Access and Analyze Broadband Measurements Collected using M-Lab

Tiziana Refice
Google

Google™

Broadband Transparency

- What's the actual performance of the broadband connectivity that you receive from your ISP? Is it what is specified in your contract?
- Is your ISP interfering (e.g., blocking, throttling) with certain applications or web sites?
- **Real numbers, not marketing**!

Broadband Transparency is (also) about

- **Measuring** end-users' broadband performance.
- **Sharing** measurement results (with end-users and policy makers).

Google

Measurement Lab (M-Lab)

- **Open**, distributed ***server platform*** for ***researchers***, to deploy tools to measure broadband performance.
  - Well-provisioned and well-connected servers.
- **Open** ***tools*** for ***users***, to test their broadband connection.
  - Required to publish source code.
- **Open,** better ***data*** for ***everyone***, to build on a common pool of network measurement data.
  - Raw data, not just aggregated numbers.

M-Lab believes that
open tools and open data access facilitate
peer-review, independent analysis & better research.

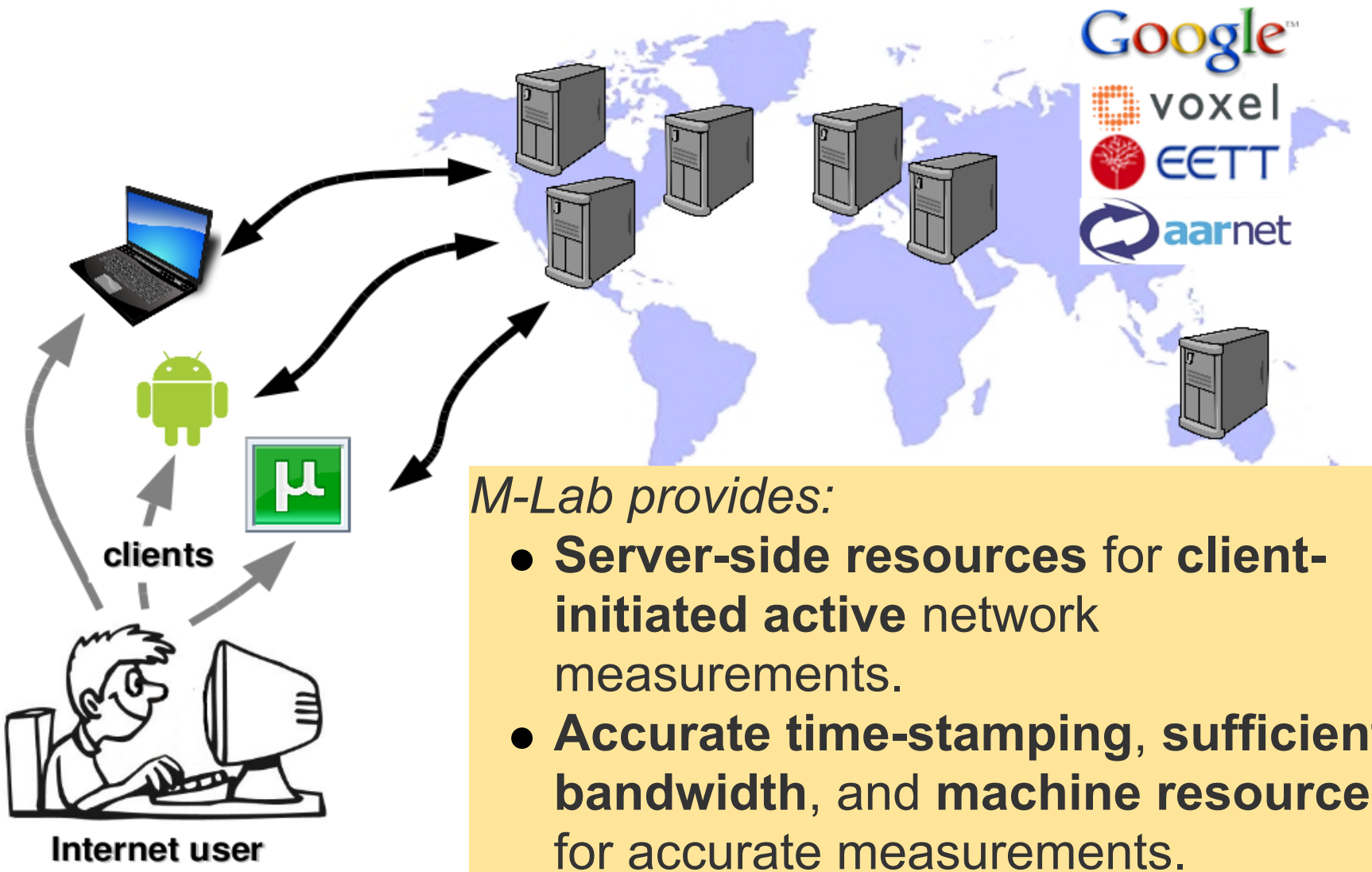
Google™

Founders & Supporting partners

- Open Technology Institute (at New America Foundation)
- PlanetLab Consortium
- Google
- Academic researchers:
  - Georgia Institute of Technology,
  - Internet2,
  - Max Planck Institute,
  - Northwestern University,
  - Pittsburgh
  - Supercomputing Center,
  - Princeton University
- Servers hosting & connectivity
  - Google, voxel, EETT, aarnet
- Client support
  - uTorrent, FCC, SamKnows
- DNS hosting & server selection
  - Princeton University
- Data hosting
  - Amazon, Google

Google™

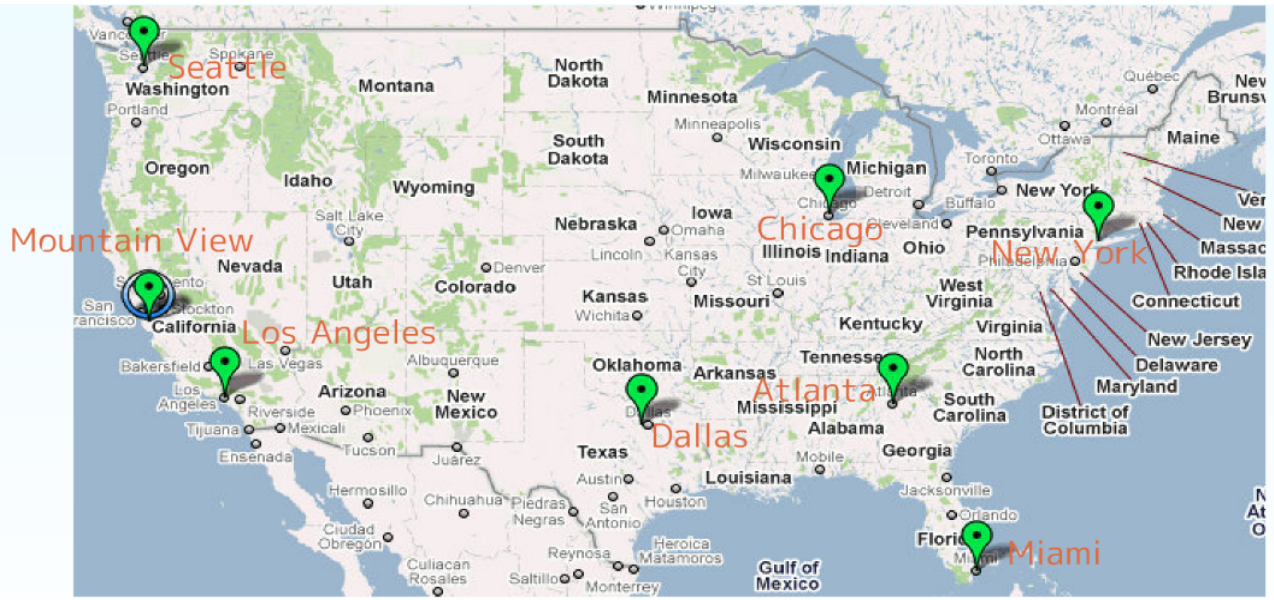
M-Lab platform



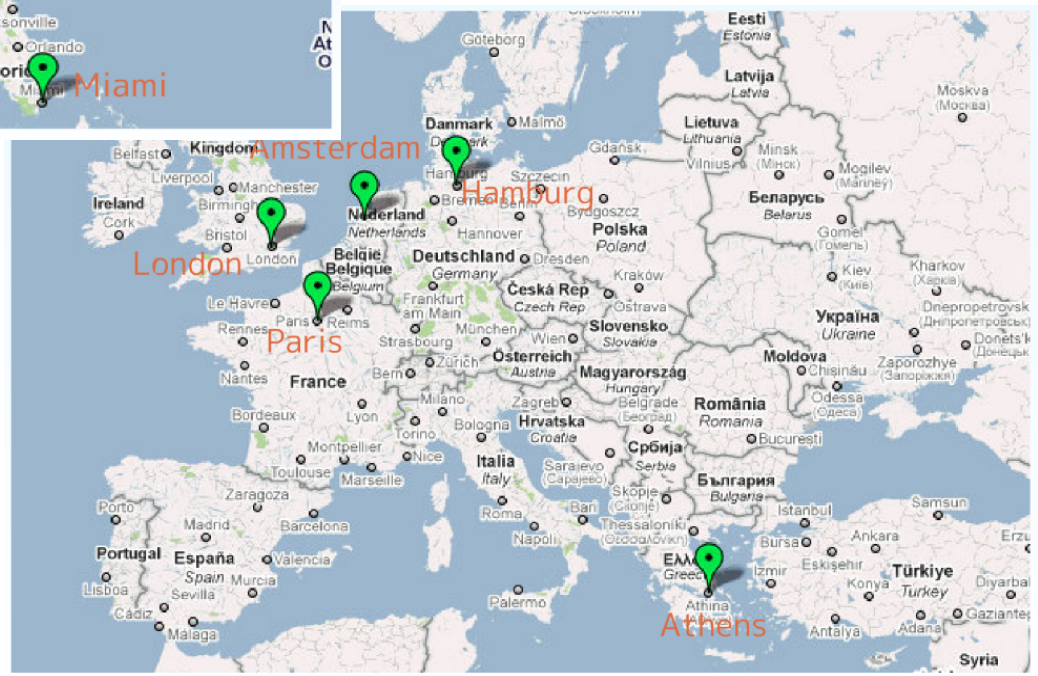
M-Lab provides:

- **Server-side resources** for **client-initiated active** network measurements.
- **Accurate time-stamping**, **sufficient bandwidth**, and **machine resources** for accurate measurements.

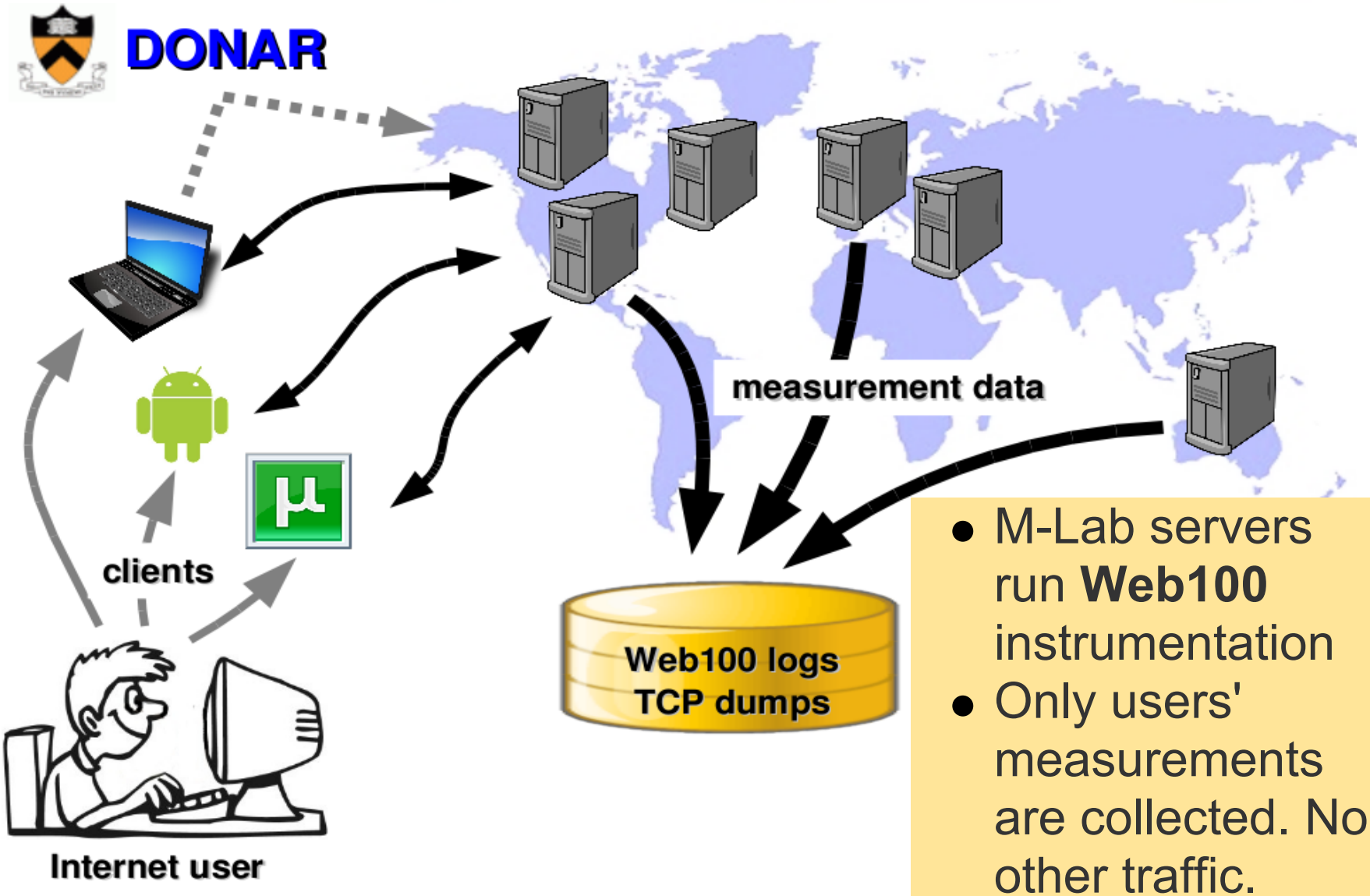
M-Lab nodes



48 servers in
16 locations
(NY and Amsterdam have
2 locations each)



Measurement & Data collection



- M-Lab servers run **Web100** instrumentation
- Only users' measurements are collected. No other traffic.

Currently available tools (1)

- **Network Diagnostic Tool** (**NDT**)

Tests user's connection speed and reports sophisticated diagnosis of problems limiting the speed.

  - Used by the **FCC Consumer Broadband Test**
  - **http://www.broadband.gov/**
  - Integrated with **µTorrent client**

- **Network Path and Application Diagnosis** (**NPAD**)

Diagnoses common problems that impact last-mile broadband networks.

- **Glasnost**

Tests whether traffic from user's applications is being rate-limited (i.e., throttled) or blocked.

INTERNET2®

PITTSBURGH SUPERCOMPUTING CENTER

Max Planck Institute for Software Systems

Google™

Currently available tools (2)

- **Pathload2**

Tests user's available bandwidth.

- **ShaperProbe**

Determines whether an ISP is performing traffic shaping.

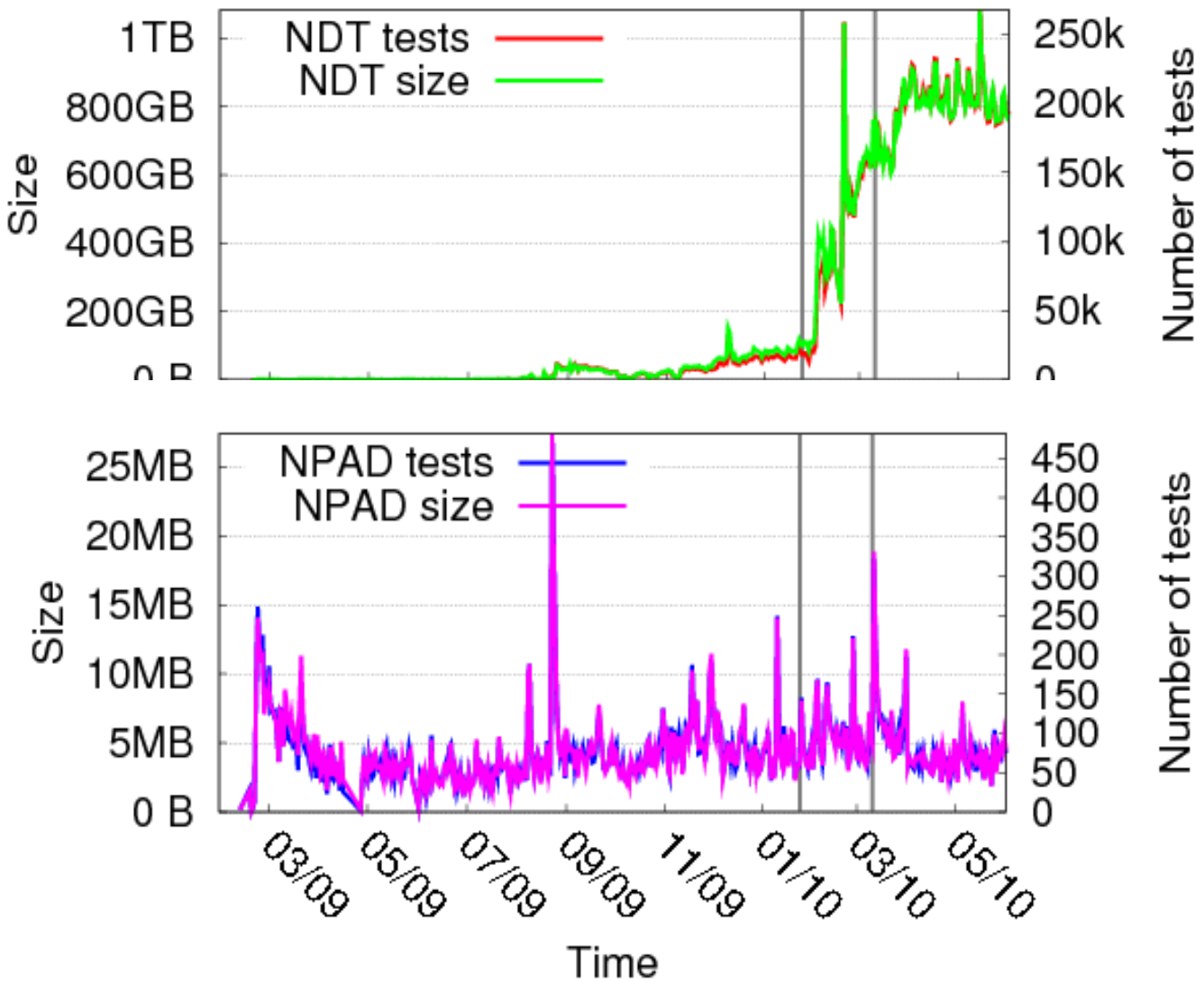
- **Windrider**

Detects whether your mobile broadband provider is performing application or service specific differentiation.

Georgia Tech

Google™

How much data? How many tests?



Jan 25 2010
uTorrent launch
Mar 11 2010
FCC launch

NDT

Tot tests: 22**M**

Tot size: 93**TB**

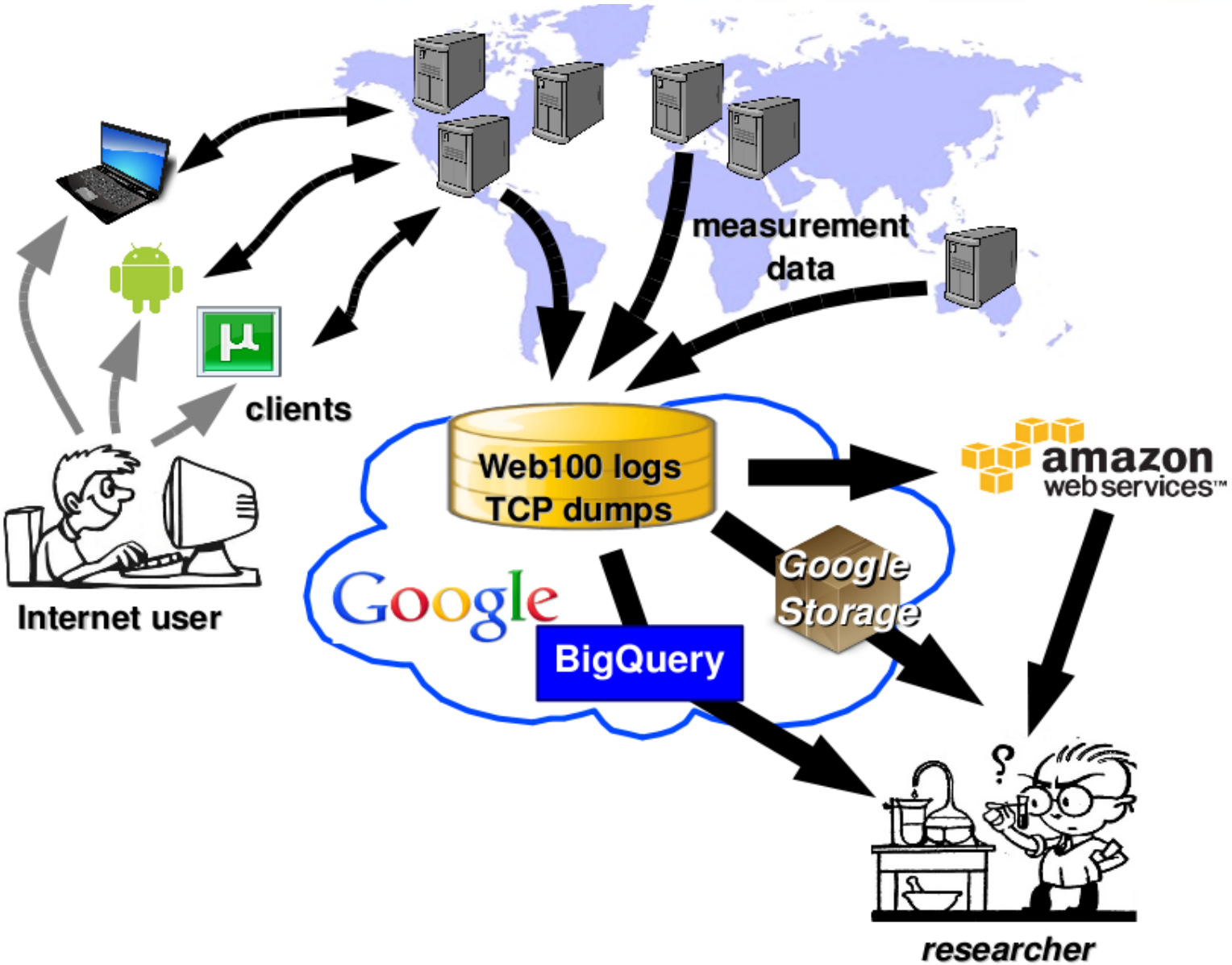
NPAD

Tot tests: **34K**

Tot size: **2GB**

Google

Data sharing



You can download the raw data ...

- M-Lab keeps all the data collected since Feb 2009.
- Data shared through
  - Amazon EC2
  - Google Storage (storage service in Google's cloud)
- Datasets currently shared:
  - Web100 logs, TCP dumps, tools' analysis results.
  - Collected by NDT, NPAD, and Glasnost.
  - *Other tools' datasets coming soon*.
- Amount of shared data (as of June 6 2010):
  - **4TB** compressed, split into daily tarballs.
  - **97TB** (!!) uncompressed.

Google™

You can download the raw data ...

- You probably want to download the raw data if you plan to run a detailed and extensive analysis.
  - *e.g., What are the traffic trends of US users?*
- But, what if you are only looking for aggregated numbers? Or if you are only interested in a small subset of the whole dataset?
  - *e.g., What's the distribution over time of the average RTT in the last two year, split by country?*
- Working with the raw data requires:
  - *Time* and *bandwidth* to *download* the data.
  - *Storage* to *keep* the data.
  - *CPU* to *analyze* the data.

Google™

... or analyze it using BigQuery

- BigQuery is a web service provided by Google that allows to run interactive analysis of huge datasets.
- Why are we using BigQuery for M-Lab?
  - **Scale** - Terabytes of data.
  - **Speed** - Analysis of billions of rows in seconds.
  - **Simplicity** - SQL-like query language.
- M-Lab data in BigQuery:
  - Web100 logs collected by NDT, NPAD. (Other datasets will be added based on your interest.)
  - **60 Billion** rows.
  - **~40 sec** (!!) to execute a query against the whole dataset.

Google™

A sample query on BigQuery:

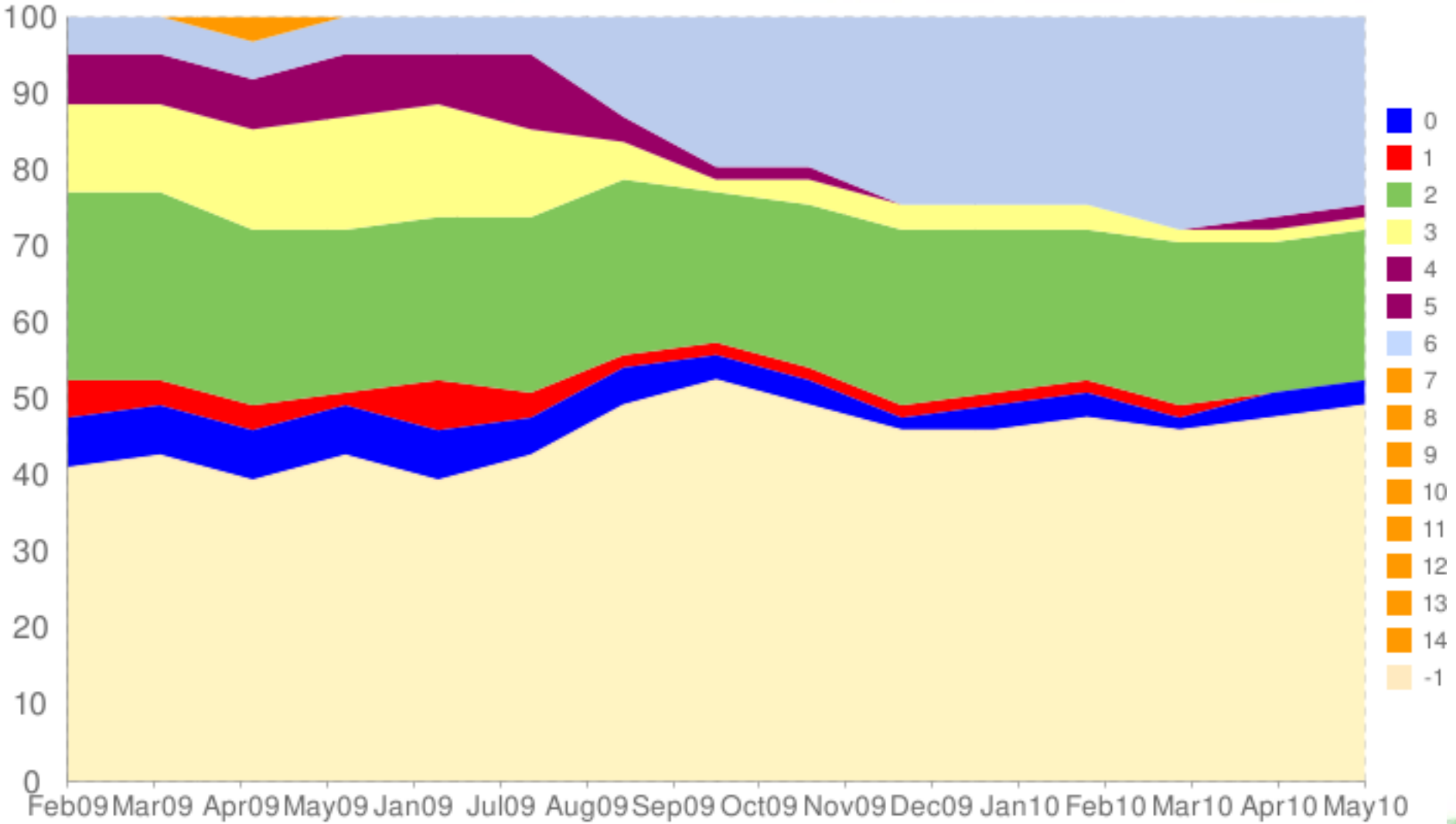
WinScaleRcvd usage across clients' IPs

```
> SELECT
>   STRFTIME_UTC_USEC(log_time, "%%Y-%%m") AS month,
>   web100_log_entry.snap.WinScaleRcvd AS WinScaleRcvd,
>   COUNT (DISTINCT web100_log_entry.connection_spec.remote_ip)
>     AS client_ip_addresses
> FROM [bigquery/samples/mlab]
> GROUP BY month, WinScaleRcvd
> ORDER BY month, WinScaleRcvd ASC;

month      WinScaleRcvd   client_ip_addresses
-------------------------------------------
  2009-02          -1              464
  2009-02           0               70
  2009-02           1               53
[...]

Execution time: 33.206 seconds
218 rows
```

Google™

A sample query on BigQuery:

WinScaleRcvd usage across clients' IPs



Google™

More information about using BigQuery to analyze M-Lab data

http://code.google.com/apis/bigquery/docs/dataset-mlab.html

In particular, more examples in this codelab:
http://code.google.com/apis/bigquery/docs/codelab-mlab.html

Google™

GET INVOLVED!

- **Analyze collected data** and share your results with the Internet community.
  - If you are interested in using BigQuery, provide **feedback** about which datasets to include in BigQuery. (Currently, only Web100 logs generated by NDT and NPAD.)
- Contribute **tools** (client- and/or server-side).
- Provide **servers** and **network connectivity** for the platform.
- Provide resources for **data hosting** and **publication**.
- Provide **funding** to support the above.

Google™

Questions?

http://www.measurementlab.net/contact

More information at
www.measurementlab.net

Google™