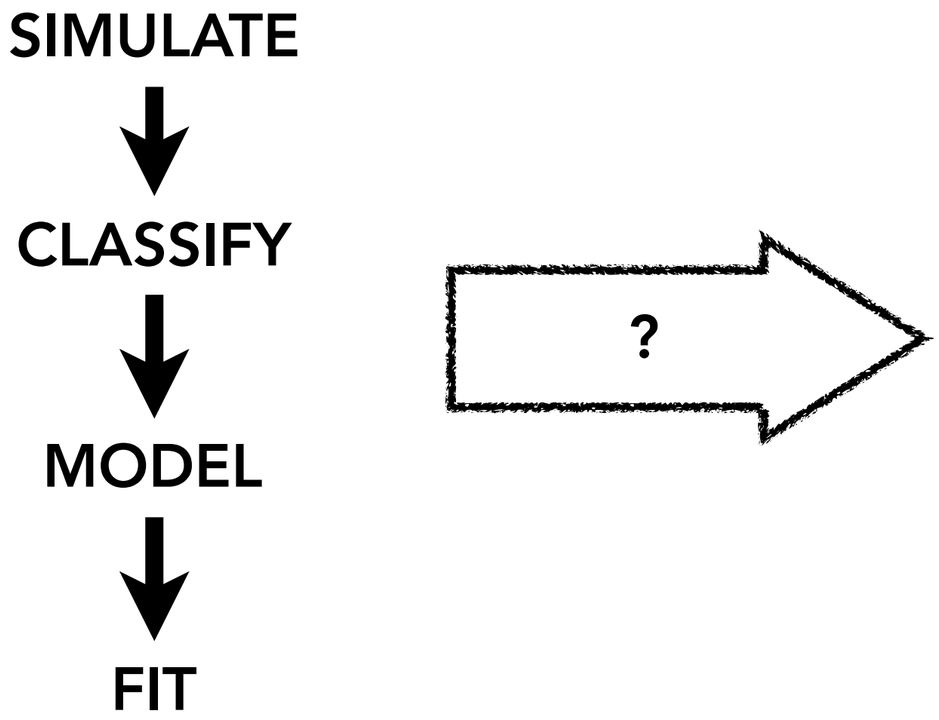
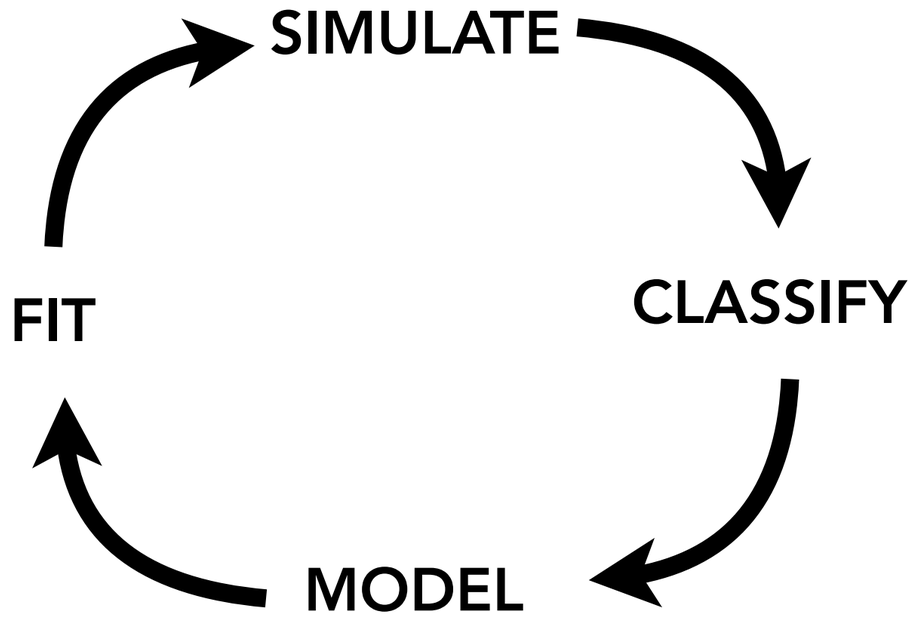
Active learning for "intelligent" simulation





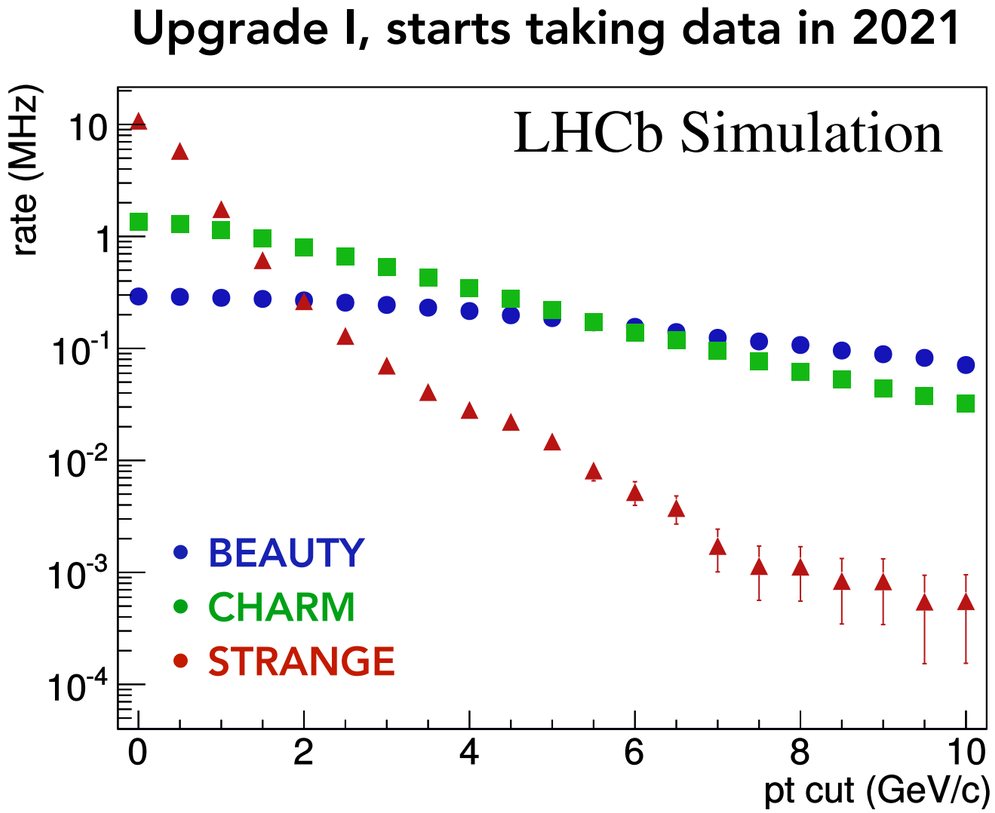
Vladimir V. Gligorov, CNRS/LPNHE

IN2P3 ML meeting, Lyon, 29.03.2018

WHAT PROBLEM DO WE NEED TO SOLVE?

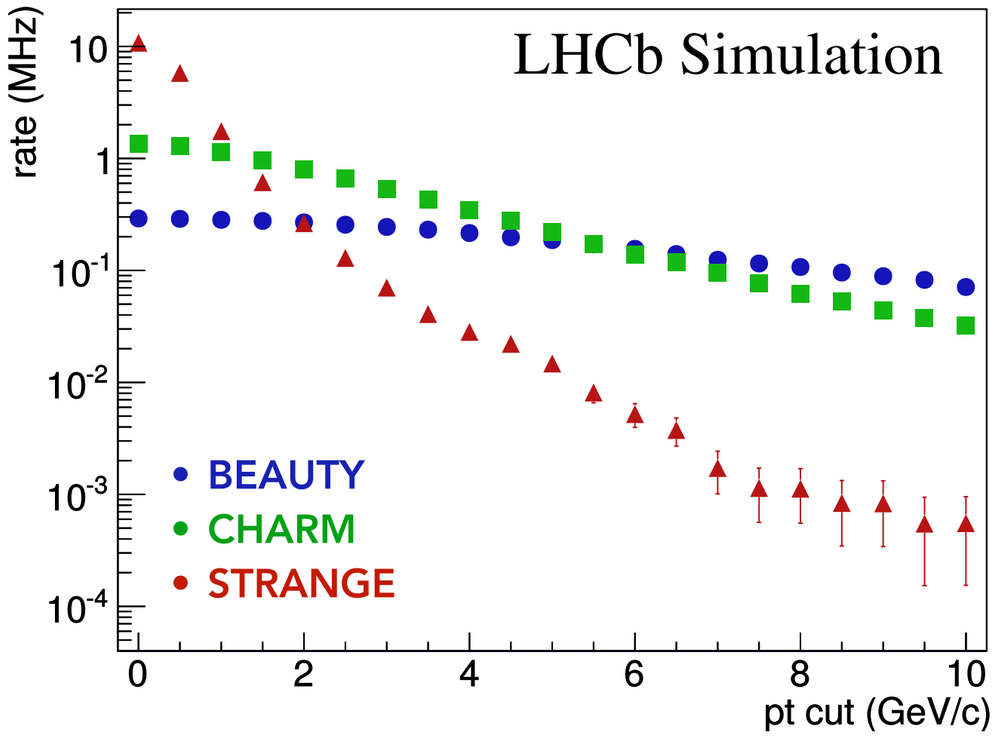
Note : I work on LHCb so will use it to illustrate ideas behind this talk, but many of these points apply equally well to ATLAS, CMS, and other experiments

LHC and HL-LHC data rates

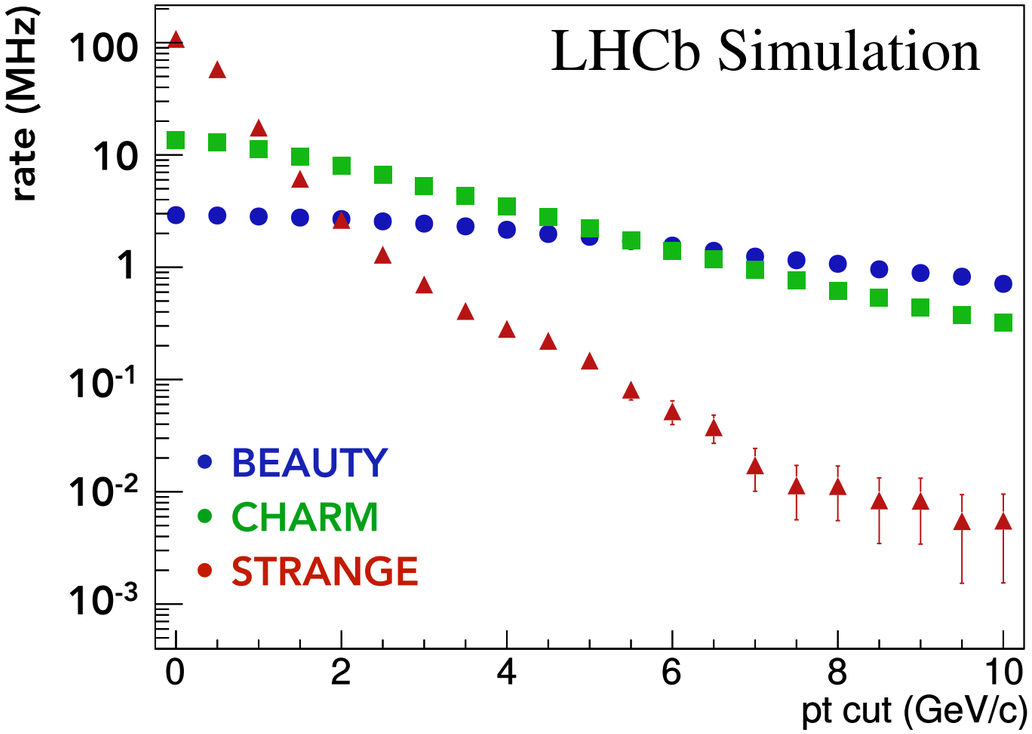


LHC and HL-LHC data rates

Upgrade I, starts taking data in 2021



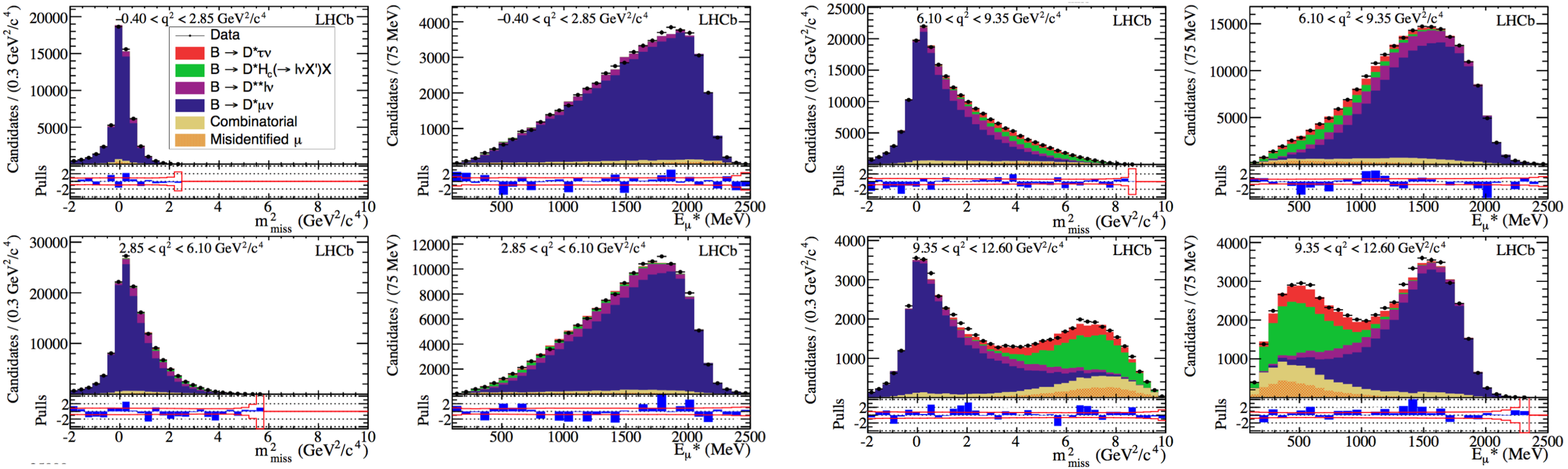
Upgrade II, starts taking data in 2032?



Signal rate is enormous : take data for O(10^6) seconds per year, so can easily accumulate tens of billions of signals even after a full selection!

Impact of simulation on physics

LHCb-PAPER-2015-025



Many of our measurements which are sensitive to New Physics require multidimensional fits to separate signal from background. Ultimate precision is driven by the *simulation* uncertainty on the signal and background shapes.

Impact of simulation on physics

LHCb-PAPER-2015-025

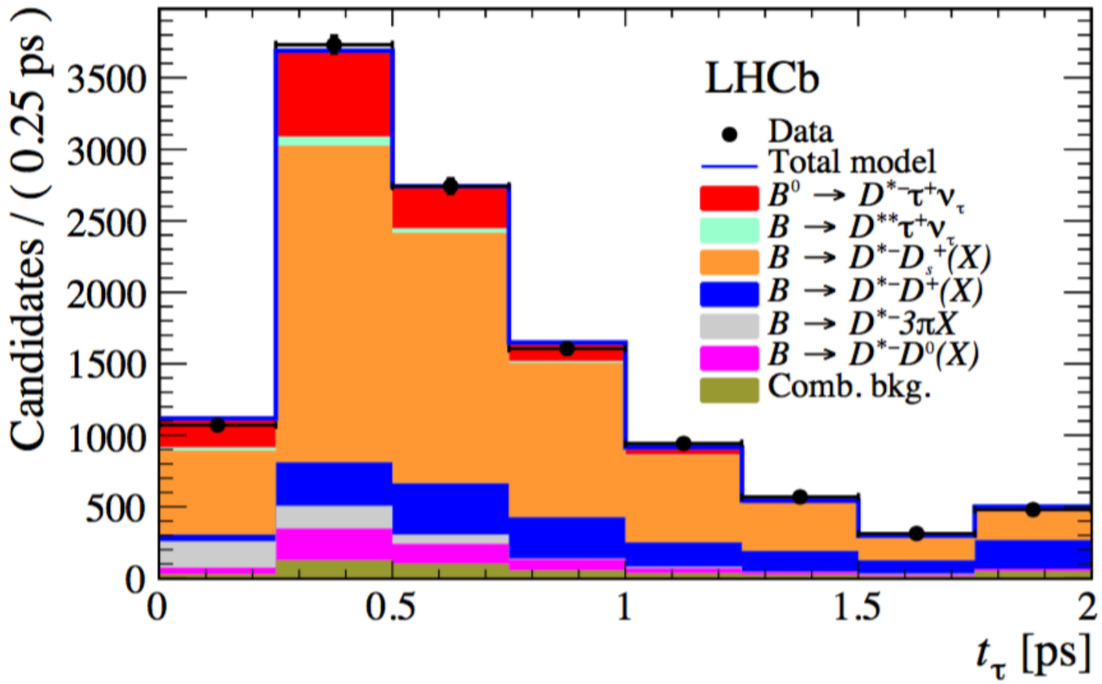
Table 1: Systematic uncertainties in the extraction of $\mathcal{R}(D^*)$.

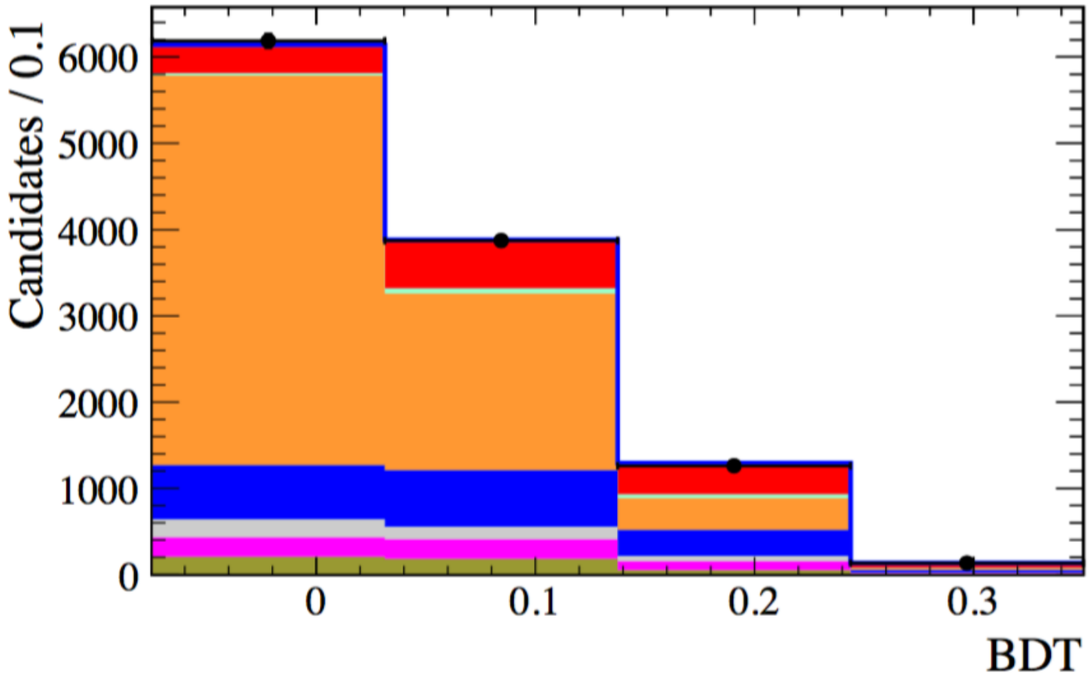
Model uncertainties	**Absolute size ($\times 10^{-2}$)**
Simulated sample size	2.0
Misidentified μ template shape	1.6
$\overline{B}^0 \to D^{*+}(\tau^-/\mu^-)\overline{\nu}$ form factors	0.6
$\overline{B} \to D^{*+}H_c(\to \mu\nu X')X$ shape corrections	0.5
$\mathcal{B}(\overline{B} \to D^{**}\tau^-\overline{\nu}_\tau)/\mathcal{B}(\overline{B} \to D^{**}\mu^-\overline{\nu}_\mu)$	0.5
$\overline{B} \to D^{**}(\to D^*\pi\pi)\mu\nu$ shape corrections	0.4
Corrections to simulation	0.4
Combinatorial background shape	0.3
$\overline{B} \to D^{**}(\to D^{*+}\pi)\mu^-\overline{\nu}_\mu$ form factors	0.3
$\overline{B} \to D^{*+}(D_s \to \tau\nu)X$ fraction	0.1
Total model uncertainty	**2.8**
Normalization uncertainties	**Absolute size ($\times 10^{-2}$)**
Simulated sample size	0.6
Hardware trigger efficiency	0.6
Particle identification efficiencies	0.3
Form-factors	0.2
$\mathcal{B}(\tau^- \to \mu^-\overline{\nu}_\mu\nu_\tau)$	< 0.1
Total normalization uncertainty	**0.9**
Total systematic uncertainty	**3.0**

Many of our measurements which are sensitive to New Physics require multidimensional fits to separate signal from background. Ultimate precision is driven by the *simulation* uncertainty on the signal and background shapes.

Impact of simulation on physics

LHCb-PAPER-2017-027





In other measurements can have backgrounds with >100x the signal rate which must be simulated very precisely. Because the efficiency to select these background events is so small (else you couldn't make the measurement) you have to simulate enormous amounts to understand shapes of surviving events.

Simulation stages and cost

Generate “true” collision and products
Cost : essentially 0

Simulate passage of events through detector
Cost : 1-100x data reconstruction

Reconstruct event
Cost : 1x data reconstruction

Often dominated by simulating passage of particles through material, for some analyses reconstruction equally important. Ignore disk space for now.

PROPOSED SOLUTIONS TO DATE

Partial detector simulation

Solution : simulate only part of the detector, e.g. the tracker

Addresses : both the cost of simulating particle passage through material, and the reconstruction cost. Amount saved depends on which detectors are excluded.

Shortcomings : cannot simulate the trigger, which requires the full detector information to process every event. Trigger emulation or data-driven corrections can be tricky depending on analysis.

Fast or parametric simulation

Solution : do not simulate passage of particles through material or reconstruct them, but rather generate reconstructed objects based on a parametarized smearing of the truth-level information.

Addresses : both the cost of simulating particle passage through material, and the reconstruction cost, reduces both to ~0.

Shortcomings : tuning the parametric simulation to describe the core of the distribution is proven to be doable, getting the tails and the correlations right is hard. No proof yet that this can be used for precision measurements without introducing systematics.

Reuse of underlying event

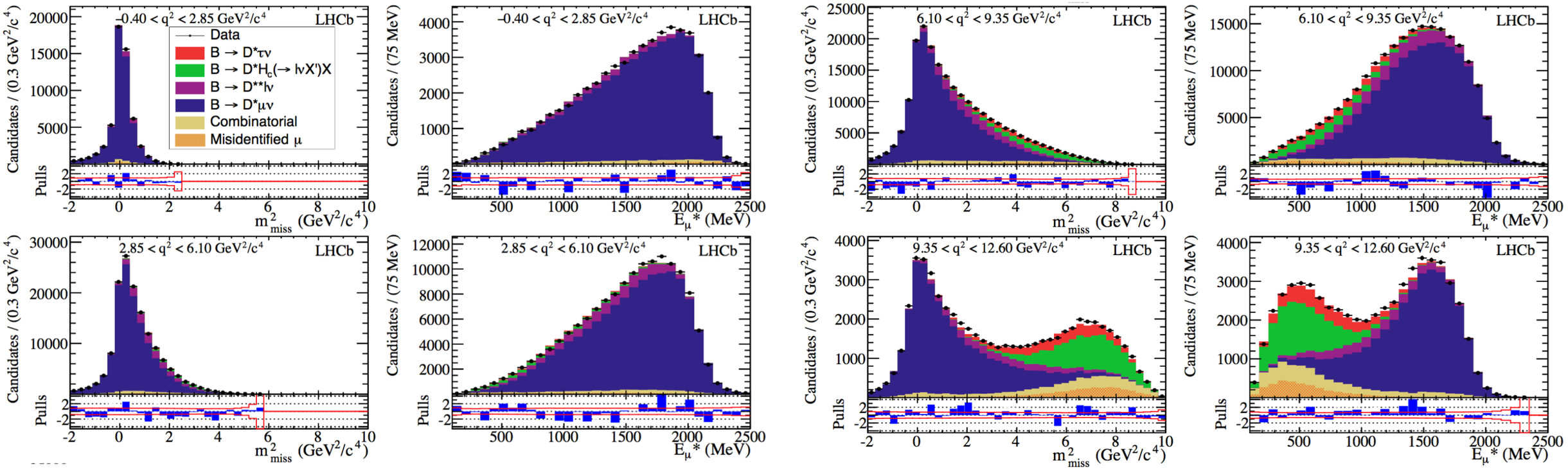
Solution : most of each event is not the interesting signal but rather pp collision byproducts, so simulate the byproducts once and reuse them while “ReDecay”-ing the signal.

Addresses : reduces the cost of simulating particle passage through material by a factor 10-100.

Shortcomings : introduces correlations between the signal properties for ReDecay-ed events, if these are large lose statistical power of signals and must generate more simulation thus reducing the cost saving.

ML FOR INTELLIGENT, NOT ONLY FAST SIM?

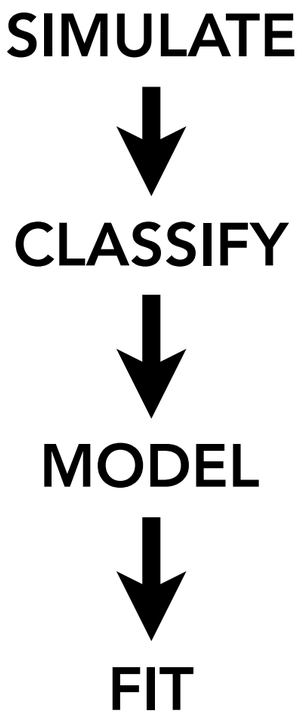
Basic idea



Q : Where do you have to know the signal shape precisely from simulation?

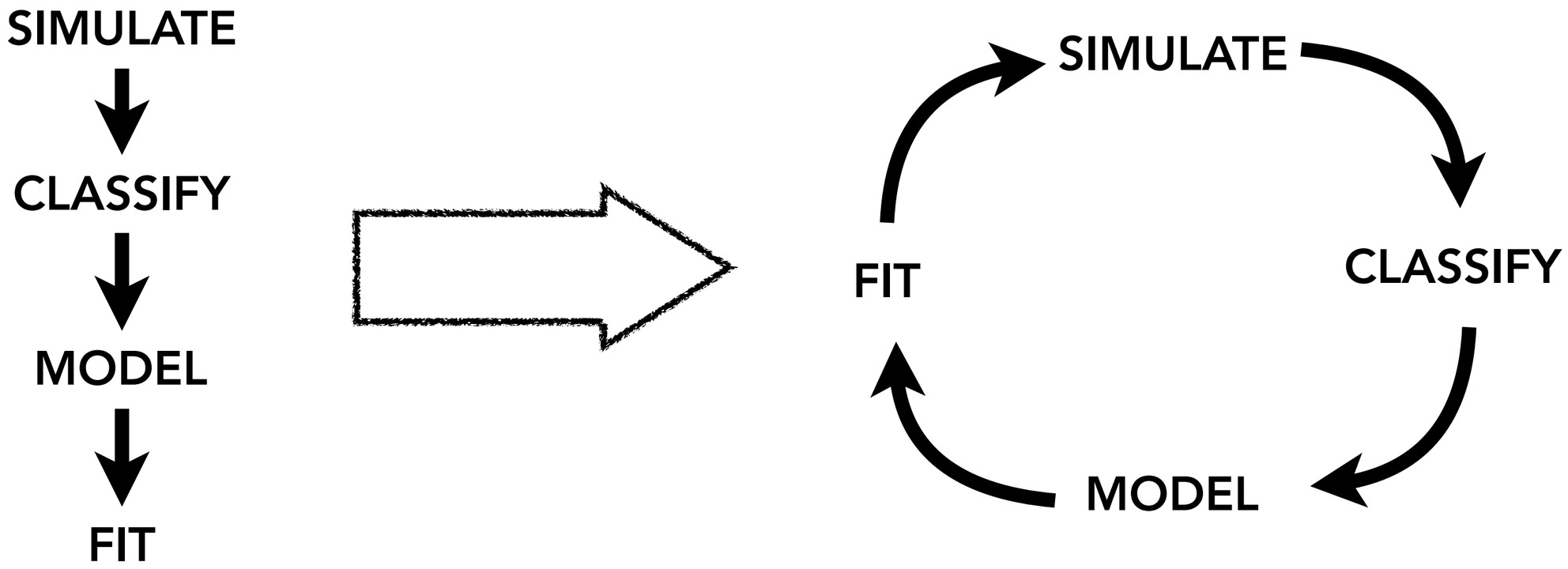
A : Only in the regions where it overlaps with the background.

Use classifier to drive generator?



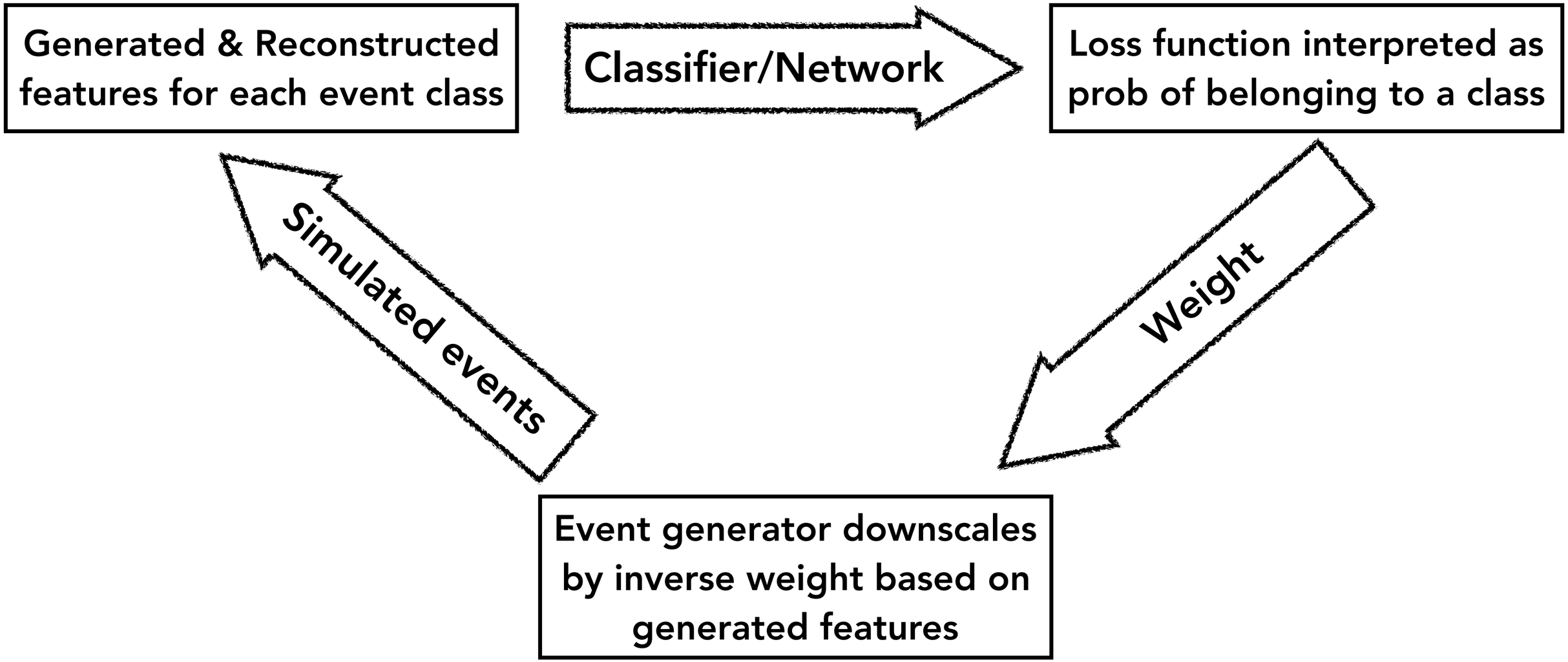
Typically seen as a linear cascade, you simulate once with some parameters, train a classifier to separate signal and background, model what passes the classifier and use these models to fit your data.

Use classifier to drive generator?



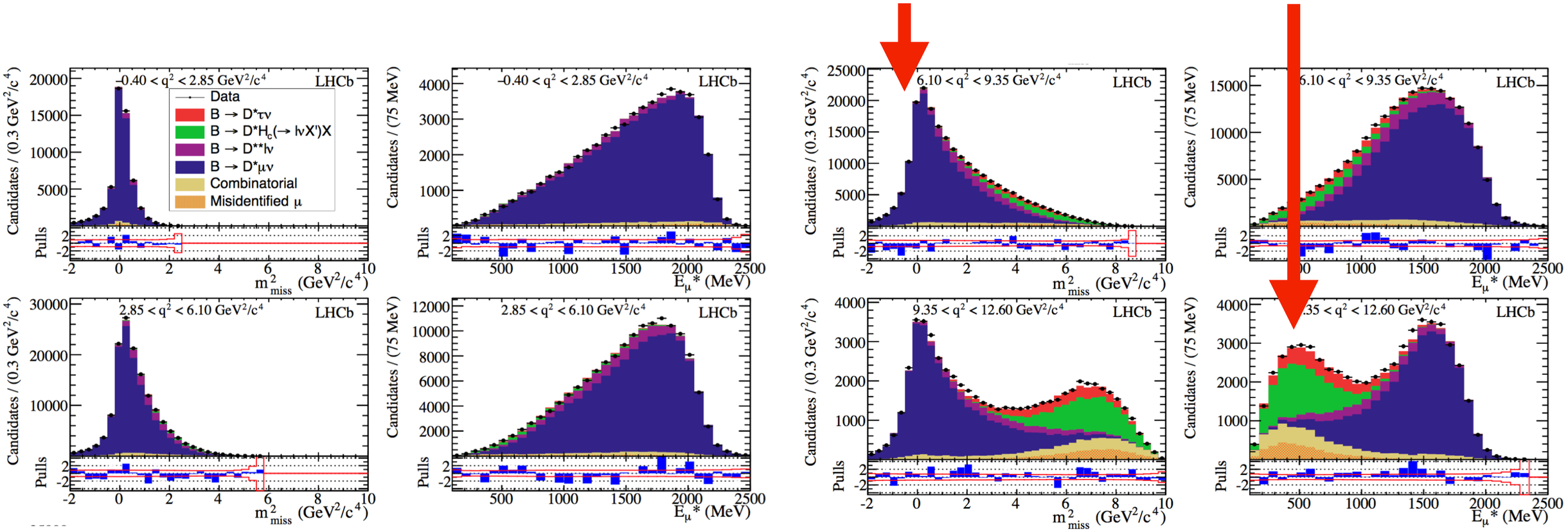
What if instead we treated this as a loop? Simulate a small amount of signal and background, teach classifier the truth-level properties of the events where signal and backgrounds overlap, then use this to weight the generation.

Input/output



Application to physics

Trade uncertainty in background region for precision in signal region?



Coming back to our physics example, you would end up with templates which were relatively less precise in regions dominated by background, and more precise in regions dominated by signal. Would this really affect measurement?

Potential advantages

Since the classifier weights events at generator level, it is trivial to calculate its efficiency, you can always make a generator-level sample of infinite size for that purpose.

Important difference with traditional generator level *cuts* : weight events by estimated classification error, do not simply cut. Can then be used even in situations where truth-level quantities do not map in a simple way onto reconstructed/fit quantities.

After the generator step, still run the best simulation of your detector possible; by definition minimizes systematic uncertainties associated with data-simulation differences.

Conclusion

Many new physics sensitive analyses are and will continue to be systematics limited by the cost of simulating events

Many different approaches being tried to address this problem, mainly based on performing either partial or parametric simulations

Propose to try a complementary approach : integrate classification of signals and backgrounds into the simulation chain and use ML to learn which kinds of events to simulate more frequently in order to achieve maximum sensitivity to the physics observables with the minimum number of simulated events.

As you can see, this is just an idea for now… assuming you don’t tell me it is very stupid :) we will try it out in the next months…

BACKUP