DHS PREDICT project: CAIDA update

- Research accomplished
- Research/data infrastructure updates
- Dataset dissemination/stats
- Open issues
- Phase 2 datasets

Recent research

- E. Kenneally, M. Bailey, D. Maughan, "*A Framework for Understanding and Applying Ethical Principles in Network and Security Research*", in the proceedings of the *Workshop on Ethics in Computer Security Research (WECSR 2010)* in January 2010
  http://www.caida.org/publications/papers/2010/framework_ethical_research/
- E. Kenneally, kc claffy, "*An Internet Data Sharing Framework For Balancing Privacy and Utility*", in the proceedings of *Engaging Data: First International Forum on the Application and Management of Personal Electronic Information*, MIT, Oct. 12-13, 2009.
  http://www.caida.org/publications/papers/2009/engaging_data/

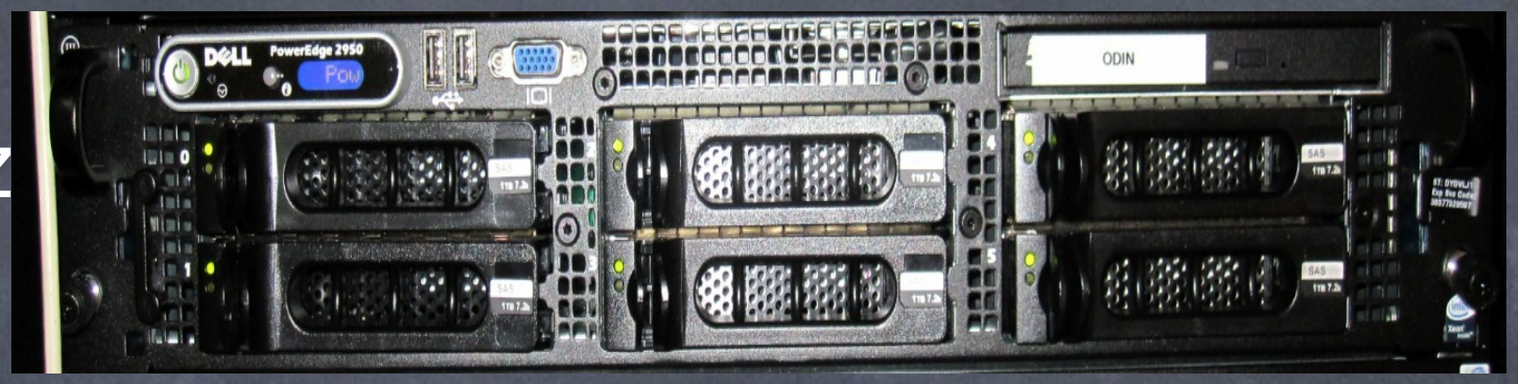
current research (cont.)

- S. Castro, M. Zhang, W. John, D. Wessels, kc claffy, "*Understanding and preparing for DNS evolution*", published in the *PAM Traffic Monitoring and Analysis Workshop*, January 2010
  http://www.pam2010.ethz.ch/TMA/cfp.html
- E. Kenneally and kc claffy, "*Dialing privacy and utility: a proposed data-sharing framework to advance Internet research*", in submission to *IEEE Security & Privacy* special issue, July 2010.

(http://www.caida.org/publications/papers/2010/)

New research infrastructure

- UCSD Telescope:
  Intel Xeon 8 x 3GHz
  32GB RAM, 4.3TB



- Data Server:
  Intel Xeon 8 x dual cor
  2.5GHz,16GB RAM,
  8TB storage

- New Ark Monitors
  38 total active: 11 IPv6
  (see topology project)

New research infrastructure (cont)

- Doors and curtains

New research infrastructure (cont)

- SDSC Machine room gets shower curtains

New research infrastructure (cont)

- SDSC innovating in energy efficient machine room air handling systems where few standards exist
  - Old machine room focuses on hot aisles using curtains to force hot air to 16 air handlers. Allows us to increase air delivery set point from 55 degrees to 75 degrees. Curtains (from Subzero Coldrooms) have fuseable links that release in case of fire so fire systems will function.
  - New machine room focuses on cold aisles using homogenous racks to contain cold air to 500 sq/ft of the 5000 sq/ft space. Requires only 6 air handlers. (Knuerr)

New machine room

Policy support: IRB (re)review

- We submitted our 2nd IRB application to UCSD in November 2009, provided requested clarifications in December and were told in January 2010 that our protocol does not qualify as human subjects research.

The Committee at the January 6, 2010 Institutional Review Board meeting reviewed your December 14, 2009 correspondence. The Committee acknowledged that the PI had provided a thoughtful and substantive response to their letter dated December 2, 2009; however, the Committee voted unanimously that the proposed activity does not satisfy the regulatory definition of human subjects research and therefore falls outside the jurisdiction of the IRB. For further clarification we offer the following:

It is the Institutional Review Board's charge to review research that presents a testable hypothesis and study design that will add to generalizable knowledge and takes into account protection of living individuals under appropriate informed consent. Based on the information submitted, the proposed activity is not actively seeking information from human subjects, but rather computer generated information. Should there be a decision in the future to develop, for research purposes, a hypothesis and other research related design methods that utilizes information gathers from individuals, the Committee recommended that a new application be submitted for IRB review.

what data do we collect?

- OC192 backbone: 8.5 TB (3.6 anonymized; 4.9 unanonymized) – curation to quarterlies will reduce
- UCSD telescope: 3.4 TB on disk (30 day window) 4.8 T on samqfs
- OC48 traces: 1.7TB (same old 2004 traces, in PREDICT)
- topology: 12.3 TB (skitter+ark uncompressed)
  - routed ipv4: 2.3TB (in PREDICT) since Sep 2007
  - routed ipv6: 275MB since Dec 2008

Total: ~30TB (as of 15 Feb 2010)

how do we curate the data?

- OC192 backbone: strip payload/L1/L2, transfer, anonymize, archive (aggregated links)
- OC48 traces: strip payload/L1/L2, anonymized w (prefix-preserve) cryptopan
- UCSD telescope: filter legitimate traffic at the router, 30 days on disk, curate custom data sets upon request
- topology: see cybersecurity project

(http://www.caida.org/home/legal)

how do we serve the data?

- OC192 backbone: report generator
  http://www.caida.org/data/realtime/passive/?monitor=equinix-chicago-dirA (also traces to academics who sign AUP)
- OC48 traces: PREDICT; academics who sign
  http://www.caida.org/data/passive/anon_internet_traces_request.xml
- UCSD telescope: PREDICT; academics who sign http://www.caida.org/data/passive/network_telescope.xml#access
- topology: PREDICT; academics who sign
  http://www.caida.org/data/active/ipv4_routed_24_topology_dataset.xml

(commercial researchers must join caida)

How do researchers use the data?

- OC192 backbone: report generator up, traffic classification, performance modeling
  http://www.caida.org/data/publications/bydataset/index.xml#passive
- OC48 traces: traffic classification, modeling, monitoring, filtering, generation, locality
  http://www.caida.org/data/publications/bydataset/index.xml#OC48
- UCSD telescope: Conficker, worm research
  http://www.caida.org/data/publications/bydataset/index.xml#Backscatter
- topology: pkt traceback, marking. DOS defense. topo and routing modeling, discovery, metrics, improvements
  http://www.caida.org/data/publications/bydataset/index.xml#Topology

how do we use the data?

- OC192 backbone: traffic classification, real time monitor, traffic symmetry, address utilization, other myths
- OC48 traces: traffic classification, modeling, p2p, (also http://www.caida.org/data/realtime/passive/?monitor=sdnap)
- UCSD telescope: traffic classification, real-time monitor (/data/realtime/telescope), lots of (and not enough..) Conficker analysis
- Topology: annotated Internet mapping http://www.caida.org/research/topology/

(www.caida.org/publications/papers/)

how many PREDICT requests for our data?

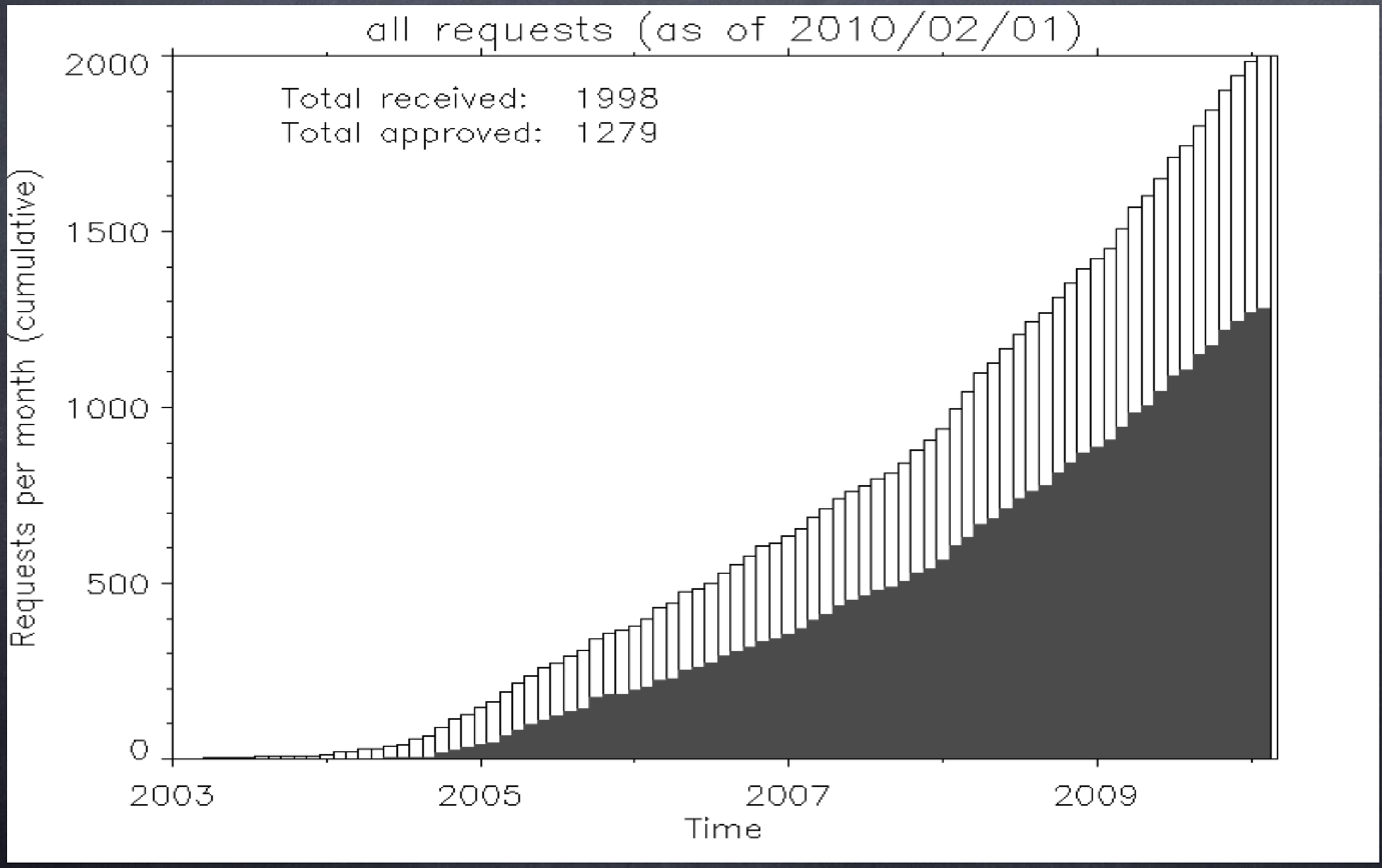
Dataset	Requests	Approved	Accessed
Backscatter	3	3	1
Passive (oc48)	2	1	0
	5	4	1

how many total requests for the data?

Dataset	Requests	Approved	Accessed	Since
Backscatter	451	241	207	Feb 2003
Passive	799	585	483	Feb 2004
Topology	614	372	290	Jul 2004
Witty	58	38	32	Mar 2008
Telescope	36	20	16	Jul 2009
DNS-RTT	40	23	18	Aug 2006
	1998	1279	1046	

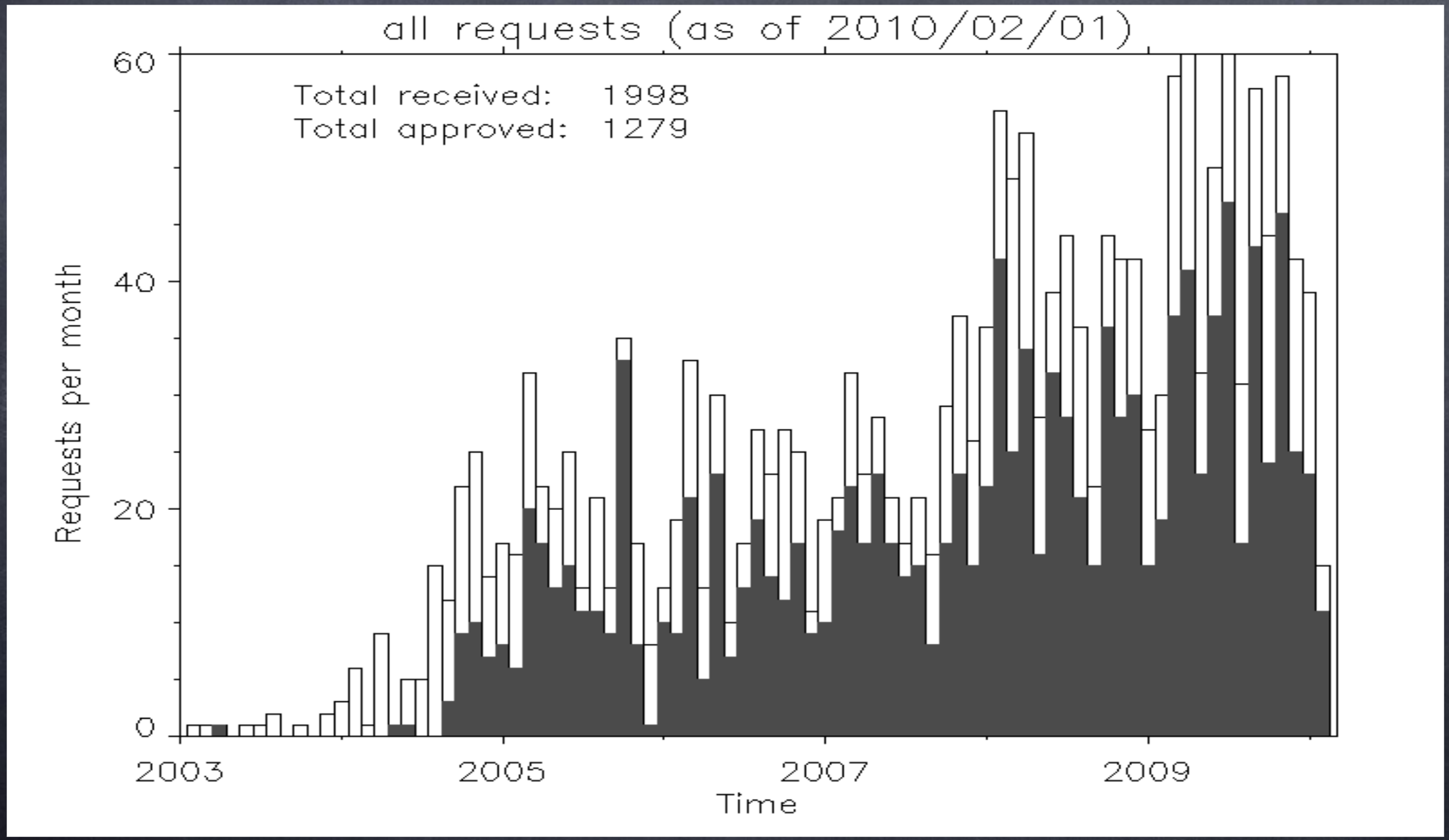
Data request stats

- All requests (cumulative)



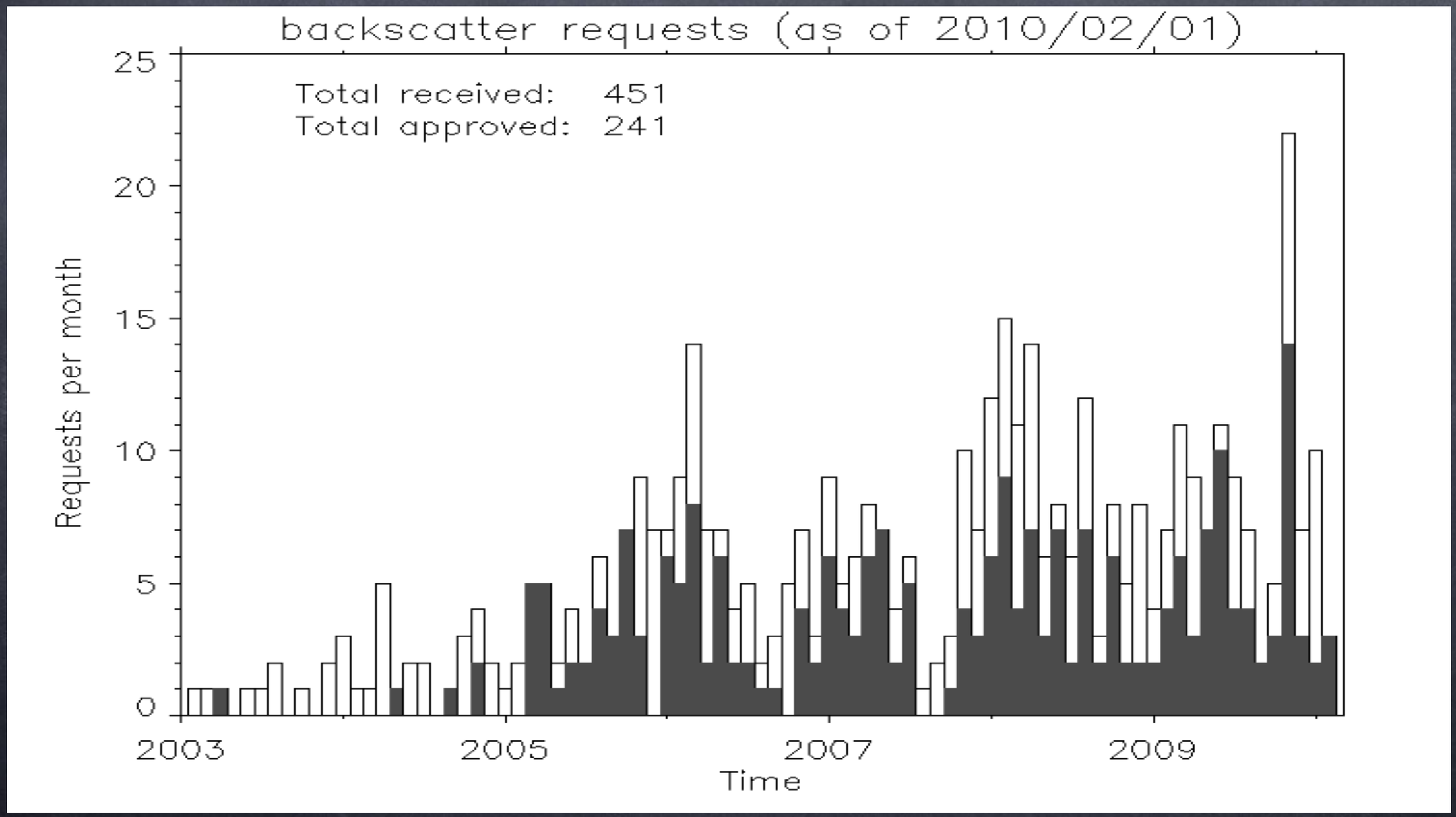
Data request stats (cont)

- All requests (monthly)



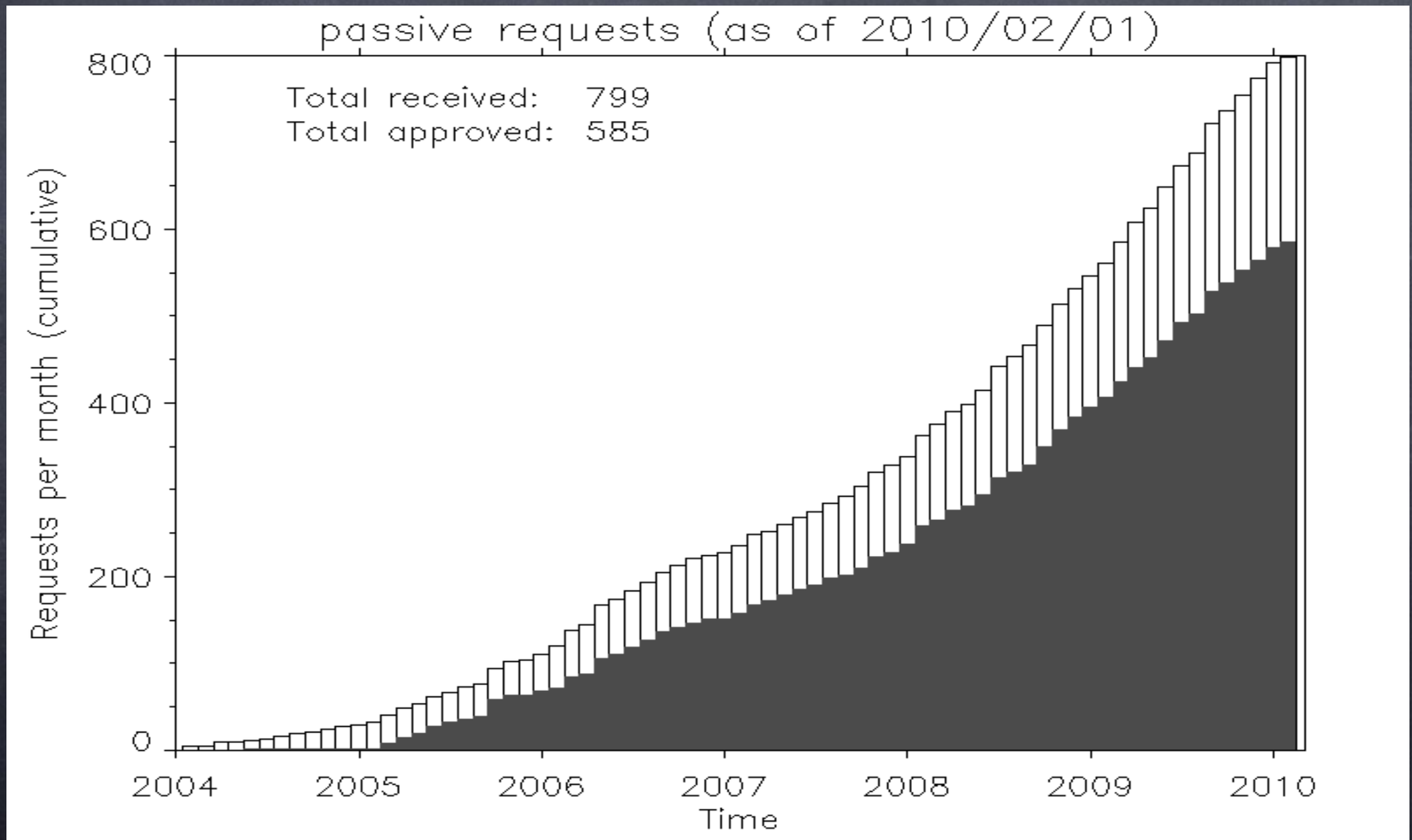
Data request stats (cont)

- Backscatter requests (monthly)



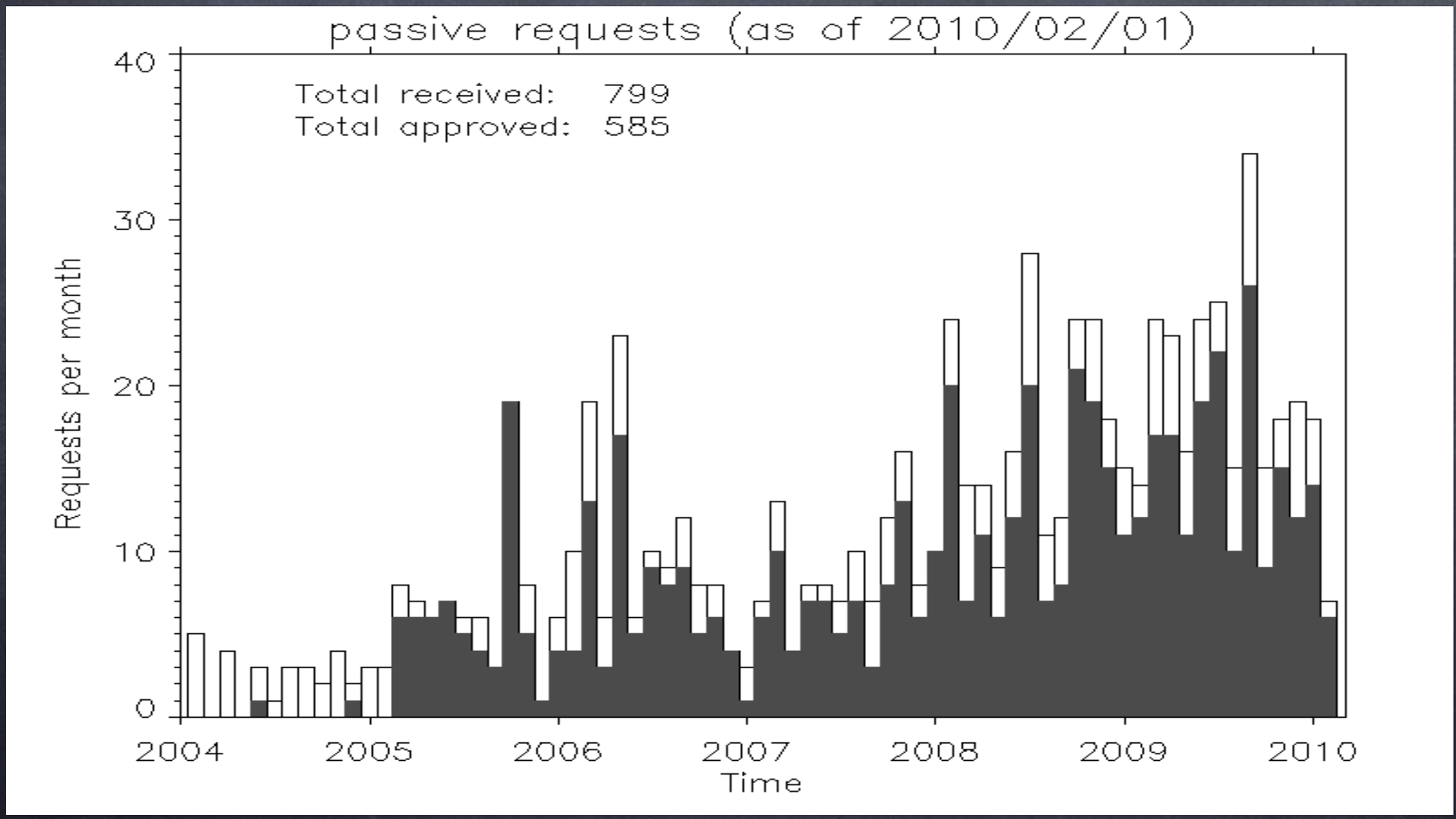
Data request stats (cont)

- Passive requests (cumulative)



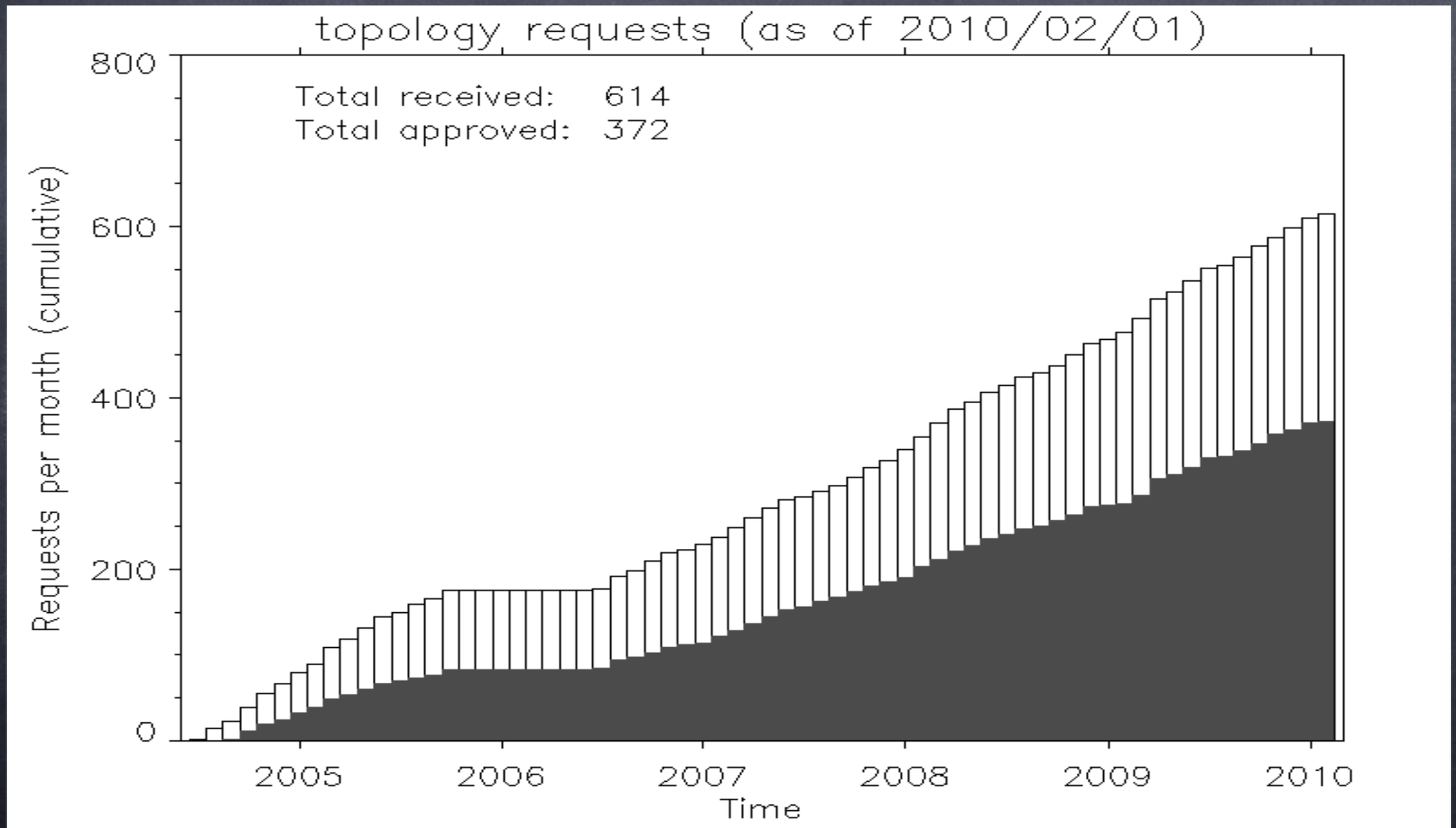
Data request stats (cont)

- Passive requests (monthly)



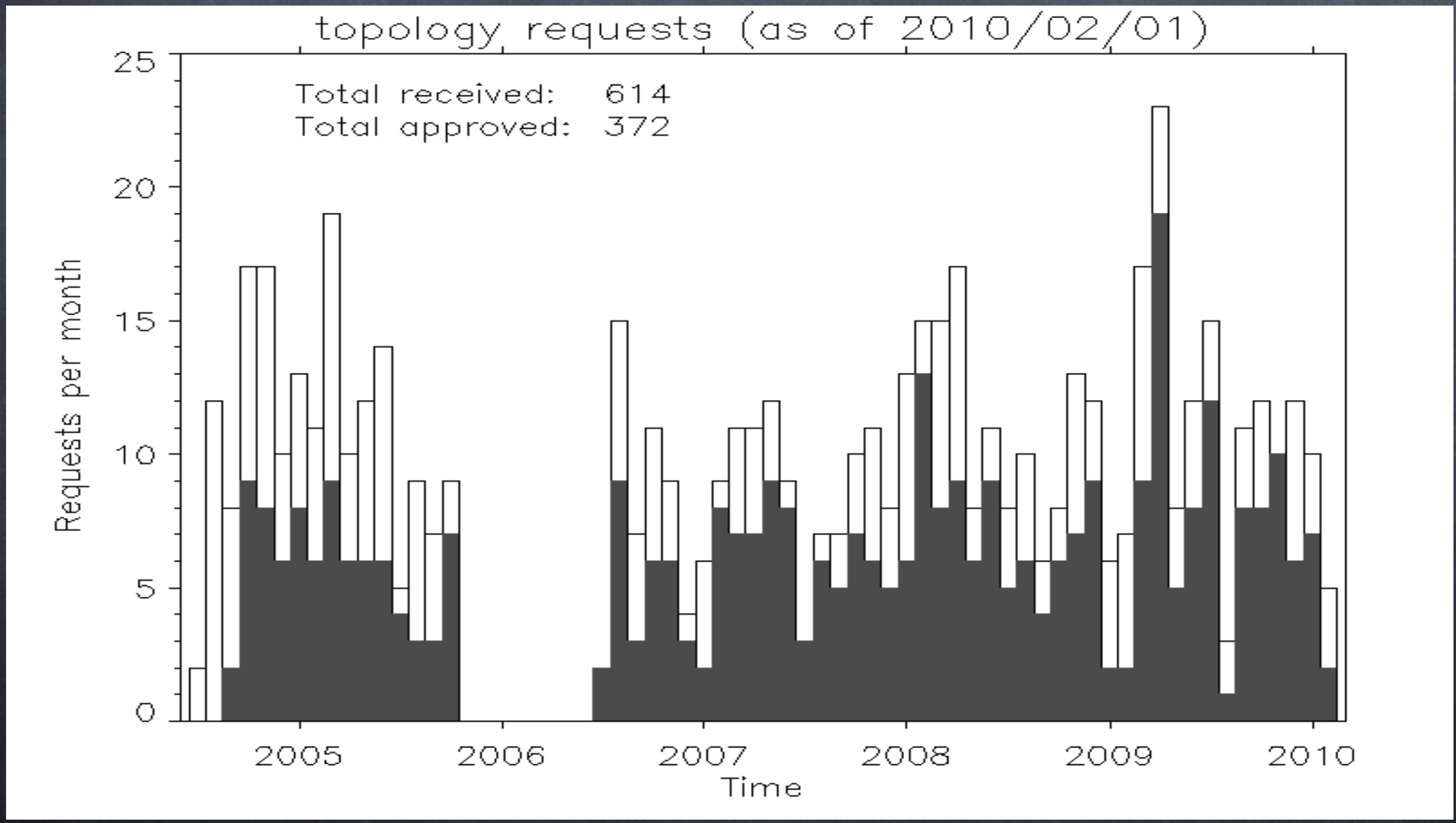
Data request stats (cont)

- Topology requests (cumulative)



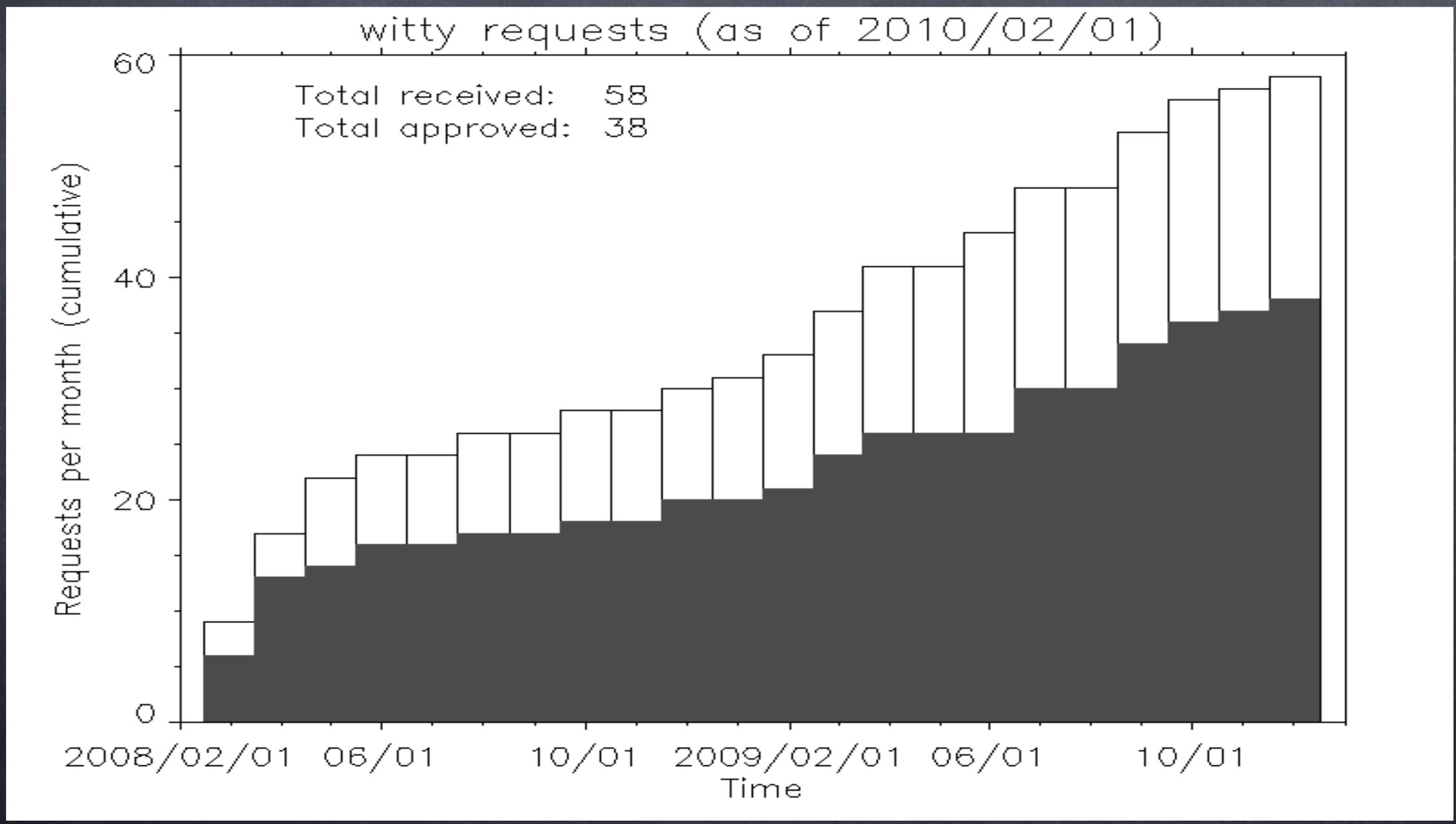
Data request stats (cont)

- Topology requests (monthly)



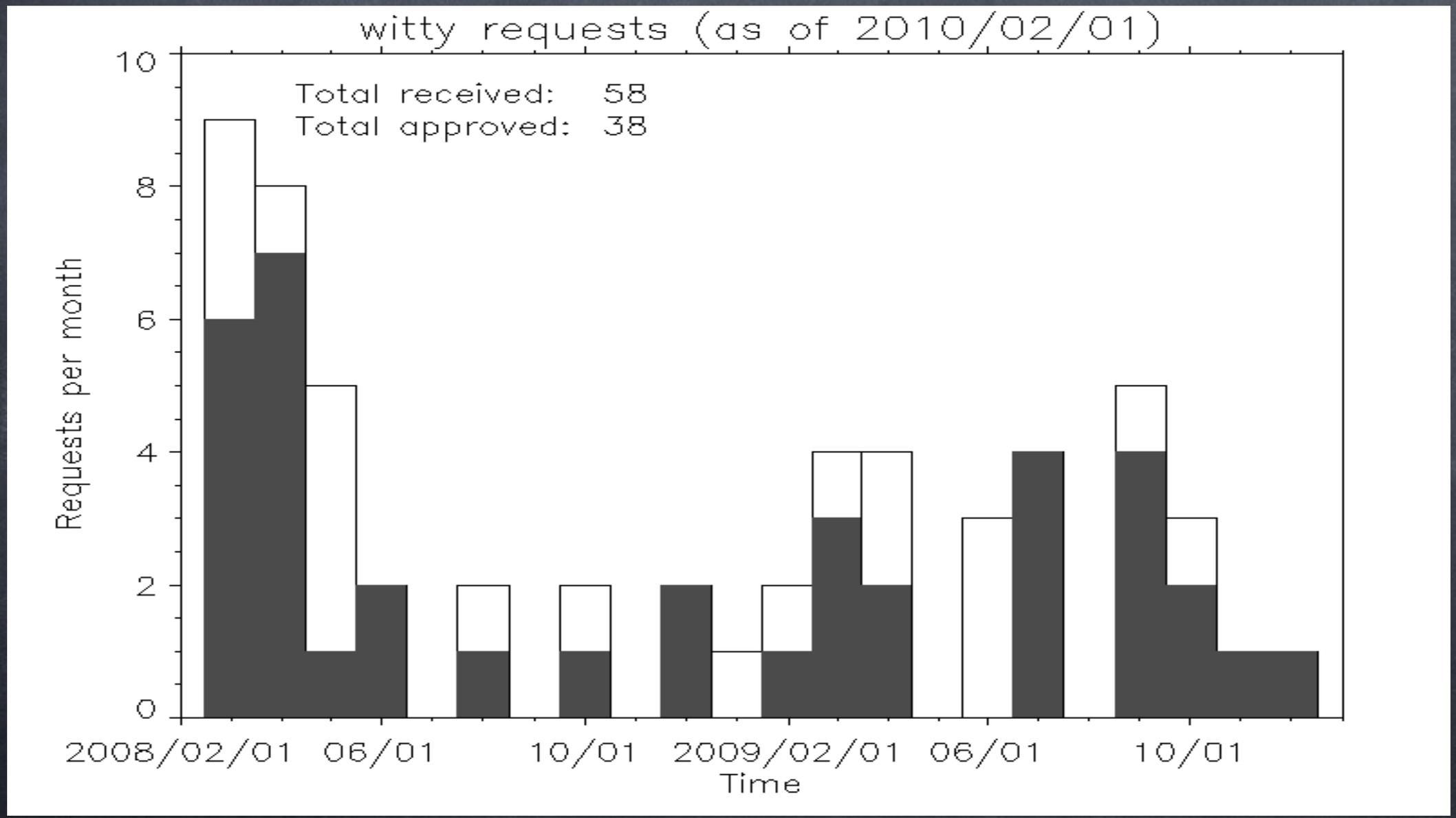
Data request stats (cont)

- Witty requests (cumulative)



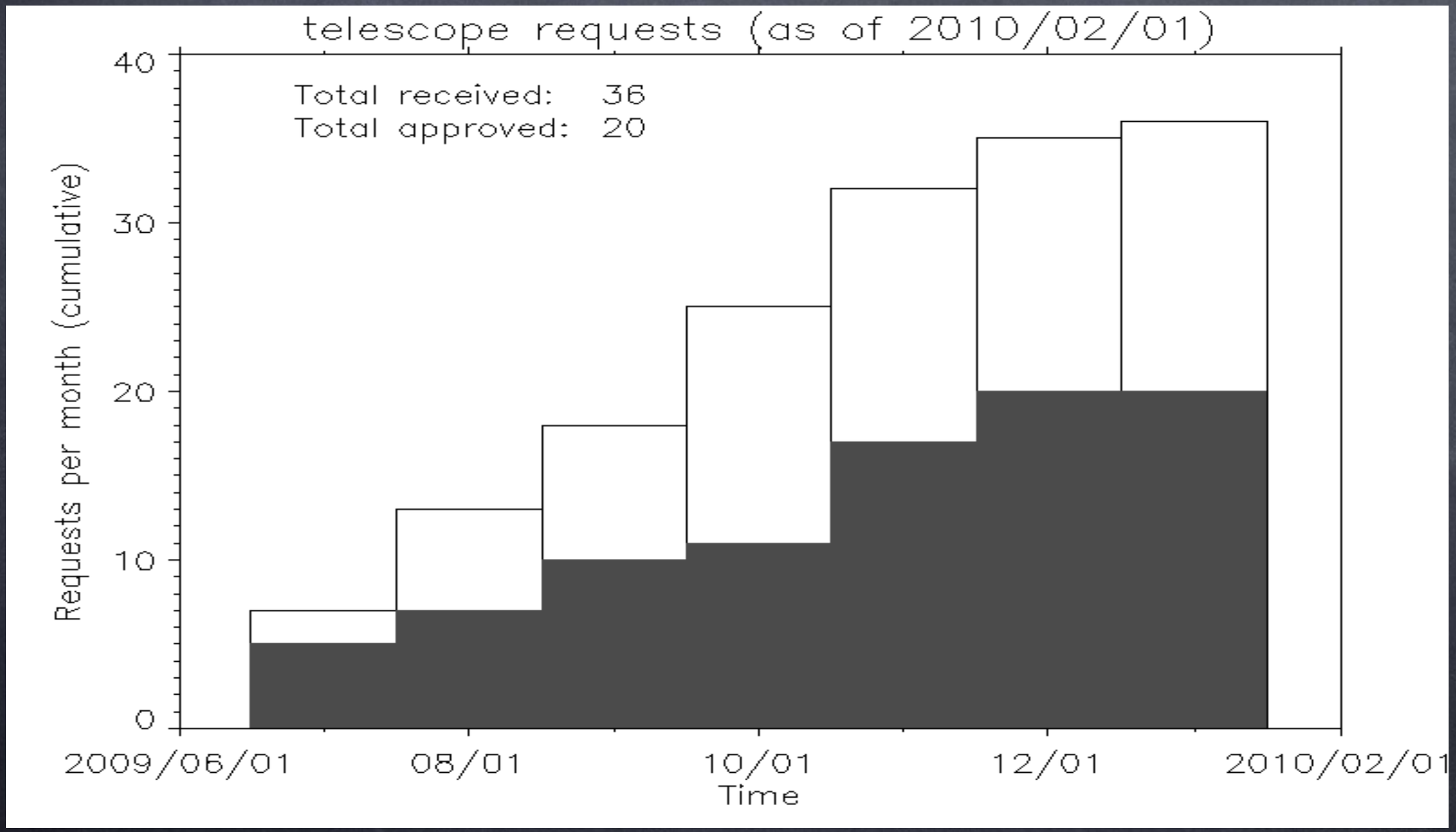
Data request stats (cont)

- Witty requests (monthly)



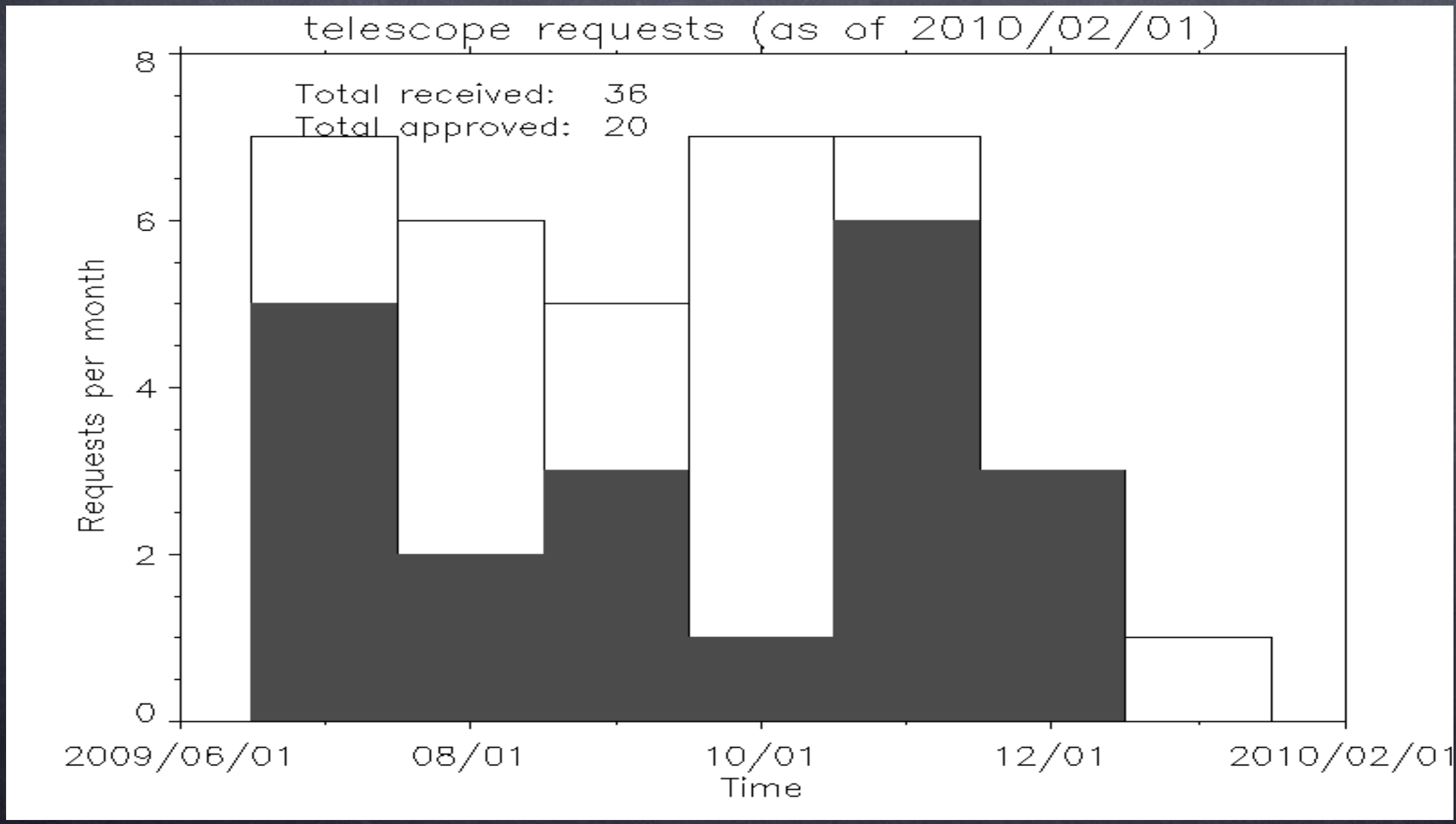
Data request stats (cont)

- Telescope requests (cumulative)



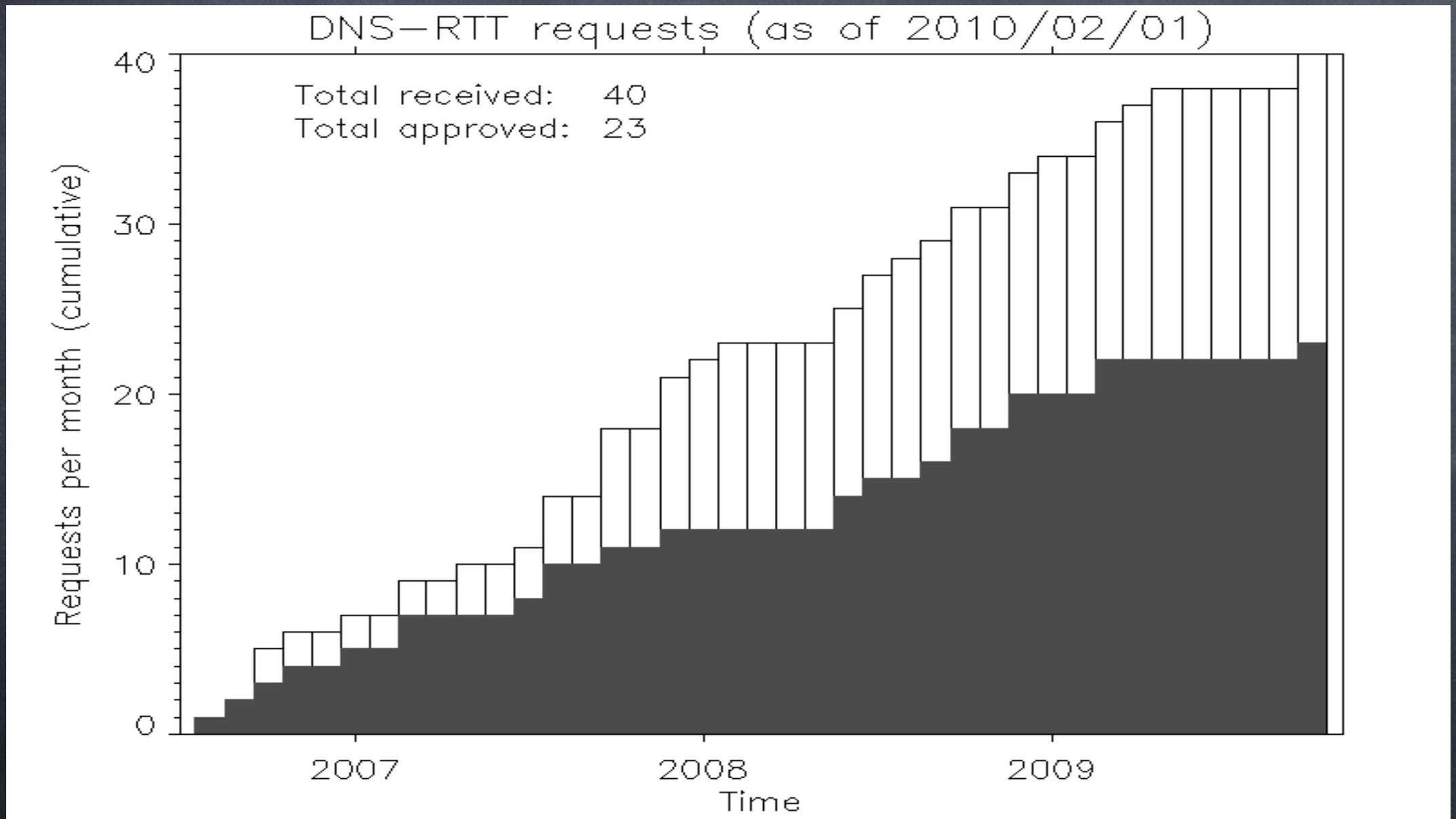
Data request stats (cont)

- Telescope requests (monthly)



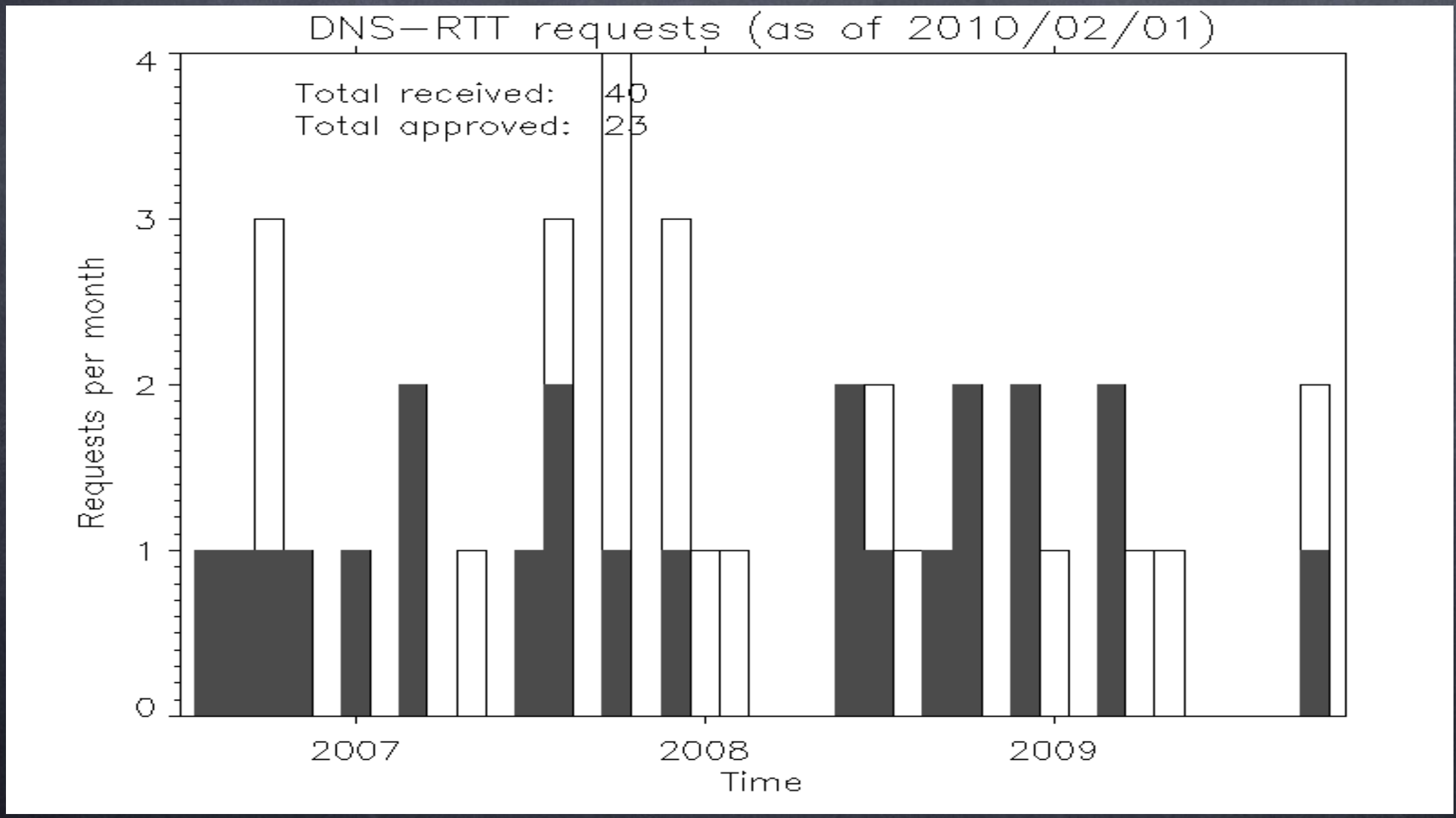
Data request stats (cont)

- DNS-RTT requests (cumulative)



Data request stats (cont)

- DNS-RTT requests (monthly)



Proposed phase 2 data sets

- Packet traces: longer traces, payload, other sites
- Internet2: better netflow, pkt traces, report gen.
  http://www.caida.org/data/realtime/passive/?monitor=sdnap
- UCSD Telescope: near real-time, unanonymized, payload

Open issues in PREDICT

- policy support: research/position papers
- privacy impact statement: needs repair
- no govt use of data: needs clarification
- no networks that serve public
- metadata catalog
- metrics for success
- community outreach: wikis, blogs, bofs, socialnets
- improved PR