
welcome
to Lecture 6
Computation and the Brain

# First:  What happened last Wednesday

# Kiran’s talk on mapping fMRI data to stimulus semantics

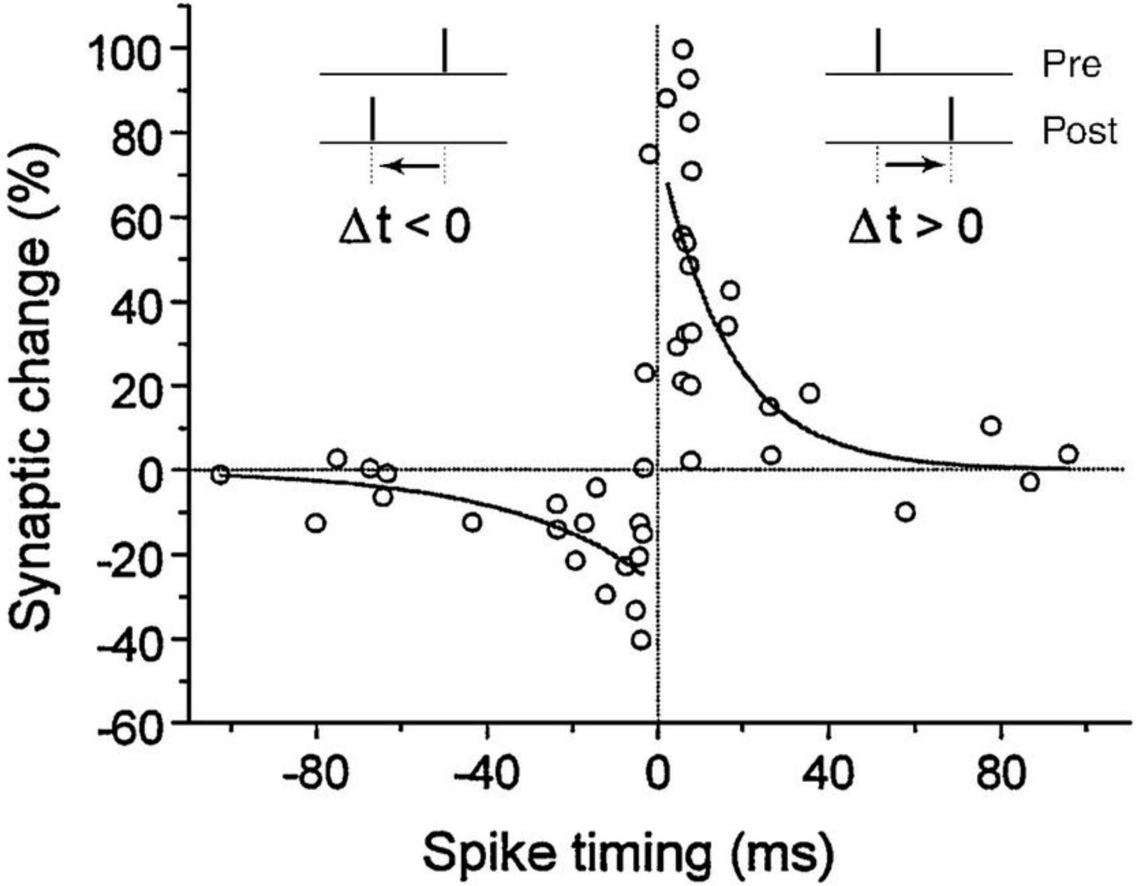
- Learned a map from fMRI data from selected areas of the brain of subjects watching a movie to a corpus of annotations on the movie
- Also: discussion of word embeddings, a useful technique that encodes words in a corpus as vectors in $R^d$

# Jacob’s talk on computation in the fly’s brain

- A treasure trove of information, ideas and project topics

# Concluding our treatment of synaptic plasticity: Spike timing-dependent plasticity (STDP)

If spike arrives in time, some gain.  Just in time, **big gain**. If it misses it, some loss. Just misses it, **big loss**.



# Bengio et al. 2016 ``Towards biologically plausible deep learning" through STDP

- Gradient descent: $\Delta x^t = \alpha\, \delta(t)\, \nabla f(x^t)$ update happens at time t
- SDTP: $\Delta w^t = \beta\, \delta(t) \nabla V(w^t)$ t is the time the spike arrives at the synapse
- Some similarity, huh?
- Idea: What if we use an STDP feedforward net to optimize some objective function whose "local derivative" is V?
- This idea is pursued in the paper; some learning can be done, but there are catches and different kinds of biological implausibility…

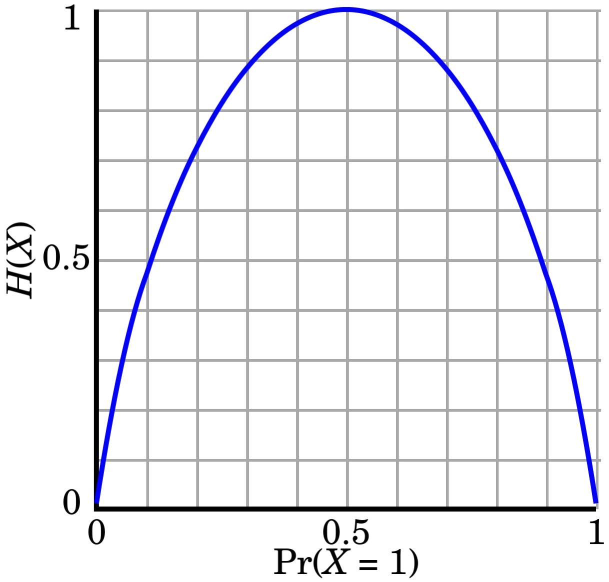
# Incidentally, my take on biological plausibility

- Deep nets are biologically plausible (in some well defined sense)
- Forward computation is of course plausible (e.g., the visual cortex)
- Backprop can be thought of as modeling evolution
- Assuming that the feedforward circuit and/or the weights are a phenotype that depends on *many* genes
- Minibatch: the collective experience of a generation
- Selection changes the allele statistics of the population

# Information Theory: Entropy of a distribution D

$$H(D) = -\Sigma_j \text{ Prob}[r_j] \log_2 (\text{Prob}[r_j])$$

Example: coin {p, 1-p}



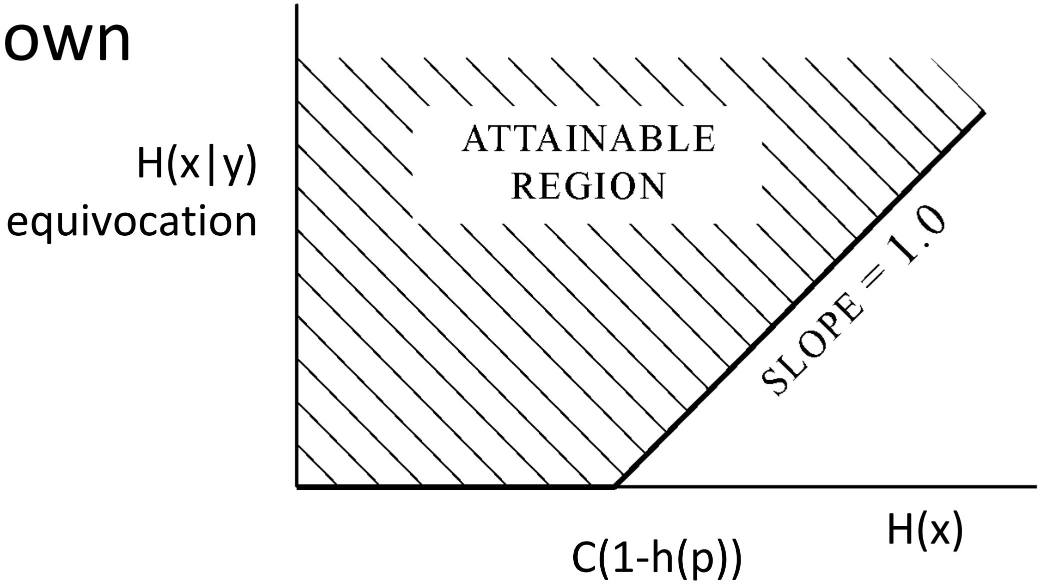
# Shannon's second theorem

**Theorem 2:** If the channel has capacity C and noise p < ½ then

(a) Any rate R < C (1– h(p)) can be achieved by coding

(b) No rate greater than C (1 – h(p)) can be achieved

(c) If equivocation – uncertainty in decoding, H(x|y) > 0 – is allowed, then the attainable region is as shown



# Shannon's second theorem, part (a)

**Theorem 2:** If a channel has capacity C and noise $p < ½$ then

(a) Any rate $R < C\,(1 - h(p))$ can be achieved by coding

**Proof of (a):** Consider a long bit string B of length m. Map each such B to a random bit string c(B) of length $m + r$, where r is the redundancy afforded by the excess of C over R. $2^m$ such codewords

Remarkably, after c(B) is received as a corrupted bit string B', B can be recovered by finding the closest codeword to B'.

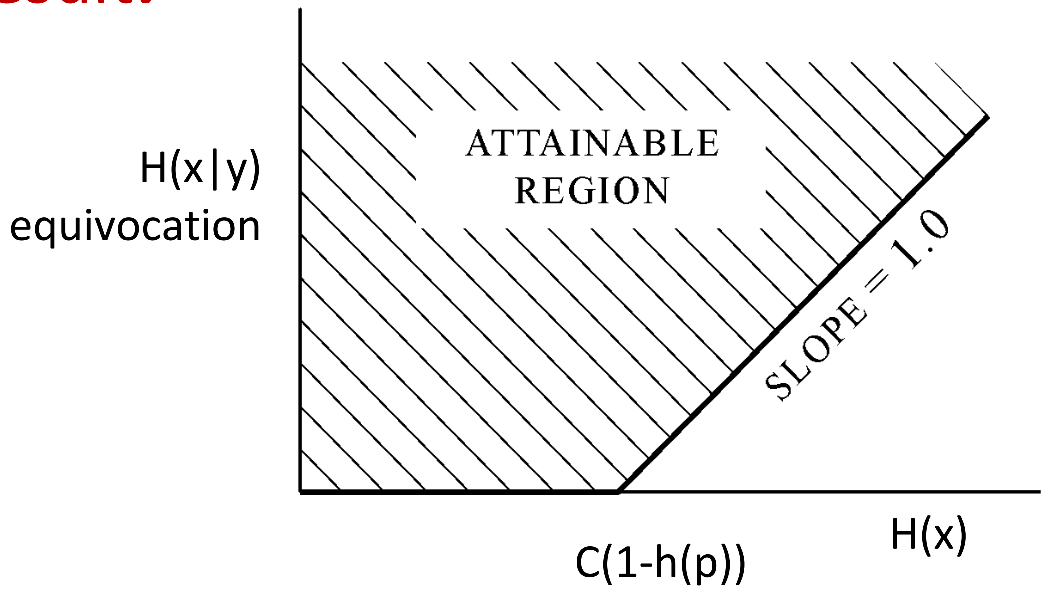
This is because these "Hamming balls" around the c(B)s with radius $p(m+r)$ are disjoint (with high probability).

# Shannon's second theorem, parts (b) and (c)

**Theorem 2:** (b) No rate greater than C (1 – h(p)) can be achieved

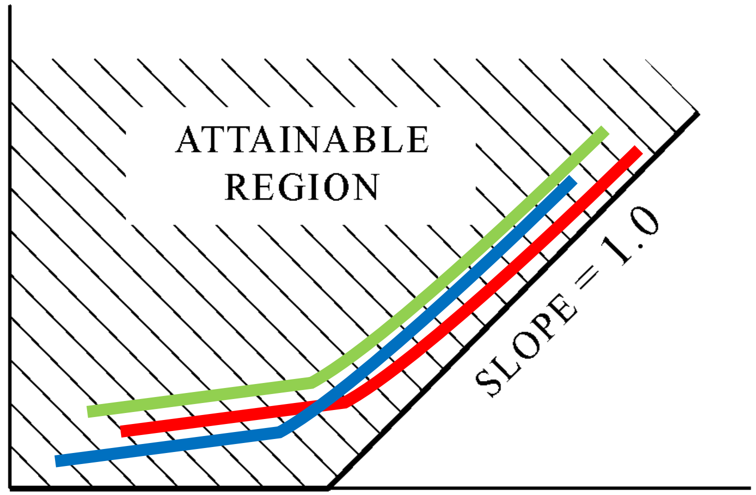
c) If equivocation is allowed, the attainable region is as shown

**Proof of (b) and (c):** If the code has redundancy less than C h(p), the balls centered at the codewords will intersect with large probability and substantial equivocation will result.



# PS: Coding Theory since Shannon

- Striving to achieve this Shannon bound with explicit codes (not randomly selected codewords)
- Reed Solomon codes, BCH codes (1950s and 1960s)
- Polar codes, turbo-codes, sparse graph codes (2000s)



# Joint, relative, and mutual entropy

- Entropy of a **joint** distribution $H(x,y)$
- Conditional entropy $H(x|y)$
- Chain rule: $H(x,y) = H(x) + H(y|x) = H(y) + H(x|y)$
- Mutual information:

  $$I(x, y) = \Sigma_{i,j} \text{Prob}[i,j] \log_2(\text{Prob}[i,j]/\text{Prob}[i]\ \text{Prob}[j])$$

- "How far from independent" are these random variables…

H(x)

H(x)

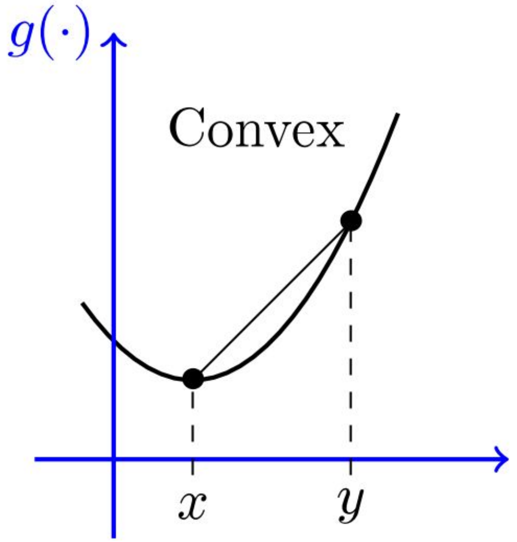
# Joint, relative, and mutual entropy (cont.)

- Mutual information:

  $$I(x, y) = \Sigma_{i,j}\, Prob[i,j] \log_2(Prob[i,j]/Prob[i]\, Prob[j])$$

- Kullback-Leibler divergence of two distributions P, Q

  $$KL(P,Q)\ (\neq KL(Q,P)) = -\Sigma_i\, P[i] \log_2 (Q[i]/P[i]) \geq 0$$




- Notice: $I(x, y) = KL\,(P(x,y), P(x)\cdot P(y))$

# A connection to Deep Learning [McAllester 2018]

- You have the distribution P(image,label) in the world
- You want to create another distribution, call it

$Q_{N(\Theta)}$(image,label) where $\Theta$ are the parameters (weight etc.) of the CNN

You want to $\max_{\Theta} E_{(\text{image,label})\sim P} \log Q_{N(\Theta)}(\text{label}|\text{image})$

Equivalently, to $\min_{\Theta} KL(P, Q_{N(\Theta)})$

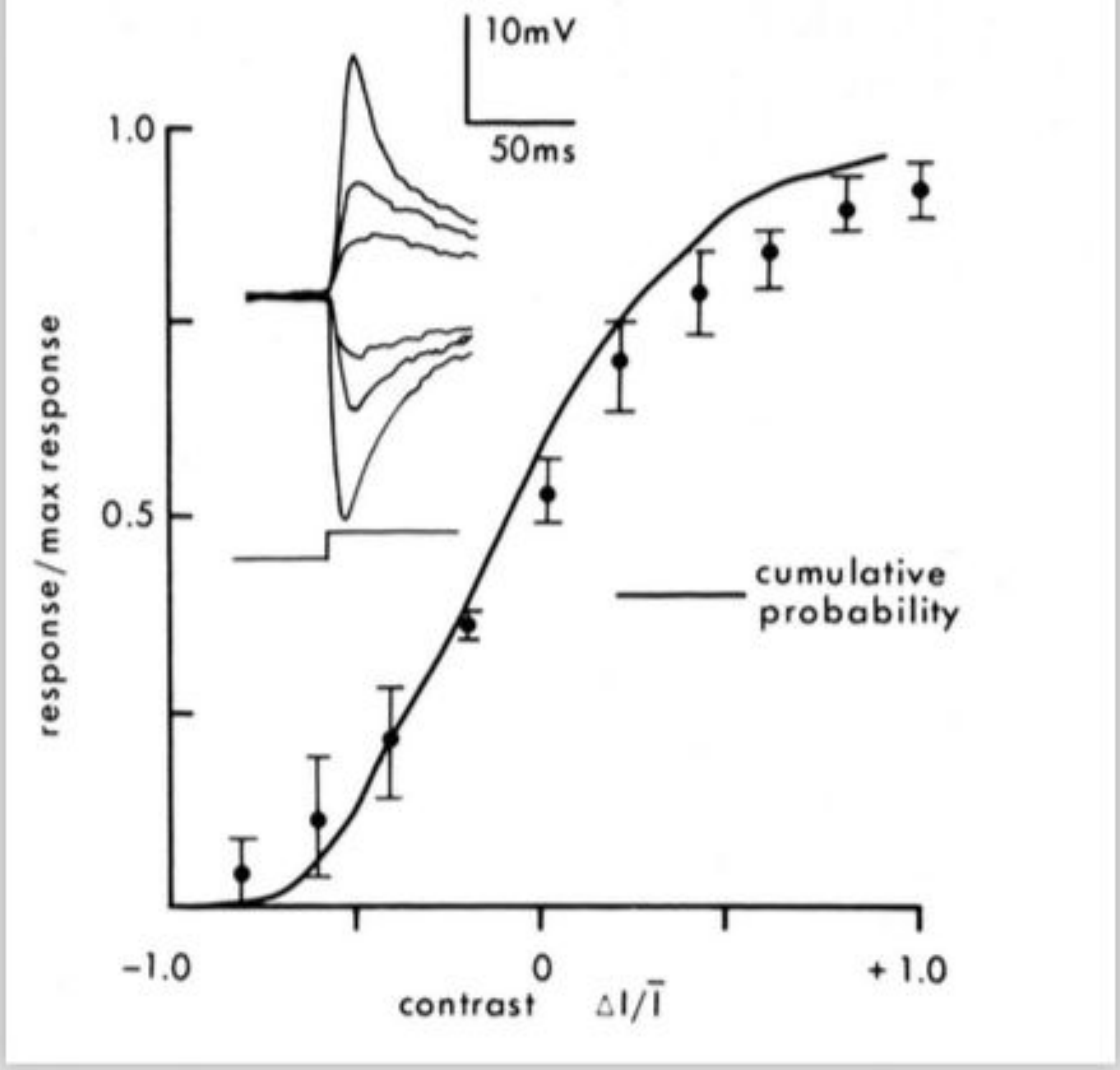
# Finally…

- Discrete information theory can be extended to continuous random variables and distributions

$$\Sigma \rightarrow \int$$

# One application: how flies encode contrast

**To maximize entropy of the encoding, the response distribution should mimic the stimulus distribution!**



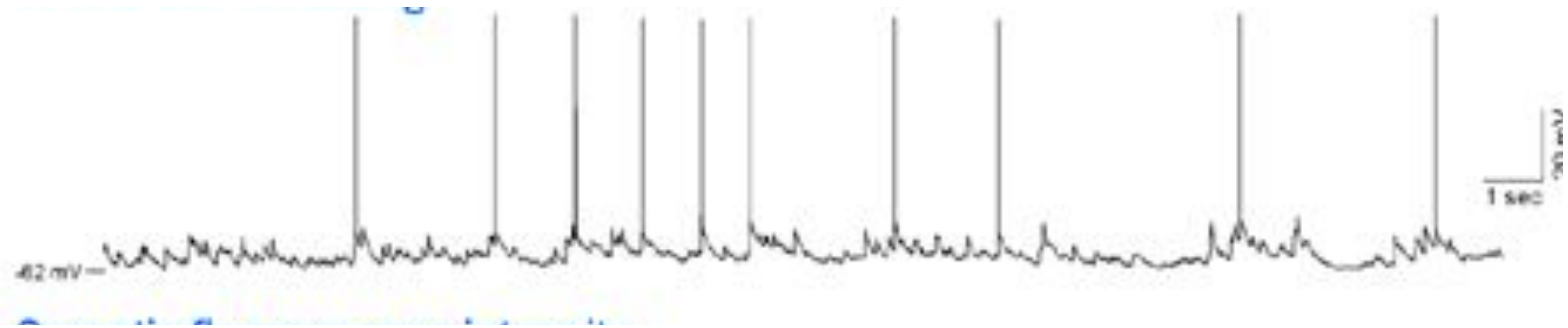
***Questions? Thoughts? Feedback?***

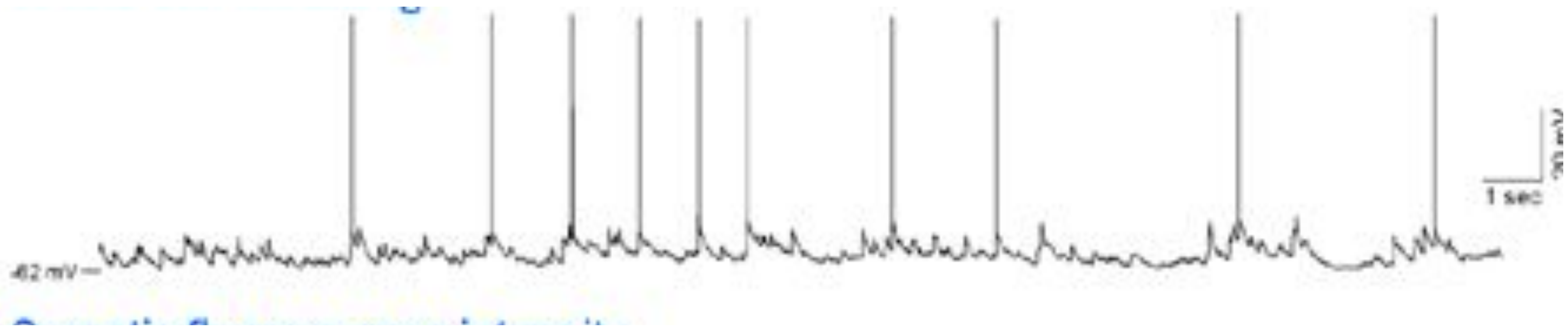
## Today:

- Continue on Information Theory and the Brain
- Introduction to Dynamical Systems
- Examples of Dynamical Systems models of the Brain

# Another example of applying Information Theory to the Brain (besides contrast coding in the fly)

- Q: What is the information contained in the spike train of a neuron responding to a stimulus?





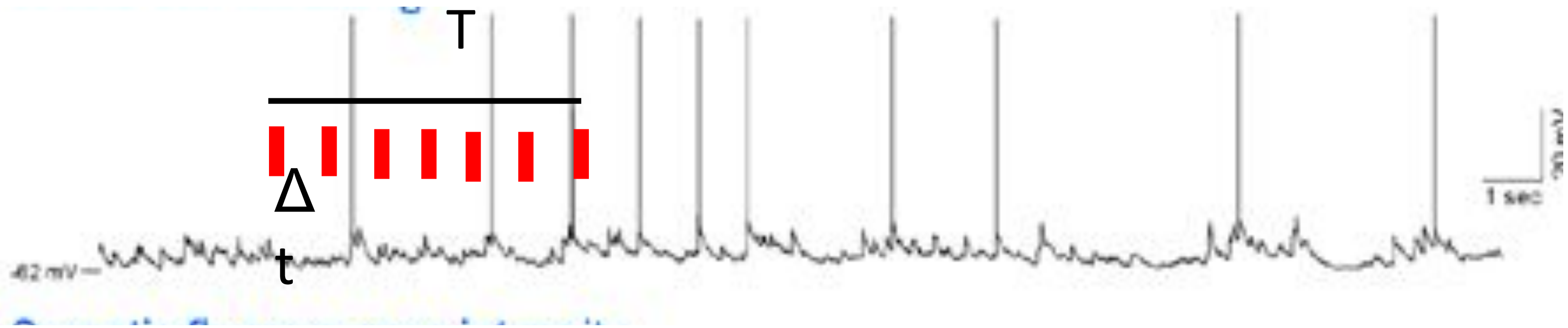

•A1: $p(t)$ = probability there is a spike in $[t, t + \Delta t]$

entropy rate → $H \lessapprox - R \int_0^\infty p(t) \log_2(p(t)\,\Delta t)\, dt$

(approximate because of possible correlations)

(choose $\Delta t$ small enough so two spikes unlikely)

For Poisson spikes: $H = R(1 - \ln(R\,\Delta t))/\ln 2$




- Better approach: Pick $\Delta t$ small, and $T = m\ \Delta t$
- $p(B)$ is the probability that $B$ in $\{0,1\}^m$ occurs in some time interval of length $T$
- Above, $m = 6$ and $B = 010011$

$$H = -\left[\Sigma_B\ p(B) \log_2 p(B)\right]/T$$

- But there is noise in the spikes

# Calculation of noise entropy

- Noise at time t: the entropy of the distribution of the different responses B(t) to the same stimulus starting at time t

$$H_{noise} = -\Delta t\,[\Sigma_B\, p(B(t)) \log_2 p(B(t))]/T^2$$

$$H_{true} = H - H_{noise}$$

- Finally, extrapolate data for $T \rightarrow \infty$

# Rieke et al. 1995 "Naturalistic stimuli increase efficiency of information transmission…"

Result: Frog auditory neurons respond with much higher $H_{true}$ to sounds that resemble frog calls than to white noise
-- despite the fact that the latter has higher entropy

# Information Theory

- Fundamental, important, useful
- Entropy, Shannon’s theorems, mutual information, KL divergence
- **Information Theory and the Brain**: extent and scope of existing results arguably somewhat below (my) expectations

NB: not to be confused with Fisher Information [Fisher and Edgeworth, 1930s], often called just "information" in Statistics

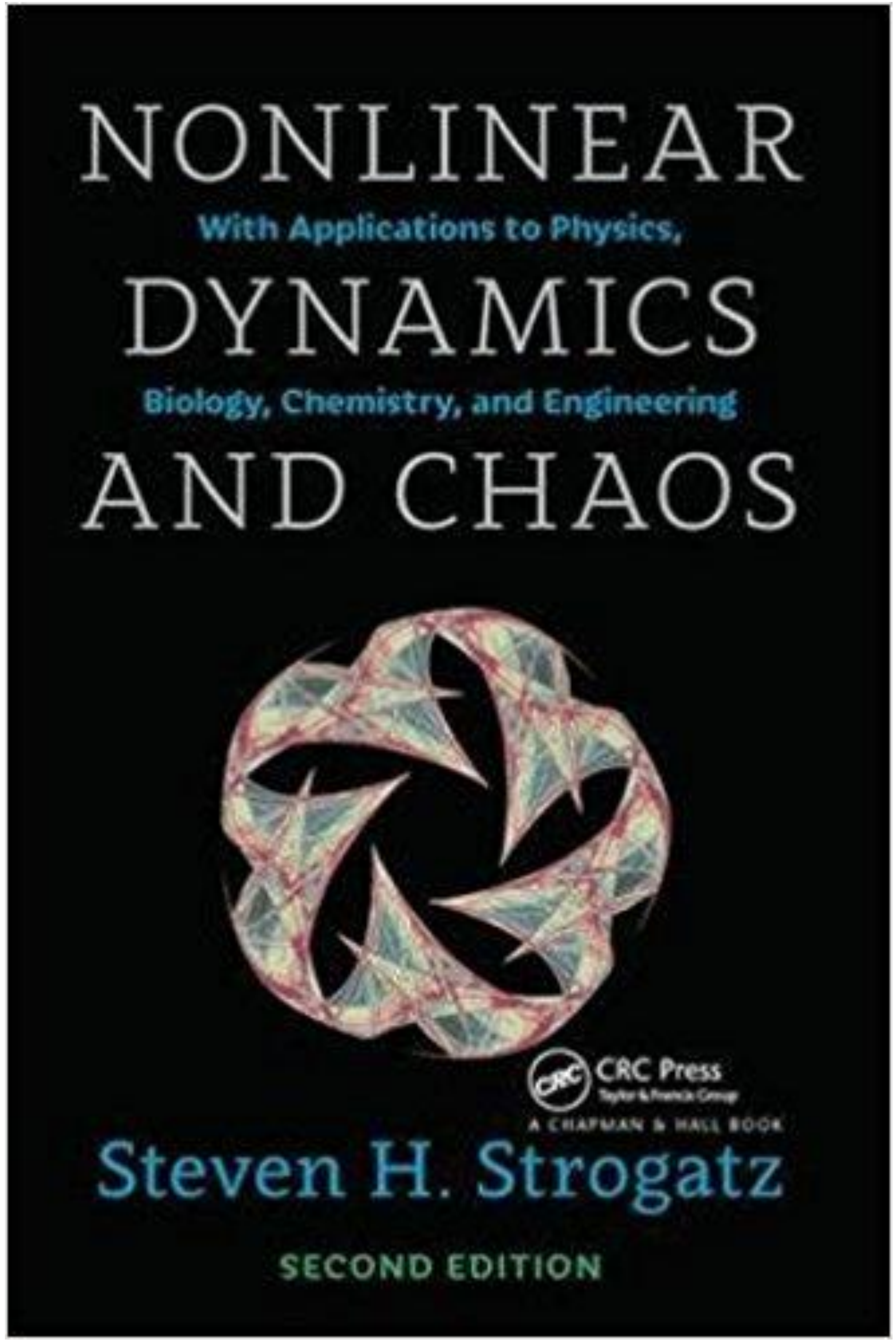
- A stimulus $\theta$ and a response x
- Q: how much information does the distribution $f(x|\theta)$ (twice diff'ble) carry about the value of $\theta$?
- Fisher information:

$$I(\theta) = E_f\left[(\partial^2/\partial\,\theta^2)\ \ln(f(x|\theta))\right]$$

- Important fact (Cramer – Rao bound): The variance of any estimate of $\theta$ is lower bounded by $\sim 1/I(\theta)$
- Occasionally used instead of entropy in the study of Brain systems

# Next: Dynamical Systems (aka Ordinary Differential Equations, ODEs)

- Find $\dot{x} = f(x)$, given the value of x at t = 0
- x(t) is an unknown function of time t, usually a vector function; $\dot{x}$ denotes dx/dt
- Linear dynamical system: $\dot{x} = A\,x$
- One dimension, solution: $x(t) = x(0)\,e^{At}$
- True for any number of dimensions…
- Linear systems are useful only as local approximations for solving nonlinear systems (helps, sometimes)




## *Excellent Book!*

# The dawn of dynamical systems: The two-body problem [Newton 1687]

- E.g., the earth and the moon (ignoring all else)

$$F(x,y) = M\ddot{x}$$
$$-F(x,y) = m\ddot{y}$$

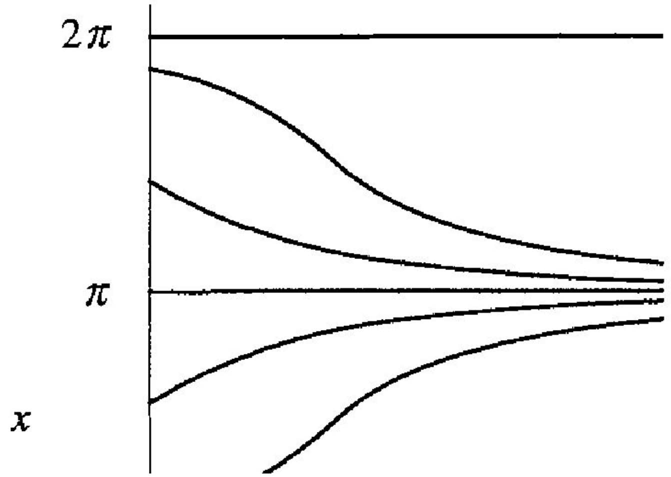
- (Second derivatives simulated by an extra equation)
- Two body problem can be solved easily
- Add: the center of mass moves with constant velocity (assume zero)
- Subtract: the vector of the two bodies moves on a plane
- Etc.

# The Three-Body Problem?
# Sun – earth –moon [Euler 1770]

- Surprise: essentially unsolvable (e.g., in closed form)
- Families of periodic solutions found, but not the full realm of solutions
- Field stuck after first success…
- Breakthrough, [Poincaré 1890s] focus on qualitative questions: “will the moon ever fly away?”
- The limit behavior of the system

# 1D systems

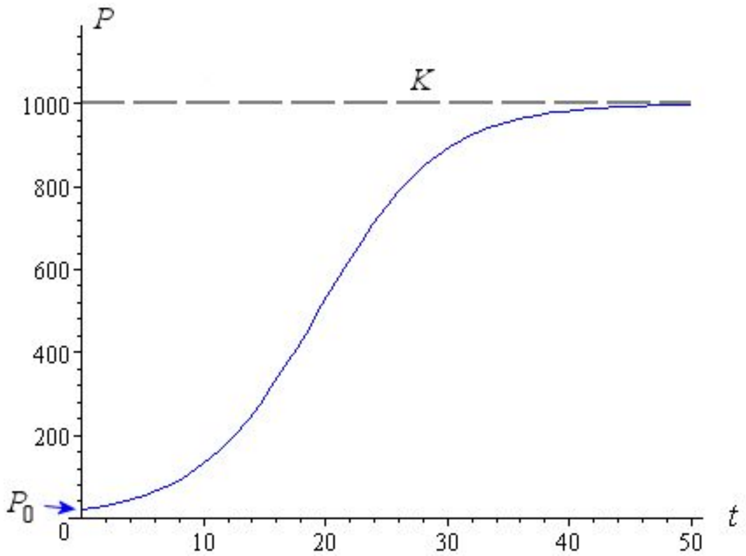
- There can be no periodic solution: only equilibria (stable/unstable) where the graph of f(x) intersects the x-axis
- Q: how does one prove convergence?
- A: potential/Lyapunov functions



# 1D systems: more examples

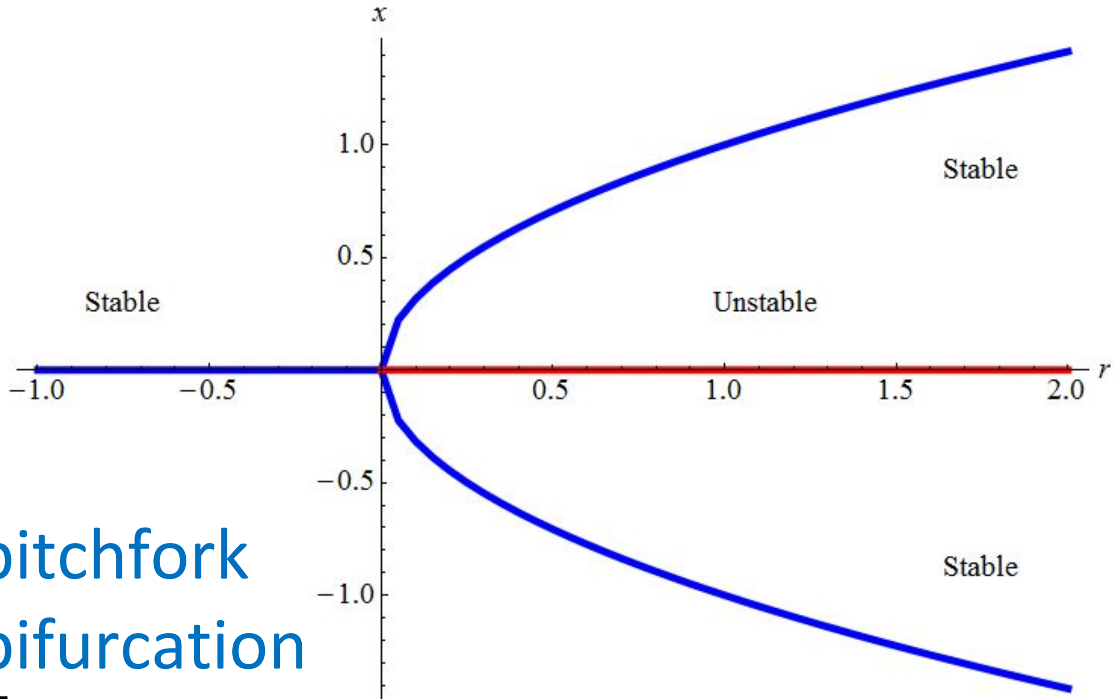
- Exponential growth. $\dot{x} = a\,x$
- The logistic equation: growth with limits

$$\dot{x} = a\,x\,(1 - x)$$



# 1D systems: bifurcation

$$\dot{x} = r\,x - x^3$$




pitchfork bifurcation

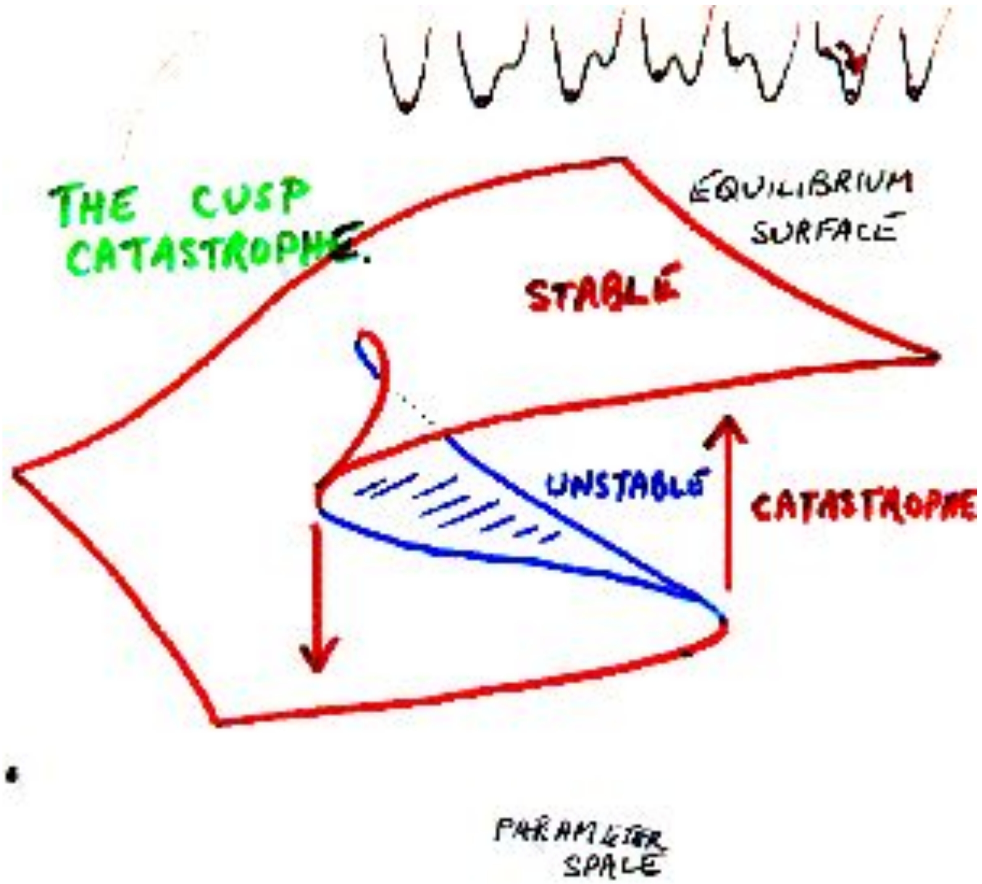
solutions no longer decay exponentially fast–

algebraic function of time (recall Exercise 2.4.9). This lethargic decay is called ***critical slowing down*** in the physics literature. Finally, when $r > 0$, the origin has become unstable. Two new stable fixed points appear on either side of the origin, symmetrically located at $x^* = \pm\sqrt{r}$.

The reason for the term "pitchfork" becomes clear when we plot the bifurcation diagram (Figure 3.4.2). Actually, pitchfork trifurcation might be a better word!

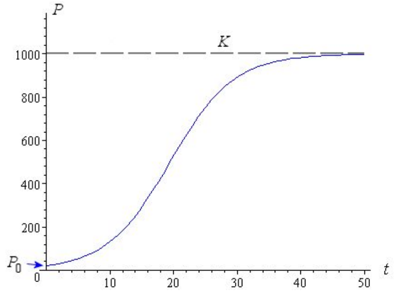
# 1D systems: imperfect (or "catastrophic") bifurcation

$$\dot{x} = h + r\,x - x^3$$



# Btw, recall the logistic equation

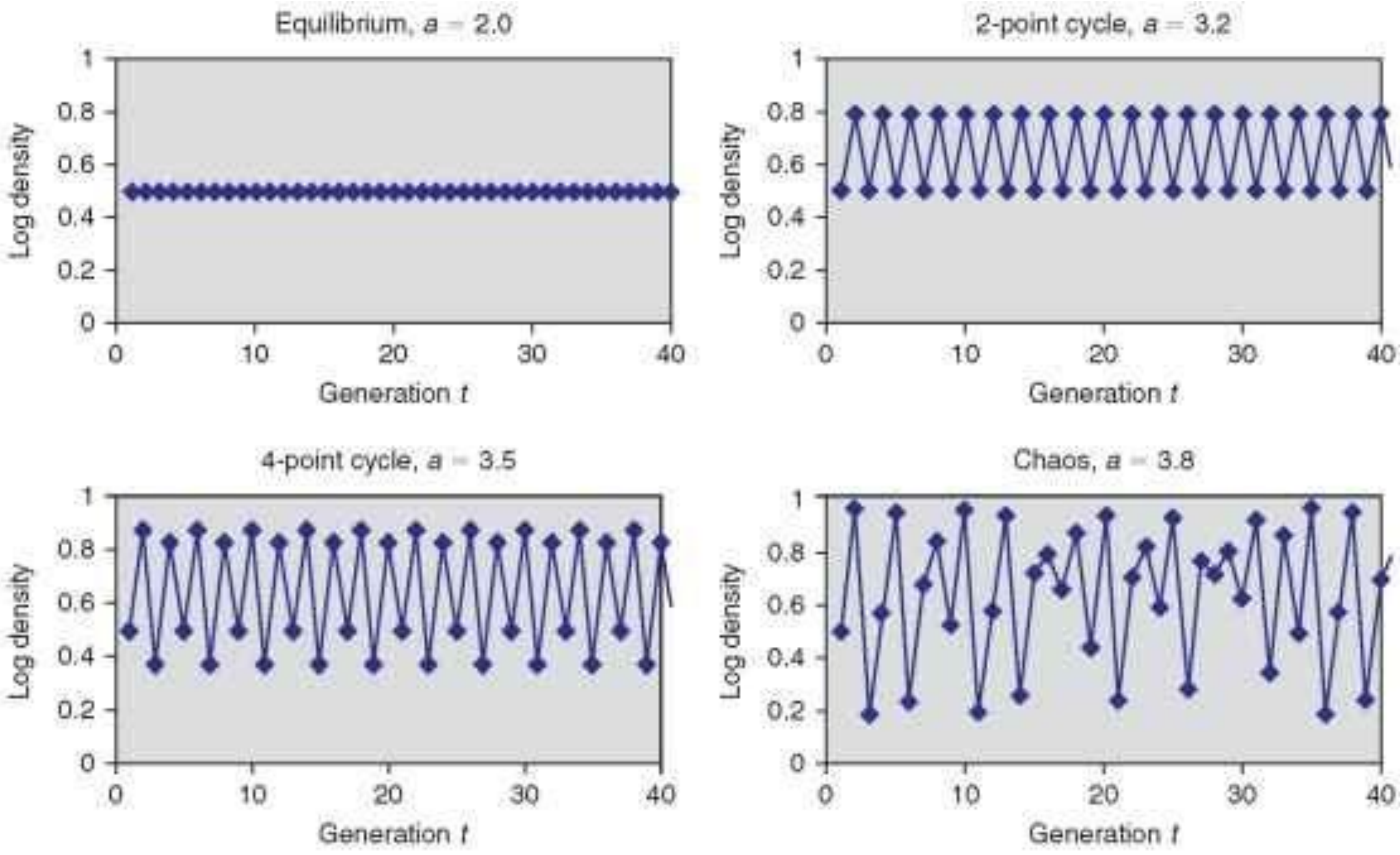
$$\dot{x} = a\,x\,(1 - x)$$




- The discrete-time logistic equation

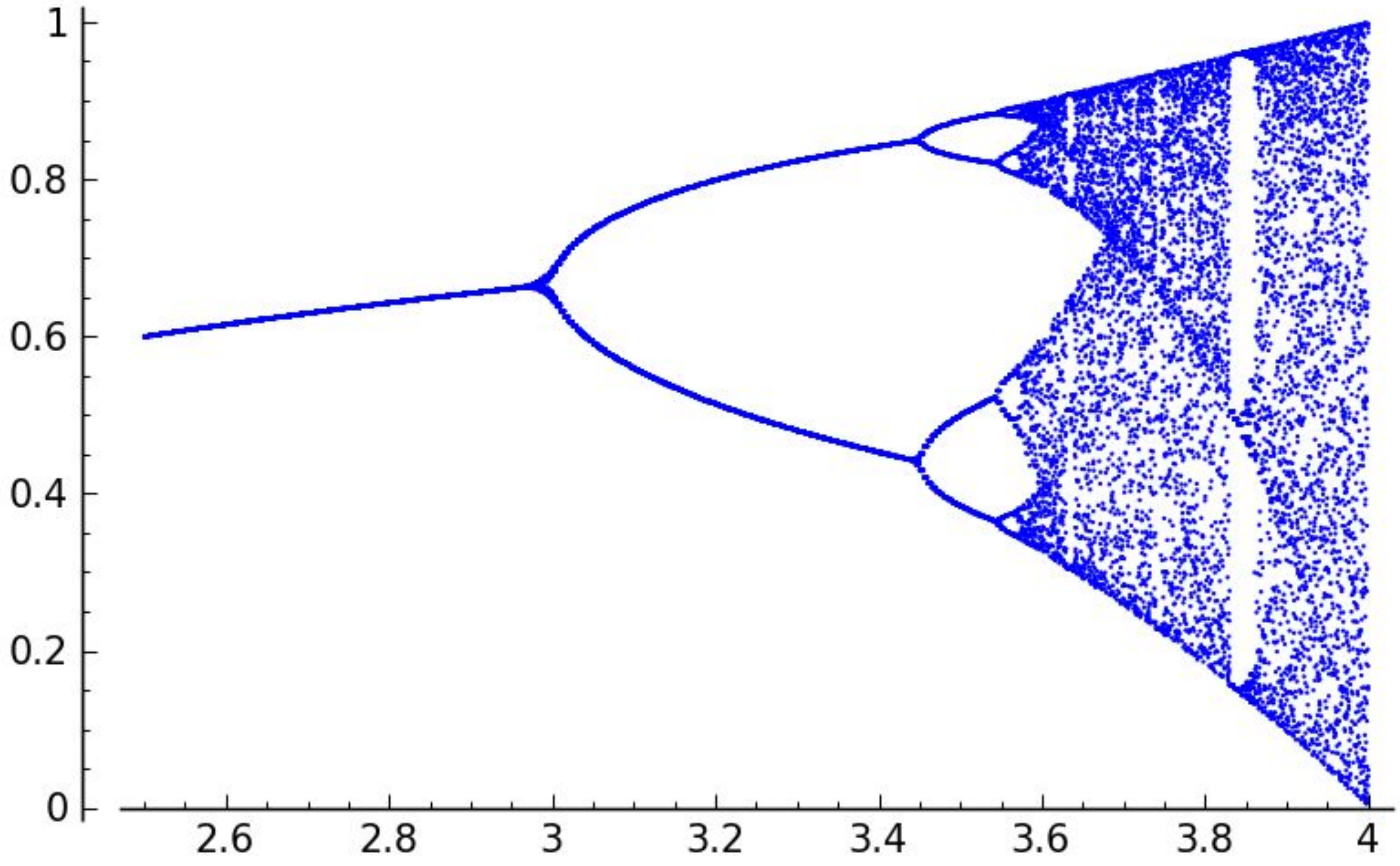
$$x^{t+1} = a\,x^{t}\,(1 - x^{t})$$

$$x^{t+1} = a\, x^t\, (1 - x^t)$$

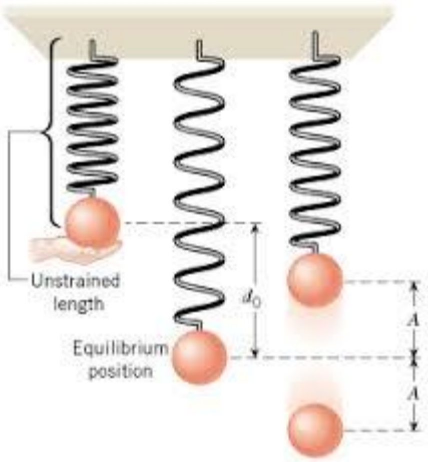


Equilibrium, a = 2.0
2-point cycle, a = 3.2
4-point cycle, a = 3.5
Chaos, a = 3.8
Log density
Generation t

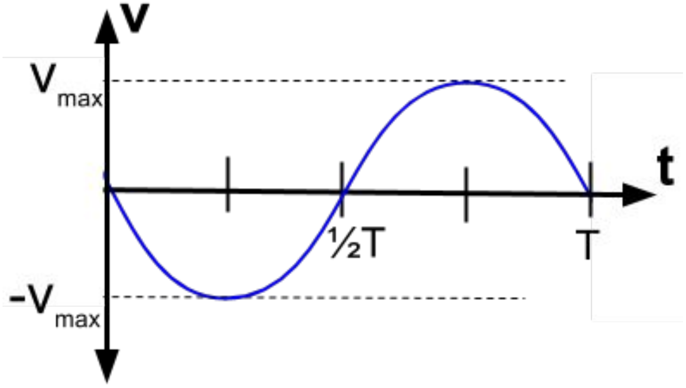
# $x^{t+1} = a\, x^t\, (1 - x^t)$: a taste of chaos…



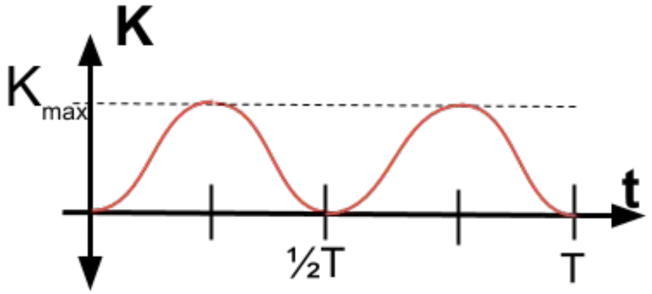
# 2D systems: periodic solutions (cycling!)




$$m\ddot{x} = - k\, x$$




- The harmonic oscillator
- Or the pendulum

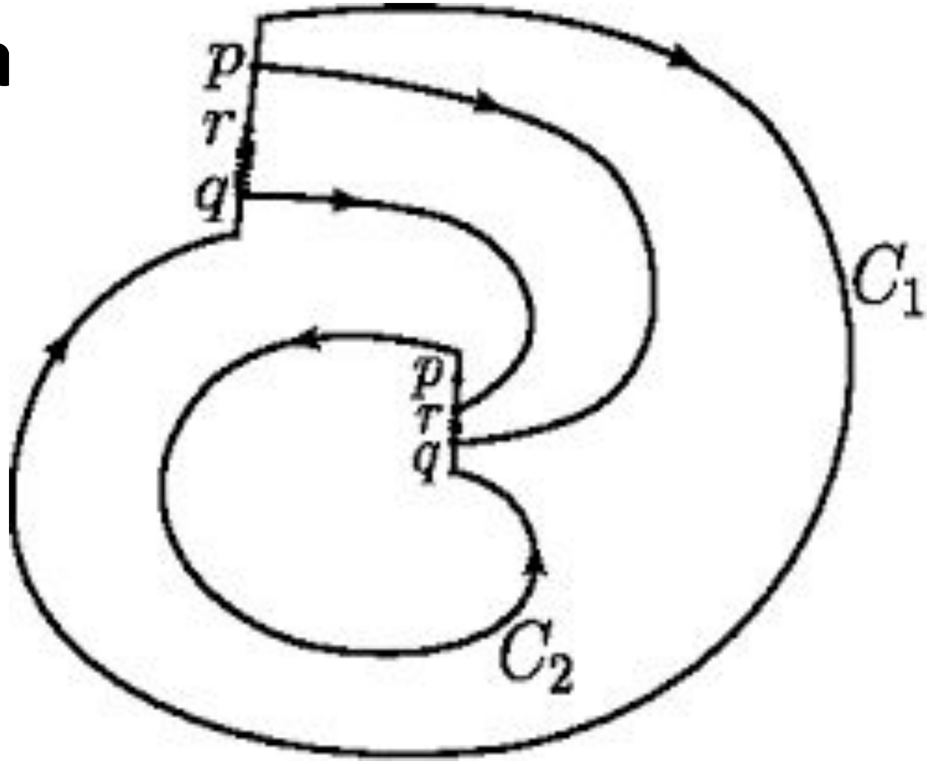


# One and two dimensions, summary

- In 1D dynamical systems, the limit behavior is equilibrium (or growth): there are no cycles
- In 2D? **Poincaré – Bendixson theorem**

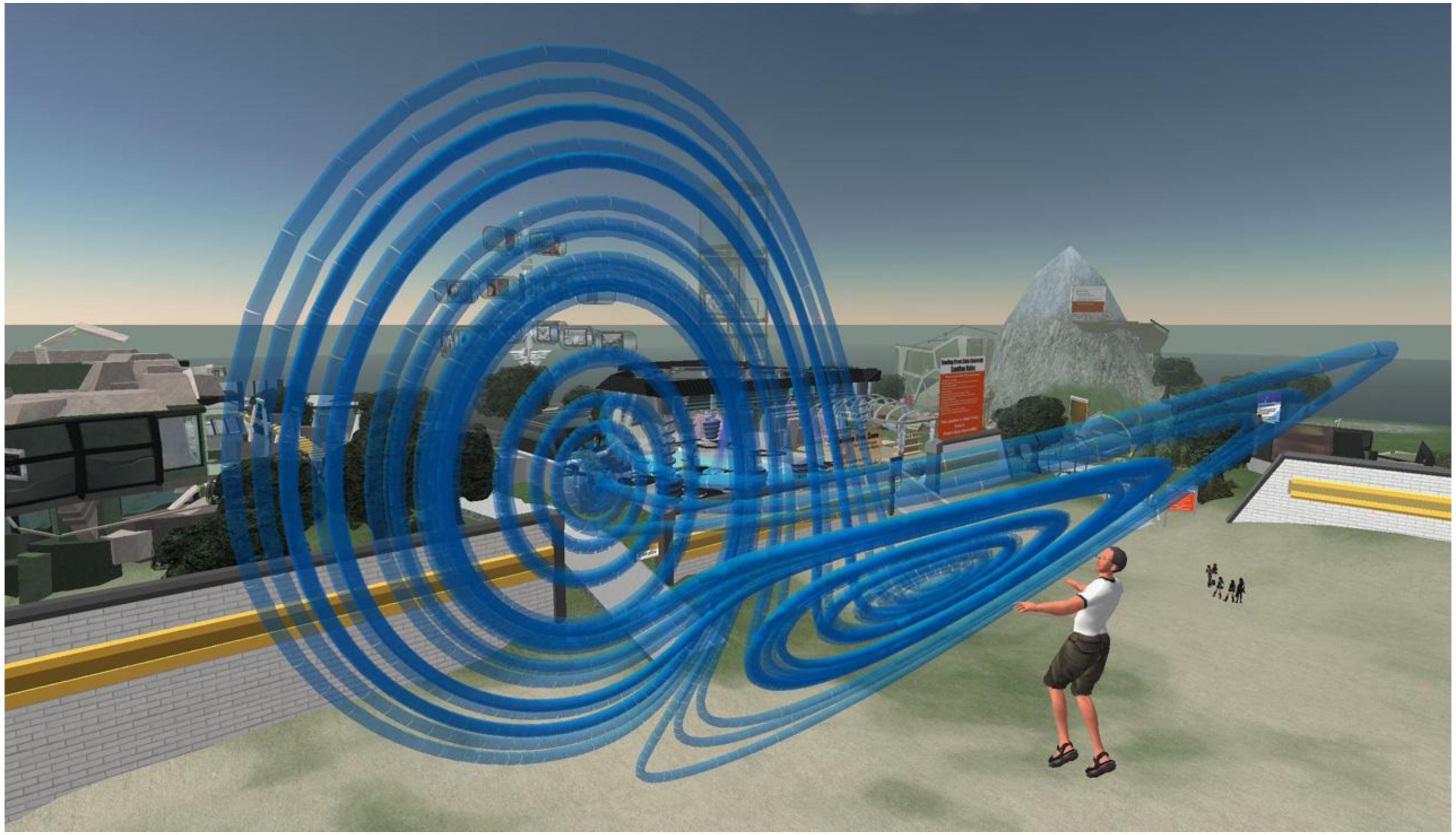
In 2D the limit behavior is either stationary (equilibrium) or periodic (cycles)*

- There can be no chaos here, the flow "restrains itself"

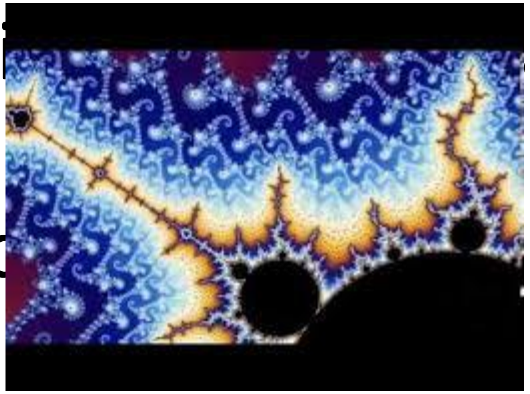


# Three dimensional dynamical systems
# Lorenz oscillator, 1963: CHAOS

$$\dot{x} = a(y - x)$$
$$\dot{y} = x(b - z) - y$$
$$\dot{z} = xy - cz$$



# What is Chaos?

- Exponentially small perturbations  in parameters and initial conditions lead to qualitatively different behaviors
- A seemingly periodic behavior repeats forever, except that the system never exactly cycles (Lorenz)
- An attractor is strange (fractal-like)
- In discrete time: there cycles of all ki  (but a cycle of period three is enough….)
- The system cannot be solved (or unc any satisfactory way

# Against chaos: Properties you want your dynamical system to have

- Conservative systems: they conserve energy, other quantities of interest
- Reversible systems: they can be "run backwards"
- Systems that have a Lyapunov function (progress toward convergence)

# The fundamental theorem of dynamical systems: "Poincare'-Bendixson envy"

- D > 2: is there a notion of a cycle so that the P-B theorem is restored (despite chaos)?
- 1900 – 1980: topologists looked for it
- Discrete time, say (continuous time follows)
- Suppose for all ε > 0 there is a N such that from x I can come back to x with a sequence of < N steps alternating with jumps of length < ε
- Call such a point x chain-recurrent

# The fundamental theorem of dynamical systems (cont.)

- **Theorem** [Conley 1984]: The domain of any dynamical system can be decomposed in the chain recurrent components (CRC) and the transient parts.  There is a Lyapunov function that drives any transient point towards the CRCs

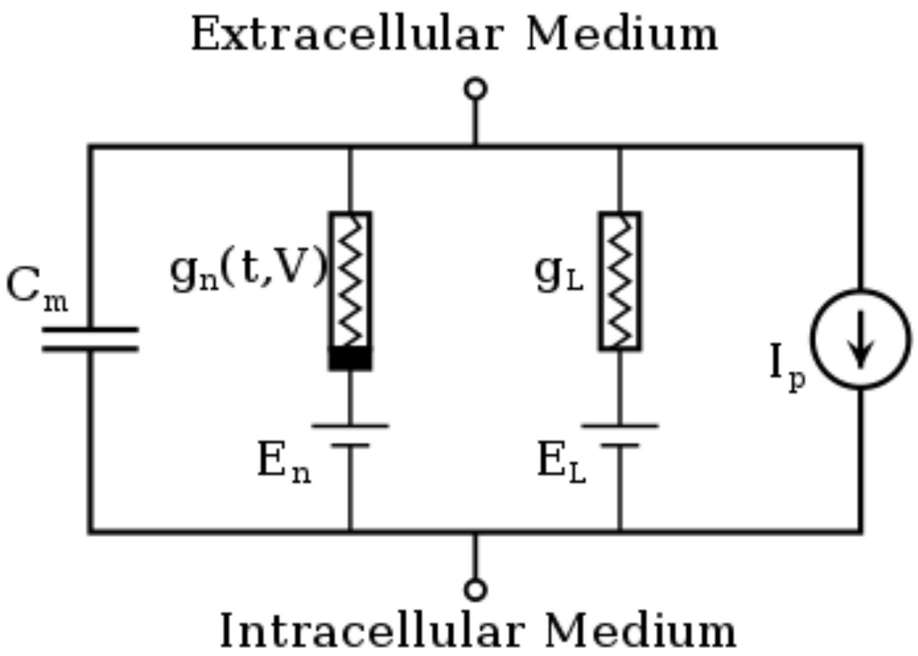
- In other words “if you squint a little, chaos goes away”

# OK, that was our quick introduction to dynamical systems

- Next: dynamical systems for modeling parts of the Brain
- Continuous and discrete
- A few representative examples
- Avoiding chaos

# We have seen one

- The Hodgkin-Huxley oscillator

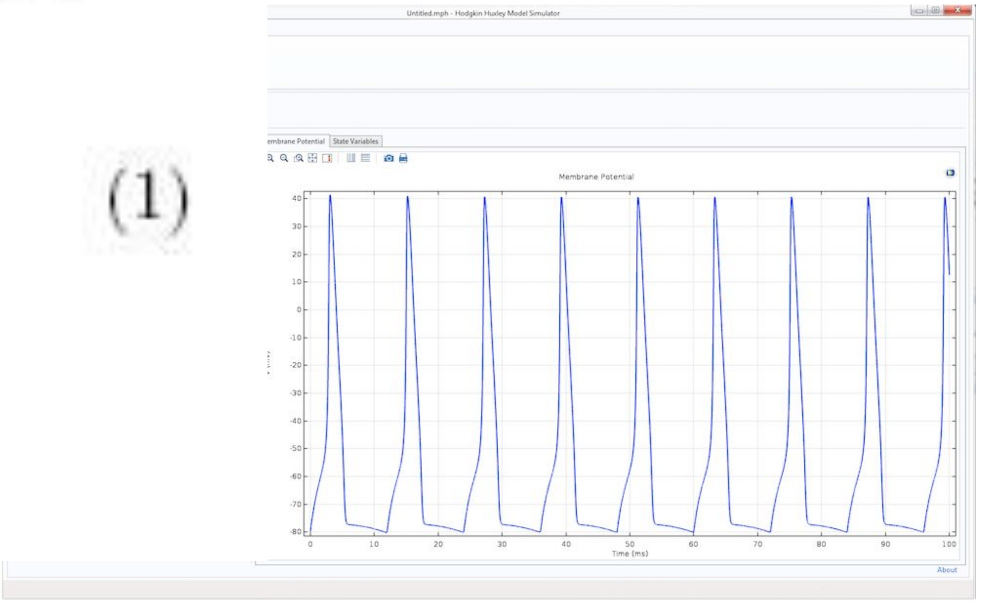



$$\frac{dv}{dt} = \frac{1}{C_m}[I - g_{Na}m^3h(v - E_{Na}) - g_k n^4(v - E_K) \\ -g_L(v - E_L)]$$

$$\frac{dm}{dt} = \alpha_m(v)(1-m) - \beta_m(v)m \qquad (1)$$

$$\frac{dm}{dt} = \alpha_m(v)(1-m) - \beta_m(v)m$$

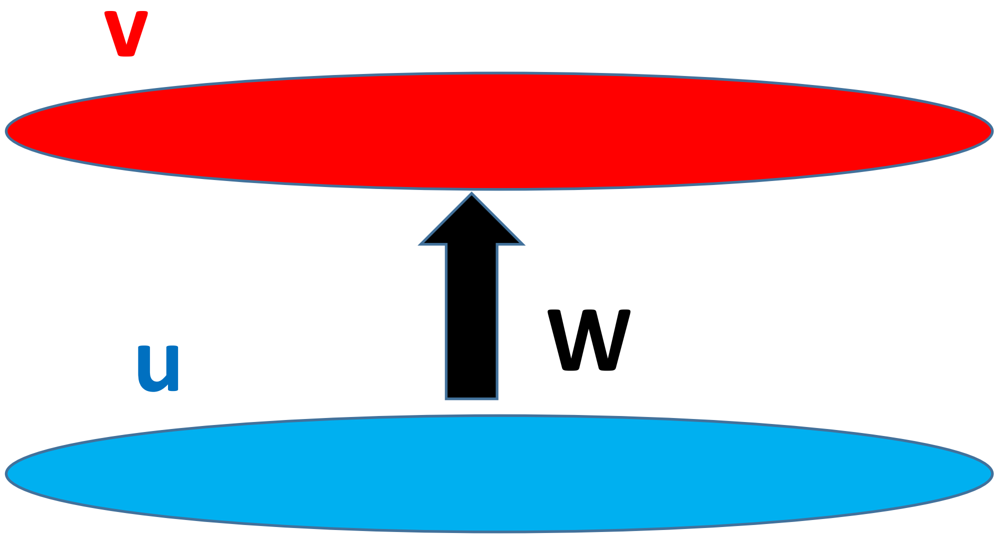
$$\frac{dh}{dt} = \alpha_h(v)(1-h) - \beta_h(v)h$$



# Feedforward network

- Two populations of neurons
- Feedforward synaptic connections
- **u**, **v**: vectors of spiking rates
- W: matrix of synaptic weights

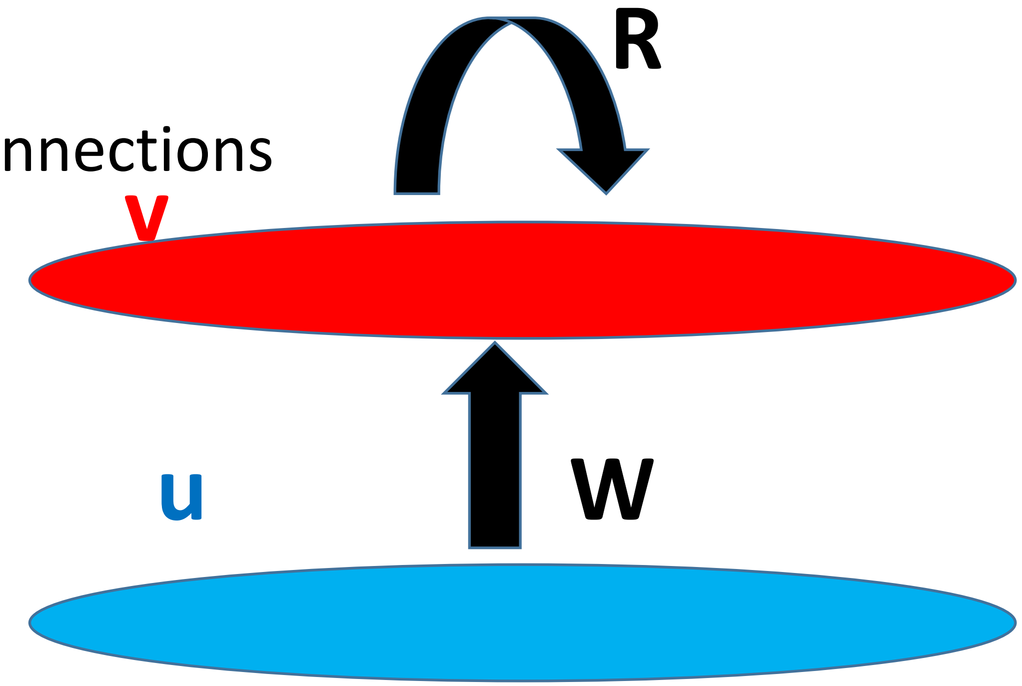
$$\tau \cdot dv/dt = -v + F(W \cdot u)$$



# Feedforward and Recurrent network

- Two populations of neurons
- Feedforward and recurrent synaptic connections
- **u**, **v**: vectors of firing rates
- **W**: matrix of synaptic weights

$$\tau \cdot dv/dt = -v + F(W \cdot u + R \cdot v)$$

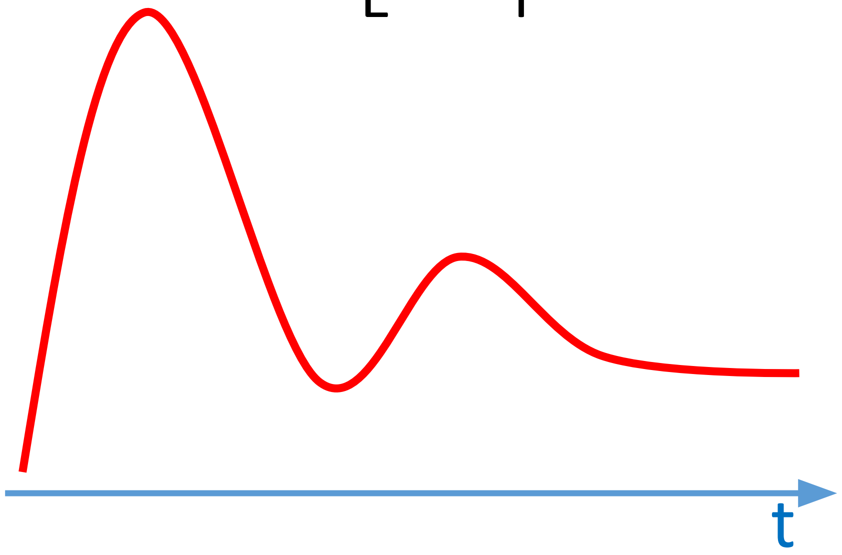


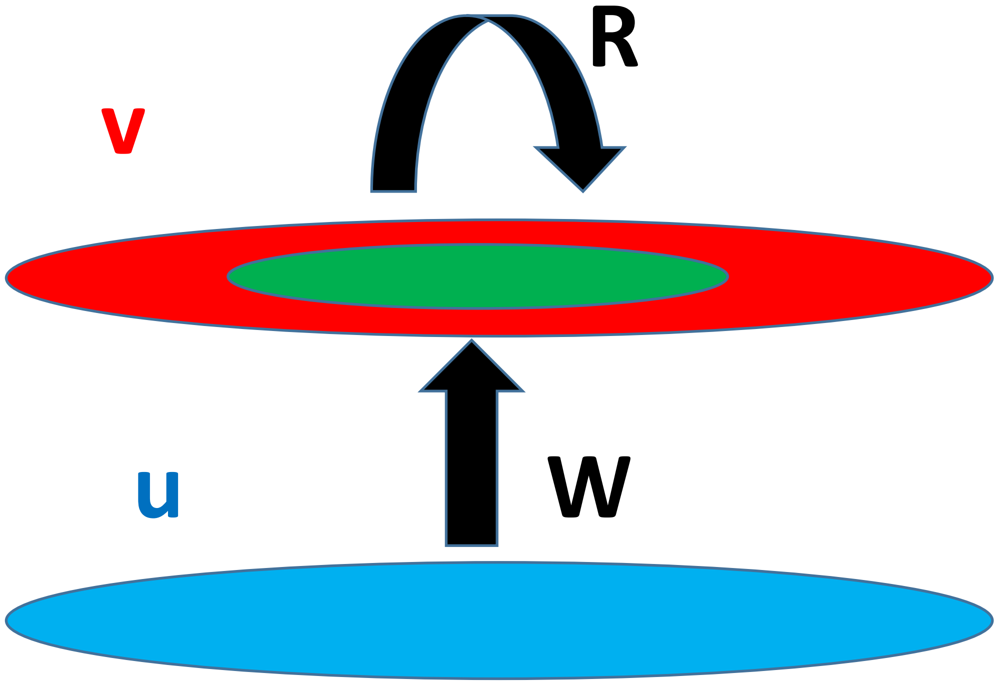

- Interesting case: Inhibitory and excitatory neurons in **RED**

Interesting case: **Inhibitory** and **excitatory** neurons (some ***negative columns*** in R, per Dale's Law)

- $\tau \cdot dv/dt = -v + F(W \cdot u + R \cdot v)$

$T_E > T_I$



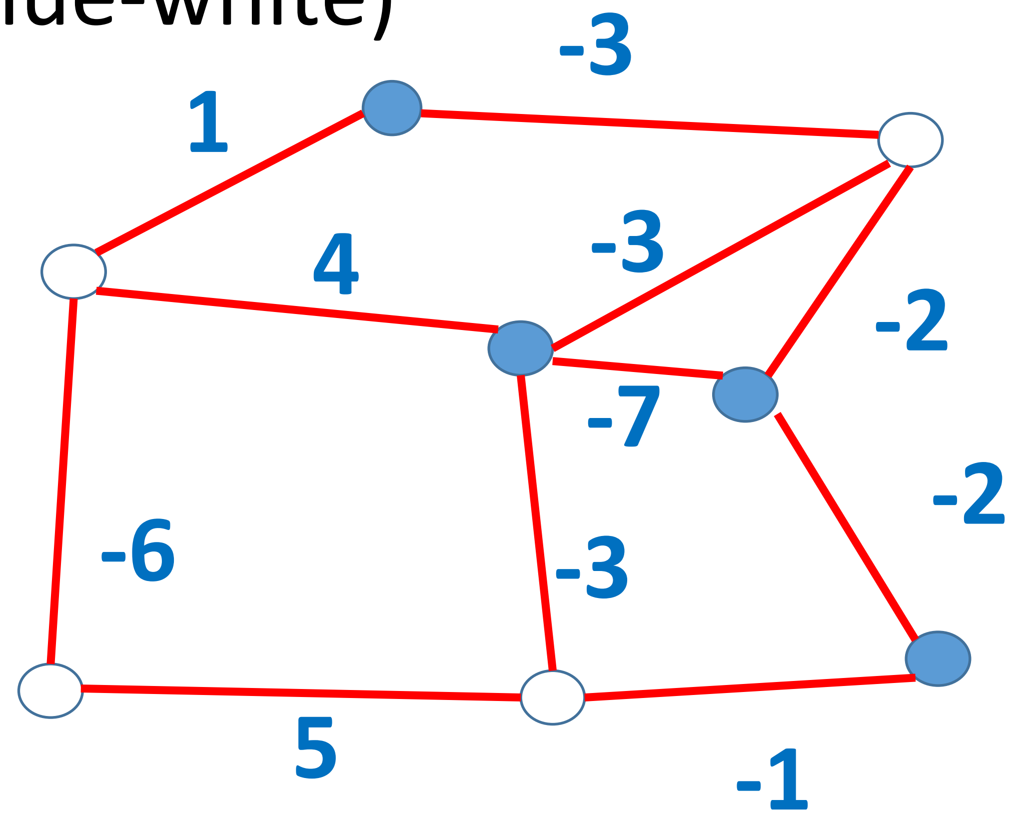
# A discrete-time system: Hopfield net

Nodes have two values: +1, -1 (blue-white)

Node i is happy if $\Sigma_j v_i v_j w_{ij} \geq 0$ $\geq \Theta_i$

Algorithm/dynamical system:

**while there is**
**an unhappy node**
**flip it**



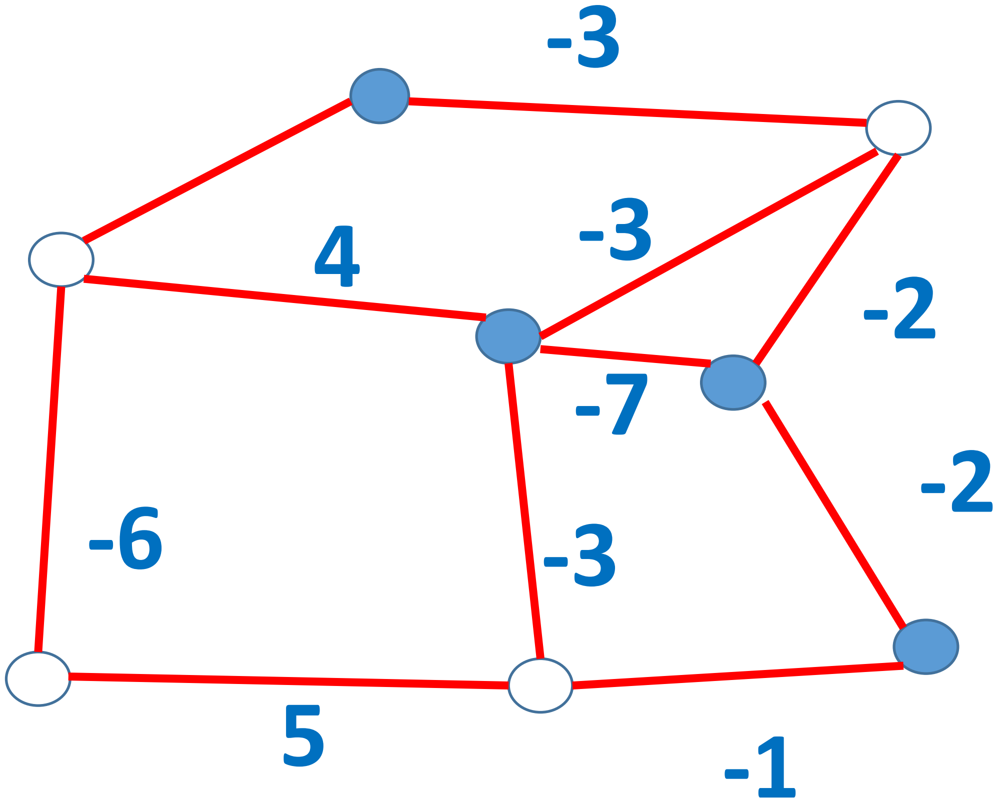
# A discrete-time system: Hopfield net

Theorem [Hopfield 1982]: Dynamical system converges

Proof: **Lyapunov** function

$$\Sigma_{i,j} v_i v_j w_{ij}$$

always increases



# Pattern completion!