Statistical learning theory and communication

E.P. Stabler, A.N.G. Kirschel, C.E. Taylor, UCLA

Bioacoustic Monitoring Workshop • James Reserve • 2008

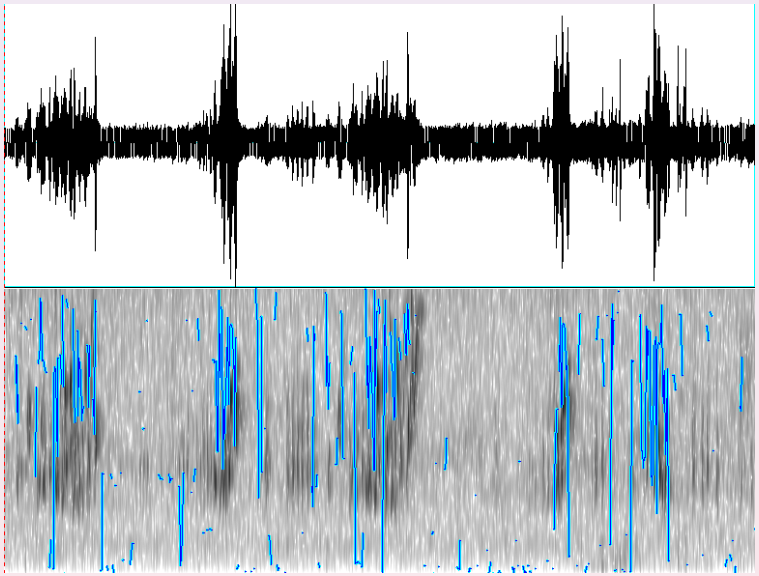
Learning from sensor data

- Models of the environment
- Individual and species identification

  (Trifa et al'08,Vallejo et al'07,Vilches et al'07)
- Organisms as situated agents. . .

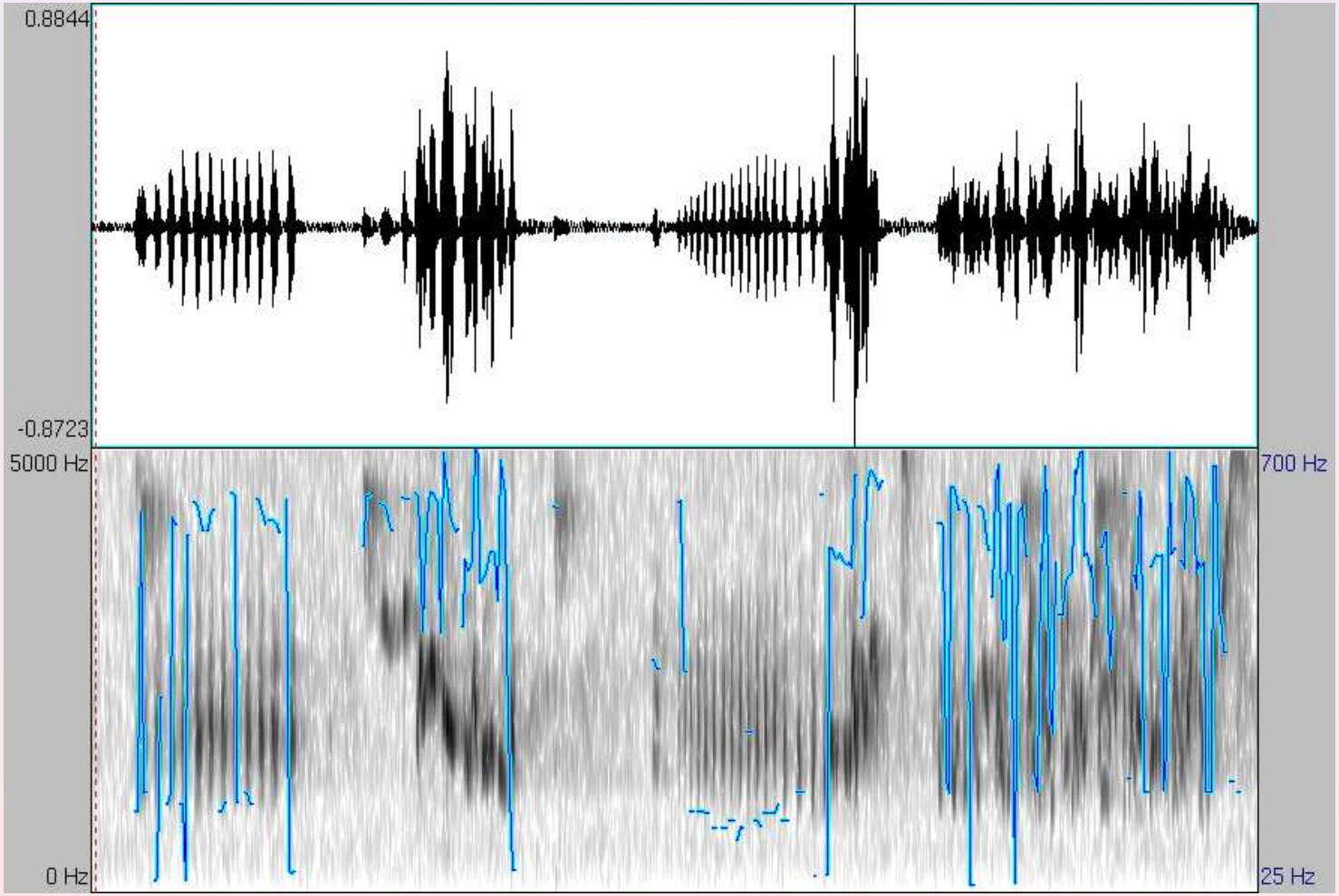
We have significant developments in data collection and computational resources, *and significant developments in modeling methods*. But we need better models!

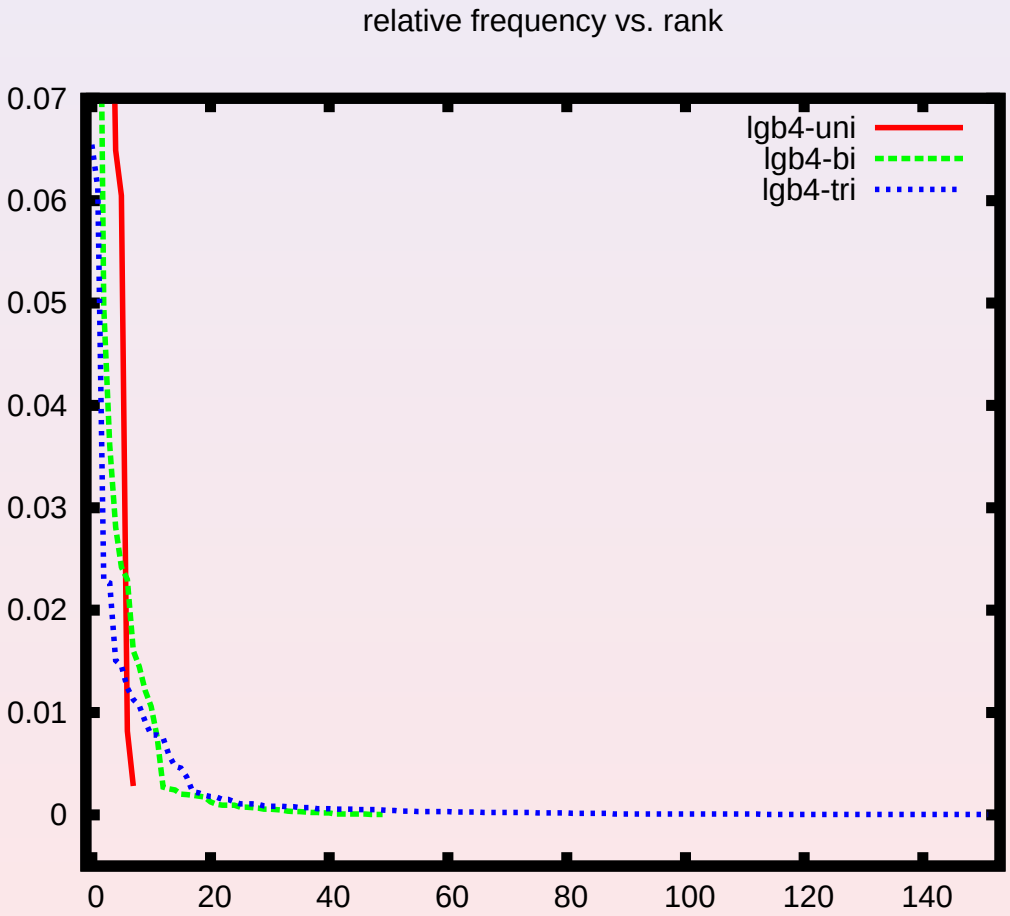
Little Greenbul:



- what is the bird doing? what sequences in these songs?
- how is the bird doing that?

>20K songs; here 4 of 6 types:



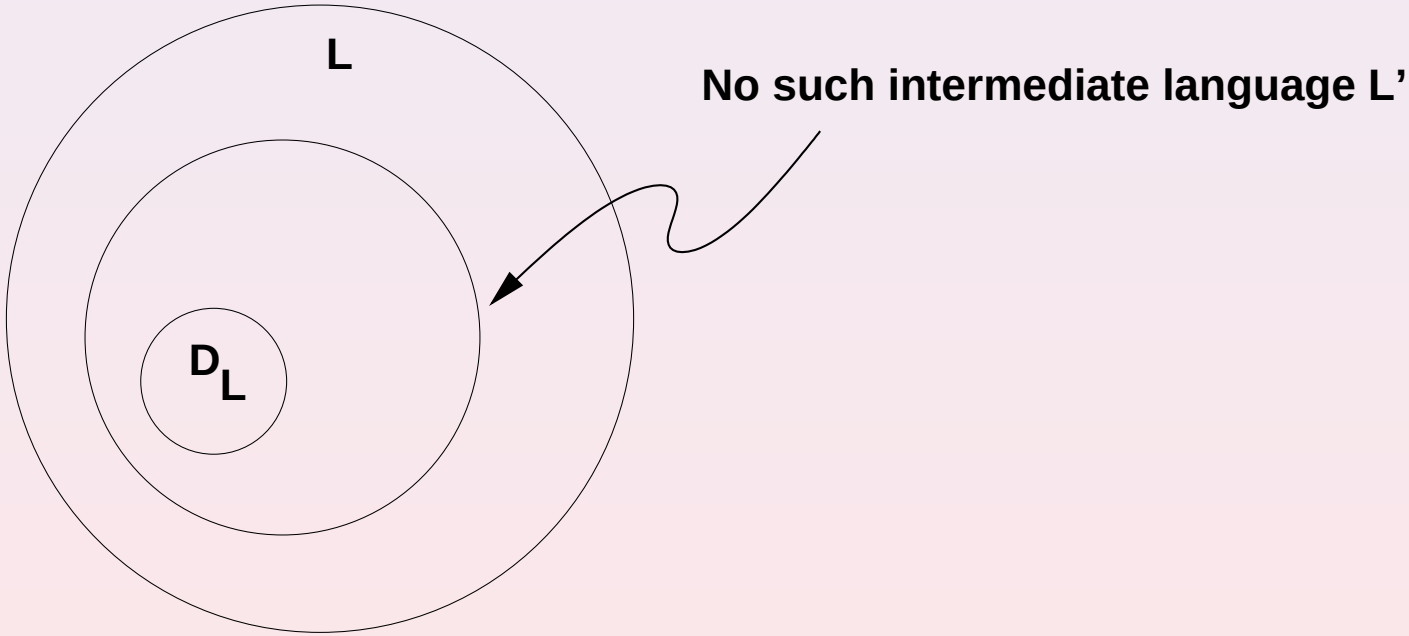


(Zipf effects: $\frac{2}{3}$ trigrams absent. 26% trigrams occur 1; 42% occur 2; >80% Ss occur 1)

learning as identification in the limit: (Gold, 1967)

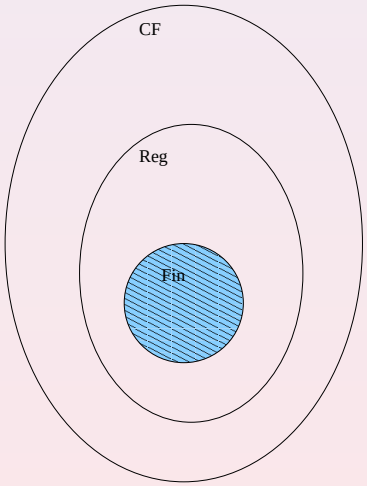
- **evidence:** an enumeration t of a dataset L
- **learner** ϕ**:** initial segments of enumerations $\longrightarrow$ model G
- ϕ **learns** t**:** iff ϕ converges to a generator G of L
- **L is learnable:** some ϕ learns every enumeration of L
- **class** $\mathcal{L}$ **is learnable** iff every L$\in \mathcal{L}$ is

(Angluin, 1980) $\mathcal{L}$ is identifiable in the limit from examples iff every L$\in \mathcal{L}$ has a finite subset D such that

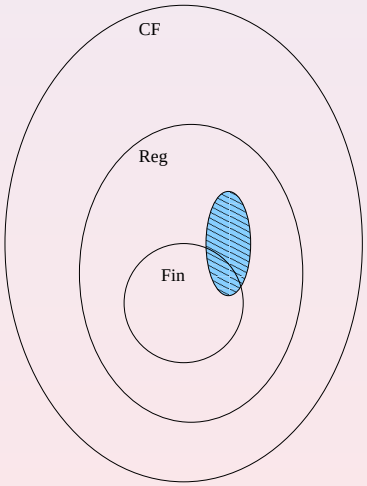


(Pitt, 1989) If $\mathcal{L}$ identifiable with $p > \frac{1}{2}$, then identifiable in the limit

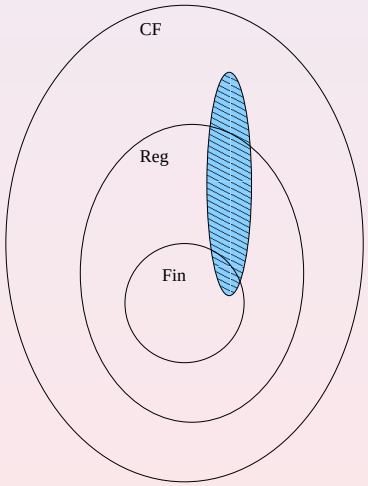
(Gold: positive results)



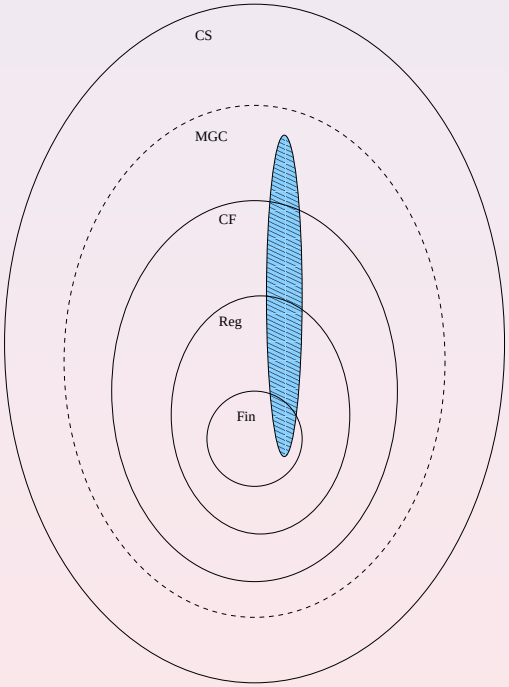
Gold 1967
finite



Angluin 1982
reversible



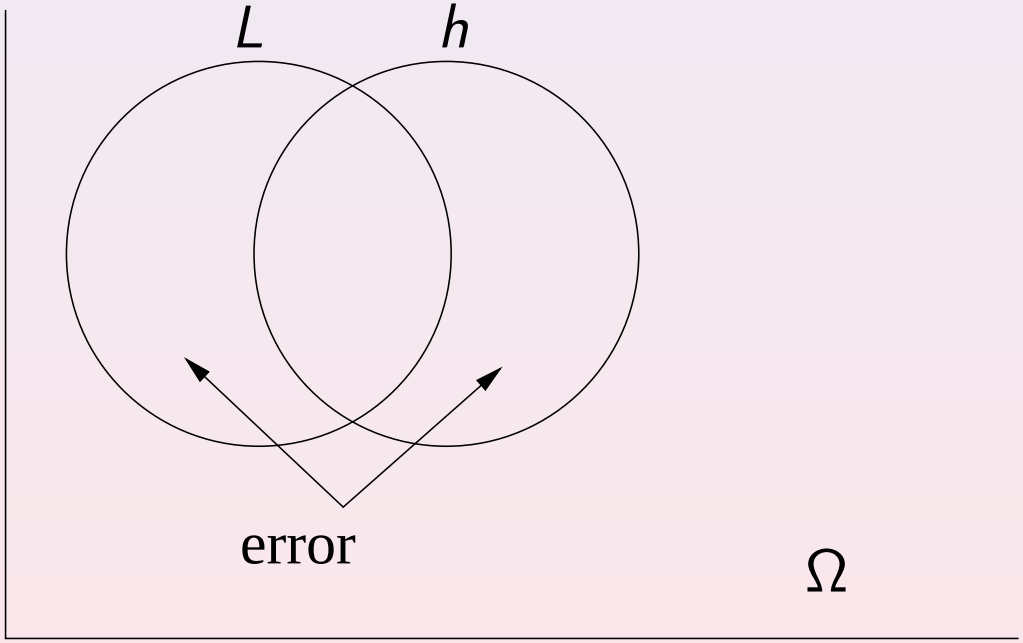
Yokomori 2003
simple CF



Stabler et al 2003
rigid PMCF

(nb: the models may be probabilistic, but the success criterion is Boolean)

Learning as (distribution-free) approximate learning (PAC)



With any μ on Ω, learner's hypothesis h has error $\leq \epsilon$, with probability $\geq 1 - \delta$, after $m(\epsilon, \delta)$ labeled samples from Ω

Learning as (distribution-free) approximate learning (PAC)

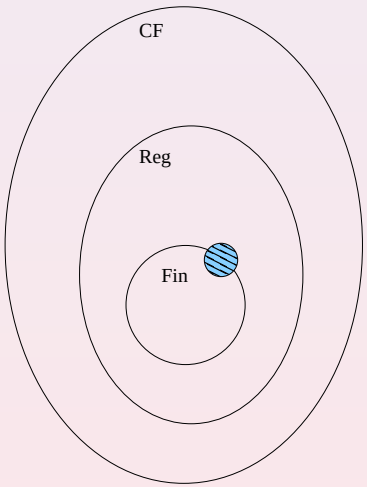
- **evidence:** $EX(L,\mu)^n$ a sample of length n drawn from Ω w.r.t. μ, labeled by target concept $L \in \mathcal{L}$
- **learner:** a function ϕ from samples to hypotheses $h \in \mathcal{H}$
- **class $\mathcal{L}$ is PAC learnable** iff $\exists\phi$ and $m : [0,1]^2 \rightarrow \mathbb{N}$ such that $\forall\mu\ \forall L \in \mathcal{L}, \forall 0 < \epsilon, \delta < 1$

$$\phi(EX(L,\mu)^{m(\epsilon,\delta)}) = h \text{ where } \mu(error_\mu(L,h) \leq \epsilon) \geq (1-\delta)$$

(Vapnik & Chervonenkis 1971, Blumer et al 1989, Alon et al. 1997, Cucker & Smale 2001, Mukherjee et al. 2004, Poggio et al. 2004,...)

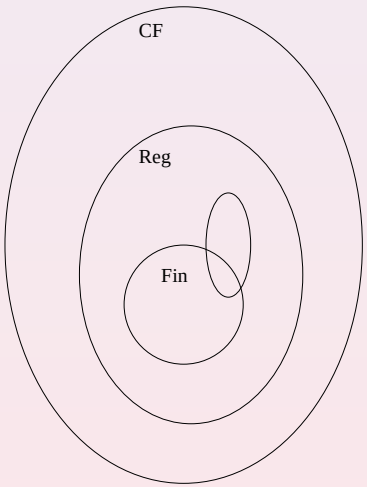
class $\mathcal{L}$ is PAC learnable
$\equiv$ ERM consistent
$\equiv$ identifiable by a 'leave-one-out stable' method
$\equiv$ $VC(\mathcal{L})$ finite
$\equiv$ uniform Glivenko-Cantelli (uGC)

(PAC, ERM and relatives: positive results)



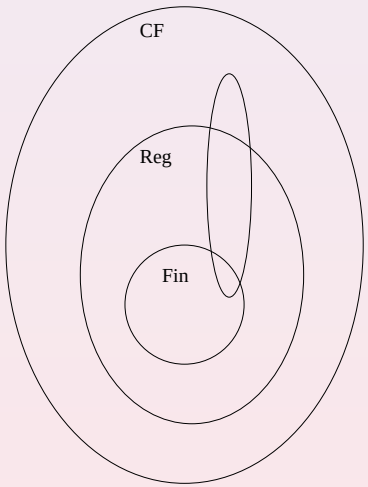
Gold 1967

~~fin~~



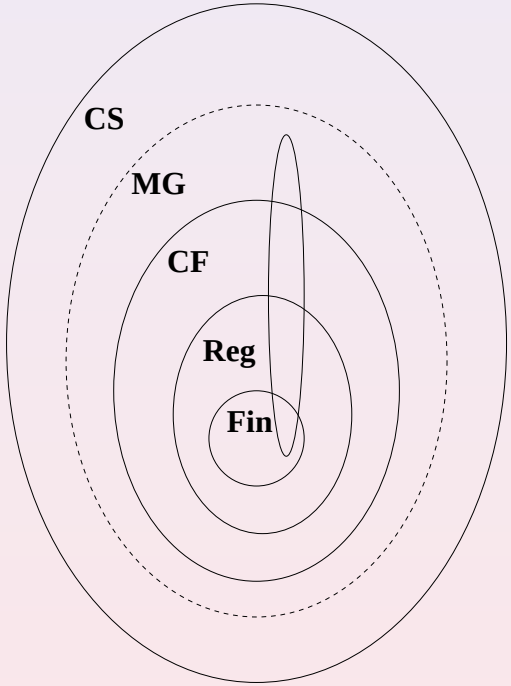
Angluin 1982

~~reversible~~



Yokomori 2003

simple CF?



Stabler et al 2003

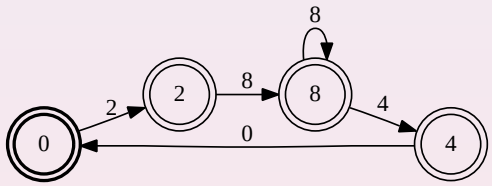
rigid PMCF?

(most (all?) classes shown earlier have ∞ VC dimension; not PAC)

bounded FSA (n-grams, Markov Models, etc)

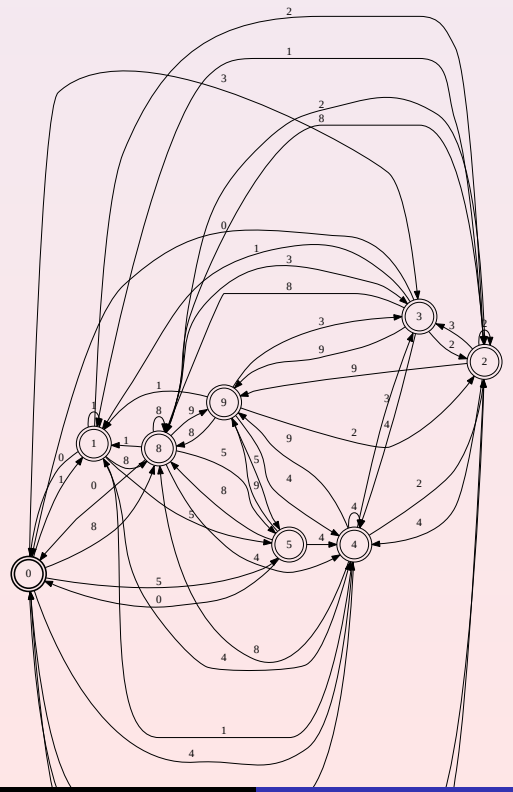
. . . (02888840,205) . . . (028888840,124) . . . (0288840,105) . . . (0288888840,36) . . . (02888888840,7) . . . (028888888840,6) . . .

	0	2	4	8
0	0.	1.	0.	0.
2	0.	0.	0.	1.
4	1.	0.	0.	0.
8	0.	0.	0.125	0.875



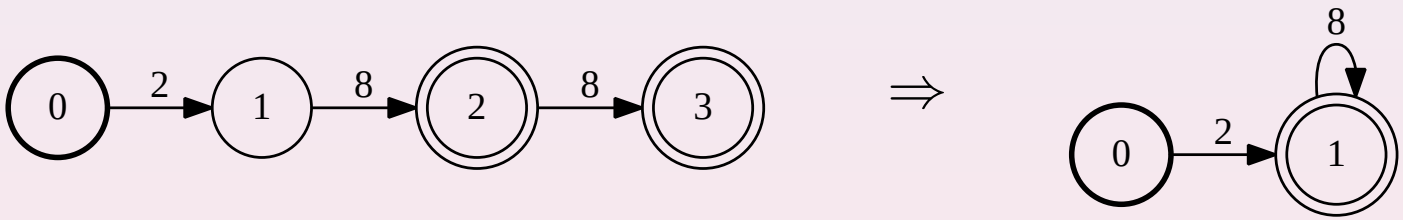
- Gold and PAC learnable
- in Greenbul song, bounds not known
- good model? decide by p(heldout) and simplicity of model

	0	1	2	3	4	5	8	9
0	0.000000	0.207968	0.330182	0.404456	0.013504	0.014855	0.029034	0.000000
1	0.013121	0.874113	0.019149	0.000000	0.091489	0.000709	0.000709	0.000709
2	0.348837	0.017054	0.009302	0.007752	0.065116	0.000000	0.528682	0.023256
3	0.225000	0.016667	0.009722	0.000000	0.022222	0.000000	0.715278	0.011111
4	0.907216	0.003436	0.017182	0.008591	0.001718	0.000000	0.012027	0.049828
5	0.363636	0.000000	0.000000	0.000000	0.030303	0.000000	0.606061	0.000000
8	0.014100	0.001356	0.010575	0.001898	0.208514	0.000542	0.761931	0.001085
9	0.000000	0.183908	0.310345	0.459770	0.022989	0.011494	0.011494	0.000000

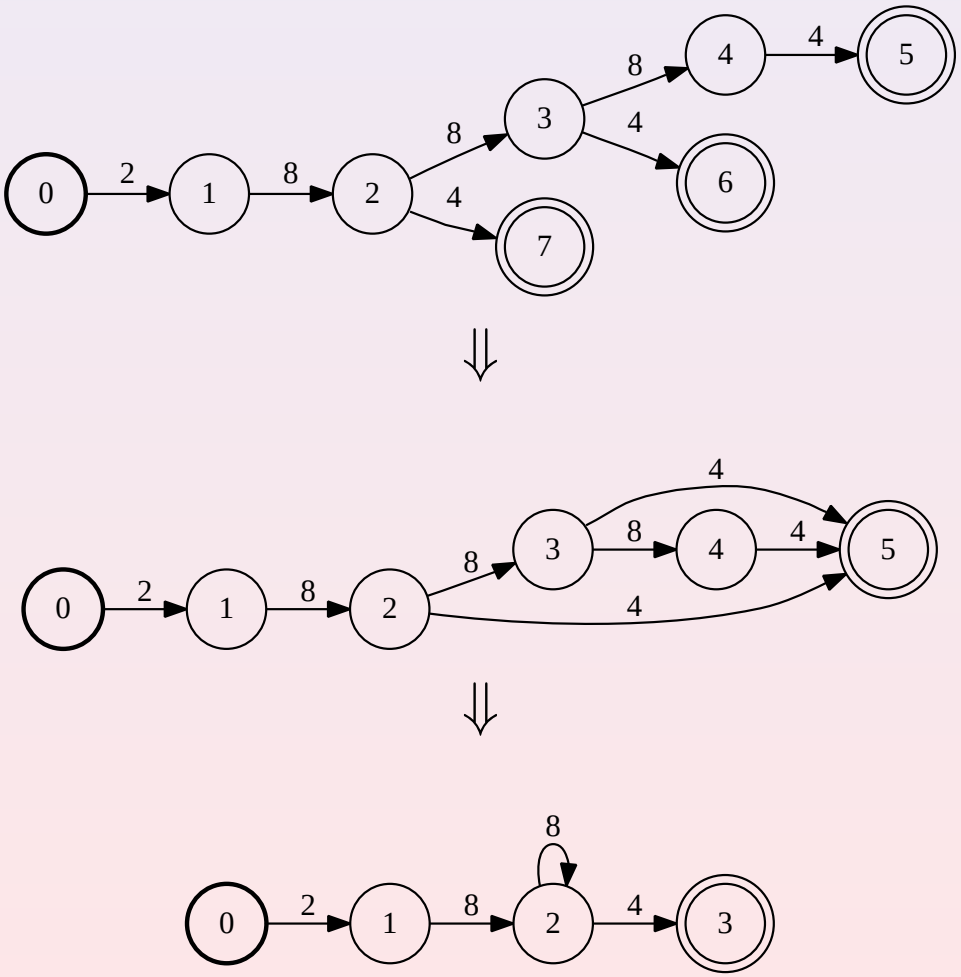


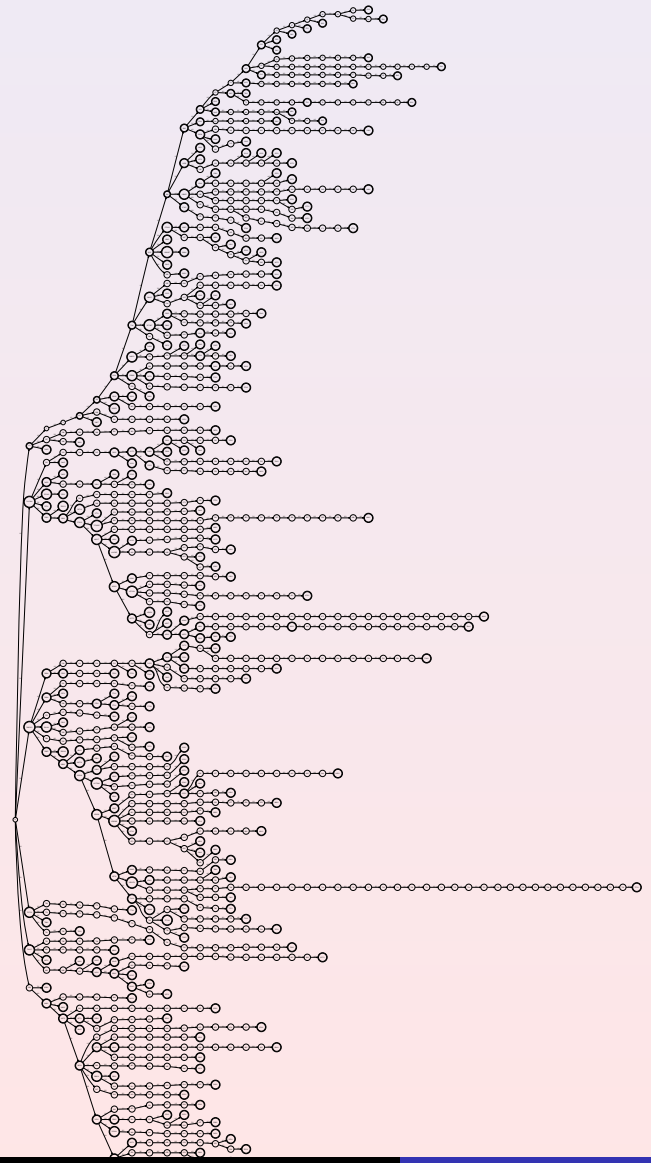
(Angluin '82)

Reversible FS (forward and backward deterministic)

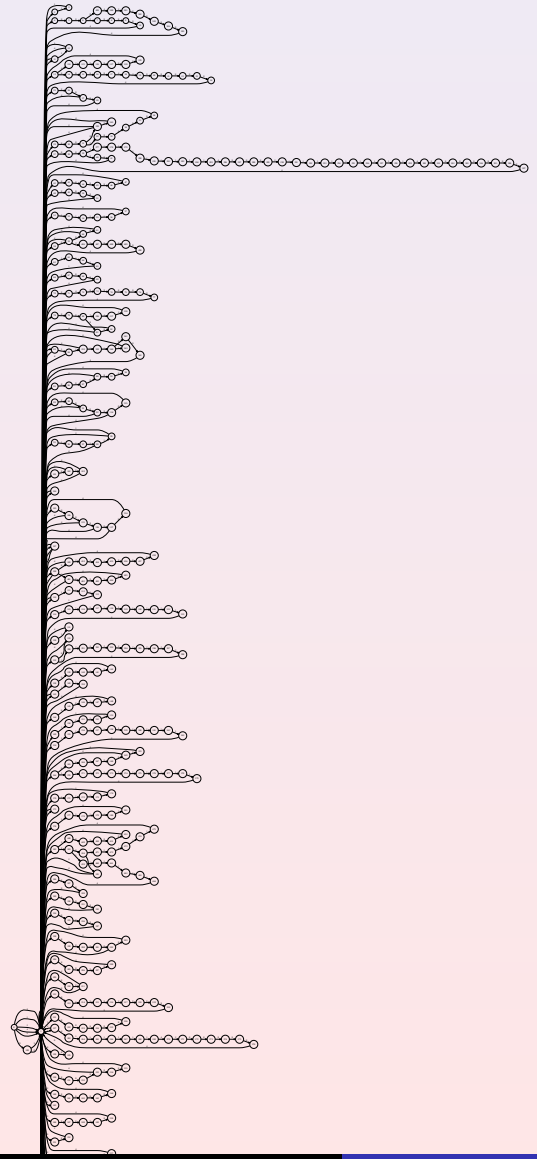


- no fixed finite bound (unlike *n*-gram, Markov models), not PAC
- applied to birdsong by (Sasahara et al'06)
- good model? compare p(heldout) (hunch:no, rev unevidenced)

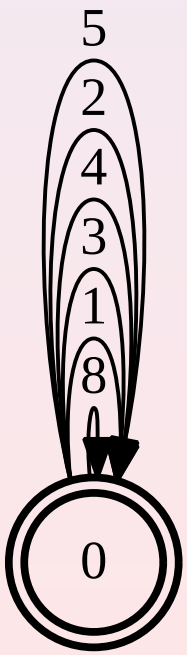

0
2
1
8
2
8
3
4
8
4
4
5
4
6
4
7
⇓
0
2
1
8
2
8
3
4
8
4
4
5
4
⇓
0
2
1
8
2
8
4
3

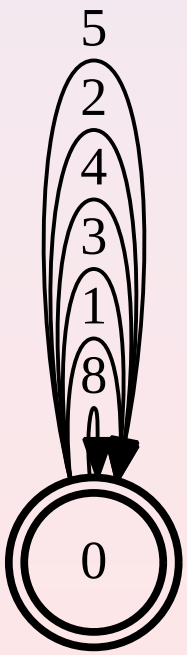


$\Rightarrow$



$\Rightarrow$


5
2
4
3
1
8
0



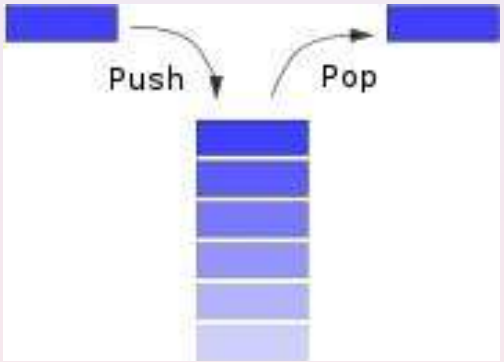
Sasahara et al propose relaxing reversibility to 'k-reversibility'...

We will look at more radical strategies...

(Yokomori'03)

Simple CF: $A \to a\alpha$ **for** $a \in \Sigma, \alpha \in N^*$**, at most one per** $a \in \Sigma$

N → * N N
N → + N N
N → - N
N → 0|1|2|...



S → 0 T E
T → 2
E → 8 E
E → 4

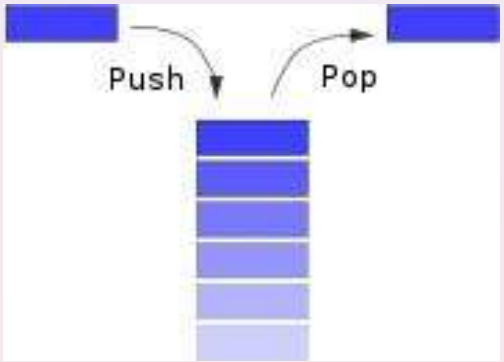
- Gentner et al'06: starlings recognize CF pattern (aabb)
- good model of Greenbul song?

...(0288884,205) ...(02888884,124) ...(028884,105) ...(028888884,36) ...(0288888884,7) ...(02888888884,6) ...

(Yokomori'03)

Simple CF: $A \to a\alpha$ **for** $a \in \Sigma, \alpha \in N^*$**, at most one per** $a \in \Sigma$

N → * N N
N → + N N
N → - N
N → 0|1|2|...



S → 0 T E
T → 2
E → 8 E
E → 4

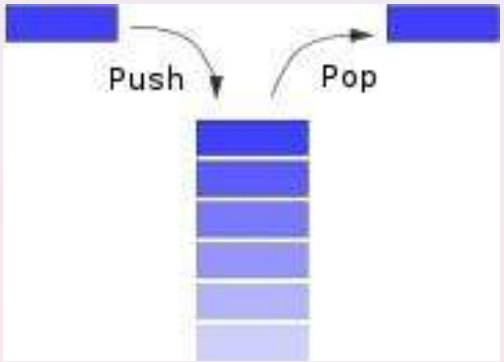
- Gentner et al'06: starlings recognize CF pattern (aabb)
- good model of Greenbul song? (no: data not generable!, and misses repetition)

...(0288884,205) ...(02888884,124) ...(028884,105) ...(028888884,36) ...(0288888884,7) ...(02888888884,6) ...

(Yokomori'03)

Simple CF: $A \to a\alpha$ **for** $a \in \Sigma, \alpha \in N^*$**, at most one per** $a \in \Sigma$

N → * N N
N → + N N
N → - N
N → 0|1|2|...



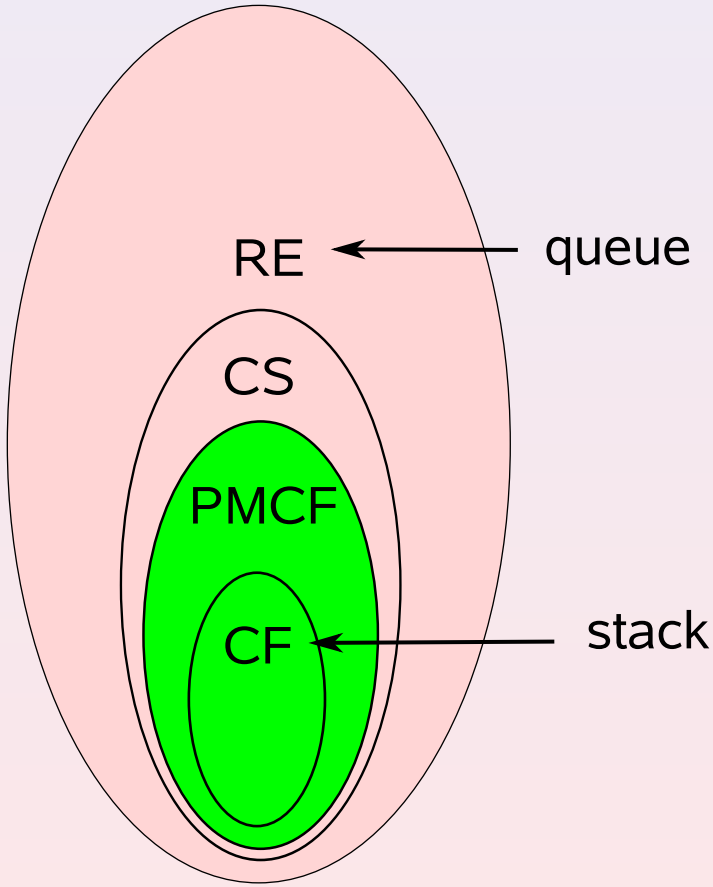
S → 0 T E
T → 2
E → 8 E
E → 4

- Gentner et al'06: starlings recognize CF pattern (aabb)
- good model of Greenbul song? (no: data not generable!, and misses repetition)

How can we get all the data and increase p(repetitions)?

...(0288884,205) ...(02888884,124) ...(028884,105) ...(028888884,36) ...(0288888884,7) ...(02888888884,6) ...

the power of the queue

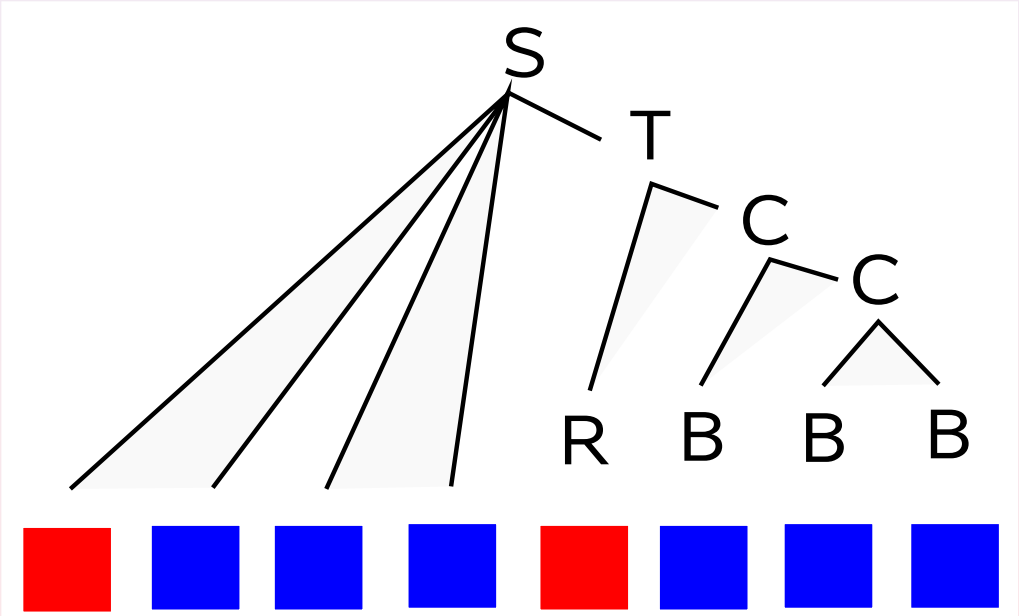


- with a queue, we have turing machine power (e.g. Li et al '93)
- idea: stack for computation + memory for input (PMCF)

(Seki et al'91,Stabler et al'03,Kobele'06)

Unambiguous PMCF

$$\begin{matrix} x \\ S \end{matrix} \longrightarrow \begin{matrix} x \\ T \end{matrix}, \quad \begin{matrix} xx \\ S \end{matrix} \longrightarrow \begin{matrix} x \\ T \end{matrix}$$

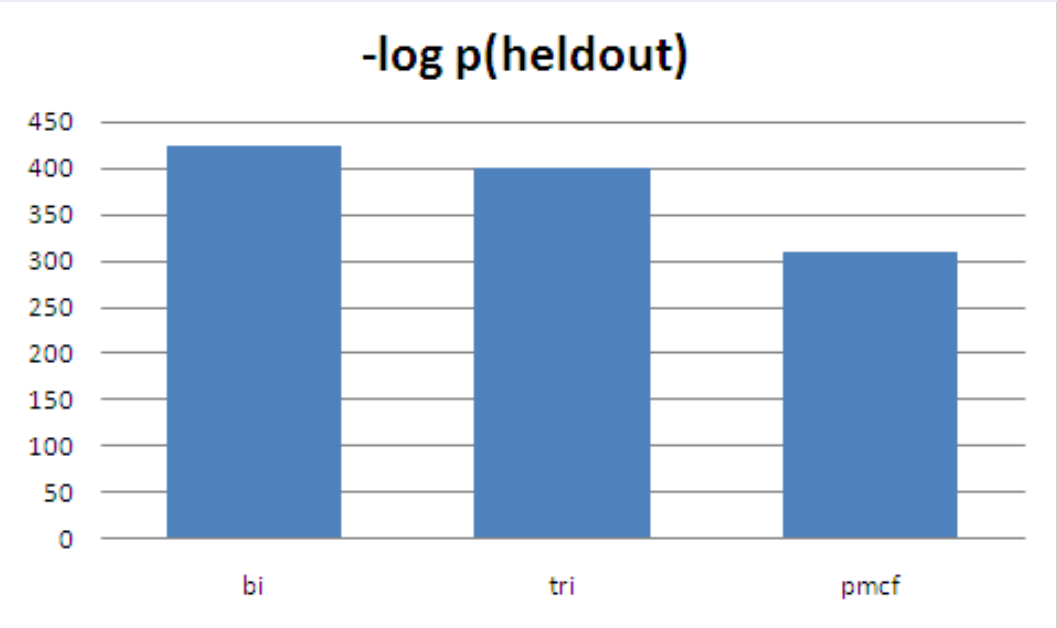


- rules allow copying, so p(copies)↑
- good model?

(hunch: maybe)

Comparison:

	-log p(heldout)
bigram	423.615064
trigram	398.922888
pmcf	308.870379



- not easy to keep such comparisons fair since the models differ in many ways,

  but these preliminary results confirm that small models with copying can fit the data very well

Prospects and future work:

- *n*-gram,Markov,PAC models to weak; too few other models
- H: birds produce and notice repetition in song sequences
  - Confirmed by our first comparison of Greenbul models

    (cf Mennill&Vehrencamp'08; Mann et al'06; Hill'04; Trainer&McDonald'95)
  - Interesting models of gestural iteration & timing

    (cf Nam et al'08; Port'03; Beek et al'02; Saltzman&Byrd'00; Wallenstein et al'95)
- Many questions we would like to answer
  - How much individual variation among birds in same locale?
  - What kinds of information are communicated by birdsong?
  - What neural mechanisms recognize/produce iteration?

Alon, Noga, Shai Ben-David, Nicolò Cesa-Bianchi, and David Haussler. 1997. Scale-sensitive dimensions, uniform convergence, and learnability. Journal of the Association for Computing Machinery, 44(4):615–631.

Angluin, Dana. 1980. Inductive inference of formal languages from positive data. Information and Control, 45:117–135.

Angluin, Dana. 1982. Inference of reversible languages. Journal of the Association for Computing Machinery, 29:741–765.

Beek, P. J., C. E. Peper, and A. Daffertshofer. 2002. Modeling rhythmic interlimb coordination: Beyond the Haken-Kelso-Bunz model. Brain and Cognition, 48(1):149–165.

Blumer, Anselm, Andrzei Ehrenfeucht, David Haussler, and Manfred K. Warmuth. 1989. Learnability and the Vapnik-Chervonenkis dimension. Journal of the Association for Computing Machinery, 36:929–965.

Cristianini, Nello and John Shawe-Taylor. 2000. Support Vector Machines and Other Kernel-Based Learning Methods. Cambridge University Press, NY.

Cucker, Felipe and Stephen Smale. 2001. On the mathematical foundations of learning. Bulletin of the American Mathematical Society, 39(1):1–49.

Duda, Richard O., Peter E. Hart, and David G. Stork. 2001. Pattern Classification. Wiley, NY.

Fitzsimmons, Lauren P., Jennifer R. Foote, Laurene M. Ratcliffe, and Daniel J. Mennill. 2008. Frequency matching, overlapping and movement behavior in diurnal countersinging interactions of black-capped chickadees. Animal Behaviour, 75:1913–1920.

Gene V. Wallenstein, J.A. Scott Kelso, Steven L. Bressler. 1995. Phase transitions in spatiotemporal patterns of brain activity and behavior. Physica D: Nonlinear Phenomena, 3-4(84):626–634.

Gold, E. Mark. 1967. Language identification in the limit. Information and Control, 10:447–474.

Hastie, T., R. Tibshirani, and J.H. Friedman. 2001. The Elements of Statistical Learning: Data Mining, Inference, and Prediction. Springer Series in Statistics. Springer, NY.

Hill, M. L. 2004. A review of hypotheses for the functions of avian duetting. Behavioral Ecology and Sociobiology, 5(5):415–430.

Kakishita, Y., K. Sasahara, T. Nishino, M. Takahasi, and K. Okanoya. 2008. Ethological data mining: An automata-based approach to extract behavioral units and rules. Data Mining and Knowledge Discovery, in press.

Kearns, Michael J. and Umesh V. Vazirani. 1994. An Introduction to Computational Learning Theory. MIT Press, Cambridge, Massachusetts.

Kobele, Gregory M. 2006. Generating Copies: An Investigation into Structural Identity in Language and Grammar. Ph.D. thesis, UCLA.

Li, Ming, Luc Longpré, and Paul Vitányi. 1993. The power of the queue. SIAM Journal on Computing, 21(4):697–712.

Mann, Nigel I., Kimberly A. Dingess, and P. J. B. Slater. 2006. Antiphonal four-part synchronized chorusing ina a Neotropical wren. Biological Letters, 2:1–4.

Mendelson, Shahar. 2004. Geometric parameters in learning theory. In Geometric Aspects of Functional Analysis, Lecture Notes in Mathematics. Springer, Berlin.

Mennill, Daniel J. and Sandra L. Vehrencamp. 2008. Context-dependent functions of avian duets revealed by microphone-array recordings and multispeaker playback. Current Biology, 18:1314–1319.

Mukherjee, Sayan, Partha Niyogi, Tomaso Poggio, and Ryan Rifkin. 2004. Learning theory: stability is sufficient for generalization and necessary and sufficient for consistency of Empirical Risk Minimization. Advances in Computational Mathematics, 25(1-3):161–193.

Nam, Hosung, Louis Goldstein, and Elliot Saltzman. 1991. Self-organization of syllable structure: a coupled oscillator model. In F. Pellegrino, E. Marisco, and I. Chitoran, editors, Approaches to Phonological Complexity. Mouton de Gruyter, Berlin, pages 361–376.

Niyogi, Partha. 2006. The Computational Nature of Language Learning and Evolution. MIT Press, Cambridge, Massachusetts.

Pitt, Leonard. 1989. Probabilistic Inductive Inference. Ph.D. thesis, University of Illinois.

Pitt, Leonard and Leslie Valiant. 1988. Computational limitations on learning from examples. Journal of the Association for Computing Machinery, 35:965–984.

Poggio, Tomaso, Ryan Rifkin, Partha Niyogi, and Sayan Mukherjee. 2004. General conditions for predictivity in learning theory. Nature, 428:419–422.

Port, Robert F. 2003. Meter and speech. Journal of Phonetics, 31(3-4):599–611.

Saltzman, Elliot and Dani Byrd. 2000. Task-dynamics of gestural timing: Phase windows and multifrequency rhythms. Human Movement Science, 19:499–526.

Sasahara, K., Y. Kakishita, T. Nishino, M. Takahasi, and K. Okanoya. 2006. A reversible automata approach to modeling birdsongs. In 15th IEEE International Conference on Computing.

Seki, Hiroyuki, Takashi Matsumura, Mamoru Fujii, and Tadao Kasami. 1991. On multiple context-free grammars. Theoretical Computer Science, 88:191–229.

Slabbekoorn, H. and T. B. Smith. 2002. Habitat-dependent song divergence in the little greenbul: an analysis of environmental selection pressures on acoustic signals. Evolution, 56:1849–1858.

Stabler, Edward P. 2004. Varieties of crossing dependencies: Structure dependence and mild context sensitivity. Cognitive Science, 93(5):699–720.

Stabler, Edward P., Travis C. Collier, Gregory M. Kobele, Yoosook Lee, Ying Lin, Jason Riggle, Yuan Yao, and Charles E. Taylor. 2003. The learning and emergence of mildly context sensitive languages. In W. Banzhaf, T. Christaller, P. Dittrich, J.T. Kim, and J. Ziegler, editors, Advances in Artificial Life. Springer, NY.

Trainer, J. M. and D. B. McDonald. 1995. Singing performance, frequency matching and courtship success of long-tailed manakins chiroxiphia linearis. Behavioral Ecology and Sociobiology, 37:249–254.

Trifa, V., A. N. G. Kirschel, C. E. Taylor, and E. Vallejo. 2008. Automated species recognition of antbirds in a mexican rainforest using hidden markov models. Journal of Acoustical Society of America, 123:2424–2431.

Valiant, Leslie G. 1984. A theory of the learnable. Communications of the Association for Computing Machinery, 27(11):1134–1142.

Vallejo, E. E., M. L. Cody, and C. E. Taylor. 2007. Unsupervised acoustic classification of bird species using hierarchical self-organizing maps. In M. Randall, H. A. Abbass, and J. Wiles, editors, Progress in Artificial Life: Third Australian Conference, ACAL2007, LNAI 4828. Springer-Verlag, Berlin.

Vapnik, V.N. and A.Y. Chervonenkis. 1971. On the uniform convergence of relative frequencies of events to their probabilities. Theory of Probability and its Applications, 16:264–280.

Yokomori, Takashi. 2003. Polynomial-time identification of very simple grammars from positive data. Theoretical Computer Science, 298:179–206.