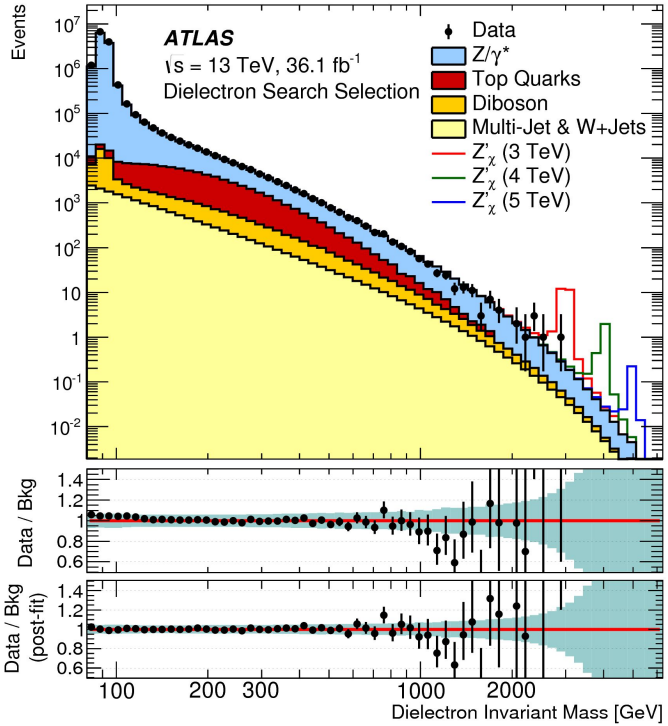
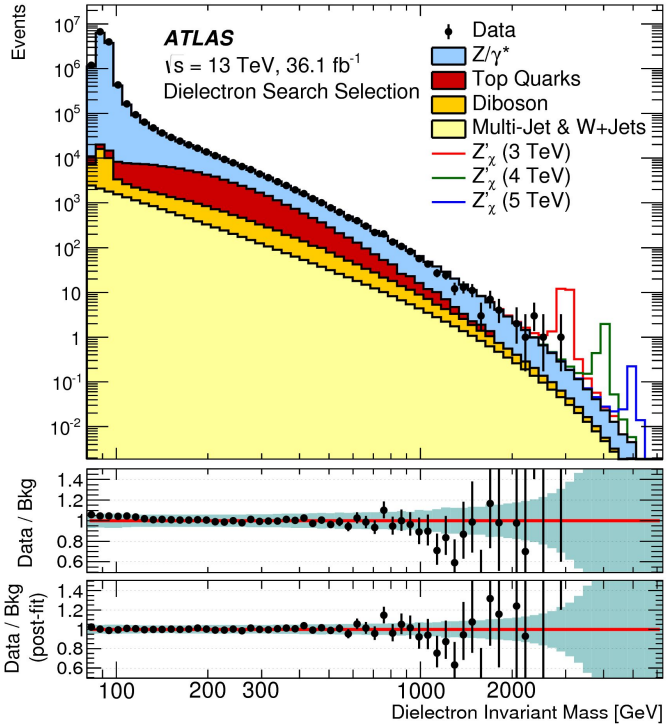
Data analysis in a nutshell

Background vs. Signal



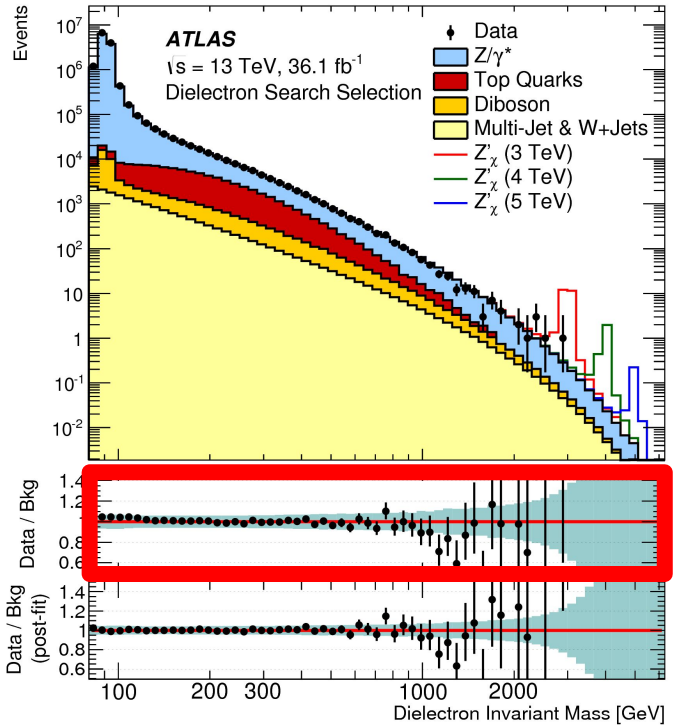
- Signal: process under study
- Background: processes yielding the same final state as signal

Background estimation



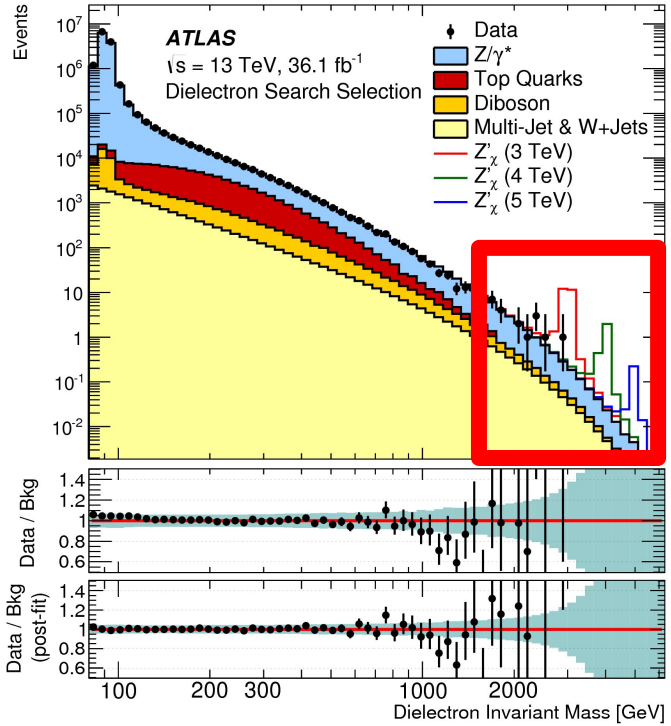
- With MC
  - generate events (-> truth-level)
  - simulate detector response (-> reco-level)
- Data-driven
  - E.g. extrapolate a fit function
  - Fake background estimation

Systematic uncertainties



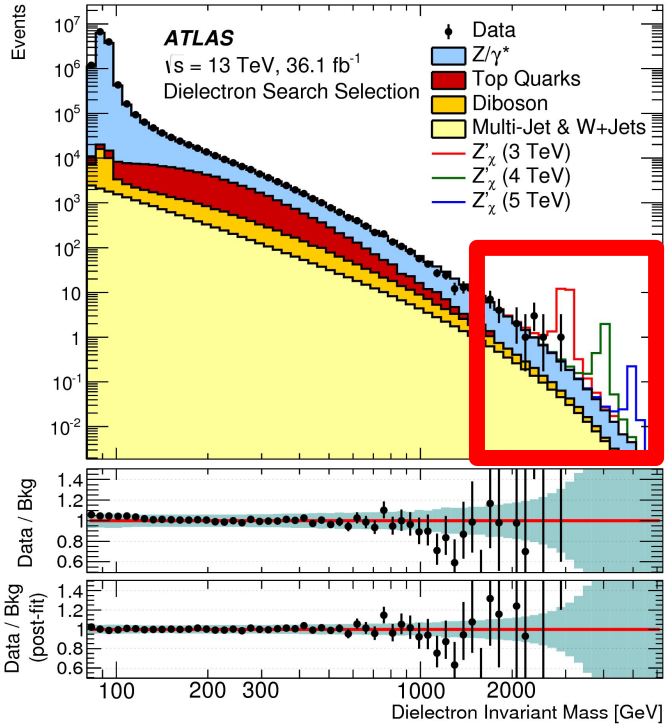
- Due to imperfect description of background and signal predictions
  - MC: differences between the simulated and the real detector
  - Data-driven: unknown parametrization
  - Theory: truncated series, model uncertainties

Sensitive variable(s), Signal Region

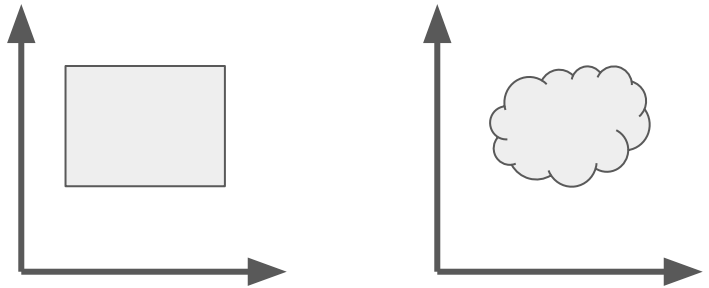


- Phase space: space in which “data live”
- Goal: find a region of phase space in which signal dominates over background
- Invariant mass

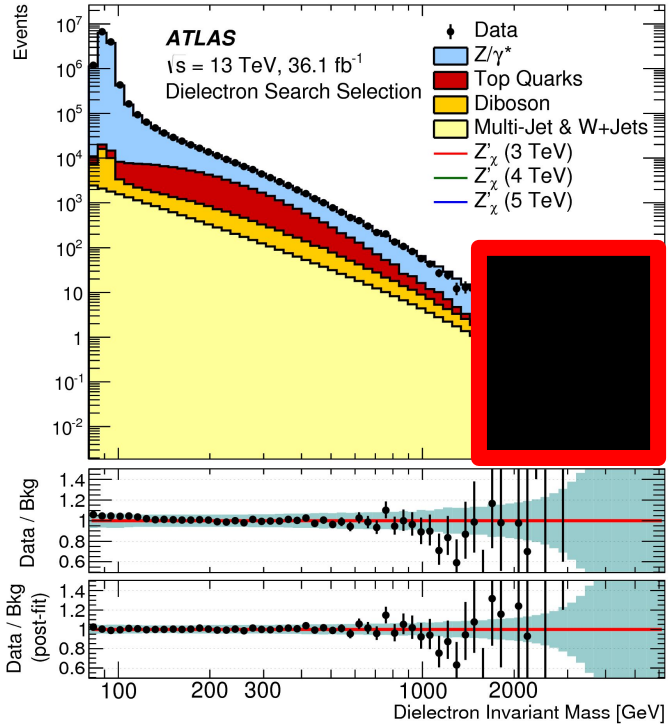
Signal Region selection



- Cuts:
  - simplest selection criteria
  - select a hyper-cuboid
- Artificial intelligence
  - a.k.a. multivariate analysis (MVA)
  - select a region of any shape
  - Neural networks, Boosted Decision Tree…

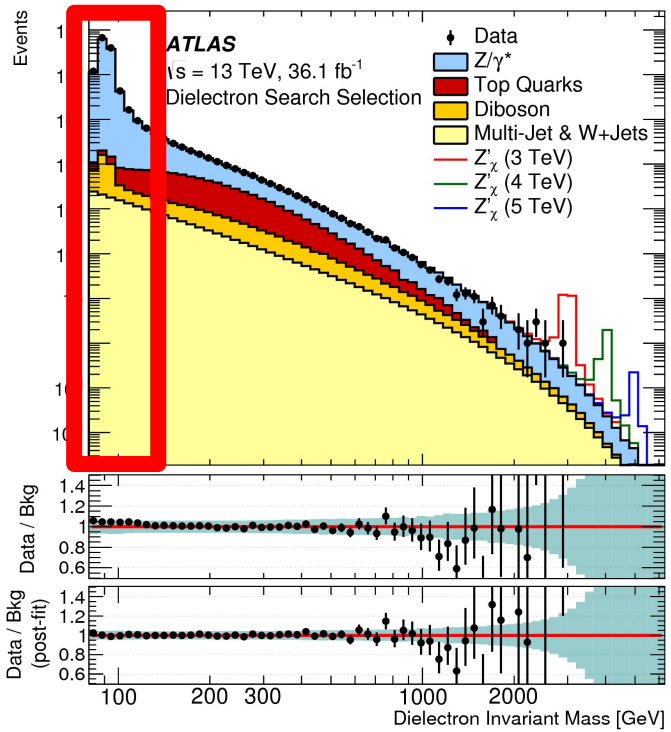


Blinded analysis



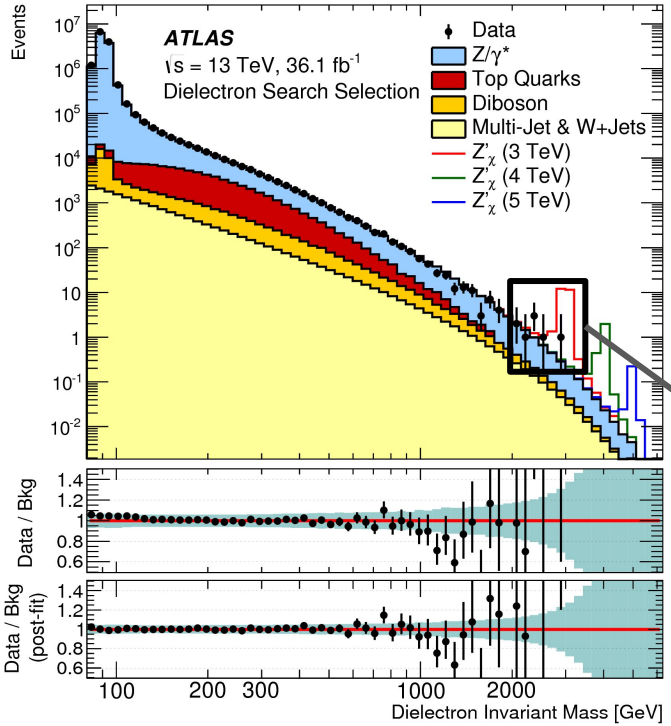
- Don’t look at data in the SR until the whole analysis strategy is defined!
- Helps avoid analyzer’s bias
  - E.g. adjust bin widths to create a peak

Control Region



- Region in which we don't expect signal events
- Check quality of background estimation with comparison to data in CRs
- Adjust normalization of background estimation so that it matches data in CRs
- Several CRs per analysis
  - Z CR, top CR, W CR...

Statistical fluctuations

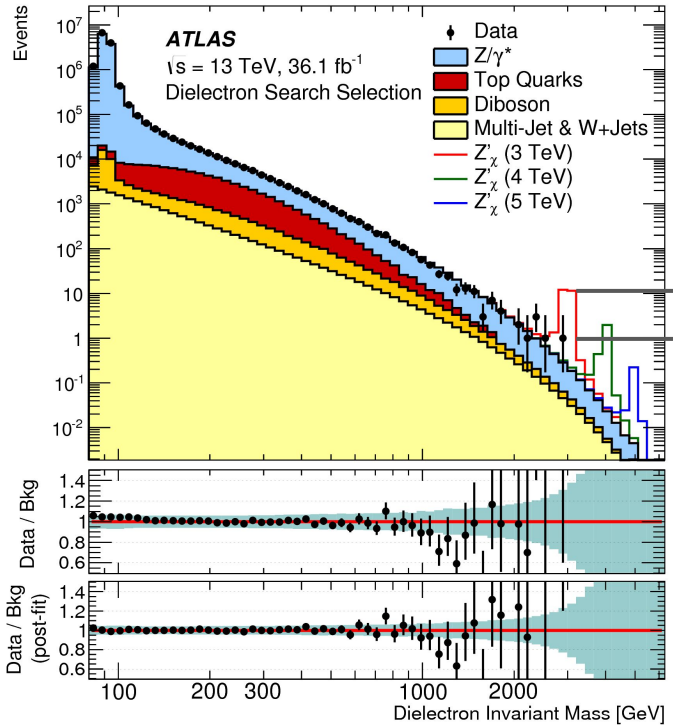


- Data is random
  - number of data events in a bin is a random variable drawn from a Poisson distribution

$$f(x) = \frac{\lambda^x}{x!} e^{-\lambda}$$

Excess of data events over the background prediction?

Extraction of Parameter(s) of Interest: fit

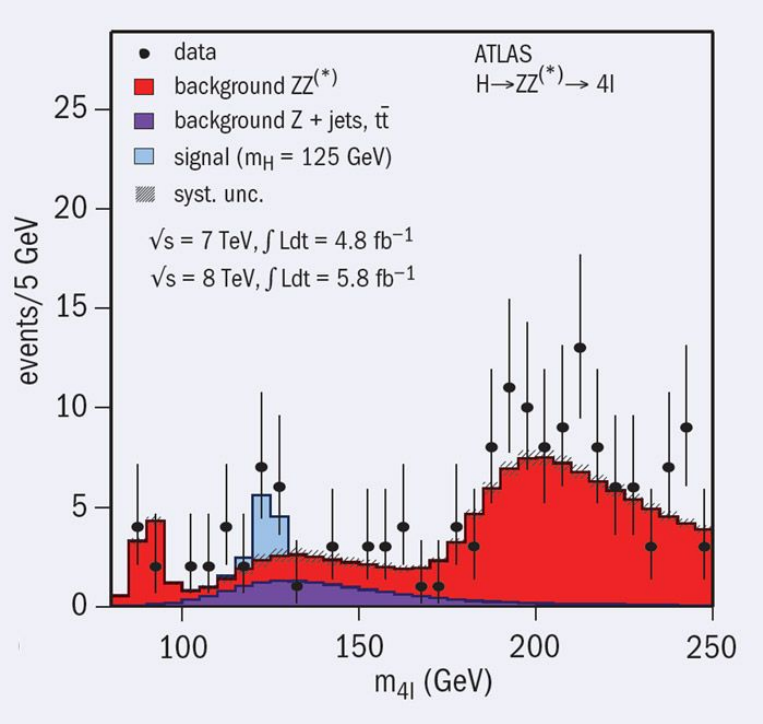


μ=1

μ=1/10

- Binned template fit
  - Most often: extract normalization of the signal template
  - find μ such that B + μS matches data the best
- Unbinned fit
  - Parametrize the probability distribution function for data
  - Background and signal predictions used “just” to find a good parametrization

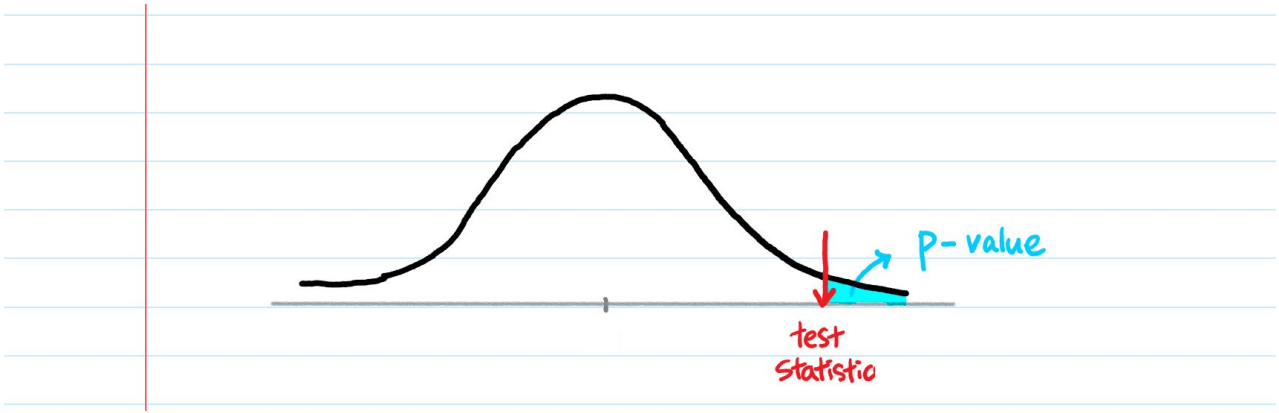
Hypothesis test



- The most important question: Does the signal exist?
  - Data will never tell us!
  - Maximum we can do is to try to rule out a hypothesis.
- Rephrase the question: Is data compatible with the hypothesis that signal doesn't exist?
  - Very incompatible in the Figure!
  - Hard to judge by eye

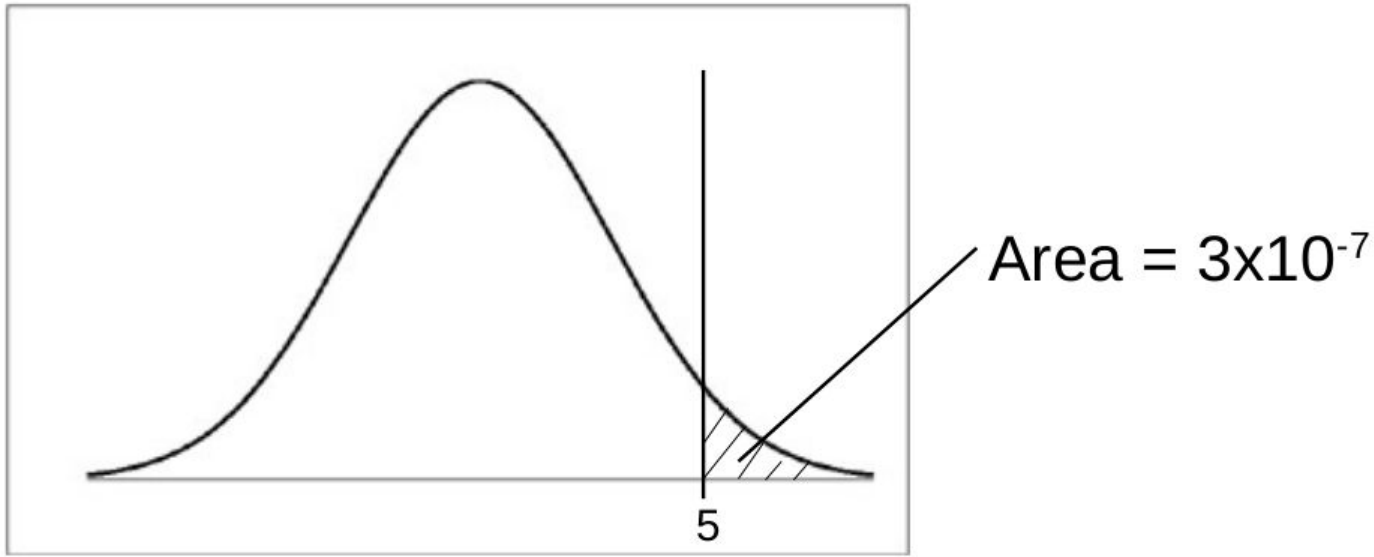
Test statistic, p-value

- Test statistic: any function of data (that quantifies “distance” of data and the tested hypothesis prediction)
  - E.g. number of data events in bins where signal is supposed to appear
  - In practice: profile likelihood ratio



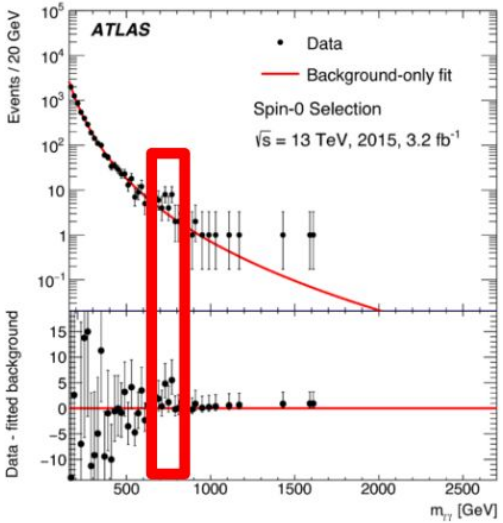
Presentation of probability

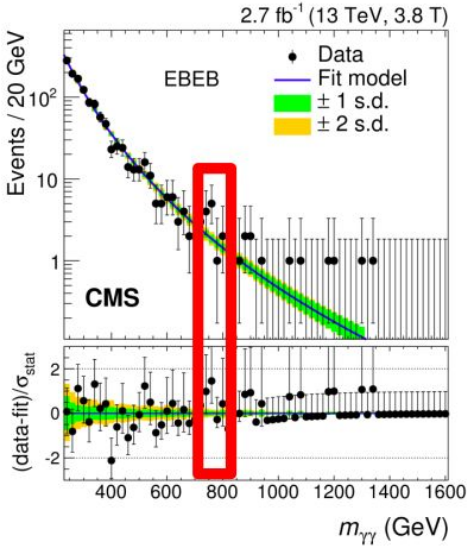
- With the use of the normal distribution, Gaus(μ=0, σ=1)



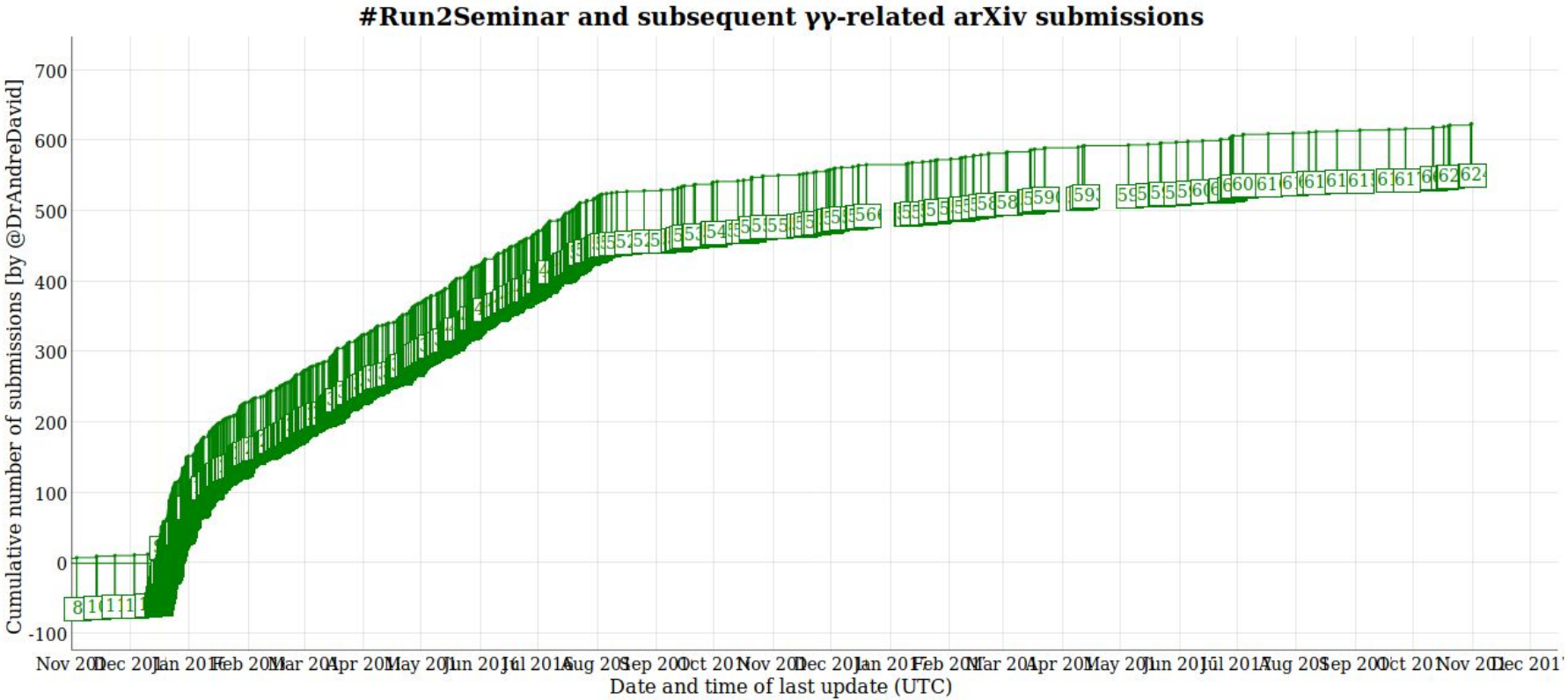
5σ is useful! Example of a false alarm

- Observed significance was 3.9σ (ATLAS) and 3.4σ (CMS)!
- Excesses gone when more data taken



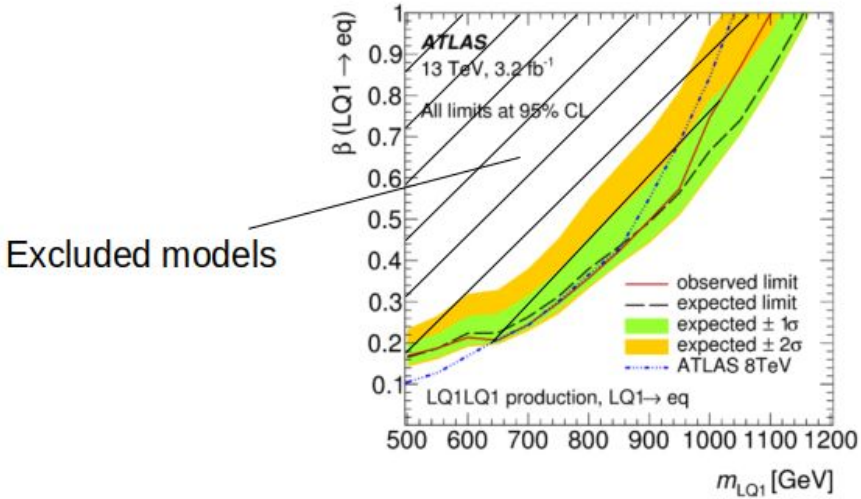


Number of papers explaining the excesses on arXiv: 624



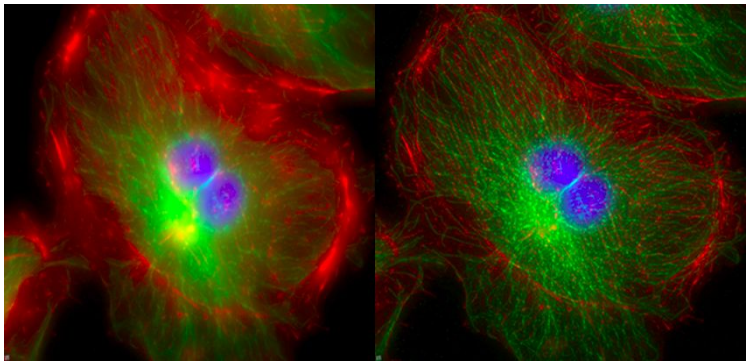
Limits

- If there’s no excess of data over the background prediction
- Limit setting: test of the hypothesis “signal exists” for models defined by some parameter values
- Exclusion of a model: probability that the signal exists is < 0.05



Different approach to measurements: Unfolding

- If we don’t have any specific model prediction - we don’t fit any specific model parameters
- We just want to remove from our spectra distortion due to the detector response
  - resolution effects
  - efficiency effects



Detector effects

- Unfolding: restore the blue curve based on the red one
- Simplest approach: invert the response matrix
  - in practice: more sophisticated methods used