# Privacy Advances in Machine Learning Systems

Katharine Jarmul
O'Reilly AI London
kjamistan

# When did consumers become concerned about privacy and computing?

# From Understanding Privacy Concerns (1992)

A 1990 Louis Harris survey commissioned by Equifax, for instance, found 71 percent of the respondents believed consumers "have lost all control over how personal information about them is used by companies"). More recently, a 1991 Gallup survey found **78 percent of the respondents described themselves as "very concerned" or "somewhat concerned" about what marketers know about them.**

Nowak et al., 1992.

How and when were people *actually* affected by privacy-unaware data collection?

# Privacy Issues in Knowledge Discovery and Data Mining (2000)

Despite collecting over $16 million USD by selling the driver-license data from 19.5 million Californian residents, the Department of Motor Vehicles in California revised its data selling policy after Robert Brado used their services to obtain the address of actress Rebecca Schaeffer and later killed her in her apartment.

Brankovic et al., 2000.

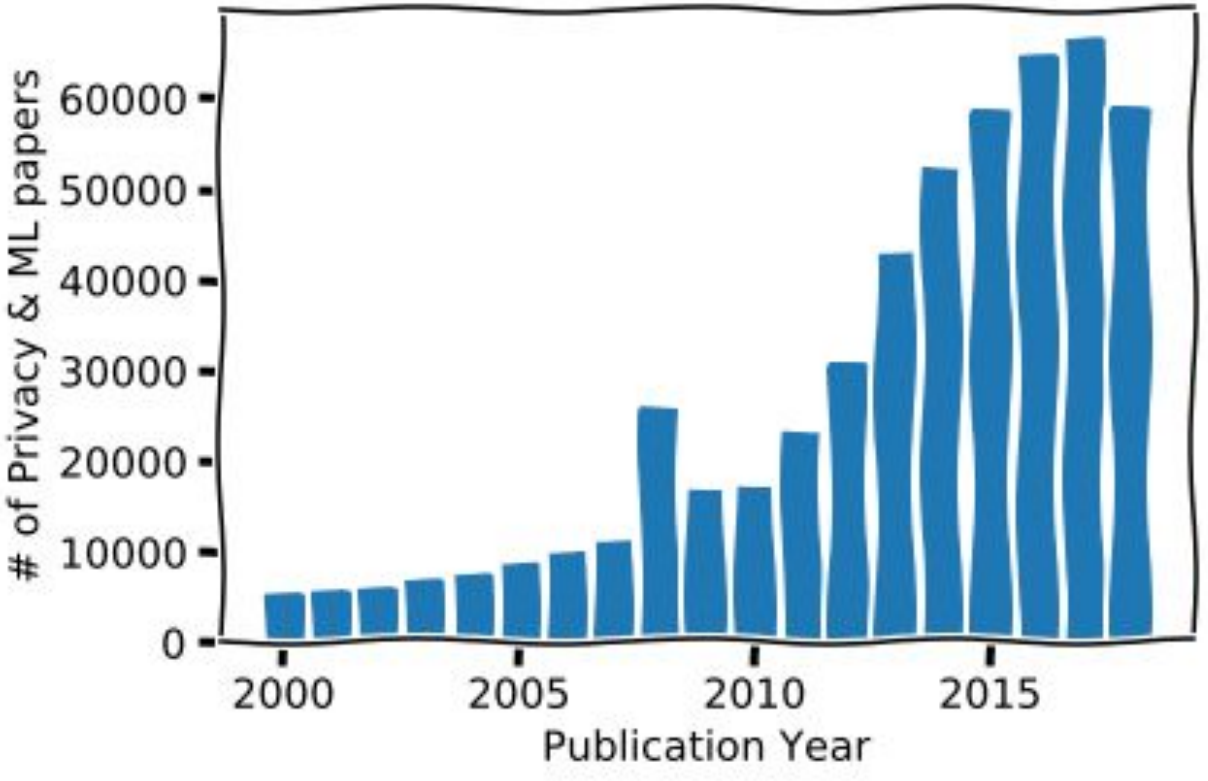
# What do machine learning and cryptography have in common?

# From Cryptography and Machine Learning (1988)

**Machine learning and cryptanalysis can be viewed as “sister fields,”** since they share many of the same notions and concerns. In a typical cryptanalytic situation, the cryptanalyst wishes to "break" some cryptosystem. Typically this means **he wishes to find the secret key used by the users of the cryptosystem, where the general system is already known.** The decryption function thus comes from a known family of such functions (indexed by the key), and the goal of the cryptanalyst is to exactly identify which such function is being used. **This problem can also be described as the problem of "learning an unknown function" (that is, the decryption function) from examples of its input/output behavior and prior knowledge** about the class of possible functions.

Rivest, 1988.

# Privacy in ML



# Defining the Problem

# Threat Model:

- Private Data Collection & Storage?
- Sharing Private Data for Training?

- Exposing Private Data via Queries or Model Access?
- Private Predictions?

# Notable Past Work

# Timeline

1978 - Concept of Homomorphic Encryption
1982 - Data Swapping
1998 - K-Anonymity
2003 - Tor Project Publicly Released
2005 - Personal Search Results (Google)
2006 - Differential Privacy
2009 - Differentially Private Logistic Regression
2010 - Full Homomorphic Encryption

# Homomorphic Encryption

Partially Homomorphic (PHE)

- Additive or multiplicative

Somewhat Homomorphic (SWHE)

- Addition and multiplication, but limited # of ops

Fully Homomorphic (FHE)

- Addition, multiplication for unbound # of ops

# Distributed Clustering

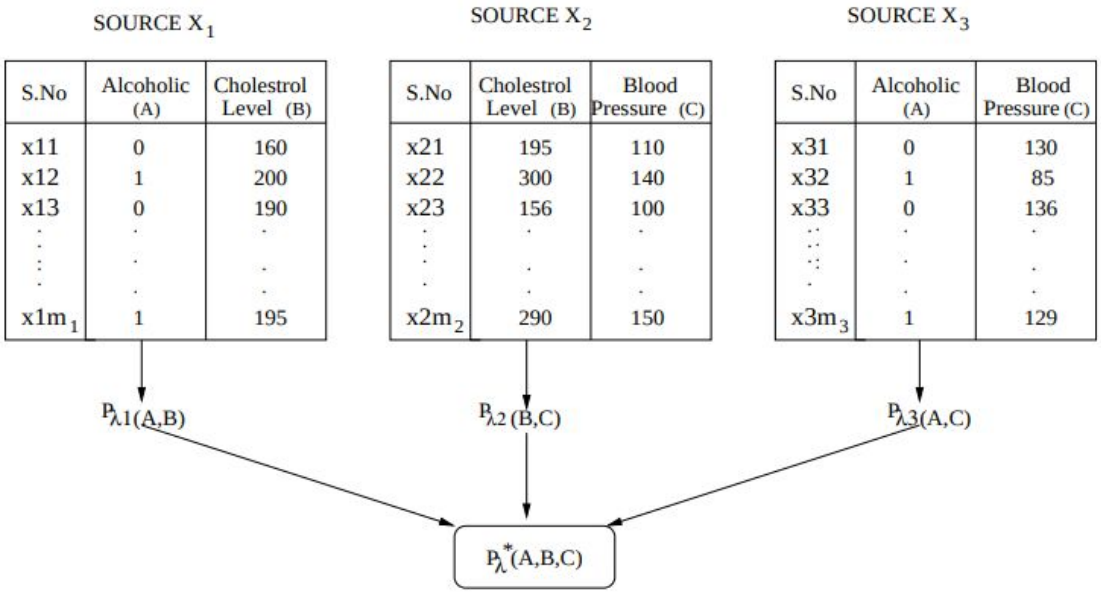


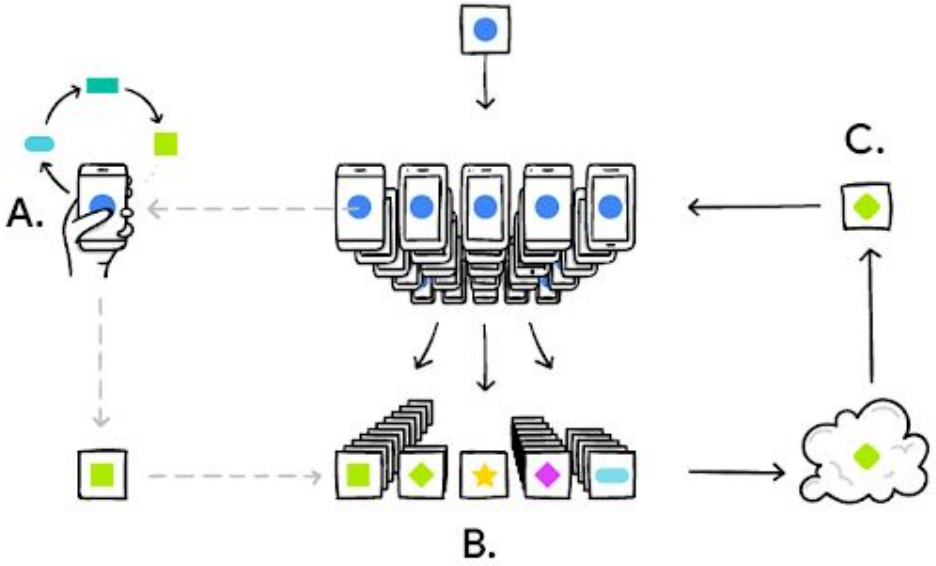

**Figure 1: Distributed learning scenario with overlapping sets of features.**

Merugu et al., 2005.

# Recent Advances in Privacy-Preserving Machine Learning

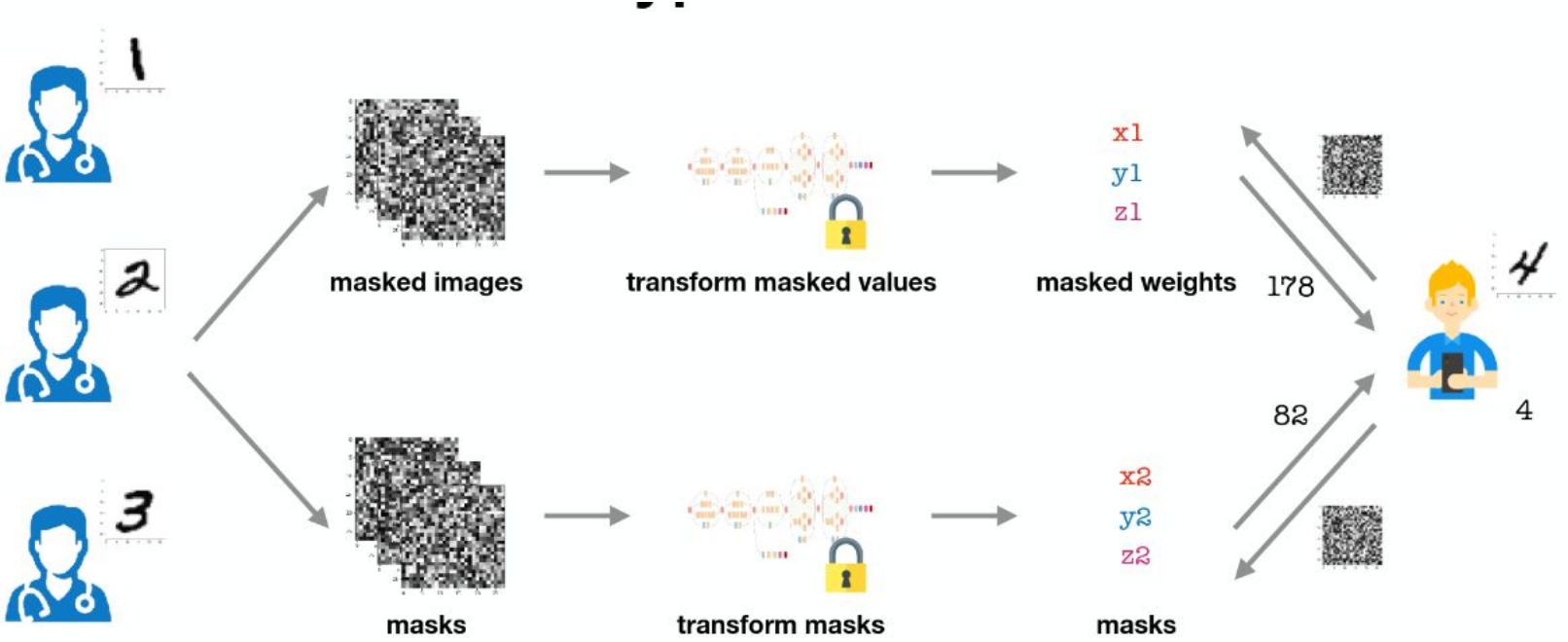
# Federated Learning

TensorFlow Federated enables developers to express and simulate federated learning systems. Pictured here, each phone trains the model locally (A). Their updates are aggregated (B) to form an improved shared model (C).




Google: tf-federated

# Encrypted Learning: Secure Multiparty Computation




DropoutLabs: tf-encrypted

# Differential Privacy

**Algorithm 1** Differentially private SGD (Outline)

**Input:** Examples $\{x_1, \ldots, x_N\}$, loss function $\mathcal{L}(\theta) = \frac{1}{N}\sum_i \mathcal{L}(\theta, x_i)$. Parameters: learning rate $\eta_t$, noise scale $\sigma$, group size $L$, gradient norm bound $C$.

**Initialize** $\theta_0$ randomly

**for** $t \in [T]$ **do**

Take a random sample $L_t$ with sampling probability $L/N$

**Compute gradient**

For each $i \in L_t$, compute $\mathbf{g}_t(x_i) \leftarrow \nabla_{\theta_t}\mathcal{L}(\theta_t, x_i)$

**Clip gradient**

$\bar{\mathbf{g}}_t(x_i) \leftarrow \mathbf{g}_t(x_i) / \max\left(1, \frac{\|\mathbf{g}_t(x_i)\|_2}{C}\right)$

**Add noise**

$\tilde{\mathbf{g}}_t \leftarrow \frac{1}{L}\left(\sum_i \bar{\mathbf{g}}_t(x_i) + \mathcal{N}(0, \sigma^2 C^2 \mathbf{I})\right)$

**Descent**

$\theta_{t+1} \leftarrow \theta_t - \eta_t \tilde{\mathbf{g}}_t$

**Output** $\theta_T$ and compute the overall privacy cost $(\varepsilon, \delta)$ using a privacy accounting method.

Abadi et al., 2015.

# Adversarial Regularization

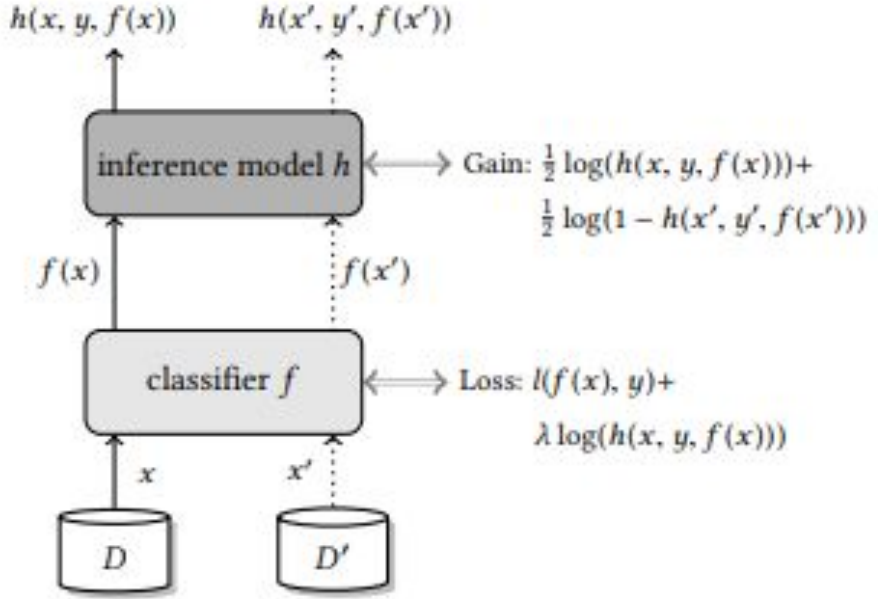



**Figure 2: Classification loss and inference gain, on the training dataset $D$ and reference dataset $D'$, in our adversarial training. The classification loss is computed over $D$, but, the inference gain is computed on both sets. To simplify the illustration, the mini-batch size is set to 1 here.**

Nasr et al., 2018.

# Encrypted Prediction Queries

| Data set | Model size | Computation | | Time per protocol | | Total running time | Comm. | Interactions |
|---|---|---|---|---|---|---|---|---|
| | | Client | Server | Compare | Dot product | | | |
| Breast cancer (2) | 30 | 46.4 ms | 43.8 ms | 194 ms | 9.67 ms | 204 ms | 35.84 kB | 7 |
| Credit (3) | 47 | 55.5 ms | 43.8 ms | 194 ms | 23.6 ms | 217 ms | 40.19 kB | 7 |

(a) Linear Classifier. Time per protocol includes communication.

| Data set | Specs. | | Computation | | Time per protocol | | Total running time | Comm. | Interactions |
|---|---|---|---|---|---|---|---|---|---|
| | $C$ | $F$ | Client | Server | Prob. Comp. | Argmax | | | |
| Breast Cancer (1) | 2 | 9 | 150 ms | 104 ms | 82.9 ms | 396 ms | 479 ms | 72.47 kB | 14 |
| Nursery (5) | 5 | 9 | 537 ms | 368 ms | 82.8 ms | 1332 ms | 1415 ms | 150.7 kB | 42 |
| Audiology (4) | 24 | 70 | 1652 ms | 1664 ms | 431 ms | 3379 ms | 3810 ms | 1911 kB | 166 |

(b) Naïve Bayes Classifier. $C$ is the number of classes and $F$ is the number of features. The Prob. Comp. column corresponds to the computation of the probabilities $p(c_i|x)$ (*cf.* Section 6). Time per protocol includes communication.

Bost et al., 2015.

# Still Unanswered Questions

# Overfitting? Model Capacity? Poor Regularization?

Table 1: The training and test accuracy (in percentage) of various models on the CIFAR10 dataset. Performance with and without data augmentation and weight decay are compared. The results of fitting random labels are also included.

| model | # params | random crop | weight decay | train accuracy | test accuracy |
|---|---|---|---|---|---|
| Inception | 1,649,402 | yes | yes | 100.0 | 89.05 |
| | | yes | no | 100.0 | 89.31 |
| | | no | yes | 100.0 | 86.03 |
| | | no | no | 100.0 | 85.75 |
| (fitting random labels) | | no | no | 100.0 | 9.78 |
| Inception w/o BatchNorm | 1,649,402 | no | yes | 100.0 | 83.00 |
| | | no | no | 100.0 | 82.00 |
| (fitting random labels) | | no | no | 100.0 | 10.12 |
| Alexnet | 1,387,786 | yes | yes | 99.90 | 81.22 |
| | | yes | no | 99.82 | 79.66 |
| | | no | yes | 100.0 | 77.36 |
| | | no | no | 100.0 | 76.07 |
| (fitting random labels) | | no | no | 99.82 | 9.86 |
| MLP 3x512 | 1,735,178 | no | yes | 100.0 | 53.35 |
| | | no | no | 100.0 | 52.39 |
| (fitting random labels) | | no | no | 100.0 | 10.48 |
| MLP 1x512 | 1,209,866 | no | yes | 99.80 | 50.39 |
| | | no | no | 100.0 | 50.51 |
| (fitting random labels) | | no | no | 99.34 | 10.61 |

Zhang et al., 2017

# Accurate, Practical Threat Modeling

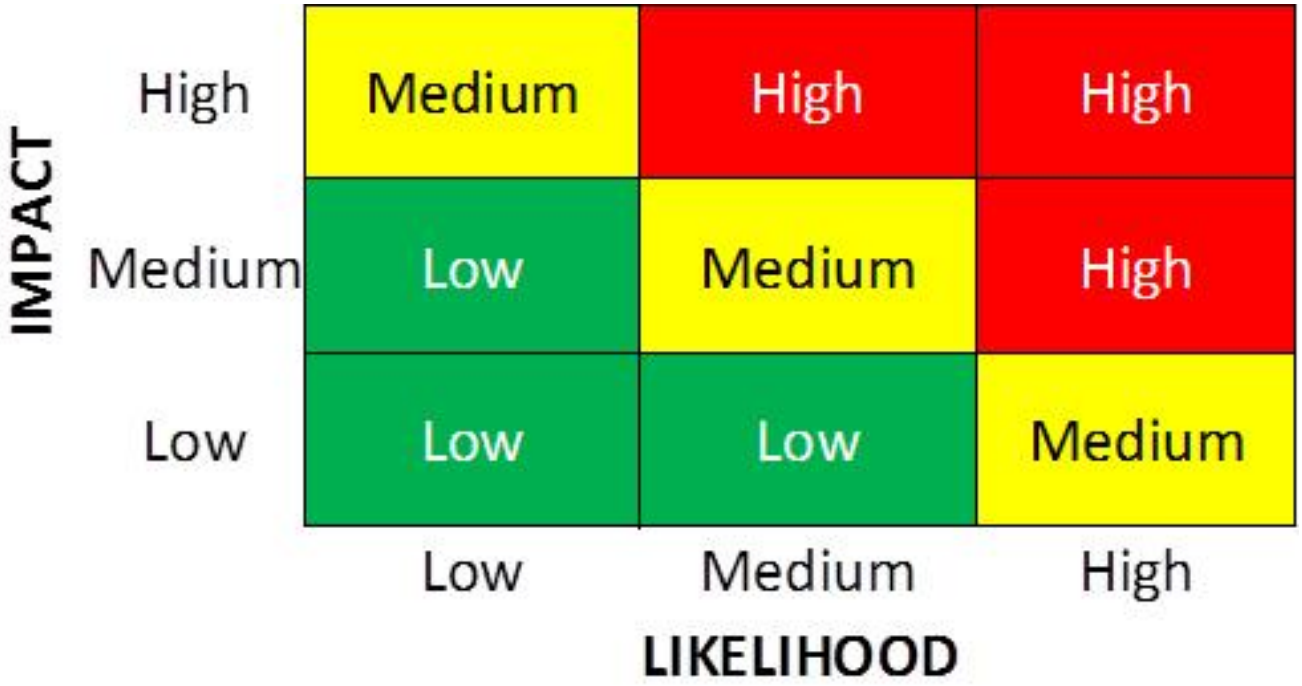



Image: https://www.pivotpointsecurity.com

# Privacy & Interpretability

**Table 2: Minority populations are more vulnerable to being revealed by the Koh and Liang method.**

| | #points | $k = 1$ | $k = 5$ | $k = 10$ |
|---|---|---|---|---|
| Whole data set | 2400 | 26% | 36% | 39% |
| Clownfish | 26 | 27% | 37% | 43% |
| Lion fish | 29 | 9% | 42% | 51% |
| Birds | 15 | 64% | 85% | 90% |

**(a) Disclosure likelihood by type in the dog/fish dataset.**

| | % of data | $k = 1$ | $k = 5$ | $k = 10$ |
|---|---|---|---|---|
| Whole data set | 100% | 34% | 64% | 77% |
| Age 0 -10 | <0.1% | 67% | 100% | 100% |
| Age 0 -20 | <1% | 20% | 58% | 92% |
| Caucasian | 74% | 34% | 64% | 77% |
| African American | 19% | 38% | 68% | 81% |
| Hispanics | 2% | 39% | 64% | 76% |
| Unknown race | 1% | 35% | 60% | 77% |
| Asian American | <1% | 25% | 64% | 89% |

**(b) Disclosure likelihood by age and race in the hospital dataset.**

Shokri et al., 2019

# Accurate Definitions of Privacy

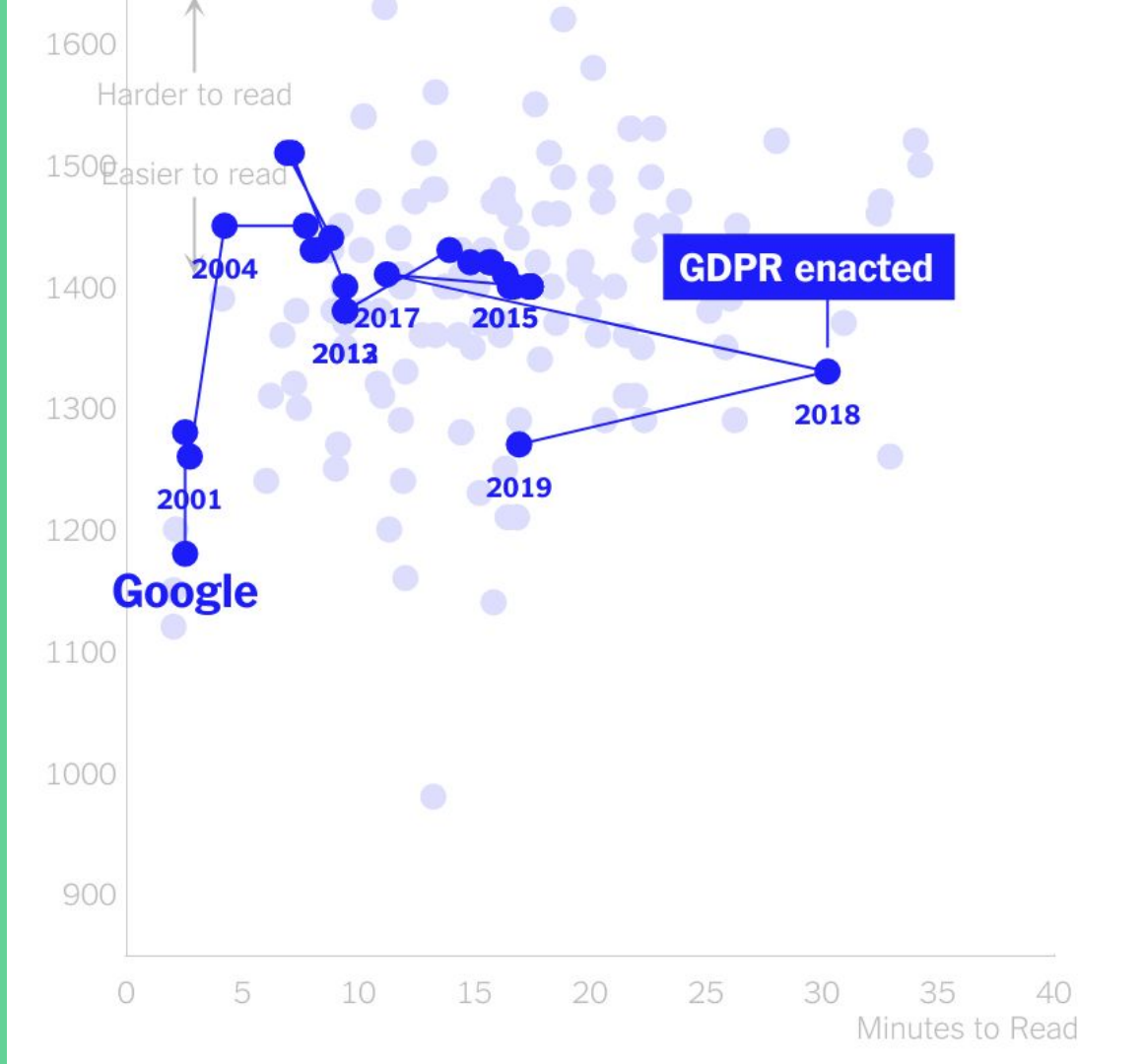
Privacy is not about control over data nor is it a property of data.  It's about **a collective understanding of a social situation's boundaries** and knowing how to operate within them.  In other words, it’s about having control over a situation. It's about understanding the audience and knowing how far information will flow.  It’s about **trusting the people, the situating, and the context.**

-- danah boyd

# Location Tracking and Privacy Policies (2008)

The work presented in this article confirms that **people are generally apprehensive about the privacy implications associated with location tracking**. It also shows that privacy preferences tend to be complex and depend on a variety of contextual attributes (e.g. relationship with requester, time of the day, where they are located). Through a series of user studies, we have found that **most users are not good at articulating these preferences.**

Sahdeh et al., 2008.


1600
Harder to read
1500
Easier to read
1400
1300
1200
1100
1000
900
2004
2017
2015
2013
GDPR enacted
2018
2001
2019
Google
0
5
10
15
20
25
30
35
40
Minutes to Read

The scientist and engineer has responsibilities that transcend his immediate situation, that in fact extend directly to future generations... We are all their trustees.

Joseph Weizenbaum, 1976

Thank you!

Questions?

- Now?
- Later?
  - katharine@kjamistan.com
  - @kjam (Twitter)

# Slide References


- Nowak et al., *Understanding Privacy Concerns*, 1992.
- Brankovic et al., *Privacy Issues in Knowledge Discovery and Data Mining*, 2000.
- Rivest, *Cryptography and Machine Learning*, 1988.
- Merugu et al., *A privacy-sensitive approach to distributed clustering,* 2004.
- Tf-federated: https://www.tensorflow.org/federated
- Tf-encrypted: https://github.com/tf-encrypted/tf-encrypted
- Abadi et al., *Deep Learning with Differential Privacy*, 2015
- Bost et al., *Machine Learning Classification over Encrypted Data*, 2015.
- Zhang et al., *Understanding Deep Learning Requires Rethinking Generalization*, 2017.
- Shokri et al., *Privacy Risks of Explaining Machine Learning Models*, 2019.
- Sadeh et al., *Understanding and Capturing People's Privacy Policies in a Mobile Social Networking Application*, 2008.
- Brankovic et al.,, *Privacy Issues in Knowledge Discovery and Data Mining*, 2000.
- NYTimes Privacy Policy Investigation: https://www.nytimes.com/interactive/2019/06/12/opinion/facebook-google-privacy-policies.html
- Weizenbaum, *Computer Power and Human Reason*, 1976.