Competence Network for Artificial Intelligence
Kompetenznetzwerk für künstliche Intelligenz
Réseau de compétences en intelligence artificielle
Rete di competenze per l'intelligenza artificiale

Data Science
Competence Center DSCC

Schweizerische Eidgenossenschaft
Confédération suisse
Confederazione Svizzera
Confederazium svizra

Swiss Confederation

Bundesamt für Statistik BFS
Office fédéral de la statistique OFS
Ufficio federale di statistica UST
Federal Statistical Office FSO



Lomas – A platform for Confidential Data Analysis

Christine Choirat, Raphaël de Fondeville,
Data Science and Artificial Intelligence
Swiss Federal Statistical Office

October 22nd, 2026

Agenda

1. Lomas – A platform for Eyes off Data Science
2. Lomas in practice

Agenda

1. **Lomas – A platform for Eyes off Data Science**
2. Lomas in practice

Secondary Use of Data for Research

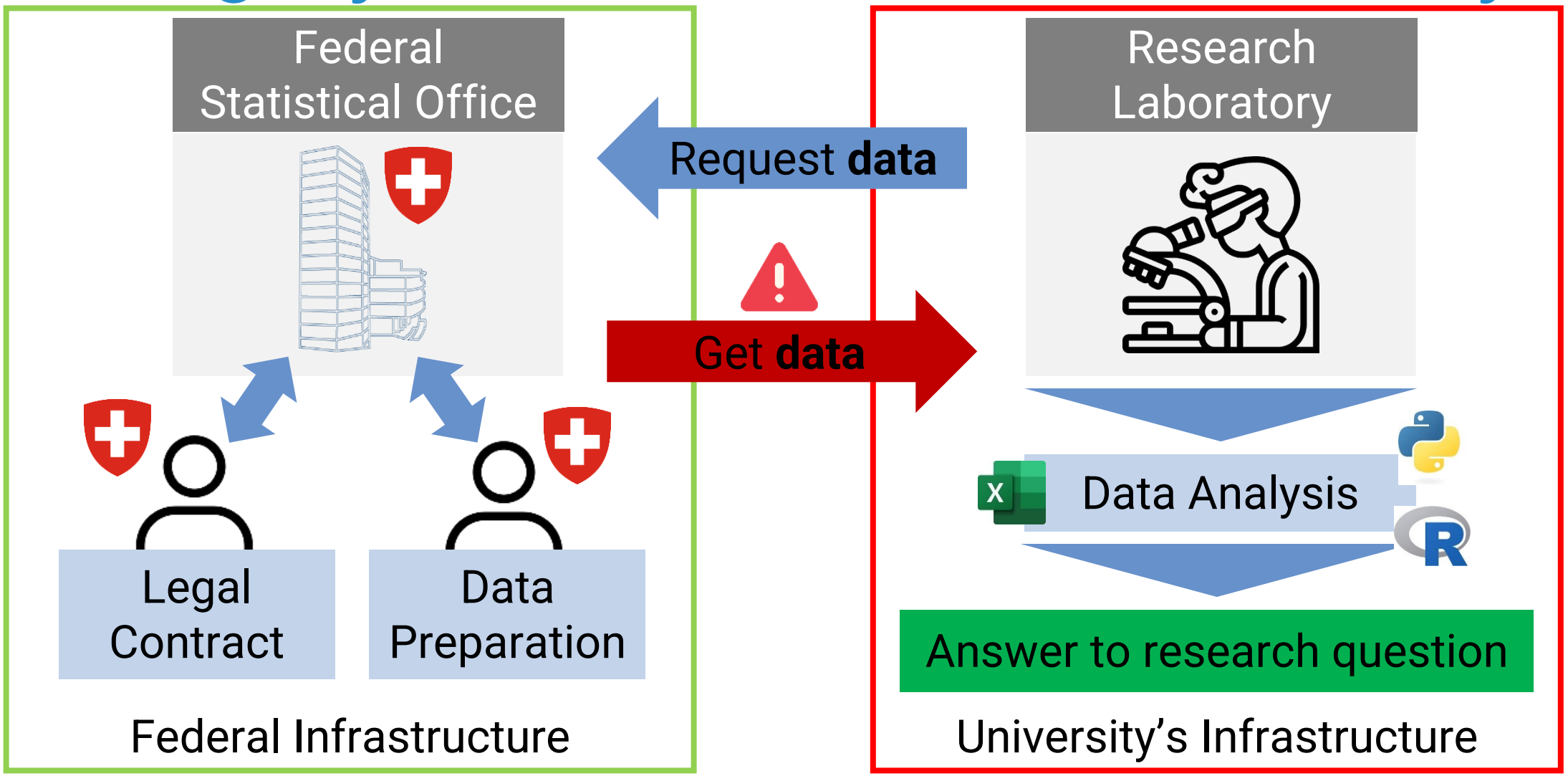
The Federal Statistical Office compelled by law to use data only for the purpose for which it has been collected.

… except …

Ordinance on Statistical Surveys

- Secondary use allowed for **research, planning or statistics**.
- Data is **anonymized** and data subject **not identifiable.**

Legacy Process at FSO: Data delivery

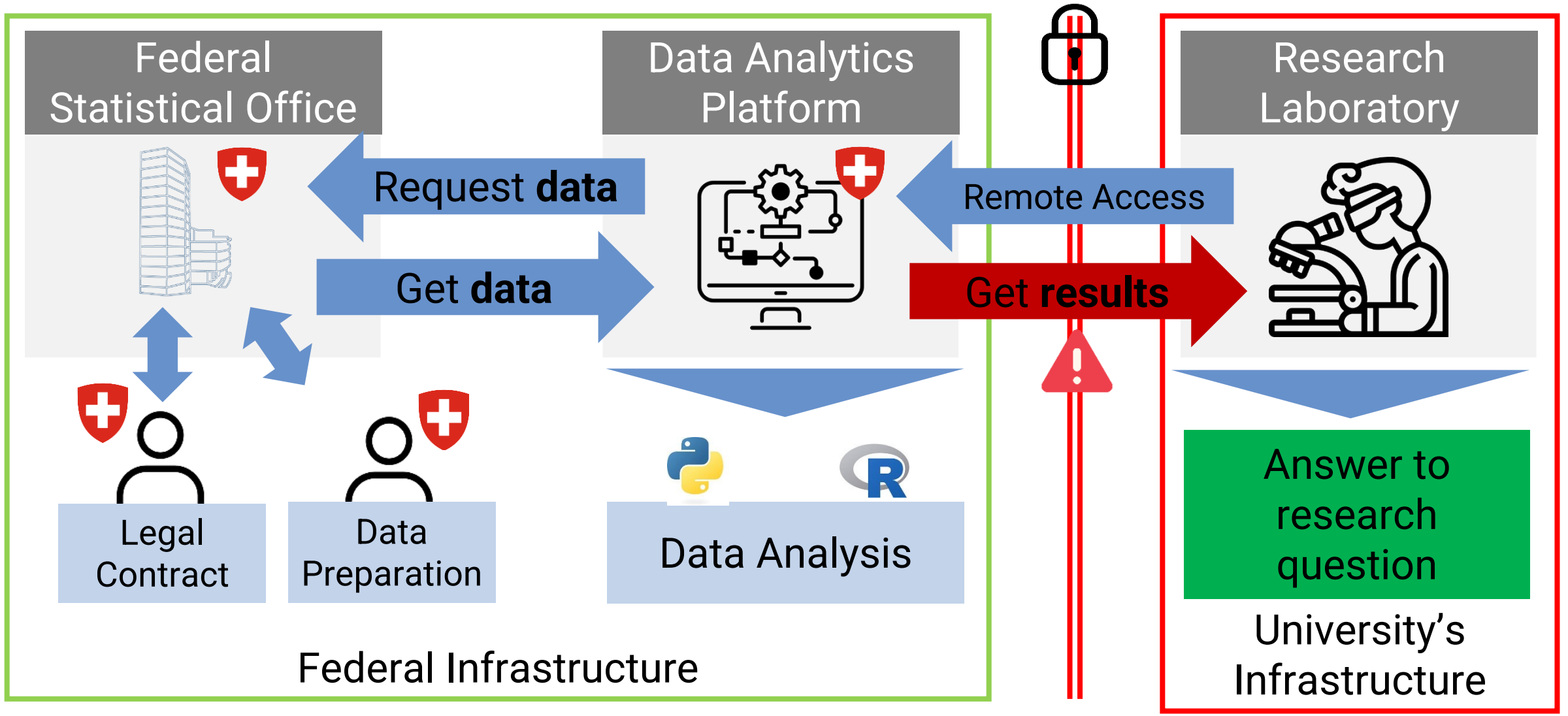


Limits of Data Delivery

Federal Statistical Office	Research Laboratories

• Costly (legal and technical staff).
• Cannot scale to increasing demand.
• Data protection relies on
  • responsability transfer,
  • pseudonymization.
• Reputational risk.
	• Slow process.
	• Scope of data delivered is limited.

Remote Access



Advantages of Remote Access

Federal Statistical Office

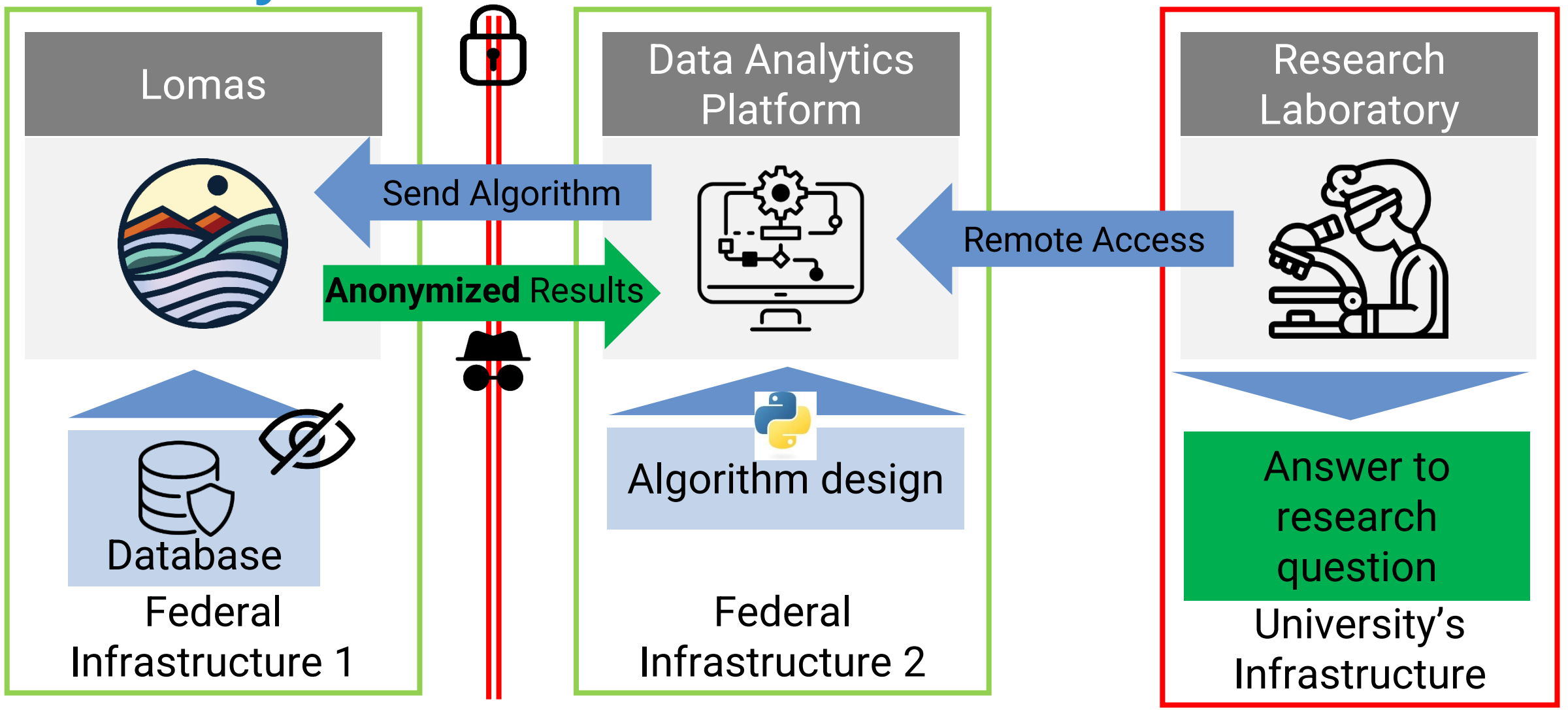
- Data remains within the Federal Network.
- Data protected against security breach on the university side.

- **BUT**
  - Administrative burden for FSO remains unchanged (scalability).
  - Data extraction risk impossible to suppress.

Research Laboratories

- Limited risk exposure.

«Eyes-off» Data Science



Advantages of «Eyes-off» Data Science

Federal Statistical Office

- Minimal human intervention.
- Can scale with infrastructure.
- Reputational risk is limited as no direct access to data.

BUT

- Necessitate to quantify and track disclosure risk of results.

Research Laboratories

- Faster process (starting to prototype is a matter of days).
- Wider access to data.
- Offers a standard environment for replicability and (statistical) reproducibility.

Lomas: The platform for confidential data analysis

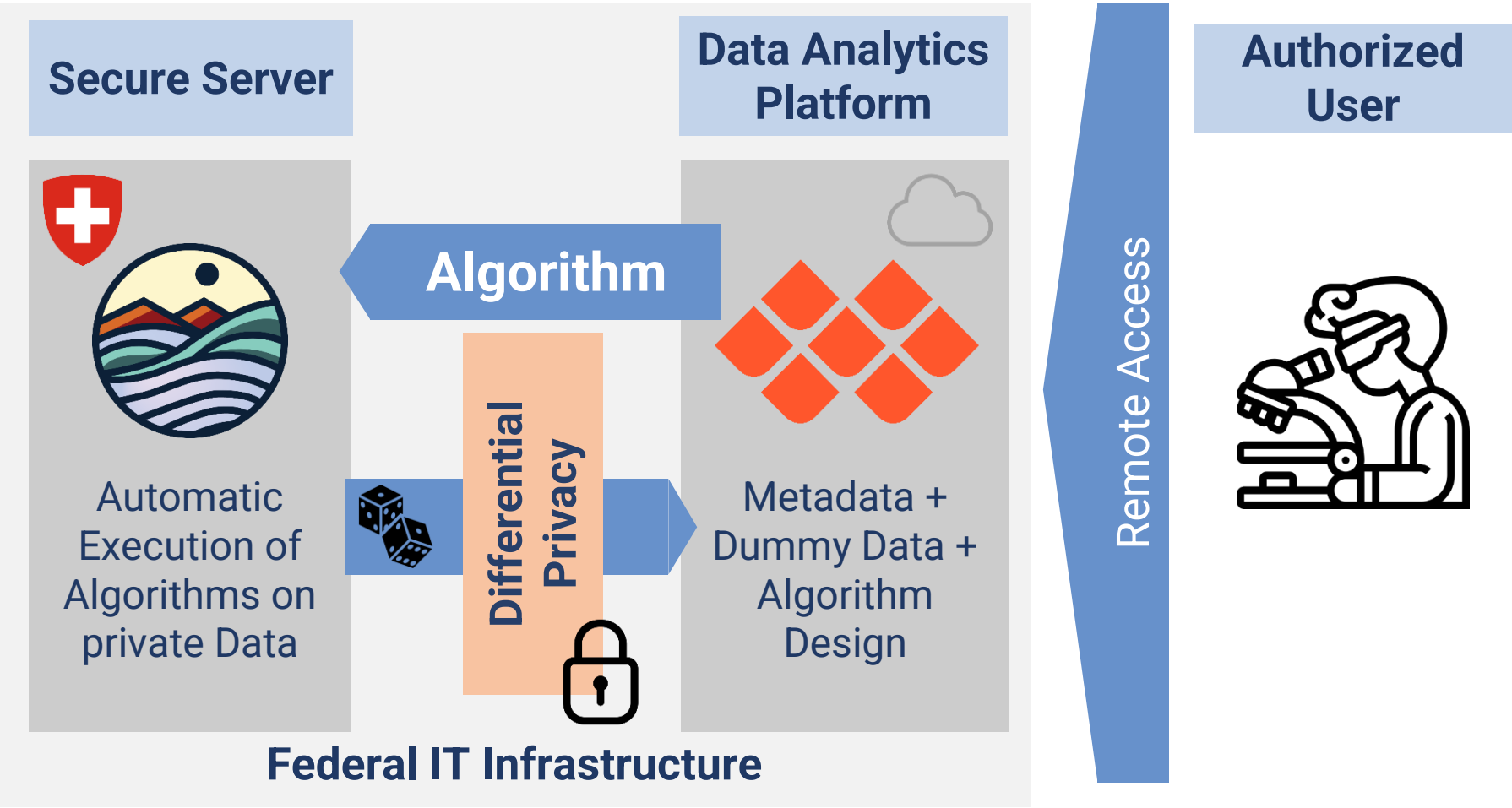
Aim

Enabling data (secondary) use (for research) at scale without data sharing and while automatically controlling disclosure risk.

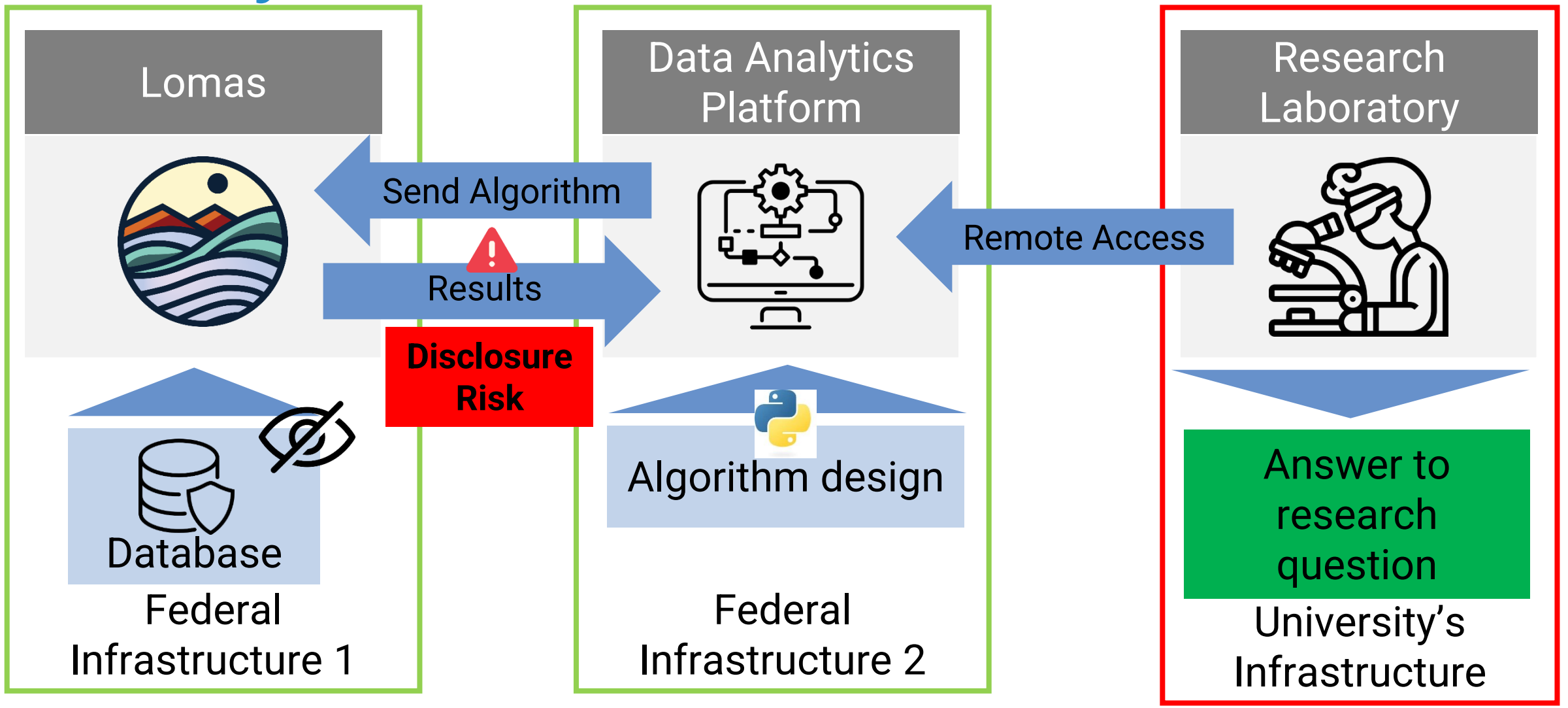
Principles

1. Remote access
2. “Eyes-off” Data Science
3. Automatic Output checking
4. Standard, modular and portable

Lomas: The platform for confidential data analysis



«Eyes-off» Data Science



Example of a DP Mechanism
The Noise Addition

Sex	Region	...	HIV?
M	1005		Postive
F	2012		Negative
F	8038		Negative
M	1503		Negative
F	1220		Positive
M	3807		Positive

Database D_{HIV+}

Differential Privacy Mechanism

Query q

q: «How many individuals are HIV positive in Bern?»

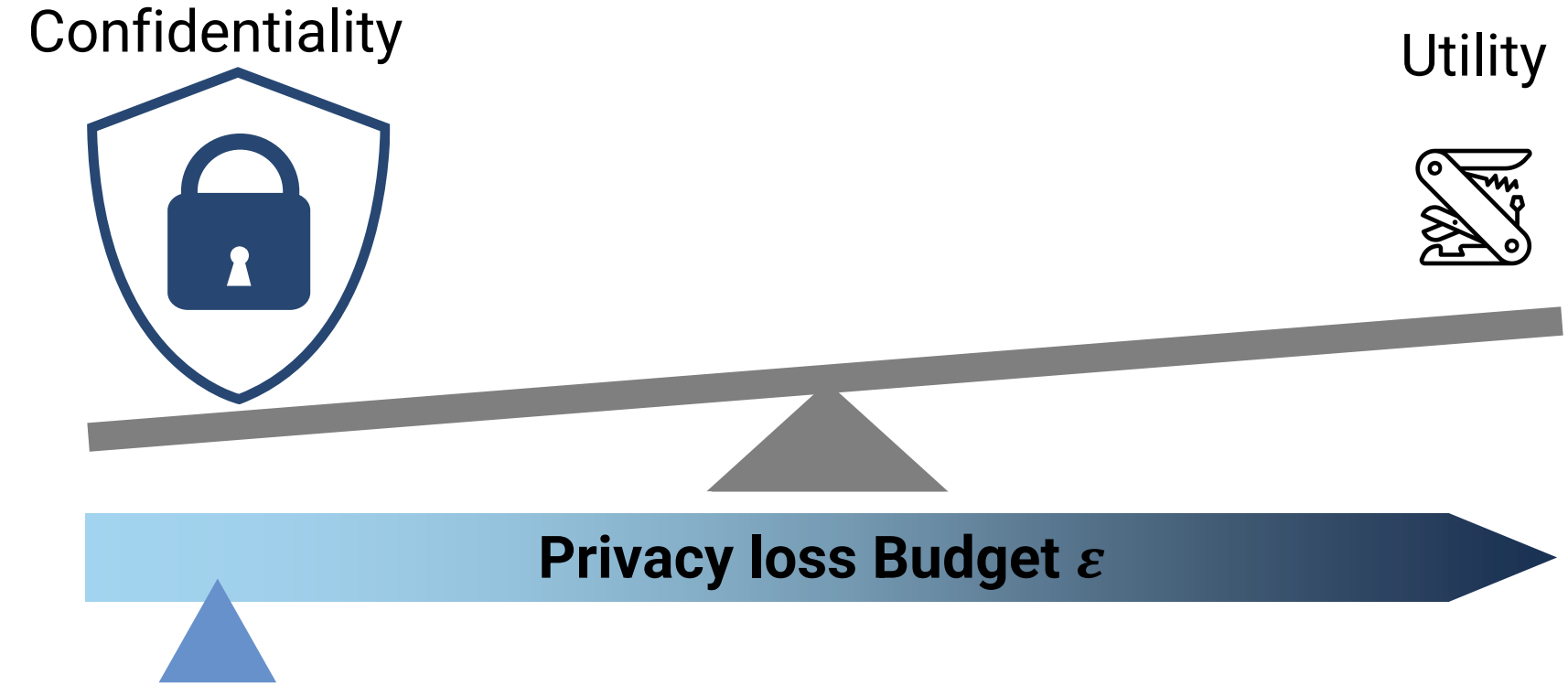
DP Mechanism Q

$$Q(D_{HIV+}) = q(D_{HIV+}) + \textbf{Noise}(\varepsilon)$$

Public Release (Adversaries)

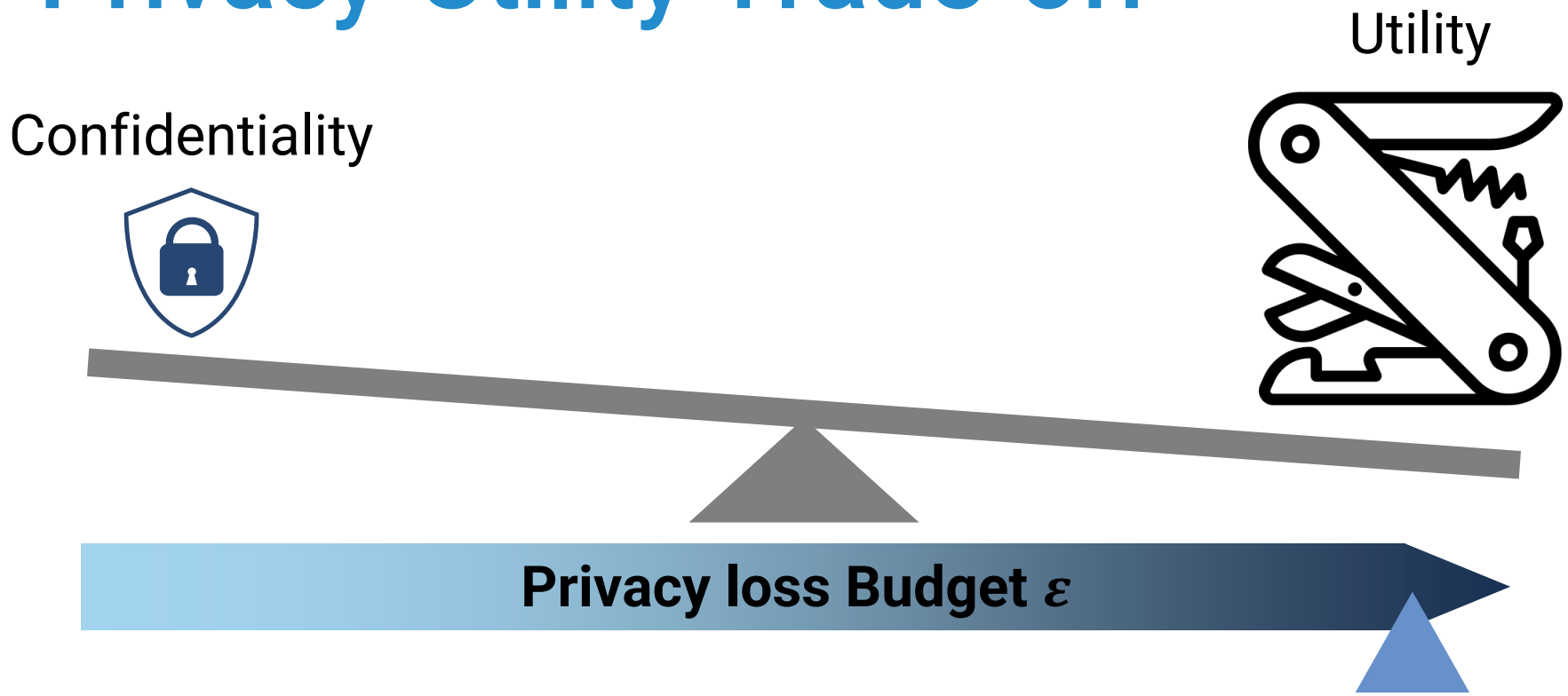


Privacy-Utility Trade-off



Differential Privacy is a tool to control the privacy-utility trade-off.

Privacy-Utility Trade-off



Differential Privacy is a tool to control the privacy-utility trade-off.

Differential Privacy: Composition

$Q_1(D)$

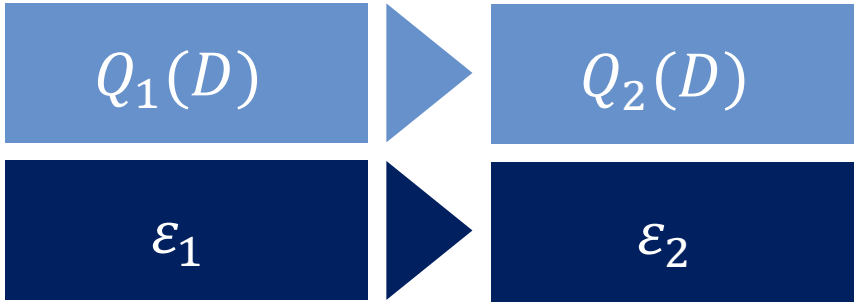
ε_1

Total privacy budget

$\varepsilon_1 \leq \varepsilon$

Overall pricacy loss budget needs to be tracked!

Differential Privacy: Composition

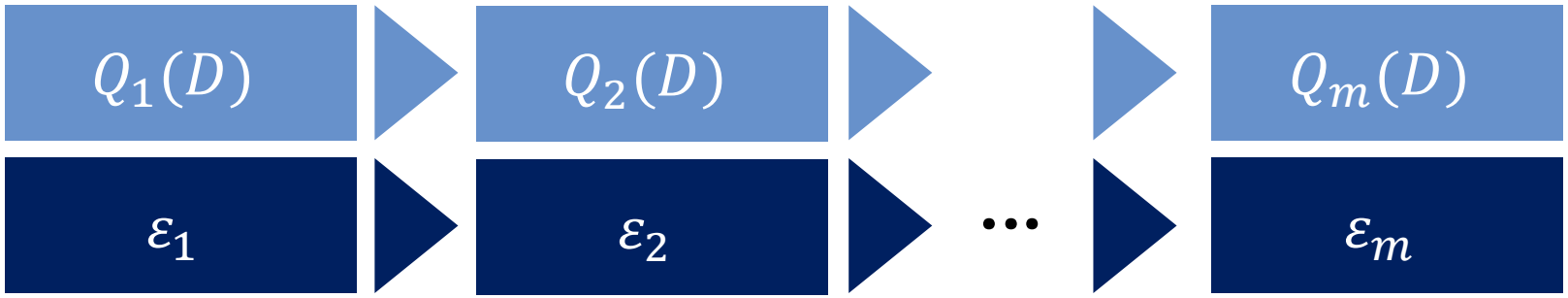


Total privacy budget

$\varepsilon_1 \leq \varepsilon$ → $\varepsilon_1 + \varepsilon_2 \leq \varepsilon$

Overall pricacy loss budget needs to be tracked!

Differential Privacy: Composition

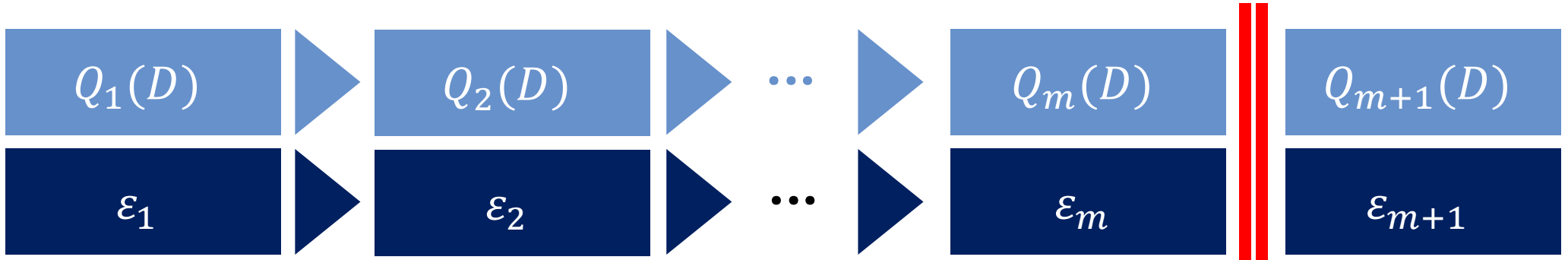


Total privacy budget

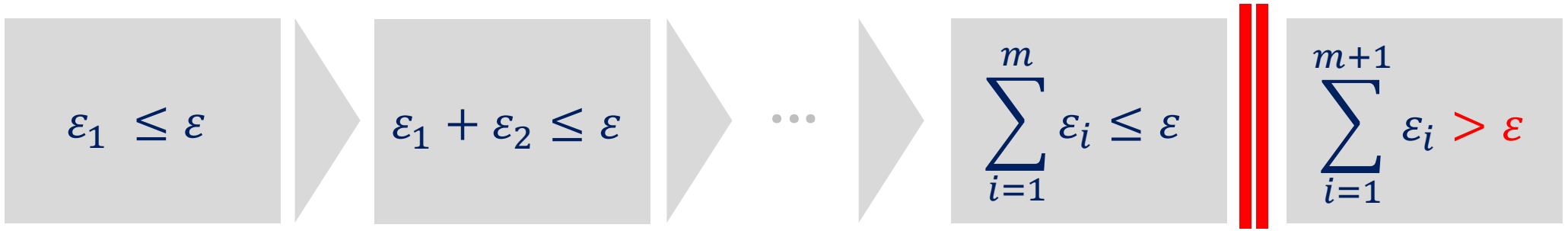
$\varepsilon_1 \leq \varepsilon$ → $\varepsilon_1 + \varepsilon_2 \leq \varepsilon$ → ... → $\sum_{i=1}^{m} \varepsilon_i \leq \varepsilon$

Overall pricacy loss budget needs to be tracked!

Differential Privacy: Composition



Total privacy budget



Overall pricacy loss budget needs to be tracked!

Statistical Reproducibility

Does the addition of noise prevent drawing reliable conlusions ?

Statistician's Answer

- Noise is added to the algorithm's output so that it does not introduce bias, only increased **variability**.
- By accounting for noise in the **experimental design**, conclusions can remain reliable even if the results are not.

Pragmatic Answer

- Noise scale is determined by the target audience:
  - Researchers: conduct analysis with arbitrarily small noise.
  - Publication: only noisy results are published.
- Offer possibility to reviewers to reproduce experiment.

Lomas: an Open-source Project

Lomas documentation provides extensive explanations for researchers, with detailed tutorials and explanation for all the supported DP libraries and queries.

All the code and other ressources are available on Github.

We are happy to organize a demo/workshop if there is an interest.

Lomas is a public good, meant to be shared and reused!

Agenda

1. Lomas – A platform for Eyes off Data Science
2. **Lomas in practice**

Lomas for secondary use of data 2025

Pilot project on real health data (hospitalization):

- Partnership with 3 Swiss universities.
- Research questions using FSO's «MS data» answered with Lomas.
- Secure deployement on federal infrastructure.

Aim

Demonstrate the FSO's ability to create and operate a platform for the secondary use of data in the health sector.

Research Project 1: Unisanté

Childbirth practices in Switzerland

The project aims to analyze the evolution of **cesarean section** (CS) practices based on various demographic, socioeconomic, and temporal factors. In particular, we aim to determine whether these **practices vary on weekends and public holidays**, which could be linked to fluctuations in hospital resource availability in Switzerland.

- Is there a pattern of reduced births on weekends and public holidays, and to what extent is this influenced by CS scheduling practices or individual characteristics?
- How do CS delivery patterns respond to school admission thresholds, and are these patterns influenced by individual characteristics?

Research Project 2: ZHAW

Prostate Cancer and Radical Prostatectomy (RP)

The project aims to understand the evolution of the practice of **radical prostatectomy** as treatment for prostate cancer and its impact on survival and subsequent treatments.

- What is the evolution of the inpatient treated prostate cancer prevalence?
- What is the evolution of prostate cancer treated with RP in Switzerland?
- What is the mean/median survival after RP?
- What are the subsequent outpatient treatments after RP?

Research Project 3: UZH

Multiple Sclerosis

The project aims to understand how the frequency, nature, and reasons for hospitalizations change in the years before and after a **Multiple Sclerosis (MS) diagnosis.**

- What is the number of MS diagnosis per year?
- What is the mean number of hospitalization per year before and after diagnosis?
- What are the comorbidities 5 years before versus after a diagnostic of MS?

Governance Framework

- Data use is approved by «project».
- Explicit experimental design are requested at application time.
- Privacy-loss budget are allocated under the «utility-first» principle.

Utility-first

Determine the minimal privacy-loss budget that needs to be spent to ensure that reliable conclusion can still be drawn.

- Utility and reliability of conclusions are evaluated by analysis of the statistical power of hypothesis tests.

Multiple sclerosis in Switzerland
Deep dive

Longitudinal study to understand the impact of a MS diagnosis on patients trajectories.

- What is the number of MS diagnosis per year?
- What is the mean number of hospitalization per year before and after diagnosis?
- What is the main treatment 5 years before versus after a diagnostic of MS?
- What are the comorbidities 5 years before versus after a diagnostic of MS?
- What is the age distribution of the patients?

Multiple sclerosis in Switzerland
Deep dive

Longitudinal study to understand the impact of a MS diagnosis on patients trajectories.

- What is the number of MS diagnosis per year? → **COUNTS**
- What is the mean number of hospitalization per year before and after diagnosis? → **AVERAGE**
- What is the main treatment 5 years before versus after a diagnostic of MS? → **COUNTS**
- What are the comorbidities 5 years before versus after a diagnostic of MS? → **COUNTS**
- What is the age distribution of the patients? → **QUANTILE**

Multiple sclerosis in Switzerland

Deep dive: query example I

Number of MS diagnosis per year

```
query = """
SELECT
    year_of_hospitalisation,
    age_group,
    COUNT(DISTINCT patient_number) AS n_patients
FROM
    df
WHERE
    after_diagnosis = TRUE
GROUP BY
    year_of_hospitalisation,
    age_group
"""
```

```
res = client.smartnoise_sql.query(
    query,
    dummy = False,
    nb_rows = NB_ROWS,
    epsilon=1.5,
    delta=0,
    mechanisms={"count": "laplace"}
)
```

	year_of_hospitalisation	n_patients
0	2010	2541
1	2011	2765
2	2012	2700
3	2013	2911
4	2014	3084
5	2015	3245

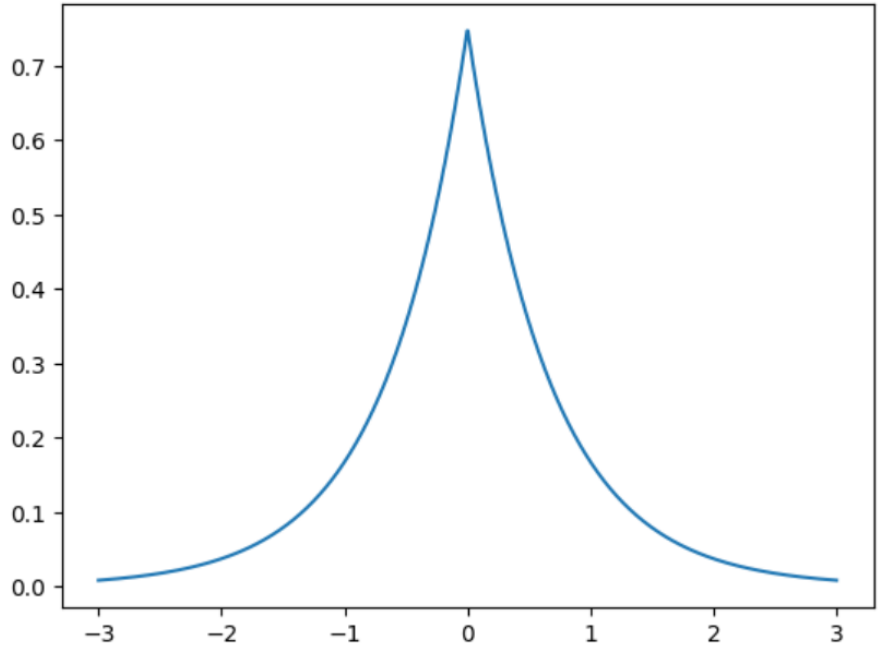
Query → Lomas → DP Protected Results

Multiple sclerosis in Switzerland

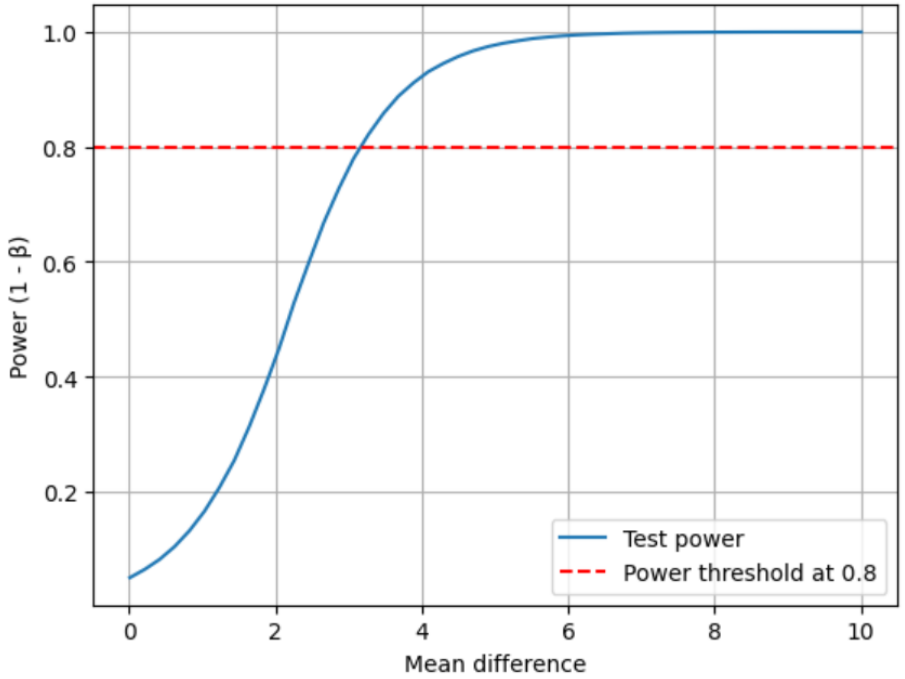
Deep dive: DP budget selection I

Utility analysis for privacy-loss budget $\varepsilon = 1.5$

Noise distribution



Statistical power between years



Multiple sclerosis in Switzerland

Deep dive: query example II

Average Number of Hospitalisations per Year after MS diagnosis

```
query = """
SELECT
    age_group,
    sex,
    region,
    after_diagnosis,
    AVG(nb_hopitalizations_in_year) AS avg_nb_hosp
FROM
    df
WHERE
    first_hospitalisation_in_year = TRUE
GROUP BY
    age_group,
    sex,
    region,
    after_diagnosis
"""
```

```
res = client.smartnoise_sql.query(
    query,
    dummy = False,
    epsilon=10,
    delta=0.01,
    mechanisms={"sum": "gaussian"}
)
```

	age_group	sex	after_diagnosis	avg_nb_hosp
0	0-9	1	False	8.267264
1	0-9	2	True	5.734952
2	10-19	2	True	2.355486
3	20-29	1	False	13.485484
4	20-29	2	False	12.163221
5	20-29	2	True	16.136499

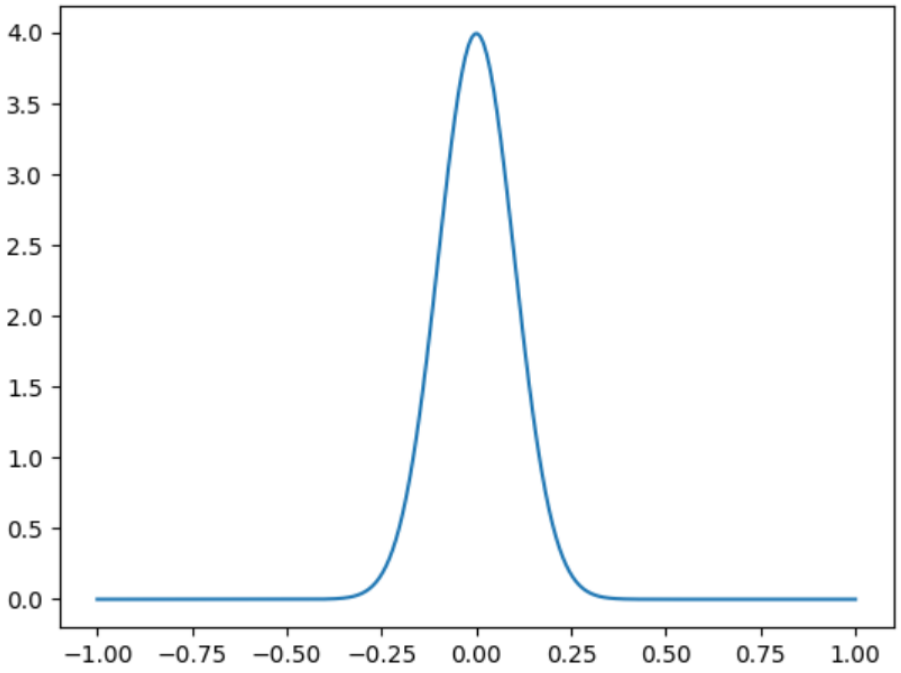
Query → Lomas → DP Protected Results

Multiple sclerosis in Switzerland

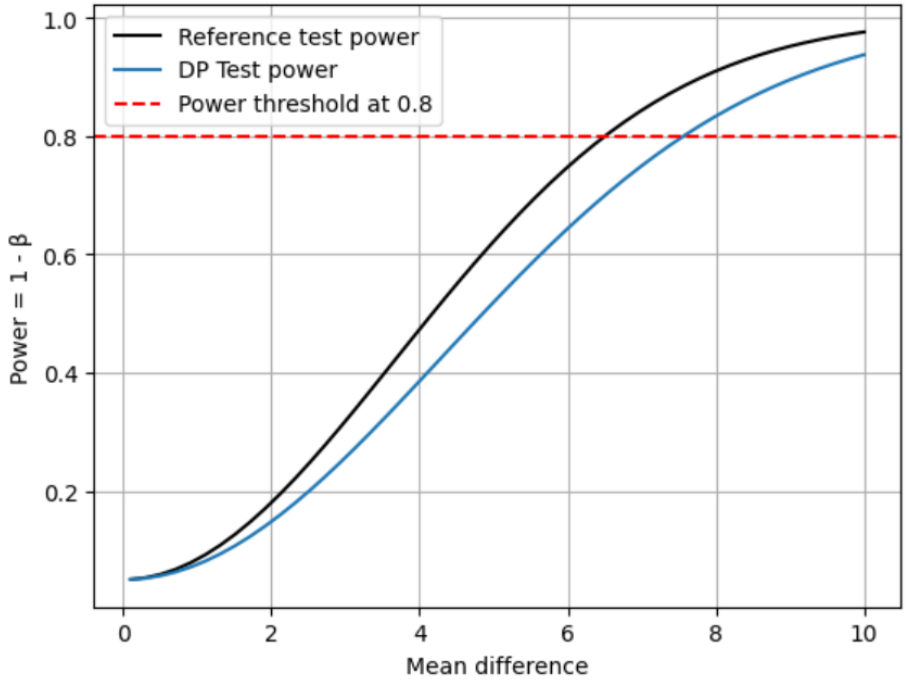
Deep dive: DP budget selection II

Utility analysis for privacy-loss budget $\varepsilon = 10$ and $\delta = 0.01$

Noise distribution



Statistical power between years



Multiple sclerosis in Switzerland

Deep dive: query example II

Median and quartiles age per year

```
plan = (
    dummy_lf
    .group_by("year_of_hospitalisation")
    .agg([
        pl.col("age").dp.median(age_candidates, scale=20.),
        pl.col("age").dp.quantile(0.25, age_candidates, scale=20.),
        pl.col("age").dp.quantile(0.75, age_candidates, scale=20.),
    ])
)
```

```
res = client.opendp.query(plan)
```

year_of_hospitalisation	median_age	Q1_age	Q3_age
i64	f64	f64	f64
2011	53.0	41.7	64.9
2017	56.0	44.3	66.5
2010	52.4	40.2	63.7
2014	54.0	41.2	65.0

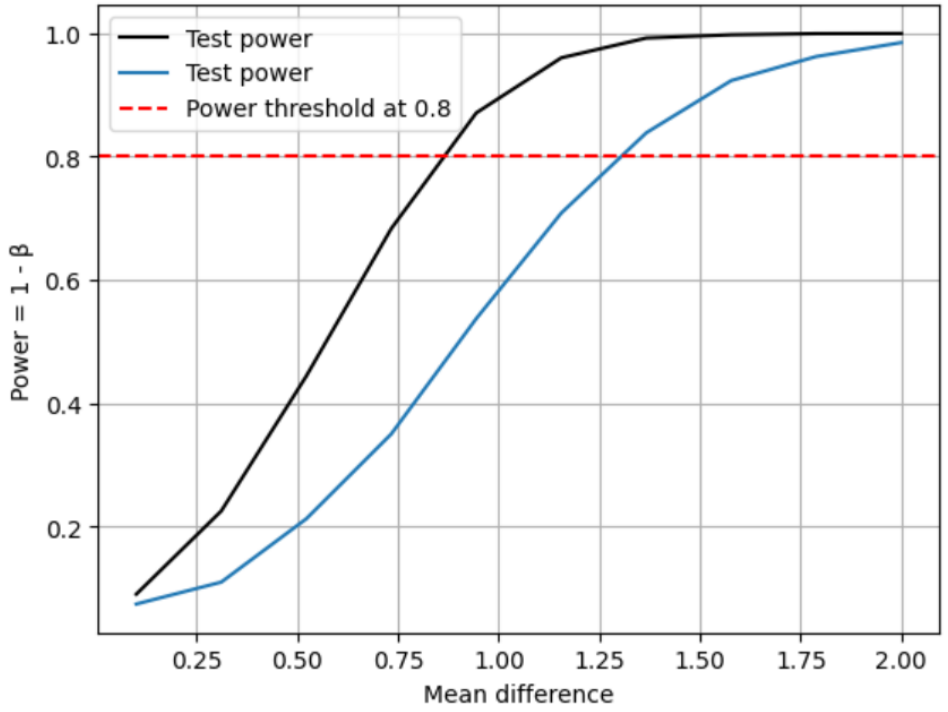
Query → Lomas → DP Protected Results

Multiple sclerosis in Switzerland

Deep dive: DP budget selection II

Utility analysis for privacy-loss budget $\sigma = 20$

Statistical power between years (not noise addition mechanism)



Improving the service

Lessons learned from the pilot

Dummy dataset generation and metadata standard

- Dummy datasets lacks consistencies in some cases, e.g., unique gender per individual.

  Solution: improve metadata to include dependencies between columns.

Data preparation required for longitunidal studies

- Longitudinal studies were possible thanks to the manual preparation of some columns.

  Solution: prepare database to ensure compatibility with lastest version of DP librairies using an automated pipeline.

Outcomes and Roadmap

Pilot's Outcomes

- Researchers were successful to generate sufficiently reliable DP-protected results for their research project.
- Lomas was still missing some important feature to ensure that the service could be provided autonomously (without manual data preparation.

Current work

- Improve metadata standard and pipeline functionalities.
- Productive deployment for hospitalization data.

Competence Network for Artificial Intelligence
Kompetenznetzwerk für künstliche Intelligenz
Réseau de compétences en intelligence artificielle
Rete di competenze per l'intelligenza artificiale

Data Science
Competence Center DSCC

Schweizerische Eidgenossenschaft
Confédération suisse
Confederazione Svizzera
Confederazium svizra

Swiss Confederation

Bundesamt für Statistik BFS
Office fédéral de la statistique OFS
Ufficio federale di statistica UST
Federal Statistical Office FSO

Questions?

www.dscc.ch
www.dscc.ch/fr
www.dscc.ch/it
www.dscc.ch/en

raphael.defondeville@bfs.admin.ch

@StatSchweiz / @statsvizzera
@statsuisse / @swissstatistics

@SwissStats

@SwissStatistics

@bfs-ofs