Data Cleansing

School of Information Studies
Syracuse University

Data Cleansing/Scrubbing

- The process of identifying and correcting bad data
- Trivial cases:
  - Replacing nulls with defaults
  - Fixing case: MIKE → Mike
  - Formatting: 3154432911 → 315-443-2911
- Advanced cases (matching)
  - Regular expressions: e-mails, IP addresses
  - Lookups on a business key
  - Fuzzy matching: Do not → Don't
  - Rule-based: [Bill, Will, William, Billy] → Bill

External Sources

- Use external datasets or web APIs to perform validation of common data.
- Examples:
  - E-mail address validation
  - Address validation
  - Postal/Zip codes
  - Country names to codes
  - GeoIP lookup/IP address validation
  - Credit card validation/Luhn check
  - Phone number validation

Error Event Schema

- A centralized dimensional model for logging failed data quality screens.
- Fact table grain is an error event.
- Dimensions are date, ETL job, quality screen source.
- A row added whenever there is a quality screening event that results in an error.
- Schema can also be used for warnings or fixes.