Department of Mathematics
University of Calabria

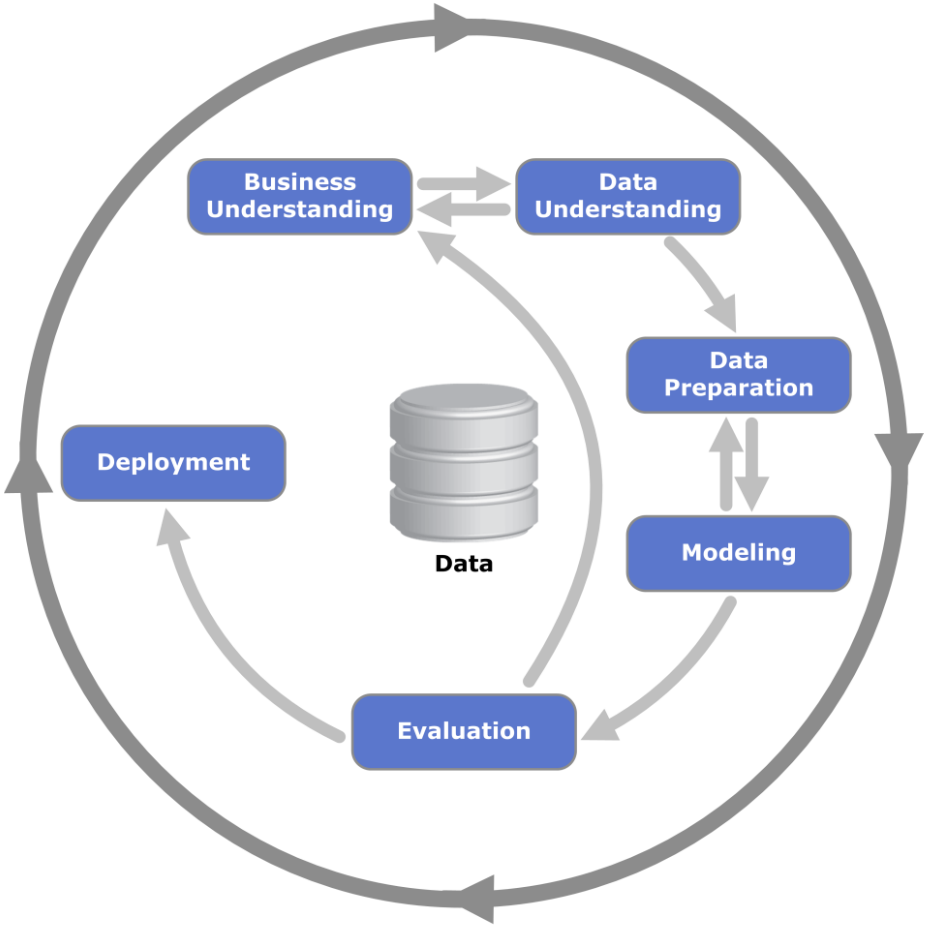
Data Warehouse and Data Mining

Module II – Data Mining

Evaluation

Ph.D. Ettore Ritacco

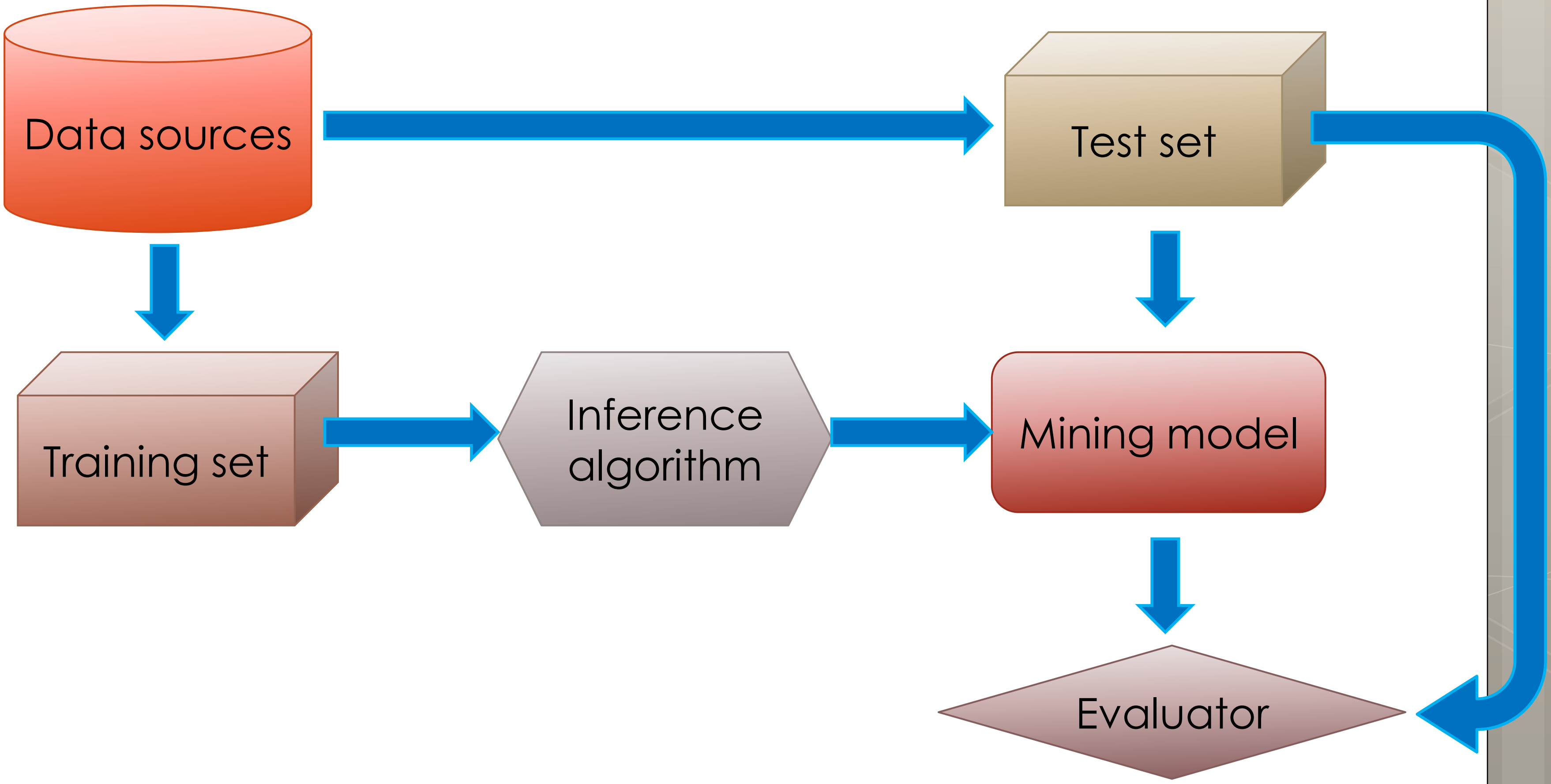
CRISP-DM Methodology



How to evaluate a model?

- Select a training set
- Build a mining model
- **Choose a quality measure**
- **Select a test set**
- **Apply the model on the test set**
- **Compute the value of the quality measure**

A simple evaluation schema

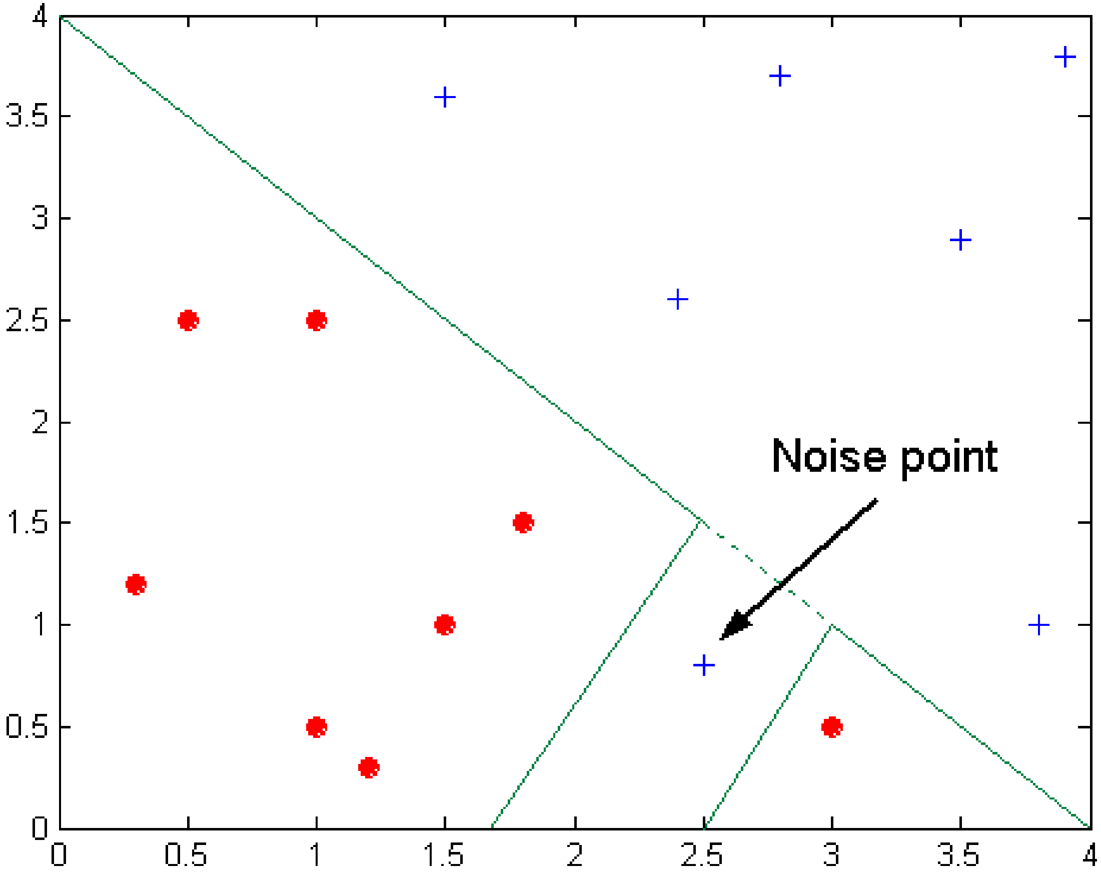


The fitting problem

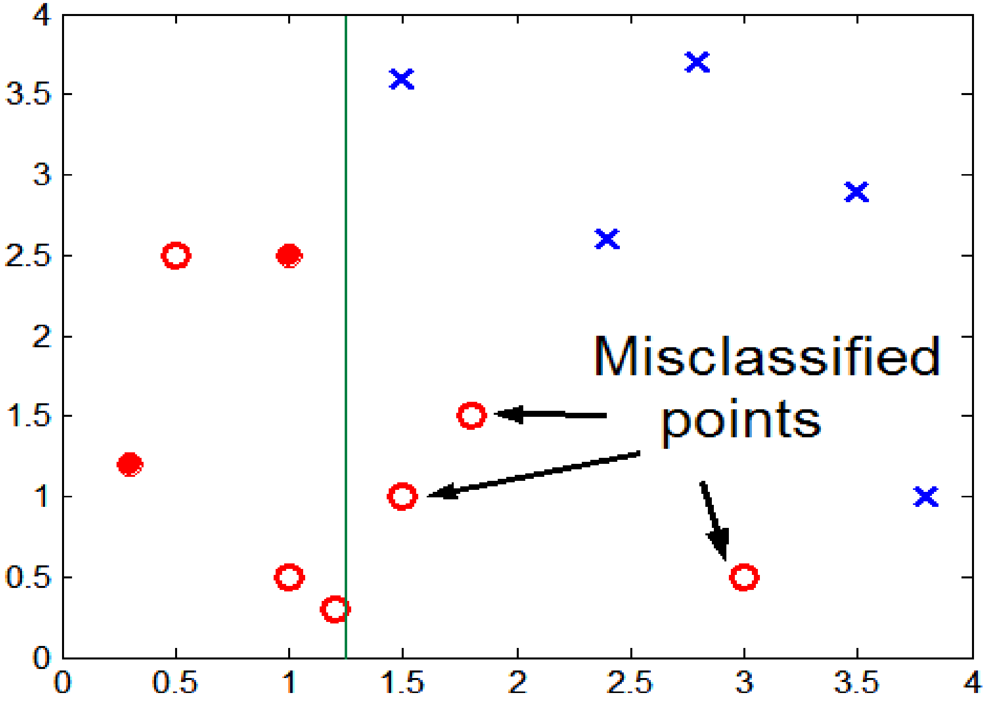
- Beyond the data analysis issues, there are challenges even in the modeling and evaluate phases in the CRISP-DM Methodology

- Namely
  - Underfitting
    - The model is too simple: the evaluation will be poor on both the training and the evaluation set
  - Overfitting
    - The model is too complex, fitting as close as it can the training data, the evaluation will be good on the training set, but poor on the evaluation set

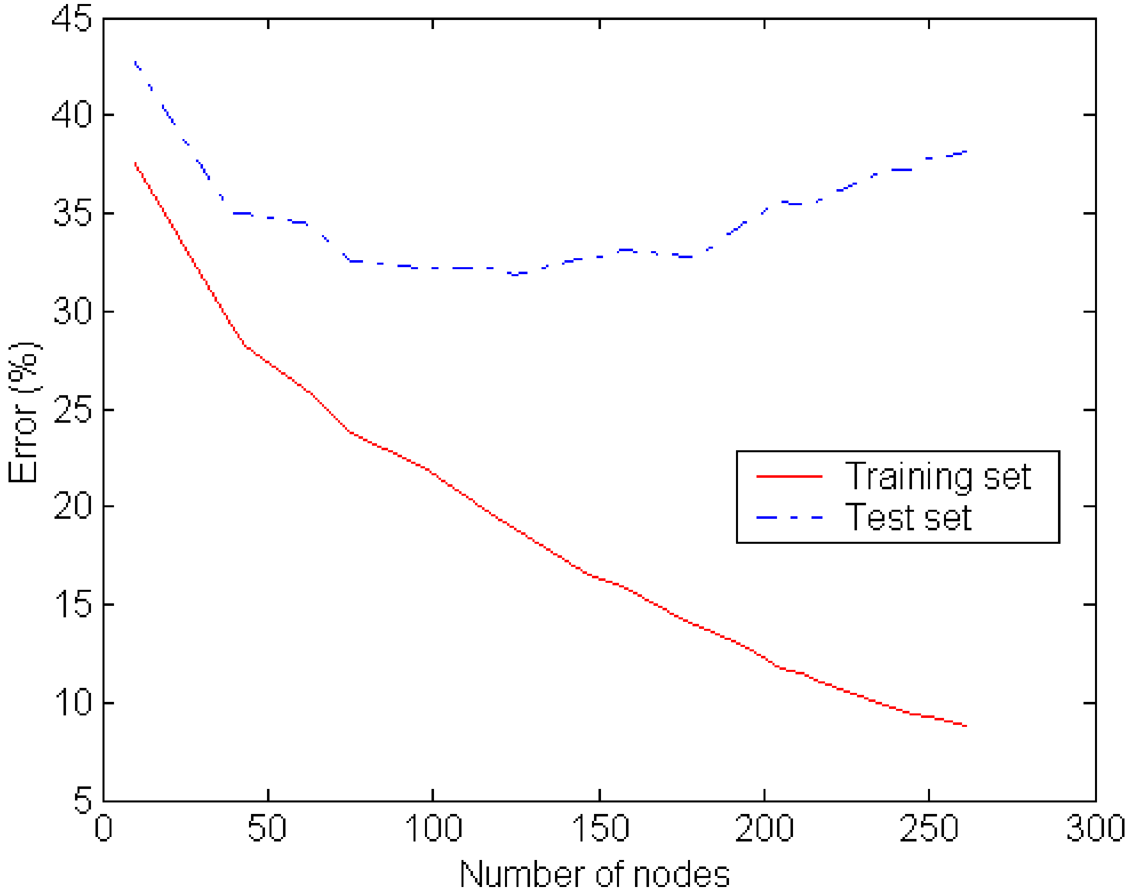
Overfitting (due to noise)



Overfitting (due to a too little dimension of the data set)



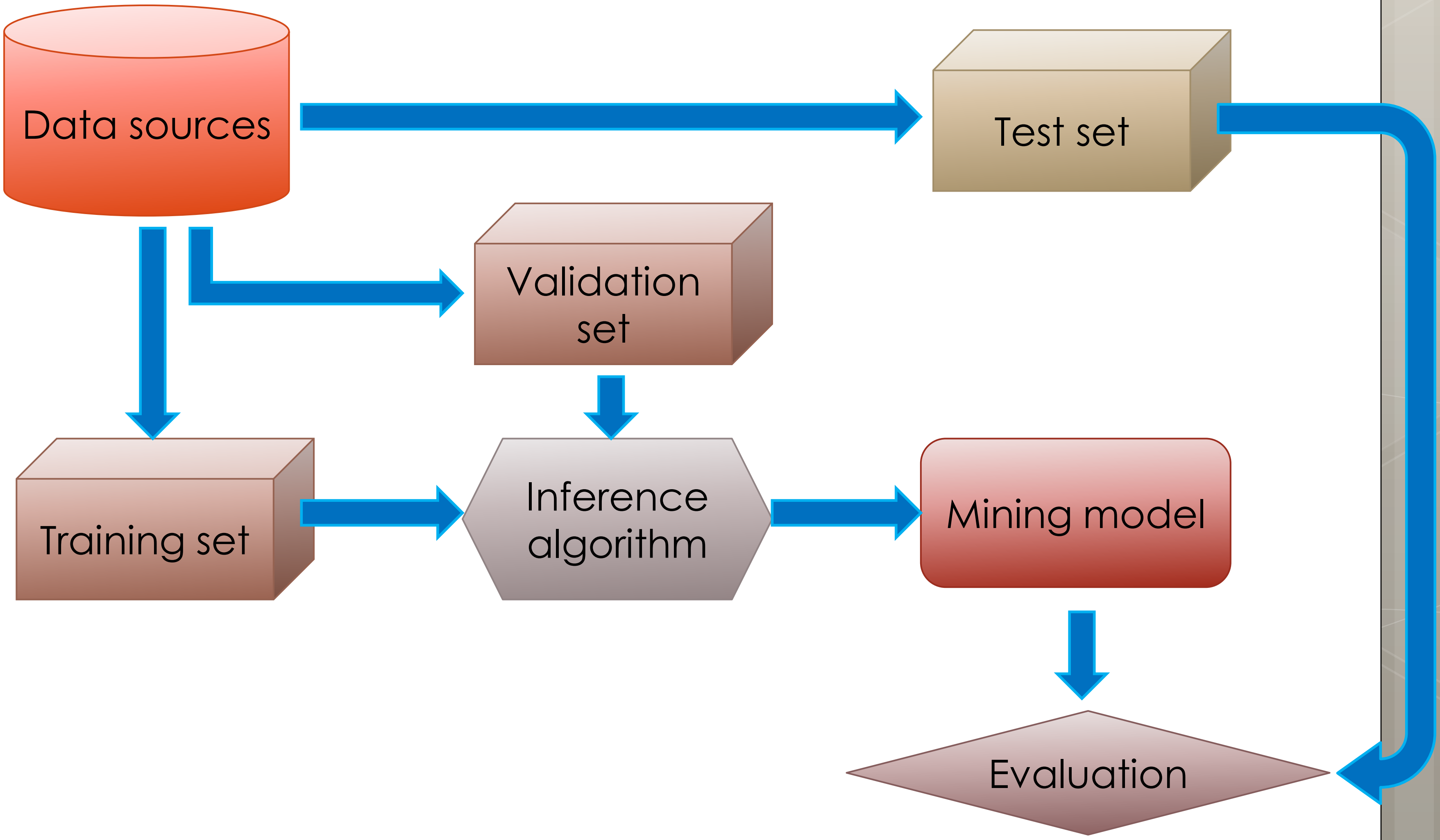
Overfitting



How to mitigate the overfittig?

- Prevention
  - A good data preparation
- Avoiding
  - Feed the building phase with further data for improving the model’s generality (e.g. online pruning)
- Recovery
  - Manipulate the model after its creation (e.g. post pruning)

How to mitigate the overfittig?



How to evaluate a model?

- Is a model that achives 70% of global accuracy a "good" model?

How to evaluate a model?

- Is a model that achives 70% of global accuracy a "good" model?
  - It dipends…

How to evaluate a model?

- Is a model that achives 70% of global accuracy a "good" model?
  - It dipends…
- Is a model that achives 95% of global accuracy a "good" model?

How to evaluate a model?

- Is a model that achives 70% of global accuracy a "good" model?
  - It dipends…
- Is a model that achives 95% of global accuracy a "good" model?
  - It dipends…

How to evaluate a model?

- We can perform only comparative evaluations.
- A “*null hypothesis*” (in other words, a *baseline*) is needed.
- We can only say, given a statistic, if a model is better then another one, in terms of the chosen statistic.

How to evaluate a model?

- The “true” error of a hypothesis *h*

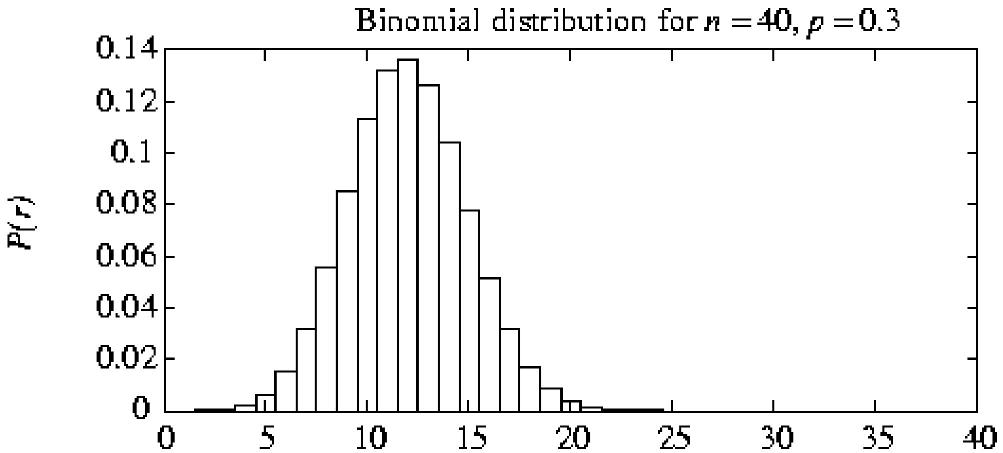
$$e(h) = \underset{x\in D}{P}\left[c(x) \neq h(x)\right]$$

- The error on our sample

$$e(h) = \frac{1}{|S|}\sum_{x\in S}\delta\left(c(x) \neq h(x)\right)$$

How to evaluate a model?

- The probability of *r* misclassifications is governed by a binomial distribution:



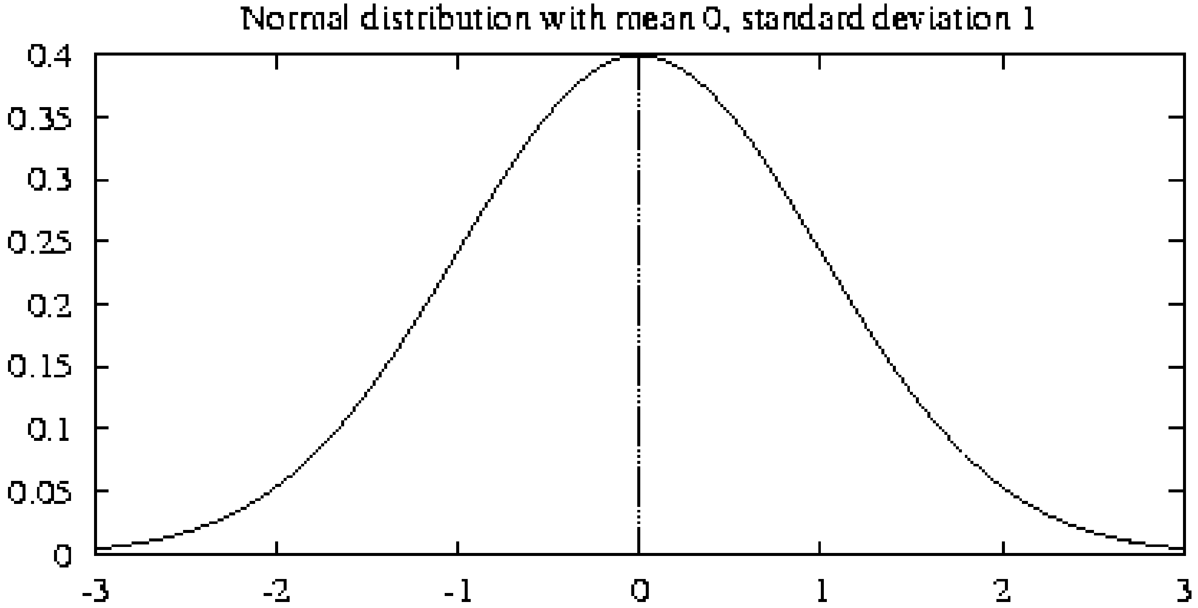
$$P(r) = \frac{|S|!}{r!\,(|S|-r)!}\,e(h)^r\,(1-e(h))^{|S|-r}$$

How to evaluate a model?

- If $|S|$ is sufficient great (typically $|S|>30$) the binomial distribution can be approximated by a normal distribution
  - Central limit theorem

How to evaluate a model?

- Normal distribution



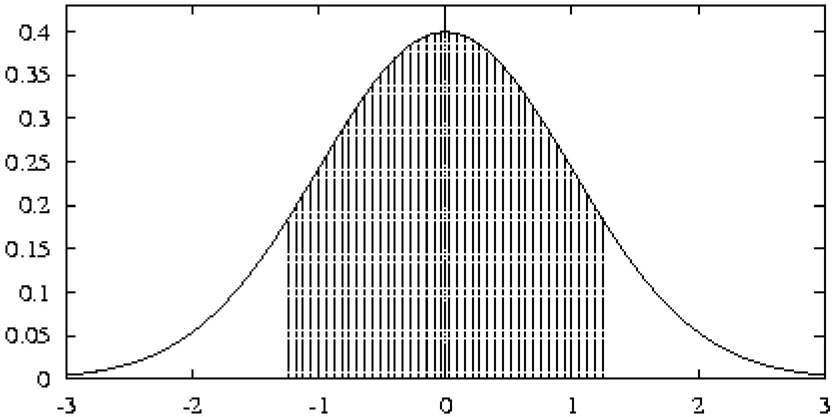
How to evaluate a model?

- Normal distribution
  - Density $$p(x) = \frac{1}{\sqrt{2\pi\sigma}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)$$
  - Cumulative $$P(a \le X \le b) = \int_a^b p(x)dx$$
  - Expected Value $$E[X] = \mu$$
  - Variance $$Var[X] = \sigma^2$$

How to evaluate a model?

- Confidence Intervals
  - Given a probability α, we are interested in finding an interval [a, b] such that

$$P(a \le X \le b) = \alpha$$



  - In the normal case

$$P(\mu - z_n\sigma \le X \le \mu + z_n\sigma) = \gamma$$

γ	50%	68%	80%	90%	**95%**	98%	99%
z_N	0.67	1.00	1.28	1.64	**1.96**	2.33	2.58

How to evaluate a model?

- Consider two hypothesis *h* and *j*…
- … and the random variable

$$d = e(h) - e(j)$$

- Choose z_n and consequently γ

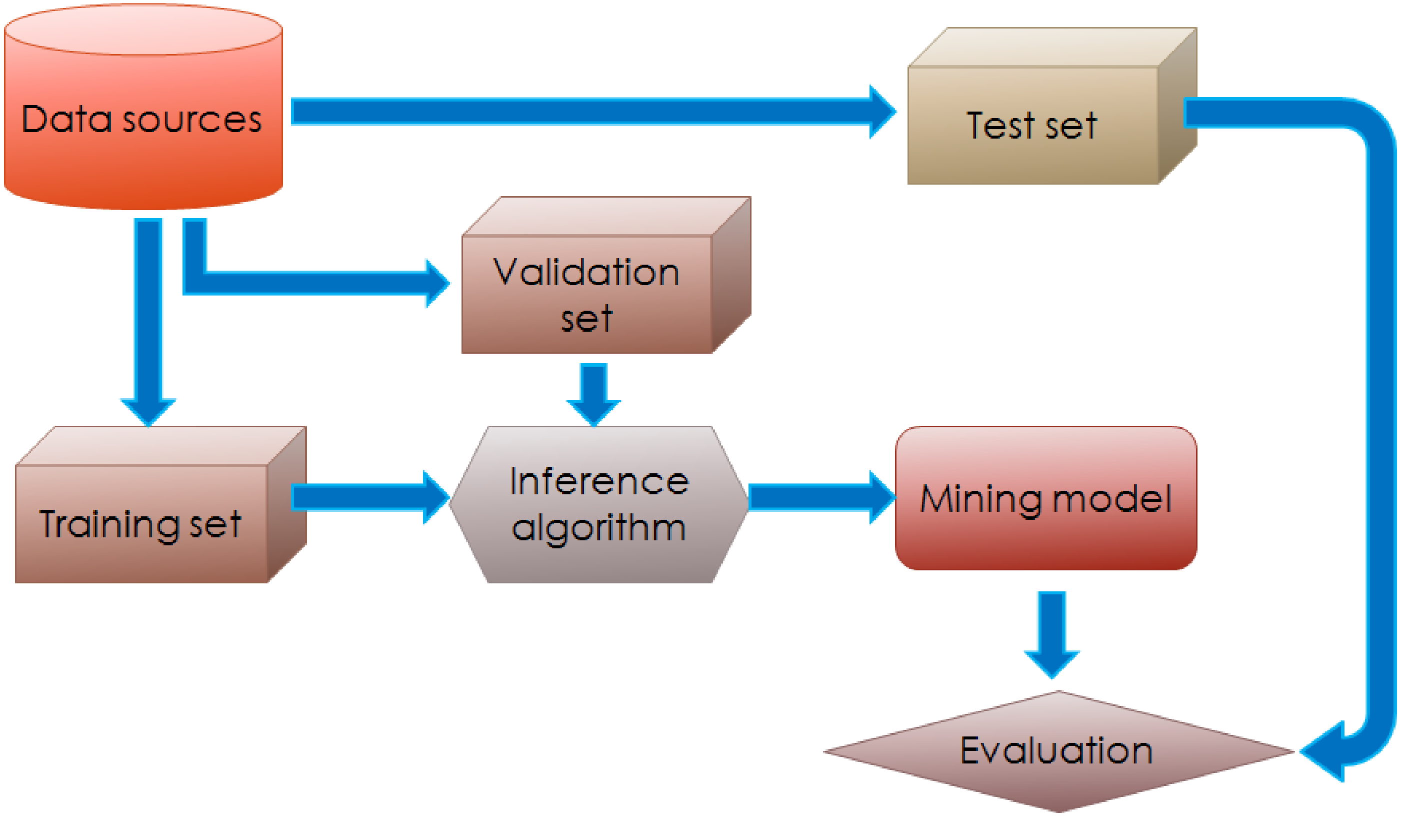
How to evaluate a model?

- Three cases: $d = e(h) - e(j)$
  - Zero is in the confidence interval of *d*
    - There is no statistical difference between *h* and *j*, with significance γ
  - The confidence interval of *d* is under Zero
    - *e(h)* is statistically lower than *e(j)*, with significance γ
  - The confidence interval of *d* is above Zero
    - *e(h)* is statistically higher than *e(j)*, with significance γ

$$P(\mu - z_n\sigma \leq X \leq \mu + z_n\sigma) = \gamma$$

Methods for model evaluation

- Hold-out



Methods for model evaluation

- Hold-out
  - Pros:
    - Fast evaluation
  - Cons:
    - Only one experiment ➔ low statistical relevance

Methods for model evaluation

- Repeated Hold-out with random sub-sampling
  - Choose *n*
  - *ResultList = { }*
  - *For 1 < i < n*
    - *Random Sampling of (with or without replacement):*
      - *Training set*
      - *Validation set*
      - *Test set*
    - *Model = buildModel(Training set, Validation set)*
    - *ResultList.add(evaluateModel(Model, Test set))*
  - *Return avg(ResultList)*

Methods for model evaluation

- Repeated Hold-out with random sub-sampling
  - Pros:
    - More statistical significance
  - Cons:
    - Slow evaluation
    - Not all the tuples are involved in the training and evaluation phase

Methods for model evaluation

- *k*-fold Cross Validation
  - *Choose k*
  - *Divide the whole dataset D in k folds (portion)*
  - *ResultList = { }*
  - *For 1 < i < k*
    - *Build Training set = D \ $fold_i$*
    - *Random sample the Validation Set from the Training Set*
    - *Training set = Training set \ Validation Set*
    - *Test set = $fold_i$*
    - *Model = buildModel(Training set, Validation set)*
    - *ResultList.add(evaluateModel(Model, Test set))*
  - *Return avg(ResultList)*

Methods for model evaluation

- *k*-fold Cross Validation
  - Pros:
    - Good statistical significance
      - the greater is *k* the better the significance
        - If $k = |D|$ Cross Validation is called leave-one-out evaluation
  - Cons:
    - Very slow evaluation
    - The *k*-fold Cross Validation needs to be stratified:
      - Each fold has to keep the same statistical properties of the whole dataset

Evaluation Metrics

- The focus is on the predictive quality of a model
  - instead of computational cost, scalability…
- Confusion Matrix

	Predicted class		
Actual class		Class = Yes	Class = No
	Class = Yes	True Positive (TP)	False Negative (FN)
	Class = No	False Positive (FP)	True Negative (TN)

Evaluation Metrics

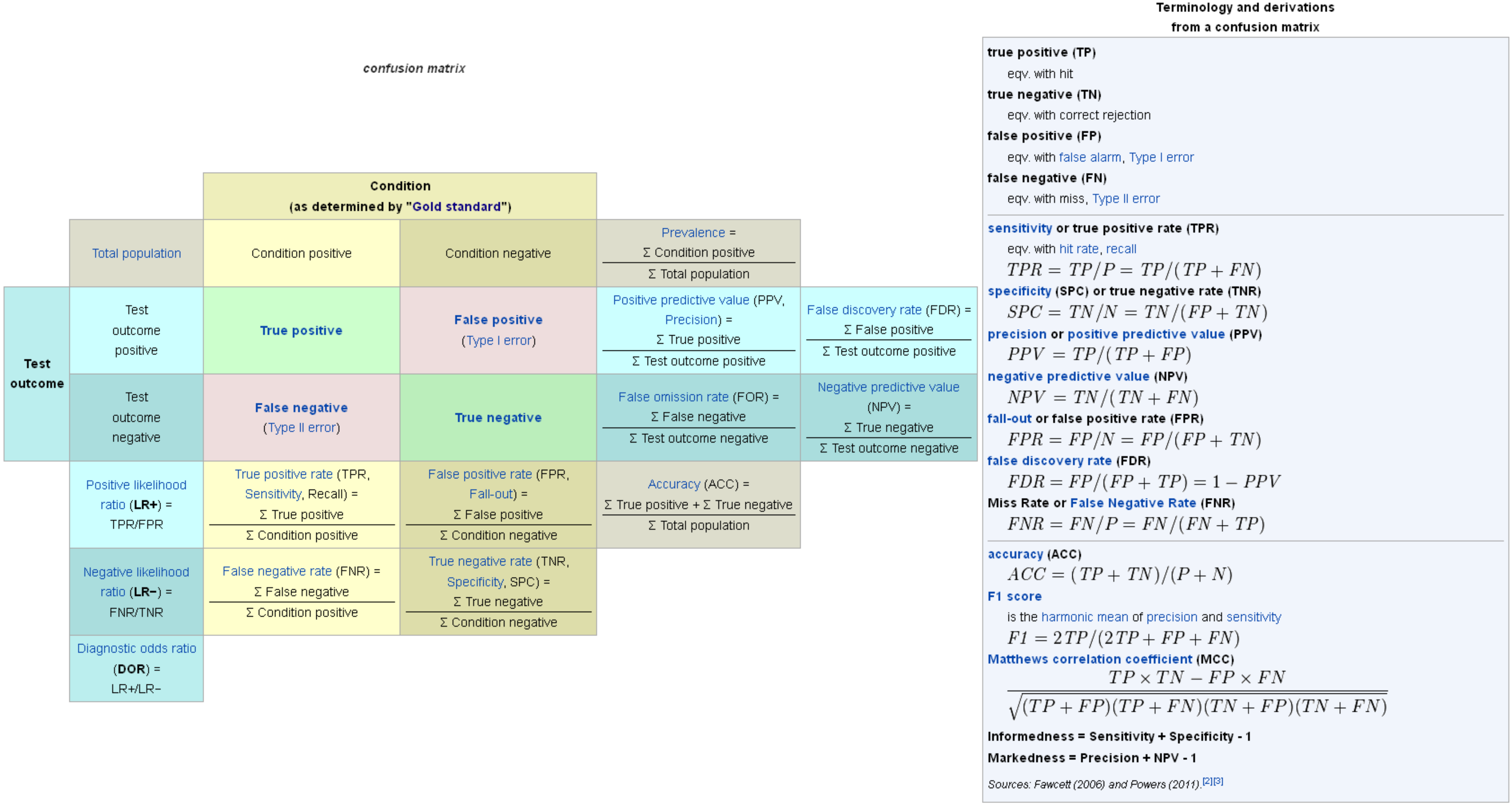
- Global Accuracy

$$accuracy = \frac{TP + TN}{TP + FN + FP + TN}$$

- Is a classifier, with a global accuracy equals to 99.9%, good?

- To be continued…

Confusion Matrix



confusion matrix

		Condition (as determined by "**Gold standard**")			
	Total population	Condition positive	Condition negative	Prevalence = $\frac{\Sigma \text{ Condition positive}}{\Sigma \text{ Total population}}$	
Test outcome	Test outcome positive	**True positive**	**False positive** (Type I error)	Positive predictive value (PPV, Precision) = $\frac{\Sigma \text{ True positive}}{\Sigma \text{ Test outcome positive}}$	False discovery rate (FDR) = $\frac{\Sigma \text{ False positive}}{\Sigma \text{ Test outcome positive}}$
	Test outcome negative	**False negative** (Type II error)	**True negative**	False omission rate (FOR) = $\frac{\Sigma \text{ False negative}}{\Sigma \text{ Test outcome negative}}$	Negative predictive value (NPV) = $\frac{\Sigma \text{ True negative}}{\Sigma \text{ Test outcome negative}}$
	Positive likelihood ratio (**LR+**) = TPR/FPR	True positive rate (TPR, Sensitivity, Recall) = $\frac{\Sigma \text{ True positive}}{\Sigma \text{ Condition positive}}$	False positive rate (FPR, Fall-out) = $\frac{\Sigma \text{ False positive}}{\Sigma \text{ Condition negative}}$	Accuracy (ACC) = $\frac{\Sigma \text{ True positive} + \Sigma \text{ True negative}}{\Sigma \text{ Total population}}$	
	Negative likelihood ratio (**LR−**) = FNR/TNR	False negative rate (FNR) = $\frac{\Sigma \text{ False negative}}{\Sigma \text{ Condition positive}}$	True negative rate (TNR, Specificity, SPC) = $\frac{\Sigma \text{ True negative}}{\Sigma \text{ Condition negative}}$		
	Diagnostic odds ratio (**DOR**) = LR+/LR−				

Terminology and derivations from a confusion matrix

true positive (TP)
eqv. with hit

true negative (TN)
eqv. with correct rejection

false positive (FP)
eqv. with false alarm, Type I error

false negative (FN)
eqv. with miss, Type II error

sensitivity or true positive rate (TPR)
eqv. with hit rate, recall

$$TPR = TP/P = TP/(TP + FN)$$

specificity (SPC) or true negative rate (TNR)

$$SPC = TN/N = TN/(FP + TN)$$

precision or positive predictive value (PPV)

$$PPV = TP/(TP + FP)$$

negative predictive value (NPV)

$$NPV = TN/(TN + FN)$$

fall-out or false positive rate (FPR)

$$FPR = FP/N = FP/(FP + TN)$$

false discovery rate (FDR)

$$FDR = FP/(FP + TP) = 1 - PPV$$

Miss Rate or False Negative Rate (FNR)

$$FNR = FN/P = FN/(FN + TP)$$

accuracy (ACC)

$$ACC = (TP + TN)/(P + N)$$

F1 score
is the harmonic mean of precision and sensitivity

$$F1 = 2TP/(2TP + FP + FN)$$

Matthews correlation coefficient (MCC)

$$\frac{TP \times TN - FP \times FN}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}}$$

Informedness = Sensitivity + Specificity - 1

Markedness = Precision + NPV - 1

Sources: Fawcett (2006) and Powers (2011).[2][3]