Challenges Related to Data Cleaning

Peter Fontana

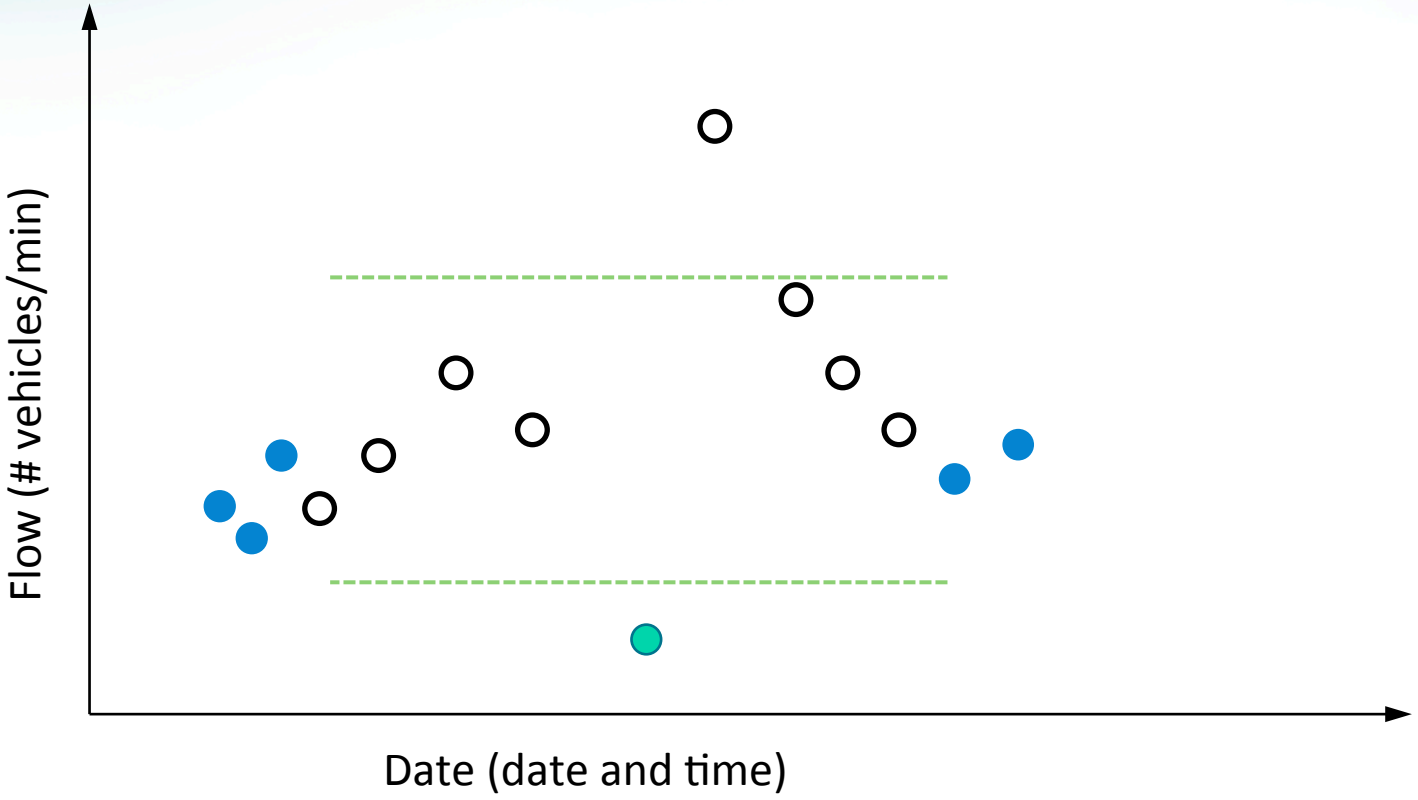
March 17–18, 2016

Dirty Data is A Problem

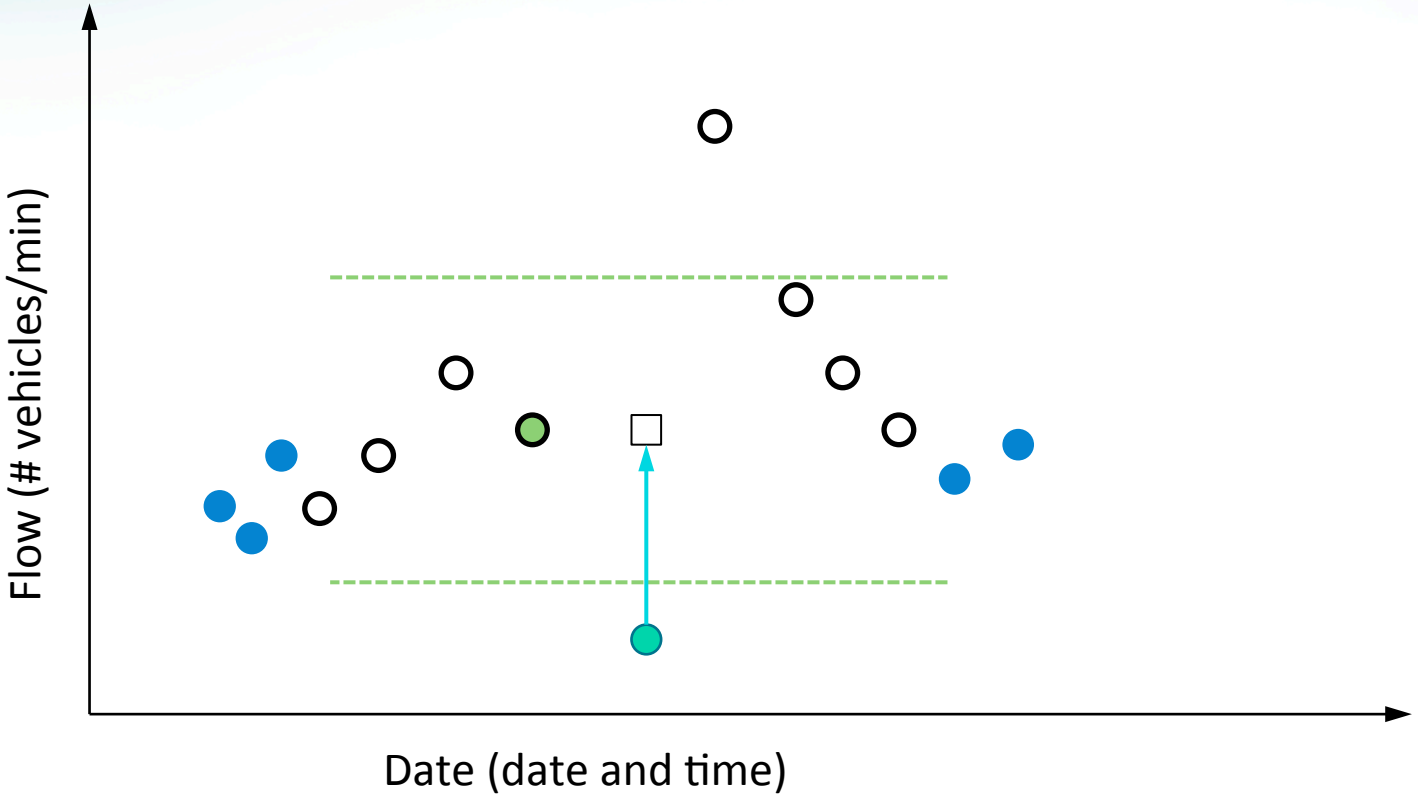
- "Yet far too much handcrafted work — what data scientists call 'data wrangling,' 'data munging' and 'data janitor work' — is still required. Data scientists, according to interviews and expert estimates, spend from 50 percent to 80 percent of their time mired in this more mundane labor of collecting and preparing unruly digital data, before it can be explored for useful nuggets."

Source: Steve Lohr, "For Big-Data Scientists, 'Janitor Work' is Key Hurdle to Insights." *The New York Times*. August 17, 2014. http://nyti.ms/1t8IzfE

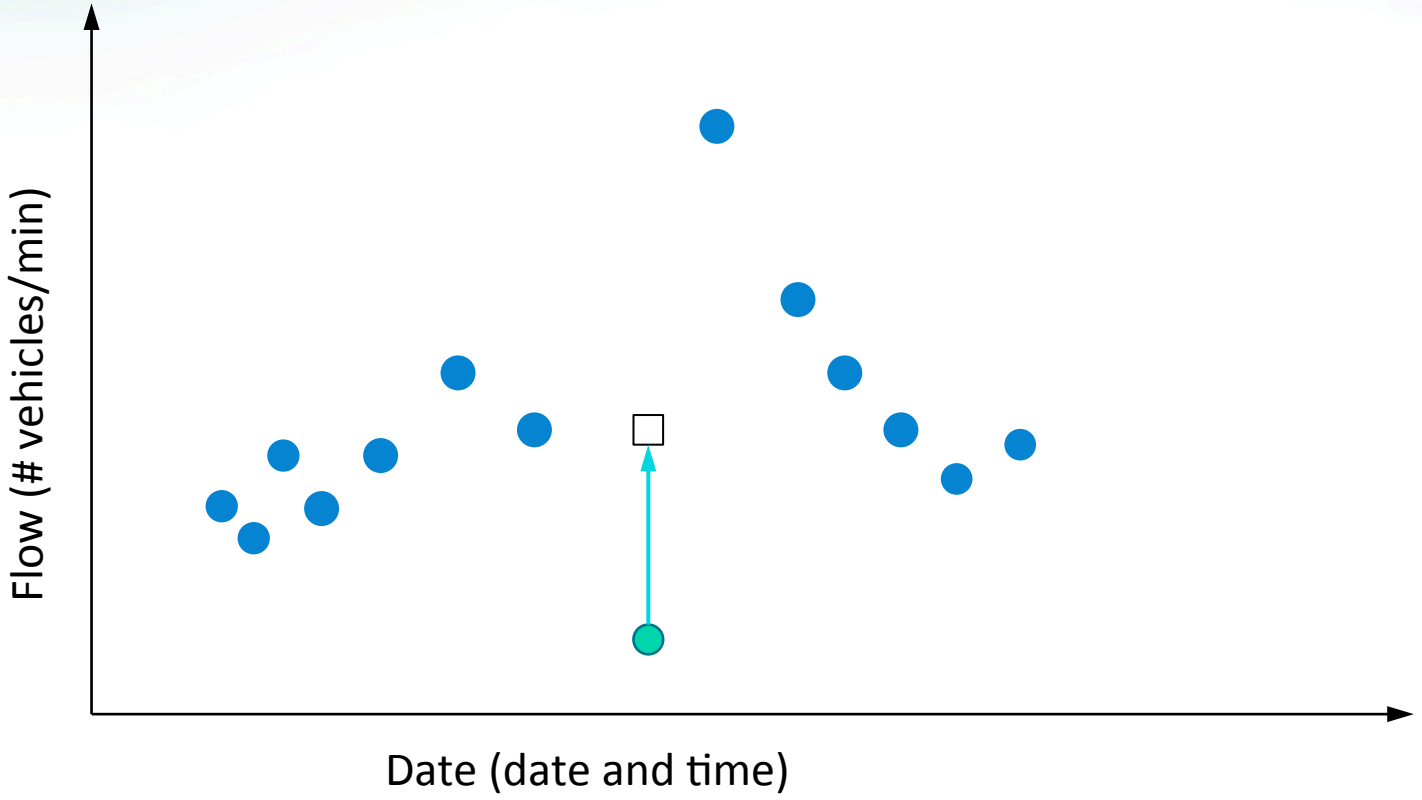
Cleaning: Anomaly Detection?



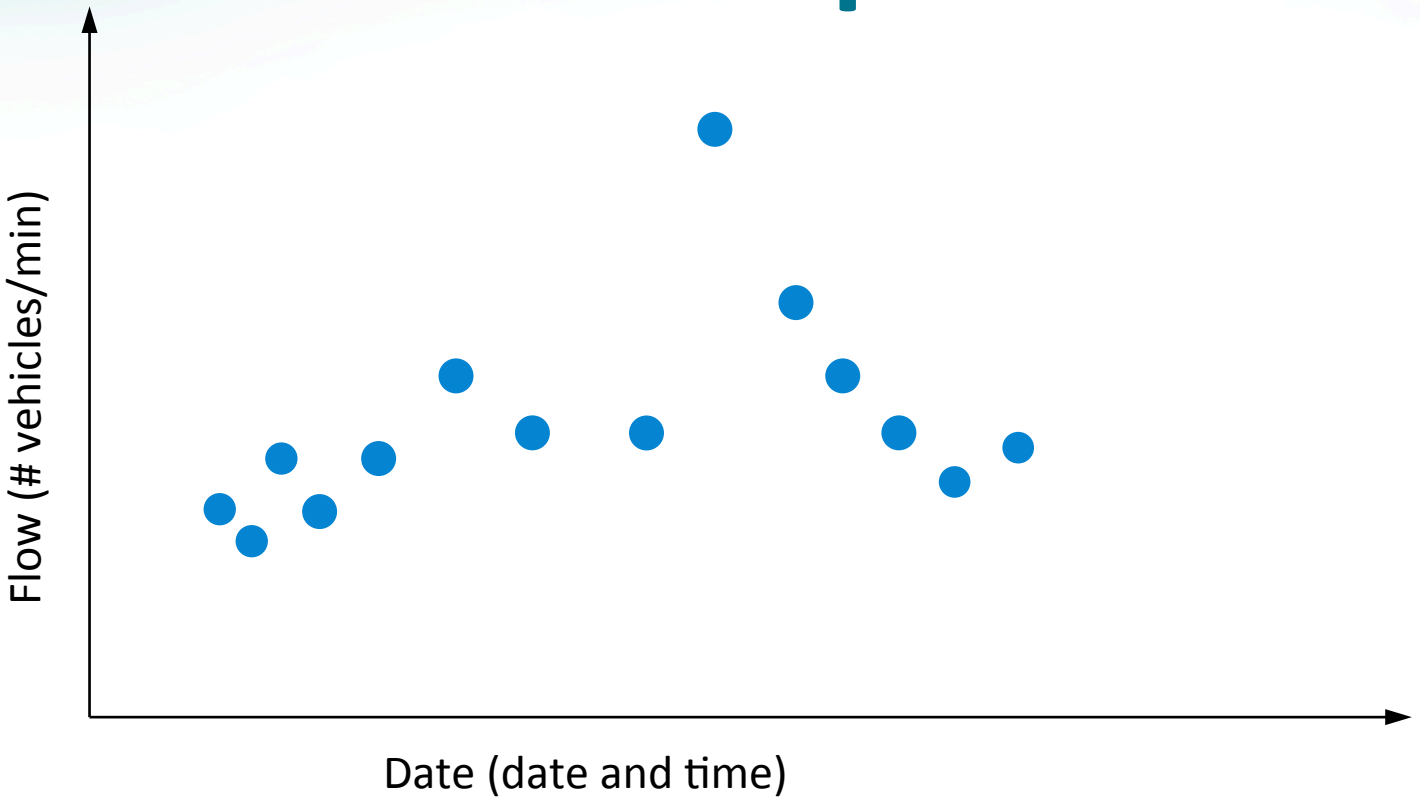
Cleaning: Must Correct Data



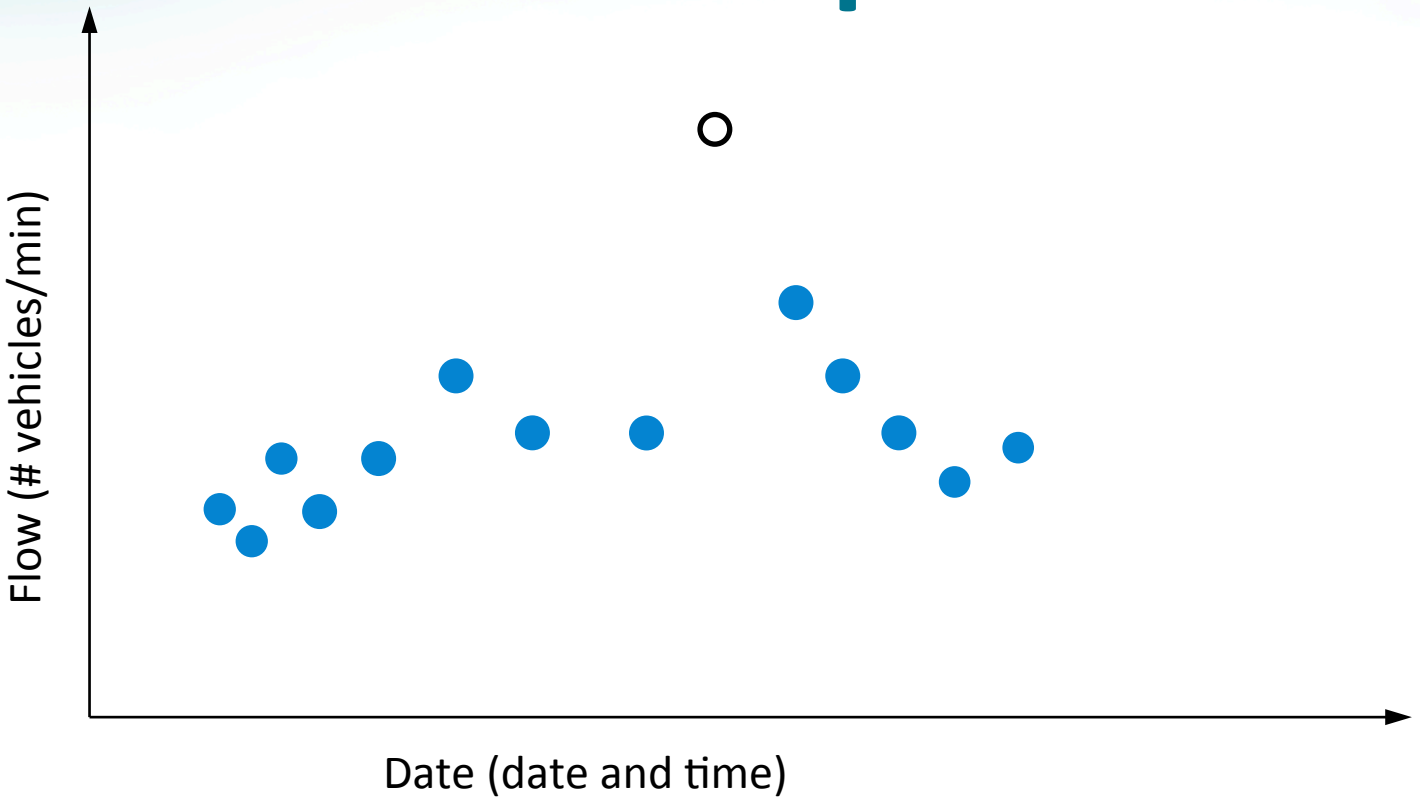
Data Uncertainty: Are the points **cleaner**?



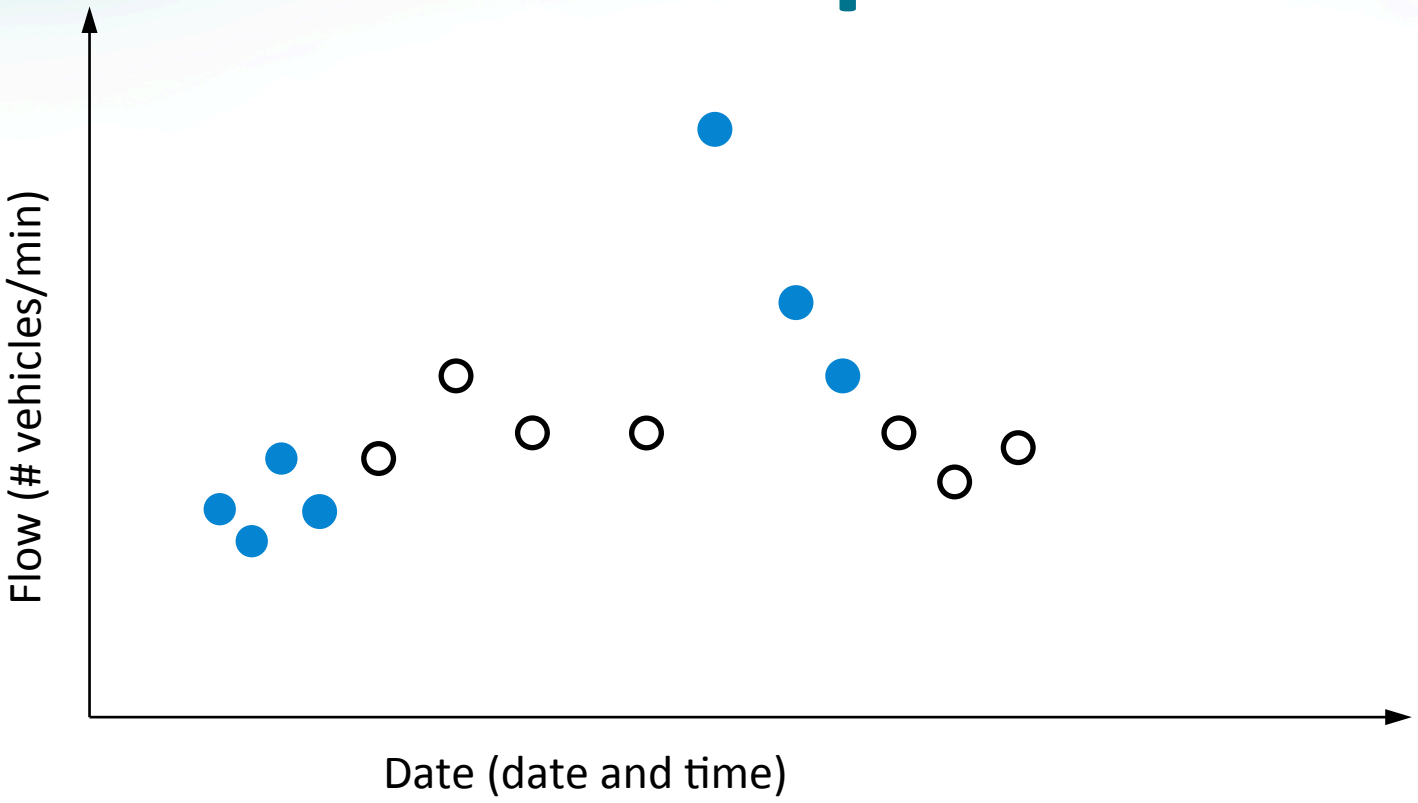
Error Propagation: Which errors are **more problematic**?



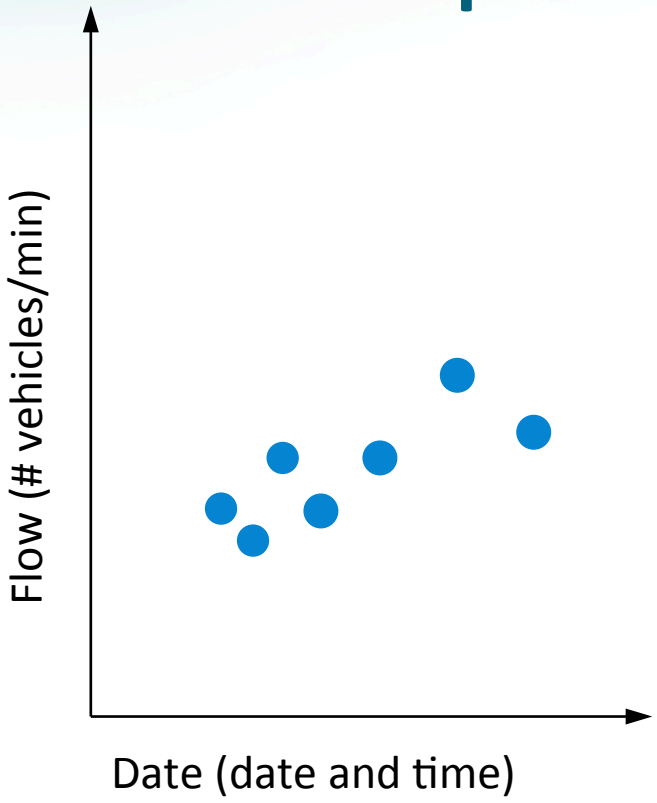
Error Propagation: Which errors are **more problematic**?

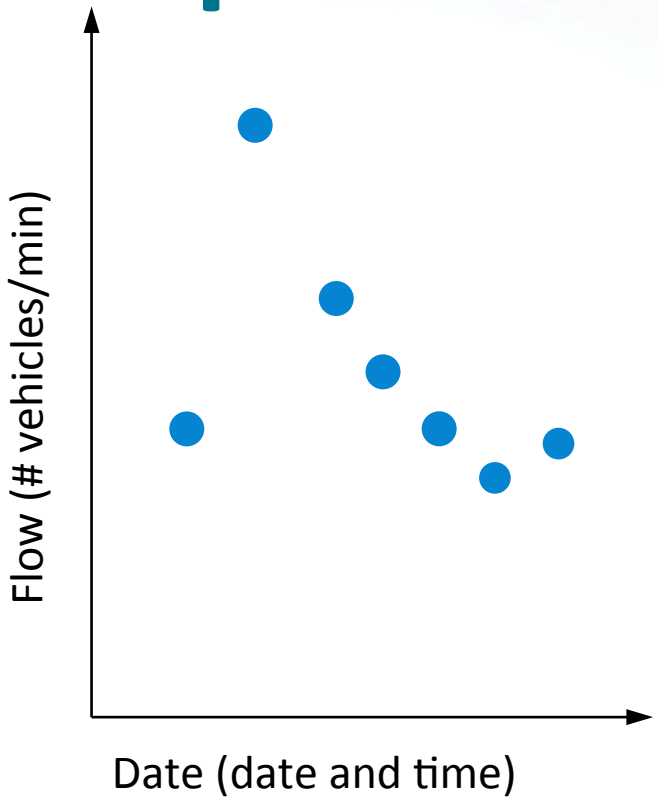


Error Propagation: Which errors are **more problematic**?

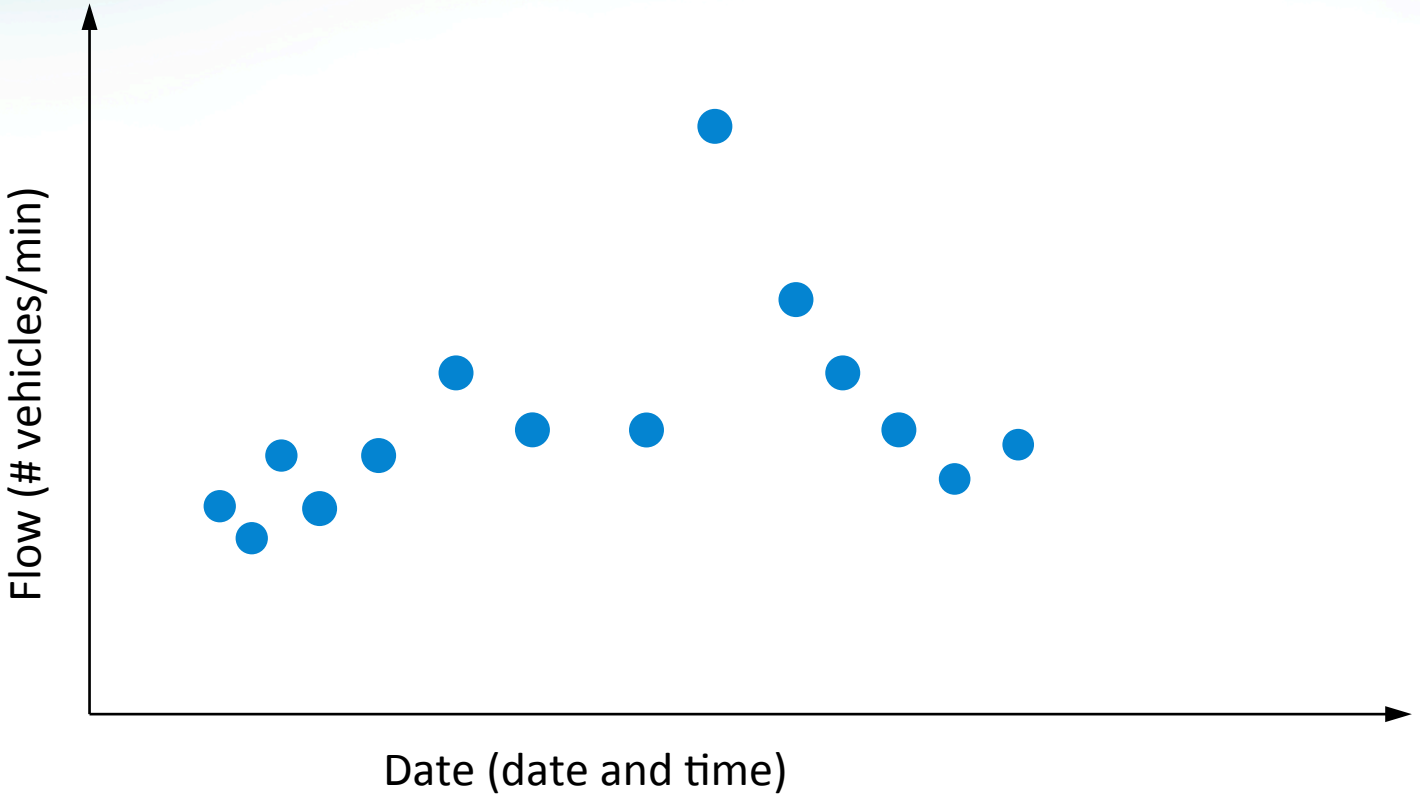


Replication: What errors inhibit experiment **replication**?

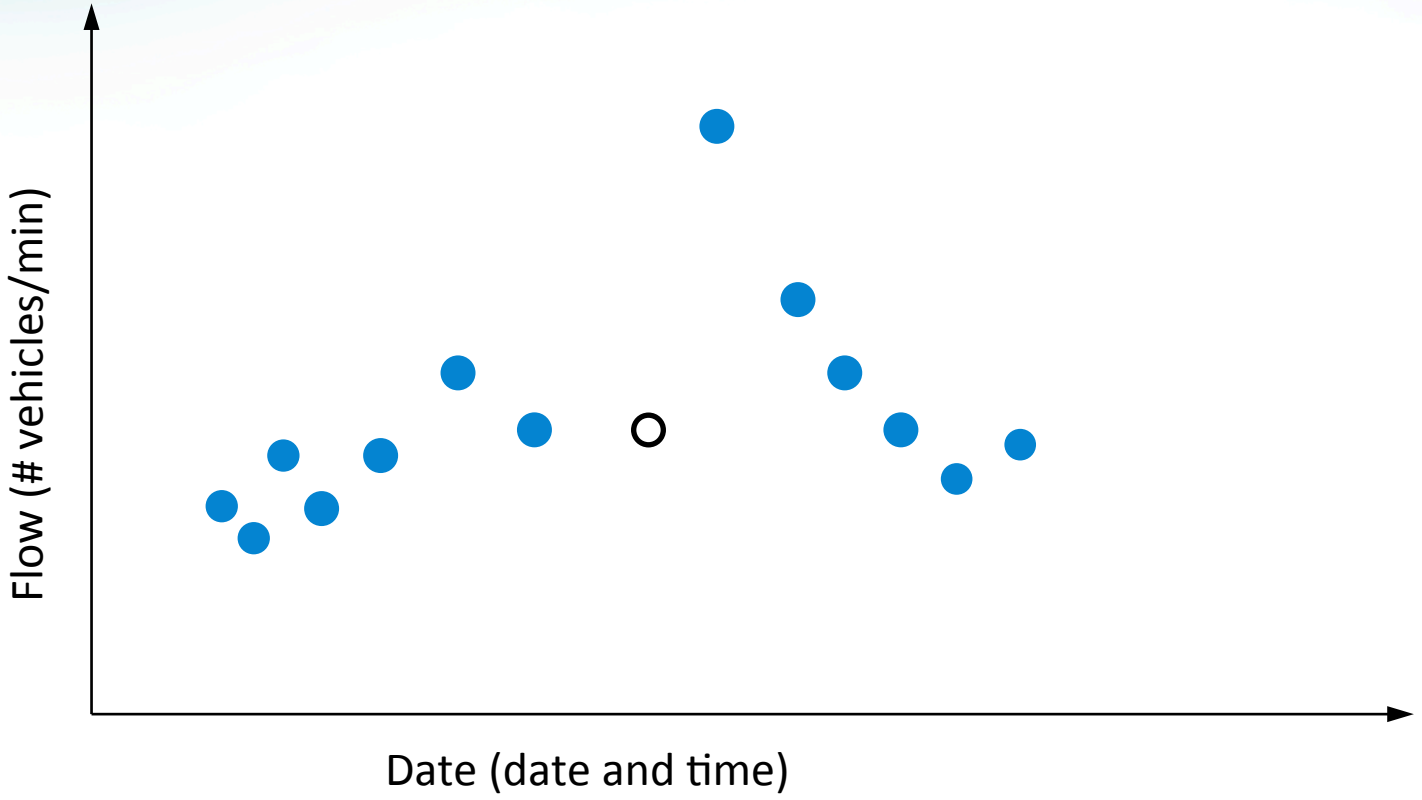




Provenance: **Which points** were cleaned?



Provenance: **Which points** were cleaned?



Alternative and Related Terms

- Data Cleaning
- Data Quality
- Data Processing
- Extract, Transform, Load
- Anomaly Detection
- Outlier Detection
- Uncertain Data
- Probabilistic Databases

Discussion Questions

- How do we tell when data is cleaner?
- What errors in data are more problematic? What algorithms are more robust to errors?
- What errors in data inhibit experiment reproduction, and how do we design experiments to mitigate the effects of these errors?
- How do we track data to identify which points have been cleaned and how they have been changed?