The Onion Router and Dark Web

And also, ML and web systems bias

Review: Ethics

- **Ethics** are a set of guiding principles for deciding whether behavior is acceptable or not
- In web systems, we must comply with ethical guidelines when it comes to data we collect and use
  - **Who owns** the data?
  - How do we **collect** the data?
  - Once collected, what will be **use** it for?

- Sometimes, data can be collected easily and via *implied consent*
  - When you search for something on Google, they'll collect your IP and search history
  - But sometimes, easily-collected data can be used for other purposes
    - In which case you may need **informed consent**

- We must gather **informed consent** when collecting data
  - The subject must understand *what* data is collected, *why*, and *how* it will be used
  - It is unethical to collect certain data without informed consent

- Unethical behavior *sometimes* is met with force of law, but often **credibility** is critical
  - Will we ever trust Equifax again? Remember the "doctor" that published the vaccines = autism study?

Half-Slide Summary? Bias

(the other half is on Dark Web)

- We already saw **Google Bombing** to influence search results
- More generally, **big data** has led to an explosion of **deep learning**
- **Biases present in data may influence machine learning models**
  - Systemic biases that influence outcomes of predictions for a variety of purposes
  - Affects performance and use of web systems

The Intercept_

HOMELAND SECURITY WILL LET COMPUTERS PREDICT WHO MIGHT BE A TERRORIST ON YOUR PLANE — JUST DON’T ASK HOW IT WORKS

Sam Biddle

December 3 2018, 1:47 p.m.

39

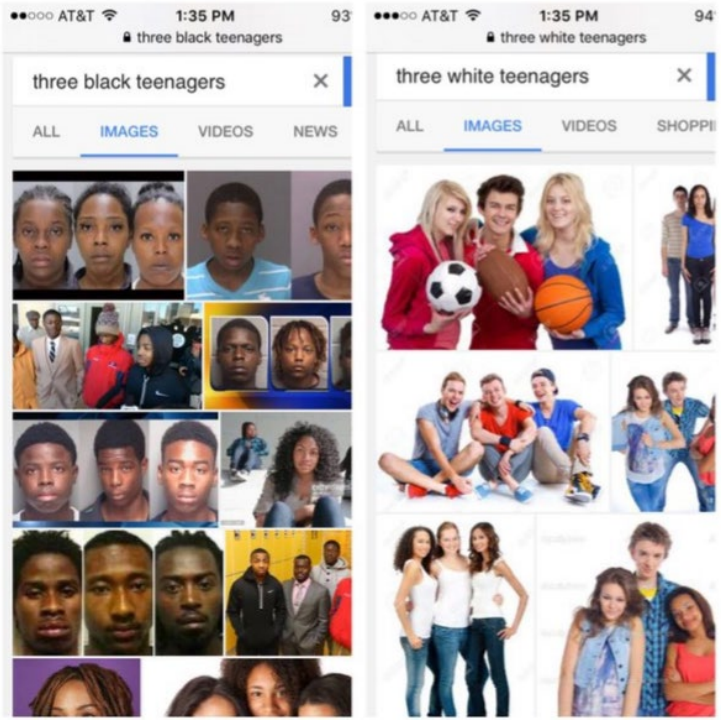
LAPD to scrap some crime data programs after criticism

By MARK PUENTE
APR 05, 2019 | 6:00 PM

https://www.latimes.com/local/lanow/la-me-lapd-predictive-policing-big-data-20190405-story.html

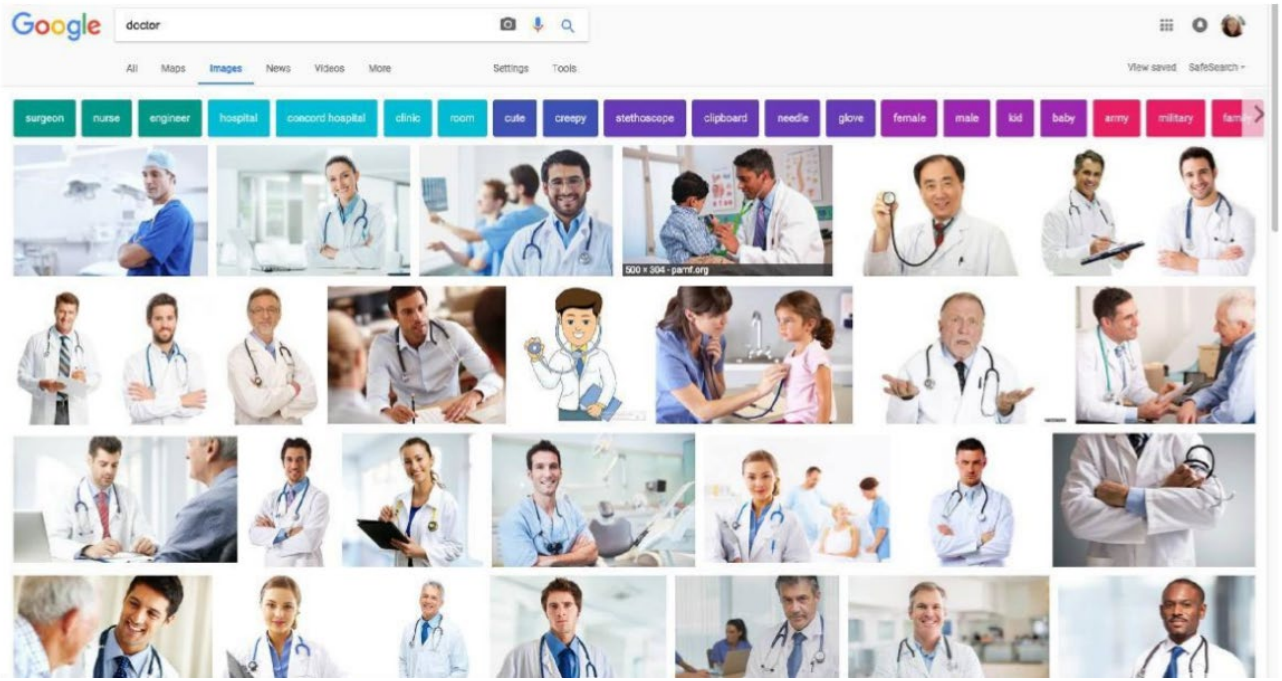
Impact of Social Stereotypes on Data

- 2016 Google queries: racial stereotypes



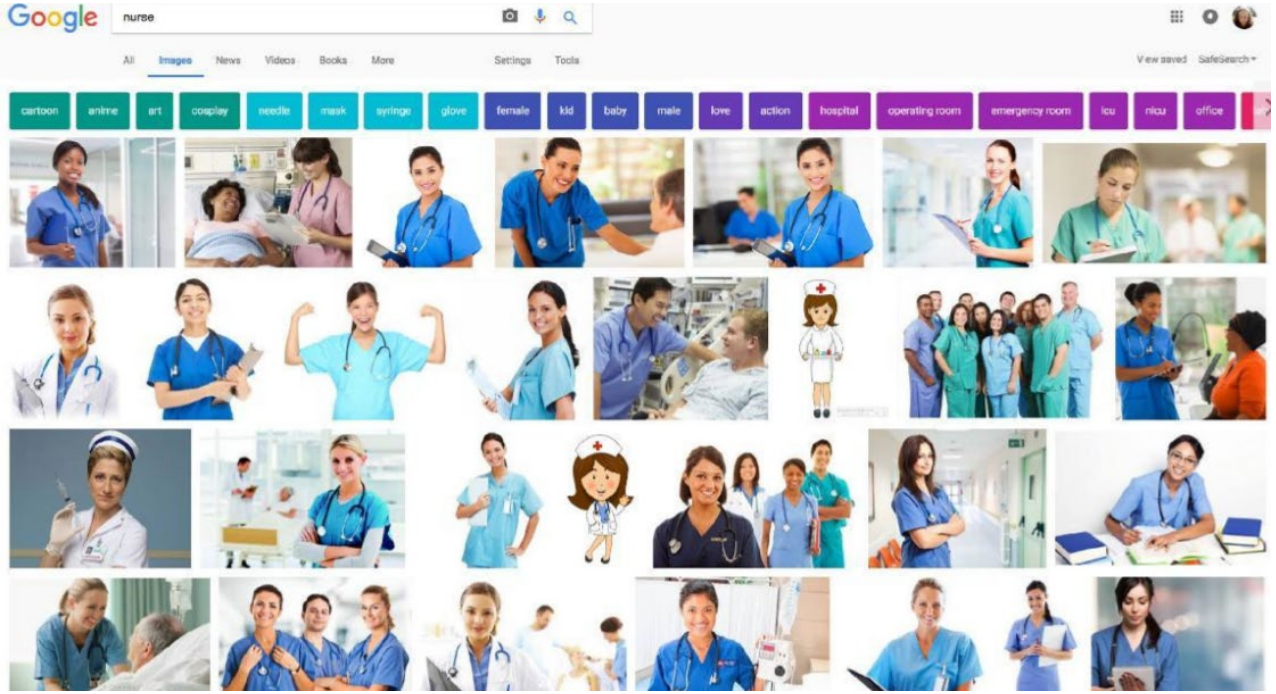
Impact of Social Stereotypes on Data

- Google query for “doctor”: race/gender/age stereotypes



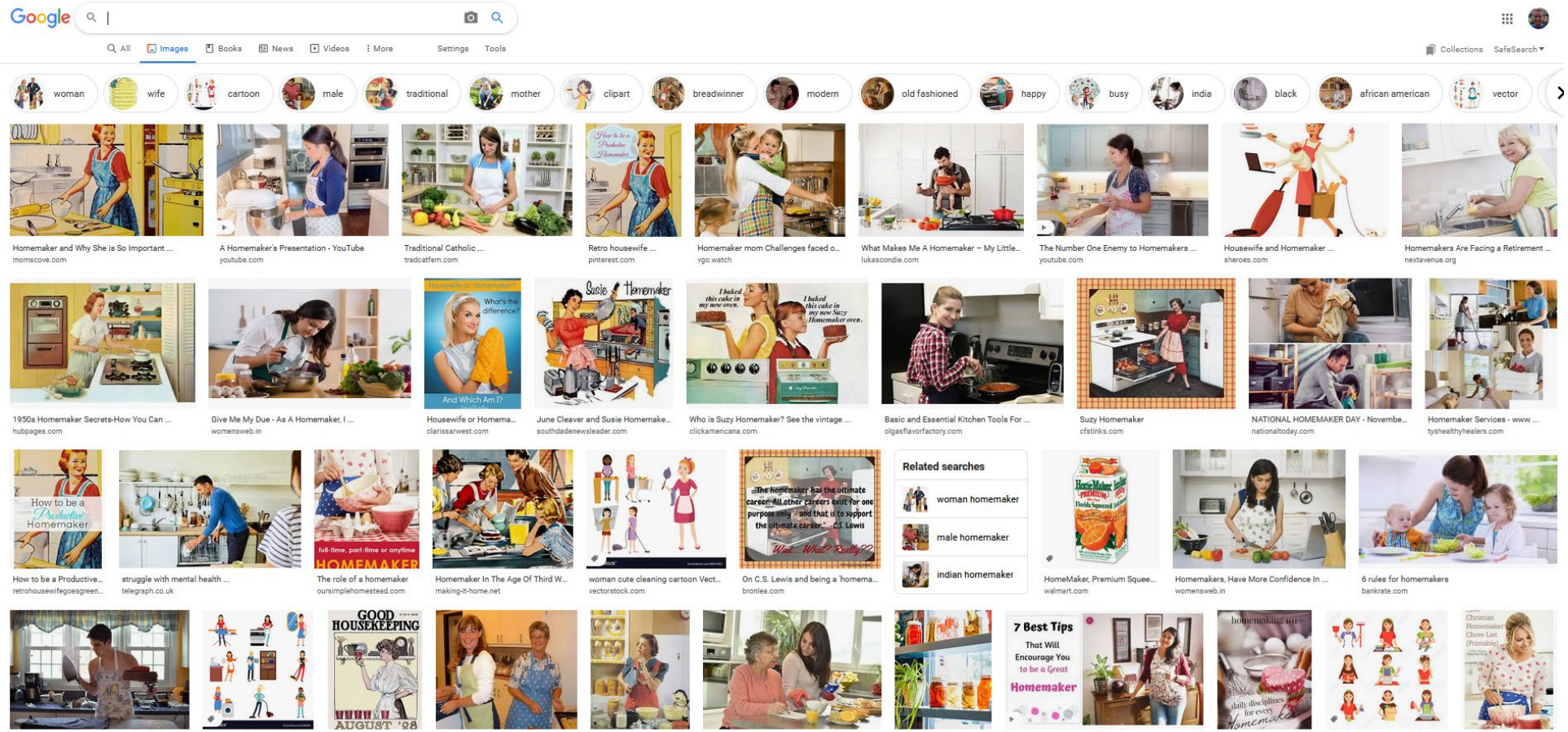
Impact of Social Stereotypes on Data

- Google query for “nurse”



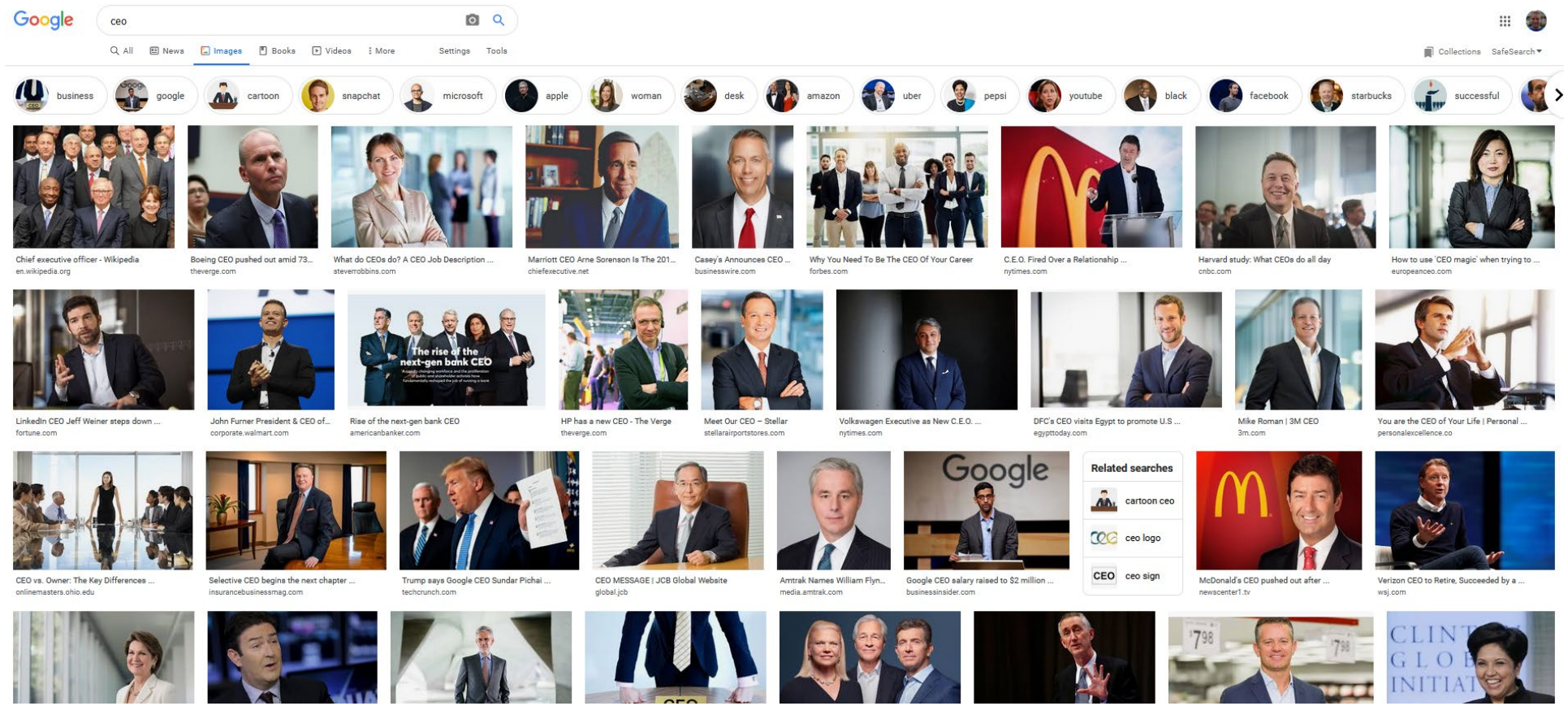
Impact of Social Stereotypes on Data

- Google query for “homemaker”



Impact of Social Stereotypes on Data

- Google query for “CEO”



Societal Stereotypes in Data

- Biased data produces biased models
  - Thus, predictions are biased as well

- Alternative thought question:
  - What is a chair?

Research on Bias in ML

- Machines learn trustworthiness and likeability traits from faces

  (Steed and Caliskan 2020)
- Self-driving cars biased against genders and races

  (Wilson, Hoffman, and Morgenstern 2019)
- Males are over-represented in the reporting of web-based news articles

  (Jia, Lansdall-Welfare, and Cristianini 2015)
- Males are over-represented in twitter conversations

  (Garcia, Weber, and Garimella 2014)
- Biographical articles about women on Wikipedia disproportionately discuss romantic relationships or family-related issues

  (Wagner et al. 2015)
- IMDB reviews written by women are perceived as less useful

  (Otterbacher 2013)

Sources of Bias

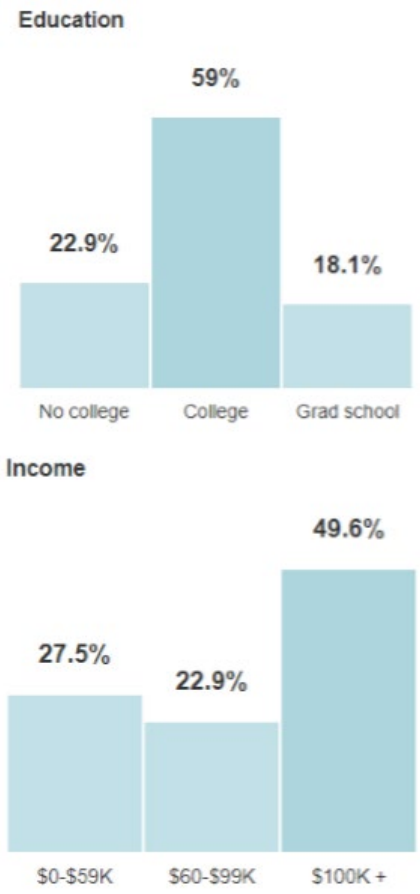
- Bias in data and sampling
  - (social biases, unrepresentative user base)
- Optimizing for a biased objective
  - (bad training)
- Inductive bias
  - (implicit assumptions made by the model itself)
- Bias amplification
  - (the model learns the “wrong” features)

Bias in Data and Sampling

- **Self-selection bias** is a statistical effect in which a group will select themselves, biasing a sample

- **Concretely:** who writes Yelp reviews? Who reads them?
  - People may not talk about things consistent with empirical measurement
  - Communities of language speakers lead to differing model performance
- What about system bias?
  - Can we tell if Yelp is biasing reviews?
    - "it would be a shame if you didn't pay us and you got a few 1-star reviews…"

Distribution of Yelp Users



Bias in Language Identification

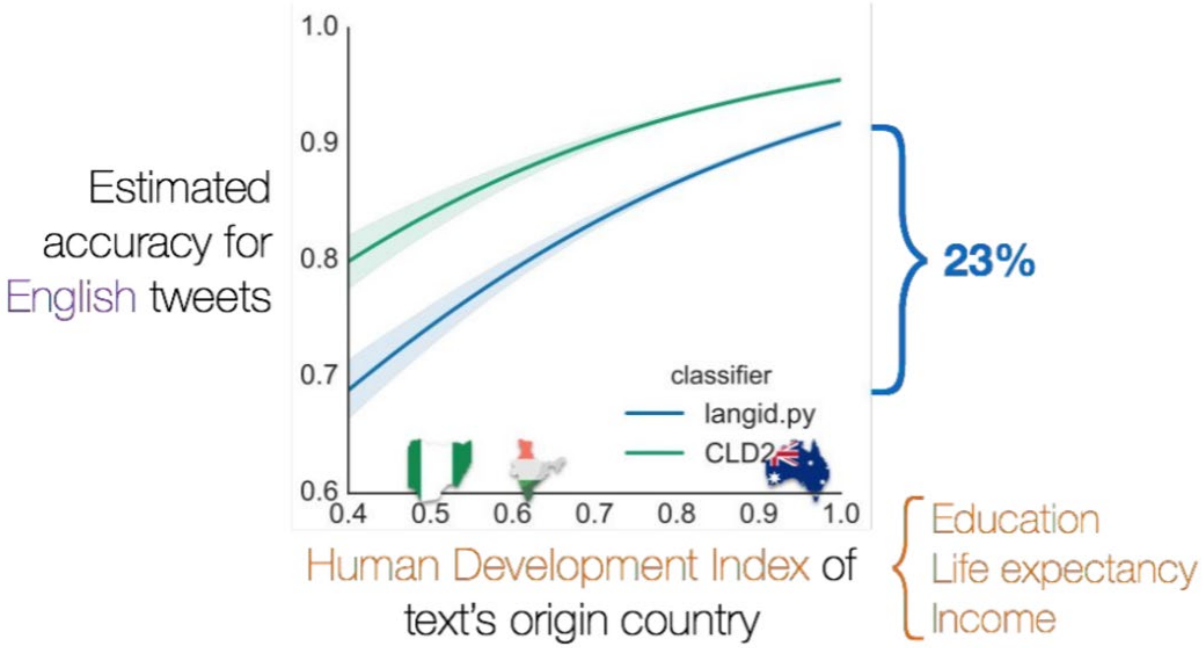
- ML application: Identifying a language give a string written in it

After language identification, we can look for keywords like flu/sick, then followup with a conclusive explanation
(maybe they're hungover)

If we can't identify the language to begin with, there's no way to extract followup semantics (i.e., we can't find keywords like flu/sick without knowing it's an English Tweet)

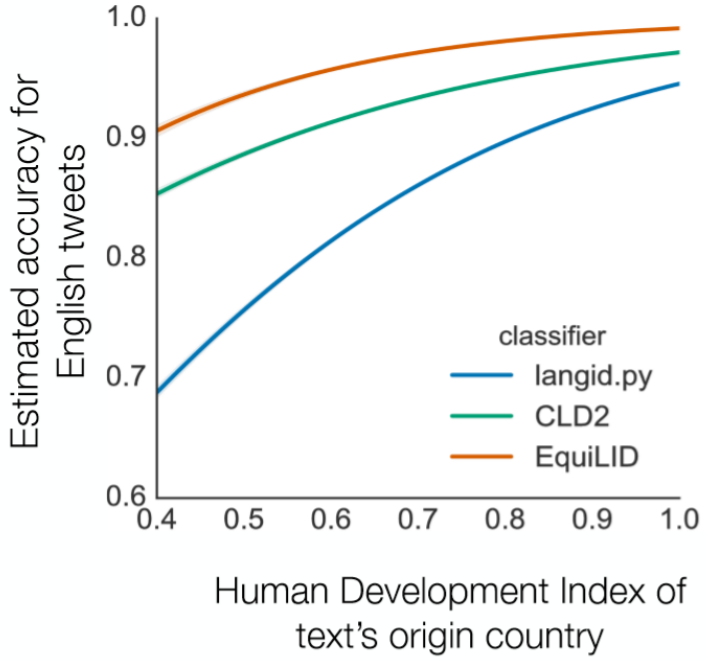
Bias in Language Identification

- Language Identification systems under-represent populations in underdeveloped countries



Bias in Language Identification

- By retraining on more representative corpora:



Objective Bias

- **Objective bias** occurs when models are asked to make predictions that actually answer a different question

- **Concretely:** “What is the **probability** that a given **person** will commit a serious **crime** in the **future** based on the **sentence given now**?”

- Example: COMPAS
  - Balanced data from people of all races (and race was not a feature)
  - **Problem:** “who will commit a crime” is not obtainable (we can’t know it ahead of time)
    - **Instead**: model was learning “who is more likely to be convicted” (notice the difference!)

Inductive Bias

- An **Inductive bias** is the result of an implicit assumption made in the construction of a given model

- **Concretely:** Datasets of words may represent biases

- Consider word2vec, an *embedding* for words
  - (gross oversimplification: fancy tf-idf scores)
  - You can use it to represent each *word* or *phrase* as a vector **based on examples of English text**
  - $\overrightarrow{man} - \overrightarrow{woman} \approx \overrightarrow{computer\ programmer} - \overrightarrow{homemaker}$

Inductive Bias in Embeddings

$$\min cos(he - she,\ x - y)\ s.t.\ ||x - y||_2 < \delta$$

Extreme *she*	Extreme *he*
1. homemaker	1. maestro
2. nurse	2. skipper
3. receptionist	3. protege
4. librarian	4. philosopher
5. socialite	5. captain
6. hairdresser	6. architect
7. nanny	7. financier
8. bookkeeper	8. warrior
9. stylist	9. broadcaster
10. housekeeper	10. magician

Gender stereotype *she-he* analogies

sewing-carpentry	registered nurse-physician	housewife-shopkeeper
nurse-surgeon	interior designer-architect	softball-baseball
blond-burly	feminism-conservatism	cosmetics-pharmaceuticals
giggle-chuckle	vocalist-guitarist	petite-lanky
sassy-snappy	diva-superstar	charming-affable
volleyball-football	cupcakes-pizzas	lovely-brilliant

Gender appropriate *she-he* analogies

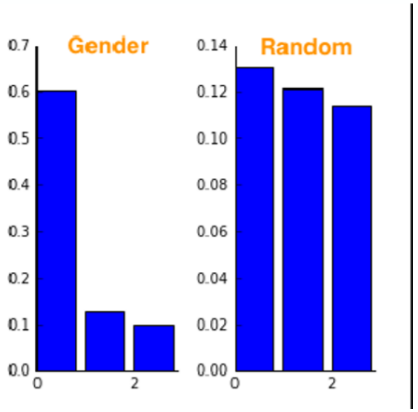
queen-king	sister-brother	mother-father
waitress-waiter	ovarian cancer-prostate cancer	convent-monastery

Figure 1: **Left** The most extreme occupations as projected on to the *she*−*he* gender direction on w2vNEWS. Occupations such as *businesswoman*, where gender is suggested by the orthography, were excluded. **Right** Automatically generated analogies for the pair *she-he* using the procedure described in text. Each automatically generated analogy is evaluated by 10 crowd-workers to whether or not it reflects gender stereotype.

Fixing Inductive Bias: Debiasing

$\overrightarrow{\text{she}} - \overrightarrow{\text{he}}$
$\overrightarrow{\text{her}} - \overrightarrow{\text{his}}$
$\overrightarrow{\text{woman}} - \overrightarrow{\text{man}}$
$\overrightarrow{\text{Mary}} - \overrightarrow{\text{John}}$
$\overrightarrow{\text{herself}} - \overrightarrow{\text{himself}}$
$\overrightarrow{\text{daughter}} - \overrightarrow{\text{son}}$
$\overrightarrow{\text{mother}} - \overrightarrow{\text{father}}$
$\overrightarrow{\text{gal}} - \overrightarrow{\text{guy}}$
$\overrightarrow{\text{girl}} - \overrightarrow{\text{boy}}$
$\overrightarrow{\text{female}} - \overrightarrow{\text{male}}$

- We can identify *gendered terms* to determine which *features* contribute to determining differences between them
  - Then, for other *non-gendered terms*, we can compute debiased distances by weighing the gendered features *less*



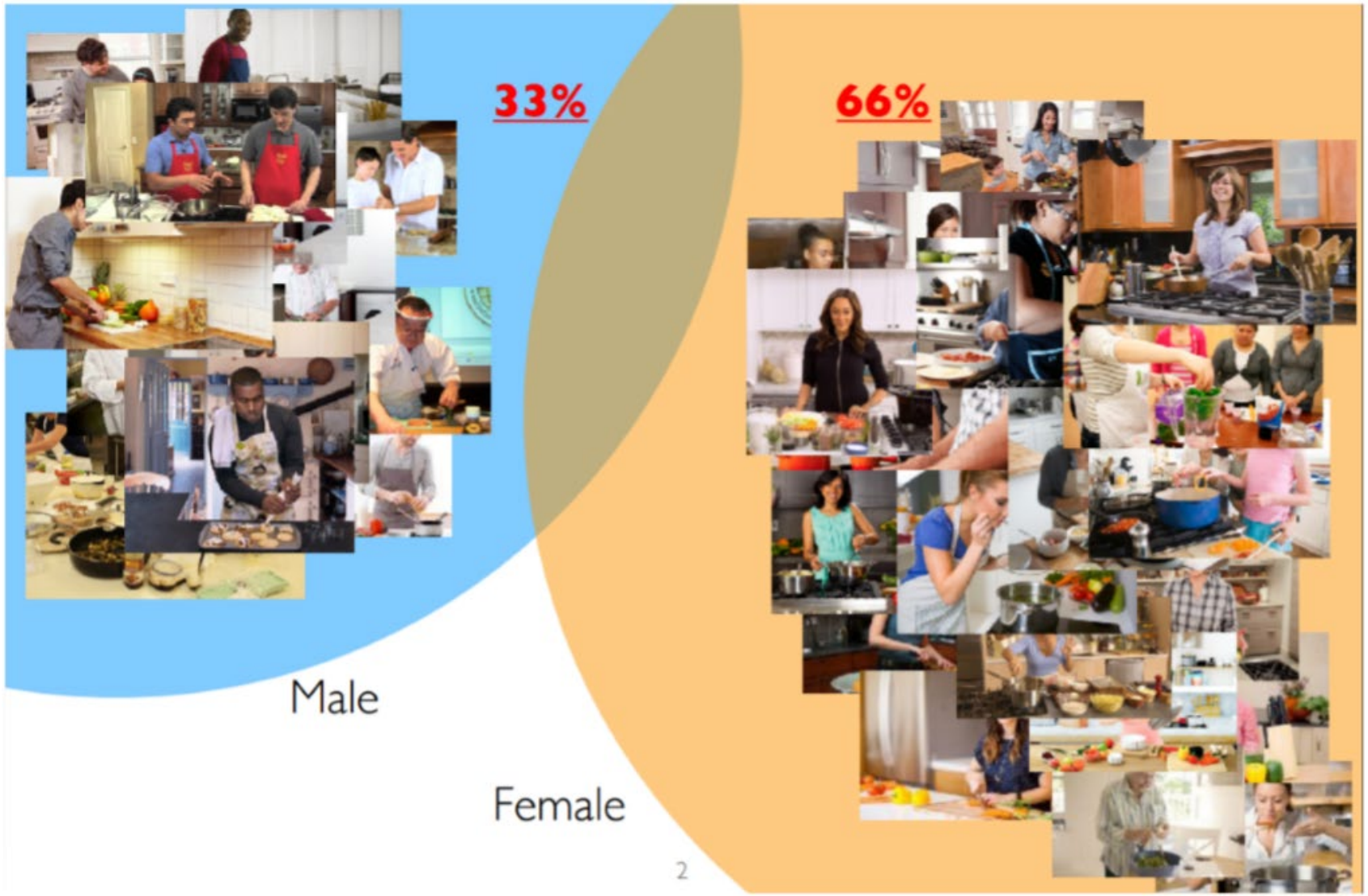
Top principal components identify gender subspace

Bias Amplification

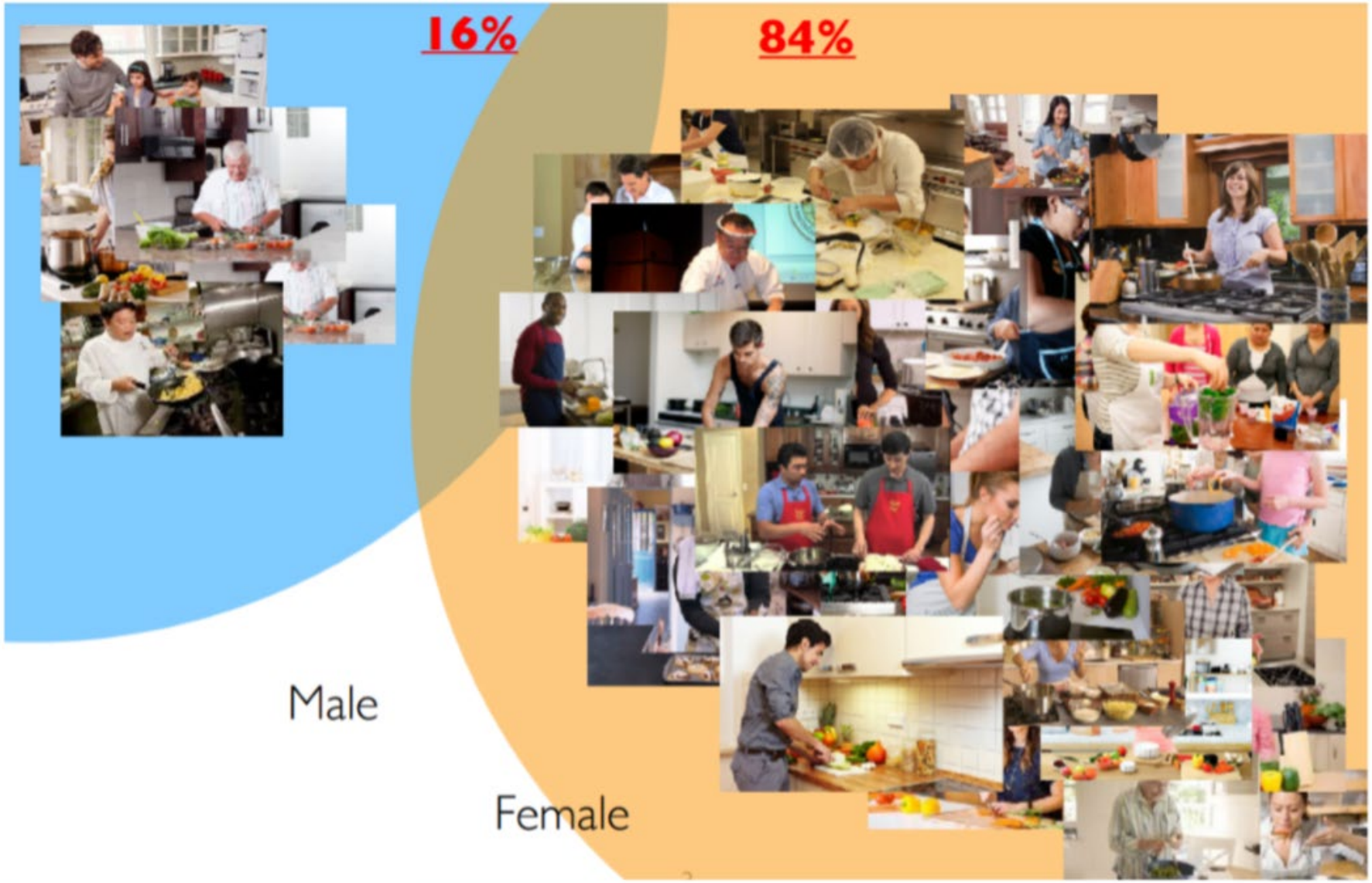
- **Bias Amplification** occurs when unrepresentative data leads a model to learn the wrong features

- **Consider:** What is a chair?
- **Concretely:** if all of your dataset contains barstools as examples of chairs, your model will learn the wrong features
  - e.g., it will only have examples of tall, backless seats near alcohol sources

Bias Amplification: Training

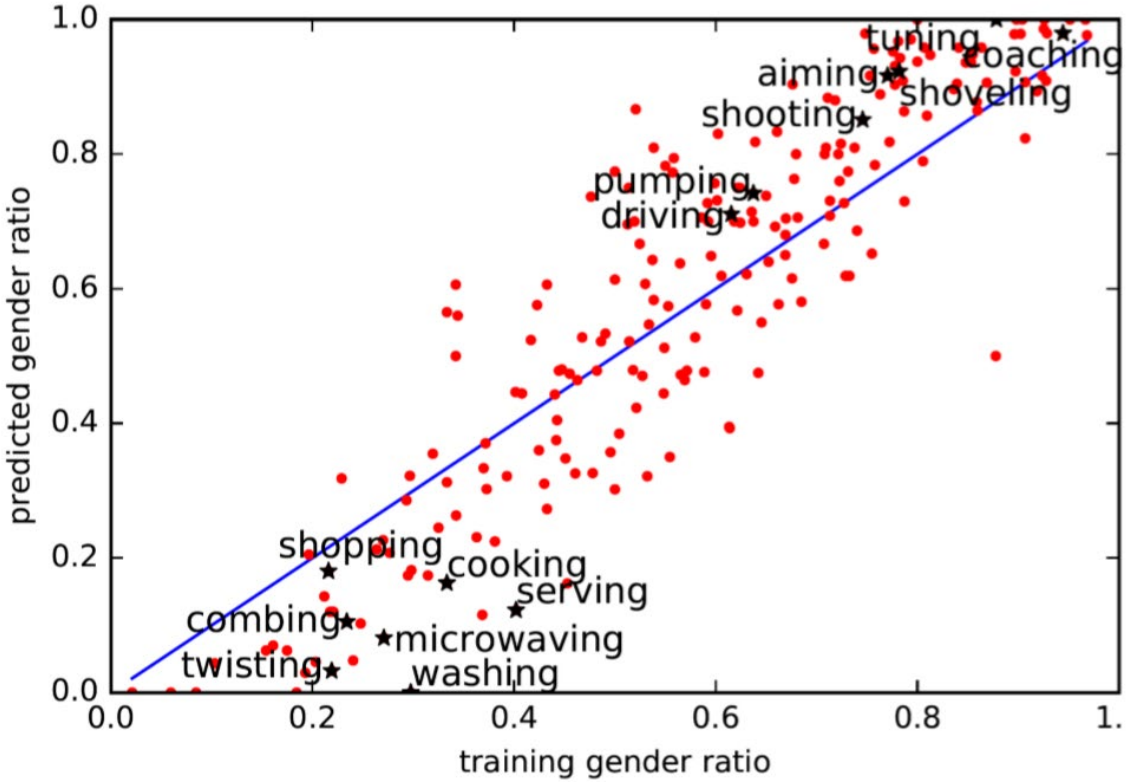


Bias Amplification: Predictions



Reducing Bias Amplification

- Find ratio of predictions made against ground truth labels
- Identify distribution of labels in dataset
- Adjust predicted outputs based on target distribution



Bias in ML and Web Systems

- Bias in data and sampling
  - (social biases, unrepresentative user base)
- Optimizing for a biased objective
  - (bad training)
- Inductive bias
  - (implicit assumptions made by the model itself)
- Bias amplification
  - (the model learns the “wrong” features)

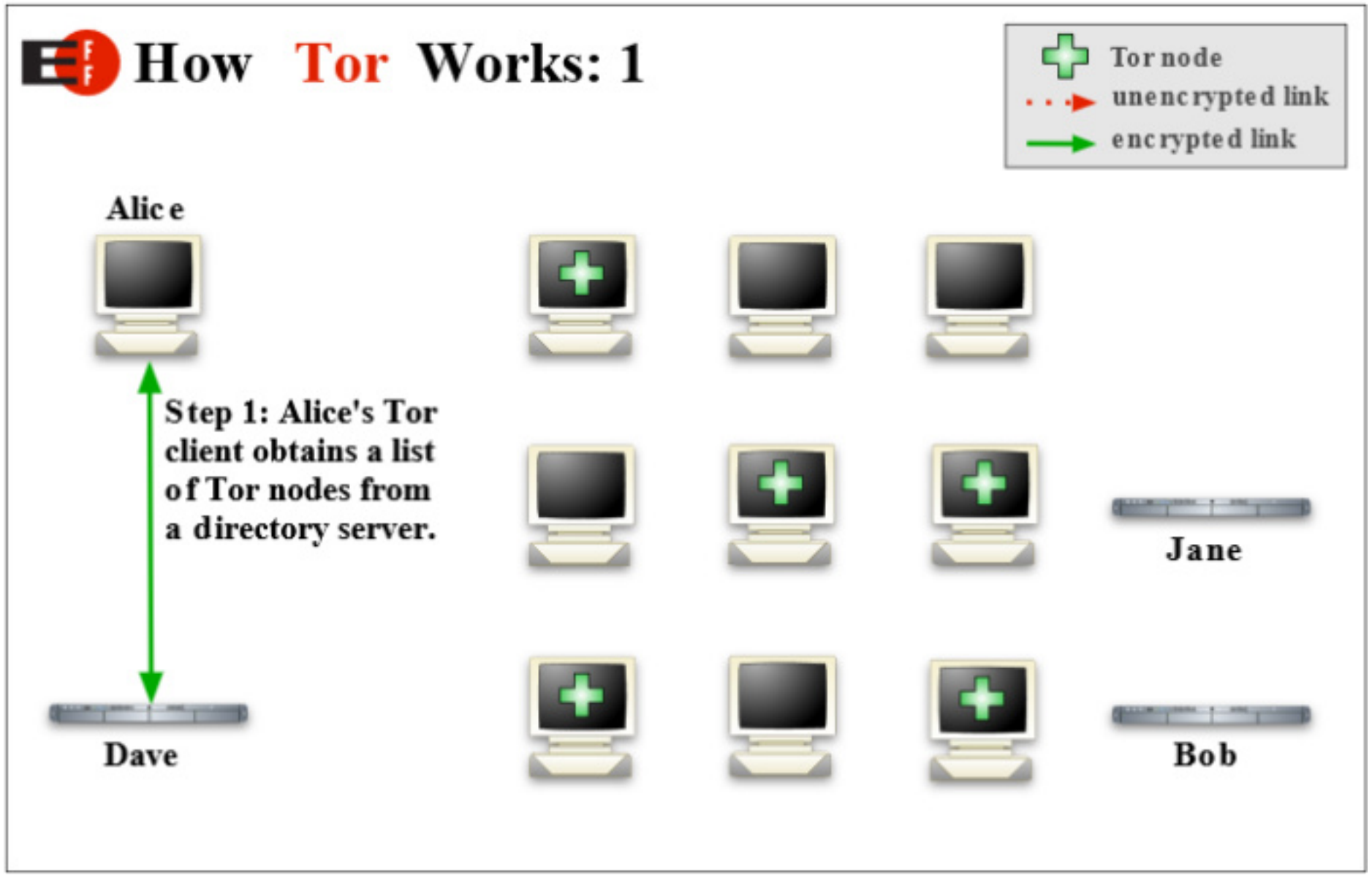
Half-Slide Summary: Dark Web

- The Internet is not “free” for all
  - Government regulations prohibit access to certain websites
  - Sedition laws prohibit certain topics or content
- People may want more privacy
  - Can we really trust soulless corporations to *ethically* collect and maintain data?
  - What if my neighbor sees my browsing history?
- We can use **The Onion Router** (TOR) to access the internet through a sequence of encrypted and relayed channels
  - TOR provides strong privacy *if* you assume no one entity controls the majority of networking nodes
  - TOR has brought with it an underground market of “Dark Web” sites that are used for illicit purposes

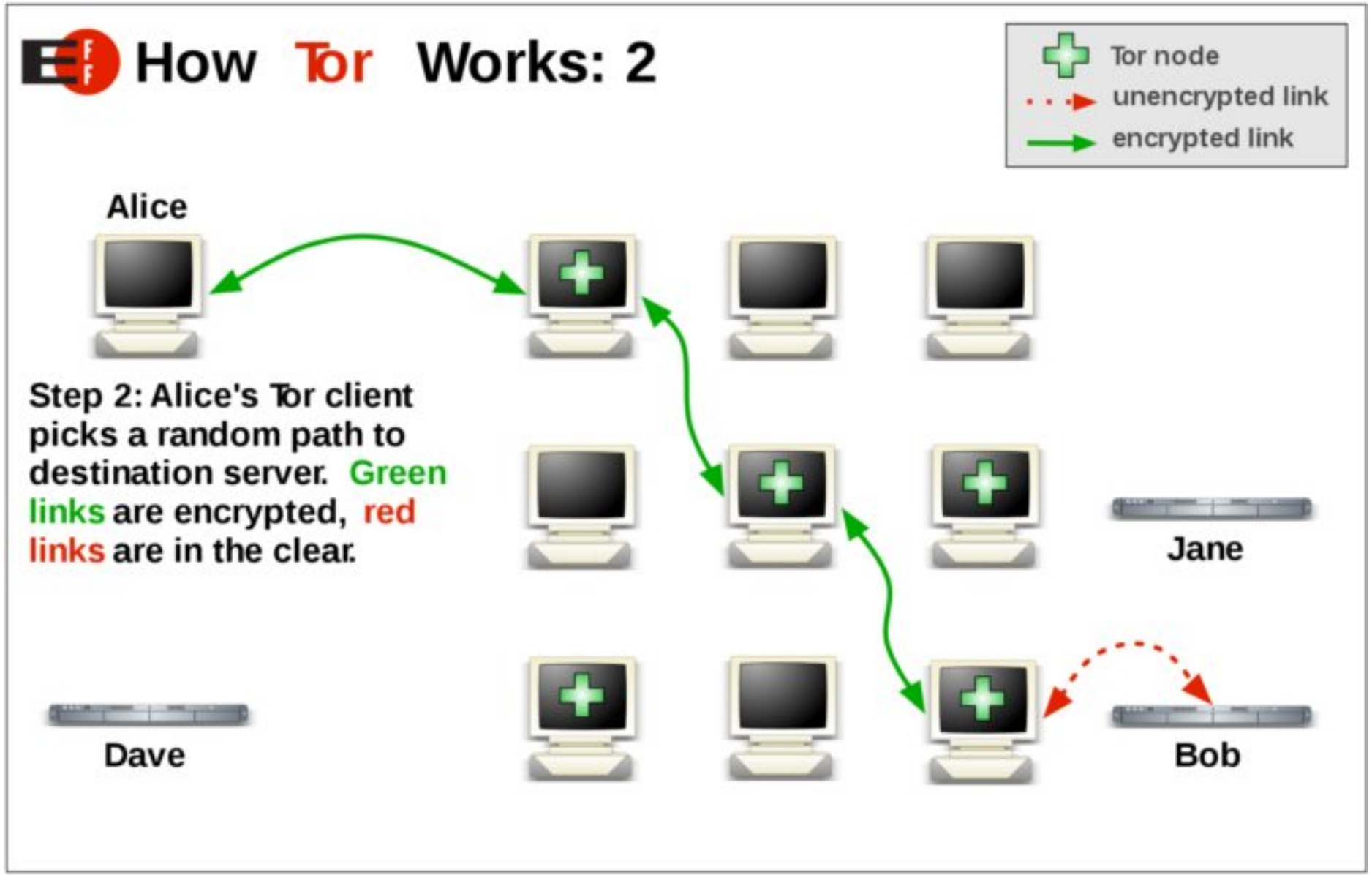
Tor

- **T**he **O**nion **R**outer
- Used to fight traffic analysis

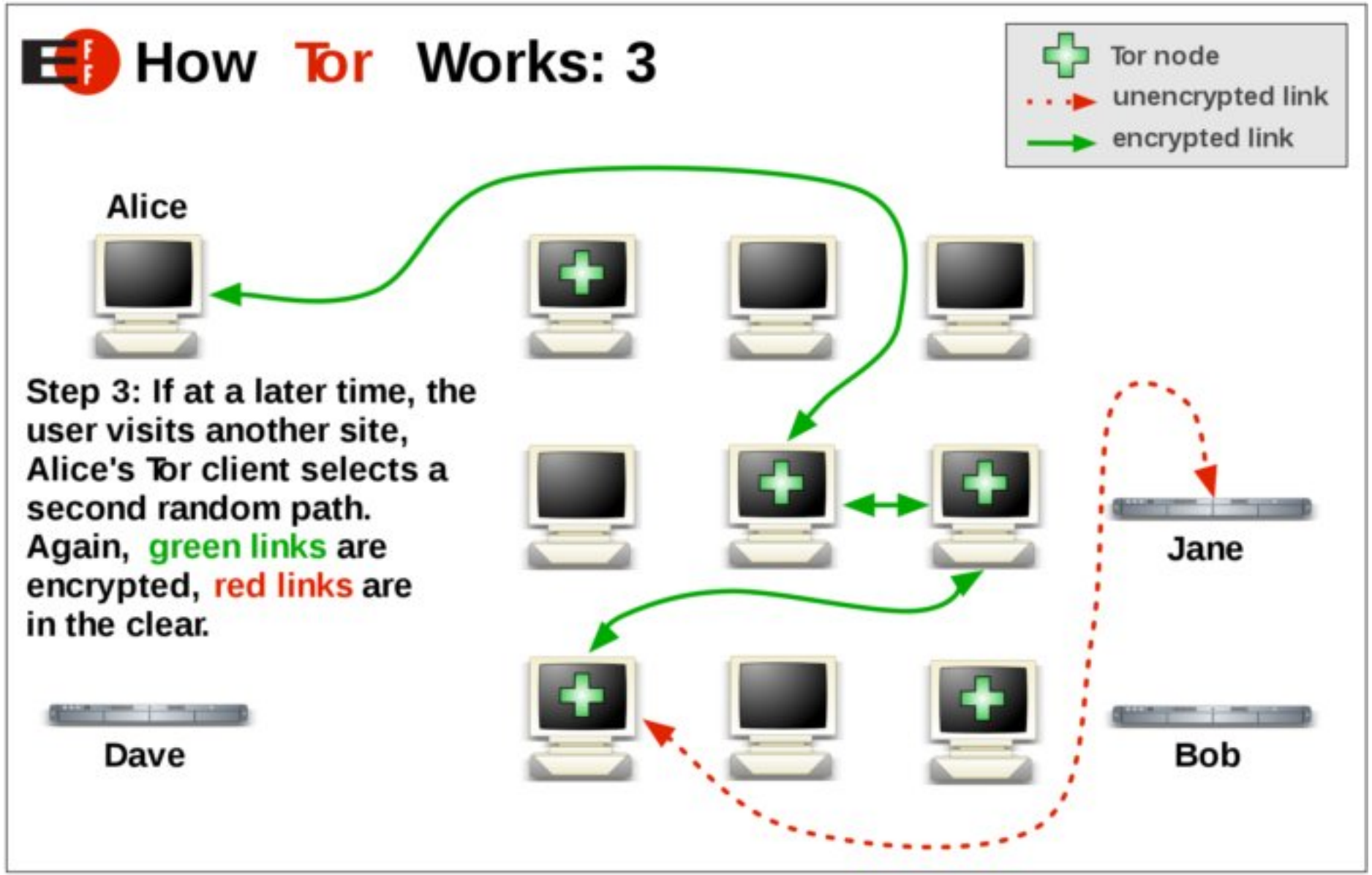
- Idea: like an onion, make your connection have layers
  - Your HTTP traffic gets routed through multiple nodes before hitting an **exit node** that connects for you
  - Hard to reconstruct who you are because each connection between each node is separately encrypted



More detail: https://2019.www.torproject.org/about/overview.html.en#thesolution



More detail: https://2019.www.torproject.org/about/overview.html.en#thesolution



More detail: https://2019.www.torproject.org/about/overview.html.en#thesolution

Tor servers

- Run by volunteers
- There's one in CSE

Tor

- Tor encrypts outgoing data multiple times, like layers of an onion
  - One encryption layer for each step in the relay
- Each layer of encryption is peeled off by a relay node in the network. At each layer:
  - Decrypt the current layer
  - Forward to the next destination

TOR vulnerabilities

- Only the first relay node knows the source IP address
- Only the last relay node knows the destination IP address
- To break anonymity, you need to surveil ALL nodes in the Tor circuit

- Also, all bets are off it you enable JS
  - Or really, any other browser fingerprinting – often, TOR clients will force you to use fixed-size browser windows, no JS, simple CSS, no posting, no cookies, etc. to prevent deanonymization
    - Also run it on a Live CD instance with a RAM disk rather than from durable storage…

Statistical correlation attack

- Let’s say that Elmo and Abby use Tor regularly
  - You control their ISP and collect frequent traffic logs with timestamps
- Also control the ISP for ILoveBigBird.com and DownWithBigBird.org
  - You also collect frequent traffic logs with timestamps
- How can you tell who is pro-BigBird vs anti-BigBird?
  - With enough log data, you can perform statistical correlation attack
  - Remember metadata from yesterday?

Statistical correlation attack: prevention

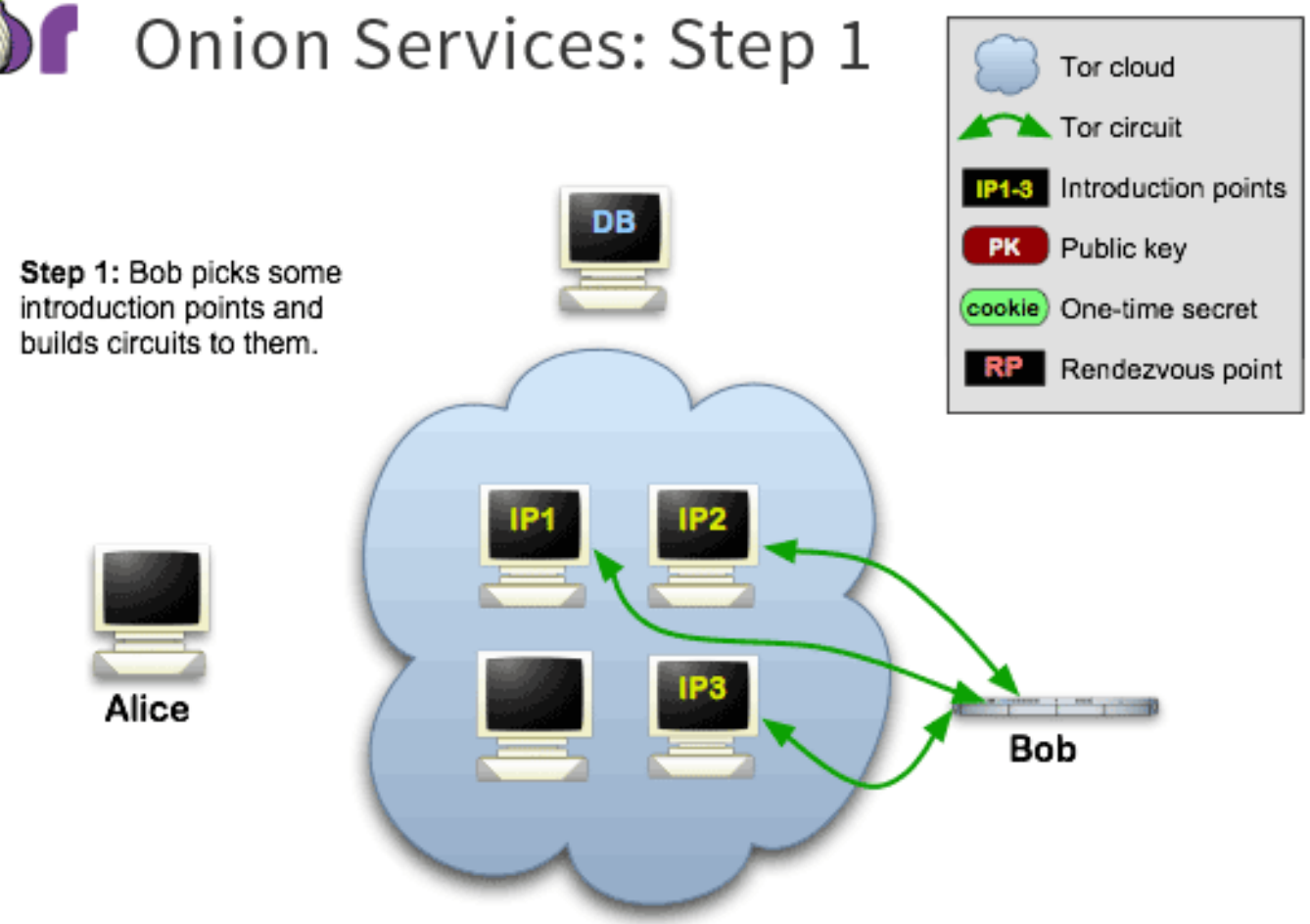
- Elmo and Abby always transmit data to Tor once per second
- If no data to send, just send NULL or random data
  - Always attempt to give the appearance of traffic so an eavesdropper can't be sure what you're doing…

- *More Tor* users conducting *more activity* on Tor reduces vulnerability

Tor services

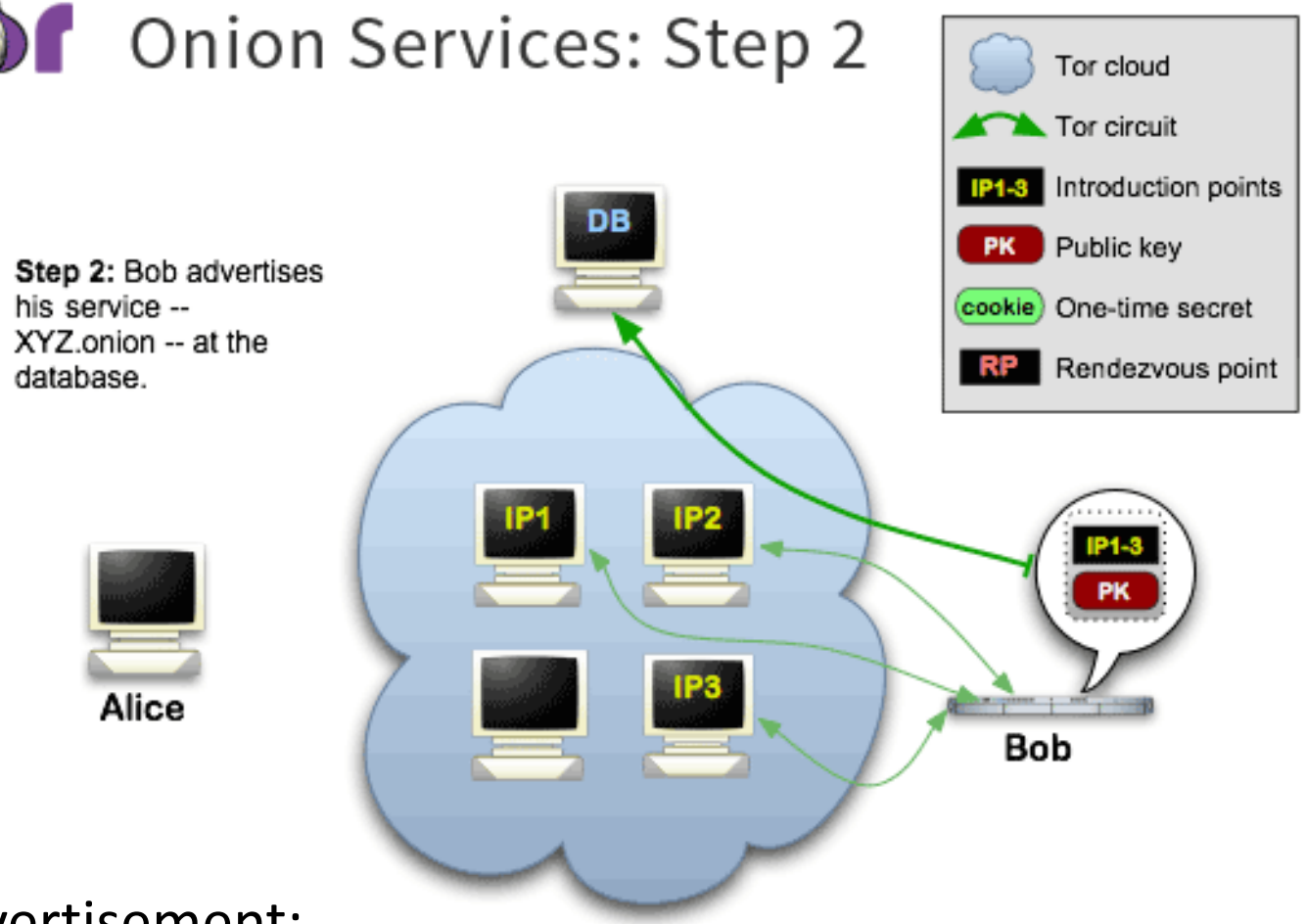
- Tor allows users to anonymously publish services
  - E.g., web pages, or a chat server
  - The exit node and target host connect normally… this sounds bad
- Service locations (*rendezvous points*) must be known to clients
  - Even though censors may want to locate, take down services
- Key idea: layer of indirection
  - *Introduction points* are Tor nodes that relay traffic from clients to services

Onion Services: Step 1



More detail: https://2019.www.torproject.org/docs/onion-services.html

Onion Services: Step 2



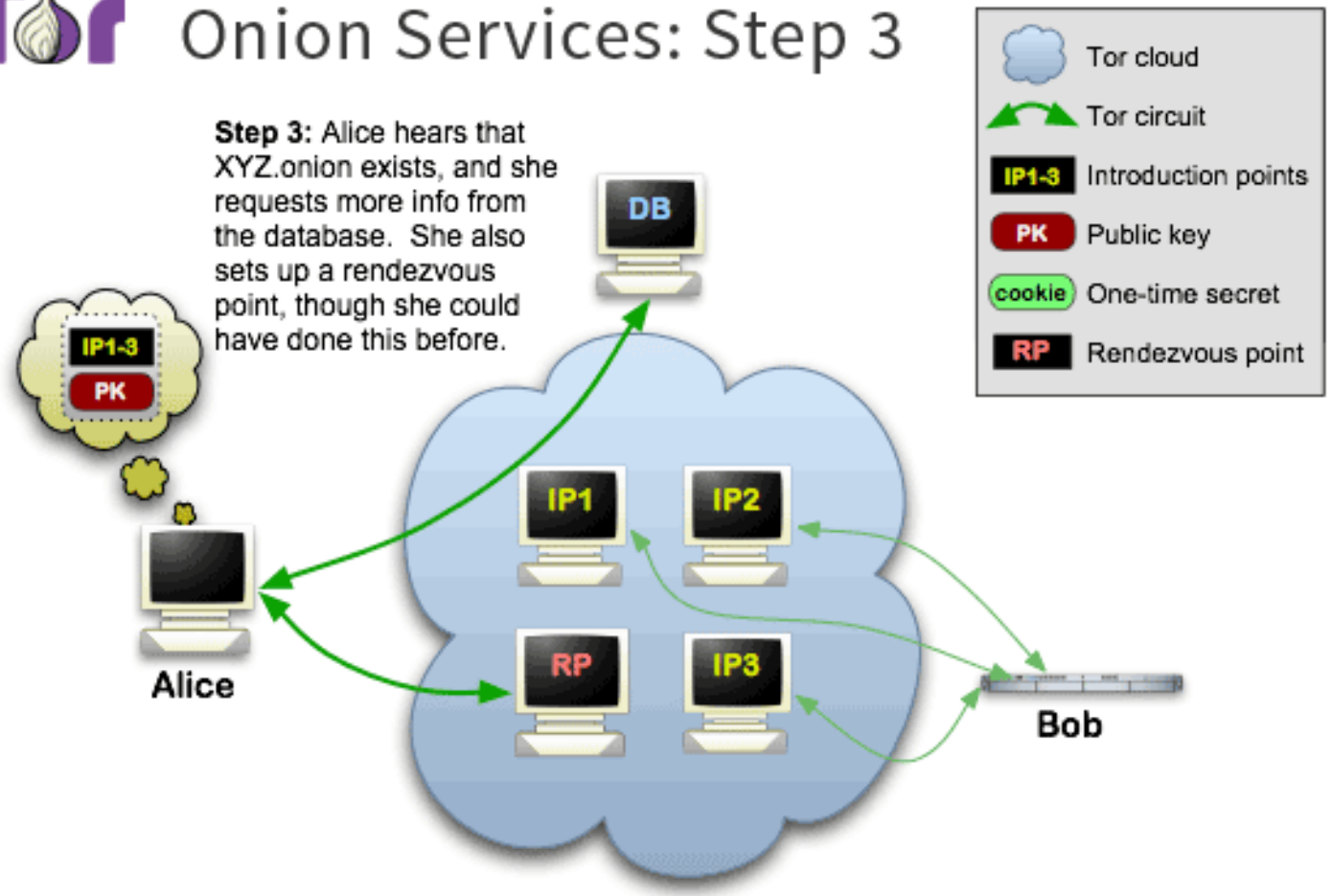
The advertisement:

- Contains Bob's public key
- Contains each intro node
- Is signed with Bob's private key

XYZ is an autogenerated name derived from Bob's public key

More detail: https://2019.www.torproject.org/docs/onion-services.html

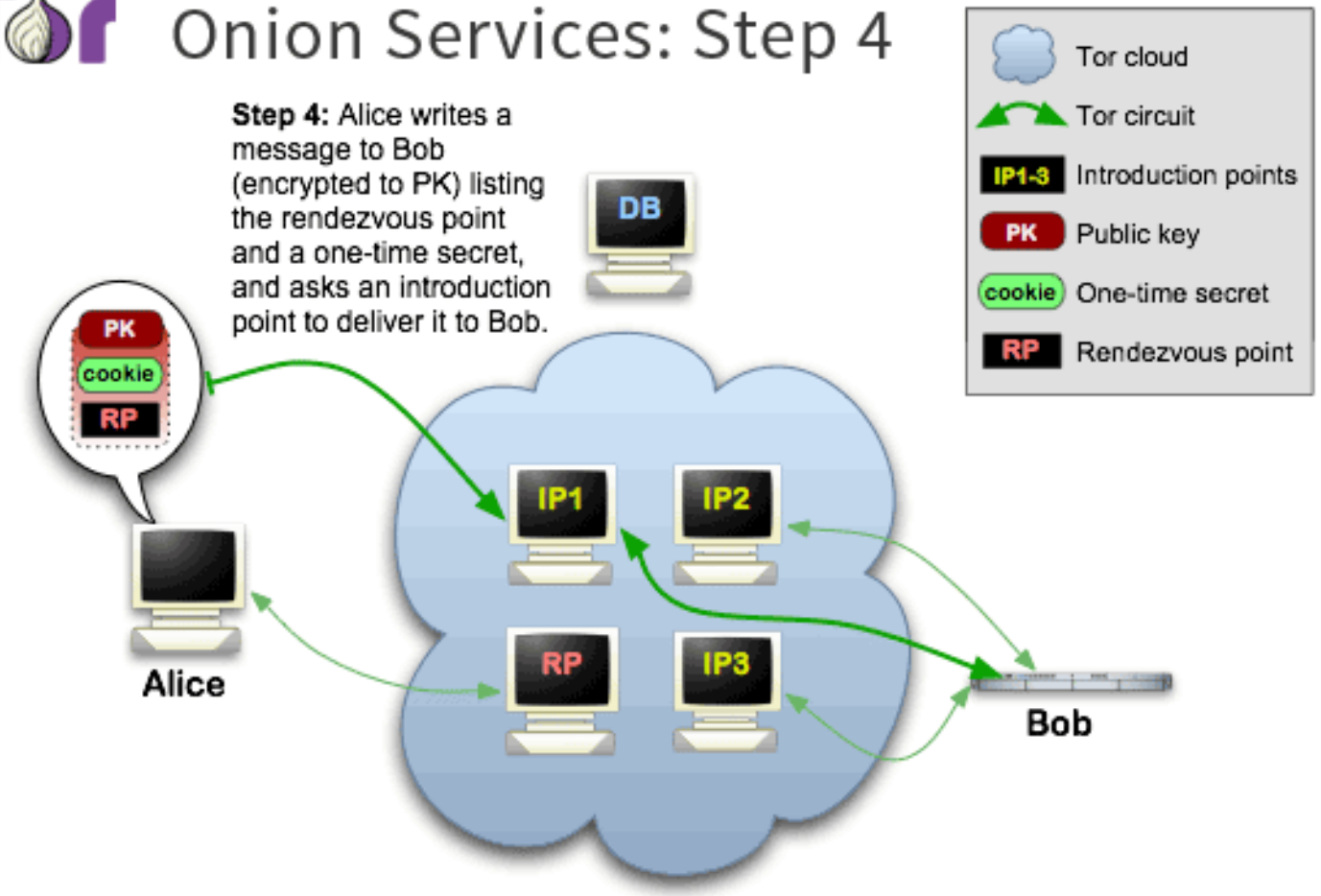
Onion Services: Step 3



- The downloaded advertisement record tells Alice where to find *Introduction Points*.
- She further chooses a random Tor node to act as a *Rendezvous Point*, and connects to it.

More detail: https://2019.www.torproject.org/docs/onion-services.html

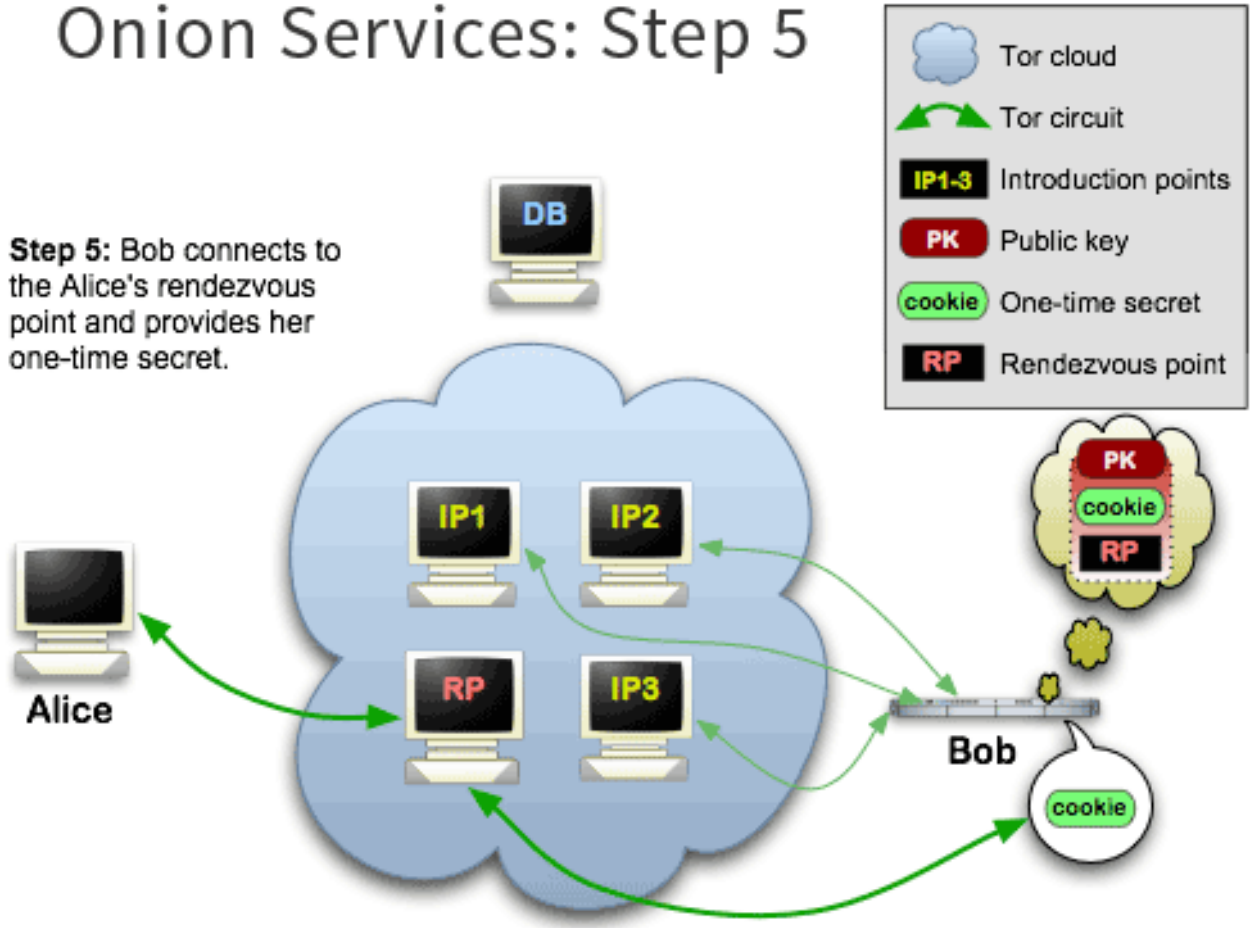
Onion Services: Step 4



All links to *Introduction Points* and *Rendezvous Points* are encrypted, and via Tor; no one can connect the message to Alice's IP address.

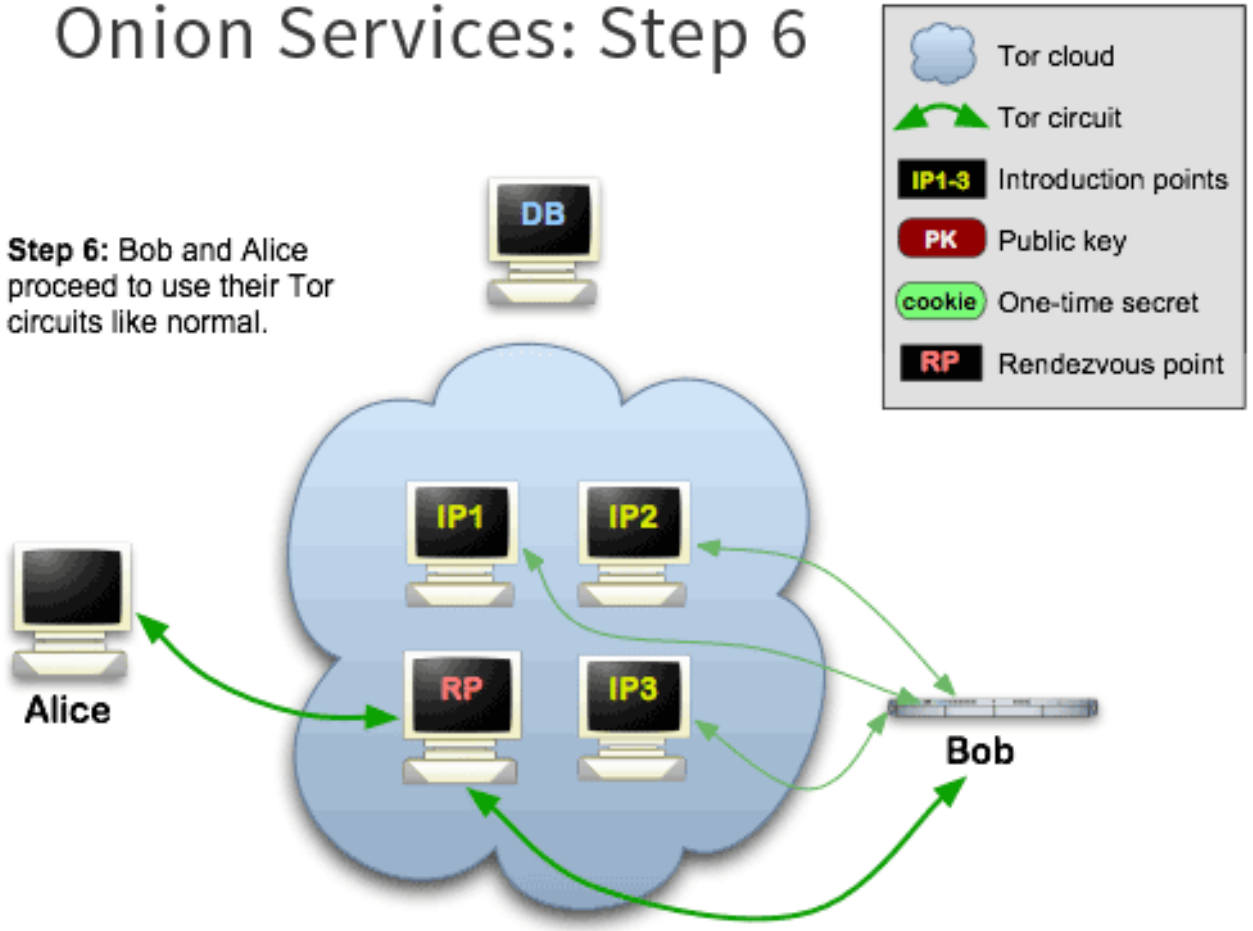
More detail: https://2019.www.torproject.org/docs/onion-services.html

Onion Services: Step 5



More detail: https://2019.www.torproject.org/docs/onion-services.html

Onion Services: Step 6



- We now have a connection where neither Alice nor Bob know each other's IP addresses
- Why bother with Rendezvous Points? No Tor node should appear to be exclusively responsible for a service.

More detail: https://2019.www.torproject.org/docs/onion-services.html

Distributed hash tables

- How can Alice find the XYZ.onion record in the first place?
- One answer: Ask a directory server
  - Centralized, so easy to take down
- Another answer: Hash-based mapping
  - If N servers, store file *foo* on server *hash(foo) % N*
  - What if we need to add a server?
  - File is now mapped to server *hash(foo) % (N+1)*

Finding Tor services

- Curated lists
  - Reddit
  - Hidden Wiki
  - ... and many others
- Hidden service search engines
  - Ahmia
  - Torch
  - Not Evil
  - ... and many others
- Search engine crawlers on dark web
  - Route crawler GET requests to .onion sites through Tor

Course Wrap Up!

- Thanks for a great semester

- Think of all you have accomplished. Your resume has gone up a level!
  - HTML, CSS front-end
  - Flask-based Python server
  - AWS integration
  - JS and React, complex PL features and asynchronous programming
  - MapReduce Framework from scratch!
  - Raw sockets, threading, reliability and fault tolerance
  - Search engine and IR
  - Databases, SQLite
  - Bash scripting, VMs, containers, virtual environments, Linux utilities

- Completed at *double* pace and entirely *remotely*!
  - This is an achievement you should be proud of