데이터 가치 거래를 위한 Data Federation 기술

2023.6

김 화 종

목차

- 데이터 가치 산정
- Data Federation
- Data Federation을 이용한 가치 산정
- 기대효과

데이터 가치 산정

데이터 가치 산정 필요성

- 현 데이터 경제 체제에는 **데이터 서비스 제공자**가 대부분의 수익을 얻고 있다
- **원천 데이터 생산자**에게도 데이터 서비스로 얻는 혜택이 나누어져야 하나 이를 위한 장치가 존재하지 않는다
- 향후 **데이터 주권** 문제는 점차 심화될 것이다
- 데이터 생산자에게도 혜택을 주기 위해서는 **데이터의 가치**를 평가하는 방법이 필요하다
- 데이터의 가치 평가는 데이터 기반 산업의 장기적인 발전을 위해서도 필요하다

→ 그러나 현실에서 데이터 가치 평가는 매우 어렵다

데이터 가치 산정이 어려운 이유

- **주관성**
  - 데이터 가치 평가는 주관적이며 누가, 언제, 어떤 목적으로 데이터를 사용하는지에 따라 데이터의 가치가 매우 달라진다
- **불명확한 품질 기준**
  - 오류율, 결측치 비율, 데이터 수명, 데이터 크기 차이 등 데이터 품질 기준을 미리 정하는 것이 어렵다
- **미래 잠재 가치 미정**
  - 데이터 기반 비즈니스가 급변하고 있어 데이터의 미래 잠재적인 활용 가치를 미리 추정할 수 없다
- **정책적 변화**
  - 개인 정보 보호 규정, 데이터 거버넌스, 윤리적 기준이 계속 달라지며 이에 따라 데이터의 가치가 달라질 수 있다

데이터 가치 산정 방안

- 데이터의 가치를 산정하는데 있어, **미리 정한 가격표** 방식이 아닌 다음과 같은 요건을 갖춘 **새로운 접근**이 필요하다
- **객관적, 상대적**
  - 데이터의 가치를 주관적, 절대적으로 평가하는 것이 아니라, 객관적, 상대적으로 평가하는 방식
- **동적인 품질 기준**
  - 데이터의 품질이 정적으로 한번에 정해지는 것이 아니라 활용 목적, 활용 결과에 따라 품질 기준이 정해지는 개념
- **미래 가치 반영**
  - 데이터가 생성되는 시점이 아니라, 미래에 어떻게 사용되는가에 따라 가치를 산정하는 방법
- **정책변화 반영**
  - 현재의 법적, 정책 기준이 아니라 데이터가 사용될 때의 정책이 반영되는 방법

제안 방안

- 앞의 네 가지 이슈를 해결하는 한가지 방안으로, 다음과 같은 데이터의 가치를 산정하는 방법을 제안

→ 데이터가 실제로 활용되어 어떤 가치를 만들었을 때, 사용된 데이터가 <u>이에 기여한 정도를 측정</u>하는 방법

- 즉, 데이터가 데이터분석, 머신러닝 등에 **사용된 시점**에,
- 모델의 성능개선에 **기여한 정도를 상대적**으로 평가하고,
- 이 기여도를 기준으로 어떤 **보상**을 하는 방안

Data Federation

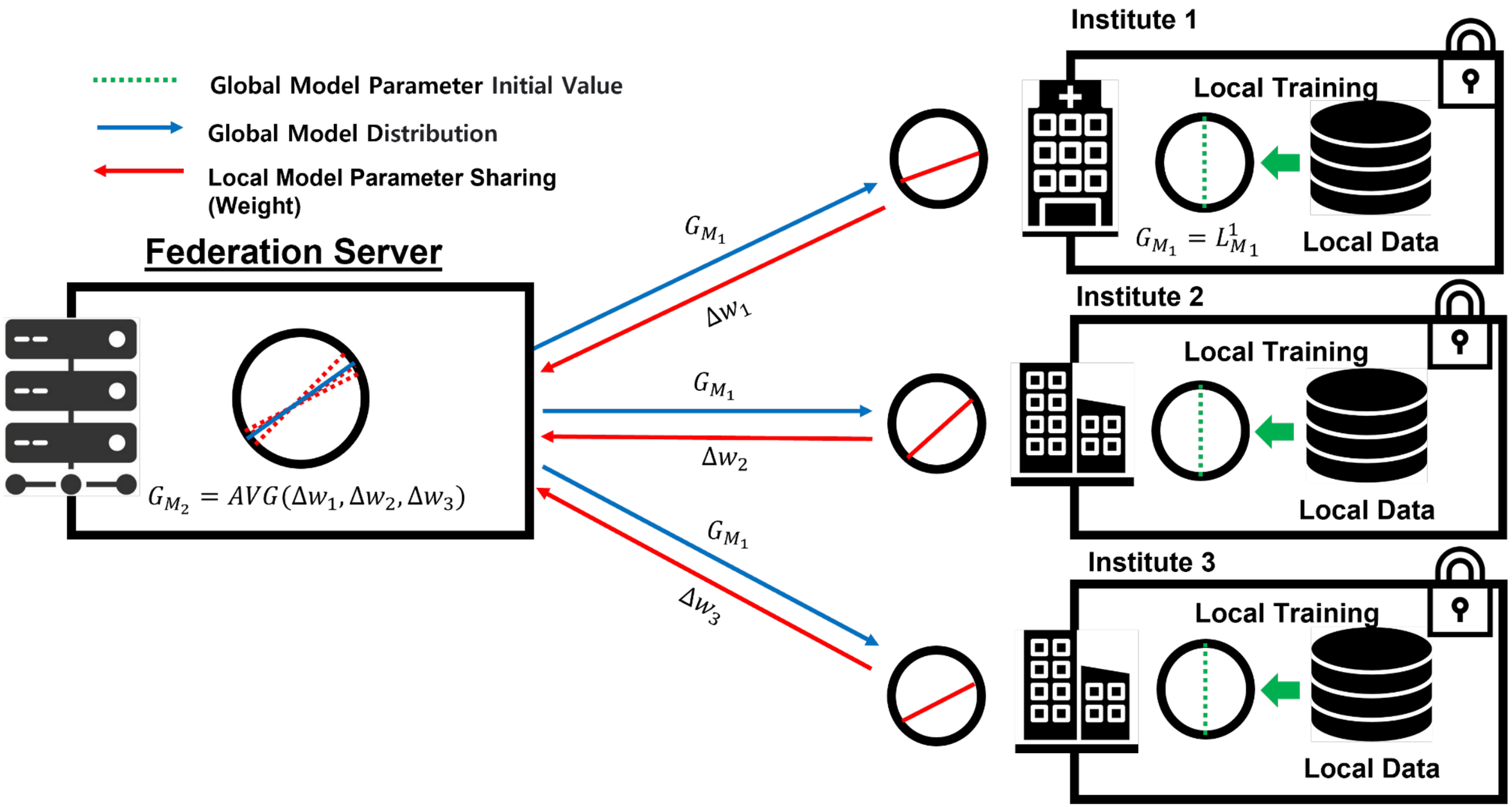
Federated Learning

- 여러 클라이언트의 데이터를 한 곳에 집적하여 머신러닝 모델을 만드는 방법은 익명화, 비실명화를 하더라도 개인정보보호 등의 문제를 피하기가 어렵다
- 이러한 문제를 해결하기 위해서 연합학습 방법이 도입되었다
- **연합학습의 정의**

  - 각 클라이언트가 학습하여 얻은 **로컬 모델의 파라미터를 서버에서 취합하여 성능이 향상된 글로벌 모델**을 만들고 이를 재배포, 파인튜닝하는 방식 (2017년 구글이 제안)

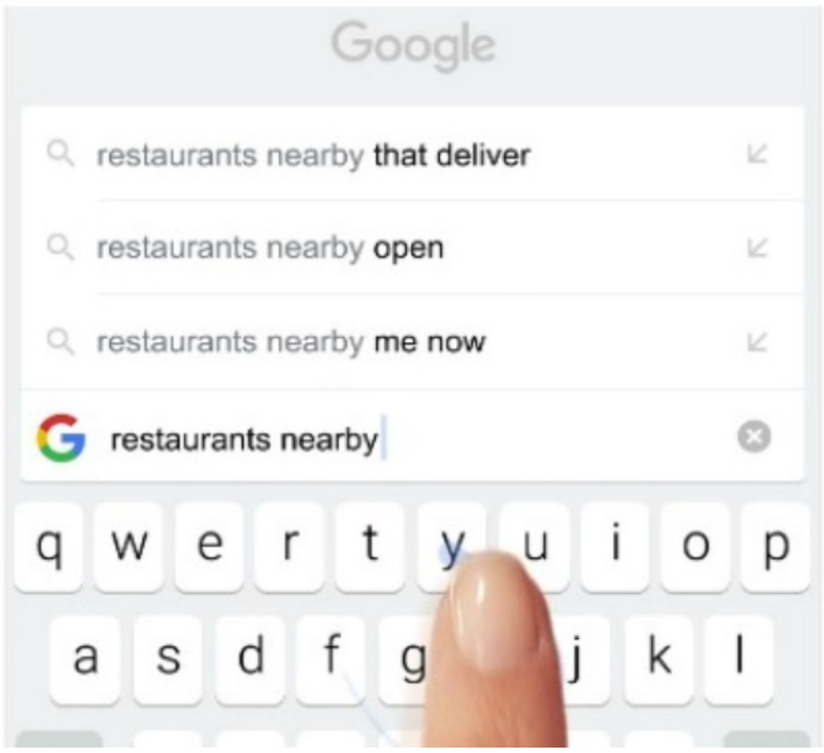
Federated Learning

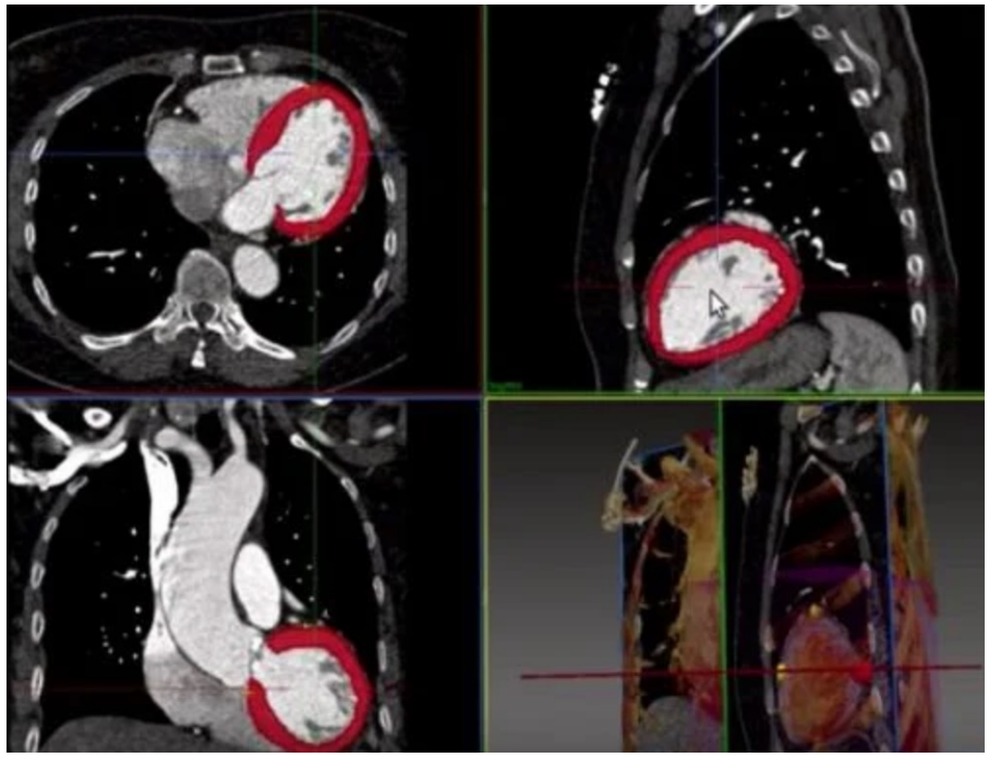
- 데이터가 이동하지 않고 모델 파라미터만 이동한다



연합학습 사례

- 구글
  - 단어 자동 완성 추천기능에 Federated Learning을 도입 (Gboard)
- Clara
  - NVIDIA, 미국 영상의학회, UCLA 헬스, 매사추세츠 제너럴 브리검 병원, 영국 킹스칼리지 런던 등이 참여
  - 영상 의료 데이터 연합 분석 AI 시스템 구축
  - 코로나19 초기 검사만으로 산소 보충이 필요한지 판단하는 AI 모델

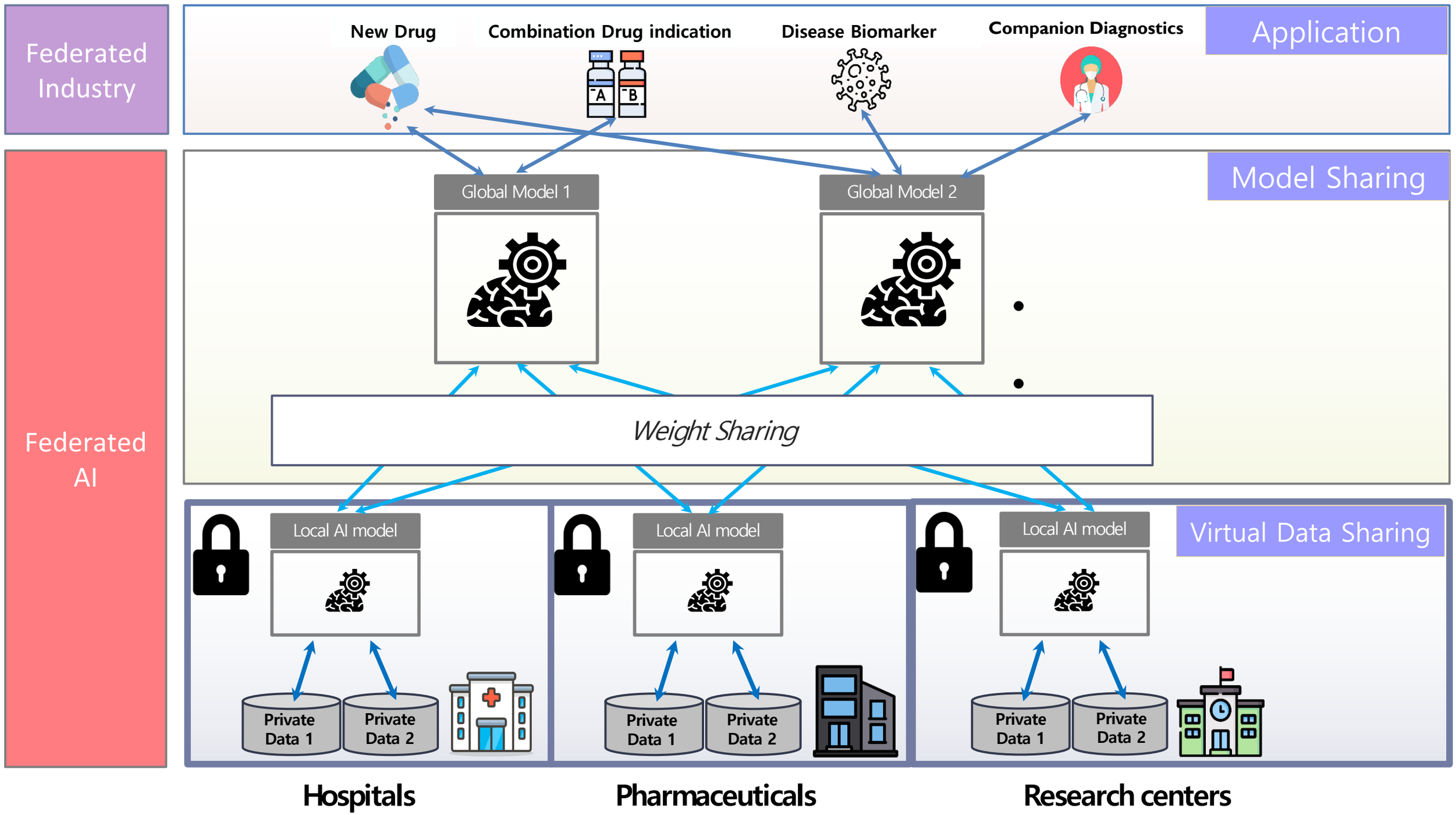




연합학습 사례

- Machine Learning Ledger Orchestration for Drug Discovery (2019~2022)
  - IMI(Innovative Medicines Initiative)에서 2천만불 지원
  - 약물 발견 예측 모델을 훈련하고 평가할 수 있는 FL 프레임워크 구축

Federated Drug Discovery



연합학습의 장점

- 각 클라이언트는 데이터를 외부로 공개하지 않으므로 개인정보보호 문제의 근본적인 해결책 제시
- 각 클라이언트가 대량의 데이터를 처리하지 않아도 됨
- 전체 데이터를 한 곳에 모아서 모델을 만든 것과 유사한 성능을 갖는 글로벌 모델을 각 클라이언트가 공유함
- 새로운 머신러닝 모델의 현장 데이터 검증 및 솔류션 배포가 수월함

Data Federation을 이용한 가치산정

Data Federation & Evaluation

- **데이터 연합**
  - 데이터가 로컬하게 단독으로 사용되는 경우가 아닌, 여러 클라이언트의 데이터를 연합 사용하는 경우를 가정
  - 머신러닝은 여러 클라이언트의 데이터를 다양하게 사용할수록 성능이 향상되므로 데이터의 공유활용을 지속적으로 요구되고 있음
- **데이터 가치 평가**
  - 데이터가 머신러닝 모델 학습에 사용되어 성능을 향상시킨 경우 이의 기여도를 평가
  - 데이터의 성능향상 기여도를 상대적으로 평가하고 이를 데이터 가치 산정에 사용

Data Federation & Evaluation

- **객관적, 상대적 평가 가능**
  - 데이터 소유자가 주관적으로 가치를 평가하는 것이 아니라, 데이터를 분석(머신러닝)에 활용한 후 성능개선에 기여한 정도를 객관적, 상대적으로 평가하는 것이 가능
- **동적인 품질 기준**
  - 데이터의 품질을 미리 정하는 것이 아니라 머신러닝 모델 성능 개선에 반영되는 정도에 따라 정해짐
- **미래 가치 반영**
  - 데이터가 생성되는 시점이 아니라, 머신러닝 모델에 활용되는 시점에 정해짐
- **정책변화 반영**
  - 연합학습은 프라이버시에 원천적을 강점이 있는 방법이며 또한 데이터가 사용될 때의 정책을 반영시킬 수 있음

기대효과

- 데이터 제공자에 대한 평가 및 인센티브 제공 가능
  - 현재 많은 기관들이 FL에 적극적으로 참여하지 못하는 가장 큰 이유는 FL 학습에 자신의 데이터를 기여한 것에 대한 명확한 평가와 보상 체계가 없기 때문임
  - 정부 및 관리 기관에서는 이러한 평가를 향후 지원을 차별화 하는데 사용할 수 있을 것임
  - 데이터 기여도 측정과 보상체계를 통해서 데이터 보유자 (Data Owner)의 참여를 적극 유도할 수 있을 것임
- 새로운 머신러닝 모델의 현장 적용 가속화
  - 새로운 머신러닝 모델에 대해서 다양한 데이터로 검증하는 것이 수월해짐
  - 우수한 솔류션의 발굴과 배포가 수월해짐

해결해야 할 과제

- 객관적인 평가 방법 개발
  - 각 클라이언트가 보유한 데이터의 글로벌 **모델 향상 기여도**를 평가하는 공정한 기술 개발이 필요함
  - 데이터 기여도 평가에 대한 표준화가 필요함
- 2단계 구조의 가치평가
  - 클라이언트가 기관인 경우, 이 기관에 포함된 각 개인 데이터의 기여도를 산정하는 방안 필요
    - 예를 들어 A, B, C, D 기관이 연합학습에 참여하였다면 A 기관이 보유한 개인별 데이터는 동일한 비중으로 나눈다
- 이를 위해서 각 개인은 데이터 수집기관과 다시 정산을 하는 2단계 구조가 필요함

감사합니다