Towards a computer-interpretable actionable formal model to encode data governance rules

Rui Zhao & Malcolm Atkinson

Centre for Intelligent Systems and their Applications
School of Informatics
University of Edinburgh

24 September 2019

Outline

Data-sharing collaboration

Data-sharing collaboration

Data-sharing collaboration

Current practice

Data-sharing and more:

- Data
- Method
- Code
- (Re-)Execution

Current practice

Data-sharing and more:

- Data
- Method
- Code
- (Re-)Execution

Di erent organizational forms:

- Centralized
  - Single-institutional
  - Cross-institutional
- Federated

Current practice

Data-sharing and more:

- Data
- Method
- Code
- (Re-)Execution

Di erent organizational forms:

- Centralized
  - Single-institutional
  - Cross-institutional
- Federated

Observed problems

- Shredded
  - Hard to trace back
  - Hard to receive update
  - Hard to reuse
- \Random" data use
  - Unintended data leaks
  - Violating T&C from the data providers

Current solution... of shredded

- Work ow + standardized description language (e.g. CWL)
- Provenance record
- Research portal / scienti c gateway

Current solution... of "random" data use

- Sensitive data
  - Restricted environment
  - Extra training
- Non-sensitive data
  - Very basic or no policy (open data)

Current~~solution~~... of\random" data use

- Sensitive data
  - Restricted environment
  - Extra training
  - Long rigorous approval procedure
- Many stages & data in between middle ground
  - No systematic support available
- Non-sensitive data
  - Very basic or no policy (open data)
  - Hidden rules

) Polarized, wasted e ort and accidents

Re-think the big picture

Proposed strategy

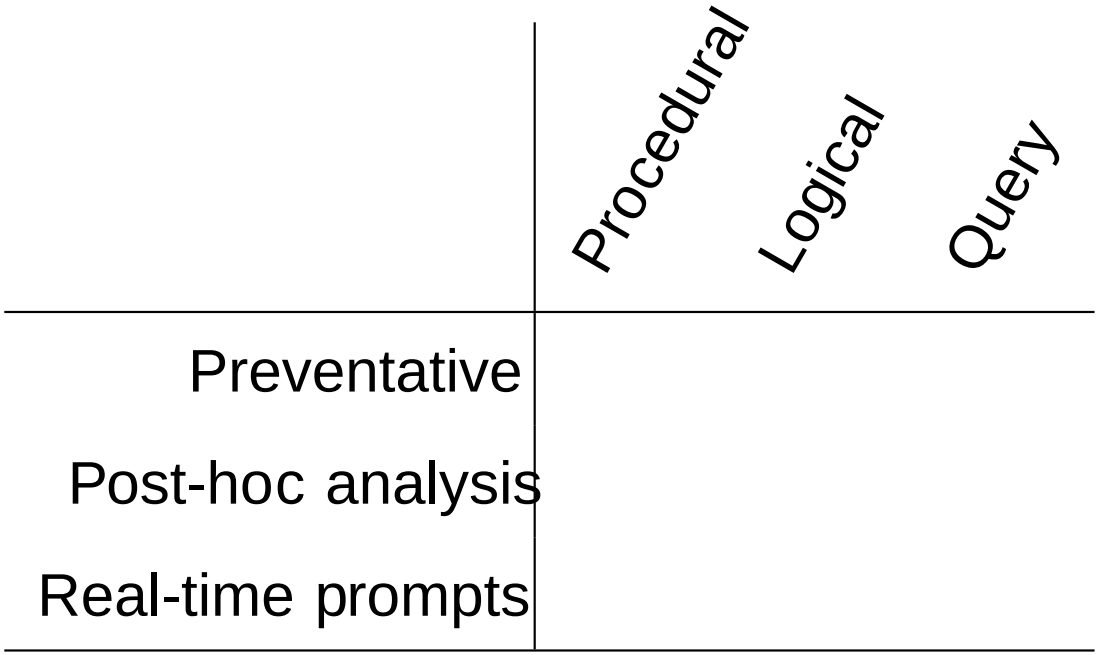
- Encode data-use rules
- Explicit
  - for humans
  - for computers
- Precise
- Computer-aided compliance

Formal model for data policies (i.e. governance rules)
+
Supporting framework

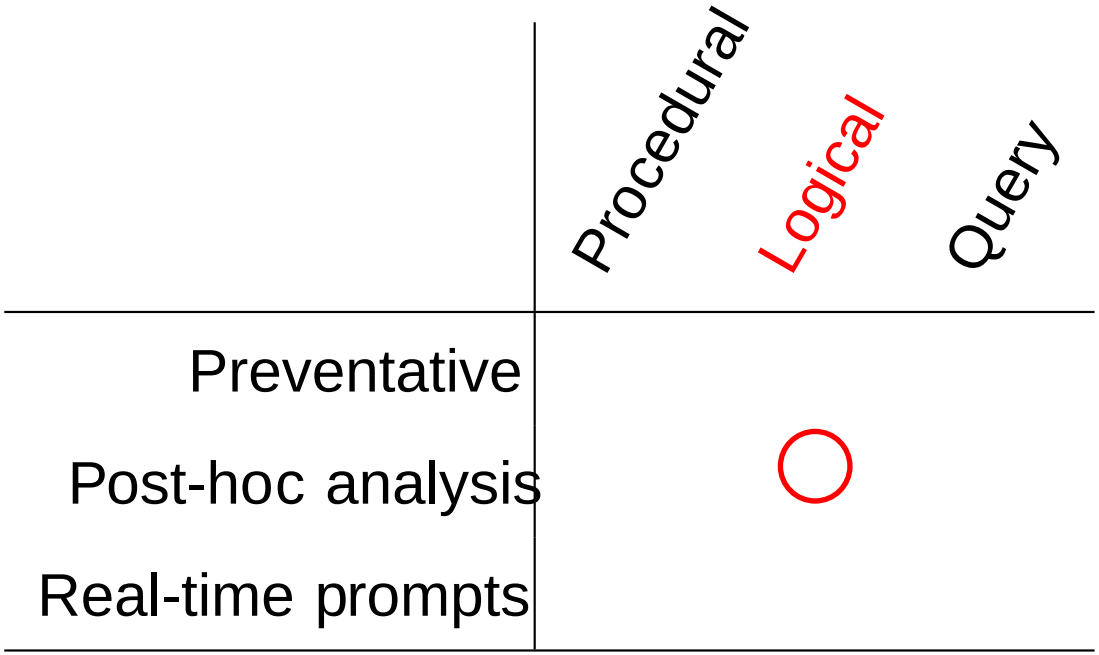
Language model requirements

- Human comprehensible
  - Easy to author
  - Easy to understand
- Computer interpretable
  - (Critical)
- Transformation needed?

Solution space



Solution space



Let rules flow with data and change with processing

Use ontologies for interoperability

- Ontologies for di erent purposes
  - Obligation type
  - Attribute class
  - Activation condition
- Rule propagation control

Provenance as input

Why provenance?

- History
- Standardized
- Extensible
- Interoperable

Provenance provides

Integrated information flow from systems, software and tools

Working rule compliance checker

- Extend datasets: arrive with data rules
- Processes have ow rules
- Intermediate & output datasets have inferred rules

Example: simple yet complicated

Data providers often want data users to...

> Account data use: \Report the number of times you use our data."

Example: simple yet complicated

> \Report the number of times you use our data."

De ne use:

- every timeas input to a process
- when asinitial input
- when resulting in æsuccessful result

De ne report time:

- whenever a new use occurs
- when nished

Example encoding

Report the number of times you use our data every time as input to a process

```
Obligation(:report :source, :source, :WhenAsInput)
Attribute(:source, "some source")
```

Report the number of times you use our data when initial input

```
Obligation(:report :source, :source, :WhenImported)
Attribute(:source, "some source")
```

Flow Rule

Merge

Activation

Future work

- Improve logical foundation
- More rule forms
  - wider range of use cases
- Supporting systems
  - operational integration & tools

Question time & Acknowledgement

Questions?

Special thanks to Christian Page, Wim Som de Cer and Luca Trani who replied to our initial survey, to Alessandro Spinuso who provided testing provenance database (and helped the survey) and to everyone who provided help.