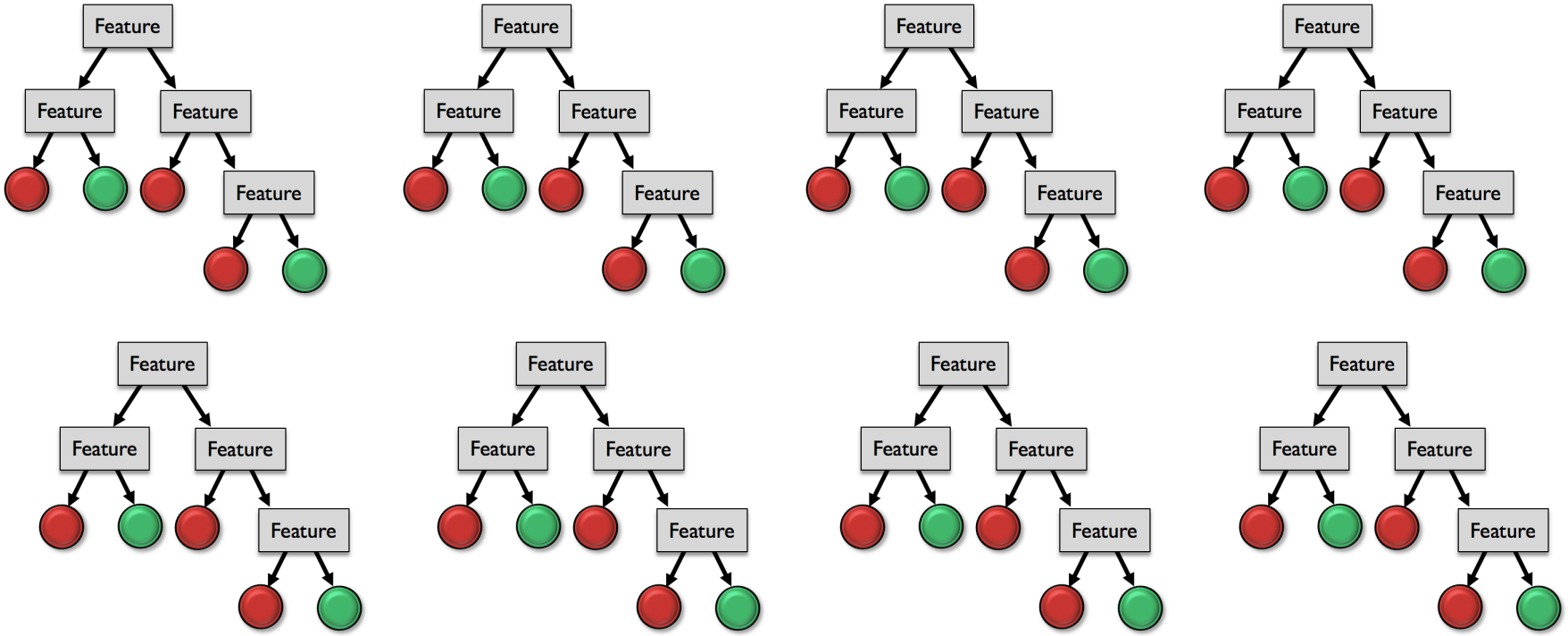
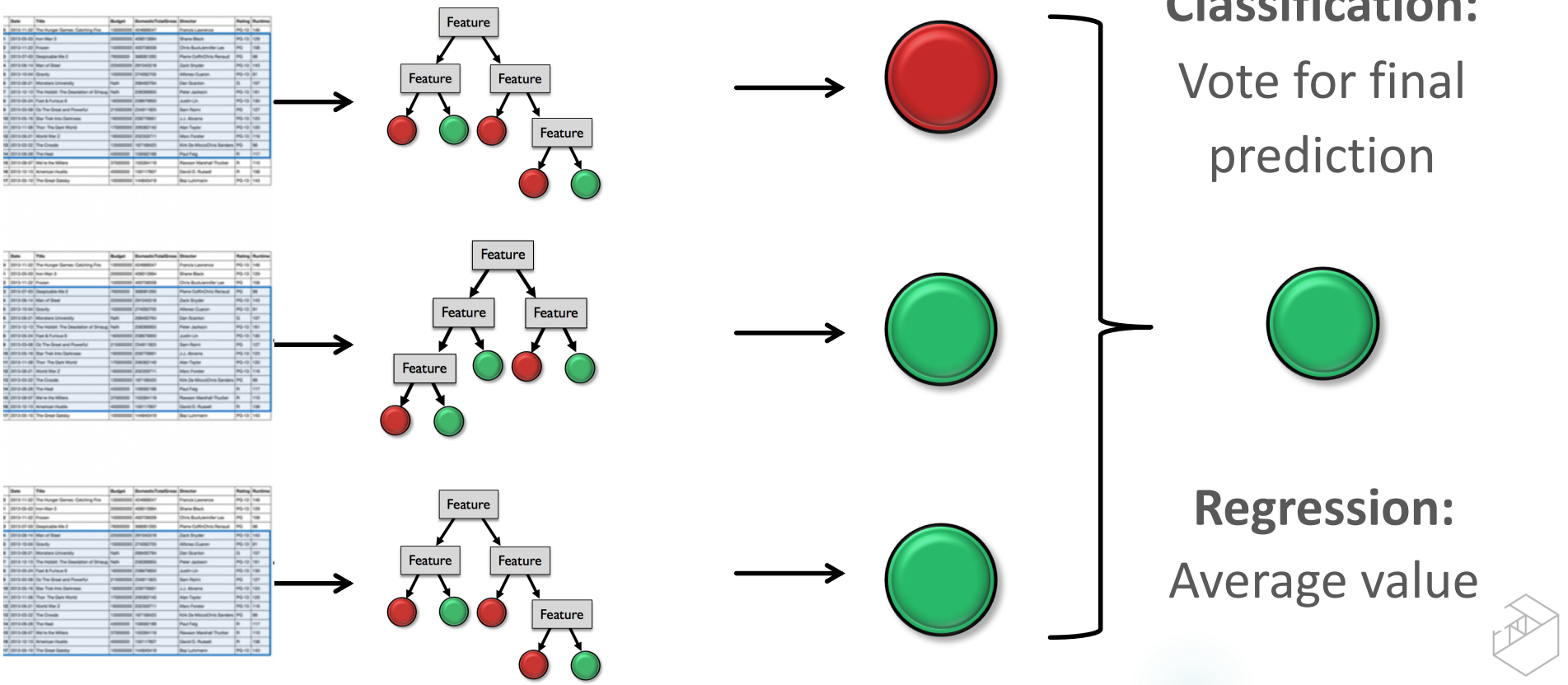
Ensembling

Bagging

Overcome Variance Issue by Ensembling

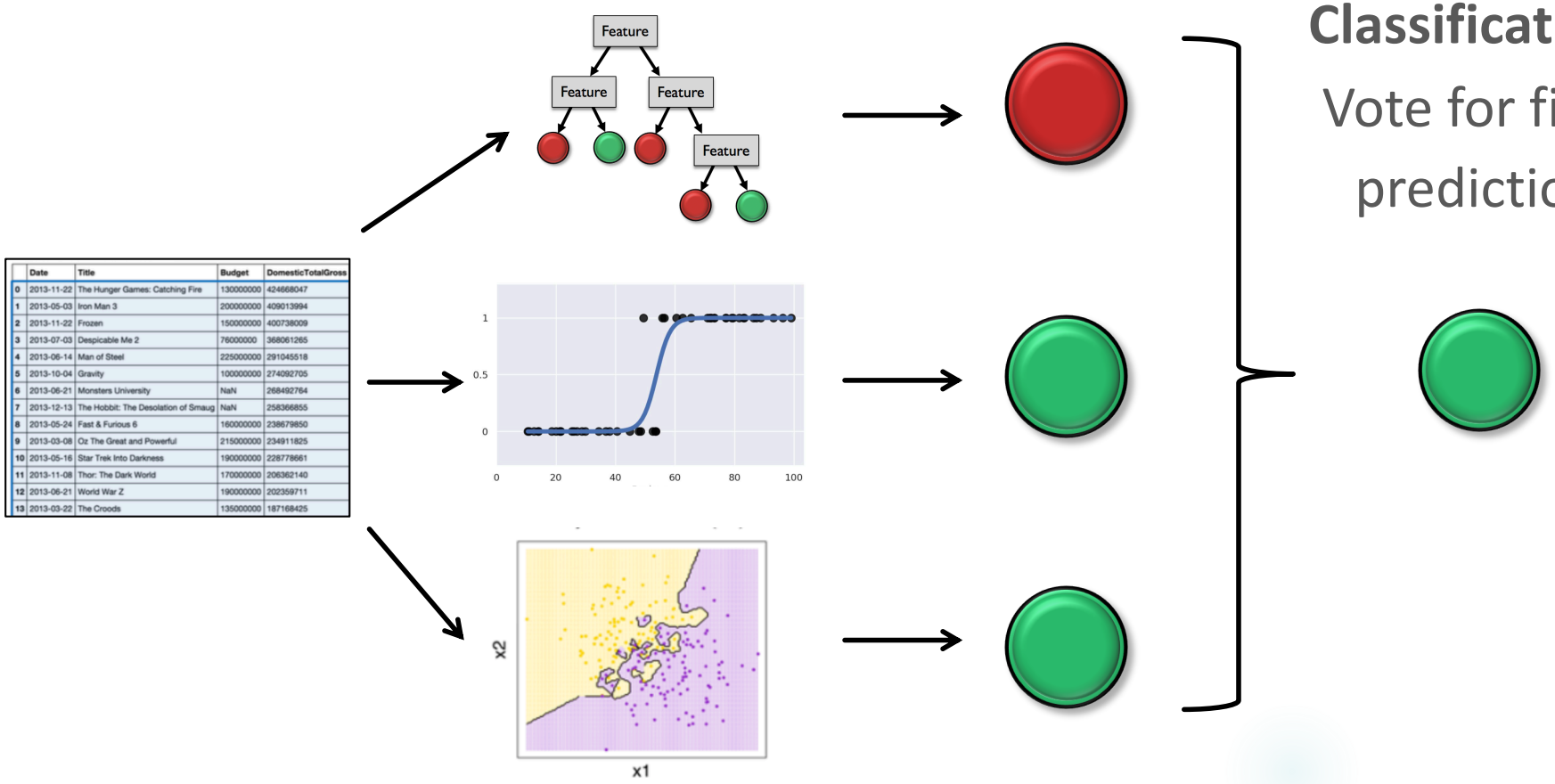


Bagging


Feature
Feature
Feature
Feature
Feature
Feature
Feature
Feature
Feature
Feature
Feature
Feature
Classification:
Vote for final prediction
Regression:
Average value

Voting Classifier

Voting Classifier



	Date	Title	Budget	DomesticTotalGross
0	2013-11-22	The Hunger Games: Catching Fire	130000000	424668047
1	2013-05-03	Iron Man 3	200000000	409013994
2	2013-11-22	Frozen	150000000	400738009
3	2013-07-03	Despicable Me 2	76000000	368061265
4	2013-06-14	Man of Steel	225000000	291045518
5	2013-10-04	Gravity	100000000	274092705
6	2013-06-21	Monsters University	NaN	268492764
7	2013-12-13	The Hobbit: The Desolation of Smaug	NaN	258366855
8	2013-05-24	Fast & Furious 6	160000000	238679850
9	2013-03-08	Oz The Great and Powerful	215000000	234911825
10	2013-05-16	Star Trek Into Darkness	190000000	228778661
11	2013-11-08	Thor: The Dark World	170000000	206362140
12	2013-06-21	World War Z	190000000	202359711
13	2013-03-22	The Croods	135000000	187168425

How to vote?

- **Max voting**
  - Majority class from hard predictions
- **Average voting**
  - Average soft probability predictions for each class
  - Pick class with highest probability
- **Weighted voting**
  - Weigh model soft (or hard) predictions before averaging

How to weigh each model?

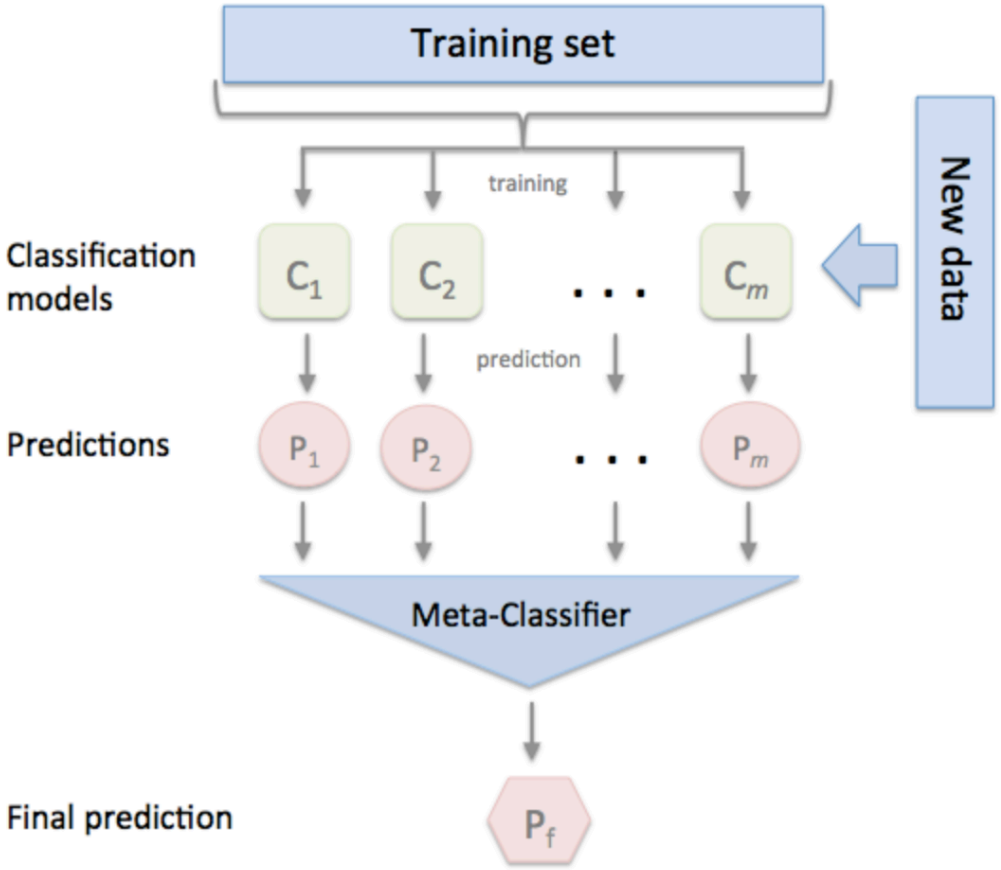
- Difficult task in general
- Tune hyperparameter weights via validation/cross-validation
- Use a stacked classifier

Stacking

What is stacking?

- Ensemble learning technique
- Meta-classifier

  A classifier for classifiers!



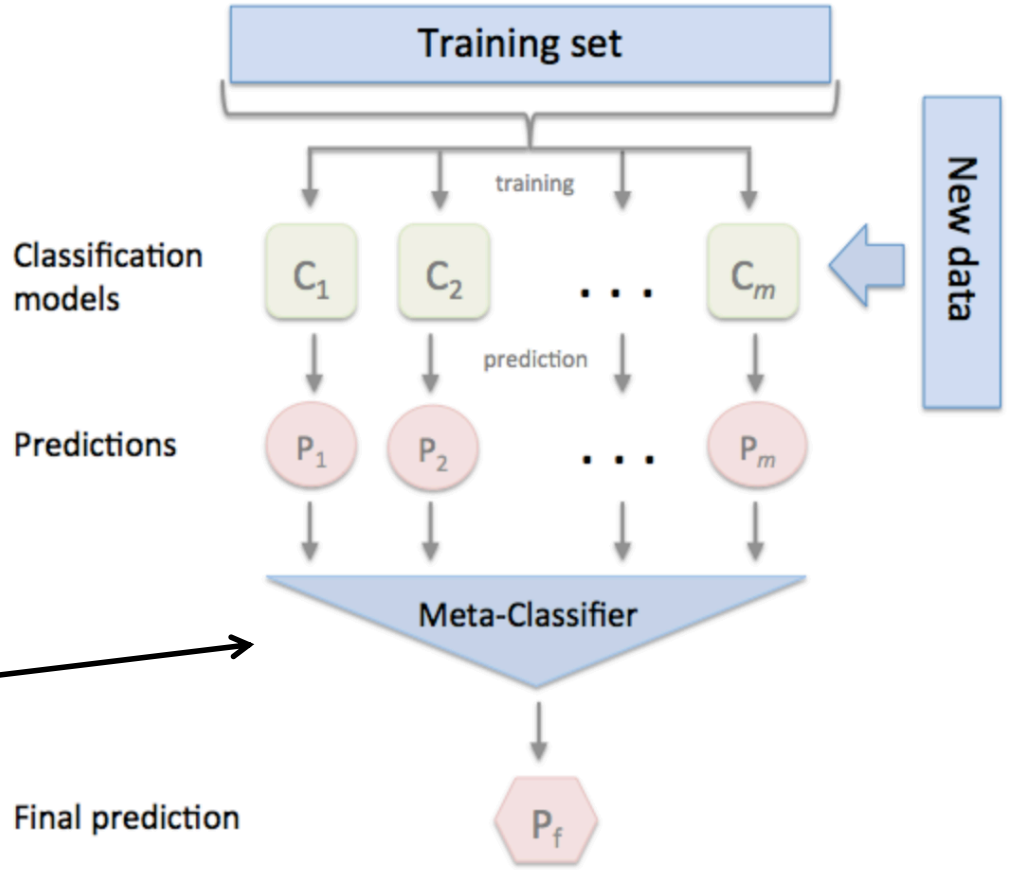
What is stacking?

Example:

Various models

(trees, logistic regression, kNN, etc.)

Logistic regression



Ensembling

ADVANTAGES

- Reduces variance
- Learn different data aspects from each model type
- Typically, better predictive power and overall model performance

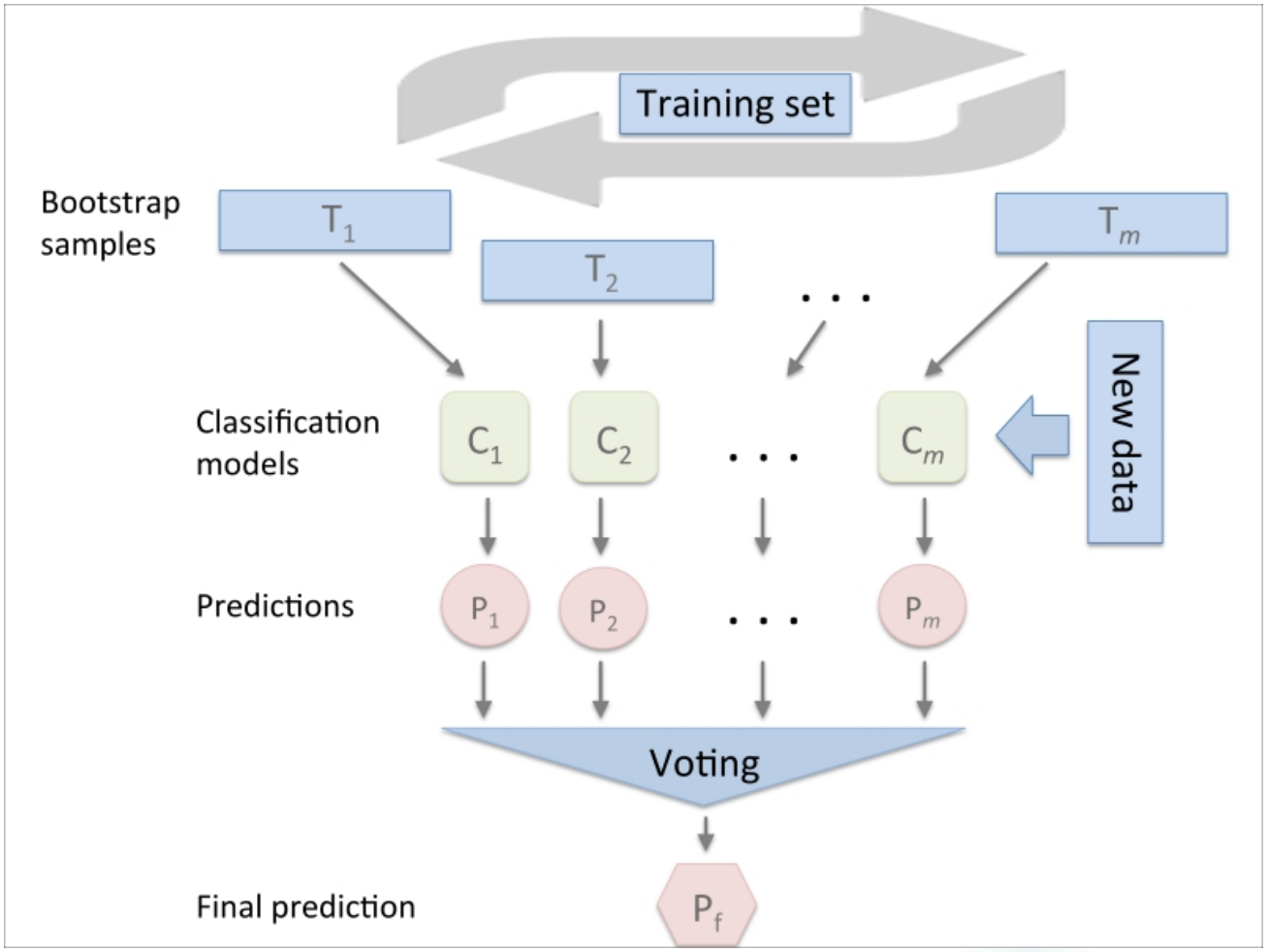
DISADVANTAGES

- Loss of interpretability
- Computationally expensive

Questions?

Appendix

Bagging


Training set
Bootstrap samples
T_1
T_2
T_m
Classification models
C_1
C_2
C_m
New data
Predictions
P_1
P_2
P_m
Voting
Final prediction
P_f

Voting Classifier