DigiTrans 2022

Practical Private Data Analysis

Nicolas Grislain
Sarus Technologies

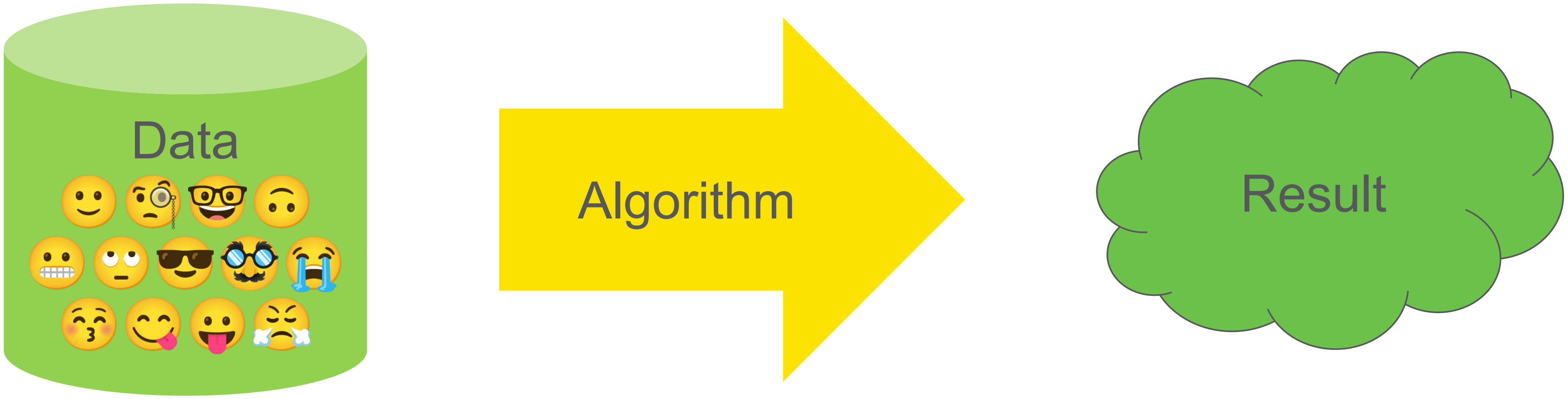
Agenda

- Private Data Analysis
- Differential Privacy
- Complex, Composed Mechanisms out of basic building blocks: DP-SGD
- Private Analysis in Practice

Private Data Analysis

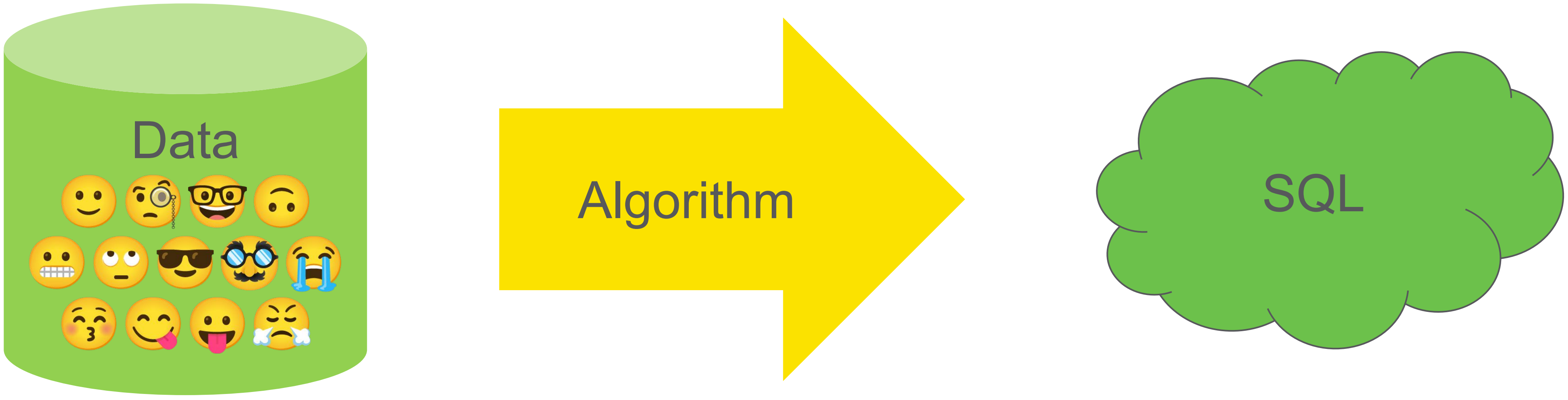
Private Data Analysis

We want to compute **something** on private data



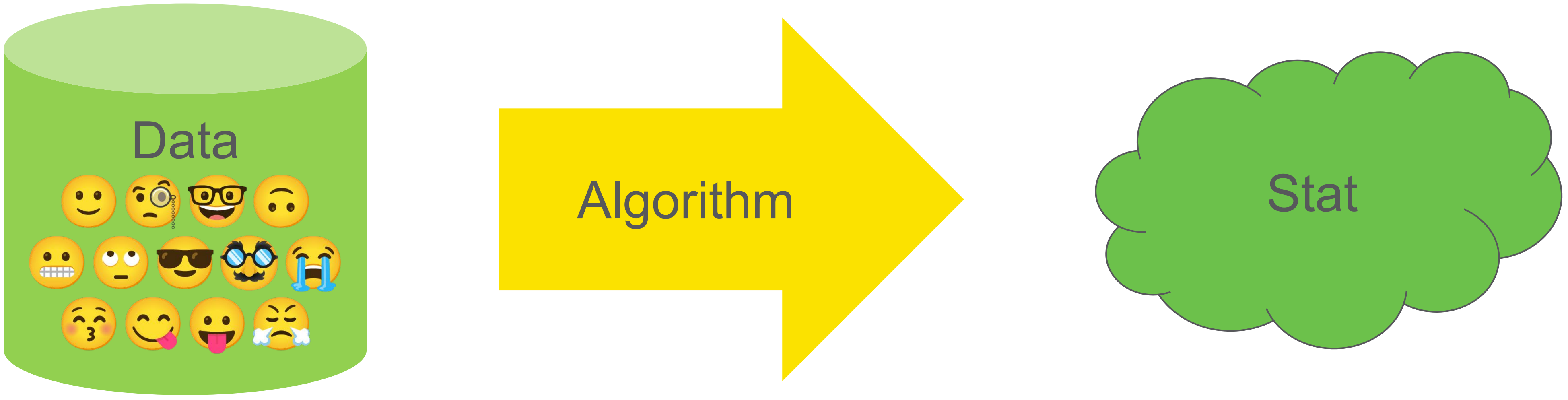
Private Data Analysis

We want to run **SQL queries** on private data



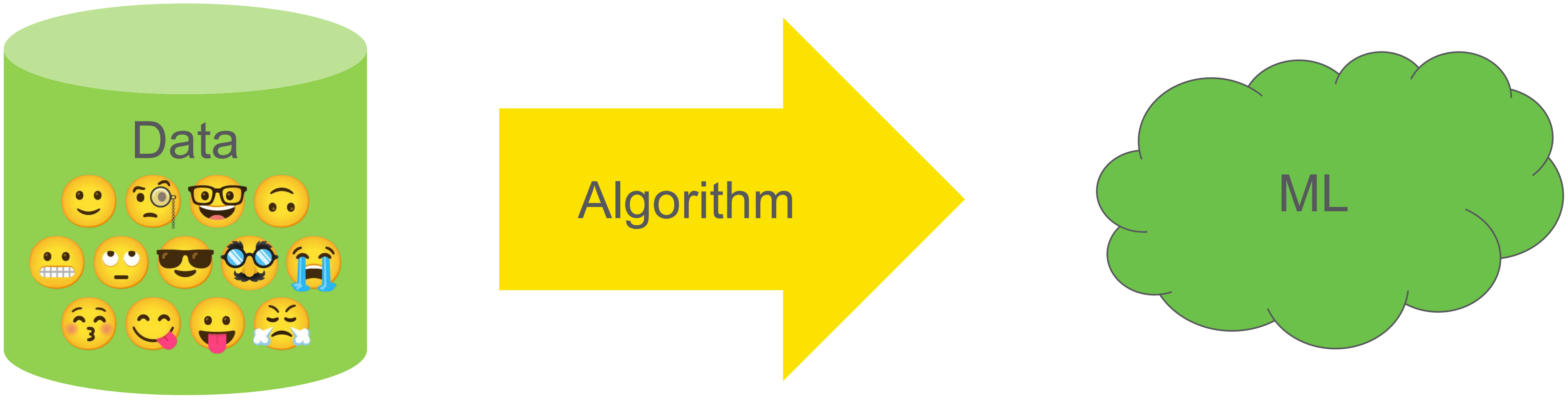
Private Data Analysis

We want to run **Logistic Regression** on private data



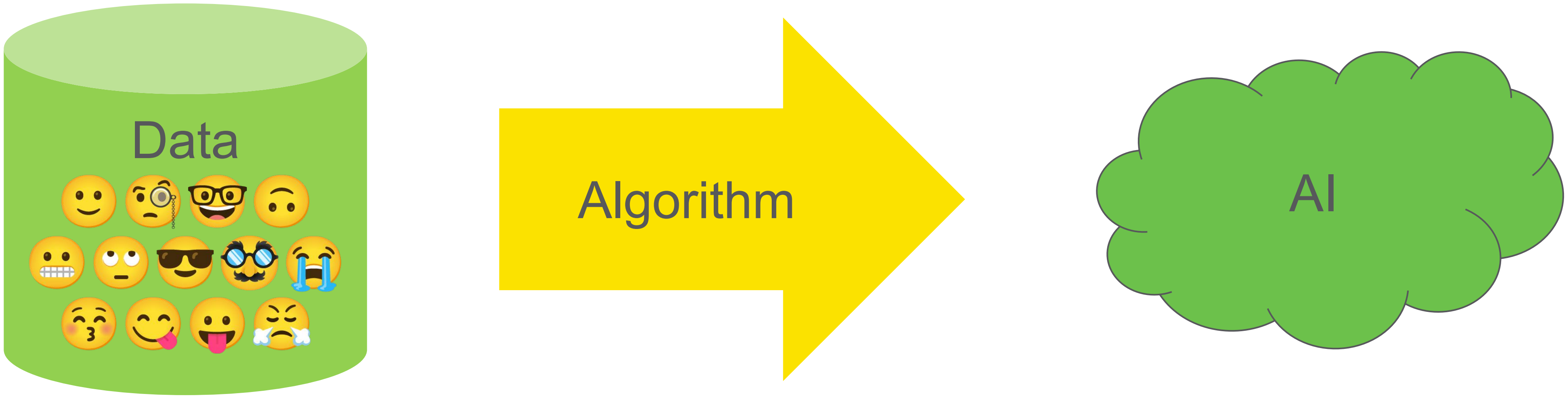
Private Data Analysis

We want to fit **Machine Learning Models** on private data



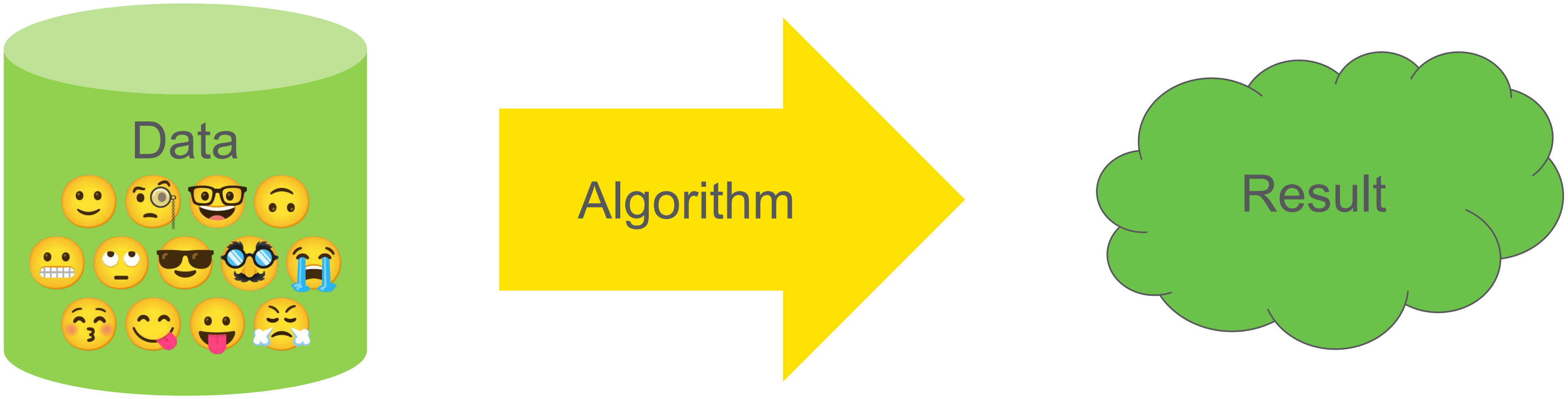
Private Data Analysis

We want to train **Neural Nets** on private data



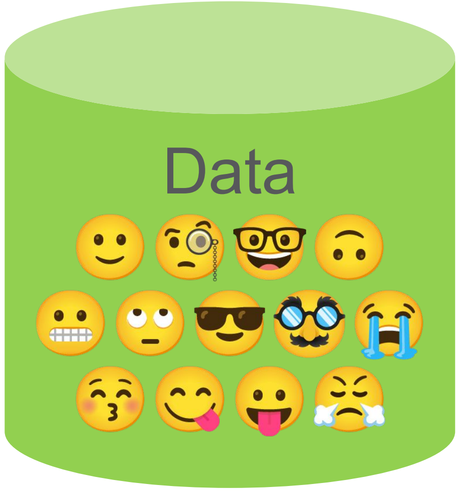
Private Data Analysis

We want to compute **something** on private data

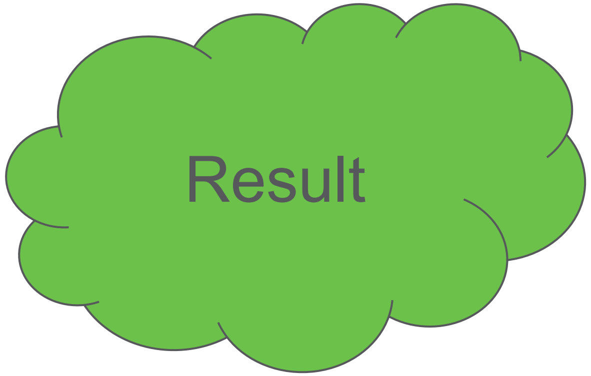


Private Data Analysis

We want to compute **something** on private data

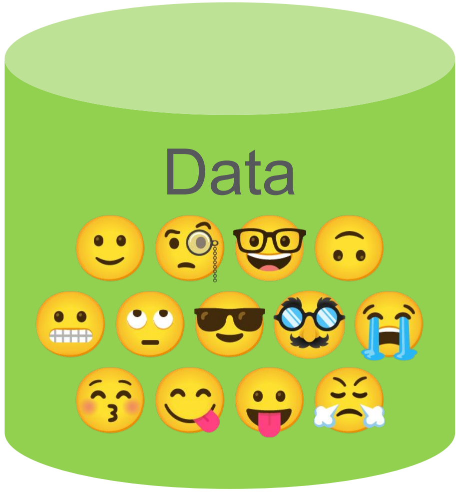


- Remove less relevant information
- Aggregate data enough
- Make sure an individual may be in many aggregates
- Use heuristics to prevent specific attack scenarios

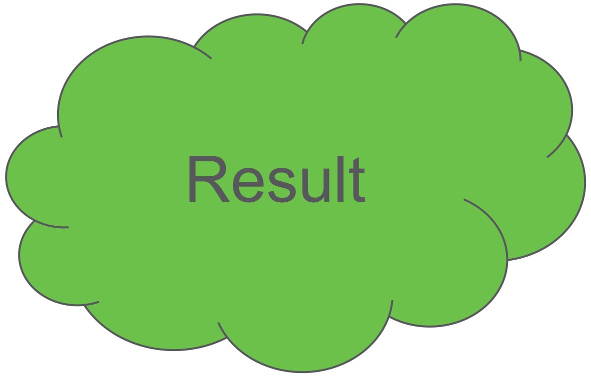


Private Data Analysis

Current practice: manual, takes time, assumptions about attackers, destroy data



- Remove less relevant in rmation
- gat
- M in be in
- eur p vent spe c attack scenarios



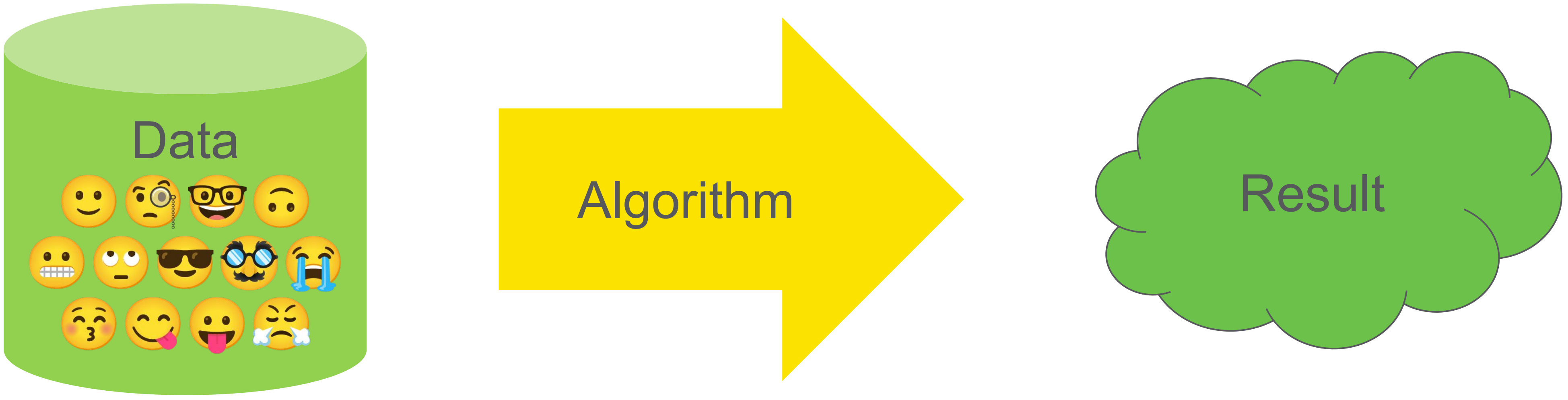
We need a better way

- Less manual
- Less destructive
- Convenient to use
- Stronger…

Differential Privacy

Private Data Analysis

We want to compute **something** on private data



Differential Privacy

Differential Privacy was introduced by Cynthia Dwork in 2006

- It bounds the difference in results for any two datasets differing by one individual used as input.

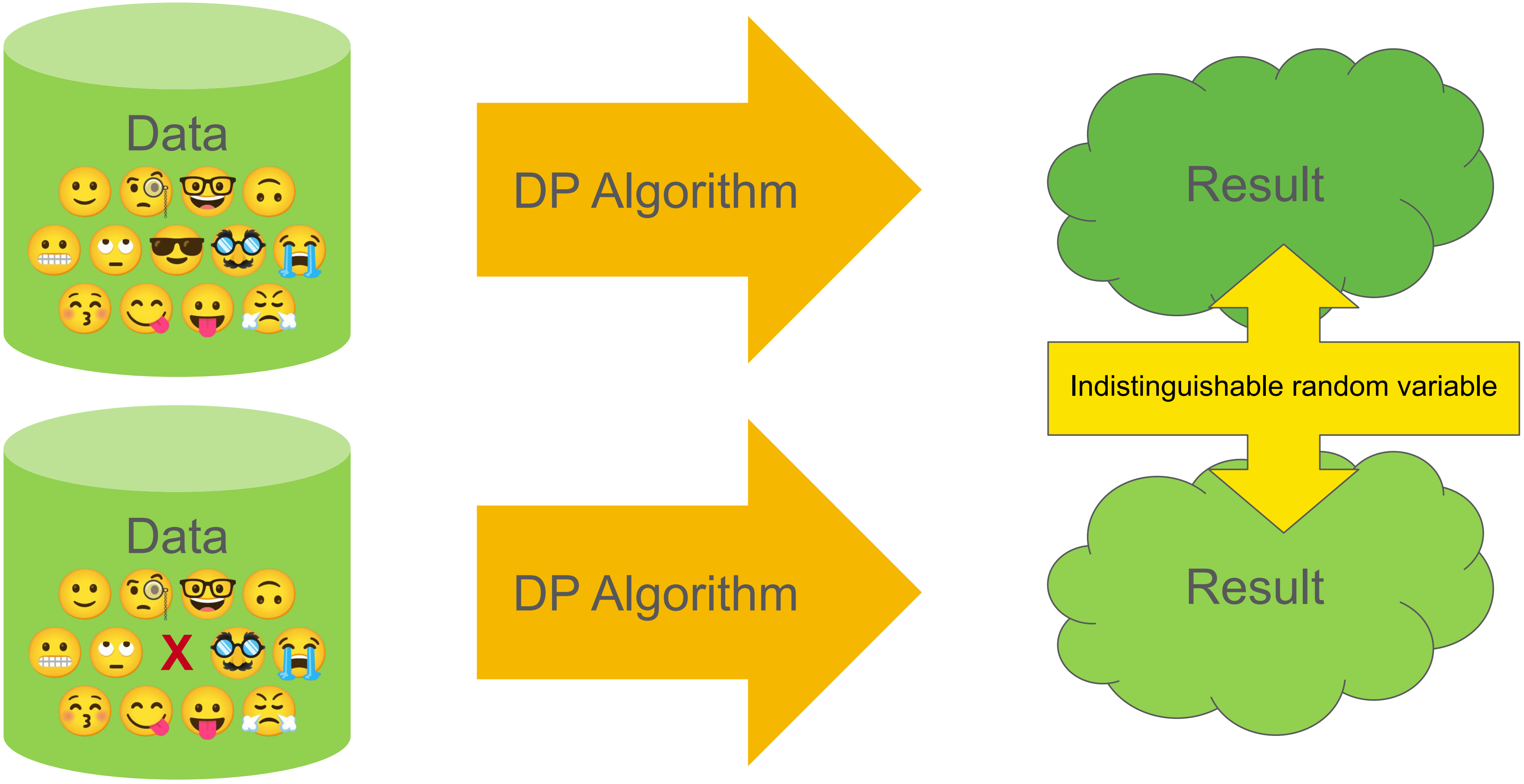
$$\Pr[\mathcal{A}(D_1) \in S] \leq \exp(\varepsilon) \cdot \Pr[\mathcal{A}(D_2) \in S]$$

- It gives a strong theoretical foundation to privacy protection
  - No bayesian inference is possible about an individual. At all.

- It does not rely on any assumption about the attacker

- It quantifies privacy loss and enables the definition of privacy budgets.

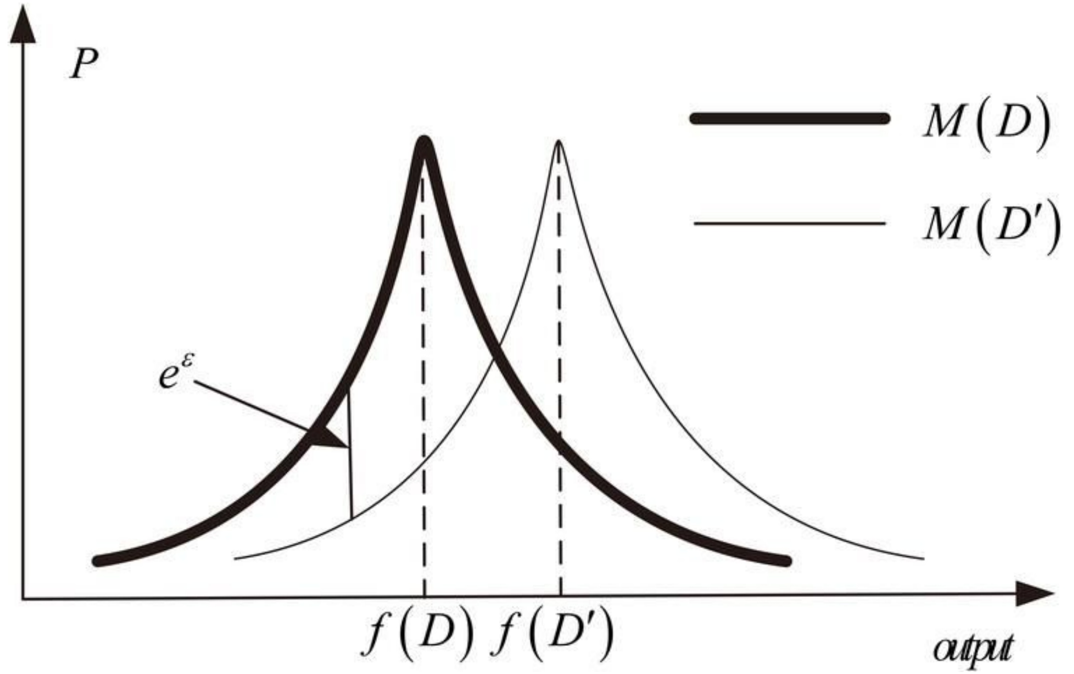
Differential Privacy randomizes the result



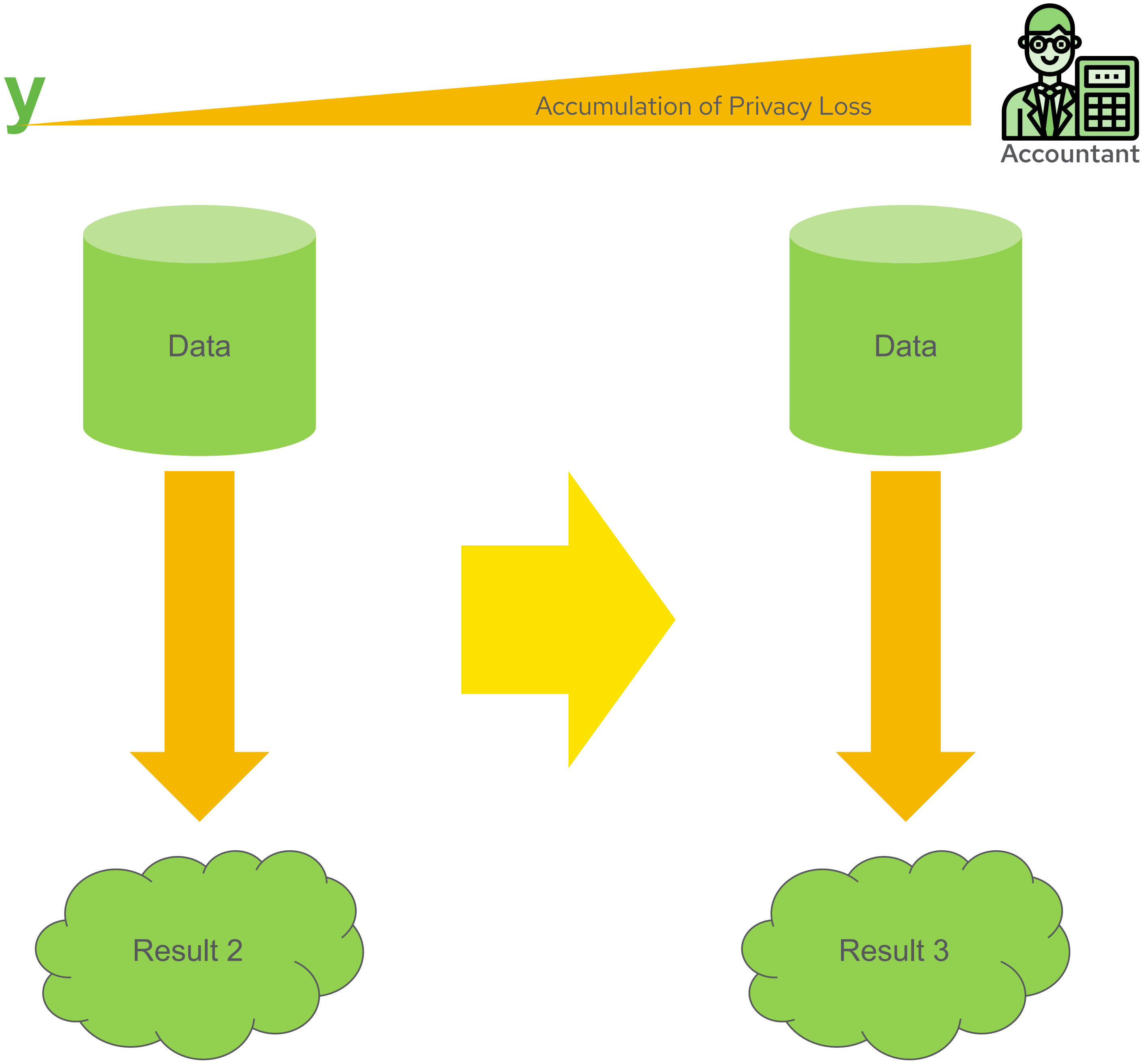
It caps the difference in distribution

Differential Privacy was introduced by Cynthia Dwork in 2006

$$\Pr[\mathcal{A}(D_1) \in S] \leq \exp(\varepsilon) \cdot \Pr[\mathcal{A}(D_2) \in S]$$



Compose adaptively

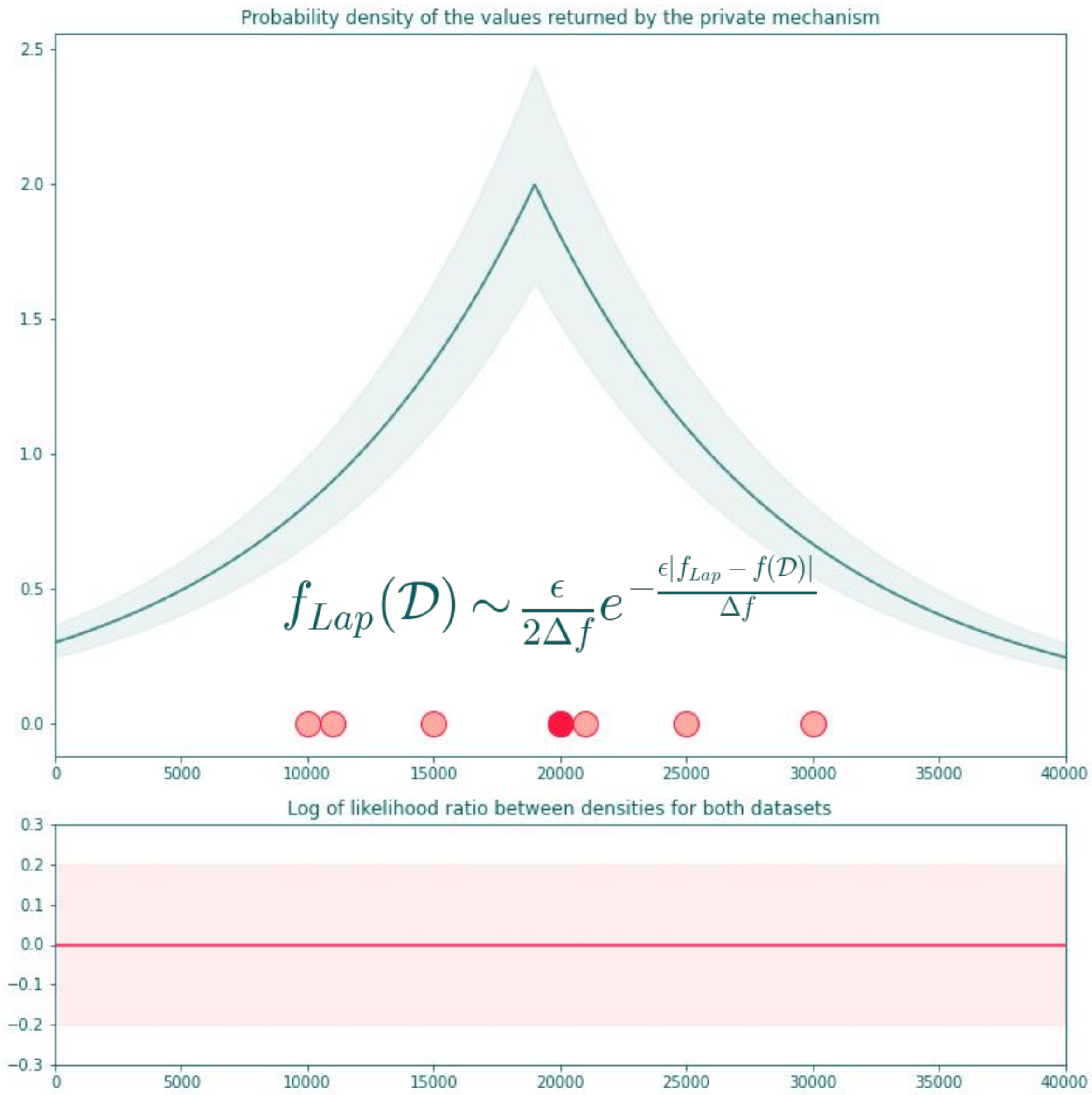


Differential Privacy

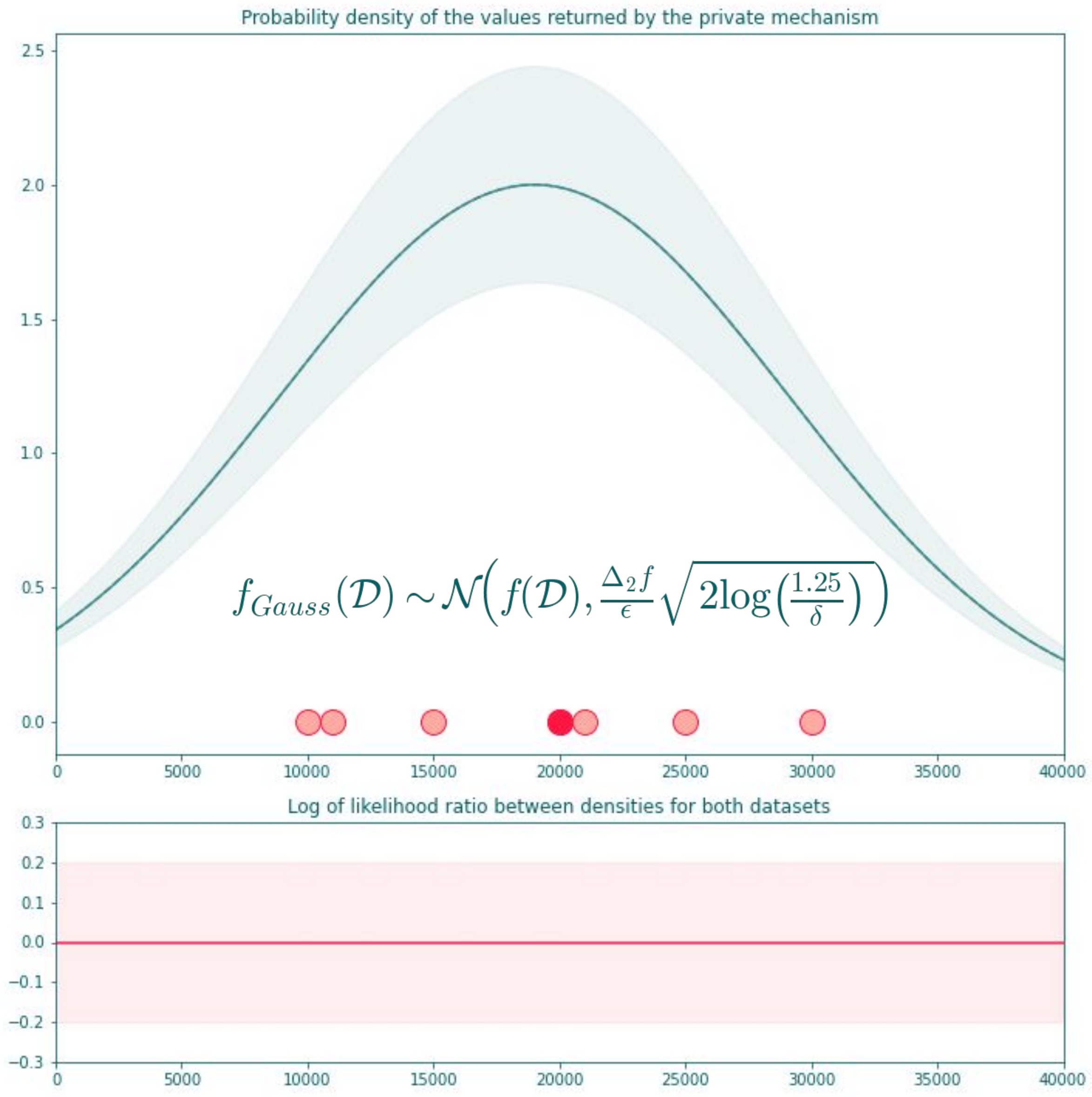
- Systematic approach
  - No need for an attack model
  - No expert judgement required
- Privacy Loss can be quantified and controlled
  - Privacy loss accumulates
  - We can have a notion of privacy budget
- DP Mechanisms can be composed
  - Complex analysis use-cases can be built out of basic building blocks

Complex, Composed Mechanisms out of basic building blocks: DP-SGD

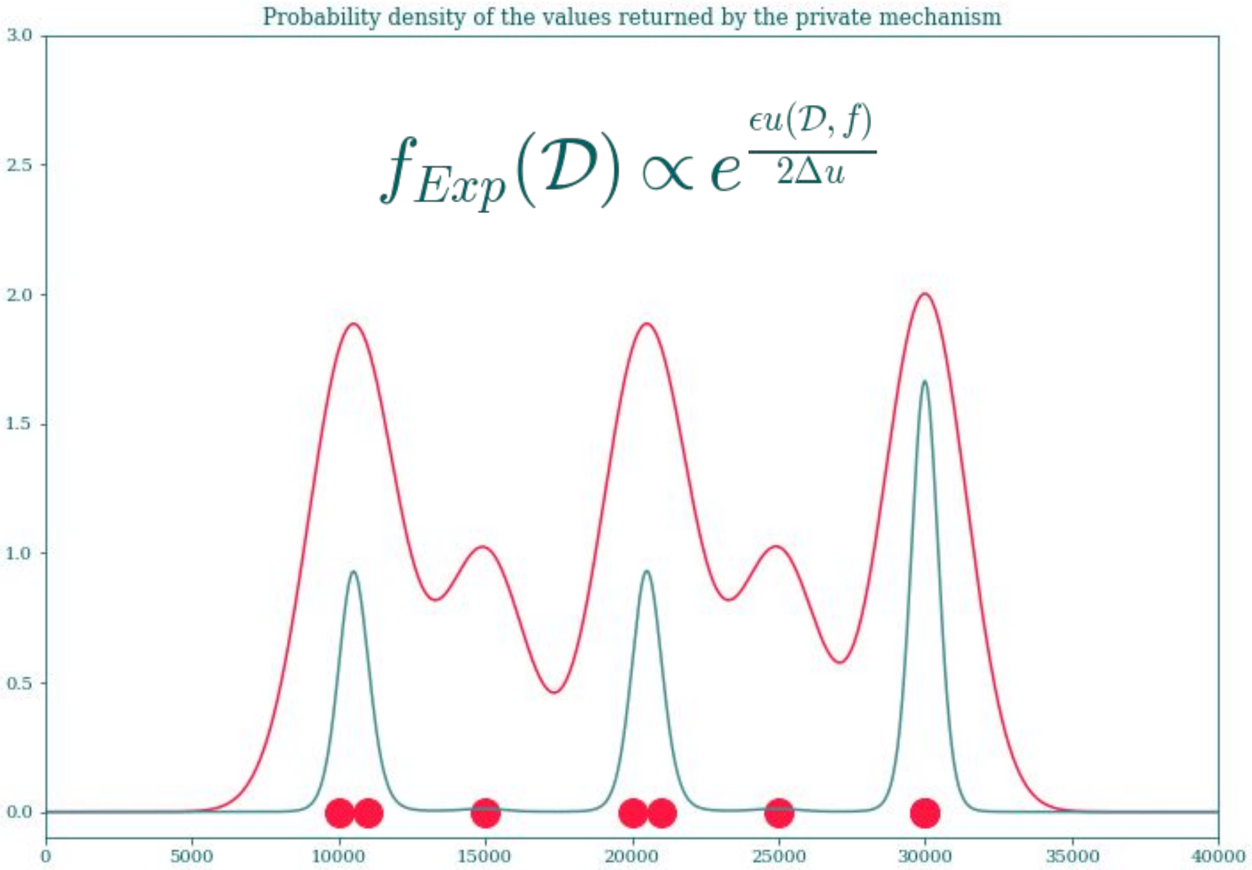
Laplace Mechanism



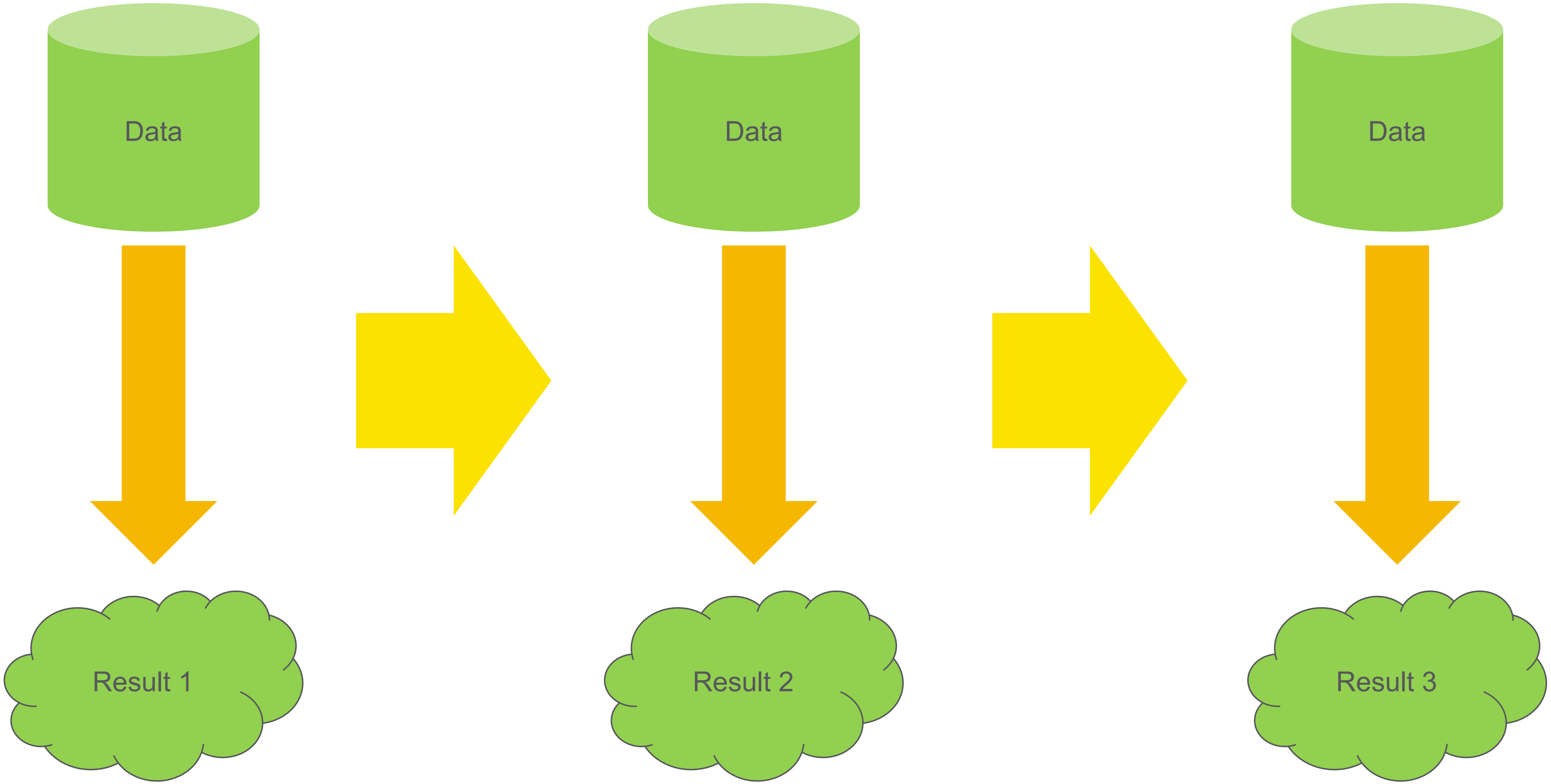
Gaussian Mechanism



Exponential Mechanism



They compose…



DP-SGD

- DP-SGD
  - [Abadi et al. 2016 - Deep Learning with Differential Privacy]
  - [Differential Privacy Series Part 1 | DP-SGD Algorithm Explained]

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 η_t, noise scale σ, group size L, gradient norm bound C.

Initialize θ_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

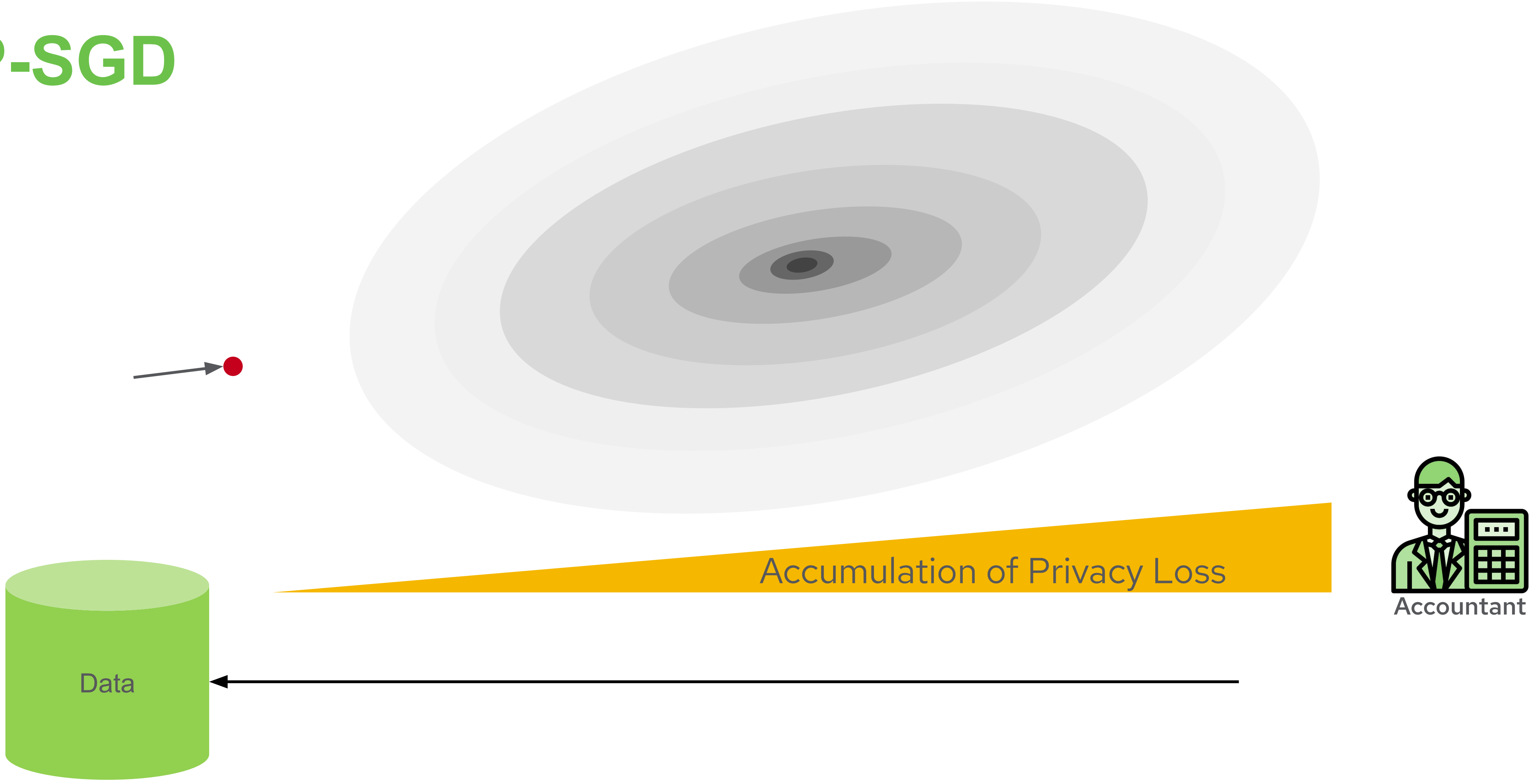
$\tilde{\mathbf{g}}_t \leftarrow \frac{1}{L}\left(\sum_i \bar{\mathbf{g}}_t(x_i) + \mathcal{N}(0, \sigma^2 C^2 \mathbf{I})\right)$

Descent

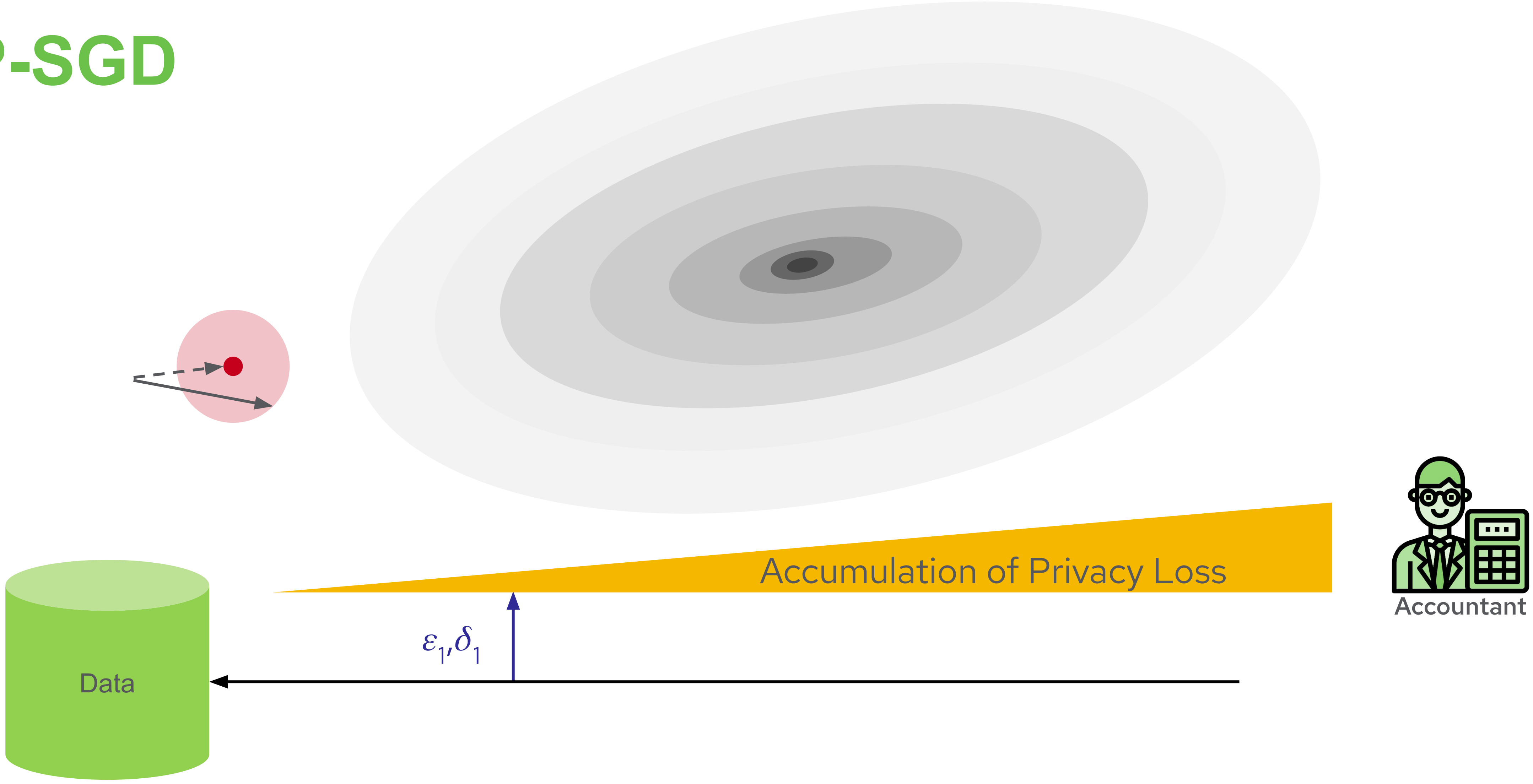
$\theta_{t+1} \leftarrow \theta_t - \eta_t \tilde{\mathbf{g}}_t$

Output θ_T and compute the overall privacy cost (ε, δ) using a privacy accounting method.

DP-SGD

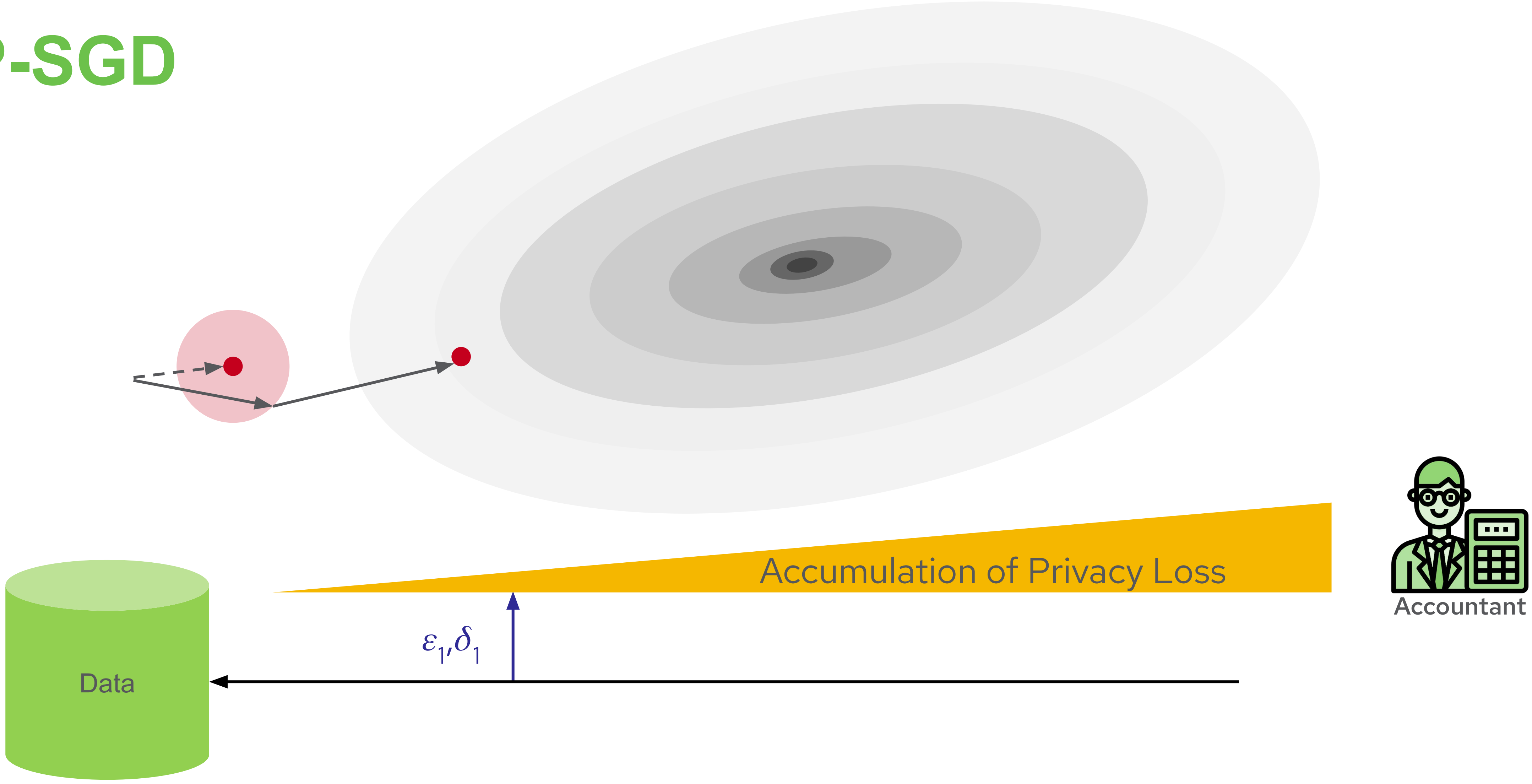


DP-SGD



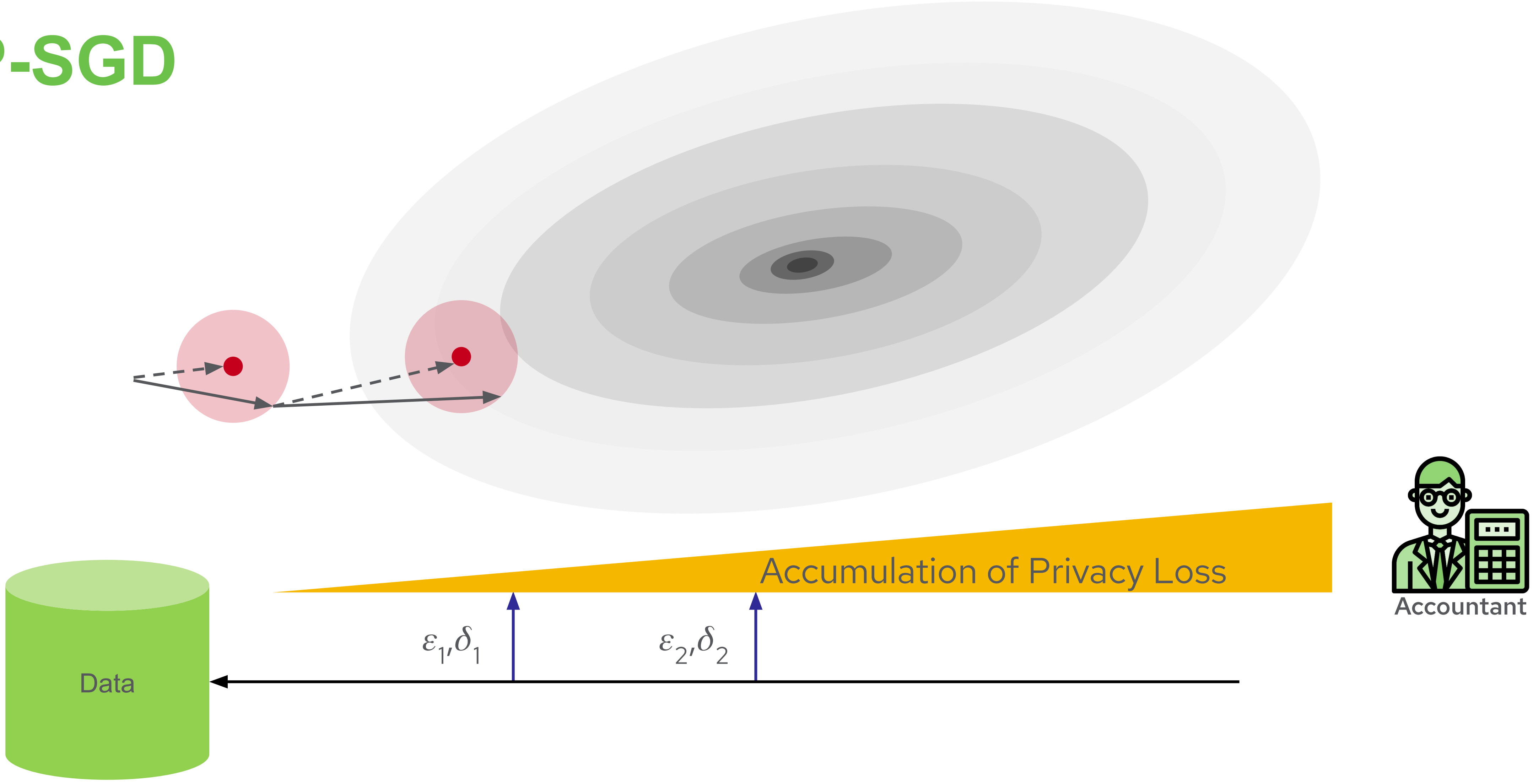
Accumulation of Privacy Loss
Accountant
ε_1, δ_1
Data

DP-SGD



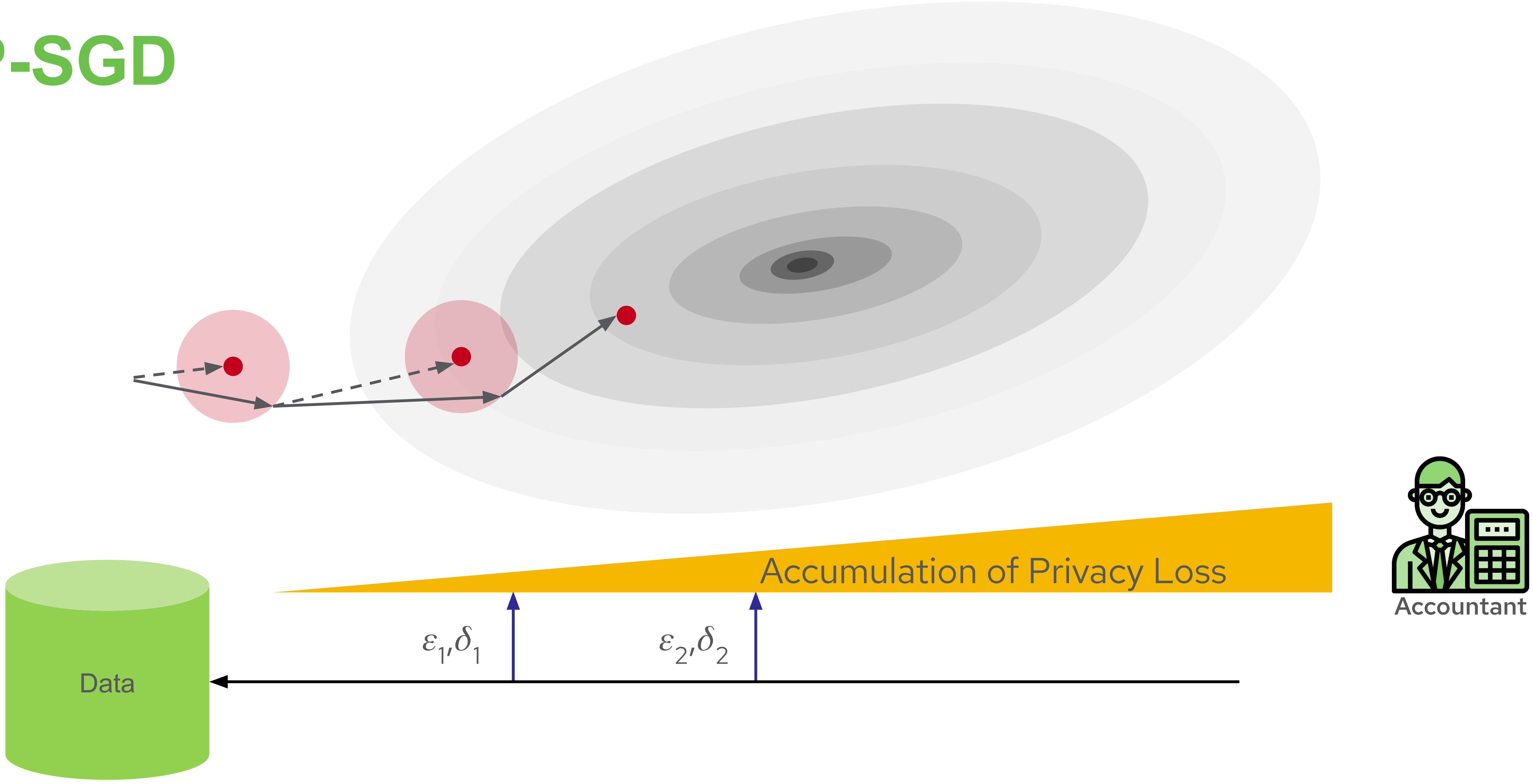
Accumulation of Privacy Loss
Accountant
ε_1, δ_1
Data

DP-SGD

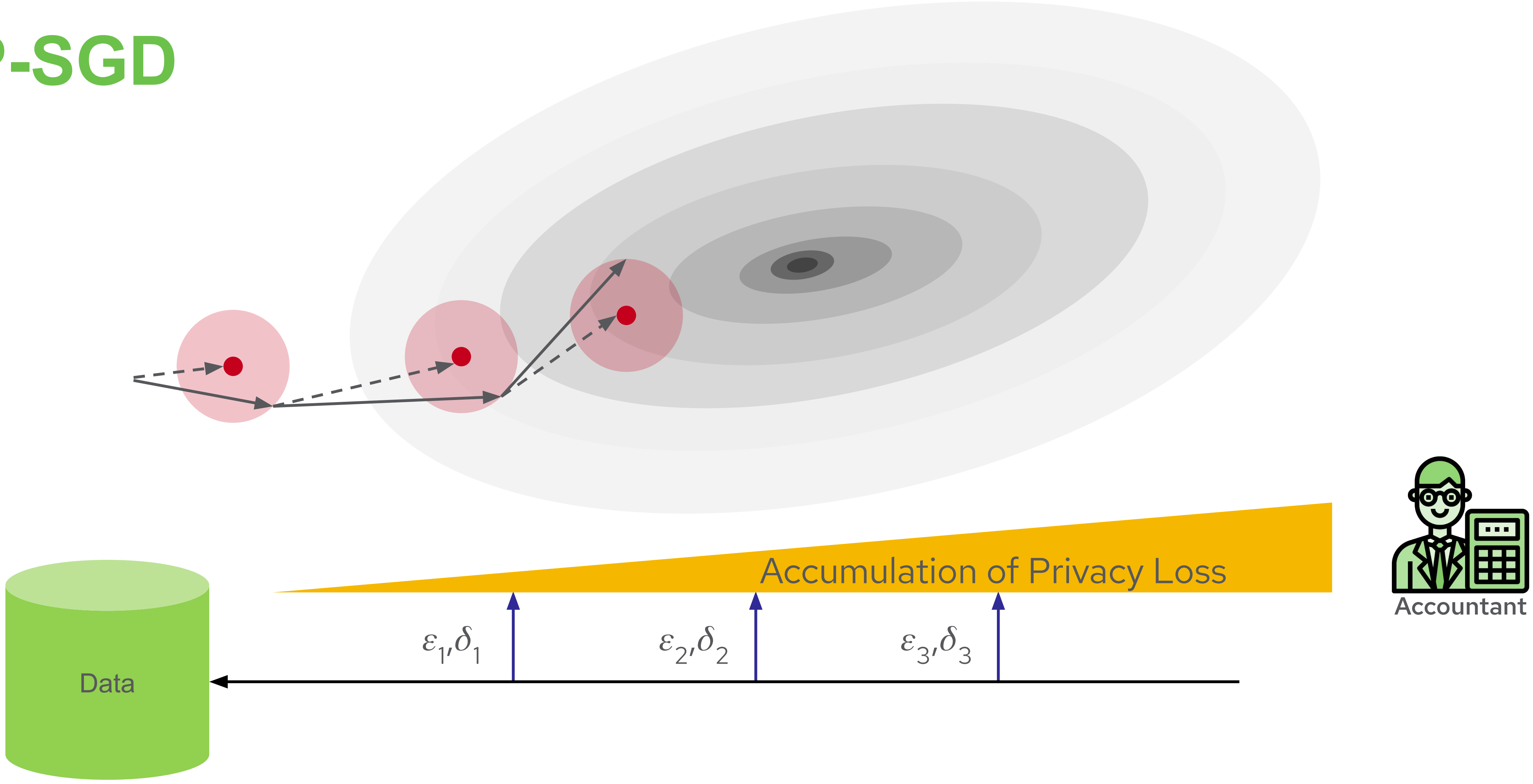


Accumulation of Privacy Loss
Accountant
Data
ε_1, δ_1
ε_2, δ_2

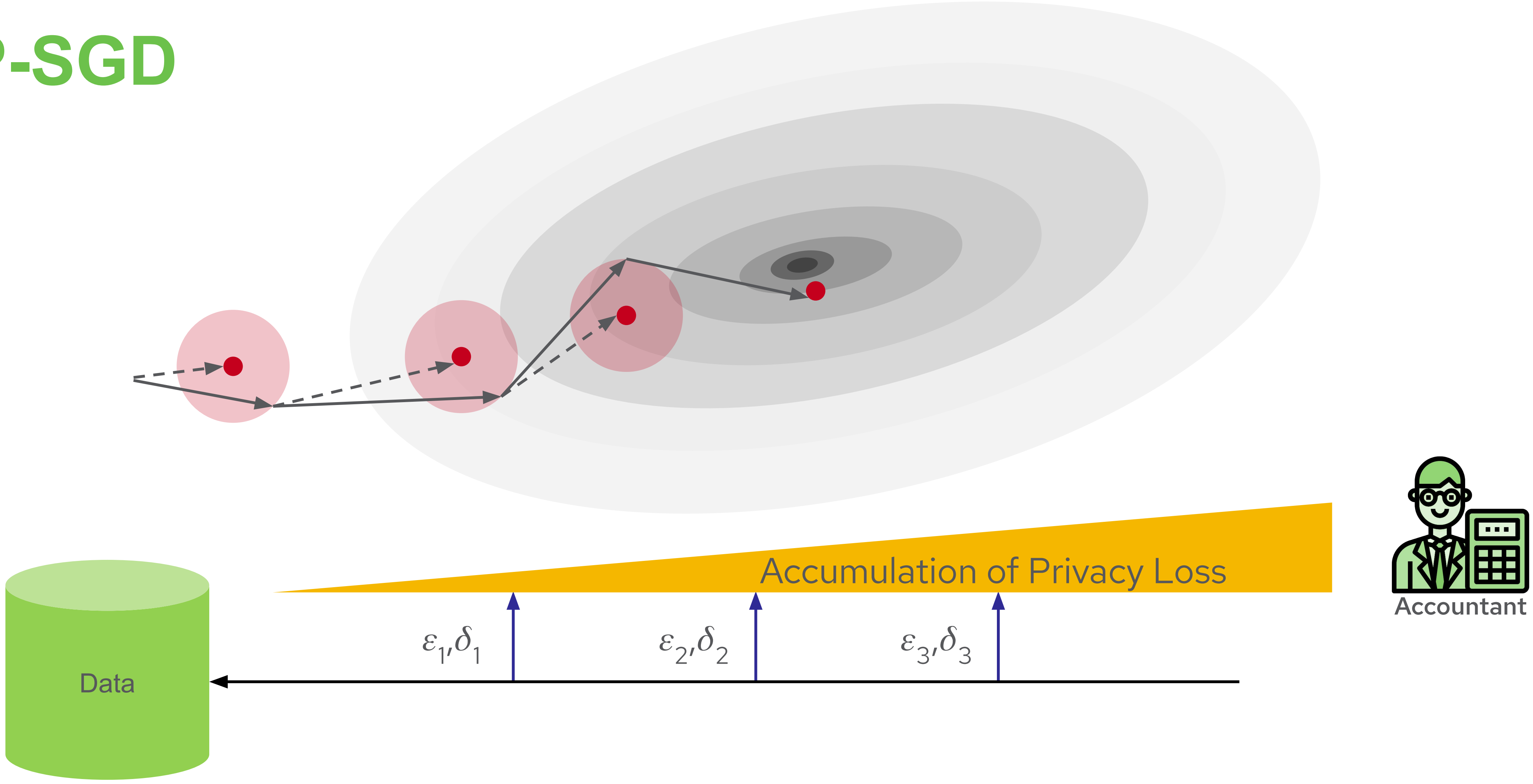
DP-SGD


Accumulation of Privacy Loss
Accountant
ε_1, δ_1
ε_2, δ_2
Data

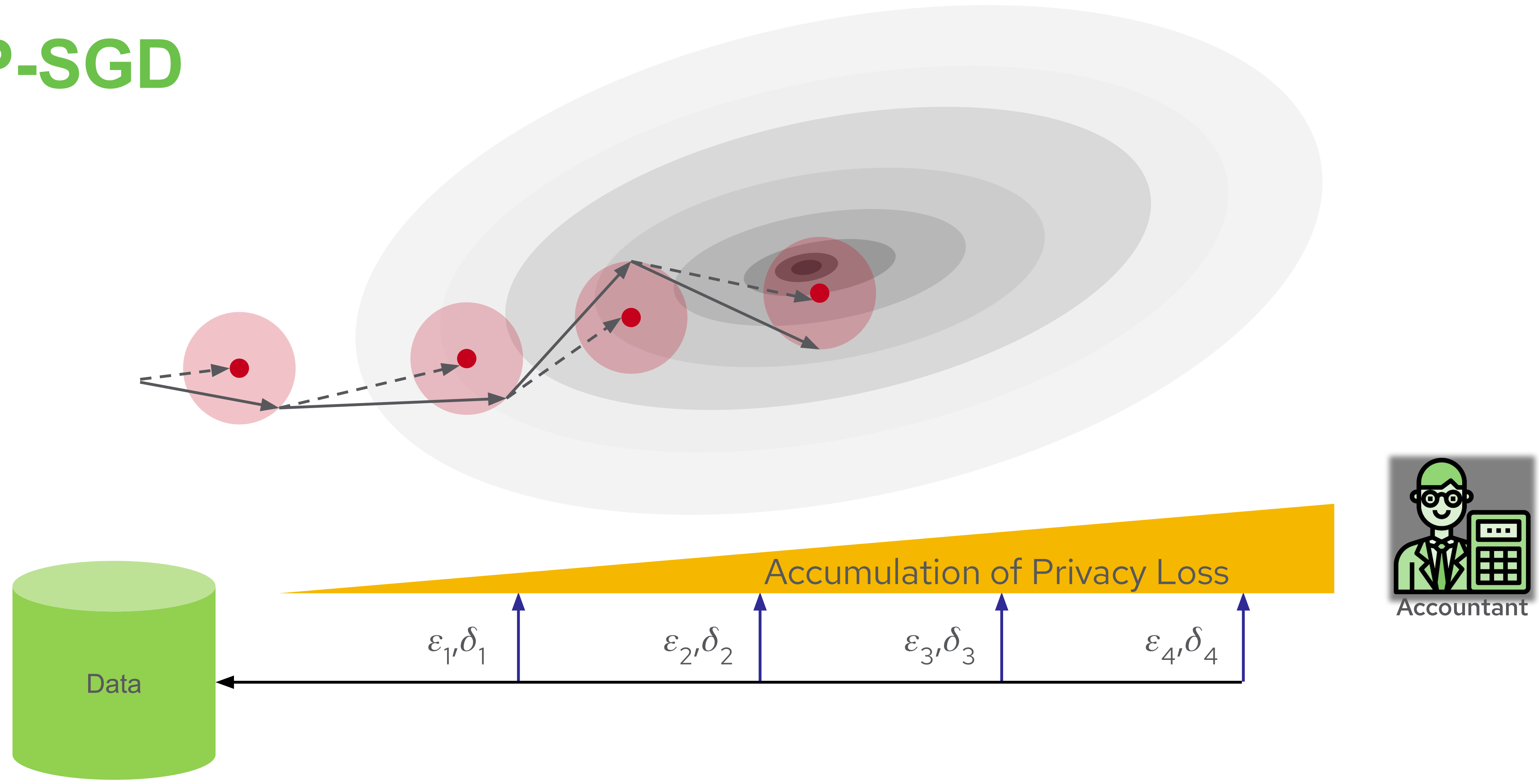
DP-SGD


Accumulation of Privacy Loss
Accountant
ε_1, δ_1
ε_2, δ_2
ε_3, δ_3
Data

DP-SGD


Accumulation of Privacy Loss
Accountant
Data
ε_1, δ_1
ε_2, δ_2
ε_3, δ_3

DP-SGD



Accumulation of Privacy Loss
Accountant
ε_1, δ_1
ε_2, δ_2
ε_3, δ_3
ε_4, δ_4
Data

Large spectrum of possible applications

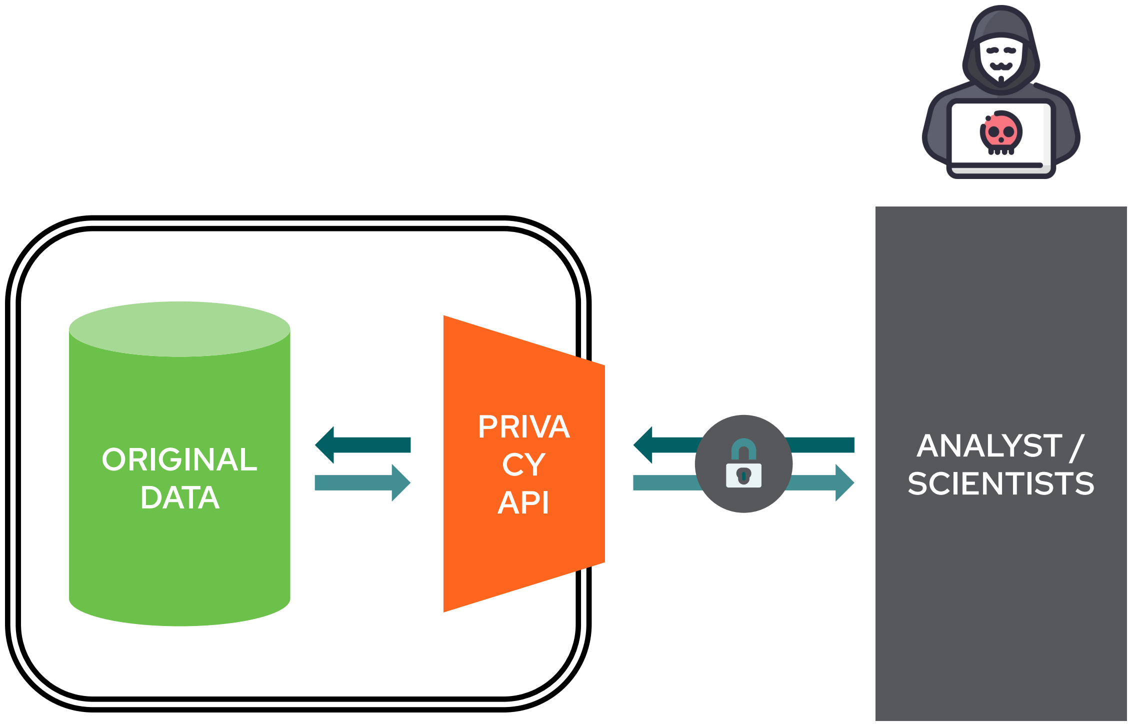
- Analytics
  - Counts, Sums, Averages, Pandas, SQL queries
- Stats
  - PCA, Linear regressions, Logistic regression
- ML
  - Random forests, Boosted trees
- AI
  - DP-SGD
  - Deep-learning

Real-world applications

- US Census bureau
  - *“2020 Census results will be protected using “differential privacy,” the new gold standard in data privacy protection.” (census.gov).* It is the elected standard that can comply with US law: The Census Bureau must keep responses completely confidential.
- Google
  - RAPPOR: Randomized Aggregatable Privacy-Preserving Ordinal Response (security.googleblog.com)
- Apple
  - Apple has adopted and further developed a technique known in the academic world as local differential privacy to do something really exciting: gain insight into what many Apple users are doing, while helping to preserve the privacy of individual users. (apple.com)
- Microsoft (or Linkedin), Uber, Facebook…

Private Analysis in Practice

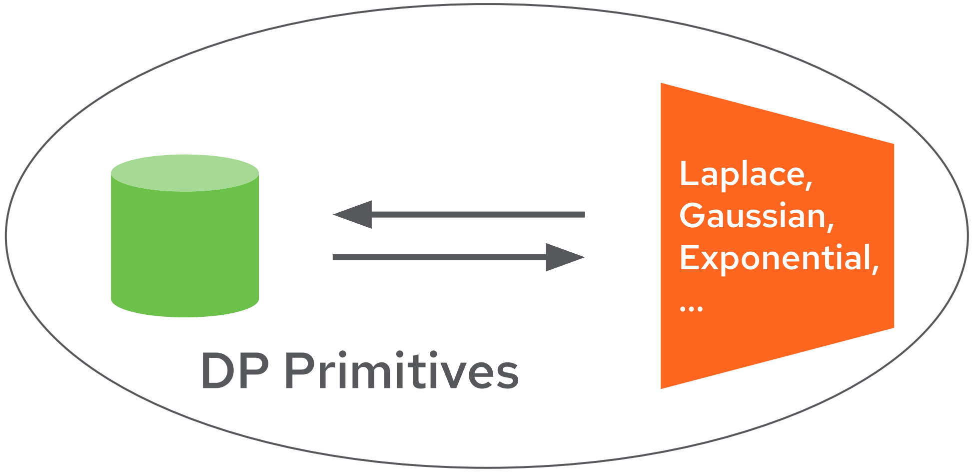
Idea: proxy all inteactions with the data



Use a *Privacy API* to:

- Access catalogs, metadata
- Submit data analysis jobs
- Get results with privacy guarantees
- Preferrably without any workflow disruption

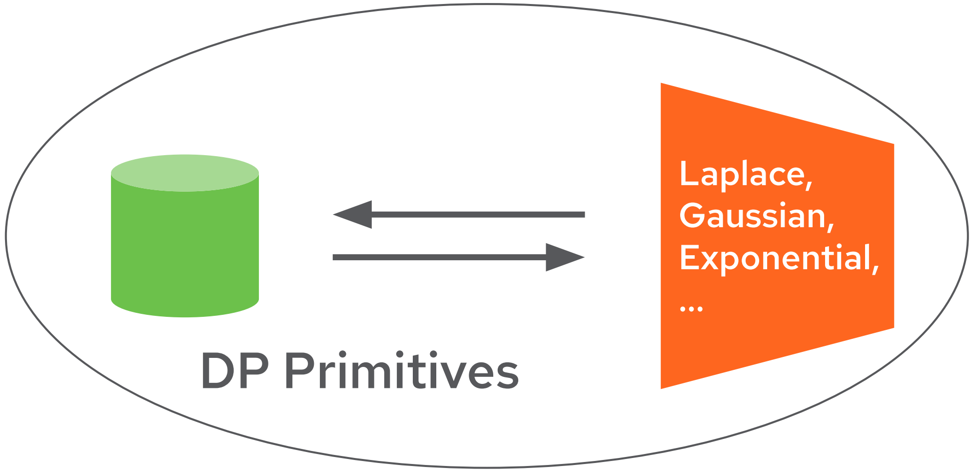
Differential privacy data products



Libraries & implementations

- Main open source libraries
  - **Smartnoise (primitives)**
  - **Google Privacy (primitives)**
  - **IBM Diffprivlib (primitives)**
  - **Others: Brubinstein/diffpriv (primitives in R)**

Differential privacy data products

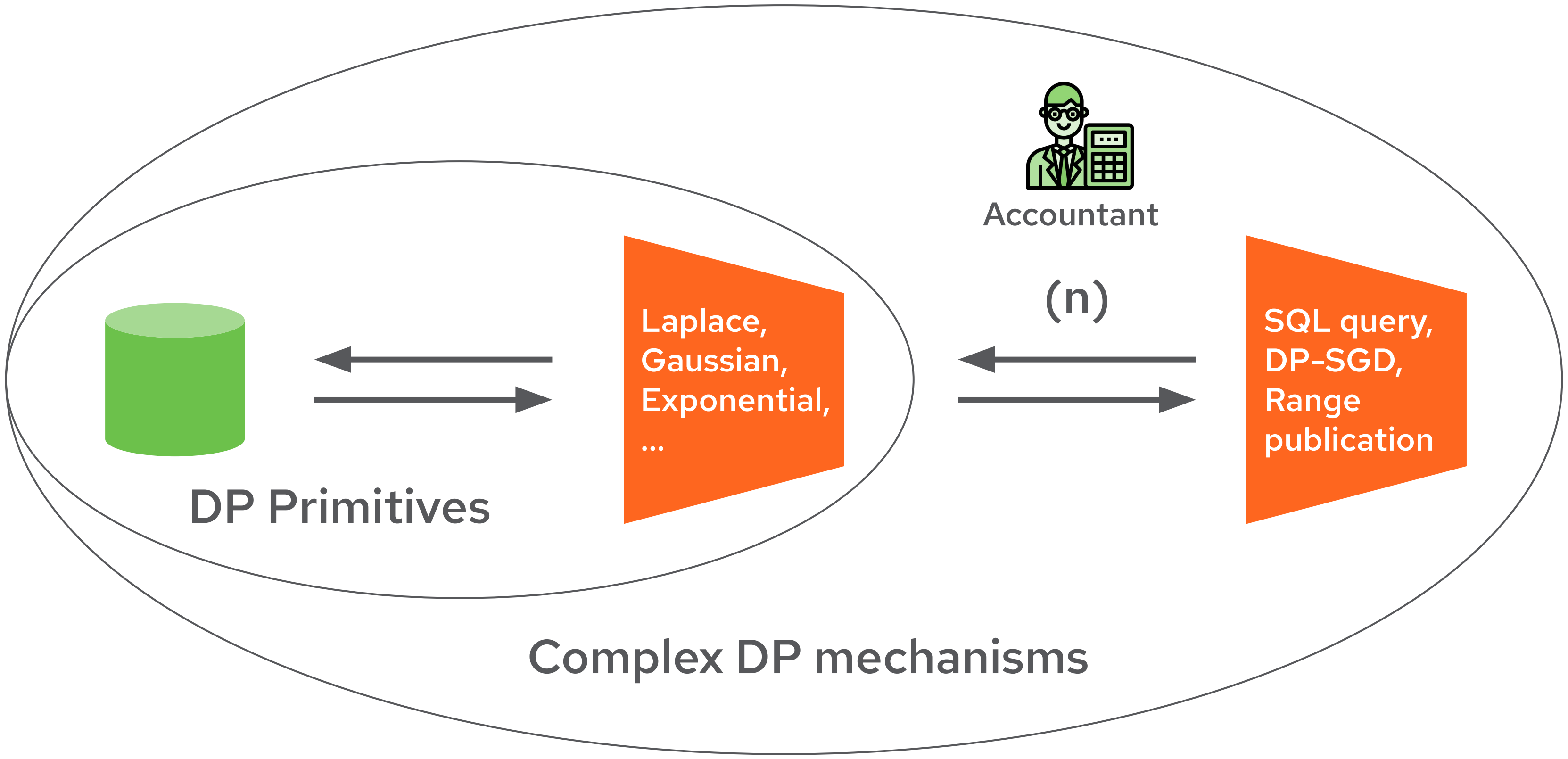


DP Primitives only enable:

- simple computations
- with specific tools
- from a trusted operator.

The result can be safely published.

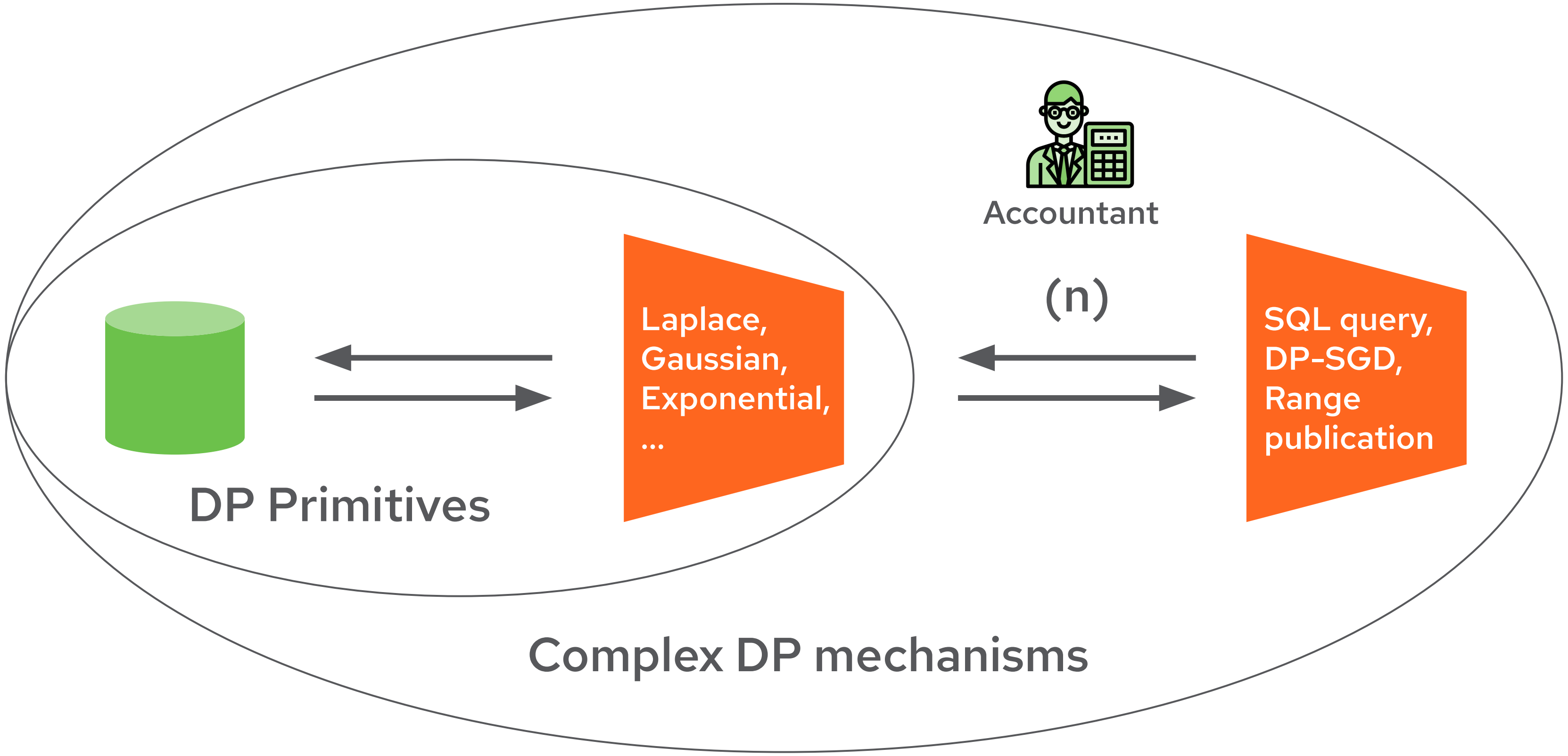
Differential privacy data products



Libraries & implementations

- Main open source libraries
  - Smartnoise (primitives), **smartnoise-sdk (SQL)**
  - Google Privacy (primitives, **SQL**), **Tensorflow-privacy (Deep Learning)**
  - IBM Diffprivlib (primitives, **ML**)
  - **Facebook Opacus (Deep Learning)**
  - Others: Brubinstein/diffpriv (primitives in R), **Uber (SQL), US census (SQL)**

Differential privacy data products



Complex DP mechanisms

- Enable complex queries: SQL, ML, AI
- Privacy loss is computed across the queries
- Result may safely be published

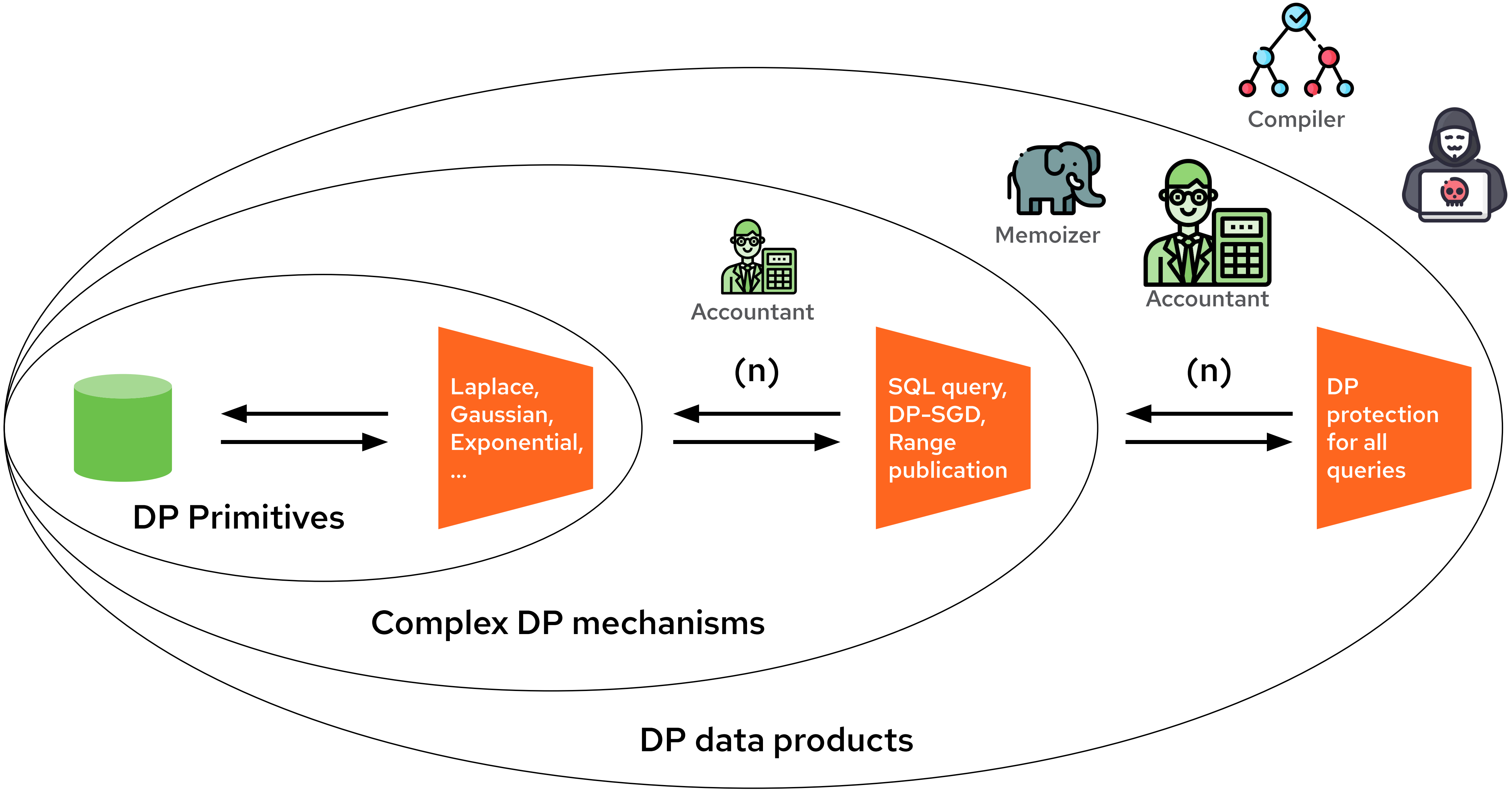
But:

- Specific tools still need to be used
- The operator still need to be trusted

What does a comprehensive framework look like?

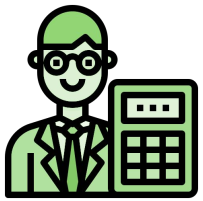
- Permissions and privacy consumption rights should be managed centrally
- Any complex queries should be available: SQL, Pandas, SkLearn, Tensorflow
- Privacy consumption should be optimized across queries
- Anyone should be able to run analysis, not just trusted users
- One should be able to use his usual tools

Differential privacy data products



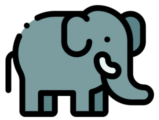
What does a comprehensive framework look like?

- Permissions and privacy consumption rights should be managed centrally
  - Provide a UI to the data owner to manage permissions
  - Enforce permission with a centrall accountant



Accountant

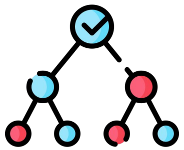
- Any complex queries should be available: SQL, Pandas, SkLearn, Tensorflow
- Privacy consumption should be optimized across queries
  - Remember past queries to save privacy on future queries



Memoizer

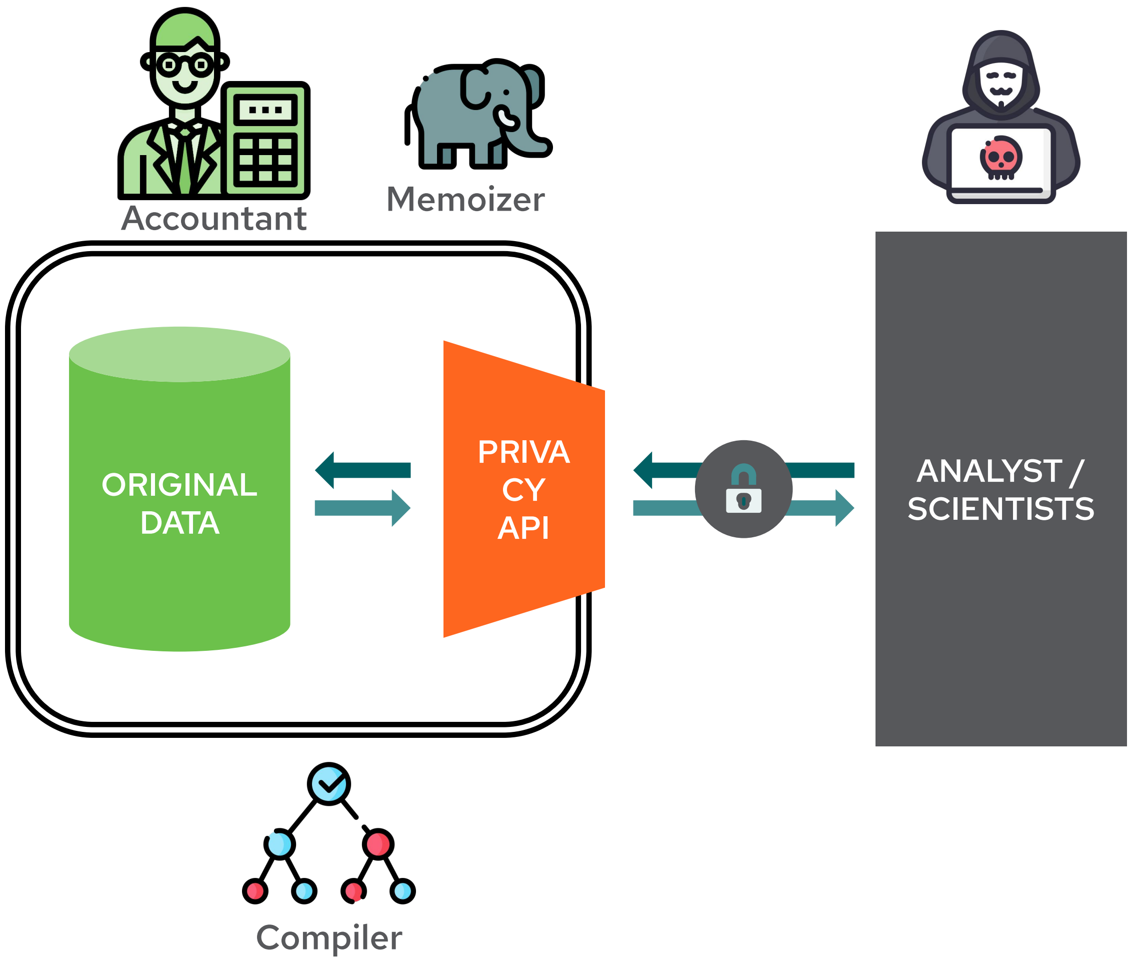
- Anyone should be able to run analysis, not just trusted users
  - The data is accessed through a proxy API

- One should be able to use his usual tools
  - A compiler is used to compile plain pandas + numpy + sklearn into DP ones



Compiler

Proxy all inteactions with the data



Use a *Privacy API* for anyone to:

- Access catalogs, metadata
- Submit data analysis jobs
- Get results with privacy guarantees
- Without any workflow disruption

In [4]:

```
# Fetch by name
dataset = client.dataset(slugname="census")

# Or fetch by id
# dataset = client.dataset(id=6)

print([feature["name"] for feature in dataset.features])
```

Out [4]:

```
['age', 'workclass', 'fnlwgt', 'education', 'education_num', 'marital_status', 'occupation', 'relationship', 'race', 'sex', 'capita
```

In [5]:

```
df = dataset.as_pandas()
```

In [6]:

```
y = df.income
```

In [8]:

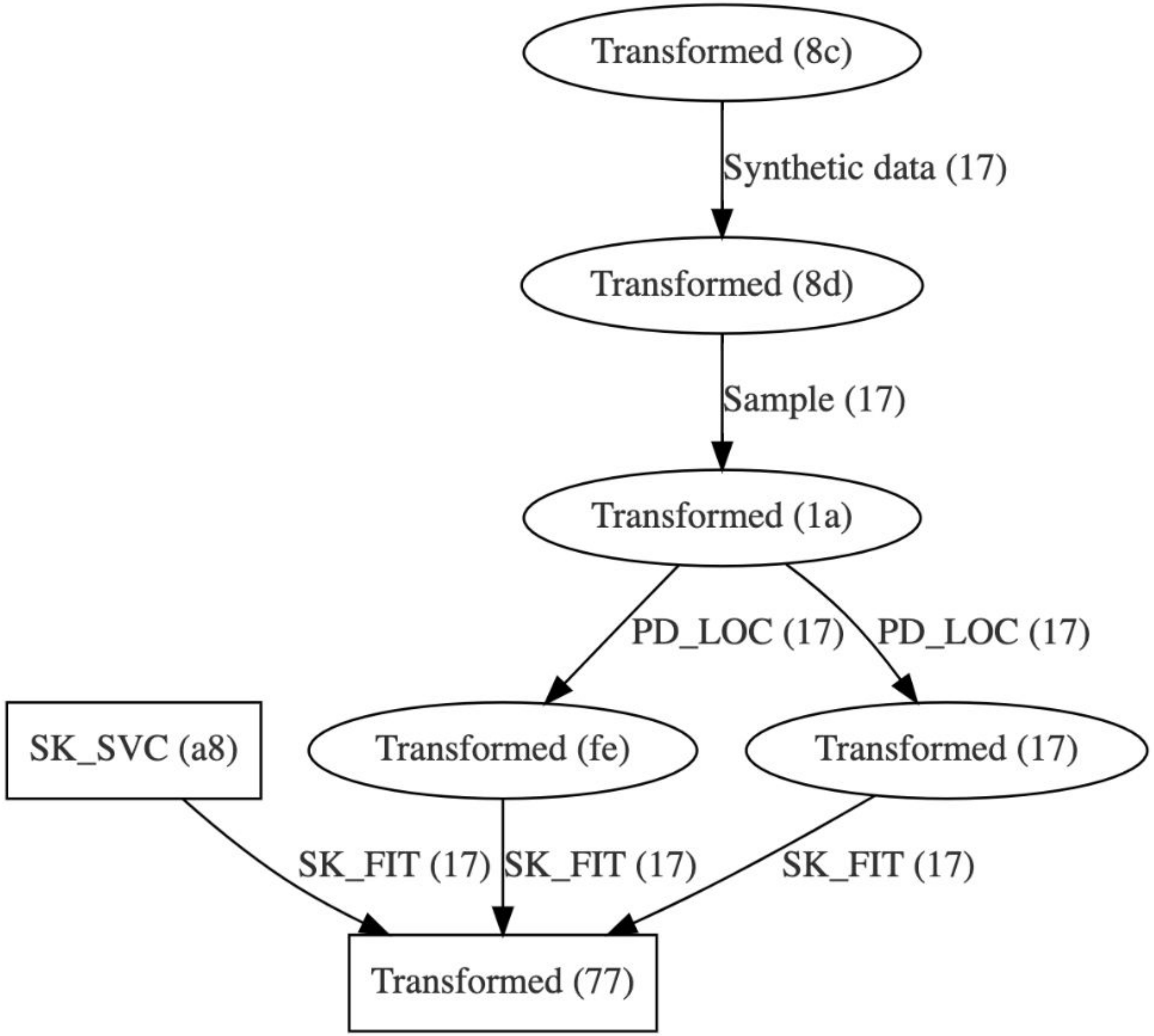
```
X = df.loc[:, ["age", "education_num", "hours_per_week"]]
```

In [9]:

```
from sarus.sklearn.svm import SVC

model = SVC()
fitted_model = model.fit(X=X, y=y)
```

Out [10]:



The data science job is analysed and compiled into a privacy-preserving equivalent

- Some operations are substituted by their DP equivalent
- Some are just executed on DP synthetic data
- DP synthetic data is used as a fall-back

Thank you!

DigiTrans 2022