Big Data: CMS's **Quality Improvement Evaluation System** (QIES) and **Proficiency Testing Database**

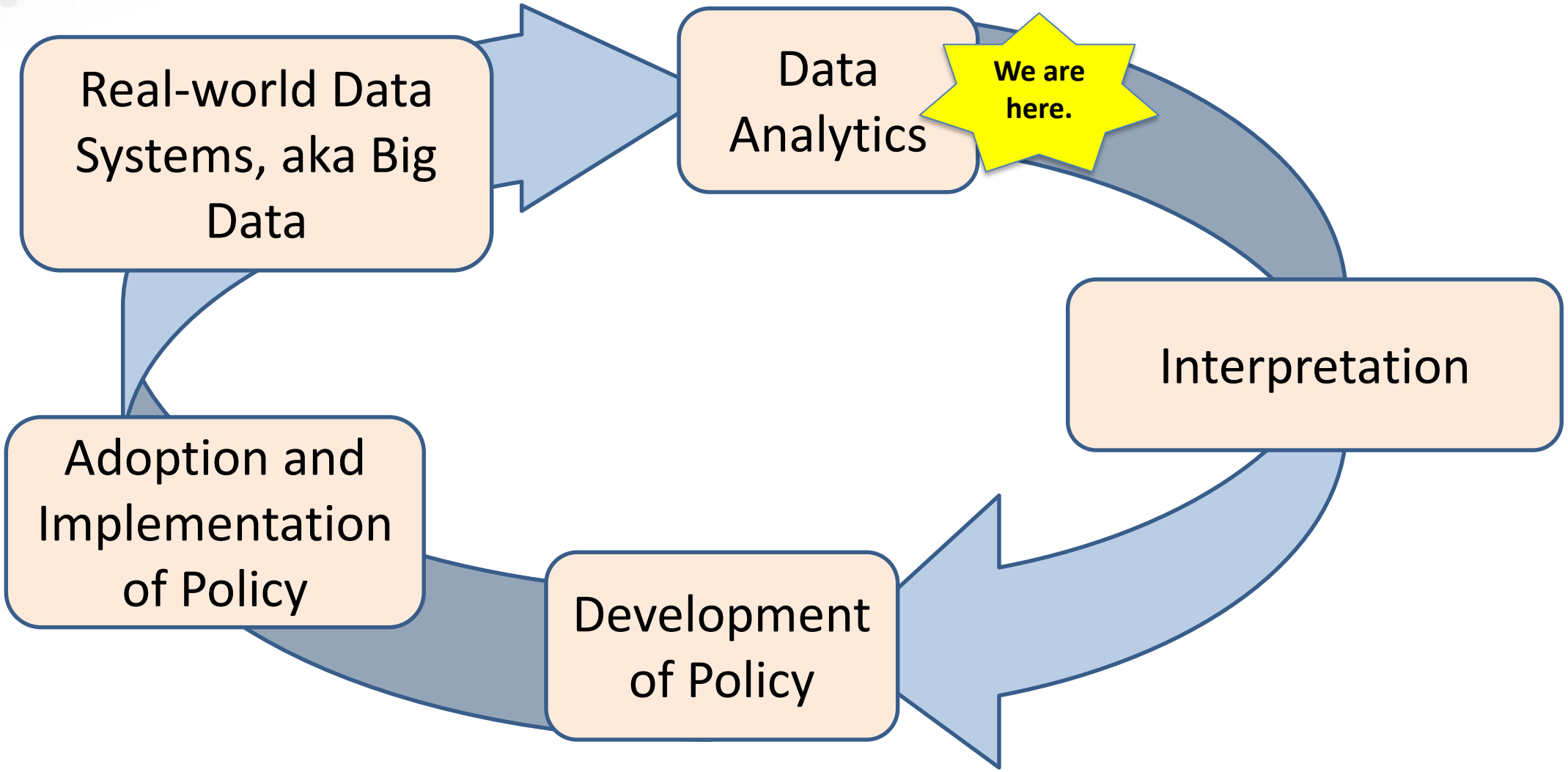
Thomas H. Taylor, Jr., P.E., M.S.
Mathematical Statistician
Division of Laboratory Systems

Excellent Laboratories, Outstanding Health

Scope of Talk – Data Analytics

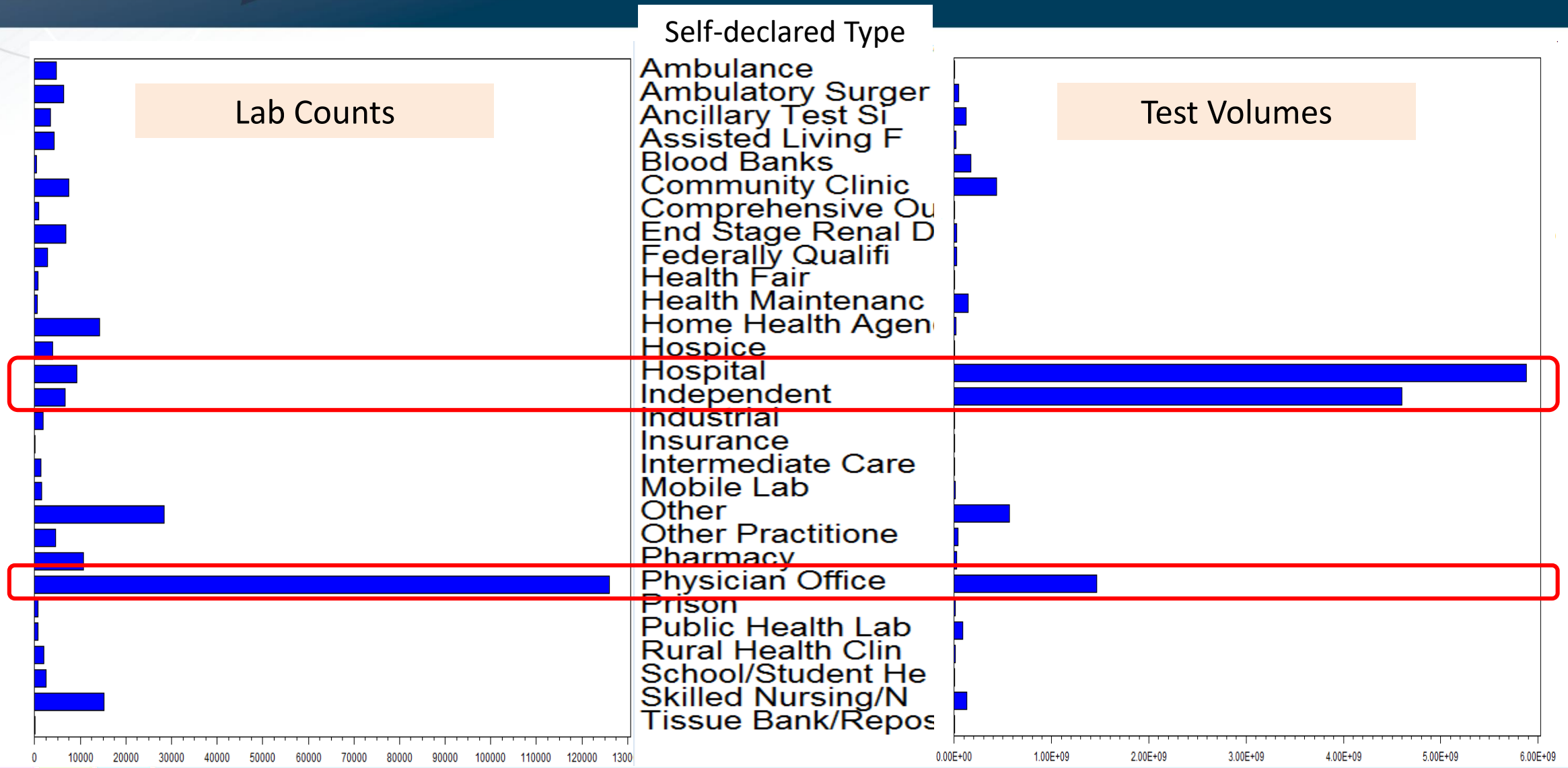
1. *Real-world Data* about CLIA Laboratories
2. CMS's QIES database *is Big Data*
   A. Certification Registration and Surveys
   B. Proficiency Testing (PT) Results
3. Challenges of using (any) Big Data
   A. Practical -IT Performance, Algorithm Development, Data Timeliness
   B. Scientific - Representative? Sample? Population? "Non-cases"
   C. "Relevance and reliability" (FDA's *Guidance* terms), aka Data Quality - Completeness, Correctness, Accuracy/Precision
   D. Regulatory – Informed Consent, HIPAA

Perspective on Data Analytics

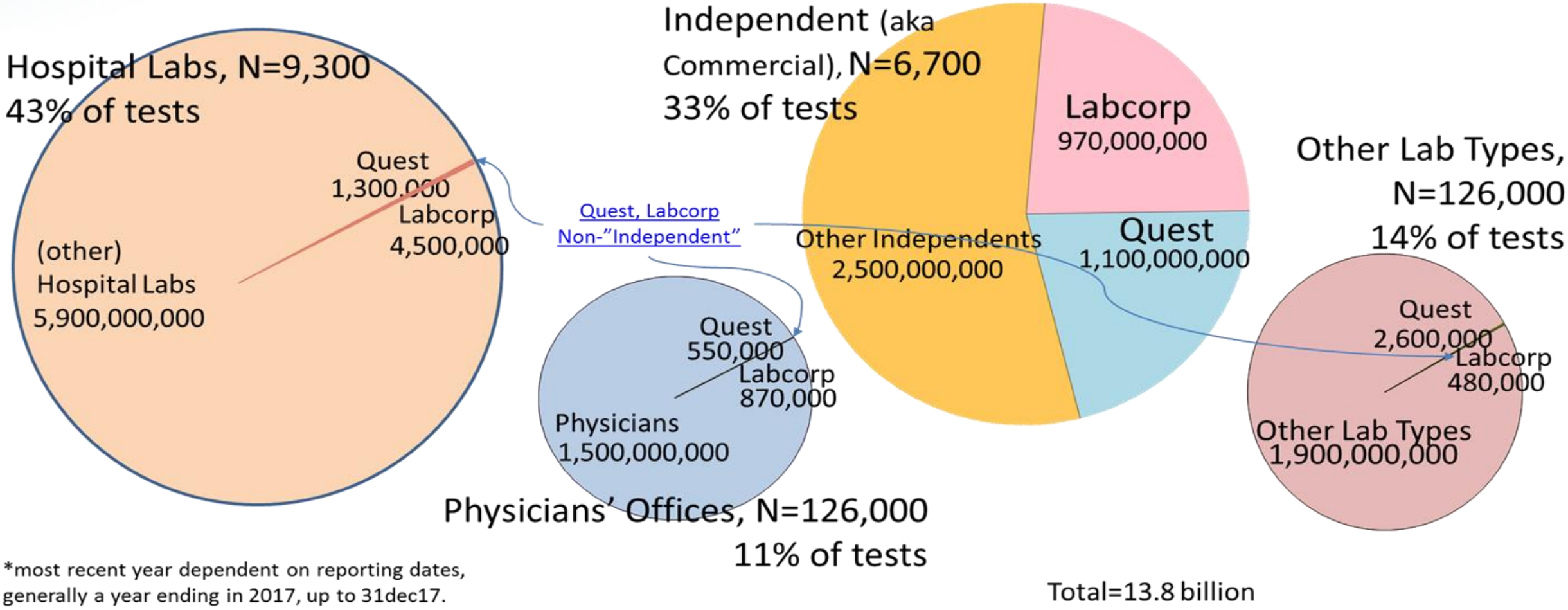


DIRECT OBSERVATIONS FROM QIES

Facility-Type: Lab Counts vs Test Volumes



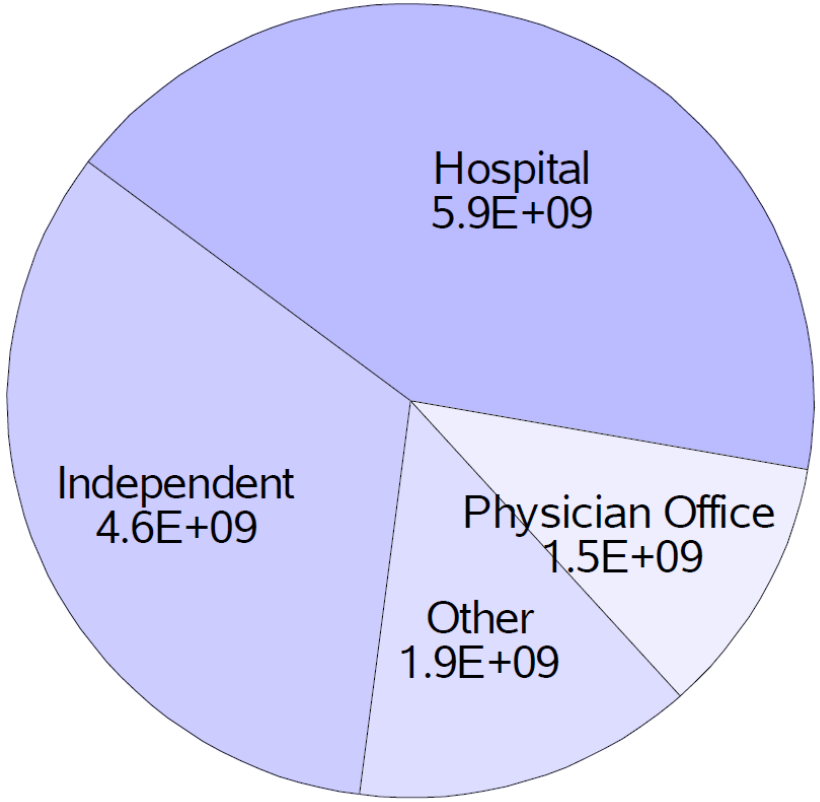
2017* Test Volume by Facility Type



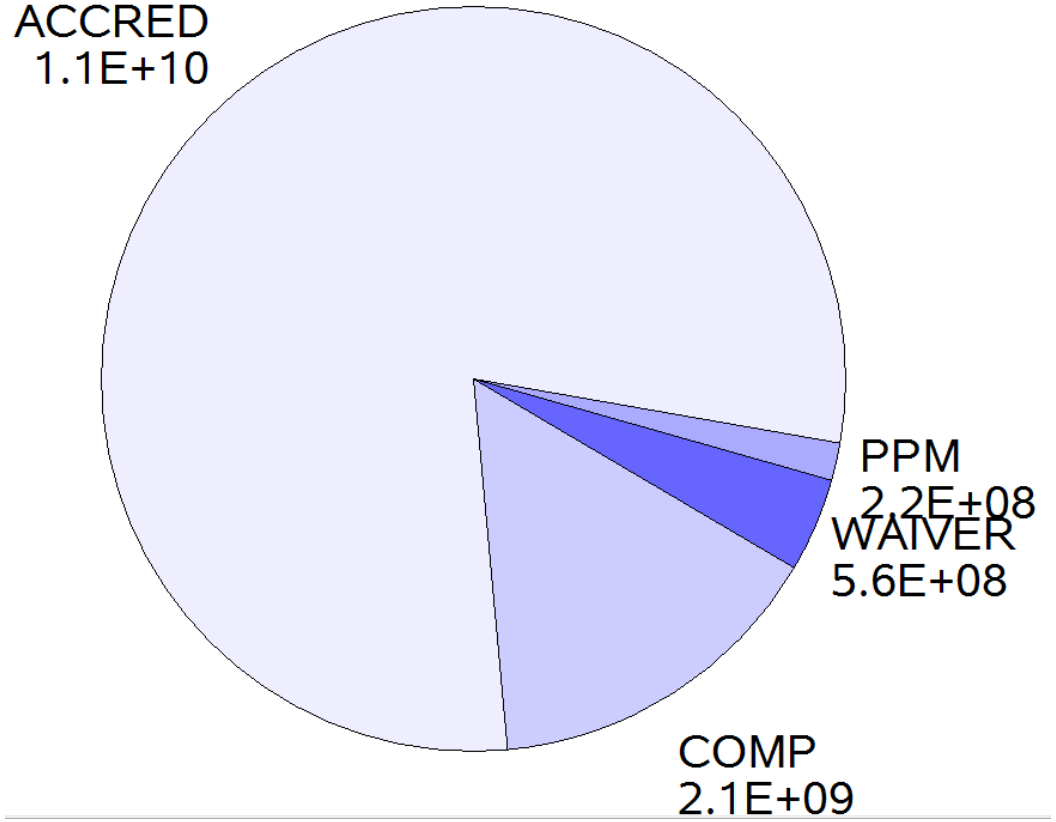
*most recent year dependent on reporting dates, generally a year ending in 2017, up to 31dec17.

Annual (2017*) Test Volume of CLIA Labs

By Facility Type EOY 2017



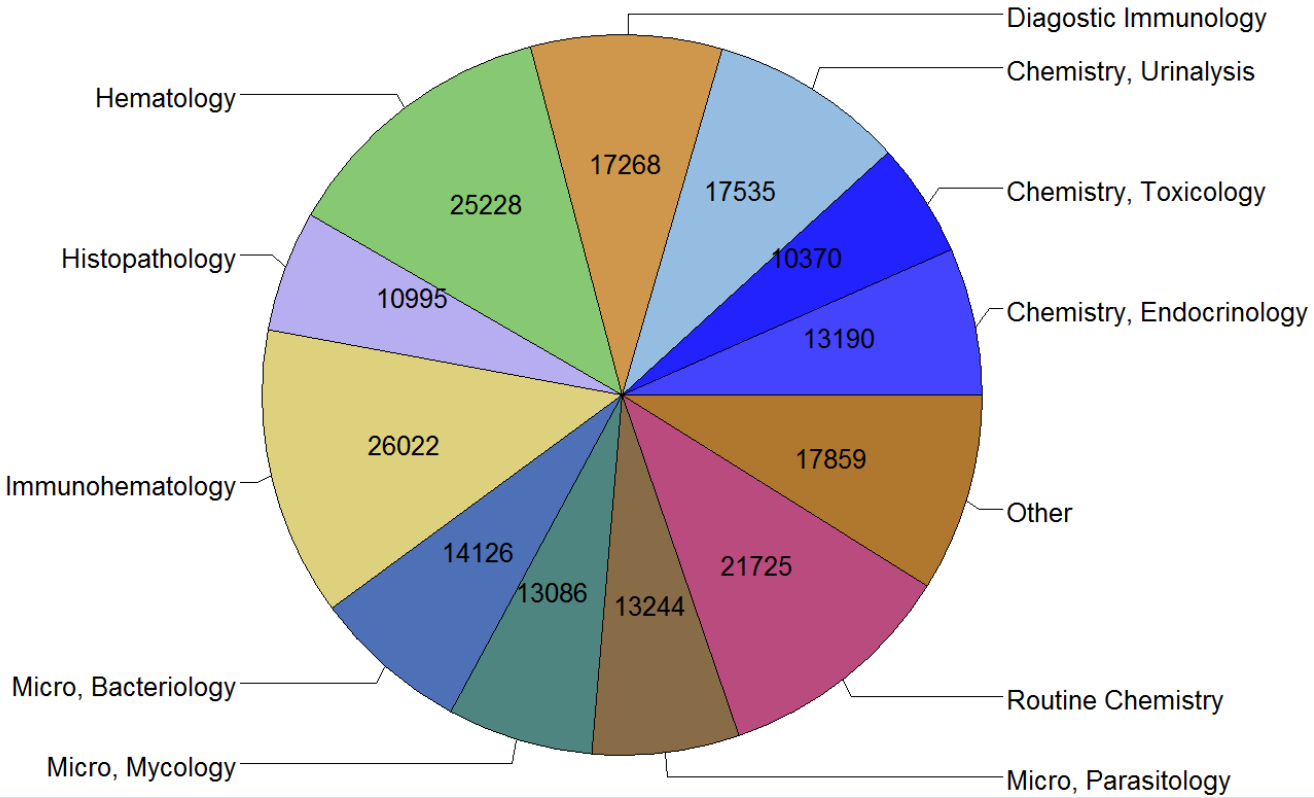
By CLIA Certificate Type EOY 2017



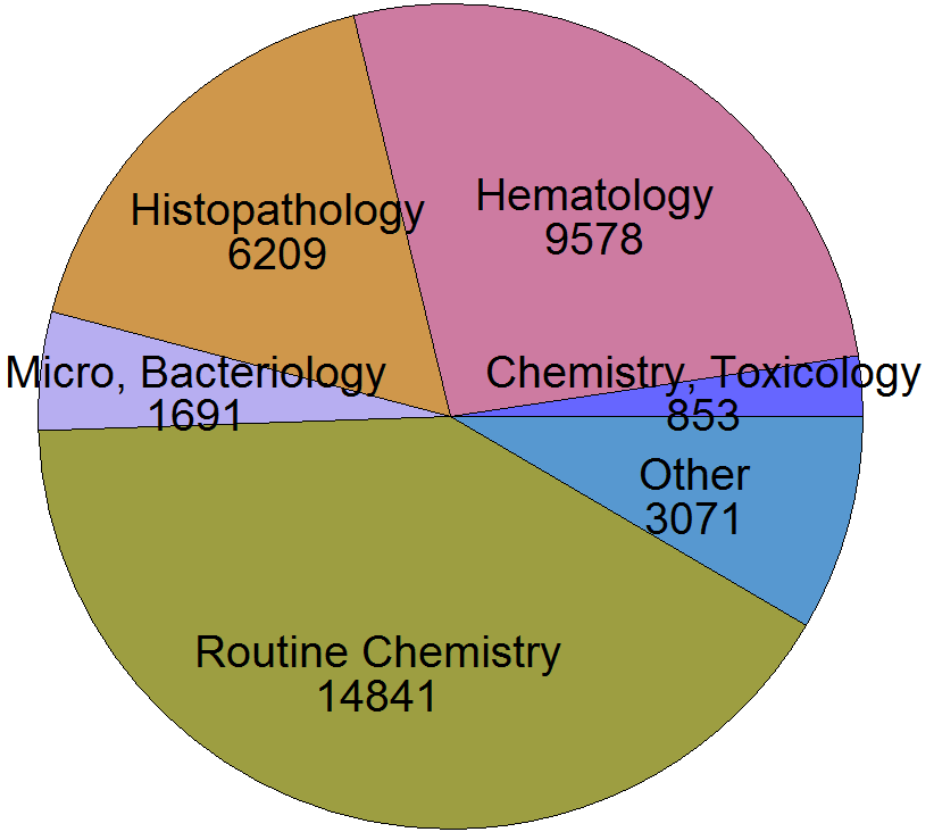
*most recent year dependent on reporting dates, generally a year ending in 2017, up to 31dec17.

Number of Labs by Specialties

Numbers of ALL Specialties Certified, Regardless of Volume

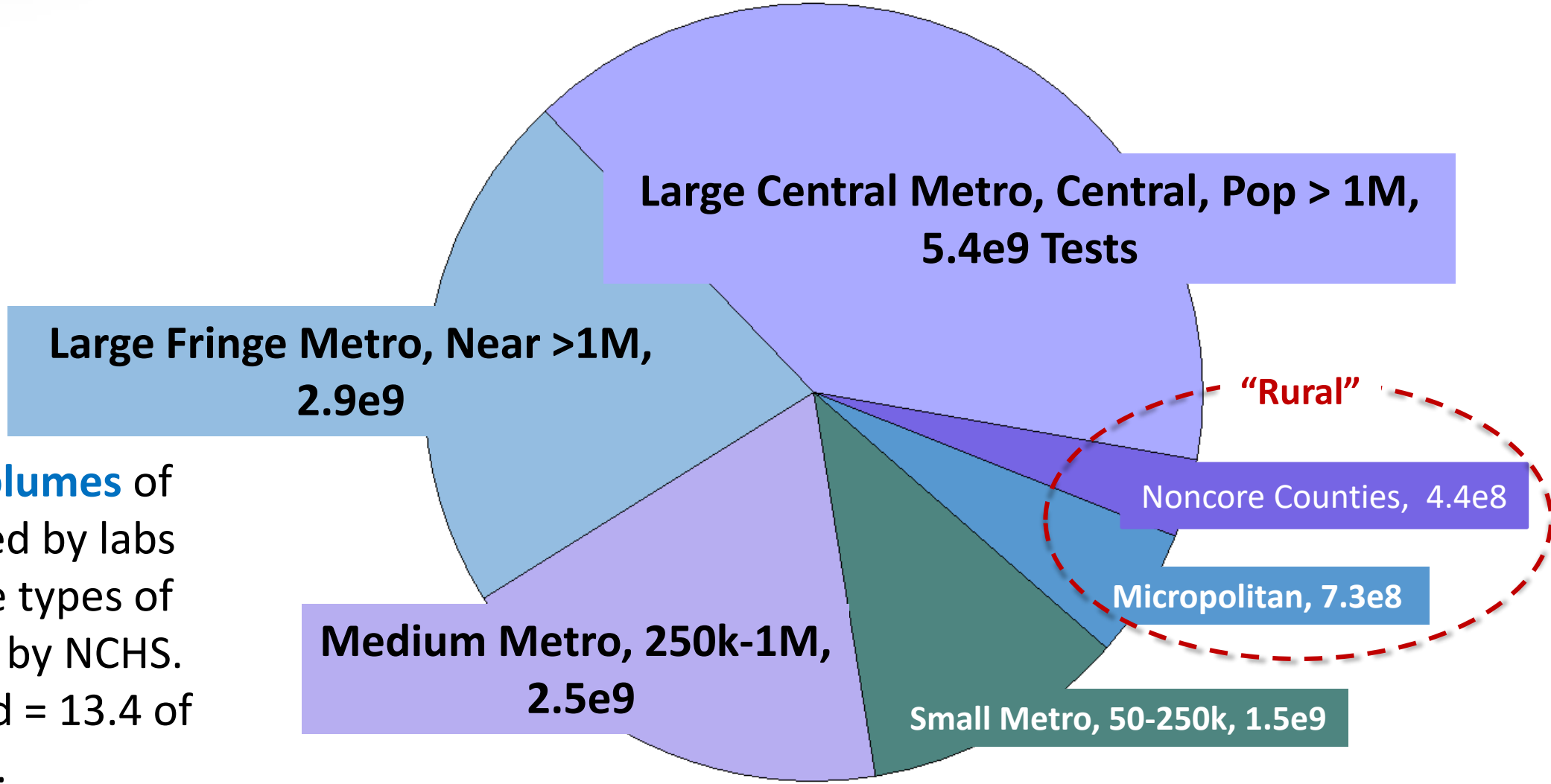


Number of Highest-volume Specialties



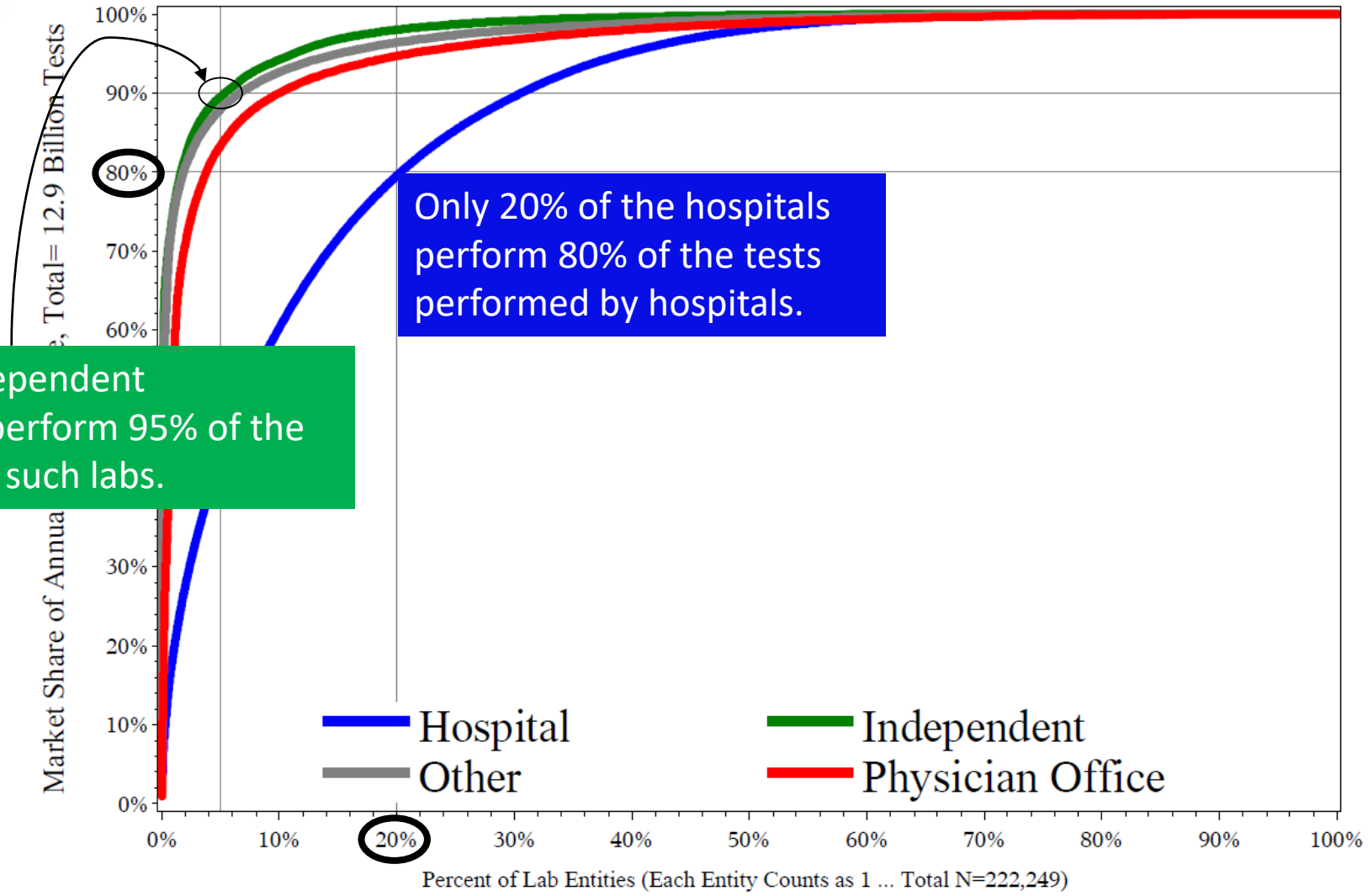
Numbers in pies are numbers of labs

Rural and Urban, National Center for Health Statistics (NCHS) Categories



These are the **volumes** of testing performed by labs certified in these types of areas as defined by NCHS. Total Categorized = 13.4 of 13.8 billion total.

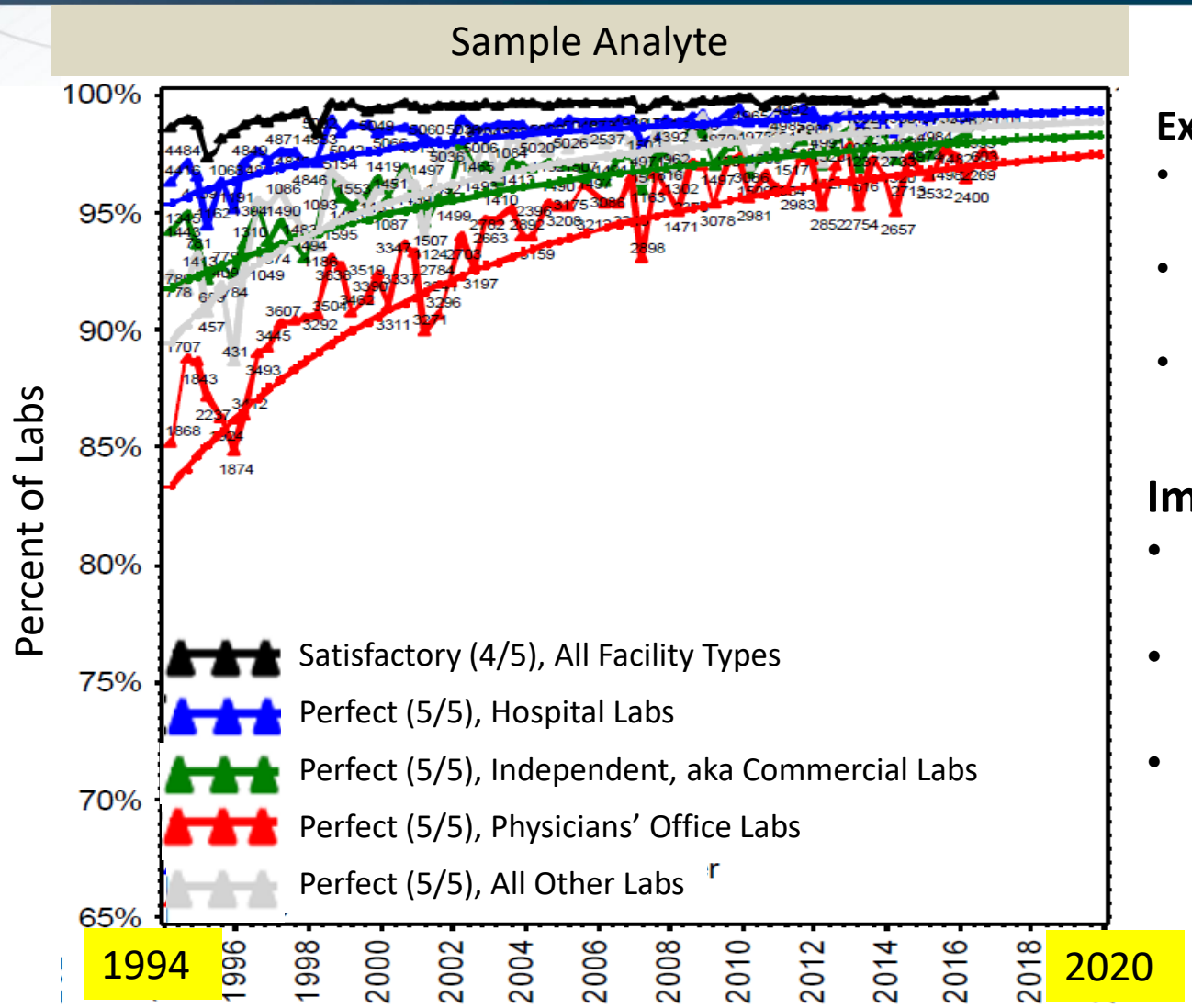
Market Distribution of Testing in U.S., 2016



Only 20% of the hospitals perform 80% of the tests performed by hospitals.

Only 5% of the independent (commercial) labs perform 95% of the tests performed by such labs.

Proficiency Testing (PT) Performance *Across Labs*



Expanded Legend

- Y-axis: Percent of Labs achieving either satisfactory (4/5=80) or perfect score (5/5=100)
- X-axis: years, 1994-2017 actual plus trend projection to 2020
- Years 2018-2020 projected *if curve-fit is appropriate*

Immediate Observations

- All facility types have improved during the CLIA period.
- Virtually all labs are satisfactory virtually all the time since about 2000 (black line).
- Hospitals have been consistently higher performing than other facility types.

NOTE: There are 86 analytes whose results are available this way.

Data Science Axioms

- A small percentage, e.g. <10%, of data that *is missing or corrupted* does not generally negate the analytical value of the remaining data.
- Fields with much greater “missingness” may still contain the best possible data set for the metric in question.
- Quality is field-by-field issue – one bad apple does not spoil the lot.

QIES Data Quality - Most of the data was present and valid most of the time. But not all.

- A total number of labs in the CLIA system, with or without various data fields in the QIES database, was 268,200 as of 12/31/17.
- The total number of labs with non-missing data for volumes and certification type was 262,600. Completeness was 98%
- The total annual test volume *for all labs reporting volume* (in any one of four possible fields) was 13.8 billion tests.
- The total test volume for labs whose reported county matched up correctly with a NCHS county name was 13.4 billion tests. 13.4/13.8=97%
- EXCEPTION: Staff figures turned out to be especially “missing”, with only 22% credible data.

Meeting Data Challenges – Cluster Analysis and Multivariable Regression Modeling, Imputation

Fact:

- 22% good data, from over 50,000 lab-samples, is still the best data we have about staff levels

Assumption:

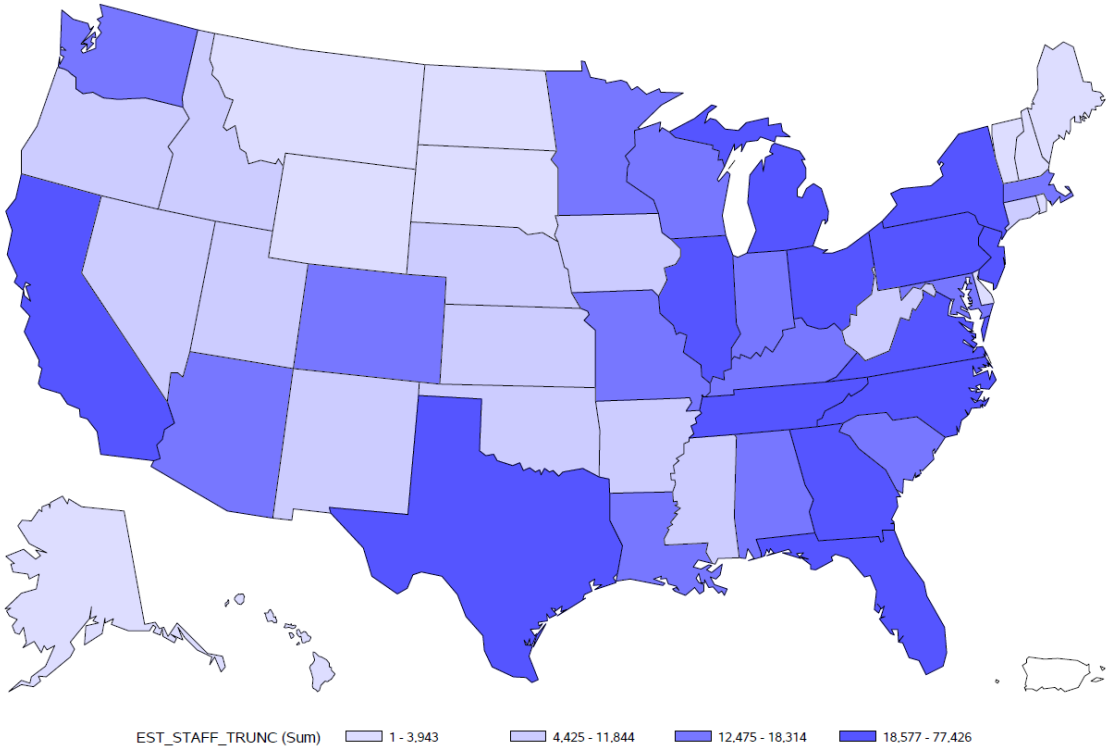
- *Tests per staff per shift* is a fundamental metric which is consistent within technologies and market segments

Approach:

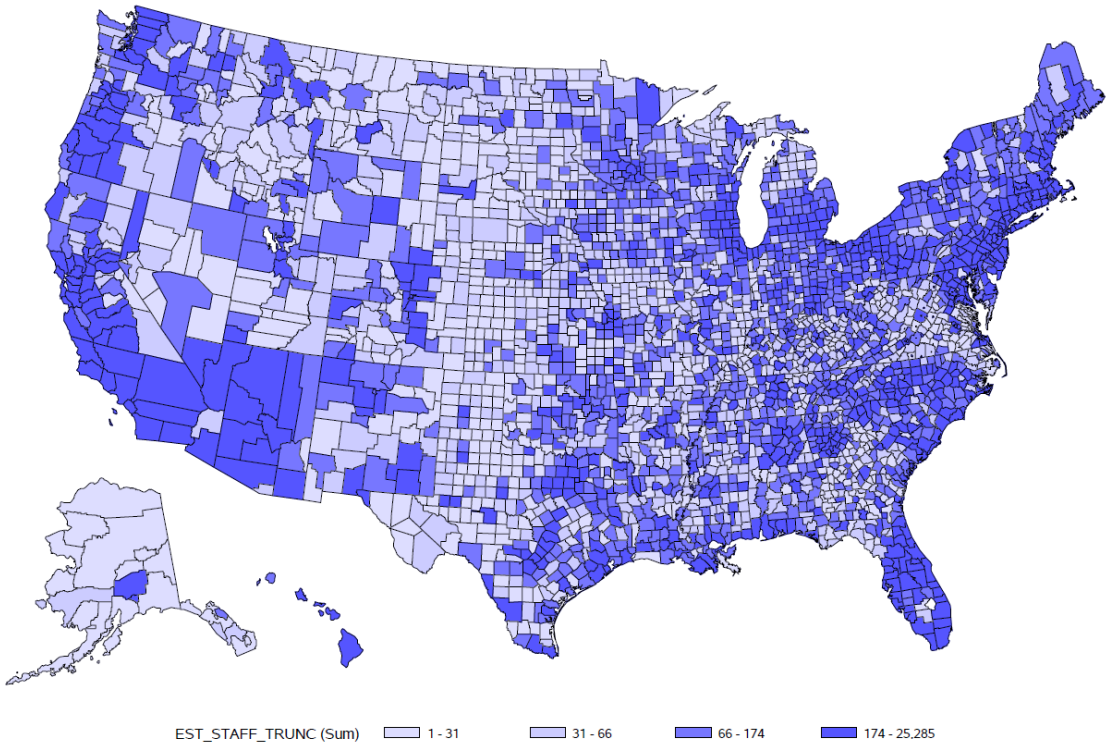
- Staff Level = *f*(technology) = *f*(proxies)
- Proxies are facility type, test volume, geography, rural/urban, specialties, et al
- QA, QA, QA!!

Workforce - Geographic Dispersion

Total Lab Staff in US States - Quartiles - ESTIMATED



Total Lab Staff in US Counties - Quartiles - ESTIMATED

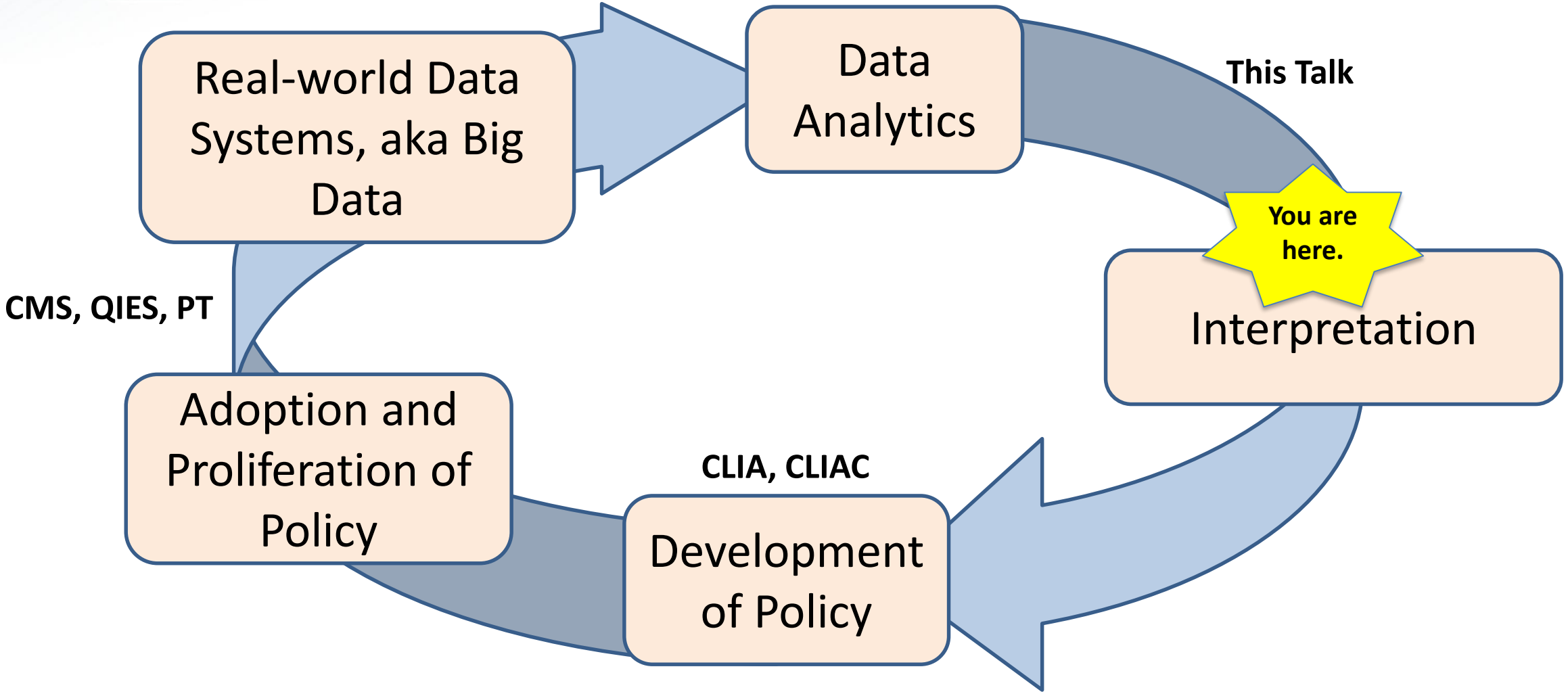


RECAP

Real-world Data

- Facility Types
- Test Volumes
- Geographic Dispersion
- Staff Levels
- Specialties
- Proficiency

Perspective on Data Analytics



Take-home Points

1. This presentation is relatively raw data; interpretation is needed.
2. The QIES database *is* Big Data.
3. Big Data presents surmountable challenges.
4. Meeting those challenges produces real-world evidence for decision making and policy development.

For more information, contact CDC
1-800-CDC-INFO (232-4636)
TTY: 1-888-232-6348 https://www.cdc.gov/

Images used in accordance with fair use terms under the federal copyright law, not for distribution.

Use of trade names is for identification only and does not imply endorsement by U.S. Centers for Disease Control and Prevention.

The findings and conclusions in this report are those of the authors and do not necessarily represent the official position of Centers for Disease Control and Prevention.