CS 563 - Advanced Computer Security: Web Privacy

Professor Adam Bates
Fall 2018

Administrative

Learning Objectives:

- Consider the difference between security and privacy
- Discuss work on browser privacy, location privacy
- Survey broad topics in the "web privacy" area

Announcements:

- Reaction paper was due today (and all classes)
- Feedback for reaction papers soon
- Next Wednesday, will discuss first "homework"

Reminder: Please put away (backlit) devices at the start of class

A Brief Note

Security versus Privacy?

A False Dichotomy

There is no ~~cloud~~ privacy
It's just someone else's ~~computer~~ security

- Personal Opinion: *Privacy is often used as a diminutive term to downplay the importance of individual security.*
- "Privacy" refers to a class of important security problems, often related to individual liberties.
- The <u>Security Triad</u> captures all privacy problems, and privacy problems can be found in all sections of the triad.

A False Dichotomy

There is no ~~cloud~~ privacy
It's just someone else's ~~computer~~ security

- Confidentiality: Who can access my personal data? Can the data I explicitly disclose be used to make sensitive inferences about me?
- Integrity: Who manages the data that I consume? Can unauthorized parties affect that data?
- Availability: Is my personal data accessible to me and other authorized partied when I need it?

Tracking Web Browsers

- Browser Tracking: The ability to associate a browser's activities at different times and on different websites.

- Cookies: Data from a website that is stored in the browser.

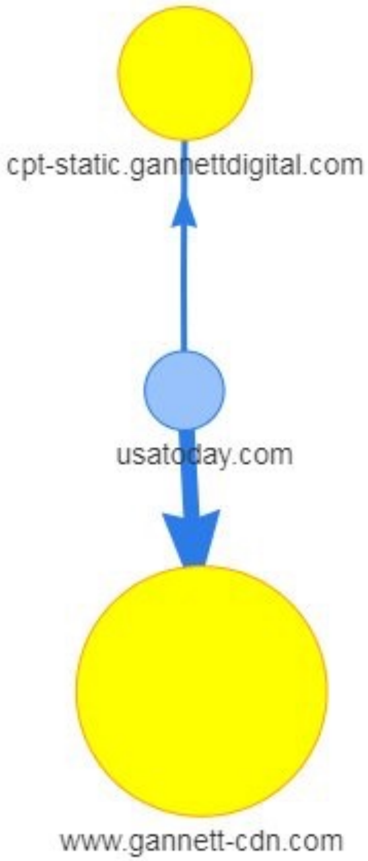
- Enables a stateful Internet

- Same-Origin Policies limit cookie's use in browser tracking.

- Supercookies: Any alternative to HTTP cookies that can be used to track browsers across multiple website.
  - Ex: ETags used in web caching (Microsoft circa 2011)

Aside: Who Cares?

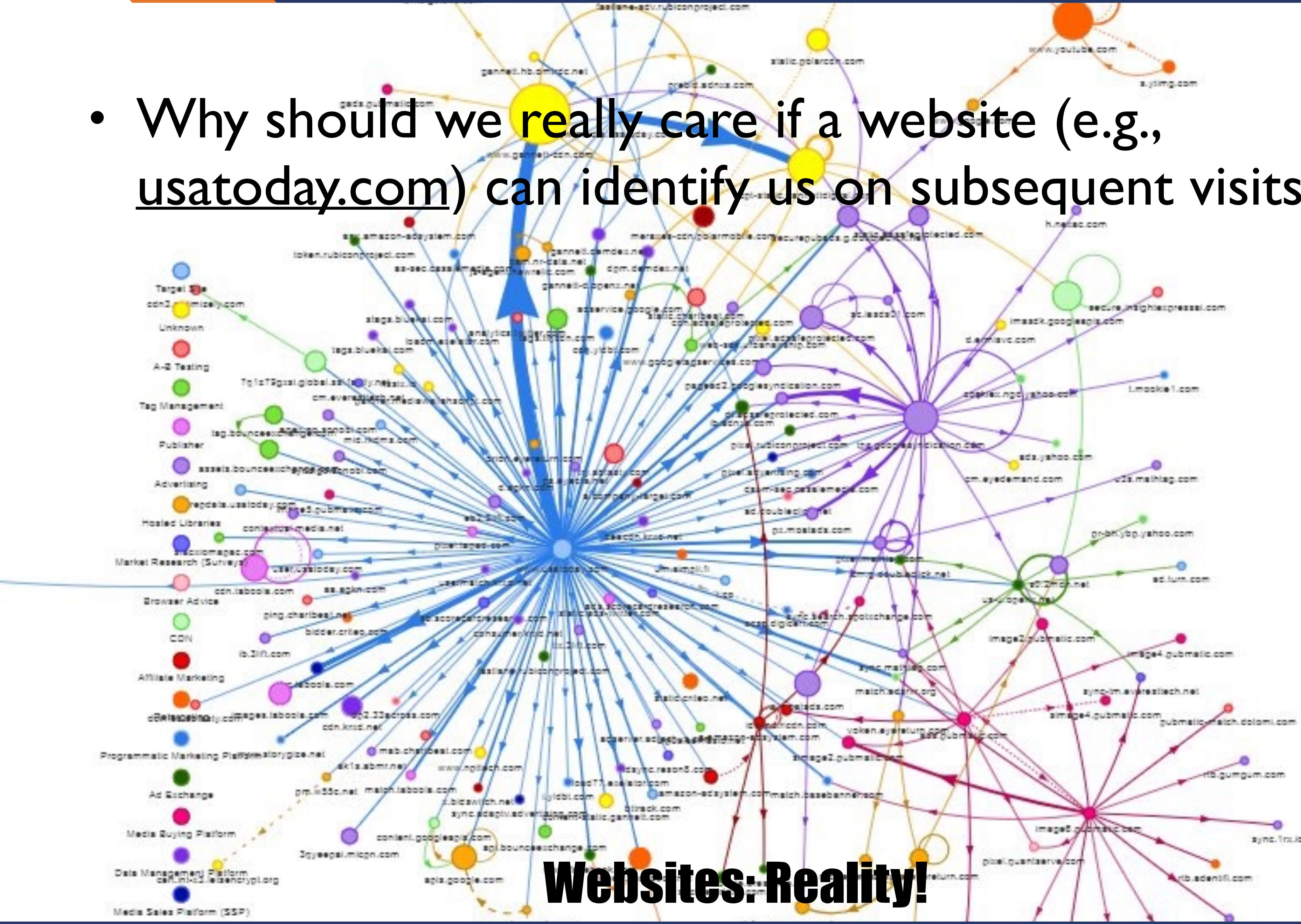
- Why should we really care if a website (e.g., usatoday.com) can identify us on subsequent visits?



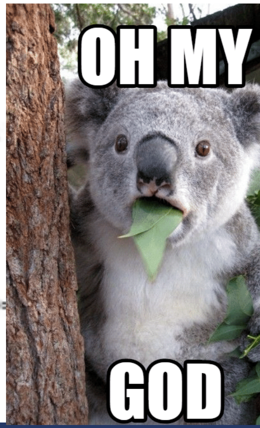
Websites: Expectation...

Aside: Who Cares?

- Why should we really care if a website (e.g., usatoday.com) can identify us on subsequent visits?

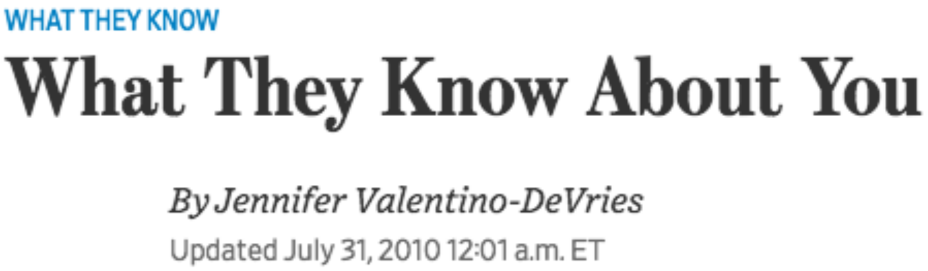


Websites: Reality!



Anti-Tracking Movement

- In 2010, more users were realizing the extent of the browser tracking problem…

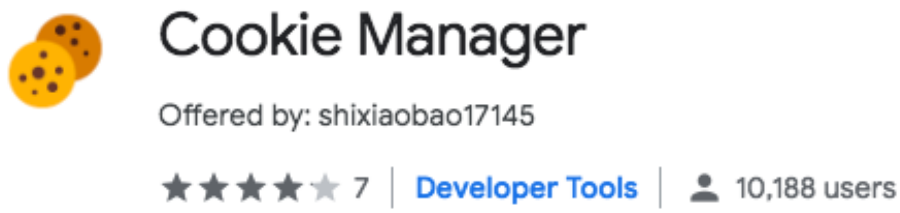


WHAT THEY KNOW

What They Know About You

By Jennifer Valentino-DeVries

Updated July 31, 2010 12:01 a.m. ET



Cookie Manager

Offered by: shixiaobao17145

★★★★☆ 7 | Developer Tools | 10,188 users

If we eradicated cookies from the Internet, would that solve the browser tracking problem?

Browser Fingerprinting

- An invisible, data-free form of browser tracking.
- Already appearing in advertising products back in 2010

41STPARAMETER
A part of Experian

- One instance of broader class of attacks against hardware and devices. You can basically fingerprint anything, and use anything to fingerprint:
  - Targets: Phones, Computers, Cameras, etc.
  - Signals: Accelerometer readings, packet arrivals, etc.

Browser Fingerprinting

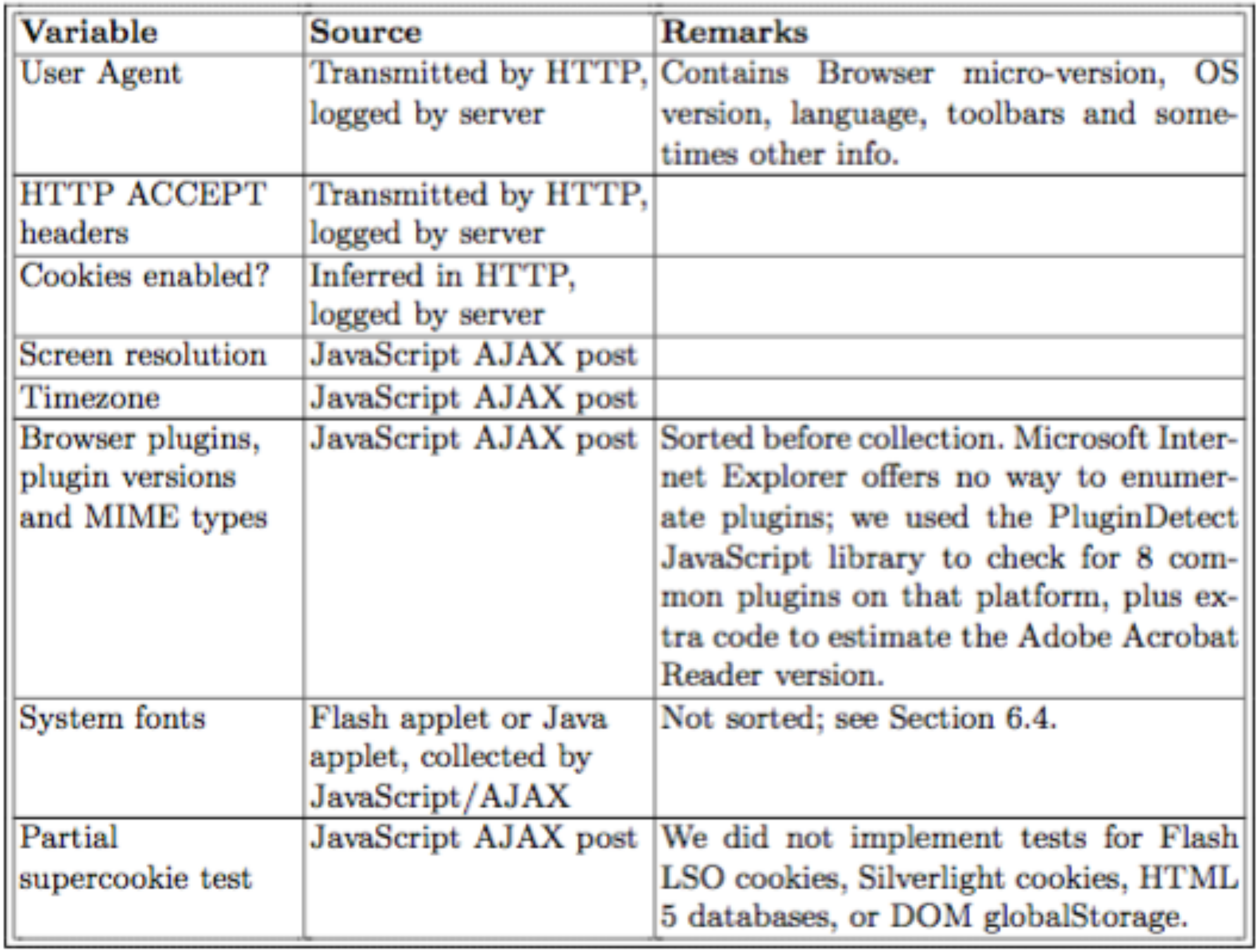
- Many possible applications for browser fingerprinting, albeit with varying levels of difficulty, including:
  - Fingerprints to differentiate NATed devices
  - Fingerprints to defeat Cookie Regenerators
  - Fingerprints at Global Identifiers

- What makes a given fingerprinting challenge easier or harder?

Enter Panoptoclick

- The EFF wanted to know how practical Internet-scale browser fingerprinting was.
- Since algorithms were proprietary, they made their own from various server-accessible browser attributes
- Invited people to visit panoptoclick.eff.org
- Analyzed entropy of resulting fingerprints to determine severity of the problem.

Panoptoclick Fingerprint



Variable	Source	Remarks
User Agent	Transmitted by HTTP, logged by server	Contains Browser micro-version, OS version, language, toolbars and sometimes other info.
HTTP ACCEPT headers	Transmitted by HTTP, logged by server	
Cookies enabled?	Inferred in HTTP, logged by server	
Screen resolution	JavaScript AJAX post	
Timezone	JavaScript AJAX post	
Browser plugins, plugin versions and MIME types	JavaScript AJAX post	Sorted before collection. Microsoft Internet Explorer offers no way to enumerate plugins; we used the PluginDetect JavaScript library to check for 8 common plugins on that platform, plus extra code to estimate the Adobe Acrobat Reader version.
System fonts	Flash applet or Java applet, collected by JavaScript/AJAX	Not sorted; see Section 6.4.
Partial supercookie test	JavaScript AJAX post	We did not implement tests for Flash LSO cookies, Silverlight cookies, HTML 5 databases, or DOM globalStorage.

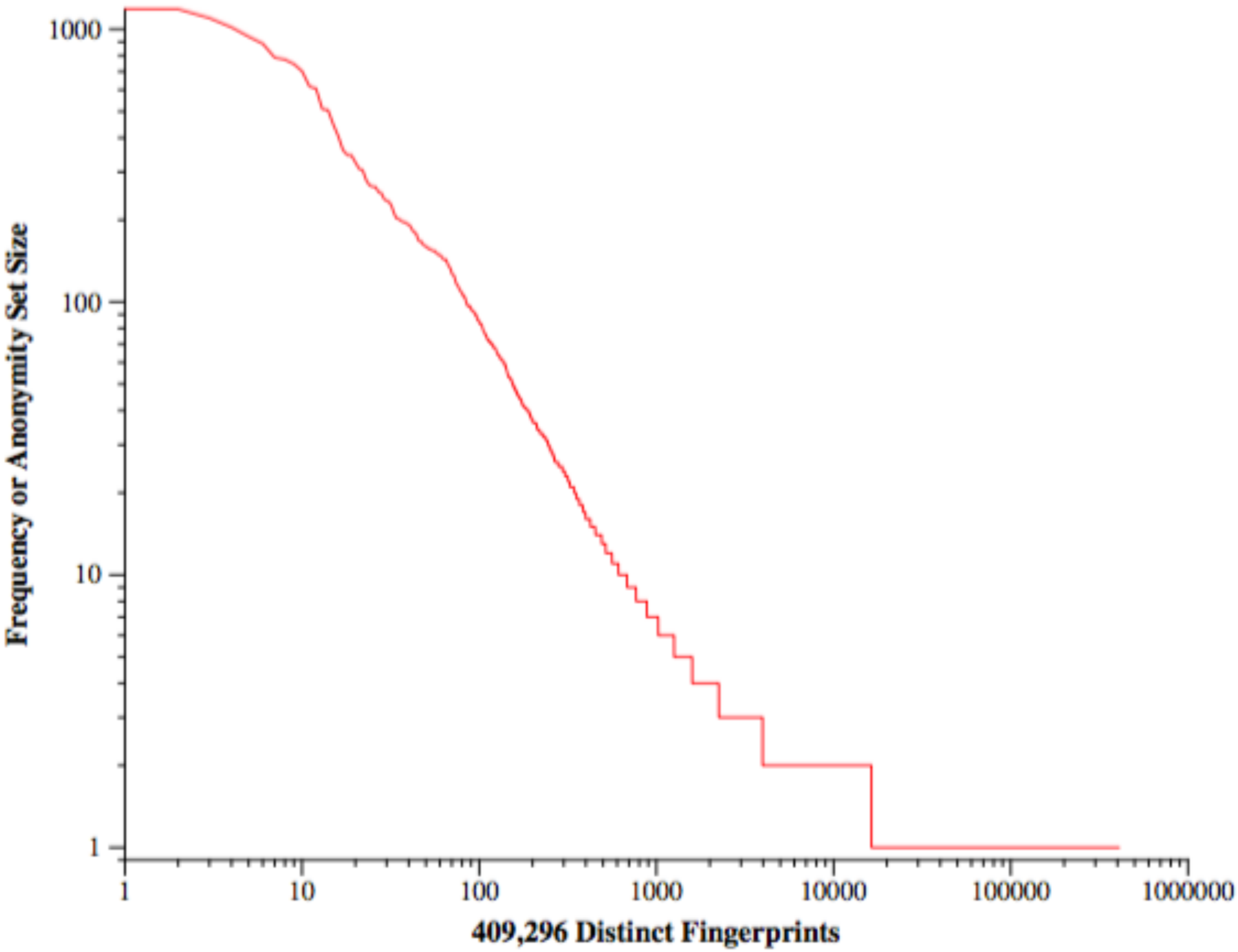
Note: Plenty of unharvested info, such as ActiveX, Silverlight, etc.

Panoptoclick Analysis

- Each feature is associated with a distribution related to Self-Information / Surprisal / Entropy (related ideas)
- I.E., how much do we learn about an object when one of its random variable(s) is sampled?
  - Each bit of information cuts space of objects in half
- Combine multiple features together, adjusting for the fact that the variables won't all be independent.
- Your browser is uniquely identifiable if the number of bits of information gained from its features is greater than the (logarithm of) the number of browsers in "the world"

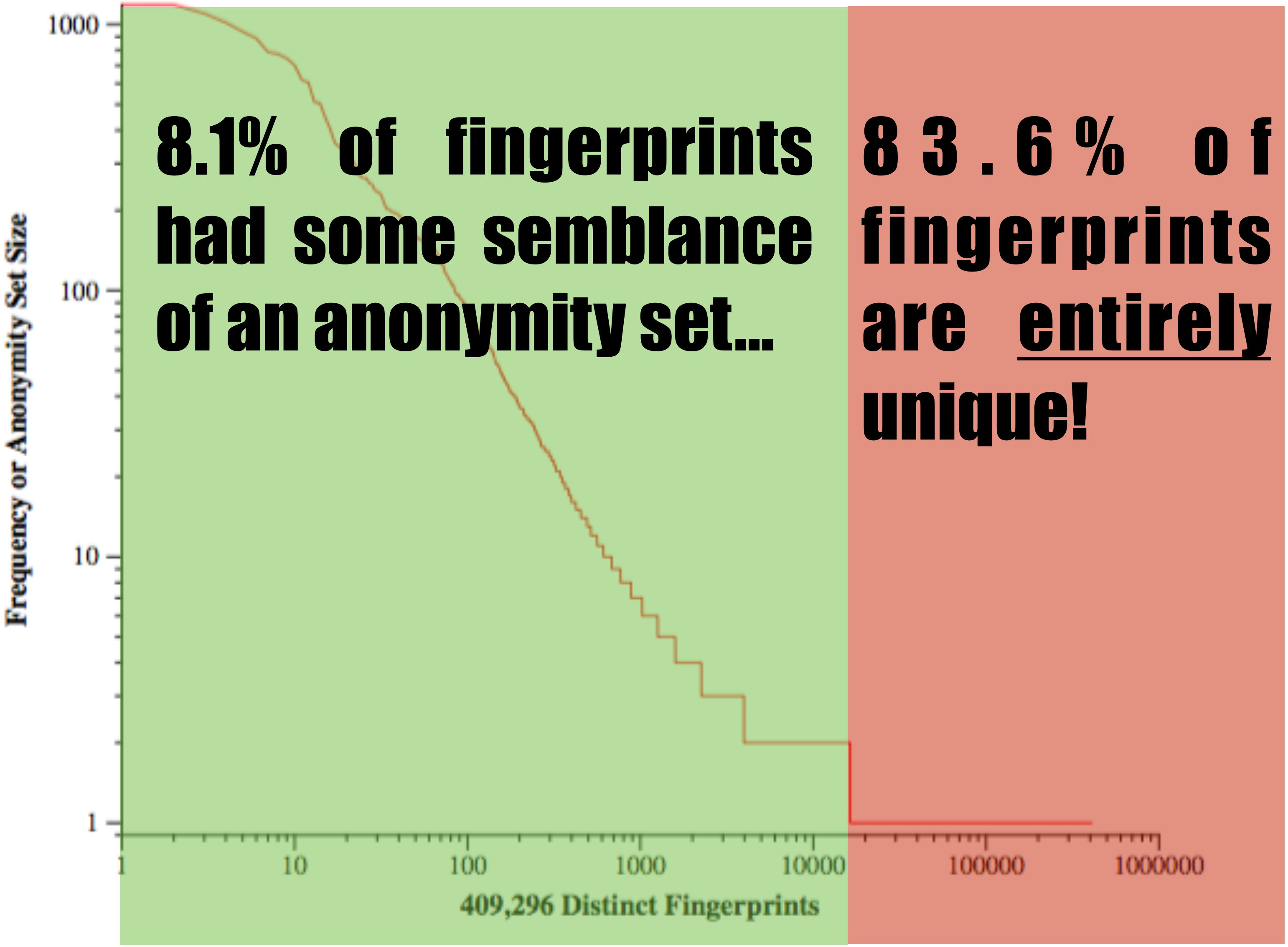
Panoptoclick Results

Of ~470,000 fingerprint instances collected…



Panoptoclick Results

Of ~470,000 fingerprint instances collected…

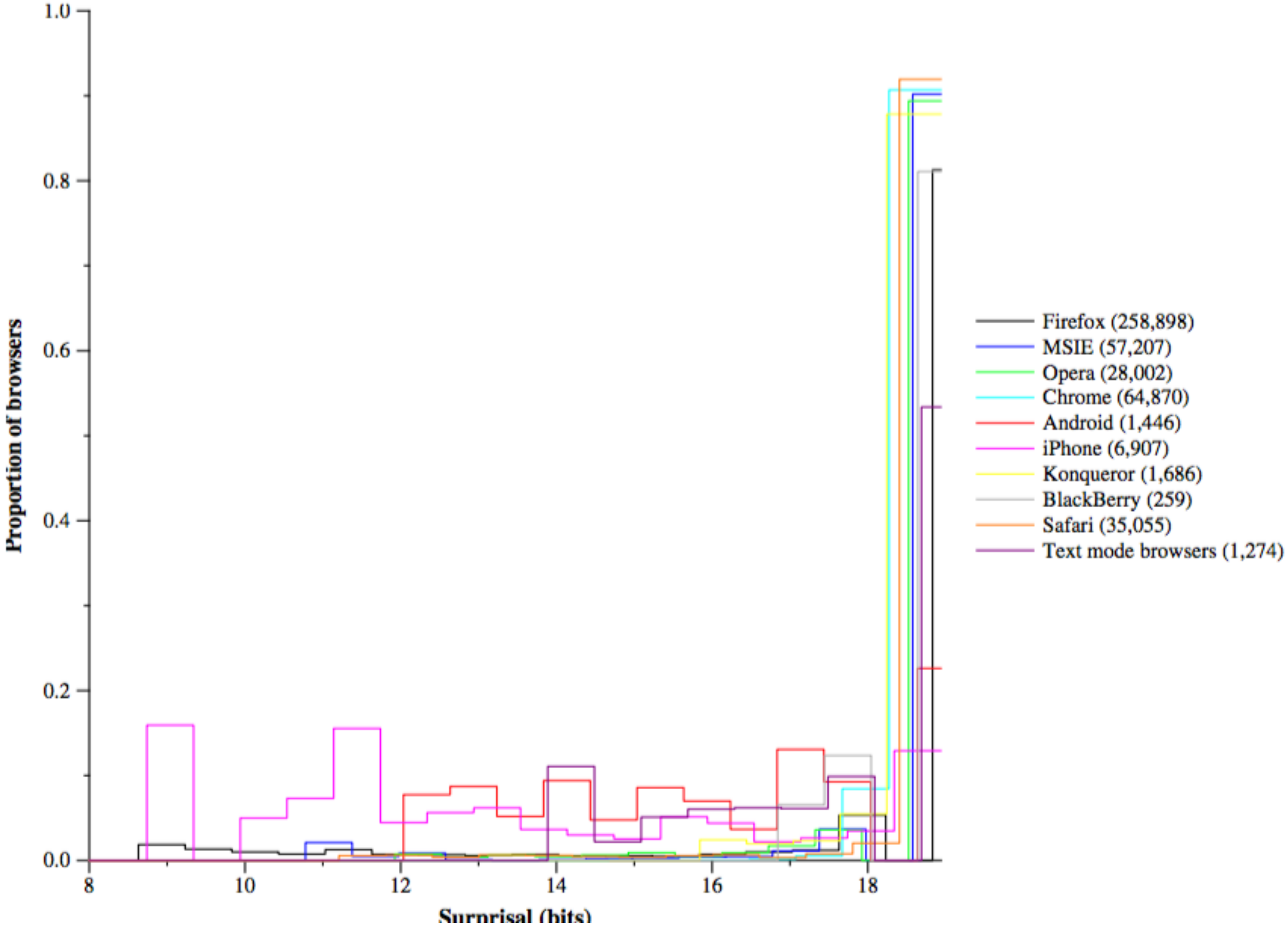


8.1% of fingerprints had some semblance of an anonymity set…

83.6% of fingerprints are entirely unique!

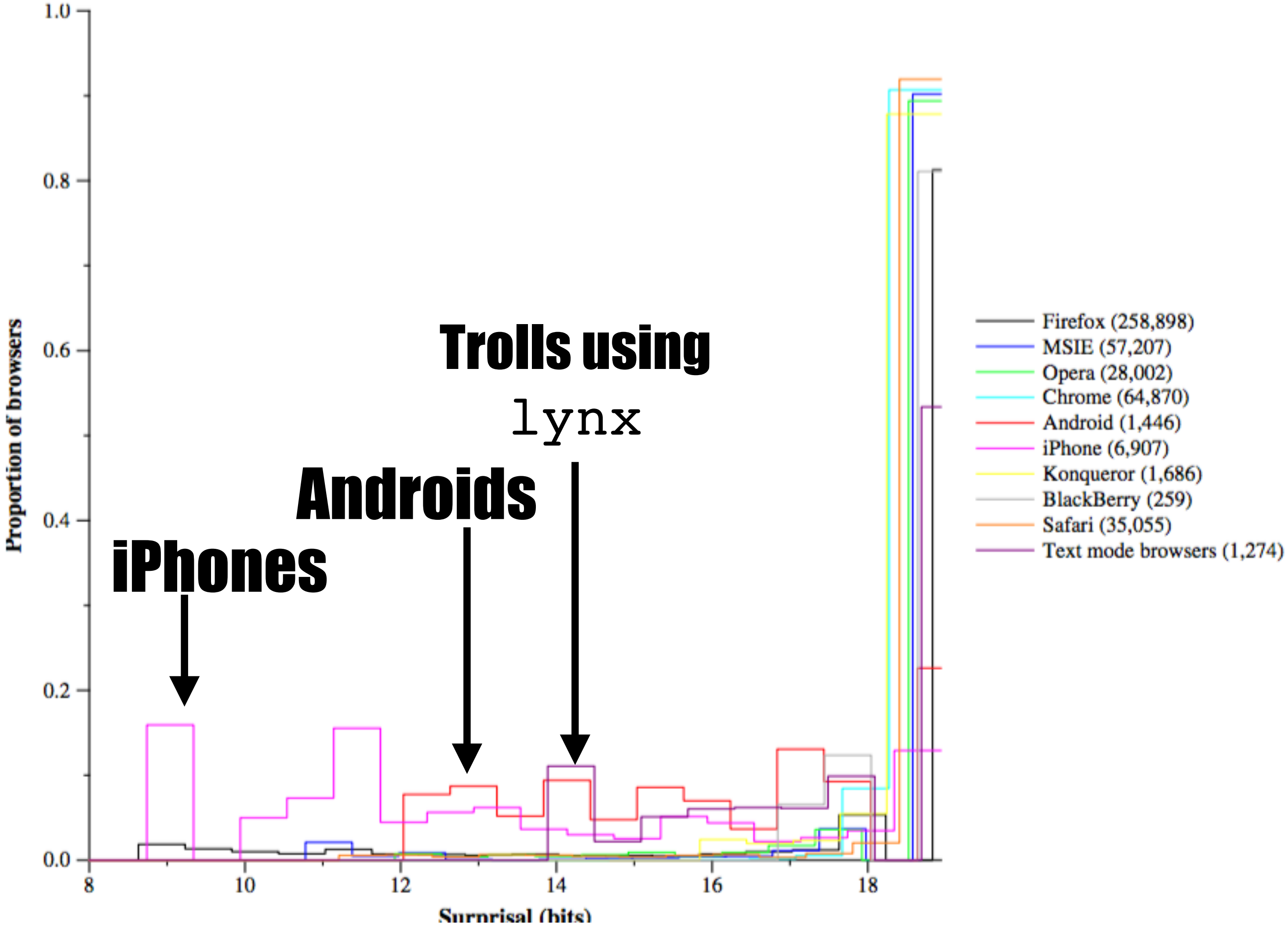
Panoptoclick Results

Where did Panoptoclick struggle?



Panoptoclick Results

Where did Panoptoclick struggle?



Trolls using `lynx`

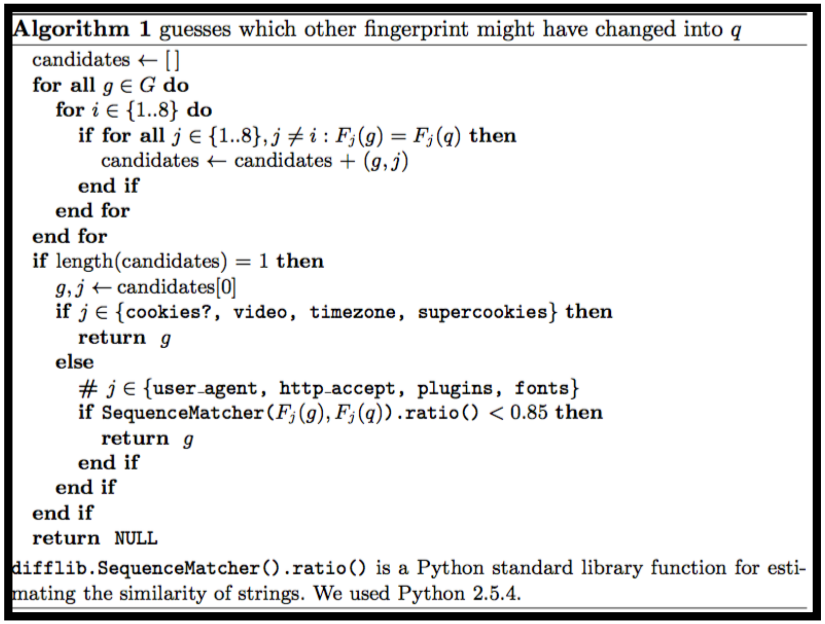
Androids

iPhones

Panoptoclick Results

Are browser fingerprints consistent?

- No! 37.4% churn
- But, probably over-reported given the EFF's clientele…
- Worse, even a crude algorithm can guess the link between two fingerprints 65% of the time (w/ 0.9% FP).



Algorithm 1 guesses which other fingerprint might have changed into q

```
candidates ← []
for all $g \in G$ do
  for $i \in \{1..8\}$ do
    if for all $j \in \{1..8\}, j \neq i : F_j(g) = F_j(q)$ then
      candidates ← candidates + $(g, j)$
    end if
  end for
end for
if length(candidates) = 1 then
  $g, j$ ← candidates[0]
  if $j \in$ {cookies?, video, timezone, supercookies} then
    return $g$
  else
    # $j \in$ {user_agent, http_accept, plugins, fonts}
    if SequenceMatcher($F_j(g), F_j(q)$).ratio() $< 0.85$ then
      return $g$
    end if
  end if
end if
return NULL
```

difflib.SequenceMatcher().ratio() is a Python standard library function for estimating the similarity of strings. We used Python 2.5.4.

Additional Observations

- The presence of Privacy Enhancing Technologies (e.g., anonymity plug-ins) often decreased anonymity set!!
  - Why?
- APIs frequently offer the ability to enumerate system information. Testable APIs would increase difficulty of fingerprinting.
- Tension between ease of debugging and difficulty of fingerprinting (e.g., fine-grained version numbers)
- Tension between expressivity of browser config and difficulty of fingerprinting (e.g., font orders)

Location Privacy

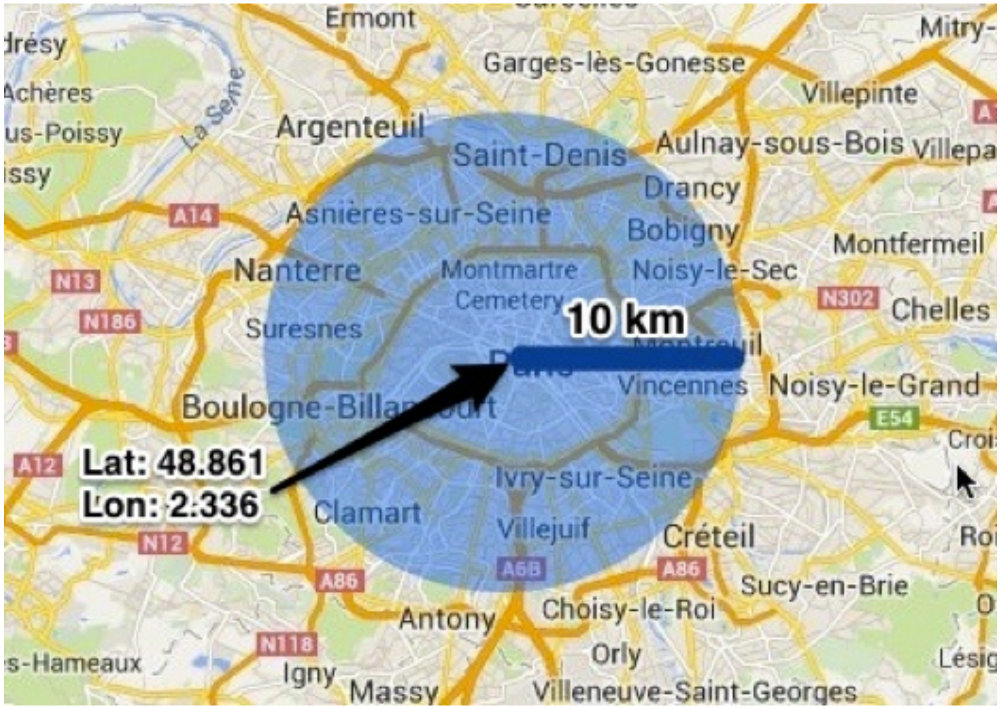
- Today, the world is lousy with location-based services (LBS), e.g., …

G
The Weather Channel
yelp
lyft

- Coarse-grained LBS: weather, advertising, events in area
- Fine-grained LBS: navigation, ride share, fitness tracking
- Untrustworthy LBS could make sensitive inferences about our identity, of even harm us in the real world!
- How can we use LBS without revealing our location?

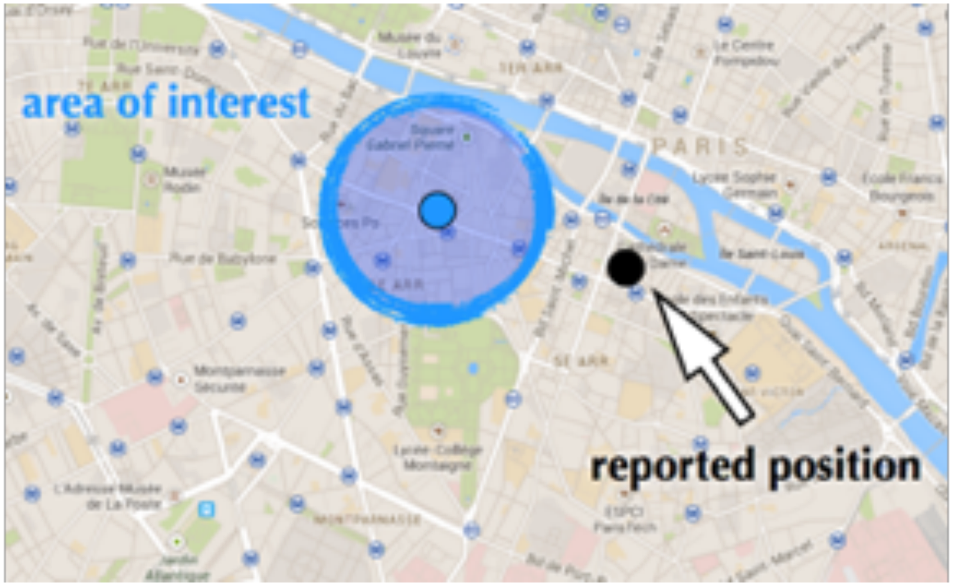
Geo-Indistinguishability (GI)

- On device, add controlled noise to user's location before sharing with LBS.
- Achieves quasi-indistinguishability within a given area
- Generalization of differential privacy for an arbitrary distance function.



"User is equally likely to be anywhere within radius r of the Eiffel Tower"

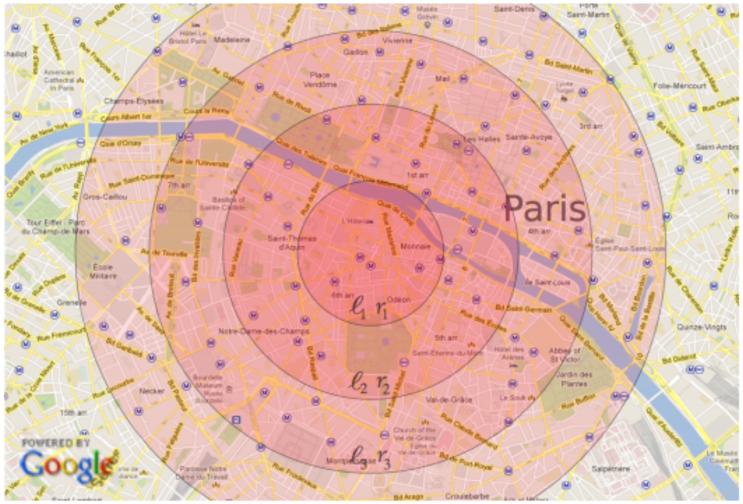
Geo-Indistinguishability (GI)



How does GI work?

- User is at location x
- User specifies radius r, level of similarity λ
- User reports some point z based on x, r, λ

Geo-Indistinguishability (GI)

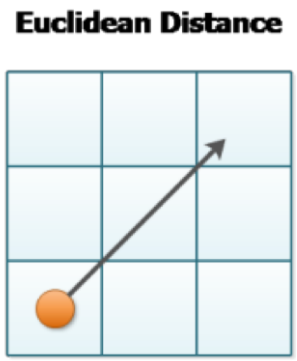


Properties of GI

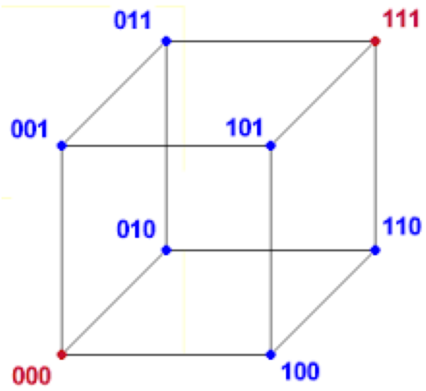
- What is point z?
  - Each point within one unit of distance within the region specified by ε is *equally likely* to be returned
- Privacy level ε is the radio of λ to r
  - If r is small, λ must be large to have high ε
  - If r is large, λ can be smaller to have high ε
  - If we fix λ and increase r, ε is greater but results are inaccurate.

compare to Differential Privacy (DP)?

- Similar to DP, GI is independent from side information of the attacker (no assumptions made about priors)
- GI uses euclidean distance instead of hamming distance
  - Euclidean Distance: spatial or linear distance between two points
  - Hamming Distance: distance between two datasets

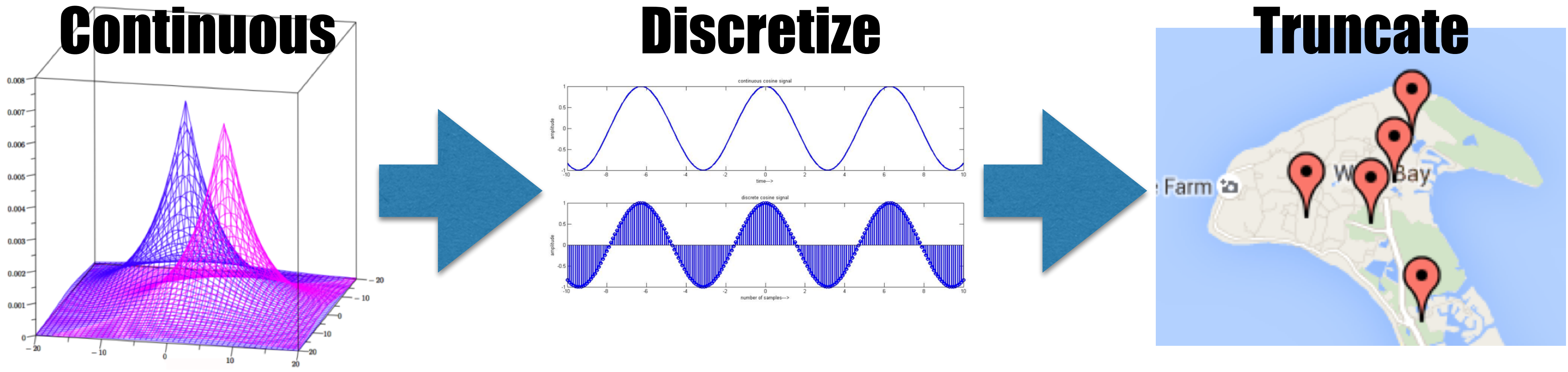


$$\sqrt{(x_1 - x_2)^2 + (y_1 - y_2)^2}$$



GI Algorithm

- Perturbate input by noise generated from Laplace distribution, yielding a probability density function from which we choose a random point.
- Map random point from the continuous domain to the nearest point in discrete domain (i.e., Lat, Long)
- Eliminate unrealistic points based based on map data



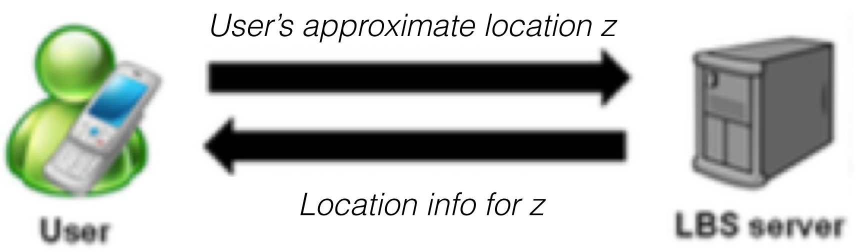
Continuous

Discretize

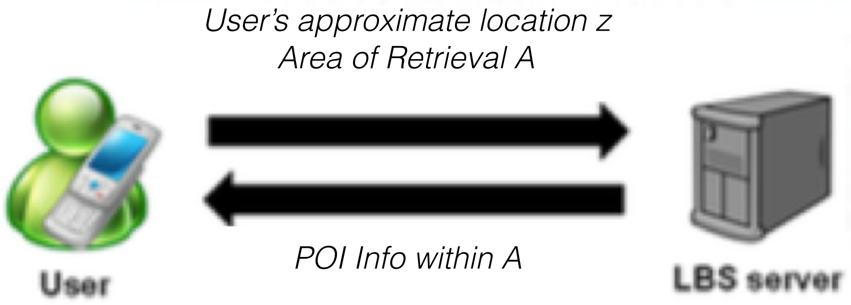
Truncate

Enhancing LBS

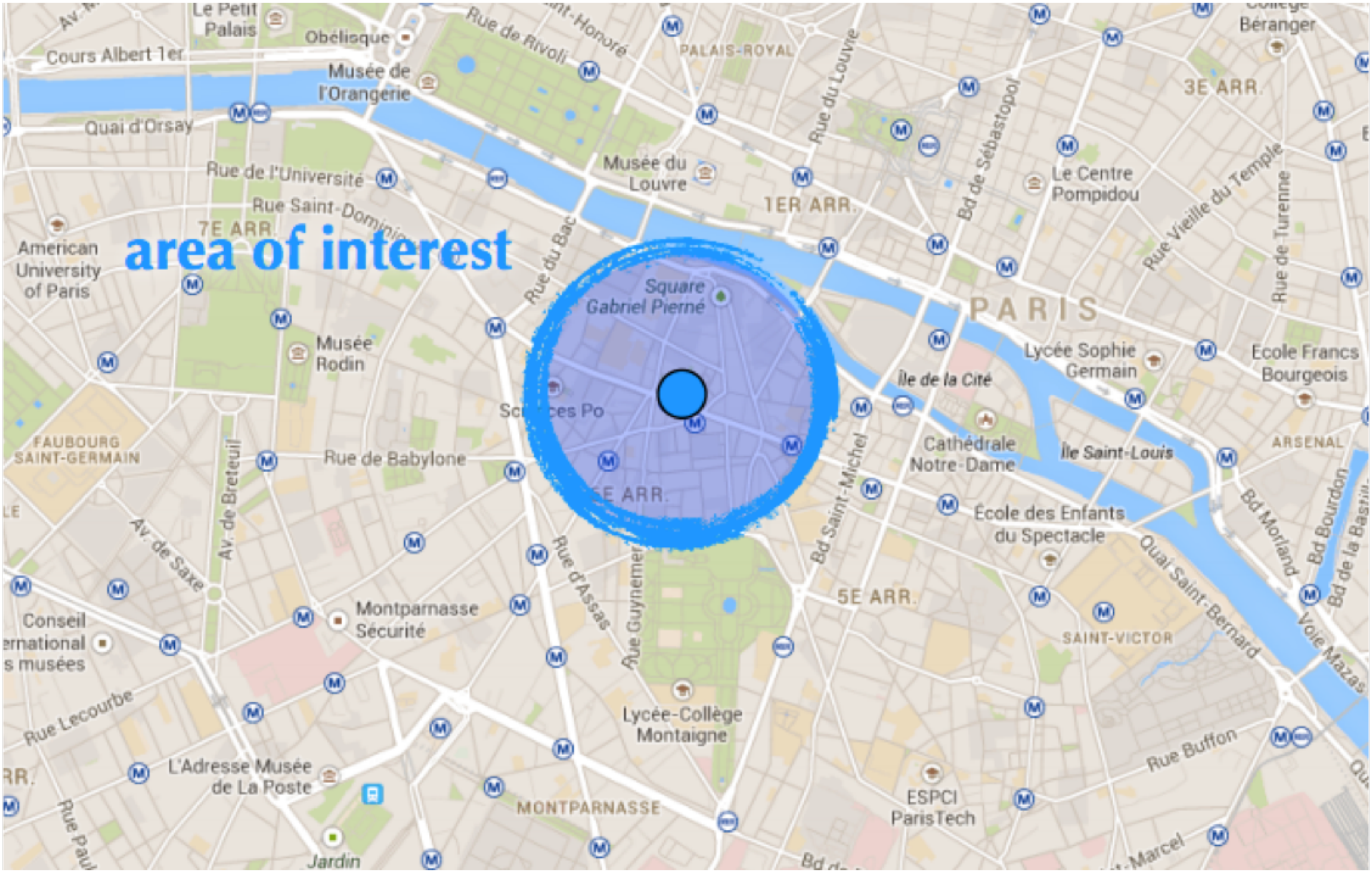
- Coarse-grained LBS: apply stock geo-indistinguishability



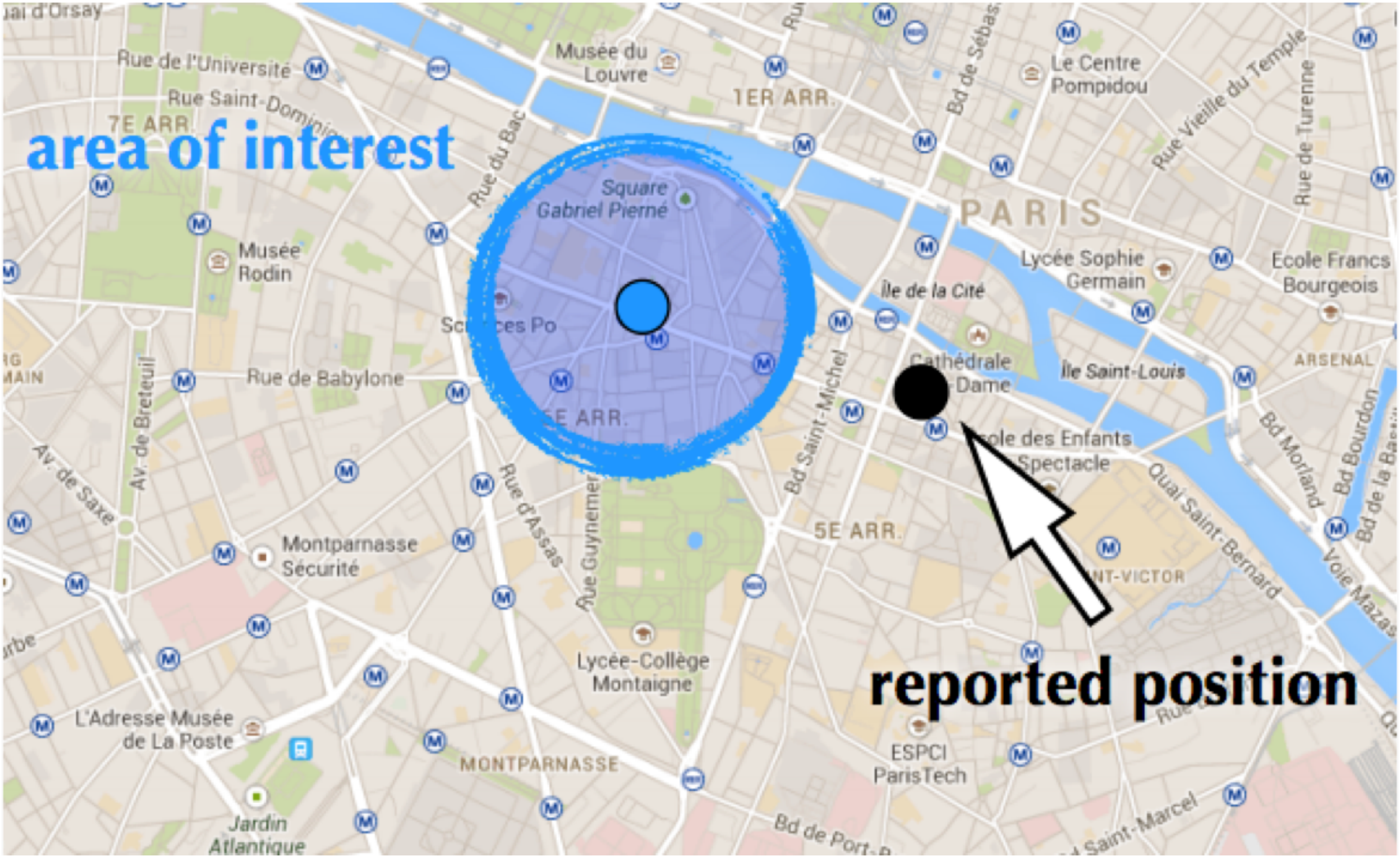
- Fine-grained LBS: Geo-Indistinguishability may be inadequate, instead specify larger area of retrieval based on z:



Fine-Grained LBS w/ GI



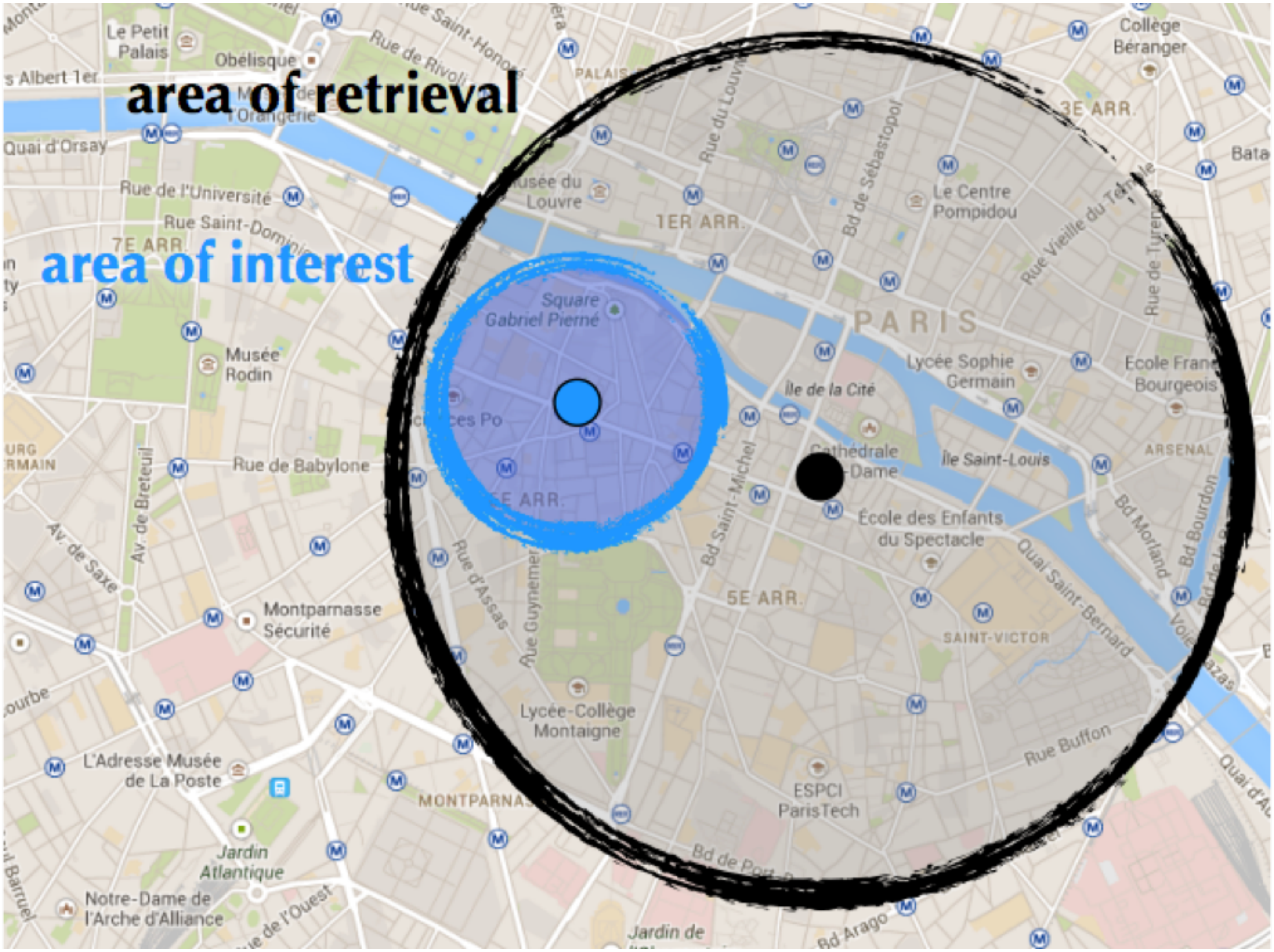
Fine-Grained LBS w/ GI



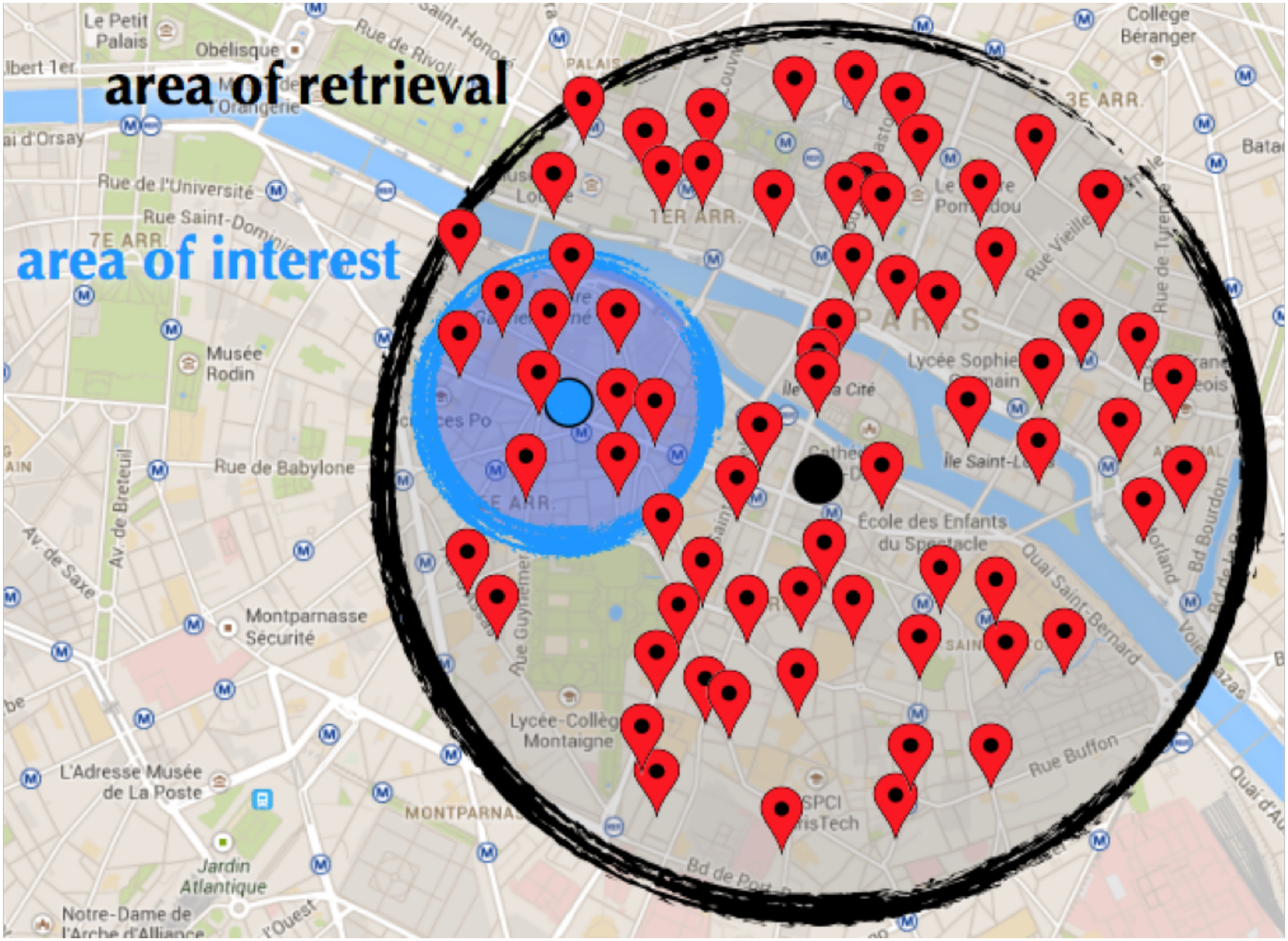
area of interest

reported position

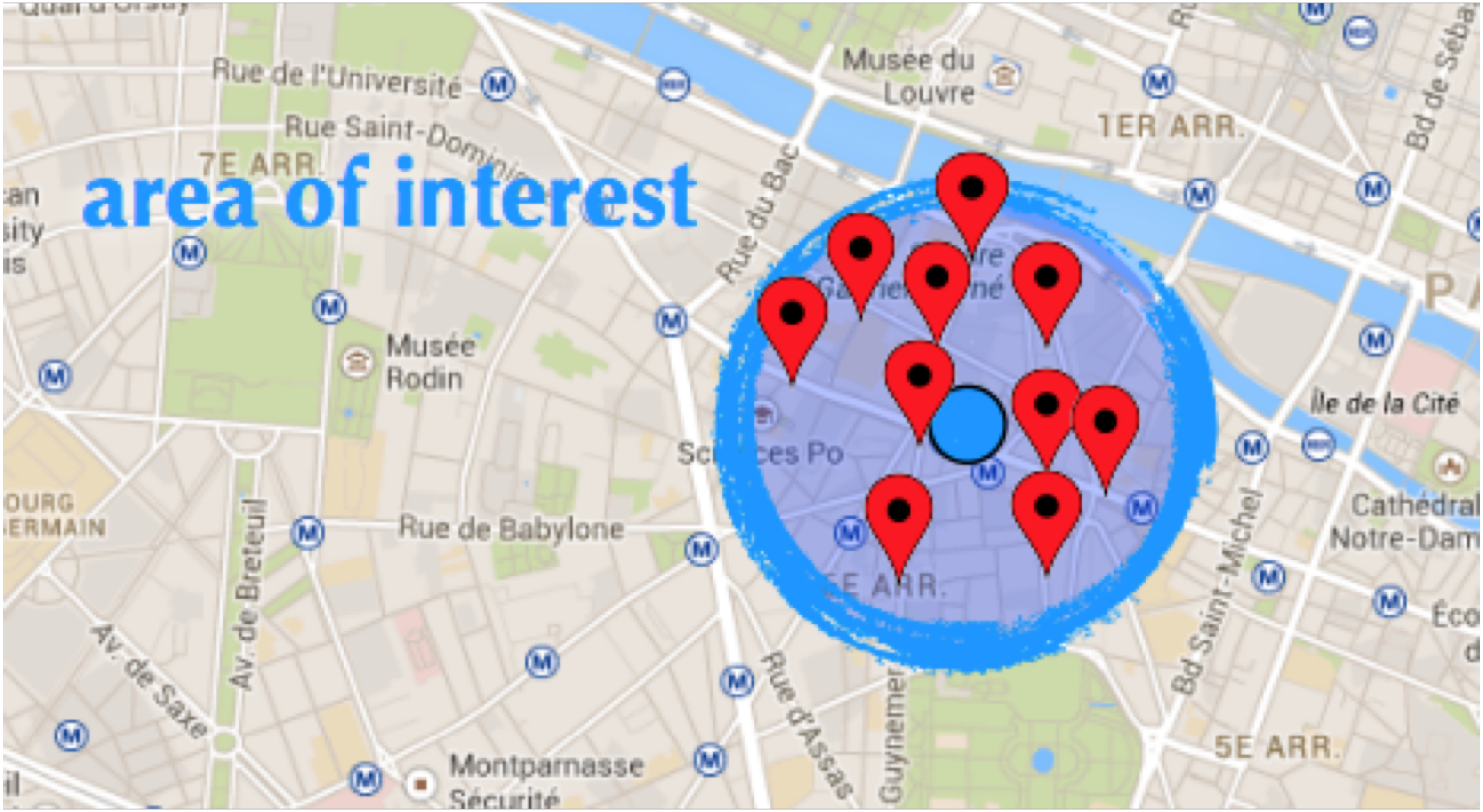
Fine-Grained LBS w/ GI



Fine-Grained LBS w/ GI



Fine-Grained LBS w/ GI



Case Study: U.S. Census

- The Census Bureau contains information in the form of (hBlock, wBlock)
  - hBlock—where the worker lives
  - wBlock—where the worker works
- Takes each point of the census data and randomizes it according to specified values of l and r

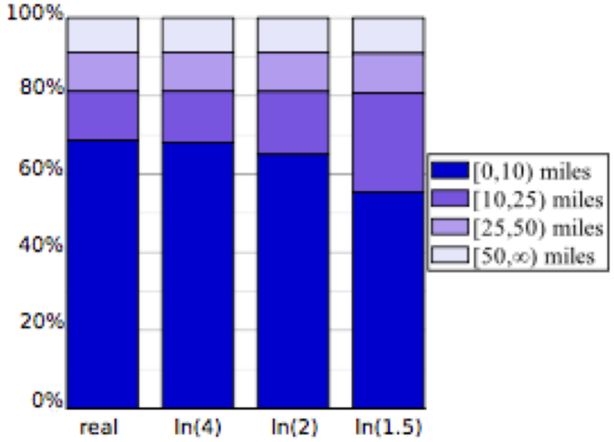
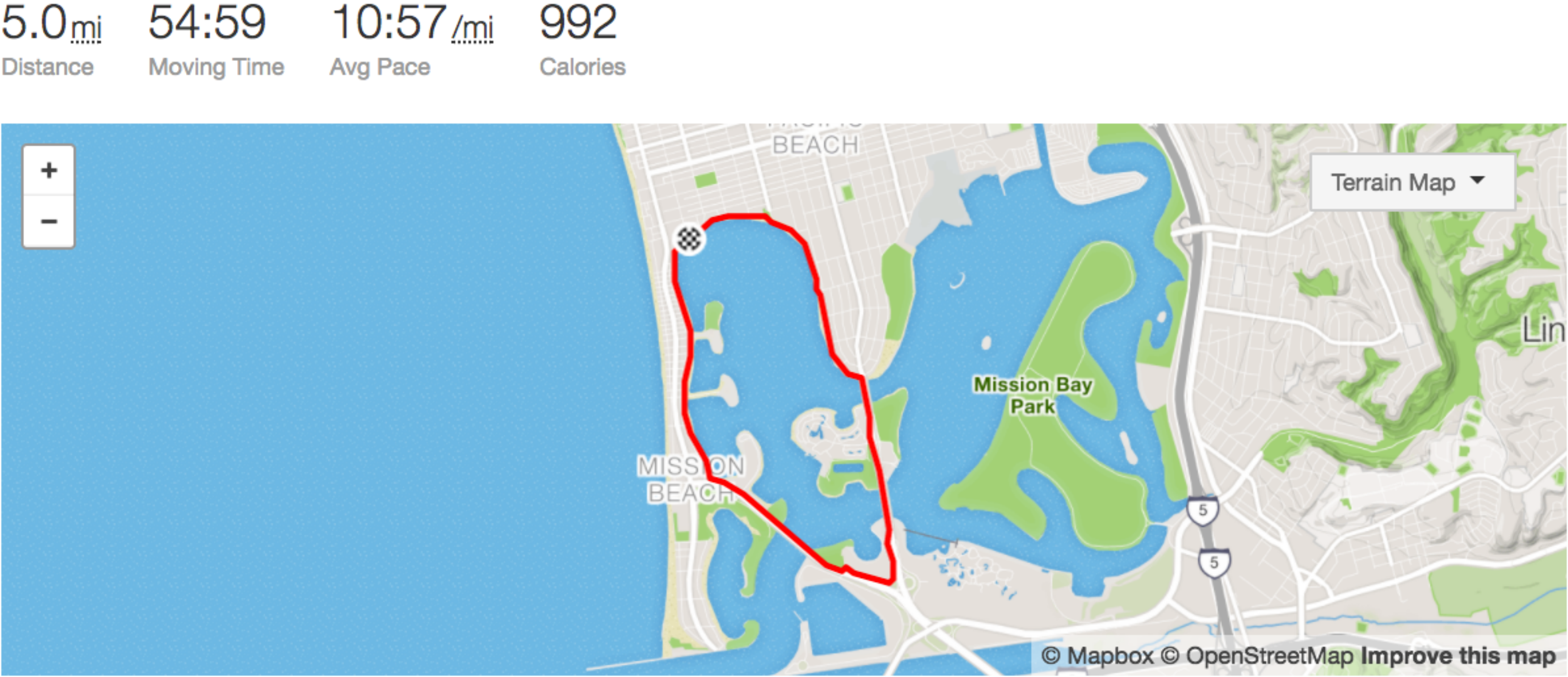


Figure 13. Home-work commute distance for $r = 1.22$ and various ℓ.

End-of-Talk Palette Cleanser…

End-of-Talk Palette Cleanser…

Endpoint Privacy Zones…

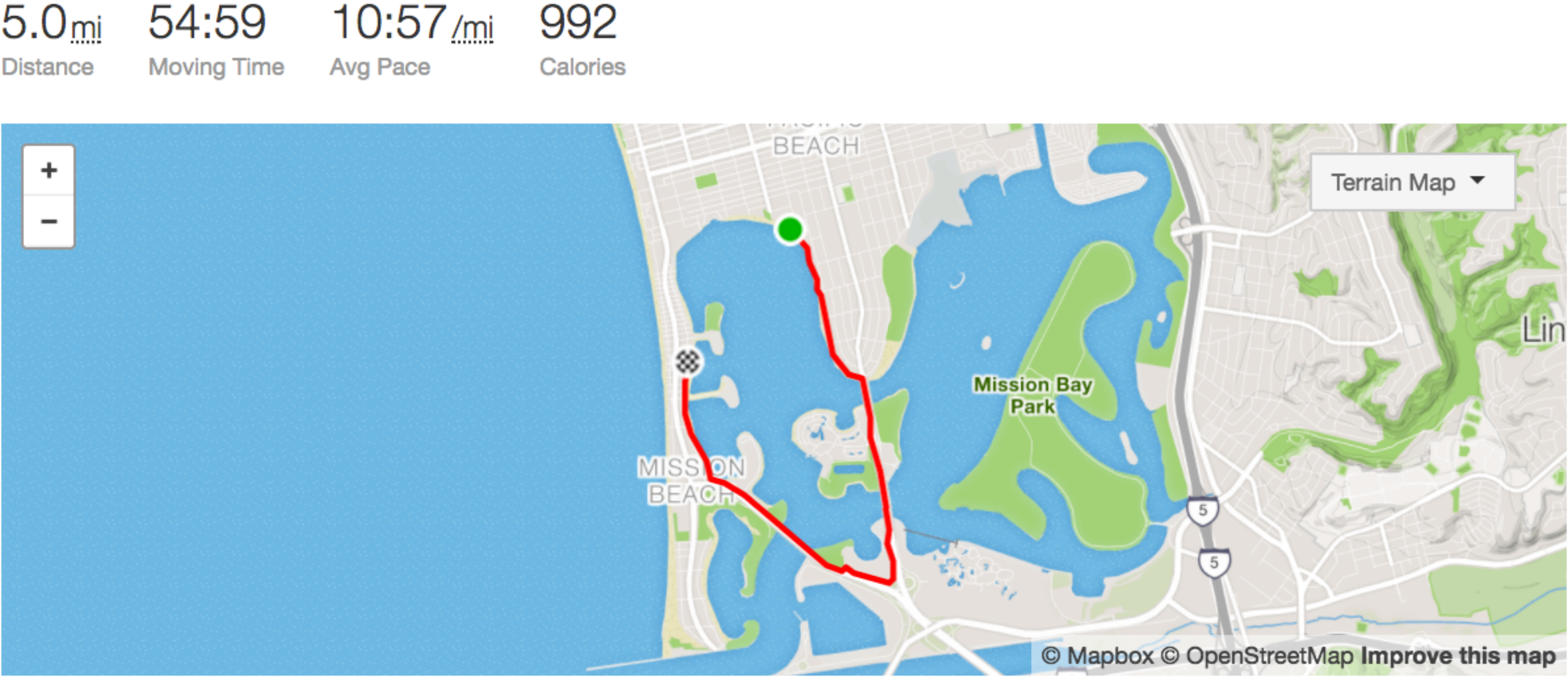


5.0mi	54:59	10:57/mi	992
Distance	Moving Time	Avg Pace	Calories

STRAVA™

End-of-Talk Palette Cleanser…

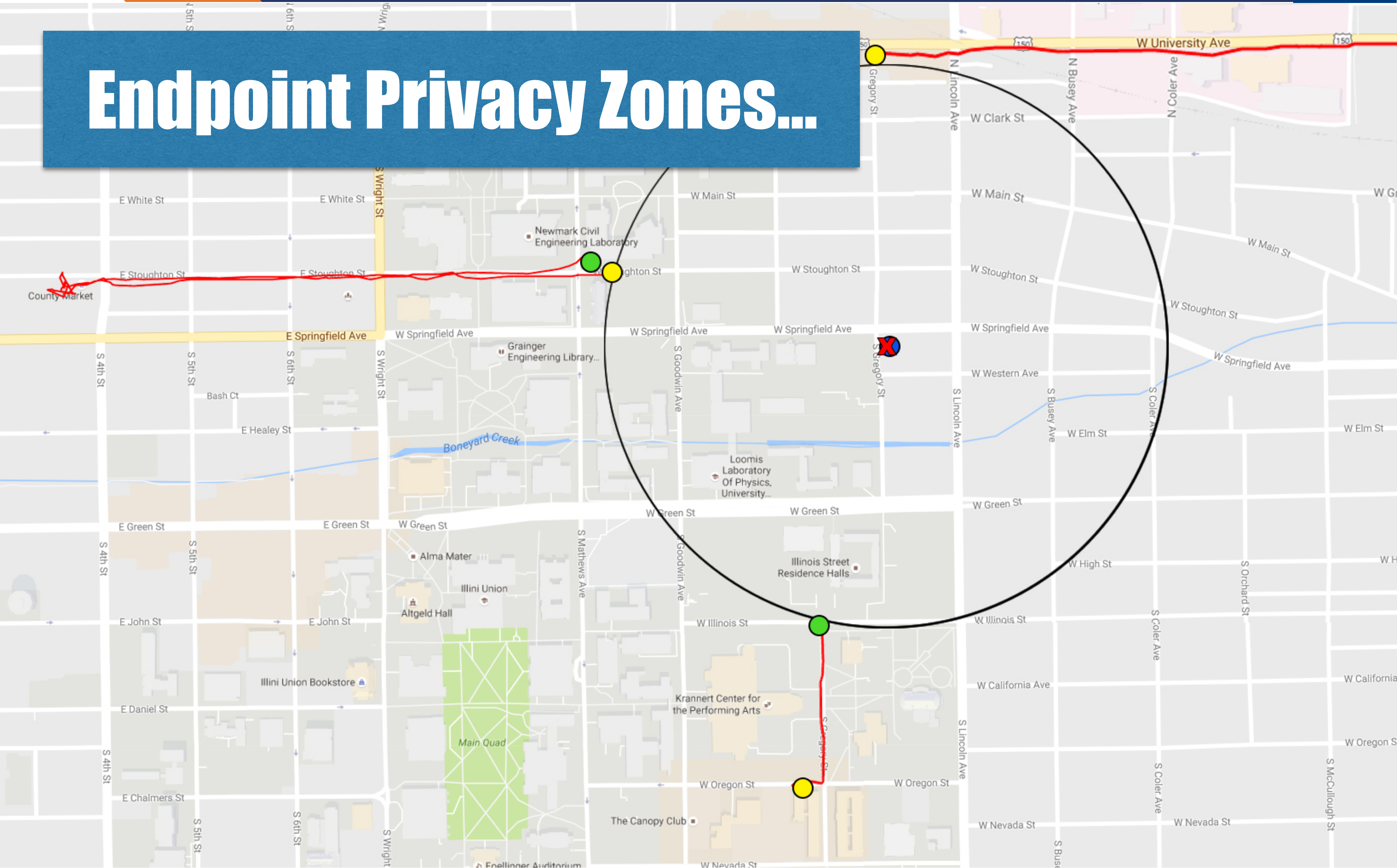
Endpoint Privacy Zones...



5.0mi	54:59	10:57/mi	992
Distance	Moving Time	Avg Pace	Calories

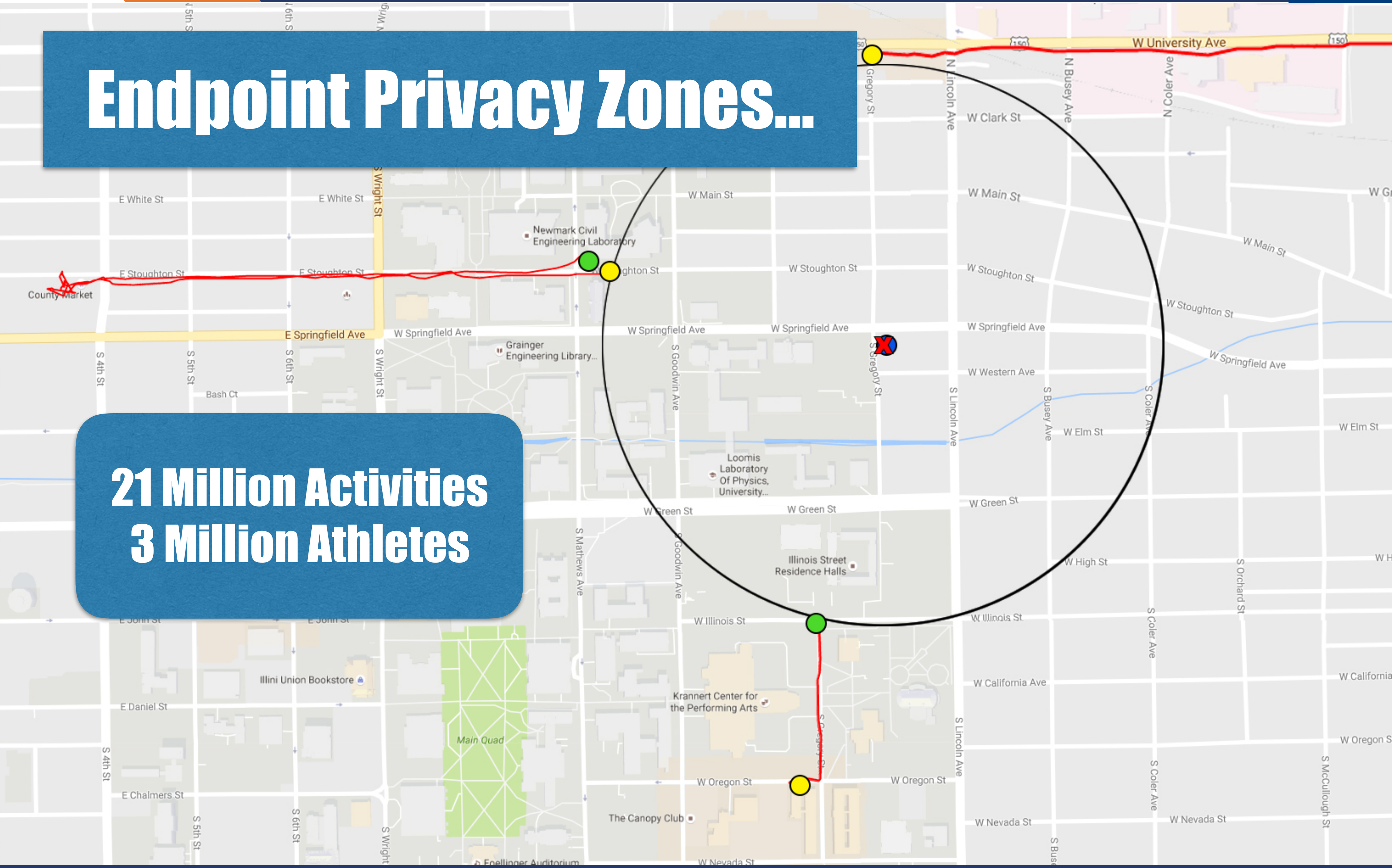
STRAVA™

End-of-Talk Palette Cleanser…



Endpoint Privacy Zones…

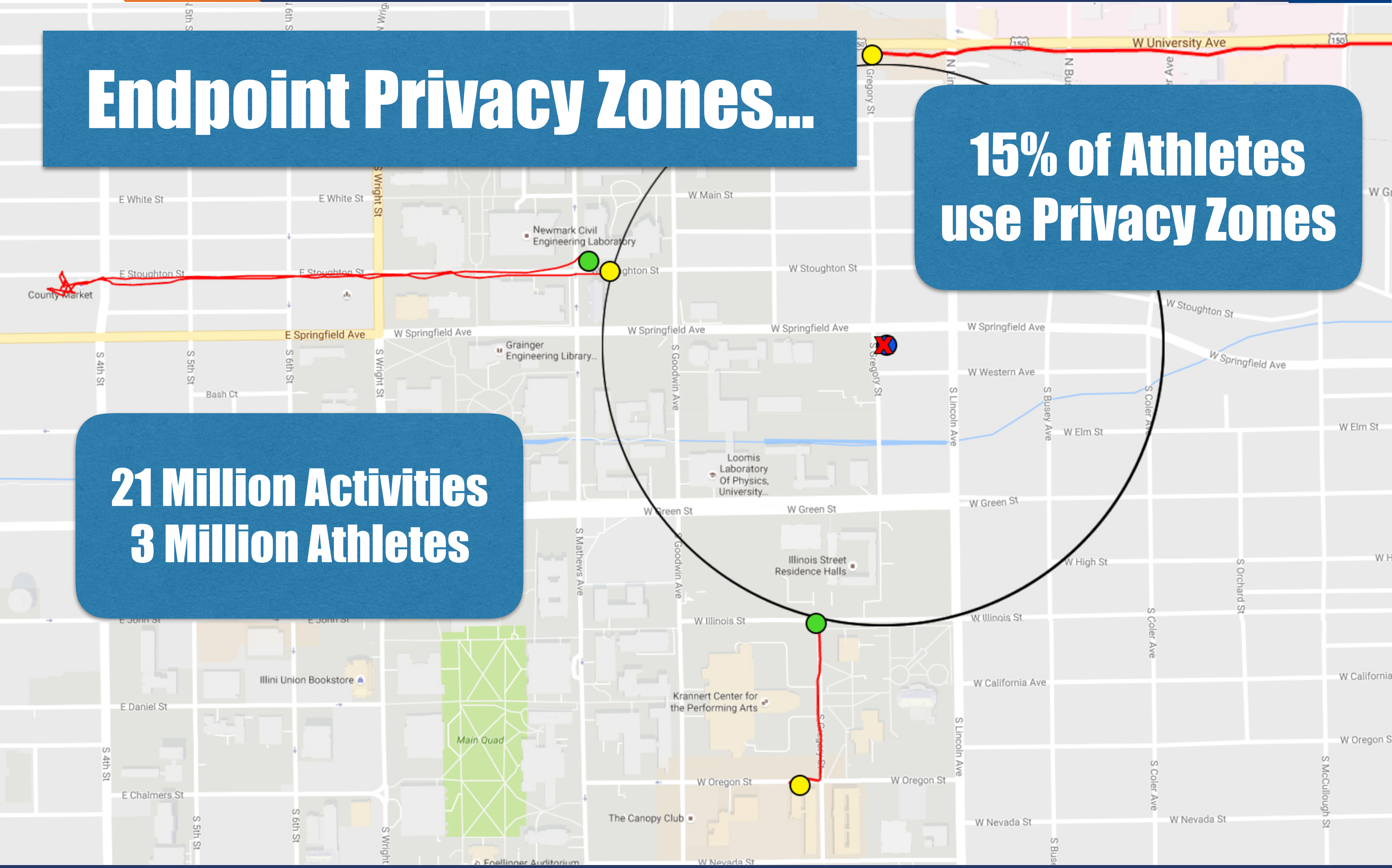
End-of-Talk Palette Cleanser…



Endpoint Privacy Zones...

21 Million Activities
3 Million Athletes

End-of-Talk Palette Cleanser…

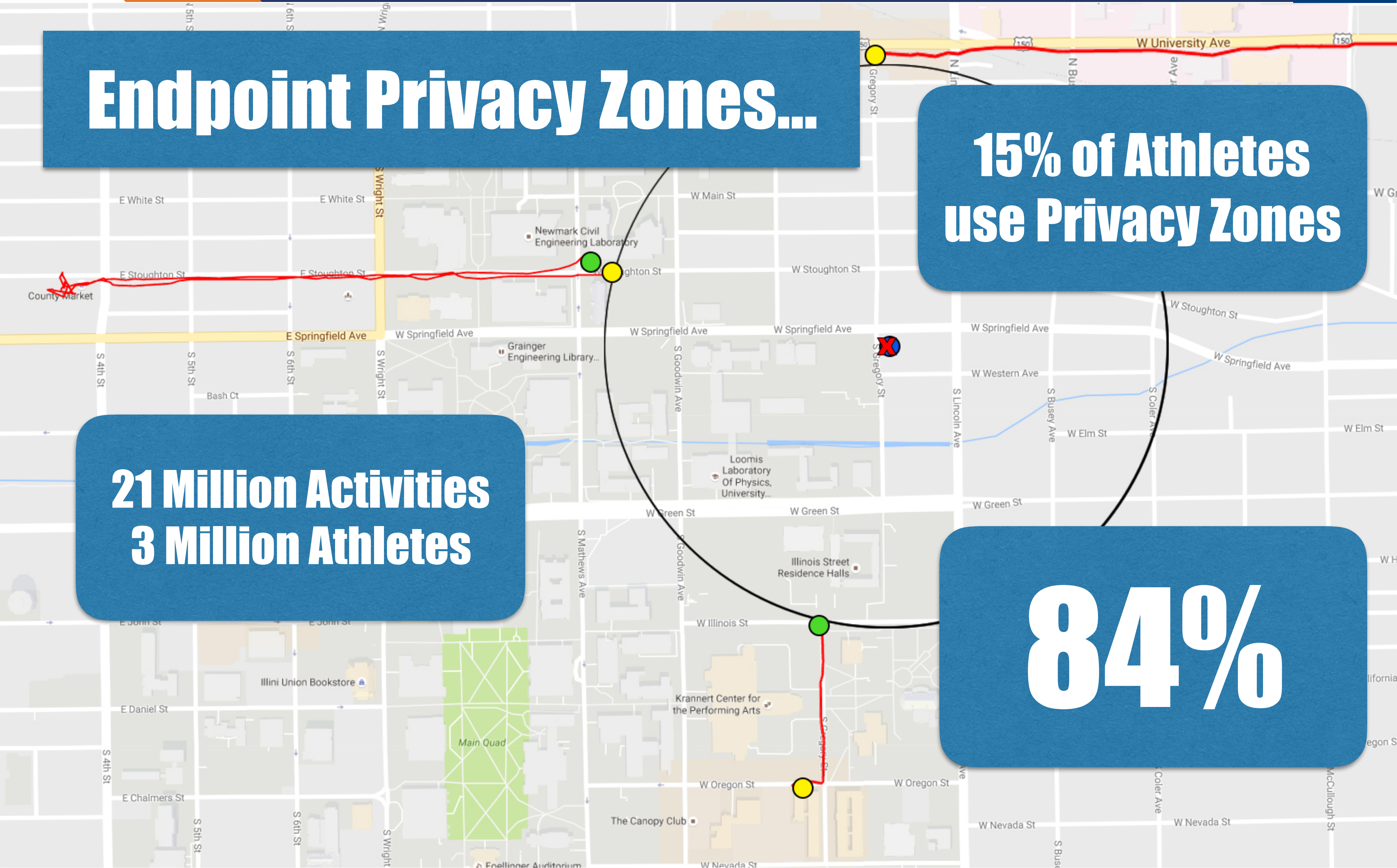


Endpoint Privacy Zones…

15% of Athletes use Privacy Zones

21 Million Activities
3 Million Athletes

End-of-Talk Palette Cleanser…



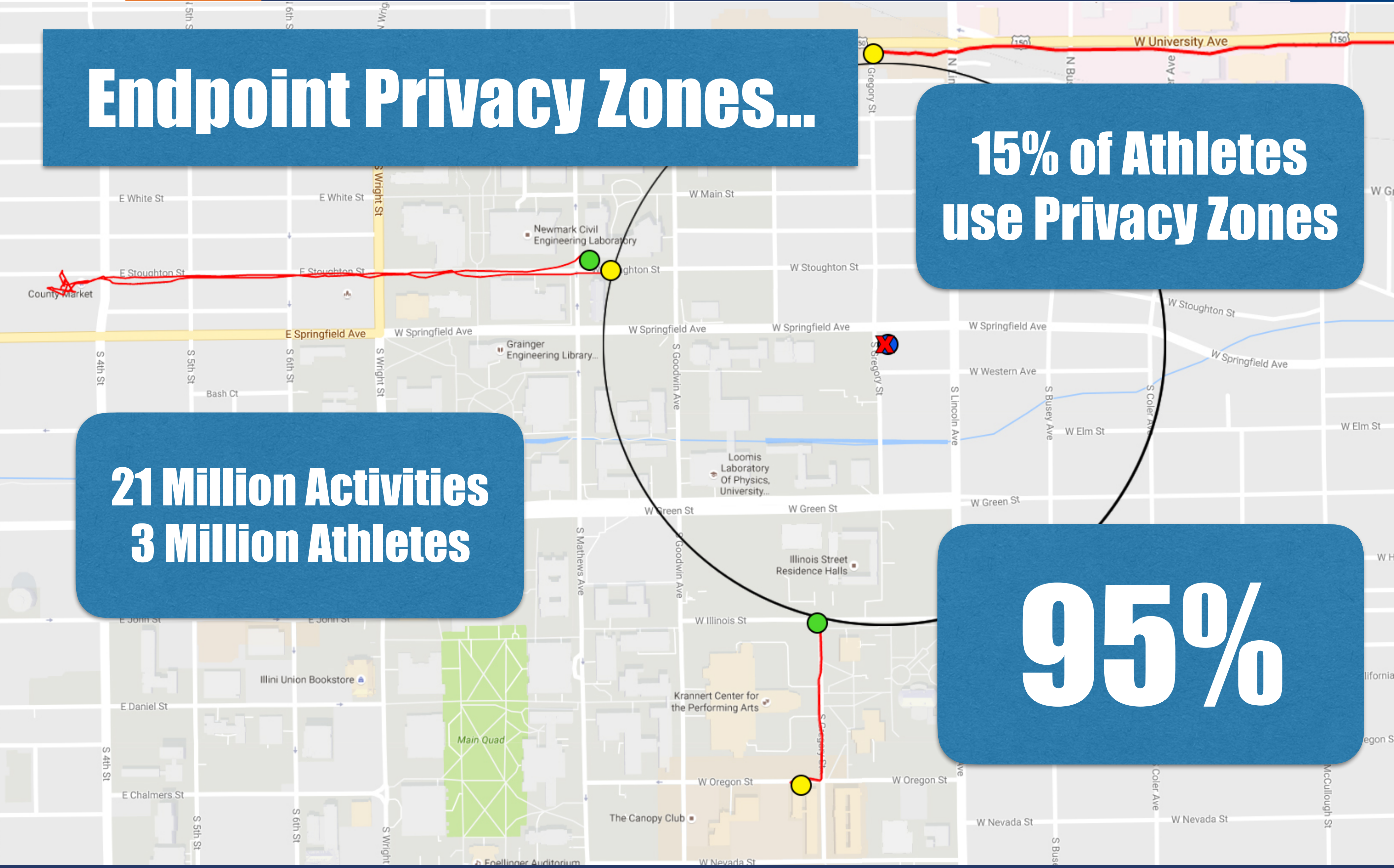
Endpoint Privacy Zones…

15% of Athletes use Privacy Zones

21 Million Activities
3 Million Athletes

84%

End-of-Talk Palette Cleanser…



Endpoint Privacy Zones…

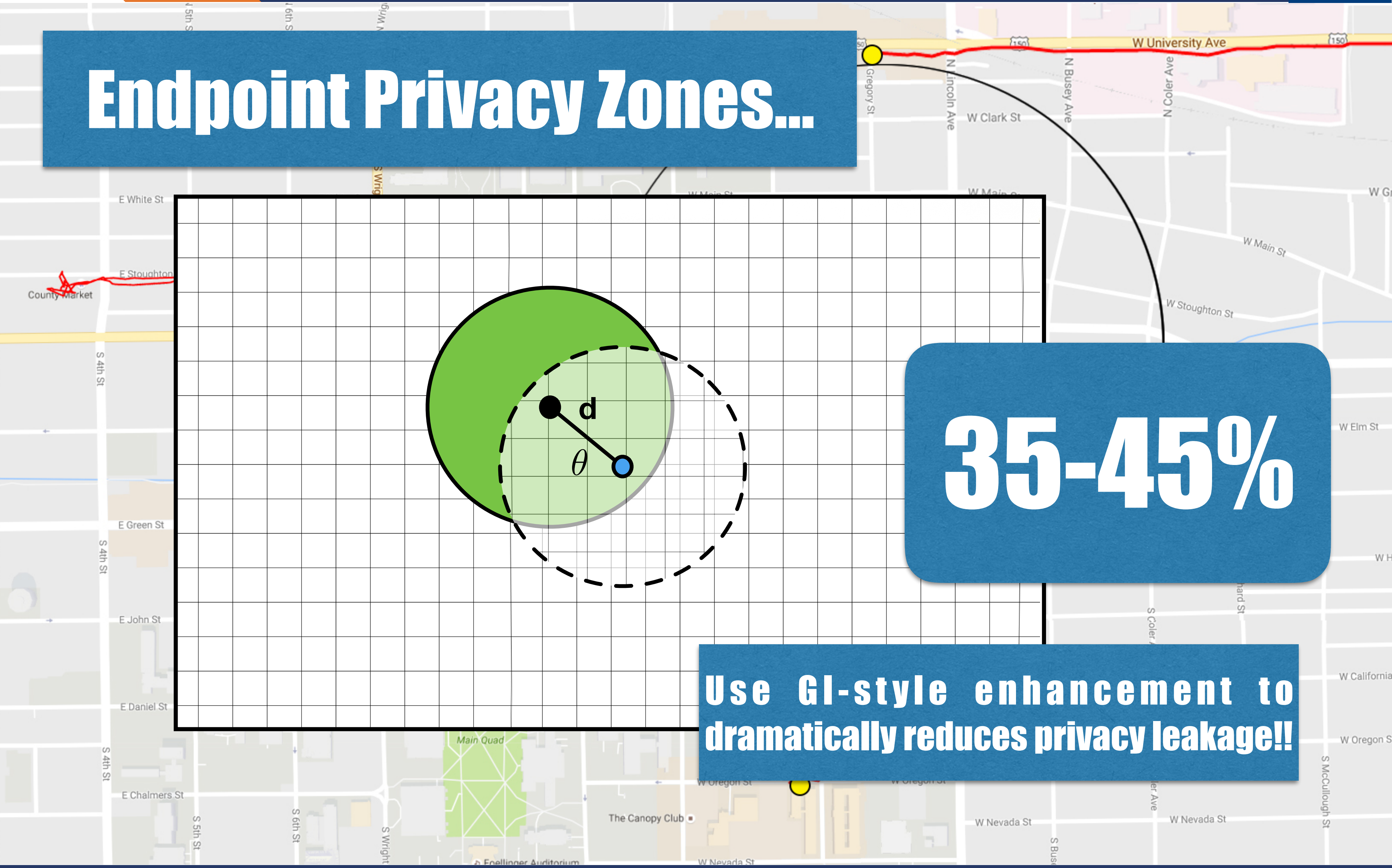
15% of Athletes use Privacy Zones

21 Million Activities
3 Million Athletes

95%

End-of-Talk Palette Cleanser…

Endpoint Privacy Zones…



35-45%

Use GI-style enhancement to dramatically reduces privacy leakage!!

Web Privacy: Looking Forward

- Where to look for privacy literature: "Big 4" security conferences (IEEE S&P a.k.a. Oakland, USENIX Security, CCS, NDSS), prestigious privacy-focused conferences (i.e., PETS).
- Hot Topics in Web Privacy (not exhaustive):
  - Fingerprinting browsers, devices, encrypted traffic
  - The WWW stack: cookies, CDNs, TLS/HTTPS adoption
  - OSNs: Policies, Features, Advertising, Inference attacks
  - Anonymity systems, secure communications, Tor
  - Data Processing: differential privacy, private stream aggregation
  - Location: Inference attacks, privacy-preserving mechanism