

Integrity of Data Samples and Results in High Energy Physics

Jeffrey D. Richman

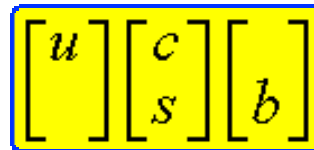
University of California, Santa Barbara

BABAR Experiment, SLAC

CMS Experiment, CERN



TM

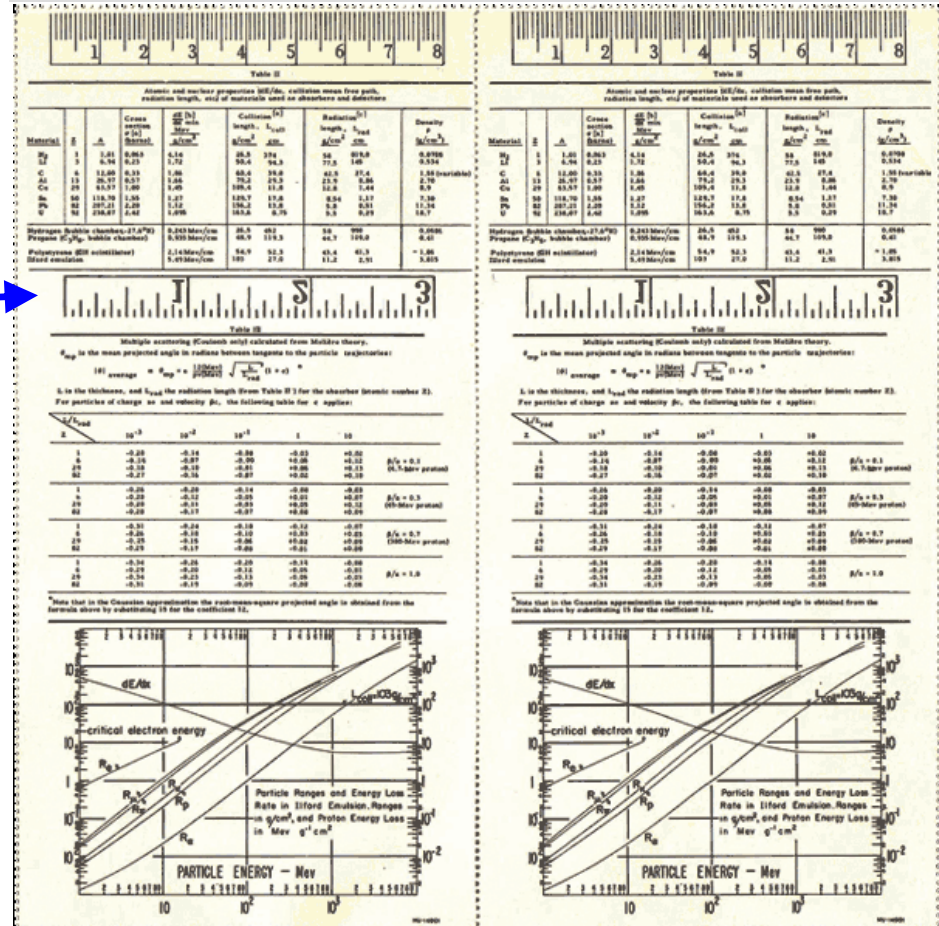


Presentation to National Academies Committee on
Assuring the Utility and Integrity of Research Data in the Digital Era
Irvine, California, September 17, 2007

The Challenge of Data Quality

- Two aspects
 - Experimental/Technical issues.
 - Human behavior issues.
- Already had difficulties reproducing ruler in 1958! →
- Particle physicists put a huge effort into maintaining the quality of data and the results derived from the data.
- This is fundamentally a hard problem: there are a lot more ways to things wrong than to do things right! (Entropy)

Particle Data Book, 1958
Lawrence Radiation Lab Report UCRL-8030



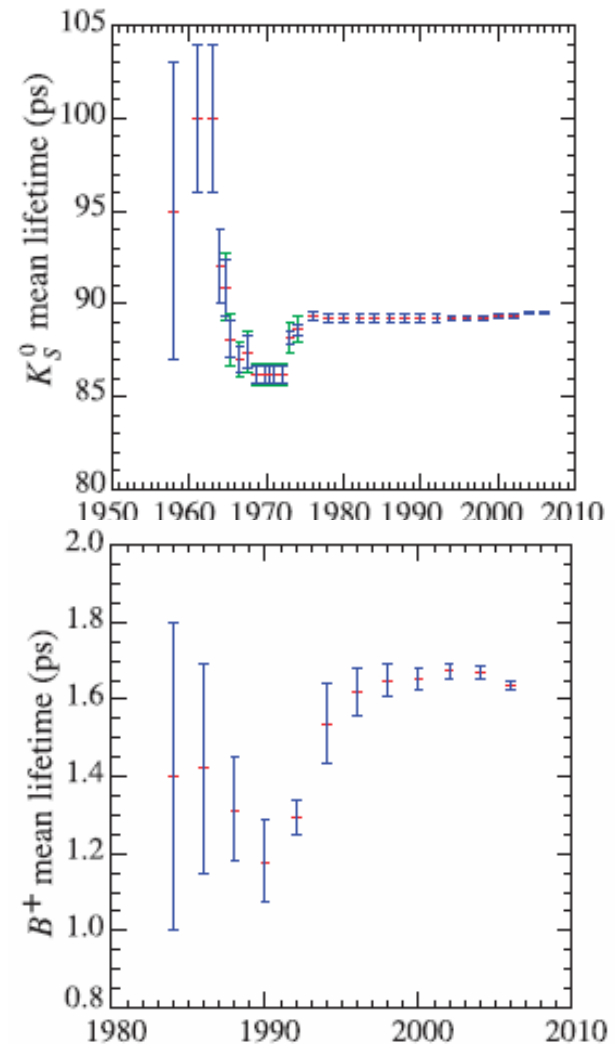
2006 edition has 1232 pages; 24,559 measurements!

a few things can go wrong...

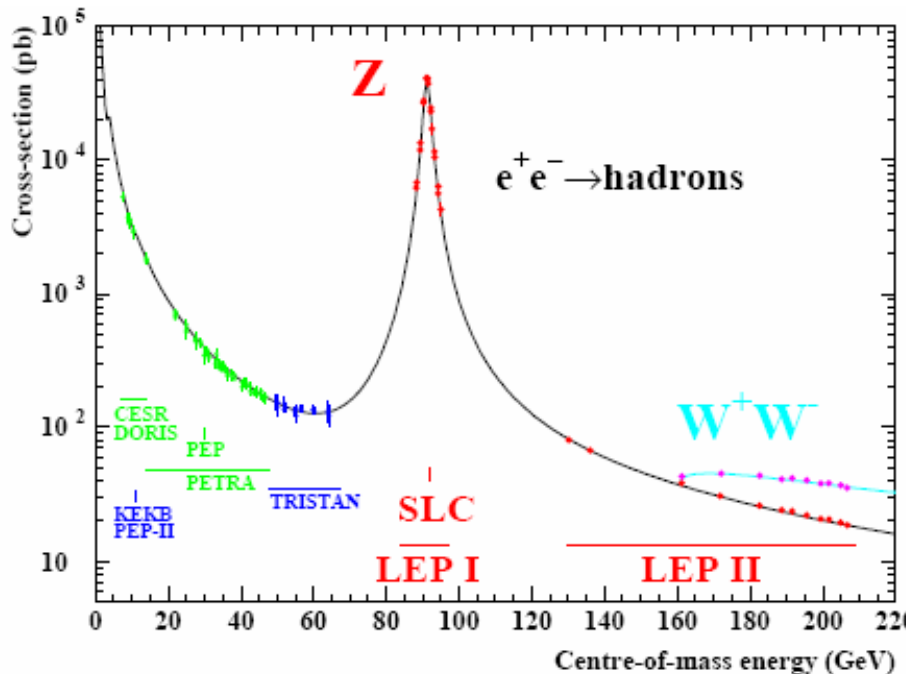
- “My answer agrees with the previous result, so it must be right.”
- “Something must be wrong with the data...the answer isn’t coming out right.”
- “We don’t need to perform a blind analysis, because we already know the answer.”
- “If this is right I could win the...”
- “Correlations?”
- “He needs to graduate now.”
- “The conference is in two weeks.”
- “Let’s see if we can enhance the significance of our signal by changing the selection requirements.”
- “If it turns out to be true, we can say we saw it first.”

I have either heard all of these or seen the direct result.

History plots from the Particle Data Group; each point is cum avg.



But there is also good in us...



Precision Electroweak Measurements on the Z Resonance

The ALEPH, DELPHI, L3, OPAL, SLD Collaborations,¹
the LEP Electroweak Working Group,²
the SLD Electroweak and Heavy Flavour Groups

Much effort was dedicated to the determination of the energy of the colliding beams. A precision of about 2 MeV in the centre-of-mass energy was achieved, corresponding to a relative uncertainty of about $2 \cdot 10^{-6}$ on the absolute energy scale. This level of accuracy was vital for the precision of the measurements of the mass and width of the Z, as described in Chapter 2. In particular the off-peak energies in the 1993 and 1995 scans were carefully calibrated employing the technique of resonant depolarisation of the transversely polarised beams [14,15]. In order to minimise the effects of any long-term instabilities during the energy scans, the centre-of-mass energy was changed for every new fill of the machine. As a result, the data samples taken above and below the resonance are well balanced within each year, and the data at each energy are spread evenly in time. The data recorded within a year around one centre-of-mass energy were combined to give one measurement at this “energy point”.

The build-up of transverse polarisation due to the emission of synchrotron radiation [16] was achieved with specially smoothed beam trajectories. Measurements with resonant depolarisation were therefore only made outside normal data taking, and typically at the ends of fills. Numerous potential causes of shifts in the centre-of-mass energy were investigated, and some unexpected sources identified. These include the effects of earth tides generated by the moon and sun, and local geological deformations following heavy rainfall or changes in the level of Lake Geneva. While the beam orbit length was constrained by the RF accelerating system, the focusing quadrupoles were fixed to the earth and moved with respect to the beam, changing the effective total bending magnetic field and the beam energy by 10 MeV over several hours. Leakage currents from electric trains operating in the vicinity provoked a gradual change in the bending field of the main dipoles, directly affecting the beam energy. The collision energy at each interaction point also depended for example on the exact configuration of the RF accelerating system. All these effects are large compared to the less than 2 MeV systematic uncertainty on the centre-of-mass energy eventually achieved through careful monitoring of the running conditions and modelling of the beam energy.

[http://lepewwg.web.cern.ch/
LEPEWWG/1/physrep.pdf](http://lepewwg.web.cern.ch/LEPEWWG/1/physrep.pdf)

Outline

Main topics

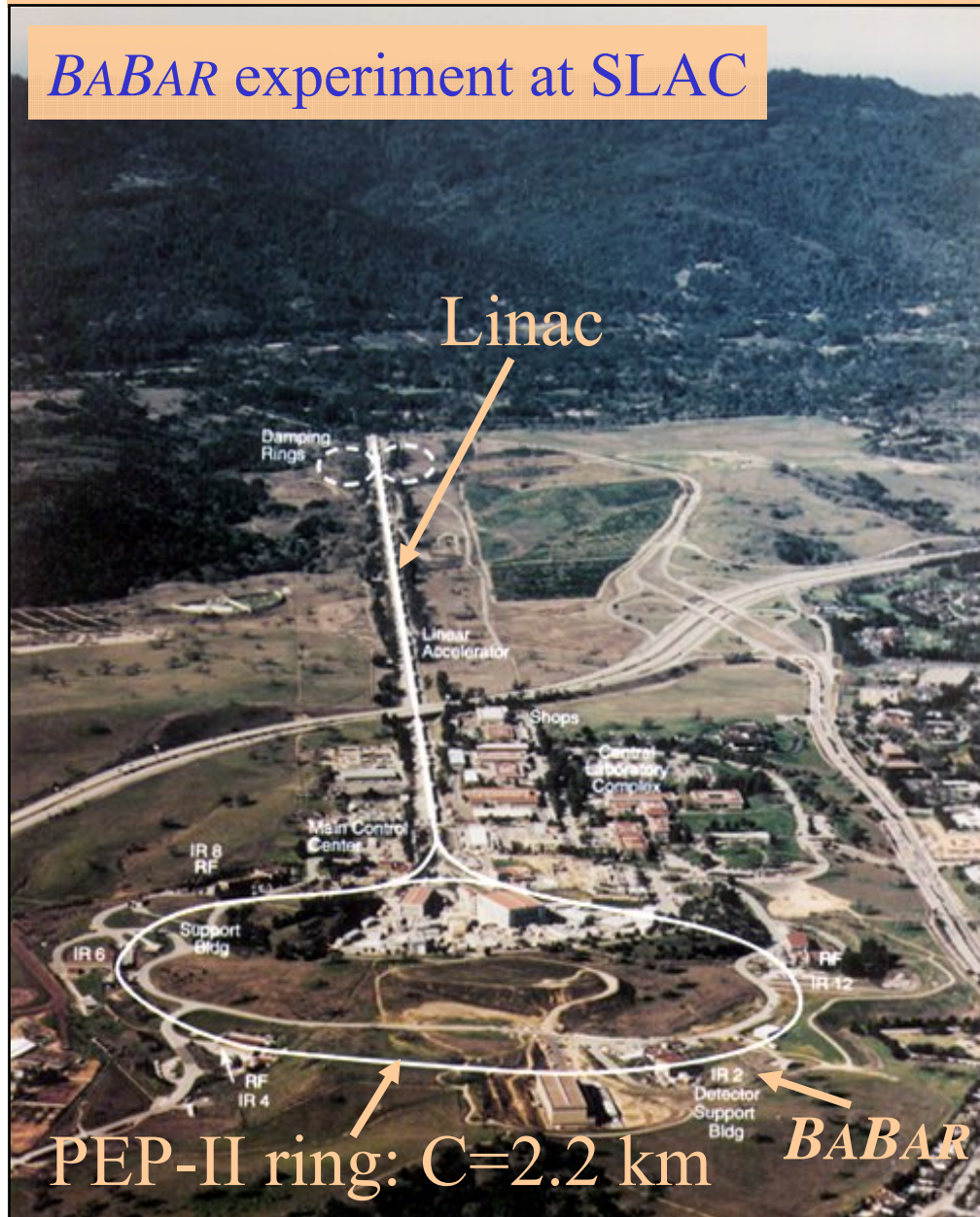
- Orientation: a few pictures of HEP experiments
- Embarrassing moments in High Energy Physics (HEP)
- What is data quality?
- How do HEP collaborations ensure data quality?
- How do high energy collaborations ensure the quality of results?
- Quick answers to panel questions
- Wisdom from Feynman

Related topics (if extra time)

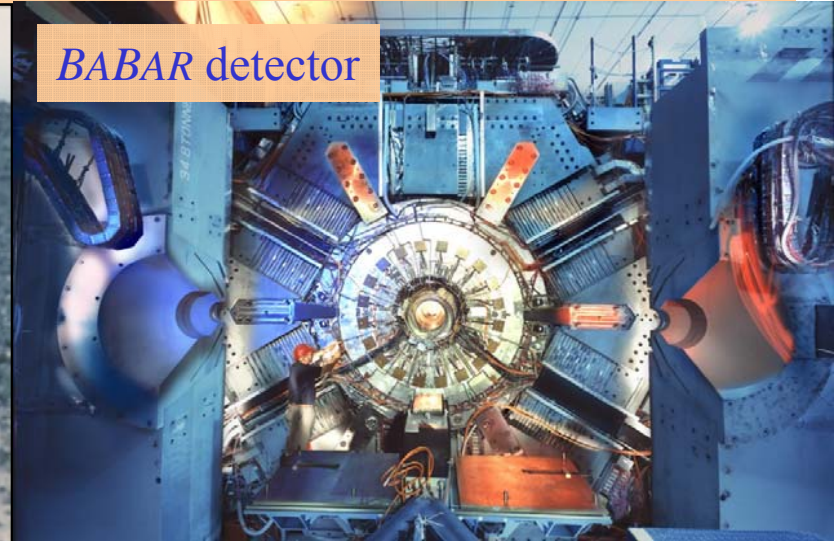
- Particle data compilations, combining scientific results from different experiments
- Long-term maintenance of data samples; public access issues

Data quality begins with high quality apparatus

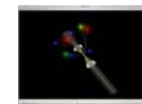
BABAR experiment at SLAC



BABAR detector



BABAR Silicon Vertex Tracker



animation

BABAR
single-event
display

2.636 GeV

electron-positron collision
point

pi+

mu-

pi-

pi+

pi-

pi+

pi+

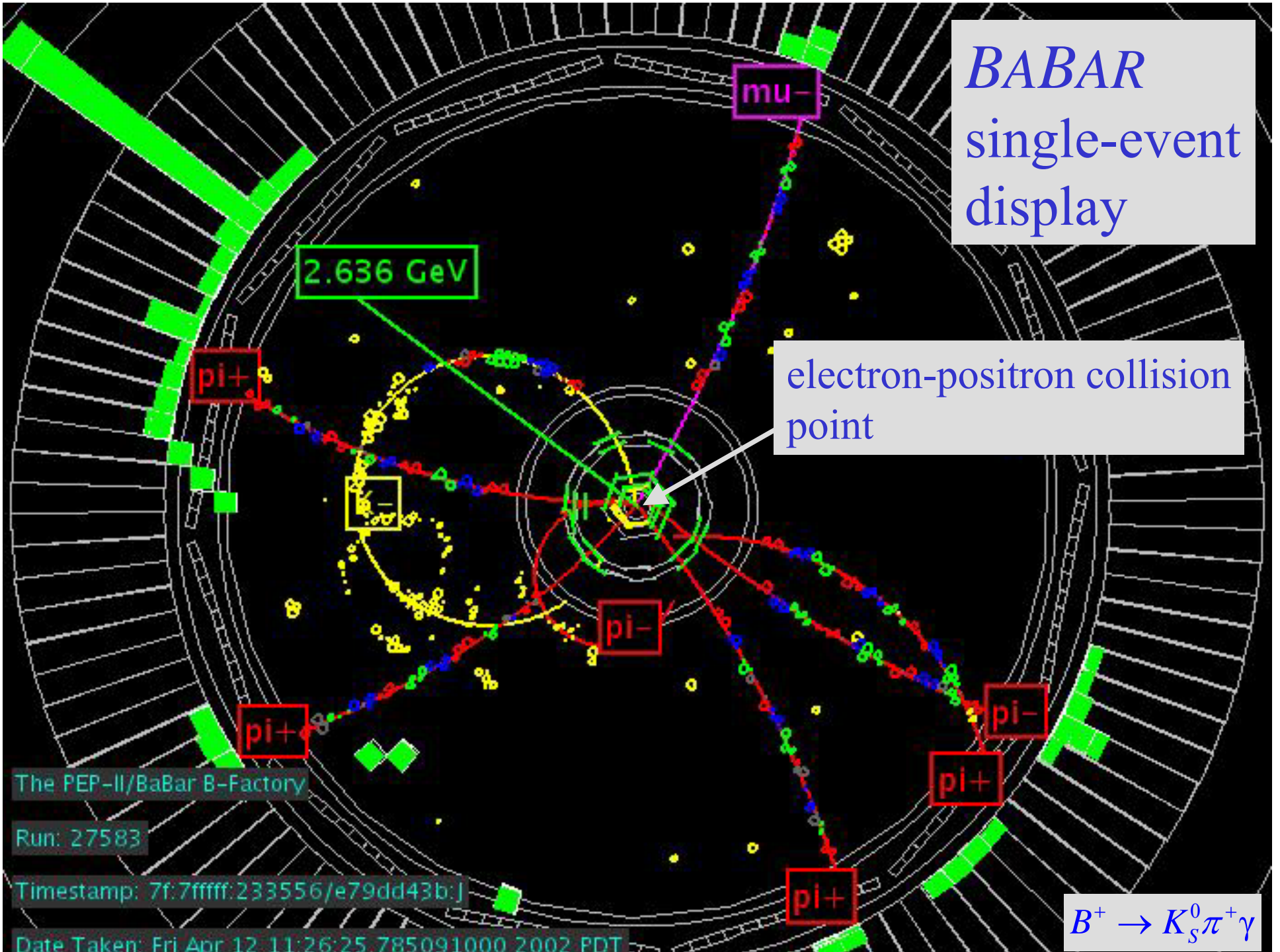
The PEP-II/BaBar B-Factory

Run: 27583

Timestamp: 7f:7ffff:233556/e79dd43b:j

Date Taken: Fri Apr 12 11:26:25.785091000 2002 PDT

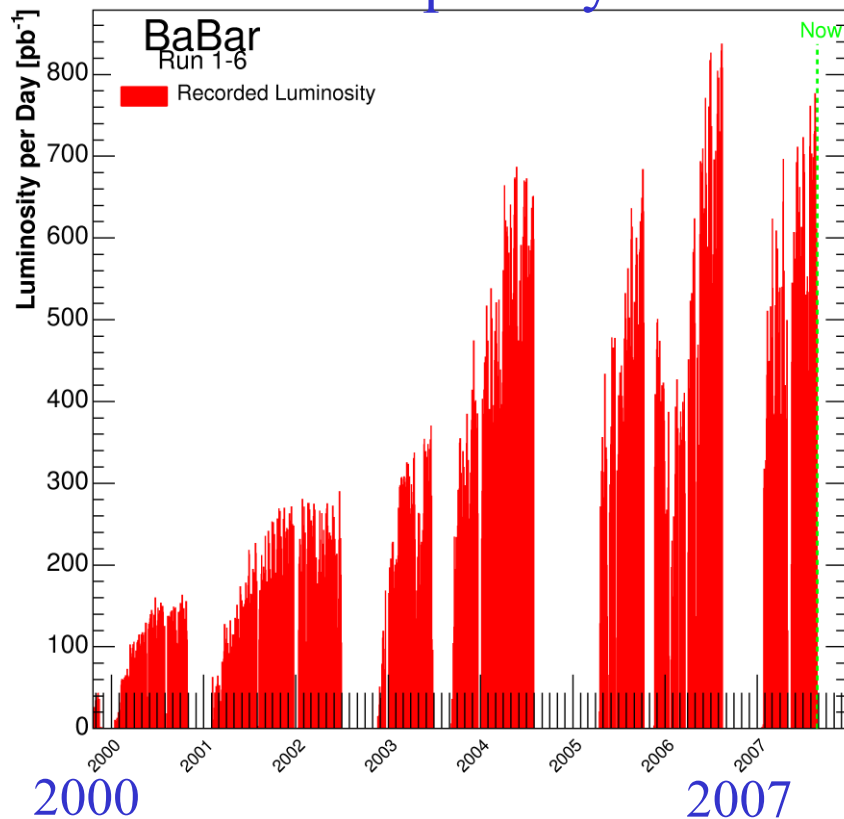
$B^+ \rightarrow K_S^0 \pi^+ \gamma$



Growth of the *BABAR* Data Sample

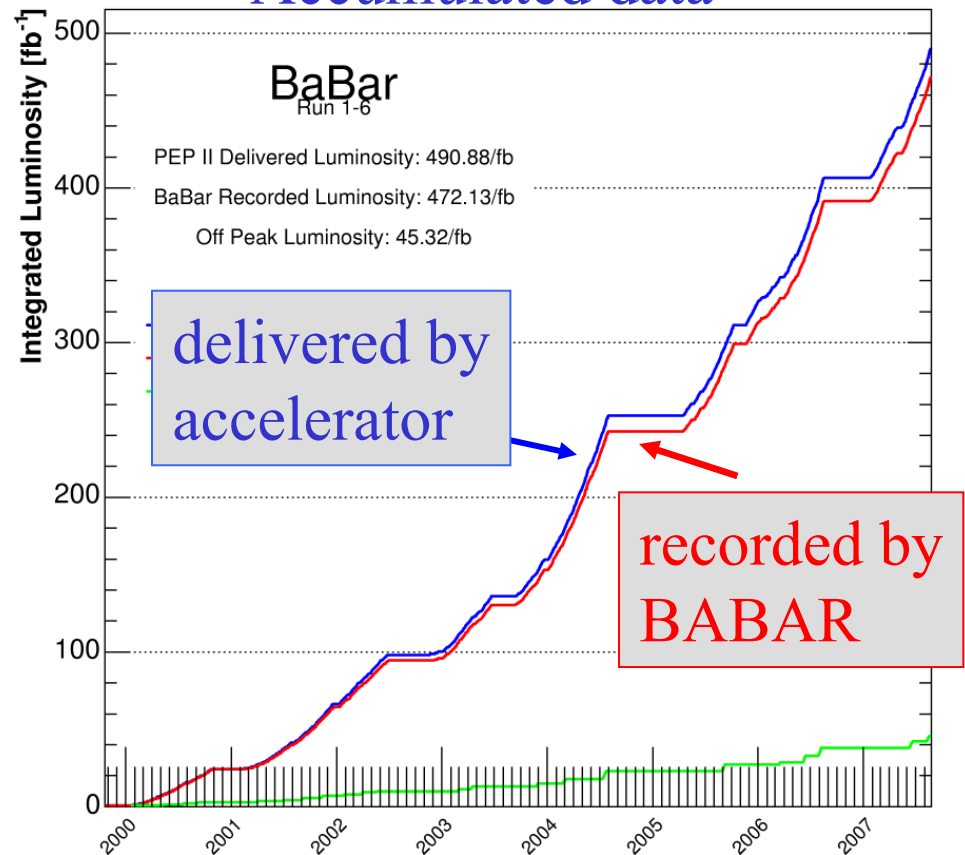
Data per day

As of 2007/08/26 00:00



Accumulated data

As of 2007/08/26 00:00



- Huge number of different particle processes in data sample.
- Published 300 journal articles so far; now about 1/week.
- Data also contain many well understood processes that can be used for calibration and crosschecking.

Characteristics of *BABAR* Data Sample

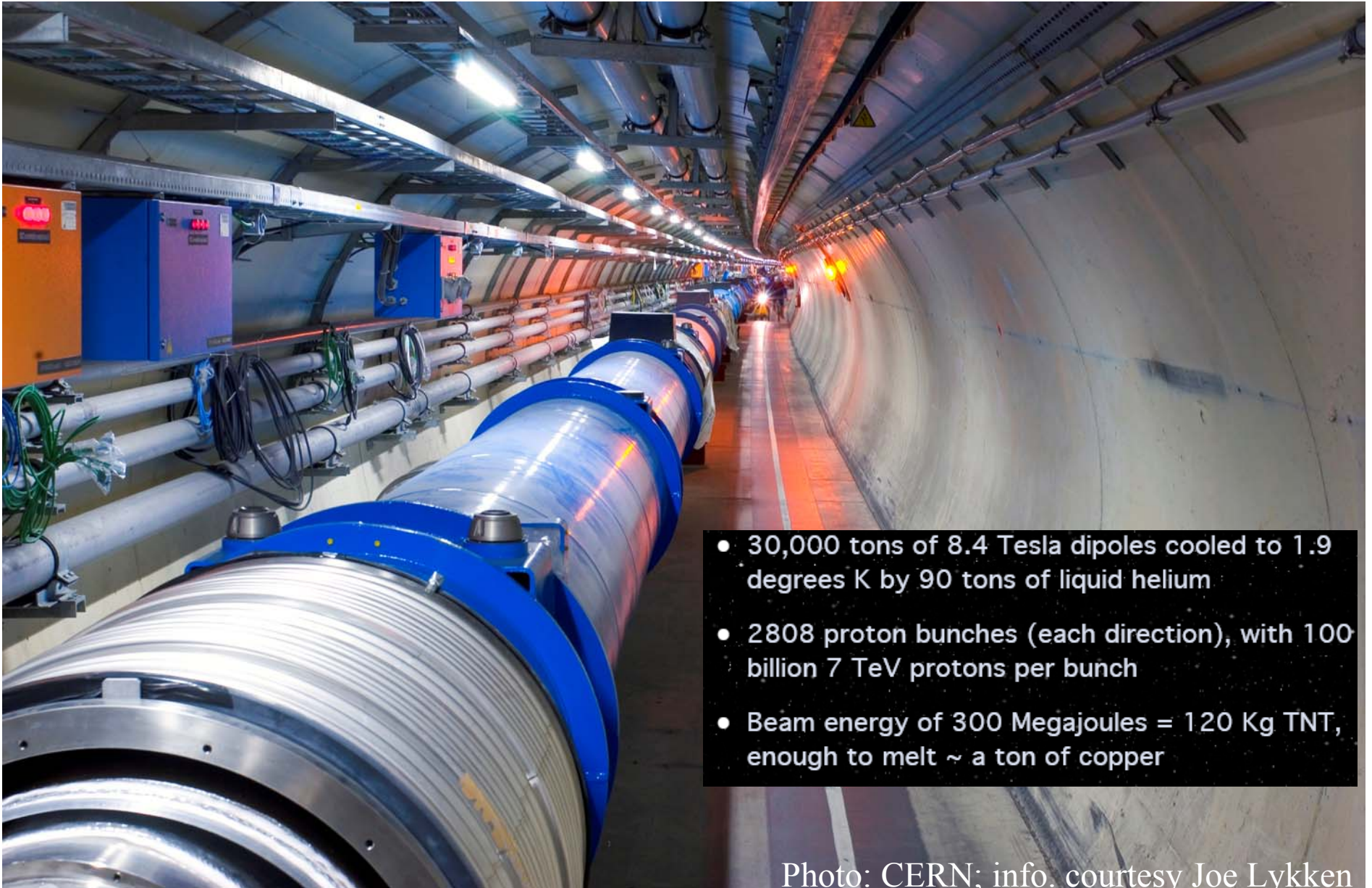
- **Data events**

- ⇒ **Running since 1999: about 32K individual run periods**
- ⇒ **Number data events (passing minimal filter): $6.6 * 10^9$**
- ⇒ **Average event size: $7.8 * 10^3$ bytes**
- ⇒ **Total sample size: $5.2 * 10^{13}$ bytes (52 TB)**

- **Simulated (“Monte Carlo”) events**

- ⇒ **Total number generic events: $4.6 * 10^9$**
- ⇒ **Total number of signal events: $4.3 * 10^9$**
- ⇒ **Total sample size: $12.1 * 10^{13}$ bytes (121 TB)**

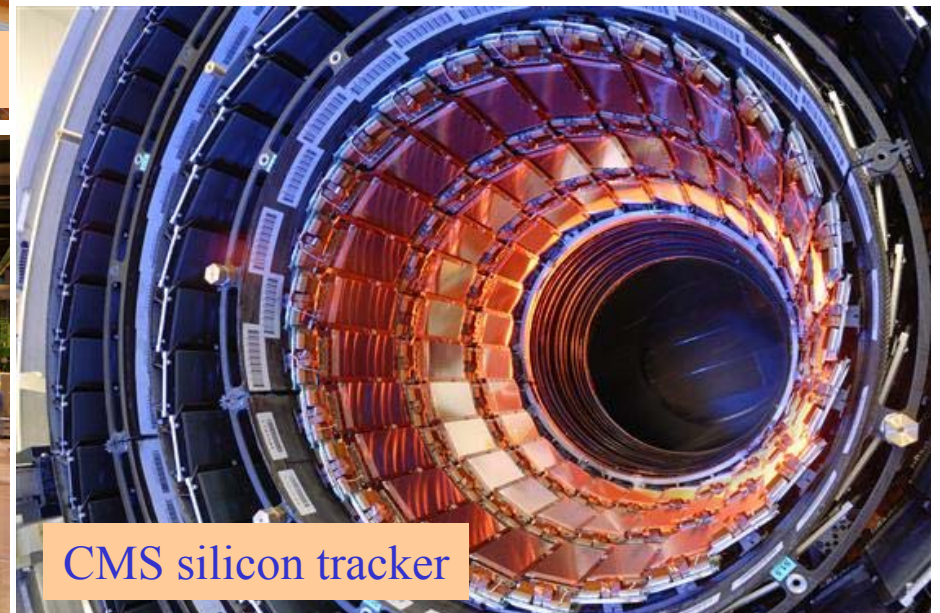
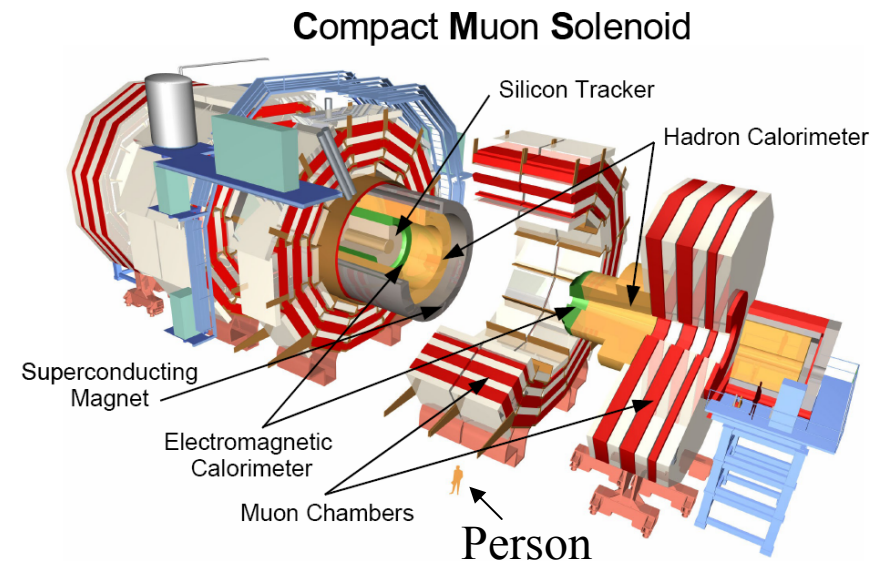
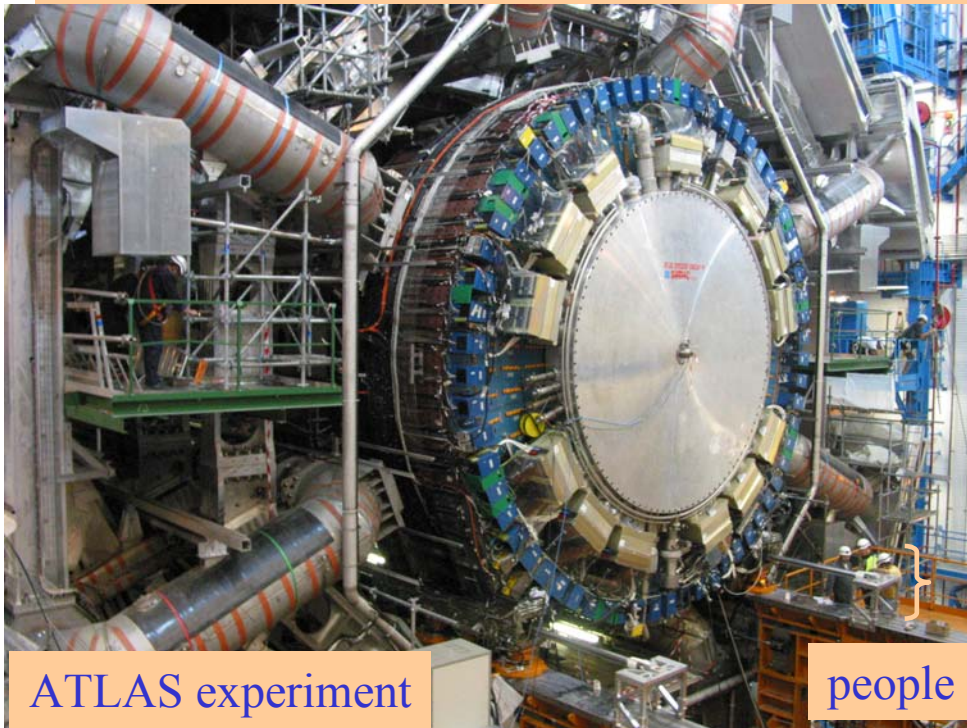
The next big thing: LHC at CERN



- 30,000 tons of 8.4 Tesla dipoles cooled to 1.9 degrees K by 90 tons of liquid helium
- 2808 proton bunches (each direction), with 100 billion 7 TeV protons per bunch
- Beam energy of 300 Megajoules = 120 Kg TNT, enough to melt ~ a ton of copper

Photo: CERN; info. courtesy Joe Lykken

ATLAS and CMS Experiments at the LHC



Embarrassing moments in particle physics

1. “Discovery” of the $\zeta(8.1)$ —Crystal Ball expt. (1984)

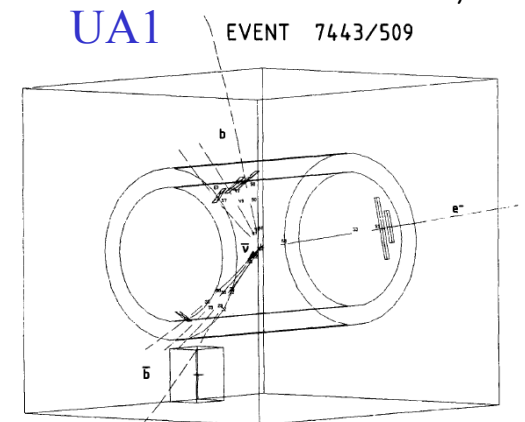
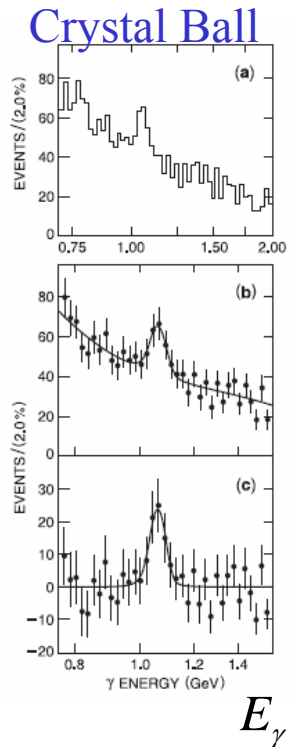
- Observation of peaks in photon-energy spectrum in two independent decay channels.
- Not confirmed in subsequent data sample
- Only presented at conferences; not published

2. “Discovery” of top quark – UA1 experiment (1984)

- Observation of 6th quark (top) incorrectly inferred from CERN experiment
- top quark finally discovered at Fermilab at much higher mass

3. “Discovery” of penta-quark states (2002-2004)

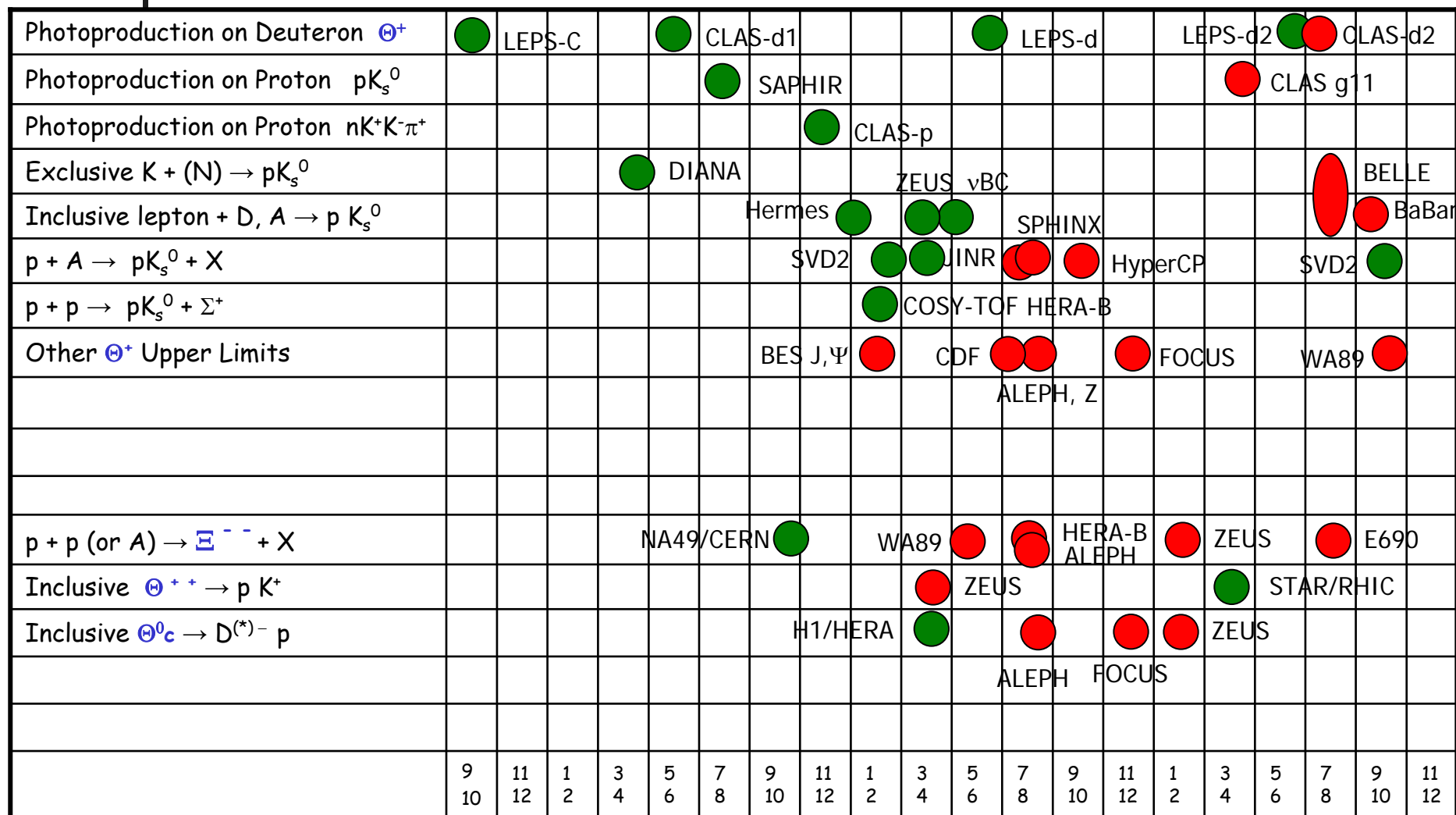
- remarkable bandwagon effect (next slide)





Slide courtesy of Reinhard Schumacher

Pentaquark Exp'ts Timeline



2002

2003

2004

2005

from Particles and Nuclei International Conference, Santa Fe, 2005

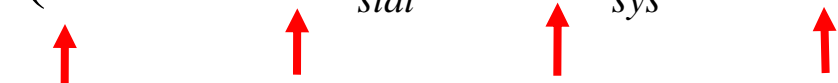
Some common problems

- People often stop looking for mistakes when they obtain a desirable result.
- Background shape or normalization estimated incorrectly.
- Backgrounds peaking under signal not correctly determined.
- Signal significance estimated incorrectly.
- Signal is created artificially as “reflection” of another signal.
- Errors determined incorrectly.
- Correlations not taken into account.
- Shapes used in fit are not adequate to describe the data.
- Bugs in program.
- Systematic errors underestimated.
- Systematic errors incomplete.
- Unstated/incorrect assumptions.
- Changes in experimental conditions not fully taken into account.
- Average of many bad measurements might not give a good measurement.

What is data quality?

Bottom Line: a scientific result based on a high-quality data sample is reproducible and unbiased, and its uncertainties are quantified.

$$\textit{MeasuredQuantity} = (4.54 \pm 0.34_{stat} \pm 0.07_{sys} \pm 0.04_{norm}) \times 10^{-3}$$

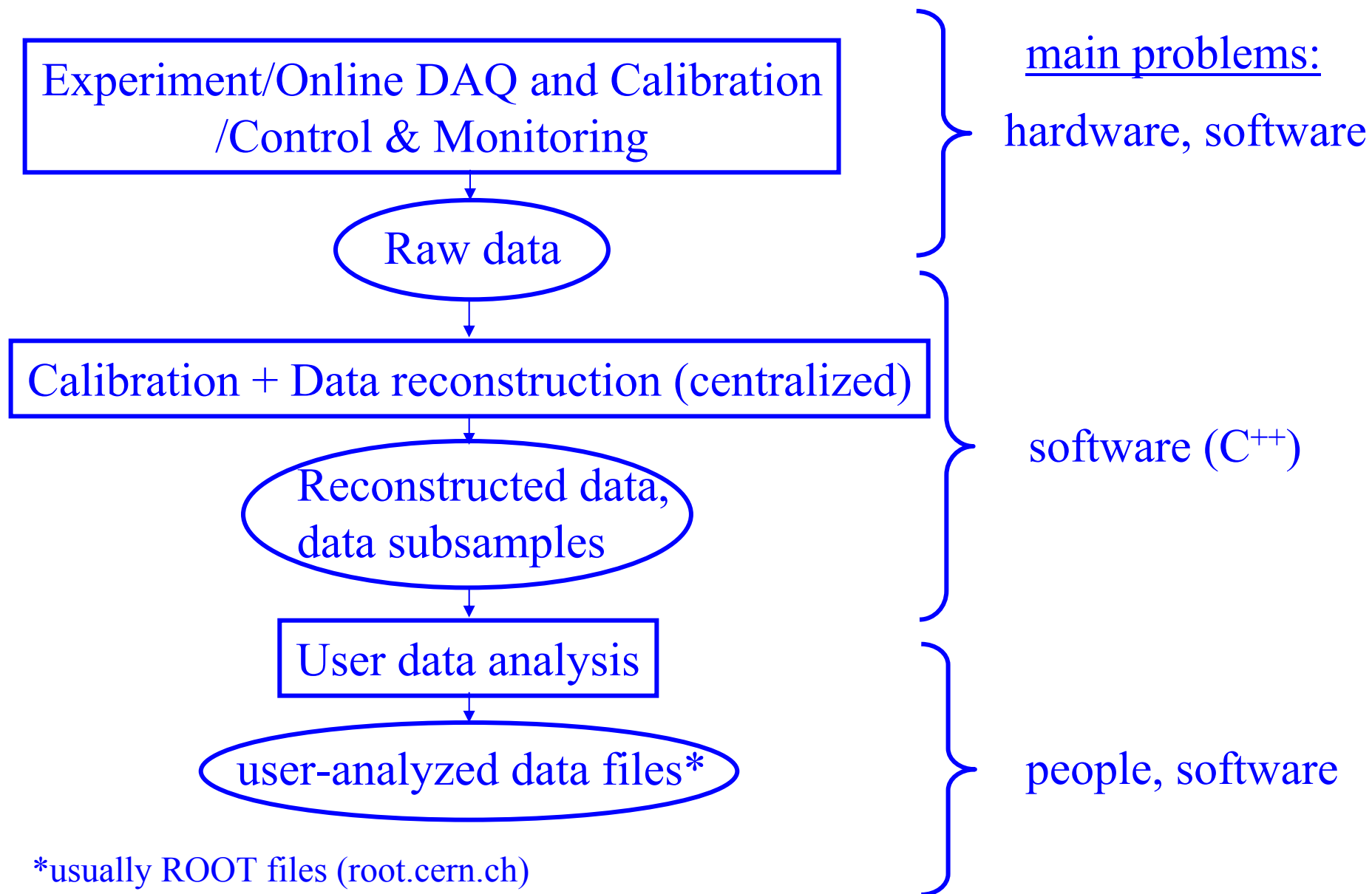

central statistical systematic systematic
value uncert. uncert. uncert. due to
external data inputs

- The individual contributions to the systematic uncertainty are stated.
- Any assumptions are stated, including numerical inputs from other experiments that might change.
- Sometimes also need to include information on error correlations.

This is a broad definition, implicitly recognizing that

- the quality of the data sample is fundamental
- but...the quality of the data analysis is critical

Stages of Data Acquisition, Processing, & Analysis



Ensuring data/result quality in HEP experiments

- Experimental conditions must be
 1. Controlled and extensively monitored (rapidly/online)
 2. Recorded/documented/archived in an automated and accessible way
- Instrument performance must be
 1. Calibrated and understood (online + offline)
 2. Recorded/documented/archived in an automated and accessible way
- Data handling and offline processing must be
 1. Controlled by automated software tools and monitored by bookkeeping tools
 2. Performed with documented software with version control
- Data analysis must be
 1. Unbiased
 2. Fully documented
 3. Subject to extensive internal review/scrutiny by experts

Maintaining data quality in *BABAR*: Data Quality Group

Data Quality Group (Hypernews bulletin board system)

(DQ already checked at some level online)

The Data Quality group forum covers data quality monitoring, reporting, planning, and tools development. The group includes efforts from conveners or representative groups. The data quality group will ensure the quality and integrity of new data, reprocessed data, monte carlo production and skim production.

Email gateway for this forum is: DQG-hn@slac.stanford.edu

⚠ The **RETIRED** messages for this forum can be found [here](#).

✚ Messages Inline Depth: ☐ 0 ☒ 1 ☐ All Outline Depth: ☐ 1 ☒ -1 ☐ +1 ☐ All

1505. [Run6 data to check for August 20 - August 26](#) by Virginia Azzolini, 8/26/07 **NEW**
1504. [Runs with SVT ODF damage](#) by [=?ISO-8859-1?Q?Jordi_Garra_Tic=F3?=>](#), 8/24/07 **NEW**
1. [Re: Runs with SVT ODF damage](#) by Nicolas Arnaud, 8/24/07 **NEW**
1503. [Weekly DQG meeting, Thursday August 23rd 2007](#) by Carlos Alberto Chavez, 8/23/07 **NEW**
1. [some short runs \(75807 & 75779\)](#) by Zafar Yasin, 8/23/07 **NEW**
- > [Re: some short runs \(75807 & 75779\)](#) by Zafar Yasin, 8/23/07 **NEW**
1502. [PR datasets tags R22c-v05 and R22d-v02 are now available](#) by Carlos Alberto Chavez, 8/21/07 **NEW**
1. [SP-Generic datasets tags R22c-v05 are now available](#) by Carlos Alberto Chavez, 8/26/07 **NEW**
1501. [Run6 data to check for August 13 - August 20](#) by Virginia Azzolini, 8/20/07 **NEW**
2. [Re: Run6 data to check for August 13 - August 20 \[BRECO&PHYS\] OK!](#) by Francesco.Renga@roma1.infn.it, 8/21/07 **NEW**
3. [Re: Run6 data to check for August 13 - August 20 \[TRK\]](#) by Yanyan Gao, 8/21/07 **NEW**
5. [Re: Run6 data to check for August 13 - August 20 \[SVT\]](#) by Joel Martinez, 8/21/07 **NEW**
7. [Re: Run6 data to check for August 13 - August 20 \[DCH & dE/dx\]](#) by Martin Simard, 8/21/07 **NEW**
9. [Re: Run6 data to check for August 13 - August 20 \[PID\]](#) by Diego Alejandro Milanes, 8/22/07 **NEW**
10. [Re: Run6 data to check for August 13 - August 20](#) by morris.570@gmail.com, 8/22/07 **NEW**
11. [Re: Run6 data to check for August 13 - August 20 \[DIRC\]](#) by justine serrano, 8/23/07 **NEW**
13. [Re: Run6 data to check for August 13 - August 20 \[TRG\]](#) by rahmat, 8/23/07 **NEW**
1. [Re: Run6 data to check for August 13 - August 20 \[TRG\]](#) by Michael Sigamani, 8/24/07 **NEW**
1500. [preliminary candidate datasets tags for R22c \(v05-c1\) and R22d \(v02-c1\) skim cycles](#) by Carlos Alberto Chavez, 8/16/07 **NEW**
1. [Re: preliminary candidate datasets tags for R22c \(v05-c1\) and R22d \(v02-c1\) skim cycles --some R22d-v02-c1 checks by](#)
- > [RE: preliminary candidate datasets tags for R22c \(v05-c1\) and R22d \(v02-c1\) skim cycles --some R22d-v02-c1 checks by](#)
- > [RE: preliminary candidate datasets tags for R22c \(v05-c1\) and R22d \(v02-c1\) skim cycles -- some R22d stragglers by Ho](#)
1499. [Final set of mini 2 mini validations](#) by Carlos Alberto Chavez, 8/16/07 **NEW**
1. [Re: Final set of mini 2 mini validations \[PHYS&BRECO\] OK!](#) by Francesco.Renga@roma1.infn.it, 8/16/07 **NEW**
2. [Re: Final set of mini 2 mini validations \[TRK\]](#) by Yanyan Gao, 8/16/07 **NEW**
4. [Re: Final set of mini 2 mini validations \[IFR\]](#) by James Morris, 8/16/07 **NEW**
5. [Re: Final set of mini 2 mini validations \[DIRC\]](#) by justine serrano, 8/16/07 **NEW**
7. [Re: Final set of mini 2 mini validations \[EMC\]](#) by Neng Xu, 8/16/07 **NEW**
9. [Re: Final set of mini 2 mini validations \[DCH & dE/dx\]](#) by Martin Simard, 8/16/07 **NEW**
11. [Re: Final set of mini 2 mini validations \[PID\]](#) by Diego Alejandro Milanes, 8/16/07 **NEW**

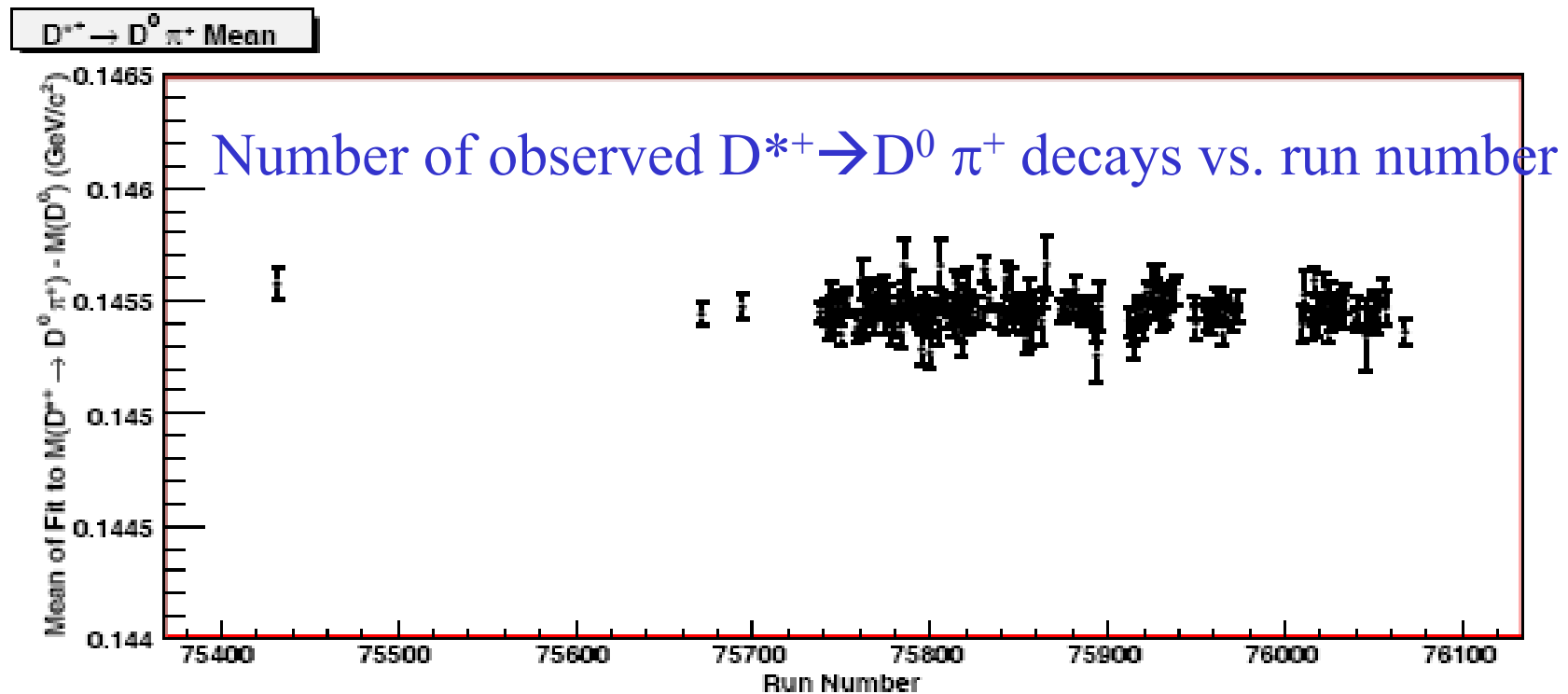
Meeting posted:
About 15 experts meet to review quality of 1 week's worth of data.
Rapid checking is very impt. to avoid problems affecting large amounts of data.

Experts from:
Management, computing, detector operations, physics analysis groups.

Find: ☒ Next ☒ Previous ☐ Highlight all ☐ Match case

Example of Data Quality Monitoring

Monitor several hundred observable high-level quantities during each run, in addition to experimental settings/conditions.



“Good” data must not selected according to whether they confirm a desired result! Completely obvious, yet it has been done.

Maintaining the quality of data analysis

- A distinct problem from “data quality”—arising from the complexity of the data sample and the data analysis procedures.
 - ⇒ Knowledge of experimental apparatus and its limitations
 - ⇒ Extensive knowledge of particle physics
 - ⇒ Knowledge of fitting/statistical methods
 - ⇒ Awareness of many pitfalls (e.g., how false signals are created).

Standardized and thoroughly tested data analysis tools

Redundancy: procedures checked using other processes in data sample.

Use of simulated data samples (“Monte Carlo”) to devise procedures without looking at the data.

Blind analysis protocols: conceal result until procedures fixed.

Rigorous requirements for review and detailed documentation.

The joy and torture of blind analysis

Blind analysis procedures have advantages and disadvantages.

- extensively used in BaBar; goal is to reduce potential bias
- applicability and usefulness depends on situation
- a complex topic worthy of a talk on its own



Late-night unblinding party in BaBar for the measurement of a matter-antimatter asymmetry.

Reference: “Blind Analysis in Nuclear and Particle Physics”, J.R. Klein & A. Roodman, Annu. Rev. Nucl. Part. Sci. 2005.55:141—63.

Life story of an internal review of a data analysis project

1. **Analysis underway: about 10 progress reports to Analysis Working Group of people performing similar studies.**
2. **Produce detailed internal document.**
3. **Formal reading of internal document by Analysis Working Group and signoff.**
4. **Appointment of Internal Review Committee (3-4 people)**
5. **Review and signoff of journal article by Internal Review Committee**
6. **Presentation to full collaboration.**
7. **Review of journal article by designated member institutions.**
8. **Electronic posting of questions and authors' responses.**
9. **Signoff on responses by review committee.**
10. **Final review by Publications Board; one-week final notice to the collaboration.**
11. **Submission of article to journal.**

Time from start to finish: 1-3 years, mostly Steps 1-3.


Management of Data Analysis: BaBar Analysis Information System (BAIS)

BAIS: Analysis Index - Mozilla Firefox

File Edit View History Bookmarks Tools Help

https://www.slac.stanford.edu/babar-internal/BAIS/info/index.html

BaBar Home Computing Detector Documentation Hypernews Organization Physics BaBar Web Search Meeting Organizer

 **BAIS: Analysis Index**

HN userid: "nichman"

Help Analysis Index Analysis Detailed Index AWG Index IRG Index Admin & Maintenance

Create new analysis entry Create new AWG Review Committees Analyst List Reviewer List

Potential Reviewer List RCLog Time Alerts

Analysis Index for all AWGs

BAIS contains 745 analyses.

745 data analysis projects!

Please either enter a search query here :

We are testing a new way of selecting BAIS entries using a pseudo-SQL command to be entered by the user. If you decide to try it, please send comments or problem reports to ecmartin@slac.stanford.edu. The usual BAIS sort and selection box remains unchanged. Thank you.

Analysis Search Form

Enter search query :

For example : find (awg = charm or cwrstart > 01-jan-1900) and name = mixing
or find awg = charm order by pubstat (note: order by can be used with awg, pubstat, name and confstat)

Search by	AWG	AWG code	Analysis Name	Analyst(s)	Publication Status	Target Pub Period	Target Journal	Conference Status
Code	awg	code	name	analyst	pubstat	targetpub	journal	confstat
Options	<input type="text" value="breco"/>	<input type="text" value="awg-yy/xx"/>	<input type="text" value="anything"/>	<input type="text" value="anything"/>	<input type="text" value="Free"/>	<input type="text" value="4/08"/>	<input type="text" value="PRL"/>	<input type="text" value="RC/CONF"/>

Search by	Target Conference	CWR start	CWR end	Keyword	RC first meeting	PAC sign-off	RClog EET	BAD #	IRG	Submission	Final Notice Start	Final Notice End
Code	targetconf	cwrstart	cwrend	key	firstmtg	signoff	eet	bad	irg	sub	fnstart	fnend
Options	<input type="text" value="CKM 2008"/>	<input type="text" value="dd-mmm-yyyy or 0 (empty date)"/>	<input type="text" value="dd-mmm-yyyy or 0 (empty date)"/>	<input type="text" value="anything"/>	<input type="text" value="dd-mmm-yyyy or 0 (empty date)"/>	<input type="text" value="yes/no"/>	<input type="text" value="yes/no"/>	<input type="text" value="number"/>	<input type="text" value="ex: 1a"/>	<input type="text" value="dd-mmm-yyyy or 0 (empty date)"/>	<input type="text" value="dd-mmm-yyyy or 0 (empty date)"/>	<input type="text" value="dd-mmm-yyyy or 0 (empty date)"/>

Or enter selection & sort criteria here :

Analysis Selection & Sort Form

Sort by: ☒ AWG Code ☐ Publication Status ☐ Analysis Name ☐ Conference Status

Sort order: ☒ Ascending ☐ Descending

Select AWG:

Data analysis projects in one subgroup

17	SemiLep-05/04	Weak annihilation in $B0 \rightarrow X1 \nu$	Gagliardi, Nicola ; Rotondo, Marcello ; Simonetto, Franco	AWG/RC	3/07	PRL	Submitted	LP2007	ICHEP, DPF, WINTER2007, WSTAT07-Q, SUMMER2007, R18, SSTAT07-L, LP07-3
18	SemiLep-05/07	Exclusive semileptonic $b \rightarrow u$ using neutrino reconstruction	Dingfelder, Jochen Christian ; Kelsey, Michael H. ; Luth, Vera	AWG/RC	2/07	PRD		LP2007	semilep, WINTER2007, WSTAT07-Q, SUMMER2007, R18
19	SemiLep-05/12	B \rightarrow Ds K l ν	Jasper, Heiko	AWG/RC	3/07		RC/PhysNote		ICHEP, WINTER2007, WSTAT07-Q, R18
20	SemiLep-05/15	Ds semileptonic decays	Oyanguren, Arantza ; Roudeau, Patrick ; Serrano, Justine	AWG/RC	3/07	PRI			
21	SemiLep-06/01	D \rightarrow pi l ν	Oyanguren, Arantza ; Roudeau, Patrick	AWG/RC	3/07				
22	SemiLep-06/02	Exclusive B- \rightarrow Xu l ν with Hadronic Tags	Bianchi, Fabrizio ; D'Orazio, Alessia ; Gallo, Francesco ; del Re, Daniele	AWG/RC	2/07	PRD-RC	Submitted	LP2007	semilep, ICHEP, HQ8, DPF, DPF06-BPHY22, AS07-1, LL07-3, SUMMER2007, R18
23	SemiLep-06/15	Exclusive B-\rightarrowD/D*(*)pi l ν with hadronic tags	Battaglia, Marco ; Lopes Pegna, David	AWG/RC	3/07	PRD	Submitted	LP2007	SUMMER2007, R22, SSTAT07-L, EPS07-4, LP07-3
24	SemiLep-06/16	b-\rightarrowclnu mixed had. moments Nx	Klose, Verena ; Schubert, Klaus R. ; Sundermann, Jan E.	AWG/RC	3/07	PRL		EPS2007	SUMMER2007, R18, SSTAT07-O
25	SemiLep-07/01	Improved Q^2-El analysis: q^2 spectra and Vub	Lacker, Heiko M. ; Volk, Alexei	AWG/RC	2/07				
26	SemiLep-04/02	D*0lnu BF and Vcb	Schubert, Jens	Conf/prelim	2/07	PRI			
27	SemiLep-05/11	b-\rightarrowclnu Hadronic Moments	Sundermann, Jan E.	Conf/prelim	2/07	EPJ			
28	SemiLep-05/13	Mixed hadronic energy and mass moments in B-\rightarrowXclnu	Klose, Verena ; Sundermann, Jan E.	Conf/prelim	3/07				
29	SemiLep-03/02	B \rightarrow D(*) tau nu branching fraction using Breco tags	Mazur, Michael Alan ; Richman, Jeffrey	CWR	2/07	PRI			
30	SemiLep-06/04	Inclusive b-\rightarrow u l ν studies with Hadronic Tags	Azzolini, Virginia ; Bozzi, Concezio ; Lacker, Heiko M. ; Menges, Wolfgang ; Petrella, Antonio ; Sacco, Roberto ; Tackmann, Kerstin ; del Re, Daniele	FN/FR	2/07	PRL			2007, R18, 07-3, LP07-3
31	SemiLep-04/05	B0 \rightarrow D*+lnu FF and Vcb	Bomben, Marco ; Cossutti, Fabio ; Della Ricca, Giuseppe	SUB	3/06	PRD			ICHEP, HQ8, DPF, 07-3, 2007, 3

In Analysis Working Group has Review Committee

Preliminary public result approved

In Collab. Wide Review

Final 1-week notice

Submitted to journal



BAIS: SemiLep-03/02 Details

HN user:
"nichman"

- Help
- Analysis Index
- Analysis Detailed Index
- AWG Index
- IRG Index
- Admin & Maintenance
- Create new analysis entry
- Create new AWG
- Review Committees
- Analyst List
- Reviewer List
- Potential Reviewer List
- RClog Time Alerts

B -> D(*) tau nu branching fraction using Breco tags — Analysis Details

Quick links to other analyses in this AWG: [SemiLep-01/01](#), [SemiLep-01/02](#), [SemiLep-01/03](#), [SemiLep-02/01](#), [SemiLep-02/02](#), [SemiLep-02/03](#), [SemiLep-02/04](#), [SemiLep-02/05](#), [SemiLep-02/06](#), [SemiLep-02/07](#), [SemiLep-02/08](#), [SemiLep-03/01](#), [SemiLep-03/03](#), [SemiLep-03/04](#), [SemiLep-03/05](#), [SemiLep-03/06](#), [SemiLep-03/07](#), [SemiLep-03/08](#), [SemiLep-03/09](#), [SemiLep-03/10](#), [SemiLep-03/11](#), [SemiLep-04/02](#), [SemiLep-04/03](#), [SemiLep-04/04](#), [SemiLep-04/05](#), [SemiLep-04/06](#), [SemiLep-04/07](#), [SemiLep-04/08](#), [SemiLep-04/09](#), [SemiLep-05/01](#), [SemiLep-05/02](#), [SemiLep-05/03](#), [SemiLep-05/04](#), [SemiLep-05/05](#), [SemiLep-05/06](#), [SemiLep-05/07](#), [SemiLep-05/08](#), [SemiLep-05/09](#), [SemiLep-05/10](#), [SemiLep-05/11](#), [SemiLep-05/12](#), [SemiLep-05/13](#), [SemiLep-05/14](#), [SemiLep-05/15](#), [SemiLep-05/16](#), [SemiLep-06/01](#), [SemiLep-06/02](#), [SemiLep-06/03](#), [SemiLep-06/04](#), [SemiLep-06/05](#), [SemiLep-06/06](#), [SemiLep-06/07](#), [SemiLep-06/09](#), [SemiLep-06/10](#), [SemiLep-06/11](#), [SemiLep-06/12](#), [SemiLep-06/13](#), [SemiLep-06/14](#), [SemiLep-06/15](#), [SemiLep-06/16](#), [SemiLep-07/01](#)

AWG Code	Analysis Name	Description	Data Information	Schedule/Timeline	Updated (By)	Created (By)
SemiLep-03/02						
<div>Publication Status</div> <div>SUB</div>	<div>Conference Status</div> <div>Submitted</div>	B -> D(*) tau nu branching fraction using Breco tags	<div>Hadronic Breco events are used to look for D(*) tau nu recoiling against the Breco. The tau is reconstructed in the leptonic decay, and both D and D* are considered. The analysis is sensitive to non-SM enhancements of the D(*) tau nu BF.</div> <div>Sources: CM2</div> <div>Samples: Run1 Run2 Run3 Run4</div> <div>Additional comment:</div>	AWG review in Feb 07, followed by RC.	04 September 2007 14:17 (davidk)	22 January 2004 16:56 (kowalews)
Analysts	Mazur, Michael Alan ; Richman, Jeffrey					
Review Committee	comm344 [HN: rev-SemiLep-03-02] (06 May 2007...) Members: Honscheid, Klaus , Kelsey, Michael H. (chair), Simonetto, Franco [View journal article RClog]					
Target Publication Period	2nd Quarter 2007					
Target Journal	Physical Review Letters (PRL)					
Target Conference	2007 Lepton-Photon 2007 (LP2007)					
Keywords	semilep, SUMMER2007, SSTAT07-O, EPS07-4, LP07-3, CORE					
	Result type:	Journal article				
	Review Committee:	comm344 [HN: rev-SemiLep-03-02] (06 May 2007...) Members: Honscheid, Klaus , Kelsey, Michael H. (chair), Simonetto, Franco				
	Contact author:	Mazur, Michael Alan				
	CWR period:	16 August 2007 to 24 August 2007				
	Final notice period:	4 September 2007 to 11 September 2007				
	Final reader(s):	Kirkby, David , Prell, Soeren A.				
	Primary BAD:	BAD 1832 , version 8				
	Authors list:	Mazur, Michael Alan , Richman, Jeffrey				

Quick responses to panel questions (1, 2)

1. **With which large data collections are you familiar and what has been your involvement with these collections?**

↪ **28 years working on 7 high-energy physics (HEP) experiments; variety of leadership and coordination roles; experience with large and complex data samples; Physics Coordinator of BaBar experiment; consultant to Particle Data Group.**

2. **What lessons can be learned from large data collections relative to how they can best be utilized and in maintaining their integrity?**

Fundamentals

- ↪ **(1) documentation, documentation, documentation**
- ↪ **(2) rapid, extensive, & transparent internal review at multiple stages**
- ↪ **(3) Data Quality (DQ) assurance as an explicit task with planning, resources, & recognition**
- ↪ **(4) DQ coordination across organizational boundaries (big science issue)**
- ↪ **(5) “strong” data sample has numerous self-calibration features**
- ↪ **(6) use of simulated data is extremely powerful tool**
- ↪ **(7) transparent and robust bookkeeping, software version control, etc.**
- ↪ **(8) attitude/culture: eternal vigilance, skepticism, and redundant checking**

Quick responses to panel questions (3)

3. What are the standards and expectations in your field with regard to data quality (e.g. peer review), communication with other scholars (e.g. making data available to other researchers), and providing data in the public domain? Are standards and expectations changing? What are the barriers to providing access (e.g. cost)?
 - ↪ Standards are generally high and have been improving over time.
 - ↪ Quality ultimately due to competition/redundancy. Large number of collaborators & competing experiments; aggressive culture in high energy physics.
 - ↪ BUT errors do occur, usually misjudgements/overoptimism/bias in data analysis, not technical/hardware failure.
 - ↪ HEP data samples are very complex; “owned” by collaboration of institutions who have defined responsibilities in building/maintaining/operating the experiment.
 - ↪ Public access/uses: educational (maybe), scientific (hard).

Quick responses to panel questions (4)

4. What are your thoughts on the questions in the committee's charge? What recommendations would you make?
 - ↪ Charge seems vague; relies on committee's expertise to sharpen the issues. (Maybe that's OK.)
 - ↪ Proposing universal standards will be difficult; maybe it's more useful to think in terms of "best practices."
 - ↪ Value: different scientific fields may be able to learn from each other's experience. Try to identify commonalities and recognize differences. Discuss successes and failures. Lessons learned.
 - ↪ Comment: emphasis on data may de-emphasizes importance of data analysis procedures. In HEP, it's a long way from the sample to the result.

Combining results from experiments & archiving results

- **Key document of particle physics community: “Review of Particle Physics”**

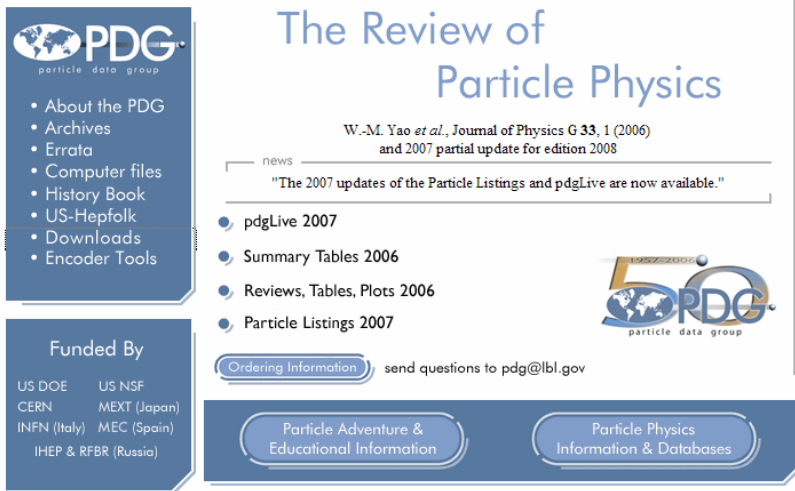
- ✚ **pocket edition: Particle Data Booklet**
- ✚ **compiled by the Particle Data Group (LBNL)**
- ✚ **many contributors throughout the HEP community; I have been a consultant.**
- ✚ **“encoders” read all HEP papers and incorporate into appropriate section of book.**

- ✚ **Averaging of results is performed.**

- **Averaging groups: need to understand correlations between uncertainties of results from different experiments**

- ✚ **Heavy Flavor Averaging Group:**
<http://www.slac.stanford.edu/xorg/hfag/>
- ✚ **LEP electroweak working group:**
<http://lepewwg.web.cern.ch/LEPEWWG/>

mirror sites: USA (LBNL) - Brazil - CERN - Indonesia - Italy - Japan (KEK) - Russia (Novosibirsk) - Russia (Protvino) - UK (Durham)



PDG
particle data group

- About the PDG
- Archives
- Errata
- Computer files
- History Book
- US-Hepfolk
- Downloads
- Encoder Tools

The Review of Particle Physics

W.-M. Yao *et al.*, *Journal of Physics G* **33**, 1 (2006)
and 2007 partial update for edition 2008

news
"The 2007 updates of the Particle Listings and pdgLive are now available."

- pdgLive 2007
- Summary Tables 2006
- Reviews, Tables, Plots 2006
- Particle Listings 2007

Ordering Information: send questions to pdg@lbl.gov

Funded By

US DOE	US NSF
CERN	MEXT (Japan)
INFN (Italy)	MEC (Spain)
IHEP & RFBR (Russia)	

Particle Adventure & Educational Information

Particle Physics Information & Databases

Copyright Information: This page and all following are copyrighted by the Regents of the University of California

Long-term preservation of HEP data samples

- If you ask a high-energy physicist whether it would be practical to preserve data samples for use far into the future, the answer most likely would be “you’ve got to be kidding!”
- Why?
 - ⇒ Extensive knowledge required to use the data sample.
 - ⇒ The data sample does not really exist independently of the software required to analyze it. You would also need to maintain the software.
 - ⇒ Since analyzing the data is so difficult, who would want to do it?
 - ⇒ Probably most of what can be learned has already been learned.
 - ⇒ Large computing resources would be required to store and use the samples.
- But it is still worth considering
 - ⇒ In several cases, the data samples are “canonical” and may never be surpassed given resource limitations.
 - ⇒ There are probably a few surprises lurking in these samples.

Wisdom from Feynman (from “Cargo Cult Science,” in “*Surely You’re Joking Mr. Feynman!*”)

- “But there is *one* feature I notice that is generally missing in cargo cult science. That is the idea that we all hope you have learned in studying science in school---we never explicitly say what this is, but just hope that you catch on by all the examples of scientific investigation. It is interesting, therefore, to bring it out now and speak of it explicitly. It’s a kind of scientific integrity, a principle of scientific thought that corresponds to a kind of utter honesty—a kind of leaning over backwards.”
- “Details that could throw doubt on your interpretation must be given, if you know them.”
- “In summary, the idea is to try to give *all* the information to help others judge the value of your contribution; not just the information that leads to judgement in one particular direction or another.”
- “But this long history of learning how to not fool ourselves—of having utter scientific integrity—is, I’m sorry to say, something that we haven’t specifically included in any particular course that I know of.”
- “The first principle is that you must not fool yourself—and you are the easiest person to fool.”

Statement of Task (I)

- **What are the growing varieties of research data? In addition to issues concerned with the direct products of research, what issues are involved in the treatment of raw data, pre-publication data, materials, algorithms, and computer codes?**
- **Who owns research data, particularly that which results from federally funded research? Is it the public? The research institution? The lab? The researcher?**
- **To what extent is a scientist responsible for supplying research data to other scientists (including those who seek to reproduce the research) and to other parties who request them? Is a scientist responsible for supplying the data, algorithms and computer codes to other scientists who request them?**
- **What challenges does the science and technology community face arising from actions that would compromise the integrity of research data? What actions should the science and technology community, journal publishers, funding agencies and universities take in response?**

Statement of Task (II)

- **What are the current standards for accessing and maintaining research data, and how should they evolve in the future? How might such standards differ for federally-funded and privately-funded research, and for research conducted in academia, industry, governmental and non-governmental organizations?**