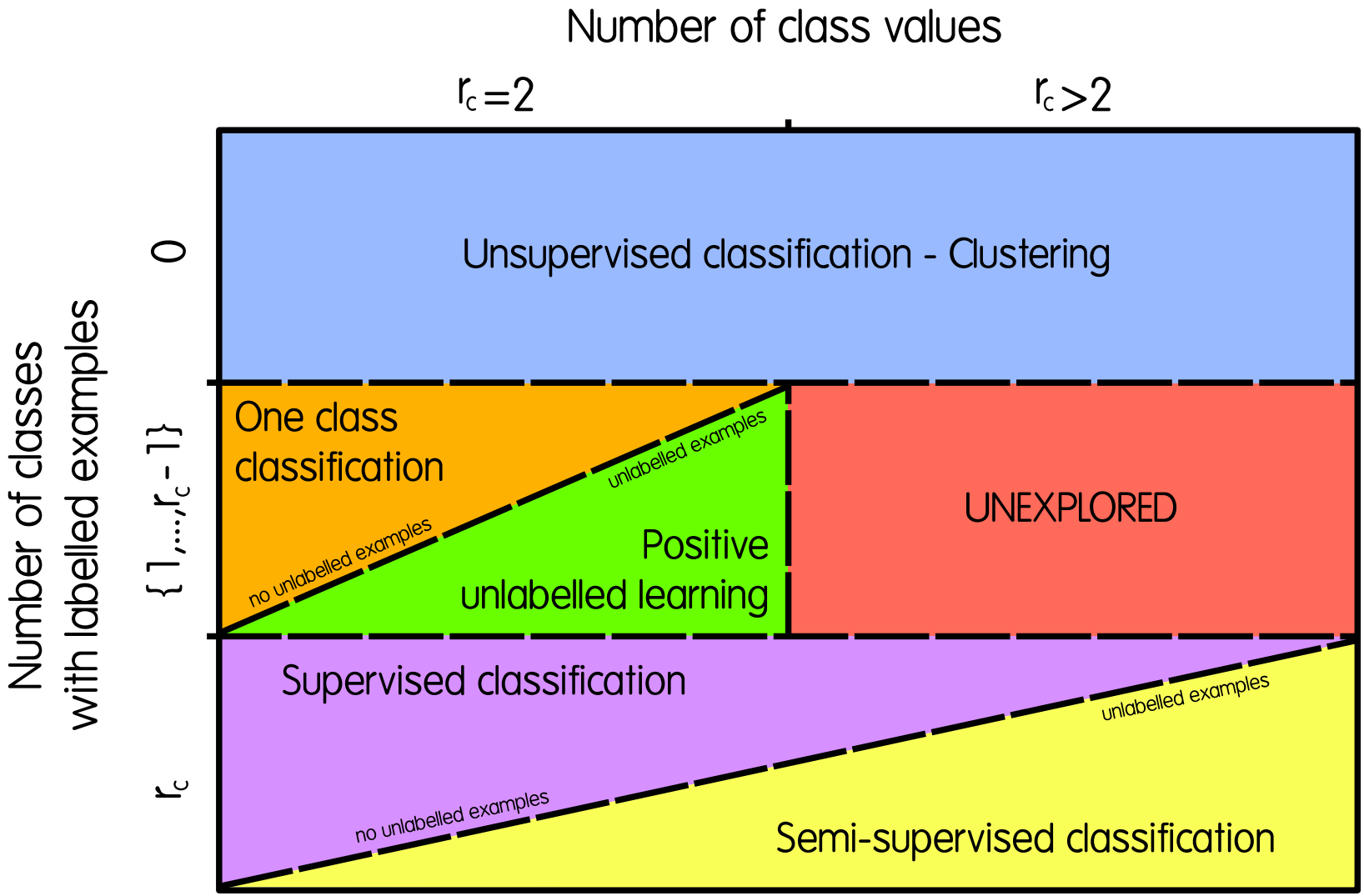
Pattern Recognition Performance Assessment

Classification general taxonomy



Data for supervised classification

Turn the phenotype knowledge to one's advantage

	X_1	X_2	$\ldots$	X_i	$\ldots$	X_n	C
$(\boldsymbol{x}^{(1)}, c^{(1)})$	x_1^1	x_2^1	$\ldots$	x_i^1	$\ldots$	x_n^1	c^1
$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\ddots$	$\vdots$	$\vdots$
$(\boldsymbol{x}^{(j)}, c^{(j)})$	x_1^j	x_2^j	$\ldots$	x_i^j	$\ldots$	x_n^j	c^j
$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\ddots$	$\vdots$	$\vdots$
$(\boldsymbol{x}^{(N)}, c^{(N)})$	x_1^N	x_2^N	$\ldots$	x_i^N	$\ldots$	x_n^N	c^N
$\boldsymbol{x}^{(N+1)}$	$x_1^{(N+1)}$	$x_2^{(N+1)}$	$\ldots$	$x_i^{(N+1)}$	$\ldots$	$x_n^{(N+1)}$	???

- The class constitutes a kind of metaknowledge of high usability
- A classifier is a function that maps instances with classes

$$\gamma : (x_1, \ldots, x_n) \rightarrow \{1, 2, \ldots, m\}$$

Classification paradigms

To enumerate some...

- Lazy family: *k* nearest neighbours
- Functions: Linear Discriminant Analysis, Regression, SVMs
- Bayesian: Naïve Bayes, TAN, FAN, *k*-DB, Bayesian Networks
- Trees: ID3, C4.5, M5
- ... more and more

Evaluation indices

General indices

Accuracy, Brier score, Cross-entropy error

Discrimination

Sensitivity, Specificity, Positive predicitive value PPV, Negative predicitive value NPV, ROC curve, Area under curve AUC, Matthews correlation coefficient MCC

Calibration

Calibration curves, Hosmer and Lemeshow goodness-of-fit

Measuring the performance of a classifier

Confusion matrix

		C True class +	C True class -
C_M Predicted class	+	a	b
	-	c	d

Figures of merit

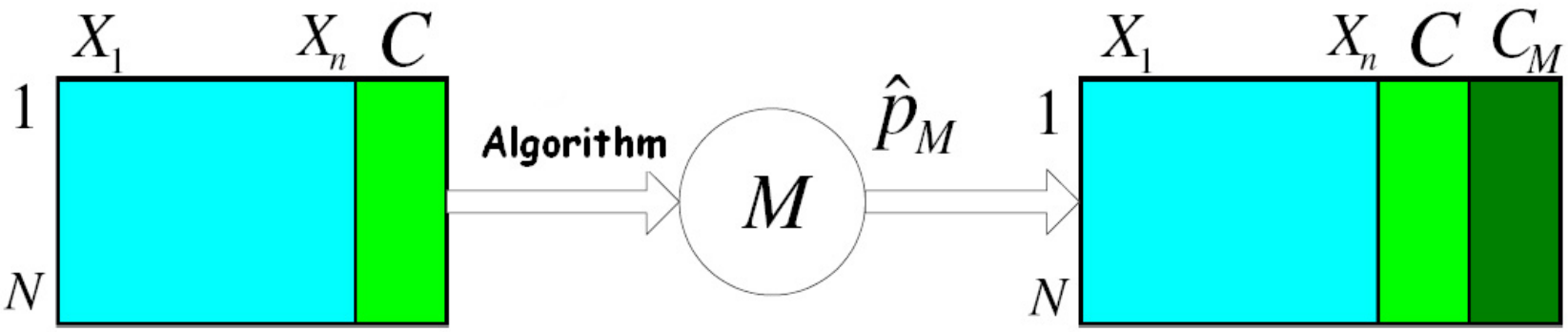
Accuracy: $\frac{a+d}{a+b+c+d}$

Error rate: $\frac{c+b}{a+b+c+d}$

Rate of true positives (sensitivity): $\frac{a}{a+c}$

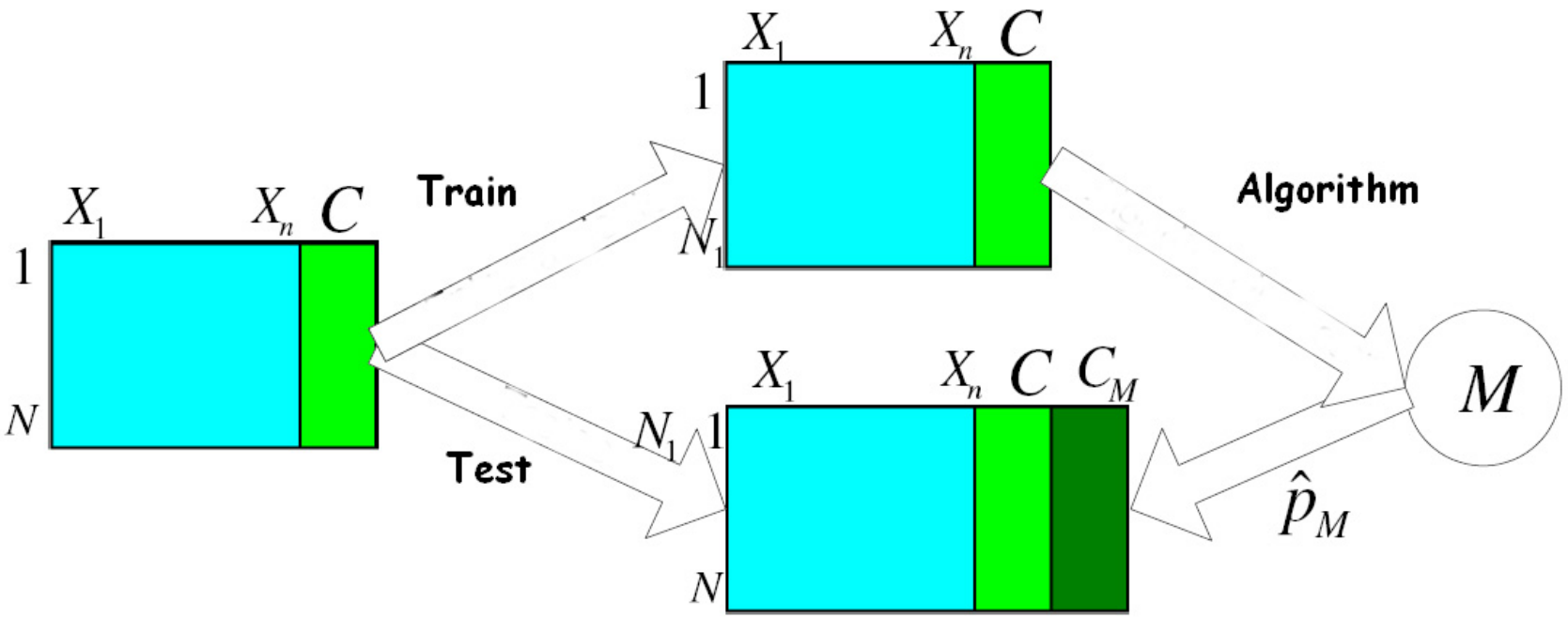
Rate of true negatives (specificity): $\frac{d}{b+d}$

Estimation methods: No honest



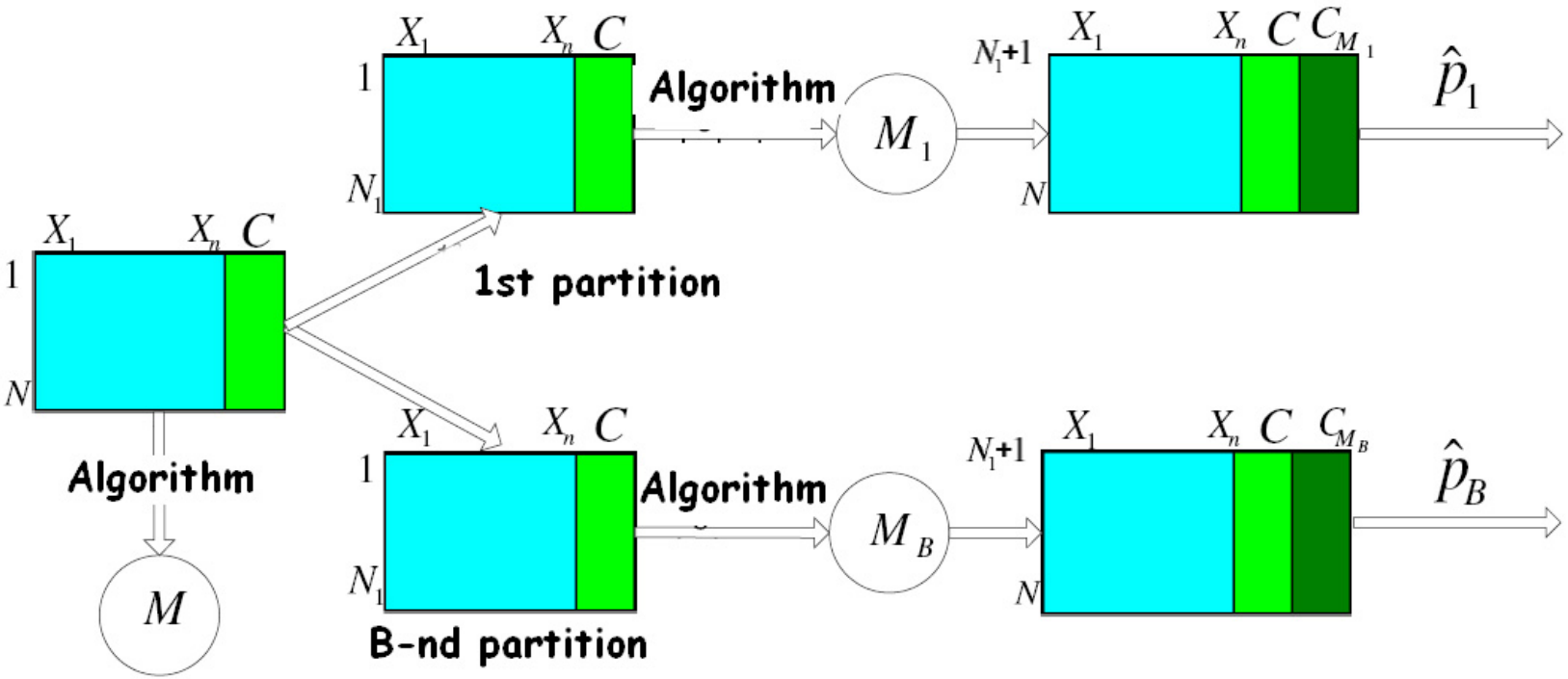
$$\hat{p}_M = \frac{1}{N} \sum_{i=1}^{N} \delta(c^{(i)} = c_M^{(i)})$$

Estimation methods: Train and test



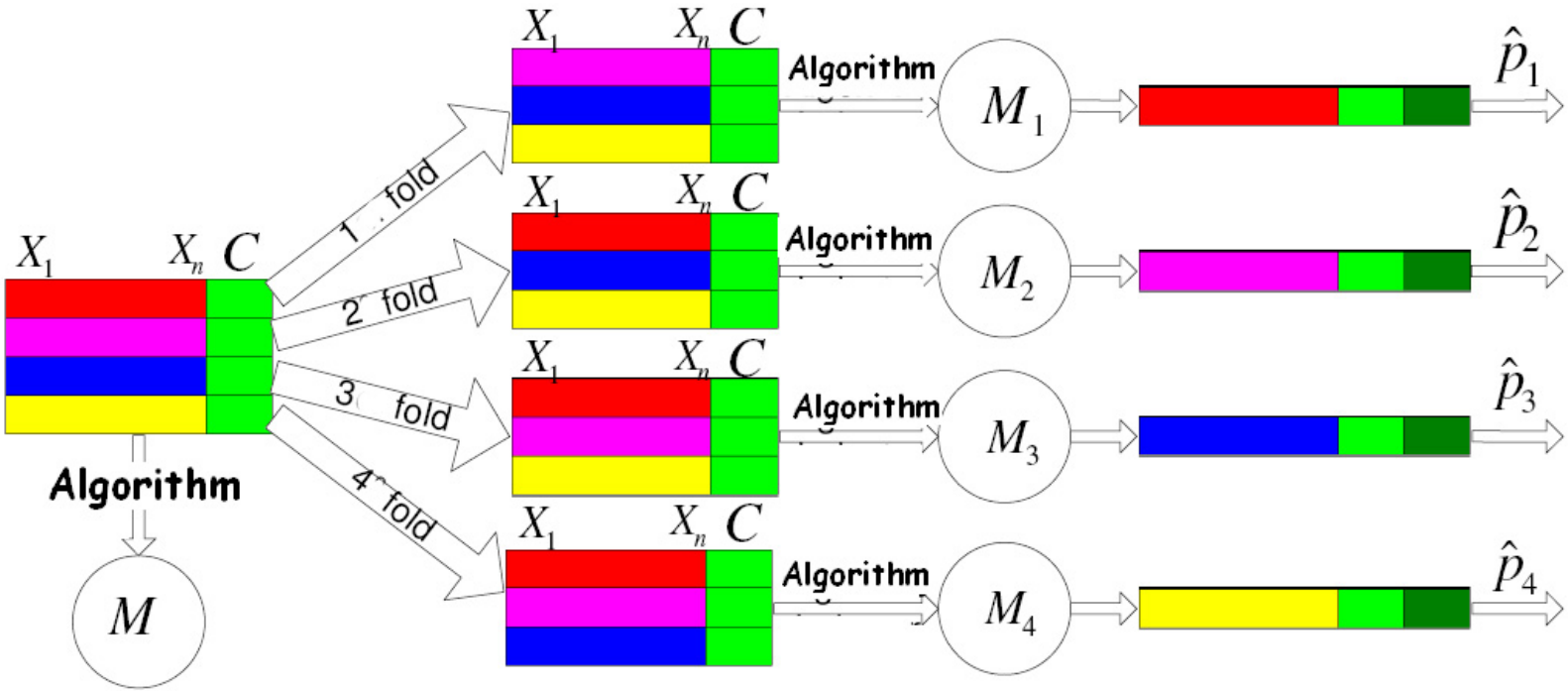
$$\hat{p}_M = \frac{1}{N - N_1} \sum_{i=1}^{N-N_1} \delta(c^{(N_1+i)} = c_M^{(N_1+i)})$$

Estimation methods: Train and test several times



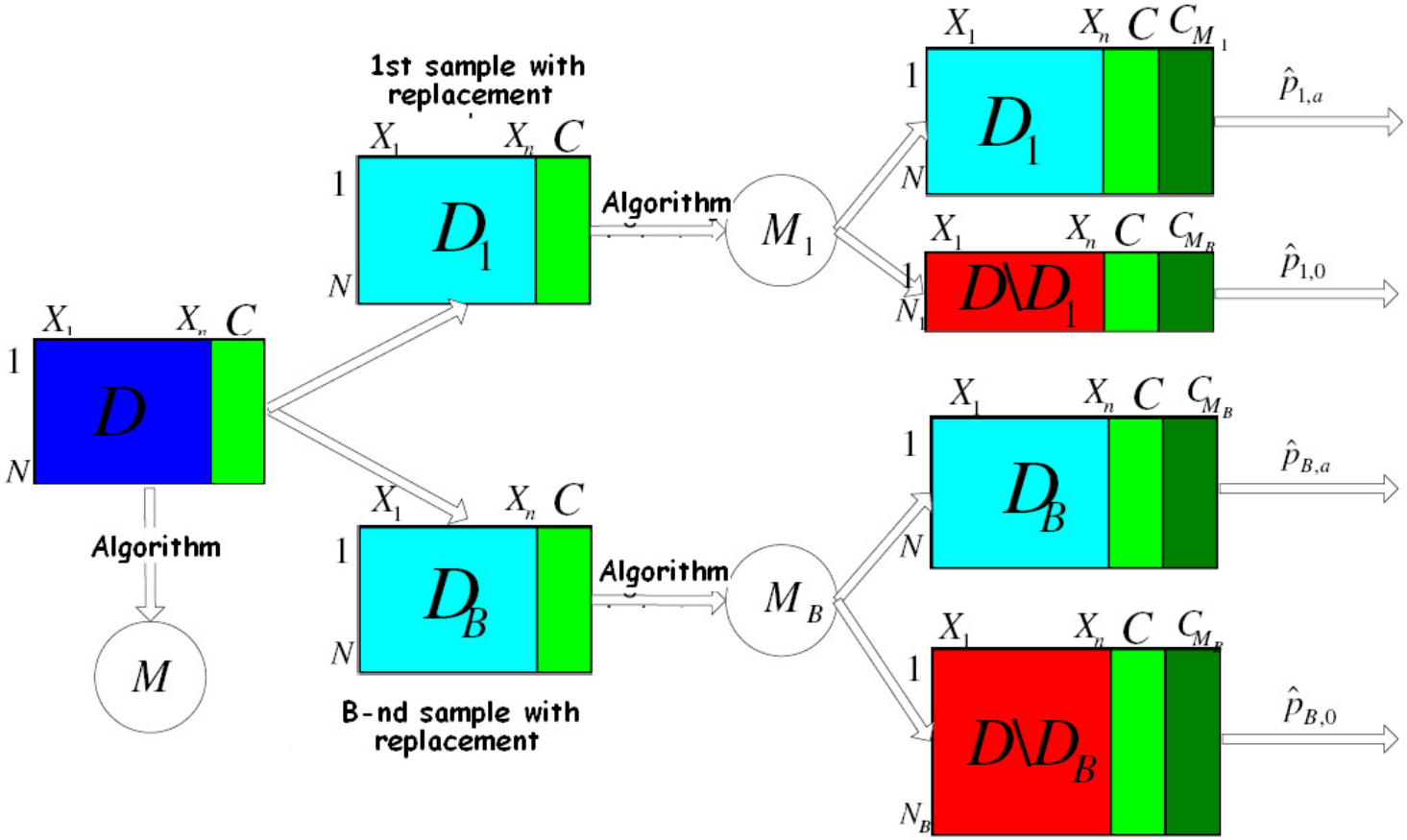
$$\hat{p}_M = \frac{1}{B}\sum_{i=1}^{B}\hat{p}_i$$

Estimation methods: k–fold cross validation



$$\hat{p}_M = \frac{1}{k} \sum_{i=1}^{k} \hat{p}_i$$

Estimation methods: 0.632 bootstrapping



$$\hat{p}_a = \frac{1}{B}\sum_{i=1}^{B}\hat{p}_{i,a} \quad \hat{p}_0 = \frac{1}{B}\sum_{i=1}^{B}\hat{p}_{i,0}$$

$$\hat{p}_M = \hat{p}_{0.632Bo} = (0.368\hat{p}_a + 0.632\hat{p}_0)$$