

The Plane distributed measurement infrastructure

Overview, insights & hindsights

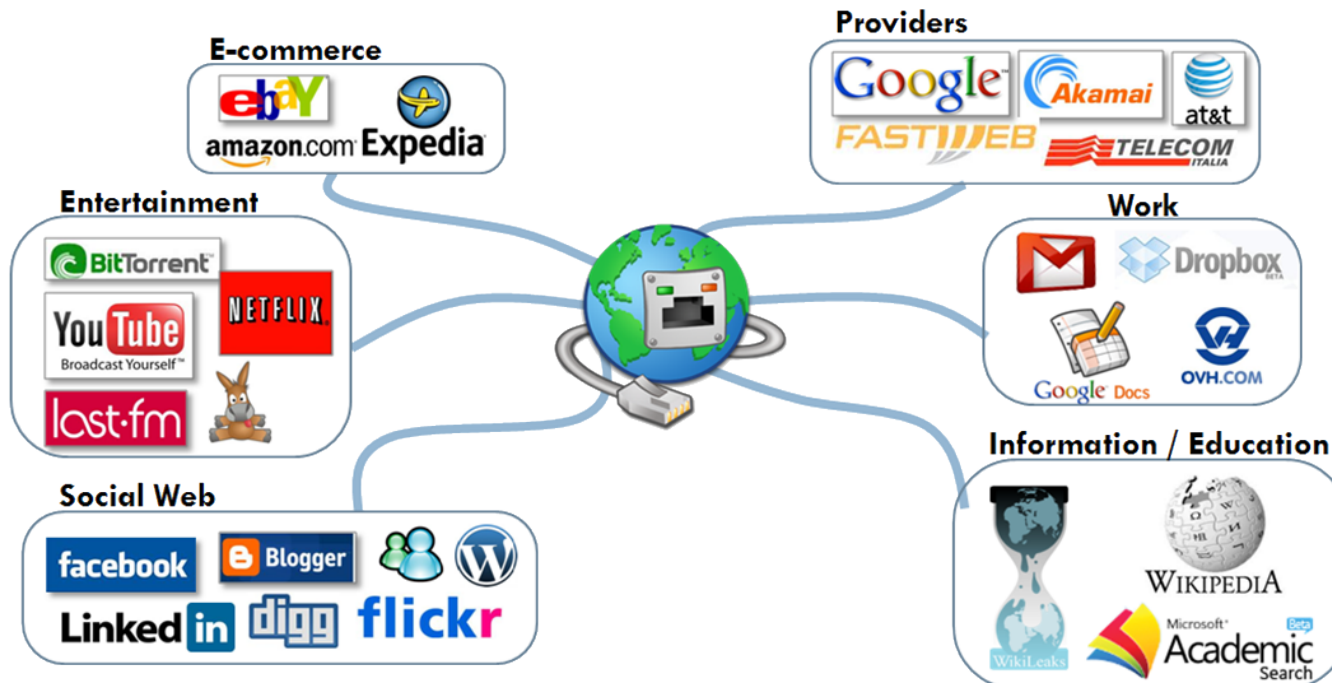
Dario Rossi
Professor

dario.rossi@telecom-paristech.fr

<http://www.telecom-paristech.fr/~drossi>



Today Internet



Why is **skype**™ not working?

Which ISP is the best in my area?

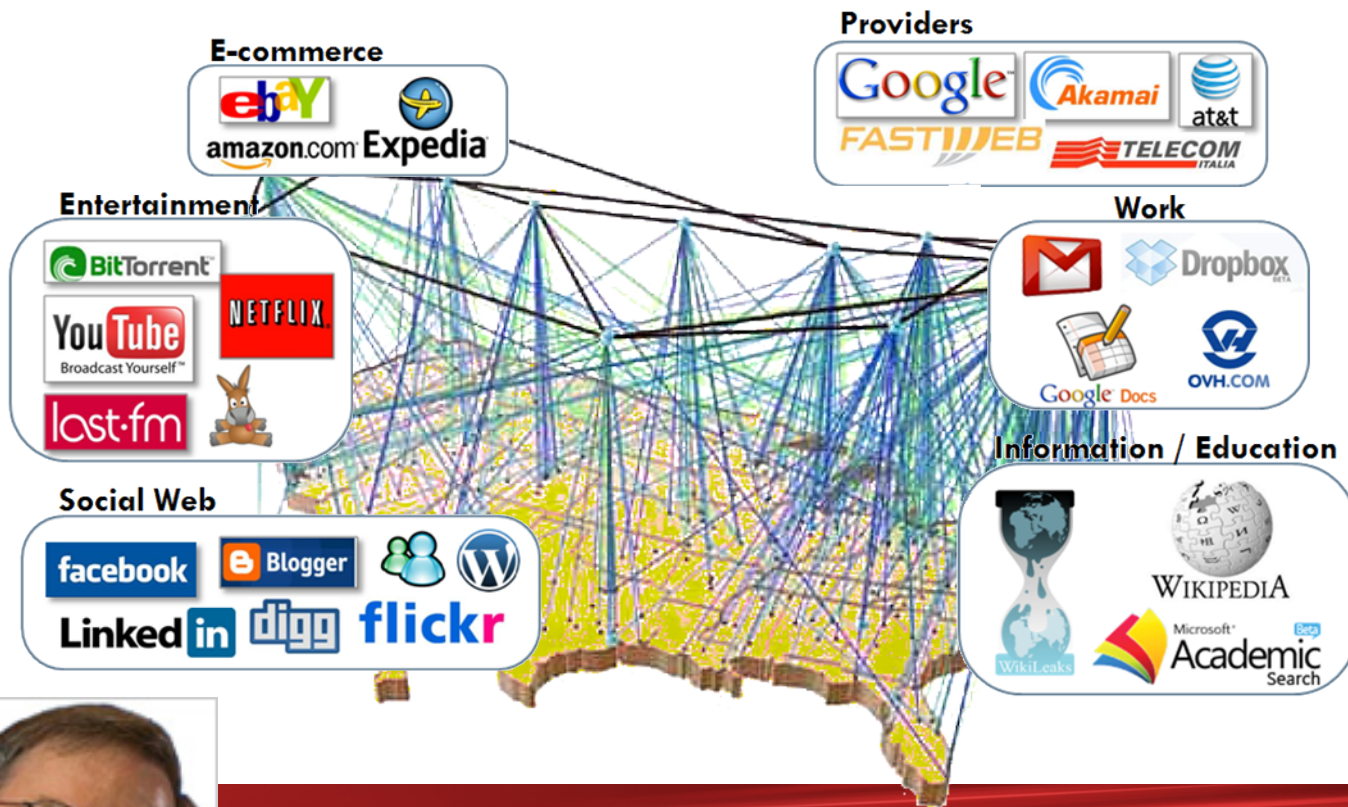
How does it work? How to do better?

How to keep users happy & engaged?

How to optimize my network for users apps?

Where is **YouTube** traffic coming from?

Today Internet



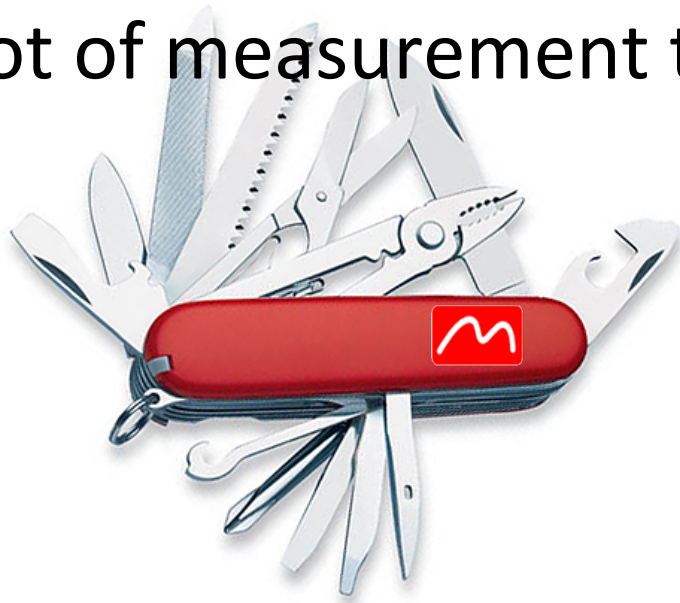
“The Internet is the first thing that humanity has built that humanity doesn't understand, the largest experiment in anarchy that we have ever had.”

Eric Schmidt – ex Google Exec. Chairman

Internet easurement

Shed light on the Internet operational obscurity

Lot of measurement tools...



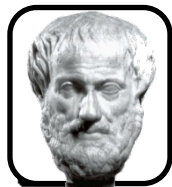
... issing orchestration

 Plane: avoid to reinvent the wheel
& assist in building automated pilots!



Two kinds of measurements

- Passive
 - Observe network traffic without interference
 - Similar to Aristotle's observational method
- Active
 - Perturb the network & measure its reaction
 - Similar to Newton's experimental method



Ἀριστοτέλης
@aristotle

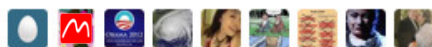
 Follow

All science is either practical, poetical or theoretical (Metaphysics)

 Reply  Retweet  Favorite

1388
RETWEETS

372
FAVORITES



Sir Isaac Newton
@newton

 Follow

Every body continues in its state of rest, or of uniform motion in a right line, unless it is compelled to change that state by forces impressed upon it (Principia)

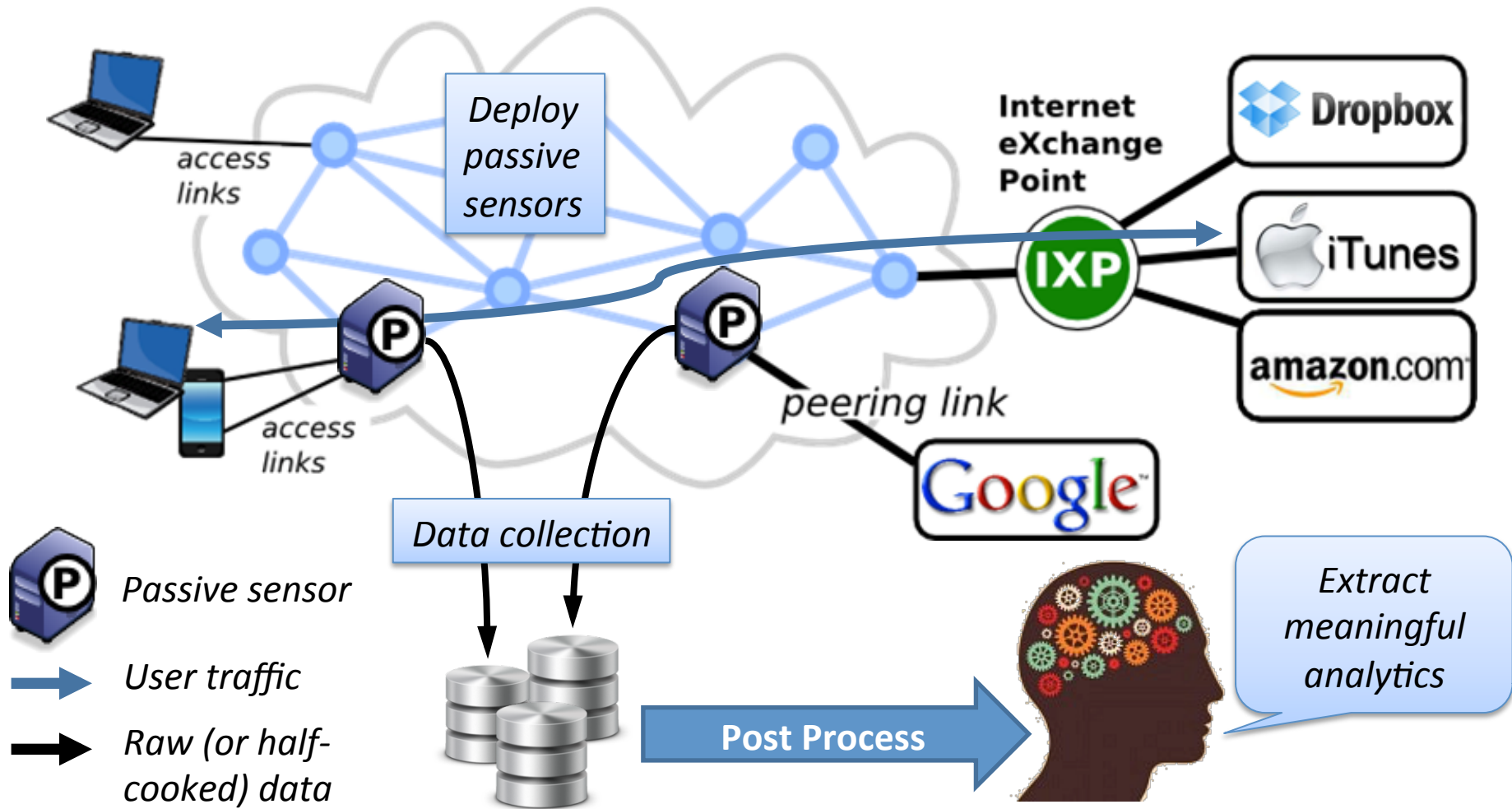
 Reply  Retweet  Favorite

2411
RETWEETS

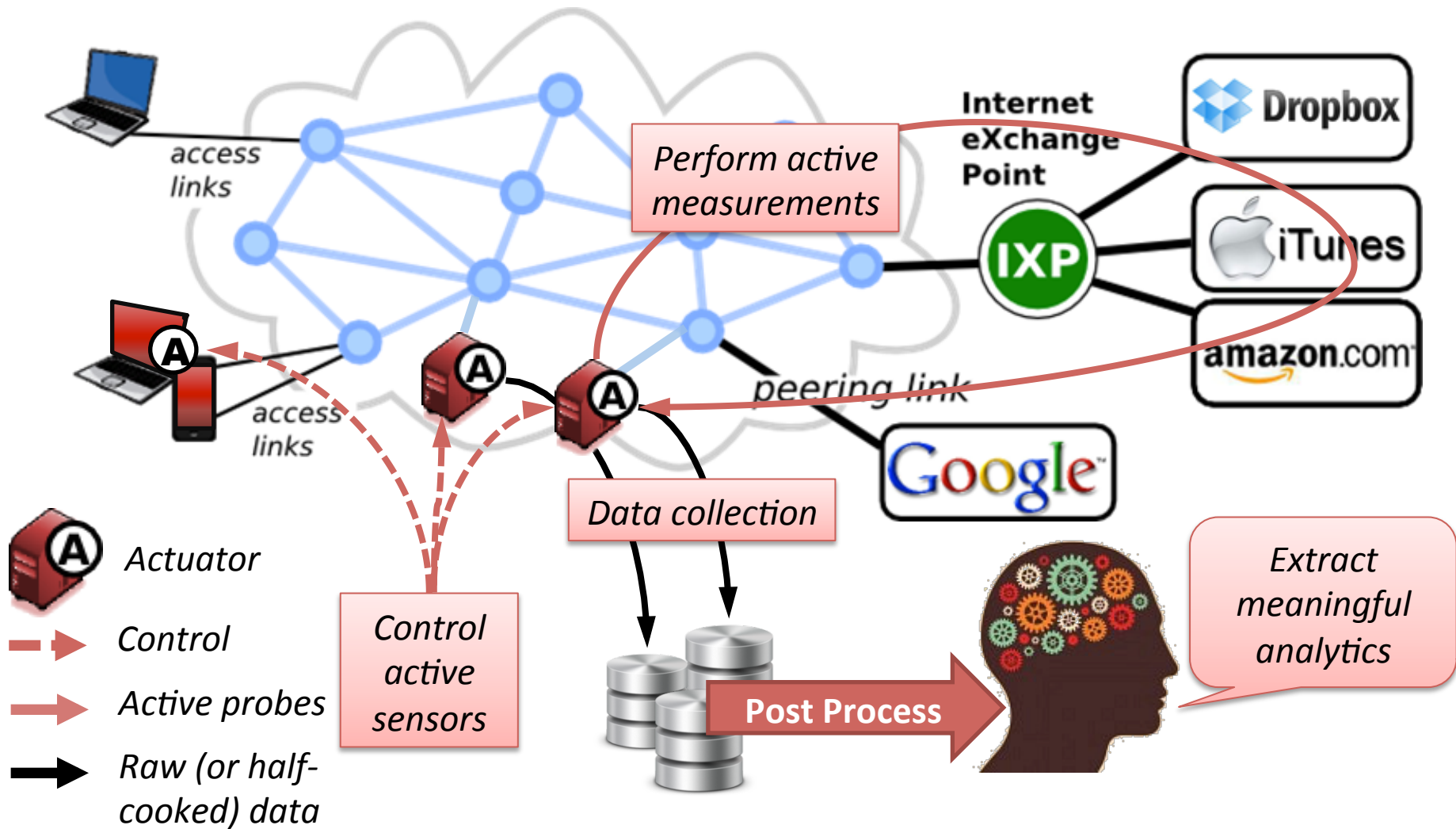
353
FAVORITES



Passive easurements



Active measurements



Merging all together

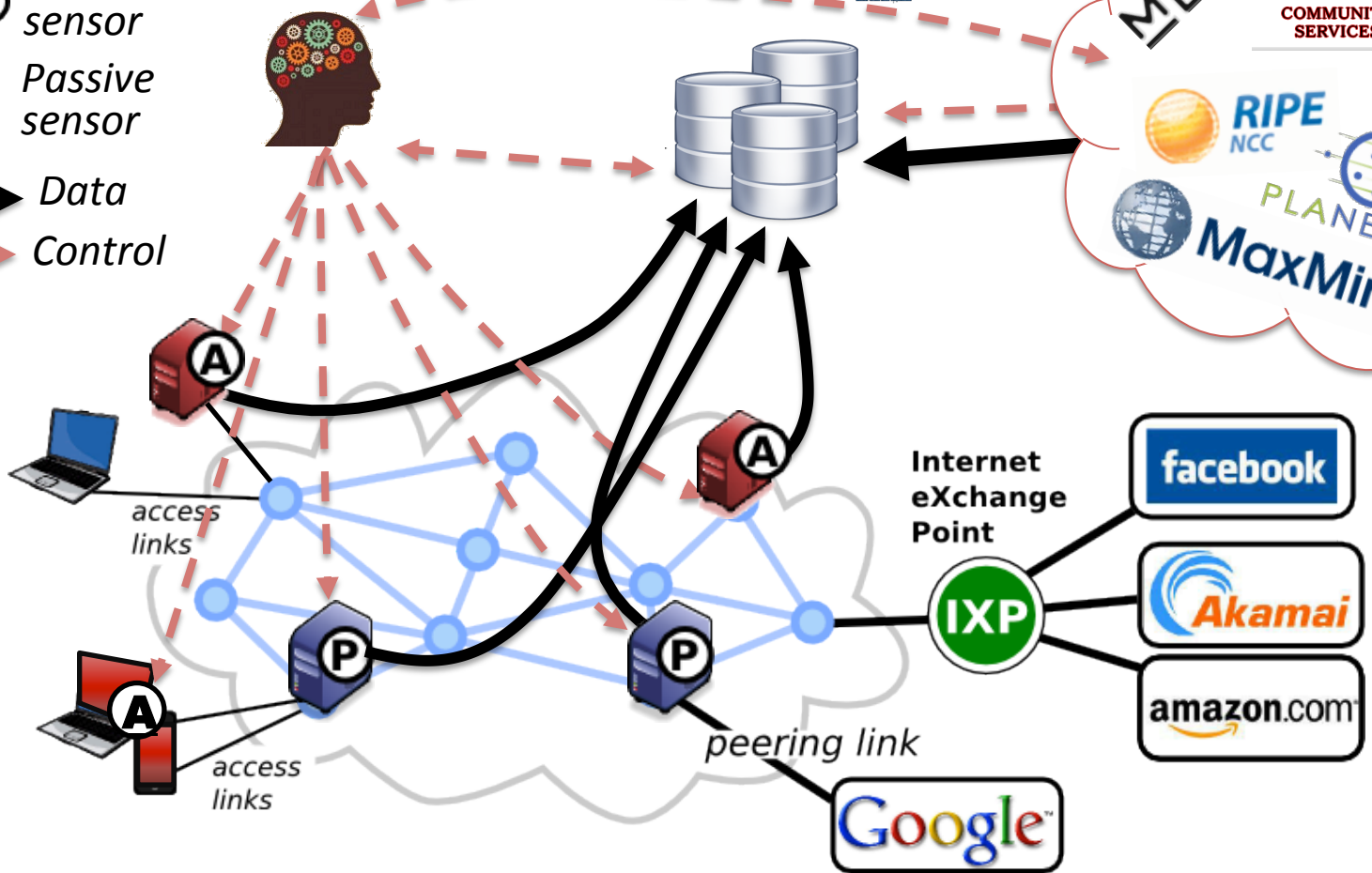
Integration with existing monitoring frameworks





Active and passive analysis for iterative root-cause-analysis



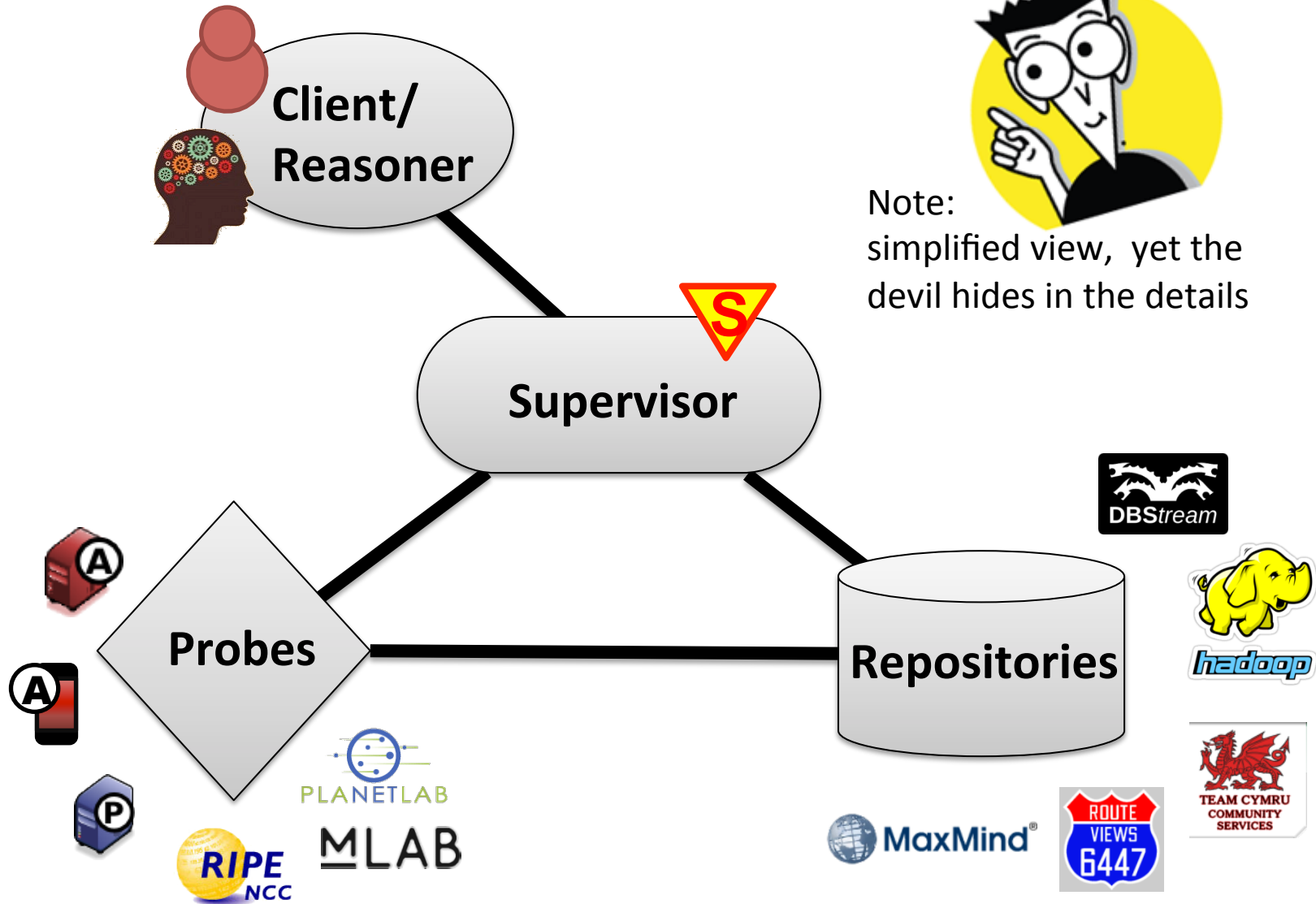
- Active sensor
- Passive sensor
- Data
- Control



(the Big Data moment)

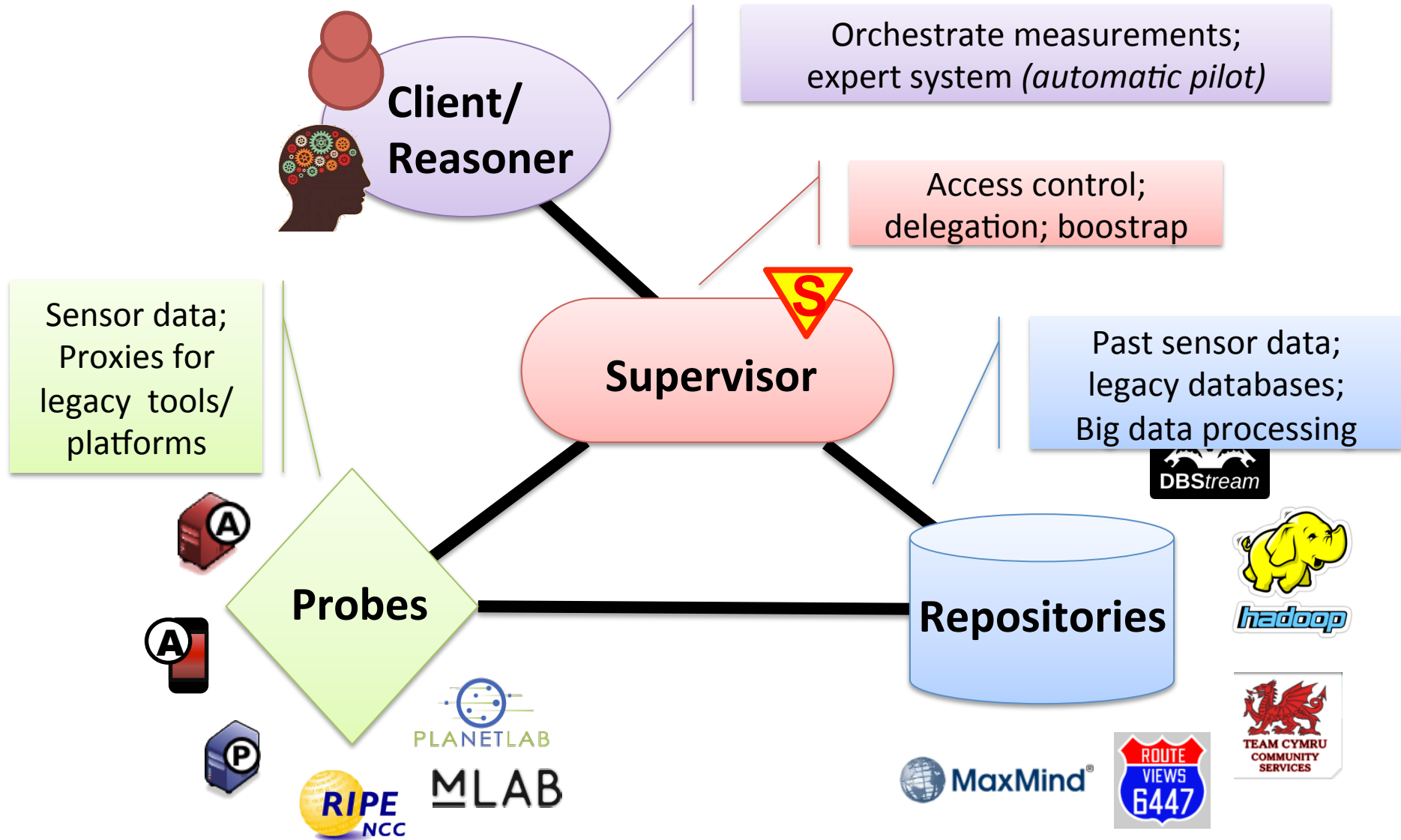
- Traffic  measurements orders of  magnitude
 - 40Gbps (157PB/yr) full-packet processing at each passive sensor, with up to $O(10^7)$ traffic classifications per second...
 - $O(10^9)$ active probes in an Internet anycast census... (more on that later if time allows)
- To be compared with
 - The Large Hadron Collider (LHC), generates ~ 25 PB/yr and $O(10^8)$ collisions per second
 - Sloan Digital Sky Survey (SDSS), generates ~ 73 TB/yr
 - Capacity of the Human genome with 2-bits bases $\sim O(10^9)$

Plane architecture walkthrough

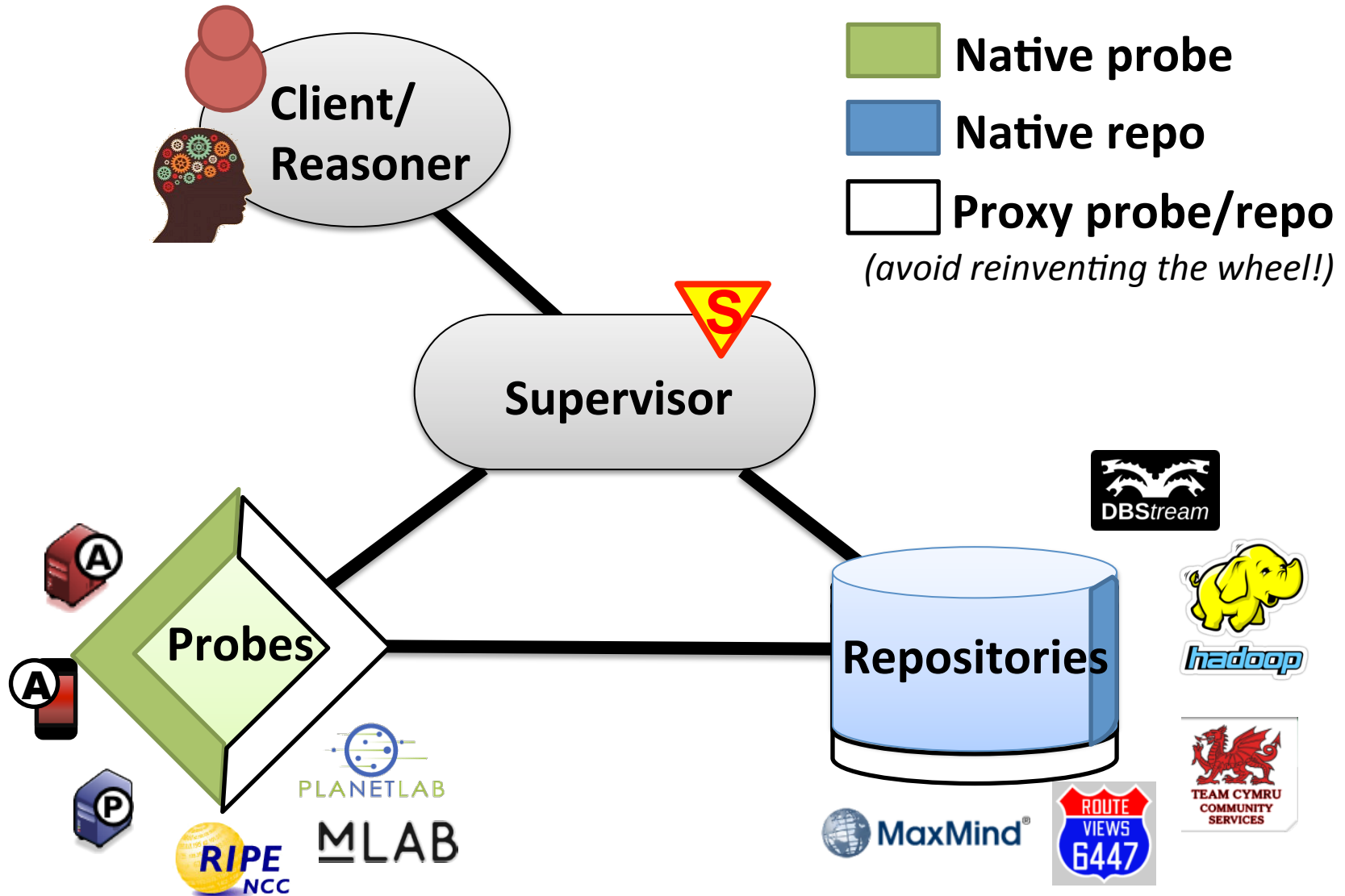


Note:
simplified view, yet the
devil hides in the details

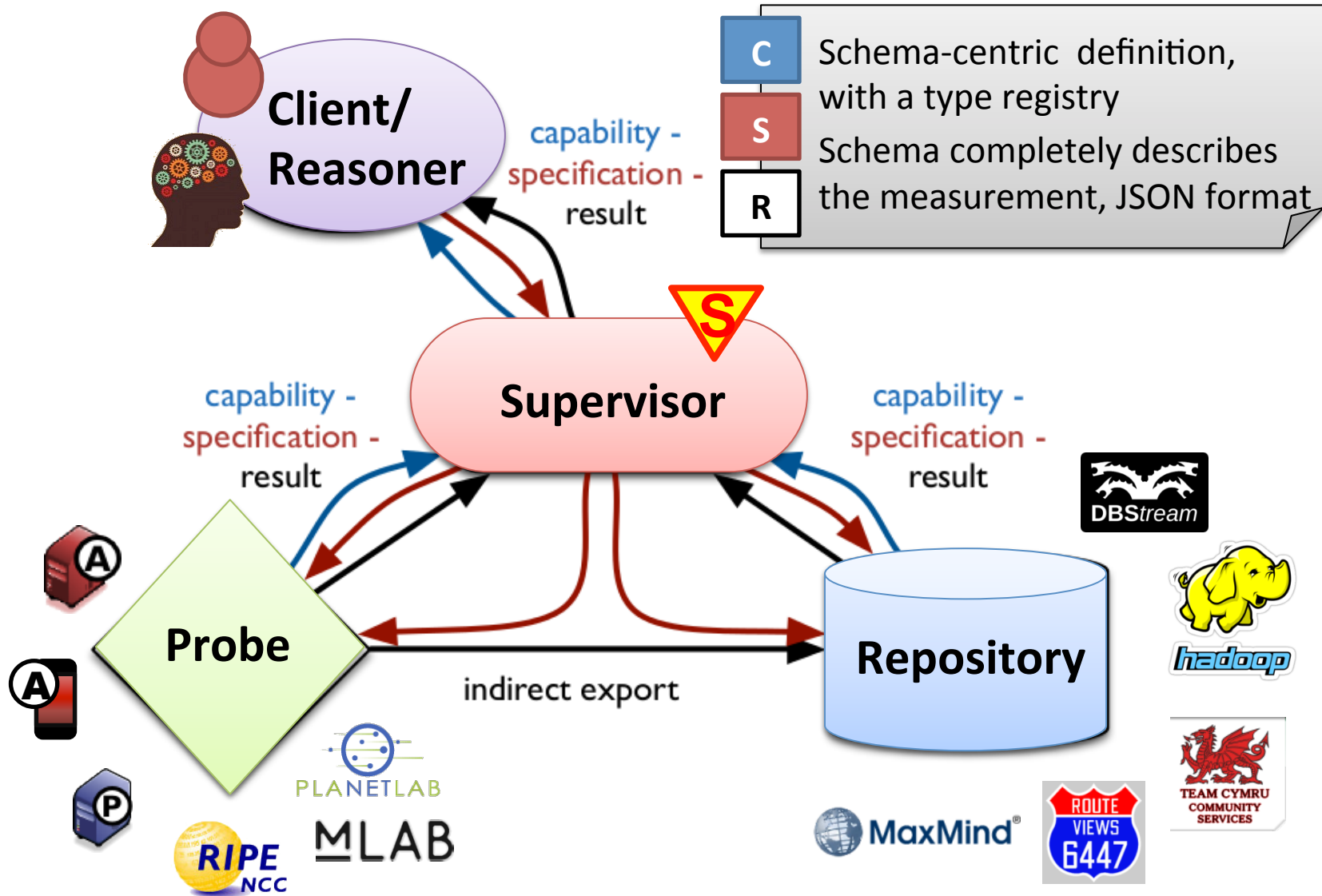
Plane architecture: entities



Plane architecture: interfaces



Plane architecture: messages



Plane architecture: messages

C

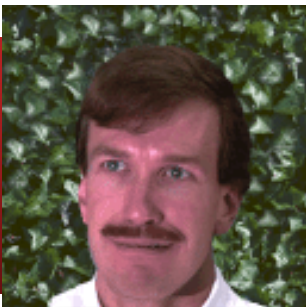
```
{
  "capability": "measure",
  "version": 0,
  "registry": http://ict-mplane.eu/reg,
  "label": "ping-aggregate"
  "when": "now .. future / 1s"
  "parameters": {"source.ip4": "137.3.1.1",
                 "destination.ip4": "*"},
  "results": ["delay.twoway.icmp.us.min",
             "delay.twoway.icmp.us.mean",
             "delay.twoway.icmp.us.max",
             "delay.twoway.icmp.count"]
}
```

S

```
{
  "specification": "measure",
  "version": 0,
  "registry": http://ict-mplane.eu/reg,
  "label": "ping-aggregate-experiment1",
  "when": "now + 30s / 1s",
  "token" : "0f31c9033f8fce0c9be41d4944"
  "parameters": {"source.ip4": "137.3.1.1",
                 "destination.ip4": "137.194.164.1"},
  "results": ["delay.twoway.icmp.us.min",
             "delay.twoway.icmp.us.mean",
             "delay.twoway.icmp.us.max",
             "delay.twoway.icmp.count"]
}
```

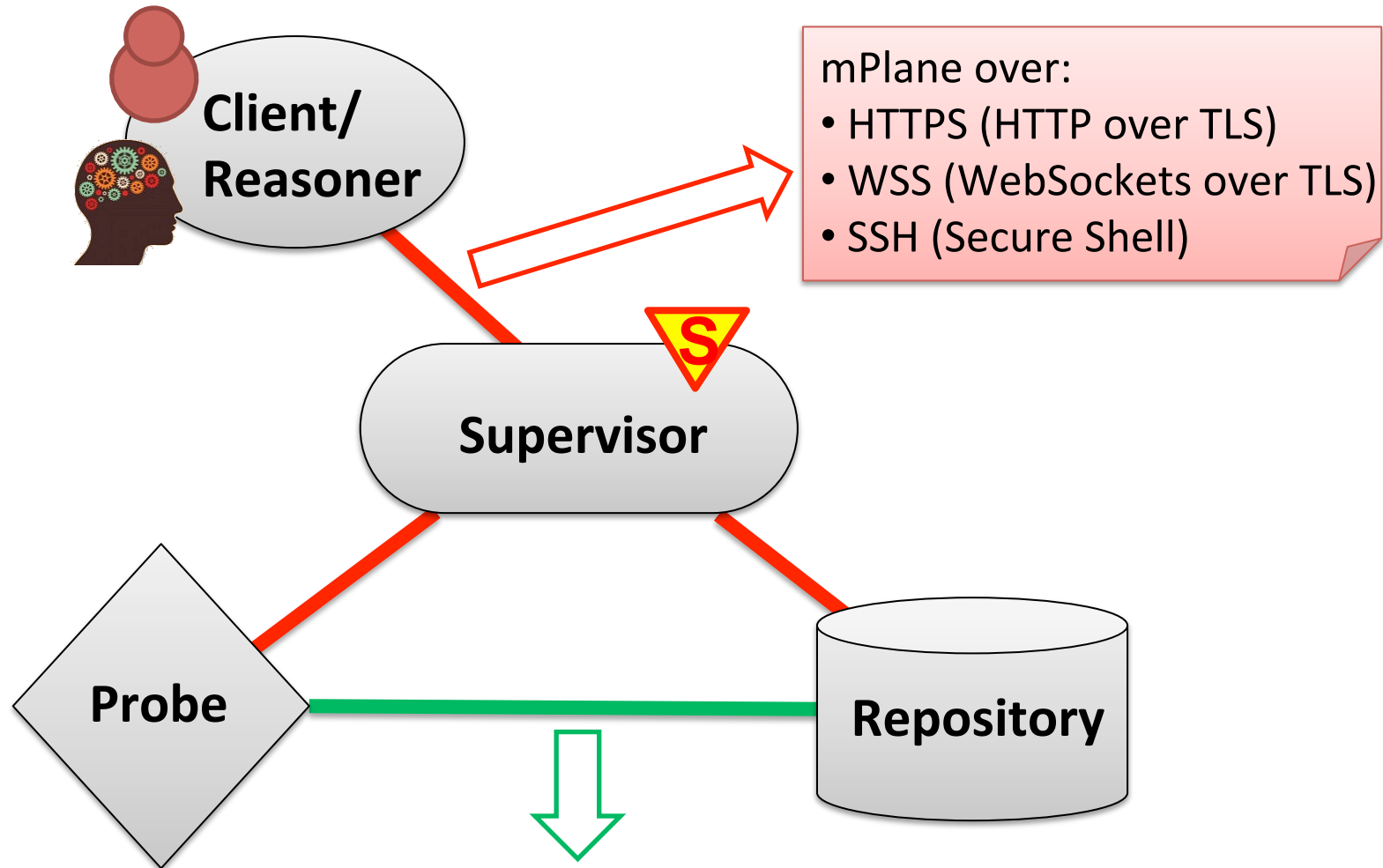
R

```
{
  "specification": "measure",
  "version": 0,
  "registry": http://ict-mplane.eu/reg,
  "label": "ping-aggregate-experiment1",
  "when": "2015-09-07 14:30:02.123 [...] /1s",
  "token" : "0f31c9033f8fce0c9be41d4944",
  "parameters": {"source.ip4": "137.3.1.1",
                 "destination.ip4": "137.194.164.1"},
  "results": ["delay.twoway.icmp.us.min",
             "delay.twoway.icmp.us.mean",
             "delay.twoway.icmp.us.max",
             "delay.twoway.icmp.count"]
  "resultsvalue" : [[ 2390,2983, 6600,30]]
}
```



**grin* If I'd known then that [ping] would be my most famous accomplishment in life, I might have worked on it another day or two and added some more options.*

Plane architecture: transport

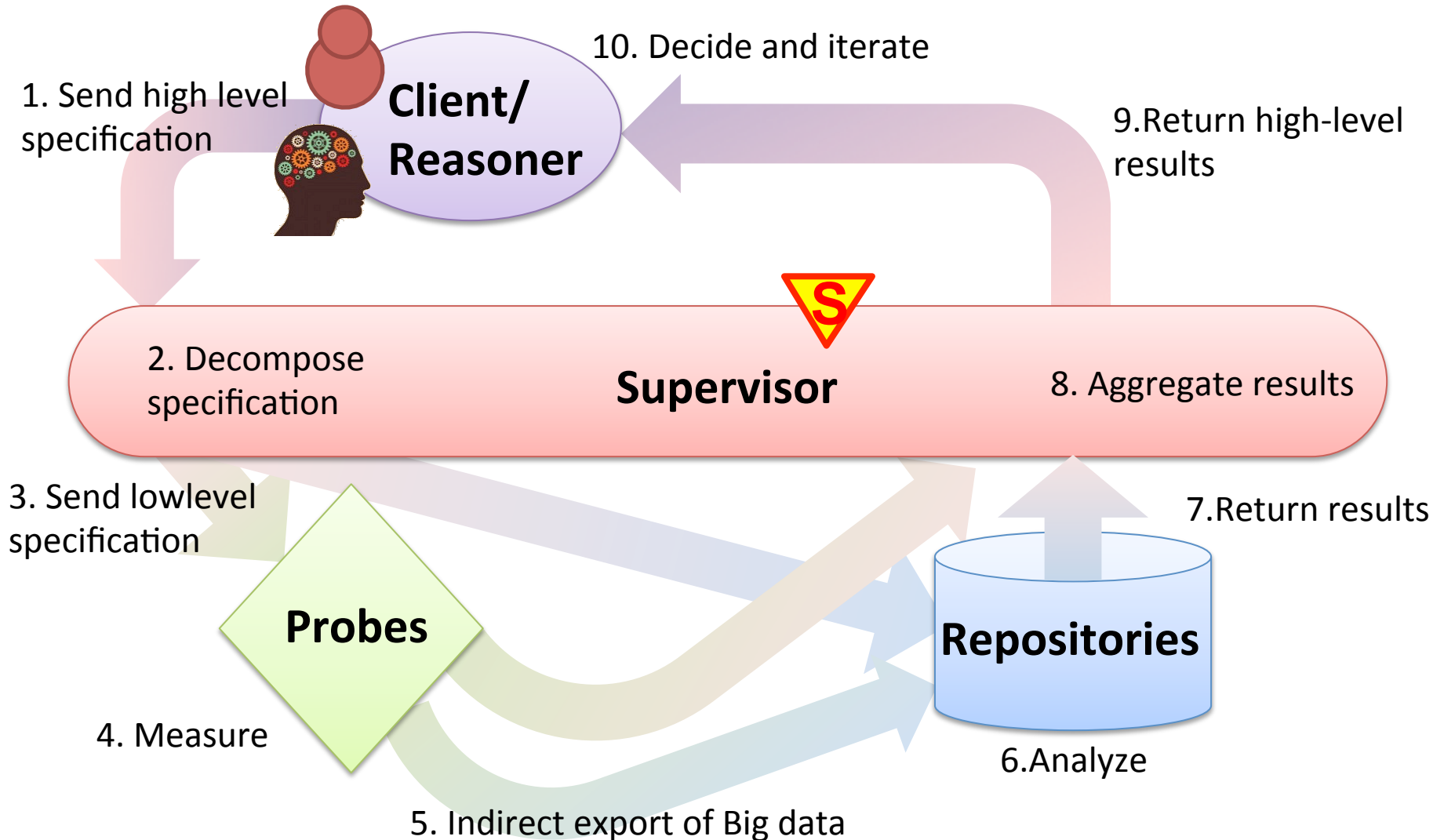


mPlane over:

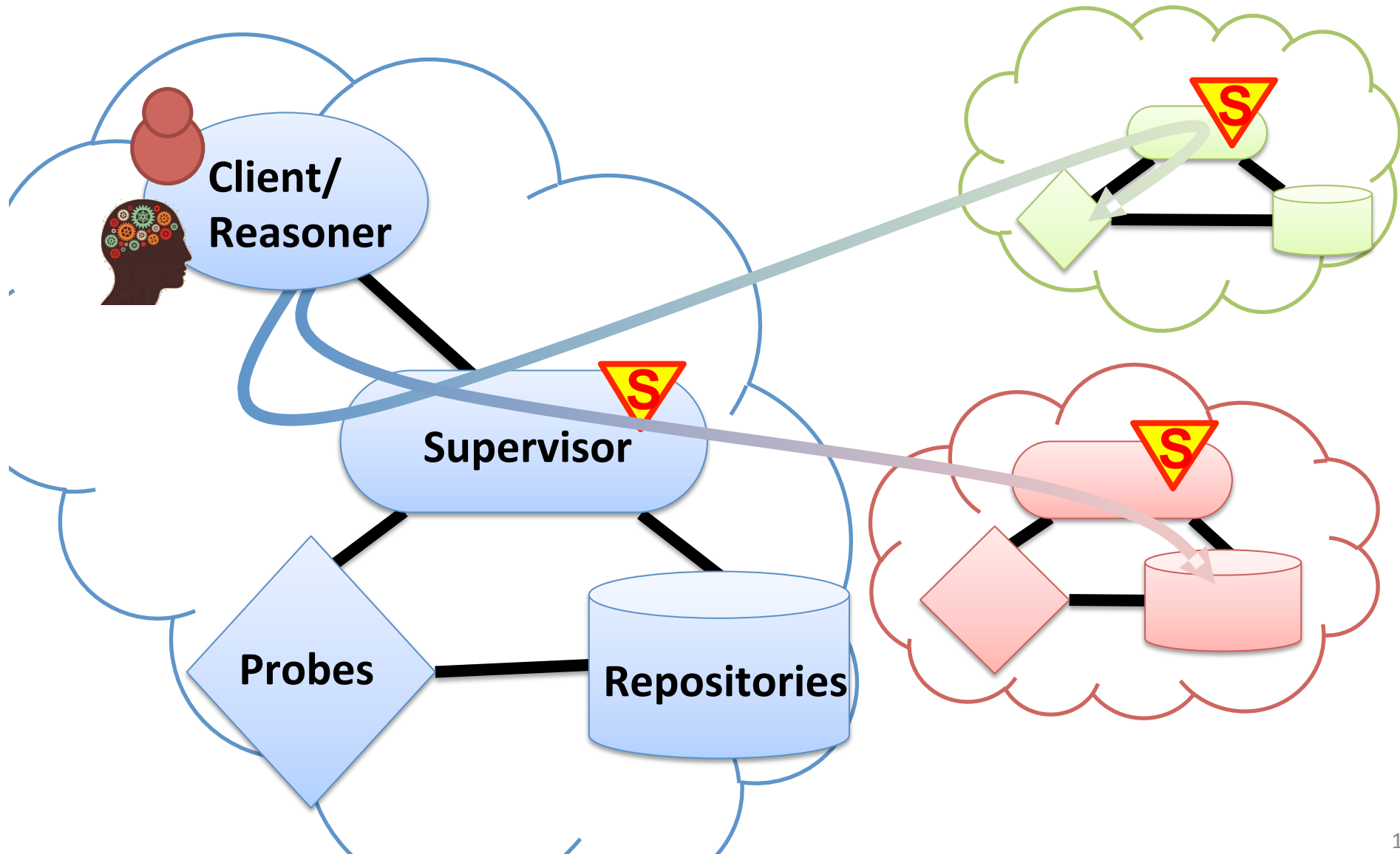
- HTTPS (HTTP over TLS)
- WSS (WebSockets over TLS)
- SSH (Secure Shell)

Indirect export uses custom protocols:
IPFIX, FTP, BitTorrent, Apache Flume, RFC1149, RFC6214, etc.



Plane architecture: workflow



Plane architecture: Inter-domain

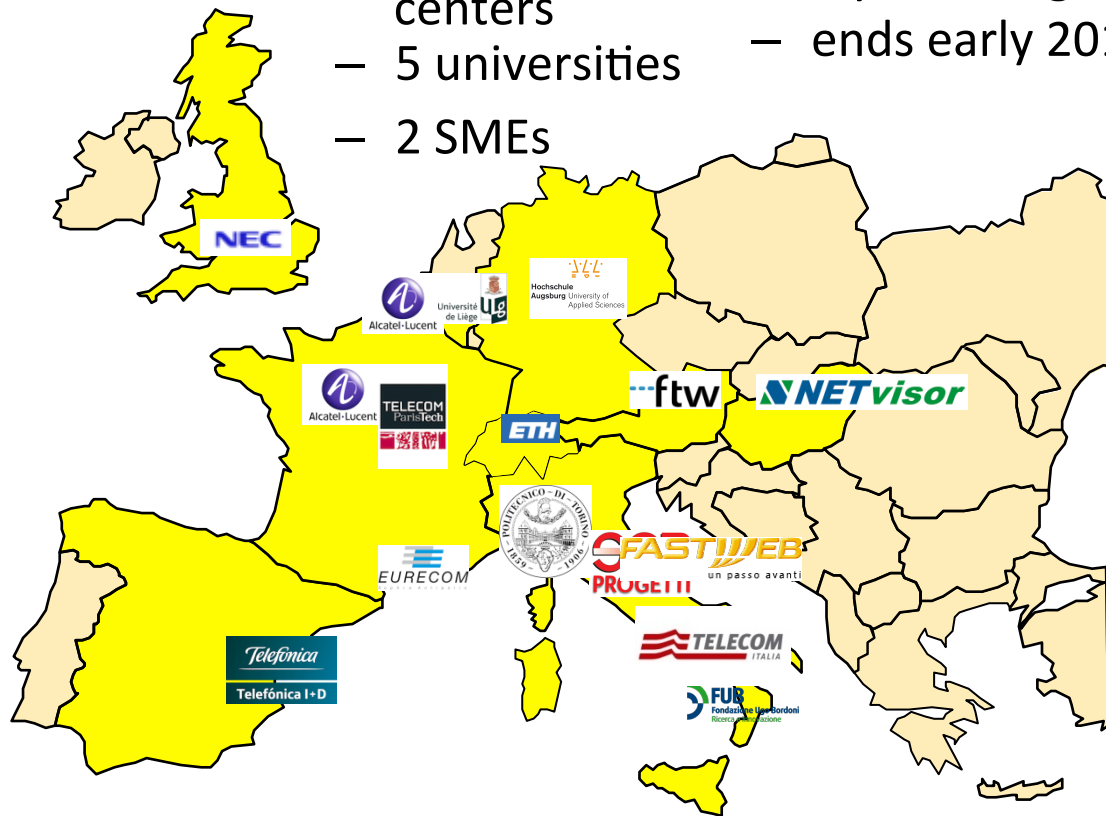



The broader measurement ecosystem

-  *“From global measurements to local management”*
 - 10 partners, 3.8 MEUR, FP7 STREP
 - More focused use case: Build a framework out of  probes
 - Knowledge sharing (e.g., joint work, Dagstuhl seminars, etc.)
- IETF Large-Scale Measurement of Broadband Performance (LMAP)
 - Defines the components, protocols, rules, etc., but does not specifically target adding “a brain” to the system
 - Common core set, Largely interoperable
- IETF IP Performance Metrics (IPPM)
 - Registry related, we use its vocabulary as much as possible
- IETF IP Flow Information Export (IPFIX)
 - Indirect export related, active contributors
- Others in scope
 - IETF DOCTORS; tcpm; ConEx; NETCONF; IRTF NMRG; ETSI STQ; ITU SG12

Plane consortium

- 16 partners
- FP7 IP
- 3 ISPs
- 6 research centers
- 5 universities
- 2 SMEs
- 11 MEUR
- 3 years long
- ends early 2016



Marco Mellia
POLITO 



Saverio Nicolini
NEC



Dina Papagiannaki
Telefonica



Ernst Biersack
Eurecom



Brian Trammell
ETH



Tivadar Szemethy
NetVisor



Andrea Fregosi
Fastweb



Dario Rossi
ENST



Fabrizio Invernizzi
Telecom Italia



Guy Leduc
Univ. Liege



Pietro Michiardi
Eurecom



Pedro Casas
FTW

mPlane achievements so far



Note:
hard to summarize!

- Standardization & dissemination
 - 3 RFCs, 5 drafts
 - 80+ papers, including 4 Best paper awards and ACM SIGCOMM IMC & CoNEXT, IEEE INFOCOM
- Software
 - <https://www.ict-mplane.eu/public/software> and <https://github.com/fp7mplane/>
(Ready-to-use Virtual Machines under preparation)
- Check the demos on **You Tube**
 - <https://www.youtube.com/channel/UCHGS6UIUKvGZTyt5DemmPaw>

“Biased sampling”:

Anycast geolocation as a showcase of mPlane achievements

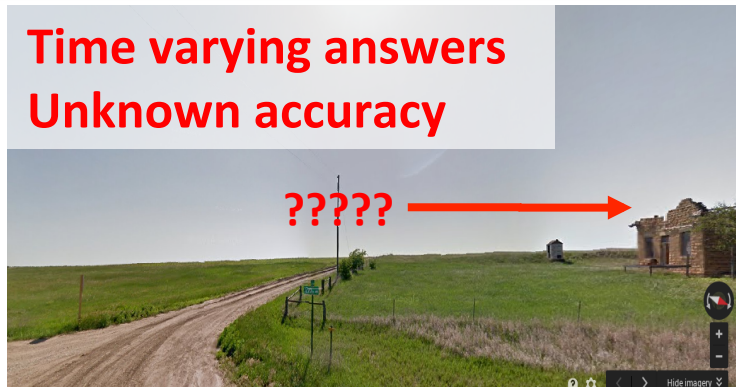
IP Anycast

- Set of equivalent replicated servers sharing the same IP address, user routed to closest replica (in BGP sense)
- Question: where are Google DNS servers 8.8.8.8?

Commercial databases

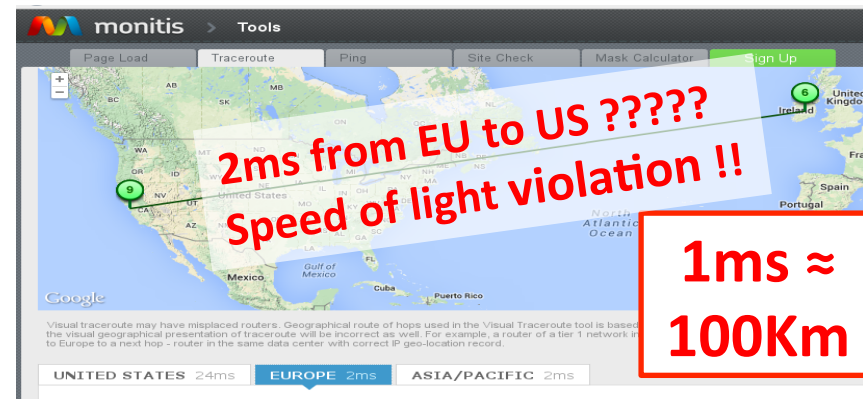
Mountain View, CA (IP2Location)
New York, NY (Geobytes)
United States (Maxmind)

Time varying answers
Unknown accuracy

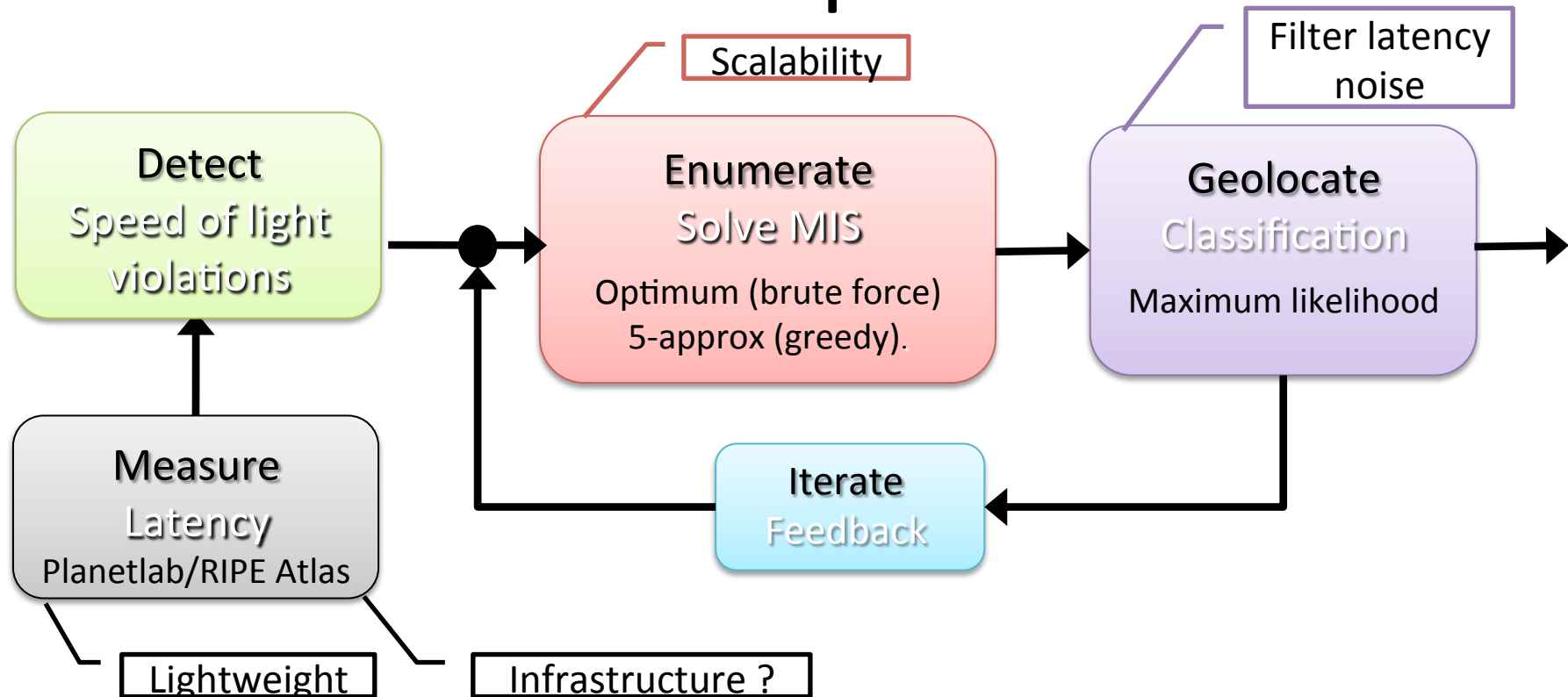


Distributed measurement

Tools using distributed measurement aren't better !



Geolocation technique



measurement infrastructure

- RIPE Atlas vs PlanetLab

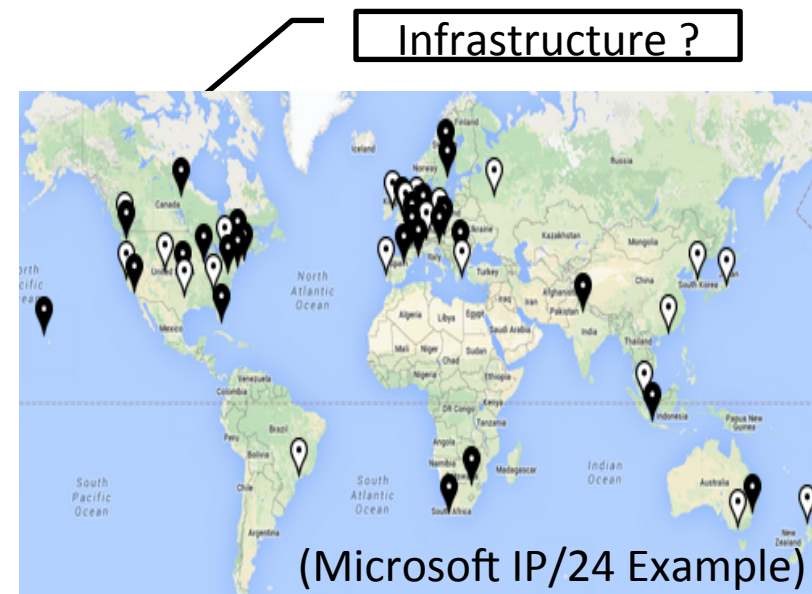
Footprint	VPs	ASes	Countries
RIPE Atlas	7k	2k	150
PlanetLab	~300	180	30

- Anecdotal understanding

- For some deployments, RIPE includes PlanetLab

- For others, the union is greater than the sum of the parts !

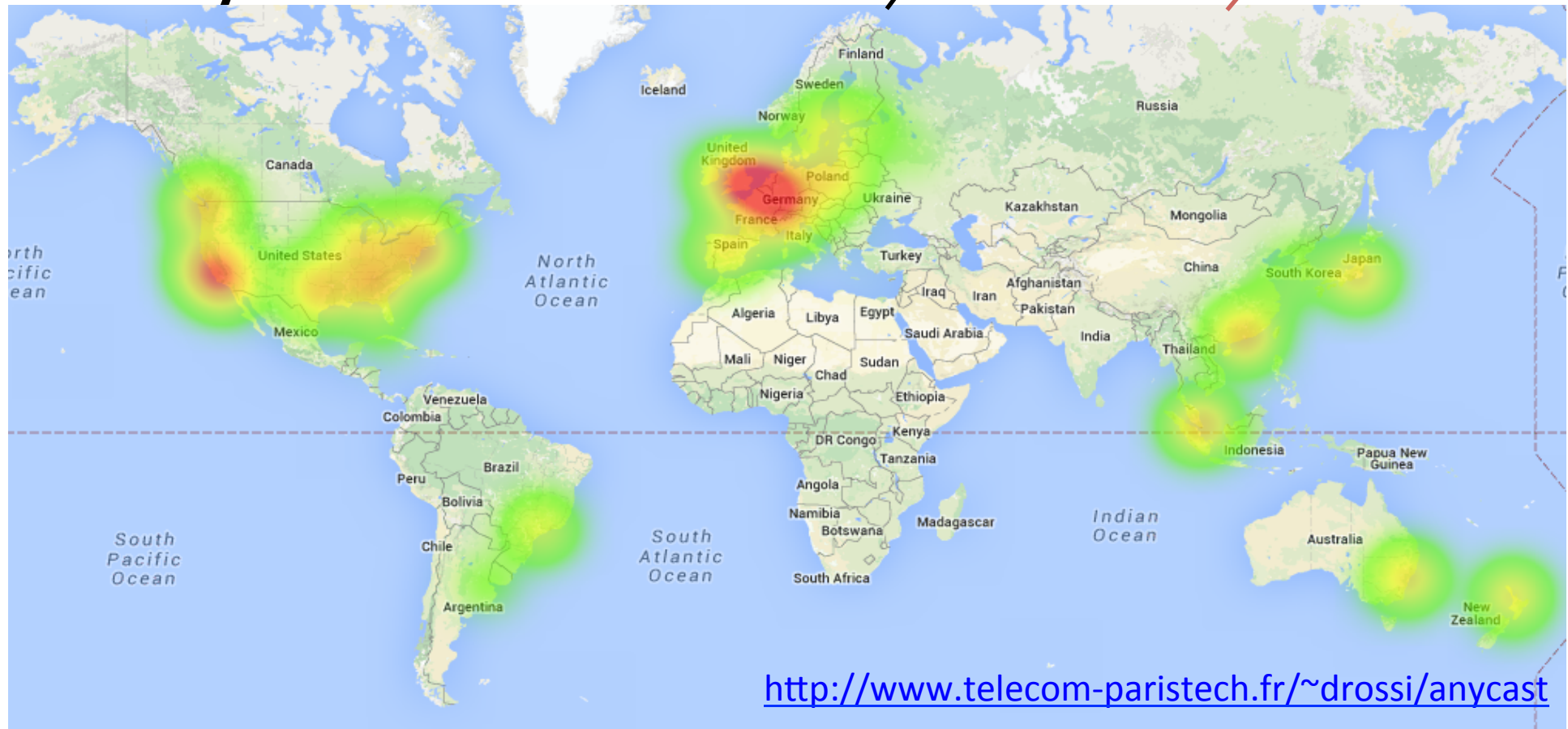
-  Plane enables/simplifies systematic studies



Anycast census

Lightweight

Scalable



- $O(10^7)$ targets x $O(10^2)$ active sensors
- $O(10^3)$ targets /sensor /second
- $O(10^3)$ targets are anycast – *needle in the IPv4 haystack*



Su ary

- Overview
 - Measurements to shed light on Internet operational obscurity
- Insights
 - Measurement plane to facilitate expression of measurement capabilities and needs
 - Allow users to concentrate effort on hard problems
 - Avoiding time consuming tasks
- Hintsights
 - Crucial to foster adoption (GitHub, ready-to-use VMs) to reach a critical mass (incentive in proxying)

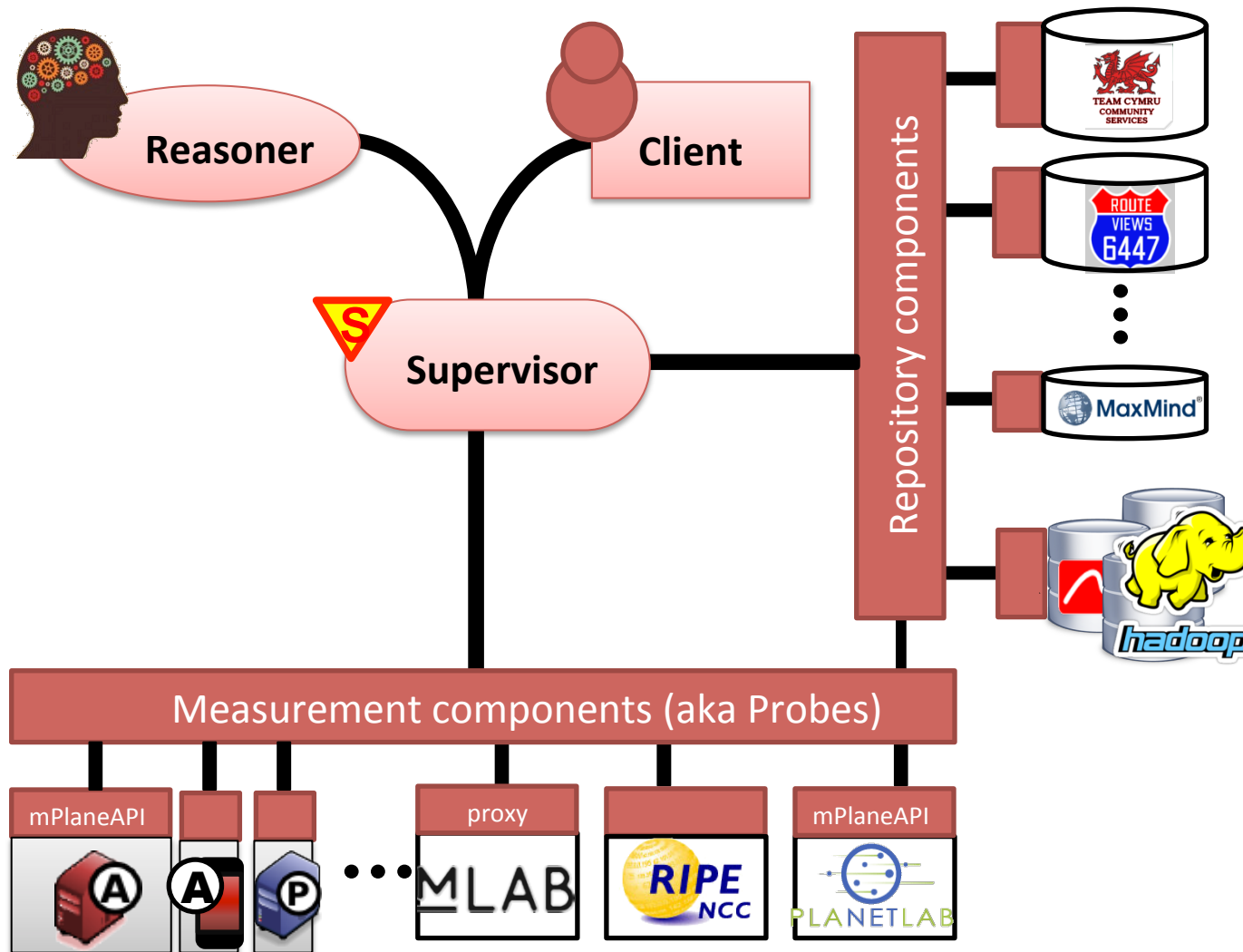


 any thanks

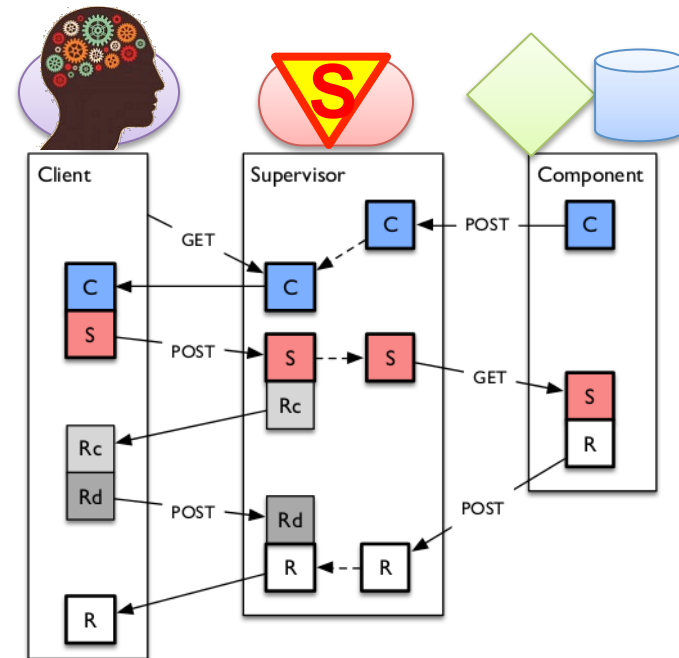
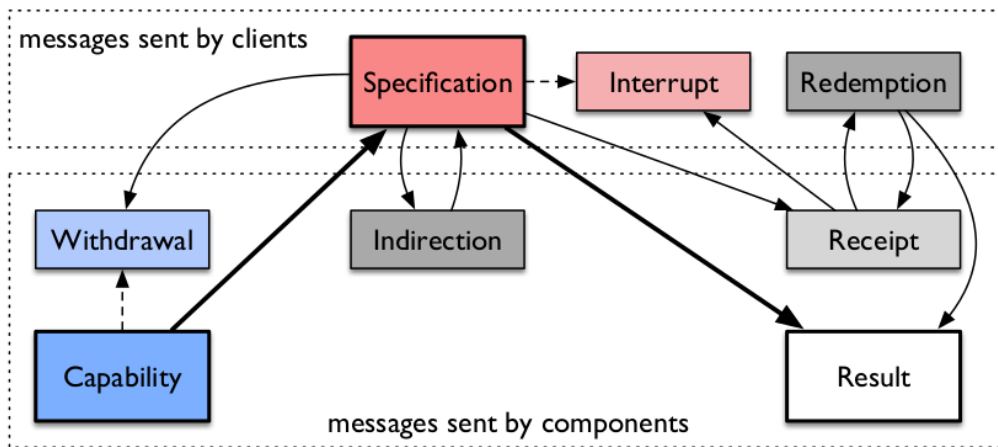
?? || //

Backup slides

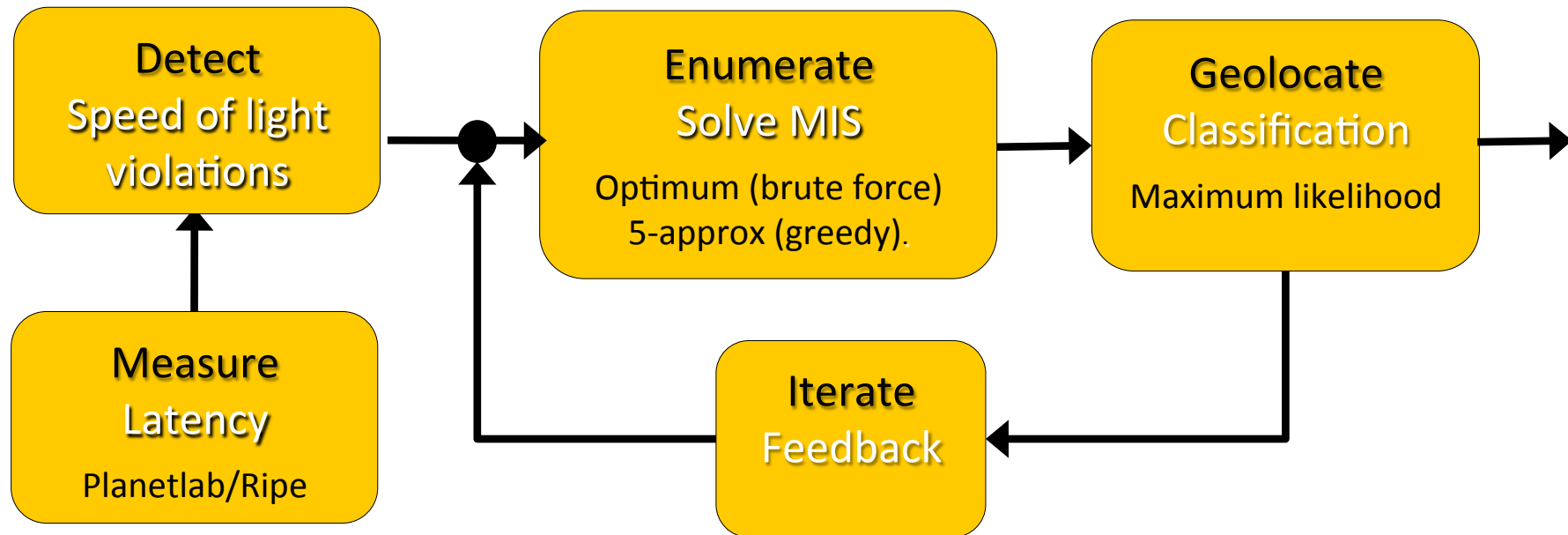
Plane architecture: simplistic view



Plane architecture: gory details

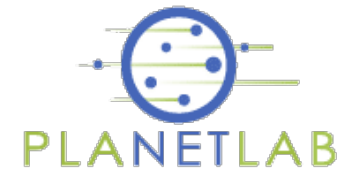
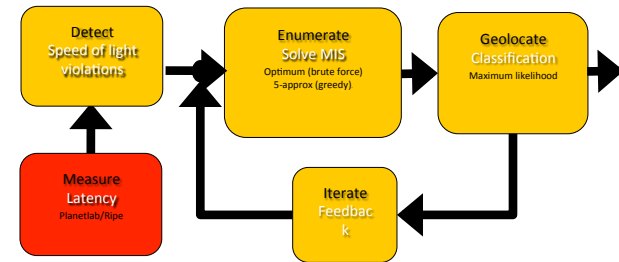


Methodology overview



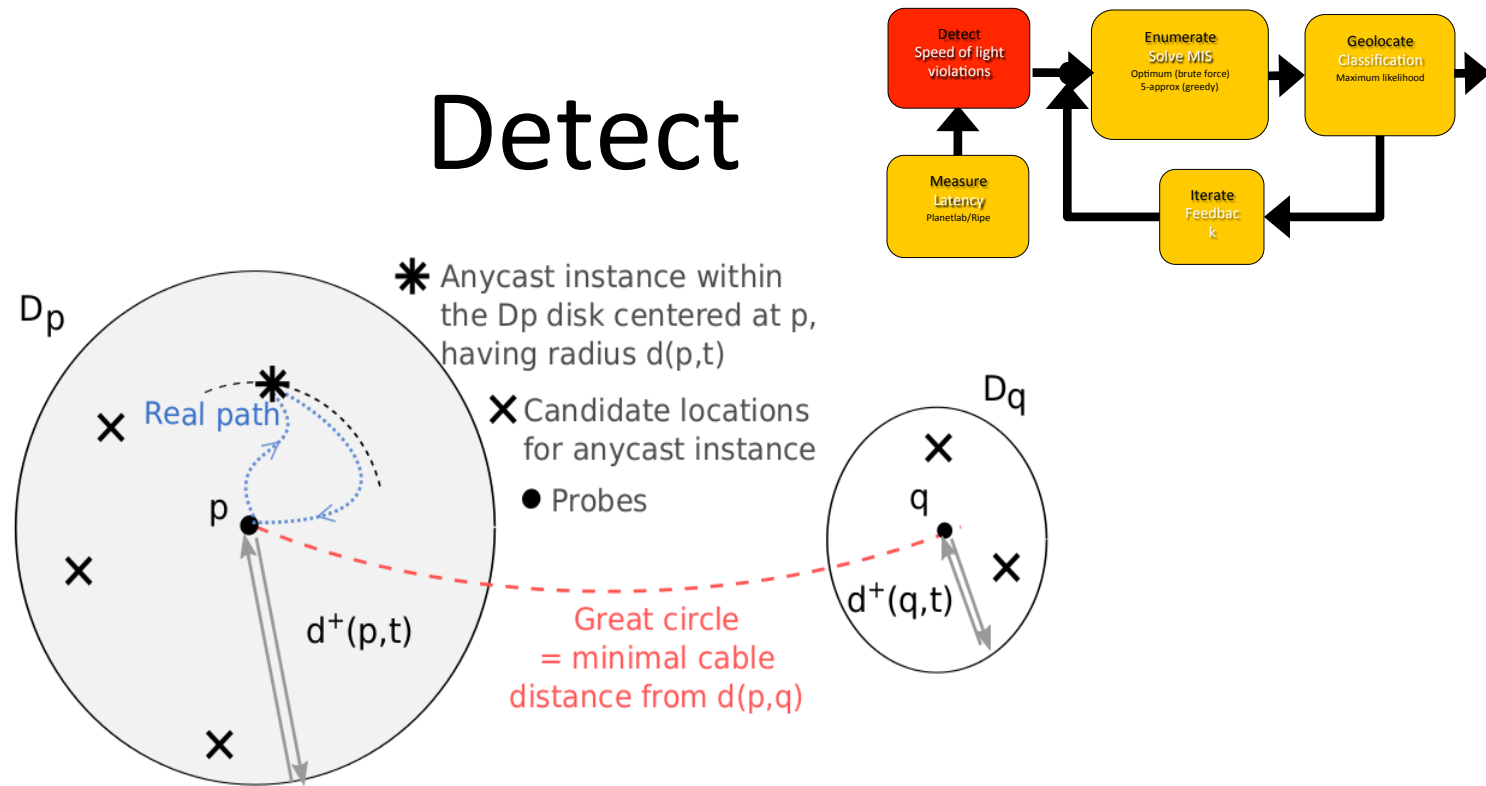
Measure

- PlanetLab
 - 300 vantage points
 - Geolocated with Spotter (ok for unicast)
 - Freedom in type/rate of measurement
ICMP, DNS, TCP-3way delay, etc
- RIPE
 - 6000 vantage points
 - Geolocated with MaxMind (ok for unicast)
 - More constrained (ICMP, traceroute)



In this talk: min over 10 ICMP samples

Detect

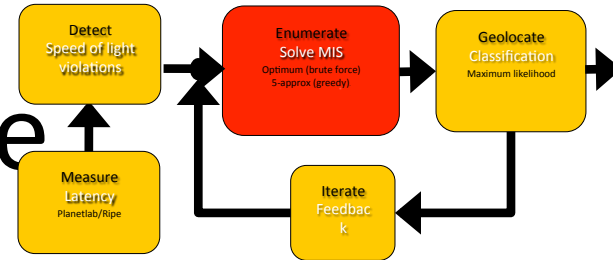


The vantage points p and q are referring to two different instances if:

$$d(p; q) > d(p; t) + d(q; t)$$

Packets cannot travel faster than the speed of light

Enumerate



- Find a **maximum independent set** \mathcal{E}
 - of discs such that:
 - Brute force (optimum) vs Greedily from smallest (5-approximation)

$$\forall \mathcal{D}_p, \mathcal{D}_q \in \mathcal{E}, \quad \mathcal{D}_p \cap \mathcal{D}_q = \emptyset$$



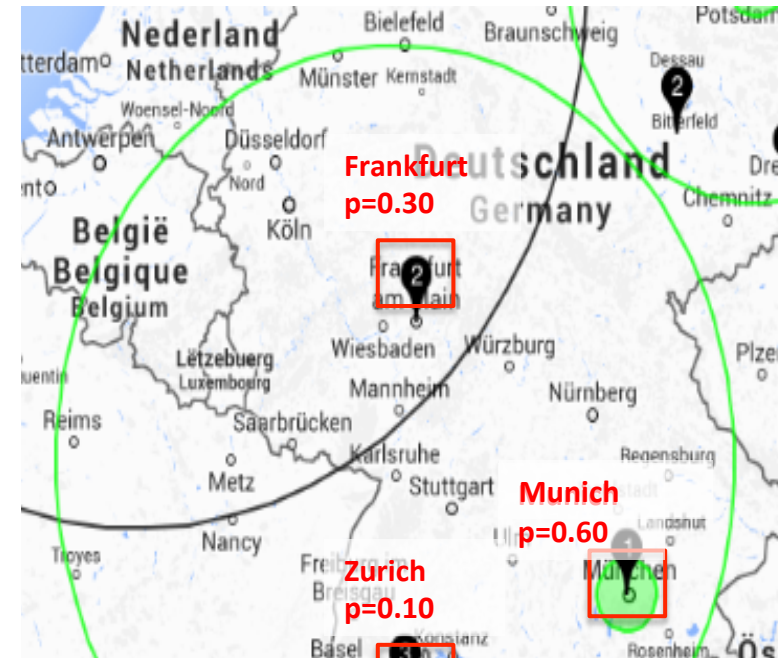
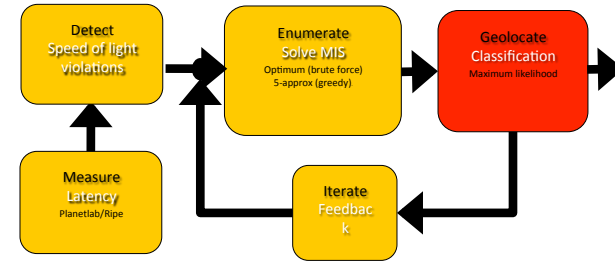
Geolocate

- Classification task
 - Map each disk D_p to **most likely** city
 - Compute likelihood (p) of each city in disk based on:

- c_i : Population of city i
- A_i : Location of ATA airport of city i
- $d(x,y)$: Geodesic distance
- α : city vs distance weighting

$$p_i = \alpha \frac{c_i}{\sum_j c_j} + (1 - \alpha) \frac{d(p, t) - d(p, A_i)}{\sum_j d(p, t) - d(p, A_j)}$$

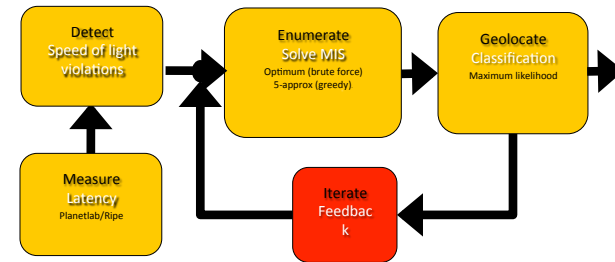
- Output policy
 - **Proportional**: Return all cities in D_p with respective likelihoods
 - **Argmax**: Pick city with highest likelihood



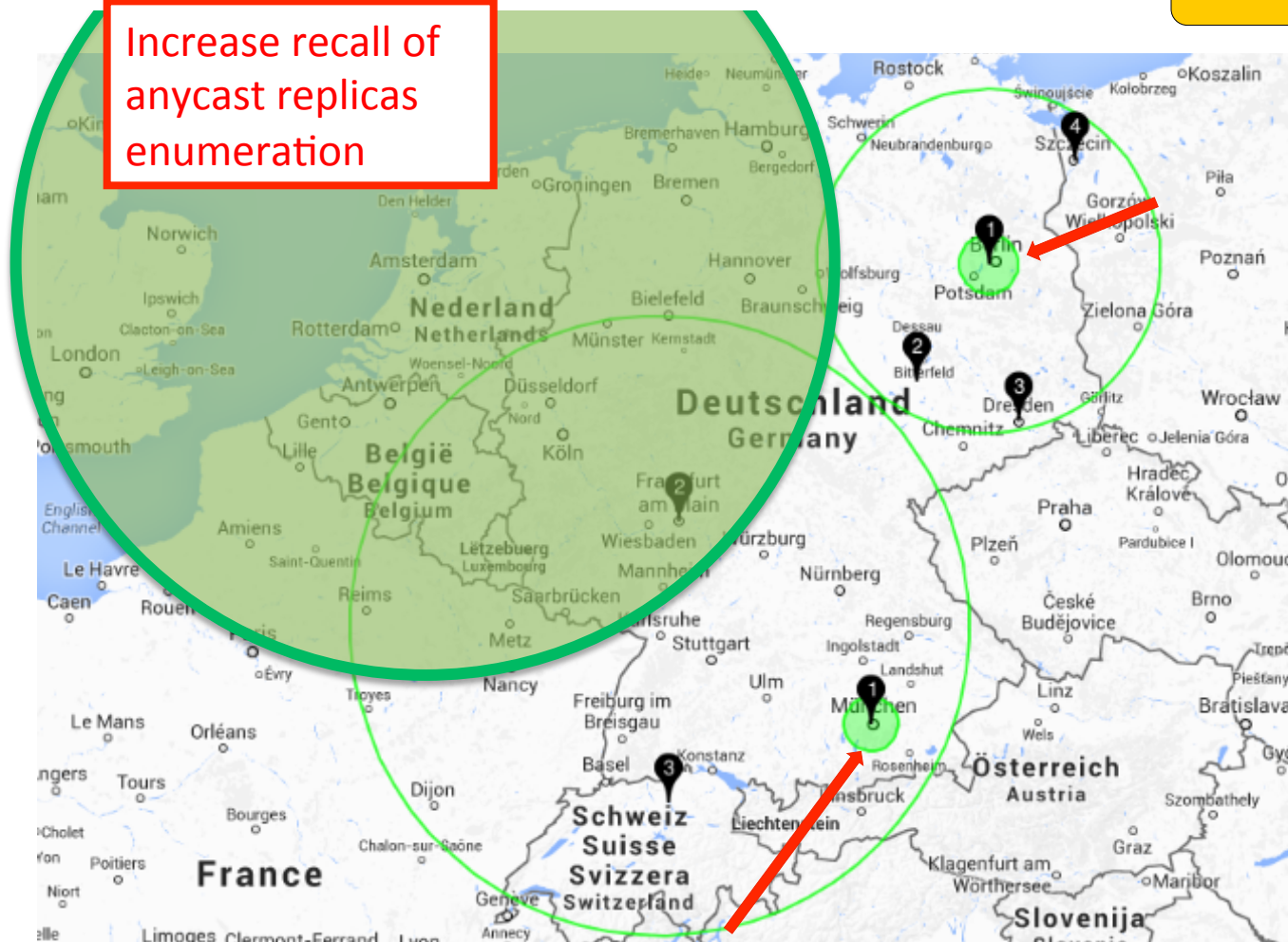
rationale: users lives in densely populated area; to serve users, servers are placed close to cities

airports: simplifies validation against ground truth (DNS)

Iterate



Increase recall of anycast replicas enumeration



- Collapse
 - Geolocated disks to city area
- Rerun
 - Enumeration on modified input set

Performance at a glance

– Protocol agnostic and lightweight

- Based on a handful of delay measurement $O(100)$ VPs
- 1000x fewer VPs than state of the art

– Enumeration

- iGreedy use 75% of the probes (25% discarded due to overlap)
- Overall 50% recall (depends on VPs; stratification is promising)

– Geolocation

- Correct geolocation for 78% of enumerated replicas
- 361 km mean geolocation error for all enumerated replicas (271 km median for erroneous classification)