AIML4OSS – Synthetic data (WP13)

Alexander Kowarik

Massimo De Cubbelis
ISTAT

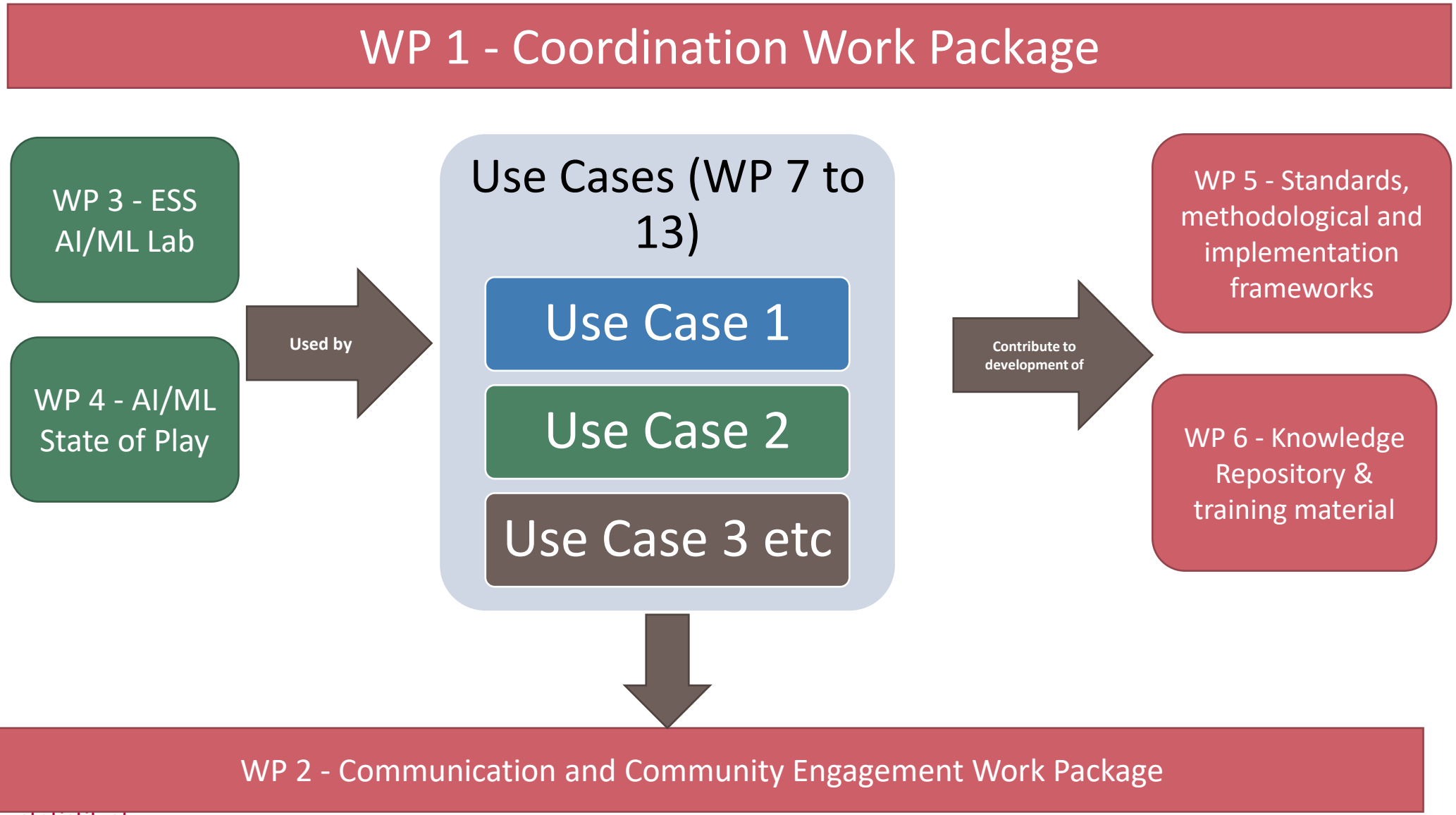
Online 27 February 2025

www.statistik.at

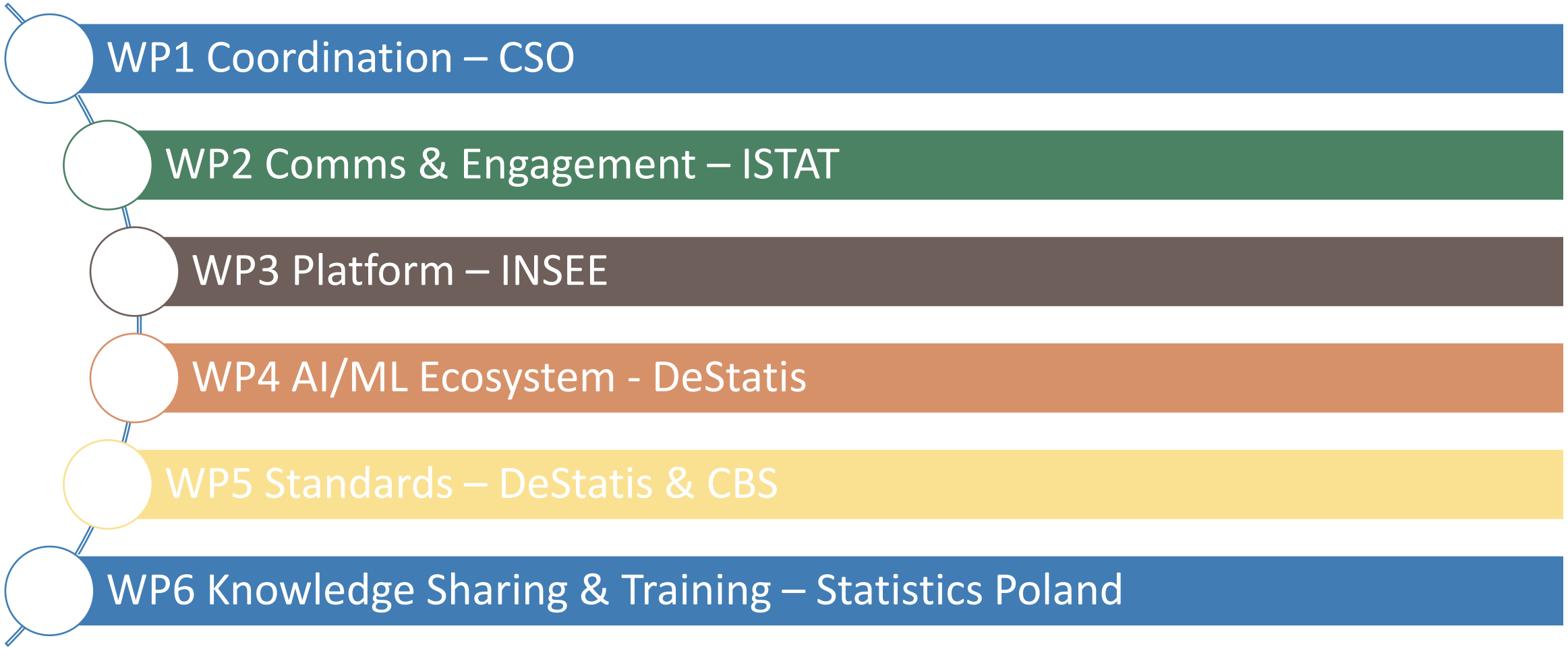
Istat
Istituto Nazionale di Statistica

STATISTICS AUSTRIA
The Information Manager

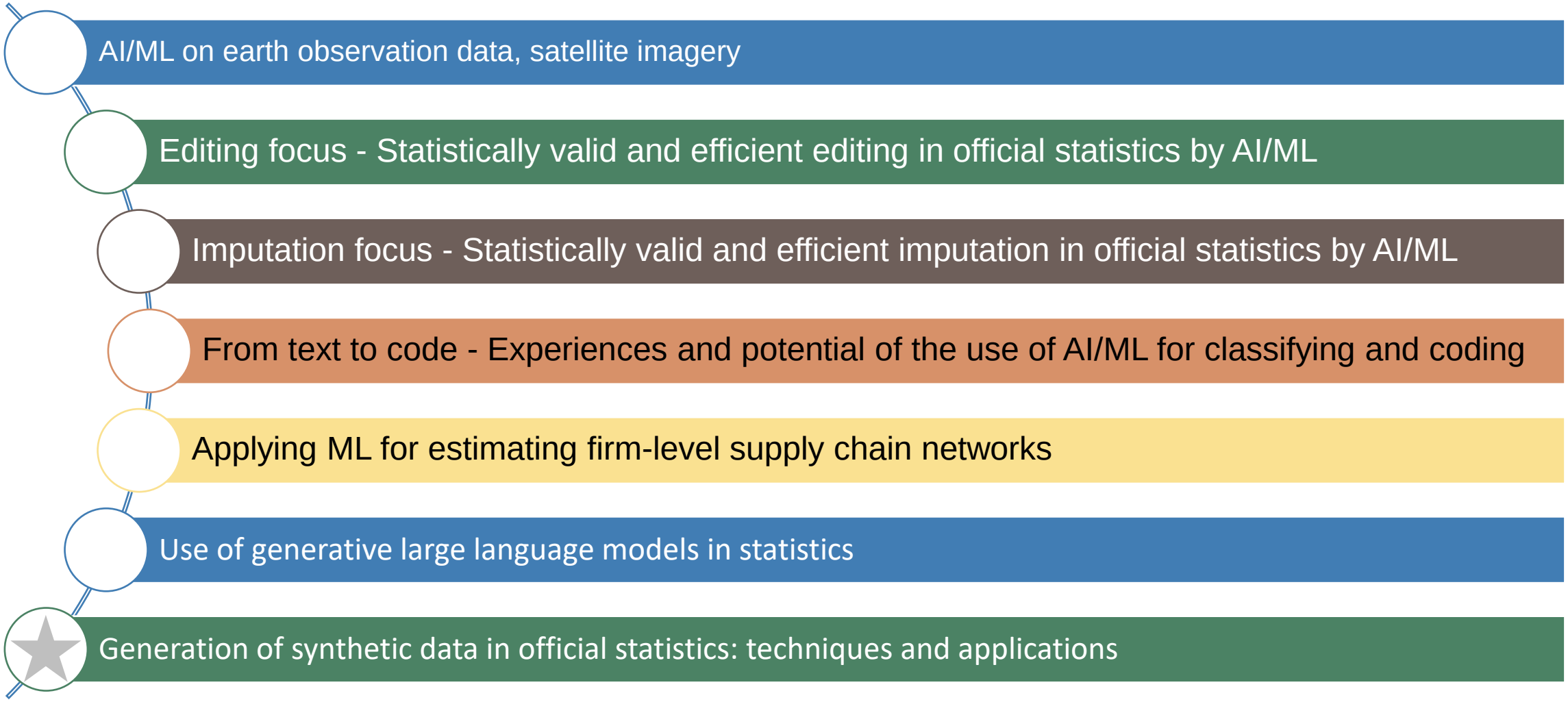
Project


WP 1 - Coordination Work Package
WP 3 - ESS AI/ML Lab
WP 4 - AI/ML State of Play
Used by
Use Cases (WP 7 to 13)
Use Case 1
Use Case 2
Use Case 3 etc
Contribute to development of
WP 5 - Standards, methodological and implementation frameworks
WP 6 - Knowledge Repository & training material
WP 2 - Communication and Community Engagement Work Package

Overarching Work Packages



Use Cases


AI/ML on earth observation data, satellite imagery
Editing focus - Statistically valid and efficient editing in official statistics by AI/ML
Imputation focus - Statistically valid and efficient imputation in official statistics by AI/ML
From text to code - Experiences and potential of the use of AI/ML for classifying and coding
Applying ML for estimating firm-level supply chain networks
Use of generative large language models in statistics
Generation of synthetic data in official statistics: techniques and applications

Who's Involved? In WP13

An Phríomh-Oifig Staidrimh Central Statistics Office

Istat Istituto Nazionale di Statistica

INSTITUTO NACIONAL DE ESTATÍSTICA STATISTICS PORTUGAL

DESTATIS wissen.nutzen.

INSEE

Statistics Poland

INE Instituto Nacional de Estadística

STATISTICS AUSTRIA The Information Manager

Work

- Investigate different AI/ML algorithms to generate synthetic data in official statistics domains
- Balancing utility and privacy
- Quality assessment: comparisons of the statistical properties, distributions, and performance metrics

- We will use the Onyxia environment (in SSPcloud provided by INSEE)

- Currently, we are collecting methods for privacy and utility assessment and to generate synthetic data in different proof-of-concepts

Purpose of Synthetic Data – first draft structure

- **Structural data set** = Software Development/Testing data set
  - Preparatory file before access to secured use file
- **Public use file**, e.g. for educational pruposes – main statistical features
- **Scientifc use file** - to gain insights in a specific reseach area
  - Special purpose scientific use file -
- **A perfect "twin"** of the reality.
  - All possible statistical analysis can be covered.
  - ML model training

Utility metrics categories – proposal by Destatis (based on Drechsler et al. 2024)

- Fit-for-purpose utility
  - First impression of the quality of the synthetic output
    - E.g. Plausibility checks, graphical evaluation of the distributions
- Global utility
  - Compare original data with protected data (distribution similarity)
    - E.g. Propensity score (can a model distinguish between original and synthetic)
- Outcome-specific utility
  - Measure utility for specific analyses
    - E.g. Comparison of GLM coefficient for a specific model

„Synthetic does not mean the data is safe“ Privacy measure - discussion

- Can privacy be measured with „only“ the original and the synthetic data as input?
- If not what kind of information is needed on the model/methodology?

- **Attack model based methods**, e.g.,
  - Membership Inference Attacks
  - Attribute Inference Attacks
- **Attribution and Disclosure Risk Metrics**, e.g.,
  - Equivalence Class Attribution Probability
- **Privacy Risk Metrics for Fully Synthetic Data**
  - Replicated uniques
- **Privacy Risk Metrics for Partially Synthetic Data**
  - Expected Match Risk

Methods to generate synthetic

- Statistical and Rule-Based Methods
- Machine Learning and Deep Learning based Methods
- Privacy-Preserving Frameworks
  - E.g. DP-safe ML or synthpop

- Deep Learning: Transformer, Generative Adversarial Networks (GANs), Variational Autoencoders (VAEs)

Next steps

- Utility and privacy measure will be assessed if they are easily availabe in R and Python, if not, selected ones will implemented.
- Define the use case each partner is running: statistical domain, methods for generation, risk/utility measures