

Enhancing Qualitative Data Coding in Federal Statistics: A Human–Machine Learning Framework

Judy H. Tang

Senior Research Associate,
Survey Management
and Design

Tom Krenzke

Vice President,
Statistics and
Data Science

Jin Hui Xu

Lead IT Associate,
Technology and
Digital Solutions

Karen Lo

Principal IT Associate,
Technology and
Digital Solutions

Challenge

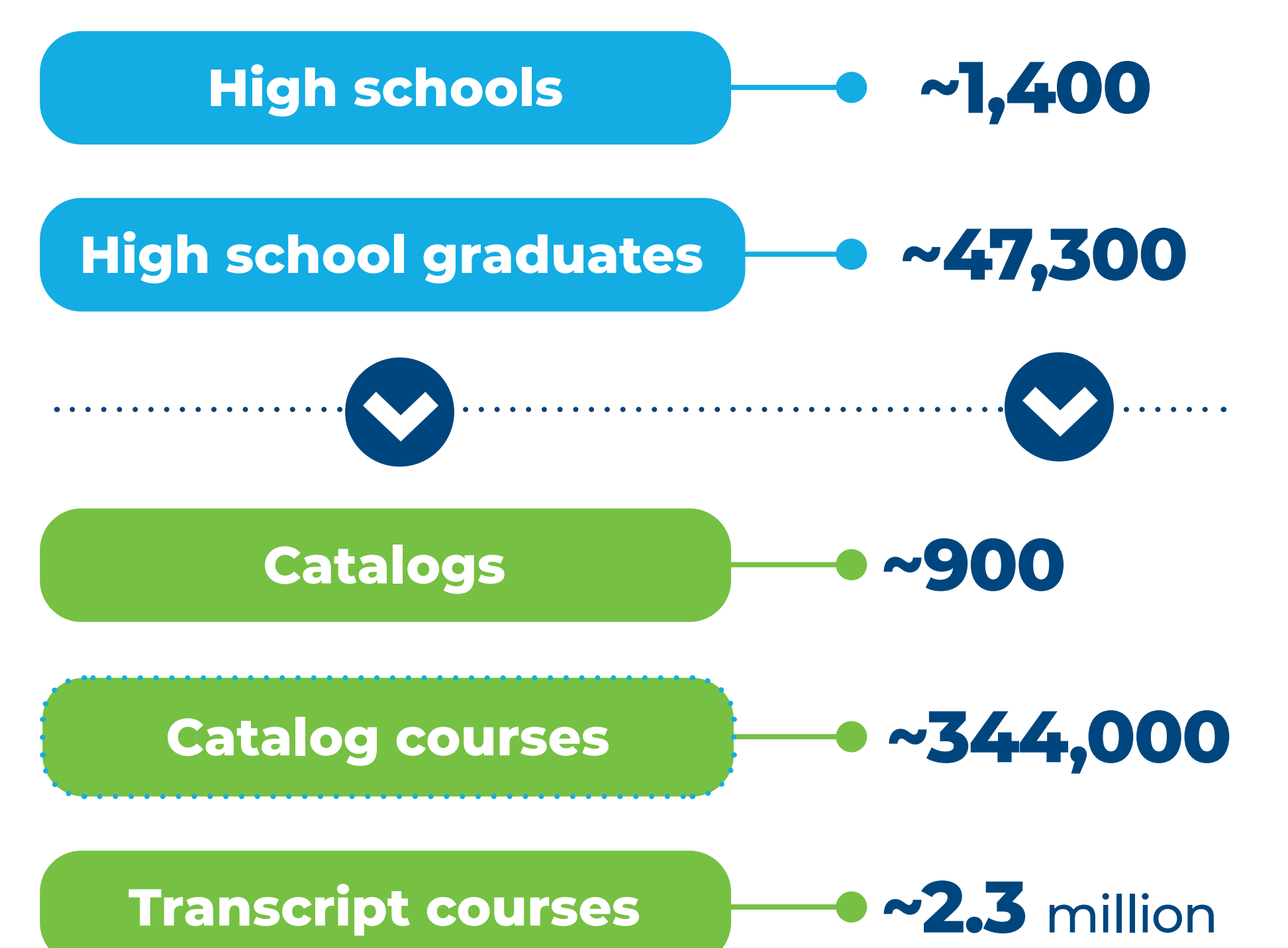
Qualitative data coding remains a persistent challenge in large-scale federal data collections. Transforming unstructured, locally defined information into standardized variables introduces:

- › Variability in terminology and descriptions
- › Ambiguity in interpretation
- › Inconsistency across coders
- › High reliance on human judgment
- › Scale of processing

Federal Context: NAEP HSTS

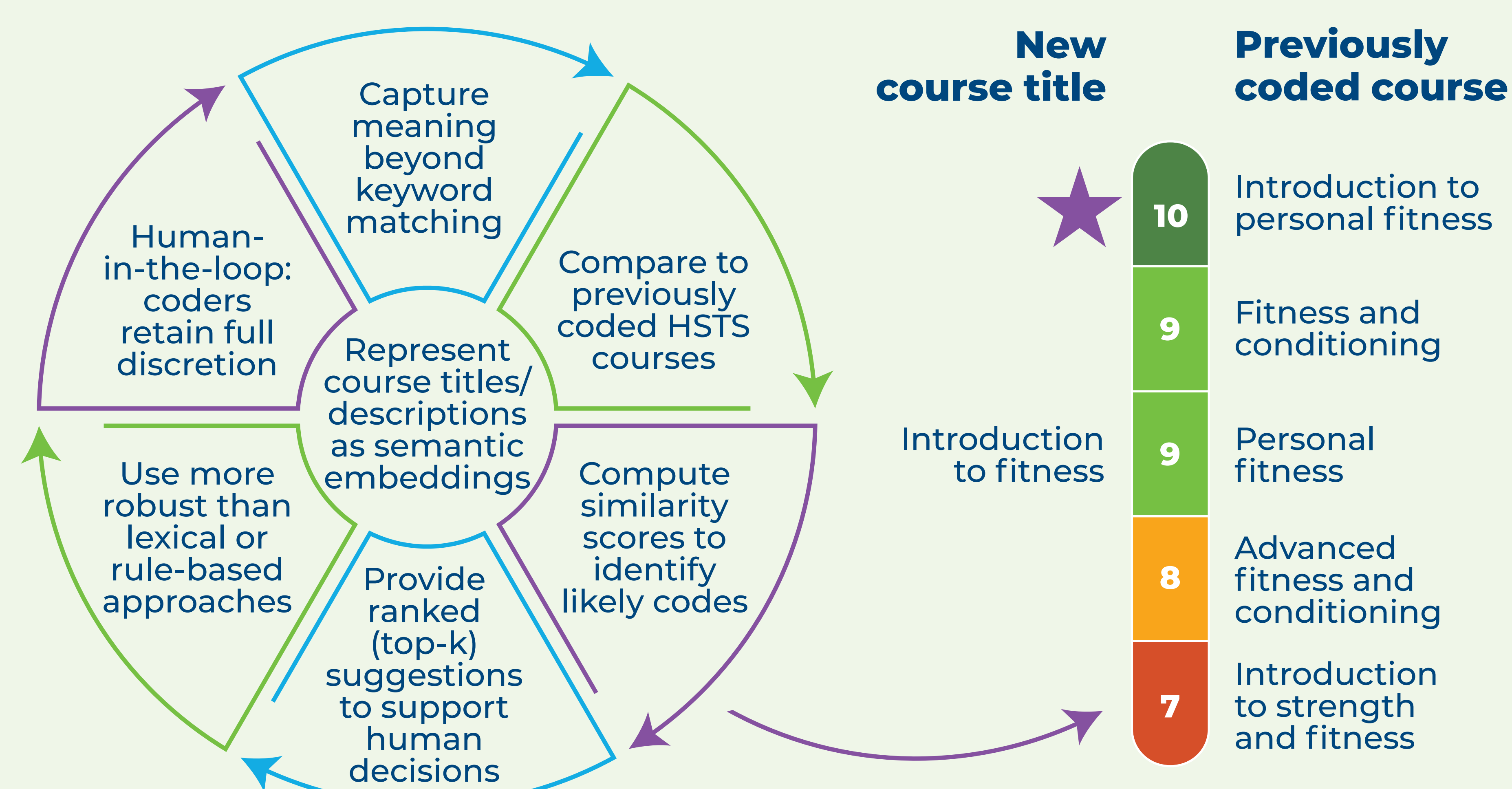
The NAEP High School Transcript Study (HSTS) is a national data collection conducted by NCES that focuses on understanding students' coursework and academic outcomes at the end of high school.

Approximate volume of data collected by HSTS in 2019:

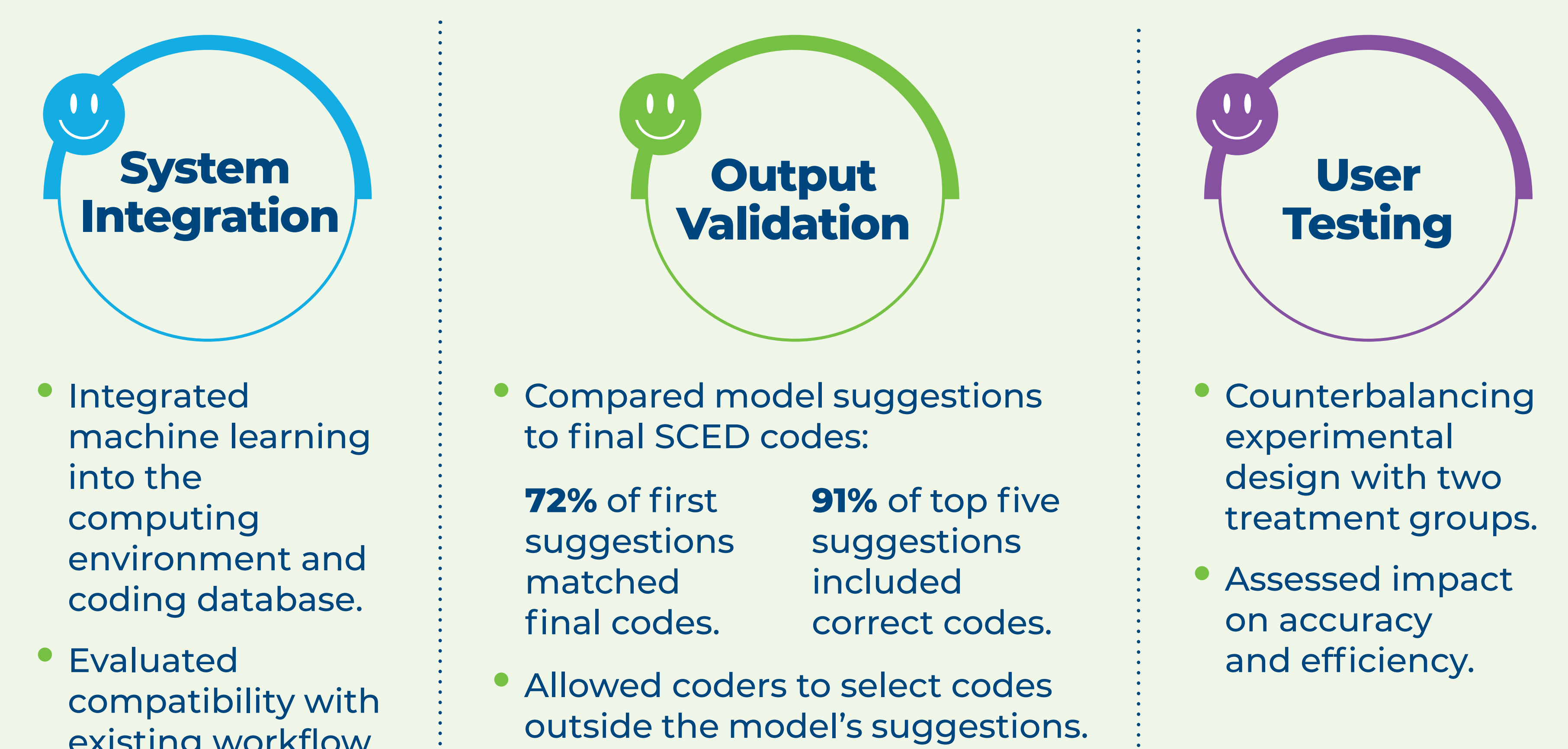


Sources: U.S. Department of Education, Institute of Education Sciences, National Center for Education Statistics (NCES), National Assessment of Educational Progress (NAEP) High School Transcript Study (HSTS), 2019.

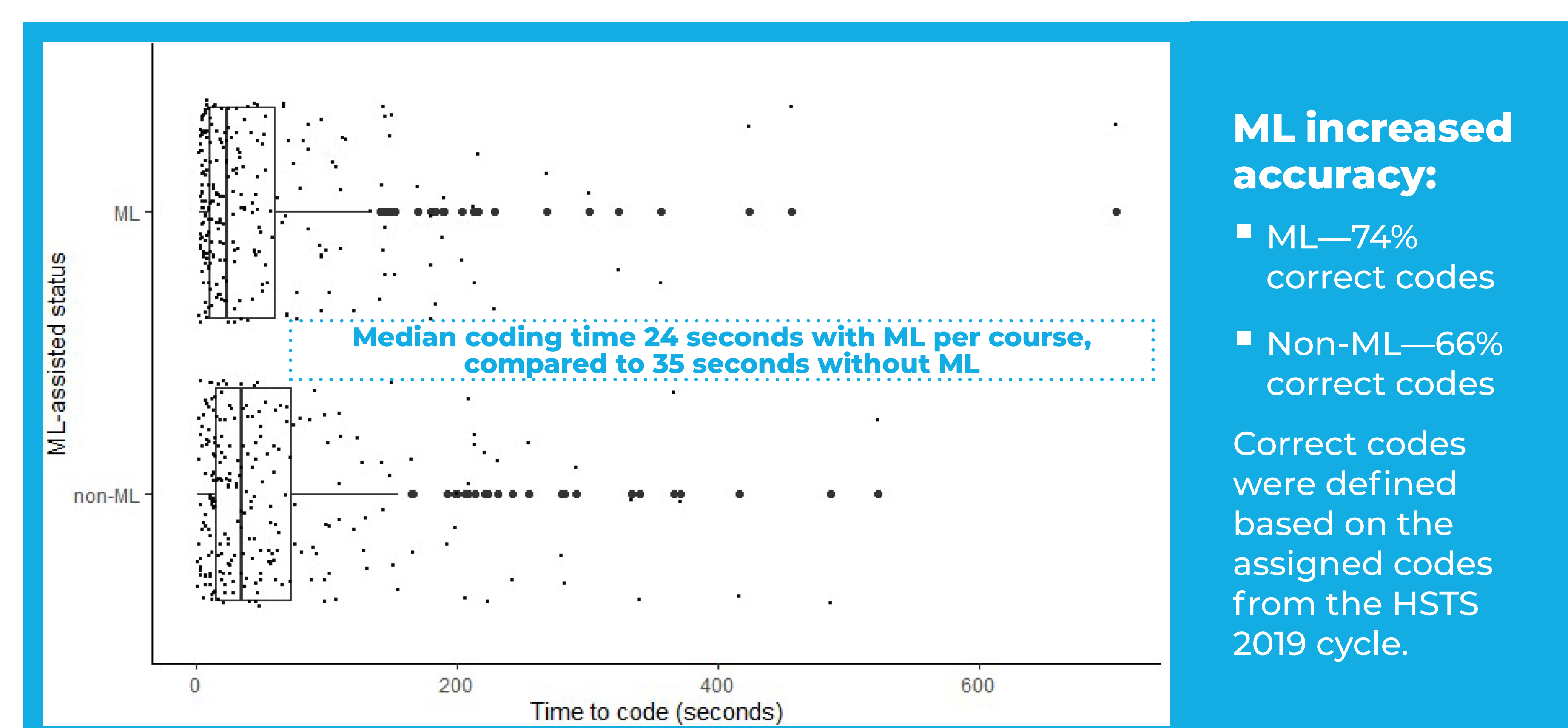
Human–Machine Learning Framework



Multi-Phase Experimental Design Evaluation



Impact on Coding Performance



Note: ML = machine learning

Conclusion: Insights and Success

- › Human plus machine learning improves performance: Outperforms either alone; top-k suggestions beat single predictions.
- › Human-in-the-loop is critical: Humans resolve ambiguous cases; performance reflects training data quality and coverage.
- › Human–machine learning framework improves federal data quality: Enhances accuracy, consistency, efficiency, reproducibility, and standardization.