Meaningful Datatypes: Ontologically-Sound Dependent Type Systems for Data Science

Riley Moher
21.04.2022
University of Toronto

The Role of Data Science

- Data Science is increasingly important and valuable
- It drives important decisions
- Understanding data is critical

What’s in a (Data) Type?

- All data is representative of real-world phenomena
- Important decisions are made based on data scientists’ results
- The nuances of the real world are reduced to integers, strings, etc

Outline

1. Why are datatypes problematic for data science?

2. What are my contributions to solving this problem?

3. What is the significance of these contributions?

4. What are the directions for future work?

Outline

1. **Why are datatypes problematic for data science?**
2. What are my contributions to solving this problem?
3. What is the significance of these contributions?
4. What are the directions for future work?

Outline

1. **Why are datatypes problematic for data science?**
   I. **Datatype Problem Classes**
   II. **The gaps in current work**

2. What are my contributions to solving this problem?

3. What is the significance of these contributions?

4. What are the next steps?

Outline

1. Why are datatypes problematic for data science?
   I. **Datatype Problem Classes**
   II. The gaps in current work
2. What are my contributions to solving this problem?
3. What is the significance of these contributions?
4. What are the directions for future work?

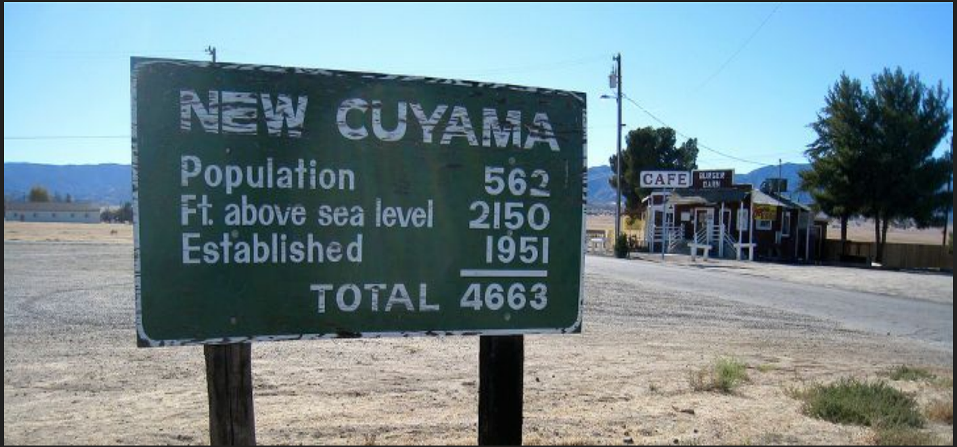
Datatype Problems

The Real World

- Implicit rules
- Complex concepts and relationships

Simple Datatypes

- Numbers are numbers*
- String OR Integer OR Float OR …
- Same datatype represents many different concepts



Datatype Problem Classes

1. Time

2. Mereology

3. Provenance

Datatype Problem Classes

1. **Time**
2. Mereology
3. Provenance

Datatype Issues: Time

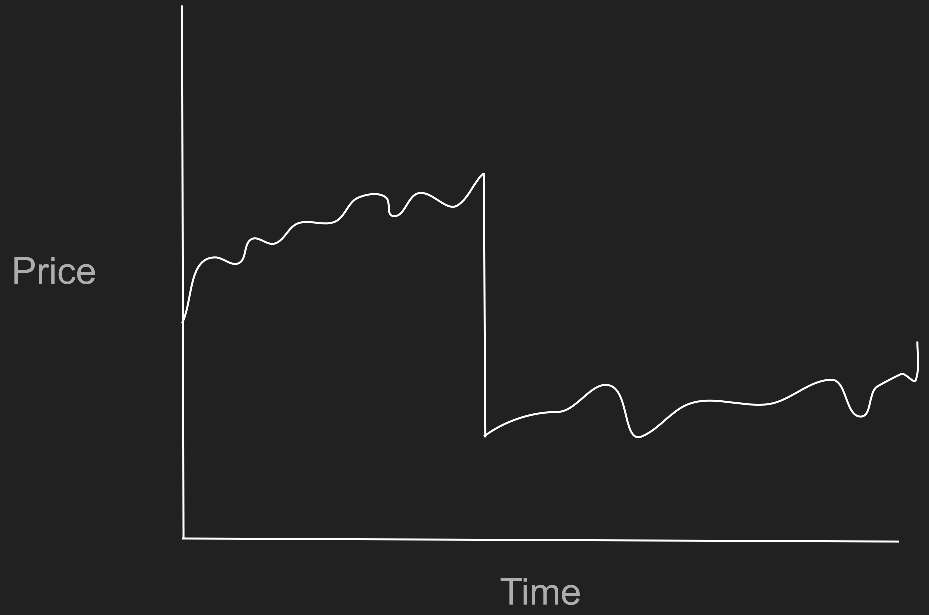
- Time is common and important in most data

- Ex: House Prices

Date	Average House Price
1960-05-20	12 000
…	…
2022-01-01	850 000

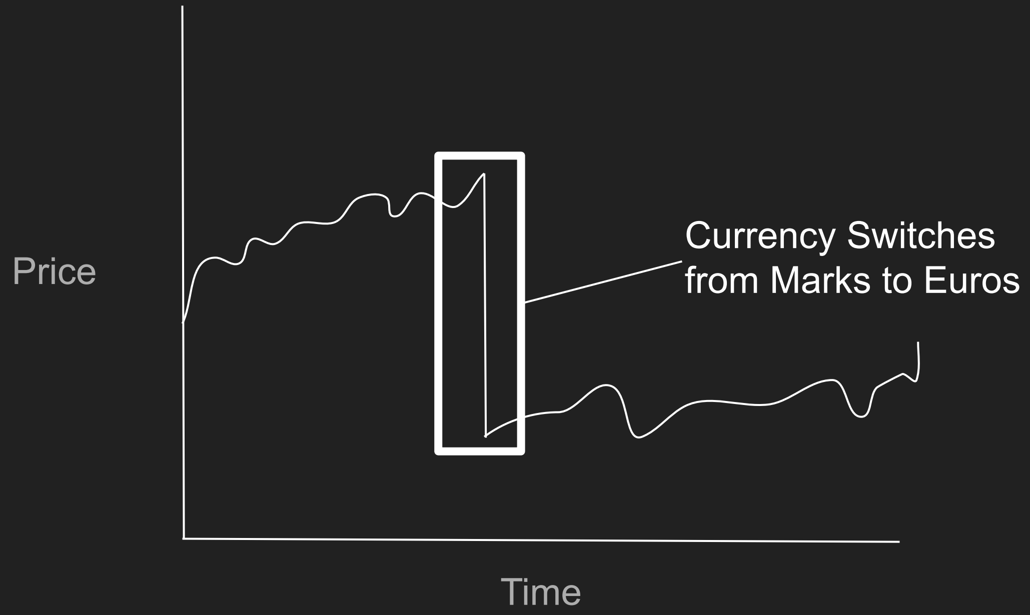
Datatype Issues: Time

- Ex: Frankfurt Stock Exchange Quote



Datatype Issues: Time

- Ex: Frankfurt Stock Exchange Quote



Datatype Problem Classes

1. Time
2. **Mereology**
3. Provenance

Datatype Issues: Mereology

- Mereology in data exists in many different forms

- Ex: COVID-19 Vaccination Data

Date	**Fully Vaccinated**	**% of Eligible Population Fully Vaccinated**
2021-06-22	1 200 000	3.5%
…	…	…
2021-07-29	1 335 000	2.6%
…	…	…
2022-05-28	1 086 100	2.2%

Datatype Issues: Mereology

- Mereology in data exists in many different forms

- Ex: COVID-19 Vaccination Data

Date	Fully Vaccinated	% of Eligible Population Fully Vaccinated
2021-06-22	1 200 000	3.5%
Children Become Part of Eligible Population		…
2021-07-29	1 335 000	2.6%
…	…	…
2022-05-28	1 086 100	2.2%

Datatype Issues: Mereology

- Mereology in data exists in many different forms

- Ex: COVID-19 Vaccination Data

Date	Fully Vaccinated	% of Eligible Population Fully Vaccinated
2021-06-22	1 200 000	3.5%
…	…	…
2021-07-29	1 335 000	2.6%
“Fully Vaccinated” definition changes to 3+ doses		…
2022-05-28	1 086 100	2.2%

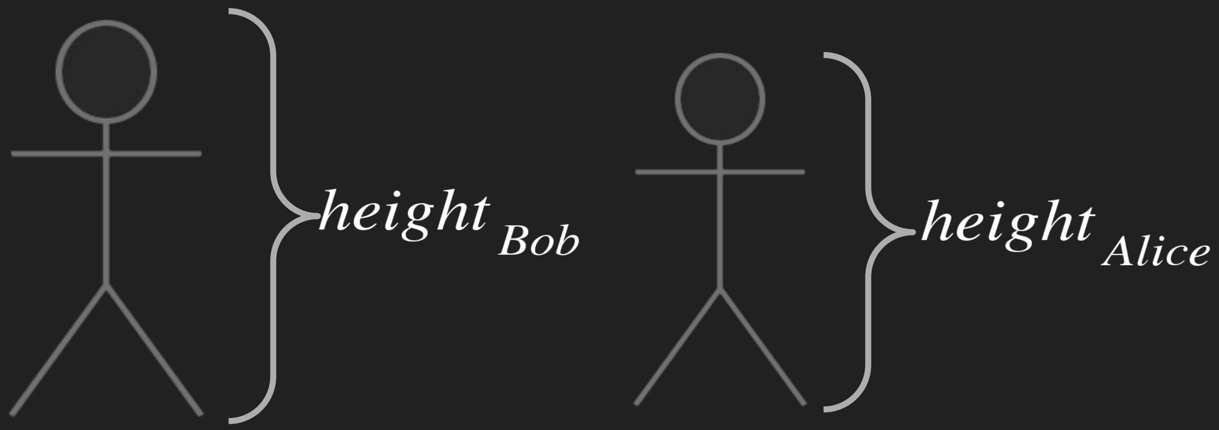
Datatype Problem Classes

1. Time
2. Mereology
3. **Provenance**

Datatype Issues: Provenance

- Data science operations transform real-world interpretations

- Ex: Height Measurements



Datatype Issues: Provenance

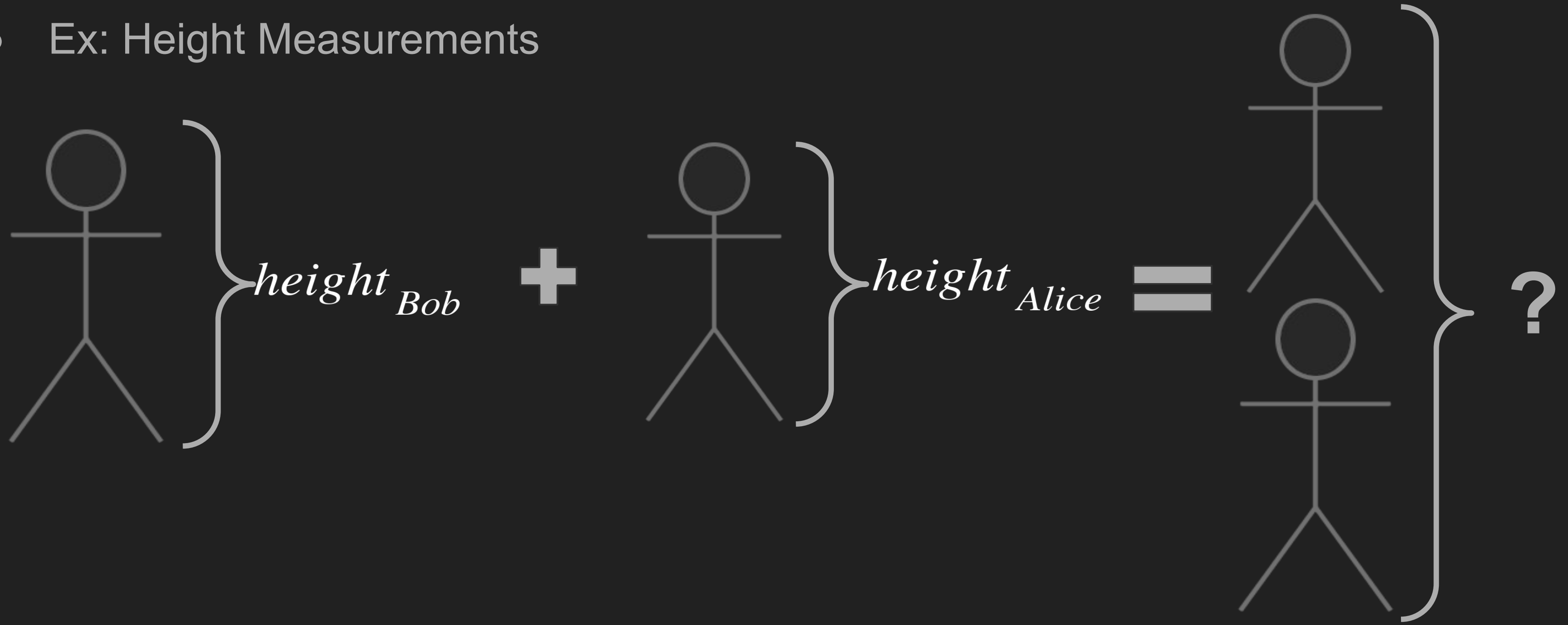
- Data science operations transform real-world interpretations

- Ex: Height Measurements

Datatype Issues: Provenance

- Data science operations transform real-world interpretations

- Ex: Height Measurements



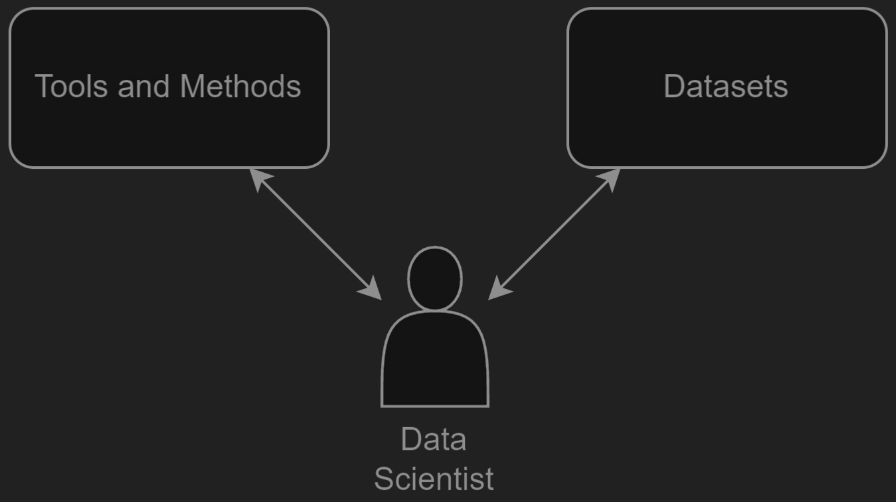
Outline

1. Why are datatypes problematic for data science?
   I. Datatype Problem Classes
   II. The gaps in current work
2. What are my contributions to solving this problem?
3. What is the significance of these contributions?
4. What are the next steps?

Datatype Issues: Current Solutions

- These approaches supplement datatypes with external knowledge and tools

- Applied in informal, ad-hoc, opaque, laborious ways



Datatype Issues: Current Solutions

1. Documentation

2. Provenance Tracking

3. Knowledge Representation

Datatype Issues: Current Solutions

1. **Documentation**

2. Provenance Tracking

3. Knowledge Representation

Documentation Standards

- Understand data through documentation standards
  - Provide list of important questions to be answered about the dataset

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

Documentation Standards

- Understand data through documentation standards
  - Provide list of important questions to be answered about the 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

Datatype Issues: Current Solutions

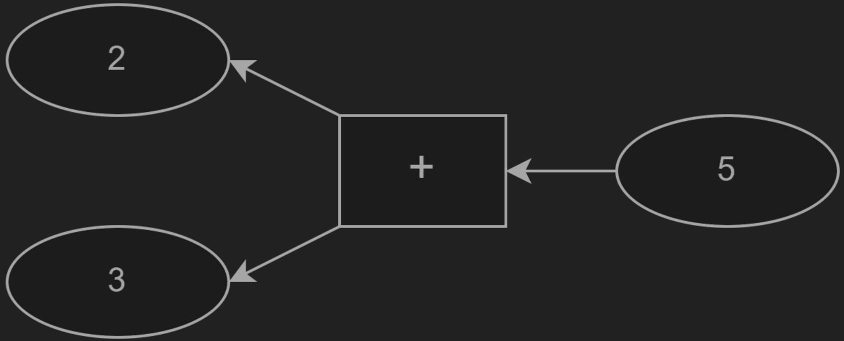
1. Documentation

2. **Provenance Tracking**

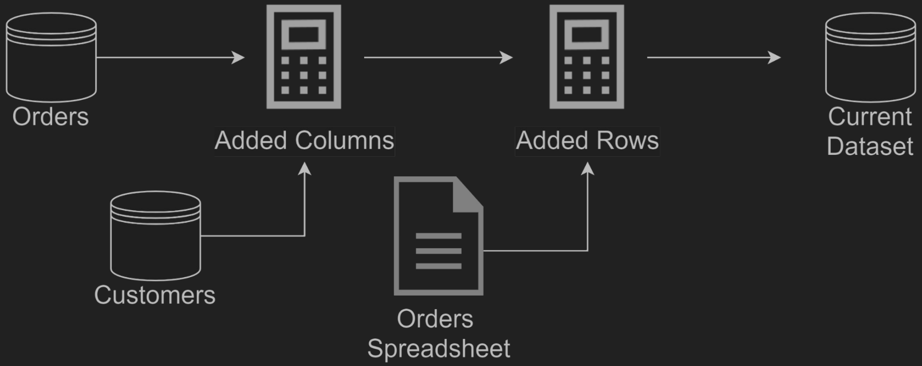
3. Knowledge Representation

Provenance Tracking

- Lineage-Provenance (What, How?)

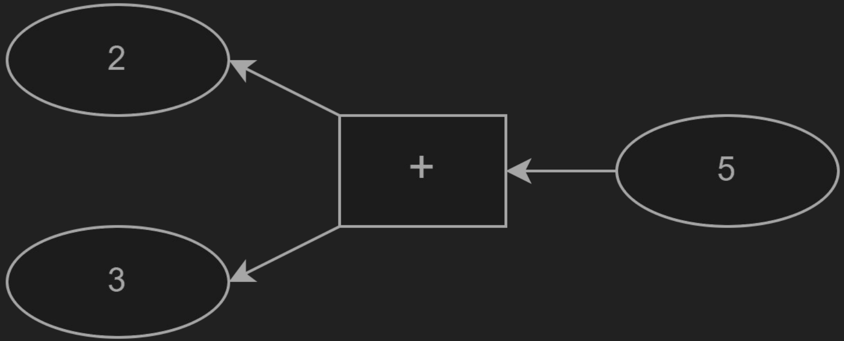


- Where-Provenance (Where From?)

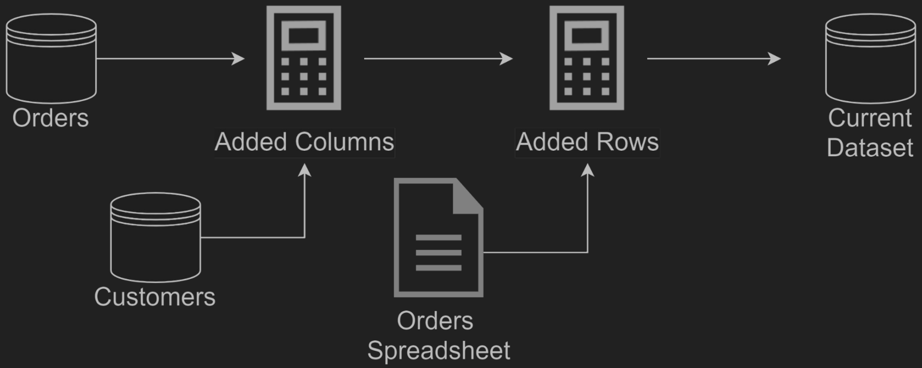


Provenance Tracking

- Lineage-Provenance (What, How?)



- Where-Provenance (Where From?)



- No automatic error detection
- Does not encode real-world semantics
- Human verification still necessary

Datatype Issues: Current Solutions

1. Documentation

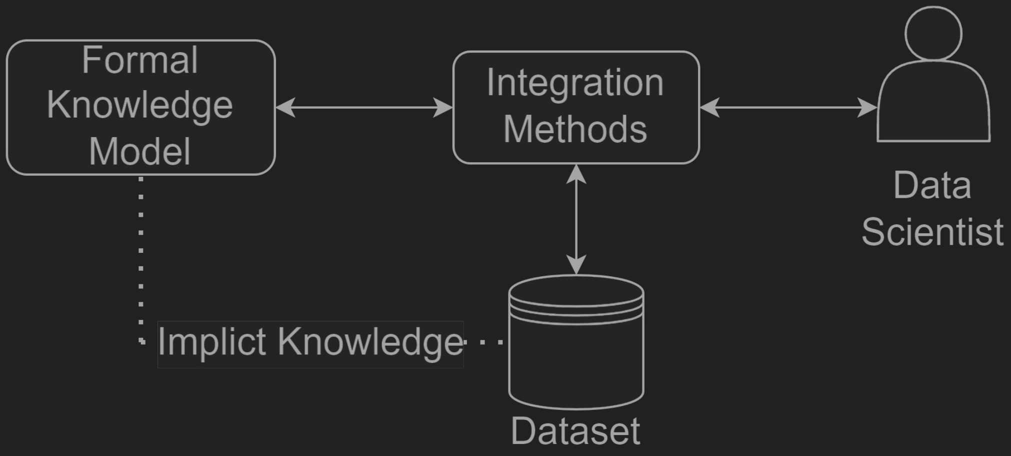
2. Provenance Tracking

3. **Knowledge Representation**

Knowledge Representation

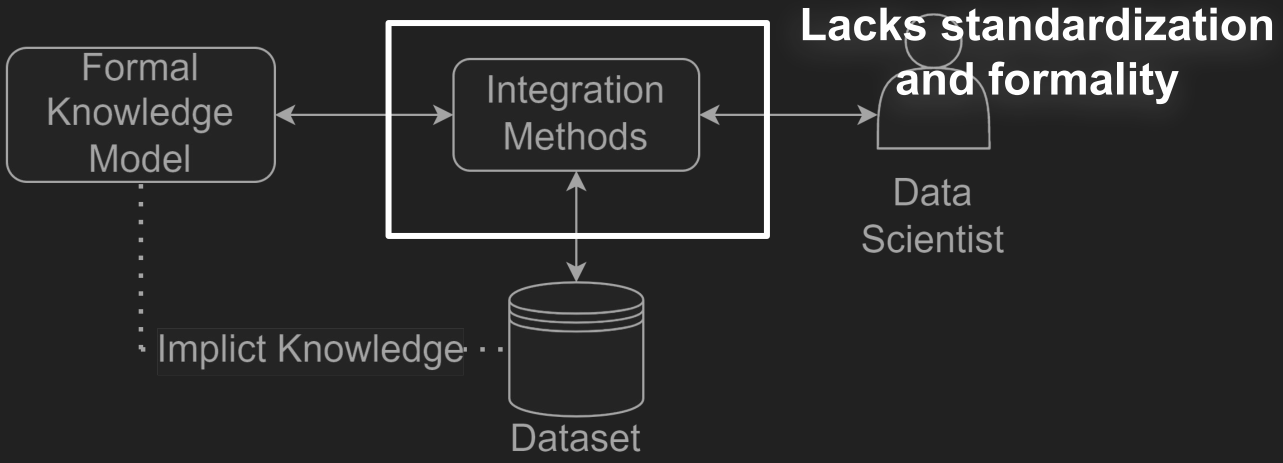
- Diverse knowledge can be represented at many levels of detail

- Integration method and role of data science operations is problematic



Knowledge Representation

- Diverse knowledge can be represented at many levels of detail
- Integration method and role of data science operations is problematic



Outline

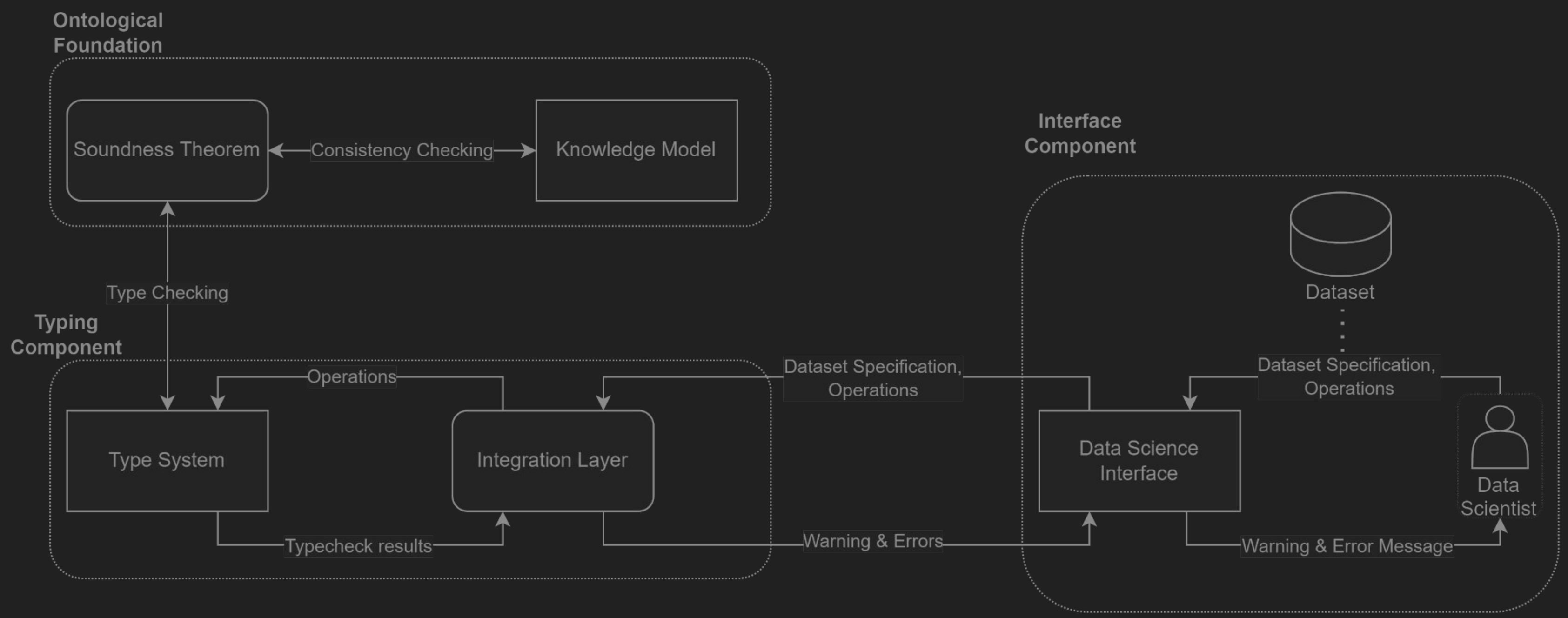
1. Why are datatypes problematic for data science?
2. **What are my contributions to solving this problem?**
3. What is the significance of these contributions?
4. What are the directions for future work?

The Meaningful Type Safety Framework (MeTS)

Ontologically-Sound Dependent Type Systems for Data Science

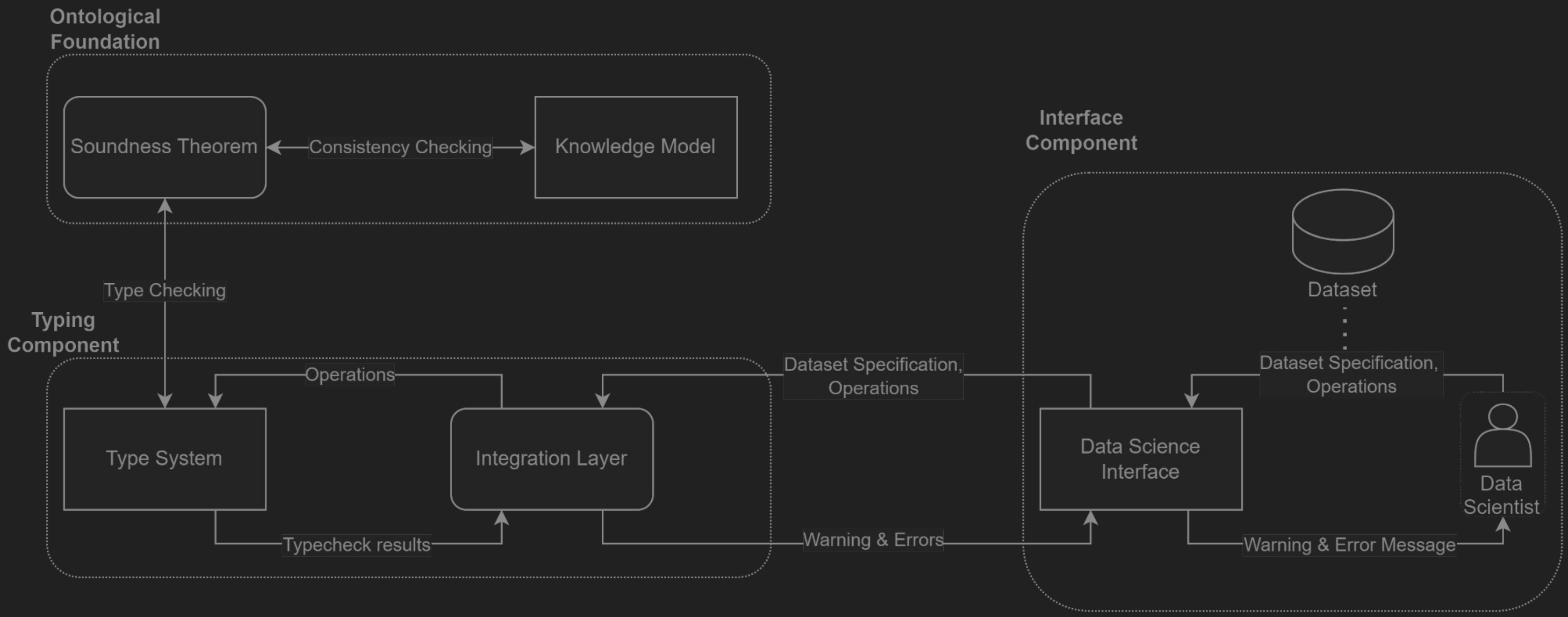
The Meaningful Type Safety Framework (MeTS)

Ontologically-Sound Dependent Type Systems for Data Science



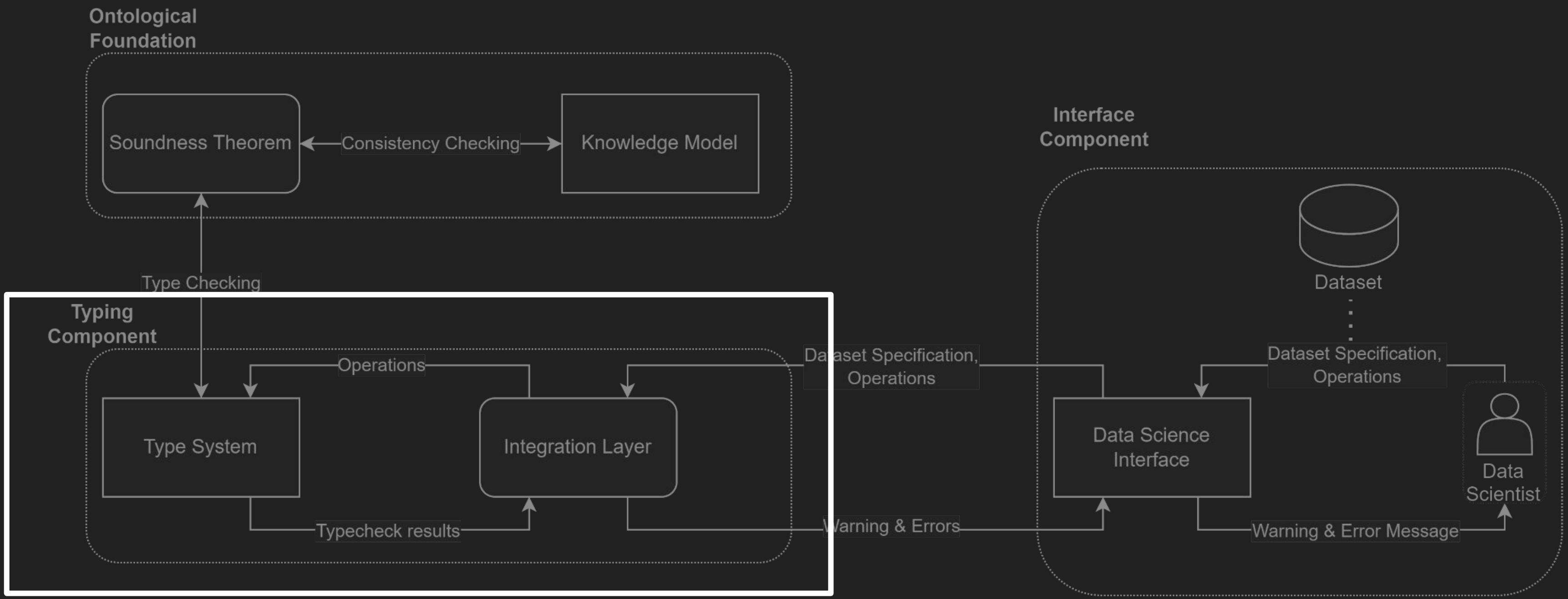
The Meaningful Type Safety Framework (MeTS)

Ontologically-Sound **Dependent Type Systems** for Data Science



The Meaningful Type Safety Framework (MeTS)

Ontologically-Sound **Dependent Type Systems** for Data Science



MeTS Type System

- Dependent pair types enforce operation preconditions

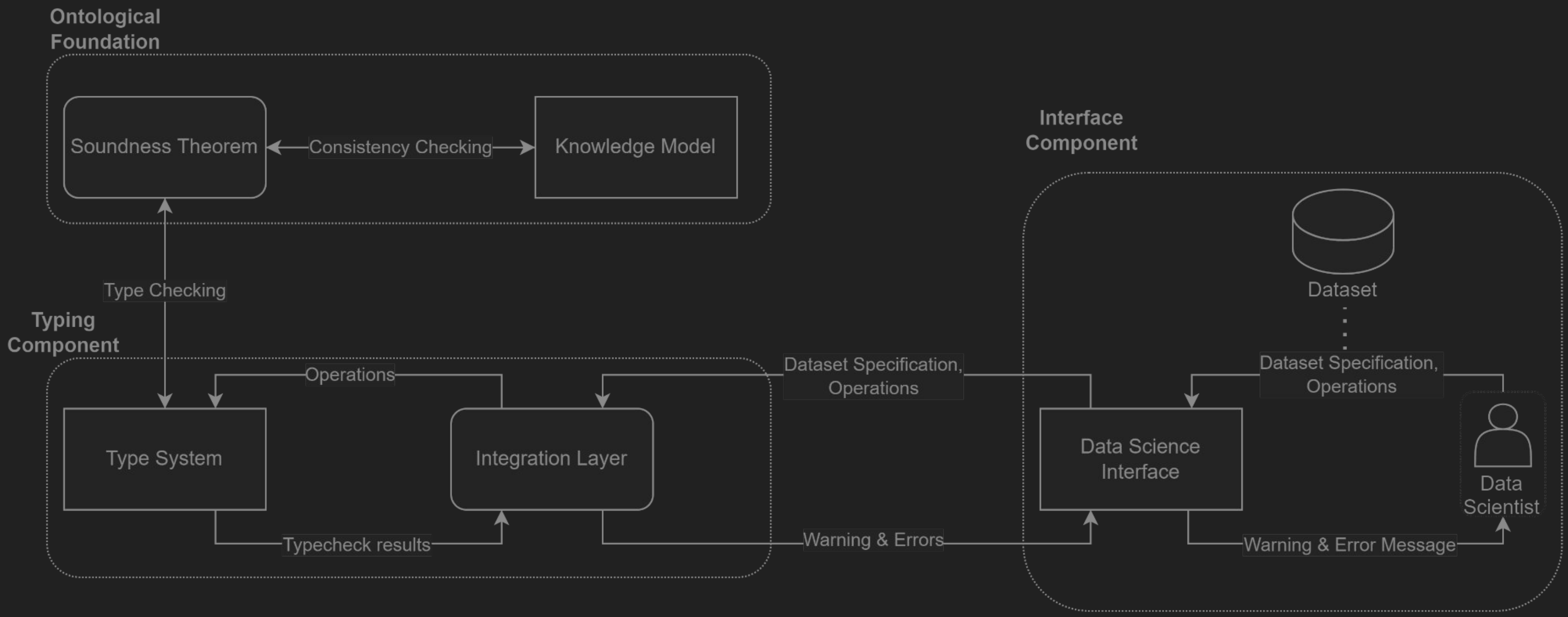
```
RegionSum : List Disjoint Region -> Region
```

- Values change after each operation : Provenance-Integrated

```
PopAvg(operands) = Avg Over operands
```

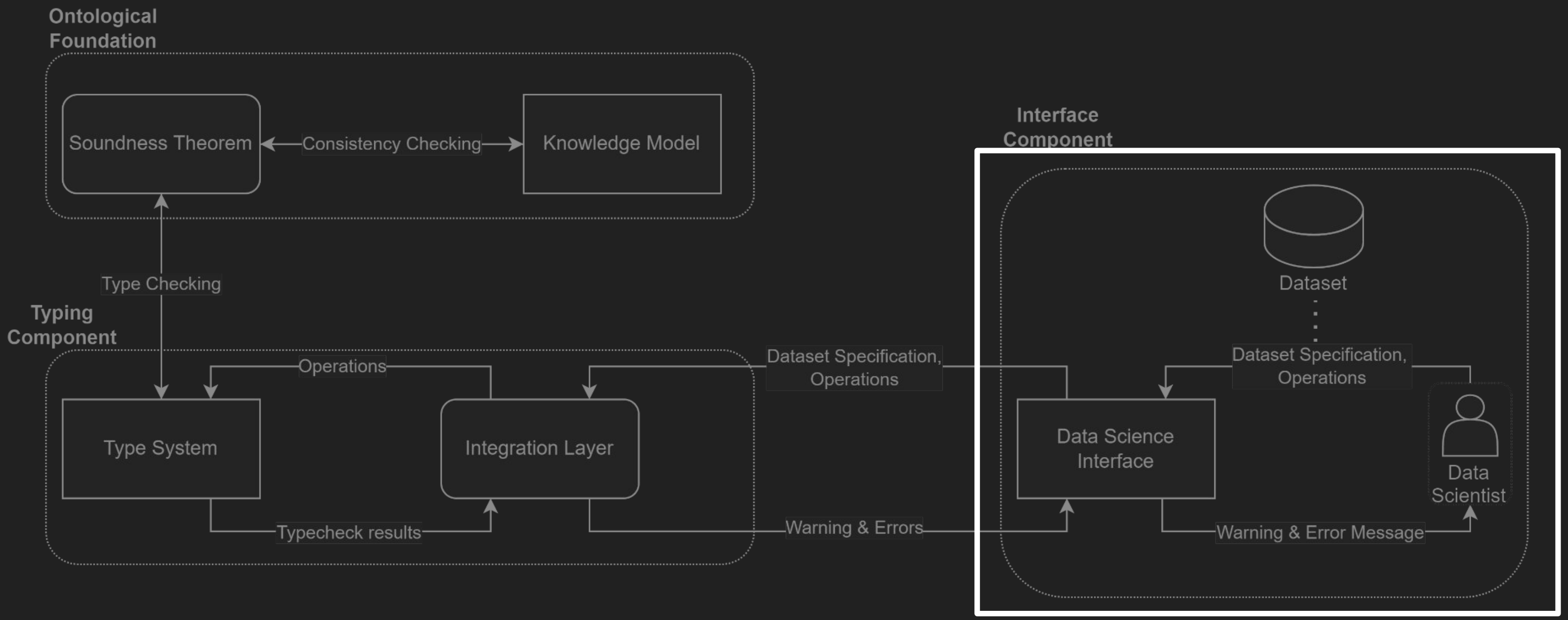
The Meaningful Type Safety Framework (MeTS)

Ontologically-Sound Dependent Type Systems **for Data Science**



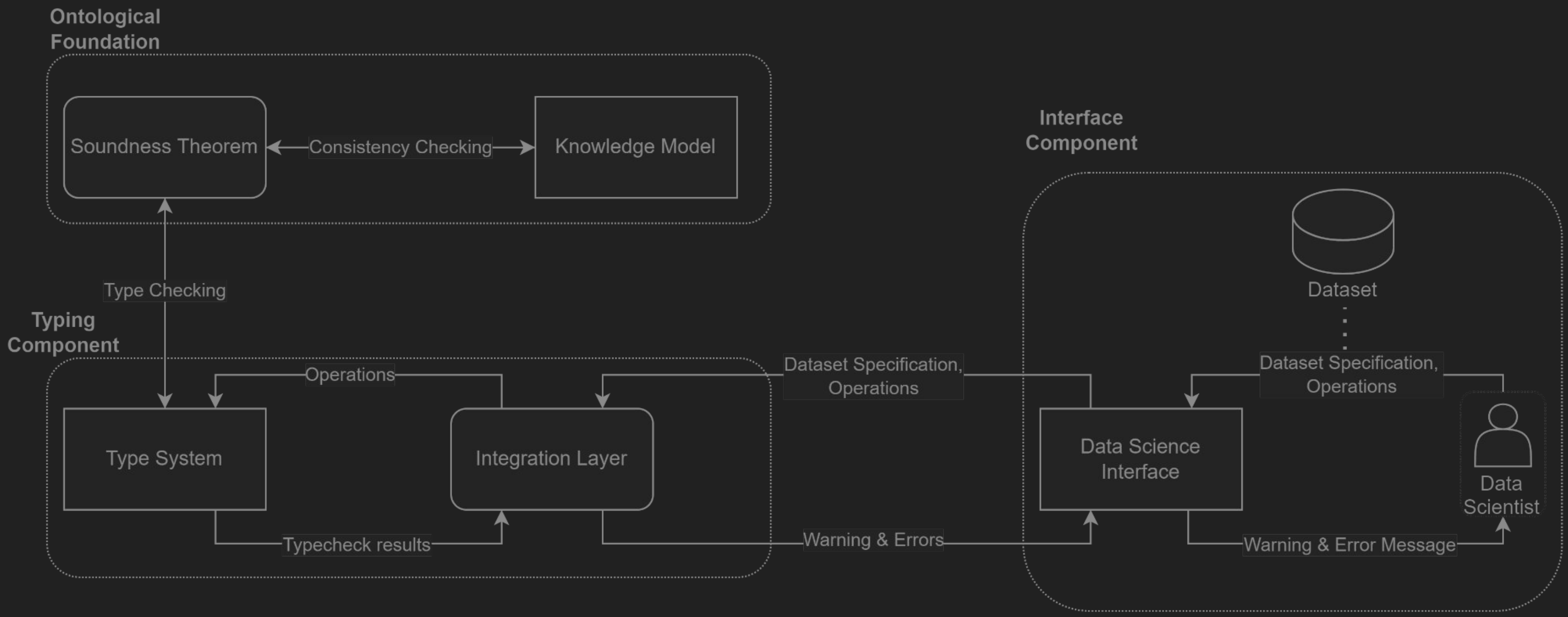
The Meaningful Type Safety Framework (MeTS)

Ontologically-Sound Dependent Type Systems **for Data Science**



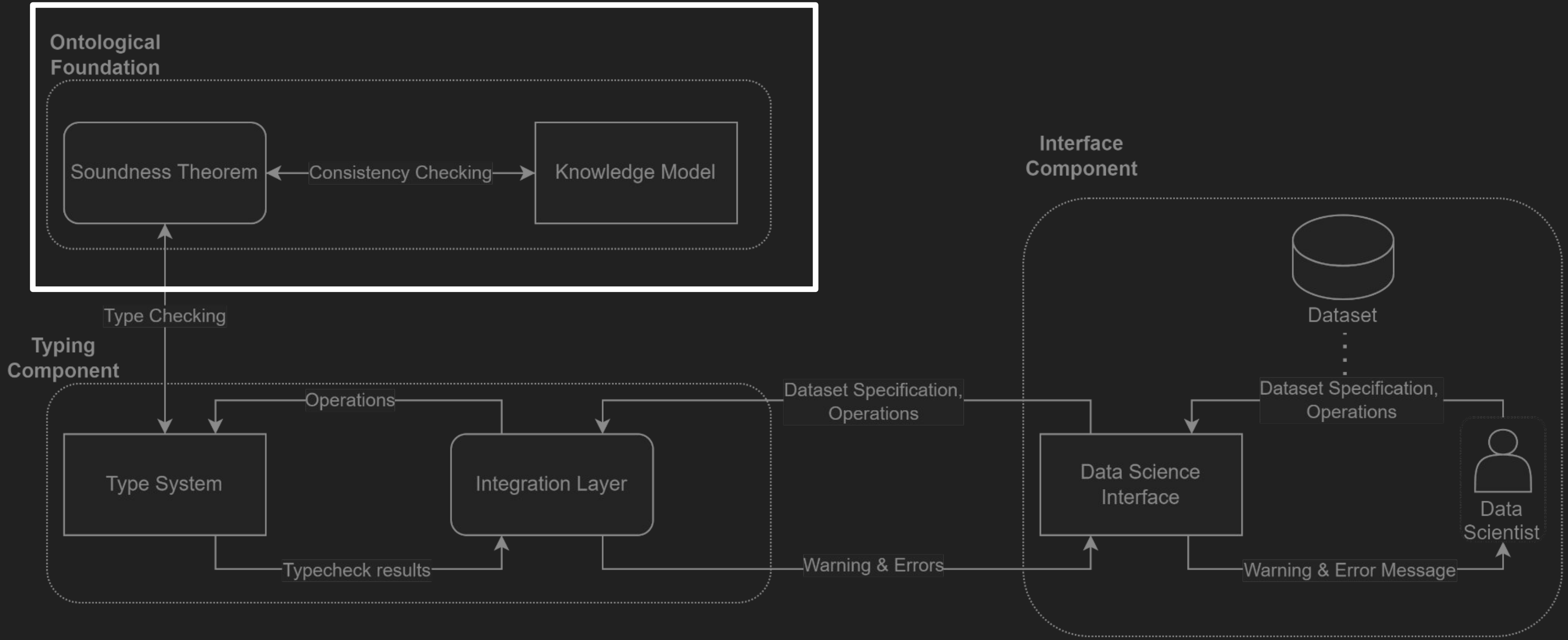
The Meaningful Type Safety Framework (MeTS)

Ontologically-Sound Dependent Type Systems for Data Science



The Meaningful Type Safety Framework (MeTS)

Ontologically-Sound Dependent Type Systems for Data Science

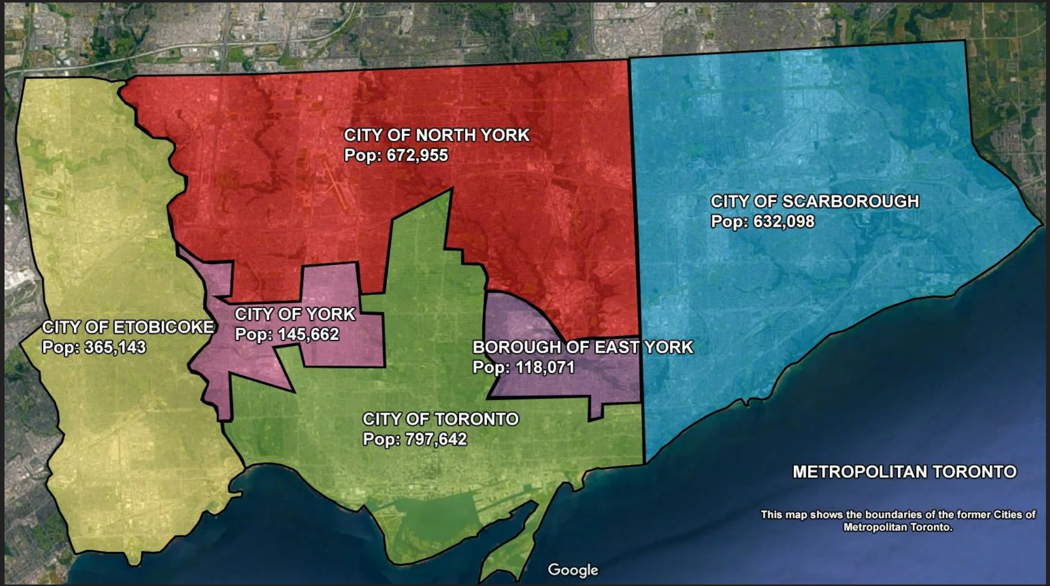


Census Data Ontology

- Represents the fundamental factors of census data
    - Movement of People
    - Crowd mereology
    - Geopolitical occupation
    - Geospatial mereology

Census Data Ontology

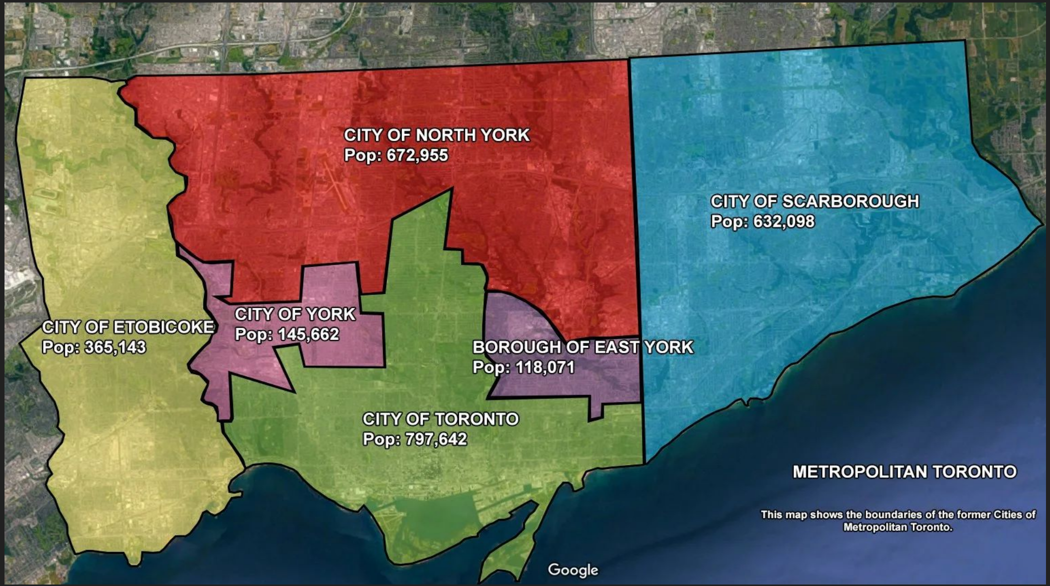
- Represents the fundamental factors of census data
  - Movement of People
  - Crowd mereology
  - Geopolitical occupation
  - Geospatial mereology



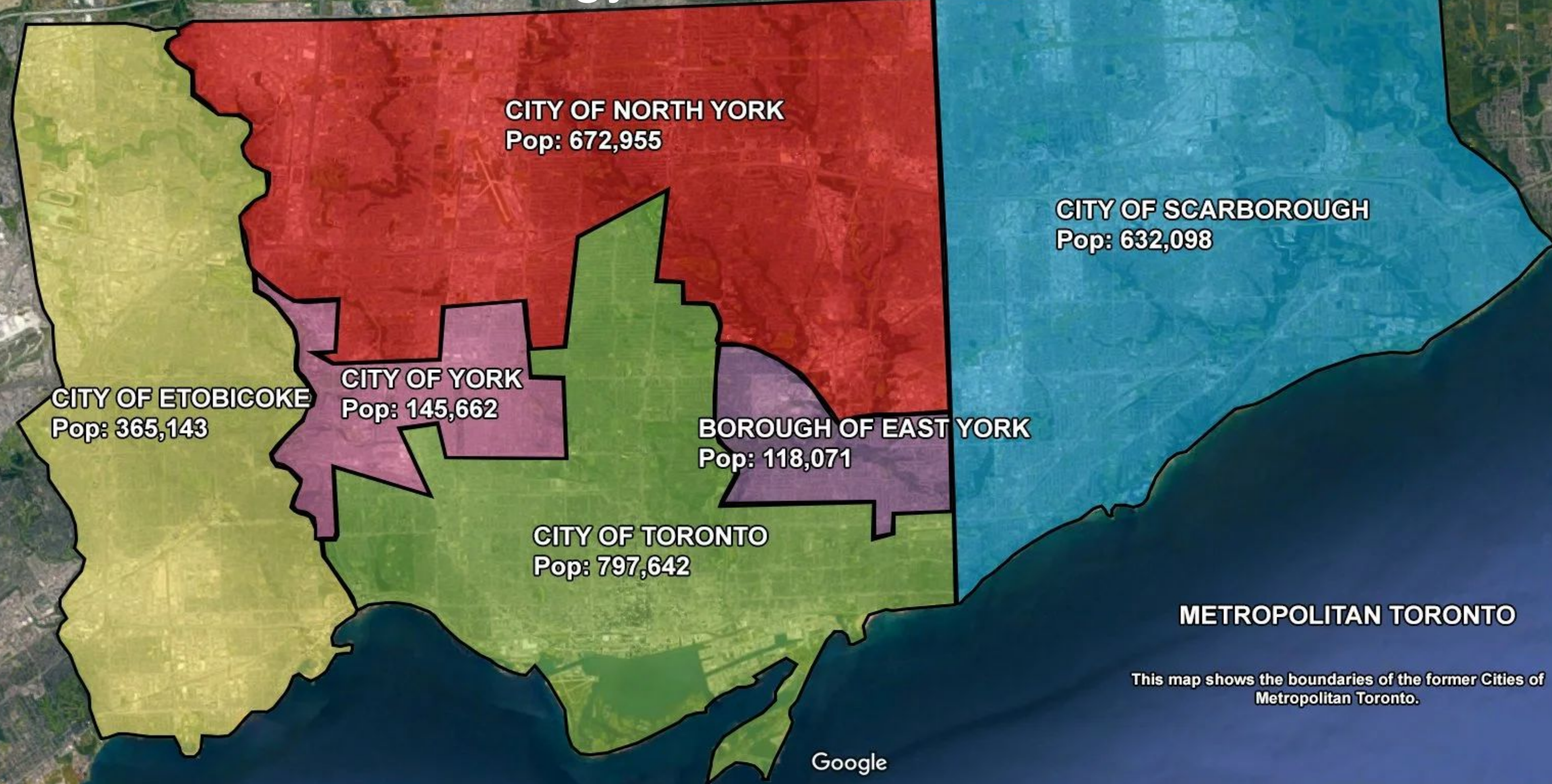
Census Data Ontology

- “I live in Toronto” - different interpretations over time

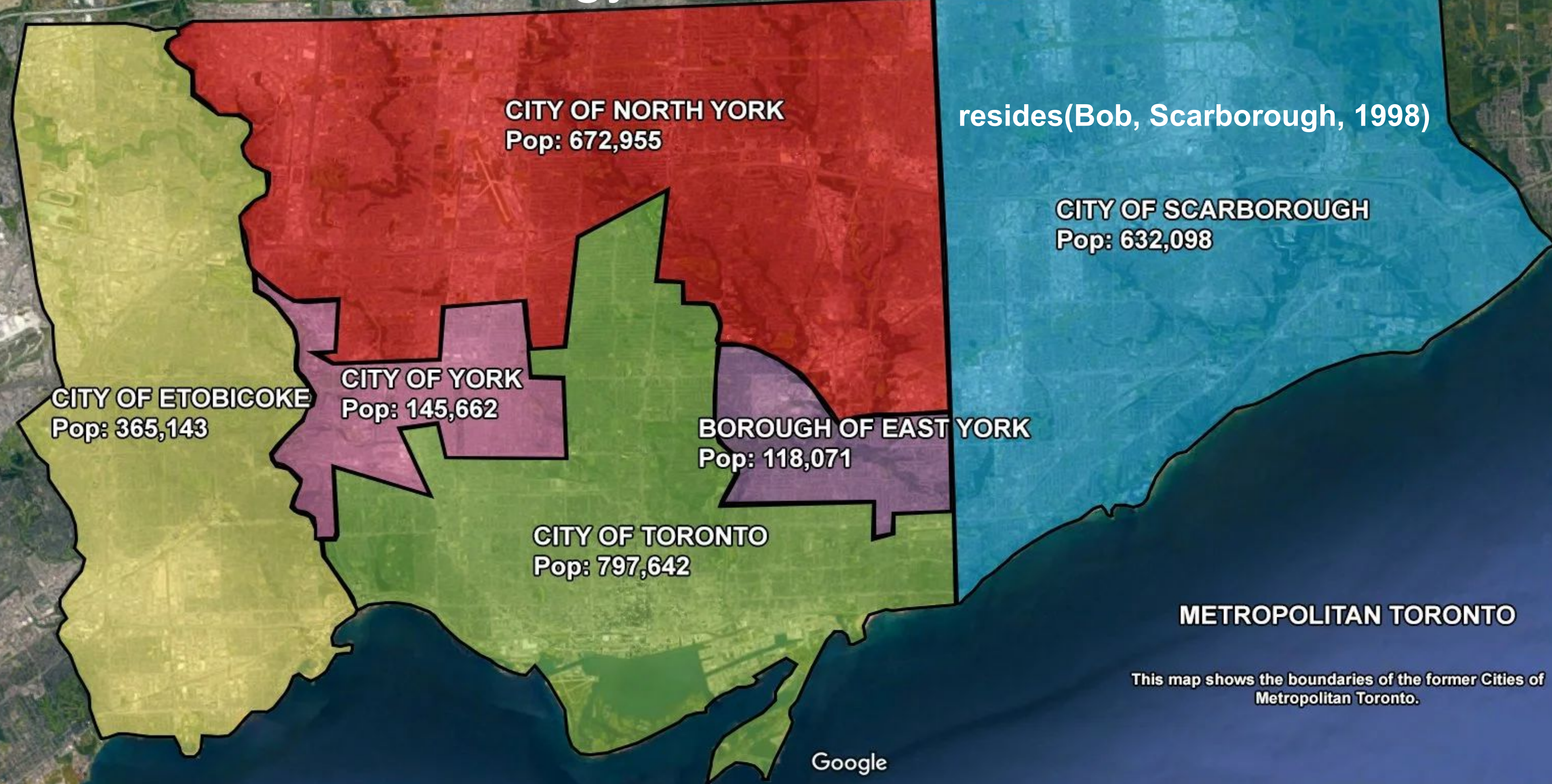
- Toronto as a region of land vs the geopolitical entity



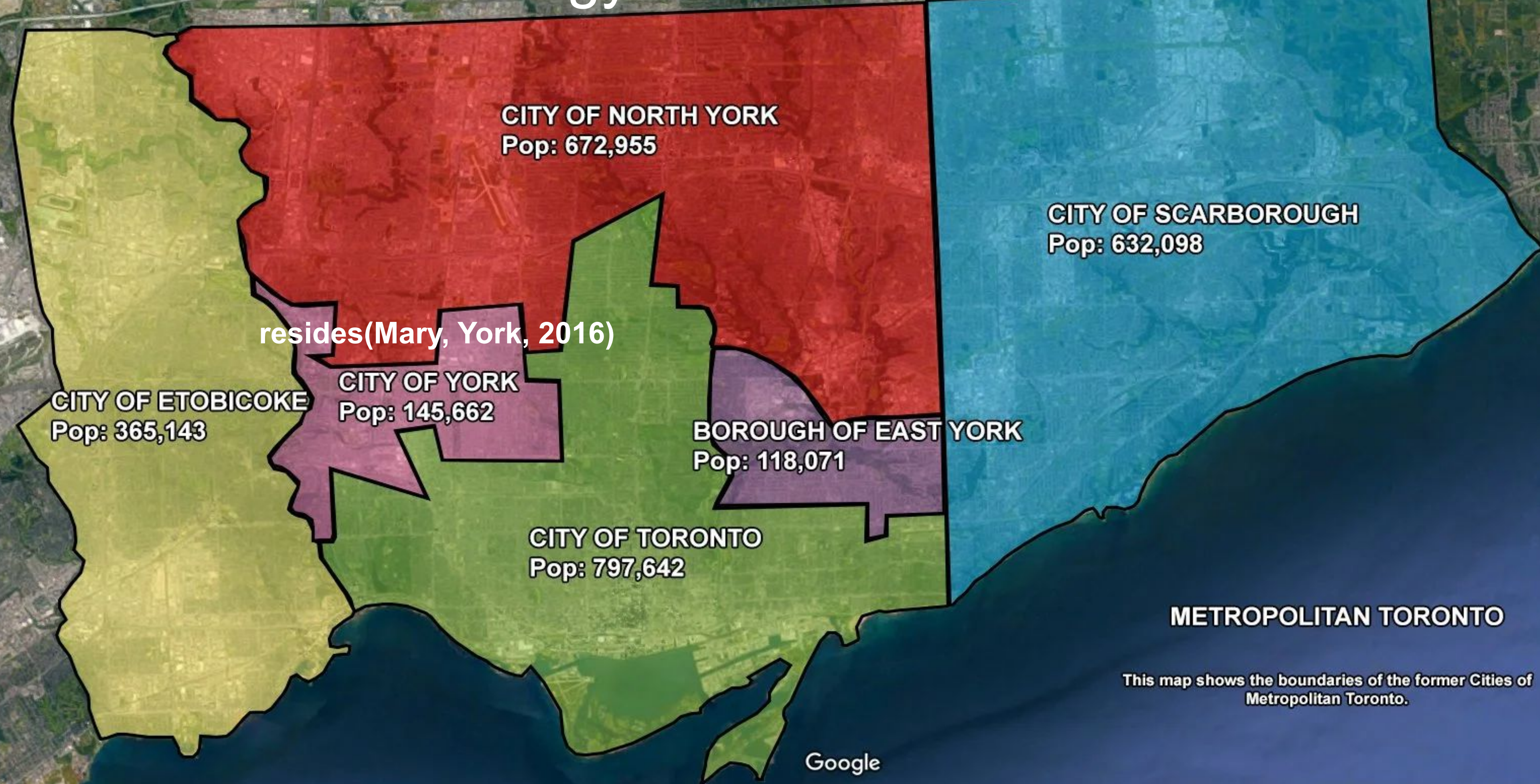
Census Data Ontology



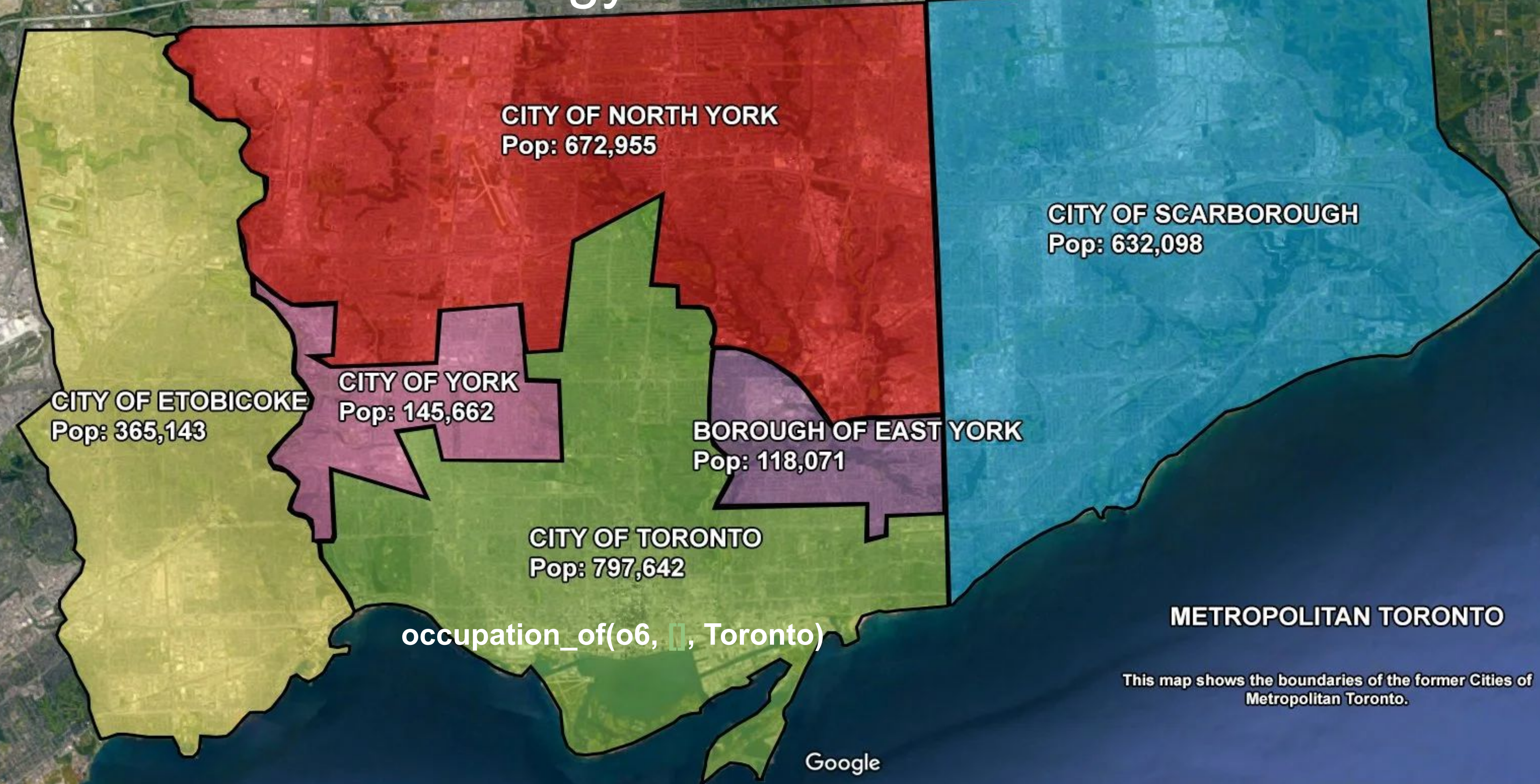
Census Data Ontology



Census Data Ontology

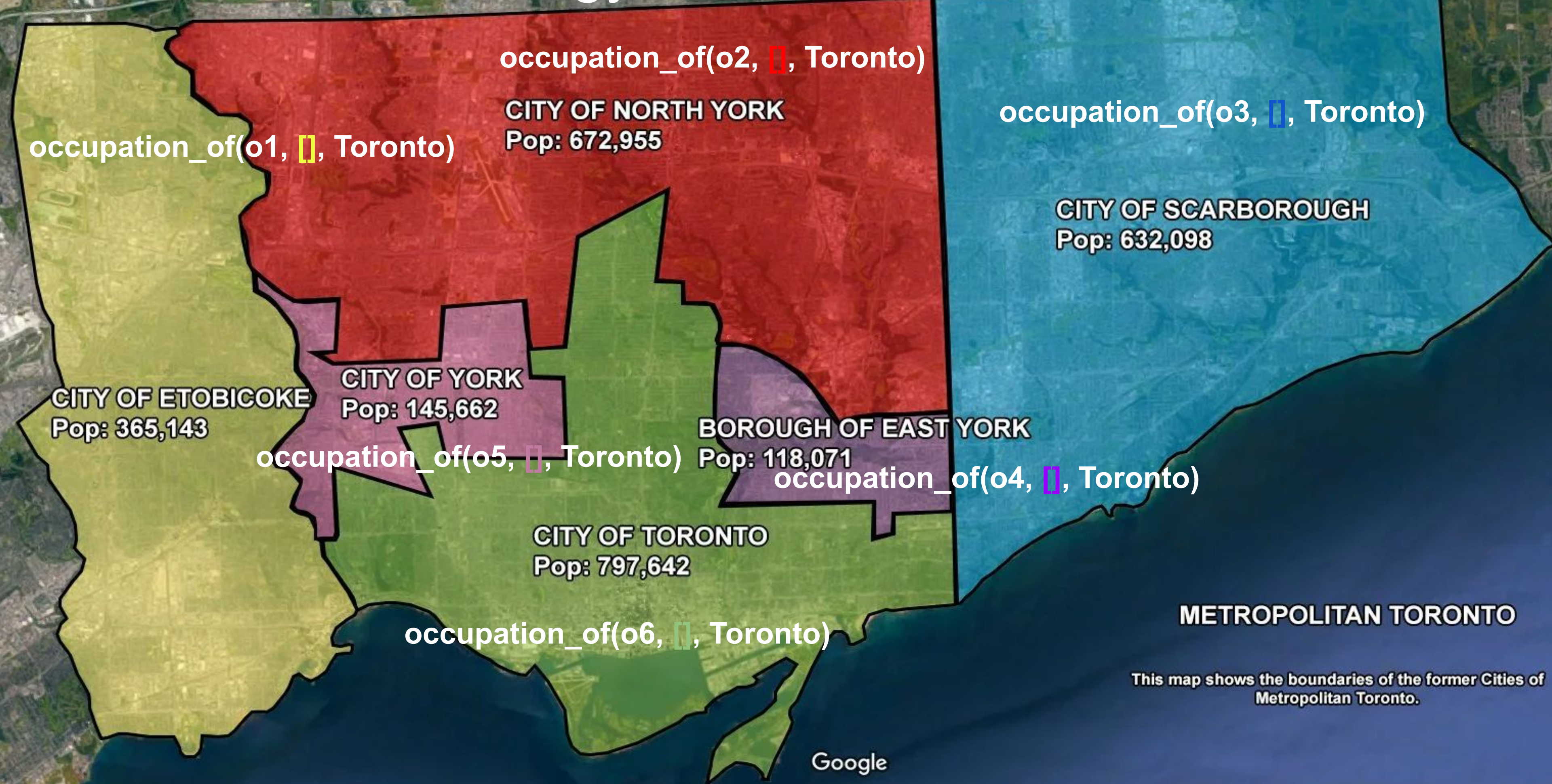


Census Data Ontology



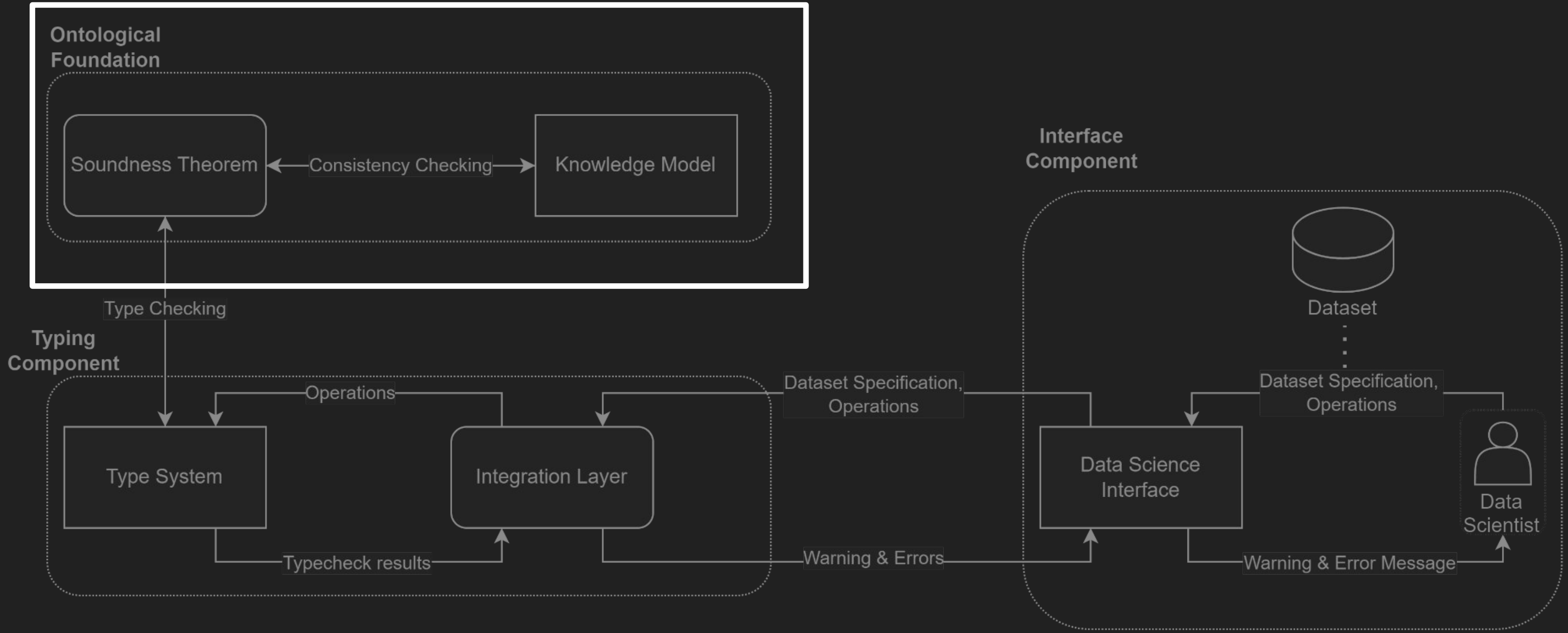
occupation_of(o6, [], Toronto)

Census Data Ontology



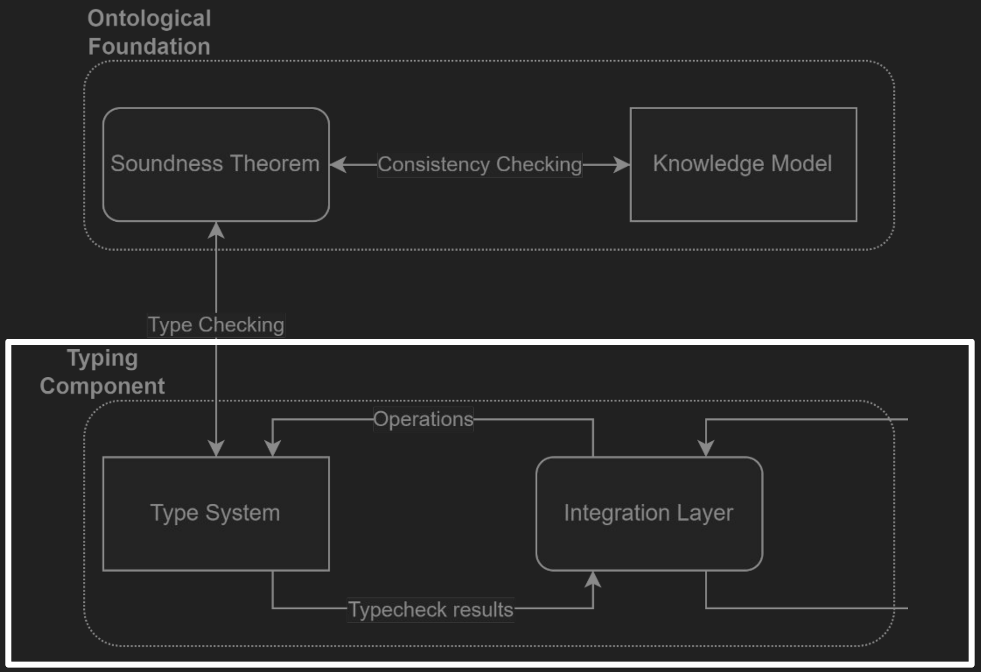
The Meaningful Type Safety Framework (MeTS)

Ontologically-Sound Dependent Type Systems for Data Science



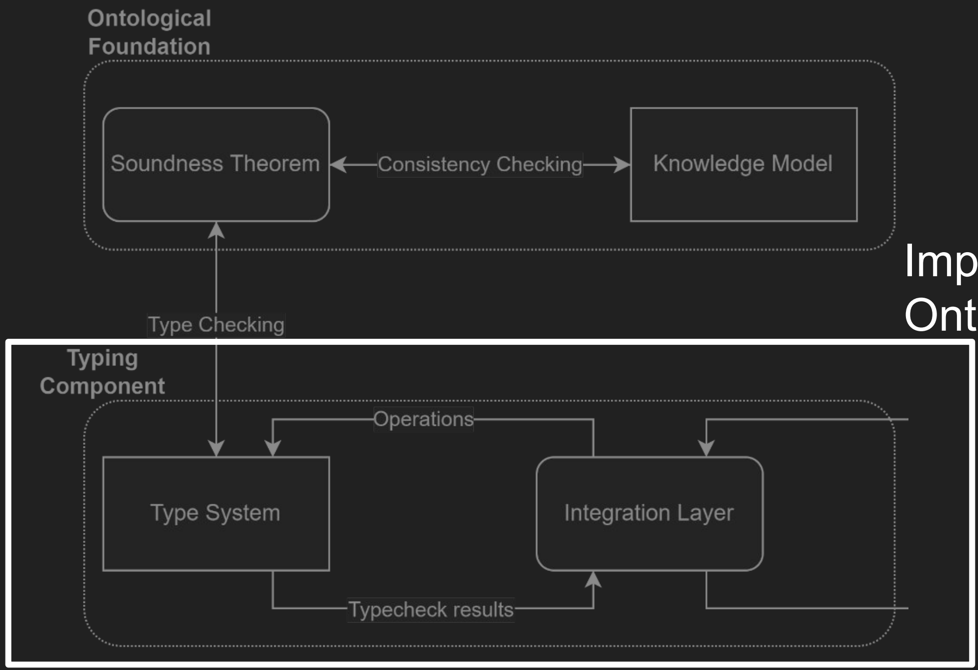
The Meaningful Type Safety Framework (MeTS)

Ontologically-Sound Dependent Type Systems for Data Science



The Meaningful Type Safety Framework (MeTS)

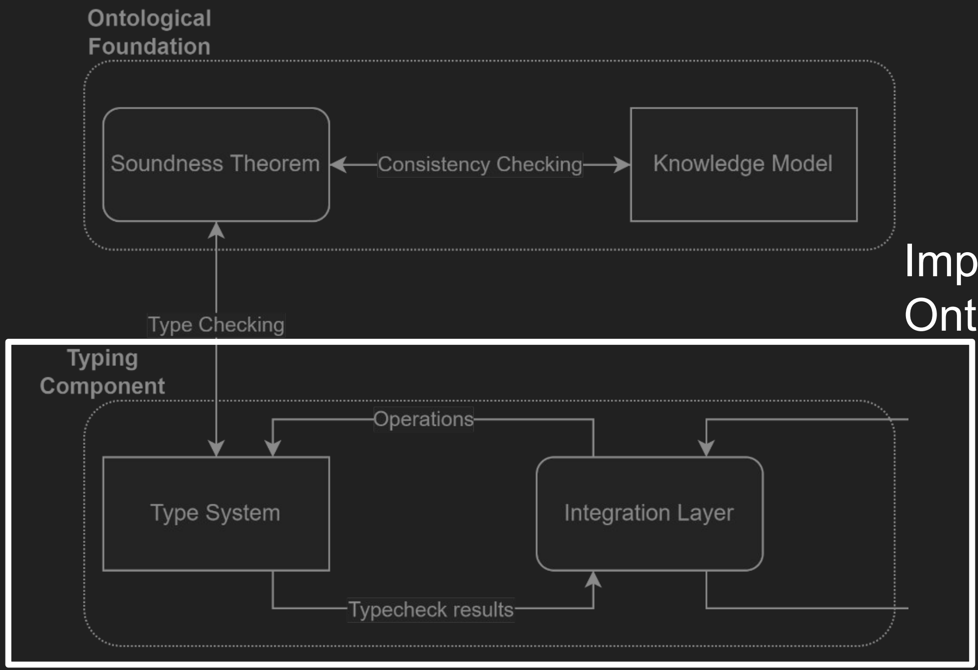
Ontologically-Sound Dependent Type Systems for Data Science



Implicit
Ontological Commitments

The Meaningful Type Safety Framework (MeTS)

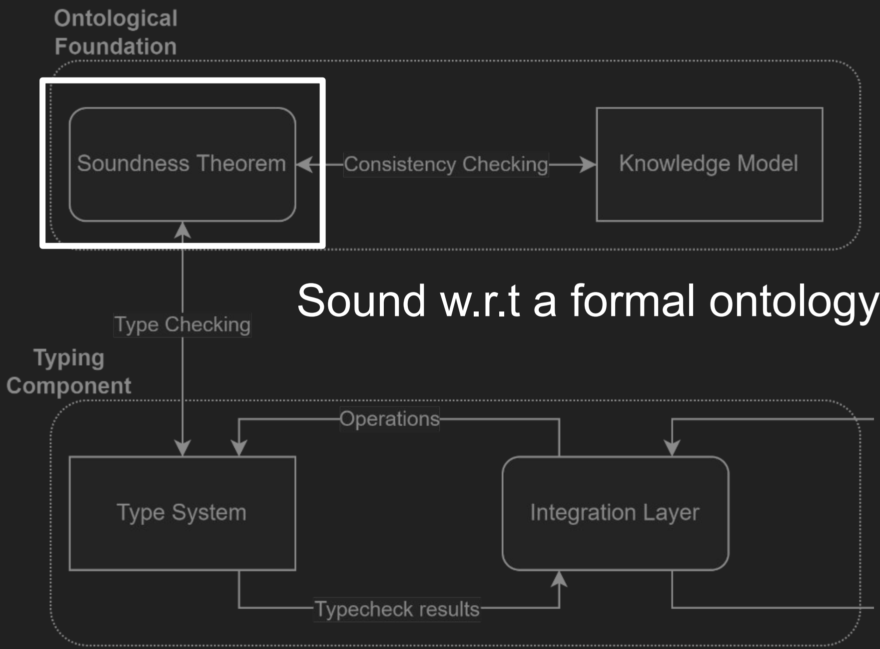
Ontologically-Sound Dependent Type Systems for Data Science



Implicit
Ontological Commitments

The Meaningful Type Safety Framework (MeTS)

Ontologically-Sound Dependent Type Systems for Data Science



Sound w.r.t a formal ontology

Outline

1. Why are datatypes problematic for data science?

2. What are my contributions to solving this problem?

3. **What is the significance of these contributions?**

4. What are the directions for future work?

Contributions & Significance

1. MeTS Framework

2. Census Data Ontology

3. MeTS Soundness Theorem

Contributions & Significance

1. **MeTS Framework**

2. Census Data Ontology

3. MeTS Soundness Theorem

Contributions & Significance

1. **MeTS Framework**
   I. Applies dependent types to model real-world knowledge for data science
   II. Elevates type safety to a meaningful result
   III. Ensures real-world interpretation is upheld throughout the data science pipeline

2. Census Data Ontology

3. MeTS Soundness Theorem

Contributions & Significance

1. MeTS Framework

2. **Census Data Ontology**

3. MeTS Soundness Theorem

Contributions & Significance

1. MeTS Framework

2. **Census Data Ontology**
   I. Models fundamental factors of census data
   II. Provides exceptional expressiveness
   III. Models census data operations

3. MeTS Soundness Theorem

Contributions & Significance

1. MeTS Framework

2. Census Data Ontology

3. **MeTS Soundness Theorem**

Contributions & Significance

1. MeTS Framework

2. Census Data Ontology

3. **MeTS Soundness Theorem**
    I. **Decouples ontological commitments from type system implementation**
    II. **Enables increased knowledge sharing**
    III. **Creates opportunities for efficient alternative reasoning**

Outline

1. Why are datatypes problematic for data science?
2. What are my contributions to solving this problem?
3. What is the significance of these contributions?
4. **What are the directions for future work?**

Future Work

1. **Correspondence Theorem**

2. Ontology for Data Quantities

3. Tractability

Future Work

1. **Correspondence Theorem**
   I. **Soundness AND completeness**
   II. **Requires detailed specification for logic and type systems**
   III. **New methods of collaboration**

2. Ontology for Data Quantities

3. Tractability

Future Work

1. Correspondence Theorem

2. **Ontology for Data Quantities**

3. Tractability

Future Work

1. Correspondence Theorem

2. **Ontology for Data Quantities**
   I. **Meta-model of existing ontologies**
   II. **Provenance and operation-centric**
   III. **Guides development of future ontologies and MeTS**

3. Tractability

Future Work

1. Correspondence Theorem

2. Ontology for Data Quantities

3. **Tractability**

Future Work

1. Correspondence Theorem

2. Ontology for Data Quantities

3. **Tractability**
   I. **Minimal change to data scientists workflow**
   II. **Integrate MeTS into existing data science tools**
   III. **Enable increased adoption**

Thank you!

Questions?

References

1. Gebru, Timnit, et al. "Datasheets for datasets." Communications of the ACM 64.12 (2021): 86-92.
2. Herschel, Melanie, Ralf Diestelkämper, and Houssem Ben Lahmar. "A survey on provenance: What for? What form? What from?." The VLDB Journal 26.6 (2017): 881-906.
3. Acar, Umut A., et al. "A Graph Model of Data and Workflow Provenance." TaPP. 2010.
4. Grüninger, Michael, et al. "Foundational Ontologies for Units of Measure." FOIS. 2018.
5. Firat, Aykut. Information integration using contextual knowledge and ontology merging. Diss. Massachusetts Institute of Technology, 2003.
6. Fox, Mark S. "The semantics of populations: A city indicator perspective." Journal of Web Semantics 48 (2018): 48-65.
7. Fox, Mark S. "An ontology engineering approach to measuring city education system performance." Expert Systems with Applications 186 (2021): 115734.
8. Albuquerque, Antognoni, and Giancarlo Guizzardi. "An ontological foundation for conceptual modeling datatypes based on semantic reference spaces." IEEE 7th International Conference on Research Challenges in Information Science (RCIS). IEEE, 2013.
9. Löh, Andres, Conor McBride, and Wouter Swierstra. "A tutorial implementation of a dependently typed lambda calculus." Fundamenta informaticae 102.2 (2010): 177-207.
10. Brady, Edwin. "Idris, a general-purpose dependently typed programming language: Design and implementation." Journal of functional programming 23.5 (2013): 552-593.

References

11. Balasubramanian, Vidhya. "InGIST: A Queryable and Configurable IndoorGIS Toolkit." Geospatial Infrastructure, Applications and Technologies: India Case Studies. Springer, Singapore, 2018. 93-105.
12. Abdelmoty, Alia I., et al. "A critical evaluation of ontology languages for geographic information retrieval on the Internet." Journal of Visual Languages & Computing 16.4 (2005): 331-358.
13. A Shallow Embedding of Pure Type Systems into First-Order Logic
14. Dapoigny, Richard, and Patrick Barlatier. "Towards Ontological Correctness of Part-whole Relations with Dependent Types." FOIS. Vol. 2010. 2010.
15. Barlatier, Patrick, and Richard Dapoigny. "A type-theoretical approach for ontologies: The case of roles." Applied Ontology 7.3 (2012): 311-356.
16. Dapoigny, Richard, and Patrick Barlatier. "Formalizing context for domain ontologies in Coq." Context in computing. Springer, New York, NY, 2014. 437-454.