

PP24

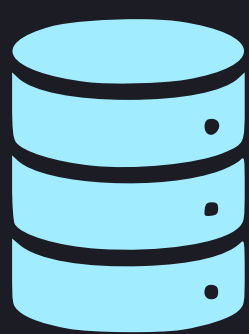
DATA ANONYMISATION & RISK ASSESSMENT

PROCESS MAP AND AUTOMATION EFFORTS

STEP 1: GATHER & PRE-PROCESS

INPUTS:
SOURCE DATA

OUTPUTS:
CLEAN DATASETS
DATASET SPECS
[VAR NAME, ATTRIBUTES]



- UNIFY VARIABLE ATTRIBUTES ACROSS DATASETS
- REMOVE SAS FORMATS, CONVERT FORMATTED VARIABLES TO CHARACTERS
- CONVERT ALL DATE VARIABLES TO CHARACTER ISO8601 FORMAT
- REPLACE NON-STANDARD VALUES WITH CONTROLLED TERMINOLOGY WHERE POSSIBLE

STEP 2: FIND IDENTIFIERS



INPUTS:
CLEAN DATASETS (1)
DATASET SPECS (1)
PhUSE DE-ID STANDARD

OUTPUTS:
UPDATED DATASET SPECS
[VAR NAME, ATTRS, +CLASSIFICATION]

- MERGE SPECS WITH PhUSE DE-ID STANDARD TO FIND DIRECT IDENTIFIERS, QUASI-IDENTIFIERS, NON-IDENTIFIERS
- DEMOG VARS - AGE, SEX, COUNTRY, REGION, RACE, ETHNICITY, ETC.
- BODY MEASUREMENT VARS: HEIGHT, WEIGHT, BMI, ETC.
- AE, CONMED, MEDHIST TERMS
- DATE AND TIME VARS
- BIRTH / DEATH DATES
- FREE ENTRY VARS, COMMENTS.
- OTHER INFO WHICH MAY HELP INFER INFORMATION

PREPARATION

STEP 3: BUILD SPECS



INPUTS:
DATASET SPECS (2)
PhUSE DE-ID STANDARD

OUTPUTS:
UPDATED DATASET SPECS
[VAR NAME, ATTRS, CLASS, +ANONYMISATION RULES]

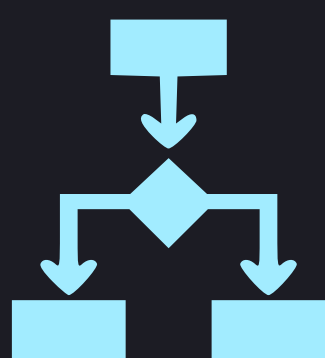
- RECODE ID
- KEEP
- DROP
- CLEAR
- OFFSET
- NEW
- DERIVE

- AGE-BANDS
- AGE-IN-BAND
- AGE89
- COUNTRY-POOL
- LOW-FREQ-POOL
- REDACT
- OTHER

STEP 4: APPLY RULES

INPUTS:
CLEAN DATASETS (1)
DATASET SPECS (3)

OUTPUTS:
ANONYMISED DATASETS



- APPLY INDIVIDUAL RULES TO EACH VARIABLE ITERATING OVER ALL COLUMNS IN SPEC
- OUTPUT ANONYMISED DATASETS, ENSURE ALL VALUES WHICH REQUIRE TRANSFORMATION HAVE BEEN REPLACED OR REMOVED

STEP 5: CALCULATE RISK

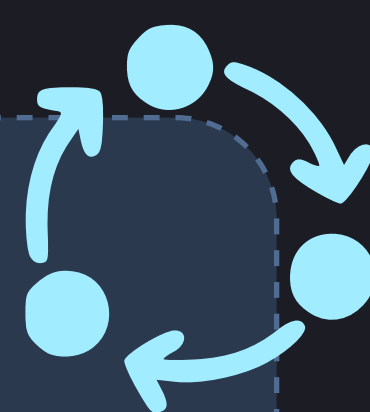
INPUTS:
ANONYMISED DATASETS (4)
REFERENCE DATASET (OPTIONAL)

OUTPUTS:
RISK CALCULATION REPORT

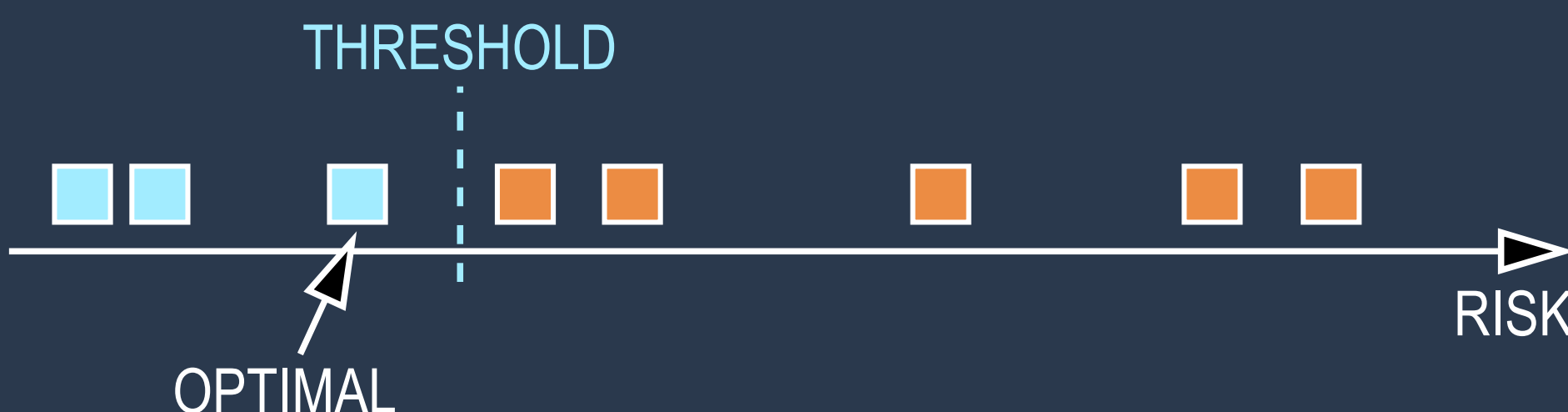


- 0: SELECT QUASI-IDENTIFIERS FOR RISK CALCULATION.
- 1: ASSESS RISK OF ATTACK.
 - RISK OF DELIBERATE ATTACK
 - PROBABILITY OF INADVERTENT RE-IDENTIFICATION
 - RISK OF DATA BREACH
- 2: CALCULATE RECORD-LEVEL INDIVIDUAL VALUES OF RISK OF SUCCESSFUL RE-IDENTIFICATION IN THE EVENT OF AN ATTACK
- 3: FIND AVERAGE AND MAXIMUM RISK (PROSECUTOR OR JOURNALIST)
- 4: CALCULATE OVERALL RISK OF SUCCESSFUL RE-IDENTIFICATION USING 1 AND 3
- 5: ASSESS K-ANONYMITY AND UNIQUENESS IN DATASET

REPEAT STEPS 3-5



AGE X **COUNTRY** X **SEX** = 24 SCENARIOS
KEEP AGE-BANDS(5) AGE-BANDS(10) DROP X KEEP COUNTRY-POOL DROP X KEEP DROP



FINDING OPTIMAL LEVEL OF ANONYMISATION

STEP 6: RUN & BUILD PACKAGE

INPUTS:
CLEAN DATASETS (1)
FINAL DATASET SPECS (3)
[VAR NAME, ATTRS, CLASS, ANONYMISATION RULES]

OUTPUTS:
FINAL ANONYMISED DATASETS IN DESIRED FORMAT
FINAL DATASET SPECS
ANONYMISATION REPORT
ZIPPED DELIVERABLE



- RERUN STEPS 3, 4, 5 USING OPTIMAL RULES
- SAVE FINAL SPECIFICATIONS AND RISK CALCULATION RESULTS AS DOCUMENTATION
- CONVERT DATASETS TO DESIRED FORMAT (SAS, XPT, CSV, XML, JSON)
- COMPRESS PACKAGE INTO SINGLE PASSWORD PROTECTED ZIP FILE FOLLOW NAMING CONVENTION (E.G. RQ001-2019-11-10.ZIP)

STEP 7: REPORT & DOCUMENT



ANONYMISATION REPORT

- LIST OF IDENTIFIERS (DIRECT AND QUASI)
- CLASSIFICATION OF ALL VARIABLES AND RULES APPLIED
- RISK CALCULATION METHODOLOGY AND REFERENCE POPULATION
- VALUES OF RISK OF ATTEMPT (DELIBERATE, INADVERTENT, BREACH, OVERALL)
- RISK CALCULATION RESULTS (ALL OR SELECTED SCENARIOS)
- IMPACT ON DATA UTILITY

PRODUCTION