Machine-Learned Exclusion Limits without Binning

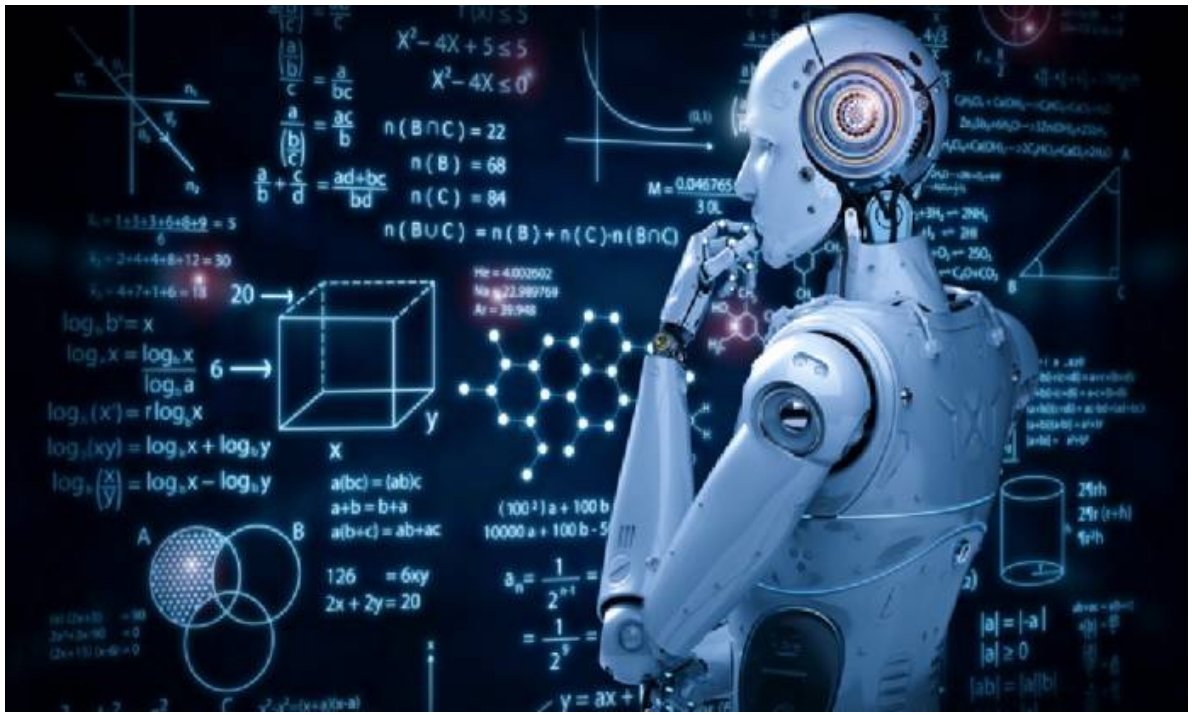
Rosa María Sandá Seoane

Instituto de Física Teórica UAM-CSIC

SATURNALIA 2022
December 21, 2022

What is Machine Learning?

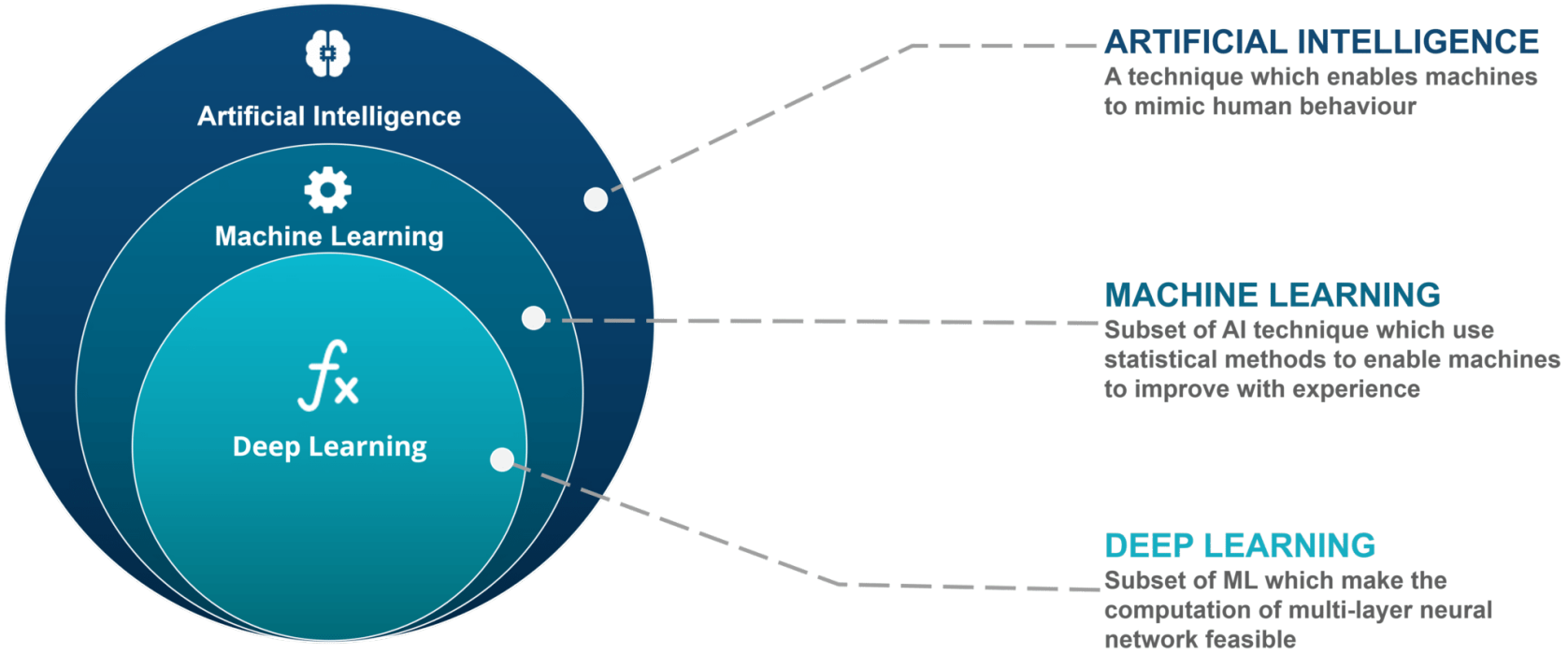
Machine Learning (ML) is the study of computer algorithms capable of building a mathematical model out of a data sample, by learning from examples.



The algorithm builds a predictive model without being explicitly programmed to do so

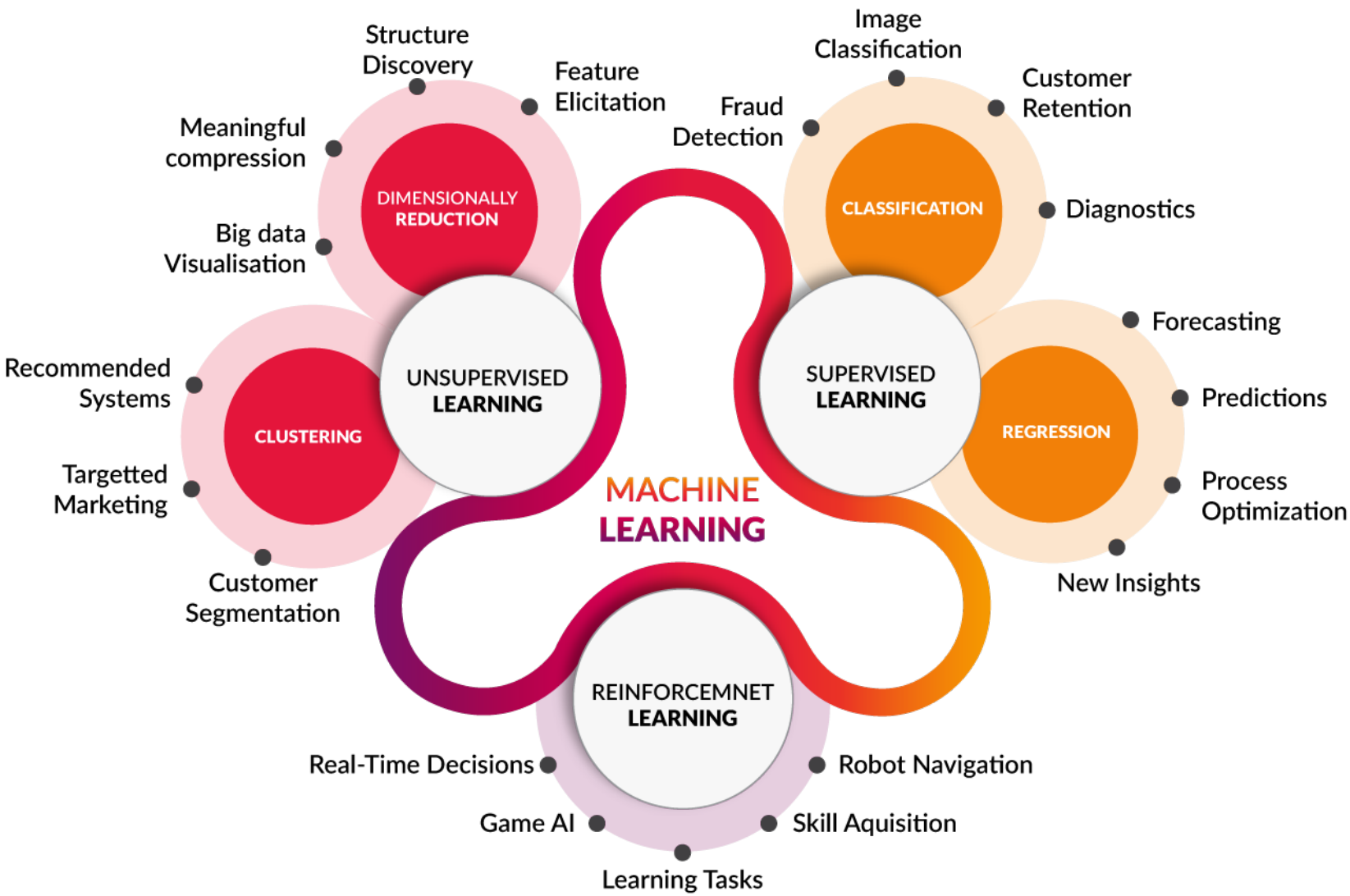
Artificial Intelligence

Intelligence $\longrightarrow$ Ability to process current information to inform future decisions



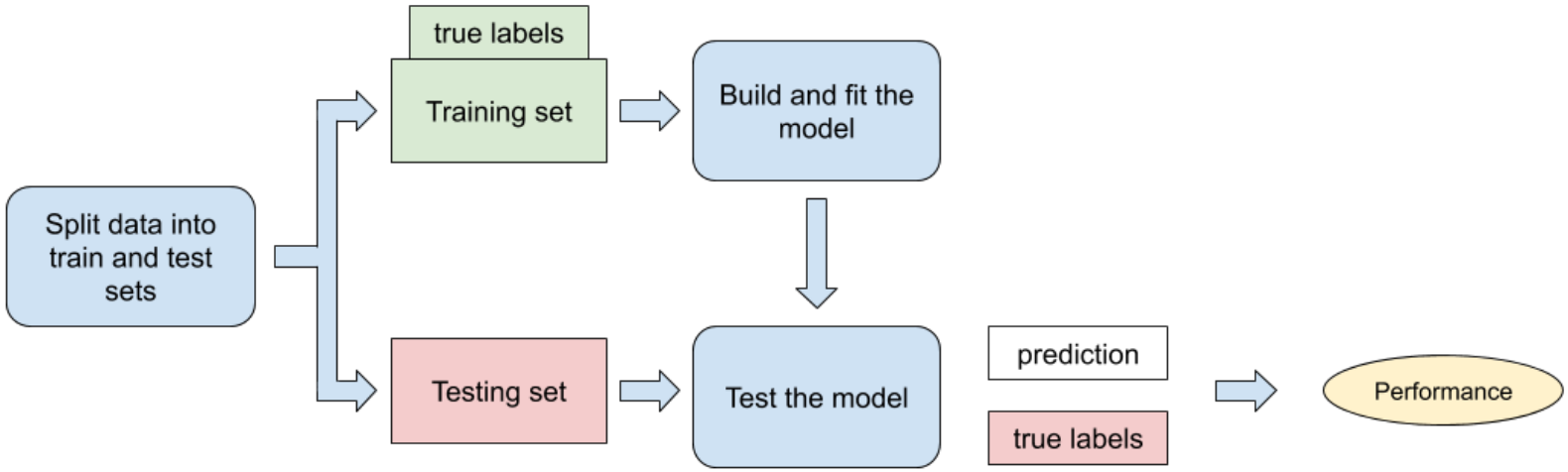
None of the systems we have nowadays are real AI! The brain learns so efficiently that no ML method can match it.

Learning Paradigms



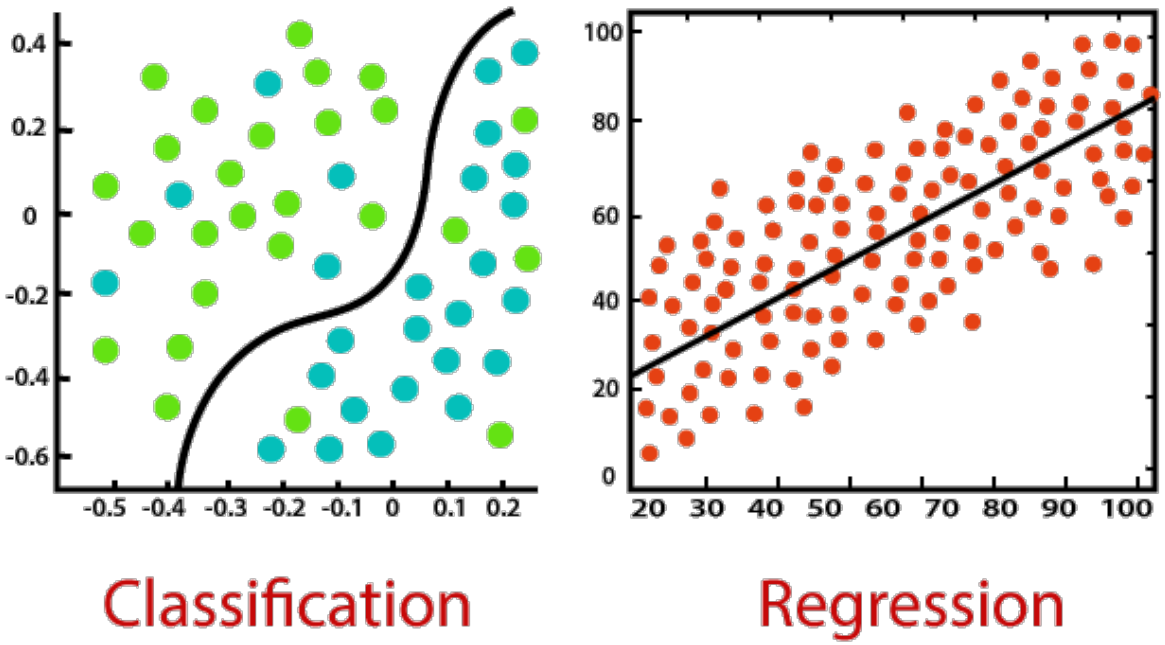
Supervised Learning

Given some labeled data $D = \{(\vec{x_1}, \vec{t_1}), ..., (\vec{x_n}, \vec{t_n})\}$ with features $\{\vec{x_i}\}$ and targets $\{\vec{t_i}\}$, the algorithm finds a mapping $\vec{t_i} = F(\vec{x_i})$



Supervised Learning

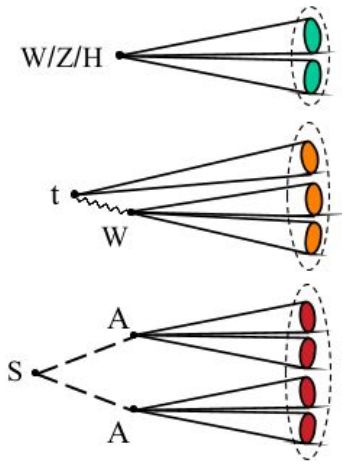
Classification: $\{\vec{t_1}, ..., \vec{t_n}\}$ (finite set of labels)
Regression: $\vec{t_i} \in \mathbb{R}^n$



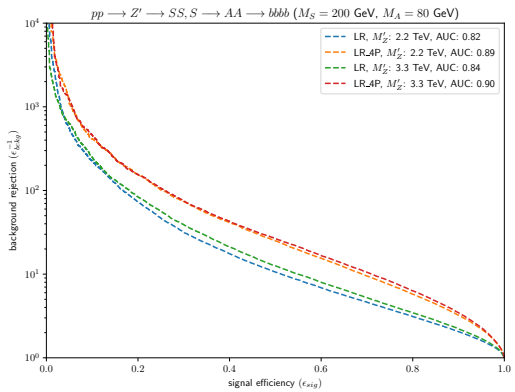
(Linear Regression is the oldest ML algorithm!)

Supervised Learning in HEP: examples

Classification: Jet Tagging (infering number of quarks and gluons inside it, or *prongs*)



Processes	Prongness	Classification
QCD	One-pronged (1P)	**Background**
$W/Z/H \to q\bar{q}$	Two-pronged (2P)	
$t \to W^+b \to q\bar{q}b$	Three-pronged (3P)	**Signal**
$S \to AA \to q\bar{q}q\bar{q}$	Four-pronged (4P)	



$\{\vec{x_i}\}$: jet mass, jet's tranverse momentum, 17 N-subjetiness variables.
$\{\vec{t_i}\} = \{0, 1\}$

(J.A. Aguilar-Saavedra, E. Arganda, F.R. Joaquim, J. Seabra, RMSS)

Method: Machine-Learned Likelihood

E. Arganda, X. Marcano, V. Martín Lozano, A. D. Medina, A. D. Perez, M. Szewc, A. Szynkman

Eur. Phys. J. C **82**, no.11, 993 (2022)

A method for approximating optimal statistical significances with machine-learned likelihoods

Ernesto Arganda,[a,b*], Xabier Marcano,[a,c†] Víctor Martín Lozano,[d,e‡] Anibal D. Medina,[b§] Andres D. Perez,[b¶] Manuel Szewc[f,g‖] and Alejandro Szynkman[b**]

[a]Instituto de Física Teórica UAM-CSIC,
C/ Nicolás Cabrera 13-15, Campus de Cantoblanco, 28049, Madrid, Spain
[b]IFLP, CONICET - Dpto. de Física, Universidad Nacional de La Plata,
C.C. 67, 1900 La Plata, Argentina
[c]Departamento de Física Teórica, Universidad Autónoma de Madrid ,
E-28049 Cantoblanco, Madrid, Spain
[d]Departament de Física Teòrica and IFIC, Universitat de València-CSIC,

E. Arganda, M. de los Rios, A. D. Perez, RMSS

PoS **ICHEP2022** (2022) 1226

PROCEEDINGS OF SCIENCE

Imposing exclusion limits on new physics with machine-learned likelihoods

Ernesto Arganda,[a,b] Martín de los Rios,[a,c] Andres D. Perez[b] and Rosa María Sandá Seoane[a,*]

[a]Instituto de Física Teórica UAM-CSIC,
C/ Nicolás Cabrera 13-15, Campus de Cantoblanco, 28049, Madrid, Spain
[b]IFLP, CONICET - Dpto. de Física, Universidad Nacional de La Plata,

E. Arganda, M. de los Rios, A. D. Perez, RMSS

arXiv: 2211.04806

Machine-Learned Exclusion Limits without Binning

Ernesto Arganda [a,b*] Andres D. Perez,[b†] Martin de los Rios [a,c‡] and Rosa María Sandá Seoane [a§]

[a]Instituto de Física Teórica UAM-CSIC,
C/ Nicolás Cabrera 13-15, Campus de Cantoblanco, 28049, Madrid, Spain
[b]IFLP, CONICET - Dpto. de Física, Universidad Nacional de La Plata,
C.C. 67, 1900 La Plata, Argentina
[c]Departamento de Física Teórica, Universidad Autónoma de Madrid,
E-28049 Cantoblanco, Madrid, Spain

Traditional vs ML search of New Physics

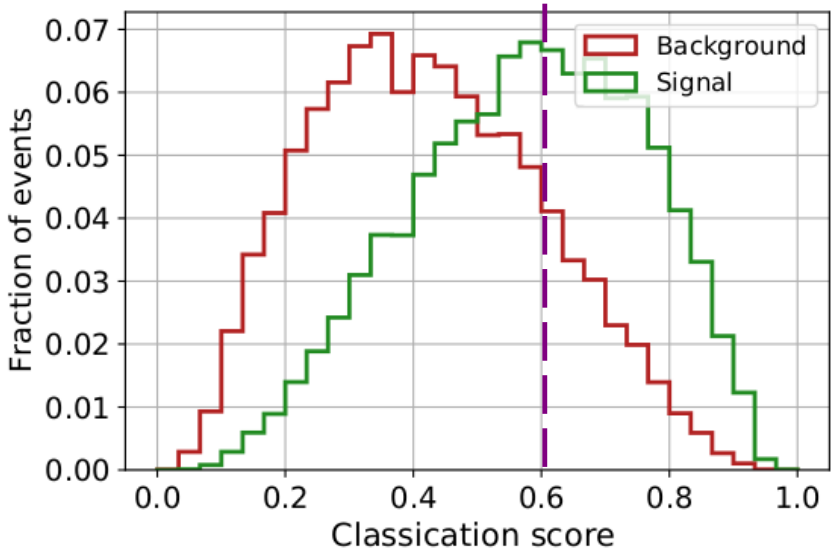
Distinguish SM (bckg) vs BSM (signal) in collider data:

- Design observables, define control regions... $\longrightarrow$ ML classifiers ✓
- For experimental significances, selection cuts $\longrightarrow$ Working points ✗

Traditional vs ML search of New Physics

Distinguish SM (bckg) vs BSM (signal) in collider data:

- Design observables, define control regions... $\longrightarrow$ ML classifiers ✓
- For experimental significances, selection cuts $\longrightarrow$ Working points ✗



Traditional vs ML search of New Physics

Distinguish SM (bckg) vs BSM (signal) in collider data:

- Design observables, define control regions... $\longrightarrow$ ML classifiers ✓
- For experimental significances, selection cuts $\longrightarrow$ Working points ✗

Is it possible to connect the ML classifier output with the standard statistical tests without defining working points?

$\longrightarrow$ Machine-Learned Likelihood (MLL) Method

Traditional vs ML search of New Physics

Distinguish SM (bckg) vs BSM (signal) in collider data:

- Design observables, define control regions... $\longrightarrow$ ML classifiers ✓
- For experimental significances, selection cuts $\longrightarrow$ Working points ✗

Is it possible to connect the ML classifier output with the standard statistical tests without defining working points?

$\longrightarrow$ Machine-Learned Likelihood (MLL) Method

Can we avoid binning the output?

$\longrightarrow$ +Kernel Density Estimators (KDE)

The MLL method

Statistical model for N independent measurements, with a high-dimensional set of observables x

$$\mathcal{L}(\mu, s, b) = p(N, \{x_i, i = 1, ..., N\}|\mu, s, b) \equiv \text{Poiss}\big(N|\mu S + B\big) \prod_{i=1}^{N} p(x_i|\mu, s, b)$$

where S (B) is the expected total signal (background) yield, and

$$p(x|\mu, s, b) = \frac{B}{\mu S + B}\, p_b(x) + \frac{\mu S}{\mu S + B}\, p_s(x)$$

The relevant to derive exclusion limits on μ (considering models with $\mu \geqslant 0$)

$$\tilde{q}_\mu = \begin{cases} 0 & \text{if } \hat{\mu} > \mu\,, \\ -2\ \text{Ln}\ \frac{\mathcal{L}(\mu,s,b)}{\mathcal{L}(\hat{\mu},s,b)} & \text{if } 0 \leqslant \hat{\mu} \leqslant \mu\,, \\ -2\ \text{Ln}\ \frac{\mathcal{L}(\mu,s,b)}{\mathcal{L}(0,s,b)} & \text{if } \hat{\mu} < 0\,, \end{cases}$$

where $\hat{\mu}$ is the parameter that maximizes the likelihood

The MLL method

Statistical model for N independent measurements, with a high-dimensional set of observables x

$$\mathcal{L}(\mu, s, b) = p(N, \{x_i, i = 1, ..., N\}|\mu, s, b) \equiv \text{Poiss}\big(N|\mu S + B\big) \prod_{i=1}^{N} p(x_i|\mu, s, b)$$

where S (B) is the expected total signal (background) yield, and

$$p(x|\mu, s, b) = \frac{B}{\mu S + B}\, p_b(x) + \frac{\mu S}{\mu S + B}\, p_s(x)$$

The relevant to derive exclusion limits on μ (considering models with $\mu \geqslant 0$)

$$\tilde{q}_\mu = \begin{cases} 0 & \text{if } \hat{\mu} > \mu \\ 2(\mu - \hat{\mu})S - 2\sum_{i=1}^{N} \text{Ln}\left(\frac{Bp_b(x_i)+\mu S p_s(x_i)}{Bp_b(x_i)+\hat{\mu} S p_s(x_i)}\right) & \text{if } 0 \leqslant \hat{\mu} \leqslant \mu \\ 2\mu S - 2\sum_{i=1}^{N} \text{Ln}\left(1 + \frac{\mu S p_s(x_i)}{Bp_b(x_i)}\right) & \text{if } \hat{\mu} < 0; \end{cases}$$

where $\hat{\mu}$ is the parameter that maximizes the likelihood

$$\sum_{i=1}^{N} \frac{p_s(x_i)}{\hat{\mu} S\, p_s(x_i) + B\, p_b(x_i)} = 1$$

The MLL method

Statistical model for N independent measurements, with a high-dimensional set of observables x

$$\mathcal{L}(\mu, s, b) = p(N, \{x_i, i = 1, ..., N\}|\mu, s, b) \equiv \text{Poiss}\big(N|\mu S + B\big) \prod_{i=1}^{N} p(x_i|\mu, s, b)$$

where S (B) is the expected total signal (background) yield, and

$$p(x|\mu, s, b) = \frac{B}{\mu S + B}\, p_b(x) + \frac{\mu S}{\mu S + B}\, p_s(x)$$

The relevant to derive exclusion limits on μ (considering models with $\mu \geqslant 0$)

$$\tilde{q}_\mu = \begin{cases} 0 & \text{if } \hat{\mu} > \mu \\ 2(\mu - \hat{\mu})S - 2\sum_{i=1}^{N} \text{Ln}\left(\frac{B p_b(x_i) + \mu S p_s(x_i)}{B p_b(x_i) + \hat{\mu} S p_s(x_i)}\right) & \text{if } 0 \leqslant \hat{\mu} \leqslant \mu \\ 2\mu S - 2\sum_{i=1}^{N} \text{Ln}\left(1 + \frac{\mu S p_s(x_i)}{B p_b(x_i)}\right) & \text{if } \hat{\mu} < 0; \end{cases}$$

where $\hat{\mu}$ is the parameter that maximizes the likelihood

$$\sum_{i=1}^{N} \frac{p_s(x_i)}{\hat{\mu} S\, p_s(x_i) + B\, p_b(x_i)} = 1$$

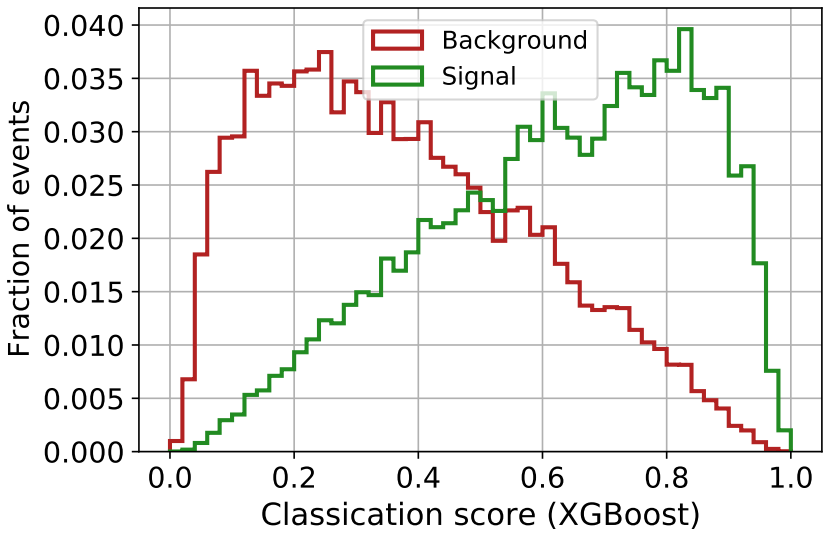
Solution: train classifier to distinguish signal from bckg with a balanced dataset. The classification score maximizing the binary cross-entropy and thus approaches

$$o(x) = \frac{p_s(x)}{p_s(x) + p_b(x)}$$

Dimensional reduction by dealing with $o(x)$ instead of x

$$p_s(x) \to \tilde{p}_s(o(x)), \qquad \text{and} \qquad p_b(x) \to \tilde{p}_b(o(x))$$

Cranmer et al, arXiv: 1506.02169



where $\tilde{p}_{s,b}(o(x))$ are the distributions of $o(x)$ for signal and background, obtained by evaluating the classifier on a set of pure signal or background events, respectively.

The relevant test statistic for exclusion limits

$$\tilde{q}_\mu = \begin{cases} 0 & \text{if } \hat{\mu} > \mu \\ 2(\mu-\hat{\mu})S - 2\sum_{i=1}^{N} \text{Ln}\left(\frac{B\tilde{p}_b(o(x_i))+\mu S\tilde{p}_s(o(x_i))}{B\tilde{p}_b(o(x_i))+\hat{\mu} S\tilde{p}_s(o(x_i))}\right) & \text{if } 0 \leqslant \hat{\mu} \leqslant \mu \\ 2\mu S - 2\sum_{i=1}^{N} \text{Ln}\left(1+\frac{\mu S\tilde{p}_s(o(x_i))}{B\tilde{p}_b(o(x_i))}\right) & \text{if } \hat{\mu} < 0; \end{cases}$$

with $\hat{\mu}$ such us

$$\sum_{i=1}^{N} \frac{\tilde{p}_s(o(x_i))}{\hat{\mu}S\,\tilde{p}_s(o(x_i)) + B\,\tilde{p}_b(o(x_i))} = 1$$

The median expected exclusion significance when the true hyphothesis is assumed to be the bckg-only one ($\mu' = 0$) is

$$\text{med}\,[Z_\mu|0] = \sqrt{\text{med}\,[\tilde{q}_\mu|0]}$$

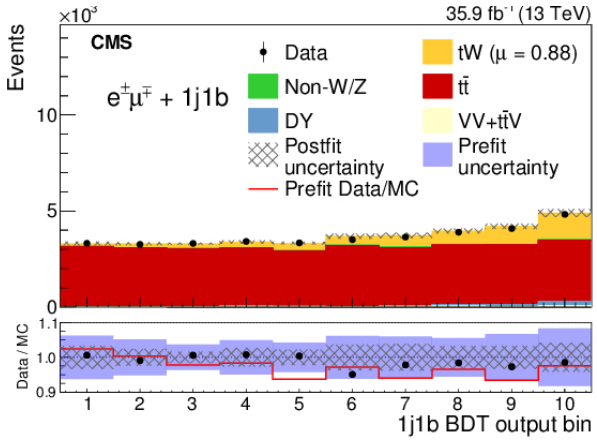
Traditional Binned-Likelihood (BL) method

$p_{s,b}(x)/\tilde{p}_{s,b}(o(x_i))$ are not known and are approximated by discrete binned distributions

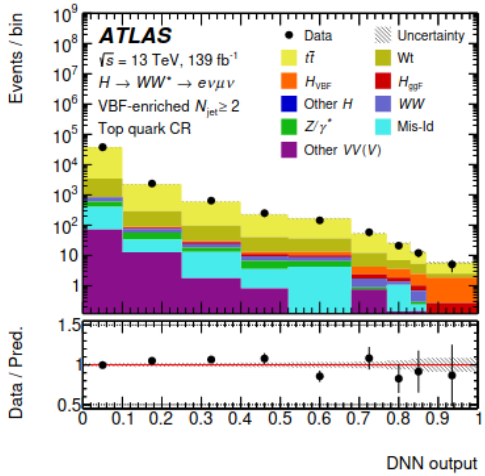
$$\mathcal{L}(\mu, s, b) = \prod_{d=1}^{D} \text{Poiss}\left(N_d | \mu S_d + B_d\right)$$

The median exclusion significance using Asimov datasets is given by

$$\text{med}[Z_\mu|0] = \sqrt{\tilde{q}_\mu|0} = \left[2\sum_{d=1}^{D}\left(B_d \ln\left(\frac{B_d}{S_d + B_d}\right) + S_d\right)\right]^{1/2} \xrightarrow[\sqrt{B}\gg 1]{S\ll B} \frac{S}{\sqrt{B}}$$



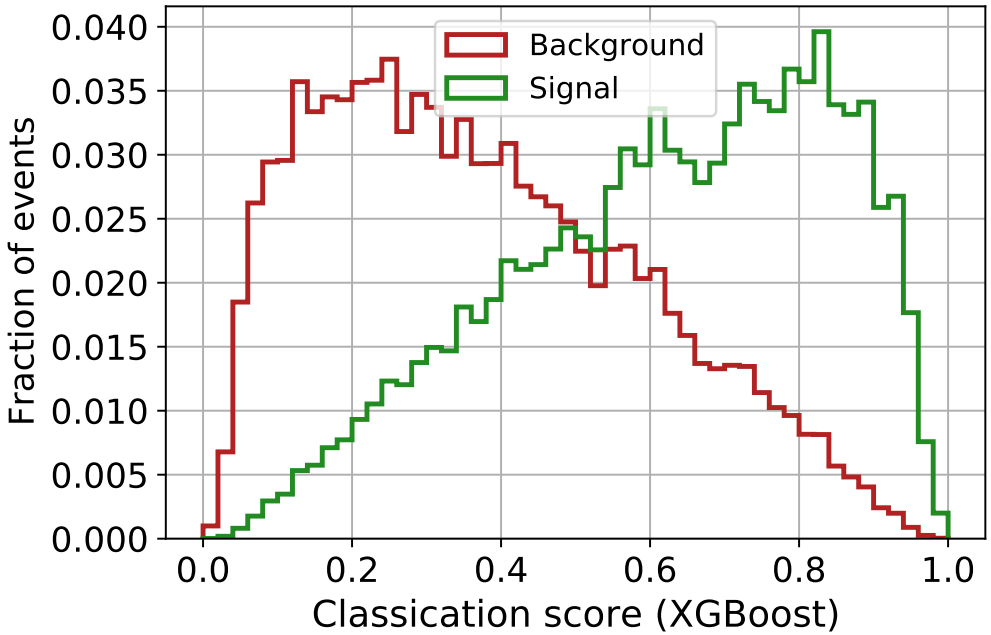
J HEP **10** (2018) 117



arXiv: 2207.00338

Density Estimation

What is the best way to extract $\tilde{p}_s(o(x))$ and $\tilde{p}_b(o(x))$?



Density Estimation

$\longrightarrow$Density estimation in a sense is the reverse of sampling: from given samples we want to retrieve the density function from which the samples were generated.

$\longrightarrow$Two types of methods for density estimation

- Parametric: model the density function as a specified functional form with a fixed number of tunable parameters.
- Non-parametric: specify a model whose complexity grows with the number of training datapoints.

Kernel Density Estimators

Kernel Density Estimators (KDE) is a non-parametric method for extracting $\tilde{p}_s(o(x_i))$ and $\tilde{p}_b(o(x_i))$

$\longrightarrow$Smoothed version of the empirical distribution $q_o(x)$ of the training data $\{x_i, i = 1, ..., N\}$

$$q_o(x) = \frac{1}{N}\sum_i^N \delta\left(x - x_i\right)$$

$\longrightarrow$We can smooth out the empirical distribution and turn it into a density by replacing each delta distribution with a smoothing kernel

$$\kappa_\epsilon(u) = \frac{1}{\epsilon^D}\,\kappa_1\left(\frac{u}{\epsilon}\right)$$

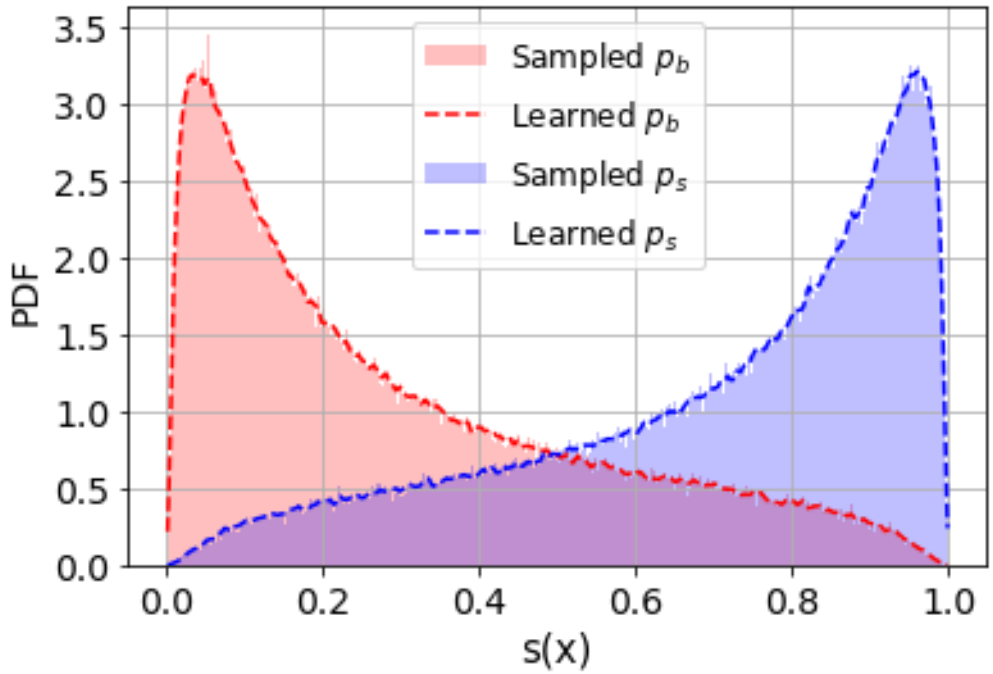
where $\epsilon > 0$ (bandwidth parameter) controls the width of the kernel and $\kappa_1(u)$ is a density function bounded from above (as $\epsilon \to 0$, $\kappa_\epsilon(u)$ approaches $\delta(u)$)

$$q_\epsilon(x) = \frac{1}{N}\sum_i^N \kappa_\epsilon\left(x - x_i\right)$$

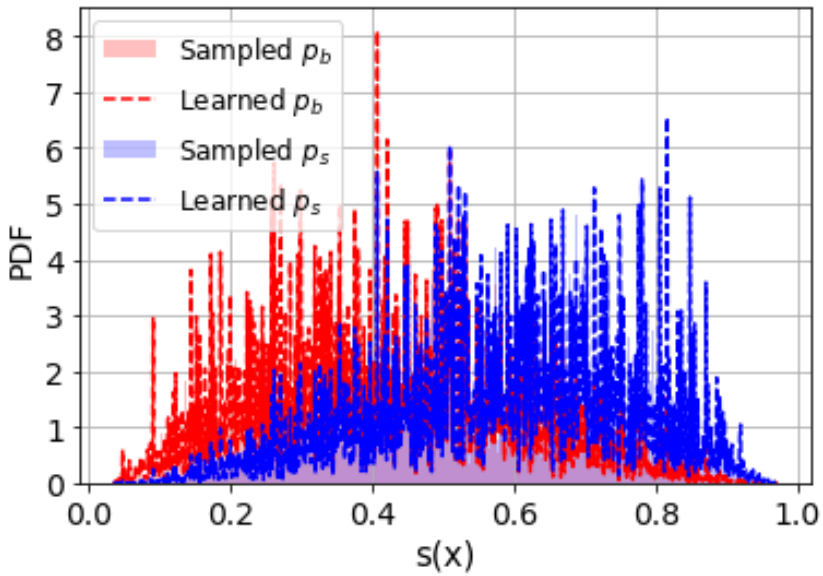
$$\tilde{p}_{s,b}(o(x)) = \frac{1}{N}\sum_{i}^{N} \kappa_\epsilon \left[o(x) - o(x_i)\right]$$

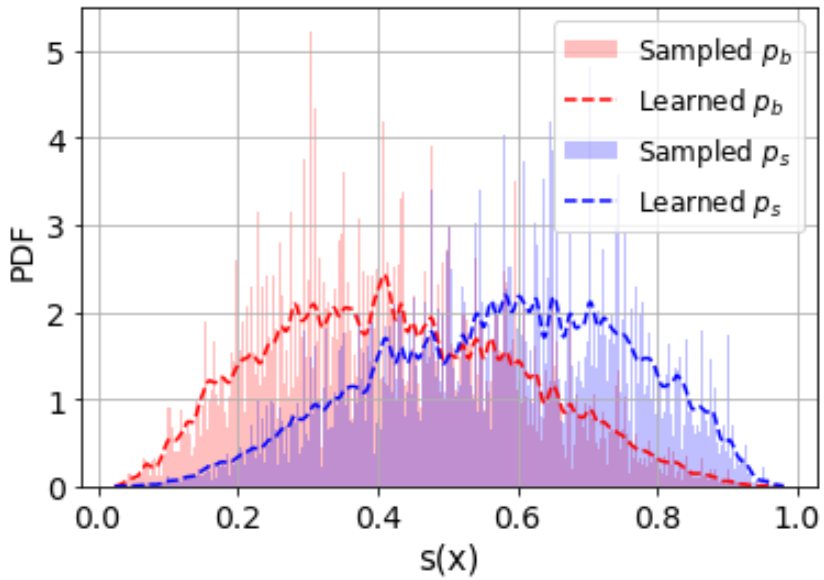
Several options for κ_ϵ, e.g.

$$\kappa_\epsilon(u) = \begin{cases} \frac{1}{\epsilon}\frac{3}{4}\left(1-(u/\epsilon)^2\right), & \text{if } |u| \leqslant \epsilon \\ 0 & \text{otherwise} \end{cases} \qquad \textit{Epanechnikov kernel}$$



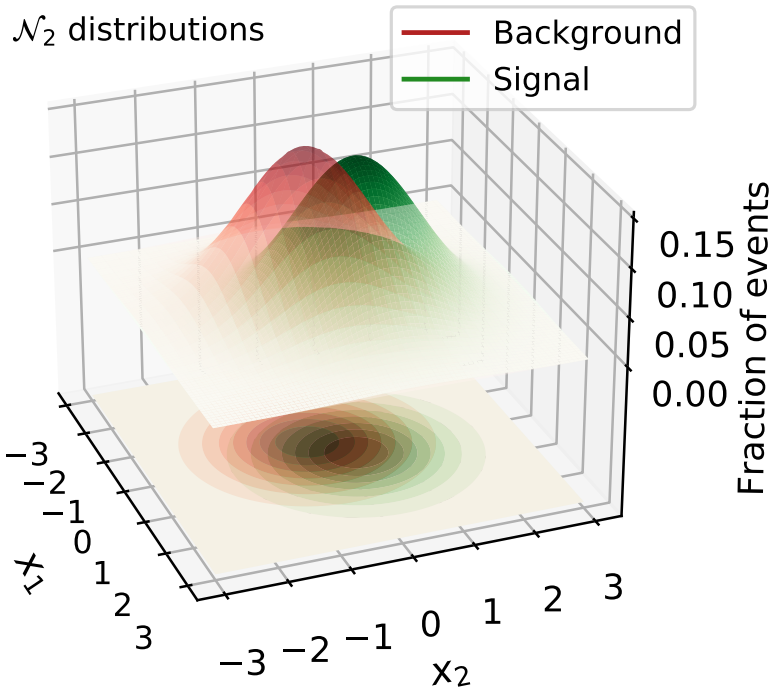
The width parameter ϵ controls the degree of smoothness, if ϵ is too low the model may overfit, whereas if ϵ is too high the model may underfit. In general, we want ϵ to be smaller the more data we have and larger the higher the dimension is.

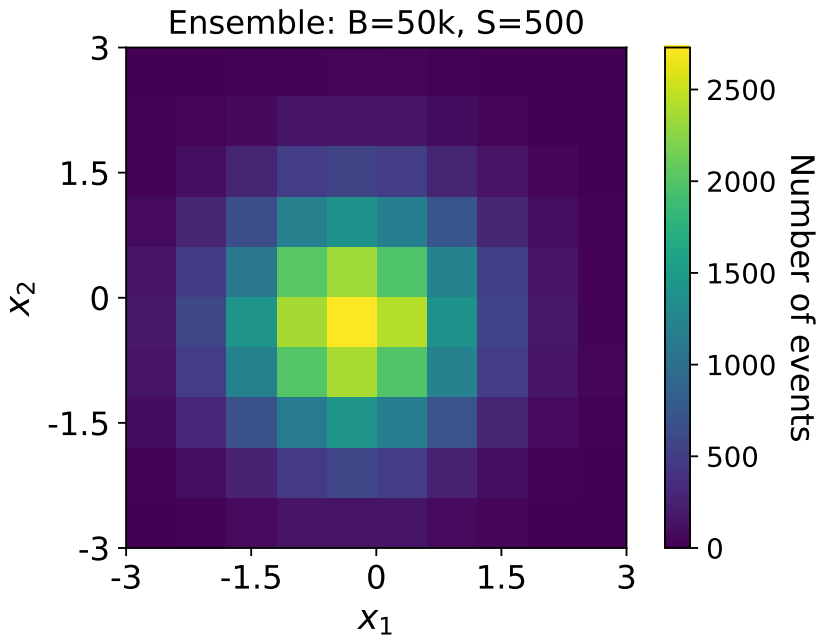




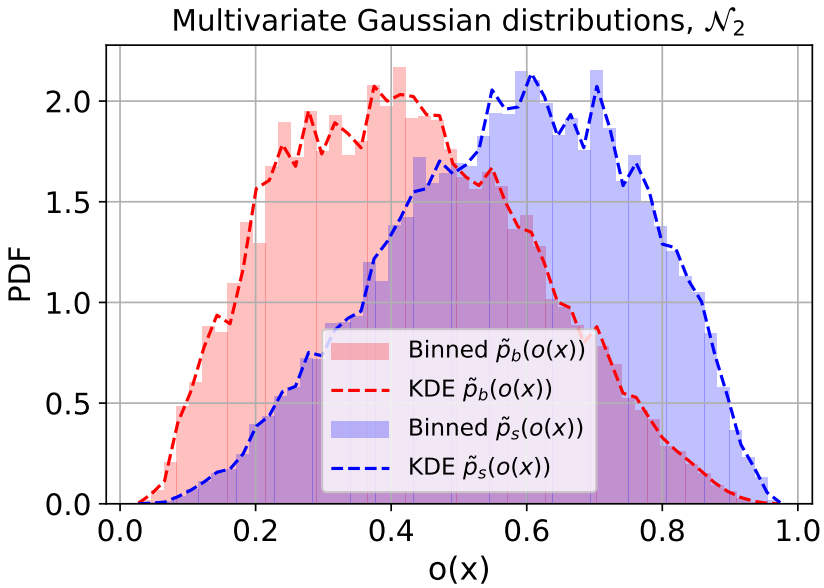
Toy Model: Multivariate Gaussian distributions

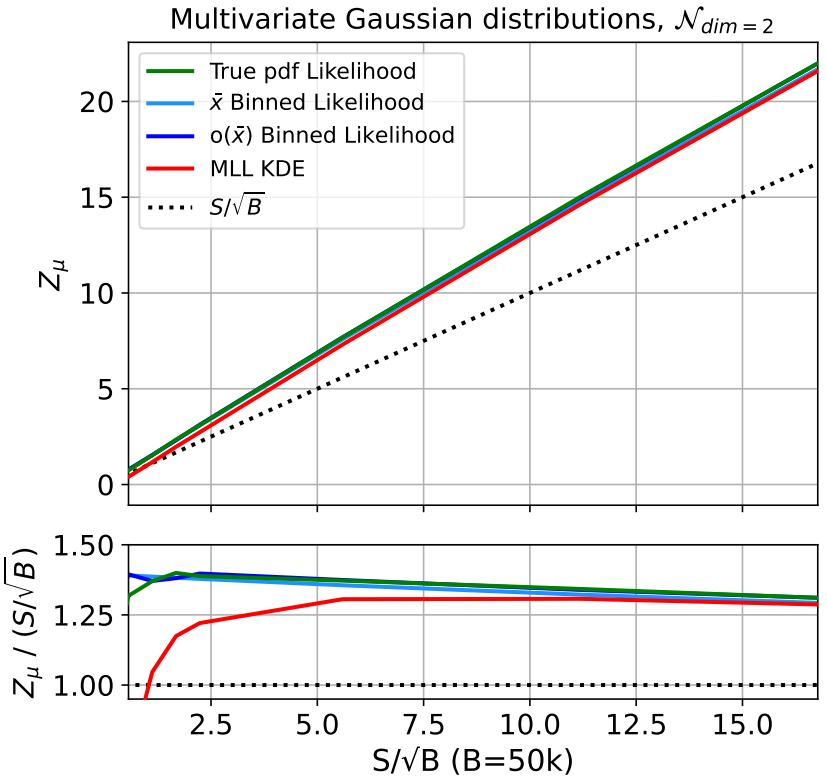
- Toy model in abstract space (x_1, x_2). Events generated by $\mathcal{N}_2(\boldsymbol{m}, \boldsymbol{\Sigma})$ (known generative functions $p_{s,b}(x)$).
- Covariance matrices $\boldsymbol{\Sigma} = \mathbb{I}_{2\times 2}$ (no correlation) and $\boldsymbol{m} = +0.3(-0.3)\ \mathbb{1}_2$ for S (B).





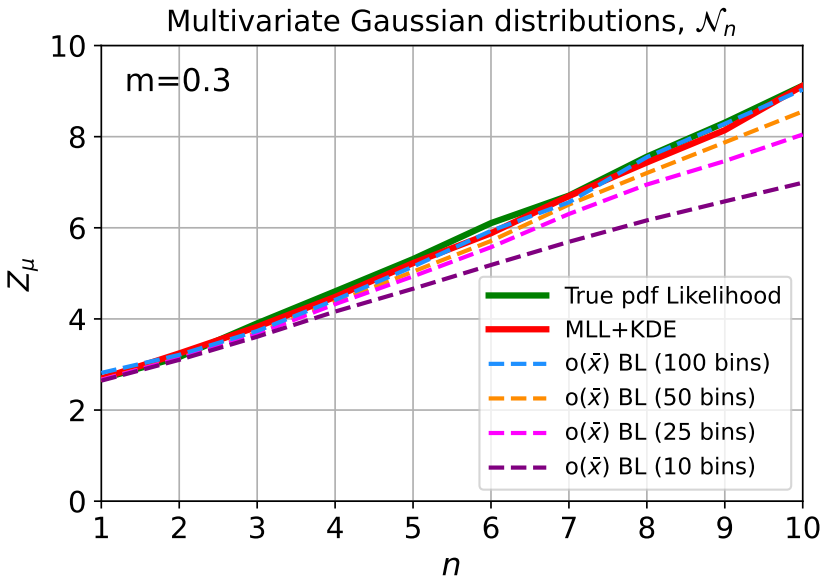
- Training of supervised per-event classifier, `XGBoost` with 1M events per class. KDE with Epanechnikov kernel.

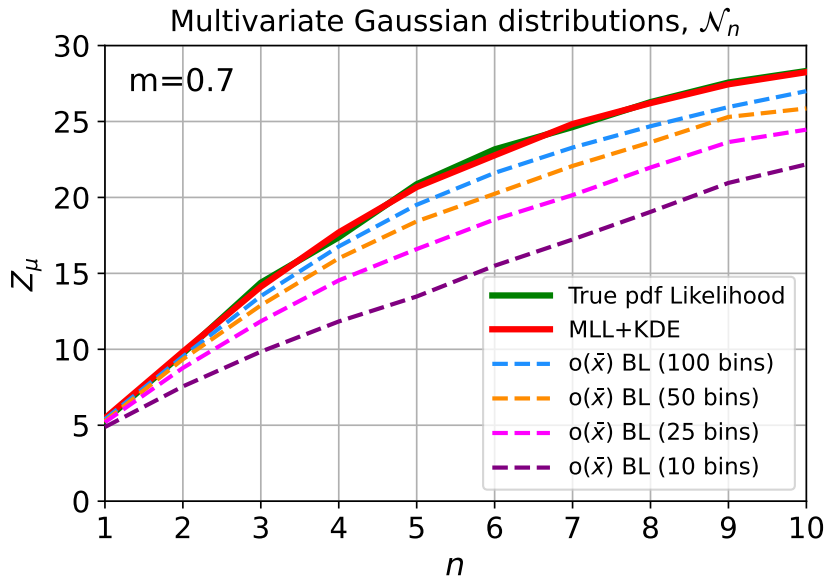




- Significances estimated with all method are all close to the true pdf scenario given the low dimensionality of the problem.
- The ML output is always 1D regardless the dimensionality of the data and can be easyly handled.

For higher dimensional data of $dim = n$ with $n > 2$, $\mathcal{N}_n(\boldsymbol{m}, \boldsymbol{\Sigma})$, $\boldsymbol{\Sigma} = \mathbb{I}_{n\times n}$, $\boldsymbol{m} = \{+0.3(-0.3), +0.7(-0.7)\}\ \mathbb{1}_n$ for S (B):





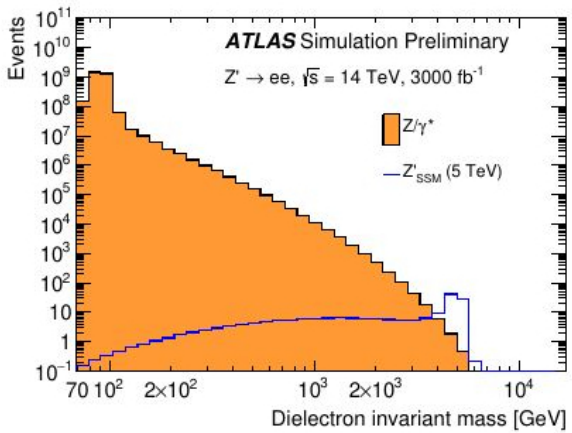
- Results with MLL+KDE method approach the ones with the true generative functions, independently of dimension and/or separation of S and B.
- BL intractable in the original space. BL in the ML output not as good as MLL+KDE. Undesirable dependence with the binning.

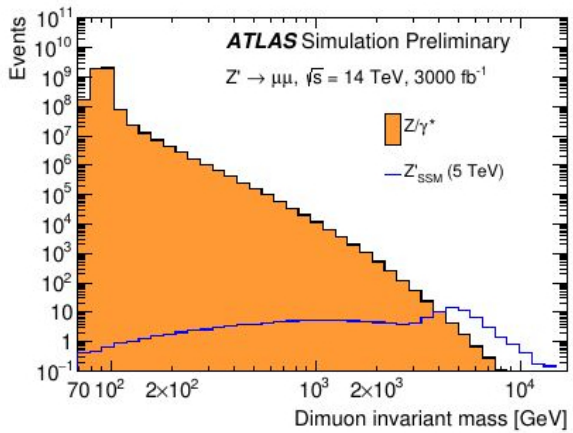
Search for a heavy SSM Z' in dilepton final states at the HL-LHC

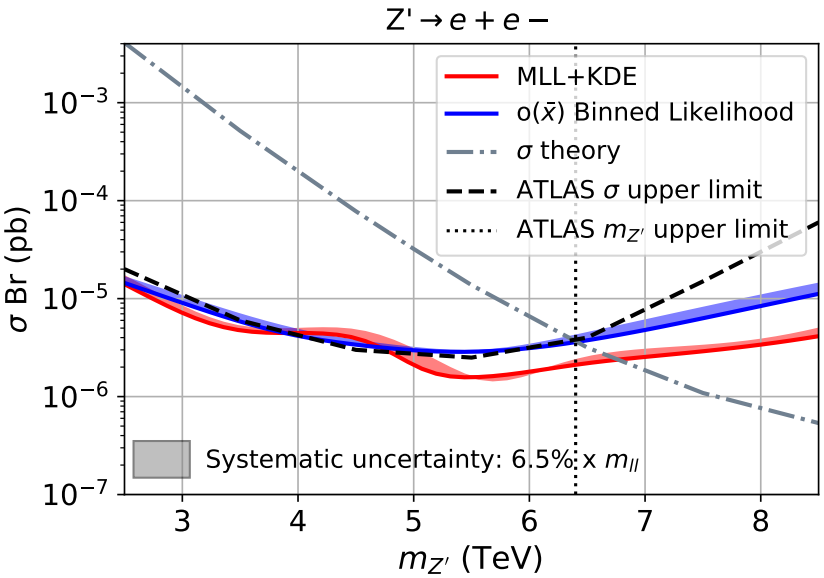
Based on ATLAS projections at $\sqrt{s} = 14$ TeV and 3000 fb^{-1} for 95% CL exclusion limits on a Z'_{SSM} (ATLAS-PHYS-PUB-2018-014).

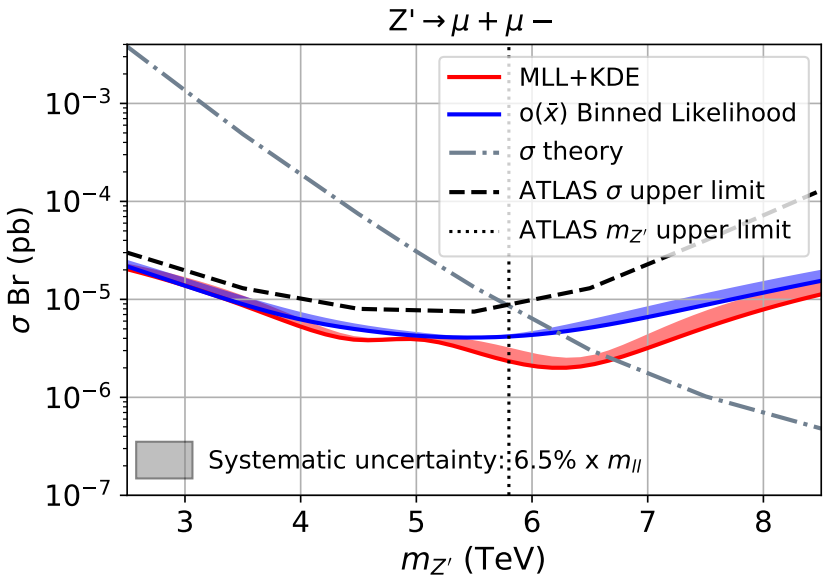
S: $p\,p \to Z' \to \ell^+\ell^-$ B: $p\,p \to Z/\gamma^* \to \ell^+\ell^-$

- Use of $|p_T|$, ϕ, and η of the final state leptons as ML inputs in each channel.
- Training of supervised per-event classifier, `XGBoost` with 1M events per class. KDE with Epanechnikov kernel.
- For saving computational resources, generation of events only with dilepton invariant mass over 1.8 TeV.







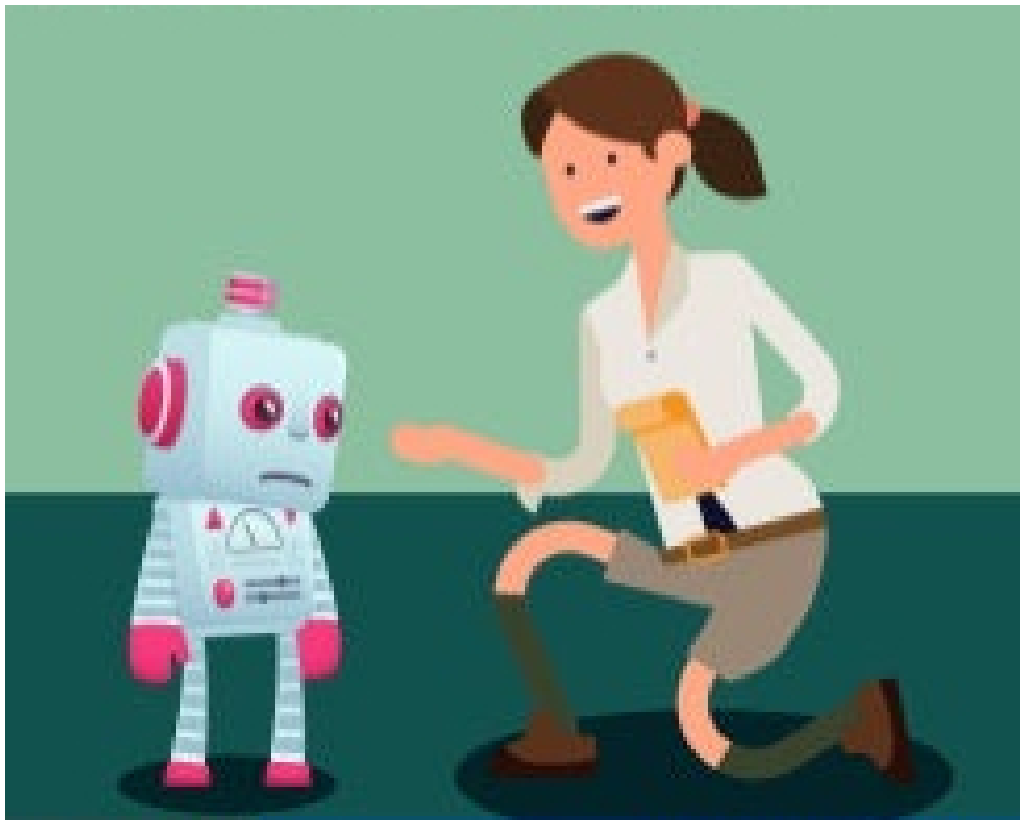


- In both channels, unbinning signal and background posteriors provide more constraining limits than binning output.
- For direct comparison with ATLAS projections, necessary to simulate full spectrum of invariant masses. Potentially biased results by possible enhancement of classifier performance.

Conclusions

- MLL method allows to obtain exclusion (and discovery) significances for additive new physics scenarios.
- Uses a single `XGBoost` classifier and its full 1D output (no working points), which allows the estimation of the S and B pdfs needed for statistical inference. Not strictly necessary to bin the output to extract the pdfs.
- Inclusion of KDE as extension of MLL method to avoid the binning of the ML classifier output.
- Improves results obtained by traditional techniques in toy models and realistic analysis, approaching (when possible) the ones computed with true generative functions.
- Possible improvements: unsupervised analysis, systematic uncertainties...

Thank you!



(special thanks to M. de los Rios and A. D. Perez)