Leveraging the Data Quality Framework Workshop

September 10, 2021

A Virtual Workshop Sponsored by the Interagency Council on Statistical Policy, Chief Data Officer Council, and Evaluation Officer Council

Hubert Hamer, USDA NASS

Opening Remarks

Mentimeter

- Go to https://www.menti.com/xau2spxsrs
- Go to menti.com and enter the code: **4081 7831**
- Scan the QR code with your mobile device

Federal Committee on
Statistical Methodology

The FCSM Framework for Data Quality

Jennifer D. Parker, Ph.D.
National Center for Health Statistics

Leveraging the Data Quality Framework
September 10, 2021

Background

- Understanding data quality is essential for data-driven decision making
  - Data users who understand the “fitness-for-use” of data products are more likely to use them appropriately
  - Higher-impact uses of data require higher quality data
- All data have strengths and weaknesses
- Data quality for surveys is relatively well-established but data quality for integrated data and other non-statistical data are less developed

Data Quality Milestones 2001-2020 (A)

WP #31

- 2001 Measuring and Reporting Sources of Error in Surveys
- Focus on reporting accuracy of survey data outputs

IQ Act

- 2001 Information Quality Act/OMB Guidelines
- Provided a framework, with a call for more detailed OMB and Agency Guidelines

Standards

- 2006 OMB Statistical Policy Directive 2: Standards and Guidelines for Statistical Surveys
- Emphasis on survey data quality

Data Quality Milestones 2001-2020 (B)

Integrated Data

- System-wide declining response rates, increasing costs
- Increased use of non survey data sources, alone or integrated with statistical survey data

Evaluations

- 2015-17 Two CNSTAT reports and the Commission on Evidence-Based Policymaking, integrated data
- New visions for Federal statistics; identified obstacles and provided recommendations for moving forward

2018 Evidence Act

- Federal Data Strategy, Foundations for Evidence Based Policymaking Act, revised OMB Information Quality Act Guidelines
- Address data quality and compatibility with integrated data

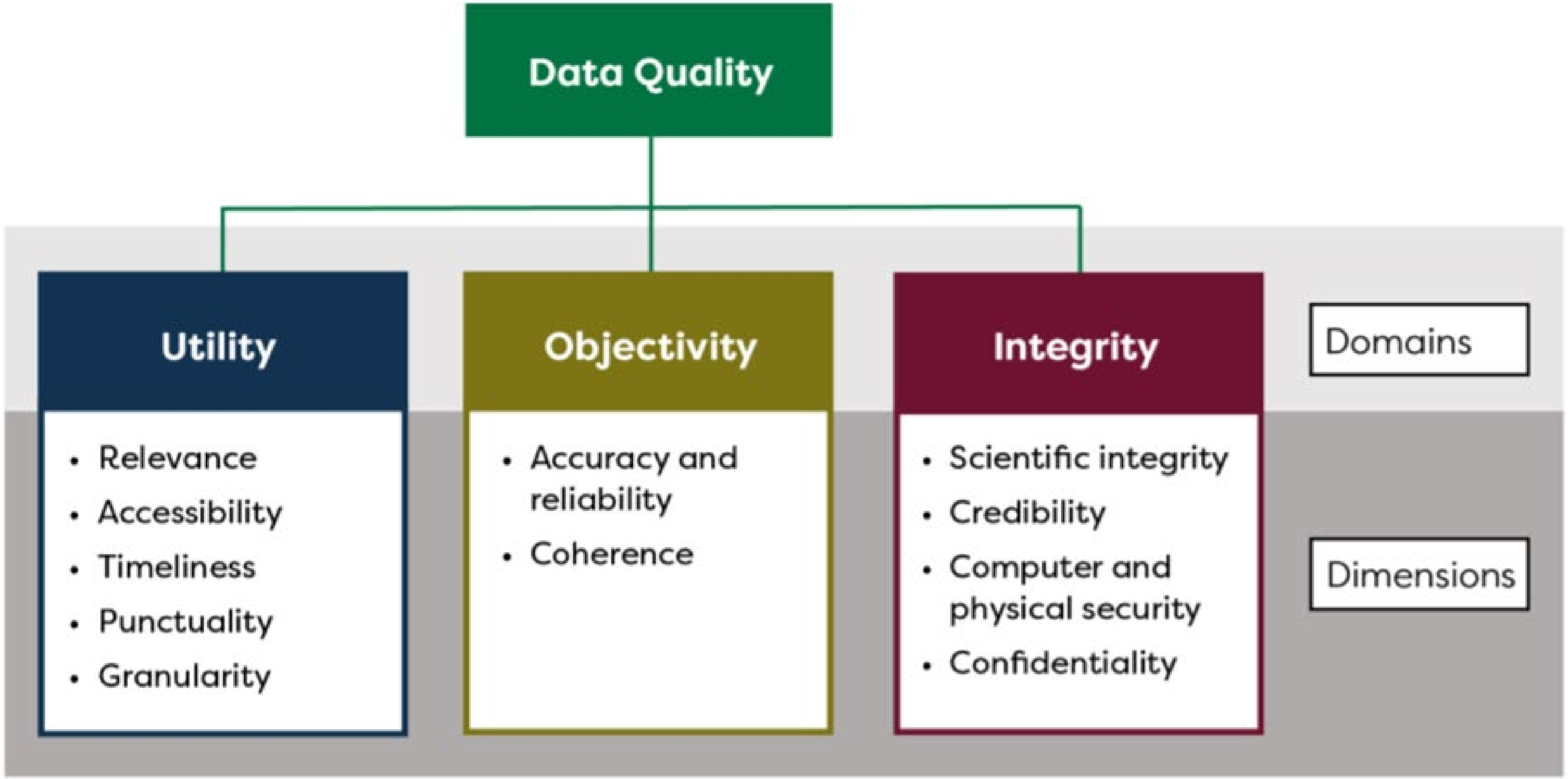
Quality Framework

- 2018-2019 seminars on data quality and integrated data
- FCSM Data Quality Framework

FCSM Framework for Data Quality

- Builds on experience of the Federal Statistical System
- Organizes the elements of data quality around the structure of the Information Quality Act
- Explains for a broad audience the importance of understanding data quality to determine fitness for purpose, identifying and mitigating key data quality threats, and evaluating trade-offs
- Provides strategies for documenting and reporting data quality

FCSM Framework for Data Quality



Data Quality

Utility
- Relevance
- Accessibility
- Timeliness
- Punctuality
- Granularity

Objectivity
- Accuracy and reliability
- Coherence

Integrity
- Scientific integrity
- Credibility
- Computer and physical security
- Confidentiality

Domains

Dimensions

Federal Committee on Statistical Methodology

Domains of Data Quality

- Utility - the extent to which information is well-targeted to valuable needs: it reflects the usefulness of the information to the intended users
- Objectivity - whether information is accurate, reliable, and unbiased, and is presented in an accurate, clear and interpretable, and unbiased manner
- Integrity – the maintenance of rigorous scientific standards and the protection of information from manipulation or influence as well as unauthorized access or revision

Dimensions of Utility - I

- Relevance: whether the data product is targeted to meet current and prospective user needs
- Credibility: the confidence that users place in data products based simply on the image of the data producer
- Accessibility: the ease with which data users can obtain an agency’s products and documentation in forms and formats that are understandable to data users.

Dimensions of Utility - II

- Timeliness: the length of time between the event or phenomenon the data describe and their availability
- Punctuality: the time lag between the actual release of the data and the planned target date for data release
- Granularity: the amount of disaggregation available for key data elements.

Dimensions of Objectivity

- Accuracy: the closeness of an estimate from a data product to its true value
  - Reliability: characterization of repeated estimates of accuracy over time
- Coherence: the ability of the data product to maintain common definitions, classification, and methodological processes, to align with external statistical standards, and to maintain consistency and comparability with other relevant data

Dimensions of Integrity

- Scientific Integrity: an environment that ensures the use of established scientific methods to produce and disseminate objective data products and shields these products from inappropriate political influence
- Computer and Physical Security: the protection of information throughout the collection, production, analysis, and development process from unauthorized access or revision to ensure that the information is not compromised through corruption or falsification
- Confidentiality: a quality or condition of information as an obligation not to disclose that information to an unauthorized party

Threats to Data Quality

- Threats can be identified for all dimensions
  - Threats can be relevant for multiple dimensions
  - Mitigating threats for one dimension can increase threats for another
- Managing trade-offs among quality dimensions is important
- Threats to quality for blended data combine threats for data inputs, blending methods, and data outputs

Assessing Data Quality

- Regularly identify threats to data quality for ongoing data collections, including when considering new source data for inclusion
  - Decisions on trade-offs among threats and mitigation measures should be considered in the context of the data’s purpose and all identified threats
  - Data quality for the intended use may differ from that for its original purpose

Conclusion

- Data quality has been long studied for statistical data, especially surveys, but is less developed for integrated and secondary-use data
- The FCSM Data Quality Framework can be used to evaluate quality for all data

Federal Committee on Statistical Methodology

Using the Framework for Data Quality

Rolf R. Schmitt, PhD
Bureau of Transportation Statistics

September 2021

The Framework for Data Quality

- Organizes the many elements of data quality around the structure of the Information Quality Act
- Provides a comprehensive and consistent terminology to describe the many aspects of data quality
- Looks overwhelming to use and burdensome to report

Don’t panic

- Many data quality threats can be dismissed after brief consideration for a data program
- There are few universal rules for weighing importance of one data quality concern over another: tradeoffs are expected
- Documentation while planning and doing what you do is a good habit that helps your successors and supports transparency

Reporting data quality

- Data quality reports as a byproduct of documenting your work
- Applies to managers of data collection programs and to analysts
- Three audiences
  - The data program manager / analyst
  - The power user
  - The occasional user or decisionmaker

Reporting data quality

- The cultural change for program managers and analysts: consider all threats and note how you address each relevant threat to inform your successor
- The manager’s notes provide a cornerstone for technical documentation for power users
- The elevator speech: describe in a few words how likely the data will misguide a decision

Tradeoffs change over time

- Covid-19 put a premium on timeliness over deliberative vetting of accuracy
- “It may be better, in the gross affairs of life, to be less precise and more prompt. Quick decisions, though they may contain a grain of error, are often better than precise decisions at the expense of time.”
  - T.C. Chamberlin, President of the University of Wisconsin, **1890**

Future work

- Additional tools to measure quality in blended data sets
- Best practices for identifying quality of data obtained from sources that lack transparency and from advanced (AI) algorithms
- Tools for harvesting data quality notes into metadata and into effective caveats for power users
- Effective labeling of carefully vetted data versus experimental data
- Communicating data quality while building trust
- Other …

Conclusion

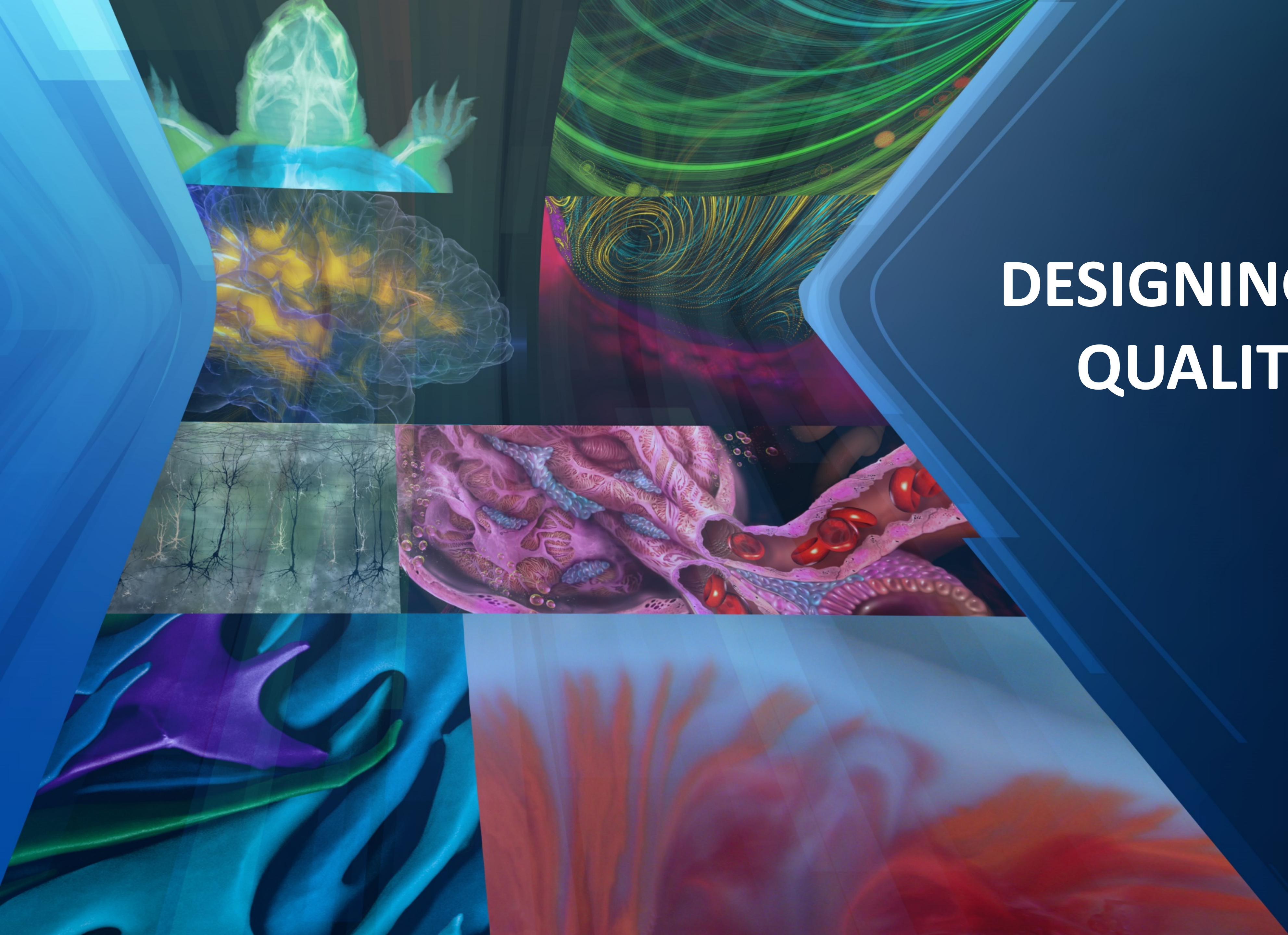
- All data have problems, but do the problems matter for the decision at hand?
- Data managers should consider all possible data quality problems, deal with problems that can reasonable be addressed, and document how they dealt each problem for their successors
- Include data quality in guides for power users and summarize the problems for an elevator speech to tell occasional users how far they can take the data without misguiding decisions that have important consequences

Conclusion

- By using the structure and terminology of the Framework, we will have a common basis for sharing information about data quality across agencies and with the public
- A common language will support transparency about our current data and analyses and a common basis for considering improvements in data and analysis

For the details

- The full report is available at: https://nces.ed.gov/fcsm/pdf/FCSM.20.04_A_Framework_for_Data_Quality.pdf



DESIGNING A DATA QUALITY POLICY

Avital Percher
Office of the Director
09/10/21

NSF's Frame of Reference

Evidence Act	Learning Agendas: "Systematic way to identify the data agencies intend to collect, use, or acquire, as well as the methods and analytical approaches to facilitate the use of evidence in policymaking"
Federal Data Strategy	The CDO shall "ensure that, to the extent practicable, the agency maximizes the use of data in the agency"
Mission (NSF Foundation Act, 1950)	To promote the progress of science; to advance the national health, prosperity, and welfare; and to secure the national defense; and for other purposes.

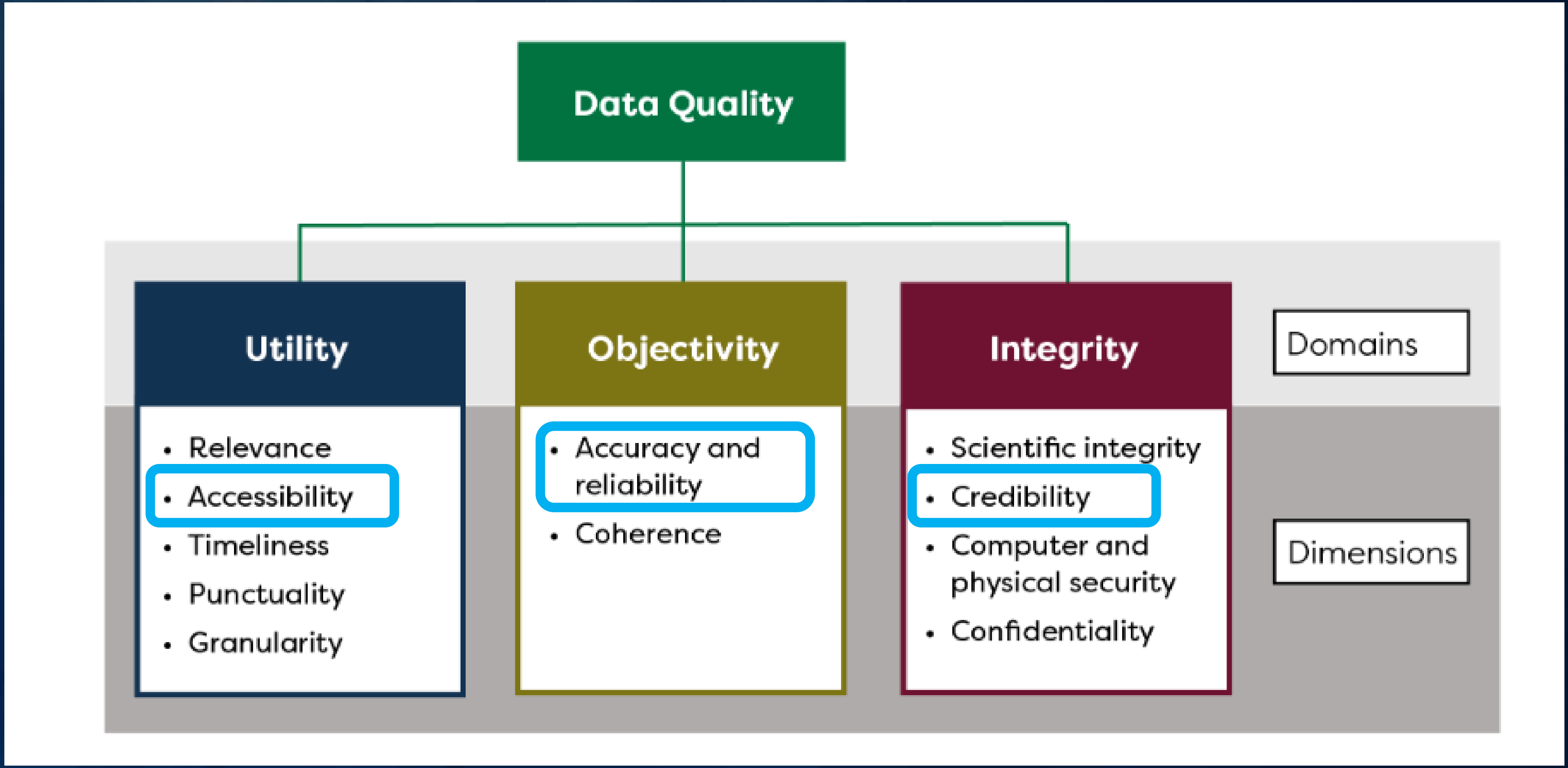
Leveraging the Data Quality Framework

"It applies to all data: data collections and data systems; restricted and public use microdata files; data products produced through data integration, modeling, harmonization and other statistical analyses; **and analysis outputs,** such as tables, estimates, graphics and reports." *FCSM Data Quality Framework*

To be a **strategic asset**, data must be **transparent**, **verified**, and **documented**...

across the **data**, **information**, **knowledge** stages.

Dimensions of the Data Quality Framework



Data Quality

Utility
- Relevance
- Accessibility
- Timeliness
- Punctuality
- Granularity

Objectivity
- Accuracy and reliability
- Coherence

Integrity
- Scientific integrity
- Credibility
- Computer and physical security
- Confidentiality

Domains

Dimensions

ATLAS Experiment at CERN

Roads

Transport

Route

To be a **strategic asset**, data must be **transparent**, **verified**, and **documented**...

across the **data**, **information**, **knowledge** stages.

Roads
∨
Data Inventory

Transport
∨
Analytic Tools

Route
∨
Analytic Outputs

Data Quality Policy

Data Inventory

Data management lifecycle is standardized, accessible and detailed

---

Data Inventory Standards

Analytic Tools

Tools processing the data are documented and vetted

Analytic Tool Standards

Analytic Outputs

Queries and analyses are documented and reproducible

Best Practices for Analytics Documentation

Data Quality Policy - Process

Data Inventory

Inventories prepared by Data Stewards

⌄

Feedback provided by internal user community

⌄

Reviewed and approved by Data Governing Body

Analytic Tools

Tools used at Enterprise level

⌄

Methods and documentation reviewed by Data Governing Body

Analytic Outputs

Output generated by office

⌄

Documented and archived internally by office standards

Thank You!

Avital Percher: apercher@nsf.gov
Dorothy Aronson (CIO/CDO): daronson@nsf.gov

Data Inventory Standards – Objectives

Defines metadata documentation standards and review process

Compliance	**Roles & Responsibilities**	**Documentation Maintenance**	**Master Metadata Schema**	**Validation**
Support Agency compliance with federal mandates	Define the roles and responsibilities in the Data Inventory Management Process	Define the requirements for maintaining metadata profiles and data dictionaries of NSF's data repositories	Define a master metadata and dictionary schema as an agency standard	Define a user inclusive validation process

Analytic Tools Standards - Objectives

Defines documentation requirements and validation process for tools used on an ‘enterprise’ level.

Community Standards

Define a community standard of excellence and support leadership’s need for trustworthy and vetted data tools.

Tool Documentation Benchmark

Establish a benchmark for tool documentation to promote development and application practices that align with community best practices.

Review and Approval

Describe a review and approval process by the EADGSC to support the NSF community’s need for tools vetted by data experts.

Best practices for analytics documentation- Objectives

Defines guidelines for documenting analytics outputs

Improved Quality Standards	**Replication**	**Knowledge Dissemination**
Enhance the quality and trustworthiness of the data collection and analysis.	Enable replication of the analysis as needed in the future, by both the office and others.	Allow the adaptation of the study to other needs of the community, increasing efficiency.

Data Inventory Standards

EDI RPTSQL PILOT STAGES

STAGE 1

Collect data lineage information from technical data stewards

Output
Draft metadata & data dictionary for RPTSQL tables

STAGE 2

Validate data with domain data steward expertise and submit for Data Governance (EAGDSC) approval

Output
Finalized metadata & data dictionary
Inputs for master data management

STAGE 3

Publish validated and approved data for internal NSF use

Output
Published & searchable data inventory (metadata & data dictionary)

[1] Stage 1 corresponds with the beginning of Q3. Activities preceding Q3 are not included in the listed stages.

Quality Considerations for Alternative Data:

A Case Study using CORP5 Data

John Bieler
Senior Economist, CPI
September 10, 2021

Familiar?

Quality Considerations for Alternative Data in the BLS Producer and Consumer Price Indexes

Crystal Konny (CPI)
Bonnie Murphy (PPI)

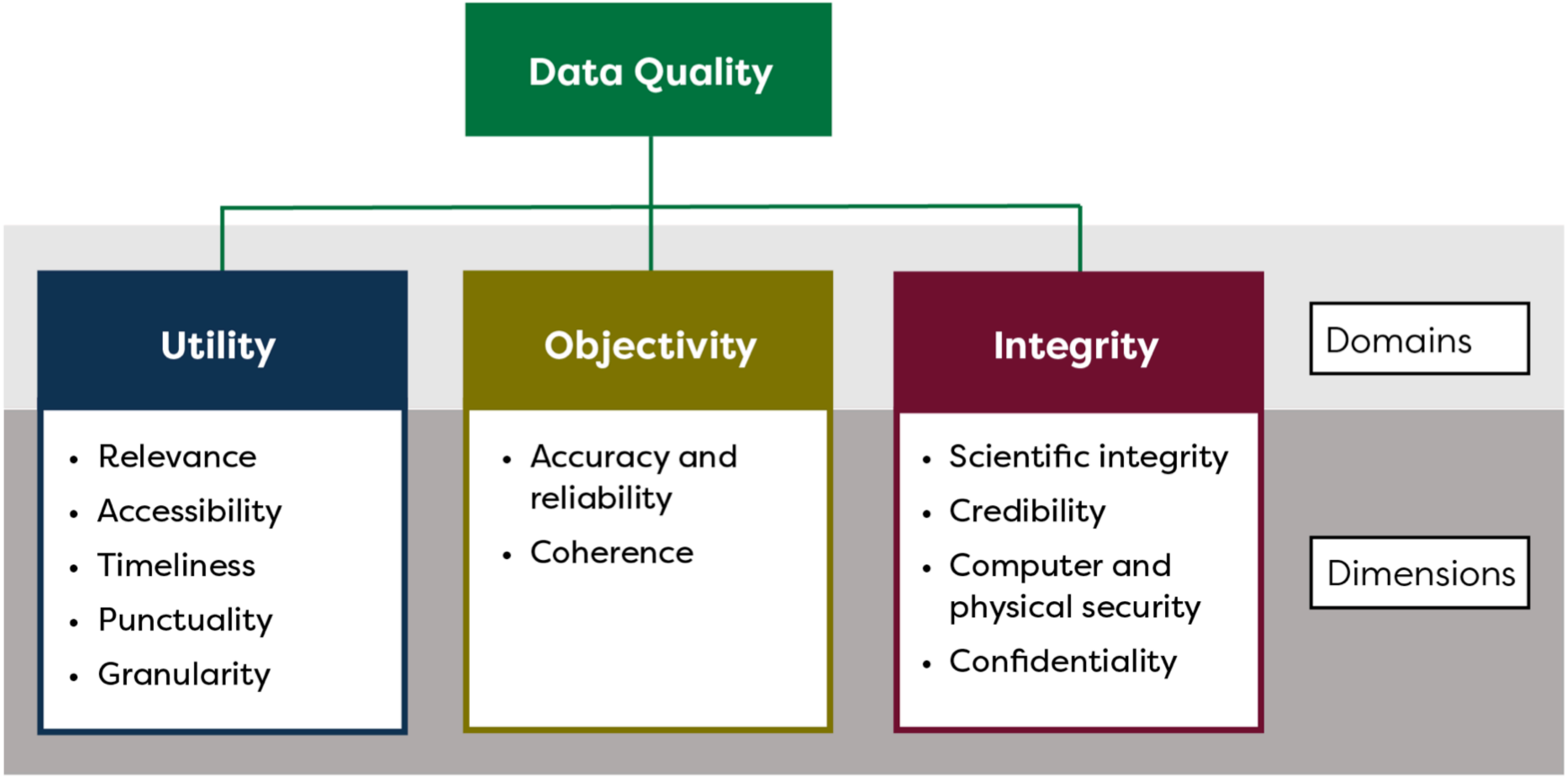
December 2017

Familiar?

Scorecard for Alternative Data

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						

The Framework



Data Quality

Utility

- Relevance
- Accessibility
- Timeliness
- Punctuality
- Granularity

Objectivity

- Accuracy and reliability
- Coherence

Integrity

- Scientific integrity
- Credibility
- Computer and physical security
- Confidentiality

Domains

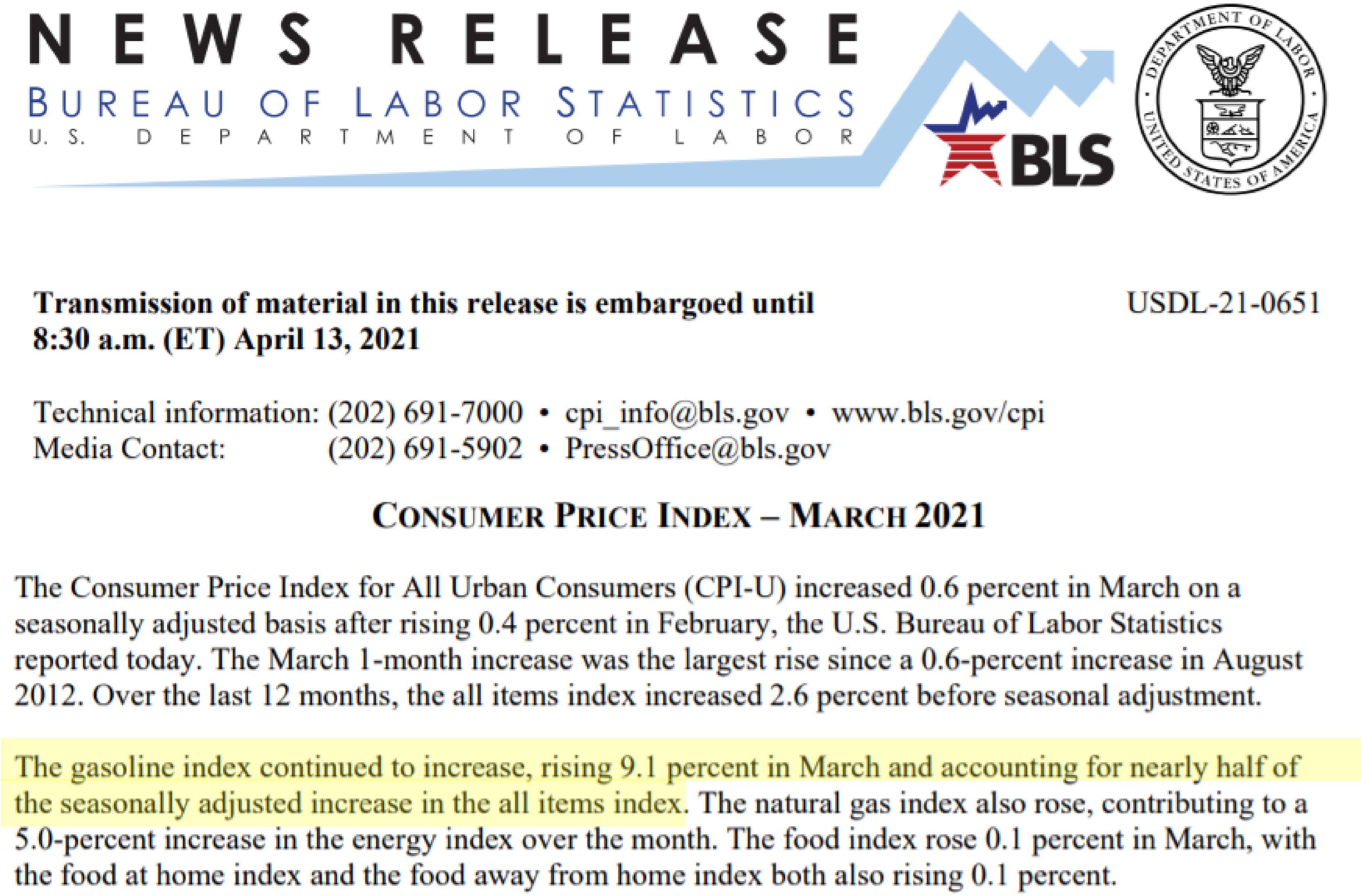
Dimensions

Background on CORP5

- CORP5 is a secondary source of gas price data
- Average of roughly 205,000 reported gas price observations every day
  - Roughly 6.23 million gas prices every month!
- Gas prices are updated in real-time
- CPI receives data the following day
- CORP5 data includes prices for three categories: Regular unleaded gasoline, Mid-grade, and Premium
- BLS obtained approval from CORP5 to use their data and began to voluntarily provide their data using a secure portal

CORP5 case study

Domain	Dimension	Definition	Question	Answer
Utility	**Relevance**	Relevance refers to whether the data product is targeted to meet current and prospective user needs.	Is the data a relevant input to our data products and measurement our measurement objective?	CORP5 provides daily gasoline prices for thousands of stations across the U.S. Produce indexes and average prices for gasoline and individual fuel types.
	Accessibility	Accessibility relates to the ease with which data users can obtain an agency's products and documentation in forms and formats that are understandable to data users.	Are the costs to access the data an effective use of resources? Will the methodology limit our ability to release data to users? How can we describe the methodology to data users?	CORP5 is providing the data on a voluntary basis. Make a public announcement in advance and provide materials on line, such as factsheets and articles.



NEWS RELEASE

BUREAU OF LABOR STATISTICS

U. S. DEPARTMENT OF LABOR

BLS

Transmission of material in this release is embargoed until 8:30 a.m. (ET) April 13, 2021

USDL-21-0651

Technical information: (202) 691-7000 • cpi_info@bls.gov • www.bls.gov/cpi
Media Contact: (202) 691-5902 • PressOffice@bls.gov

CONSUMER PRICE INDEX – MARCH 2021

The Consumer Price Index for All Urban Consumers (CPI-U) increased 0.6 percent in March on a seasonally adjusted basis after rising 0.4 percent in February, the U.S. Bureau of Labor Statistics reported today. The March 1-month increase was the largest rise since a 0.6-percent increase in August 2012. Over the last 12 months, the all items index increased 2.6 percent before seasonal adjustment.

The gasoline index continued to increase, rising 9.1 percent in March and accounting for nearly half of the seasonally adjusted increase in the all items index. The natural gas index also rose, contributing to a 5.0-percent increase in the energy index over the month. The food index rose 0.1 percent in March, with the food at home index and the food away from home index both also rising 0.1 percent.

CORP5 case study cont.

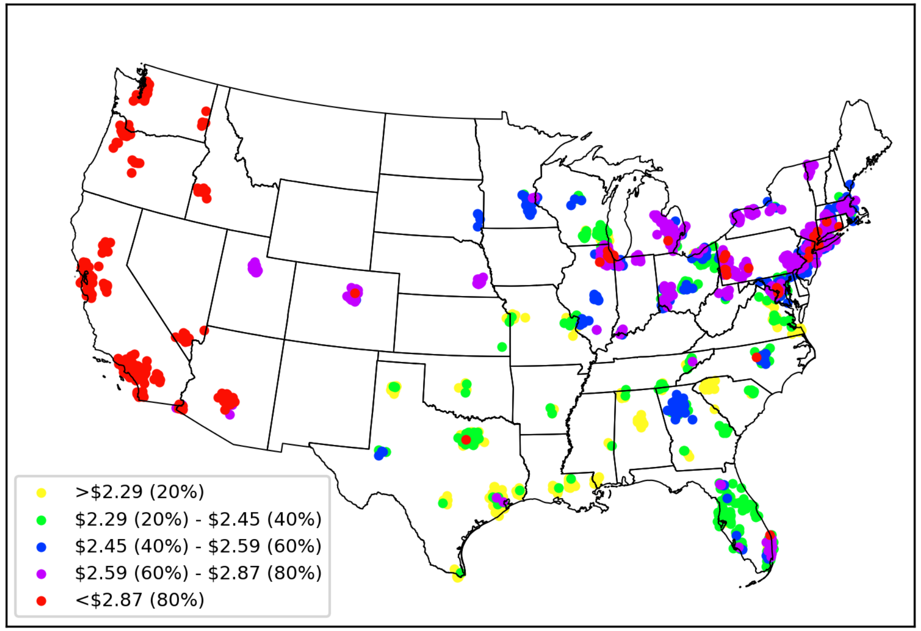
Domain	Dimension	Definition	Question	Answer
Utility	**Timeliness**	Timeliness is the length of time between the event or phenomenon the data describe and their availability.	Are the data representative of the index reference period?	Yes, daily prices across the month.
	Punctuality	Punctuality is measured as the time lag between the actual release of the data and the planned target date for data release.	Can the methodology be implemented within the typical production processing schedule? What is the probability and impact on the production schedule due to delayed delivery of data or unexpected time needed to process data?	Yes, CORP5 will be implemented into the current production schedule. We are currently parallel testing. Based on multi year evaluation period, the probability of an impact is low.

Corp5 case study cont.

Domain	Dimension	Definition	Question	Answer
Utility	**Granularity**	Granularity refers to the amount of disaggregation available for key data elements. Granularity can be expressed in units of time, level of geographic detail available, or the amount of detail available on any of a number of characteristics (e.g. (demographic, socio-economic).	Is there adequate data to support the current level of granularity in data products? Is there sufficient data to adequately protect confidentiality?	Yes, we will produce price indexes and average price products at the same level of granularity. Yes, thousands of gas stations protecting confidentiality.

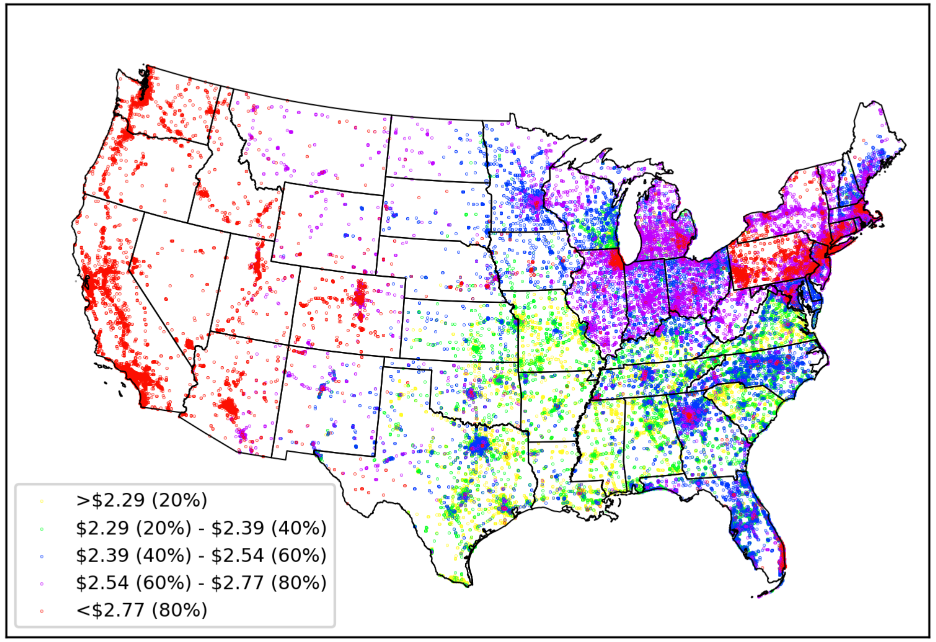
Comparison of regular gas prices

CPI


Gas prices (reg. grade) - November 6, 2019 in CPI samples

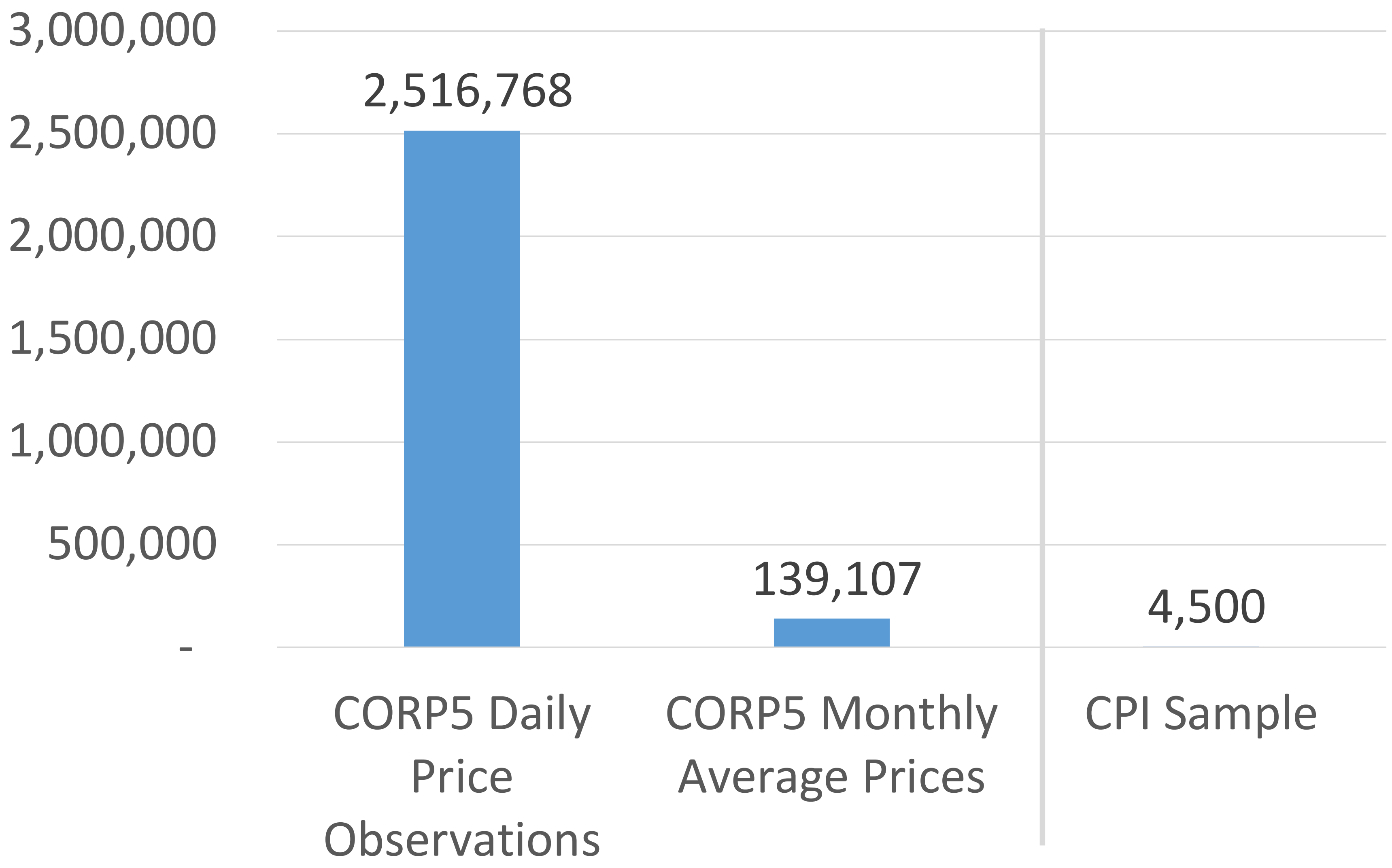
● >$2.29 (20%)
● $2.29 (20%) - $2.45 (40%)
● $2.45 (40%) - $2.59 (60%)
● $2.59 (60%) - $2.87 (80%)
● <$2.87 (80%)

CORP 5


Gas prices (reg. grade) - November 6, 2019

○ >$2.29 (20%)
○ $2.29 (20%) - $2.39 (40%)
○ $2.39 (40%) - $2.54 (60%)
○ $2.54 (60%) - $2.77 (80%)
○ <$2.77 (80%)

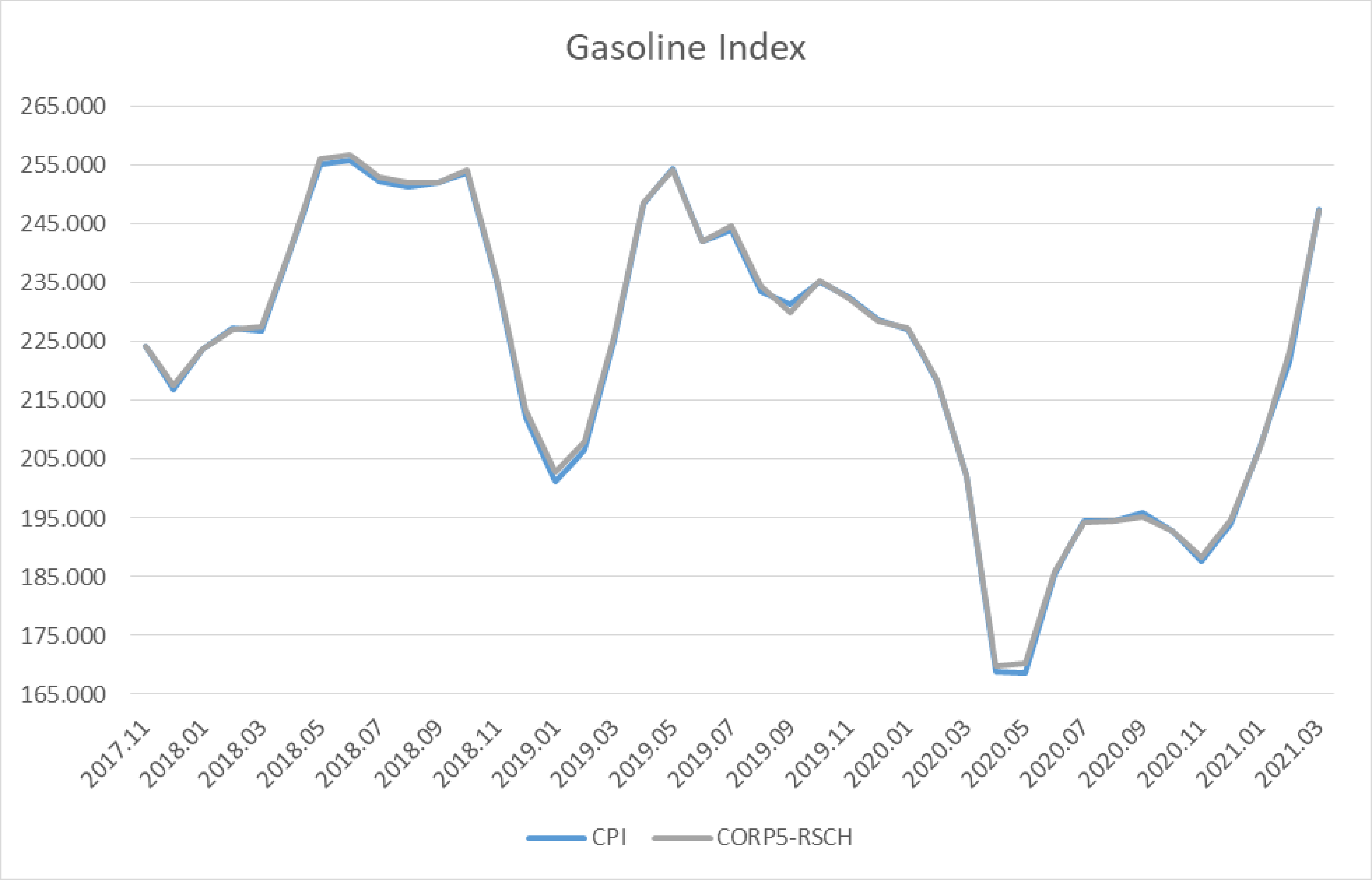
Comparing number of prices



CORP5 case study cont.

Domain	Dimension	Definition	Question	Answer
Objectivity	**Accuracy and reliability**	Accuracy measures the closeness of an estimate from a data product to its true value. Reliability, a related concept, characterizes the consistency of results when the same phenomenon is measured or estimated more than once under similar conditions.	Any concerns with the qualitative assessment of total measurement error?	No, research results compared favorably to the CPI Gasoline index at the U.S. Level.
	Coherence	Coherence is defined as the ability of the data product to maintain common definitions, classification, and methodological processes, to align with external statistical standards, and to maintain consistency and comparability with other relevant data.	Does the methodology impact the ability to compare CPI data with external sources? Is the methodology coherent with other CPI methodologies (not just what it is replacing)?	No, the methodology is still comparable with external sources. Yes, a mix of geomeans and Laspeyres index methodology. Added additional aggregation steps.

CORP5 Research – Differences never greater than 1% at U.S. level for gasoline



CORP5 Stored Meta Data

AREA	AREA_DESC	NUM_PR_OBS	NUM_RELATIVES	NUM_PHYS_LOCATIONS
0000	U.S.	9,576,611	129,755	50,049
N000	Non-Self-Representing PSUs	3,745,235	42,976	16,753
S000	Self-Representing PSUs	5,831,376	86,779	33,296
S12A	New York-Newark-Jersey City, NY-NJ-PA	265,885	10,187	3,932
S23A	Chicago-Naperville-Elgin, IL-IN-WI	901,102	6,855	2,462
S49A	Los Angeles-Long Beach-Anaheim, CA	447,457	7,086	2,408

CORP5 case study cont.

Domain	Dimension	Definition	Question	Answer
Integrity	**Scientific integrity**	Scientific integrity refers to an environment that ensures adherence to scientific standards and use of established scientific methods to produce and disseminate objective data products and one that shields these products from inappropriate political influence.	What is the probability and impact of the data provider (either maliciously or unintentionally) interfering with the data in a way that impacts estimates?	The probability is low and the impact is low. There is no incentive for the data provider to manipulate the data.
	Credibility	Credibility characterizes the confidence that users place in data products based simply on the qualifications and past performance of the data producer.	Review the output of index simulations. The more a simulation deviates from production, the more of an understanding approvers would like to have of the cause of differences.	Often cited source in news organizations and widely accepted by users as a credible source of price information.

CORP5 case study cont.

Domain	Dimension	Definition	Question	Answer
Integrity	**Computer and physical security**	Computer and physical security of data refers to the protection of information throughout the collection, production, analysis, and development process from unauthorized access or revision to ensure that the information is not compromised through corruption or falsification.	What is the probability and impact of risks of a loss of data or data quality issues due to technical issues?	The fallback plan is to use CPI collected data.

CORP5 case study cont.

Domain	Dimension	Definition	Question	Answer
Cost effectiveness (CPI addition)			Are the new data and methods cost effective relative to the data and methods they are replacing? Include development costs contracting costs, data collection costs, data storage, and maintenance costs.	Using the CORP5 data is cost neutral at this point.

Contact Information

John Bieler

Senior Economist

Consumer Price Index

(202) 691-5407

bieler.john@bls.gov

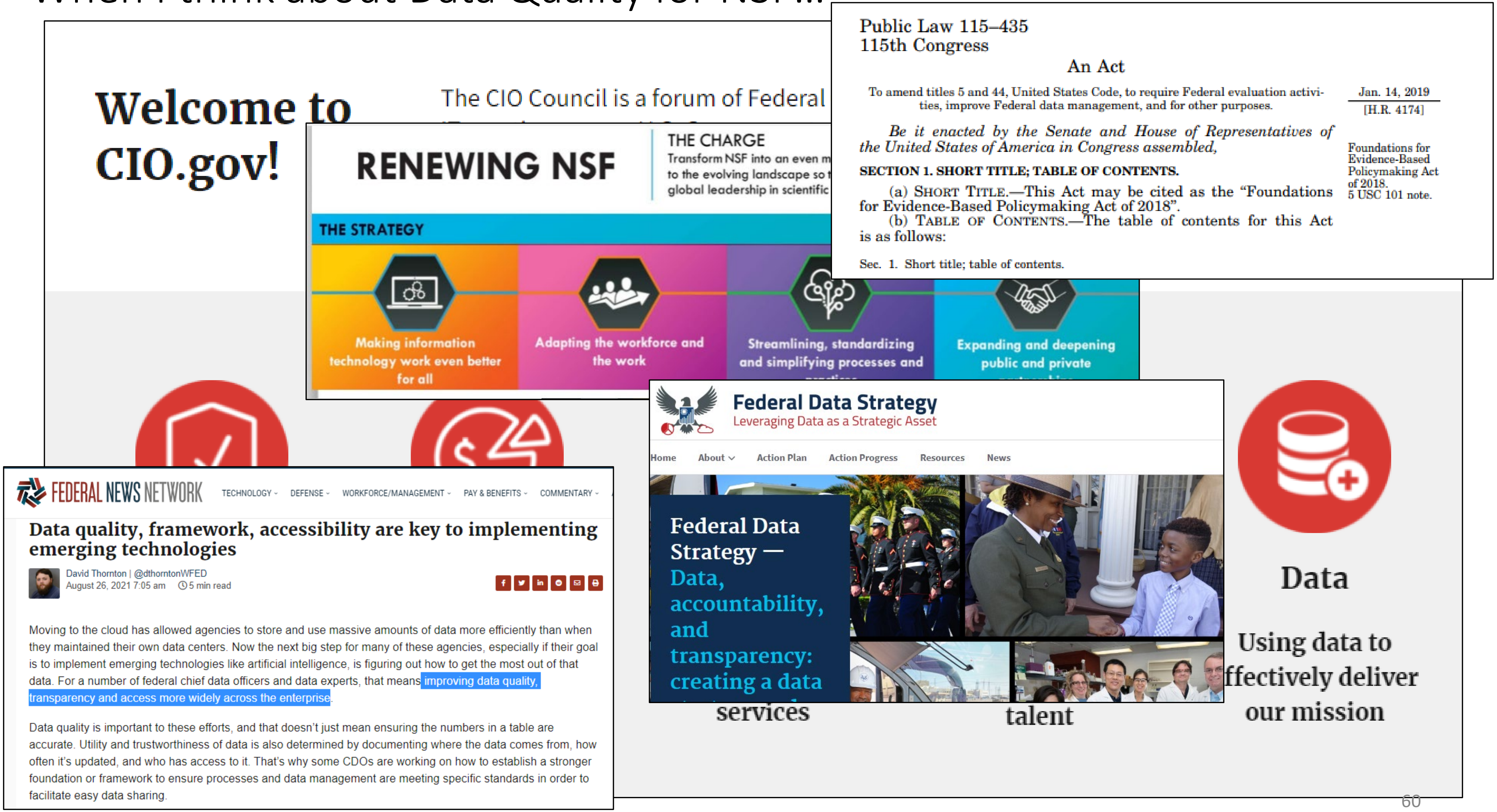
Data Quality Evolution

Dorothy Aronson

Chief Information Officer/Chief Data Officer

09/10/21

When I think about Data Quality for NSF…



Welcome to CIO.gov!

The CIO Council is a forum of Federal

RENEWING NSF

THE CHARGE

Transform NSF into an even m
to the evolving landscape so t
global leadership in scientific

THE STRATEGY

- Making information technology work even better for all
- Adapting the workforce and the work
- Streamlining, standardizing and simplifying processes and
- Expanding and deepening public and private

Public Law 115–435
115th Congress

An Act

To amend titles 5 and 44, United States Code, to require Federal evaluation activities, improve Federal data management, and for other purposes.

Jan. 14, 2019
[H.R. 4174]

Be it enacted by the Senate and House of Representatives of the United States of America in Congress assembled,

Foundations for Evidence-Based Policymaking Act of 2018.
5 USC 101 note.

SECTION 1. SHORT TITLE; TABLE OF CONTENTS.

(a) SHORT TITLE.—This Act may be cited as the “Foundations for Evidence-Based Policymaking Act of 2018”.

(b) TABLE OF CONTENTS.—The table of contents for this Act is as follows:

Sec. 1. Short title; table of contents.

Federal Data Strategy
Leveraging Data as a Strategic Asset

Home About Action Plan Action Progress Resources News

Federal Data Strategy — *Data, accountability, and transparency: creating a data*

Data

Using data to ffectively deliver our mission

services

talent

FEDERAL NEWS NETWORK

TECHNOLOGY DEFENSE WORKFORCE/MANAGEMENT PAY & BENEFITS COMMENTARY

Data quality, framework, accessibility are key to implementing emerging technologies

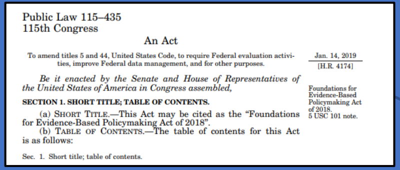
David Thornton | @dthorntonWFED
August 26, 2021 7:05 am 5 min read

Moving to the cloud has allowed agencies to store and use massive amounts of data more efficiently than when they maintained their own data centers. Now the next big step for many of these agencies, especially if their goal is to implement emerging technologies like artificial intelligence, is figuring out how to get the most out of that data. For a number of federal chief data officers and data experts, that means improving data quality, transparency and access more widely across the enterprise.

Data quality is important to these efforts, and that doesn’t just mean ensuring the numbers in a table are accurate. Utility and trustworthiness of data is also determined by documenting where the data comes from, how often it’s updated, and who has access to it. That’s why some CDOs are working on how to establish a stronger foundation or framework to ensure processes and data management are meeting specific standards in order to facilitate easy data sharing.

Source: Data quality, framework, accessibility are key to implementing emerging technologies | Federal News Network

When I think about Data Quality for NSF…



Public Law 115–435
115th Congress

An Act

To amend titles 5 and 44, United States Code, to require Federal evaluation activities, improve Federal data management, and for other purposes.

Jan. 14, 2019
[H.R. 4174]

Be it enacted by the Senate and House of Representatives of the United States of America in Congress assembled,

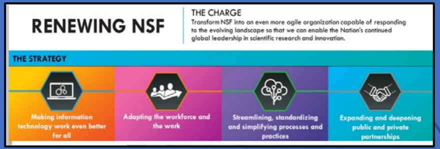
SECTION 1. SHORT TITLE; TABLE OF CONTENTS.

(a) Short Title.—This Act may be cited as the "Foundations for Evidence-Based Policymaking Act of 2018".

(b) Table of Contents.—The table of contents for this Act is as follows:

Sec. 1. Short title; table of contents.

Foundations for Evidence-Based Policymaking Act of 2018.
5 USC 101 note.



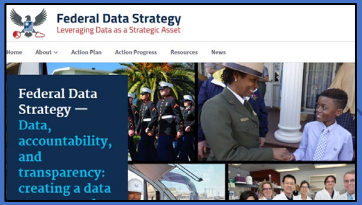
RENEWING NSF

THE CHARGE

Transform NSF into an even more agile organization capable of responding to the evolving landscape so that we can enable the Nation's continued global leadership in scientific research and innovation.

THE STRATEGY

- Making information technology work even better for all
- Adapting the workforce and the work
- Streamlining, standardizing and simplifying processes and practices
- Expanding and deepening public and private partnerships

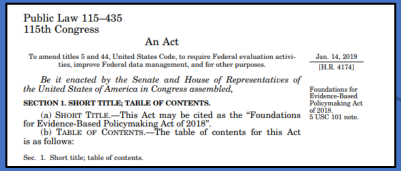


Federal Data Strategy
Leveraging Data as a Strategic Asset

Federal Data Strategy — Data, accountability, and transparency: creating a data

CIO

When I think about Data Quality for NSF…



Public Law 115–435
115th Congress

An Act

To amend titles 5 and 44, United States Code, to require Federal evaluation activities, improve Federal data management, and for other purposes.

Jan. 14, 2019
[H.R. 4174]

Be it enacted by the Senate and House of Representatives of the United States of America in Congress assembled,

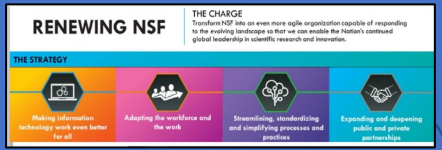
Foundations for Evidence-Based Policymaking Act of 2018.
5 USC 101 note.

SECTION 1. SHORT TITLE; TABLE OF CONTENTS.

(a) Short Title.—This Act may be cited as the "Foundations for Evidence-Based Policymaking Act of 2018".

(b) Table of Contents.—The table of contents for this Act is as follows:

Sec. 1. Short title; table of contents.



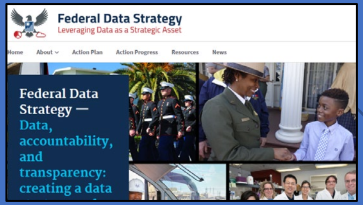
RENEWING NSF

THE CHARGE

Transform NSF into an even more agile organization capable of responding to the evolving landscape so that we can enable the Nation's continued global leadership in scientific research and innovation.

THE STRATEGY

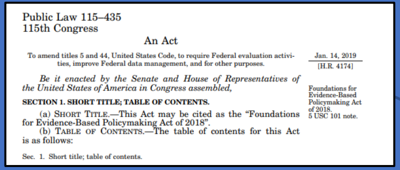
- Making information technology work even better for all
- Adapting the workforce and the work
- Streamlining, standardizing and simplifying processes and practices
- Expanding and deepening public and private partnerships



Federal Data Strategy
Leveraging Data as a Strategic Asset

Federal Data Strategy — Data, accountability, and transparency: creating a data

When I think about Data Quality for NSF…



Public Law 115–435
115th Congress

An Act

To amend titles 5 and 44, United States Code, to require Federal evaluation activities, improve Federal data management, and for other purposes.

Jan. 14, 2019
[H.R. 4174]

Be it enacted by the Senate and House of Representatives of the United States of America in Congress assembled,

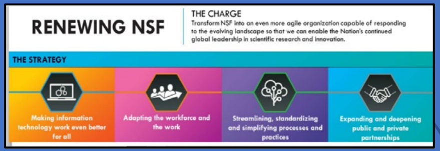
SECTION 1. SHORT TITLE; TABLE OF CONTENTS.

(a) Short Title.—This Act may be cited as the "Foundations for Evidence-Based Policymaking Act of 2018".

(b) Table of Contents.—The table of contents for this Act is as follows:

Foundations for Evidence-Based Policymaking Act of 2018.
5 USC 101 note.

Sec. 1. Short title; table of contents.



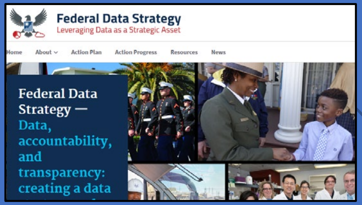
RENEWING NSF

THE CHARGE

Transform NSF into an even more agile organization capable of responding to the evolving landscape so that we can enable the Nation's continued global leadership in scientific research and innovation.

THE STRATEGY

- Making information technology work even better for all
- Adapting the workforce and the work
- Streamlining, standardizing and simplifying processes and practices
- Expanding and deepening public and private partnerships



Federal Data Strategy

Leveraging Data as a Strategic Asset

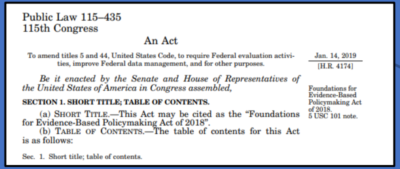
Federal Data Strategy — Data, accountability, and transparency: creating a data

CIO

Ut oh…

When I think about Data Quality for NSF…

Scientific Integrity



Public Law 115–435
115th Congress

An Act

To amend titles 5 and 44, United States Code, to require Federal evaluation activities, improve Federal data management, and for other purposes.

Jan. 14, 2019
[H.R. 4174]

Be it enacted by the Senate and House of Representatives of the United States of America in Congress assembled,

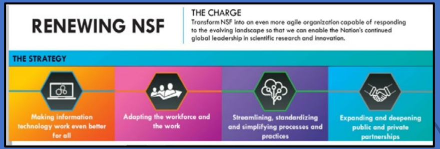
SECTION 1. SHORT TITLE; TABLE OF CONTENTS.

(a) Short Title.—This Act may be cited as the "Foundations for Evidence-Based Policymaking Act of 2018".

Foundations for Evidence-Based Policymaking Act of 2018.
5 USC 101 note.

(b) Table of Contents.—The table of contents for this Act is as follows:

Sec. 1. Short title; table of contents.



RENEWING NSF

THE CHARGE

Transform NSF into an even more agile organization capable of responding to the evolving landscape so that we can enable the Nation's continued global leadership in scientific research and innovation.

THE STRATEGY

- Making information technology work even better for all
- Adapting the workforce and the work
- Streamlining, standardizing and simplifying processes and practices
- Expanding and deepening public and private partnerships

Computer and Physical Security

Confidentiality

Credibility

Federal Data Strategy
Leveraging Data as a Strategic Asset

Home
About
Action Plan
Action Progress
Resources
News

Federal Data Strategy — Data, accountability, and transparency: creating a data

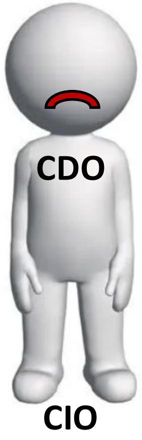
Punctuality

Relevance

Granularity

Accessibility

Reliability



What does "data quality" even mean?

Accuracy

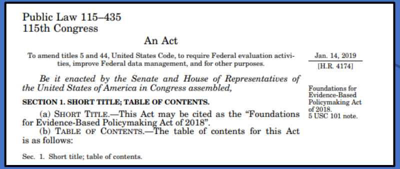
Timeliness

Coherence

FCSM offers a framework.

Scientific Integrity

Confidentiality



Public Law 115–435
115th Congress

An Act

To amend titles 5 and 44, United States Code, to require Federal evaluation activities, improve Federal data management, and for other purposes.

Jan. 14, 2019
[H.R. 4174]

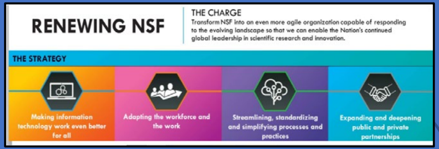
Be it enacted by the Senate and House of Representatives of the United States of America in Congress assembled,

Foundations for Evidence-Based Policymaking Act of 2018.
5 USC 101 note.

SECTION 1. SHORT TITLE; TABLE OF CONTENTS.

(a) Short Title.—This Act may be cited as the "Foundations for Evidence-Based Policymaking Act of 2018".
(b) Table of Contents.—The table of contents for this Act is as follows:

Sec. 1. Short title; table of contents.



RENEWING NSF

THE CHARGE
Transform NSF into an even more agile organization capable of responding to the evolving landscape so that we can enable the Nation's continued global leadership in scientific research and innovation.

THE STRATEGY

- Making information technology work even better for all
- Adapting the workforce and the work
- Streamlining, standardizing and simplifying processes and practices
- Expanding and deepening public and private partnerships

Computer and Physical Security

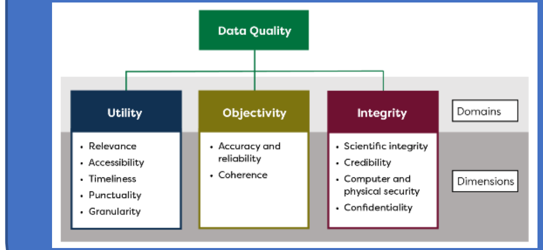
Federal Data Strategy
Leveraging Data as a Strategic Asset

Home
About
Action Plan
Action Progress
Resources
News

Federal Data Strategy — Data, accountability, and transparency: creating a data

Punctuality

Credibility



Data Quality

Utility	Objectivity	Integrity
Relevance Accessibility Timeliness Punctuality Granularity	Accuracy and reliability Coherence	Scientific integrity Credibility Computer and physical security Confidentiality

cessibility

Reliability

Granularity

Accuracy

CDO

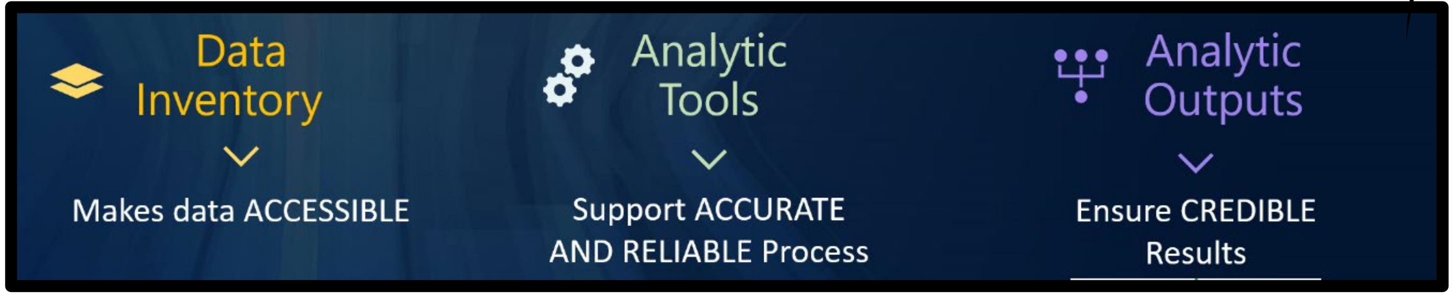
What does "data quality" even mean?

Timeliness

CIO

FCSM
Federal Committee on Statistical Methodology

NSF aligns within the framework.



Data Inventory

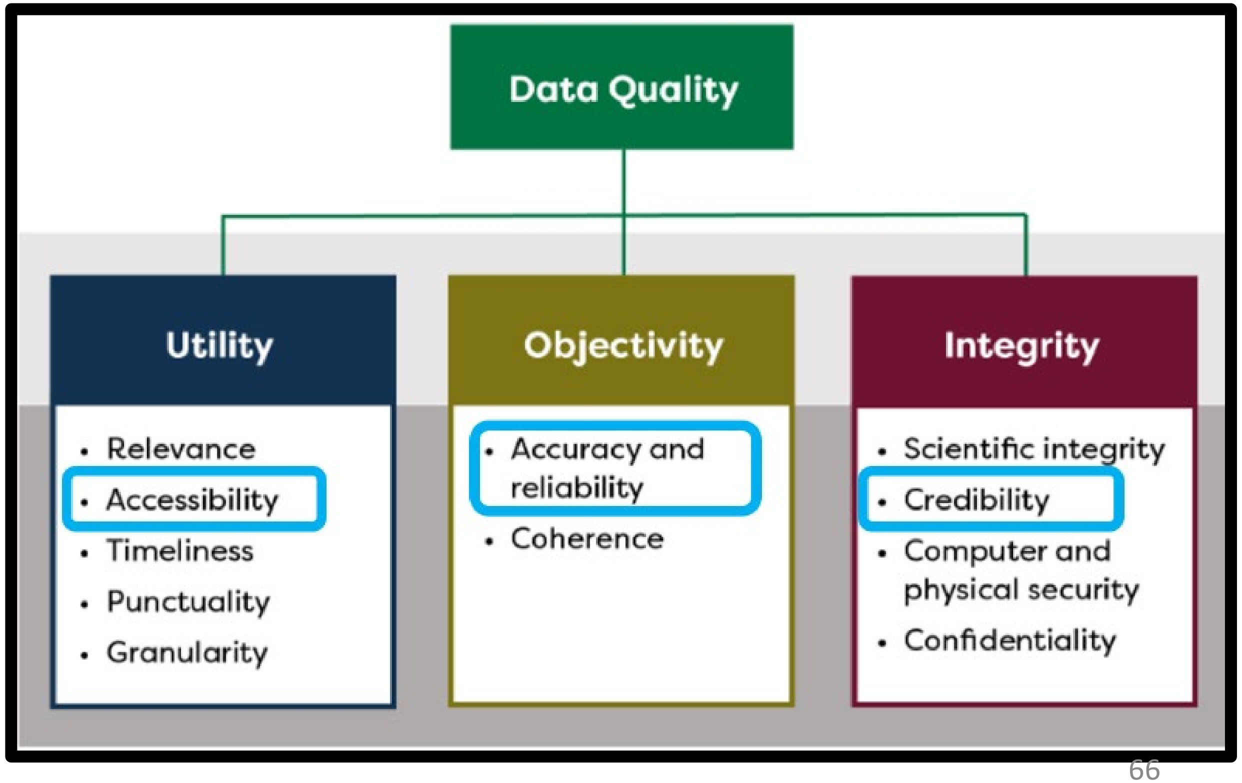
Makes data ACCESSIBLE

Analytic Tools

Support ACCURATE AND RELIABLE Process

Analytic Outputs

Ensure CREDIBLE Results

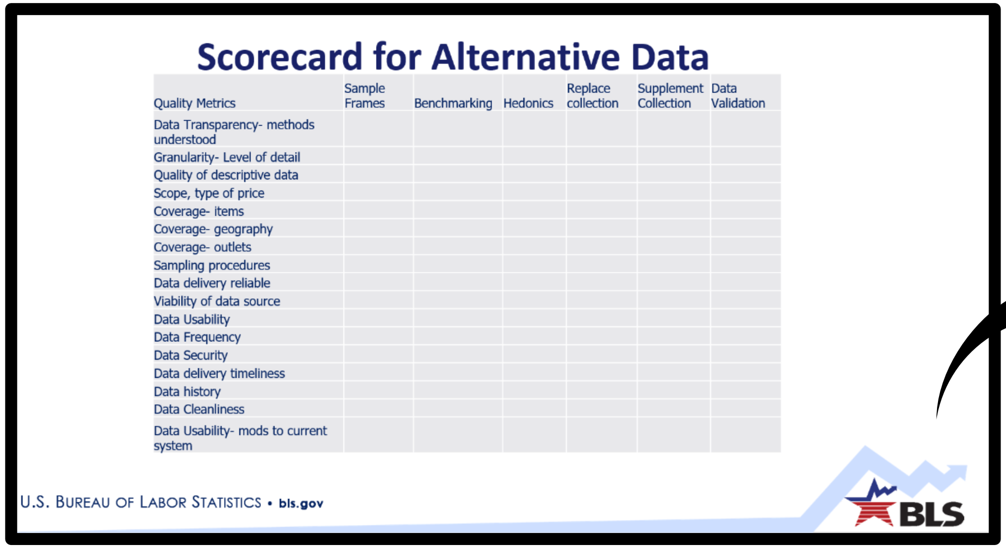


Data Quality

Utility
- Relevance
- Accessibility
- Timeliness
- Punctuality
- Granularity

Objectivity
- Accuracy and reliability
- Coherence

Integrity
- Scientific integrity
- Credibility
- Computer and physical security
- Confidentiality

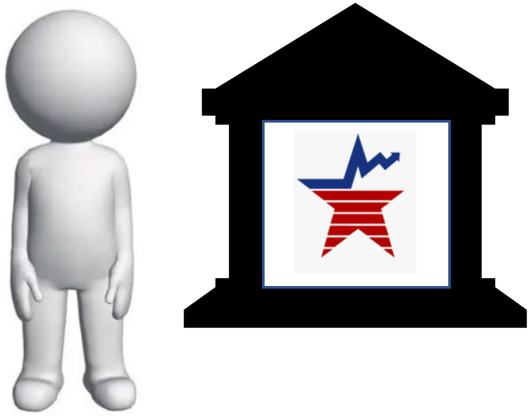


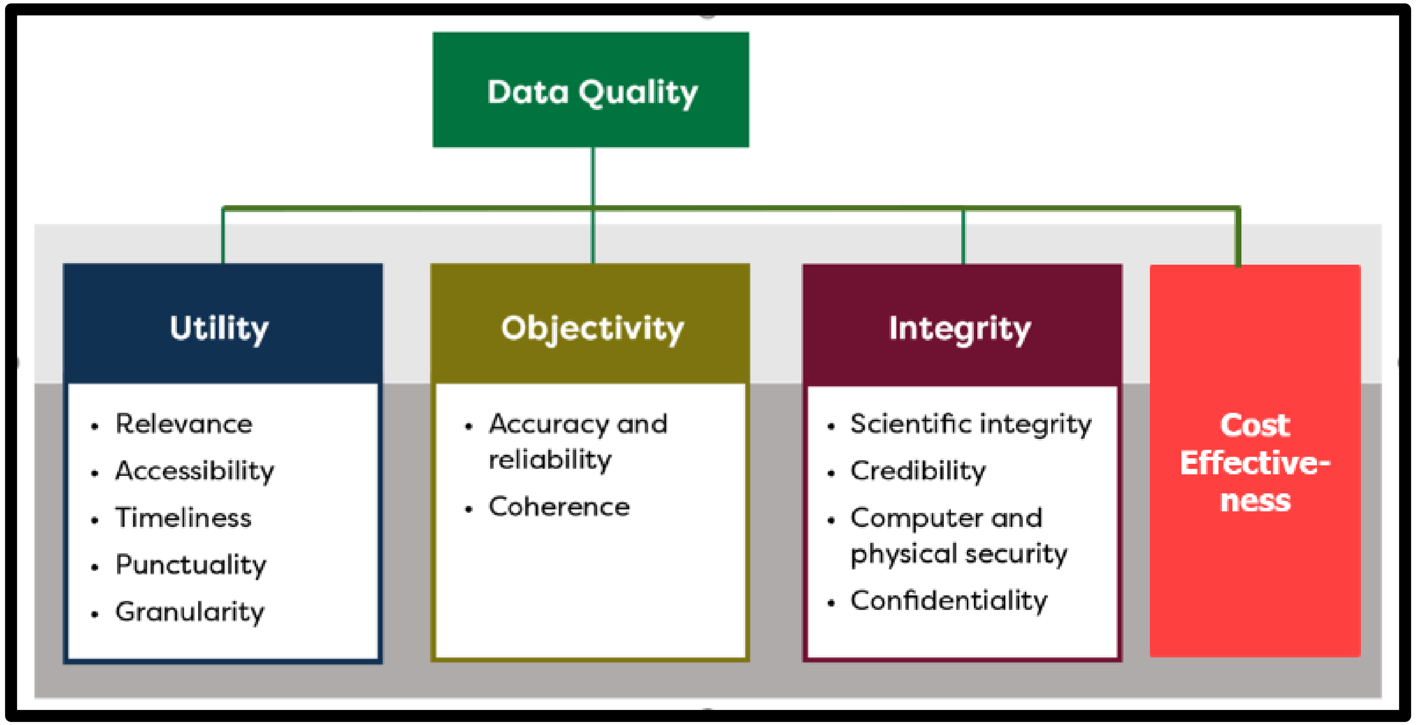
Scorecard for Alternative Data

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						

U.S. Bureau of Labor Statistics • bls.gov

BLS demonstrates alignment with the framework.





Data Quality

Utility
- Relevance
- Accessibility
- Timeliness
- Punctuality
- Granularity

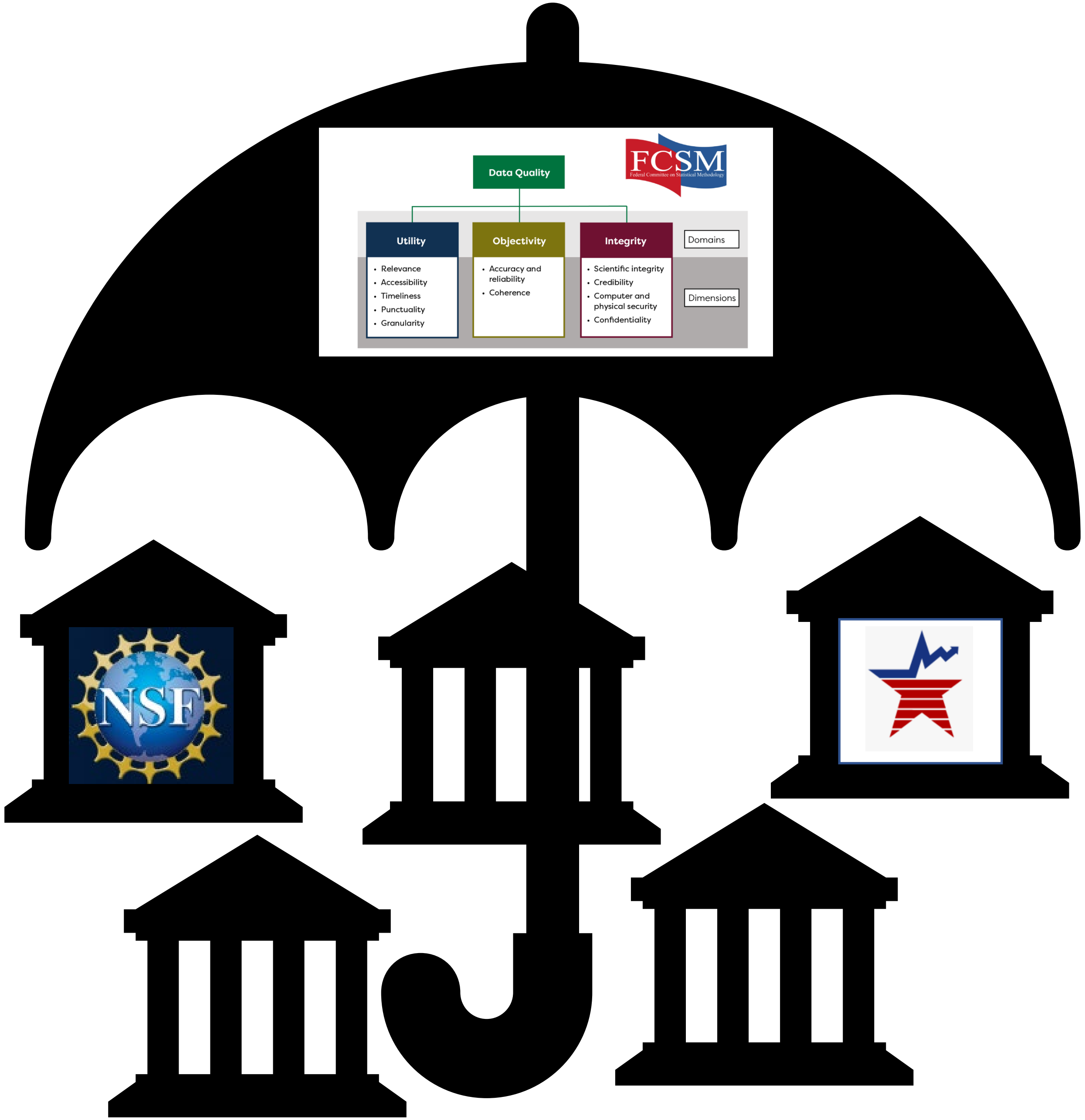
Objectivity
- Accuracy and reliability
- Coherence

Integrity
- Scientific integrity
- Credibility
- Computer and physical security
- Confidentiality

Cost Effective-ness

Framework provides a common language…

…allowing necessary variation to fit mission.

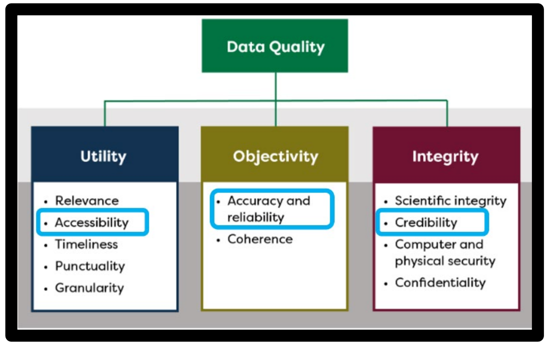

FCSM
Federal Committee on Statistical Methodology
NSF

Data Quality

- **Utility**
  - Relevance
  - Accessibility
  - Timeliness
  - Punctuality
  - Granularity
- **Objectivity**
  - Accuracy and reliability
  - Coherence
- **Integrity**
  - Scientific integrity
  - Credibility
  - Computer and physical security
  - Confidentiality

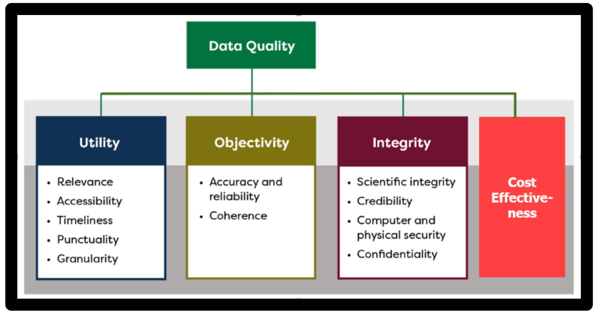
Domains

Dimensions



Data Quality

- **Utility**
  - Relevance
  - Accessibility
  - Timeliness
  - Punctuality
  - Granularity
- **Objectivity**
  - Accuracy and reliability
  - Coherence
- **Integrity**
  - Scientific integrity
  - Credibility
  - Computer and physical security
  - Confidentiality



Data Quality

- **Utility**
  - Relevance
  - Accessibility
  - Timeliness
  - Punctuality
  - Granularity
- **Objectivity**
  - Accuracy and reliability
  - Coherence
- **Integrity**
  - Scientific integrity
  - Credibility
  - Computer and physical security
  - Confidentiality
- **Cost Effectiveness**

NSF’s Data Quality: Lessons Learned

CENTER THE END USER

BOTTOM-UP AND TOP-DOWN TACTICS

UNDERSCORE AGENCY GOALS AND ALIGN WITH FEDERAL POLICIES

FOCUS ON PROGRESS OVER PERFECTION

NSF’s Data Quality Initiatives: Challenges and Solutions

ISSUE	CHALLENGE ⚠	SOLUTION ✓
Getting Started	When creating a policy from scratch there is significant time spent collecting artifacts.	Use the numerous existing resources to create the basis for the policy.
Maintaining Scope	Through the drafting and review process several other policy needs were identified.	Instead of incorporating the ideas into the draft policy, log the ideas for future policy development efforts.
Establishing an Inclusive Process	Numerous stakeholders have an interest in supporting the development of the policy.	Small teams assist in policy development. Iterative and inclusive review process. Tailored briefings for senior staff.
Making the Change Stick	Implementing a new policy requires buy-in across the agency.	(In process) Imbed Data Governance Group in policy implementation. Work to build a policy and tools that provide value to stakeholders.

BLS’ CORP5 Case Study: Standout Points

Integrating existing frameworks and resources (e.g., The Framework and the Scorecard for Alternative Data)

Showcasing Scaled impact across the alternative data sources

Strengthening evidence-building efforts by using secondary data sources to supplement agency data

Contact Information

Dorothy Aronson

CIO/CDO

NSF

daronson@nsf.gov

703.292.4299