Westat®

Improving Lives Through Research®

Privacy Preserving Data Processing

IPAC Conference
July 25, 2023

Tom Krenzke

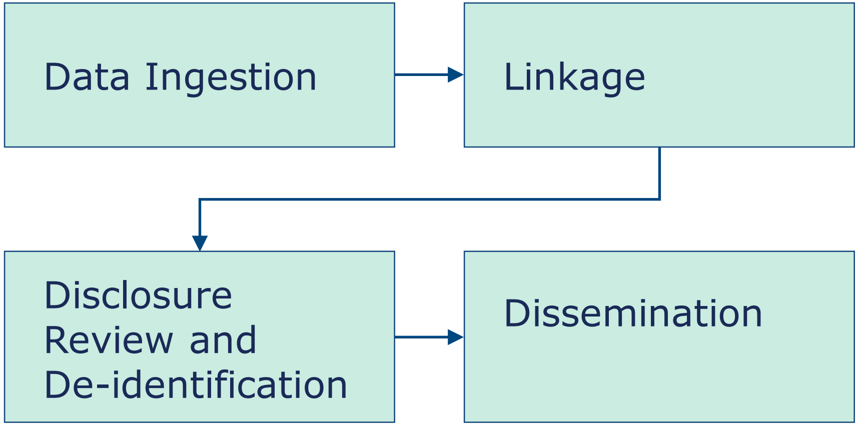
Westat

The views presented are those of the author(s) and do not represent the views of any Government Agency/Department or Westat.

Disclosure Avoidance Guidance Objectives

- Protect individuals from re-identification and maintain data confidentiality
- Protect the integrity of the data
- Ensure compliance
  - Consent
  - Authorization for EHR data
  - State and federal regulations
  - Your own program’s protocols

A Data Sharing Scenario



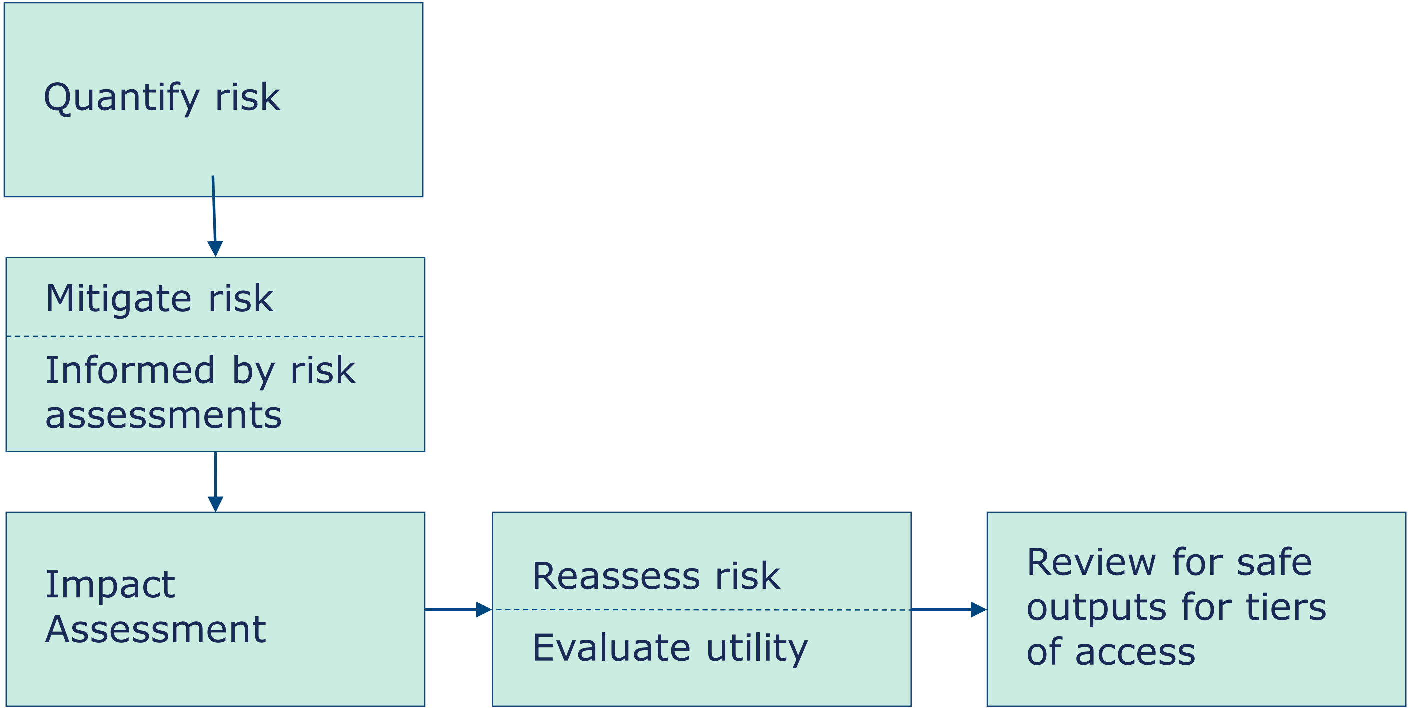
Data Ingestion

Linkage

Disclosure Review and De-identification

Dissemination

General Disclosure Avoidance Process



Quantify risk

Mitigate risk

Informed by risk assessments

Impact Assessment

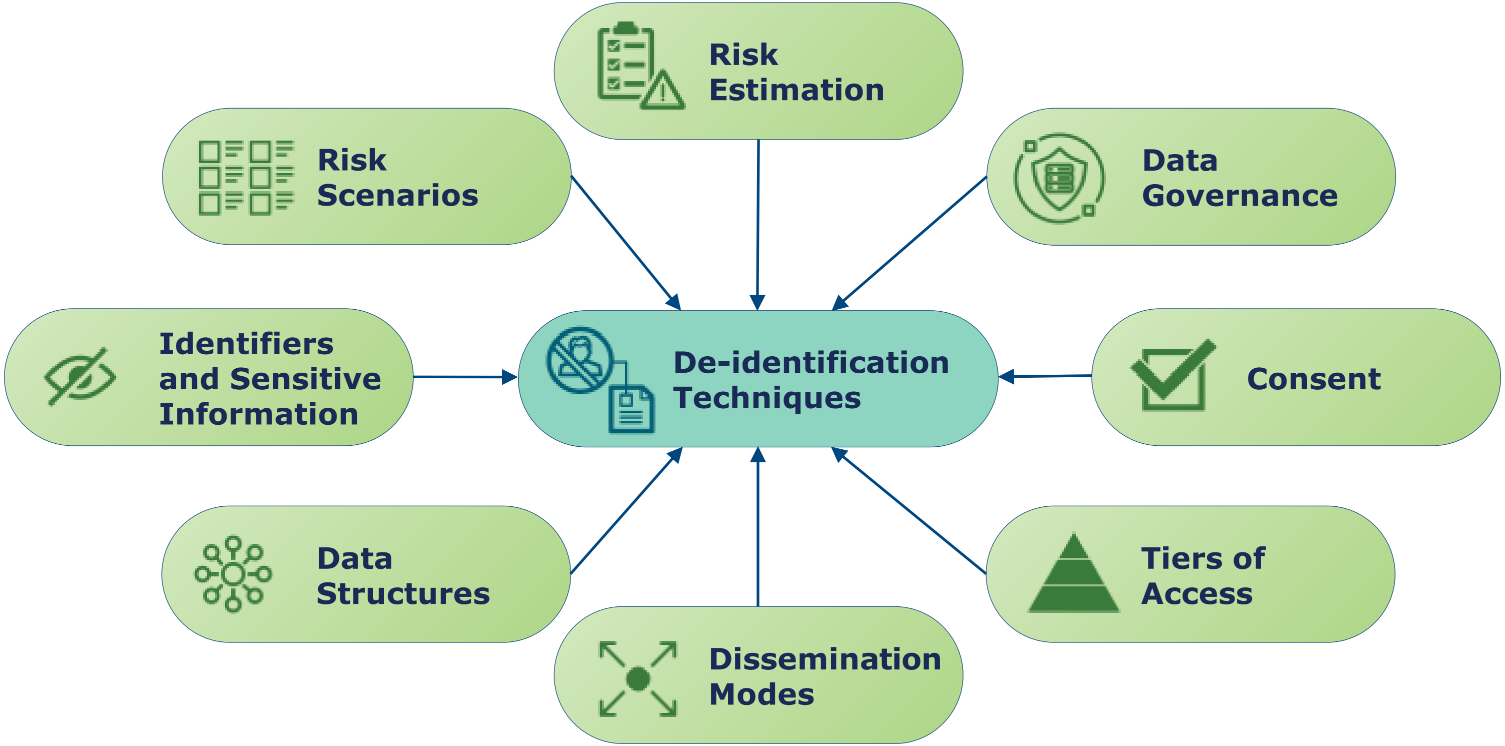
Reassess risk

Evaluate utility

Review for safe outputs for tiers of access

De-identification techniques are different for each tier of access

Informed Data Treatments



Risk Estimation

Risk Scenarios

Data Governance

Identifiers and Sensitive Information

De-identification Techniques

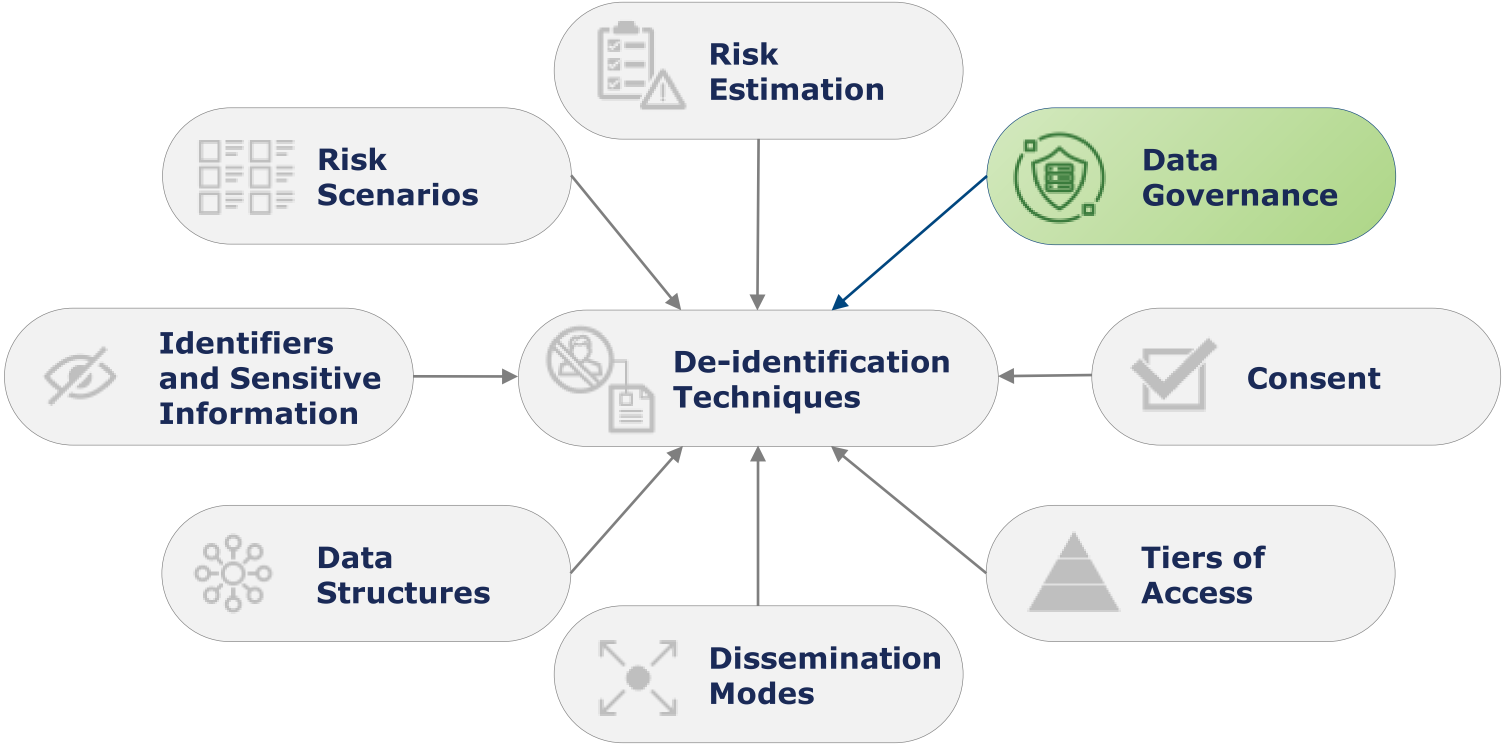
Consent

Data Structures

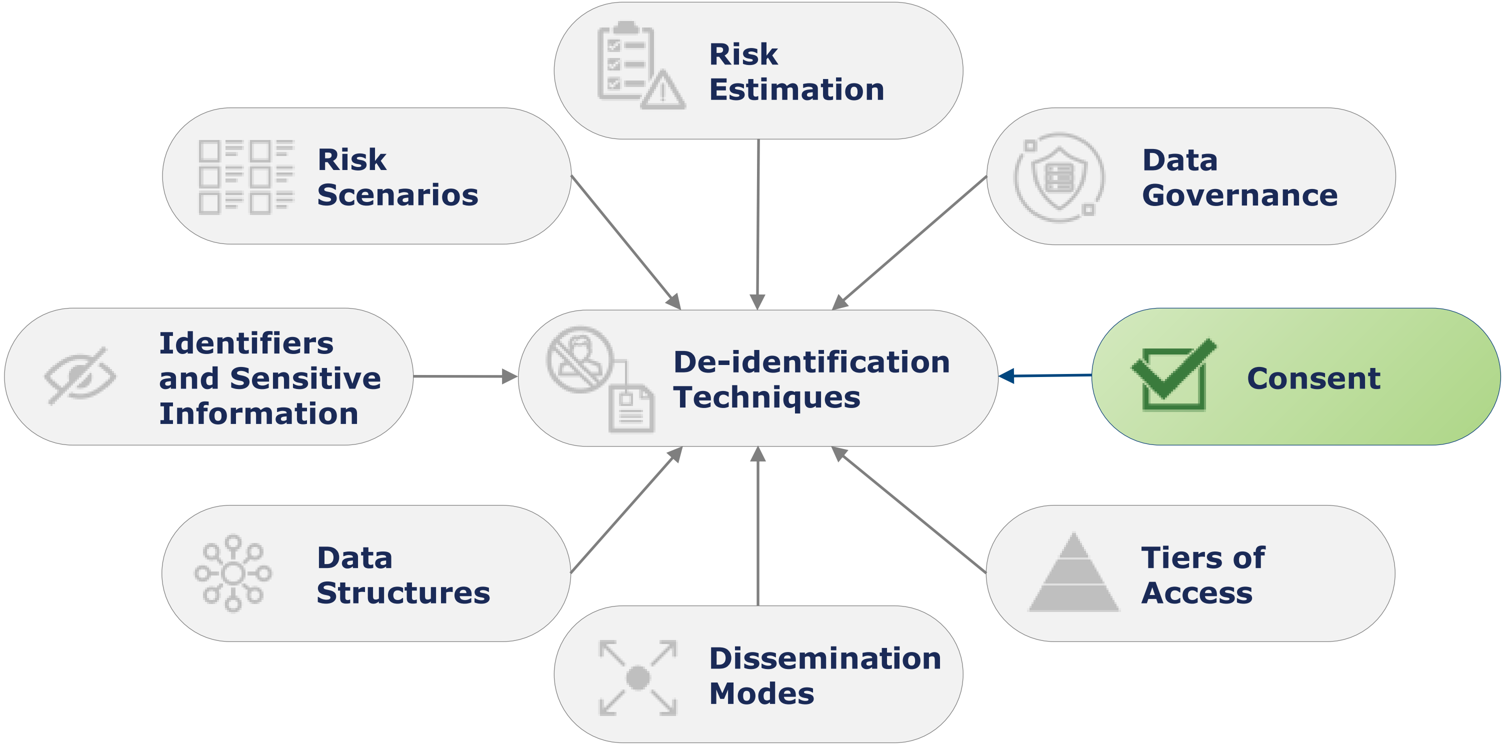
Tiers of Access

Dissemination Modes

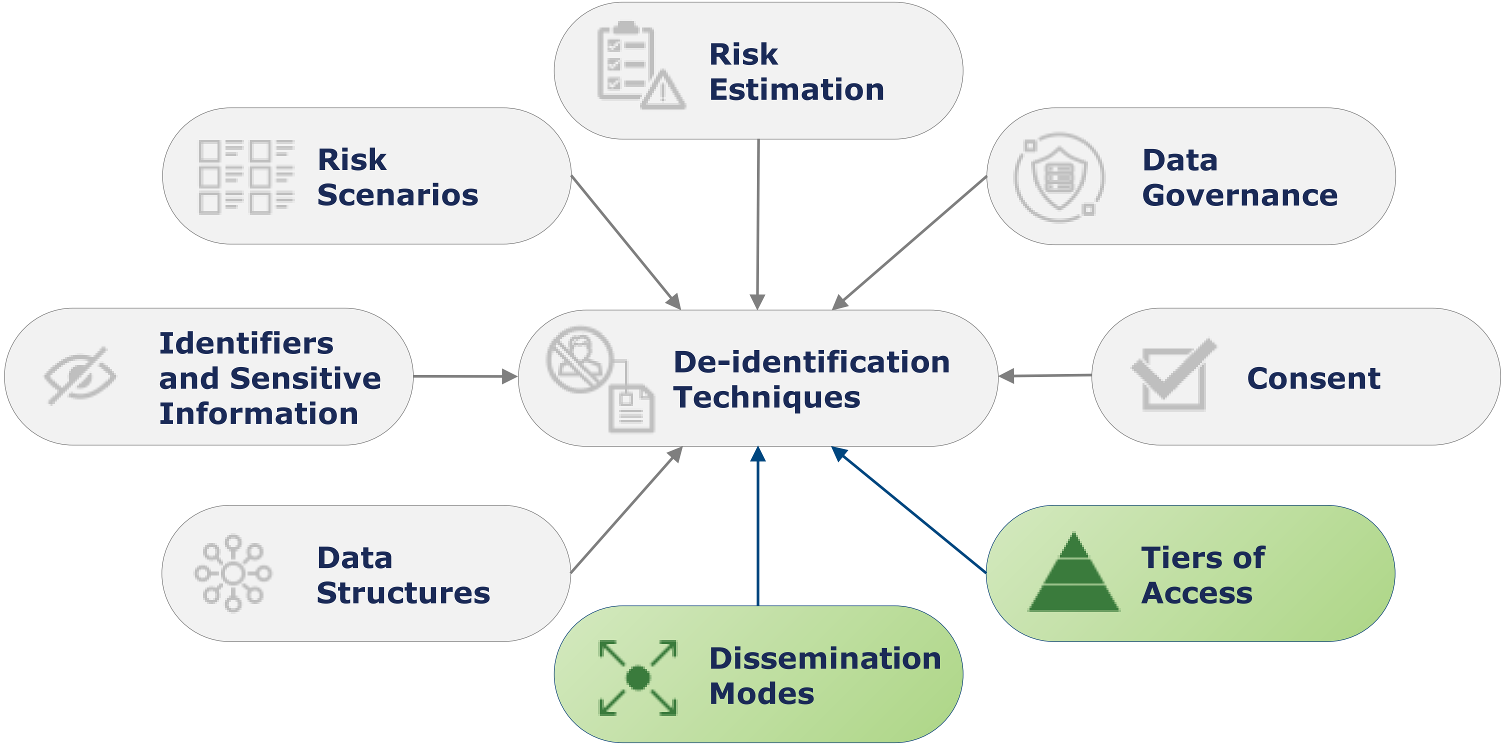
Data Governance



Consent


Risk Estimation
Risk Scenarios
Data Governance
Identifiers and Sensitive Information
De-identification Techniques
Consent
Data Structures
Dissemination Modes
Tiers of Access

Data Access and Dissemination Modes



Risk Estimation

Risk Scenarios

Data Governance

Identifiers and Sensitive Information

De-identification Techniques

Consent

Data Structures

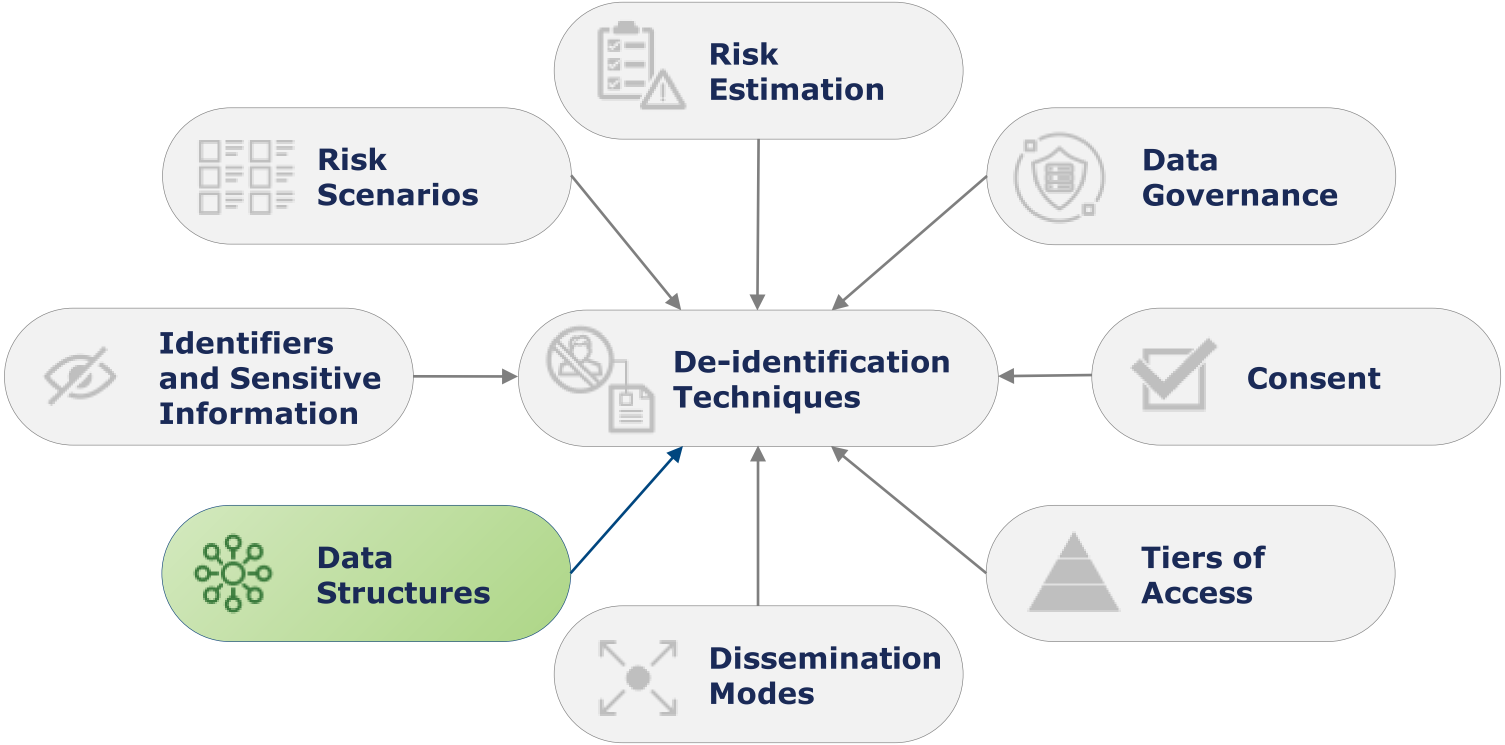
Tiers of Access

Dissemination Modes

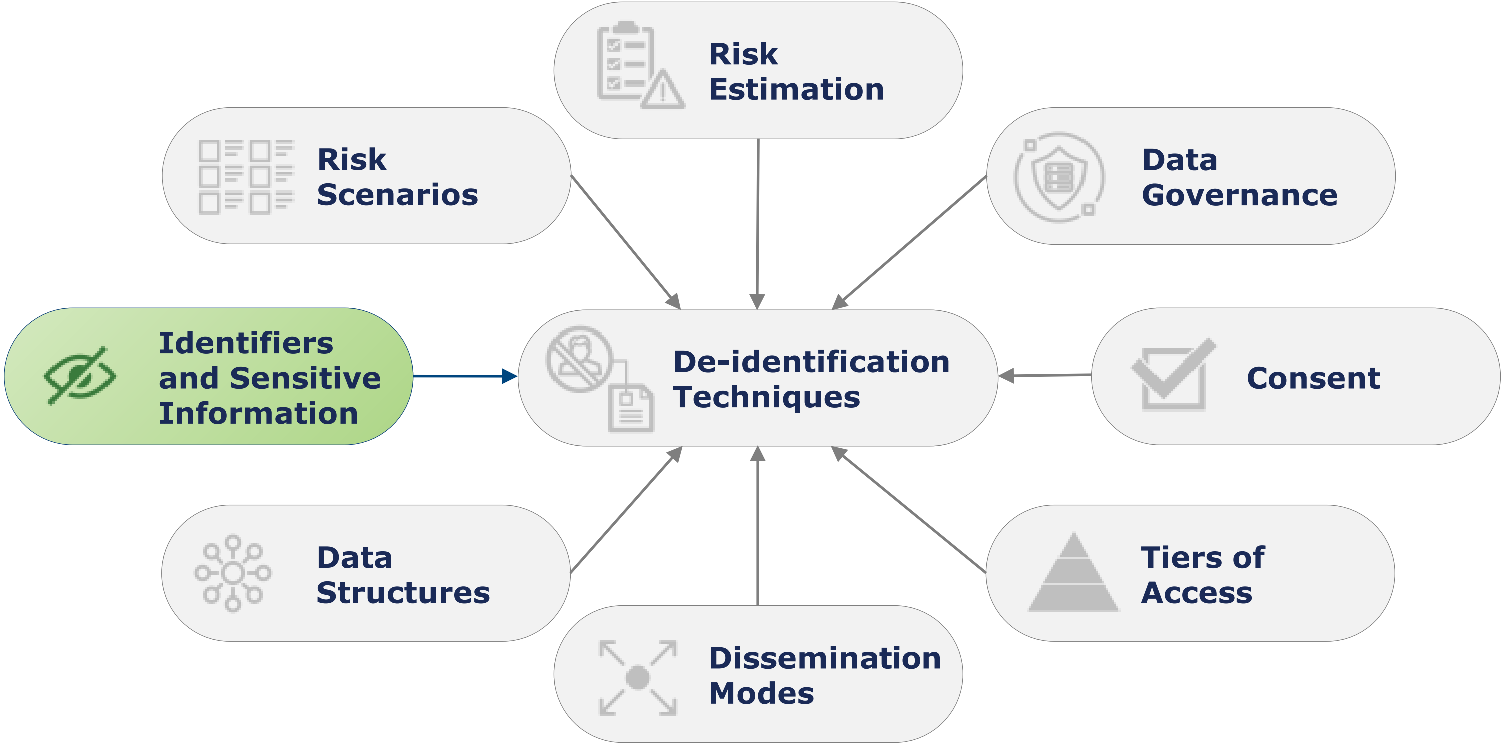
Modes and Access Levels

- Microdata
  - Public use files (PUFs) and restricted use files (RUFs)
- Tables
  - Static tables
  - Flexible table generator – UI tool
- Access to RUF
  - Licenses
  - FSRDC
  - Virtual access
- Consistency in results between data products

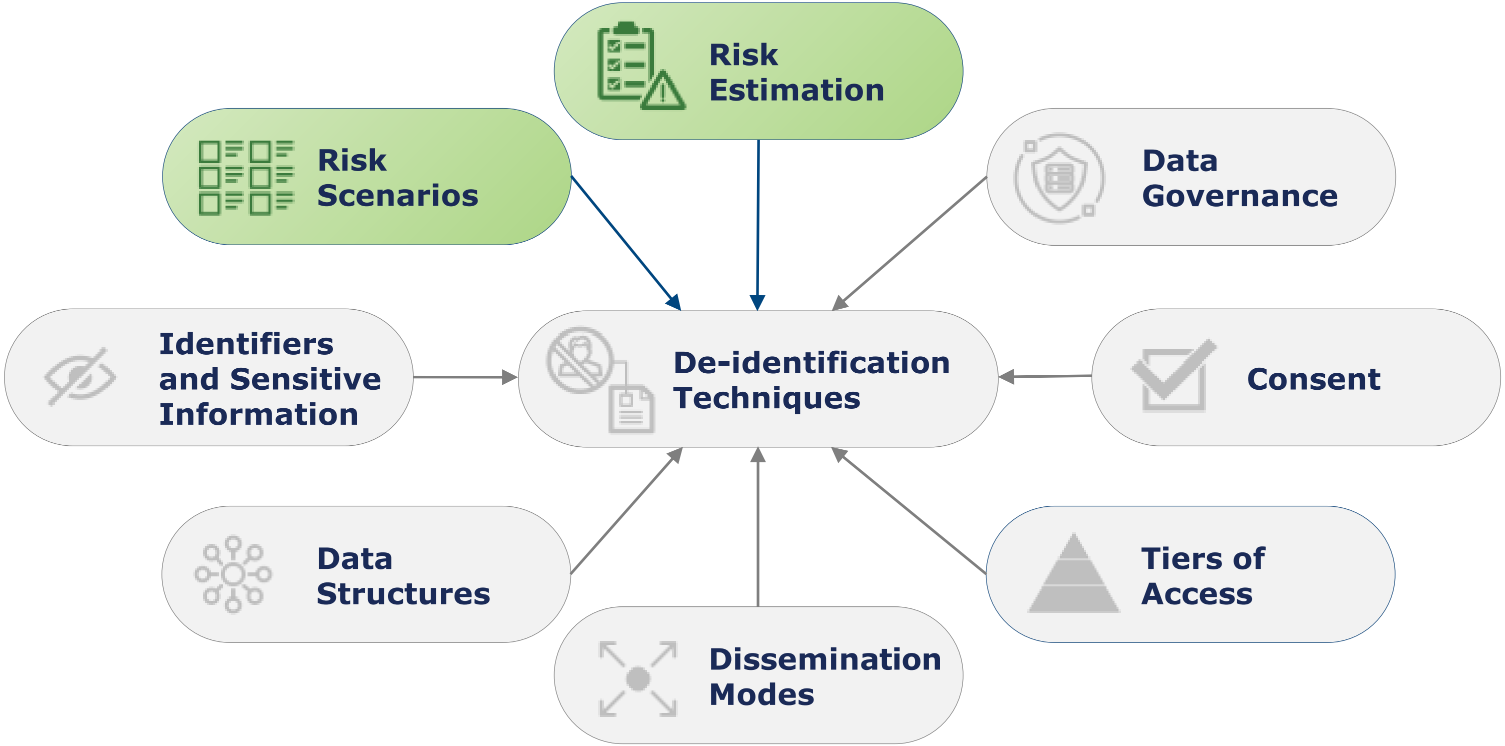
Data Structures



Identifiers and Sensitive Information



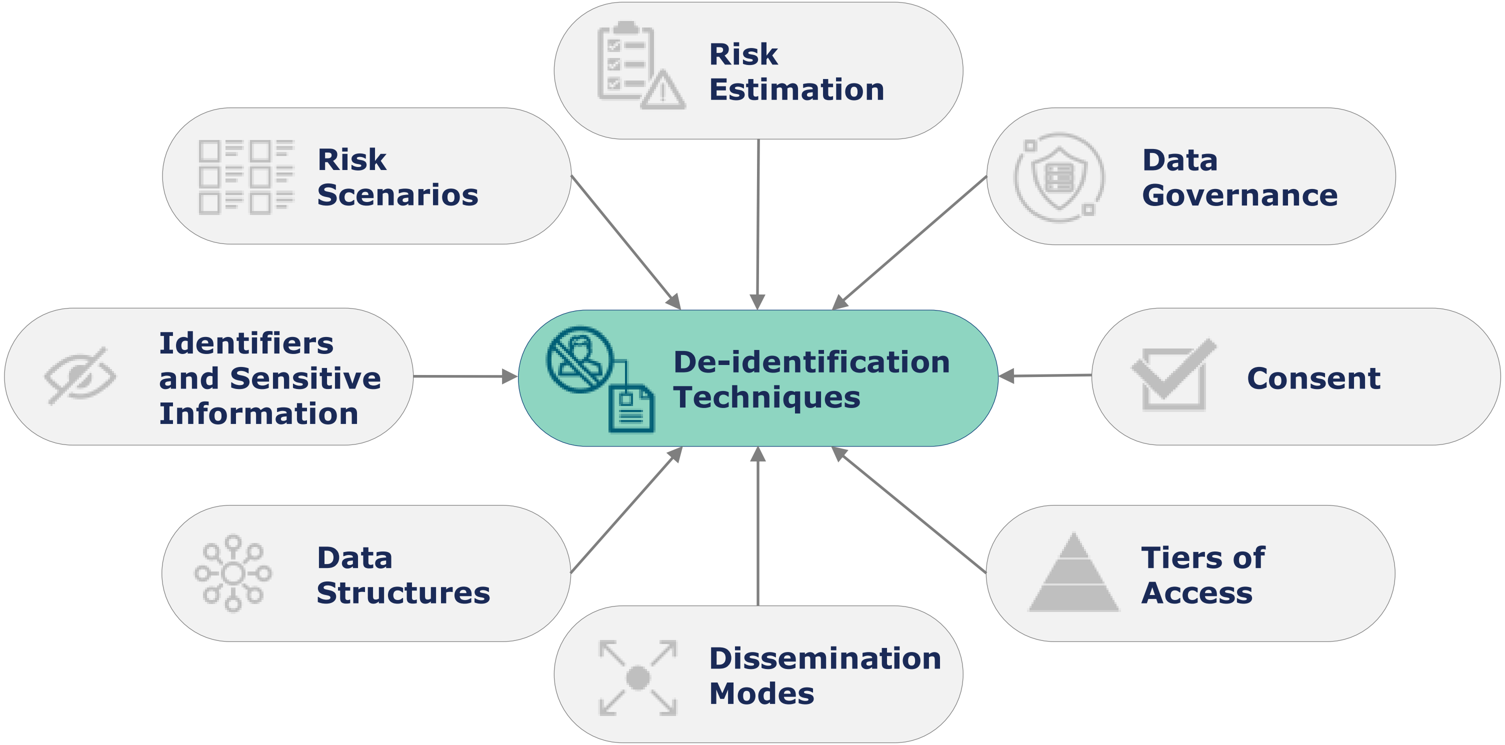
Risk Scenarios and Quantification



Risk Assessment Approaches

- Probabilistic record linkage
  - Match your file with publicly available data
  - Use variables in common to each file
- Model-assisted (Skinner and Shlomo, 2008)
- Exhaustive tabulations
  - Conduct n-way crosstabs & count cells with low frequency to identify
    - Categories of variables to recode
    - Variables to suppress
    - Which records to give higher chance to perturb
  - *SDCNway* (R package on the CRAN network)

Data Treatments



De-identification Techniques

- Data coarsening (suppression, recodes, rounding)
  - All data coarsening procedures result in a loss of information
  - If risk remains, leads to either data access restrictions or perturbation
- Controlled random treatments
  - Noise infusion
  - Perturbation (swapping, model-assisted constrained hotdeck)
  - Synthetic data
- Tables (cell suppression, others), table generators
- Others

Example of Mix of Treatments and Tiers of Access

- Public tier – Aggregates and summaries
- Intermediate tier -- Microdata
  - Treatments – e.g., fewer variables released, data coarsening, perturbation or synthetic data for sensitive data
  - Table generator, verification server
- Restricted tier -- Microdata
  - Less treatments than intermediate tier
  - More variables – more demographics, EHR and survey data

Reference

- Skinner, C.J. and Shlomo, N. (2008). Assessing Identification Risk in Survey Microdata Using Log-linear Models. Journal of American Statistical Association, 103, 989–1001.

Thank You