ENABLING PRIVACY-PRESERVING ANALYSES ON FEDERATED HEALTHCARE DATA

Ton Peters, Daniël Worm, June 16th 2022

janssen KNL TNO innovation for life

AGENDA

- Introduction and why collaborate
- Aim
- Problem & Solution
- LANCELOT
- HERACLES as new initiative
- Conclusions and next steps

COLLABORATION BETWEEN TNO, IKNL AND JANSSEN

We aim to reduce impact of cancer by using AI to find the right treatments for the right patient at the right time and place

IKNL

- IKNL maintains the Netherlands Cancer Registry (NCR) to enable people to reflect on cancer care and prevention.
- Creates technologies to safely extract insights from distributed databases.

janssen

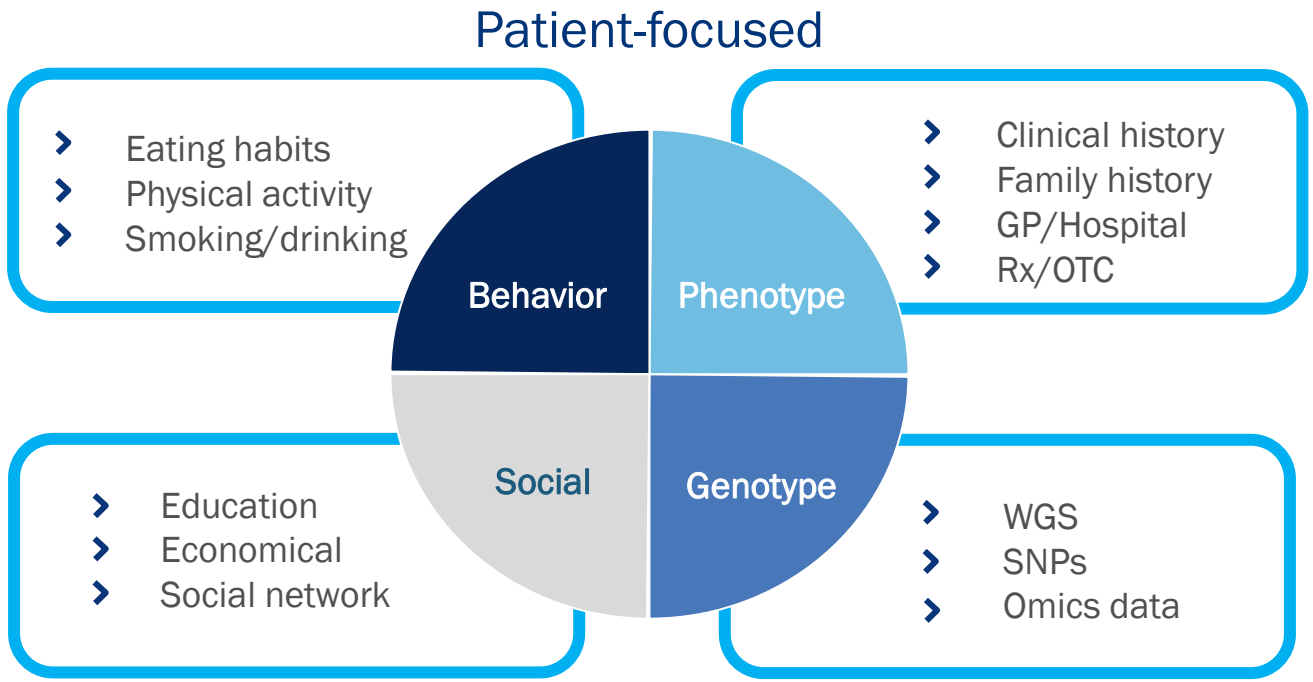
- Need for
  - evidence for personalized treatments (VBHC)
  - advanced methodologies for research (RWE)
- To overcome
  - bias due to GDPR
  - administrative burden of registries and clinical trials

TNO

- Create innovations by connecting people and knowledge
- Developing ICT solutions to enable privacy preserving analyses

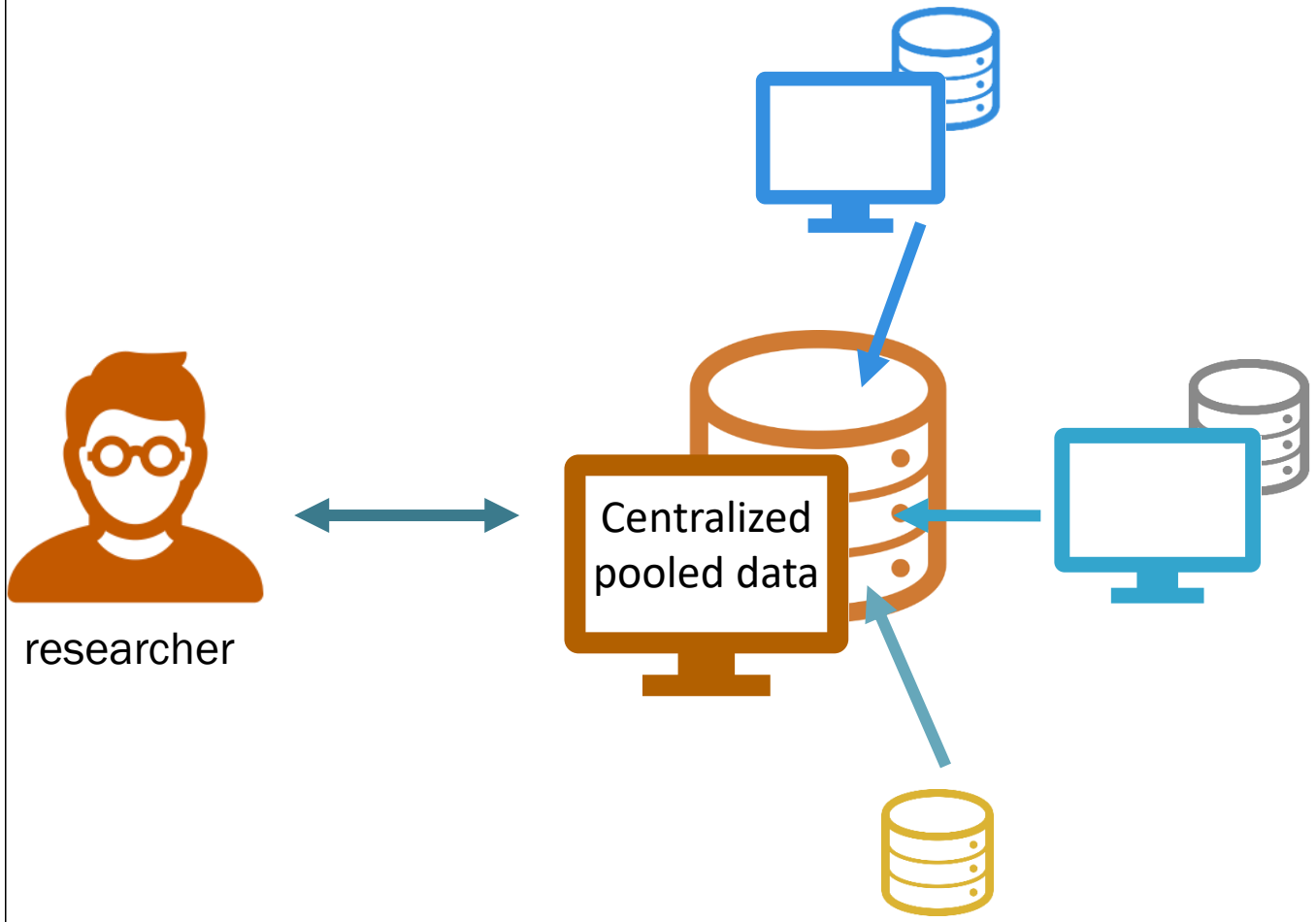
VALUE BASED HEALTH CARE
PERSONALIZED DRIVEN APPROACH

- Research in health care requires
  - FAIR data*
  - Harmonized data
  - Enriched data to get better and new insights
  - GDPR proof data
  - Captured with limited administrative burden



* data should be Findable, Accessible, Interoperable, Reusable

PROBLEM
TRADITIONAL APPROACH NOT SUFFICIENT



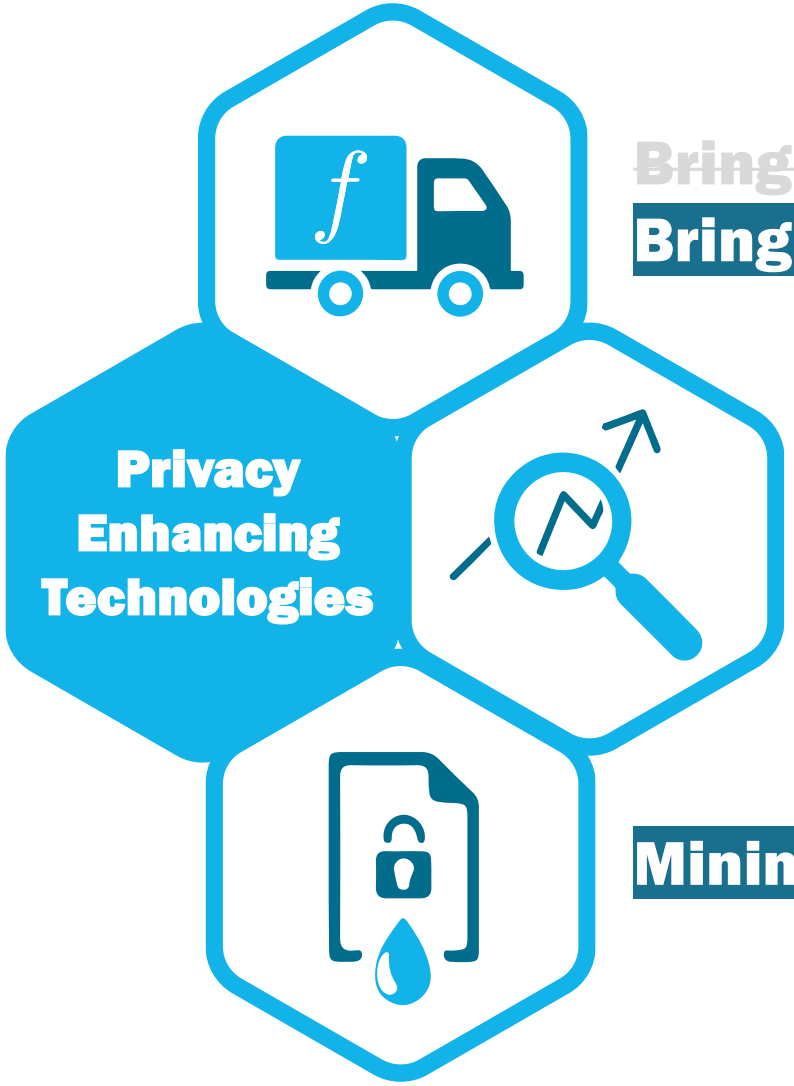
- Datasets are often limited and too small
- Compare regions not possible
- Informed Consent mandatory

DATA CAN OFTEN NOT “JUST” BE SHARED OR COMBINED

GDPR principles

- Data minimization
- Proportionality
- Data protection

PRIVACY ENHANCING TECHNOLOGIES OFFER A SOLUTION



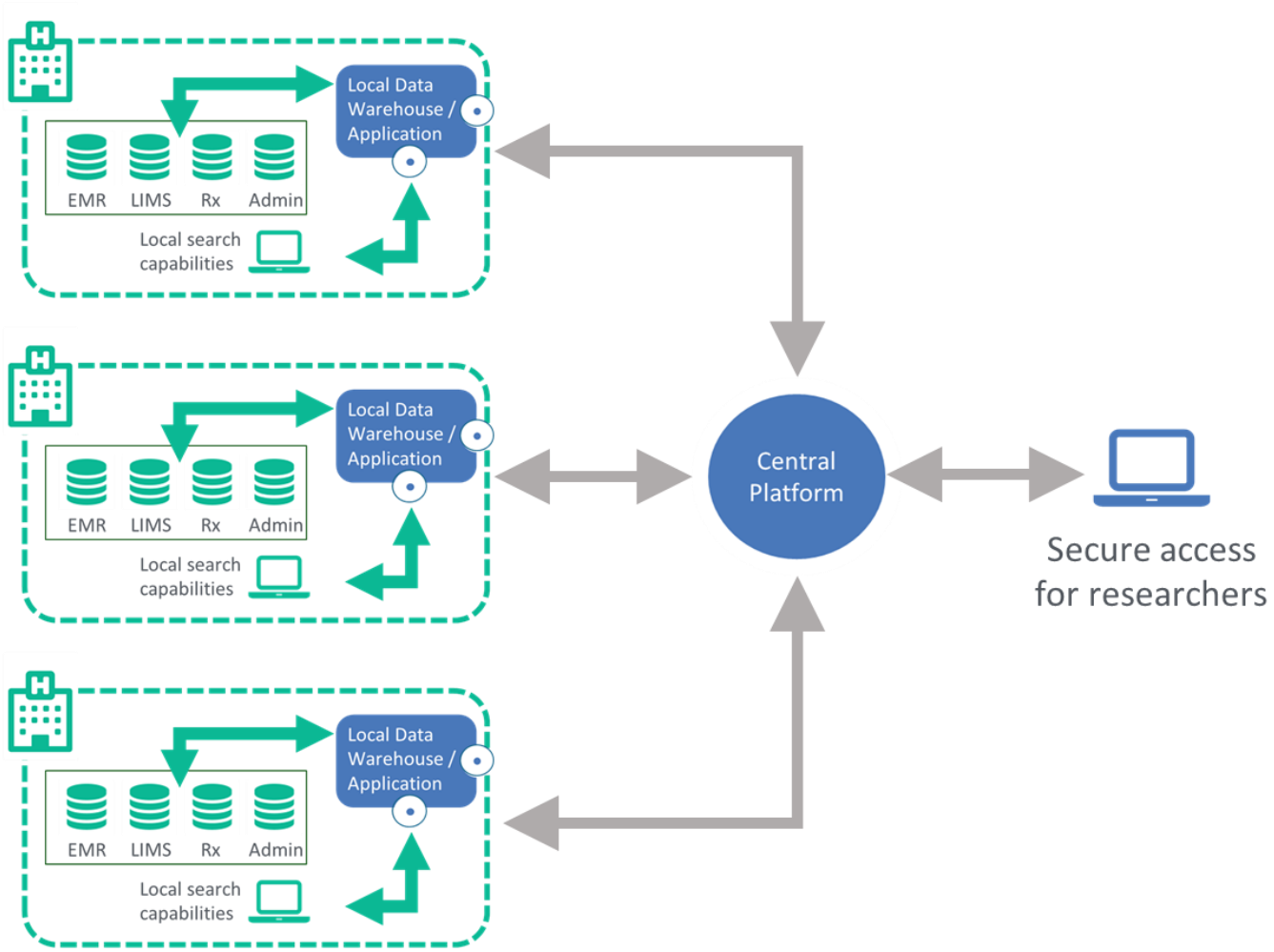
~~Bring data to the algorithms~~

Bring algorithms to the data

Maximize the potential of multiple datasets without sharing data

Minimize data leaks and privacy risks

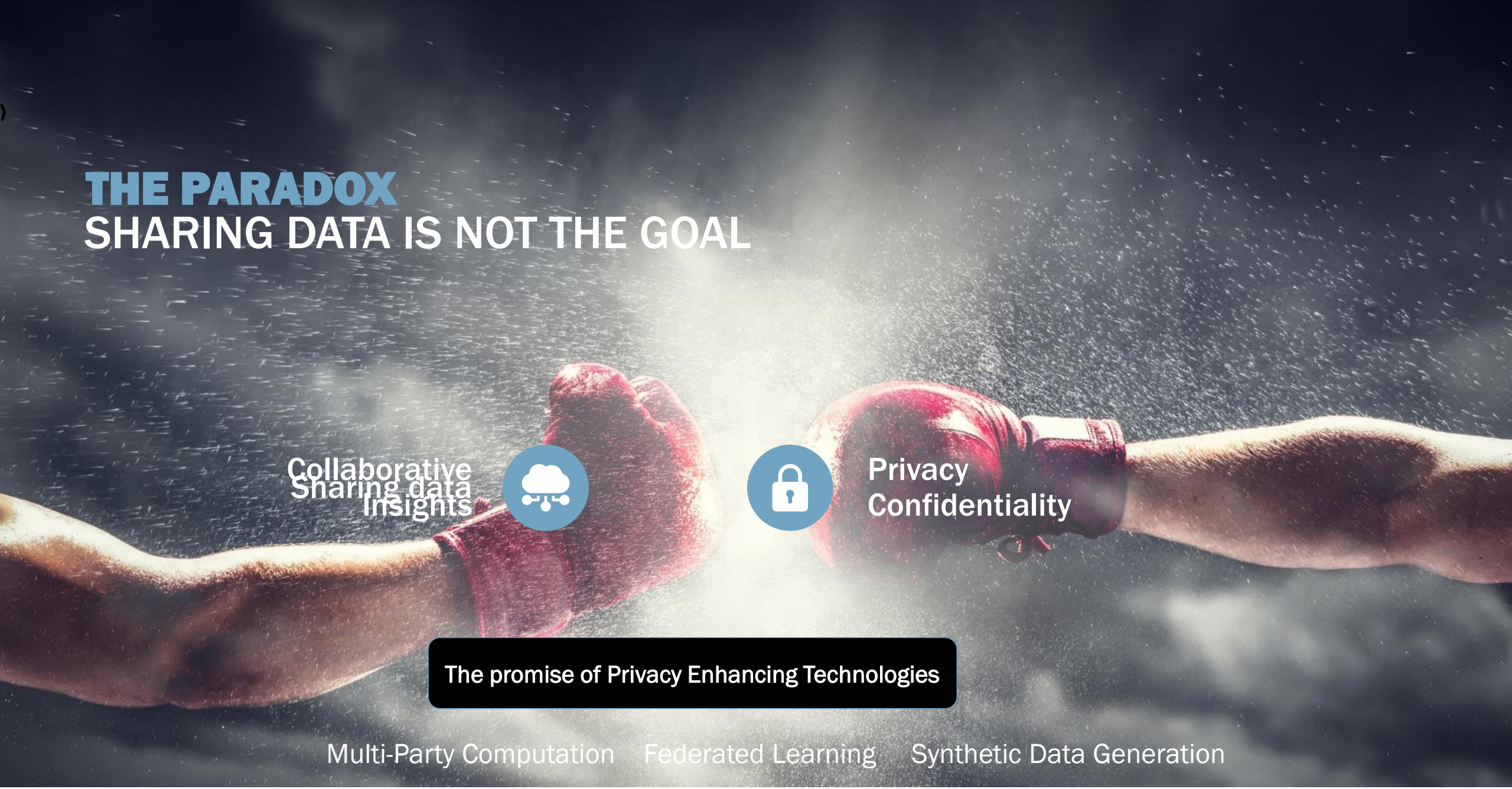
WHAT ARE FEDERATED DATA NETWORKS?



Benefits of federated networks

- Data remains under the control of the data owner
- Locally required legal and ethical approvals apply
- No patient level data leaves the owner's site, only aggregated counts, thereby ensuring patient privacy
- GDPR – *'Privacy by Design'*
- Analysis is "brought to the data" rather than creating central data repository
- Use of common data model allows for efficient search / analysis across multiple data sets
- Requires close collaboration with data owners which builds trust

THE PARADOX
SHARING DATA IS NOT THE GOAL

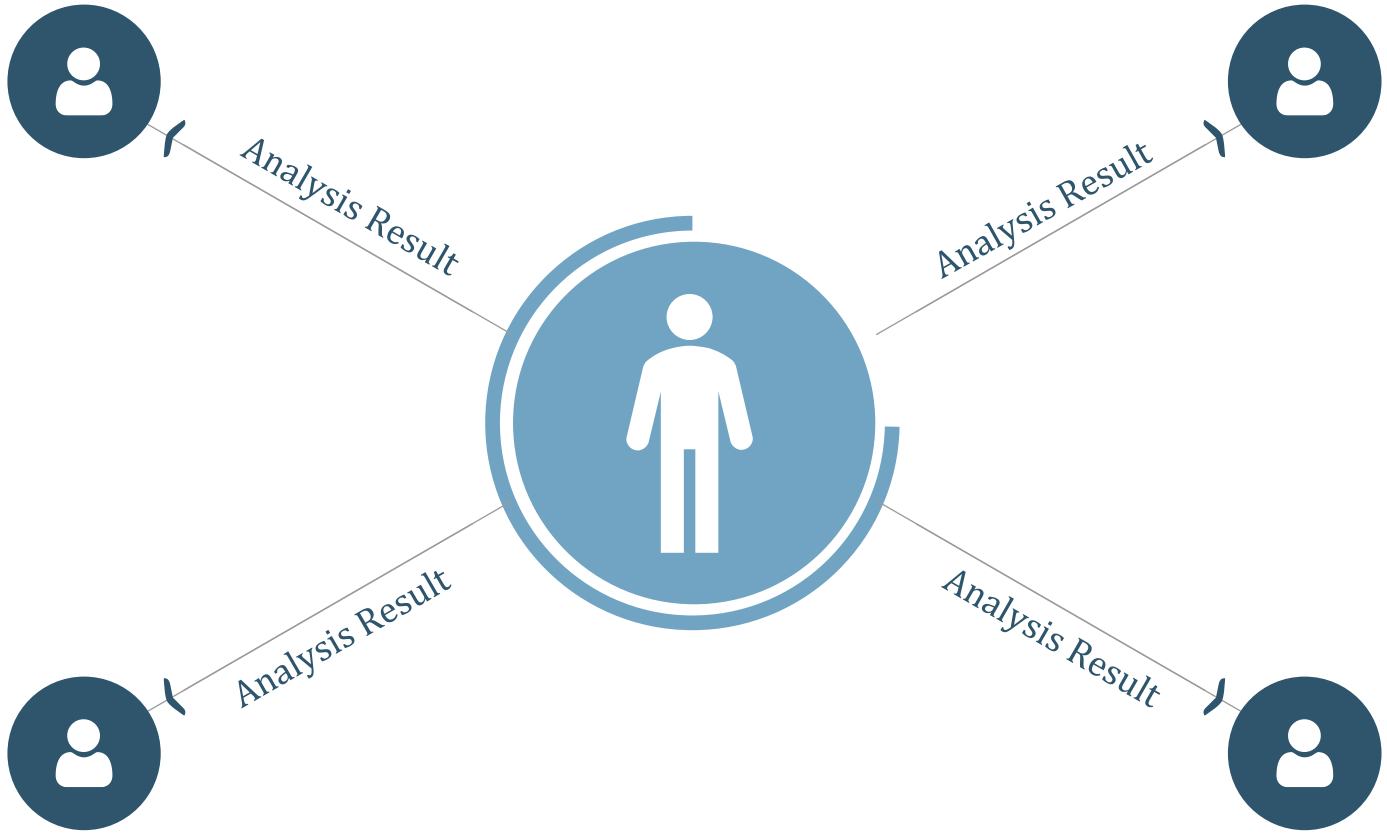


WHAT IS MULTI-PARTY COMPUTATION (MPC)?

INSTEAD OF THE TRADITIONAL SOLUTION (SHARING DATA)...

WHAT IS MULTI-PARTY COMPUTATION (MPC)?

INSTEAD OF THE TRADITIONAL SOLUTION (SHARING DATA)...

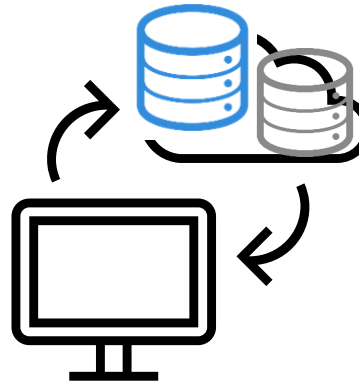


WHAT IS MULTI-PARTY COMPUTATION (MPC)?
A DECENTRALIZED SOLUTION BASED ON CRYPTOGRAPHY

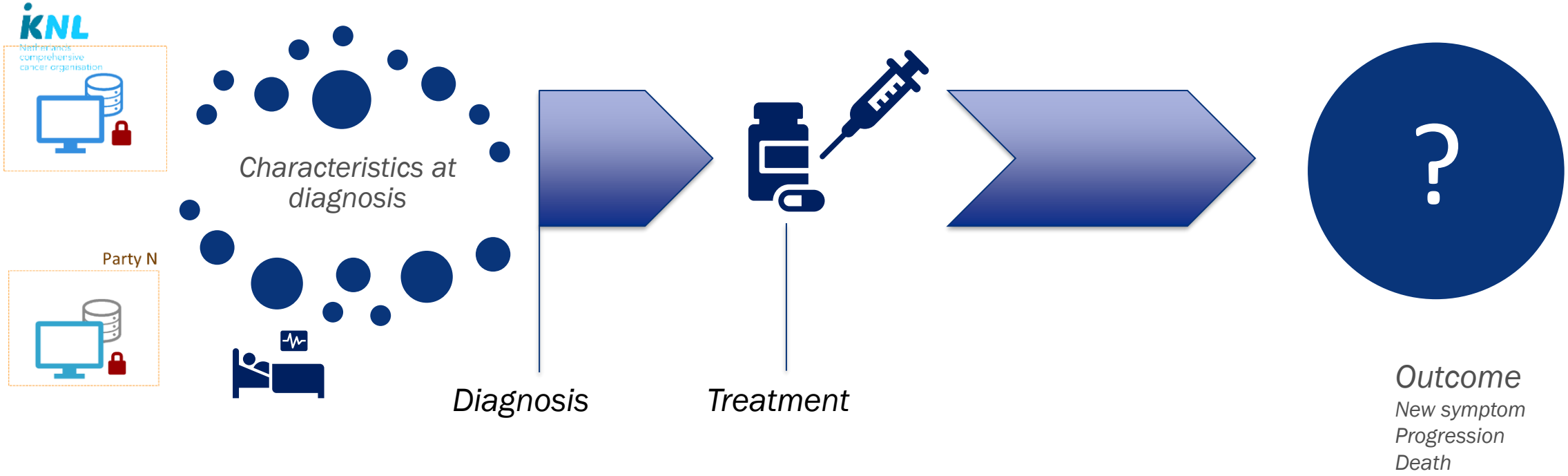
Joint computation on sensitive data using cryptographic techniques:

- Collaborative insight is obtained
- Underlying data is not revealed
- Security by design

LANCELOT: NON-SMALL CELL LUNG CANCER USE CASE



Predict **occurrence of long-term endpoints** such as **death** or **progression** of the disease from variables characterizing the patient cohort.

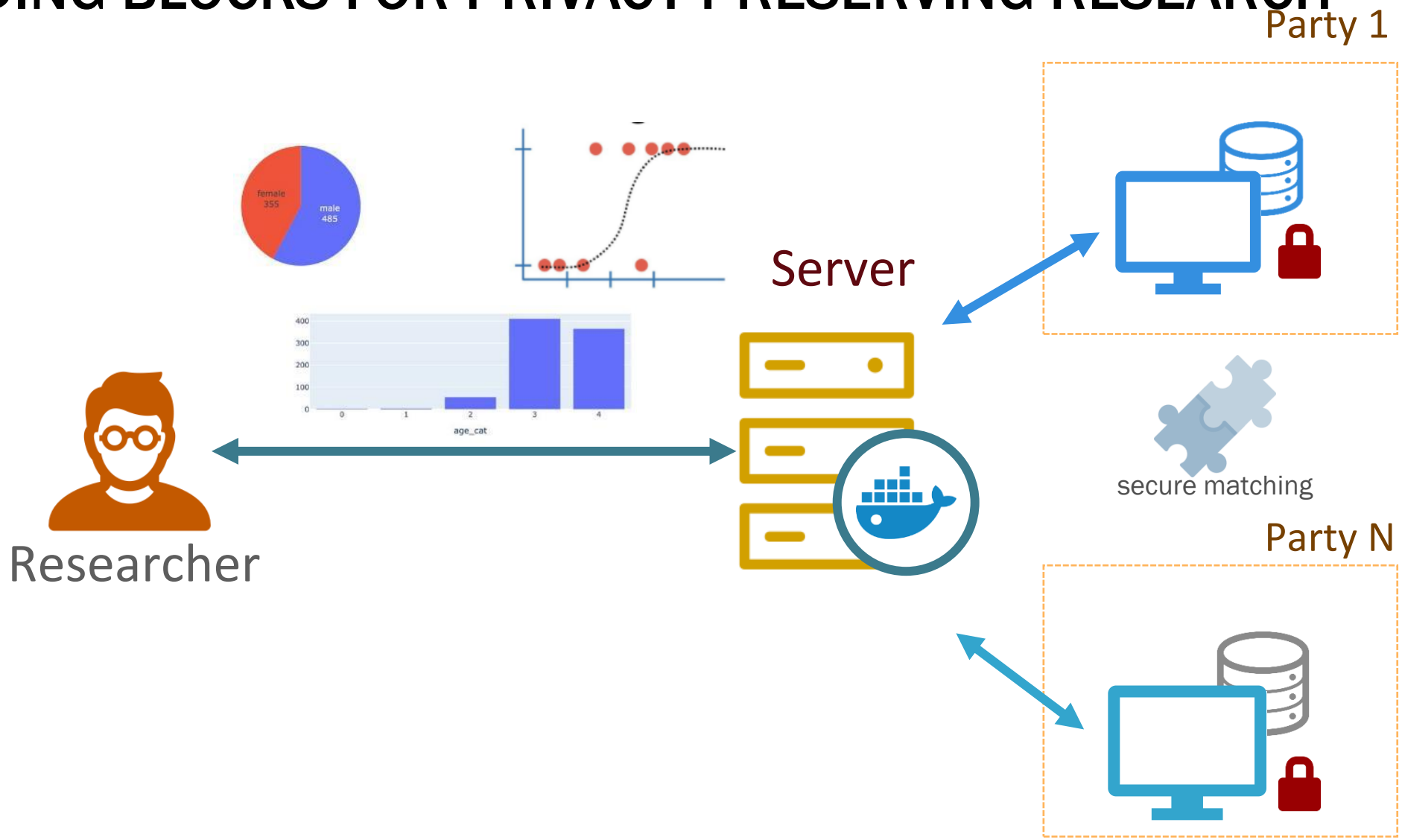


Data on demographics, disease, treatment, survival, distributed at different organisations.

Realistic synthetic data generated resembling real patient data.

LANCELOT: ANALYSIS OF DISTRIBUTED DATA

BUILDING BLOCKS FOR PRIVACY-PRESERVING RESEARCH



LANCELOT: SECURE MATCHING

- Data is **vertically partitioned** and sensitive
- The different data sets need to be matched
  - Without unique (patient) identifier
  - With (typing) mistakes
  - Without seeing the other data sets

Identifier	Feature A
Harry	Male
Alice	Female
Susan	Female
Bob	Male
Olivia	Female

Identifier	Feature B
Harry	32
Alice	25
Susan	53
Bob	64
Olivia	42

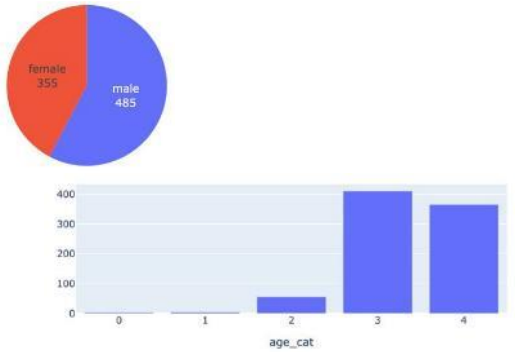
Vertically partitioned data

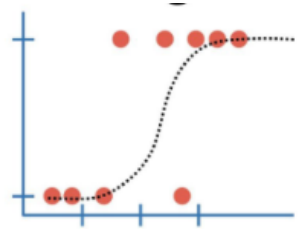
- We developed a secure solution for approximate matching, published it open source
  - Matching patients based on different features
  - Taking a subset of (typing) mistakes into account
  - https://github.com/TNO-MPC/protocols.secure_inner_join

- Testing on real data show that high percentage of real matches are detected.

LANCELOT: SECURE ANALYSES

- Only analysis outcomes should be revealed to anyone:
  - Aggregate statistics
  - Trained logistic regression model





- New open source MPC solutions for secure statistics and secure logistic regression
  - https://github.com/TNO-MPC/mpyc.statistics
  - https://github.com/TNO-MPC/mpyc.secure_learning

- Computation time: within hours (for 50.000 patients). Scales linearly in number of patients.

- Accuracy comparable to non-secure version.

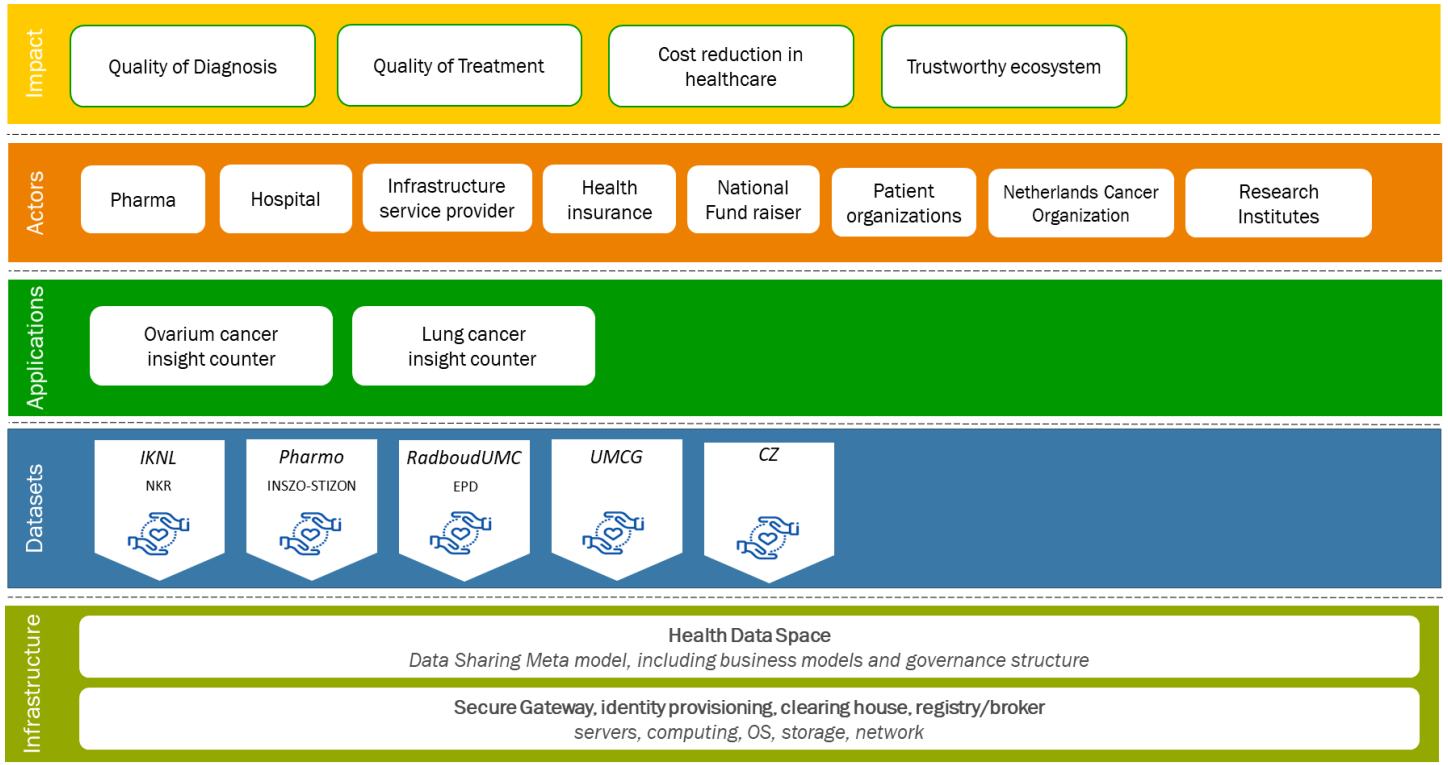
IMPORTANT FUTURE CHALLENGES

- Increasing scalability and applicability (faster, more algorithms, generic, easier usage)
  - Combining and integrating different technologies
- Legal questions
  - Jurisprudence needed for such new technologies
  - These solutions contribute to GDPR principles such as data minimization, proportionality, data protection
- Data quality essential (FAIR)
  - Garbage in = garbage out
- Data governance frameworks needed
  - Incorporating legal, organisational and ethical aspects with the technical solutions
- More testing on real data (pilots) needed

NEW INITIATIVE: HERACLES PROJECT (2022-2024)

RESEARCH PROJECT WITH MANY DIFFERENT ORGANIZATIONS

Towards a privacy-preserving and trustworthy data infrastructure enabling new insights and models from federated datasets, resulting in smart applications that actors can use to generate impact in the health domain.



CONCLUSIONS AND NEXT STEPS

- Novel techniques allow data to be harnessed without compromising privacy and confidentially
  - Enables privacy-preserving analyses in the health domain to create value based health care
- LANCELOT developed new generic open source solutions enabling secure approximate matching, secure statistics and secure machine learning, tested on synthetic data.
- These solutions positively contribute to GDPR principles
- Follow-up in HERACLES program

TNO innovation for life

For more questions:

Daniël Worm
daniel.worm@tno.nl
+31621134584