

# Stacking for supervised learning

Niall Rooney,  
NIKEL, University of Ulster

# Ensemble learning

- 1 Postulate multiple hypotheses to explain the data
- 1 Shortcomings of single model learning algorithms (Dietterich , 2002)
  - Statistical problem
  - Computational problem
  - Representational problem

# Ensemble learning

- 1 Generalization Error: Bias + Variance
  - **Bias**: how close the algorithm's average prediction is close to the target
  - **Variance** : how much the algorithm's predictions "bounces round" for different training sets
  - a model which is too simple, or too inflexible, will have a large bias
  - a model which has too much flexibility will have high variance

# Ensemble learning

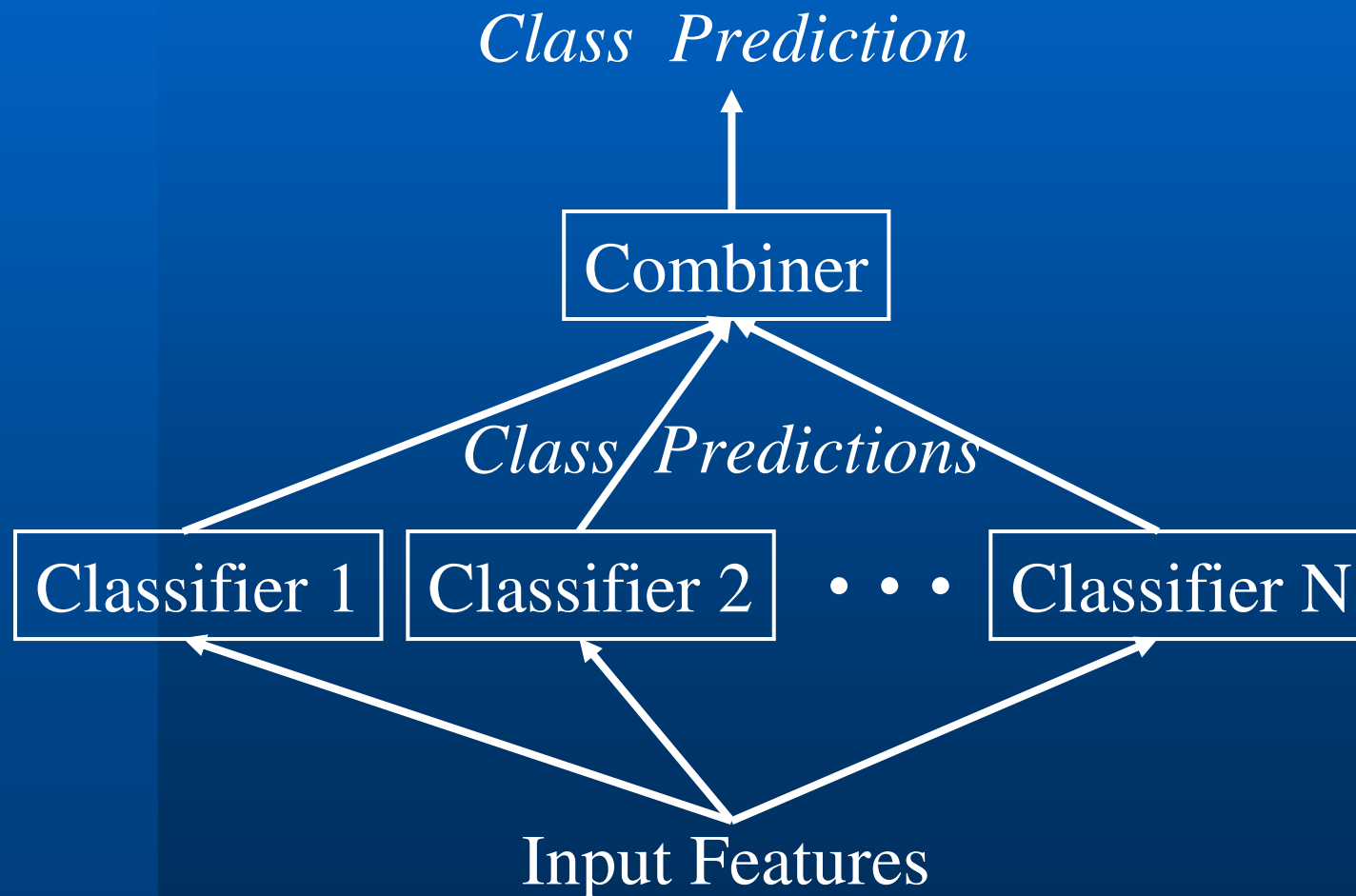
## 1 Generalization Error: Ensembles

- Ensembles reduce bias and/or variance
- Ensembles to be effective – need diverse and accurate base models
- Diversity measured by level of variability in base members predictions (for regression)

# Ensemble learning

- § Homogeneous learning
  - **data sampling, feature sampling, randomization, parameter settings**
- § Heterogeneous learning
  - **Same data, different learning algorithms**

# Ensemble Learning



# Ensemble learning

## *Methods of combination:*

Voting, Weighting, Selection

Mixture of experts

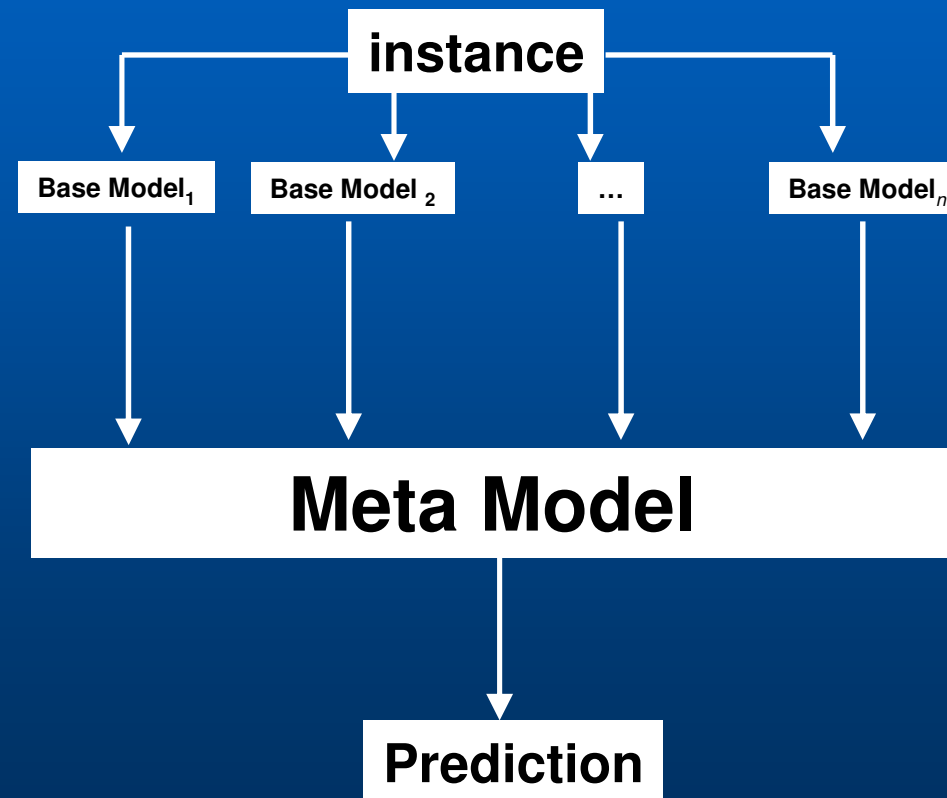
Error-correcting output codes

Bagging

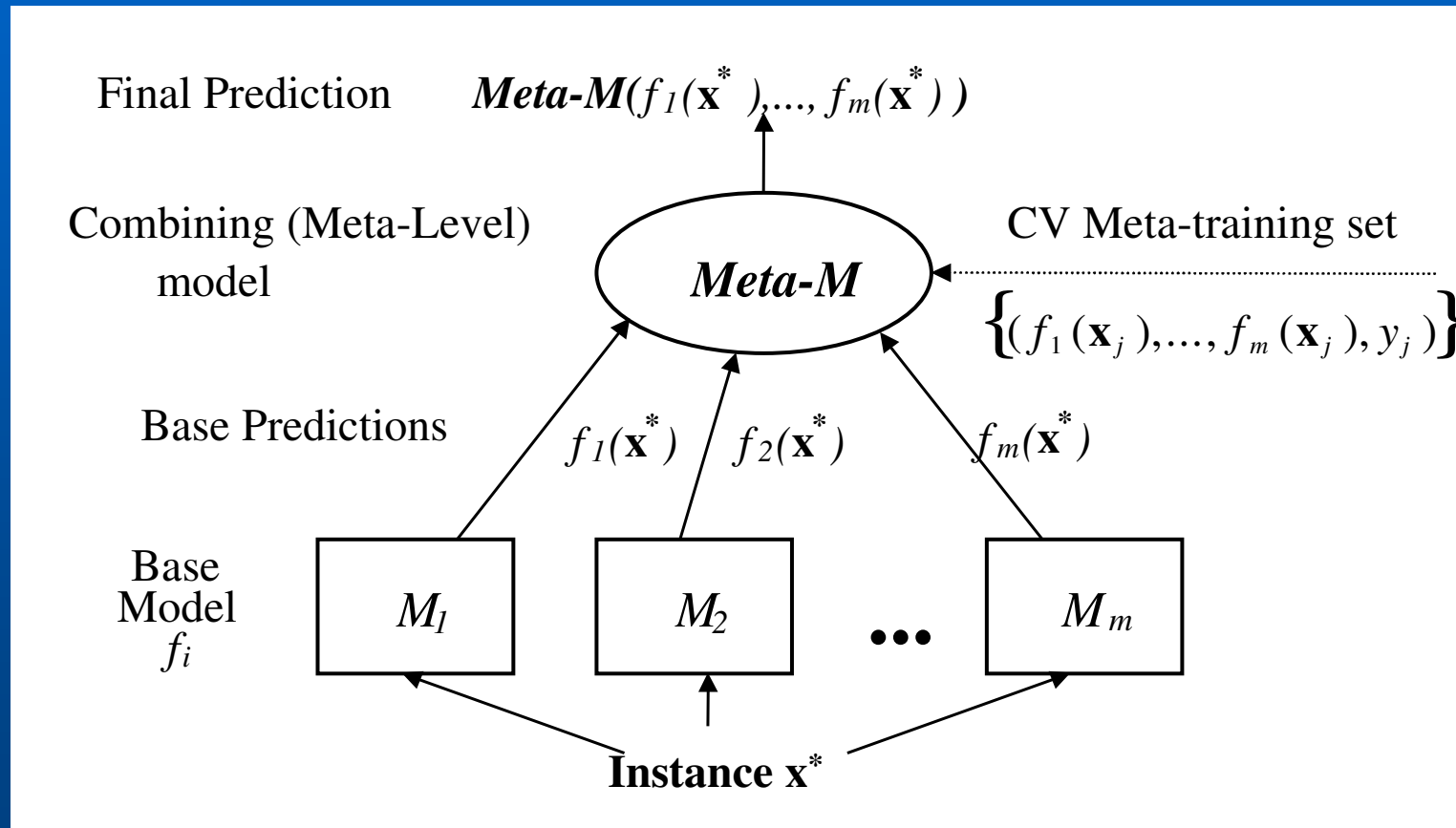
Boosting

Stacking

# Ensemble Learning: Stacking



# Meta Technique: SR



# Stacking for classification

- § Use class distributions from base classifiers rather than class predictions

$$\{(P_1(C_1 | x), \dots, P_1(C_k | x), \dots, P_m(C_1 | x), \dots, P_m(C_k | x), y)\}$$

- § Choice of Meta-classifier:

Multi-response linear regression

- For a classification with  $m$  class values,  $m$  regression problems
- Only use probabilities related to class  $C_j$  to predict class  $C_j$

# Stacking for classification

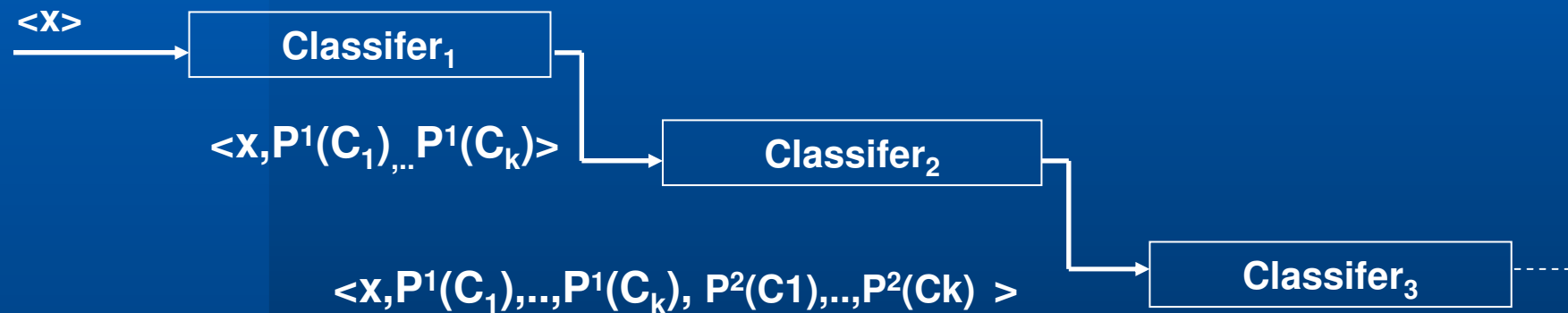
- § Different “type” of base classifiers
- § Multi-response model trees used to guarantee better performance than Selecting best classifier

# Stacking for regression

- § Linear regression requires non-negative weights
- § Model trees meta-learner
- § Homogeneous Stacking using random feature sub-sets
- § Feature sub-sets can be improved upon using hill-climbing or GA techniques

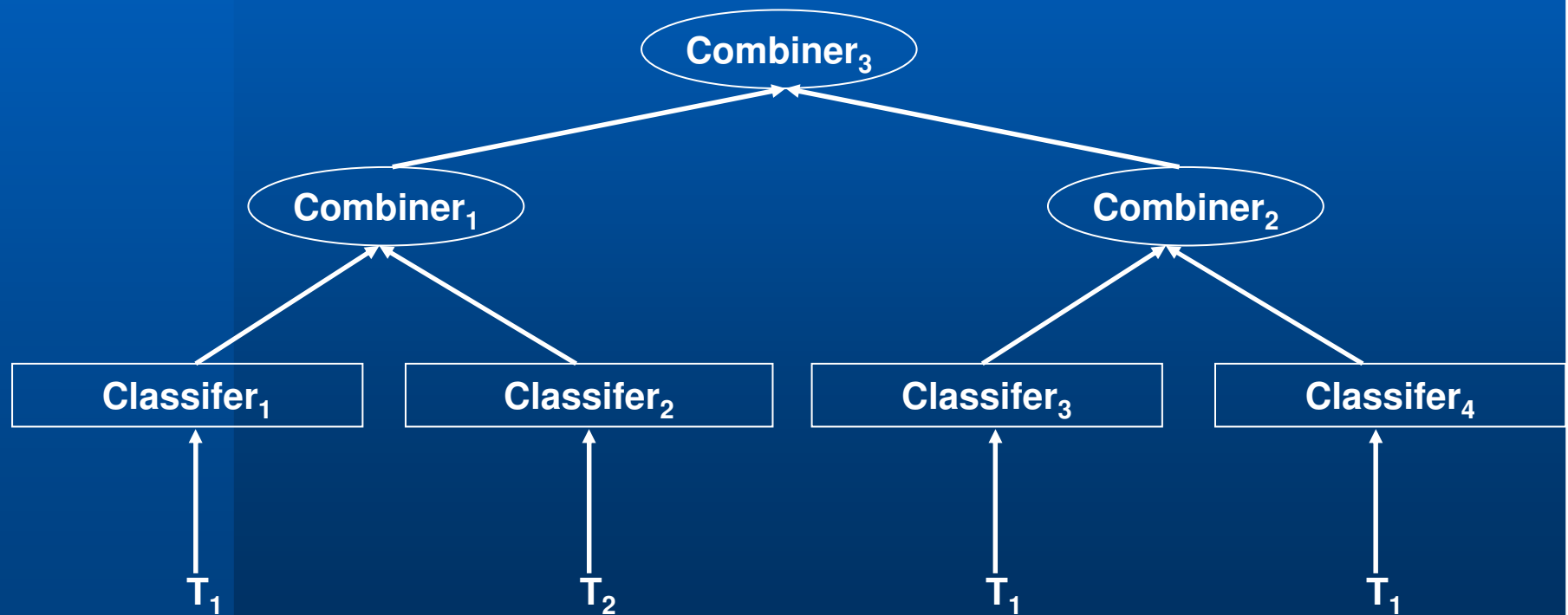
# Related techniques: Multiple meta-levels

## Cascade Generalization



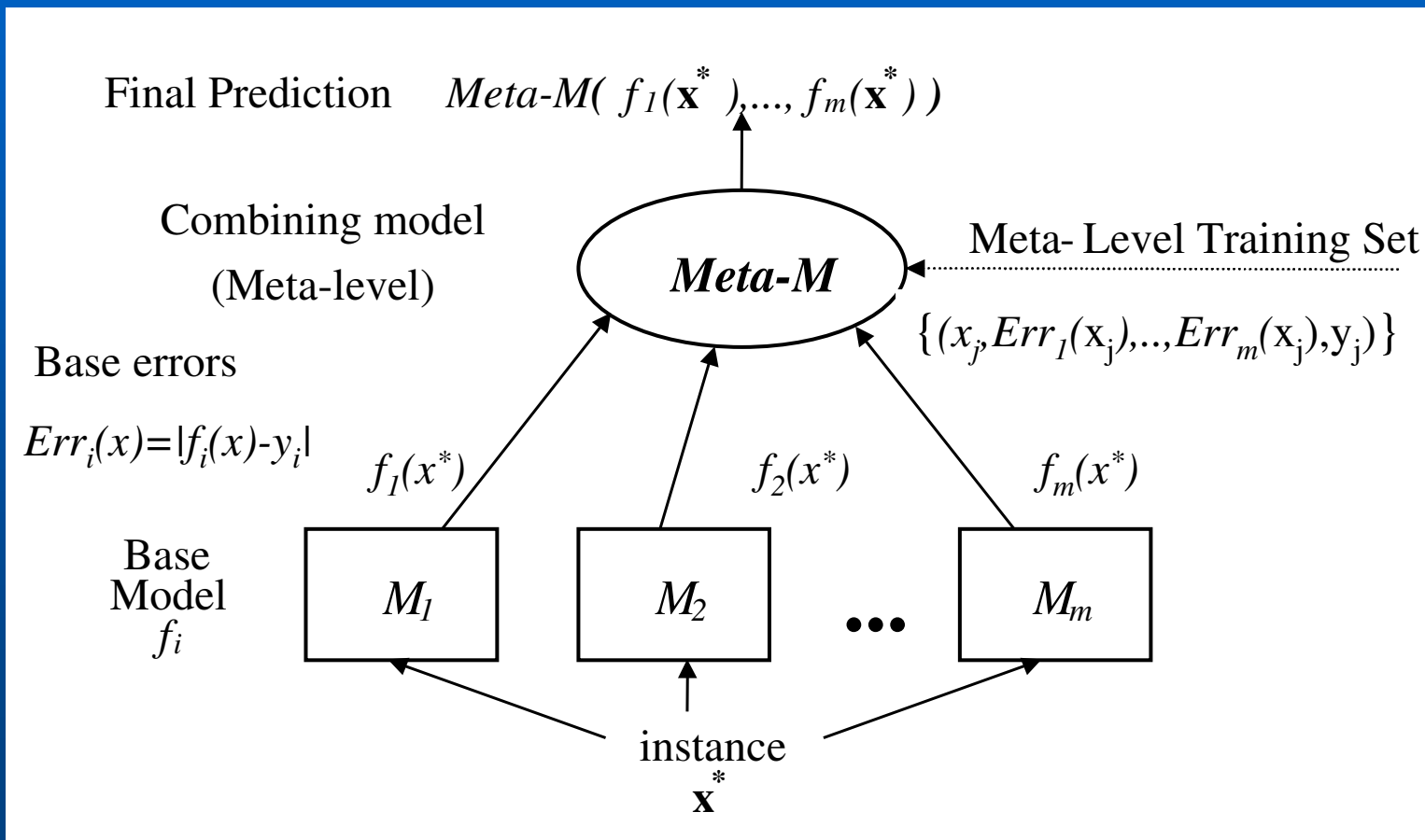
# Related techniques: Multiple meta-levels

## Combiner Trees



Disjoint training sets

# Related Techniques: Dynamic Integration



# Dynamic Integration

*Meta-M*

**Meta Model - distance weighted k-NN**

- 1 NN – set of k nearest meta-instances
- 1 For each member find cumulative error of each model

# Dynamic Integration

- 1 Dynamic Selection (DS)
  - choose the model with lowest cumulative error
- 1 Dynamic Weighting (DW)
  - combine the models with weights based on their cumulative error
- 1 Dynamic Weighting with Selection (DWS)
  - combine the models as DW but exclude models if they have larger than median cumulative error

# Applications

- 1 Distributed data mining
- 1 Intrusion detection
- 1 Concept drift

# Key papers

- 1 Wolpert, D. H.: Stacked Generalization. *Neural Networks*, 5 (1992) 241-259
- 1 Breiman, L.: Stacked Regressions. *Machine Learning*, 24 (1996) 49-64
- 1 Dietterich, T. G.: Ensemble Methods in Machine Learning. *Lecture Notes in Computer Science*, 1857 (2000) 1-15
- 1 Dzeroski, S., & Zenko, B.: Is Combining Classifiers with Stacking Better than Selecting the Best One? *Machine Learning*, 54 (2004) 255-273
- 1 Ting, K. M., & Witten, I. H.: Issues in Stacked Generalization. *Journal of Artificial Intelligence Research*, 10 (1999) 271-289