Perspectives on privacy-preserving data sharing tools for biomedical research

Dr. Jean Louis Raisaro
Head of Clinical Data Science Group
Biomedical Data Science Center

UNIL | Université de Lausanne
Faculty of Biology and Medicine

About me: From medicine, to computer science and back

Early life

Family of medical doctors

2006-2012

AD NYE VEJE AALBORG UNIVERSITET

UNIVERSITÀ DI PAVIA

BS and MS in Biomedical Engineering and Medical Informatics

SNOMED CT The global language of healthcare

2012-2018

PhD in Computer Science

EPFL

HARVARD MEDICAL SCHOOL

MGH 1811

2018-2020

MedCO Collective protection of medical data

CHUV

CHUV Post-doc

AI

2020-now

CHUV CDS group lead

AI

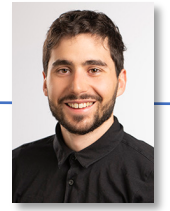
CHUV

Clinical Data Science (CDS) group

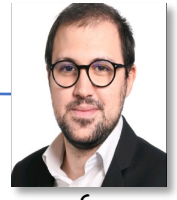
Dr. Jean Louis Raisaro, PhD
Group leader

Jérémie Despraz,MS
Principal data scientist

He Xu, PhD
Senior data scientist



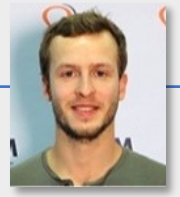
Jules Fasquelle, MS
Data privacy scientist



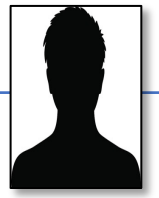
Bayrem Kabaachi, MS
Data privacy scientist

SDSC EPFL

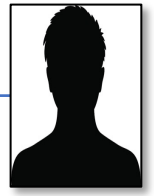
Snezana Nektarijevic, PhD
Embedded data scientist



Raphael Matusiak, MS
Embedded data scientist



MS Student intern



MS Student intern

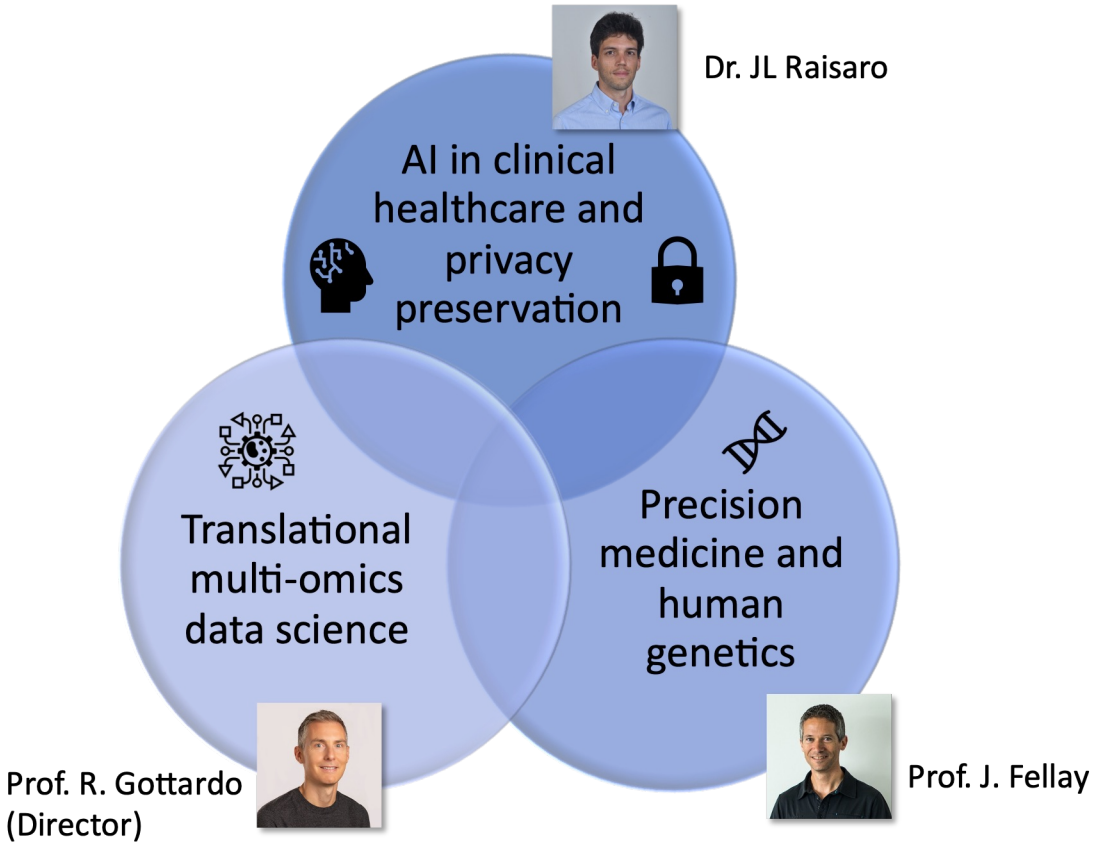
Core team

Extended team

- Anonymization and privacy-enhancing technologies for health data
- Multimodal AI and machine learning for clinical decision support
- Natural language processing
- Data interoperability and knowledge representation
- MLOps for automation and operationalization of data science workflows

Part of the CHUV "Biomedical Data Science Center"

https://www.chuv.ch/en/bdsc/



30+ collaborators

BDSC's Mission

Accelerating the organization and exploitation of biomedical big data to enable personalized medicine.

Key principles

Patient centricity

- Respect for privacy
- Patient engagement
- Consent

Data centricity

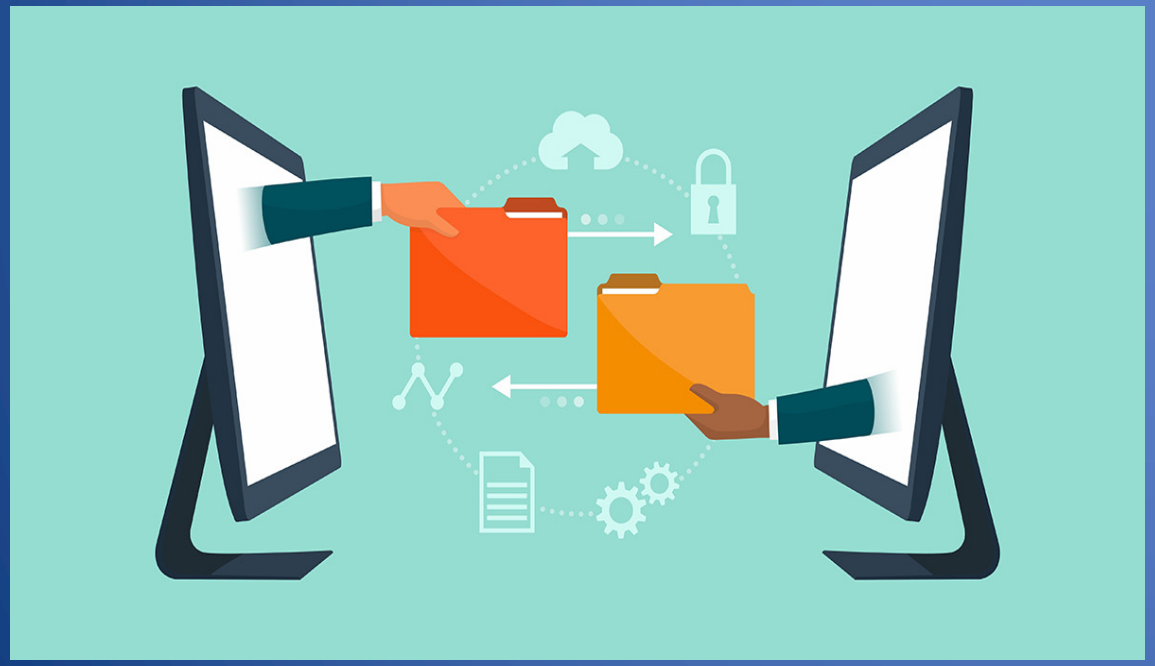
- FAIR data guiding principals: Findable, Accessible, Interoperable, Reusable

Modern and principled tools

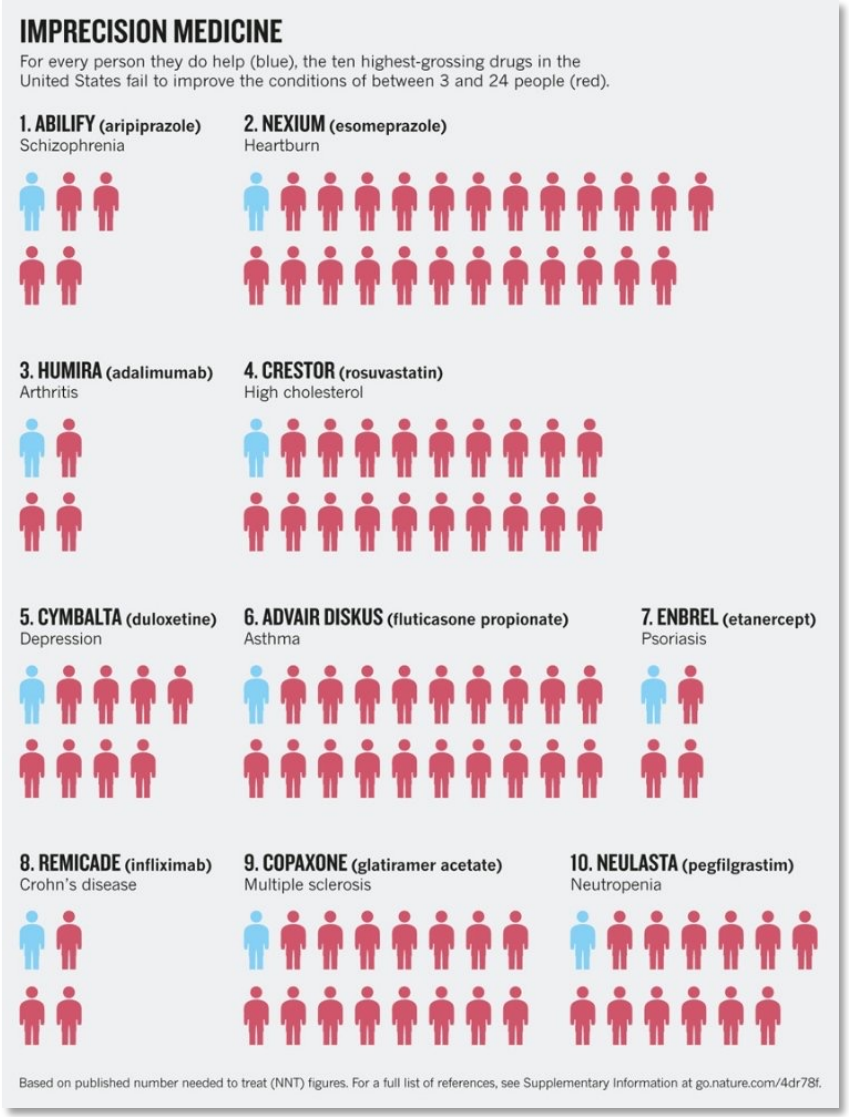
- State of the art secure infrastructures to facilitate data access and management
- Principled and mathematically-based

Our two complementary research areas

- Privacy-preserving health data sharing
- AI-based clinical decision support tools



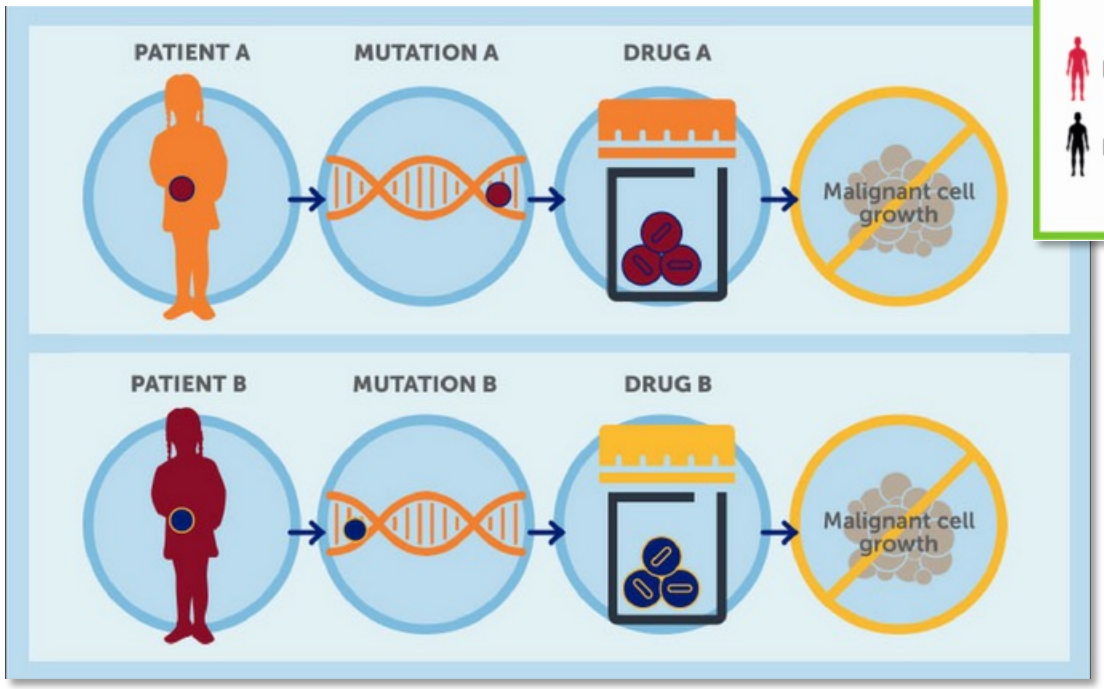
From imprecision medicine …

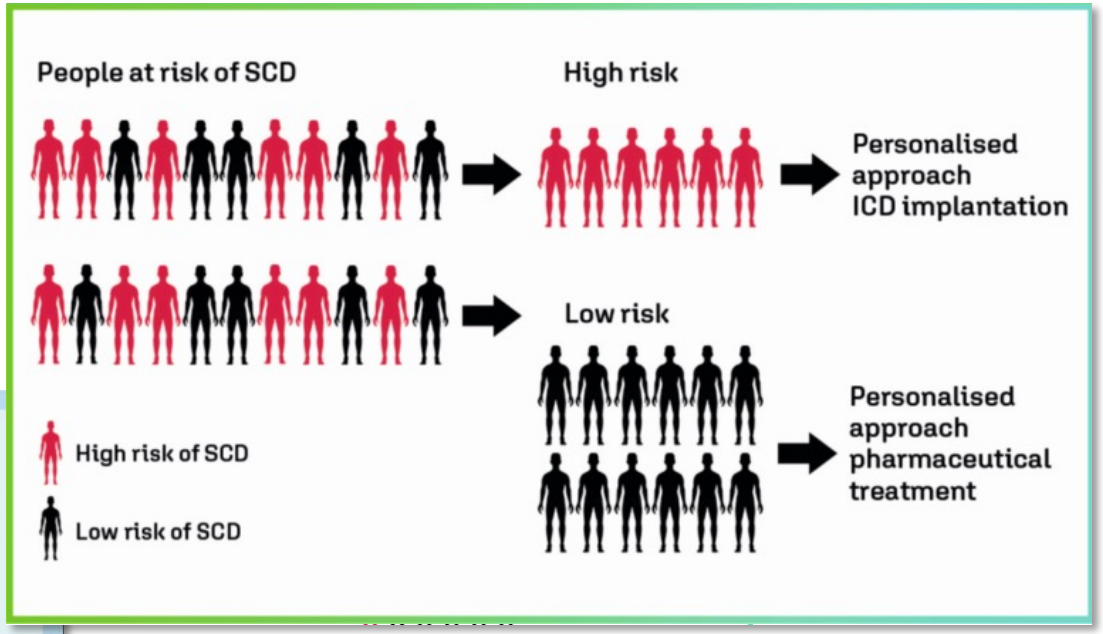


For every person they do help (blue), the ten highest-grossing drugs fail to improve the conditions of between 3 and 24 people (red)

… to precision medicine

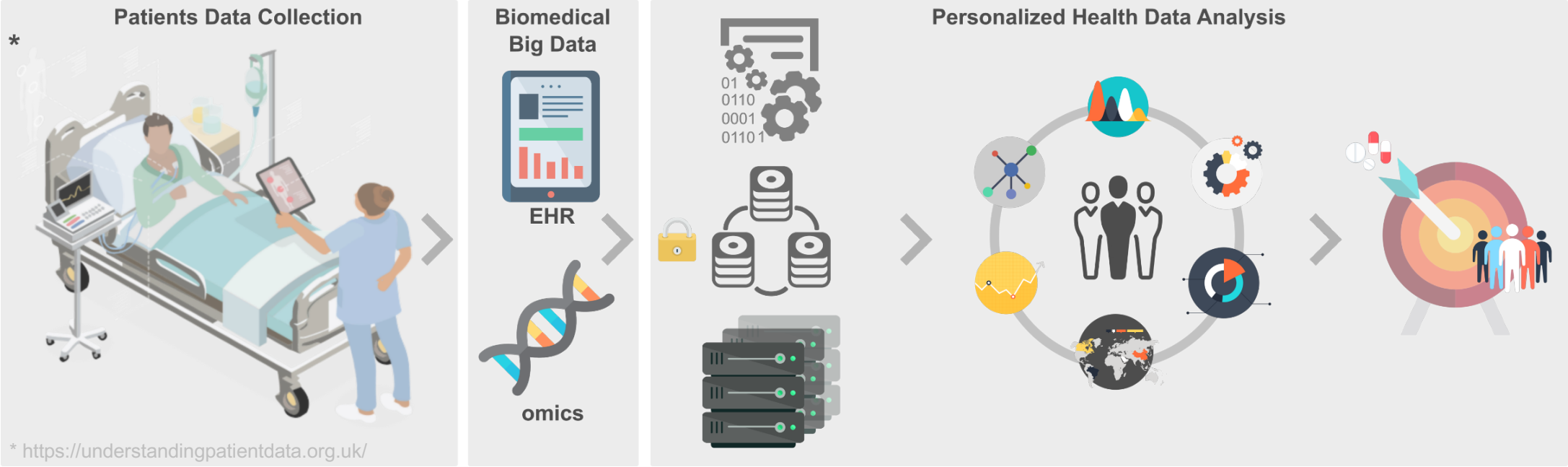
- Targeted cancer treatments (e.g., immunotherapy)
- Prevention of sudden cardiac death (SCD)



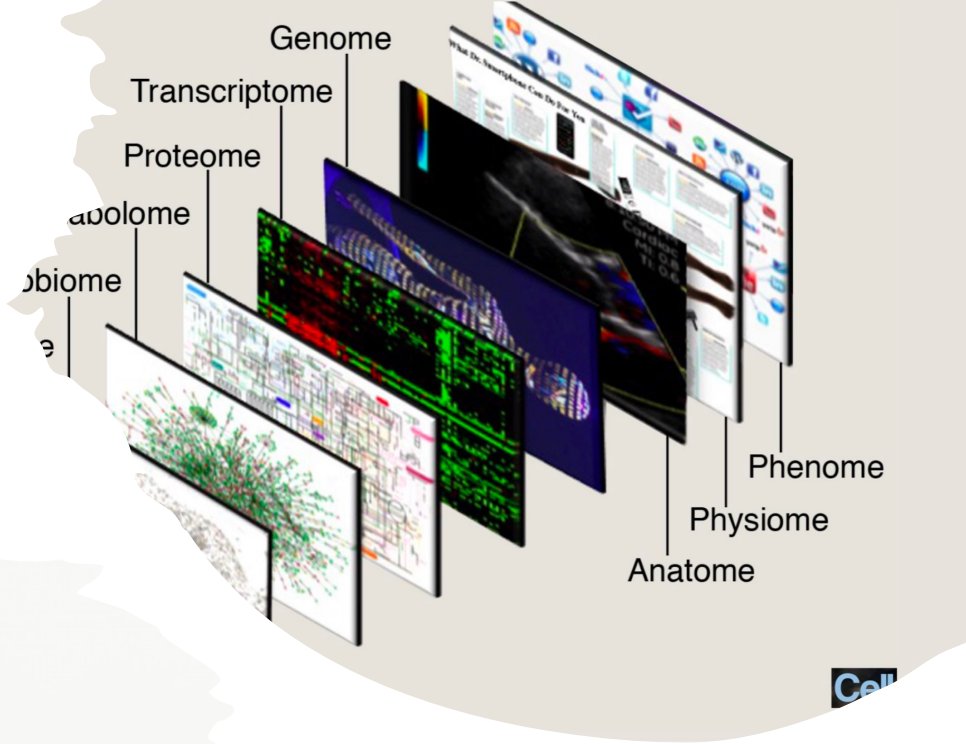


The end goal

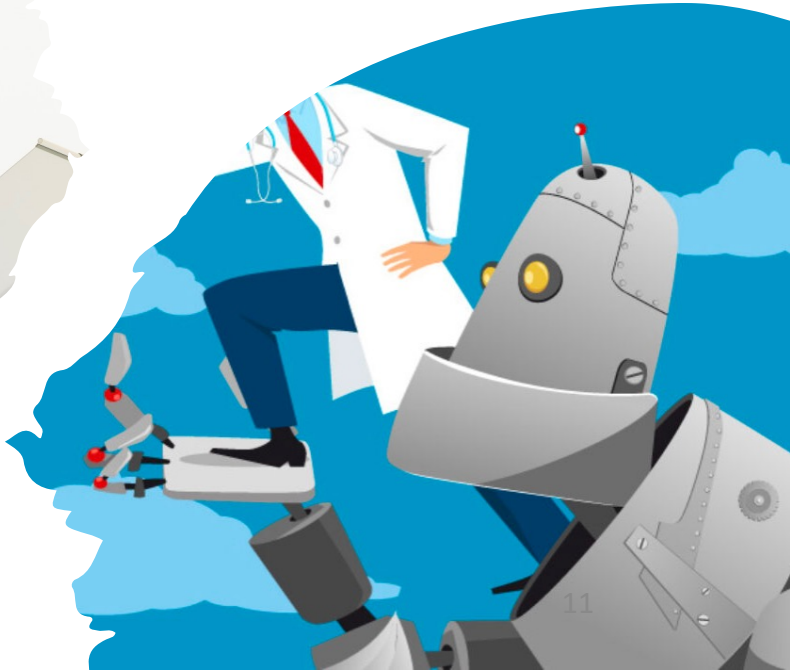
Provide patients with the right treatment at the right time through advanced data analysis using large volumes of medical data (big data) and artificial intelligence tools.



Many open challenges still ahead of us



- Data sharing and privacy
- Data standardization and interoperability
- Transparency and reliability of data and AI algorithms
- Patient safety and accountability
- Human-AI interaction and workforce displacement
- Education of an AI-literate workforce



nature

View all journals | Search | Log in

Explore content | About the journal | Publish with us | Subscribe | Sign up for alerts | RSS feed

nature > technology features > article

TECHNOLOGY FEATURE | 03 October 2022 | Correction 04 October 2022

Taking the pain out of data sharing

Despite agreeing to make raw data available, some authors fail to comply. The right strategies and platforms can ease the task.

Matthew Hutson

nature

Explore content | About the journal | Publish with us | Subscribe

nature > technology features > article

TECHNOLOGY FEATURE | 09 January 2023 | Correction 12 January 2023

The reproducibility issues that haunt health-care AI

Health-care systems are rolling out artificial-intelligence tools for diagnosis and monitoring. But how reliable are the models?

Emily Sohn

nature

Explore content | About the journal | Publish with us | Subscribe

nature > news > article

NEWS | 21 June 2022

Many researchers say they'll share data — but don't

Reasons included a lack of informed consent or ethics approval to share; misplaced data; and that others had moved on from the project.

Clare Watson

Sharing health data is (extremely) hard

Technical challenges:

- Vendor lock-in EHR systems
- Lack of semantic interoperability
- Cyber security

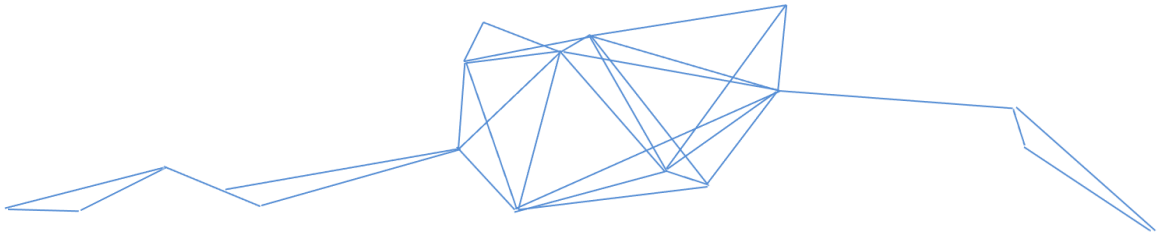
Cultural challenges:

- Data ownership
- Reputation

Legal and ethical challenges:

- Stringent regulations
- Consent
- Incompatibility of regulations across jurisdictions

Solutions in place today at CHUV/UNIL for open data



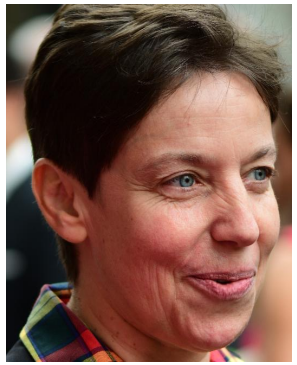
Cécile Lebrand, PhD

Responsable du service de soutien pour la gestion des publications et des données de recherche FBM
Consultante recherche Open Science FBM
Bibliothèque universitaire de médecine
Tél. ++41 (0)21 314 50 81
Cecile.Lebrand@chuv.ch
https://www.bium.ch/en/publication-open-access/

Data steward, UNIRIS, UNIL
researchdata@unil.ch
https://www.unil.ch/uniris/home/menuinst/donnees-de-recherche.html

UNIL | Université de Lausanne
Faculté de biologie et de médecine

Dr. Patrick Furrer



Dr. Cécile Lebrand

RISK-BASED DE-IDENTIFICATION: A POSSIBLE WAY FORWARD?

Reminder on the legal context

Country	Legal basis	Further use of data (genetic/non-genetic)
	• Human Research Act (HRA) • Human Research Ordinance (HRO)	• Uncoded • Coded • Anonymized
	• Federal Act on Data Protection (FADP) • Data Protection Regulations of Swiss cantons	• Personal • Anonymous
	• General Data Protection Regulation (GDPR)	• Pseudonymized • Anonymized
	• Health Insurance Portability and Accountability Act (HIPAA) Privacy Rule	• De-identified

Coding vs. Anonymization

Ordinance on Human Research with the Exception of Clinical Trials

(Human Research Ordinance, HRO)

– **Art. 25** Anonymisation

[1] For the anonymisation of biological material and health-related personal data, all items which, when combined, would enable the data subject to be identified without disproportionate effort, must be irreversibly masked or deleted.

[2] In particular, the name, address, date of birth and unique identification numbers must be masked or deleted.

– **Art. 26** Coding

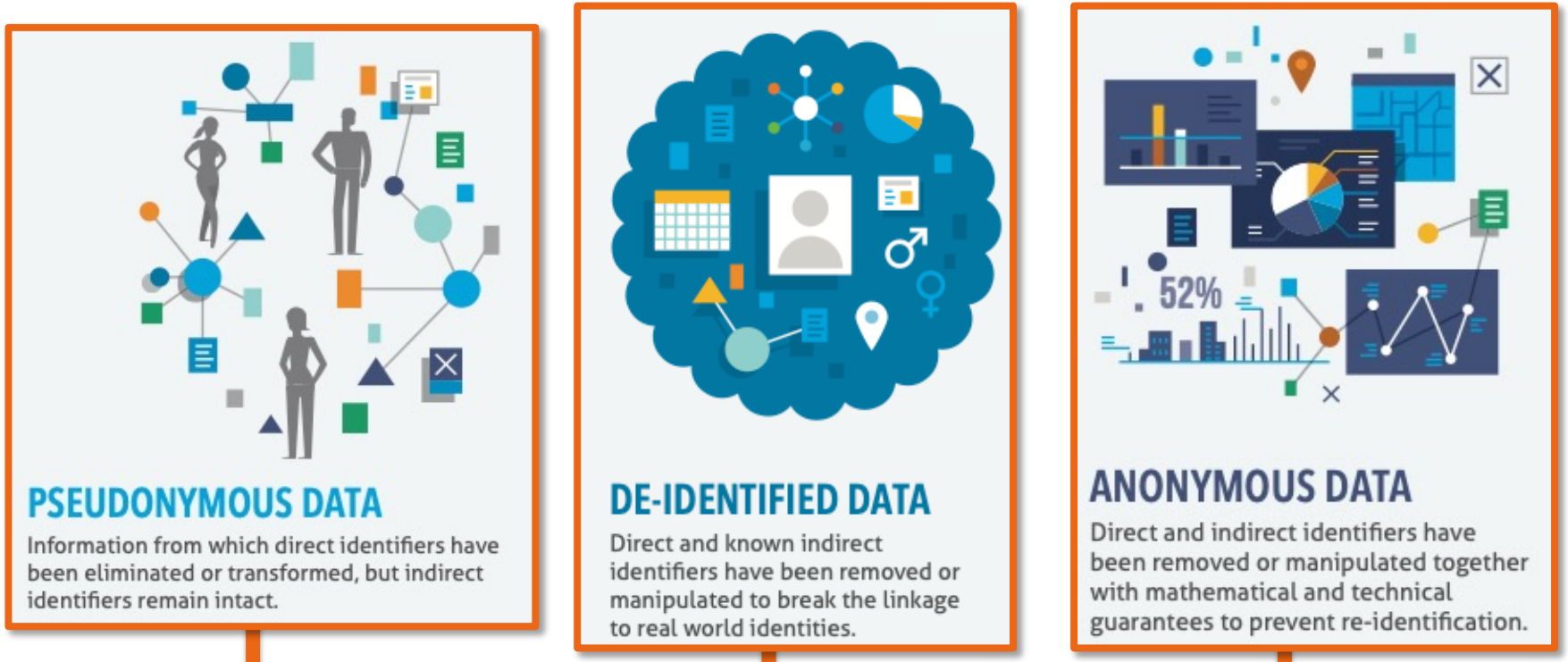
[1] Biological material and health-related personal data are considered to be correctly coded in accordance with Article 32 paragraph 2 and Article 33 paragraph 2 HRA if, from the perspective of a person who lacks access to the key, they are to be characterised as anonymised.

[2] The key must be stored separately from the material or data collection and in accordance with the principles of Article 5 paragraph 1, by a person to be designated in the application who is not involved in the research project.

Data are supposed to be truly **anonymized**, if re-identification of a person is **only possible with a disproportionate effort**.

Coded or pseudonymized data are de-identified data which **are still considered as personal data**.

Data identifiability continuum (technical definitions)

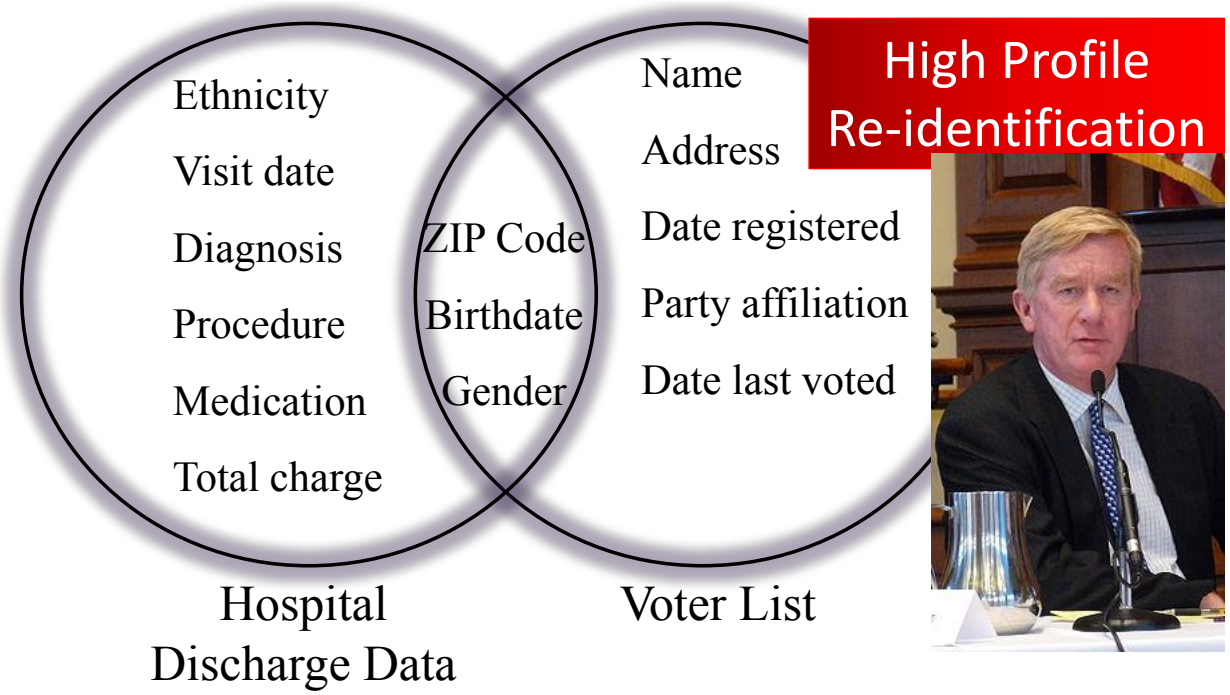


Personally identifiable data

Truly anonymized data

A “Quasi-identifier” conundrum

Sweeney. Journal of Law, Medicine, & Ethics. 1997



Credit: Prof. Bradley Malin, Vanderbilt University

5-Digit US ZIP Code

+ Birthdate

+ Gender

63-87% of USA estimated to be unique

Re-identification risk as a function of data risk and context risk

$$P(reid) \approx P(attack) * P(reid \mid attack)$$

Context

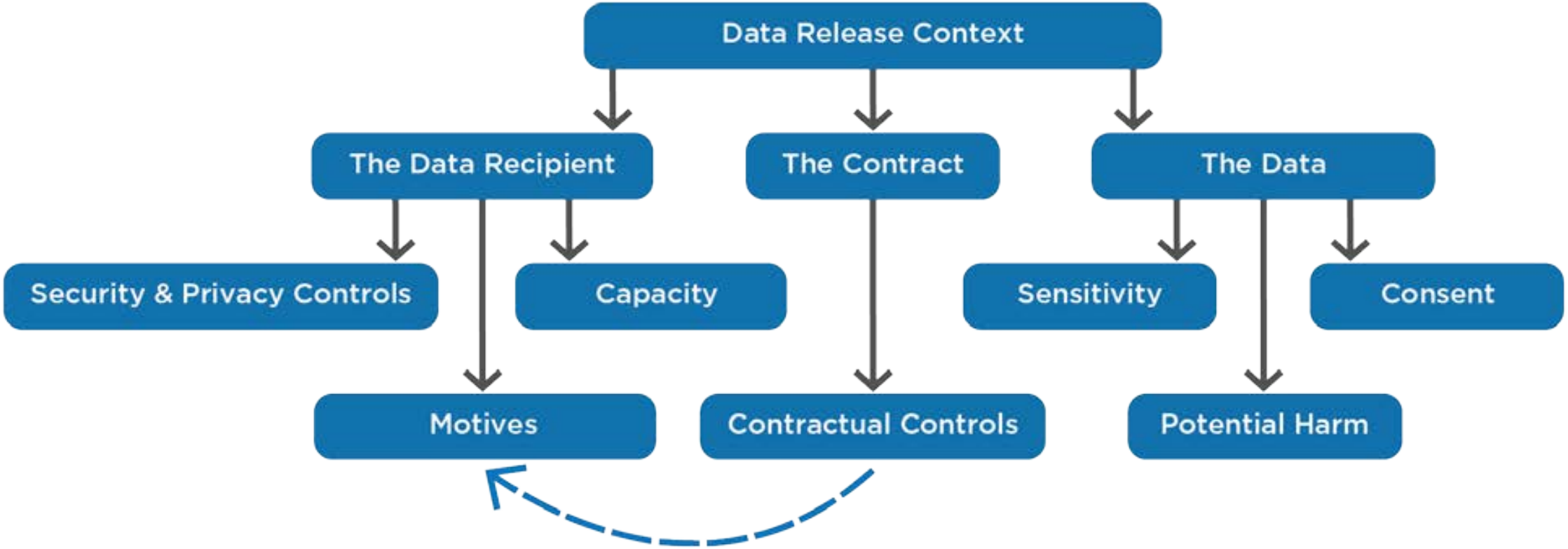
- Use Agreements
- Pay for Access
- Unique Login / Pass
- Audit

Data

- Uniqueness
- Replicability
- Availability

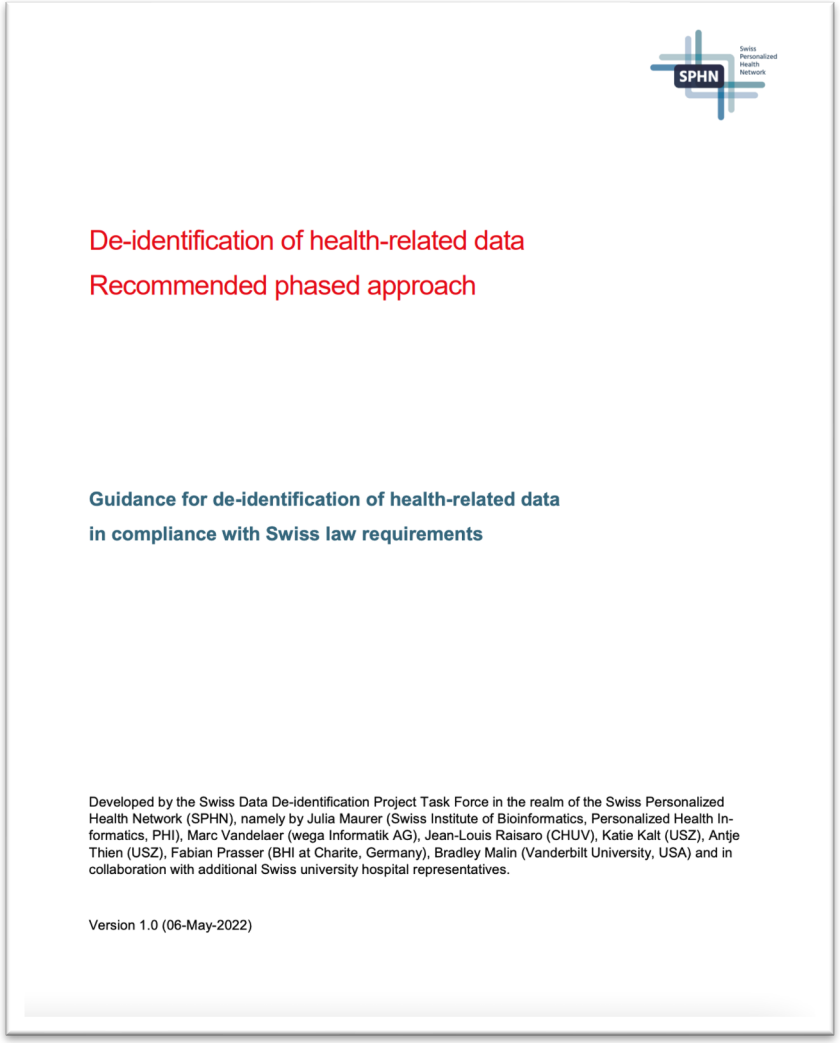
Credit: Prof. Bradley Malin, Vanderbilt University

What do we mean by context?



Credit: Prof. Khaled El Emam, Ottawa University

Data de-identification in the SPHN



De-identification of health-related data

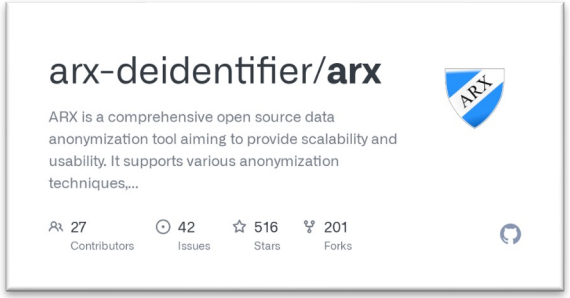
Recommended phased approach

Guidance for de-identification of health-related data in compliance with Swiss law requirements

Developed by the Swiss Data De-identification Project Task Force in the realm of the Swiss Personalized Health Network (SPHN), namely by Julia Maurer (Swiss Institute of Bioinformatics, Personalized Health Informatics, PHI), Marc Vandelaer (wega Informatik AG), Jean-Louis Raisaro (CHUV), Katie Kalt (USZ), Antje Thien (USZ), Fabian Prasser (BHI at Charite, Germany), Bradley Malin (Vanderbilt University, USA) and in collaboration with additional Swiss university hospital representatives.

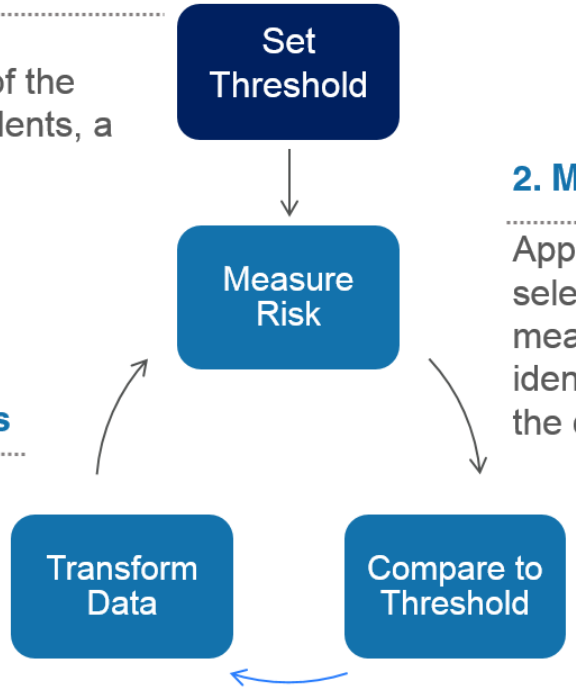
Version 1.0 (06-May-2022)

A phased and iterative process



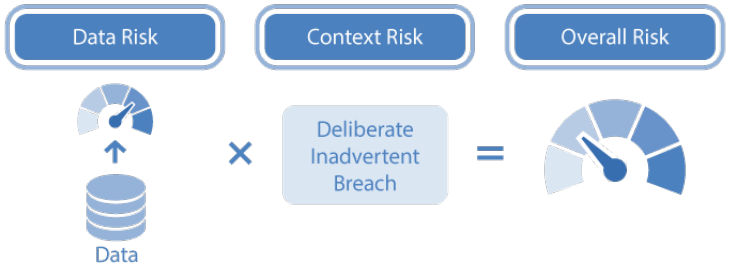
1. Set Risk Threshold

Based on the characteristics of the data and precedents, a quantitative risk threshold is set.



2. Measure Risk

Appropriate metrics are selected and used to measure re-identification risk from the data.



3. Evaluate Risk

Compare the measured risk against the threshold to determine if it is above or below it.

4. Apply Transformations

If the measured risk does not meet the threshold, specific transformations are applied to reduce the risk.

SPHN Risk assessment framework

- Multi-dimensional assessment (data + context)
- Developed by experts on data protection and data privacy from the 5 Swiss University Hospitals, SIB and international experts
- Presented to Swissethics

IT-security and contractual measures

The tab "IT-security and contractual measures" contains questions related to geographic risk, contracts and policies, cohort size, data access, infrastructure and security. Selected answers are associated with a risk level leading to a risk weight and risk value per question. Note that some answers contain a notification "yes, condition needs explicit description for ethics approval". This means, it is recommended to raise attention to this issue in the project proposal and describing the condition explicitly when applying for ethics approval. For example, sending health-related data outside of Switzerland is accompanied with higher or lower data protection measures and has to be evaluated carefully. For describing the condition explicitly the comment field might be used.
Fill in blue cells with an "x" on the appropriate answer "Yes" or "No". Please note that only either yes or no should be selected, but that, if specifically mentioned here below, some questions allow multiple sele

Additional information: The BioMedIT network usage for a secure data transfer and hosting reduces significantly the risk profile and accounts for a lower total risk score.

Additional infor answers associ Switzerland with

No. of high risk answers:	0
BioMedIT:	NO
Total Risk	0

Question #	Topic	Answer #	Possible answers	Condition needs explicit description for ethics approval		Yes	No	Risk level 0 = Null 1 = Low 2 = Medium 3 = High	Risk weight 1 = Lowest 10 = Highest	Risk value (Risk Level * Risk weight)
Geographic risk										
C-01	**Where is the health-related data planned to be stored and processed?**	C-01-01	In Switzerland		Select as many as apply			1	5	**0**
		C-01-02	In EU	Yes				2		
		C-01-03	Outside of Switzerland and EU with adequate safeguards	Yes				2		
		C-01-04	Outside of Switzerland and EU without adequate safeguards	Yes				3		
Contracts and policies										
C-02	**What is covered by the legal agreement regulating the conditions under which data are disclosed to the parties (recipients)?**	C-02-01	The legal agreement forbids the recipient from disclosing the data to third parties		Select as many as apply					
		C-02-02	The legal agreement allows audits of the recipient's data management practices							
		C-02-03	The legal agreement stipulates that regular third party privacy and security audits may be performed at the recipient site and of the recipient's practices							
		C-02-04	The legal agreement imposes strong limits on the linkage of then provided health-related data with other administrative or clinical data sources							
		C-02-05	The legal agreement associates penalties in case of health-related data misuse by the recipient							
		C-02-06	There is no legal agreement set up	Yes						
C-03	**Are there IT security and privacy policies in effect at the data recipient site?**	C-03-01	The recipient has written data privacy and IT security policies		Select as many as apply					
		C-03-02	There is a person on the data recipient side responsible for data privacy							
C-04	**Is there a legal agreement between the data recipient (i.e. the Data Controller) and its external processor?**	C-04-01	There is no legal agreement set up and its data processor(s)							
		C-04-02	There is a legal agreement between the data recipient and its data processor(s)							

Template change history | Version history | Project overview | ITsecurity+contractual measures | Dat

Data de-identification

The tab "Data de-identification" provides information about variables used in the project and de-identification rule chosen to mitigate the risk of re-identification. It differentiates between demographic and administrative, multimedia and genomic variables and DICOM a a risk level leading to a risk weight and value per question.
Note that some answers contain a notification "yes, condition needs explicit description for ethics approval". This means, it is recommended to raise attention to this issue in the project proposal and describing the condition explicitly when applying for ethics approva name is essential for the project, it has to be justified in the study protocol sent to the ethics committee. For describing the condition explicitly the comment field might be used.

Additional information: The number of high risk rules summarizes answers associated with a risk level of 3, such as keeping one or more direct identifiers (patient's name) or keeping hardware identifying attributes when using DICOM files

k rules:	0
Total Risk	0

Variable #	Identifying and quasi-identifying variables	De-identification Rule #	De-identification rule description	Selected Rule	Condition needs explicit description for ethics approval	Risk level 0 = null 1 = low 2 = medium 3 = high	Relative risk-weight per identifier/quasi-identifier 1 = Lowest 10 = Highest	Risk value per identifier/quasi-identifier (Normalized risk ranking * Risk weight)
Demographic and administrative variables								
D-01	**Direct identifiers (e.g., name, phone number social security number, email address, medical record number, patient-ID, sample-ID, license number, address)**	D-01-01	Identifiers are suppressed (Note: only applicable for structured data)			0	10	**0**
		D-01-02	Identifiers are replaced by pseudonym			1		
		D-01-03	Original values of one or more direct identifiers are kept		Yes	3		
D-02	**Dates in the patient record (dates of birth and death excluded)**	D-02-01	Dates are suppressed or replaced with a surrogate date (default)			0	3	**0**
		D-02-02	Dates are shifted by a random number of days within +/- 365 days or generalized to the year (i.e. provide year only, suppress day/month)			1		
		D-02-03	Dates are shifted by a random number of days within +/- 90 days (one quarter offset to preserve seasonality) or generalized to quarter and year			1		
		D-02-04	Dates are shifted by a random number of days within +/- 30 days (one month offset to preserve seasonality) or generalized to month and year			2		
		D-02-05	Dates are shifted by a random number of days within +/- 7 days (default; one week offset)			2		
		D-02-06	Original dates are kept		Yes	3		
D-03	**Date of birth**	D-03-01	Date of birth is suppressed or shifted by a random number of days (default)			0	6	**0**
		D-03-02	Only the year of the original birth date is kept			1		
		D-03-03	Only the year and month of the original birth date are kept			2		
		D-03-04	Full original date of birth is kept (dd/mm/yyyy)		Yes	3		
D-04	**Date of death**	D-04-01	Date of death is suppressed or shifted by a random number of days (default)			0	6	**0**
		D-04-02	Only the year of the original death date is kept			1		
		D-04-03	Only the year and month of the original death date are kept			1		
		D-04-04	Full original date of death is kept (dd/mm/yyyy)			2		
		D-04-05	Full original date of birth is kept (dd/mm/yyyy)		Yes	3		
D-05	**Age at admission / death**	D-05-01	Age is suppressed (default)			0	5	**0**
		D-05-02	Age in generalized in groups of 5 or more years			1		
		D-05-03	Original age is kept except for people with more than 89y old who are put in the age class "90y+"			2		

Additional information: If one of those dates is shifted, the date of birth and date of death should be shifted in the same way.

Template change history | Version history | Project overview | ITsecurity+contractual measures | Data de-identification | Project risk profile | +

SPHN Risk assessment framework

Categorization of risk score thresholds		
Low (Risk score = 1)	**Medium** (Risk score = 2)	**High** (Risk score = 3)
< 129	129 to 258	> 258
< 105	105 to 210	> 210
Project risk score thresholds		
< 0,51	**0,51 to 1,00**	**> 1,00**

k profile					
			Risk value subtotal	**Category weight**	**Risk score**
infrastructure and security)					
Number of high risk answers	6		94	50%	1
BioMedIT usage	No				
Data (demographic and administrative, multimedia, genomic variables and DICOM attributes)					
Number of high risk rules:	5		112	50%	2
Risk assessment outcome					
Number of high risk	11	Total Risk Score:			0,75

- Multi-dimensional assessment (data + context)
- Developed by experts on data protection and data privacy from the 5 Swiss University Hospitals, SIB and international experts
- Presented to Swissethics

https://sphn.ch/network/data-coordination-center/de-identification/

Next step: RDeID SPHN Demonstrator project

Contact | News | Funding | Ongoing projects | Grant Documents | DTUA | Documents | English | Search www.sphn.ch

SPHN Swiss Personalized Health Network

Menu

SPHN supports 11 Demonstrator projects with CHF 4.3 M

In 2022, SPHN launched a call for Demonstrator projects. These projects will test the infrastructures, processes, and data resources established in the realm of SPHN to demonstrate their added value for the network and to identify the remaining gaps. From a total of 30 project applications, 11 Demonstrator projects were selected for funding.

Two types of Demonstrator projects are supported: On the one hand, projects that test the practical application of SPHN infrastructure components in medical research and/or expand their use in the network. On the other hand, projects that demonstrate the added value of SPHN-compliant data resources from the university hospitals for personalized health research.

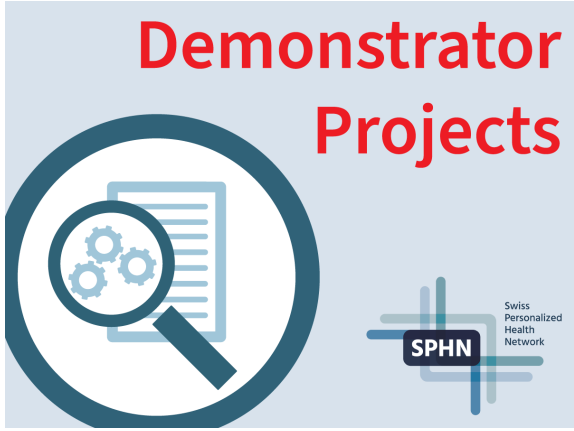
The SPHN International Advisory Board (IAB) selected the supported projects based on the following main criteria: use of SPHN infrastructures, added value for the network and personalized health research, and feasibility.

The following 11 projects will start in 2023. The names listed are the respective principal investigators:

- **SwissPedDW - Swiss Pediatric Data Warehouse,** Prof. Christoph Berger (University Children's Hospital Zurich, KiSpi)
- **Therapy-related myeloid neoplasms after cytotoxic treatment,** Dr. Sabine Blum (CHUV)
- **SwissPedHealth – PReparing PERsonalizEd PEdiatRic PRimaRy caRE (PREPP),** Prof. Jan Bonhoeffer (University Children's Hospital Basel, UKBB)
- **INFRA: INFection Radar**, Dr. Olga Endrich (Insel)
- **Accelerating detection of neonatal sepsis (ADONIS): a machine learning-based approach,** Prof. Eric Giannoni (CHUV)
- **Cohort demonstrator: Full integration of a national cohort into the SPHN infrastructure,** Dr. Michael Koller (USB)
- **Using routine health care data to facilitate clinical cohort studies (SPHN-SPAC)**, Prof. Claudia Kuehni (UniBE)
- **Smart SNOMED Search for SPHN (S4),** Prof. Christian Lovis (HUG/UniGE)
- **RDeID: Risk-based de-identification platform for health-related data**, Dr. Jean Louis Raisaro (CHUV)
- **EVIGAITCP**, Dr. Morgan Sangeux (University Children's Hospital Basel, UKBB)
- **Swiss Network of Wearables (SNOW),** Solange Zoergiebel (CHUV)

A total of CHF 4.3 million will be allocated to the Demonstrator projects. Each project will be supported with up to CHF 500'000 and has a maximum duration of 18 months. SPHN funding requires matching contributions by the participating institutions.

More information about the Demonstrator projects can be found here.



CHUV

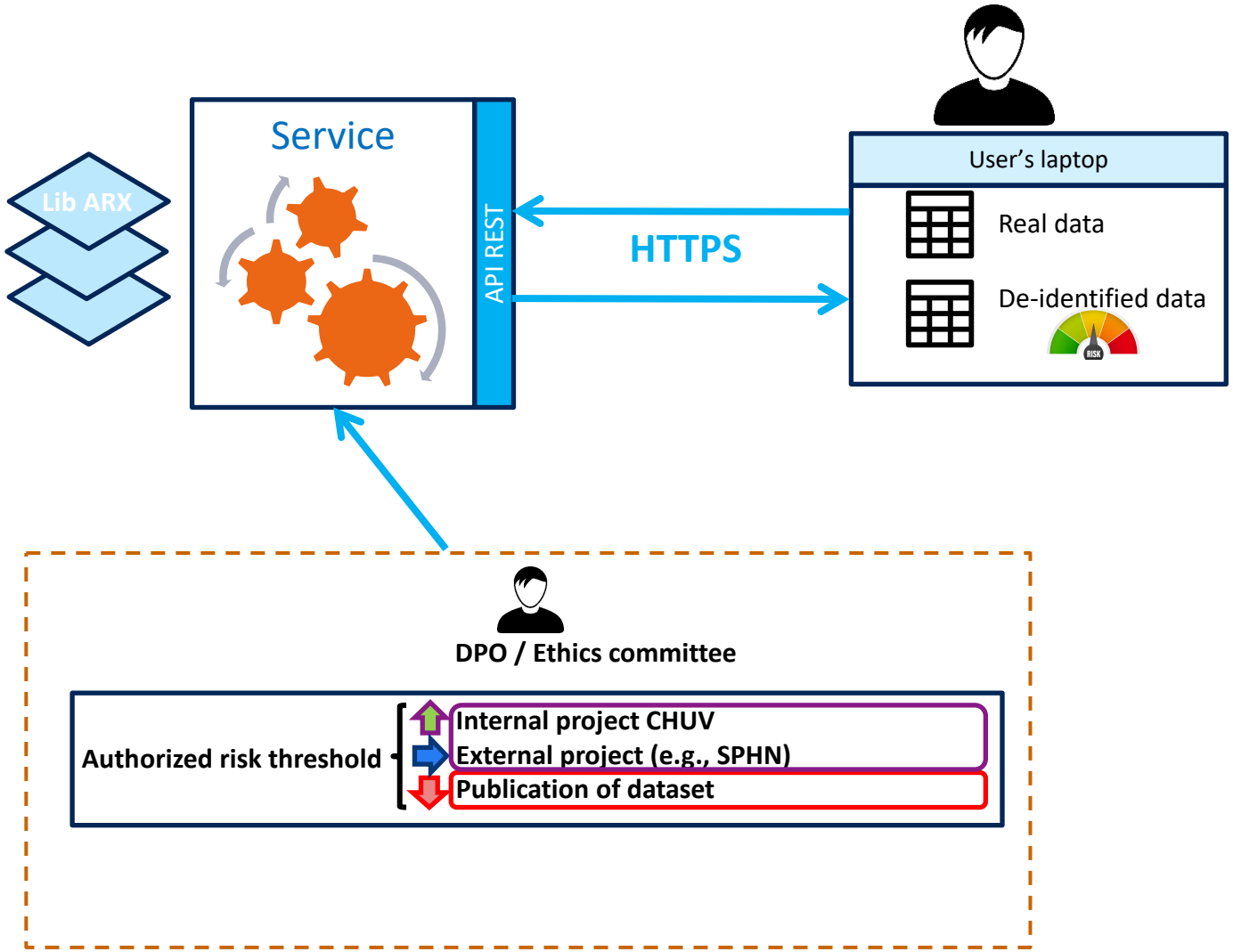
USZ

SIB Swiss Institute of Bioinformatics

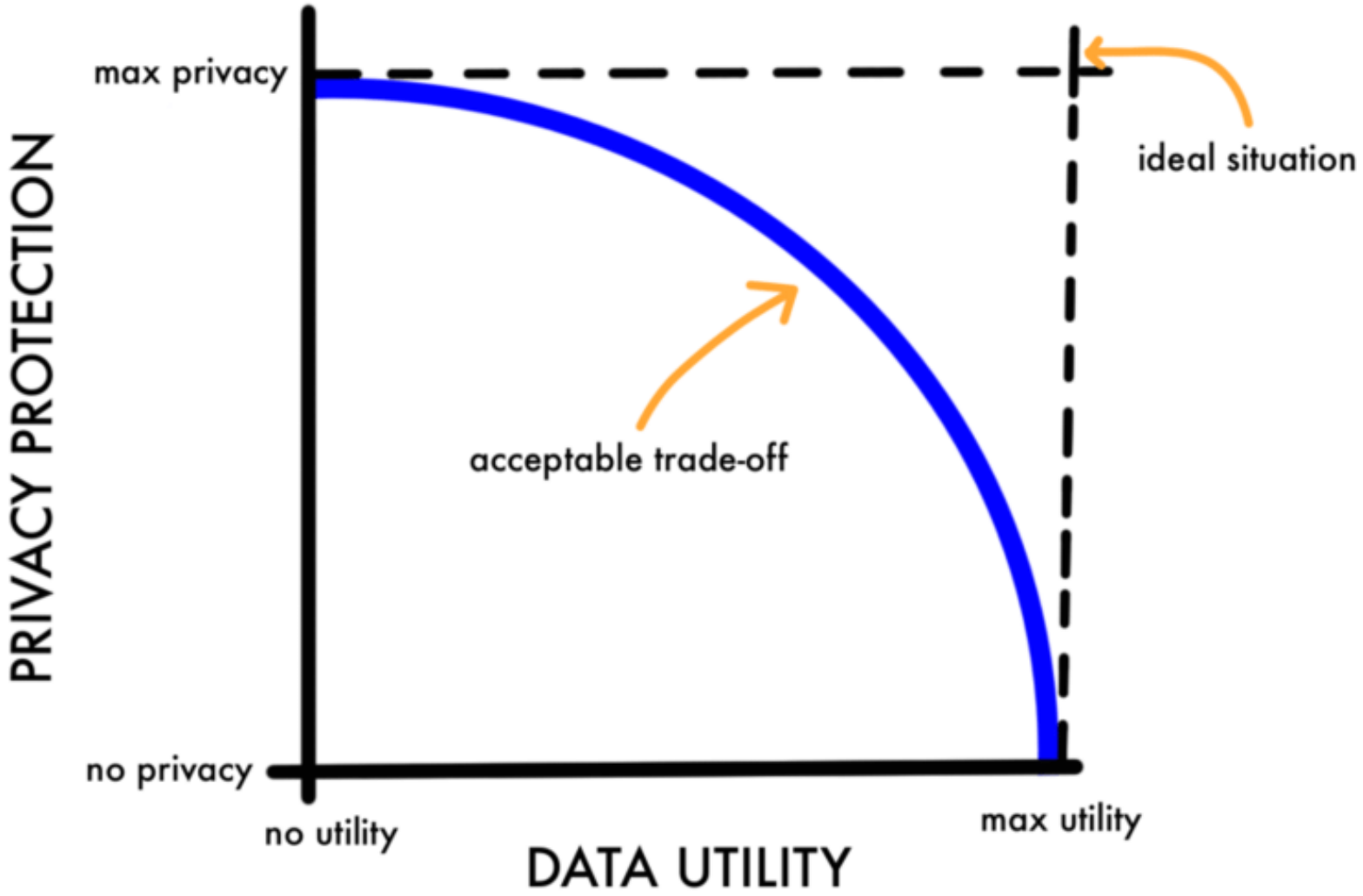
CHARITÉ UNIVERSITÄTSMEDIZIN BERLIN

canton de vaud

Goals of RDeID: automation and validation

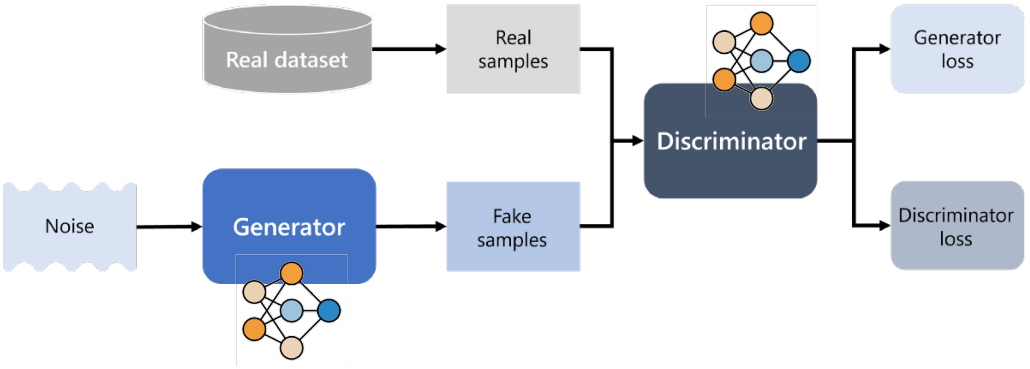


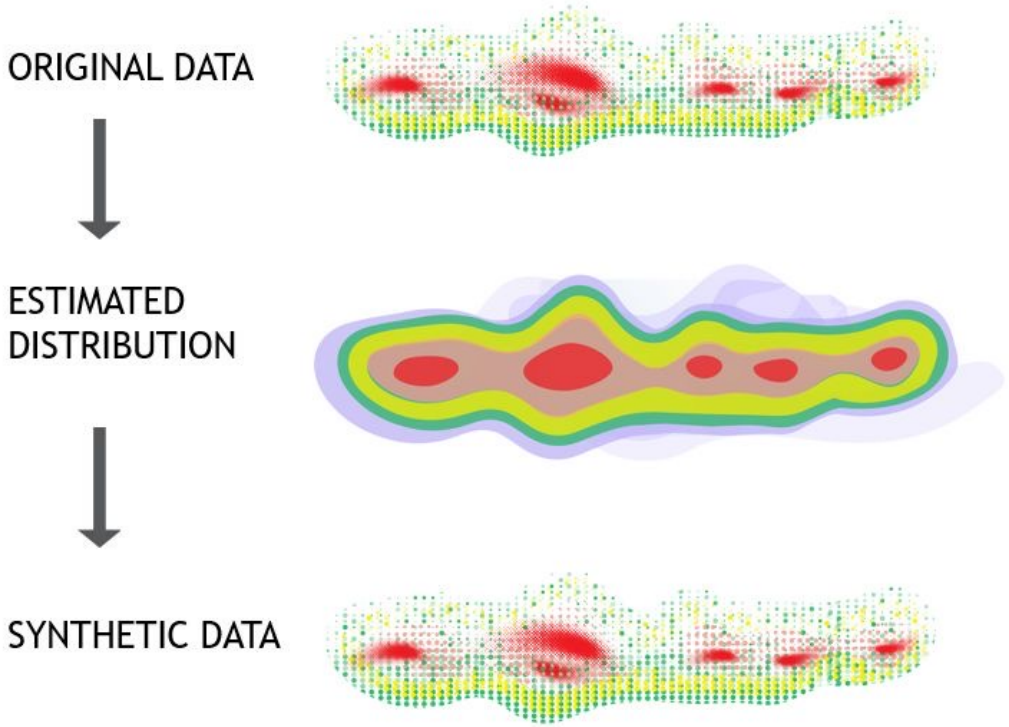
Data privacy vs. data utility trade-off



OTHER SOLUTIONS?

Synthetic data: a promising solution to alleviate the concerns on the privacy-utility trade-off

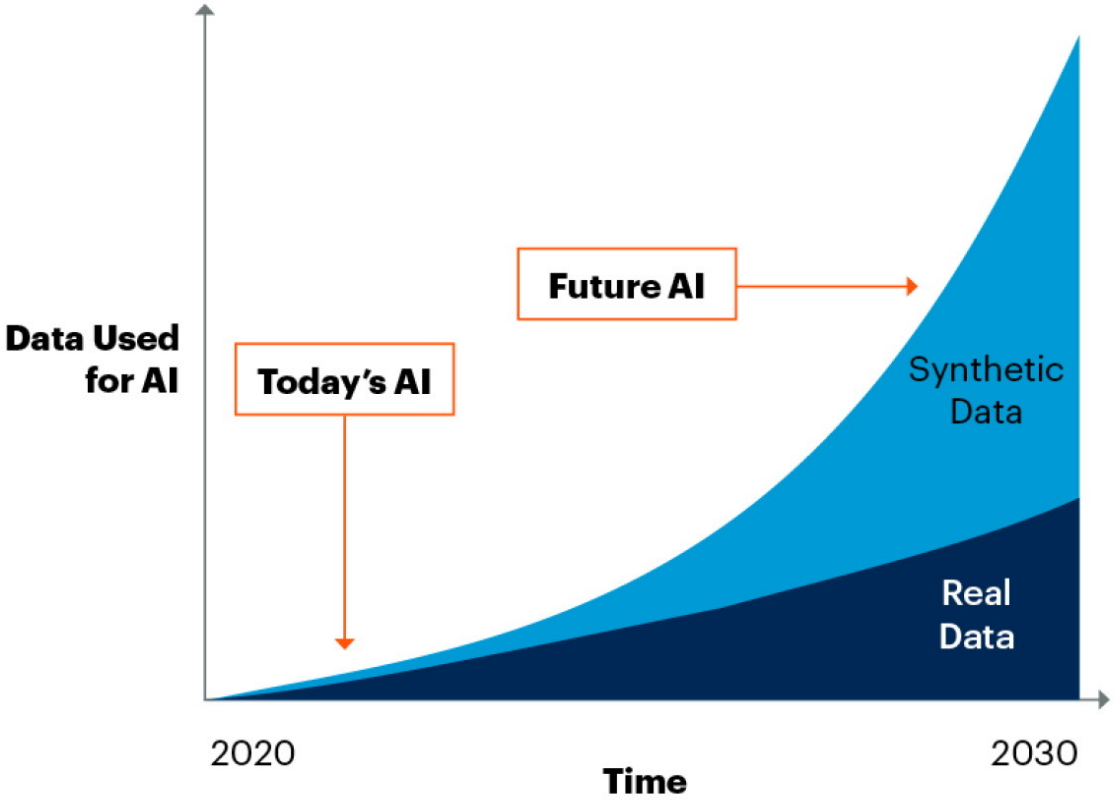




- Can address privacy concerns associated with real data
- Can address bias in real data with synthetic data diversification
- Can be a cost-effective approach for creating large datasets

Great expectations…

By 2030, Synthetic Data Will Completely Overshadow Real Data in AI Models



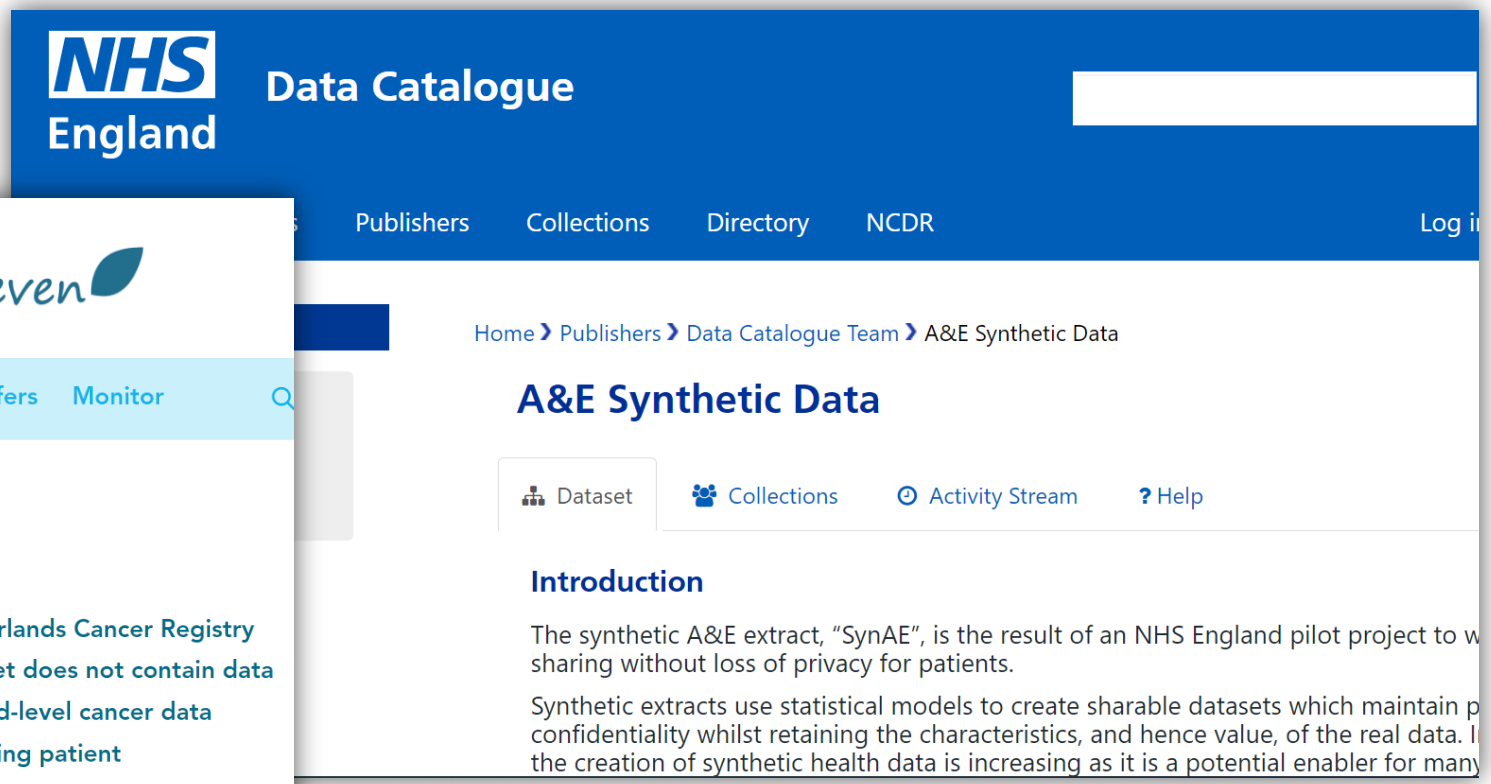
- Artificially Generated Data
- Generated From Simple Rules, Statistical Modelling, Simulation and Other Techniques

- Obtained From Direct Measurements
- Constrained by Cost, Logistics, Privacy Reasons

Source: Gartner
750175_C

Gartner

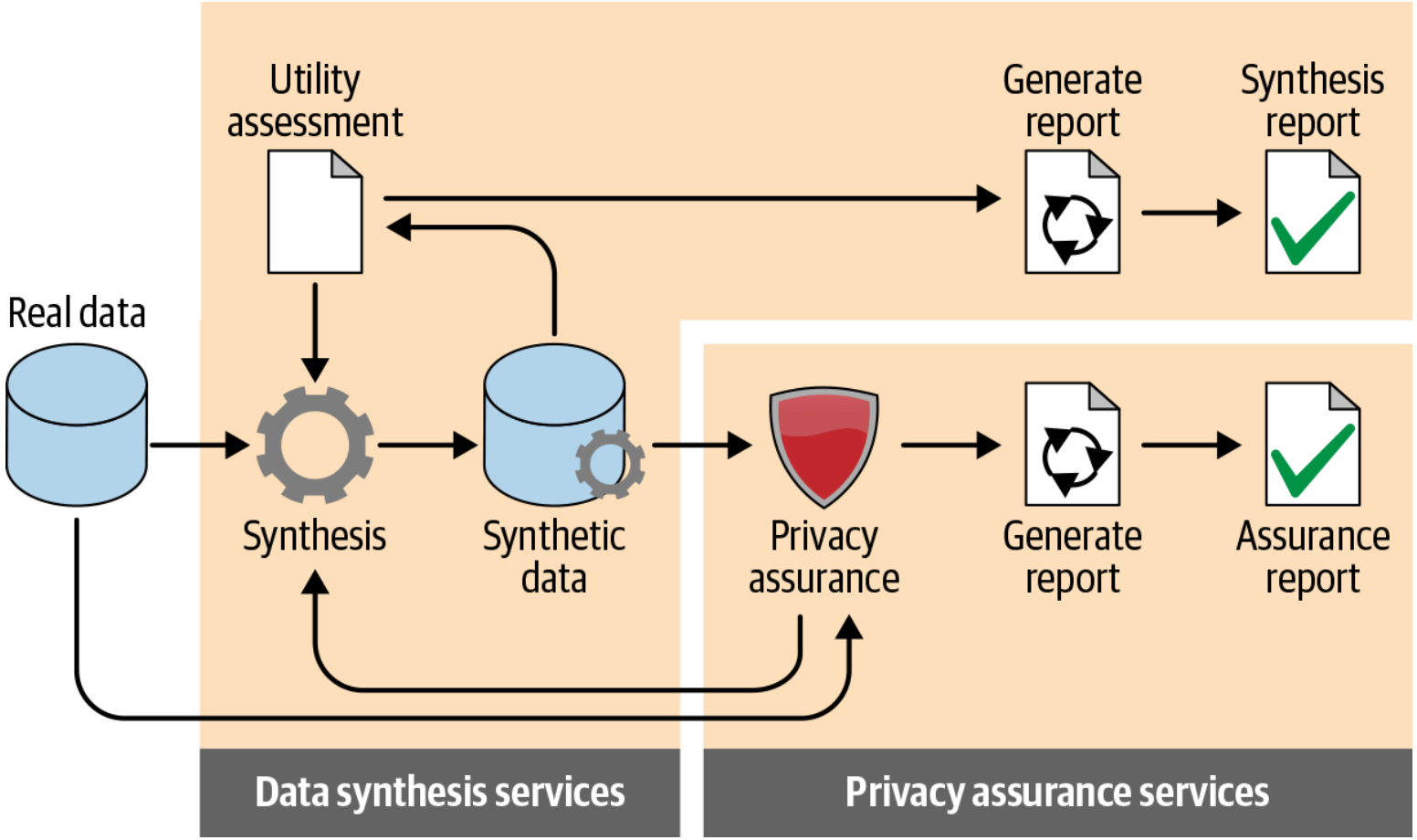
A few real-world use case examples…



Synthetic dataset (iknl.nl)

A&E Synthetic Data - Datasets - NHS England Data Catalogue

… But we need to carefully and systematically evaluate the residual risks and utility



Take-home messages

A new **biomedical data science center** has been created at CHUV to accelerate research and innovation around digital health and precision medicine

Work in progress to develop a **new privacy-preserving data sharing tools** at CHUV to facilitate research and open science

This work cannot be done in “silos” and **we need a collaborative effort** between, researchers, computer scientists, legal experts and regulators to validate and adopt these new approaches