MultiTask learning, teach multiple operations on the same input
  - Caruana: patient mortality predictions improve if same predictor must also learn to predict ICU status, WBC, etc.

Some questions and conjectures

One learning mechanism or many?

- Humans:
  - Implicit and explicit learning (unaware/aware)
  - Radically different time constants in synaptic changes (minutes) versus long term memory consolidation (days)
- Machines:
  - Inductive, data-intensive algorithms
  - Analytical compilation, knowledge + data

Conjecture:

In humans two very different learning processes.

Implicit largely inductive, Explicit involves self-explanation

Predicts: if an implicit learning task can be made explicit, it will be learnable from less data

Can Hebbian Learning Explain it All?

- Humans:
  - It is the only synapse-level learning mechanism currently known
  - It is also known that new neurons grow, travel, and die

Conjecture:

Yes, much of human learning will be explainable by Hebbian learning, just as much of computer operation can be explained by modeling transistors. Even two different learning mechanisms.

But much will need to be understood at an architectural level. E.g., what architectures could implement goal supervised learning in terms of Hebbian mechanisms?

What is Learned, What Must be Innate?

We don’t know. However, we do know:

- Low level perceptual features can emerge from unsupervised exposure to perceptual stimuli [e.g., M. Lewicky].
  - Natural visual scenes → Gabor filters similar to those in visual cortex
  - Natural sounds → basis functions similar to those in auditory cortex
- Semantic object hierarchies can emerge from observed ground-level facts
  - Neural network model [McClelland et al]
- ML models can help determine what representations can emerge from raw data.