データの量と質がAIの答えを変える

学習に使うデータが変わると、AIの答えはどう変わるのだろう

課題：「りんご」か「ほかの果物」かを見分けるAIを、3通りのデータで学習させる。

新しい画像：  緑のりんご

A 少ないデータ　3枚

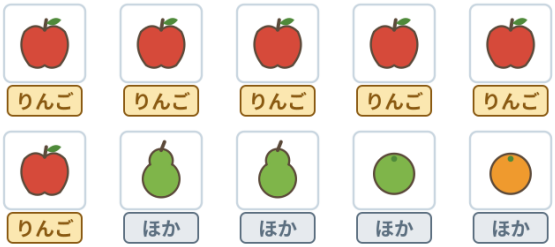


AI 学習したAI

△ **答え：決めきれない**
確からしさ：りんご 55%・ほか 45%（架空）

例が少なく、見分け方を
十分に学べない

B かたよったデータ　10枚



AI 学習したAI

× **答え：ほかの果物**
確からしさ：ほか 85%（架空）

緑のものはすべて「ほか」
だったので、「赤い＝りんご」
と学んでしまった

C 多くて多様なデータ　10枚



AI 学習したAI

○ **答え：りんご**
確からしさ：りんご 92%（架空）

色や形がいろいろな例から、
りんごの特徴を学べた

データを集めるときに考えること

- **量** 例は足りているか
- **かたより** 色・形・向き・場面などに、かたよりはないか
- **確かめ** 学習に使わないテストデータで確かめる

集め方を決めるのは人。データのかたよりは、人を不公平に扱う答え（バイアス）につながることもある。

※ 模式図。データの数と確からしさの数値は架空。実際の学習には、もっと多くのデータを使う。