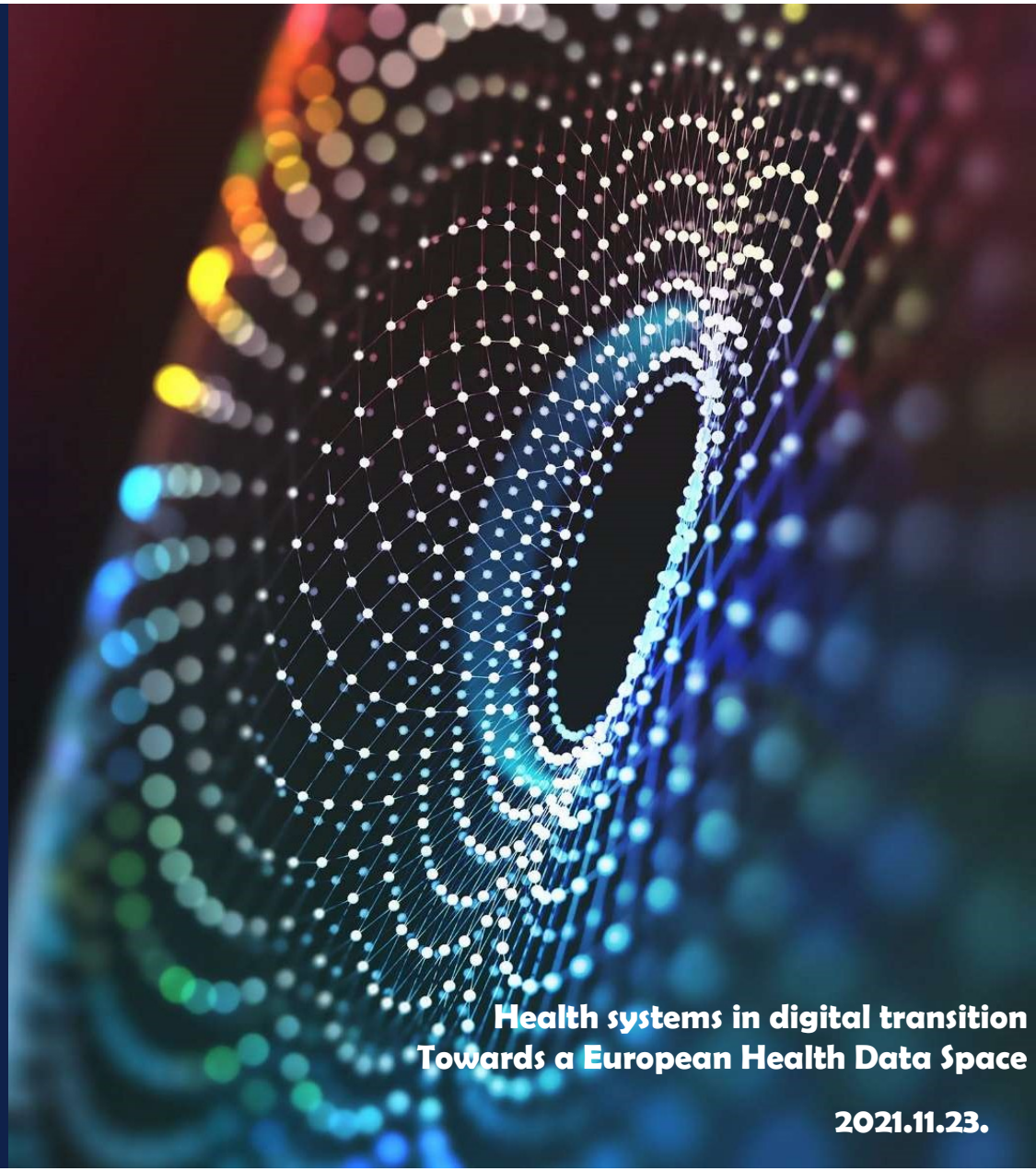
E-GROUP
SOFTWARE & BEYOND

Artificial intelligence and secondary use of health data

Data spaces for better healthcare research

Akos Tenyi, PhD
Head of Smart Data & Analytics Business Line
E- group Hungary



Health systems in digital transition
Towards a European Health Data Space

2021.11.23.


AI = everything (and more)
Increasing impact in developing healthcare

AI and Secondary Use of Health Data

AI in healthcare research

AI For Medical Diagnosis

- Disease vs no disease
- Classification of disease severity
- Imaging data – tumor vs no tumor

AI For Medical Prognosis

- Predict risks in the general population
- Risk of readmission, mortality…

AI in healthcare

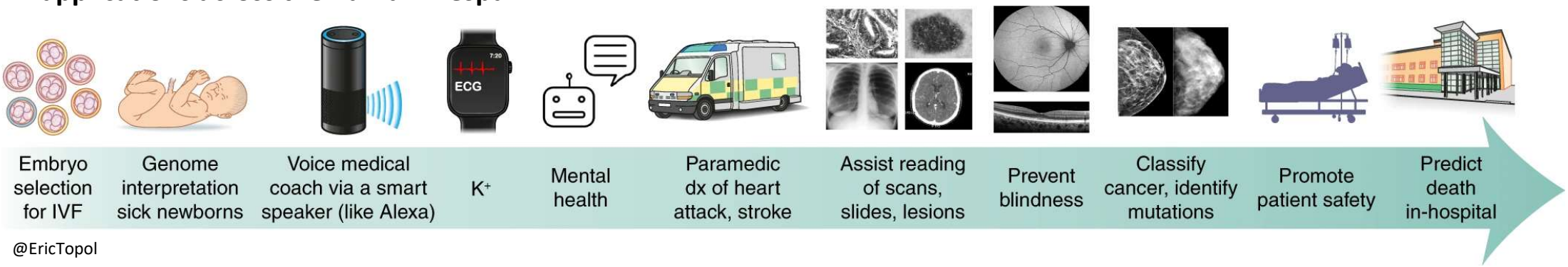
AI For Medical Treatment

- Estimate treatment effects
- Information extraction from medical reports

AI for Healthcare Optimisation

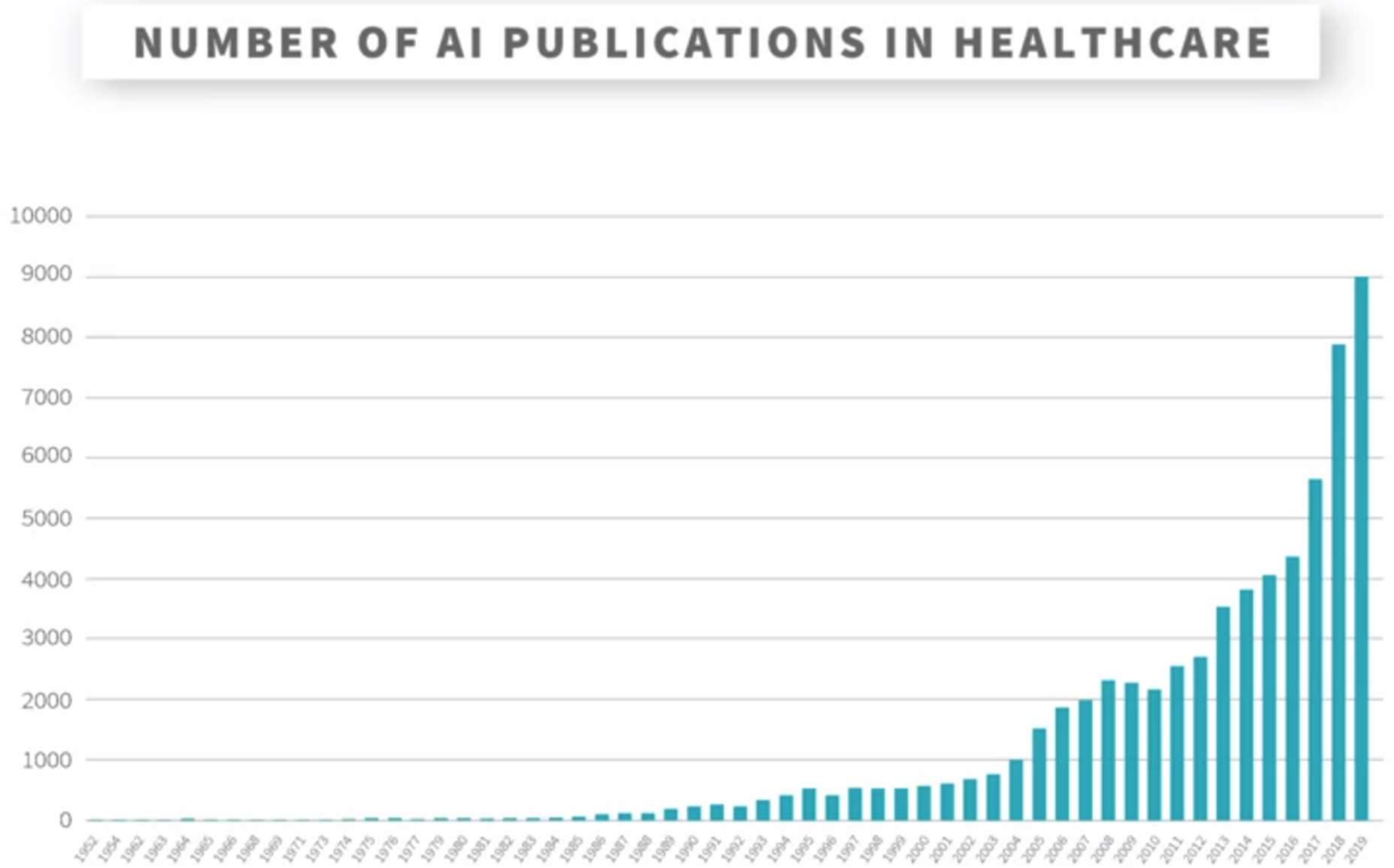
- Optimise waiting times
- Intelligent assistants

AI applications across the human lifespan

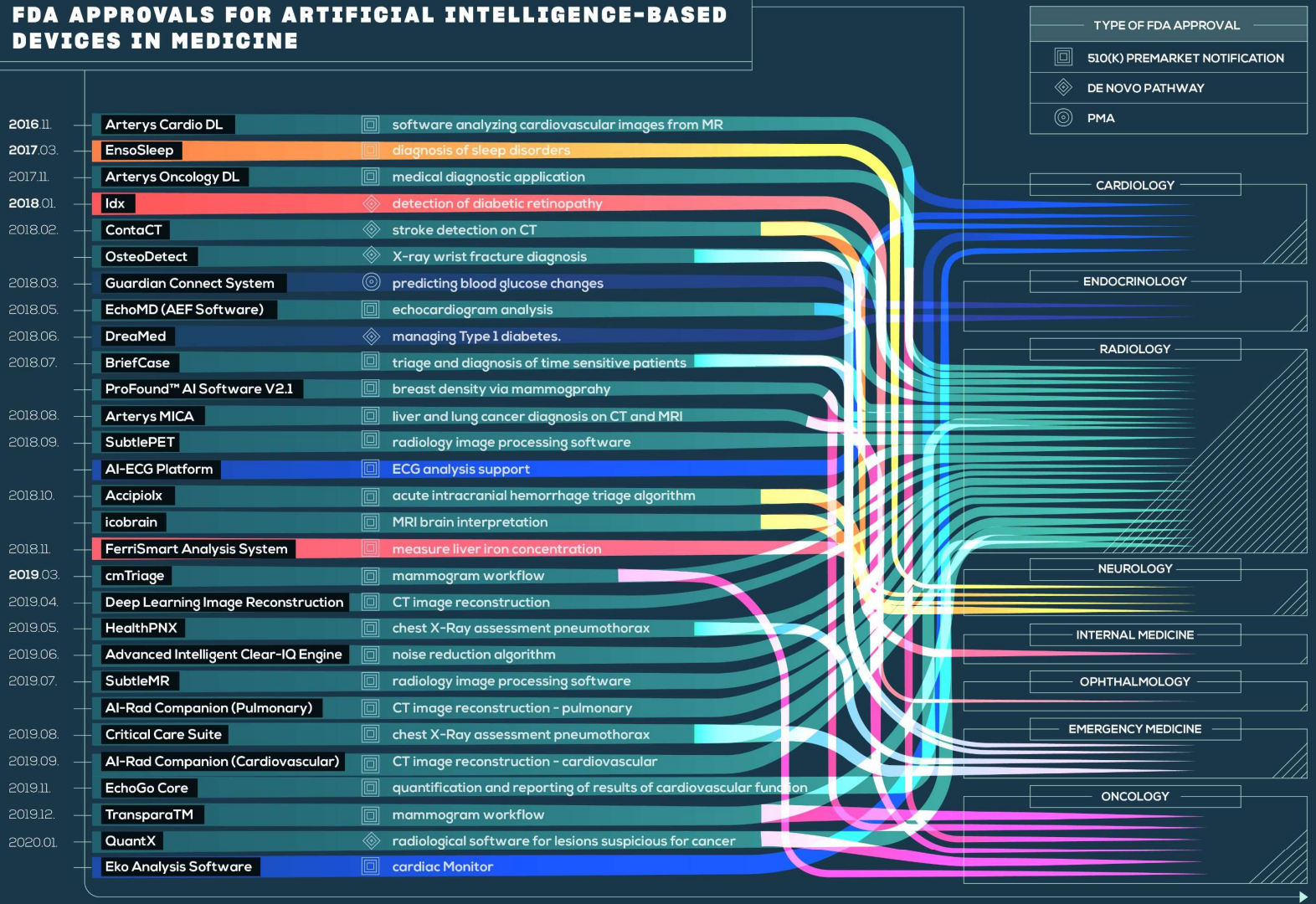


@EricTopol

AI and Secondary Use of Health Data



FDA APPROVALS FOR ARTIFICIAL INTELLIGENCE-BASED DEVICES IN MEDICINE



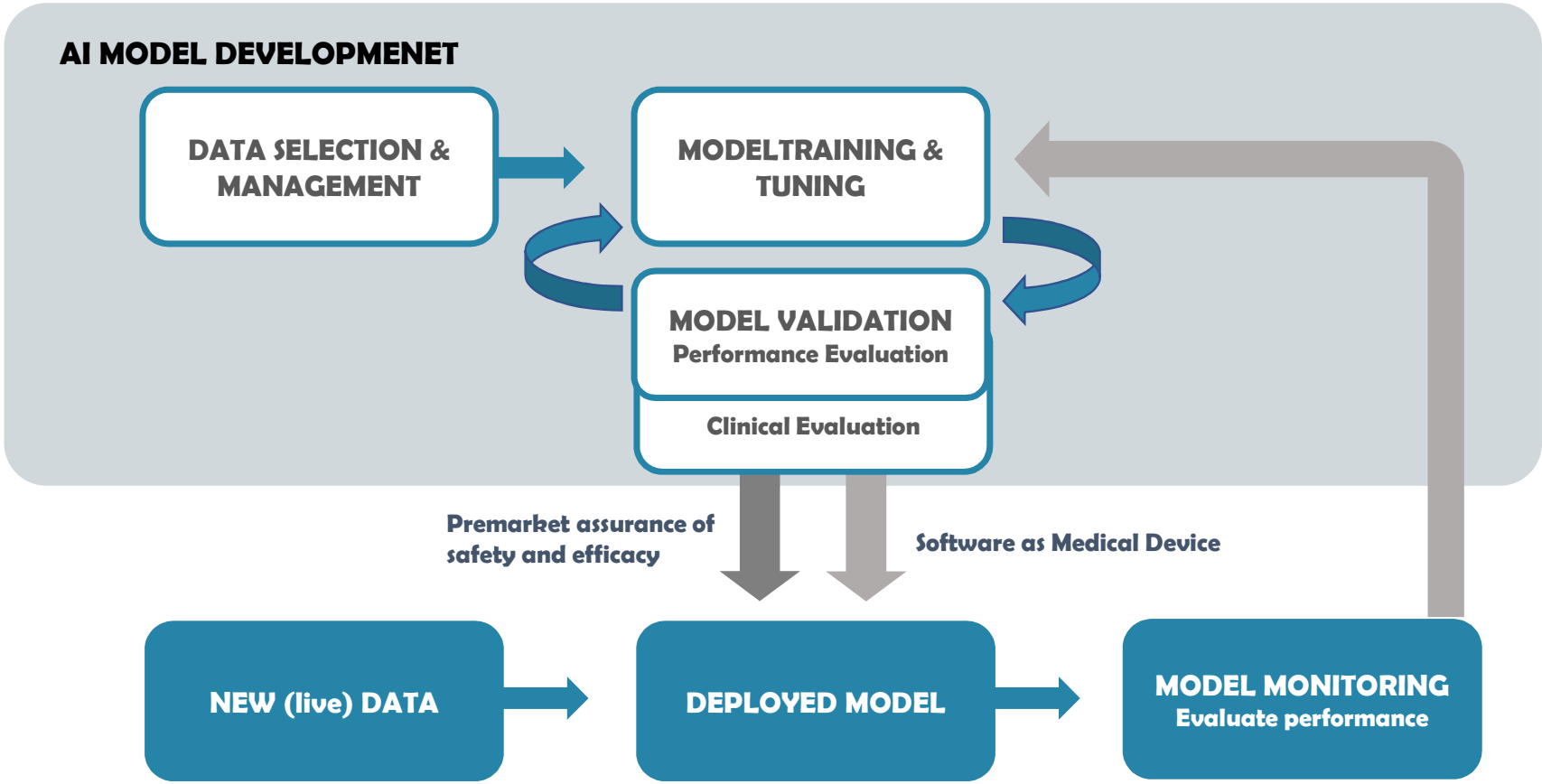
Around 30 FDA approved AI tools

Vs

~14.000 ICD9 codes

TMF
THE MEDICAL FUTURIST

AI PRODUCT LIFE CYCLE



AI and Secondary Use of Health Data

AI PRODUCT LIFE CYCLE

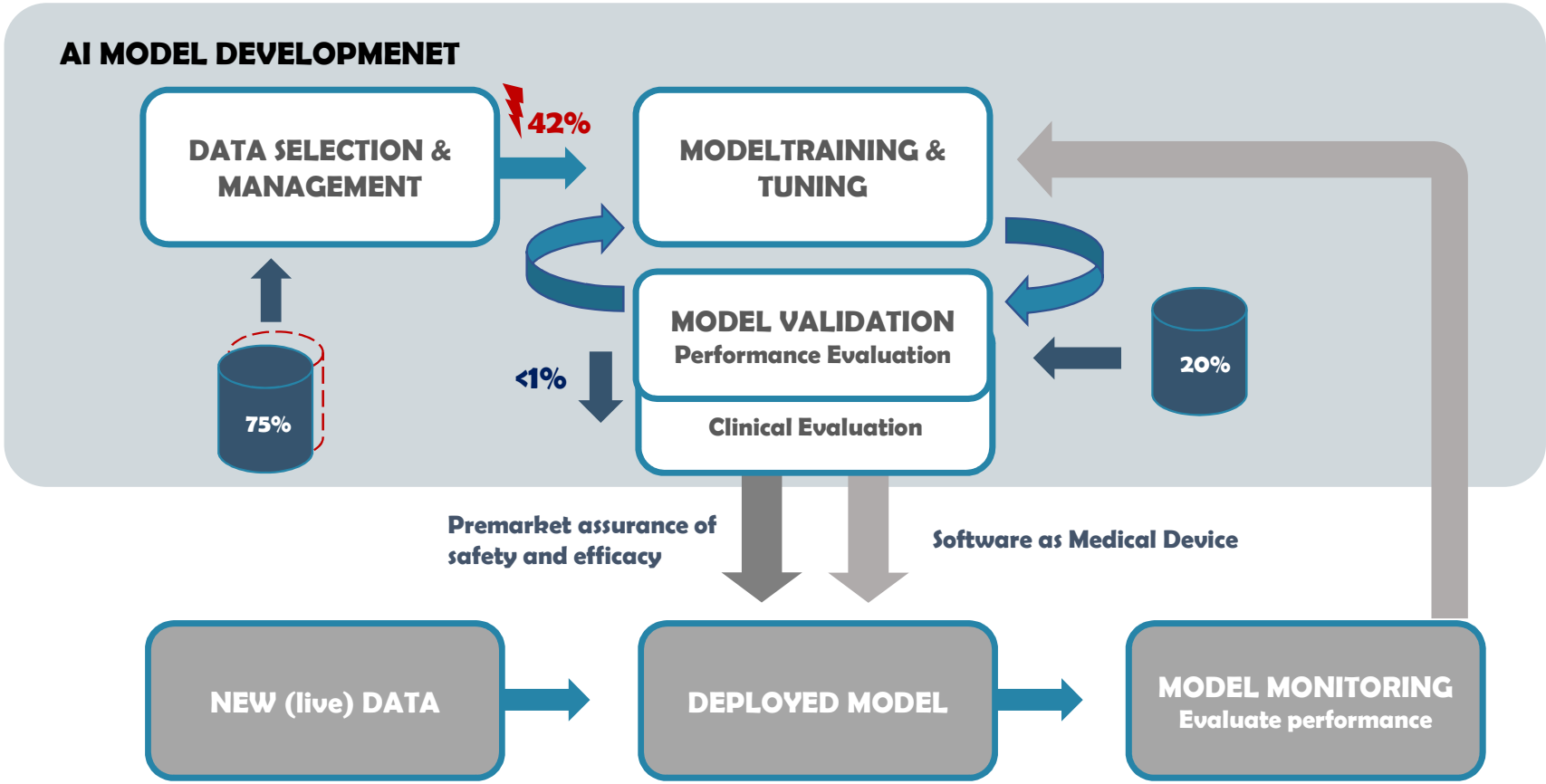
COVID19 as a use case

232 COVID19 models
- **Risk prediction models**
- **Diagnostic models**
- **Prognostic models**

Models are at high risk of bias

Main source of bias

- **Non-representative population (42%)**
- **Modest sample size**
- **Lacking (independent) validation (80%, 22%)**
- **Technical errors in analysis**



L. Wynants *et al.*, "Prediction models for diagnosis and prognosis of covid-19: systematic review and critical appraisal," *BMJ*, p. m1328, 2020.

Real world performance monitoring

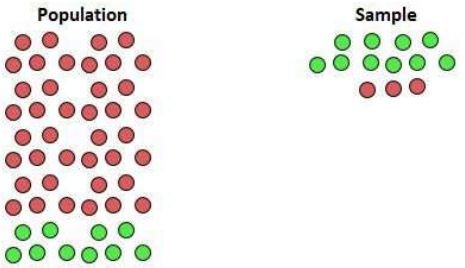
Characteristics of 'Bad Data' for Machine Learning

1. Insufficient Quantity of Training Data

- "these results suggest that we may want to reconsider the tradeoff between spending time and money on algorithm development versus spending it on corpus development." (Microsoft, 2001)

2. Non-representative Training Data

- In statistics, sampling bias is a bias in which a sample is collected in such a way that some members of the intended population have a lower or higher sampling probability than others.
- AI models often perform poorly on populations that are not represented in the training data. It is critical for AI training data to mirror the populations for which model are ultimately serving.
- E.g. image-based diagnostic task in 2019 71% used a patient cohort from one of three states: California, Massachusetts or New York. Thirty four states did not contribute data, point to huge patient underrepresentation.



3. Poor-Quality Data

4. Irrelevant Features

ACCESS TO HEALTH DATA FOR RESEARCH

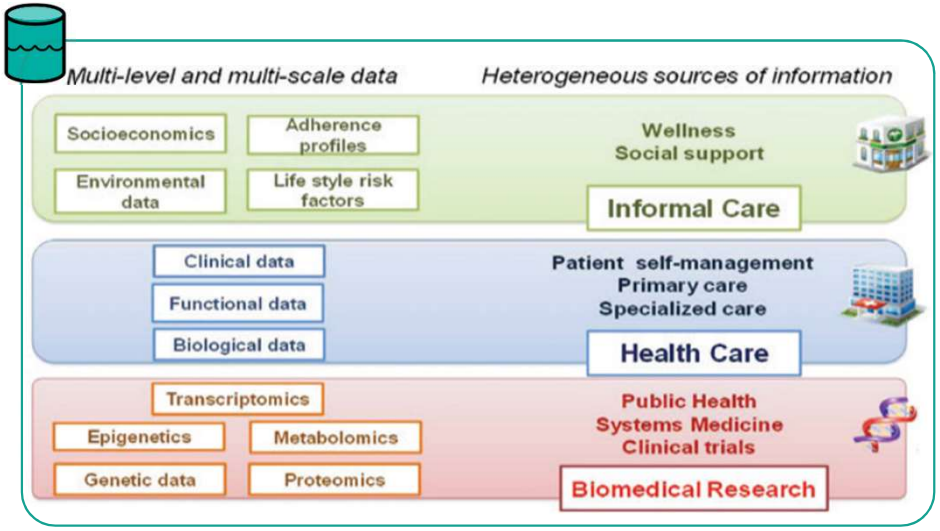
IN PRACTICE

- Increase reuse of health data
- Decrease fragmentation of healthcare data

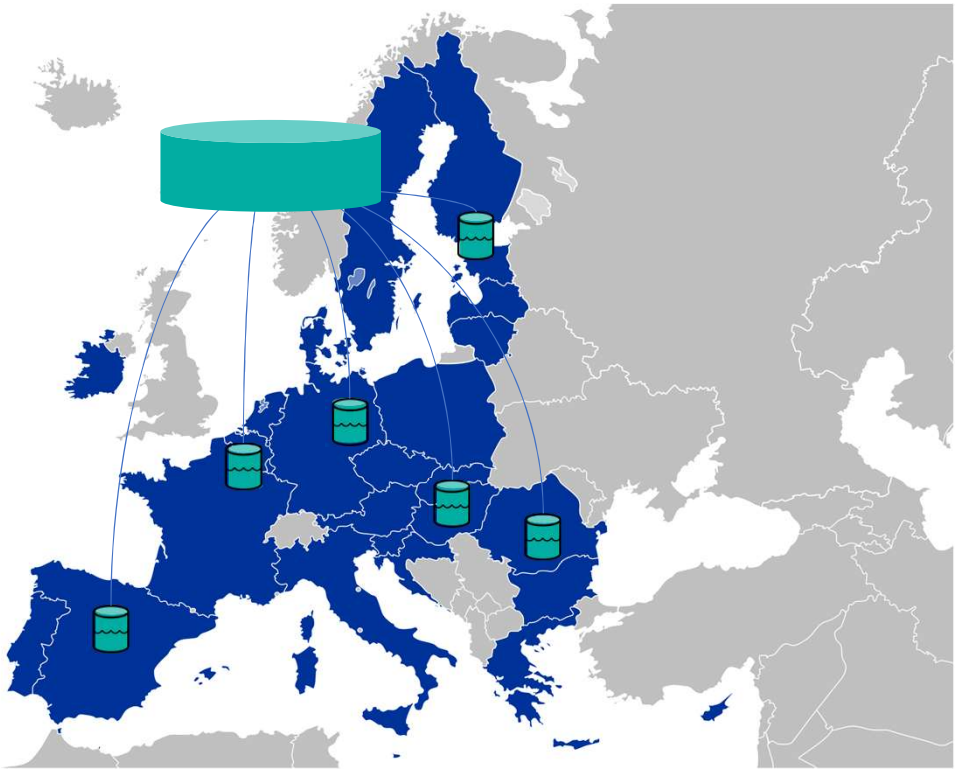
AI and Secondary Use of Health Data

Healthcare data integration

Vertical data integration



Horizontal data integration



InnoHealth DATALAKE
E-Group & University of Pécs

DLX

Roca, J., Tenyi, A., & Cano, I. (2019). Paradigm changes for diagnosis: using big data for prediction. *Clinical Chemistry and Laboratory Medicine (CCLM)*, *57*(3), 317–327.

Problem with current AI solutions

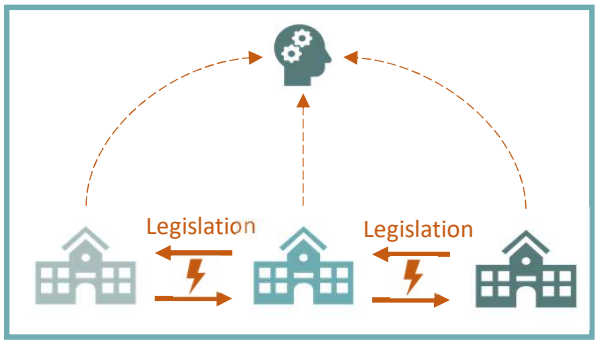
Assumption		Reality in healthcare
Centralized dataset	⇄	*Different data owners*
Homogenous data		*Heterogenous data*

With conventional approaches we need to...

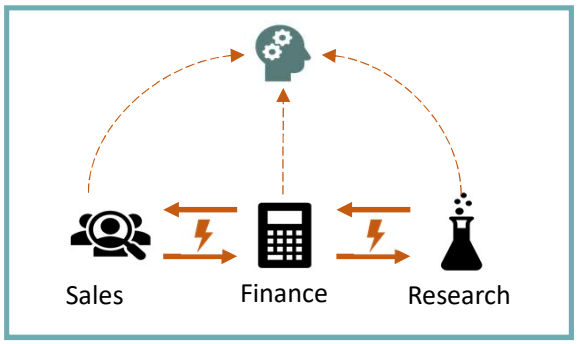
Move the data from the origin place

Solve privacy issues

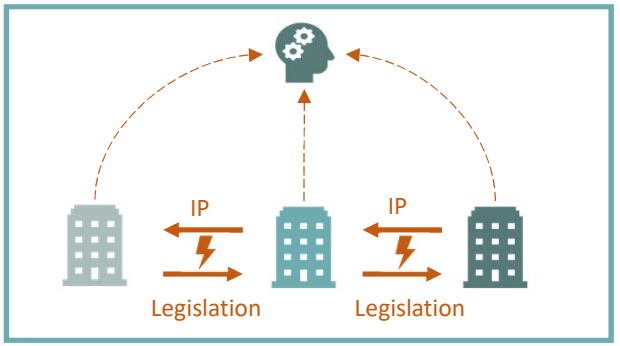
Cross-border



Cross-department



Cross-enterprise

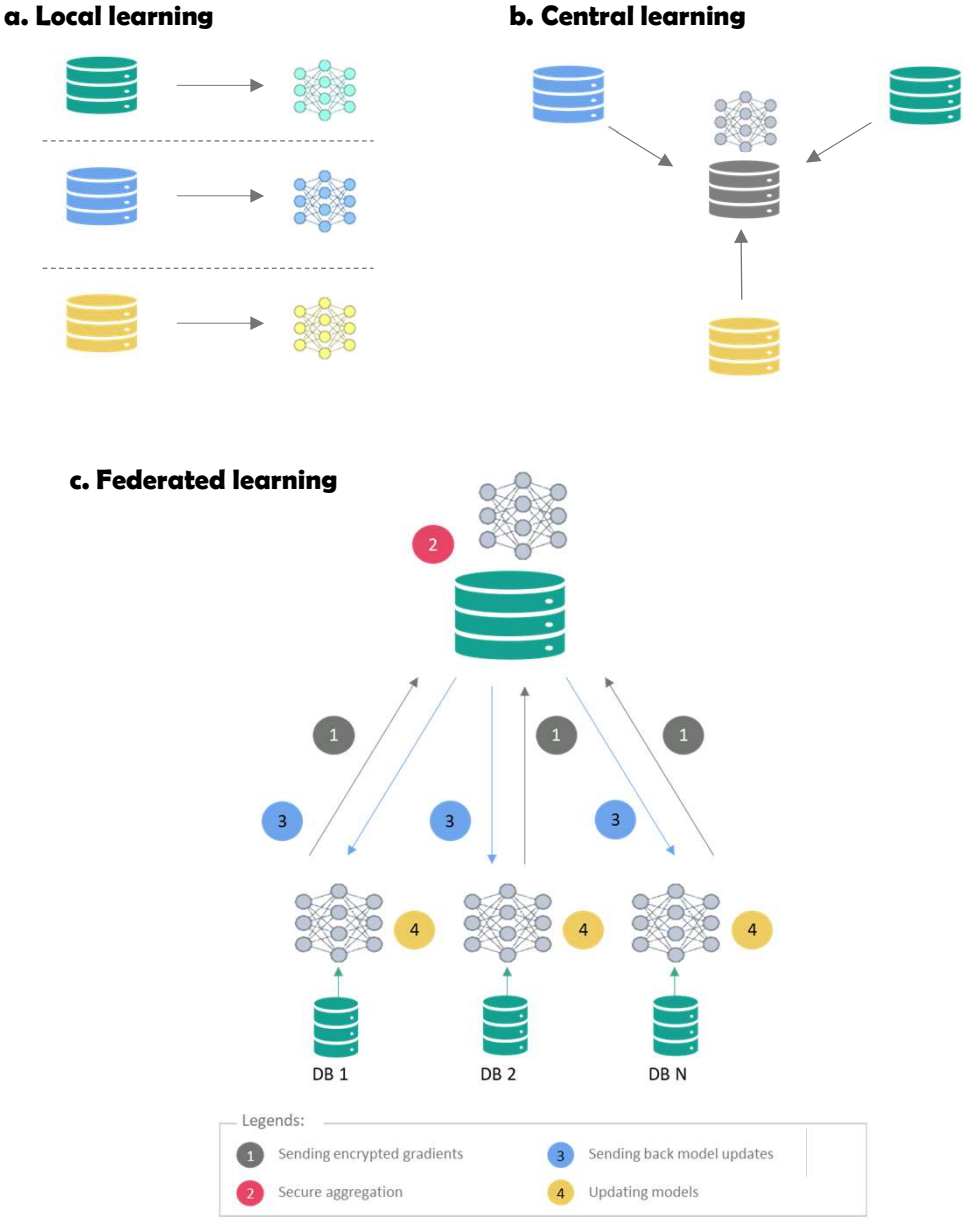


AI and Secondary Use of Health Data

FEDERATED LEARNING
Brings Analytics To The Data

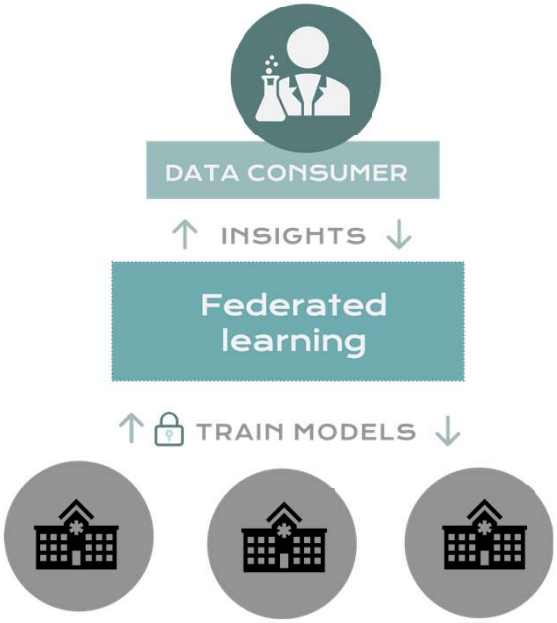
Concept Technique which trains machine learning algorithms across multiple decentralized servers

- Solution **brings the code to the data,** despite conventional machine learning which brings the data to the code
- **No need for centralized data** from a centralized storage (local disc, cloud, etc.)
- Private data can be kept in the origin place at different data holder
- Local model **privacy preserving**
- Secure and fast access to data

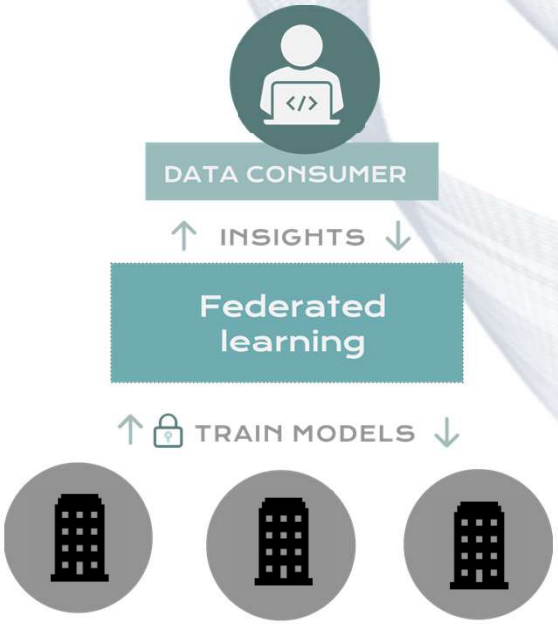


AI and Secondary Use of Health Data

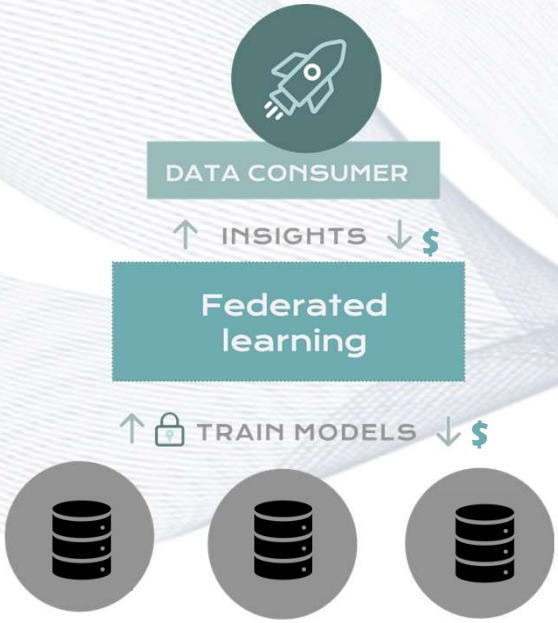
Federated learning business functionalities



...rare disease (RD) centers create a data coalition for analysing distributed sensitive RD registries using artificial intelligence methods.

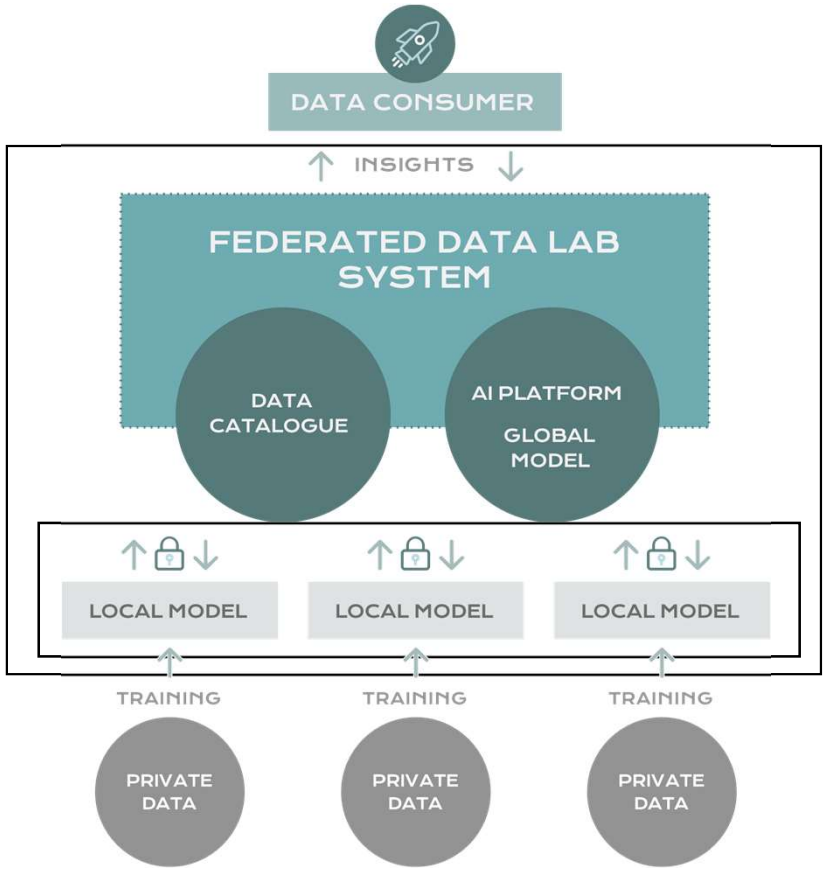


...federated network of national biobanks to integrate divided genomic data assets for analysis and to increase impact and utility.

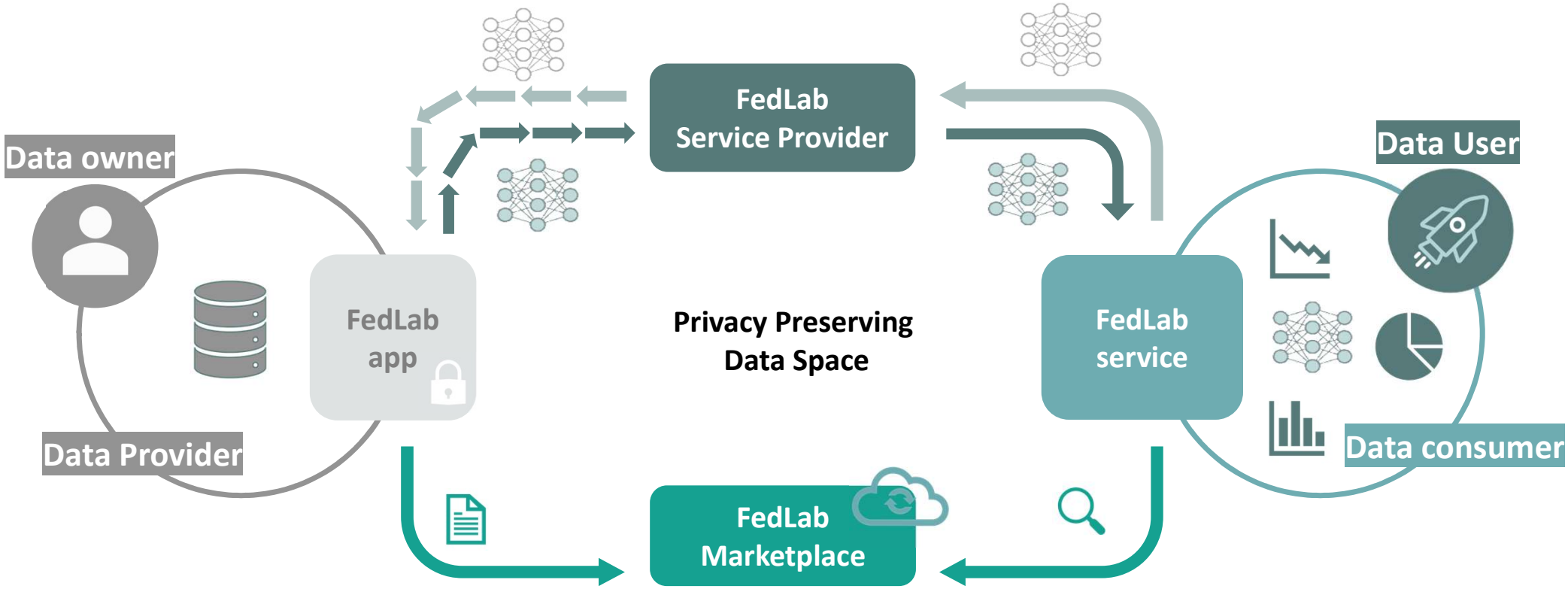


...federated data market to access sensitive health data for AI model validation.

AI and Secondary Use of Health Data



Federated learning business functionalities



Let's build together dataspaces!

https://www.egroup.hu/

akos.tenyi@egroup.hu