Cold Spring Harbor Laboratory
MEETINGS & COURSES PROGRAM

Unsupervised Feature Discovery by Neural Networks with Disynaptic Recurrent Inhibition

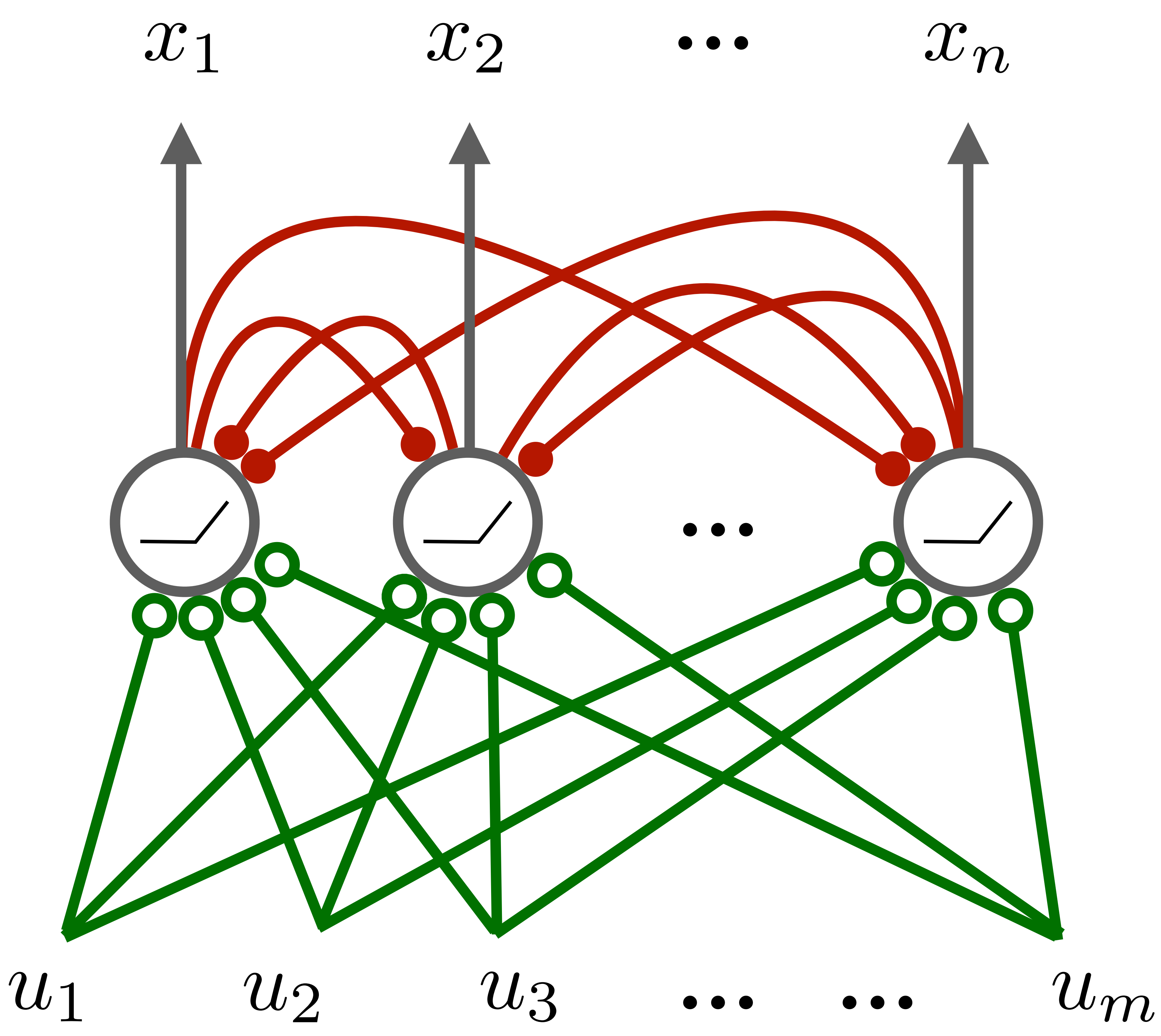
"Tony" Runzhe Yang, Kyle Luther, Sebastian Seung

Nov. 10, 2020

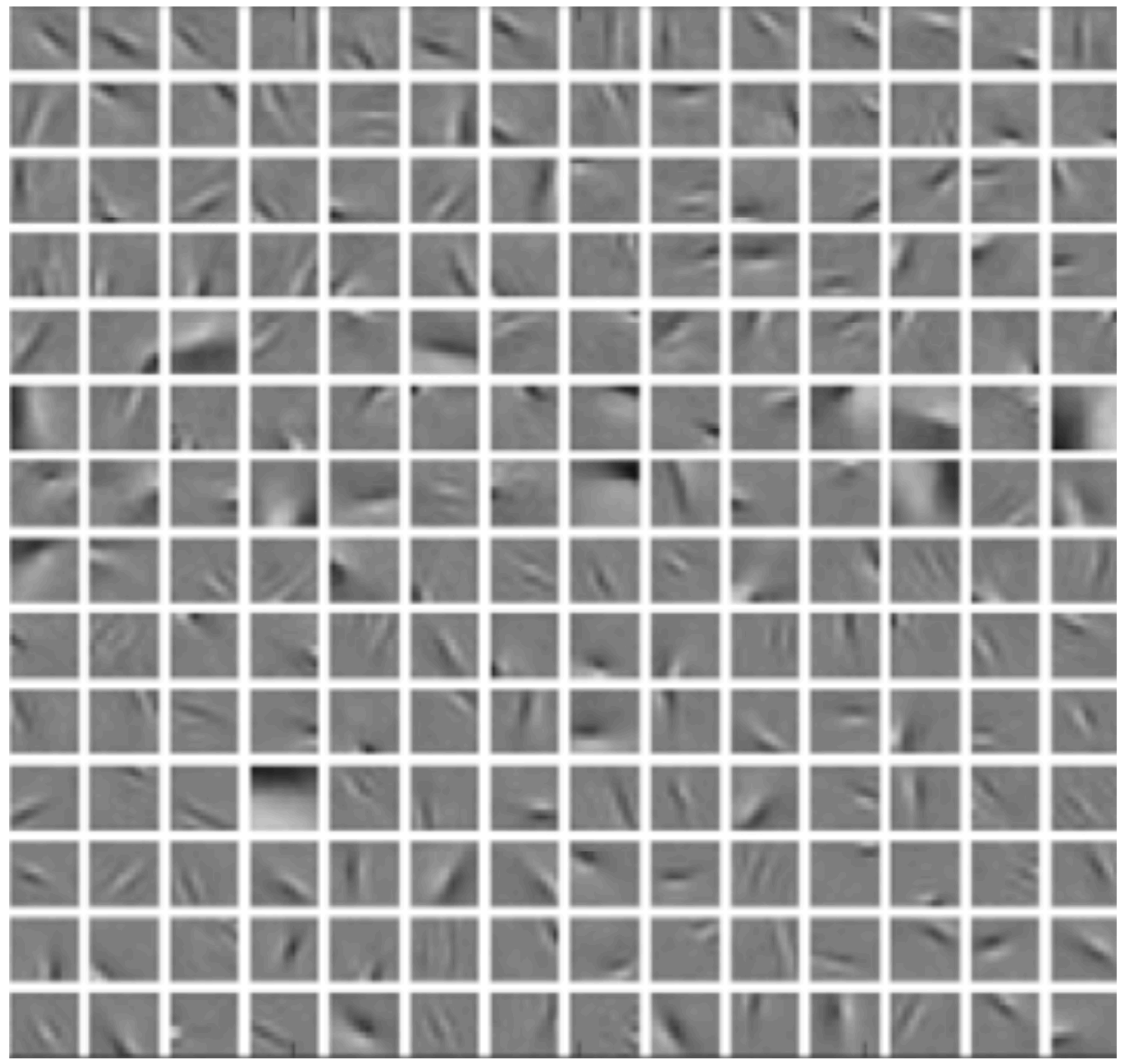
PRINCETON UNIVERSITY

How to learn diverse features?

Neural Networks with All-to-All Anti-Hebbian Inhibition Are Common



- **All-to-All Net** [Földiák 1990; Pehlevan and Chklovskii, 2014; Hu et al, 2014; Seung and Zung, 2017] **:** Neurons directly inhibits each other to encourage feature diversity.



[Hu et al, 2014]

All-to-all inhibition is not biologically plausible.

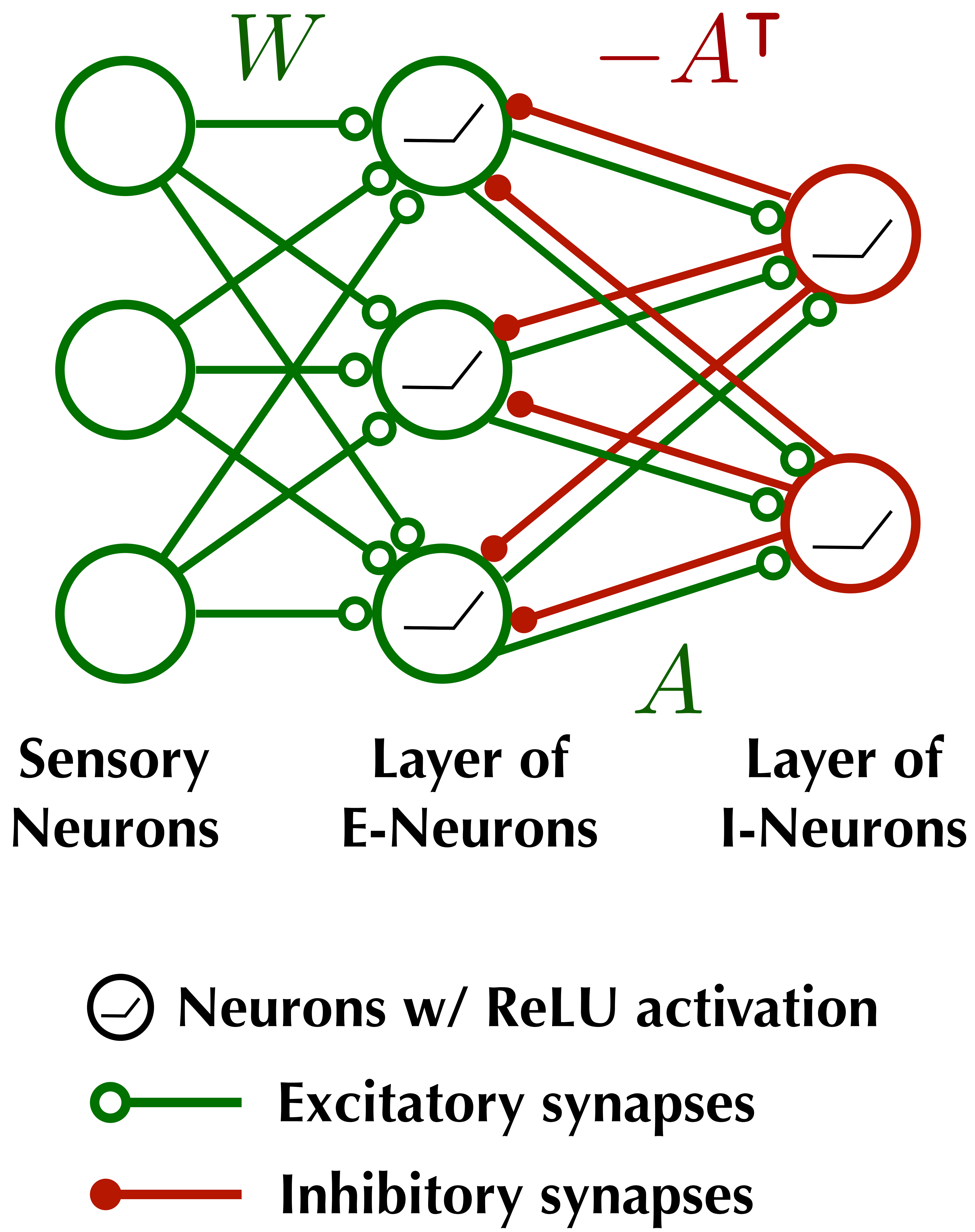
Can we learn diverse features with only a few inhibitory neurons?

Today we will show

- A model with **a few inhibitory neurons** that learns **diverse features**.
- Brain-inspired, **biologically plausible**, **unsupervised** learning.
- Explore potential application to a **language task**.

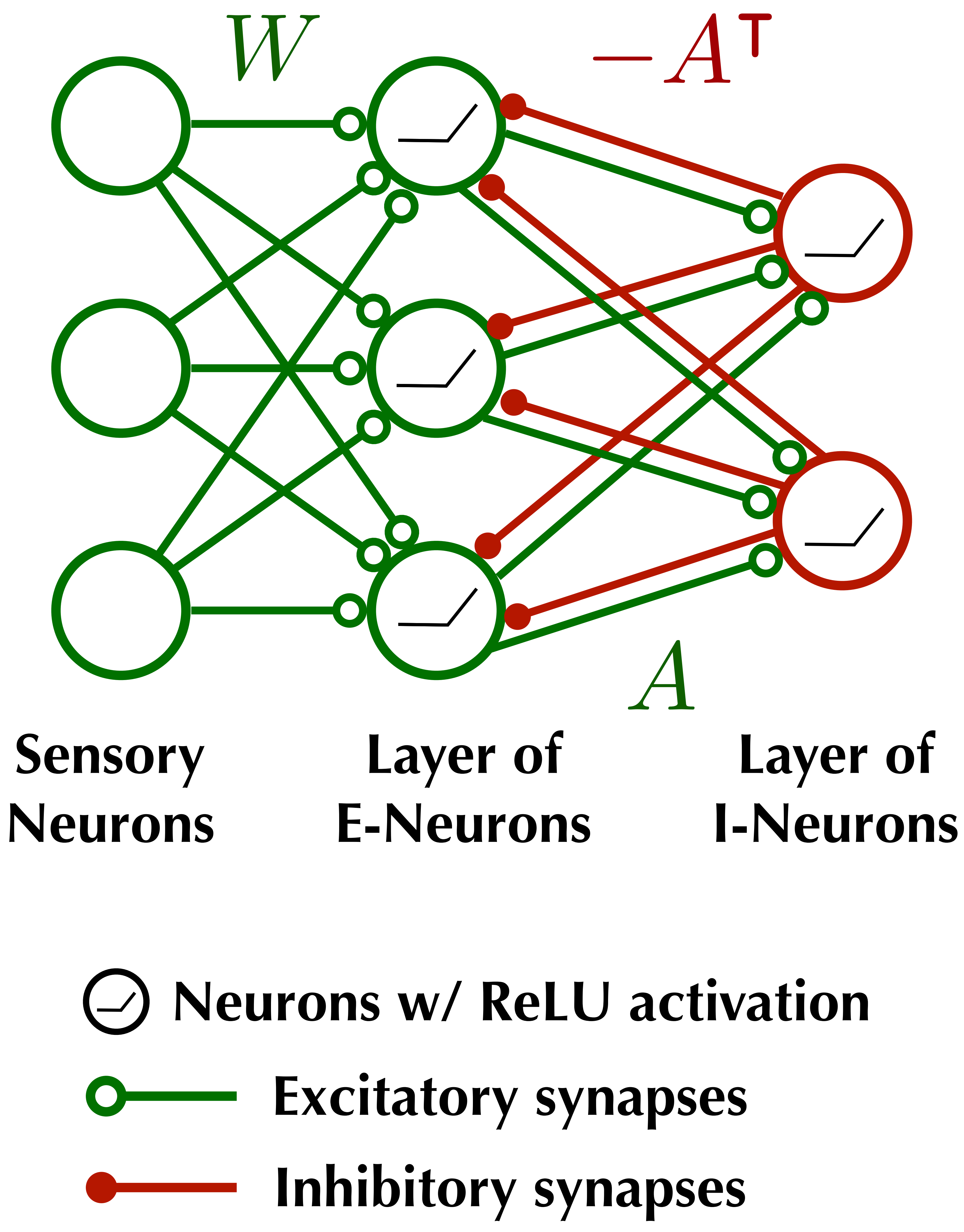
A Brain-Inspired Architecture with Disynaptic Recurrent Inhibition

[Seung, 2019]



A Brain-Inspired Architecture with Disynaptic Recurrent Inhibition

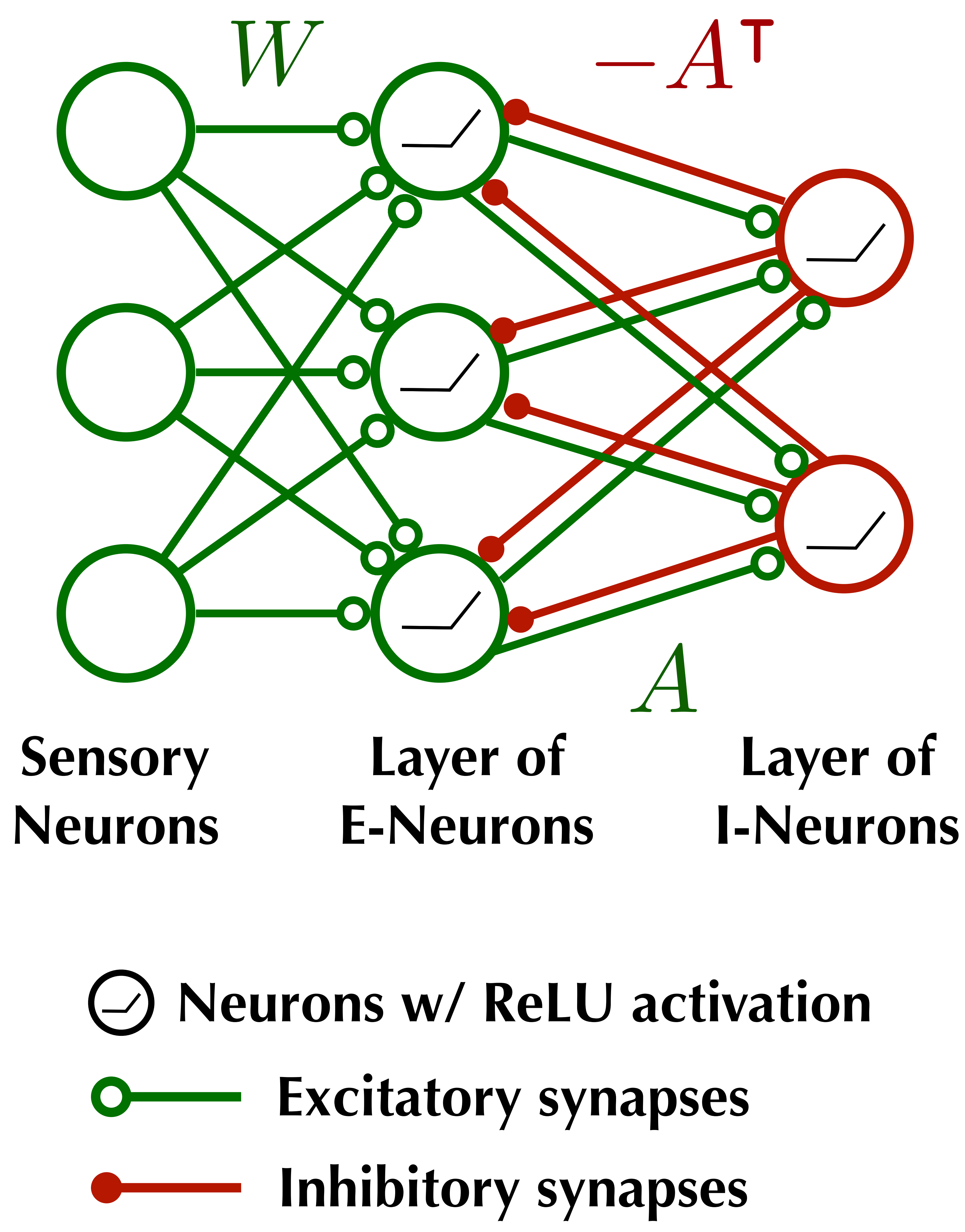
[Seung, 2019]



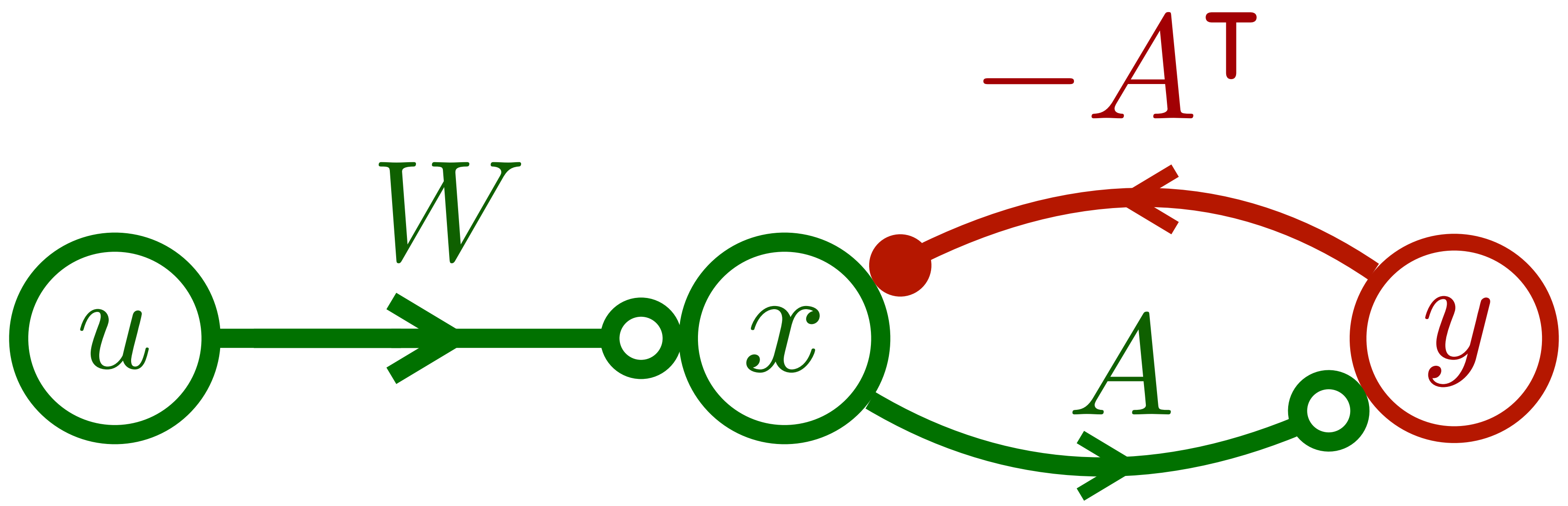
- **"Dale's Law"** [Eccles, 1954] : signs of outgoing synaptic weights of a neuron are either non-negative or non-positive.

A Brain-Inspired Architecture with Disynaptic Recurrent Inhibition

[Seung, 2019]

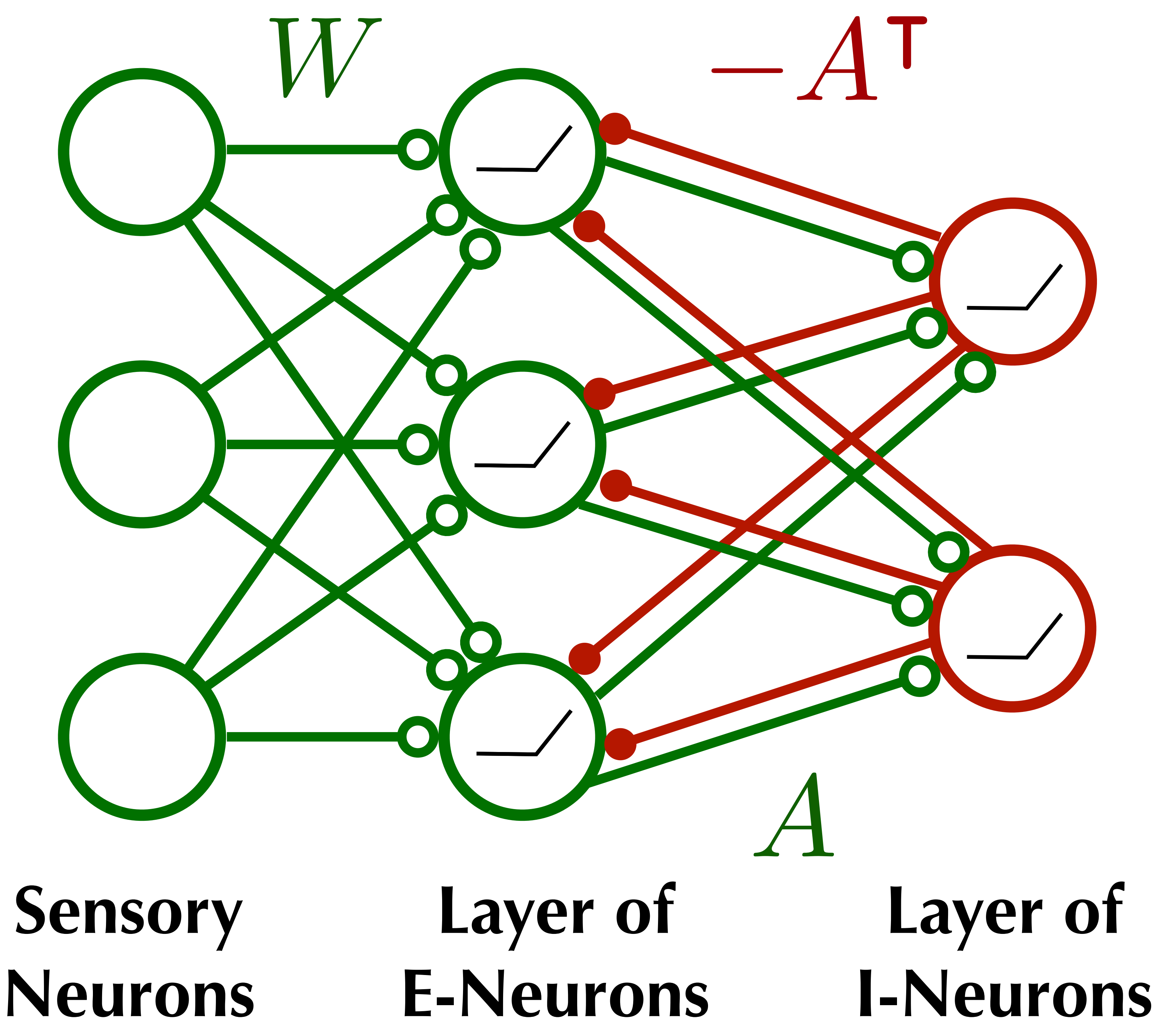


- **"Dale's Law"** [Eccles, 1954] : signs of outgoing synaptic weights of a neuron are either non-negative or non-positive.
- **Feedforward excitation + anti-symmetric reciprocal excitatory-inhibitory connections** [Znamenskiy et al., 2018]



A Brain-Inspired Architecture with Disynaptic Recurrent Inhibition

[Seung, 2019]

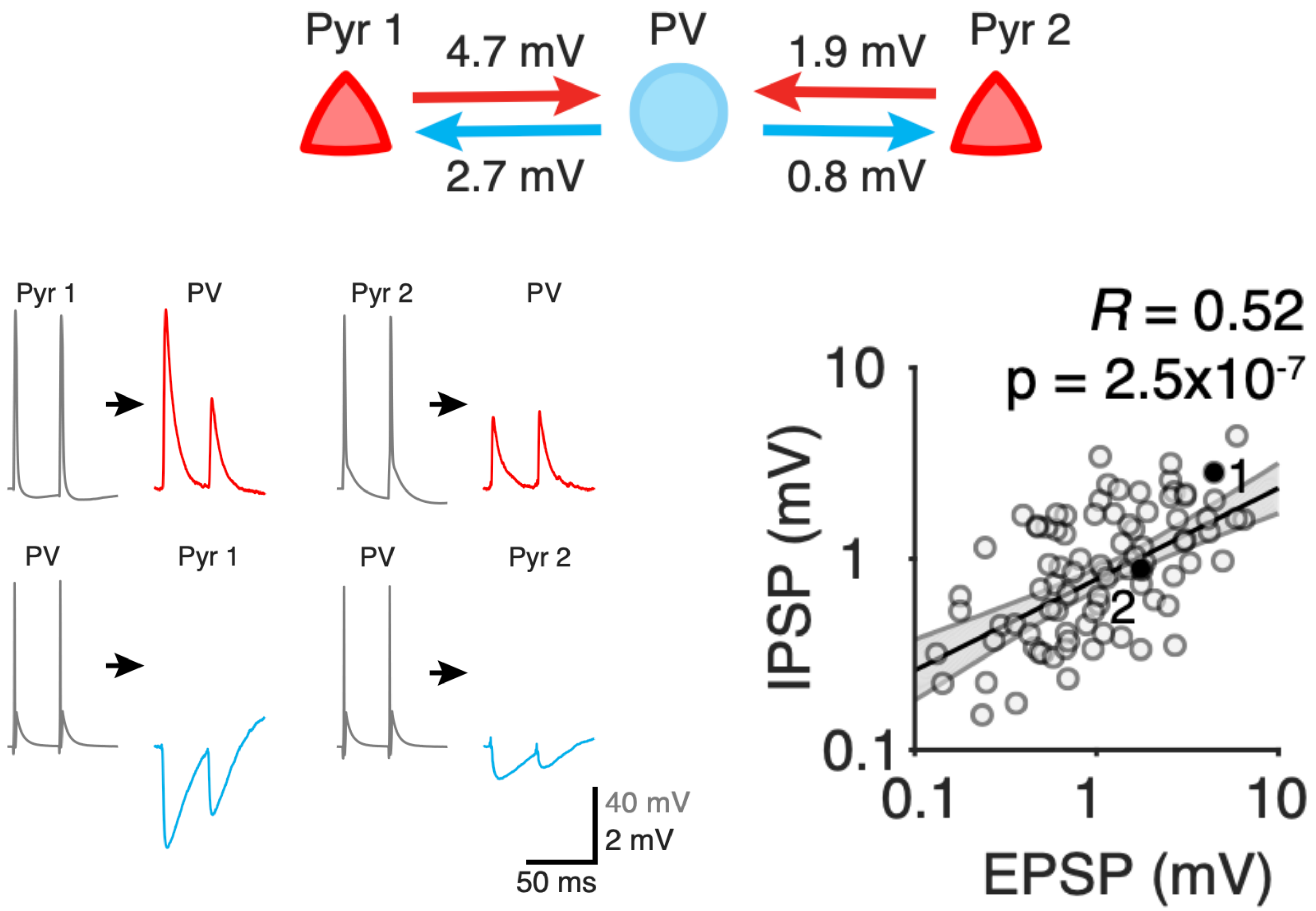


Neurons w/ ReLU activation

Excitatory synapses

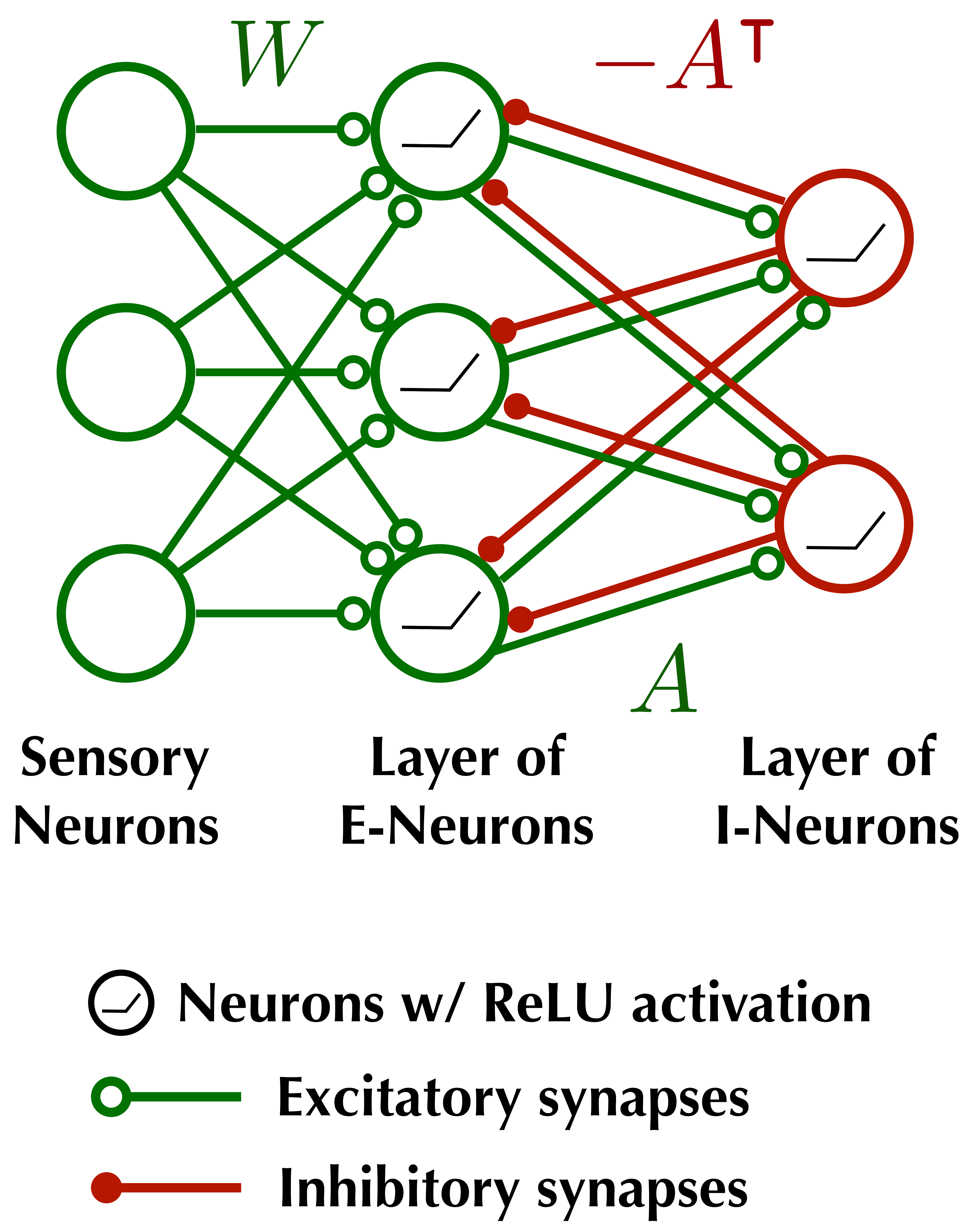
Inhibitory synapses

- **“Dale’s Law”** [Eccles, 1954] : signs of outgoing synaptic weights of a neuron are either non-negative or non-positive.
- **Feedforward excitation + anti-symmetric reciprocal excitatory-inhibitory connections** [Znamenskiy et al., 2018]



A Brain-Inspired Architecture with Disynaptic Recurrent Inhibition

[Seung, 2019]



- **"Dale's Law"** [Eccles, 1954] : signs of outgoing synaptic weights of a neuron are either non-negative or non-positive.

- **Feedforward excitation + anti-symmetric reciprocal excitatory-inhibitory connections** [Znamenskiy et al., 2018]

- Given a sensory input ***U***, compute steady neural activities ***X*** and ***Y***:

$$X_{it} = \frac{1}{\Lambda_{ii}} \left[\sum_a W_{ia} U_{at} - \sum_\alpha A_{\alpha i} Y_{\alpha t} \right]^+$$

ReLU

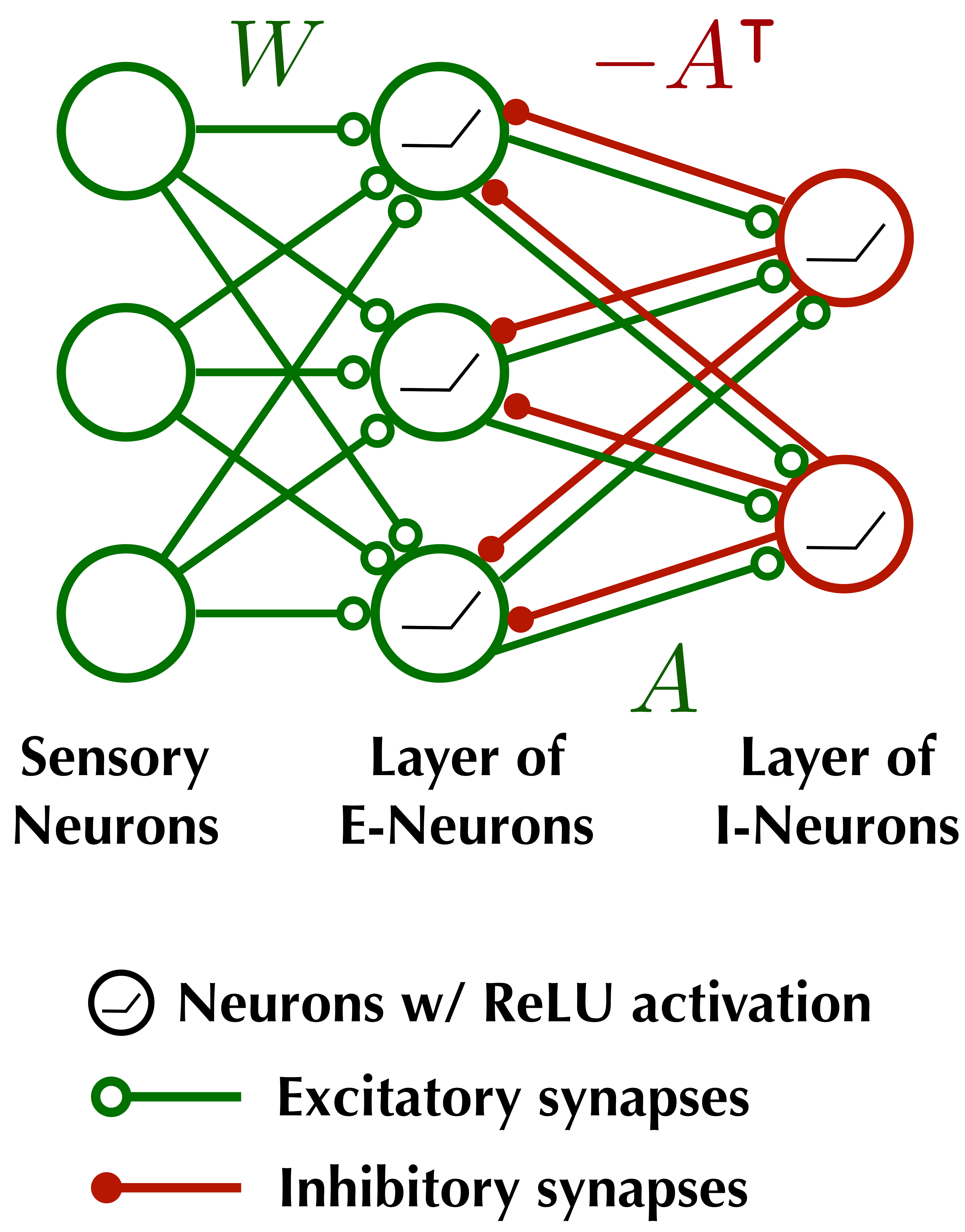
feedforward input

self-regulation / sensitivity of neuron *i* to inputs

feedback inhibition

A Brain-Inspired Architecture with Disynaptic Recurrent Inhibition

[Seung, 2019]

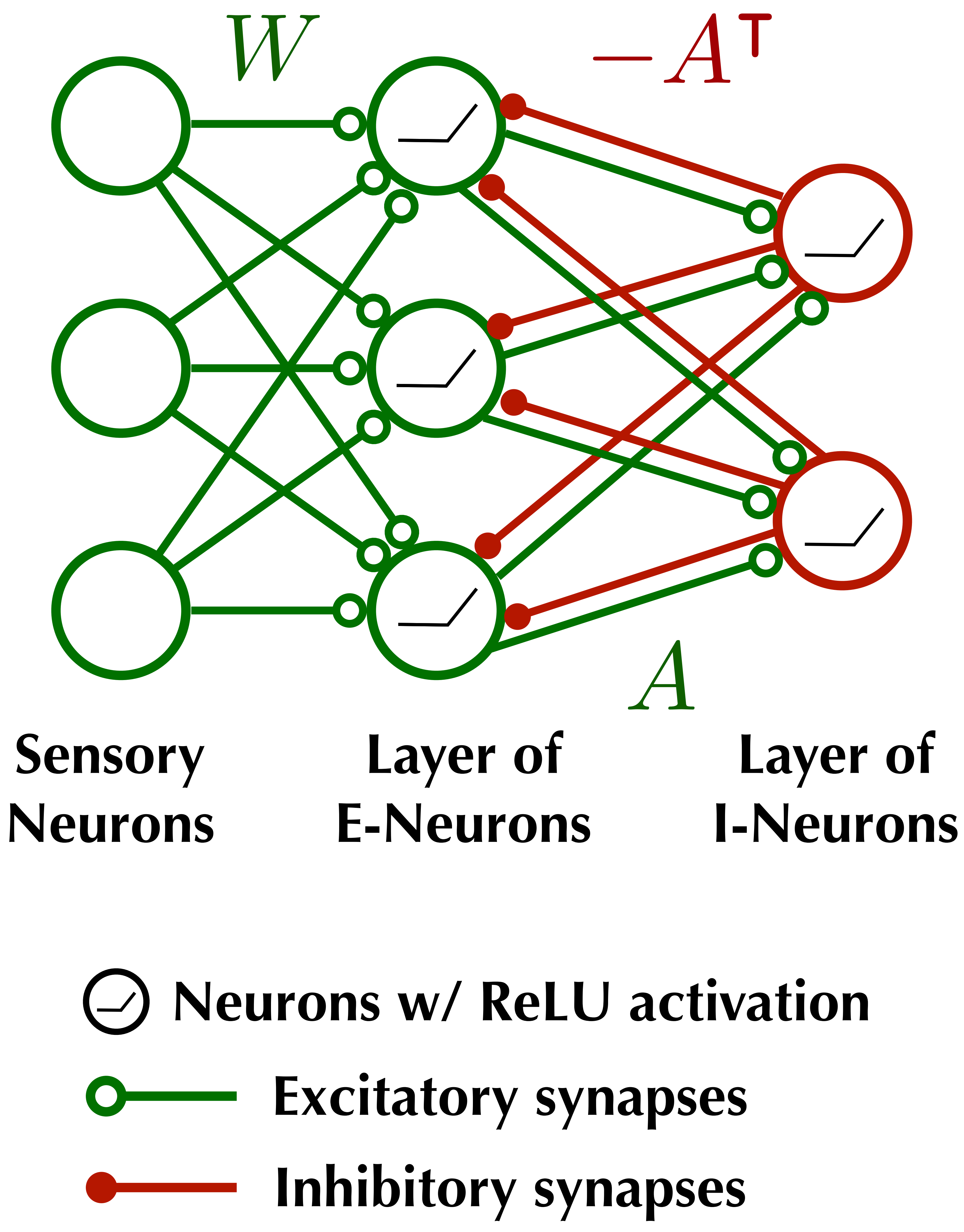


- **"Dale's Law"** [Eccles, 1954] : signs of outgoing synaptic weights of a neuron are either non-negative or non-positive.
- **Feedforward excitation + anti-symmetric reciprocal excitatory-inhibitory connections** [Znamenskiy et al., 2018]
- Given a sensory input ***U***, compute steady neural activities ***X*** and ***Y***:

feedforward excitation

$$Y_{\alpha t} = \sum_i A_{\alpha i} X_{it}$$

A Brain-Inspired Architecture with Disynaptic Recurrent Inhibition

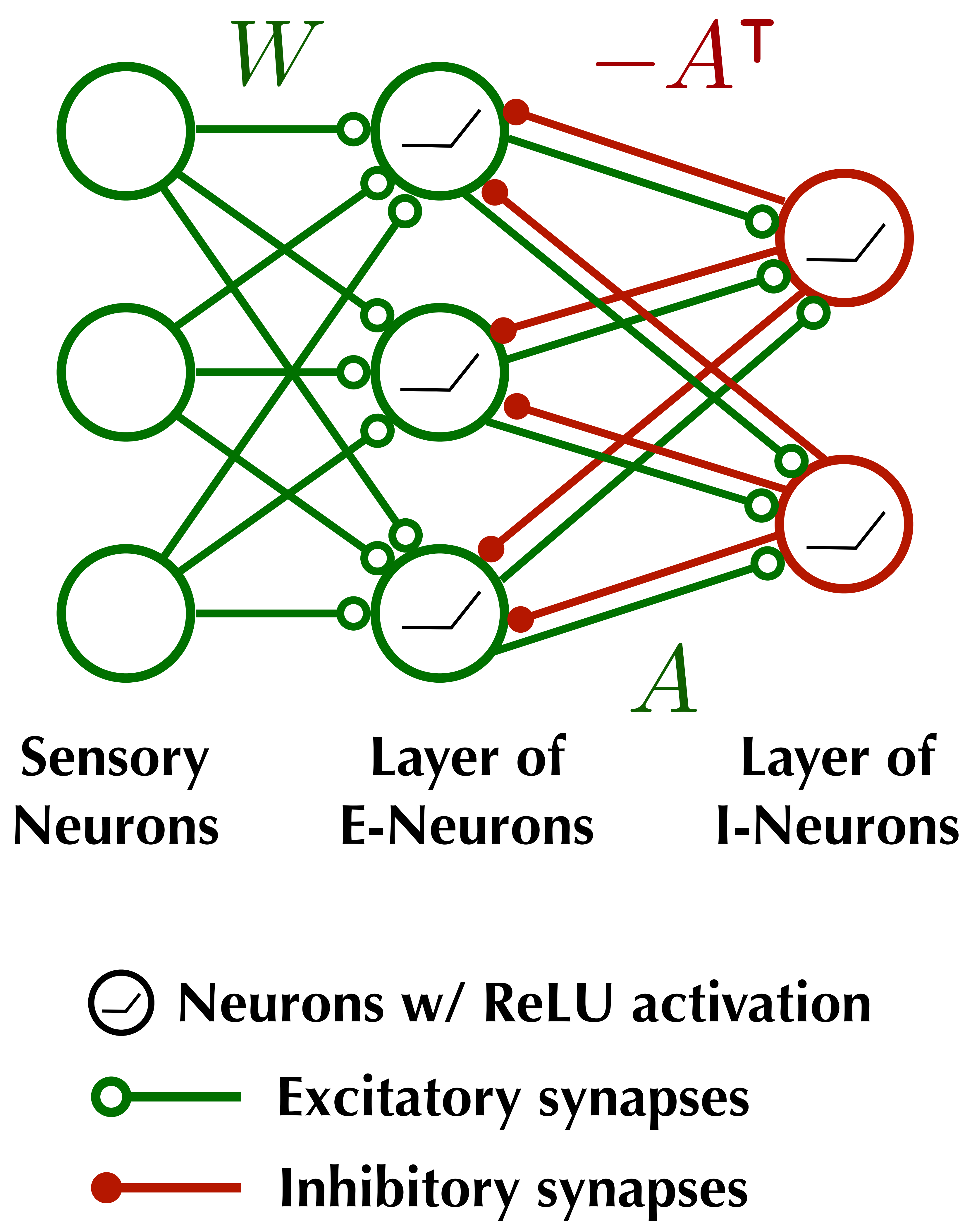


- **Local Learning Rule**: Hebbian and Anti-Hebbian Plasticity **[Földiák, 1990]**

$$\Delta W_{ia} \propto X_{it} U_{at} - \phi(W)_{ia}$$

$$\Delta A_{\alpha i} \propto Y_{\alpha t} X_{it} - \psi(A)_{\alpha i}$$

A Brain-Inspired Architecture with Disynaptic Recurrent Inhibition



- **Local Learning Rule**: Hebbian and Anti-Hebbian Plasticity [Földiák, 1990]

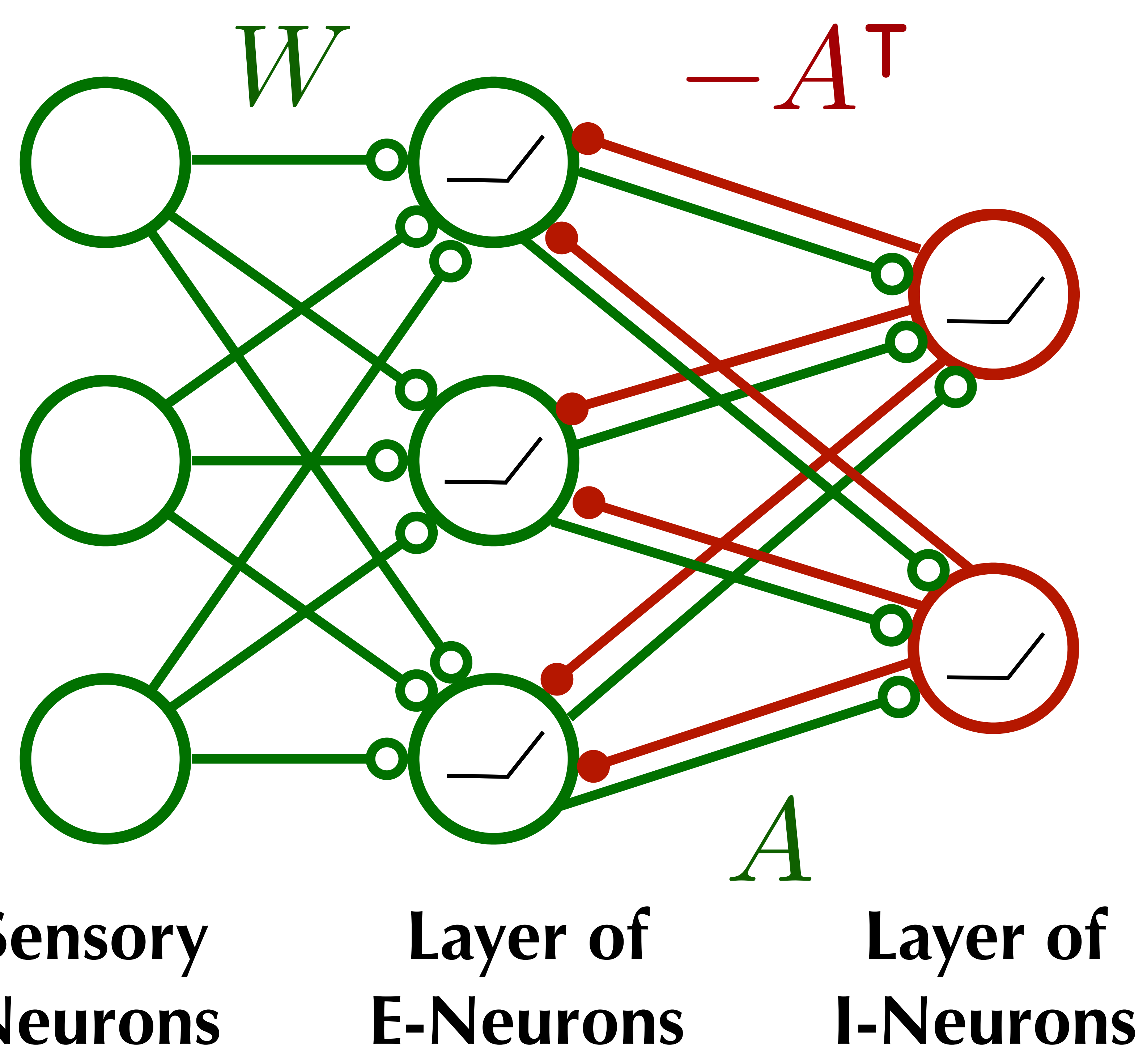
$$\Delta W_{ia} \propto X_{it} U_{at} - \phi(W)_{ia}$$

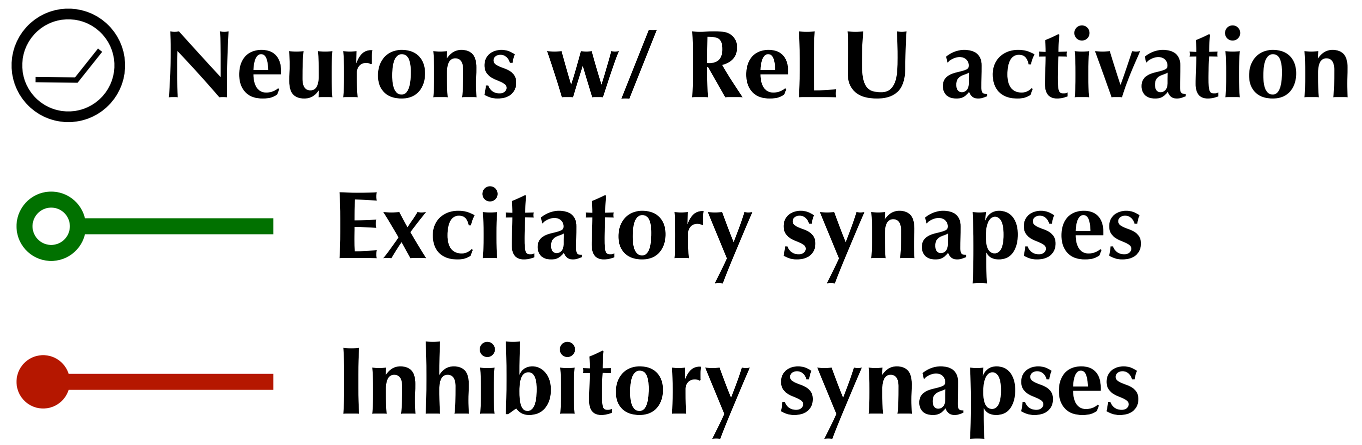
$$\Delta A_{\alpha i} \propto Y_{\alpha t} X_{it} - \psi(A)_{\alpha i}$$

Correlation

Weight Decay (explain later)

A Brain-Inspired Architecture with Disynaptic Recurrent Inhibition





- **Local Learning Rule**: Hebbian and Anti-Hebbian Plasticity [Földiák, 1990]

$$\Delta W_{ia} \propto X_{it} U_{at} - \phi(W)_{ia}$$

$$\Delta A_{\alpha i} \propto Y_{\alpha t} X_{it} - \psi(A)_{\alpha i}$$

Correlation

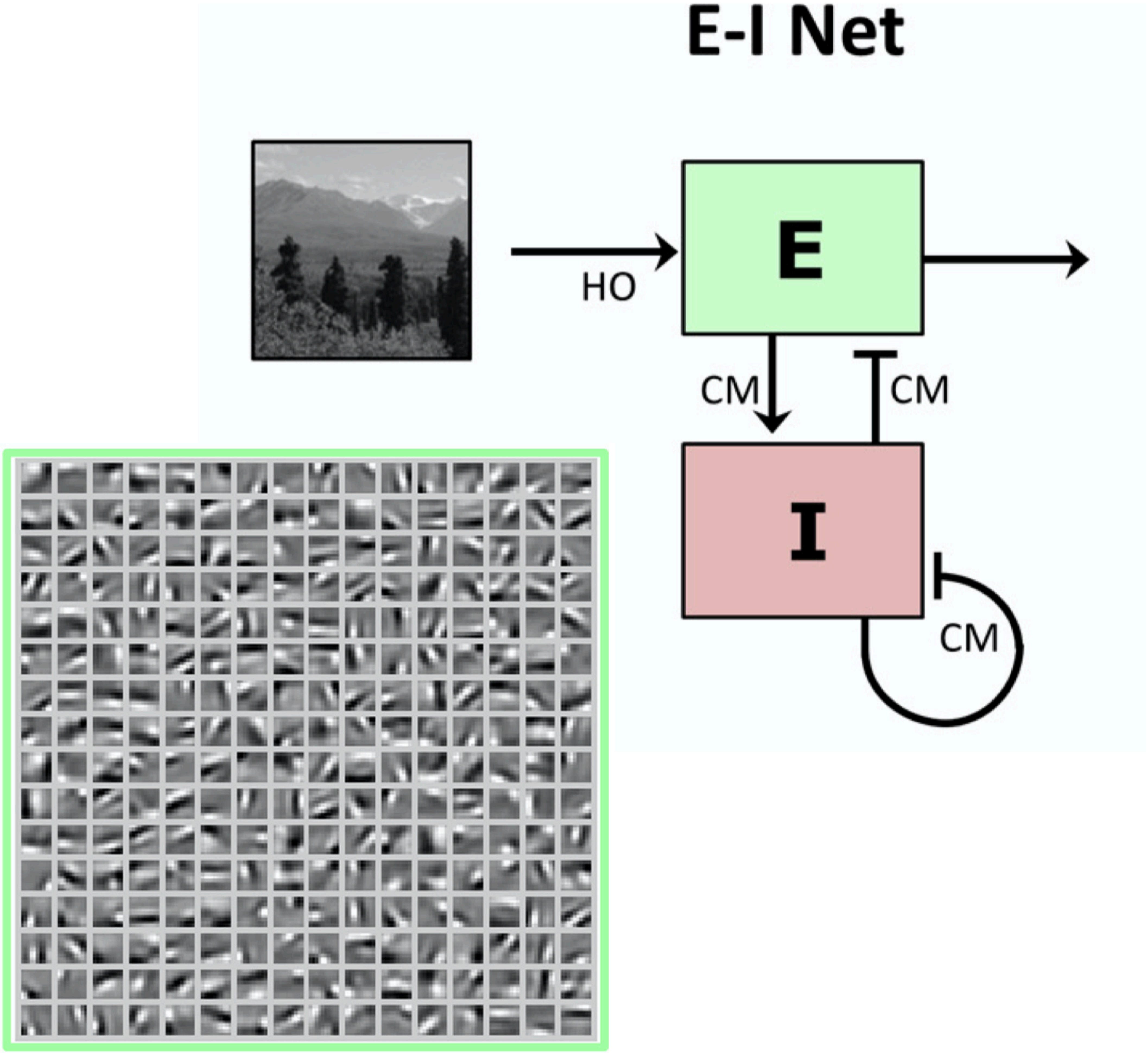
Weight Decay (explain later)

- "Effective Objective": ~ **"Softened" Correlation Game** [Luther, Yang, & Seung, 2019]:

$$\max_{X \geq 0} \left\{ \Phi^* \left(\frac{XU^\intercal}{T} \right) - \frac{1}{2} \Psi^* \left(\frac{XX^\intercal}{T} \right) \right\}$$

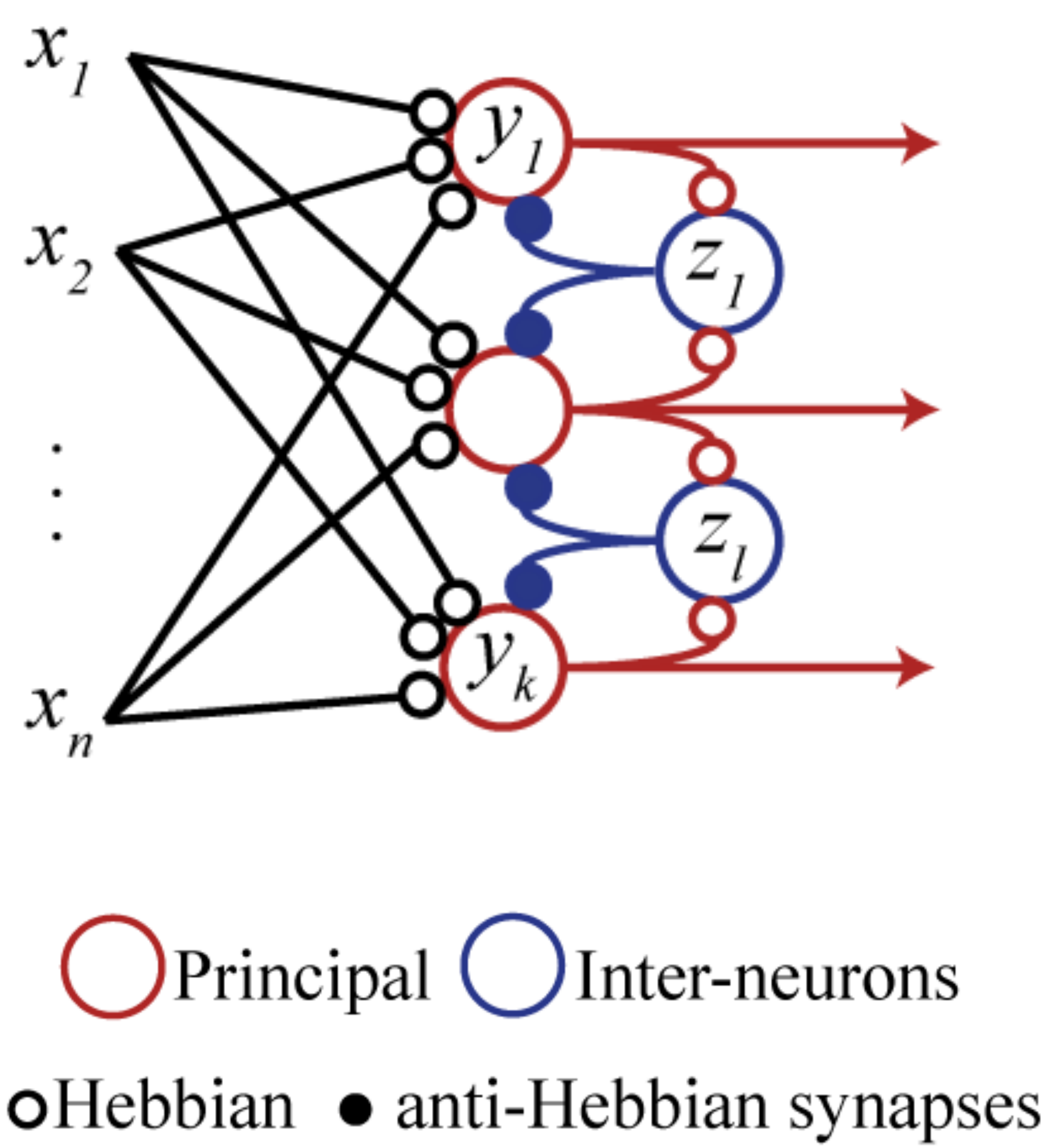
input-output correlation   **output-output correlation**

Related work



- **King el al.'s E-I Net** [King et al., 2013] **:** I-neurons decorrelate the activity of the E-neurons by suppressing redundant spiking activity.
- It's a spiking network so that it's hard to analysis the network's computational objective.

Related work



[Pehlevan & Chklovskii, 2015]

- **Constrained Similarity-Matching**

  [Pehlevan and Chklovskii, 2015] **:**

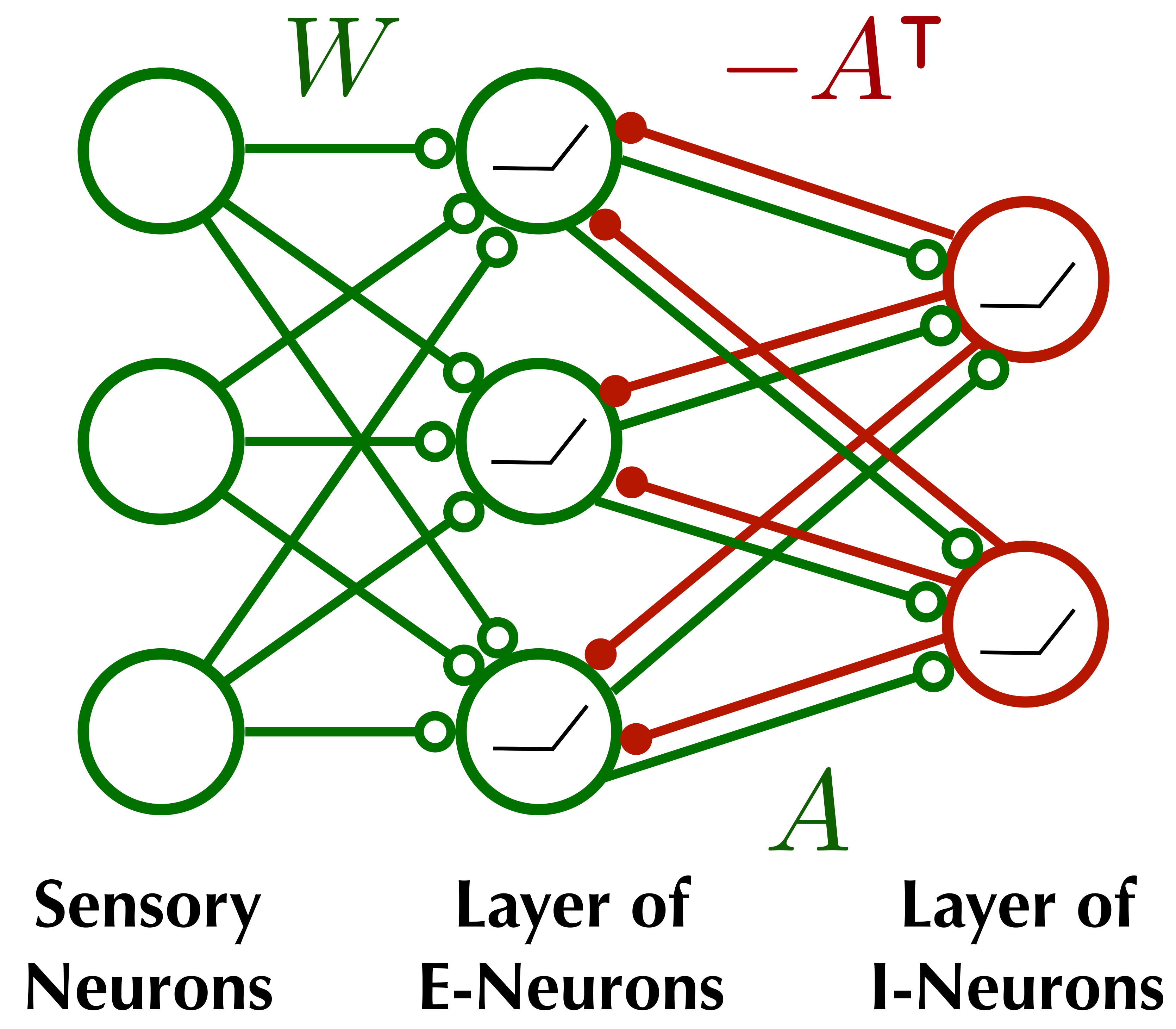
  I) interaction mediated by interneurons

  II) rate-based model

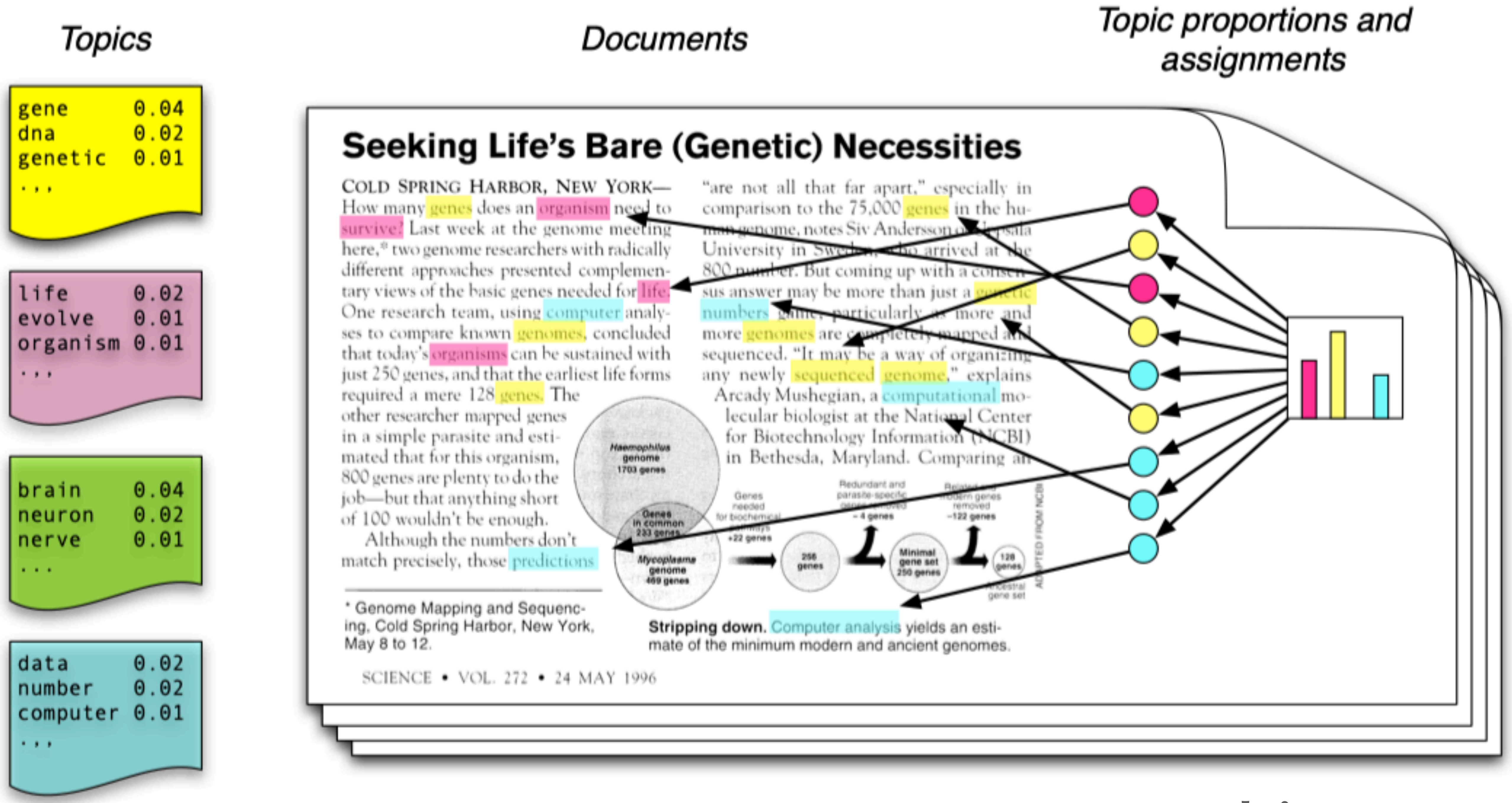
  III) derived from a constrained similarity principle

  IV) neurons are linear

What're potential ML applications?



Task Description of Topic Models

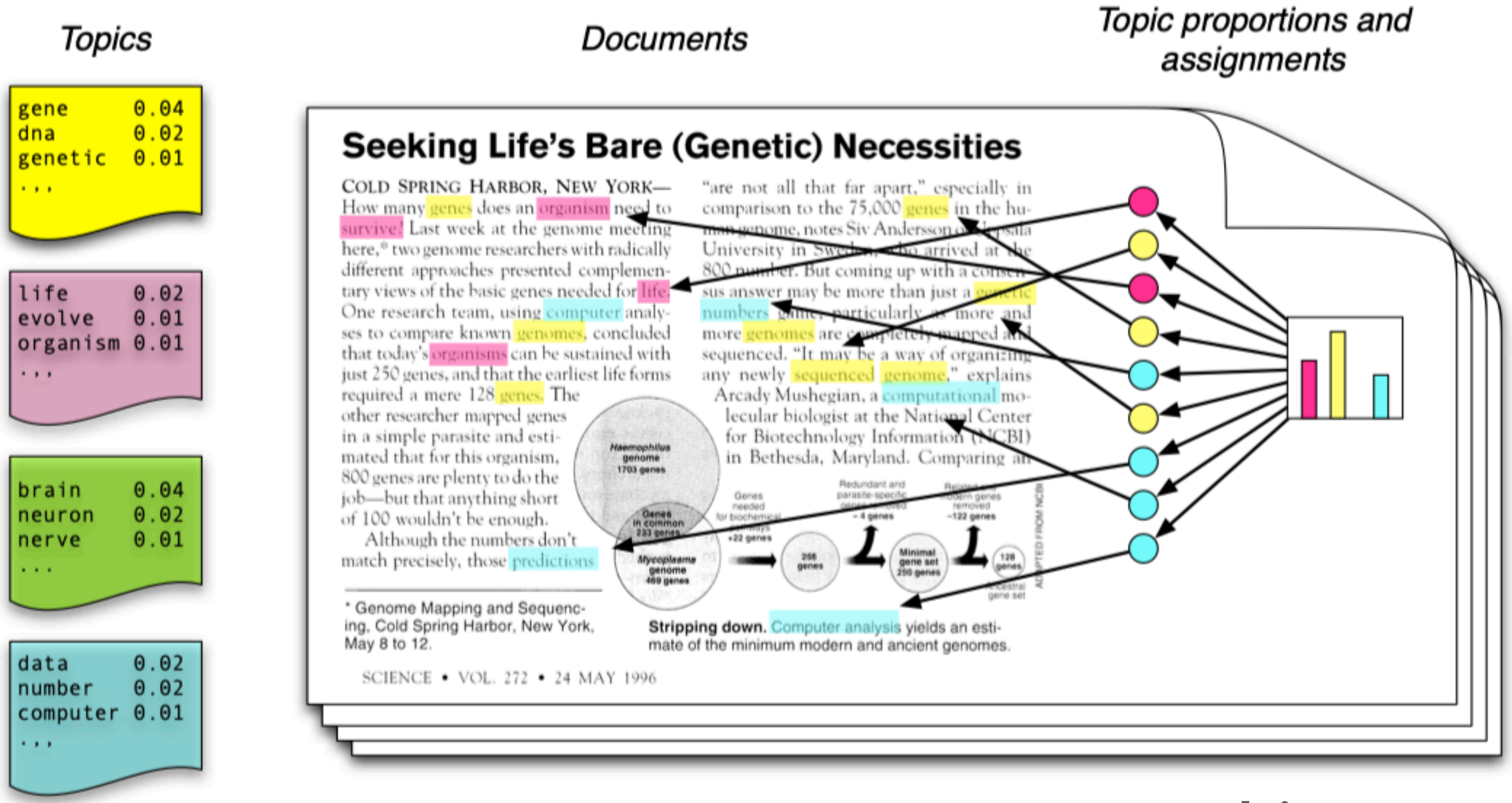


[Blei, 2012]

Generative View:

- Each **document** is a mixture of **topics**.
- Each **topic** is a distribution of **words**.
- Each **word** is drawn from one of those **topics**.

Task Description of Topic Models



[Blei, 2012]

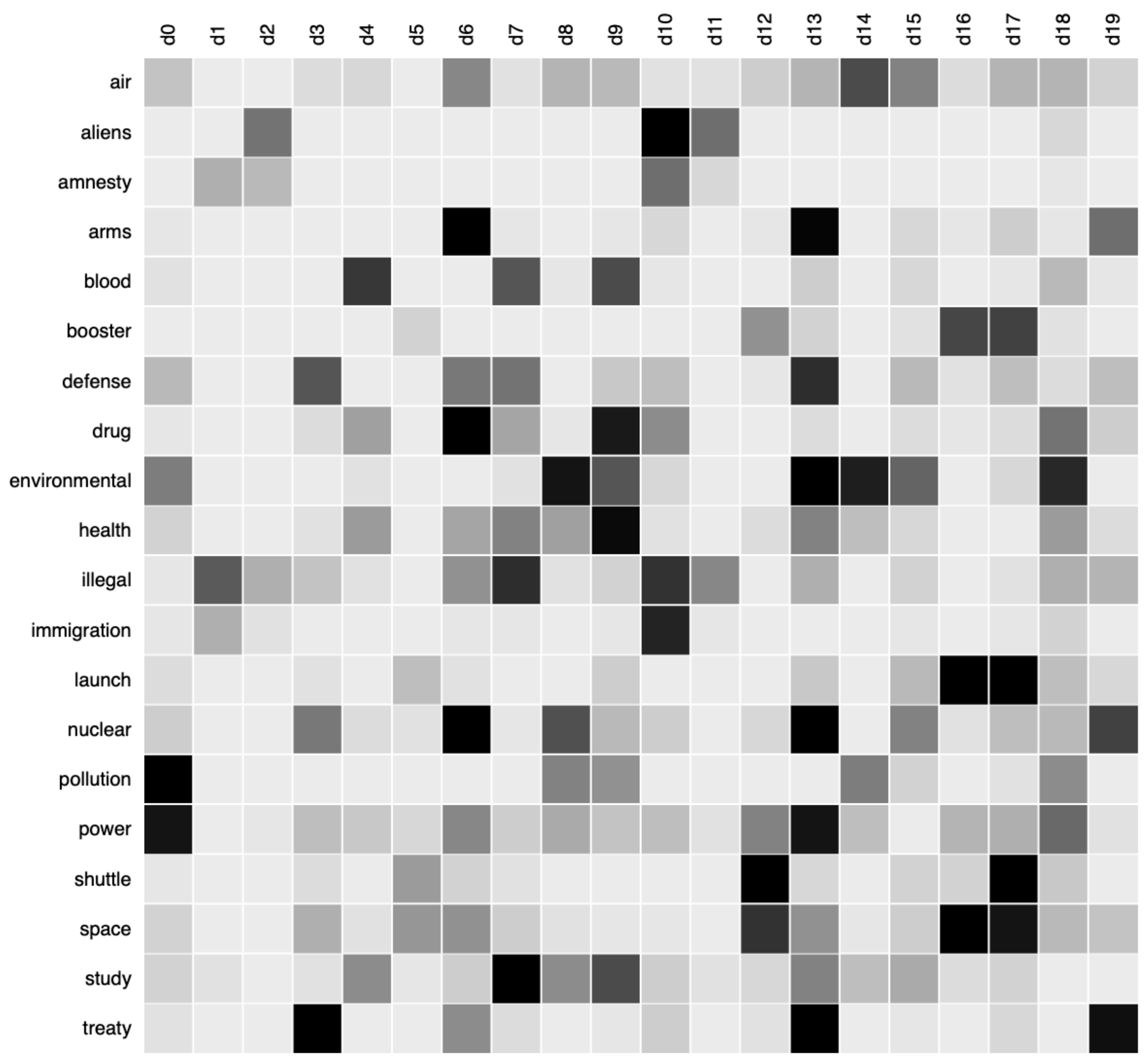
Generative View:

- Each **document** is a mixture of **topics**.
- Each **topic** is a distribution of **words**.
- Each **word** is drawn from one of those **topics**.

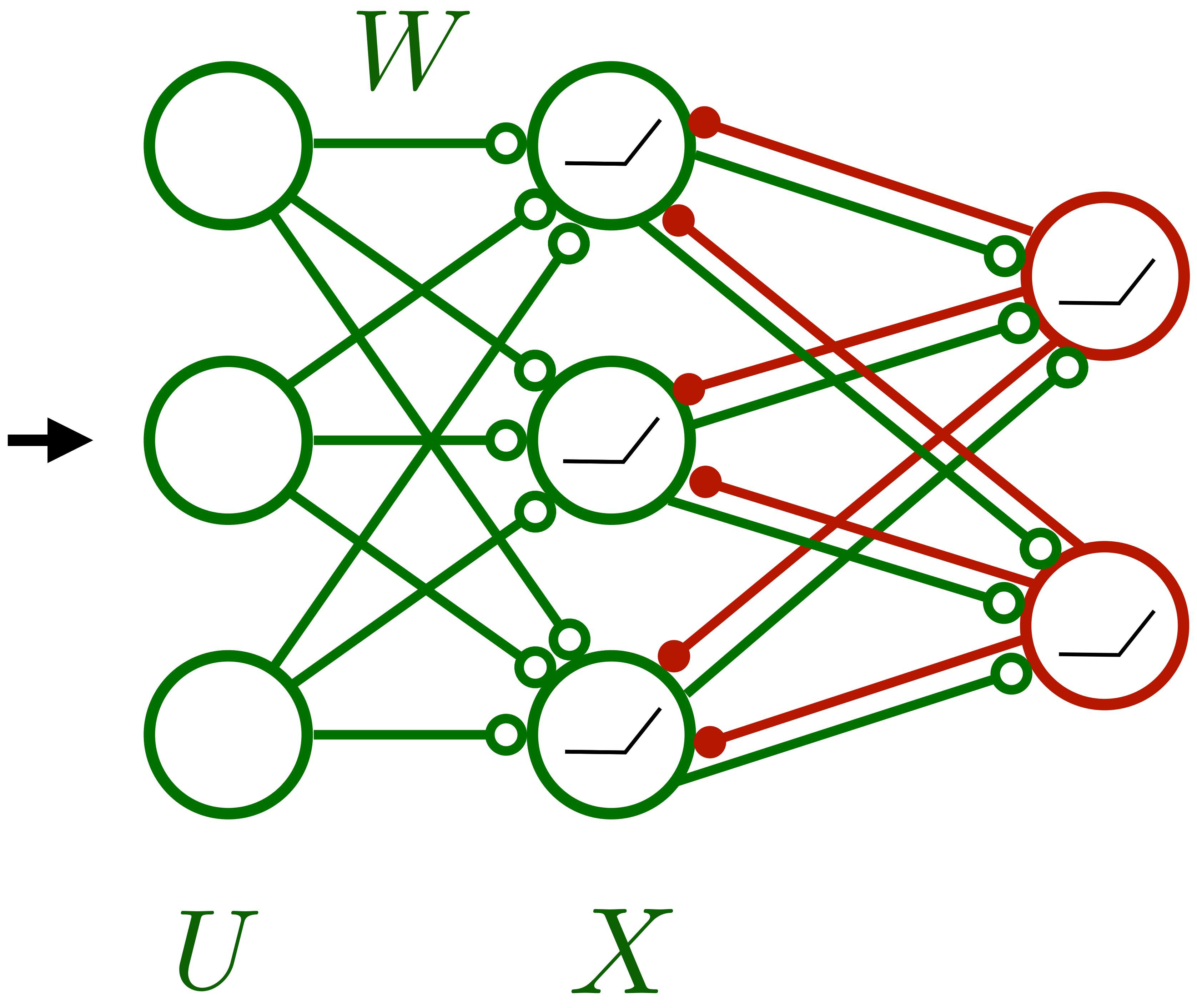
Task of Topic Modeling:

- Given documents, extract topics

Applying Disynaptic Neural Network to Topic Models



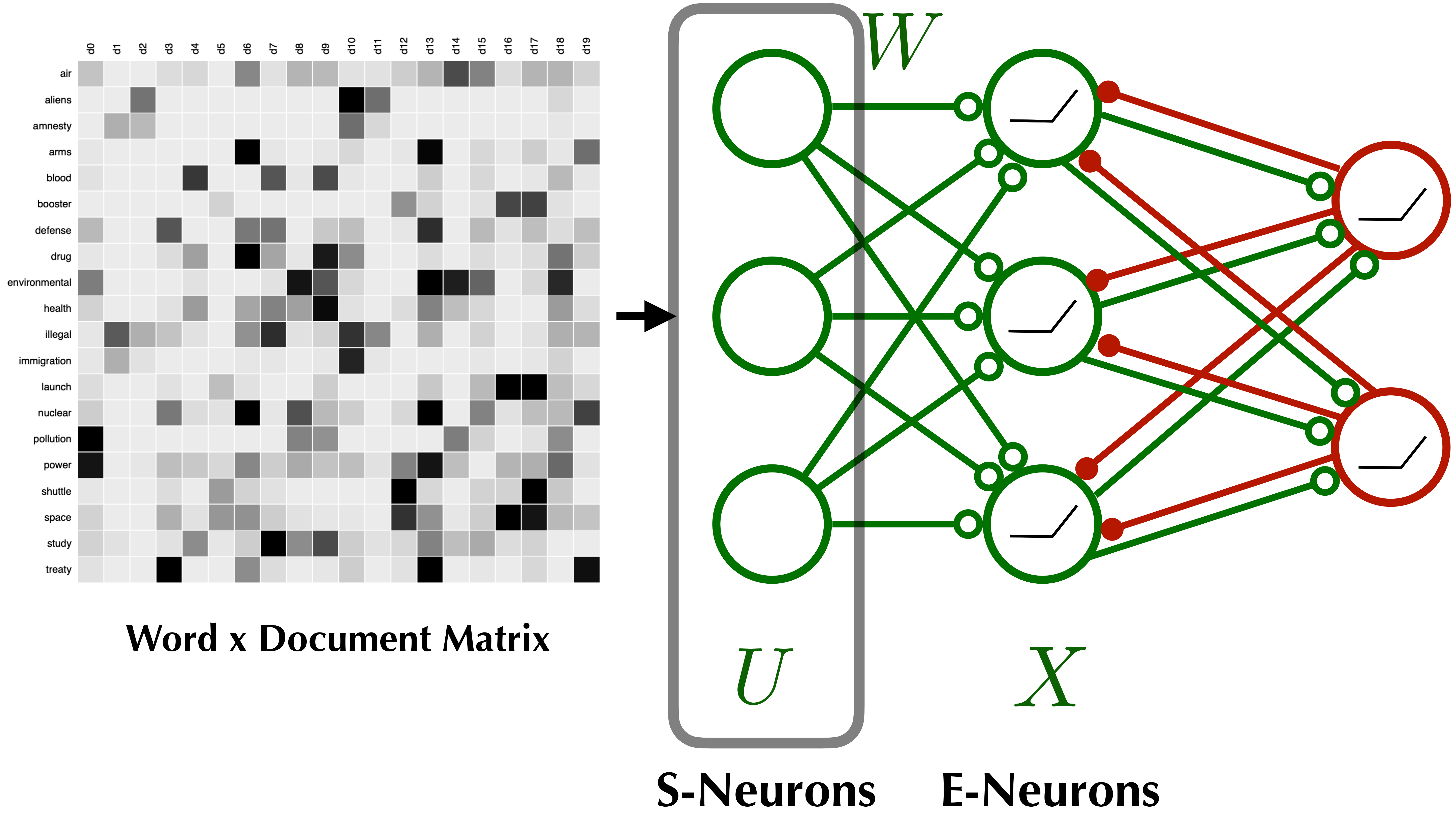
Word x Document Matrix



S-Neurons E-Neurons

Non-Generative Method:

Applying Disynaptic Neural Network to Topic Models

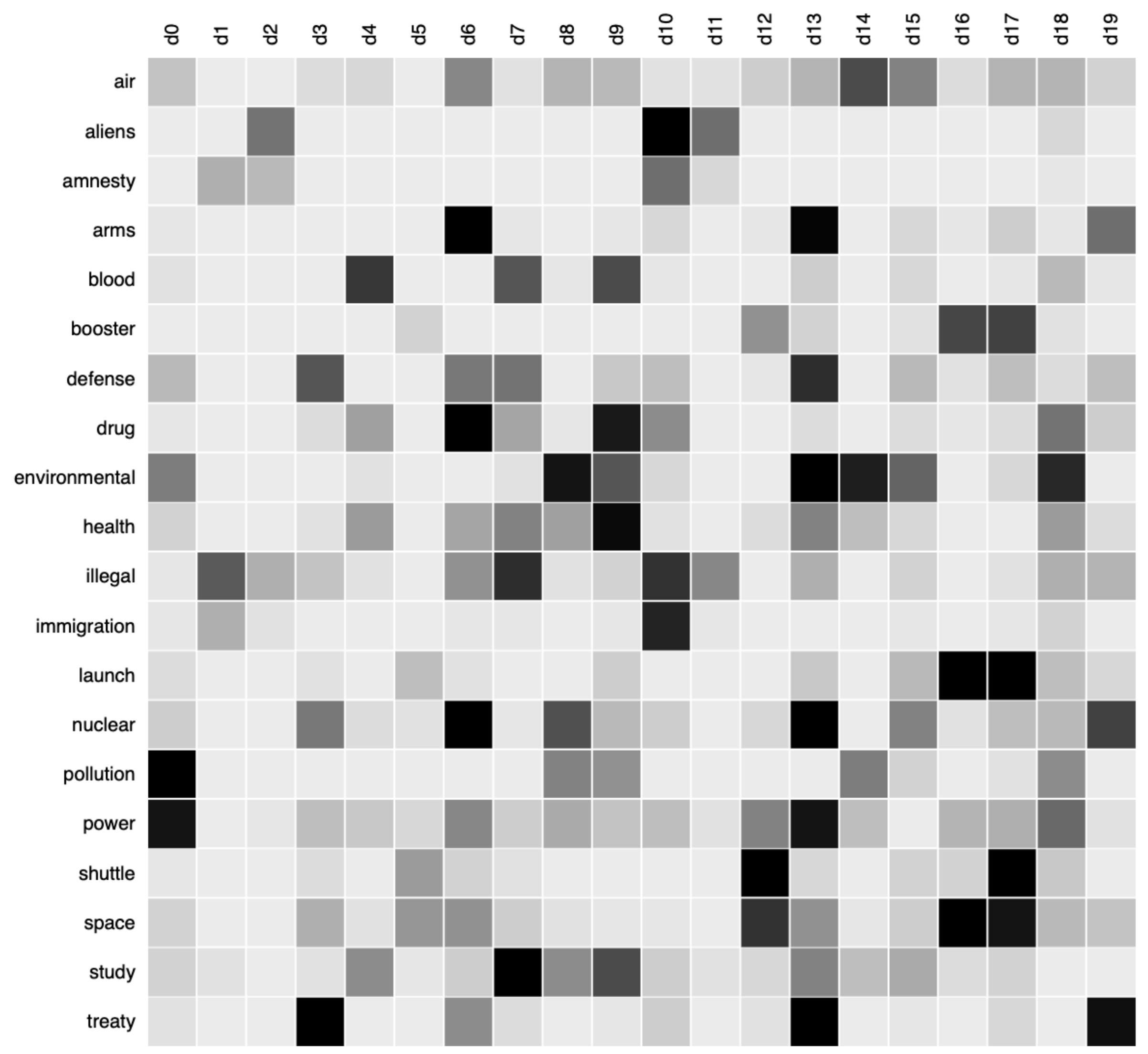


Non-Generative Method:

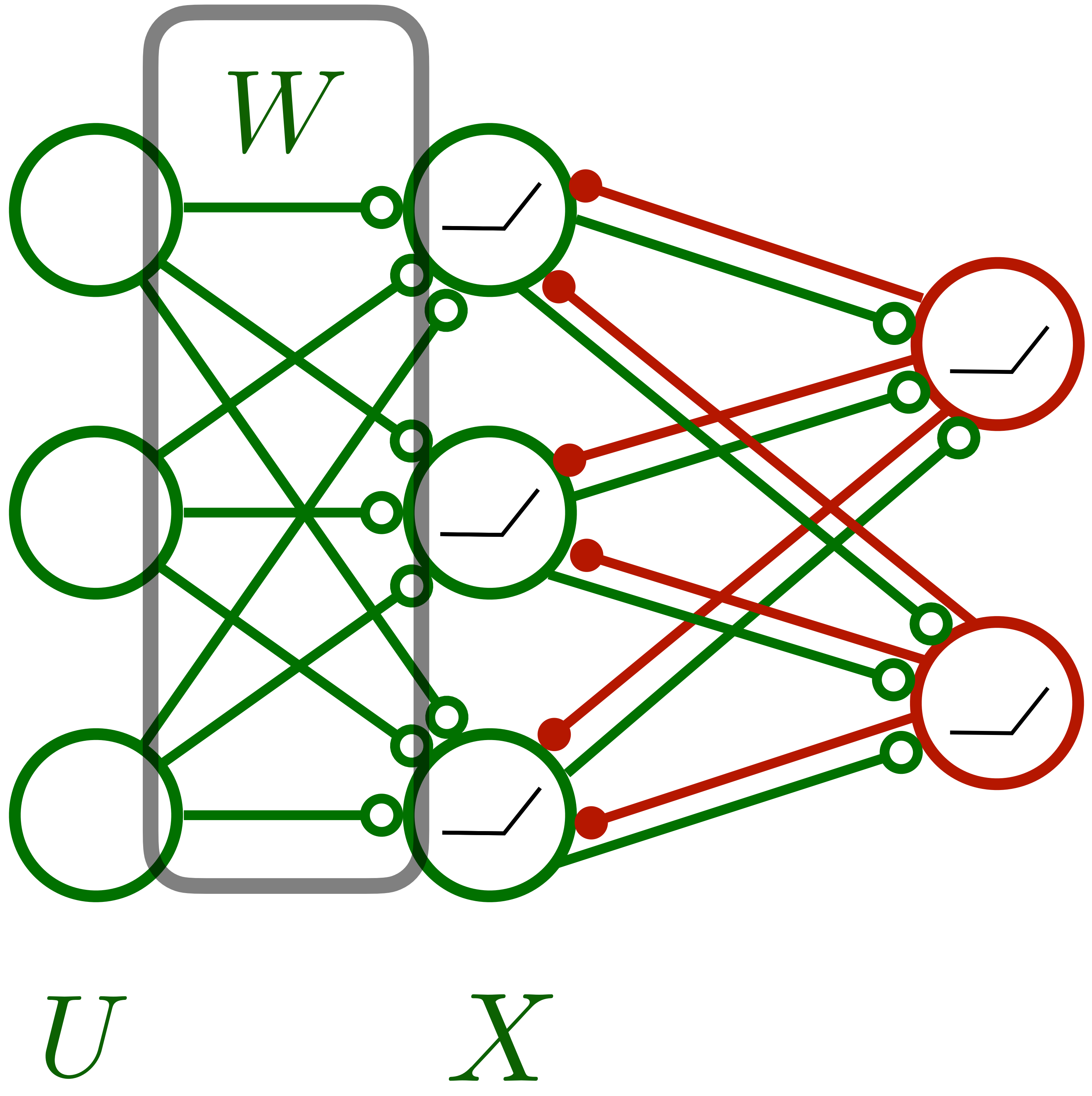
- Input $\boldsymbol{U_{\cdot t}}$ is the *t-th* document in the bag-of-words representation.

 # S-Neuron = Size of vocabulary

Applying Disynaptic Neural Network to Topic Models



Word x Document Matrix



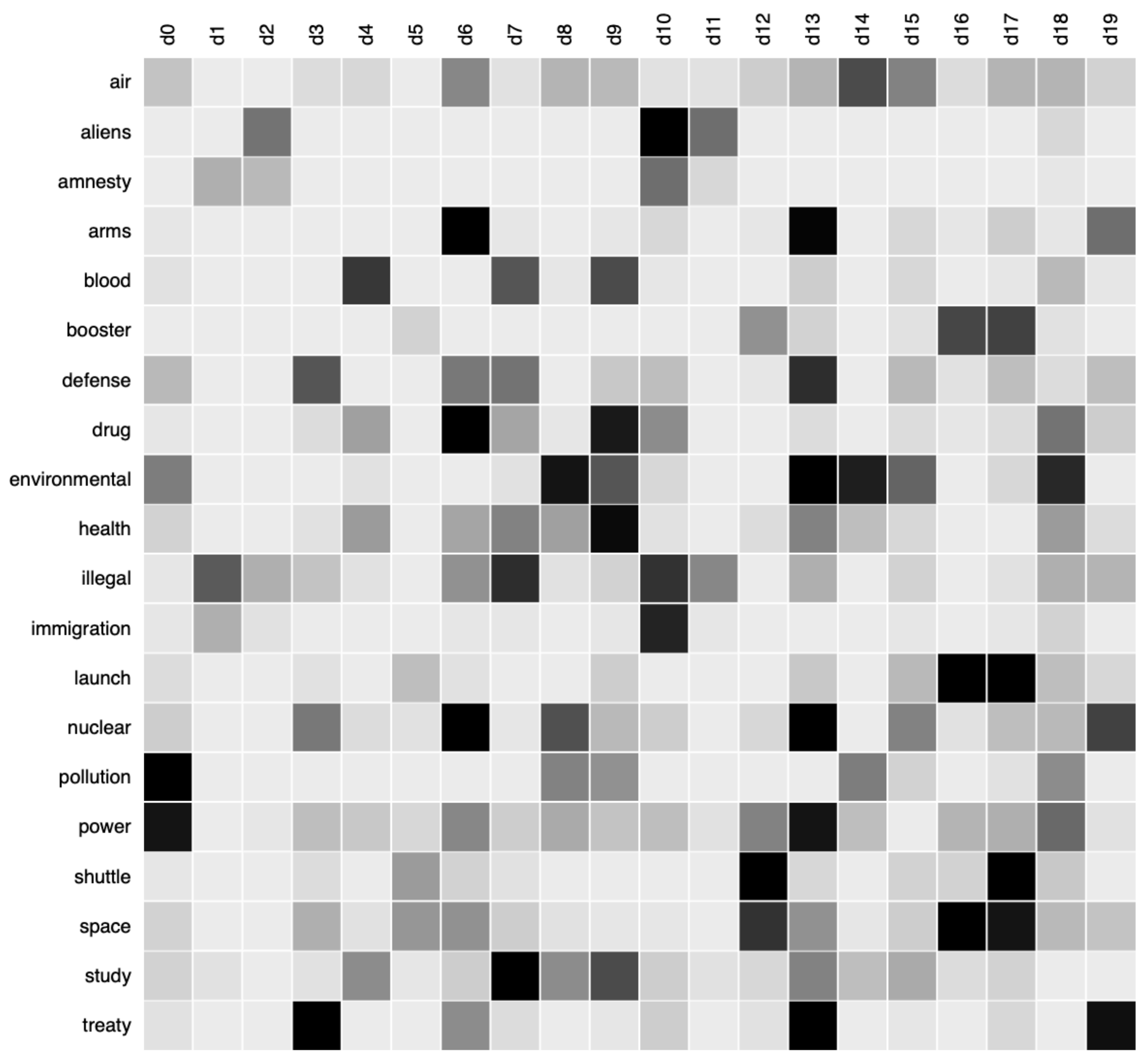
S-Neurons **E-Neurons**

Non-Generative Method:

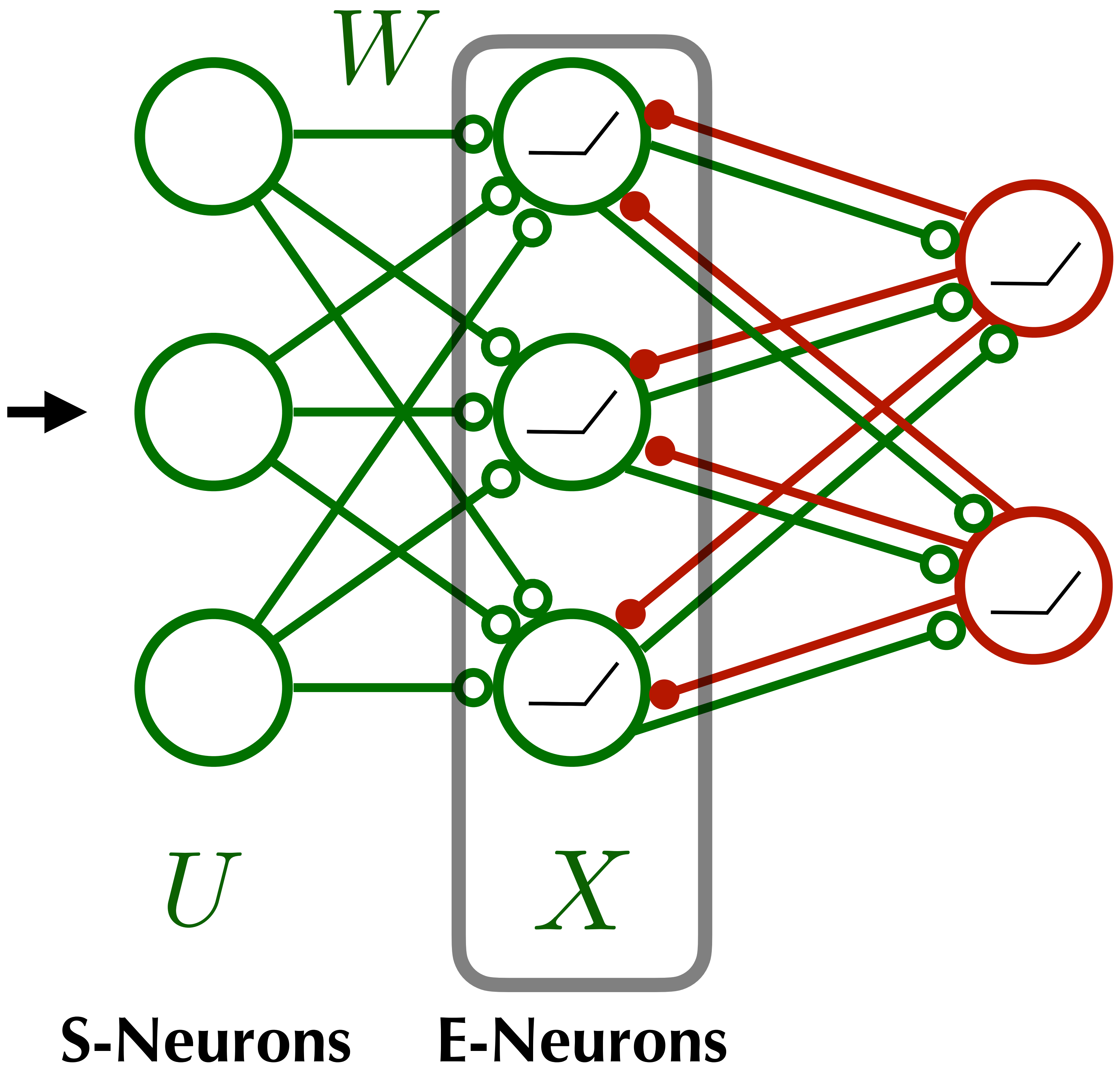
- Input $U_{\cdot t}$ is the *t-th* document in the bag-of-words representation.
- Learned S-E connections $W_{i\cdot}$ is the *i-th* topic (relevance to each word).

 # E-Neuron = Number of Topics

Applying Disynaptic Neural Network to Topic Models



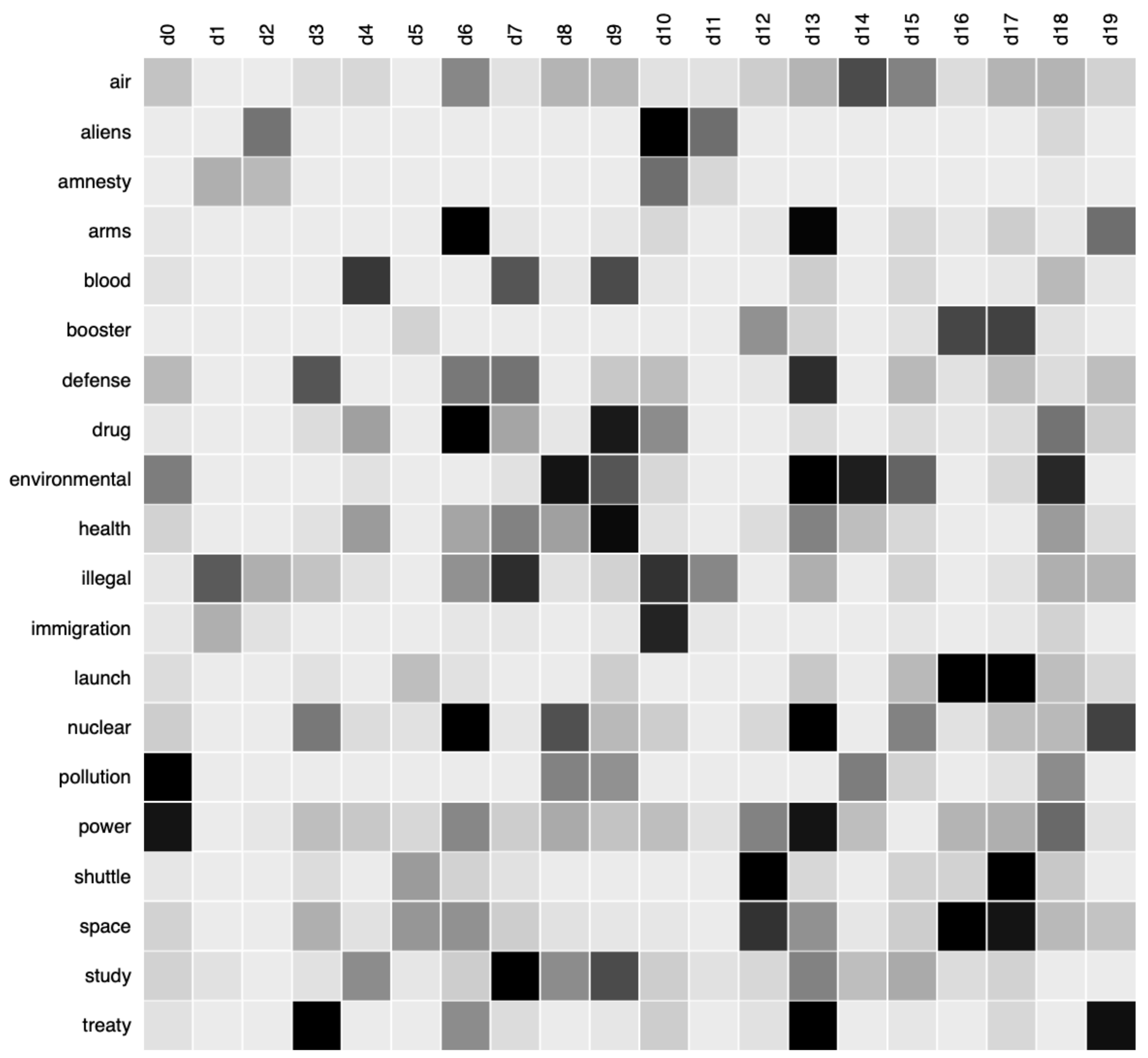
Word x Document Matrix



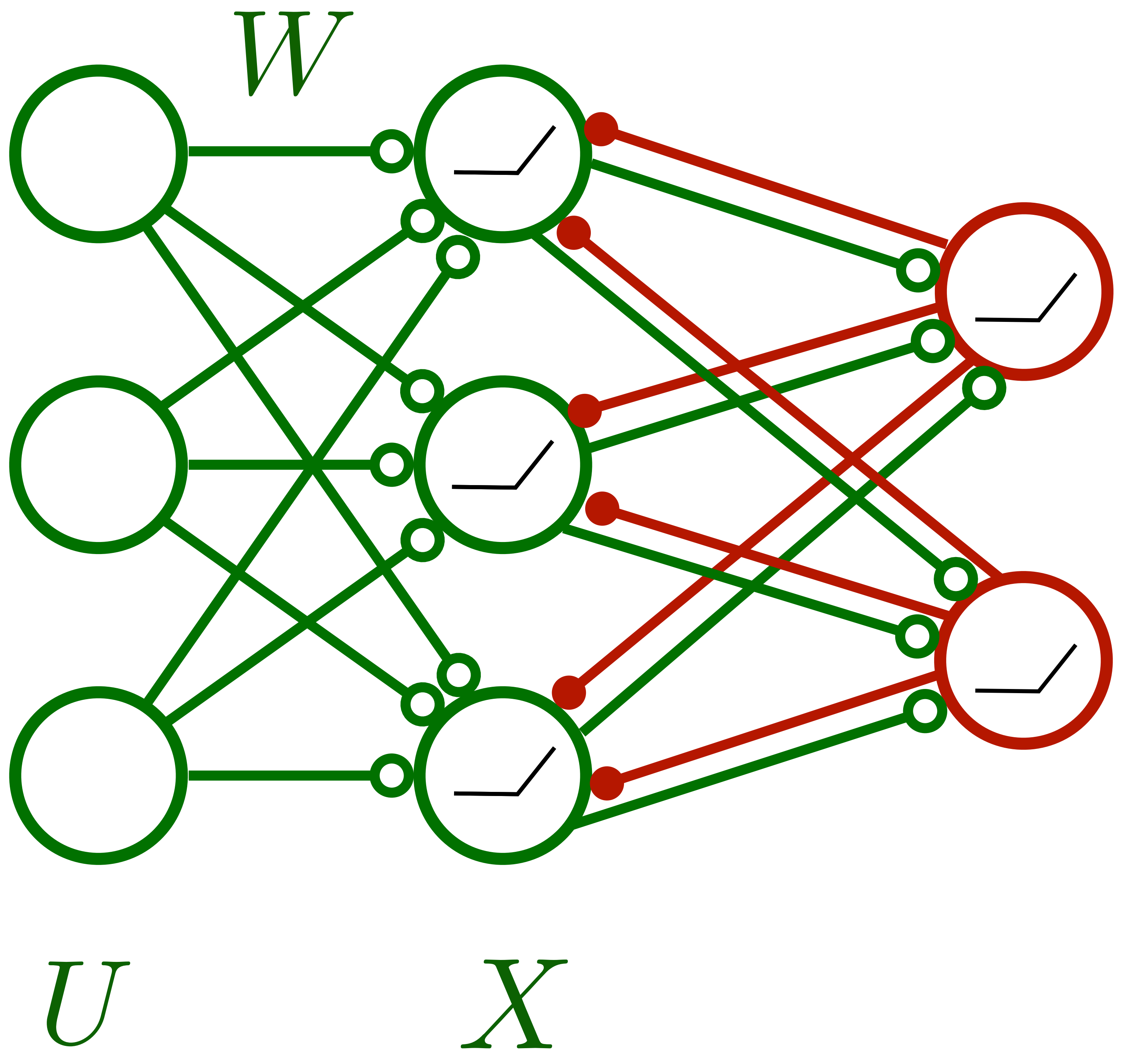
Non-Generative Method:

- Input $\boldsymbol{U_{\cdot t}}$ is the *t-th* document in the bag-of-words representation.
- Learned S-E connections $\boldsymbol{W_{i\cdot}}$ is the *i-th* topic (relevance to each word).
- E-neuron activity $\boldsymbol{X_{it}}$ is the score of topic assignment.

Applying Disynaptic Neural Network to Topic Models



Word x Document Matrix



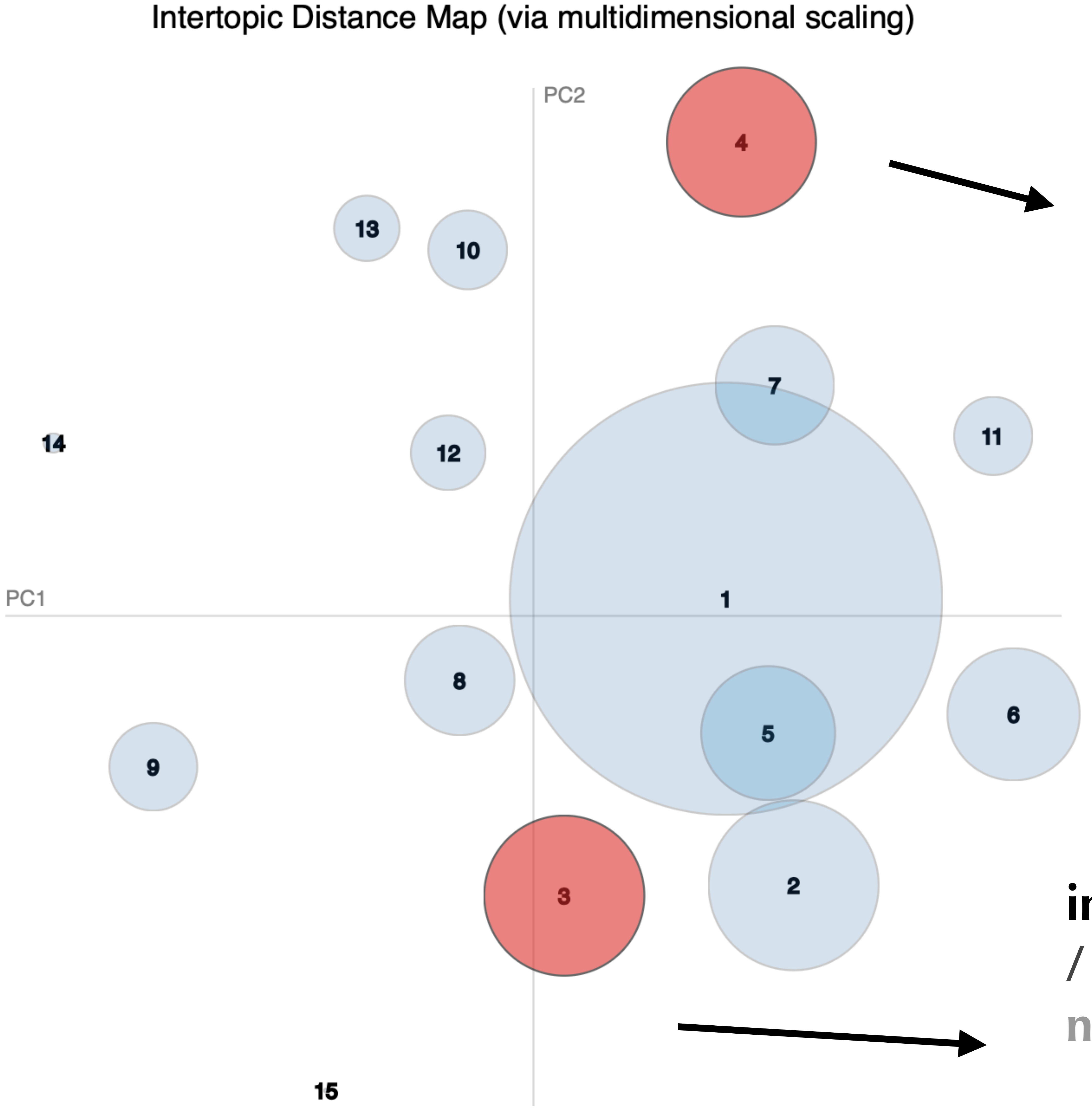
S-Neurons **E-Neurons**

Non-Generative Method:

- Input $U_{\cdot t}$ is the *t-th* document in the bag-of-words representation.
- Learned S-E connections $W_{i\cdot}$ is the *i-th* topic (relevance to each word).
- E-neuron activity X_{it} is the score of topic assignment.

Maximizing topic-document correlation, while minimizing topic-topic correlation.

Emerging Topics in a Network with Disynaptic Recurrent Inhibition



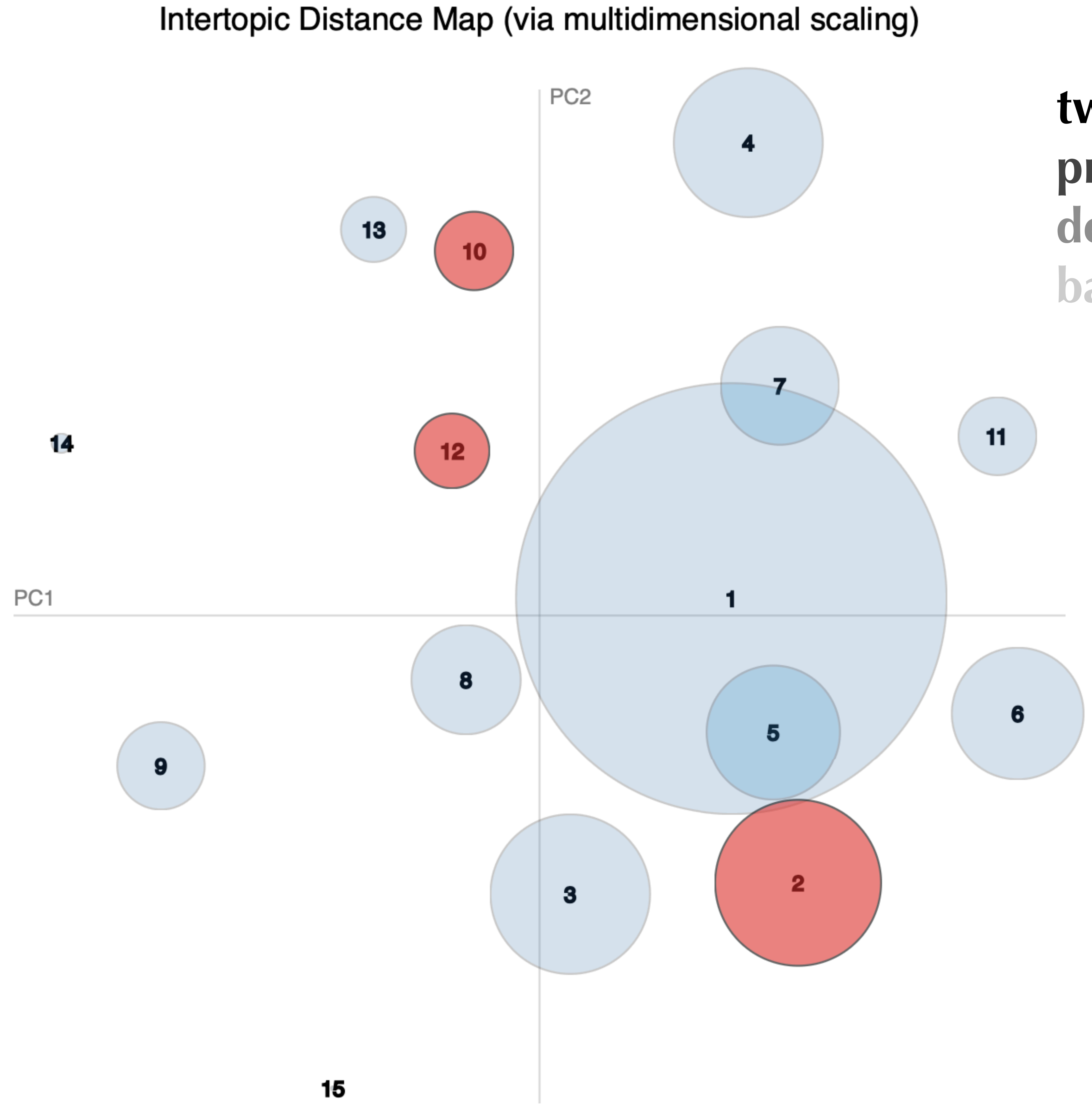
Topic 4: “College”

student / college / university / teacher / school / painting / exam / score / continue / ap / art / movement / data / high / performance / education / class / categorical / food / give

Topic 3: “Image Classification”

image / file / label / classification / letter / class / model / one / trained / improving / can / photo / recognition / network / set / contains / training / help / example / use

Emerging Topics in a Network with Disynaptic Recurrent Inhibition



Topic 10: “Twitter”

tweet / trump / donald / twitter / text / speech / time / presidential / someone / user / content / using / debate / election / sentiment / day / id / clinton / based / date

Topic 12: “Election”

election / vote / party / campaign / presidential / candidate / result / political / state / contribution / position / content / voting / constituency / federal / data / contains / expenditure / commission / finance

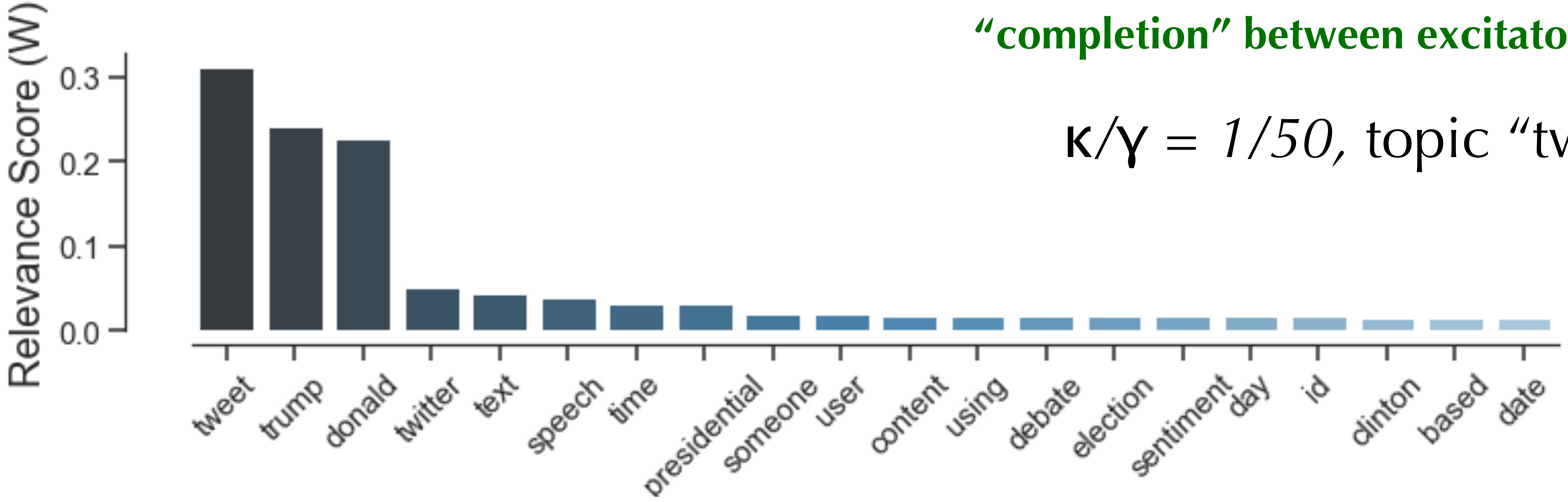
Topic 2: “Games”

game / player / team / match / pokemon / play / season / league / data / every / point / stats / played / can / com / information / per / card / result / number

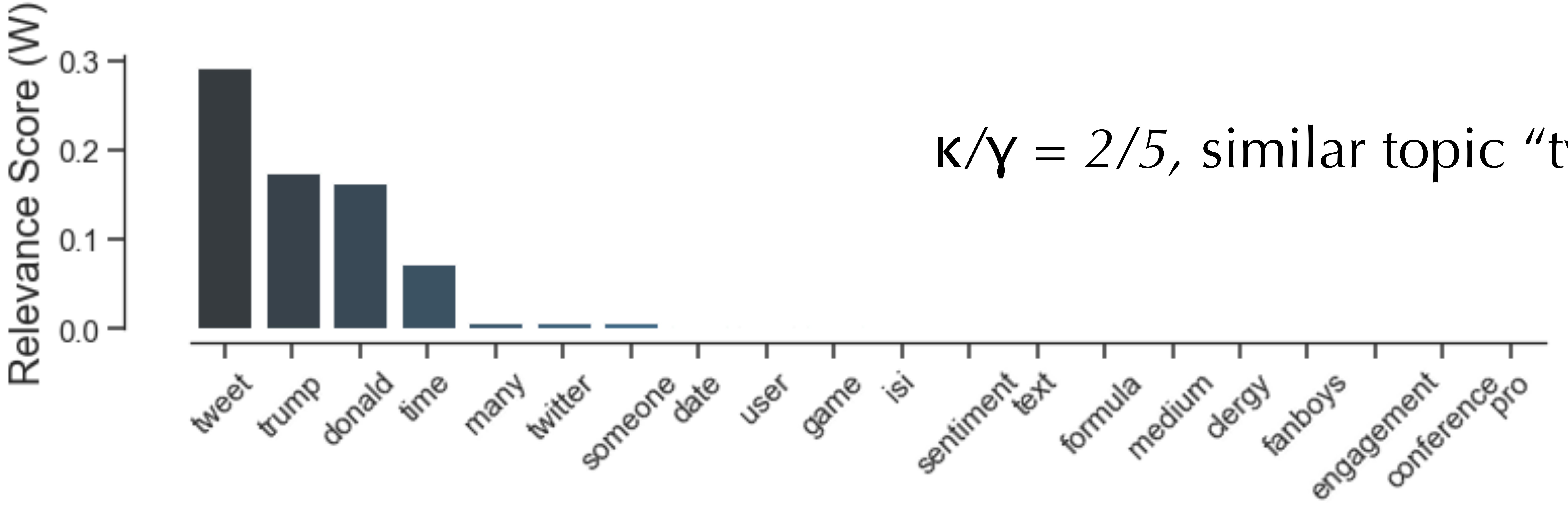
Controlling topic-word sparsity

$$\phi(W)_{ia} := \frac{\partial \Phi(W)}{\partial W_{ia}} = \gamma W_{ia} + \kappa \sum_b W_{ib}$$

"completion" between excitatory synapses



κ/γ = *1/50*, topic "tweet"



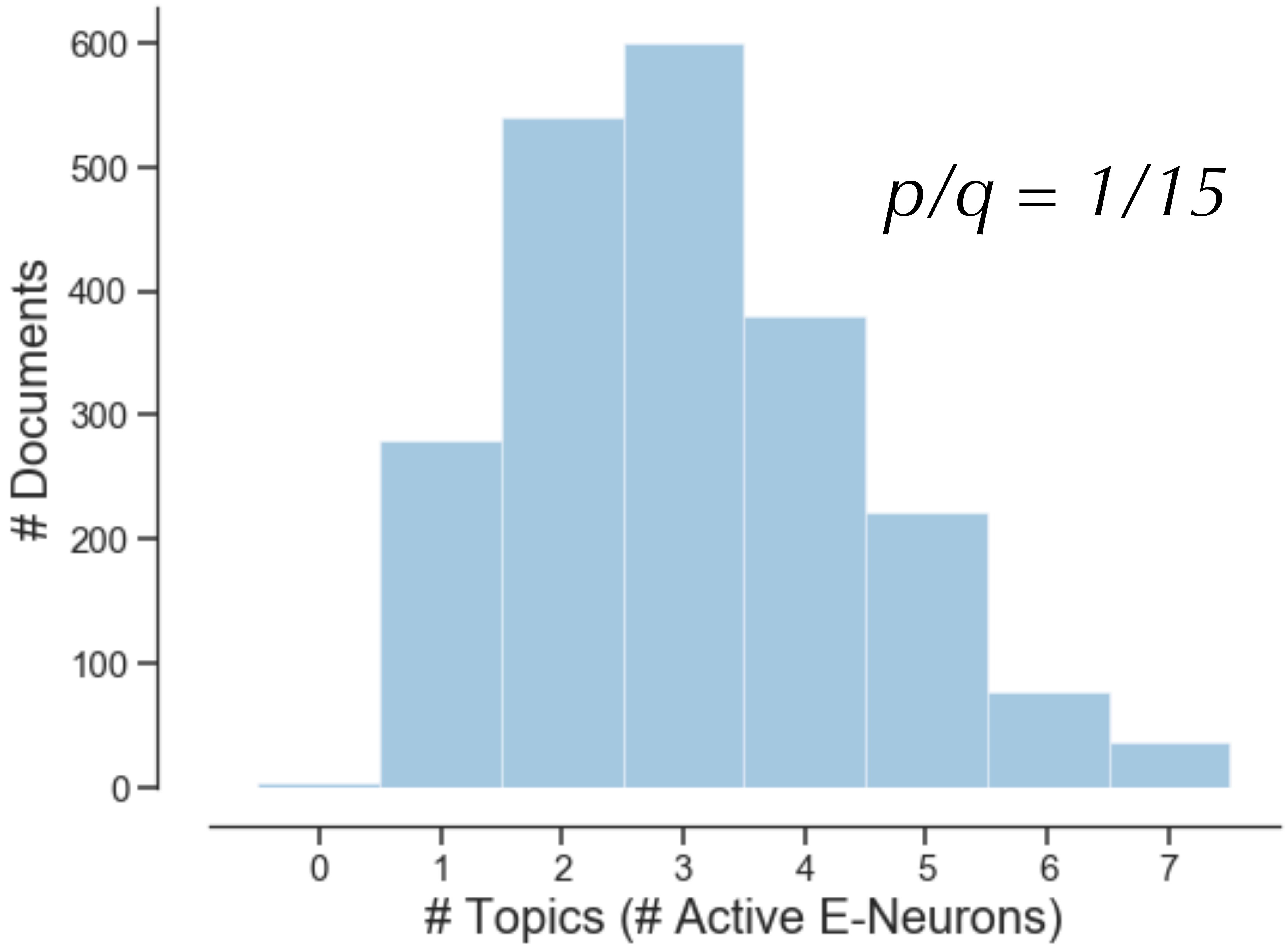
κ/γ = *2/5*, similar topic "tweet"

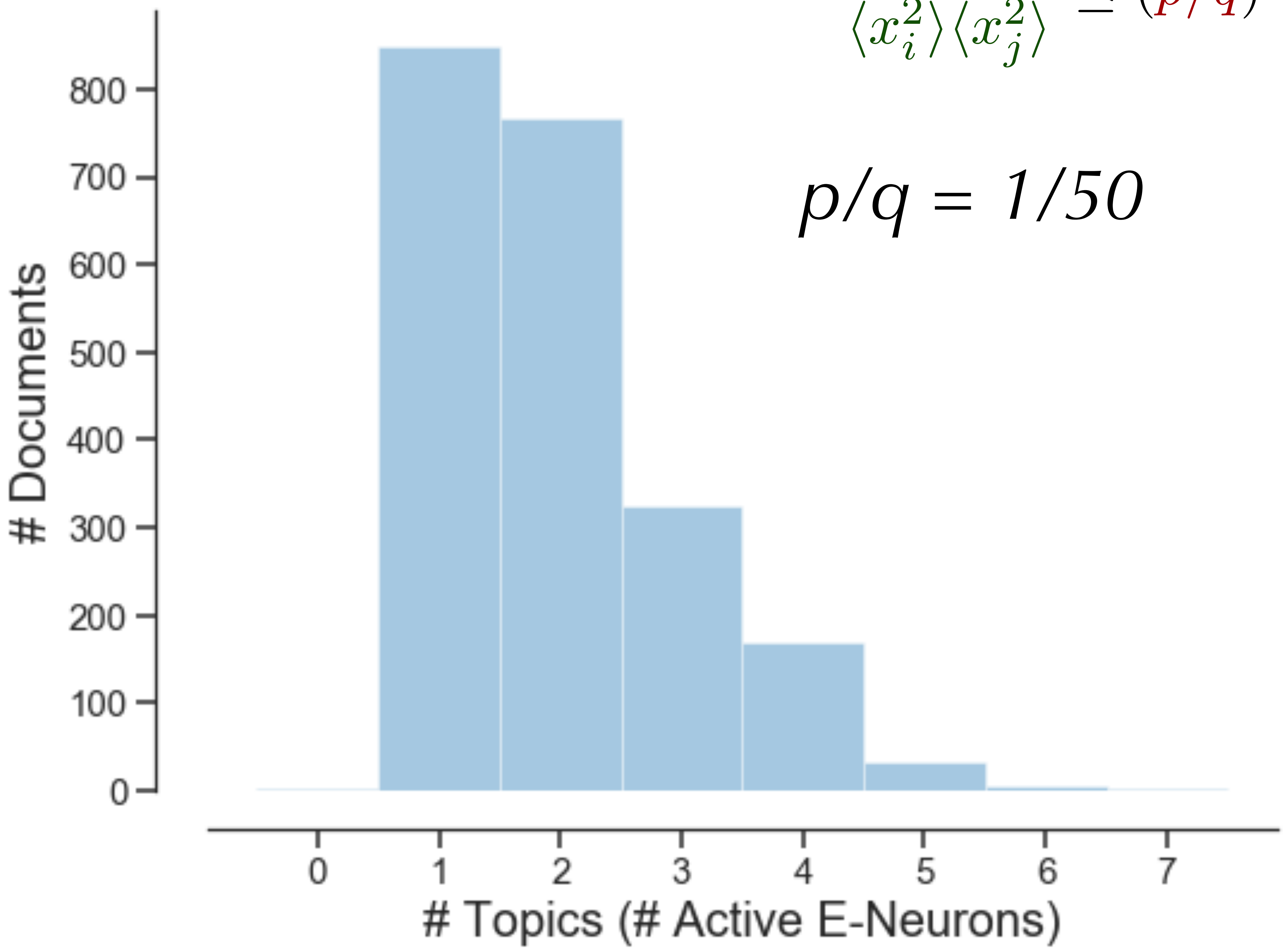
sparse feature when κ/γ is large — fewer key words for each topic

Controlling document-topic sparsity

$$\psi(A)_{\alpha i} := \frac{\partial \Psi(\Lambda + A^\intercal A)/2}{\partial A_{\alpha i}} = (q^2 - p^2)A_{\alpha i} + p^2 \sum_i A_{\alpha i}$$

$$\sim \frac{\langle x_i x_j \rangle}{\langle x_i^2 \rangle \langle x_j^2 \rangle} \leq (p/q)^2$$





strong decorrelation when *p/q* is small — sparser topic assignment.

Discussion

- Neural networks with disynaptic recurrent inhibition can approximate the **“softened” correlation game** principle.
- With only **a few inhibitory neurons** it can learn **diverse features**.
- Application to **topic models** shows that our neural network can discover topics which are similar to LDA with **controllable sparsity**.
- Future work:
  - Potential efficiency gain of a non-generative model?
  - Learning semantic embeddings of words?