A Critical Evaluation of Website Fingerprinting Attacks

Marc Juarez[1] Sadia Afroz[2] Gunes Acar[1]
Claudia Diaz[1] Rachel Greenstadt[3]

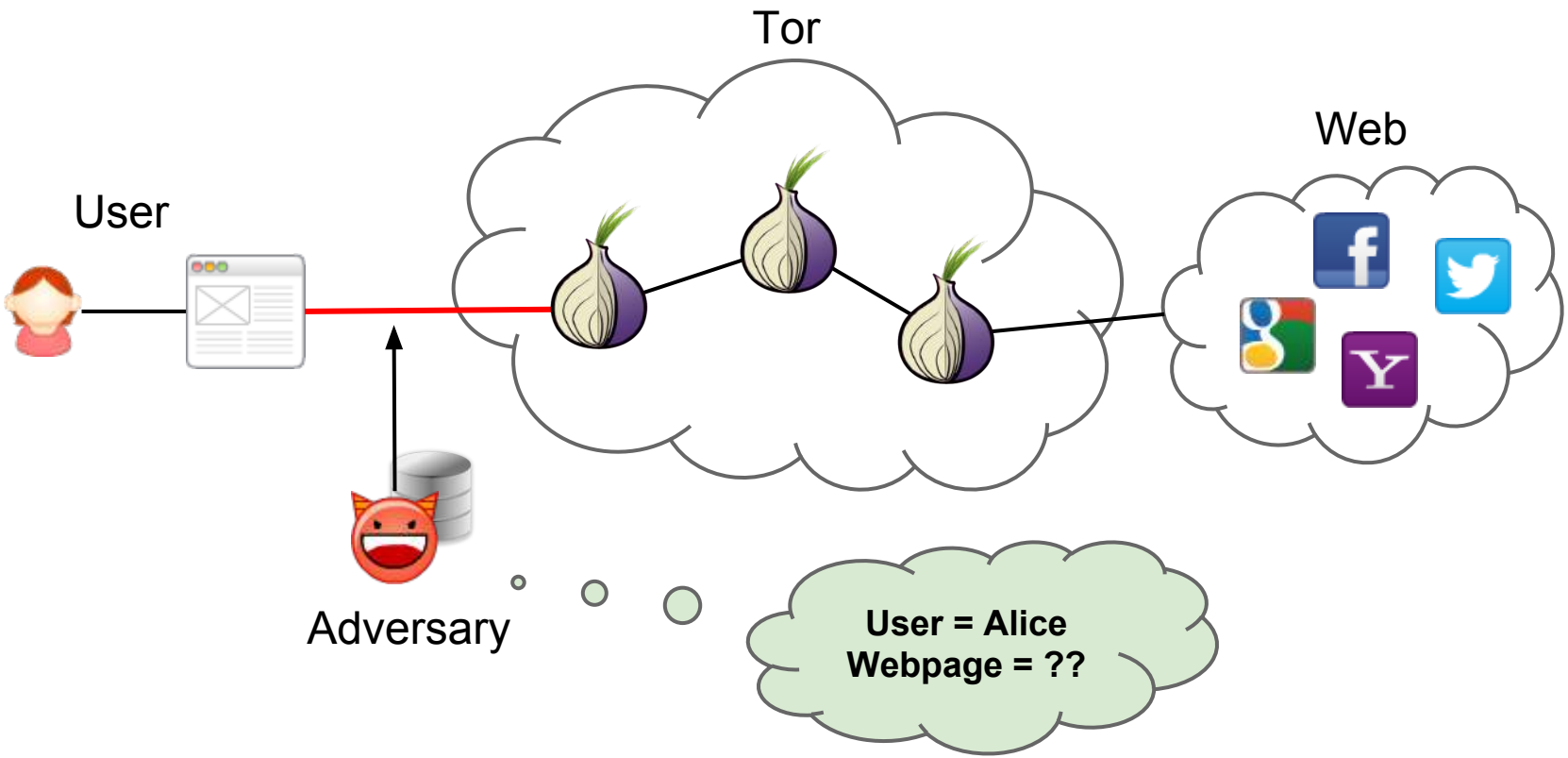
[1]KU Leuven, ESAT/COSIC and iMinds, Leuven, Belgium

[2]UC Berkeley, US

[3]Drexel University, US

CCS 2014, Scottsdale, AZ, USA, November 4, 2014

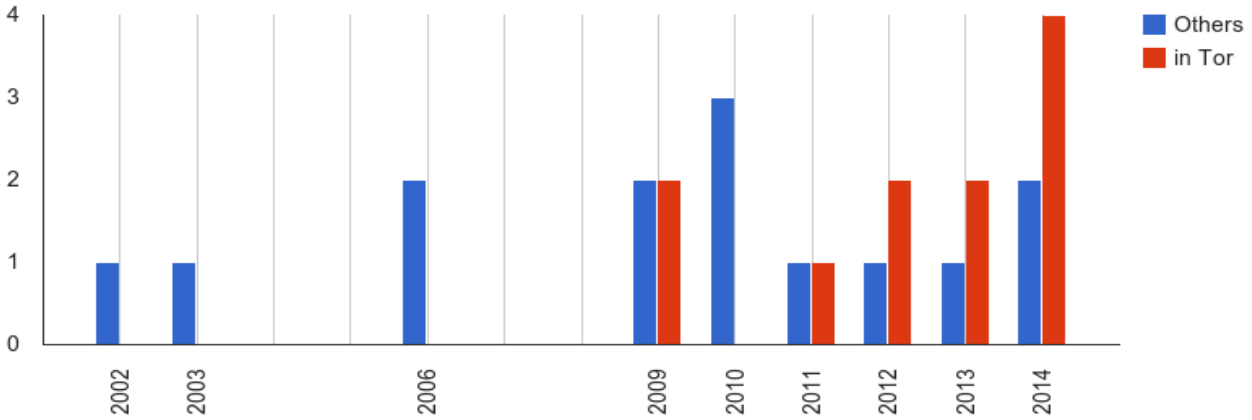
Introduction: how does WF work?



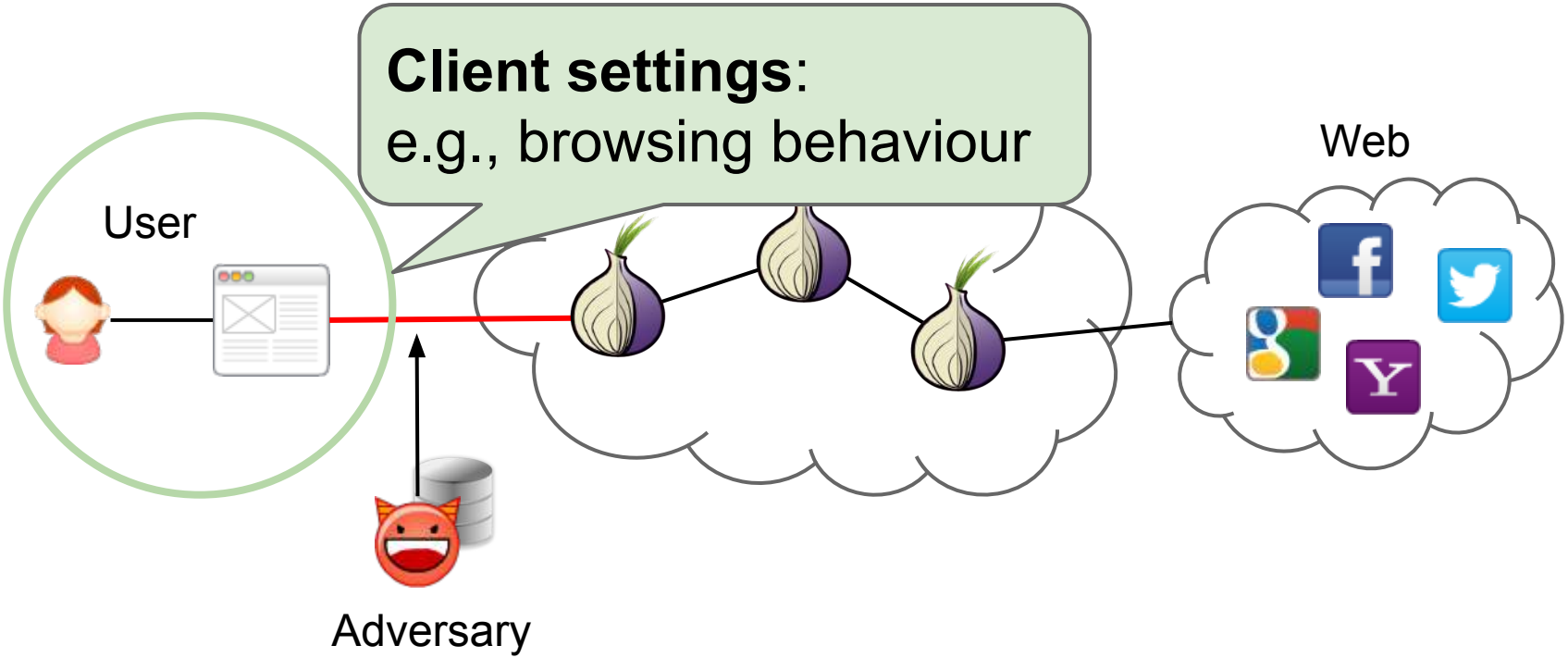
Why is WF so important?

- Tor as the most advanced anonymity network
- Allows an adversary to discover the browsing history
- Series of successful attacks
- Low cost to the adversary

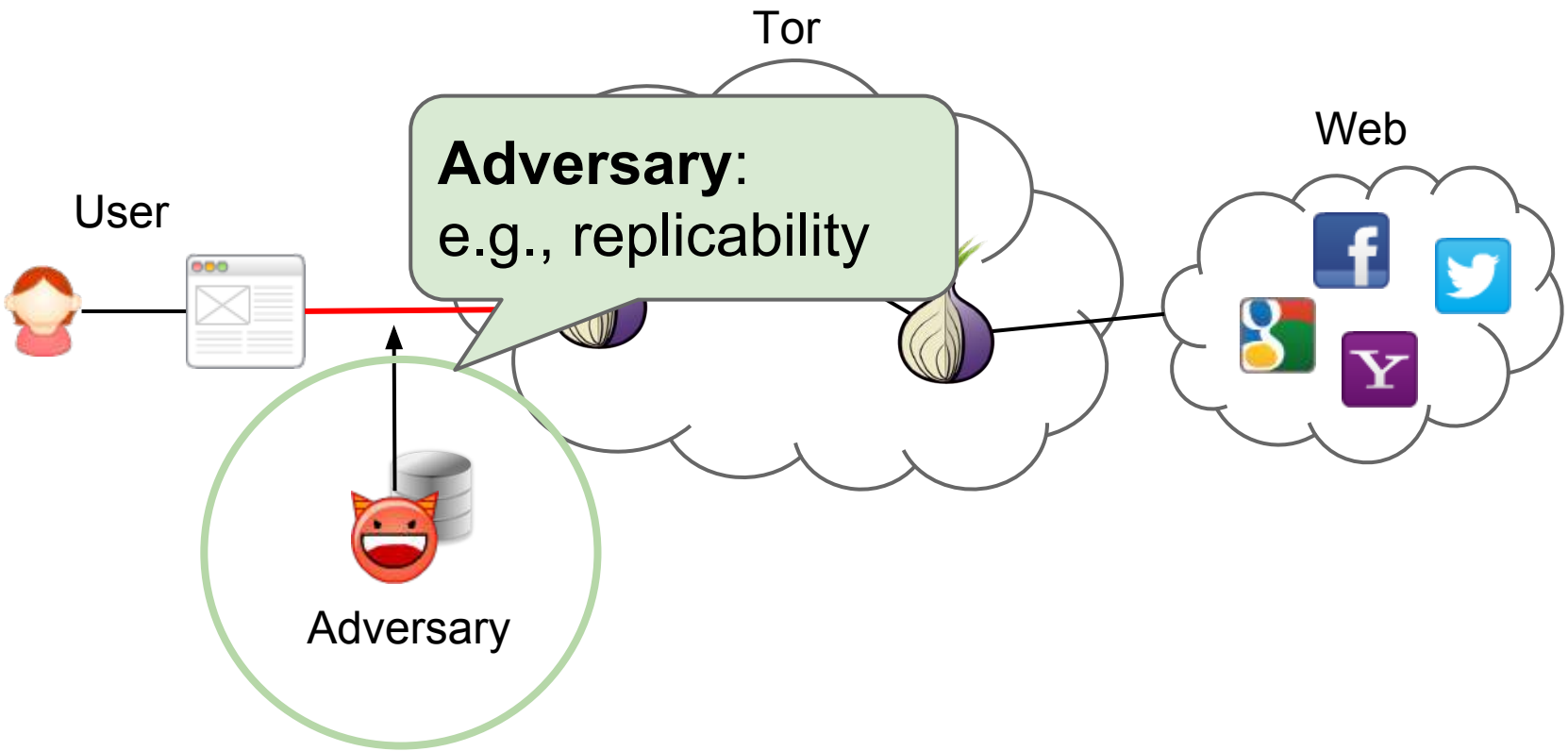
Number of top conference publications on WF (25)



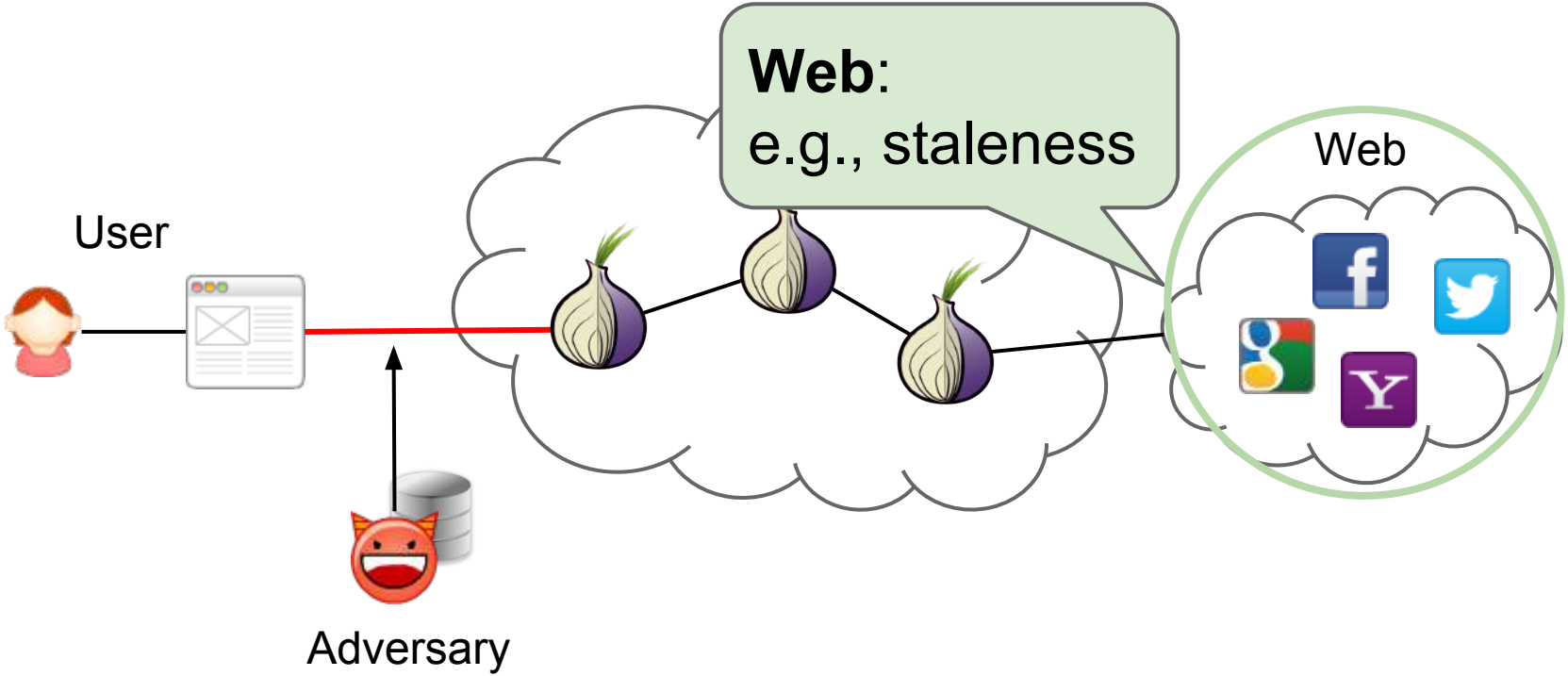
Introduction: unrealistic assumptions



Introduction: unrealistic assumptions



Introduction: unrealistic assumptions



Contributions

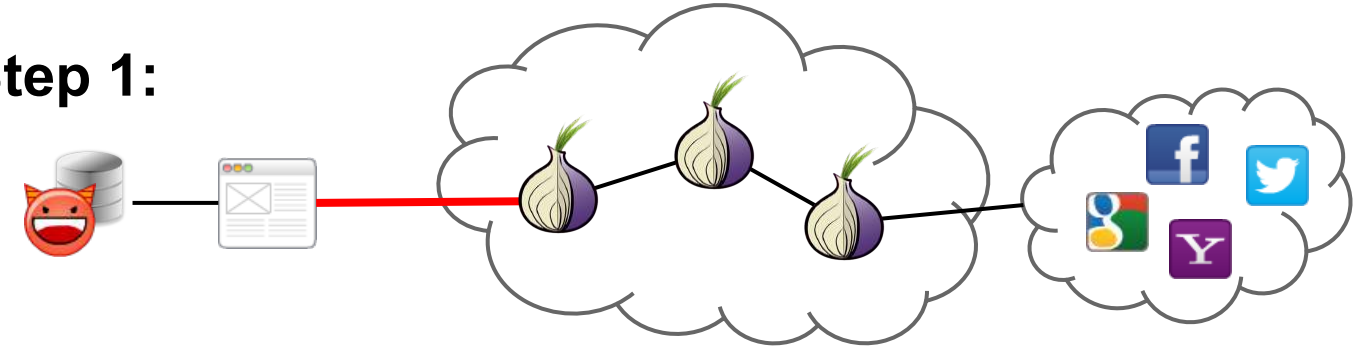
- A critical analysis of the assumptions
- Evaluation of variables that affect accuracy
- An approach to reduce false positives
- A model of the adversary’s cost

Methodology

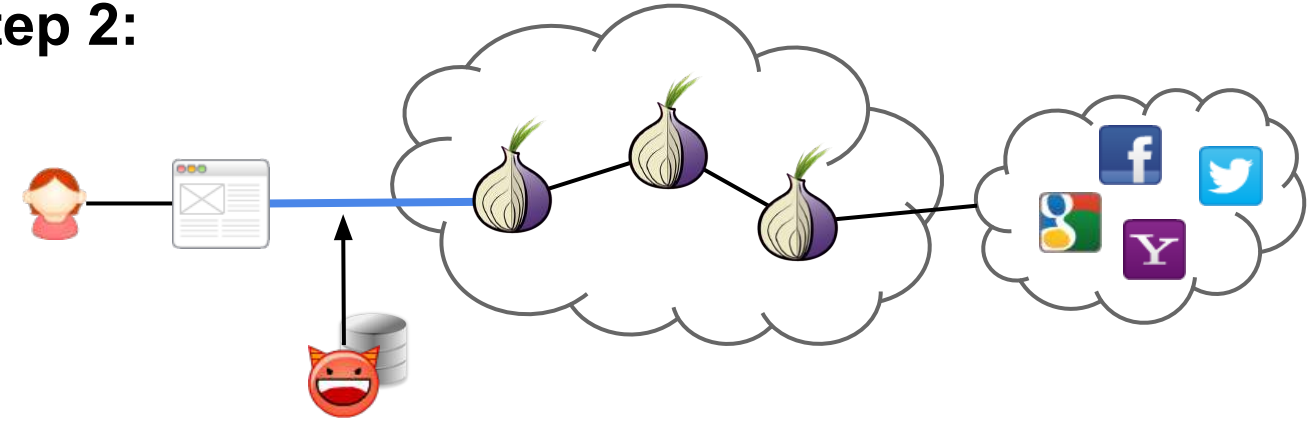
- Based on Wang and Goldberg's
  - Batches and k-fold cross-validation
  - Fast-levenshtein attack (SVM)
- Comparative experiments
  - Key: isolate variable under evaluation (e.g., TBB version)

Comparative experiments: example

- **Step 1:**

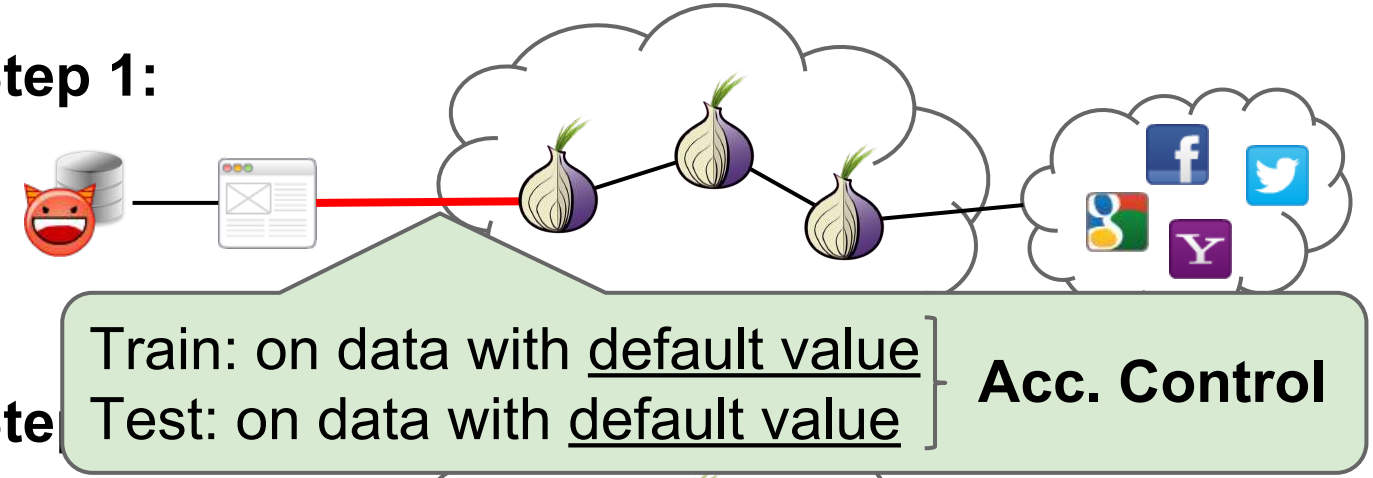


- **Step 2:**

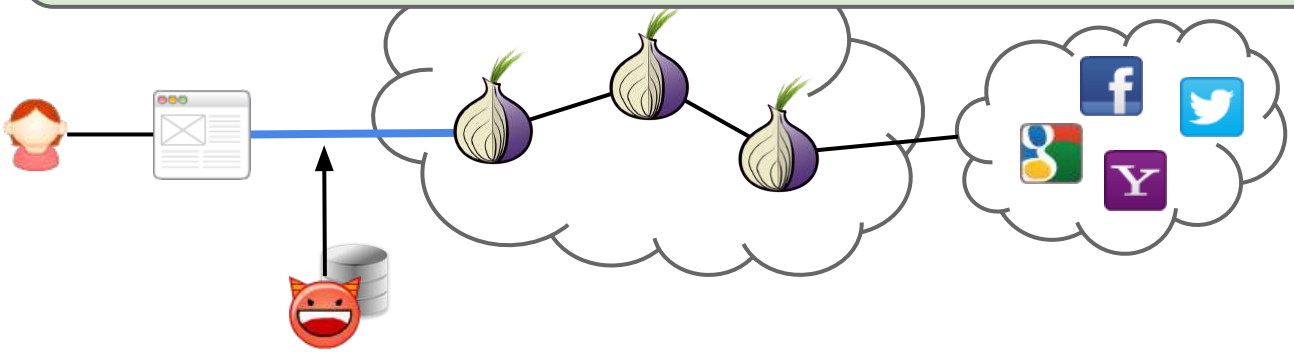


Comparative experiments: example

- **Step 1:**

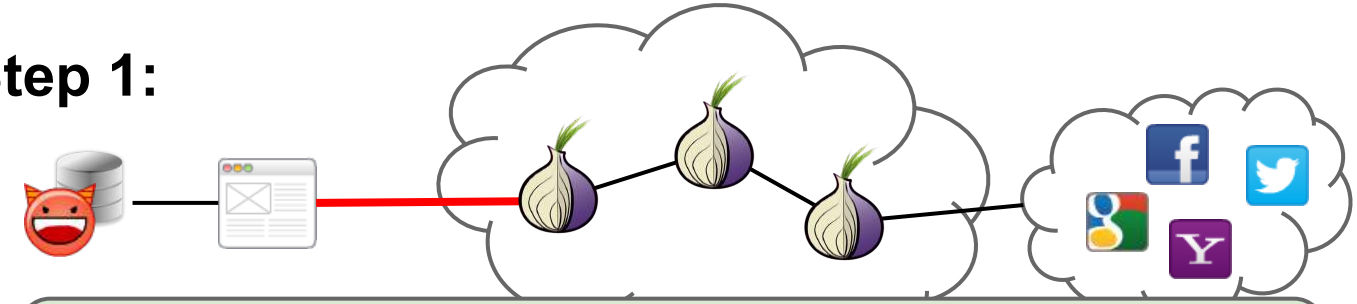


- **Ste**



Comparative experiments: example

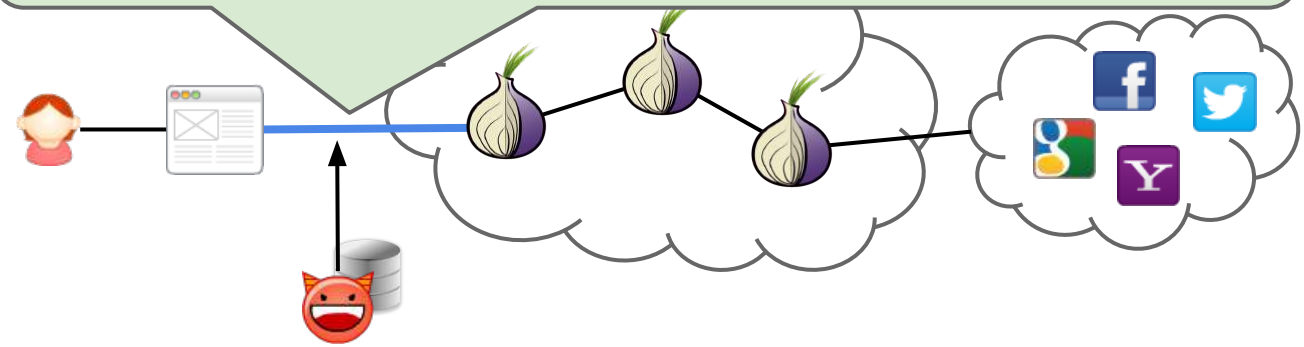
- **Step 1:**



- **St**

Train: on data with default value
Test: on data with value of interest

Acc. Test



Datasets

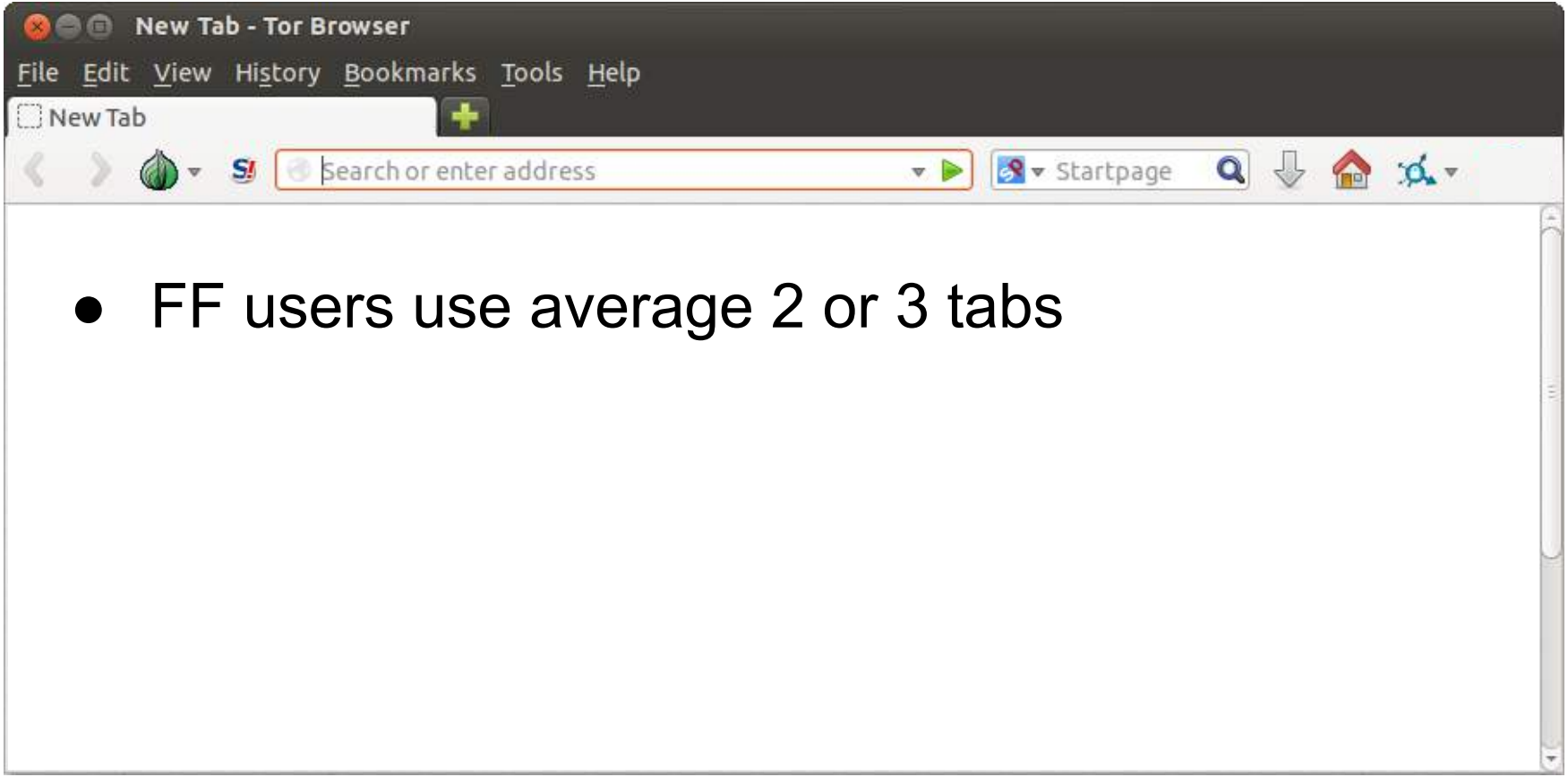
- *Alexa Top Sites*
- *Active Linguistic Authentication Dataset* (ALAD)
  - **Real-world** users (80 users, 40K unique URLs)
  - Training on Alexa and testing on ALAD?

Datasets

- *Alexa Top Sites*
- *Active Linguistic Authentication Dataset* (ALAD)
  - **Real-world** users (80 users, 40K unique URLs)
  - Training on Alexa and testing on ALAD?

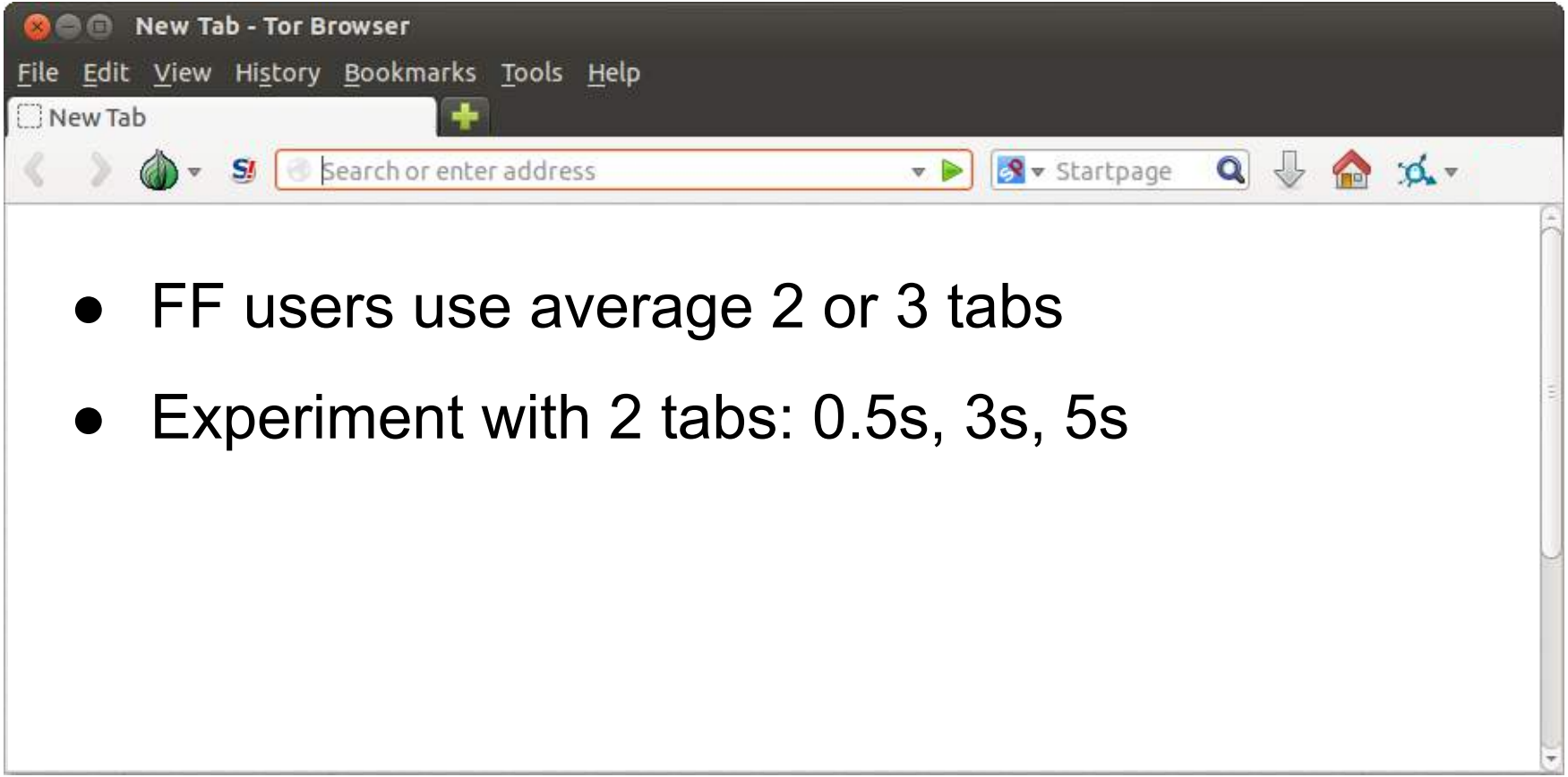
45% not in Alexa top **100** ⇨ Prohibitive number of FPs

Experiments: multitab browsing



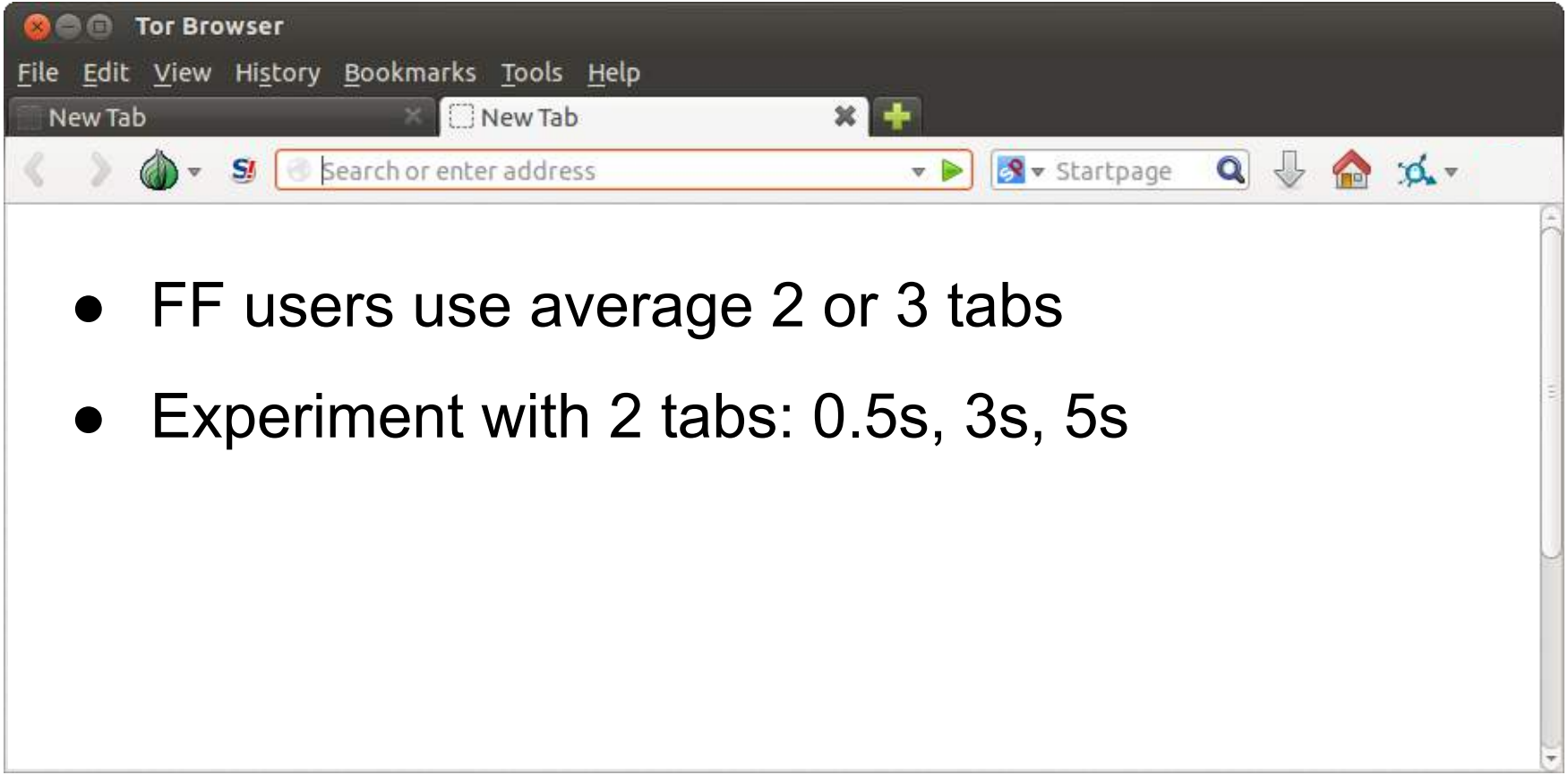
- FF users use average 2 or 3 tabs

Experiments: multitab browsing



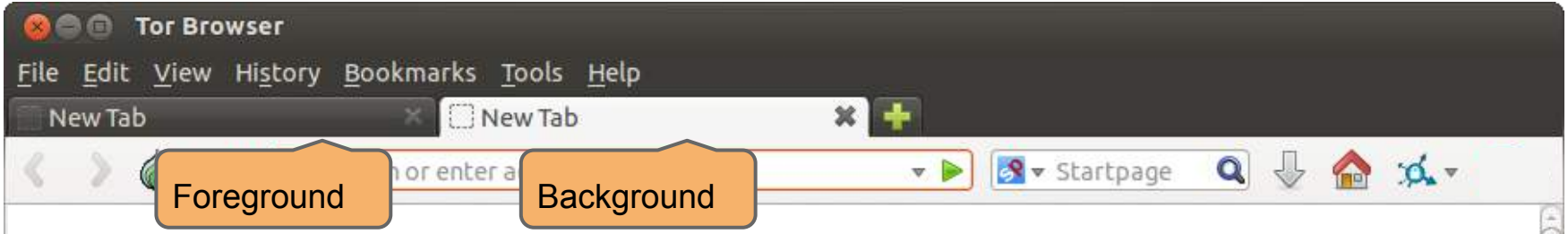
- FF users use average 2 or 3 tabs
- Experiment with 2 tabs: 0.5s, 3s, 5s

Experiments: multitab browsing



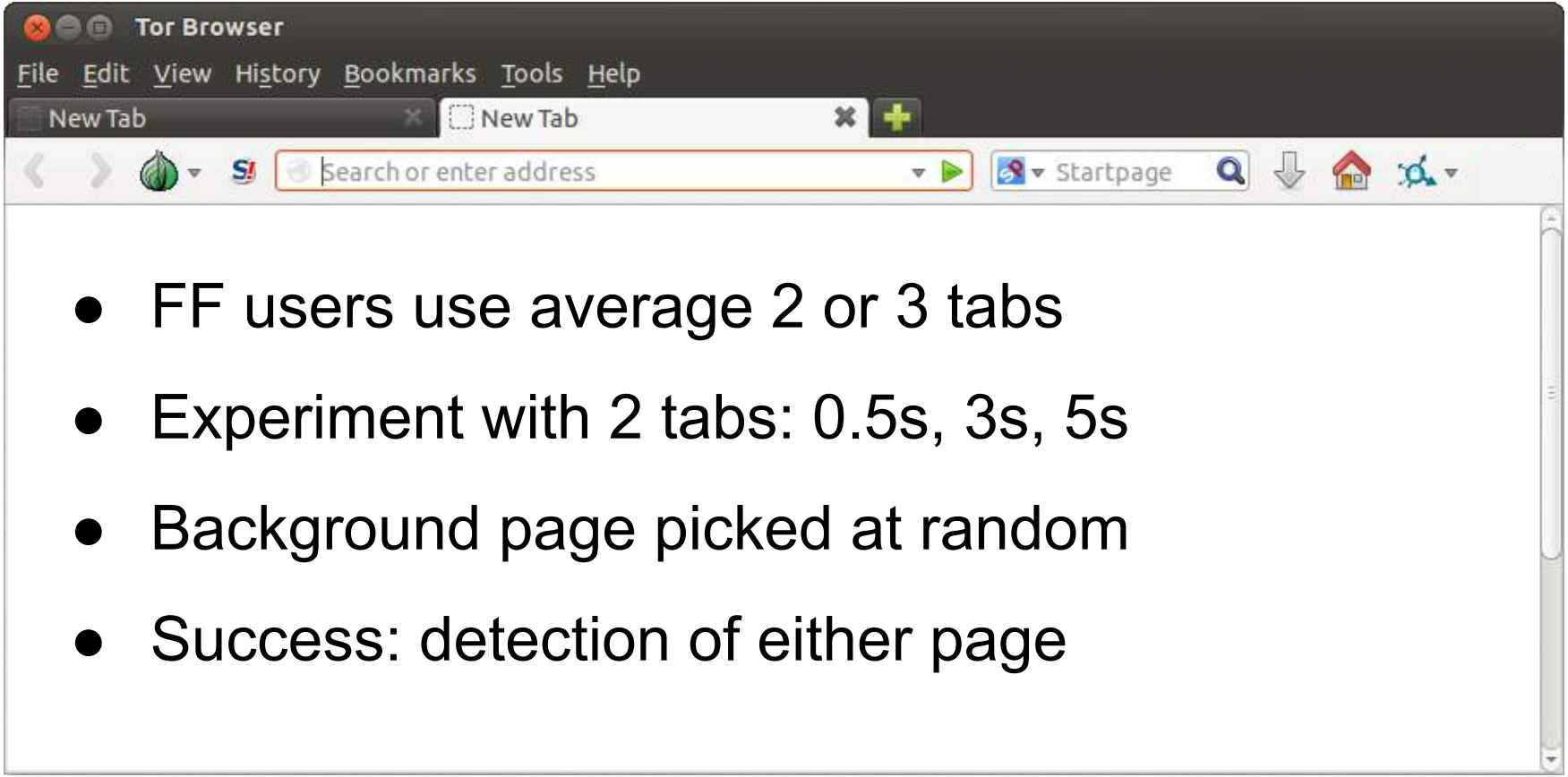
- FF users use average 2 or 3 tabs
- Experiment with 2 tabs: 0.5s, 3s, 5s

Experiments: multitab browsing



- FF users use average 2 or 3 tabs
- Experiment with 2 tabs: 0.5s, 3s, 5s
- Background page picked at random

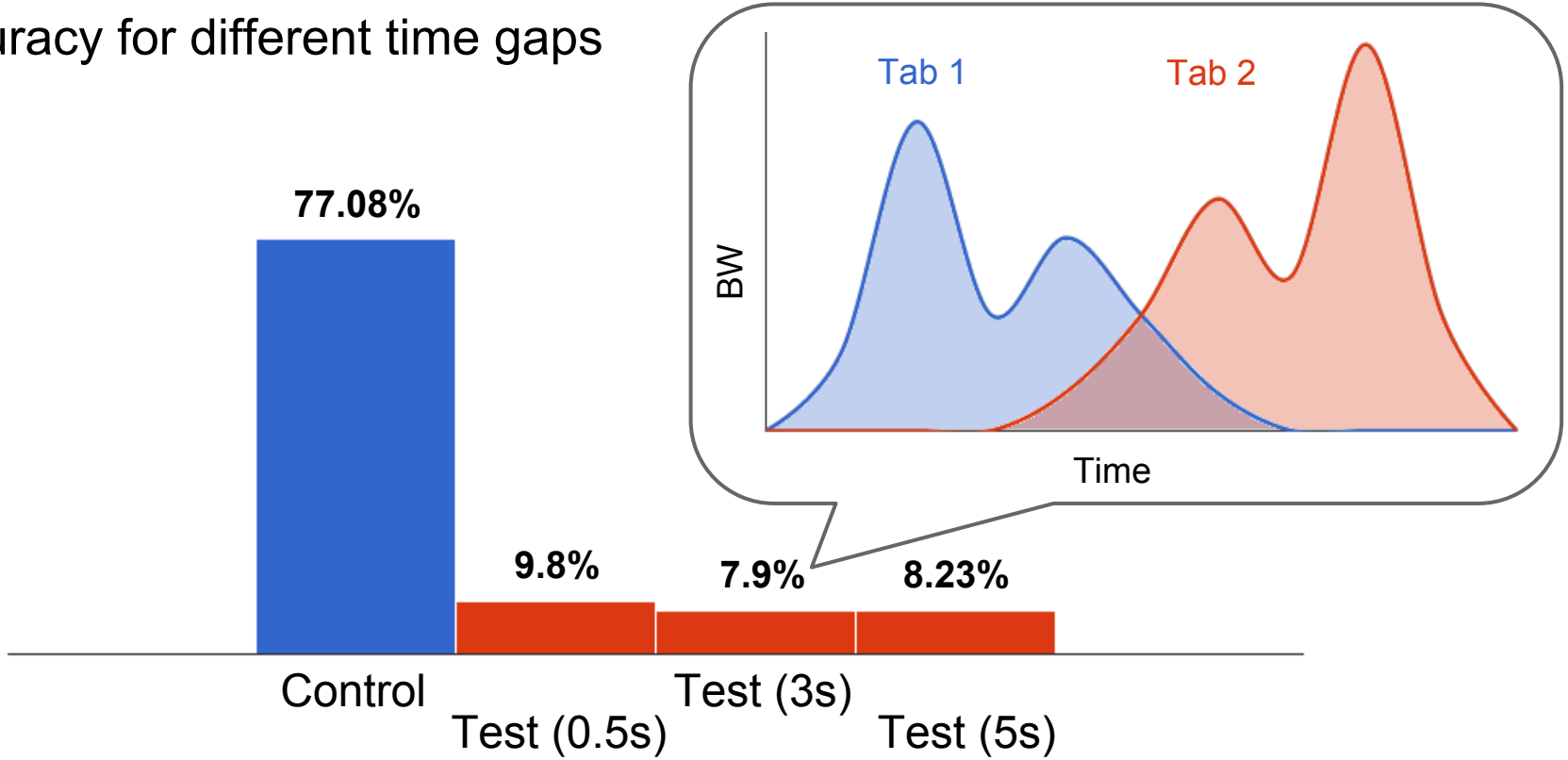
Experiments: multitab browsing



- FF users use average 2 or 3 tabs
- Experiment with 2 tabs: 0.5s, 3s, 5s
- Background page picked at random
- Success: detection of either page

Experiments: multitab browsing

Accuracy for different time gaps

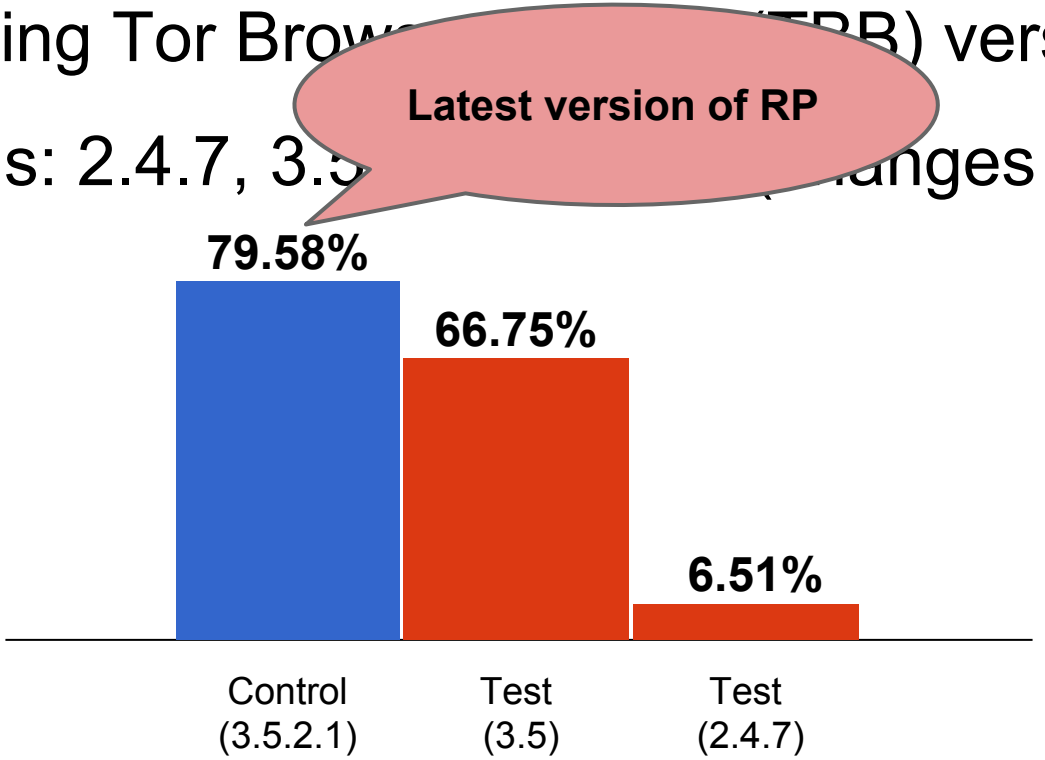


Experiments: TBB versions

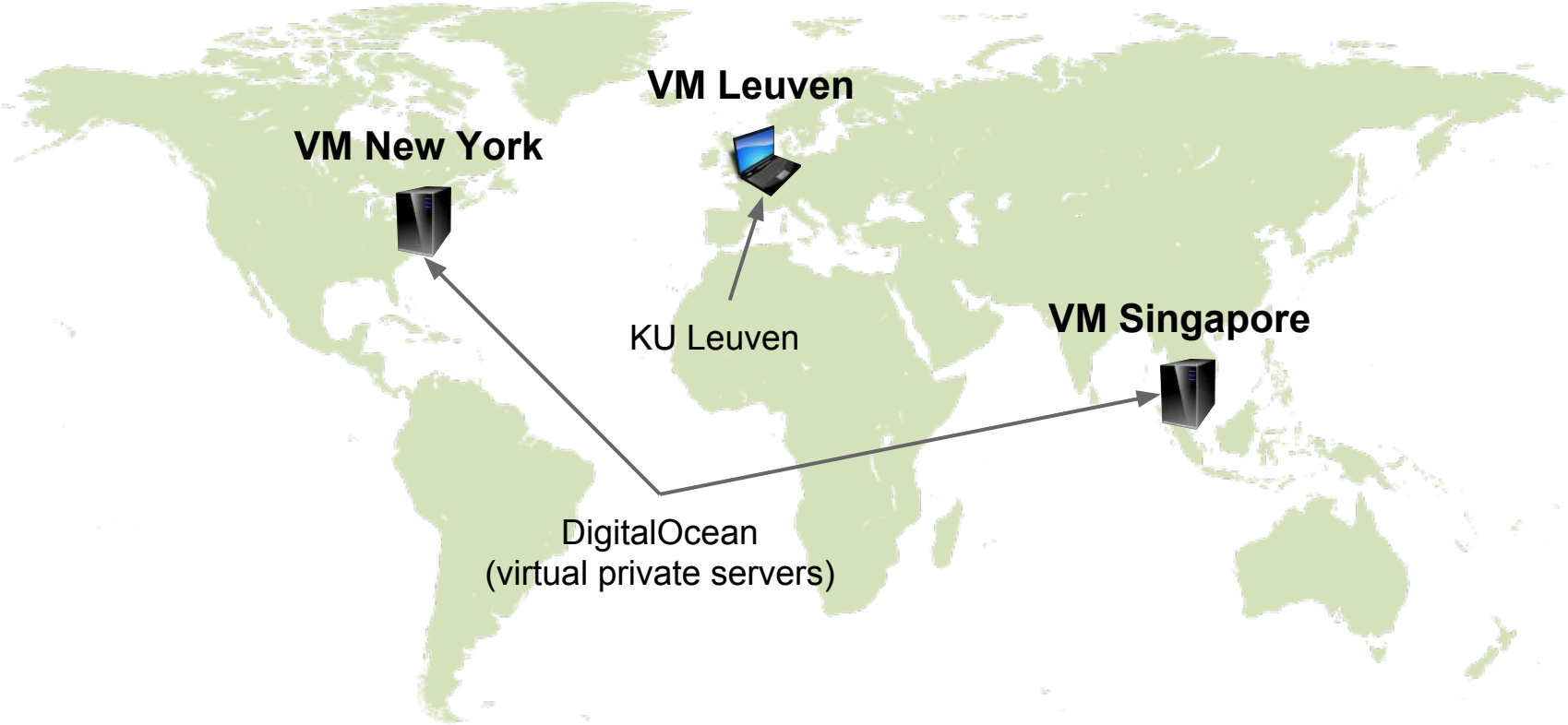
- Coexisting Tor Browser Bundle (TBB) versions
- Versions: 2.4.7, 3.5 and 3.5.2.1 (changes in RP, etc.)

Experiments: TBB versions

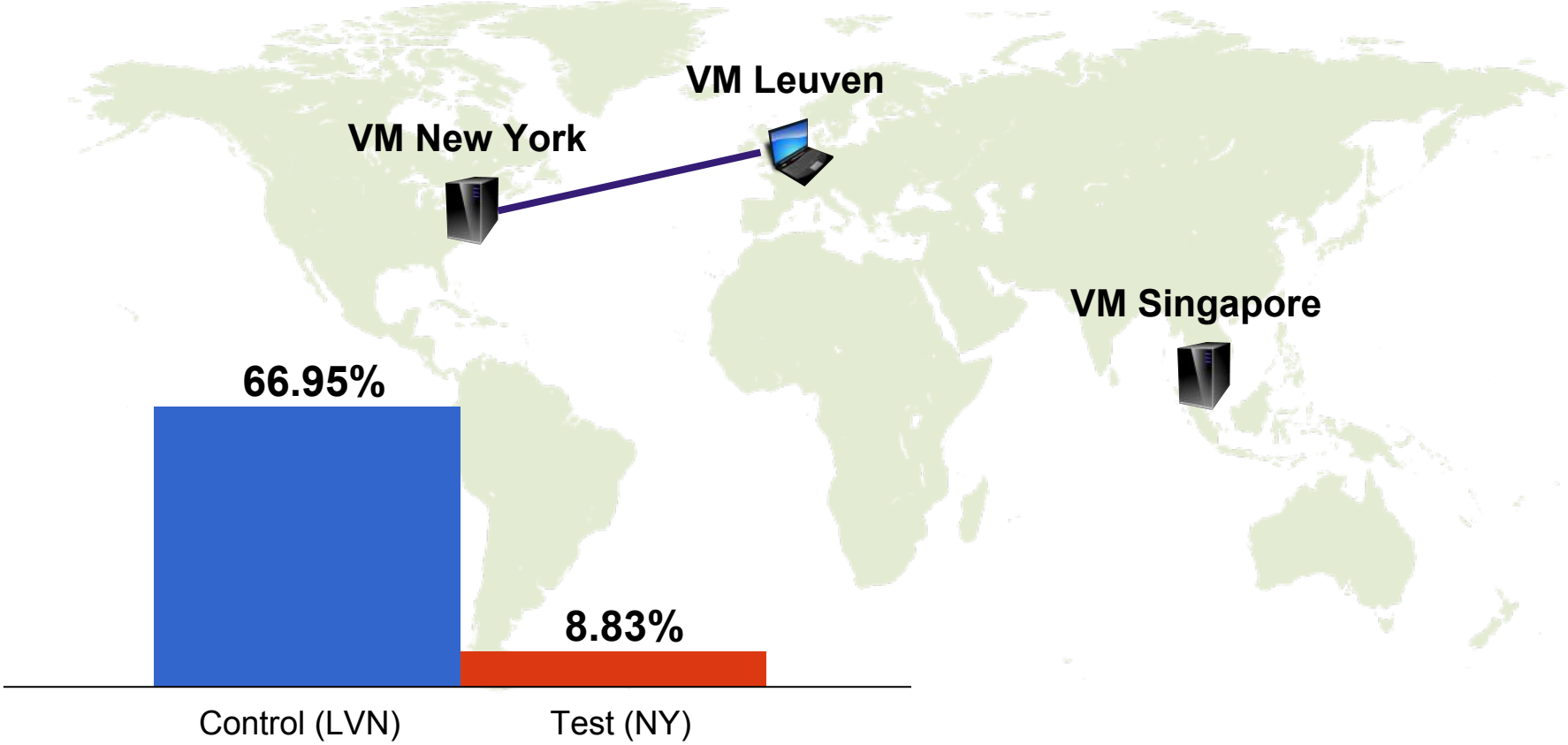
- Coexisting Tor Brow TBB) versions
- Versions: 2.4.7, 3.5 anges in RP, etc.)



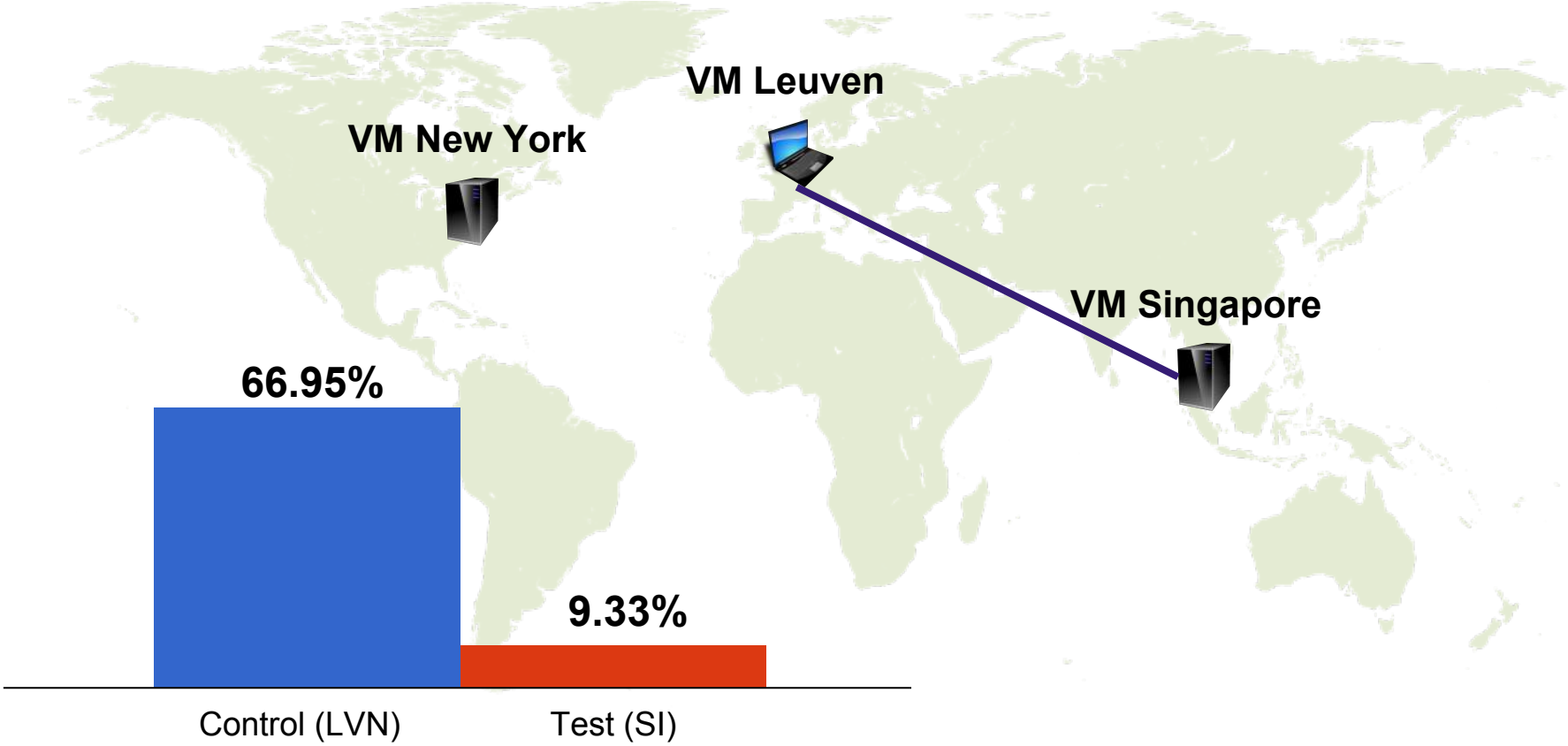
Experiments: network conditions



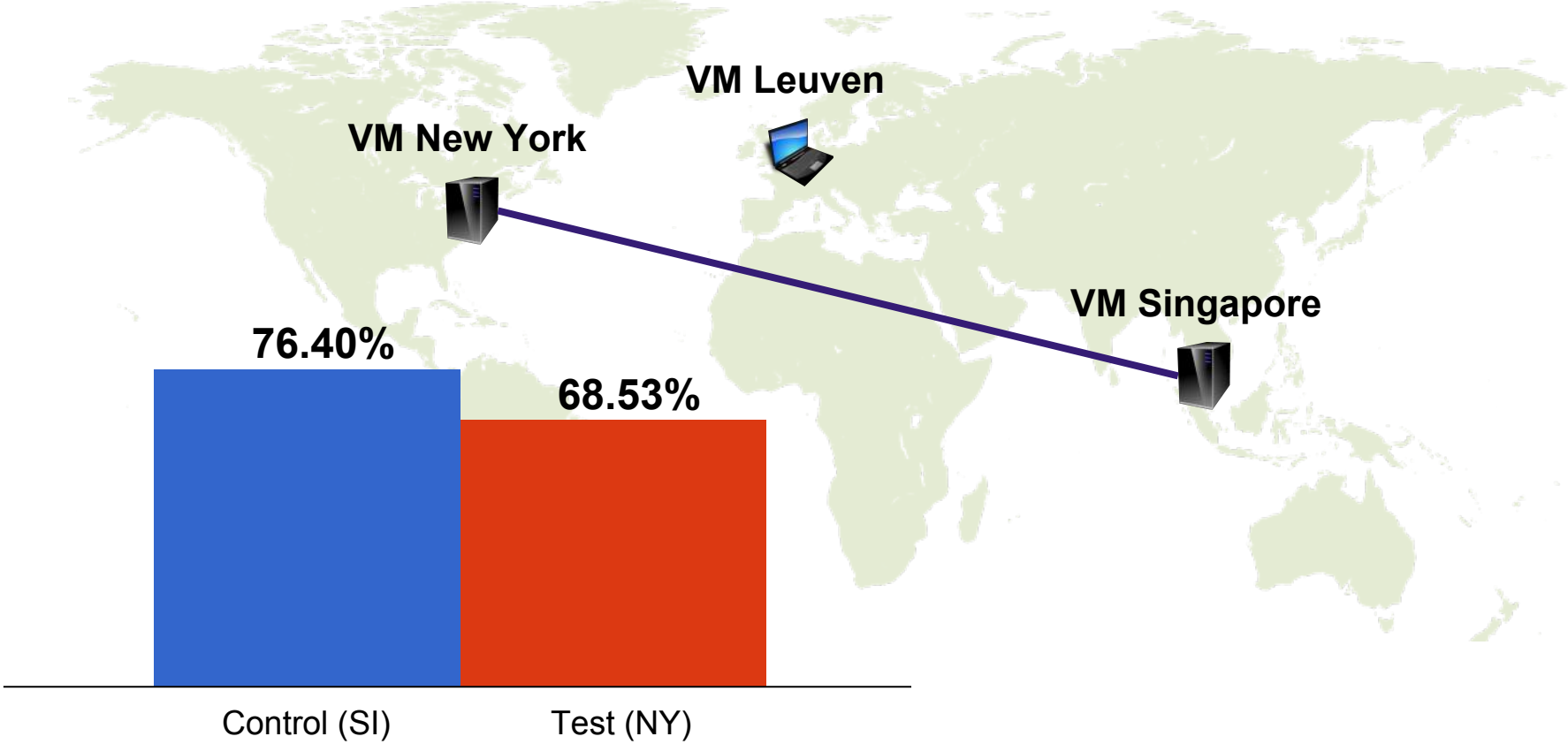
Experiments: network conditions



Experiments: network conditions



Experiments: network conditions

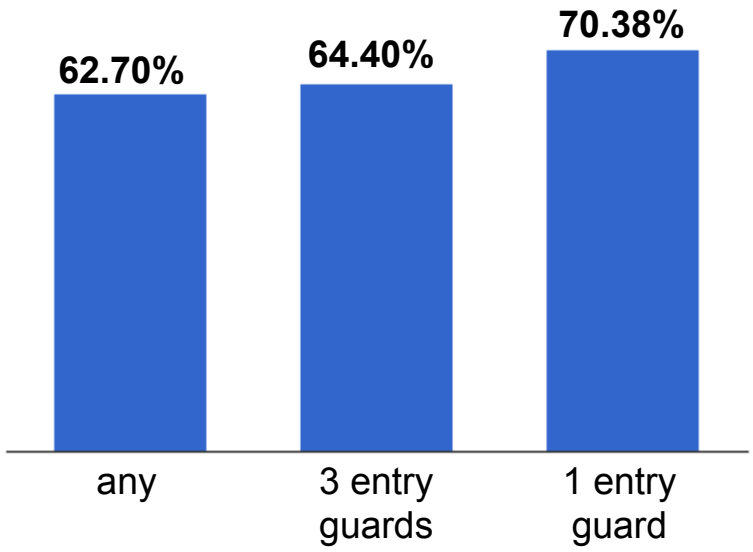


Experiments: entry guard config.

- What entry config. works better for training?
- 3 configs.:
  - Fix 1 entry guard
  - Pick entry from a list of 3 entries guards (default)
  - Pick entry from all possible entries guards (Wang and Goldberg)

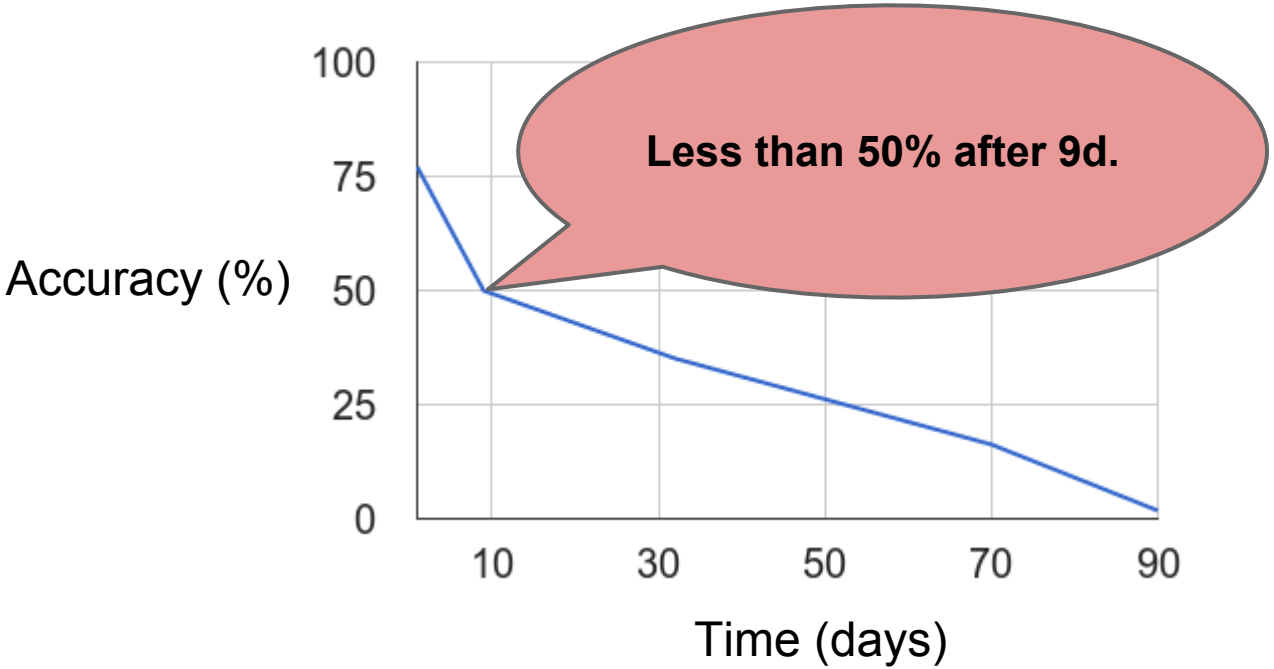
Experiments: entry guard config.

Accuracy for different entry guard configurations

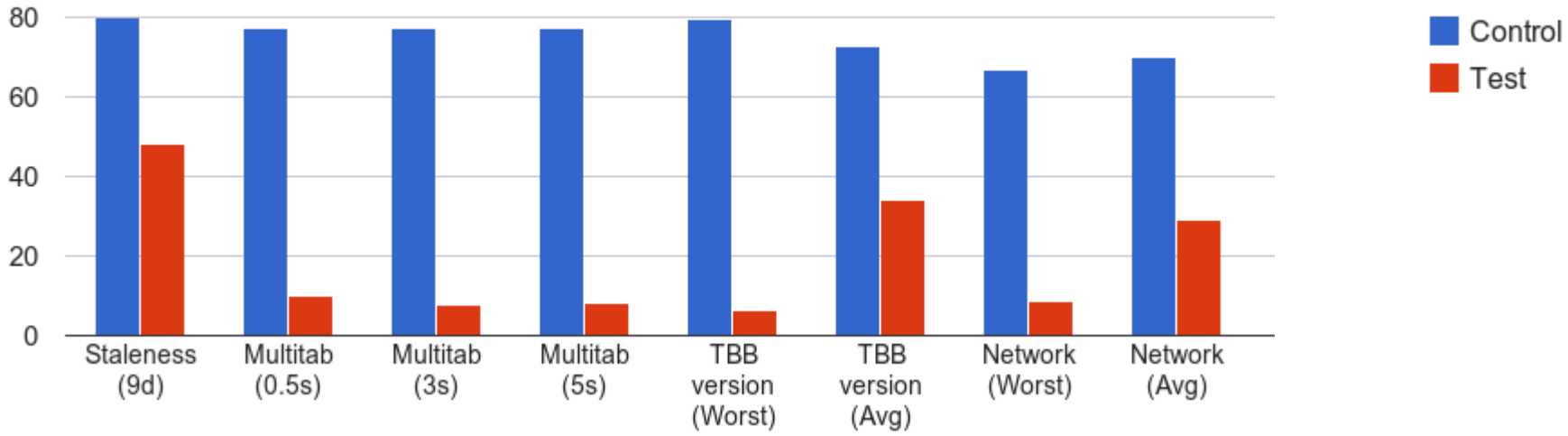


Experiments: data staleness

Staleness of our collected data over 90 days



Summary


80
60
40
20
0
Staleness (9d)
Multitab (0.5s)
Multitab (3s)
Multitab (5s)
TBB version (Worst)
TBB version (Avg)
Network (Worst)
Network (Avg)
Control
Test

The base rate fallacy: example

- Breathalyzer test:
  - **0.88** identifies truly drunk drivers (true positives)
  - **0.05** false positives
- Alice gives positive in the test
  - What is the probability that she is indeed drunk? (**BDR**)
  - Is it 0.95? Is it 0.88? Something in between?

The base rate fallacy: example

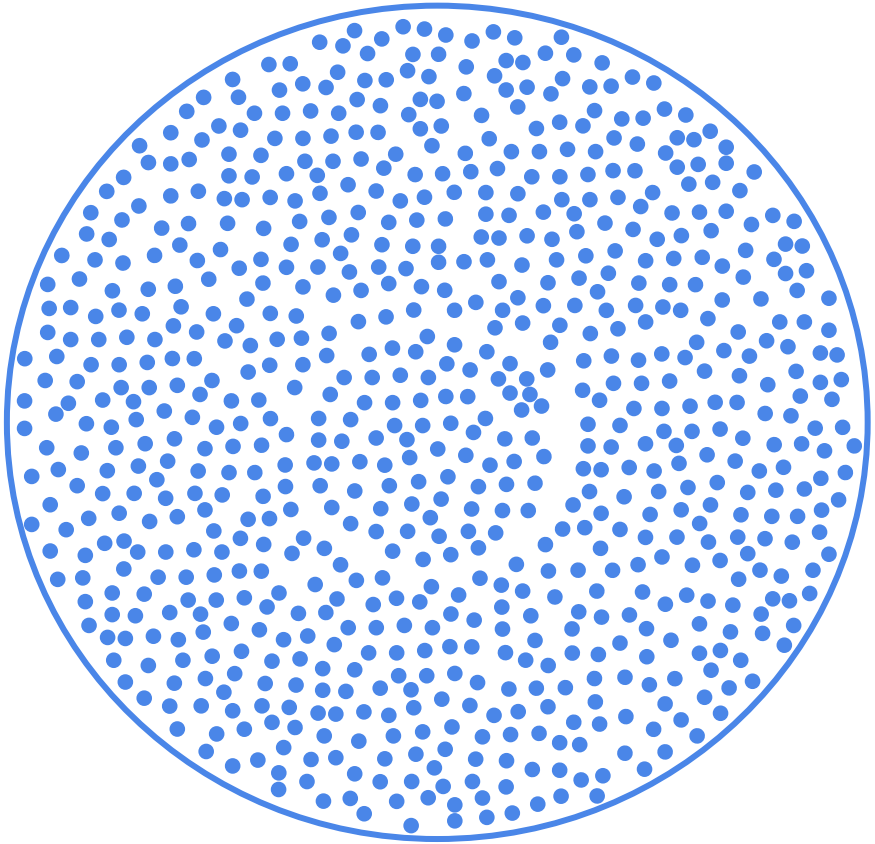
- Breathalyzer test:
  - **0.88** identifies truly drunk drivers (true positives)
  - **0.05** false positives
- Alice gives positive in [illegible]
  - What is the p[illegible] indeed drunk? (**BDR**)
  - Is it 0.95? Is [illegible]ng in between?

Only 0.1!

The base rate fallacy: example

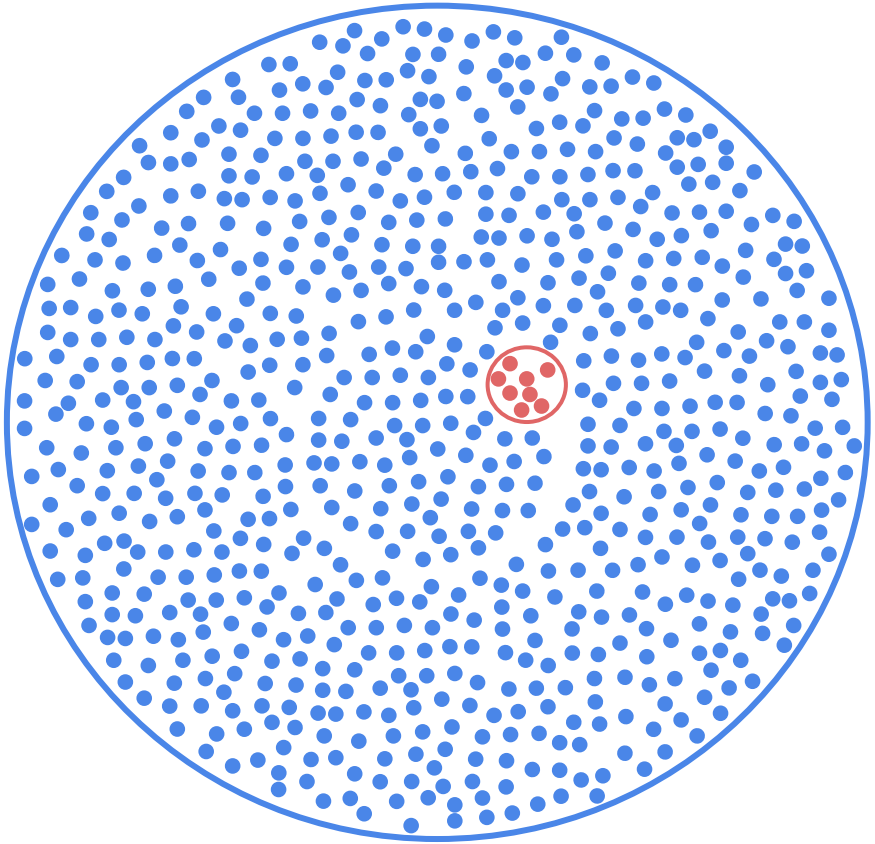
- Circumference represents the world of drivers.

- Each dot represents a driver.



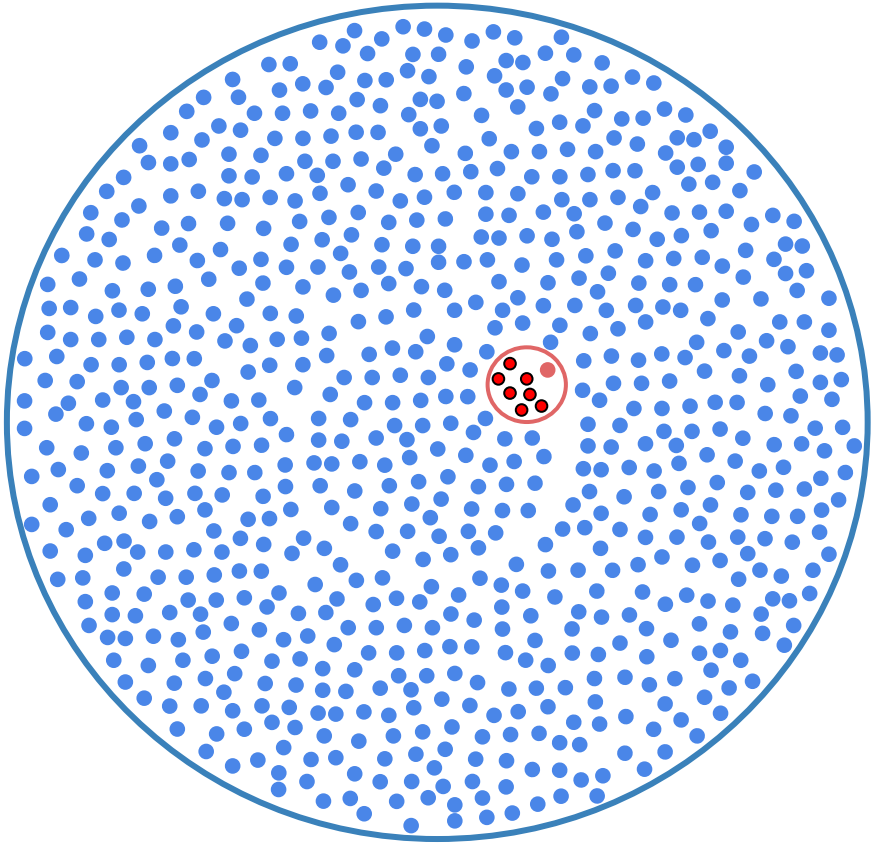
The base rate fallacy: example

- 1% of drivers are driving drunk (**base rate or prior**).



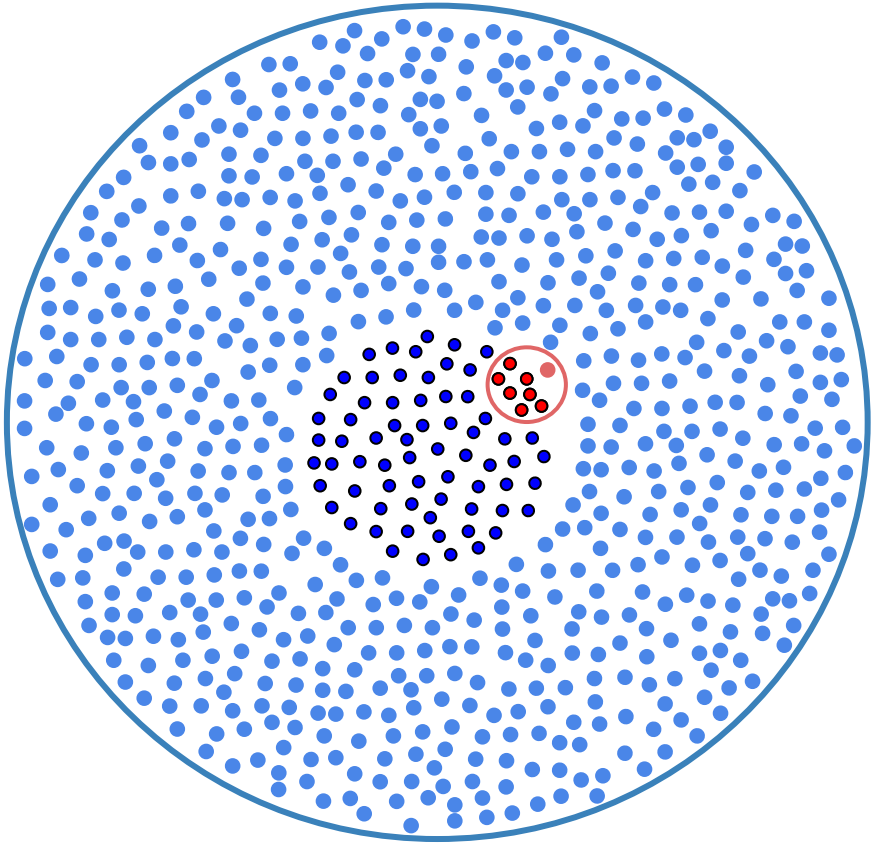
The base rate fallacy: example

- From drunk people 88% are identified as drunk by the test



The base rate fallacy: example

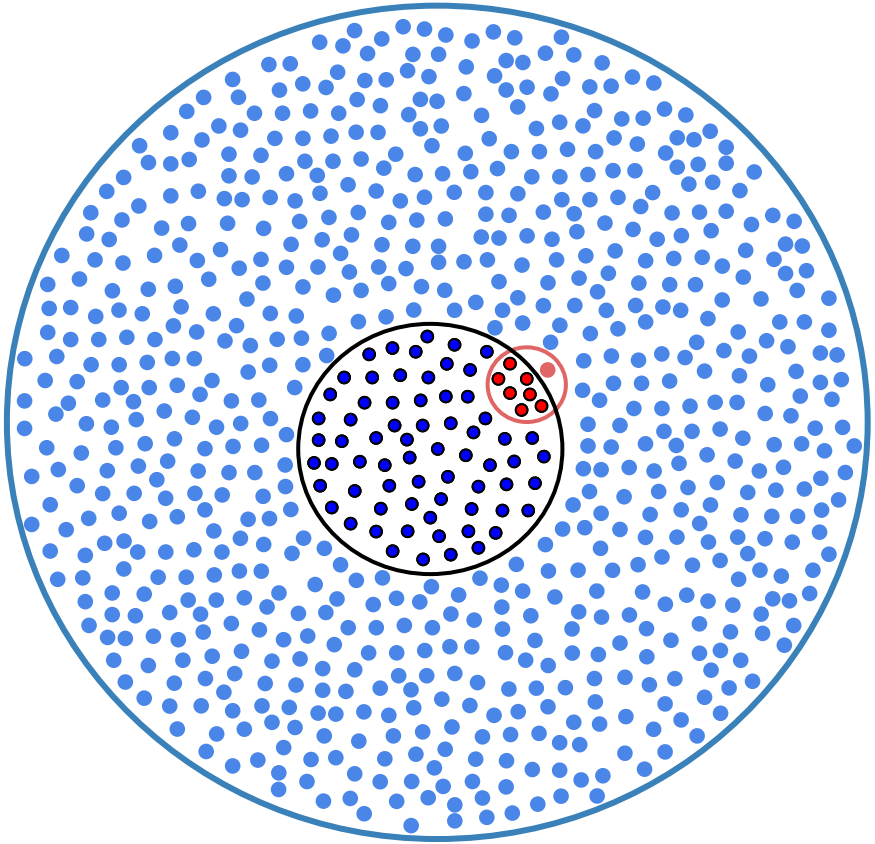
- From the not drunk people, 5% are erroneously identified as drunk



The base rate fallacy: example

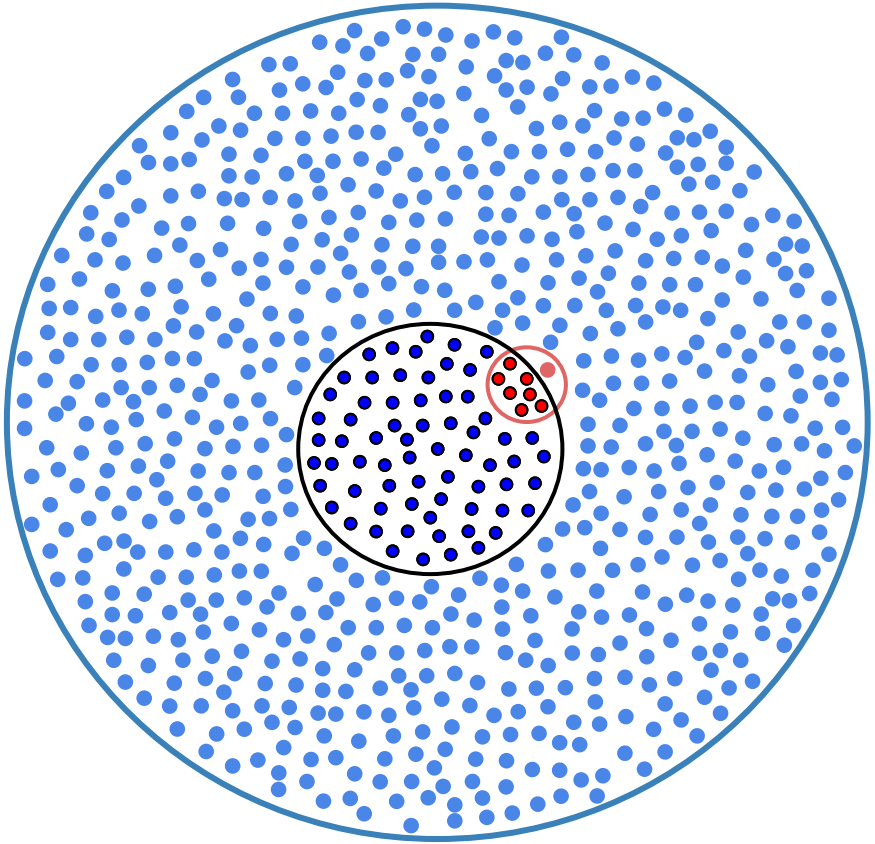
- Alice must be within the black circumference
- Ratio of red dots within the black circumference:

BDR = 7/70 = **0.1 !**



The base rate fallacy in WF

- Base rate must be taken into account
- In WF:
  - Blue: webpages
  - Red: monitored
  - Base rate?



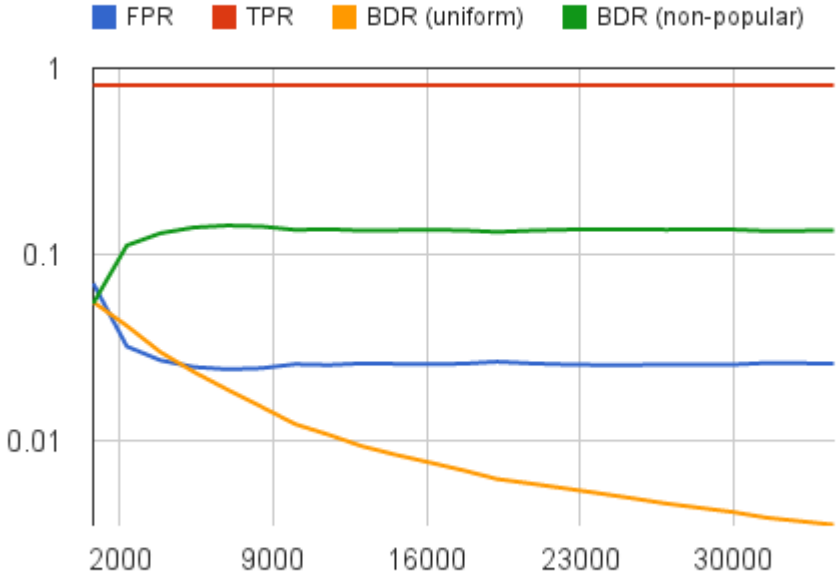
The base rate fallacy in WF

- Probability of visiting a monitored page?
- “false positives matter a lot”[1]
- Experiment: 35K world

[1]Mike Perry, “A Critique of Website Traffic Fingerprinting Attacks”, Tor project Blog, 2013. https://blog.torproject.org/blog/critique-website-traffic-fingerprinting-attacks.

Experiment: BDR in a 35K world

- Uniform world
- Non-popular pages from ALAD



Size of the world

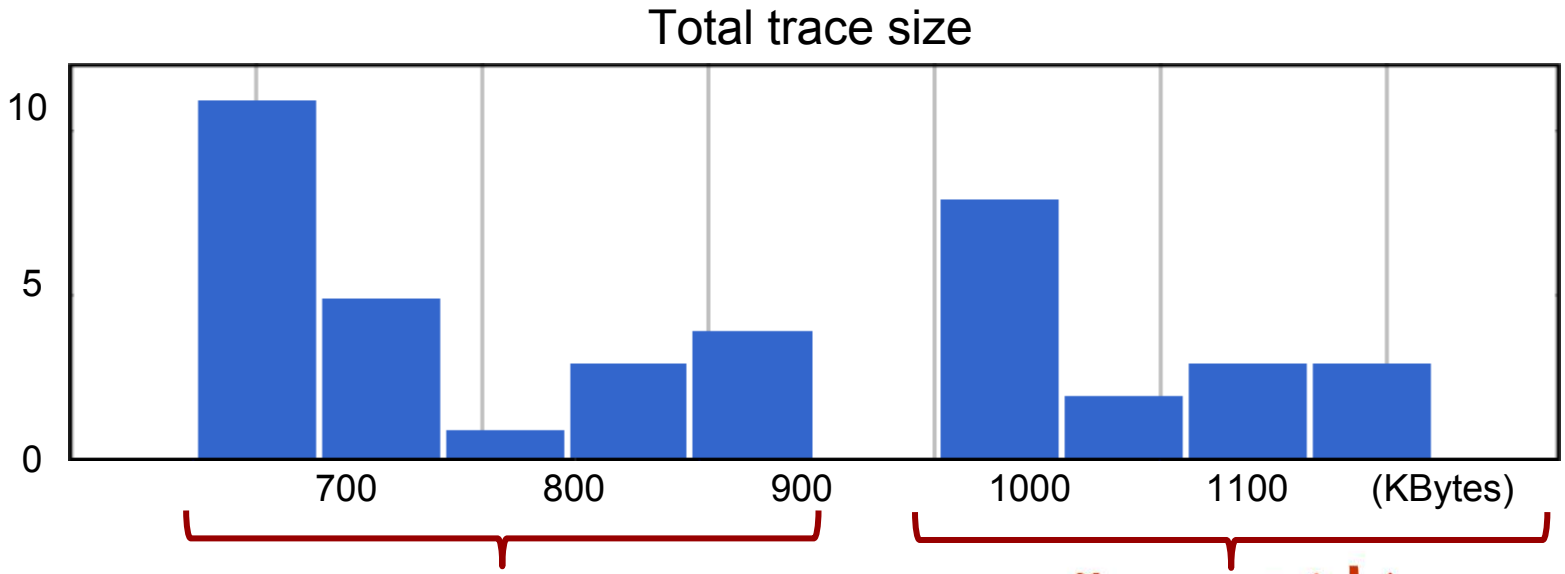
Classify, but verify

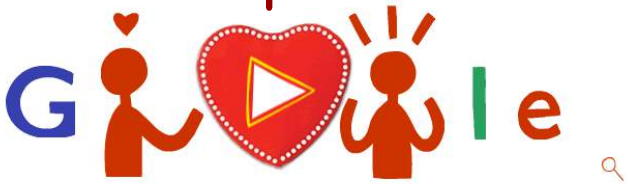
- Verification step to test classifier confidence
- Number of FPs reduced **397-42 (400)**
- But BDR is still very low for non popular pages

Cost for the adversary

- Adversary’s cost will depend on:
  - Number of pages

Versions of a page: St Valentine's doodle





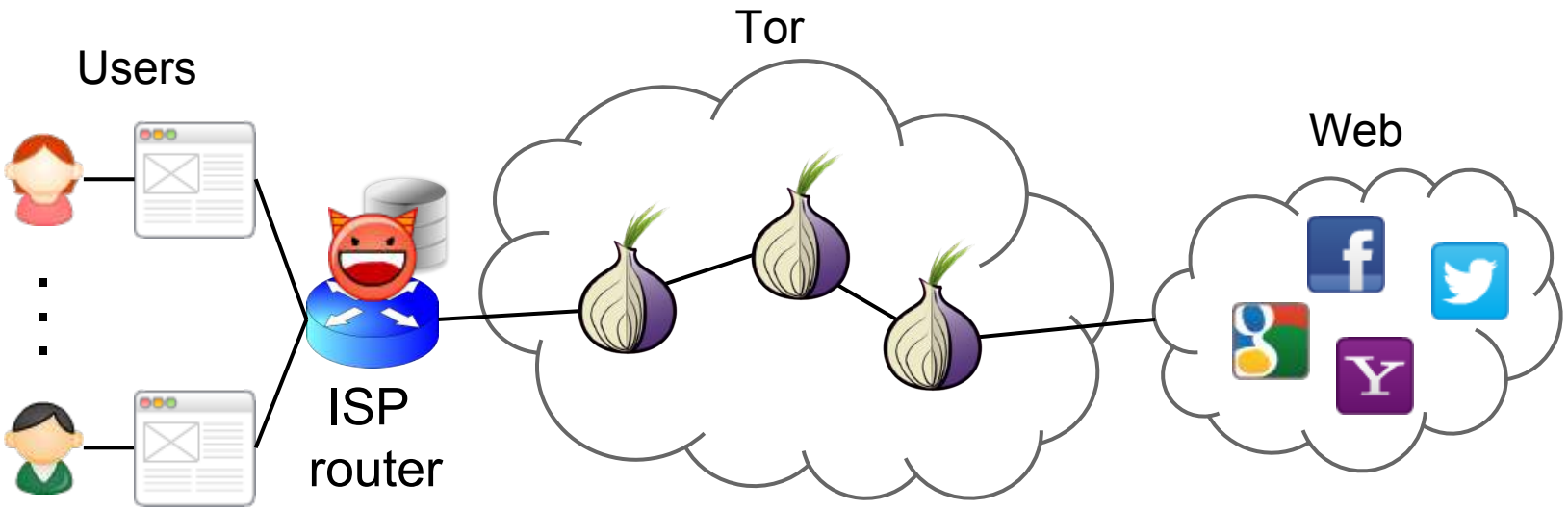
13 Feb 2013

14 Feb 2013

Cost for the adversary

- Adversary’s cost will depend on:
  - Number of pages
  - Number of targets

Non-targeted attacks



Cost for the adversary

- Adversary’s cost will depend on:
  - Number of pages
  - Number of targets
  - Training and testing complexities

Cost for the adversary

- Adversary's cost will depend on:
  - Number of pages
  - Number of targets
  - Training and testing complexities
- To maintain a successful WF system is costly

Limitations

- We took samples and may not be representative of all possible practical scenarios
- Variables difficult to control
  - Time gap
  - Tor circuit

Conclusions

- WF attack fails in realistic conditions
- We do not completely dismiss the attack
- Attack can be enhanced at a greater cost
- Defenses might be cheaper in practice

Thank you for your attention.

Questions?