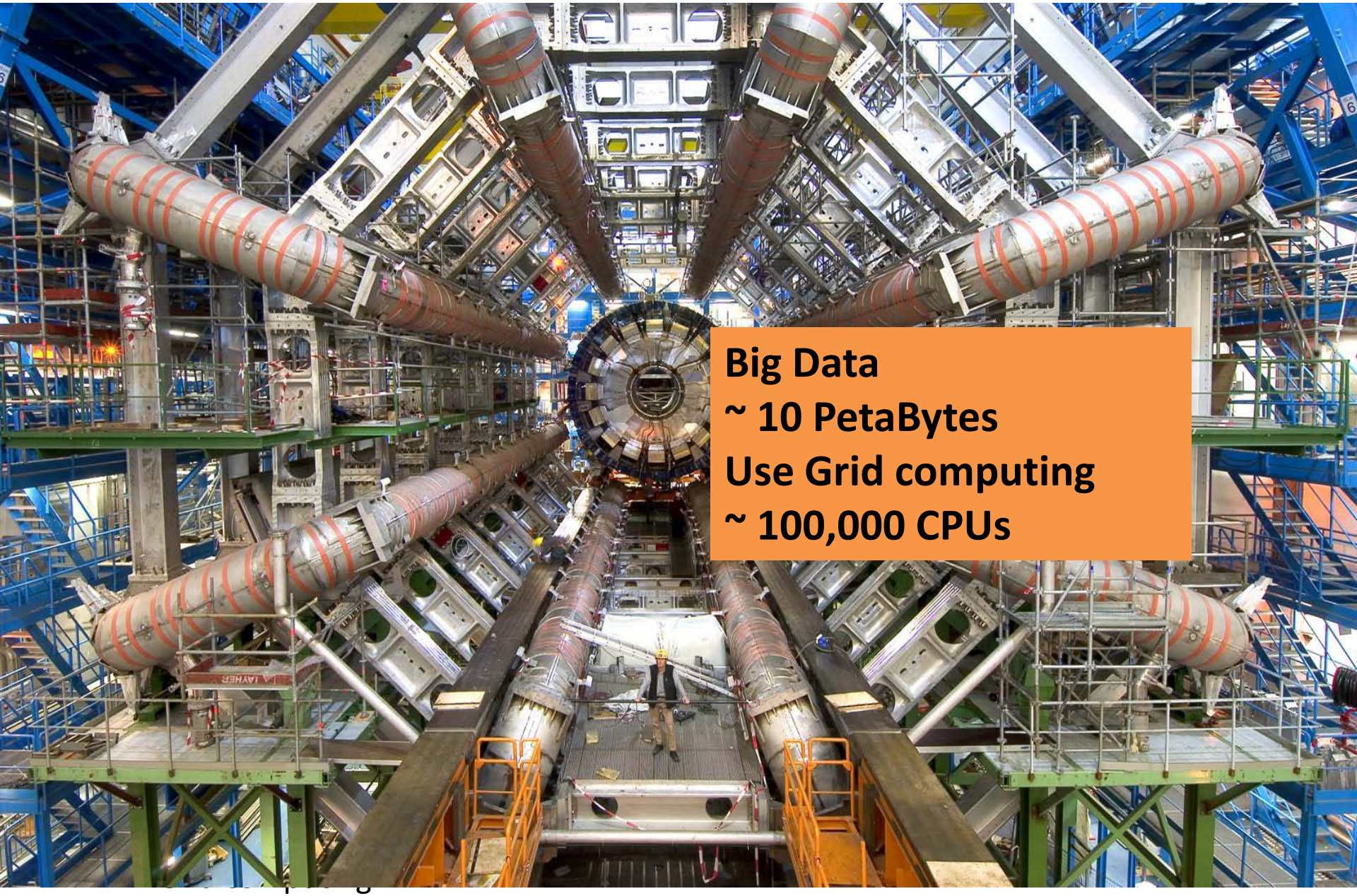
Big Data and Particle Physics

- **Example of big data analysis: Higgs boson discovery @LHC**
  - **How do we ensure reliability of the results?**
- **Possible applications to biomedical sciences?**

Tony Weidberg (Particle Physics, Oxford University)

Talk Outline

- Case Study from particle physics: Higgs boson discovery
  - Statistical procedures used
  - Organisational procedures for checking results
- Possible applications to biomedical sciences
  - Statistical procedures
  - Collaboration


Big Data
~ 10 PetaBytes
Use Grid computing
~ 100,000 CPUs

Case study: Higgs boson discovery

- Scientific procedures
  - Statistical
  - Blind analysis
  - 5 sigma threshold for discovery
- Levels of checking
  - Low level x-checks
  - Sub-group
  - Working groups
  - Editorial Board
  - (Several further levels)
  - Collaboration
  - Refereed journals
  - Confirmation by another experiment.

Blind Analysis

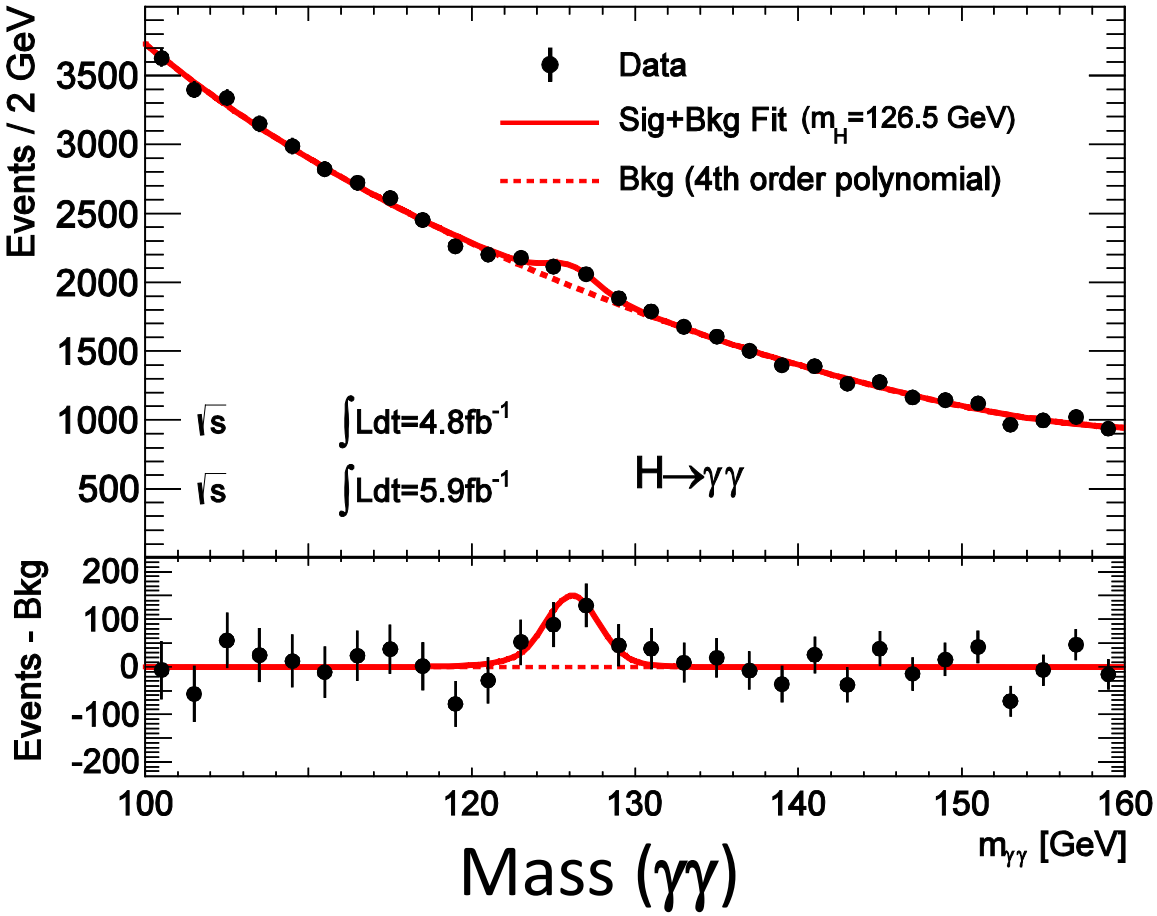
- **Avoid finding spurious signals in very large data sets ➔ use blind analysis:**
  - **Monte Carlo simulations for signals and backgrounds**
  - **Optimise analysis (separation of signal from background) using Monte Carlo samples**
  - **Review analysis and then "open box" and look at data without changing analysis**
  - **Warning: this is a very simplified description!**

Higgs Boson

- **In the Standard Model of particle physics Higgs boson gives mass to other elementary particles.**
- **Use high energy proton-proton collisions to try to produce Higgs bosons.**
- **Reconstruct decay products and use $E=mc^2$**
- **Look for peak in mass spectra at mass of Higgs boson (m_H)**

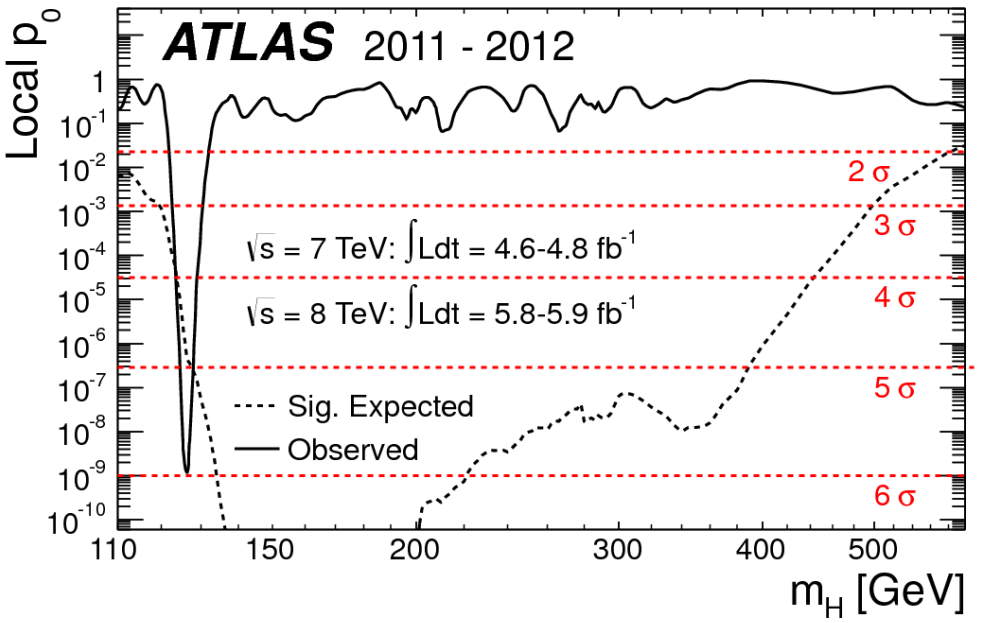
Higgs boson discovery

- **One channel: H→γγ is the "bump" a statistical fluctuation or significant evidence for a Higgs boson?**



Statistical Procedures

- **Frequentist approach:**
  - **Define probability p_0 that if experiment were repeated infinite number of times that we would see a larger discrepancy with the no-signal model than in the actual data set.**
- **Combine all channels**
- **Plot p_0 vs Higgs mass (m_H)**
- **Look Elsewhere Effect**
- **5 σ rule**



Checking Results

- Internal
  - Low level x-checks
  - Sub-group
  - Working groups
  - Editorial Board
  - (Several further levels)
  - Collaboration
- Refereed journals
- Confirmation by another experiment.

Works well because of scientific culture in which everybody is encouraged to give critical feedback

Significance for Biomedical Sciences-1

- **Statistical procedures to maximise reproducibility:**
  - **Blind analysis avoids finding spurious effects ➔ use it.**
  - **Estimate power of proposed experiments, don't do low power experiments!**
  - **95% confidence level to claim an effect seems very low, increase it?**

Significance for Biomedical Sciences-2

- **Cultural and organisational:**
  - **Create healthy scientific culture**
  - **Junior members of teams are encouraged to make critical comments about experiments and analysis.**
    - **Question: do graduate students and post-docs feel they will be rewarded for this or are they afraid it will harm their careers?**

Significance for Biomedical Sciences-3

- **Collaboration:**
  - **Difficult for one team to obtain sufficiently large sample sizes ➔ solution is collaboration**
  - **Easy to arrange collaboration with different universities and labs using web**
    - **Regular video meetings**
    - **Collaboration material on Sharepoint or Twiki or similar**
    - **Much more rigorous internal checking**
    - **In my opinion one good study is much better than a meta-analysis of several "low statistics" studies.**

Significance for Biomedical Sciences-4

- **Checking results**
  - **Verifying or falsifying an existing result is first class science and should be funded properly.**
  - **Which group was first not really important, we want to get reliable results.**

- **I hope some of these suggestions are useful and will provoke questions and discussion …**