What’s in a (Data) Type? Meaningful Type Safety for Data Science

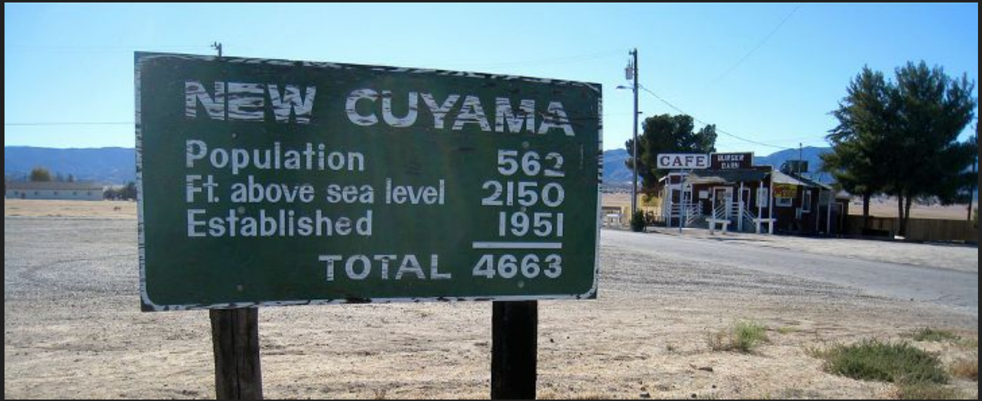
Riley Moher

06.12.2021

Data-Driven Decision Making Lab
University of Toronto

What's in a (Data) Type?

- Data Scientists work with Data that represent real-world concepts
- Important decisions are made based on data scientists' results
- The nuances of these concepts are reduced to integers, strings, etc



Overview

1. **How datatypes fail to typify data**

2. **Why current solutions don’t work**

3. **A framework for meaningful type safety**

Overview

1. **How datatypes fail to typify data**
   I. Mereology
   II. Time
   III. Provenance

2. **Why current solutions don’t work**

3. **A framework for meaningful type safety**

Mereological Troubles

- Mereology is part-whole relation
- Legs are part of a table, Toronto is part of Ontario
- Definition of ‘eligible’ COVID vaccine population changes, comprises new age groups
- Mereology not formally defined within the data, needs to be integrated manually

Time Complications

- Time is a common and important factor in most data

- Data is observed, collected, updated at specific timepoints

- Time is an additional layer of complexity, mereology changes over time
    - Toronto 1985 vs Toronto 2021

- Manual intervention still necessary, time is only values, no reasoning being done

Provenance

- As data is changed, so are the concepts it represents
  - Population vs average population
  - Density = Mass / Volume

- Take Physical Quantities, Units as an example
  - Most approaches just enforce same units and conversion

- Bob’s Height + Mary’s Height : What is this quantity?

Overview

1. How datatypes fail to typify data

2. Why current solutions don't work

   I. Improving Documentation

   II. Data Provenance

   III. Knowledge Modelling

3. A framework for meaningful type safety

Overview

1. How datatypes fail to typify data
2. Why current solutions don't work
   1. Improving Documentation
   2. Data Provenance
   3. Knowledge Modelling
3. A framework for meaningful type safety

Developing Decent Dictionaries

- To better understand data, we can create better documentation standards
  - A meaningful list of questions to be answered about a given dataset

Motivation	Composition	Collection Process	Maintenance
...?	…?	…?	…?

- We have a more complete picture of the dataset, however:
  - Description is still in natural language
  - Description is static
  - Description is not machine readable

Overview

1. How datatypes fail to typify data

2. Why current solutions don't work

    I. Improving Documentation

    II. Data Provenance

    III. Knowledge Modelling

3. A framework for meaningful type safety

Prospering with Provenance?

- To avoid confusion about semantics, keep track of how our data changes

- Provenance is mainly discussed in two forms:
  - Lineage: What is the data’s history of operations?
  - Where-provenance: What data sources were combined to arrive here?

- Provenance can give us additional info, however:
  - Provenance information won’t warn us of potential errors
  - Provenance information doesn’t ensure initial understanding

Overview

1. How datatypes fail to typify data

2. Why current solutions don't work

   I. Improving Documentation

   II. Data Provenance

   III. Knowledge Modelling

3. A framework for meaningful type safety

Opportunity for Ontologies?

- Use an ontology as an interlingua for interoperability
  - Allows us to define one ontology and map others to it
  - Requires knowledge modeling experts to maintain

- Ontology Oriented Programming
  - Ontologies integrated into programming languages
  - These tools are not very mature and unstable

- Actual integration varies widely between disciplines & software tools

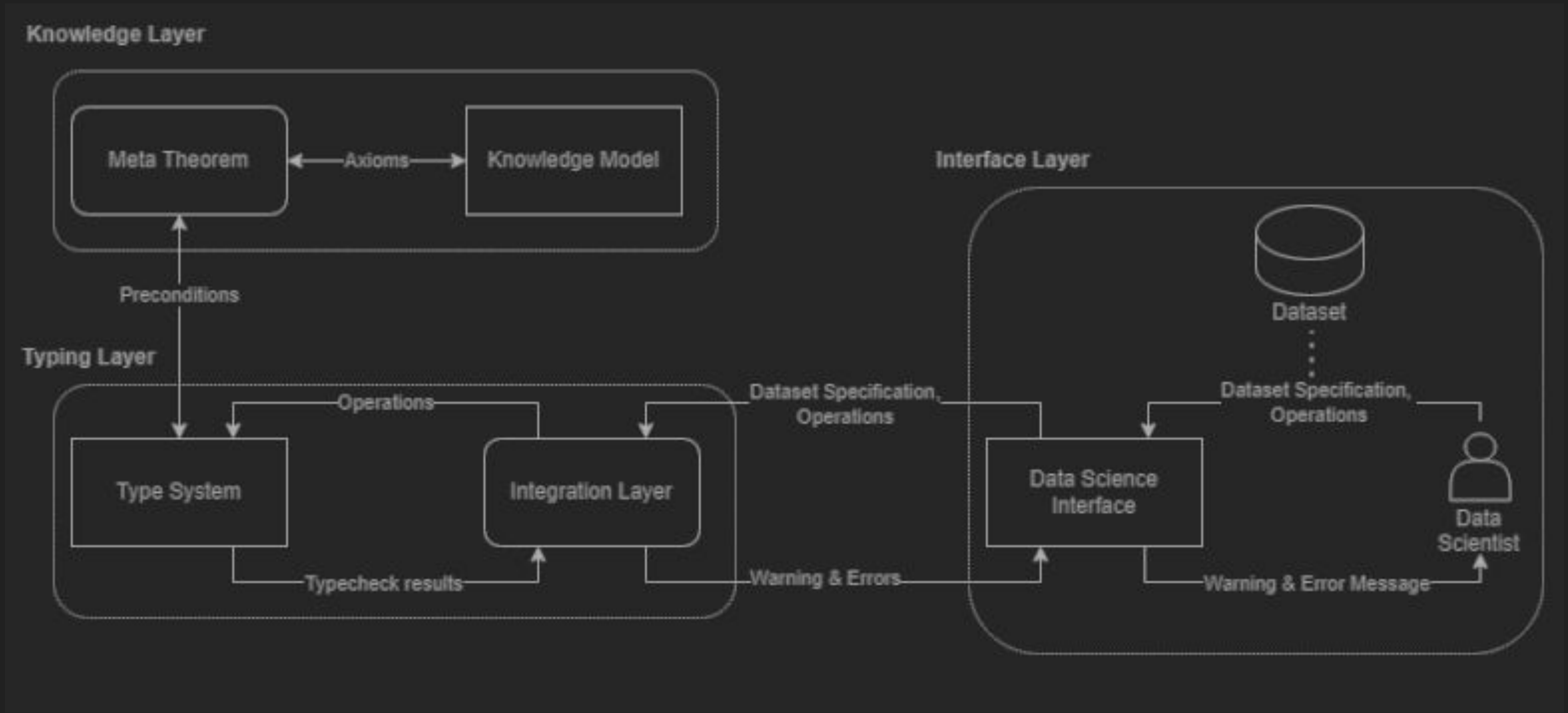
Overview

1. How datatypes fail to typify data
2. Why current solutions don’t work
3. A framework for meaningful type safety
    I. Architecture
    II. Knowledge Layer
    III. Typing Layer
    IV. Interface Layer

Overview

1. How datatypes fail to typify data
2. Why current solutions don't work
3. **A framework for meaningful type safety**
   I. **Architecture**
   II. Knowledge Layer
   III. Typing Layer
   IV. Interface Layer

Framework Architecture



Overview

1. How datatypes fail to typify data
2. Why current solutions don't work
3. **A framework for meaningful type safety**
    I. Architecture
    II. **Knowledge Layer**
    III. Typing Layer
    IV. Interface Layer

Knowledge Layer

- Formal model of concepts represented in the dataset
- Correspondence between program and logic
- Provides justification for modelling decisions, separate ontological commitments from implementation

Overview

1. How datatypes fail to typify data
2. Why current solutions don’t work
3. **A framework for meaningful type safety**
   I. Architecture
   II. Knowledge Layer
   III. **Typing Layer**
   IV. Interface Layer

Typing Layer

- Leverage Dependent Typing
  - Construct types which depend on values

- Tuples (m,n) where m < n

$$\sum_{m:\mathbb{N}} \sum_{n:\mathbb{N}} ((m < n) = True)$$

- The type of operations can enforce pre-conditions, post-conditions

Meaningful Types

- Operations enforce relationships between their operands

- Values change after each operation : Provenance-Integrated
    - Averaging populations produces an “Average Population”

```
Plus : List Disjoint Populations -> Population
…

Subtraction : Pop1, Pop2 BothSameKind -> Population
```

Overview

1. How datatypes fail to typify data
2. Why current solutions don’t work
3. **A framework for meaningful type safety**
   I. Architecture
   II. Knowledge Layer
   III. Typing Layer
   IV. **Interface Layer**

Interface Layer

- Data Scientists should not need to adopt whole new skillsets

- Logic-Based Type System is integrated into data science tools
  - Pandas: meaningful types library
  - Tableau: meaningful types plugin

- Data Scientist specifies the concepts contained in the data
  - Small additional work upfront will pay dividends

Forward-Looking Thoughts

- A complete framework is a big piece of work
- Data Scientists, Type Theorists, and Knowledge Modellers can learn from each other
    - Bridging the gap enriches us all
- Meaningful Data Science is Important
    - Reduce bias
    - Make informed decisions
    - Save lives