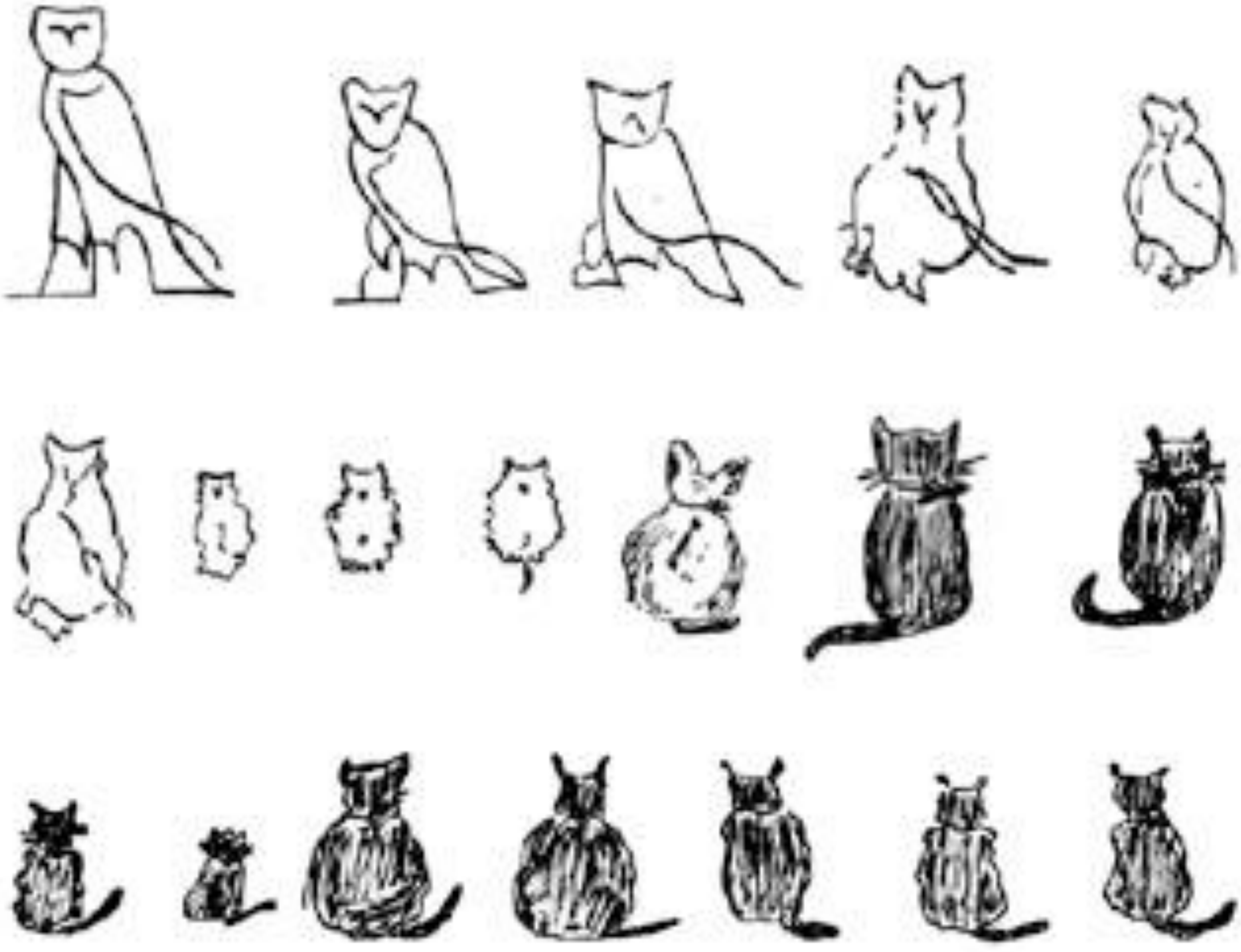
Geometry, Dynamics, and Natural Algorithms

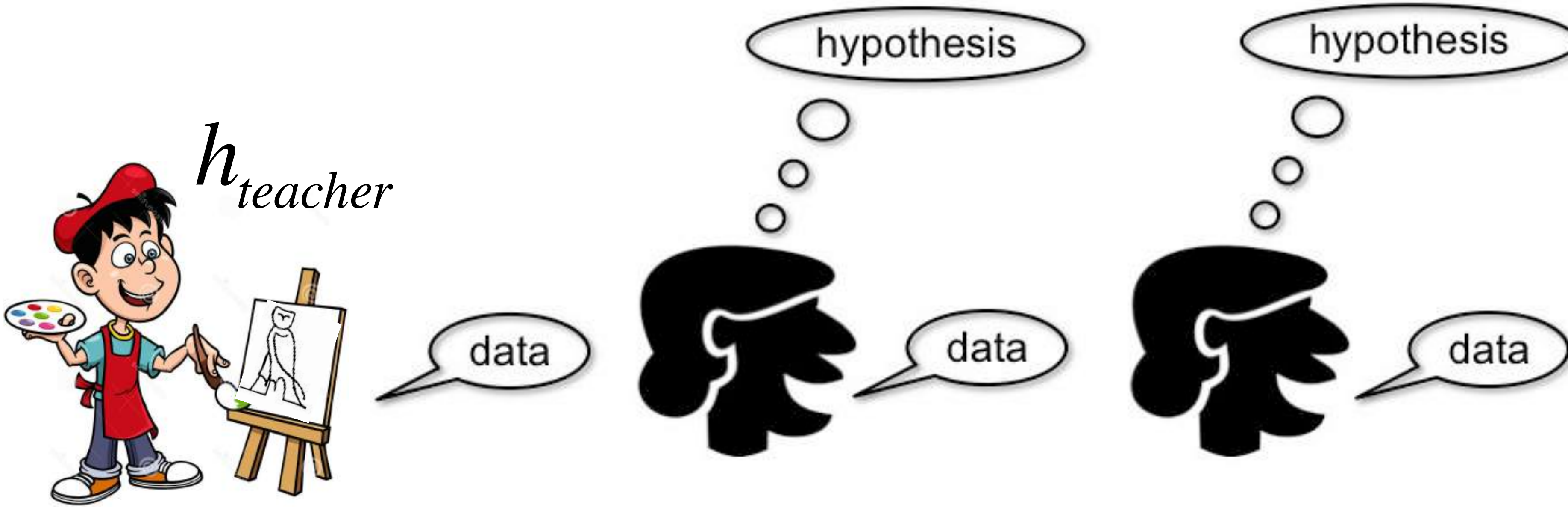
Bernard Chazelle

Princeton University

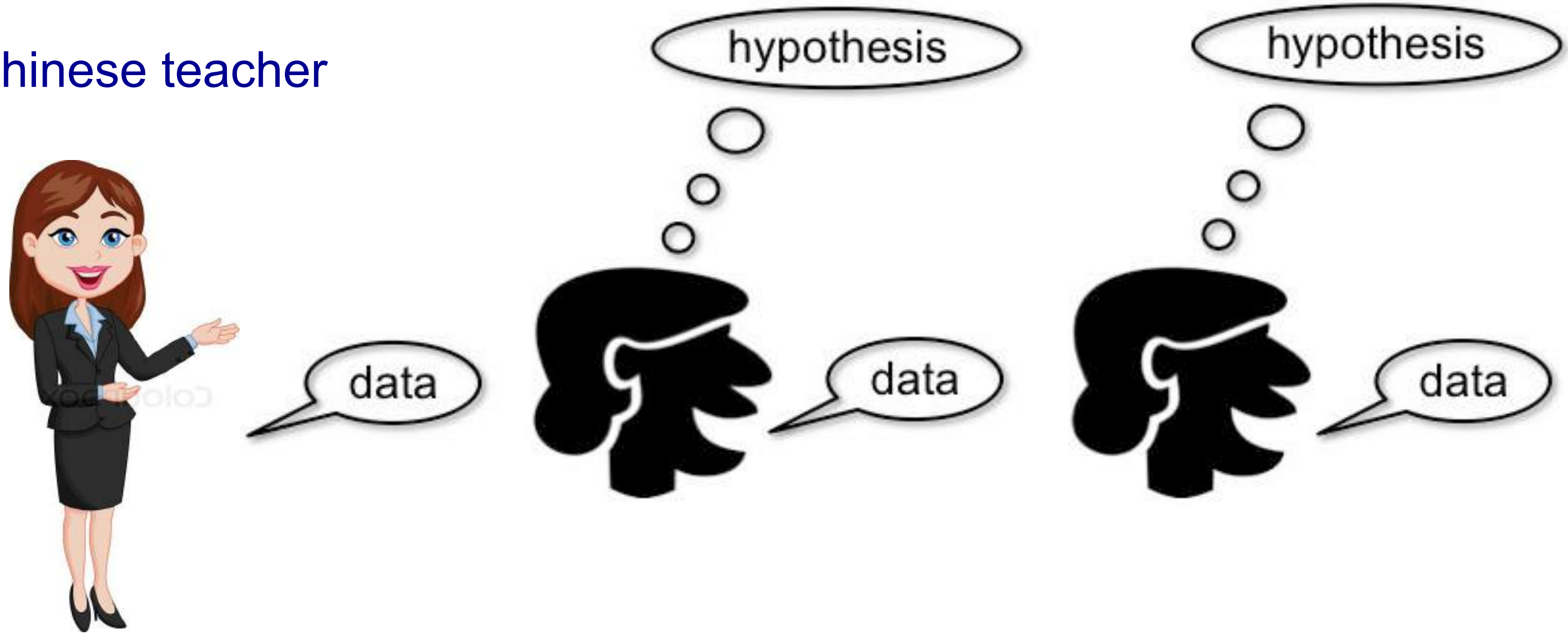
Frederic C. Bartlett 1932



Iterated learning



Chinese teacher



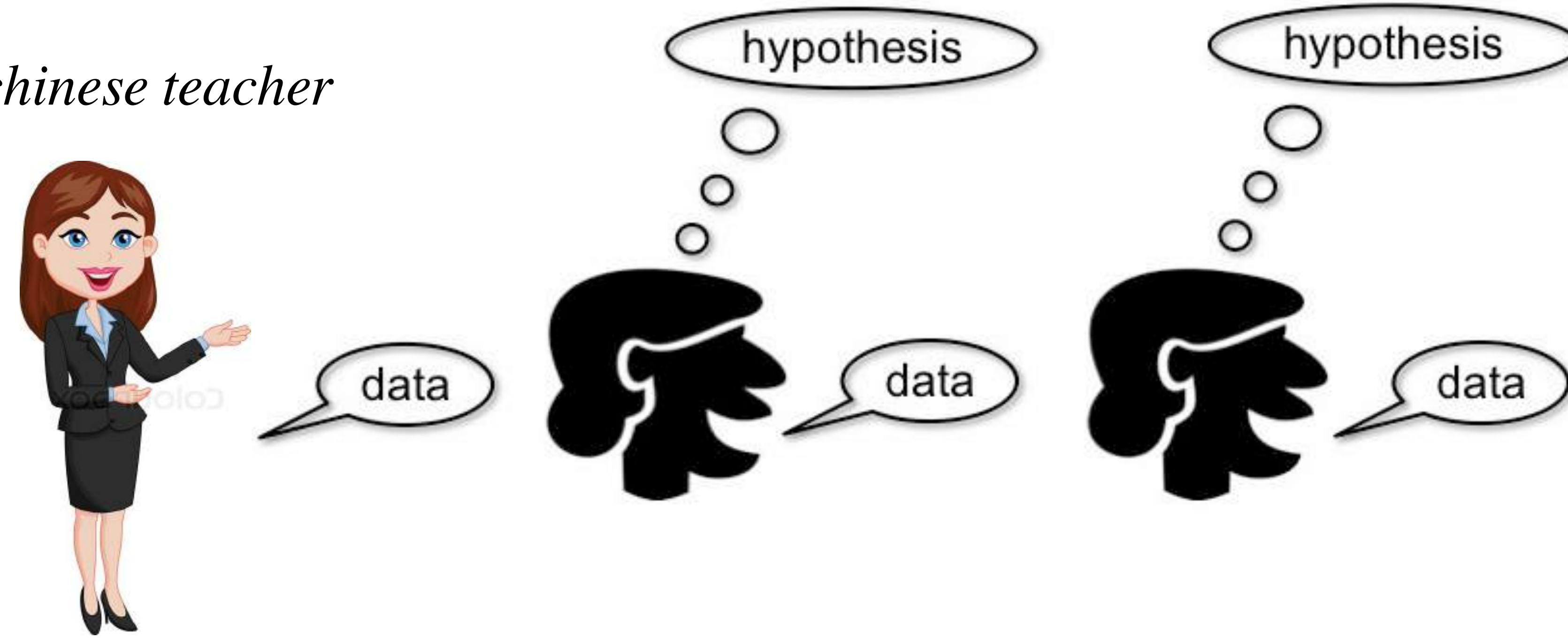
Kalish & Griffiths (2005)

prior $\mathbb{P}[h]$

likelihood $\mathbb{P}[\text{data} \mid h]$

$\mathbb{P}[h \mid \text{data}]$ $\mathbb{P}[h \mid \text{data}]$

$h_{\textit{chinese teacher}}$



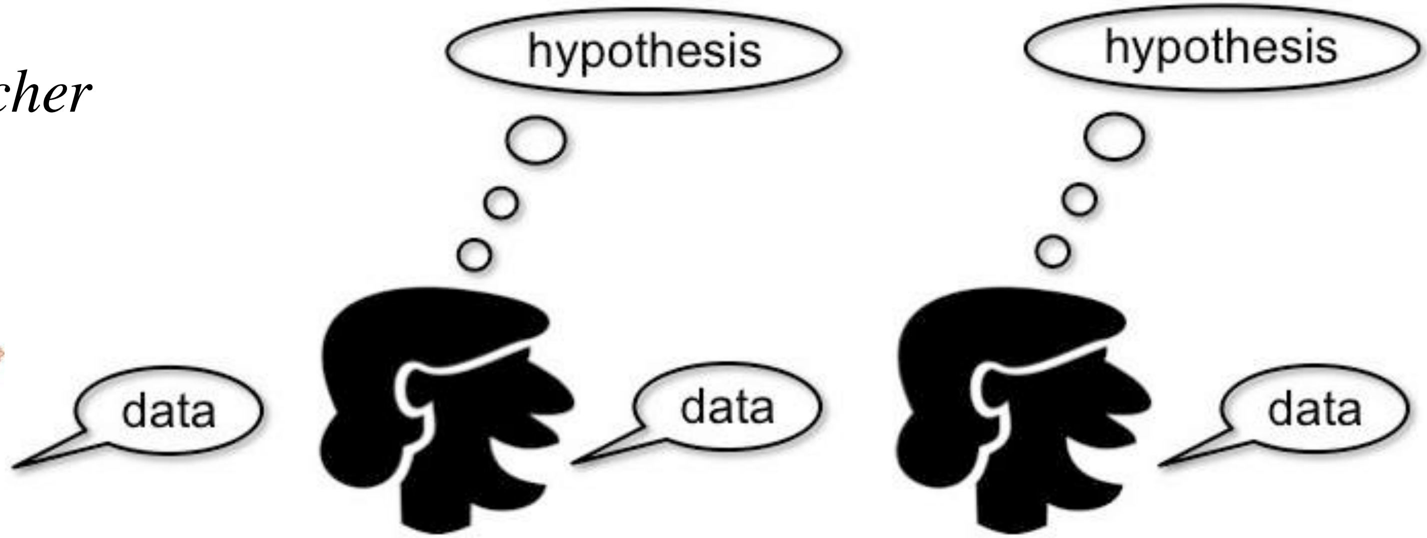
Kalish & Griffiths (2005) Gibbs sampling: mixes to prior

prior $\mathbb{P}[h]$

likelihood $\mathbb{P}[\text{data} \mid h]$

$\mathbb{P}[h \mid \text{data}]$ $\mathbb{P}[h \mid \text{data}]$

$h_{\textit{chinese teacher}}$



m_1 bits m_2 bits m_3 bits

with C. Wang (2016) Lengthen learning sessions

$$m_1 < m_2 < m_3 < \cdots$$

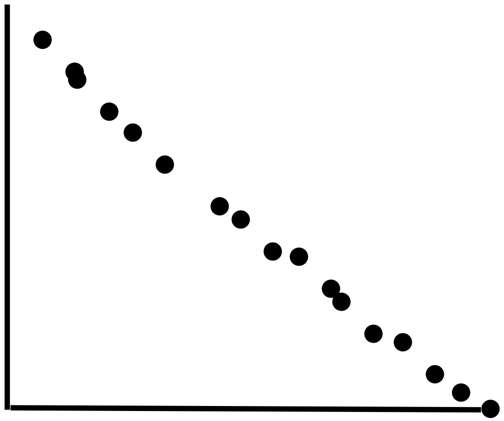
$(\varepsilon, \delta) - \text{sustainability}$

For *any* learner, with prob > $1-\varepsilon$,

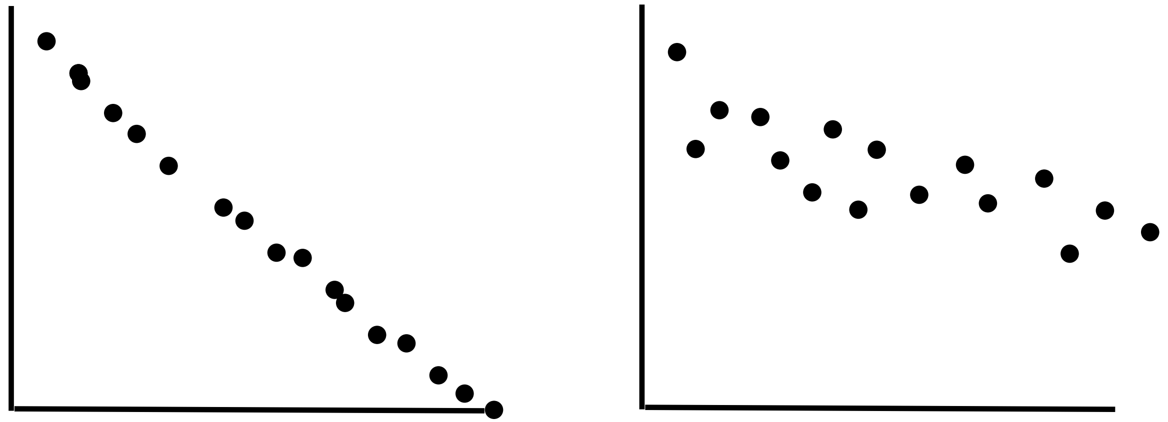
total variation between $h_{teacher}$ and random h

from posterior is at most δ.

$$m_t = \frac{1}{\delta^2} \ln \frac{t}{\varepsilon}$$

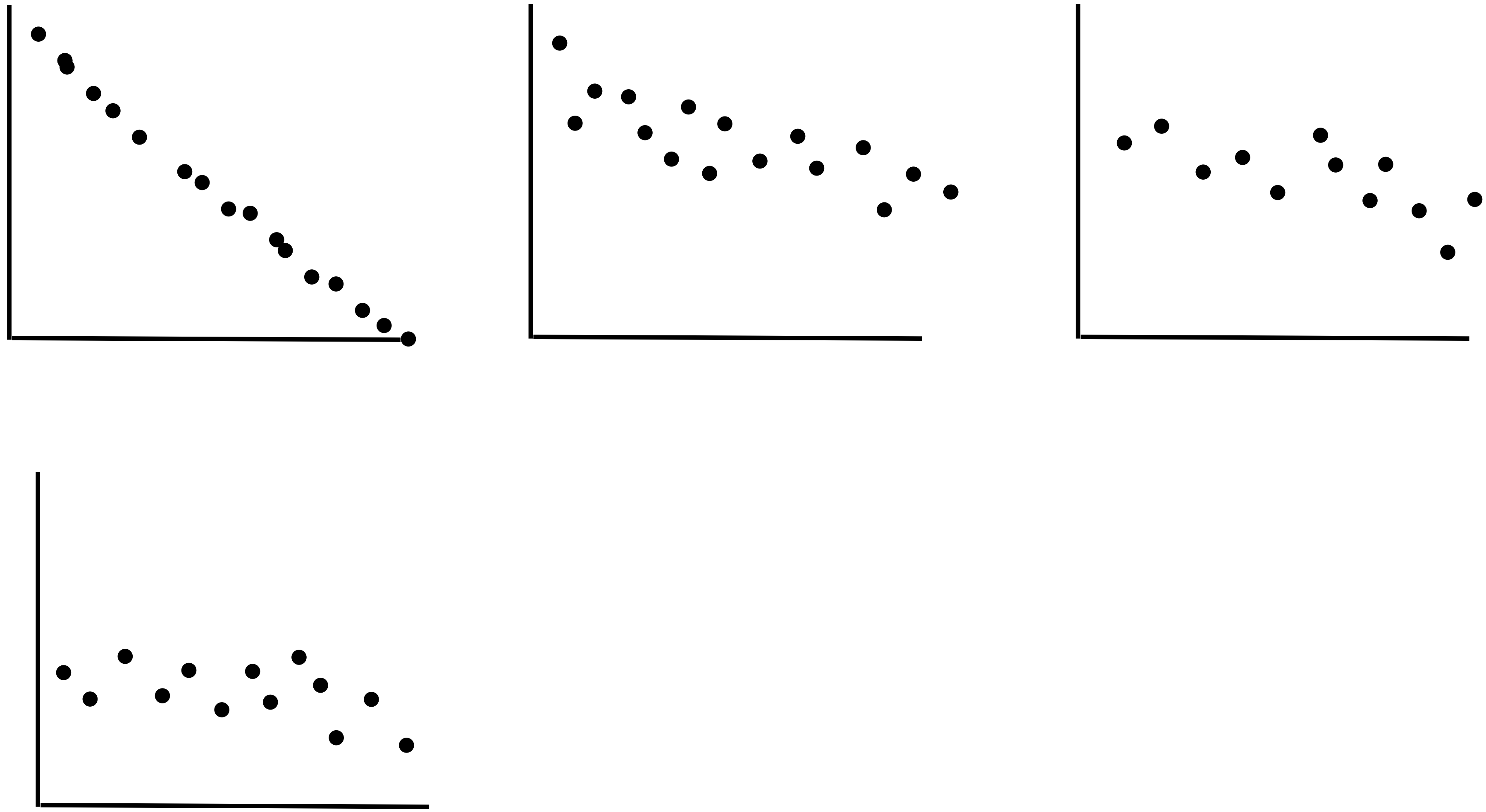


Kalish, Griffiths, Lewandowsky (2007)

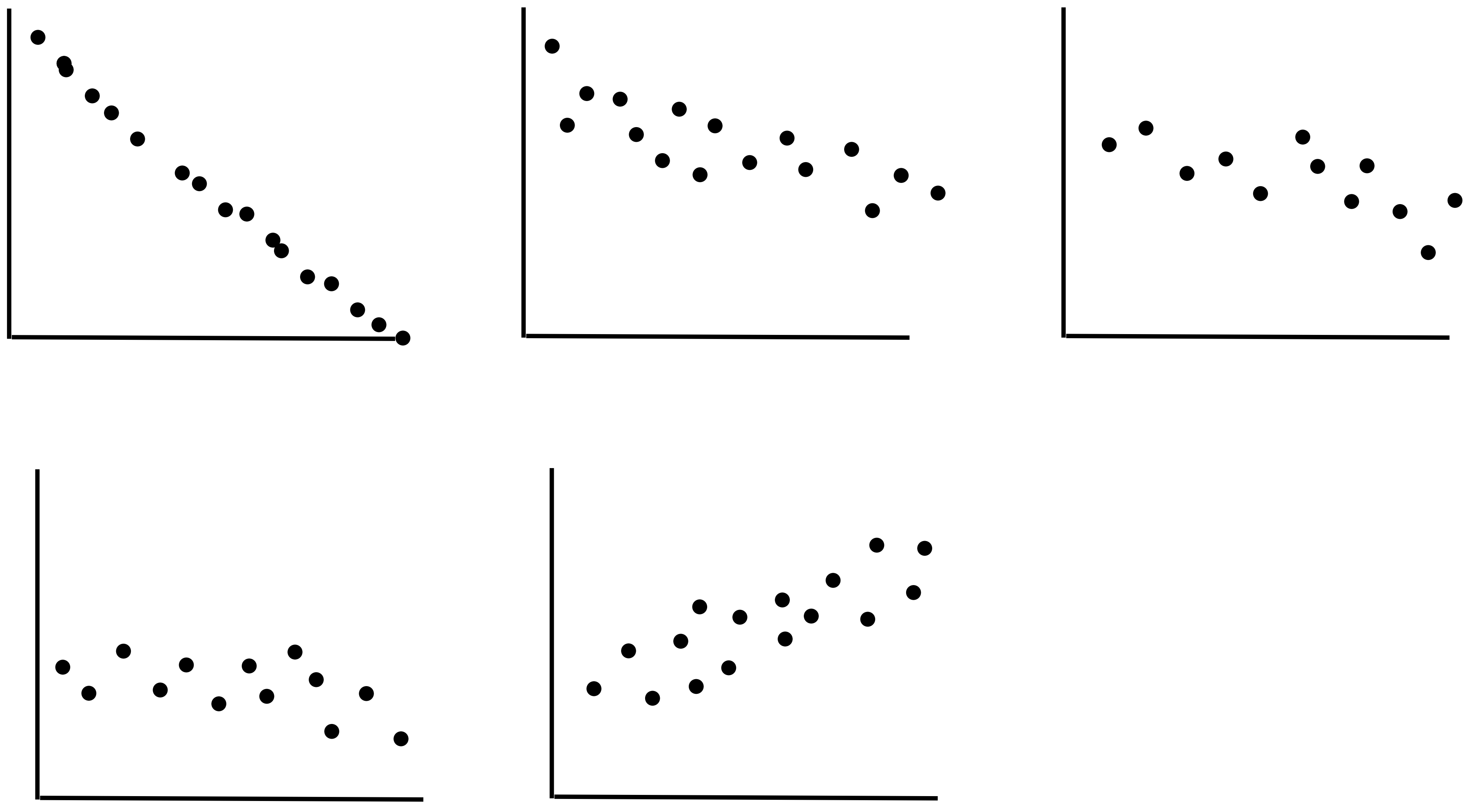


Kalish, Griffiths, Lewandowsky (2007)

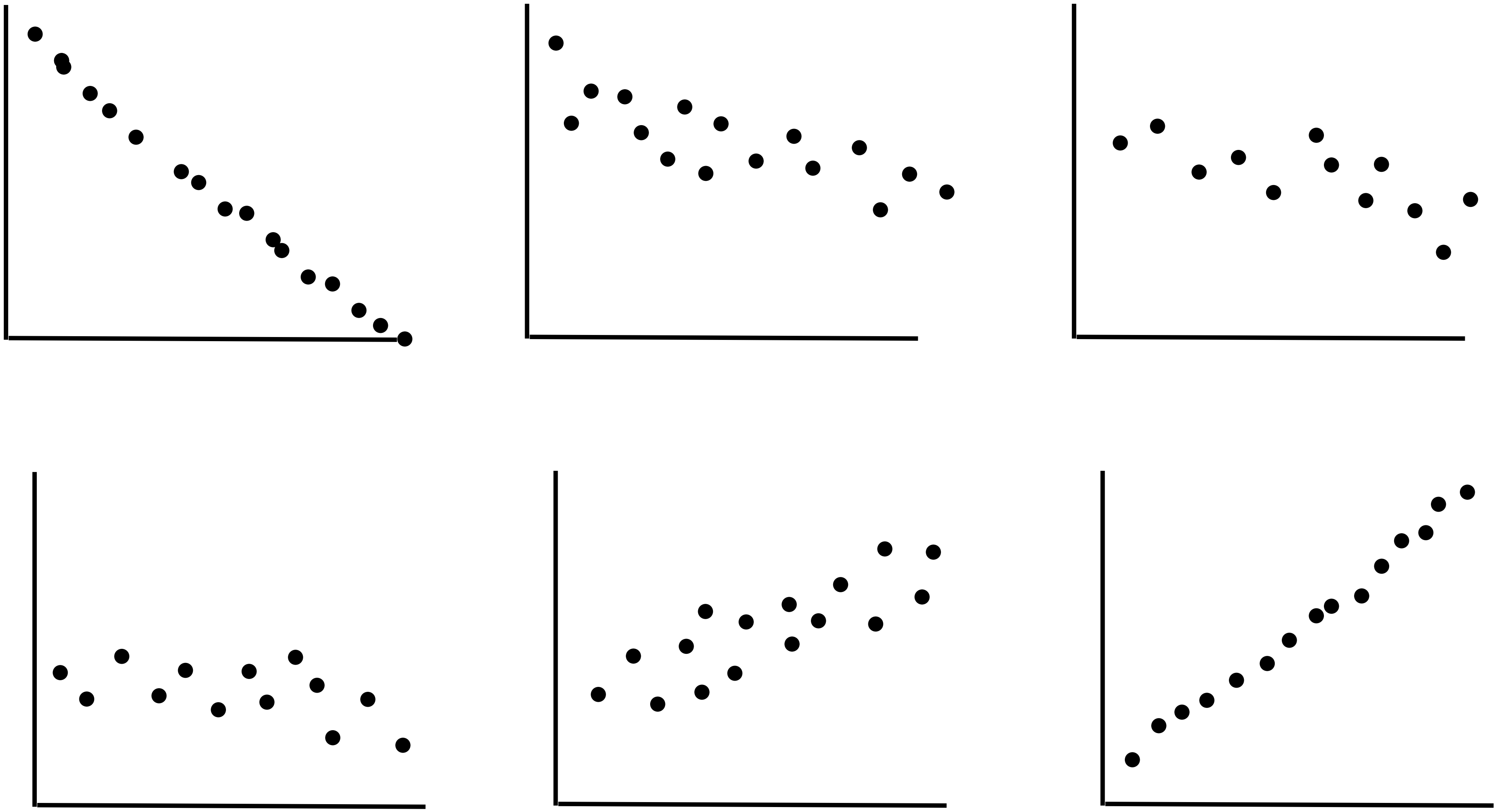
Kalish, Griffiths, Lewandowsky (2007)



Kalish, Griffiths, Lewandowsky (2007)



Kalish, Griffiths, Lewandowsky (2007)



Kalish, Griffiths, Lewandowsky (2007)

LINEAR REGRESSION IN $\mathbb{R}^d$

$$y = h^T x + \mathcal{N}(0, \sigma^2)$$

prior $\mathbb{P}[h] \sim \mathcal{N}(\bar{\mu}, \bar{\sigma}^2 I_d)$

teacher $\mathbb{P}[h] \sim \mathcal{N}(\mu_0, \bar{\sigma}^2 I_d)$

likelihood $\mathbb{P}[y \mid X, h] \sim \mathcal{N}(Xh, \sigma^2)$

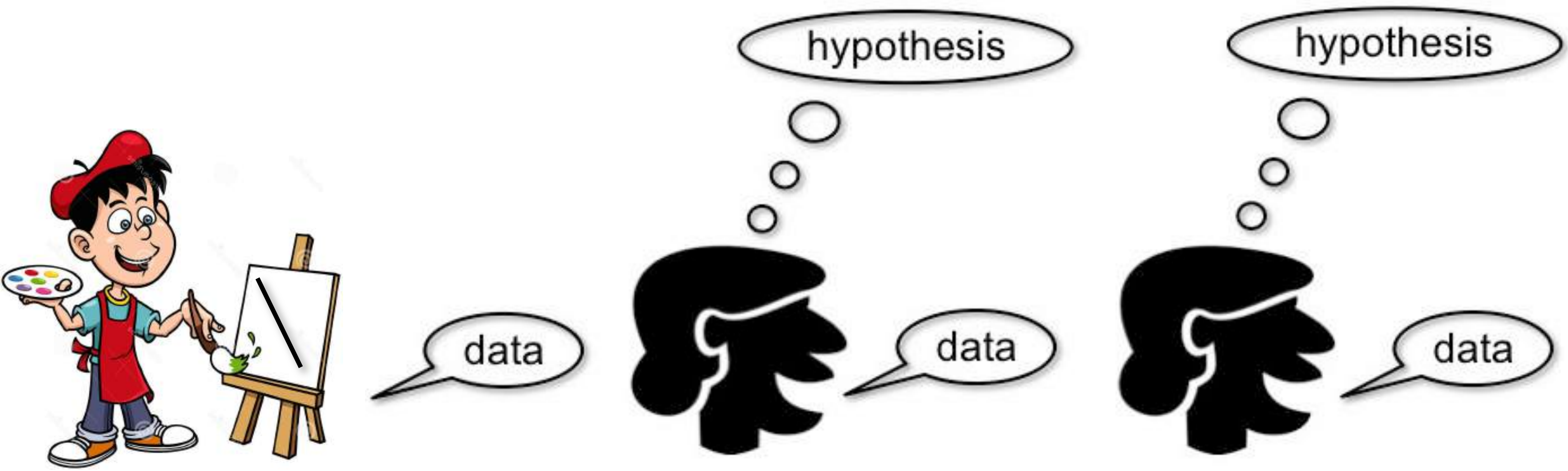
$$y = Xh + \text{noise}$$

$$\text{prior } \mathbb{P}[h] \sim \mathcal{N}(\bar{\mu}, \bar{\sigma}^2 I_d)$$

$$\text{teacher } \mathbb{P}[h] \sim \mathcal{N}(\mu_0, \bar{\sigma}^2 I_d)$$

$$\text{likelihood } \mathbb{P}[y \mid X, h] \sim \mathcal{N}(Xh, \sigma^2)$$

$$y = Xh + \text{noise}$$

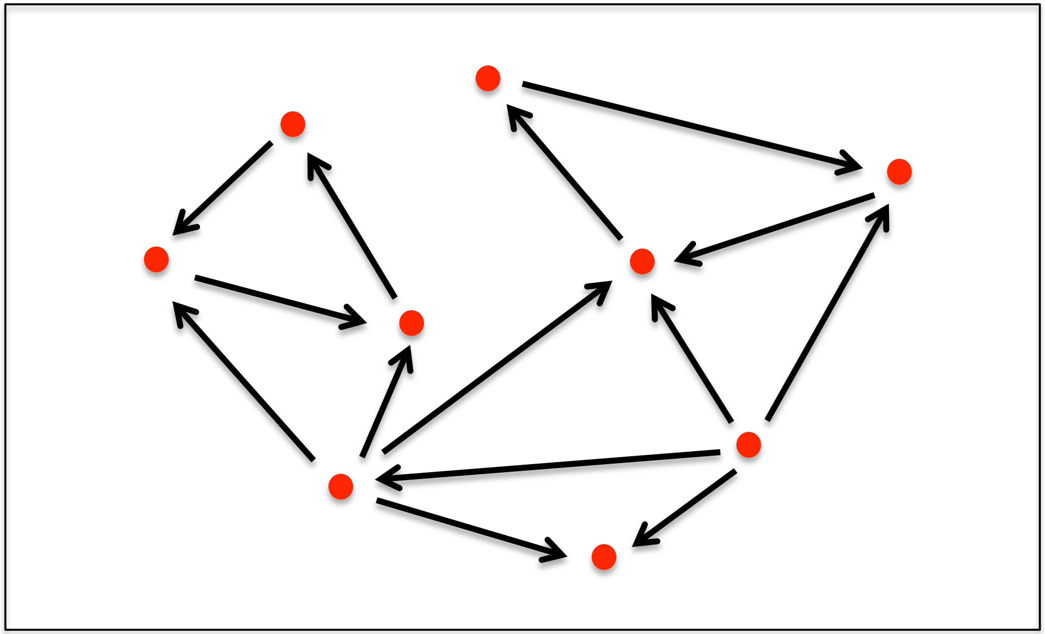


$$m_t \approx \frac{\| \mu_0 - \bar{\mu} \|_2}{\delta} \left(\frac{\sigma}{\bar{\sigma}}\right)^2 t^{1+c} + d \log \frac{t+1}{\varepsilon}$$

Sustained iterated learning requires keeping system out of equilibrium

Markov chain

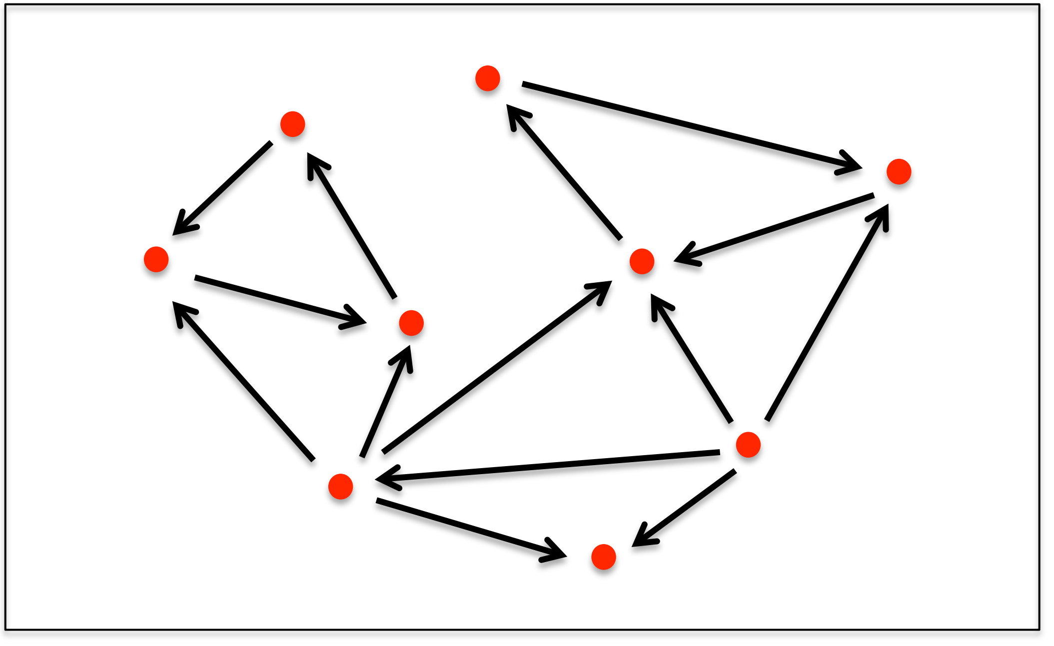
always reaches equilibrium



Markov chain

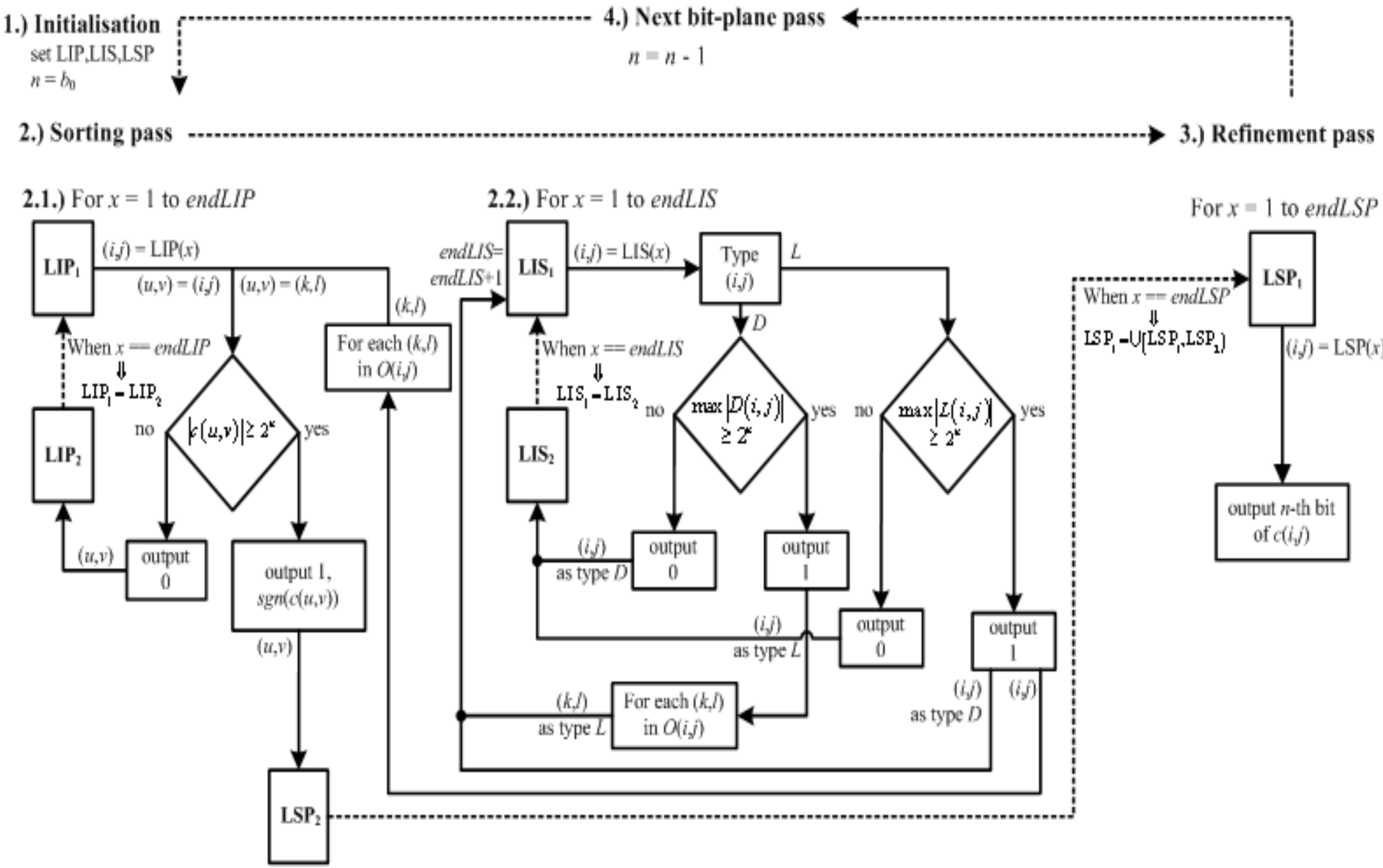
for sustainability
keep out of equilibrium

by injecting free energy

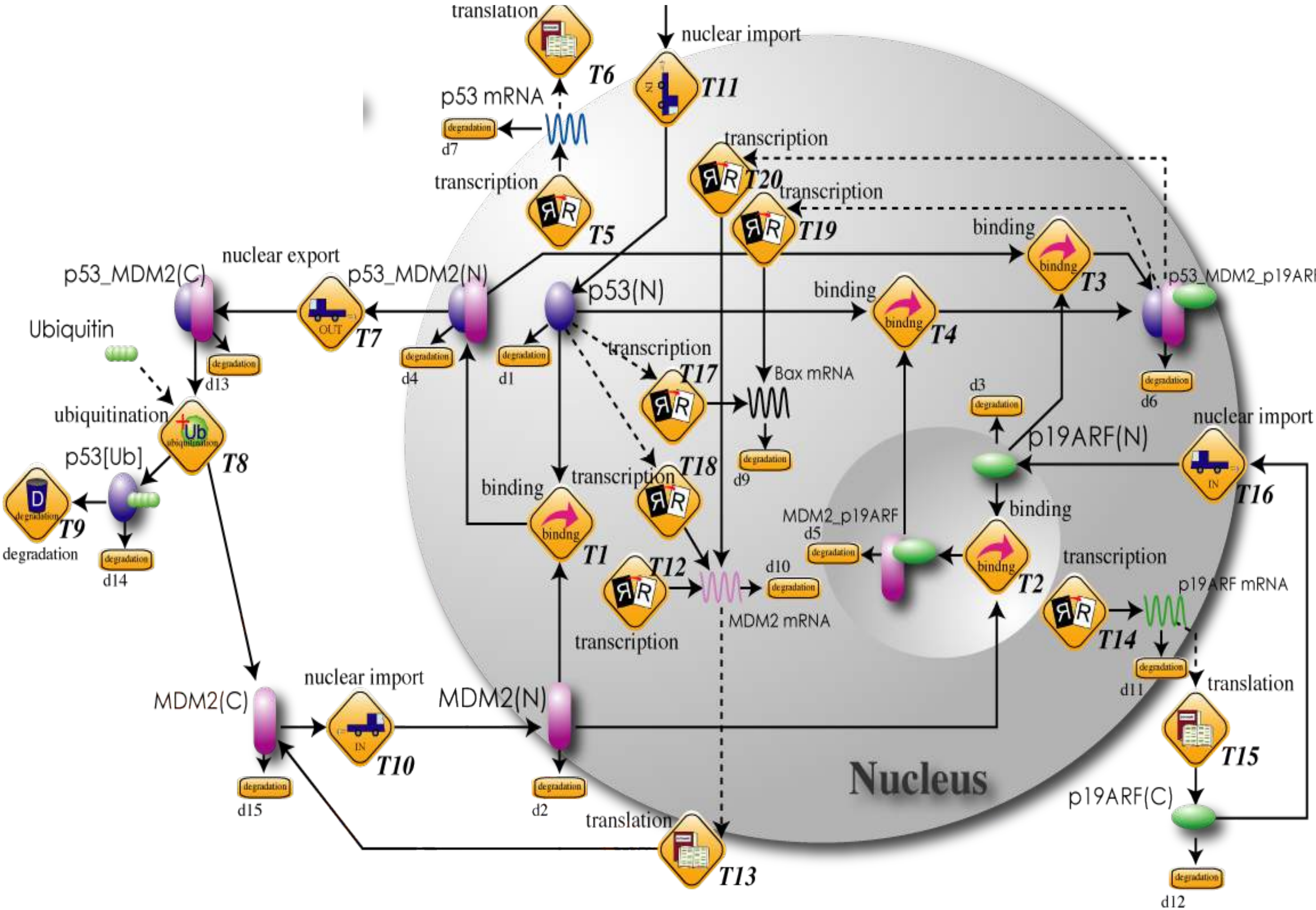


natural algorithm

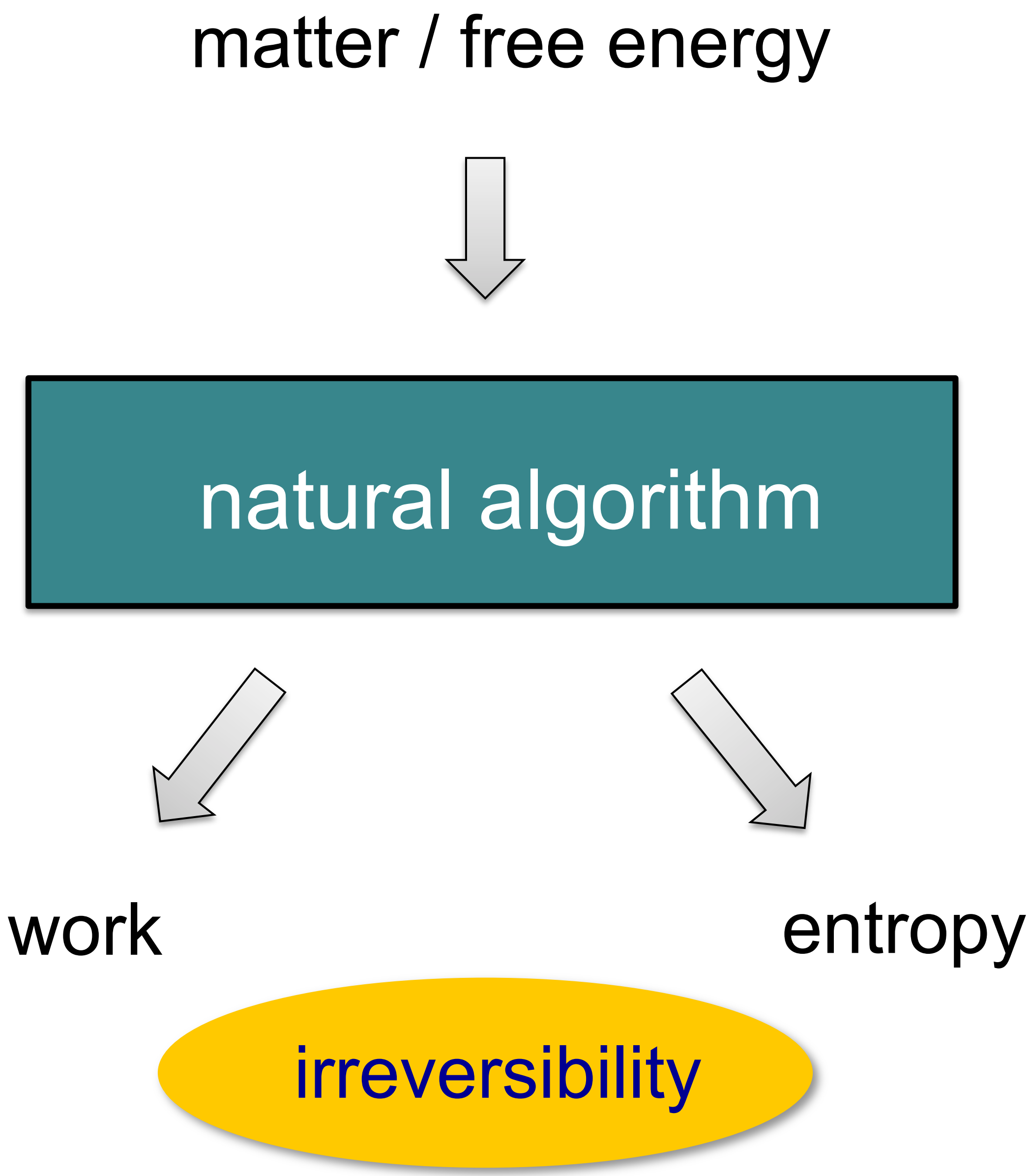
classical algorithm



natural algorithm



natural algorithm


matter / free energy
natural algorithm
work
entropy
irreversibility

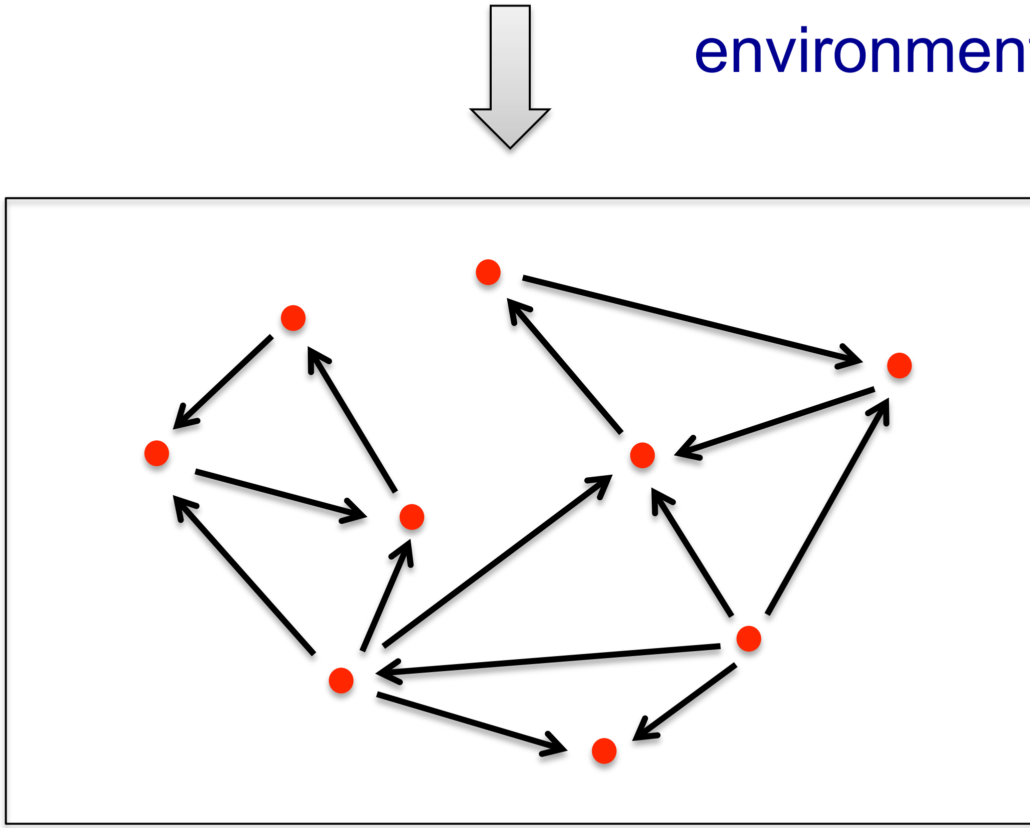
signals

driving fields

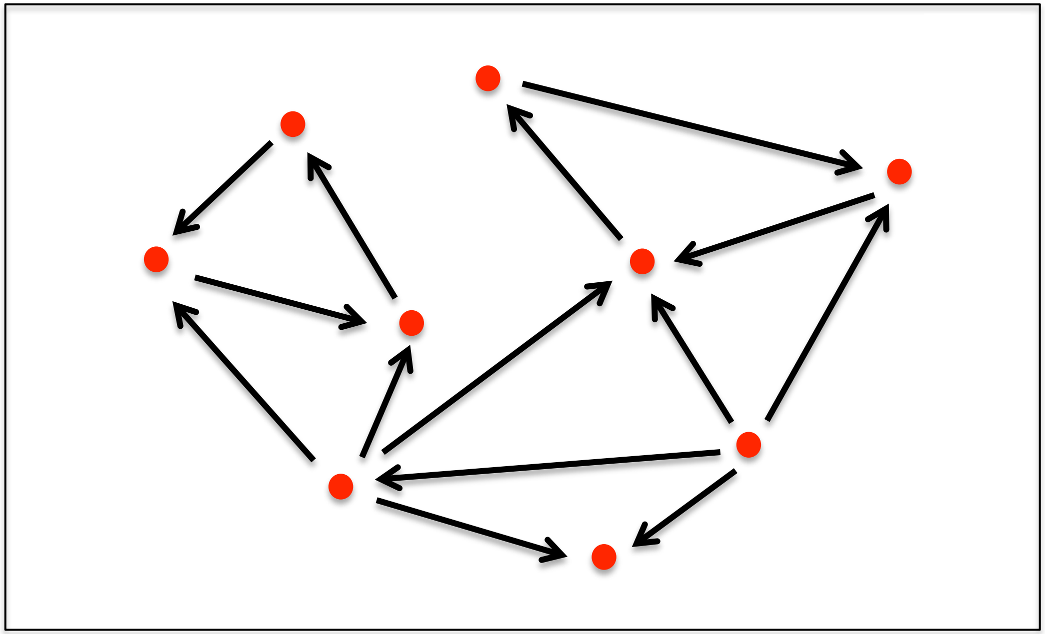
carbon sources

environmental fluctuations

matter / free energy



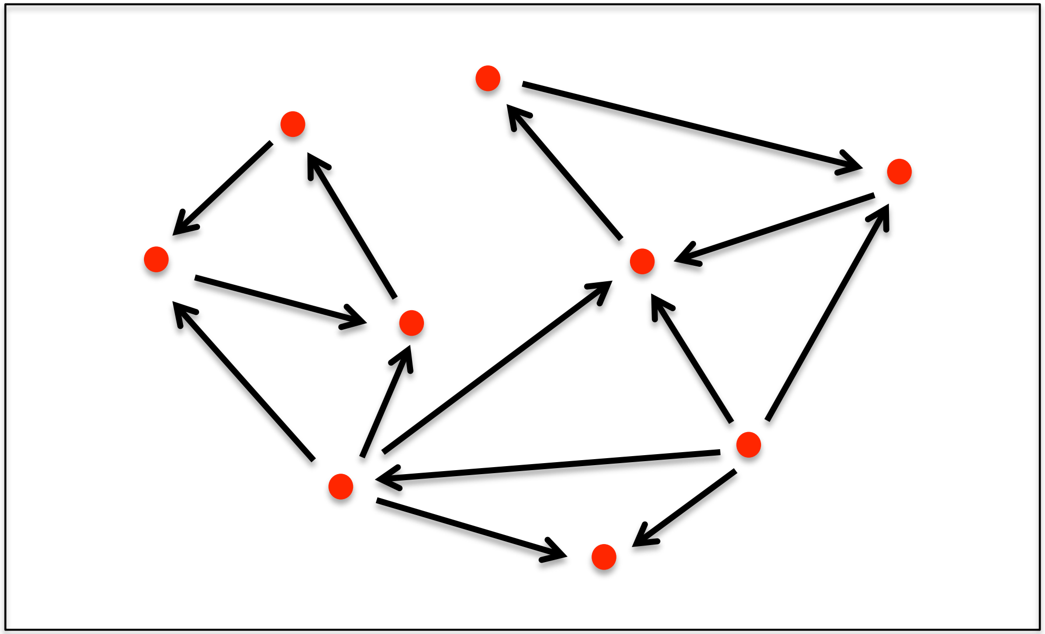
Influence Systems



Each node is an agent; at any time, it is in a state and:

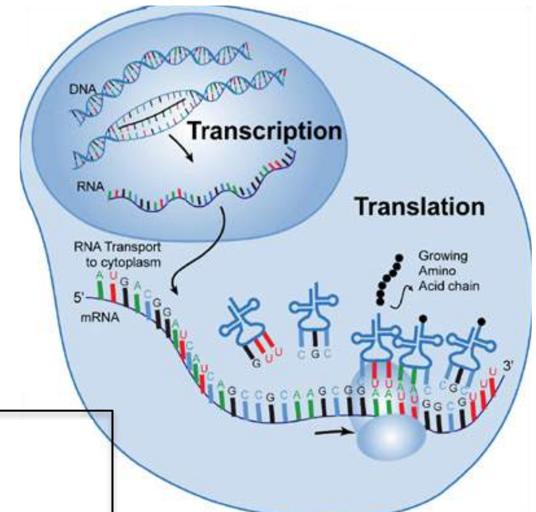
- *it picks its in-edges*
- *it updates its new state*

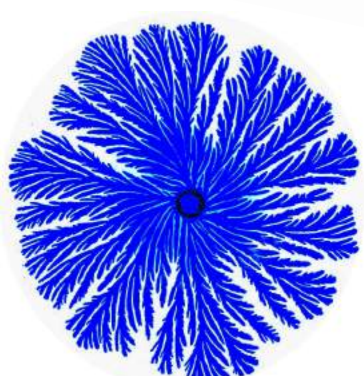
} these rules form the agent's *type*

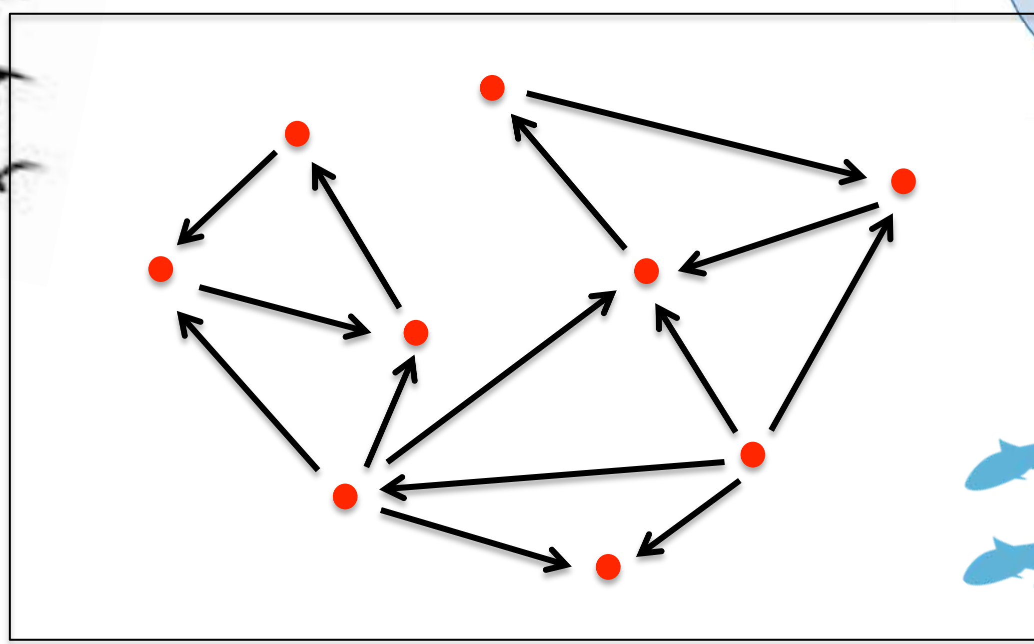


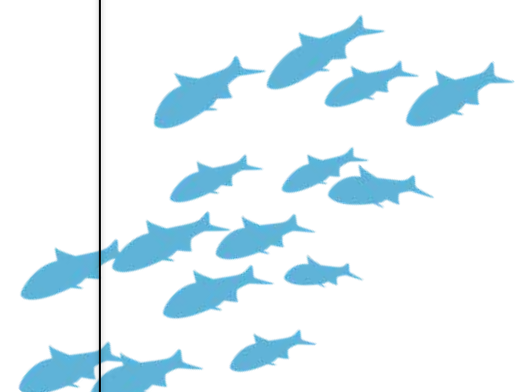
Each node is an agent; at any time, it is in a state and:

- *it picks its in-edges*
- *it updates its new state*



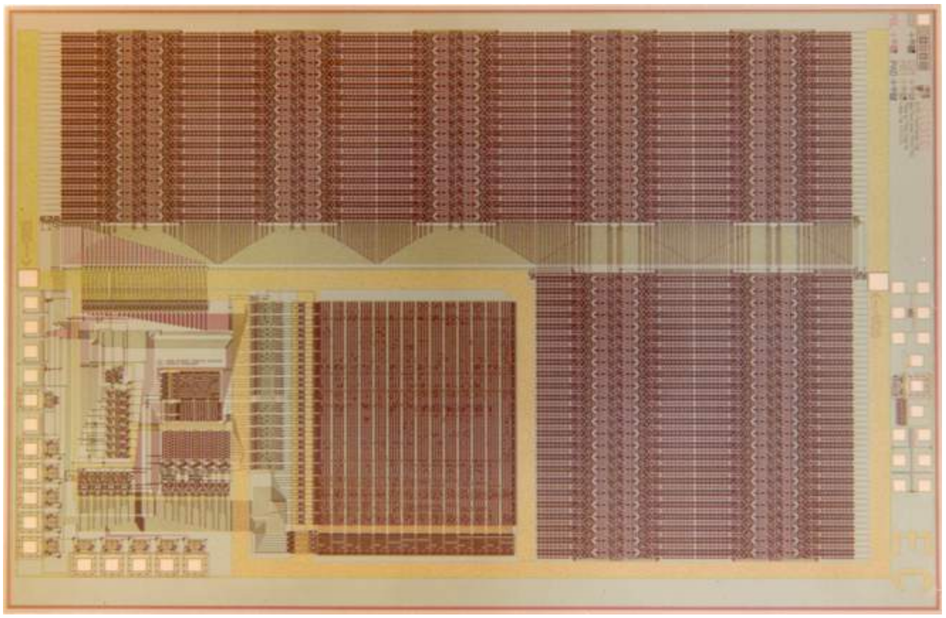






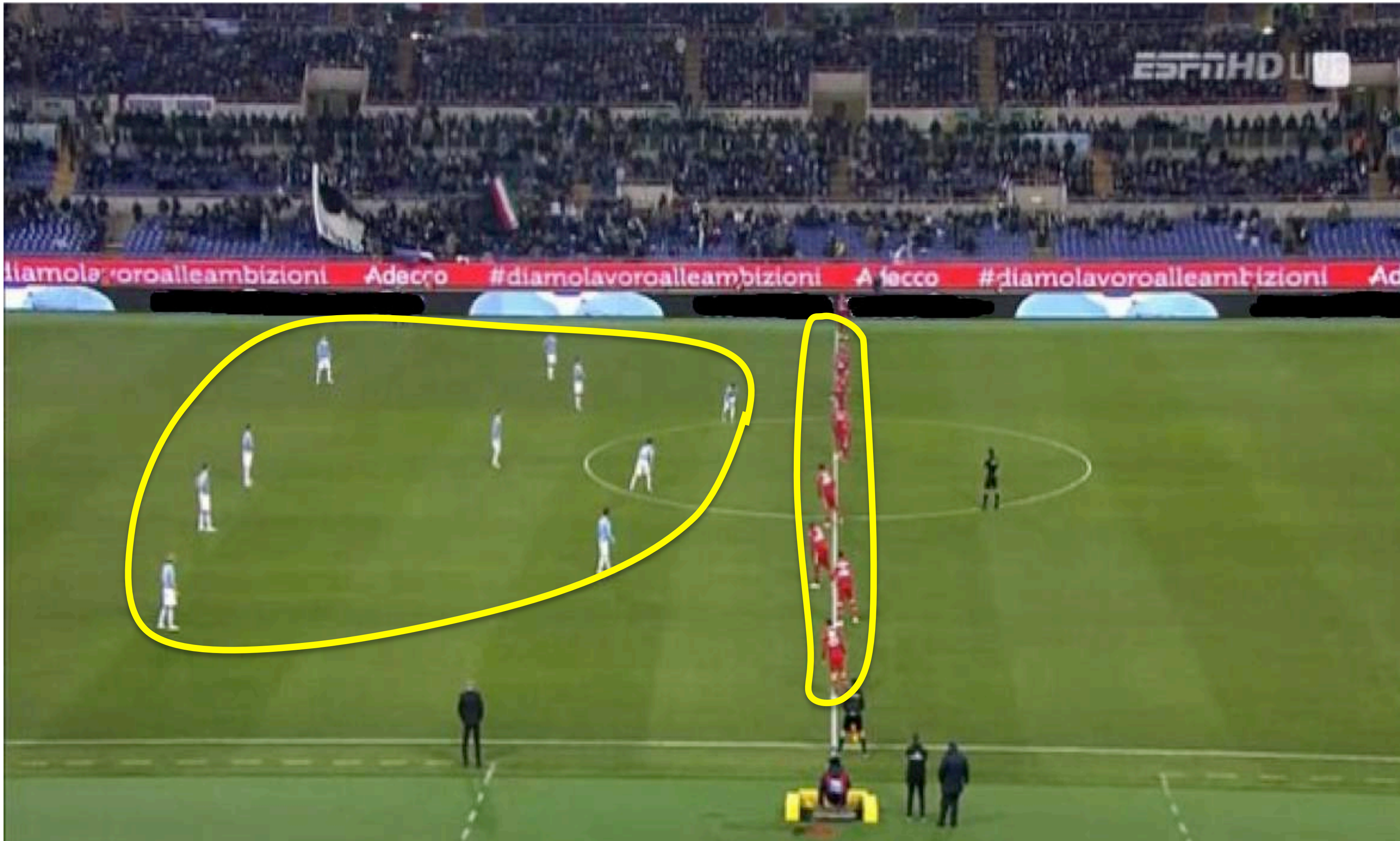
- ## Why theory ?

The Big Data pipe dream



Semantic renormalization

coarse-graining
dimension reduction
hierarchical graph clustering
abstraction

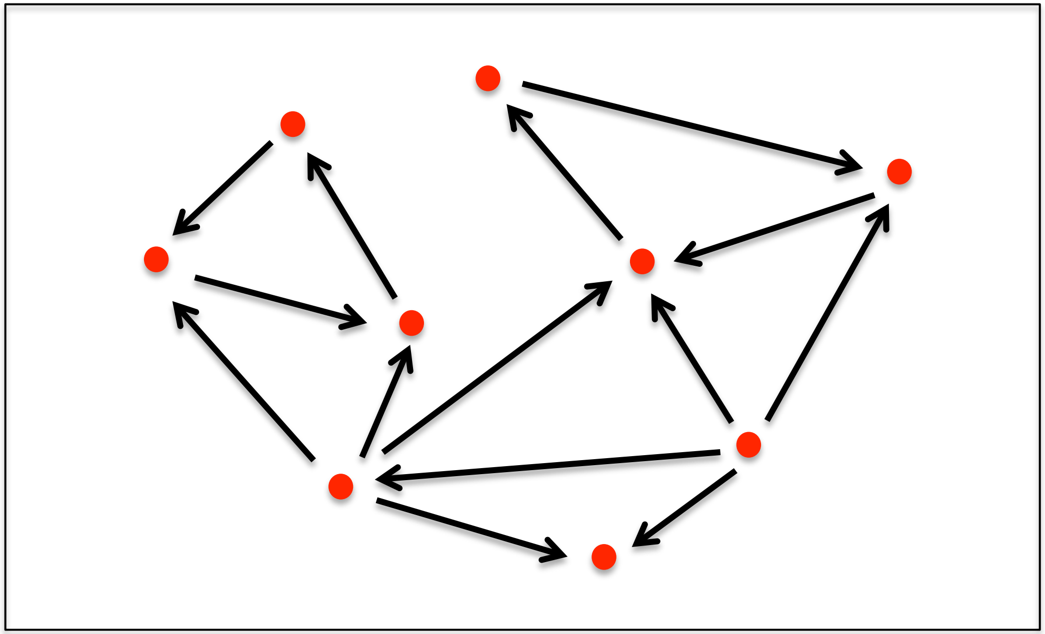

ESPNHD
#diamolavoroalleambizioni
Adecco

Clustering is path dependent

- Closed system

Markov chain

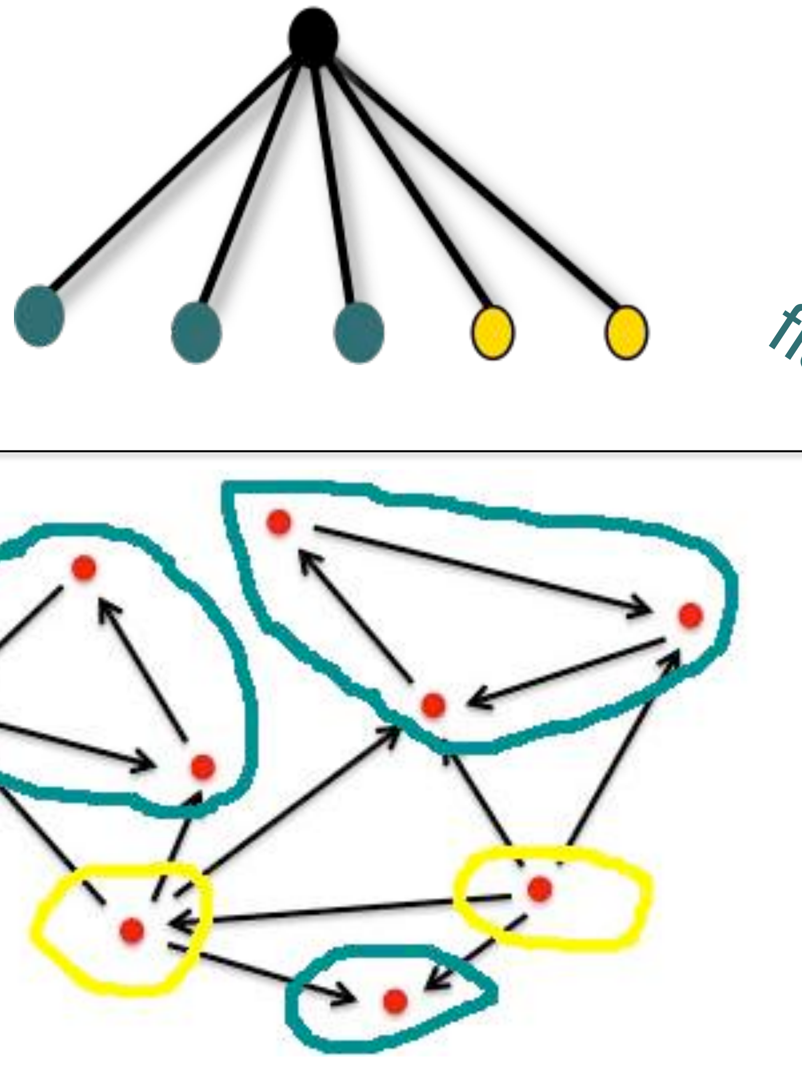
classify all orbits



- Closed system

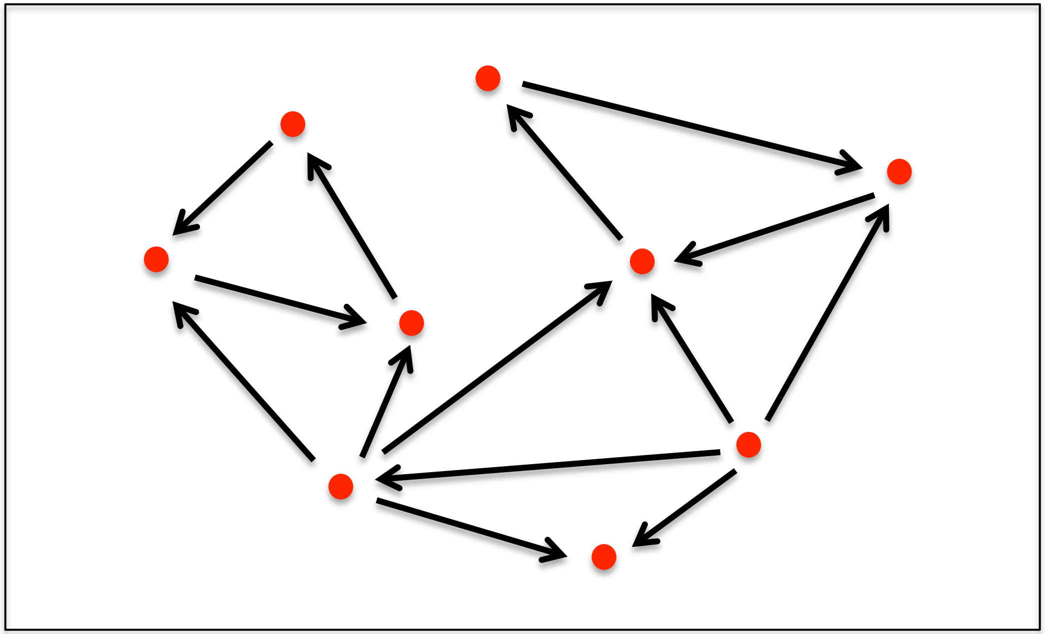
Markov chain decomposition

two kinds of super-agents



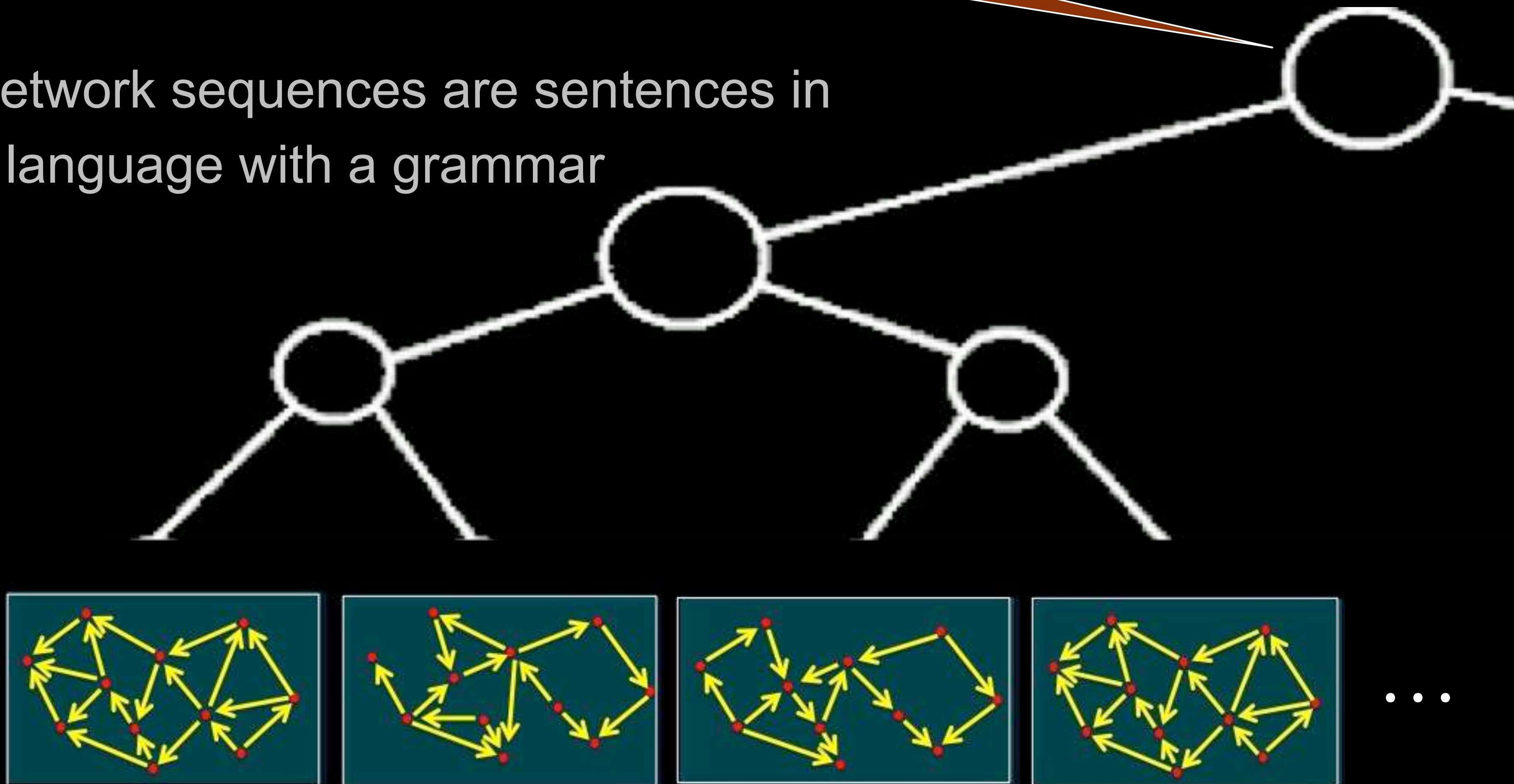
flat hierarchy

- Open system *dynamics of dynamic network*
- Track flow of information to parse network sequence *deep renormalization hierarchy*

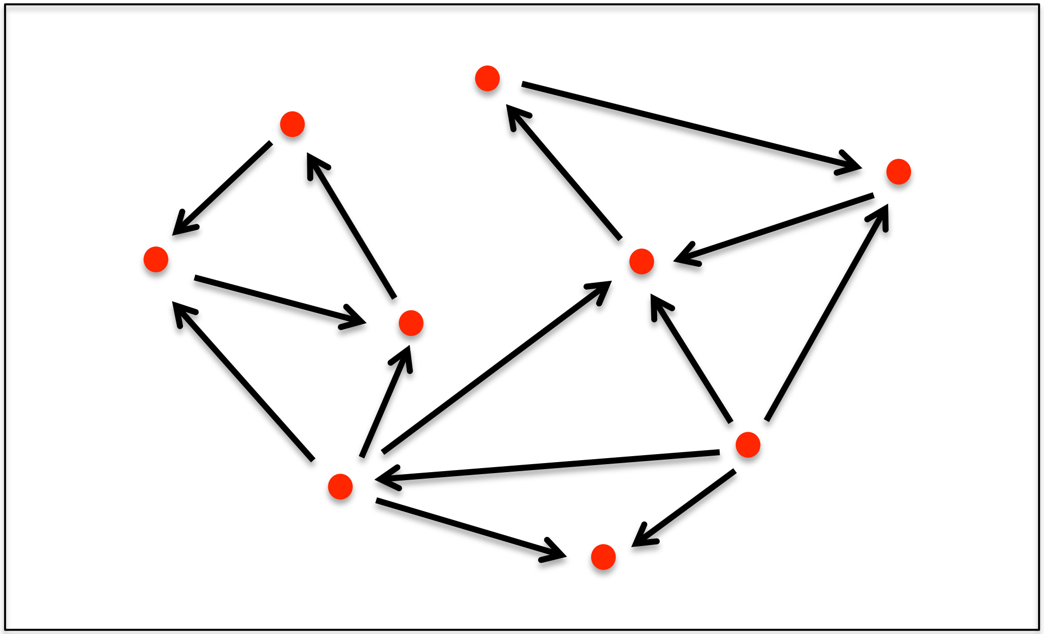


coarse-grained network

Network sequences are sentences in a language with a grammar

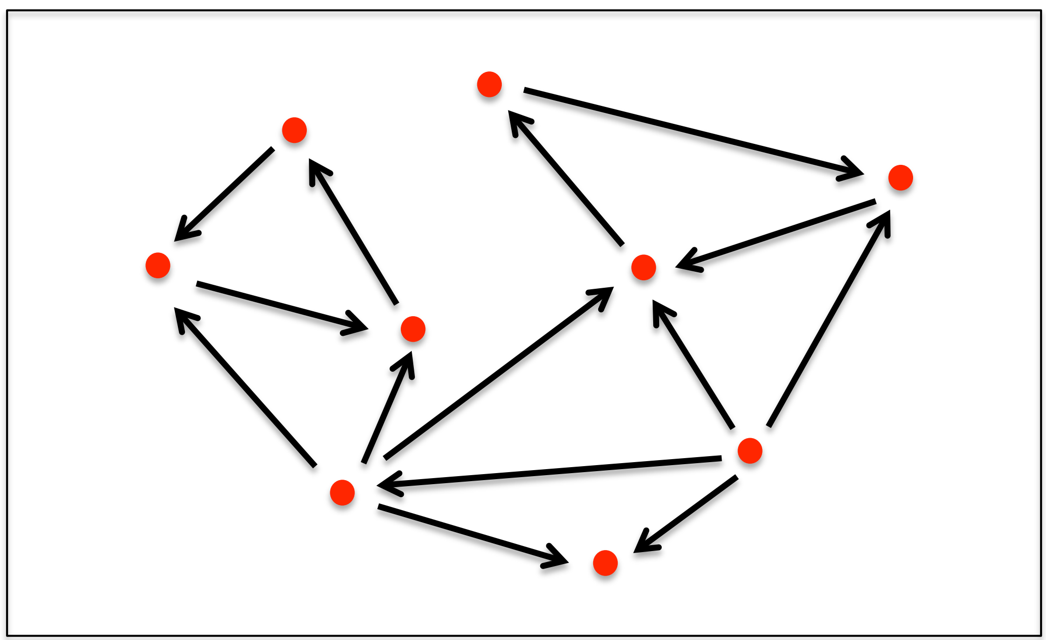


Theorem *Under some conditions, almost all orbits are limit cycles.*

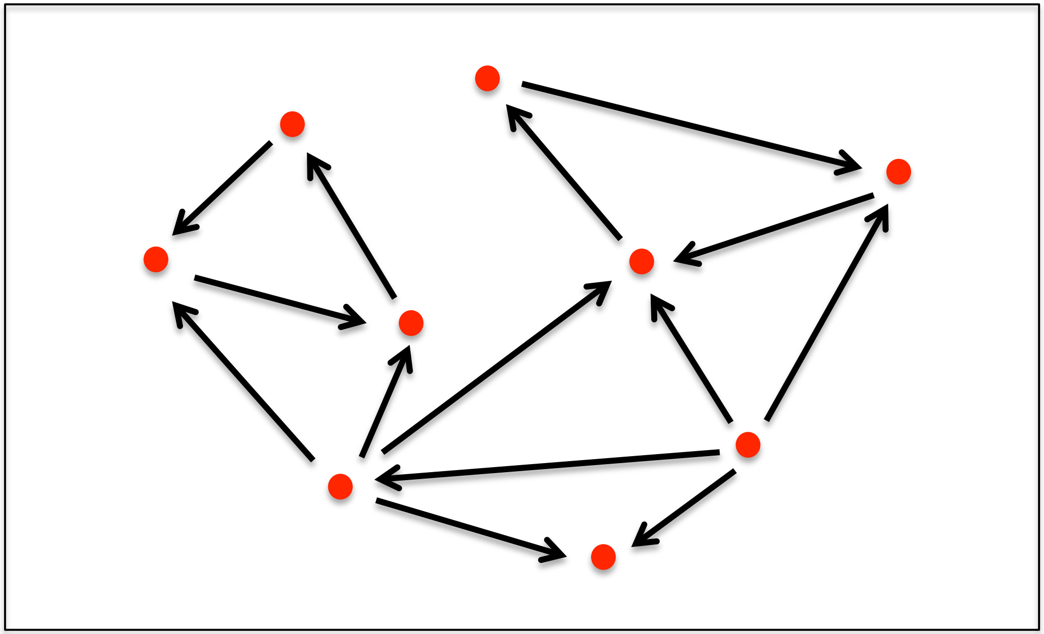


Mixed timescales and archival mechanisms

ratios > 10^{15}

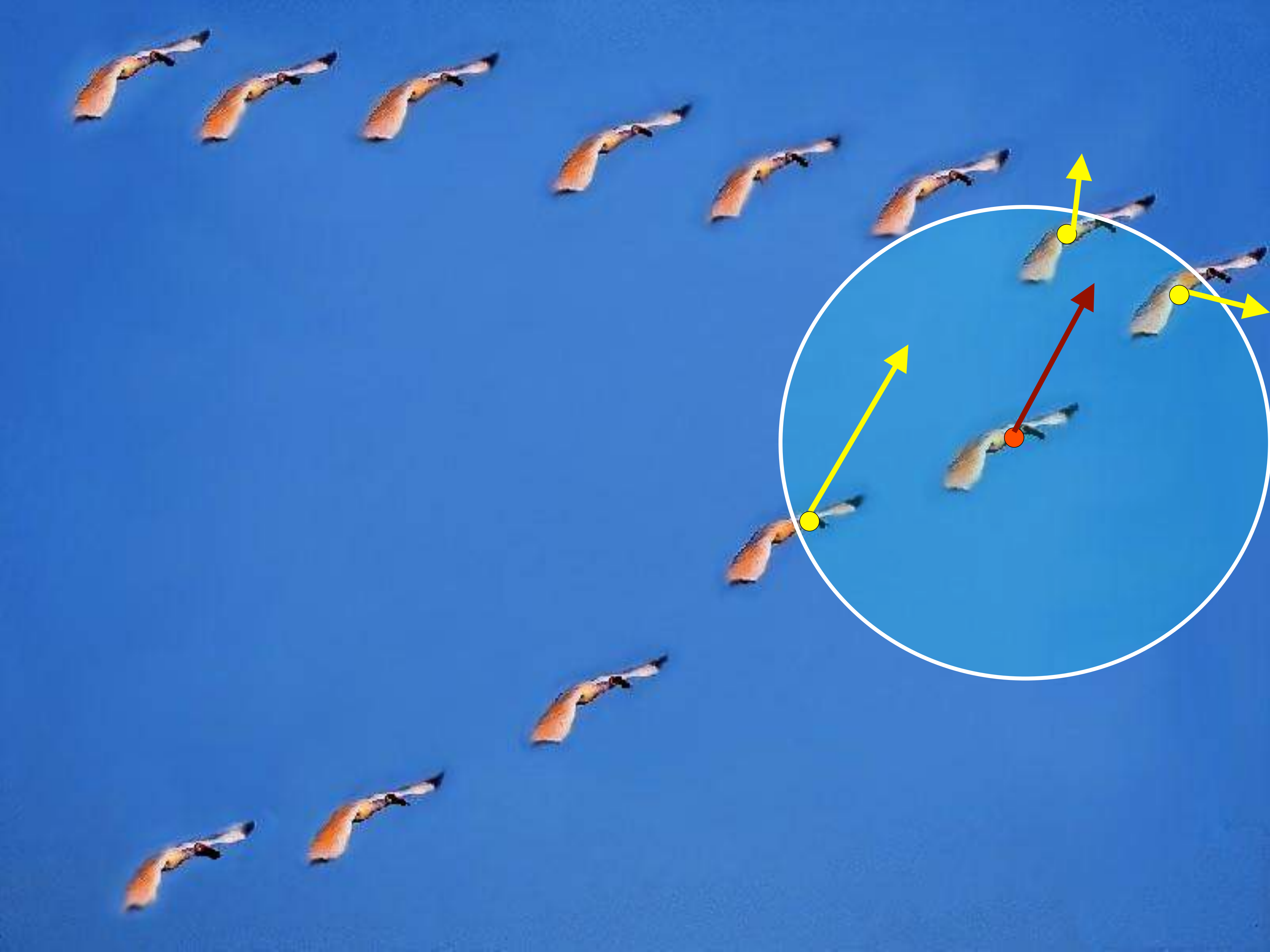


Emergence of memory



Vicsek-Cucker-Smale model


radius *R*




at time t

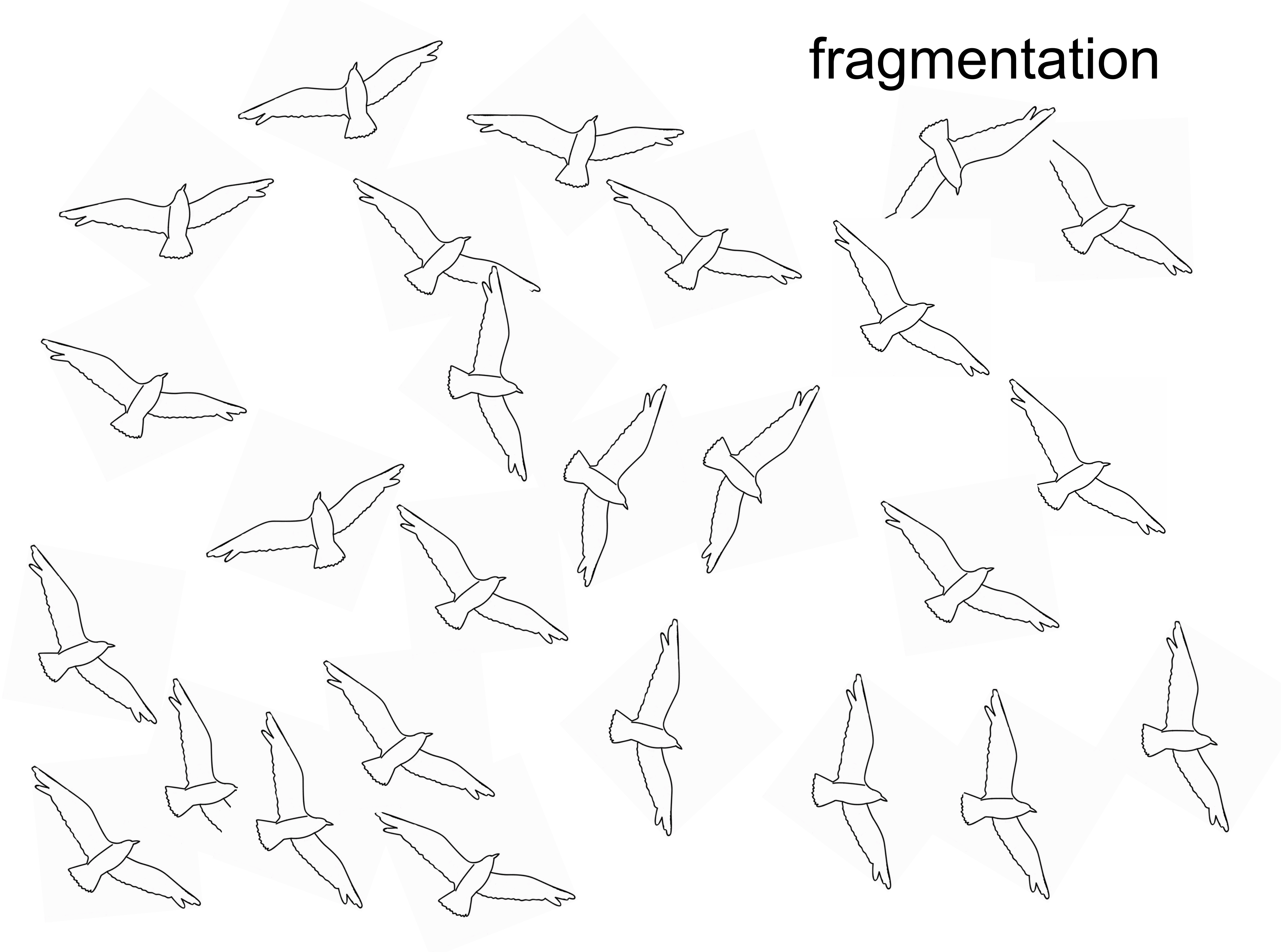

at time t


at time t+1

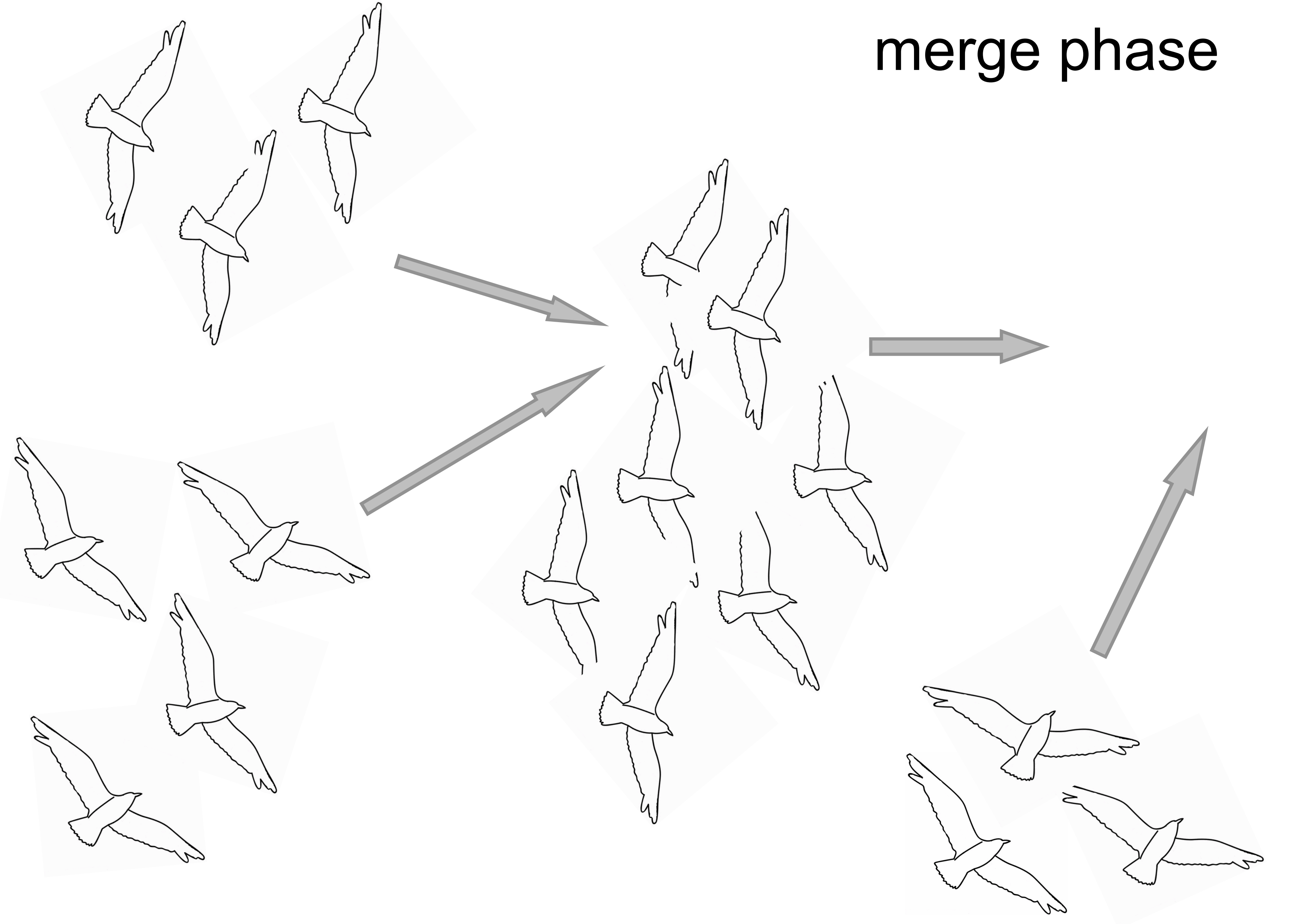
[C 2009]

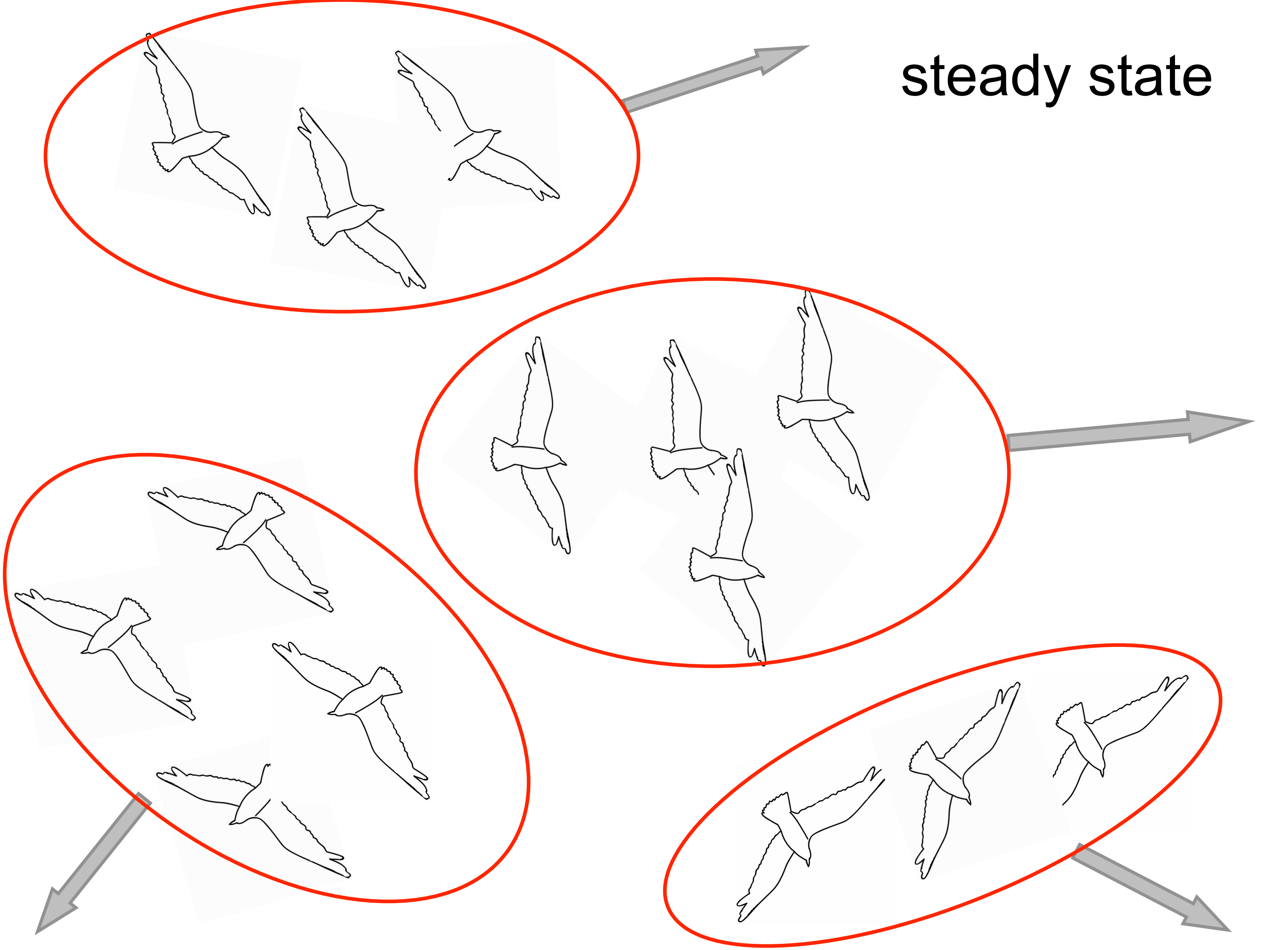
Graph eventually settles

fragmentation



merge phase




steady state

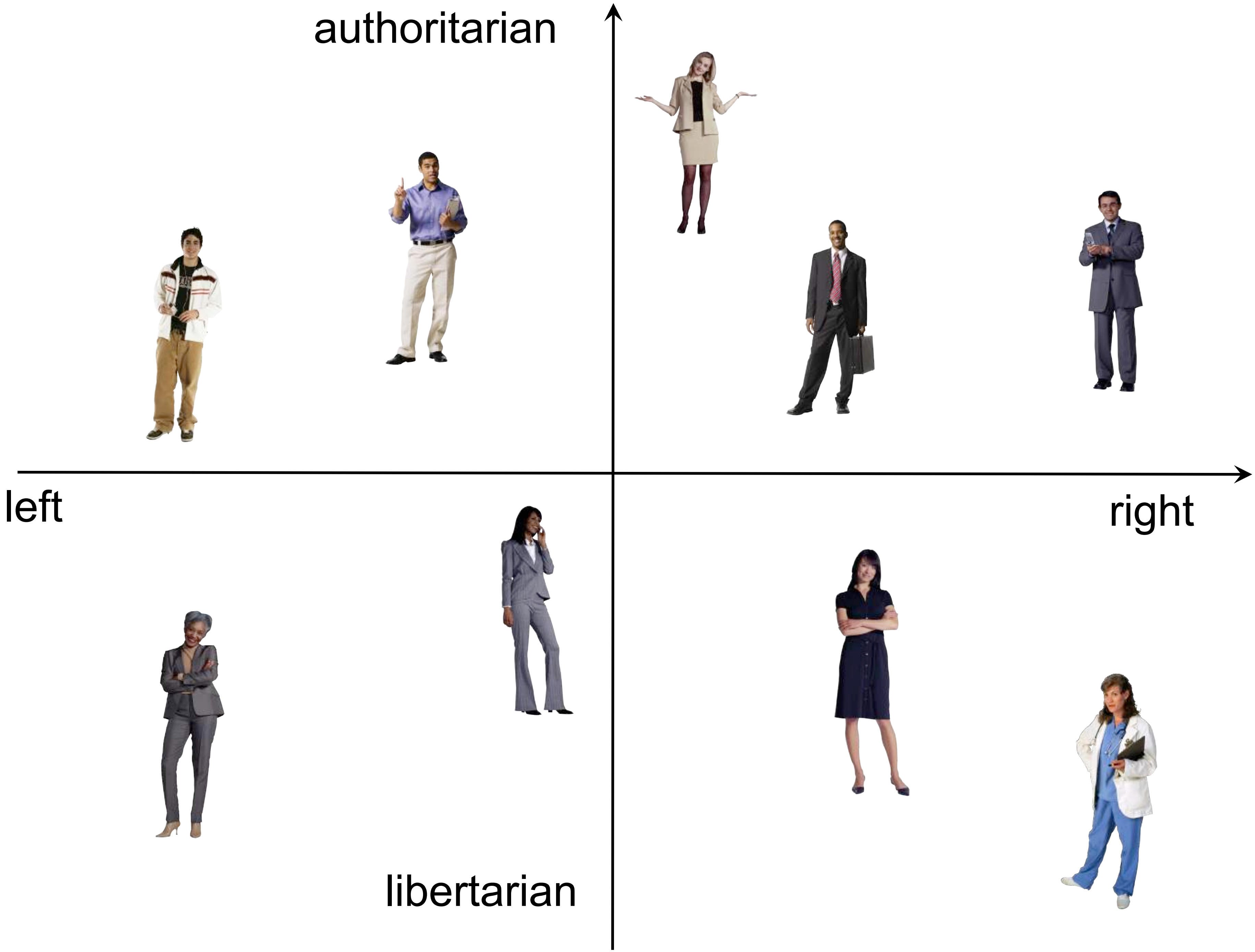
Worst case convergence time

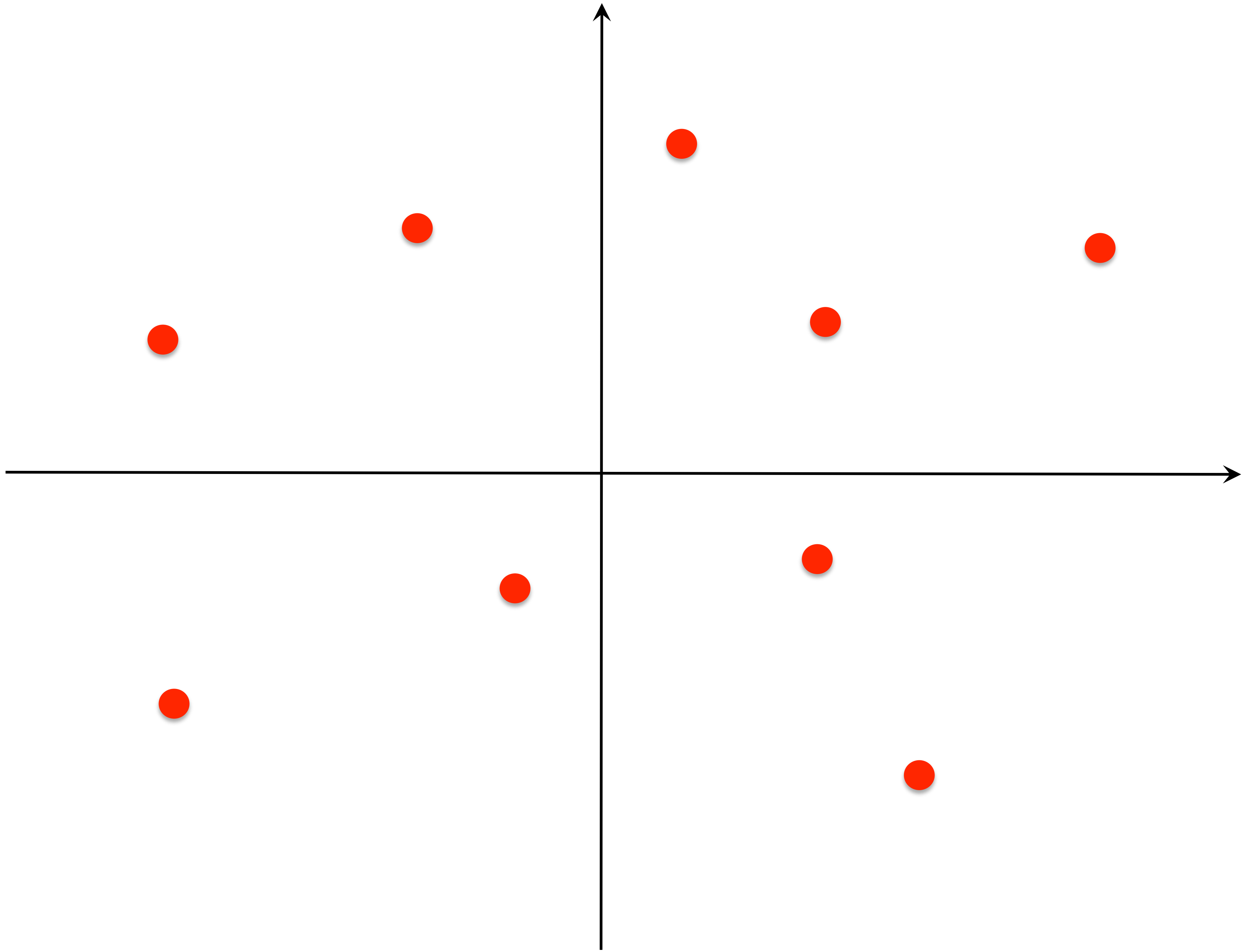
$$2^{2^{2^{\cdot^{\cdot^{\cdot^{2^{2^{2}}}}}}}}$$

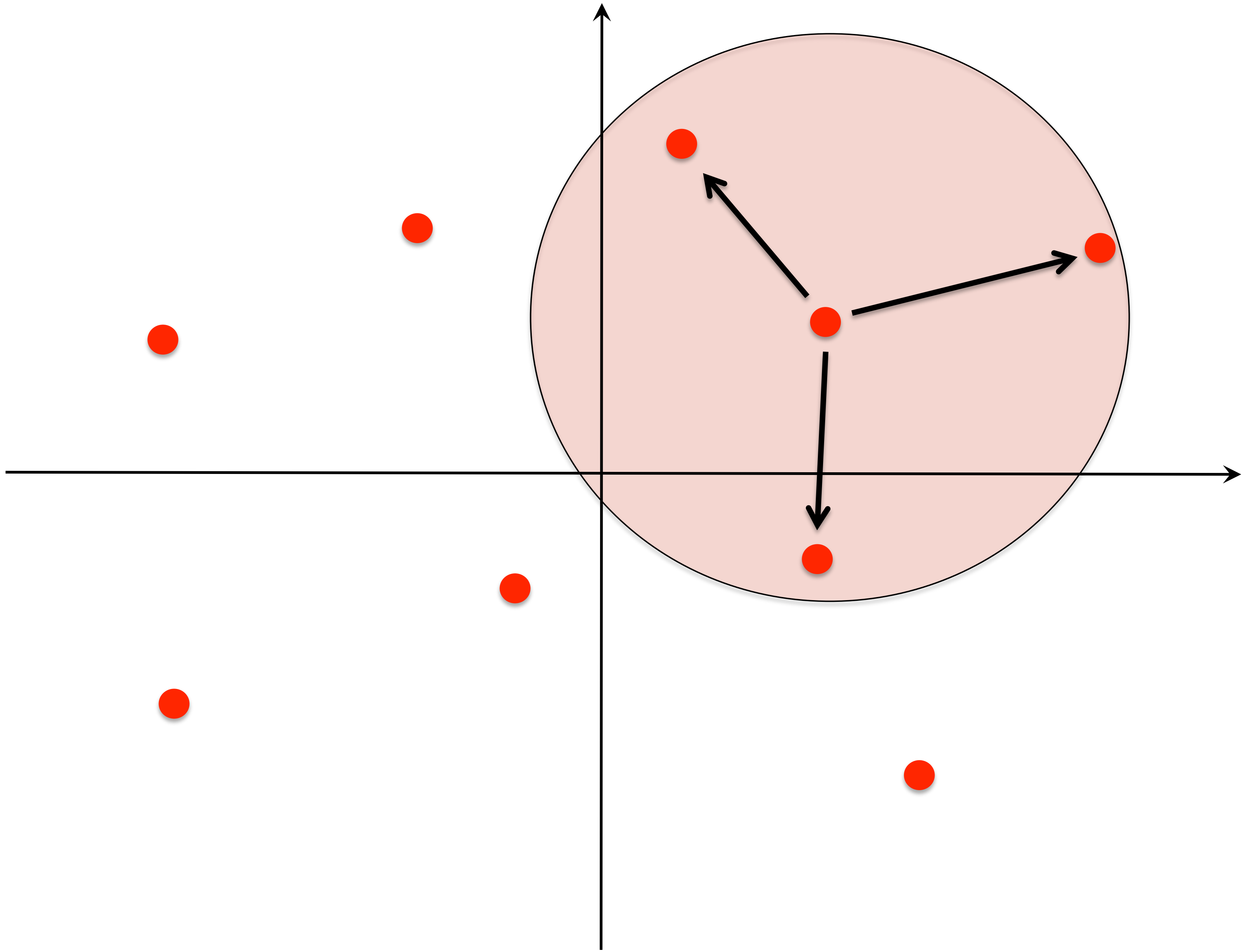
$\sim \log \#\,\text{birds}$

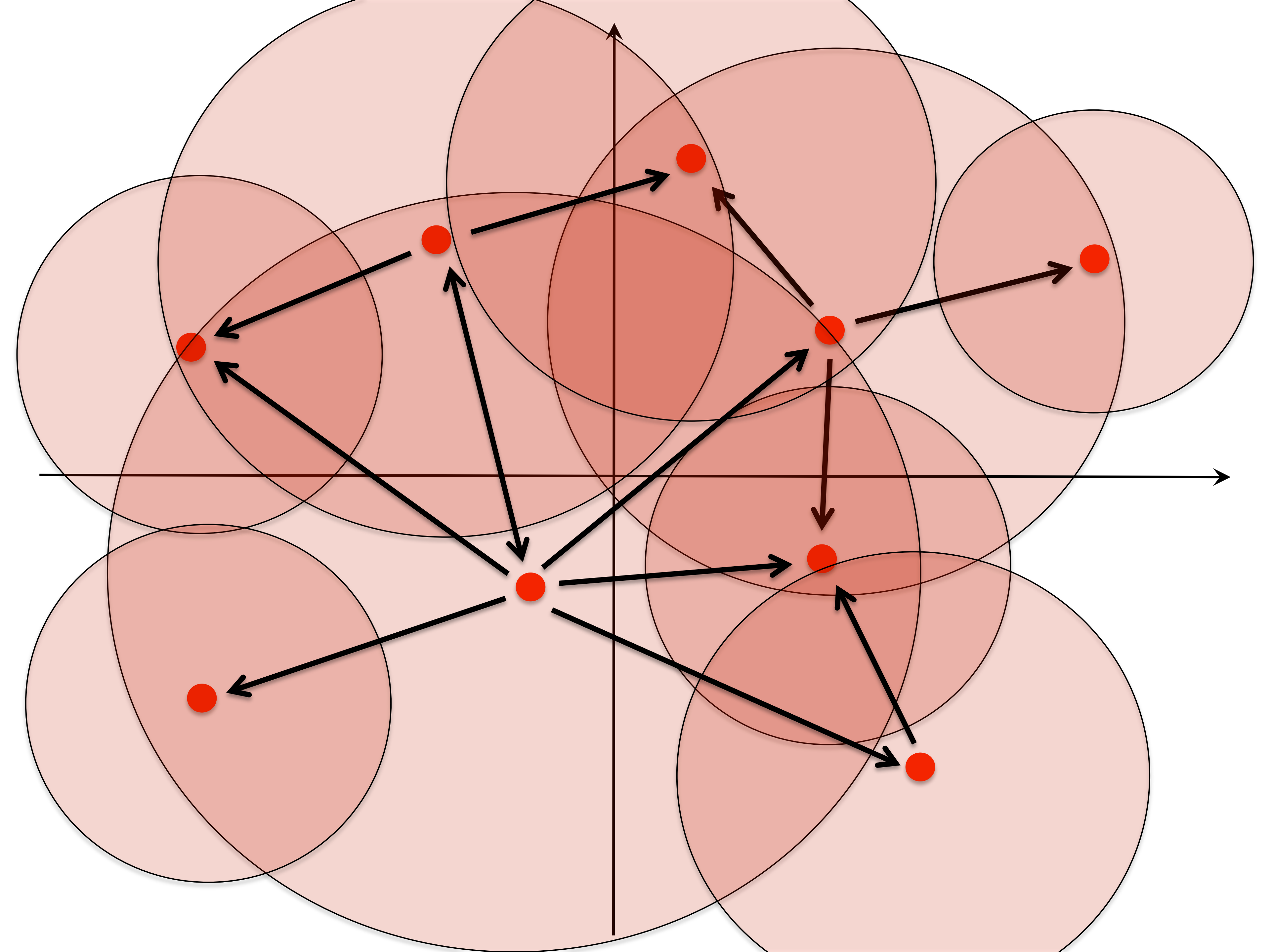
It is optimal !

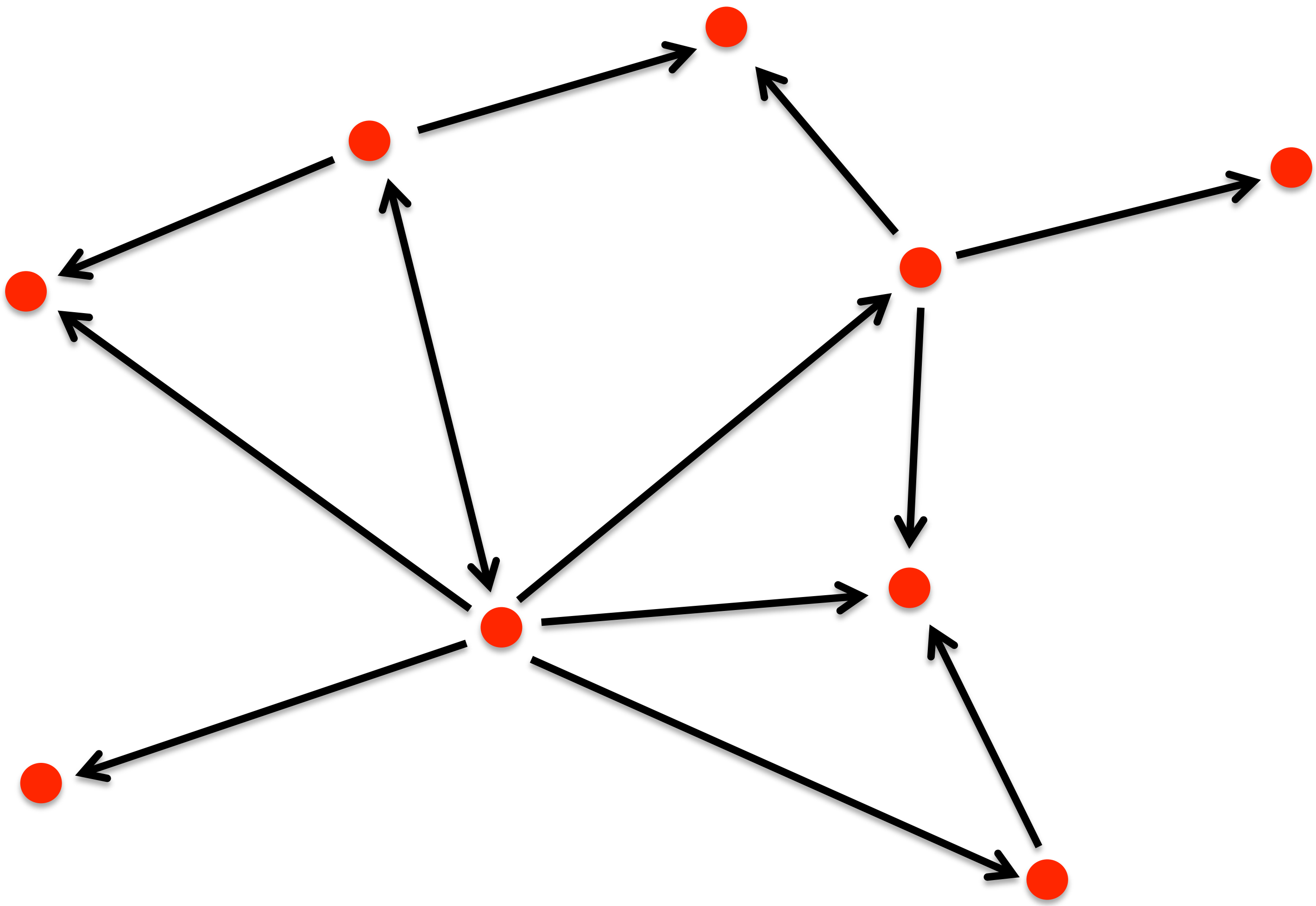
Hegselmann-Krause systems


authoritarian
left
right
libertarian

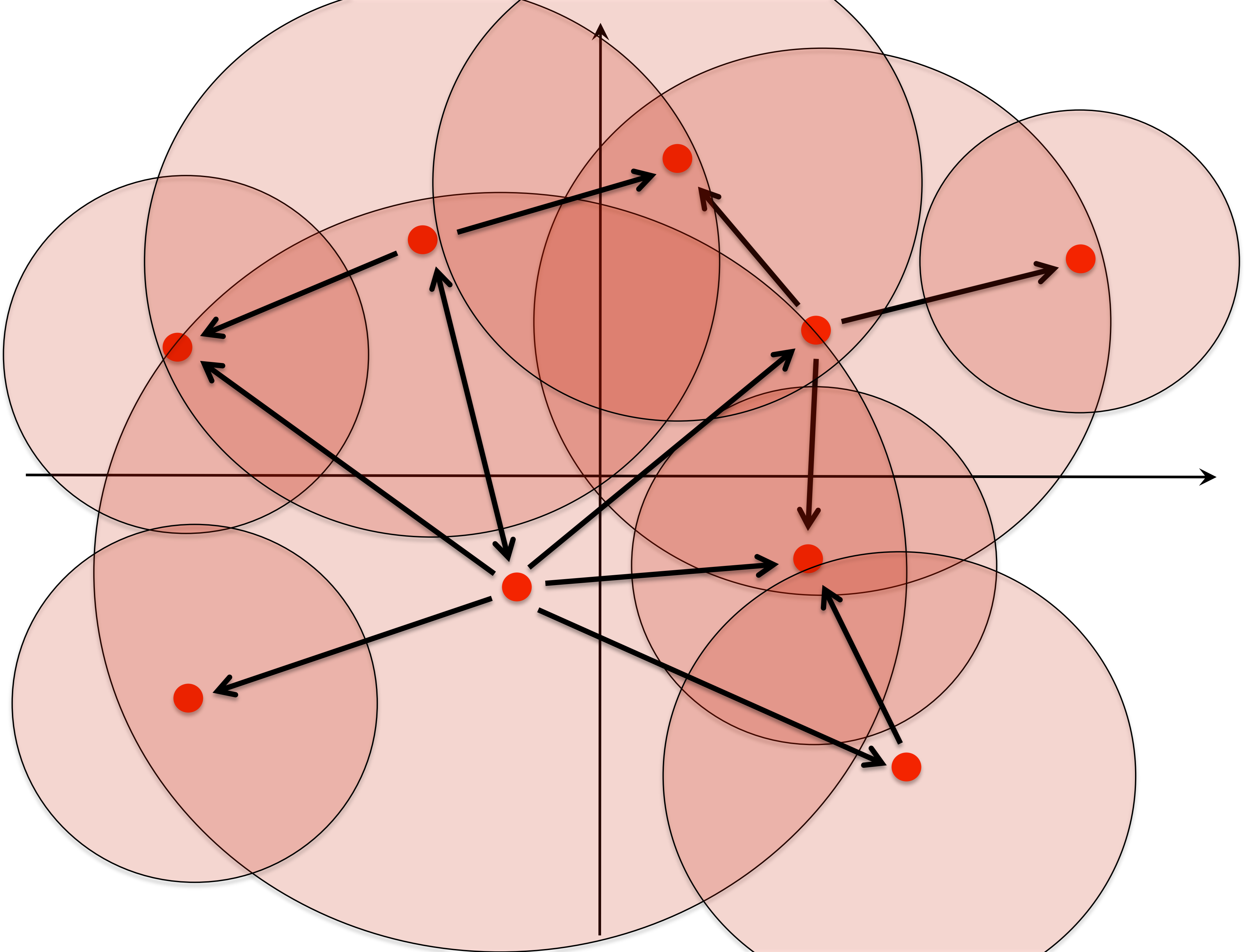




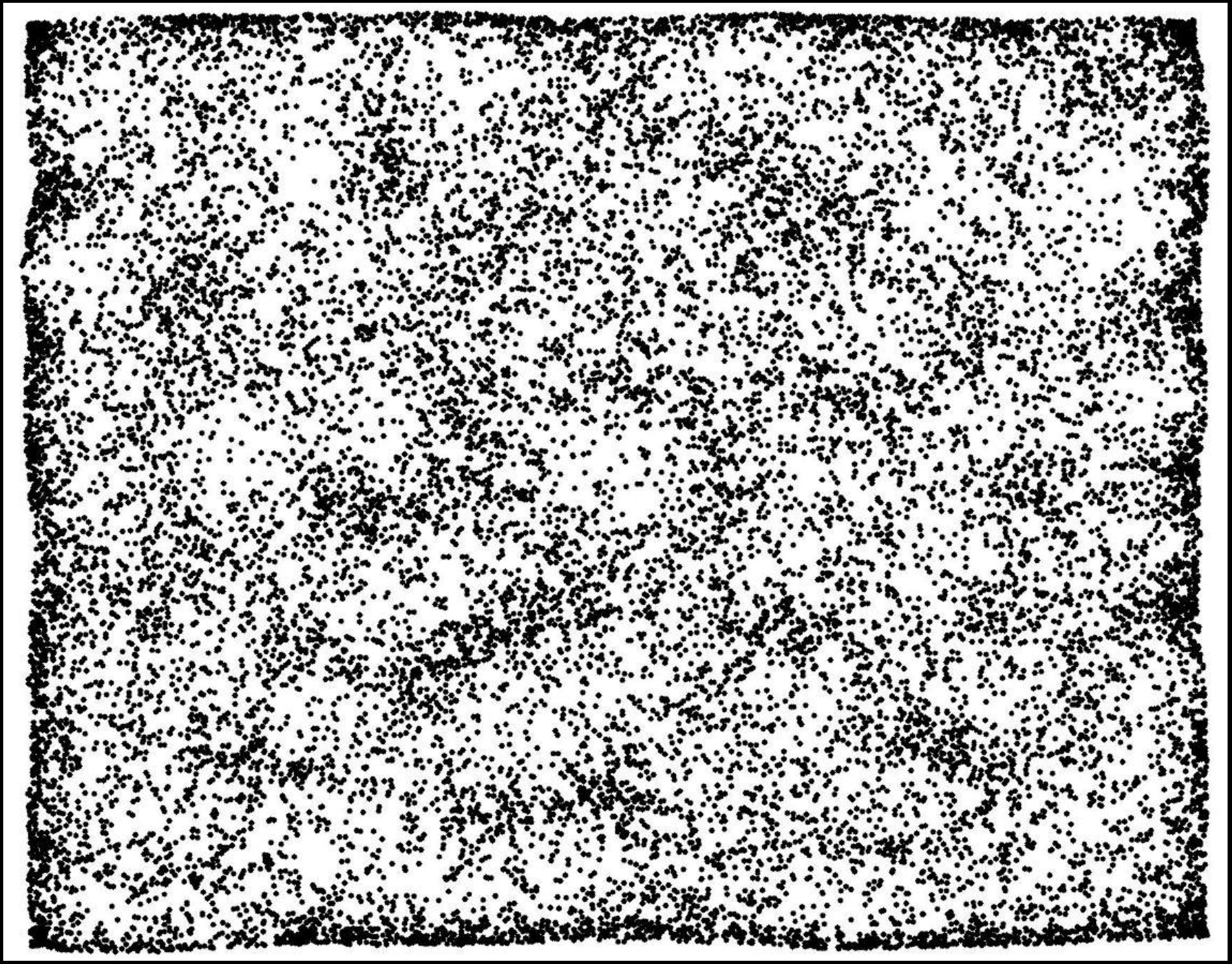


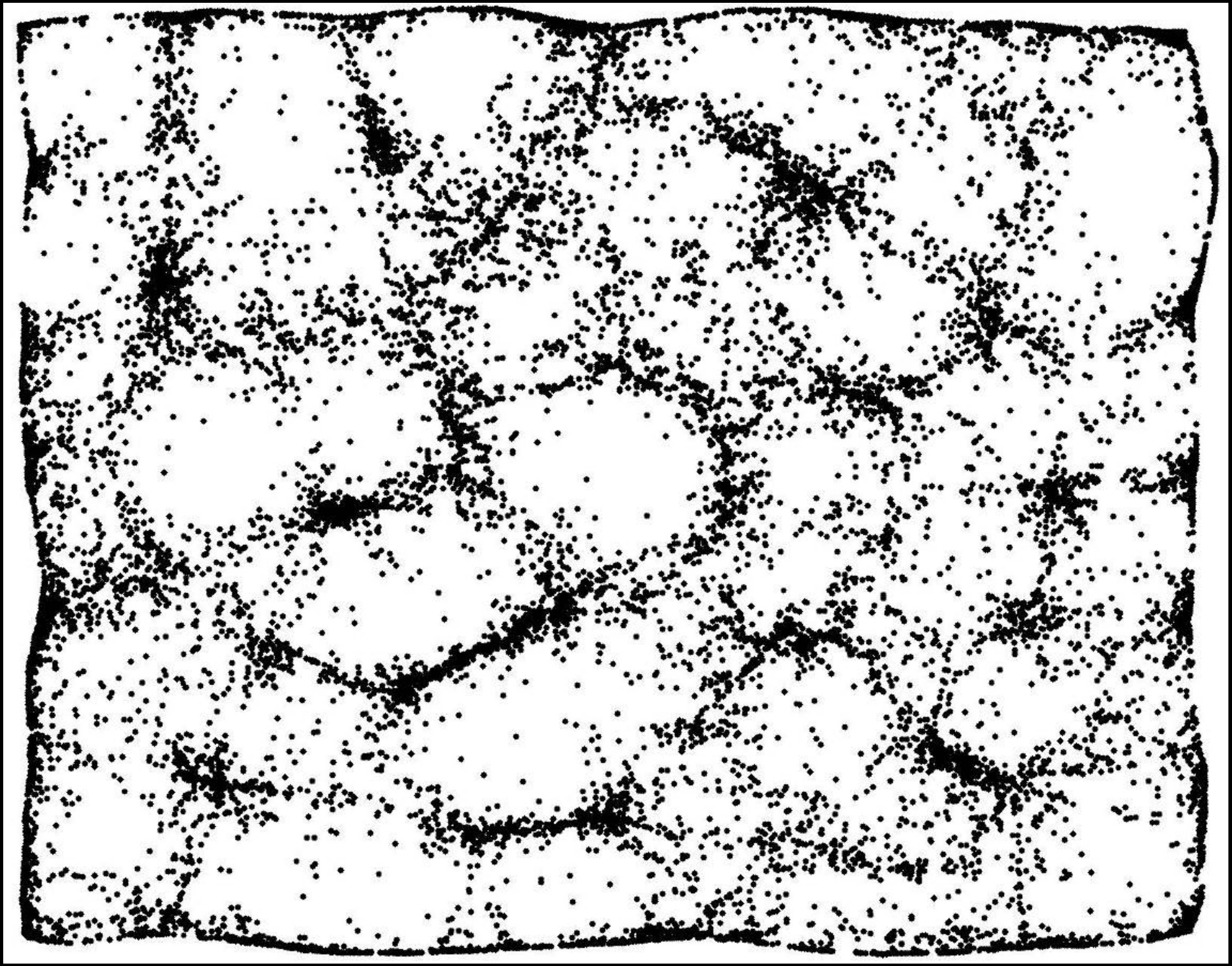


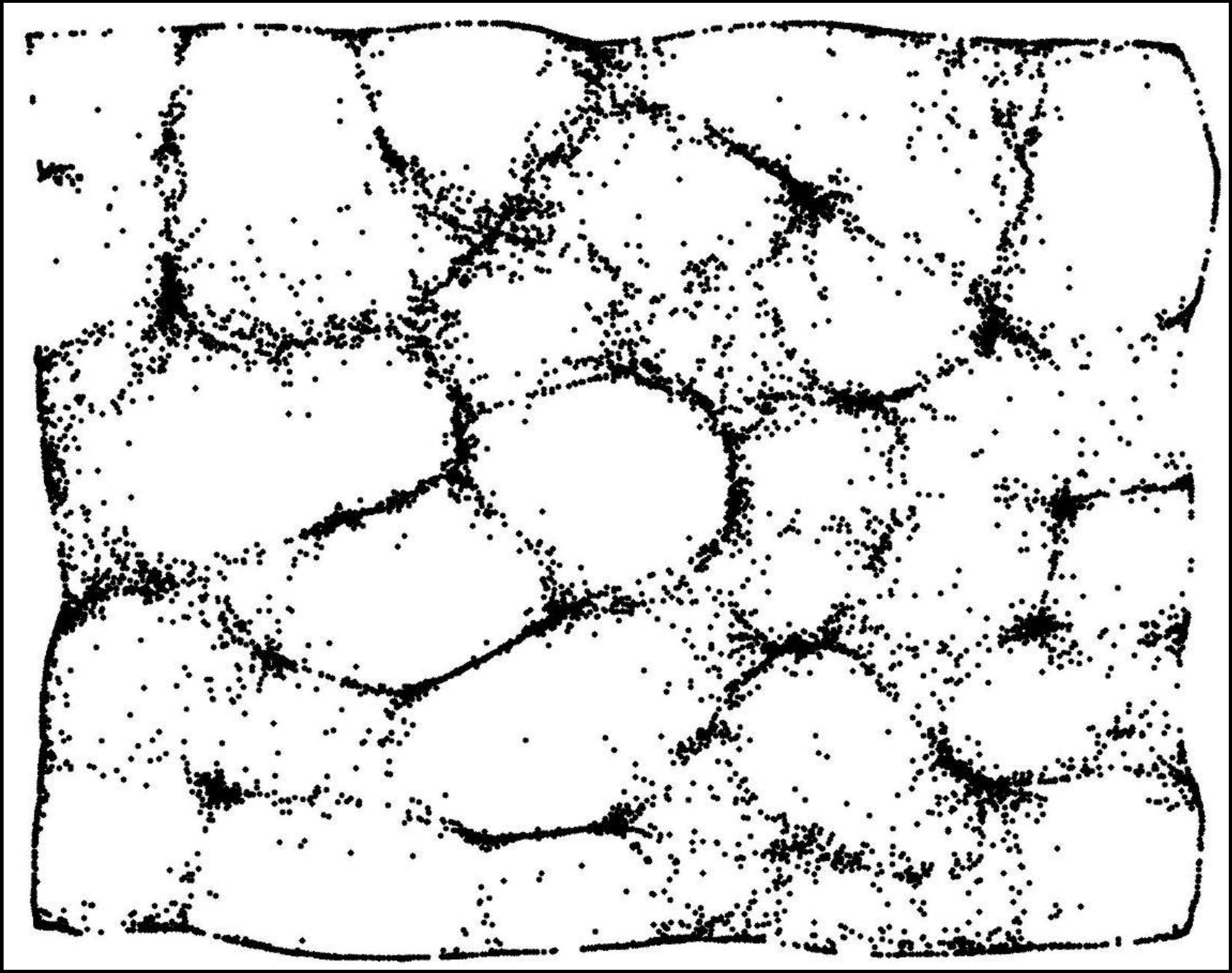
Each agent moves to mass center of neighbors

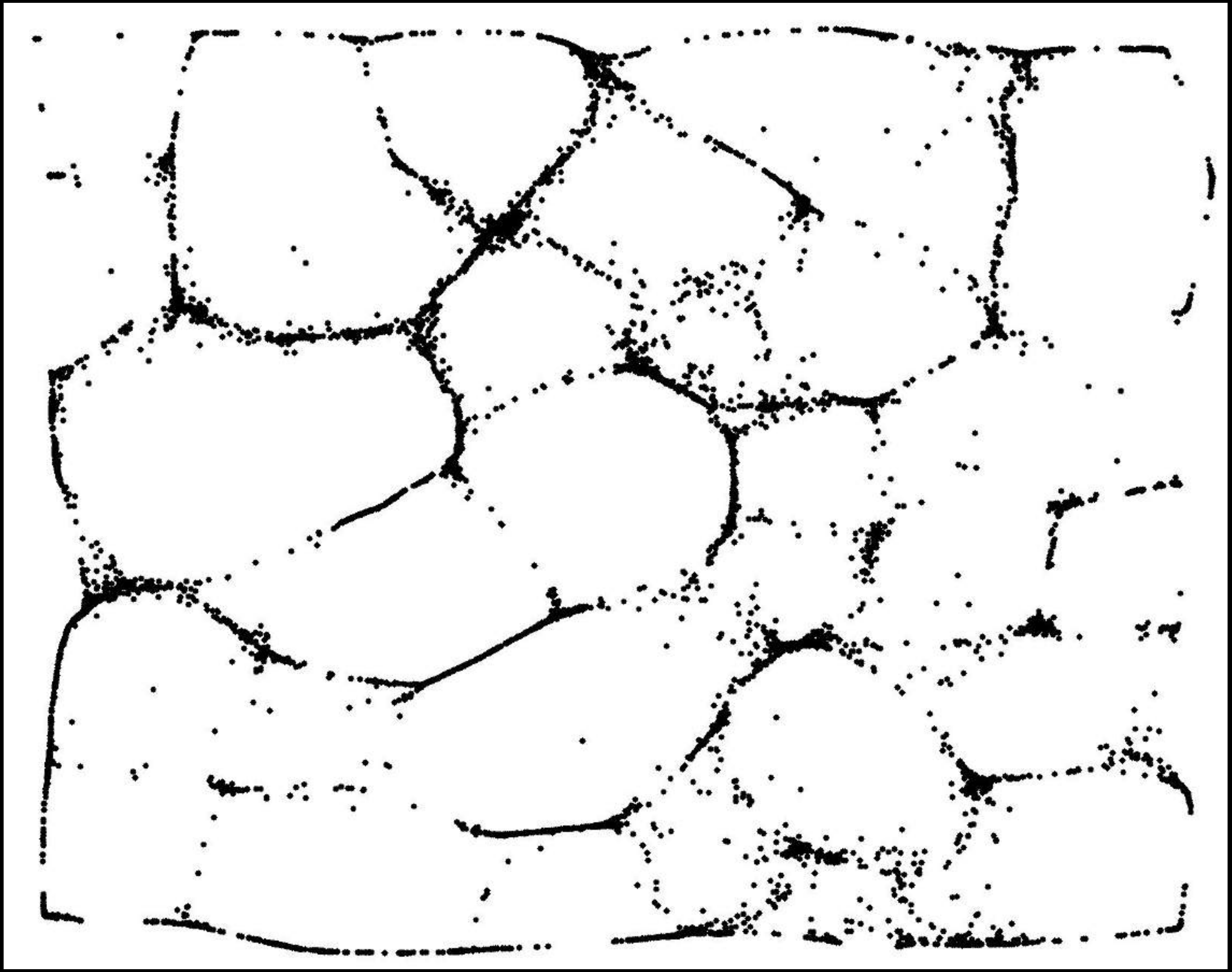


20,000 agents



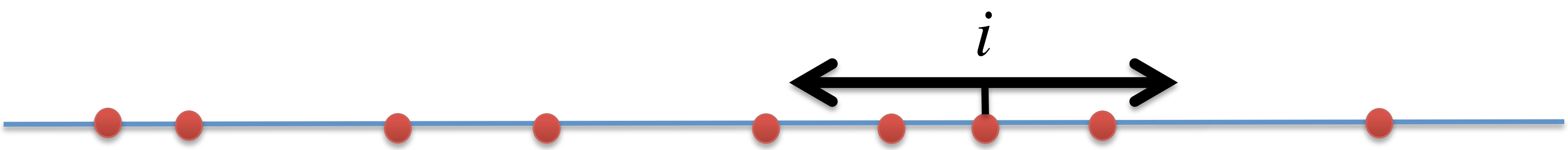






Hegselmann-Krause in 1D

$$\begin{cases} N_i(t) = \{\, j : |x_i(t) - x_j(t)| \leq R \,\} \\[2ex] x_i(t+1) = \dfrac{1}{|N_i(t)|} \displaystyle\sum_{j \in N_i(t)} x_j(t) \end{cases}$$



with Bhattacharyya, Braverman, Nguyen (2013) $O(n^3)$

Wedin-Hegarty (2015) $\Omega(n^2)$

Charron-Bost, Függer, Nowak (2015) – on circle

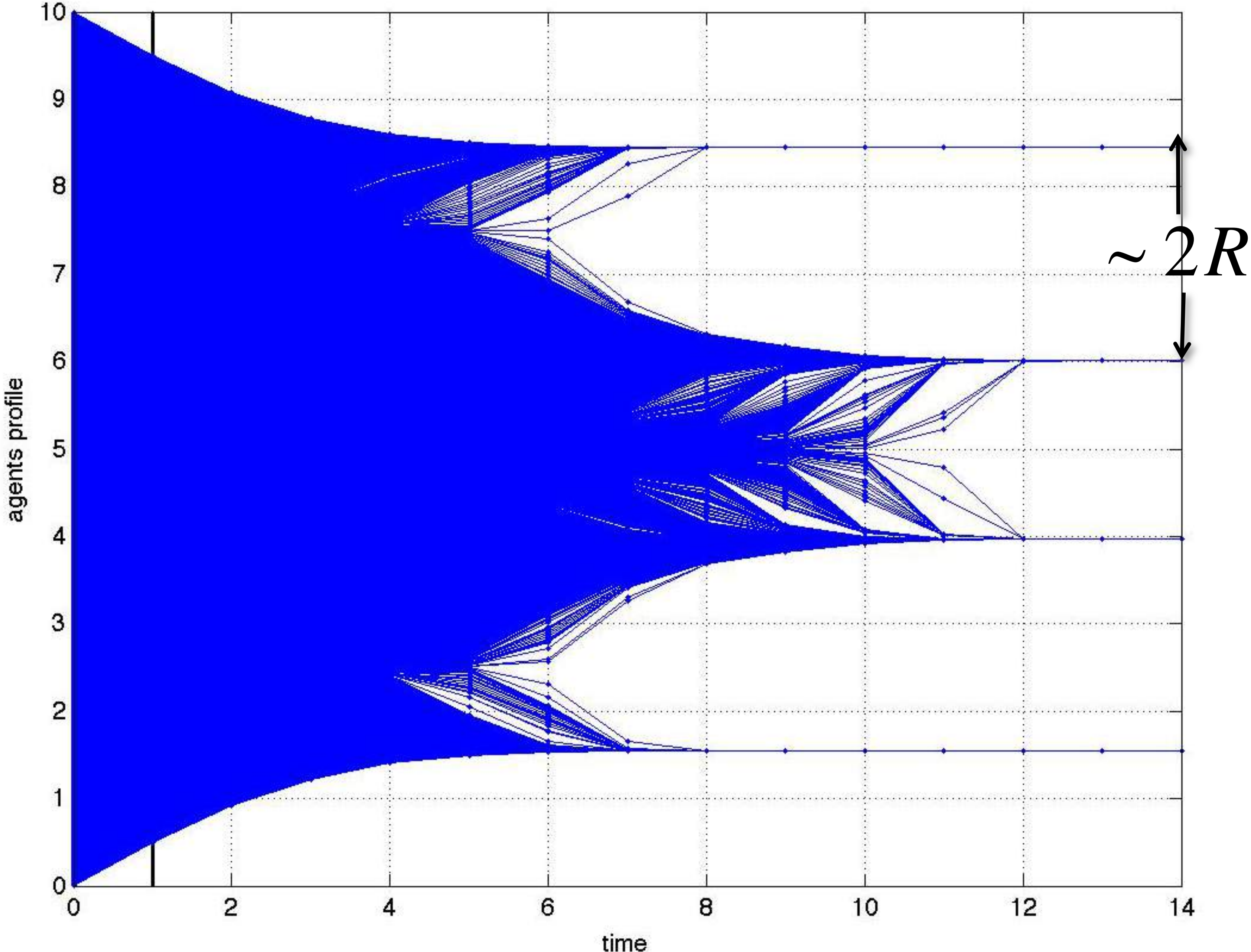
2R-Conjecture

Blondel, Hendrickx, Tsitsiklis (2007)

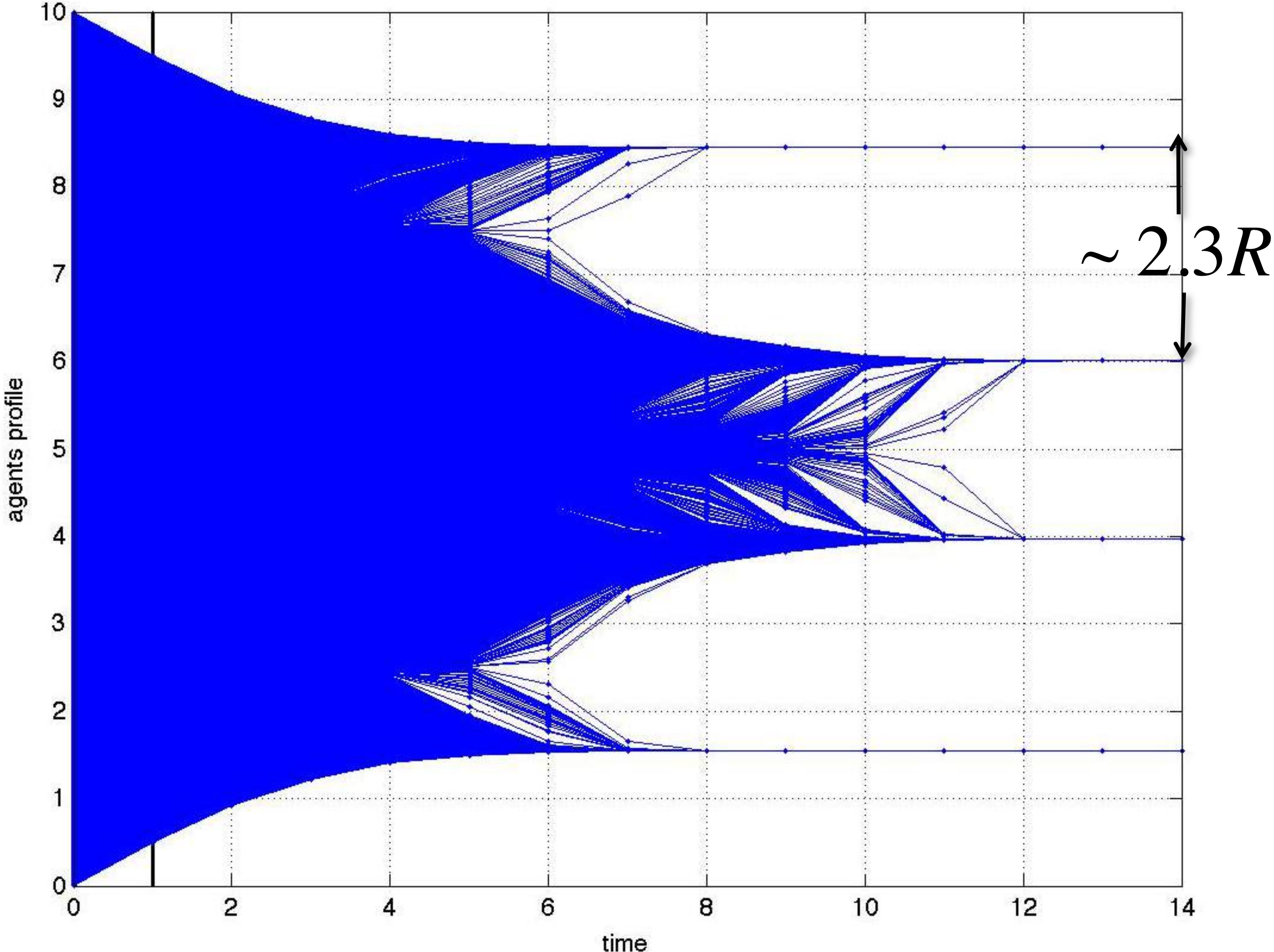
$$\begin{cases} N_i(t) = \left\{ j : | x_i(t) - x_j(t) | \leq R \right\} \\ x_i(t+1) = \dfrac{1}{| N_i(t) |} \displaystyle\sum_{j \in N_i(t)} x_j(t) \end{cases}$$

Random initial configuration

Final clusters are 2*R* away



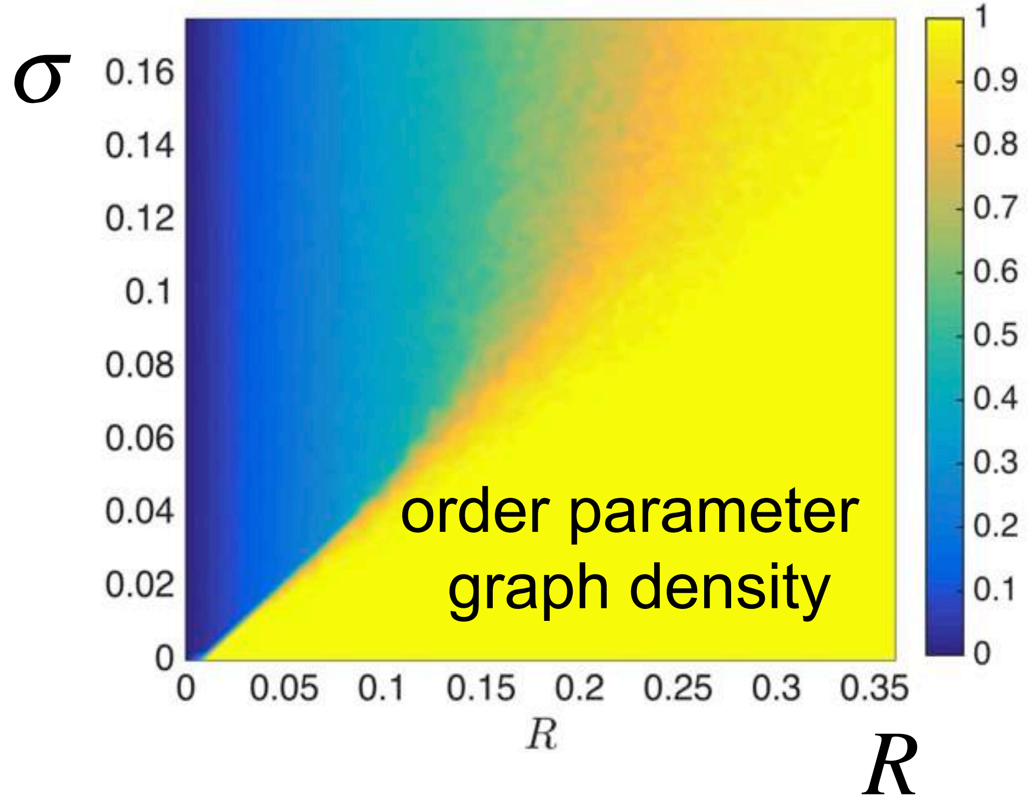
Simulation: final clusters are ~ 2.3*R* away



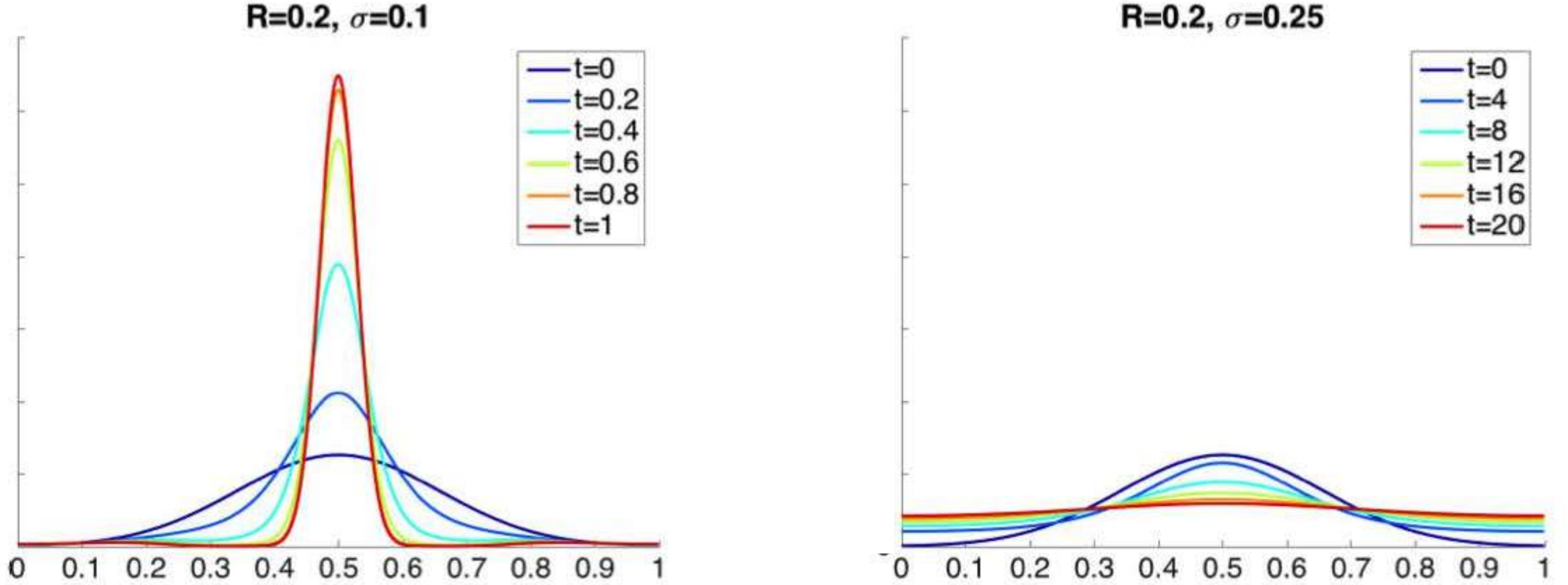
SDE :

$$dx_i = -\frac{1}{N} \sum_{j:\,|x_i - x_j| \le R} (x_i - x_j)\, dt + \sigma\, dW_t^{(i)}$$

Phase transition



Clustering vs. diffusion



Fokker-Planck PDE

Take thermodynamic limit

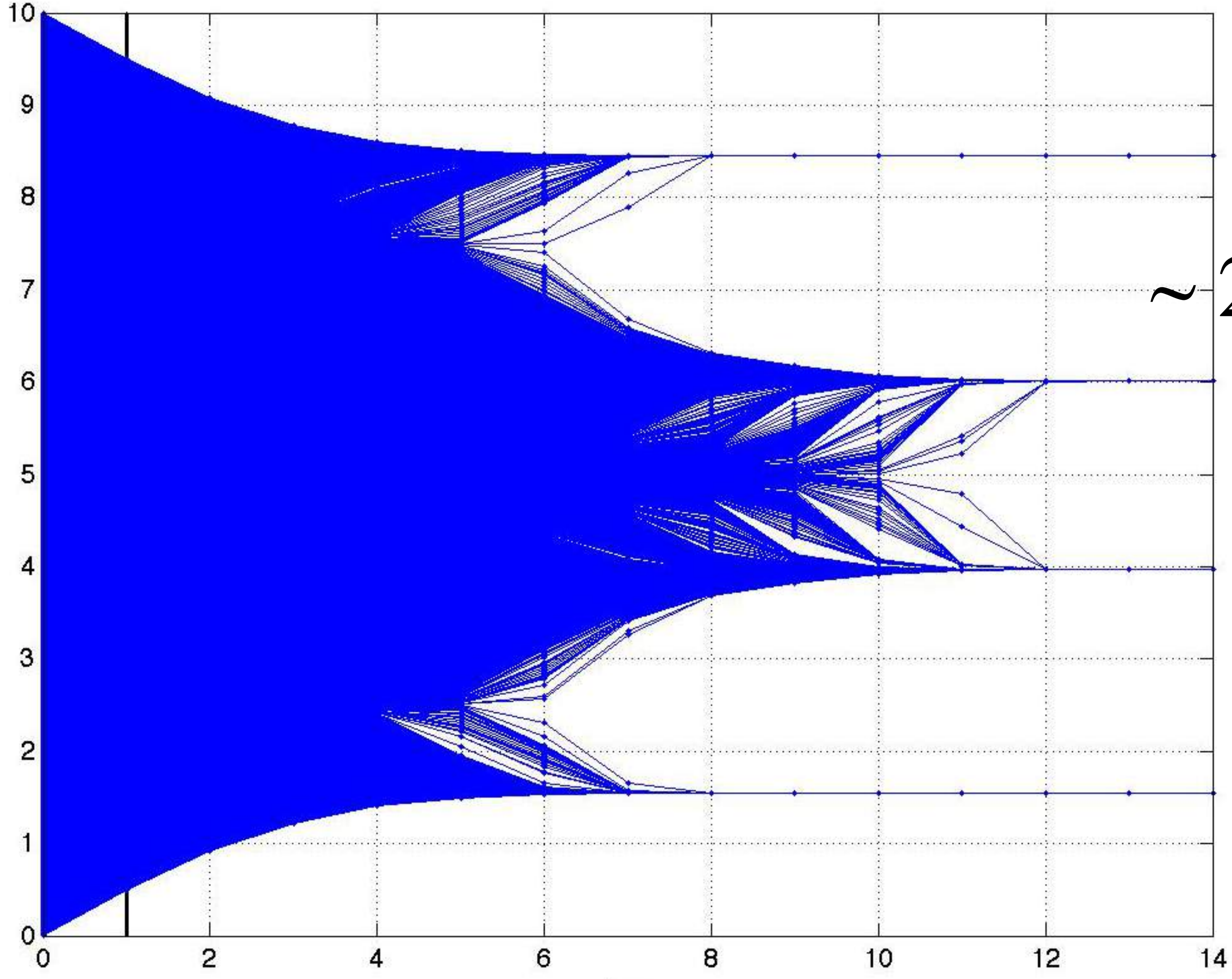
$$
\begin{cases}
\rho(x,t) := \lim_{n\to\infty} \frac{1}{n} \sum \delta_{x_j(t)}(dx) \\[2em]
\rho_t(x,t) = \left(\rho(x,t) \int (x-y)\rho(y,t)\mathbf{1}_{|y-x|\le R}\, dy \right)_x \\
\qquad\qquad + \frac{\sigma^2}{2} \rho_{xx}(x,t)
\end{cases}
$$

J. Garnier, G. Papanicolaou, T-W. Yang (2015)

Fokker-Planck PDE

$$\begin{cases} \rho(x,t) := \lim_{n\to\infty} \frac{1}{n} \sum \delta_{x_j(t)}(dx) \\ \rho_t(x,t) = \left(\rho(x,t) \int (x-y)\rho(y,t)\mathbf{1}_{|y-x|\le R})\,dy \right)_x \\ \qquad + \frac{\sigma^2}{2}\rho_{xx}(x,t) \end{cases}$$

with Q. Jiu, Q. Li, C. Wang (2015) *Well-posed*



$\sim 2.29R$

with Q. Li, Weinan, E., C. Wang (2015)

Perturbation Method

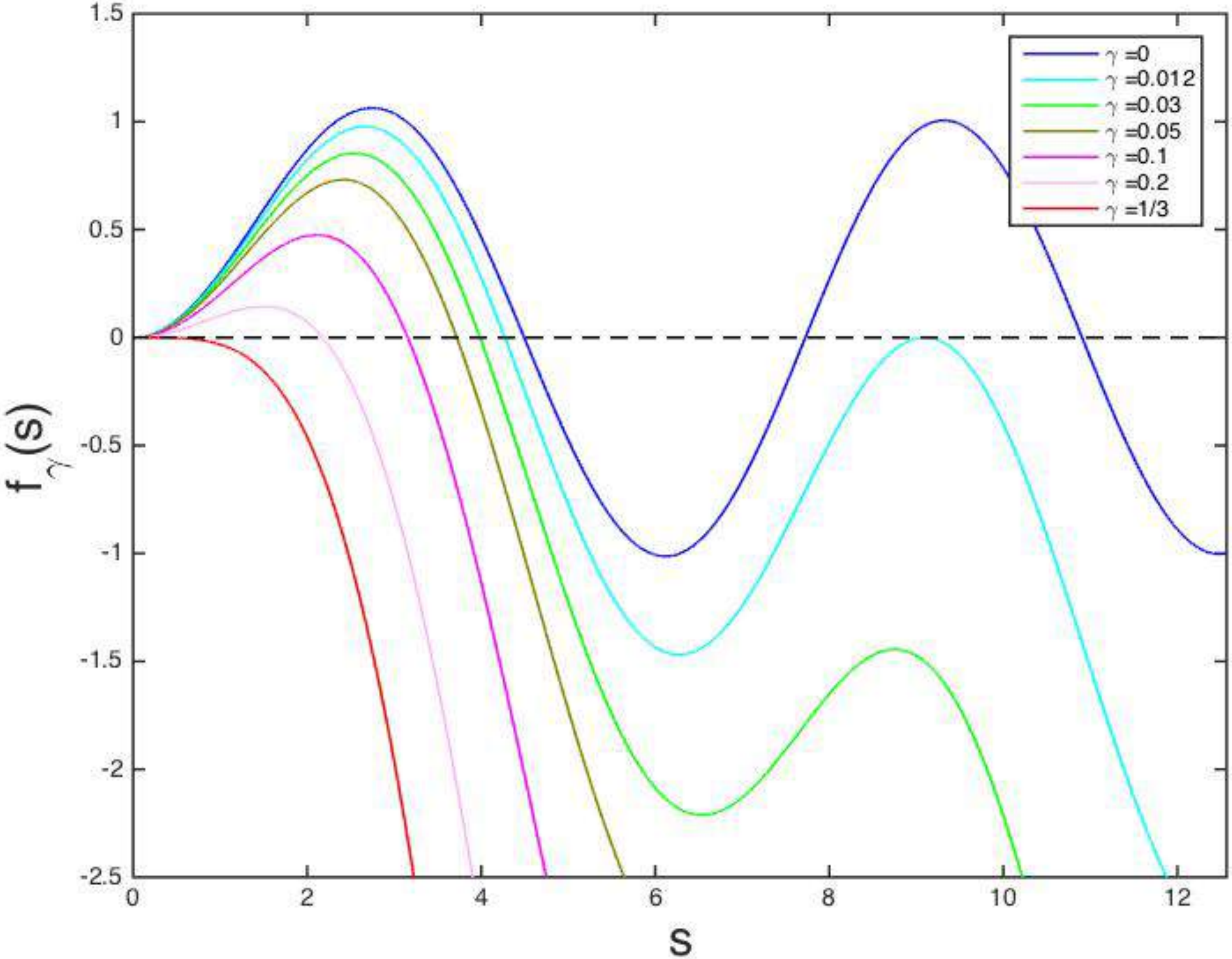
Ansatz $$\rho(x,t) = 1 + p(t)e^{2\pi ikx}$$

$$\frac{dp}{dt} = 2pRf_\gamma(s)$$

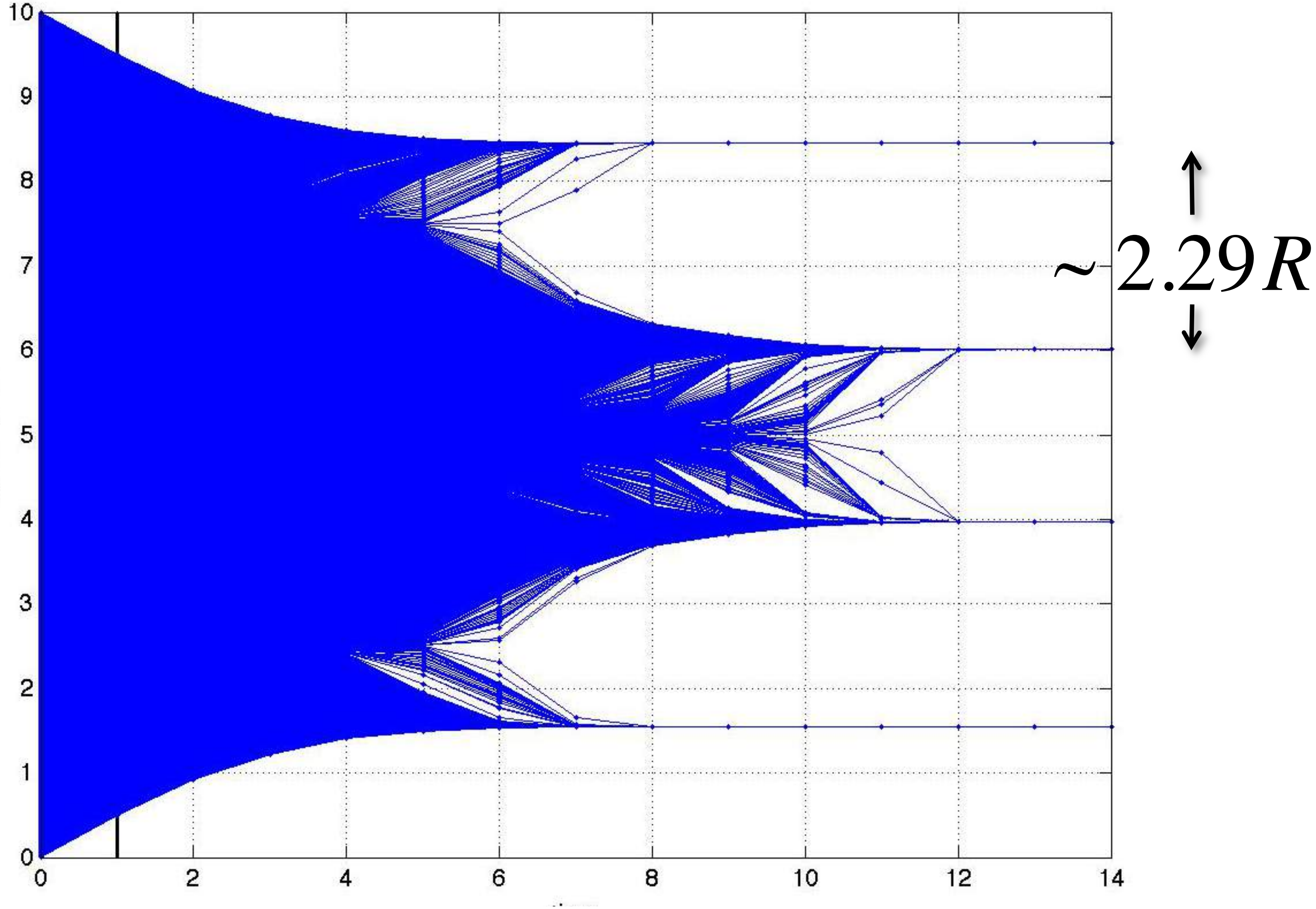
$$f_\gamma(s) = \frac{\sin s}{s} - \cos s - \gamma s^2$$

$$s = 2\pi kR; \quad \gamma = \sigma^2 / 4R^2$$

Intercluster distance $\frac{2\pi R}{s}$, *where* $\frac{df_0(s)}{ds} = 0$



γ =0
γ =0.012
γ =0.03
γ =0.05
γ =0.1
γ =0.2
γ =1/3
1.5
1
0.5
0
-0.5
-1
-1.5
-2
-2.5
0
2
4
6
8
10
12
f_γ(s)
s



10
9
8
7
6
5
4
3
2
1
0
0
2
4
6
8
10
12
14
~2.29 R

Curse of mid-dimensionality

low-dim

dyn. sys.

calculus

mid-dim

natural algorithms

dynamic graphs

high-dim

stat. mech.

universality

Merci !