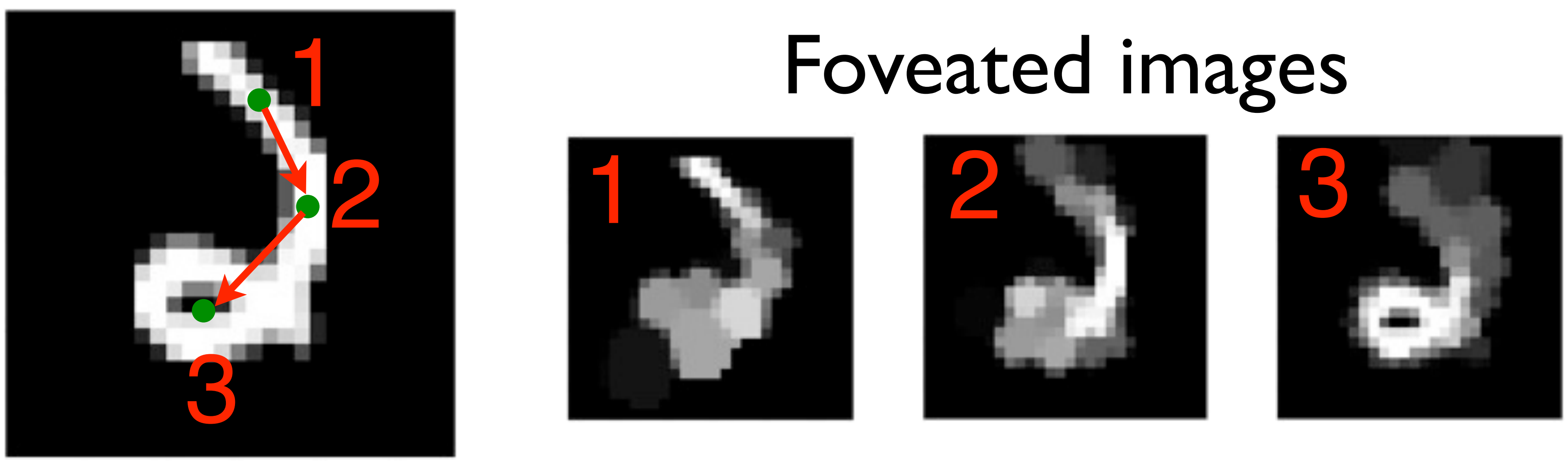
Learning to combine foveal glimpses with a third-order Boltzmann machine

Hugo Larochelle and Geoffrey Hinton
University of Toronto

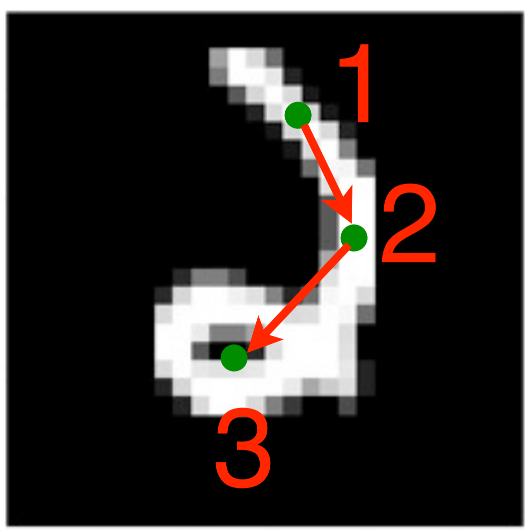
Introduction

- Human vision has the two following characteristics
  - ★ Uses an intelligent "fixation point strategy"
  - ★ Based on a retina with variable spatial resolution

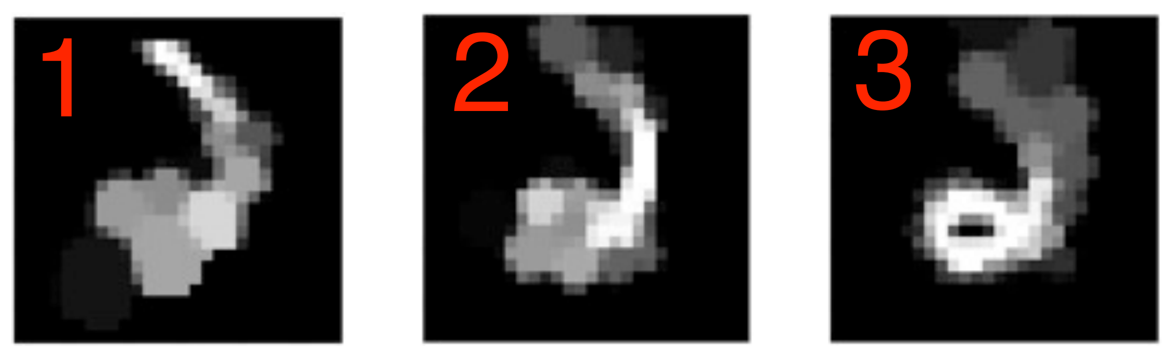


Introduction

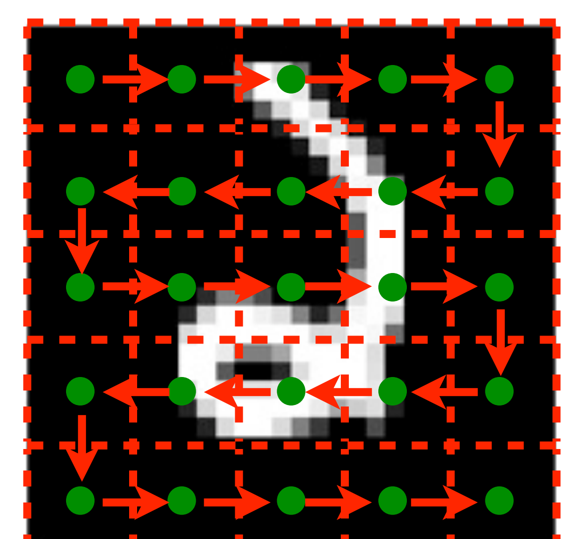
- Human vision has the two following characteristics
  - ★ Uses an intelligent "fixation point strategy"
  - ★ Based on a retina with variable spatial resolution



Foveated images

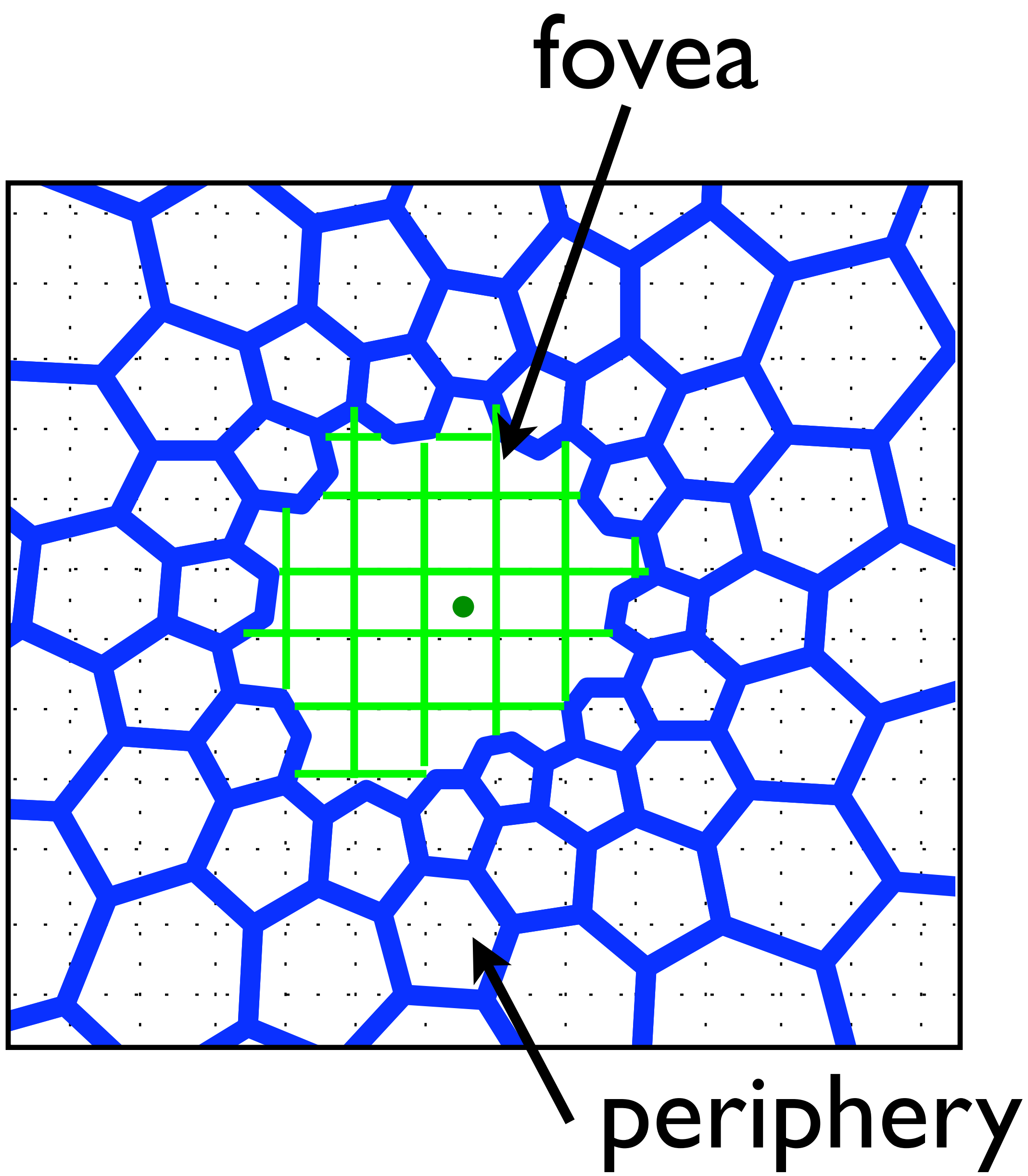


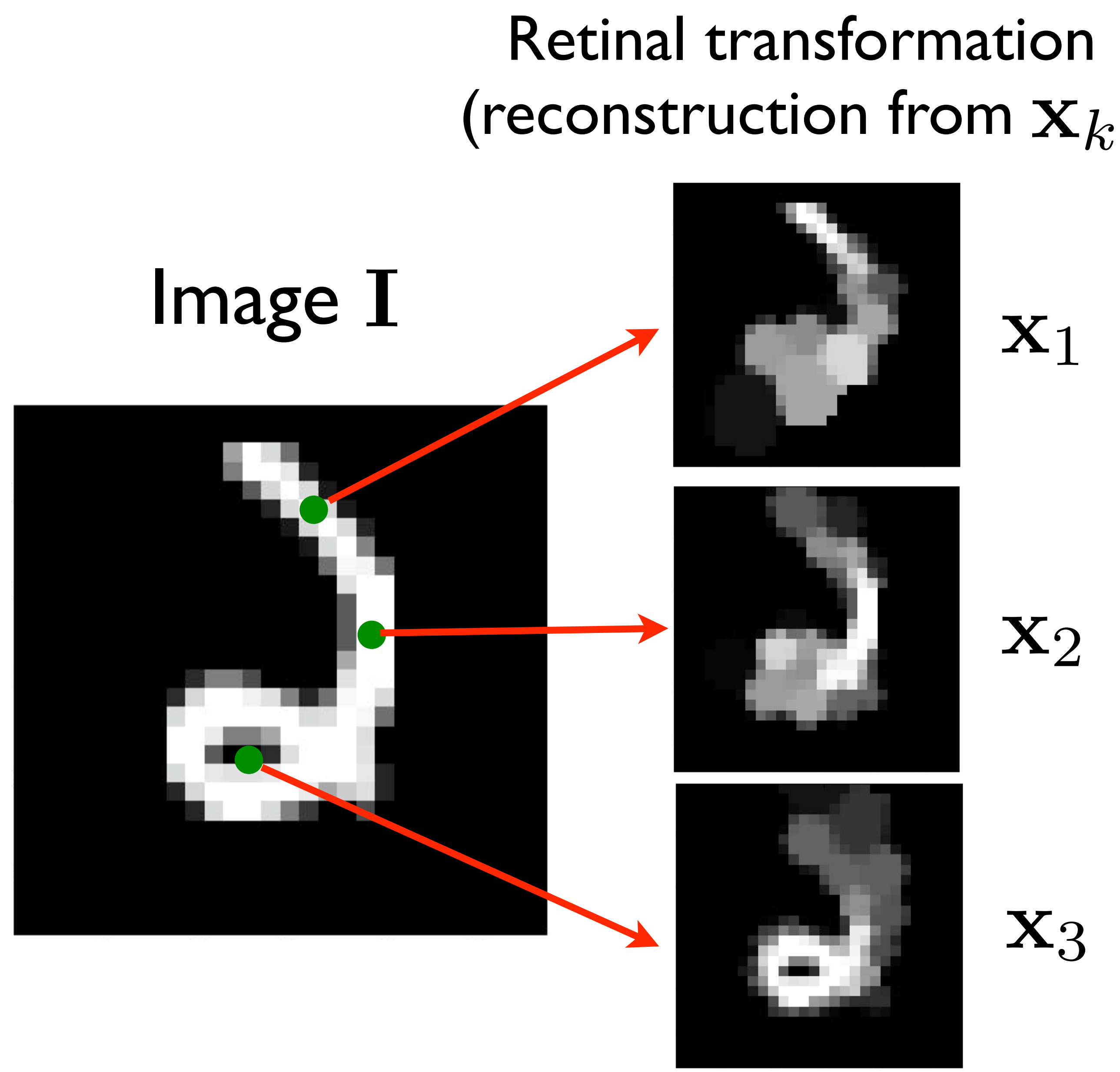
- Many vision systems are instead based on a uniform resolution retina and "fixate everywhere"



How to look: retinal transformation

- Information from the input image is extracted based on a retinal transformation ("glimpse")



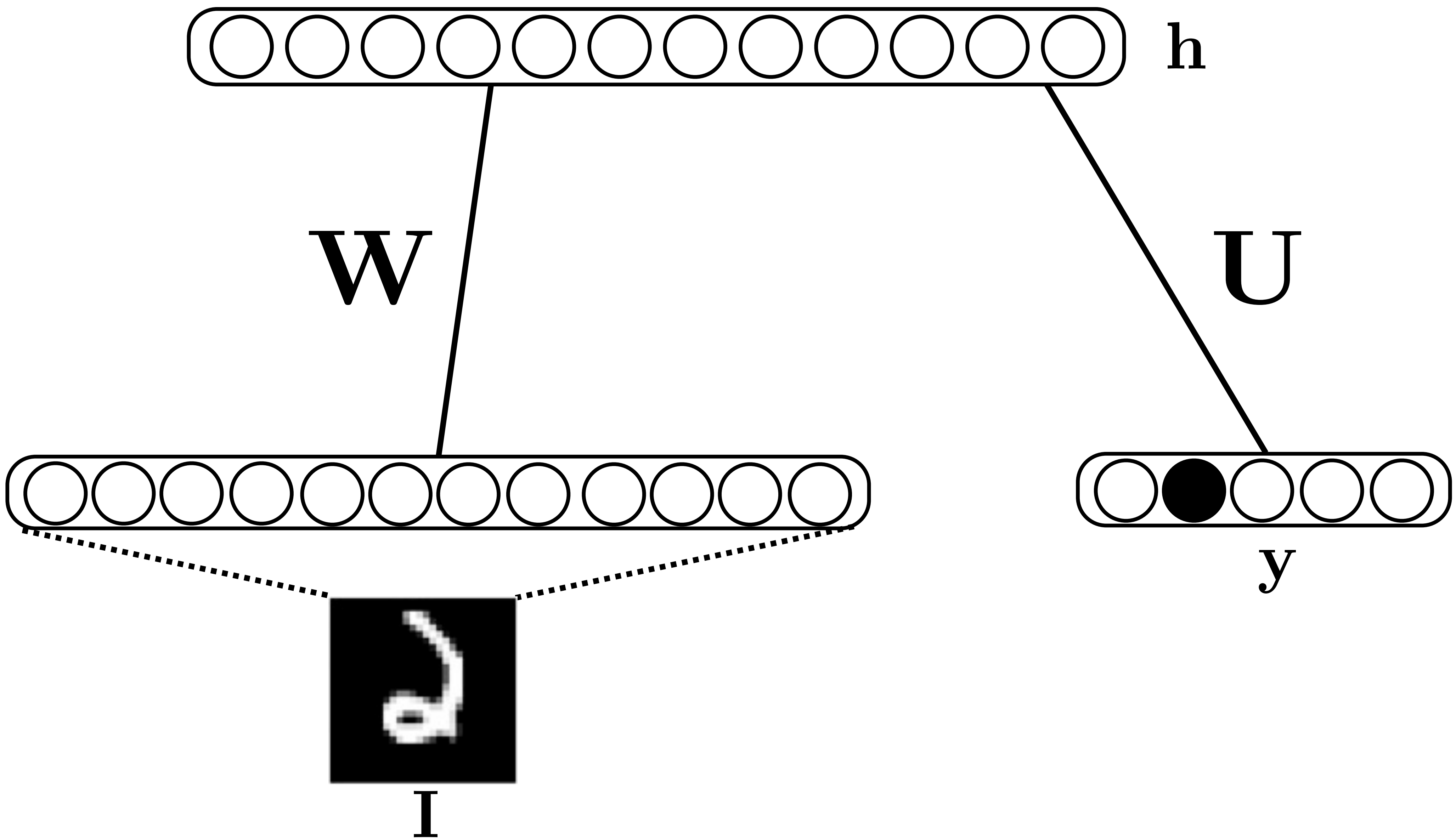


Components of the system

- **Recognition component (RBM)**
- Attentional component (controller)

Restricted Boltzmann Machine (RBM)

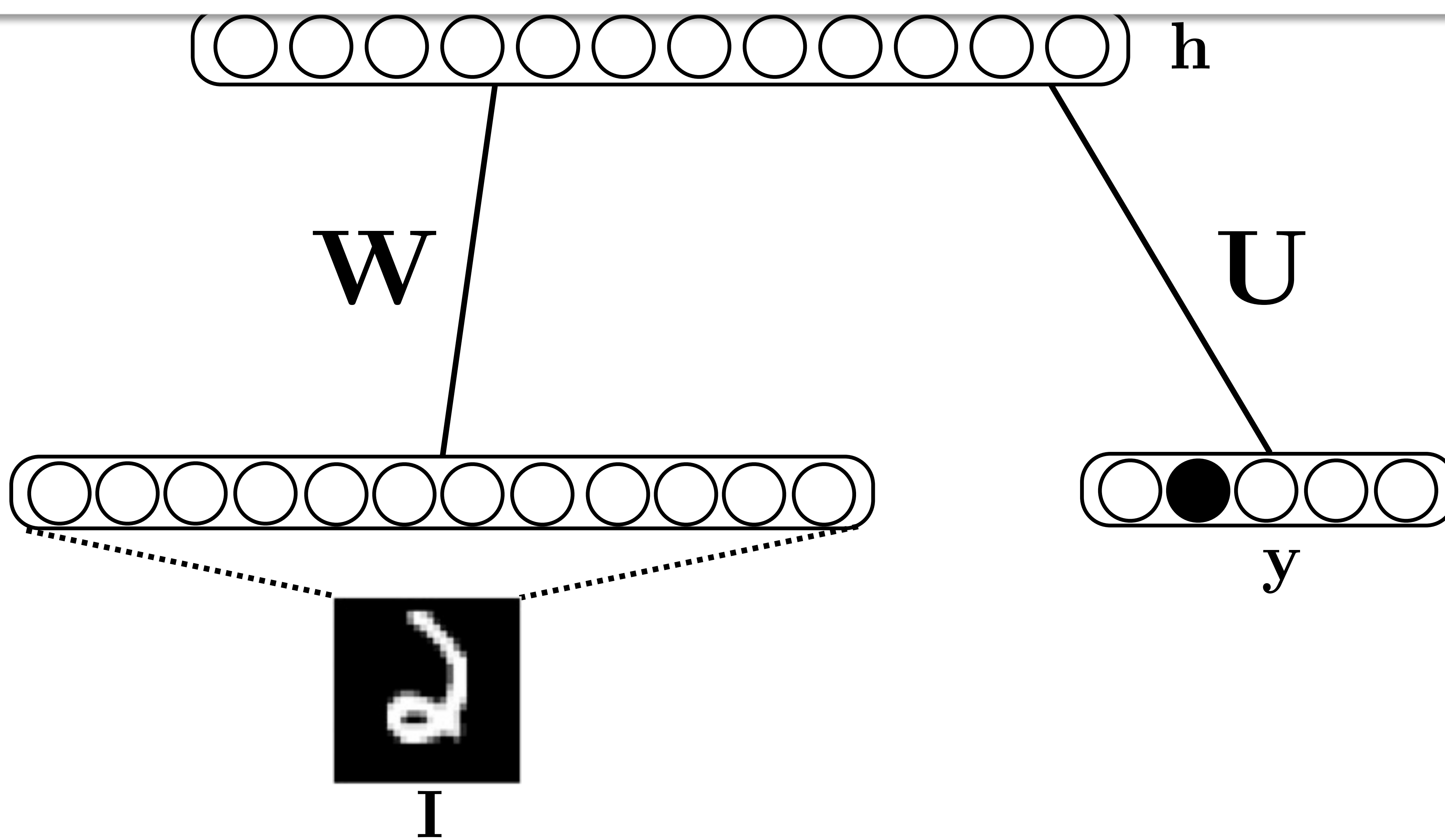
- Classification from the whole image $\mathbf{I}$



Key facts

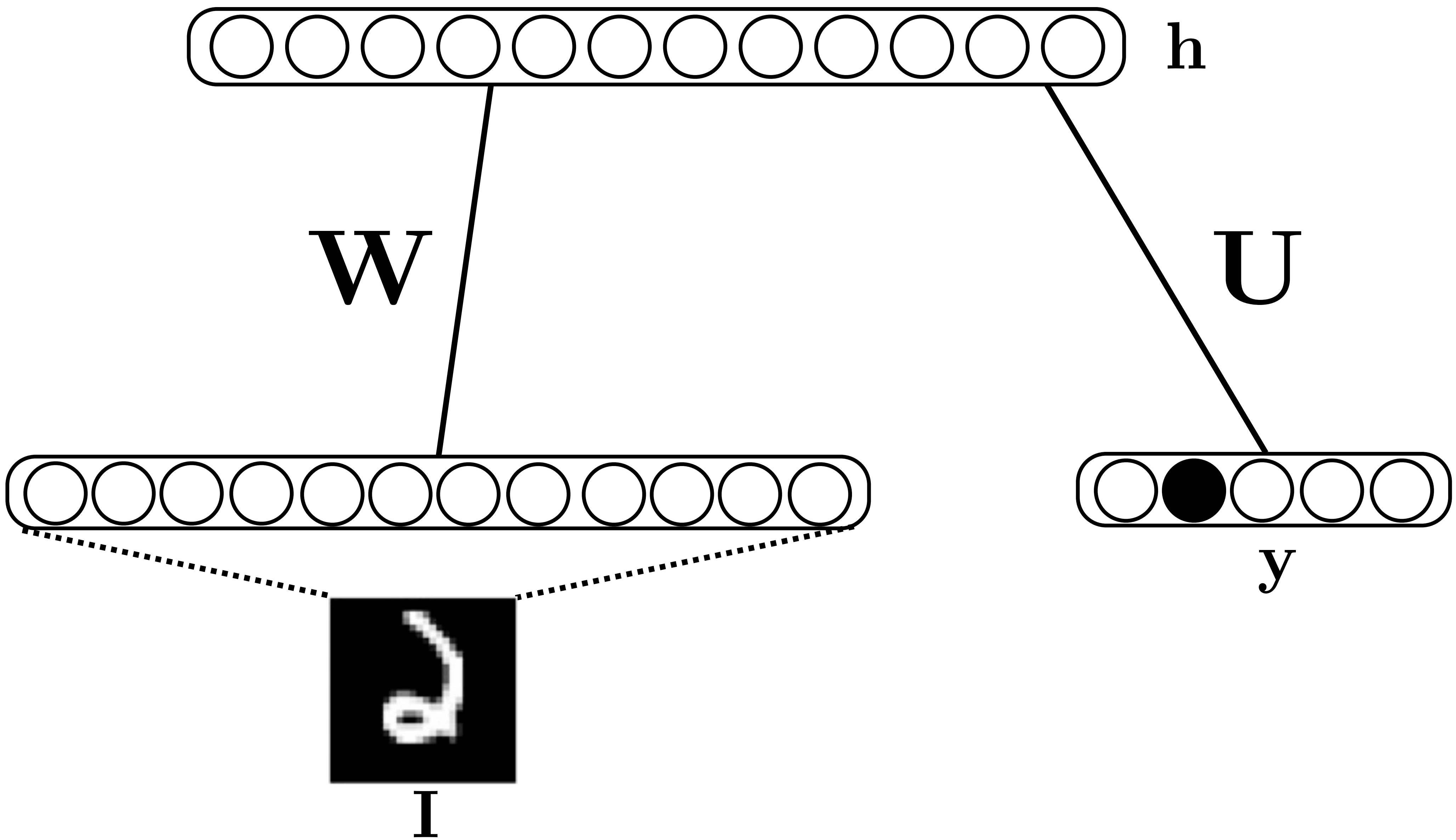
Energy function: $E(\mathbf{y}, \mathbf{I}, \mathbf{h}) = -\mathbf{h}^\top \mathbf{W}\ \mathbf{I} - \mathbf{h}^\top \mathbf{U}\ \mathbf{y}$

Probability: $p(\mathbf{y}, \mathbf{I}, \mathbf{h}) = \exp(-E(\mathbf{y}, \mathbf{I}, \mathbf{h}))/Z$



Restricted Boltzmann Machine (RBM)

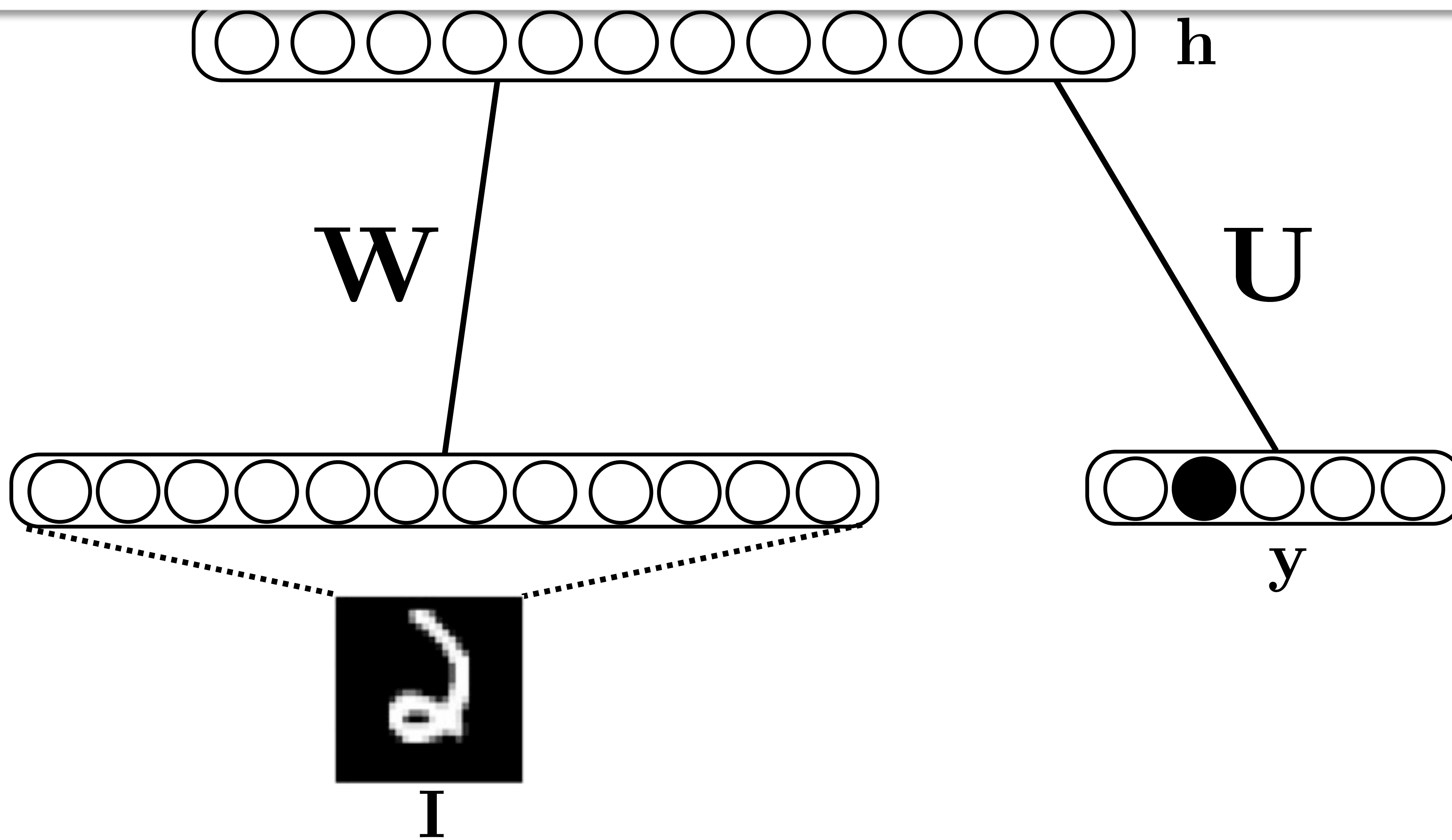
- Classification from the whole image $\mathbf{I}$



Inference is easy:

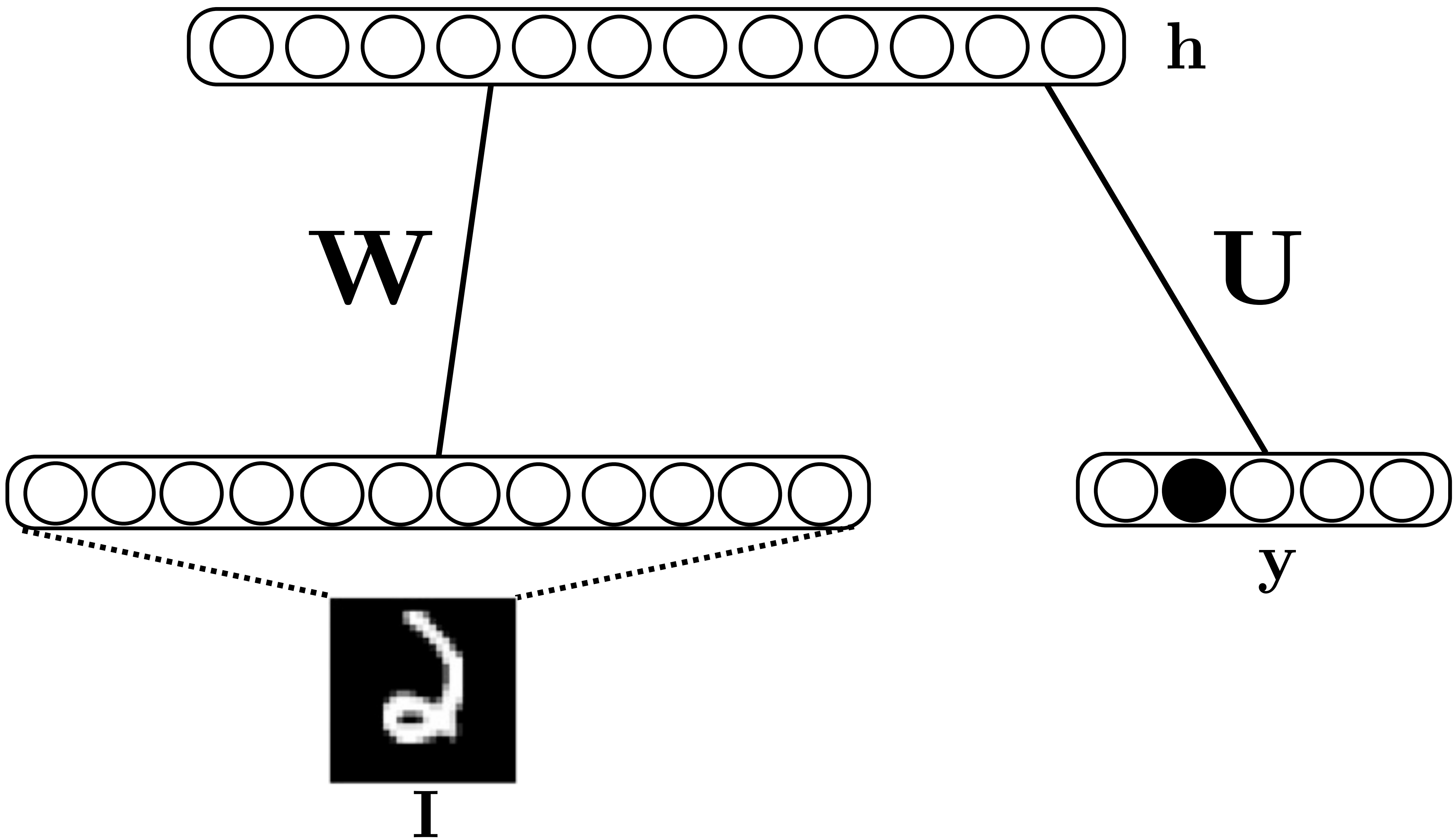
[up] $$p(\mathbf{h}|\mathbf{I},\mathbf{y}) = \prod_j p(h_j|\mathbf{I},\mathbf{y})$$

[down] $$p(\mathbf{I},\mathbf{y}|\mathbf{h}) = p(\mathbf{y}|\mathbf{h}) \prod_i p(I_i|\mathbf{h})$$



Restricted Boltzmann Machine (RBM)

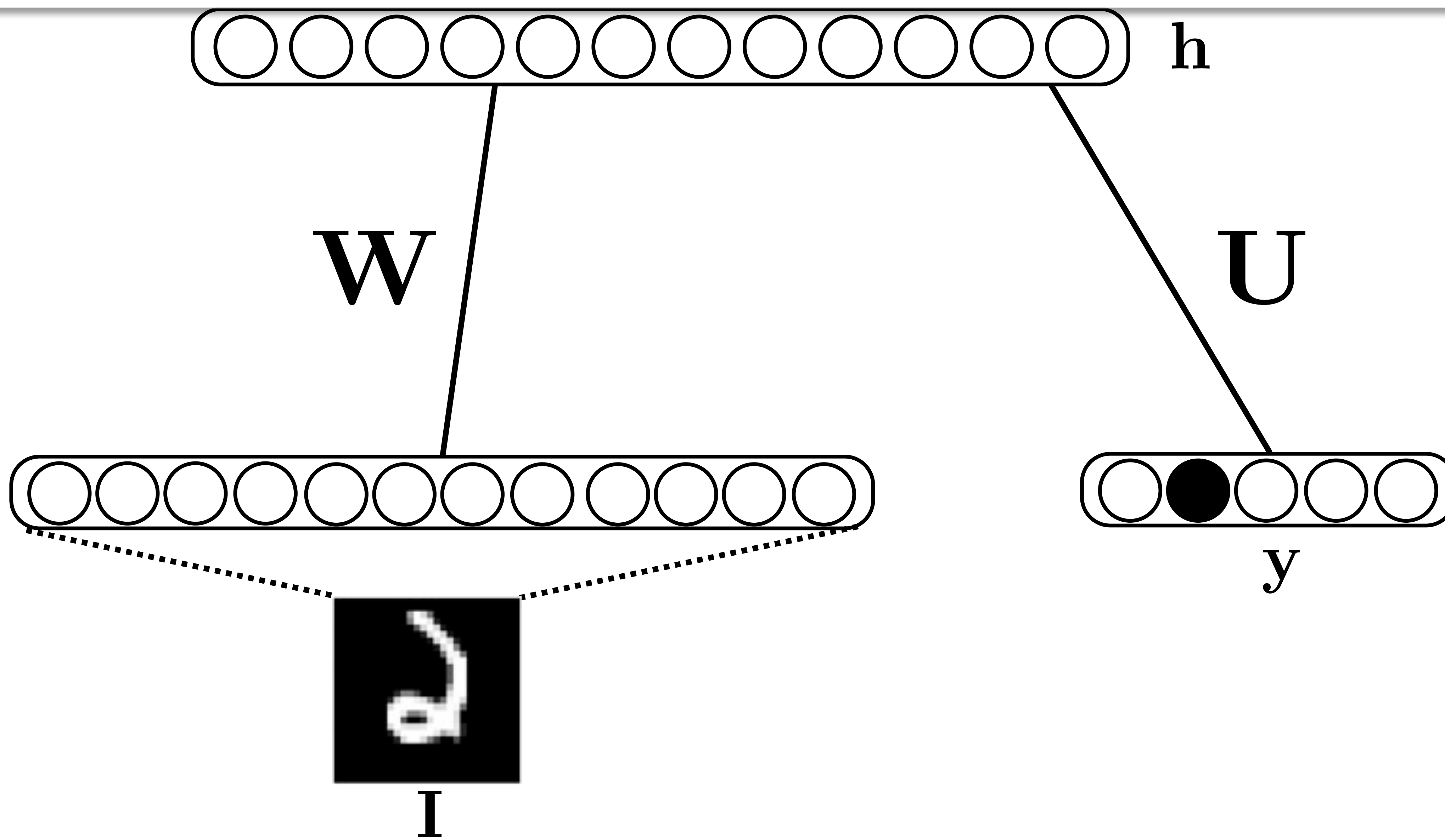
- Classification from the whole image $\mathbf{I}$



Classification is easy:

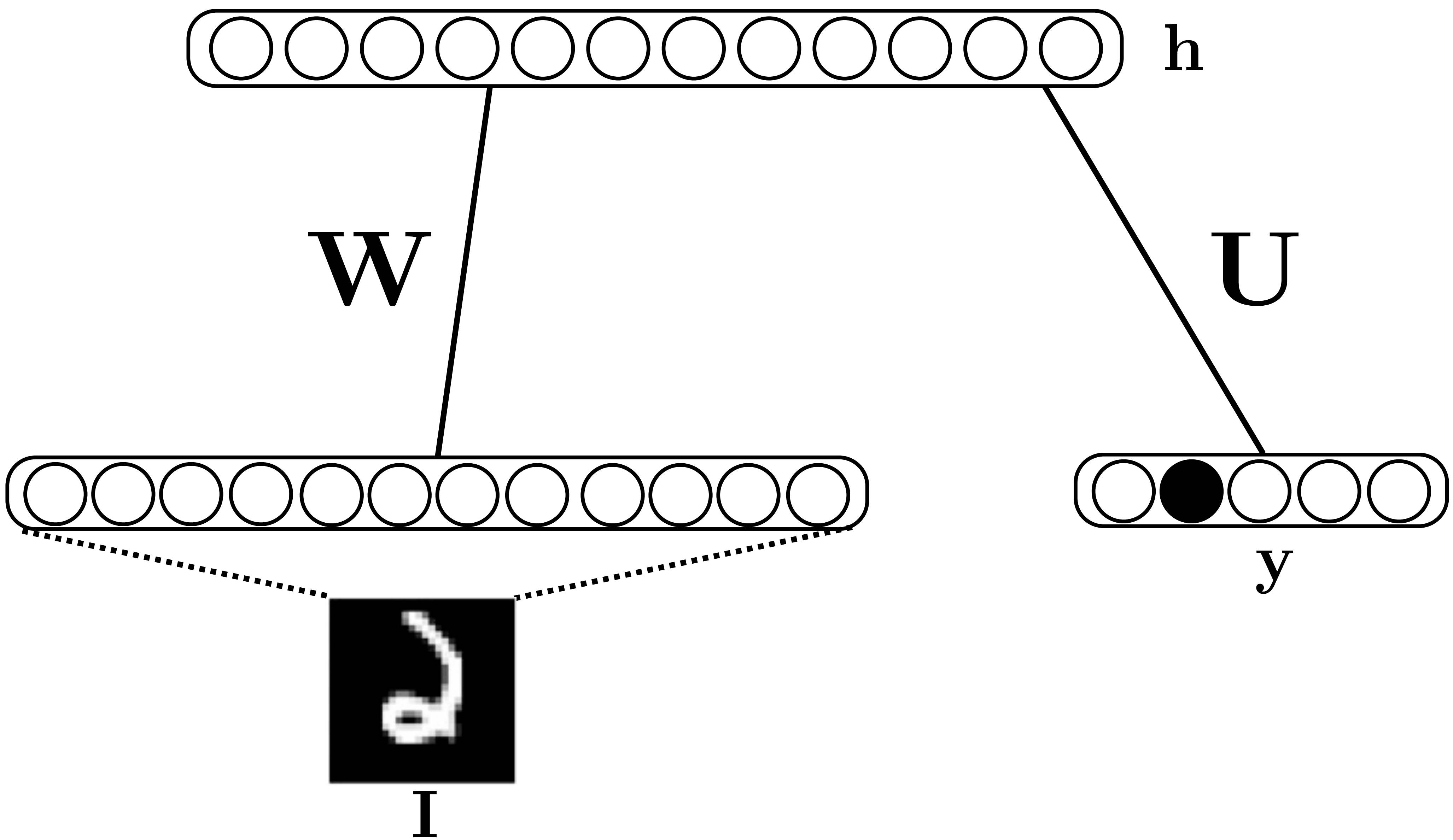
$$p(\mathbf{y}|\mathbf{I}) = \exp(-F(\mathbf{y},\mathbf{I}))/\sum_{\mathbf{y}^*}\exp(-F(\mathbf{y}^*,\mathbf{I}))$$

$$F(\mathbf{y},\mathbf{I}) = -\sum_j \log(1+\exp(\mathbf{W}\ \mathbf{I}+\mathbf{U}\ \mathbf{y}))$$



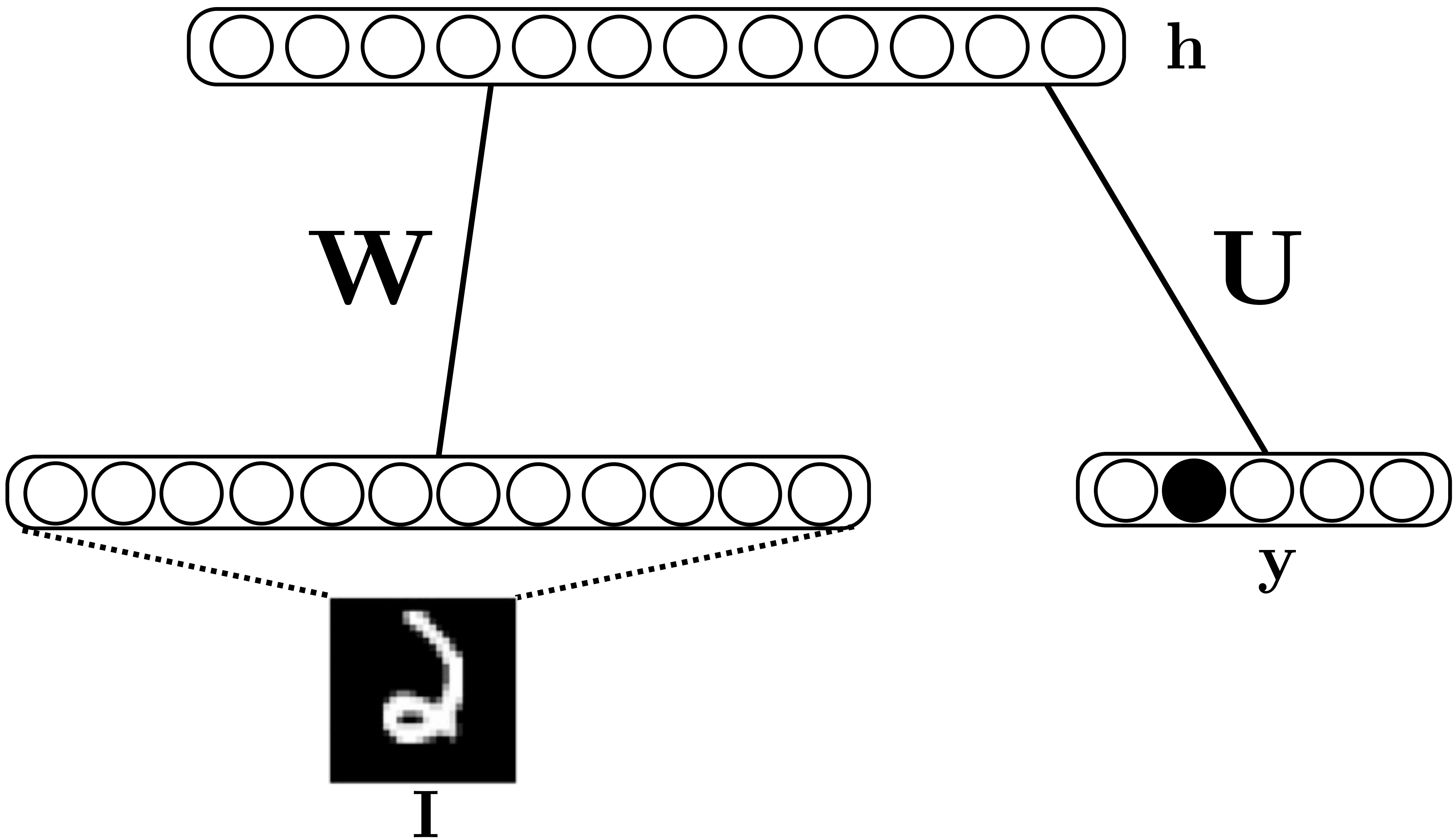
Restricted Boltzmann Machine (RBM)

- Classification from the whole image $\mathbf{I}$



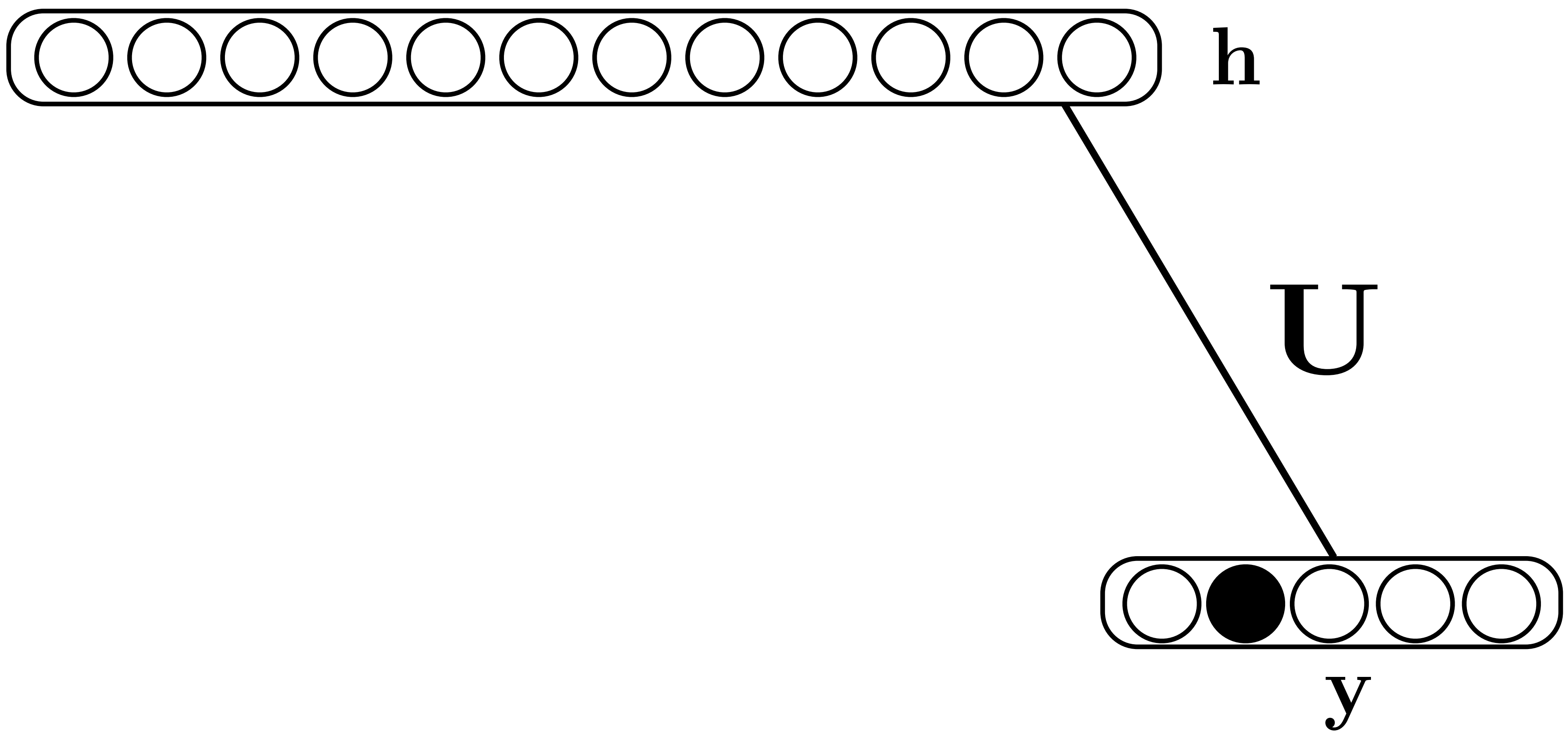
Restricted Boltzmann Machine (RBM)

- Classification from the whole image $\mathbf{I}$



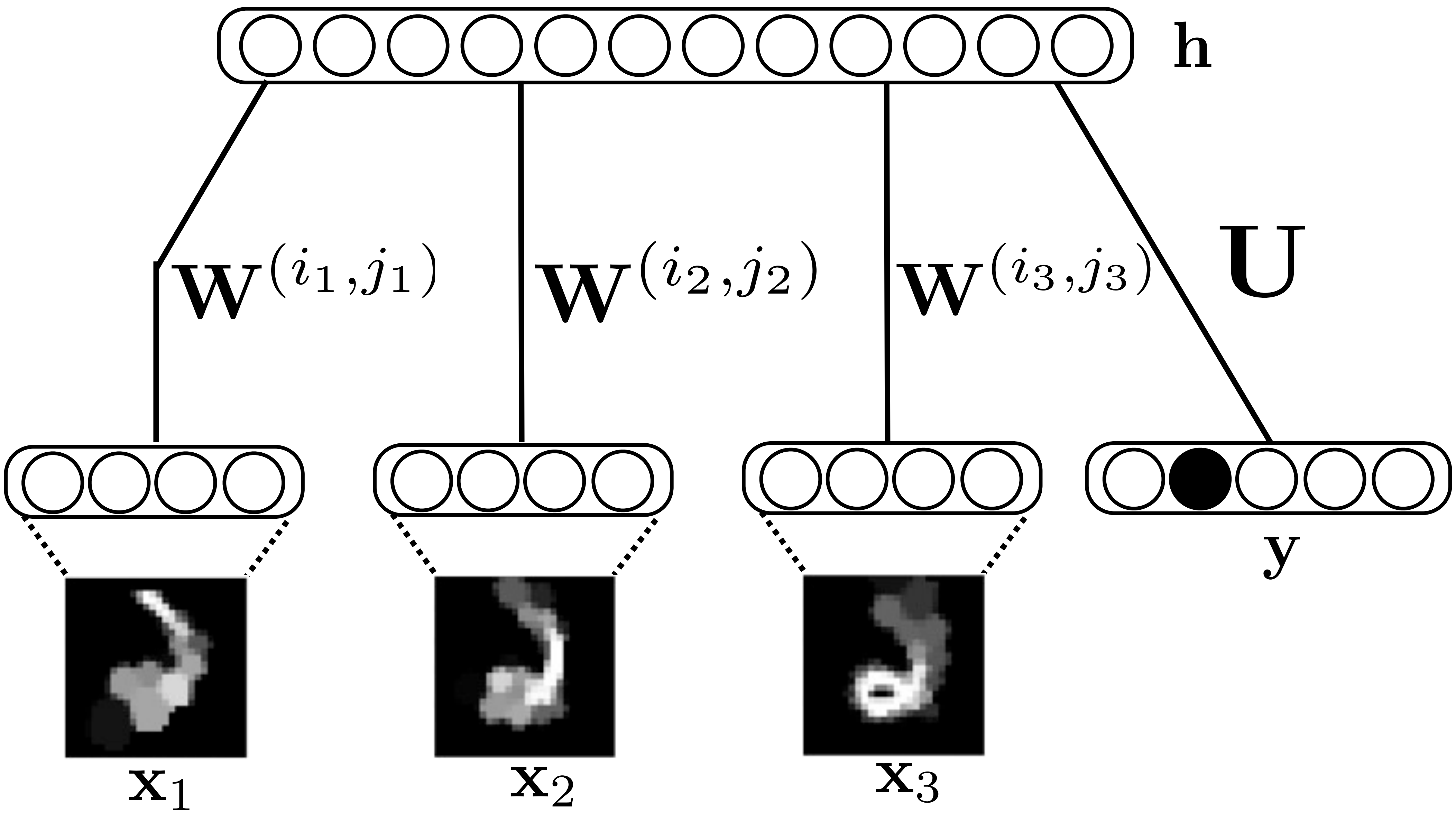
Restricted Boltzmann Machine (RBM)

- Classification from the whole image $\mathbf{I}$



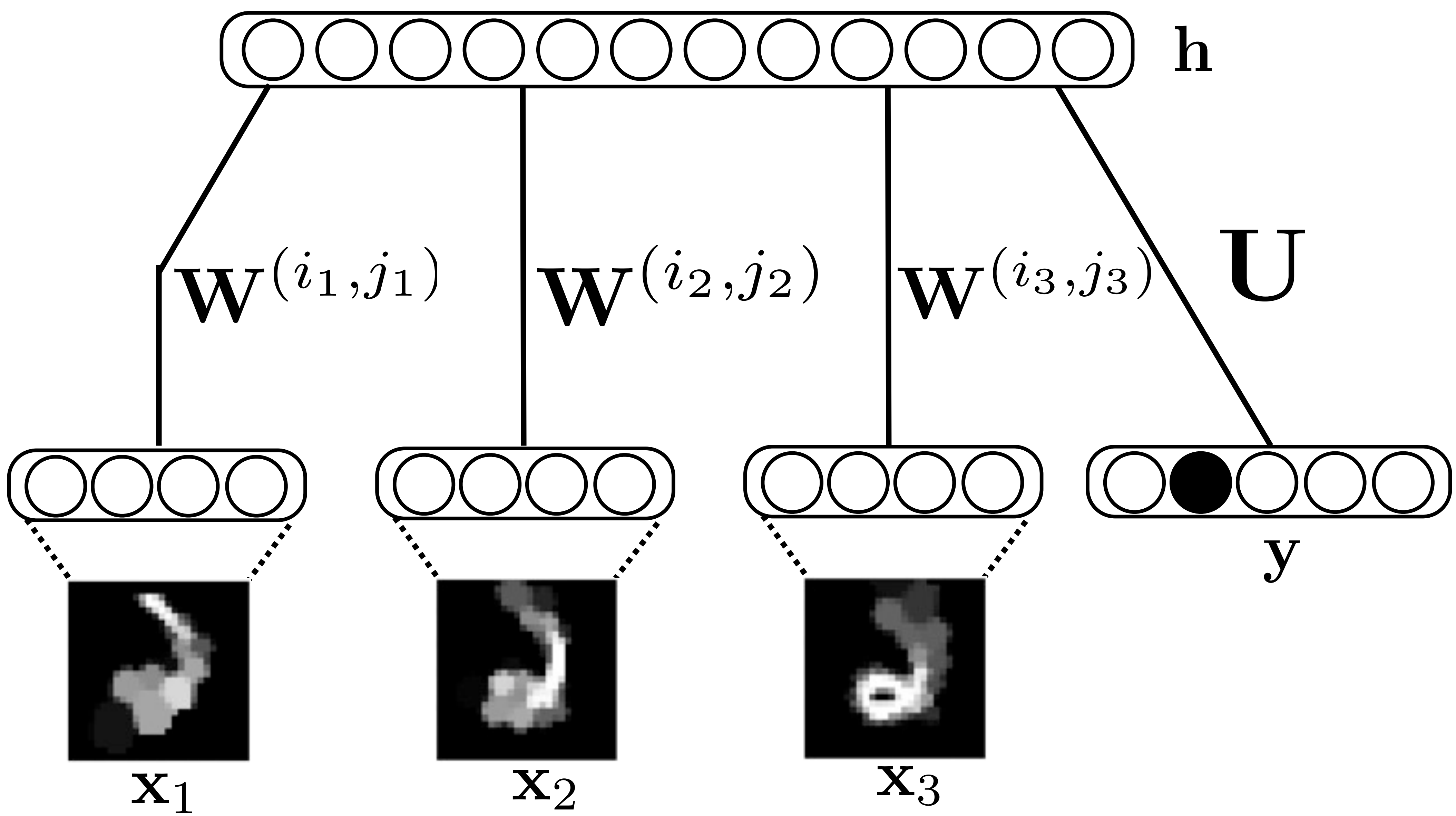
Restricted Boltzmann Machine (RBM)

- Classification from the K=3 fixations on $\mathbf{I}$



Multi-fixation RBM

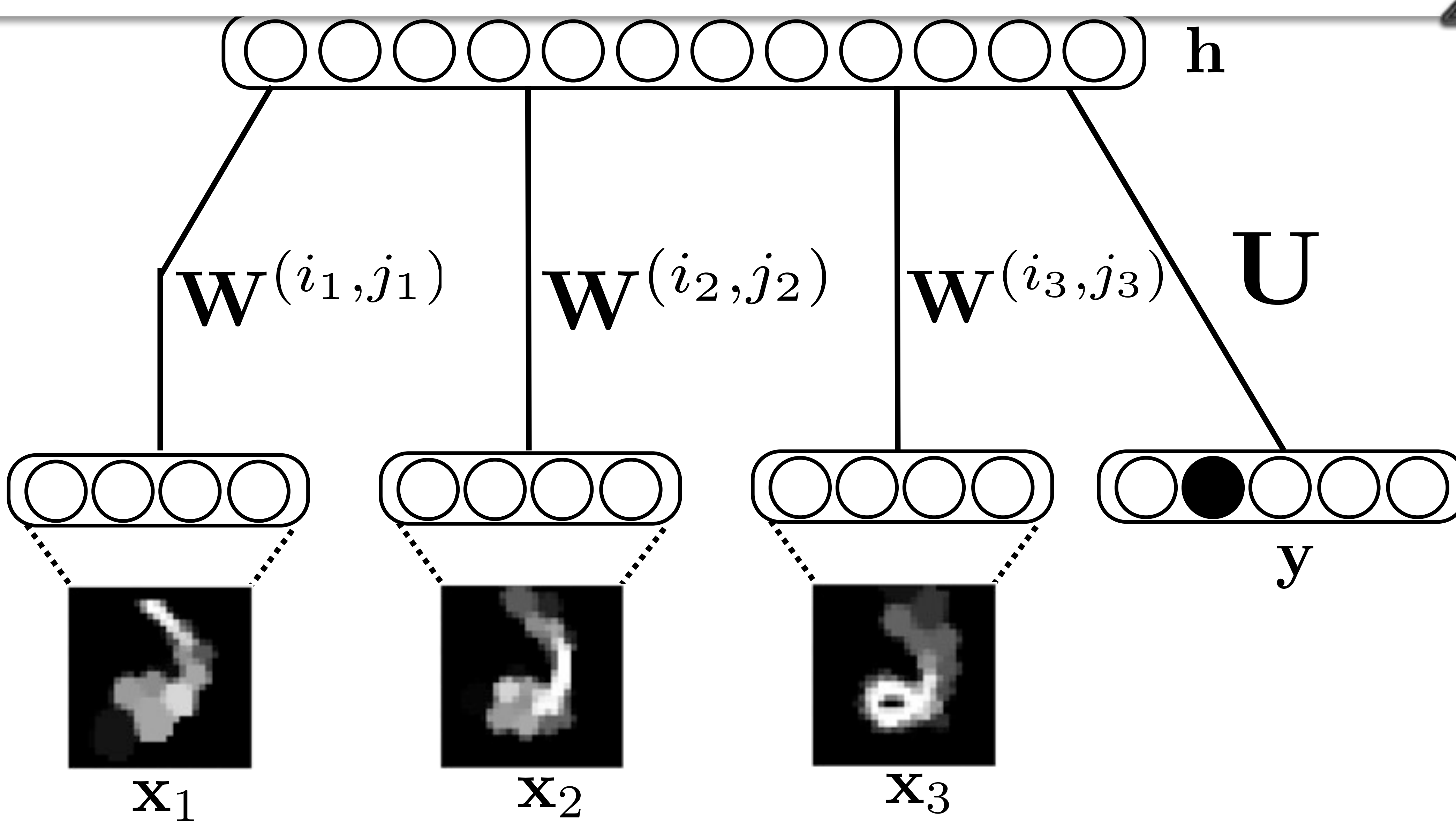
- Classification from the K=3 fixations on $\mathbf{I}$



Weight Factorization

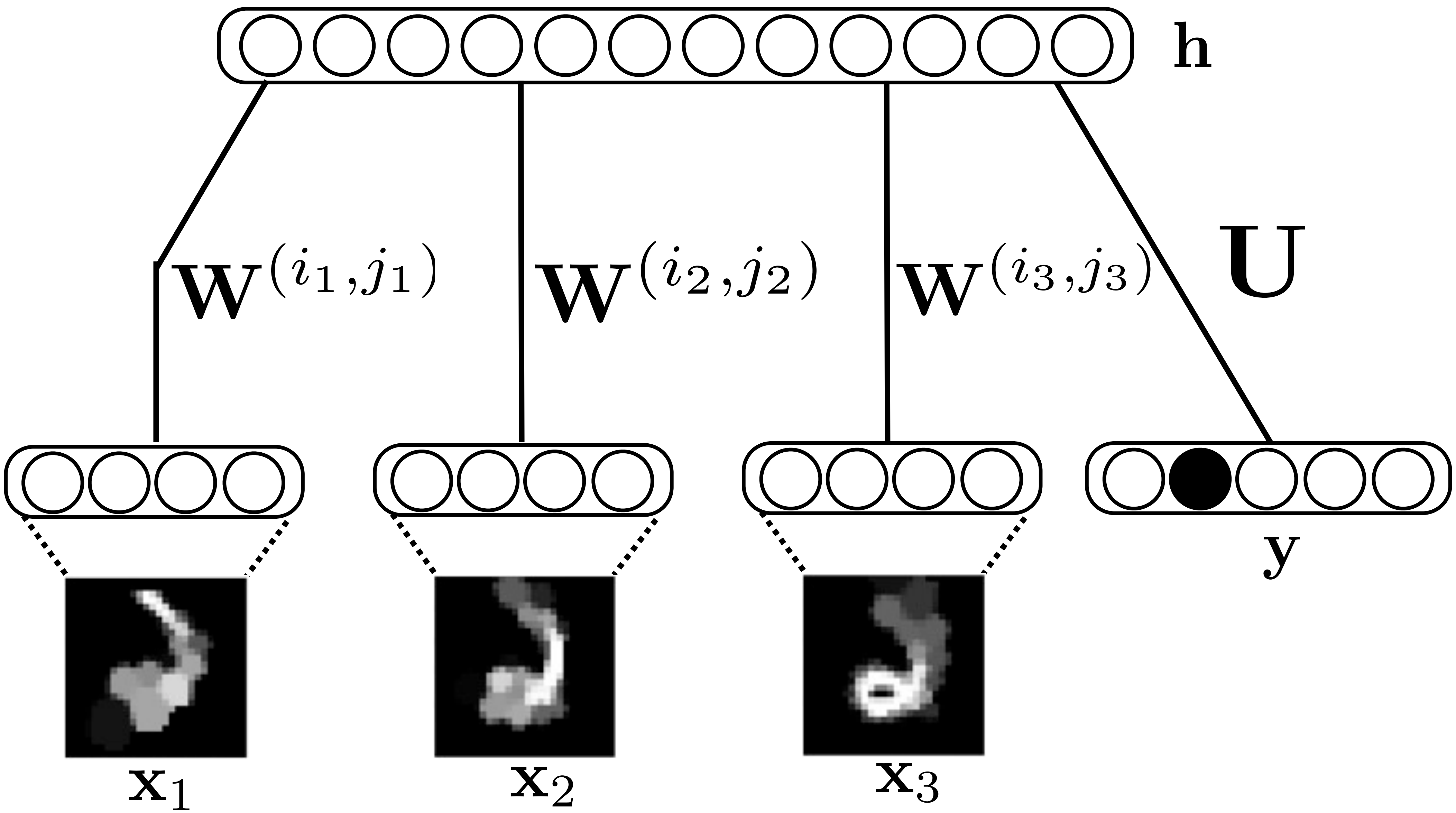
pooling gating filters

$$\mathbf{W}^{(i_k, j_k)} = \mathbf{P}\ \mathrm{diag}(\mathbf{z}^{(i_k, j_k)})\ \mathbf{F}$$



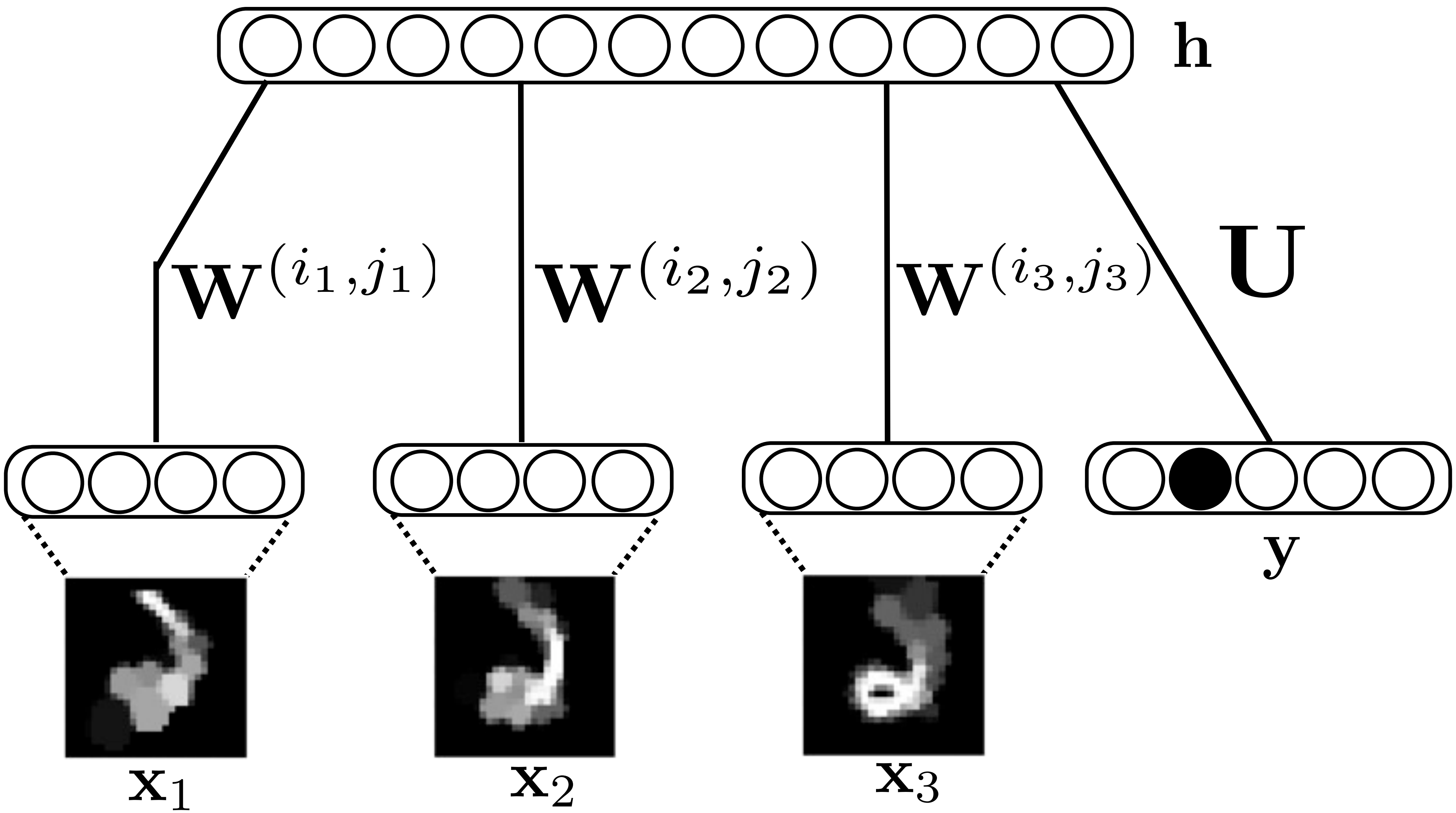
Multi-fixation RBM

- Classification from the K=3 fixations on I



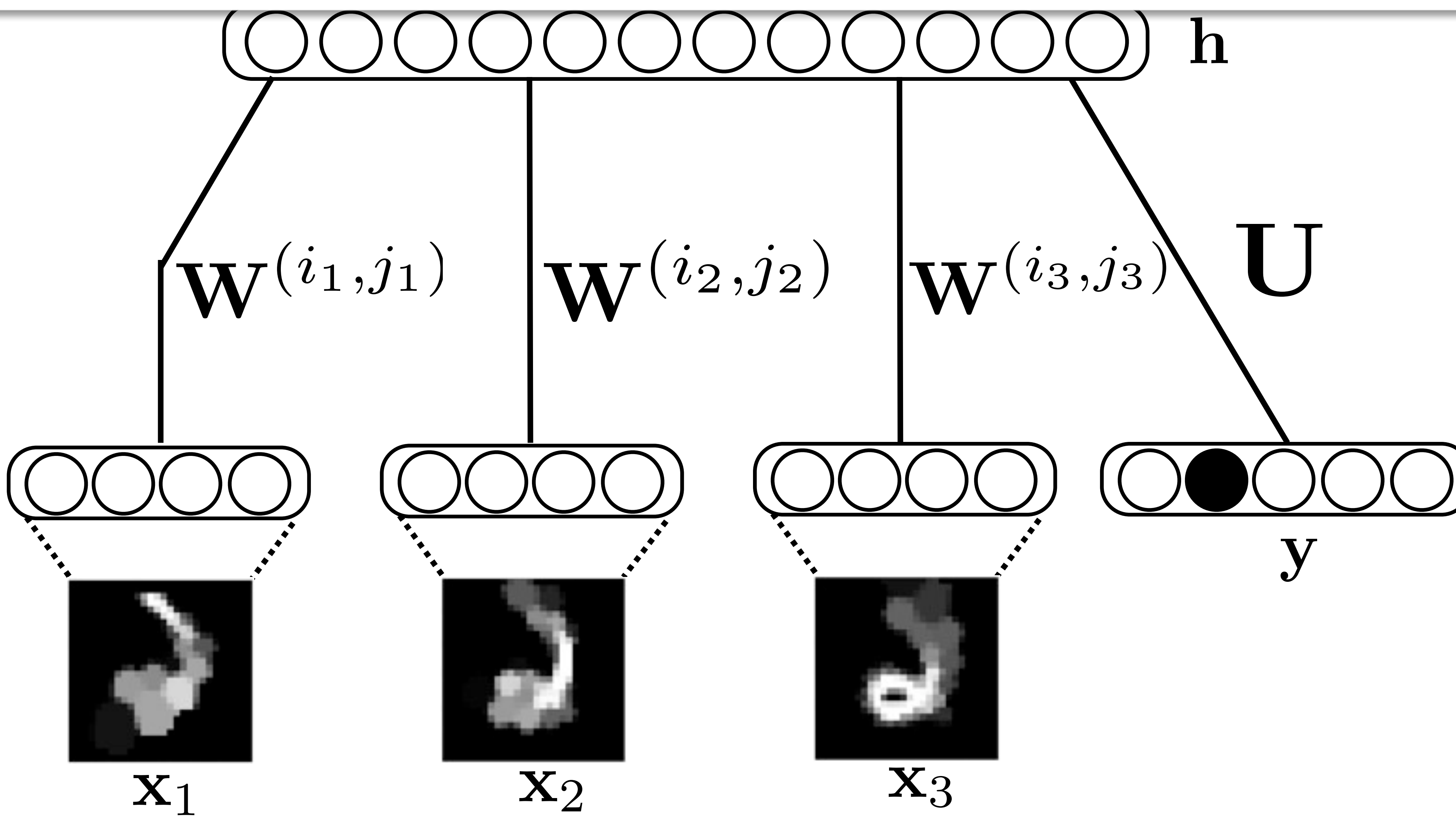
Multi-fixation RBM

- Classification from the K=3 fixations on I



Weight Factorization

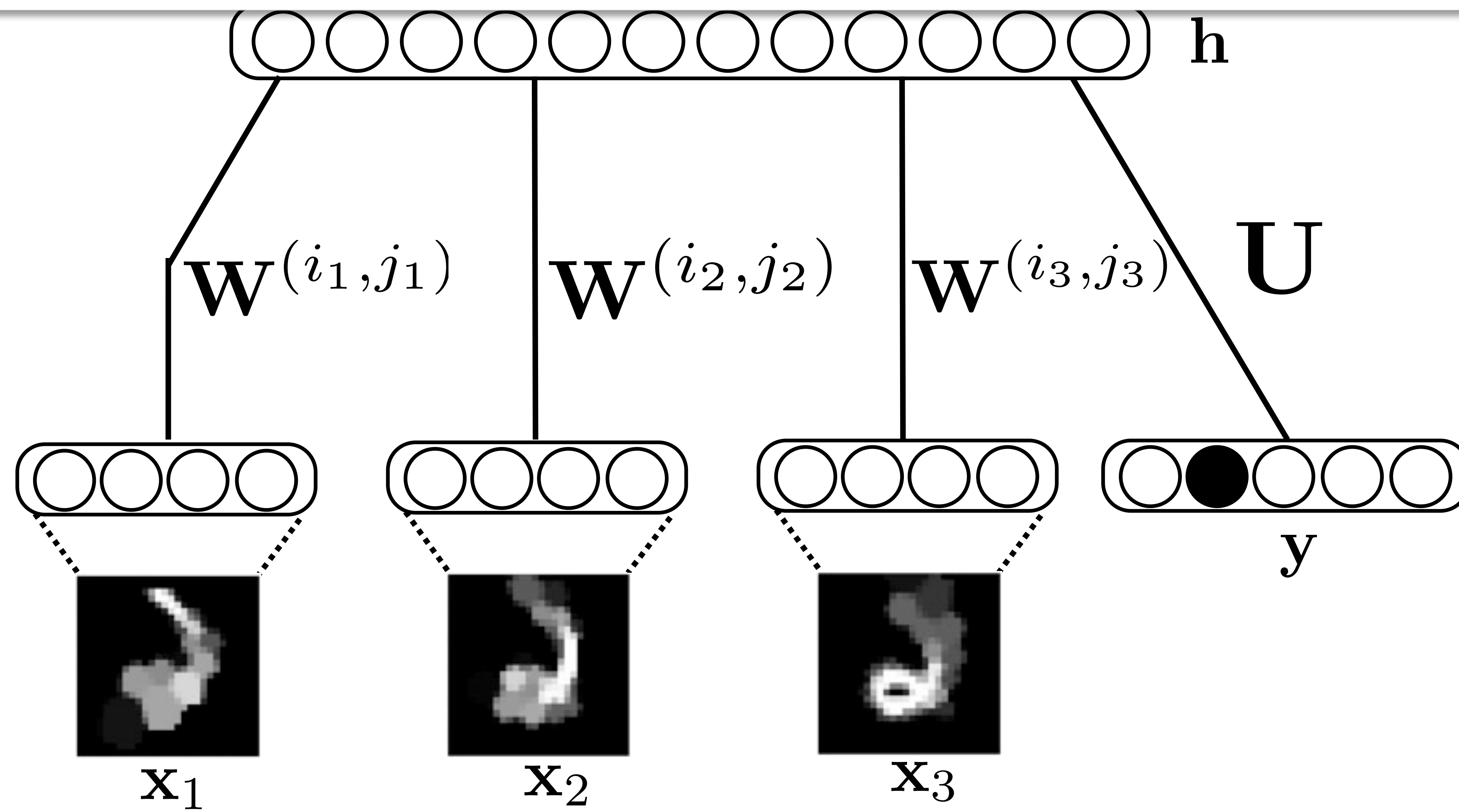
$$\mathbf{W}^{(i_1,j_1)}\ \mathbf{x}_1 = \qquad\qquad \mathbf{x}_1$$



Weight Factorization

filters

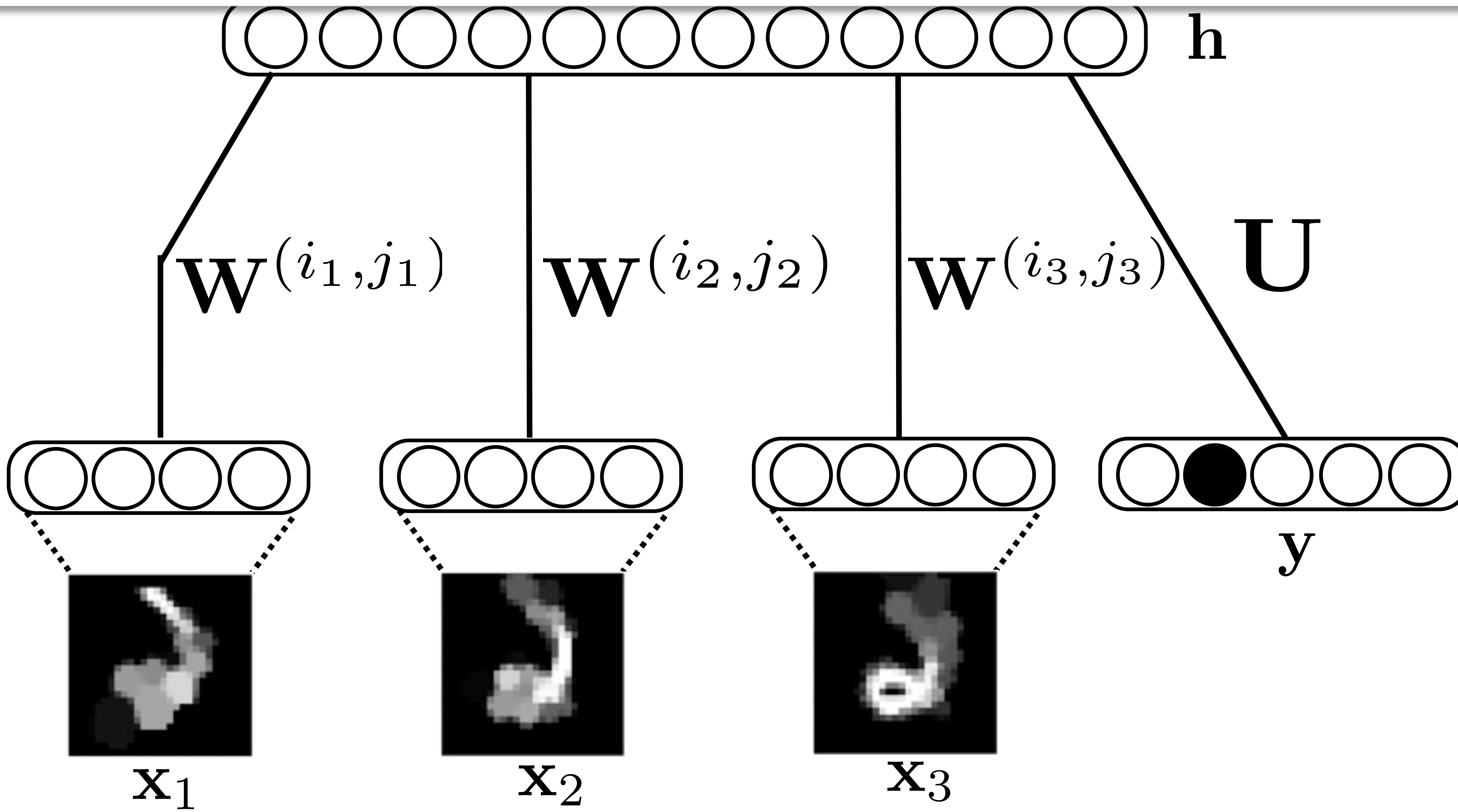
$$\mathbf{W}^{(i_1,j_1)}\ \mathbf{x}_1 = \qquad (\mathbf{F}\ \mathbf{x}_1\ )$$



Weight Factorization

gating filters

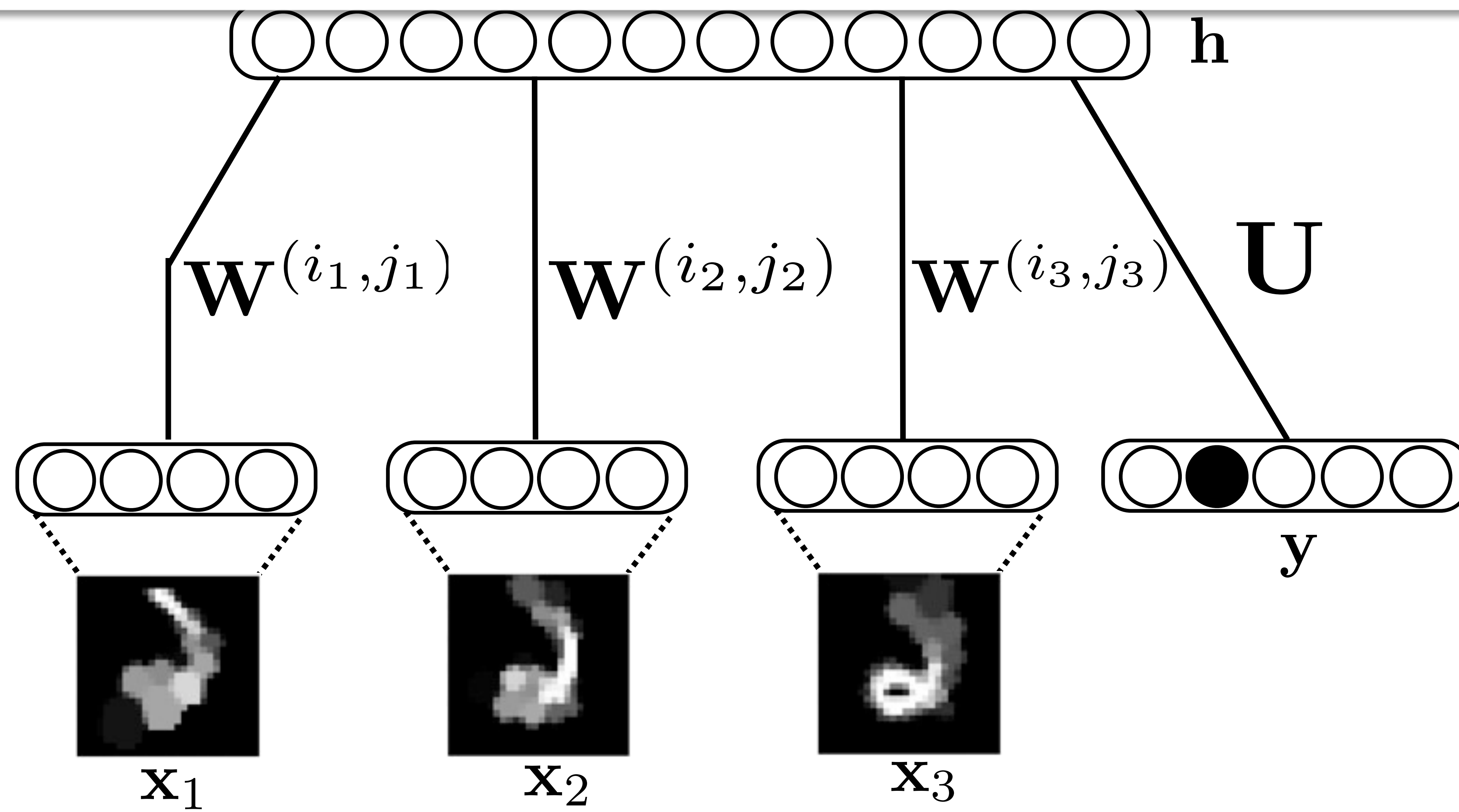
$$\mathbf{W}^{(i_1,j_1)}\,\mathbf{x}_1 = \left(\mathbf{z}^{(i_1,j_1)} \odot (\mathbf{F}\,\mathbf{x}_1)\right)$$



Weight Factorization

pooling gating filters

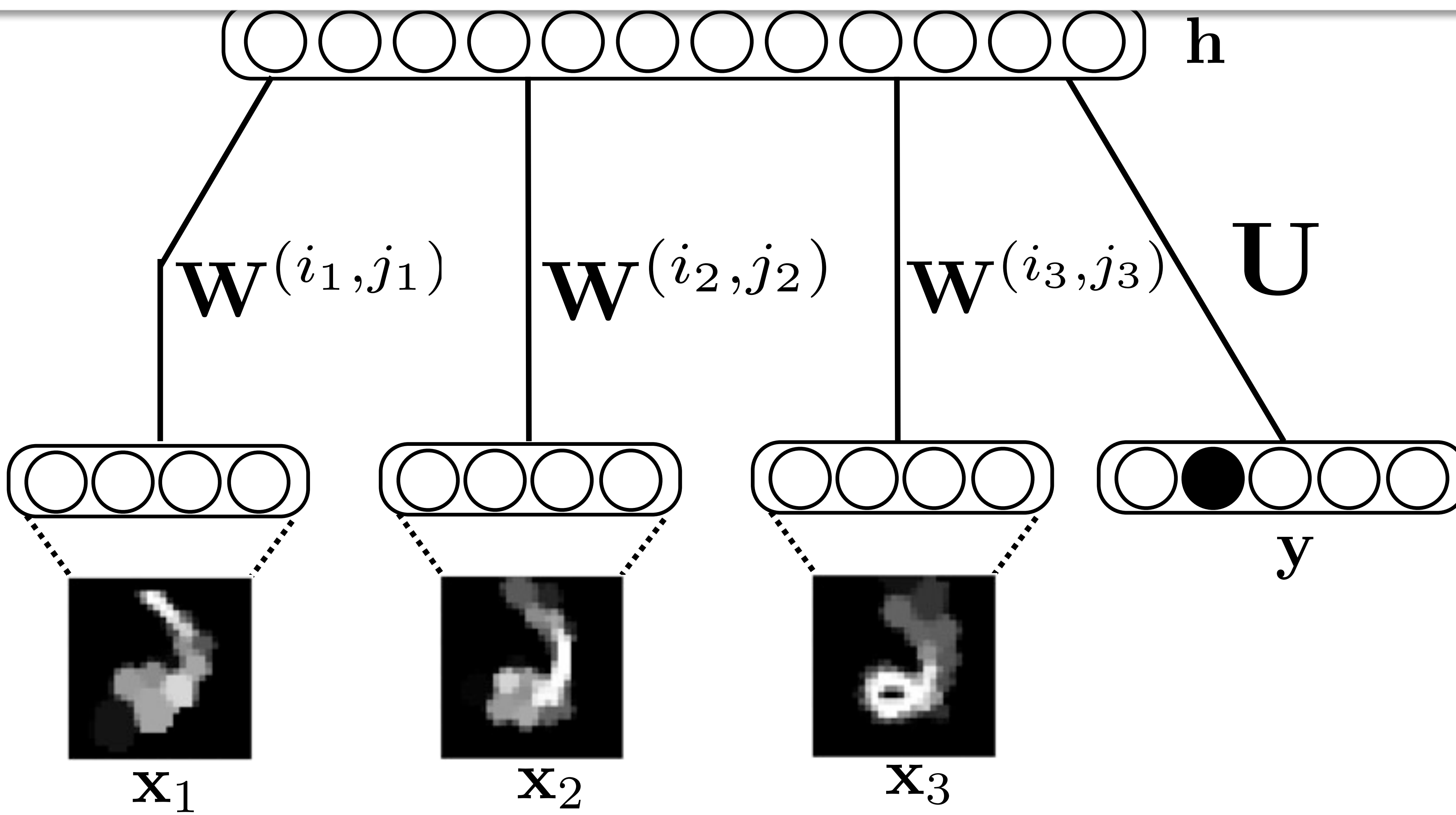
$$\mathbf{W}^{(i_1,j_1)}\ \mathbf{x}_1 = \underbrace{\mathbf{P}}_{\text{pooling}}\Big(\underbrace{\mathbf{z}^{(i_1,j_1)}}_{\text{gating}} \odot (\underbrace{\mathbf{F}}_{\text{filters}}\ \mathbf{x}_1)\Big)$$



Weight Factorization

pooling gating filters

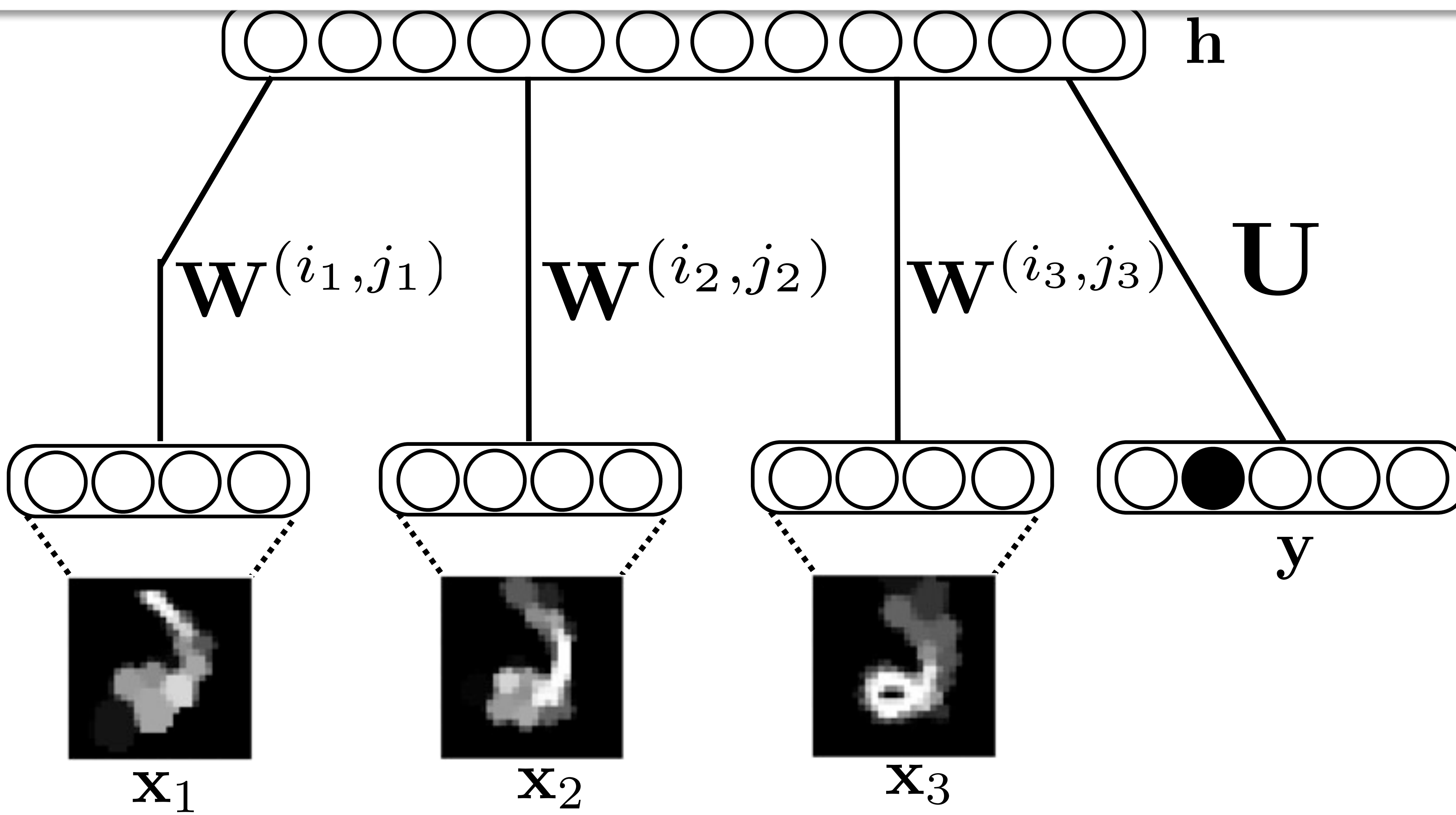
$$\mathbf{W}^{(i_2,j_2)}\,\mathbf{x}_2 = \mathbf{P}\left(\mathbf{z}^{(i_1,j_1)} \odot \left(\mathbf{F}\,\mathbf{x}_2\right)\right)$$



Weight Factorization

pooling gating filters

$$\mathbf{W}^{(i_2,j_2)}\ \mathbf{x}_2 = \mathbf{P}\left(\mathbf{z}^{(i_2,j_2)} \odot (\mathbf{F}\,\mathbf{x}_2)\right)$$



Training objectives

Discriminative: $\mathcal{C}_{\text{disc}} = -\log p(\mathbf{y}^t | \mathbf{x}^t_{1:K})$

Generative: $\mathcal{C}_{\text{gen}} = -\log p(\mathbf{y}^t, \mathbf{x}^t_{1:K})$

Hybrid: $\mathcal{C}_{\text{hybrid}} = \mathcal{C}_{\text{disc}} + \alpha \mathcal{C}_{\text{gen}}$

Bouchard & Triggs, 2004

Hybrid-sequential: $\sum_{k=1}^{K} -\log p(\mathbf{y}^t | \mathbf{x}^t_{1:k}) - \alpha \log p(\mathbf{y}^t, \mathbf{x}^t_k | \mathbf{x}^t_{1:k-1})$

Components of the system

- Recognition component (RBM)
- **Attentional component (controller)**

Where to look: learning the controller

- Given $k-1$ fixations, where should the k^{th} one be

Where to look: learning the controller

- Given $k-1$ fixations, where should the k^{th} one be

<u>Summary vector $\mathbf{s}_k$</u>

- ⋆ previous fixation positions
- ⋆ $p(h_j = 1 | \mathbf{x}_{1:k-1})$

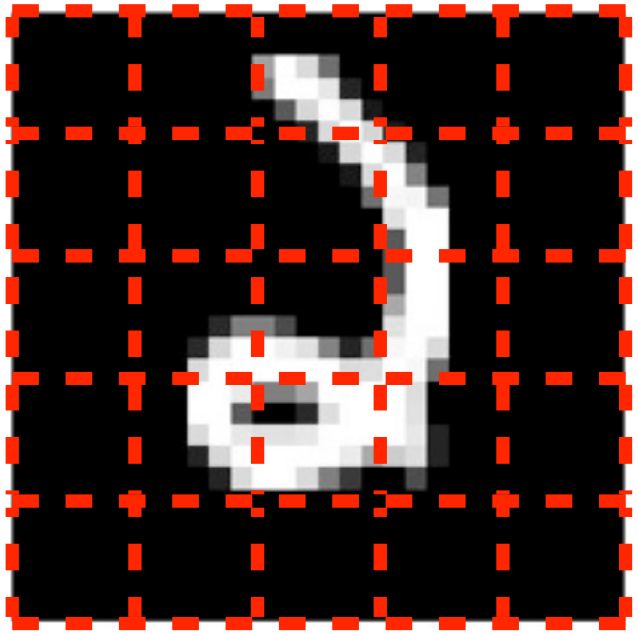
Where to look: learning the controller

- Given $k-1$ fixations, where should the k^{th} one be

<u>Summary vector $\mathbf{s}_k$</u>

★ previous fixation positions

★ $p(h_j = 1|\mathbf{x}_{1:k-1})$



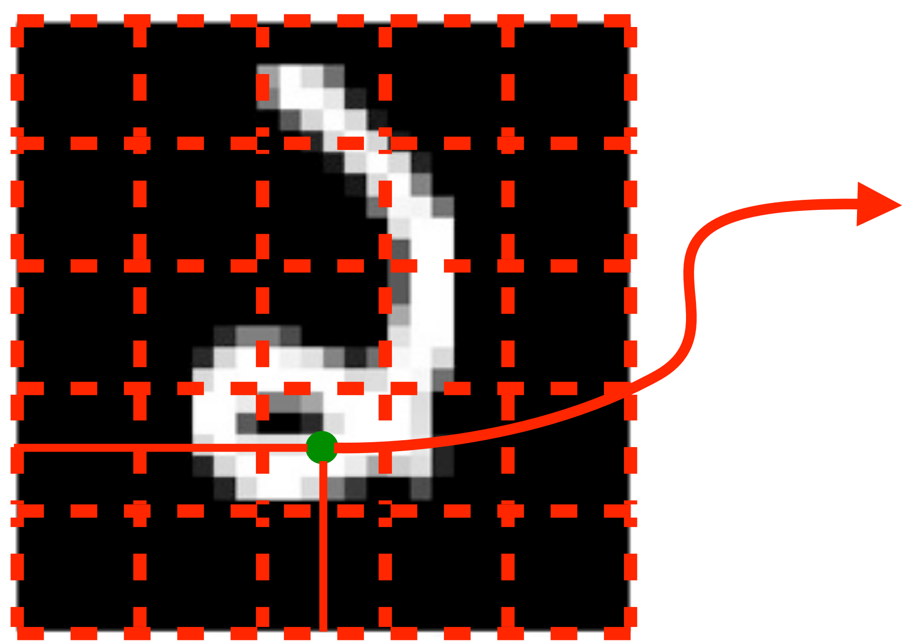
Where to look: learning the controller

- Given $k-1$ fixations, where should the k^{th} one be

<u>Summary vector $\mathbf{s}_k$</u>

★ previous fixation positions

★ $p(h_j = 1 | \mathbf{x}_{1:k-1})$



<u>Scoring function</u>

$$f(\mathbf{s}_k, (i_k, j_k))$$

$$||$$

$$\mathbf{v}^{(i_k, j_k)^\top} \mathbf{s}_k$$

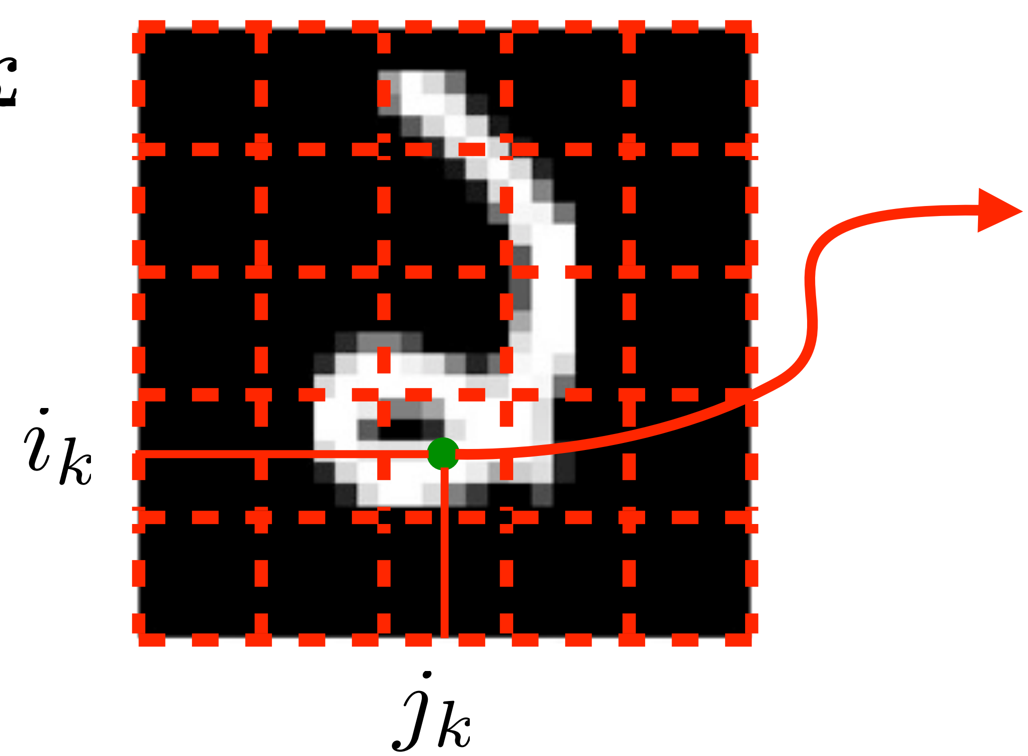
Where to look: learning the controller

- Given $k-1$ fixations, where should the k^{th} one be

<u>Summary vector $\mathbf{s}_k$</u>

★ previous fixation positions

★ $p(h_j = 1 | \mathbf{x}_{1:k-1})$



<u>Scoring function</u>

$$f(\mathbf{s}_k, (i_k, j_k)) = \mathbf{v}^{(i_k, j_k)^\top} \mathbf{s}_k$$

- Training objective of scoring function:

$$|f(\mathbf{s}_k, (i_k, j_k)) - \log p(\mathbf{y} | \mathbf{x}_{1:k-1}, \mathbf{x}_k)|$$

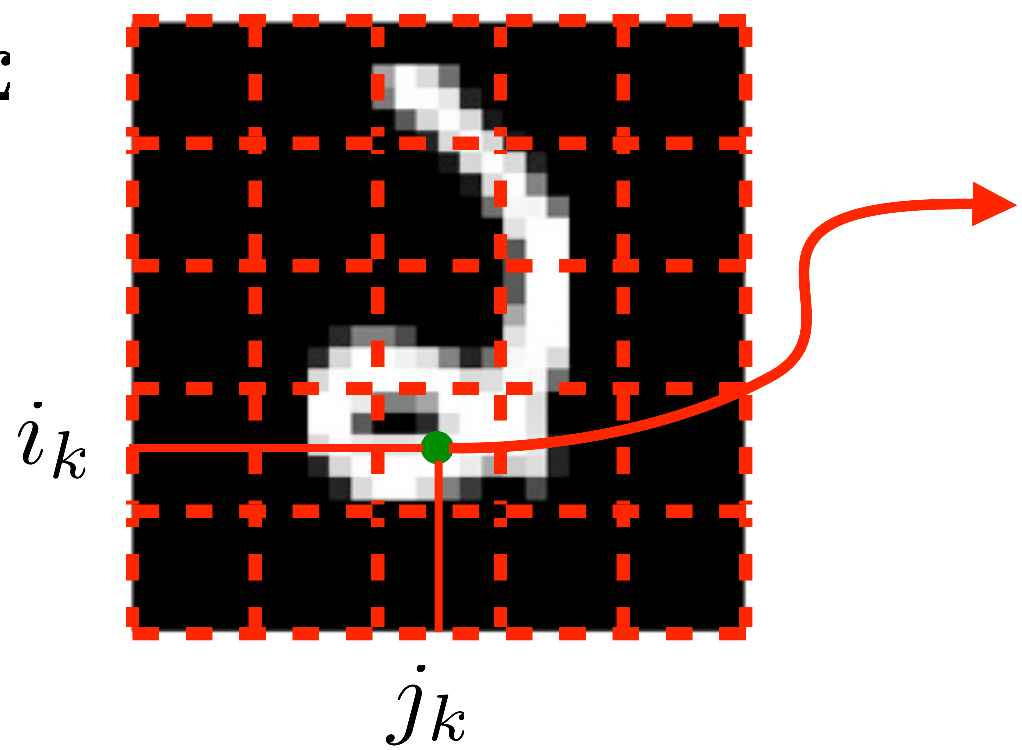
Where to look: learning the controller

- Given $k-1$ fixations, where should the k^{th} one be

<u>Summary vector $\mathbf{s}_k$</u>

★ previous fixation positions

★ $p(h_j = 1 | \mathbf{x}_{1:k-1})$



<u>Scoring function</u>

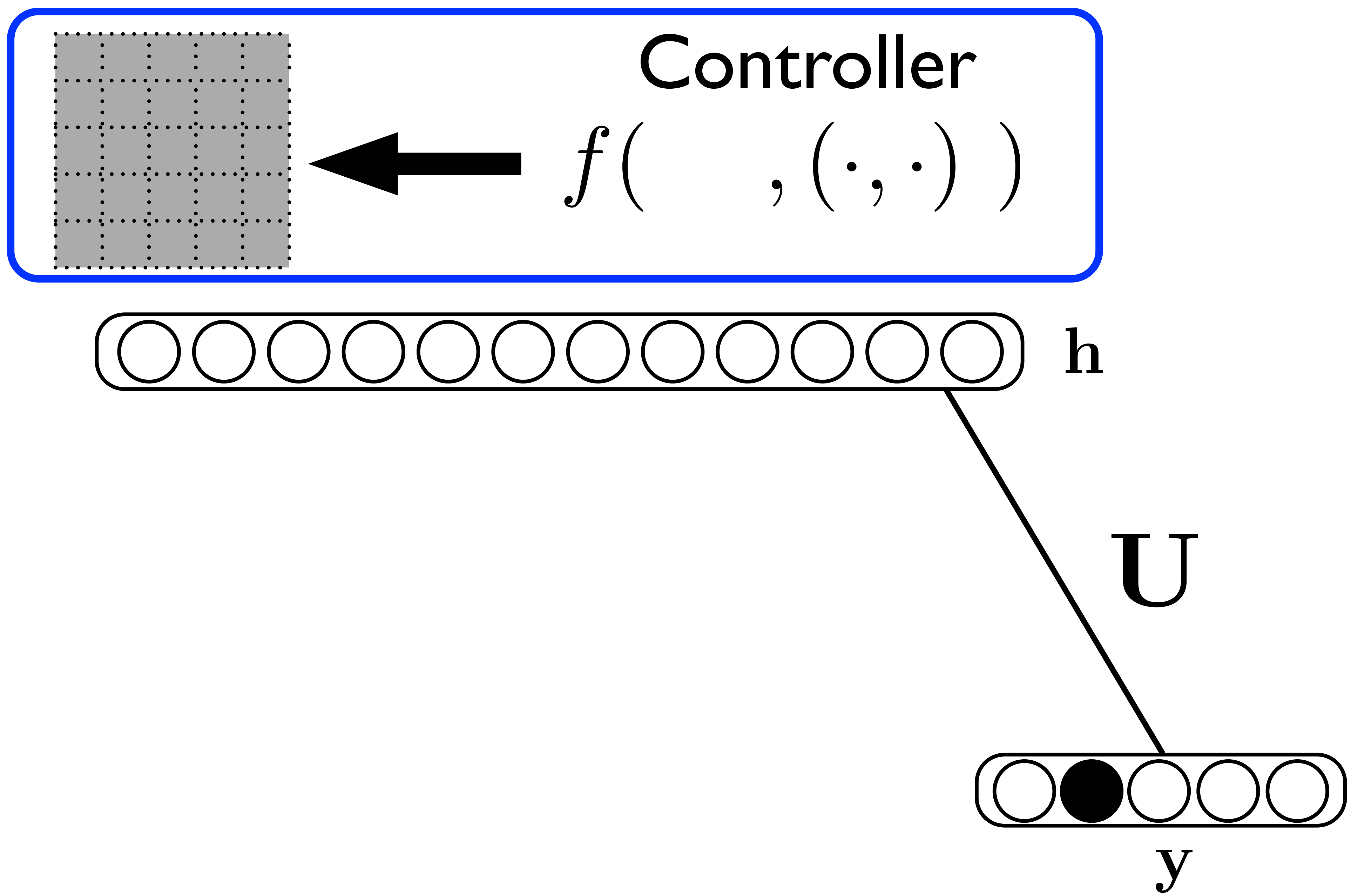
$$f(\mathbf{s}_k, (i_k, j_k))$$
$$||$$
$$\mathbf{v}^{(i_k, j_k)^\top} \mathbf{s}_k$$

- Training objective of scoring function:

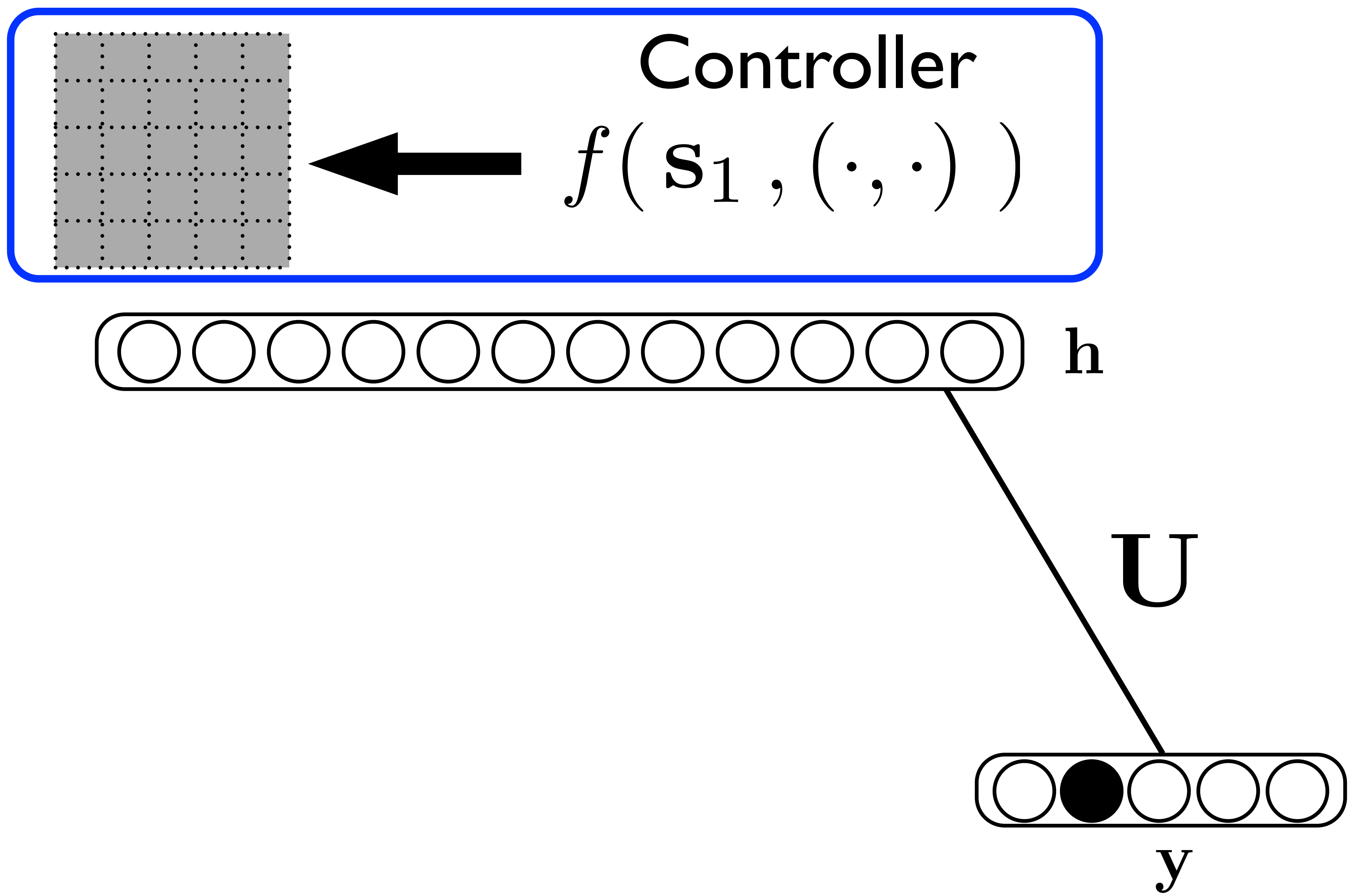
$$|f(\mathbf{s}_k, (i_k, j_k)) - \log p(\mathbf{y}|\mathbf{x}_{1:k-1}, \mathbf{x}_k)|$$

- Controller distribution: $\exp(\ f(\mathbf{s}_k, (i_k, j_k))\ )/Z(\mathbf{s}_k)$

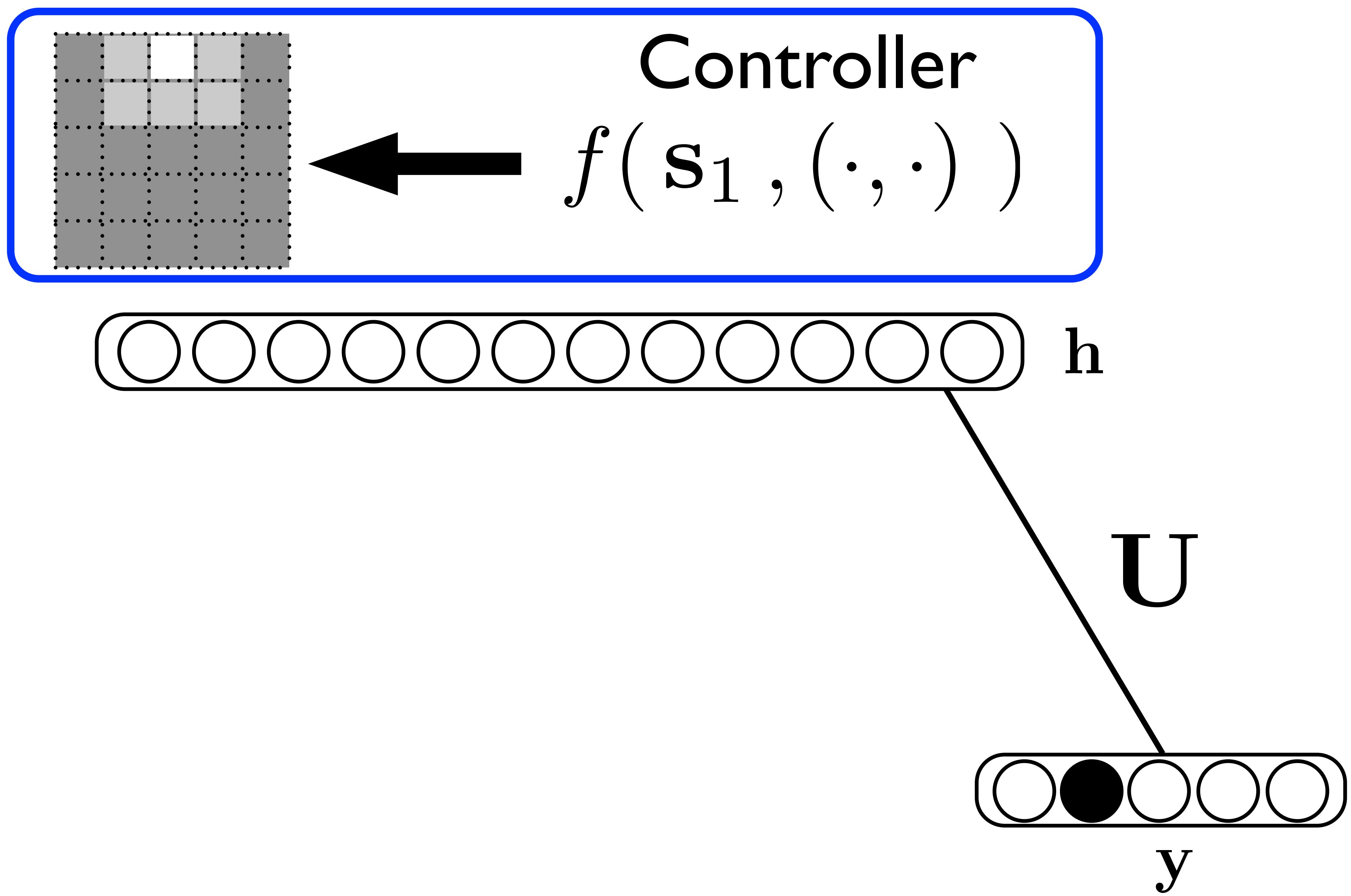
Putting it all together



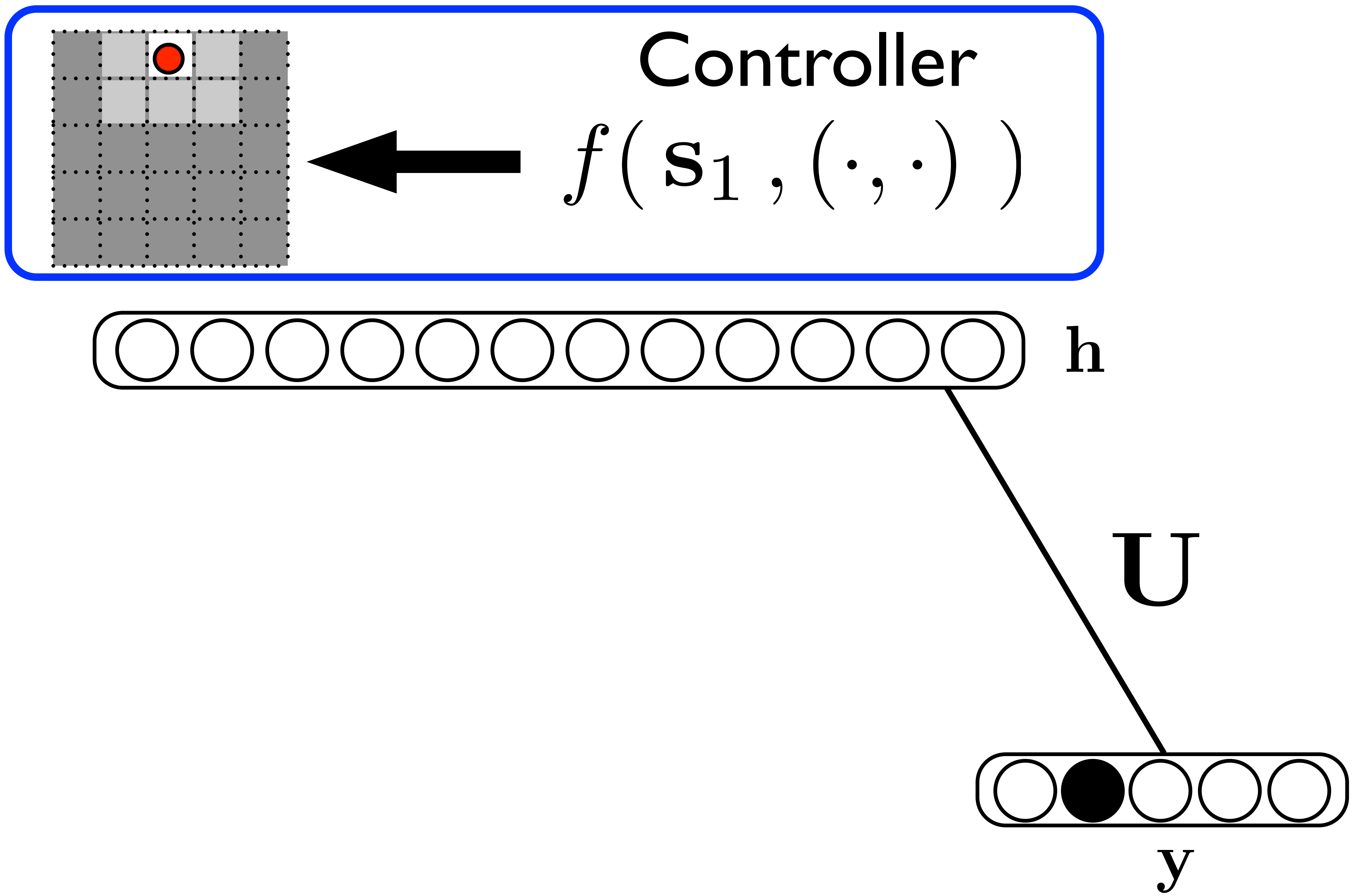
Putting it all together



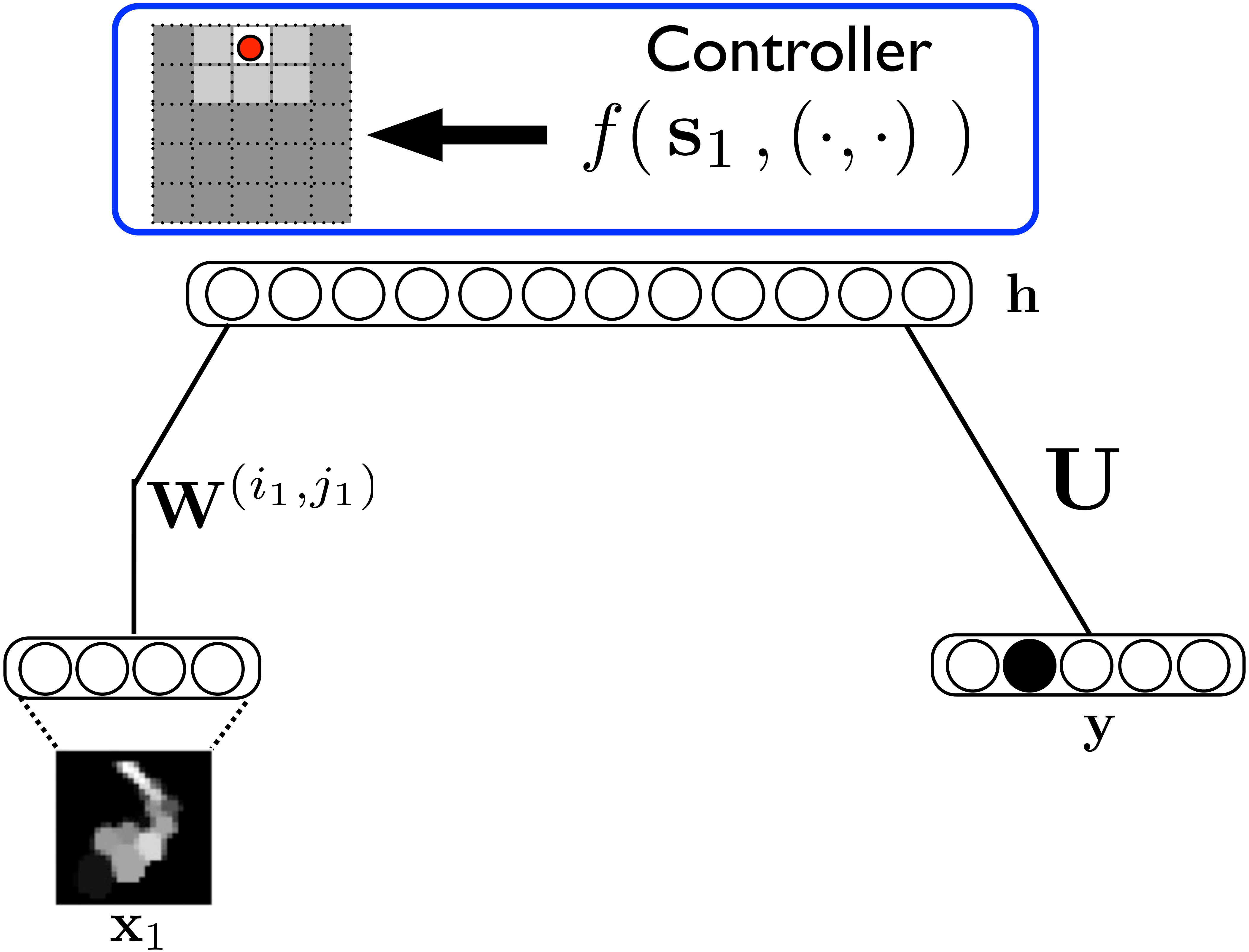
Putting it all together



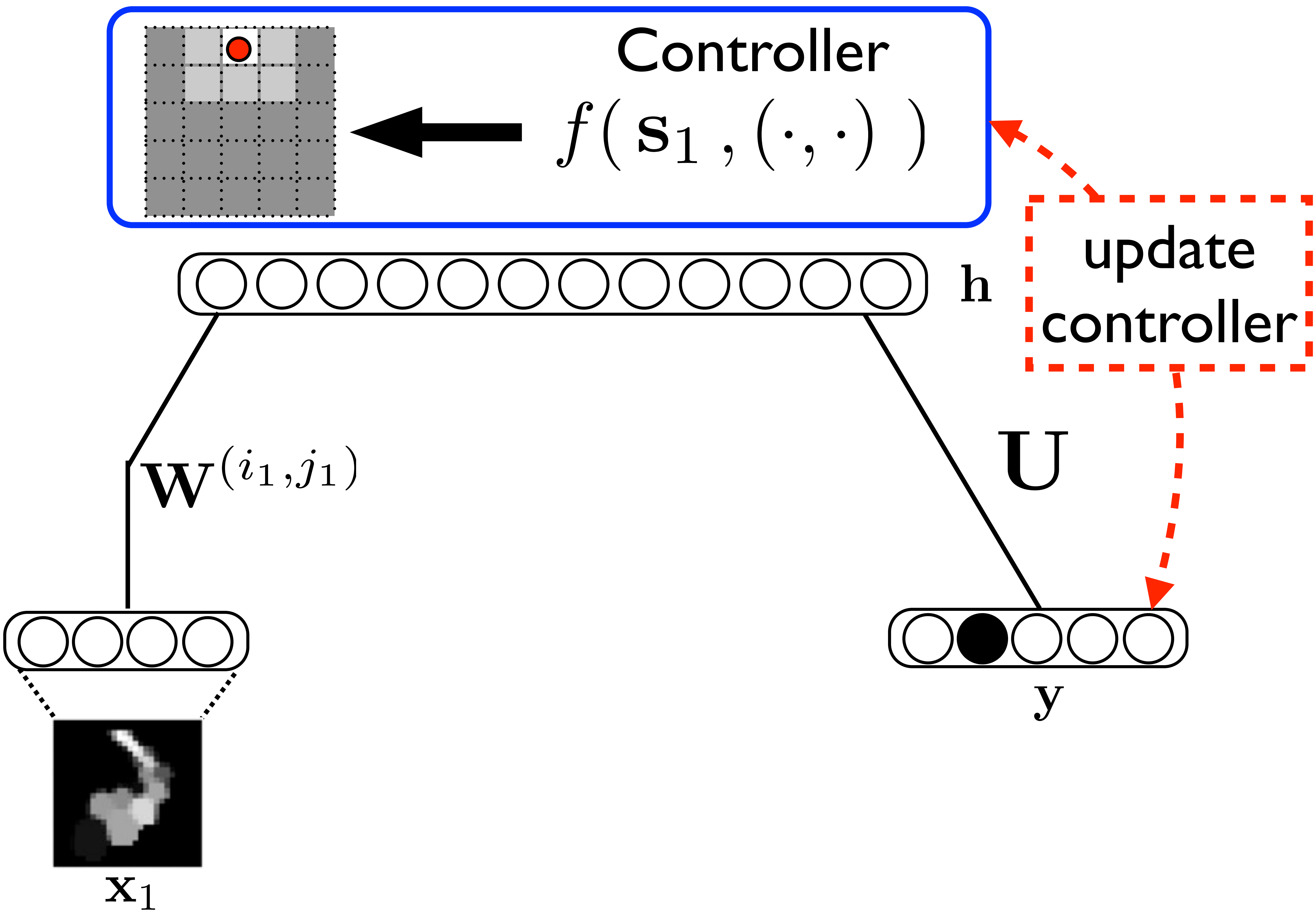
Putting it all together



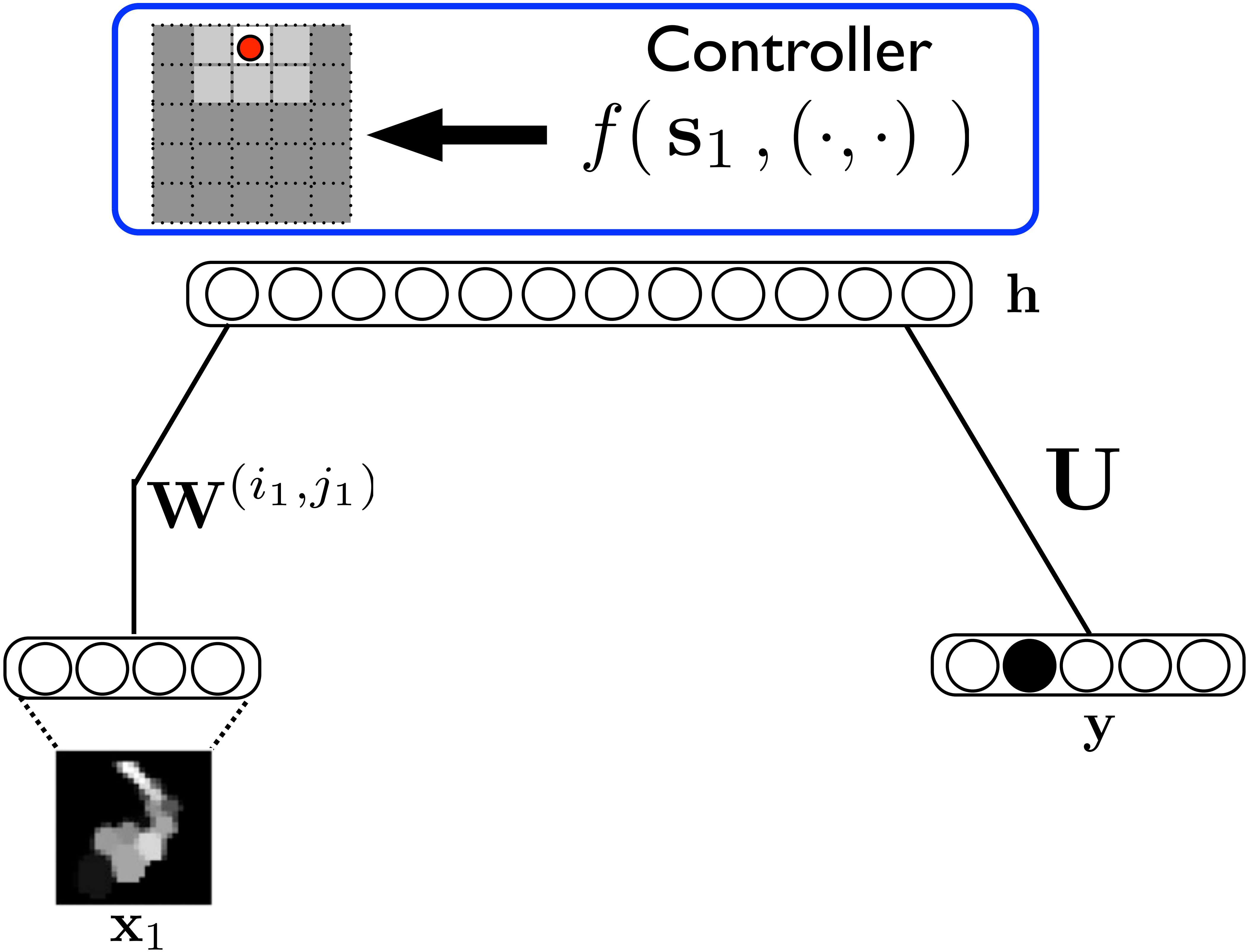
Putting it all together



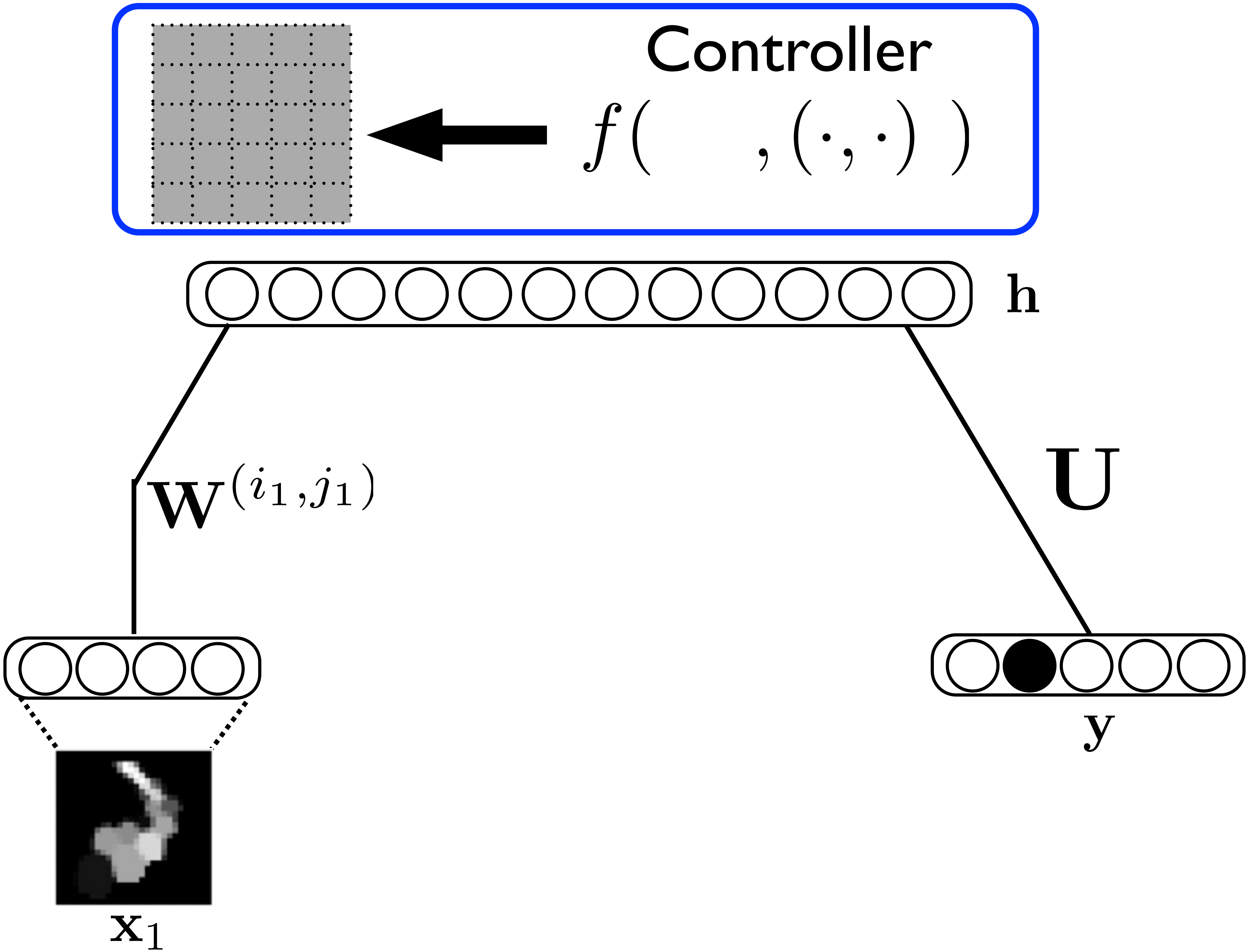
Putting it all together



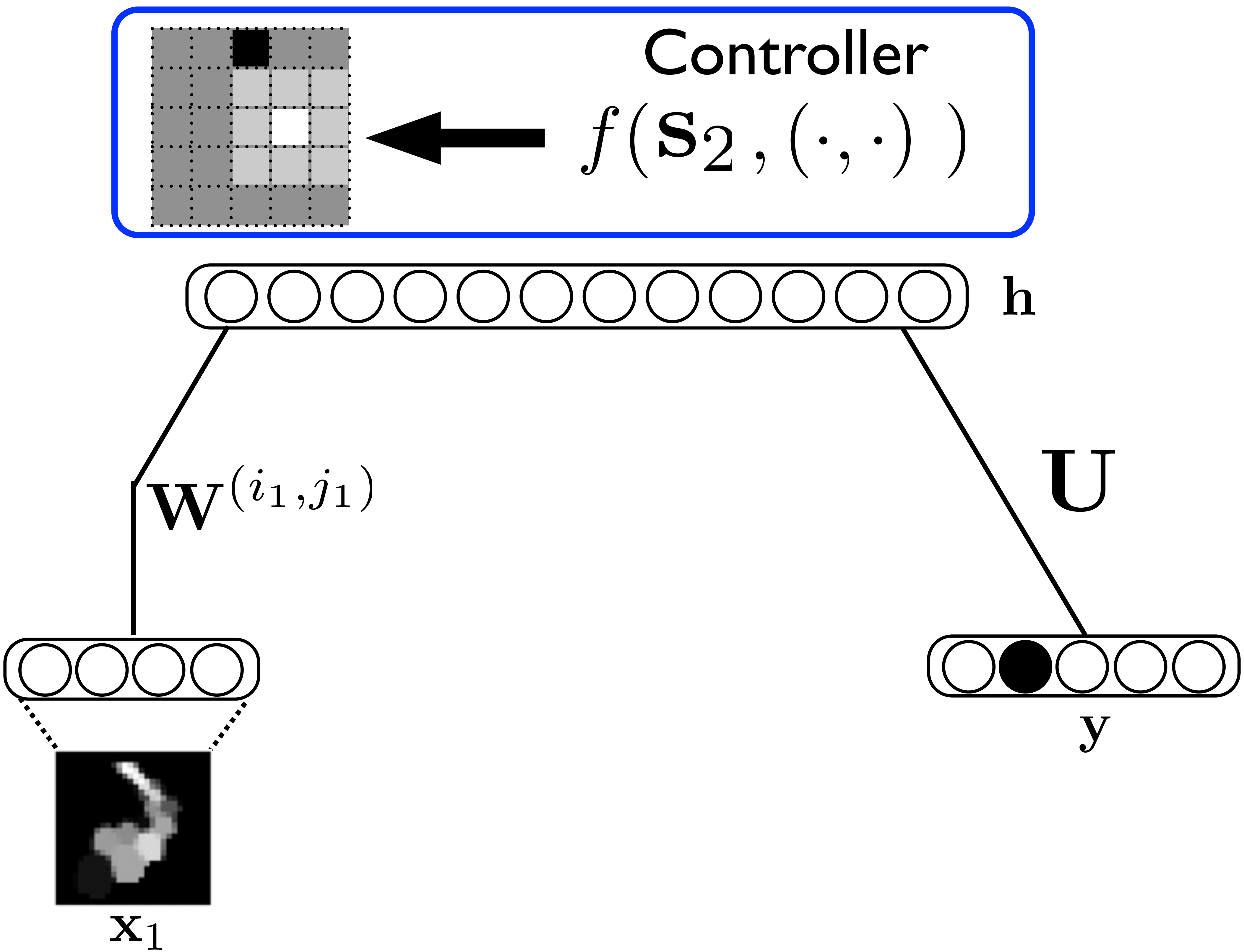
Putting it all together



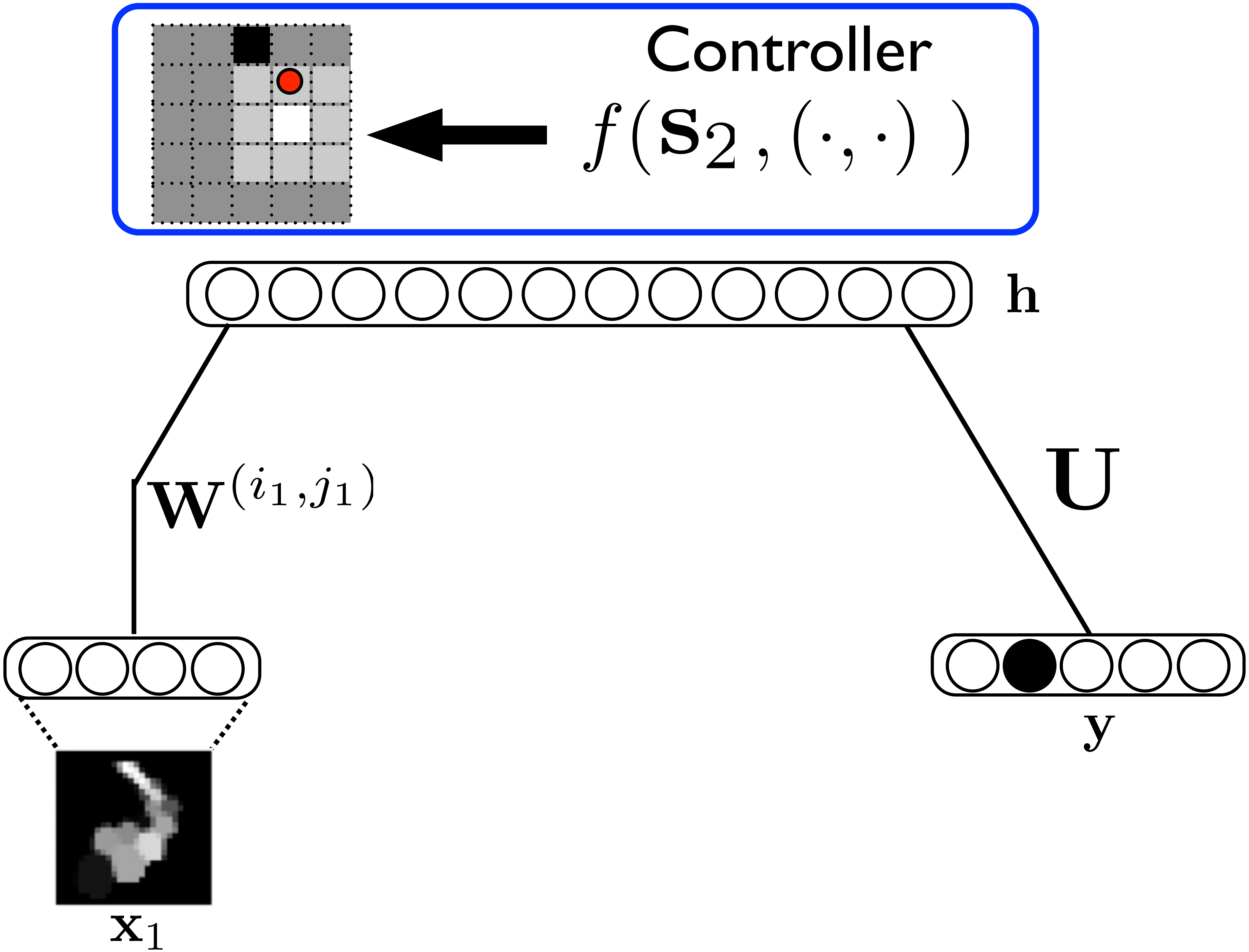
Putting it all together



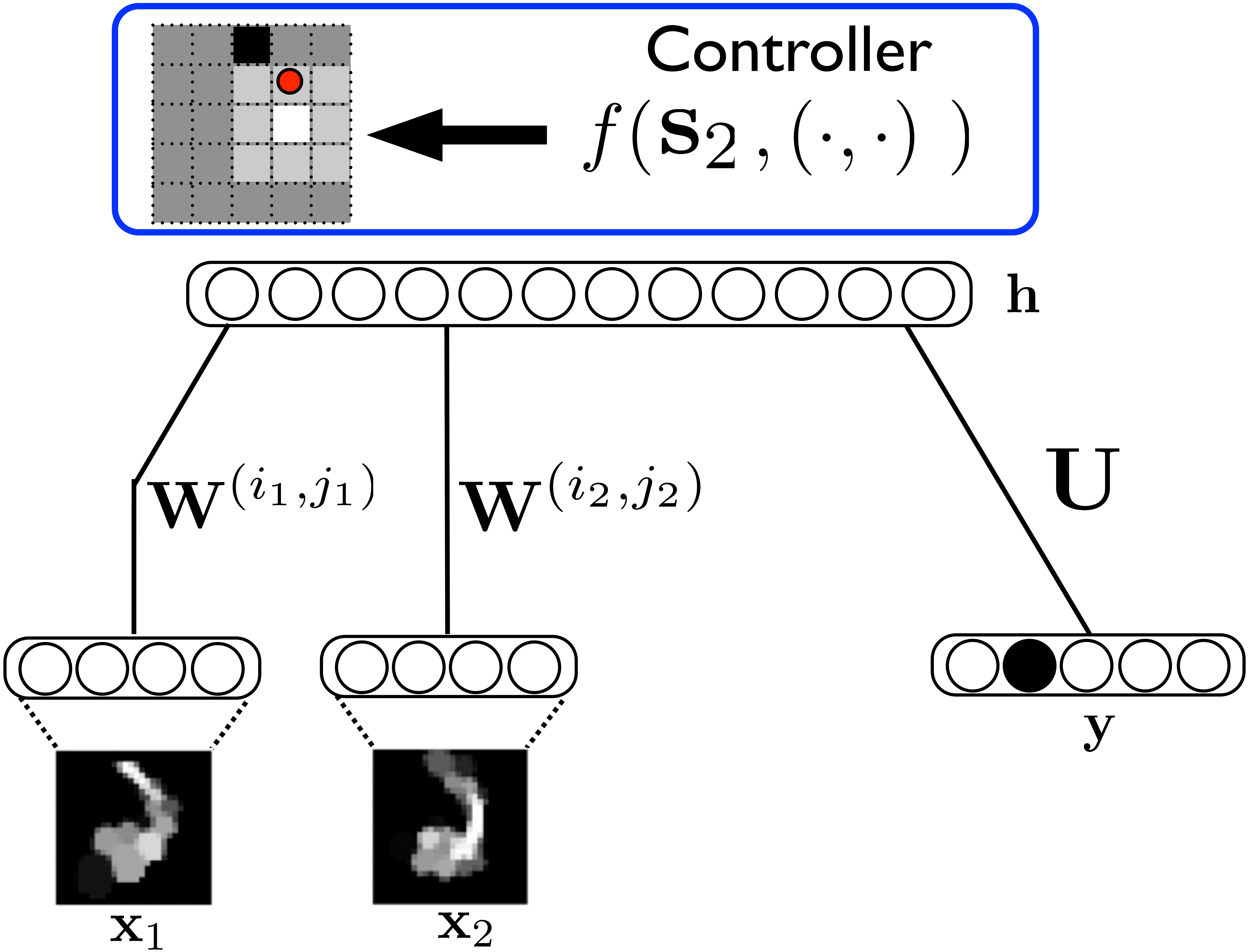
Putting it all together



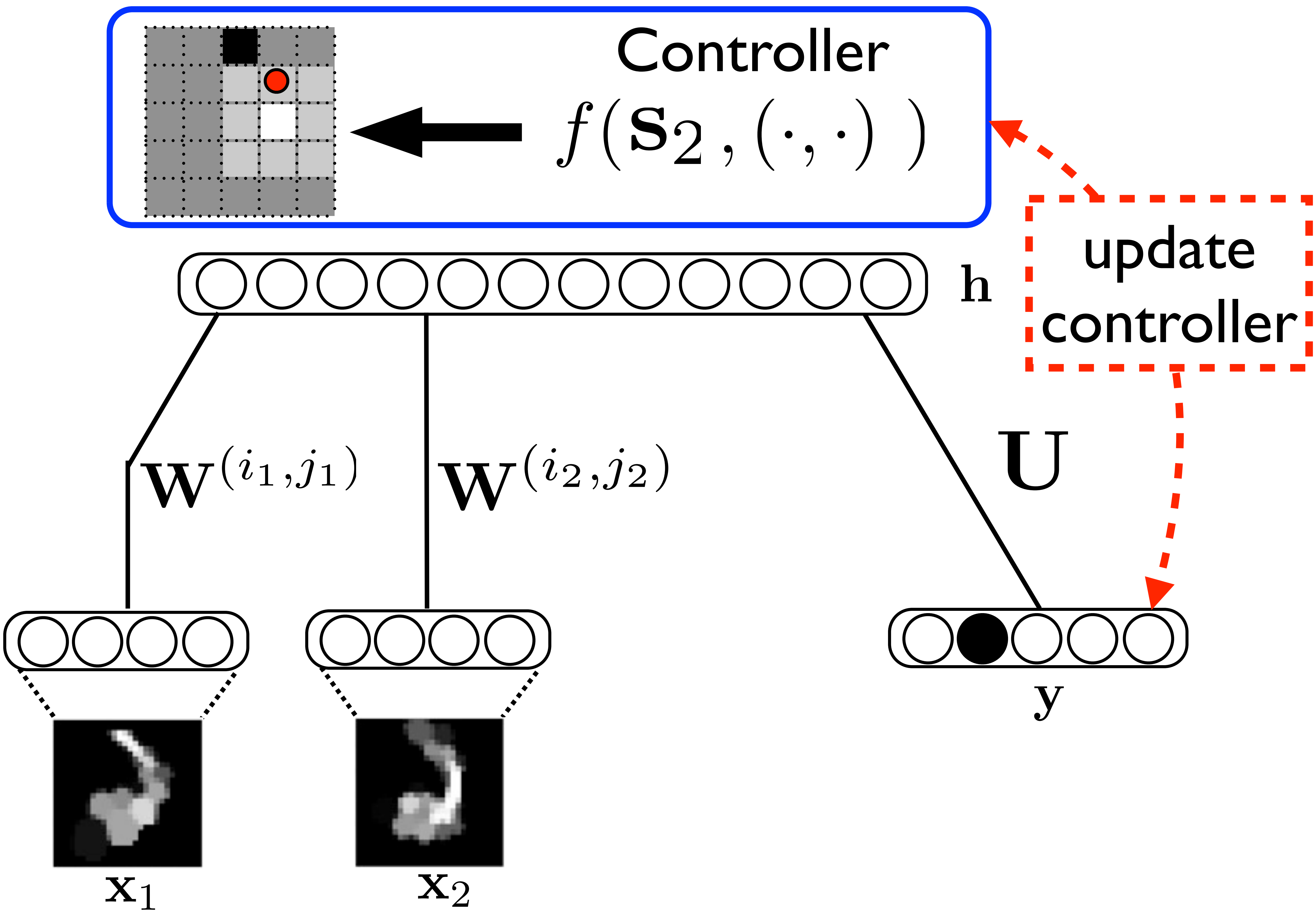
Putting it all together



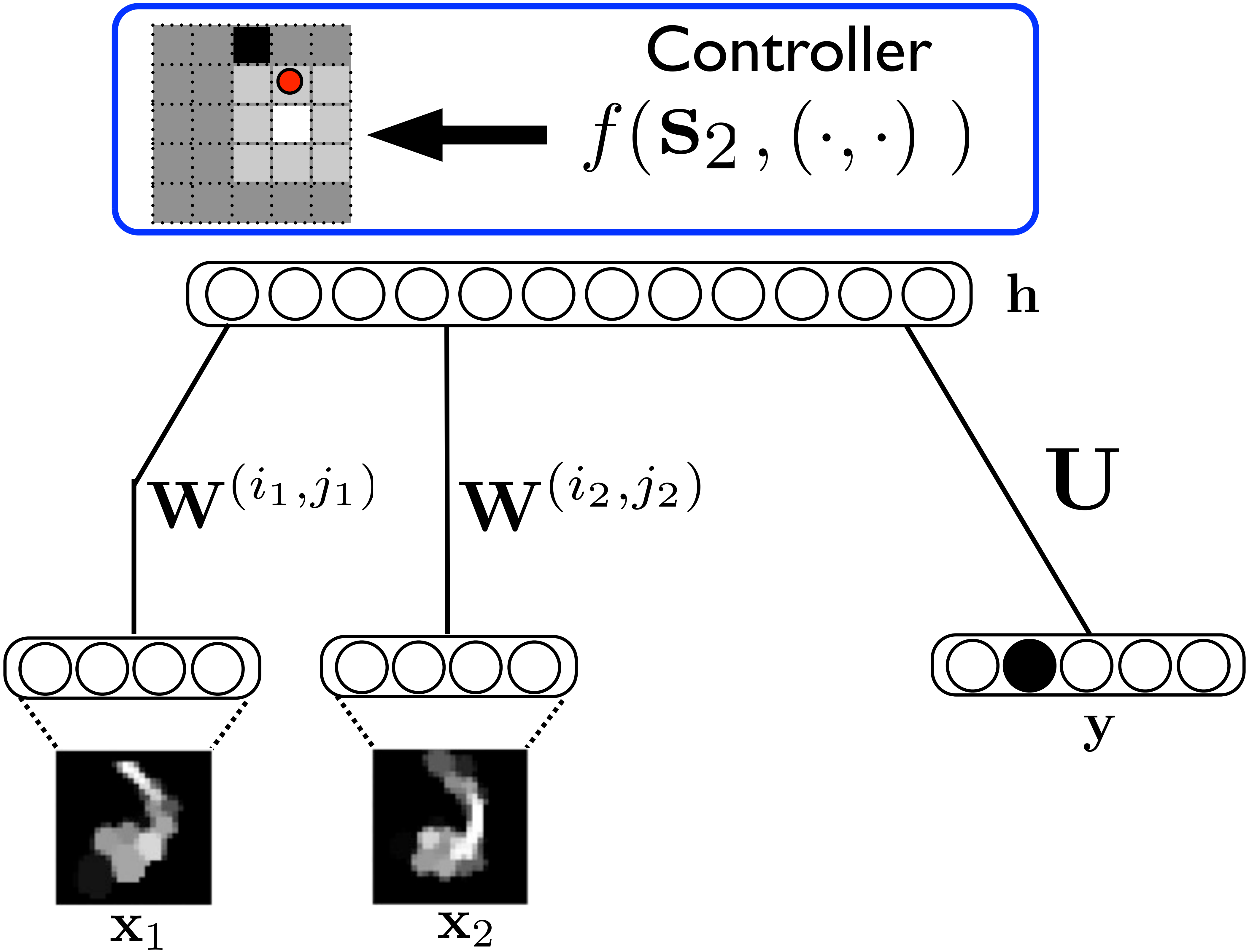
Putting it all together



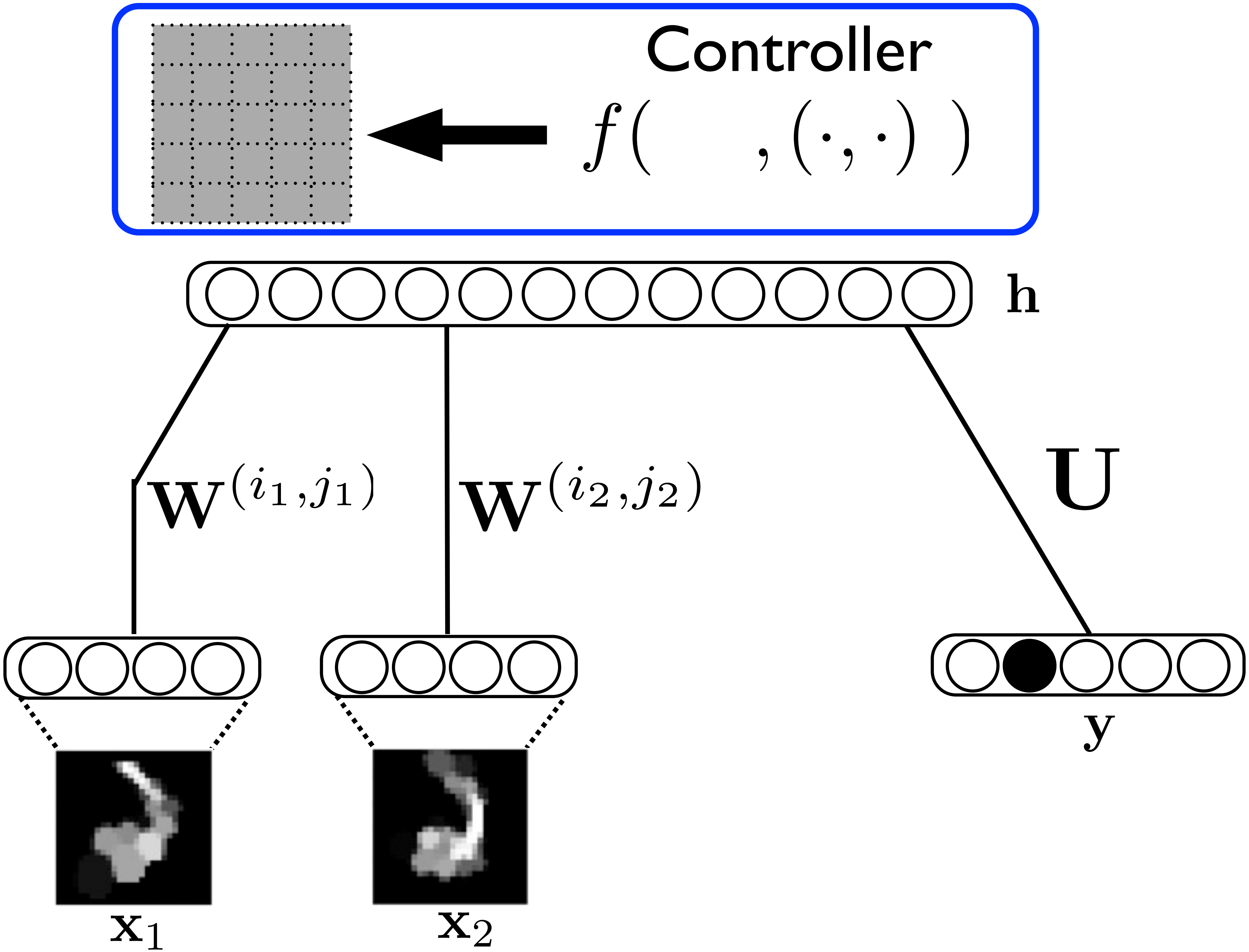
Putting it all together



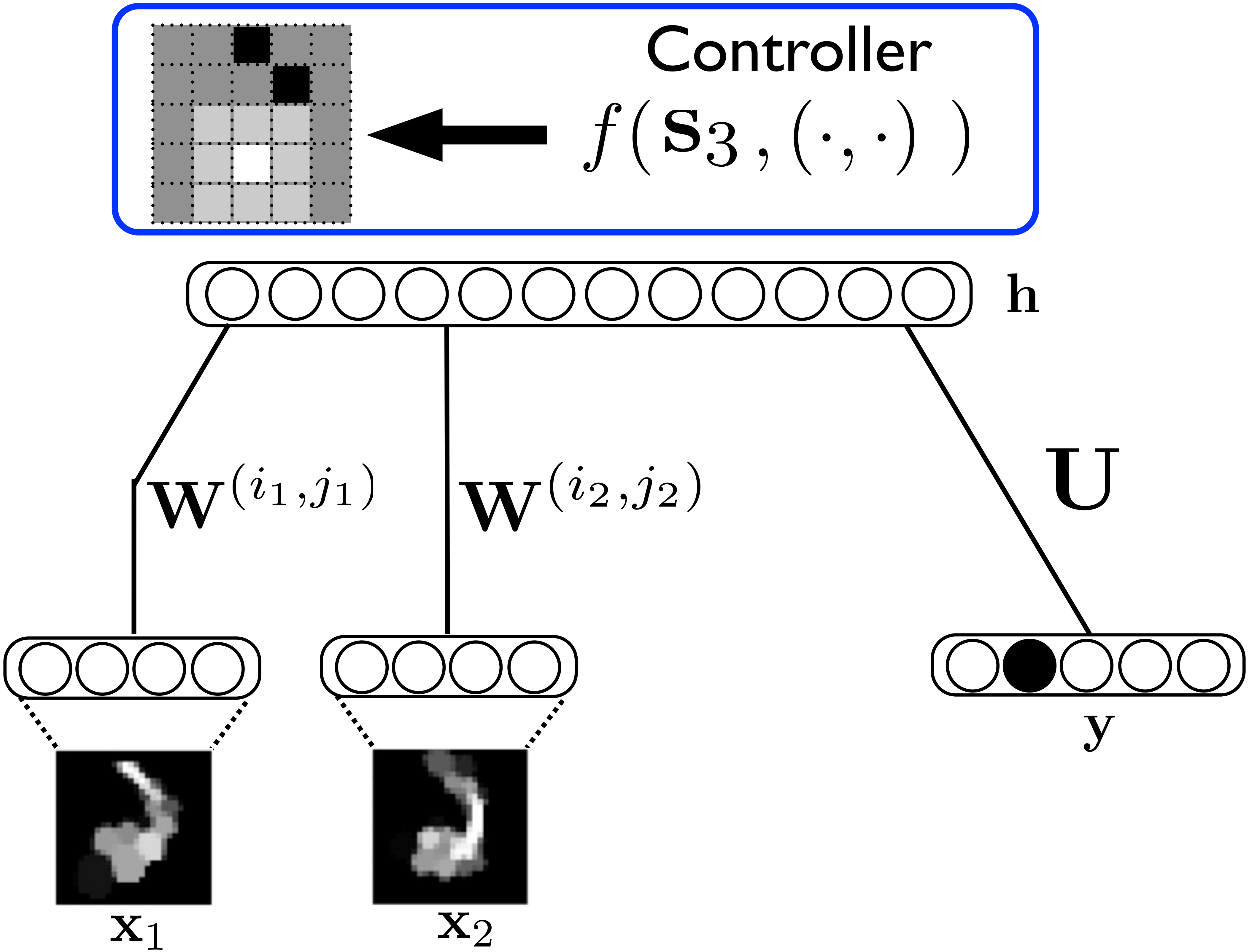
Putting it all together



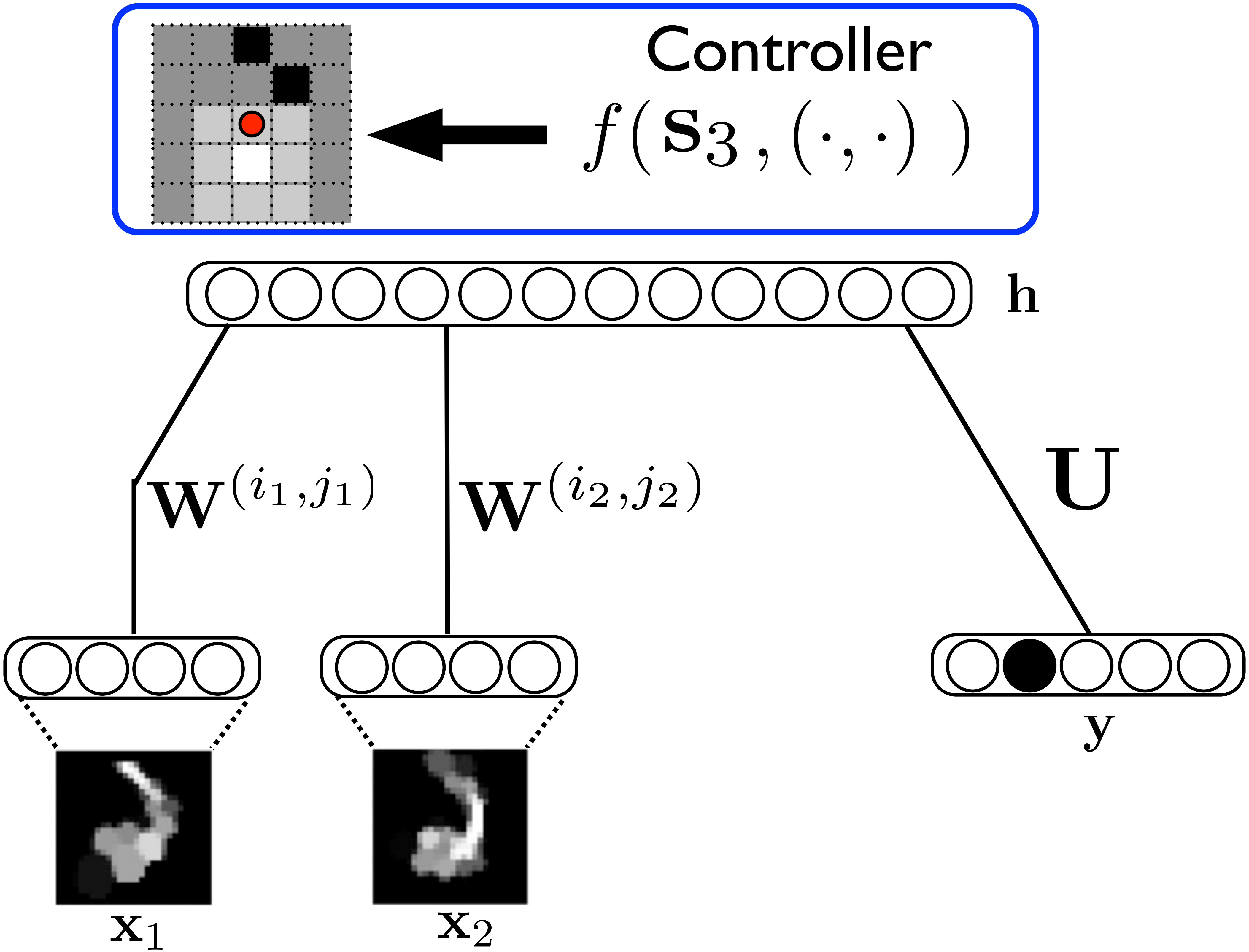
Putting it all together



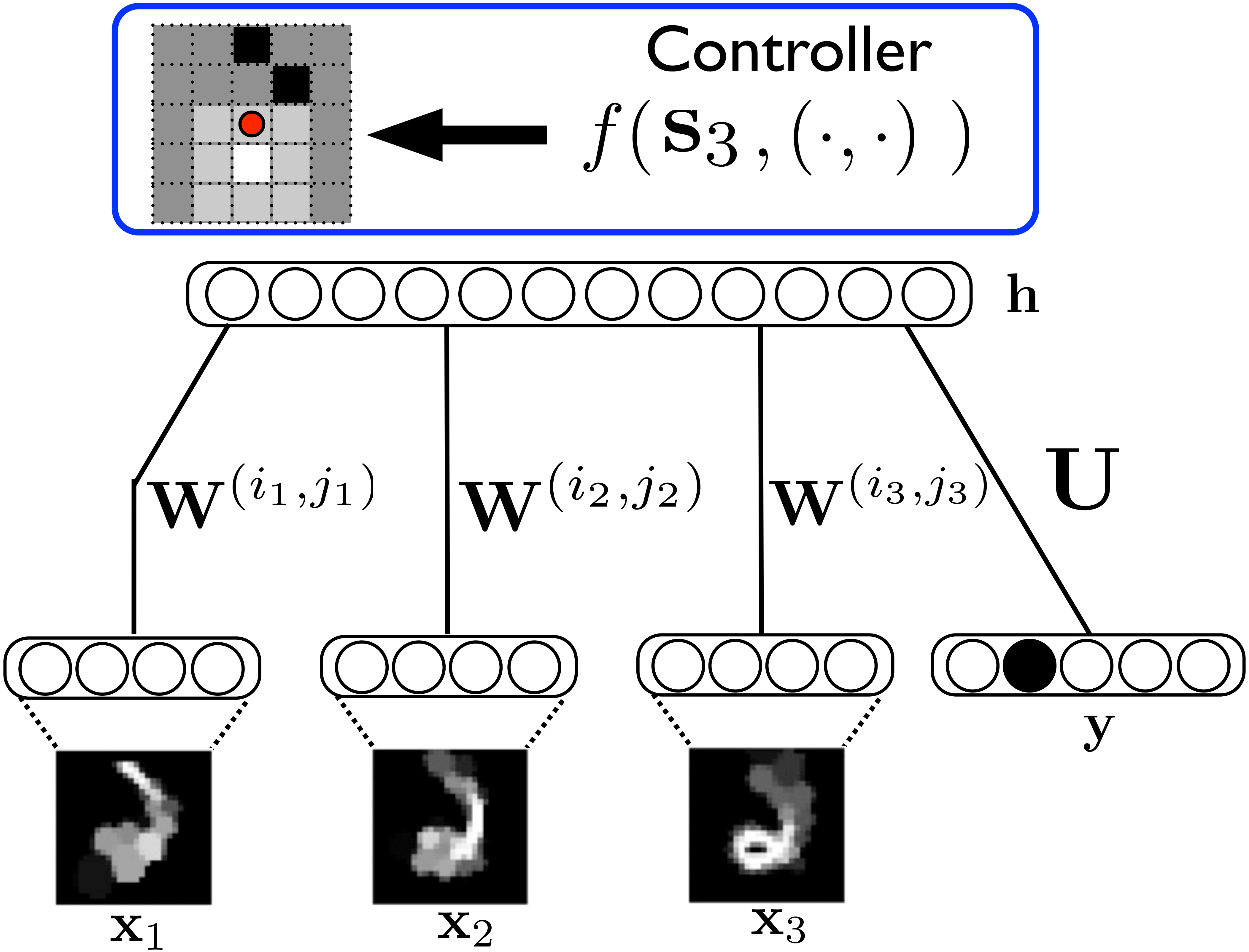
Putting it all together



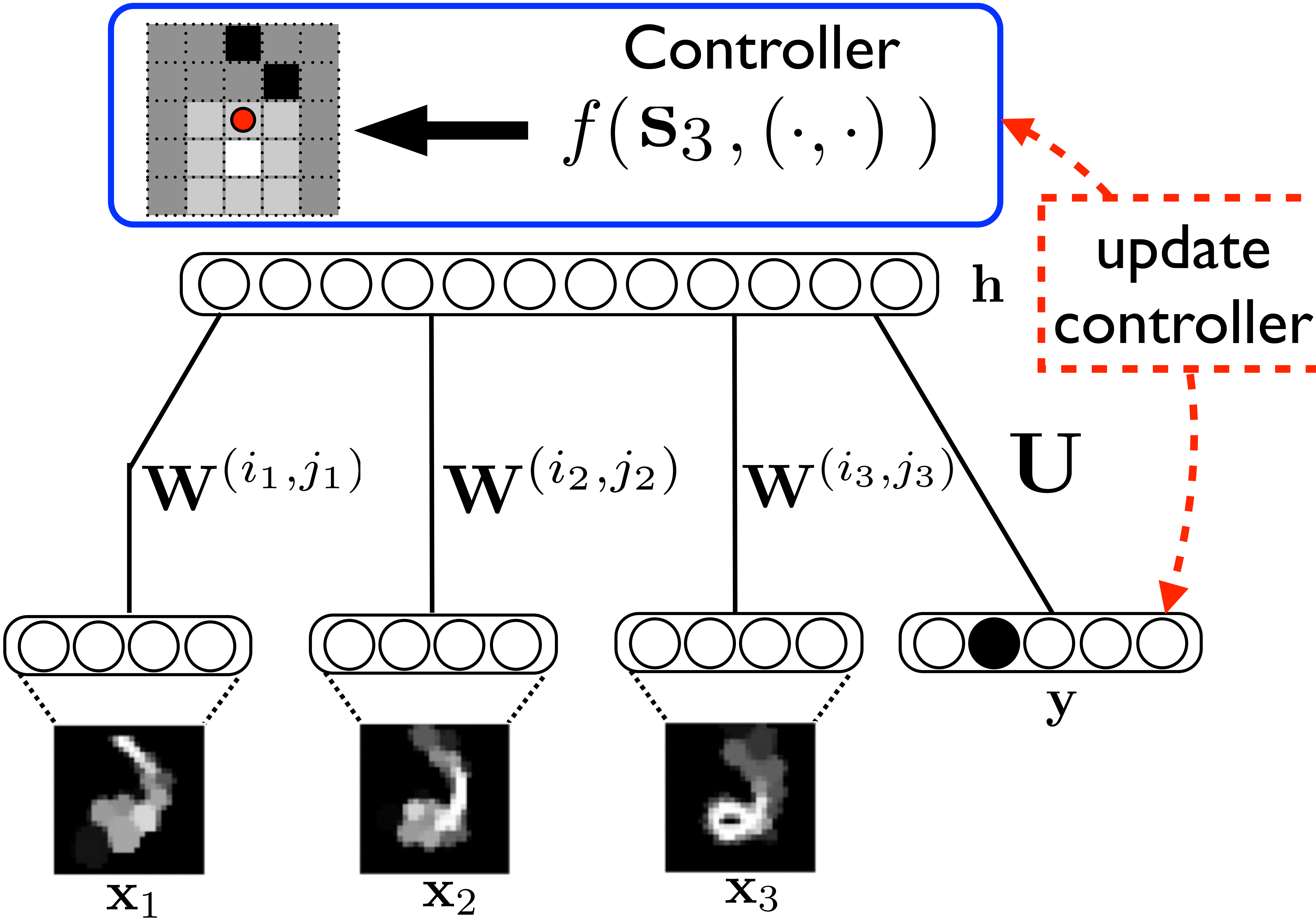
Putting it all together



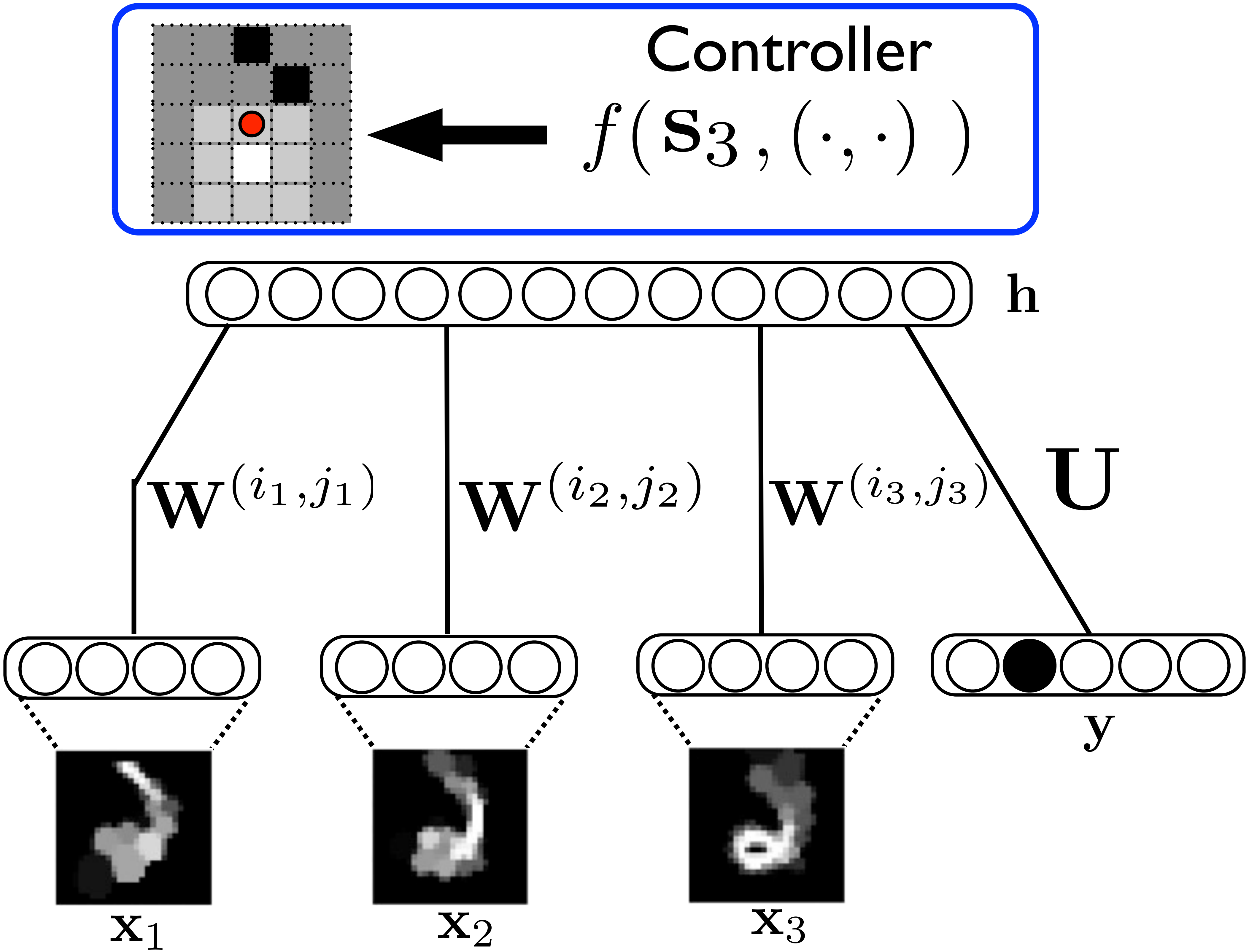
Putting it all together



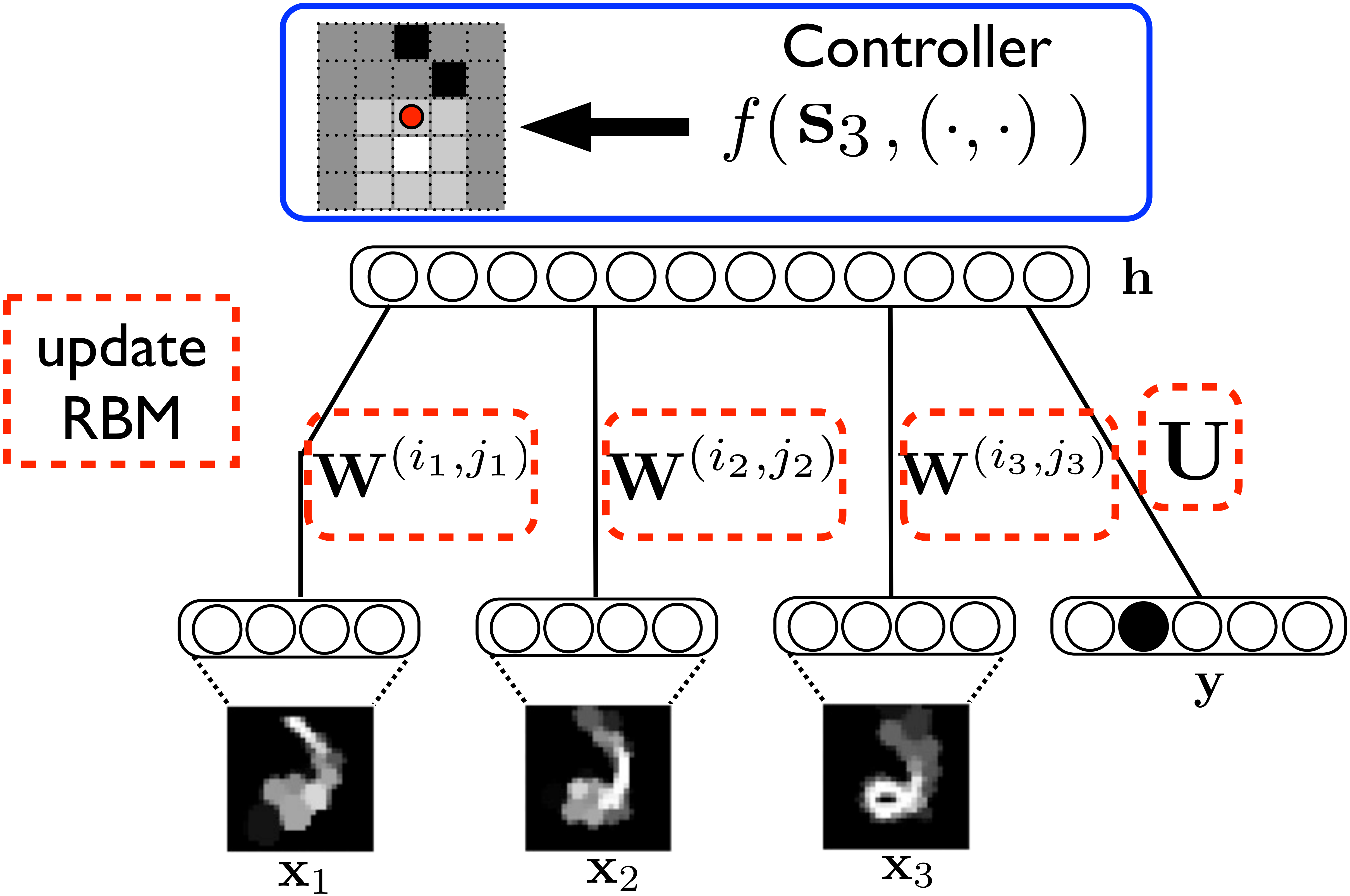
Putting it all together



Putting it all together



Putting it all together



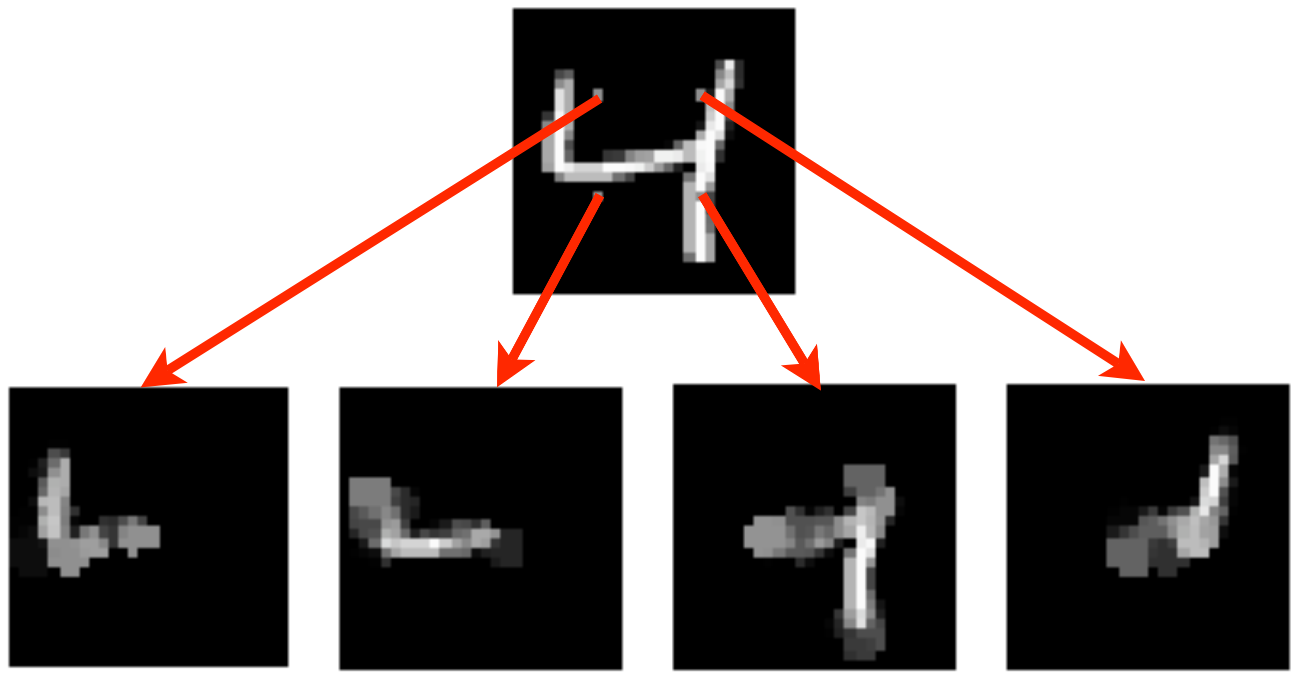
Related work

- Alpaydin (NIPS 1996):
  - ★ neural net accumulating fixations
  - ★ based on a fixed saliency map
- Kanan and Cottrell (CVPR 2010):
  - ★ learned saliency map
  - ★ non-parametric nearest neighbor recognition
- Our work:
  - ★ joint training of a recognition component (RBM) and an attentional component (controller)
  - ★ explicitly avoids looking everywhere (unlike saliency maps on high resolution image)

Experiments

- Evaluating the Multi-fixation RBM
- Evaluating the controller
- Evaluating the whole system

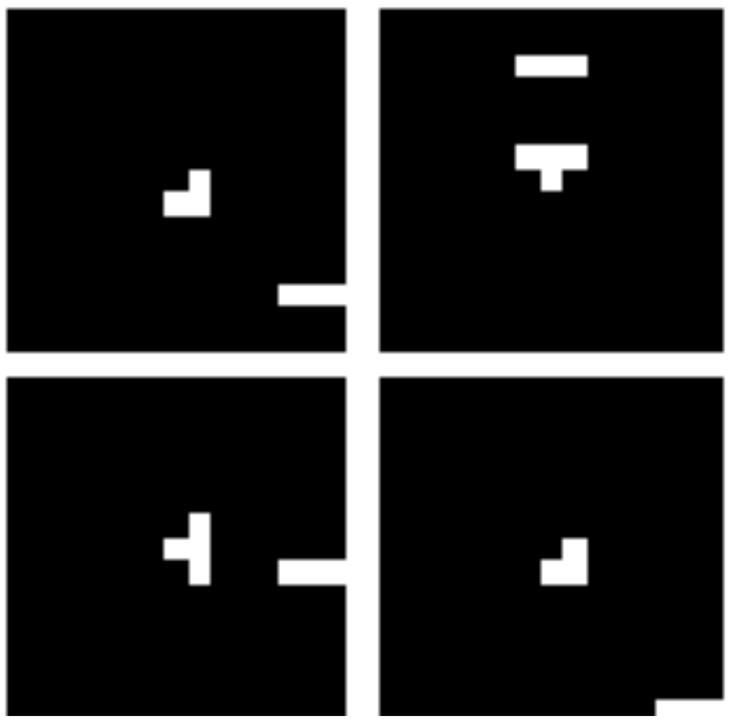
Experiment 1: MNIST (subset) with 4 fixations



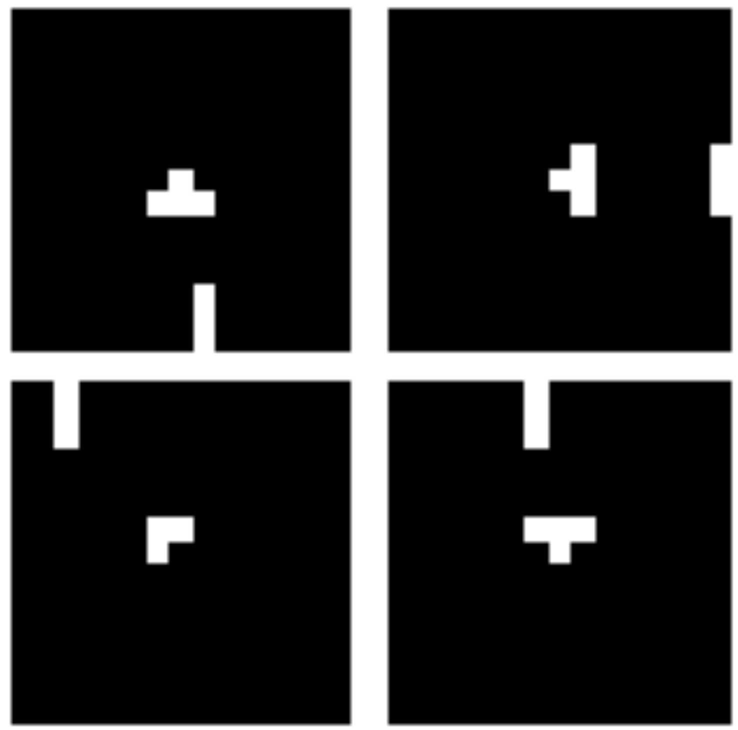
Model	Error
NNet+RBM [22]	3.17% (± 0.15)
SVM [21]	3.03% (± 0.15)
Multi-fixation RBM (hybrid)	3.20% (± 0.15)
Multi-fixation RBM (hybrid-sequential)	2.76% (± 0.14)

Experiment 2: Synthetic dataset

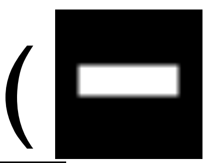
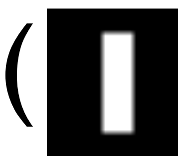
Positive examples



Negative examples



Task

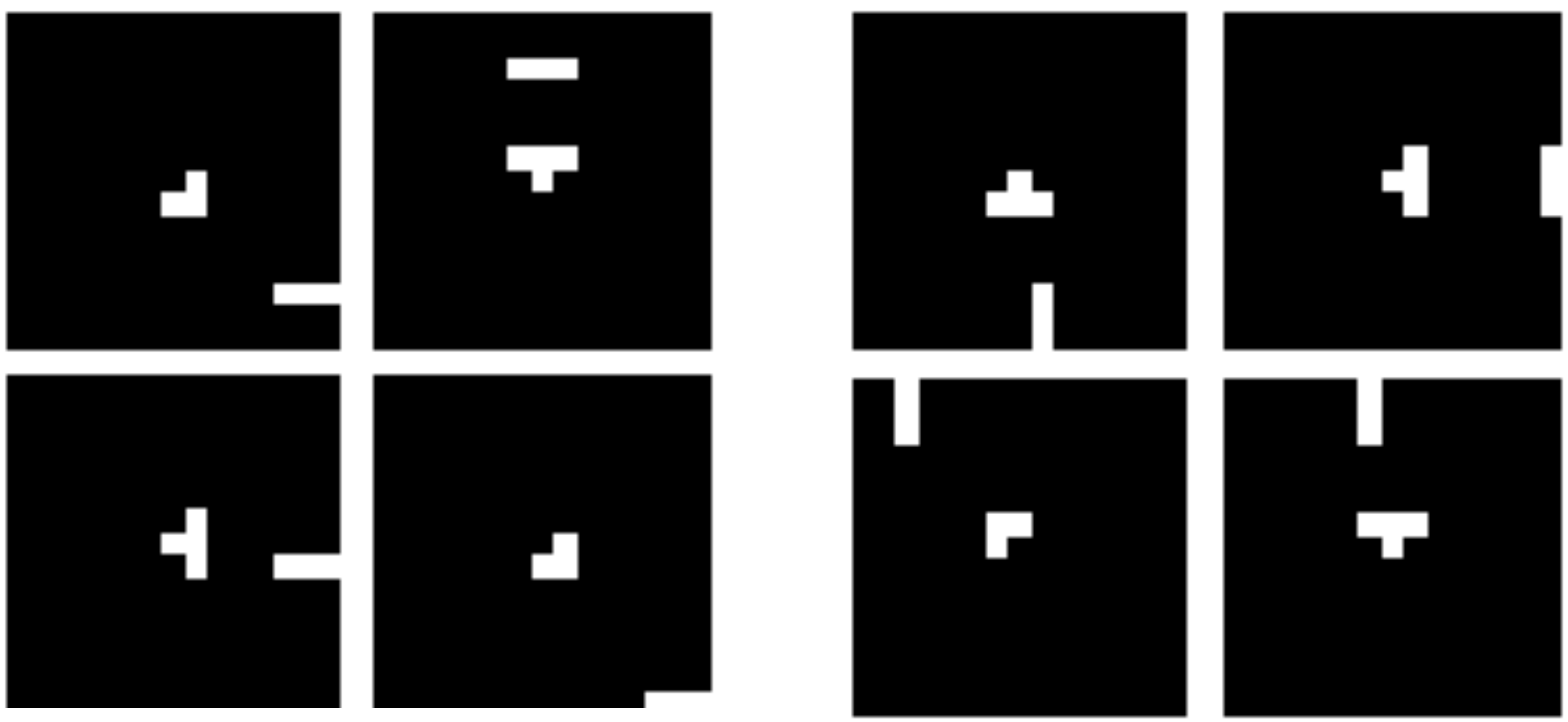
- Identify presence of horizontal () or vertical () bars
- Symbol at center says where the bar is

 = top left

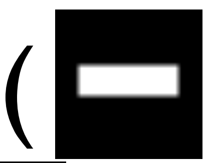
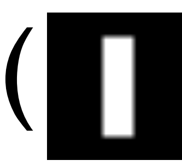
 = right

Experiment 2: Synthetic dataset

Positive examples   Negative examples



Task

- Identify presence of horizontal () or vertical () bars
- Symbol at center says where the bar is

 = top left

 = right

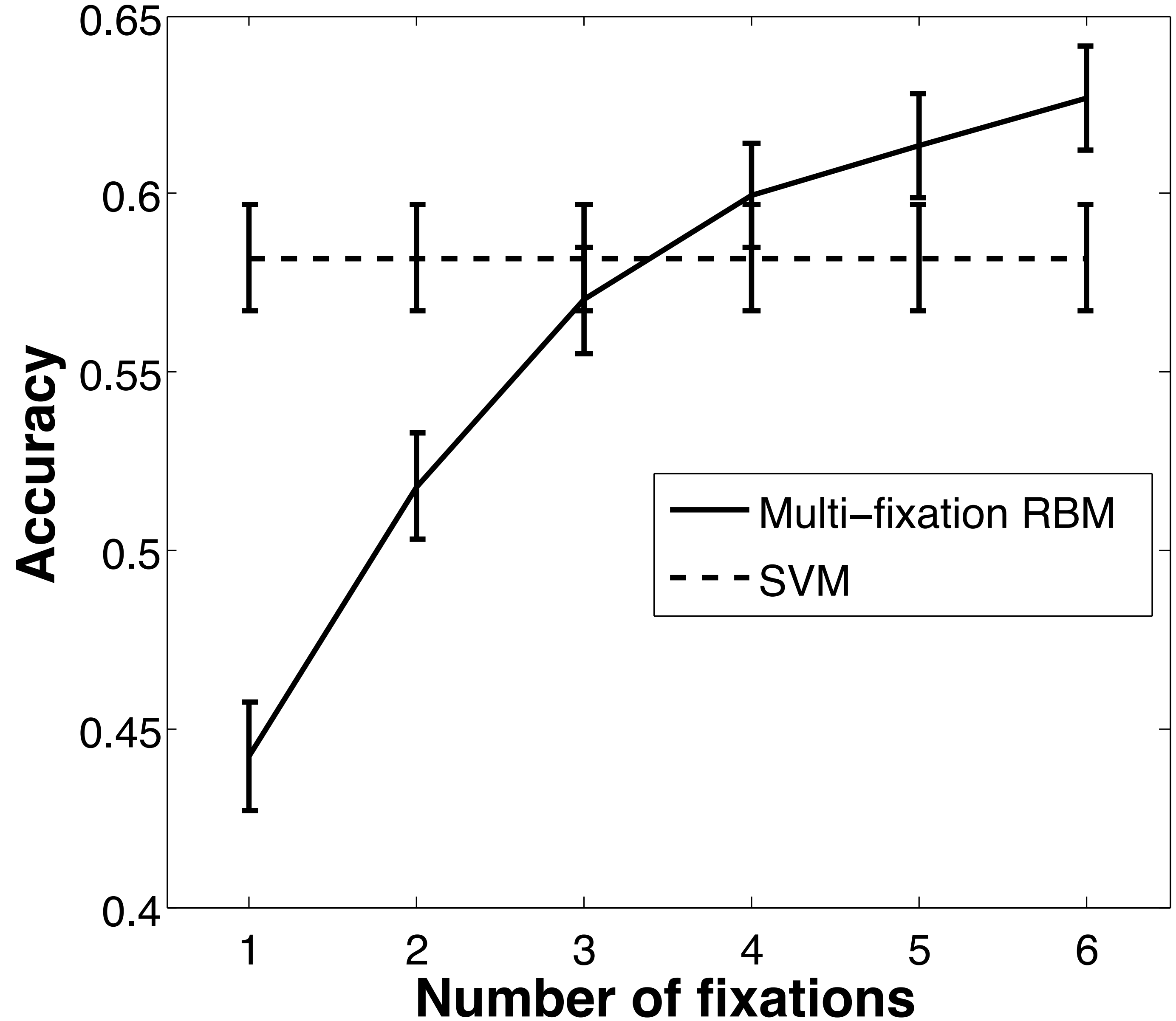
Results

1. Hybrid training solves this problem perfectly
2. Discriminative training fails (50%)

Experiment 3:
Facial expression recognition

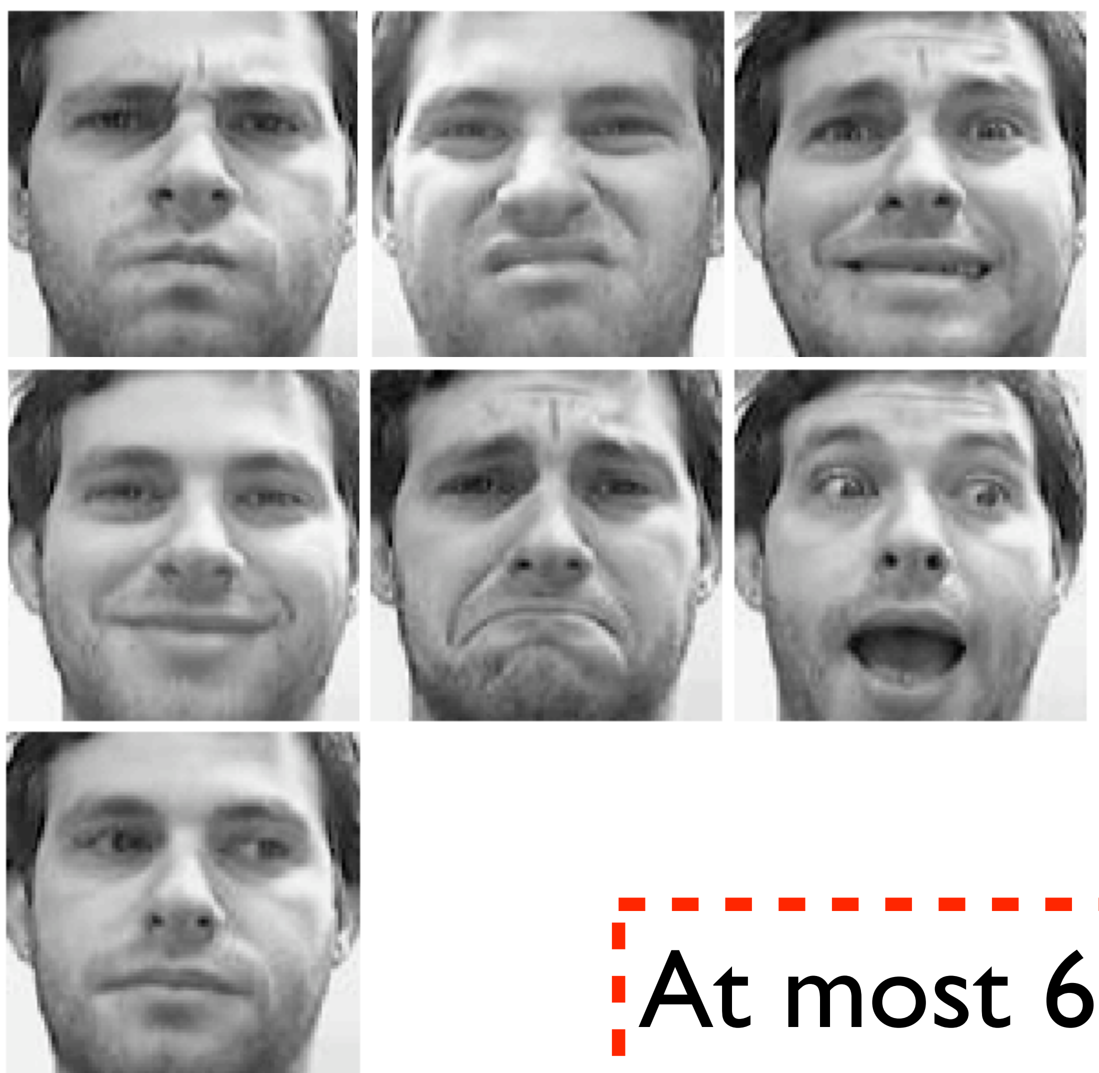
Examples

Results

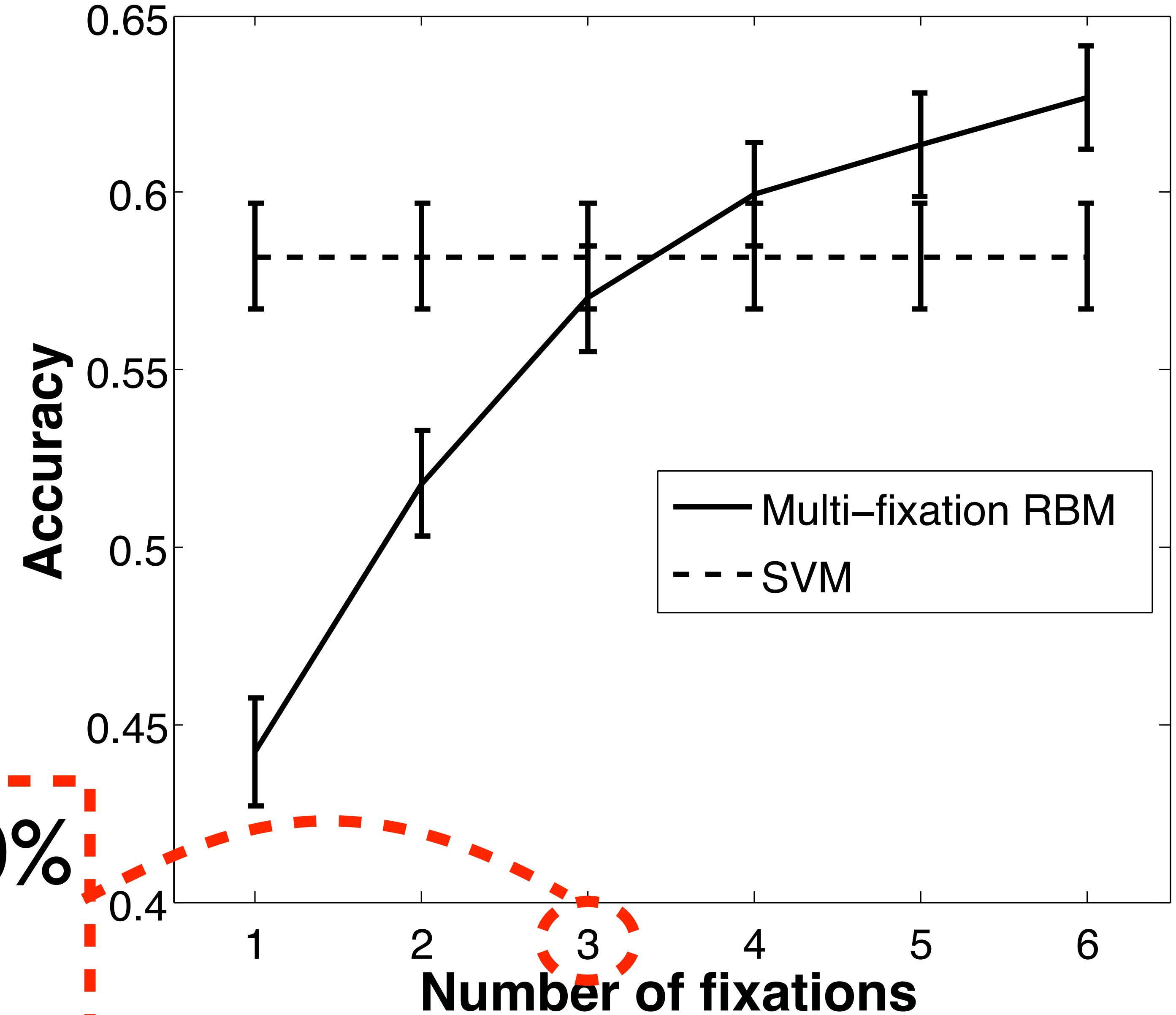


Experiment 3:
Facial expression recognition

Examples



Results



At most 60% of image

Experiment 3:
Facial expression recognition

Full image + sequence of fixations

Individual glimpses

Conclusion

- Investigated a model for jointly learning a recognition and attentional component using a Boltzmann machine

- Future work:
  - ★ impact of retinal rep. on performance
  - ★ improvement to controller algorithm
  - ★ multitask learning

Thank you!