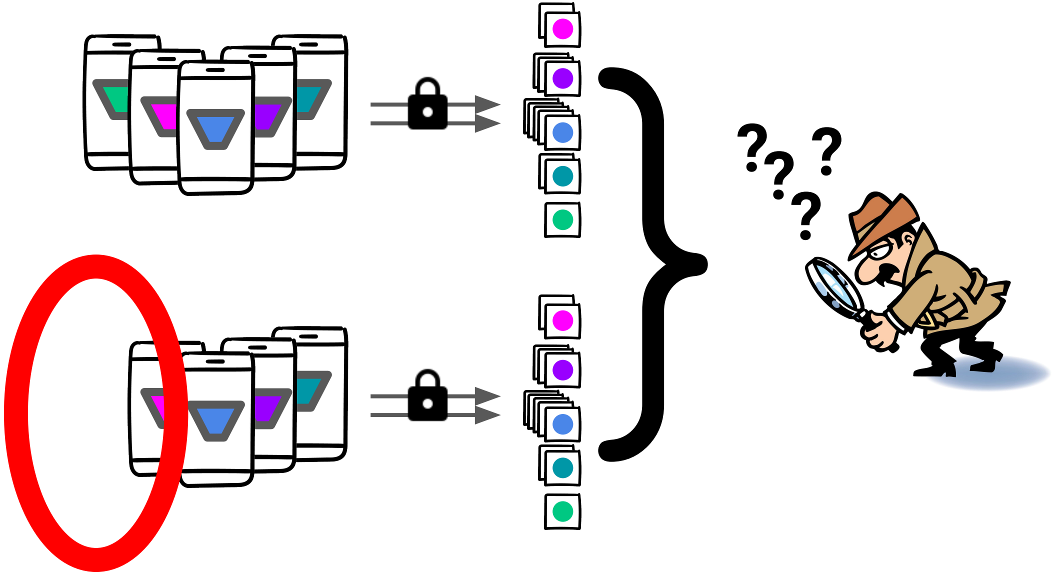
Machine Learning at Scale with Differential Privacy in TensorFlow

Nicolas Papernot
Google Brain

@NicolasPapernot

What is privacy?



Why should we care?

Training data:
The SSN of Alice is 1234
The SSN of Bob is 5678

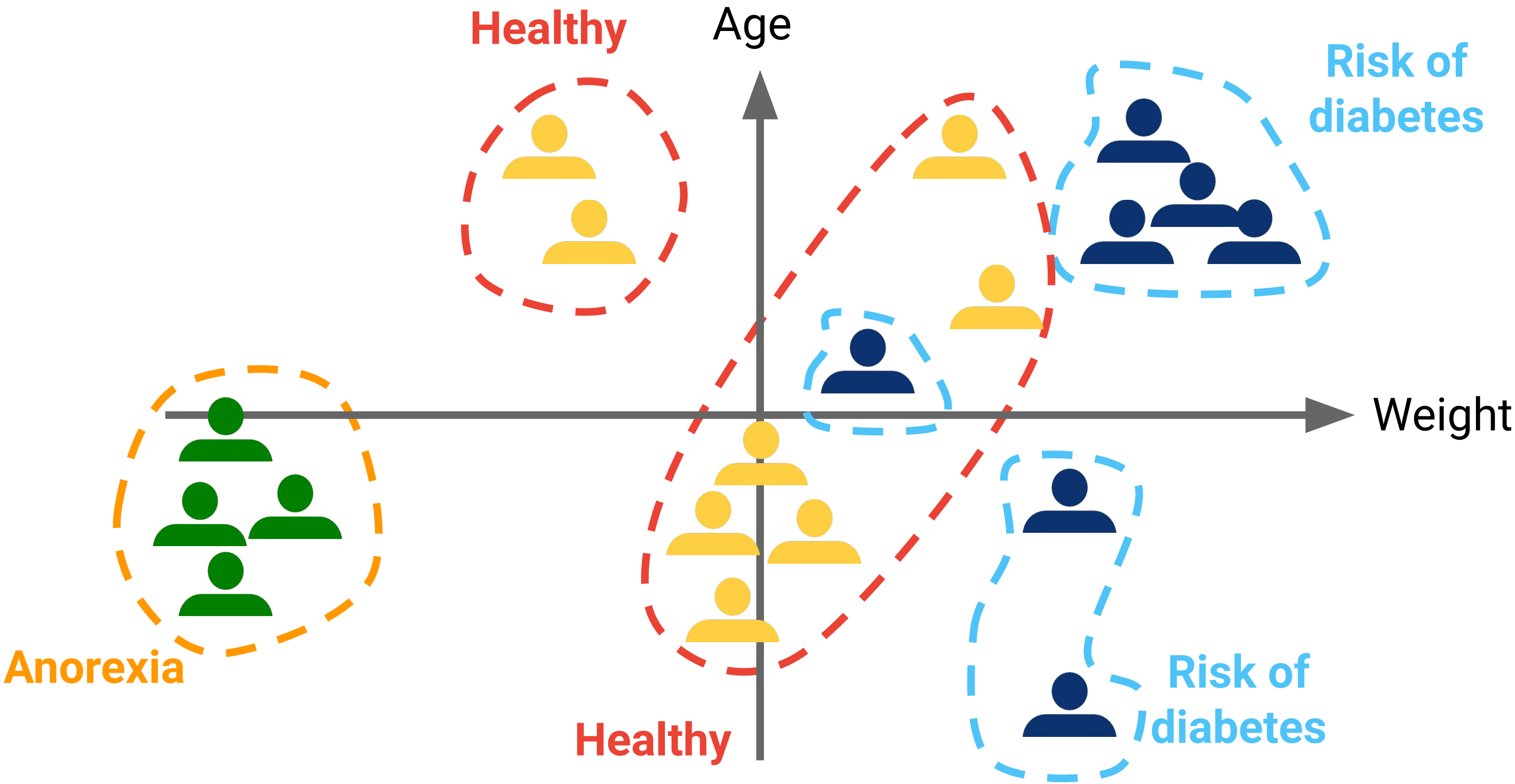
Membership inference attacks (Shokri et al.)

Was "*The SSN of Alice is 1234*" in the training data?

Extraction of memorized training data (Carlini et al.)

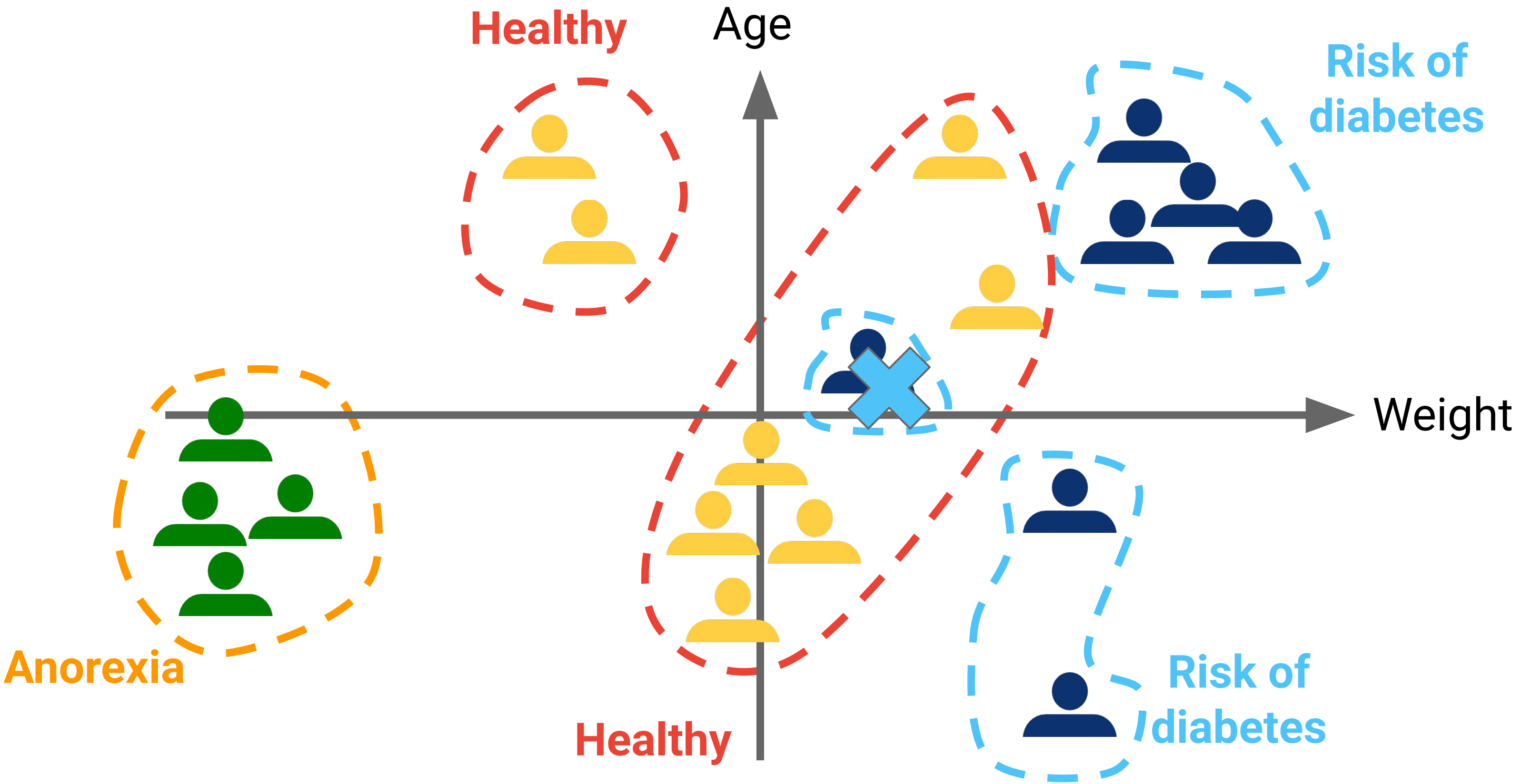
Complete "*The SSN of Alice is …*"

Machine learning is not magic



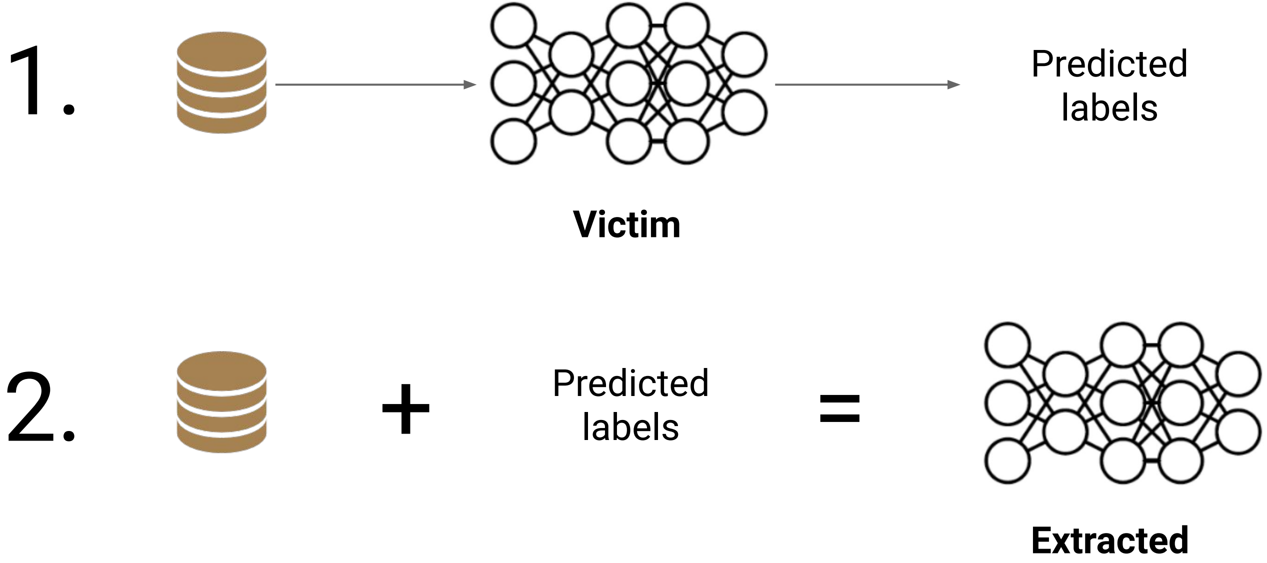
Google

Machine learning is not magic



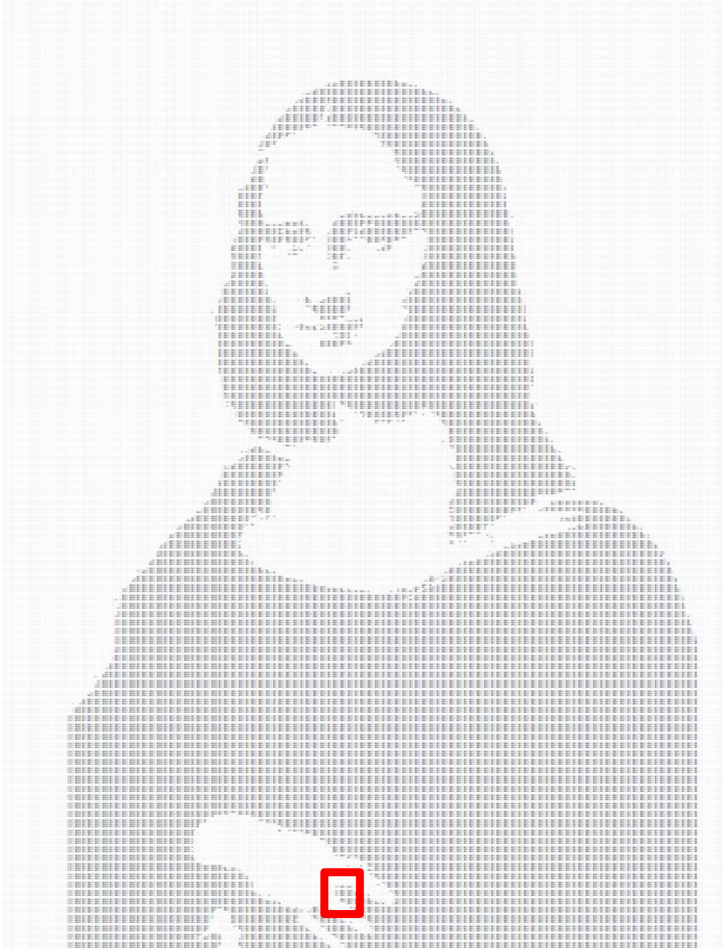
Why should we care? (part 2)

Attackers may gain white-box access to the model through **model extraction**

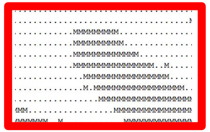


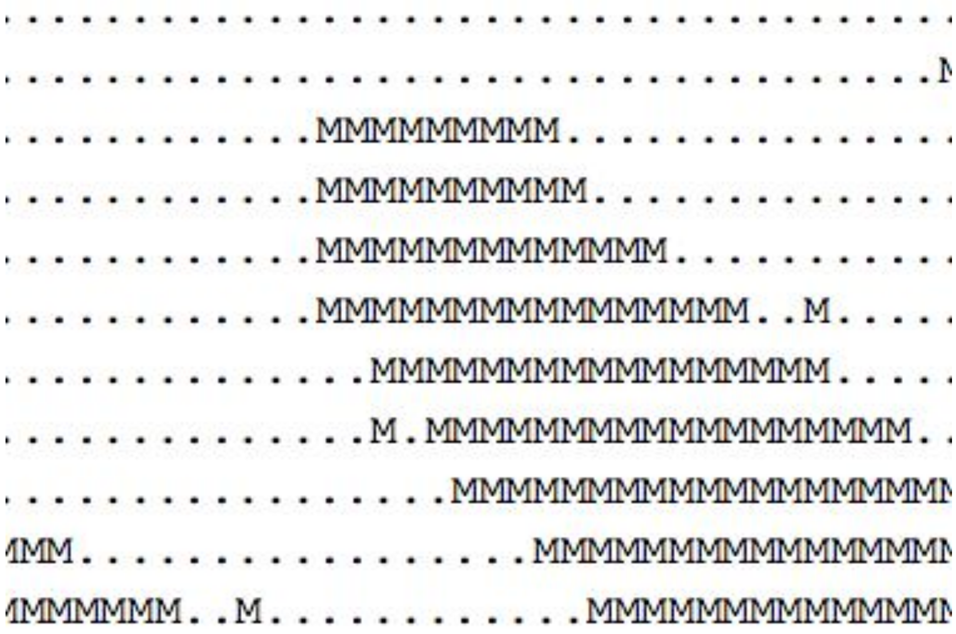
Stealing Machine Learning Models via Prediction APIs (Tramer et al.)
Practical Black-Box Attacks against Machine Learning (Papernot et al.)

A Metaphor For Private Learning



An Individual’s Training Data





An Individual's Training Data

Each bit is flipped with probability 25%

```
.......M........MM.M.......MMM.M..
......................MM...MMMM...
....M..MM.MM..MMM.M.MM.M...M..MM..
.MM......MMM....MMMMMMMMM...M...MM
..M....M........MM..MMMMMMM...M...
M.......M..MM.MMMMMMMMMMMMMMM....M
.....M.....M.M.M.MMMMMM...MMMMM...
...M.....M.MM.M.MM..M..M..MM.MMMMM
M...M.M.....M.M..M..MMM.MMMMM.MMMM
.MMM.M....M.M.M........MMMMMMMMM.M
```

Big Picture Remains!

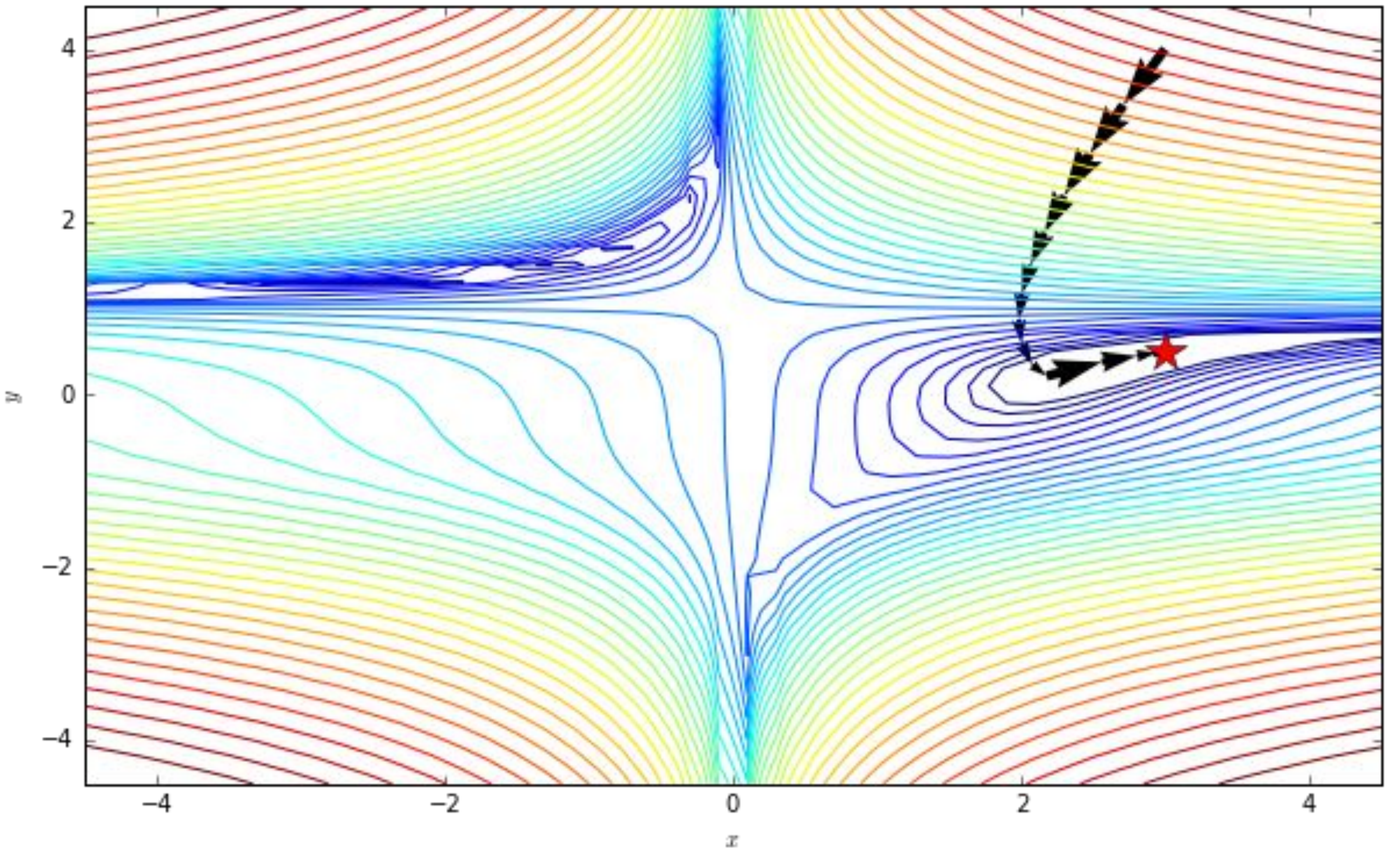
Google

How to train a model with SGD?

```
Initialize parameters θ

For t = 1..T do
  Sample batch B of training examples
  Compute average loss L on batch B
  Compute average gradient of loss L wrt parameters θ



  Update parameters θ by a multiple of gradient average
```

How to train a model with differentially private SGD?

```
Initialize parameters θ

For t = 1..T do
  Sample batch B of training examples
  Compute per-example loss L on batch B
  Compute per-example gradients of loss L wrt parameters θ
  Ensure L2 norm of gradients < C by clipping
  Add Gaussian noise to average gradients (as a function of C)
  Update parameters θ by a multiple of noisy gradient average
```



Source: ruder.io

How does it work in practice with TensorFlow?

SGD

DP-SGD with clipping and noising

```
# Before

vector_loss = tf.nn.sparse_softmax_cross_entropy_with_logits(
    labels=labels,
    logits=logits)
optimizer = tf.train.GradientDescentOptimizer(
    learning_rate=FLAGS.learning_rate)
train_op = optimizer.minimize(loss=tf.reduce_mean(vector_loss))

# After

vector_loss = tf.nn.sparse_softmax_cross_entropy_with_logits(
    labels=labels,
    logits=logits)
optimizer = privacy.VectorizedDPSGD(
    l2_norm_clip=FLAGS.l2_norm_clip,
    noise_multiplier=FLAGS.noise_multiplier,
    learning_rate=FLAGS.learning_rate)
train_op = optimizer.minimize(loss=vector_loss)
```

Optimizer receives scalar loss

Optimizer receives vector loss

How to choose hyper-parameters?

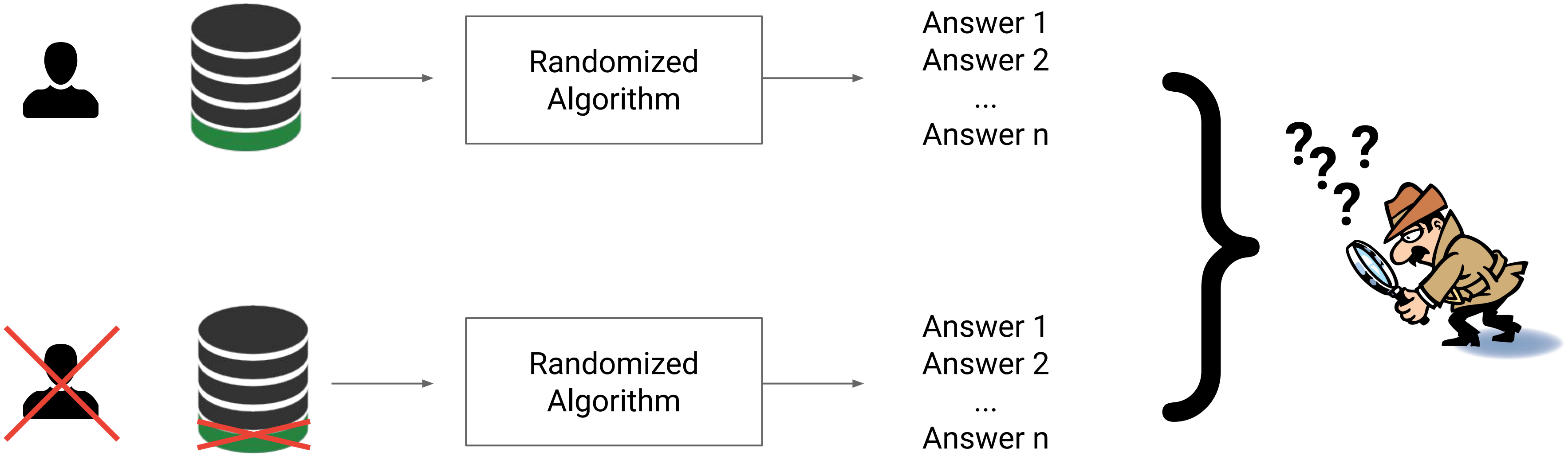
`l2_norm_clip`: maximum Euclidean norm of each individual gradient that is computed on an individual training example.

This parameter bounds the optimizer's sensitivity to individual training points. In general, the lower the value, the stronger the privacy.

`noise_multiplier`: controls how much noise is sampled and added to gradients before they are applied by the optimizer.

Generally, more noise results in better privacy (often, but not necessarily, at the expense of lower utility).

Differential privacy: a gold standard



$$Pr[M(d) \in S] \leq e^{\varepsilon} Pr[M(d') \in S]$$

IACR:3650 (Dwork et al.)

How to interpret results?

TensorFlow Privacy provides a toolkit for analyzing the privacy guarantees obtained by DP-SGD using the framework of differential privacy. Privacy guarantees are rigorous and independent of the training data, they depend on:

- Number of steps (how many batches of data are sampled to train)
- Probability of sampling each batch (i.e., batch size / number of train points)
- `noise_multiplier` parameter

See `tensorflow/privacy/analysis/compute_dp_sgd_privacy.py`

Example on MNIST

Privacy

ε	□	Accuracy
1.19	10^{-5}	95.0%
3.01	10^{-5}	96.6%
7.10	10^{-5}	97.0%
∞	0	99.0%

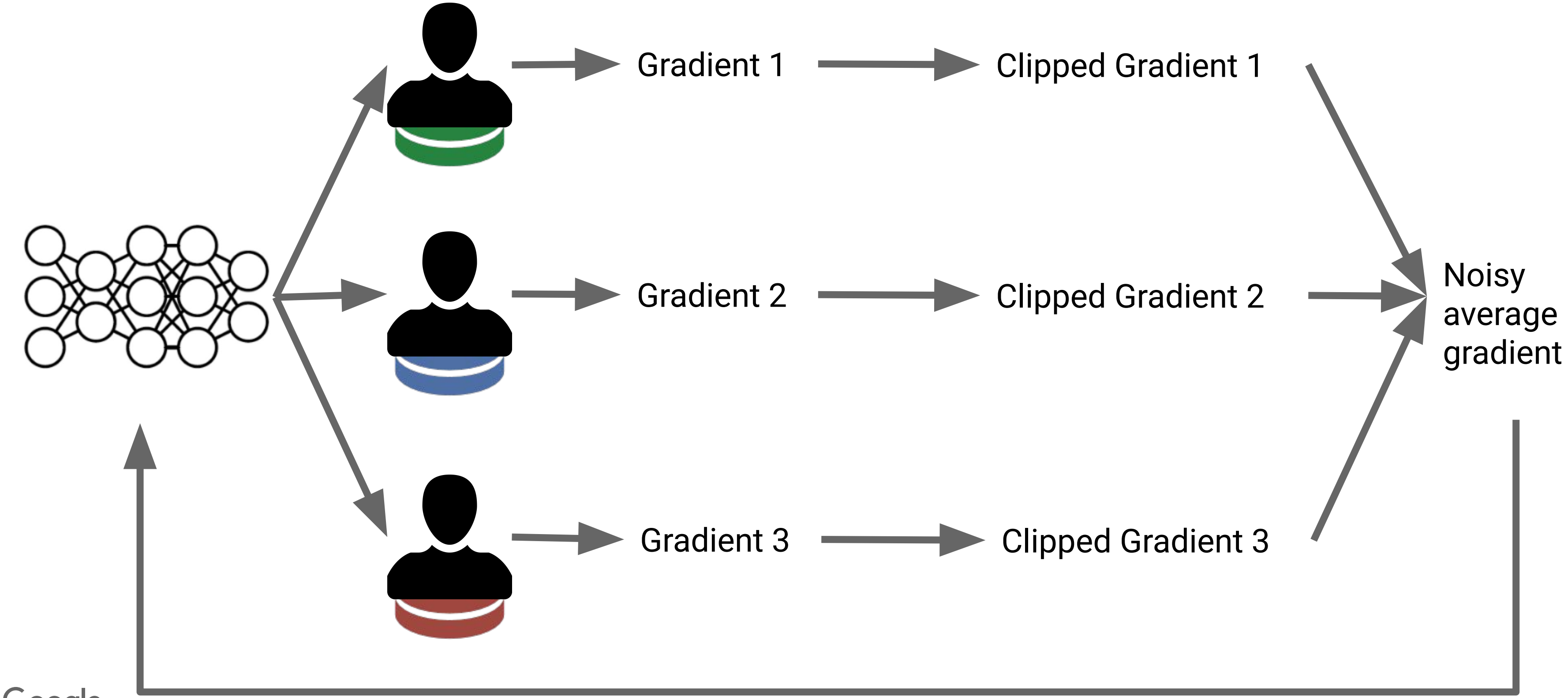
Accuracy

What are benefits of DP-SGD beyond privacy?

Google

Prototypical Examples in Deep Learning: Metrics, Characteristics, and Utility (Carlini, Erlingsson, Papernot)

What if the data is not centralized?



TF Privacy Library: **github.com/tensorflow/privacy**

Blog: **cleverhans.io**
- Privacy and machine learning: two unexpected allies?
- Machine Learning with Differential Privacy in TensorFlow

Email: **nicolas@papernot.fr**

Google