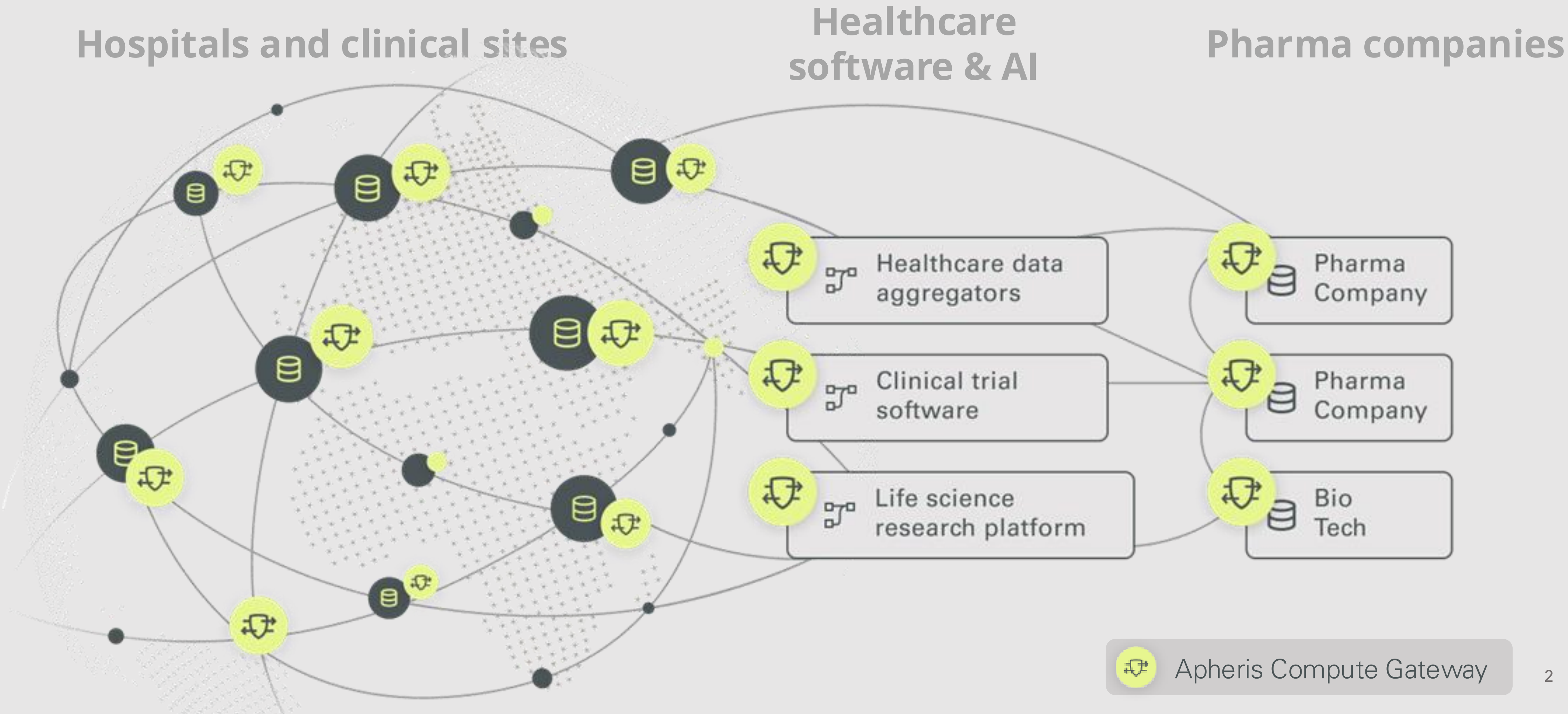
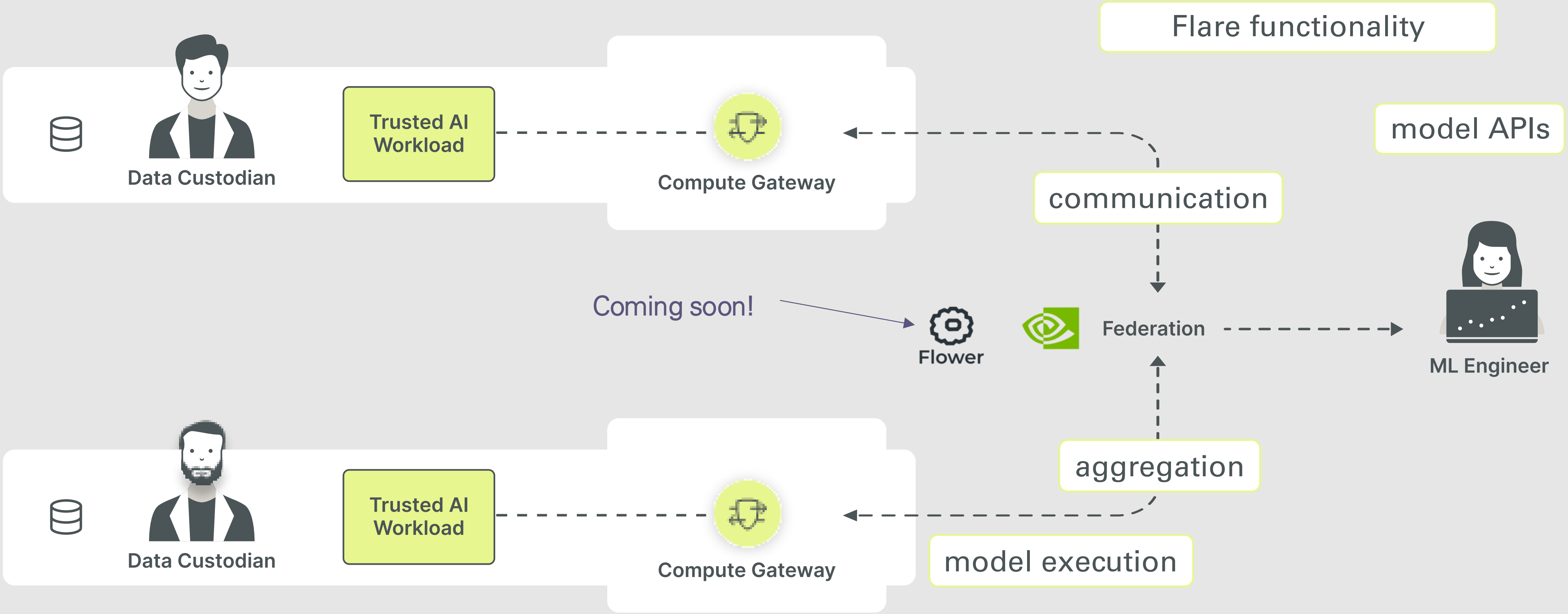
Federated data collaboration in BioPharma leveraging NVIDIA FLARE and BioNeMo

Ellie Dobson, Robin Roehm
Apheris

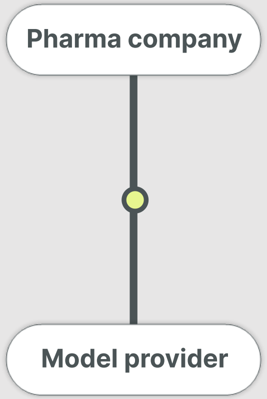
First, a few words about Apheris.
We safely connect models and data



We use NVFlare (and soon Flower-on-Flare) for a strong foundation on the federation side

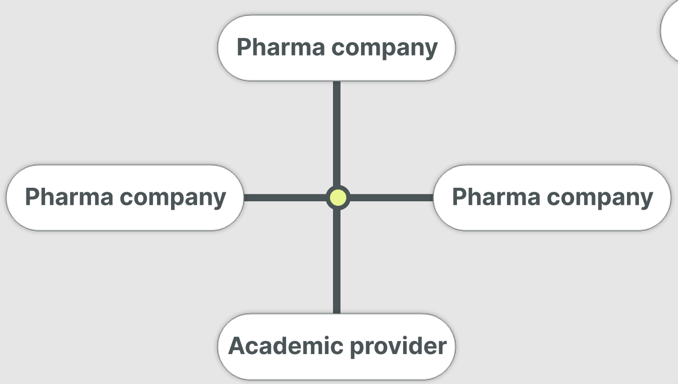


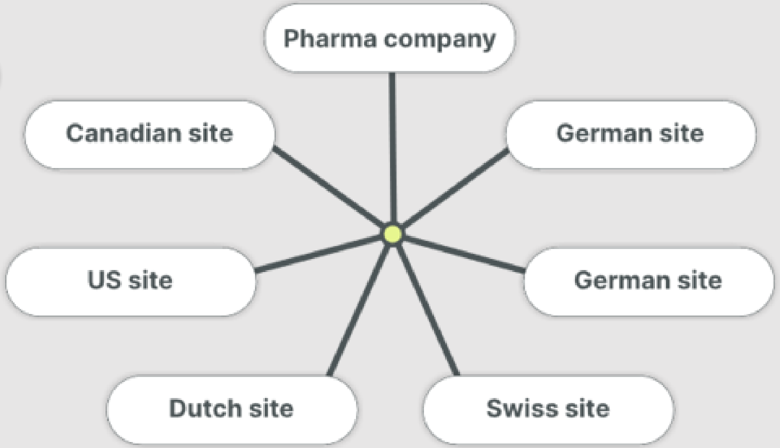
But federation alone cannot solve the issue of trust in collaborative setups



Return on investment is hard to assess upfront

Maintaining compliance slows things down





Privacy is a complex beast for deep learning

How to protect and separate both model and data IP?

https://doi.org/10.1038/s42256-024-00813-x

Federated learning is not a cure-all for data ethics

Marieke Bak, Vince I. Madai, Leo Anthony Celi, Georgios A. Kaissis, Ronald Cornet, Menno Maris, Daniel Rueckert, Alena Buyx & Stuart McLennan

Check for updates

Although federated learning is often seen as a promising solution to allow AI innovation while addressing privacy concerns, we argue that this technology does not fix all underlying data ethics concerns. Benefiting from federated learning in digital health requires acknowledgement of its limitations.

Large amounts of routinely collected health data are needed to facilitate
the development of machine learning (ML) and artificial intelligence
(AI) applications in med
protection regulations c
Federated learning (FL)
tion that can allow AI in
The term 'federated' re
a federation of particip
multiple independent c
data, and then combine
a central server. In this w
need to combine data in
FL has turned out to
using standard data for
ing[2,3]. However, FL is a t
technical consideration
tions. We are concerned
digital health, for examp
ing bodies, to use FL as a
sufficient acknowledge
tant ethical and practica
those deploying it. In thi
FL should be coupled w

TL;DR

- Is federation truly privacy preserving?
- Accuracy, bias and fairness
- Transparency and explainability

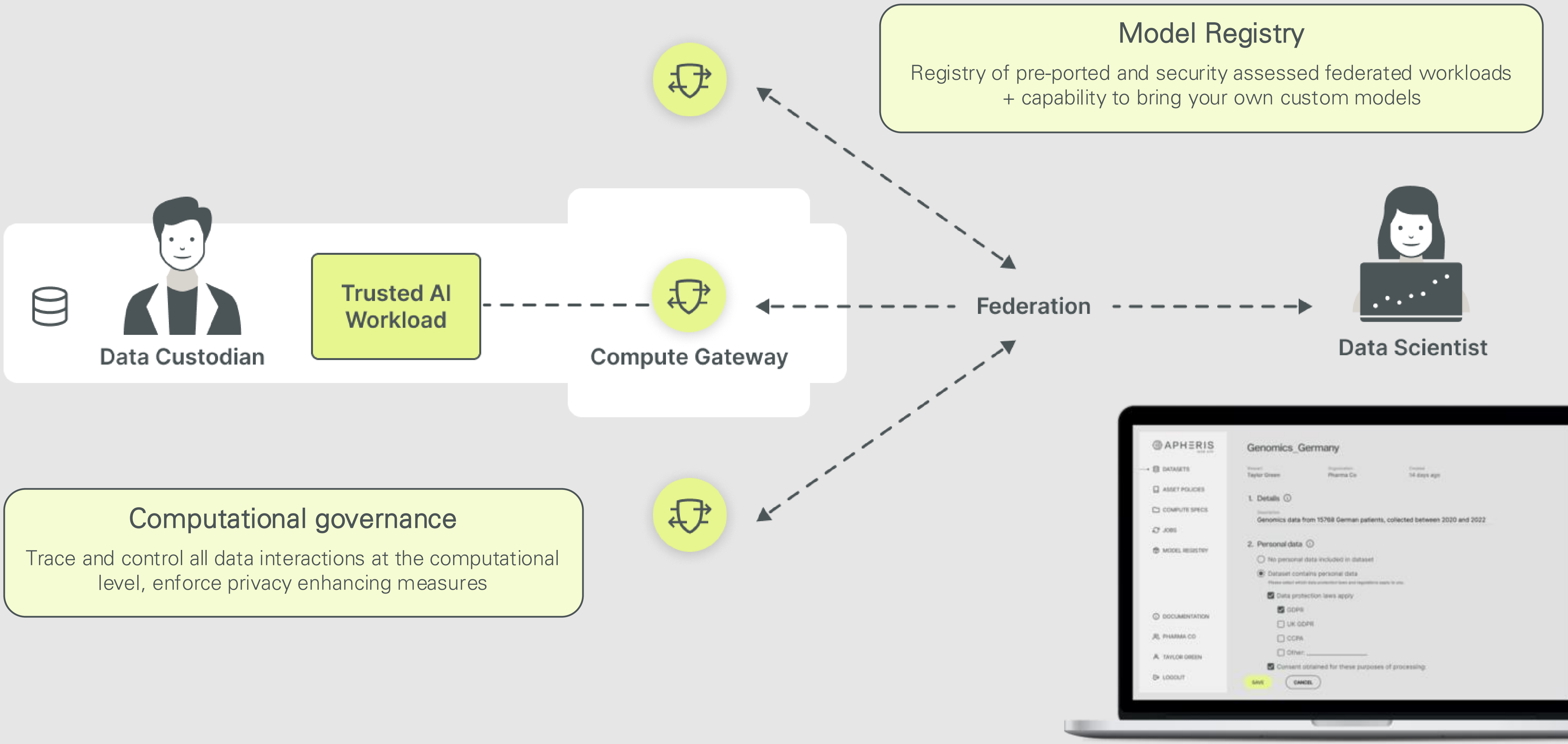
Truly privacy-preser
It is important to realize
The fact that sensitive data are not transferred to a central point in FL mitigates certain privacy concerns regarding data transfer and allows organizations to avoid administratively burdensome data transfer agreements, but it does not relieve the data holders from their data protection responsibilities[1]. Ethical and legal justifications for data collection and sharing are still needed (that is, patient consent or research exemptions), and data holders remain responsible for implementing governance and technical measures to securely protect the data locally on the nodes (for example, appropriate access control and encryption). Indeed, because this responsibility is spread out over multiple partners instead of being allocated to a central controller, the risk that no one takes proper responsibility for data governance is increased.

Furthermore, although FL removes the option of stealing the data from a central database, it does not give any formal guarantees against re-identification[4]. Numerous studies have shown that it is possible to reverse engineer privacy-sensitive data from the model information that is shared in FL (see, for example, ref. 5). Thus, the term 'privacy-preserving' is not entirely appropriate for FL. The bottom line is that an ML model (or its parameters/gradients) trained on private data must itself be considered private data. Therefore, approaches such as differential privacy (for example, adding noise locally) remain needed, yet such approaches may represent a trade-off between data privacy and the utility of the models[6]. This trade-off arises in cases in
hat is, not only
g to infer data-
ed in a fair and
nonetary cost.

by FL might be
nay be affected
at add noise to
to be acknowl-
acy can also be
rial actors aim-
uch anomalies
f the local data
ness of FL, and
r, the accuracy
s the different
practices and
ningful dataset
additional and
tiative is being

es can lead to
d fix algorith-
mic bias in models trained on pooled data, it is even more difficult when learning is distributed and each 'node' sees only its own data[6]. Bias is often caused by the under-representation of (minority) populations in the source data, or by limited access to specific variables for these groups. Although this, too, is not a problem unique to FL, it can in some cases be exacerbated by FL. For example, the FL training process can contribute to bias when the different local models are combined ('fused') using a weighted average: if larger datasets are weighed more highly during the fusion, bias in those datasets is amplified. Another issue arises when certain types of data are not used in the global model because only specific centers have them. If only the

We solve for trust between collaborating parties



https://build.nvidia.com/explore/healthcare

We benefit from the BioNeMo suite and its integration with Flare

DeepMind

NVIDIA

AlphaFold 2
Predicts the 3D structure of a protein from its amino acid sequence.

MolMM
Controlled generation, finding molecules with the right properties.

MegaMolBart
Molecular generator

DiffDock
Predicts the 3D structure of how a molecule interacts with a protein

Meta

ESM2
Generates embeddings of proteins from their amino acid sequences

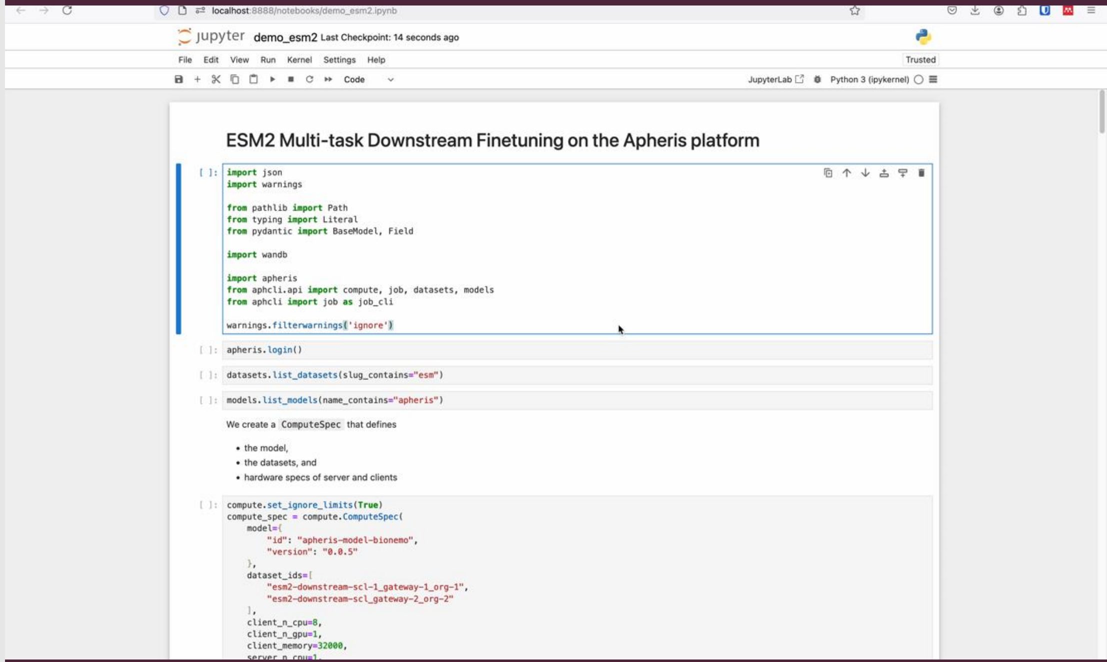
ESMFold
Derivative of ESM2; predicts the 3rd structure of a protein from it's amino acid sequence

ProteinMPNN
Predicts amino acid sequences for protein backbones

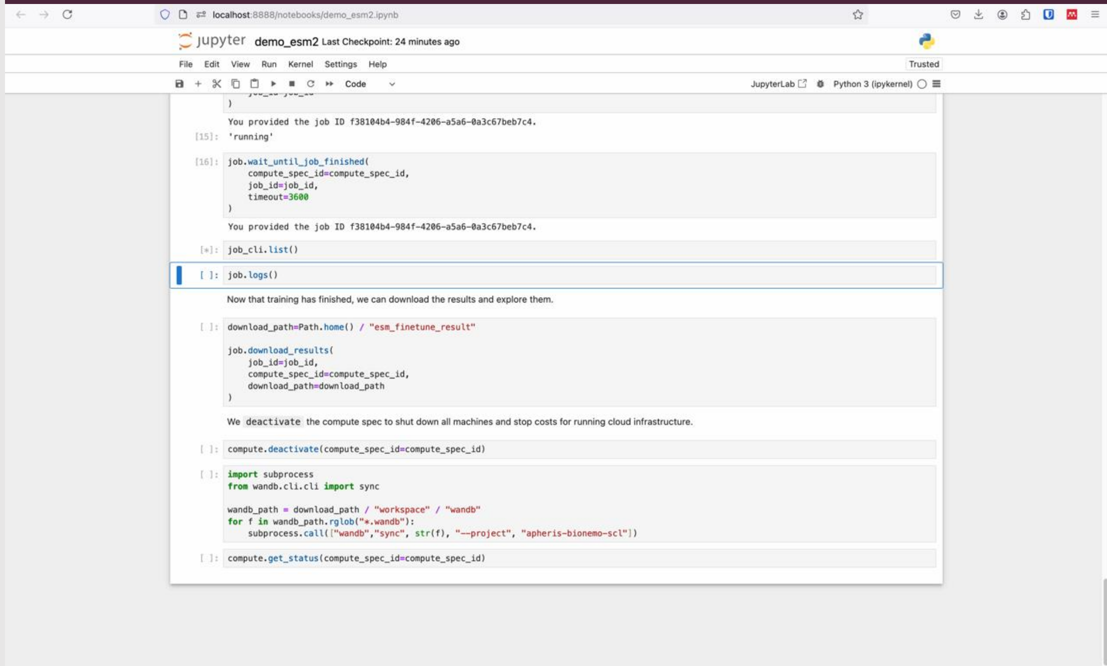
INSTITUTE FOR Protein Design UNIVERSITY of WASHINGTON

RoseTTAFold
Generates protein backbones for protein binder design

Job approval



Job execution



Powering the AISB Consortium to Revolutionize AI Drug Discovery

Apheris provides the tech layer for the Artificial Intelligence Structural Biology (AISB) Consortium, an unprecedented collaboration among AbbVie, Boehringer Ingelheim and Johnson & Johnson aimed at transforming AI drug discovery. State-of-the-art AI models will be trained and evaluated on unique data from multiple biopharma companies without exposing proprietary information.

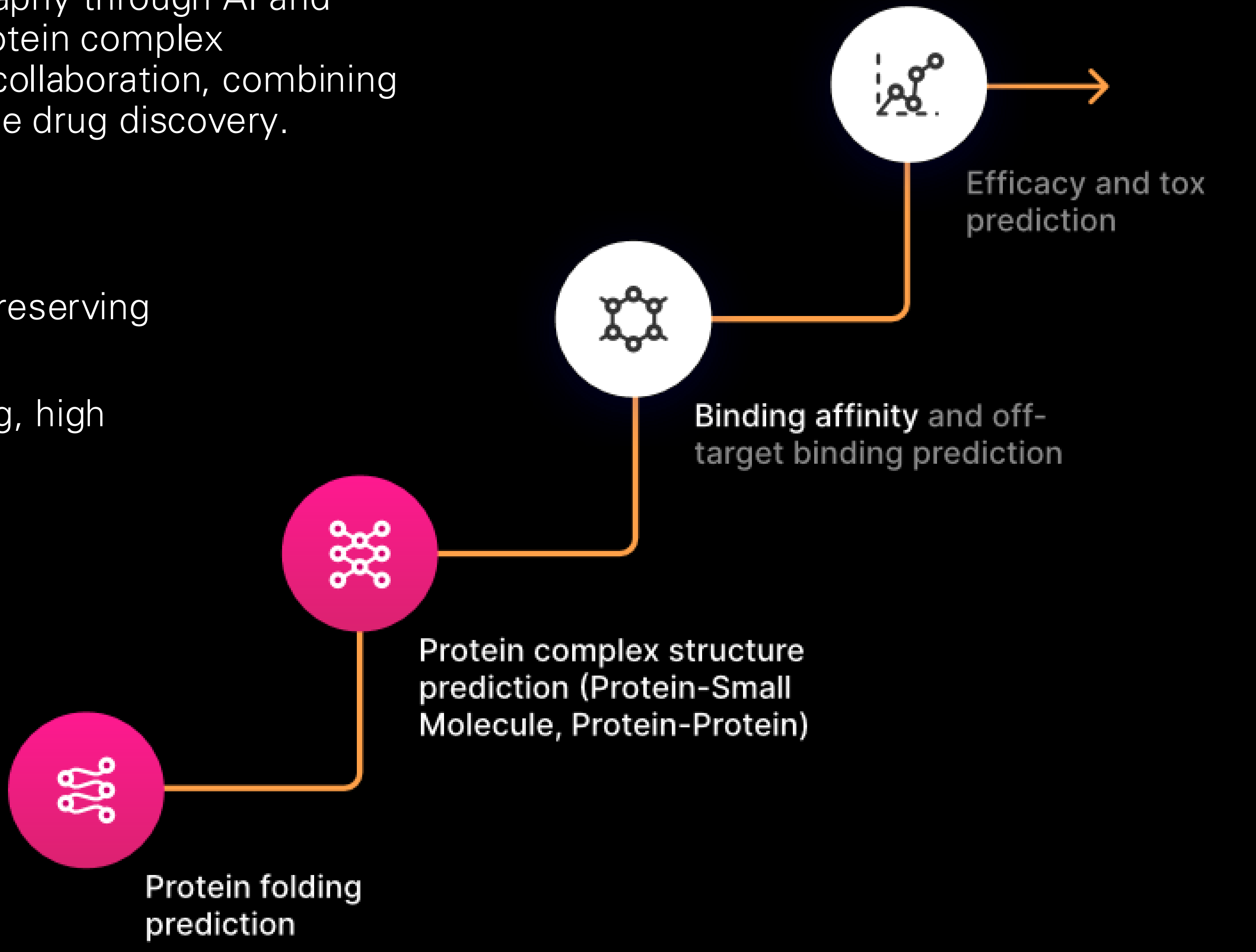
APHERIS abbvie Boehringer Ingelheim Johnson&Johnson

VISIT THE PAGE

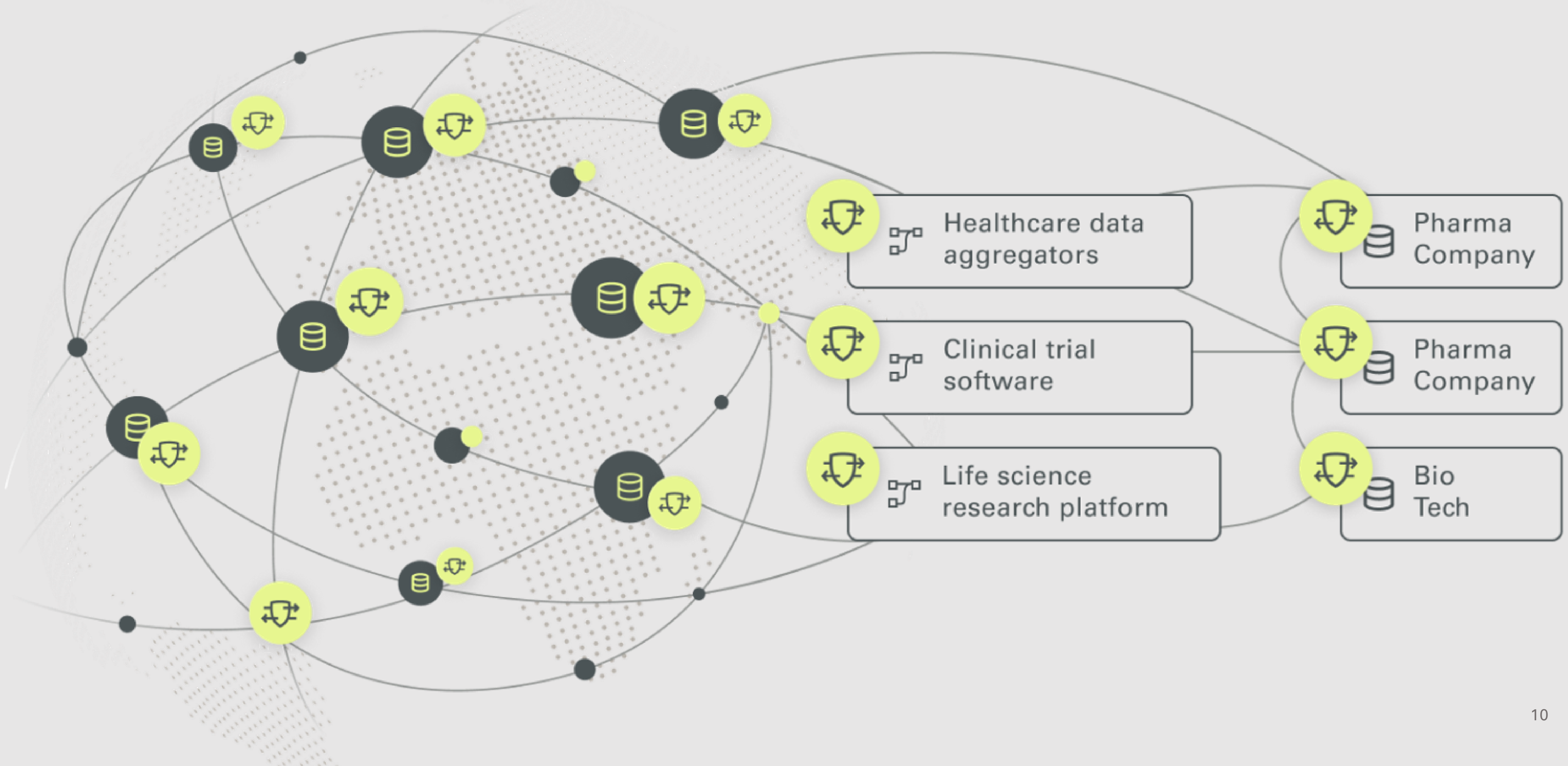
The AISB Consortium aims to leverage secure, federated learning to ensure both data privacy and AI model performance – with the ultimate goal of accelerating the application of AI in molecule design by achieving precision akin to X-ray crystallography through AI and machine learning in predicting protein complex structures. Our vision thrives on collaboration, combining our collective expertise to redefine drug discovery.

- Data available for training while preserving confidentiality/IP
- Infrastructure (Federated Learning, high performance compute)
- Leading edge AI algorithms
- Partnership between pharma and tech

Accelerate the step-change transformation in drug discovery powered by an E2E *AI/ML*



A combination of data, models, federated computing and governance can join the data-modeller network up



Thanks for listening

e.dobson@apheris.com
r.roehm@apheris.com

Legal