Bias and Variance
in
Machine Learning

Pierre Geurts

Université de Liège

Octobre 2002

Content of the presentation

- Bias and variance definitions
- Parameters that influence bias and variance
- Variance reduction techniques
- Decision tree induction

Content of the presentation

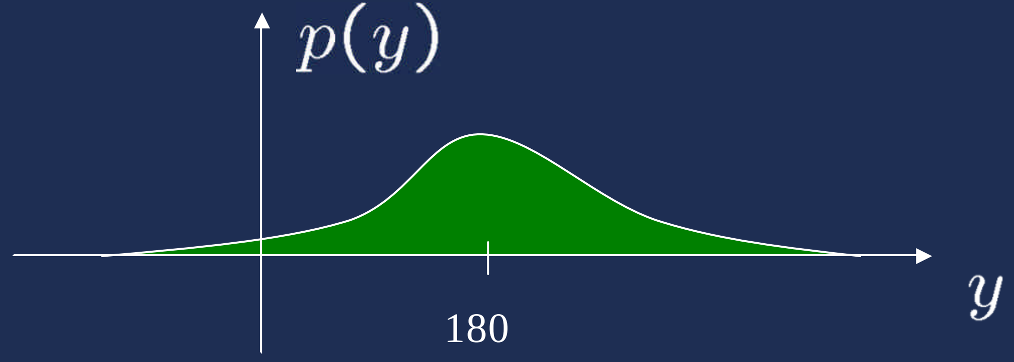
- Bias and variance definitions:
  - A simple regression problem with no input
  - Generalization to full regression problems
  - A short discussion about classification
- Parameters that influence bias and variance
- Variance reduction techniques
- Decision tree induction

Regression problem - no input

- Goal: predict as well as possible the height of a Belgian male adult
- More precisely:
  - Choose an error measure, for example the square error.
  - Find an estimation *y* such that the expectation:

$$E_y\{(y - \widehat{y})^2\}$$

over the whole population of Belgian male adult is minimized.



Regression problem - no input

- The estimation that minimizes the error can be computed by taking:

$$
\begin{aligned}
& \frac{\partial}{\partial y'} E_y\{(y-y')^2\} = 0 \\
\Leftrightarrow \quad & E_y\{-2.(y-y')\} = 0 \\
\Leftrightarrow \quad & E_y\{y\} - E_y\{y'\} = 0 \\
\Leftrightarrow \quad & y' = E_y\{y\}
\end{aligned}
$$

- So, the estimation which minimizes the error is $E_y\{y\}$. In AL, it is called the Bayes model.
- **<u>But</u>** in practice, we cannot compute the exact value of $E_y\{y\}$ (this would imply to measure the height of every Belgian male adults).

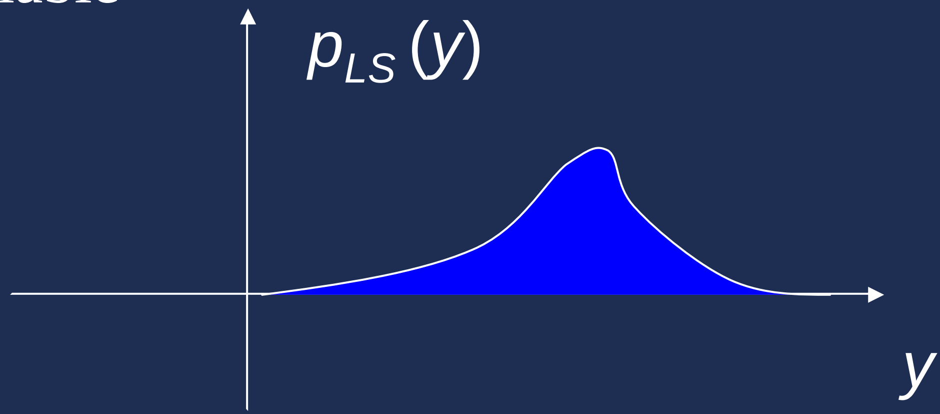
Learning algorithm

- As $p(y)$ is unknown, find an estimation y from a sample of individuals, $LS=\{y_1,y_2,\ldots,y_N\}$, drawn from the Belgian male adult population.
- Example of learning algorithms:
  - $$\hat{y}_1 = \frac{1}{N}\sum_{i=1}^{N} y_i$$
  - $$\hat{y}_2 = \frac{\lambda 180 + \sum_{i=1}^{N} y_i}{\lambda + N}, \lambda \in [0; +\infty[$$

(if we know that the height is close to 180)

Good learning algorithm

- As *LS* are randomly drawn, the prediction *y* will also be a random variable



- A good learning algorithm should not be good only on one learning sample but in average over all learning samples (of size *N*) $\Rightarrow$ we want to minimize:

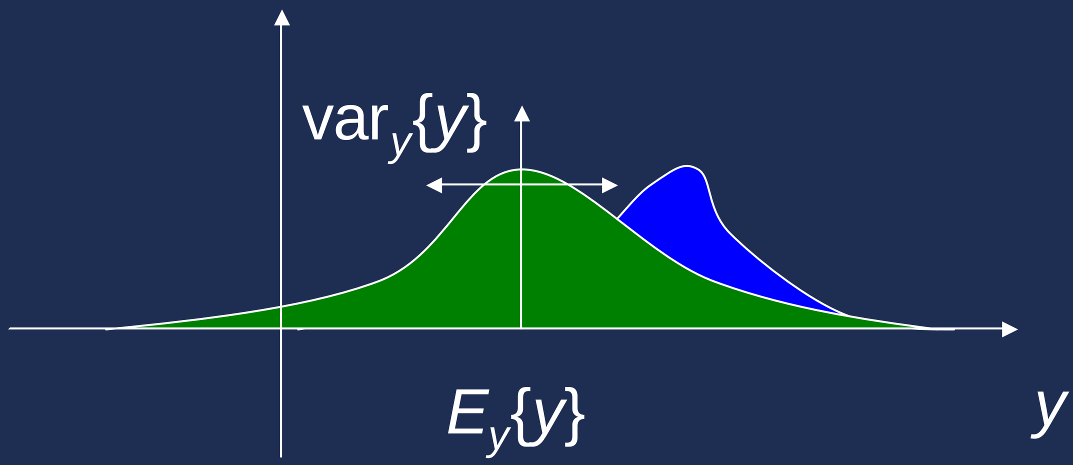
$$E = E_{LS}\{E_y\{(y - \hat{y})^2\}\}$$

- Let us analyse this error in more detail

Bias/variance decomposition (1)

$$
\begin{aligned}
& E_{LS}\{E_y\{(y-\hat{y})^2\}\} \\
= \; & E_{LS}\{E_y\{(y - E_y\{y\} + E_y\{y\} - \hat{y})^2\}\} \\
= \; & E_{LS}\{E_y\{(y - E_y\{y\})^2\}\} + E_{LS}\{E_y\{(E_y\{y\} - \hat{y})^2\}\} \\
+ \; & E_{LS}\{E_y\{2(y - E_y\{y\})(E_y\{y\} - \hat{y})\}\} \\
= \; & E_y\{(y - E_y\{y\})^2\} + E_{LS}\{(E_y\{y\} - \hat{y})^2\} \\
+ \; & E_{LS}\{2(E_y\{y\} - E_y\{y\})(E_y\{y\} - \hat{y})\} \\
= \; & E_y\{(y - E_y\{y\})^2\} + E_{LS}\{(E_y\{y\} - \hat{y})^2\}
\end{aligned}
$$

Bias/variance decomposition (2)



$$E = \underbrace{\boldsymbol{E_y\{(y - E_y\{y\})^2\}}} + E_{LS}\{(E_y\{y\} - y)^2\}$$

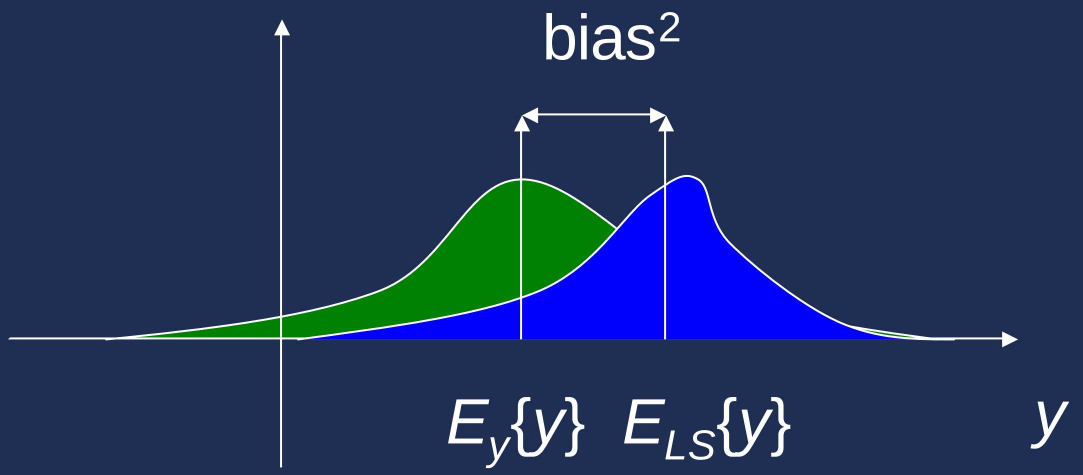
$= \text{residual error} = \text{minimal atteinable error}$

$= \text{var}_y\{y\}$

Bias/variance decomposition (3)

$$
\begin{aligned}
& E_{LS}\{(E_y\{y\} - \widehat{y})^2\} \\
= \;& E_{LS}\{(E_y\{y\} - E_{LS}\{\widehat{y}\} + E_{LS}\{\widehat{y}\} - \widehat{y})^2\} \\
= \;& E_{LS}\{(E_y\{y\} - E_{LS}\{\widehat{y}\})^2\} + E_{LS}\{(E_{LS}\{\widehat{y}\} - \widehat{y})^2\} \\
+ \;& E_{LS}\{2(E_y - E_{LS}\{\widehat{y}\})(E_{LS}\{\widehat{y}\} - \widehat{y})\}\} \\
= \;& (E_y\{y\} - E_{LS}\{\widehat{y}\})^2 + E_{LS}\{(\widehat{y} - E_{LS}\{\widehat{y}\})^2\} \\
+ \;& 2(E_y - E_{LS}\{\widehat{y}\})E_{LS}\{(E_{LS}\{\widehat{y}\} - \widehat{y})\}\} \\
= \;& (E_y\{y\} - E_{LS}\{\widehat{y}\})^2 + E_{LS}\{(\widehat{y} - E_{LS}\{\widehat{y}\})^2\}
\end{aligned}
$$

Bias/variance decomposition (4)

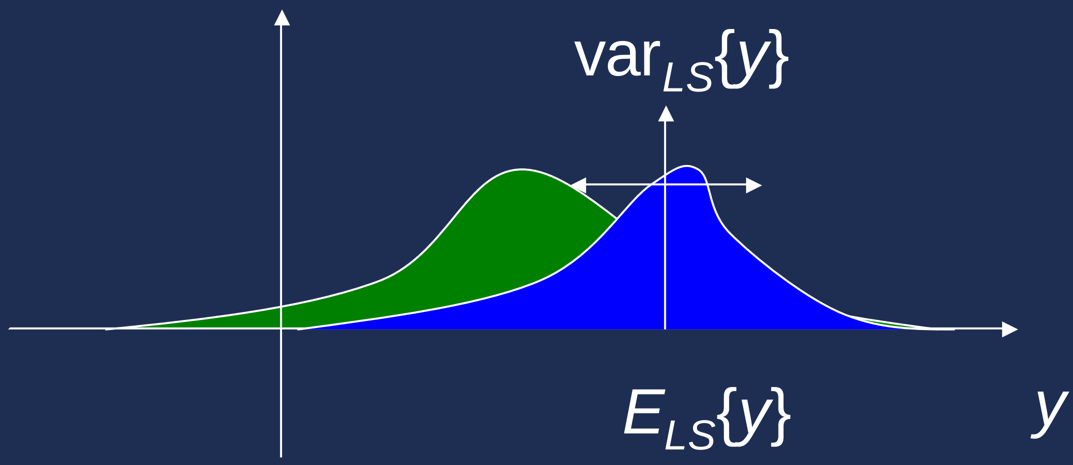


E= var$_y\{y\}$ + $\mathbf{(E_y\{y\}-E_{LS}\{y\})^2}$ + …

$E_{LS}\{\boldsymbol{y}\}$ = average model (over all LS)

bias^2 = error between bayes and average model

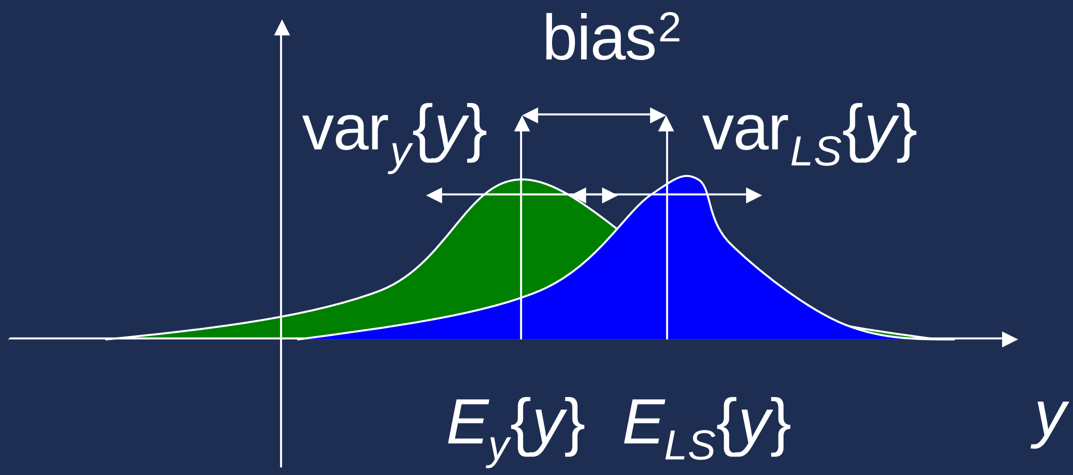
Bias/variance decomposition (5)



$E = \mathrm{var}_y\{y\} + \mathrm{bias}^2 + \boldsymbol{E_{LS}\{(y - E_{LS}\{y\})^2\}}$

$\mathrm{var}_{LS}\{\boldsymbol{y}\}$ = estimation variance = consequence of over-fitting

Bias/variance decomposition (6)



$$E = \text{var}_y\{y\} + \text{bias}^2 + \text{var}_{LS}\{\boldsymbol{y}\}$$

Our simple example

- $\hat{y}_1 = \frac{1}{N}\sum_i y_i$
  - $\text{bias}^2 = (E_y\{y\} - E_{LS}\{\hat{y}_1\})^2 = 0$
  - $\text{var}_{LS}\{\hat{y}_1\} = \frac{1}{N}\text{var}_y\{y\}$

From statistics, y_1 is the best estimate with zero bias

- $\hat{y}_2 = \frac{\lambda 180 + \sum_i y_i}{\lambda + N}$
  - $\text{bias}^2 = \frac{\lambda}{\lambda+N}(E_y\{y\} - 180)^2$
  - $\text{var}_{LS}\{\hat{y}_2\} = \frac{N}{(\lambda+N)^2}\text{var}_y\{y\}$

So, the first one may not be the best estimator because of variance (There is a bias/variance tradeoff w.r.t. λ)

Bayesian approach (1)

- Hypotheses :
  - The average height is close to 180cm:

$$P(\bar{y}) = A \exp\left(-\frac{(\bar{y}-180)^2}{2\sigma_{\bar{y}}}\right)$$

  - The height of one individual is Gaussian around the mean:

$$P(y_i|\bar{y}) = B \exp\left(-\frac{(y_i-\bar{y})^2}{2\sigma_y}\right)$$

- What is the most probable value of $\bar{y}$ after having seen the learning sample ?

$$\hat{y} = \arg\max_{\bar{y}} P(\bar{y}|LS)$$

Bayesian approach (2)

$$
\begin{aligned}
\hat{y} &= \arg\max_{\bar{y}} P(\bar{y}|LS) \\
&= \arg\max_{\bar{y}} P(LS|\bar{y})P(\bar{y}) \\
&= \arg\max_{\bar{y}} P(y_1, ..., y_N|\bar{y})P(\bar{y}) \\
&= \arg\max_{\bar{y}} \prod_{i=1}^{N} P(y_i|\bar{y})P(\bar{y}) \\
&= \arg\min_{\bar{y}} -\sum_{i=1}^{N} \log(P(y_i|\bar{y})) - \log(P(\bar{y})) \\
&= \arg\min_{\bar{y}} \sum_{i=1}^{N} \frac{(y_i - \bar{y})^2}{2\sigma_y^2} + \frac{(\bar{y} - 180)^2}{2\sigma_{\bar{y}}^2} \\
&= \ldots \\
&= \frac{\lambda 180 + \sum_i y_i}{\lambda + N} \text{ with } \lambda = \frac{\sigma_y^2}{\sigma_{\bar{y}}^2}.
\end{aligned}
$$

Bayes theorem and *P*(*LS*) is constant

Independence of the learning cases

Regression problem – full (1)

- Actually, we want to find a function $y(\underline{x})$ of several inputs => average over the whole input space:

- The error becomes:

$$E_{\underline{x},y}\{(y - \hat{y}(\underline{x}))^2\}$$

- Over all learning sets:

$$\begin{aligned} E &= E_{LS}\{E_{\underline{x},y}\{(y - \hat{y}(\underline{x}))^2\}\} \\ &= E_{\underline{x}}\{E_{LS}\{E_{y|\underline{x}}\{(y - \hat{y}(\underline{x}))^2\}\}\} \\ &= E_{\underline{x}}\{\mathrm{var}_{y|\underline{x}}\{y\}\} + E_{\underline{x}}\{\mathrm{bias}^2(\underline{x})\} + E_{\underline{x}}\{\mathrm{var}_{LS}\{\hat{y}(\underline{x})\}\} \end{aligned}$$

Regression problem – full (2)

$$E_{LS}\{E_{y|\underline{x}}\{(y-y(\underline{x}))^2\}\} = \text{Noise}(\underline{x}) + \text{Bias}^2(\underline{x}) + \text{Variance}(\underline{x})$$

- Noise(x) = $E_{y|\underline{x}}\{(y-h_B(\underline{x}))^2\}$
  Quantifies how much y varies from $h_B(\underline{x}) = E_{y|\underline{x}}\{y\}$, the Bayes model.
- Bias²(x) = $(h_B(\underline{x})-E_{LS}\{y(\underline{x})\})^2$:
  Measures the error between the Bayes model and the average model.
- Variance(x) = $E_{LS}\{(y(\underline{x})-E_{LS}\{y(\underline{x})\})^2\}$:
  Quantify how much $y(\underline{x})$ varies from one learning sample to another.

Illustration (1)

- Problem definition:
  - One input x, uniform random variable in [0,1]
  - $y=h(x)+\epsilon$ where $\epsilon \sim N(0,1)$

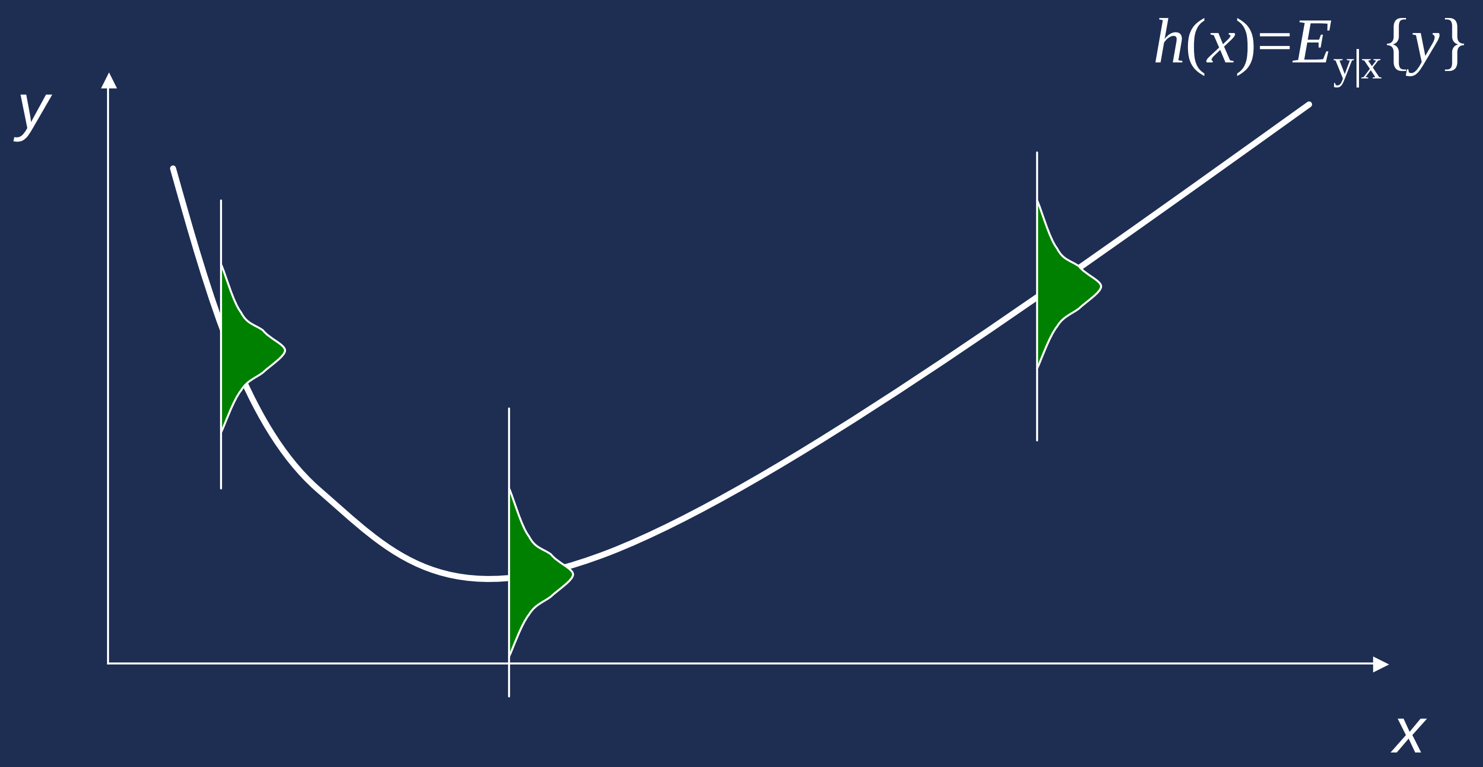


Illustration (2)

- Small variance, high bias method

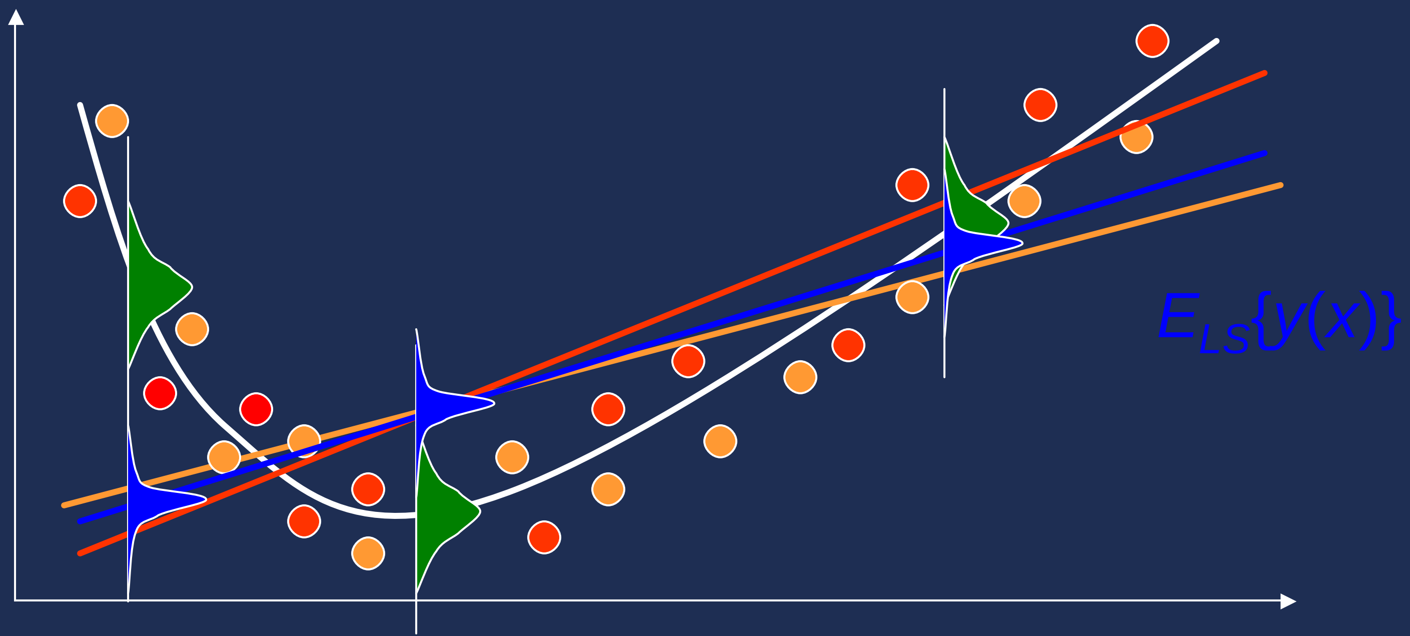
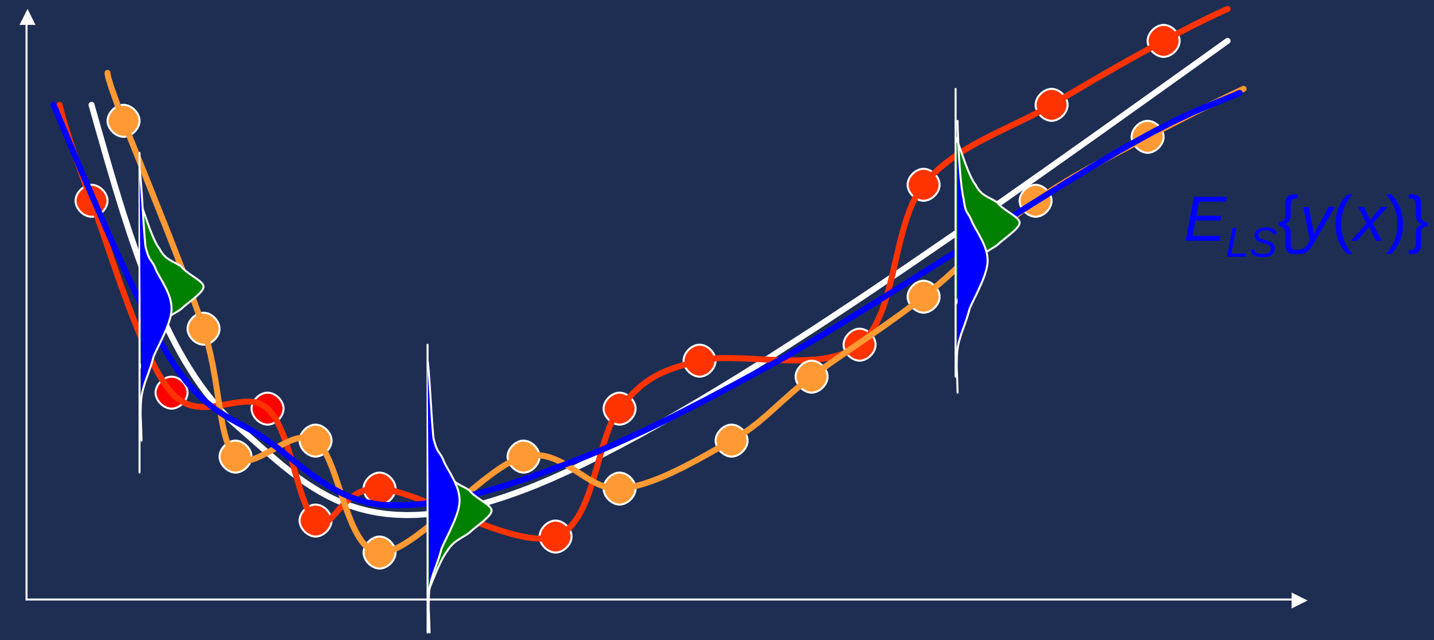
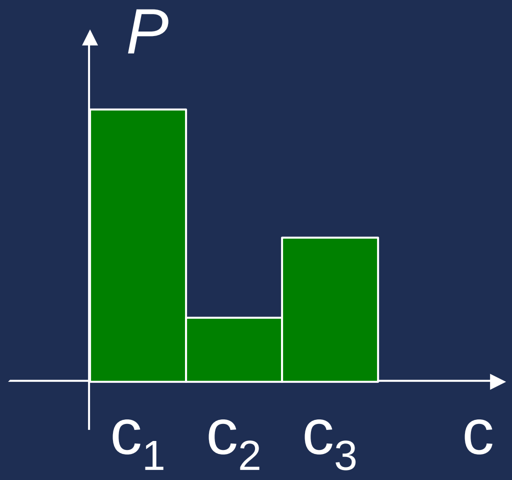


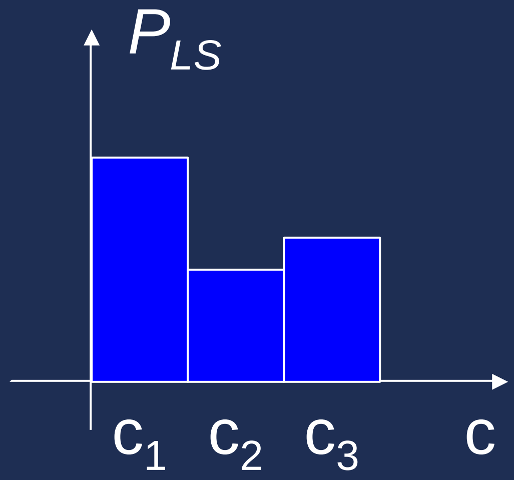
Illustration (3)

- Small bias, high variance method



Classification problem (1)





$\mathrm{err}(c,\mathsf{c})=1(c \neq \mathsf{c}) \Rightarrow PE=E_{\mathrm{LS}}\{E_{\mathrm{c}}\{1(c \neq \mathsf{c})\}\}$

Bayes model = c_{B} = arg $\max_c P(c)$

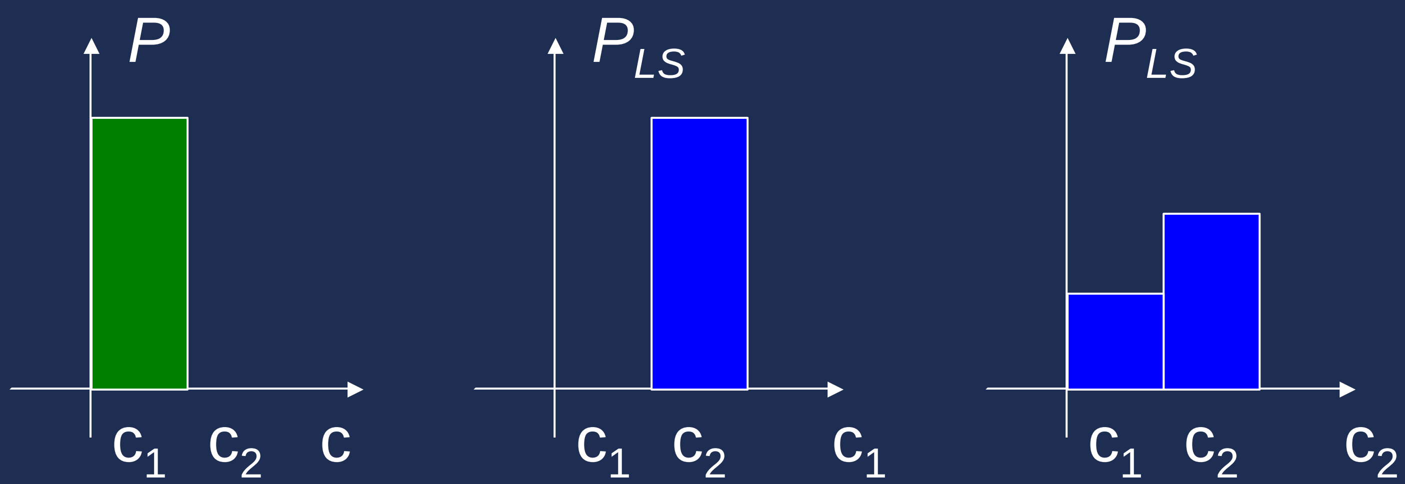
Residual error = $1 - P(c_{\mathrm{B}})$

Average model = c_{LS} = arg $\max_c P_{LS}(c)$

bias=$1(c_{\mathrm{B}} \neq c_{\mathrm{LS}})$

Classification problem (2)

- Important difference : A more unstable classification may be beneficial on biased cases (such that $c_B \neq c_{LS}$)
- Example: method 2 is better than method 1 although more variable



Content of the presentation

- Bias and variance definitions
- Parameters that influence bias and variance
  - Complexity of the model
  - Complexity of the Bayes model
  - Noise
  - Learning sample size
  - Learning algorithm
- Variance reduction techniques
- Decision tree induction

Illustrative problem

- Artificial problem with 10 inputs, all uniform random variables in [0,1]
- The true function depends only on 5 inputs:

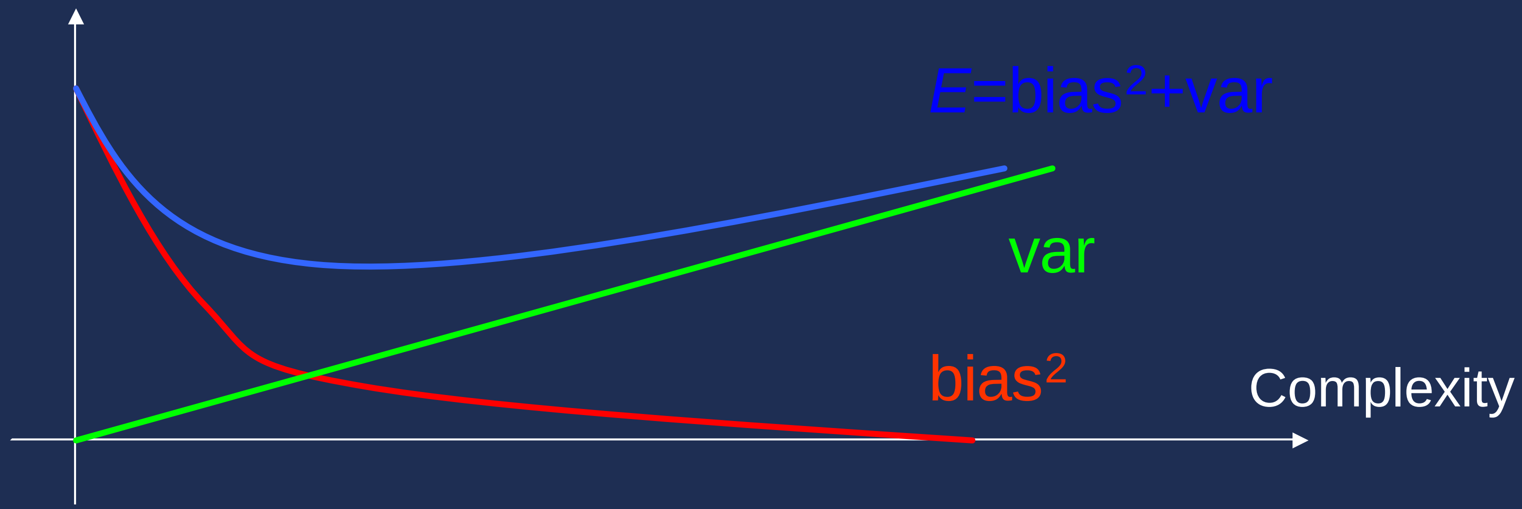
$$y(\mathbf{x})=10.\sin(\pi.x_1.x_2)+20.(x_3-0.5)^2+10.x_4+5.x_5+\varepsilon,$$

where ε is a N(0,1) random variable

- Experimentation:
  - $E_{LS} \Rightarrow$ average over 50 learning sets of size 500
  - $E_{\underline{\mathbf{x}},y} \Rightarrow$ average over 2000 cases

  $\Rightarrow$ Estimate variance and bias (+ residual error)

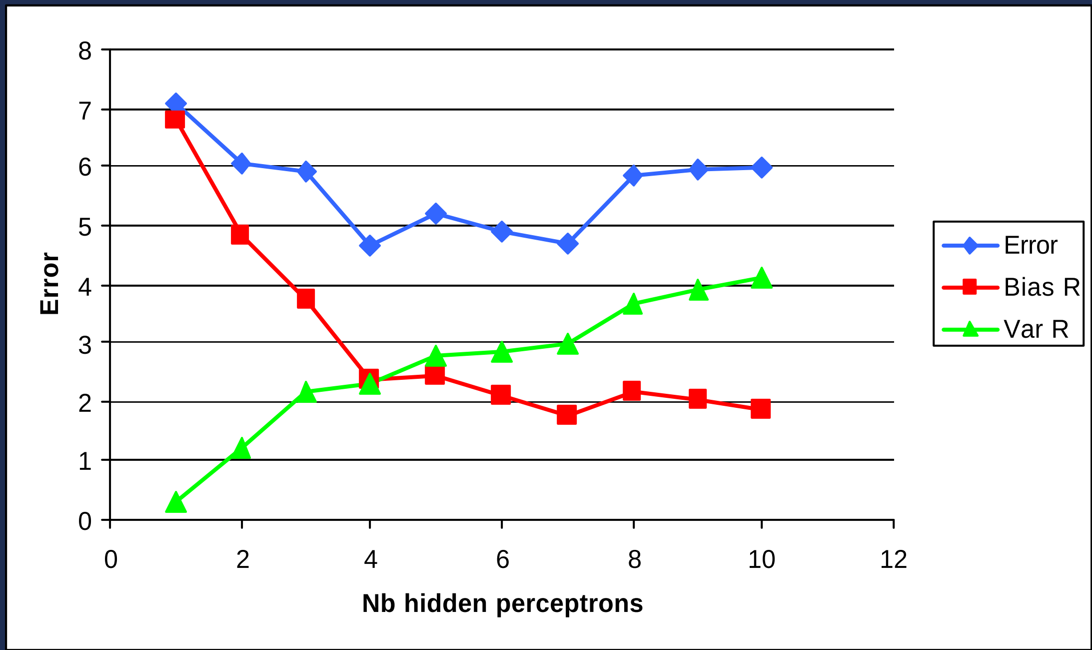
Complexity of the model



Usually, the bias is a decreasing function of the complexity, while variance is an increasing function of the complexity.

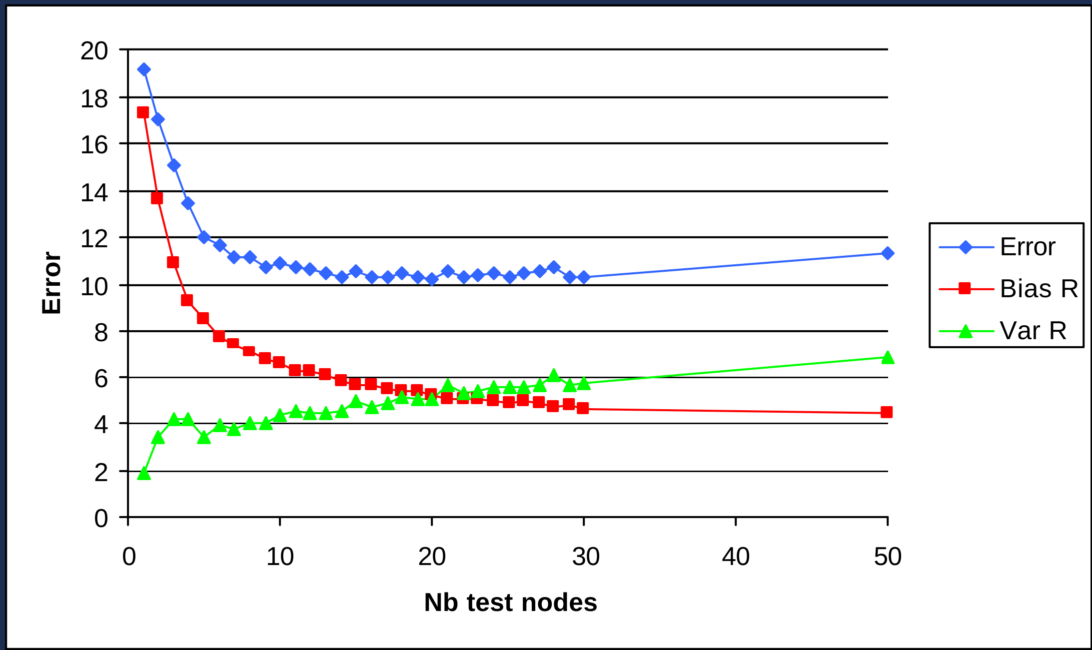
Complexity of the model – neural networks

- Error, bias, and variance w.r.t. the number of neurons in the hidden layer



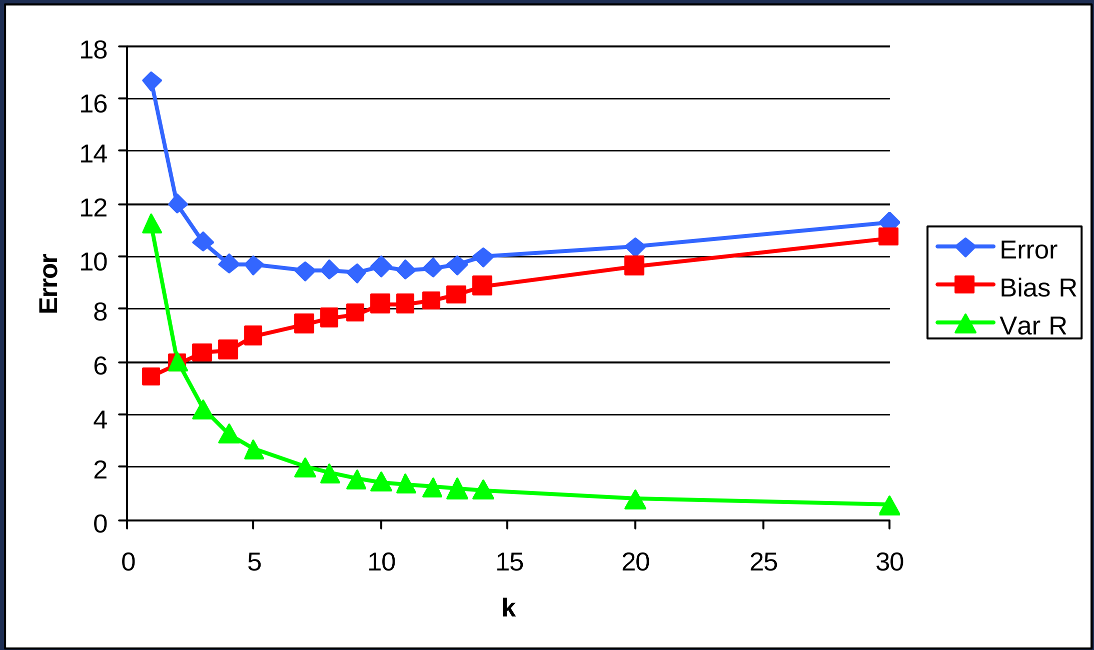
Complexity of the model – regression trees

- Error, bias, and variance w.r.t. the number of test nodes



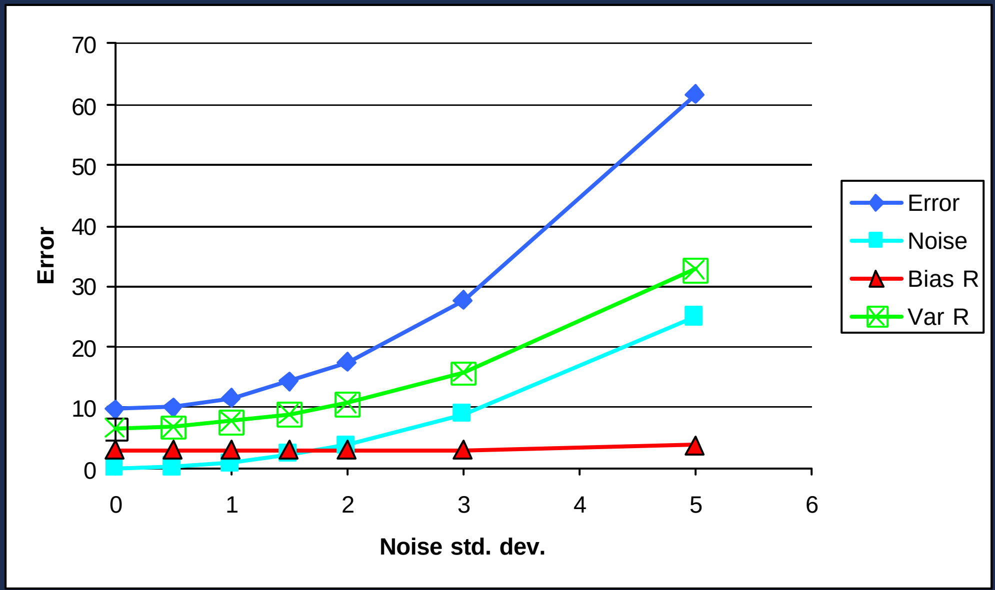
Complexity of the model – k-NN

- Error, bias, and variance w.r.t. k, the number of neighbors



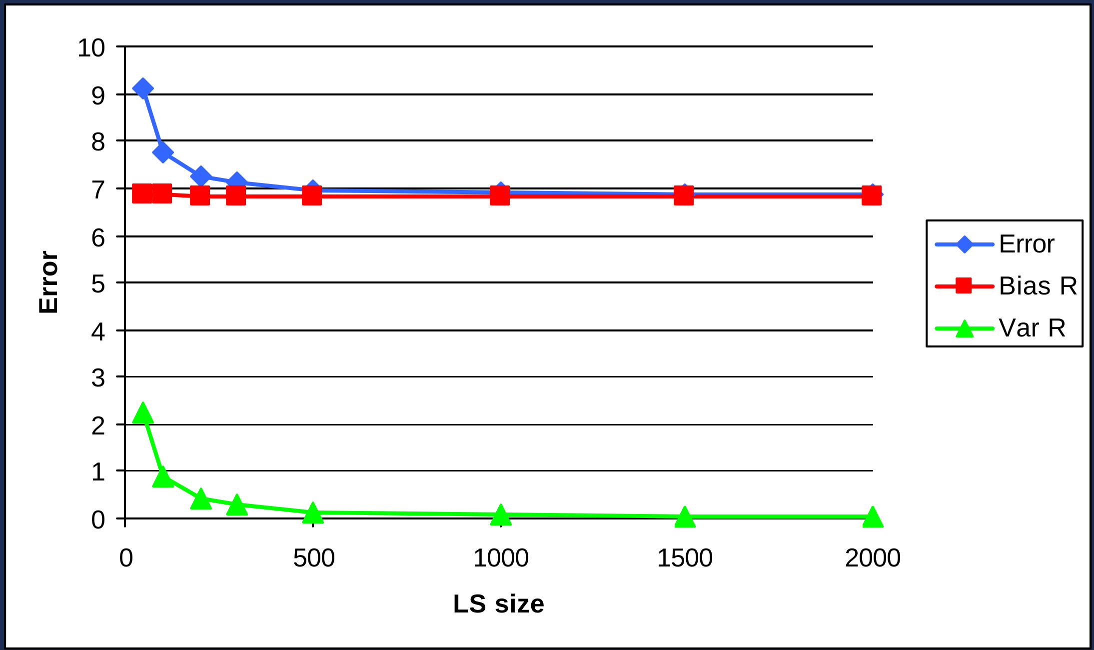
Learning problem

- Complexity of the Bayes model:
  - At fixed model complexity, bias increases with the complexity of the Bayes model. However, the effect on variance is difficult to predict.
- Noise:
  - Variance increases with noise and bias is mainly unaffected.
  - E.g. with regression trees



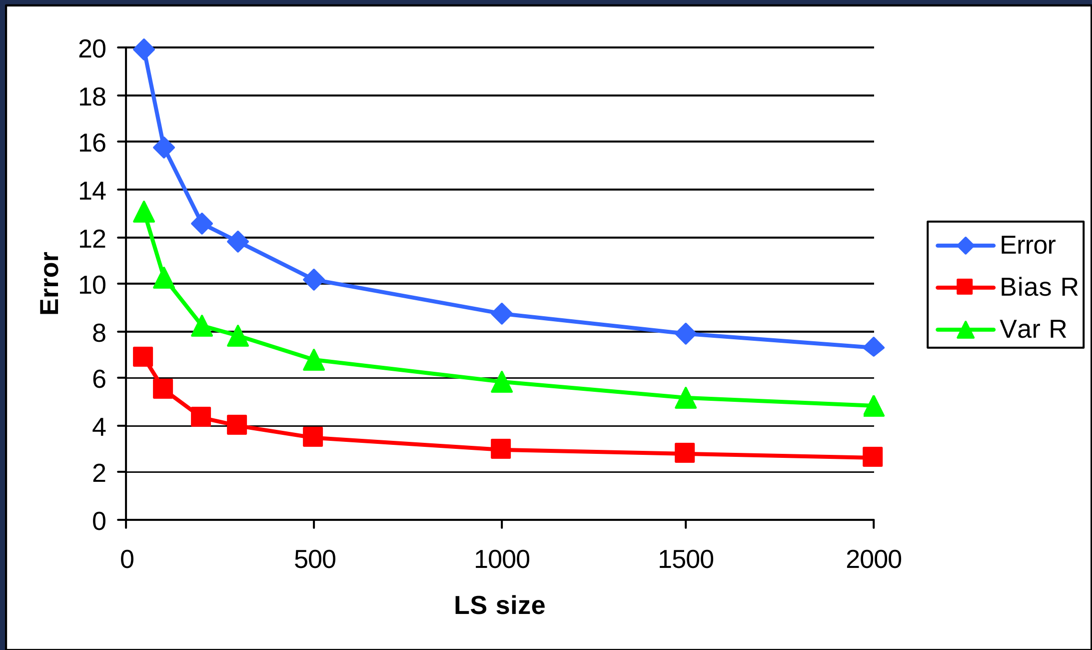
Learning sample size (1)

- At fixed model complexity, bias remains constant and variance decreases with the learning sample size. E.g. linear regression



Learning sample size (2)

- When the complexity of the model is dependant on the learning sample size, both bias and variance decrease with the learning sample size. E.g. regression trees



Learning algorithms – linear regression

Method	Err^2	$Bias^2$+Noise	Variance
Linear regr.	7.0	6.8	0.2
k-NN (k=1)	15.4	5	10.4
k-NN (k=10)	8.5	7.2	1.3
MLP (10)	2.0	1.2	0.8
MLP (10 – 10)	4.6	1.4	3.2
Regr. Tree	10.2	3.5	6.7

- Very few parameters : small variance
- Goal function is not linear : high bias

Learning algorithms – k-NN

Method	Err^2	$Bias^2$+Noise	Variance
Linear regr.	7.0	6.8	0.2
k-NN (k=1)	15.4	5	10.4
k-NN (k=10)	8.5	7.2	1.3
MLP (10)	2.0	1.2	0.8
MLP (10 – 10)	4.6	1.4	3.2
Regr. Tree	10.2	3.5	6.7

- Small k : high variance and moderate bias
- High k : smaller variance but higher bias

Learning algorithms - MLP

Method	Err^2	$Bias^2$+Noise	Variance
Linear regr.	7.0	6.8	0.2
k-NN (k=1)	15.4	5	10.4
k-NN (k=10)	8.5	7.2	1.3
MLP (10)	2.0	1.2	0.8
MLP (10 – 10)	4.6	1.4	3.2
Regr. Tree	10.2	3.5	6.7

- Small bias
- Variance increases with the model complexity

Learning algorithms – regression trees

Method	Err^2	$Bias^2$+Noise	Variance
Linear regr.	7.0	6.8	0.2
k-NN (k=1)	15.4	5	10.4
k-NN (k=10)	8.5	7.2	1.3
MLP (10)	2.0	1.2	0.8
MLP (10 – 10)	4.6	1.4	3.2
Regr. Tree	10.2	3.5	6.7

- Small bias, a (complex enough) tree can approximate any non linear function
- High variance (see later)

Content of the presentation

- Bias and variance definition
- Parameters that influence bias and variance
- Variance reduction techniques
  - Introduction
  - Dealing with the bias/variance tradeoff of one algorithm
  - Averaging techniques
- Decision tree induction

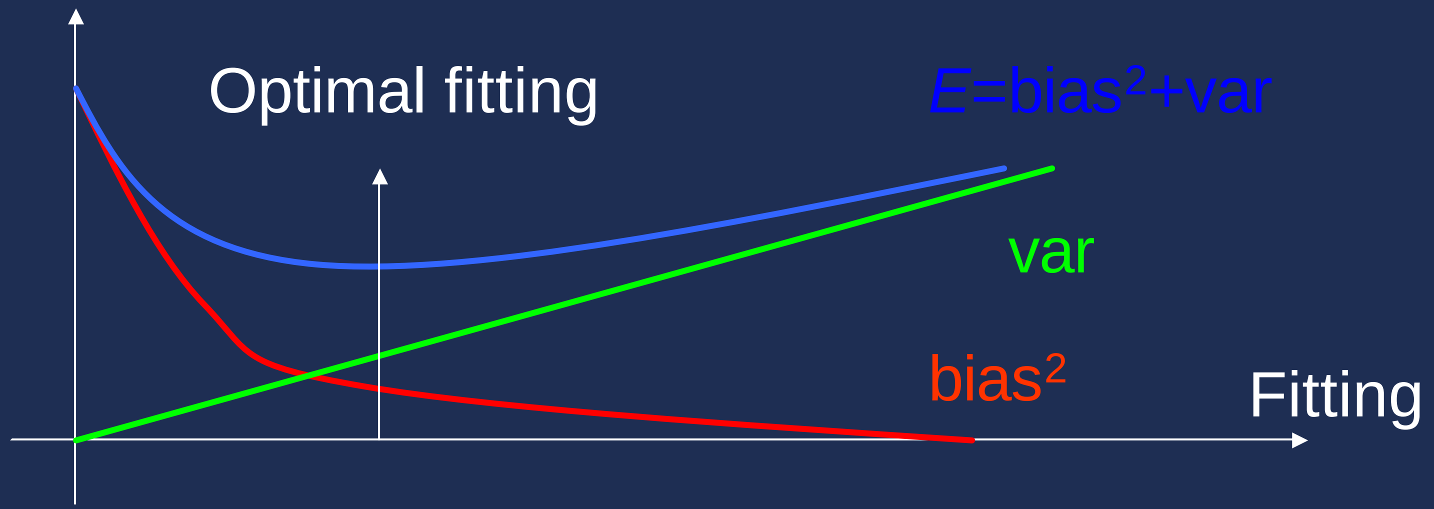
Variance reduction techniques

- In the context of a given method:
  - Adapt the learning algorithm to find the best trade-off between bias and variance.
  - Not a panacea but the least we can do.
  - Example: pruning, weight decay.
- Averaging techniques:
  - Change the bias/variance trade-off.
  - Universal but destroys some features of the initial method.
  - Example: bagging.

Variance reduction: 1 model (1)

- General idea: reduce the ability of the learning algorithm to over-fit the *LS*
  - Pruning
    - reduces the model complexity explicitly
  - Early stopping
    - reduces the amount of search
  - Regularization
    - reduce the size of hypothesis space

Variance reduction: 1 model (2)



- Bias^2 ≈ error on the learning set, E ≈ error on an independent test set
- Selection of the optimal level of fitting
  - a priori (not optimal)
  - by cross-validation (less efficient)

Variance reduction: 1 model (3)

- Examples:
  - Post-pruning of regression trees
  - Early stopping of MLP by cross-validation

Method	E	Bias	Variance
Full regr. Tree (488)	10.2	3.5	6.7
Pr. regr. Tree (93)	9.1	4.3	4.8
Full learned MLP	4.6	1.4	3.2
Early stopped MLP	3.8	1.5	2.3

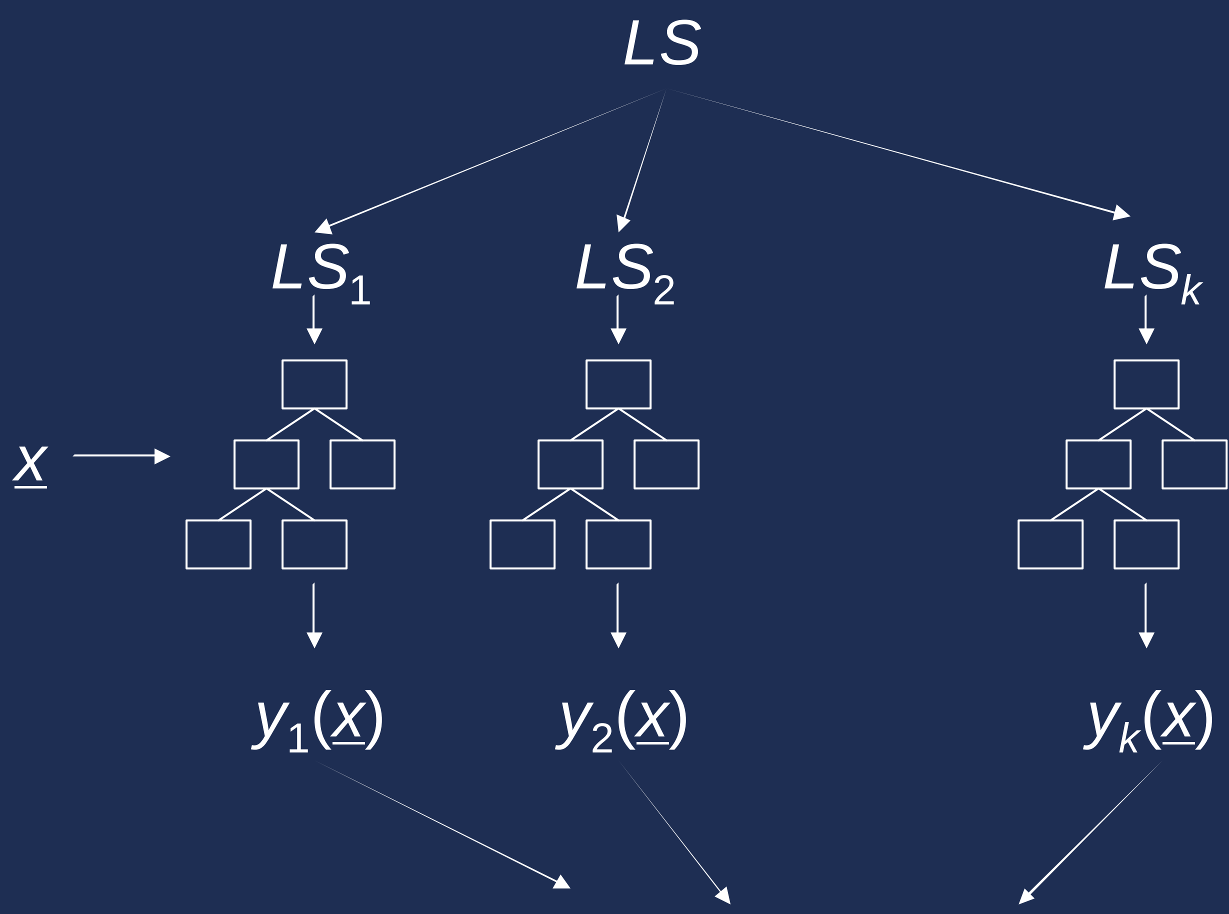
- As expected, reduces variance and increases bias

Variance reduction: bagging (1)

$$E_{LS}\{Err(\underline{x})\}=E_{y|\underline{x}}\{(y-h_B(\underline{x}))^2\}+(h_B(\underline{x})-E_{LS}\{y(\underline{x})\})^2+E_{LS}\{(y(\underline{x})-E_{LS}\{y(\underline{x})\})^2\}$$

- **Idea :** the average model $E_{LS}\{y(\underline{x})\}$ has the same bias as the original method but zero variance
- **Bagging** (**B**ootstrap **AGG**regat**ING**) :
  - To compute $E_{LS}\{y(\underline{x})\}$, we should draw an infinite number of *LS* (of size N)
  - Since we have only one single *LS*, we simulate sampling from nature by bootstrap sampling from the given *LS*
  - Bootstrap sampling = sampling with replacement of *N* objects from *LS* (*N* is the size of *LS*)

Variance reduction: bagging (2)



$$y(\underline{x}) = 1/\mathrm{k}.(y_1(\underline{x})+y_2(\underline{x})+\ldots+y_k(\underline{x}))$$

Variance reduction: bagging (3)

- Application to regression trees

Method	E	Bias	Variance
3 Test regr. Tree	14.8	11.1	3.7
Bagged	11.7	10.7	1.0
Full regr. Tree	10.2	3.5	6.7
Bagged	5.3	3.8	1.5

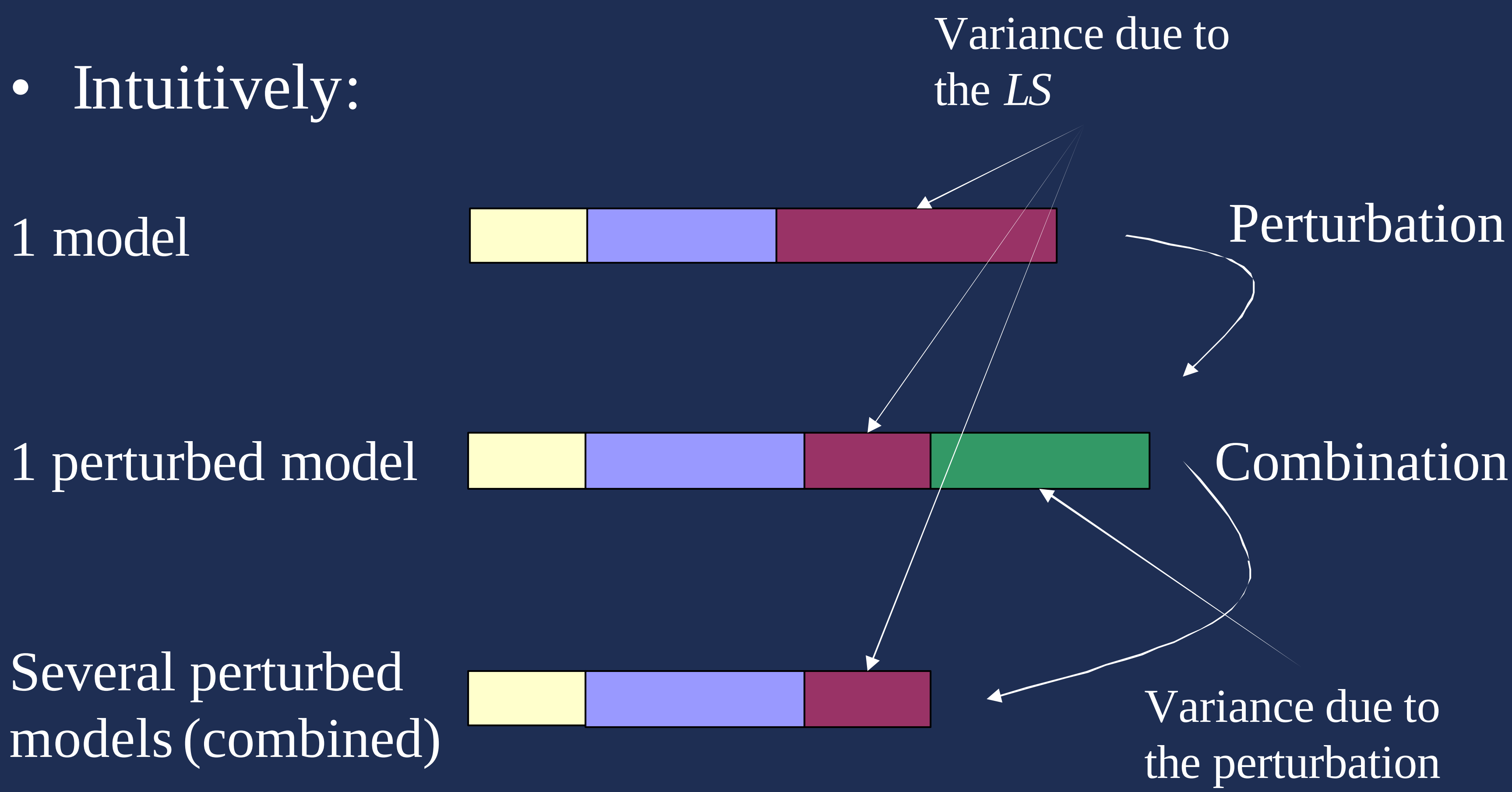
- Strong variance reduction without increasing the bias (although the model is much more complex than a single tree)

Variance reduction: averaging techniques

- Perturb and Combine paradigm:
  - Perturb the learning algorithm to obtain several models.
  - Combine the predictions of these models
- Examples:
  - Bagging: perturb learning sets.
  - Random trees: choose tests at random (see later).
  - Random initial weights for neural networks
  - …

Averaging techniques: how they work ?

- Intuitively:



- The effect of the perturbation is difficult to predict

Dual idea of bagging (1)

- Instead of perturbing learning sets to obtain several predictions, directly perturb the test case at the prediction stage
- Given a model $y(.)$ and a test case $\underline{x}$:
  - Form k attribute vectors by adding Gaussian noise to $\underline{x}$: $\{\underline{x}+\underline{e}_1, \underline{x}+\underline{e}_2, \ldots, \underline{x}+\underline{e}_k\}$.
  - Average the predictions of the model at these points to get the prediction at point x:

$$1/k.(y(\underline{x}+\underline{e}_1)+ y(\underline{x}+\underline{e}_2)+\ldots+ y(\underline{x}+\underline{e}_k)$$

- Noise level ? (variance of Gaussian noise) selected by cross-validation

Dual idea of bagging (2)

- With regression trees:

Noise level	E	Bias	Variance
0.0	10.2	3.5	6.7
0.2	6.3	3.5	2.8
0.5	5.3	4.4	0.9
2.0	13.3	13.1	0.2

- Smooth the function *y*(.).
- Too much noise increases bias. There is a (new) trade-off between bias and variance.

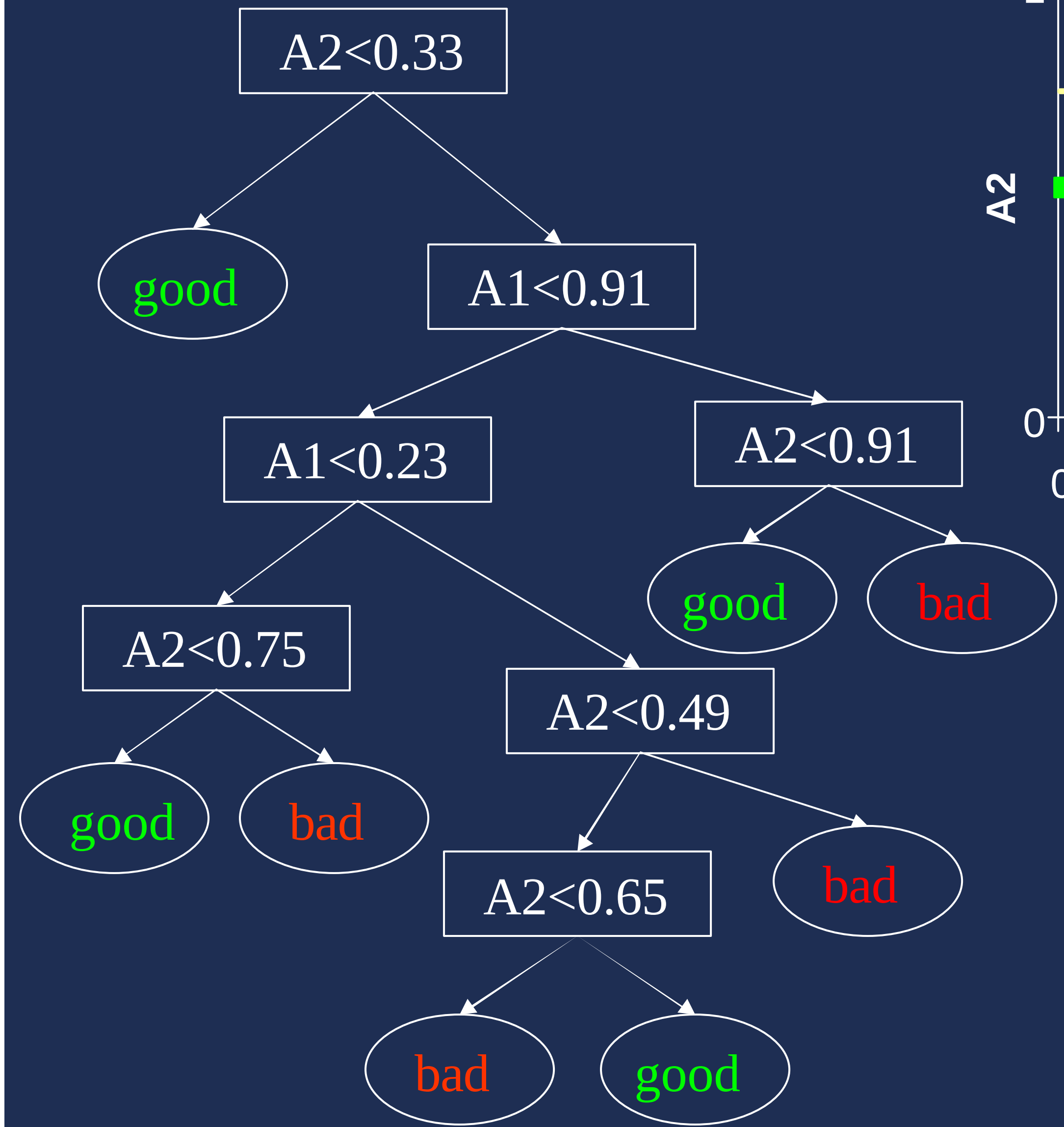
Conclusion

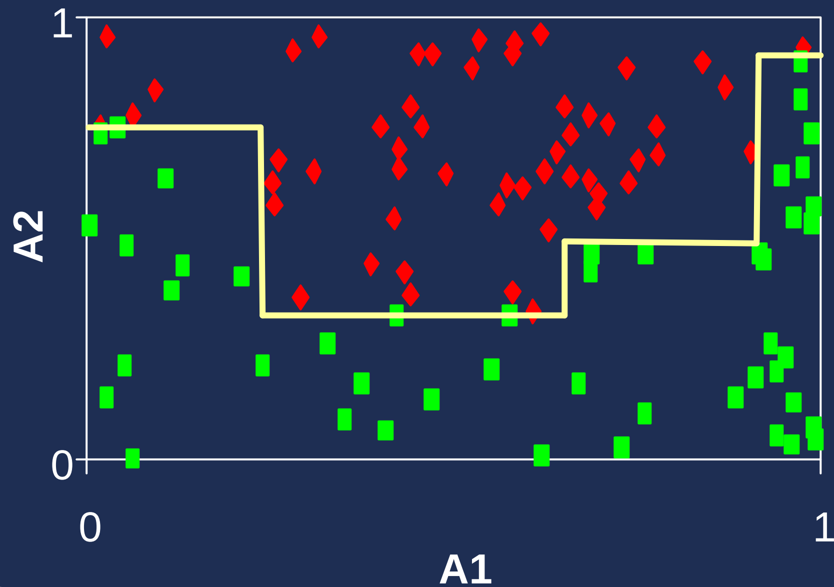
- Variance reduction is a very important topic:
  - To reduce bias is easy, but to keep variance low is not as easy.
  - Especially in the context of new applications of machine learning to very complex domains: temporal data, biological data, Bayesian networks learning…
- Interpretability of the model and efficiency of the method are difficult to preserve if we want to reduce variance significantly.
- Other approaches to variance reduction: Bayesian approaches, support vector machines

Content of the presentation

- Bias and variance definitions
- Parameters that influence bias and variance
- Variance reduction techniques
- Decision tree induction
  - Induction algorithm
  - Study of decision tree variance
  - Variance reduction methods for decision trees

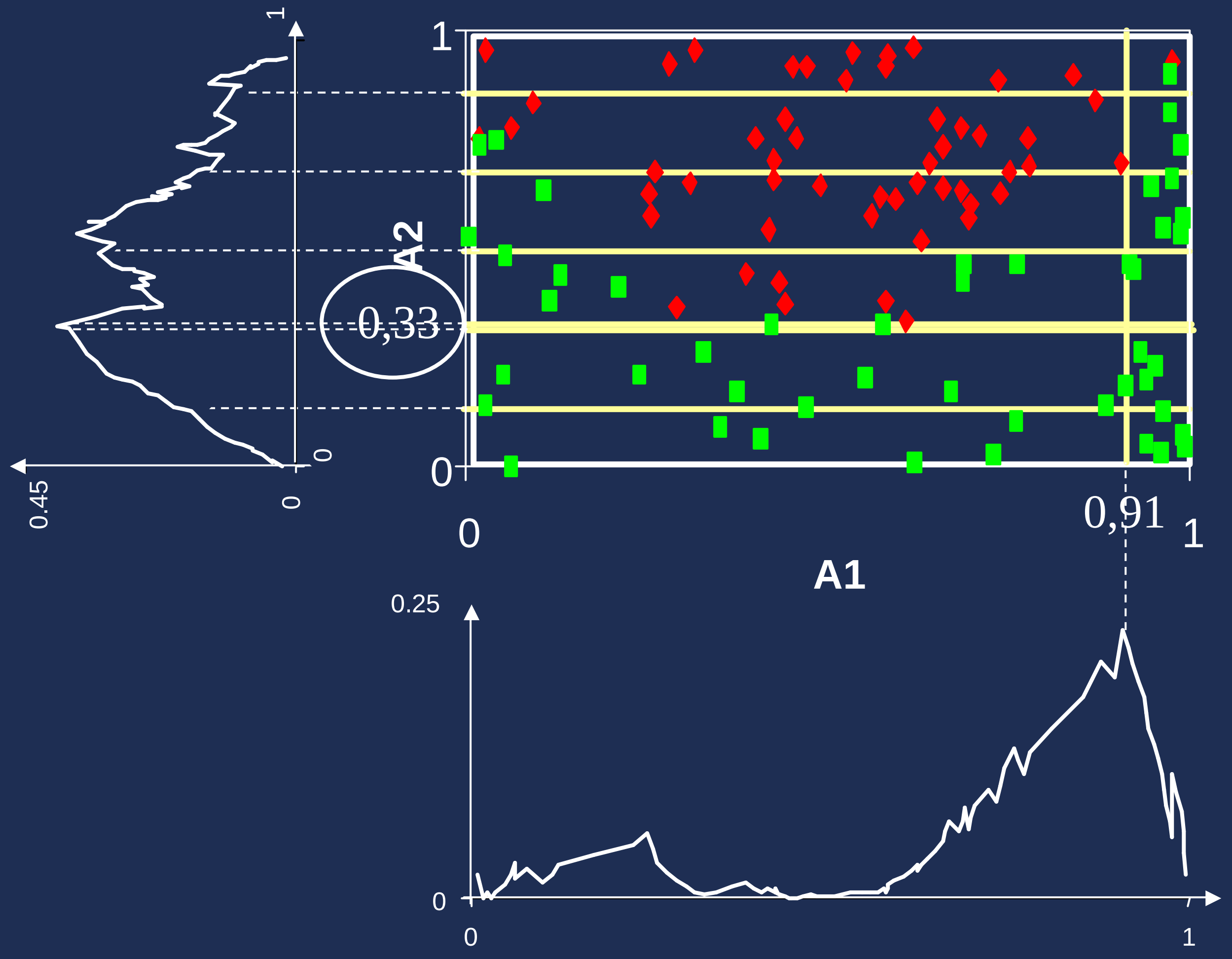
Arbre de décision: famille de modèle



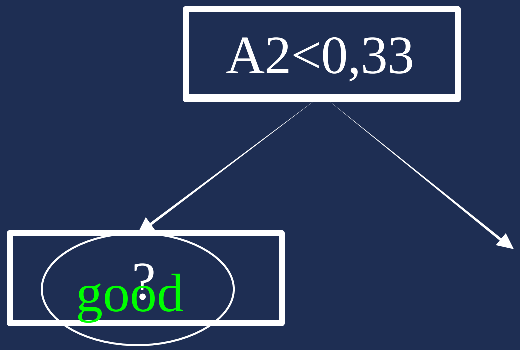


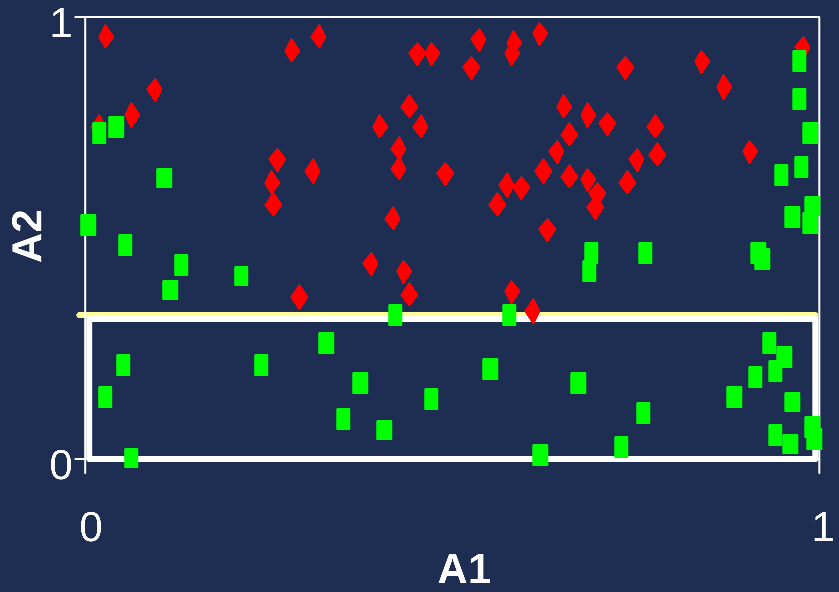
Arbre de décision: méthode d'induction

A2<0,33

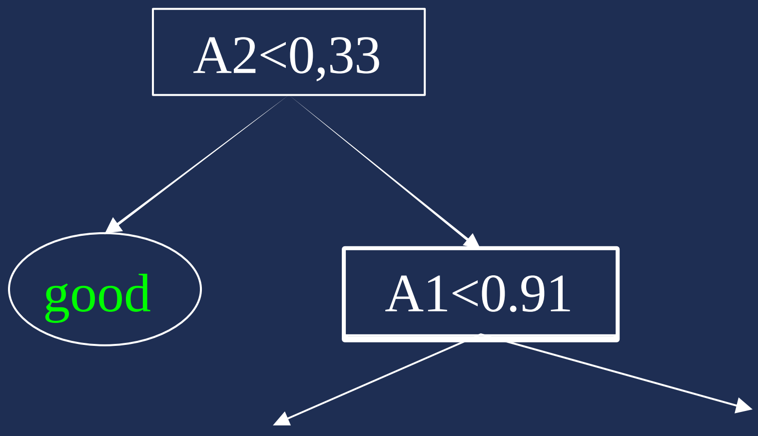


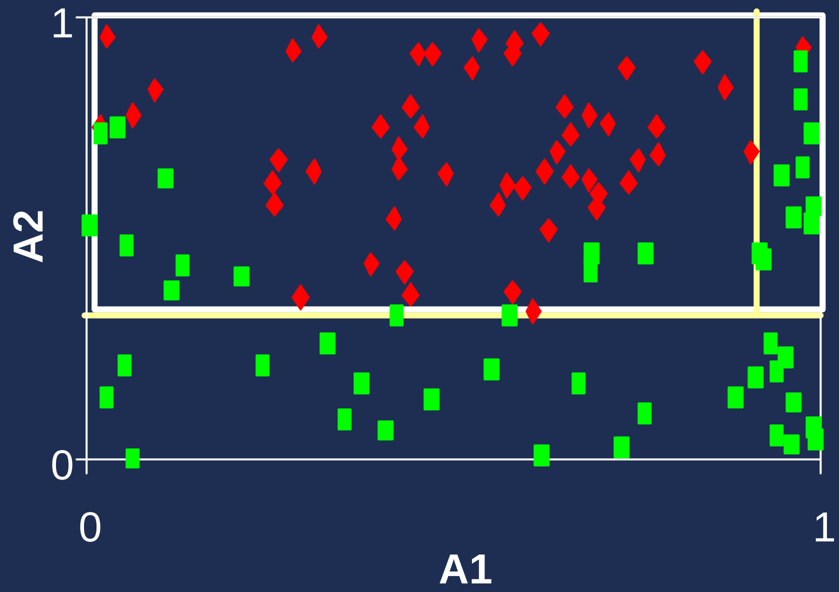
Arbre de décision: méthode d'induction



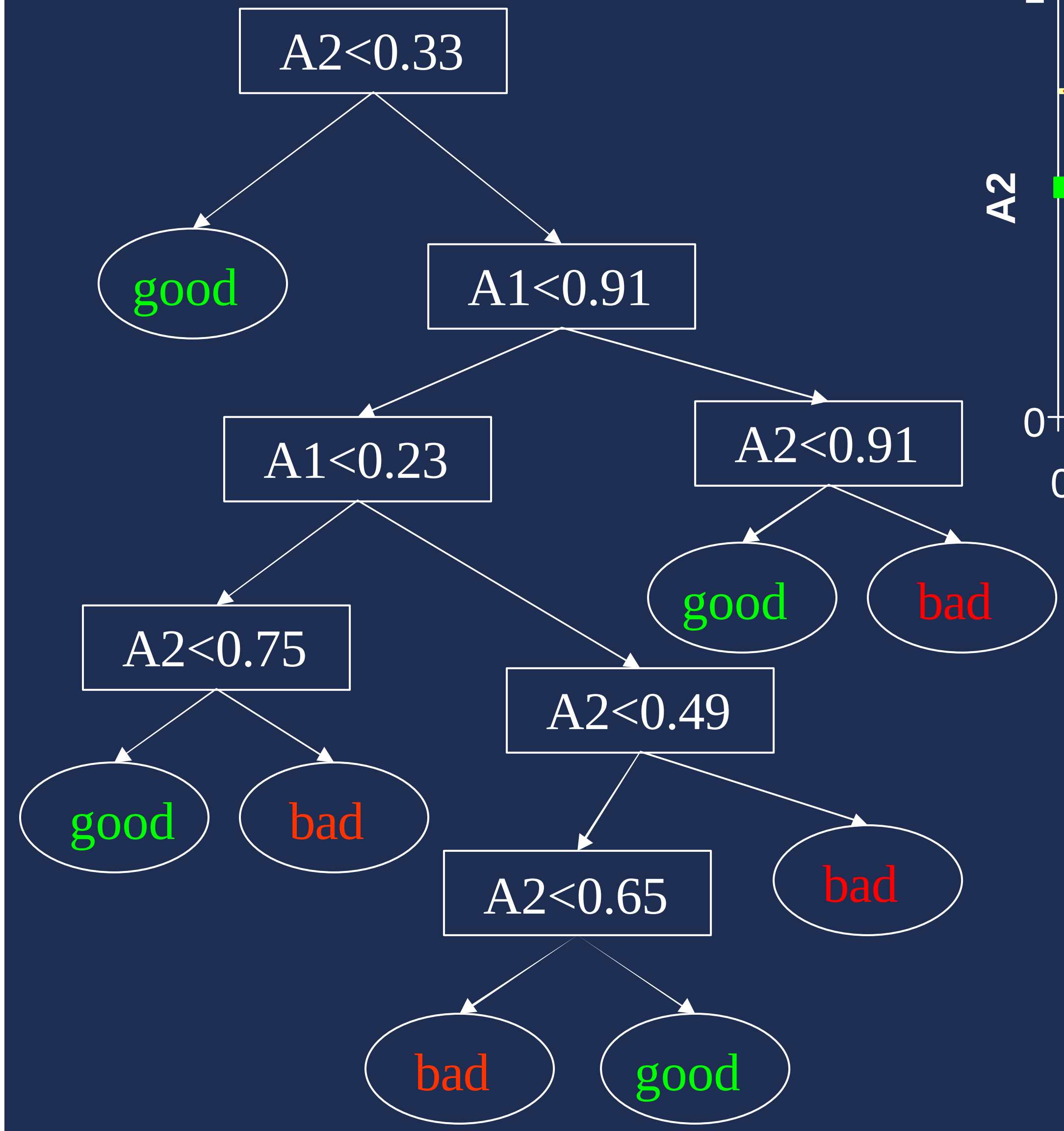


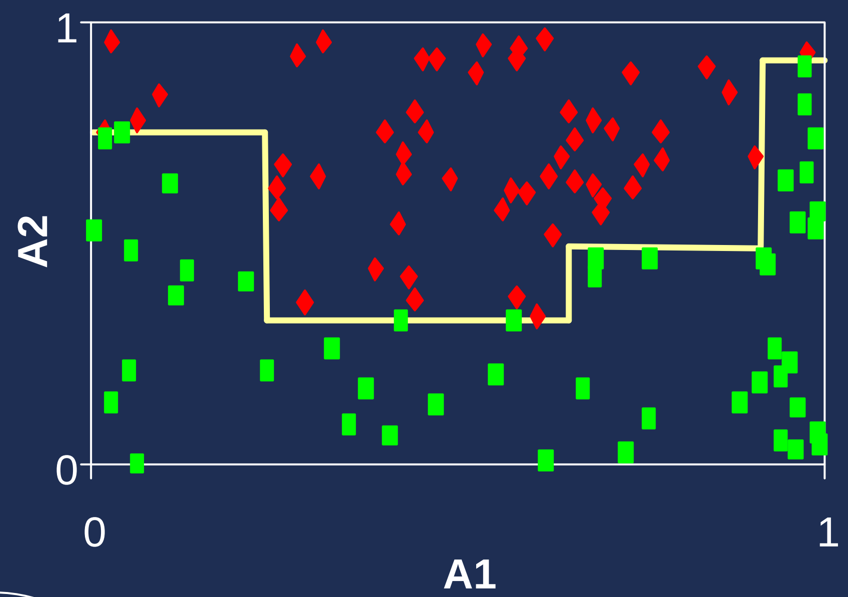
Arbre de décision: méthode d'induction





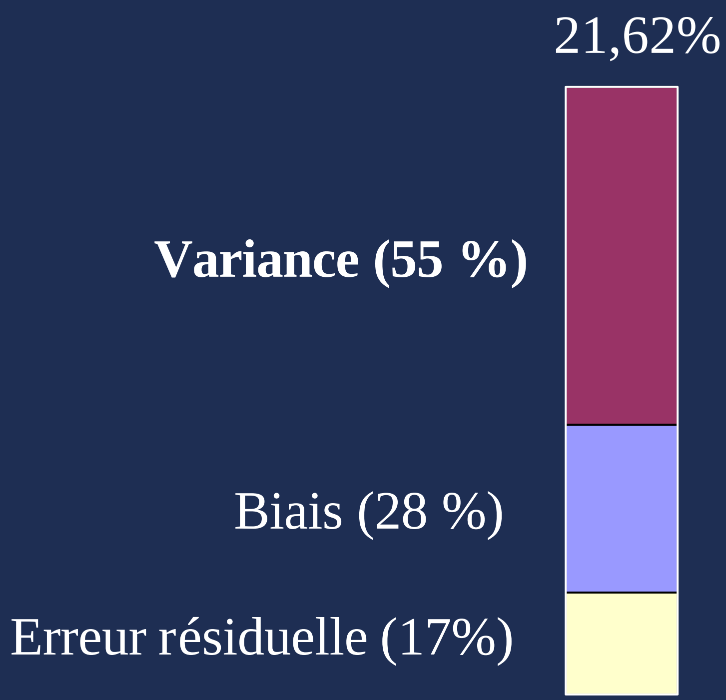
Arbre de décision: méthode d'induction





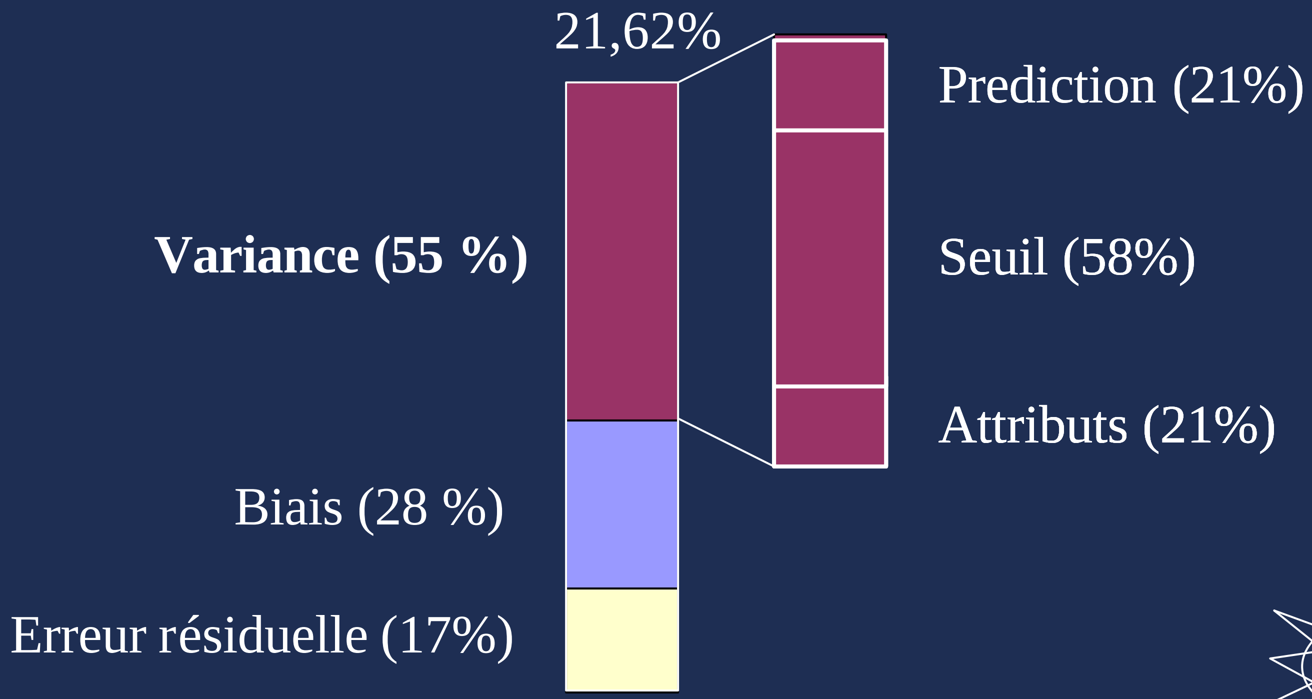
Impact de la variance sur l'erreur

- Estimation de l'erreur sur 7 problèmes différents
- Impact de la variance mesuré par la décomposition biais/variance



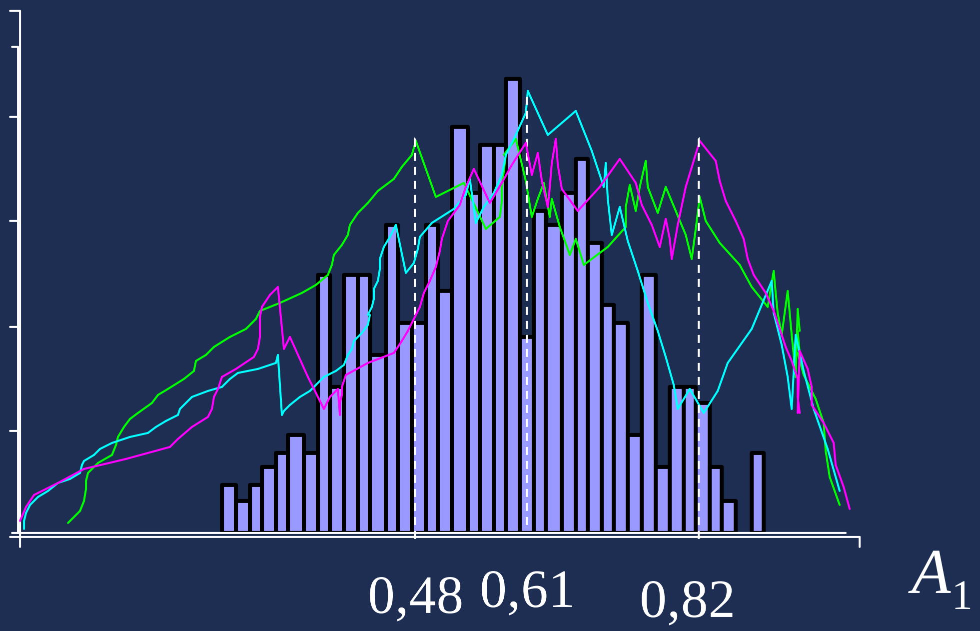
Impact de la variance sur l’erreur

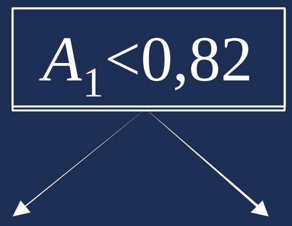
- Sources de variance = choix qui dépendent de l’échantillon



Variance des paramètres

- Expérimentations pour mettre en évidence la variabilité des paramètres avec l’échantillon
- Par exemple, le choix du seuil:





⇒Les paramètres sont très variables

⇒Remet en question l’interprétabilité de la méthode

Synthèse

	Précision	Interprétabilité	Efficacité
Arbres complets	**Moyen**	**Bon**	**Très bon**

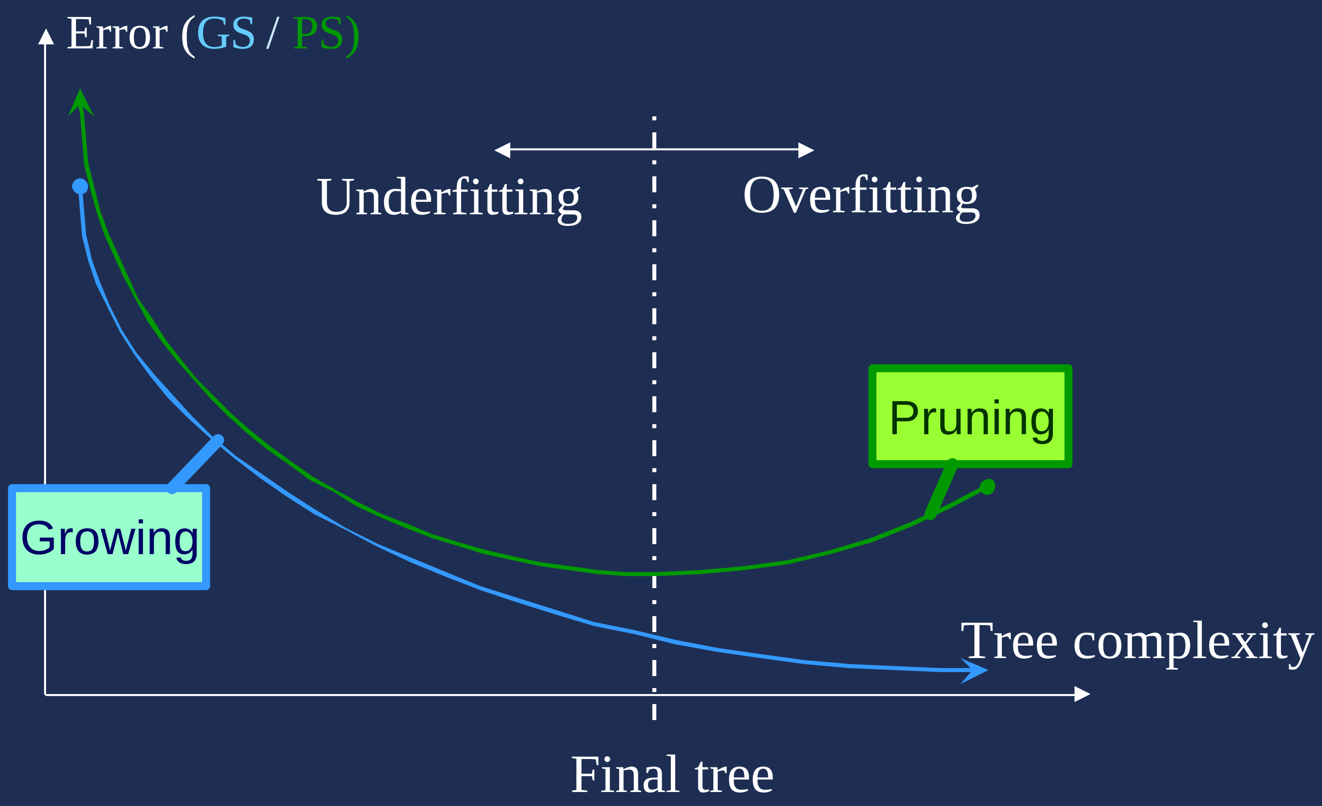
Méthode de réduction de variance

Trois approches:

- Améliorer l'interprétabilité d'abord
  - Élagage
  - Stabilisation des paramètres
- Améliorer la précision d'abord
  - Bagging
  - Arbres aléatoires
- Améliorer les deux (si possible)
  - Dual perturb and combine

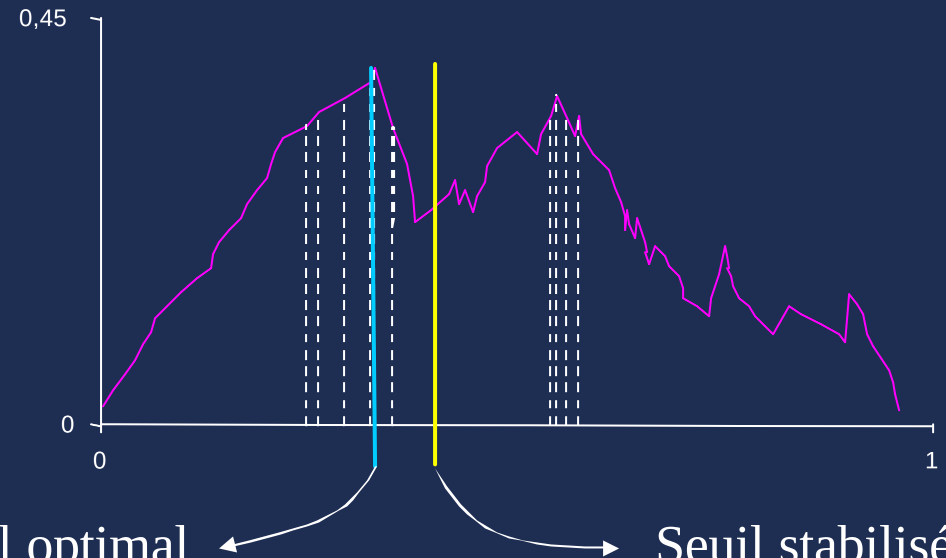
Élagage

- Détermine la taille appropriée de l ’arbre à l’aide d’un ensemble indépendant de l’ensemble d’apprentissage



Stabilisation des paramètres

- Plusieurs techniques pour stabiliser le choix des seuils de discrétisation et des attributs testés
- Un exemple de technique pour stabiliser le seuil: moyenne des *n* meilleurs seuils

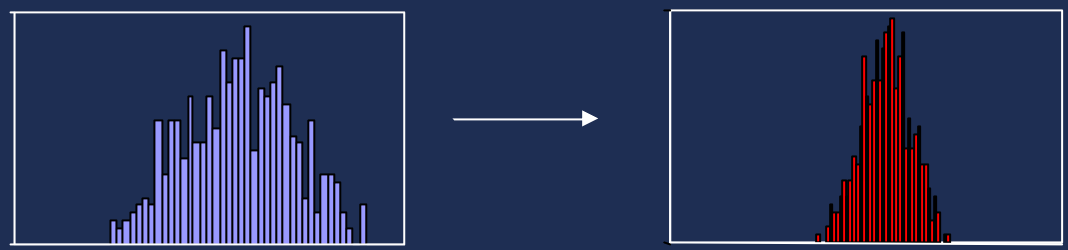


Seuil optimal

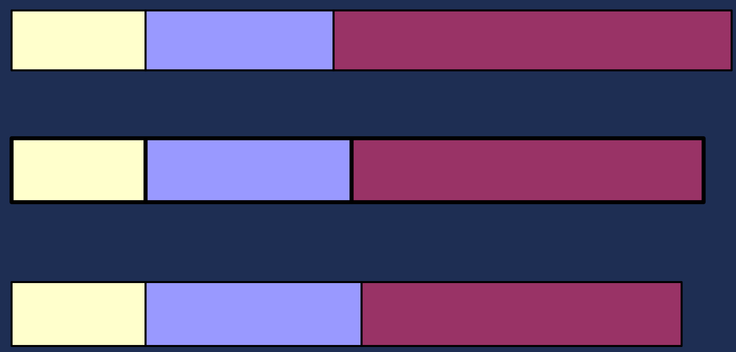
Seuil stabilisé

Stabilisation des paramètres

- Effet important sur l'interprétabilité:
  - L'élagage réduit la complexité de ..%
  - la stabilisation réduit la variance du seuil de 60 %



- Effet limité sur l'erreur: variance ↓ mais biais ↑



Arbre complet 21,61 %

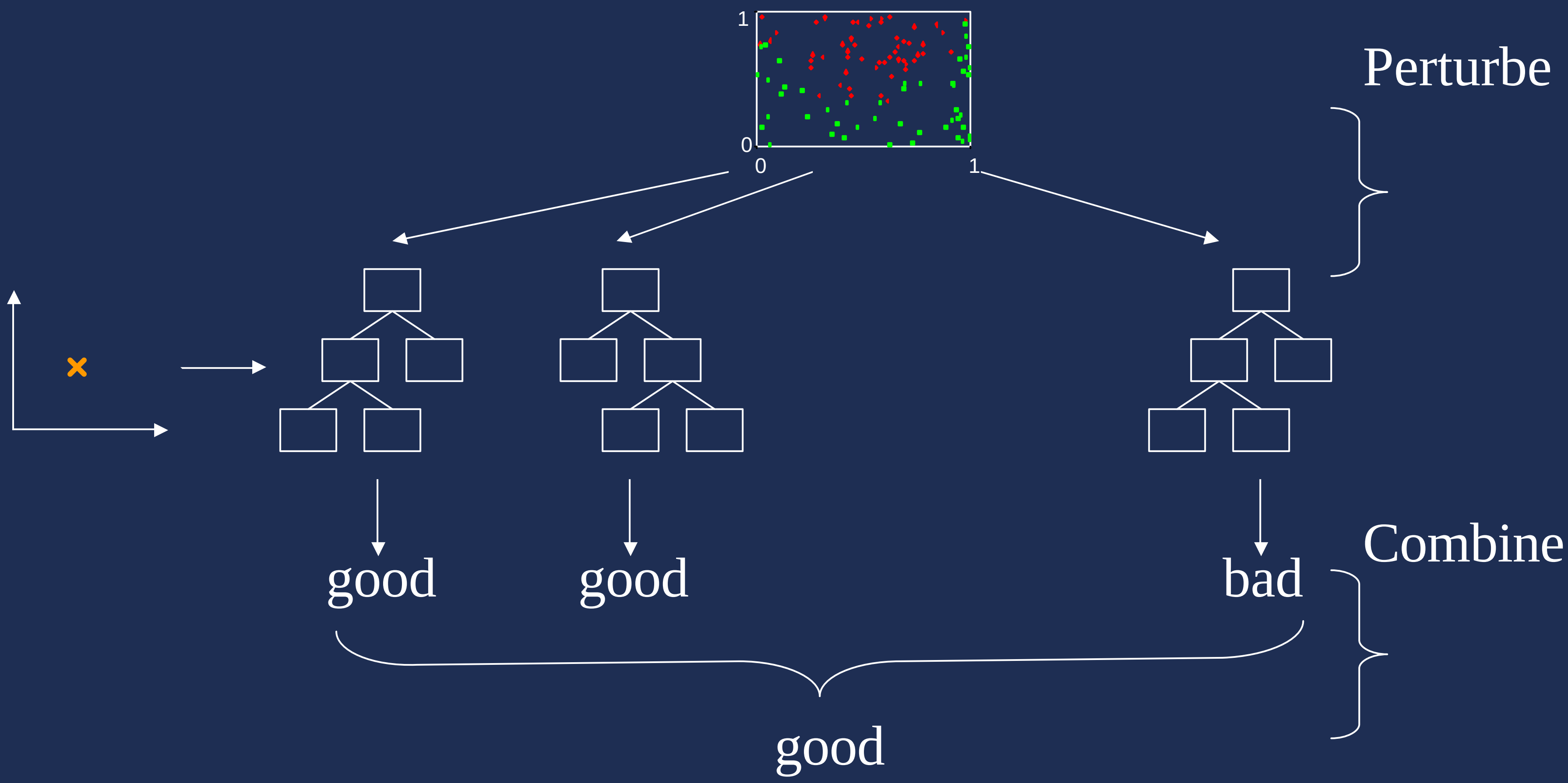
Arbre élagué 20,65 %

Élagage + Stabilisation 20,05 %

Synthèse

	Précision	Interprétabilité	Efficacité
Arbres complets	Moyen	Bon	Très bon
Élagage+Stabilisation	**Moyen**	**Très bon**	**Très bon**

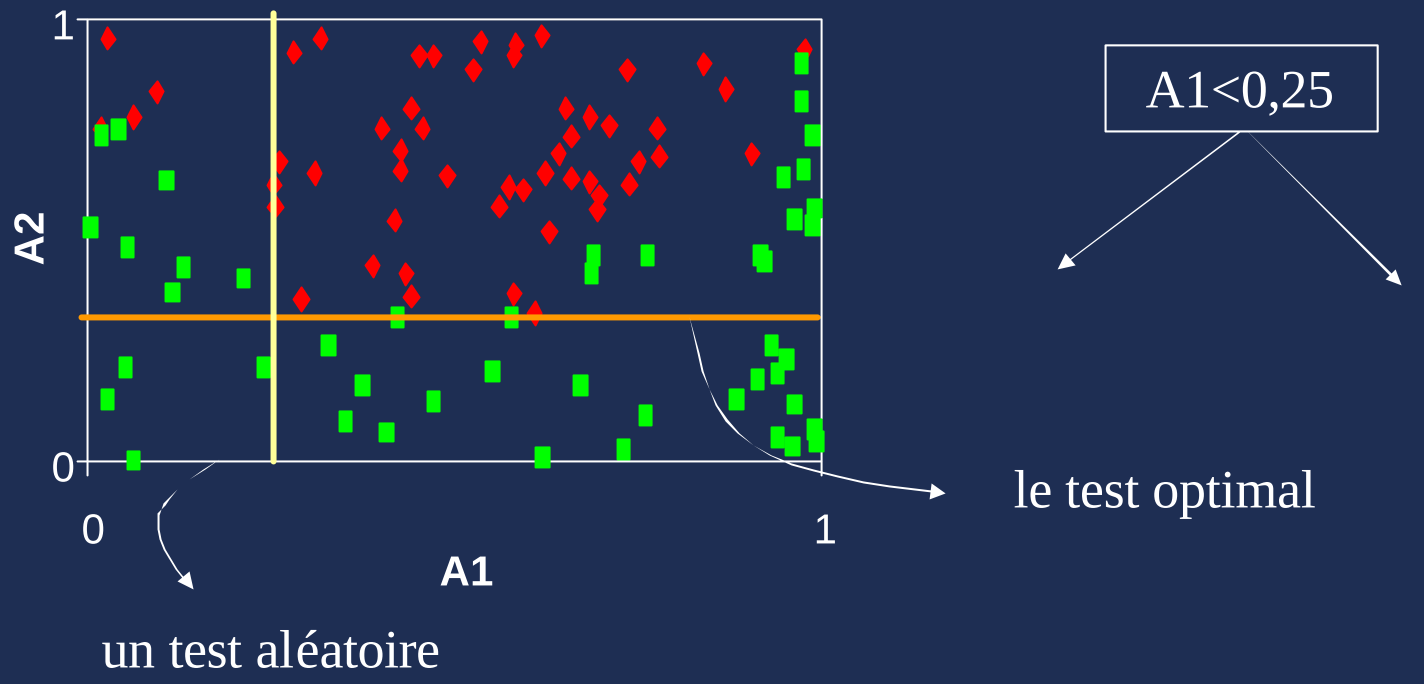
Agrégation de modèles



Exemple: le **bagging** utilise le rééchantillonnage

Arbres aléatoires: induction

- “Imite” la très grande variance des arbres en tirant un attribut et un seuil au hasard

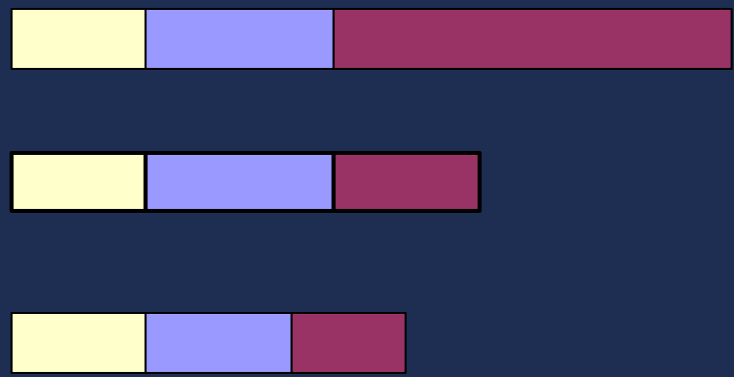


⇒ On agrège plusieurs arbres aléatoires

Arbres aléatoires: évaluation

- Effet sur la précision:

Arbre complet		21,61%
Bagging		14,02%
25 arbres aléatoires		11,84%

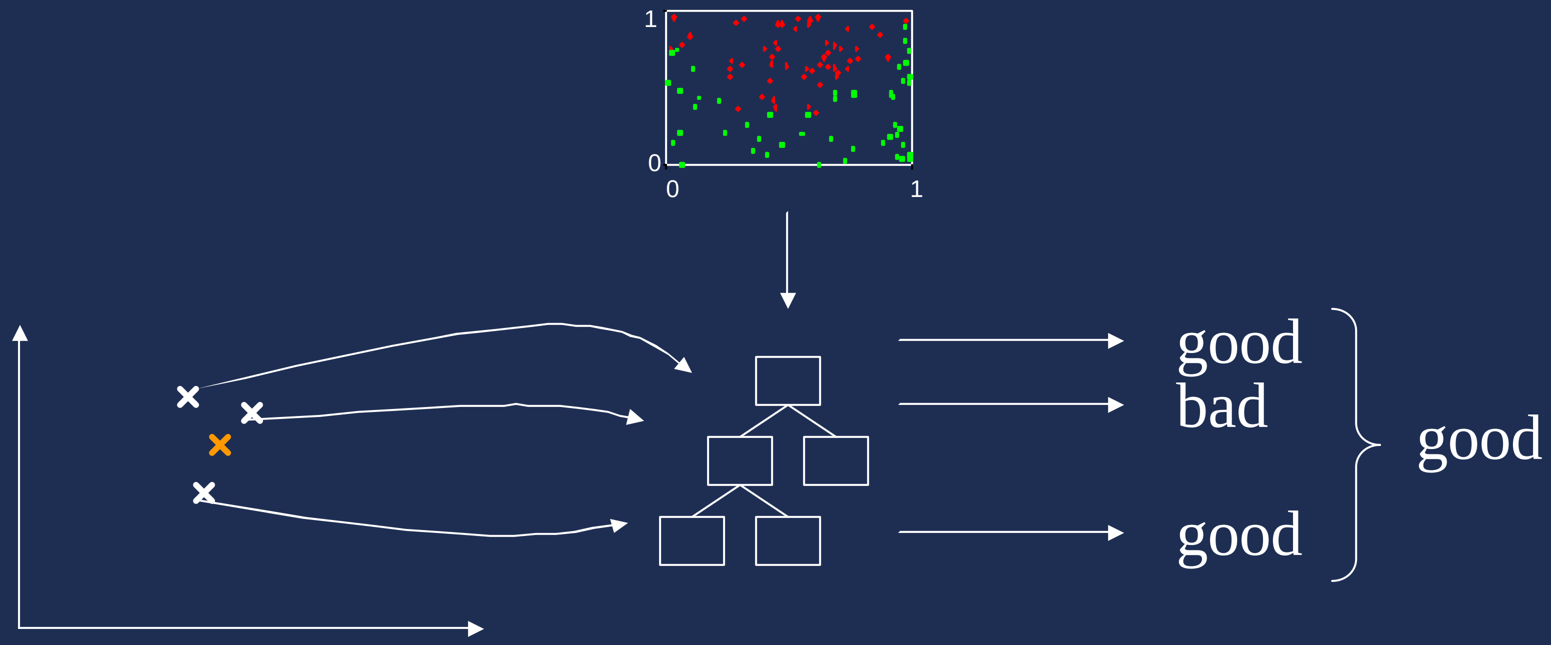


- Diminution de l'erreur due essentiellement à une diminution de la variance

Synthèse

	Précision	Interprétabilité	Efficacité
Arbres complets	Moyen	Bon	Très bon
Élagage + Stabilisation	Moyen	Très bon	Très bon
Bagging	Très bon	Mauvais	Moyen
Arbres aléatoires	**Très bon**	**Mauvais**	**Très bon**

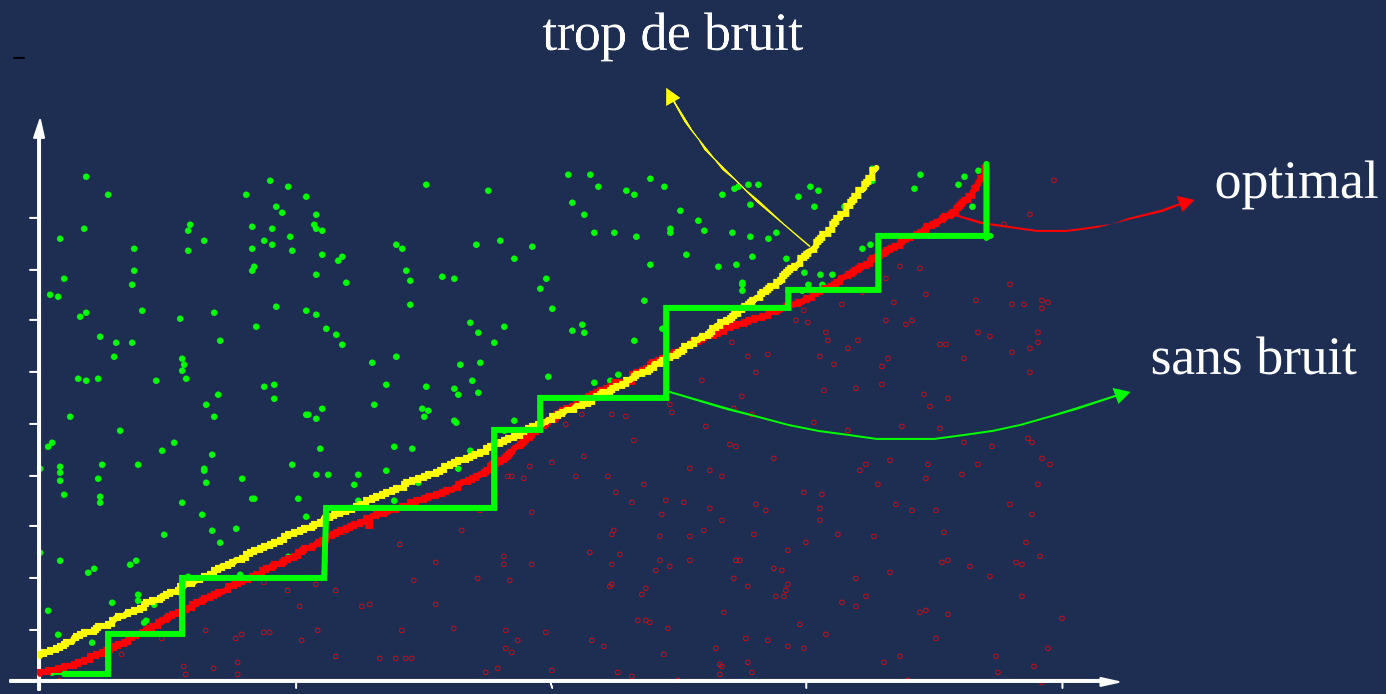
Dual Perturb and Combine



- Perturbation lors du test avec un seul modèle
- Ajout d'un bruit Gaussien indépendant à chacune des coordonnées

Dual Perturb and Combine

- Compromis biais/variance en fonction du niveau de bruit



- Détermination du niveau de bruit optimal sur un échantillon indépendant

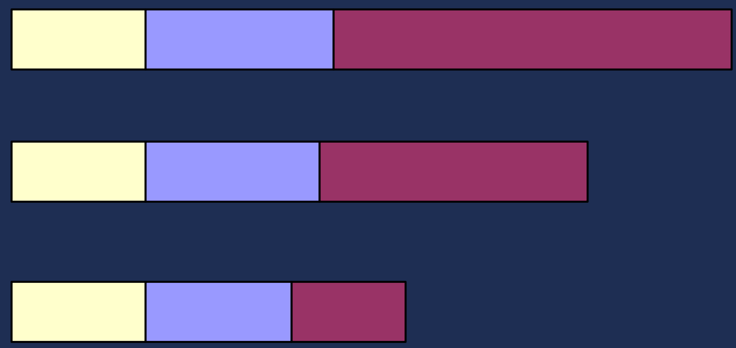
Dual Perturb and Combine

- Résultats en terme de précision

Arbre complet — 21,61%

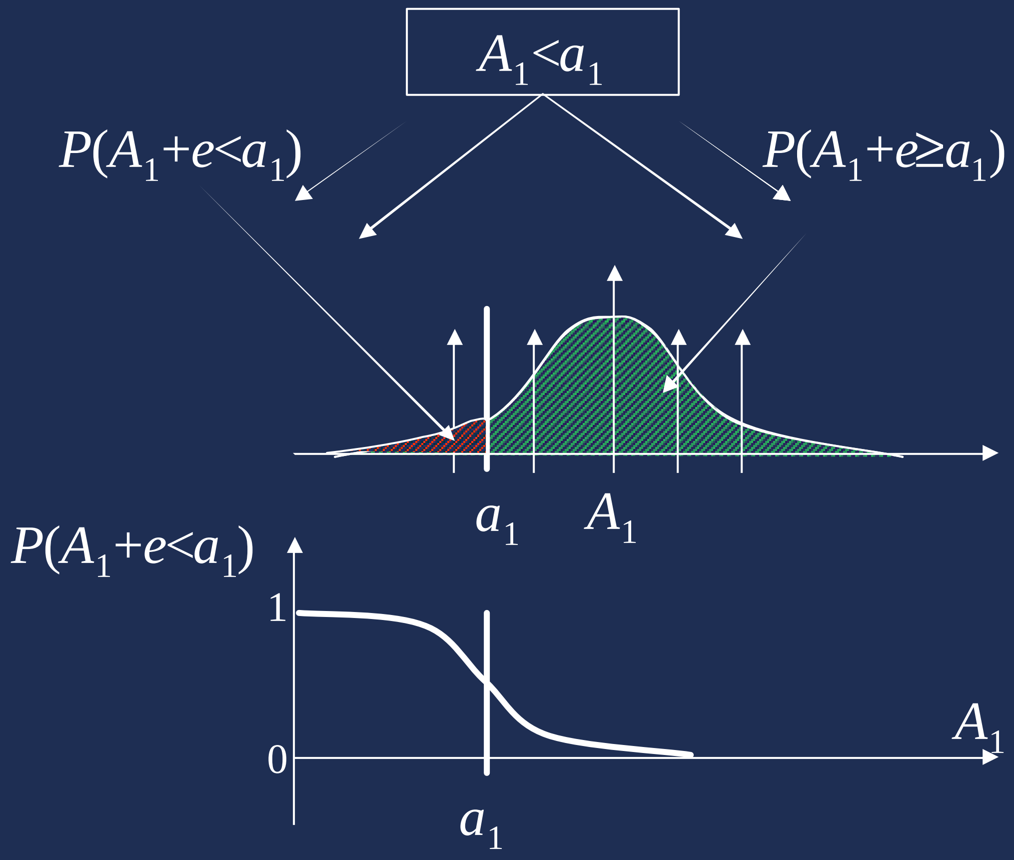
Dual P&C — 17,26%

25 arbres aléatoires — 11,84%

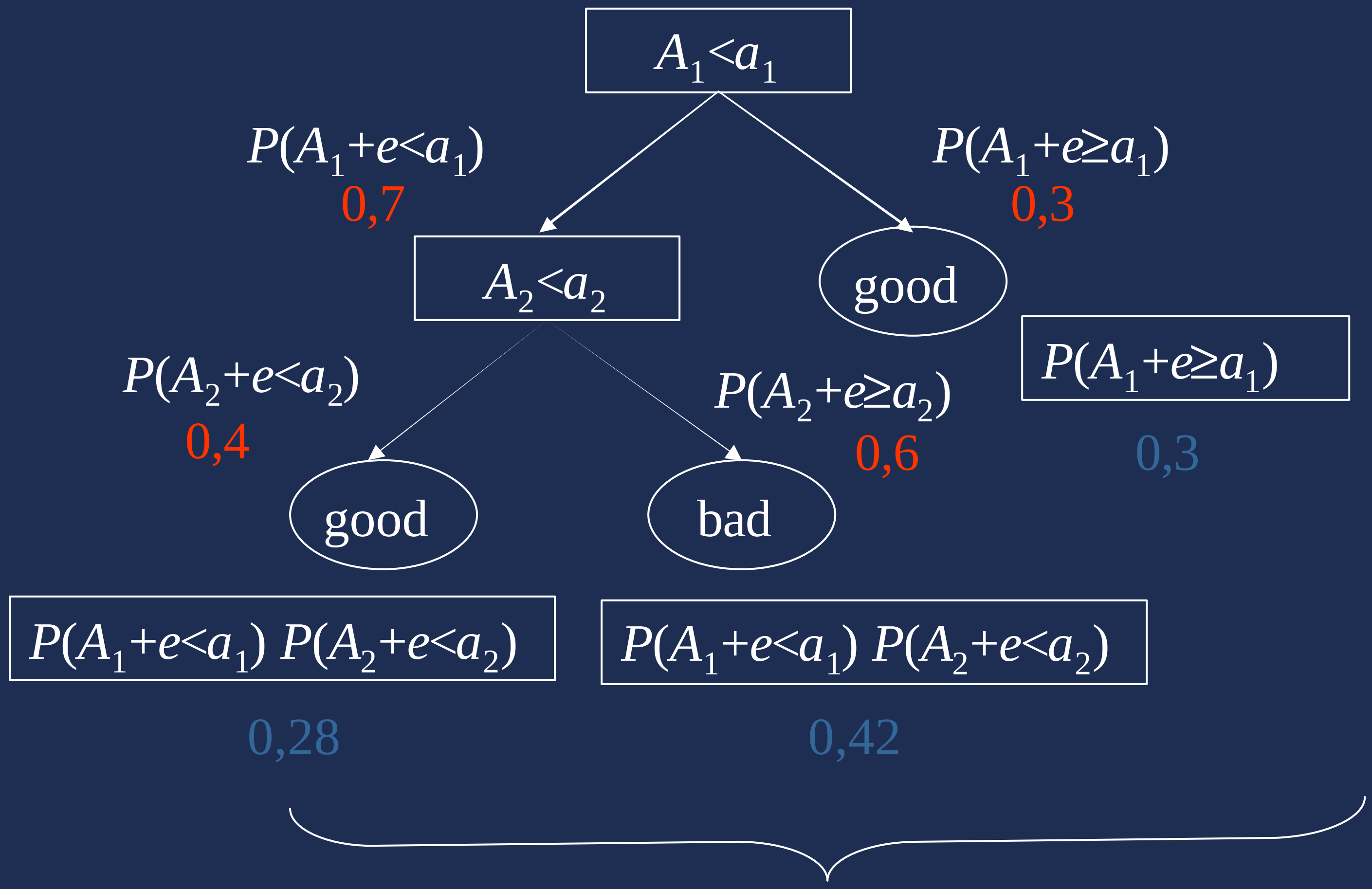


- Impact: réduction de la variance essentiellement
- Entre les arbres et les arbres aléatoires

Dual P&C = arbres flous



Dual P&C = arbres flous



Good (0,58 contre 0,42)

Synthèse

	Précision	Interprétabilité	Efficacité
Arbres complets	Moyen	Bon	Très bon
Élagage + Stabilisation	Moyen	Très bon	Très bon
Bagging	Très bon	Mauvais	Moyen
Arbres aléatoires	Très bon	Mauvais	Très bon
Dual P&C	**Bon**	**Bon**	**Bon**

Synthèse

	Précision	Interprétabilité	Efficacité
Arbres complets	Moyen	Bon	Très bon
Élagage + Stabilisation	Moyen	**Très bon**	Très bon
Bagging	Très bon	Mauvais	Moyen
Arbres aléatoires	**Très bon**	Mauvais	**Très bon**
Dual P&C	**Bon**	**Bon**	Bon