C2D: Conclave Cloud Dataverse

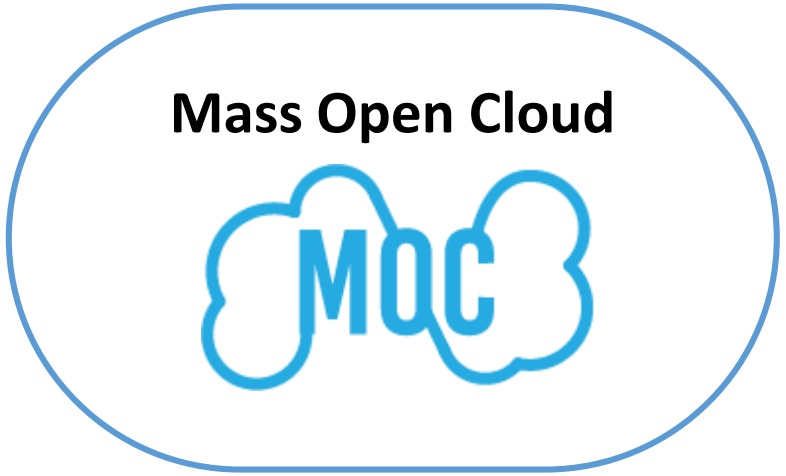
Privacy-Preserving Scientific Data Analysis in an Open Cloud

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

Data Utility vs Data Privacy

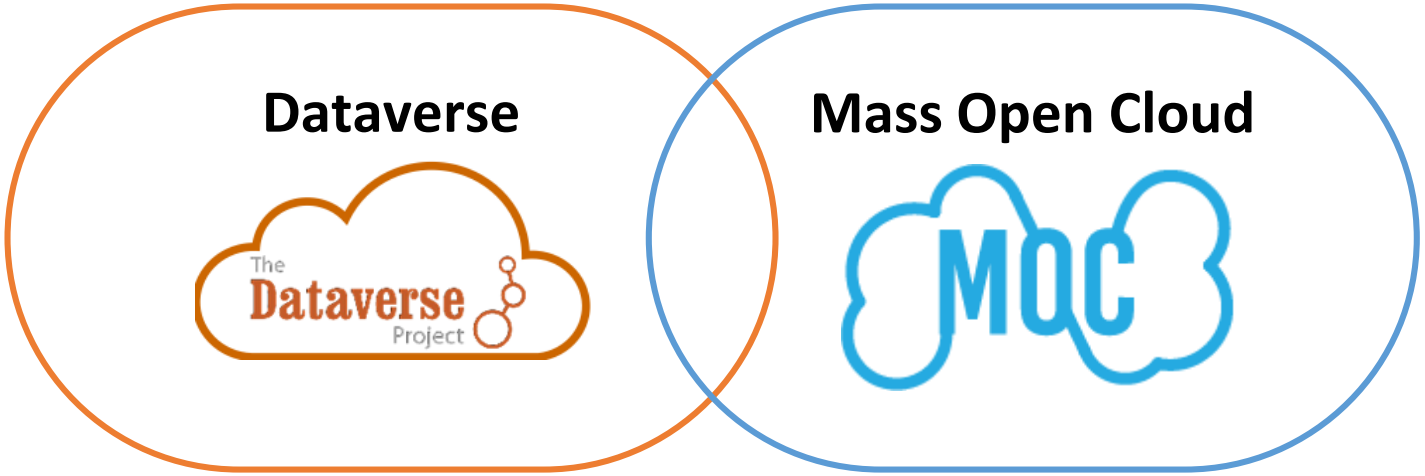
- **Companies** in MA **want to compute average salary** differences across genders, ethnicities, ... **without exposing average salary of any company**
- **Tier-1 trauma centers** in Boston **want to generate aggregate reports about cases they service without revealing any patient data**
  - E.g. how many trauma cases they serviced during the marathon bombing
- **Researchers in hospitals** want to **generate aggregate statistics about rare diseases across multiple hospitals without revealing patient data**
- **Companies** want to **run data analytics in the public cloud** but do not **trust a single public cloud provider**

Privacy-Preserving Scientific Data Analysis in an Open Cloud



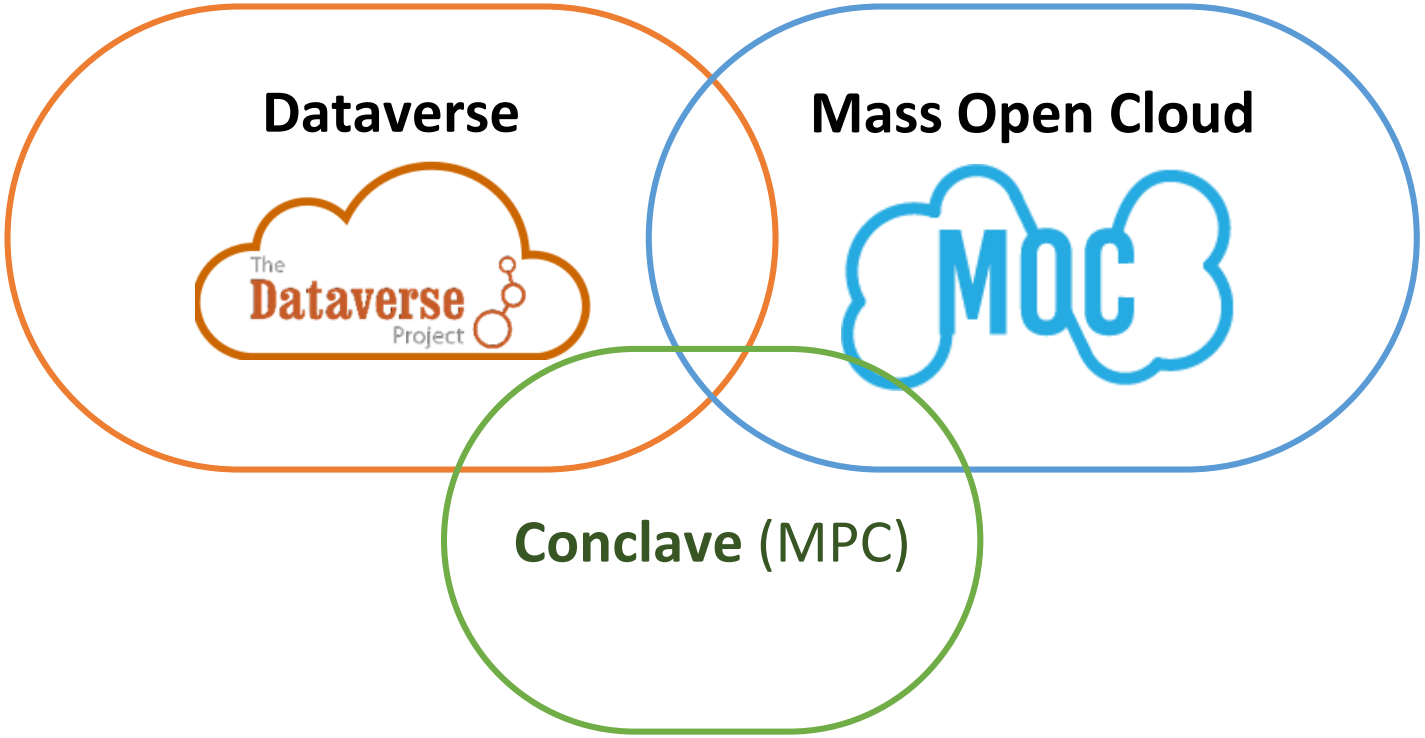
- Multi-vendor public cloud datacenter
- Collaborative effort: 5 universities, government, industry

Privacy-Preserving Scientific Data Analysis in an 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



Valuable

share data → new social insights

Toxic

silo data → safeguard privacy

Images: Facebook, Wikipedia

Valuable

share data → new social insights

and

Toxic

silo data → safeguard privacy

MPC *enables* **secure data analysis** *for* **social good**

Conclave: MPC for relational queries on big data

MPC query compilation from (unannotated) relational queries

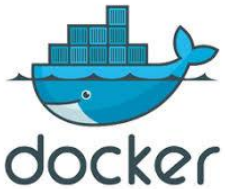
- *Static analysis* to minimize MPC use while maintaining security
- *Trust annotations* to indicate when data sharing in the clear is acceptable for even better performance

Prototype implementation that:

- Connects to existing backend data stacks like Spark and Hadoop
- Scales 4 magnitudes higher than most MPC engines (~100 GB range)

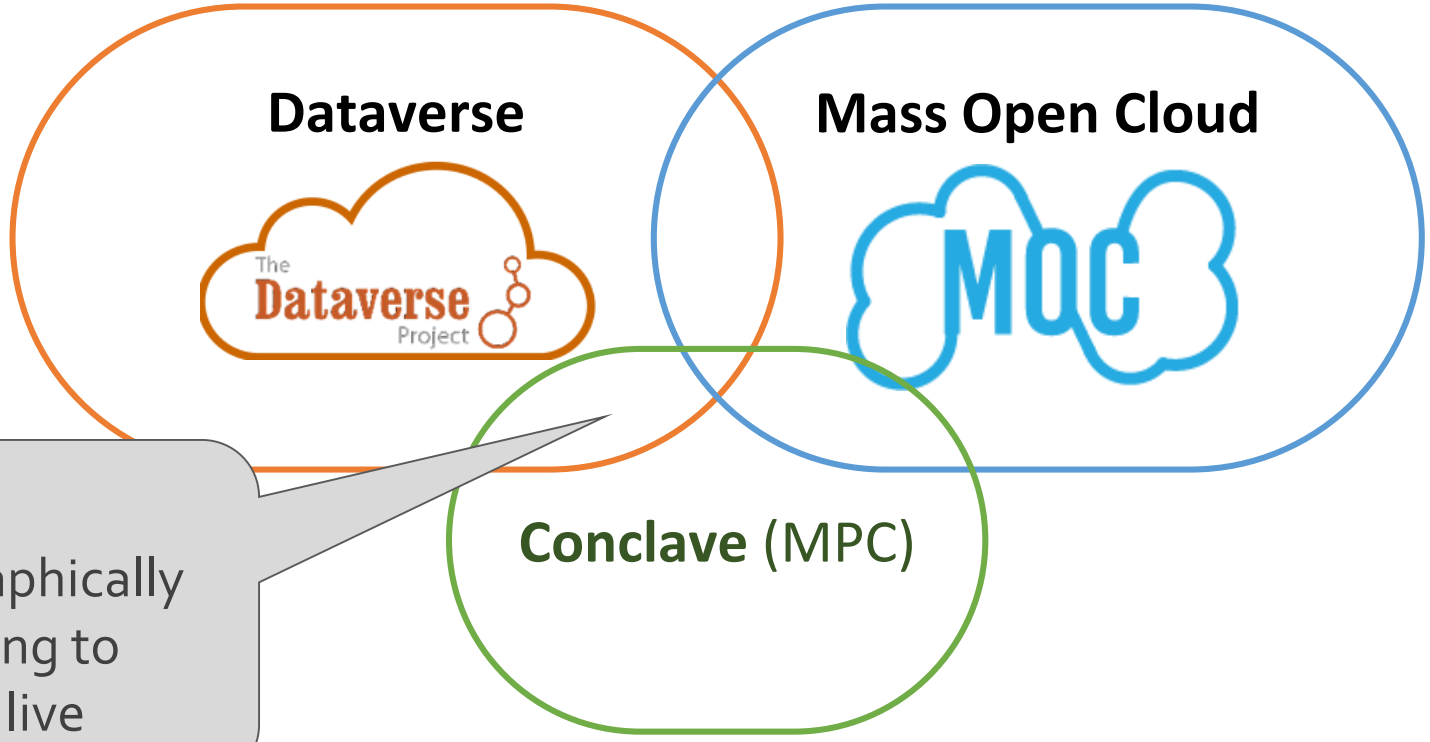
Code at `https://github.com/cici-conclave`

The C2D framework



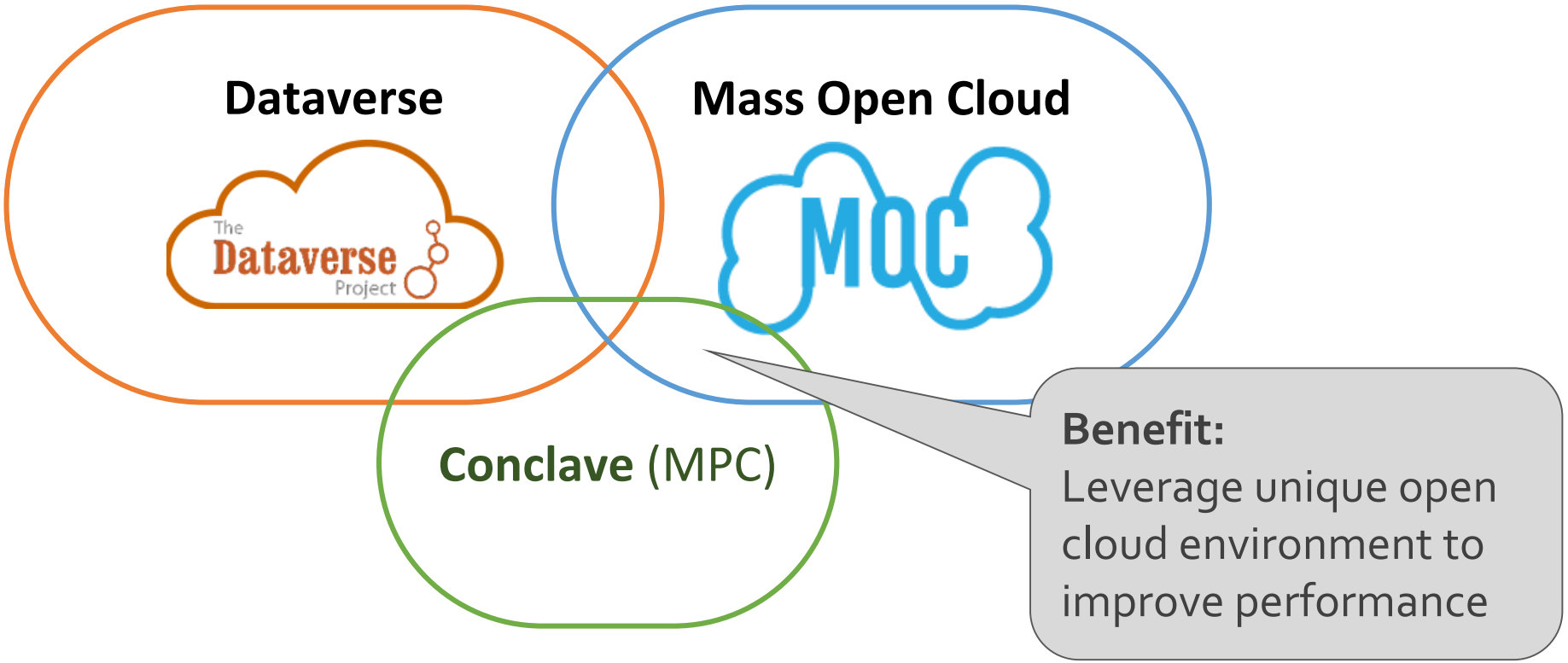
- C2D framework runs on containers
  - Each container stores data owned by a single project
  - Containers never share data with each other
- Built an OpenShift / K8s container orchestration product with
  - In-built job framework
  - Capability to manage slack resources on MOC
- Integrate with Elastic Secure Infrastructure to build trusted secure bare-metal enclaves for parties is ongoing
- Demo video at `https://youtu.be/_vEJmd_rO-0`

Benefits of integration

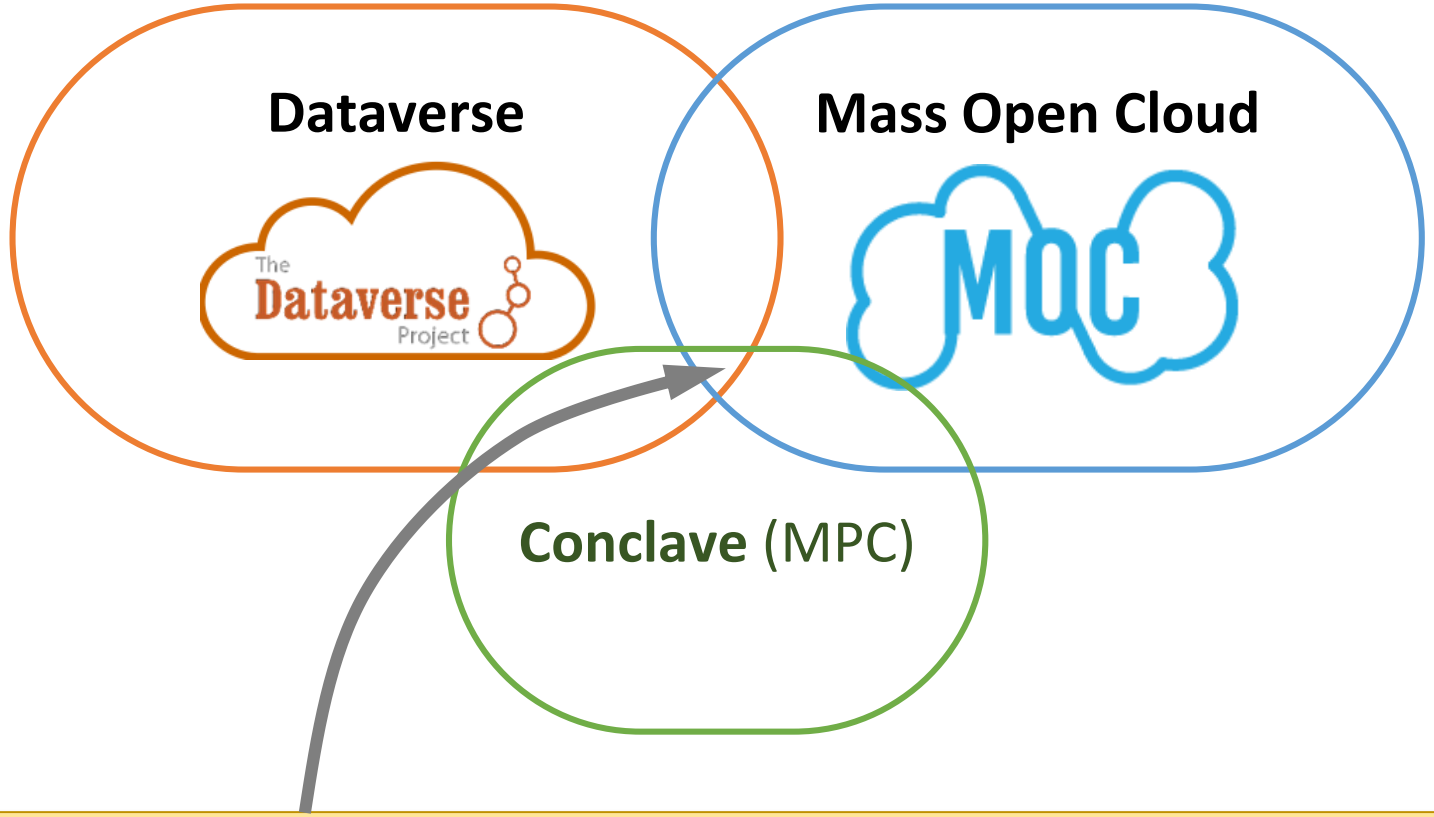


Benefit:
Bring cryptographically secure computing to where the data live

Benefits of integration



Benefits of integration



Synergistic payoff: Separate the responsibilities and amortize the effort of each expert (developers, IT staff, privacy experts, etc.)

Thanks!