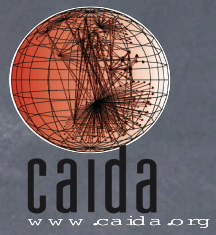
DHS PREDICT project: CAIDA update

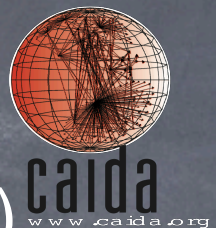


- Data collection updates
- Dataset dissemination statistics
- Other activities
- Open issues

Marina Fomenkov, *CAIDA*

December 1, 2010

Data collection - passive



- OC192 backbone: 11.5 TB compressed (8.5 TB in July)
- 2007-2009 data, Jan-Oct 2010
  - 5.4 anonymized
  - 6.1 unanonymized
- Problems:
  - no data in May
  - incomplete data in June and August
    - Chicago monitors were unreachable
  - persistent hardware problems w. Chicago monitor
- Plans:
  - package as the 2010 annual dataset
    - strip payload/L1/L2, transfer, anonymize, archive
  - collect 1 hr trace per mo = 200-250 GB
  - keep a quarterly sample - select the best quality

Data collection - passive

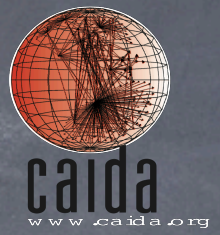
- UCSD telescope: 3.3 TB compressed (30 days window)
  - most recent month - “live” on disk
  - the previous month - backup on samqfs
    - current: Dec 2009 - Nov 2010
    - 30.1 TB compressed
  - applied for NSF funding
    - analysis
    - develop automated triggers and alerts
    - curate custom data sets upon request
    - explore “near real-time”, “bring code to the data” data sharing
- OC48 traces: 1.7 TB (2004 traces, anonymized, in PREDICT)

Data collection - active

- old *skitter* data (in PREDICT): 4 TB
  - discontinued in February 2008
- current Ark data: IPv4 topology 3.6 TB, IPv6 topology 2.2 TB
  - 53 monitors in 30 countries, 16 IPv6 capable
  - data curation:
    - create derivative data sets
    - aggregate in ITDK
      - router-level topologies: nodes and links
      - hostnames
      - router-to-AS assignment
      - geographical information

    http://www.caida.org/data/active/internet-topology-data-kit/
- applied for NSF funding to curate/analyze/annotate IPv6 data

how do we serve the data?



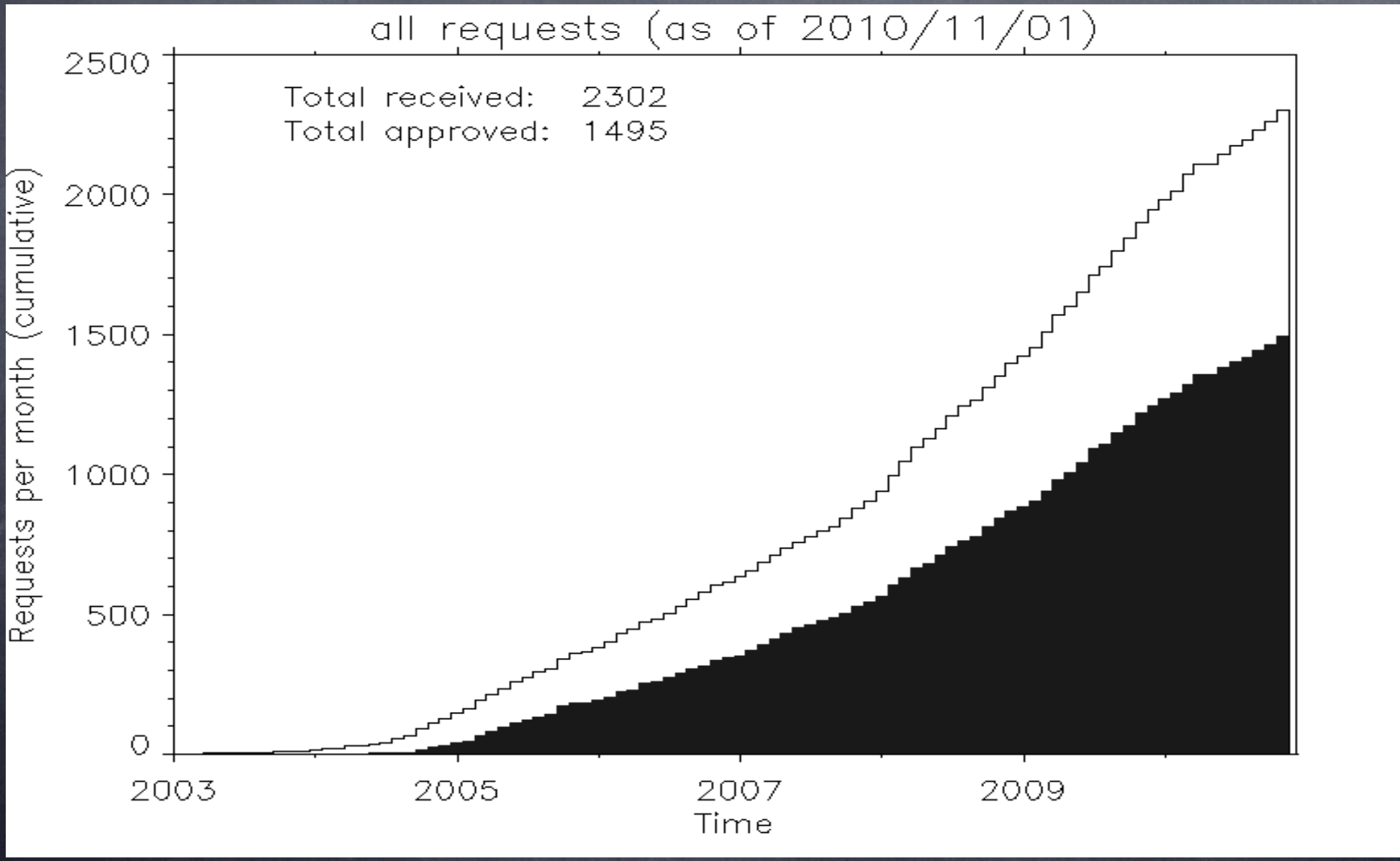
- PREDICT (OC48 traces, topology from *skitter*, telescope)
- Academics who sign AUP (OC192, topology from Ark, telescope)
- Derived data sets are publicly available (i.e., AS-links)
- Commercial researchers must join CAIDA
- Aggregated statistics online:
- OC192 backbone:
  - report generator: http://www.caida.org/data/realtime/passive/?monitor=equinix-chicago-dirA
- topology:
  - Ark statistics: http://www.caida.org/projects/ark/statistics/index.xml
  - For each monitor: path dispersion (AS and IP), path length distribution, RTT distribution, RTT vs. distance, median RTT per country

Requests for the data, 2010/2009

Dataset	Requests	Approved	Accessed	Served since
Backscatter	53/101	33/62	22/45	Feb 2003
Passive	136/242	104/181	89/151	Feb 2004
Topology	132/136	74/90	48/63	Jul 2004
Witty	12/28	10/18	9/14	Mar 2008
Telescope	23/35	17/20	13/16	Jul 2009
DNS-RTT	5/7	3/3	2/3	Aug 2006
	471/549	241/376	183/292	

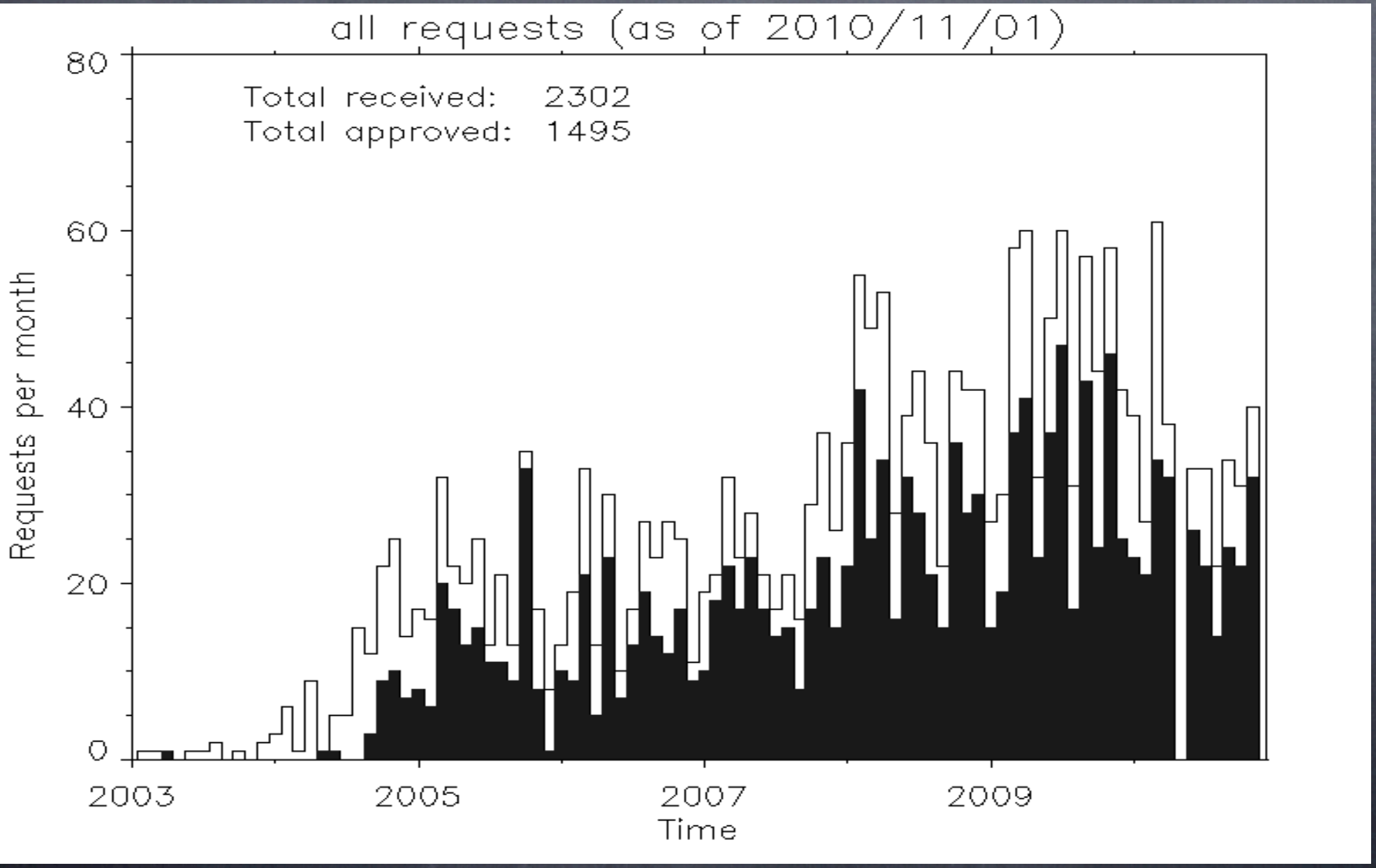
Data request stats

- All requests (cumulative)



Data request stats (cont)

- All requests (monthly)



Data Set Popularity

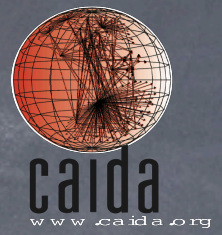
1st best - OC192 and OC48 traces

- popularity: requested 378 times, accessed 240 times (in 2009/2010)
- who used it: 201 .edu, 98 .cn, 38 .uk, 26 .com (since 2004) …
  - and 45 more domains

2nd best - topology data

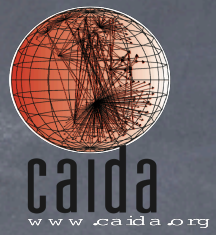
- popularity: requested 218 times, accessed 96 times (in 2009/2010)
- who used it: 212 .edu, 91 .cn, 30 .uk, 24 .kr, 22 .jp (since 2004) …
  - and 51 more domains

Publications using CAIDA data



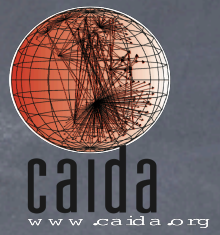
- OC192 and OC48 traces: traffic classification, performance modeling, monitoring, filtering, generation, locality
  http://www.caida.org/data/publications/bydataset/index.xml#passive
  - 76 publications (54 from data in PREDICT)
- UCSD telescope: Conficker, worm research
  http://www.caida.org/data/publications/bydataset/index.xml#Backscatter
  - 26 publications (all from data in PREDICT)
- topology: pkt traceback, marking, DOS defense, topo and routing modeling, discovery, metrics, improvements
  http://www.caida.org/data/publications/bydataset/index.xml#Topology
  - 55 publications (44 from data in PREDICT)

Recent publications



- E. Kenneally and kc claffy, *Dialing privacy and utility: a proposed data-sharing framework to advance Internet research*, *IEEE Security & Privacy* special issue, July 2010.
- A. Dianotti and kc claffy, *Obstacles and challenges to traffic classification*, submitted to IEEE Network.
- *AIMS-2 workshop report* published in ACM SIGCOMM CCR Online, October 2010.
- E. Kenneally presented *Can Network Science Help Re-write the Privacy Playbook* at the W3C Workshop on Privacy and Data Usage Control, October 2010

Meta-data for packet traces



- OC192 data: 2008-2009, Jan-Oct 2010
  - an hour-long trace every month
  - usually, 3rd Thursday, 13:00 - 14:00 UTC
- OC48 data: 2002-2003

- Publicly available statistics:
  - Date, start time, stop time
  - Numbers of IPv4, IPv6, unknown packets
  - Transmission rate in pkts/s, bits/s
  - Link utilization (%)
  - Average packet size & graph of packet size distribution

http://www.caida.org/data/passive/trace_stats/

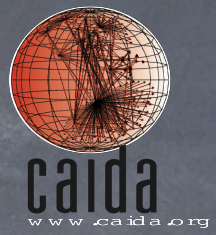
Phase II Data Sets

- Provided data set descriptions for:
- OC192 backbone: 2007-2010
- UCSD telescope: near real time
- topology: Ark data (ongoing)
  - IPv4 Routed /24 Topology dataset
  - IPv4 Routed /24 DNS Names dataset
  - IPv6 Routed Topology dataset
- topology: updated ITDK 2010

Revisions of CAIDA policies

- Telescope data (near real-time data set)
  - different from previous packaged data
  - simplified and streamlined the AUP language
  - Immediate use by postdoc A. Dianotti and his student
- ARK hosting sites
  - changed the document from Site AUP to Memorandum of Cooperation
  - began using for new sites in September 2010
  - gradually update already participating sites
- Passive data collection MOC
  - Currently under review (almost finished)

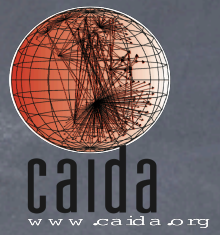
Analysis of CAIDA AUPs



- 4 categories of data - different levels of sensitivity
  - real-time telescope data
  - passive traces
  - active traces
  - derived topology
- Uncontrolled proliferation
  - 7 data request forms
  - 22 data set web pages
  - 22 README files

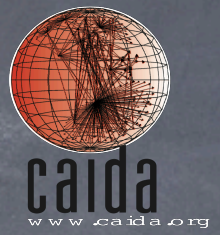
Goal: create a master AUP

Analysis of CAIDA AUPs



- Access conditions
  - Accreditation, validation, transparency
- Use restriction
  - Purpose, probing, other
- Disclosure obligations
  - Publication, 3rd party transfer, attribution
- Enforcement
  - Compliance, attestation
- Corrections / amendments
  - Measurement error notifications
- Disposition
  - Account closure, renewal
- Policy Vehicle: AUP, MOA, MOC…

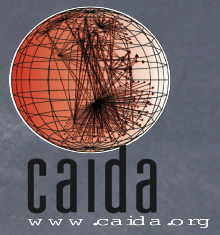
CAIDA Marketing Efforts



- CAIDA web site
  - Annual reports, Program Plan, Project web page
  - will blog about Phase II
- Presentations
- Publications
- Connections
- CAIDA workshops
- NSF channels
  - Broader Impact activity
  - Synergy in proposals
  - Workshops

How to google for PREDICT?

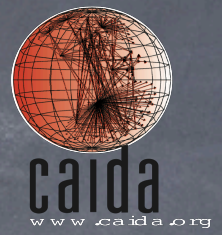
Necessary conditions of success



- Convenience
- Marketing
- Regular updates with newest data

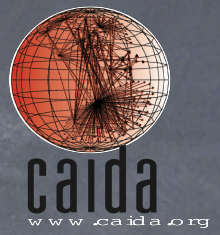
Will Phase II be the right answer?

Open issues for Phase II



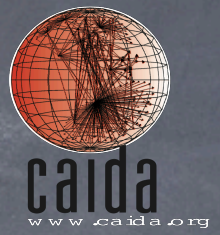
- Improve the Portal - both “how it looks” and “how it works”
  - Version 4.1 was a disappointment…
- Revise meta-data to be made public - at this meeting?
- List of keywords - where? Or when?
- MOA revisions - we will need time!
  - At least 30 days to produce 1st draft
  - At least 30 days for iterative editing
  - Current Action Plan says December 31st…

Open issues for Phase II (cont.)



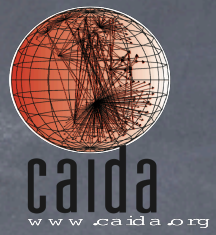
- How to organize meta-data? - not an easy problem!
  - how many data sets? tens? hundreds?
  - presentation
  - hierarchy
  - scalability
  - searchability
- Data categories descriptions - fix? (or eliminate?)
  - may be redundant if actual meta-data are posted
  - already too many and will grow
  - standard template
  - coherent technical editing

Other Open Issues



- Policy Section for the Portal - yes or no? or later?
- Metrics to track progress?
- PREDICT: 2.6 rq/mo
- CAIDA: 45 rq/mo, 27 appr/mo (not counting publicly available)
- PREDICT marketing “1-pager” - status?
- Canonical Data Sets - status?
- Privacy Impact Assessment statement - status?

Next PI meeting



- CAIDA offers to host
- When?

Welcome to sunny (or rainy) San Diego!