phuse·share
connect·share·advance

Single Day Event

How to protect individuals privacy in the future age of artificial intelligence (AI)

Yuanlin MA,
Sanofi

Disclaimer

The views and opinions expressed in this presentation are those of the author(s) and do not necessarily reflect the official policy or position of PHUSE.

Agenda

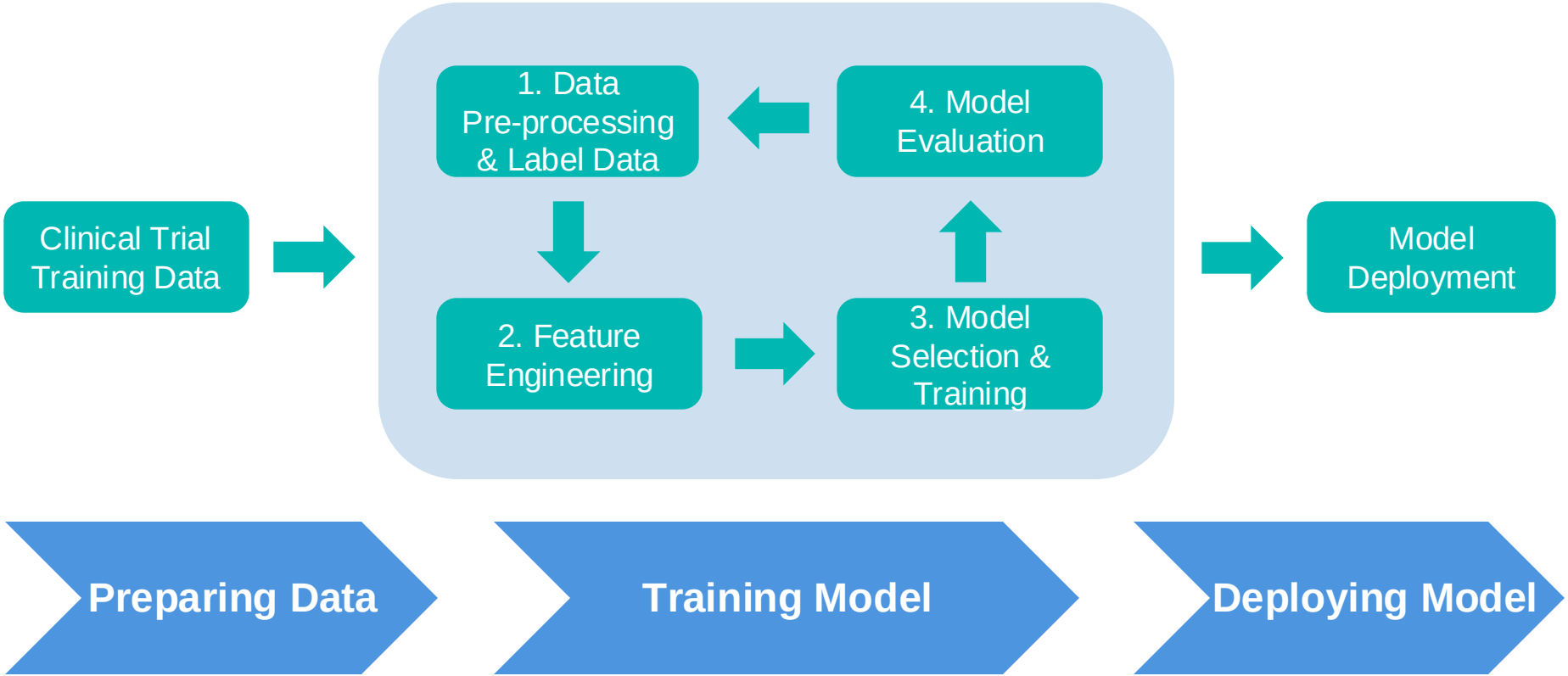
01 Why need *data anonymization*?

02 What is *data anonymization*?

03 How to process *data anonymization*?

04 Where is *data anonymization* going?

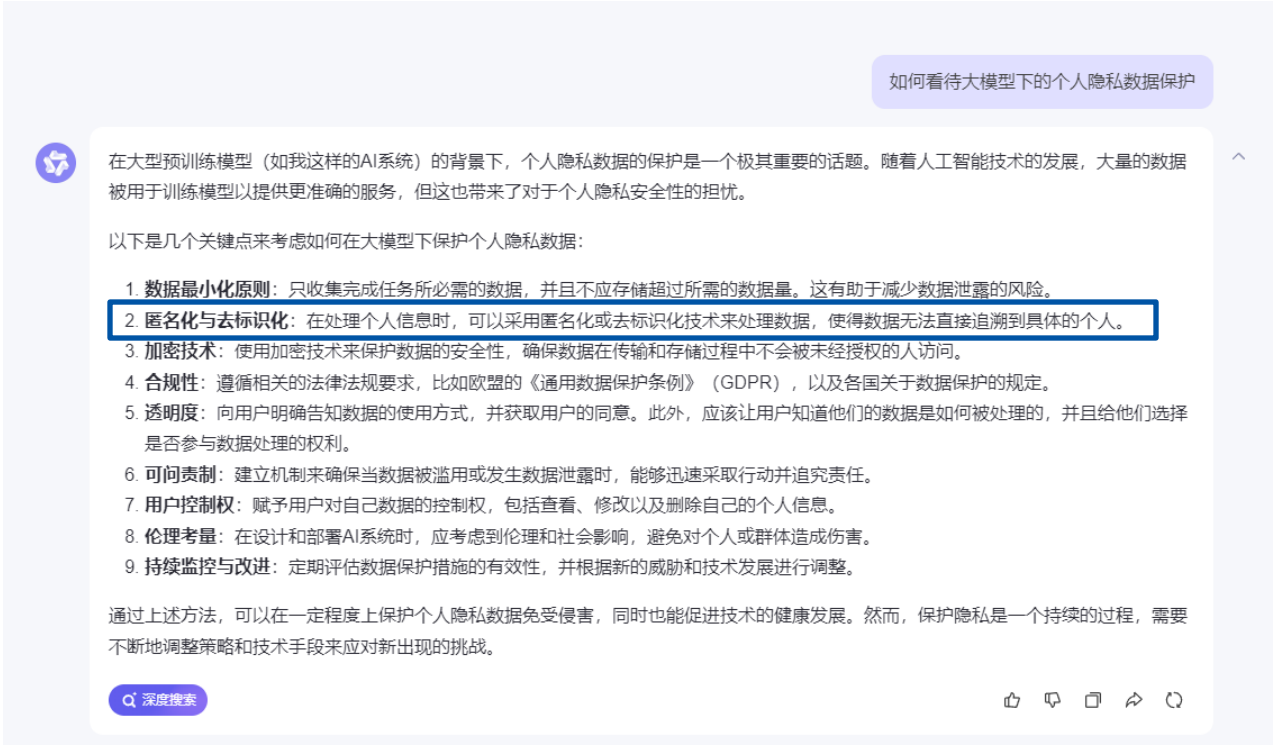
Why? – AI Process



Why? – AI's Opinion

To tongyi.ai:

- How to think about the protection of personal privacy data under the Big Model?



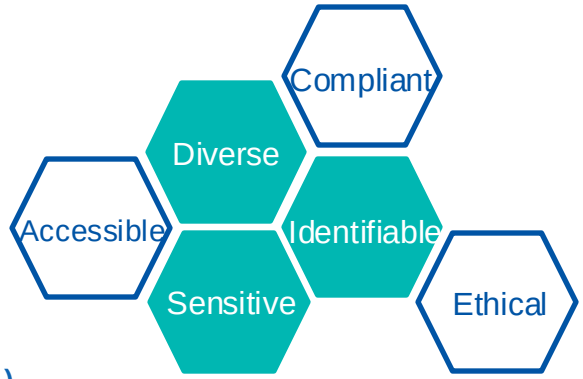
Why? – Clinical Trial Data

- Primary Use

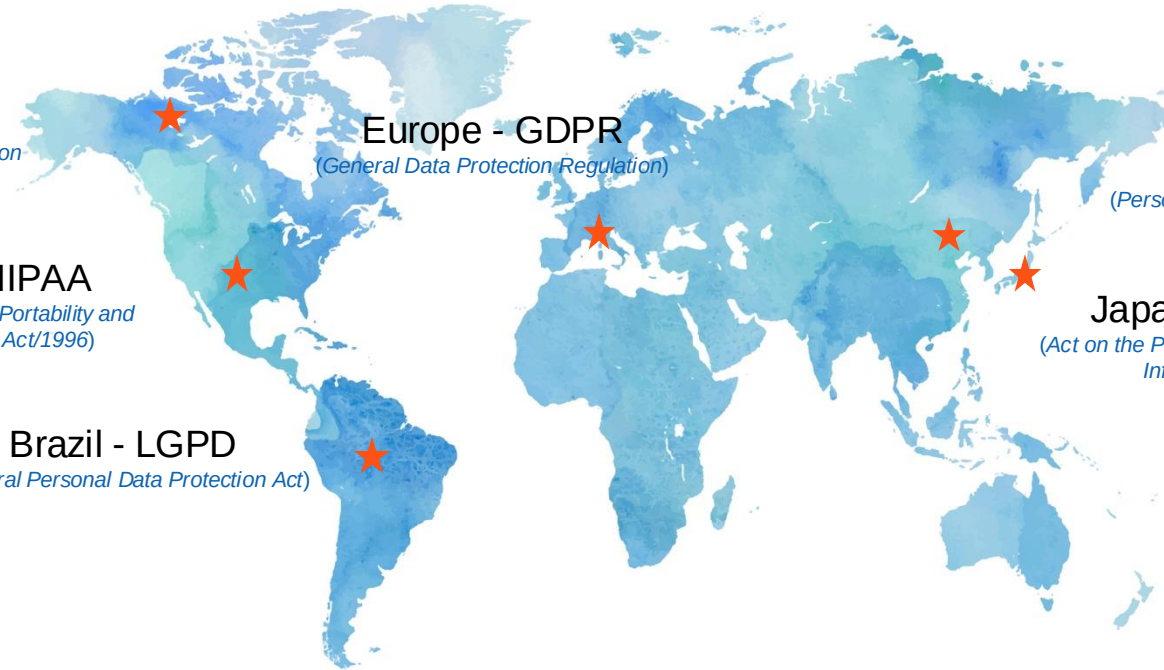
  Uses regardless of the time point with the intention to support the same project as agreed by project team needed.

- Secondary Use

  Uses and disclosures that are different from the purpose(s) for which the data were collected as described in a clinical trial protocol and informed consent form.



Why? - Regulation



Canada - PIPEDA
(*The Personal Information Protection and Electronic Documents Act*)

Europe - GDPR
(*General Data Protection Regulation*)

China - PIPL
(*Personal Information Protection Law*)

USA - HIPAA
(*Health Insurance Portability and Accountability Act/1996*)

Japan - APPI
(*Act on the Protection of Personal Information*)

Brazil - LGPD
(*General Personal Data Protection Act*)

What? - Concept

- Anonymization

  A process of turning data into a form that does not identify individuals and where identification is not likely to take place.

- Pseudonymization

  A process of both removing the association with a data subject and adding an association between a particular set of characteristics relating to the data subject and one or more pseudonyms.

What? - Concept

- Direct Identifier

  Data that can be used to uniquely identify an individual without additional information or cross-linking other information that is in the public domain.

- Quasi-identifier

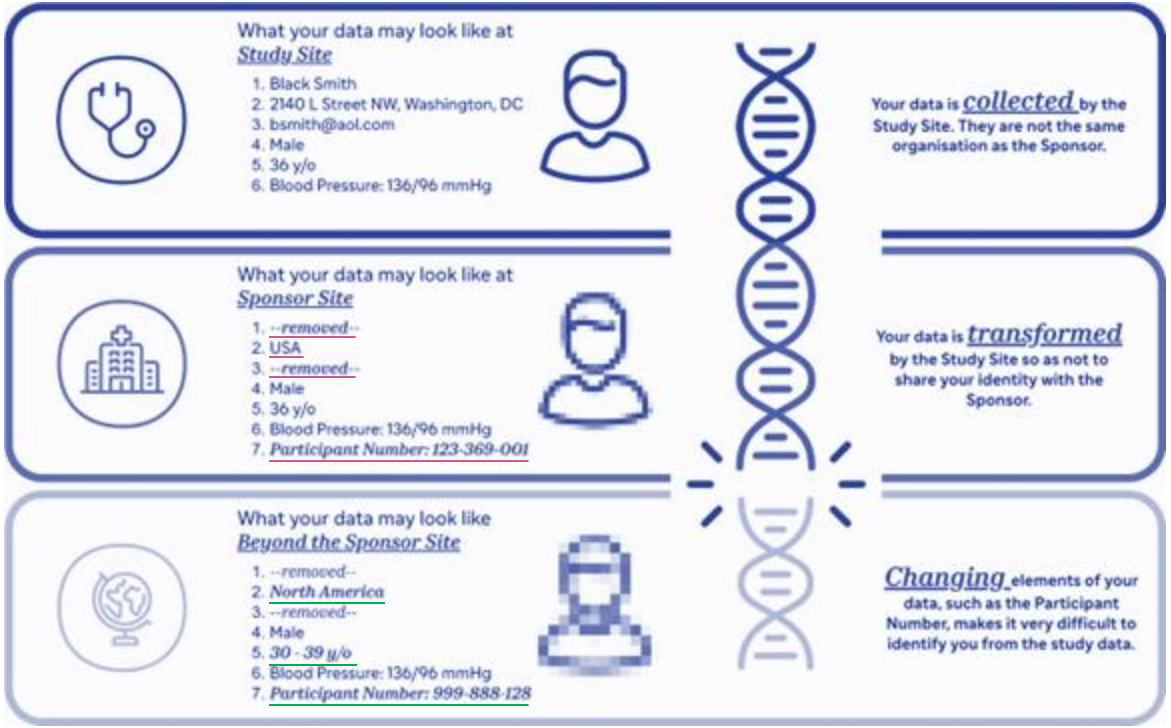
  Data that in connection with other information can be used to identify an individual with high probability.

HIPAA

18 Direct Identifiers

- Names
- Geographic subdivisions smaller than state
- All elements of dates except year
- Telephone numbers
- Fax numbers
- E-mail addresses
- Social Security numbers
- Medical record numbers
- Health plan beneficiary numbers
- Account numbers
- Certificate/license numbers
- Vehicle identifiers/serial numbers
- Device identifiers/serial numbers
- URLs
- IP addresses
- Biometric identifiers
- Full face photographic images
- Any other unique identifying number or code

What? - Process



Directly Identifiable data

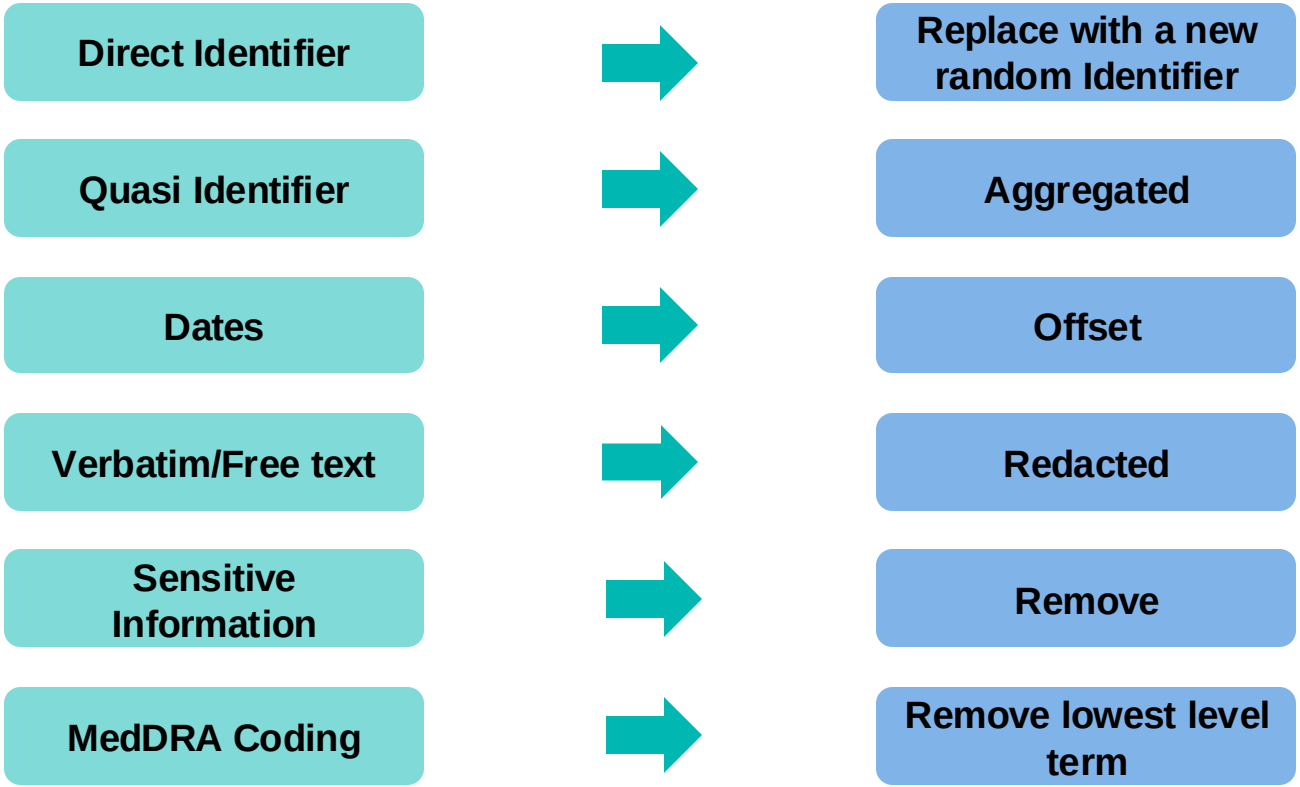
Pseudonymized data

Anonymized data

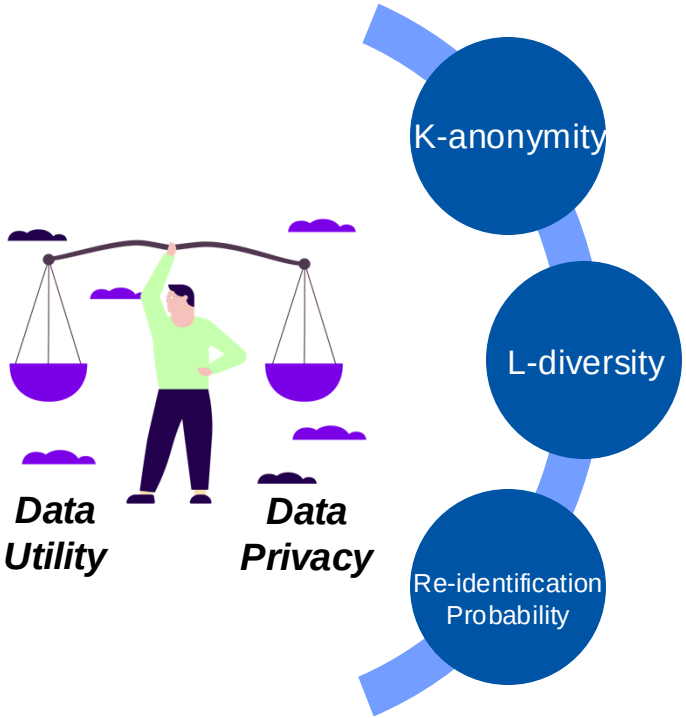
How? - Reference

PHUSE & Transcelerate

How? - Rule-based



How? - Quantitative Risk-based



There are at least *k* records within each equivalence class

Quasi-Identifiers: Age, Weight, Sex, Ethnic origin, Geographical location

Threshold: k=2 (except for up to 5% study population to keep unique)

Based on *k-anonymity* approach, ensure that records within each equivalence class have sufficient diversity

Domains: AE, CM, MH

Threshold: L- diversity < 3

The average risk of re-identification based on study population

$Pr(re-id) = \Pr(re-id|attempt) \times \Pr(attempt)$

Threshold: $Pr(re-id) < 0.09$

How? - Other Consideration

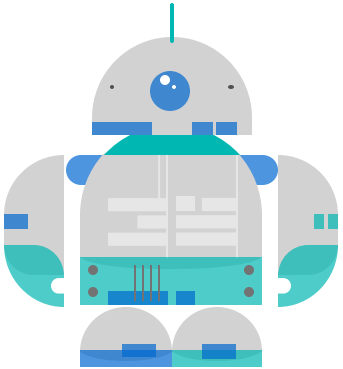
- Questionnaires
- Biomarker Data
- Rare Disease
- Small Sample Size Trial
- Trials With Dependencies
- Copyright License
- Pooling
- ….

Where to go?

AI Detect Potential Sensitive Information

Anonymization Report Leveraging Generative AI

Smart Automation



AI help QC

New Approach / Standard in Questionnaire, Biomarker

- THANK YOU -

Questions

The Global Healthcare Data Science Community

Contact Channels
UK +44 1843 609600
office@phuse.global
phuse.global

Social Media
@phusetwitta
/phusebook
/phusetube
/company/phuse