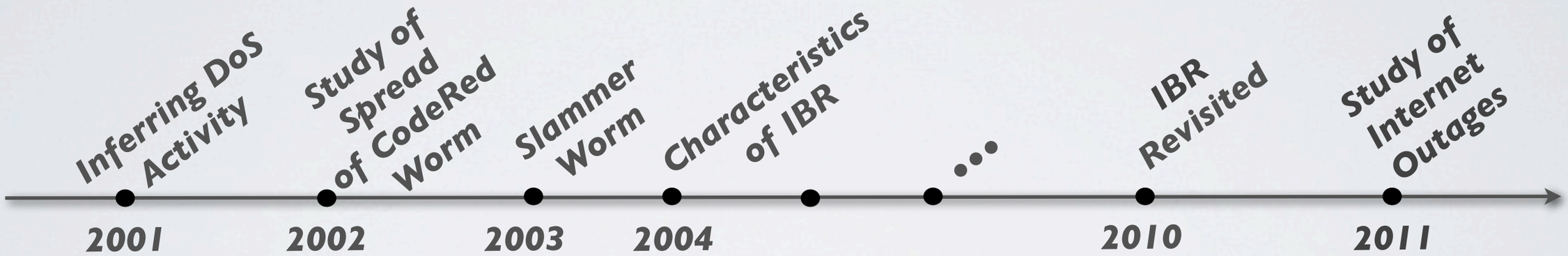
CORSARO

Alistair King, Alberto Dainotti
alistair@caida.org, alberto@caida.org
CAIDA

MOTIVATIONS

(for the scientists)

- Several researchers have used the UCSD Network Telescope



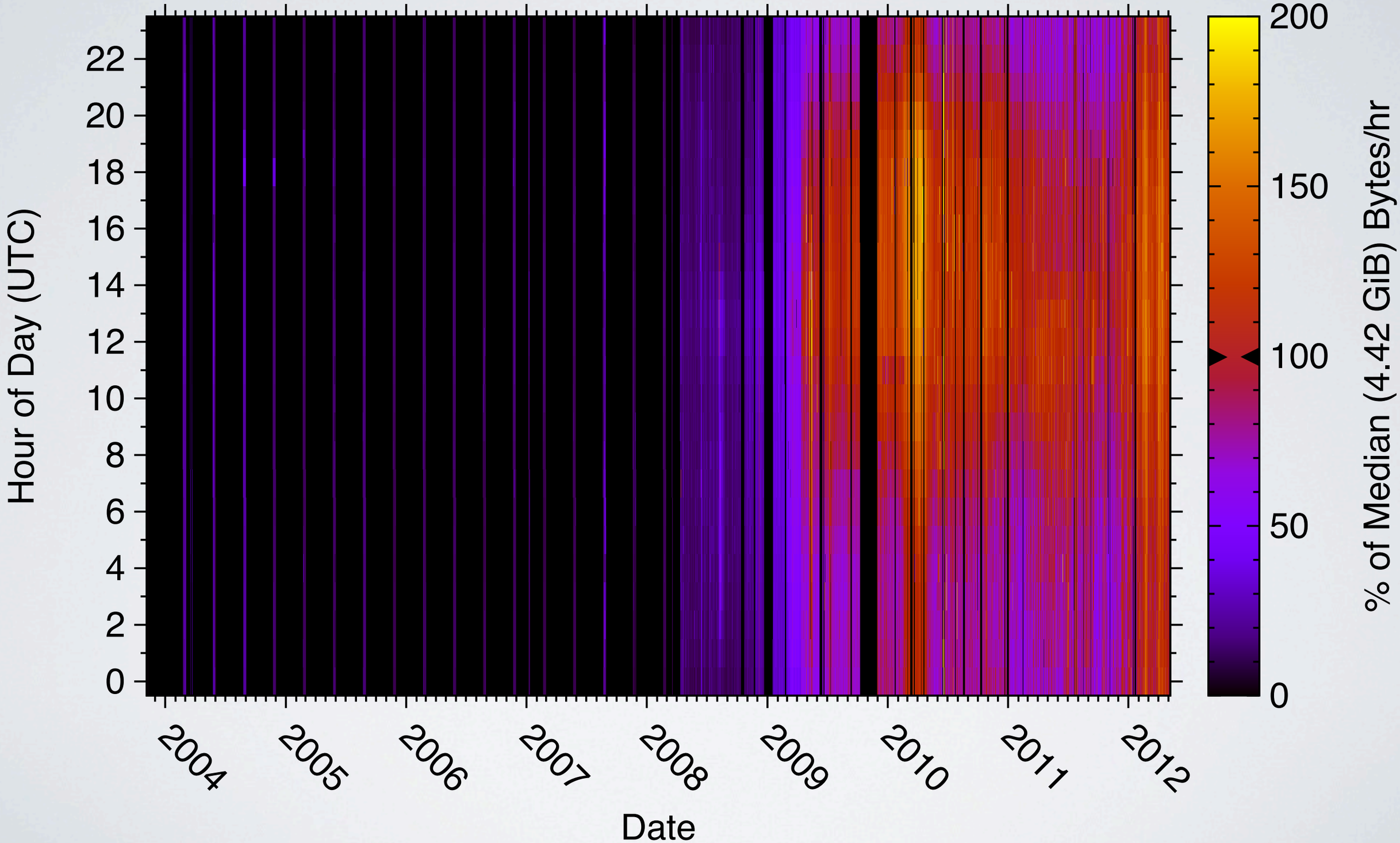
- Patchwork of tools and ad-hoc scripts
- All analysis has been with 'roll your own' code
- All results have been in 'proprietary' formats and locations
- There is no unified framework for analyzing darknet data

MOTIVATIONS

(for the people who pay the bills)

- Desperate times call for desperate measures
- For a decade CAIDA has enjoyed **free** (and virtually unlimited) **archival** of scientific data
- **No Longer!**
- We had **>100 TiB** of gzip pcap data from **2003-2011** stored on SDSC's tape archive

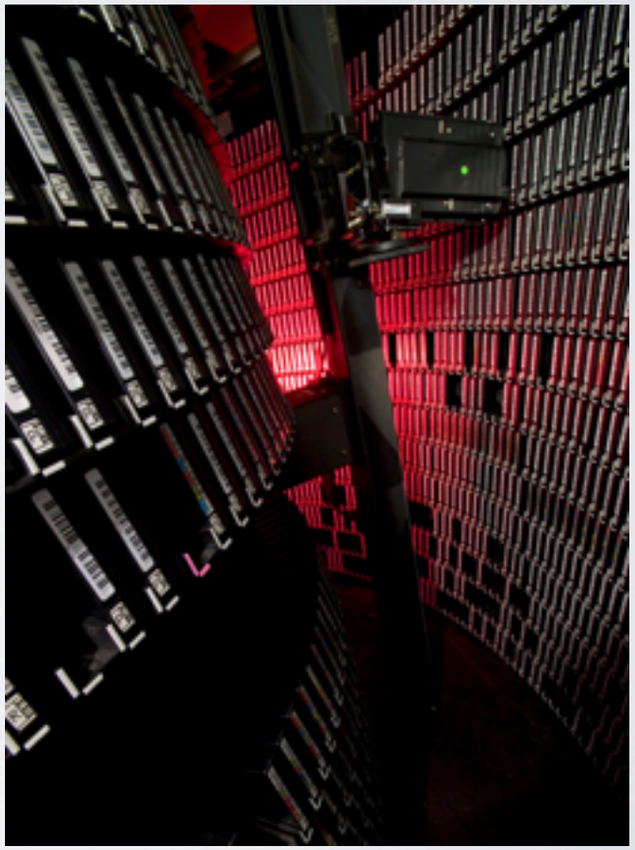
9 YEARS OF DATA



STORAGE

(the cold hard facts)

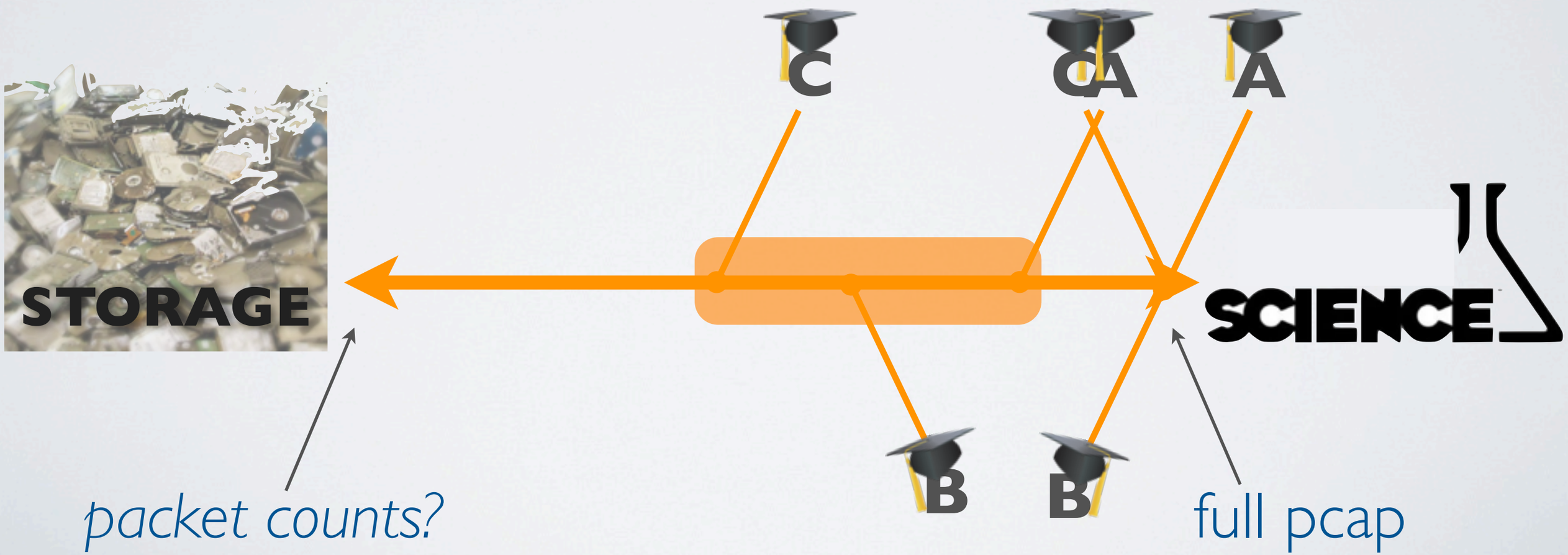
- 112 TiB -- 2,809,506,377,709+ pkts -- 36,025 hourly pcap files
- Most files are on tape
- And, it just keeps on coming!
  - ~3TiB per month
- We can't afford to store the existing data, let-alone keep up with the new data



TUG O'WAR

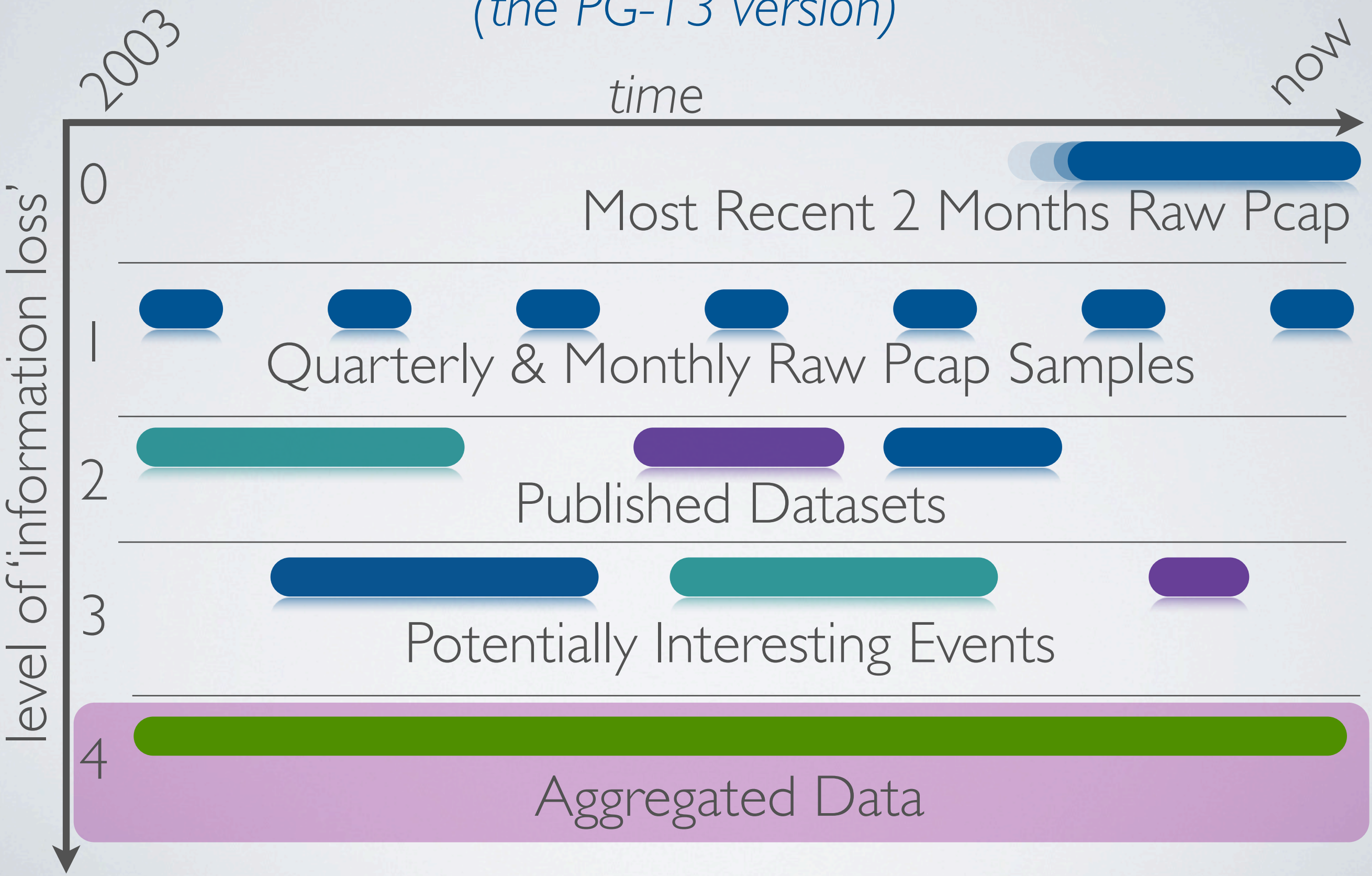
(the balancing act)

We asked several scientists what was required if we **must aggregate data**...



STRATIFIED STORAGE

(the PG-13 version)



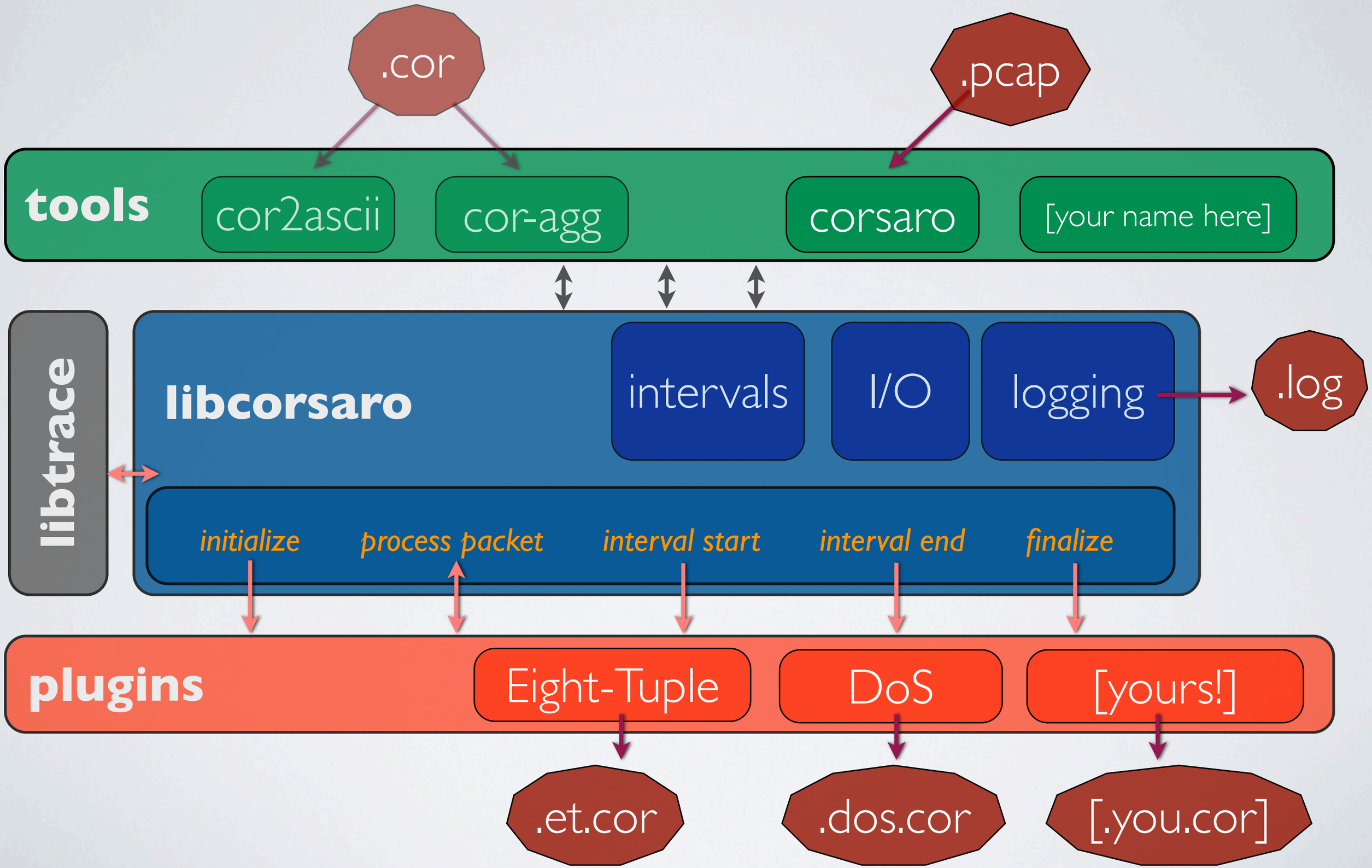
IL CORSARO

- We need a tool that can...
  - **Do Good Things** with every packet
  - Help **Minimize Storage** Costs
  - Do it very **Efficiently**
  - Be **Easy and Useful** for researchers to extend

KEY GOALS

- Compression
- Speed
- Easily Usable
- Portable
- Extensible
- Reliable

A PICTURE


.cor
.pcap
tools
cor2ascii
cor-agg
corsaro
[your name here]
libtrace
libcorsaro
intervals
I/O
logging
.log
initialize
process packet
interval start
interval end
finalize
plugins
Eight-Tuple
DoS
[yours!]
.et.cor
.dos.cor
[.you.cor]

LINEAGE

(and not reinventing the wheel)

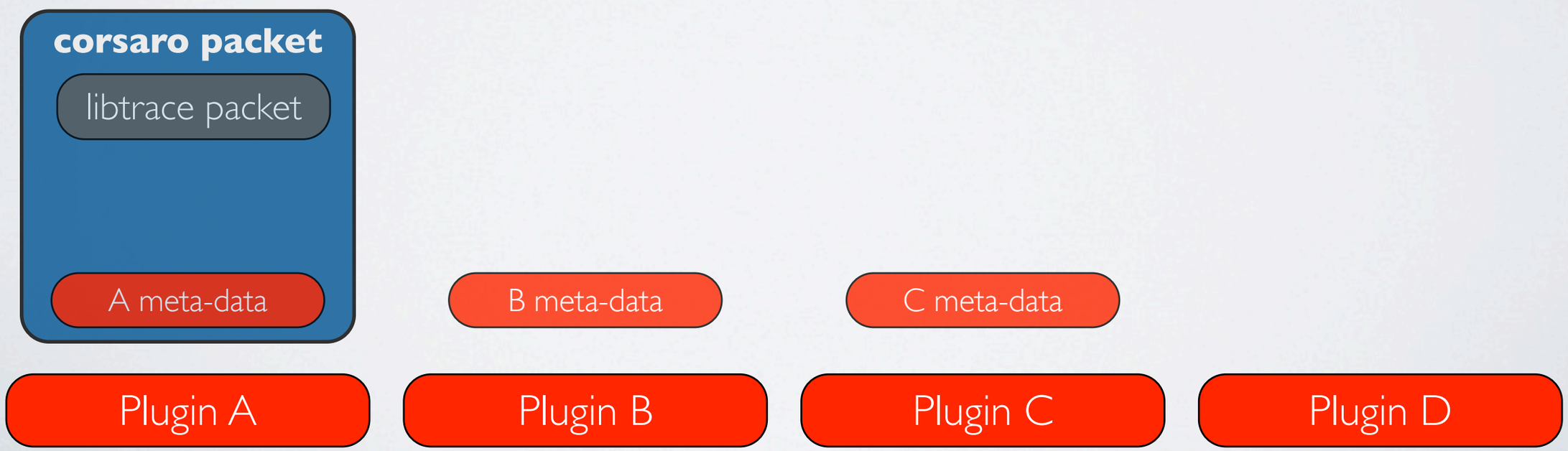
- *framework.c*
  - A proof-of-concept darkcap analysis engine by Alberto Dainotti
- *libtrace*
  - Library for trace processing by WAND group
  - Multi-threaded, actively developed/supported
  - *http://research.wand.net.nz/software/libtrace.php*
- *libwandio*
  - Library for threaded, compressed file IO.
  - Comes as part of *libtrace* (since 3.0.14)

COMPRESSION

- **Aggregates** data into **intervals**.
  - Trade-off time resolution for reduction of redundant data.
- Highly **optimized binary output**.
  - Carefully sorted to exploit characteristics of *gzip*
- Provides transparent **output compression** to plugins.
  - Both *bzip* and *gzip* supported.

SPEED
(and efficiency)

- *Libtrace* is designed for speed (zero copy, caching, etc)
- All IO is threaded to take advantage of modern hardware
  - E.g. Corsaro with *bzip* runs as fast as when it uses *gzip*
- Minimize rework by plugins:



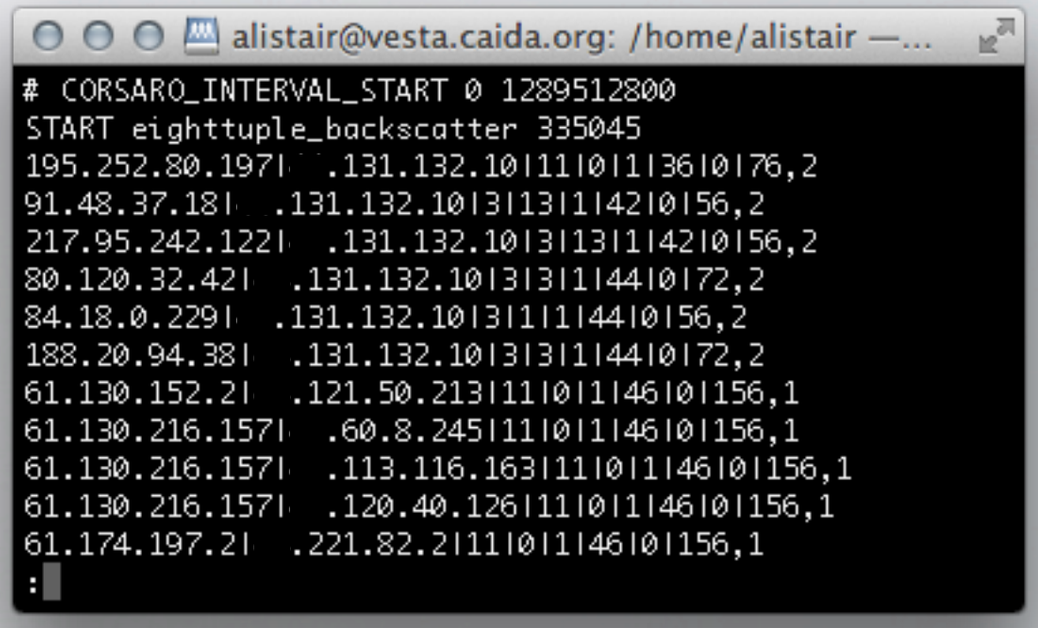
BACK TO THE SCIENCE

- We have identified three main types of plugin:
  - General purpose aggregation.
  - Specialized Analysis.
  - "I need to know x **right now**"

THE EIGHT-TUPLE

(the penicillin of aggregated data)

- A **general purpose aggregation** plugin for Corsaro.
- The **Eight-Tuple satisfies** several **common analysis needs**
- Features:
  - Source IP, Dest IP, Source Port, Dest Port, Protocol, TCP Flags, TTL, IP Length
  - Per-interval key/value pair:
    key => EightTuple
    value => Packet Count *(for the interval)*
  - Also keyed on the packet classification (e.g. backscatter)

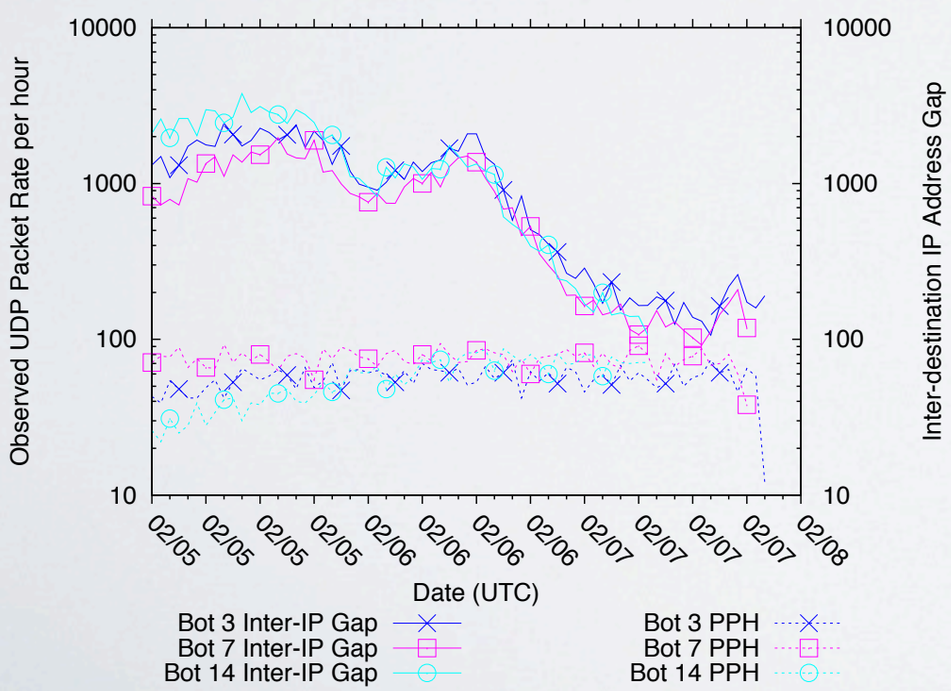


- **>80% compression** from .pcap.gz using 1 minute aggregation intervals

PUTTING IT TO USE

\cite{eight-tuple}

- **Eight-Tuple data** and **Corsaro** heavily **used for analysis** in two recent IMC papers:
  - *"Analysis of a '10' Stealth Scan from a Botnet"* - A. Dainotti *et al.*
  - *"Entropy-based Classification of IP Darkspace Events"* - T. Zseby *et al.*



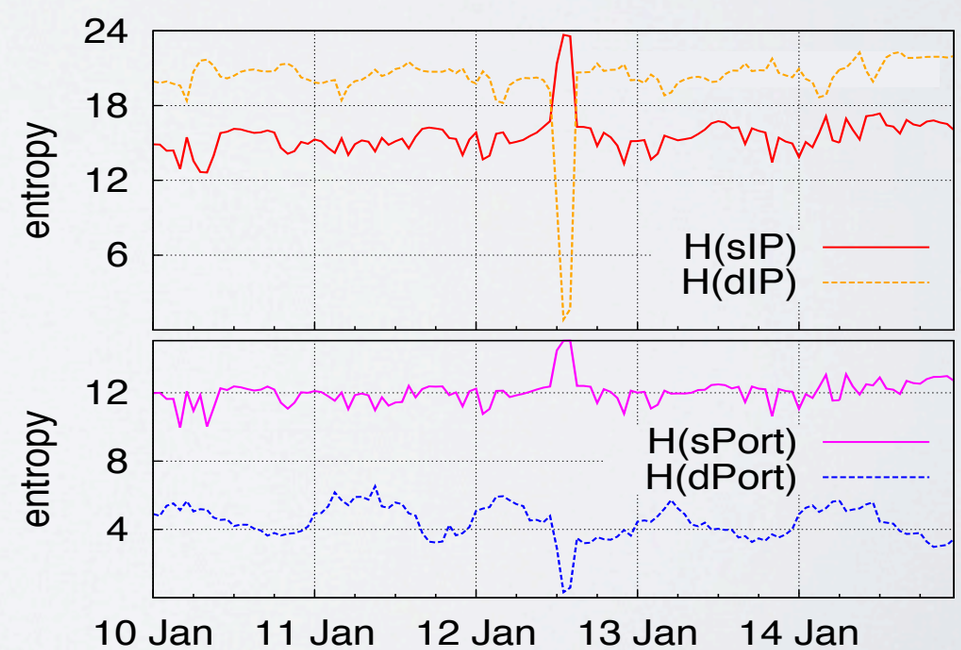


Figure 4: ***Entropy during TCP Probe***

SPECIALIZED ANALYSIS

(for that special code in your life)

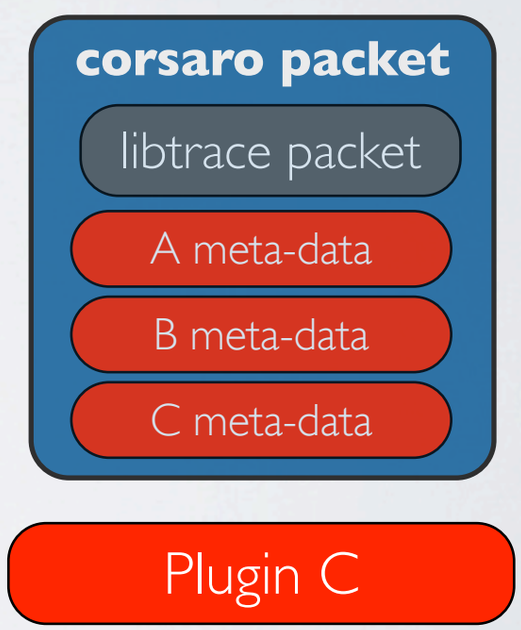
- Corsaro supports **highly-specialized analysis** plugins
- **Existing code** that does **something complicated** can **leverage Corsaro's features**
- As an example, we ported our *new_rsdos** tool:
  - DoS detection algorithm
  - Optimized for speed and output compression
  - Identifies potential "Attack Vectors" and records statistics about the attack
  - Preserves the 'initial' packet for later inspection

*see http://www.caida.org/publications/papers/2001/BackScatter/

AD-HOC ANALYSIS

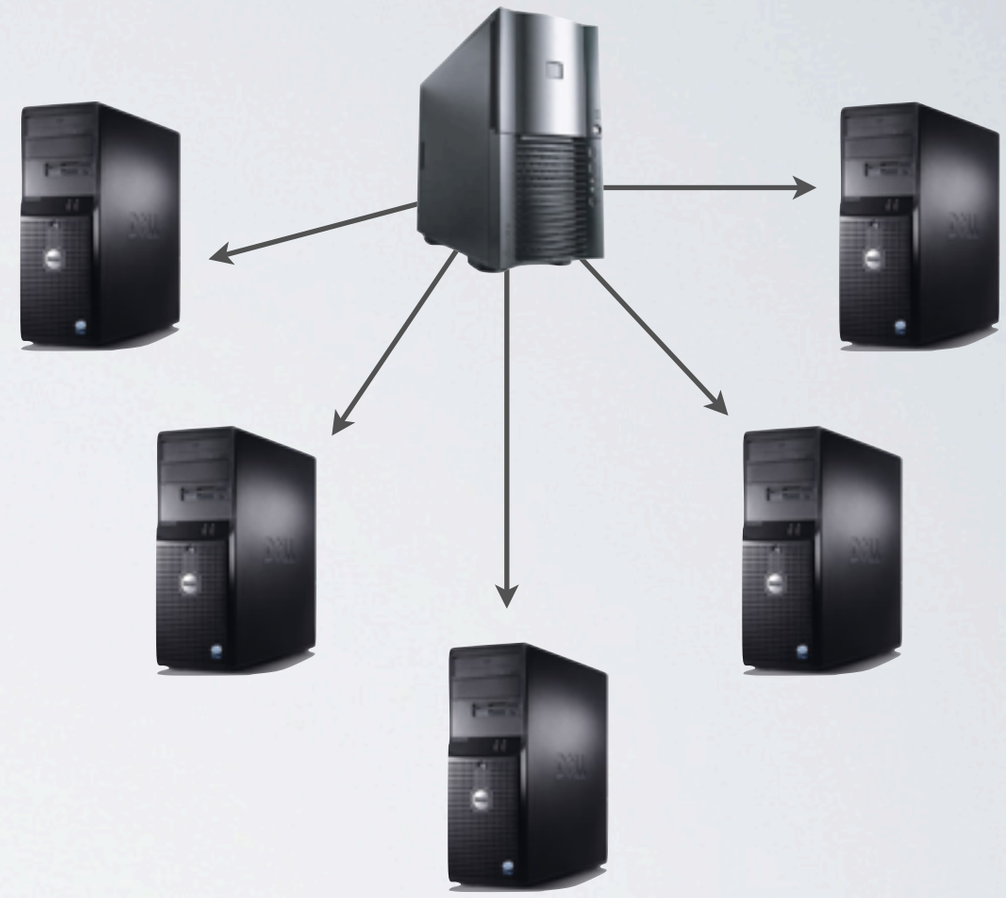
(agile research)

- **Parsing** tcpdump **ASCII** output is **slow and error prone**
- Corsaro makes it **quick and easy** to add a **new plugin**
- **E.g.** we wanted to know **# packets** and **# unique source IPs**, that are **not part of a DoS** attack, in an hour:
  - In **< 1 hour**, we had a plugin - it runs **fast**
  - For free we got:
    - DoS identification by a prior plugin (**chained results**)
    - Threaded I/O
    - Output is compressed
    - Adaptable interval lengths (e.g. we now want daily counts)



CORSARO IN ACTION

(getting it done)



- Corsaro has been in active use at CAIDA since Feb 2012
  - FreeBSD, Linux, Mac OSX, Solaris X
- Combined Corsaro and Marinda (http://www.caida.org/projects/ark/)
- Used an ad-hoc cluster to **process 100 TiB** data in **down to 15 TiB**
- Has been run with over **30,000 hours** of pcap

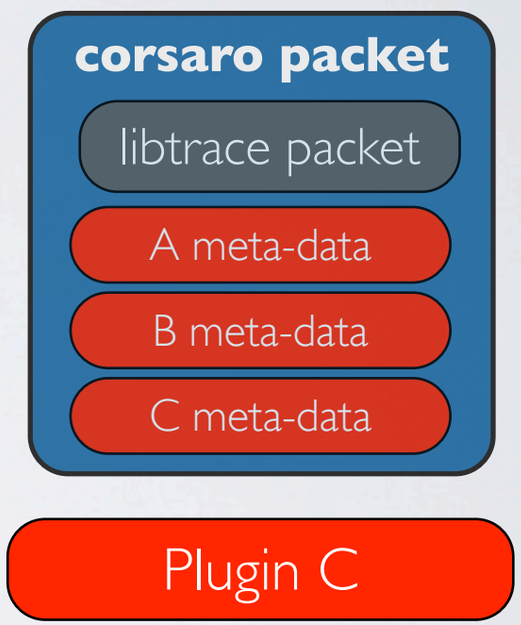
BEFORE **YOU** GET YOUR HANDS ON IT...

corsaro-1.0.0.tar.gz

- More extensive documentation
  - Currently we only have doxygen API docs
- Finish input API to process Corsaro output
  - Currently only the eight-tuple data is supported
- Remove some code specific to our /8 telescope

WHERE ARE WE GOING?

- Extend Corsaro to provide realtime packet capture, analysis and archival of darkspace data.
- Geolocation and AS-mapping plugins for populating packet meta-data
- Realtime reporting and visualization
- Data sharing
- Efficient Indexing for fast searches
- IPv6



ACK && QUESTIONS

(we would love some suggestions)

- **Dan Andersen** - for tirelessly maintaining and provisioning CAIDA machines well beyond their intended purposes.
- **Emile Aben** - for relentlessly pursuing Good Science
- **Tanja Zseby** - for valuable input along the way, and for being an eager (and sometimes unfortunate) pre-alpha user.
- **NERSC** - for agreeing to archive every pcap file at the last minute.
- **SDSC** - for being patient while we moved, providing compute resources, and for storing all that data for all those years.