Analyse de données au LHC

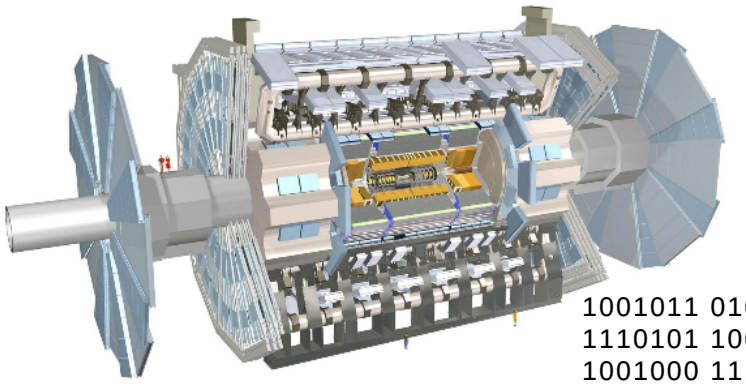
Nikola Makovec

Ecole thématique IN2P3 d'instrumentation
"De la physique au détecteur" 2017

LABORATOIRE
DE L'ACCÉLÉRATEUR
LINÉAIRE

Des données brutes à la publication

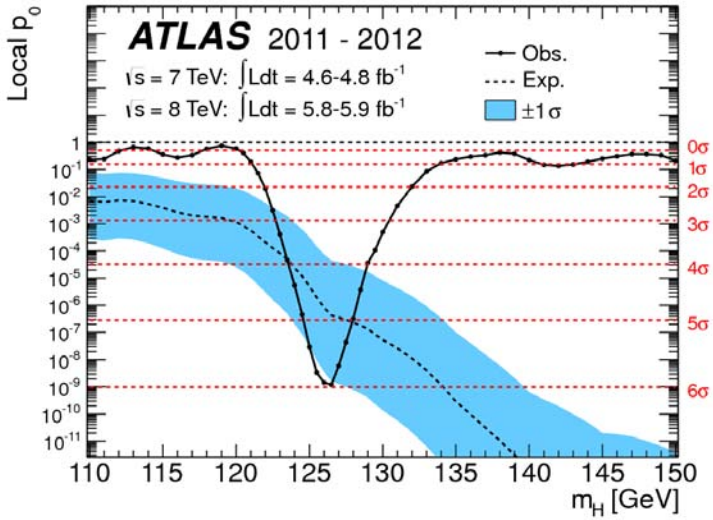
Détecteurs
(Cours P. Puzo)



1001011 01000101
1110101 10001110
1001000 11101110
1101000 10001001
1010110 00100010
...

Analyse
de données

Physique
(Cours F. Ledroit)

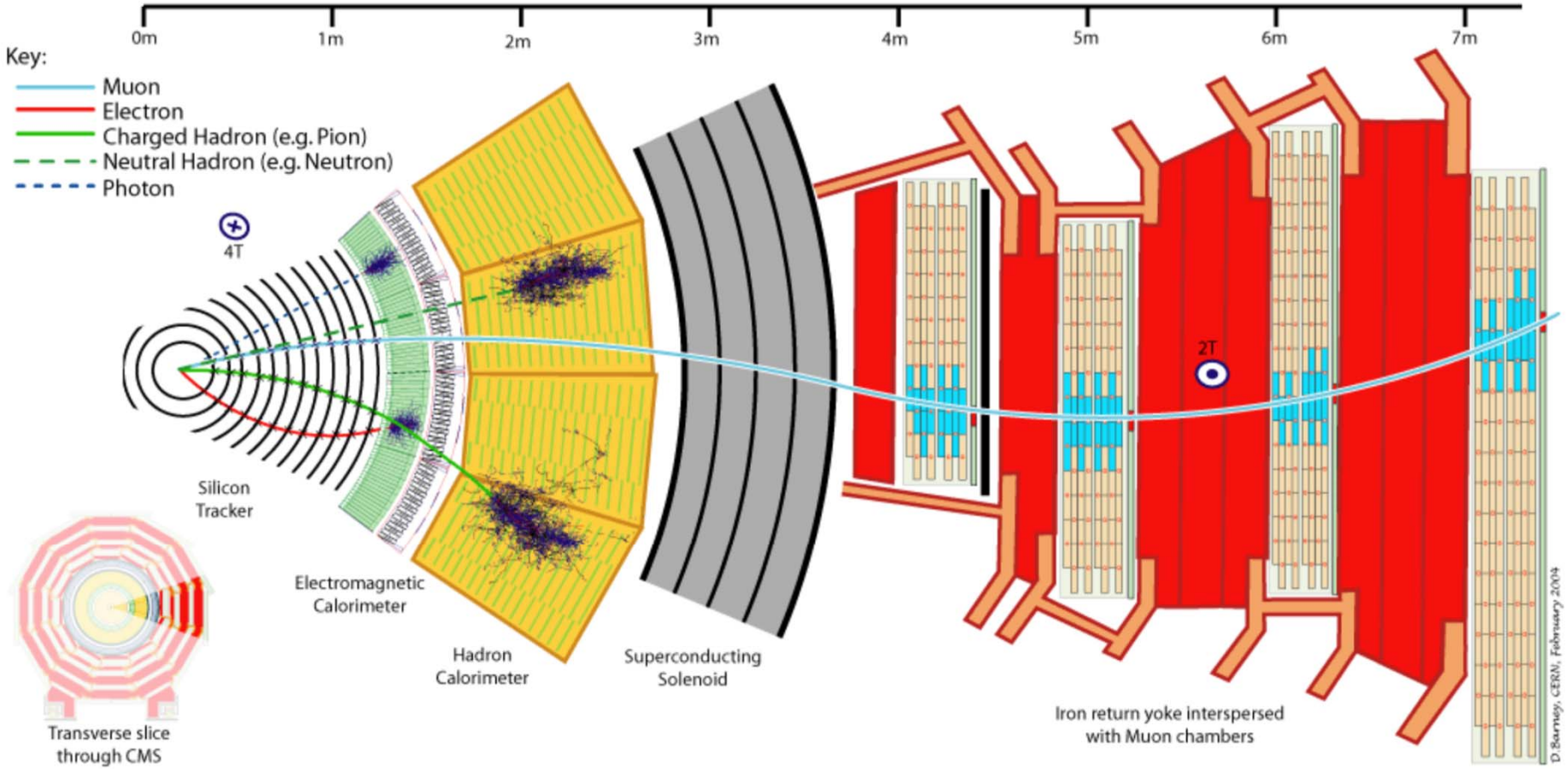


- Plan:
  - La reconstruction: vue d'ensemble
  - La grille
  - **La simulation**
  - **La reconstruction: l'exemple des électrons**
  - Qualité des données
  - Mesure de la section efficace de prod. du boson Z
  - La recherche du boson de Higgs
  - Organisation

Parenthèses statistiques:

- La méthode Monte-Carlo
- Estimation de paramètres
- Analyse multivariée
- Calcul de significance

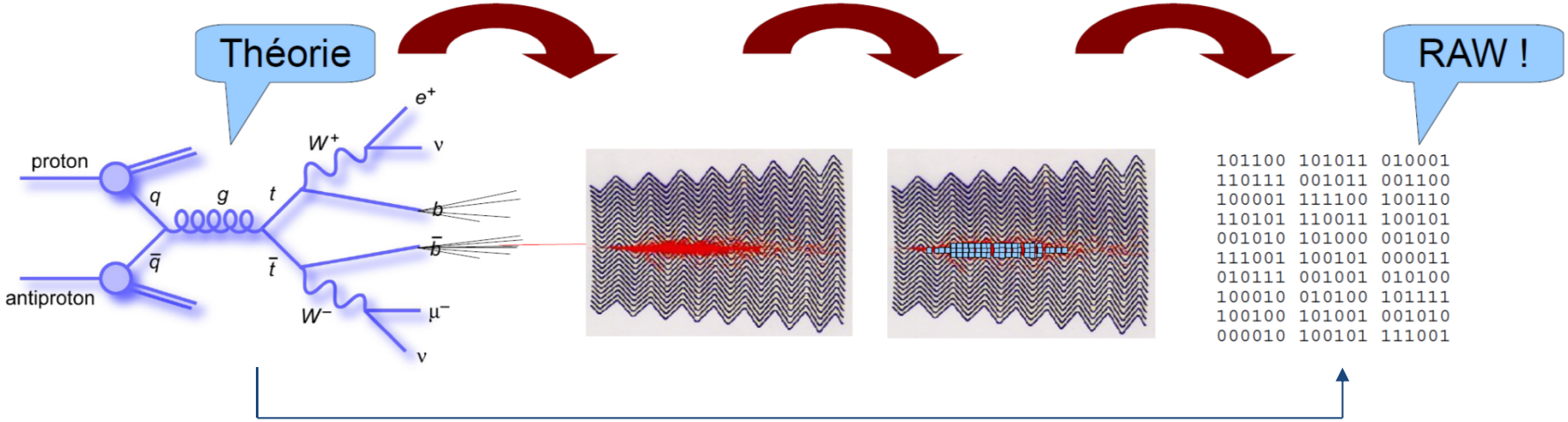
CMS


0m
1m
2m
3m
4m
5m
6m
7m
Key:
Muon
Electron
Charged Hadron (e.g. Pion)
Neutral Hadron (e.g. Neutron)
Photon
4T
2T
Silicon Tracker
Electromagnetic Calorimeter
Hadron Calorimeter
Superconducting Solenoid
Iron return yoke interspersed with Muon chambers
Transverse slice through CMS
D.Barney, CERN, February 2004

La simulation

La simulation

- La simulation est un outil central en physique des particules:
  - Développement de nouveaux détecteurs (ex: ILC)
  - Développement du software
  - Analyse de données
- 3 étapes:
  - Génération → liste de particules et leurs quadri-vecteurs (E,px,py,pz)
  - Simulation du détecteur → énergie et temps dans les zones actives (“hit”)
  - Numérisation → Format de données identiaues aux données réelles



+information vraie

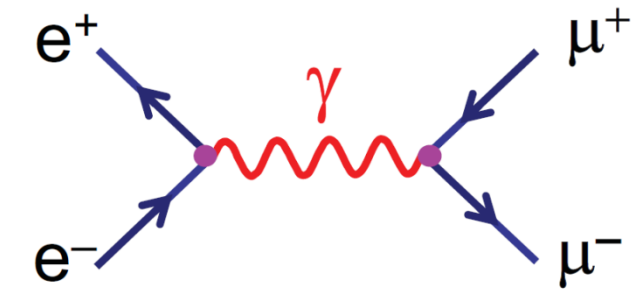
Génération: un exemple simple

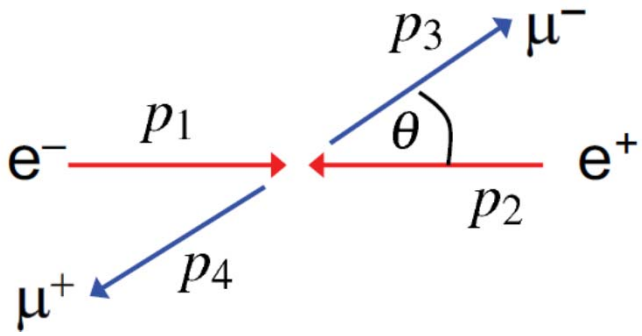
- Processus QED: $e^+e^- \rightarrow \mu^+\mu^-$
- L'état final peut être décrit par 2 variables: θ et φ
- La section efficace différentielle:

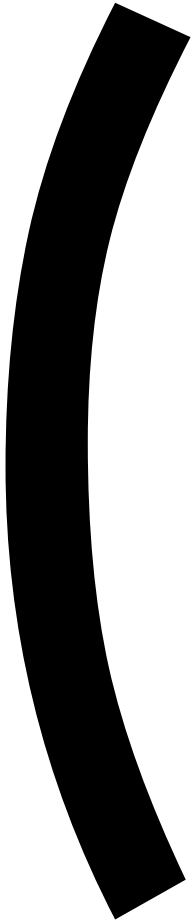
$$\frac{d\sigma}{d\cos\theta\, d\varphi} = \frac{\alpha_{em}^2}{4s}(1+\cos^2\theta)$$

- Section efficace totale:

$$\sigma = \frac{\alpha_{em}^2}{4s}\int_{\Omega}(1+\cos^2\theta)d\cos\theta\, d\varphi = \frac{4\pi\alpha_{em}^2}{3s}$$





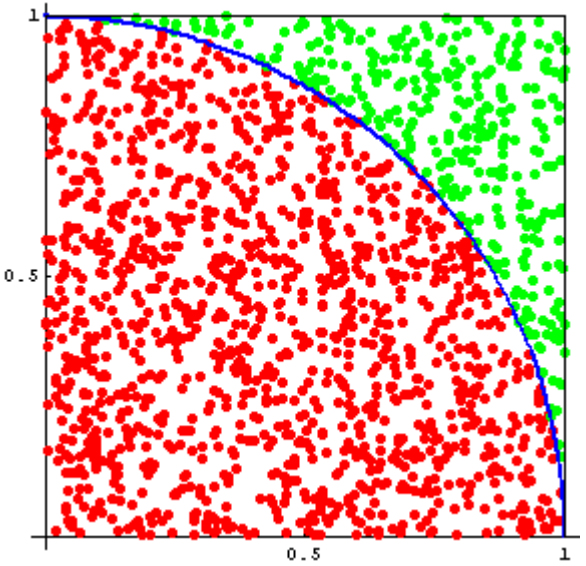


La méthode Monte Carlo

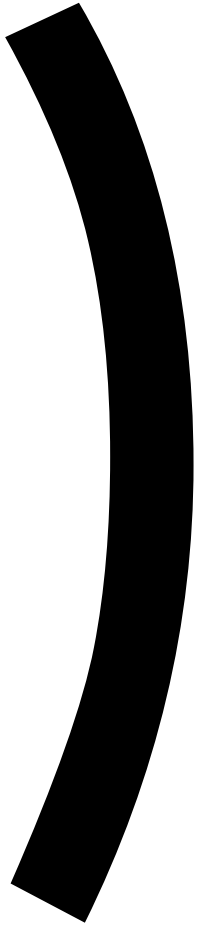
- Définition: méthode algorithmique visant à calculer une valeur numérique approchée en utilisant des procédés aléatoires

- Exemple: le calcul de π
  - Tire aléatoirement N couples de point (x,y) entre [0,1]
  - Accepte seulement les couples satisfaisant $x^2+y^2 \leq 1$
  - En faisant le rapport du nombre de points dans le disque au nombre de tirages, on obtient une approximation du nombre $\pi/4$
  - L'erreur sur l'estimateur décroit en $1/\sqrt{N}$

- Méthode facilement généralisable pour un plus grand nombre de dimensions et pour n'importe quelle hypervolume

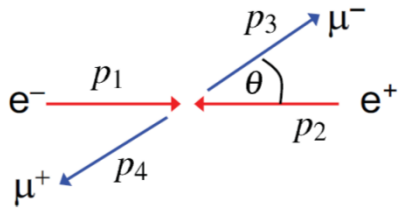


$$\hat{\pi} = 4\frac{N_{accepté}}{N_{tirage}}$$



Génération: un exemple simple

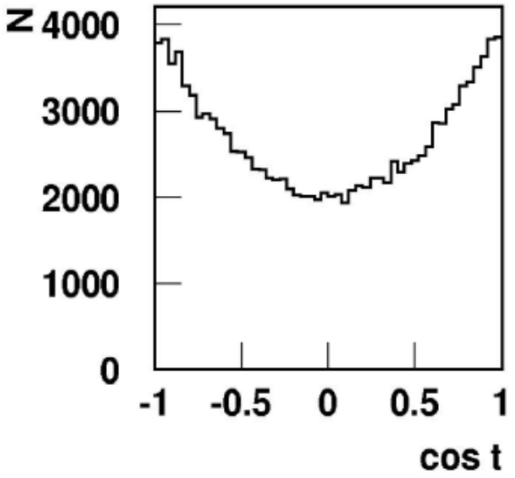
- Processus QED: $e^+e^- \rightarrow \mu^+\mu^-$



- L'état final peut être décrit par 2 variables: θ et φ

- La section efficace différentielle:

$$\frac{d\sigma}{d\cos\theta\, d\varphi} = \frac{\alpha_{em}^2}{4s}(1+\cos^2\theta)$$



- Section efficace totale:

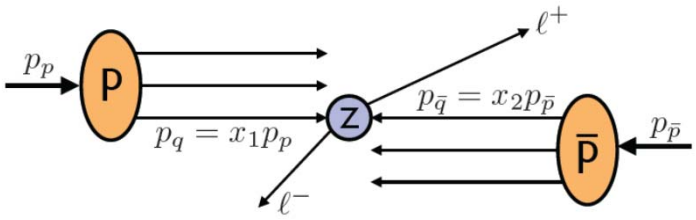
$$\sigma = \frac{\alpha_{em}^2}{4s}\int_{\Omega}(1+\cos^2\theta)d\theta\, d\varphi = \frac{4\pi\alpha_{em}^2}{3s}$$

- L'intégral peut également être calculée numériquement avec la méthode Monte-Carlo qui fournit en plus une liste d'"événements" (θ_i et φ_i) lors du tirage

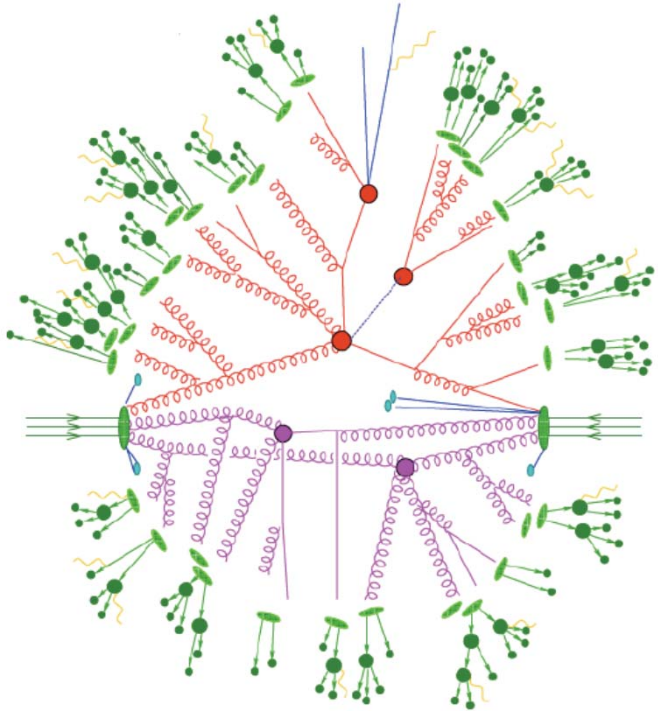
→ générateur de particules qui reproduit le caractère aléatoire de la mécanique quantique

Génération

- Auprès d'un collisionneur hadronique, la procédure est plus compliquée à cause de l'utilisation de QCD
  - PDF, hadronisation, MPI,…



- Il existe de très nombreux générateurs:
  - Pythia, Herwig, Sherpa, MadGraph, Powheg, MC@NLO, ALPGEN, Tauala, Photos, EvtGen,...
  - Généralistes ou spécialisés
  - Avec des niveaux de précisions différents
  - Processus du modèle standard ou au-delà
  - Aucun n'est idéal pour toutes les utilisations faites aux LHC



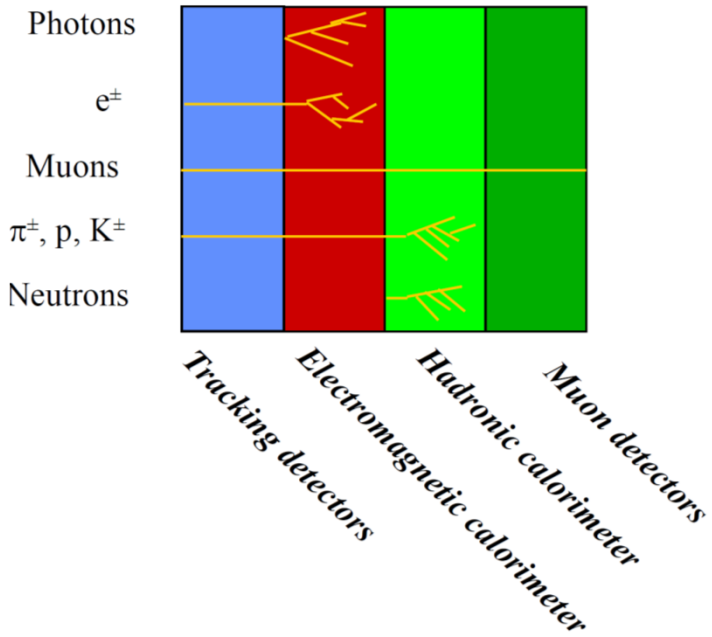
Simulation du détecteur

- Les particules des processus produits par les générateurs sont propagées dans une représentation virtuelle du détecteur

- C'est la phase dite de "transportation" :
  - Geant4

- Pour cela, il faut pouvoir modéliser
  - Nos détecteurs :
    - Géométrie, matériaux, volumes sensibles
  - Les particules, avec leur propriétés
    - Leptons, hadrons, etc.
  - La physique des interactions particules-matière
    - Modéliser les « processus physiques »

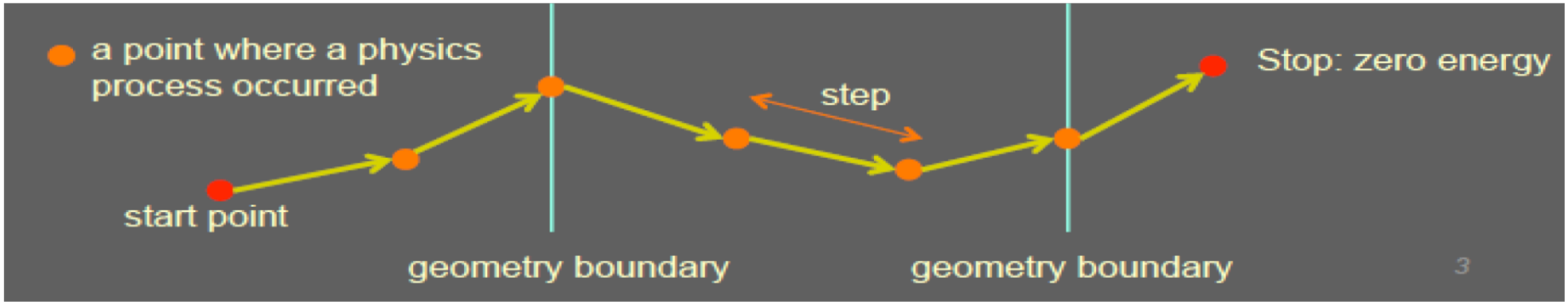
- Basée également sur la méthode Monte-Carlo



Simulation du détecteur: Geant4

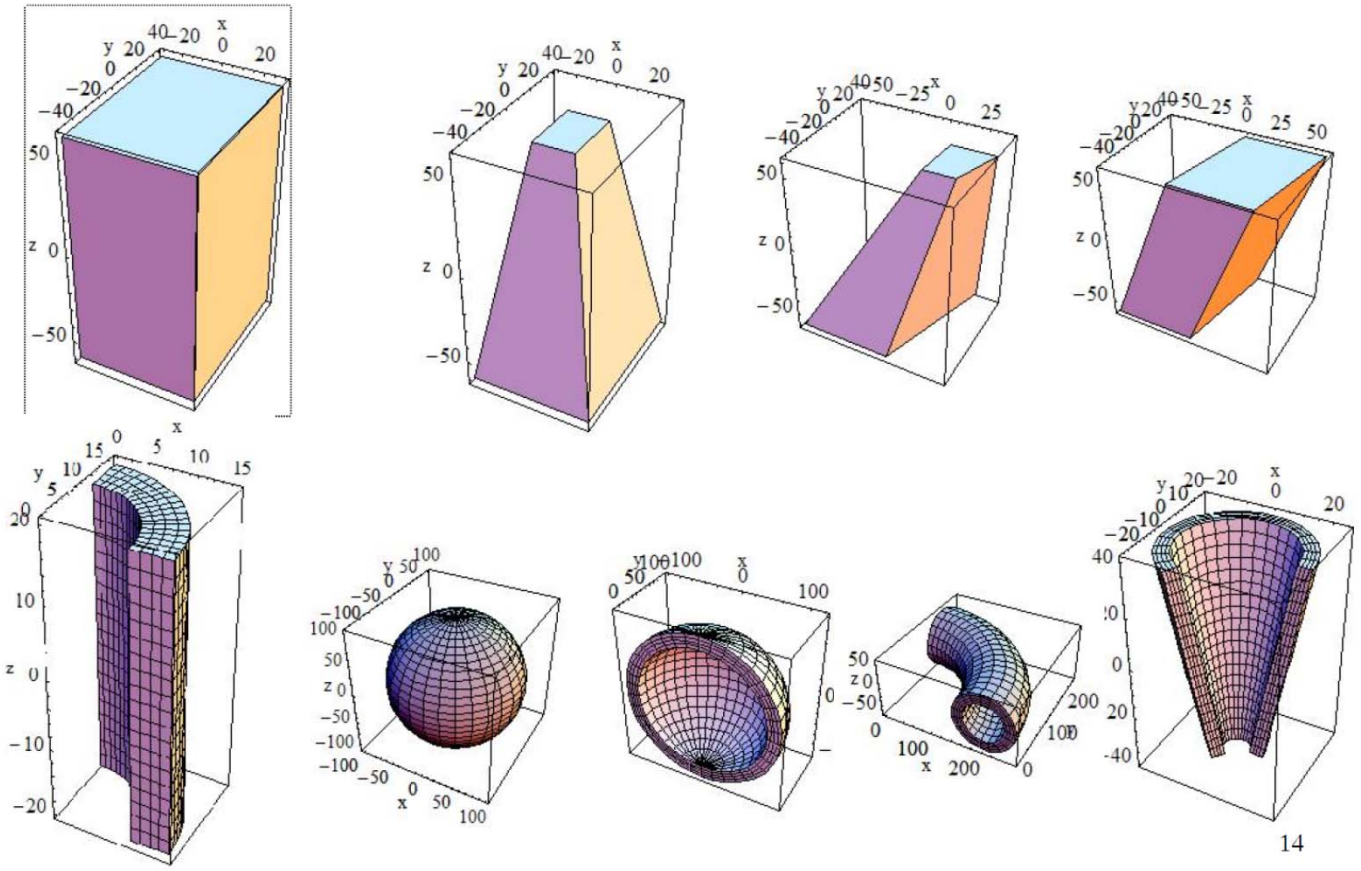
- Geant4 ("GEometry ANd Tracking")
- Initié par le CERN en 1994
- Développé par une collaboration internationale (~100 membres)
- Logiciel de simulation des interactions particules-matières, à vocation "généraliste"
- Technologie Orienté-Objet (C++)
- Ouvert et gratuit

- Transporter une particule pas-à-pas en prenant en compte les interactions avec des matériaux et des champs électromagnétiques externes jusqu'à ce que la particule
  - Perde totalement son énergie cinétique
  - Disparaisse par interaction
  - Sorte d'un volume de simulation



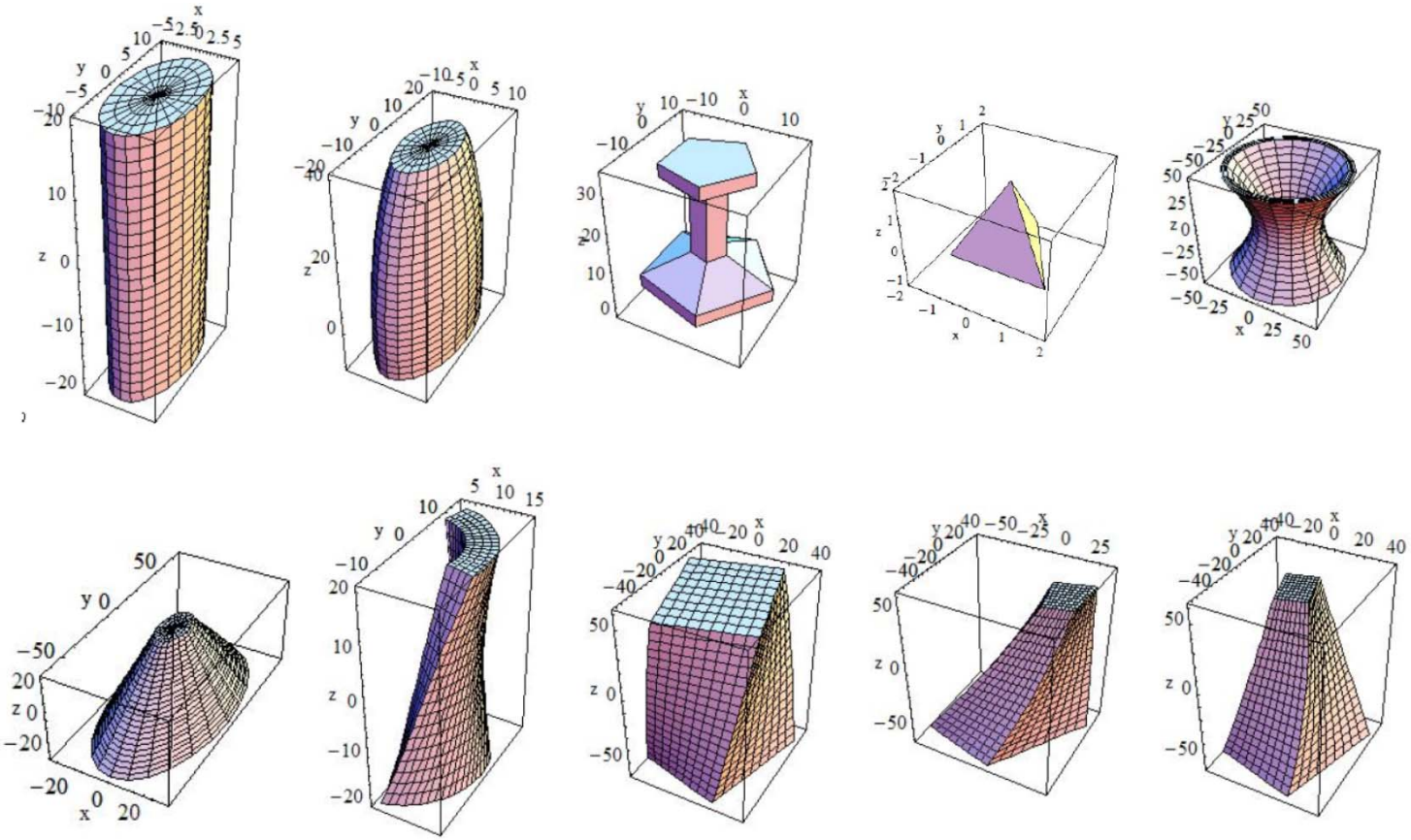
Geant4: géometrie

- De nombreuses possibilités pour décrire la géométrie du détecteur
  - Combiner des éléments géométriques de base (boîte, cylindre, trapèze, etc.)
  - Représentation par des opérations booléennes, etc.
- De nombreux matériaux prédéfinis

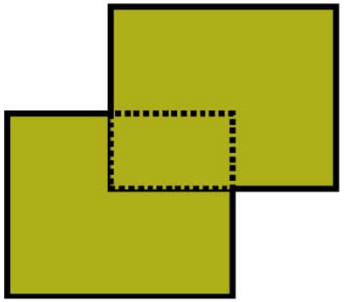


Geant4: géometrie

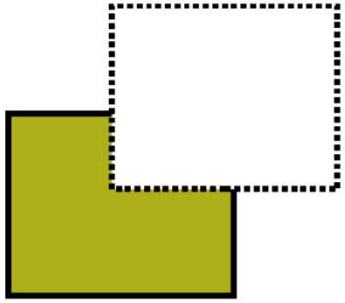
- De nombreuses possibilités pour décrire la géométrie du détecteur
  - Combiner des éléments géométriques de base (boîte, cylindre, trapèze, etc.)
  - Représentation par des opérations booléennes, etc.
- De nombreux matériaux prédéfinis



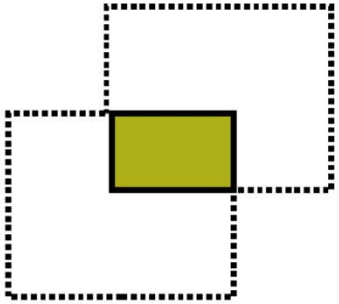
Geant4: géometrie



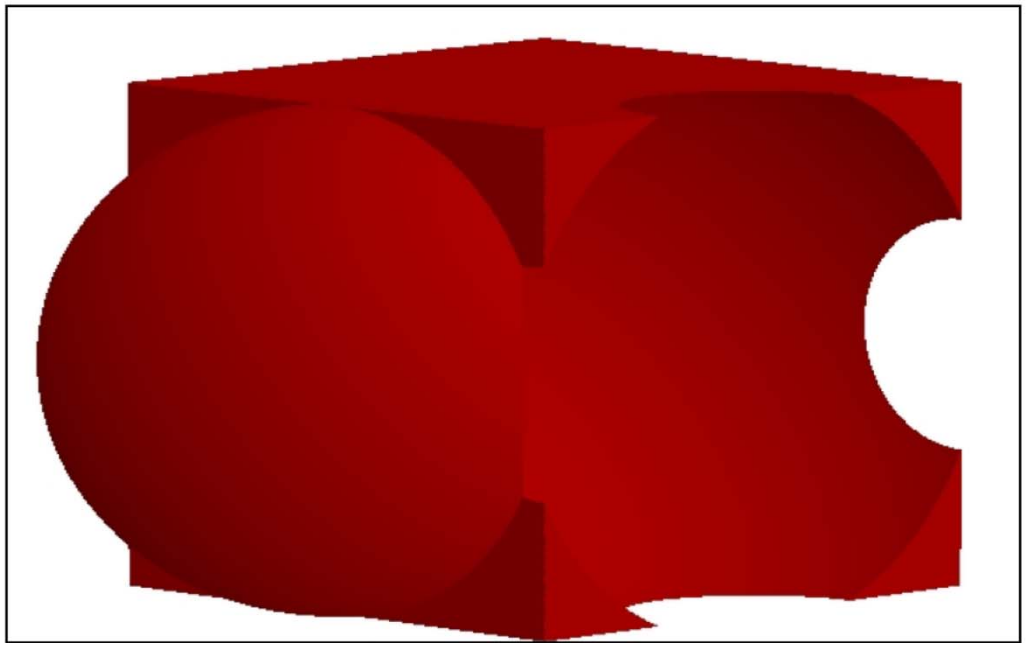
Union



Subtraction



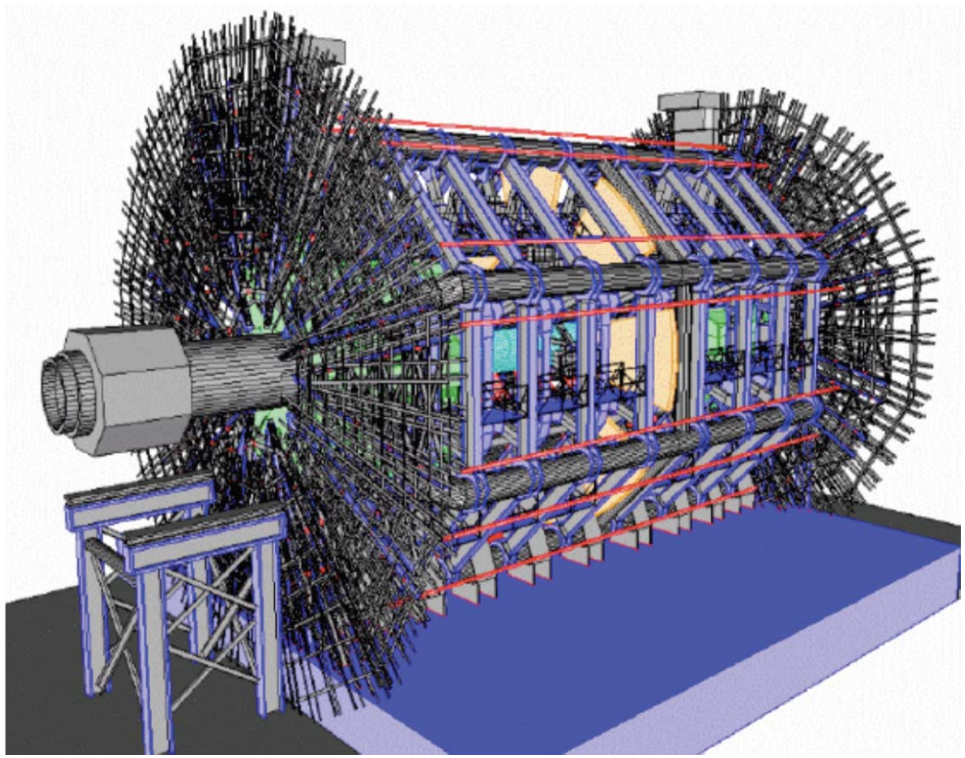
Intersection

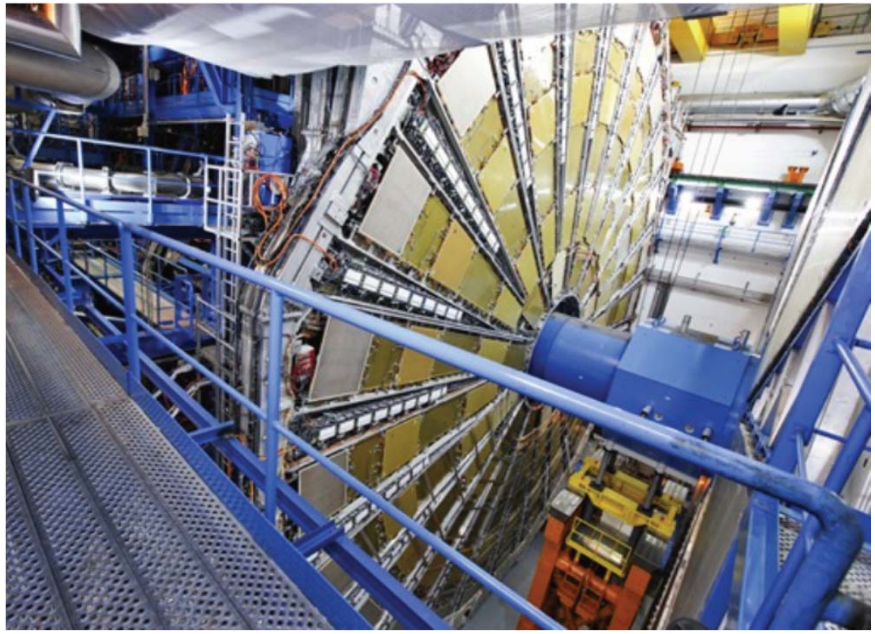


+symétrie

Geant4: géometrie

- Exemple : ATLAS ~ 5 millions de volumes

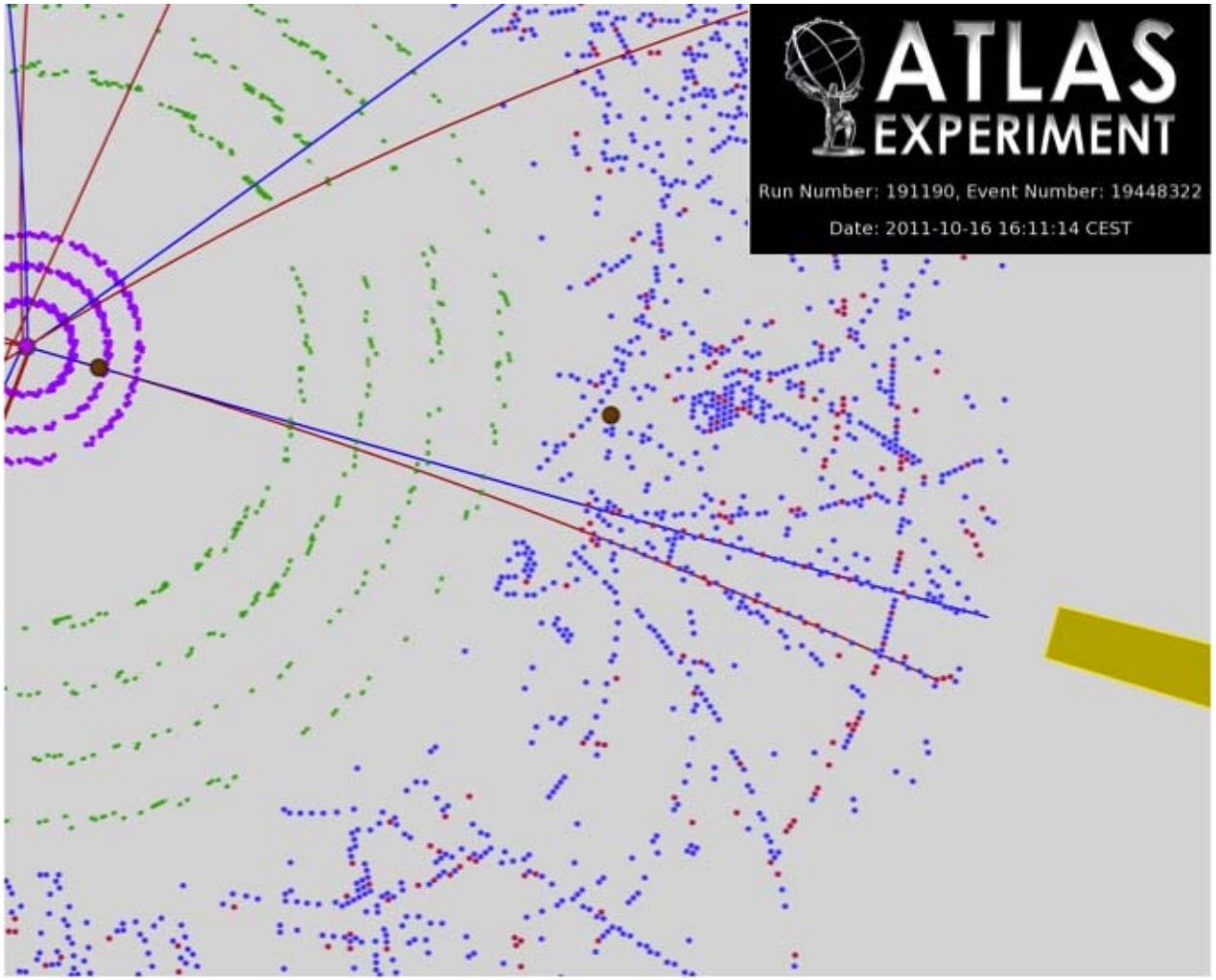


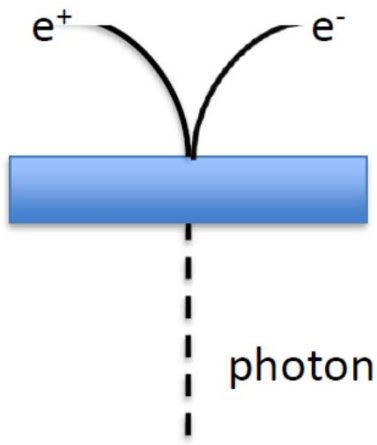


- Le challenge est de faire cela correctement!

Voir la géometrie du trajectographe

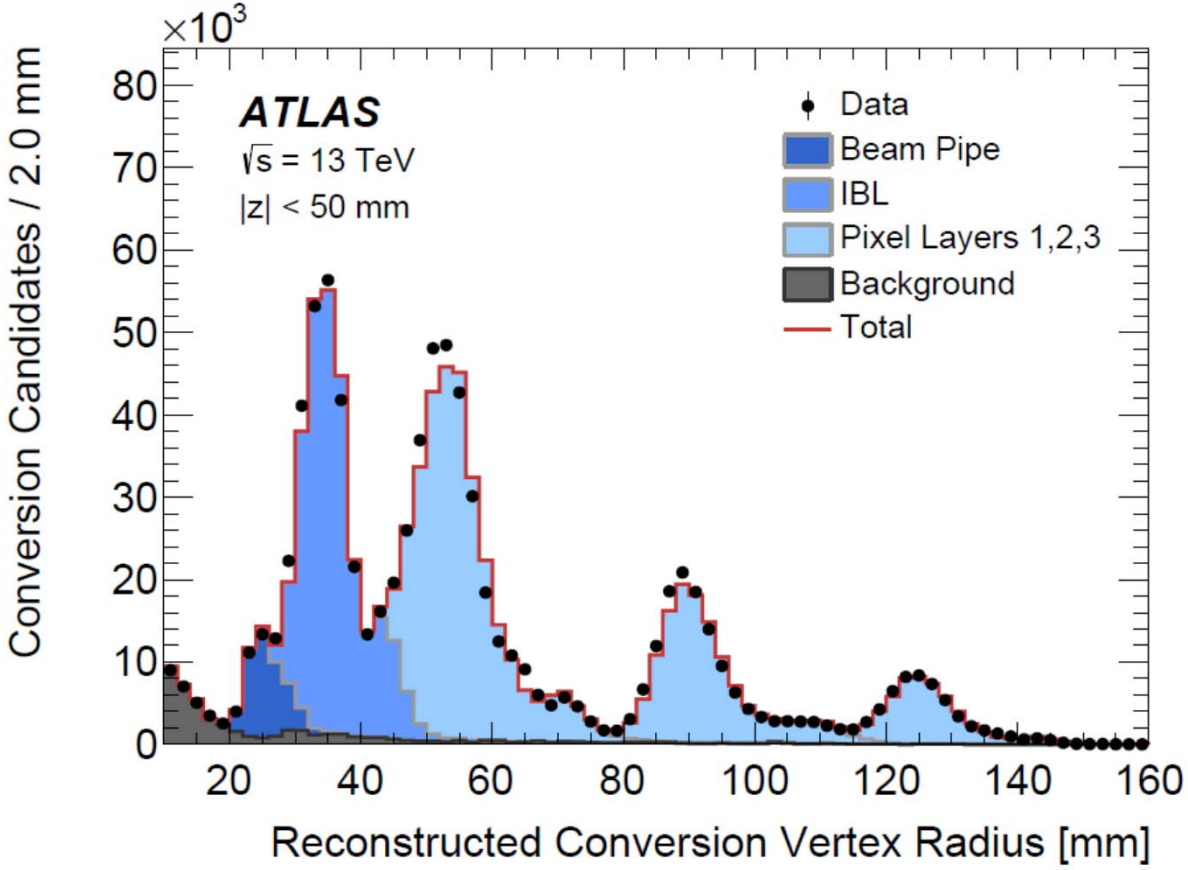
- Avec des photons converties

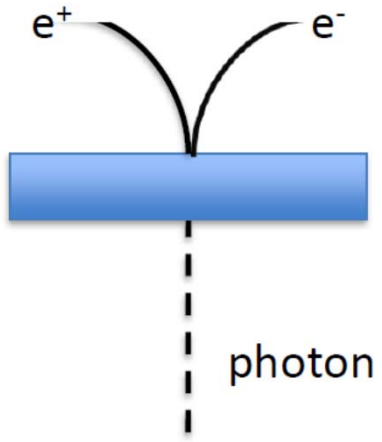




Voir la géometrie du trajectographe

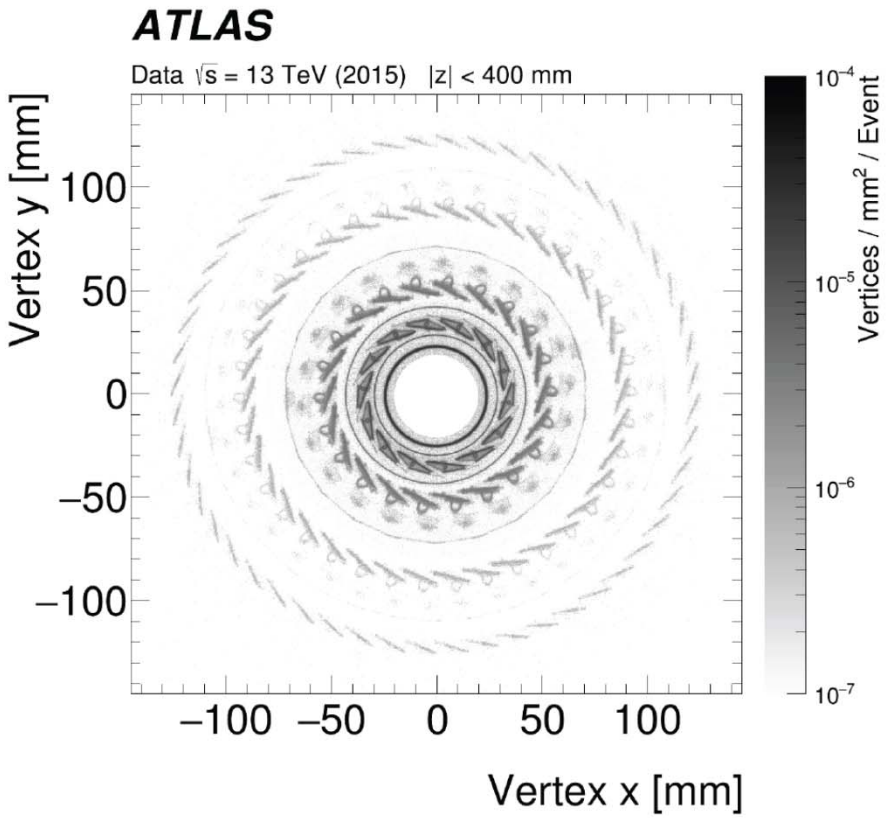
- Avec des photons converties

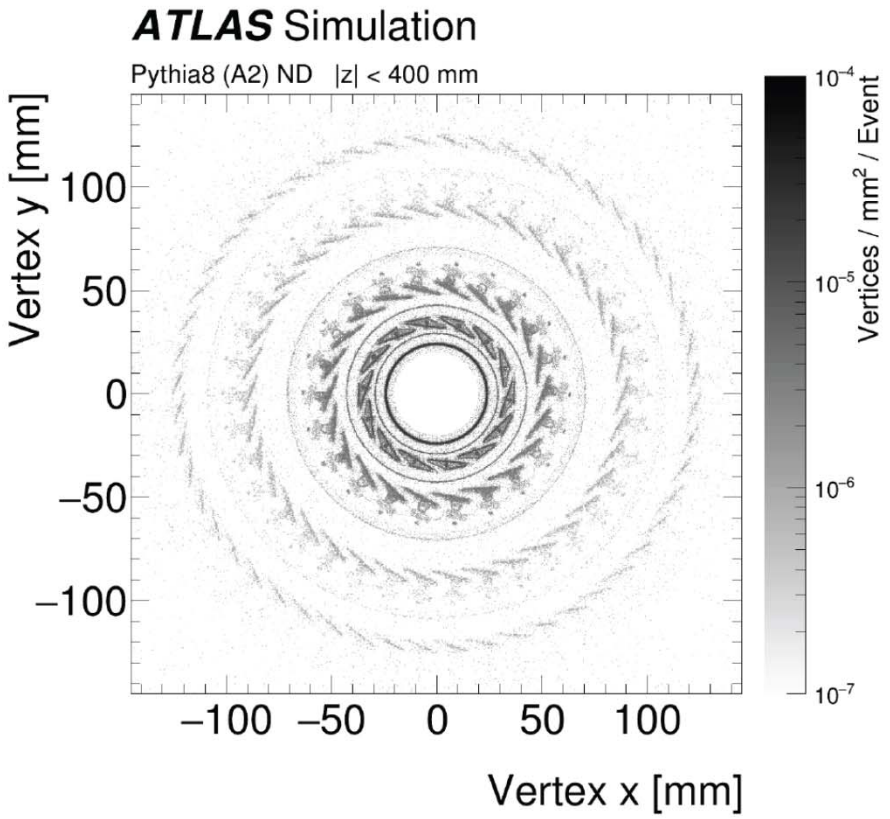




Voir la géometrie du trajectographe

- Avec des désintégrations hadroniques

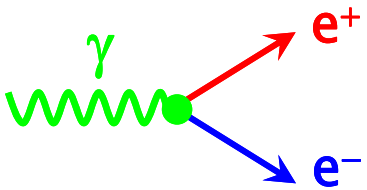




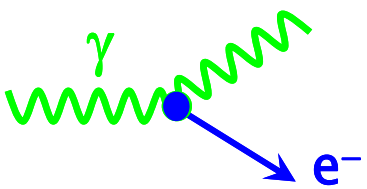
Exemple de processus

Cours de P. Puzo

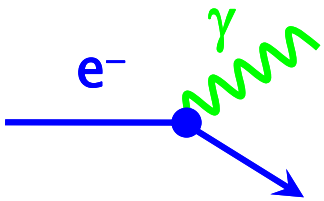
Conversion



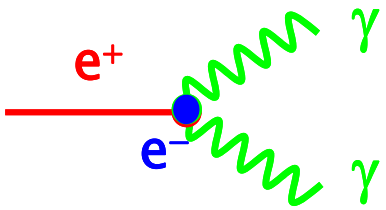
Compton



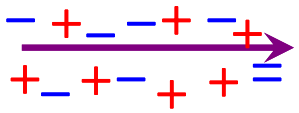
Bremsstrahlung



Annihilation



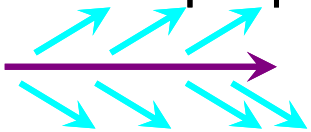
Ionisation



Particule chargée

Cherenkov

Photons optiques



Particule chargée

(dont la vitesse est > à la vitesse de la lumière dans le milieu)

+ interactions hadroniques

Numérisation

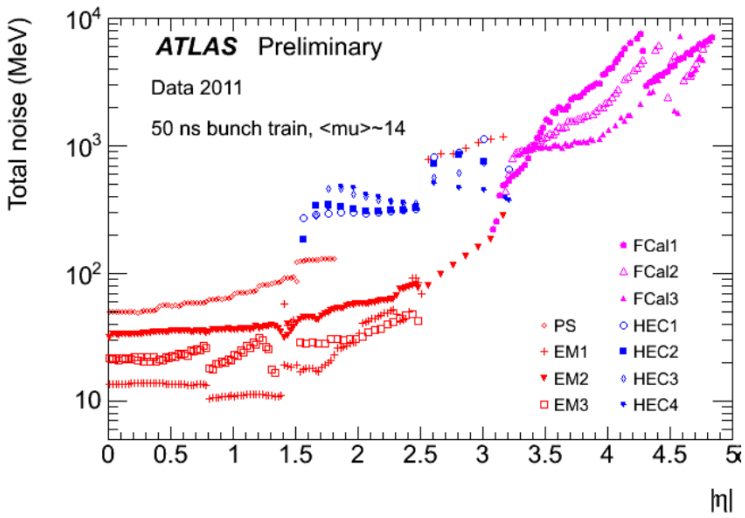
- L'énergie déposée dans les volumes de détection est utilisée pour fabriquer des données semblables aux données réelles

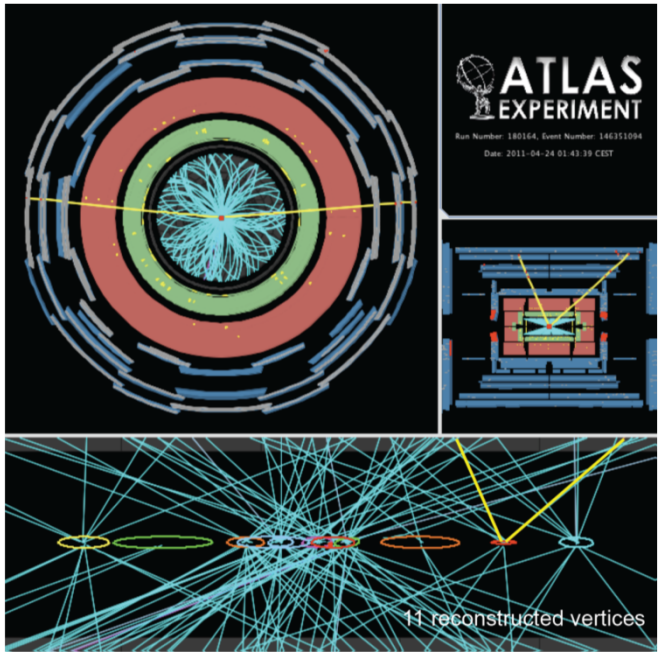
- A partir de ces dépôts, on va par exemple calculer la collection de lumière dans un cristal, la charge collectée dans un gaz, etc..

- Et calculer la réponse de la voie d'électronique correspondante

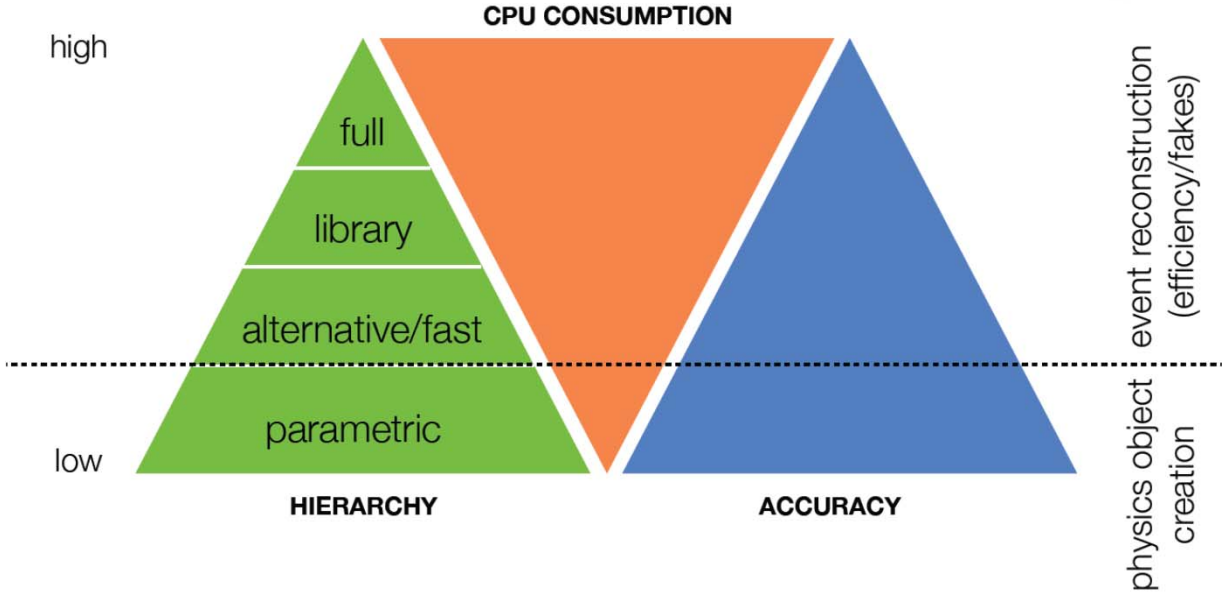
- Rajouter le bruit de l'électronique

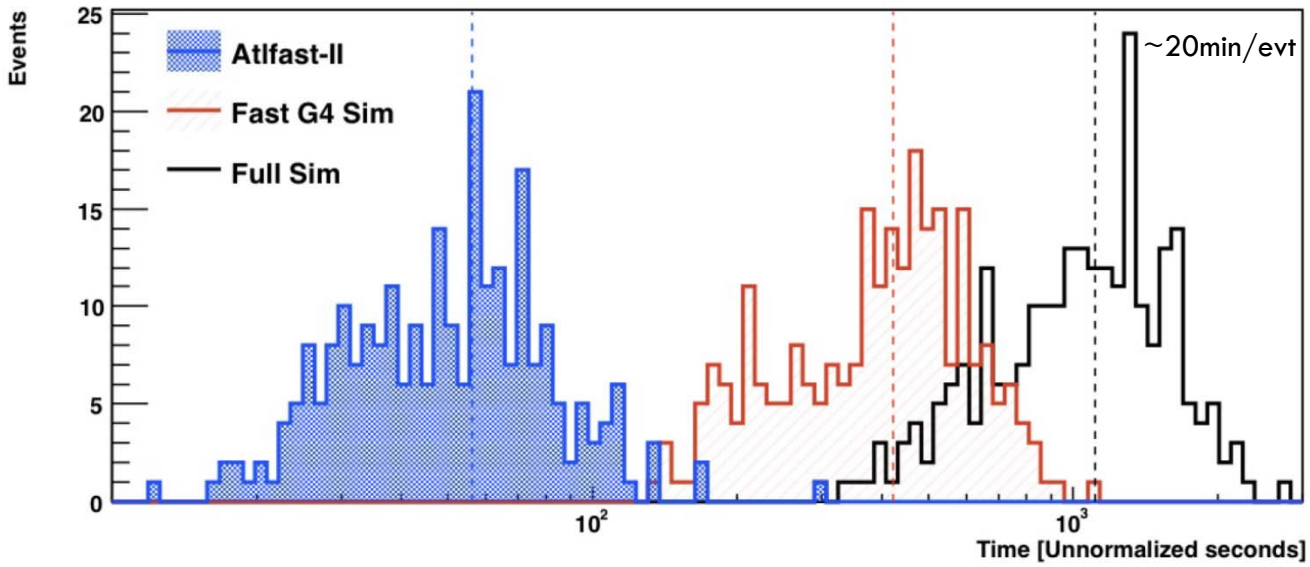
- Rajouter les événements d'empilement





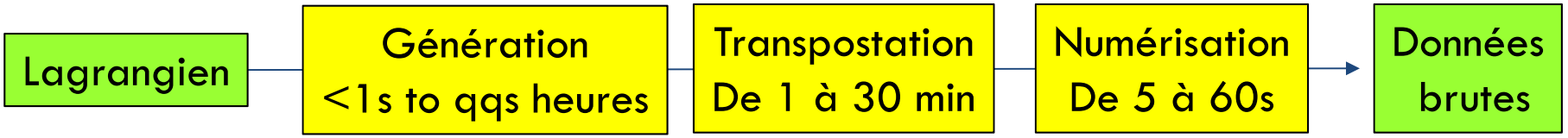
Simulation rapide





Simulation: conclusion

- Trois étapes pour la simulation:



- Les challenges principaux de la simulation sont:
  - Précision
  - Performance CPU
  - Validation

- Une alternative à la simulation complète (~20min/evt) est la simulation rapide (paramétrisation des gerbes) qui est moins précise

- Les imperfections de la modélisation doivent être étudiées afin d'estimer les incertitudes induites sur les analyses de physique

La reconstruction: l'exemple des electrons

et des positrons

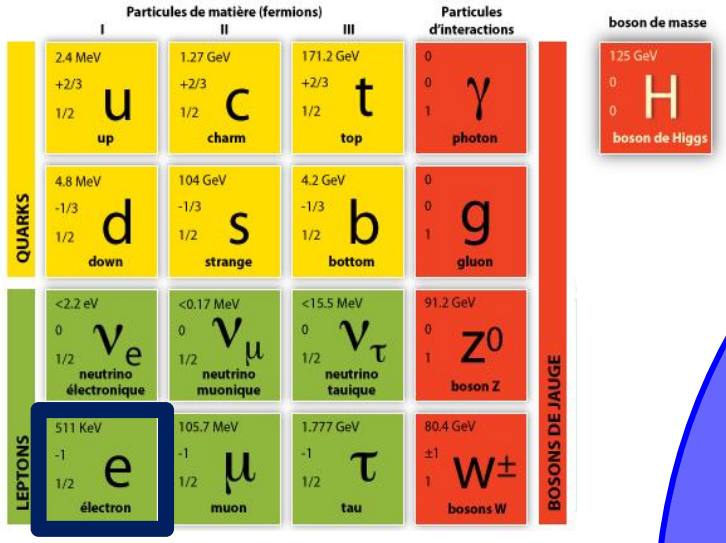
Facteurs de mérite de la reconstruction

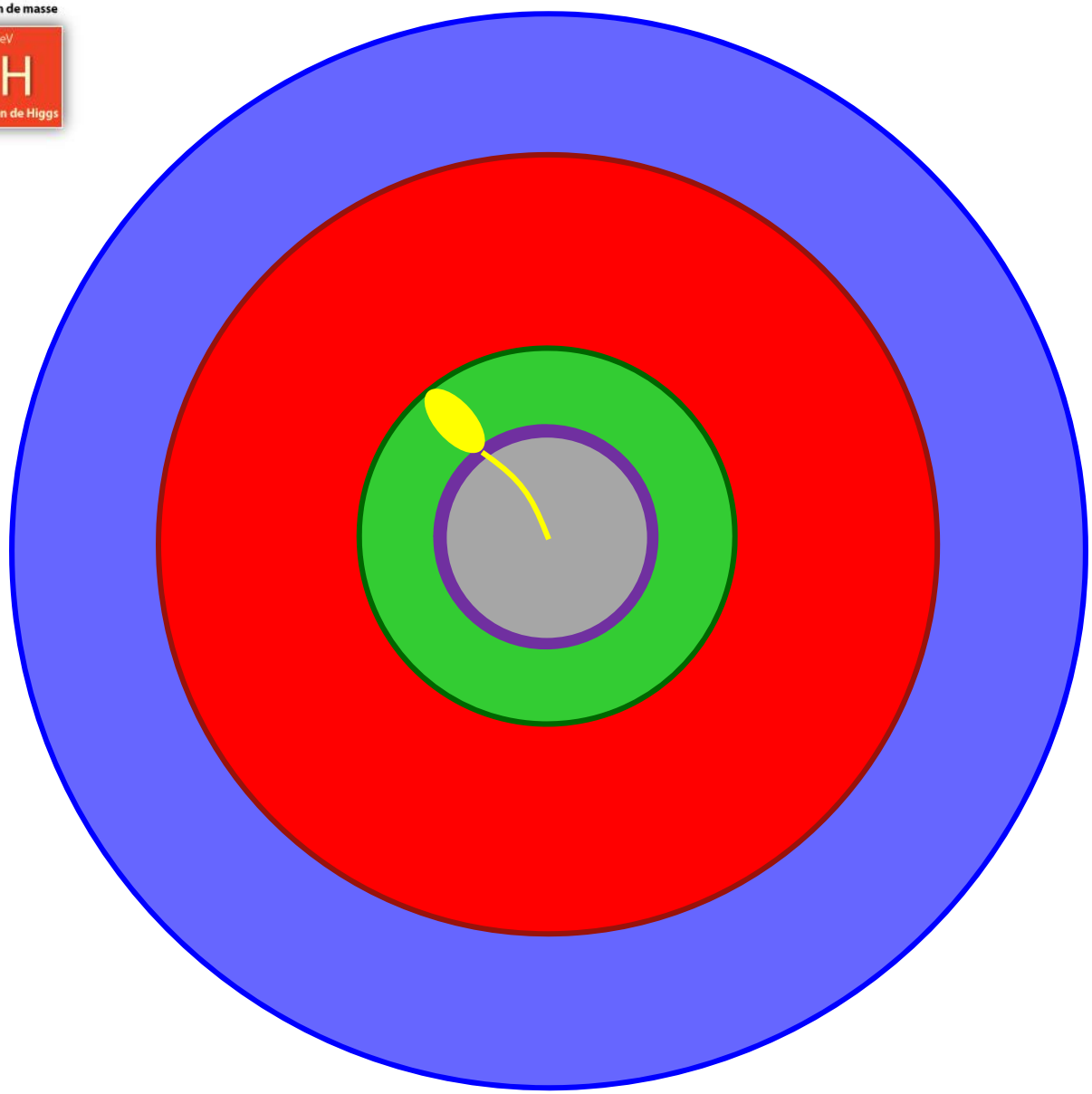
- Facteurs de mérite
  - Efficacité de reconstruction et sélection
  - Taux de mauvaise identification
  - Linéarité
  - Résolution
  - Stabilité en fonction de l'empilement
  - Ressource informatique nécessaire (temps de calcul, mémoire)

- Lors du développement d'un nouveau détecteur, il faut s'assurer que les **performances** soient les meilleurs possibles en s'appuyant sur **la simulation**

- Lorsque le détecteur est construit il faut **déterminer ces grandeurs dans les données!**

Electron

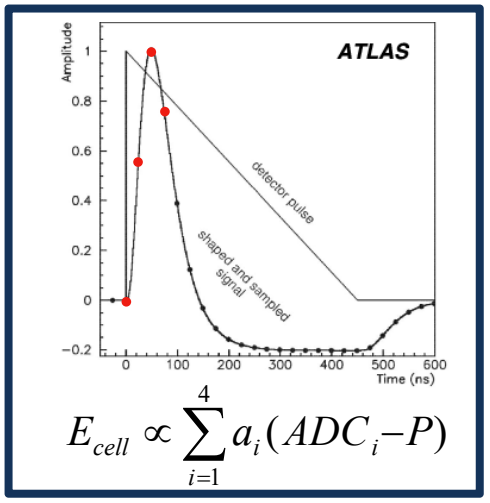




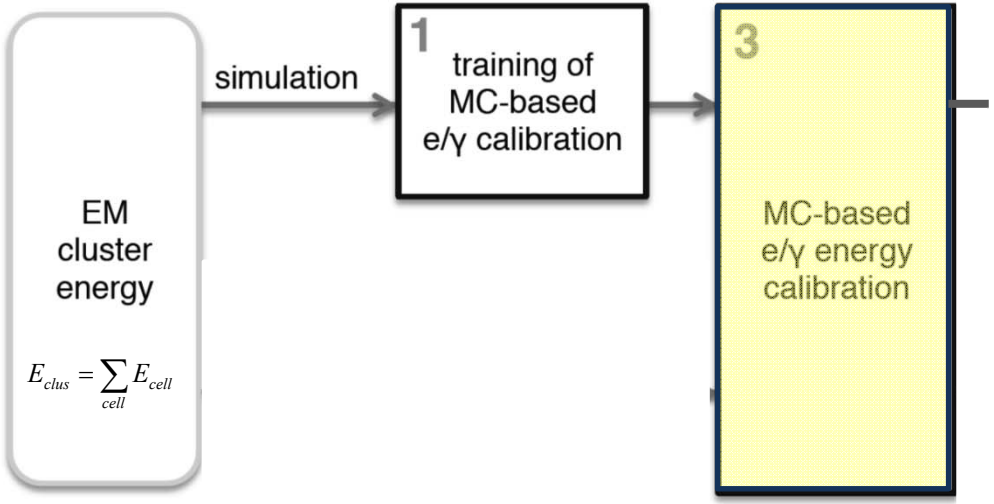
Trajectographe
Solénoïde
Calorimètre EM
Calorimètre Hadronique
Spectromètre à muon

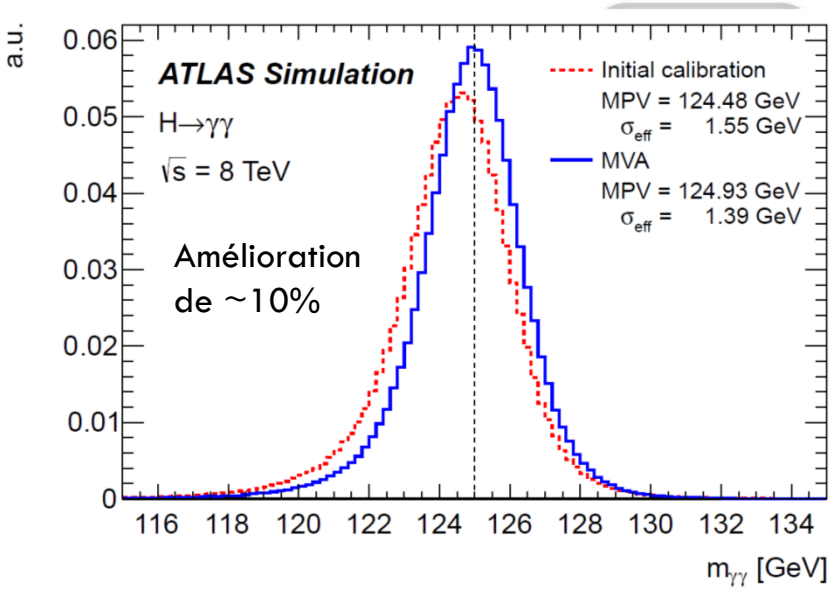
La procédure d'étalonnage

- La procédure corrige:
  - Pertes dans les matériaux inactifs
  - Fuites latérales d'énergie
  - Inhomogénéité en φ et η
- BDT utilisant avec la simulation:
  - Minimisation de la résolution en énergie
  - Nécessite une très bonne connaissance de la simulation du détecteur!
- Taille de la correction entre 5 et 15%



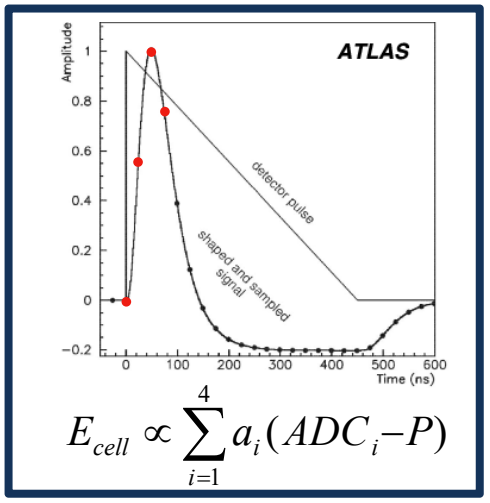
$$E_{cell} \propto \sum_{i=1}^{4} a_i (ADC_i - P)$$

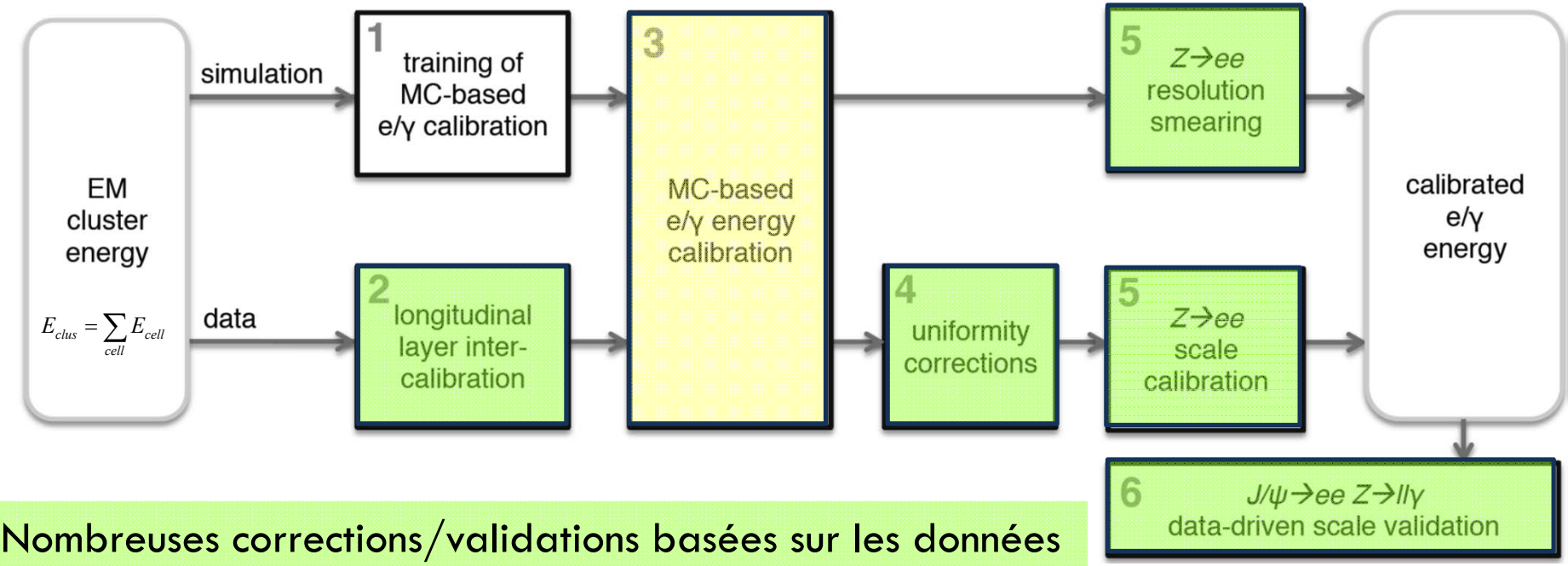




La procédure d'étalonnage

- La procédure corrige:
  - Pertes dans les matériaux inactifs
  - Fuites latérales d'énergie
  - Inhomogénéité en φ et η
- BDT utilisant avec la simulation:
  - Minimisation de la résolution en énergie
  - Nécessite une très bonne connaissance de la simulation du détecteur!
- Taille de la correction entre 5 et 15%



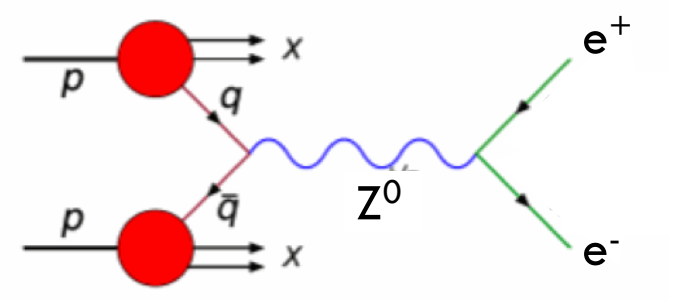


Nombreuses corrections/validations basées sur les données

Etalonnage avec la masse du Z

- La masse du Z est connue avec une très grande précision depuis LEP (91.1876(21) GeV)

$$M_{12} = \sqrt{2E_1E_2(1 - cos\theta_{12})}$$

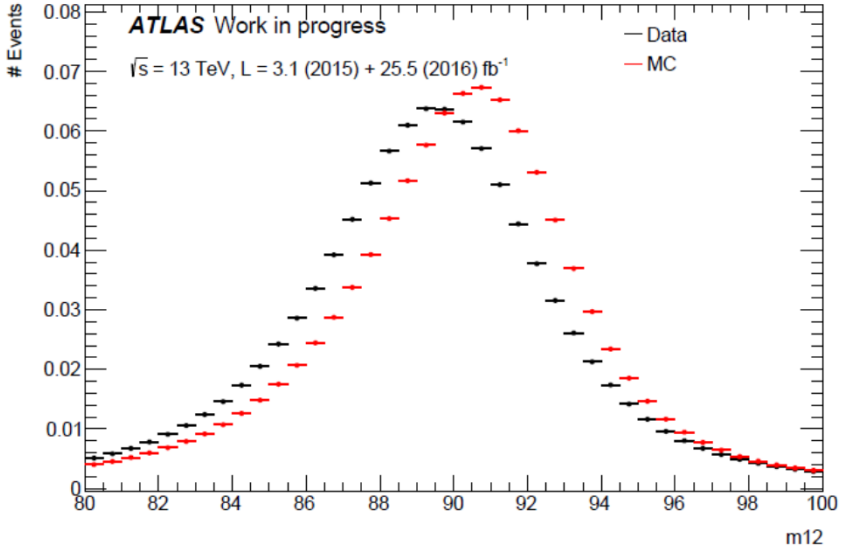


- Position du maximum

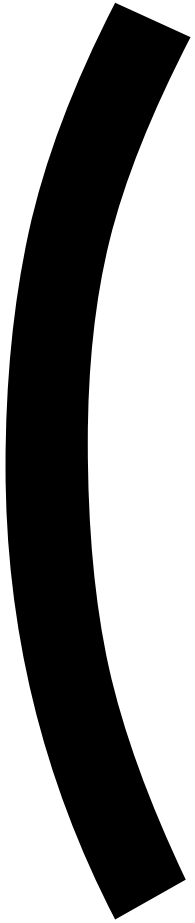
→échelle d'énergie des électrons

- Largeur de la distribution

→ résolution en énergie

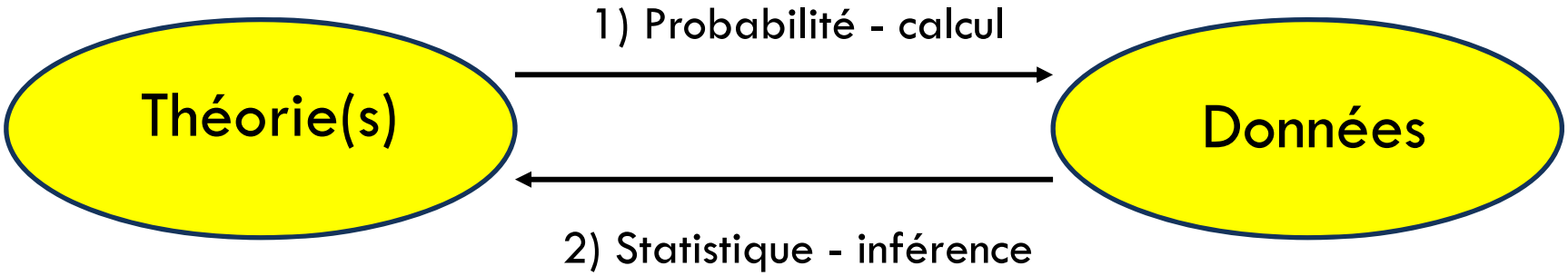


- Utilisation de la méthode du maximum de vraisemblance pour extraire les facteurs corrections

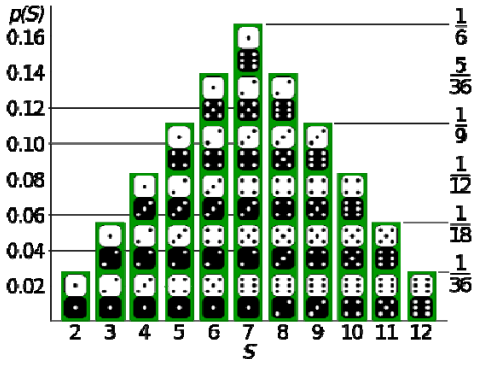


La problématique des statistiques

- En physique des particules, on s'intéresse à des processus ayant un caractère aléatoire (mécanique quantique)

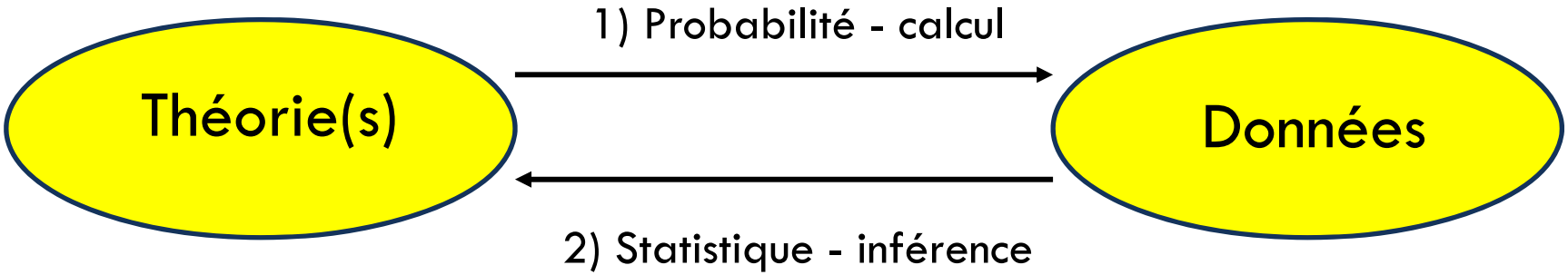


- Exemple: jets de dés
  1) Ayant 2 dés à 6 faces non pipés, quel est la probabilité de chaque occurrence?
  2) Etant donné le résultat de 20 jets de 2 dés à 6 faces, est ce que les dés sont pipés?



La problématique des statistiques

- En physique des particules, on s'intéresse à des processus ayant un caractère aléatoire (mécanique quantique)

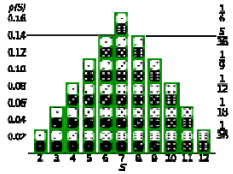


- Dans une analyse de physique, on veut:
  - Remonter à la distribution sous-jacente à partir des données. En général on a un modèle pour ces données, mais les paramètres contrôlant la densité de probabilité sont inconnus (estimation de paramètres).
  - Tester si une théorie est consistante avec les données (test d'hypothèses)

Estimateur au maximum de vraisemblance

- La **fonction de vraisemblance** est une fonction de probabilités conditionnelles qui décrit les valeurs x_i d'une loi statistique en fonction des paramètres θ_i supposés connus.

$$L(x_1,\ldots,x_n \mid \theta_1,\ldots,\theta_k) = \prod_{i=1}^{N} f(x_i \mid \theta_1,\ldots,\theta_k)$$



Mesures «fixes»

Paramètres que l'on veut déterminer

Densité de probabilité (associe une probabilité à chaque issue possible d'une expérience aléatoire)

Hypothèse: les x_i sont indépendants et identiquement distribués entre eux

Estimateur au maximum de vraisemblance

- La **fonction de vraisemblance** est une fonction de probabilités conditionnelles qui décrit les valeurs x_i d'une loi statistique en fonction des paramètres θ_i supposés connus.

$$L(x_1,\ldots,x_n \mid \theta_1,\ldots,\theta_k) = \prod_{i=1}^{N} f(x_i \mid \theta_1,\ldots,\theta_k)$$

- Pour un lot de données fixé (x_1, …, x_n), les valeurs (θ_1, …, θ_k) qui maximisent la fonction de vraisemblance maximal sont des estimateurs des paramètres (θ_1, …, θ_k)

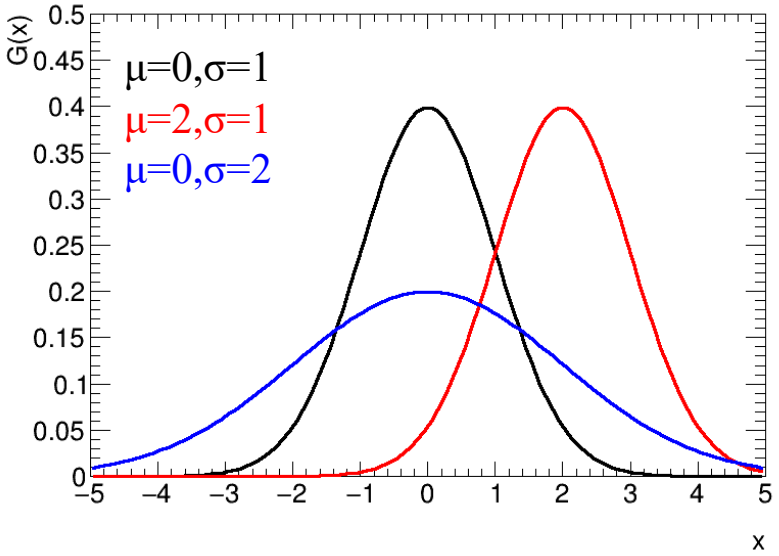
- Les estimateurs ainsi obtenus ont de nombreuses bonnes propriétés dans la limite des grands nombres:
  - Non biaisé
  - Faible variance

- En pratique, on minimise -ln L(x_1, …, x_n | θ_1, …, θ_k)

La distribution gaussienne

$$G(x;\mu,\sigma) = \underbrace{\frac{1}{\sqrt{2\pi\sigma^2}}}\, e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

Facteur de normalisation pour que l'aire soit égale à 1



- E[x]=μ
- V[x]=σ²
- Elle est également appelé la distribution normale
- Loi gaussienne centrée réduite ou loi gaussienne standard: μ=0 et σ=1

Le théorème central limite

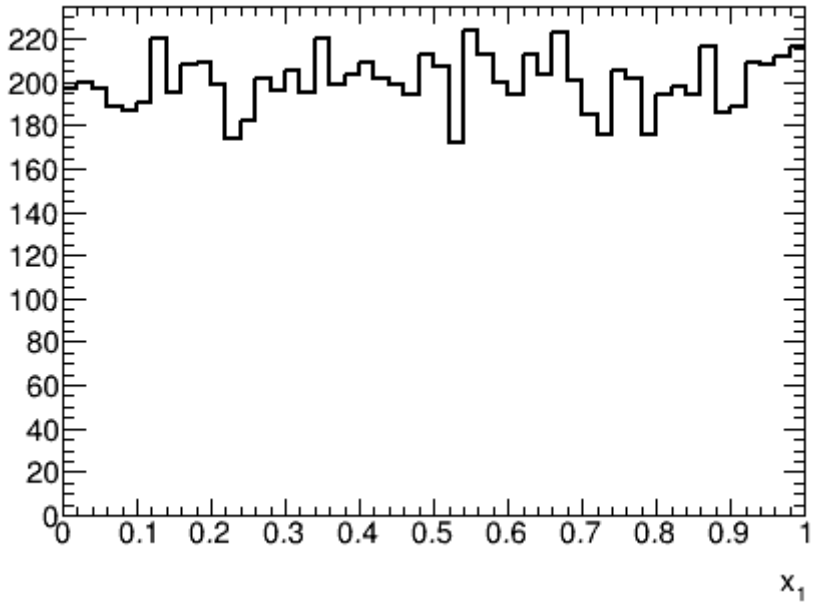
- Théorème
  - Si une variable aléatoire X est formée de la somme de N variables aléatoires x_i indépendantes, de moyenne μ_i et de variance σ_i, alors X est distribuée suivant une gaussienne de valeur moyenne $\mu = \sum_{i=1,N}\mu_i$ et de variance $\sigma^2 = \sum_{i=1,N}\sigma_i^2$, dans la limite $N \rightarrow \infty$

Le théorème central limite

- Théorème
  - Si une variable aléatoire X est formée de la somme de N variables aléatoires x_i indépendantes, de moyenne μ_i et de variance σ_i, alors X est distribuée suivant une gaussienne de valeur moyenne $\mu = \sum_{i=1,N}\mu_i$ et de variance $\sigma^2 = \sum_{i=1,N}\sigma_i^2$, dans la limite $N \to \infty$

- Exemple avec des variables x_i tirées aléatoirement de façon uniforme entre 0 et 1

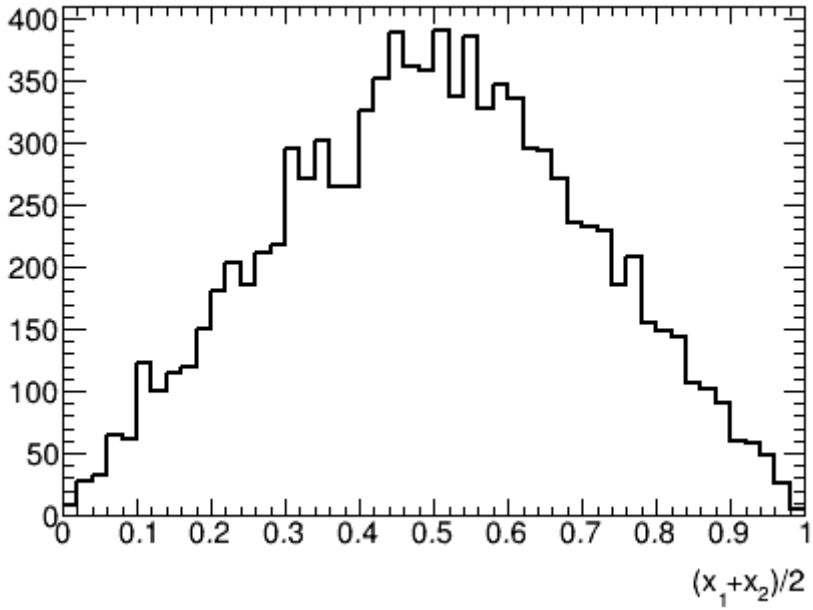
100000 tirages



Le théorème central limite

- Théorème
  - Si une variable aléatoire X est formée de la somme de N variables aléatoires x_i indépendantes, de moyenne μ_i et de variance σ_i, alors X est distribuée suivant une gaussienne de valeur moyenne $\mu = \sum_{i=1,N}\mu_i$ et de variance $\sigma^2 = \sum_{i=1,N}\sigma_i^2$, dans la limite $N \to \infty$

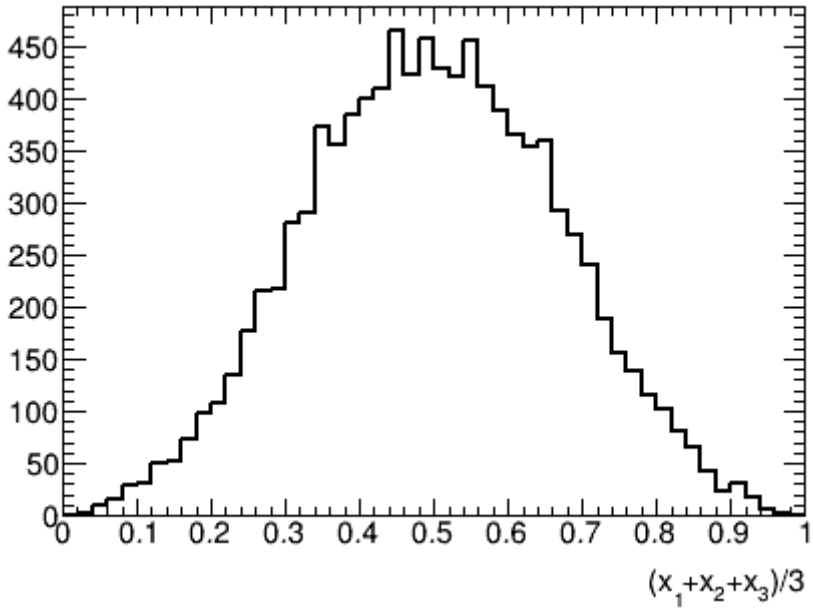
- Exemple avec des variables x_i tirées aléatoirement de façon uniforme entre 0 et 1



Le théorème central limite

- Théorème
  - Si une variable aléatoire X est formée de la somme de N variables aléatoires x_i indépendantes, de moyenne μ_i et de variance σ_i, alors X est distribuée suivant une gaussienne de valeur moyenne $\mu = \sum_{i=1,N}\mu_i$ et de variance $\sigma^2 = \sum_{i=1,N}\sigma_i^2$, dans la limite $N \to \infty$

- Exemple avec des variables x_i tirées aléatoirement de façon uniforme entre 0 et 1

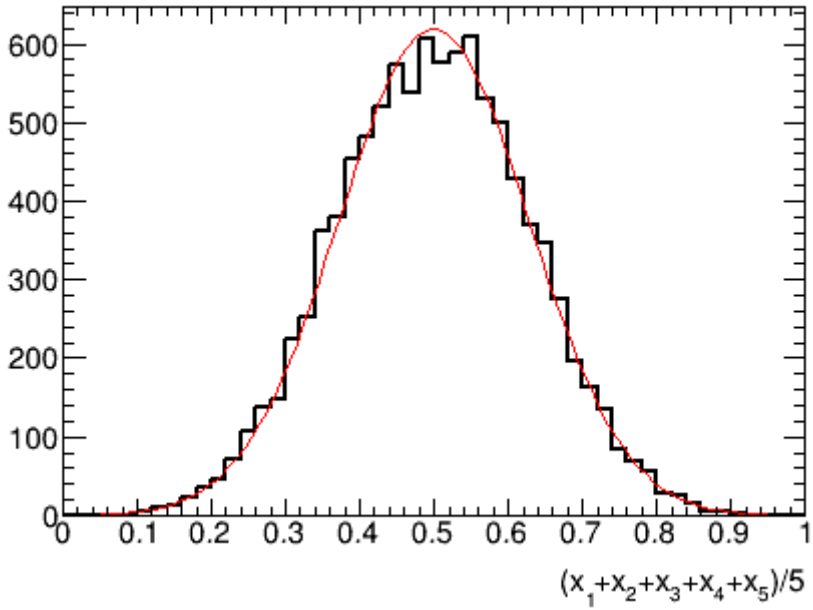


Le théorème central limite

- Théorème
  - Si une variable aléatoire X est formée de la somme de N variables aléatoires x_i indépendantes, de moyenne μ_i et de variance σ_i, alors X est distribuée suivant une gaussienne de valeur moyenne $\mu = \sum_{i=1,N}\mu_i$ et de variance $\sigma^2 = \sum_{i=1,N}\sigma_i^2$, dans la limite $N \to \infty$

- Exemple avec des variables x_i tirées aléatoirement de façon uniforme entre 0 et 1

Pour N=5, la distribution ressemble à une gaussienne

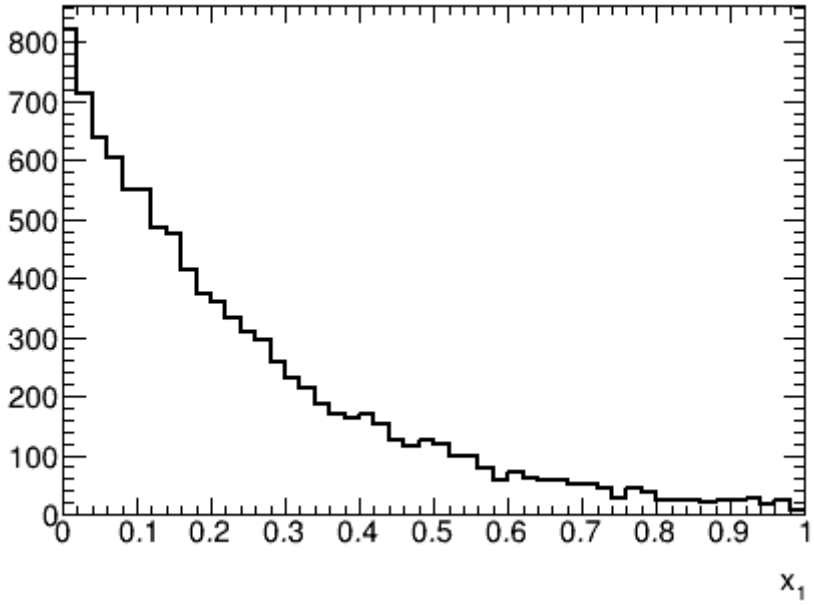


$\sigma \sim 1/\sqrt{(12x5)}$

Le théorème central limite

- Théorème
  - Si une variable aléatoire X est formée de la somme de N variables aléatoires x_i indépendantes, de moyenne μ_i et de variance σ_i, alors X est distribuée suivant une gaussienne de valeur moyenne $\mu = \sum_{i=1,N}\mu_i$ et de variance $\sigma^2 = \sum_{i=1,N}\sigma_i^2$, dans la limite $N \to \infty$

- Exemple avec des variables x_i tirées aléatoirement selon une loi exponentielle entre 0 et 1

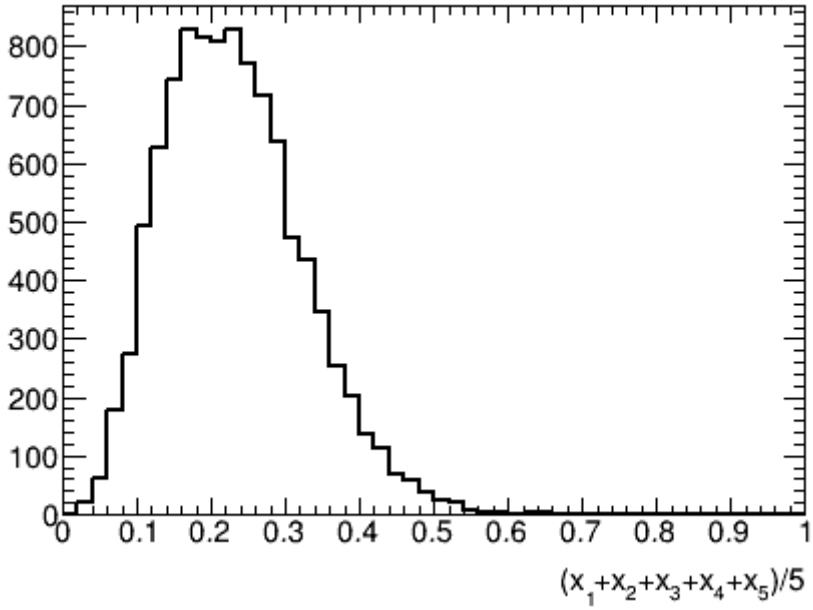


Le théorème central limite

- Théorème
  - Si une variable aléatoire X est formée de la somme de N variables aléatoires x_i indépendantes, de moyenne μ_i et de variance σ_i, alors X est distribuée suivant une gaussienne de valeur moyenne $\mu = \sum_{i=1,N}\mu_i$ et de variance $\sigma^2 = \sum_{i=1,N}\sigma_i^2$, dans la limite $N \to \infty$

- Exemple avec des variables x_i tirées aléatoirement selon une loi exponentielle entre 0 et 1

Dans ce cas, pour N=5, la distribution ne ressemble pas à une gaussienne

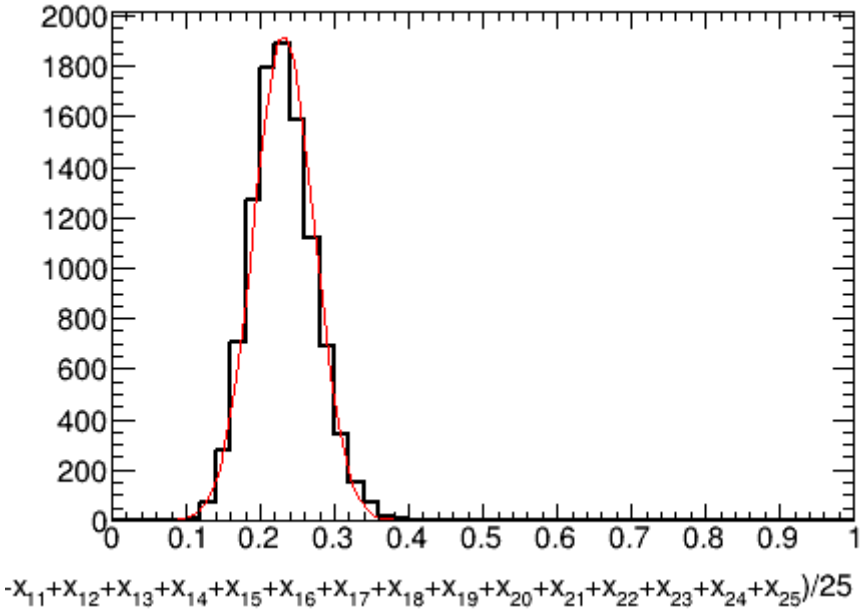


Le théorème central limite

- Théorème
  - Si une variable aléatoire X est formée de la somme de N variables aléatoires x_i indépendantes, de moyenne μ_i et de variance σ_i, alors X est distribuée suivant une gaussienne de valeur moyenne $\mu = \sum_{i=1,N}\mu_i$ et de variance $\sigma^2 = \sum_{i=1,N}\sigma_i^2$, dans la limite $N \to \infty$

- Exemple avec des variables x_i tirées aléatoirement selon une loi exponentielle entre 0 et 1

Pour N=25, la distribution ressemble presque à une gaussienne



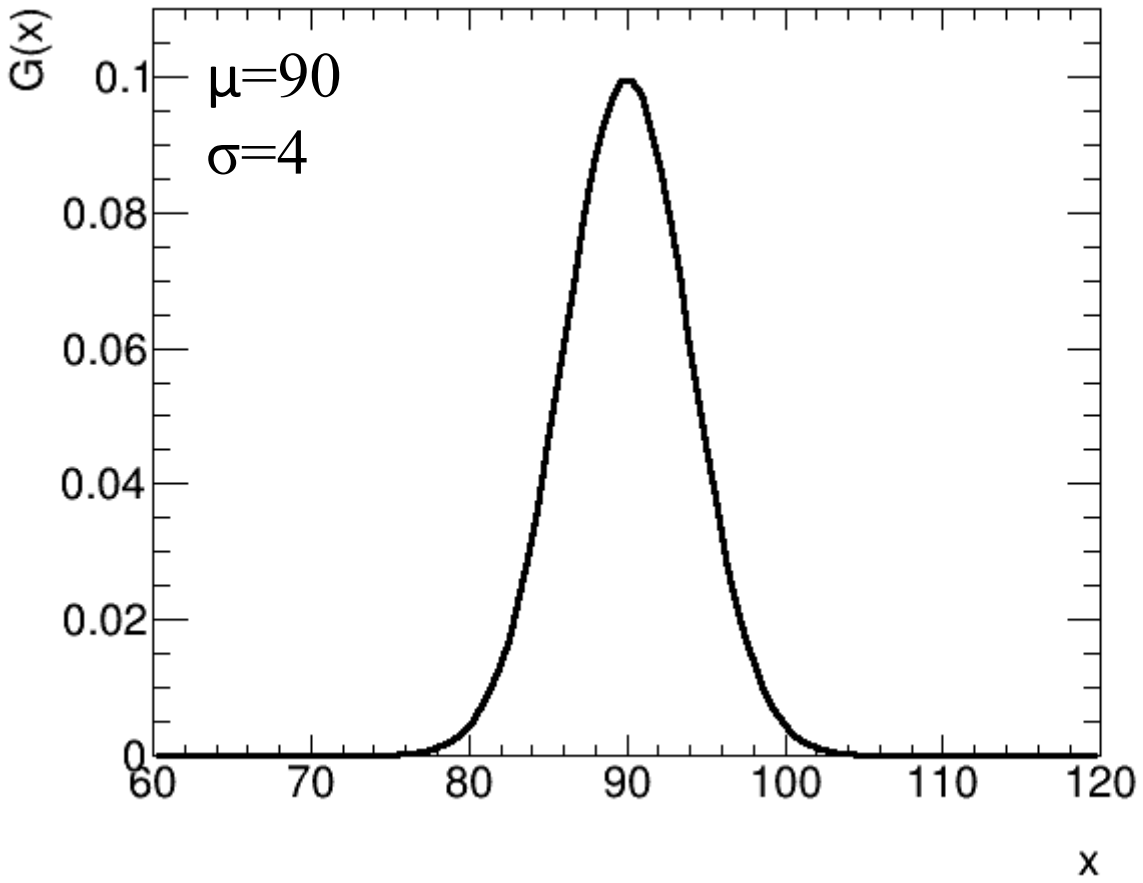
Le théorème central limite

- Théorème
  - Si une variable aléatoire X est formée de la somme de N variables aléatoires x_i indépendantes, de moyenne μ_i et de variance σ_i, alors X est distribuée suivant une gaussienne de valeur moyenne $\mu = \sum_{i=1,N}\mu_i$ et de variance $\sigma^2 = \sum_{i=1,N}\sigma_i^2$, dans la limite $N \rightarrow \infty$

- Conséquence:
  - C'est pour cela que les physiciens aiment les gaussiennes
  - Résultat d'une mesure (énergie, position…) peut-être vu comme provenant d'une suite de contributions, chacune apportant sa fluctuation
    - Fluctuation du nombres de traces déposant de l'énergie dans un cristal, fluctuation des dépôts d'énergie de ces traces, fluctuation du bruit électronique, etc…
  - Cela n'a rien de surprenant si en bout de chaîne on se retrouve avec des quantités qui ont des distributions gaussiennes

Estimateur au maximum de vraisemblance

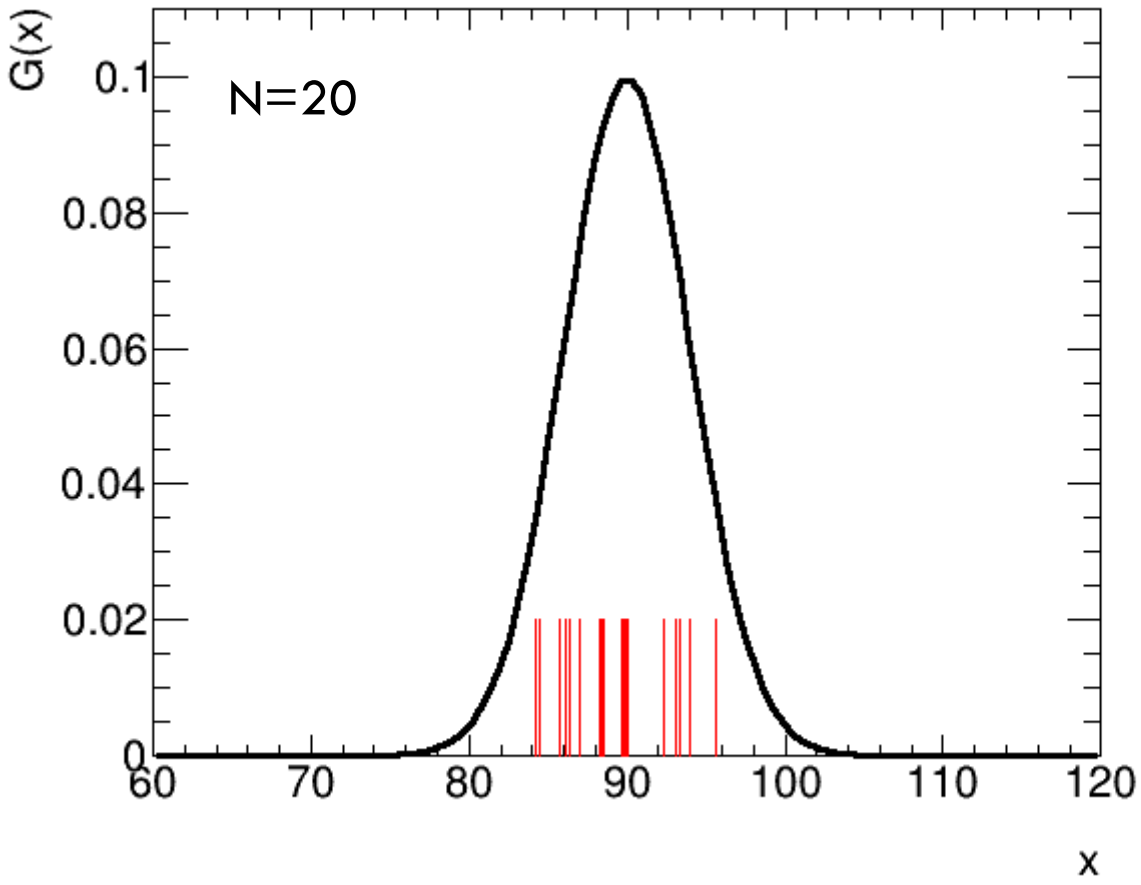
- Supposons que l'on collecte les *N* valeurs $\{x_i\}$ et que la distribution sous-jacente est une gaussienne



Estimateur au maximum de vraisemblance

- Supposons que l'on collecte les N valeurs $\{x_i\}$ et que la distribution sous-jacente est une gaussienne

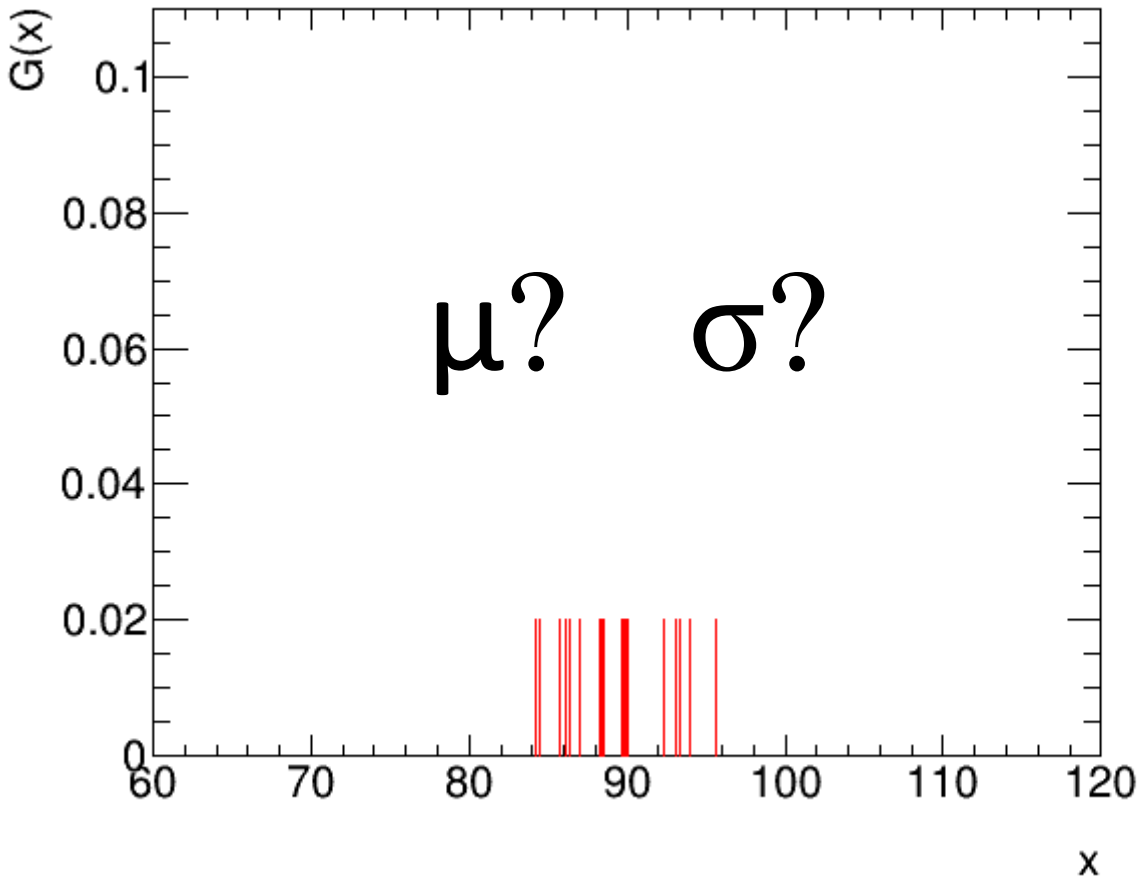
Mesure:
1. 94.0
2. 88.3
3. 93.1
4. 89.9
5. 93.3
6. 89.8
7. 86.4
8. 89.7
9. 90.0
10. 88.4
11. 95.6
12. 86.1
13. 89.8
14. 84.2
15. 85.8
16. 84.4
17. 93.1
18. 87.1
19. 92.3
20. 88.5



Estimateur au maximum de vraisemblance

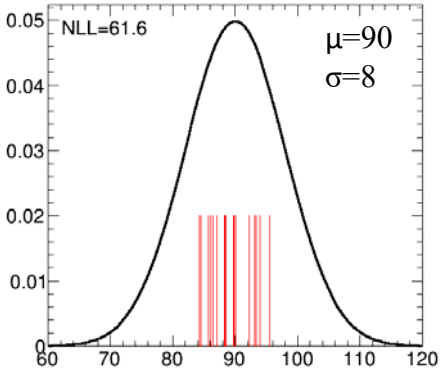
- Supposons que l'on collecte les N valeurs $\{x_i\}$ et que la distribution sous-jacente est une gaussienne

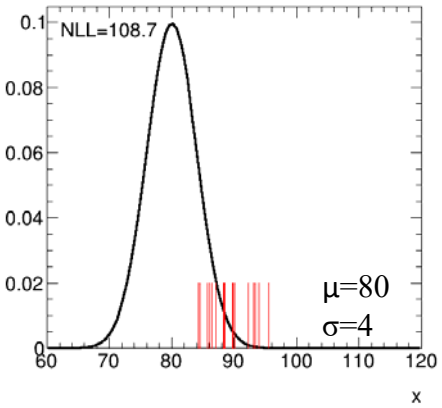
Mesure:
1. 94.0
2. 88.3
3. 93.1
4. 89.9
5. 93.3
6. 89.8
7. 86.4
8. 89.7
9. 90.0
10. 88.4
11. 95.6
12. 86.1
13. 89.8
14. 84.2
15. 85.8
16. 84.4
17. 93.1
18. 87.1
19. 92.3
20. 88.5

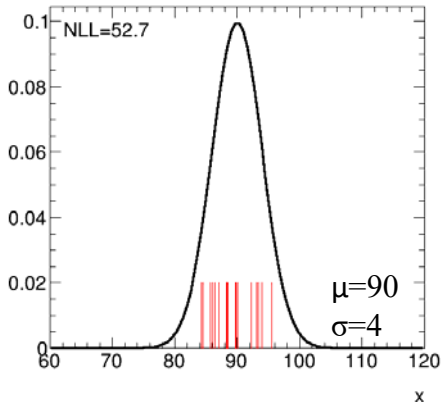


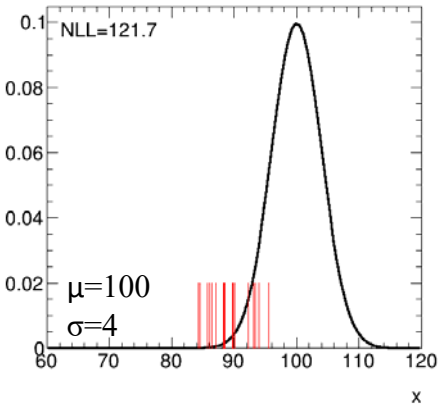
Estimateur au maximum de vraisemblance

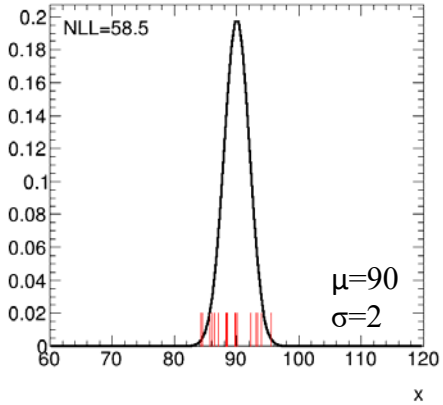
$$-\ln L(\mu,\sigma) = -\ln\left(\prod_{i=1}^{N} \frac{e^{-\frac{(x_i-\mu)^2}{2\sigma^2}}}{\sqrt{2\pi\sigma^2}}\right)$$











Estimateur au maximum de vraisemblance

- Dans ce cas, la vraisemblance est:

$$L(\mu,\sigma) = \prod_{i=1}^{N} \frac{e^{-\frac{(x_i-\mu)^2}{2\sigma^2}}}{\sqrt{2\pi\sigma^2}}$$

- En pratique, on préfère minimiser:

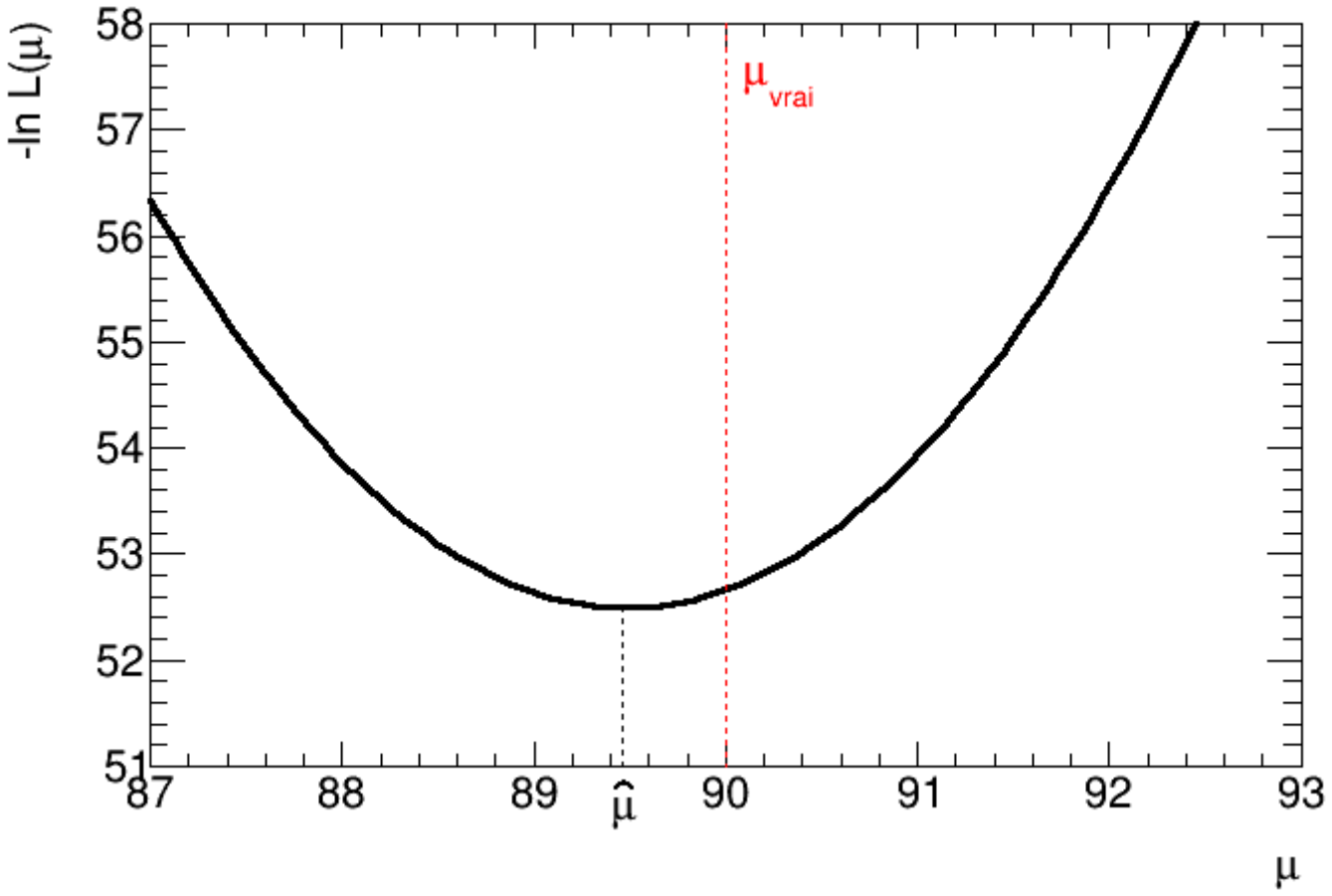
$$-\ln L(\mu,\sigma) = -N \ln \frac{1}{\sqrt{2\pi\sigma^2}} + \sum_{i=1}^{N} \frac{(x_i-\mu)^2}{2\sigma^2}$$

- La minimisation de $-\ln L(\mu,\sigma)$ donne:

$$\begin{cases} \dfrac{\partial -\ln L(\mu,\sigma)}{\partial \mu} = 0 \\ \dfrac{\partial -\ln L(\mu,\sigma)}{\partial \sigma} = 0 \end{cases} \Rightarrow \begin{cases} \hat{\mu} = \dfrac{1}{n}\sum_{i=0}^{n} x_i \\ \hat{\sigma} = \sqrt{\dfrac{1}{n}\sum_{i=0}^{n} (x_i - \hat{\mu})^2} \end{cases}$$

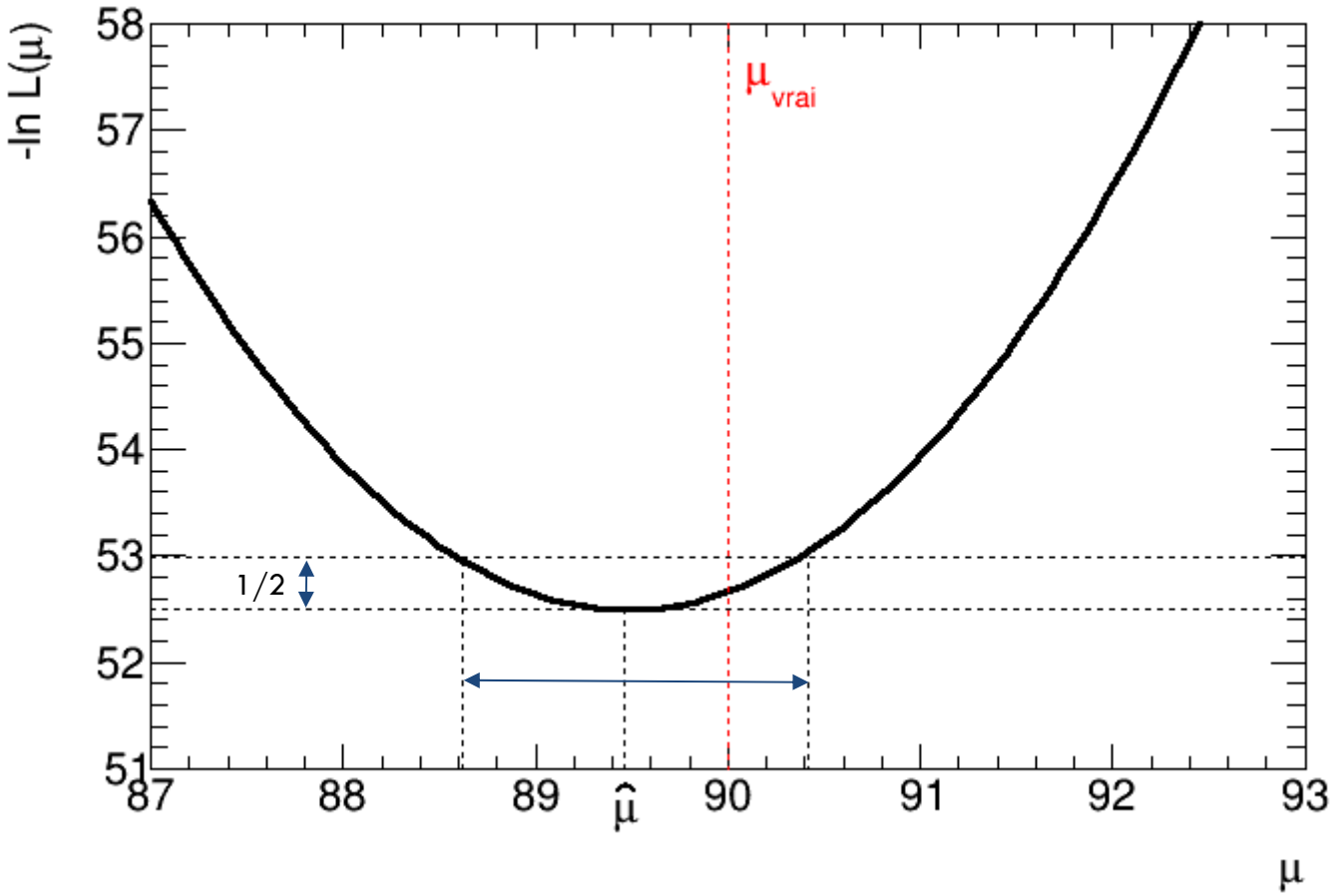
Estimateur au maximum de vraisemblance

Supposons que σ soit connu:



$\hat{\mu}=89.5$

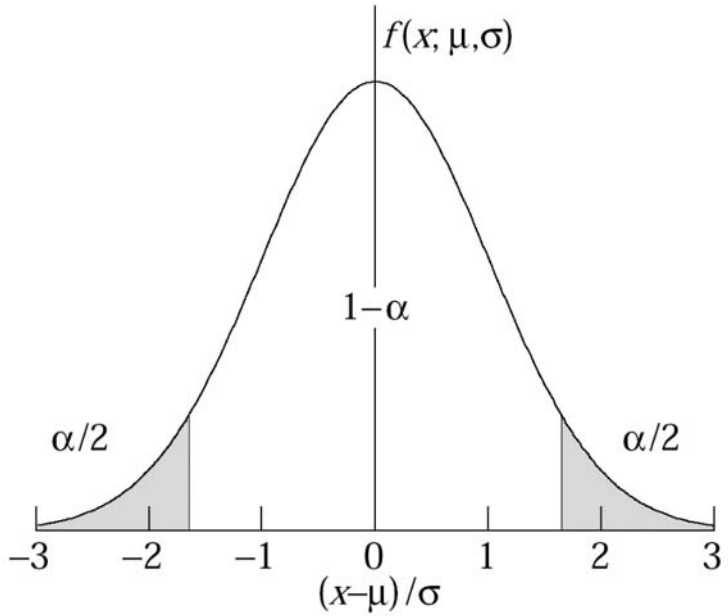
Estimateur au maximum de vraisemblance



$\hat{\mu}=89.5\pm1.0$ (Intervalle de confiance à 68% ou 1σ)

Si on répéte cette expérience, dans 68% des cas, l'intervalle de confiance va contenir la valeur vraie

Estimateur au maximum de vraisemblance

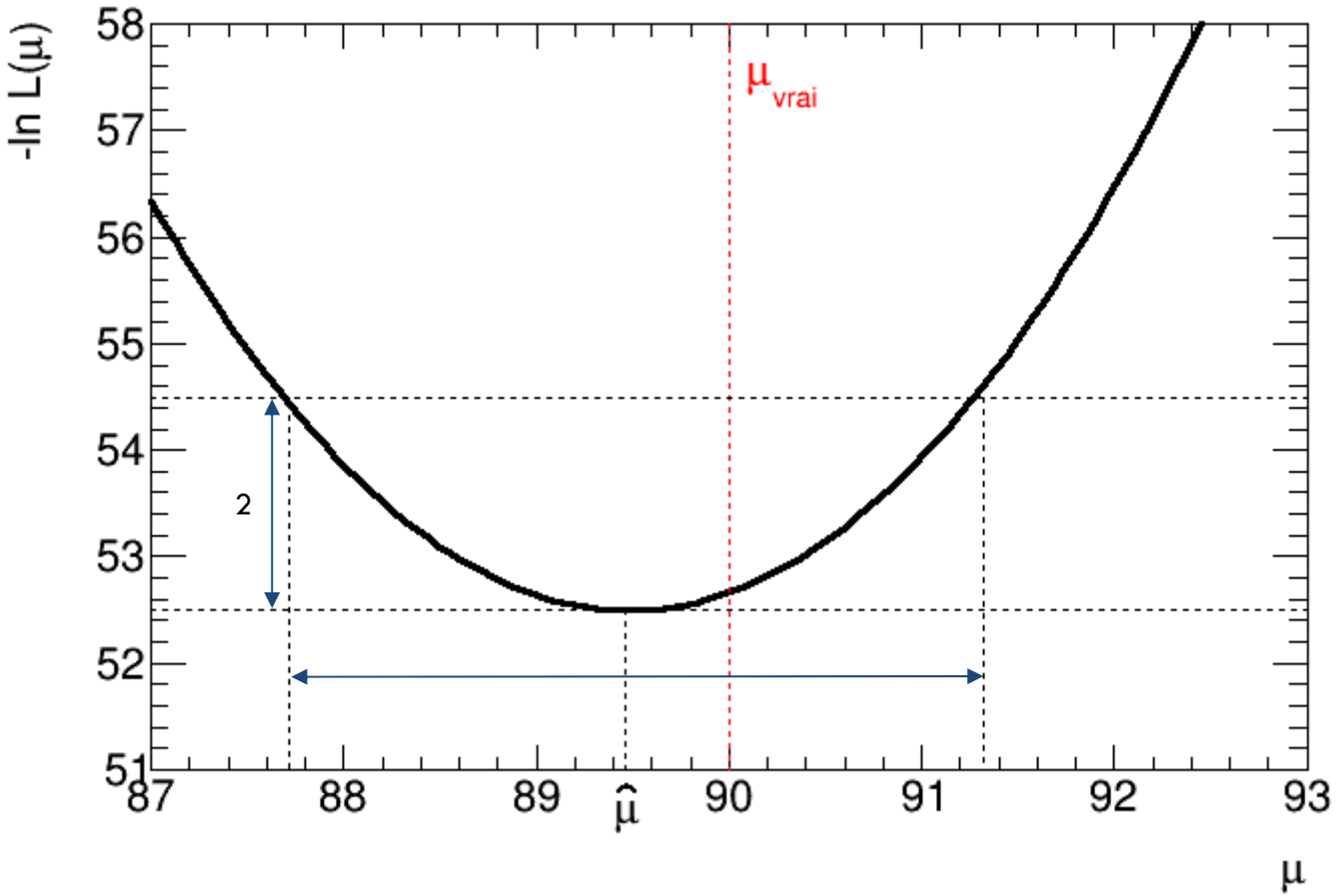


α (%)	δ
31.73	1σ
4.55	2σ
0.27	3σ
6.3×10^{-3}	4σ
5.7×10^{-5}	5σ
2.0×10^{-7}	6σ

$\hat{\mu}=89.5\pm1.0$ (Intervalle de confiance à 68% ou 1σ)

Si on répéte cette expérience, dans 68% des cas, l'intervalle de confiance va contenir la valeur vraie

Estimateur au maximum de vraisemblance

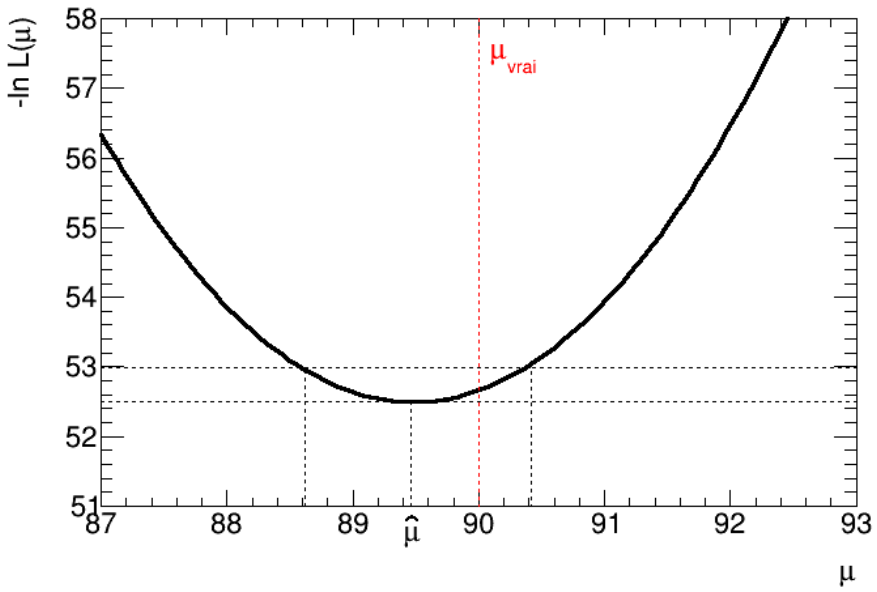


$\hat{\mu}=89.5\pm1.9$ (Intervalle de confiance à 95% ou 2σ)

Si on répéte cette expérience, dans 95% des cas, l'intervalle de confiance va contenir la valeur vraie

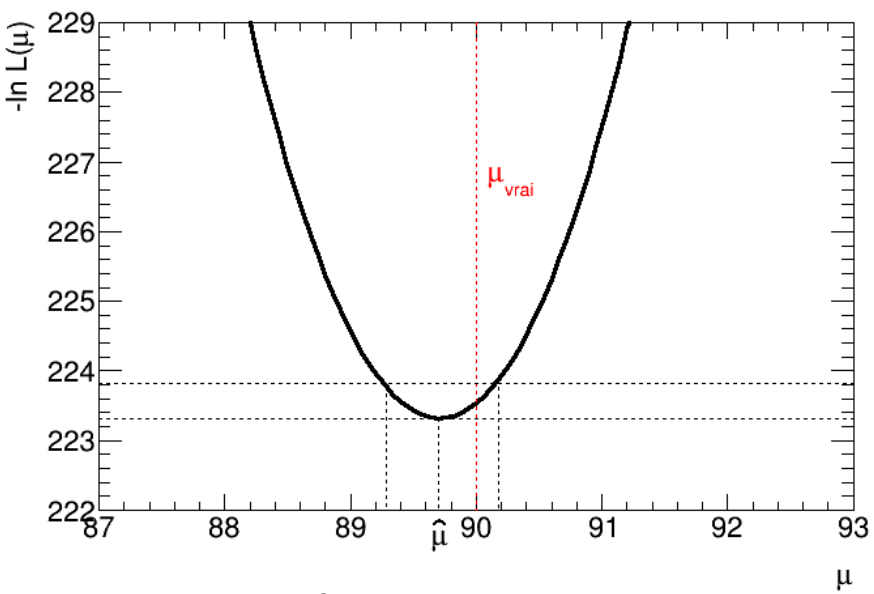
Estimateur au maximum de vraisemblance

N=20



$\hat{\mu}=89.5\pm1.0$

N=80

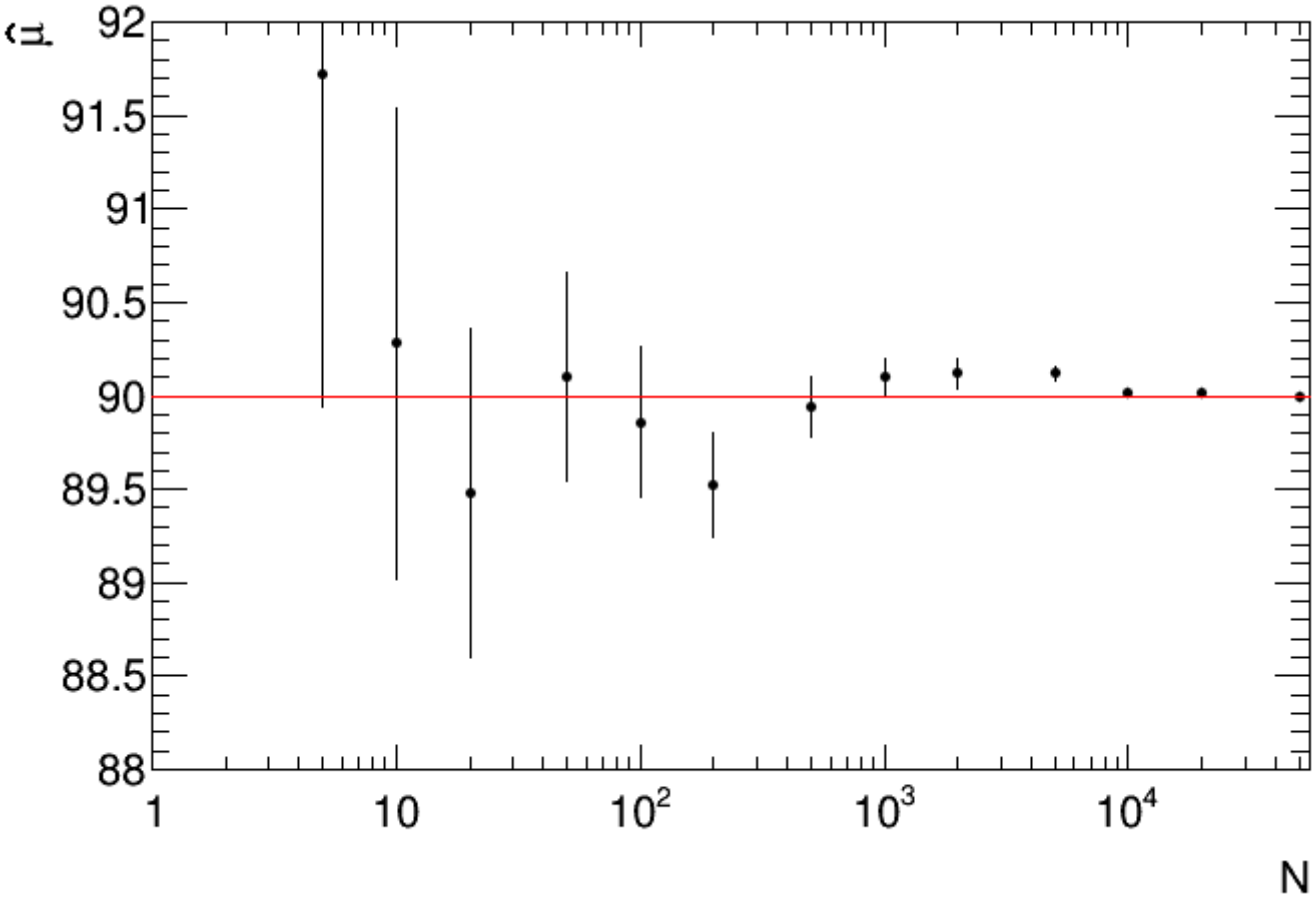


$\hat{\mu}=89.7\pm0.5$

- En passant de N=20 à N=80, l'incertitude est divisé par un facteur 2
- De façon générale, l'incertitude décroit en $1/\sqrt{N}$

Estimateur au maximum de vraisemblance

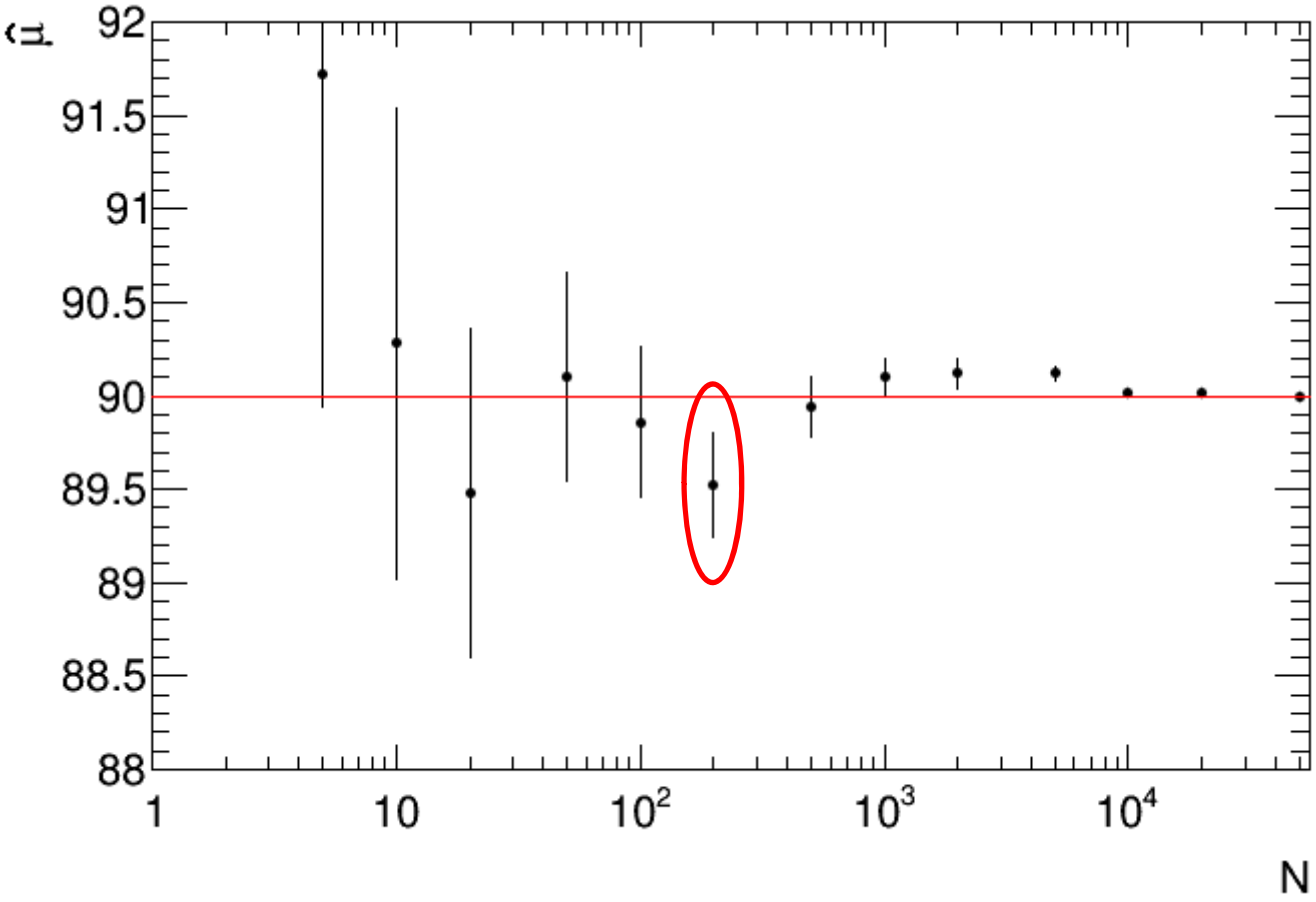
Intervalle de confiance à 1 σ



- L'incertitude décroit en $1/\sqrt{N}$
- L'estimateur de μ tend vers la vraie valeur de μ

Estimateur au maximum de vraisemblance

Intervalle de confiance à 1 σ

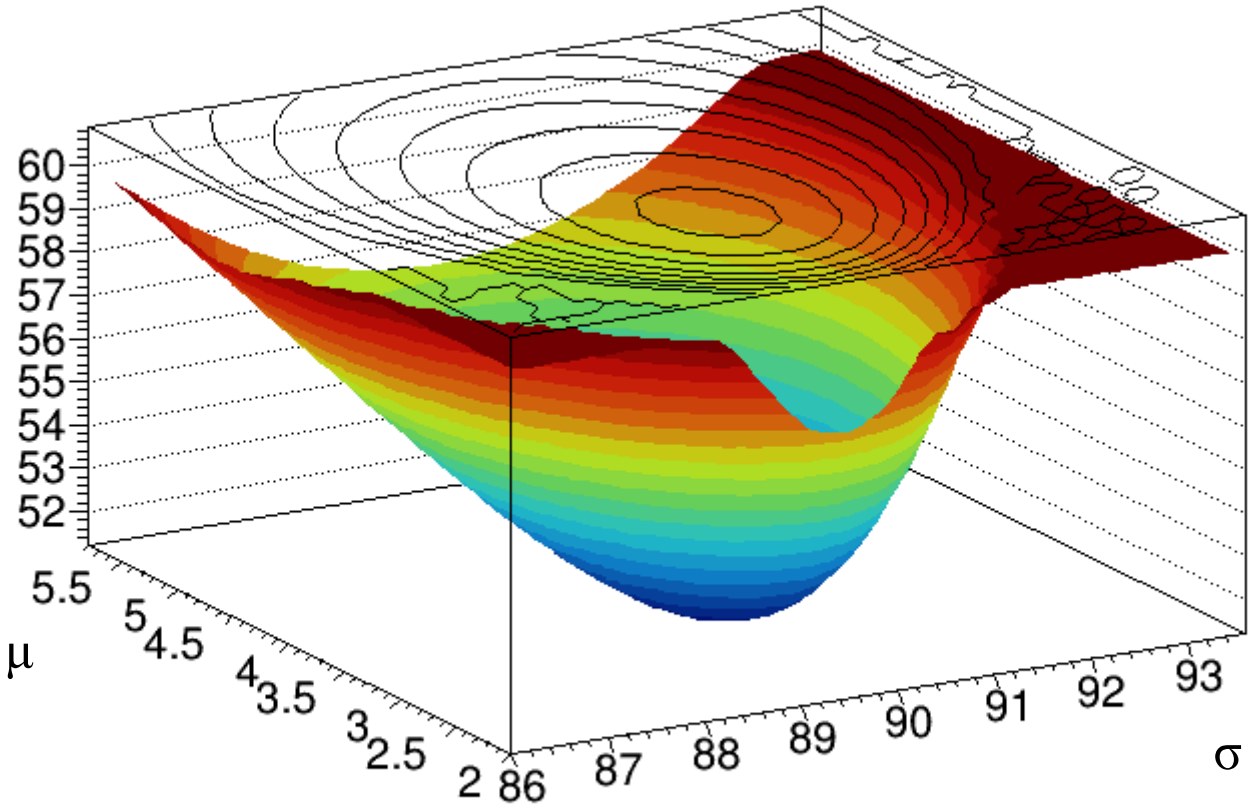


- L'incertitude décroit en $1/\sqrt{N}$
- L'estimateur de μ tend vers la vraie valeur de μ

Estimateur au maximum de vraisemblance

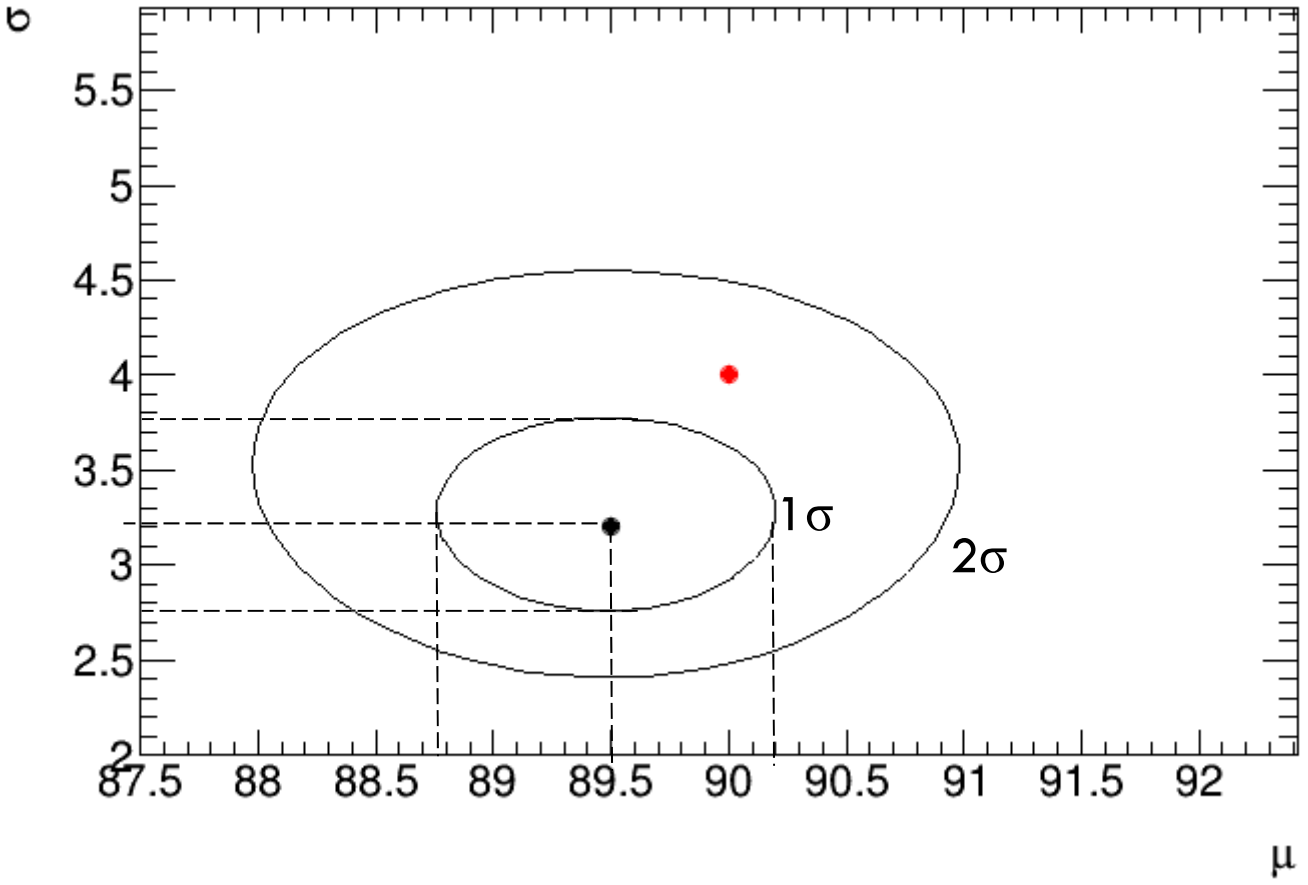
- Et maintenant avec σ et μ inconnus

$-\ln L(\mu,\sigma)$



N=20

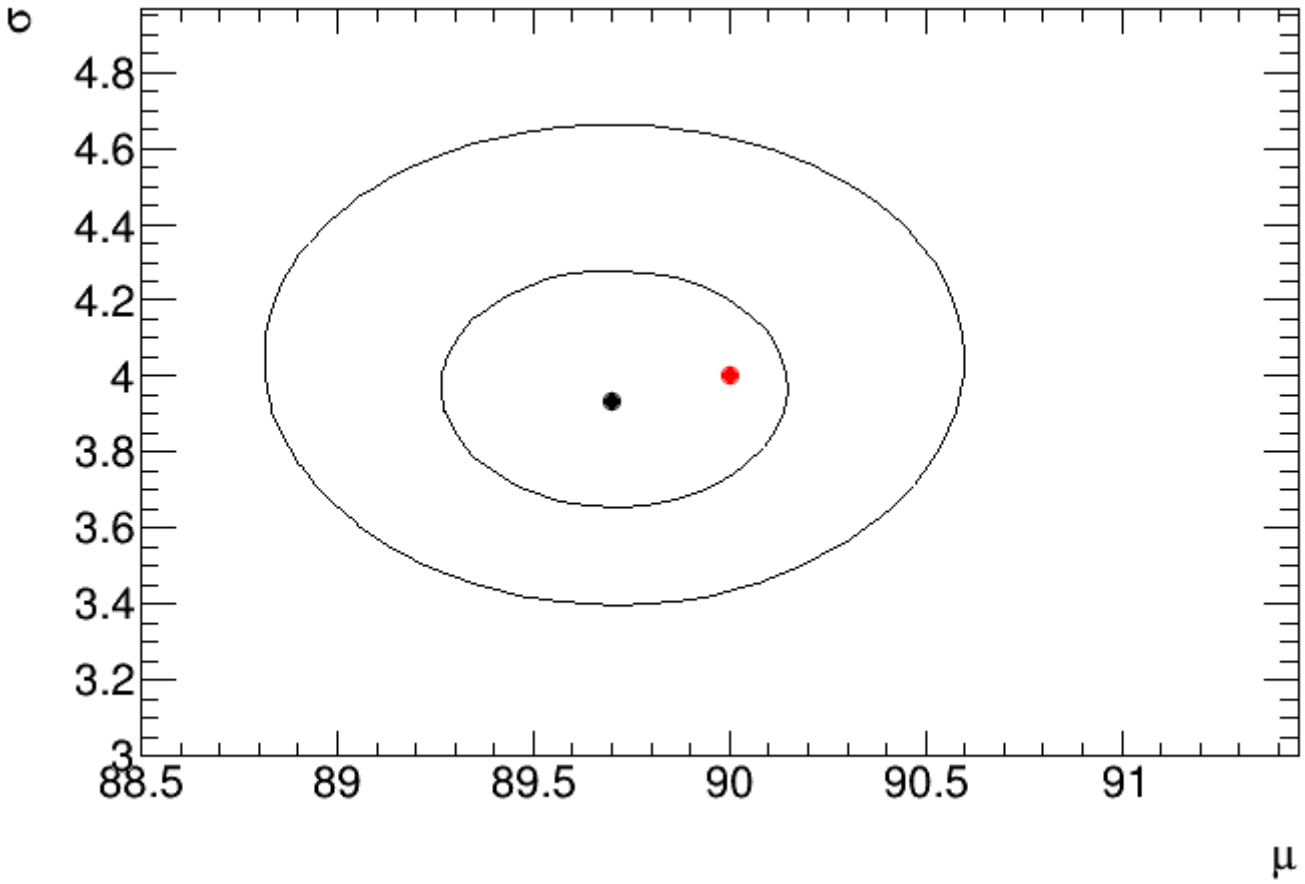
Estimateur au maximum de vraisemblance



Incertitudes asymétriques sur σ du au petit N

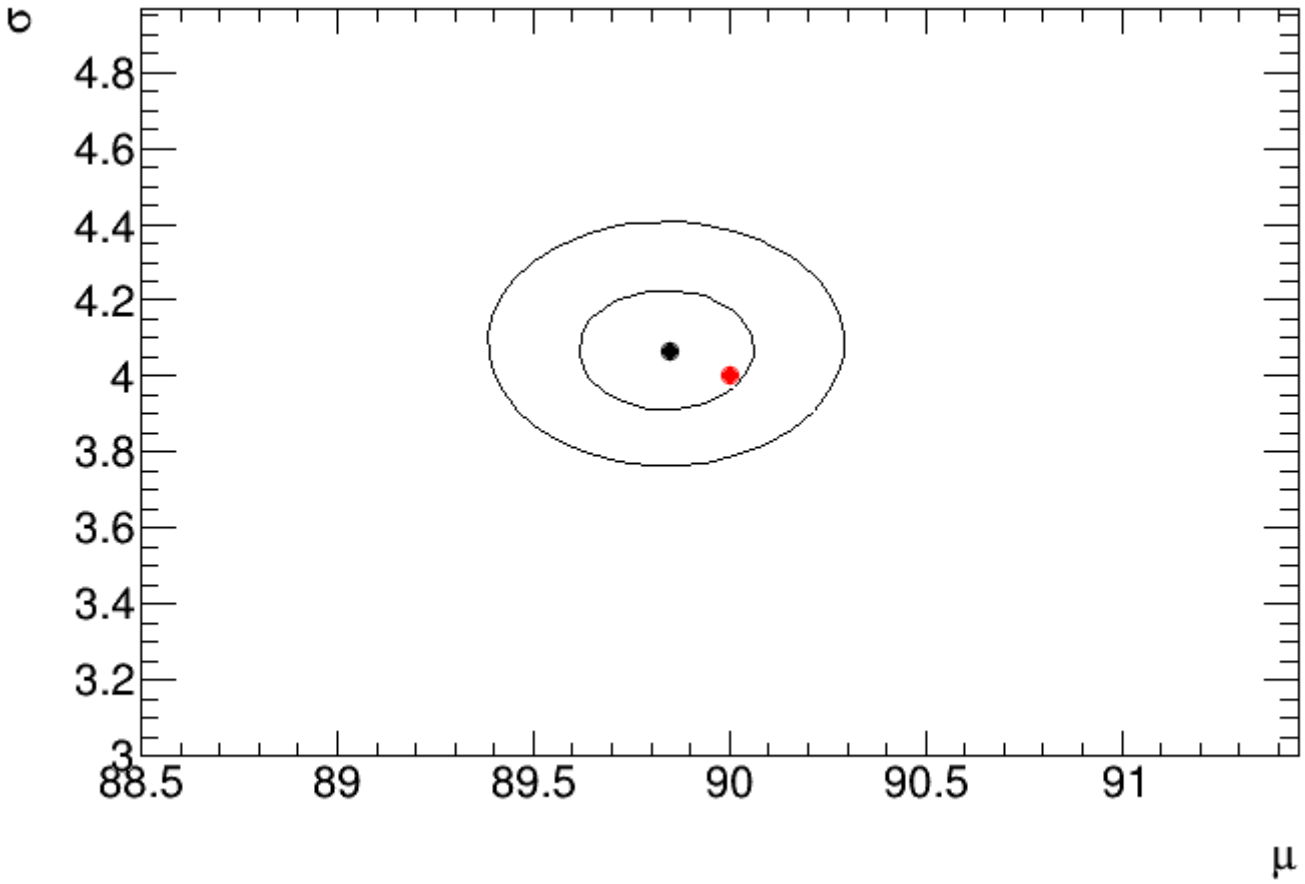
N=20

Estimateur au maximum de vraisemblance



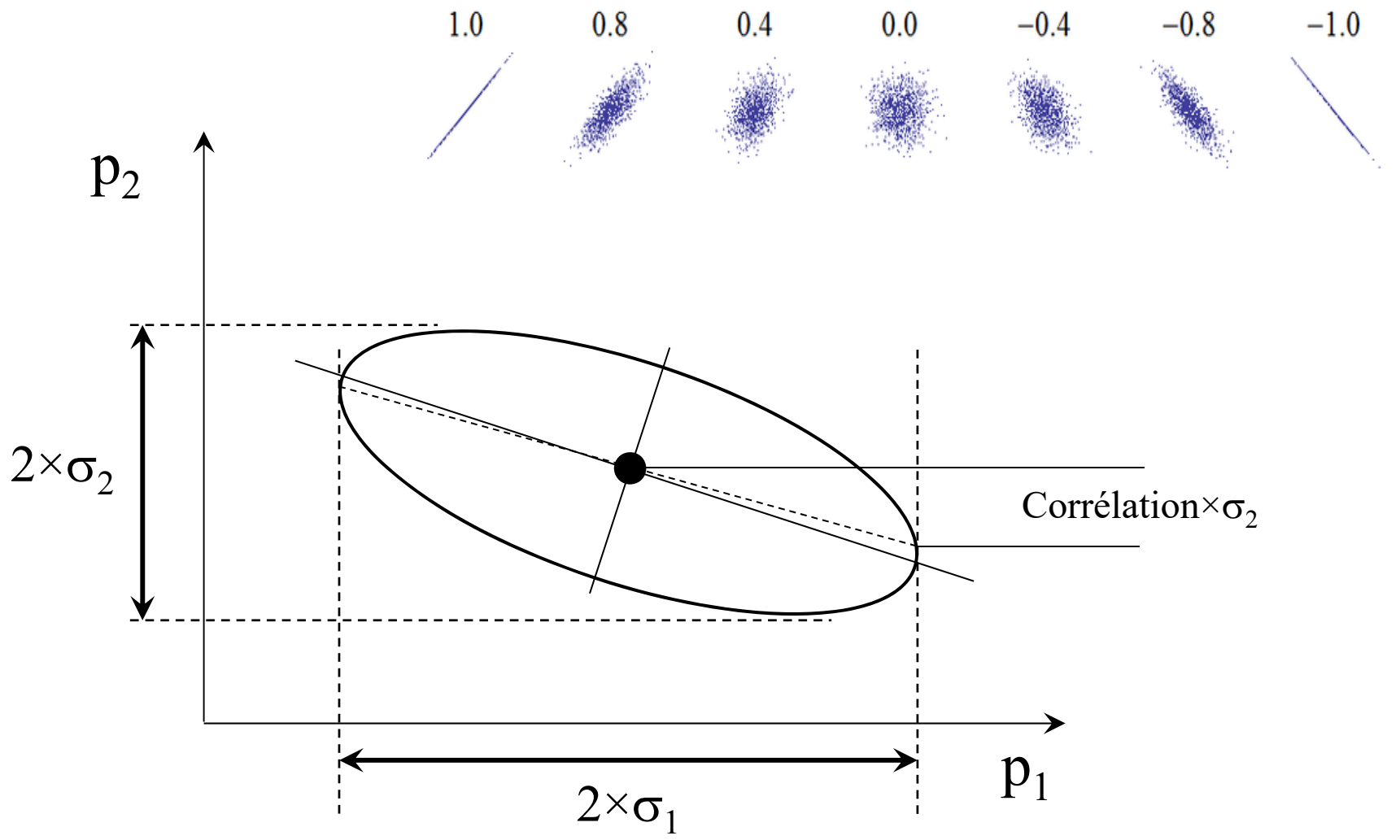
N=80 incertitudes divisées par un facteur ~2 comparé à N=20

Estimateur au maximum de vraisemblance



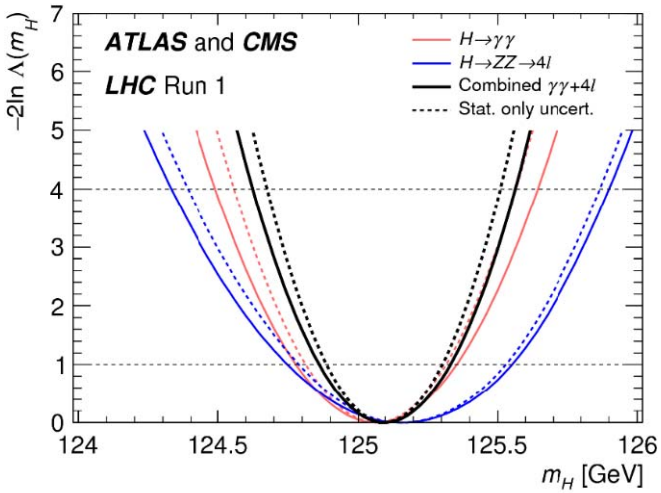
N=320 incertitudes divisées par un facteur 2 comparé à N=80

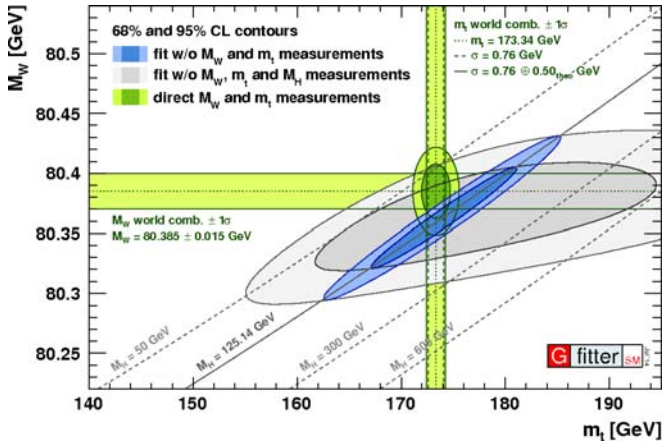
Estimateur au maximum de vraisemblance

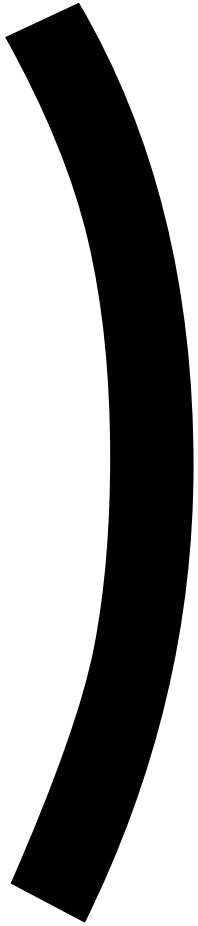


Avant de revenir à la calibration

- La méthode du maximum de vraisemblance est une méthode efficace pour trouver des estimateurs de paramètres inconnues
  - L'incertitude statistique décroit en $1/\sqrt{N}$
- Elle est utilisée intensivement et dans des cas beaucoup plus complexes (modèle, nb de paramètres,…) et la minimisation est faite à l'aide du programme `minuit`
- Il existe d'autres méthodes pour construire des estimateurs:
  - Méthode du Chi2
  - Méthodes des moments



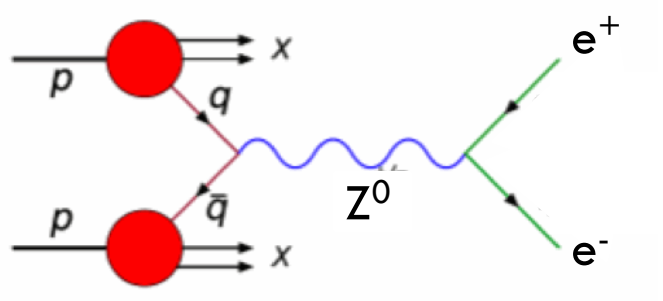




Etalonnage avec la masse du Z

- La masse du Z est connue avec une très grande précision depuis LEP (91.1876(21) GeV)

$$M_{12} = \sqrt{2E_1E_2(1 - cos\theta_{12})}$$

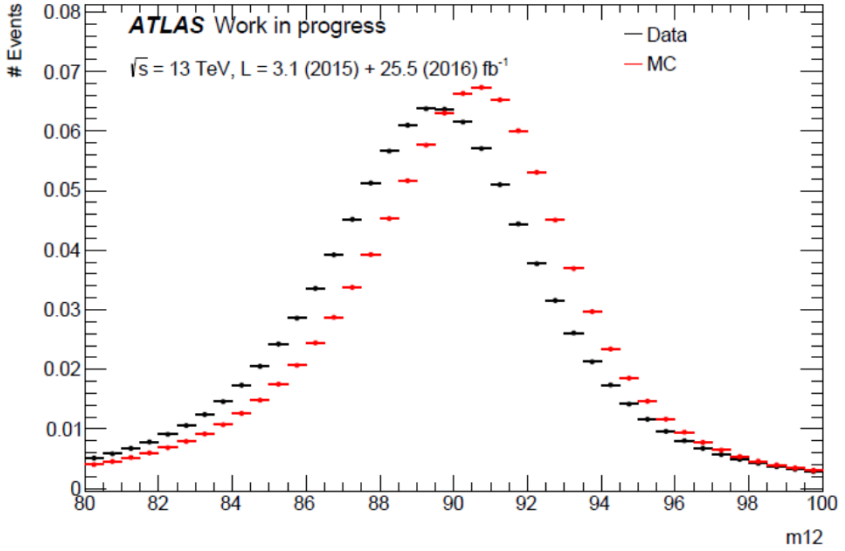


- Position du maximum

→échelle d'énergie des électrons

- Largeur de la distribution

→ résolution en énergie

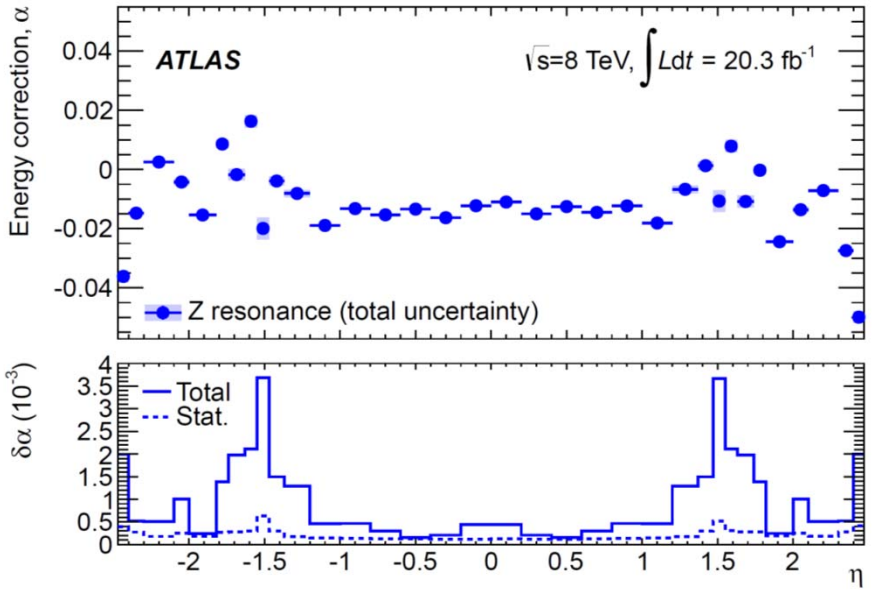


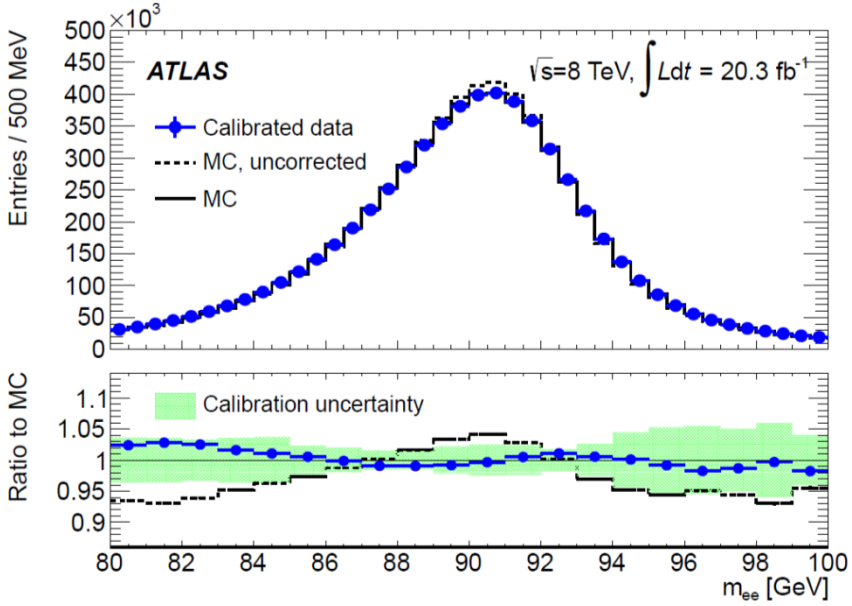
- Utilisation de la méthode du maximum de vraisemblance pour extraire les facteurs corrections

$$E_{meas} = E_{true}\,(1 + \alpha)$$

Etalonnage avec la masse du Z

- Facteur de corrections calculés en fonction de la pseudo-rapidité
- Incertitude total dominée par les incertitudes systématiques de la méthode (test de fermeture avec la simulation)

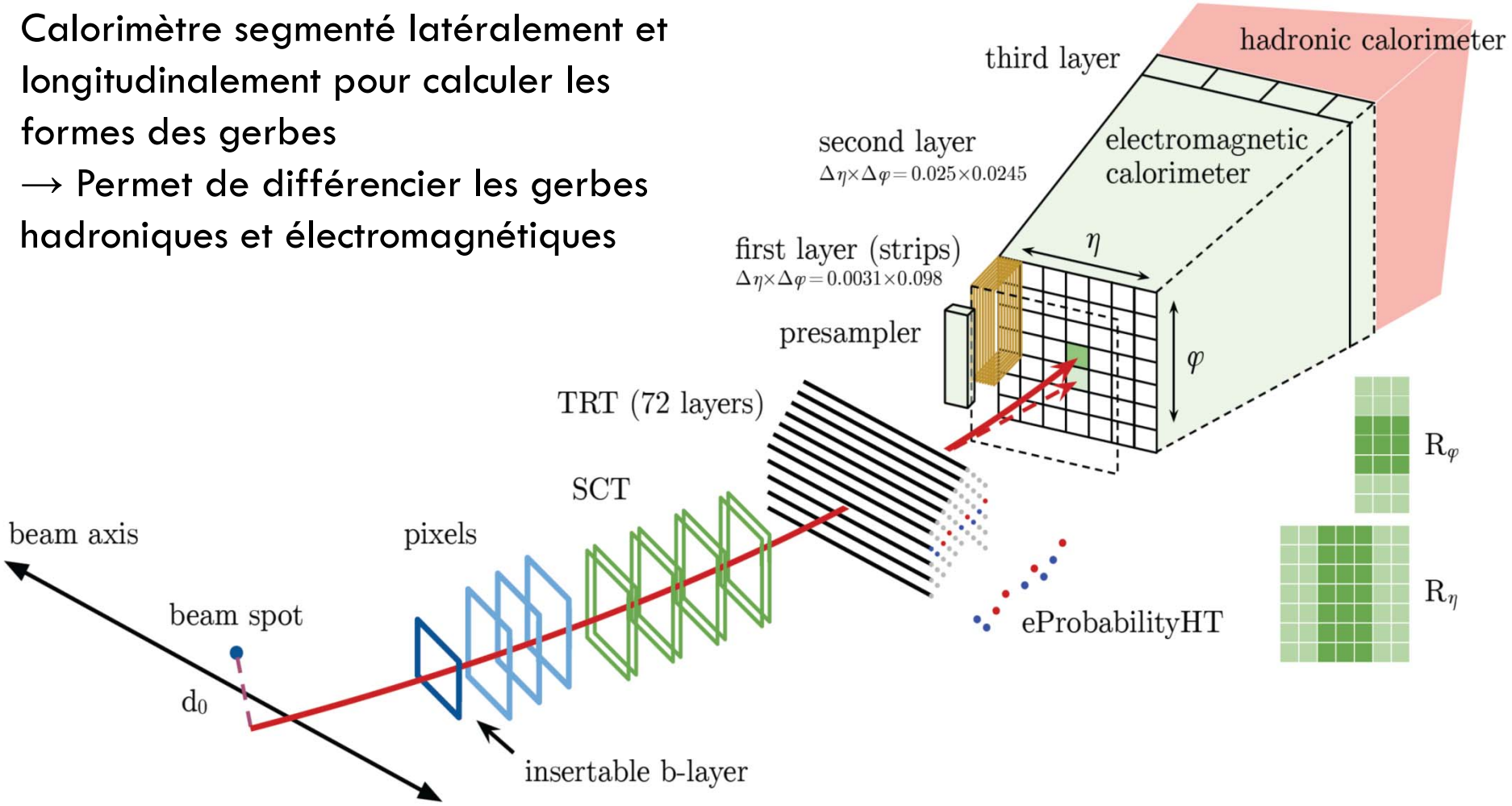




La reconstruction et la sélection des électrons

Calorimètre segmenté latéralement et longitudinalement pour calculer les formes des gerbes

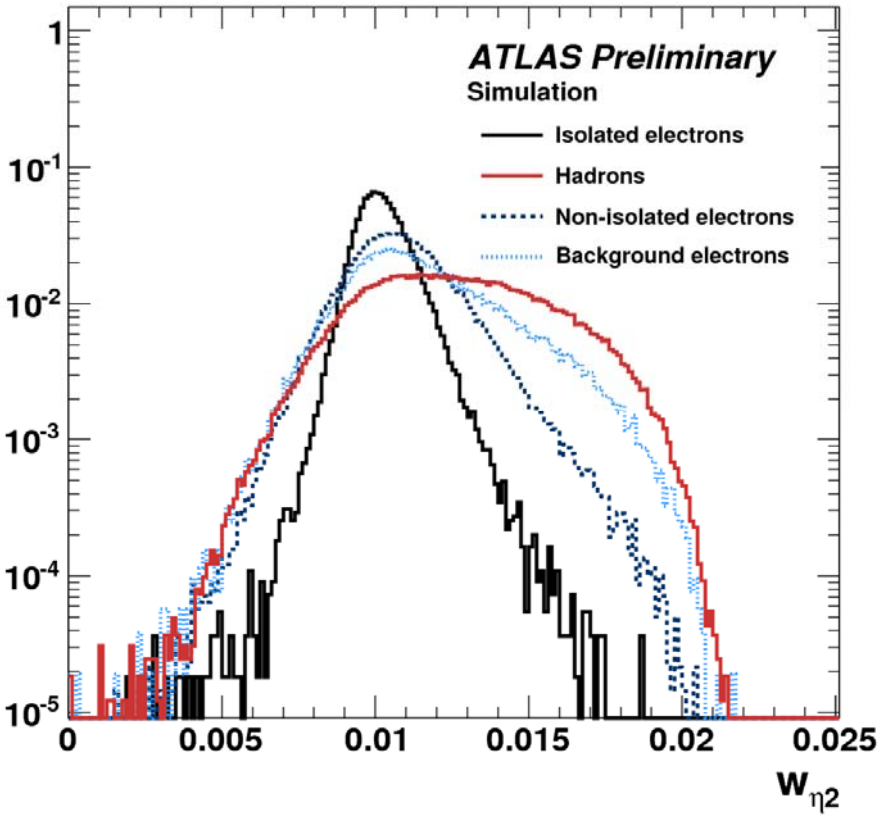
→ Permet de différencier les gerbes hadroniques et électromagnétiques



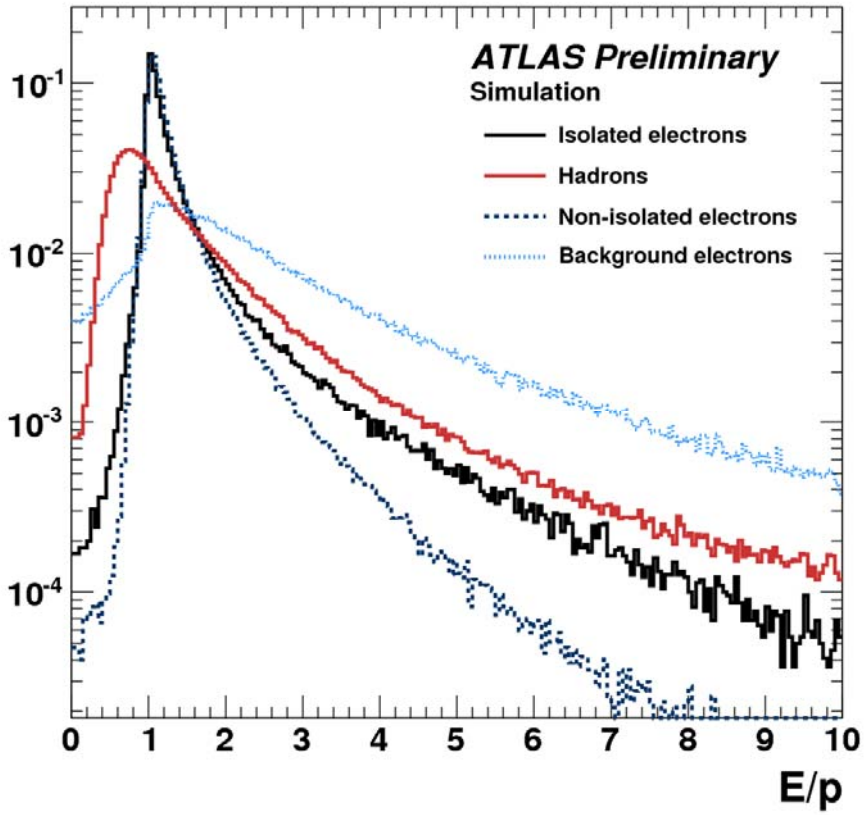
Les informations du trajectographe sont également utilisées pour identifier les électrons

Exemple de variables discriminantes

Extension latérale de la gerbe

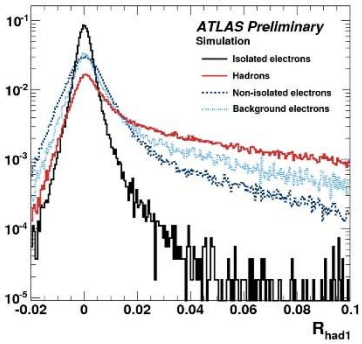


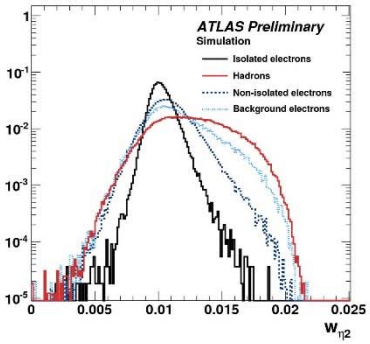
Rapport E/p

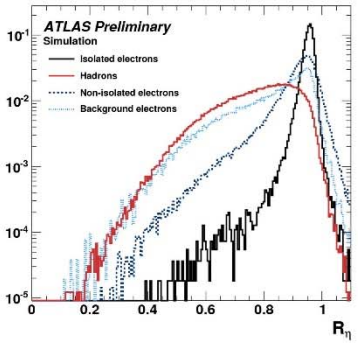


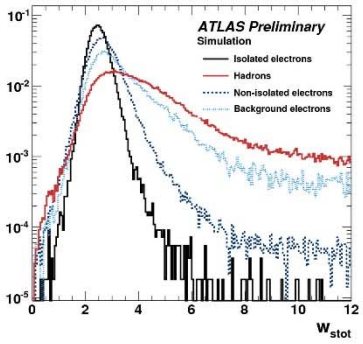
$$w_{\eta 2} = \sqrt{\frac{\sum_i E_i \eta_i^2}{\sum_i E_i} - \left(\frac{\sum_i E_i \eta_i}{\sum_i E_i}\right)^2},$$

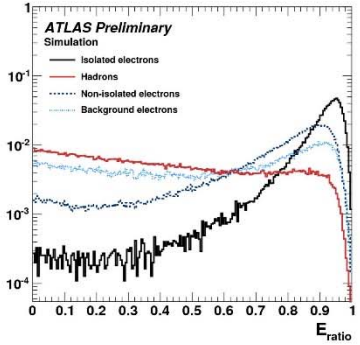
Encore plus de variables discriminantes

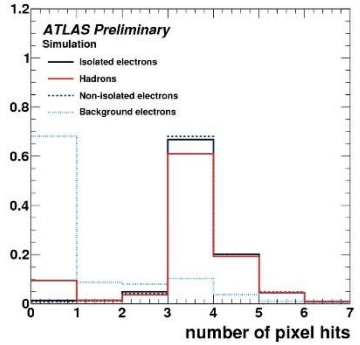


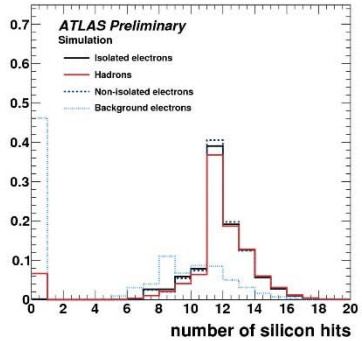


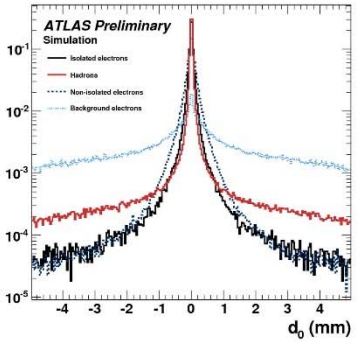


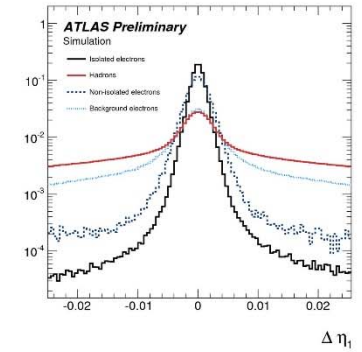


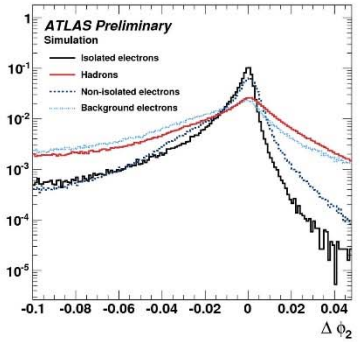


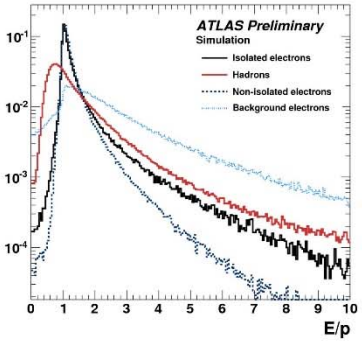


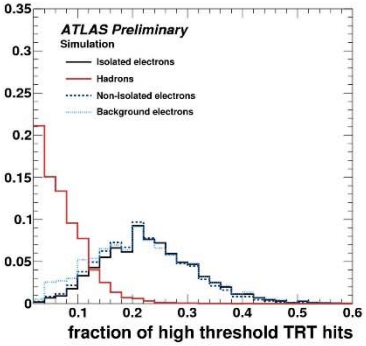


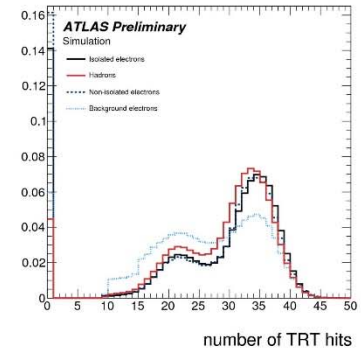


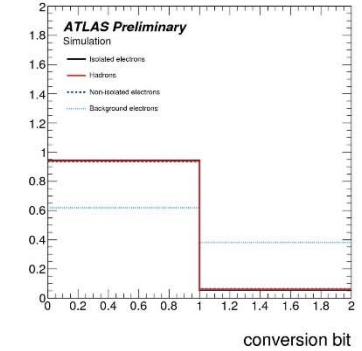


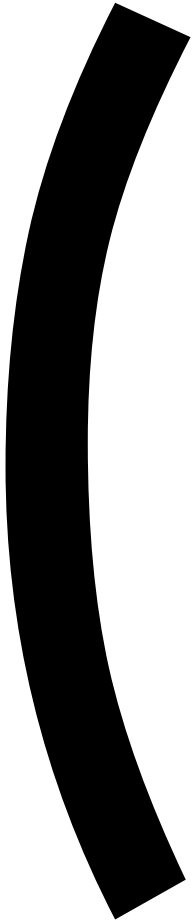






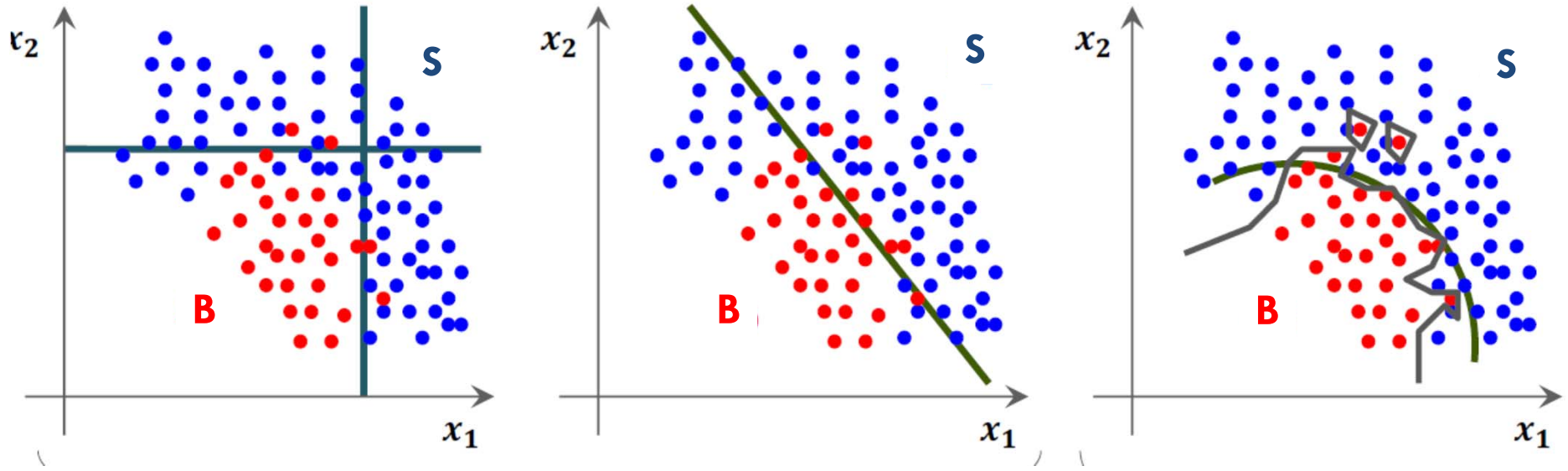






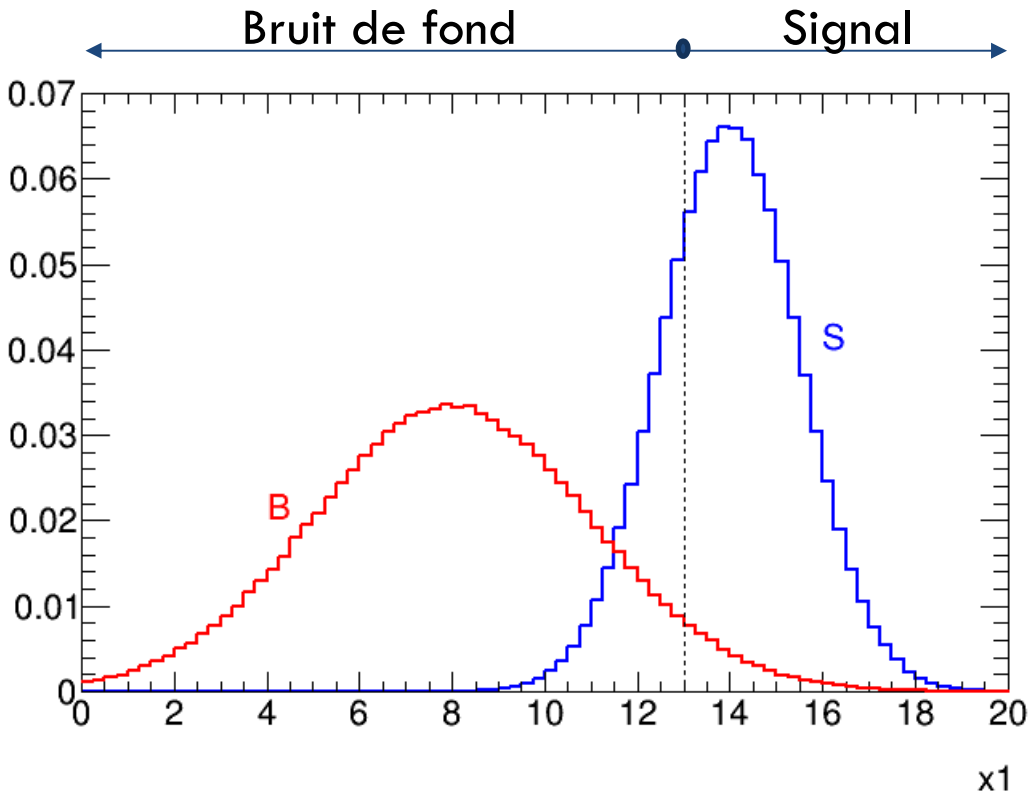
Test d'hypothèse

- Supposons un lot de données avec deux catégories B ou S (souvent noté H_0 et H_1)
  - Ex: S=electron et B=jet
- Les variables discriminantes sont $x_1, x_2, \ldots, x_n$
- Comment sélectionner le signal et rejeter le bruit de fond?



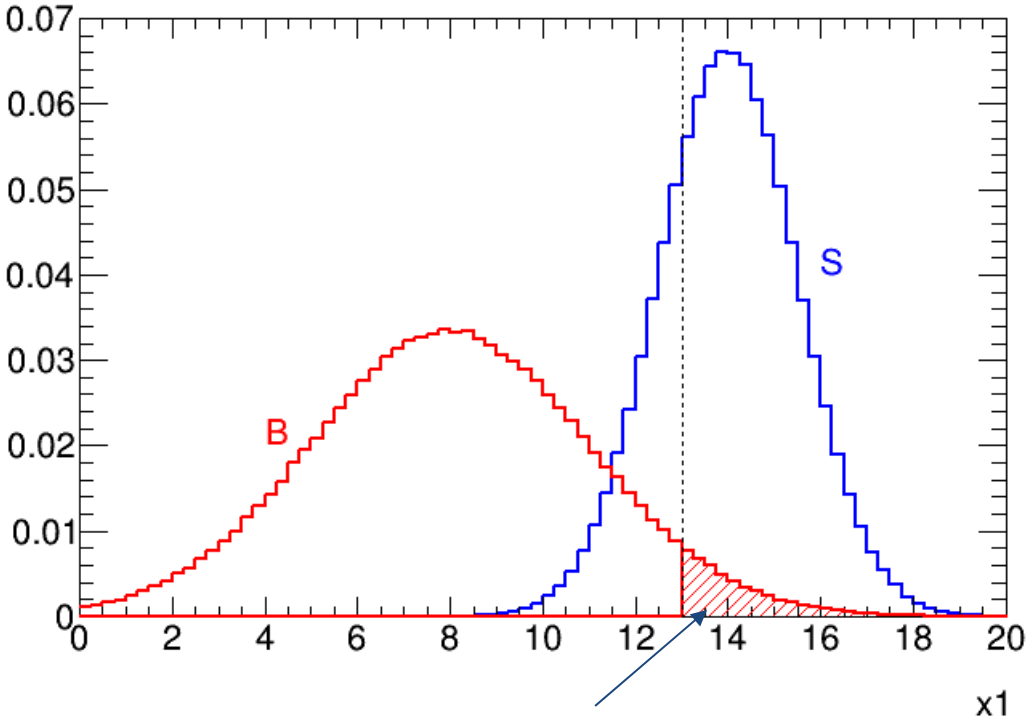
Test d'hypothèse

- Il y a deux façons de se tromper lors d'un test statistique :
  - Erreur de type-1: rejeter l'hypothèse B alors qu'elle est vraie
  - Erreur de type-2: retenir l'hypothèse B alors qu'elle est fausse



Test d'hypothèse

- Il y a deux façons de se tromper lors d'un test statistique :
  - Erreur de type-1: rejeter l'hypothèse B alors qu'elle est vraie
  - Erreur de type-2: retenir l'hypothèse B alors qu'elle est fausse

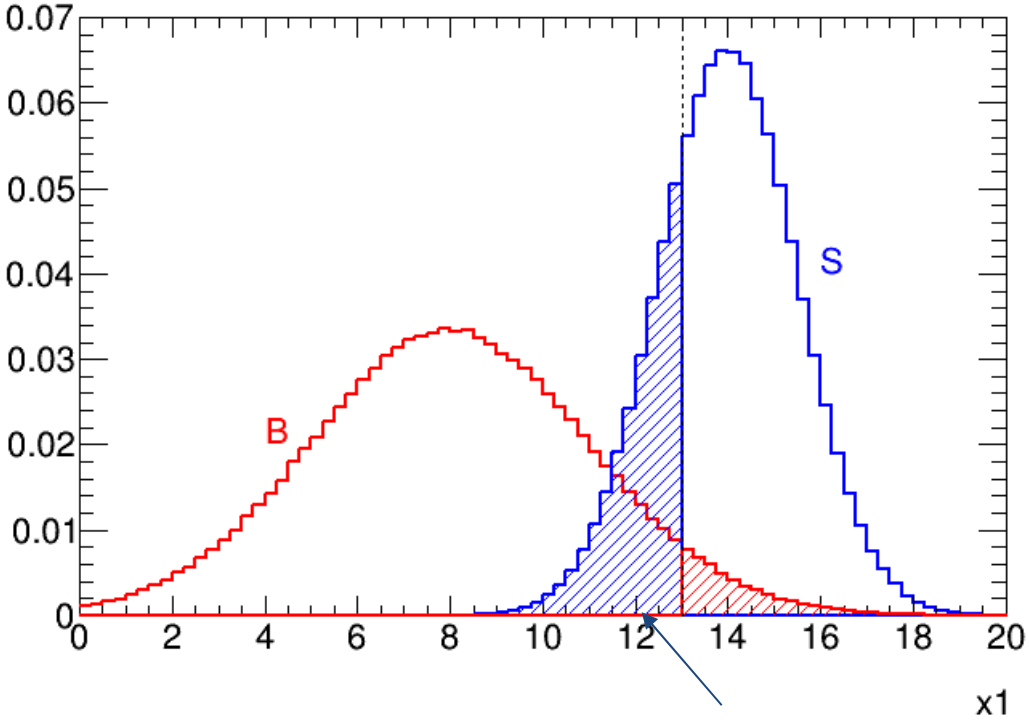


Taux d'erreur de type-1 (noté α)

Bruit de fond identifié comme du signal

Test d'hypothèse

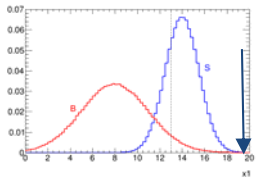
- Il y a deux façons de se tromper lors d'un test statistique :
  - Erreur de type-1: rejeter l'hypothèse B alors qu'elle est vraie
  - Erreur de type-2: retenir l'hypothèse B alors qu'elle est fausse

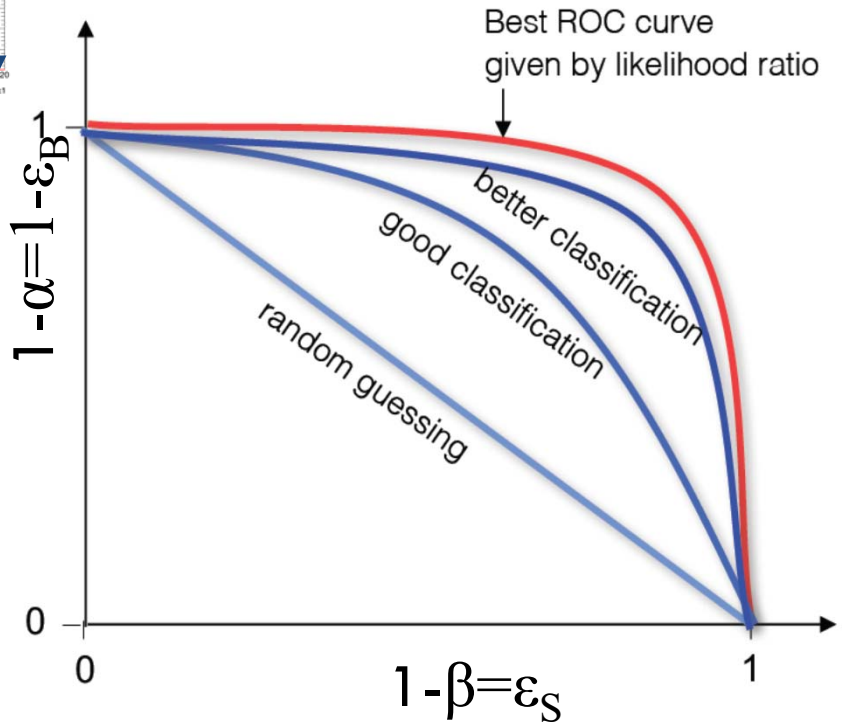


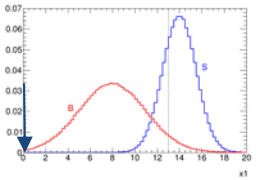
Taux d'erreur de type-2 (noté β)

1- β: efficacité de sélection du signal également appelé puissance du test

ROC (Receiver Operation Characteristic)







Choix du point de fonctionnement dépend du cas d'utilisation et de l'abondane du signal et du bruit de fond

- Maximisation de $N_S/\sqrt{N_B}$ ou de $N_S/\sqrt{(N_B+N_S)}$
- Grande pureté de selection (mesure de précision)
- Grande efficacité de selection (trigger)

Lemme de Neyman-Pearson

- Le test statistique le plus puissant est le rapport de vraisemblance:

$$y(\vec{x}) = \frac{L(\vec{x} \mid S)}{L(\vec{x} \mid B)}$$

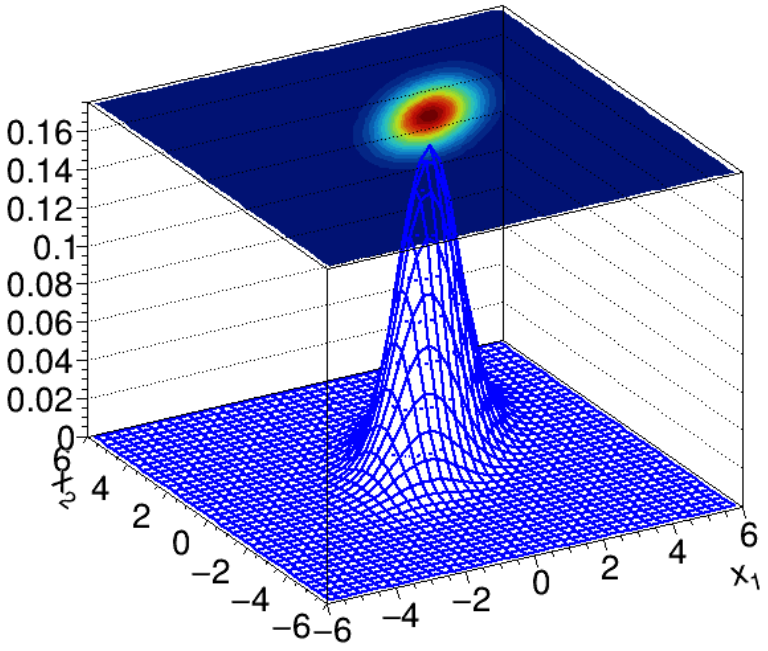
- Pour une valeur de α donnée, il aura la plus grande puissance (1-β).
- On peut également utiliser y' qui a l'avantage d'être compris entre 0 et 1

$$y'(\vec{x}) = \frac{L(\vec{x} \mid S)}{L(\vec{x} \mid S) + L(\vec{x} \mid B)}$$

Lemme de Neyman-Pearson: exemple

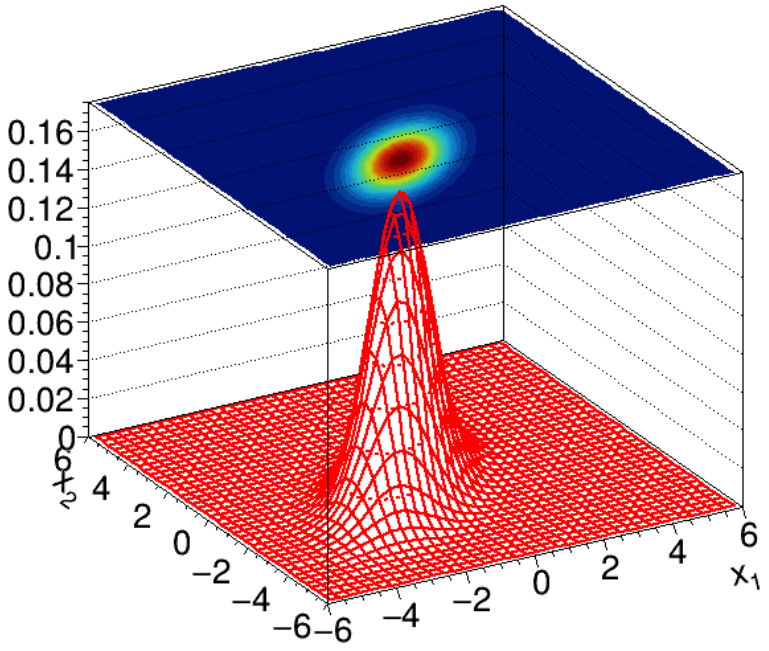
- 2 variables discriminantes: x_1 et x_2
- Les densités de probabilités pour les deux hypothèses sont:

Hypothèse S



$$L(x_1, x_2 \mid S)$$

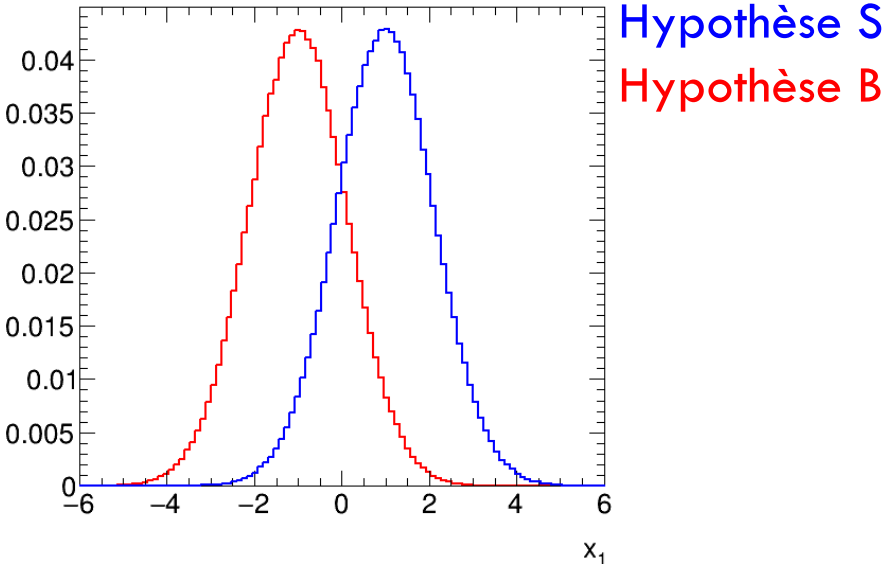
Hypothèse B

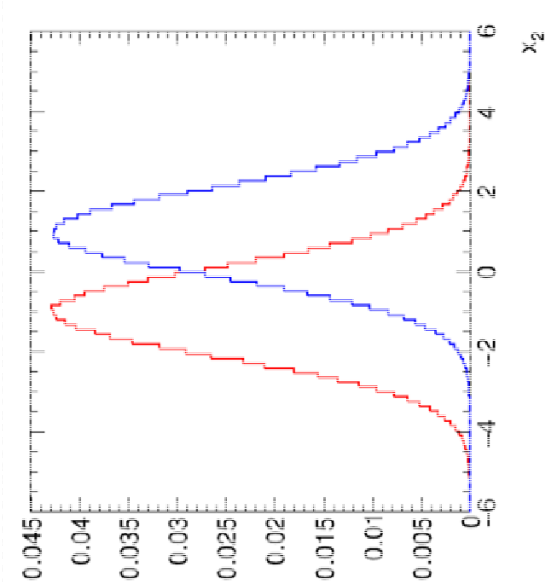


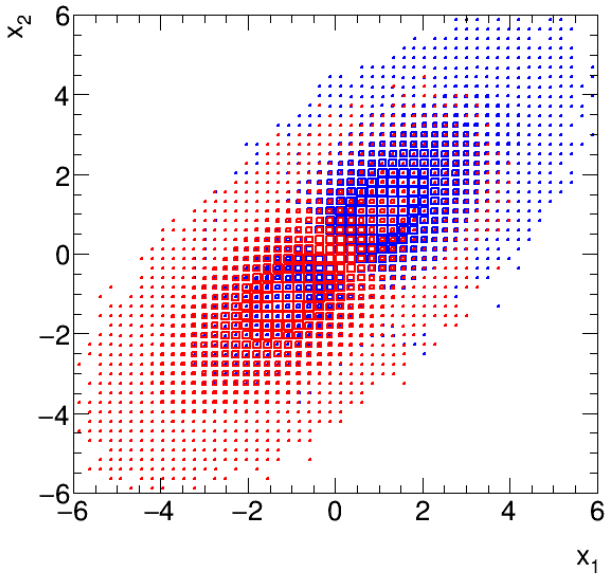
$$L(x_1, x_2 \mid B)$$

Lemme de Neyman-Pearson: exemple

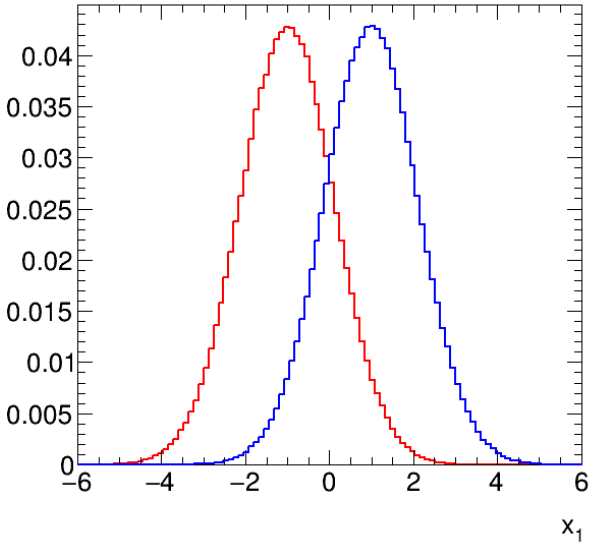
- Les variables x1 et x2 sont corrélées linéairement
- Comment sélectionner le signal et rejeter le bruit de fond?

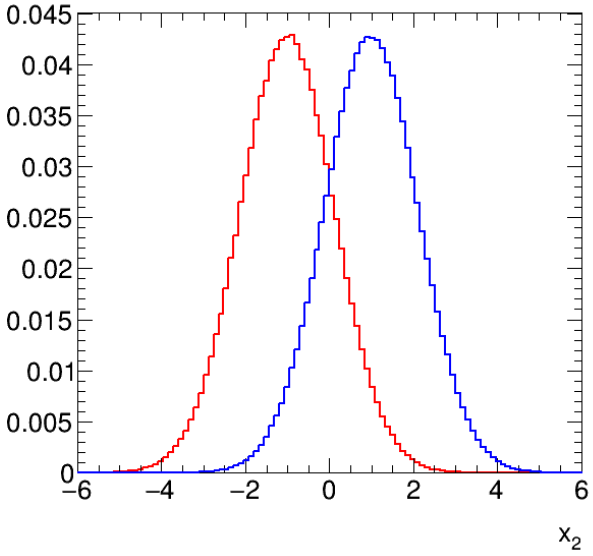




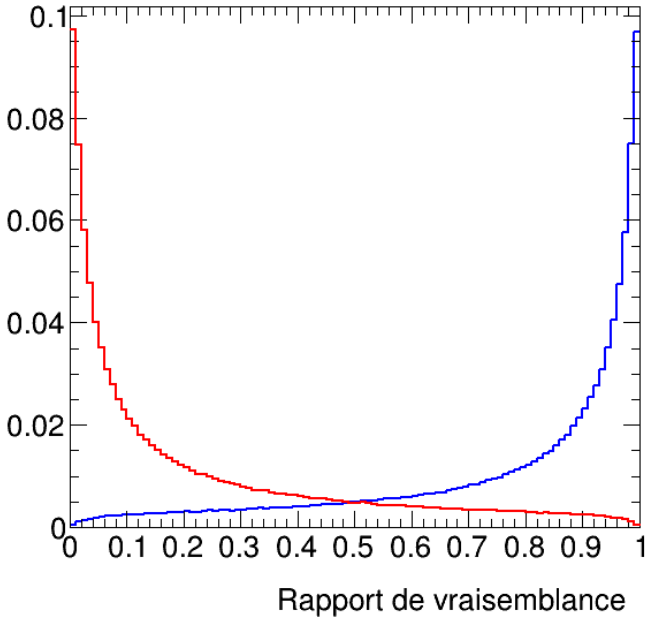


Lemme de Neyman-Pearson: exemple

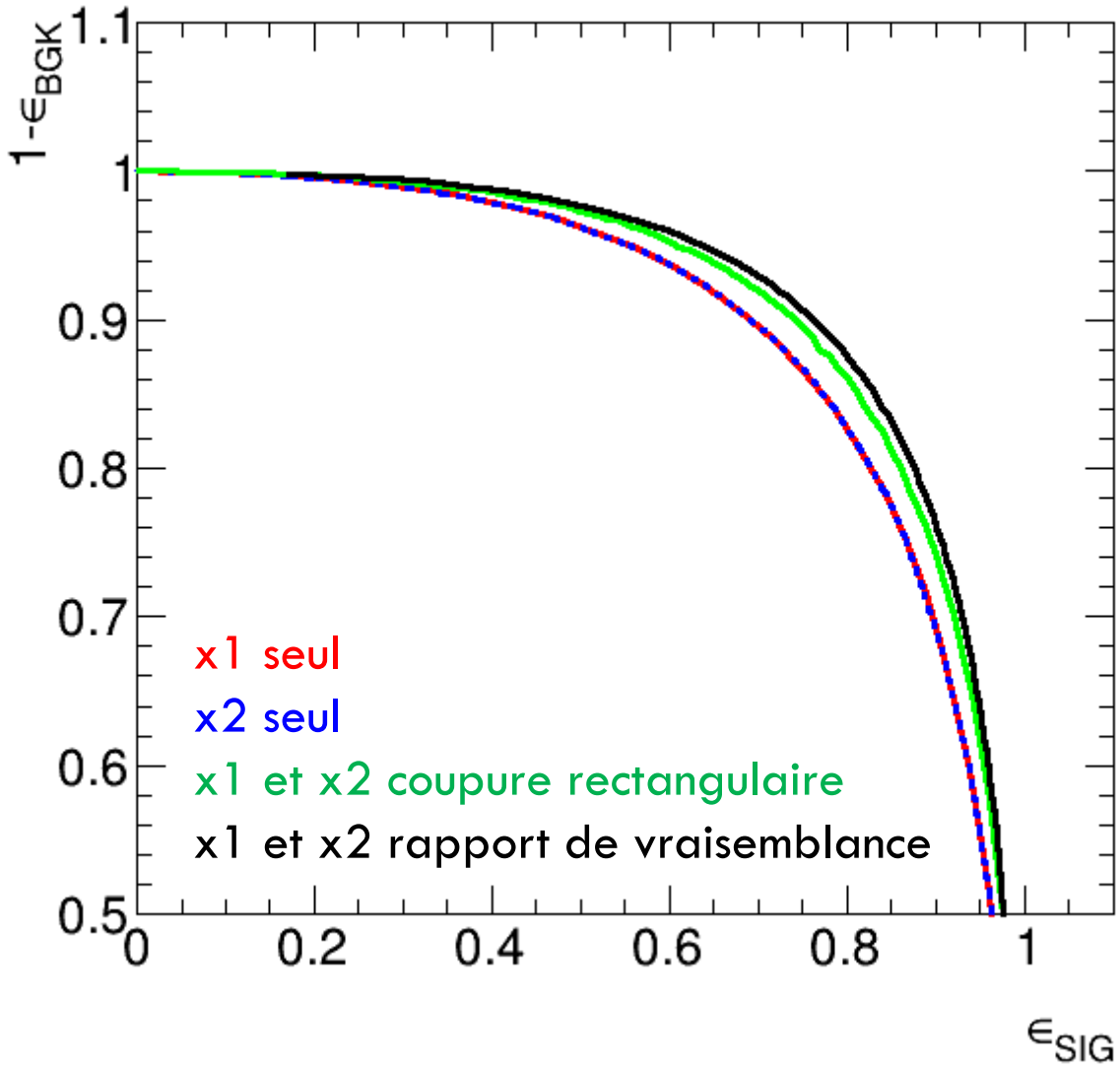




$$y(x_1, x_2) = \frac{L(x_1, x_2 \mid S)}{L(x_1, x_2 \mid S) + L(x_1, x_2 \mid B)}$$

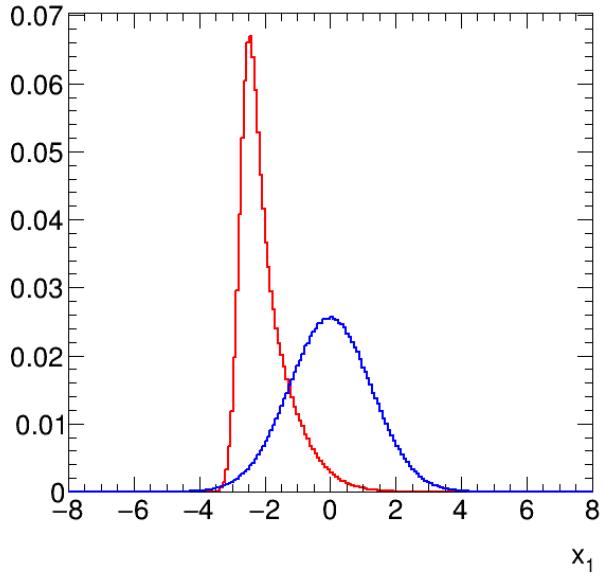


Lemme de Neyman-Pearson: exemple

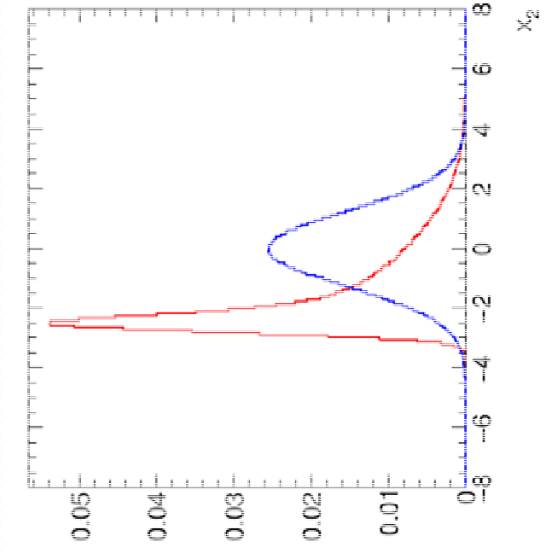


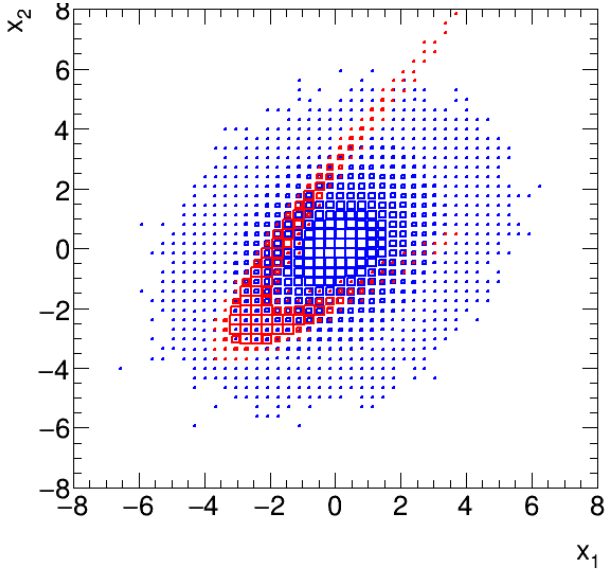
Lemme de Neyman-Pearson: autre exemple

- Corrélation non-linéaire pour le bruit de fond
- Comment sélectionner le signal et rejeter le bruit de fond?

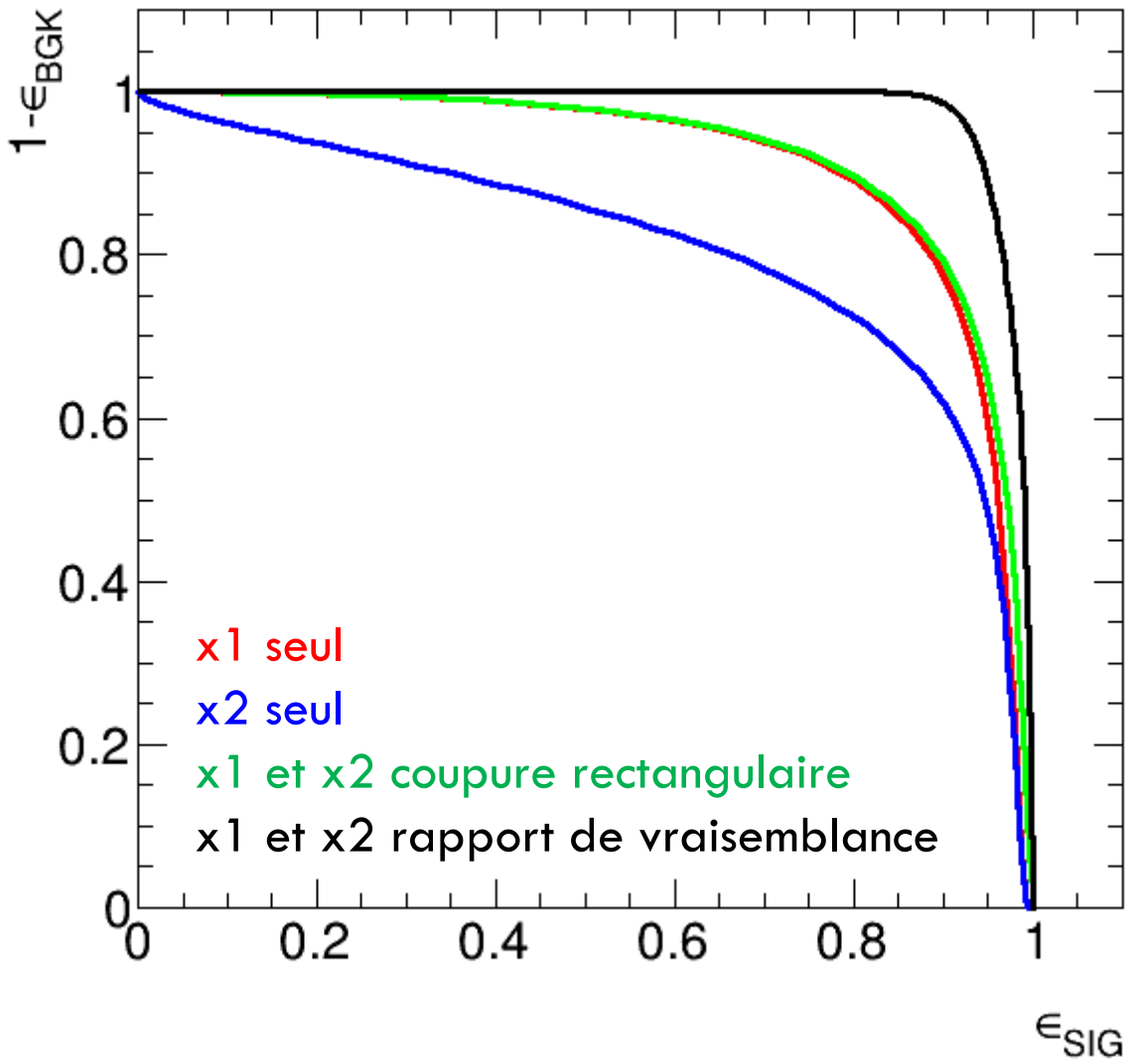


Hypothèse S
Hypothèse B



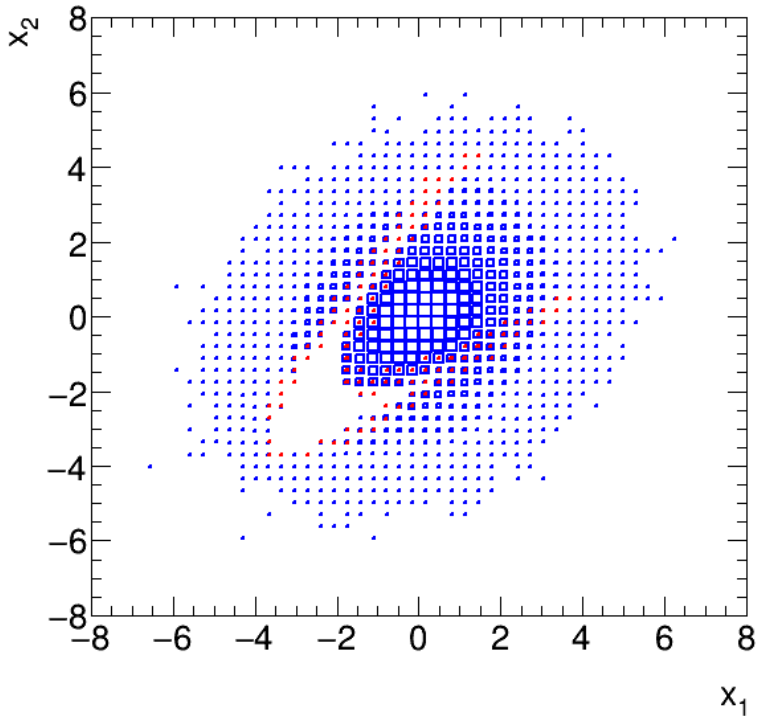


Lemme de Neyman-Pearson: autre exemple

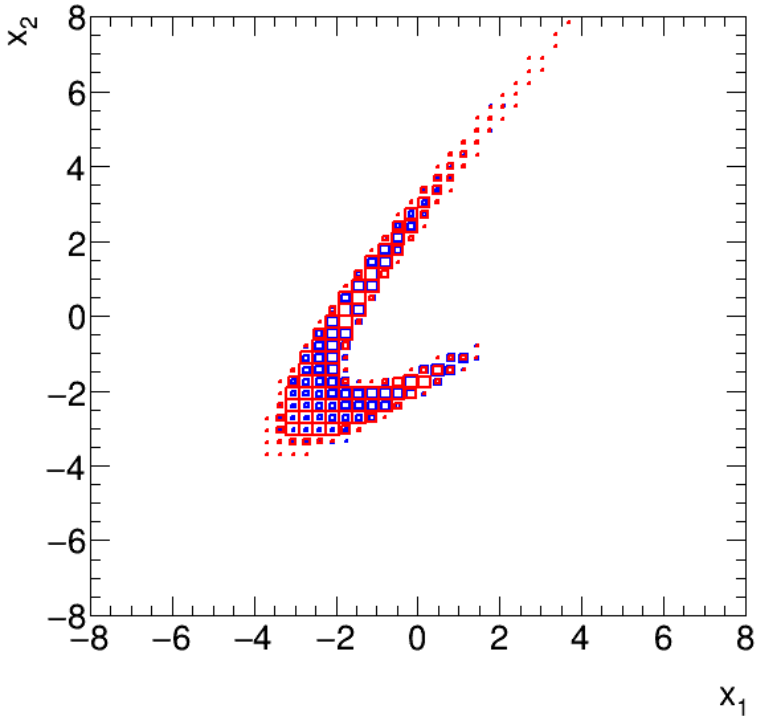


Lemme de Neyman-Pearson: autre exemple

$y(x_1, x_2) > 0.8$

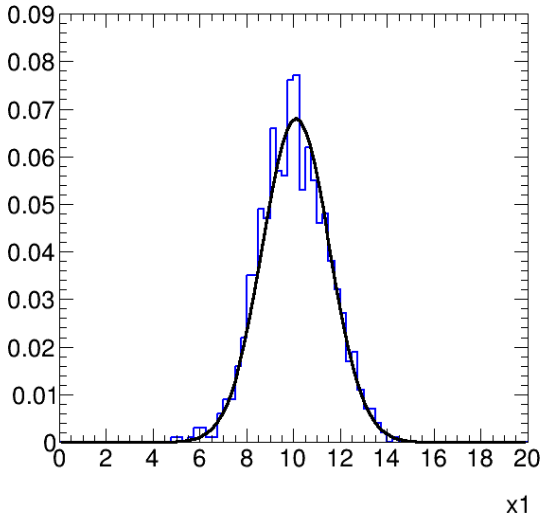


$y(x_1, x_2) < 0.8$

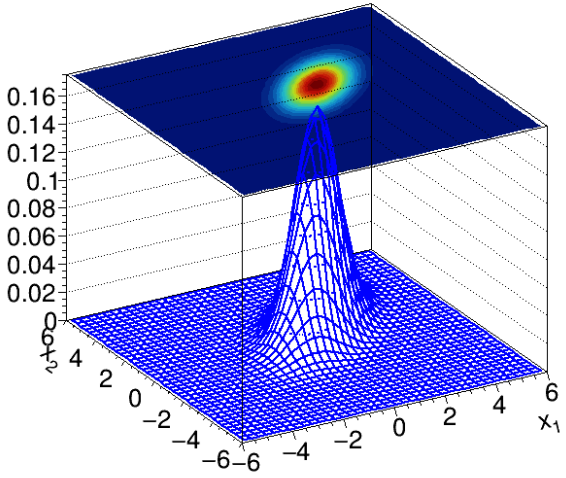


Lemme de Neyman-Pearson: limitation

- Dans la majorité des cas en physique des particules, on ne connait pas de formules explicites pour $L(x|H_0)$ et $L(x|H_1)$



- On pourrait utiliser la simulation Monte-Carlo pour avoir une approximation de ces fonctions. Malheureusement, pour un grand nombre de variables, cette méthode est prohibitif à cause du grand nombre d'événements nécéssaire
  - Pour N variables et pour M bins par variables, le nombre total de bins à remplir serait M^N.



- Si les variables sont faiblement corrélées, on peut utiliser:

$$y(\vec{x}) = \frac{\prod_{i=1}^{N} pdf(x_i \mid S)}{\prod_{i=1}^{N} pdf(x_i \mid B)}$$

Apprentissage automatique

Statistique

- Trouver une fonction discriminante y(x) qui permet de séparer le signal du bruit du fond de la façon la plus puissante

- De nombreuses méthodes existent: discriminant de Fisher, machines à vecteur de support, réseaux (profond) de neurones, méthode des k plus proches voisin, arbres de décisions,…
  - C'est un domaine en plein essor utilisé pour la reconnaissance d'image, la robotique, la traduction automatique, les jeux (AlphaGo),..

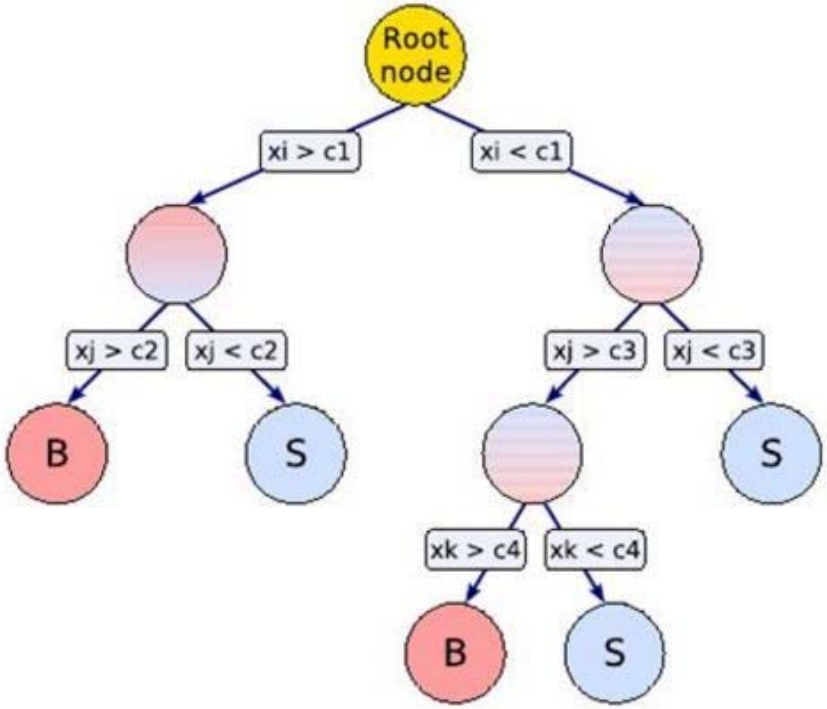
- Ces méthodes sont souvent vu comme des « boites noires » de part leurs complexités mais sont très performantes

- En physique des particules, on utilise des techniques d'apprentissage supervisés
  - En utilisant la simulation où la catégorisation est connue

Arbres de décision

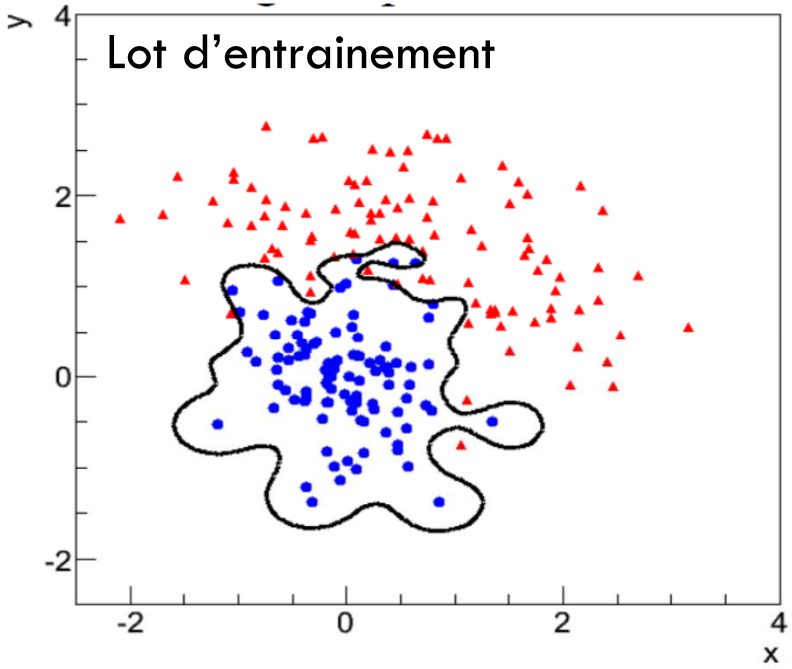
- Application séquentielle de coupure séparant les données en nœud jusqu'aux feuilles finales qui classifient un événement comme bruit de fond ou signal

- Procédure:
  - On part de la racine
  - On sépare les données d'entrainement en 2 lots en utilisant la variable et la coupure la plus performante
  - Critère de séparation: pureté*(1-pureté)
  - On réitere la procedure jusqu'à satisfaire un critère(s) de fin (Nb d'événement dans le nœud, Profondeur/pureté max,…)
  - Classifie une feuille en comptabilisant le nombre d'événement pour chaque catégorie (vote majoritaire)
  - Une fois l'entrainement terminé, on peut utiliser l'arbre de décision pour d'autres lots de données (test ou vrai données)



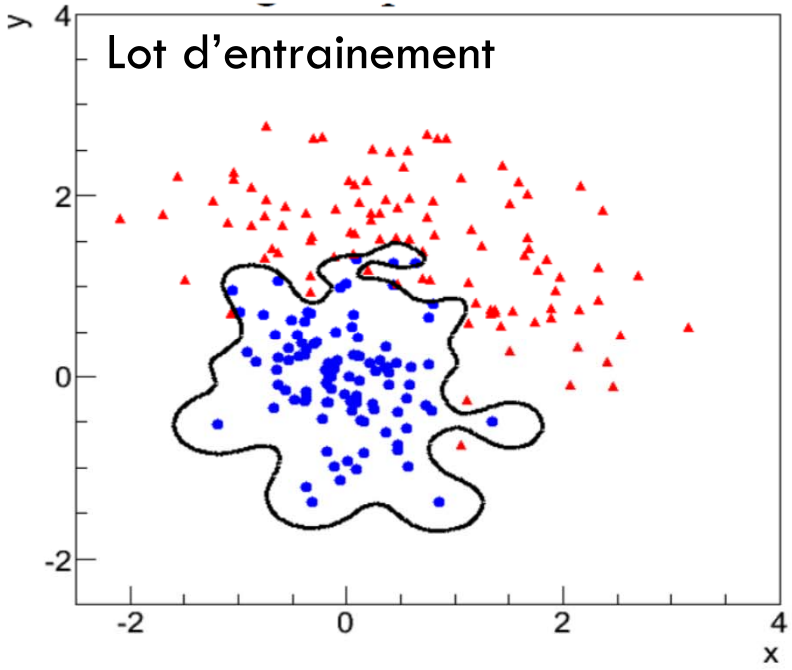
Arbres de décision

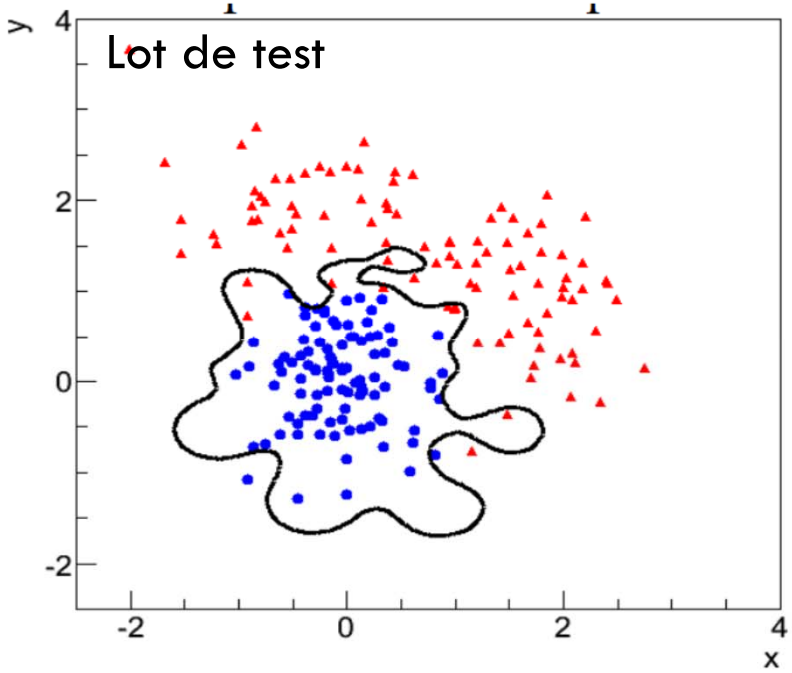
- Le surentrainement est un problème des méthodes complexes et puissantes qui peuvent être sensibles aux fluctuations statistiques dans les lots de simulation utilisés pour l'entrainement



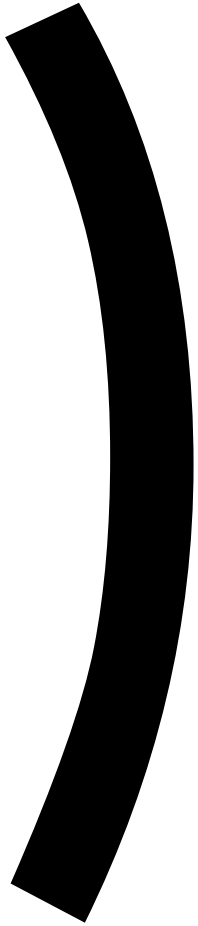
Arbres de décision

- Le surentrainement est un problème des méthodes complexes et puissantes qui peuvent être sensibles aux fluctuations statistiques dans les lots de simulation utilisés pour l'entrainement

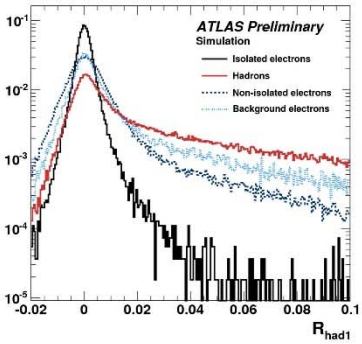


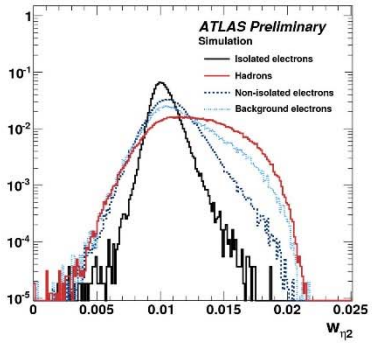


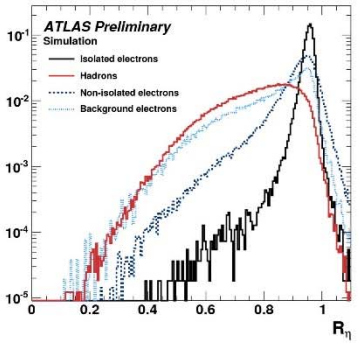
- Il existe des méthodes pour minimiser le problème du surentrainement (ex: boosting)

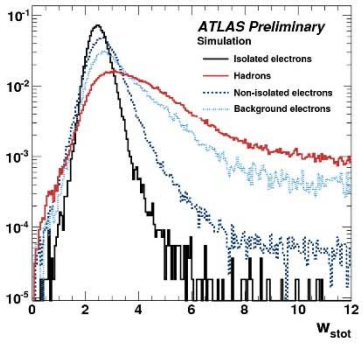


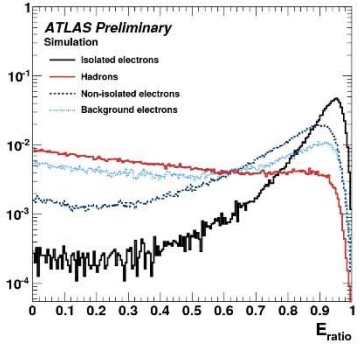
Encore plus de variables discriminantes

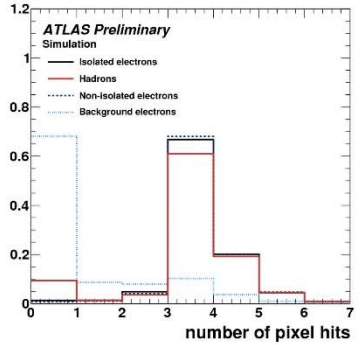


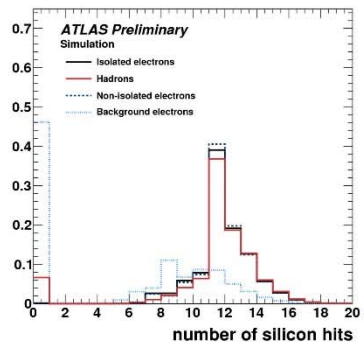


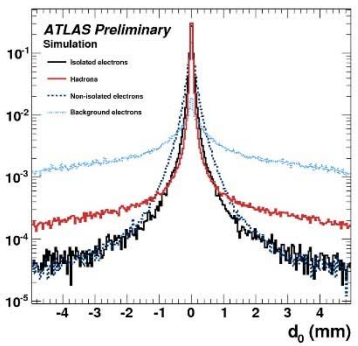


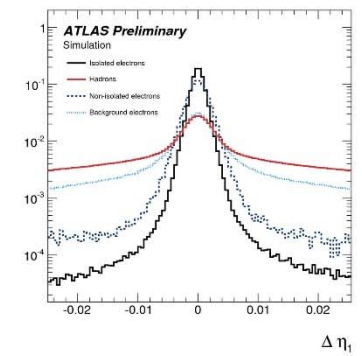


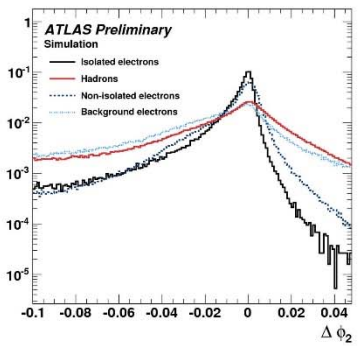


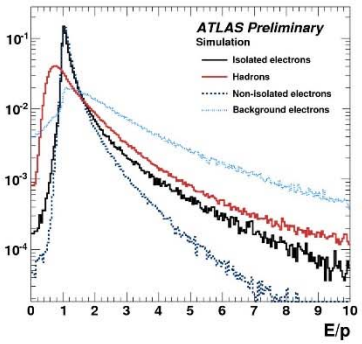


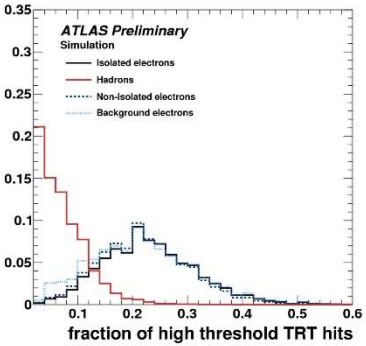


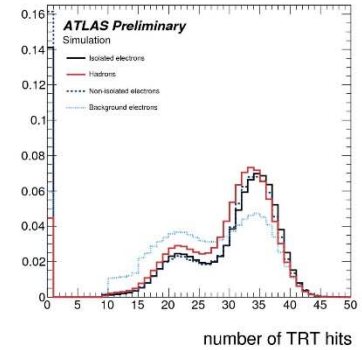


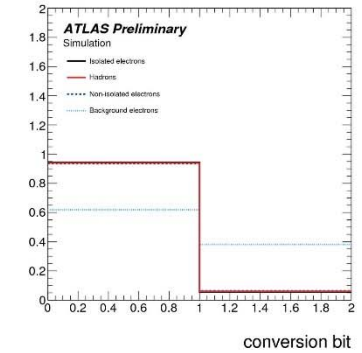




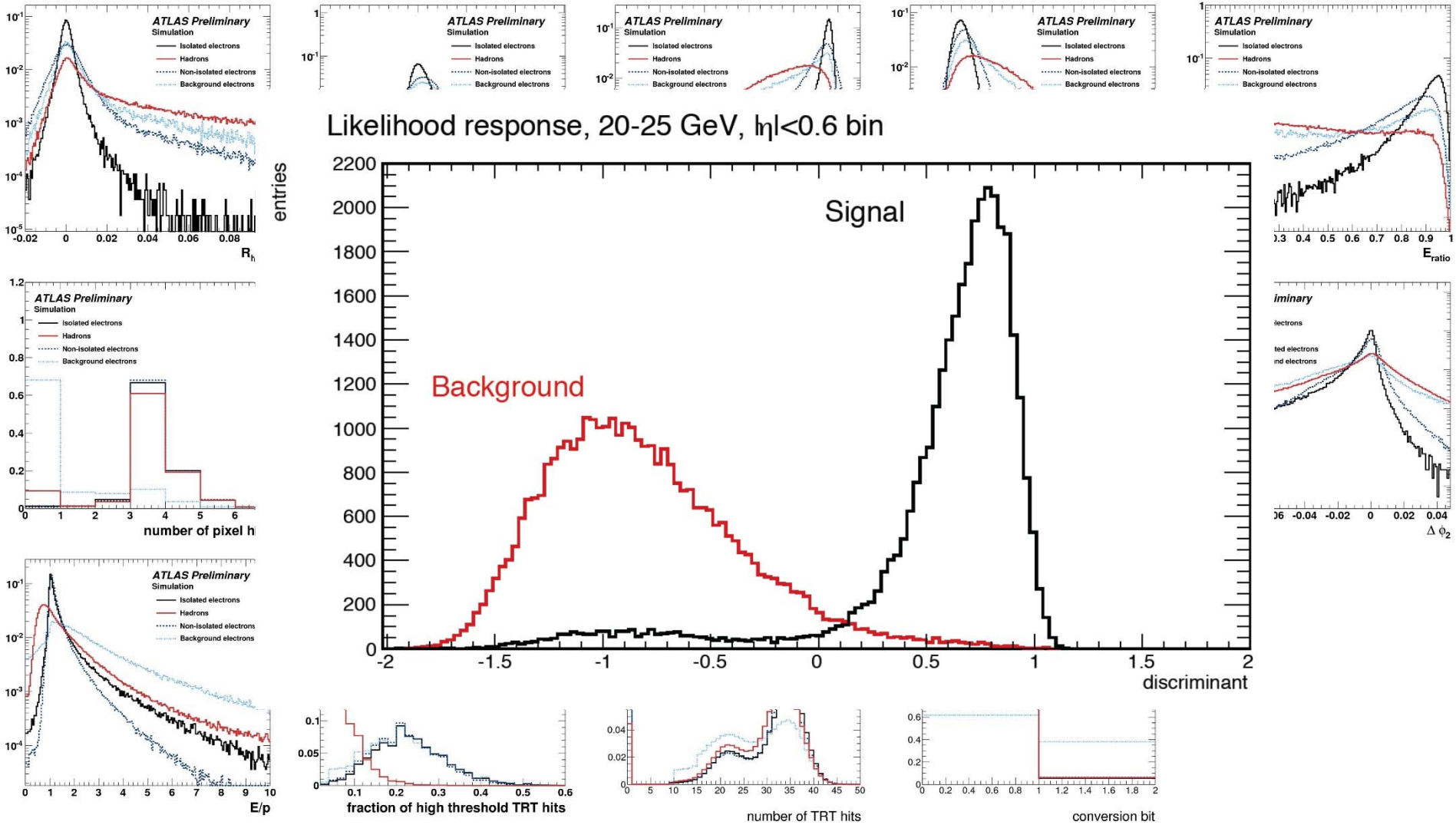




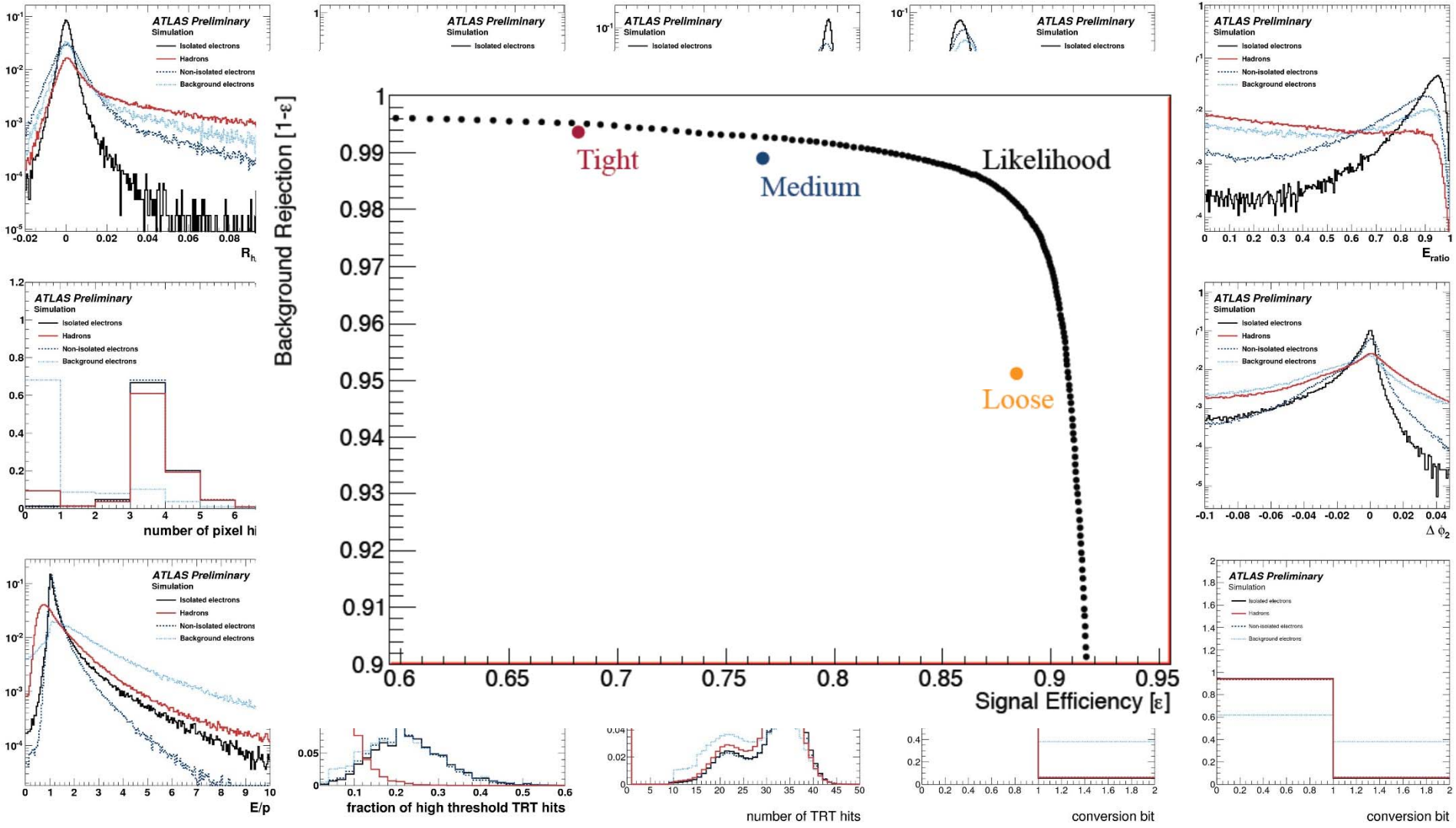




Encore plus de variables discriminantes

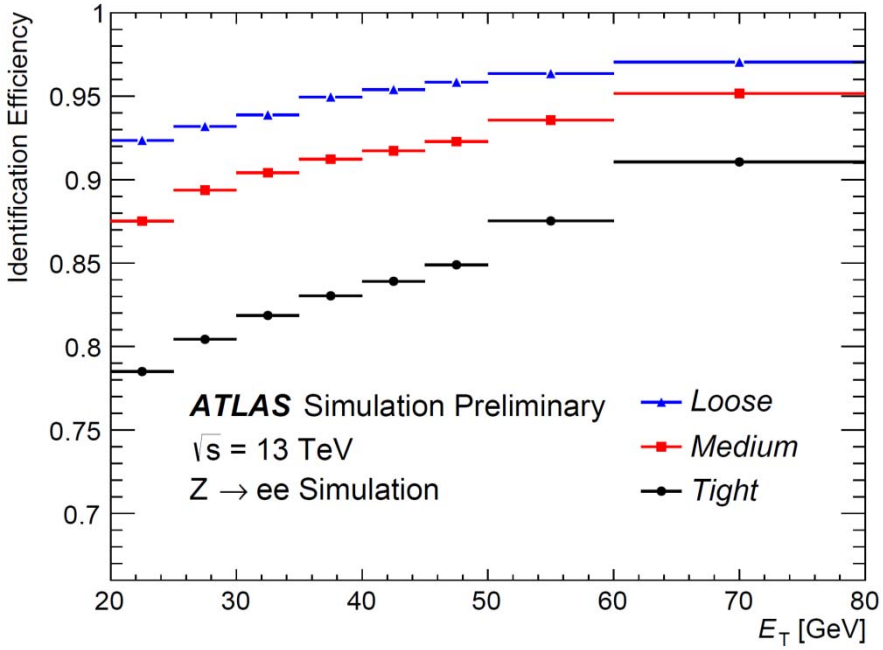


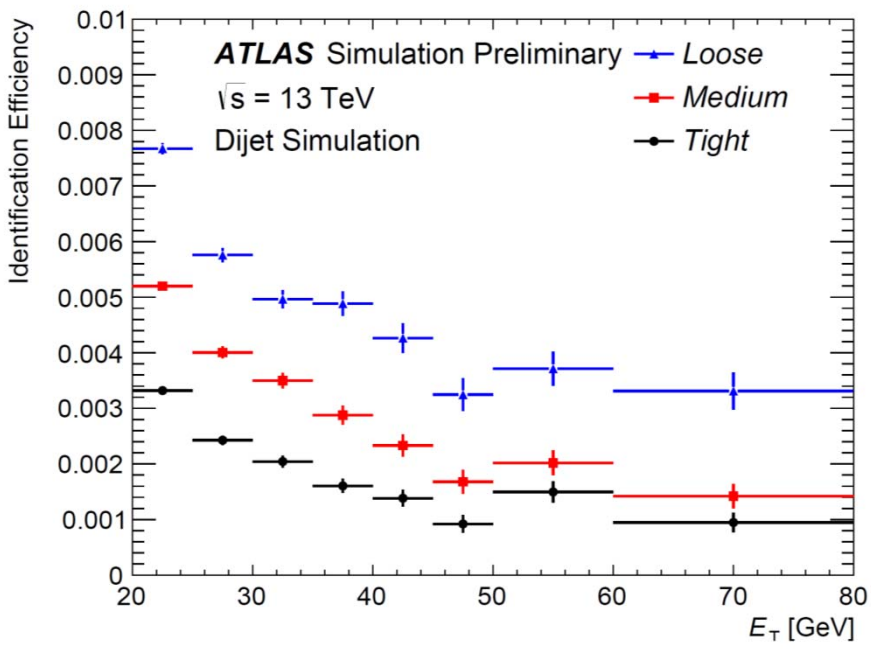
Rejection vs efficacité



Efficacité d'idenfication et rejection...

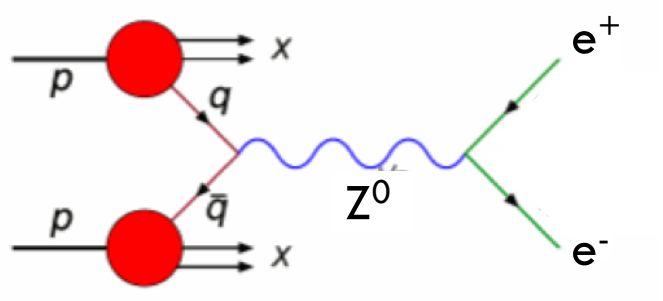
- Les performances des algorithmes de reconstruction et de selection des électrons peuvent mesurées dans la simulation en utilisant les information vraies.
- Performance des trois points de fonctionnement pour la selection des electrons:
  - Tight: Meilleur rejection mais moins bonne efficacité
  - Loose: Meilleur efficacité mais moins bonne rejection



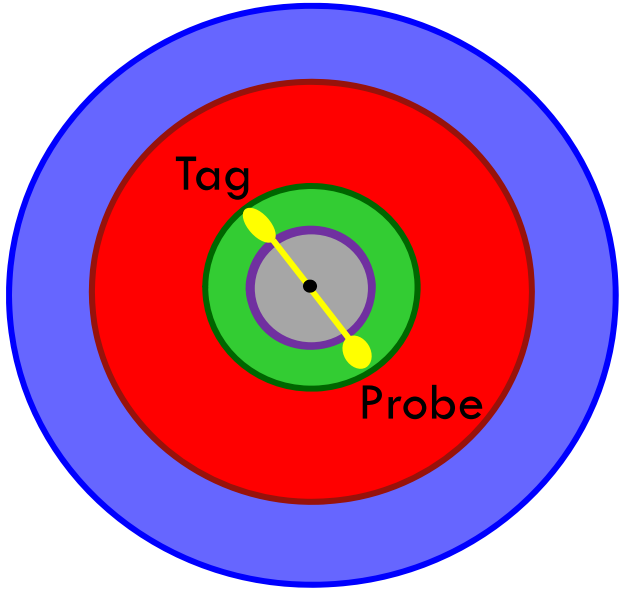


...et maintenant dans les données!

- La simulation peut différer de la réalité et il faut donc être capable de mesurer les performances directement dans les données sans l'utilisation de l'information vraie!

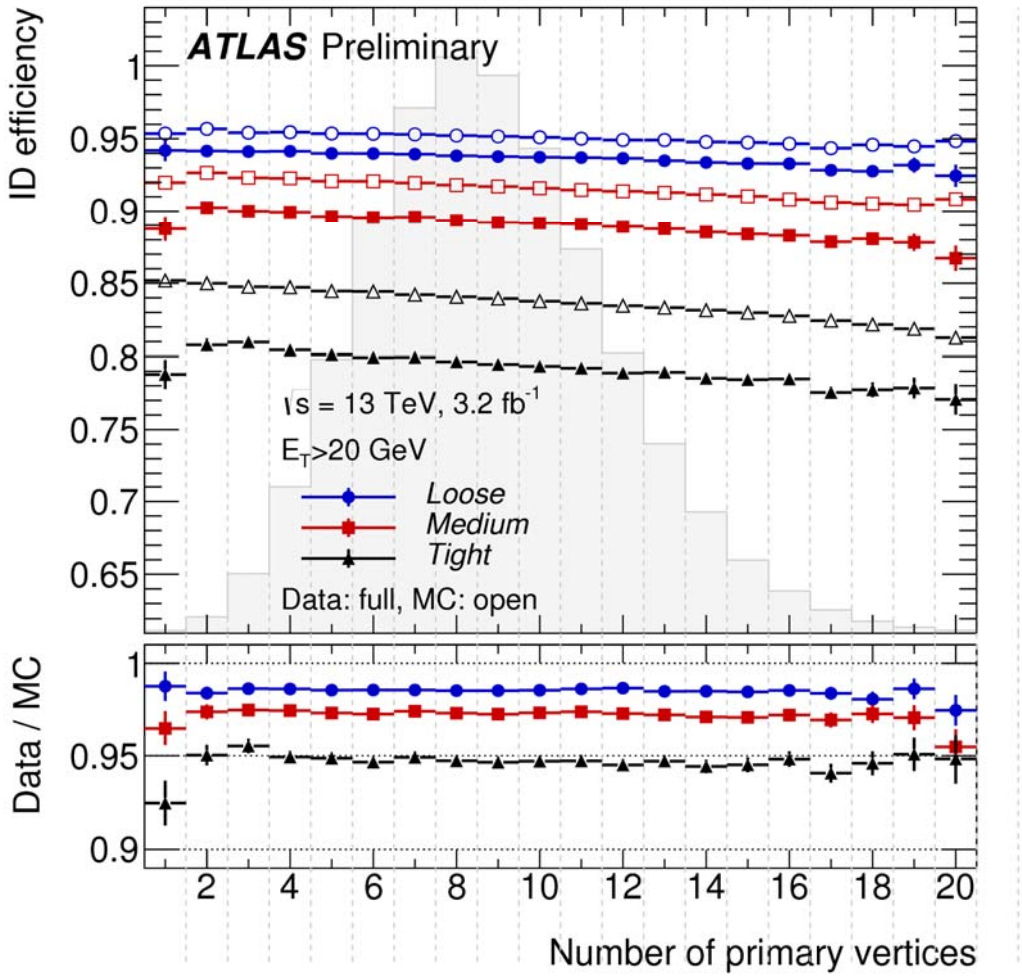


- Méthode «tag and probe» avec des bosons Z:
  - Sélection topologique
  - Tag: selection Tight, isolation,...
  - Probe: Lot pur(*) non biaisé d'électron



(*) si le lot n'est pas pur, il existe des méthodes pour soustraire les contributions des bruits de fond

...et maintenant dans les données!



- Différence entre données et simulation de l'ordre de qqs pourcents
- → Comprendre la/les source(s) de la différence et modifier la simulation
- → Alternative « rapide»: Correction ad hoc de la simulation

Résumé

- L'objectif de la reconstruction est d'obtenir les meilleurs performances possibles:
  - Efficacité de reconstruction et sélection
  - Taux de mauvaise identification
  - Linéarité
  - Résolution
  - Stabilité en fonction de l'empilement

- Ces performances doivent être mesurées et validées à partir des données avec des erreurs les plus petites possibles

- En cas de désaccord entre les données et la simulation
  1. Comprendre l'origine de ces différences et modifier la simulation (géométrie, modélisation des processus d'interaction,…)
  2. Si ce n'est pas possible, appliquer une correction ad-hoc pour reproduire les données