



United States Department of Agriculture

Transparent Quality Reporting in the Integration of Multiple Data Sources: Quality of Input Data

Mark Prell
Economic Research Service

FCSM Research Conference
March 7, 2018

*The views expressed are those of the author and should not be attributed to
the Economic Research Service or USDA*



Outline

- How an economist thinks about quality
- Workshop #1: Review of expert panel recommendations on data quality frameworks
- Workshop #1: Selected Findings (Input Data)
- Tradeoffs
- Moving forward



Outline

- **How an economist thinks about quality**
- Workshop #1: Review of expert panel recommendations on data quality frameworks
- Workshop #1: Selected Findings (Input Data)
- Tradeoffs
- Moving forward



Quality of Asparagus



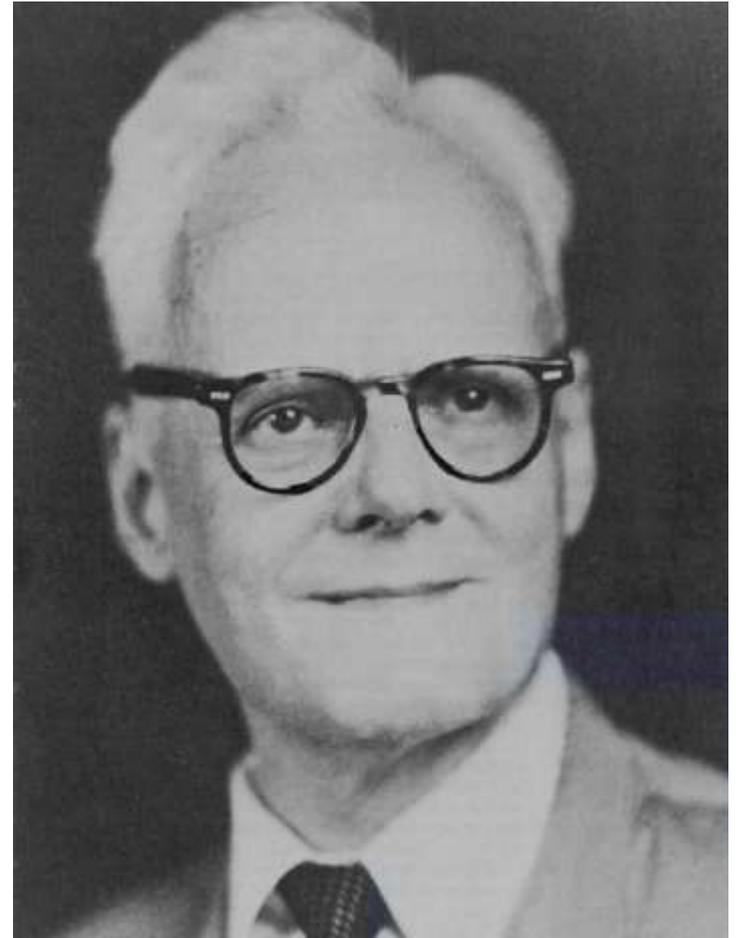
Frederick Waugh

QUALITY FACTORS INFLUENCING VEGETABLE
PRICES¹

FREDERICK V. WAUGH
MASSACHUSETTS DIVISION OF MARKETS

“Prices vary according to differences in the quality or the appearance of the Individual lots sold.”

Prices varied \$4.50 to \$12.00 per bushel box on single day.



3 “Dimensions” or “Characteristics” of Quality for Asparagus

- Color (greener is better)
- Size
- Uniformity



- Ex. of finding: price premium of 38.5 cents (per dozen bunches) per inch of green
- Color was the “most important factor”



Generalizable Lessons of Waugh study for Quality of Data

- Quality is multi-dimensional
- Consumers value different dimensions by different amounts



Dimensions cited in literature on data quality

Dimension	# cited	Dimension	# cited	Dimension	# cited
Accuracy	25	Format	4	Comparability	2
Reliability	22	Interpretability	4	Conciseness	2
Timeliness	19	Content	3	Freedom from bias	2
Relevance	16	Efficiency	3	Informativeness	2
Completeness	15	Importance	3	Level of detail	2
Currency	9	Sufficiency	3	Quantitativeness	2
Consistency	8	Usableness	3	Scope	2
Flexibility	5	Usefulness	3	Understandability	2
Precision	5	Clarity	2		

Source: Wand and Wang (1996)



Dimensions cited in literature on data quality

Dimension	# cited	Dimension	# cited	Dimension	# cited
Accuracy	25	Format	4	Comparability	2
Reliability	22	Interpretability	4	Conciseness	2
Timeliness	19	Content	3	Freedom from bias	2
Relevance	16	Efficiency	3	Informativeness	2
Completeness	15	Importance	3	Level of detail	2
Currency	9	Sufficiency	3	Quantitativeness	2
Consistency	8	Usableness	3	Scope	2
Flexibility	5	Usefulness	3	Understandability	2
Precision	5	Clarity	2		

Source: Wand and Wang (1996)



Dimensions cited in literature on data quality

Dimension	# cited	Dimension	# cited	Dimension	# cited
Accuracy	25	Format	4	Comparability	2
Reliability	22	Interpretability	4	Conciseness	2
Timeliness	19	Content	3	Freedom from bias	2
Relevance	16	Efficiency	3	Informativeness	2
Completeness	15	Importance	3	Level of detail	2
Currency	9	Sufficiency	3	Quantitativeness	2
Consistency	8	Usableness	3	Scope	2
Flexibility	5	Usefulness	3	Understandability	2
Precision	5	Clarity	2		

Source: Wand and Wang (1996)



Generalizable Lessons of Waugh study for Quality of Data

- Quality is multi-dimensional
- Consumers value several dimensions (although not all dimensions matter equally)
- ???



Outline

- How an economist thinks about quality
- **Workshop #1: Review of expert panel recommendations on data quality frameworks**
- Workshop #1: Selected Findings (Input Data)
- Tradeoffs
- Moving forward



Robert Groves

(Georgetown University)

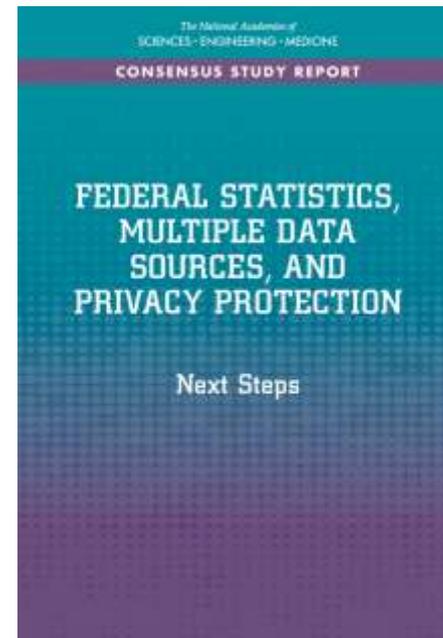
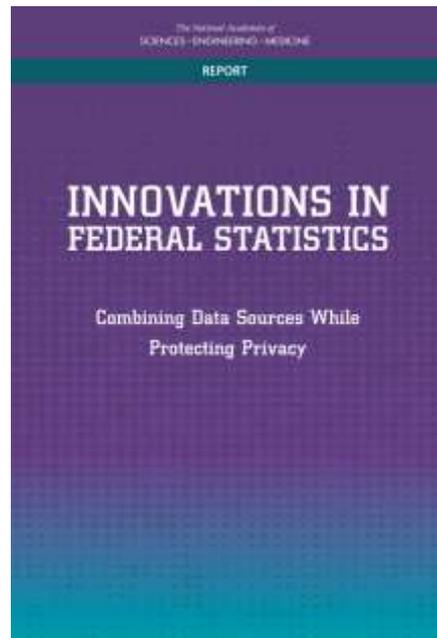
*“Advancing the Paradigm of Combining Data Sources –
Considerations from the National Academy of Sciences”*



Robert Groves

(Georgetown University)

“Advancing the Paradigm of Combining Data Sources – Considerations from the National Academy of Sciences”



Robert Groves

CNSTAT Panel Recommendations

- Recommendation 6-1.

Federal statistical agencies should adopt a

broader framework for statistical information **than**
total survey error . . .



Various definitions of quality

“Perhaps the most general and widely quoted is Juran and Gryna’s (1980) definition as simply ‘fitness for use’.”

Biemer and Lyberg,
Intro. to Survey Quality,
(2003)

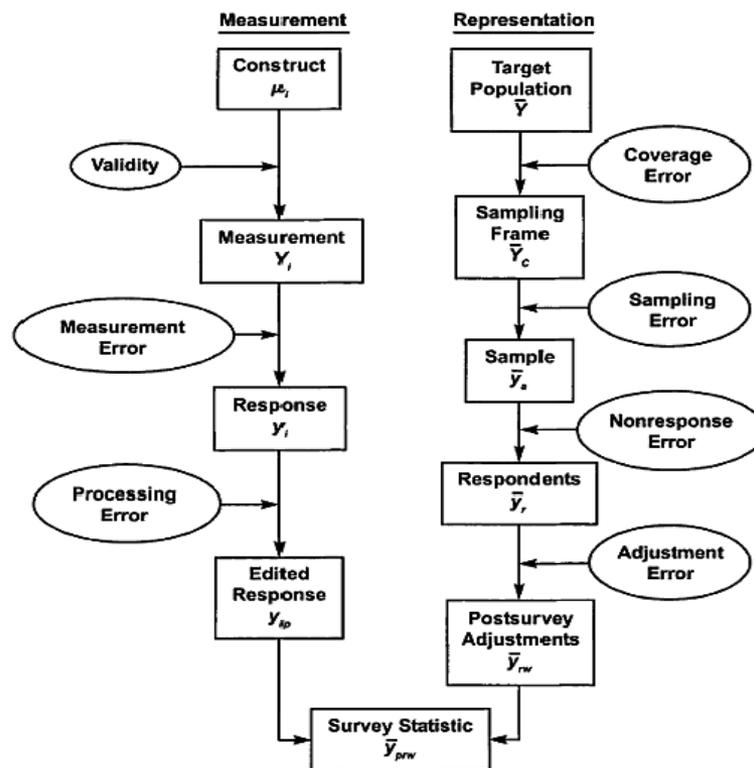


Total Survey Error (TSE)

(different versions)

TSE models portrays “types” or “sources” of errors.

Ultimately, a focus on “accuracy”



Source: Groves et al., Survey Methodology, 2nd ed. (2009)

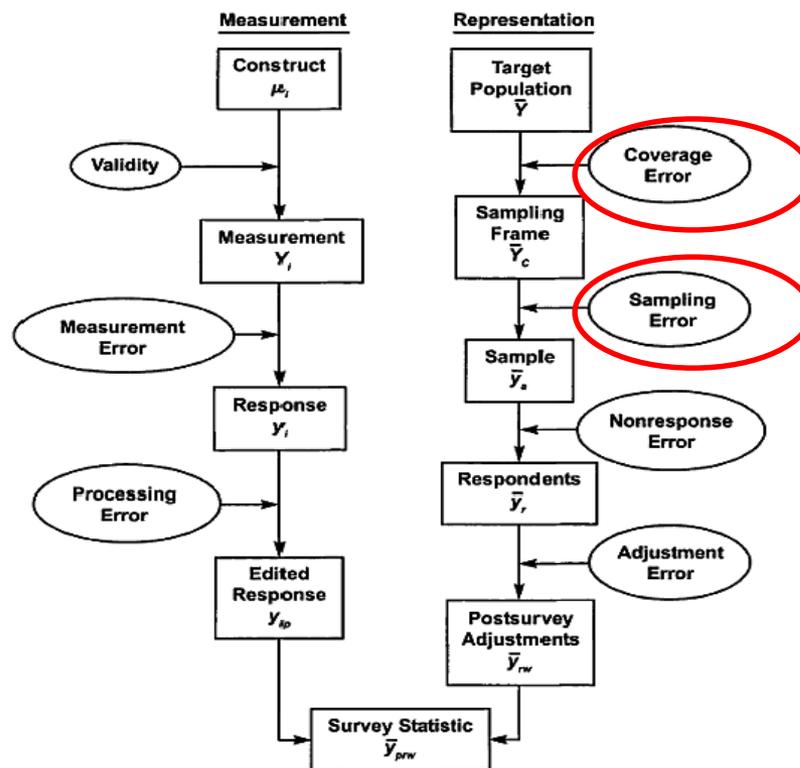


Total Survey Error (TSE)

(different versions)

TSE models portrays “types” or “sources” of errors.

Ultimately, a focus on “accuracy”



Source: Groves et al., Survey Methodology, 2nd ed. (2009)



Robert Groves

CNSTAT Panel Recommendations

- Recommendation 6-1.

Federal statistical agencies should adopt a broader framework for statistical information than total survey error to **include additional dimensions** that better capture user needs, such as **timeliness, relevance, accuracy, accessibility, coherence, integrity, privacy, transparency, and interpretability.**



Robert Groves

CNSTAT Panel Recommendations

- Recommendation 6-2

Federal statistical agencies should outline and evaluate the **strengths and weaknesses of alternative data sources** on the basis of a comprehensive quality framework, and, if possible, quantify the quality attributes and make them **transparent to users** . . .



Robert Groves

CNSTAT Panel Recommendations

- Recommendation 6-2 (continued)
 - ... **Agencies should focus more attention** on the **tradeoffs between different quality aspects**, such as, trading precision for timeliness and granularity, rather than focusing primarily on accuracy.



Outline

- How an economist thinks about quality
- Workshop #1: Review of expert panel recommendations on data quality frameworks
- **Workshop #1: Selected Findings (Input Data)**
- Tradeoffs
- Moving forward



Multiple Sources of Data

	Data Source	
	Government	Private-Sector
Structured	censuses probability surveys	academic surveys market research surveys
	administrative records	commercial transactions bank and credit card records medical records
	other: traffic sensors weather sensor water quality sensors	e-commerce mobile phone location GPS
Semi-structured	web-scraped quantitative data web logs	logs, web logs text messages and e-mail
Unstructured	satellite images traffic videos blogs and comments	Facebook pictures and videos Internet searches

- Source: Groves et al., *Innovations in Federal Statistics* (2017)



Multiple Sources of Data

	Data Source	
	Government	Private-Sector
Structured	censuses probability surveys	academic surveys market research surveys
	administrative records	commercial transactions bank and credit card records medical records
	other: traffic sensors weather sensor water quality sensors	e-commerce mobile phone location GPS
Semi-structured	web-scraped quantitative data web logs	logs, web logs text messages and e-mail
Unstructured	satellite images traffic videos blogs and comments	Facebook pictures and videos Internet searches

- Source: Groves et al., *Innovations in Federal Statistics* (2017)



Multiple Sources of Data

	Data Source	
	Government	Private-Sector
Structured	censuses probability surveys	academic surveys market research surveys
	administrative records	commercial transactions bank and credit card records medical records
	other: traffic sensors weather sensor water quality sensors	e-commerce mobile phone location GPS
Semi-structured	web-scraped quantitative data web logs	logs, web logs text messages and e-mail
Unstructured	satellite images traffic videos blogs and comments	Facebook pictures and videos Internet searches

- Source: Groves et al., *Innovations in Federal Statistics* (2017)



Multiple Sources of Data

	Data Source	
	Government	Private-Sector
Structured	censuses probability surveys	academic surveys market research surveys
	administrative records	commercial transactions bank and credit card records medical records
	other: traffic sensors weather sensor water quality sensors	e-commerce mobile phone location GPS
Semi-structured	web-scraped quantitative data web logs	logs, web logs text messages and e-mail
Unstructured	satellite images traffic videos blogs and comments	Facebook pictures and videos Internet searches

- Source: Groves et al., *Innovations in Federal Statistics* (2017)



Michael Berning and David Sheppard

(U.S. Census Bureau)

“Quality of Administrative Records as Source Data”

- Data Acquisition and Curation (DAC) manages over 150 interagency agreements to share administrative data and outside survey data sources
- DAC evaluates each source using a *Data Quality Assessment Tool for Administrative Data* (FCSM #46, 2013) with qualitative and quantitative dimensions.



Michael Berning and David Sheppard

(U.S. Census Bureau)

- Six dimensions: **relevance, accessibility, coherence, interpretability, accuracy, and institutional environment**
- Will source will be available and consistent over time?
- What are laws governing use of data? (**accessibility**)
- Do states have different methods to key data? (**coherence**)



Multiple Sources of Data

	Data Source	
	Government	Private-Sector
Structured	censuses probability surveys	academic surveys market research surveys
	administrative records	commercial transactions bank and credit card records medical records
	other: traffic sensors weather sensor water quality sensors	e-commerce mobile phone location GPS
Semi-structured	web-scraped quantitative data web logs	logs, web logs text messages and e-mail
Unstructured	satellite images traffic videos blogs and comments	Facebook pictures and videos Internet searches

- Source: Groves et al., *Innovations in Federal Statistics* (2017)



Bonnie Murphy and Crystal Konny (BLS)

“Quality Considerations for Administrative Data Used for the Producer Price Index (PPI) & Consumer Price Index (CPI)”

CPI and PPI use administrative (“alternative”) data:

- create sampling frames
- supplement, validate, and impute survey data
- reduce cost, save respondent burden
- improve dimensions of **accuracy** and **timeliness**



Bonnie Murphy and Crystal Konny (Alternative Data Matrix)



Bonnie Murphy and Crystal Konny

(Alternative Data Matrix)

Quality Metrics	Sample Frames	Benchmarking	Hedonics	Replace collection	Supplement Collection	Data Validation
Data Transparency- methods understood						
Granularity- Level of detail						
Quality of descriptive data						
Scope, type of price						
Coverage- items						
Coverage- geography						
Coverage- outlets						
Sampling procedures						
Data delivery reliable						
Viability of data source						
Data Usability						
Data Frequency						
Data Security						
Data delivery timeliness						
Data history						
Data Cleanliness						
Data Usability- mods to current system						



Bonnie Murphy and Crystal Konny

(Alternative Data Matrix)

Quality Metrics	Sample Frames	Benchmarking	Hedonics	Replace collection	Supplement Collection	Data Validation
Data Transparency- methods understood						
Granularity- Level of detail						
Quality of descriptive data						
Scope, type of price						
Coverage- items						
Coverage- geography						
Coverage- outlets						
Sampling procedures						
Data delivery reliable						
Viability of data source						
Data Usability						
Data Frequency						
Data Security						
Data delivery timeliness						
Data history						
Data Cleanliness						
Data Usability- mods to current system						



Bonnie Murphy and Crystal Konny

(Alternative Data Matrix)

Quality Metrics	Sample Frames	Benchmarking	Hedonics	Replace collection	Supplement Collection	Data Validation
Data Transparency- methods understood						
Granularity- Level of detail						
Quality of descriptive data						
Scope, type of price						
Coverage- items						
Coverage- geography						
Coverage- outlets						
Sampling procedures						
Data delivery reliable						
Viability of data source						
Data Usability						
Data Frequency						
Data Security						
Data delivery timeliness						
Data history						
Data Cleanliness						
Data Usability- mods to current system						



Bonnie Murphy and Crystal Konny

(Alternative Data Matrix)

Quality Metrics	Sample Frames	Benchmarking	Hedonics	Replace collection	Supplement Collection	Data Validation
Data Transparency- methods understood						
Granularity- Level of detail						
Quality of descriptive data						
Scope, type of price						
Coverage- items						
Coverage- geography						
Coverage- outlets						
Sampling procedures						
Data delivery reliable						
Viability of data source						
Data Usability						
Data Frequency						
Data Security						
Data delivery timeliness						
Data history						
Data Cleanliness						
Data Usability- mods to current system						



Mary Muth

(RTI International)

“Assessment of Commercial Store and Household Scanner Data: Methods, Content, and Cautions”

ERS has used retail food price data to:

- construct Quarterly Food-at-Home Price Database
- obtain consistent product descriptions and price characteristics of local markets for National Household Food Acquisition and Purchase Survey (FoodAPS)
- calculate cost of Thrifty Food Plan, basis for benefits in SNAP (food stamps).



Mary Muth

(RTI International)

Coherence and comparability issues. Differences in the data provided by stores on:

- store-brand products
- detail for individual stores (vs. aggregate sales for all locations)
- prices affected by coupons or loyalty card discounts
- random-weight products (e.g., produce)



Multiple Sources of Data

	Data Source	
	Government	Private-Sector
Structured	censuses probability surveys	academic surveys market research surveys
	administrative records	commercial transactions bank and credit card records medical records
	other: traffic sensors weather sensor water quality sensors	e-commerce mobile phone location GPS
Semi-structured	web-scraped quantitative data web logs	logs, web logs text messages and e-mail
Unstructured	satellite images traffic videos blogs and comments	Facebook pictures and videos Internet searches

- Source: Groves et al., *Innovations in Federal Statistics* (2017)



Roberto Rigobon

(MIT, National Bureau of Economic Research)

“Web-scraped Data: Consideration of Quality Issues for Federal Statistics”

- Using web-scraped data to:
 - study international pricing practices
 - produce alternative measures of inflation



Roberto Rigobon

(MIT, National Bureau of Economic Research)

- Web-scraping collects data from documents, images and descriptions on websites.
- Advantages: non-intrusive, automated (low cost)
- Disadvantages:
 - As websites change structure, need to re-program
 - Representativeness, sample selection (**accuracy**)
 - Reliable source in future (**comparability** over time)

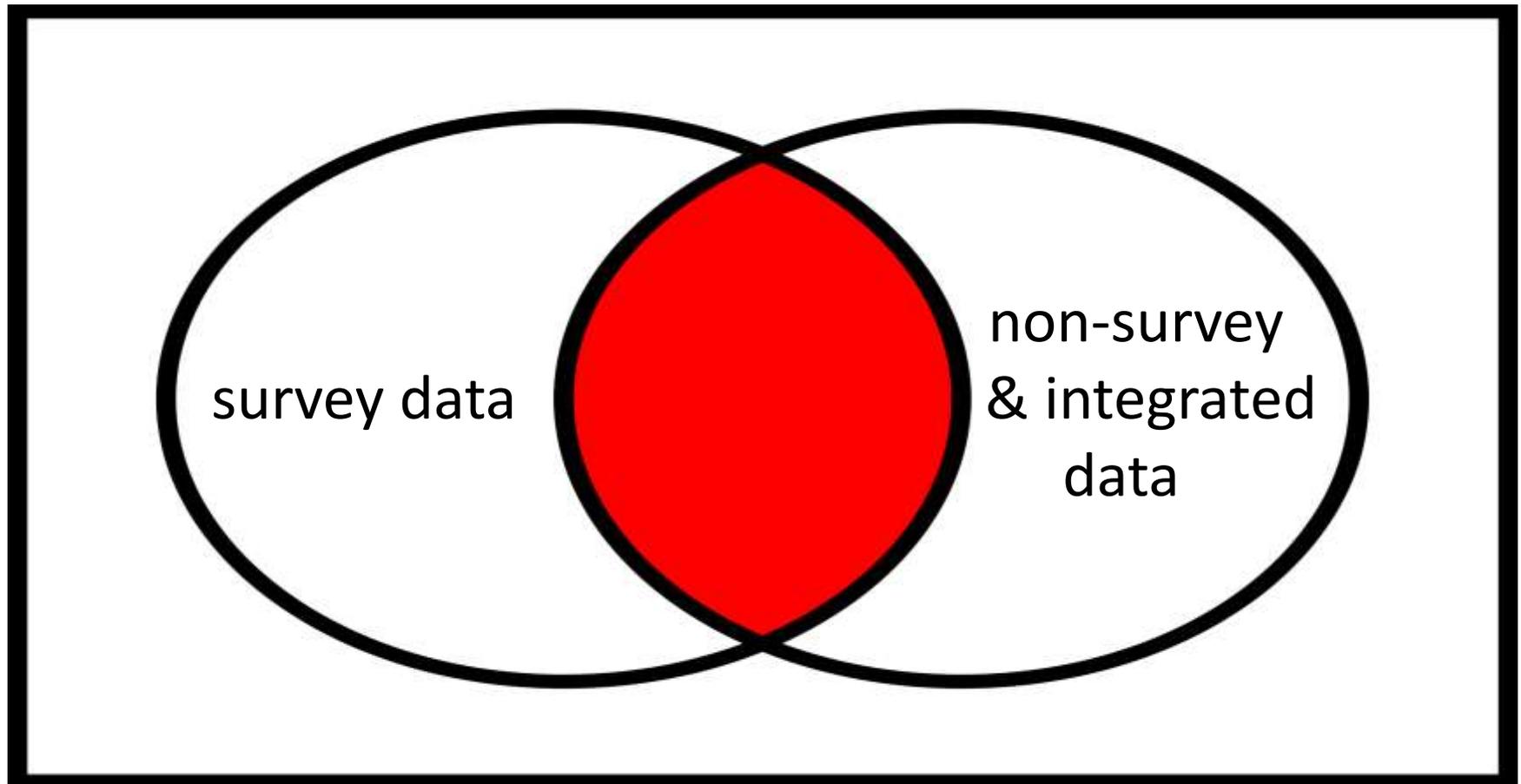


Lessons from presentations

- Data quality dimensions developed for **survey** data quality applicable to **non-survey** data (admin data, commercial transactions, web-scraped data)



Concepts, metrics of data quality **shared** by both types of data



Lessons from presentations

- Data quality dimensions developed for **survey** data quality applicable to **non-survey** data (admin data, commercial transactions, web-scraped data)
- Federal statistical agencies already consider dimensions other than accuracy for non-survey data (although the “language” of dimensions may be rare)

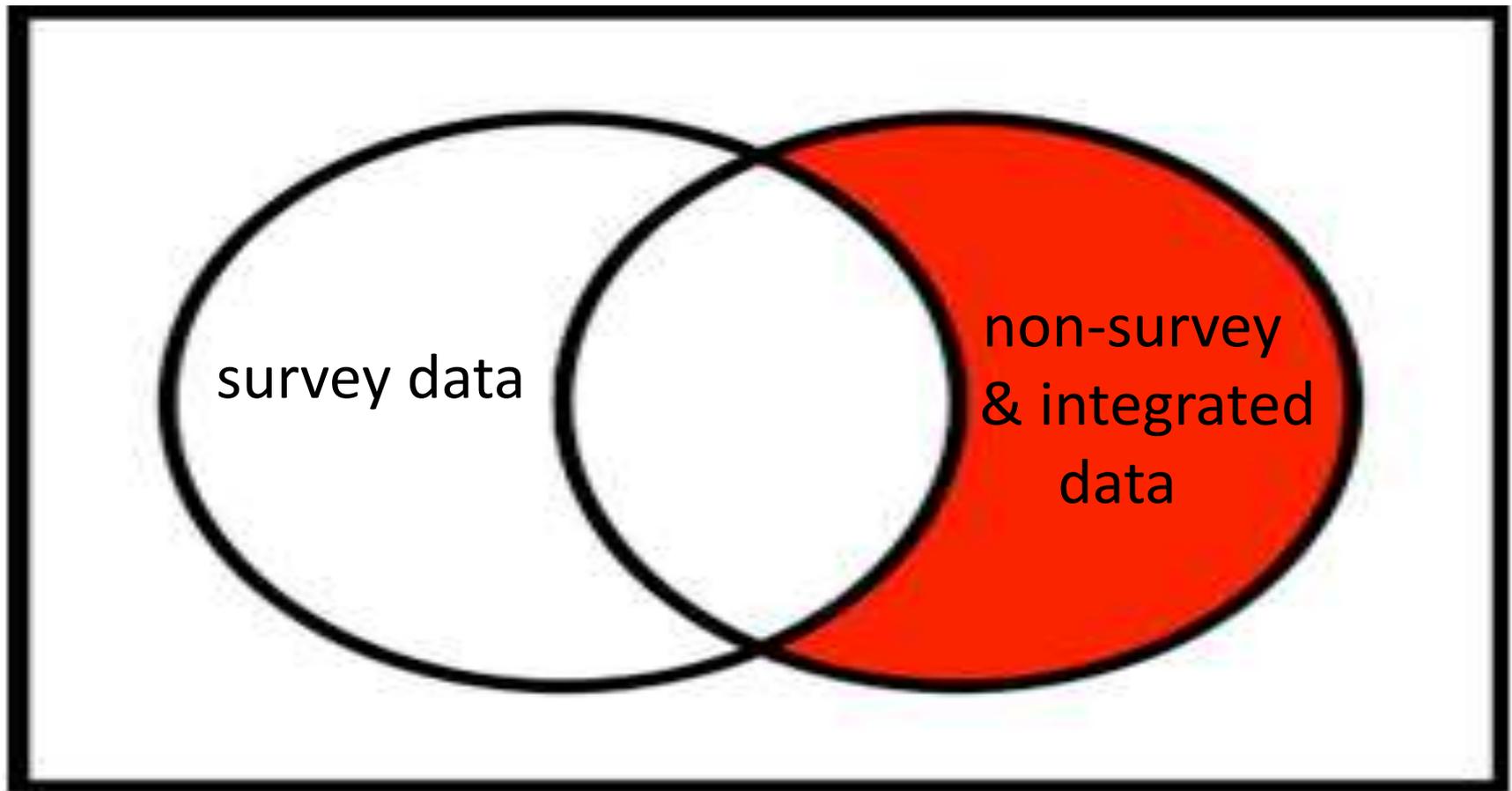


Lessons from presentations

- Data quality dimensions developed for **survey** data quality applicable to **non-survey** data (admin data, commercial transactions, web-scraped data)
- Federal statistical agencies already consider dimensions other than accuracy for non-survey data (although the “language” of dimensions may be rare)
- Possible that some new concepts of data quality might be needed to describe non-survey and integrated data (especially unstructured data)



Concepts, metrics of data quality **unique** to non-survey & integrated data



Outline

- How an economist thinks about quality
- Workshop #1: Review of expert panel recommendations on data quality frameworks
- Workshop #1: Selected Findings (Input Data)
- **Tradeoffs**
- Moving forward



CNSTAT Panel Recommendation

- Recommendation 6-2 (continued)
 - ... **Agencies should focus more attention** on the **tradeoffs between different quality aspects**, such as, trading precision for timeliness and granularity, rather than focusing primarily on accuracy.



CNSTAT Panel Recommendation

- Recommendation 6-2 (continued)
 - ... **Agencies should focus more attention** on the **tradeoffs between different quality aspects**, such as, trading **precision** for timeliness and **granularity**, rather than focusing primarily on accuracy.



CNSTAT Panel Recommendation

- Recommendation 6-2 (continued)
 - ... **Agencies should focus more attention** on the **tradeoffs between different quality aspects**, such as, trading **precision** for timeliness and **granularity**, rather than focusing primarily on accuracy.
- Agency decisions on tradeoffs are difficult, in part, because there are **multiple users with DIFFERENT valuations** of those tradeoffs.



Different users have different valuations of dimensions

- “[T]he importance of different characteristics varies among users”
Biemer and Lyberg, *Intro. to Survey Quality*, (2003)
- Implication: Reporting is key.
 - A statistical agency can report on precision and granularity so that **each user can make his or her own assessment** of quality (fitness-for-use for that particular user’s needs)

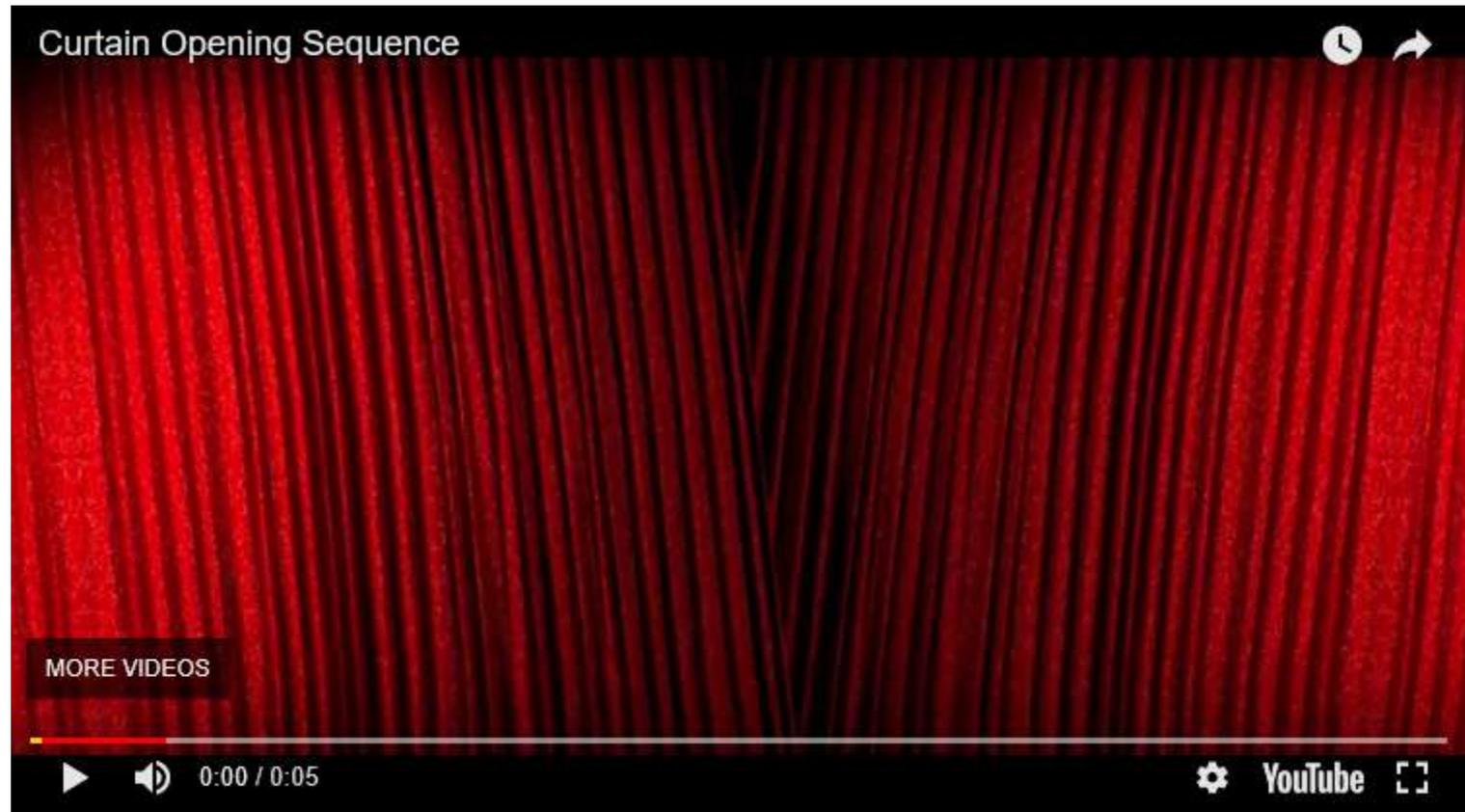


Generalizable Lessons of Waugh study for Quality of Data

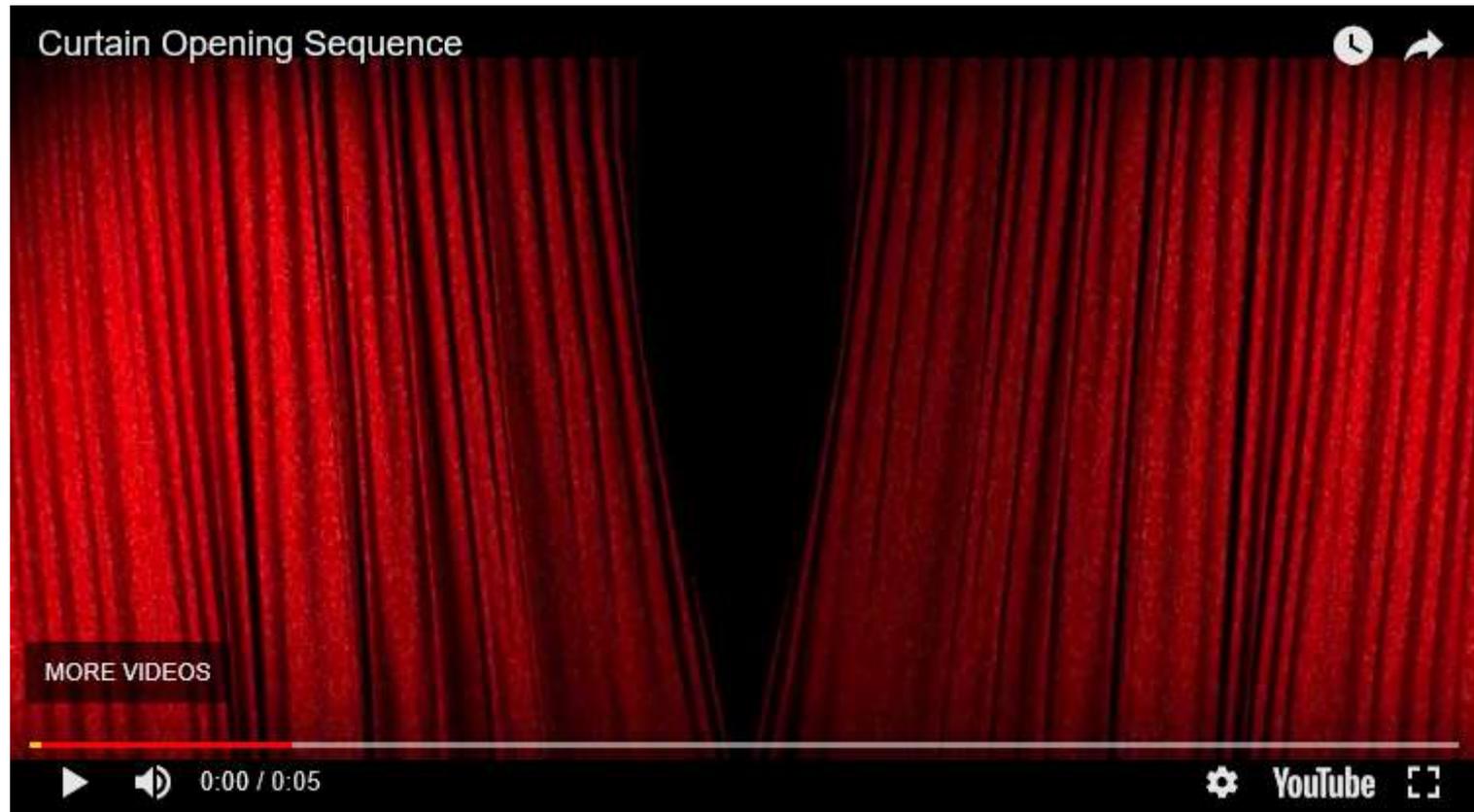
- Quality is multi-dimensional
- Consumers value several dimensions (although not all dimensions matter equally)
- ???



The Last Lesson from Waugh



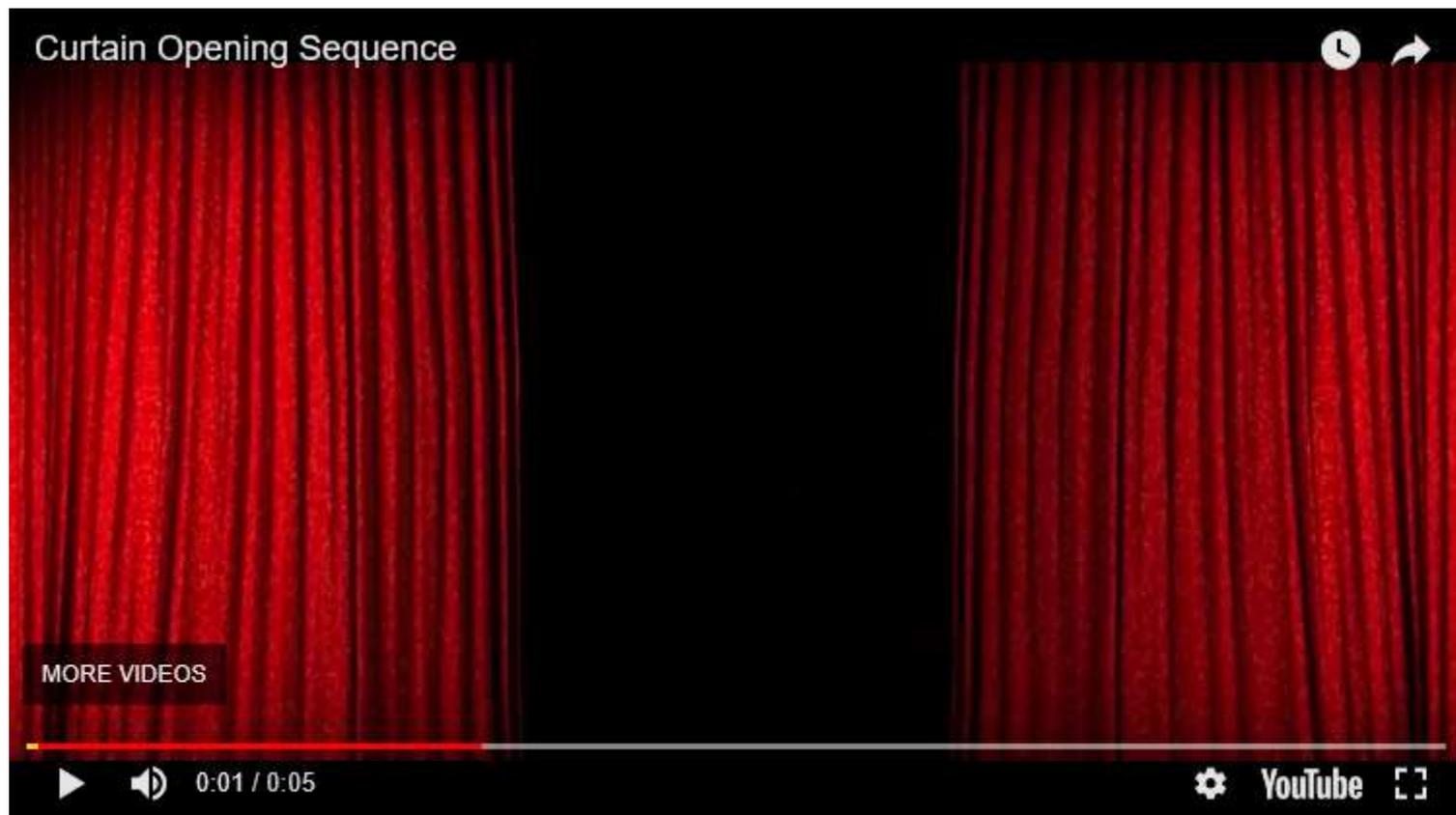
The Last Lesson from Waugh



Source: Miklos (2011)



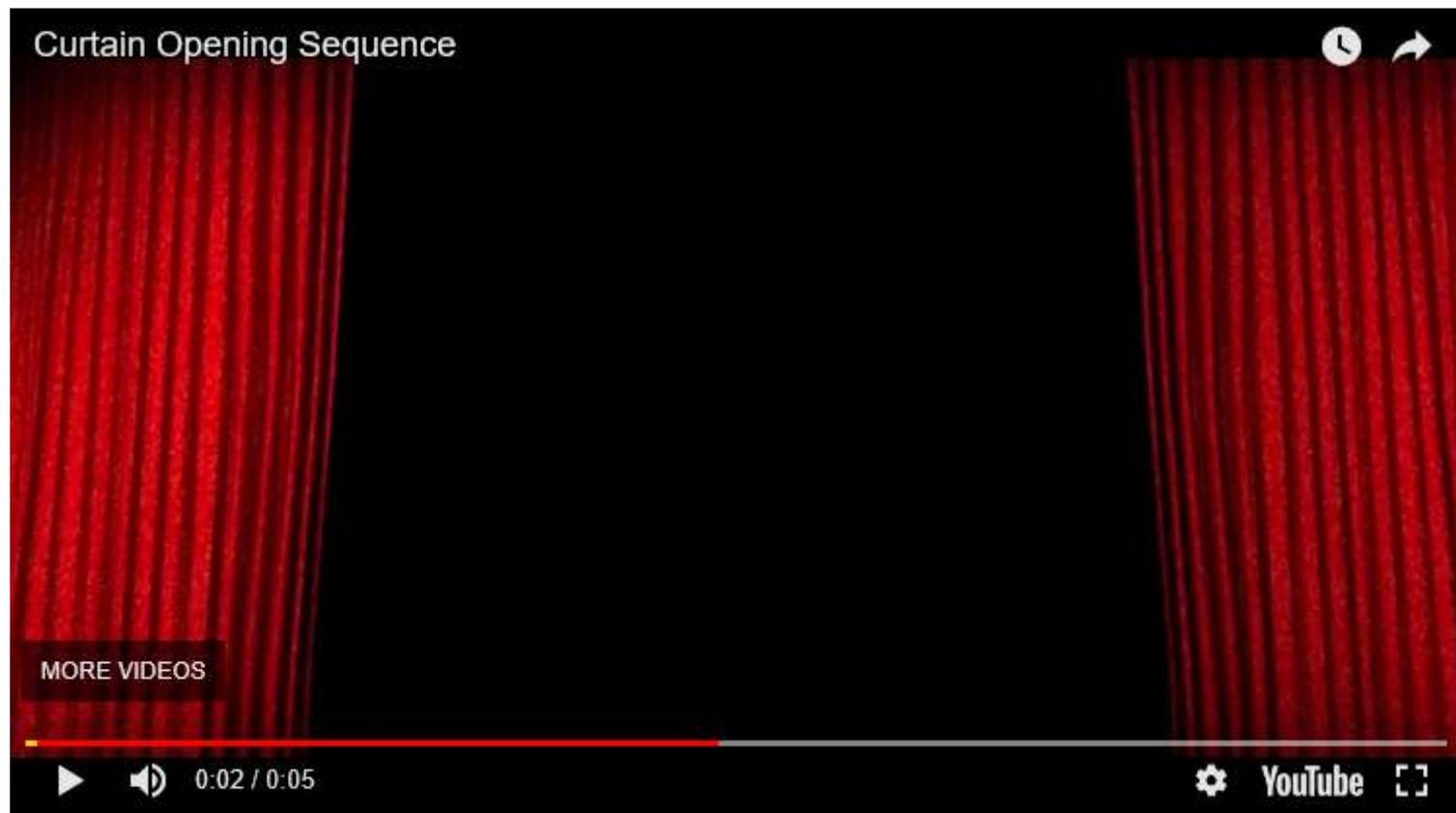
The Last Lesson from Waugh



Source: Miklos (2011)



The Last Lesson from Waugh



Source: Miklos (2011)



The Last Lesson from Waugh



Source: Miklos (2011)



The Last Lesson from Waugh



Source: Miklos (2011)



Generalizable Lessons of Waugh study for Quality of Data

- Quality is multi-dimensional
- Consumers value several dimensions (although not all dimensions matter equally)
- **Farmers face quality tradeoffs in production—and so do statistical agencies**



Economic view of quality tradeoffs: TWO perspectives

- (1) Tradeoff from a **user's perspective (fitness-for-use)**

“I value some granularity, and I am willing to give up some precision to get it”

- (2) Tradeoff from **agency perspective (technically and financially feasible)**

“We can provide more granularity, but with fixed budget we have to reduce precision”



Outline

- How an economist thinks about quality
- Workshop #1: Review of expert panel recommendations on data quality frameworks
- Workshop #1: Selected Findings (Input Data)
- Tradeoffs
- **Moving forward**



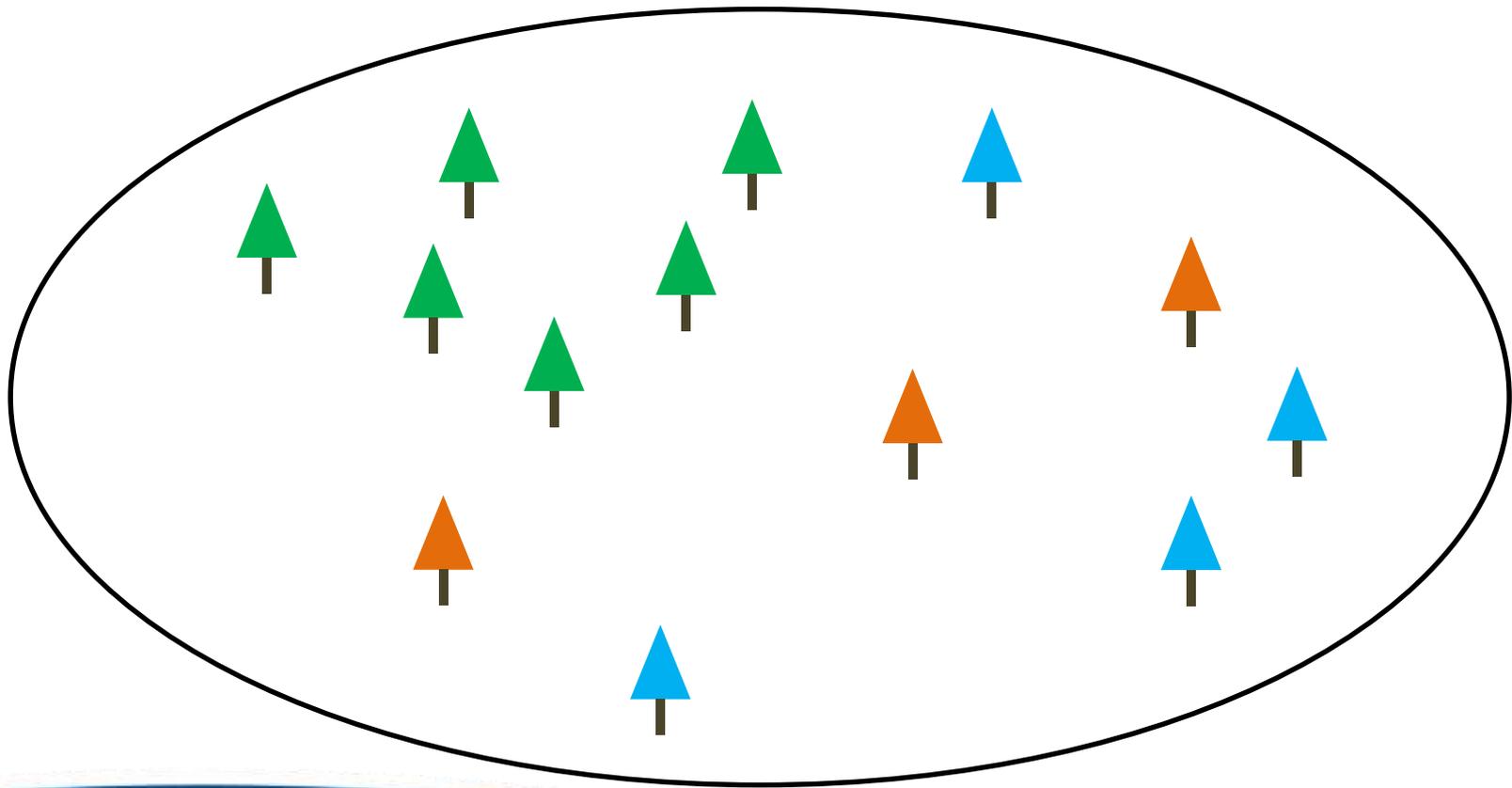
Two possible implications of transparent reporting

1. Re-organize extant information and metrics, by dimension (translation to new terminology)
 - Example: when reporting on how data were collected differently from different units, agency can describe that as “coherence” issue
 - relatively smaller burden



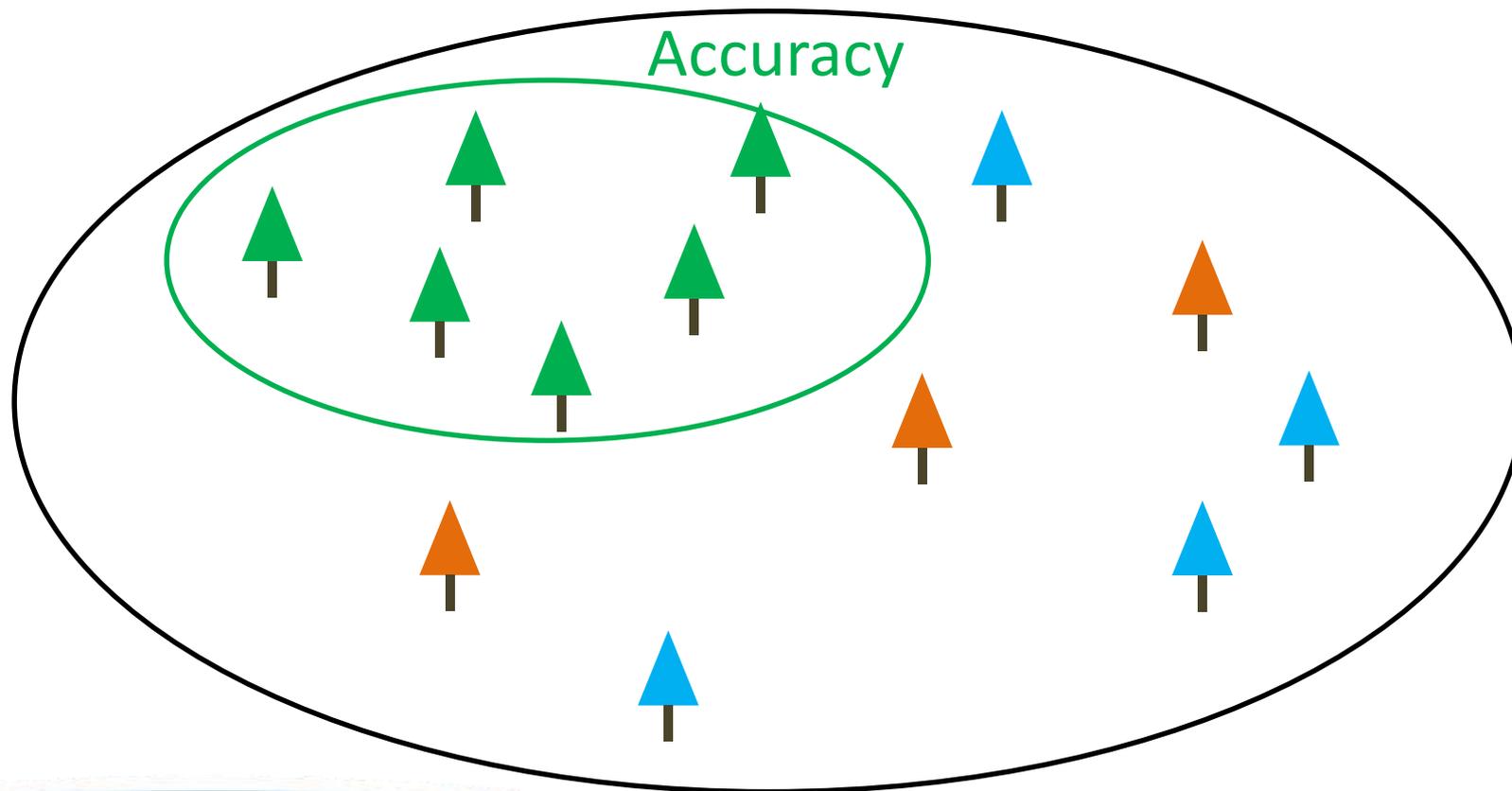
Trees-Groves-Forest of Information Reported on Data Quality

Current Forest



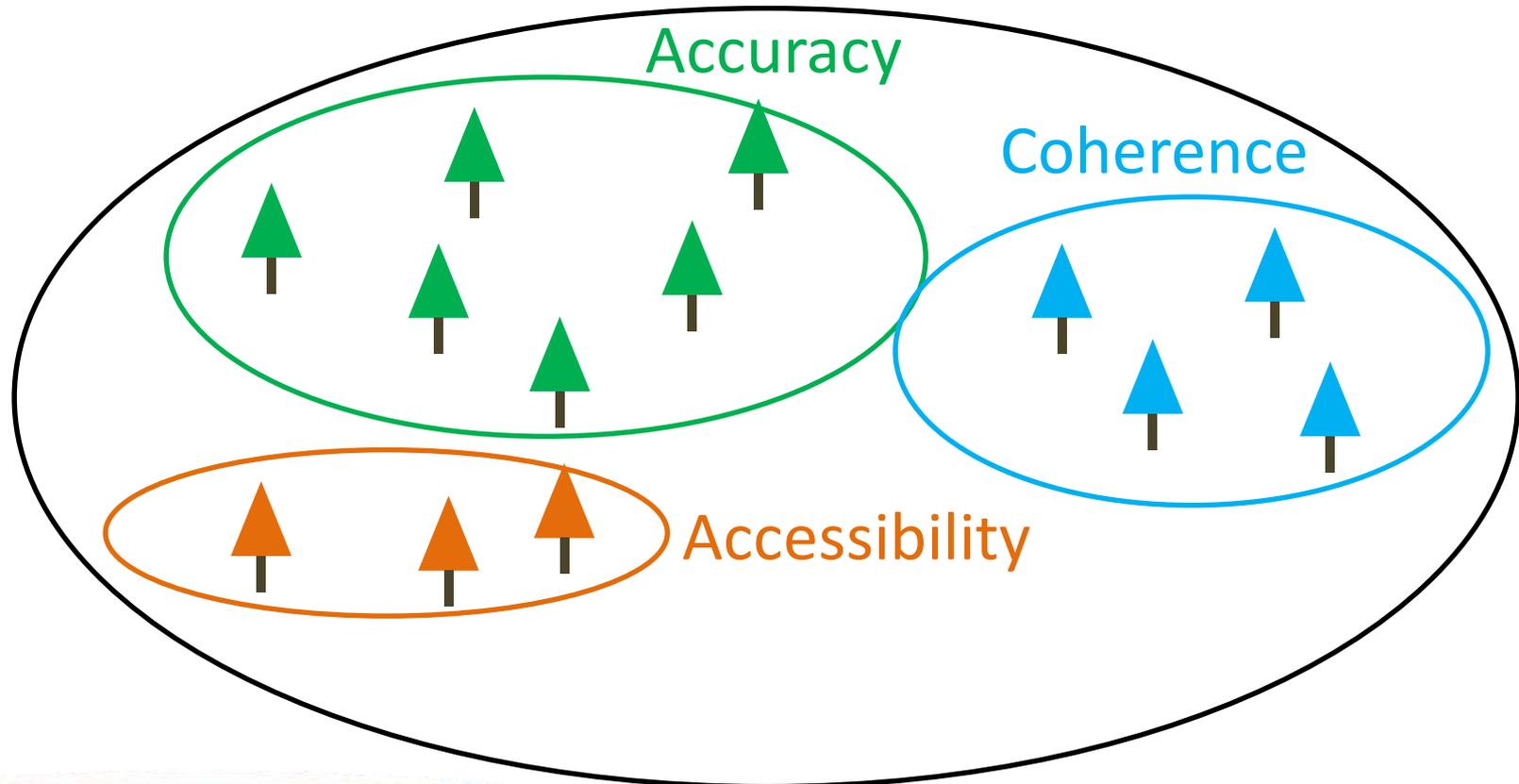
Trees-Groves-Forest of Information Reported on Data Quality

Current Forest



Trees-Groves-Forest of Information Reported on Data Quality

Re-organized Forest



Multiple dimensions in use

Quality report for ESS Labor
Force Survey 2015 (2017)

Ch 3. Relevance

Ch 4. Accuracy

Ch 5. Timeliness

Ch 6. Accessibility and
Clarity

Ch 7. Comparability

Ch 8. Coherence



Two possible implications of transparent reporting

2. Report additional detail (metrics, narratives)

- Example: (Berning and Sheppard)

- ✓ Degree of missingness

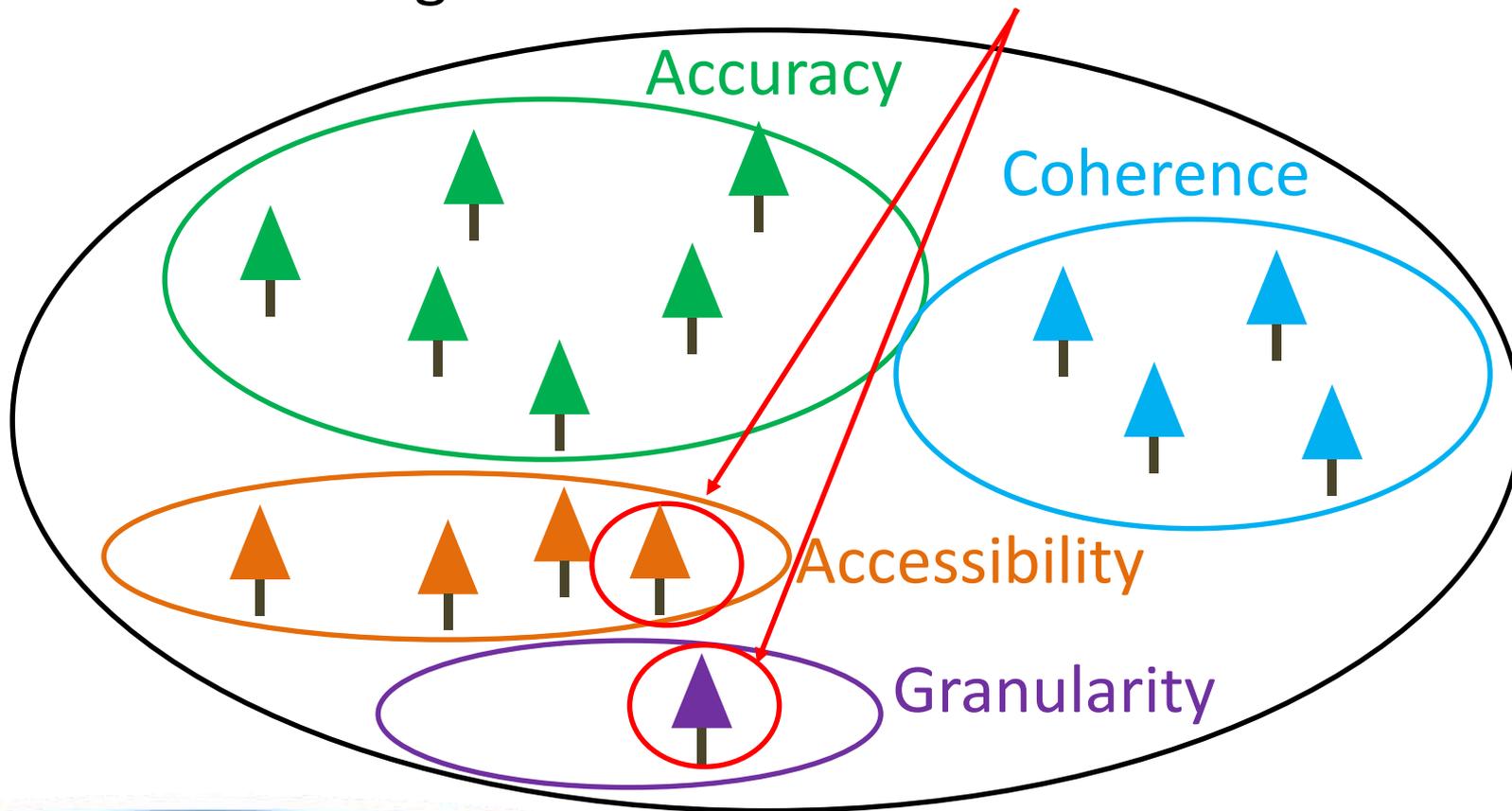
- ✓ Frequency of invalid values

- Example: documentation adds to “accessibility”



Trees-Groves-Forest of Information Reported on Data Quality

Re-organized Forest **with New Trees**



Two possible implications of transparent reporting

2. Report additional detail (metrics, narratives)

- Examples (Berning and Sheppard)

 - ✓ Degree of missingness

 - ✓ Frequency of invalid value

- **relatively larger burden**



Two possible implications of transparent reporting

2. Report additional detail (metrics, narratives)

- Examples (Berning and Sheppard)

 - ✓ Degree of missingness

 - ✓ Frequency of invalid value

- **relatively larger burden**

- **relatively larger opportunity to improve agency-user communication**



3 Take-away lessons

1. **Workshop.** Data quality concepts developed for survey data quality can be successfully applied to non-survey data; some unique concepts may be needed for non-survey and integrated data.
2. **Quality tradeoffs.** Two perspectives: user and agency (technical tradeoff, given budget)
3. **Moving forward.** To extent agencies already report on dimensions other than accuracy, fuller reporting has smaller implications for agencies; to extent need to increase details, larger implications.



Having whetted your appetite for quality...



On to presentations about
processing and output...

Thank you!

mprell@ers.usda.gov

(And thanks to colleagues from the Joint Program in Survey Methodology who assisted the FCSM working group to bring the highlights of the workshop to you: Katharine Abraham, Frauke Kreuter, Alexandra Brown, Andrew Caporaso)

