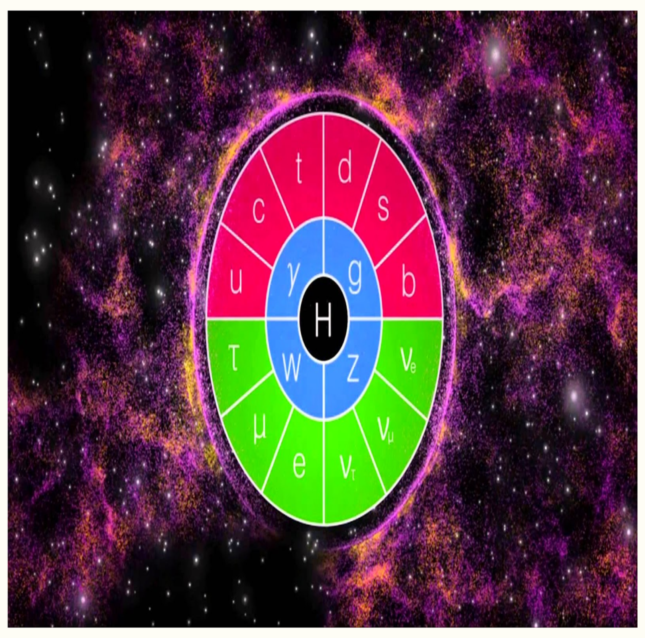
# $H \to \tau\tau$ Mass Estimation Using Boosted Regression Trees


**Sina Bahrasemani, Dugan O'Neil, Quentin Buat**

Simon Fraser University; Atlas Collaboration, CERN

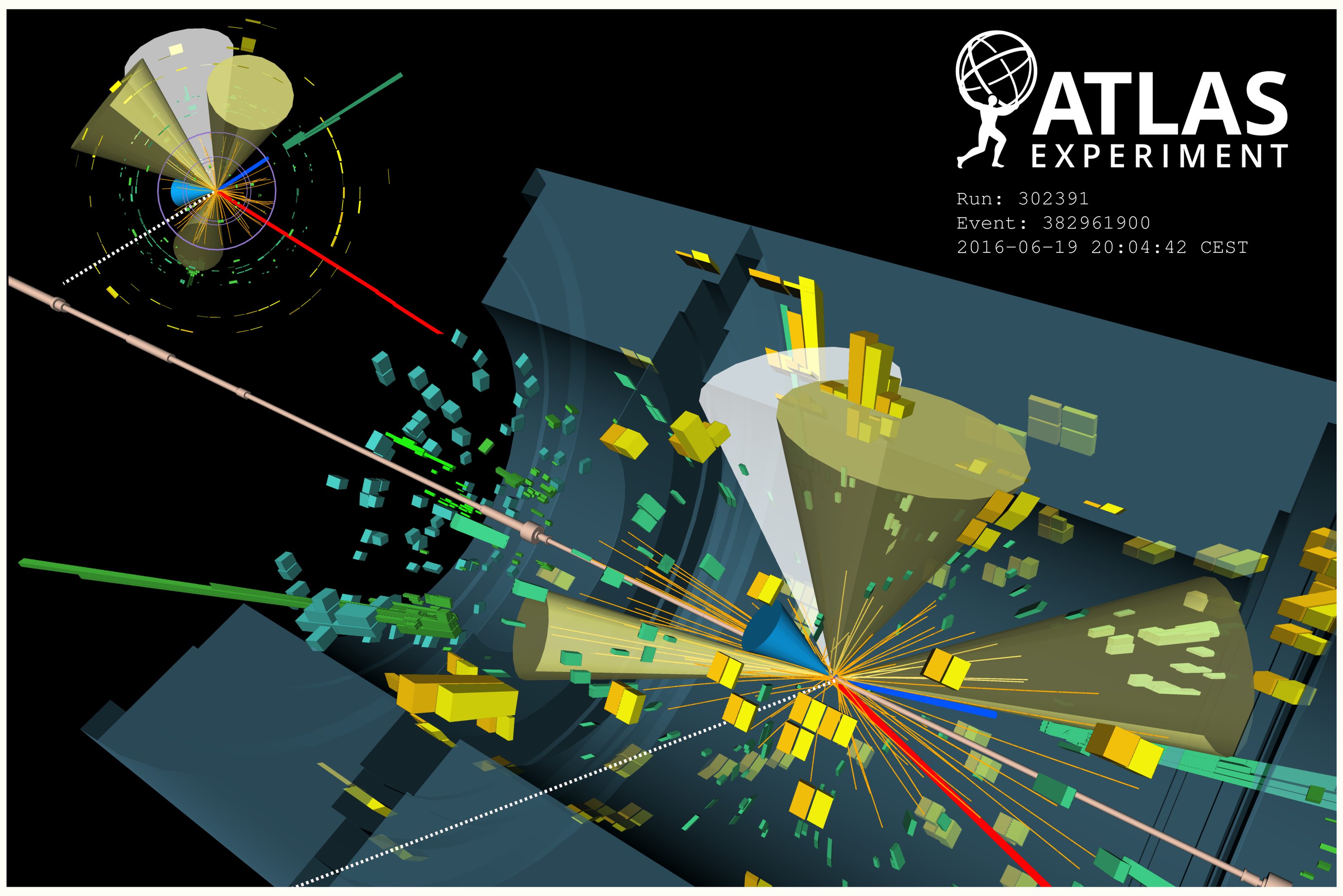
sbahrase@sfu.ca doniel@sfu.ca qbuat@sfu.ca


## The God Particle

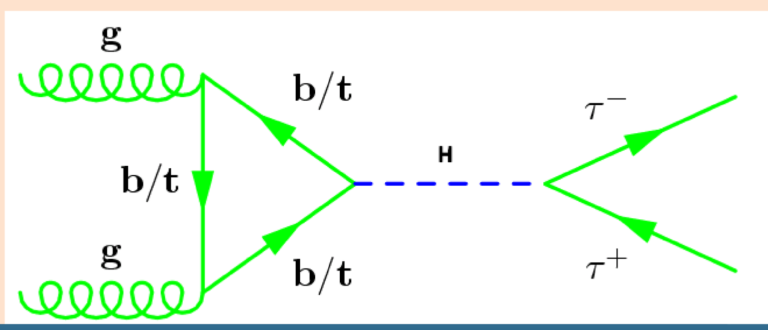



- **Matter particles**: the building blocks of matter occur in two basic types called quarks and leptons.
- **Forces & carrier particles:** particles of matter transfer discrete amounts of energy by exchanging bosons with each other.
- **Higgs particle is the Golden Glue of Standard Model of Particle Physics**
- **Higgs particle gives mass to elementery particles**

## Higgs Physics at LHC




- **LHC is colliding proton-proton beams with $E_{com} = 13TeV$**
- **LHC produces $\approx 6 \times 10^8$ inelastic collisions per second**
- **From $10^9$ pp-collisions $\approx ONE$ Higgs Boson produced**
- **$H \to \tau\tau$ has the highest B.R. among leptons $\Rightarrow$ key to understand Higgs couplings to leptons.**
- **Mass is one the best handles to extract signals from enormuos backgrounds**
- **$H \to \tau\tau$ Difficult final state: 2 to 6 neutrinos, depending on the decay sub-channel**

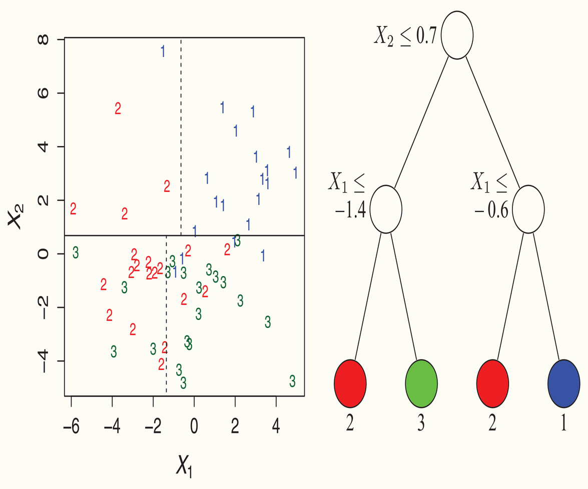



## $M_{H\to\tau\tau}$ Estimators

- **Transverse Mass:** $M_T = \sqrt{m_1^2 + m_2^2 + 2(E_{T,1}E_{T,2} - \vec{p}_{T,1}.\vec{p}_{T,2})}$
  - ▷ Simplest calculation, well-defined for every events
  - ▷ Very limited, not using all kinematics
- **Collinear Mass:**
  - ▷ Assume neutrinos are emitted in the same direction as visible decays $\to$ reduce the phase space to more boosted events
  - ▷ Better resolution than the transverse mass
- **Missing Mass Calculator:**
  - ▷ Use kinematics of the events from simulation to 'guess' the most likely $M_{\tau\tau}$ value event-by-event
  - ▷ Resolvingan under-constrained system of $6to8$ unknowns $\to$ scanning the phase space of key parameters
  - ▷ Scan $E_{T,x}, E_{T,y}, \phi_1, \phi_2$ & derive a weight function
  - ▷ Use this weight to evaluate $M_{\tau\tau}$ for each event

## Regression Trees

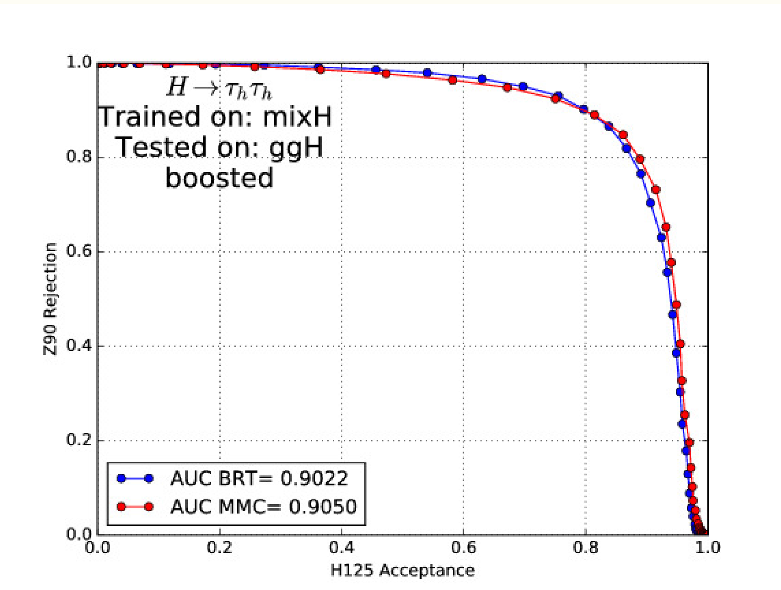
- **Linear regressions are global models:** a single predictive formula which holds all over dataset range.
- **Correlations among the variables makes the prediction very complicated & The number of the free parameters grows very fast.** $\Longrightarrow$ nonlinear regression
- **An alternative to the nonlinear regression is to partition the sample to smaller sub-samples recursively until get chunks simple enough to model.**
- **Prediction trees use the tree to represent the recursive partition.**
- **Regression Trees can be quite sensitive to the statistical fluctuation in the input samples, ...**
- **It's possible to use an ensemble of weak regressor (decision trees) and combine them to get a stronger regressor.**

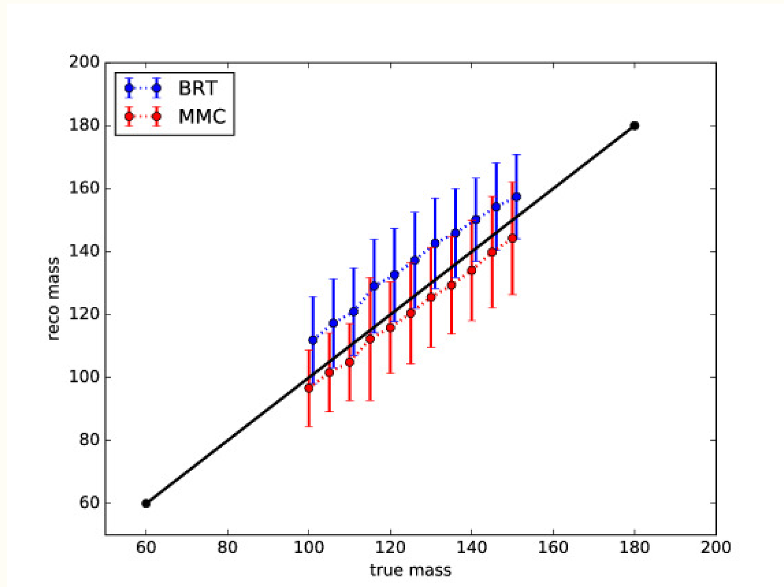



## Trainig Regression Tree for $M_{H\tau\tau}$

- **Goal:**
  - ▷ Predicting the invariant mass of ditau system using the visible tau decay products kinematics.
- **Training sample:**
  - ▷ $pp \to H[X] \to \tau\tau[X']$ Monte Carlo (full ATLAS detector with 5GeV spacing) $60GeV \le M_H \le 200GeV$ ($\approx 20k$ events per mass point)
- **Training Algorithm:**
  - ▷ **Machine Learning Software:** Scikit-Learn
  - ▷ **Splitting nodes condition:** Minimum square error
  - ▷ **Stoping condition:** Minimum leaf size $> 0.02\%$ of the sample size

## BRT $M_{H\to\tau\tau}Calculator$




- BRT mass calculator is quite competitive with MMC mass calculator
- Predicted mass from BRT and MMC is shifted by a constant.
- The mass distribution resolution is almost the same
- BRT and MMC have almost the same power in separating singal from background

For rare processes, like $H \to \tau\tau$ separating signal from enormous background is absolutely crucial. Also invariant mass distribution can be directly used to search for new resonances

## Conclusion

- ☑ **BRT has reached the physics performance of the most commonly used di-tau mass calculator; the MMC**
- ☑ **BRT is roughly $3000x$ faster than MMC.**
  - ▷ Mass reconstruction is the single slowest part of doing the analysis of ATLAS $H \to \tau\tau$ events.
  - ▷ re-tuning the MMC for changing experimental conditions is very slow (months) while, retraining the BRT is super-fast (minutes)
  - ▷ Being so fast $\Rightarrow$ There is the potential in future to exceed the physics performance