Advanced Section #5: Ensemble Methods and Mixture of Experts

CS109A Introduction to Data Science

Pavlos Protopapas, Kevin Rader and Chris Tanner

Presentation prepared by Cecilia Garraffo for CS109A 2020

Ensemble methods and Mixture of experts combine models in order to obtain a more accurate and/or more robust model

Ensemble methods → combines models cooperatively
- → homogeneous models
  - Voting and Averaging
  - Bagging and Boosting
  - covered in class, reviewed here today
- → heterogeneous models
  - Blending
  - Stacking

Mixture of experts → combines models by specialization → heterogeneous models
- experts + gating network

covered here today

Outline

- Intuition for ensemble methods (cooperative) and mixture of experts (specialization)
- Simple ensemble methods for classification and regression: voting and averaging - homogeneous learners
- More ensemble methods: bagging, boosting - homogeneous learners- and blending and stacking - heterogeneous learners
- Mixture of experts and hierarchical mixture of experts

Ensemble Methods

Team work

Ensemble methods use a combination of simpler learners (any model trained on data) to improve predictions. Combining models usually results in a more precise model (often among the top rankings of many machine learning competitions, including Netflix, KDD 2009, Kaggle's competitions)

Ensemble Methods

Team work

Ensemble methods use a combination of simpler learners (any model trained on data) to improve predictions. Combining models usually results in a more precise model (often among the top rankings of many machine learning competitions, including Netflix, KDD 2009, Kaggle’s competitions)

Why does it work? Intuition:

The famous jelly bean experiment by Prof. Marcus du Sautoy (Anah Veronica - Medium)

Ensemble Methods

Team work

Ensemble methods use a combination of simpler learners (any model trained on data) to improve predictions. Combining models usually results in a more precise model (often among the top rankings of many machine learning competitions, including Netflix, KDD 2009, Kaggle’s competitions)

Why does it work? Intuition:

The famous jelly bean experiment by Prof. Marcus du Sautoy (Anah Veronica - Medium)

- Jelly beans jar
- Asked 160 people to guess how many
- Answers ranged from 400 up to 50,000
- The average was 4514
- The true number of beans was 4510!

Mixture of Experts

Specialist

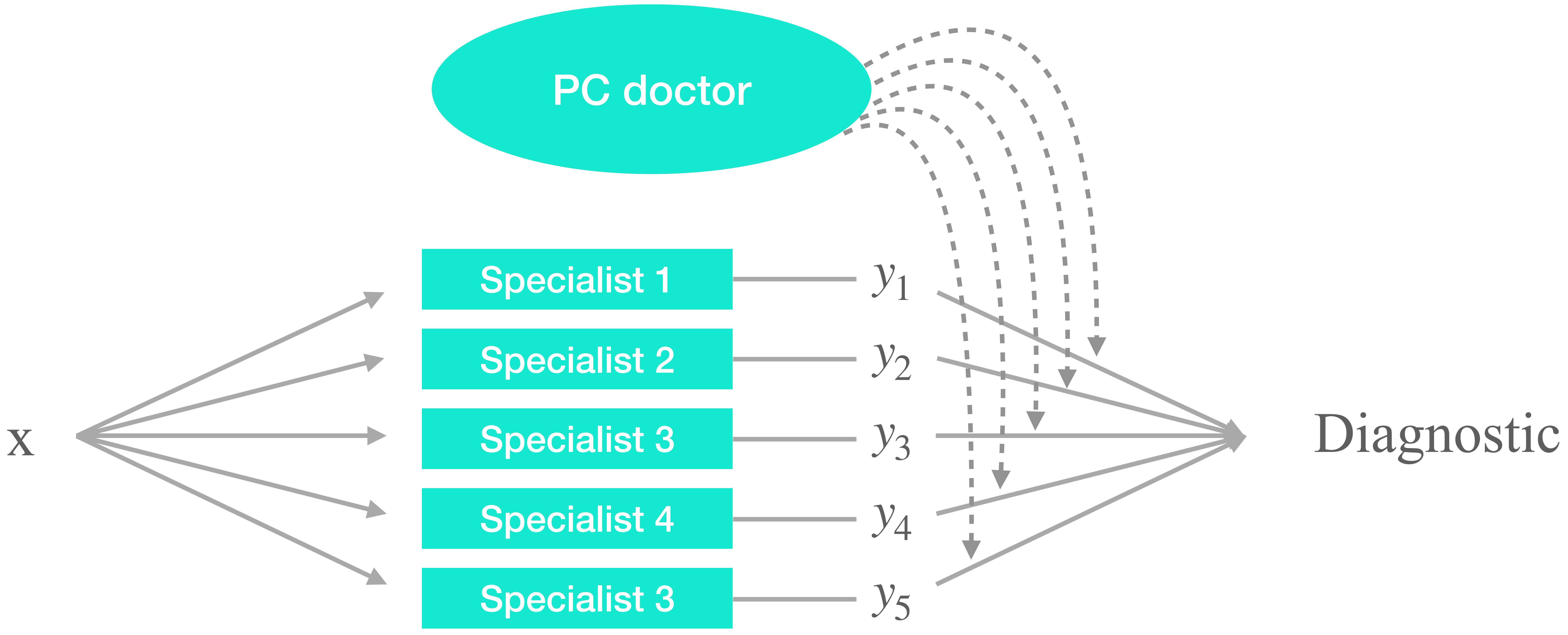
Uses multiple simple learners, each of which specializes on a different part of the data, plus a manager model that will decide which specialist to use for each input data.

Mixture of Experts

Specialist

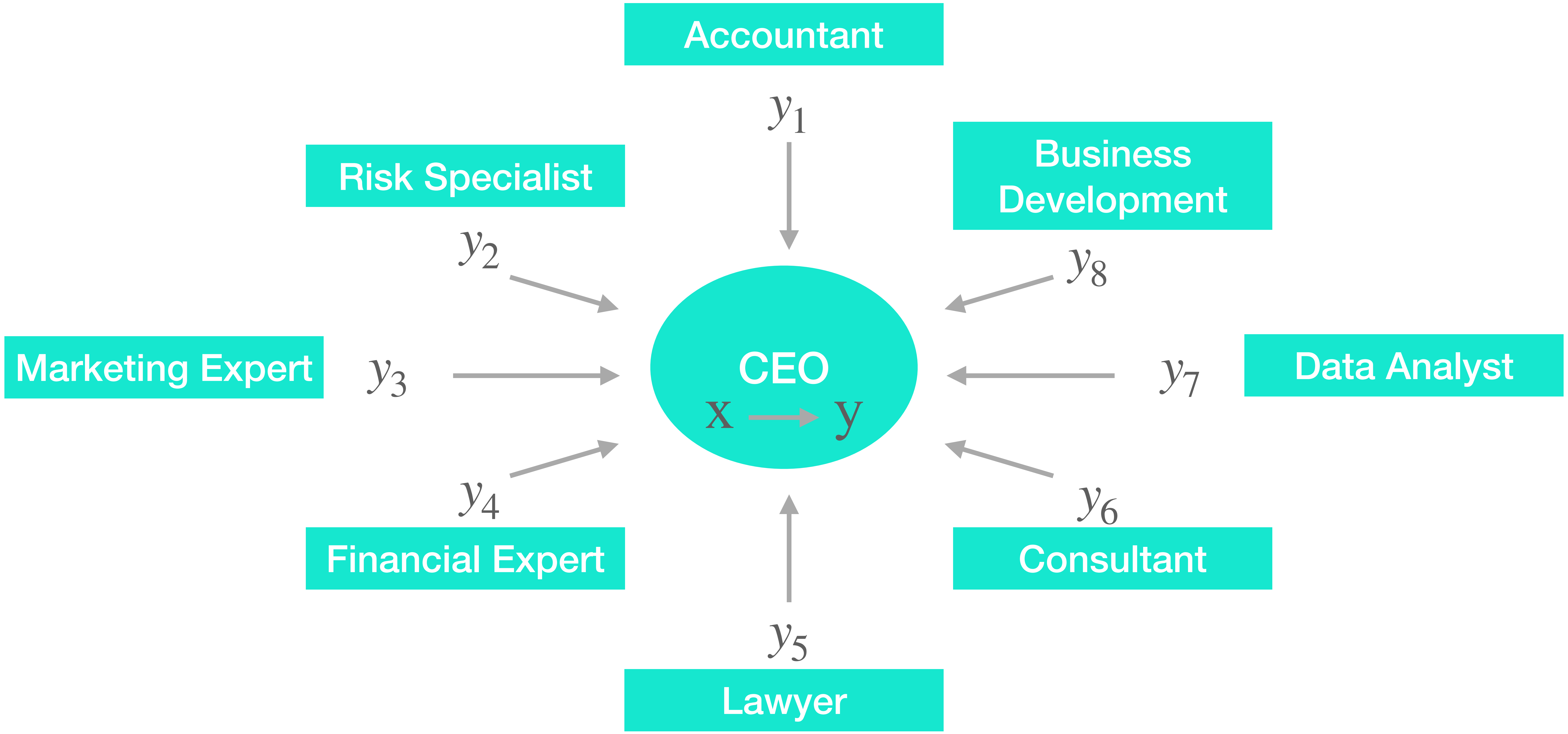
Uses multiple simple learners, each of which specializes on a different part of the data, plus a manager model that will decide which specialist to use for each input data.

Intuition:



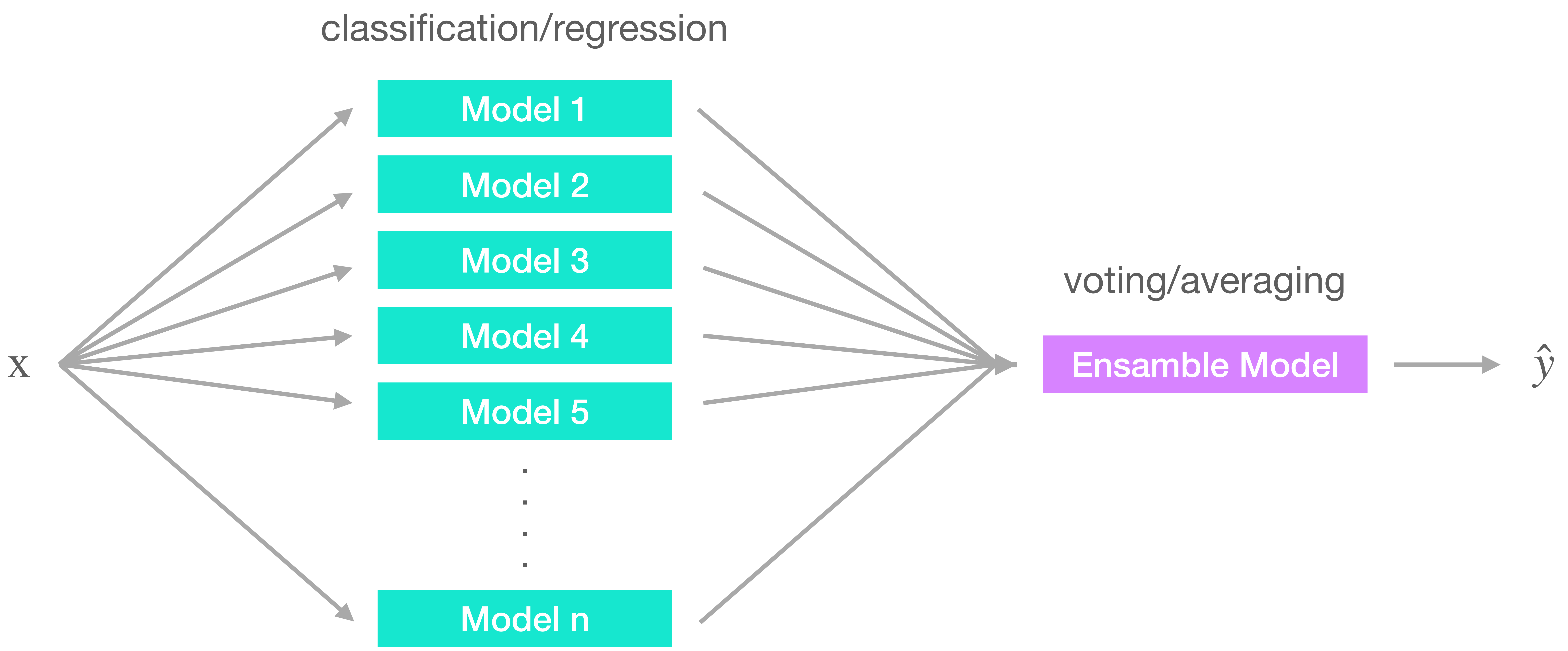
Mixture of Experts

Specialist: Intuition



Ensemble Methods

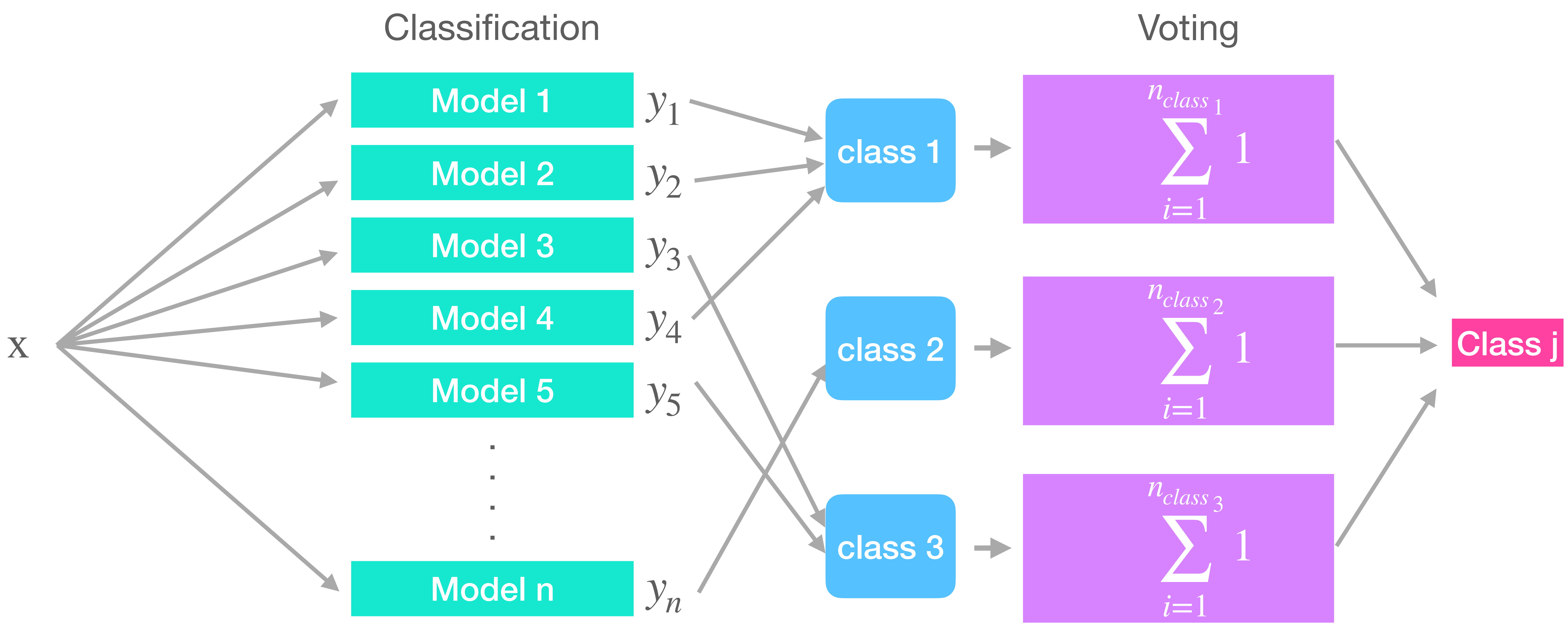
Voting and Averaging



- Models can be trained with different splits of the same dataset and same algorithm (homogeneous) or with the same dataset and different algorithms

Ensemble Methods

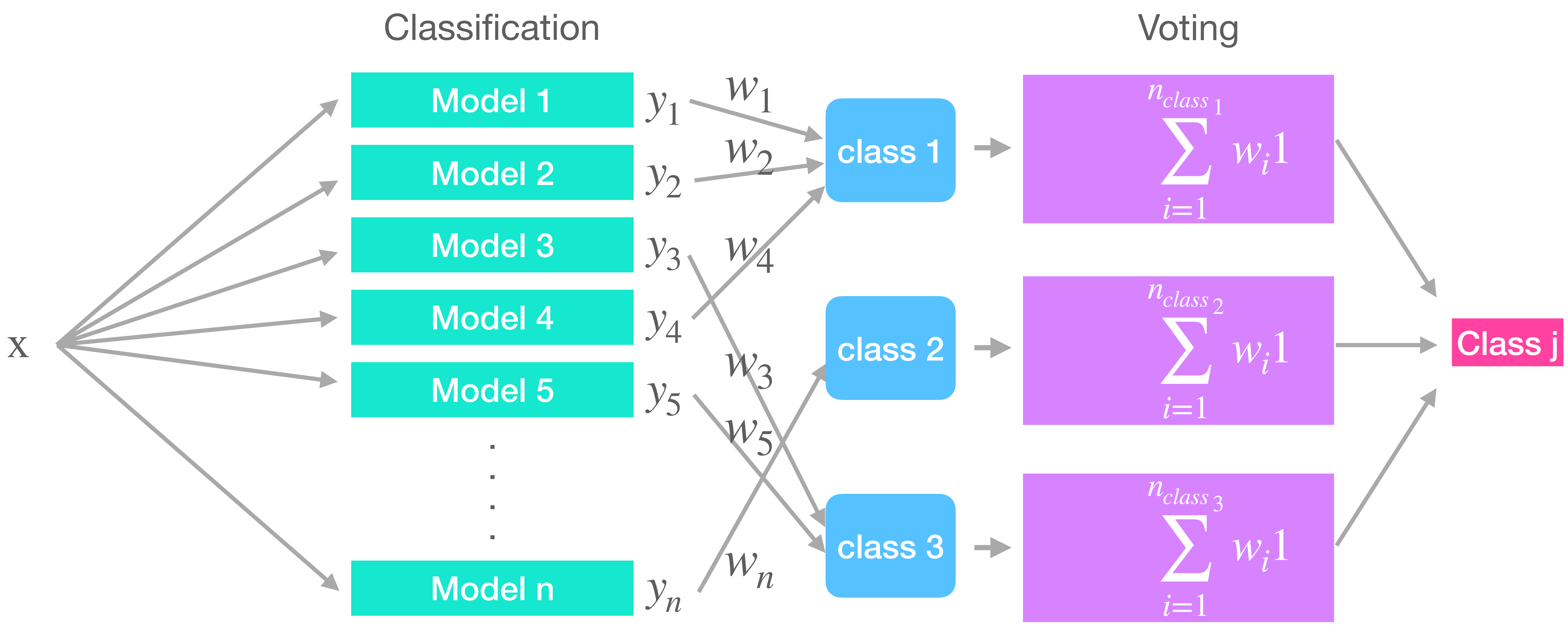
Majority voting



A class gets more than half of the votes, the prediction is called a “stable prediction”. Otherwise, the prediction results less reliable and it is sometimes called “plurality voting”.

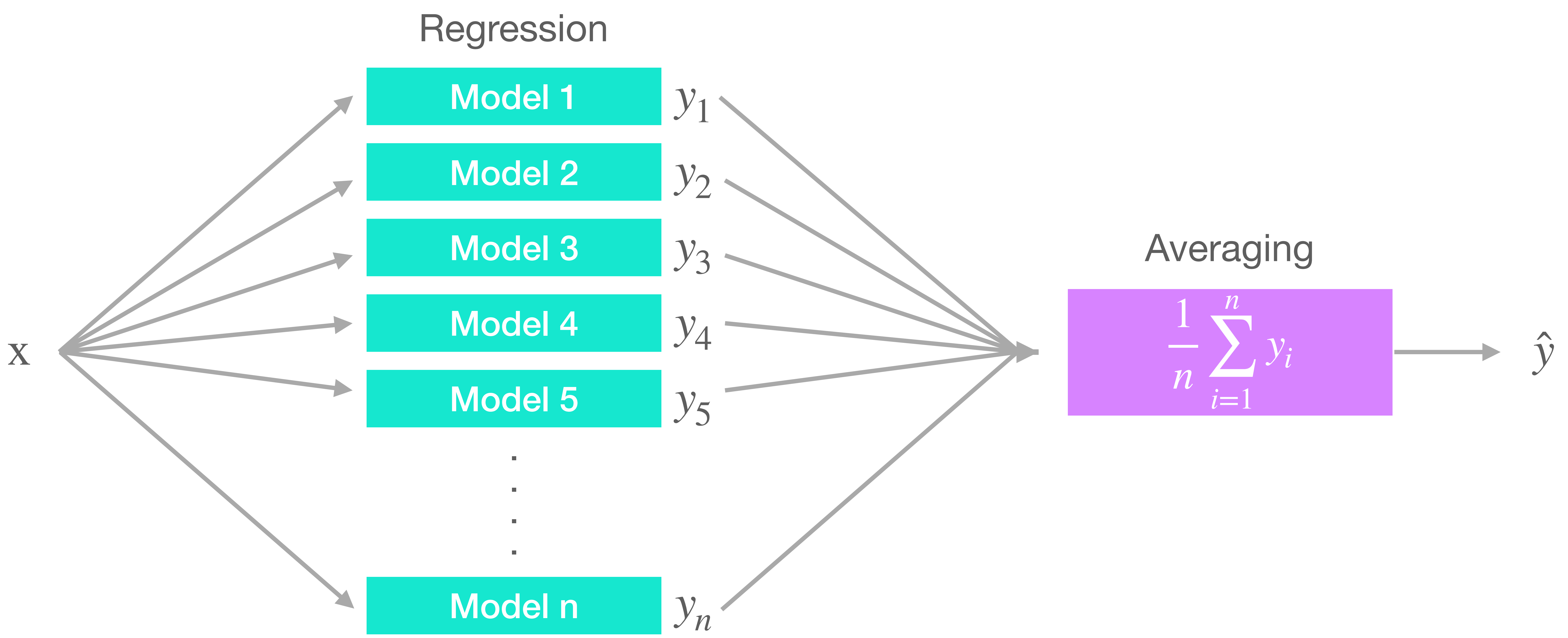
Ensemble Methods

Weighting voting



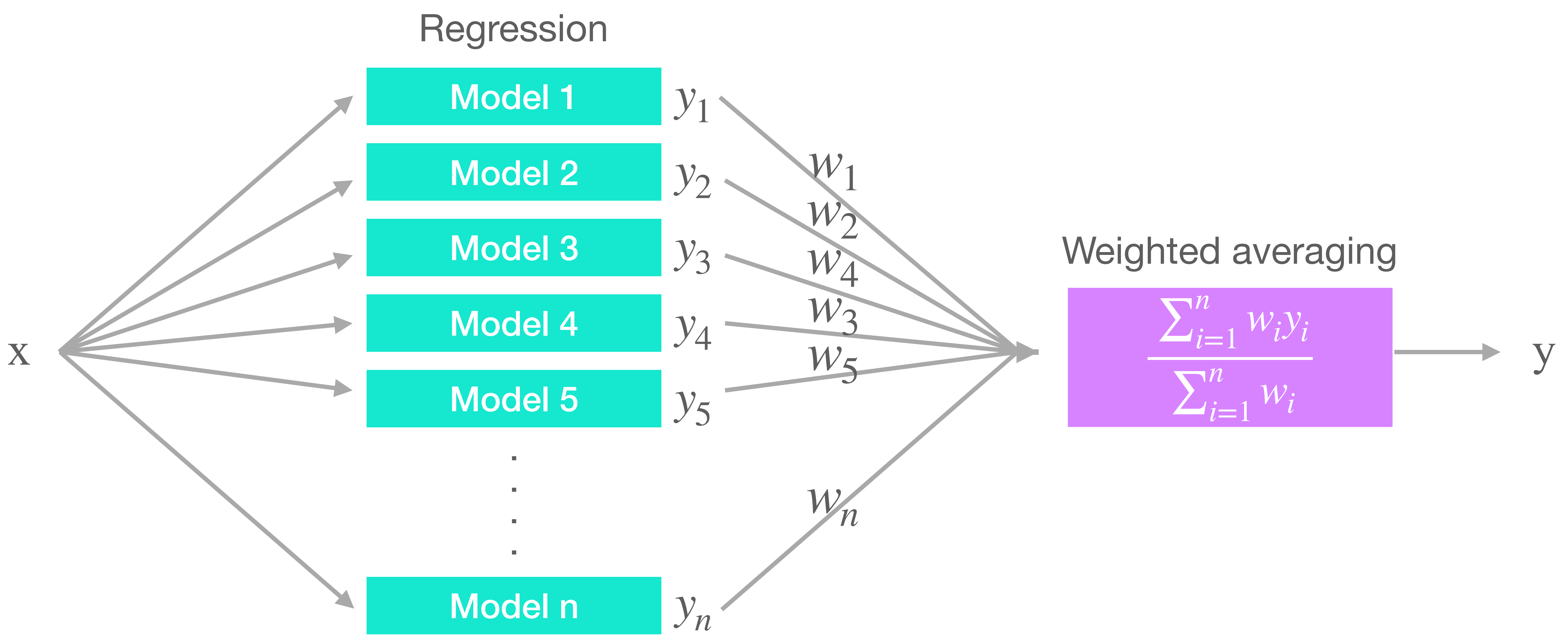
Ensemble Methods

Simple averaging



Ensemble Methods

Weighted averaging



Bagging

Bootstrap aggregating for lower variance: usually homogeneous weak learners, independently learned in parallel and combined using some kind of deterministic averaging process

If n is large -> representativity

If n >> l -> independence

Boosting

Usually uses homogeneous weak learners, learns them sequentially - a base model depends on the previous ones

Figure adapted from Sirakorn - Wikimedia.org

Outline

- Intuition for ensemble of models and mixture of experts ✓
- Simple ensemble of models for classification and regression: voting and averaging
- More ensemble of models: bagging, boosting, blending and stacking
- Mixture of experts and hierarchical mixture of experts

Outline

- Intuition for ensemble of models and mixture of experts ✓
- Simple ensemble of models for classification and regression: voting and averaging ✓
- More ensemble of models: bagging, boosting, blending and stacking
- Mixture of experts and hierarchical mixture of experts

Outline

- Intuition for ensemble of models and mixture of experts ✓
- Simple ensemble of models for classification and regression: voting and averaging ✓
- More ensemble of models: $\overbrace{\text{bagging, boosting}}^{\text{homogeneous}}$ ✓, blending and stacking
- Mixture of experts and hierarchical mixture of experts

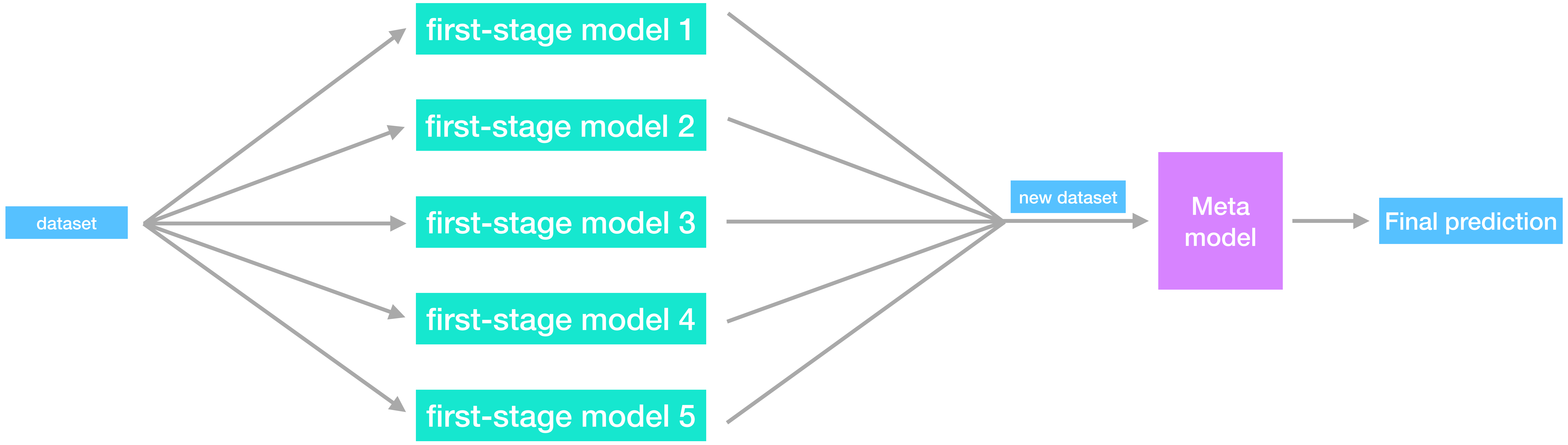
Outline

- Intuition for ensemble of models and mixture of experts ✓
- Simple ensemble of models for classification and regression: voting and averaging ✓
- More ensemble of models: $\underbrace{\text{bagging, boosting}}_{\text{homogeneous}}$ ✓, blending and stacking

  less variance   less bias
- Mixture of experts and hierarchical mixture of experts

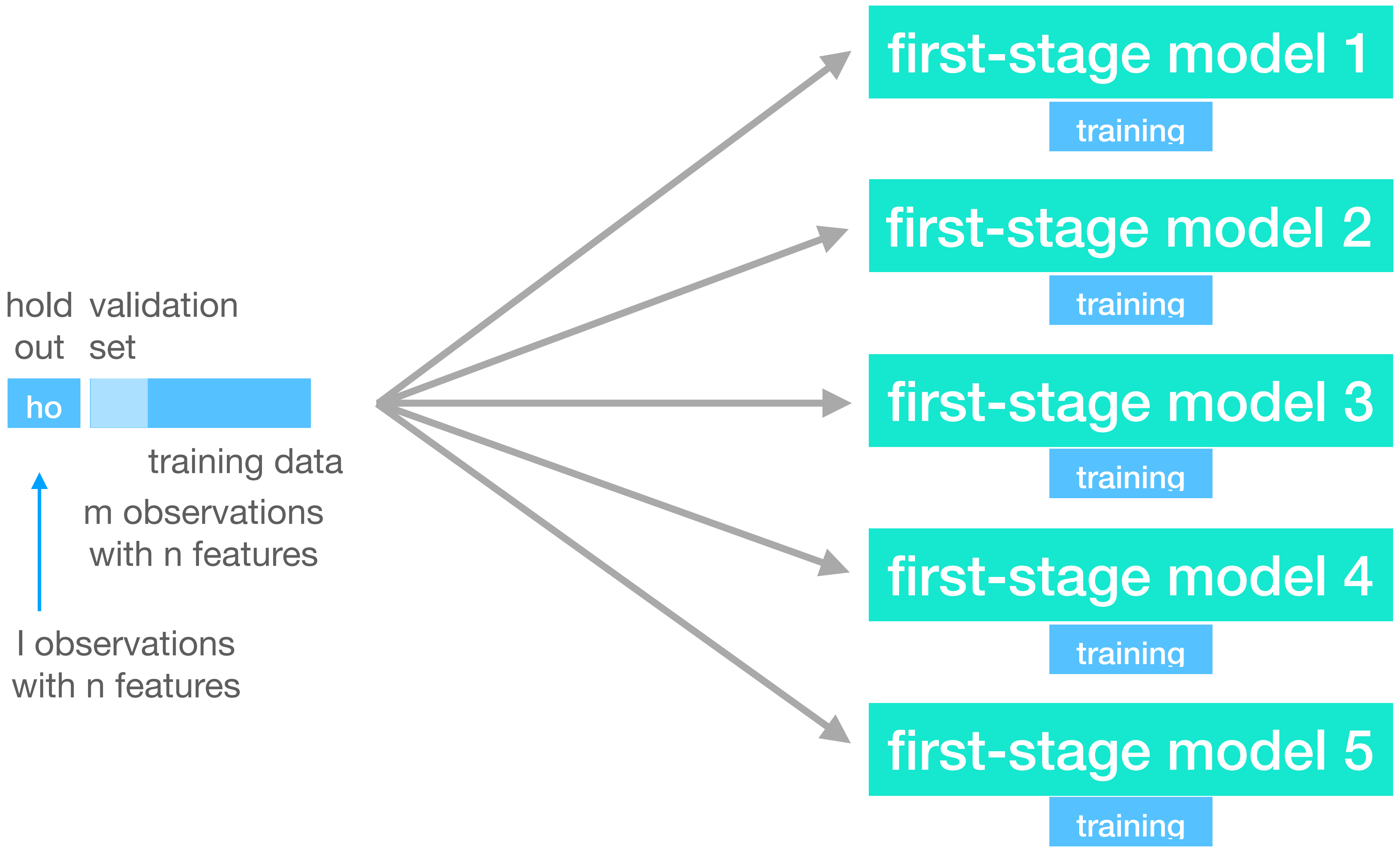
Blending

Independent, parallel, heterogeneous weak learners, by training a meta-model to output a prediction - less bias



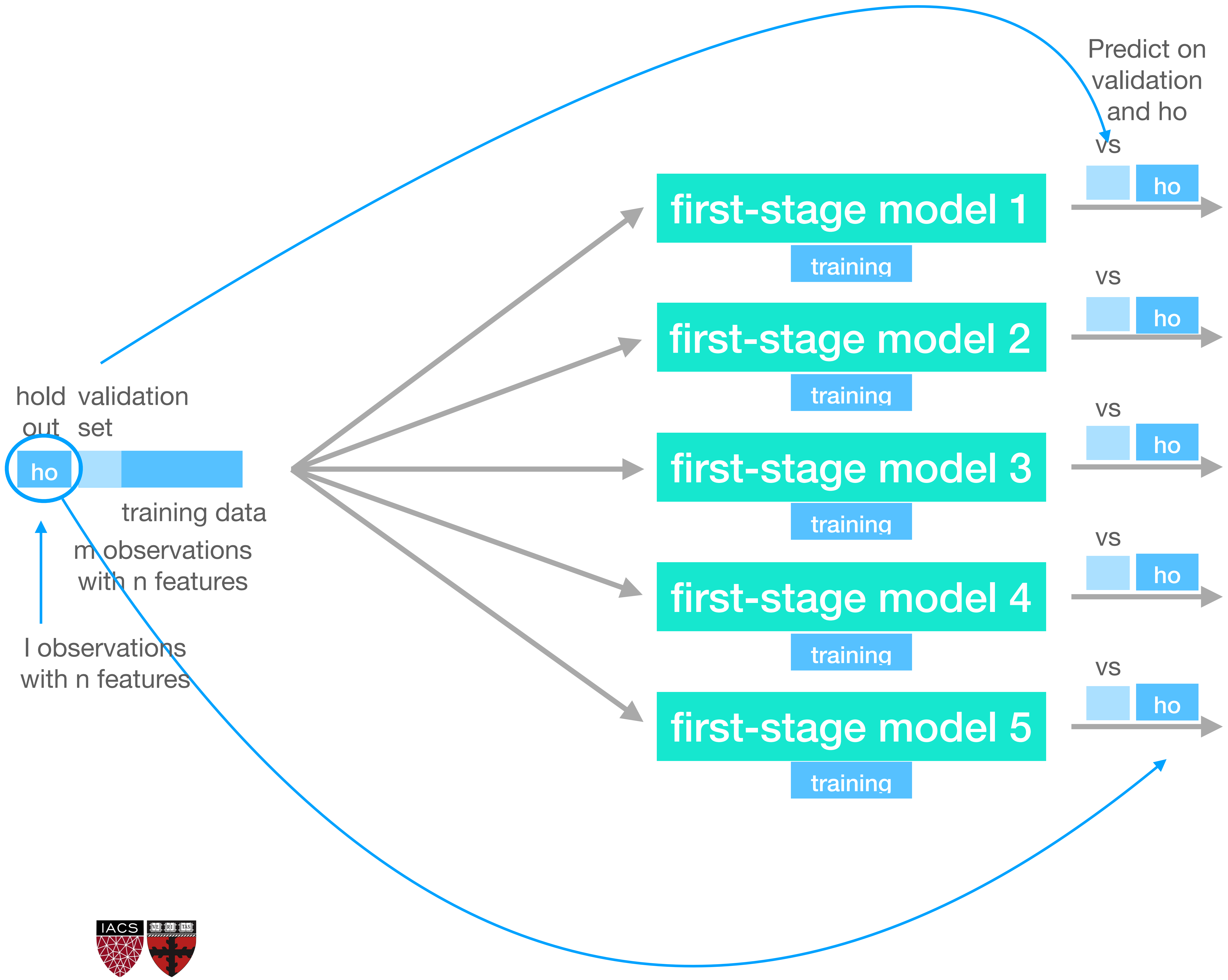
Blending

Independent, parallel, heterogeneous weak learners, by training a meta-model to output a prediction - less bias



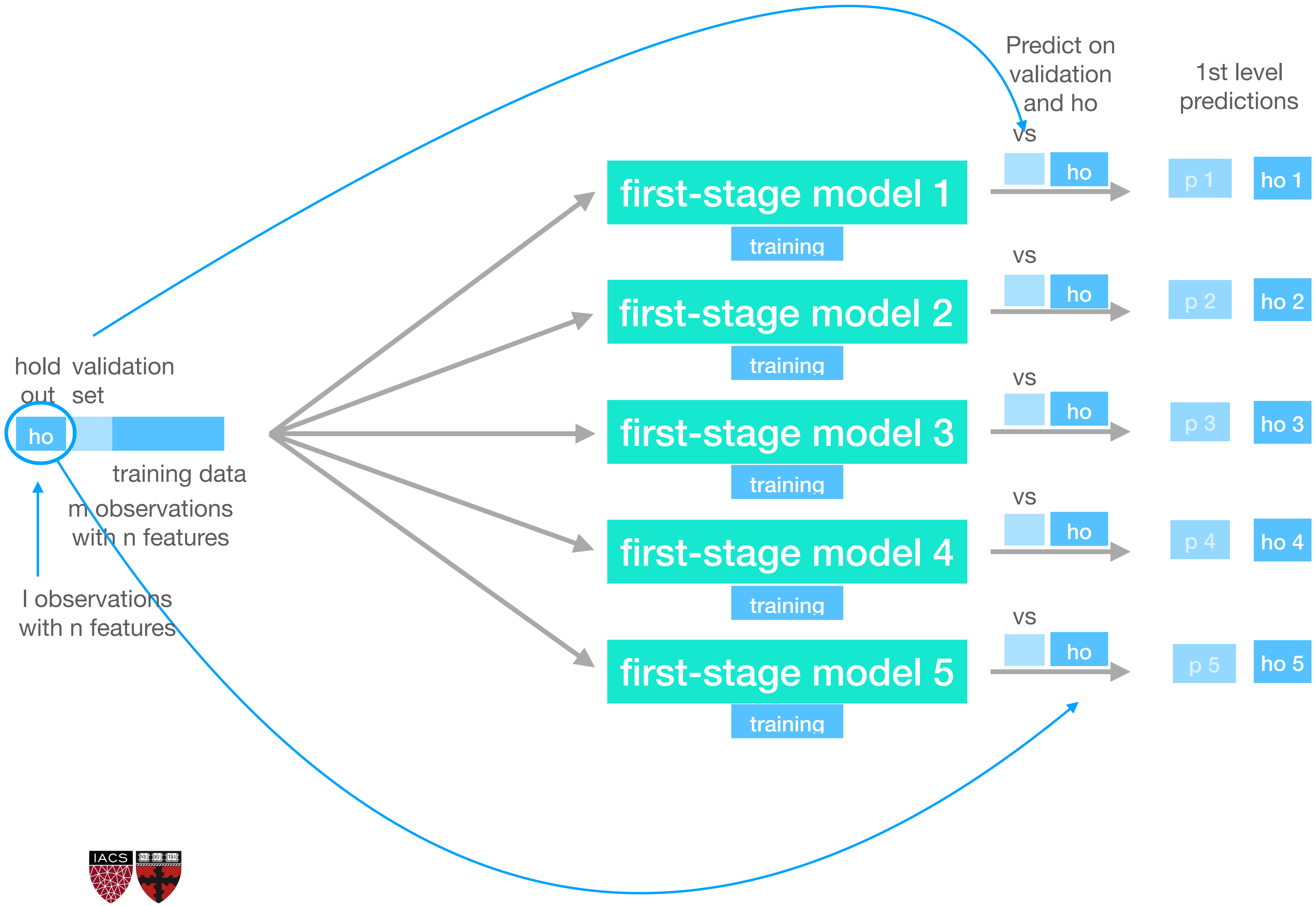
Blending

Independent, parallel, heterogeneous weak learners, by training a meta-model to output a prediction - less bias



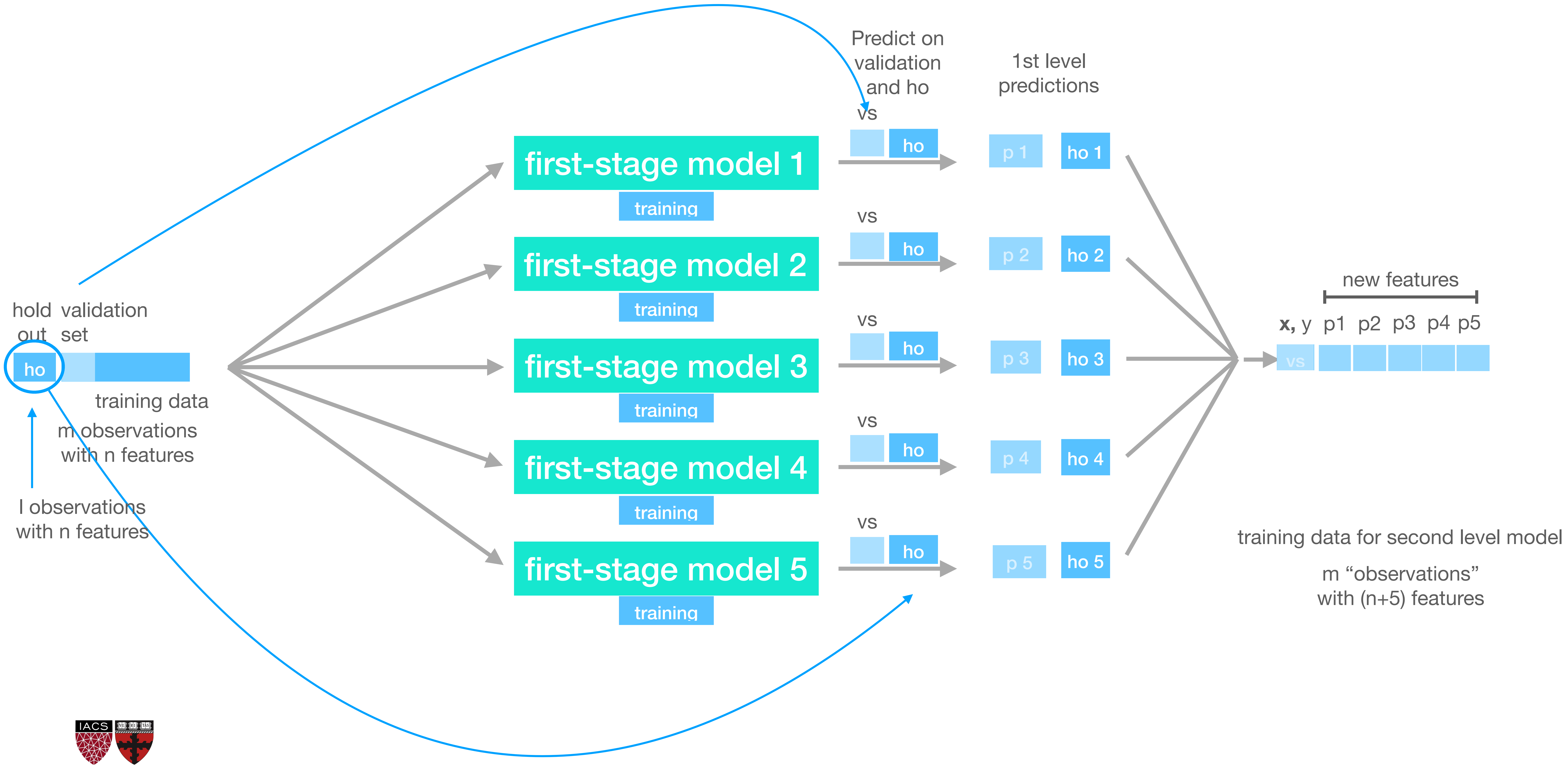
Blending

Independent, parallel, heterogeneous weak learners, by training a meta-model to output a prediction - less bias



Blending

Independent, parallel, heterogeneous weak learners, by training a meta-model to output a prediction - less bias



Blending

Independent, parallel, heterogeneous weak learners, by training a meta-model to output a prediction - less bias

Blending

Independent, parallel, heterogeneous weak learners, by training a meta-model to output a prediction - less bias

Blending

Independent, parallel, heterogeneous weak learners, by training a meta-model to output a prediction - less bias

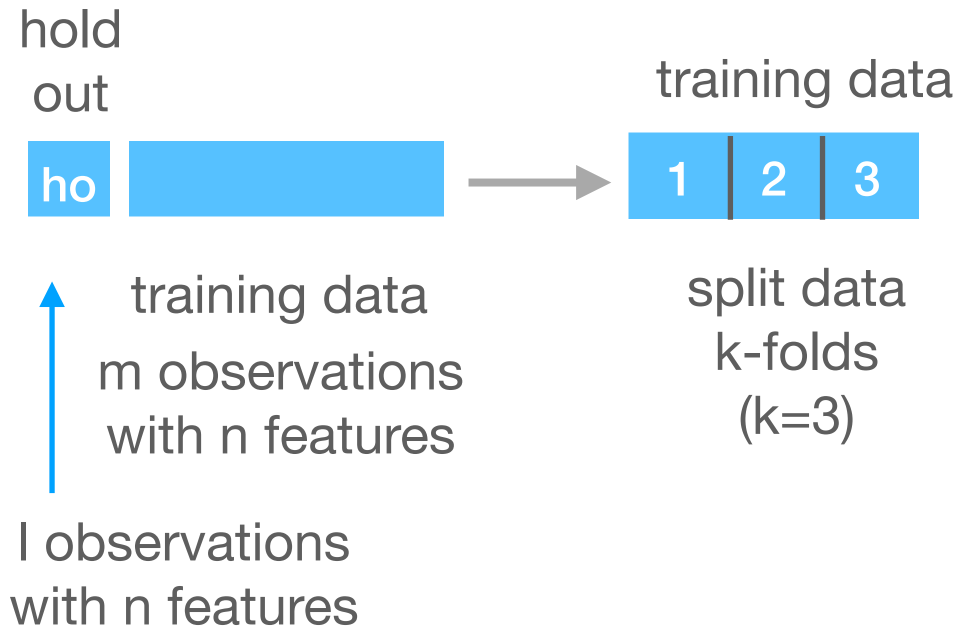
Blending

Independent, parallel, heterogeneous weak learners, by training a meta-model to output a prediction - less bias

Stacking

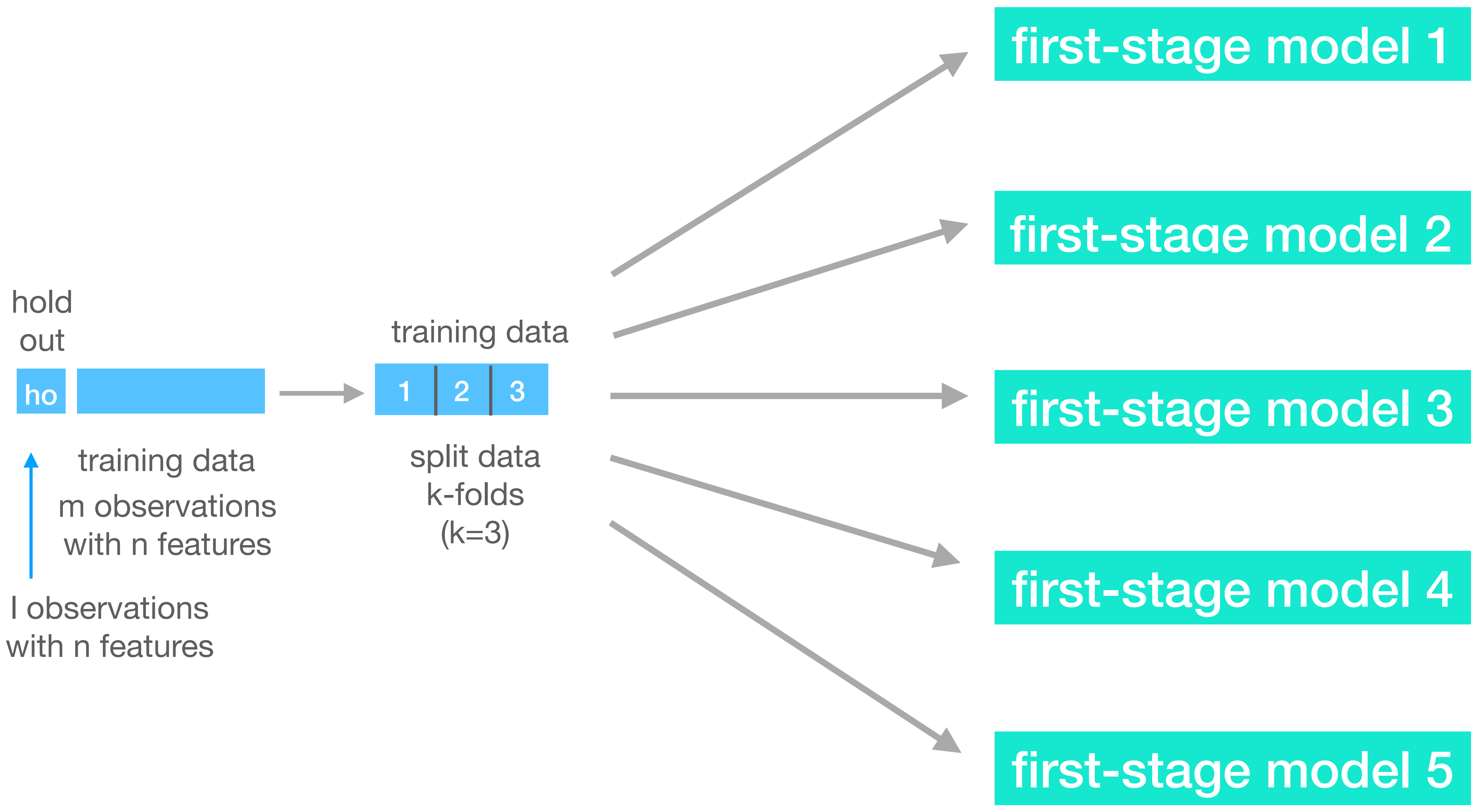
Similar to Blending, except each model is now used to make out of distribution predictions.

“Stacked generalization works by deducing the biases of the generalizer(s) with respect to a provided learning set. This deduction proceeds by generalizing in a second space whose inputs are (for example) the guesses of the original generalizers when taught with part of the learning set and trying to guess the rest of it, and whose output is (for example) the correct guess.” Stack Generalization 1992 - D. Wolpert



Stacking

Similar to Blending, except each model is now used to make out of distribution predictions.



Stacking

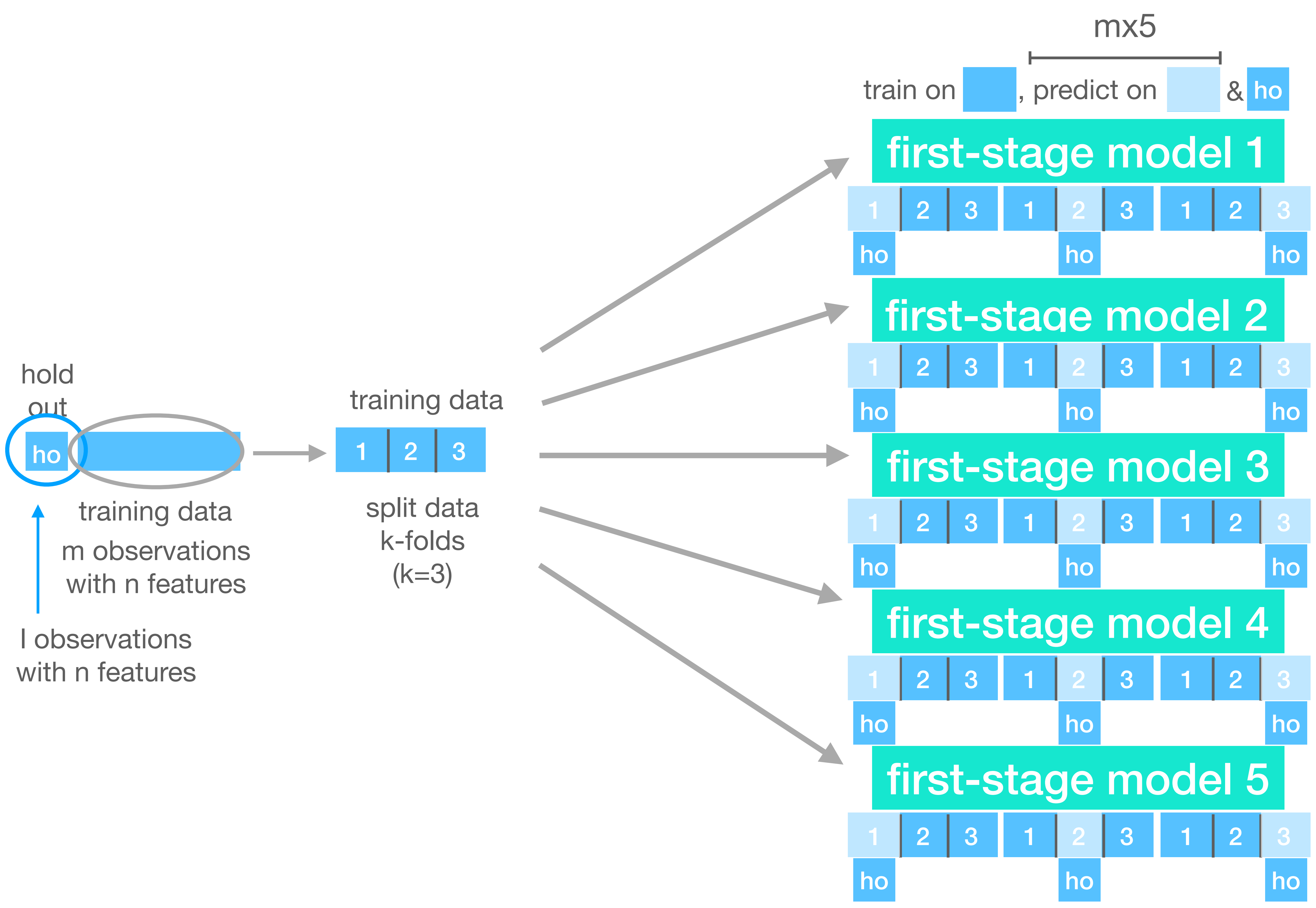
Similar to Blending, except each model is now used to make out of distribution predictions.

Stacking

Similar to Blending, except each model is now used to make out of distribution predictions.

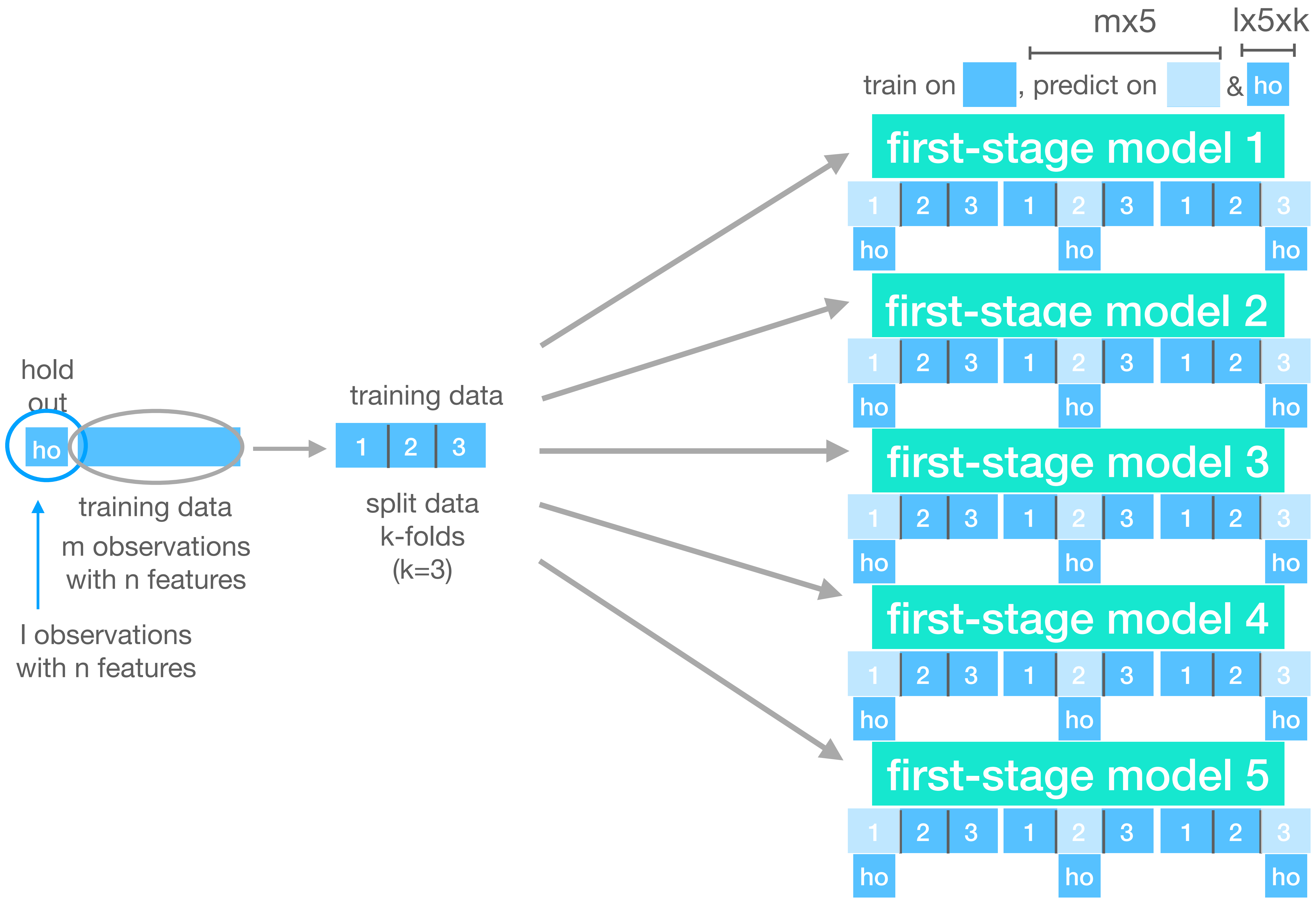
Stacking

Similar to Blending, except each model is now used to make out of distribution predictions.



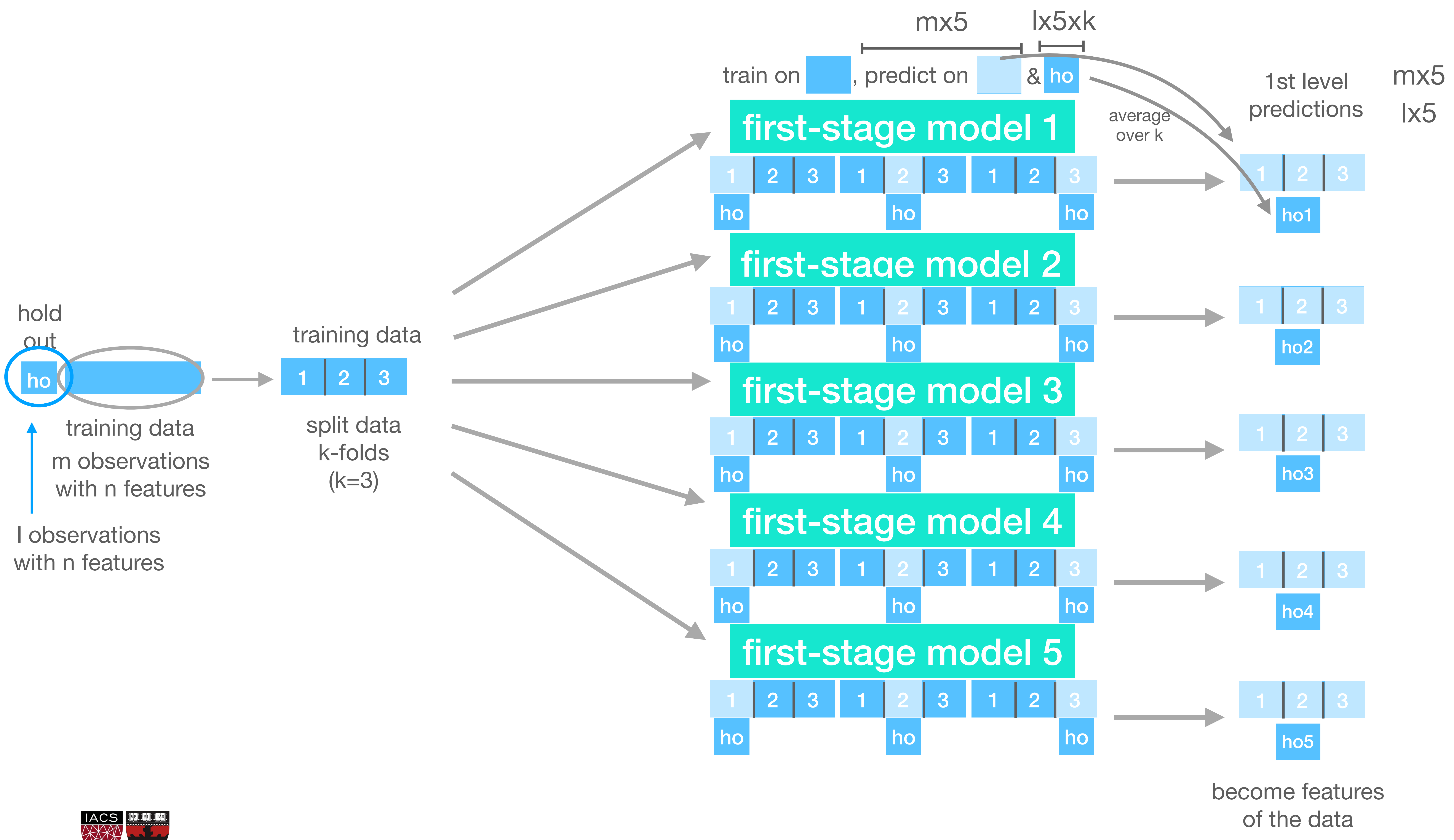
Stacking

Similar to Blending, except each model is now used to make out of distribution predictions.



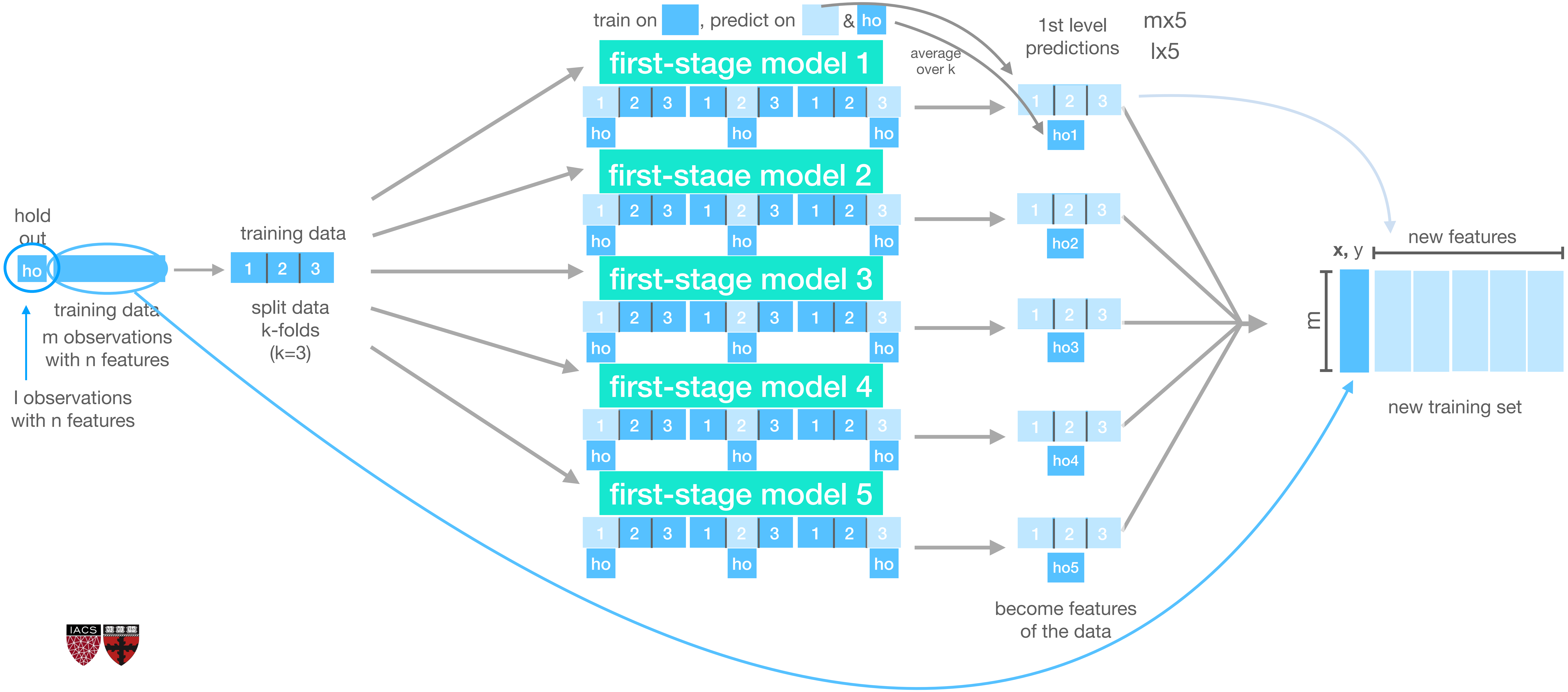
Stacking

Similar to Blending, except each model is now used to make out of distribution predictions.



Stacking

Similar to Blending, except each model is now used to make out of distribution predictions.



Stacking

Similar to Blending, except each model is now used to make out of distribution predictions.

Stacking

Similar to Blending, except each model is now used to make out of distribution predictions.

Outline

- Intuition for ensemble of models and mixture of experts ✓
- Simple ensemble of models for classification and regression: voting and averaging ✓
- More ensemble of models: $\underbrace{\text{bagging, boosting}}_{\text{homogeneous}}$, $\underbrace{\text{blending and}}_{\text{heterogeneous}}$ stacking ✓

  homogeneous ✓ heterogeneous

  less variance less bias
- Mixture of experts and hierarchical mixture of experts

Mixture of Experts

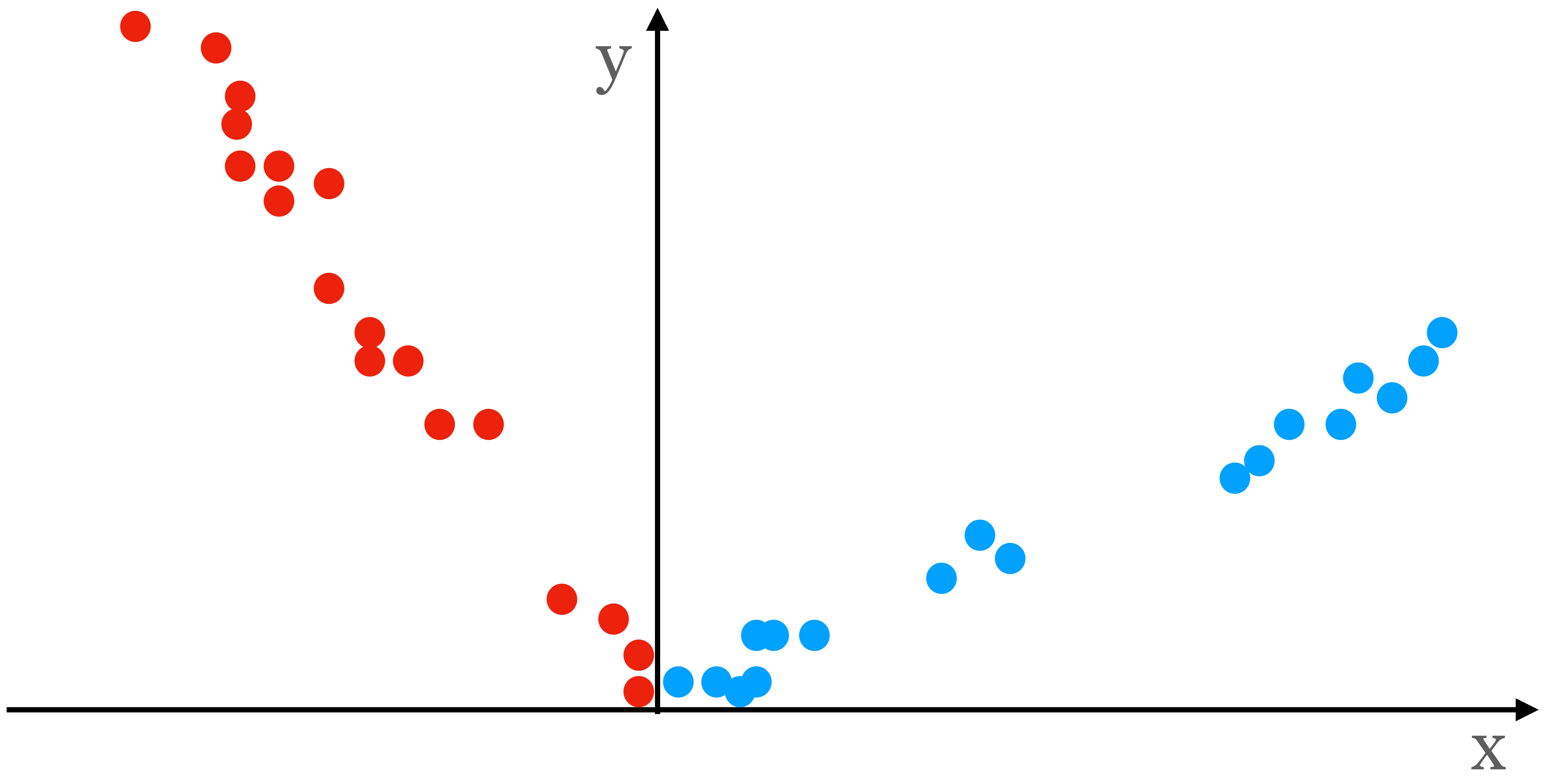
Specialization instead of cooperation

Good if the dataset contains several different regimes which have different relationships between input and output. Covers different input regions with different learners

Mixture of Experts

Specialization instead of cooperation

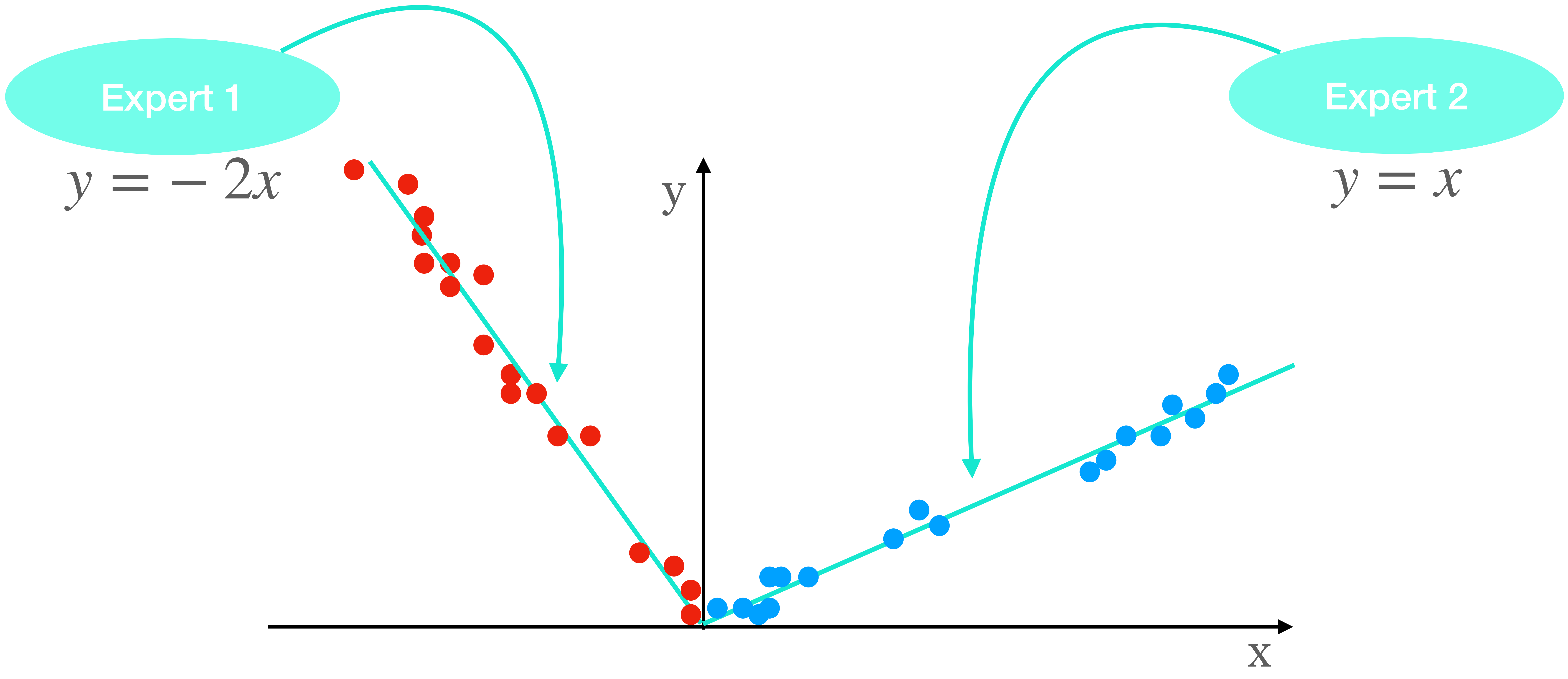
Good if the dataset contains several different regimes which have different relationships between input and output. Covers different input regions with different learners



Mixture of Experts

Specialization instead of cooperation

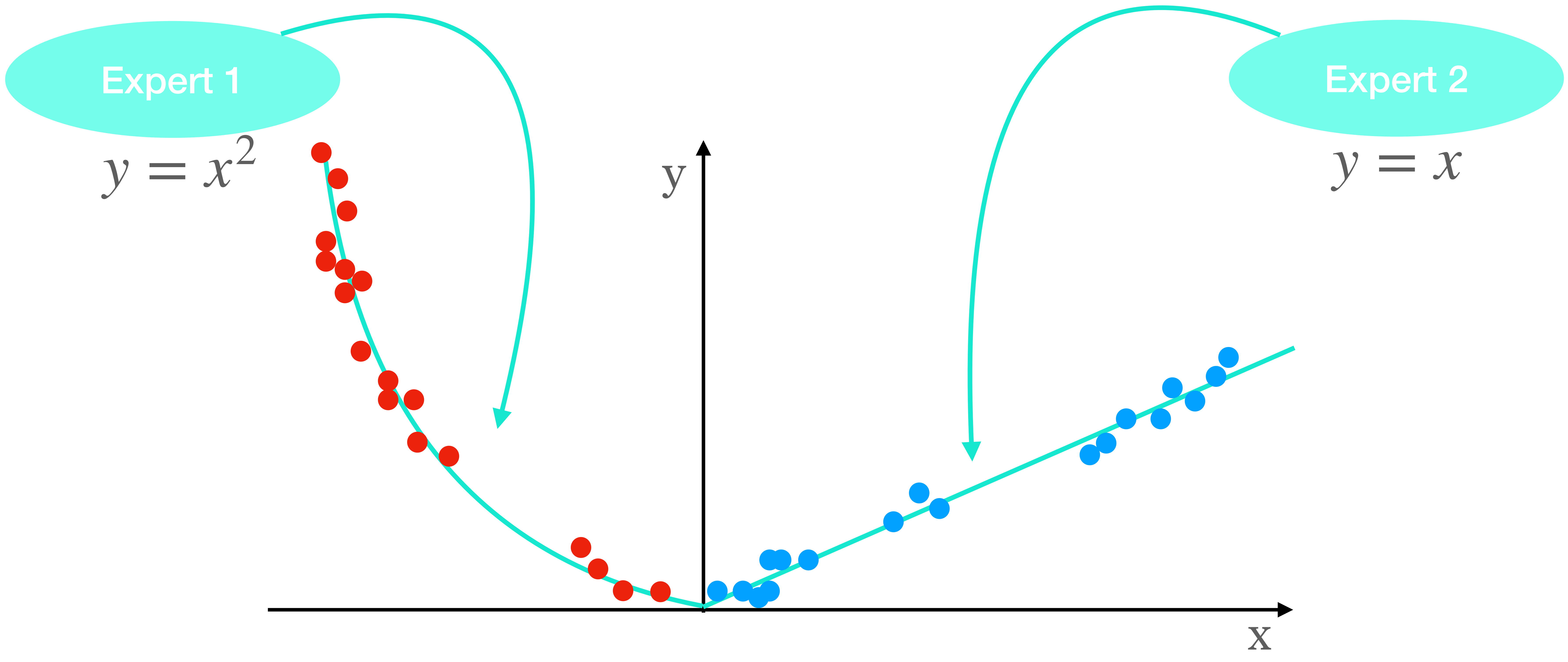
Good if the dataset contains several different regimes which have different relationships between input and output. Covers different input regions with different learners



Mixture of Experts

Specialization instead of cooperation

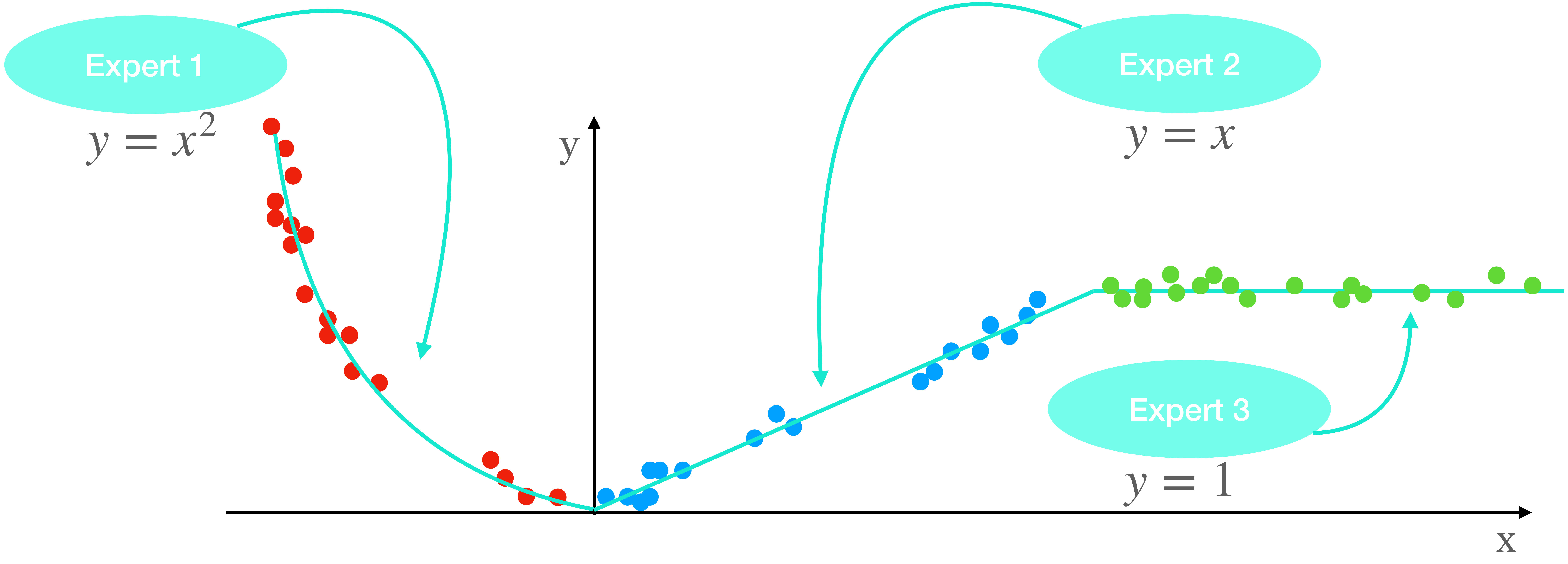
Good if the dataset contains several different regimes which have different relationships between input and output. Covers different input regions with different learners



Mixture of Experts

Specialization instead of cooperation

Good if the dataset contains several different regimes which have different relationships between input and output. Covers different input regions with different learners



Mixture of Experts

Specialization instead of cooperation

Good if the dataset contains several different regimes which have different relationships between input and output. Covers different input regions with different learners

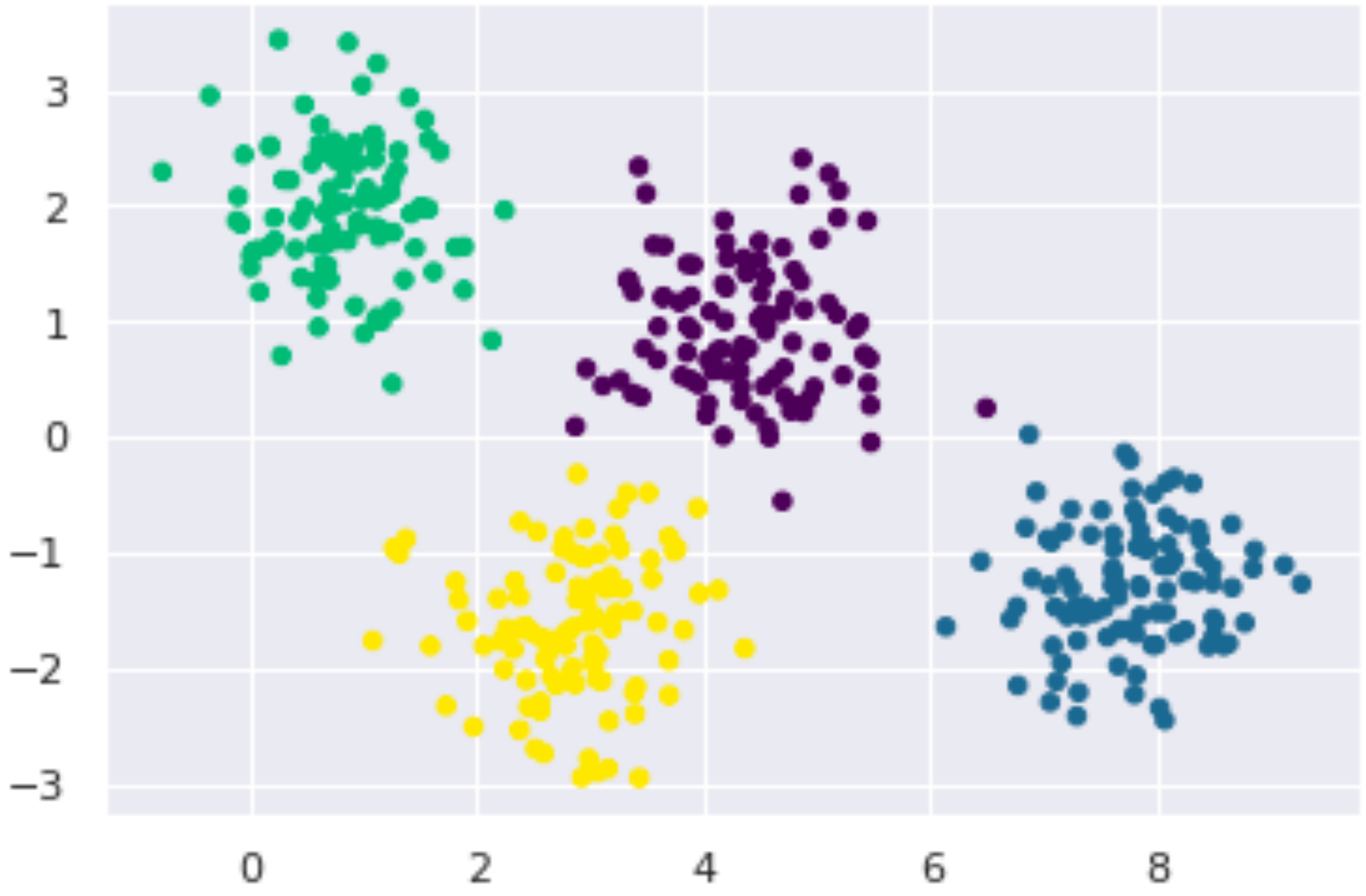
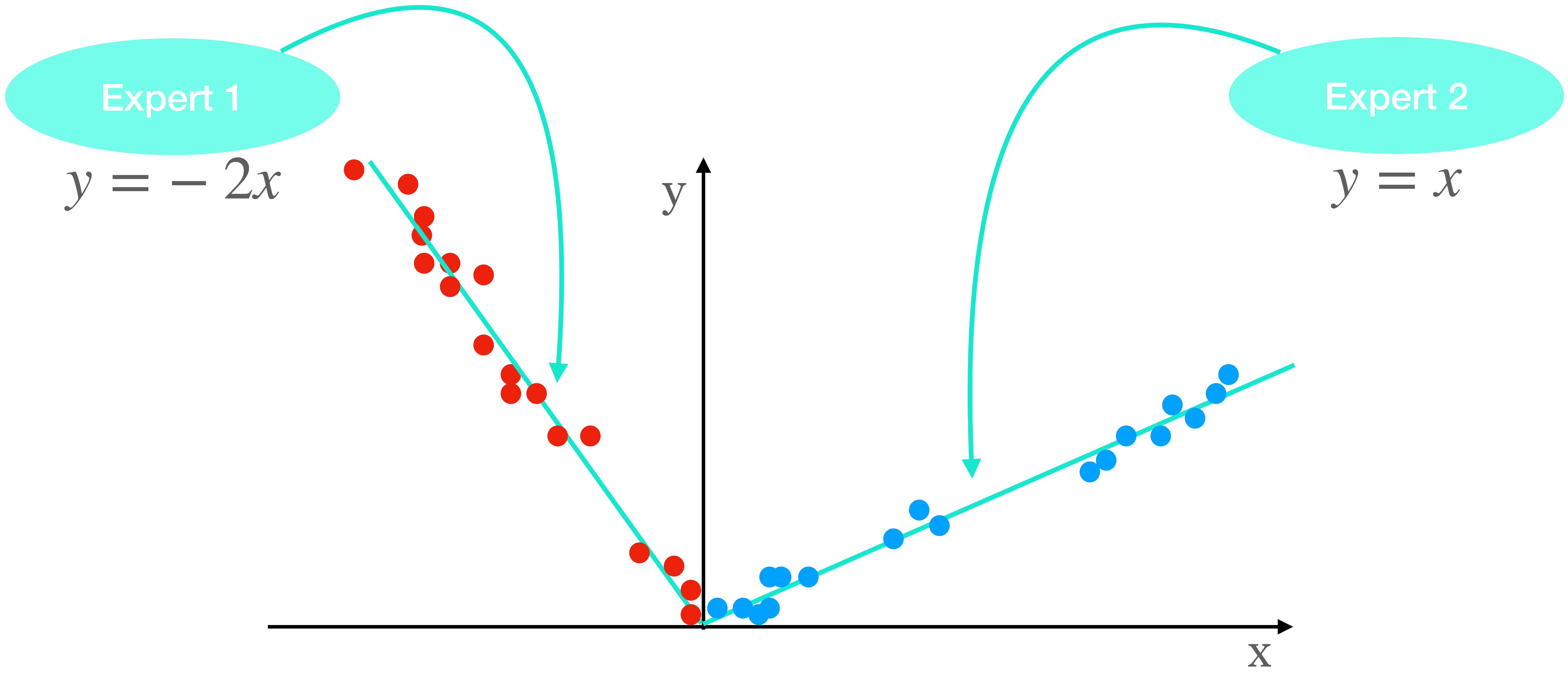


Image credit: Data Flair

Mixture of Experts

Specialization instead of cooperation

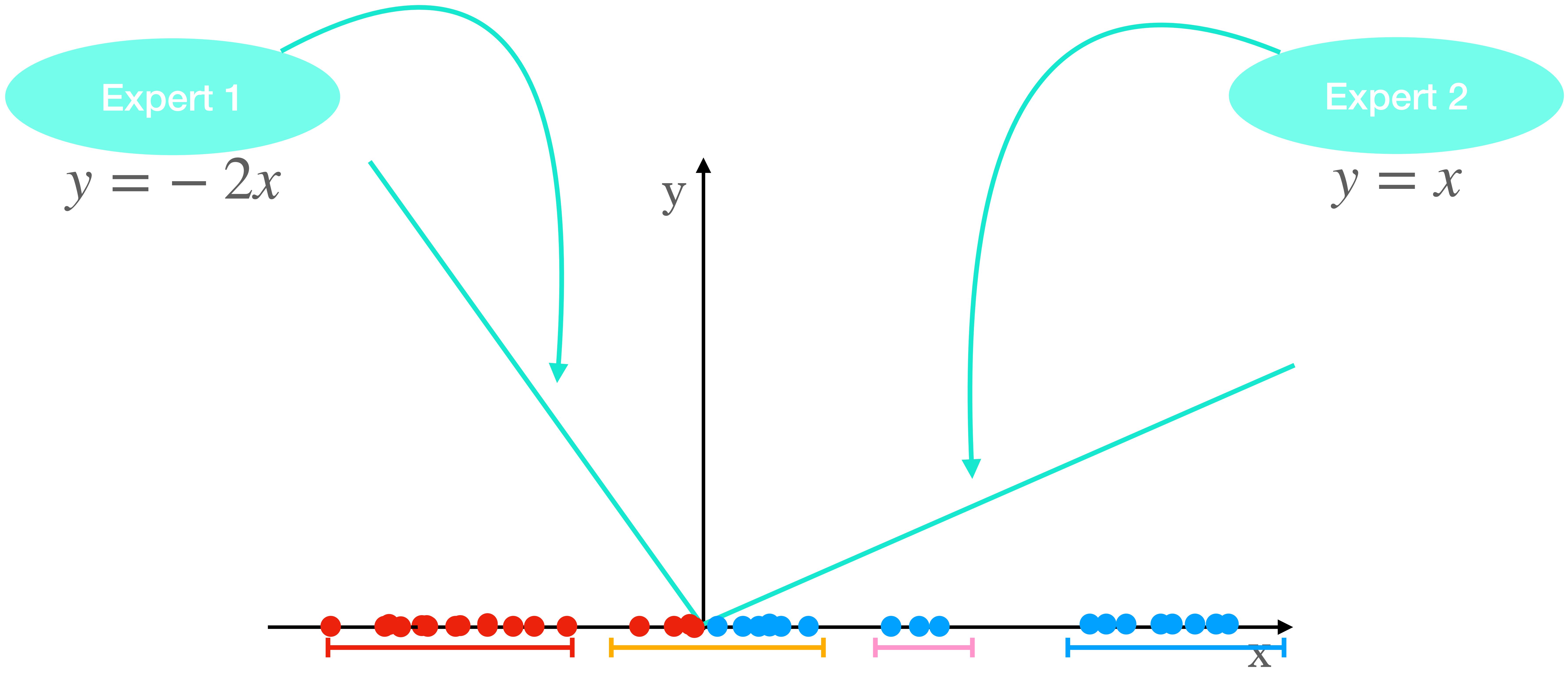
Good if the dataset contains several different regimes which have different relationships between input and output. Covers different input regions with different learners



Mixture of Experts

Specialization instead of cooperation

Good if the dataset contains several different regimes which have different relationships between input and output. Covers different input regions with different learners



Mixture of Experts

Specialization instead of cooperation

Good if the dataset contains several different regimes which have different relationships between input and output. Covers different input regions with different learners

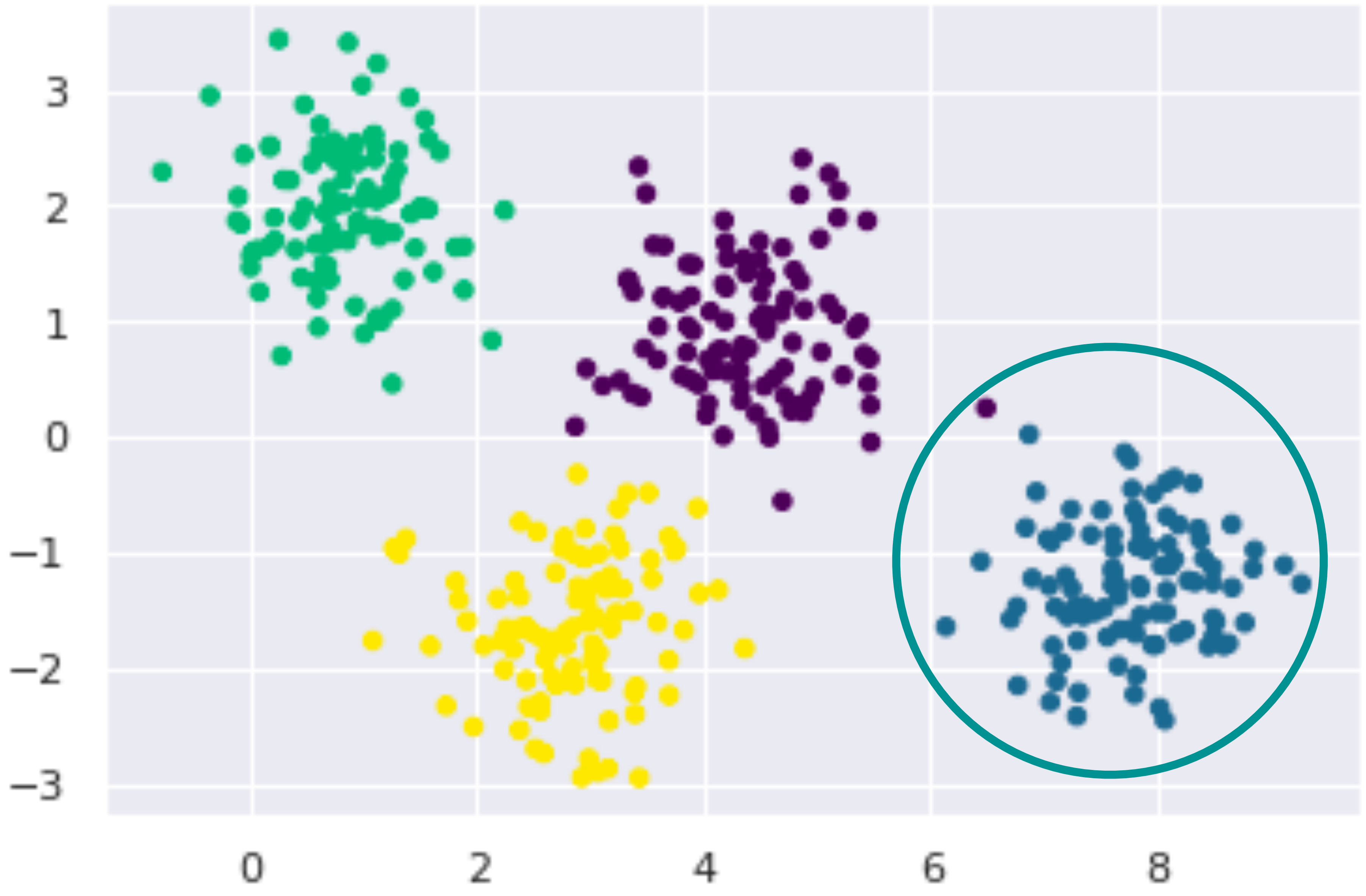
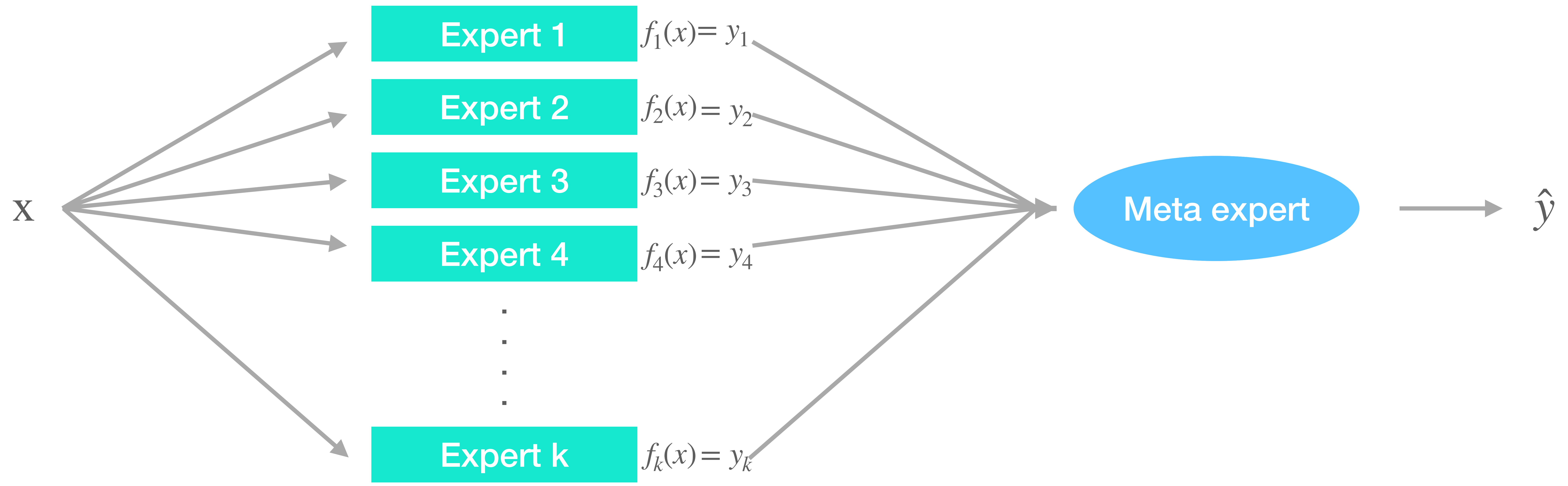


Image credit: Data Flair

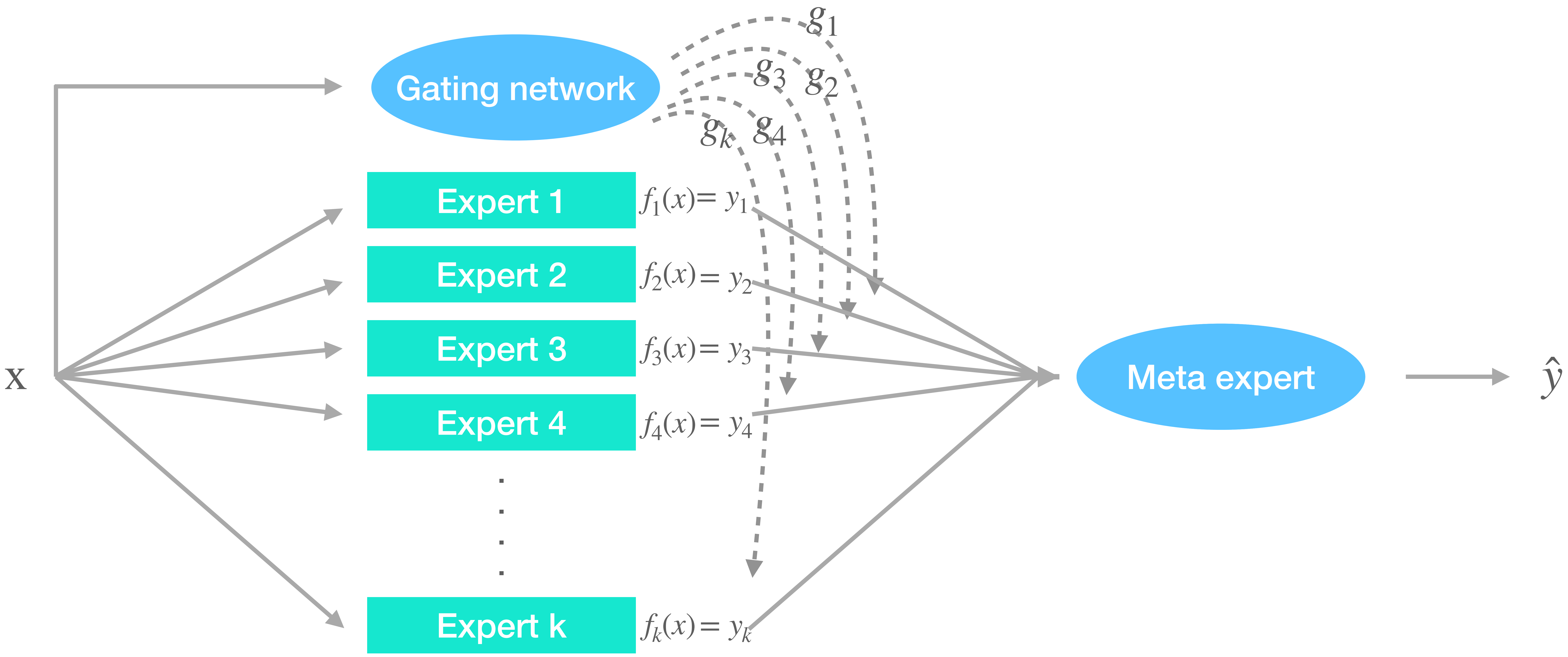
Mixture of Experts

Uses different learners or combination of experts for different regions of the input data.



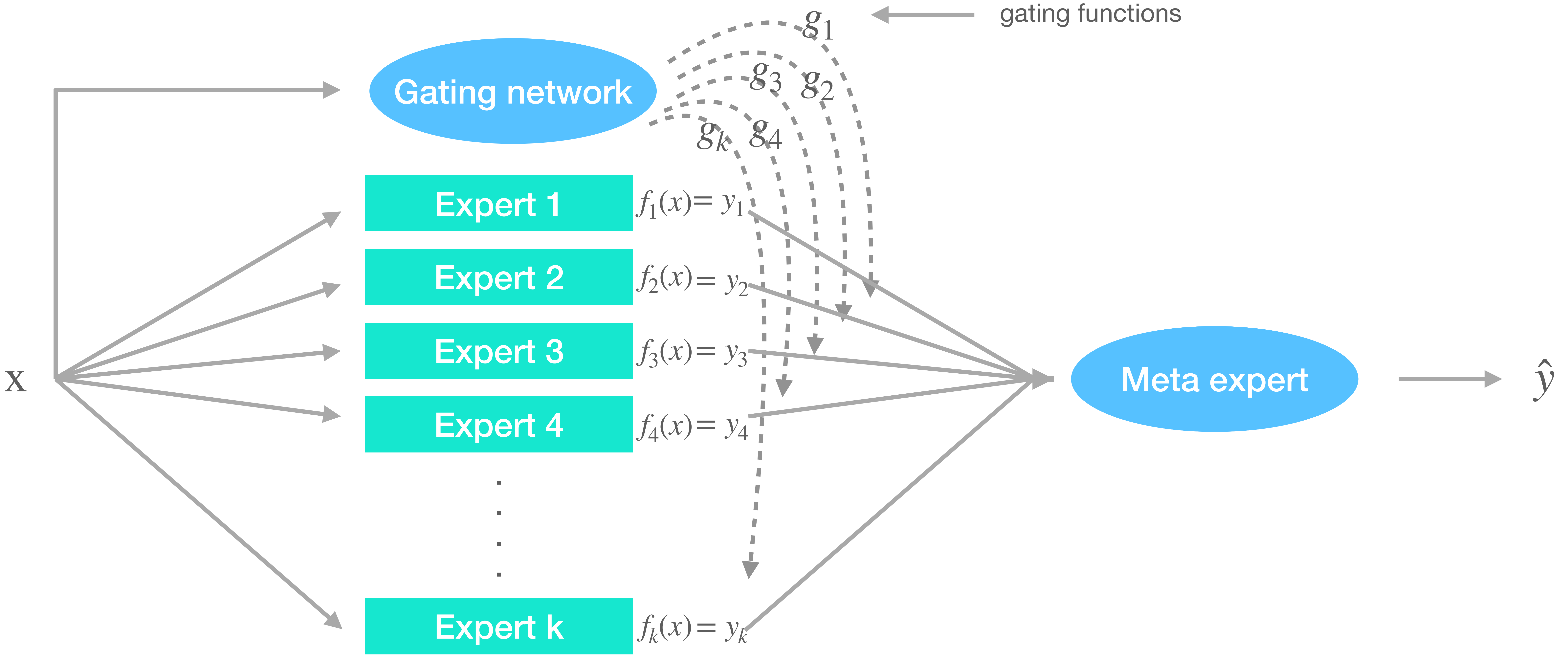
Mixture of Experts

Uses different learners or combination of experts for different regions of the input data.



Mixture of Experts

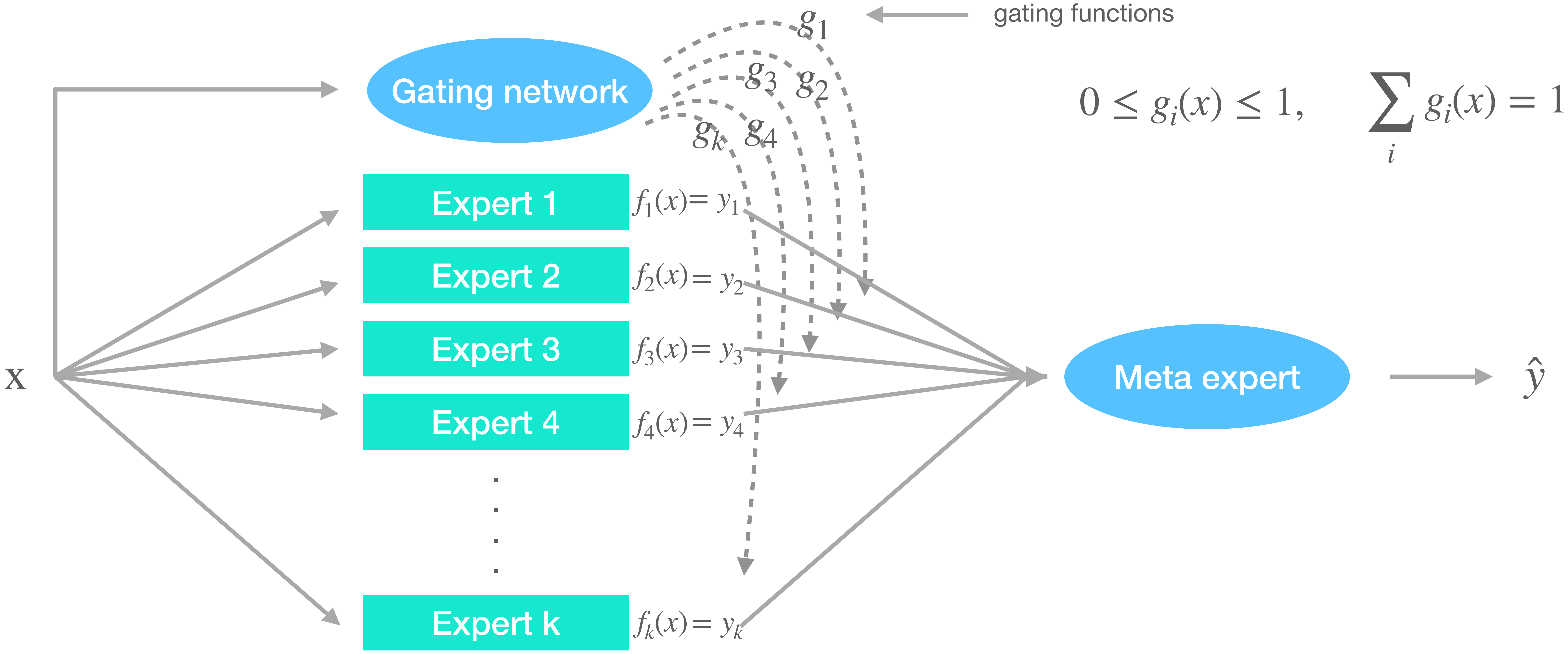
Uses different learners or combination of experts for different regions of the input data.



The models are mixed using a gating network that decides what combination of experts to use

Mixture of Experts

Uses different learners or combination of experts for different regions of the input data.

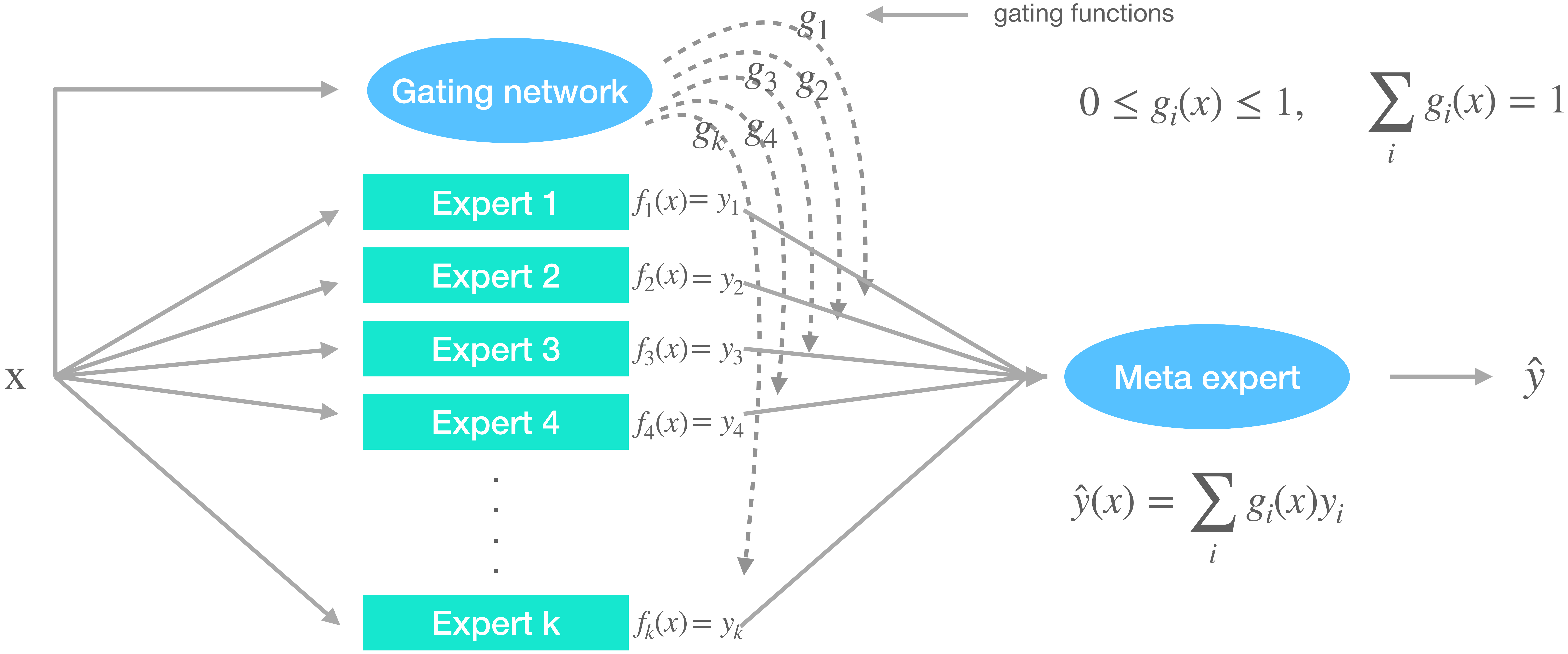


$$0 \le g_i(x) \le 1, \quad \sum_i g_i(x) = 1$$

The models are mixed using a gating network that decides what combination of experts to use

Mixture of Experts

Uses different learners or combination of experts for different regions of the input data.



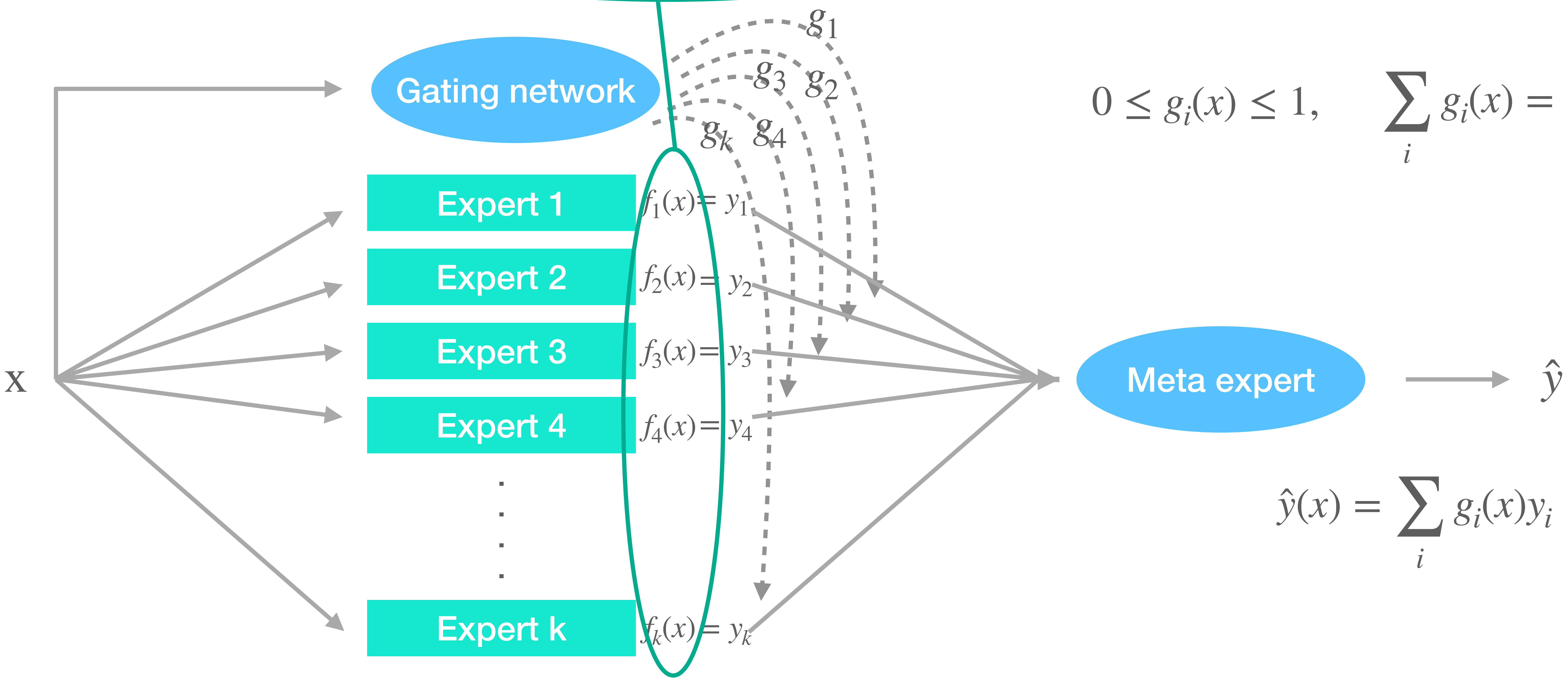
$$0 \leq g_i(x) \leq 1, \quad \sum_i g_i(x) = 1$$

$$\hat{y}(x) = \sum_i g_i(x) y_i$$

The models are mixed using a gating network that decides what combination of experts to use

Mixture of Experts

Learning involves learning the parameters of each expert

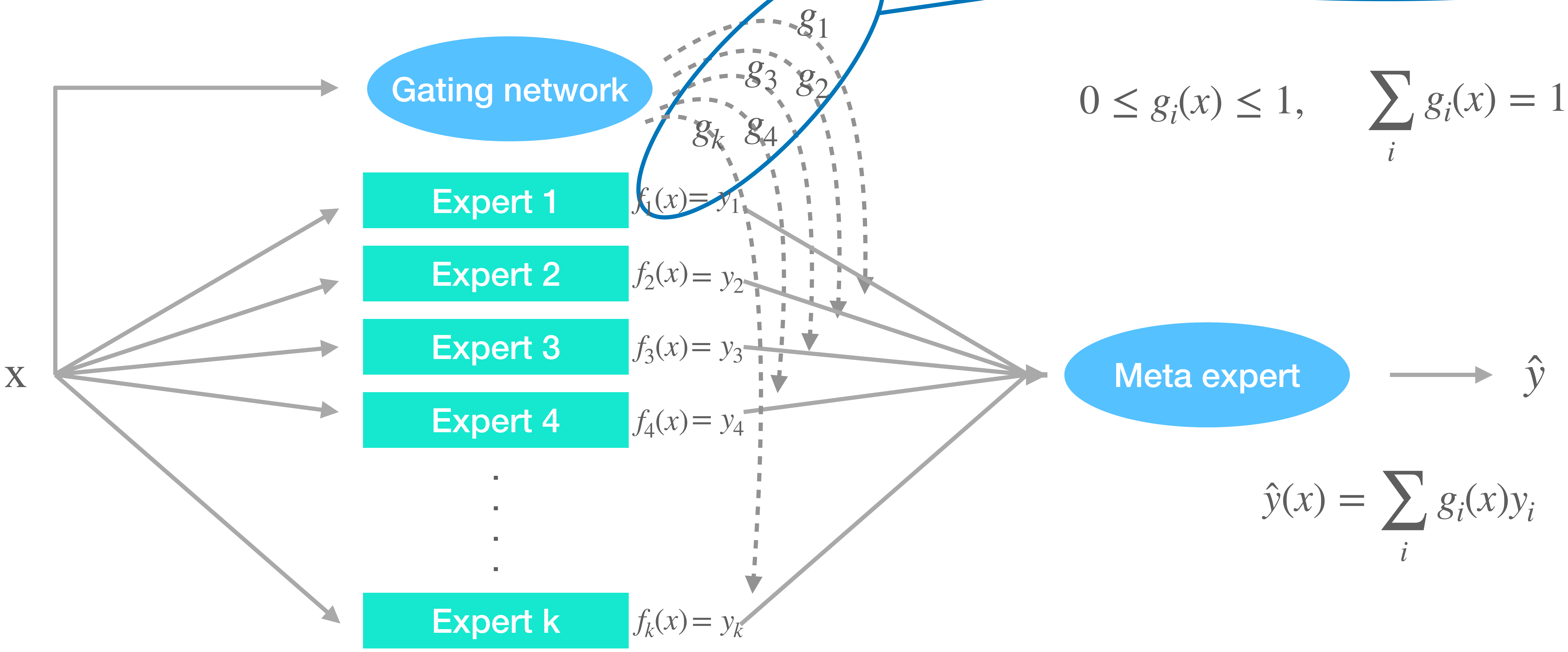


$$0 \le g_i(x) \le 1, \quad \sum_i g_i(x) = 1$$

$$\hat{y}(x) = \sum_i g_i(x) y_i$$

Mixture of Experts

Learning involves learning the parameters of each expert and the parameters of the gating network

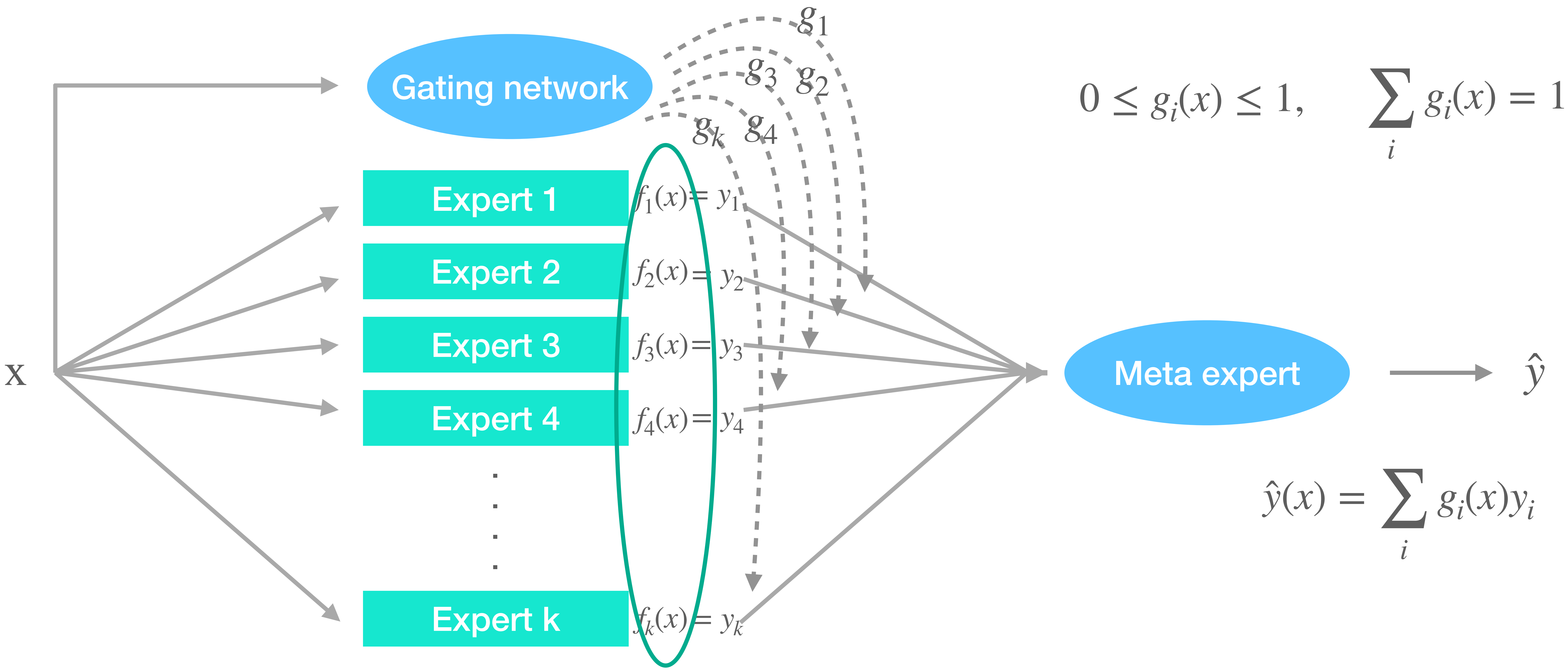


$$0 \leq g_i(x) \leq 1, \quad \sum_i g_i(x) = 1$$

$$\hat{y}(x) = \sum_i g_i(x) y_i$$

Mixture of Experts

Learning involves learning the parameters of each expert and the parameters of the gating network



$$0 \leq g_i(x) \leq 1, \quad \sum_i g_i(x) = 1$$

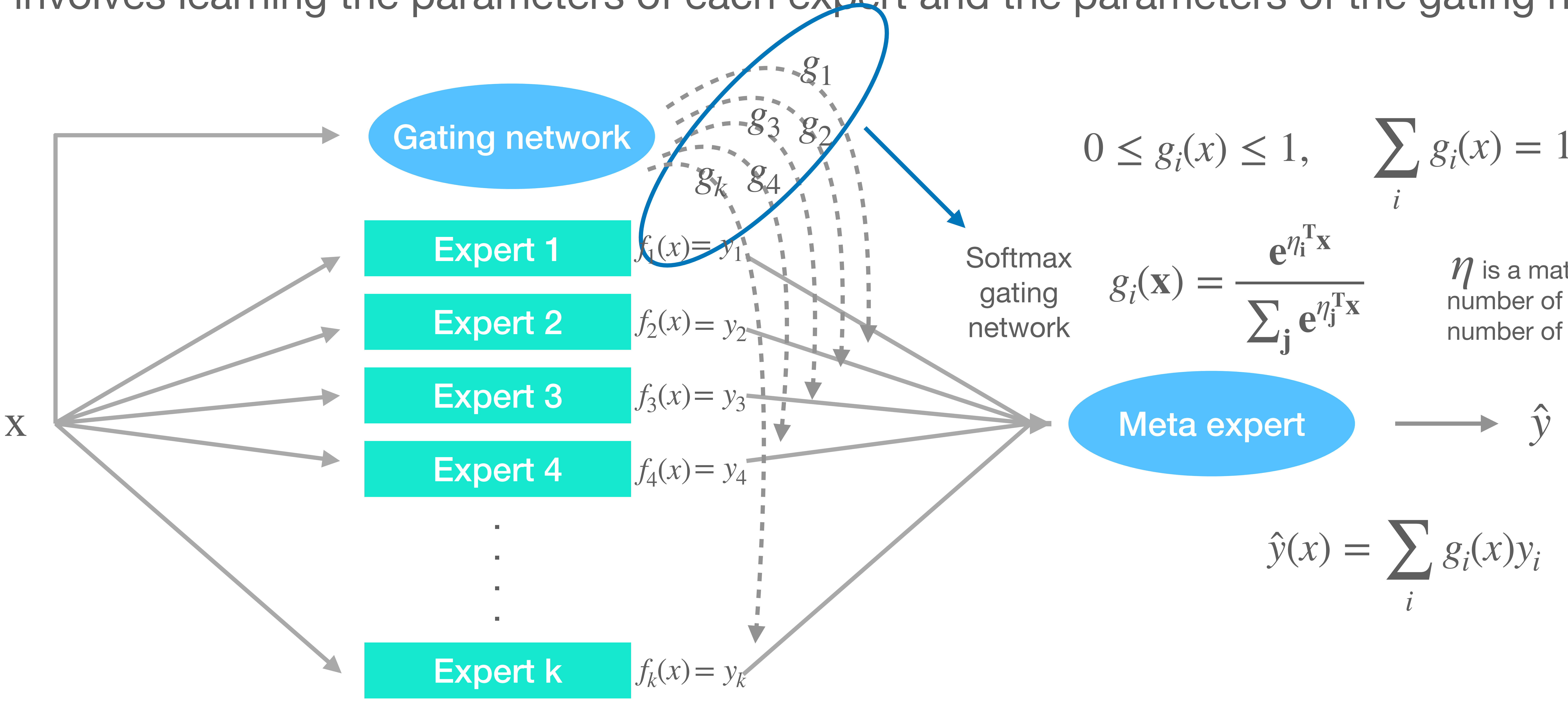
$$\hat{y}(x) = \sum_i g_i(x) y_i$$

$$f_i(x) = \theta_{\mathbf{i}}^{\mathbf{T}} \mathbf{x}$$

Set of linear experts

Mixture of Experts

Learning involves learning the parameters of each expert and the parameters of the gating network



$$0 \le g_i(x) \le 1, \quad \sum_i g_i(x) = 1$$

$$g_i(\mathbf{x}) = \frac{\mathbf{e}^{\eta_i^{\mathbf{T}}\mathbf{x}}}{\sum_{\mathbf{j}} \mathbf{e}^{\eta_j^{\mathbf{T}}\mathbf{x}}}$$

η is a matrix of dim: number of experts by number of features

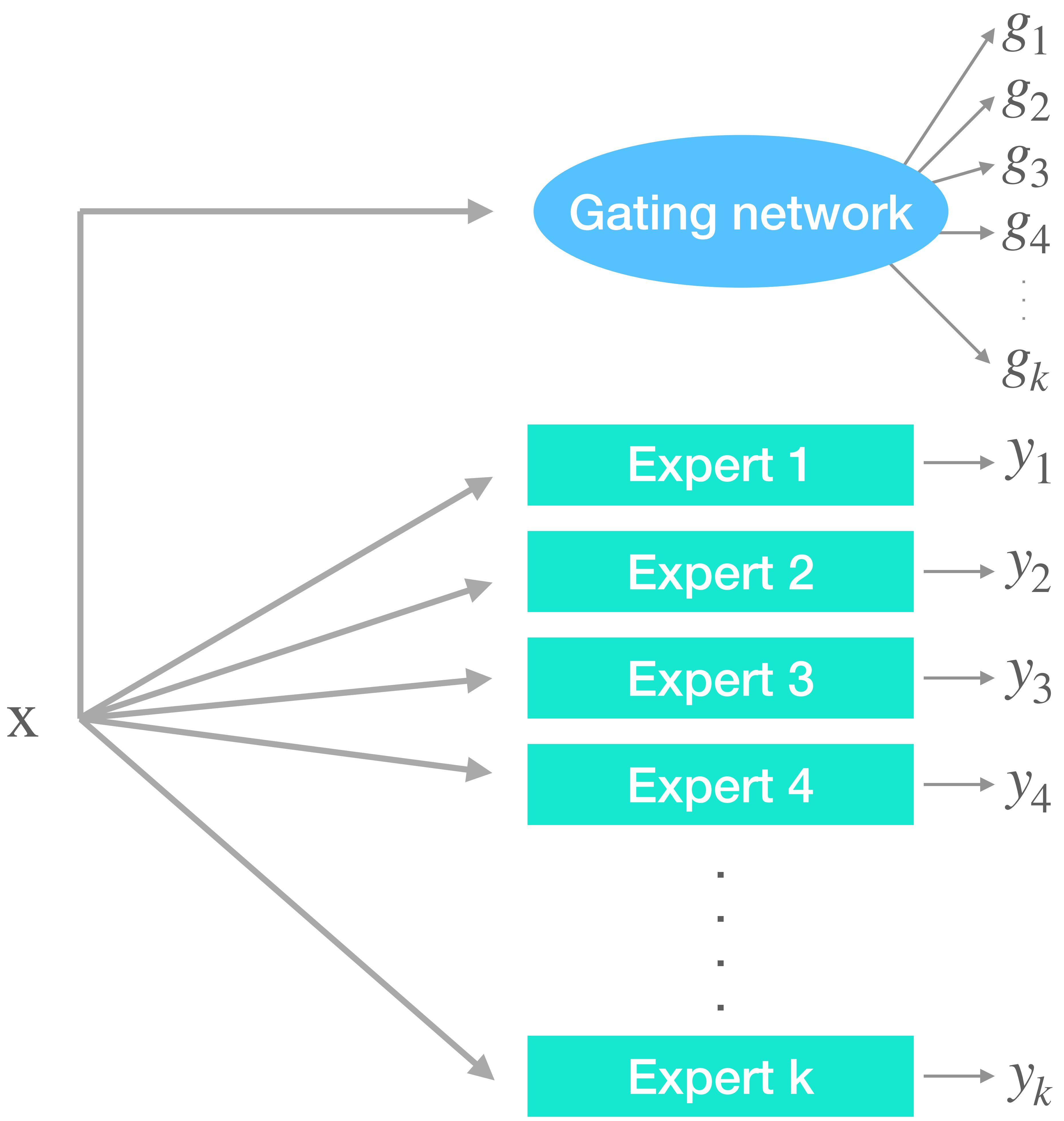
$$\hat{y}(x) = \sum_i g_i(x) y_i$$

$$f_i(x) = \theta_{\mathbf{i}}^{\mathbf{T}}\mathbf{x}$$

Set of linear experts

Mixture of Experts

Learning involves learning the parameters of each expert and the parameters of the gating network



$$0 \leq g_i \leq 1,$$

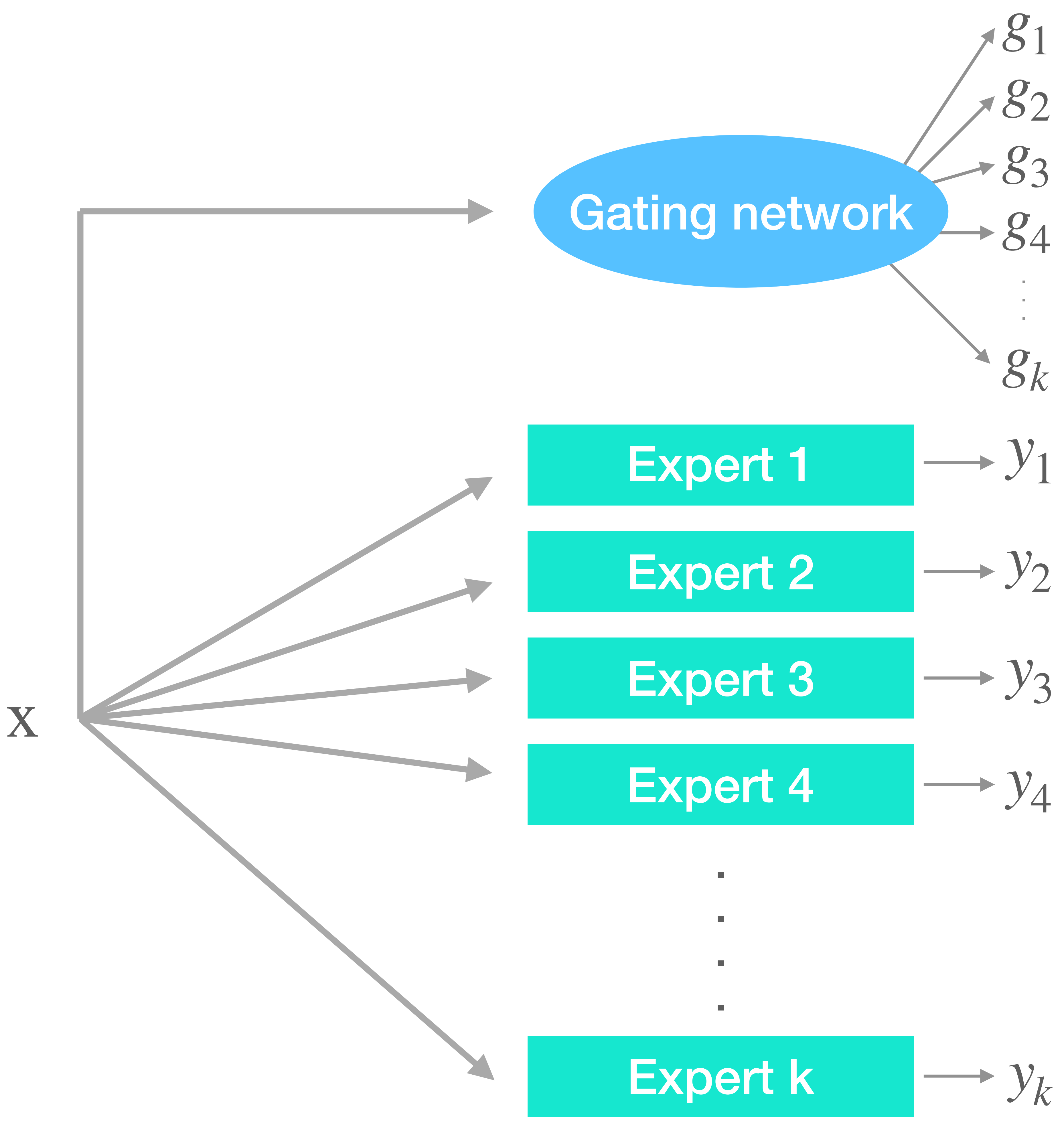
$$\sum_i g_i = 1$$

$$g_i(\mathbf{x}) = \frac{\mathbf{e}^{\eta_i^T \mathbf{x}}}{\sum_j \mathbf{e}^{\eta_j^T \mathbf{x}}}$$

η is a matrix of dim: number of experts by number of features

Mixture of Experts

Learning involves learning the parameters of each expert and the parameters of the gating network



$$0 \leq g_i \leq 1,$$

$$\sum_i g_i = 1$$

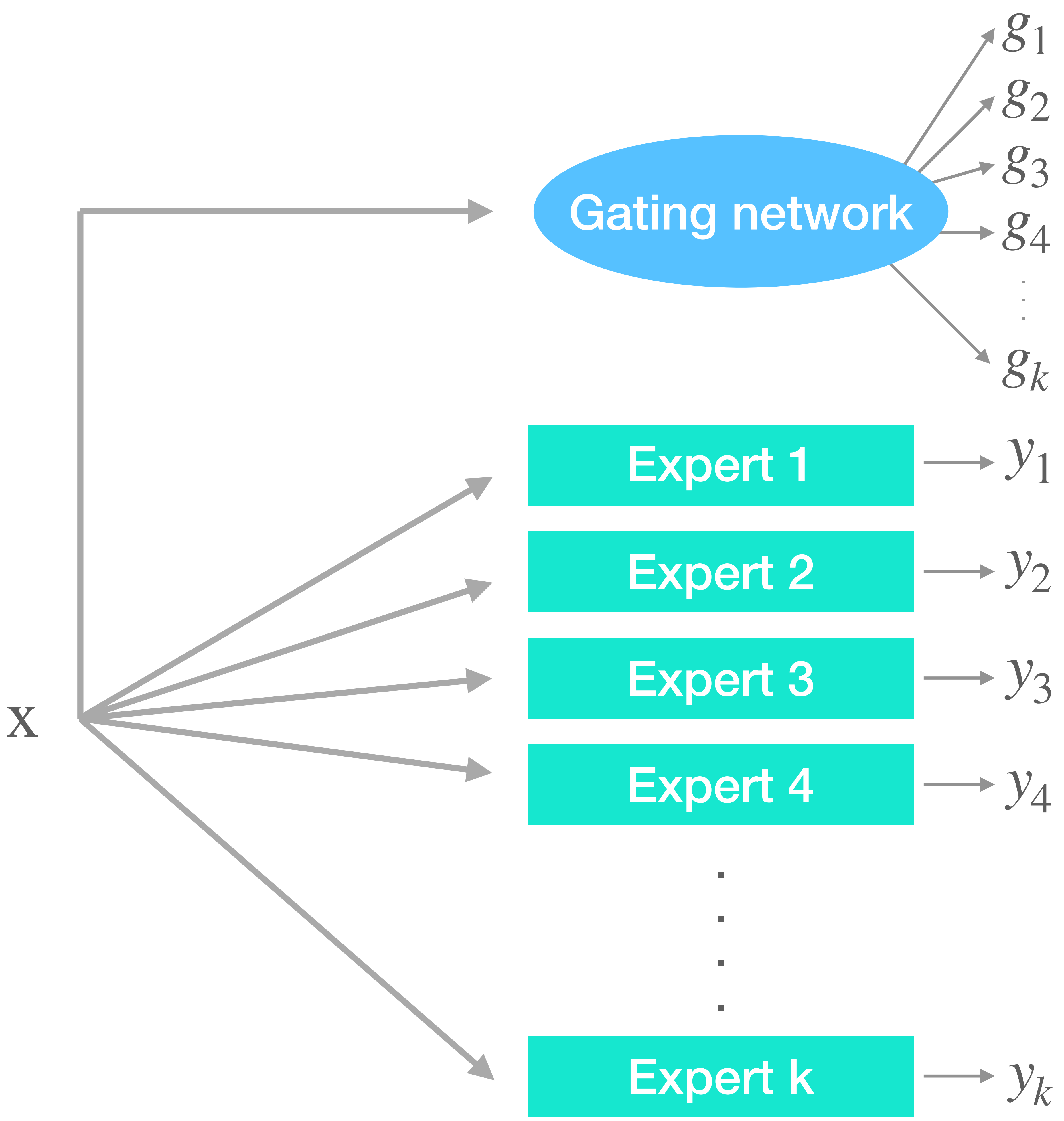
$$g_i(\mathbf{x}) = \frac{\mathbf{e}^{\eta_i^{\mathrm{T}}\mathbf{x}}}{\sum_{\mathbf{j}} \mathbf{e}^{\eta_j^{\mathrm{T}}\mathbf{x}}}$$

η is a matrix of dim: number of experts by number of features

Cooperation $\longrightarrow$ $L = (y - \frac{1}{n}\sum_{i=1}^{n} \hat{y}_i)^2$

Mixture of Experts

Learning involves learning the parameters of each expert and the parameters of the gating network



$0 \leq g_i \leq 1,$

$$\sum_i g_i = 1 \qquad g_i(\mathbf{x}) = \frac{\mathbf{e}^{\eta_\mathbf{i}^\mathbf{T}\mathbf{x}}}{\sum_\mathbf{j} \mathbf{e}^{\eta_\mathbf{j}^\mathbf{T}\mathbf{x}}}$$

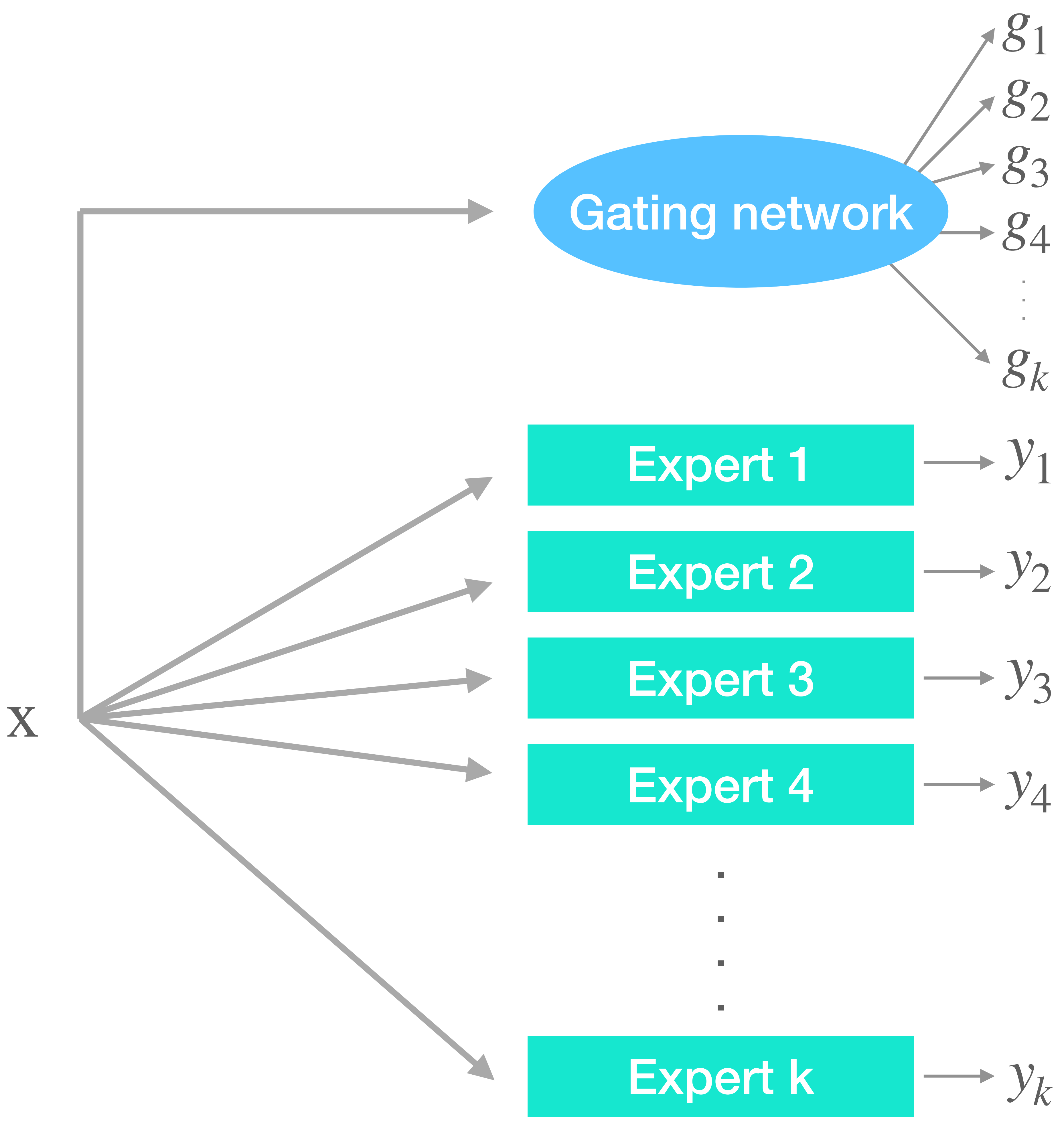
η is a matrix of dim: number of experts by number of features

Cooperation $\longrightarrow$ $\cancel{L = (y - \frac{1}{n}\sum_{i=1}^{n} \hat{y}_i)^2}$

Specialization $\longrightarrow$ $L = \sum_{i=1}^{n} g_i (y - \hat{y}_i)^2$

Mixture of Experts

Learning involves learning the parameters of each expert and the parameters of the gating network



$$0 \leq g_i \leq 1,$$

$$\sum_i g_i = 1$$

$$g_i(\mathbf{x}) = \frac{\mathbf{e}^{\eta_\mathbf{i}^\mathbf{T}\mathbf{x}}}{\sum_\mathbf{j} \mathbf{e}^{\eta_\mathbf{j}^\mathbf{T}\mathbf{x}}}$$

Simple loss function for training

$$L = \sum_i g_i (y - \hat{y}_i)^2$$

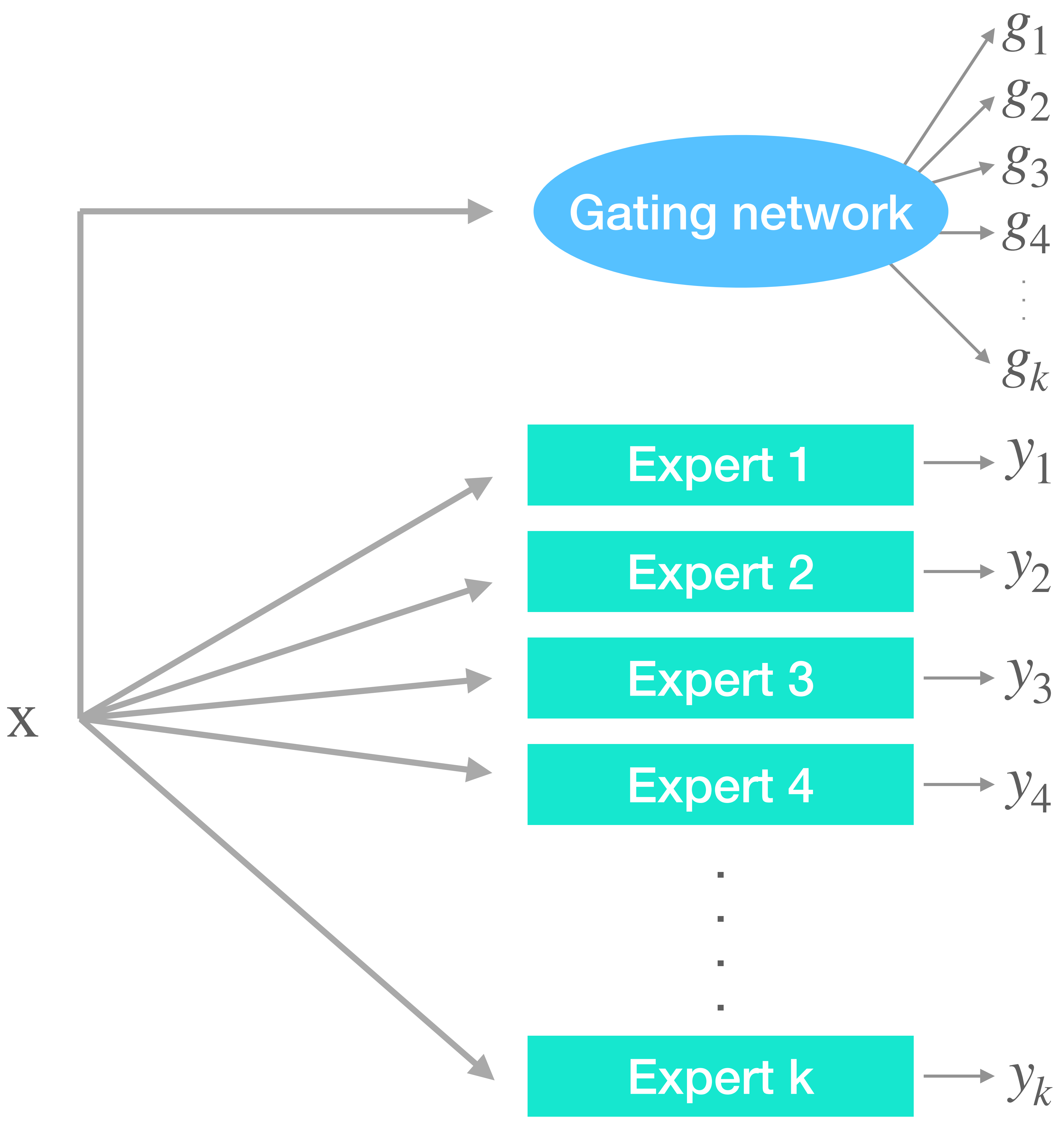
Combined predictor

$$\hat{y} = \sum_i g_i y_i$$

$$y_i = \theta_\mathbf{i}^\mathbf{T}\mathbf{x}$$

Mixture of Experts

Learning involves learning the parameters of each expert and the parameters of the gating network



$$0 \le g_i \le 1,$$

$$\sum_i g_i = 1$$

$$g_i(\mathbf{x}) = \frac{\mathbf{e}^{\eta_\mathbf{i}^\mathbf{T}\mathbf{x}}}{\sum_\mathbf{j} \mathbf{e}^{\eta_\mathbf{j}^\mathbf{T}\mathbf{x}}}$$

Combined predictor

$$\hat{y} = \sum_i g_i y_i$$

$$y_i = \theta_\mathbf{i}^\mathbf{T}\mathbf{x}$$

Simple loss function for training

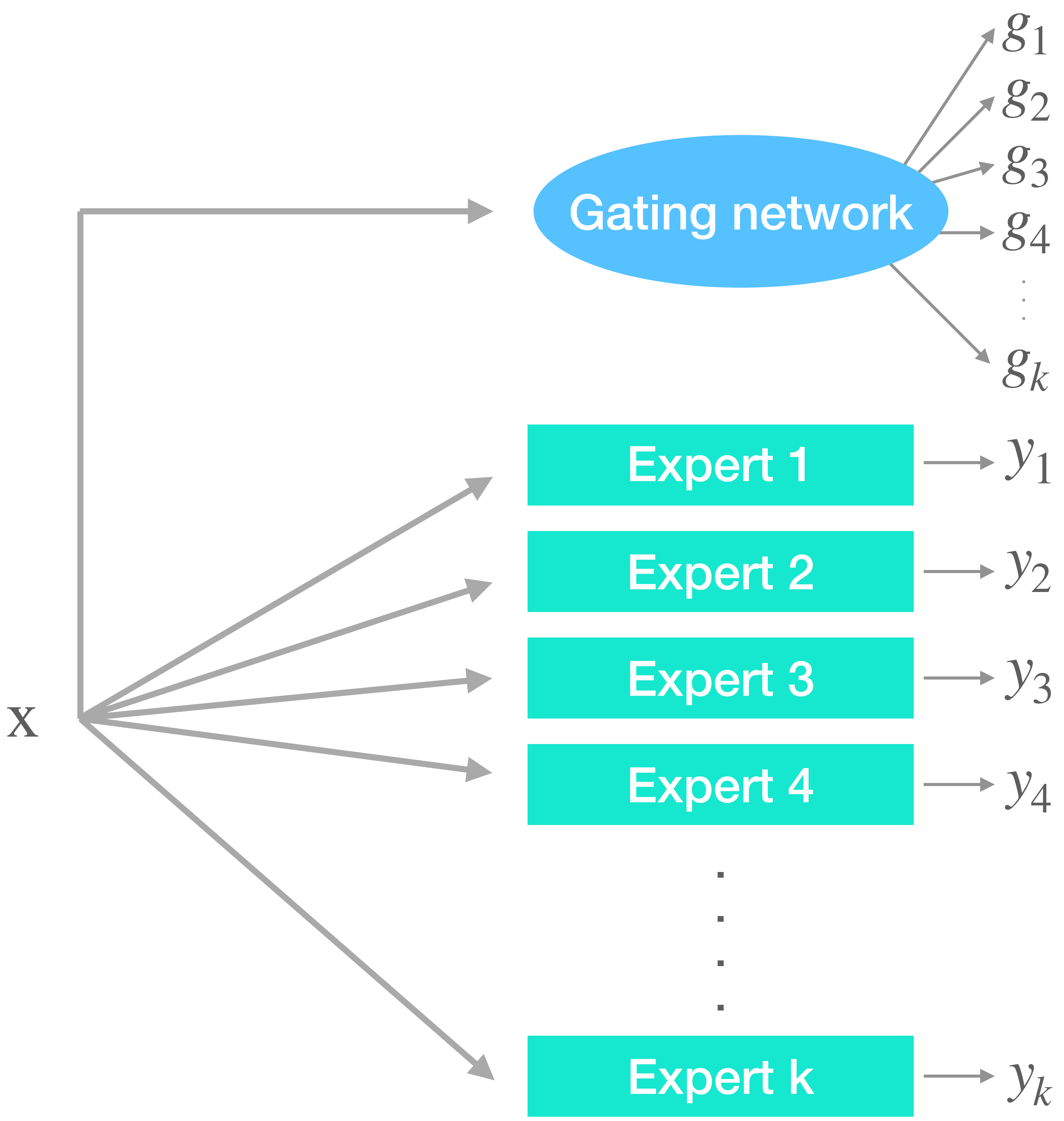
$$L = \sum_i g_i (y - \hat{y}_i)^2$$

Signal for training each expert

$$\frac{\partial L}{\partial \theta_i} = \frac{\partial L}{\partial y_i}\frac{\partial y_i}{\partial \theta_i} \propto g_i(y - y_i)\frac{\partial y_i}{\partial \theta_i}$$

Mixture of Experts

Learning involves learning the parameters of each expert and the parameters of the gating network



$$0 \leq g_i \leq 1,$$

$$\sum_i g_i = 1$$

$$g_i(\mathbf{x}) = \frac{\mathbf{e}^{\eta_{\mathbf{i}}^{\mathbf{T}}\mathbf{x}}}{\sum_{\mathbf{j}} \mathbf{e}^{\eta_{\mathbf{j}}^{\mathbf{T}}\mathbf{x}}}$$

Simple loss function for training $L = \sum_i g_i(y - \hat{y}_i)^2$

Combined predictor

$$\hat{y} = \sum_i g_i y_i$$

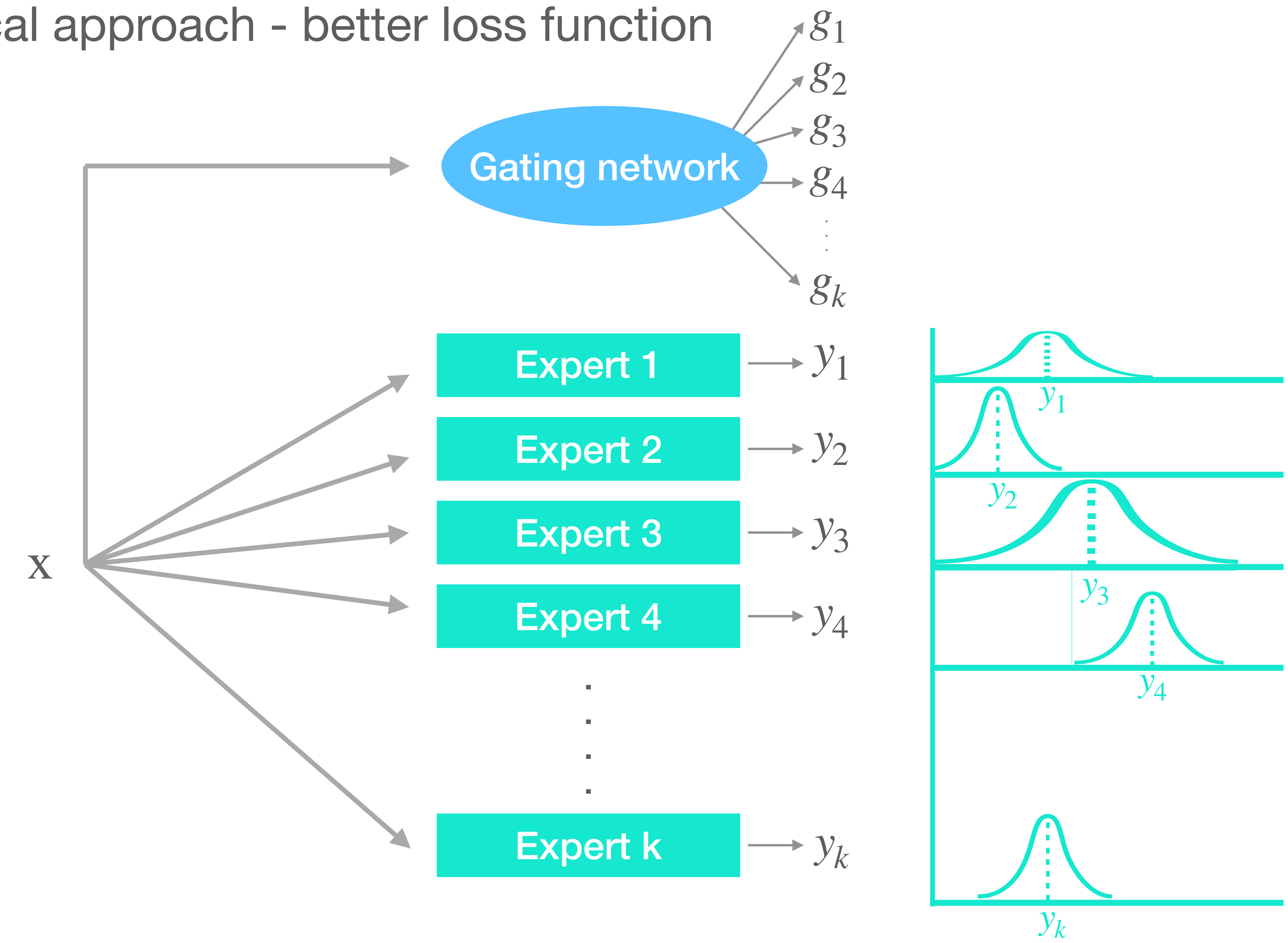
$$y_i = \theta_{\mathbf{i}}^{\mathbf{T}}\mathbf{x}$$

Signal for training each expert $\dfrac{\partial L}{\partial \theta_i} = \dfrac{\partial L}{\partial y_i}\dfrac{\partial y_i}{\partial \theta_i} \propto g_i(y - y_i)\dfrac{\partial y_i}{\partial \theta_i}$

Signal for training the gating net $\dfrac{\partial L}{\partial \eta_i} \propto g_i[(y - y_i)^2 - L]$

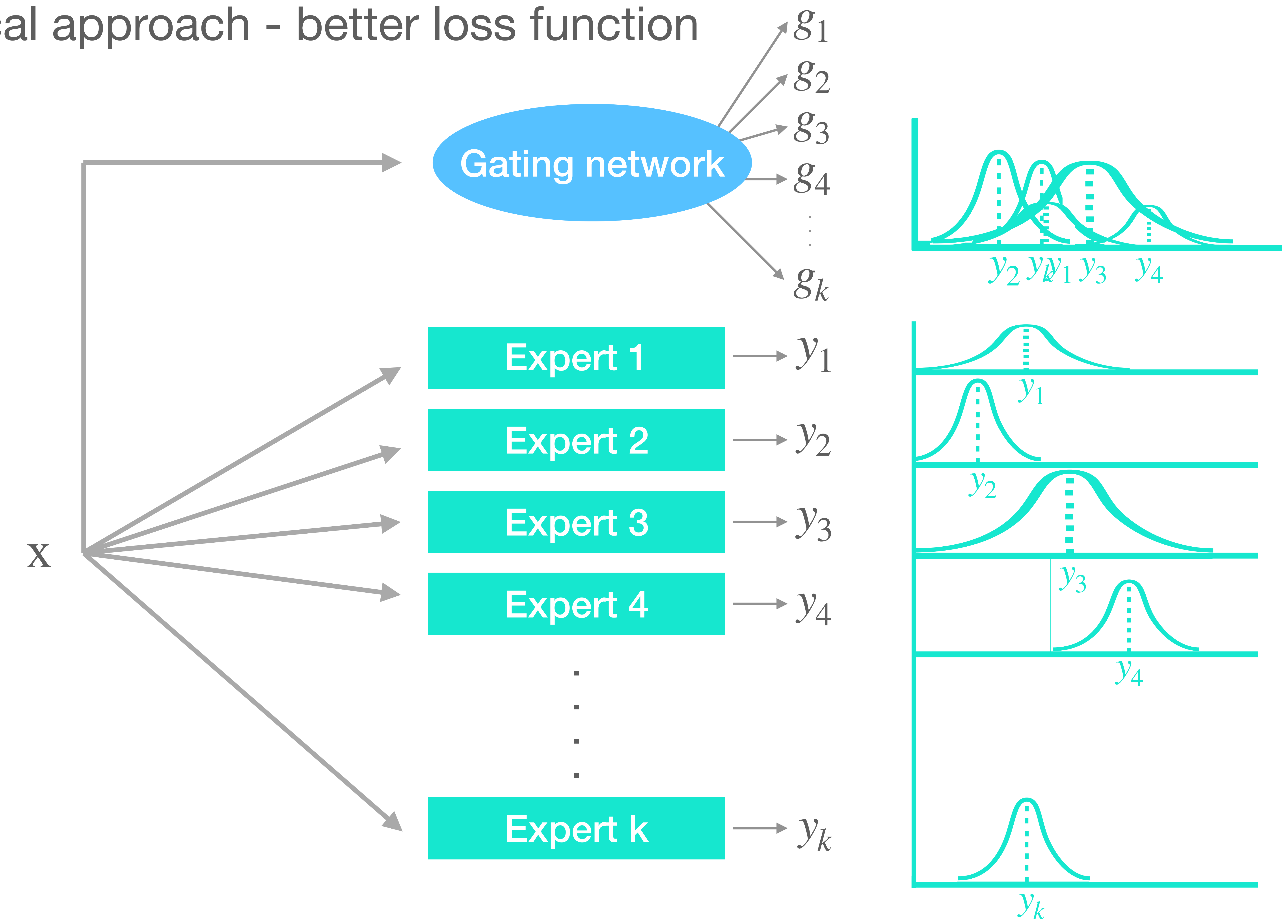
Mixture of Experts

Making more assumptions about how the data came to be, we get a likelihood function that gives a statistical approach - better loss function



Mixture of Experts

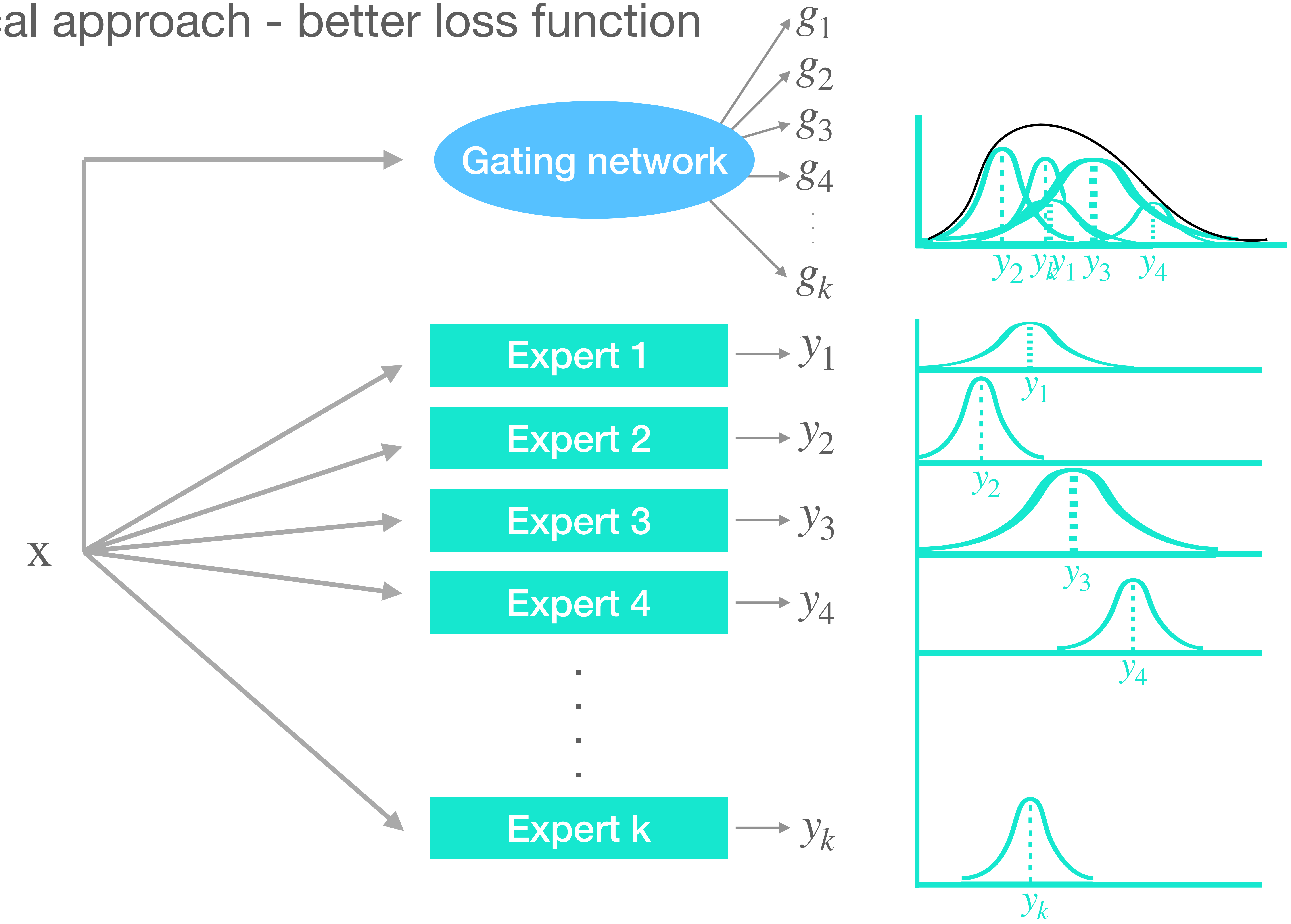
Making more assumptions about how the data came to be, we get a likelihood function that gives a statistical approach - better loss function



$$\hat{y} = \sum_i g_i y_i$$

Mixture of Experts

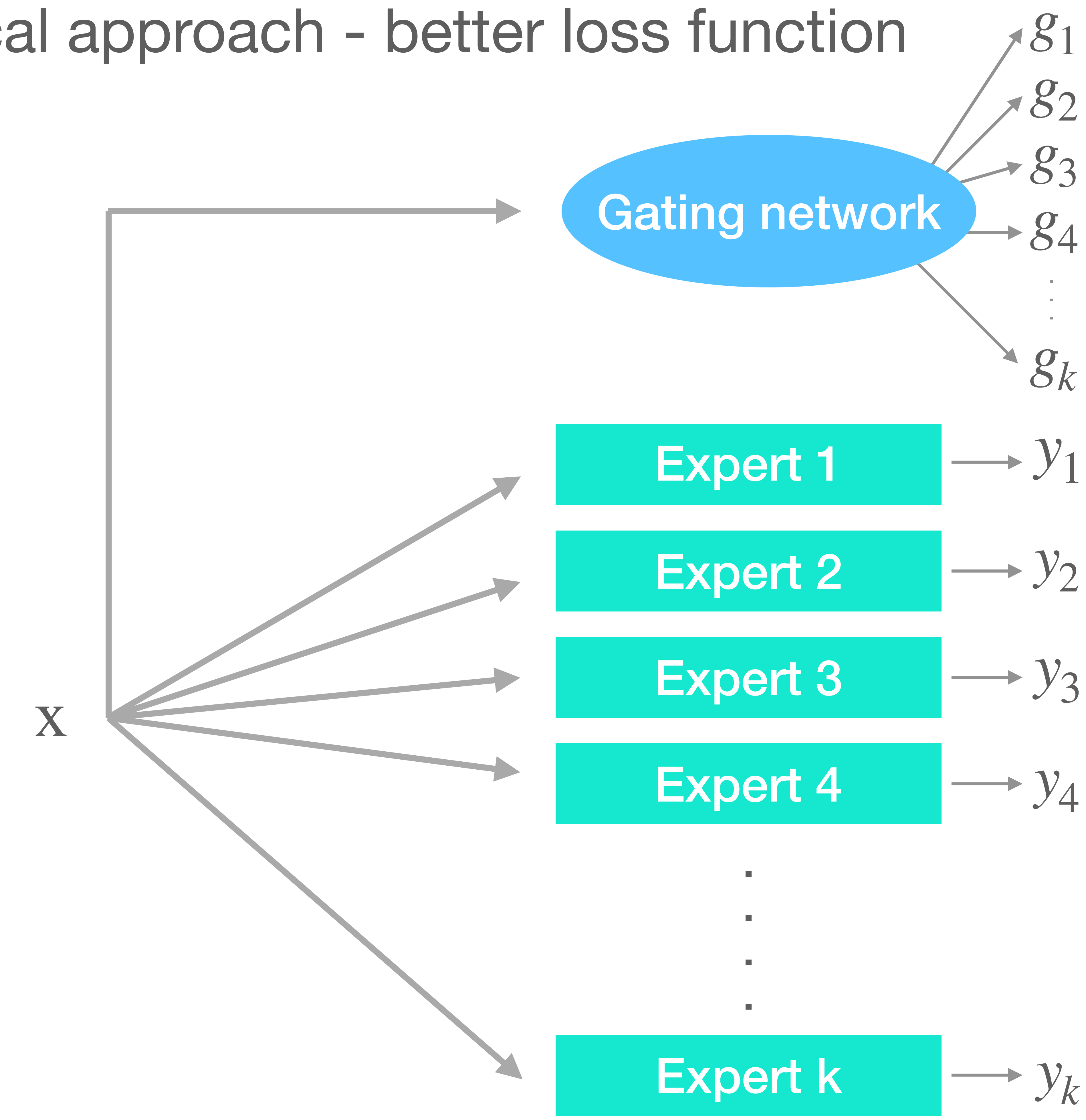
Making more assumptions about how the data came to be, we get a likelihood function that gives a statistical approach - better loss function

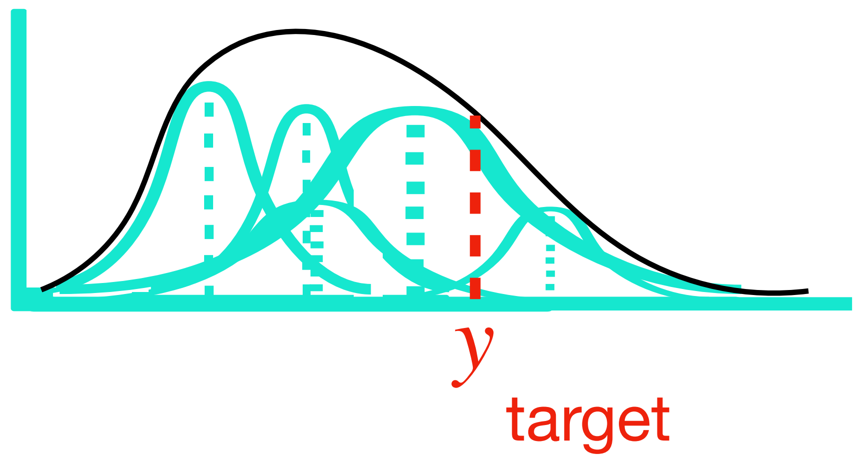


$$\hat{y} = \sum_i g_i y_i$$

Mixture of Experts

Making more assumptions about how the data came to be, we get a likelihood function that gives a statistical approach - better loss function

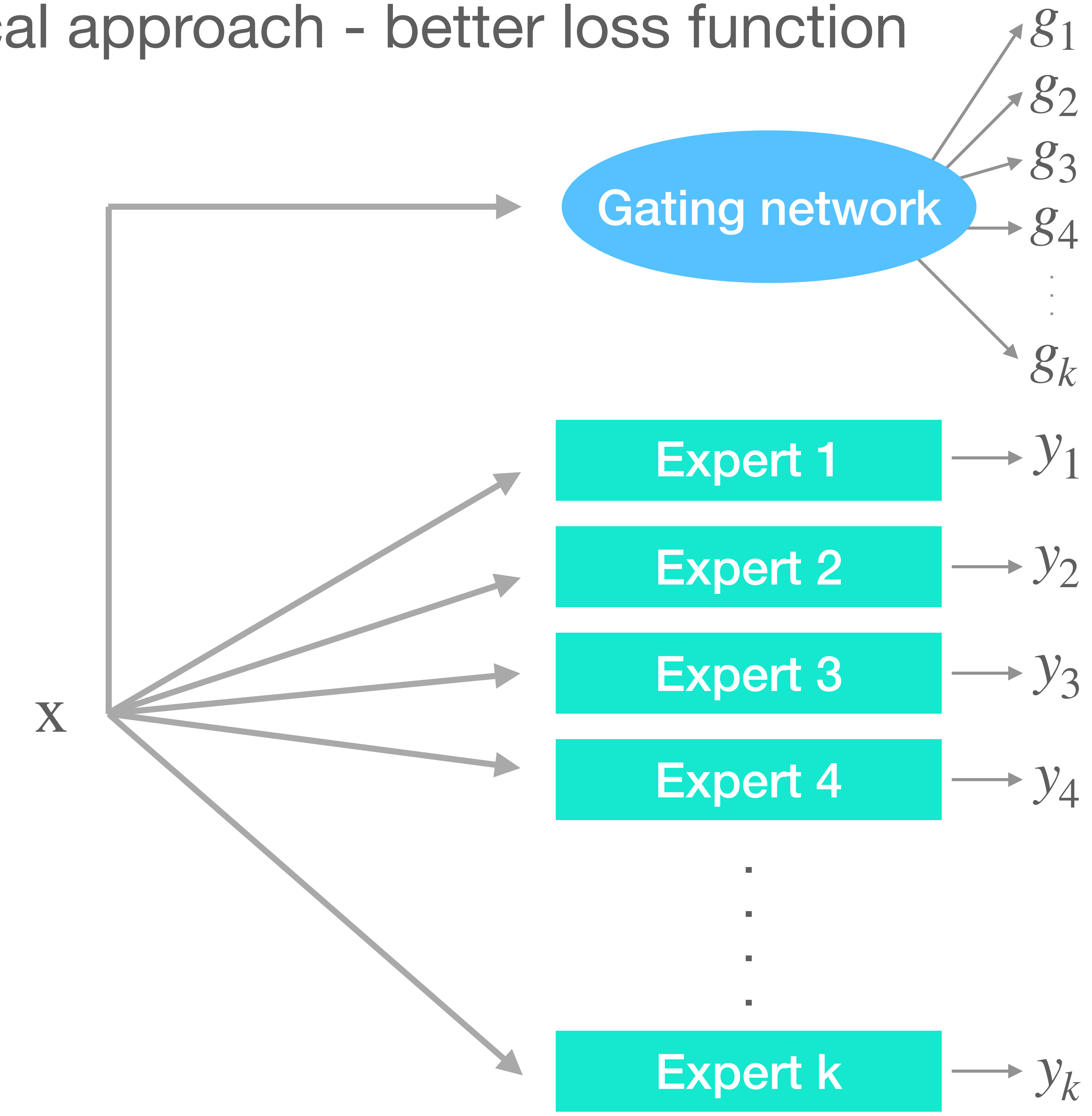


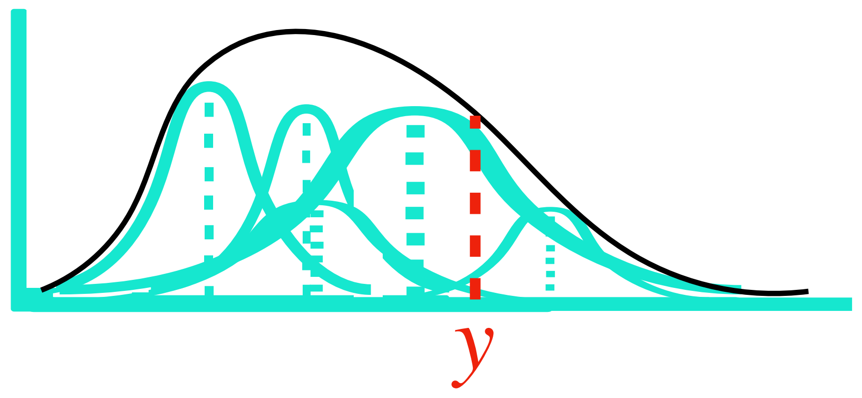


$$\hat{y} = \sum_i g_i y_i$$

Mixture of Experts

Making more assumptions about how the data came to be, we get a likelihood function that gives a statistical approach - better loss function



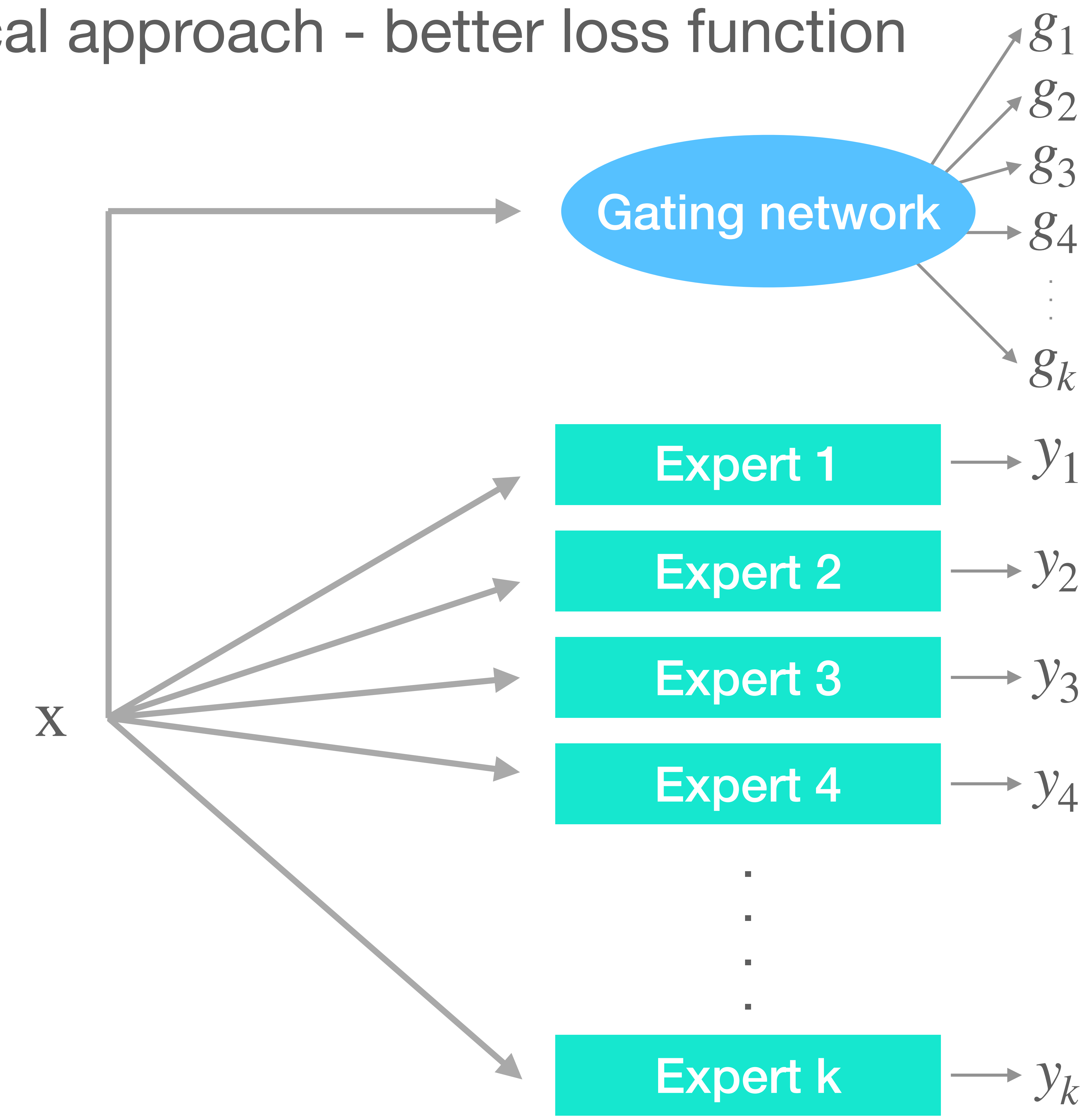


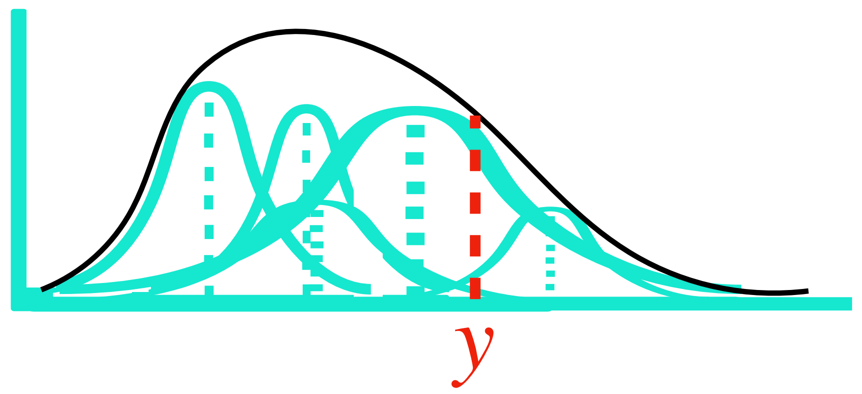
$$\hat{y} = \sum_i g_i y_i$$

$$p(y\,|\,MoE) = \sum_i g_i \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}(y-y_i)^2}$$

Mixture of Experts

Making more assumptions about how the data came to be, we get a likelihood function that gives a statistical approach - better loss function





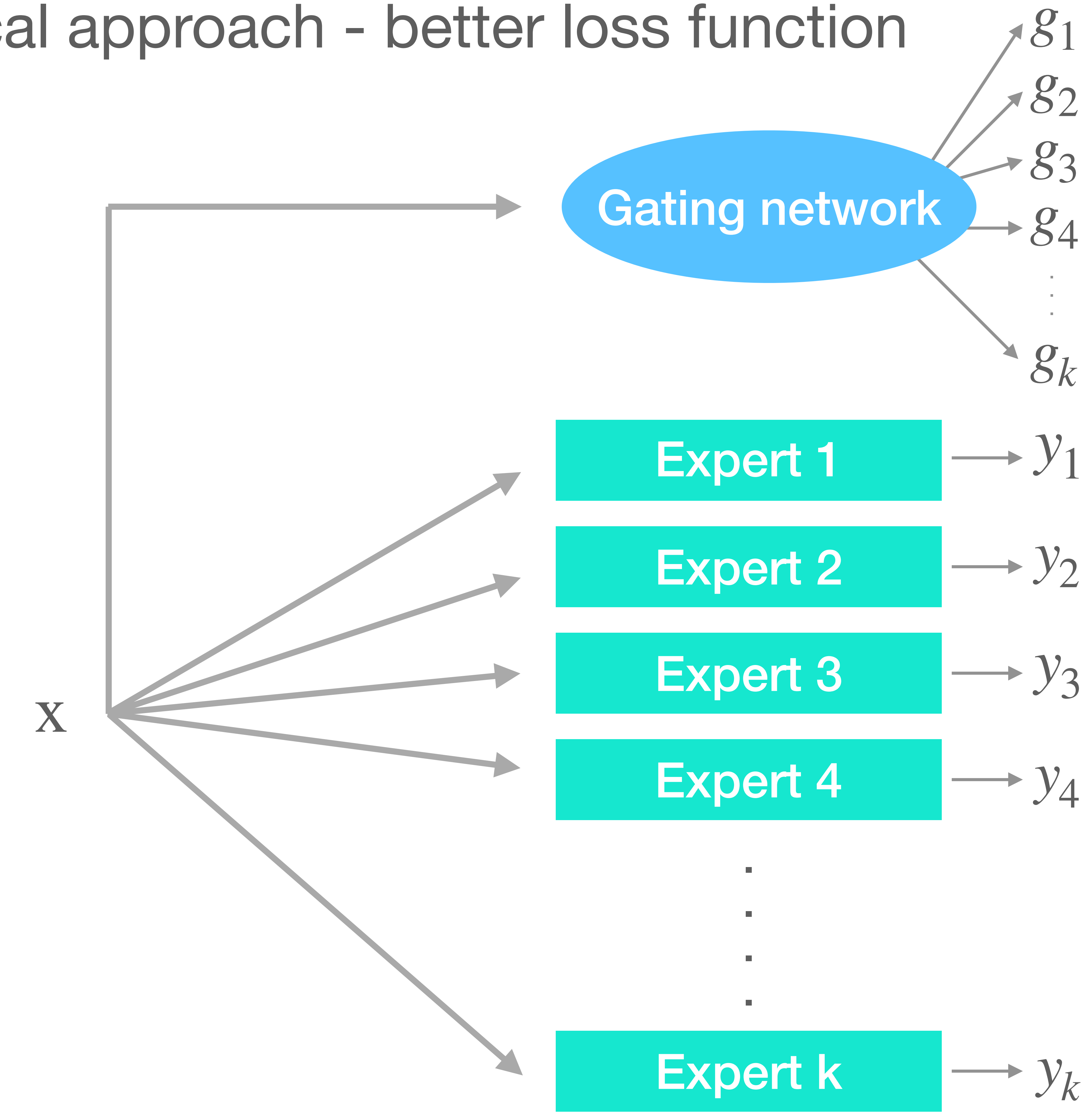
$$\hat{y} = \sum_i g_i y_i$$

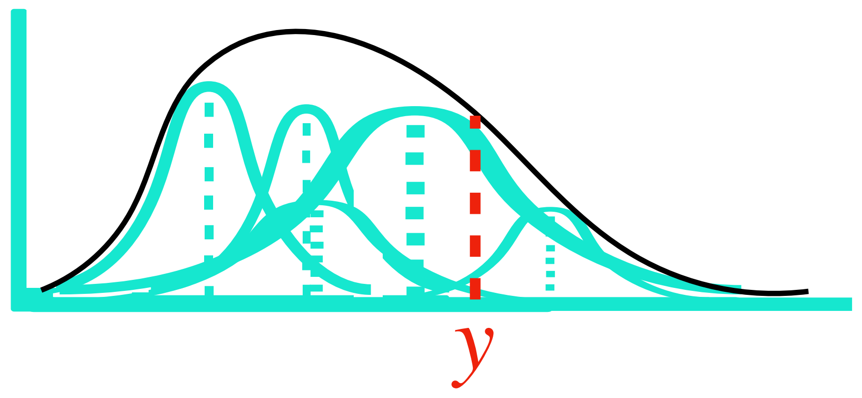
$$p(y \mid MoE) = \sum_i g_i \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}(y - y_i)^2}$$

$$L = -\,log(p(y \mid MoE))$$

Mixture of Experts

Making more assumptions about how the data came to be, we get a likelihood function that gives a statistical approach - better loss function





$$\hat{y} = \sum_i g_i y_i$$

$$p(y\,|\,MoE) = \sum_i g_i \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}(y-y_i)^2}$$

Linear experts

$$y_i = \theta_{\mathbf{i}}^{\mathbf{T}}\mathbf{x}$$

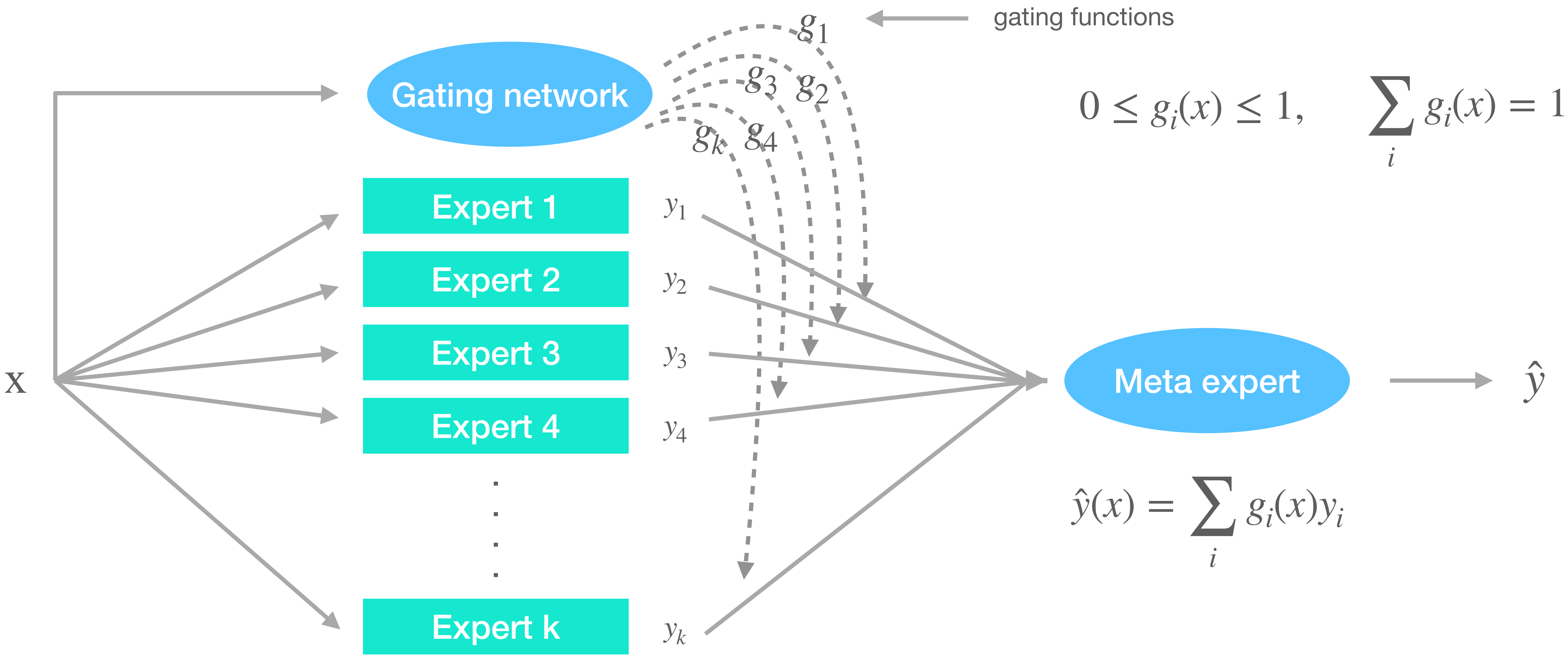
$$L = -\,log(p(y\,|\,MoE))$$

Softmax gating network

$$g_i(\mathbf{x}) = \frac{\mathbf{e}^{\eta_{\mathbf{i}}^{\mathbf{T}}\mathbf{x}}}{\sum_{\mathbf{j=1}}^{\mathbf{k}} \mathbf{e}^{\eta_{\mathbf{j}}^{\mathbf{T}}\mathbf{x}}}$$

Mixture of Experts

Uses different learners or combination of experts for different regions of the input data.



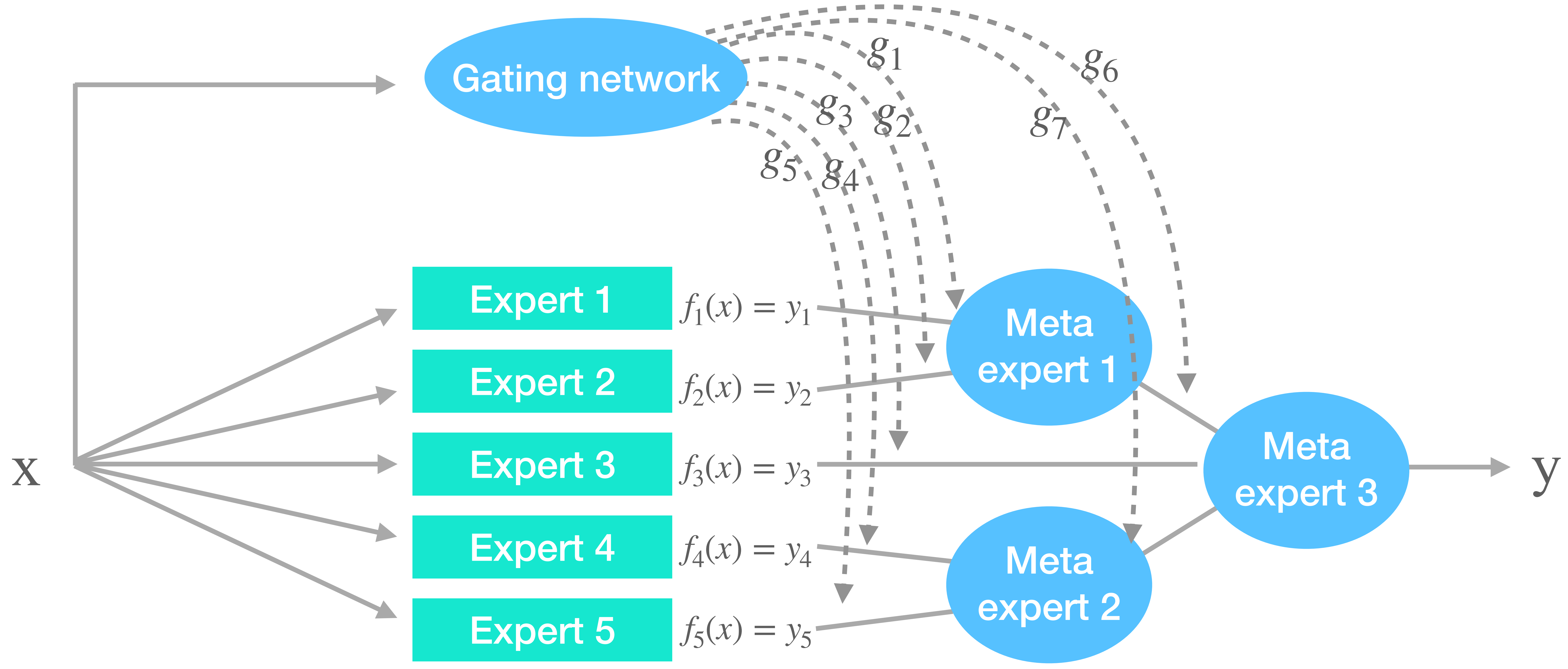
$$0 \leq g_i(x) \leq 1, \qquad \sum_i g_i(x) = 1$$

$$\hat{y}(x) = \sum_i g_i(x) y_i$$

The models are mixed using a gating network that decides what combination of experts to use

Hierarchical Mixture of Experts

If the output is conditioned on multiple levels of probabilistic gating functions, the mixture is called a hierarchical mixture of experts



Summary

Ensamble Models- cooperation:

Simple:

- **Voting**: simple and weighted
- **Averaging**: simple and weighted

Less simple:

- **Bagging**: independent, parallel, homogeneous weak learners, combined with some deterministic averaging process - less variance
- **Boosting**: sequential, homogeneous weak learners, combined in a deterministic, adaptive way (a base model depends on the previous ones) - less bias
- **Blending**: independent, parallel, heterogeneous weak learners, by training a meta-model to output a prediction - less bias
- **Stacking**: same as blending but k-folding the training data - less bias

Mixture of Experts - specialization:

For data that was generated with different models, or whose description depends on the input-output regime. Heterogeneous models, trained in different regions of the data, combined by a gating network that decides the probability that a given input is best described by a certain expert.

Questions?