



ÉCOLE POLYTECHNIQUE
FÉDÉRALE DE LAUSANNE



Information Security and Privacy (COM-402)

Part 6: Privacy enhancing technologies

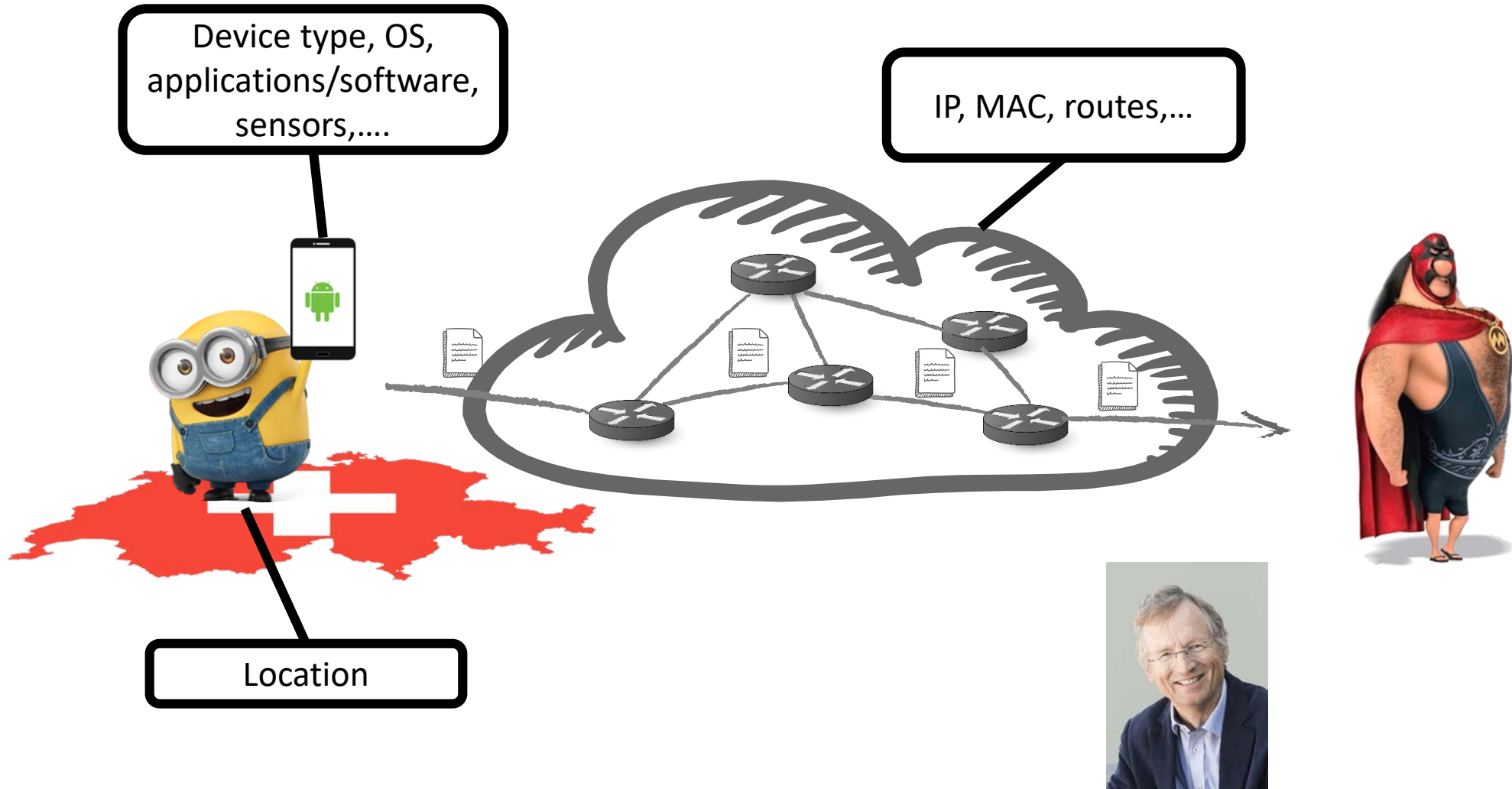
Metadata protection

Carmela Troncoso

SPRING Lab

carmela.troncoso@epfl.ch

Beyond data publication... privacy when data is in transit



Computing privately on data

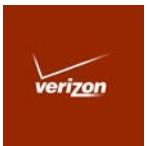
The adversary is anyone and VERY powerful



Intelligence
agencies



The Boss



ISPs



SysAdmins



Your
Parents



Your
Children



Your Roomates



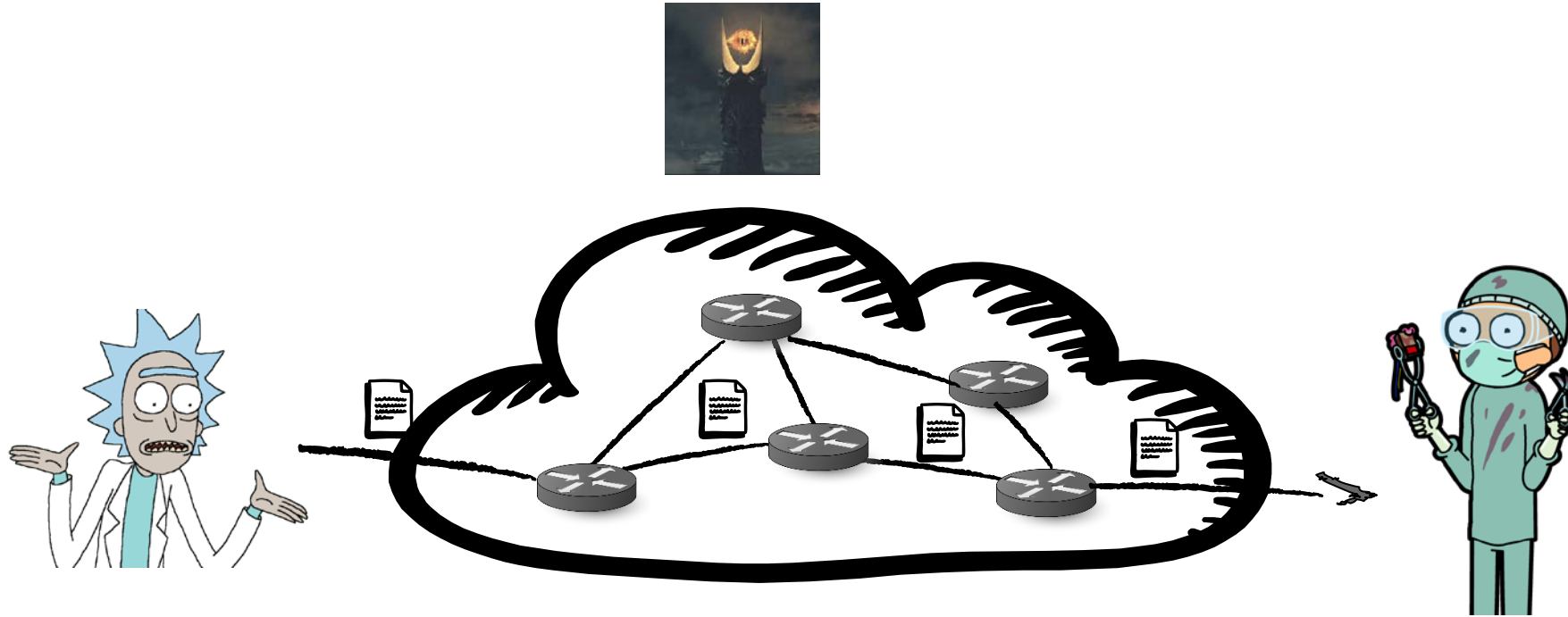
Anybody
curious



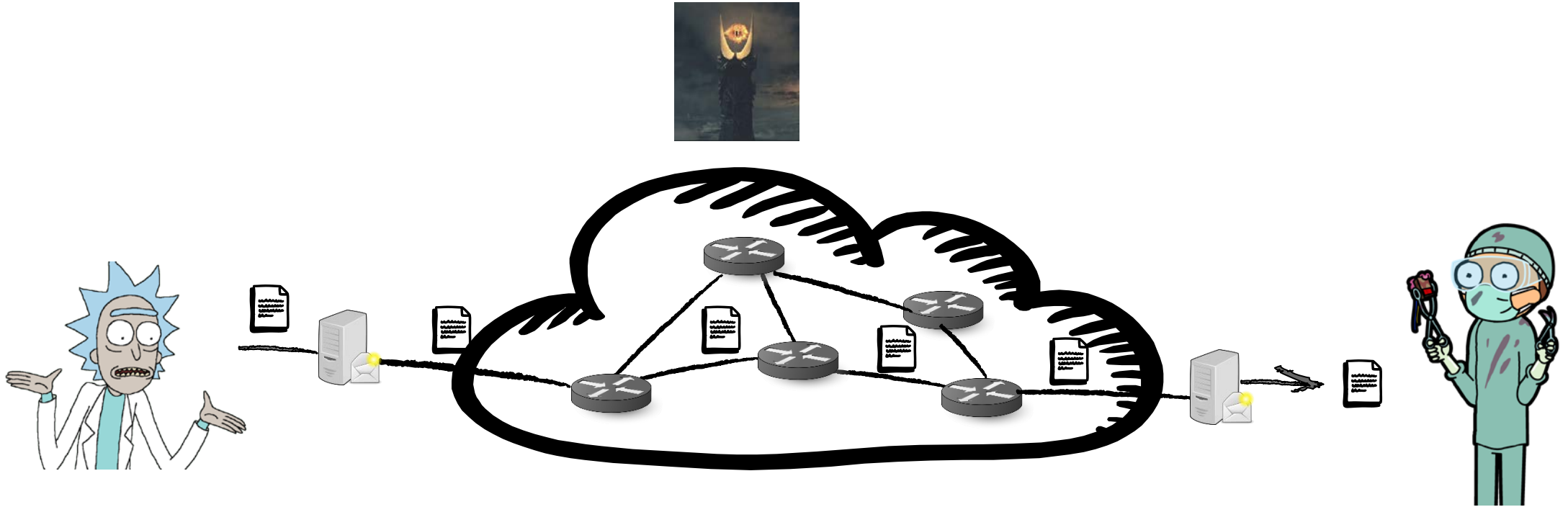
Dear Dr. Morty,
Can we change my
chemo appointment?
Rick



End to End Encryption



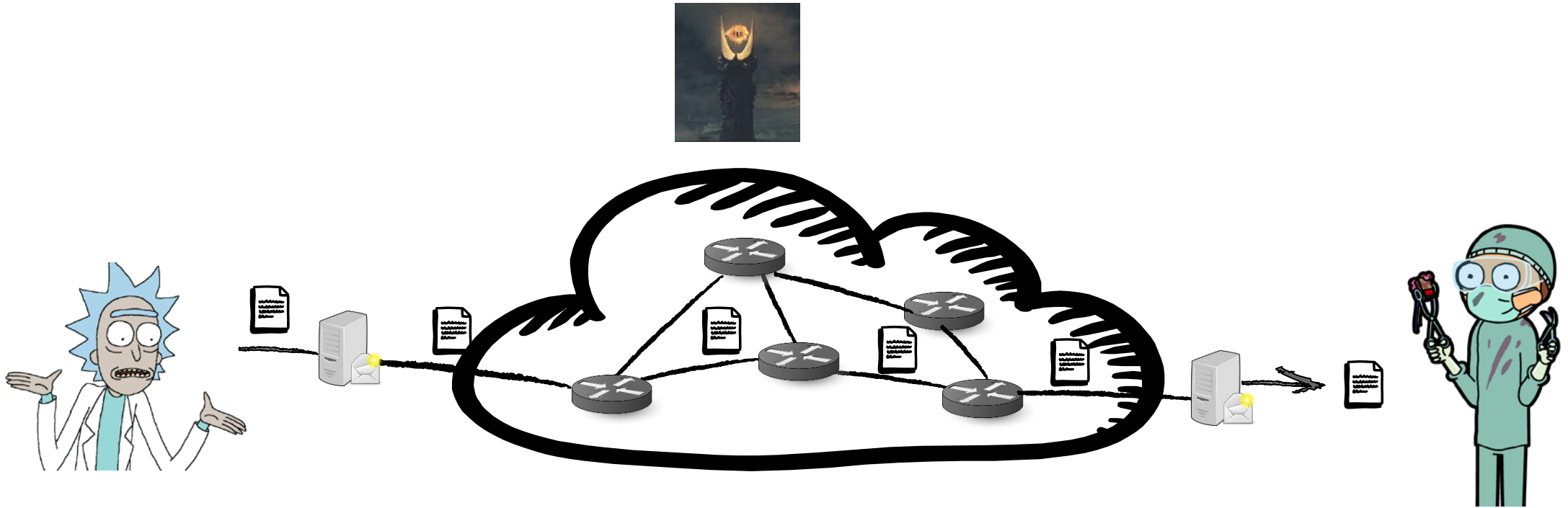
End to End Encryption



**Cryptography → Confidentiality!
(and integrity and authenticity)**

End to End Encryption

What is an End?



**Cryptography → Confidentiality!
(and integrity and authenticity)**

End to End Encryption



Perfect forward secrecy

Cryptography → Confidentiality!

BUT WHAT IF SOMEONE FORCES YOU TO DISCLOSE THE KEY?



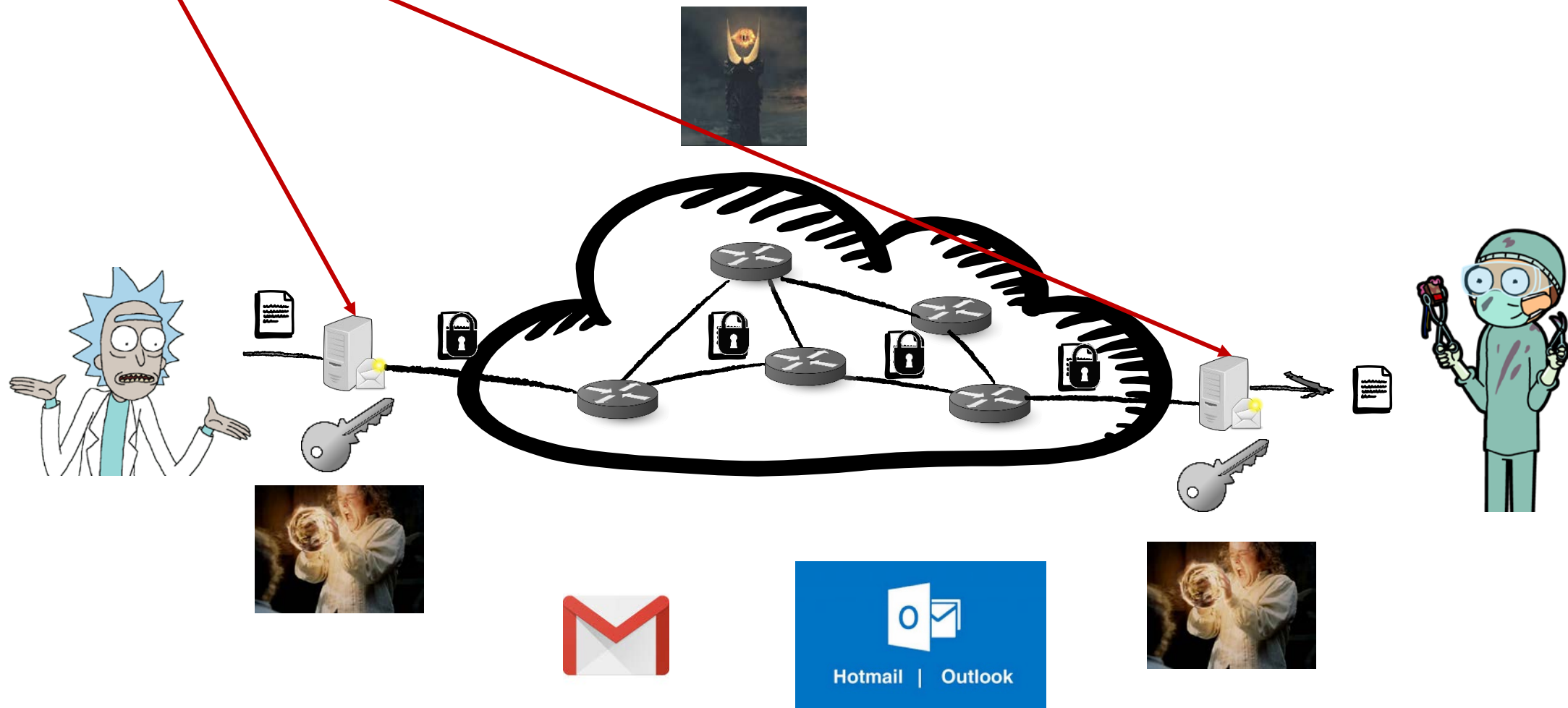
- 1) Start with keys that allow Alice to authenticate Bob.
 - Public key encryption
- 2) Alice and Bob create fresh public keys and exchange them
- 3) They establish fresh shared keys, and talk secretly
 - Diffie Hellman
- 4) Once done, they delete the shared keys.

ONE-TIME USE KEYS:
EPHEMERAL KEYS

AFTER A CONVERSATION IS OVER
NO-ONE CAN DECRYPT WHAT WAS SAID!!!


PLAUSIBLE DENIABILITY!!

End to End Encryption



The problem is Traffic Analysis



Bit Position: 0		4	8	16	24	31
Version	IHL	Type of Service		Total Length		
Identification				Flags	Fragment Offset	
Time to Live		Protocol		Header Checksum		
Source IP Address						
Destination IP Address						
IP Options (optional)					Padding	
						

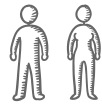
IPv4 Header (RFC 791, 1981)

*Same for Ethernet, TCP,
SMTP, IRC, HTTP, ...*

Traffic WHAT?

Wikipedia: traffic analysis is the process of intercepting and examining messages in order to deduce information from patterns in communication

Making use of “just” traffic data of a communication (aka metadata) to extract information
(as opposed to analyzing content or perform cryptanalysis)



Identities of
communicating parties



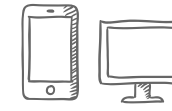
Timing, frequency,
duration



Location



Volume



Device

MILITARY ROOTS

M. Herman: “These non-textual techniques can establish **targets' locations**, order-of-battle and **movement**. Even when messages are not being deciphered, traffic analysis of the target's Command, Control, Communications and intelligence system and its patterns of behavior provides indications of his **intentions** and **states of mind**”

WWI: British troops finding German boats.

WWII: assessing size of German Air Force, fingerprinting of transmitters or operators (localization of troops).



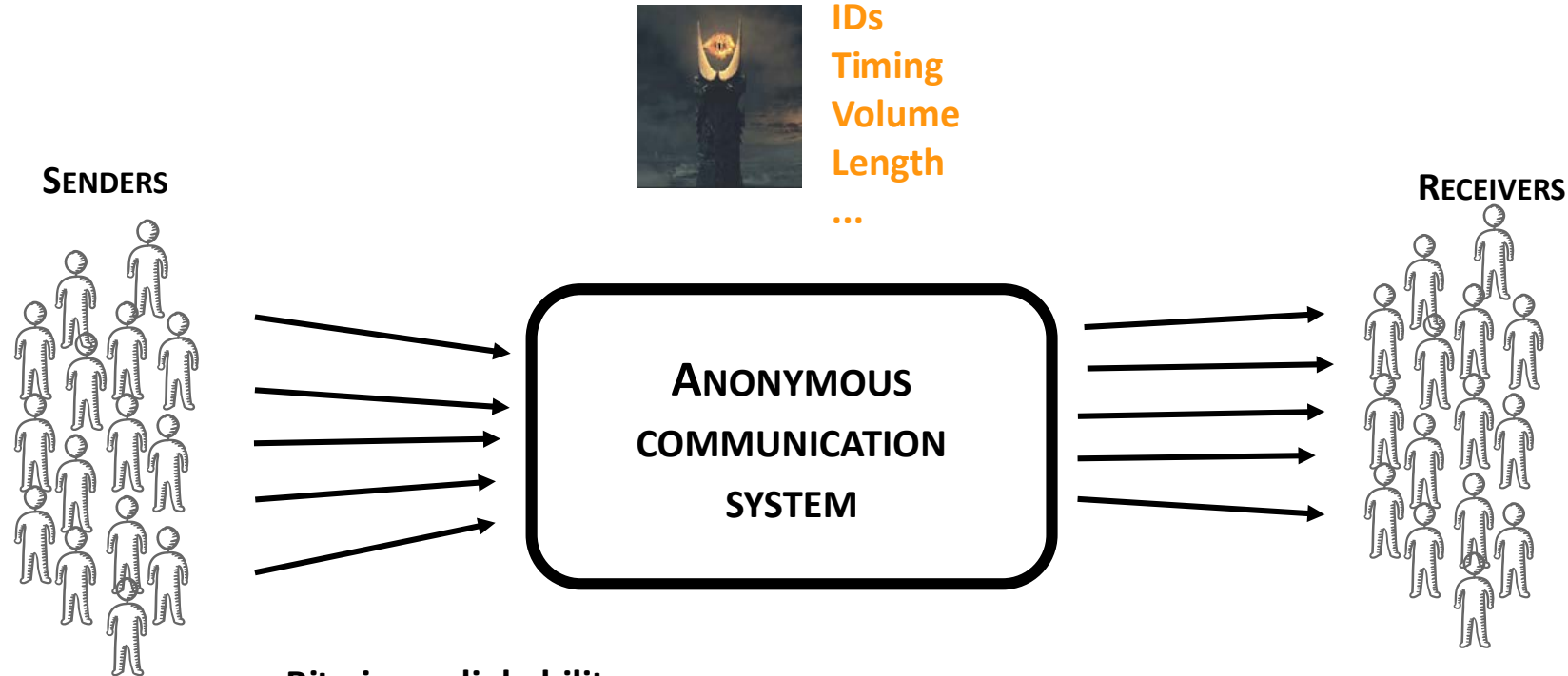
NOWADAYS

Diffie&Landau: “Traffic analysis, not cryptanalysis, is the backbone of communications intelligence”

Stewart Baker (NSA): “metadata **absolutely tells you everything about somebody's life**. If you have enough metadata, you don't really need content.”

Tempora, MUSCULAR → XkeyScore, PRISM

Anonymous communications – Abstract model



Bitwise unlinkability

Use cryptography to make inputs and outputs to the anonymous communication systems appearance (bits) different

(re)packetizing + (re)schedule

Destroy patterns (traffic analysis resistance)

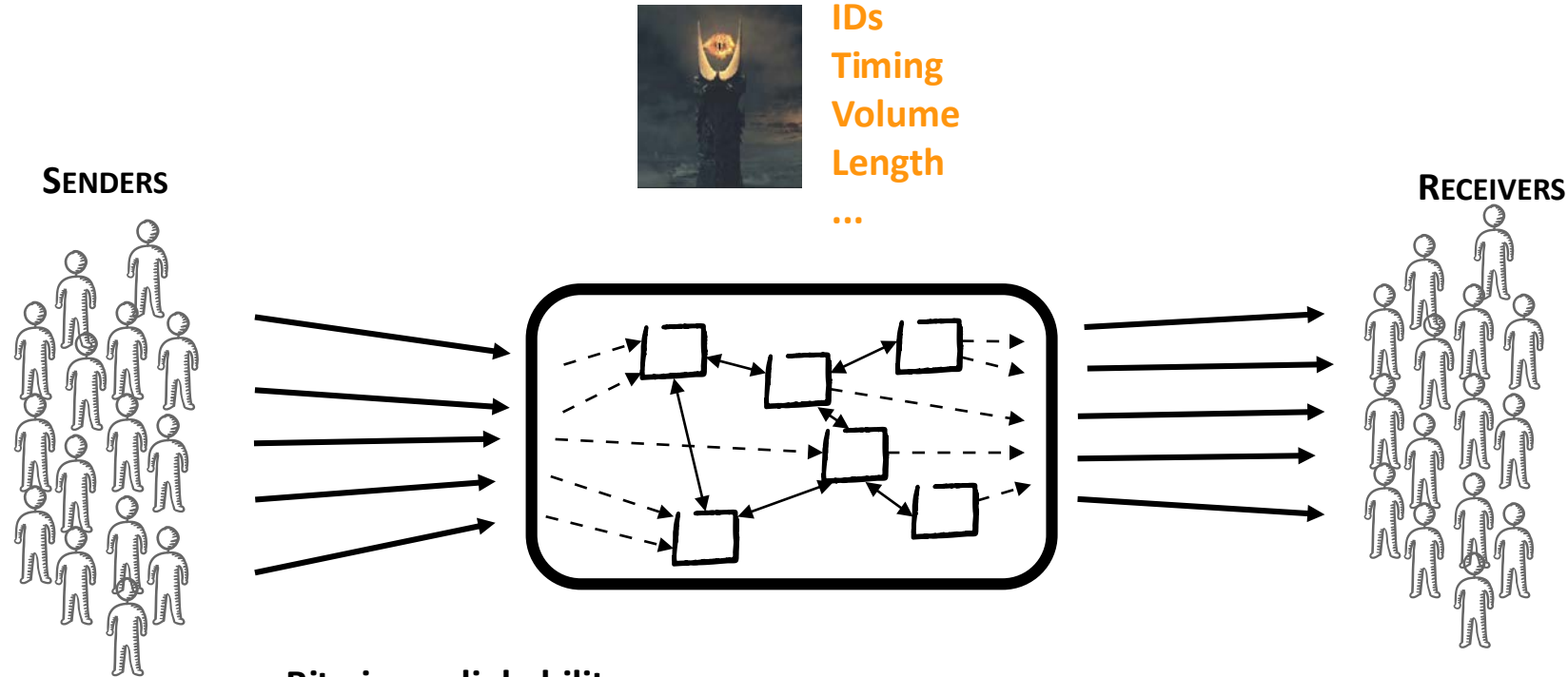
ONE-PROXY PROBLEMS

LOW THROUGHPUT

CORRUPT PROXY OR PROXY HACKED / COERCED

REAL CASE: PENET.FI VS THE CHURCH OF SCIENTOLOGY (1996)

Anonymous communications – Abstract model



Bitwise unlinkability

Use cryptography to make inputs and outputs to the anonymous communication systems appearance (bits) different

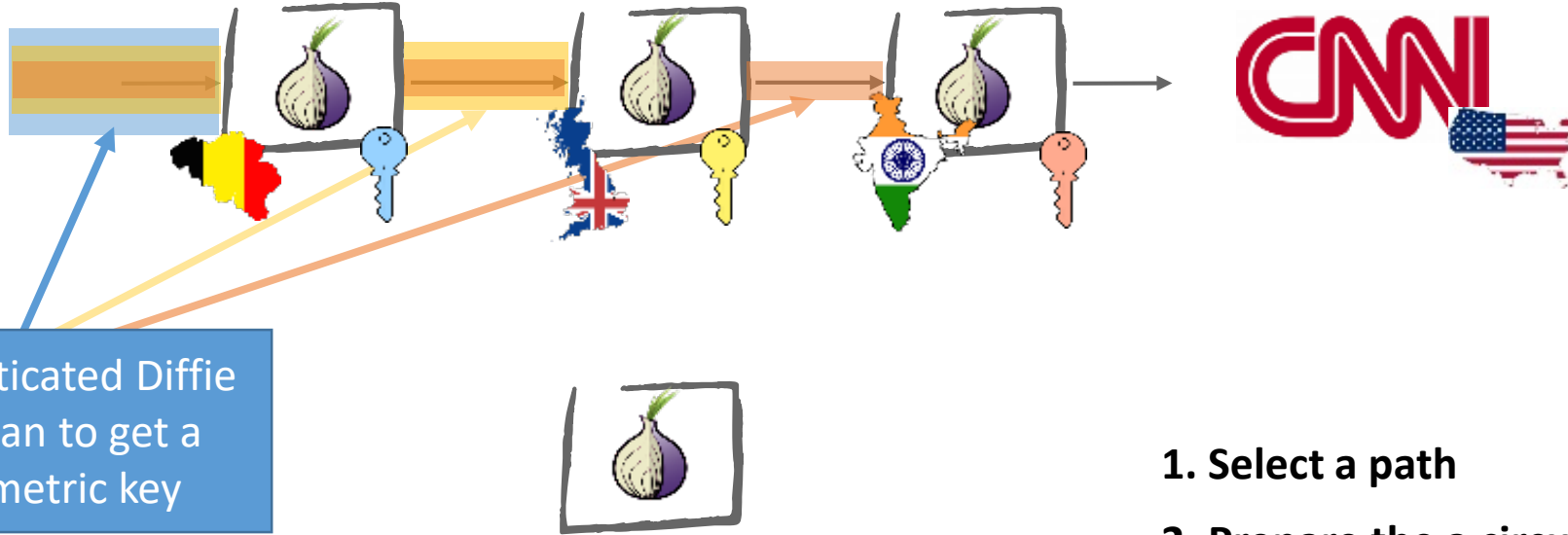
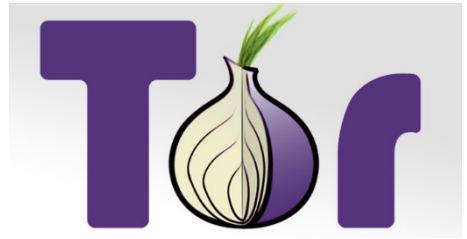
(re)packetizing + (re)schedule + (re)routing

Destroy patterns (traffic analysis resistance)

Load balancing

Distribute trust

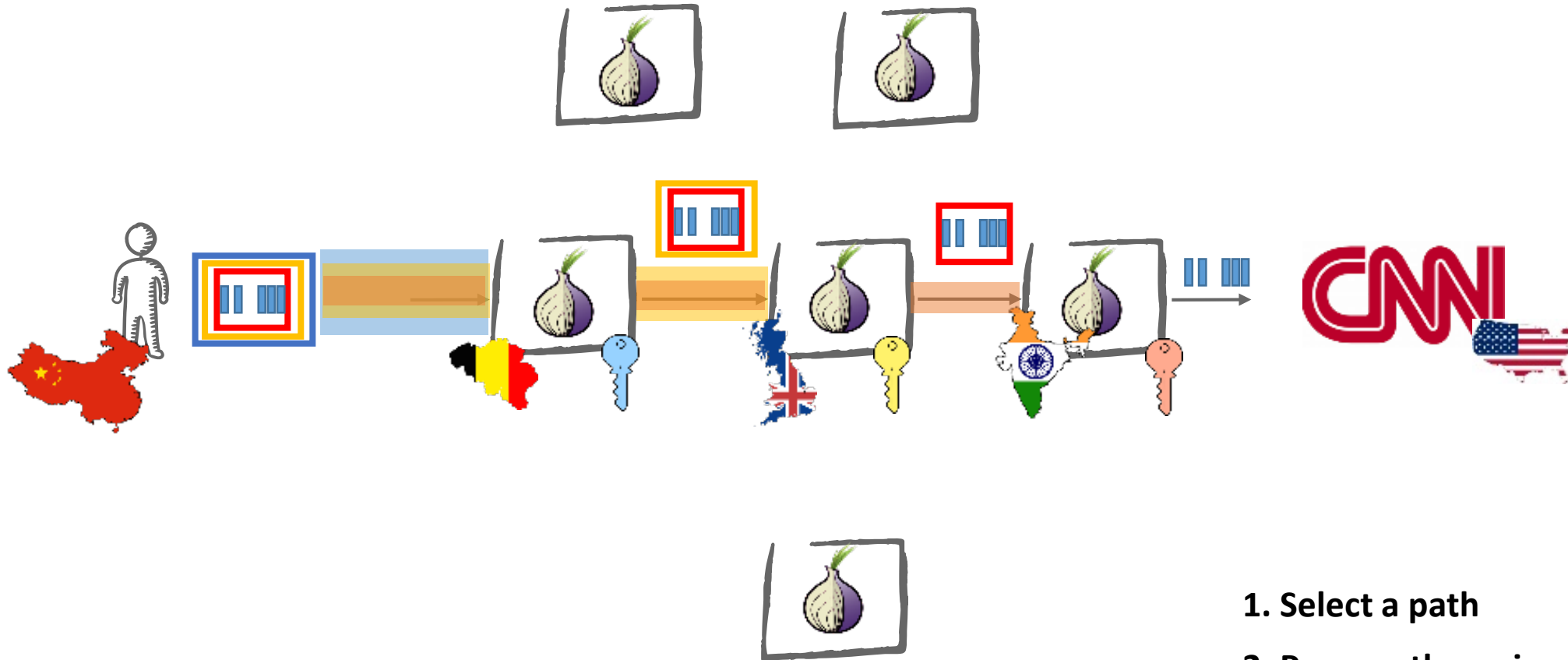
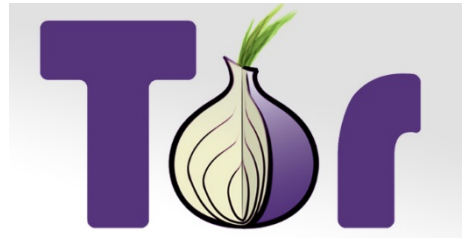
The Tor network – Onion routing



Authenticated Diffie
Hellman to get a
symmetric key

1. Select a path
2. Prepare the a circuit

The Tor network – Onion routing



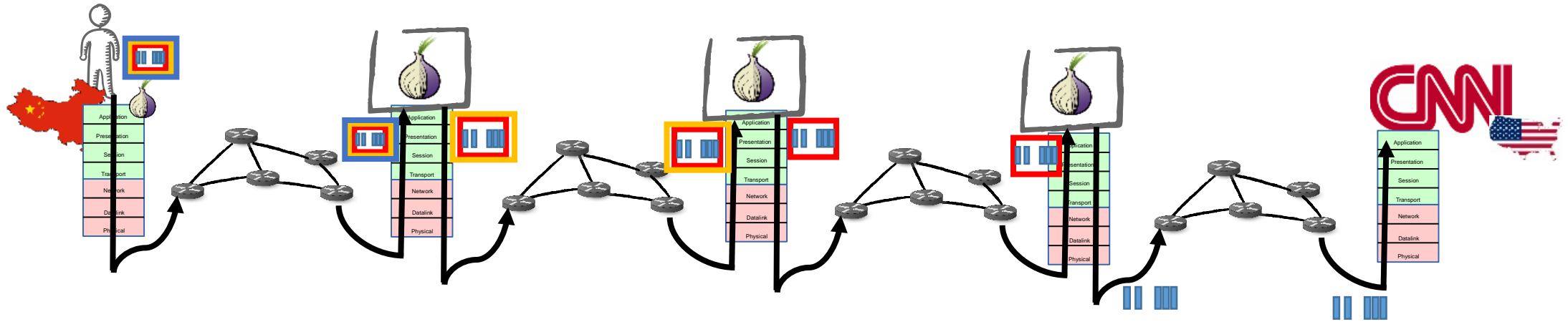
1. Select a path
2. Prepare the a circuit
3. Send stream

Anonymous communication networks are overlay networks

Nodes in anonymous communication networks (e.g., onion routers in Tor) are **not** internet routers. They work at the application layer!

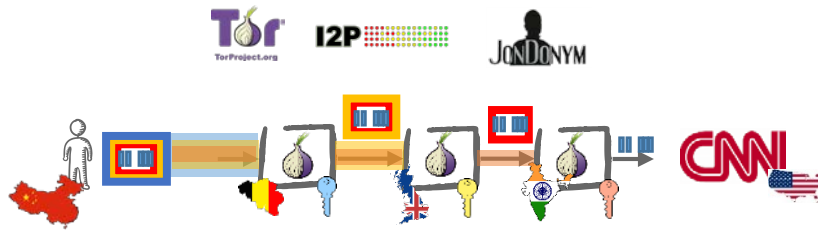
(overlay network = a computer network that is built on top of another network)

A more realistic view of how Tor traffic travels would be this



Anonymous communications out there

LOW LATENCY 

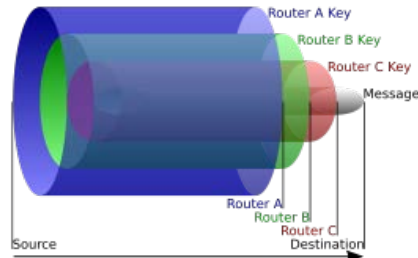


Web browsing, Instant Messaging, streaming

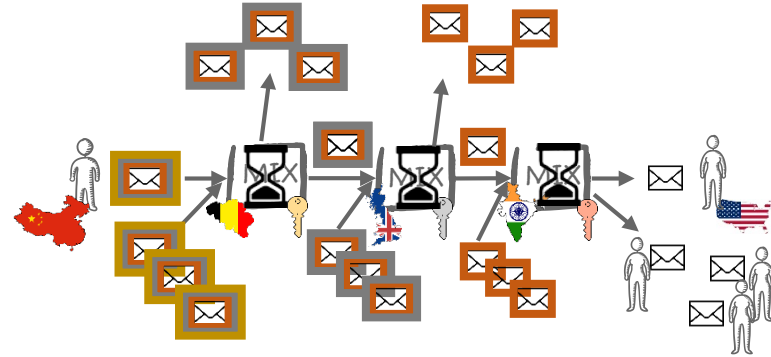
STREAM-based:



**fixed
for the
stream**



HIGH LATENCY 



Email, Voting

MSG-based:

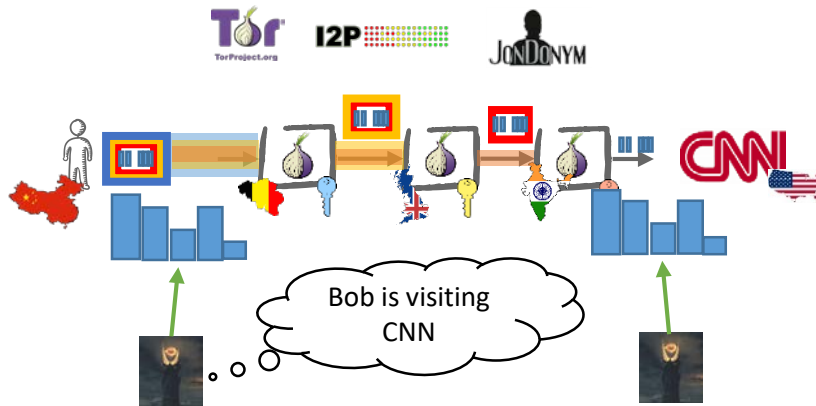


vary every message

One route per message + delays
(slower!)

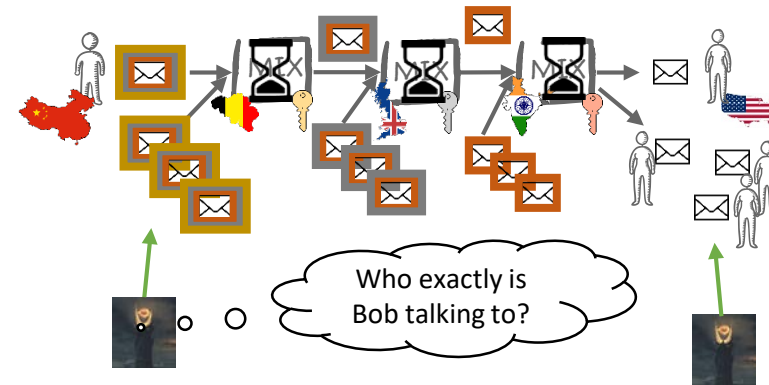
Anonymous communications out there

LOW LATENCY 



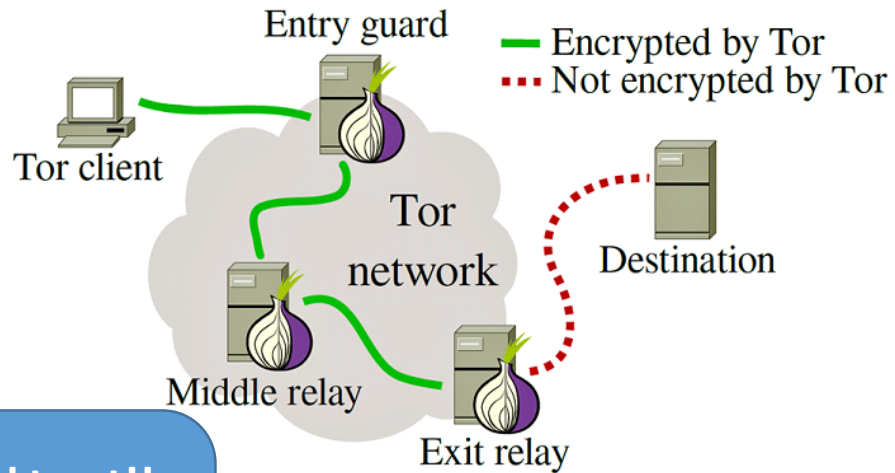
Cannot resist **Global** Adversary
(Tor assumes that the adversary cannot
see both edges)

HIGH LATENCY 



Global Adversary resistance
at the cost of latency
(and long term patterns revealed)

Anonymous communications vs. VPN



Decentralized trust!!
Provides privacy as long as the adversary cannot see both edges

Centralized trust. No anonymity vs the VPN, or anyone seeing the VPN



Takeaways

When thinking about end-to-end encryption it is important to think:

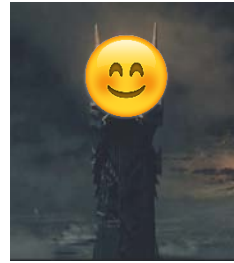
- Who are the ends of the communication vs. who is the adversary
- Forward secrecy, what happens if one key is compromised

Encryption is great, but for privacy **protecting traffic metadata is as important**

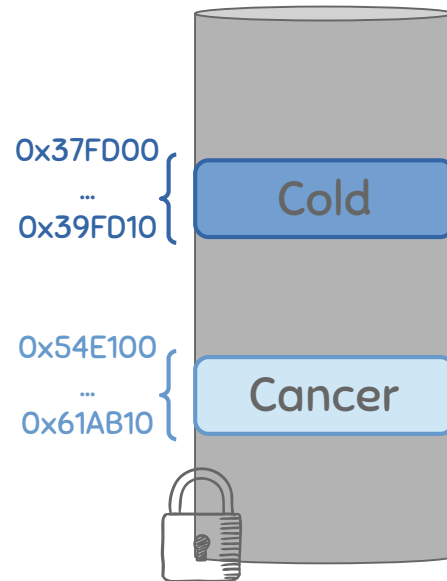
Low-latency communications – fast, but **only protect from partial adversaries**

High-latency communications – slow, but **protect from global adversaries**

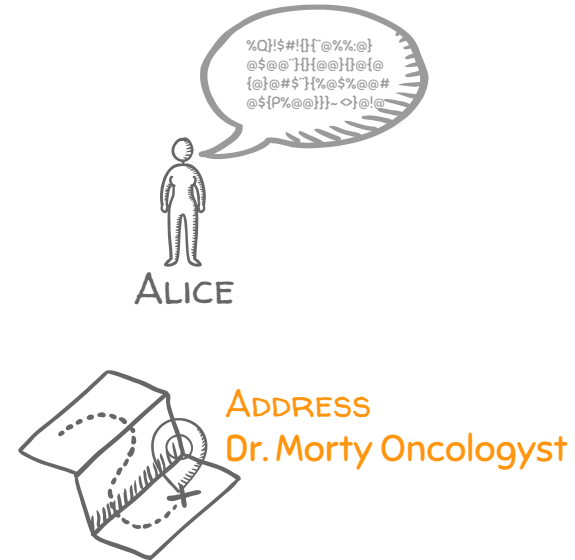
Other metadata is also sensitive!!



Implicit data is as
important as explicit data!



The hardware, software, firmware
of a device is a unique identifier
that encodes information
Example: medical app installed.



The address where data is stored may reveal
information about the content.

Example: medical database with patients with mild
and severe diseases in different locations

The address where an action happens may reveal
information about the action / user.

Example: sending a message from an Oncologist
clinic reveals information about the sender

Tracking anonymous users

The cookie zoo



Normal cookies
persistent identifiers

Third party cookies
webs establish cookies for other webs
webs have identifiers for other webs

Evercookies!

Tracking anonymous users without cookies



Would tracking be solved??



Fingerprinting Web Browsers

- User agent
- HTTP ACCEPT headers
- Browser plug-ins
- MIME support
- Clock skew
- Installed fonts
- Cookies enabled?
- Browser add-ons
- Screen resolution



**Mobile phones are also
fingerprintable!**



PANOPTICCLICK^{3.0}

Is your browser safe against tracking?

When you visit a website, online trackers and the site itself may be able to identify you – even if you've installed software to protect yourself. It's possible to configure your browser to thwart tracking, but many people don't know how.

Panopticlick will analyze how well your browser and add-ons protect you against online tracking techniques. We'll also see if your system is uniquely configured—and thus identifiable—even if you are using privacy-protective software. However, we only do so with your explicit consent, through the TEST ME button below.

TEST ME

☒ Test with a real tracking company [what's this?](#)

Only **anonymous data** will be collected through this site.

Panopticlick is a research project of the Electronic Frontier Foundation. EFF operates Panopticlick in the United States, which may not provide as much privacy protection as your home country. Panopticlick is part of an effort to illustrate the problem with tracking techniques, and help get stronger privacy protections for everyone. [Learn more.](#)

SHARE ON FACEBOOK

SHARE ON TWITTER

SHARE ON GOOGLE+



Location privacy: Points of interest (POIs)

specific location that someone may find useful or interesting

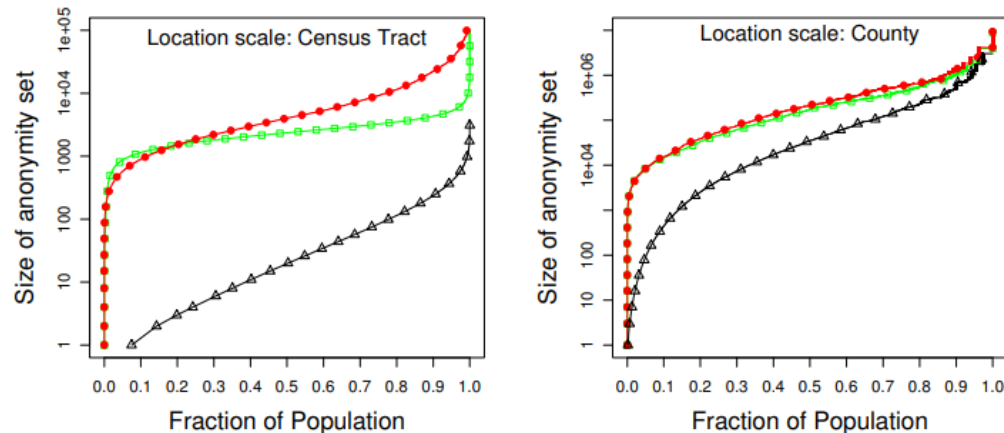
Why are POIs important?

- **Movements are unique** [De Montjoye et al 2013] [De Montjoye et al 2015]

4 spatio-temporal points are enough to uniquely identify 95% of people in a mobile phone database of 1.5M people and to identify 90% of people in a credit card database of 1M people

- **Home and Work: unique identifier** [Golle & Partridge 2009]

individual's anonymity set in the U.S. working population is 1, 21 and 34,980, for locations known at the granularity of a census block, census tract and county respectively



Location privacy: Points of interest (POIs)

specific location that someone may find useful or interesting

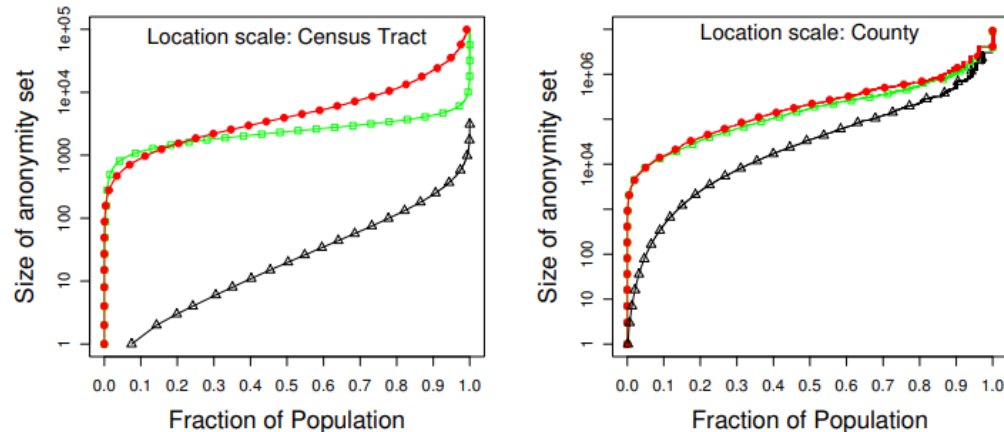
Why are POIs important?

- **Movements are unique** [De Montjoye et al 2013] [De Montjoye et al 2015]

4 spatio-temporal points are enough to uniquely identify 95% of people in a mobile phone database of 1.5M people and to identify 90% of people in a credit card database of 1M people

- **Home and Work: unique identifier** [Golle & Partridge 2009]

individual's anonymity set in the U.S. working population is 1, 21 and 34,980, for locations known at the granularity of a census block, census tract and county respectively



Location privacy: Points of interest (POIs)

specific location that someone may find useful or interesting

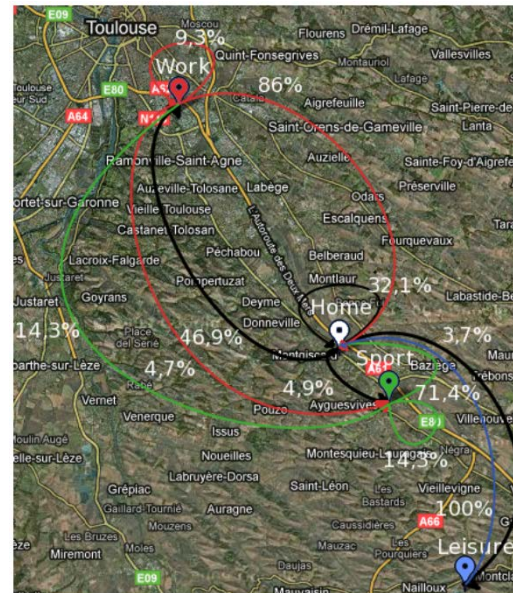
Why are POIs important?

- **N-top locations: unique identifiers** [Zhang & Bolot 2011]

[call records] *"top 2" locations likely correspond to home and work locations, the "top 3" to home, work, and shopping/school/commute path locations*

- **Where a user will move next** [Gambs et al 2012]

Accuracy for the prediction of the next location in the range of 70% to 95%



**Hidden Markov Model
movement patterns**

Location privacy: Points of interest (POIs)

specific location that someone may find useful or interesting

Why are POIs important?

- **Learning about users' motivation** [Bilogrevic et al 2015]

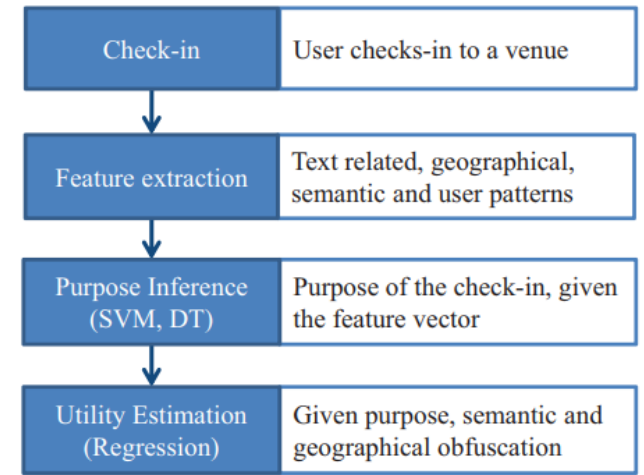
43 % correct classification (22% baseline predict most likely "Inform I am here")

Interesting utility impact study [complementary to this presentation]

- **Learning Demographics and other Patterns**

[Pang and Zhang 2017] [Felbo et al 2017][Cho et al 2010] [Liao et al 2005] [Liao et al 2007]

Machine-learning based frameworks



Location privacy: Defenses

- **Perturbation:** report a perturbed noise for the location
 - How to add the noise? (remember that the adversary knows!)
- **Generalization:** report a larger region instead of a point, i.e., reduce precision
- **Hiding:** do not report every single location
 - How to hide? What about recurrent patterns?
- **Add dummy locations:** hide the real trips among dummies
 - How to create the dummies in an undistinguishable manner?

Takeaways

Metadata is as important as content

Anonymous communications protect traffic data

- Low latency (Tor) vs. High latency (Mixes)

- Tradeoff performance for security (stronger adversary)

Tracking anonymous users is easy

- Devices reveal / collect too much information

Location is very revealing

- Hard to defend, correlations enable inferences