DMDB
WiSe 2019/20
Ticket to Class 5

Name: Moser Markus

Questions based on the Udacity course "Data Wrangling with MongoDB", lession 5.

1. What are common sources of dirty data?

   - user entry errors
   - no/poorly applied coding standards
   - different schemas
   - Legacy systems (encoded differently, with disk space in mind)
   - evolving applications
   - no unique identifiers
   - data migration (cost in tranformation)
   - programmer error
   - corruption in transmission

2. What are 5 measures for data quality?

   5 • validity: conforms to a schema
   3 • accuracy: conforms to gold standard
   1 • completeness: all records? (hard to measure)
   4 • consistency: matches other data
   2 • uniformity: same units (distance, weight, etc.)

guessed rank of difficulty (1=hardest)

3. What is the suggested blueprint for cleaning?

1. audit your data (programmatically check quality => report, maybe statistical analysis for outliers)
2. create a data cleaning plan
   (a) identify causes (of dirty data)
   (b) define operations (that will correct data)
   (c) test
3. execute the plan
4. manually correct remaining
5. [go back to step 1] iterate, maybe 2+ times

4. *Auditing validity* is about determining what the

constraints

are on individual fields and checking to make sure the field values

adhere to those constraints.

- mandatory/unique fields (in online form)
- foreign key constraints
- cross field constraints (startday before endday)
- datatype/range
- regex
- set membership/enum