Educational Big Data and the Digital Learner

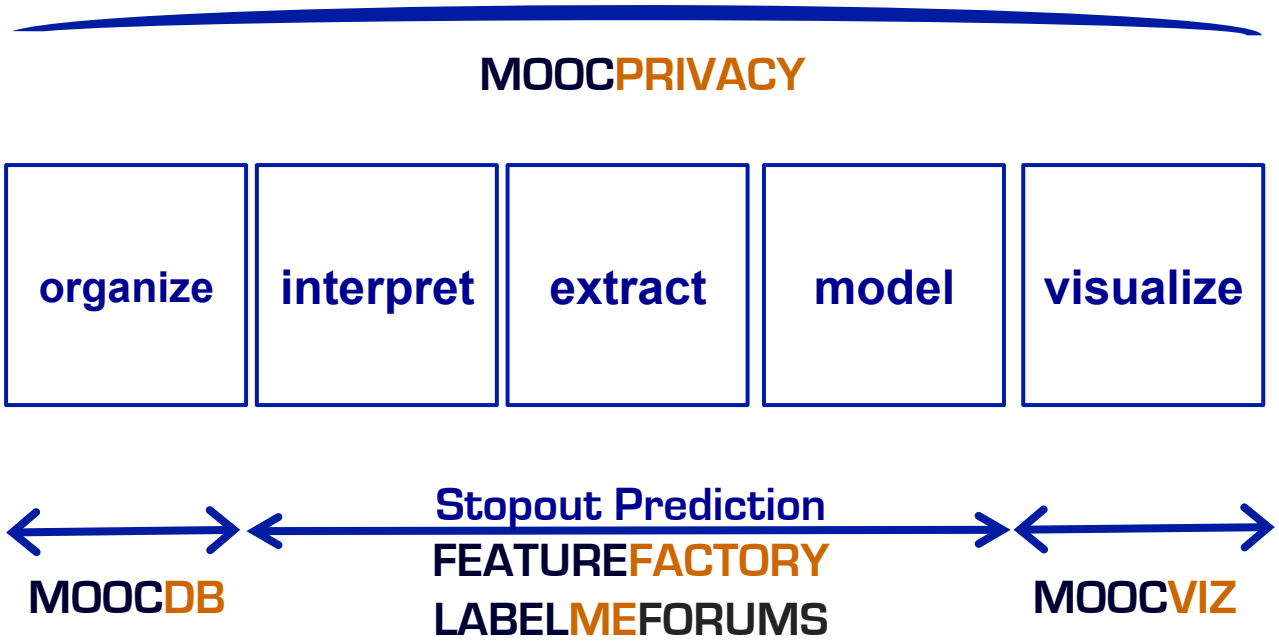
Privacy and Access Considerations

Una-May O'Reilly
ALFA Group
Computer Science & AI Lab
MIT

ALFA
ANYSCALE LEARNING FOR ALL

CSAIL

ALFA MOOCdb Project



Personalization of Learning

Teaching /Learning Research

Enabling technology

- MOOCdb project targets a Data Science Commons concept
- Includes multiple projects and open-source software that transform data and enable MOOC-related data science

ALFA
ANYSCALE LEARNING FOR ALL

MOOCDB Data Science Commons

CSAIL

Sharing MOOC data science software

The moocdb project

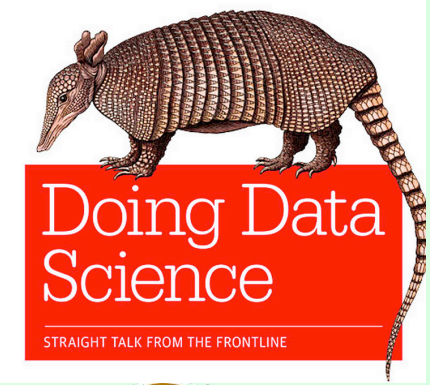
The hub becomes software, NOT DATA!
This is a mooc data science commons!
moocdb.csail.mit.edu
- moocViz
- featureFactory
- labelMe Text

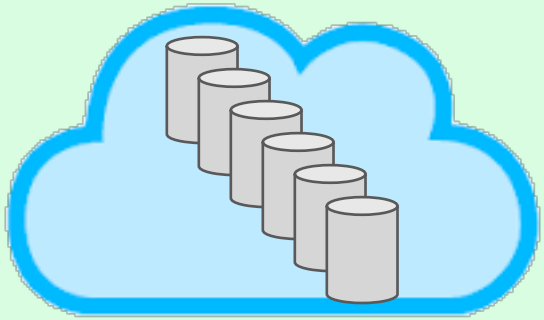
- Mooc data will be spatially distributed
- Mooc data access will be controlled locally
  - Chances of a big Data hub are low…
  - Maybe meta-data can be shared

Another Strategy

- MoocDB concept of software sharing shifts privacy into checking software outputs
- Return to the goal of openly sharing data
  - Must return to the issues of privacy protection
    - » Risk of re-identification through linking
- In this context:
  - Technology is integral
    - » But it's only one element of the story…
  - Start by bringing clarity to the complexity of the system by identifying the stakeholders

The Stakeholders





Instructor
Institution of instructor

coursera
edX

Platform provider

Analysts

Subject/provider Controller?

Controllers

ALFA
ANYSCALE LEARNING FOR ALL

CSAIL

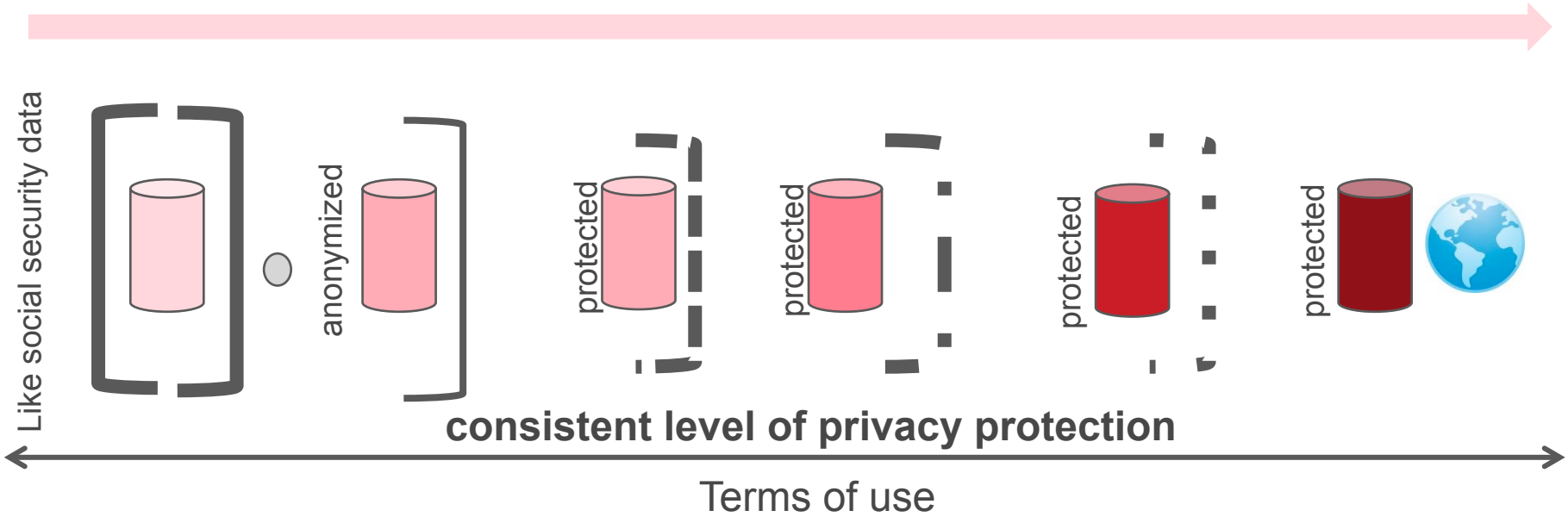
Sharing Data and Protecting Privacy
The Present

Two extremes

1. **Rich, raw data – highly vetted**
2. **Used, simple data - open**

One Possible Future

- **Sharing via gradualism: a consistent protection level provided while**
  - **Access increases**
  - **protection mechanism potentially changes:**
    - **» This depends on nature of data being shared**
      - **We are interested in sharing transformed, advanced research-ready data**



ALFA
ANYSCALE LEARNING FOR ALL

CSAIL

Data Privacy and Big Data

- **Study from policy, access control and technology vantages in global, broad contexts**

ALFA
ANYSCALE LEARNING FOR ALL

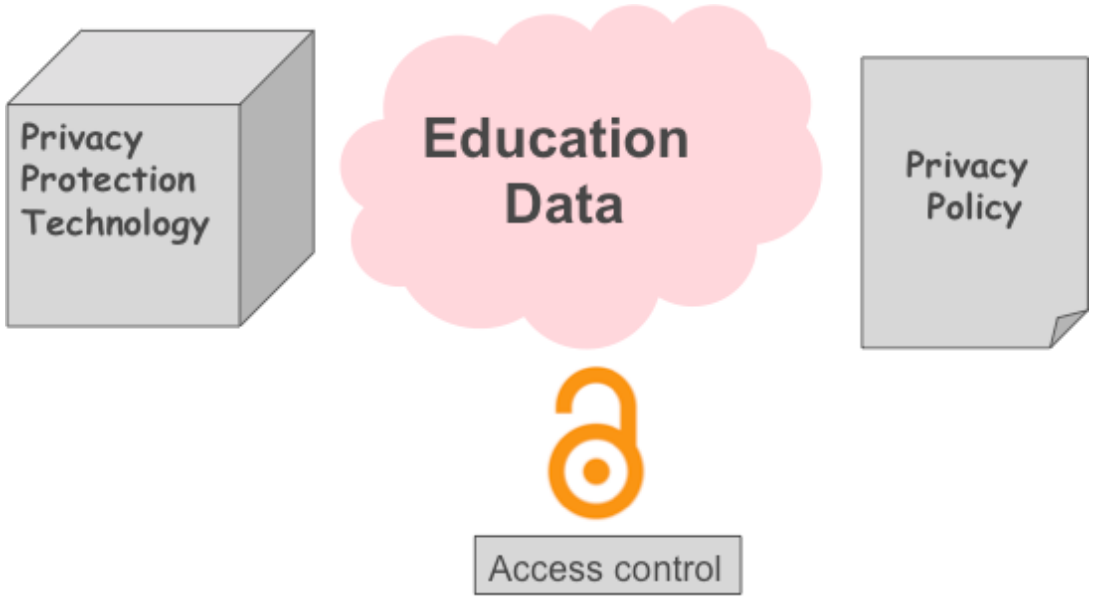
CSAIL

Closer to home: USA

- **BigData@CSAIL initiative**
  - **Workshops on data privacy ->Report**
  - **Working group report, sub-group on online learning**

- **Helped me to formulate a working context and its requirements for progress**

A Working Context



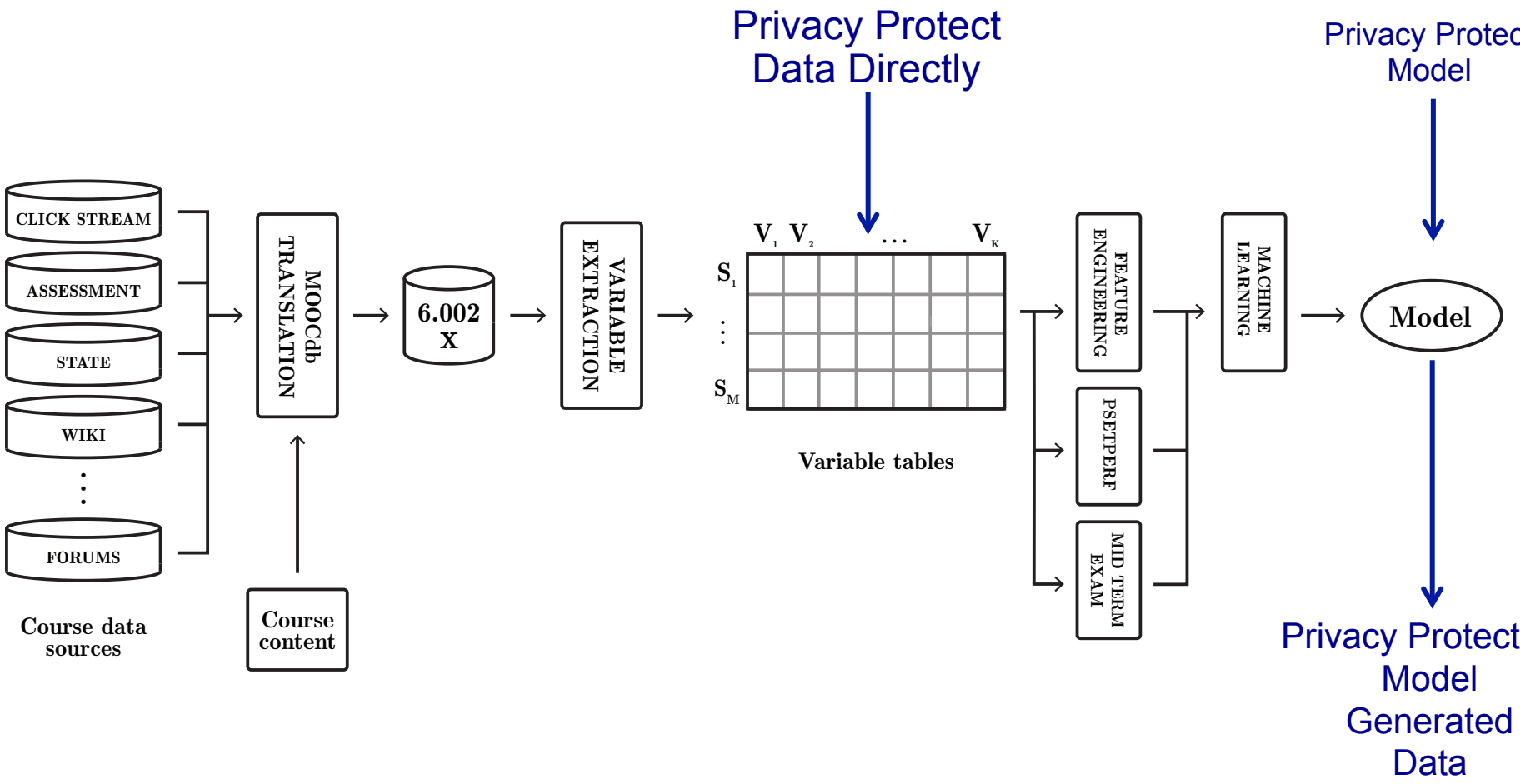
Requirements

- **Identify the data we want to share**
- **Assess: technology, control and policy practices, state of art**
- **Integrative Roadmap!!**

ALFA
ANYSCALE LEARNING FOR ALL

CSAIL

Sharing Research-Ready Transformed Data



Sharing transformed data that enables collaboration and cross-institute research will enable a MOOCDB data science commons concept

ALFA
ANYSCALE LEARNING FOR ALL

CSAIL

Assessment

- **Technology**
  - **Gap between practitioners' needs and technology maturity**
- **Policy**
  - **Asilomar**
- **Controllers**
  - **MIT Registrar "challenge"**

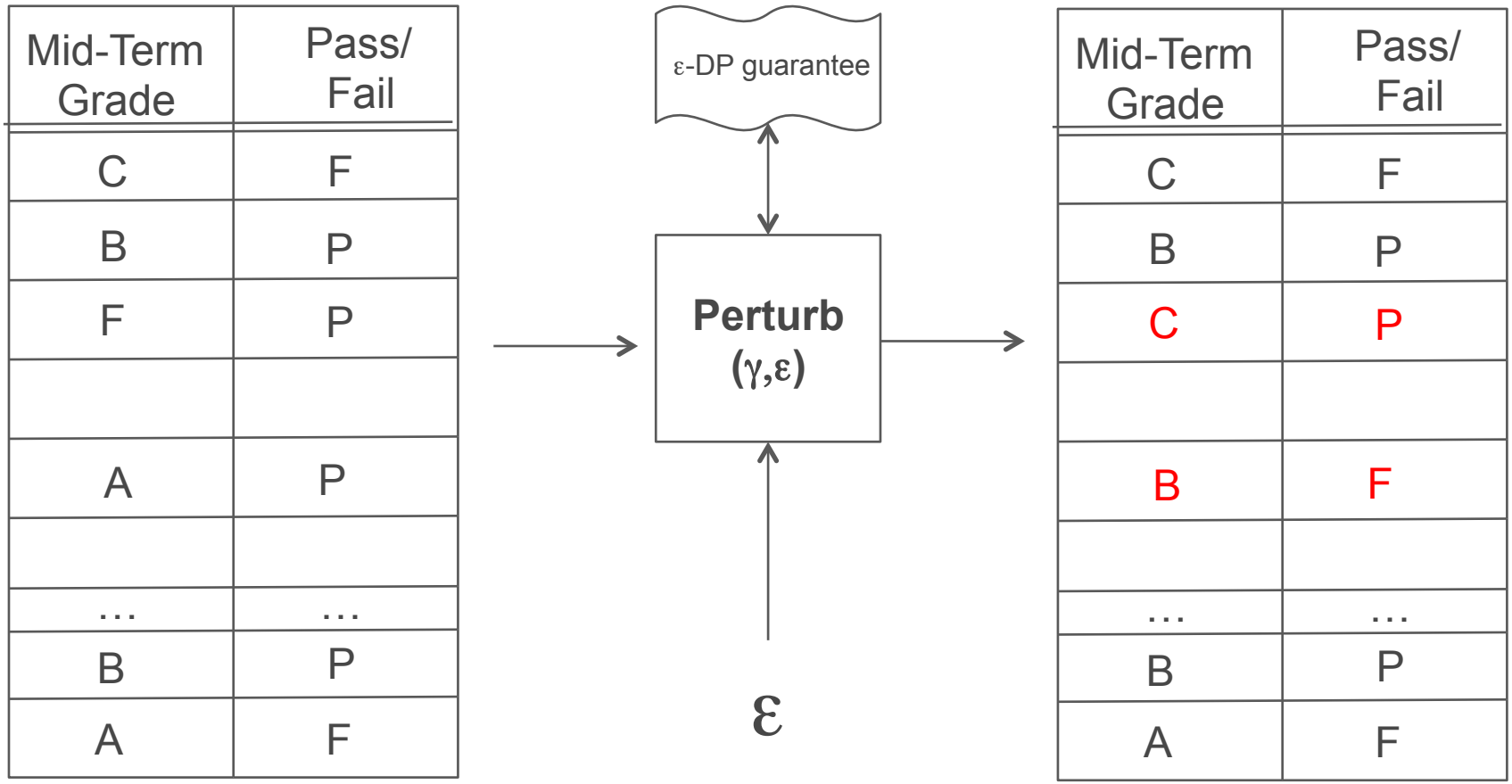
ALFA
ANYSCALE LEARNING FOR ALL

CSAIL

Roadmap

- **Design in progress!**
- **Started a modest investigation thread**
  - **Theory/Social Science Sharing/Privacy-protected data**

Sanitizing the Data Directly



Truthful data

Protected Data

Post-Randomization

ALFA
ANYSCALE LEARNING FOR ALL

CSAIL