Semantics in Data Science

What does it all mean?

Riley Moher

04.02.2021
Data-Driven Decision Making Lab
University of Toronto

Data Semantics: Catastrophic Consequences

- Data science lacks a consistent semantics
  - Columns improperly labelled
  - Heterogeneous datasets not properly combined

- In Academia, this is research funding down the drain in wasted effort

- In healthcare domain, lack of semantics means:
  - Poor interoperability
  - Missing, incomplete patient data
  - Loss of life

Overview

1. **Why ignoring data semantics causes problems**

2. **How current solutions give an incomplete answer**

3. **How we could address data semantics**

Overview

1. **Why Ignoring data semantics causes problems**
    I. Lack of context in columns
    II. Machine learning

2. **How current solutions give an incomplete answer**

3. **How we could address data semantics**

Lack of Context in Columns

- You have been given two datasets with some data dictionaries

CompanyA

Quarter	Profit	Operating Costs
Q1-2014	158 000 000	35 000 000
Q2-2014	200 000 000	38 000 000
...	...	...

Data Dictionary A

CompanyB

Quarter	Profit	Operating Costs
Q1-2014	160 000 000	26 000 000
Q2-2014	300 000 000	33 000 000
...	...	...

Data Dictionary B

- Which company is better to invest in?
  - B: lower costs, higher profit

Lack of Context in Columns

- You have been given two datasets with some data dictionaries

CompanyA

Quarter	**Profit (Net)**	Operating Costs
Q1-2014	158 000 000	35 000 000
Q2-2014	200 000 000	38 000 000
...	...	...

Data Dictionary A

CompanyB

Quarter	**Profit (Gross)**	Operating Costs
Q1-2014	160 000 000	26 000 000
Q2-2014	300 000 000	33 000 000
...	...	...

Data Dictionary B

- The columns were labelled the same, but represented different concepts

Lack of Context in Columns

- Mislabelled columns arise in many domains and cause lots of issues
  - Toronto 1985 vs Toronto 2021 neighbourhood definitions
  - Admissions to Sunnybrook hospital vs admissions to hospitals in Ontario
  - Price to earnings ratio: earnings for entire year, or just last quarter?

- Semantic Heterogeneity can be spatial, temporal, aggregational, etc.

- Semantic Heterogeneity is not limited only to columns

Semantics in Machine Learning

- You wish to train a CHF classification model on ECG data
- The data has some pre-processing applied to it by a colleague

Time	Signal	CHF
0.25	6.27	0
0.50	4.99	0
0.25	6.35	1
0.5	5.12	1
...	...	...

- You get a train accuracy of 99.3%, and a test accuracy of 97.8%.

Semantics in Machine Learning

- You wish to train a CHF classification model on ECG data
- The data has some pre-processing applied to it by a colleague

Time	Signal	CHF
0.25	6.27	0
0.50	4.99	0
0.25	**6.35**	**1**
0.5	**5.12**	**1**
...	...	...

Downsampled

- You tried to merge data that had different processes applied

Semantics in Machine Learning

- Semantic Heterogeneity is especially troublesome in machine learning
  - Classes having distinct operations applied to them
  - Training and testing on heterogenous data
  - Test data filled with observations later than training samples

- Data operations themselves have implicit assumptions and semantics
  - Summation implies non-overlapping samples (don’t double count)

- Current solutions do not solve the complete problem

Overview

1. Ignoring data semantics causes problems
2. **How current solutions give an incomplete answer**
    I. **Improving Documentation**
    II. **Data Provenance**
    III. **Type Theory**
    IV. **Ontologies**
3. How we could address data semantics

Overview

1. Ignoring data semantics causes problems
2. **Current solutions get some things right, but don't work**
    I. **Improving Documentation**
    II. Data Provenance
    III. Type Theory
    IV. Ontologies
3. How we could address data semantics

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. **Ignoring data semantics causes problems**
2. **How current solutions give an incomplete answer**
    I. Improving Documentation
    II. **Data Provenance**
    III. Type Theory
    IV. Ontologies
3. **How we could address data semantics**

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. **Ignoring data semantics causes problems**
2. **How current solutions give an incomplete answer**
    I. Improving Documentation
    II. Data Provenance
    III. Type Theory
    IV. Ontologies
3. **How we could address data semantics**

Turning to Types

- Common programming languages have primitive type safety

```
A = 6 + “hello”
> Error: Cannot add int and String

String one = 1;
> Error: 1 is not of type String
```

- Can we make types smarter?
  - Semantic Datatypes > Primitive Datatypes

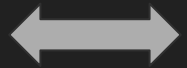
Turning to Types

- Leverage the **curry-howard correspondence**
  - Correspondence between proofs and programs

Program

```
Data Point = Point Int Int

makePoint :: Int -> Int -> Point
makePoint x y = Point x y
```



Logic

$$\frac{x: \text{Int} \quad y: \text{Int}}{\texttt{makePoint x y} : \text{Point}}$$

- We can construct programs based on type logic and vice versa

- How many data scientists have you heard say “I use Haskell all the time!”?

Overview

1. **Ignoring data semantics causes problems**
2. **Current solutions get some things right, but don’t work**
    I. Improving Documentation
    II. Data Provenance
    III. Type Theory
    IV. Ontologies
3. **How we could address data semantics**

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. Ignoring data semantics causes problems

2. How current solutions give an incomplete answer

3. **How we could address data semantics**

Forward-Looking Thoughts

- How can we create a framework that features:
  - Semantic Datatypes
  - Provenance-integrated types
  - Data science tool support
  - Semantics of data operations

- Ultimately, data semantics are part of a decision support system.

- We cannot use data semantics to take the science out of data science
  - But we can prevent serious issues

… and maybe save a few lives