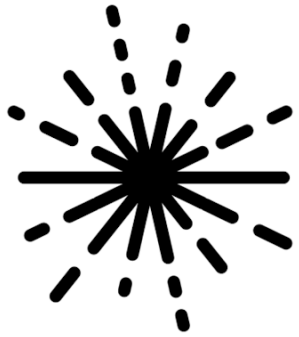


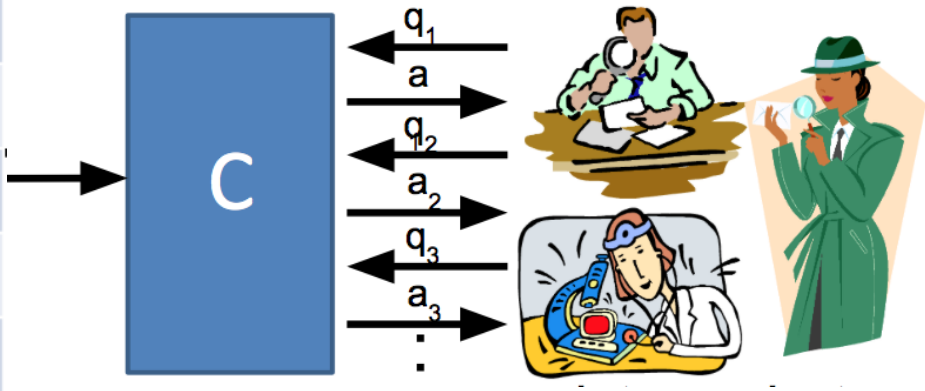
CS208: Applied Privacy for Data Science
The Local Model: Implementations

School of Engineering & Applied Sciences
Harvard University

April 5, 2022

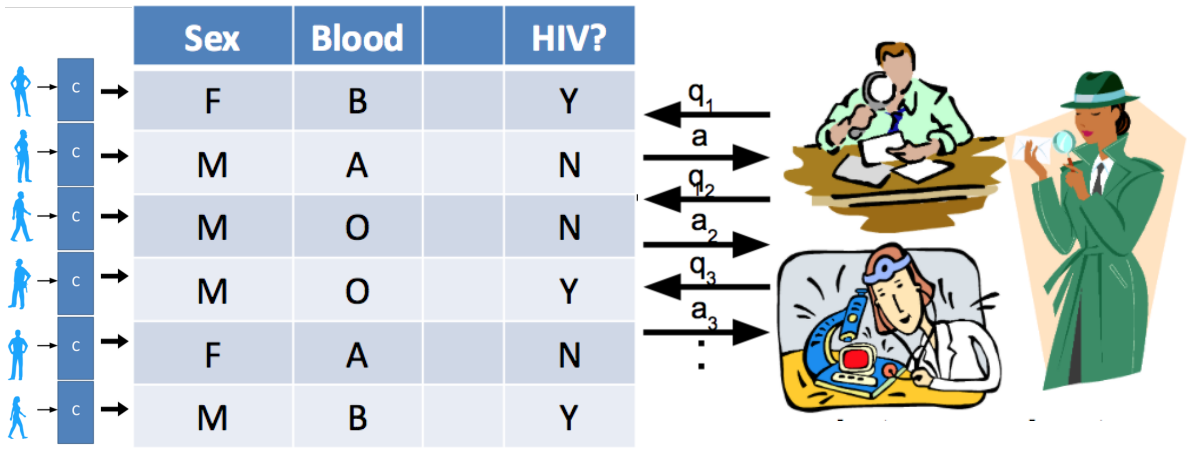
Central Model

Sex	Blood		HIV?
F	B		Y
M	A		N
M	O		N
M	O		Y
F	A		N
M	B		Y



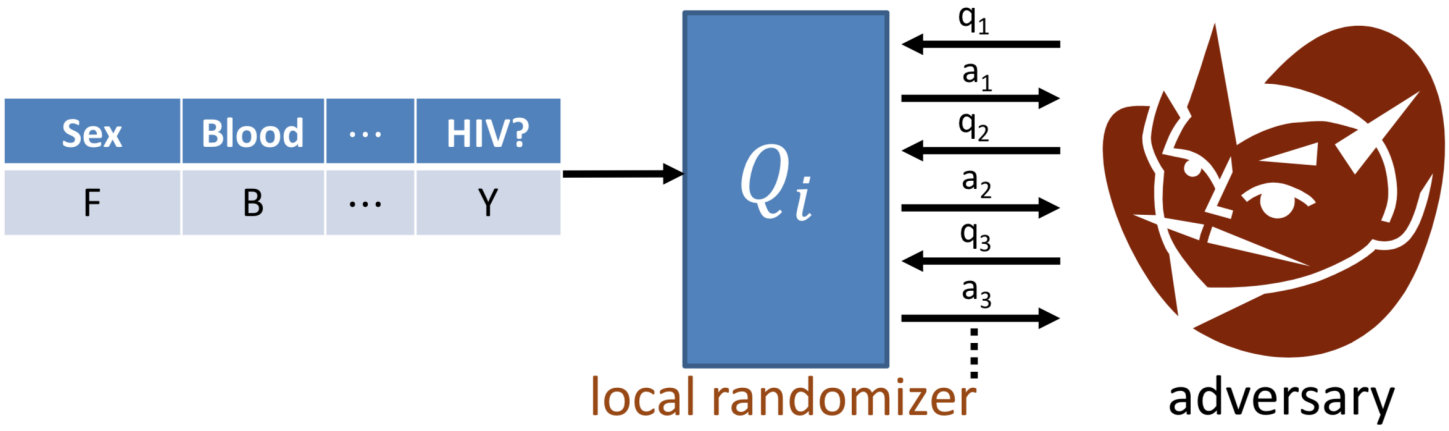
Data Repository Curator Analysts

Local Model



Subjects Repository Analysts

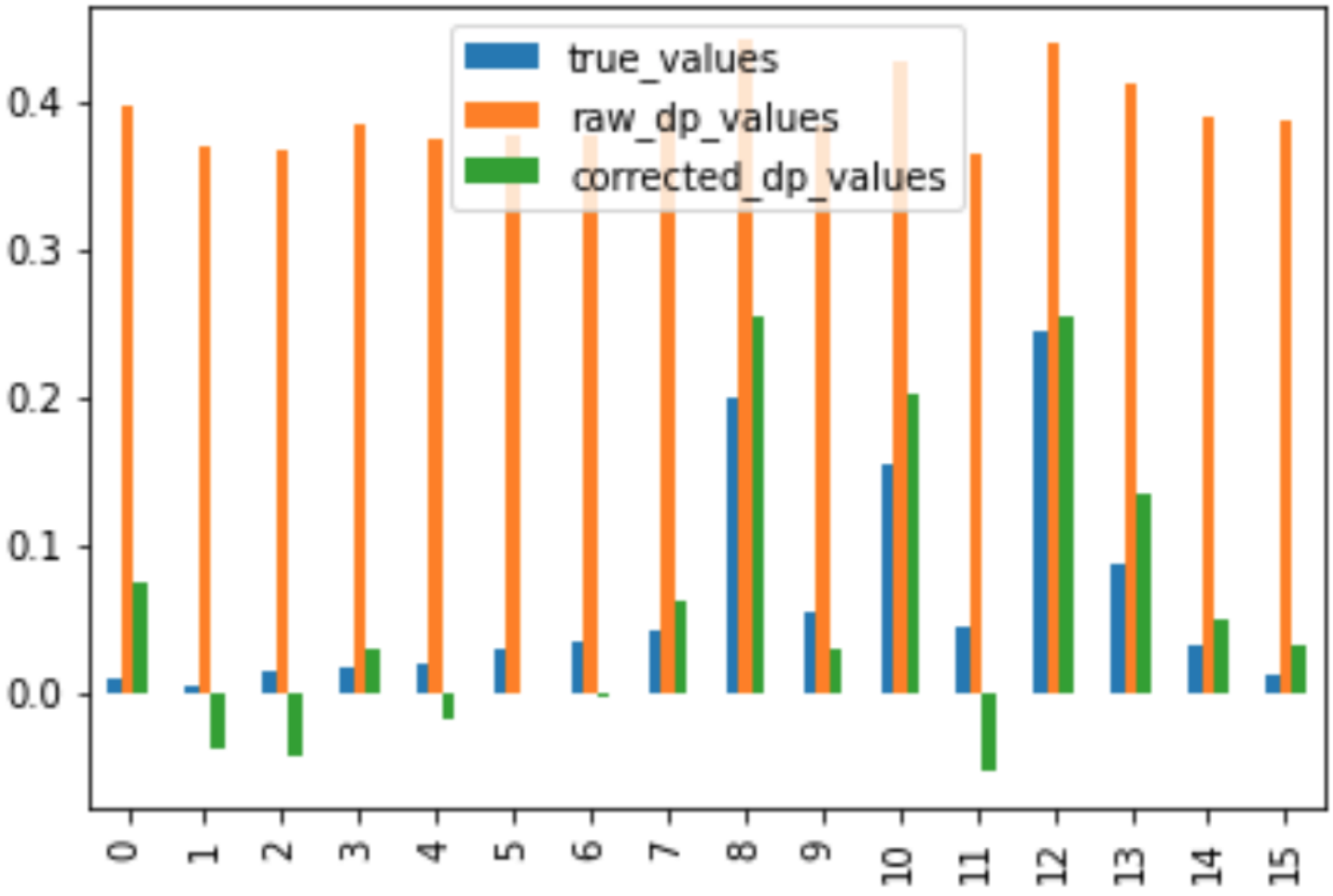
Local DP



Require: for all i, x_i, x_i' ~~differing on one row,~~ all strategies A

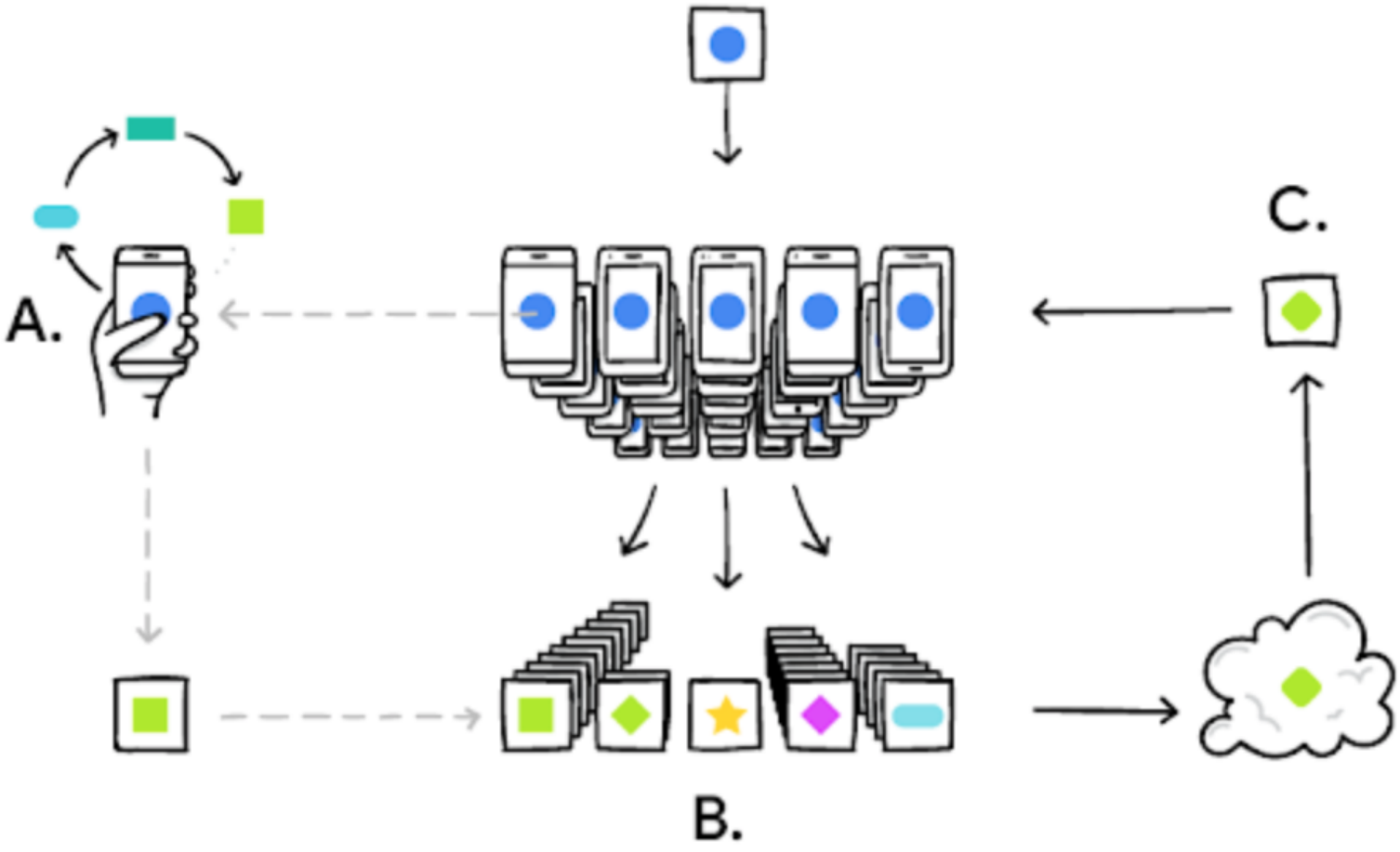
$$\Pr[A \text{ outputs YES after interacting w}/Q_i(x_i)]$$
$$\leq e^{\varepsilon} \cdot \Pr[A \text{ outputs YES after interacting w}/Q_i(x_i')] + \delta$$

Local Model for Integers/Histograms



$\epsilon = 1, N = 2,000$

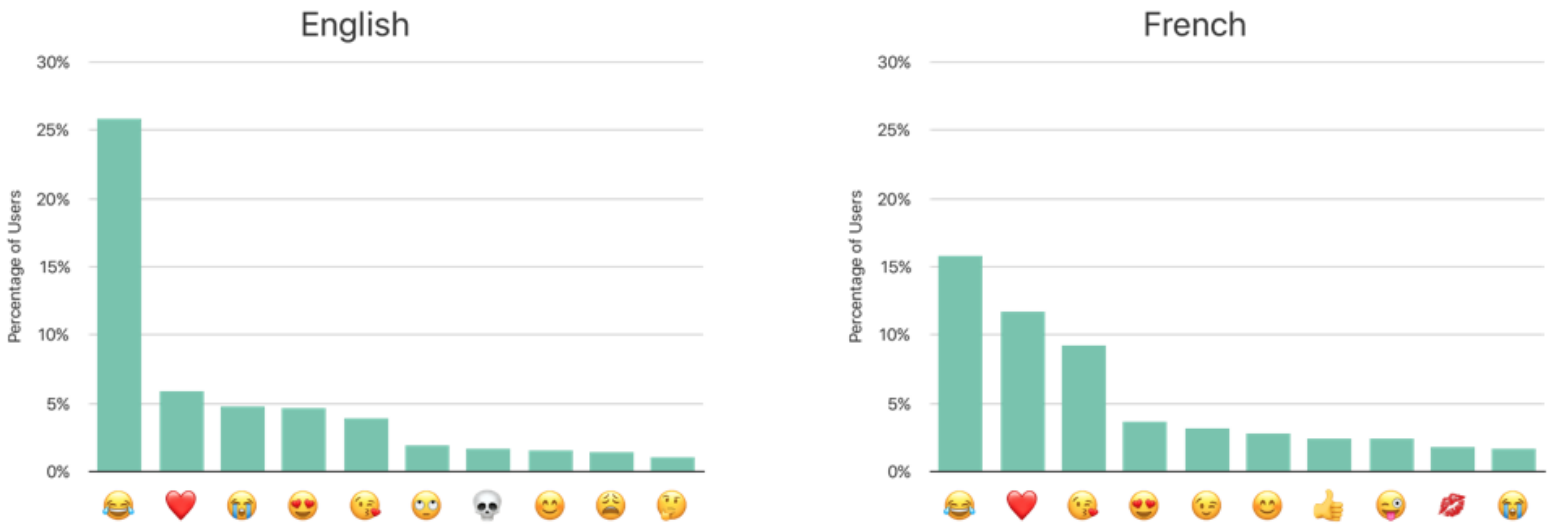
Federation of Histograms



from: `https://ai.googleblog.com/2017/04/federated-learning-collaborative.html`

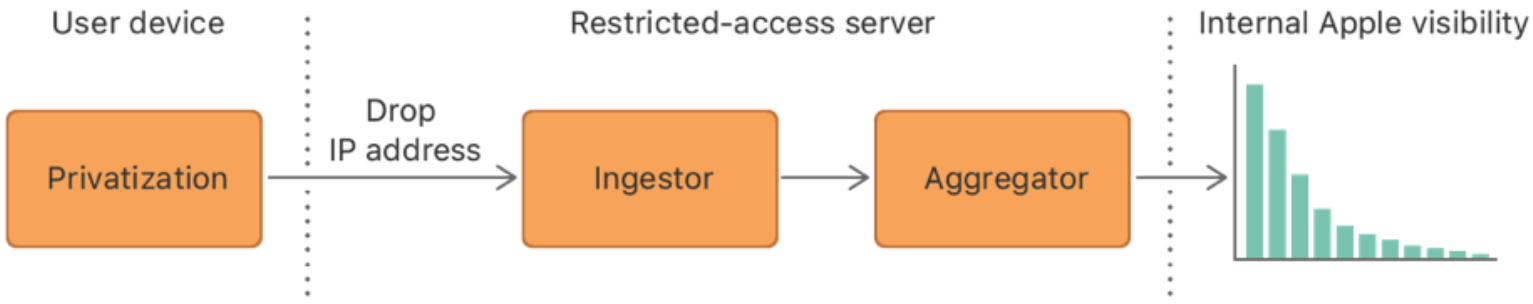
Federation of Histograms

Figure 6. Emojis in Different Keyboard Locales.



Federation of Histograms

Figure 1. System Overview.



Discovering Unknown Values (Apple)

- Method described so far requires server to decide on a small set values $v_1, \ldots, v_s$ to estimate frequencies of.
- Goal: find unanticipated frequent values v
- Idea: reconstruct v one symbol at a time
  - do RR on $h(x_i)||x_i[j]$ for each bitposition $j = 1, \ldots, \log D$.
  - $\hat{f}[w||\sigma_j]$ is large ⇒ there probably is a frequently occurring value $v \in \{1, \ldots, D\}$ such that $h(v) = w$ and $v[j] = \sigma_j$.
  - $\hat{f}[w||\sigma_1], \ldots, \hat{f}[w||\sigma_{\log D}]$ large ⇒ $v = \sigma_1 \cdots \sigma_{\log D}$

Discovering Unknown Values (Apple)

Algorithm 2 Client-Side $\mathcal{A}_{\mathsf{client-CMS}}$

Require: Data element: $d \in \mathcal{D}$; ϵ, $\mathcal{H}$.

1. Sample j uniformly at random from $[k]$.
2. Initialize a vector $\boldsymbol{v} \leftarrow -\mathbf{1} \in \mathbb{R}^m$.
3. Set $v_{h_j(d)} \leftarrow 1$.
4. Sample $\boldsymbol{b} \in \{-1, +1\}^m$, where each b_ℓ is i.i.d. where $\Pr[b_\ell = +1] = \frac{e^{\epsilon/2}}{e^{\epsilon/2}+1}$.
5. $\tilde{\boldsymbol{v}} \leftarrow (v_1 b_1, \cdots, v_m b_m)$.
6. **return** $\tilde{\boldsymbol{v}}$, index j.

Algorithm 9 Client-Side $\mathcal{A}_{\mathsf{client-SFP}}$

Require: String: $\boldsymbol{s} \in \mathcal{D}$; privacy parameters: (ϵ, ϵ'), hash functions $(\mathcal{H}, \mathcal{H}')$, and h with output size 256.

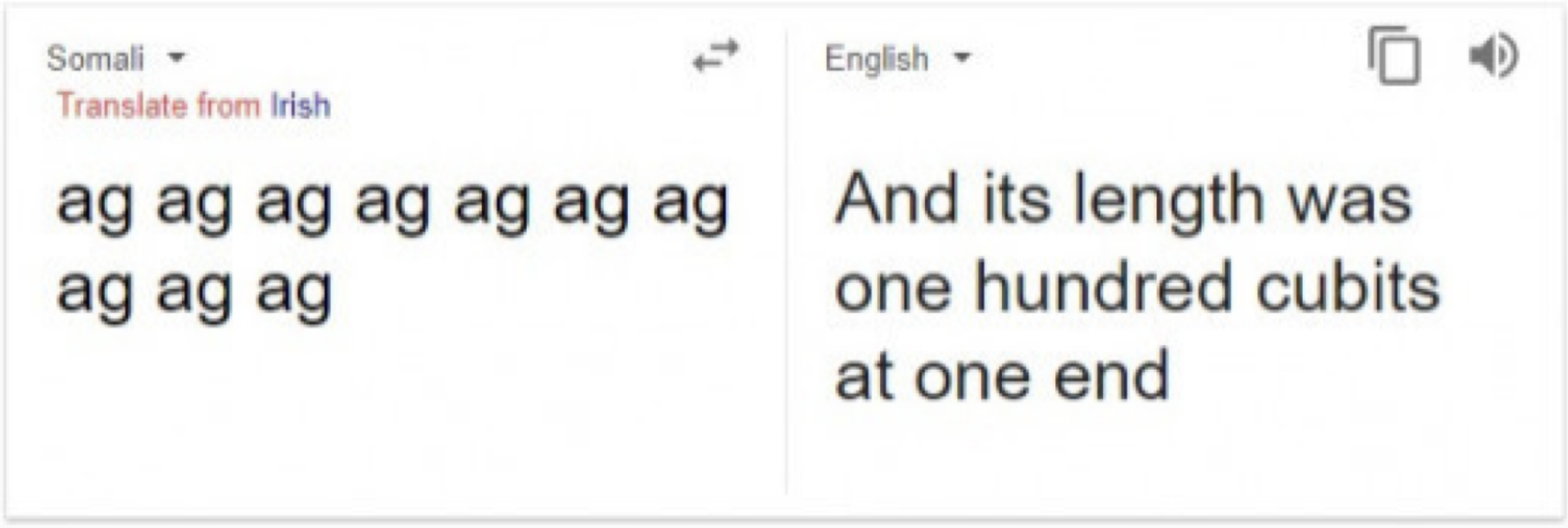
1. Sample ℓ uniformly at random from $\{1, 3, 5, 7, 9\}$.
2. Set $\boldsymbol{r} \leftarrow h(\boldsymbol{s}) \| \boldsymbol{s}[\ell : \ell + 1]$.
3. **return** $(\mathcal{A}_{\mathsf{client-CMS}}(\boldsymbol{r}, \epsilon', \mathcal{H}'), \mathcal{A}_{\mathsf{client-CMS}}(\boldsymbol{s}, \epsilon, \mathcal{H}), \ell)$.

Discovering Unknown Values (Apple)

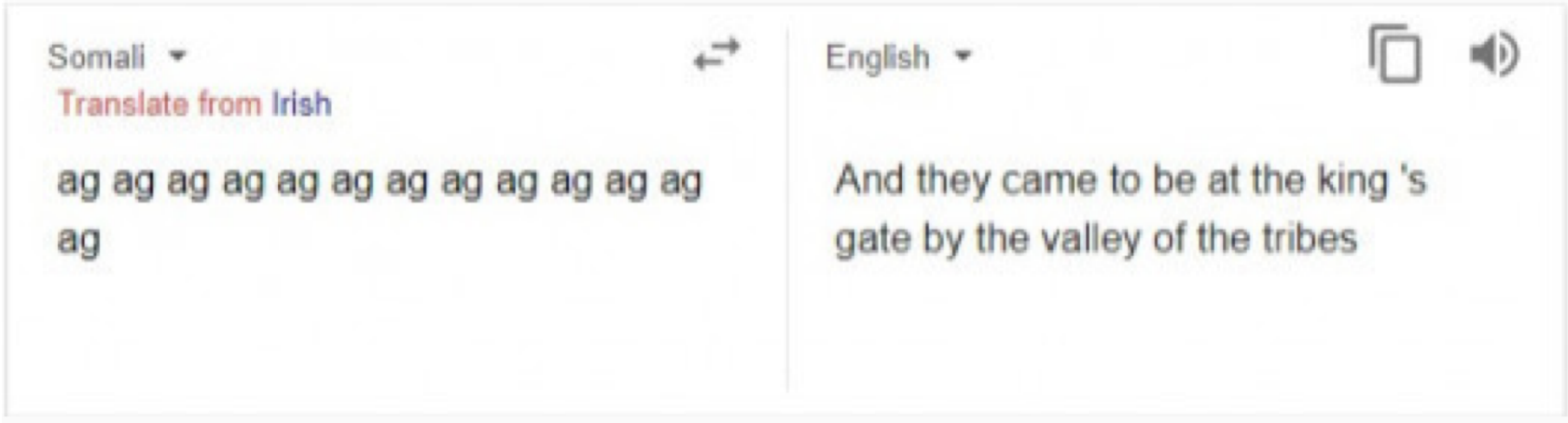
```
Actual Names:
setosa
 versicolor
 virginica
      temp.l
temp.b   1   2   3   4   5   6   7   8   9  10
         0   0   0   0   0   0 264 236 220 316
     a   0   0   0   0   0 247   0   0  89   0
     c   0   0   0   0   0  79   0  98   0   0
     e   0 354   0   0   0   0   0   0   0   0
     g   0   0   0  99   0   0   0   0   0   0
     i   0  72   0   0 188   0  80   0   0   0
     l   0   0   0   0   0   0   0 105   0   0
     n   0   0   0   0   0  82   0   0   0   0
     o   0   0   0 219   0   0 101   0  87   0
     r   0   0 190   0   0   0   0   0   0  98
     s 244   0   0  81 250   0   0   0   0   0
     t   0   0 243   0   0   0   0   0   0   0
     v 185   0   0   0   0   0   0   0   0   0
 [1] "s" "e" "t" "o" "s" "a" ""  ""  ""  ""
      temp.l
temp.b   1   2   3   4   5   6   7   8   9  10
         0   0   0   0   0   0 104  94  89 324
     a   0   0   0   0   0 125   0   0 228   0
     c   0   0   0   0   0 251   0 254   0   0
     e   0 335   0   0   0   0   0   0   0   0
     g   0   0   0 238   0   0   0   0   0   0
     i   0 243   0   0 493   0 231   0   0   0
     l   0   0   0   0   0   0   0 260   0   0
     n   0   0   0   0   0 234   0   0   0   0
     o   0   0   0  86   0   0 258   0 260   0
     r   0   0 488   0   0   0   0   0   0 234
     s 100   0   0 222  92   0   0   0   0   0
     t   0   0  81   0   0   0   0   0   0   0
     v 519   0   0   0   0   0   0   0   0   0
 [1] "v" "e" "r" "g" "i" "c" "o" "l" "o" ""
```

Federated Learning

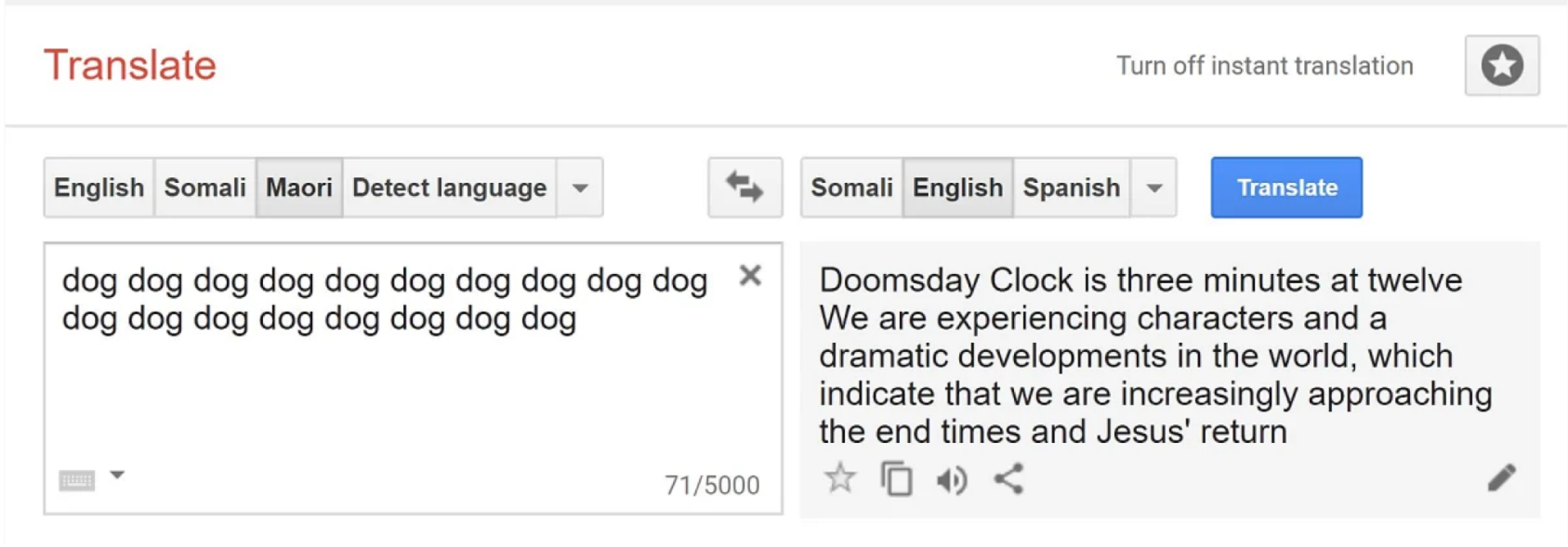
WHEN YOU TRAIN PREDICTIVE MODELS ON INPUT FROM YOUR USERS, IT CAN LEAK INFORMATION IN UNEXPECTED WAYS.



from Ilya Mironov



from Ilya Mironov



from Ilya Mironov

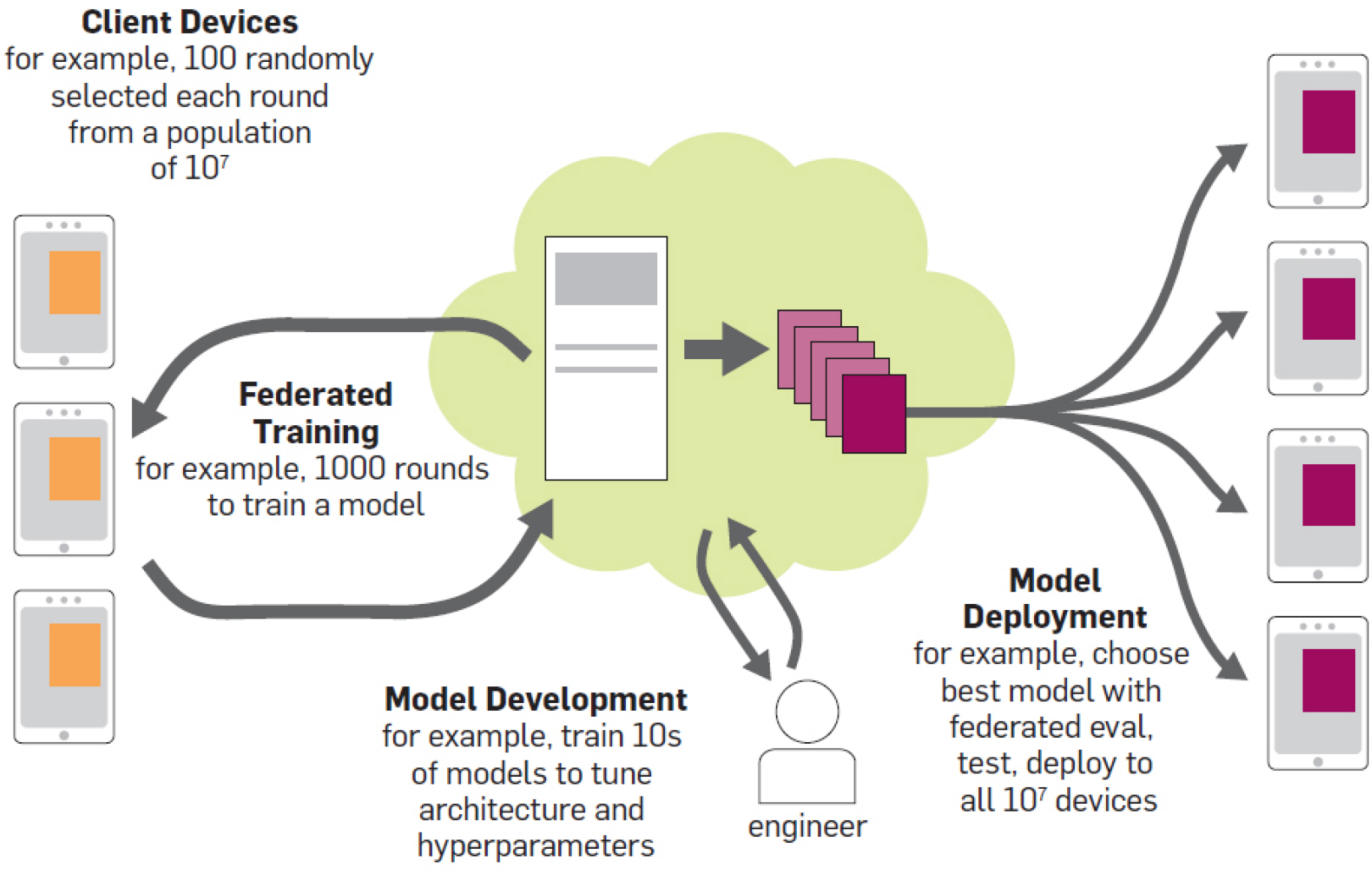
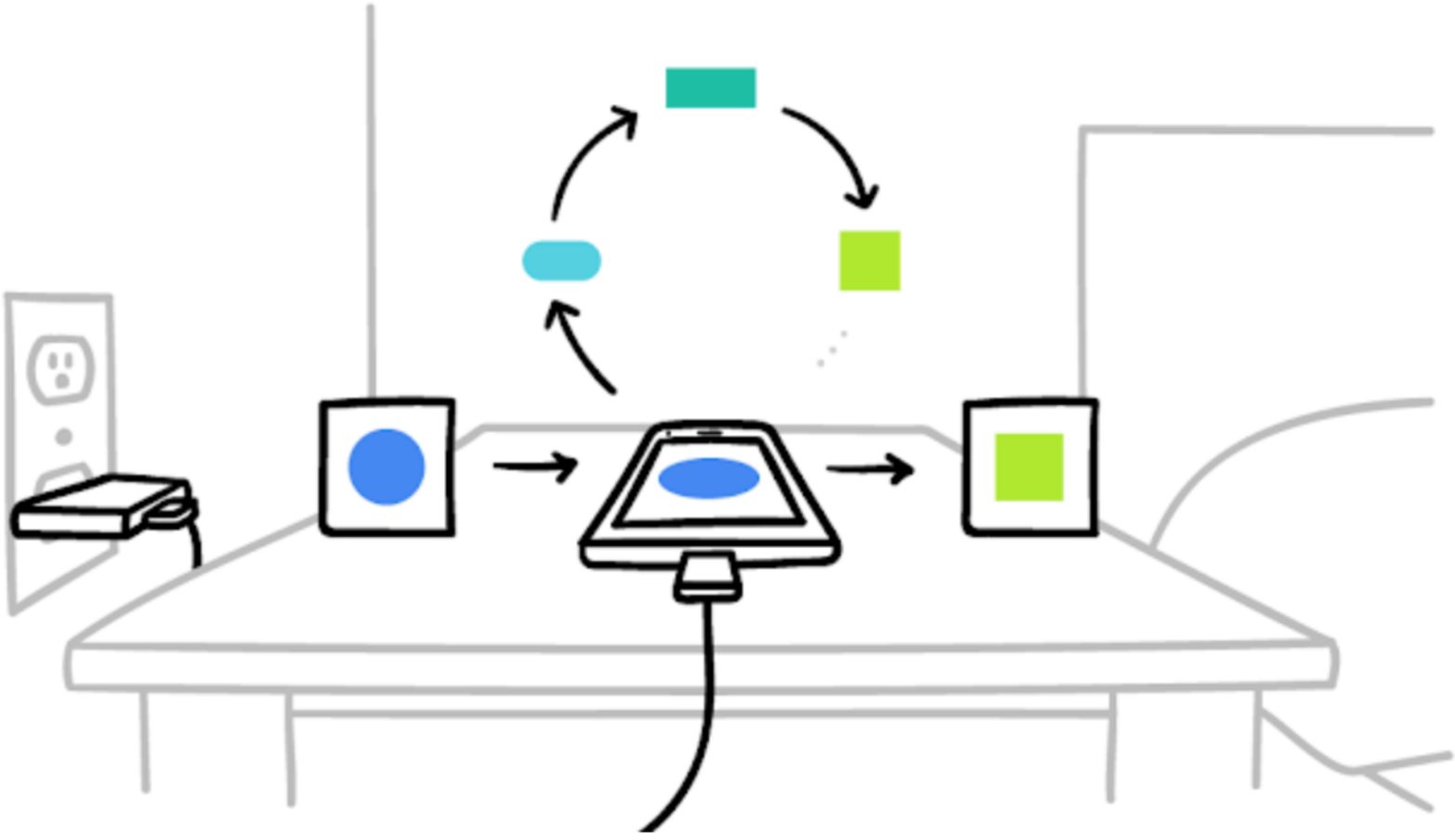

Client Devices
for example, 100 randomly selected each round from a population of 10^7
Federated Training
for example, 1000 rounds to train a model
Model Development
for example, train 10s of models to tune architecture and hyperparameters
engineer
Model Deployment
for example, choose best model with federated eval, test, deploy to all 10^7 devices

TABLE 1: **TYPICAL FL SETTINGS AND OF TRADITIONAL DISTRIBUTED LEARNING**

	DATACENTER DISTRIBUTED LEARNING	CROSS-SILO FEDERATED LEARNING	CROSS-DEVICE FEDERATED LEARNING
Setting	Training a model on a large but "flat" dataset. Clients are compute nodes in a single cluster or datacenter.	Training a model on siloed data. Clients are different organizations (e.g., medical or financial) or datacenters in different geographical regions.	The clients are a very large number of mobile or IoT devices.
Data distribution	Data is centrally stored, so it can be shuffled and balanced across clients. Any client can read any part of the dataset.	Data is generated locally and remains decentralized. Each client stores its own data and cannot read the data of other clients. Data is not independently or identically distributed.	
Orchestration	Centrally orchestrated.	A central orchestration server/service organizes the training, but never sees raw data.	
Distribution scale	Typically 1 - 1000 clients.	Typically 2 - 100 clients.	Up to 10^{10} clients.
Client properties	Clients are reliable and almost always available to participate in computations. Clients may be directly addressed, and can maintain state across computation rounds.		Clients are often unavailable and can only be accessed by random sampling from available devices. For large populations a single client will typically only participate once in a given computation.



Your phone participates in Federated Learning only
when it won't negatively impact your experience.

from: https://ai.googleblog.com/2017/04/federated-learning-collaborative.html