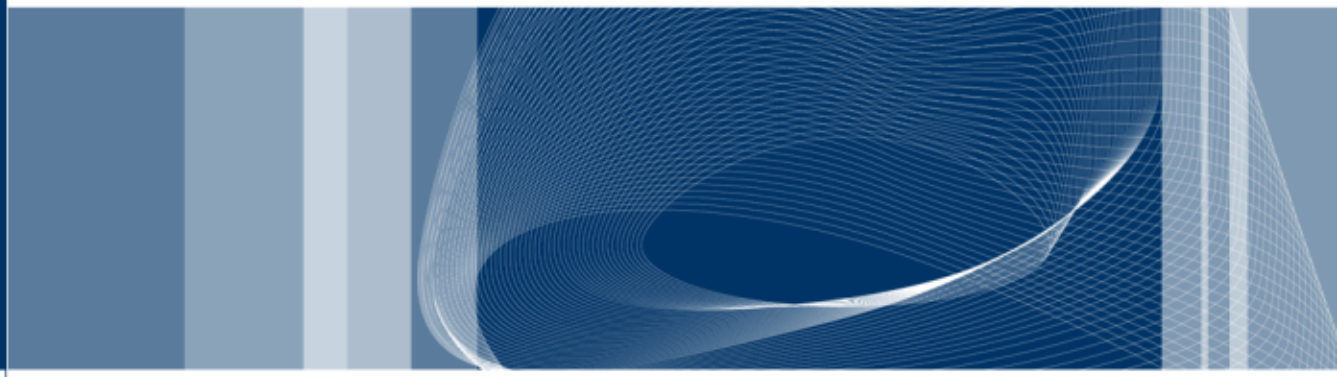
POLITECNICO DI MILANO



Towards DevOps for Privacy-by-Design in Data-Intensive Applications: A Research Roadmap

Context

- Growing interest in Big Data and **data-intensive computing**.
- **Data privacy** is a huge concern!
- Data privacy as a **primary quality aspect** for data-intensive applications.

POLITECNICO DI MILANO

Privacy Assurance

- **Privacy policy**: determining how third parties can access and use data.
- Defining, **monitoring** and enforcing privacy policies.
- Many existing privacy enhacing technologies:
  - Data anonyization
  - Encryption
  - **Attribute-based access control**

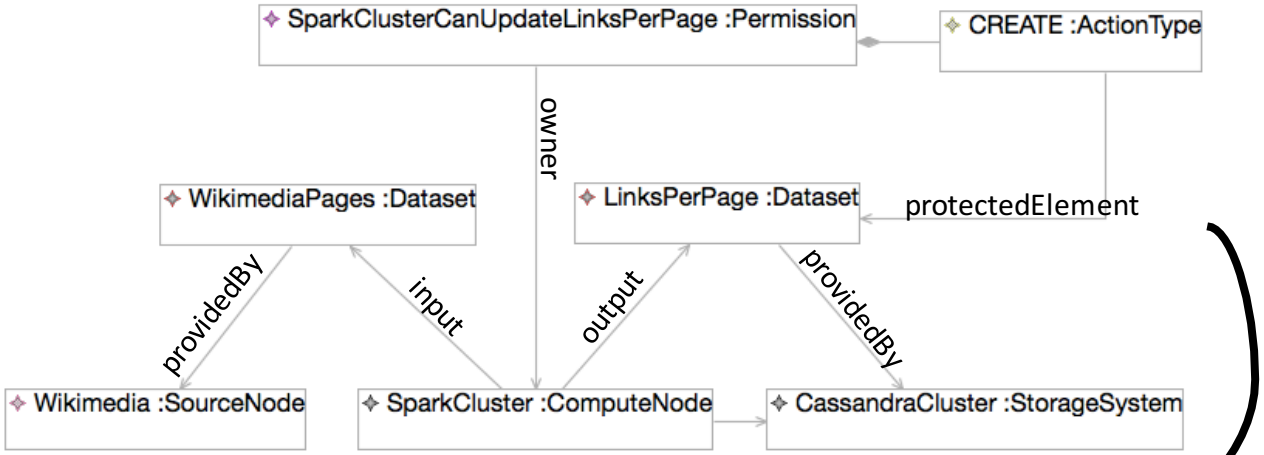
POLITECNICO DI MILANO

Our Solution

- **Tool prototype**: model-driven DevOps trace checking of temporal-based data access policies for data-intensive applications.

POLITECNICO DI MILANO

An Example Scenario

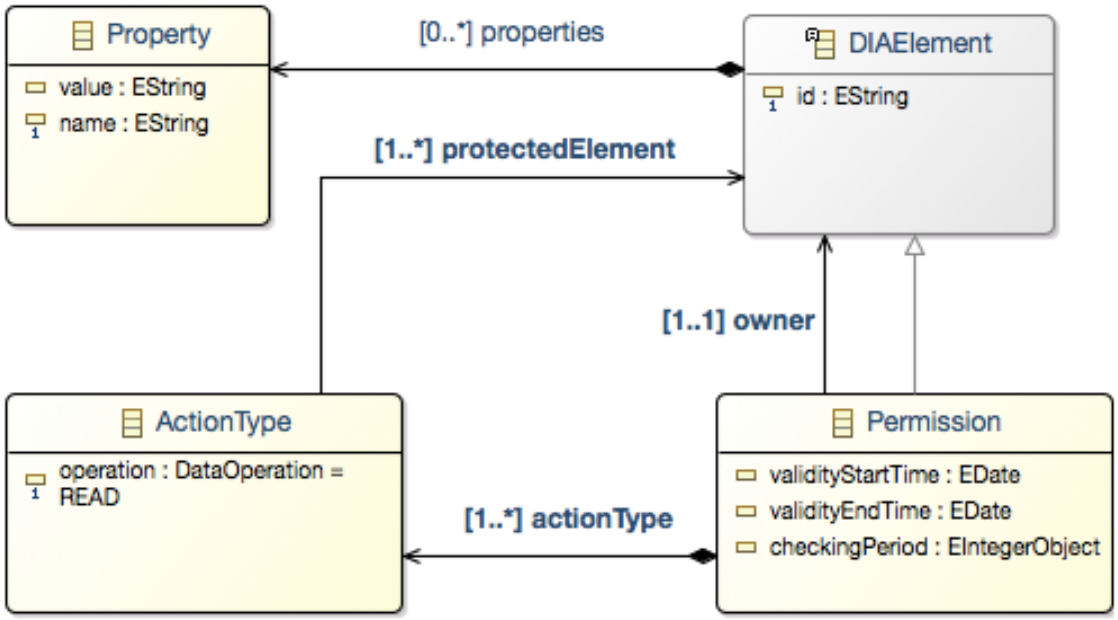


Update(SparkCluster, LinksPerPage, CassandraCluster)
→ $\boldsymbol{P}_{[20,2]}$*START*

```
0  START
2  Elem(SparkCluster)
5  Elem(CassandraCluster)
7  Elem(WikimediaPages)
12 Elem(LinksPerPage)
15 Read(SparkCluster, WikimediaPages, CassandraCluster)
16 Update(SparkCluster, LinksPerPage, CassandraCluster)
22 Update(SparkCluster, LinksPerPage, CassandraCluster)
25 Create(SparkCluster, LinksPerPage,CassandraCluster)
30 Read(SparkCluster, WikimediaPages,CassandraCluster)
```


POLITECNICO DI MILANO

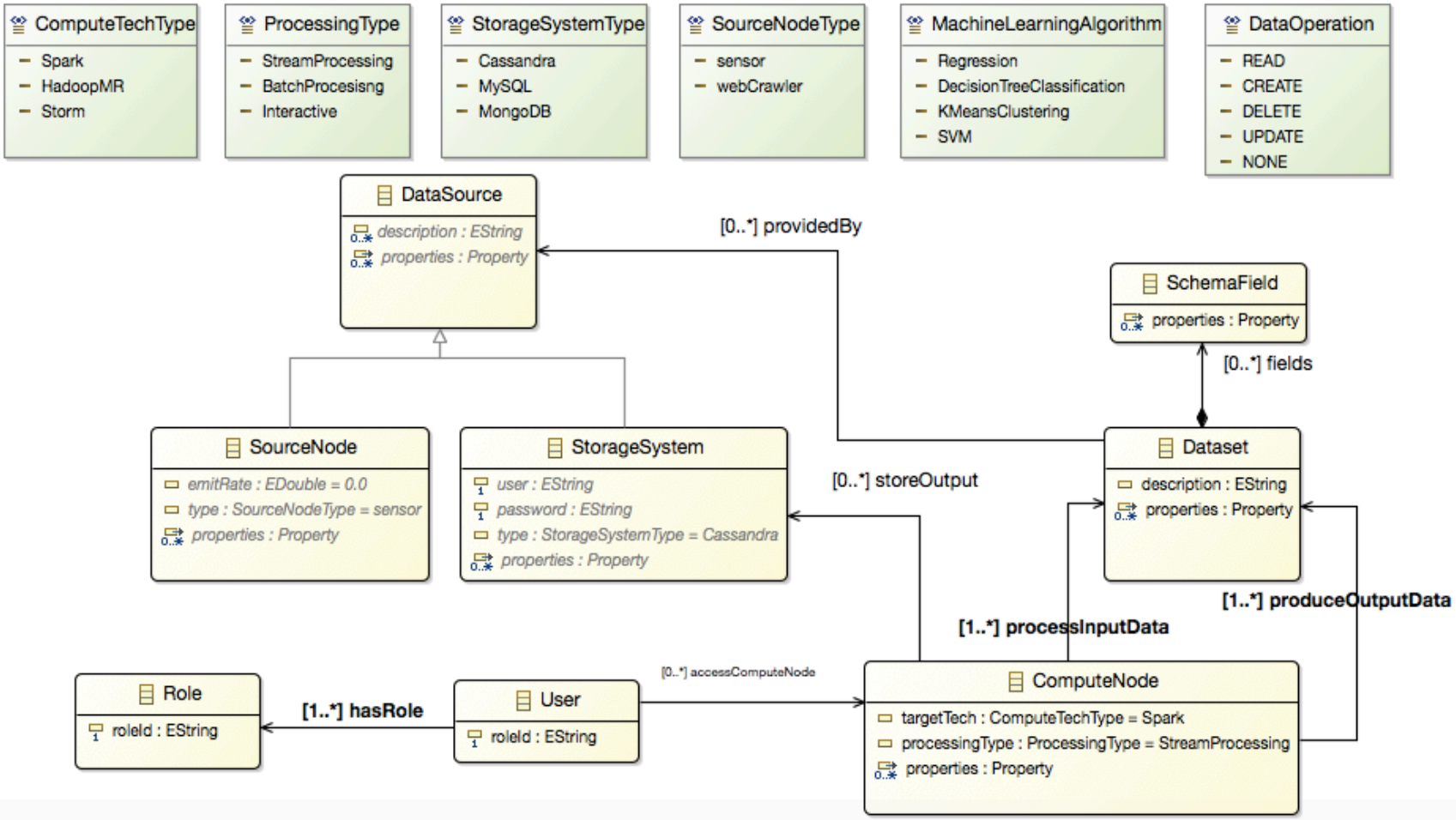
Modeling Language (1)



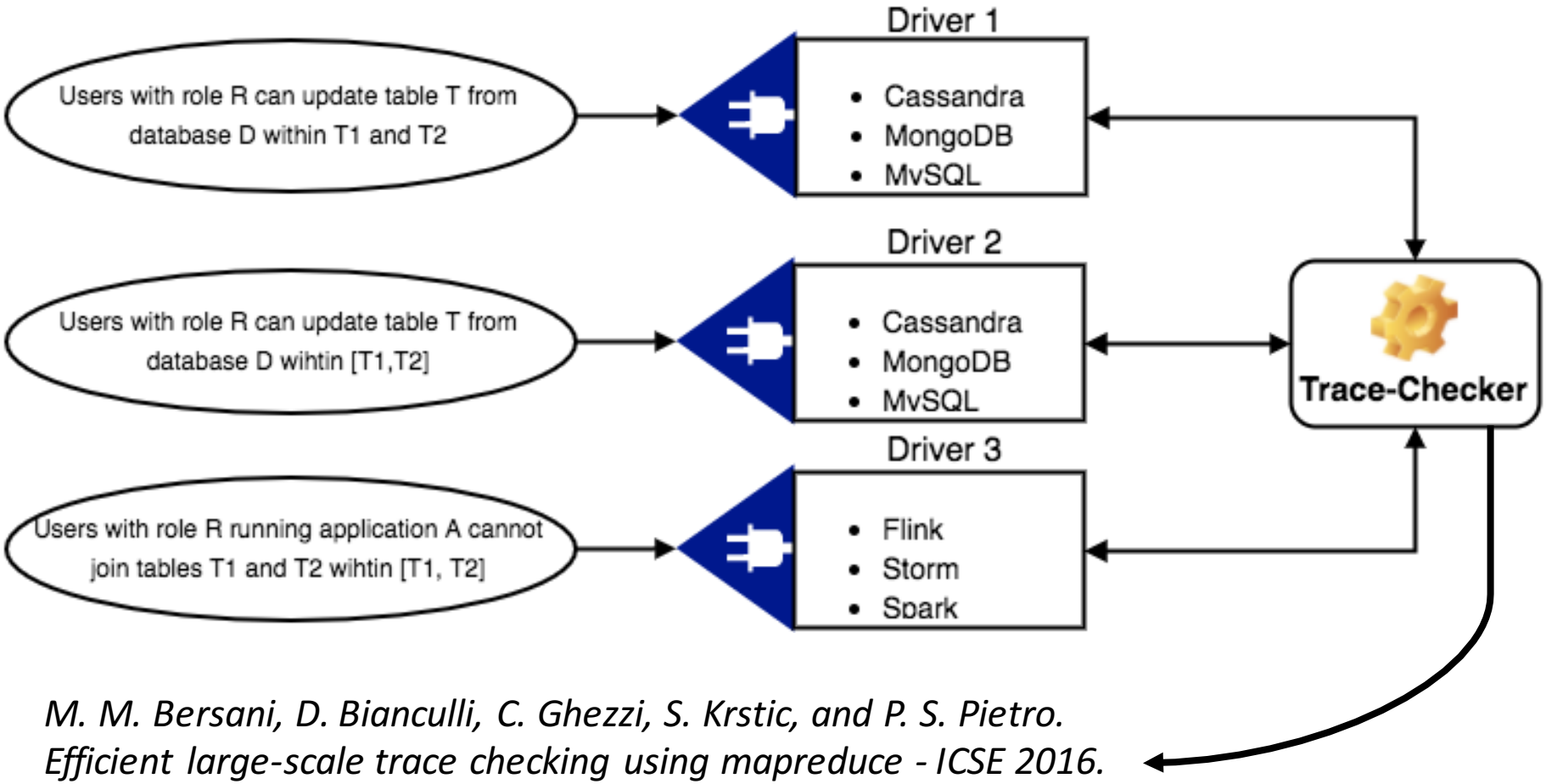
[Ispired by Role Based Access Control and **SecureUML**.]

POLITECNICO DI MILANO

Modeling Language (2)



Trace-Checking Service



M. M. Bersani, D. Bianculli, C. Ghezzi, S. Krstic, and P. S. Pietro. Efficient large-scale trace checking using mapreduce - ICSE 2016.

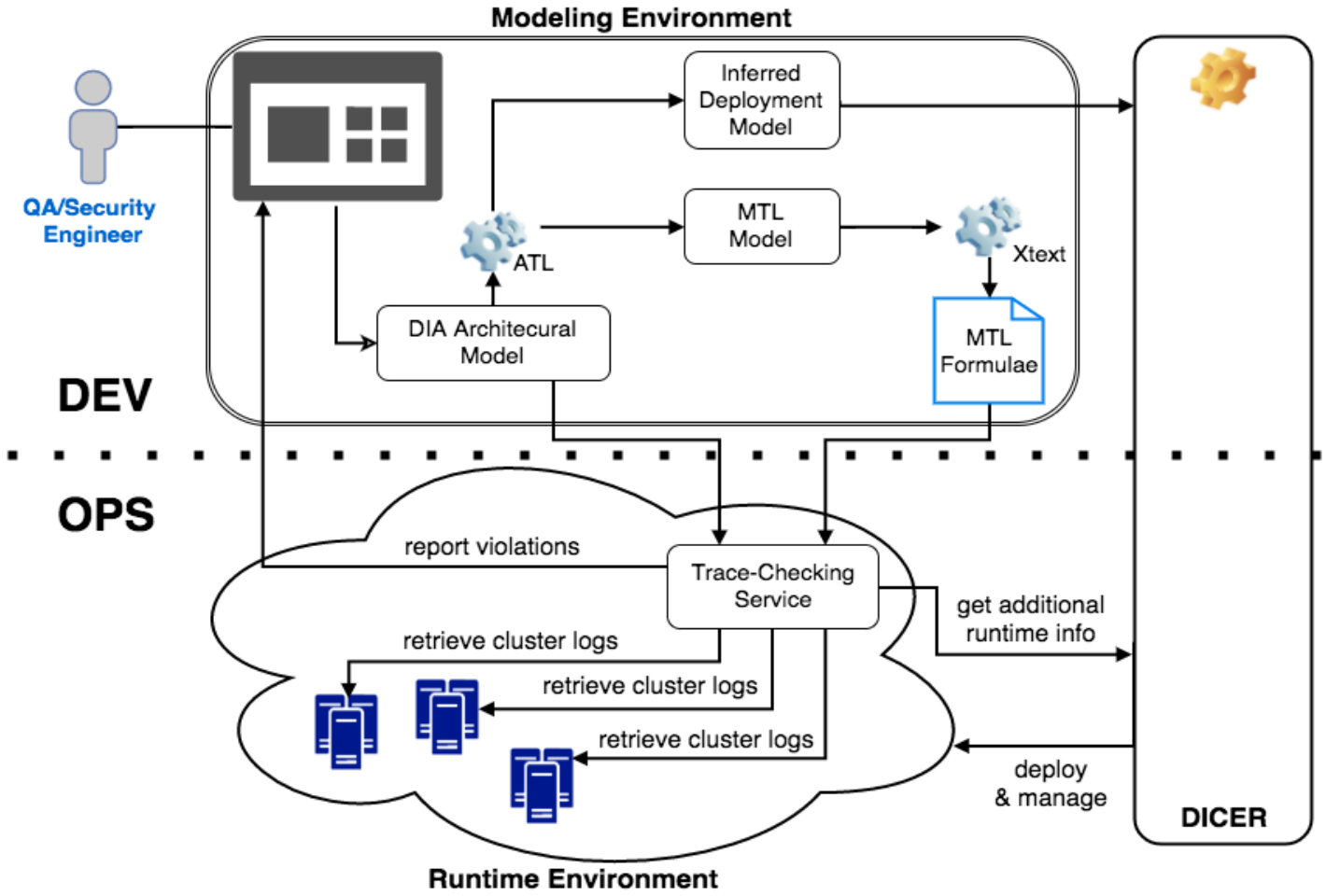
POLITECNICO DI MILANO

Driver for Apache Cassandra

- Able to check permissions of type *"Only within T1 and T2 any application running on cluster X can execute action A on column C of dataset D, where D is stored in Cassandra."*
- Exploit Cassandra **built-in tracing features**.
- Build traces of events by querying Cassandra for the received queries.

POLITECNICO DI MILANO

Prototype Architecture



Threat to validity

- Trace checking Drivers have to provide a lot of **system-level events**.
- Checking privacy violations is not enough! Privacy for big data **must be guaranteed**.
- Are there other approaches rather than periodic trace checking that better fit the problem?

POLITECNICO DI MILANO

Future Work

- Systematic **evaluation** of the proposed solution.
- Towards privacy-aware **UML modeling**: exploit and extend SecureUML to express privacy policies.
- How to **react** to privacy violations?
- **Extended support** for privacy-awareness: continous modeling and deploying of access control and data anonymization policies.

POLITECNICO DI MILANO

Towards a Research Roadmap

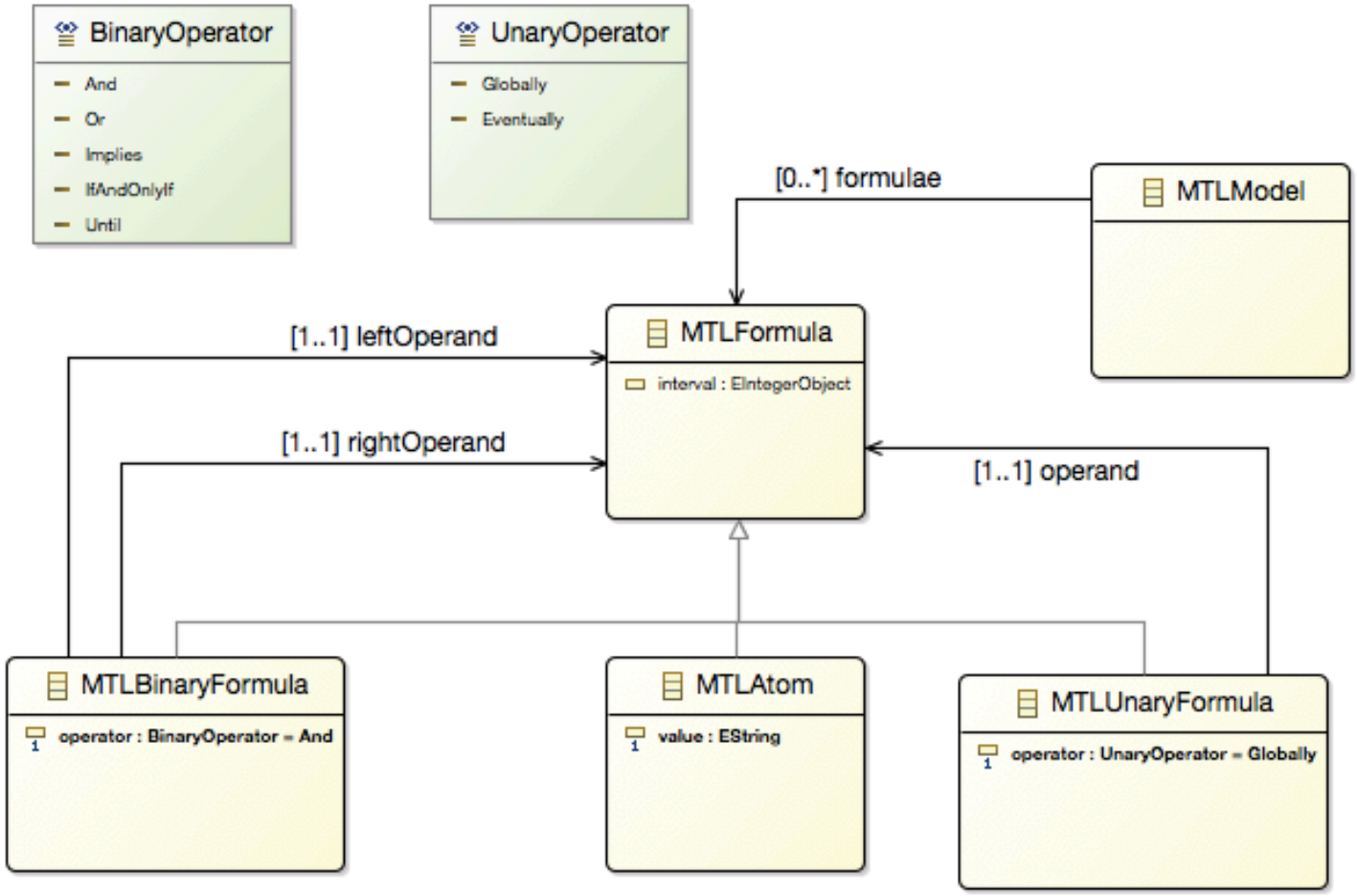
- The problem with privacy in big data is **much bigger** than this… We need to
  1. Understand and model new privacy aspects.
  2. Provide technologies and tools for privacy-by-design (e.g. formal verification of privacy models).
  3. Adapt traditional privacy enhancing technologies to the current technological environment.
  4. Extend big data technologies with built-in mechanisms to support privacy.

POLITECNICO DI MILANO

Q&A

POLITECNICO DI MILANO

MTL Metamodel



POLITECNICO DI MILANO

Recent Developments

- Complicated driver-based implementation.
- Simplify the trace-checking service and gain in flexibility at the expense of applications trasparency.
- **Application instrumentation** with tracing features. No need for specific drivers to retrieve traces.
- Applications properly log events about their data accesses over which MTL formulae can be directly verified.

POLITECNICO DI MILANO

Conclusion

- A DevOps tool prototype to support the model-driven trace checking of privacy-aware data-intensive applications.
- Preliminary experimentation in industry looks promising.
- Definition of a research roadmap towards guaranteeing privacy in big data.

POLITECNICO DI MILANO