AI for Health
An ITU Focus Group
In collaboration with WHO

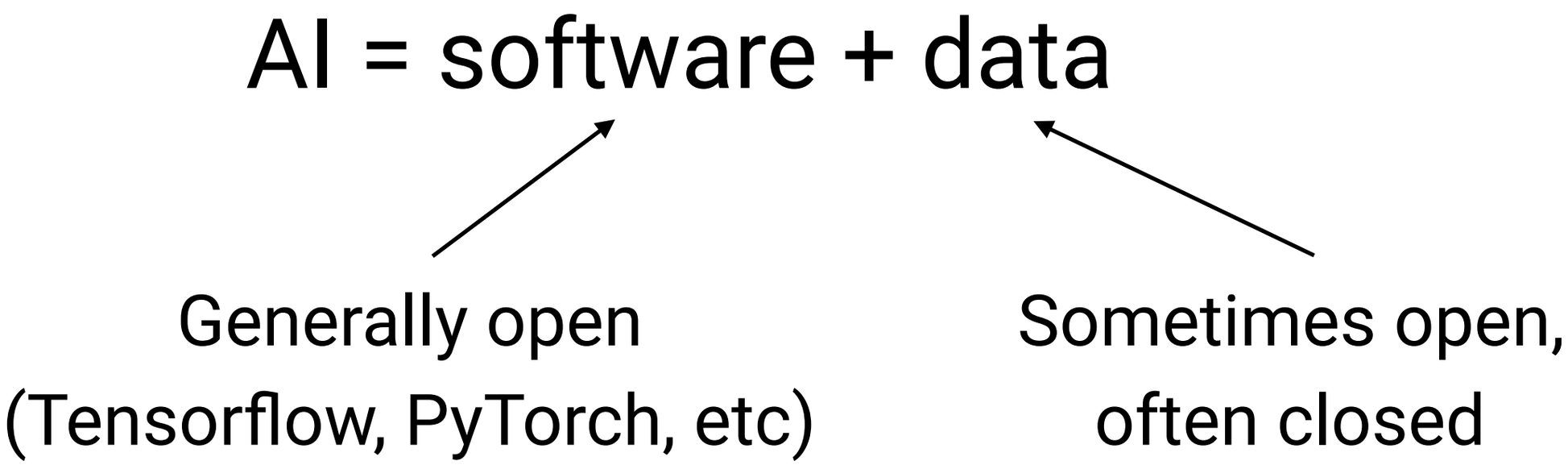
On Benchmarking

Prof. Marcel Salathé, Digital Epidemiology Lab, EPFL
marcel.salathe@epfl.ch
@marcelsalathe

ÉCOLE POLYTECHNIQUE
FÉDÉRALE DE LAUSANNE

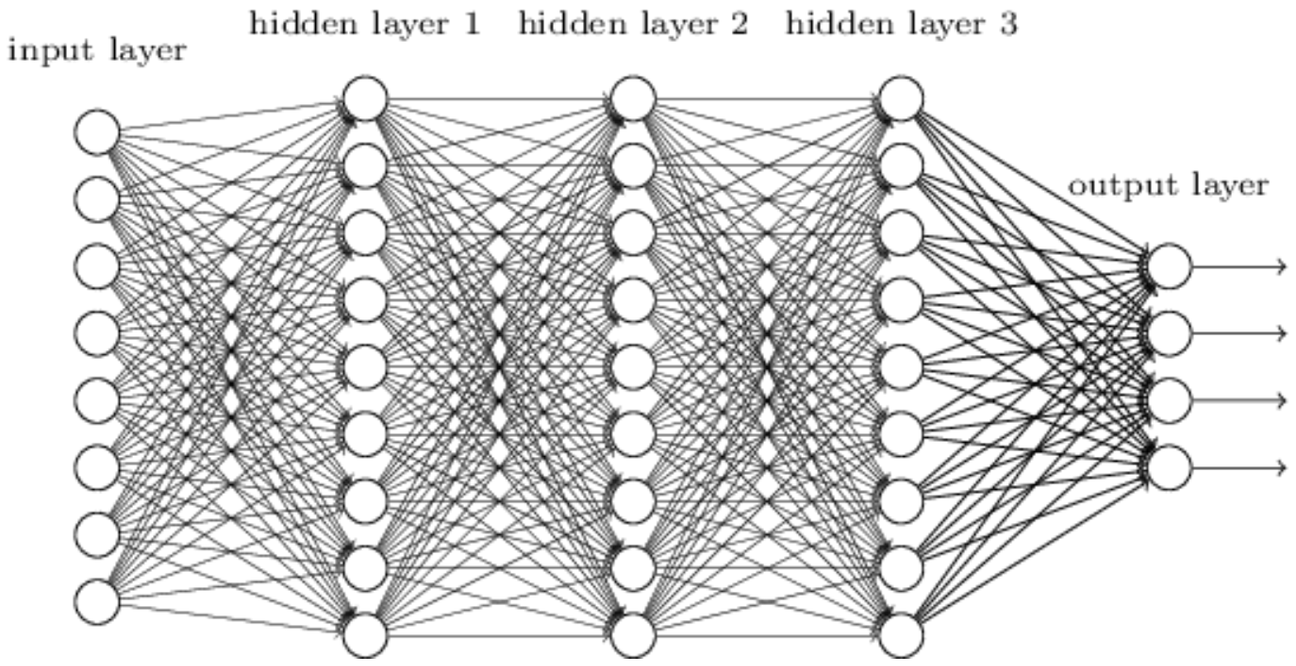
AI evaluation: status quo

Modern AI - generally based on deep learning (artificial neuronal networks). These networks are trained on data.



AI evaluation: status quo

Modern deep learning is generally “end-to-end”: from the input layer to the output layer, there is no domain expertise needed for the training of the network.



Labels

AI evaluation: status quo

Consequence: very “permissible” and accessible field - essentially, everyone can train deep learning networks provided s/he finds some good data to train on.

Advantage: extremely dynamic field, welcoming to outsiders

Disadvantage: There are no established ways to compare AI models - they exist as code, papers, apps, etc. (“Wild West”)

AI evaluation: status quo

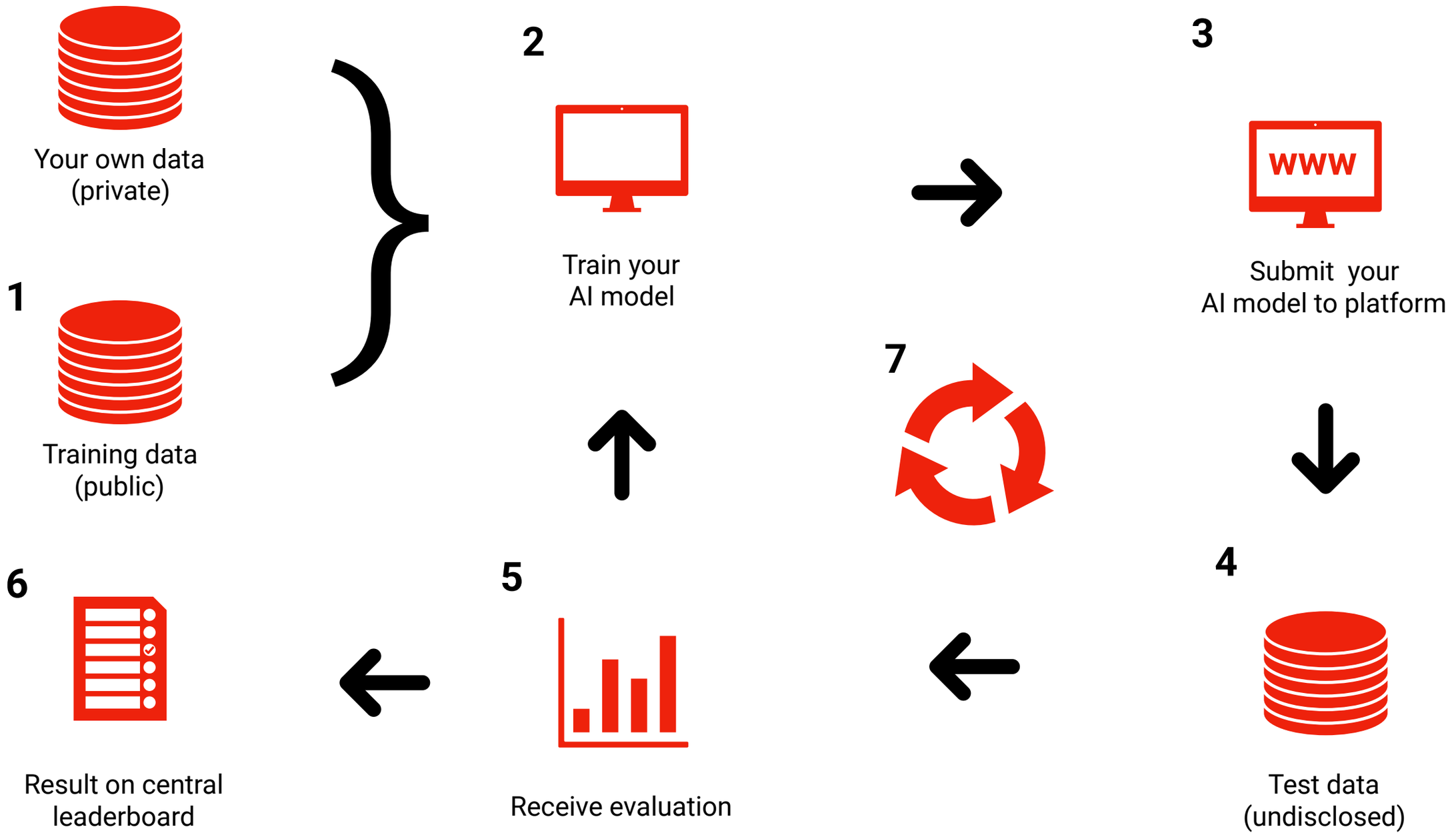
If you are a regulator (or in the policy making process in any shape or form), how are you going to deal with this?

Good Solution: We need benchmarks.

Ideal Solution: We need benchmarks that relevant stakeholders can agree upon.

That is why we need technology experts and policy experts to work together (hence the FG AI4H).

AI evaluation: proposed framework



AI evaluation: proposed framework

Your own data
(private)

Training data
(public)

1. Training modern AI models requires high quality data as input.

The FG can help identify high quality, open data sets to make the AI development more accessible and inclusive.

AI evaluation: proposed framework

Train your
AI model

2. Participants build AI models based on public data and other (private) data sources

Submit your
AI model to platform

3. FG will provide an online platform where participants can submit their AI model for evaluation.

AI evaluation: proposed framework

Test data
(undisclosed)

4. The model will be evaluated on a undisclosed test set. The FG oversees the process to ensure the highest integrity, quality, and confidentiality of this data set.

Receive evaluation

5. Following the evaluation of the test set, participants will receive the results of the evaluation.

AI evaluation: proposed framework

Result on central leaderboard

6. The results can be shown on a central leaderboard. This allows the global community to check the current state-of-the art performances in the field.

7. The evaluation process can be designed to be ongoing, as it enables stability and continuation.

Benchmarking: a two-step process

Working Group Technical Requirements:

WG coordinates a “Call for Algorithms: Feasibility Phase” and runs a preliminary benchmark to assess the feasibility of AI addressing the health problem identified by the WG “Health Requirements”.

A positive outcome of this feasibility assessment will then result in an evaluation managed by the WG “Evaluation”.

Benchmarking: a two-step process

Working Group Evaluation:

Following a positive outcome of a feasibility assessment by WG “Technical Requirements”, the WG “Evaluation” will develop a “Call for Algorithms: Evaluation Phase”.

It will define all aspects of evaluation, and manage the undisclosed test data set, as well as other work related to evaluation.

Benefits of proposed model

- ✓ Process is open and inclusive
- ✓ Process adds substantial clarity to the field (no more guessing about “what is the current status in the field?”, “how well does my algorithm perform?”)
- ✓ Being able to execute code allows for undisclosed test sets, and for replicable results.