Find and Study User Actions from HTTP Traces

User Actions = webpage URLs explicitly requested by a user

1 - Find User Actions from HTTP traces
2 - Analyze User Actions
3 - Model user navigation behaviour

➢ Show how users surf the Web
➢ Build model of real users to:
  ➢ Forecast their behaviors
  ➢ Enhance webpage ranking algorithms

1 - Find User Actions

Users in households   Network traffic passively observed   Webpage URLs

Tstat
HTTP trace

PoP

Domain A
Page A
Page B

Domain B
Page A
Page B
Page C

Data from 20k users, 3 years

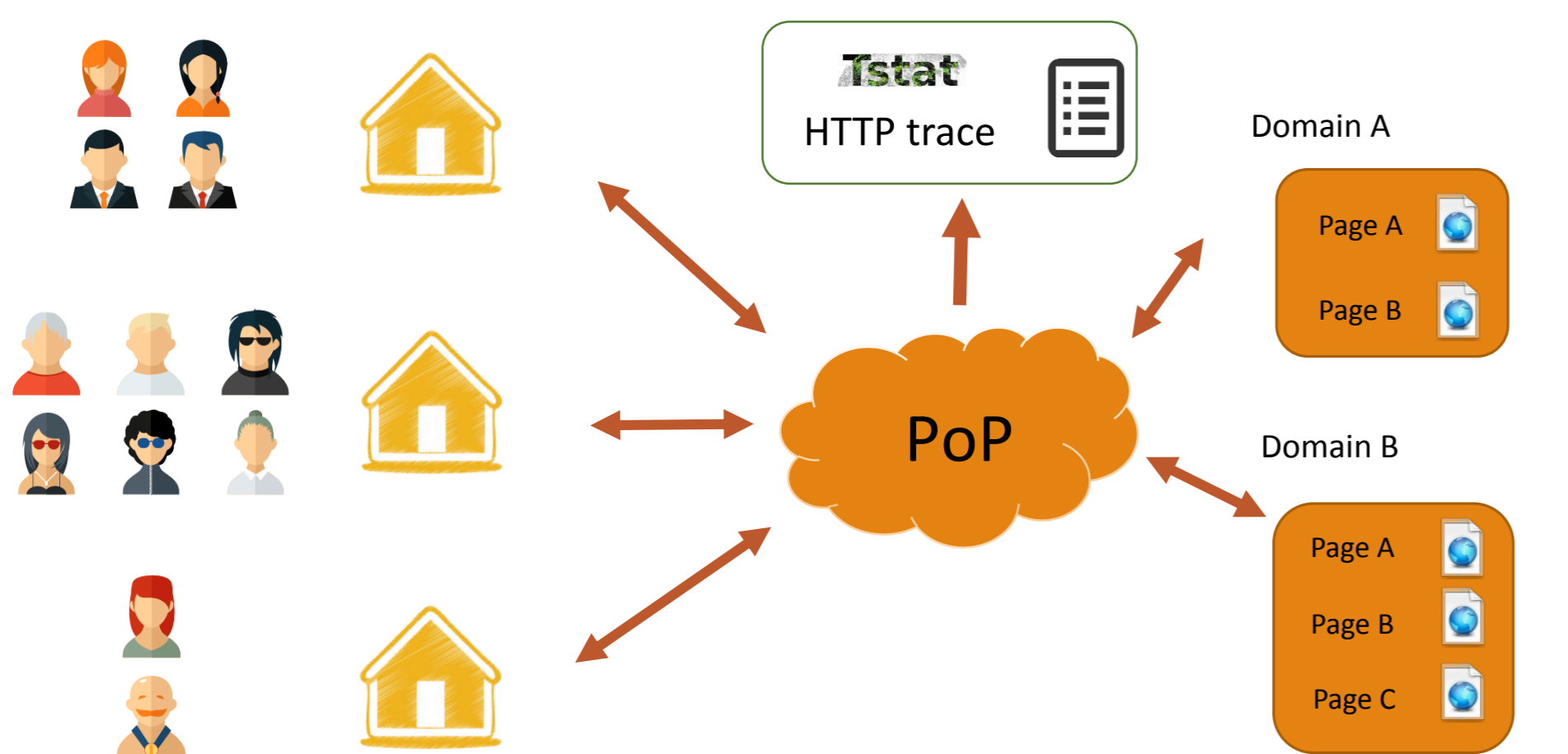Challenge: detect User Actions – just 2% of entries in HTTP traces

Approach: **machine learning**
automatic, more precise, adapt to changes

Datasets for learning: Classifiers:
Volunteers and synthetic   Decision tree, **random forest**, bayesian network, NN

Features selection: refered requests, inter-request time, content type,…
Training, validation, testing with precision 70-90% and recall 90-95%

Output:
Graphs of user surfing history

coolvideo.com
myfavouriteblog.com   facebook.com
anotherblog.com   google.com
news.com
newspaper.com   sportnews.com

2 – Analyze User Actions

➢ Quantify visited webpages and domains
  ➢ Paths followed by the users
  ➢ Clusterization of the browsing behaviour
➢ Impact of devices
➢ Identification of prominent content promoters
  ➢ Role of social networks and search engines

Examples:

Time between two User Actions

Webpages weekly visited per device
PC
Tablet
Smartphone

Webpages and domains per path
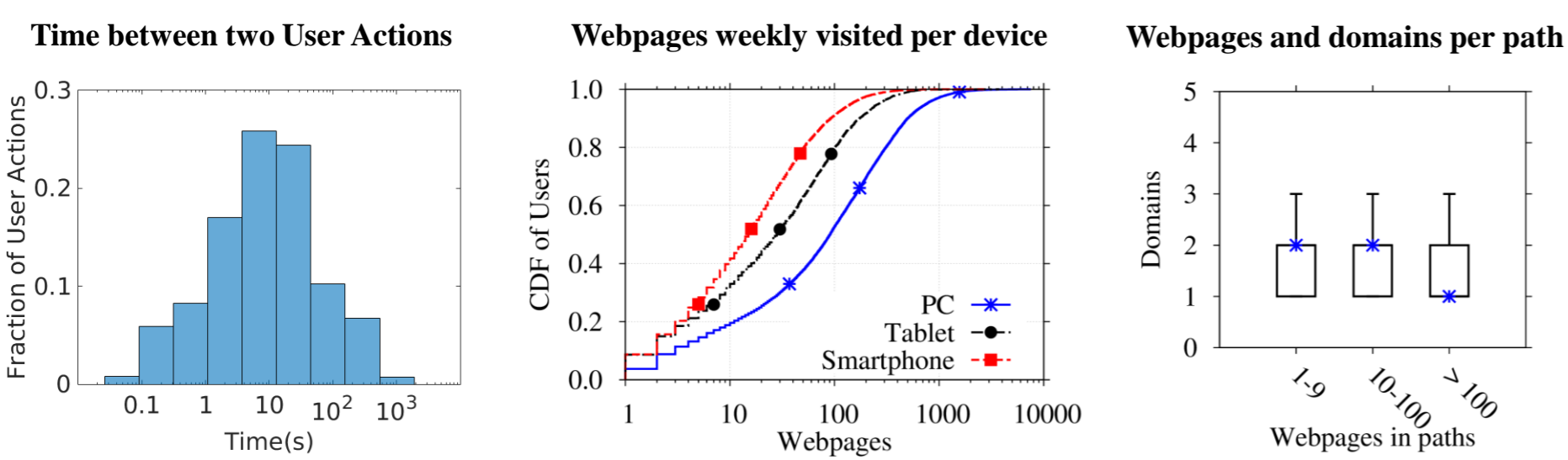
Compare Hot and Cold Web

➢ Hot Web: part of Web visited through User Actions
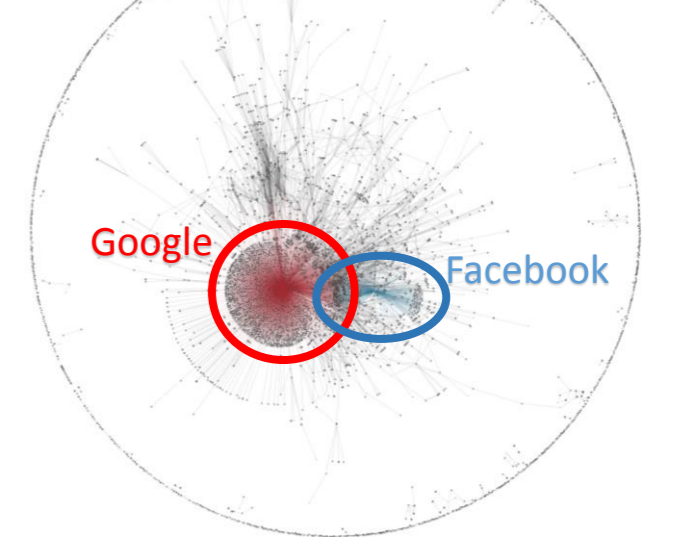➢ Cold Web: active crawling of Web static structure

Built an automatic multi-threaded **crawler** to obtain Cold Web

➢ Many webpages quickly disappear
➢ Most of the hyperlinks are never followed
➢ Hot Web is based on search engines and social networks, with personalized and dynamical pages
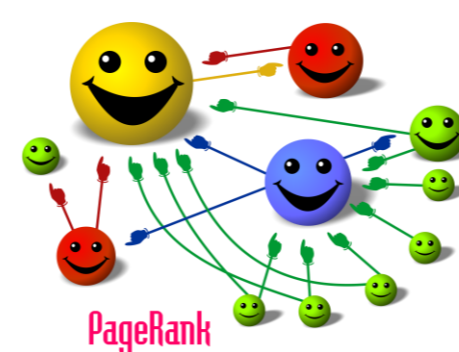
Webpages still existing (LB)
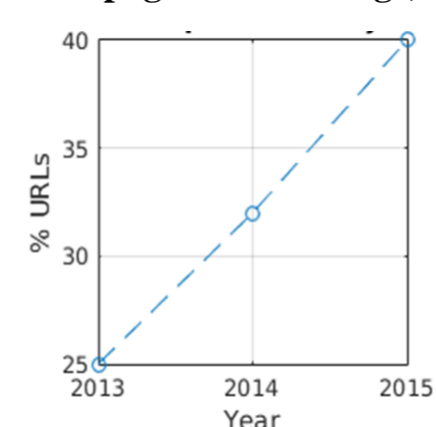
Visualization of Hot Web
Google   Facebook

3 - Model user behavior

Random surfer model ➡ **PageRank**

Likelihood that a user randomly walking will arrive at a certain webpage

Popularity of webpages for random surfers

Comparing with User Actions, random surfer model is not adequate

PageRank

Real surfer model ➡ **Improved PageRank**

Users' walk ≠ random walk
Information of User Actions and Hot Web:

➢ Where to start navigation
➢ How to move among webpages weighting links
➢ How to deal with multiple browsers tabs
➢ Where to stop depending on the webpage and surfer's path

Luca Vassio

luca.vassio@polito.it
lucavassio.wordpress.com

Supervisor: Marco Mellia

POLITECNICO DI TORINO