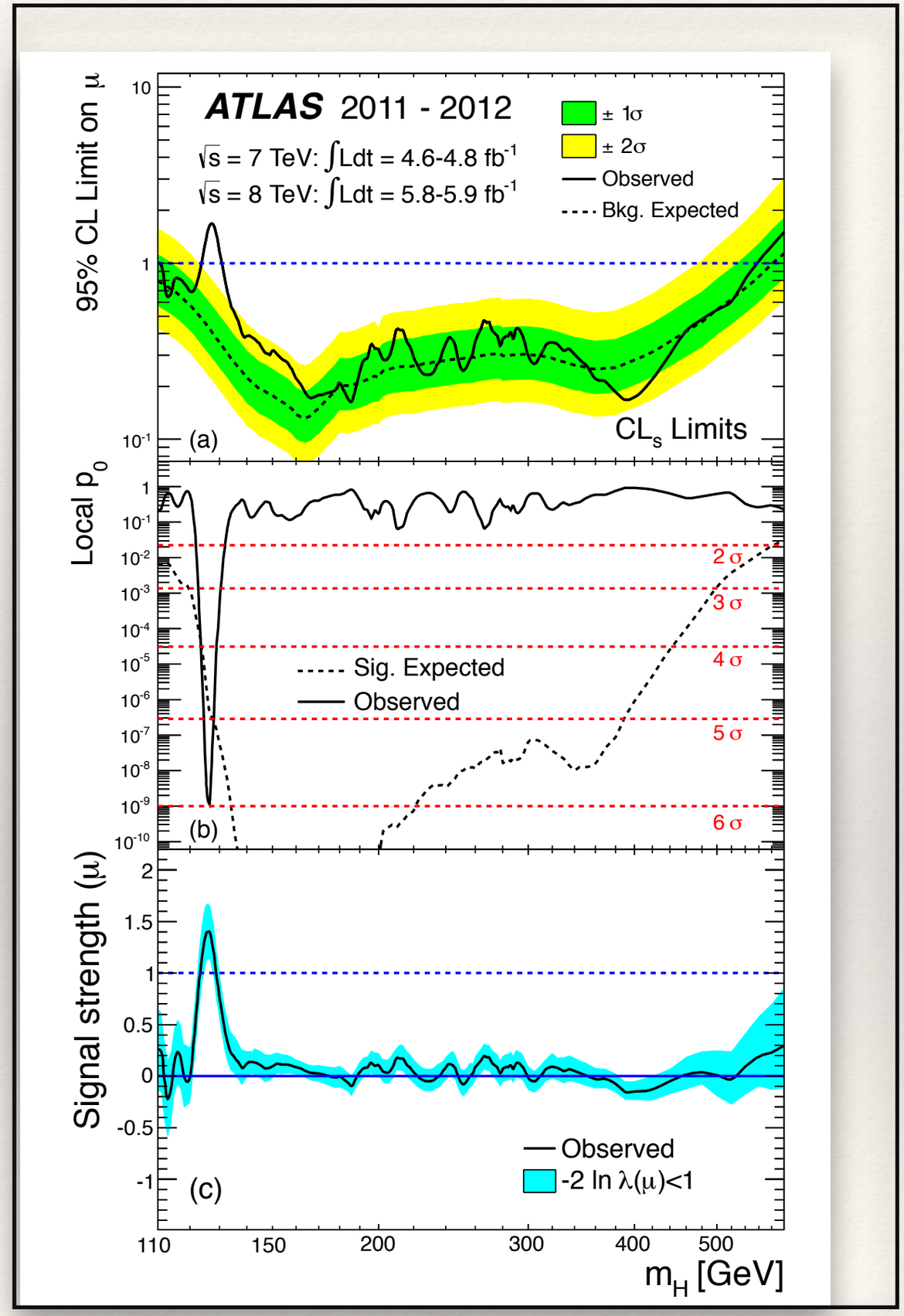
Kent W. Staley, Dept. of Philosophy, Saint Louis University

Selection, Significance, and Signification: Issues in High Energy Physics

Boston University
"Revisiting the Foundations of Statistics in the Era of Big Data: Scaling Up to Meet the Challenge"
February 21, 2014

Questions

- Is the use by HEP of a methodology of significance testing warranted?
- What warrants that practice?
- How might the use and rationale for significance testing by HEP in the past guide the development of new statistical methods for the future?



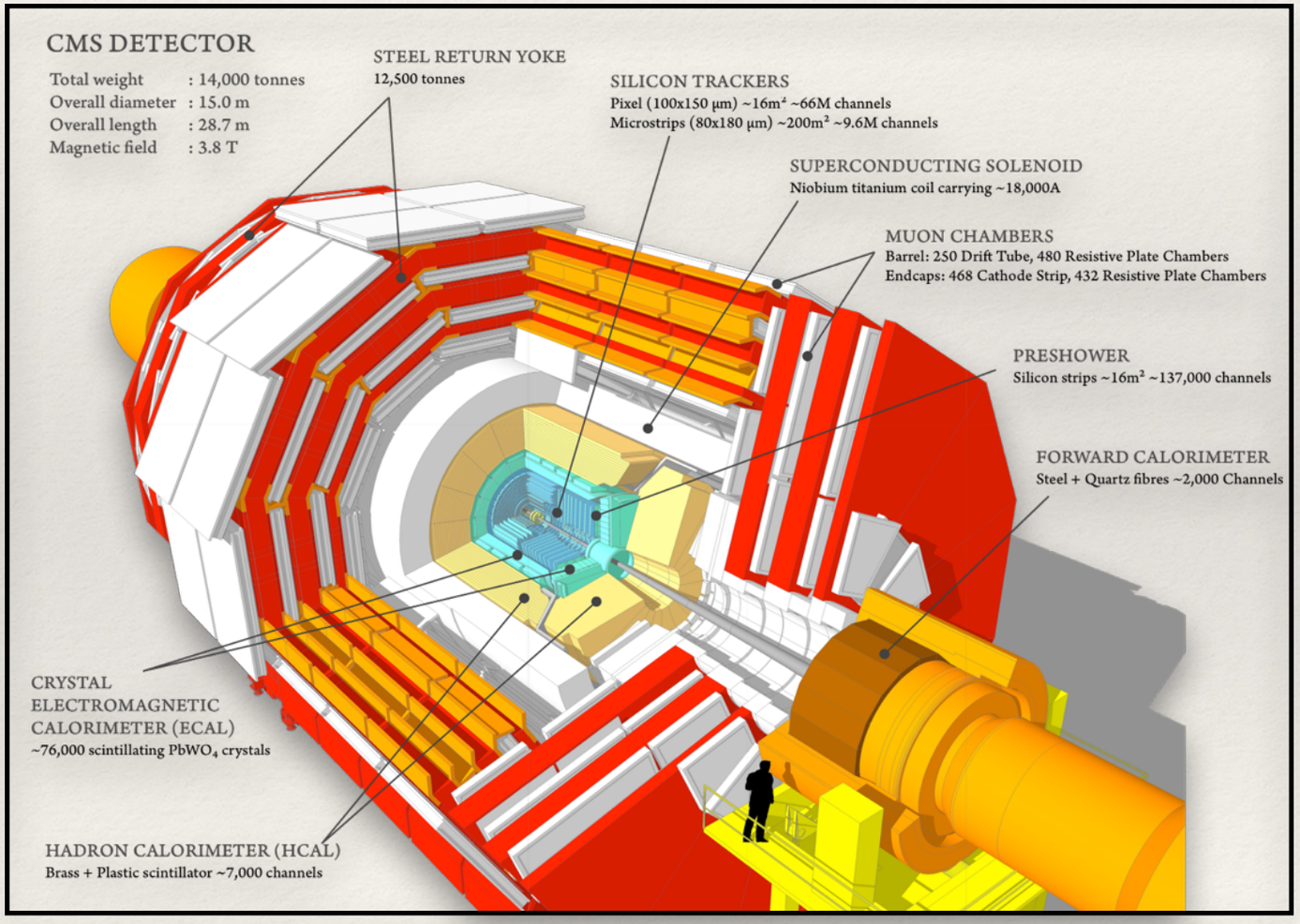
Outline

- Introduction: July 4, 2012
- Reception of the Higgs announcement
- The roles of null hypotheses in HEP
- Problems for some views about significance testing
- Pragmatism in HEP statistics and in philosophy
- Conclusion: Looking ahead

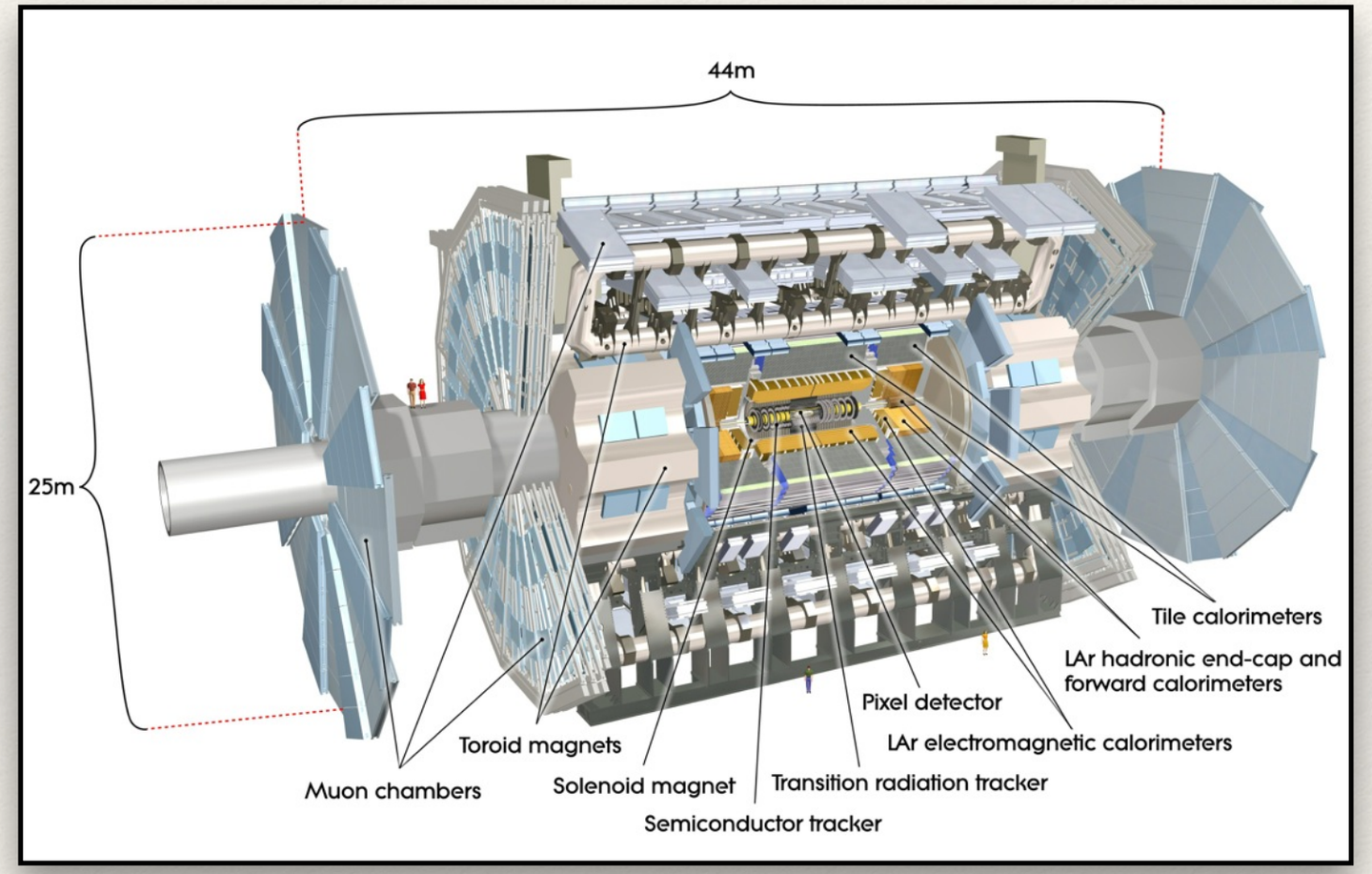
Introduction: July 4, 2012

CMS

- “CMS observes an excess of events of approximately 125 GeV with a statistical significance of five standard deviations … above background expectations.”
  - CMS Press Release, “Observation of a New Particle with a Mass of 125 GeV”



ATLAS



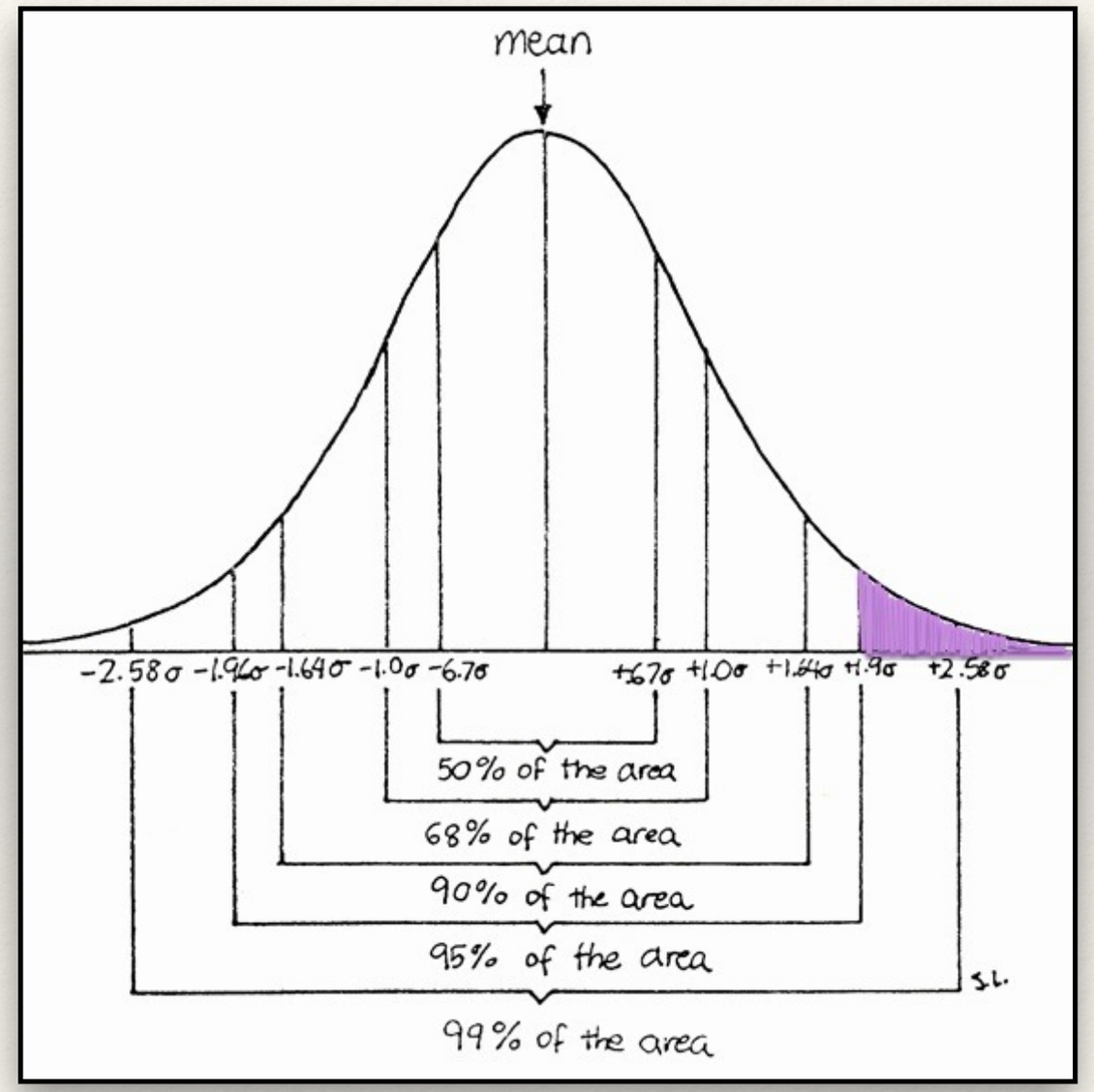
- "We observe in our data clear signs of a new particle, at the level of 5 sigma, in the mass region around 126 GeV"
  - ATLAS spokesperson Fabiola Gianotti, quoted in ATLAS Press Release, "Latest Results from ATLAS Higgs Search"

Significance testing methodology

- a null hypothesis H_0
  - e.g., $\mu = 0$, where μ denotes the mean value in a population
- a test statistic $d(\mathbf{X})$
  - $d(\mathbf{X})$ has a known distribution under H_0
  - $d(\mathbf{X})$ is a distance measure
  - larger values of $d(\mathbf{X})$ indicate stronger evidence of departure from what is expected if H_0 is true.
- from the data: $d(\mathbf{X}) = d(\mathbf{x}_0)$
- the p-value is $\Pr(d(\mathbf{X}) \geq d(\mathbf{x}_0); H_0)$

From *p*-values to σ's

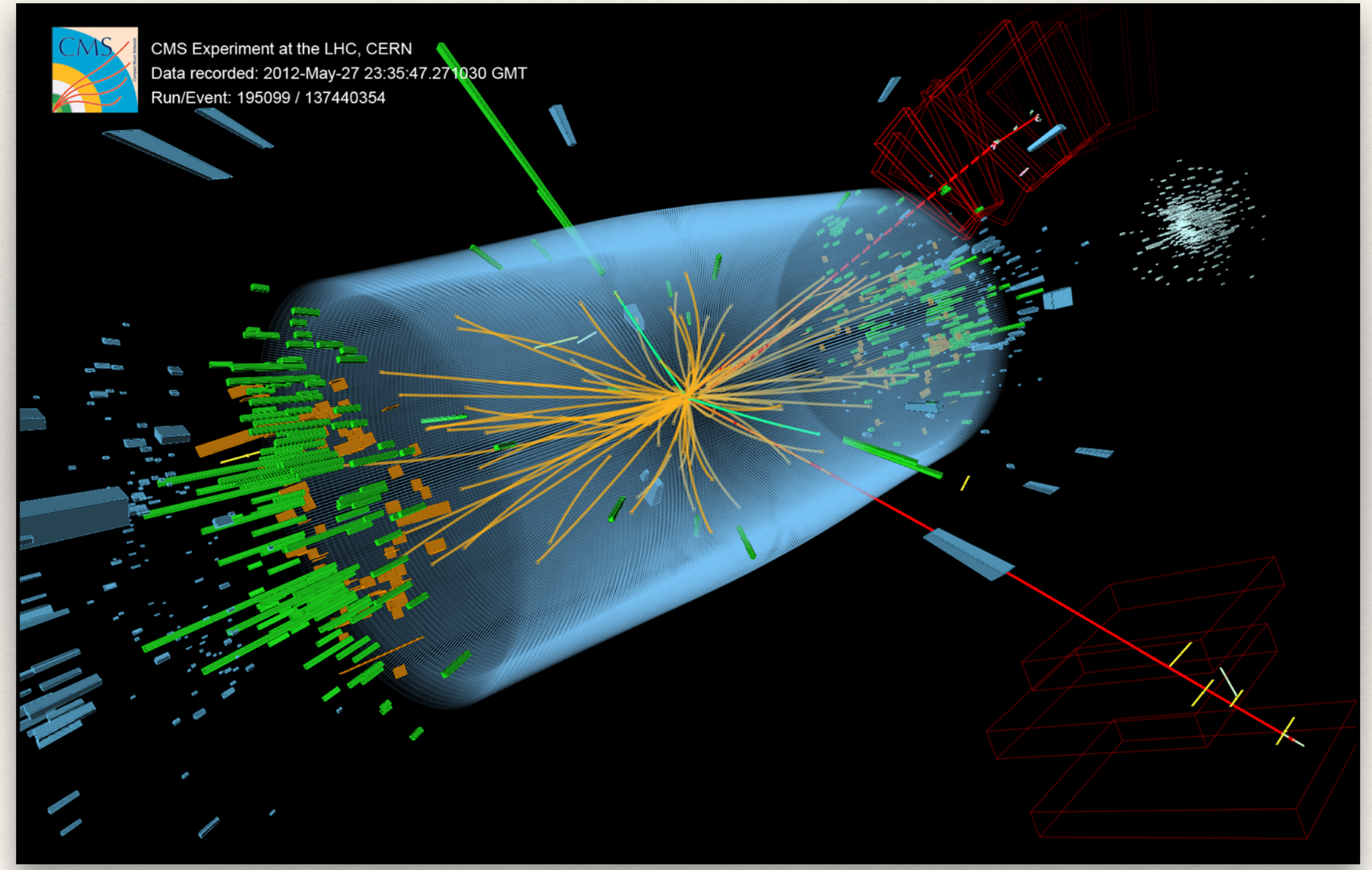
- The *p*-value can be converted to σ's by indicating the number of σ's from the null expectation value to $d(\mathbf{x}_0)$.



The null test component of the Higgs search

- A one-sided significance test:
- H_0: $\mu = 0$; H_1: $\mu > 0$
- where μ is the Higgs "signal strength" for a Higgs boson with a given mass.

Search methodology in HEP



Search methodology in HEP

- Identify physical signature (decays into known particles identifiable via measured properties)
- Operationalize physical signature using data selection criteria ("cuts") for identifying *candidate events*
- Estimate background (partly theoretical/simulated, partly data-driven): Expected number of candidate events.
- Test null ("background only") hypothesis H_0 against alternative H_1 ("signal + background").

Search methodology in HEP

- The typical test statistic is the likelihood ratio

$$d(\mathbf{X}) = -2\ln\frac{Pr(\mathbf{X}|H_0)}{Pr(\mathbf{X}|H_1)}$$

- The *p*-value is then calculated:

$$p = Pr(d(\mathbf{X}) \geq d(\mathbf{x_0})|H_0)$$

A Bayesian alternative

$$p(\mu|x) = \frac{L(x|\mu)\pi(\mu)}{\int L(x|\mu)\pi(\mu)d\mu}$$

Reception of the Higgs announcement

Press

- New York Times (04.07.12): “Both groups said that the likelihood that their signal was a result of a chance fluctuation was less than one chance in 3.5 million, ‘five sigma,’ which is the gold standard in physics for a discovery."

Press

- Reuters (04.07.12): "Five sigma, a measure of probability reflecting a less than one in a million chance of a fluke in the data, is a widely accepted standard for scientists to agree the particle exists."

Blogging Bayesians

Tony O'Hagan, on the ISBA Forum:

- "Why such an extreme evidence requirement? We know from a Bayesian perspective that this only makes sense if (a) the existence of the Higgs boson (or some other particle sharing some of its properties) has extremely small prior probability and/or (b) the consequences of erroneously announcing its discovery are dire in the extreme. Neither seems to be the case, so why 5-sigma?"

Blogging Bayesians

Tony O'Hagan, on the ISBA Forum:

- "Rather than ad hoc justification of a *p*-value, it is of course better to do a proper Bayesian analysis. Are the particle physics community completely wedded to frequentist analysis? If so, has anyone tried to explain what bad science that is?"

The Roles of null hypotheses in HEP

Null hypotheses that are "taken seriously" (credible nulls)

- The rate at which protons decay is exactly zero.
- The speed at which neutrinos (and other things) travel is always $< c$.
- There are no scalar leptoquarks.
- Any "Beyond-the-Standard-Model" (BSM) search would be an example of a test of such hypotheses.

Null hypotheses that were widely disbelieved (incredible nulls)

- There is no top quark (FNAL ca. 1994)
- Top quarks are produced only in pairs and never as single particles (FNAL ca. 2008)
- There is no Higgs boson (CERN ca. 2011)
- "The main case in which we place little prior belief on the null is an artificial case in which the null hypothesis is the Standard Model with a missing piece!"
  - R. Cousins, "The Jeffreys-Lindley Paradox and Discovery Criteria in High Energy Physics"

Problems for some views about significance testing

The "sub-conscious Bayes factor" interpretation

- A high-significance threshold (like 5σ) is only warranted when the prior Bayesian probability of the alternative hypothesis (π_1) is very low.
- O'Hagan: 5σ "only makes sense if the existence of the Higgs boson … has extremely small prior probability….it is of course better to do a proper Bayesian analysis"
- HEP uses the same threshold for credible and incredible nulls, though some have called for reforming this practice.
  - See Louis Lyons, "Discovering the Significance of 5σ"

The parsimony interpretation

- Null hypotheses are typically "no effect" hypotheses. They are therefore more parsimonious than alternative hypotheses and this warrants giving them "the benefit of the doubt"
  - No effect hypotheses are not necessarily more parsimonious.

Pragmatism in HEP statistics and in philosophy

A pragmatic alternative

- The purpose of an experiment is to answer a question.
  - It is an opportunity to learn something.
- The choice of a null hypothesis should be guided by its appropriateness to the question at hand.
- The overall design of the experiment should be guided by the *learning goals* of the experiment, the *possible errors* that investigators confront, and their *practical consequences, including those that bear on related inquiries*.

Rationales for the 5σ standard

- huge investment
- "so much can go wrong" (systematics)
- the Look-Elsewhere-Effect (LEE)
- resilience: protection against an excess of early luck
  - "It is not so hard to lose a sigma with added data or other changes to an analysis."
    - Joe Incandela, former CMS spokesperson, personal communication

The LEE

- The Higgs boson mass m_H was not known in advance. So the location of the excess to be reported was not fixed in advance.
- m_H a nuisance parameter
- 5σ applied to *local* significance
  - p calculated as a function of m_H
  - The minimum p_{min} of this function is the *local p-value*
  - Corresponding significance is *local significance*.

The LEE

- But the probability p_{real} of getting a result that yields the calculated value of p_{min} somewhere in a range of possible masses (e.g., possible values of m_{H}) is greater than p_{min}.
- 5σ standard is applied to this nominal local significance.

Is this cheating?

- Why not report the true p-value p_{real} instead?
- The "true" p-value is ill-defined because of the ambiguity of the space of possible discoveries.
- A *global* p-value is also reported, which is calculated relative to some specified (but somewhat arbitrary) mass range.

Local and global Higgs significances

- CMS:
  - local: 5.0σ
  - global: 4.6σ for the search range 115-130 GeV
  - global: 4.5σ for the search range 110-145 GeV
- ATLAS:
  - local: 5.9σ
  - global: 5.3σ for the search range 110-150 GeV
  - global: 5.1σ for the search range 110-600 GeV

July, 2012

Consequences of error

- Deciding to announce "Observation" has practical consequences
  - for future data analysis: from searching to measuring
  - for public relations: holding a press conference and having a lot of attention directed your way
    - and drawing attention to the enormous resources expended in pursuit of the answer to the question addressed in this experiment

Behaviorism?

- The design and implementation of statistical procedures in HEP has been guided by a concern with the learning goals of particular experiments, the consideration of the most salient possible errors in those experiments, and the weighing of the consequences of such errors.
- Does this amount to treating these statistical procedures as simply devices for making decisions rather than evaluating evidence?

Behaviorism?

"We are inclined to think that as far as a particular hypothesis is concerned, no test based upon the theory of probability can by itself provide any valuable evidence of the truth or falsehood of that hypothesis….

- J. Neyman and E. Pearson, "On the Problem of the Most Efficient Tests of Statistical Hypotheses"

Jerzy Neyman

Behaviorism?

Egon Pearson

"Without hoping to know whether each separate hypothesis is true or false, we may search for rules to govern our behaviour with regard to them, in following which we insure that, in the long run of experience, we shall not be too often wrong…. Such a rule tells us nothing as to whether in a particular case *H* is true when [the rule says to accept it] or false when [the rule says to reject it]."

- J. Neyman and E. Pearson, "On the Problem of the Most Efficient Tests of Statistical Hypotheses"

Behaviorism?

- Royall's three questions:
  1. What should I believe?
     - Bayesianism
  2. What should I do?
     - Frequentism
  3. How should I interpret this body of observations as evidence?
     - Likelihoods

Pragmatism

"How to Make Our Ideas Clear" (1878)

- Clarity
- Distinctness
- "a third grade of clearness of apprehension"

Charles Sanders Peirce

The pragmatic maxim

"Consider what effects, which might conceivably have practical bearings, we conceive the object of our conception to have. Then, our conception of these effects is the whole of our conception of the object."

- The pragmatic orientation that I wish to highlight can be regarded as the application of this maxim – or something close to it – to the outcomes of statistical inferences.

Pragmatism, not behaviorism

- Behaviorism becomes problematic when it seeks to *reduce* the process of inquiry to decision-making.
  - reducing "theoretical aspects of science to technology and decision-making" (Isaac Levi)
- But pragmatism is a means of achieving greater clarity:
  - What is your aim?
  - What would it be like to get it wrong?
  - What's at stake?

Conclusion: Looking ahead

The statistical future of HEP?

- Significance testing has limitations
  - but has been used in its limited role to good effect
- 5σ as a uniform and rigid rule is pragmatically inappropriate
  - but is already not used as a uniform and rigid rule
  - reforms for more flexible standards have been proposed (taking into account impact, LEE, systematics, “sub-conscious Bayes factor”)

The statistical future of HEP?

- Continuing innovation in statistical methods
  - both frequentist and Bayesian
- Not “anything goes” but “consider the effects”

thank you!