SDAL SOCIAL & DECISION ANALYTICS LABORATORY

Data Science for the Public Good

CHARACTERIZING ADMINISTRATIVE DATA QUALITY: A NEW TOOL

Adrienne Rogers (Virginia Tech)
Project Team : Aaron Schroeder (SDAL), Lin Tan, Isabel Bradburn (Human Development, VT)

Overview

Nationally, states collect student records from all public schools for reporting and accountability purposes.

- These data are available **through statewide longitudinal data systems (SLDS)**.
- States use SLDS data to conduct research to inform policy.
- **Virginia Tech (VT) developed a template** to help researchers better understand the administrative data elements they are interested in and better select the data for use.
- Working with the Virginia Department of Education (VDOE), VT iteratively tested and refined the template.

VLDS

Student Record Collection
VLDS Data Collection reporting requirements of the No Child Left Behind (NCLB) Act. Includes student demographics, special needs information, classroom environment, etc.

Data set Summary:
Variable Count: 95 Total Record Count: 42,134,526

Creating the Template

Creating Rules

Created a set of rules for each variable in the dataset by using the data dictionary. These rules were coded using SQL.

Ex. “SELECT variable FROM data set WHERE variable IS NULL’”

Automating Publication

In order to apply the code for these rules to a variety of variables, the template was created in R Markdown to allow the user to rerun the document for different variables by only changing the ‘column name’ section.

Formatting Design

Using HTML code for layout and R for plots enabled a systematic, easy-to-use format

Template Data Element Checks

Completeness	the number of missing values and the percent not null
Validity	the number of values not defined by the data dictionary and the percent of those which are
Uniqueness	the frequency of unique values
Record Consistency	the number of records not behaving by the rules defined in the data dictionary and the percent which are behaving
Longitudinal Consistency	the percent of variables which remain appropriately consistent across record collection periods

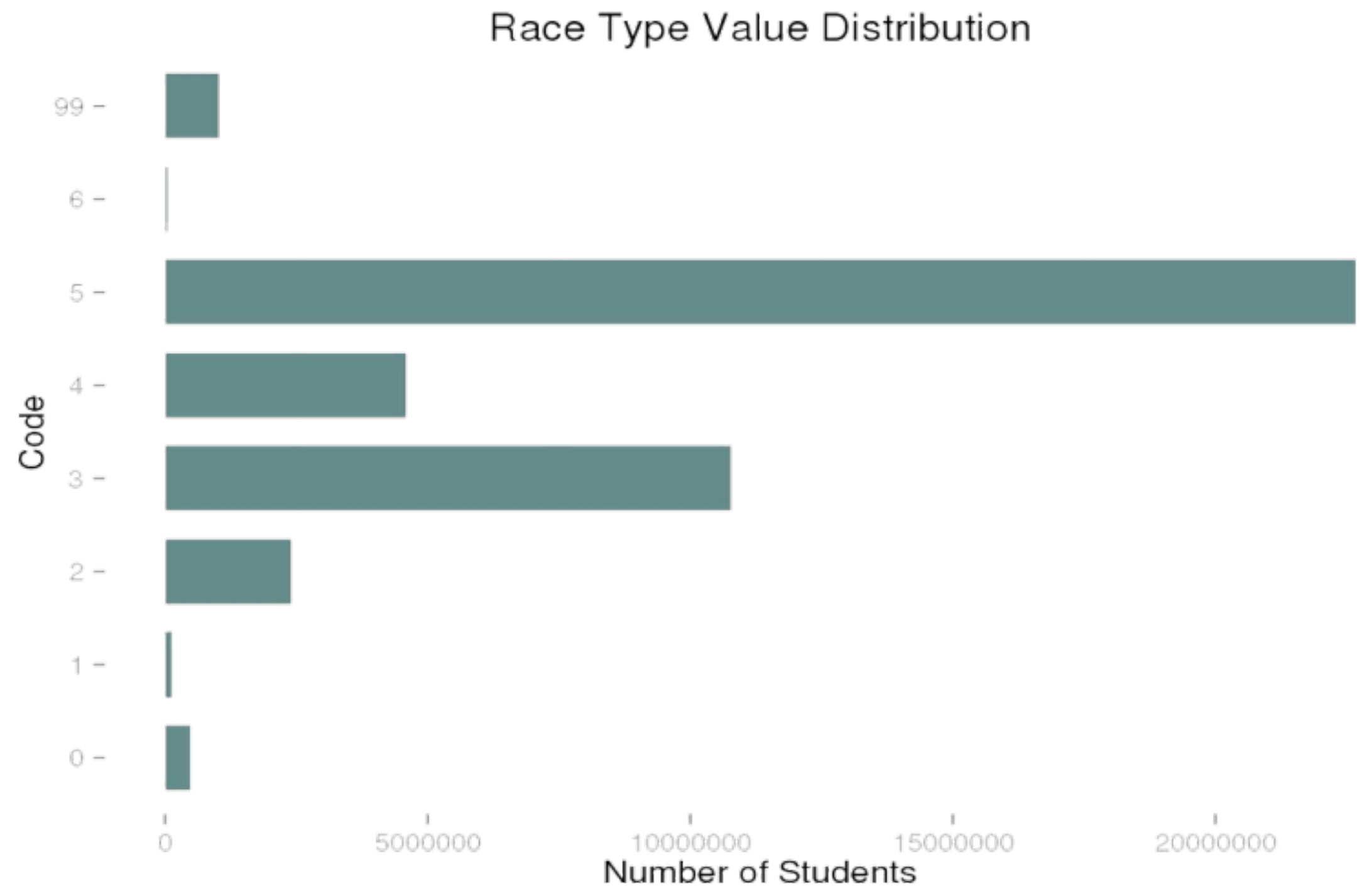
“Race Type” Variable Summary Example

Definition: a code for the one or more races a student identifies with (often prone to changes over time)
Dates Collected: 2005-2015

- During the collection period, starting with the 2010-2011 school year, the “Race Type” definition expanded to allow students to choose multiple races and separately specify Hispanic ethnicity.
- In this process categories were eliminated. Eliminated categories included the ‘Unspecified or Unknown’ category and the ‘Hispanic’ category. A student with Hispanic ethnicity is now captured under the ethnic flag variable.
- Due to this change, “Race Type” is one of the few variables which undergoes all data element checks, including both record and longitudinal consistencies. “Race Type” is expected to have longitudinal inconsistencies due to both the change and the nature of personal reporting.

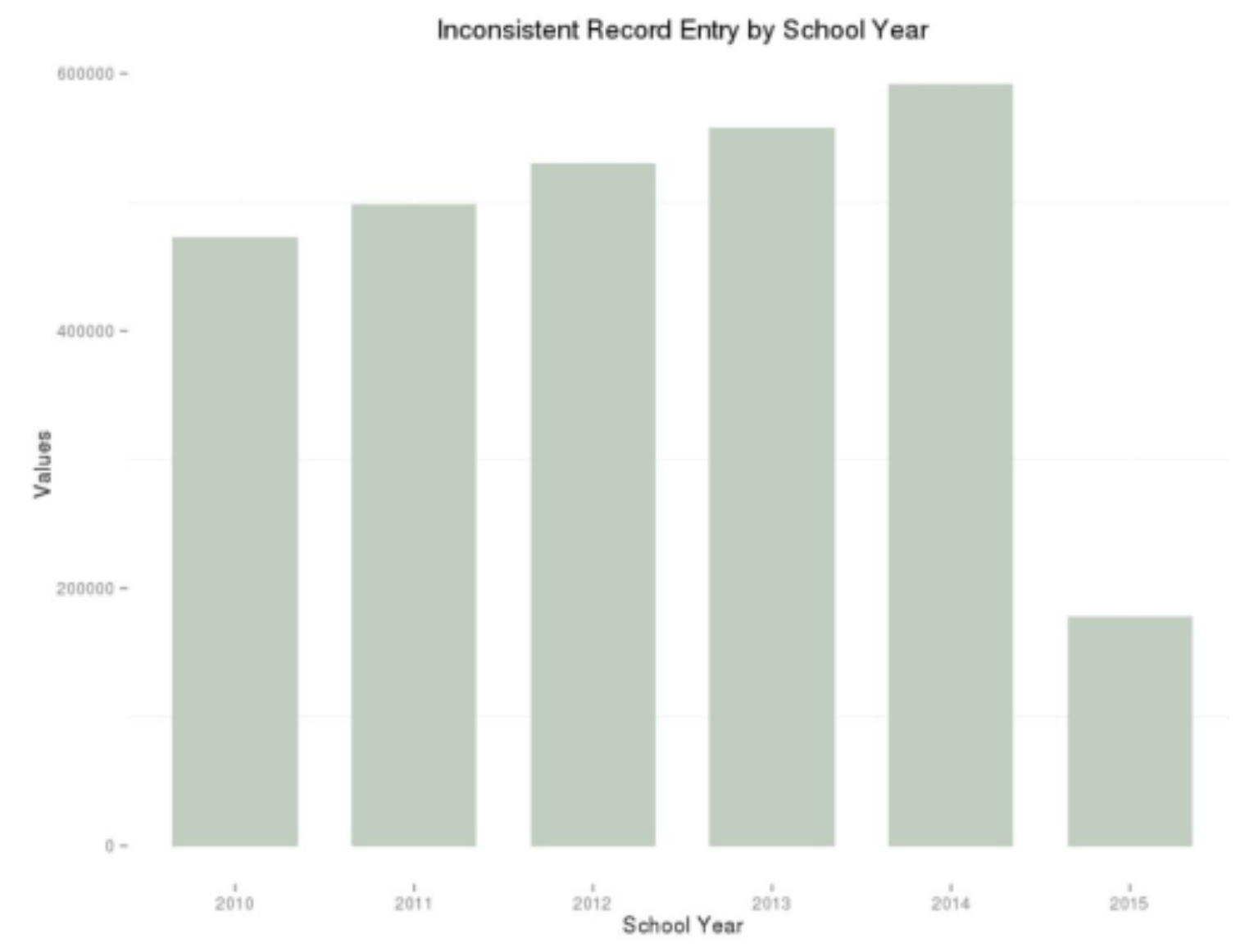
Number of Unique Values: 8

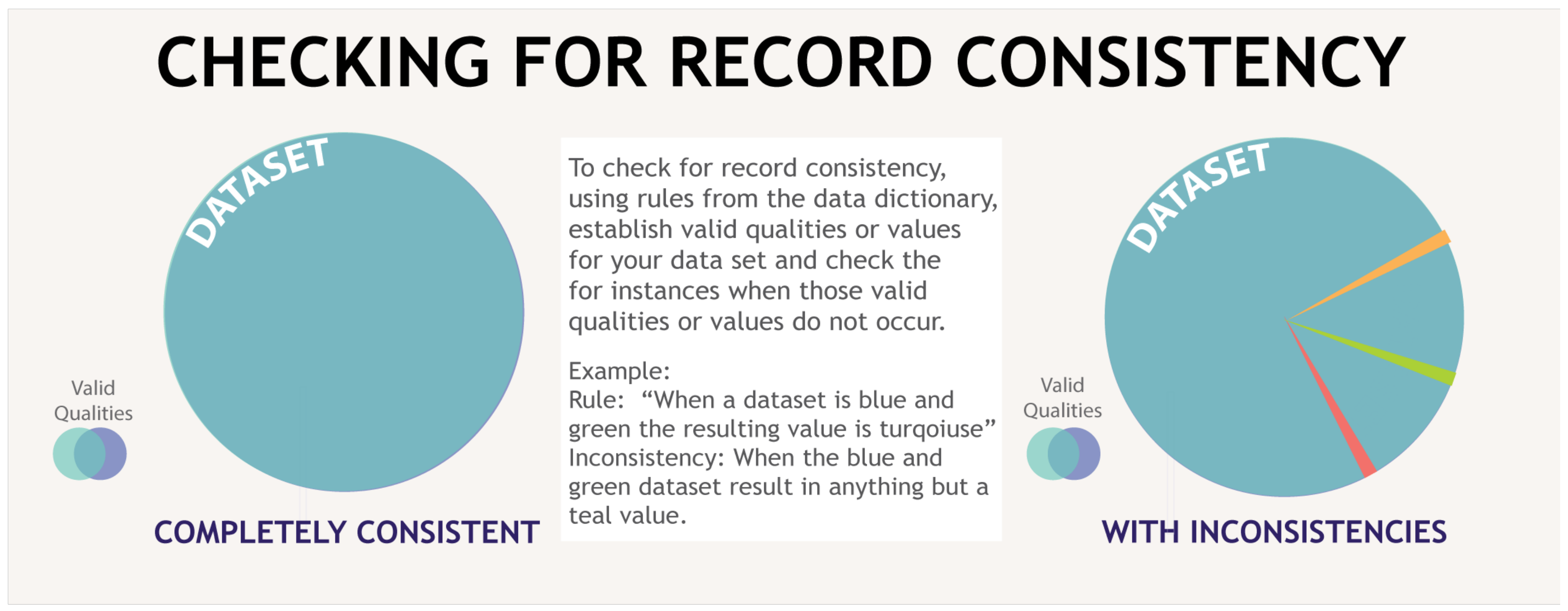
value	value_description
99	Multiple race types reported
1	American Indian or Alaskan Native
2	Asian
3	Black or African American
4	Hispanic (valid before 2010 school year)
5	White
6	Native Hawaiian or Other Pacific Islander
0	Unspecified or Unknown (valid before 2010 school year)



Defining and Testing Record Consistency Rules

- Variables are often interrelated. In the case of the variable “Race Type”, certain codes cannot be used for certain school years, due to changes in the recording process. To ensure data quality, the dataset must be checked against these rules.
- For “Race Type”, record consistency shows how quickly the reporting changes are catching on and being properly recorded.





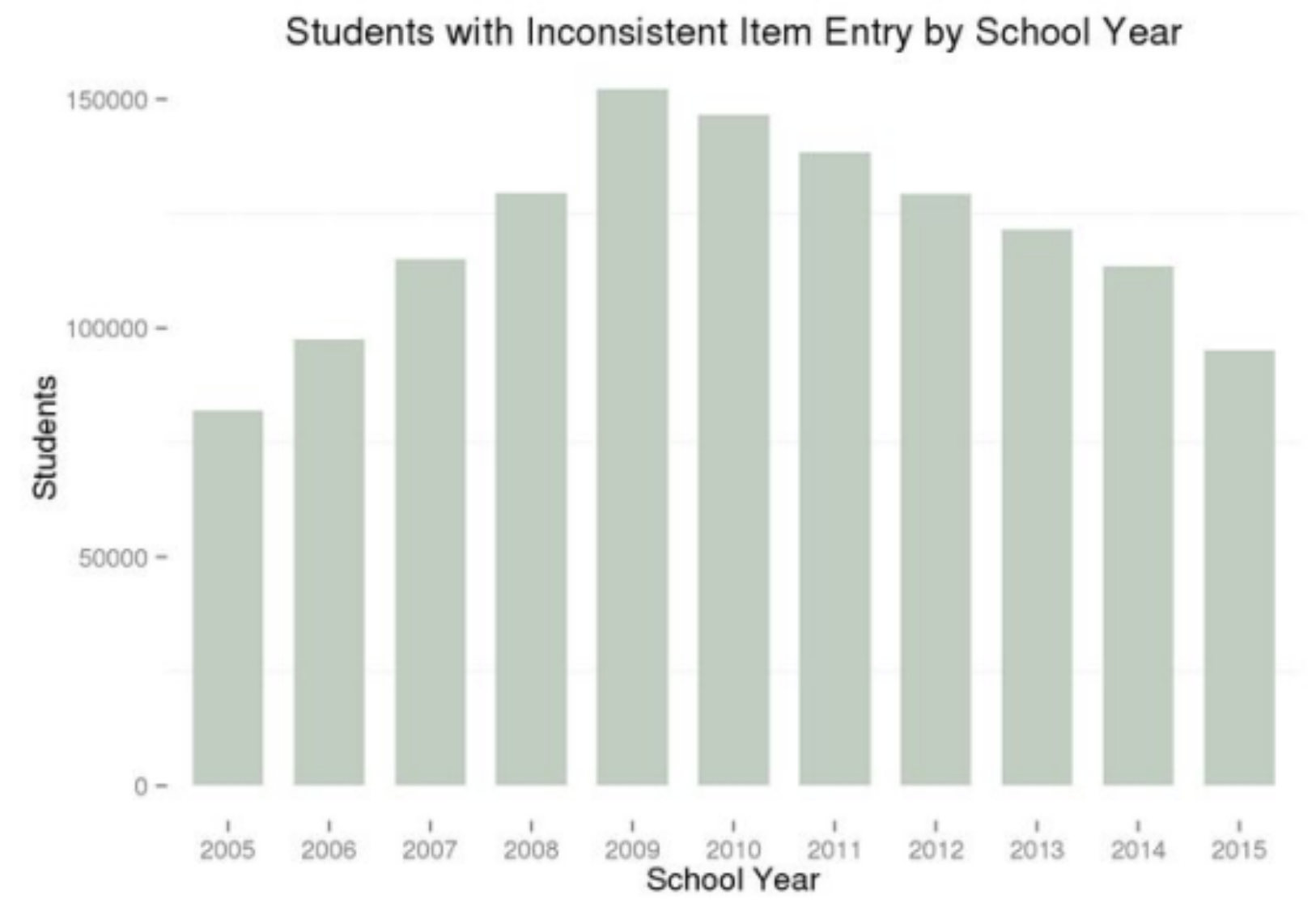
CHECKING FOR RECORD CONSISTENCY

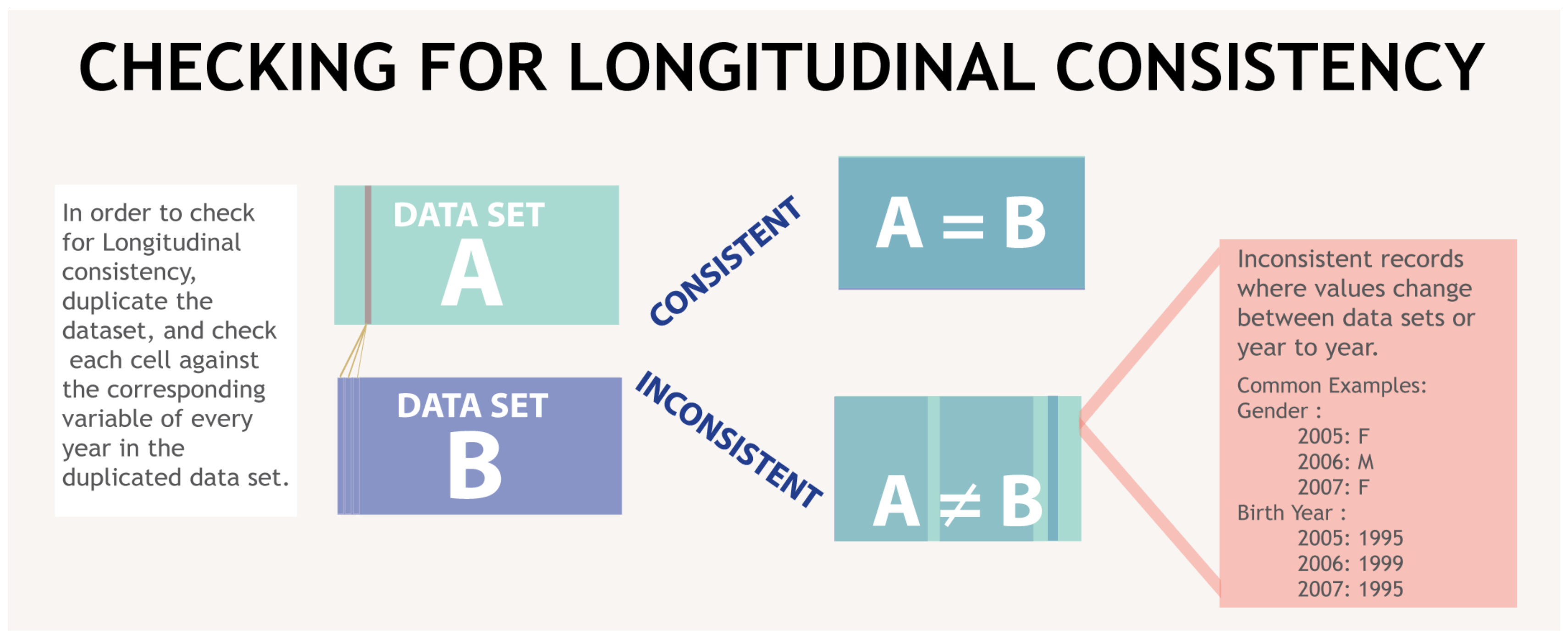
To check for record consistency, using rules from the data dictionary, establish valid qualities or values for your data set and check the for instances when those valid qualities or values do not occur.

Example:
Rule: “When a dataset is blue and green the resulting value is turqoiuse”
Inconsistency: When the blue and green dataset result in anything but a teal value.

Testing Longitudinal Consistency

- For longitudinal studies, it is possible to see if demographic records are the same every year, e.g. age, gender, grade.
- For “Race Type”, longitudinal consistency is assessed to see whether records reacted to the changes in reporting policy.
- Around 2010 (the year the Decennial Census is collected), the data show the greatest number of inconsistencies and then decline steadily over time.





CHECKING FOR LONGITUDINAL CONSISTENCY

In order to check for Longitudinal consistency, duplicate the dataset, and check each cell against the corresponding variable of every year in the duplicated data set.

Inconsistent records where values change between data sets or year to year.

Common Examples:
Gender :
2005: F
2006: M
2007: F
Birth Year :
2005: 1995
2006: 1999
2007: 1995

VirginiaTech Biocomplexity Institute

Acknowledgment: NSF REU, supplement to *Building Research Infrastructure and Community to Advance Social Scientific and Educational Use of Administrative Data (#1338491)*

bi.vt.edu/sdal