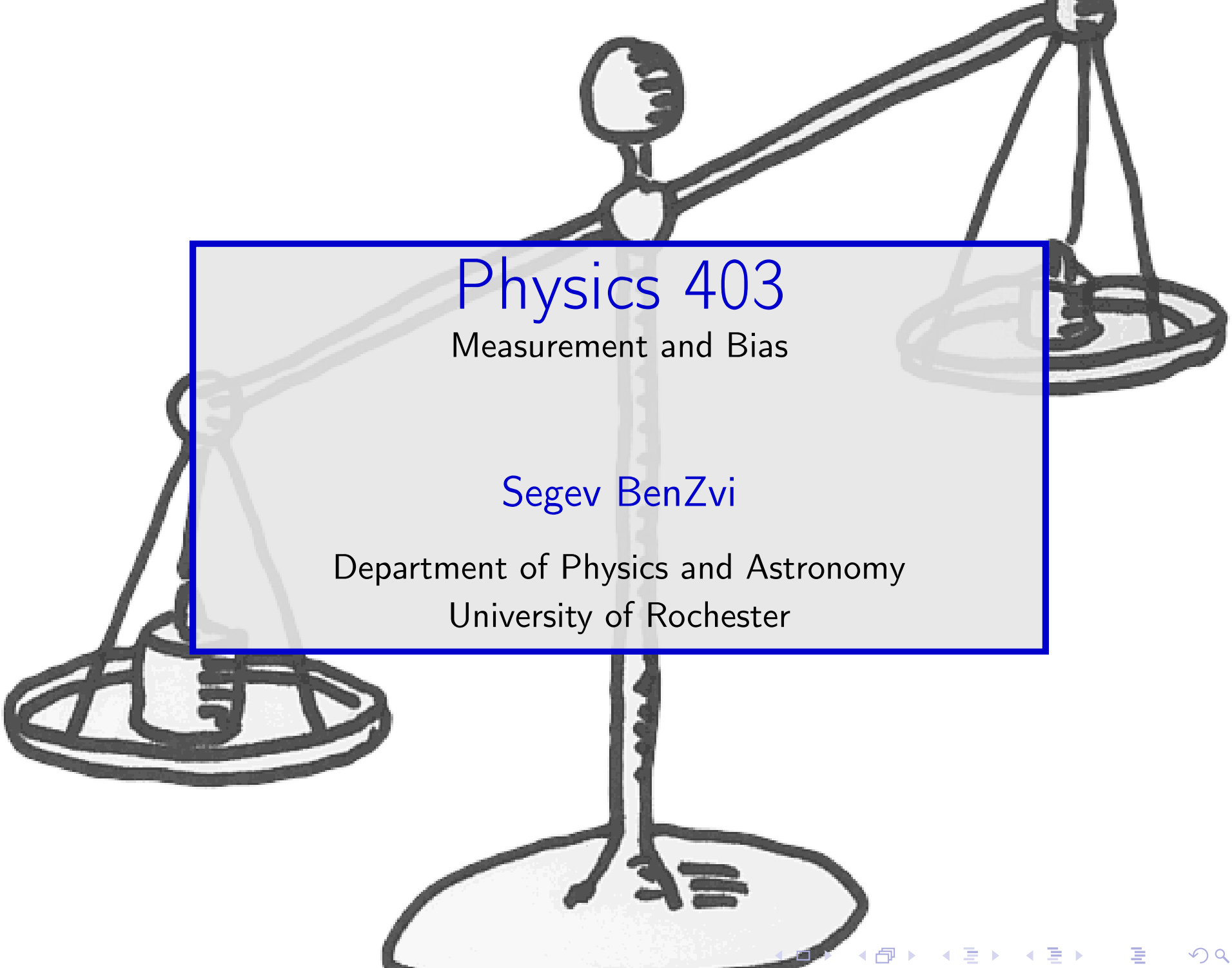


Physics 403

Measurement and Bias

Segev BenZvi

Department of Physics and Astronomy
University of Rochester

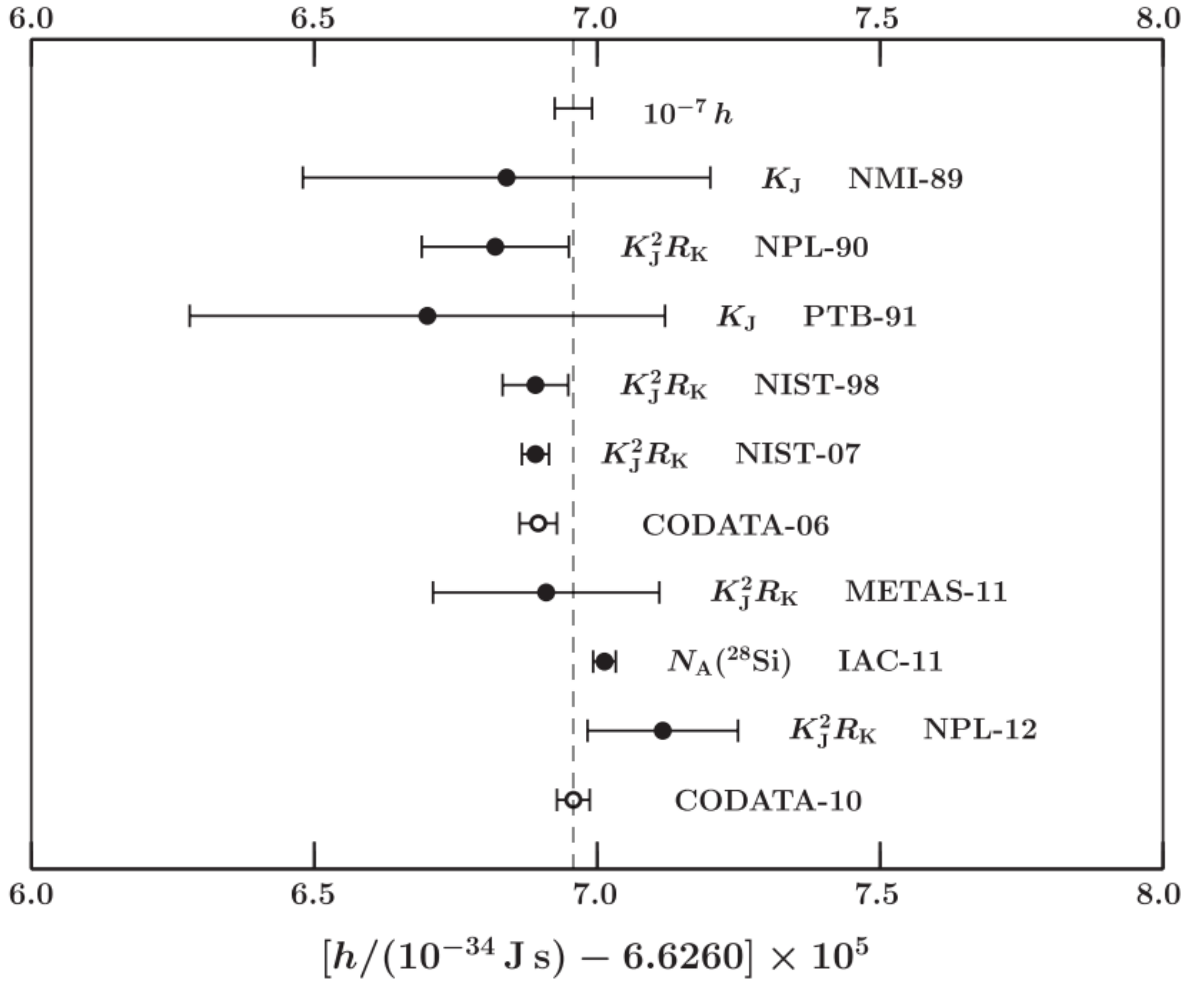
Table of Contents

Measurements and Inference in Physics

- We have discussed many statistical techniques in this course
- Ultimately what we are trying to do is run statistical tests and then make decisions based on those
- Applying a particular method of inference is easy. The hard part is interpreting the data
- **Ex.**: you make a measurement and it agrees with theory and/or a previous result. End of story?
- **Ex.**: you make a measurement and it disagrees with previous results. Is something wrong?
- Depending on your answer to these questions, you can very easily bias a result
- The application of a statistical technique is the beginning of the discussion, not the end

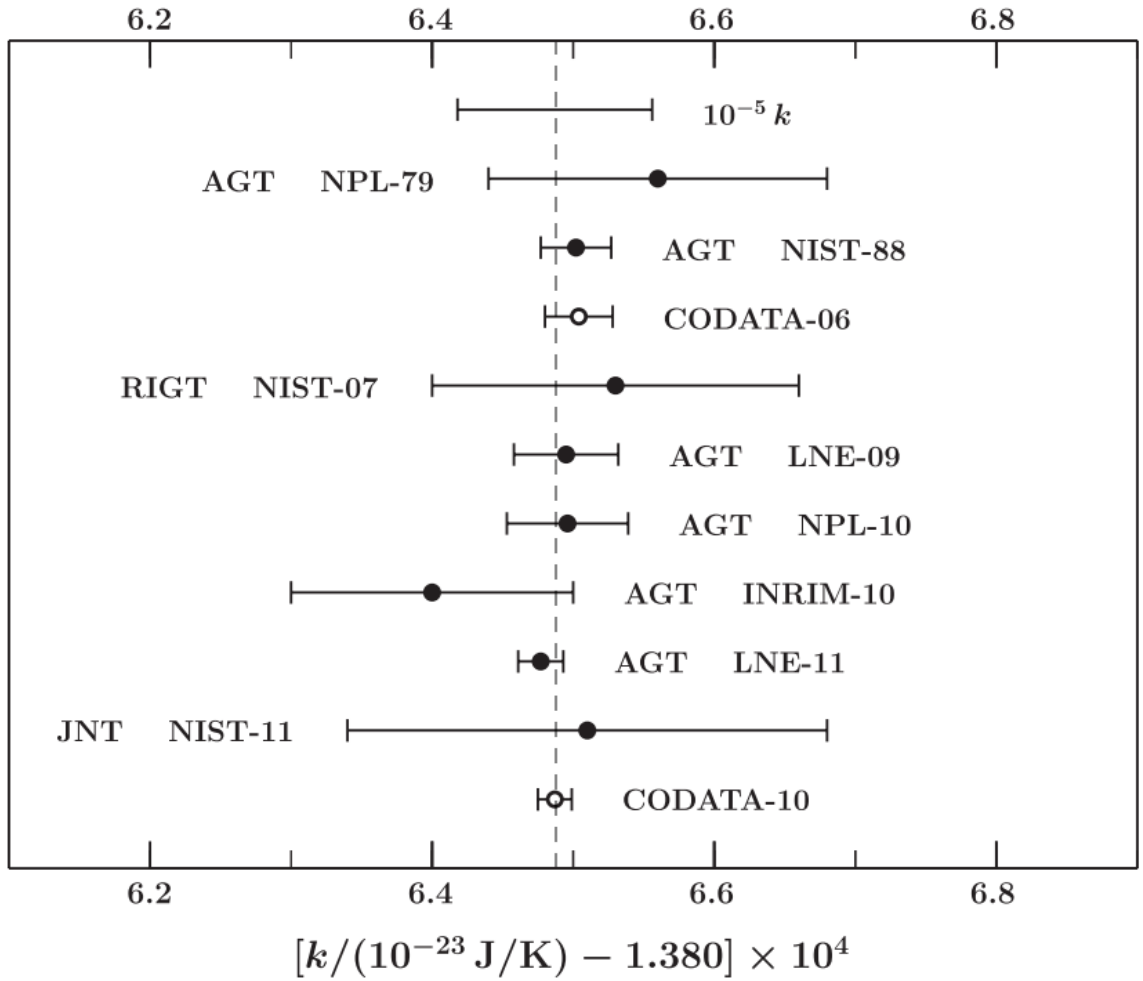
Measurements of h over 25 Years

Changes in the accepted value of h over time [1]



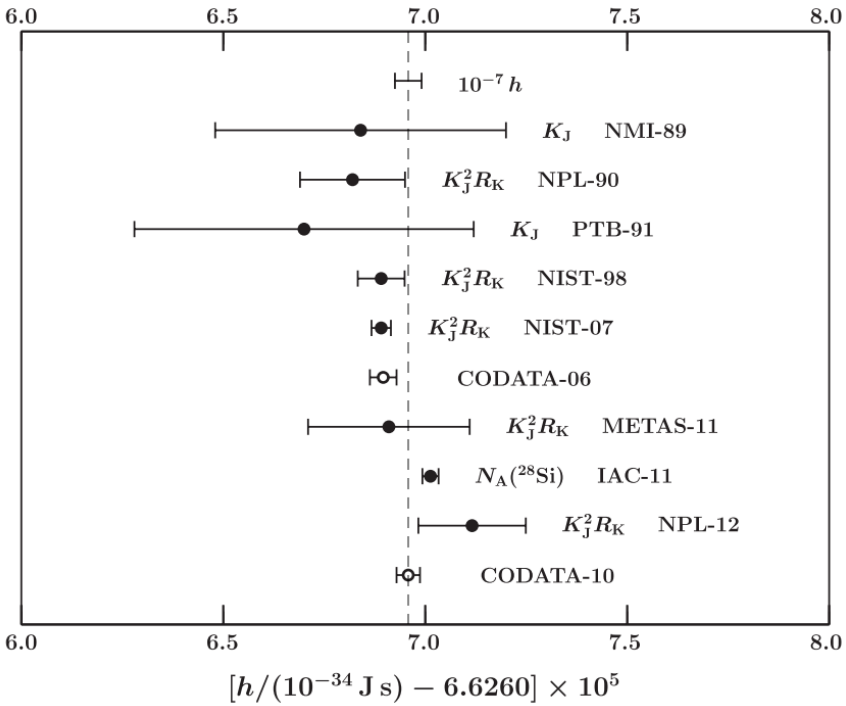
Measurements of k_B over 25 Years

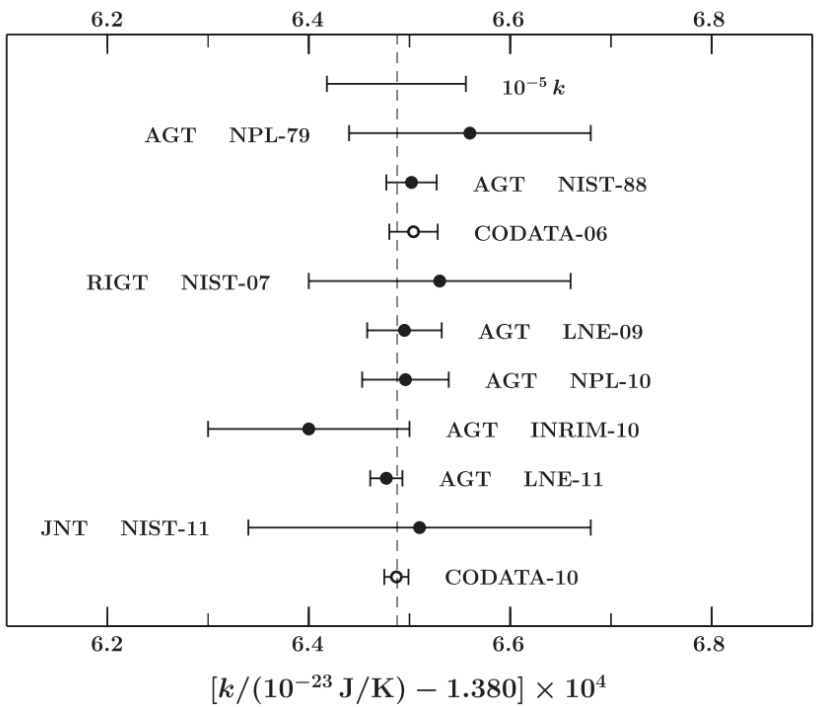
Changes in the accepted value of k_B over time [1]



Measurements of h and k_B

Measurements of h and k_B indicate random scatter about a central value, with no clear trend in the reported results

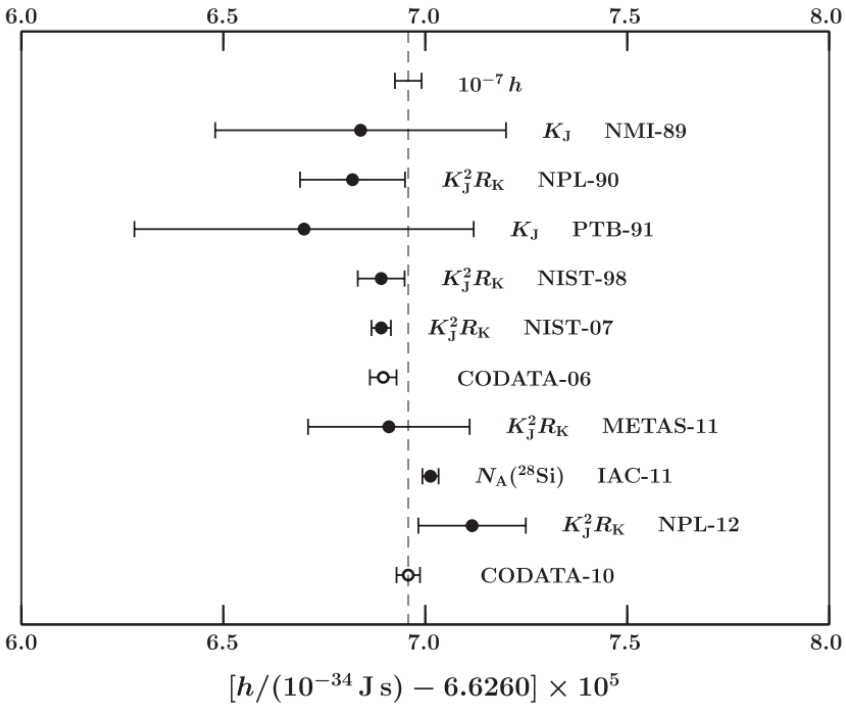


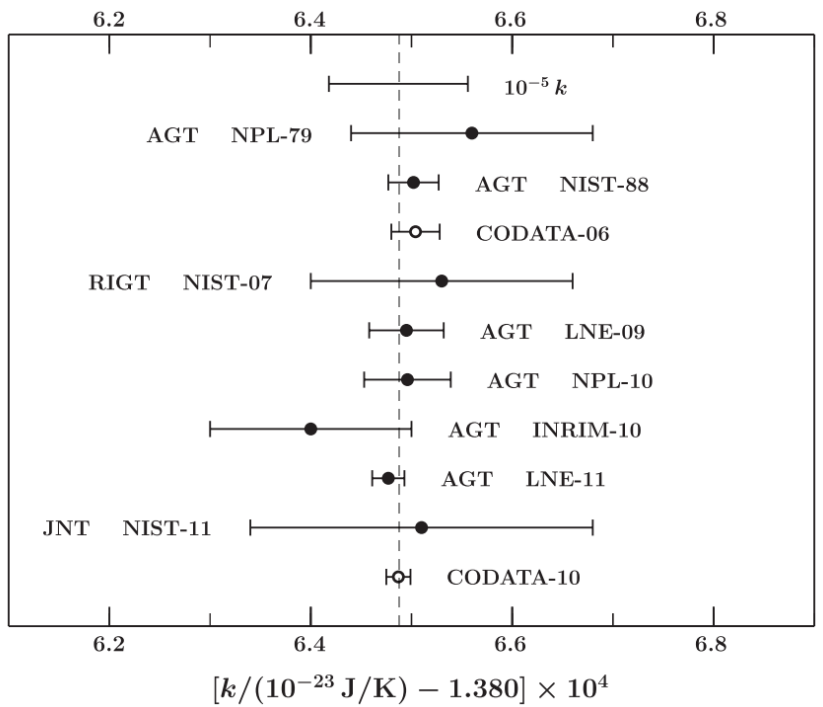


The CODATA-10 values have small uncertainties w.r.t. direct measurements. Why?

Measurements of h and k_B

Measurements of h and k_B indicate random scatter about a central value, with no clear trend in the reported results

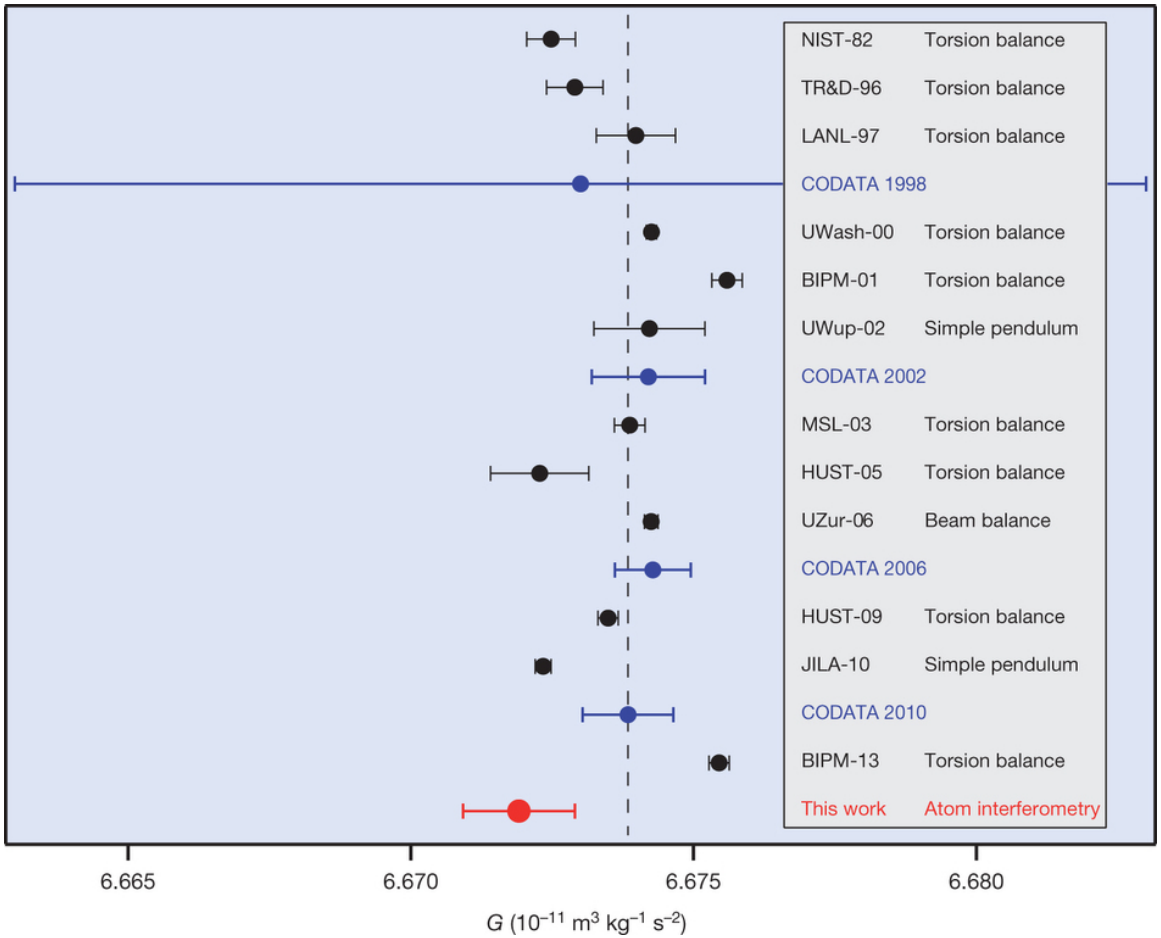




The CODATA-10 values have small uncertainties w.r.t. direct measurements. Why? CODATA h and k_B are inferred from N_A, not directly measured.

Measurements of G

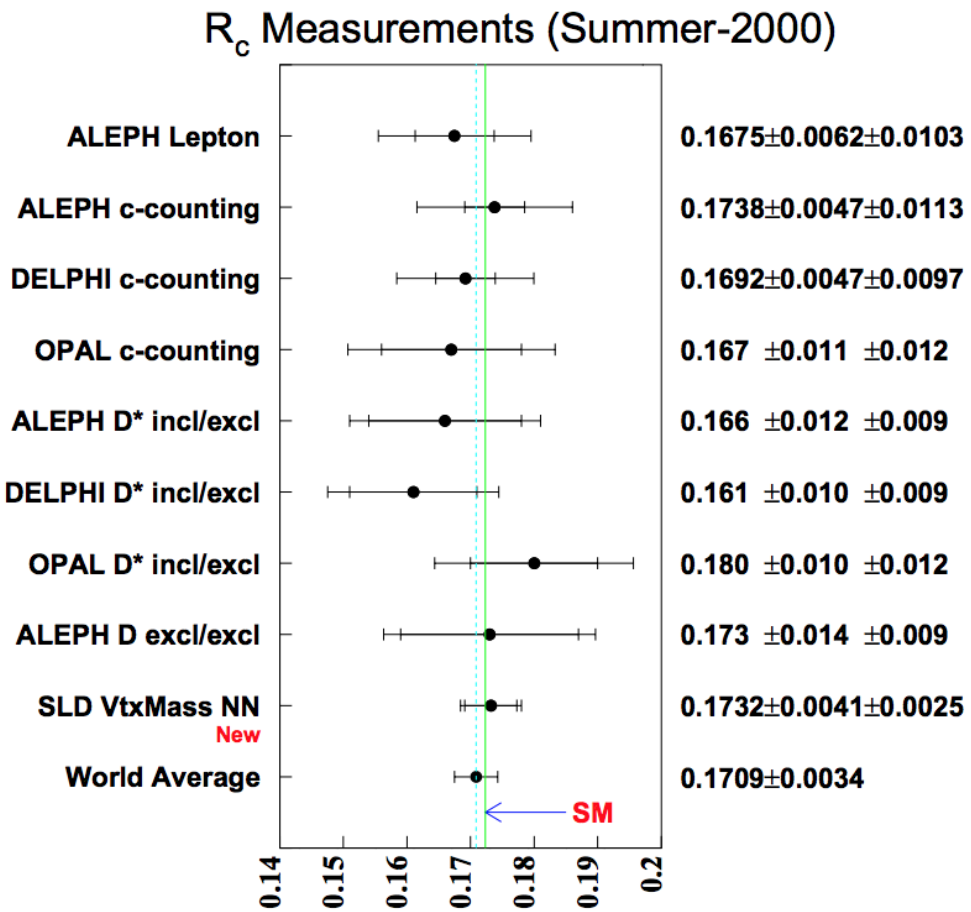
Recent measurements of G with CODATA-10 adjustments [2]



Notice anything odd about the uncertainties?

Do the Measurements Agree Too Well?

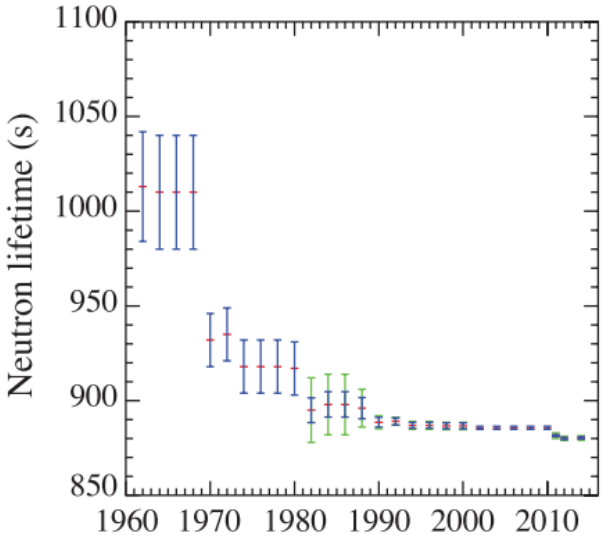
Measurements of R_c (Z^0 coupling to quarks) at SLAC and LEP [3]:

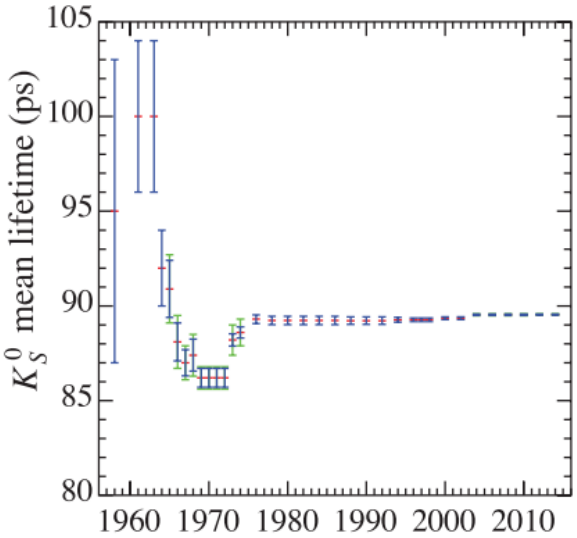


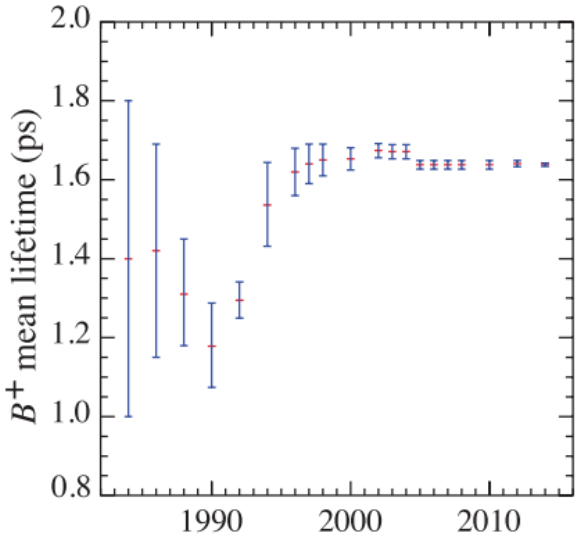
$\chi^2/\text{NDF} = 3.25/8$ (stat. only); $p = 0.08$, or $< 2\sigma$. Seems reasonable

Timeline of Selected Lifetime Measurements

Every year the Particle Data Group (PDG) publishes history plots of selected particle properties [4]





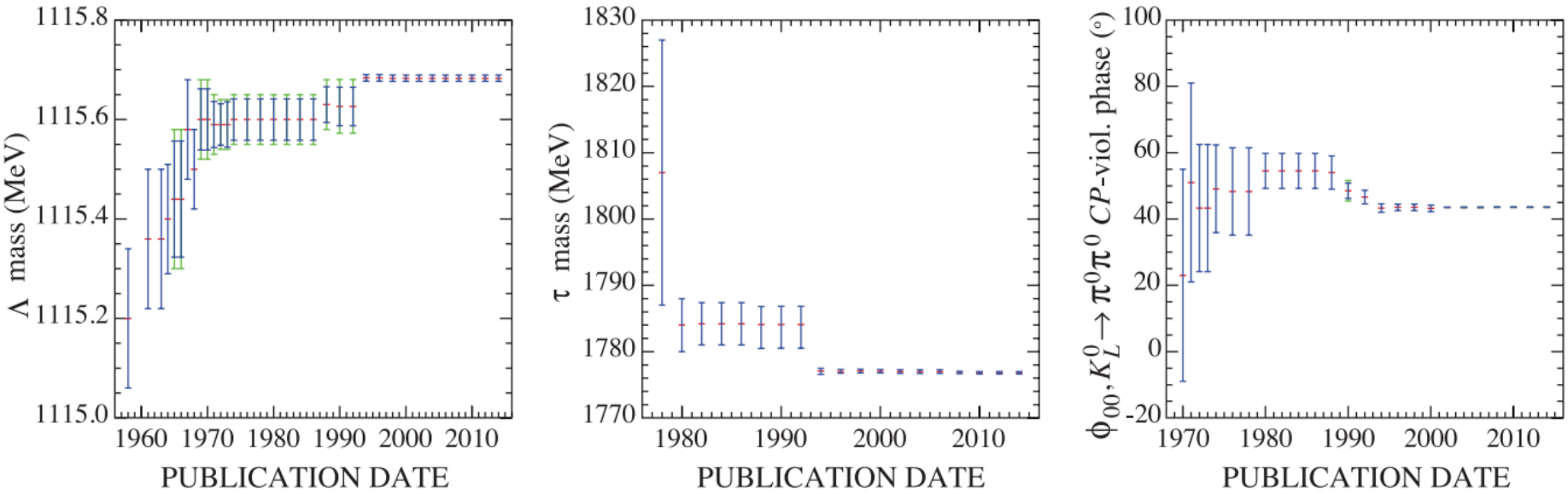


Our expectation is that the measurements should be scattered with Gaussian uncertainties about the world average

Instead, the history plots seem to exponentially decay!

Timeline of Selected Mass Measurements

More particle data [4]. The data are randomly scattered about the prior reported measurement rather than the world average, leading to an exponential distribution



This is known as the bandwagon effect [5]. Somehow the measurements are not really independent. Any thoughts?

Table of Contents

Confirmation Bias

We have a tendency to focus on evidence that conforms to prior expectations, and subtlely exclude contrary data

> *It's interesting to look at the history of measurements of the charge of an electron, after Millikan. If you plot them as a function of time, you find that one is a little bit bigger than Millikan's, and the next one's a little bit bigger than that, and the next one's a little bit bigger than that, until finally they settle down to a number which is higher.*
>
> *Why didn't they discover the new number was higher right away? It's a thing that scientists are ashamed of – this history – because it's apparent that people did things like this: When they got a number that was too high above Millikan's, they thought something must be wrong – and they would look for and find a reason why something might be wrong. When they got a number close to Millikan's value they didn't look so hard. And so they eliminated the numbers that were too far off, and did other things like that. We've learned those tricks nowadays, and now we don't have that kind of a disease.*

– R.P. Feynman, 1974

Origins of Bias

Feynman suggests we've learned our lesson since Millikan's day, but the PDG history plots show otherwise.

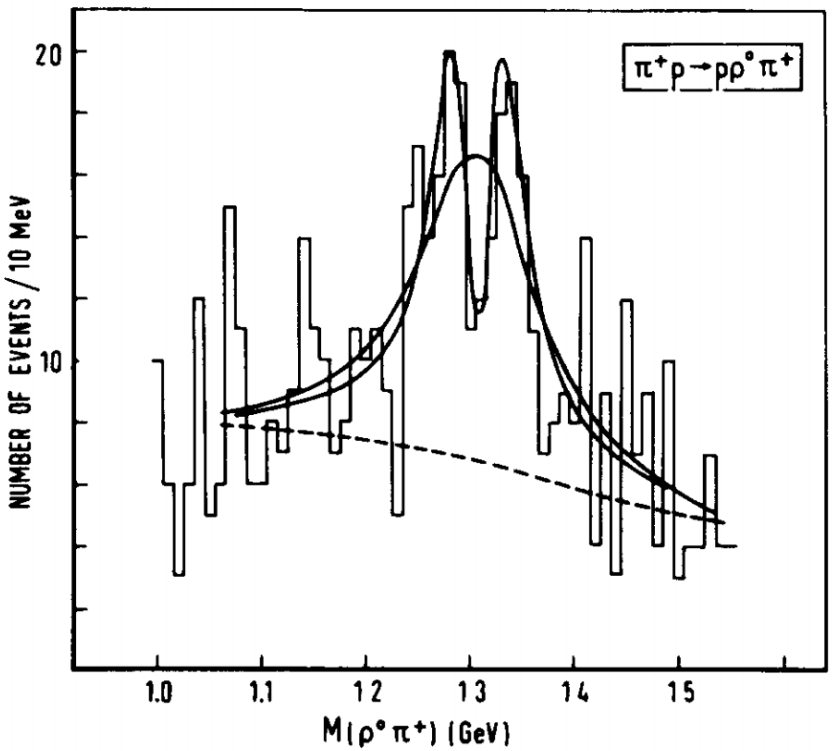
How does bias sneak into results so easily? From Harrison [6]:

- Event selection (a.k.a. cuts) is unconsciously tuned to get the "right" result
- The investigator looks for extra systematic effects when the numbers do not come out "right"
- If a result disagrees with expectations, comprehensive checks of the analysis [7] and experiment are performed. If the answer came out "right," the checks are not so comprehensive
- Stopping bias: the decision to publish or look for more data is made after doing the analysis and seeing if the results conform to expectations

Even well-intentioned investigators are prone to these kinds of mistakes.

Example: Data Selection Bias

Evidence for "split peak" in the A_2 spectrum from $\pi^- + p \to p + MM^-$ [8]:



Not seen in follow up data. Later revealed that the investigators checked their data and found "something wrong" in the instrument every time the split was not seen, so they tossed those data. But they didn't do the same checks when the split was seen!

Hidden Trials

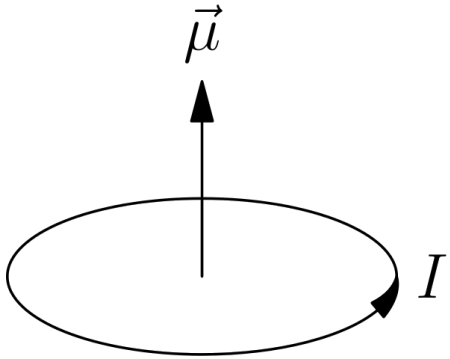
- We are human beings and will always be biased by prior results and theoretical expectations
- The big issue: things can go wrong when you put off checks and decision making until after you look at your data
- These *a posteriori* decisions are known as hidden trials. You take multiple opportunities to get the "right answer," affecting your conclusions in an insidious and hard-to-quantify manner
- Solution: do some contigency planning. Think about the checks you want to do before taking data, then make sure you do them

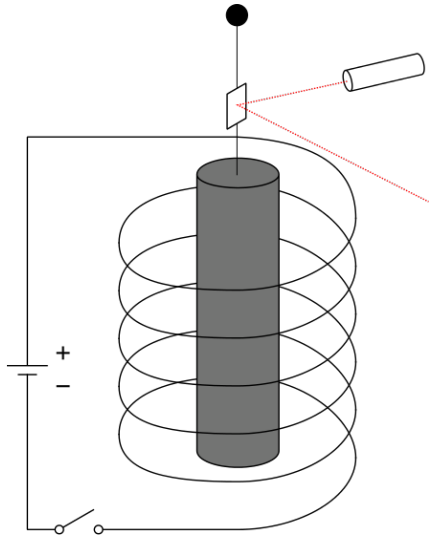
Example

Before looking at the data, ask the question: what checks would I do if I observed a $> 3\sigma$ anomaly? Then **do those checks anyway**.

Example: Stopping Bias

Einstein-de Haas effect: test Ampere's hypothesis that ferromagnetism is caused by little current loops from orbiting electrons





Setup: suspend an iron cylinder by a rope, reverse the field in the cylinder, and measure the torque

$$\boldsymbol{\mu} = I \cdot \pi r^2 \hat{\boldsymbol{\mu}} = \frac{e}{2\pi r/v} \cdot \pi r^2 \hat{\boldsymbol{\mu}} = \frac{1}{2} evr \hat{\boldsymbol{\mu}} = \frac{e}{2m} \boldsymbol{L}$$

$$g = \frac{|\boldsymbol{\mu}|}{|\boldsymbol{L}|} = \frac{e}{2m}$$

Example: Stopping Bias

- Einstein and de Haas performed the measurement in 1915 and observed $g = 1.02 \pm 0.10$, in agreement with the classical prediction [9]
- **Problem**: the classical prediction is wrong by a factor of 2
- The measurement had a lot of systematic effects. As these were reduced in follow-up experiments, people observed $g \to 2$ (eventually explained by Dirac)
- So how did Einstein and de Haas measure g nearly 10σ away from the currently accepted value?
- They tried to account for systematic effects, but probably stopped when their result agreed with the classical expectation
- **Solution**: be careful about whittling away systematic effects until you reach the "right" answer. Again, the issue is doing an *a posteriori* analysis of the data

Table of Contents

Blindness

Blindness is standard practice in clinical trials:

> *Double blind procedures are now universally taught and employed in the fields that experiment on human subjects: clinical research and psychology. It has long been recognized in those fields that, to avoid experimenter bias, it is not sufficient just to hide certain information from the human subjects; the information must also be hidden from the experimenters by the experimenters themselves until after the analysis is complete [10]*

Since the mid-1990s it's become increasingly popular in high-energy physics

Concept: there is no placebo effect to worry about, but as in a clinical trial, information about the results that may lead to a bias is hidden from the investigator until after the test is made

This is a good way to eliminated many sources of bias. But doing a blind analysis is a challenge

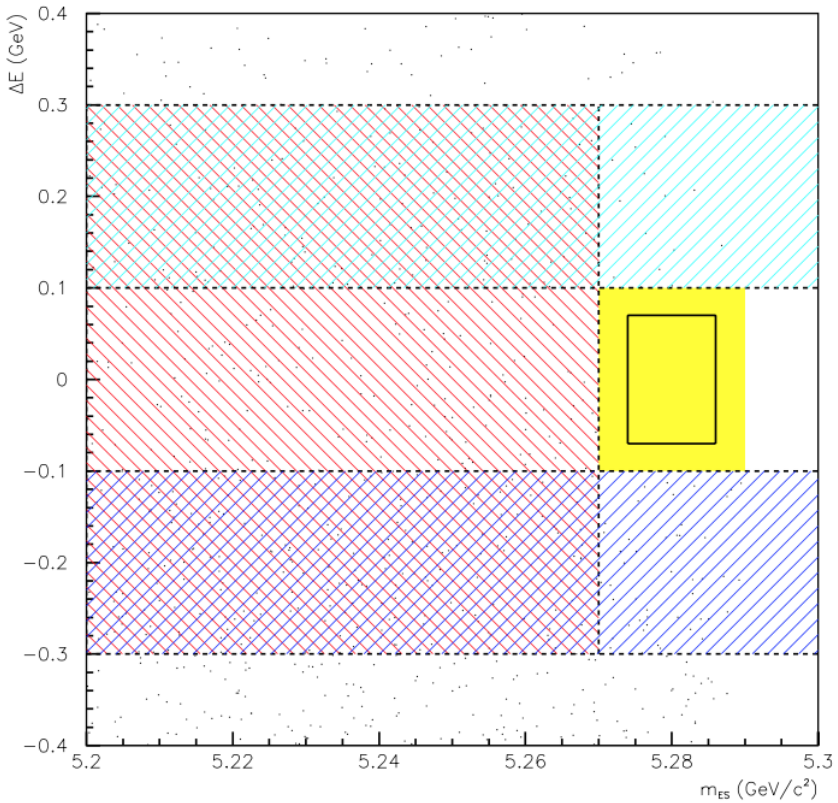
Common Approaches for Blind Analysis

The following techniqes are the most common methods for applying a blind analysis in physics, in order from most to least restrictive:

- **Hide the Answer**: hide events in a "signal region" where results are expected to occur. Perform diagnostics and tuning on data outside the signal region. Once cuts are finalized open the signal "box" to get the answer
- **Shift the Answer**: shift the answer by a random unknown offset Δ. This allows two independent groups to analyze the same data and compare answers without tuning on the real result. Once cuts are finalized remove the random offset Δ
- **Split the Data**: perform a non-blind analysis on a subset of the data, using it like a sandbox for tuning cuts. Once cuts are finalized, apply them to the blinded part of the data set.
- **Data Challenge**: hide all the data and tune cuts using Monte Carlo. Only works if Monte Carlo describes data well enough

Example: Hide the Answer

$B^0 \to \rho^{\mp}\pi^{\pm}$ Search at BABAR



Sidebands provide on-resonance estimates of backgrounds.

- Search for a rare B^0 decay [6]
- Analysis in two kinematic variables:

$$m_{\mathrm{ES}} = \sqrt{\left(\frac{s}{2} + \boldsymbol{p}_0 \cdot \boldsymbol{p}_B\right)^2 / E_0^2 - p_B^2}$$

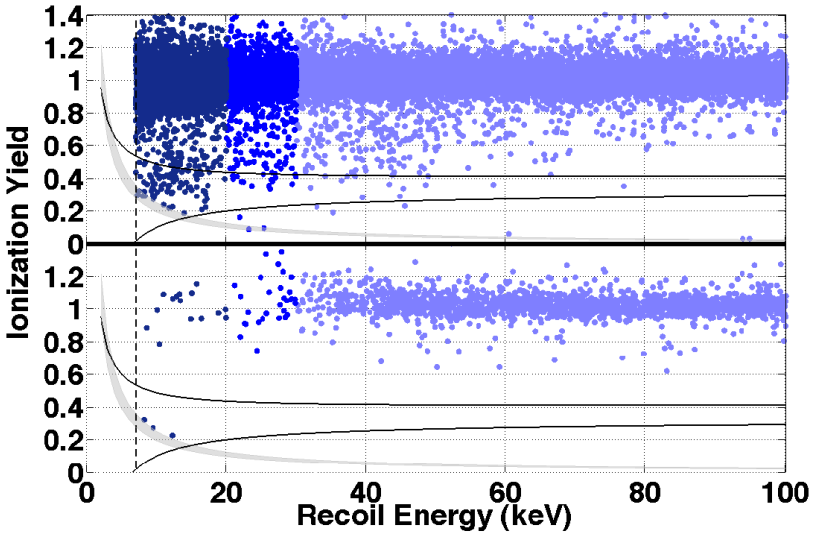
$$\Delta E = E_B^* - \sqrt{s}/2$$

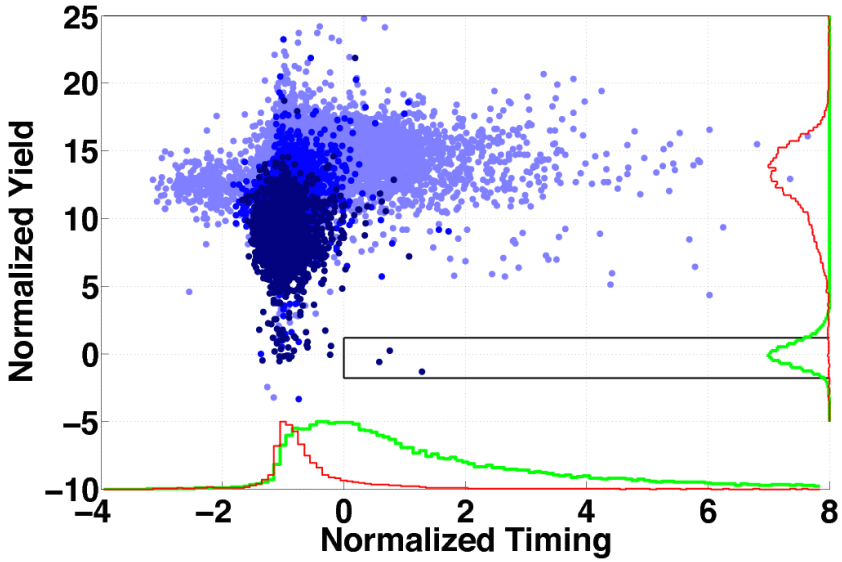
- Yellow box: signal blinding region, with optimized signal box drawn inside
- Shaded regions: signal sidebands for checking the shape and normalization of the background in m_{ES} and ΔE

Example: Hide the Answer

WIMP Nuclear Recoil Search at CDMS II

Looking for WIMPs with the Cryogenic Dark Matter Search (CDMS) experiment, which uses low-temperature Si detectors to identify dark matter interacting in the detector

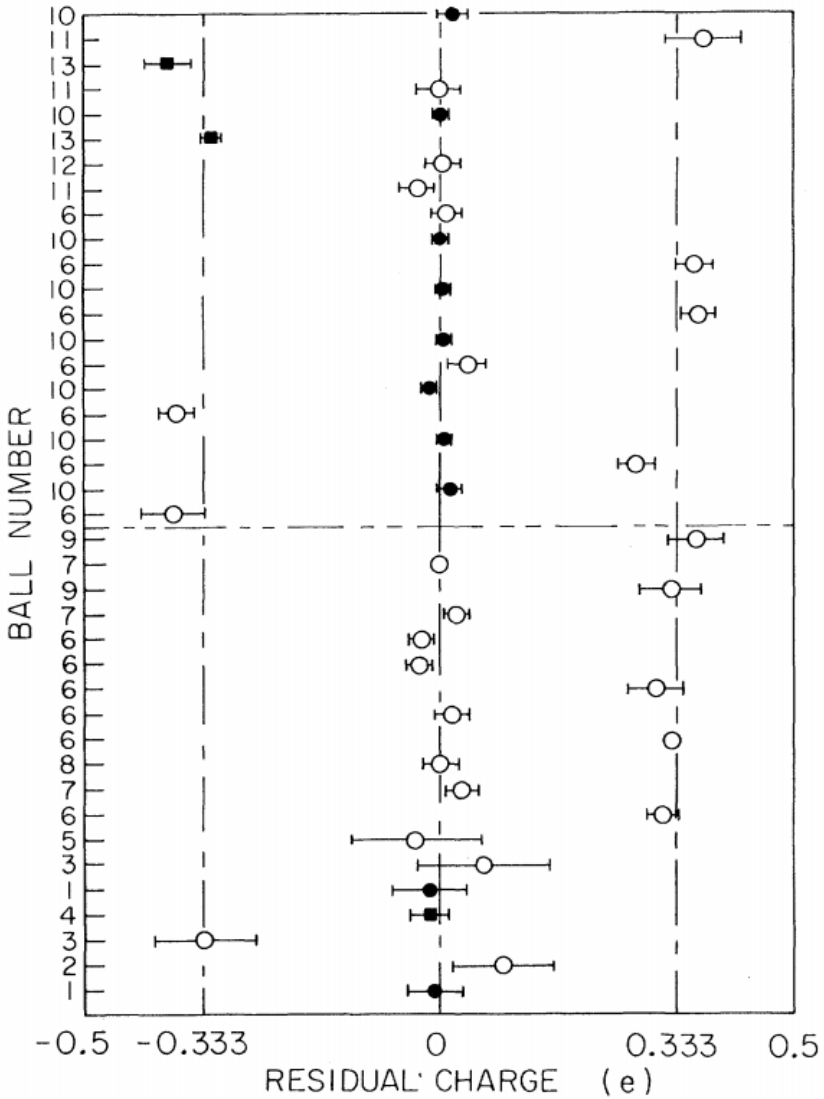




Tiny signal with large backgrounds from neutrons (cosmogenic and radioactive) and surface electron recoils. Results: 3 WIMP candidates in signal region after unblinding; 5.4% chance probability they are due to background [11]

Example: Shift the Answer

Observation of Free Fractional Charges

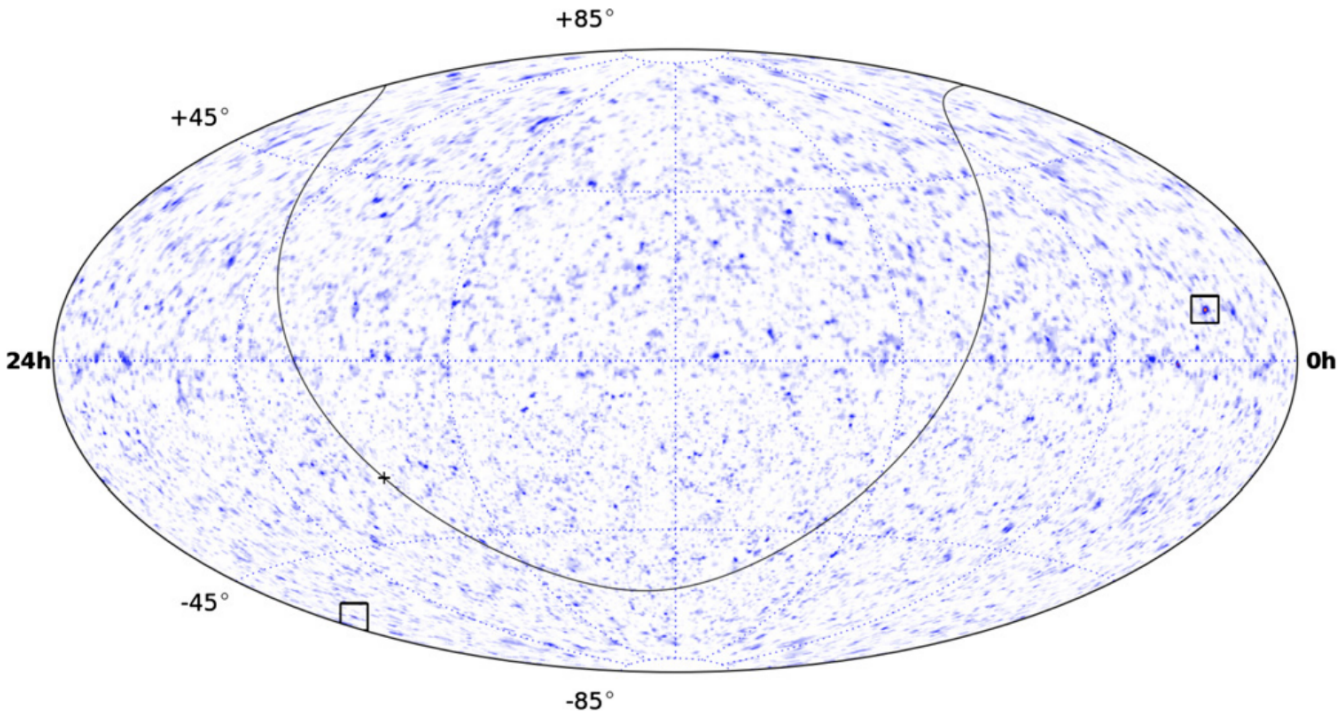


- 1981: measurements of levitated superconducting Nb spheres indicate the presence of charges of free charges of $\pm 1/3\ e$ [12]
- These results were measured in several studies by the same group
- Proposal: add a random value to the measured charge and repeat the analyis. Remove the random value when the analysis is complete
- Unblinded result: did not confirm "discovery" of free fractional charge [13]. An unconscious selection effect was previously at play

Example: Shift the Answer

Astrophysical Neutrino Searches

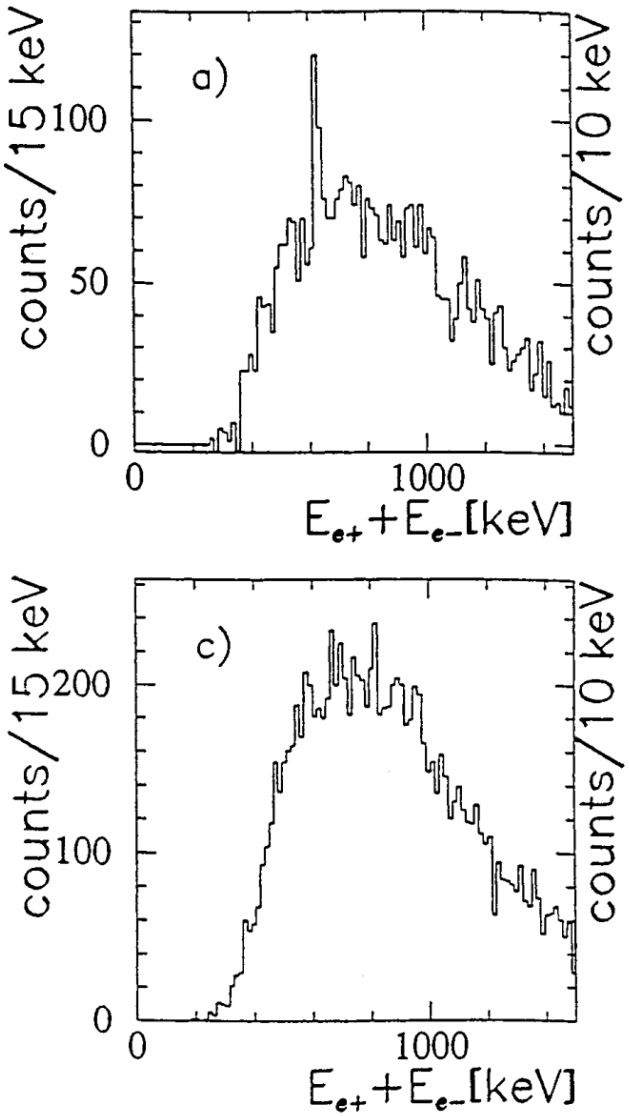
In IceCube, cuts on ν event size and direction could be tuned to enhance a clustering signal, so those data are blinded [14]



Specific procedure: local zenith angles are accessible but azimuth angles get a random offset so you can't observe the sky coordinates of events until cuts are locked in. No way to tune cuts on the hotspots

Example: Split the Data

Search for Low-Mass e^+e^- States

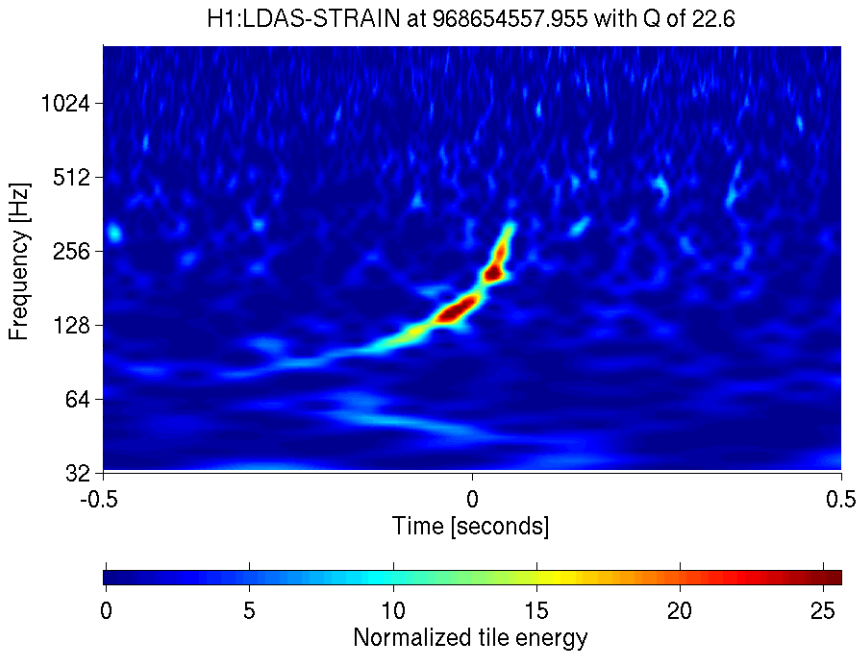


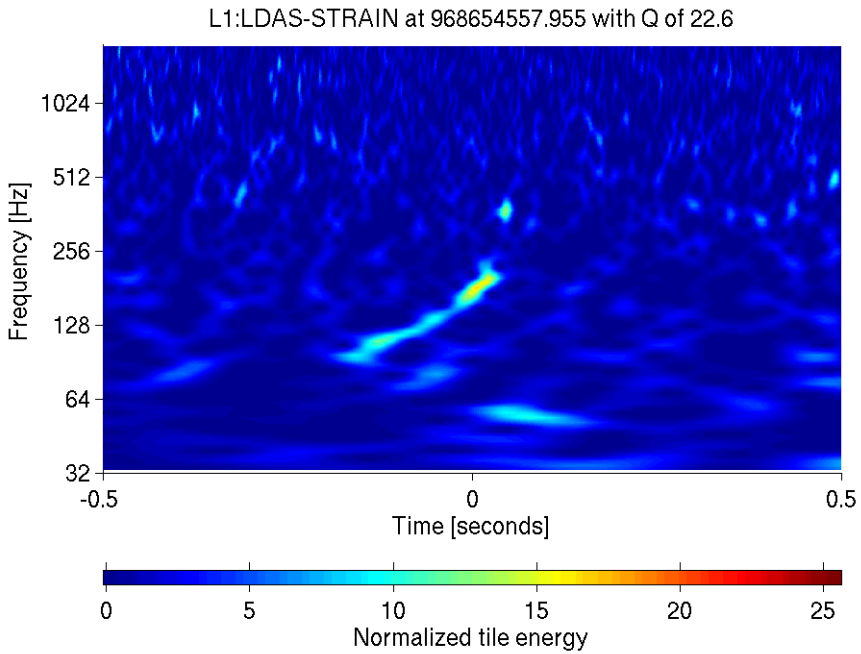
- Very narrow lines in e^+e^- spectra from heavy-ion collisions were claimed starting in the 1980s [15]
- 1996: EPOS collaboration conducts an analysis by splitting the data into two parts [16]
- **Part 1**: cuts tuned to enhance the peak in the e^+e^- spectrum
- **Part 2**: same cuts applied to a fresh data set. No peak observed
- Conclusion: previous experiments had inadvertently produced the peaks by tuning on fluctuations in their data!

Example: Data Challenge

Adding Fake Events to LIGO Data

Insertion of fake binary merger injected directly into the data stream of the LIGO detector, 2011:





Produces an end-to-end stress test of the analysis, including any biases introduced by human beings

What Blind Analysis Can and Cannot Do

Blind analyses can:

- Offer some protection against biases created by *a posteriori* analyses of data
- Cause you to think of your behavior as an experimenter before taking data

Note: you don't have to plan for every single contingency before taking data. Run the experiment as you like. The step to avoid is looking at the data and only then deciding what to do based on the result.

Blind analysis cannot:

- Protect against cheating (i.e., trying to look "into the box")
- Protect against fraud
- Guarantee that you won't make mistakes

What to do if the Blind Analysis Goes Wrong

Some problems you can encounter in a blind analysis:

- You open the signal box and observe many more events than expected due to a forgotten background. (Happened repeatedly in early diffuse ν analyses in IceCube)
- You remove a hidden offset and the answer makes no sense because you forgot a correction (also common)
- Blinding actually biased the analysis in one direction

If these things happen, it's best to disclose the outcome. It's fine to fix the mistake and report the corrected answer, but note publicly that it is an *a posteriori* result.

Excellent suggestion from Scott Oser: it also helps to have a written contingency plan to follow in case something goes wrong

Summary

- Ultimately, our statistical analyses are just the start of the conversation about the significance of a result
- Results in physics are prone to investigator bias because of the risk of *a posteriori* effects in your analysis
- These effects are insidious and impossible to quantify. Because they can mimic some of the issues caused by the look-elsewhere effect they are called hidden trials
- Blindness is a way to avoid hidden trials by forcing you to make your analysis and publication plans before looking into the data
- But, it's not a panacea and won't protect you from mistakes or bad luck
- The best solution for dealing with bad luck is honesty

References I

[1] Peter J. Mohr, Barry N. Taylor, and David B. Newell. "CODATA Recommended Values of the Fundamental Physical Constants: 2010". In: *Rev.Mod.Phys.* 84 (2012), pp. 1527–1605. arXiv: 1203.5425 [physics.atom-ph].

[2] G. Rosi et al. "Precision Measurement of the Newtonian Gravitational Constant Using Cold Atoms". In: *Nature* 510 (2014), p. 518. arXiv: 1412.7954 [physics.atom-ph].

[3] D. Su. "R(b), R(c) measurements at SLD and LEP-1". In: *Proc. XXXth ICHEP*. Osaka, Japan, 2000, pp. 632–636.

[4] K.A. Olive et al. "2014 Review of Particle Physics". In: *Chin. Phys.* C38 (2014), p. 090001.

[5] M. Jeng. "Bandwagon effects and error bars in particle physics". In: *Nucl.Instrum.Meth.* A571 (2007), pp. 704–708.

References II

[6] Paul Harrison. "Blind Analyses". In: *Conf. on Adv. Stat. Techniques in Particle Physics*. Durham, England, 2002, p. 278.

[7] Roger Barlow. "Systematic Errors: Facts and Fictions". In: *Conf. on Adv. Stat. Techniques in Particle Physics*. Durham, England, 2002, pp. 134–144. arXiv: hep-ex/0207026 [hep-ex].

[8] K. Boeckmann et al. "Decay-properties of the *a2(1300)*-meson". In: *Nucl.Phys.* B16 (1970), pp. 221–238.

[9] A. Einstein and W.J. de Haas. "Experimental proof of the existence of Ampere's molecular currents". In: *Proc. Koninklijke Akademie van Wetenschappen te Amsterdam*. Vol. 18. Amsterdam, Netherlands, 1915, pp. 696–711.

[10] Heinrich, Joel. *Benefit of Blind Analysis Techniques (CDF Memo)*. 2003. URL: http://www-cdf.fnal.gov/physics/statistics/notes/cdf6576_blind.pdf.

References III

[11] R. Agnese et al. "Silicon Detector Dark Matter Results from the Final Exposure of CDMS II". In: *Phys.Rev.Lett.* 111.25 (2013), p. 251301. arXiv: 1304.4279 [hep-ex].

[12] G.S. Larue, J.D. Phillips, and W.M. Fairbank. "Observation of Fractional Charge of (1/3)e on Matter". In: *Phys.Rev.Lett.* 46 (1981), pp. 967–970.

[13] L. Lyons. "Quark Search Experiments at Accelerators and in Cosmic Rays". In: *Phys.Rept.* 129 (1985), p. 225.

[14] M.G. Aartsen et al. "Searches for Extended and Point-like Neutrino Sources with Four Years of IceCube Data". In: (2014). arXiv: 1406.6757 [astro-ph.HE].

[15] Allan Franklin. "Selectivity and the Production of Experimental Results". In: *Arch. Hist. Exact Sci.* 53 (1998), pp. 399–485.

References IV

[16] R. Ganz et al. “Search for e+ e- pairs with narrow sum - energy distributions in heavy ion collisions”. In: *Phys.Lett.* B389 (1996), pp. 4–12.