Lesson 5: Data Qualiyy

**Data Cleaning**

* Data cleaning is **iterative** 🡪 detect THEN correct bad data
* May be designed incorrectly, text where we should have numeric data, may be numbers “out-of-range”, may not be in correct structure, outliers to standard distribution, wrong date format
* Source of Dirty Data 🡪 any time there is human involvement
* User entry errors
* No or poorly applied coding standards
* Different schemas used for same type of data (eyeColor vs. eye\_color)
* Legacy systems
* Evolving applications + systems 🡪 business needs change so data changes
* No unique identifiers
* Lost data during transformation
* Programming error
* Corruption in transmission
* Formal Measures of Data Quality
* **Validity** 🡪 measure to which entries in a dataset conforms to a defined schema/other constraints we may have
* **Accuracy** 🡪 degree to which entries conform to gold-standard data (all street addresses in dataset actually exist 🡪 must compare our data to data we trust)
* **Completeness** 🡪 do we have all records we should have? 🡪 actually measuring is difficult
* **Consistency w/in data 🡪** may have multiple records w/ some overlap in contained data
* Want to ensure consistently among fields that represent the same data across systems
* **Uniformity** 🡪 all values for certain things use same units
* Blueprint for Data Cleaning
* **Audit** data **🡪** programmatically check data using defined rules + create report on data quality
* May also run statistical analysis to check for outliers
* Use audit info to create data cleaning plan 🡪 ID any causes of dirty data (situation-specific)
* Define set of operations to correct data (situation-specific)
* Test to make sure data cleaning plan does what we need it to do
* Execute the actual plan (script to run defined operations)
* It’s (most) likely we cannot programmatically clean data entirely, so we manually correct it somehow (if necessary)
* May have to iterate this whole process multiple times since humans may be involved again at the final step