C2D: Conclave Cloud Dataverse

Privacy-Preserving Scientific Data Analysis in an Open Cloud

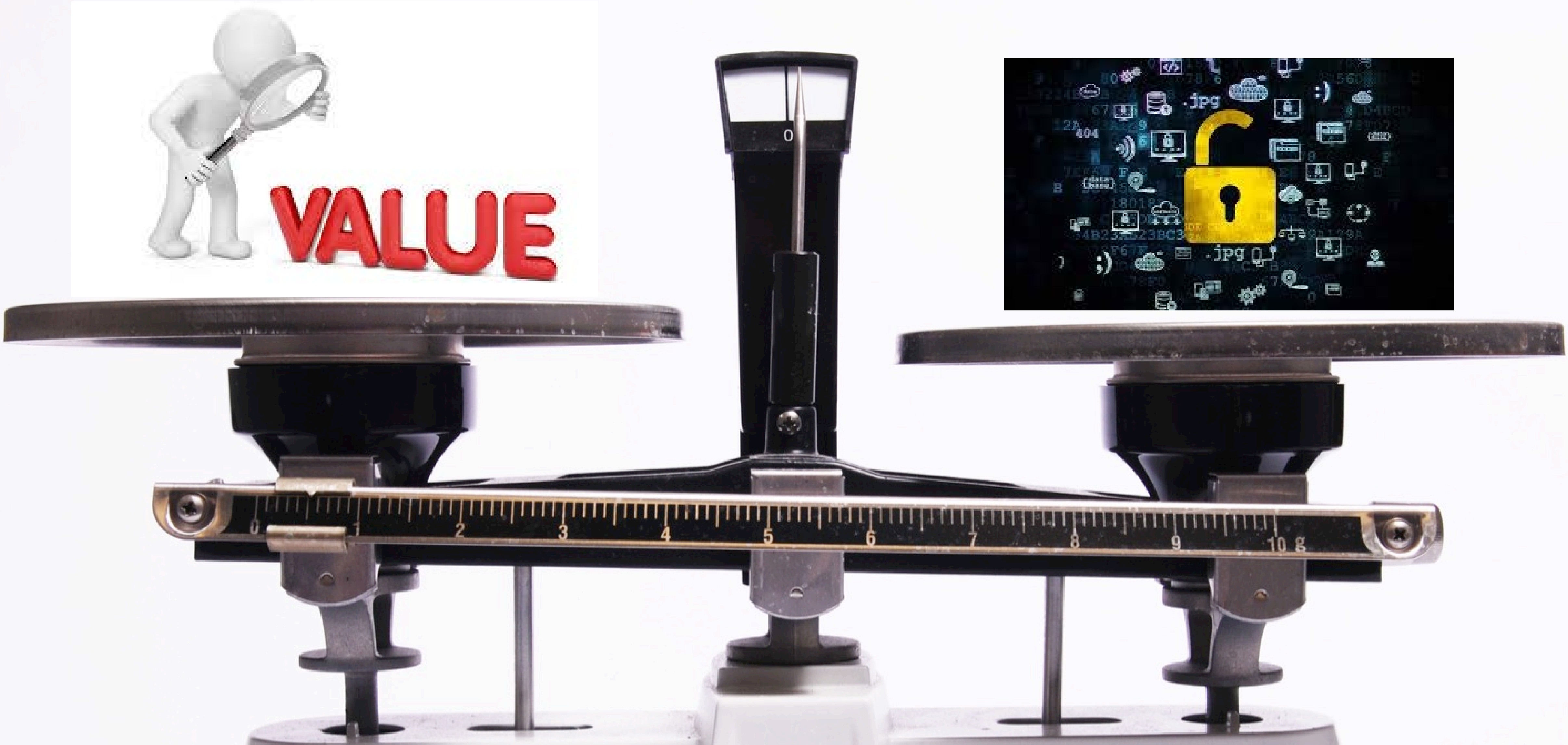
Ben Getchell, Mayank Varia, Andrei Lapets, Ata Turk, Orran Krieger,
Robert Bartlett Baron, Nicolas Haddad, Parul Singh

C2D Use Cases:

- **Tier-1 trauma centers** in Boston **want to join reports about cases they service without revealing any patient data**
    - E.g. how many trauma cases they serviced during the marathon bombing
- **Researchers in hospitals** want to **pool data across multiple hospitals about rare diseases without revealing patient data**

Sharing data

Protecting data



Images: Facebook, Wikipedia

Secure Multiparty Computation (MPC):

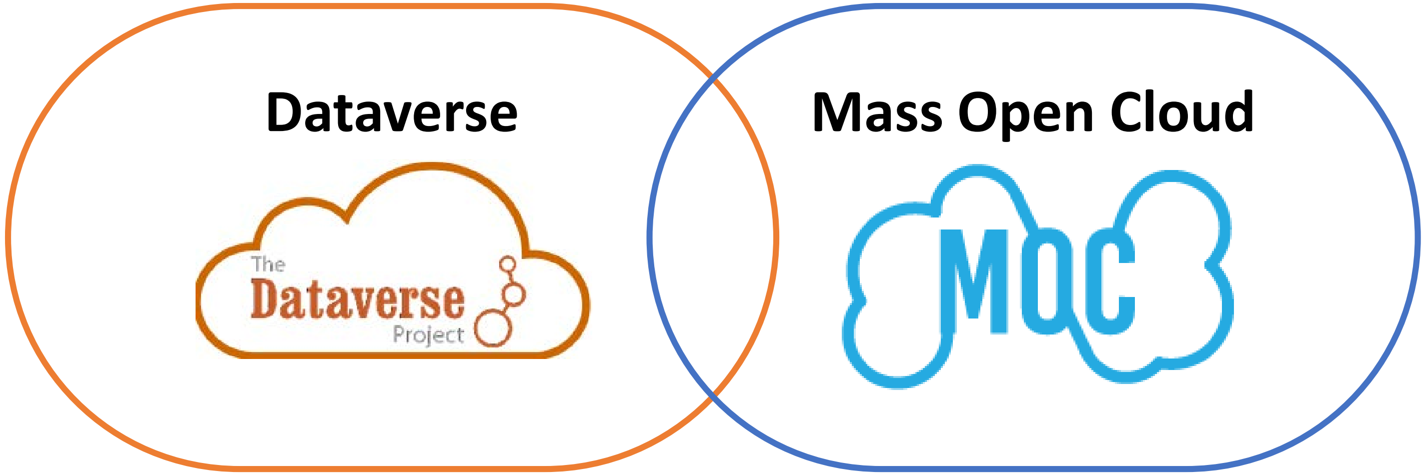
- **Securely** compute and analyze data with collaborators.
- Each contributor’s data is never shared in the clear **with anyone.**
- Only the result of the computation is revealed.

Sharing data

and

Protecting data

Privacy-Preserving Scientific Data Analysis in an Open Cloud

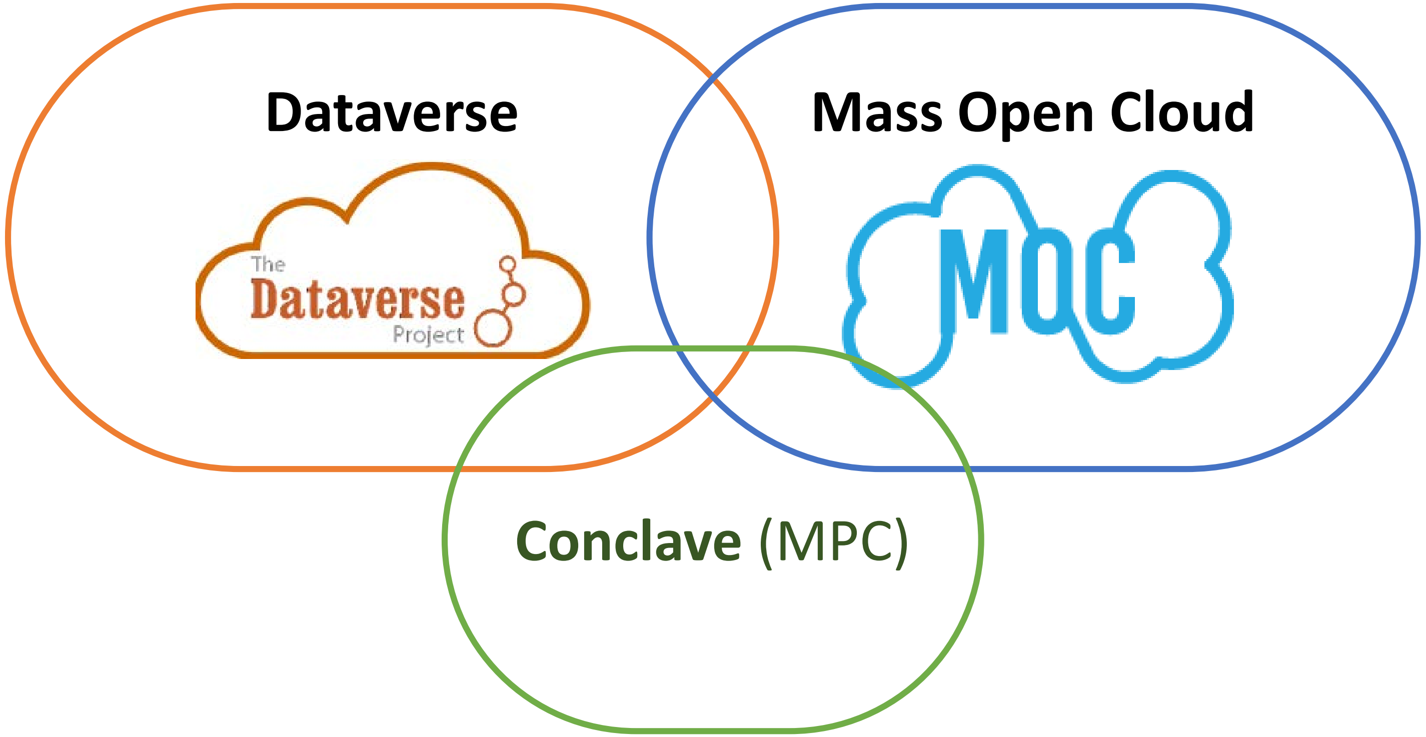


Dataverse

Mass Open Cloud

- Open-source platform for data repositories
- Mechanisms to control access
- Incentives to share and credit use of data

Privacy-Preserving Scientific Data Analysis in an Open Cloud



Dataverse

Mass Open Cloud

Conclave (MPC)

Conclave: scalable MPC

- **Relational workflows**
    - SQL-like query language

- **Minimize MPC**
    - Automatically determine local and MPC barriers

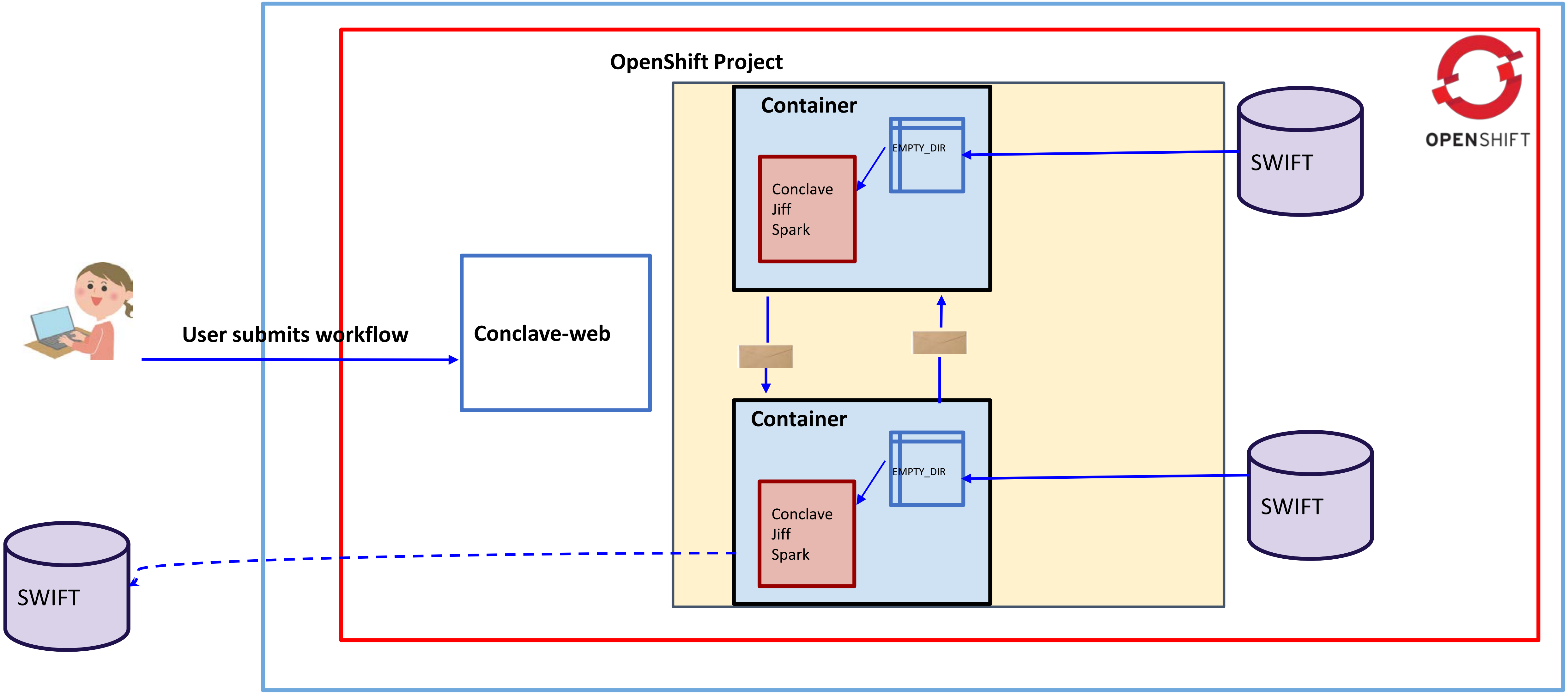
- **Currently**:
    - Connects to existing backend data stacks (e.g. – Spark)
    - Scales 4 magnitude higher than most MPC engines (~100GB range)
    - Code at https://github.com/cici-conclave

The C2D framework

- **Runs on containers**
  - Each container stores data owned by a single project
  - Containers never share data with one other, and are deleted when a computation terminates

- **OpenShift / K8s**
  - Pods are spawned for computations

The C2D framework



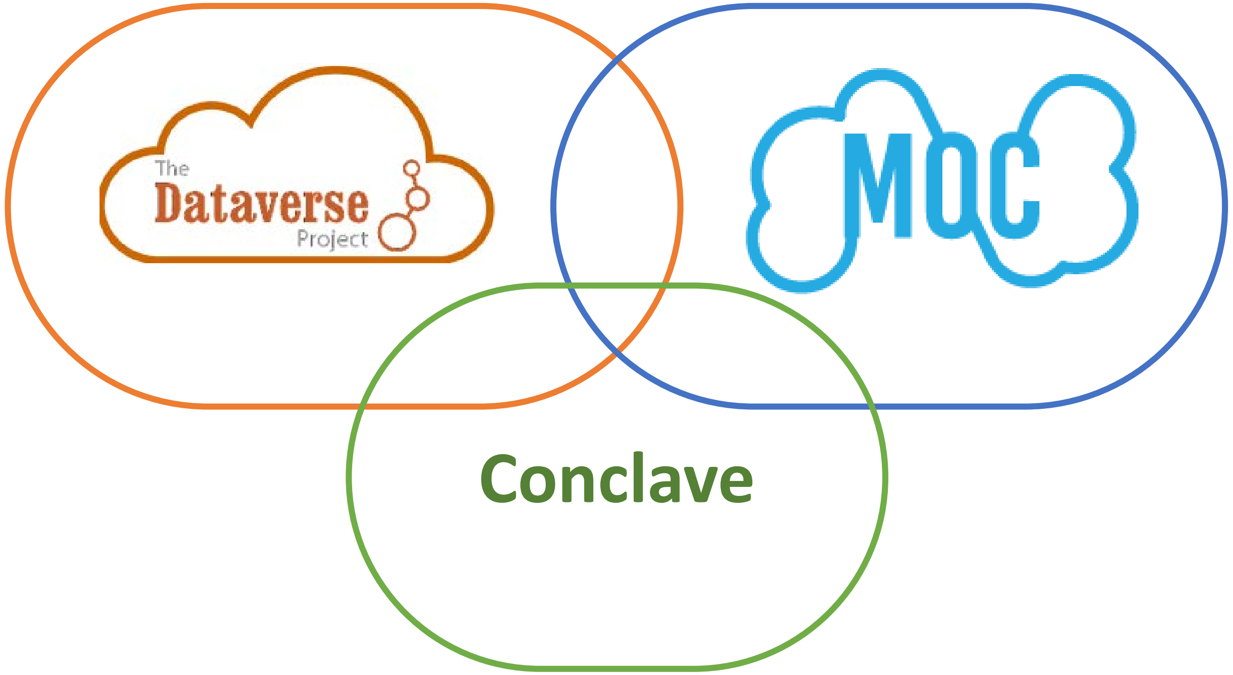
Ongoing work:

- **Privacy engine**
  - Allow data owners to restrict which kinds of computations can be run on their data
- **Dataverse integration**
  - Currently using Swift
- **Computations across OpenShift deployments**
  - Pods with a user’s data will only be run on a deployment associated with that user

Summing up

- **MPC can alter the way we do data science**
  - No need to choose between data sharing and privacy
  - Unique insights for social good
- **C2D on the MOC can do this**
  - Brings MPC to where the data already lives
  - Separate cryptographic details from user

Thanks!



Conclave