Search for New Physics in Rare Decays

Phsyics as a Game

"The present situation in physics is as if we know chess, but we don't know one or two rules."

— Richard P. Feynman

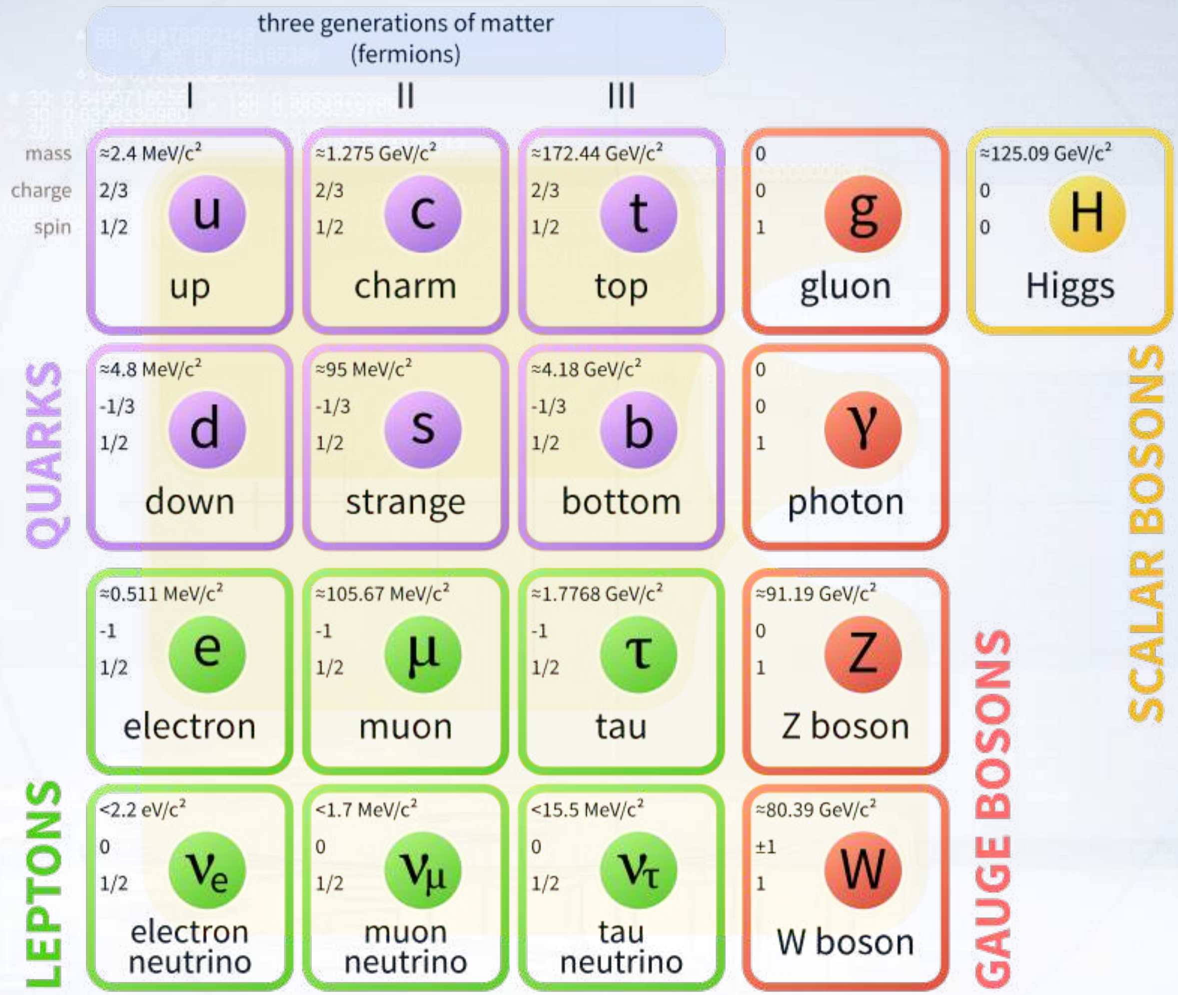
Phsyics as a Game

"The present situation in physics is as if we know chess, but we don't know one or two rules."

— Richard P. Feynman

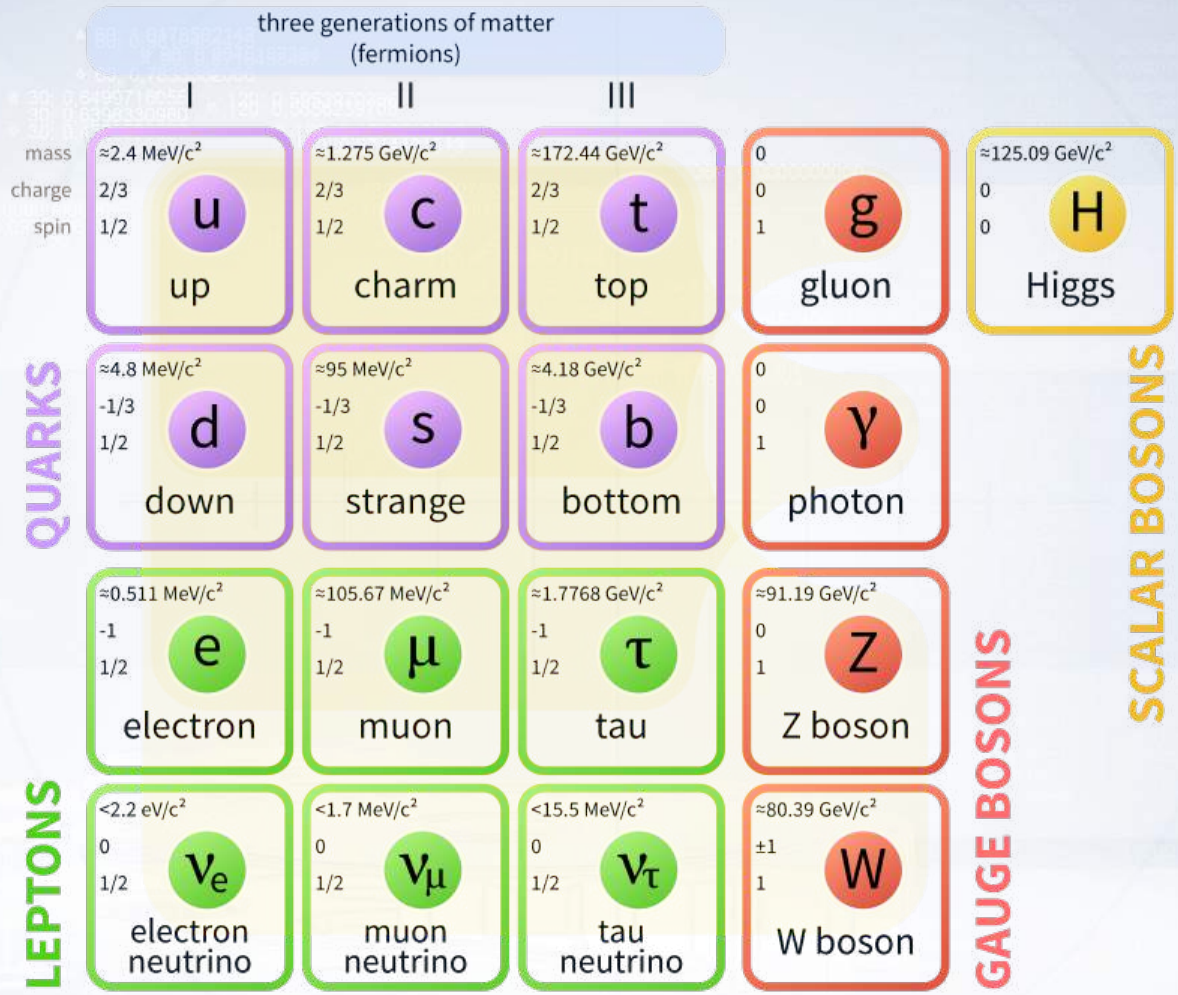
http://bit.ly/2DVIl4p

Standard Model (SM) of Elementary Particles



https://simple.wikipedia.org/wiki/Standard_Model

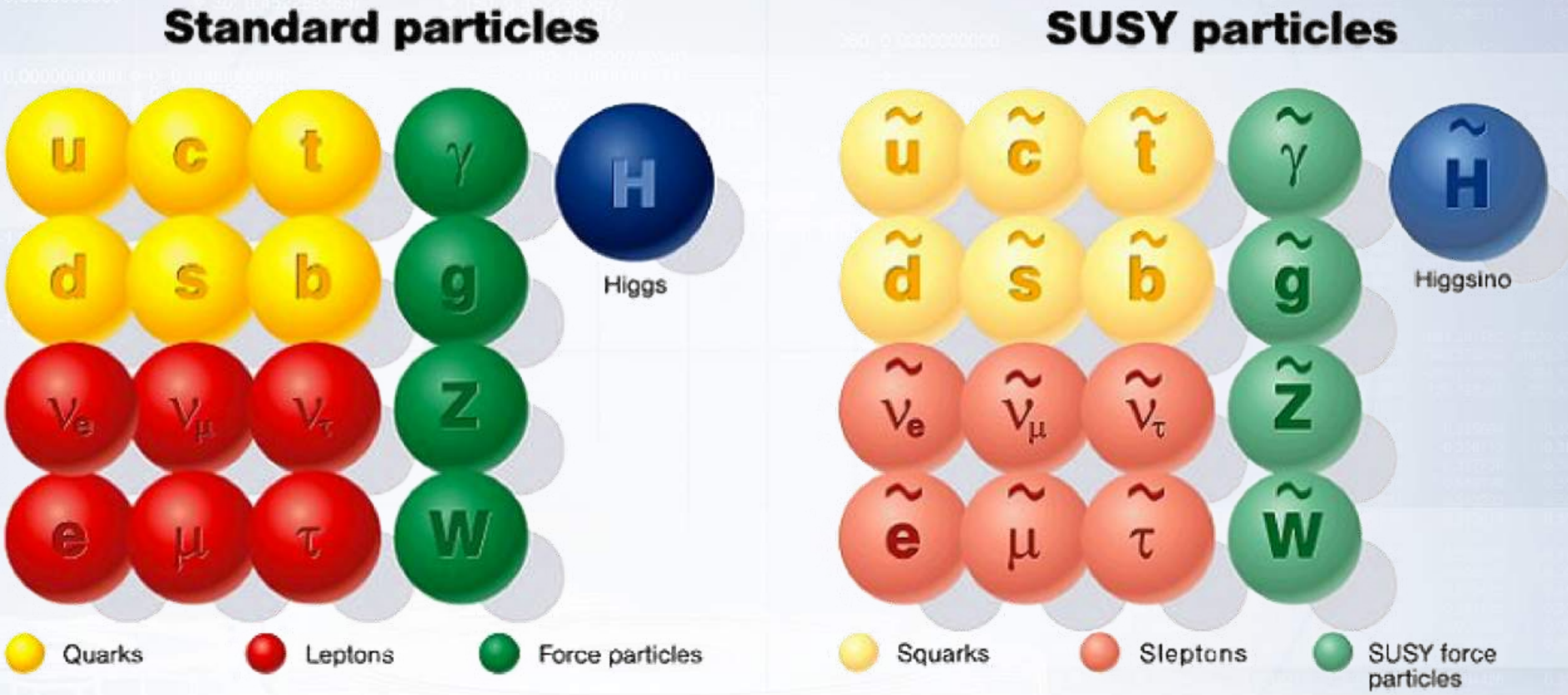
Standard Model (SM) of Elementary Particles



$$\mu^- \rightarrow e^- + \bar{\nu}_e + \nu_\mu$$

https://simple.wikipedia.org/wiki/Standard_Model

Super Symmetry (SUSY)



- Explains dark matter phenomena
- Capable of unifying 3 basic forces: electromagnetic, weak and strong

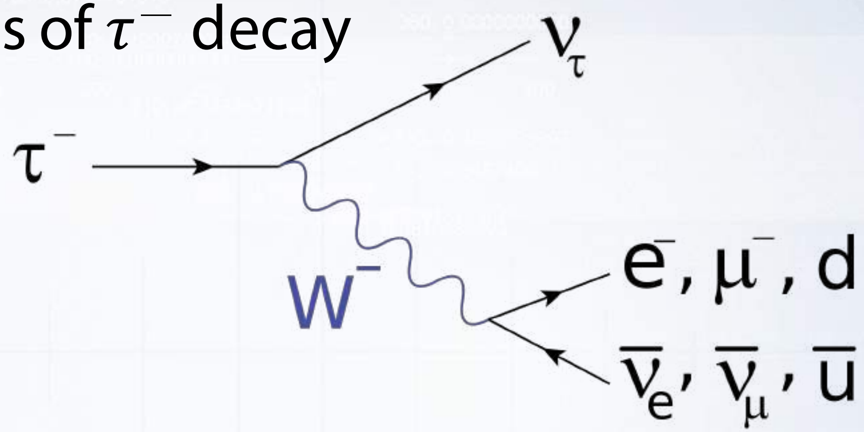
Alternative Model Testing

- Pick an alternative model (hypothesis) X, that a) explains an unknown phenomena i.e. dark matter and b) has different predictions for specific measurable quantity (e.g. *branching fraction* - B) for specific decay D;
- Estimate theoretical predictions of SM and X for B: B_{SM}, B_X;
- Make an experiment observing D and measure the actual value for the quantity – B_{obs} and the undecertainty σ;
- If B_{obs} is too far* from B_{SM} and close to $B_x \rightarrow$ Hail to the X! if it is too far from B_X, forget X. Otherwise wait for more data.

*) as a distance here we could use p-value for example. By convention, if it lies outside of 5- σ interval (probability of random fluctuation is <0.0000003) from B_{SM}, it is considered far enough.

Branching Fraction

- A particle can decay into different products, for example possible modes of τ^- decay



- Different modes have different probabilities or branching fractions (B), e.g.:
  - 17.83±0.04: electron e^- and electron antineutrino
  - 17.41±0.04: muon μ^- and muon antineutrino
  - 25.52±0.09: two pions π^-,π_0

http://bit.ly/2vuIKfk

Lepton Flavour Violation

Symmetry Invariants and Conservation Laws

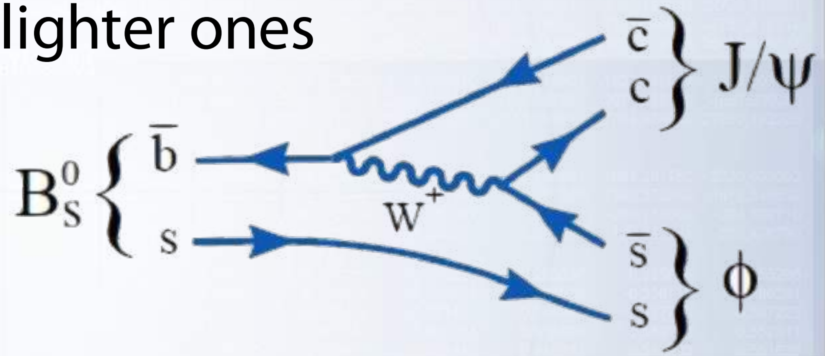
- One of the most profound theorems: Emmy Noether Theorem on relation between transformation invariance and physics quantity conservation law

Transformation Invariance	Symmetry	Conservation Law of
Time translation	Time uniformity	Energy
Space Translation	Space uniformity	Momentum
Space Rotation	Space isotropy	Angular momentum
Time and CP	Time isotropy	CP
Gauge symmetry	Gauge invariance	Charge, lepton number, …

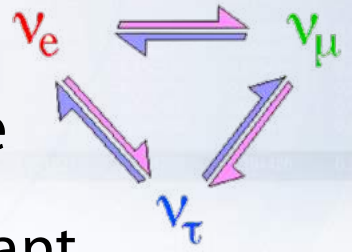
https://wiki2.org/en/Symmetry_in_physics+Newton

Lepton Flavour Violation (LFV)

- Flavour = Generation
- Rich phenomenology when weak interaction is involved
  - Quarks can transform into lighter ones



  - Neutrinos can transform into each other



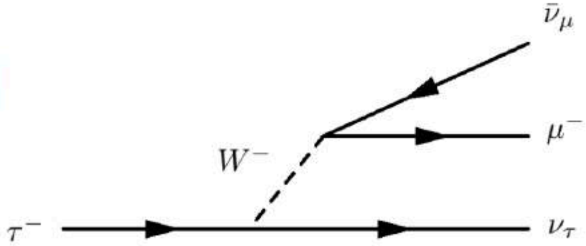
- SM predictions for charged LFV are negligible
- SUSY predictions for charged LFV are significant

$$\mu^- \to e^- + \gamma$$

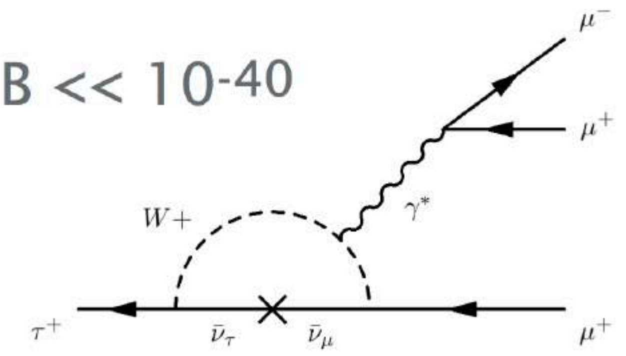
Ideal Ground to Search for New Physics

Tau decay & cLFV

SM

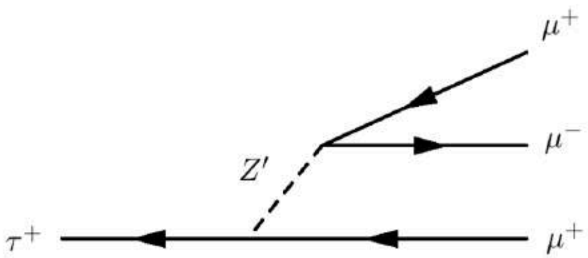


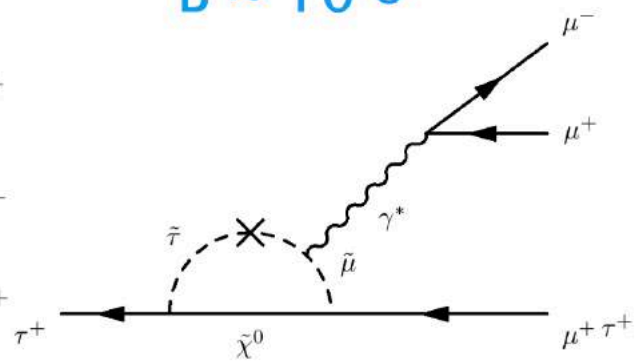
B << 10-40

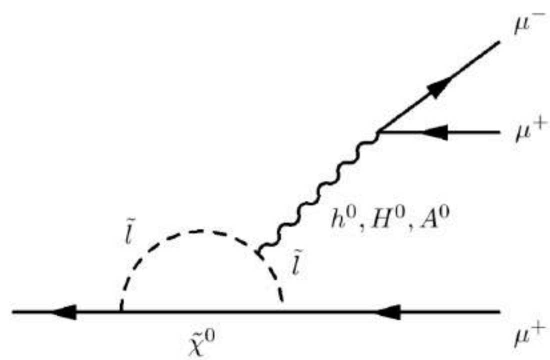


New physics

B $\overset{?}{\sim}$ 10-8







Large mass → many possible LFV final states
→ closer to high-mass new physics

Gerco Onderwater, NUFACT2014

http://cds.cern.ch/record/1751523/

Analysis Strategy

Signal-like channel:
- Real
- Simulated

Normalization channel:
- Real
- Simulated

Trigger (muon, secondary vertex)

Selection by angle, momentum, etc

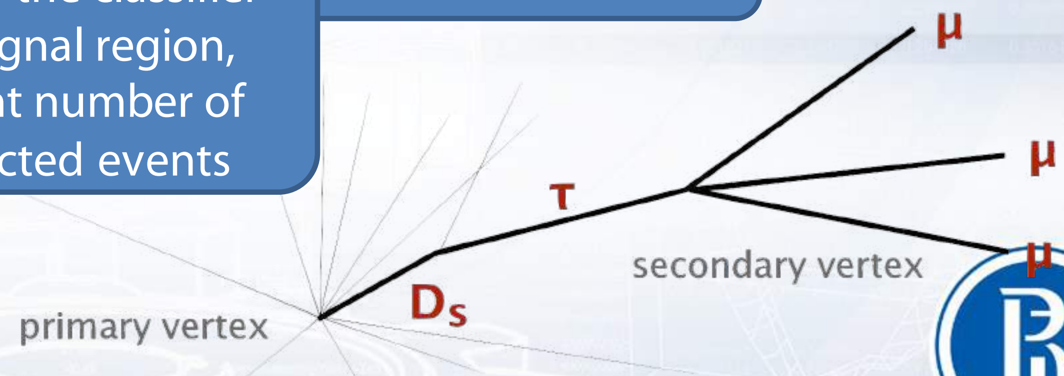
Hide signal region (Blinding)

Train classifier for final selection

Estimate background

Apply the classifier to signal region, count number of selected events

Compare the number with estimated background and normalize



Blinding motivation

Signal region - mass spectrum region with high probability of signal, i.e. $P(signal|X)$ is very different from $P(background|X)$ at the signal region.

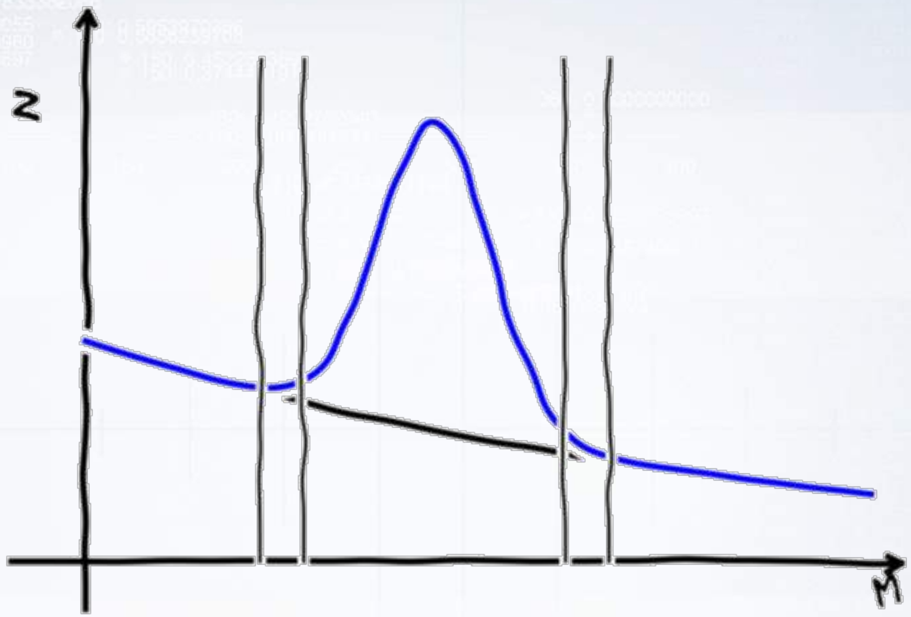
$$P \propto \frac{1}{((\mathbf{p}_Z)^2 - m_Z^2)^2 + \epsilon} = \frac{1}{((\mathbf{p}_{\mu^+} + \mathbf{p}_{\mu^-})^2 - m_Z^2)^2 + \epsilon}$$

This region is hidden during analysis to avoid psychological (experimentatlist) bias, i.e.

- Which cut should be applied?
- When stop analysis/searching for bug?

“We’re more than one sigma from zero; we have to look at it some more, because we must be doing something wrong…”

Blinding motivation



blue - the hypothetical signal distribution,
black - the distribution for the background.
The innermost region is the signal region.

arXiv:physics/0312102v1
DOI: 10.1146/annurev.nucl.55.090704.151521 DOI: 10.1088/0954-3899/28/10/312.

ML signal/background separation

Features include (good for signal/background discrimination):

- vertex fit quality, displacement from primary vertex, track quality, track isolation

Samples for training:

- Monte-Carlo simulated (MC) for signal
- Real data for background
- Similar channel $D_s \to \phi\,(\mu^+\mu^-)\,\pi^-$ used for calibration and normalization of the classifier

(Proxy) Metric:

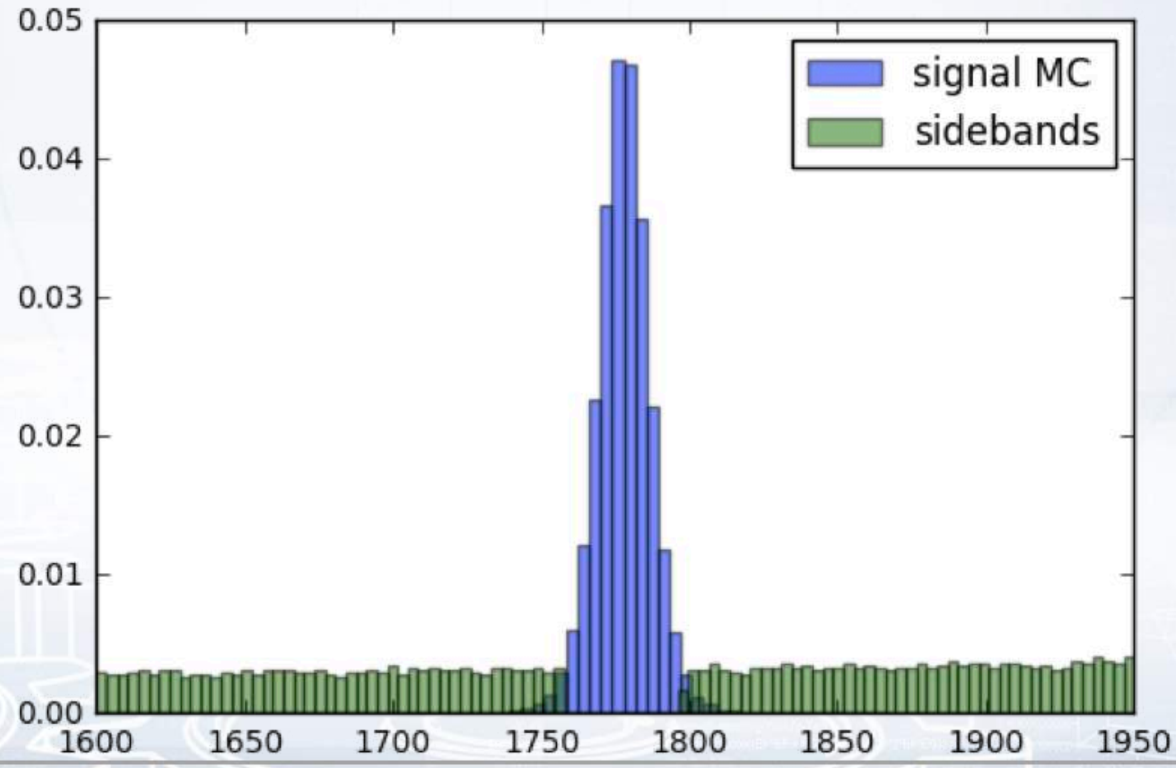
- ROC AUC

ML signal/background separation

Signal – peaking shape (e.g. Gaussian)

Background - exponential-like shape

What problems can you spot?



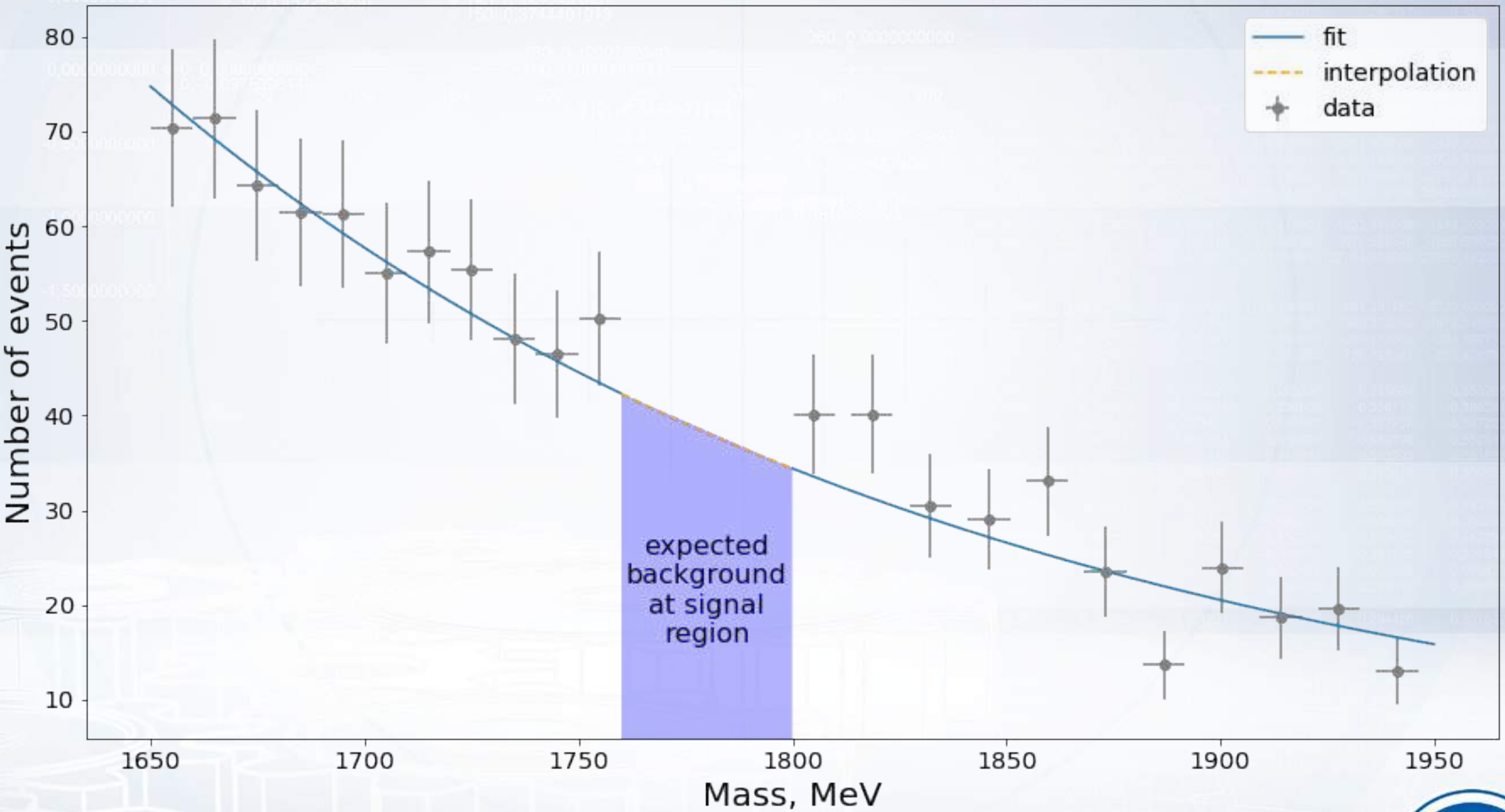
OK, we've got a model, what do we do with it?

- Get the best (significant) threshold value:
  - E.g. $t = \text{argmax}(\text{TPR}(t)^2 / \text{FPR}(t))$, (approximate)
- TPR – true positive rate (or signal fraction), FPR – false positive rate (background recognized as signal)
- Apply to real-data sample (still blinded)
- Estimate the amount of background events in the signal region (see next slide)
- Unblind: 1) apply classifier to signal region and count events, N_{sig} 2) apply it to signal region of normalization channel, N_{cal}
- Check hypothesis p-value and depending on it estimate branching ratio or upper limit

Expected Background Estimation

- Apply selection to sidebands
- Assume parametric pdf for combinatorial background, like exponential (background model)
- Fit the model to real data in the sidebands; check that PDF performs well (e.g. using χ^2 criteria)
- Extrapolate the model to the blind region and compute the area under this extrapolation
- Estimate expected number of background events in the blind region

Expected Background Estimation (illustration)



Normalization - I

Branching fraction for $\tau^- \to \mu^-\mu^+\mu^-$ normalized to $D_s^- \to \Phi(\mu^+\mu^-)\pi^-$

$$B = \frac{N(\tau \to \mu\mu\mu)}{N(\tau)} = factor \times \frac{N_{sig}}{N_{cal}}$$

$$N(\tau) \to 1/f_\tau^{D_s} N(D_s \to \tau \bar{\nu}_\tau)$$

calculated

$$N(D_s \to \tau \bar{\nu}_\tau) \to B(D_s \to \tau \bar{\nu}_\tau) N(D_s)$$

literature

$$N(D_s) \to N(D_s \to \phi(\mu\mu)\pi) / B(D_s \to \phi(\mu\mu)\pi)$$

literature

$$B(D_s \to \phi(\mu\mu)\pi) \to B(D_s \to \phi(KK)\pi) B(\phi \to \mu\mu) / B(\phi \to KK)$$

Must further include trigger, selection & reconstruction **efficiencies**

Gerco Onderwater, NUFACT2014

PLB 724, 36 (2013)

Normalization – II

Branching fraction for $\tau^- \to \mu^-\mu^+\mu^-$ normalized to $D_s^- \to \phi(\mu^+\mu^-)\pi^-$

$$\mathcal{B}(\tau^- \to \mu^-\mu^+\mu^-)$$

$$= \underbrace{\mathcal{B}(D_s^- \to \phi(\mu^+\mu^-)\pi^-)}_{(1.33\pm0.12)\times10^{-5}} \times \frac{\overbrace{f_\tau^{D_s}}^{0.78\pm0.05}}{\underbrace{\mathcal{B}(D_s^- \to \tau^-\bar{\nu}_\tau)}_{0.0561\pm0.0024}}$$

$$\times \underbrace{\frac{\epsilon_{\mathrm{cal}}^{\mathrm{REC\&SEL}}}{\epsilon_{\mathrm{sig}}^{\mathrm{REC\&SEL}}}}_{1.49\pm0.12} \times \underbrace{\frac{\epsilon_{\mathrm{cal}}^{\mathrm{TRIG}}}{\epsilon_{\mathrm{sig}}^{\mathrm{TRIG}}}}_{0.753\pm0.037} \times \underbrace{\frac{N_{\mathrm{sig}}}{N_{\mathrm{cal}}}}_{48076\ \pm\ 840}$$

$$= \underbrace{\alpha}_{(4.34\pm0.65)\times10^{-9}} \times N_{\mathrm{sig}}$$

PLB 724, 36 (2013)

Analysis Outcome

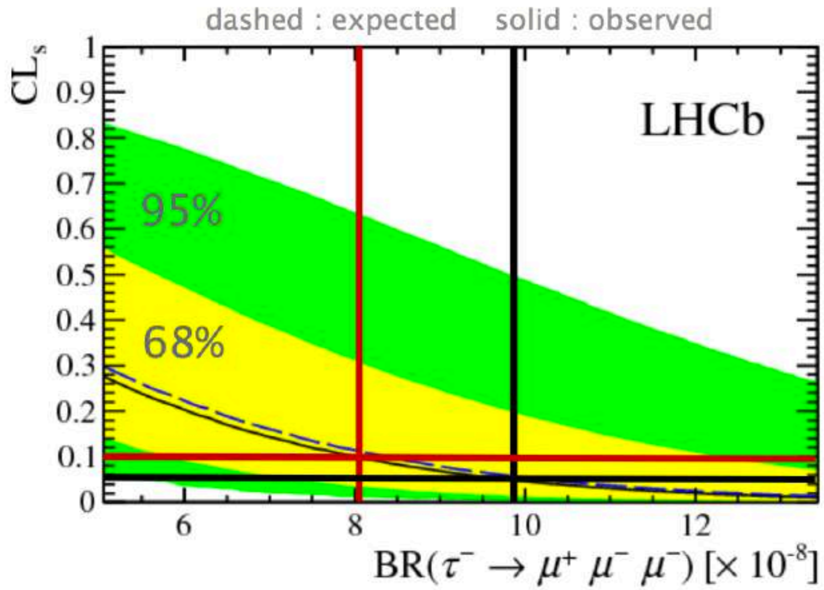
- No significant evidence for an excess of events
- CL_s method used to extract upper limit
  Likelihood ratio **signal+background** vs **background-only**

$B(\tau^- \to \mu^-\mu^+\mu^-) < 8.0\ (9.8) \times 10^{-8}$

@ 90% (95%) C.L

Belle 2.1×10^{-8} @ 90% C.L.
PLB 687, 139 (2010)

BaBar 3.3×10^{-8} @ 90% C.L.
PRD 81, 111101(R) (2010)



PLB 724, 36 (2013)

Classifier Constraints

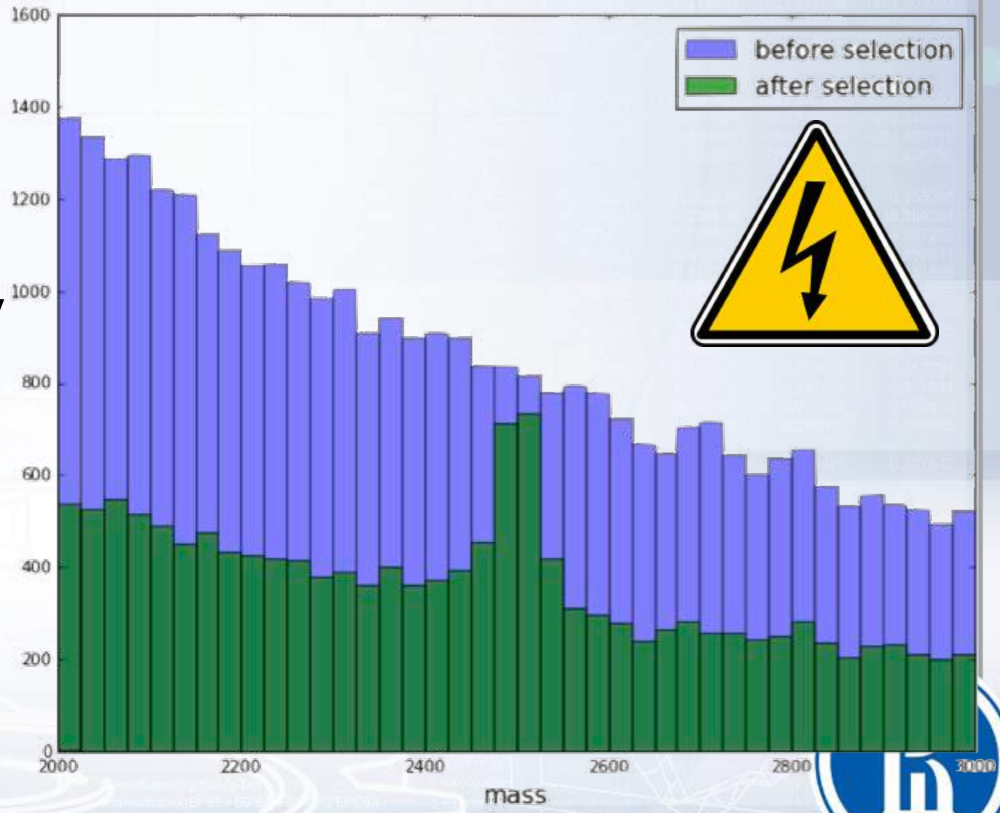
Classifier Restrictions

Uniformity

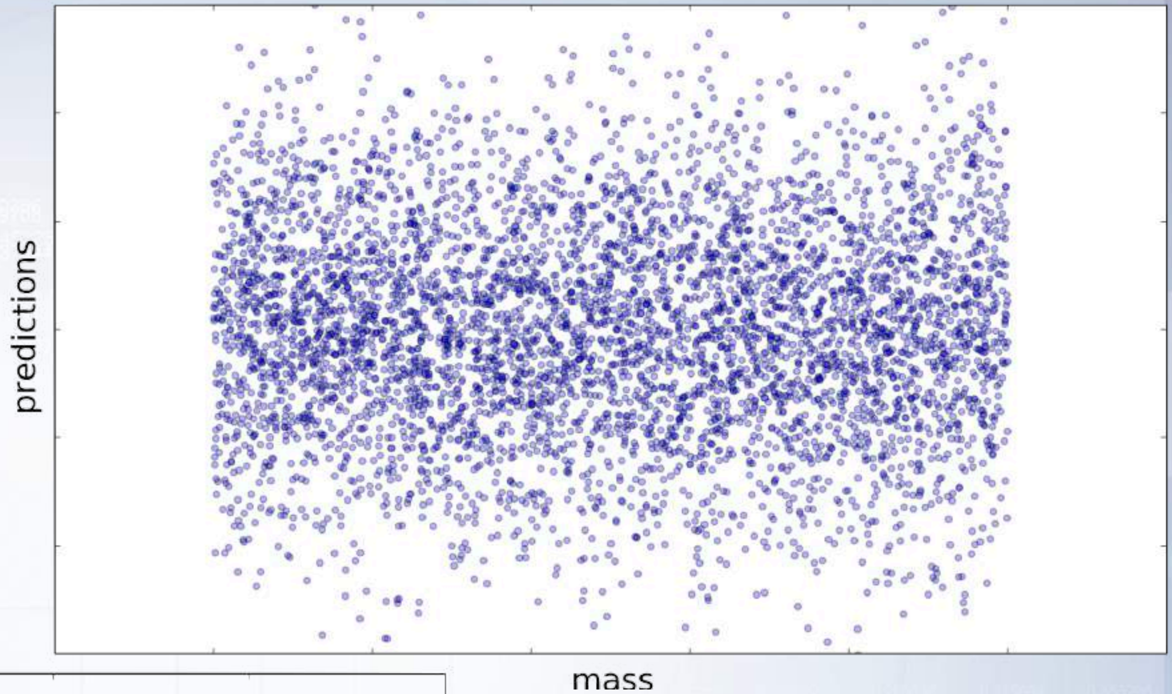
Correlation between classifier prediction and mass can lead to false peaks which spoil (bias) event counting.

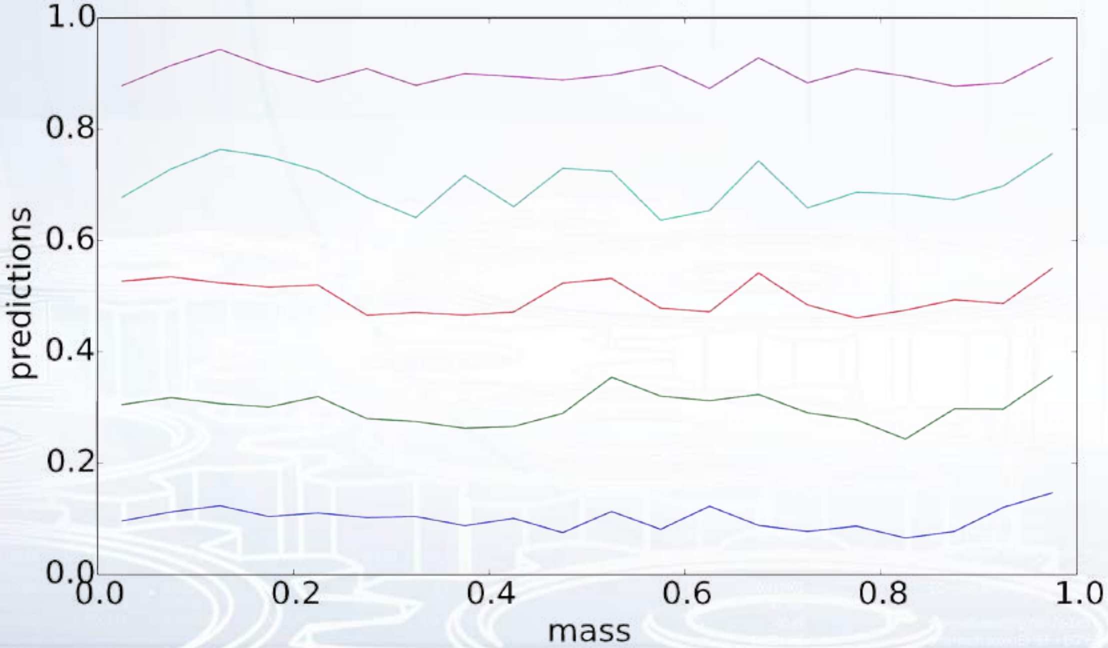
Agreement

In training dataset signal is represented by simulated events and background by real. Thus classifier might pick MC-specific features, which also bias counting.



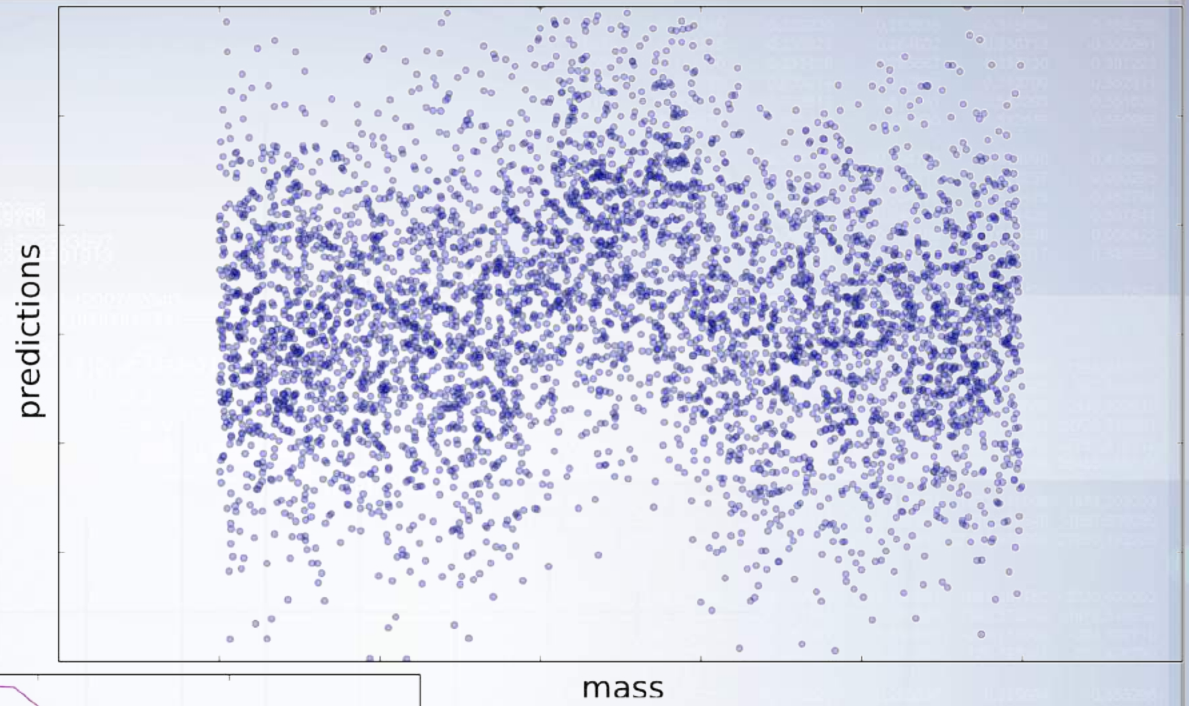
Uniformity

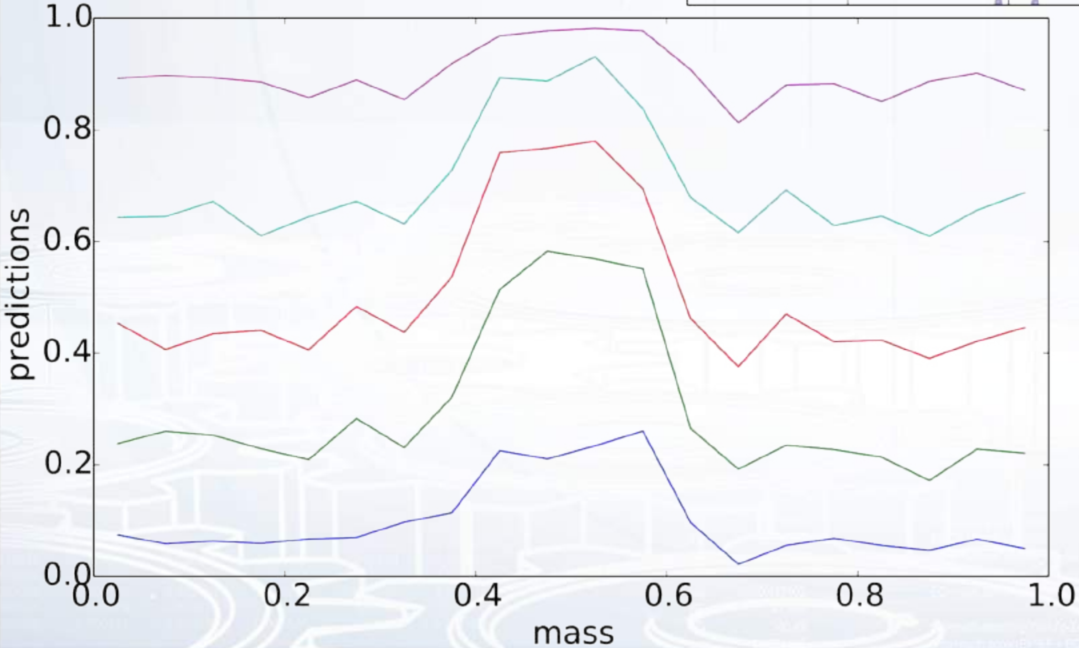




Each line shows PDF of events selected by
TPR = {0.1, 0.3, 0.5, 0.7, 0.9}

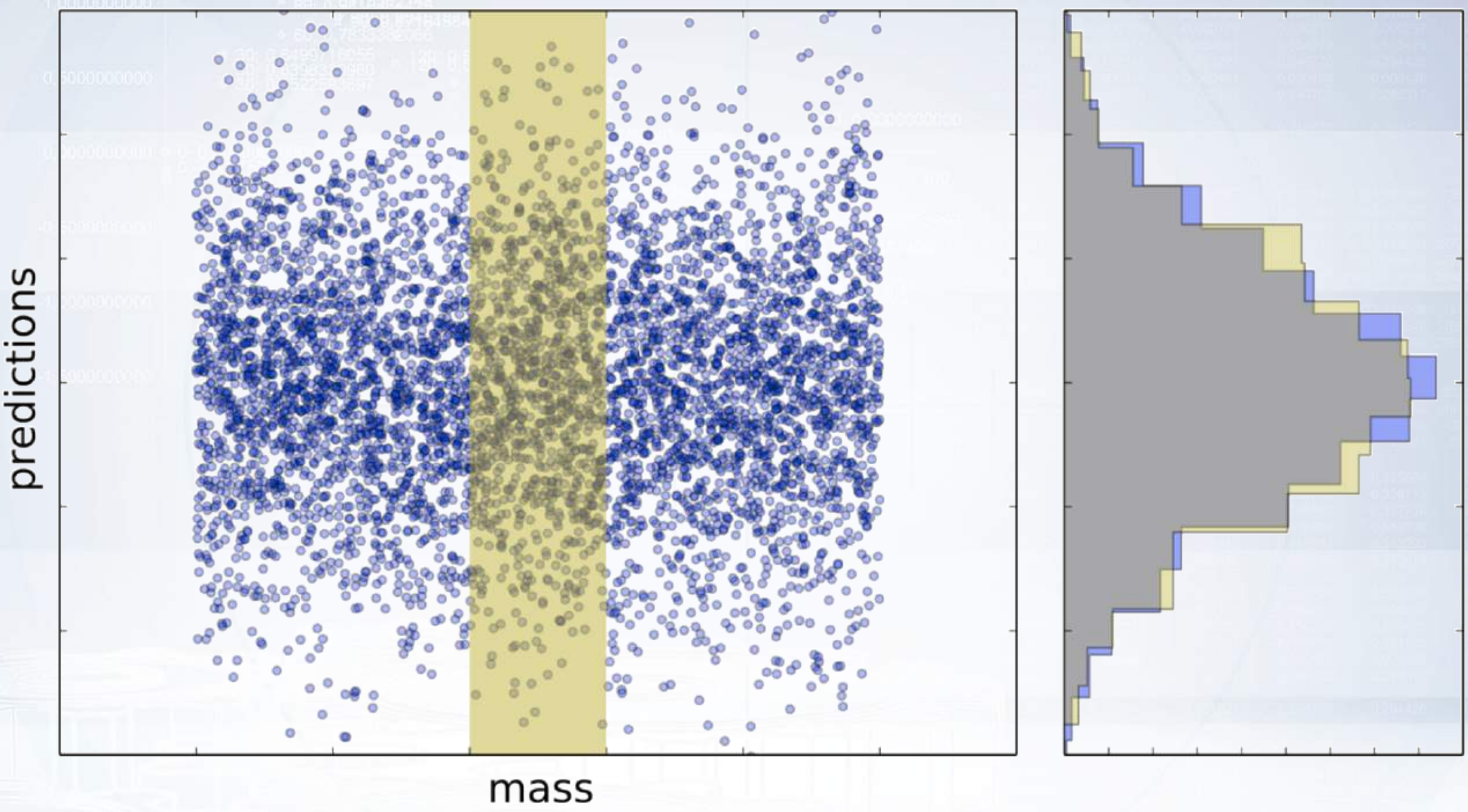
Non-Uniformity





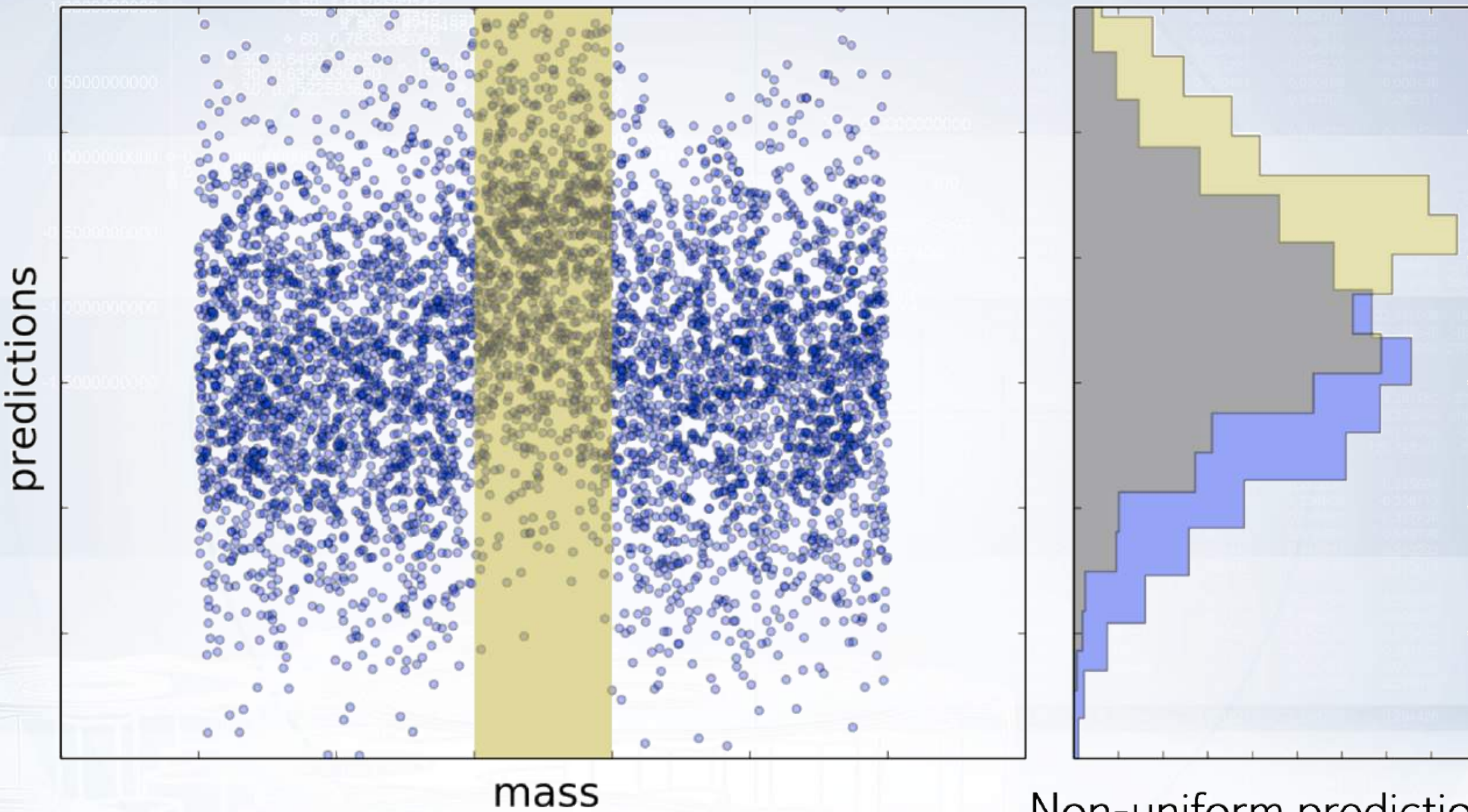
Each line shows PDF of events selected by
TPR = {0.1, 0.3, 0.5, 0.7, 0.9}

Non-Uniformity Measure



Uniform predictions

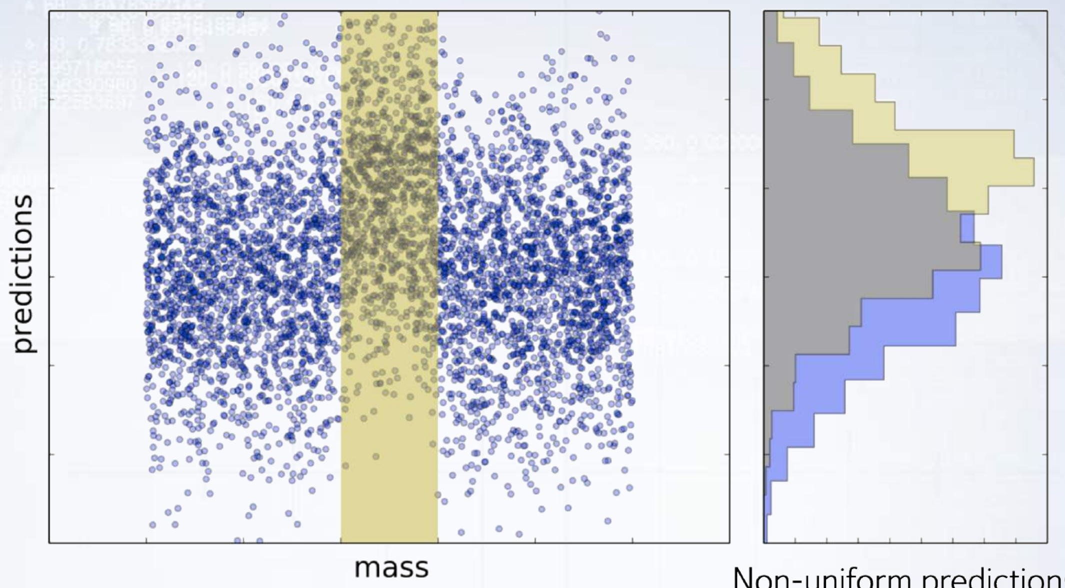
Non-Uniformity Measure



Non-uniform predictions (peak in highlighted region)

uniformity = no dependence between mass and predictions

Non-Uniformity Measure



Non-uniform predictions (peak in highlighted region)

$$\mathrm{CvM} = \sum_{\mathrm{region}} \int \left| F_{\mathrm{region}}(s) - F_{\mathrm{global}}(s) \right|^2 dF_{\mathrm{global}}(s)$$

Cramer-von Mises test (integral characteristic), where

F_{region} – CDF for region distribution (yellow)

F_{global} – CDF for global distribution (blue)

Uniformity Check

- random predictions and mass can be considered independent variables;
- assume null-hypothesis: mass and predictions are independent;
- generate distribution of CvM value under null-hypothesis by repeating many times the following steps:
  - generate random predictions;
  - compute CvM value;
- choose p-value and compute corresponding CvM value.

Basic Approach

Reduce the set of features used in training: leave only those, which do not correlate with the mass:

- It is simple and it works;
- But omitting those features we loose classification power.

Can we modify ML algorithm to use all features, but provide uniform background efficiency (FPR)/signal efficiency (TPR) along the mass?

Gradient Boosting Recap

Gradient Boosting greedily builds an ensemble of estimators

$$D(x) = \sum_j \alpha_j d_j(x)$$

That minimize given loss function. Those losses could be:

- MSE: $$\mathcal{L} = \sum_i (y_i - D(x_i))^2$$
- AdaLoss: $$\mathcal{L} = \sum_i e^{-y_i D(x_i)}, \qquad y_i = \pm 1$$
- LogLoss: $$\mathcal{L} = \sum_i \log(1 + e^{-y_i D(x_i)}), \qquad y_i = \pm 1$$

Each term in the ensemble approximates the residuals between true y_i and all the preceding terms.

uBoostBDT

Aims to get **FPR**$_{region}$=const:

- Fix target efficiency, for example **FPR$_{target}$**=30%, and find corresponding threshold
- Train a tree, its decision function is $d(x)$
- Increase weight for misclassified: $w_i \leftarrow w_i \exp(-\alpha y_i d(x_i)))$
- Increase weight of background events in the regions with high **FPR**

$$w_i \leftarrow w_i \exp\left(\beta(\mathrm{FPR}_{\mathrm{region}} - \mathrm{FPR}_{\mathrm{target}})\right)$$

Thus we achieve **FPR$_{region}$**=30% in all regions

Computationally complex, and may get biased.

uBoostGB + FlatnessLoss

- Why not minimize CvM with Gradient Descent?
  … we can't compute the gradient!

- CvM approximation:

$$\mathcal{L}_{FL} = \sum_{region} \int |F_{region}(s) - F_{global}(s)|^2 ds$$

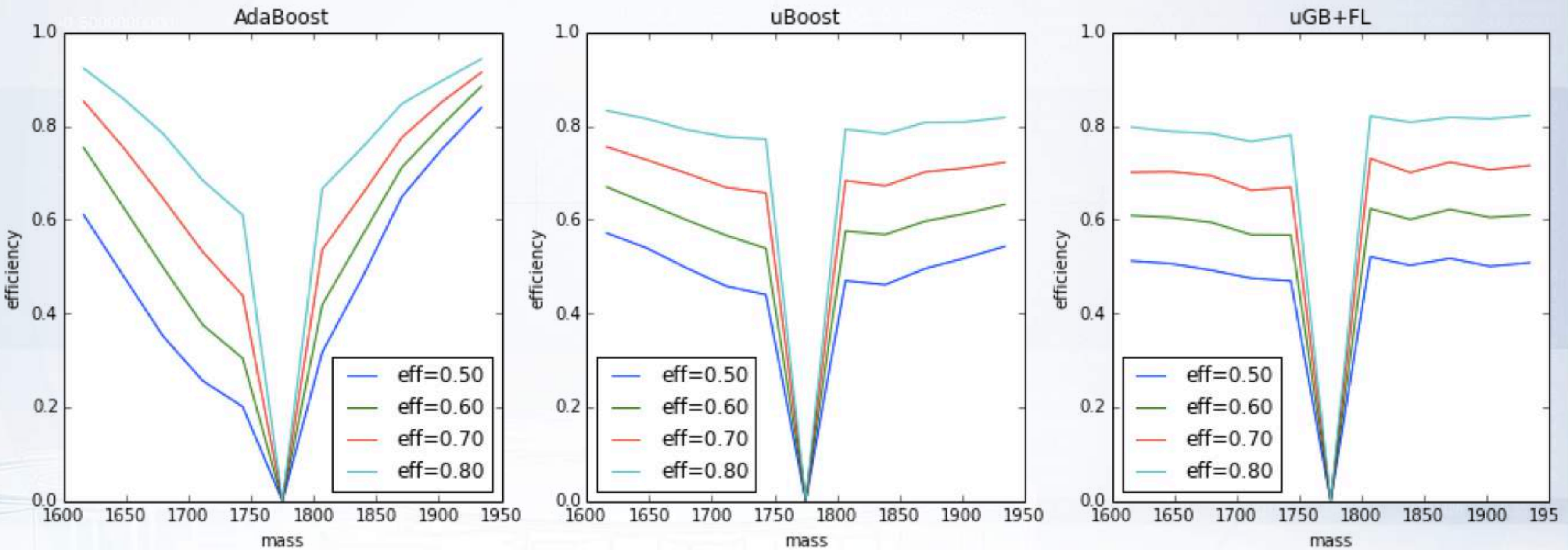
$$\frac{\partial}{\partial D(x_i)} \mathcal{L}_{FL} \sim 2(F_{region}(s) - F_{global}(s))|_{s=D(x_i)}$$

- Add approximate CvM to a loss function (regularize):

$$\mathcal{L} = \mathcal{L}_{adaloss} + \alpha \mathcal{L}_{FL}$$

arXiv:1410.4140

Improving Uniformity



- uBoostGB+FL is faster and allows for trade-off between quality / uniformity

Adversarial Approach

Chase_2017.03.22_IML.pdf (стр. 21 из 37)

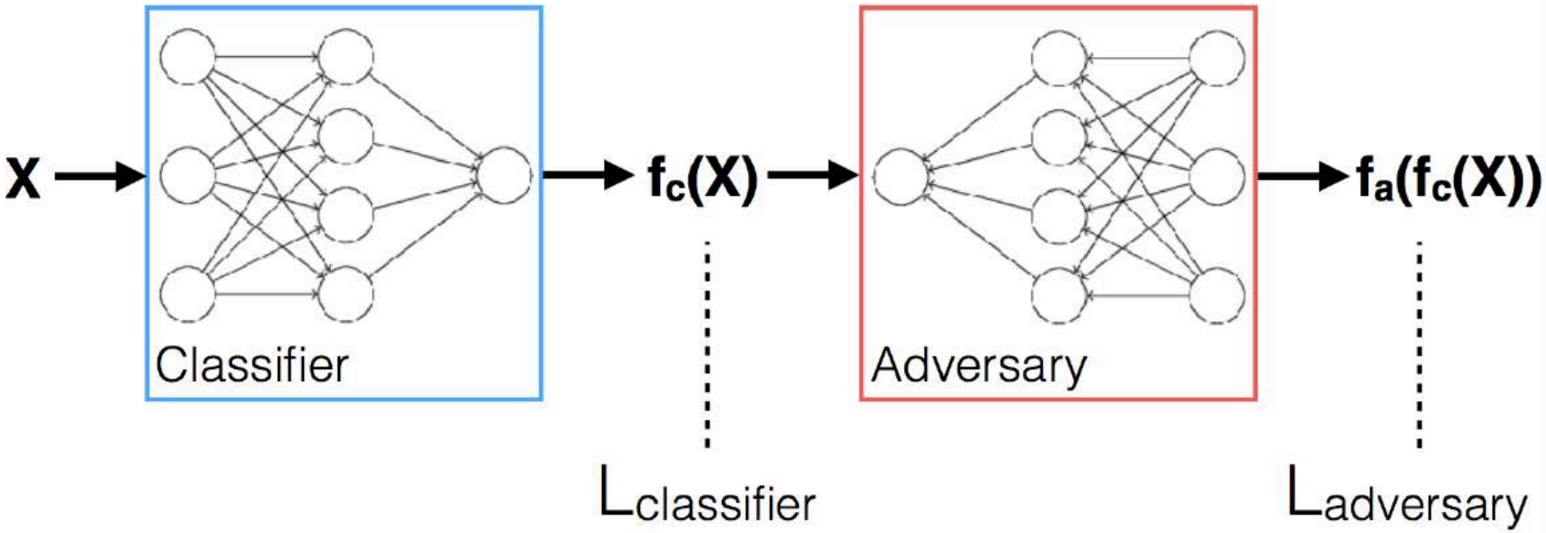
Adversarial Decorrelation

Simultaneously minimize:

$$L_{adversary}$$

and

$$L_{tagger} = L_{classifier} - \lambda L_{adversary}$$



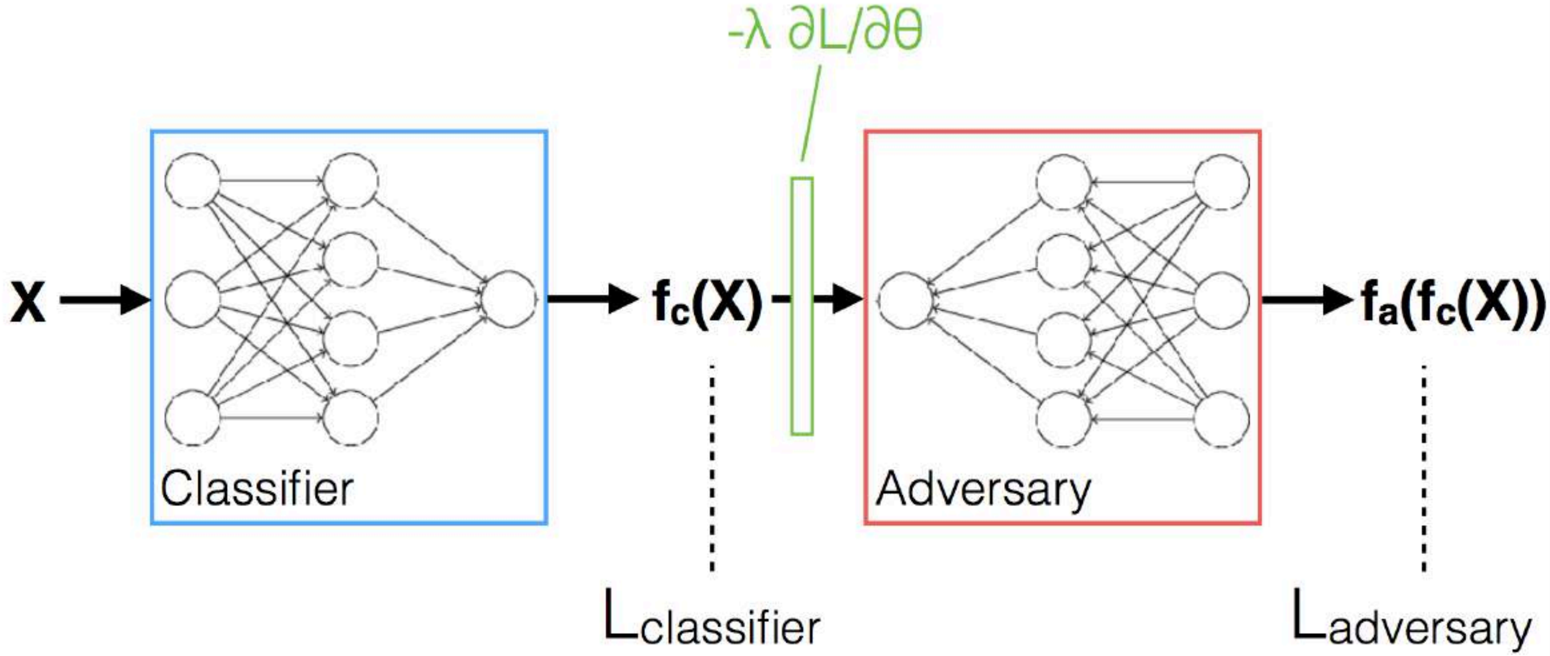
http://bit.ly/2GefGtq

Adversarial Approach

Chase_2017.03.22_IML.pdf (стр. 23 из 37)

Training

- Simultaneous optimization achieved with **gradient scaling layer**
- Signal events are given zero weight in adversary loss



http://bit.ly/2GefGtq

Adversarial Approach

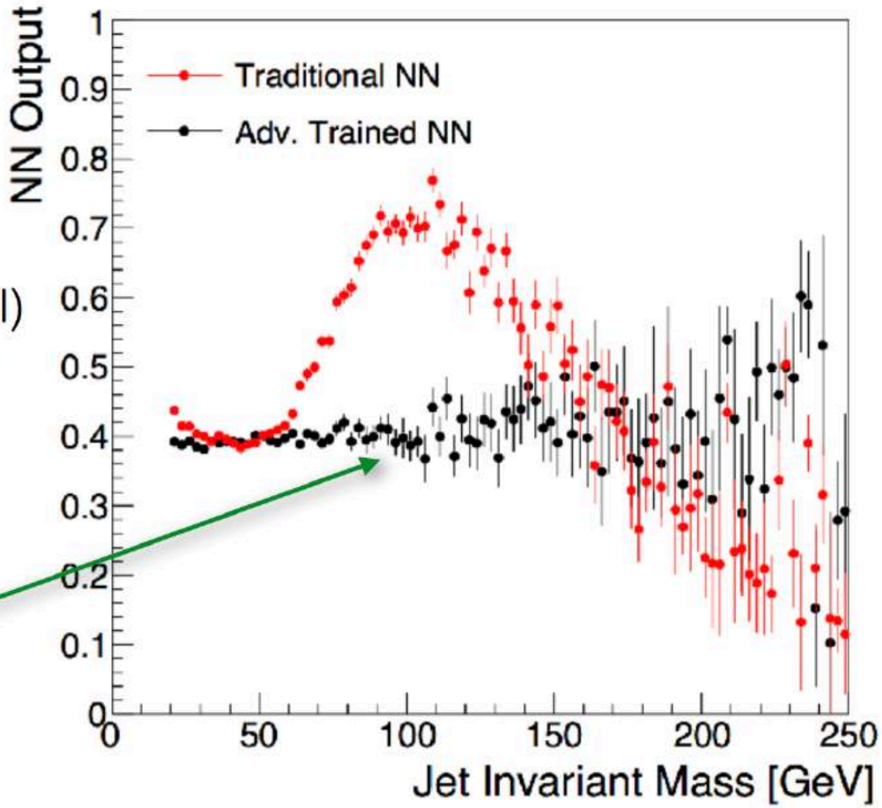
Chase_2017.03.22_IML.pdf (стр. 24 из 37)

Results

Training on ~200k MC events:

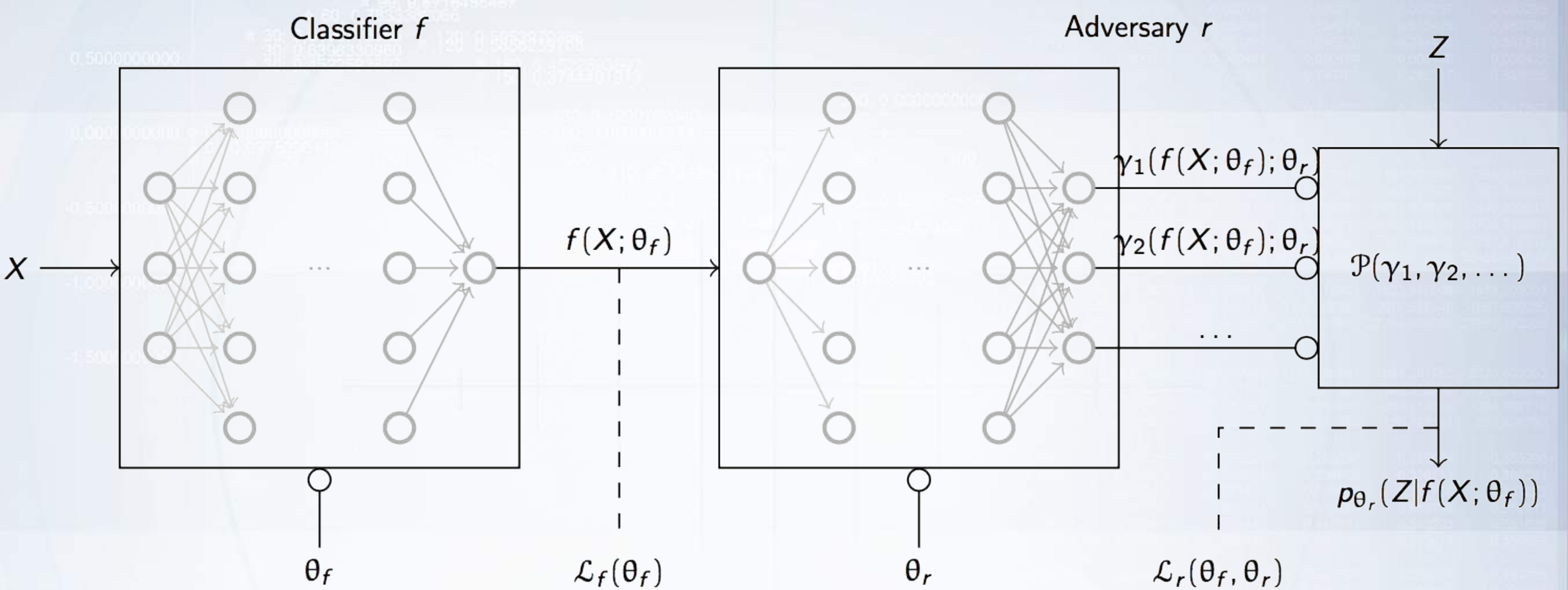
Sherpa γ+jet (BG)
MG5 γ+Z' (Signal)
Pythia + Delphes (Both)

✓ Tagger profile much flatter



http://bit.ly/2GefGtq

Going Deeper with Adversarial Training



$$\hat{\theta}_f, \hat{\theta}_r = \arg\min_{\theta_f}\max_{\theta_r} \mathcal{L}_f(\theta_f) - \mathcal{L}_r(\theta_f,\theta_r)$$

Here the adversary part identifies PDF *parameters* that can be used to infer Z (decor. feature) from classifier *f* output.

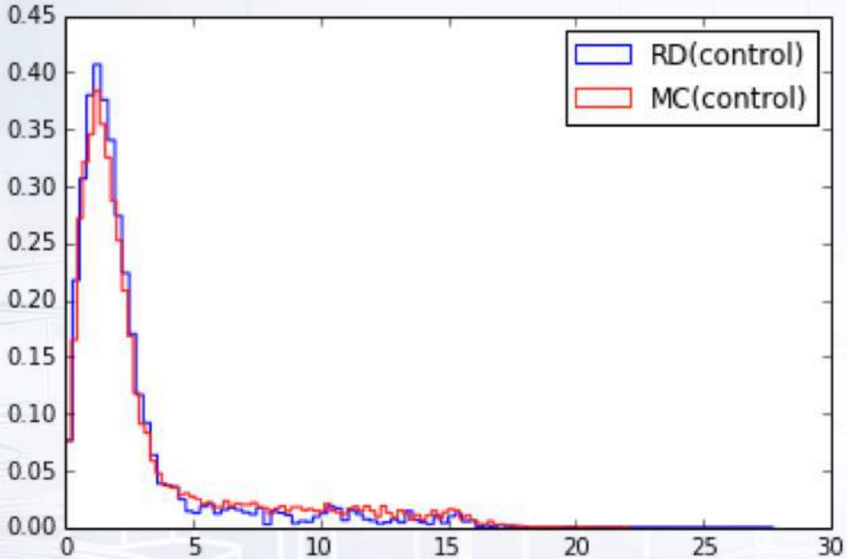
Intuitively, *r* penalizes *f* so it is impossible to reconstruct Z.

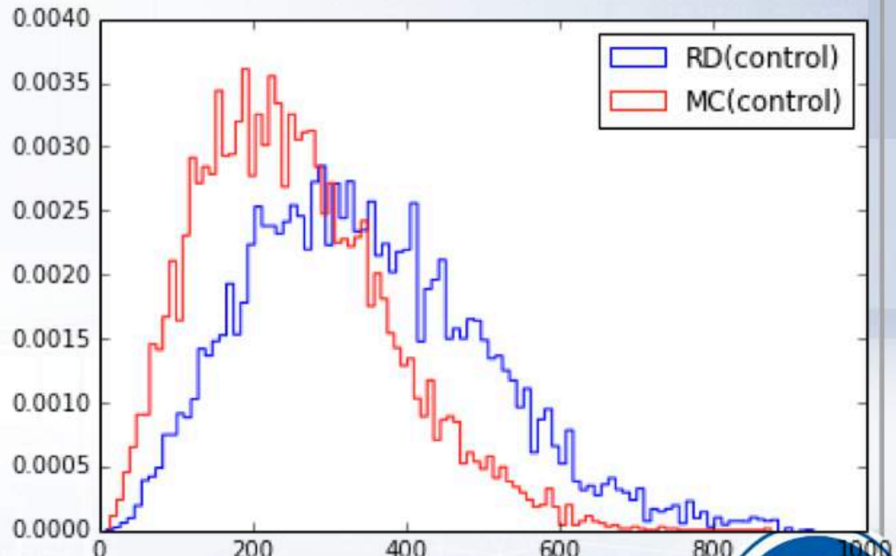
G. Louppe et al, http://bit.ly/2GejPgY

Data vs Simulation Agreement

Real Data vs Simulation Agreement

- Classifier is trained for simulated signal vs real background
- Not all features are perfectly simulated: MC and real data disagree (plots below)
- **Problem**: clasiffier efficiency might be overestimated





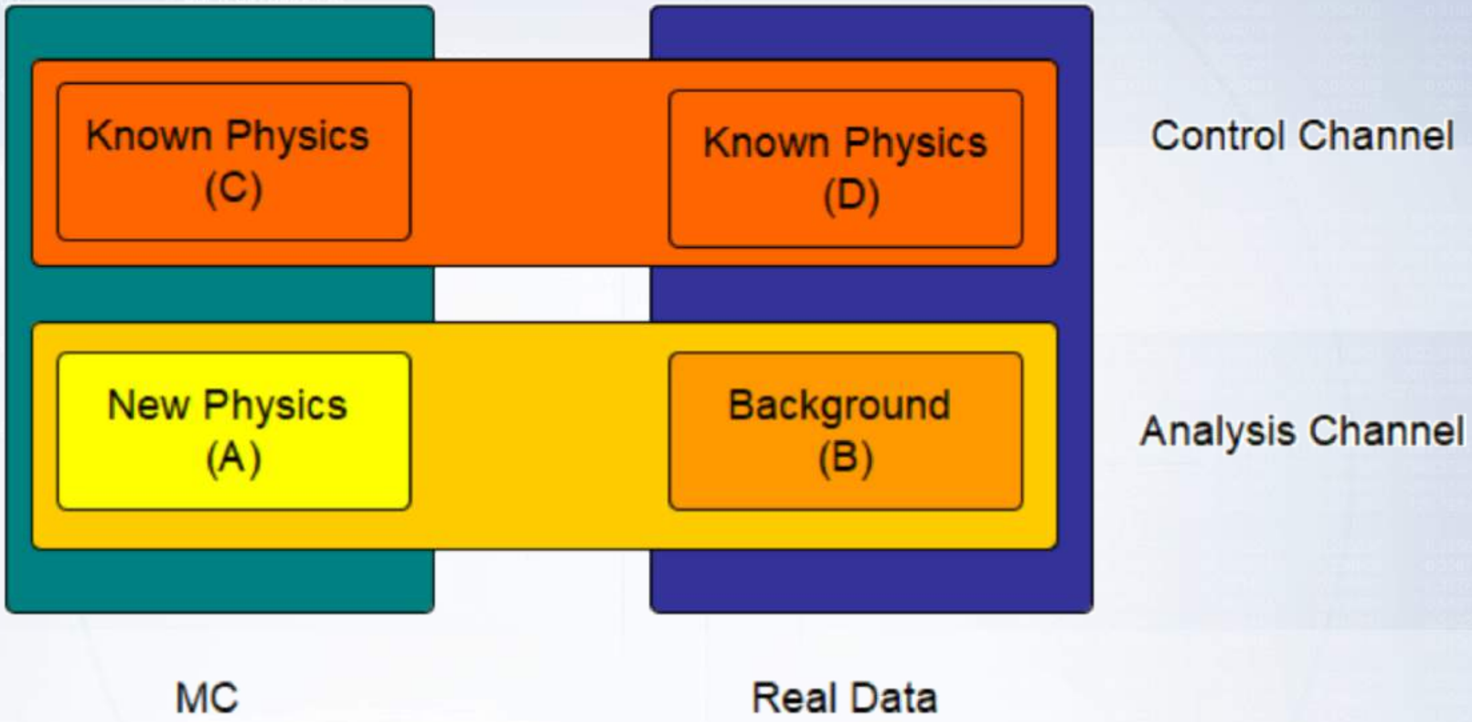
Approach

- Pick control channel of similar topology: $D_s \to \phi\,(\mu^+\mu^-)\,\pi^-$;
- $D_s \to \phi\,(\mu^+\mu^-)\,\pi^-$ is a much better known channel, can be extracted from the data;
- Compare performance of classifier on simulated and real samples using Kolmogorov-Smirnov test:

$$T=sup_x\left|F_1(x)-F_2(x)\right|$$

- Demand the distance to be below certain margin. How can we include this criteria in the training loop?

Data Doping

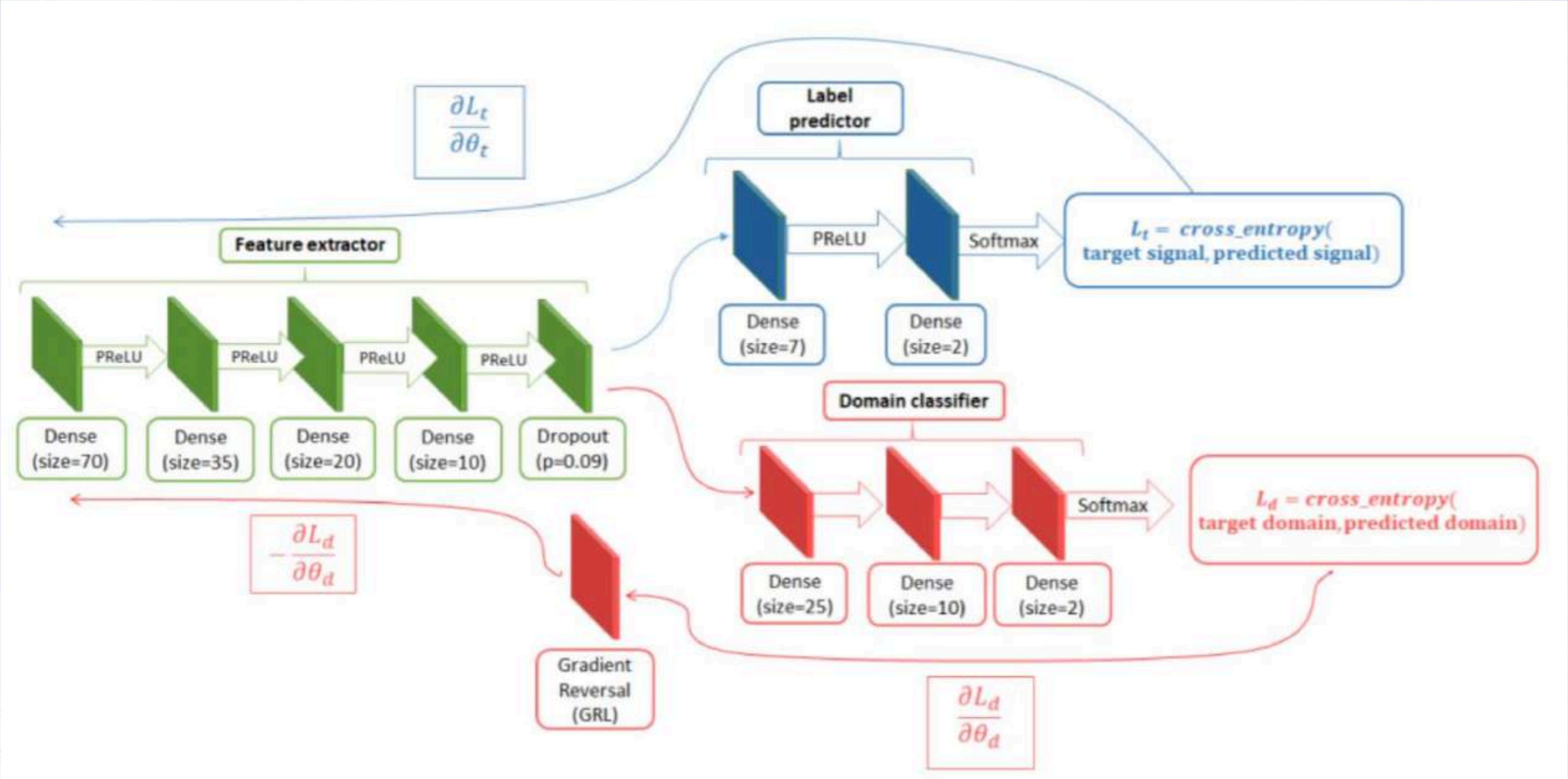


Let classifier discriminate A from B, not C from D:

- Add fraction of simulated signal (C) to the training sample with 'background' label ('*doping*')

Vicens Gaitan: http://bit.ly/2pLgVcy

Domain Adaptation by Gradient Reversal



- Feature extractor – builds meaningful representation (features);
- Label predictor – discriminates between signal and background;
- Domain classifier – discriminates between MC and real.

Domain Adaptation by Gradient Reversal

Let's compare the techniques on $\tau \to \mu\mu\mu$ dataset:

Model / Metric	Mass-aware Classifier	Data Doping	Domain-adaptation
AUC (truncated)	**0.999**	0.9744	0.979
KS (< 0.09)	0.18	0.087	**0.06**
CvM (< 0.002)	0.0008	0.0011	**0.0008**

- By varying learning rate for feature extractor and domain classifier it is possible to trafeoff classifier quality to degree of agreement.
- More details you can find in the paper by A. Ryzhikov et al: http://bit.ly/2GHCs06

Conclusion

- Finding New Physics (NP) is one of the LHC goals
- Search for NP in rare decays: compare models predictions with experiment measurements
- Complicated strategy
- Metric is not trivial:
    - Sensitivity (ROC AUC + TPR^2/FPR)
    - Uniformity and Agreement
- $\tau \rightarrow \mu\mu\mu$ decay is waiting for the discovery
- Described ML techniques still can be helpful!
    - Classifiers + special loss or Adversarial Networks

GHHHK

HOUJIX