Some not-so-pretty admissions about dealing with Internet measurements

Vern Paxson

ICSI Center for Internet Research (ICIR)
International Computer Science Institute
Berkeley, CA

vern@icir.org

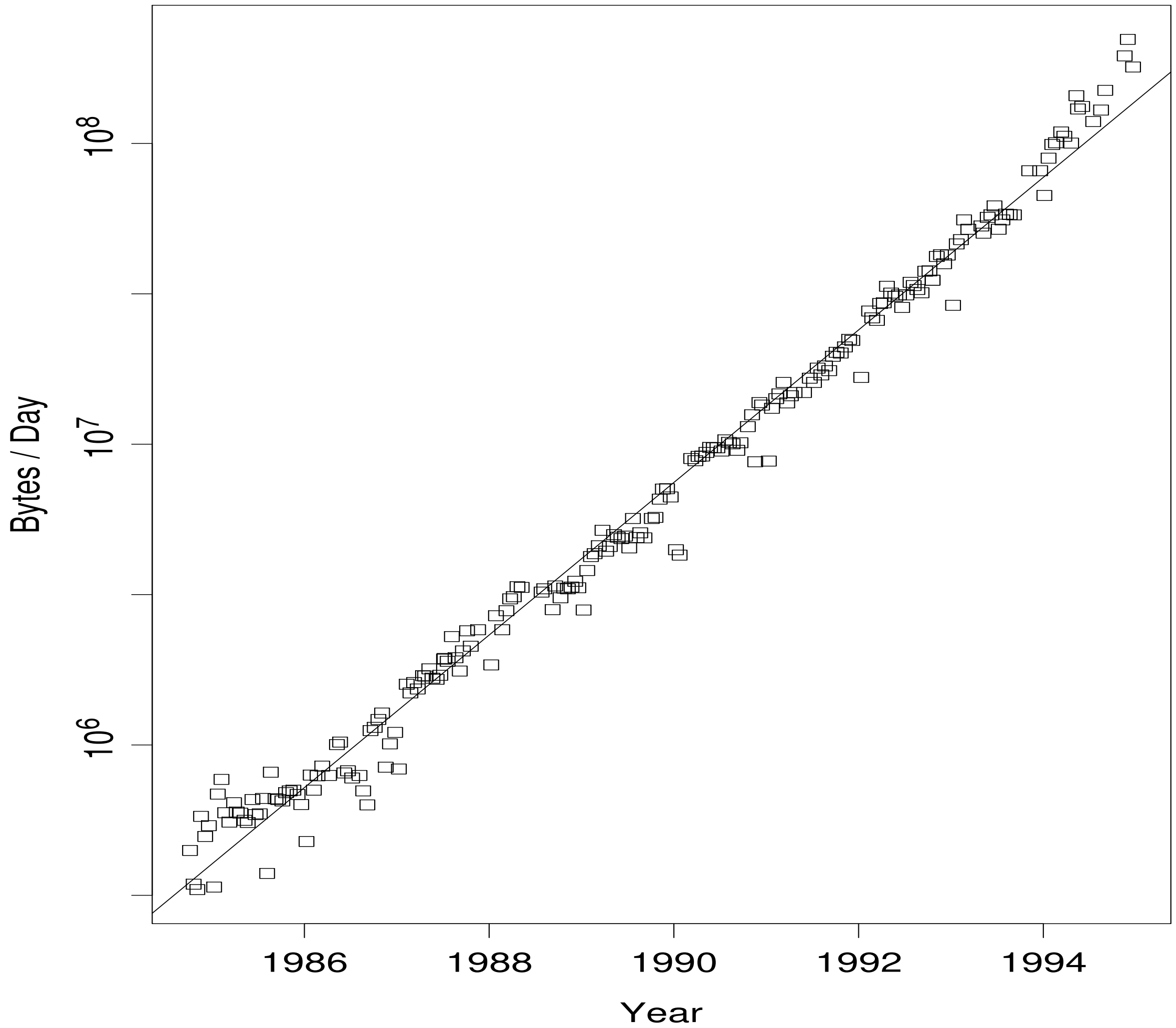
February 27, 2003

Disclaimers:

- There are no new research results in this talk.
- Many of the problems discussed are <u>mundane</u>.
- The *point* concerns the need for <u>discipline.</u>
  We know how fun that is.
- Much of this also applies to Internet simulation.
- Much of this also applies to large-scale systems work in general.

Admission #1: longitudinal data is highly ad hoc.

USENET Bulletin Board Traffic Volume

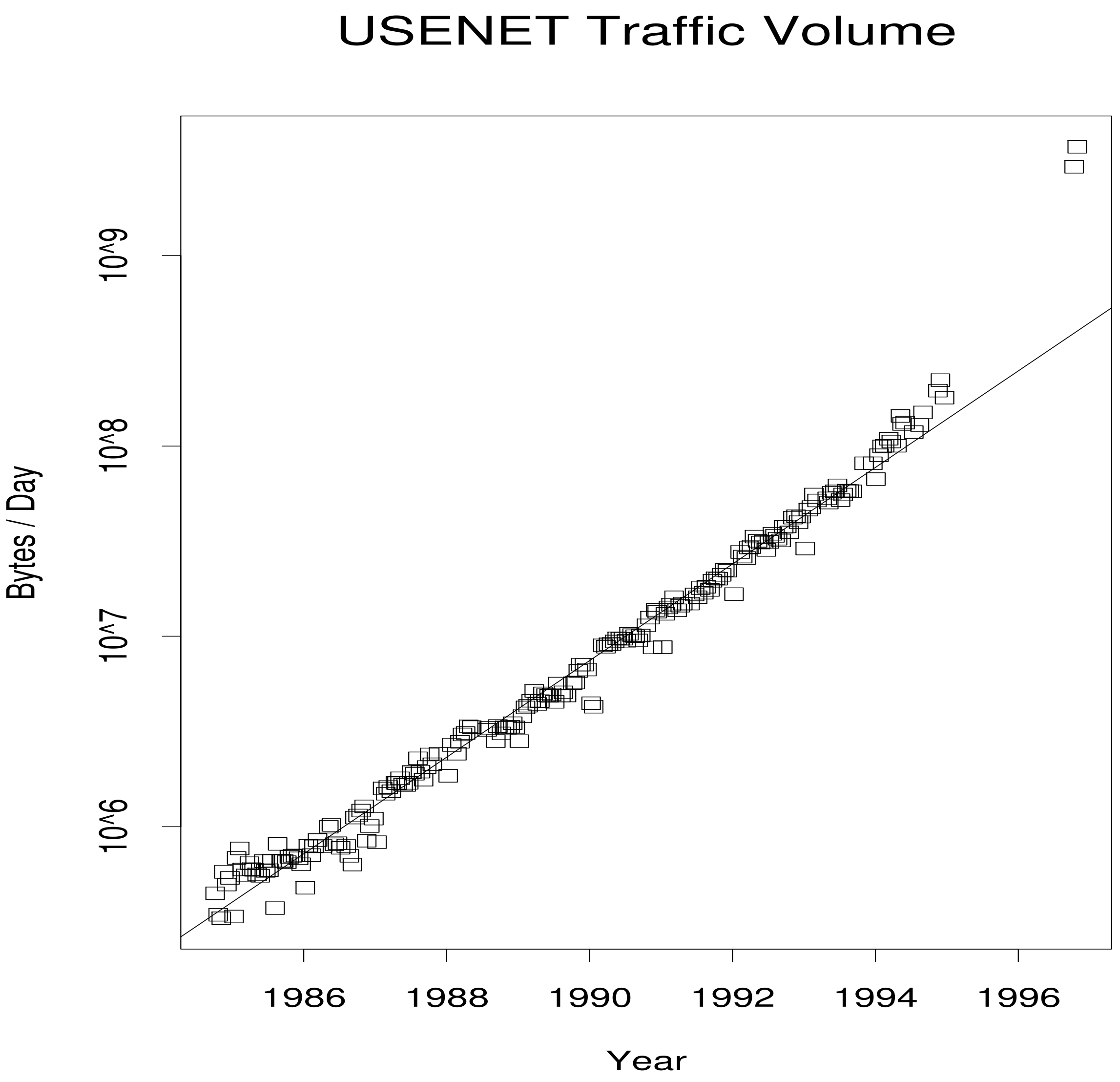


Courtesy of Rick Adams.

```
Date:  Fri, 12 Feb 1993 23:23:09 -0500 (EST)
From:  rick@uunet.uu.net (Rick Adams)
Subject:  Re:  past USENET readership summaries available somewhere?
To:  vern@ee.lbl.gov (Vern Paxson)
```

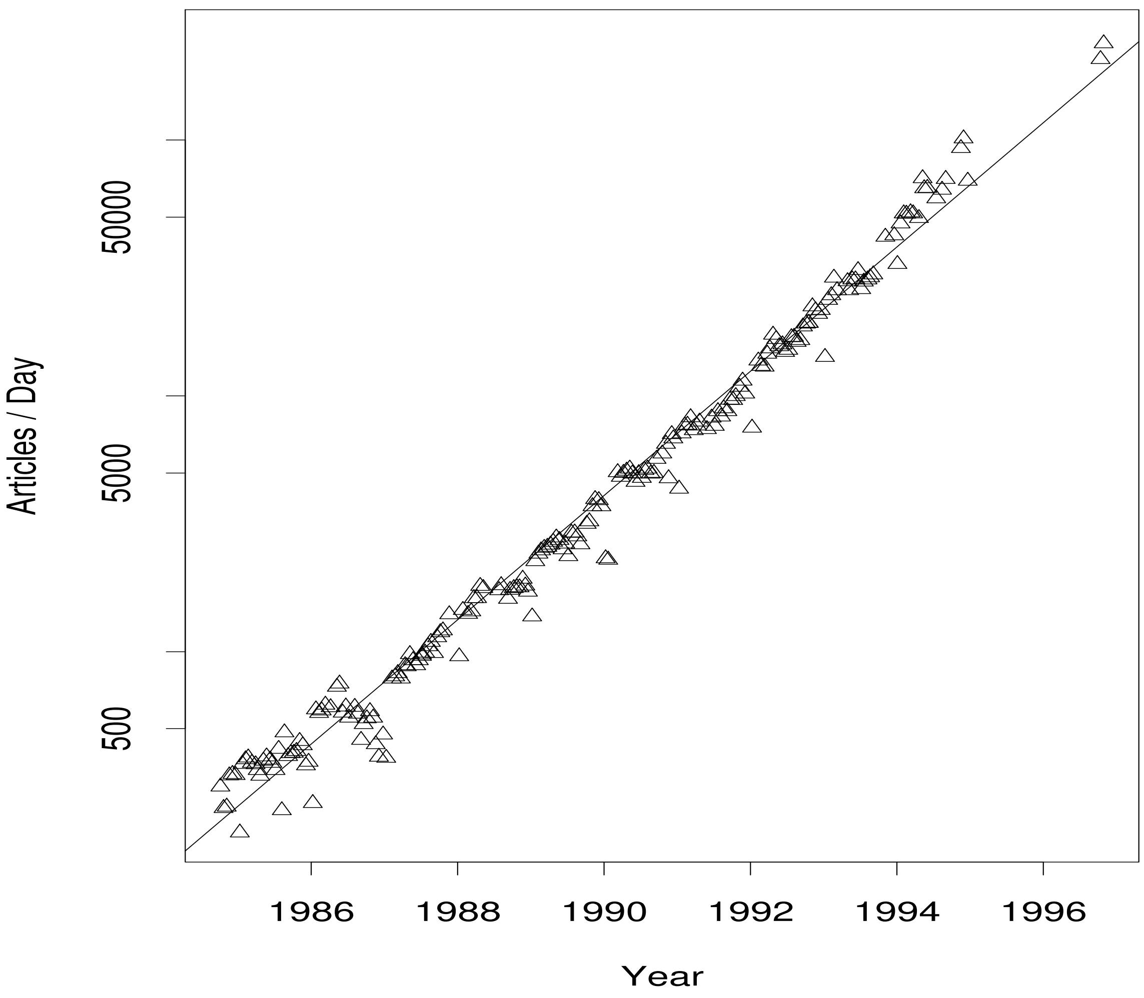
Date totarticles totbytes nsites nusers ngroups

```
841008 4241 6248063 606 1594 158
841022 3484 4826974 593 1398 153
841108 3529 4624868 624 1399 161
841122 4685 8097405 726 1780 157
...
910722 124152 250940767 12929 32790 1631
910808 118350 237868698 12940 31861 1602
910822 127466 255032480 13634 33731 1651 312170882
910909 124208 245649645 13503 33411 1657 301578027
...
940209 736124 1390635284 48119 152136 8825 1782584533
940223 734372 1406224179 48466 152994 9068 1803914508
940314 744204 1527020542 48605 164284 8432 1936945208
940328 736136 1480642940 47404 160123 8269 1892062974 42.8
940425 706233 1403401809 47057 159858 9366 1801979105 56.6616
940515 1011667 2015836594 55027 193072 9233 2583910000 53.5367
940522 923578 1824627700 55798 187601 9377 2337427010 35.31626419
...
```



Data courtesy of Rick Adams and David Lawrence.

USENET Traffic Volume



Data courtesy of Rick Adams and David Lawrence.

Admission #2: our measurement tools lie to us.

Packet filters can:

- fail to record packets
  - because the kernel ran out of buffer
  - because the NIC ran out of buffer
    * *because the packet was dropped*
    * *because they took another path*

Packet filters can:

- reorder packets
  - because there are multiple kernel queues
  - because there are multiple NICs
  - * *because the network reordered them*
  - * *because we confused cause and effect*

Packet filters can:

- replicate packets
  - because both the kernel and the NIC record them
  - * *because the network replicated them*
  - * *because the sender transmitted duplicates*

- capture packets that they're not supposed to

  "`tcp[4:4] == 0 and tcp[4:4] > 0`"

Clocks can:

- be arbitrarily off from true time
- be jumped forward or backward
- fail to move
- run arbitrarily fast or slow
  - *e.g., a factor of five*
- timestamps convey no sense of their actual accuracy

(GPS is <u>not</u> a simple fix here: hard to set up (hardware/software), hard to understand just how accurate it is.)

Admission #3: our mental models lie to us.

The problem of *misconception*:

E.g., measuring packet loss by counting retransmissions.

E.g., measuring Web fetches that hit hidden caches.

E.g., `ttcp` with large socket buffers, small data volume: what it measures is <u>time to fill the kernel buffer</u>, not transmission time.

E.g., Mark Allman's 10 msec to establish a TCP connection with a host 100 msec away, transfer data to it, close it down . . . but *the remote machine was powered off*!

Calibration:

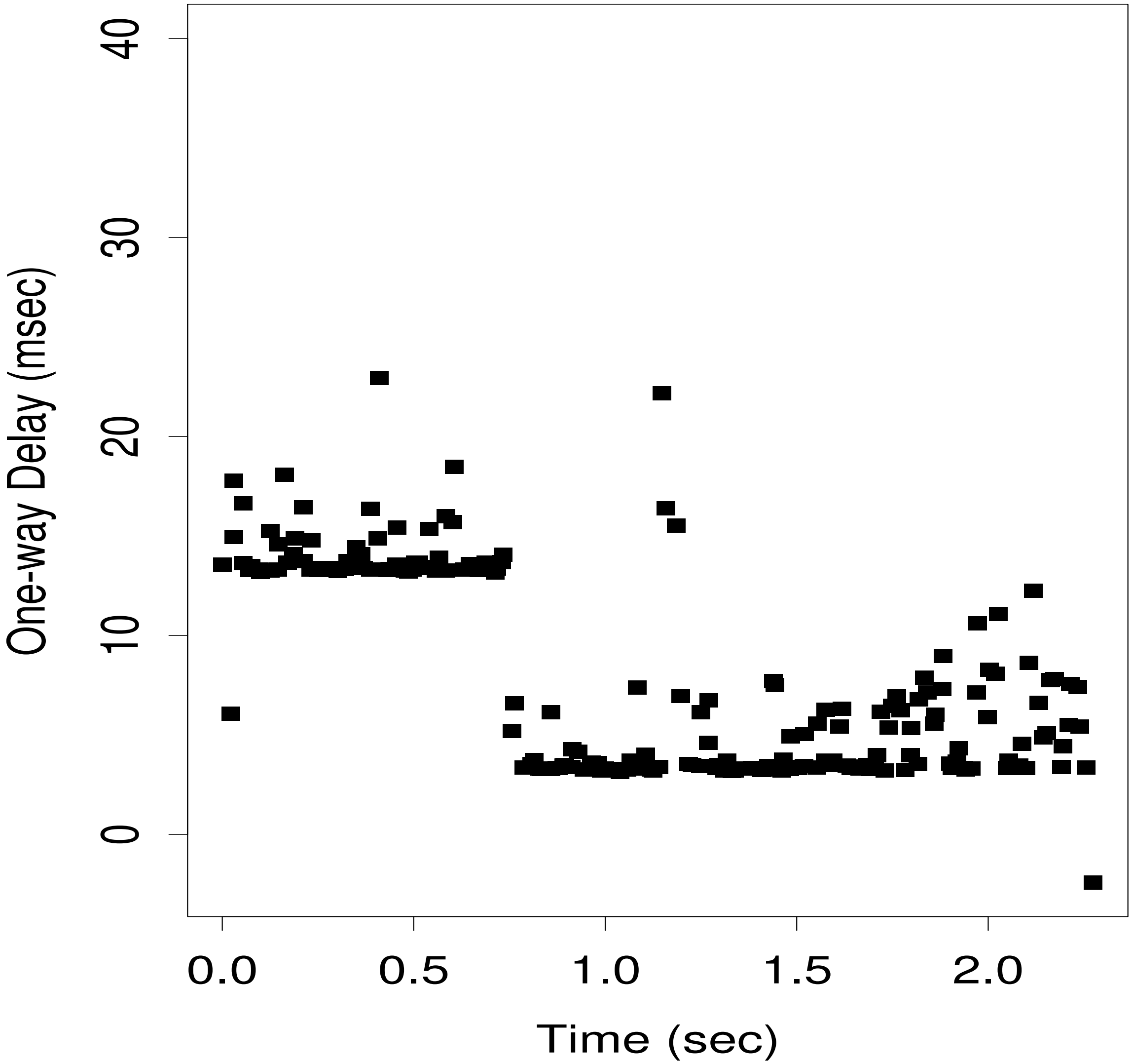
We can often detect measurement errors *if we have enough additional information*.

E.g., protocol information:

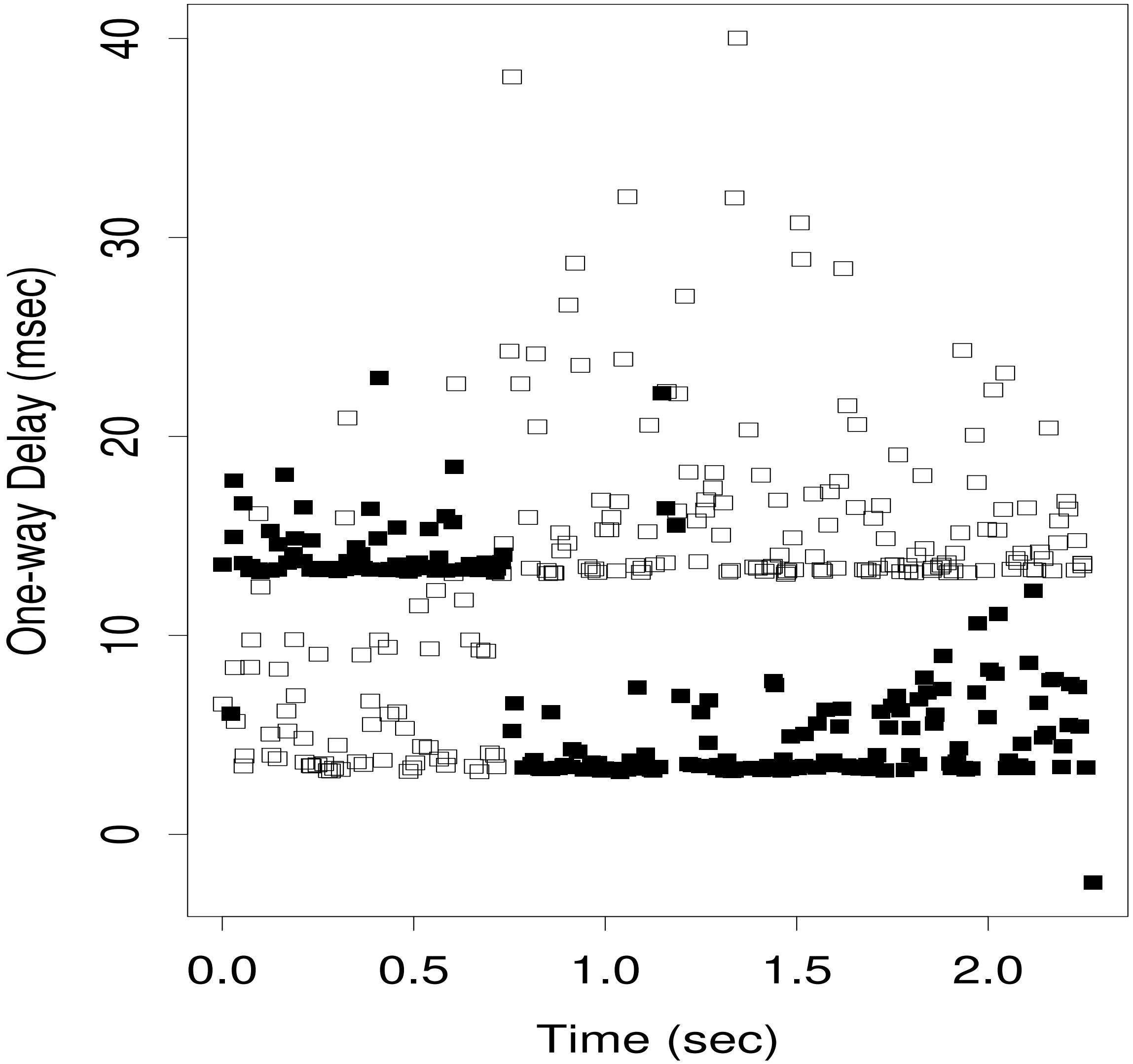
- if a TCP receiver acknowledges data never sent, the packet filter must have dropped the sent data.
- * *(except, this isn't true)*

E.g., other measurement perspectives:

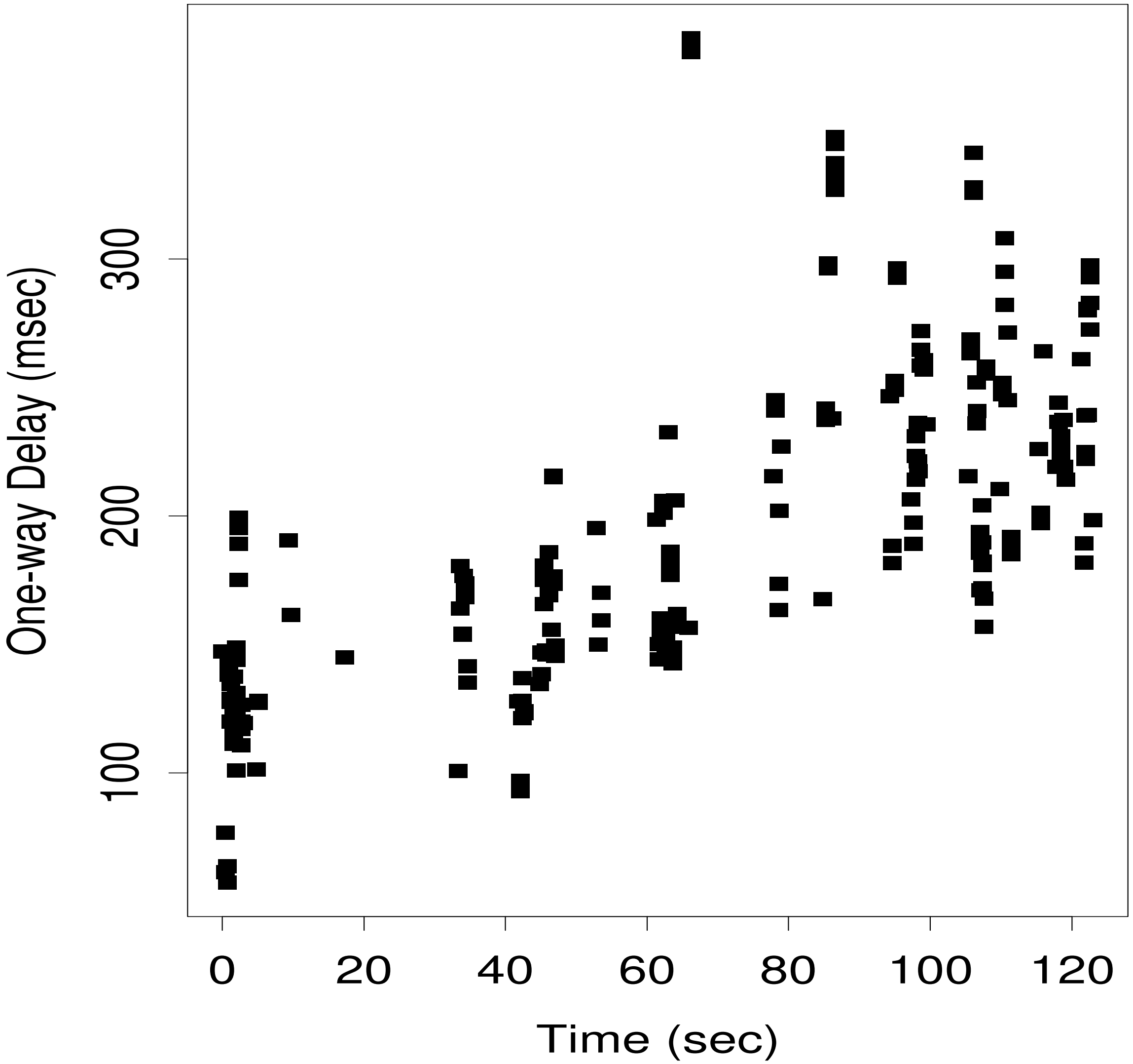
- errors in a single clock are often undetectable, but apparent when comparing clocks.

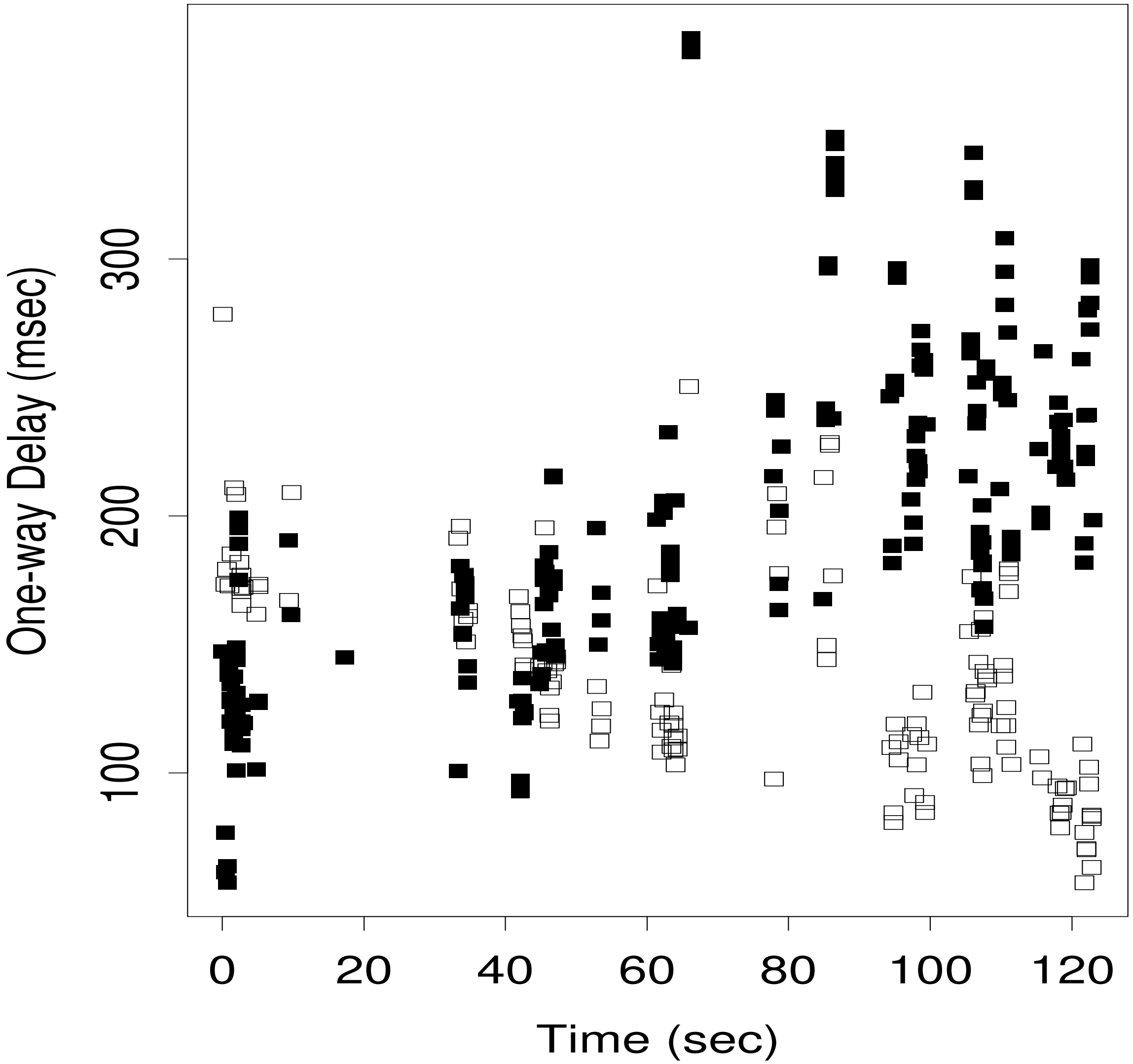


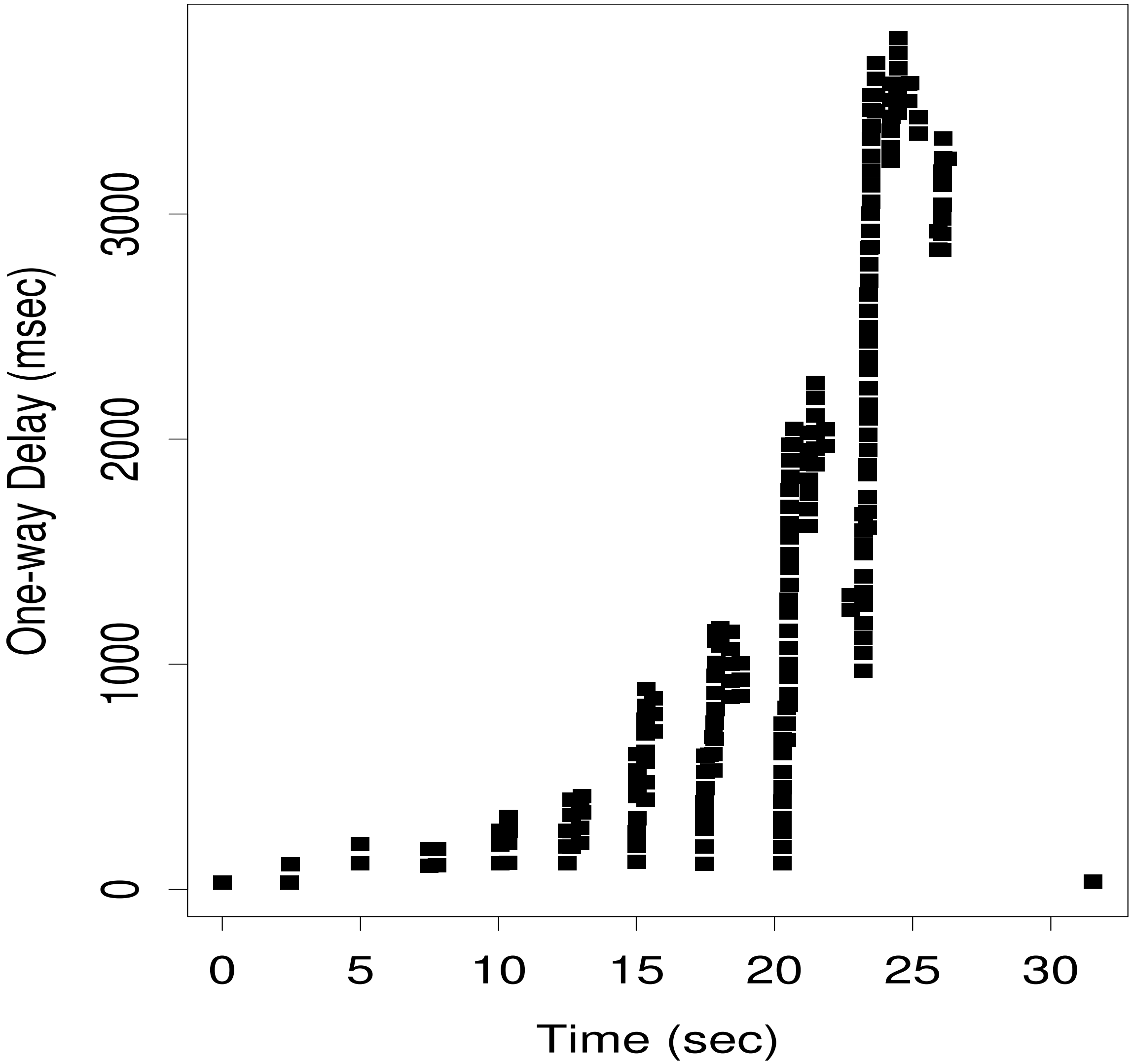
One-way Delay (msec)
0
10
20
30
40
0.0
0.5
1.0
1.5
2.0
Time (sec)



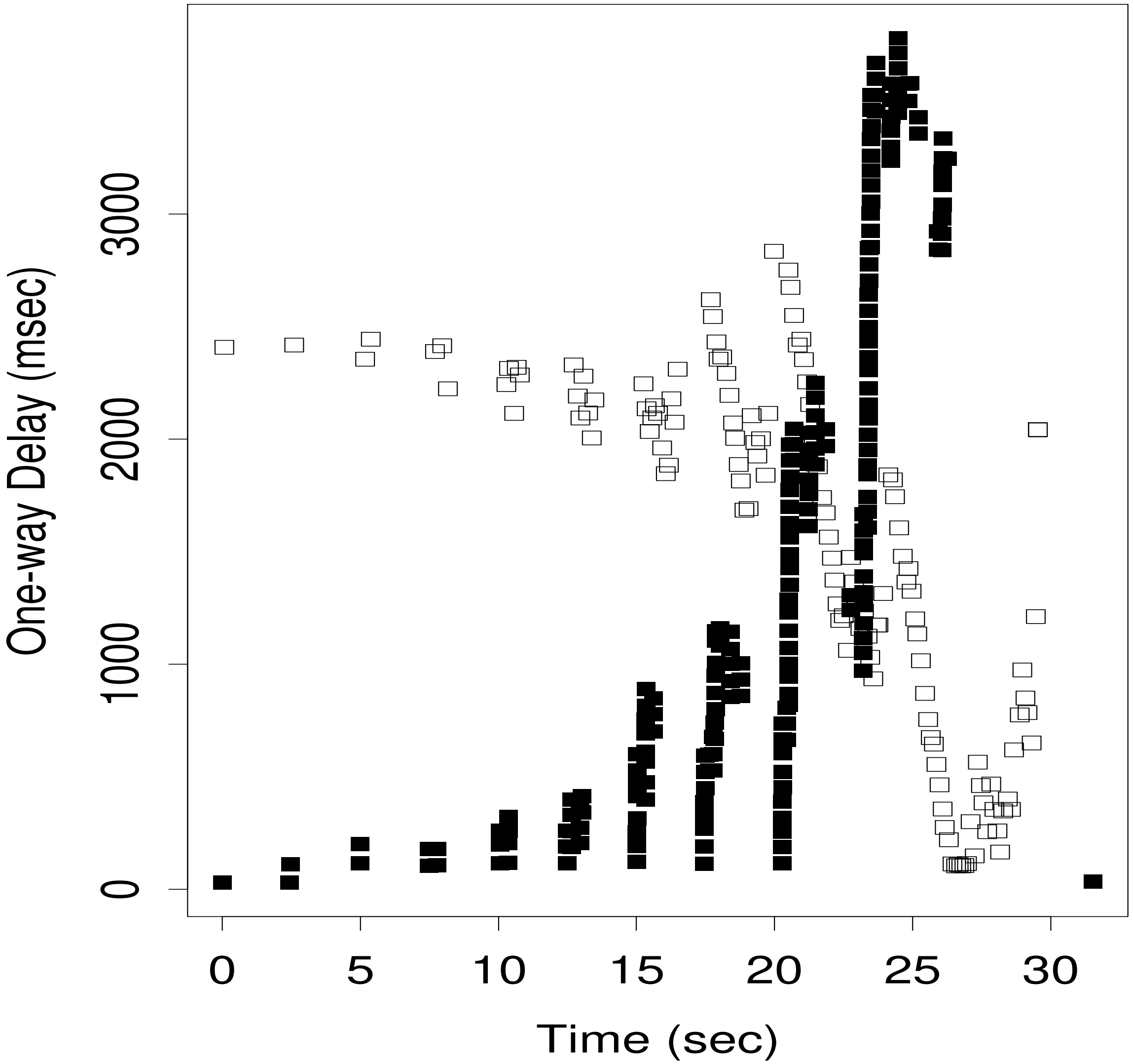
One-way Delay (msec)
0
10
20
30
40
0.0
0.5
1.0
1.5
2.0
Time (sec)


One-way Delay (msec)
100
200
300
0
20
40
60
80
100
120
Time (sec)


One-way Delay (msec)
100
200
300
Time (sec)
0
20
40
60
80
100
120



One-way Delay (msec)
0
1000
2000
3000
0
5
10
15
20
25
30
Time (sec)



One-way Delay (msec)
0
1000
2000
3000
Time (sec)
0
5
10
15
20
25
30

Calibration, con't:

So whenever we can measure something two different ways, it's incumbent upon us to do so in an attempt to calibrate.

However:

Admission #3$'$: measurements made two different ways always disagree.

Admission #3$''$: if something "can't" happen, and we check whether it has, sometimes indeed it has.

Calibration, con't:

And, even if/when we figure out why:

Admission #4: measurements carry no indication of quality.
(e.g., no room in `tcpdump` save file for [flaky] drop info)

Admission #4': measurements lack auxiliary information.
(e.g., database of IP $\Rightarrow$ hostname mappings at time of measurement)

Admission #4": we lack a *culture of calibration*.

Admission #5: our measurements are not representative.

The Internet exhibits *immense heterogeneity*: across different sites, at a single site over (not much) time, in the range of conditions encountered / performance achieved.

The Internet *keeps changing*: sustained exponential growth, repeated arrival of explosively popular new apps.

<u>There is no such thing as “typical.”</u>

$\Rightarrow$ Leads to *search for invariants*.

- Need highly diverse, longitudinal measurement datasets.
- Need measurement infrastructures.
- Need “limited scope” warning flags attached to datasets.

Admission #6: our analysis results are not reproducible.

Irreproducible results:

It is rare, though sometimes occurs, that raw measurements are made available to other researchers for further analysis *or for confirmation*.

It is more rare that analysis tools and scripts are made available, particularly in a coherent form that you can actually get to work.

It is even rarer that measurement glitches, “outliers,” analysis fudge factors, etc., are detailed.

Irreproducible results, con't:

Privacy issues and fear of being scooped are only part of the problem here; incoherence and (I bet) fear of exposure of analysis flaws are another.

Such *inter-scientist irreproducibility* is not good science.

Note: I'm *not* talking here about being able to redo an experiment and confirm/validate previous results. I'm talking about taking the *same* data and re-analyzing it in presumably the *same* way in order to confirm/validate the analysis of the data.

Irreproducible results, con't:

Referee report phrase I dread the most: "you need to redo the analysis with the following (modest) change."

Step #1: rerun without the tweak.

Likelihood this yields the same numbers as in the draft paper:

$$P[\text{same results}] = \frac{1}{\sigma\sqrt{2\pi}}\int_{n}^{\infty} \frac{-(x-\mu)^2}{2\sigma^2} dx$$

As n grows large enough, you rack up a lot of σ.

I.e., large datasets are rife with weird eccentricities.

Irreproducible results, con’t:

$\Rightarrow$ Intra-scientist irreproducibility.

(Again, this is about reproducing the analysis from the same data, not reproducing the experiment with new data.)

Irreproducible results, con't:

Of course, if you skip step #1 and just rerun the analysis with the requested tweak, life is easier, and *no one the wiser*.

But is it good science?

Related problem: ability to redo analysis or conduct follow-on analysis decays with time.

- mental context decays
- Internet shifts (hostnames change; configurations morph)

What's missing: *systematic approach to reduction / analysis*.

$\Rightarrow$ "paper trail" for analysis chain

particularly when bugs are fixed

One example of a paper trail:

- Enforce discipline of using a single (master) script that builds all analysis results from the raw data.
- Maintain all intermediary/reduced forms of the data as explicitly <u>ephemeral</u>.
- Maintain a *notebook* of what was done and to what effect.
- Use version control for scripts & notebook.

$\Rightarrow$ But also really needs: ways to visualize what's changed in analysis results after a re-run.

Admission #7: the large-scale measurements required for representative/longitudinal analysis overwhelm our home-brew data management.

Overwhelming data management:

- Difficulty of finding items with particular properties. (Meta-data embedded in filenames!)
- Risk of enormous *edit–compile–debug* cycles.
- Logistics of sharing become formidable.
- Disks fill up, inodes exhausted, data traversals hugely expensive.

  e.g., 35 days of *mflect* data = 175,472 files

  e.g., FreeBSD can't read & write huge files simultaneously
- Statistical tests overwhelmed by scale.
- Statistical software overwhelmed by scale.

Admission #8: even though taking Internet measurements is hard, it's too easy.

Two general approaches to measurement:

- **Research-driven: you have specific interests.**

  Plus: you focus on gathering particular info. you need.

  Plus: $\exists$ notion of what constitutes sound measurement.

  Minus: your notion will often be flawed (requires iteration).

  Minus: you may design data to support conclusion.

Two general approaches to measurement, con't:

- **Measurement-for-the-sake-of-measurement.**

  Plus: often vital for longitudinal studies.

  Plus: can foster serendipity.

  Minus: data often lacks care to avoid errors/artifacts.

  Minus: data often lacks ways to address these post facto.

  Minus: data is not infrequently completely useless/busted.

  Minus: can foster *glib analysis*.

  (same problem arises for public measurement repositories)

Admission #9: we don’t know how to measure real traffic.

Reality checks:

High volume links problematic to tap.

- need well-crafted measurement tools
- need well-understood *sampling methodologies*

Packet headers and esp. *packet contents* very problematic.

- legal / ethical / privacy / commercial issues
- subpoena bait

Anonymization technology is limited:

- need selective (e.g. local/remote) anonymization
- need (*sometimes*) ability to refer back to non-anonymized
- need *application-aware* anonymization
- need *configurable reduction agents*

Action items:

- Measurements need *pedigrees* describing what's known about them, what's been done to them, how to navigate.
- Need *audit trails* and portable *analysis scripting languages* (including dependency propagation) to support intra-scientist and inter-scientist reproducibility.
- General need for well-managed *meta-data*.
- Need to understand sampling implications and tech. better.
- Need anonymization tools & reduction agents.

Action items, con't:

Need a culture of, and a passion for, **sound measurement**.

$\Rightarrow$ Willingness to go the extra mile to find out you're wrong.

Measurement as: <u>science</u>.

Measurement as: <u>discipline</u>.

Culture, con’t:

Culture req’t #1: a willingness to invest $$$ in databases, etc.

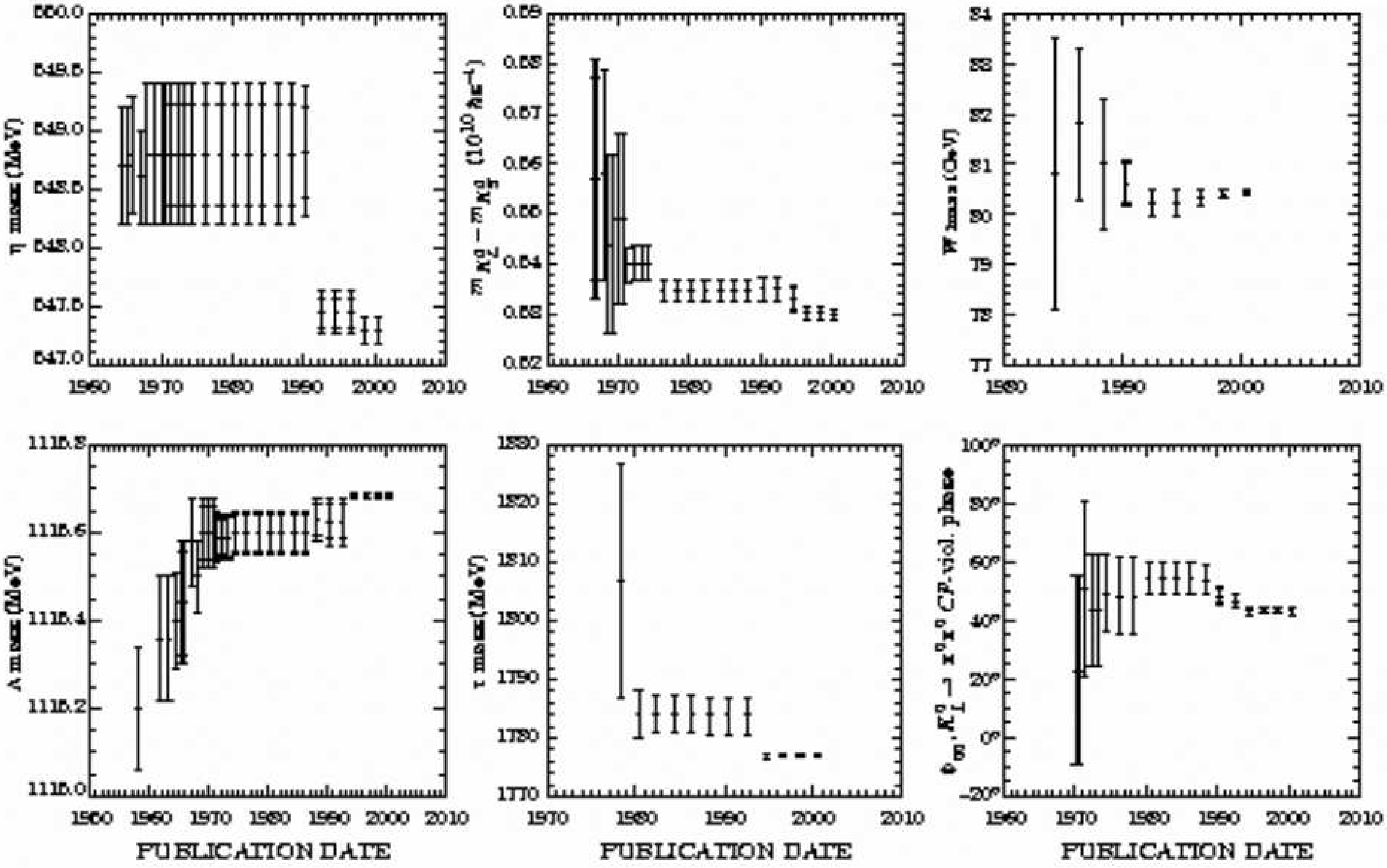
Culture req’t #2: a willingness to bounce manuscripts that don’t include the raw data + analysis scripts.

Culture req’t #3: a willingness to publish “mundane” work on measurement management.

Perhaps driven by:

Culture req’t #4: an emphasis on reproducible results.

What can we learn here from other sciences?



η mass (MeV)
550.0
549.5
549.0
548.5
548.0
547.5
547.0
1960
1970
1980
1990
2000
2010
$m_{K_L^0} - m_{K_S^0}$ (10^{10} $\hbar$ s^{-1})
0.59
0.58
0.57
0.56
0.55
0.54
0.53
0.52
W mass (GeV)
84
83
82
81
80
79
78
77
Λ mass (MeV)
1115.8
1115.6
1115.4
1115.2
1115.0
τ mass (MeV)
1830
1820
1810
1800
1790
1780
1770
ϕ_{+-}, $K_L^0 \to \pi^+\pi^-$ CP-viol. phase
100°
80°
60°
40°
20°
0°
−20°
PUBLICATION DATE
PUBLICATION DATE
PUBLICATION DATE

Some observations:

Problems with databases: schema management; lack of a warm feeling of being able to directly see the raw data. Would using XML for meta-data help here?

Should we consider teaching courses in which methodical (script-driven, database-driven, revision-controlled) analysis is *the* way it's taught?

Is it really worth the extra effort?

How much has slipshod measurement really cost us, particularly when what we’re measuring is so elusive/variable/changing anyway?

How can we know?

Is it really worth the extra effort?, con't

Clearly, we need to gauge what sort of precision & accuracy is appropriate for the problem at hand, and sometimes the answer is indeed "not much."

(This gets trickier when our measurements might be put to unforeseen uses in the future.)

Is it really worth the extra effort?, con't

But that said:

- Errors *add up*.
- Care in measurement often makes you more thoughtful about analysis, too.
- Seems likely slipshod measurement has cost us some significant dead ends, energy poorly spent, misleading mental models.

Acknowlegments:

Thanks to many for comments, additions, suggestions:

Andy Adams, Mark Allman, Paul Barford, Sally Floyd, Mark Handley, Eddie Kohler, Balachander Krishnamurthy, Al Lee, Jitendra Padhye, Scott Shenker, Walter Willinger . . .