Challenges Related to Data Alignment
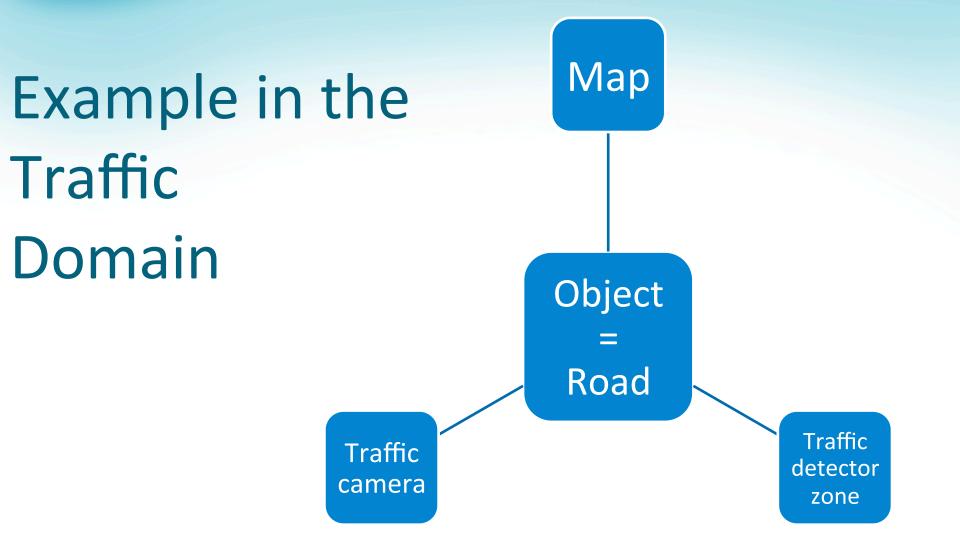
Marion Le Bras
March 17–18, 2016

Defining Alignment

"Alignment refers to the process of **relating different instances of the same object**, e.g., a word with the corresponding visual object, or time stamps associated with two different time series."
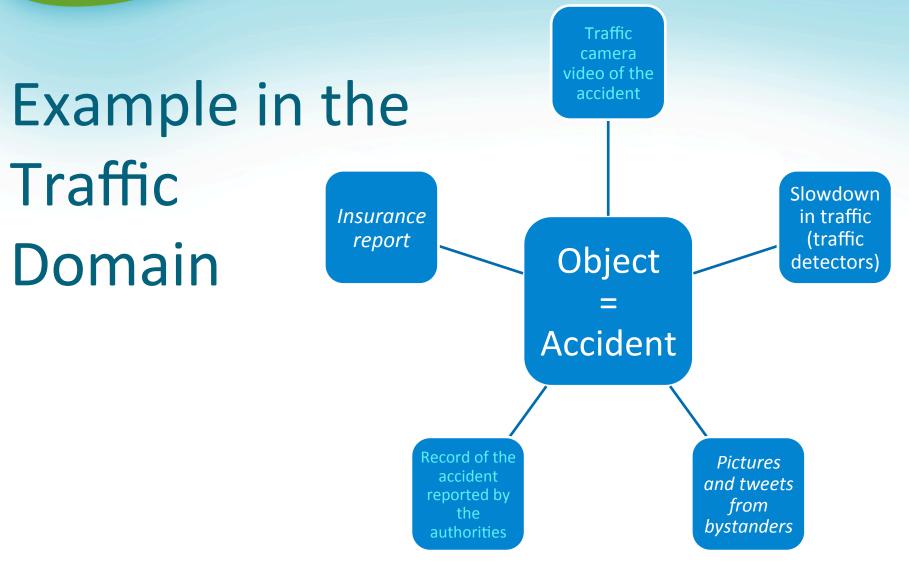
Dorr, Bonnie J., et al. "The NIST IAD Data Science Research Program." Data Science and Advanced Analytics (DSAA), 2015. 36678 2015. IEEE International Conference on. IEEE, 2015.

Example in the Traffic Domain

Map

Object = Road

Traffic camera

Traffic detector zone

Example in the Traffic Domain

Traffic camera video of the accident

Insurance report

Object = Accident

Slowdown in traffic (traffic detectors)

Record of the accident reported by the authorities

Pictures and tweets from bystanders

Provenance	Origin of raw data? Is it current? What processes were applied to derive the data from its original sources?
Heterogeneity	How to use data from multiple large heterogeneous datasets?
Predictive Analytics	How to identify and distinguish trends from random fluctuation to provide a calibrated forecast of future value?
Knowledge Assimilation	How might algorithms understand data, e.g., infer causality?
Big Data Replicability	How to reproduce experimental findings given that truth may be hard to find, consistently?
Visualization of Knowledge	How to visually represent knowledge for decision making?
Data Uncertainty	How to handle gaps in knowledge due to potential for untrustworthy or inaccurate data?
Mitigating Error Propagation	How can algorithms mitigate cascading of error through data processing steps?
Data Privacy and Security	How do we manage data and develop algorithms in the face of privacy and security concerns/policies?

Questions

- What is an alignment problem that you face in your domain?

- How do you deal with it?