# ODE-Inspired Analysis for the Biological Version of Oja's Rule in Solving Streaming PCA

Chi-Ning Chou
Harvard

Mien Brabeeba Wang
MIT

COLT 2020

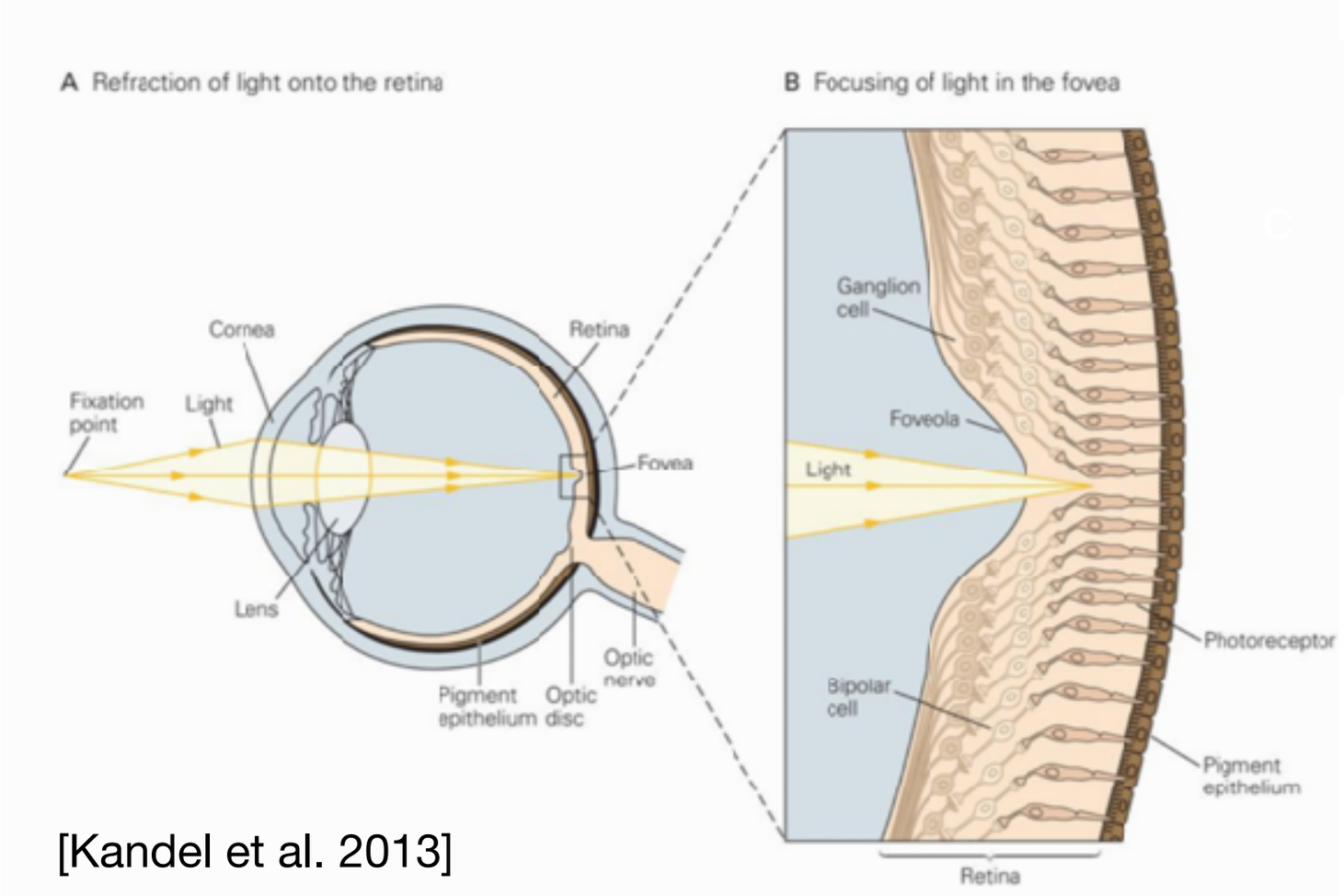
# ODE-Inspired Analysis for the Biological Version of Oja's Rule in Solving Streaming PCA

Chi-Ning Chou
Harvard

Mien Brabeeba Wang
MIT

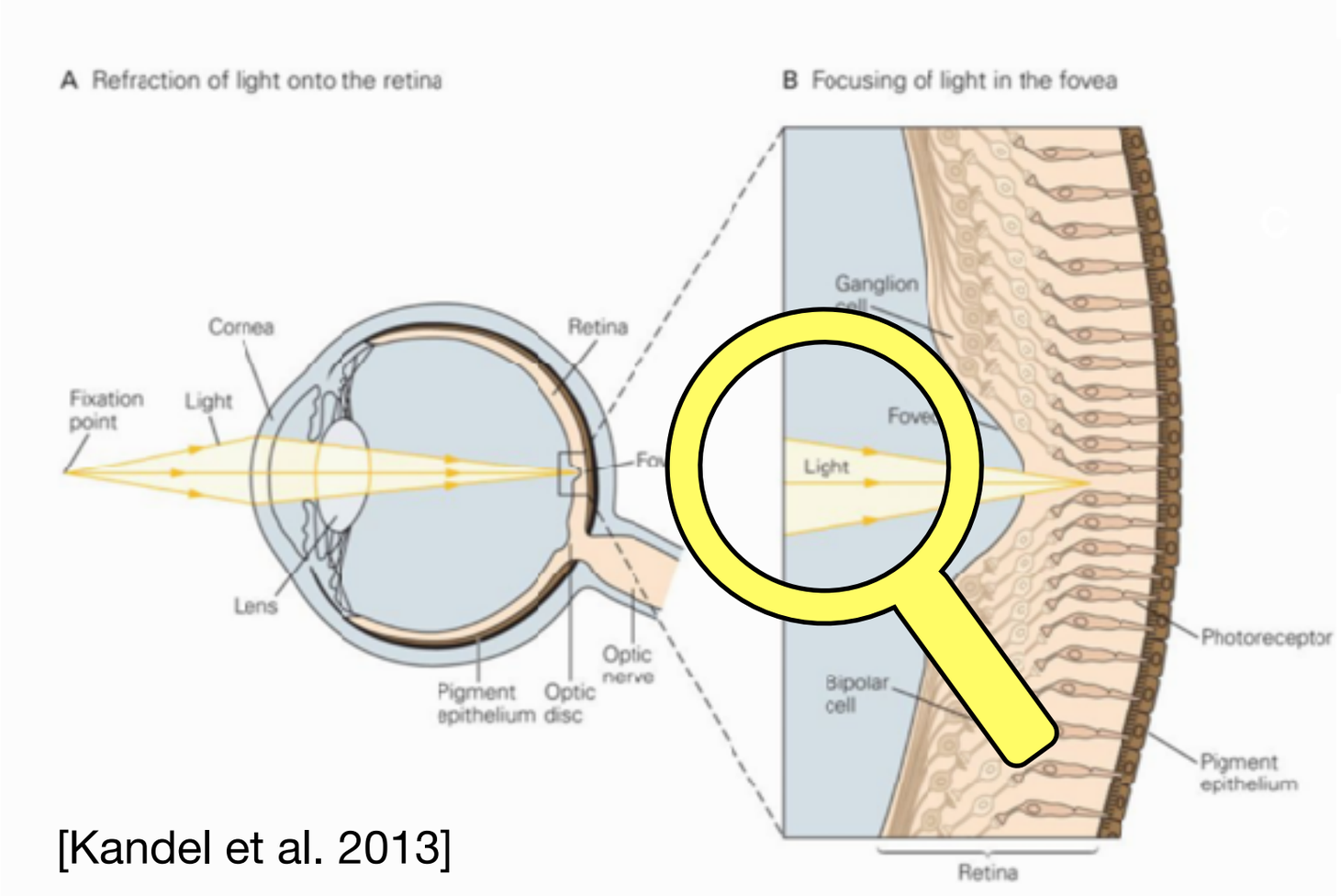
COLT 2020

# Retina




[Kandel et al. 2013]

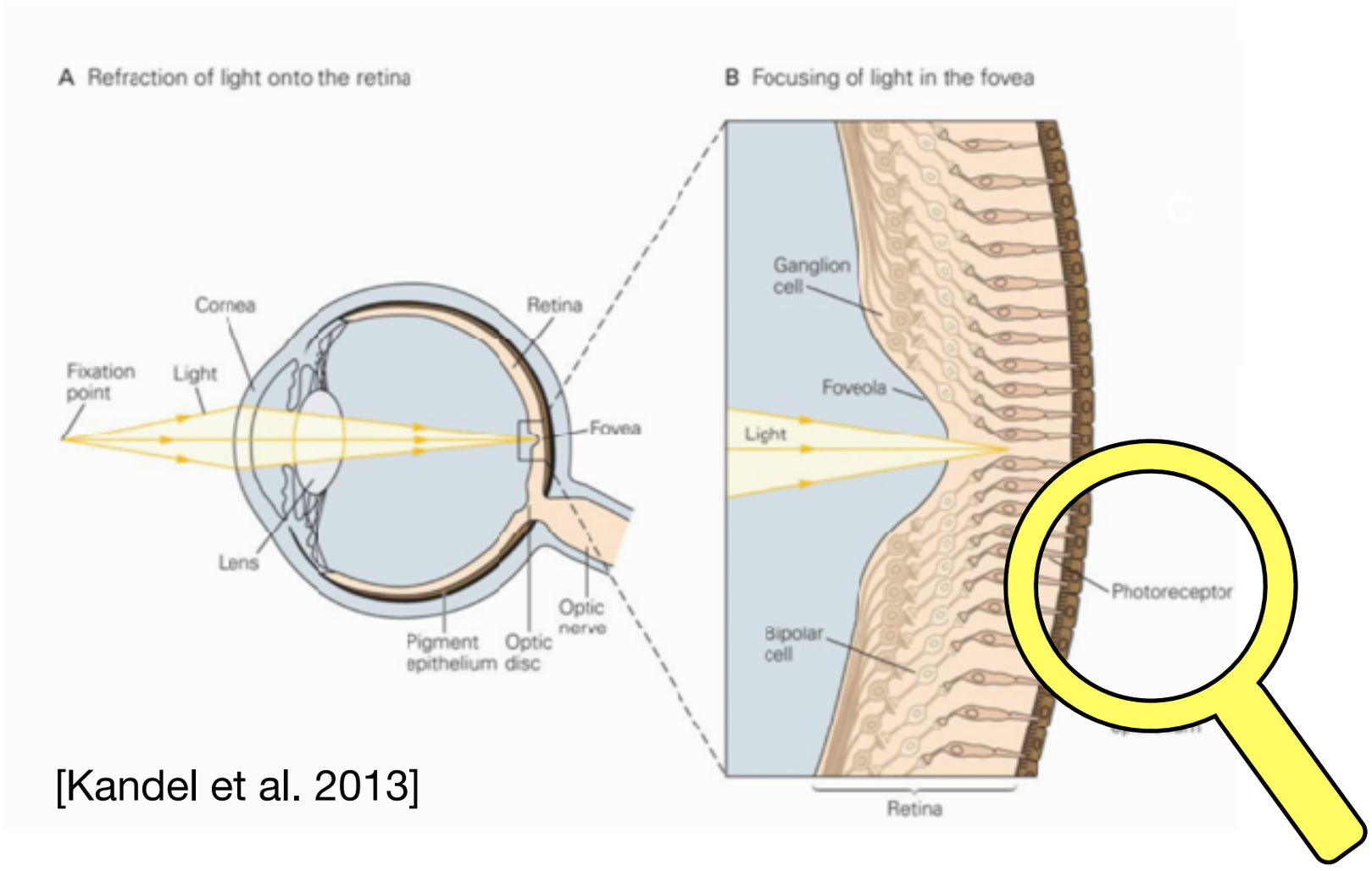
# Retina




[Kandel et al. 2013]

**Light**

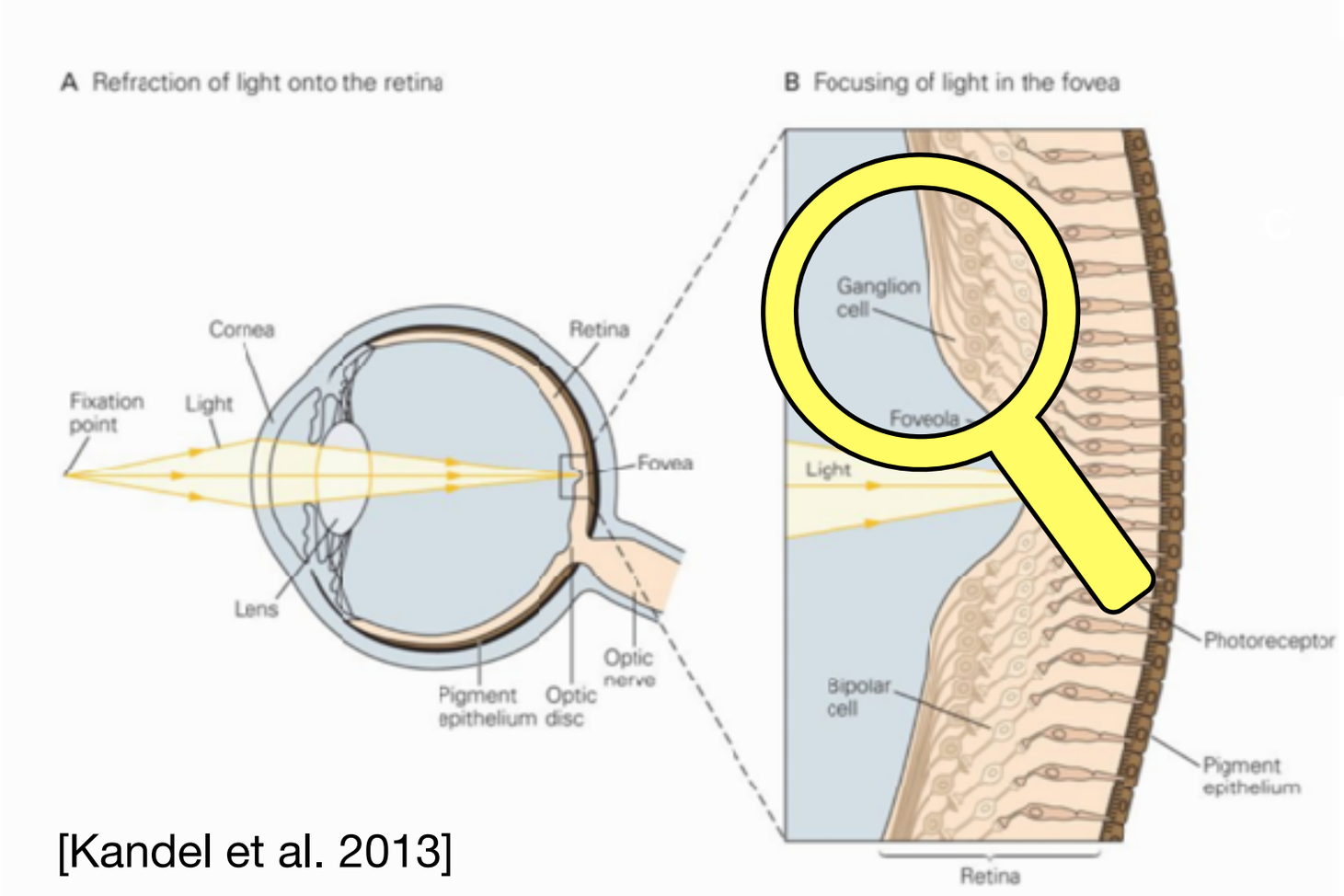
# Retina




[Kandel et al. 2013]

**Light** → **Photoreceptors**

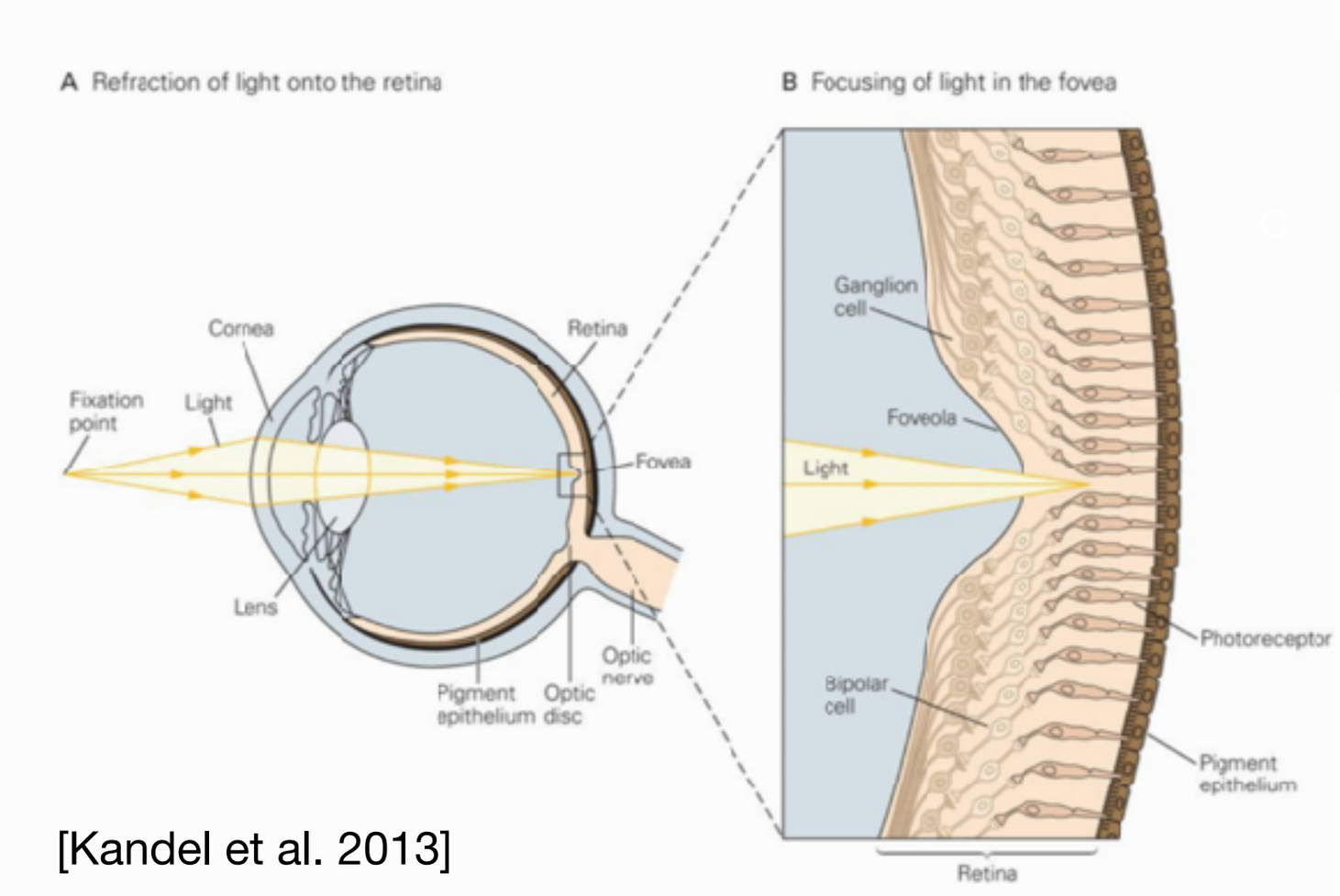
# Retina




[Kandel et al. 2013]

**Light** → **Photoreceptors** → **Ganglion cells**

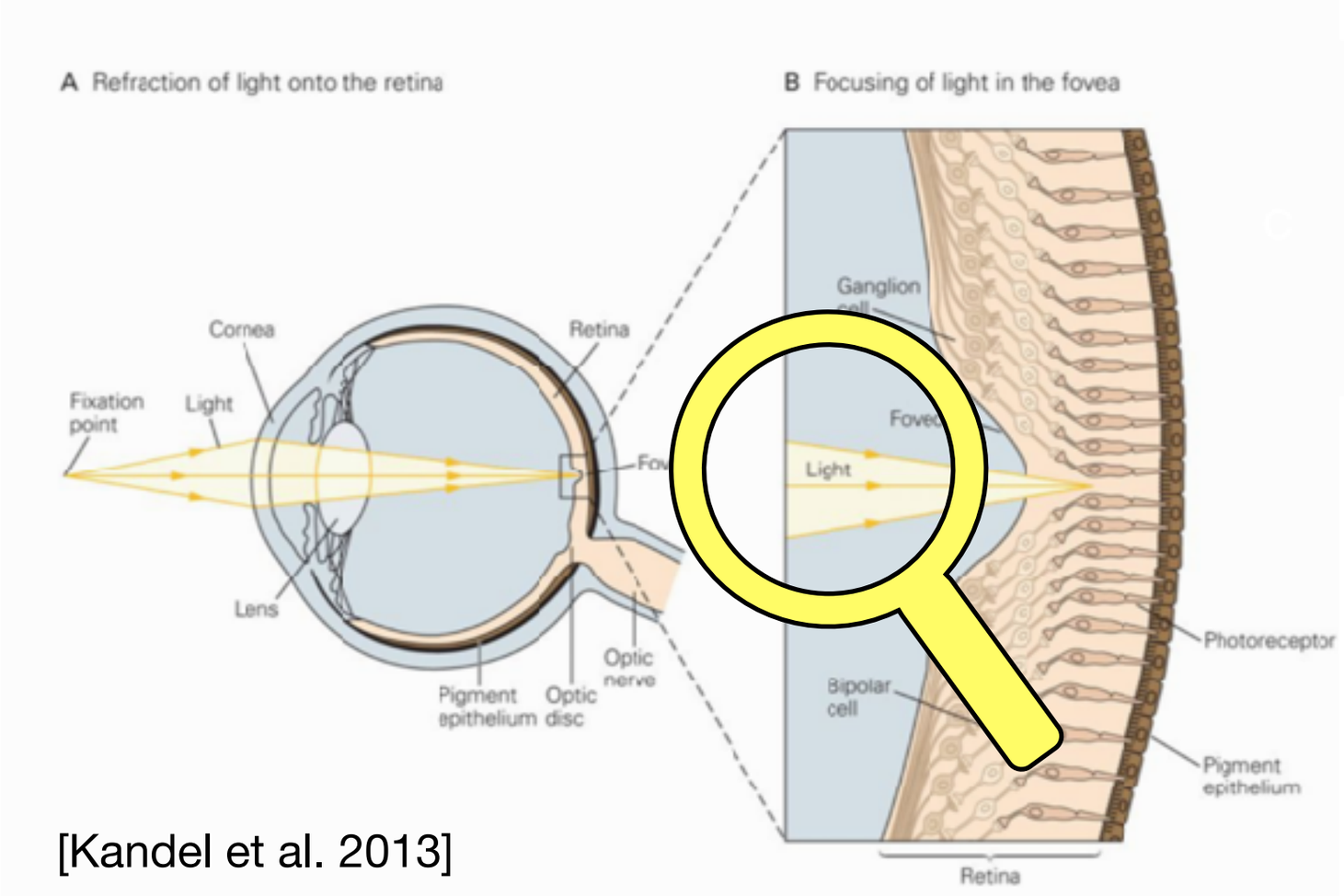
# Retina




[Kandel et al. 2013]

**Light** → **Photoreceptors** → **Ganglion cells** → **Cortex**

# Retina

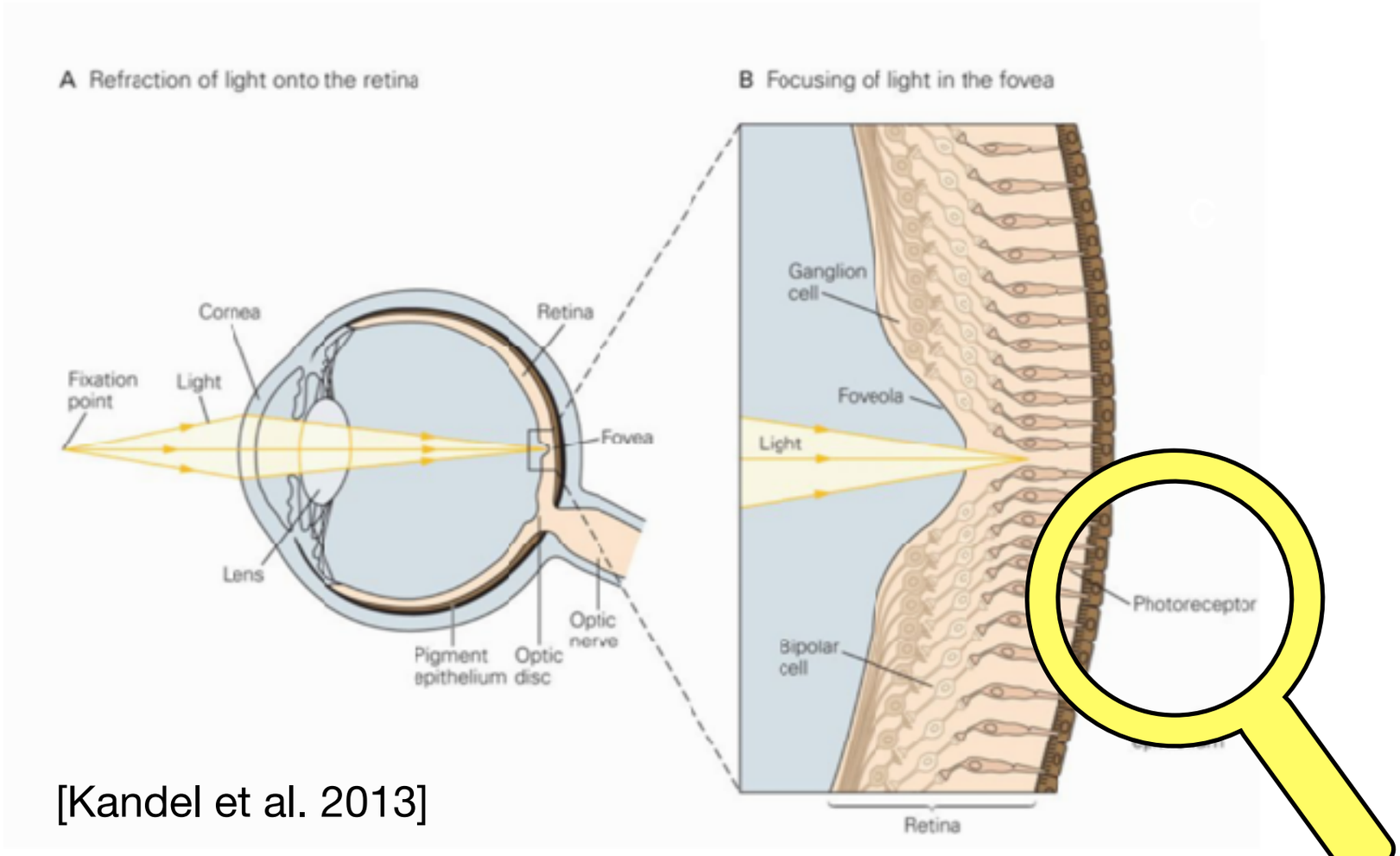



[Kandel et al. 2013]

**Light** → **Photoreceptors** → **Ganglion cells** → **Cortex**

(GBs per sec)

# Retina

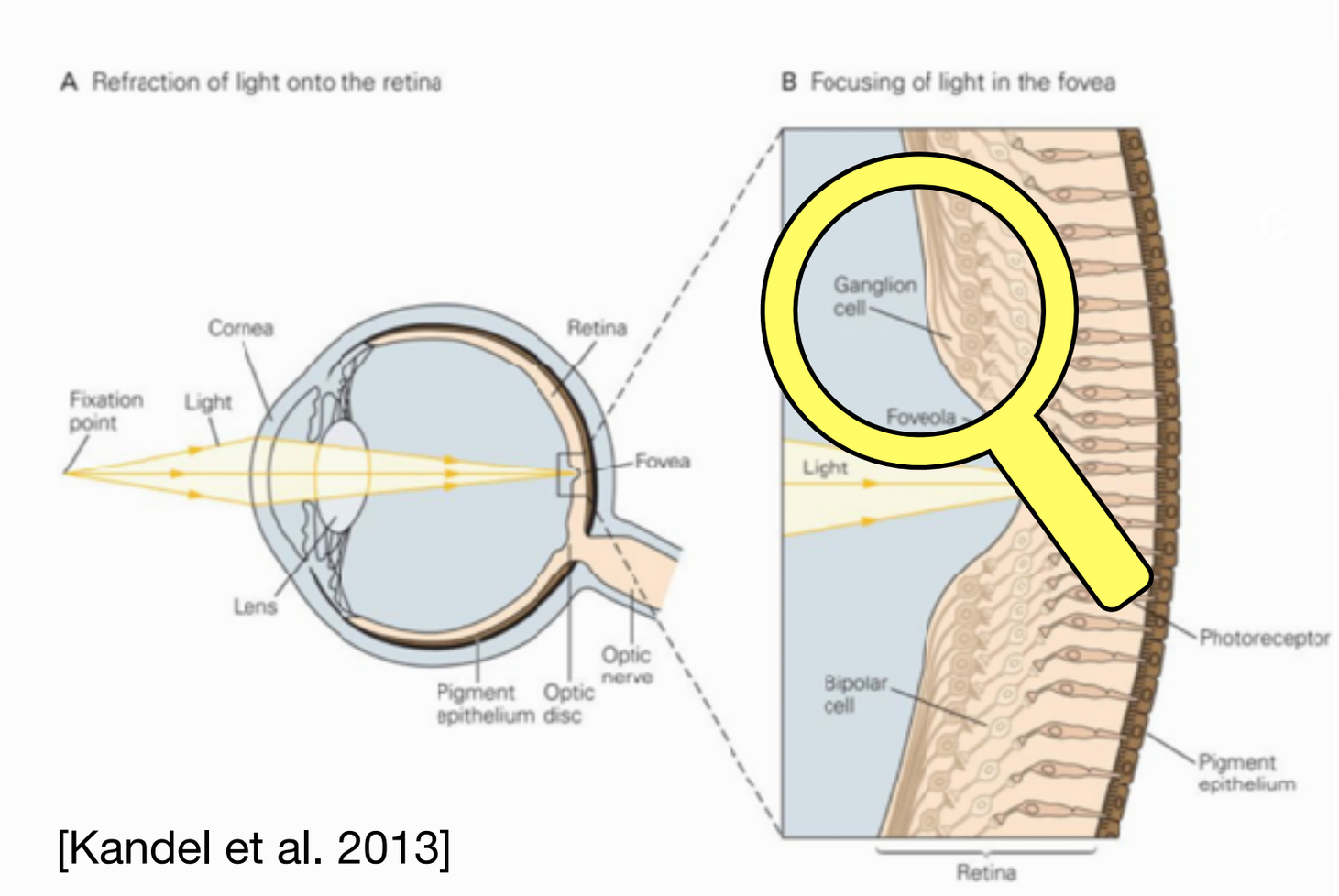



[Kandel et al. 2013]

**Light** → **Photoreceptors** → **Ganglion cells** → **Cortex**

(GBs per sec) (**100** millions)

# Retina

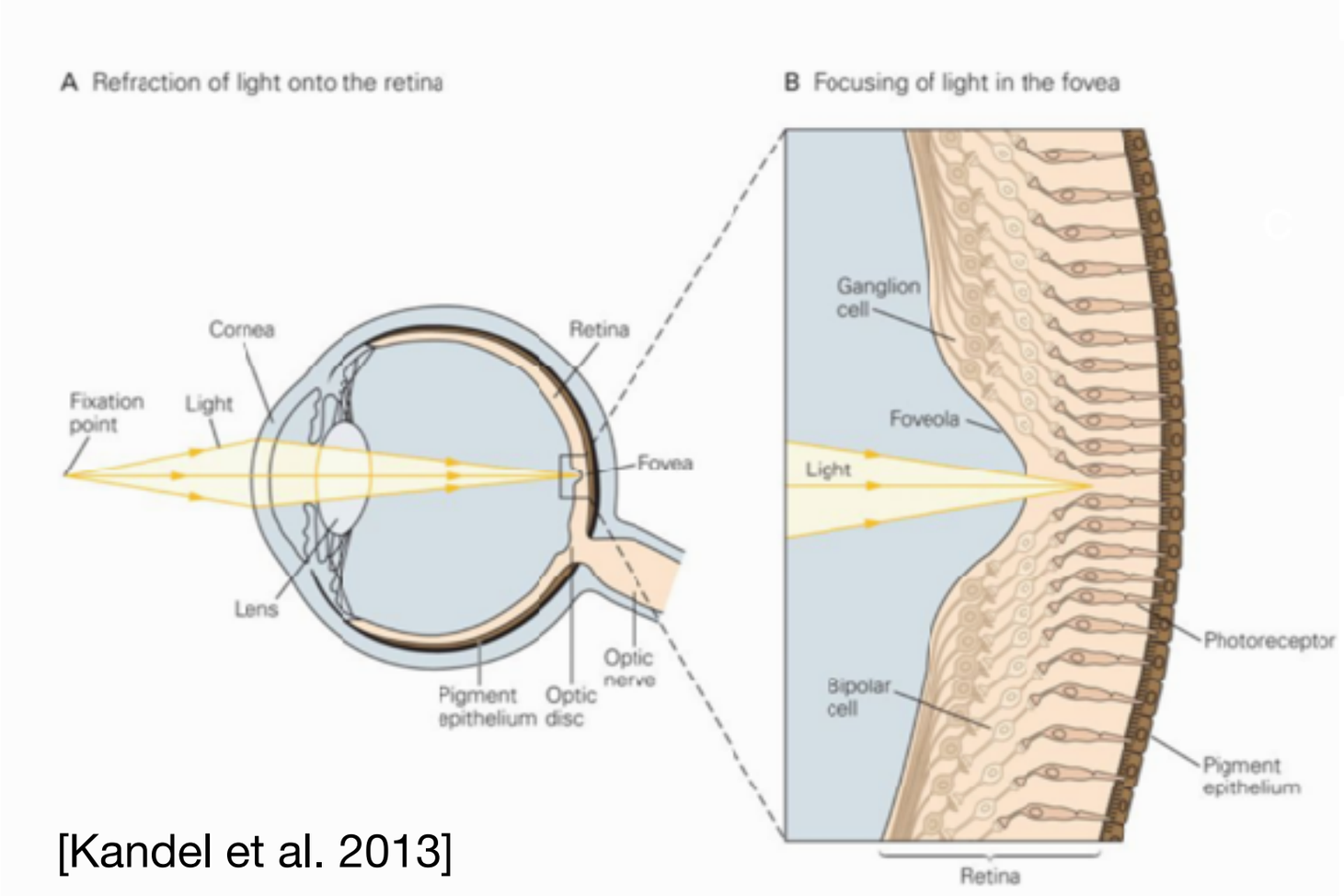



[Kandel et al. 2013]

**Light** → **Photoreceptors** → **Ganglion cells** → **Cortex**

(GBs per sec) (**100** millions) (**1** million)

# Retina




[Kandel et al. 2013]

**Light** → **Photoreceptors** → **Ganglion cells** → **Cortex**

(GBs per sec) (**100** millions) ? (**1** million)

# Adaptation in Retina

# Adaptation in Retina

*What if the environment changed?*

# Adaptation in Retina

*What if the environment changed?*

# Adaptation in Retina

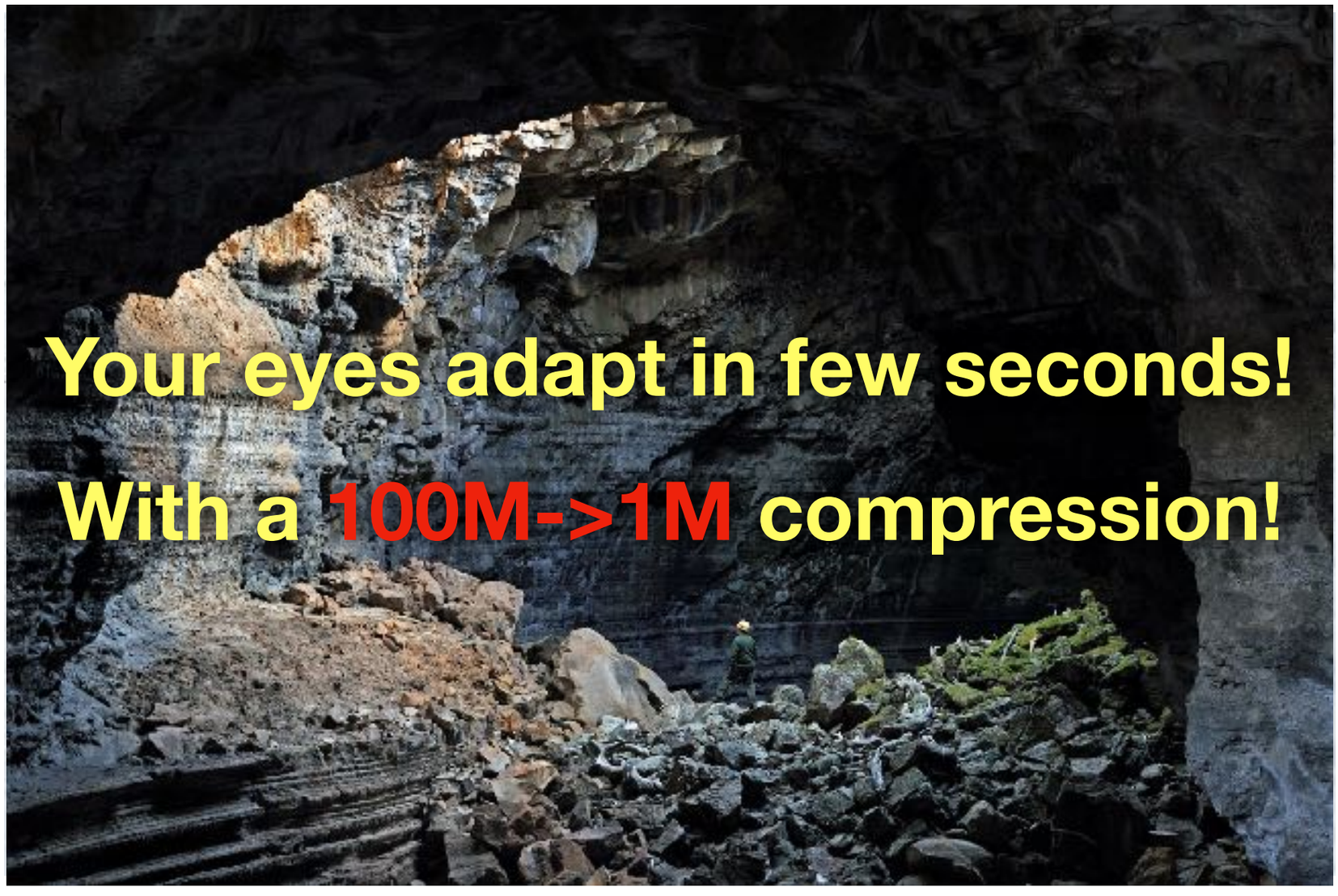
*What if the environment changed?*

# Adaptation in Retina

*What if the environment changed?*

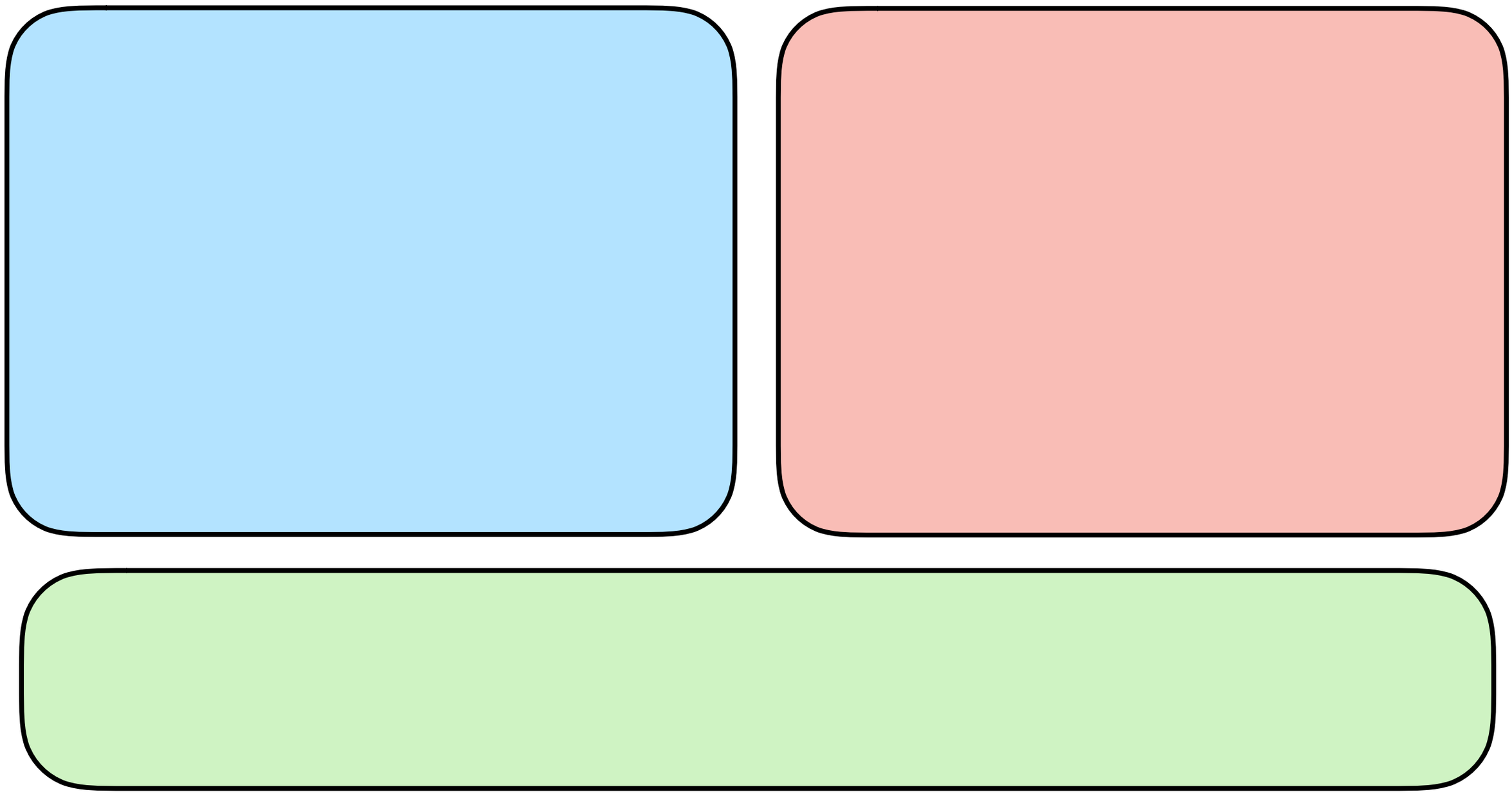
# Adaptation in Retina

*What if the environment changed?*



# What’s the Compression Mechanism in Retina?

# What’s the Compression Mechanism in Retina?



# What's the Compression Mechanism in Retina?

**A Computational Task**

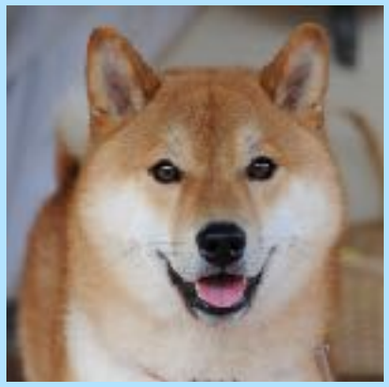
(Capturing experimental observation)

# What's the Compression Mechanism in Retina?

## A Computational Task

(Capturing experimental observation)

vs.

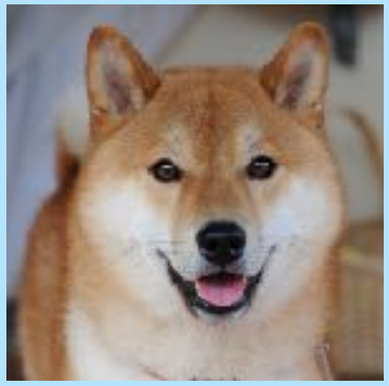


# What's the Compression Mechanism in Retina?

## A Computational Task

(Capturing experimental observation)

vs.



## A Mathematical Model

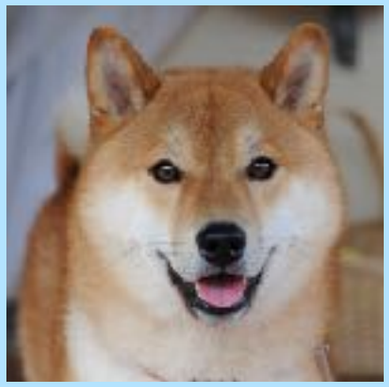
(Subject to biological constraints)

# What's the Compression Mechanism in Retina?

## A Computational Task

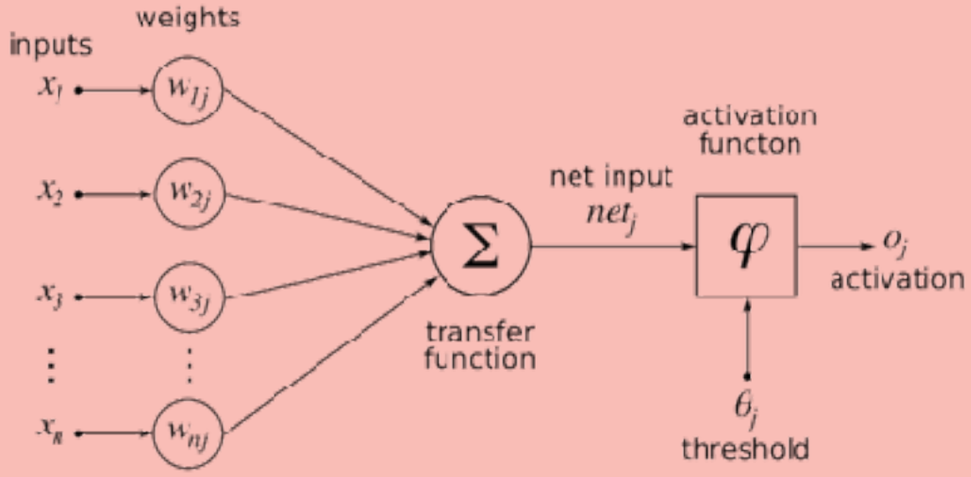
(Capturing experimental observation)

**vs.**



## A Mathematical Model

(Subject to biological constraints)



# What's the Compression Mechanism in Retina?

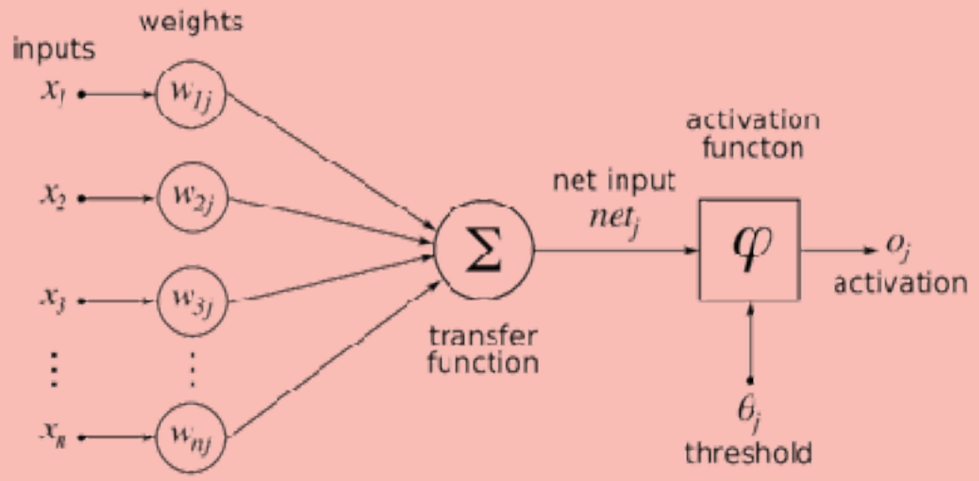
## A Computational Task

(Capturing experimental observation)

**vs.**

## A Mathematical Model

(Subject to biological constraints)




## Biologically-Realistic Timescale

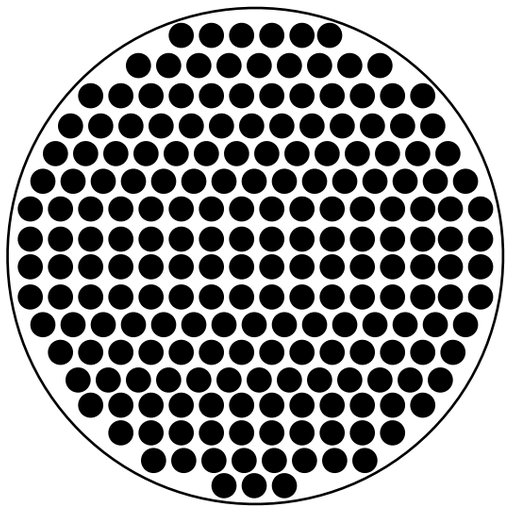
(Adaptation happens in few seconds)

# Receptive Field

## (of a Ganglion Cell)

# Receptive Field

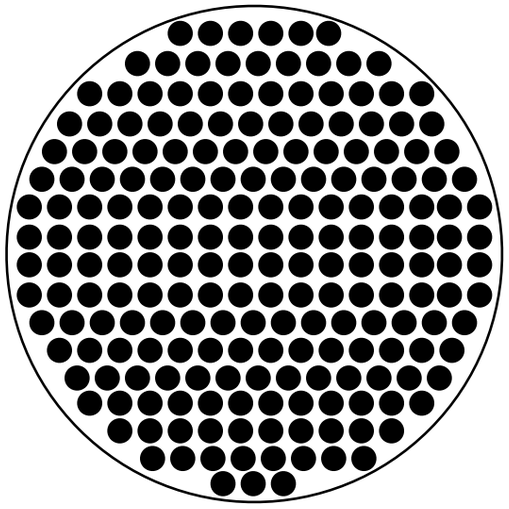
## (of a Ganglion Cell)



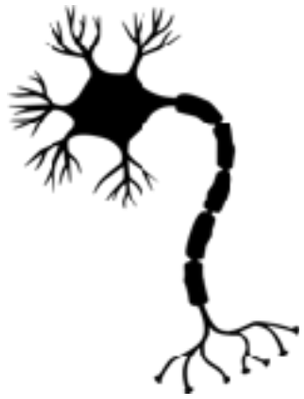
**Photoreceptors**

# Receptive Field

## (of a Ganglion Cell)



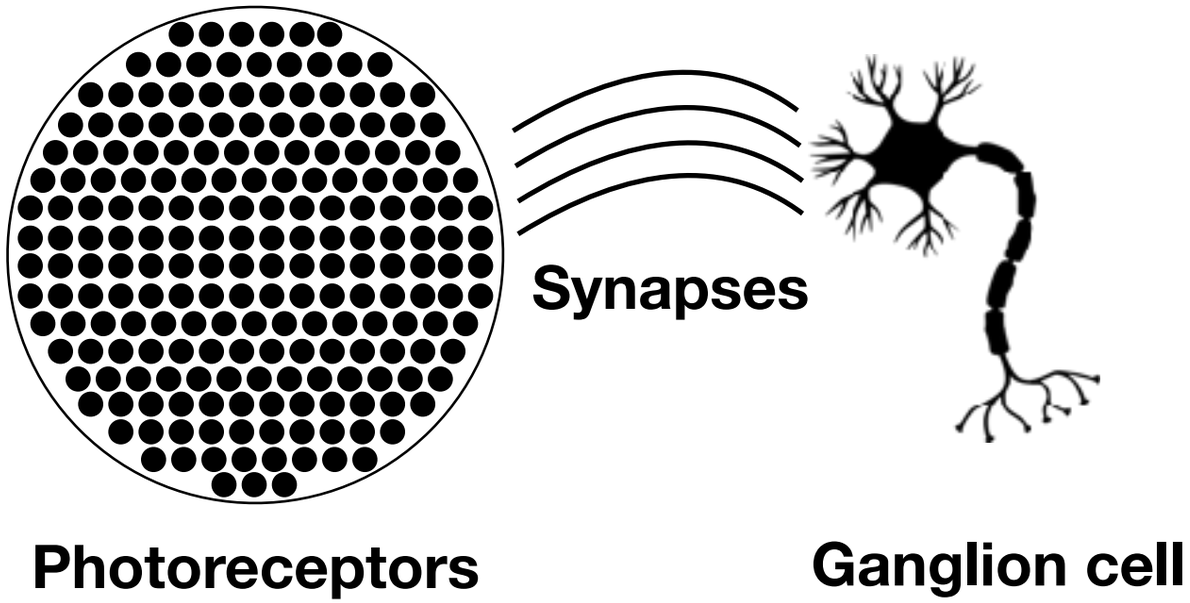
**Photoreceptors**



**Ganglion cell**

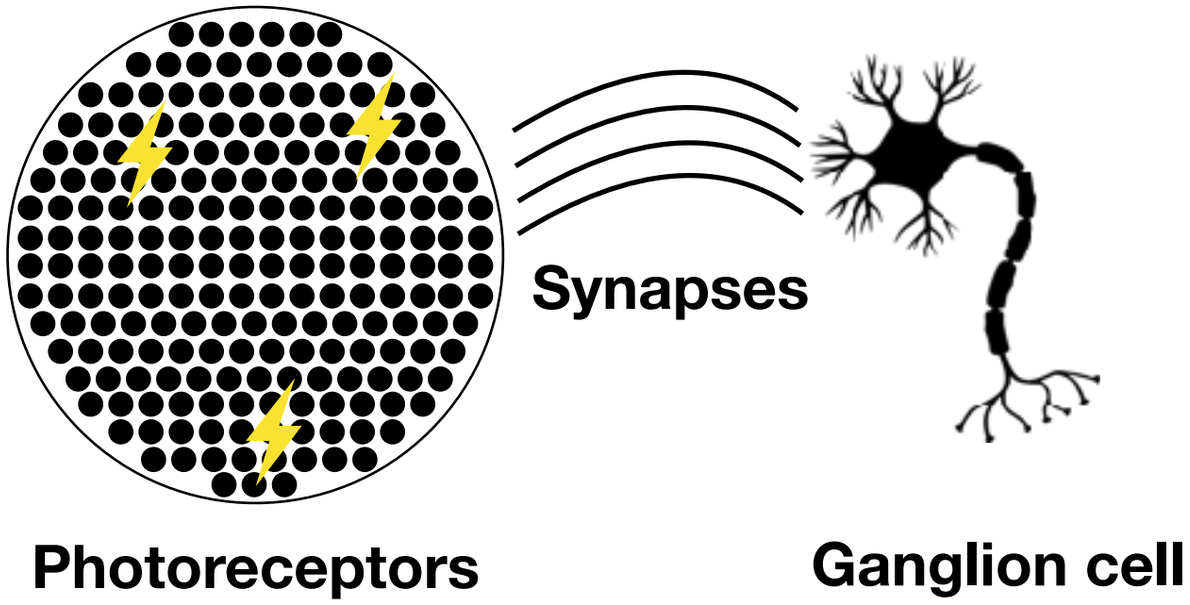
# Receptive Field

## (of a Ganglion Cell)



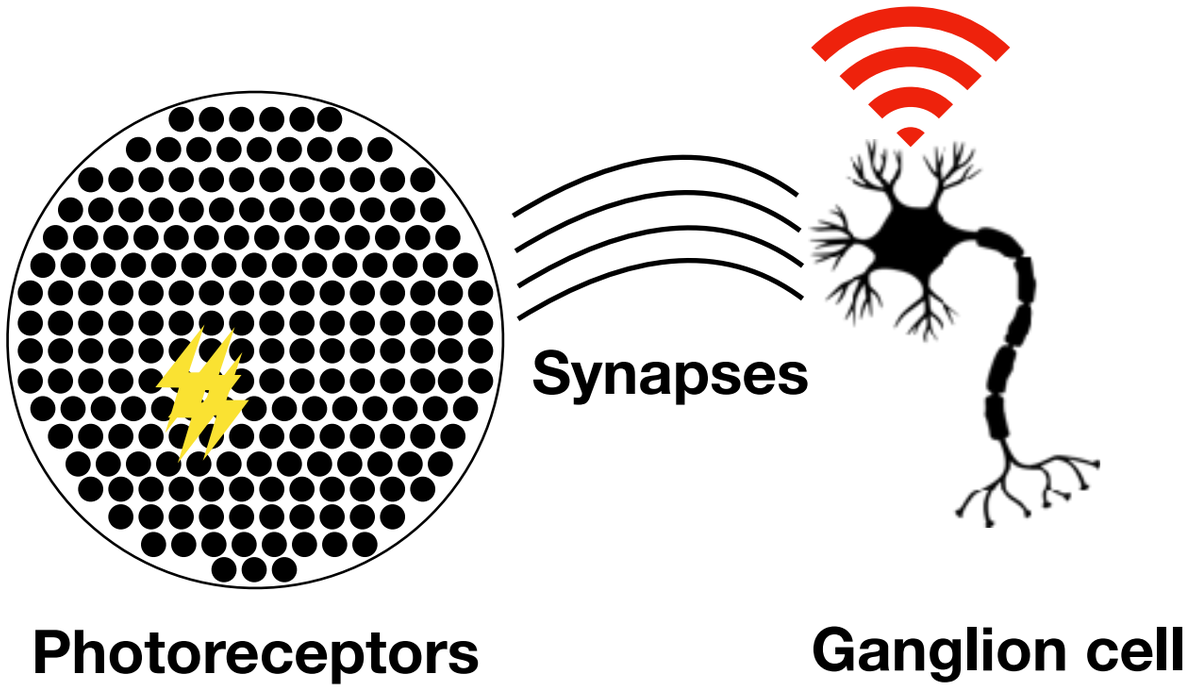
# Receptive Field

## (of a Ganglion Cell)



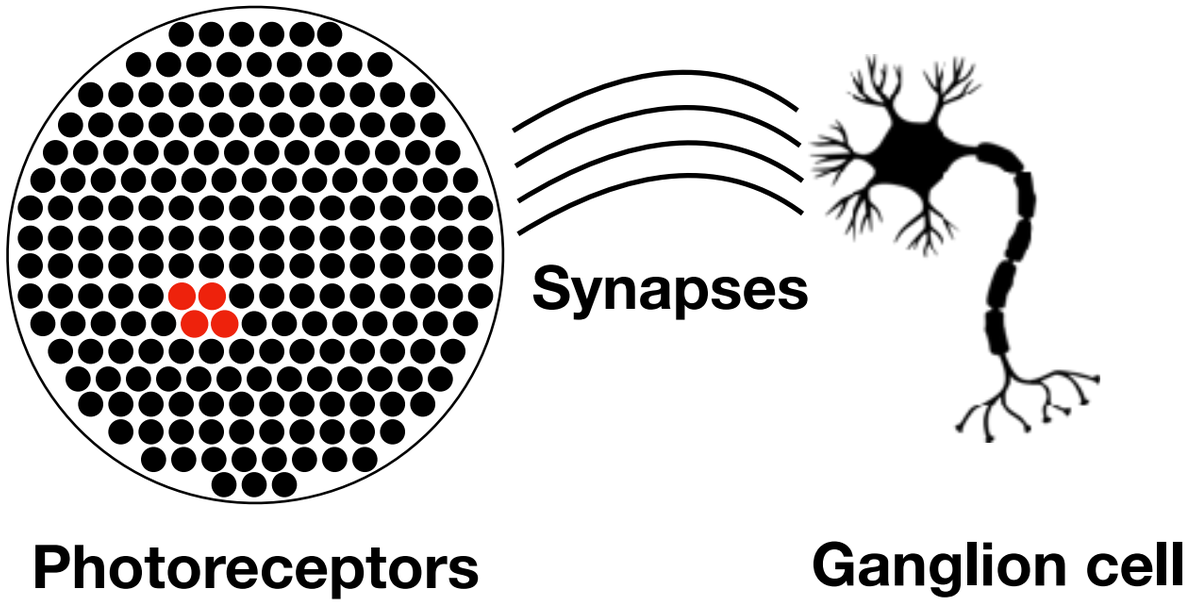
# Receptive Field

## (of a Ganglion Cell)



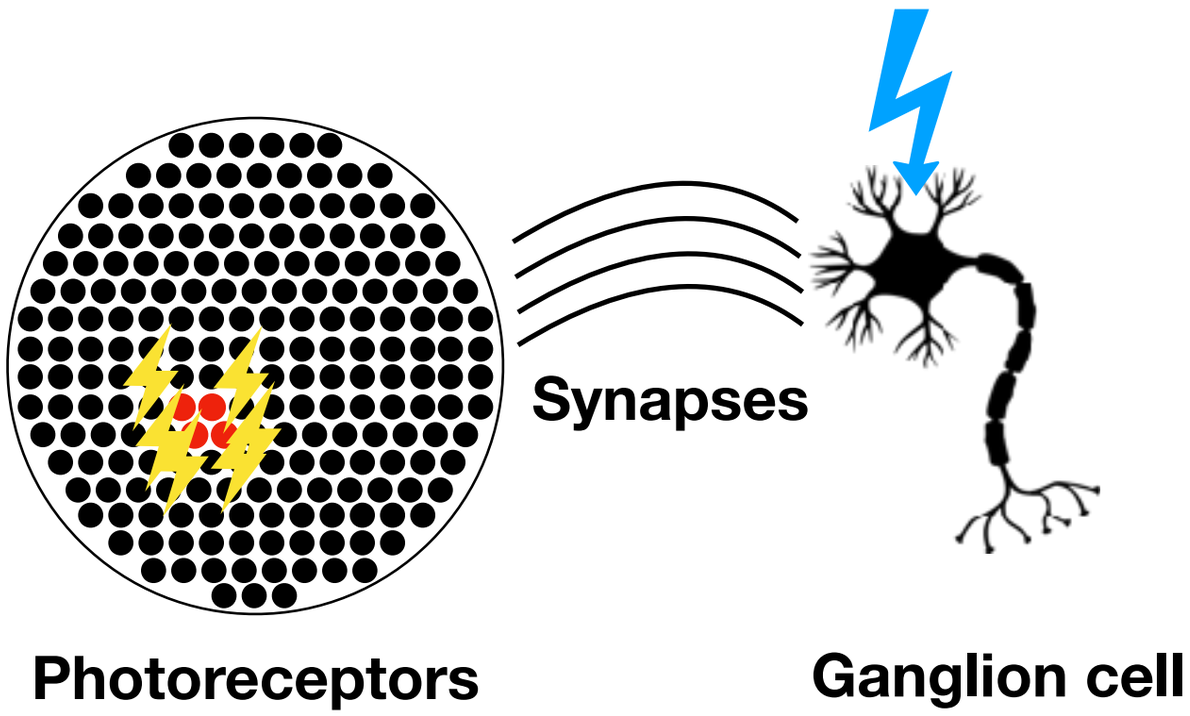
# Receptive Field

## (of a Ganglion Cell)



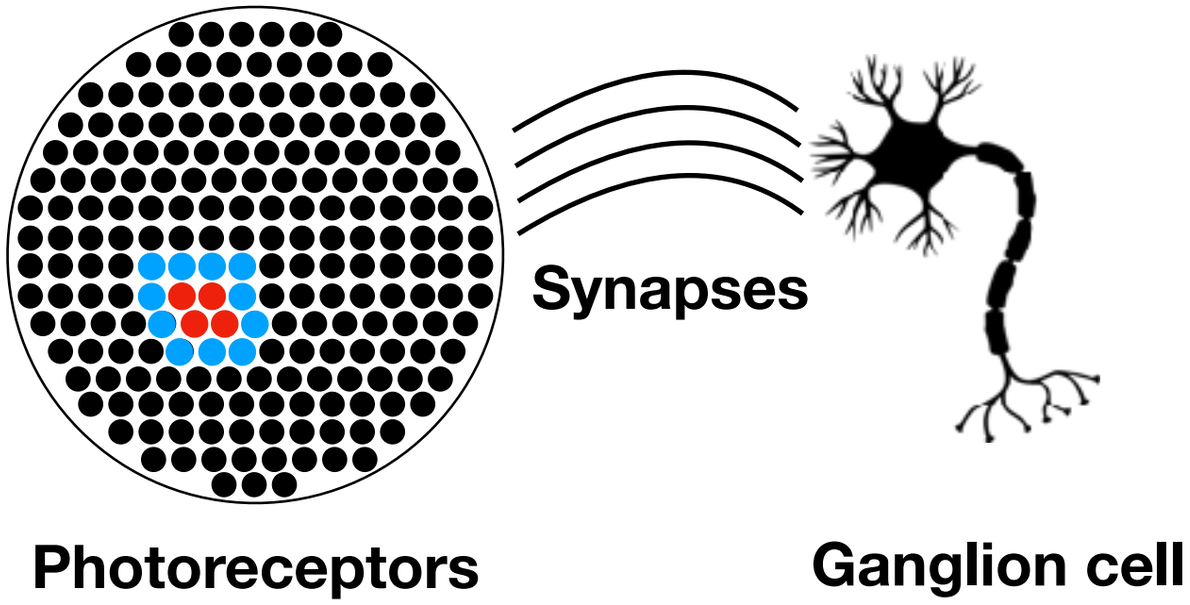
# Receptive Field

## (of a Ganglion Cell)



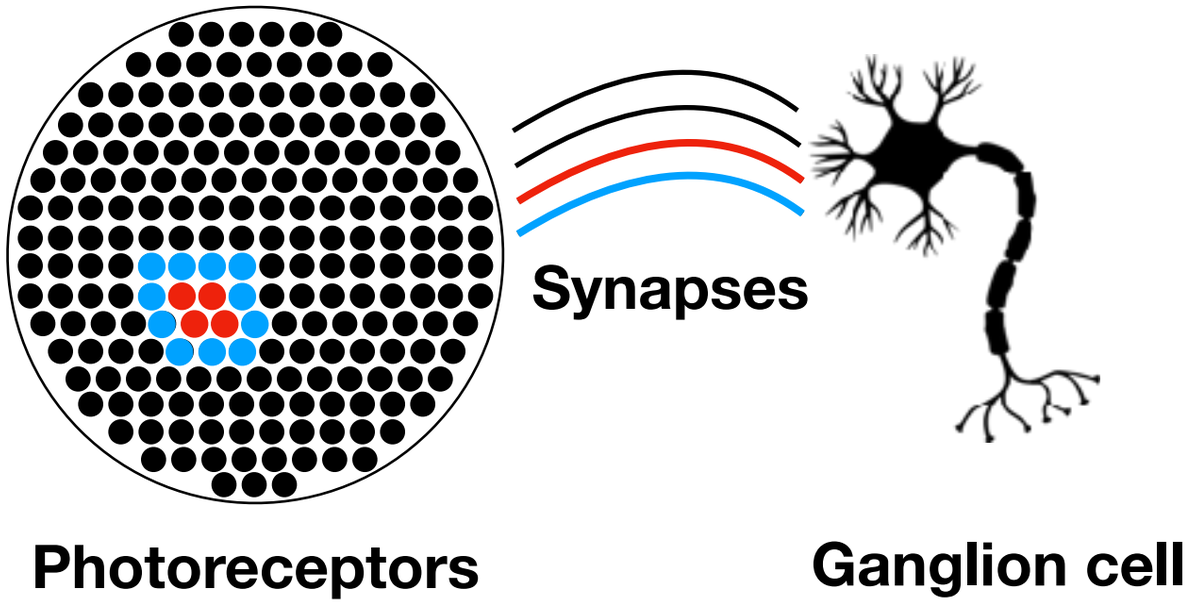
# Receptive Field

## (of a Ganglion Cell)



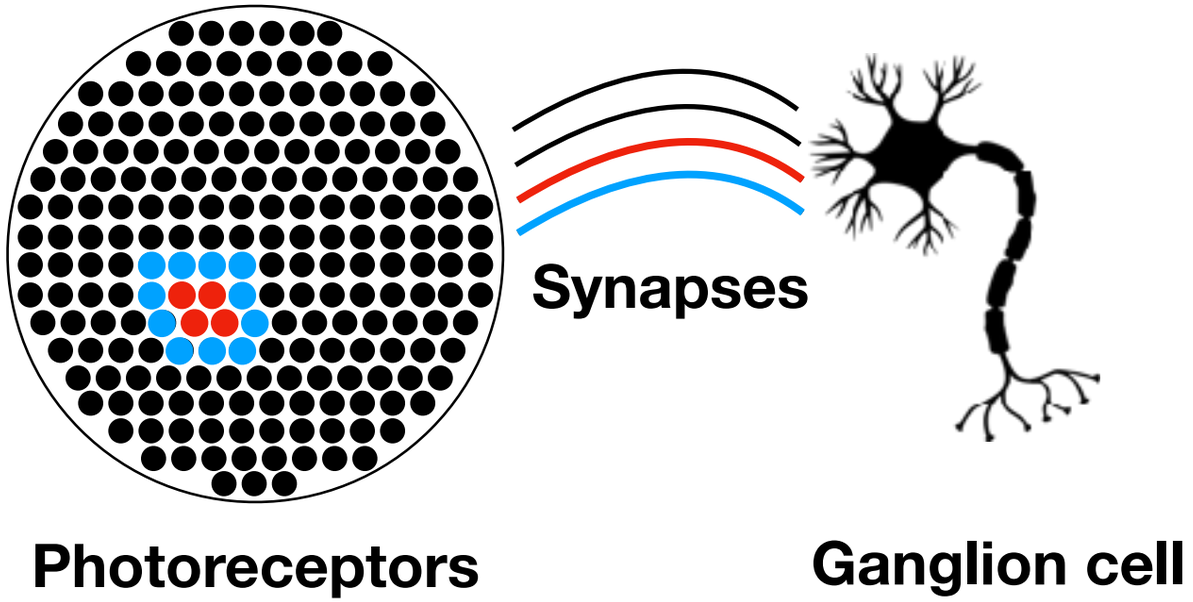
# Receptive Field

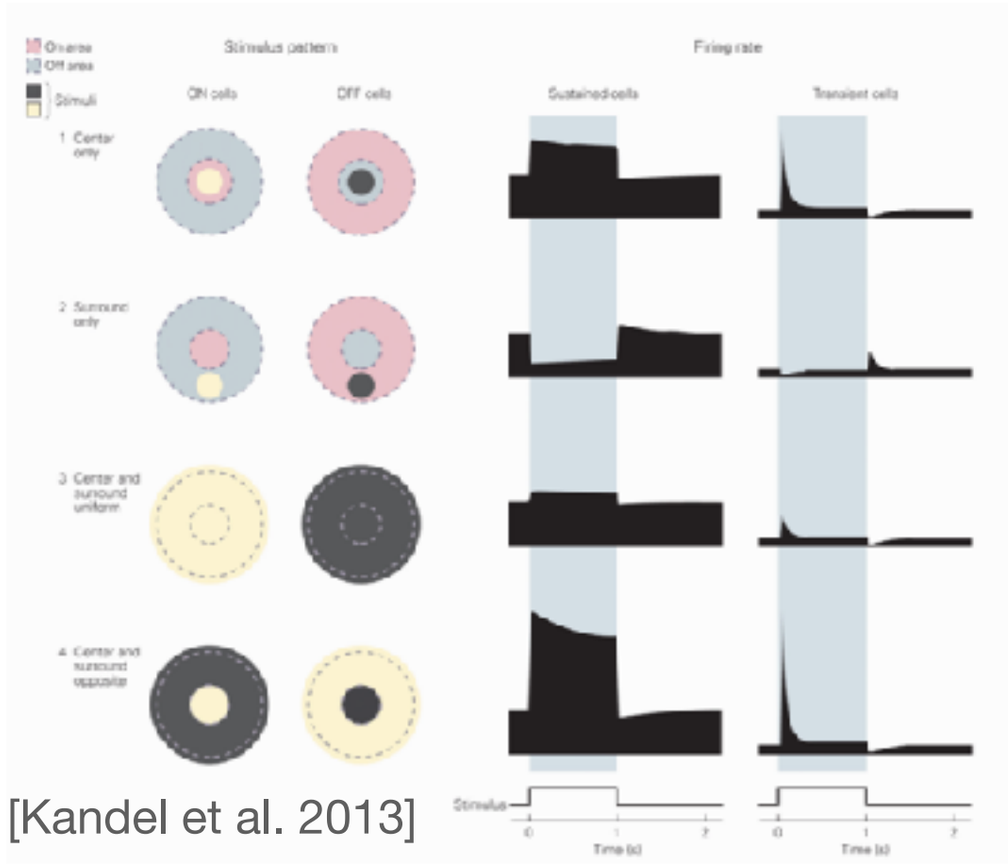
## (of a Ganglion Cell)



# Receptive Field

## (of a Ganglion Cell)



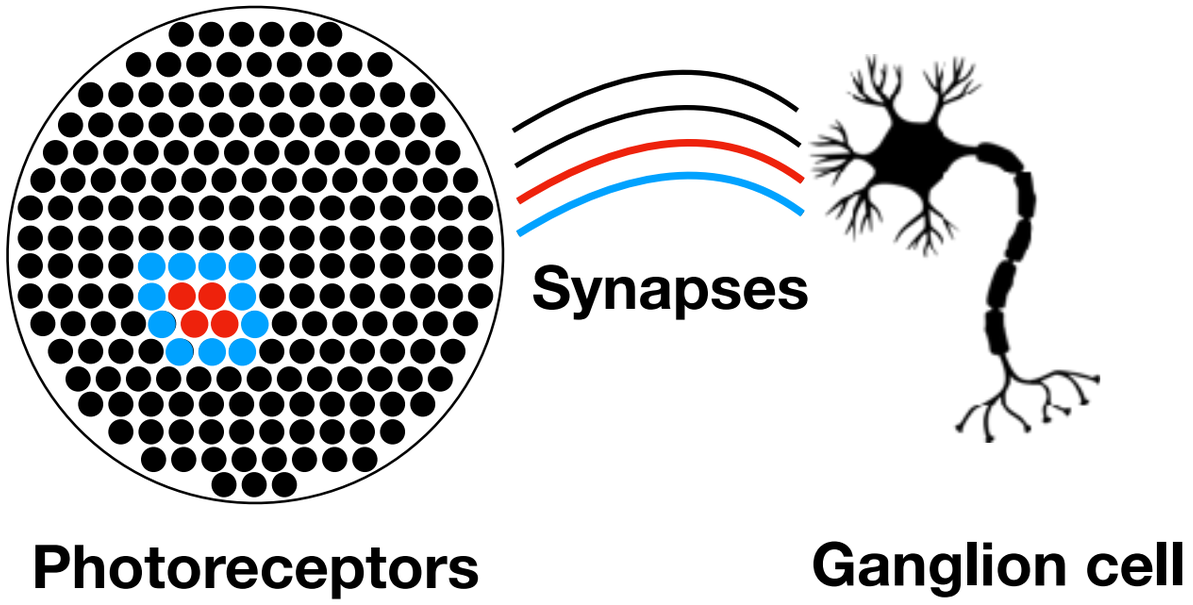



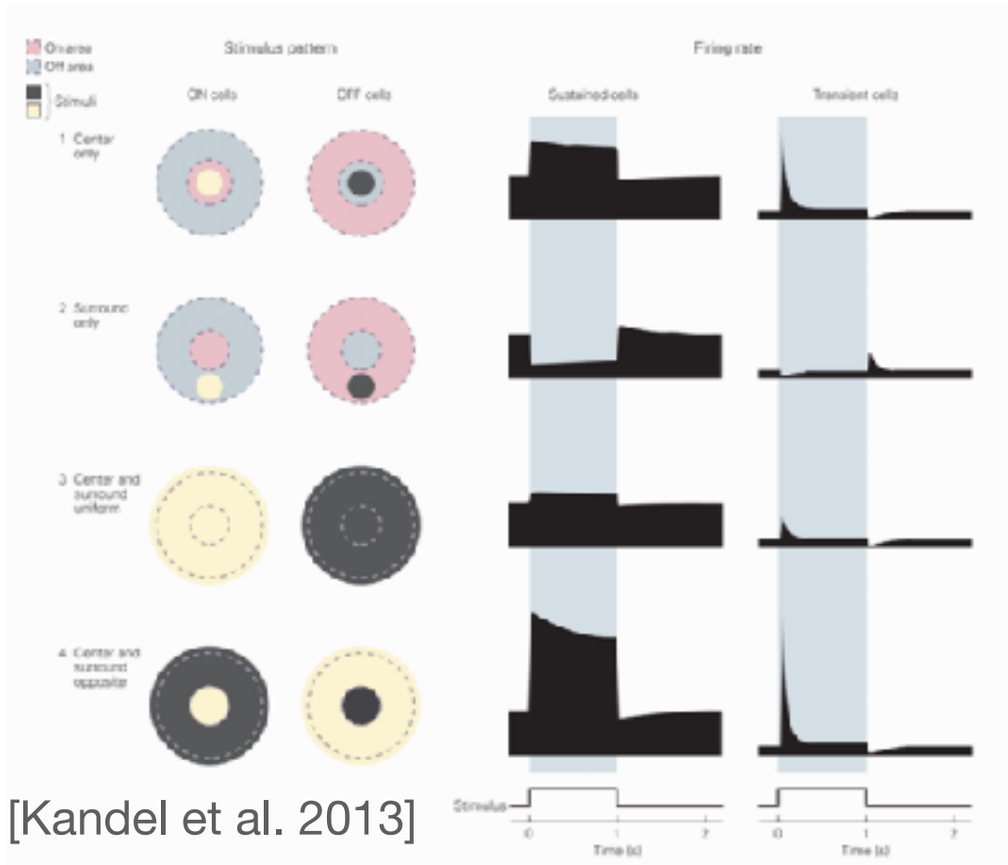

[Kandel et al. 2013]

- A signature of a ganglion cell is the center-surround receptive field.

# Receptive Field

## (of a Ganglion Cell)




[Kandel et al. 2013]

- A signature of a ganglion cell is the center-surround receptive field.
- [Atick, Redlich 1990] Principal component analysis (PCA) induces center-surround receptive fields and maximize information.

# Streaming PCA

# Streaming PCA

- **Unknown:** A distribution $\mathcal{D}$ over the unit sphere of $\mathbb{R}^n$.

# Streaming PCA

- **Unknown:** A distribution $\mathcal{D}$ over the unit sphere of $\mathbb{R}^n$.
- **Covariance matrix**: $\Sigma = \mathop{\mathbb{E}}_{\mathbf{x} \sim \mathcal{D}} [\mathbf{x}\mathbf{x}^\top] \in \mathbb{R}^{n \times n}$.

# Streaming PCA

- **Unknown:** A distribution $\mathcal{D}$ over the unit sphere of $\mathbb{R}^n$.
- **Covariance matrix**: $\Sigma = \mathop{\mathbb{E}}_{\mathbf{x}\sim\mathcal{D}}[\mathbf{x}\mathbf{x}^\top] \in \mathbb{R}^{n\times n}$.

**Input:** $\mathbf{x}_1, \ldots, \mathbf{x}_T \sim \mathcal{D}$ independent samples in a stream.

# Streaming PCA

- **Unknown:** A distribution $\mathcal{D}$ over the unit sphere of $\mathbb{R}^n$.
- **Covariance matrix**: $\Sigma = \mathop{\mathbb{E}}_{\mathbf{x}\sim\mathcal{D}}[\mathbf{x}\mathbf{x}^\top] \in \mathbb{R}^{n\times n}$.

**Input:** $\mathbf{x}_1, \ldots, \mathbf{x}_T \sim \mathcal{D}$ independent samples in a stream.

**Goal:** Using $O(n)$ space to approximate the top eigenvector $\mathbf{v}_1$ of the covariance matrix $\Sigma$.

# Streaming PCA

- **Unknown:** A distribution $\mathcal{D}$ over the unit sphere of $\mathbb{R}^n$.
- **Covariance matrix**: $\Sigma = \mathop{\mathbb{E}}_{\mathbf{x}\sim\mathcal{D}}[\mathbf{x}\mathbf{x}^\top] \in \mathbb{R}^{n\times n}$.

**Input:** $\mathbf{x}_1, \ldots, \mathbf{x}_T \sim \mathcal{D}$ independent samples in a stream.

**Goal:** Using $O(n)$ space to approximate the top eigenvector $\mathbf{v}_1$ of the covariance matrix $\Sigma$.

- A classic and well-studied non-convex optimization problem.

# Streaming PCA

- **Unknown:** A distribution $\mathcal{D}$ over the unit sphere of $\mathbb{R}^n$.
- **Covariance matrix**: $\Sigma = \mathop{\mathbb{E}}_{\mathbf{x}\sim\mathcal{D}}[\mathbf{x}\mathbf{x}^\top] \in \mathbb{R}^{n\times n}$.

**Input:** $\mathbf{x}_1, \ldots, \mathbf{x}_T \sim \mathcal{D}$ independent samples in a stream.

**Goal:** Using $O(n)$ space to approximate the top eigenvector $\mathbf{v}_1$ of the covariance matrix $\Sigma$.

- A classic and well-studied non-convex optimization problem.

**How does the retina implement streaming PCA?**

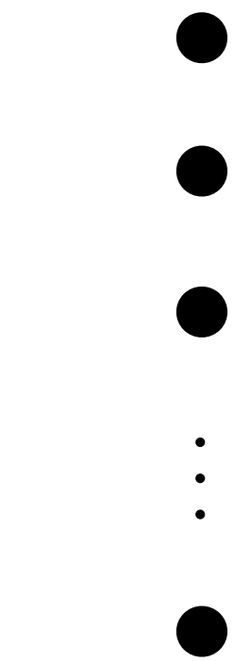
# A (Simplified) Mathematical Model for Retina

# A (Simplified) Mathematical Model for Retina

A feedforward 2-layer circuit with single output neuron.

# A (Simplified) Mathematical Model for Retina

A feedforward 2-layer circuit with single output neuron.



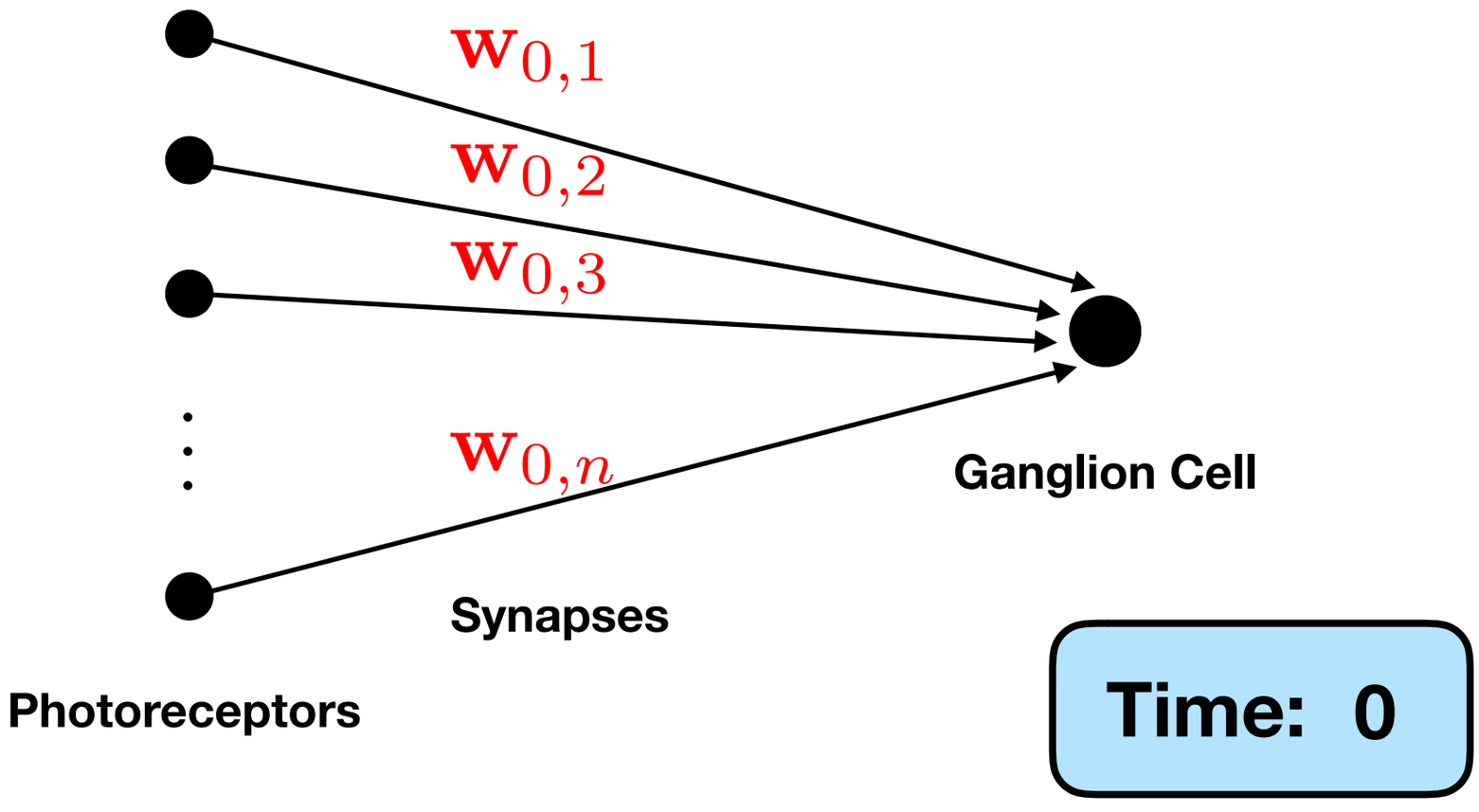
**Photoreceptors**

# A (Simplified) Mathematical Model for Retina

A feedforward 2-layer circuit with single output neuron.

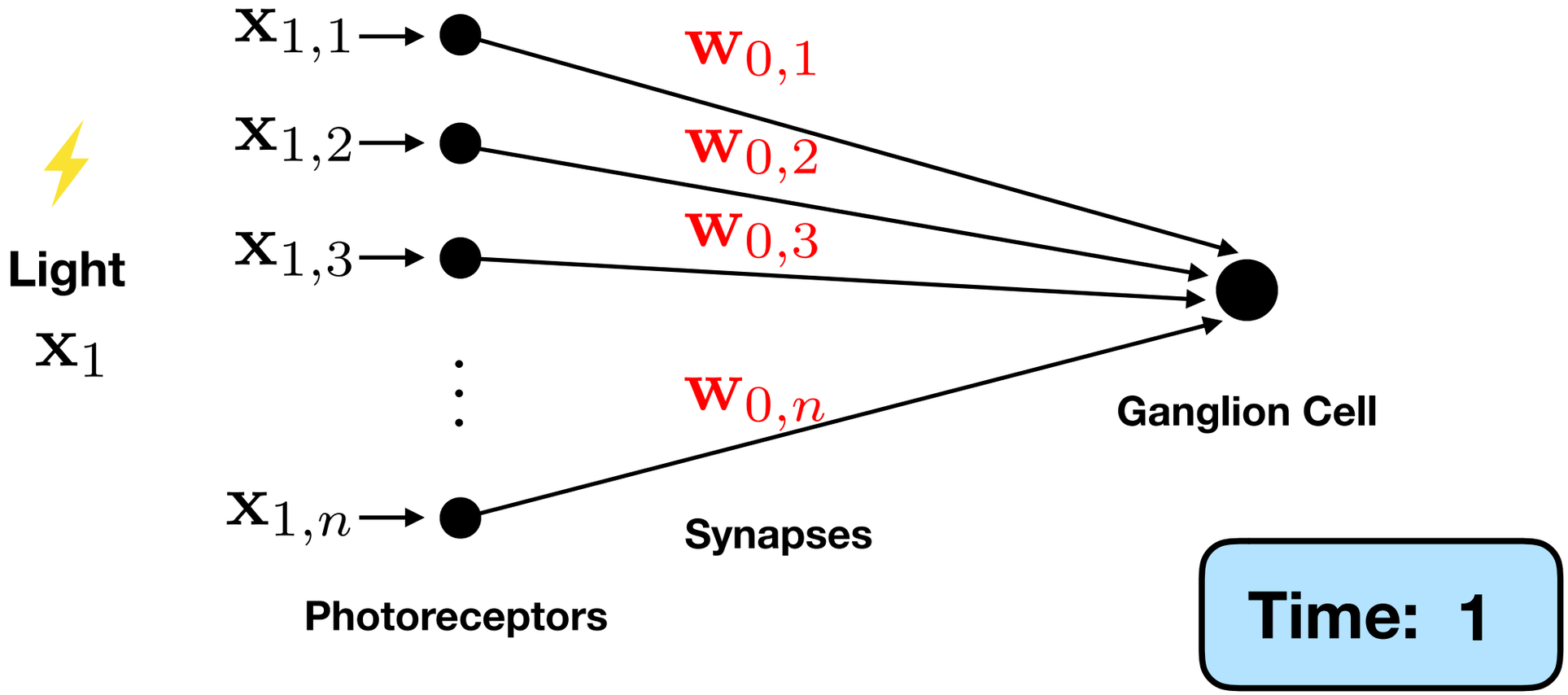
# A (Simplified) Mathematical Model for Retina

A feedforward 2-layer circuit with single output neuron.



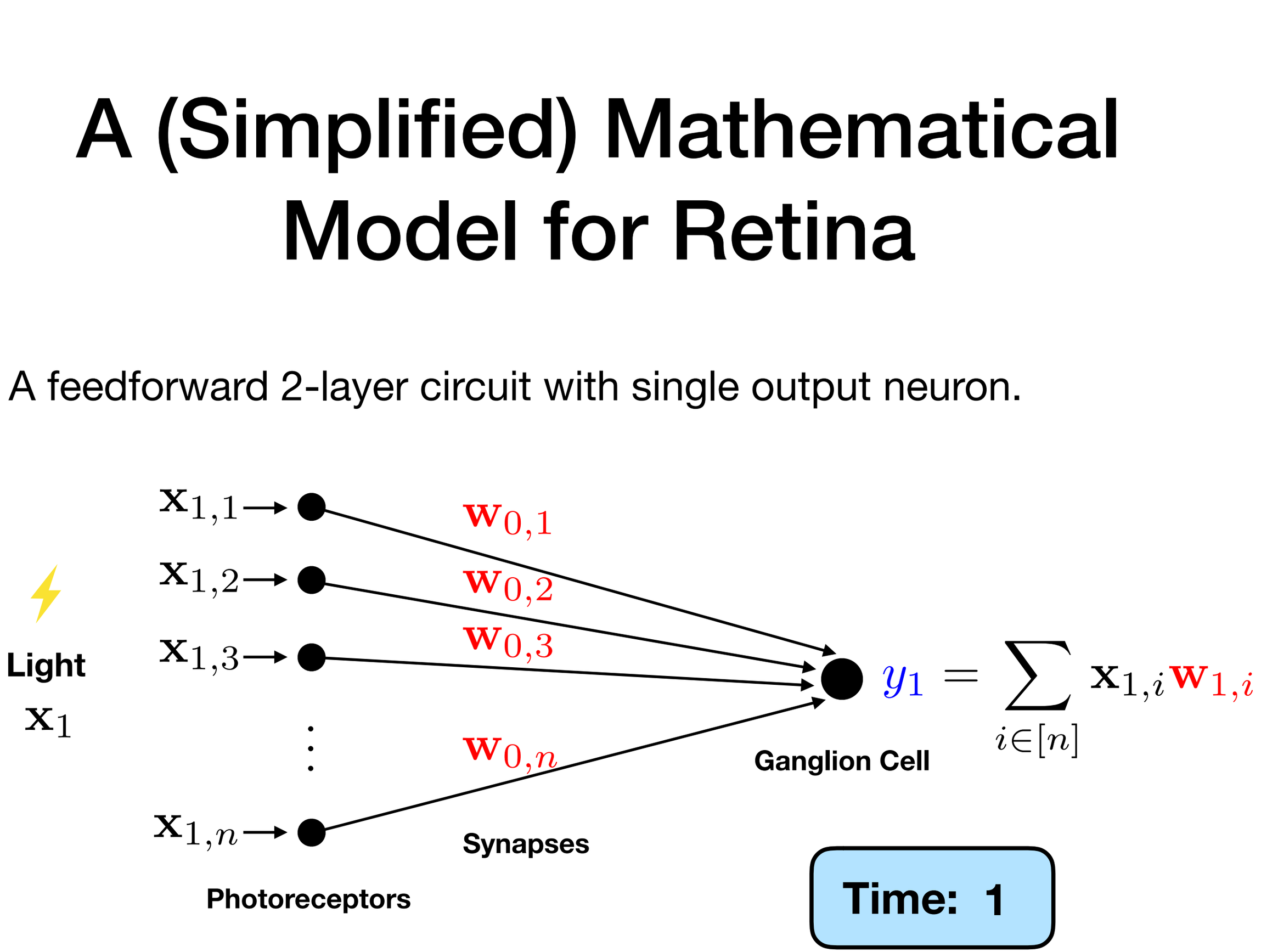
# A (Simplified) Mathematical Model for Retina

A feedforward 2-layer circuit with single output neuron.



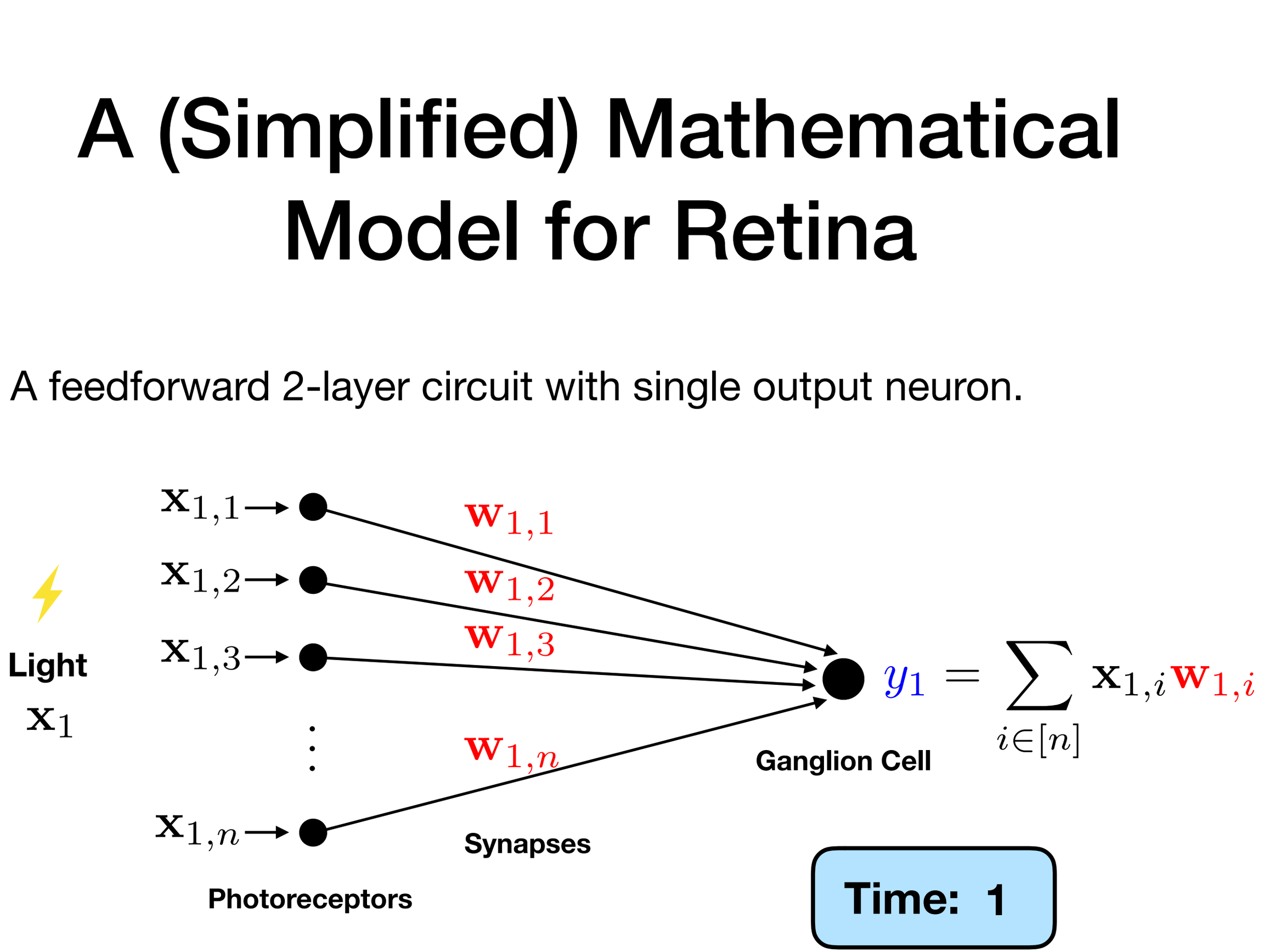
# A (Simplified) Mathematical Model for Retina

A feedforward 2-layer circuit with single output neuron.



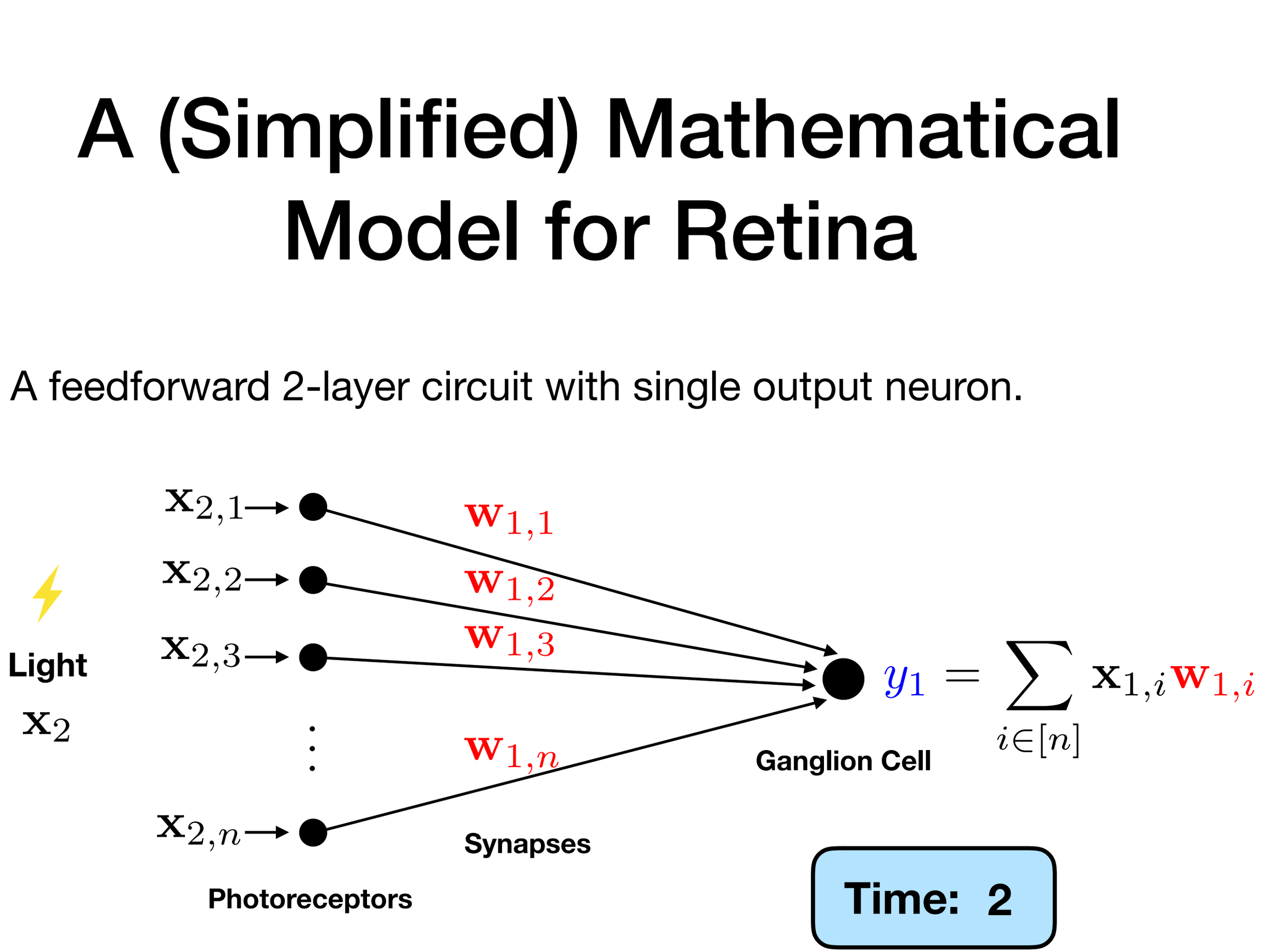
# A (Simplified) Mathematical Model for Retina

A feedforward 2-layer circuit with single output neuron.



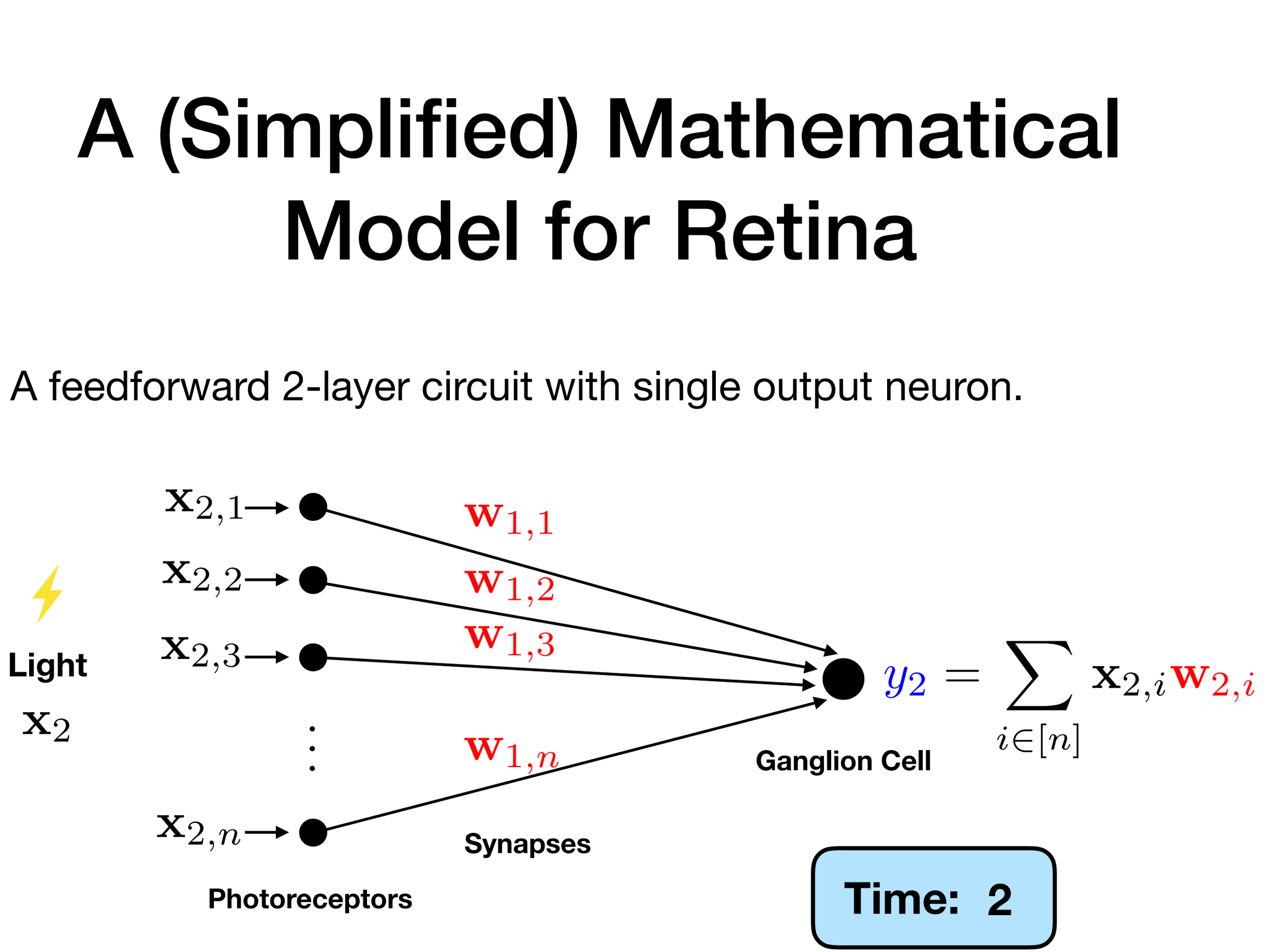
# A (Simplified) Mathematical Model for Retina

A feedforward 2-layer circuit with single output neuron.



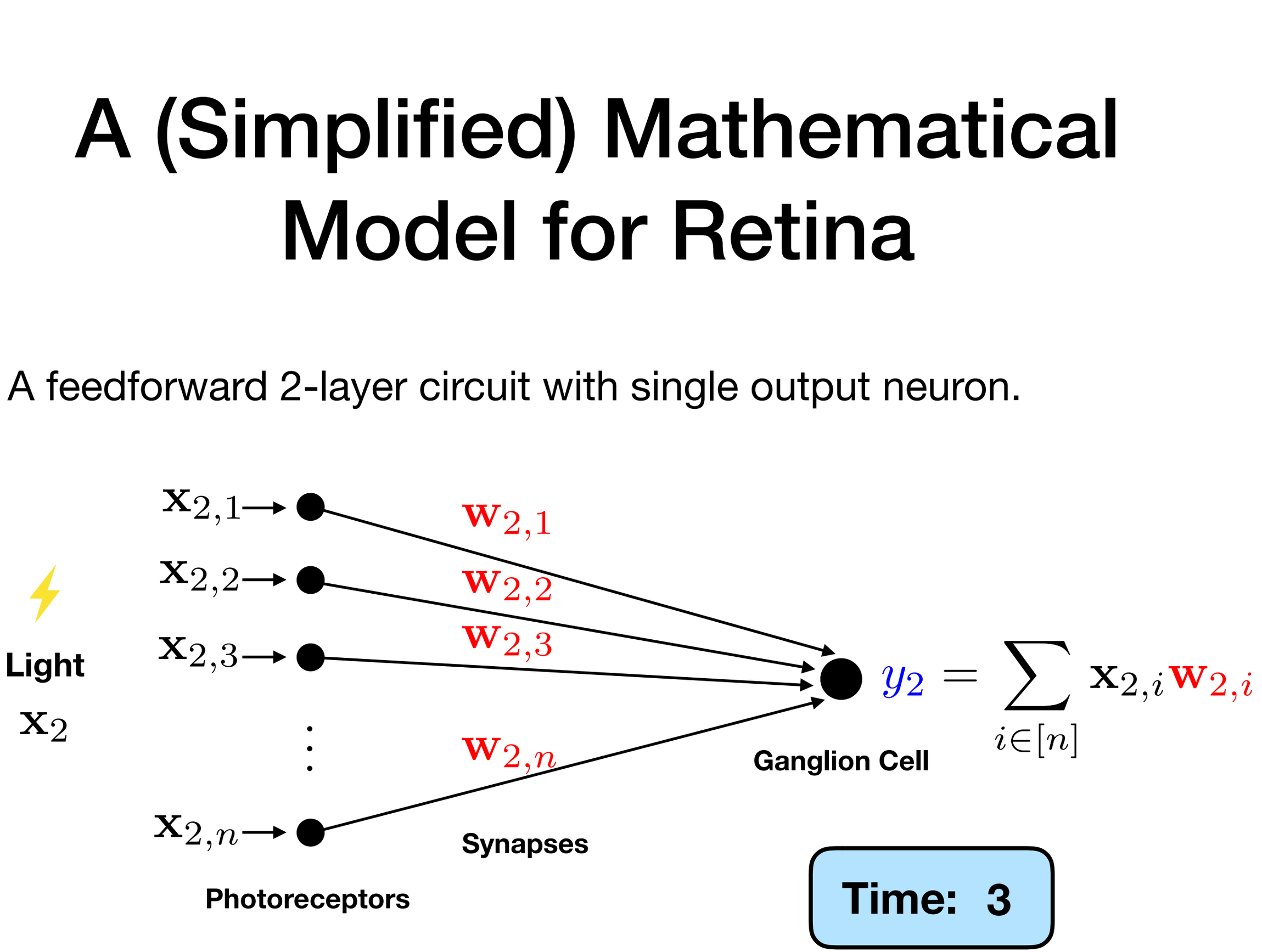
# A (Simplified) Mathematical Model for Retina

A feedforward 2-layer circuit with single output neuron.



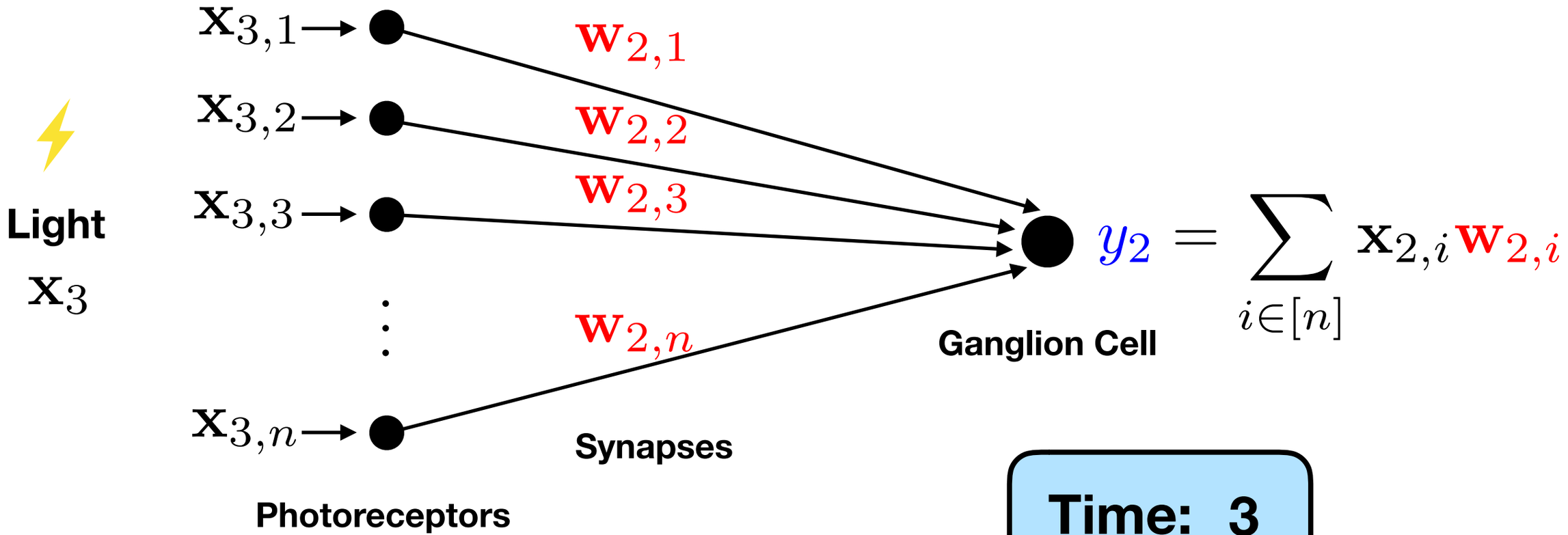
# A (Simplified) Mathematical Model for Retina

A feedforward 2-layer circuit with single output neuron.



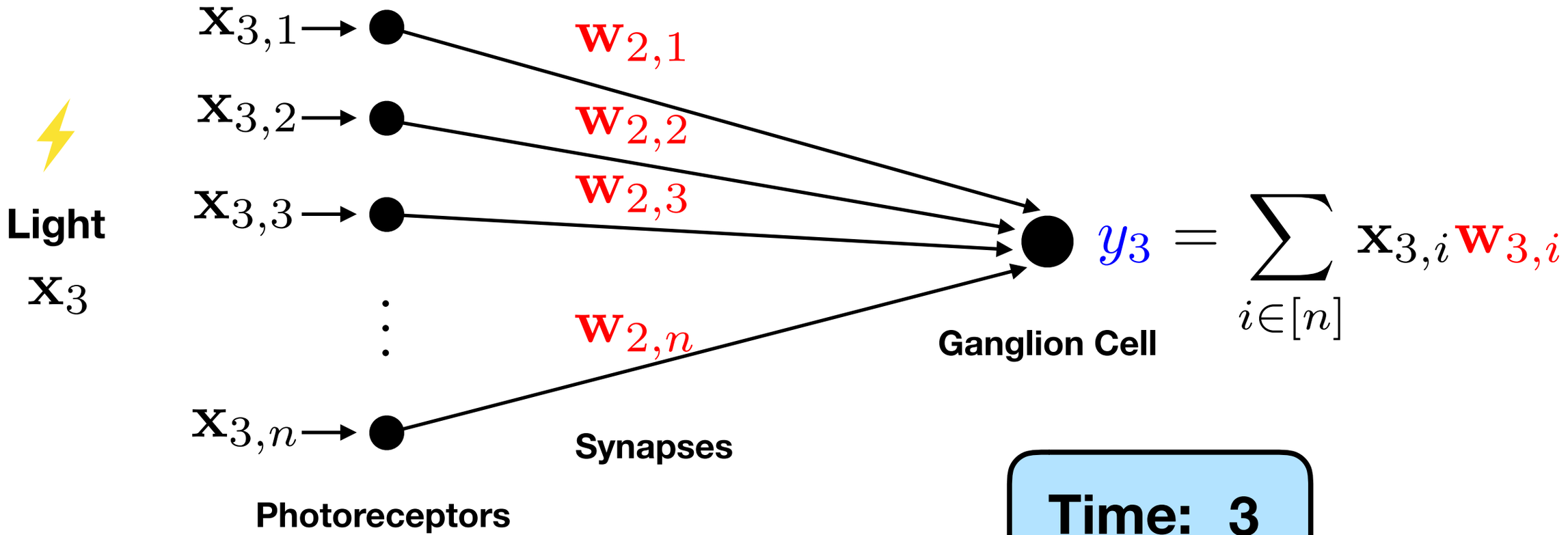
# A (Simplified) Mathematical Model for Retina

A feedforward 2-layer circuit with single output neuron.



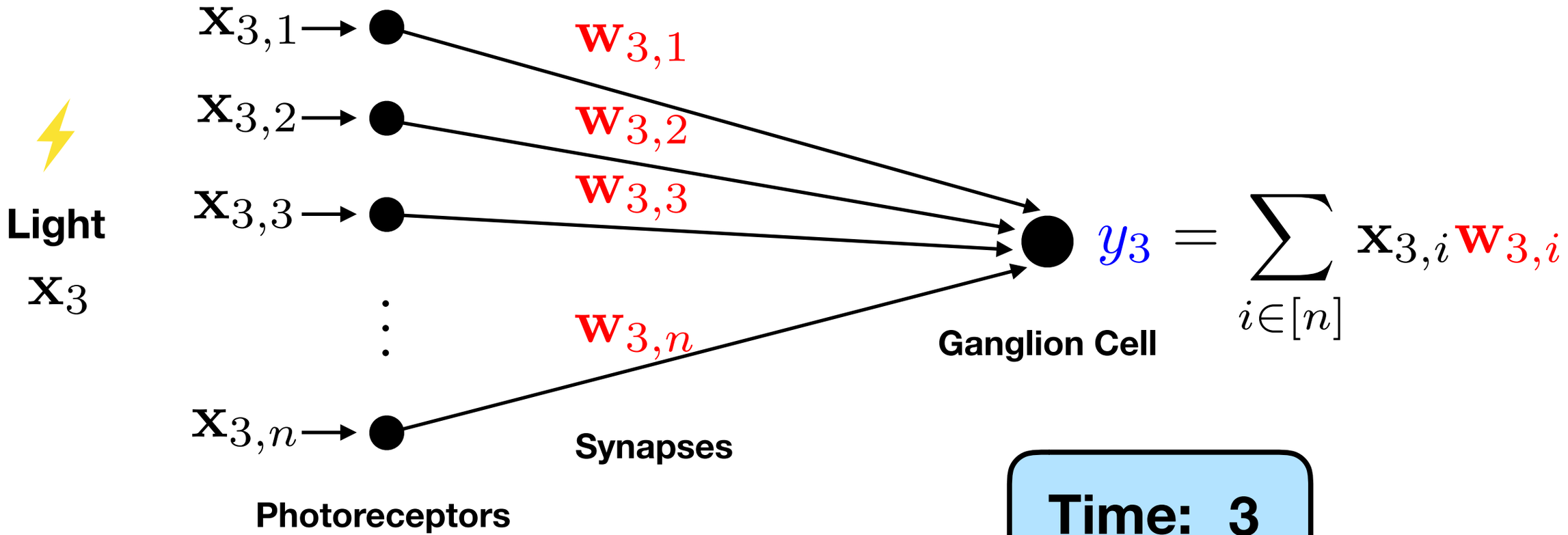
# A (Simplified) Mathematical Model for Retina

A feedforward 2-layer circuit with single output neuron.



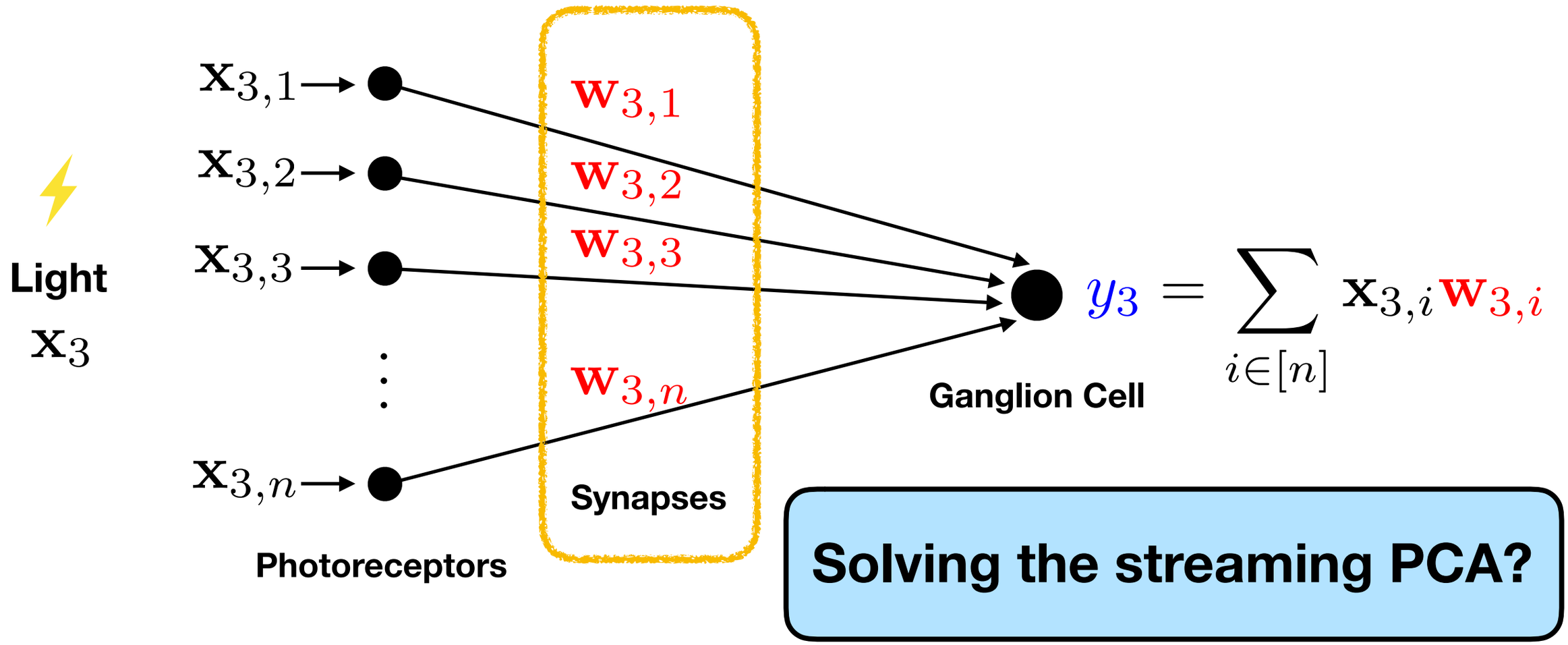
# A (Simplified) Mathematical Model for Retina

A feedforward 2-layer circuit with single output neuron.



# A (Simplified) Mathematical Model for Retina

A feedforward 2-layer circuit with single output neuron.




**Solving the streaming PCA?**

# Synaptic Learning

# Synaptic Learning

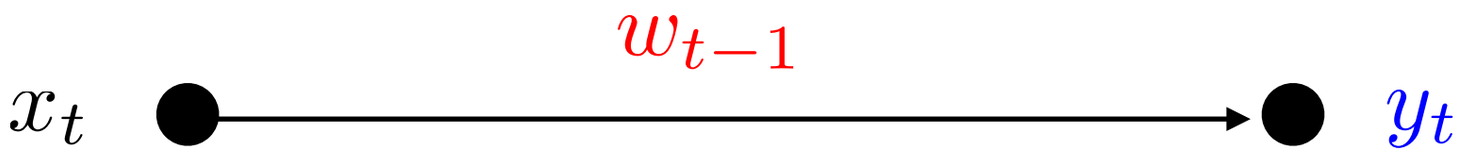
- The synapse varies w.r.t. time, a.k.a., *synaptic plasticity*.

# Synaptic Learning

- The synapse varies w.r.t. time, a.k.a., *synaptic plasticity*.
- The update should be biologically-plausible, *e.g.,* local.

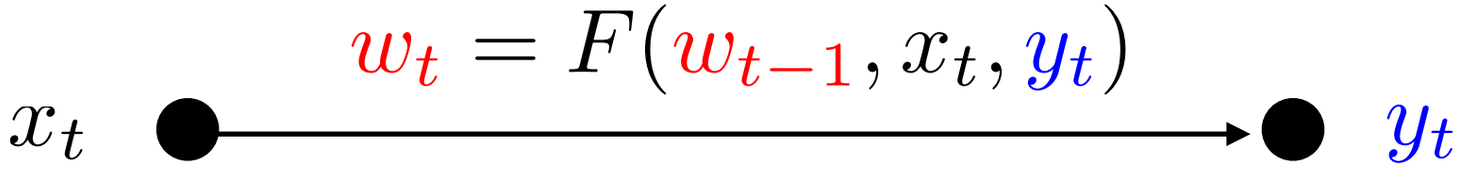
# Synaptic Learning

- The synapse varies w.r.t. time, a.k.a., *synaptic plasticity*.
- The update should be biologically-plausible, *e.g.,* local.



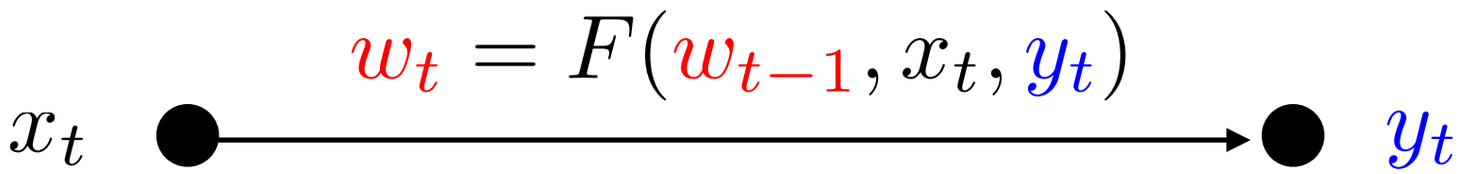
# Synaptic Learning

- The synapse varies w.r.t. time, a.k.a., *synaptic plasticity*.
- The update should be biologically-plausible, *e.g.,* local.



# Synaptic Learning

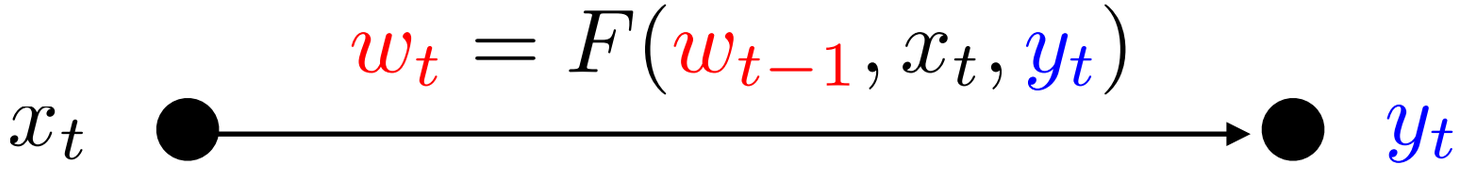
- The synapse varies w.r.t. time, a.k.a., *synaptic plasticity*.
- The update should be biologically-plausible, *e.g.,* local.




- **The Hebbian plasticity** [Hebb 1949]**:** Update is local and fire together wire together, e.g., $w_t = w_{t-1} + x_t y_t$.

# Synaptic Learning

- The synapse varies w.r.t. time, a.k.a., *synaptic plasticity*.
- The update should be biologically-plausible, *e.g.,* local.




- **The Hebbian plasticity** [Hebb 1949]**:** Update is local and fire together wire together, e.g., $w_t = w_{t-1} + x_t y_t$.
- **The Homeostatic plasticity** [Turrigiano 2008]: Stabilize the network by normalizing the strength of incoming synapses.

# What's the Compression Mechanism in Retina?

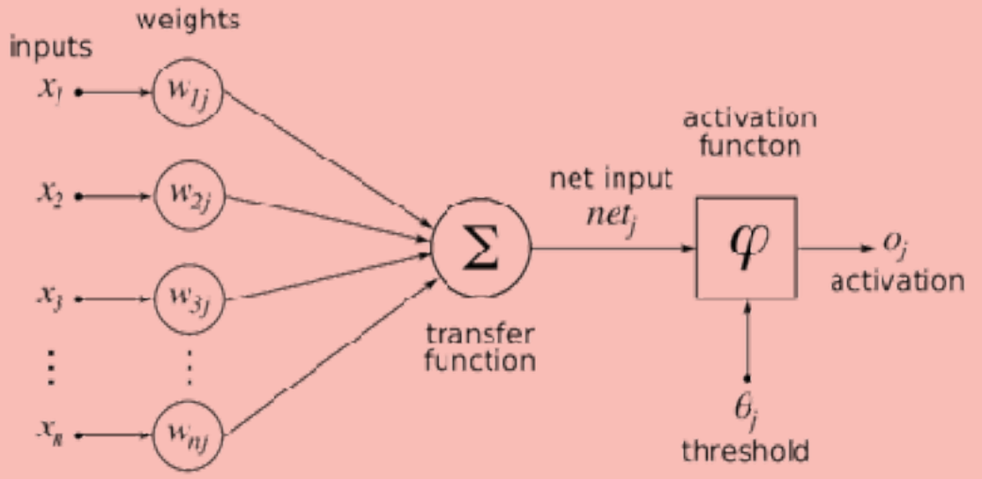
## A Computational Task

(Capturing Experimental Observation)

**vs.**

## A Mathematical Model

(Subject to biological constraints)




## Biologically-Realistic Timescale

(Adaptation happens in few seconds)

# What's the Compression Mechanism in Retina?

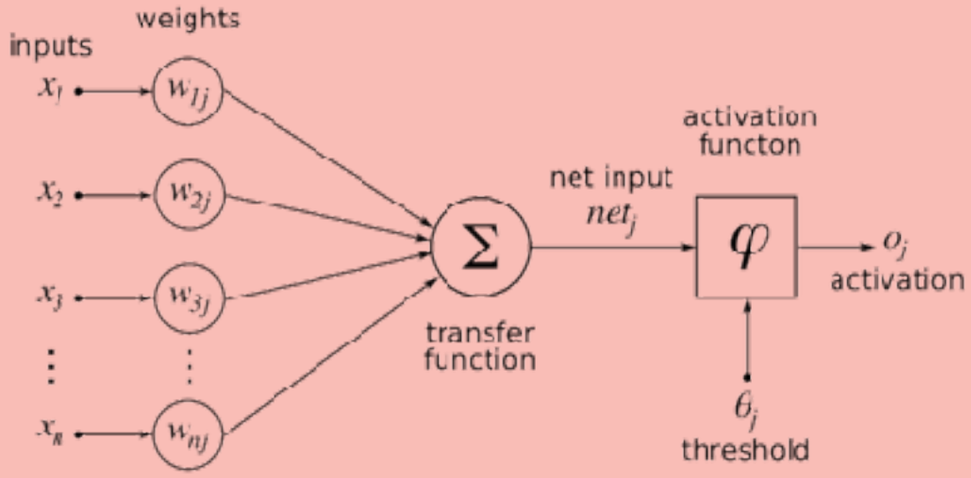
## A Computational Task

(Capturing Experimental Observation)

**Streaming PCA**

## A Mathematical Model

(Subject to biological constraints)




## Biologically-Realistic Timescale

(Adaptation happens in few seconds)

# What's the Compression Mechanism in Retina?

**A Computational Task**

(Capturing Experimental Observation)

**Streaming PCA**

**A Mathematical Model**

(Subject to biological constraints)

**Synaptic Learning**

**Biologically-Realistic Timescale**

(Adaptation happens in few seconds)

# What's the Compression Mechanism in Retina?

**A Computational Task**

(Capturing Experimental Observation)

**Streaming PCA**

**A Mathematical Model**

(Subject to biological constraints)

**Synaptic Learning**

**Biologically-Realistic Timescale**

(Adaptation happens in few seconds)

**Tiny Dependency on #Photoreceptors!**

# Oja's Rule [Oja 1982]

# Oja's Rule [Oja 1982]

**Homeostatic Plasticity**
(Synaptic scaling)

**Hebbian Plasticity**
(Local update)

# Oja's Rule [Oja 1982]

**Homeostatic Plasticity**
(Synaptic scaling)

**Hebbian Plasticity**
(Local update)

**Idea**: Mimic power method.

# Oja's Rule [Oja 1982]

**Homeostatic Plasticity**
(Synaptic scaling)

**Hebbian Plasticity**
(Local update)

**Idea**: Mimic power method.

Learning rate

$$\mathbf{w}_t \leftarrow (I + \eta_t \cdot \mathbf{x}_t \mathbf{x}_t^\top)\mathbf{w}_{t-1}$$

# Oja's Rule [Oja 1982]

**Homeostatic Plasticity**
(Synaptic scaling)

**Hebbian Plasticity**
(Local update)

**Idea**: Mimic power method.

Learning rate

$$\mathbf{w}_t \leftarrow (I + \eta_t \cdot \mathbf{x}_t \mathbf{x}_t^\top)\mathbf{w}_{t-1}$$

Violate **synaptic scaling**! How about normalization?

# Oja's Rule [Oja 1982]

**Homeostatic Plasticity**
(Synaptic scaling)

**Hebbian Plasticity**
(Local update)

**Idea**: Mimic power method.

Learning rate

$$\mathbf{w}_t \leftarrow (I + \eta_t \cdot \mathbf{x}_t \mathbf{x}_t^\top)\mathbf{w}_{t-1}$$

Violate **synaptic scaling**! How about normalization?

$$\mathbf{w}_t \leftarrow \frac{(I + \eta_t \cdot \mathbf{x}_t \mathbf{x}_t^\top)\mathbf{w}_{t-1}}{\|(I + \eta_t \cdot \mathbf{x}_t \mathbf{x}_t^\top)\mathbf{w}_{t-1}\|_2}$$

# Oja's Rule [Oja 1982]

**Homeostatic Plasticity**
(Synaptic scaling) ✓

**Hebbian Plasticity**
(Local update)

**Idea**: Mimic power method.

Learning rate

$$\mathbf{w}_t \leftarrow (I + \eta_t \cdot \mathbf{x}_t \mathbf{x}_t^\top)\mathbf{w}_{t-1}$$

Violate **synaptic scaling**! How about normalization?

$$\mathbf{w}_t \leftarrow \frac{(I + \eta_t \cdot \mathbf{x}_t \mathbf{x}_t^\top)\mathbf{w}_{t-1}}{\|(I + \eta_t \cdot \mathbf{x}_t \mathbf{x}_t^\top)\mathbf{w}_{t-1}\|_2}$$

# Oja's Rule [Oja 1982]

**Homeostatic Plasticity**
(Synaptic scaling)

**Hebbian Plasticity**
(Local update)

$$\mathbf{w}_t \leftarrow \frac{(I + \eta_t \cdot \mathbf{x}_t \mathbf{x}_t^\top)\mathbf{w}_{t-1}}{\|(I + \eta_t \cdot \mathbf{x}_t \mathbf{x}_t^\top)\mathbf{w}_{t-1}\|_2}$$

# Oja's Rule [Oja 1982]

**Homeostatic Plasticity**
(Synaptic scaling)

**Hebbian Plasticity**
(Local update)

$$\mathbf{w}_t \leftarrow \frac{(I + \eta_t \cdot \mathbf{x}_t \mathbf{x}_t^\top)\mathbf{w}_{t-1}}{\|(I + \eta_t \cdot \mathbf{x}_t \mathbf{x}_t^\top)\mathbf{w}_{t-1}\|_2}$$

The update is not local, e.g., updating $\mathbf{w}_{t,1}$ requires knowledge of $\mathbf{w}_{t,2}$. How about Taylor's expansion?

# Oja's Rule [Oja 1982]

**Homeostatic Plasticity**
(Synaptic scaling)

**Hebbian Plasticity**
(Local update)

$$\mathbf{w}_t \leftarrow \frac{(I + \eta_t \cdot \mathbf{x}_t \mathbf{x}_t^\top)\mathbf{w}_{t-1}}{\|(I + \eta_t \cdot \mathbf{x}_t \mathbf{x}_t^\top)\mathbf{w}_{t-1}\|_2}$$

The update is not local, e.g., updating $\mathbf{w}_{t,1}$ requires knowledge of $\mathbf{w}_{t,2}$. How about Taylor's expansion?

$$\mathbf{w}_t = \mathbf{w}_{t-1} + \eta_t y_t \cdot (\mathbf{x}_t - y_t \mathbf{w}_{t-1})$$

where $y_t = \mathbf{x}_t^\top \mathbf{w}_{t-1}$ is the correlation.

# Oja's Rule [Oja 1982]

**Homeostatic Plasticity**
(Synaptic scaling)

**Hebbian Plasticity**
(Local update)

$$\mathbf{w}_t \leftarrow \frac{(I + \eta_t \cdot \mathbf{x}_t \mathbf{x}_t^\top)\mathbf{w}_{t-1}}{\|(I + \eta_t \cdot \mathbf{x}_t \mathbf{x}_t^\top)\mathbf{w}_{t-1}\|_2}$$

The update is not local, e.g., updating $\mathbf{w}_{t,1}$ requires knowledge of $\mathbf{w}_{t,2}$. How about Taylor's expansion?

$$\mathbf{w}_t = \mathbf{w}_{t-1} + \eta_t y_t \cdot (\mathbf{x}_t - y_t \mathbf{w}_{t-1})$$

where $y_t = \mathbf{x}_t^\top \mathbf{w}_{t-1}$ is the correlation.

# Oja's Rule [Oja 1982]

**Homeostatic Plasticity**
(Synaptic scaling)

**Hebbian Plasticity**
(Local update)

$$\mathbf{w}_t \leftarrow \frac{(I + \eta_t \cdot \mathbf{x}_t \mathbf{x}_t^\top)\mathbf{w}_{t-1}}{\|(I + \eta_t \cdot \mathbf{x}_t \mathbf{x}_t^\top)\mathbf{w}_{t-1}\|_2} \quad \textbf{(ML Oja' Rule)}$$

The update is not local, e.g., updating $\mathbf{w}_{t,1}$ requires knowledge of $\mathbf{w}_{t,2}$. How about Taylor's expansion?

$$\mathbf{w}_t = \mathbf{w}_{t-1} + \eta_t y_t \cdot (\mathbf{x}_t - y_t \mathbf{w}_{t-1}) \quad \textbf{(Bio Oja' Rule)}$$

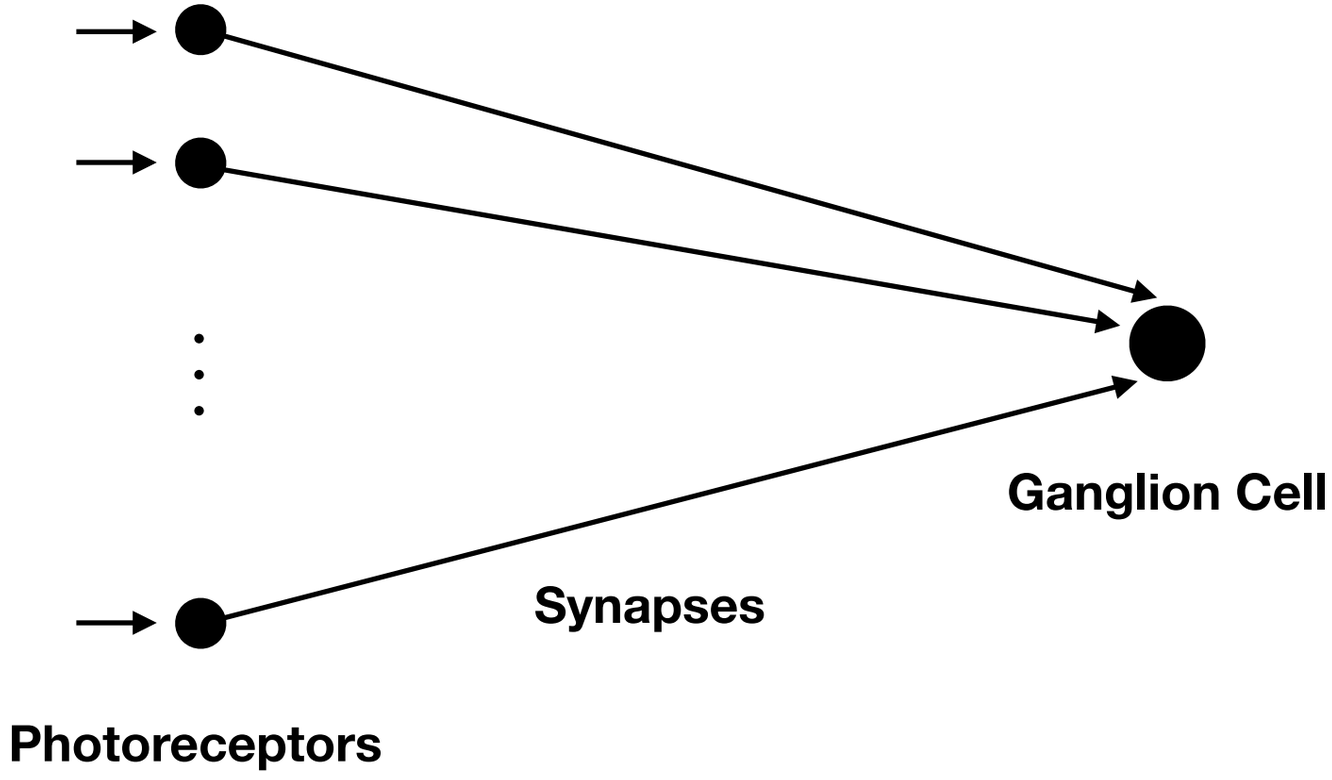
where $y_t = \mathbf{x}_t^\top \mathbf{w}_{t-1}$ is the correlation.

# Oja's Rule & Retina

The update rule is $\mathbf{w}_t = \mathbf{w}_{t-1} + \eta_t y_t \cdot (\mathbf{x}_t - y_t \mathbf{w}_{t-1})$
where $y_t = \mathbf{x}_t^\top \mathbf{w}_{t-1}$ is the correlation.

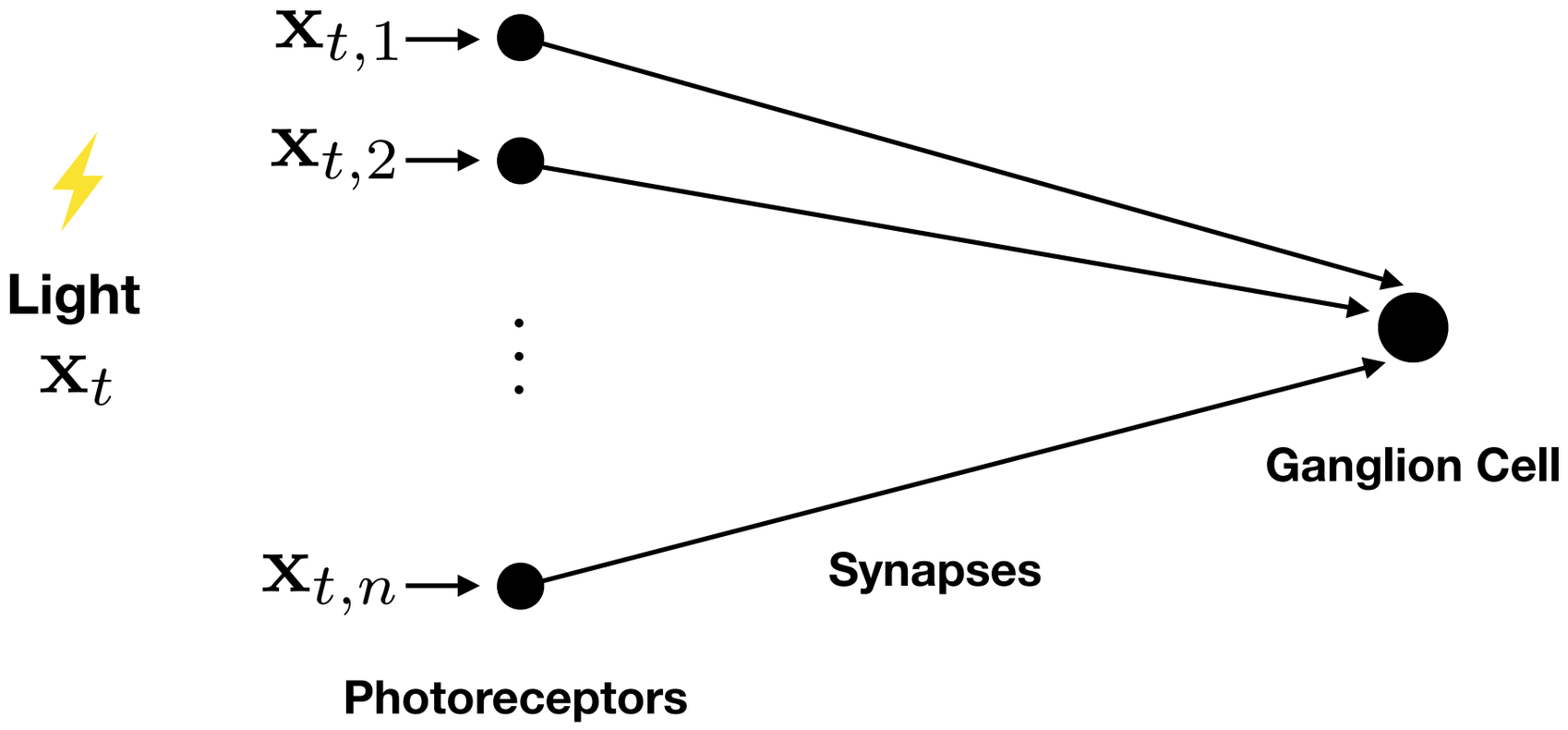
# Oja's Rule & Retina

The update rule is $\mathbf{w}_t = \mathbf{w}_{t-1} + \eta_t y_t \cdot (\mathbf{x}_t - y_t \mathbf{w}_{t-1})$
where $y_t = \mathbf{x}_t^\top \mathbf{w}_{t-1}$ is the correlation.



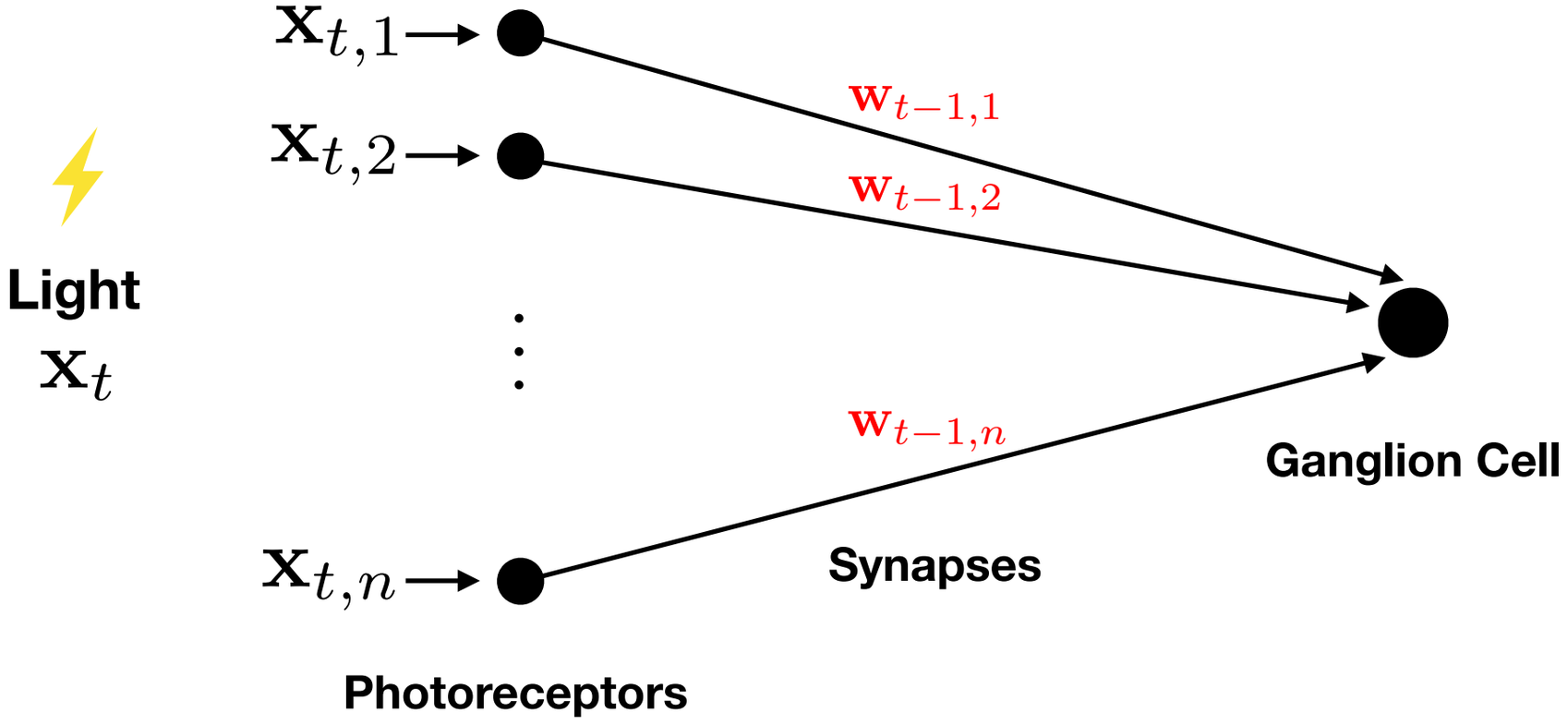
# Oja's Rule & Retina

The update rule is $\mathbf{w}_t = \mathbf{w}_{t-1} + \eta_t y_t \cdot (\mathbf{x}_t - y_t \mathbf{w}_{t-1})$
where $y_t = \mathbf{x}_t^\top \mathbf{w}_{t-1}$ is the correlation.



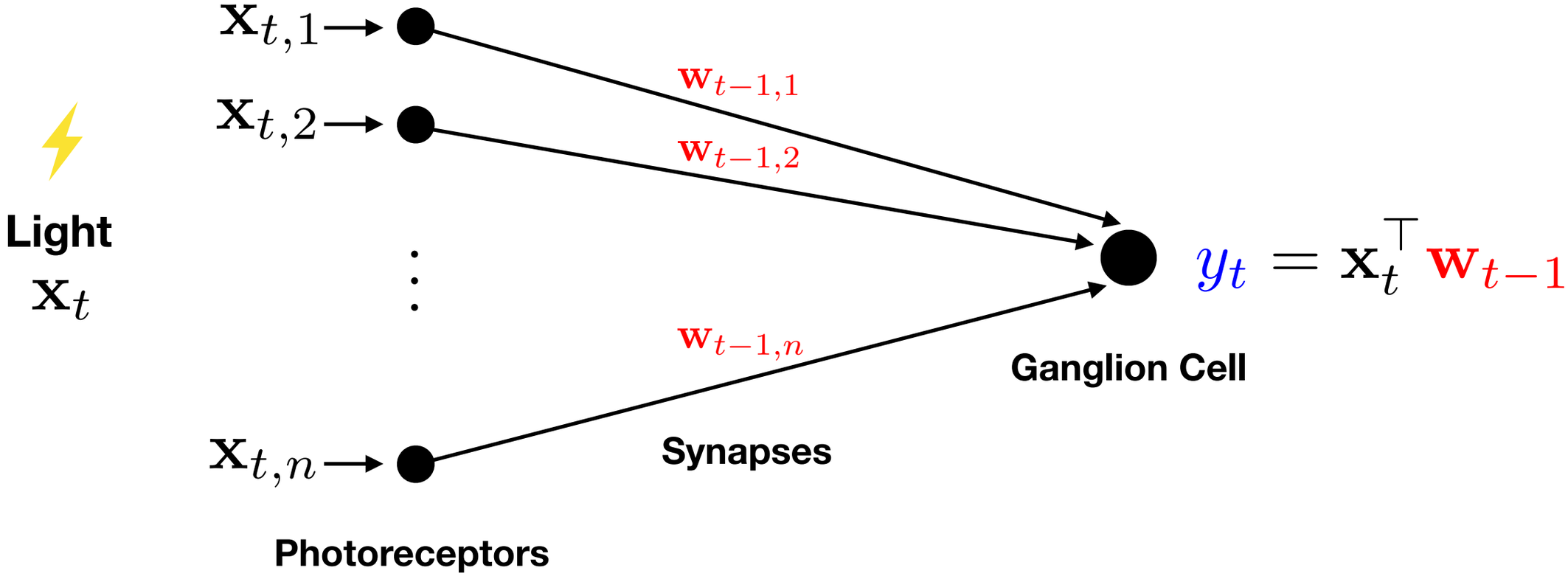
# Oja's Rule & Retina

The update rule is $\mathbf{w}_t = \mathbf{w}_{t-1} + \eta_t y_t \cdot (\mathbf{x}_t - y_t \mathbf{w}_{t-1})$
where $y_t = \mathbf{x}_t^\top \mathbf{w}_{t-1}$ is the correlation.



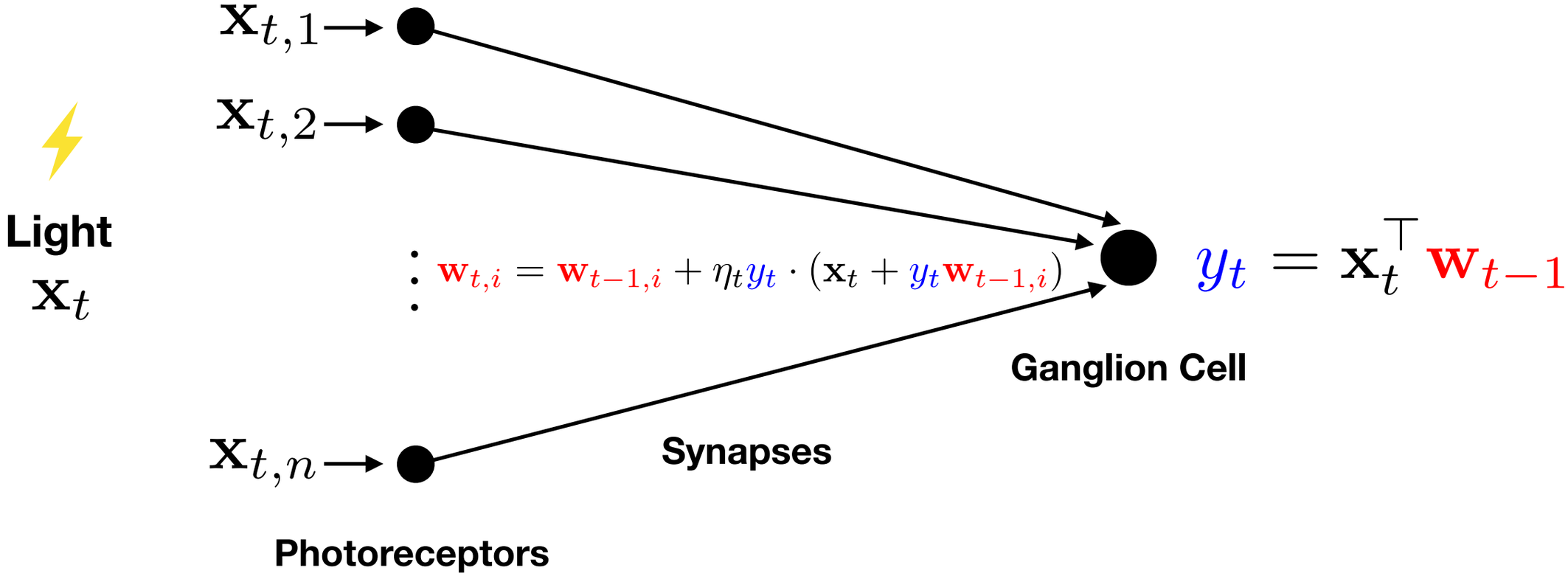
# Oja's Rule & Retina

The update rule is $\mathbf{w}_t = \mathbf{w}_{t-1} + \eta_t y_t \cdot (\mathbf{x}_t - y_t \mathbf{w}_{t-1})$
where $y_t = \mathbf{x}_t^\top \mathbf{w}_{t-1}$ is the correlation.



# Oja's Rule & Retina

The update rule is $\mathbf{w}_t = \mathbf{w}_{t-1} + \eta_t y_t \cdot (\mathbf{x}_t - y_t \mathbf{w}_{t-1})$
where $y_t = \mathbf{x}_t^\top \mathbf{w}_{t-1}$ is the correlation.



# Prior Work

# Prior Work

Only **convergence in the limit** result had been known 😳

# Prior Work

Only **convergence in the limit** result had been known 😳

The higher order term makes the analysis difficult 😢

# Prior Work

Only **convergence in the limit** result had been known 😳

The higher order term makes the analysis difficult 😢

**Theorem ([Duflo 2013], informal).**

Let $\mathbf{w}_0$ be a random unit vector and $\{\eta_t\}$ be the learning rate. If (i) $\sum_t \eta_t = \infty$ and (ii) $\sum_t {\eta_t}^2 < \infty$, then

$$\lim_{t \to \infty} \langle \mathbf{w}_t, \mathbf{v}_1 \rangle^2 = 1$$

almost surely where $\mathbf{v}_1$ is the normalized top eigenvector of the covariance matrix.

# Main Theorem

# Main Theorem

We prove the **first convergence rate** analysis

# Main Theorem

We prove the **first convergence rate** analysis and match the information-theoretic lower bound 🙂

# Main Theorem

We prove the **first convergence rate** analysis and match the information-theoretic lower bound 🙂

**Theorem (Oja's rule efficiently solves streaming PCA).**

Let $\mathbf{w}_0$ be a random unit vector and $\epsilon, \delta > 0$. There exist learning rate $\{\eta_t\}$ with $\eta_t = \Theta(1/\log t)$ and

$$T = \tilde{\Theta}\left(\frac{\lambda_1}{\min\{\epsilon, \delta^2\} \cdot (\lambda_1 - \lambda_2)^2}\right)$$

where $\lambda_1, \lambda_2$ are the top two eigenvalues, such that

$$\Pr\left[\exists t \geq T,\ \frac{\langle \mathbf{w}_t, \mathbf{v}_1 \rangle^2}{\|\mathbf{w}_t\|_2^2} < 1 - \epsilon\right] < \delta\,.$$

# Main Theorem

We prove the **first convergence rate** analysis and match the information-theoretic lower bound 🙂

**Theorem (Oja's rule efficiently solves streaming PCA).**

Let $\mathbf{w}_0$ be a random unit vector and $\epsilon, \delta > 0$. There exist learning rate $\{\eta_t\}$ with $\eta_t = \Theta(1/\log t)$ and

$$T = \tilde{\Theta}\left(\frac{\lambda_1}{\min\{\epsilon, \delta^2\} \cdot (\lambda_1 - \lambda_2)^2}\right)$$

where $\lambda_1, \lambda_2$ are the top two eigenvalues, and

$$\Pr\left[\exists t \geq T, \ \frac{\langle \mathbf{w}_t, \mathbf{v}_1 \rangle^2}{\|\mathbf{w}_t\|_2^2} < 1\right.$$

**This improves the best known result on streaming PCA!**

# Main Theorem

We prove the **first convergence rate** analysis and match the information-theoretic lower bound 🙂

**Theorem (Oja's rule efficiently solves streaming PCA).**

Let $\mathbf{w}_0$ be a random unit vector and $\epsilon, \delta > 0$. There exist learning rate $\{\eta_t\}$ with $\eta_t = \Theta(1/\log t)$ and

$$T = \tilde{\Theta}\left(\frac{\lambda_1}{\min\{\epsilon, \delta^2\} \cdot (\lambda_1 - \lambda_2)^2}\right)$$

where $\lambda_1, \lambda_2$ are the top two eigenvalues, such that

$$\Pr\left[\exists t \geq T,\ \frac{\langle \mathbf{w}_t, \mathbf{v}_1 \rangle^2}{\|\mathbf{w}_t\|_2^2} < 1 - \epsilon\right] < \delta\,.$$

# Main Theorem

We prove the **first convergence rate** analysis and match the information-theoretic lower bound 🙂

**Theorem (Oja's rule efficiently solves streaming PCA).**

Let $\mathbf{w}_0$ be a random unit vector and $\epsilon, \delta > 0$. There exist learning rate $\{\eta_t\}$ with $\eta_t = \Theta(1/\log t)$ and

$$T = \tilde{\Theta}\left(\frac{\lambda_1}{\min\{\epsilon, \delta^2\} \cdot (\lambda_1 - \lambda_2)^2}\right)$$

where $\lambda_1, \lambda_2$ are the top two eigenvalues, such that

$$\Pr\left[\exists t \geq T, \ \frac{\langle \mathbf{w}_t, \mathbf{v}_1 \rangle^2}{\|\mathbf{w}_t\|_2^2} < 1 - \epsilon\right] < \delta\,.$$

# Main Theorem

We prove the **first convergence rate** analysis and match the information-theoretic lower bound 🙂

**Theorem (Oja's rule efficiently solves streaming PCA).**

Let $\mathbf{w}_0$ be a random unit vector and $\epsilon, \delta > 0$. There exist learning rate $\{\eta_t\}$ with $\eta_t = \Theta(1/\log t)$ and

$$T = \tilde{\Theta}\left(\frac{\lambda_1}{\min\{\epsilon, \delta^2\} \cdot (\lambda_1 - \lambda_2)^2}\right)$$

where $\lambda_1, \lambda_2$ are the top two eigenvalues, such that

$$\Pr\left[\exists t \geq T,\ \frac{\langle \mathbf{w}_t, \mathbf{v}_1 \rangle^2}{\|\mathbf{w}_t\|_2^2} < 1 - \epsilon\right] < \delta\,.$$

# Biological Perspectives

# Biological Perspectives

Retina adapts to illumination, contrast, spatial frequency etc. in few seconds despite high-dimensional inputs [Shapley 1994].

# Biological Perspectives

Retina adapts to illumination, contrast, spatial frequency etc. in few seconds despite high-dimensional inputs [Shapley 1994].

## Efficient Coding Principle

## Continual Learning

# Biological Perspectives

Retina adapts to illumination, contrast, spatial frequency etc. in few seconds despite high-dimensional inputs [Shapley 1994].

## Efficient Coding Principle

[Barlow 1961] the main goal of a sensory system is to maximize the information in the neural encoding.

## Continual Learning

# Biological Perspectives

Retina adapts to illumination, contrast, spatial frequency etc. in few seconds despite high-dimensional inputs [Shapley 1994].

## Efficient Coding Principle

[Barlow 1961] the main goal of a sensory system is to maximize the information in the neural encoding.

~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~

**Q: What's the underlying dynamic?**

~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~

## Continual Learning

# Biological Perspectives

Retina adapts to illumination, contrast, spatial frequency etc. in few seconds despite high-dimensional inputs [Shapley 1994].

## Efficient Coding Principle

[Barlow 1961] the main goal of a sensory system is to maximize the information in the neural encoding.

~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~

**Q: What's the underlying dynamic?**

~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~

We provide the first provable explanation for fast adaptation.

## Continual Learning

# Biological Perspectives

Retina adapts to illumination, contrast, spatial frequency etc. in few seconds despite high-dimensional inputs [Shapley 1994].

## Efficient Coding Principle

[Barlow 1961] the main goal of a sensory system is to maximize the information in the neural encoding.

~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~

**Q: What's the underlying dynamic?**

~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~

We provide the first provable explanation for fast adaptation.

## Continual Learning

Adaptation happens throughout lifetime.

# Biological Perspectives

Retina adapts to illumination, contrast, spatial frequency etc. in few seconds despite high-dimensional inputs [Shapley 1994].

## Efficient Coding Principle

[Barlow 1961] the main goal of a sensory system is to maximize the information in the neural encoding.

~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~

**Q: What's the underlying dynamic?**

~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~

We provide the first provable explanation for fast adaptation.

## Continual Learning

Adaptation happens throughout lifetime.

~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~

**Q: How is it possible to continually have efficient adaptation?**

~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~

# Biological Perspectives

Retina adapts to illumination, contrast, spatial frequency etc. in few seconds despite high-dimensional inputs [Shapley 1994].

## Efficient Coding Principle

[Barlow 1961] the main goal of a sensory system is to maximize the information in the neural encoding.

~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~

**Q: What's the underlying dynamic?**

~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~

We provide the first provable explanation for fast adaptation.

## Continual Learning

Adaptation happens throughout lifetime.

~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~

**Q: How is it possible to continually have efficient adaptation?**

~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~

- The learning rate remains large and the guarantee is for-all-time.
- Specifically, we have $\sum_t {\eta_t}^2 = \infty$.

# Few Words on the Proofs

# Few Words on the Proofs

The higher order term makes the previous analysis (for other streaming PCA algorithm) fail in analyzing Oja's rule 😢

$$\mathbf{w}_t = \mathbf{w}_{t-1} + \eta_t y_t \mathbf{x}_t - \eta_t {y_t}^2 \mathbf{w}_{t-1}$$

# Few Words on the Proofs

The higher order term makes the previous analysis (for other streaming PCA algorithm) fail in analyzing Oja's rule 😢

$$\mathbf{w}_t = \mathbf{w}_{t-1} + \eta_t y_t \mathbf{x}_t - \eta_t {y_t}^2 \mathbf{w}_{t-1}$$

# Few Words on the Proofs

The higher order term makes the previous analysis (for other streaming PCA algorithm) fail in analyzing Oja's rule 😢

$$\mathbf{w}_t = \mathbf{w}_{t-1} + \eta_t y_t \mathbf{x}_t - \eta_t {y_t}^2 \mathbf{w}_{t-1}$$

**Continuous Analysis**

**Stopping Time Framework**

# Few Words on the Proofs

The higher order term makes the previous analysis (for other streaming PCA algorithm) fail in analyzing Oja's rule 😢

$$\mathbf{w}_t = \mathbf{w}_{t-1} + \eta_t y_t \mathbf{x}_t - \eta_t y_t^2 \mathbf{w}_{t-1}$$

**Continuous Analysis**

*Suggest the right way to analyze!*

**Stopping Time Framework**

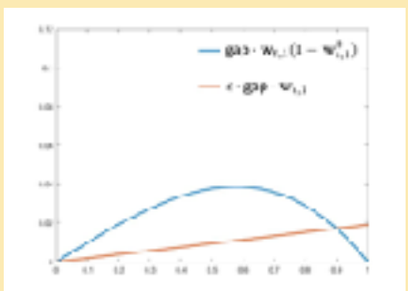
# Few Words on the Proofs

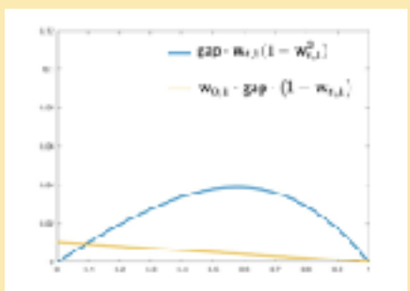
The higher order term makes the previous analysis (for other streaming PCA algorithm) fail in analyzing Oja's rule 😢

$$\mathbf{w}_t = \mathbf{w}_{t-1} + \eta_t y_t \mathbf{x}_t - \eta_t y_t^2 \mathbf{w}_{t-1}$$

## Continuous Analysis

*Suggest the right way to analyze!*





## Stopping Time Framework

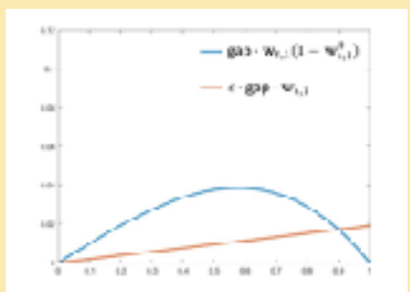
# Few Words on the Proofs

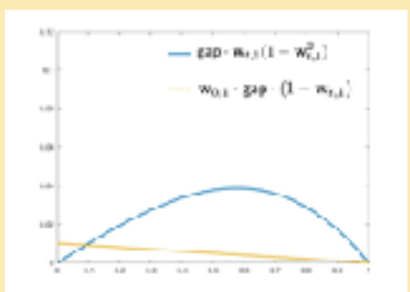
The higher order term makes the previous analysis (for other streaming PCA algorithm) fail in analyzing Oja's rule 😢

$$\mathbf{w}_t = \mathbf{w}_{t-1} + \eta_t y_t \mathbf{x}_t - \eta_t y_t^2 \mathbf{w}_{t-1}$$

## Continuous Analysis

*Suggest the right way to analyze!*





## Stopping Time Framework

*Flexible and nearly optimal!*

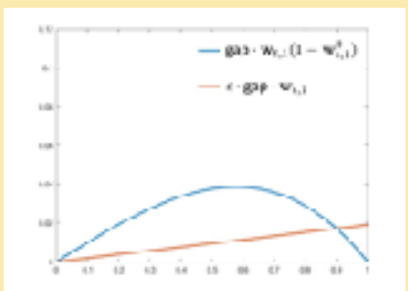
# Few Words on the Proofs

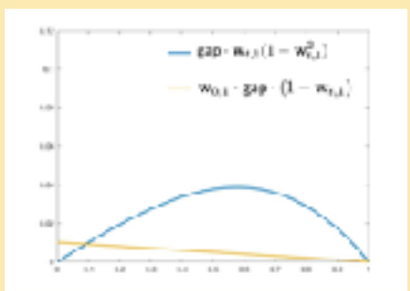
The higher order term makes the previous analysis (for other streaming PCA algorithm) fail in analyzing Oja's rule 😢

$$\mathbf{w}_t = \mathbf{w}_{t-1} + \eta_t y_t \mathbf{x}_t - \eta_t y_t^2 \mathbf{w}_{t-1}$$

## Continuous Analysis

*Suggest the right way to analyze!*





## Stopping Time Framework

*Flexible and nearly optimal!*

**Step 1: Linearization & Moment Analysis**

**Step 2: Improvement Analysis**

**Step 3: Interval Analysis**

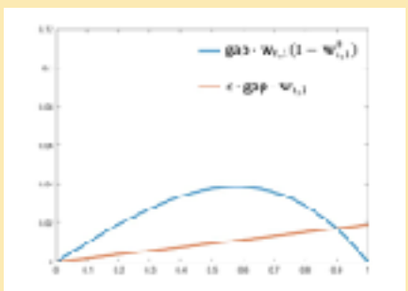
# Few Words on the Proofs

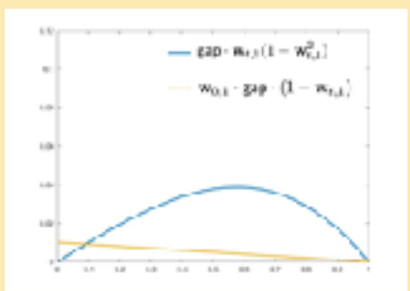
The higher order term makes the previous analysis (for other streaming PCA algorithm) fail in analyzing Oja's rule 😢

$$\mathbf{w}_t = \mathbf{w}_{t-1} + \eta_t y_t \mathbf{x}_t - \eta_t y_t^2 \mathbf{w}_{t-1}$$

## Continuous Analysis

*Suggest the right way to analyze!*





## Stopping Time Framework

*Flexible and nearly optimal!*

**Step 1: Linearization & Moment Analysis**

**Step 2: Improvement Analysis**

**Step 3: Interval Analysis**

See the the full version [arXiv 1911.02363v2] or a follow-up paper [arXiv 2006.06171] focusing on the proof framework/techniques!

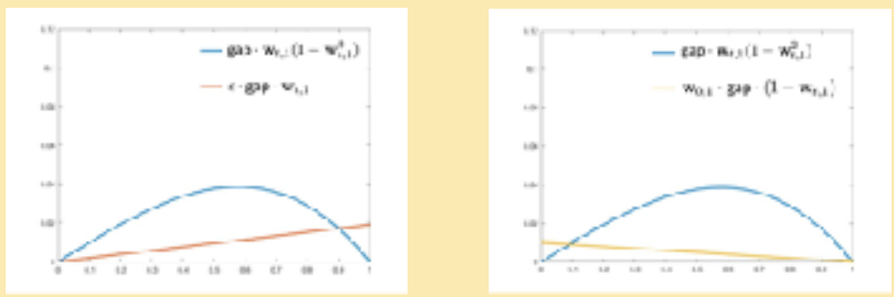
# Few Words on the Proofs

The higher order term makes the previous analysis (for other streaming PCA algorithm) fail in analyzing Oja's rule 😢

$$\mathbf{w}_t = \mathbf{w}_{t-1} + \eta_t y_t \mathbf{x}_t - \eta_t y_t^2 \mathbf{w}_{t-1}$$

## Continuous Analysis

*Suggest the right way to analyze!*



## Stopping Time Framework

*Flexible and nearly optimal!*

**Step 1: Linearization & I ent An sis**

**Step 2: Improvement A**

**Step 3: Interval**

**The technique improves 3 different ML problems!**

See the the full version [arXiv 1911.02363v2] (
[arXiv 2006.06171] focusing on the pr

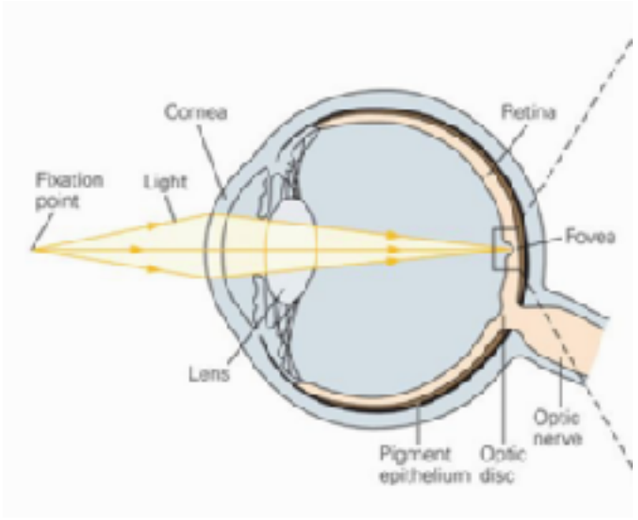
# Future Endeavor

# Future Endeavor

*When CS & Math Meet Neuroscience!?*

# Future Endeavor

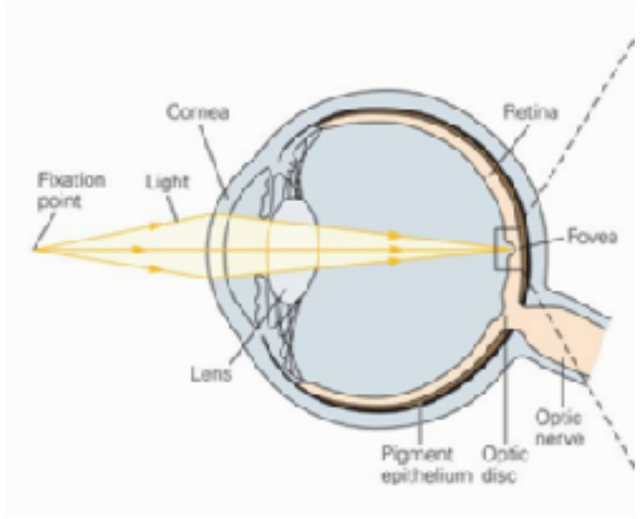
*When CS & Math Meet Neuroscience!?*




**Sensory System**
**Information Theory**

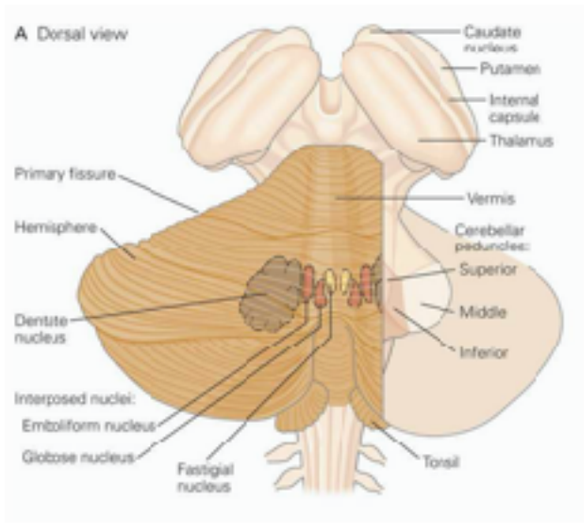
# Future Endeavor

*When CS & Math Meet Neuroscience!?*



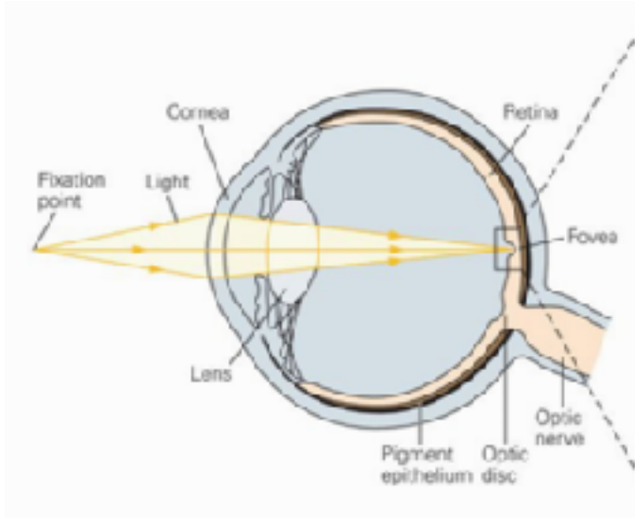

**Sensory System**
**Information Theory**

**Cerebellum**
**Supervised Learning**



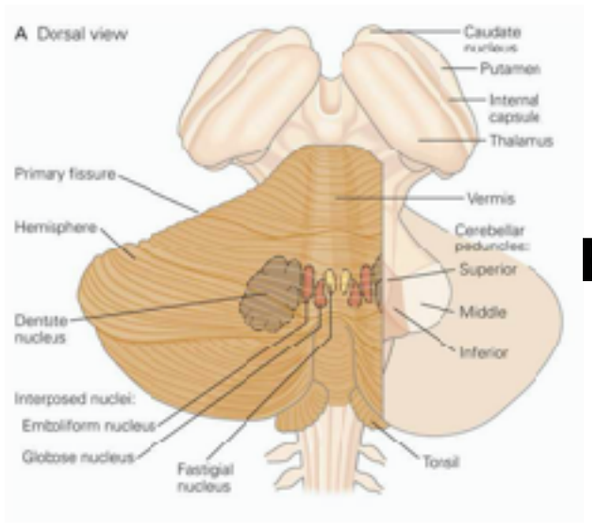
# Future Endeavor

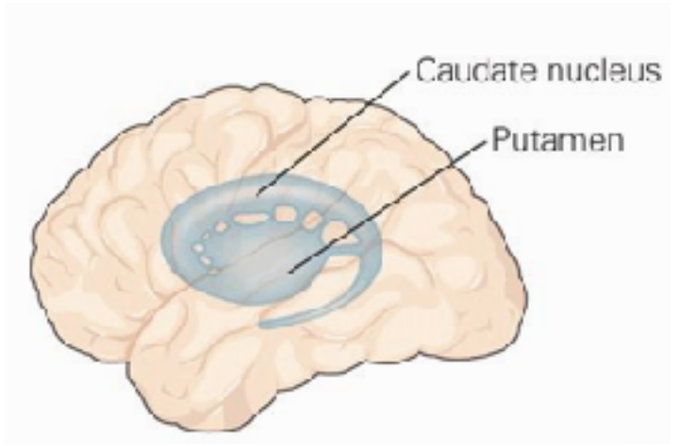
*When CS & Math Meet Neuroscience!?*




**Sensory System**
**Information Theory**

**Cerebellum**
**Supervised Learning**

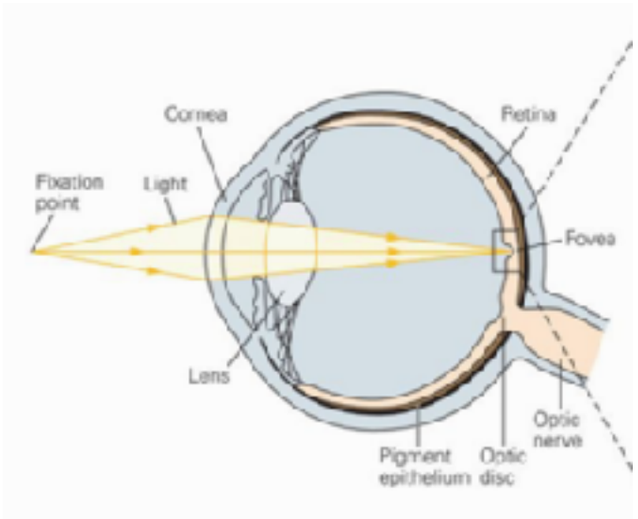



**Basal Ganglia**
**Reinforcement Learning**

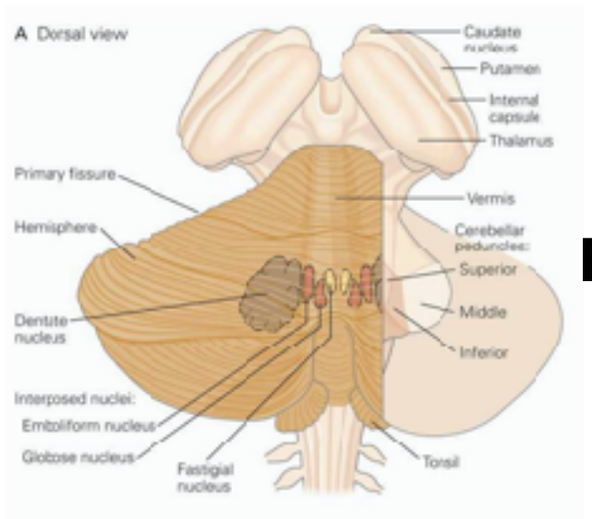
# Future Endeavor

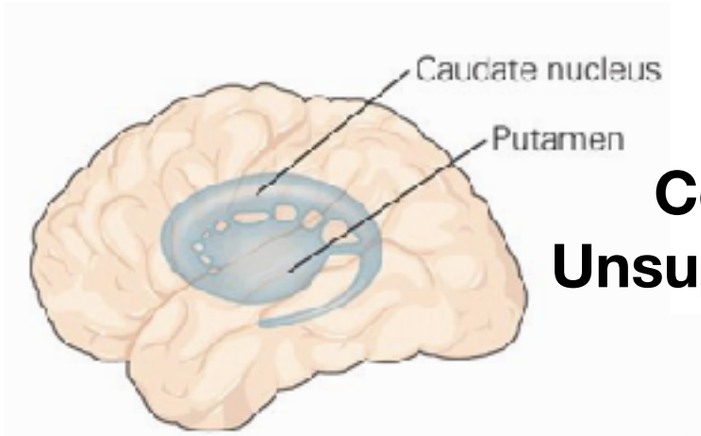
*When CS & Math Meet Neuroscience!?*




**Sensory System**
**Information Theory**

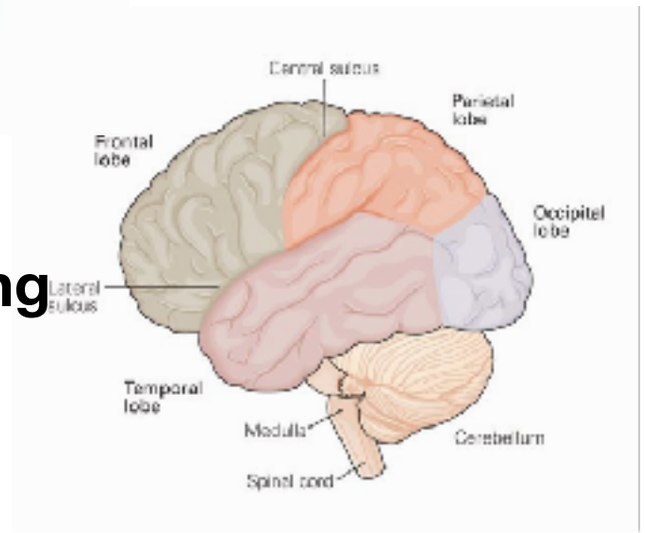
**Cerebellum**
**Supervised Learning**



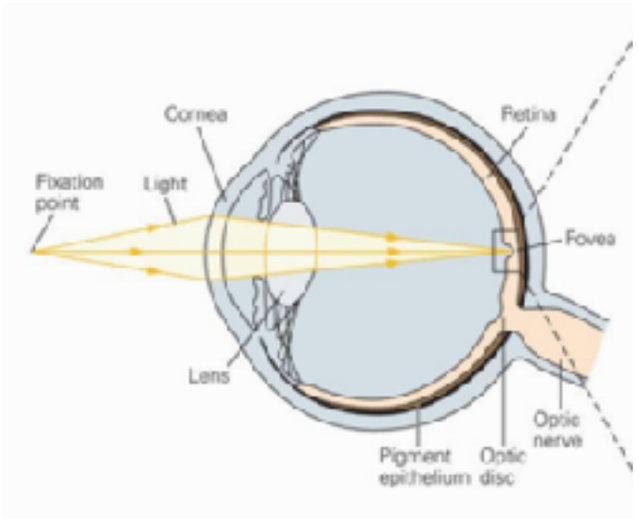

**Basal Ganglia**
**Reinforcement Learning**

**Cerebral Cortex**
**Unsupervised Learning**



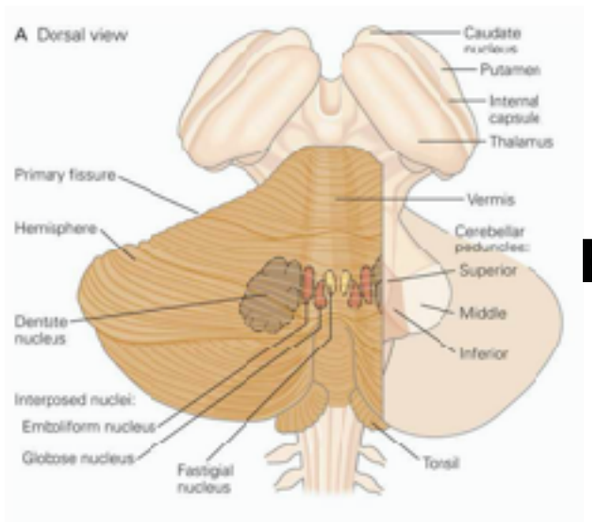
# Future Endeavor

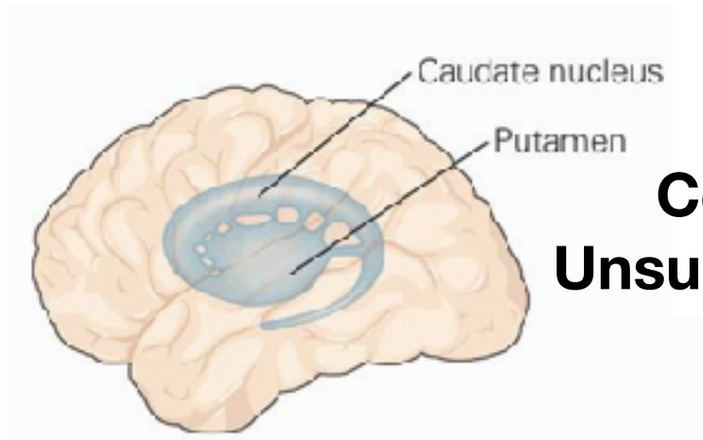
*When CS & Math Meet Neuroscience!?*




**Sensory System**
**Information Theory**

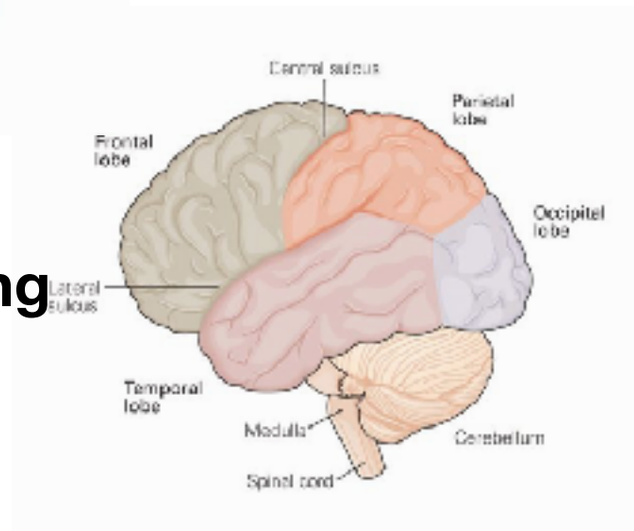
**Cerebellum**
**Supervised Learning**




**Basal Ganglia**
**Reinforcement Learning**

**Cerebral Cortex**
**Unsupervised Learning**




**Theoretical Insights** ⇄ **Biological Phenomenons**

# Thank you!