USDA Economic Research Service
U.S. DEPARTMENT OF AGRICULTURE

Potential Value of Data and Free Access to These Data

Spiro Stefanou
Administrator
USDA Economic Research Service

Amos Golan
Research Professor of Economics
American University

Federal Committee on Statistical Methodology
2024 Research and Policy Conference
College Park, MD October 22-24, 2024
F-5: Appraise, Assess, & Apply: Advancing Alternative Data Use

Motivation and Objective

- Motivation
  - Better understand how to value the meaning and quality of information, with emphasis on observed information - data.
  - Develop an applicable method for evaluating the *potential value* of datasets (and access to these data).

- Objectives
  - Develop an approach (set of measures) for evaluating (or approximating) the *potential value* of datasets ('Option Value')That set needs to be applicable, interpretable and (relatively) simple to compute and evaluate
  - It must be *independent* of the inference
  - Test the approach with different (private and public) datasets
  - Develop a way to also evaluate the value of (free) access to that data
  - Converting the potential value into a monetary value

More on Potential Value:

- The overall value that society may obtain from a certain data set, assuming all the information and knowledge embedded in that data are extracted.
- •(It is not a value based on past use of the data, but rather the *complete* potential of that data, if indeed it will materialize.)
- •These measures are independent of the inferential approaches used (or to be used) when converting the information in the data into knowledge.

Brief Background and Basic Definitions

Information:

- Anything that informs us (a bit circular…)
- This thing that informs us – it reduces the bounds of uncertainty about possible outcomes/inferences/decisions…
- It is anything that effects our estimates (meaningful content)
- Anything that may affects our preferences and behavior

Note on "Having information": a weaker notion than having knowledge, or even beliefs.

- For the applied researcher who is interested in modeling, inference and learning this means that information is anything that may affect one's estimates, the uncertainties about these estimates, or decisions. It is "meaningful content."

Note on 'Truth': we take the view that, in general, information is true and is not intended to be false, though it is frequently noisy and imperfect, and its meaning may be subject to interpretational and processing errors

Remarks on Information

Information can be:

- Noisy and imperfect – subject to processing errors.
- Extrinsic or intrinsic.
- Absolute or relative.

Objective or subjective.

  - Objective: A physical law, an undisputed fact or an observed action.
  - Subjective: Relative to the person using the information (or making the judgement).
  - Include intuitions, assumptions and interpretations.

Valuing Information

- Value is always relative and subjective (relative to the user or decision maker, or…)
- •It cannot be absolute
- •It cannot be unique
- •Value also depends on meaning (but meaning is subject to interpretation) and meaning is also affected by context (the circumstances that form the setting for an event, statement, or idea, and in terms of which it can be fully understood and assessed)

Potential Value: Basic Building Blocks

The idea:

We want our measure to satisfy a minimal set of requirements (attributes)

- These requirements are organized as three hierarchical building blocks, each containing several attributes.
- Some of these attributes are measurable and can be objectively quantified.
- Others are qualitative or ordinal and at times subjective, and some are fuzzier.

Potential Value: Basic Building Blocks (cont'd)

The Blocks:

- At the bottom of the hierarchy is the first building block: *Data Reliability, Integrity, and Accuracy*. It comprises measures identifying the basic attributes of the data, including basic statistics.

- The second building block is *Data Quality*.
  - It comprises objective and quantitative measures as well as more complex attributes related to the meaning and semantics of the data.

- The third building block, at the highest level, is the *Potential Value of Data*. It comprises the first two building blocks as well as other relative (subjective) attributes related to meaning and importance.
- Note: That last part is the most complicated, as we must use meaning to evaluate value, but for 'meaning' we need a context.

More on the Blocks

The overall structure

- We require the *Reliability* block to exceed a minimal level. Otherwise, the data may be unusable. The exact minimal level depends on the potential use of the data.

- Conditional on that, the *Quality* block is calculated. If it satisfies our desires, we calculate the potential value according to the *Value* block.

- **Note:** Keeping in mind that data are scarce resources, it is usually impractical and illogical to provide these minimal thresholds.

The Building Blocks

Basic Block 3: Potential Value of Data	• Elements of the first two building blocks through the perspective of potential. • Attributes relate to meaning and importance. • Context dependent.
Basic Block 2: Data Quality	• Objective and quantitative measures. • Additional attributes that relate to semantics.
Basic Block 1: Data Reliability, Integrity and Accuracy	• Identify the basic attributes of the data and dataset. • Minimal satisfaction of attributes required.

Attributes of Building Block 1:
Data Reliability, Integrity & Accuracy

- Information and Entropy of a *Single* Random Variable
  - Entropy: Function of a variable's probability distribution; Free of semantics
  - Information: Inversely related to probability; Rare occurrences hold more information
- Information and Entropy of *Multiple* Random Variables
  - Joint, Marginal and Conditional Entropy
- Entropy Convergence
  - Entropy sequentially calculated
  - Used to test data integrity
  - Natural distribution and frequency of digits; confirms data integrity

Attributes of Building Block 1:
Data Reliability, Integrity & Accuracy

- The Shannon Limit
  - Max compressibility of bits of information w/o loss of information
  - Potential predictability

- Condition Number
  - Measures degree of collinearity; are data suitable for inference

- Benford's Law
  - Natural distribution and frequency of digits; confirms data integrity

Attributes of Building Block 2:
Data Quality

- Completeness, Quality and Documentation
- Representativeness, Trust and Believability
- Age of Data and Dataset
- Accessibility
- Interpretability and Heterogeneity
- Dependencies, Predictability and Embedded Information

Attributes of Building Block 3
Potential Value of Data

- Semantics, Meaning, Importance and Inference
- Extremes, Relationships, Questions, and Updated Inference and Uniqueness
- Timeliness, Availability and Consistence
- Heterogeneity, Quality, Size and Representativeness
- Expected Outcomes and Time Horizon

Three Case Studies

Case 1: Textbook example of movie release data (Greene, 2011)

- Small dataset with less than 70 observations (i.e., movies).

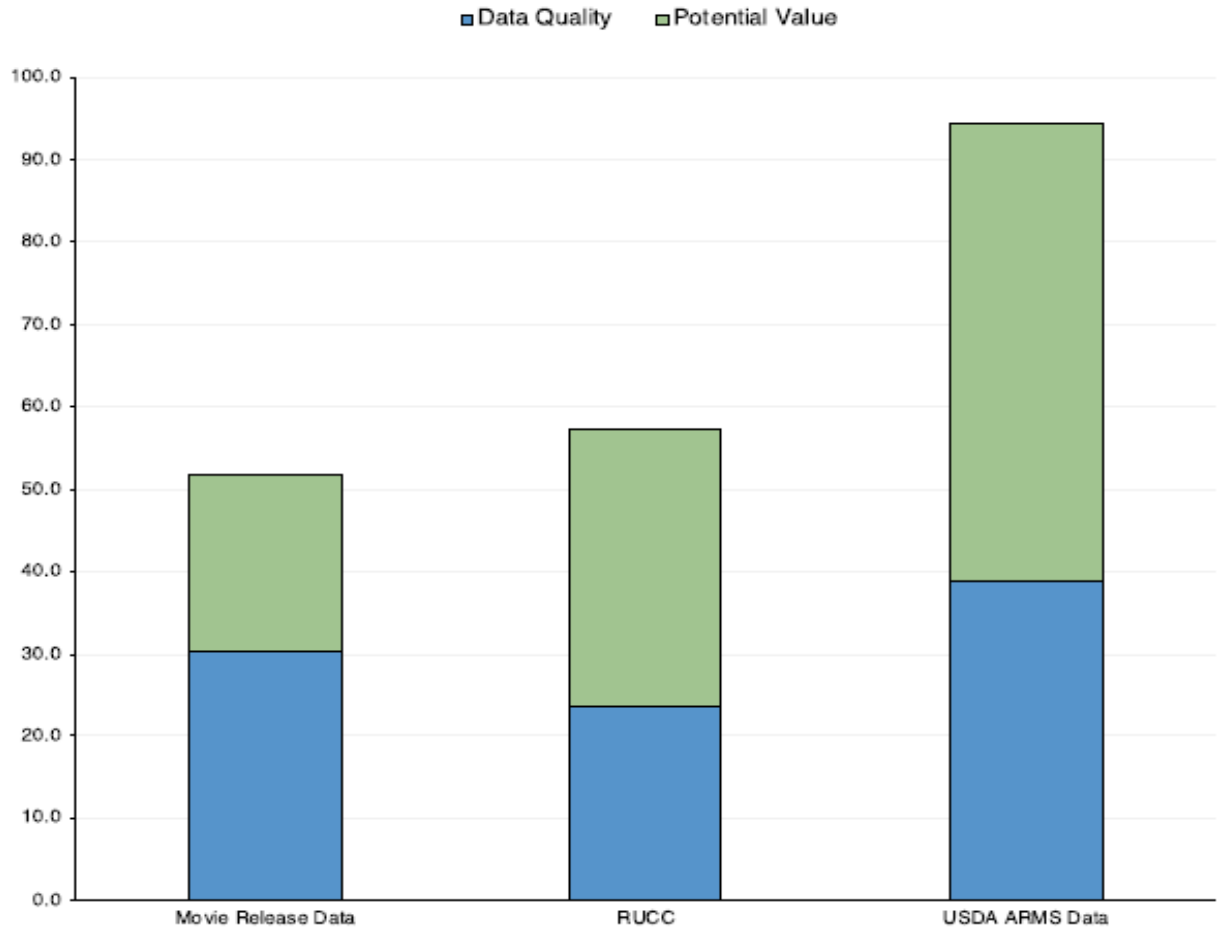
Case 2: Rural Urban Continuum Codes (RUCC)

- Short panel (wide) with 5 periods (one year per decade).
- Over 2 thousand observations at the county level.
- Identifiers and only one primary variable.

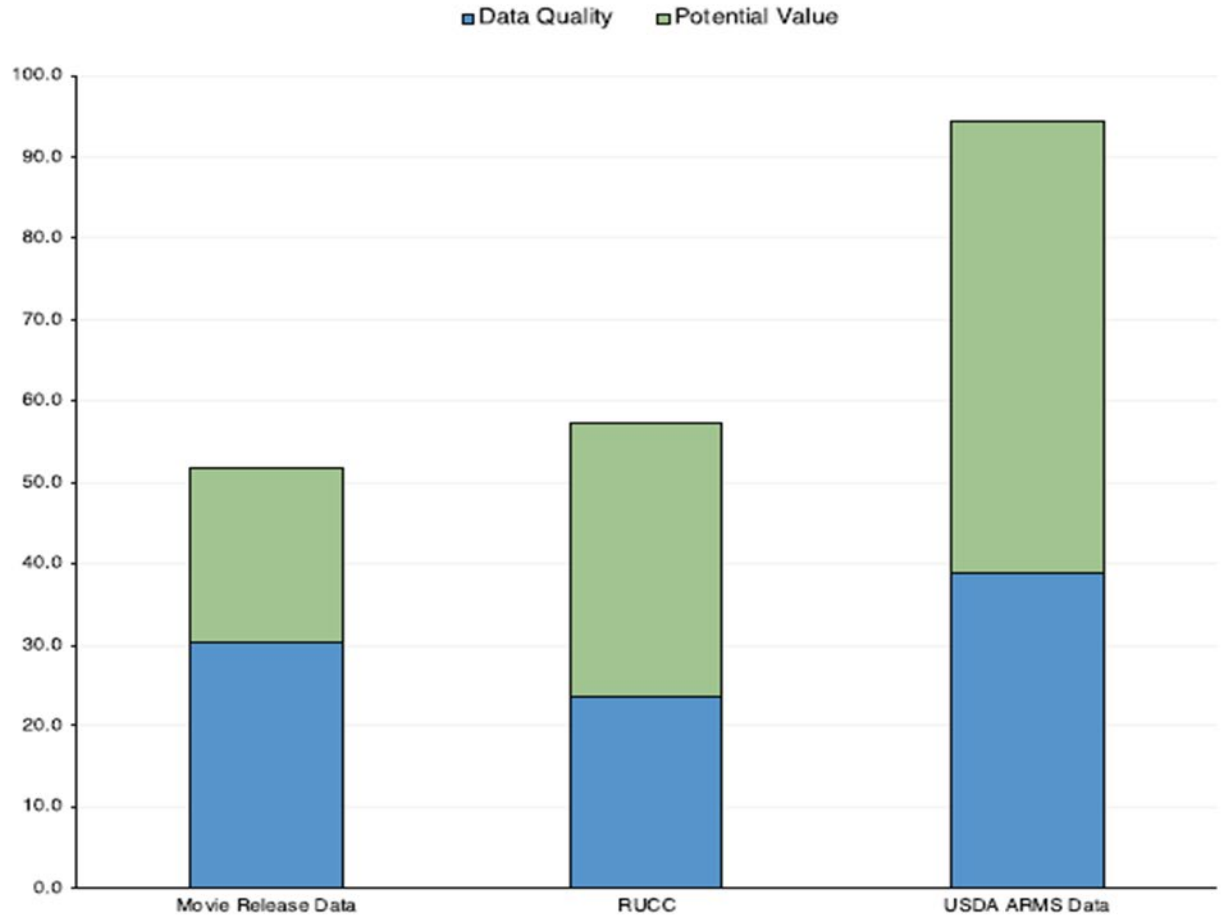
Case 3: Income and expenditure data from Agricultural Resource Management Survey (ARMS)

- Panel (long) with 19 years .
- Covers 15 states.
- Over two dozen different variables, which are more complex and interrelated.

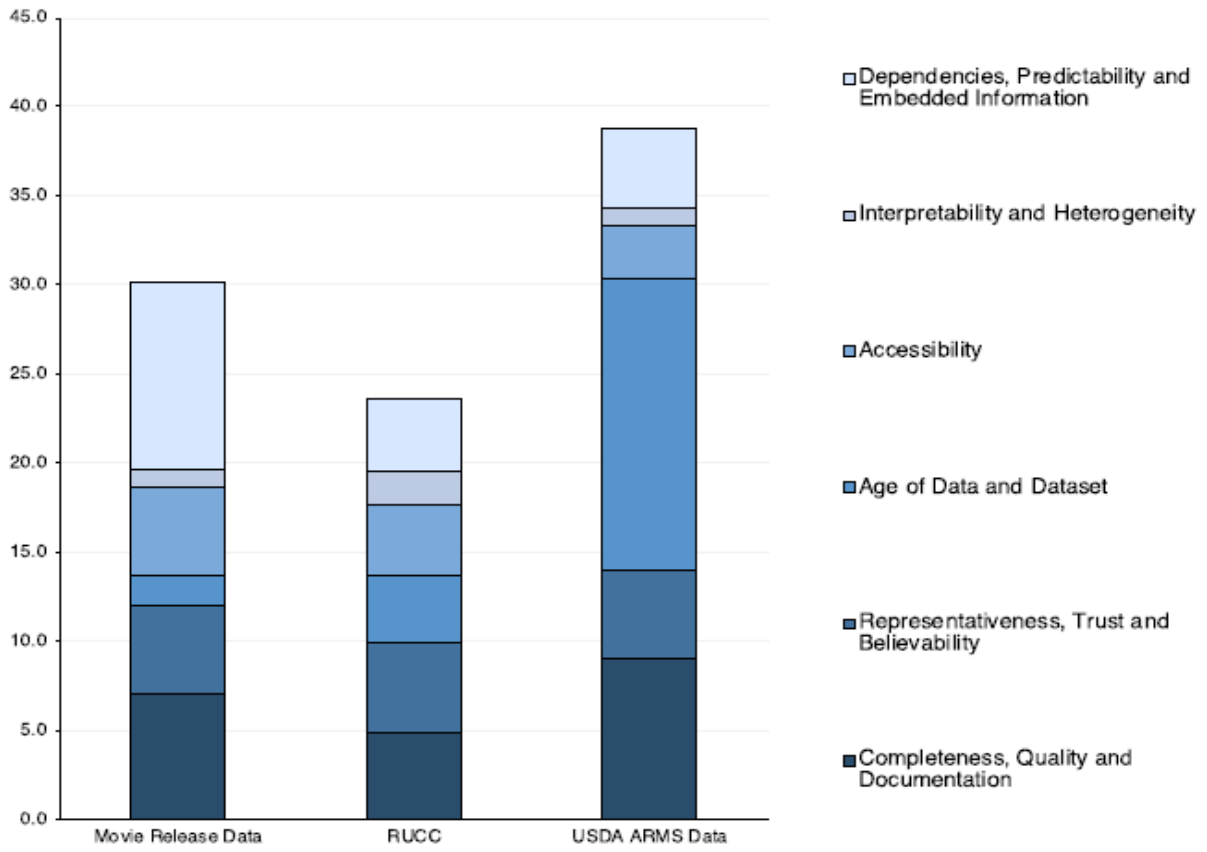
Quality and Value Blocks



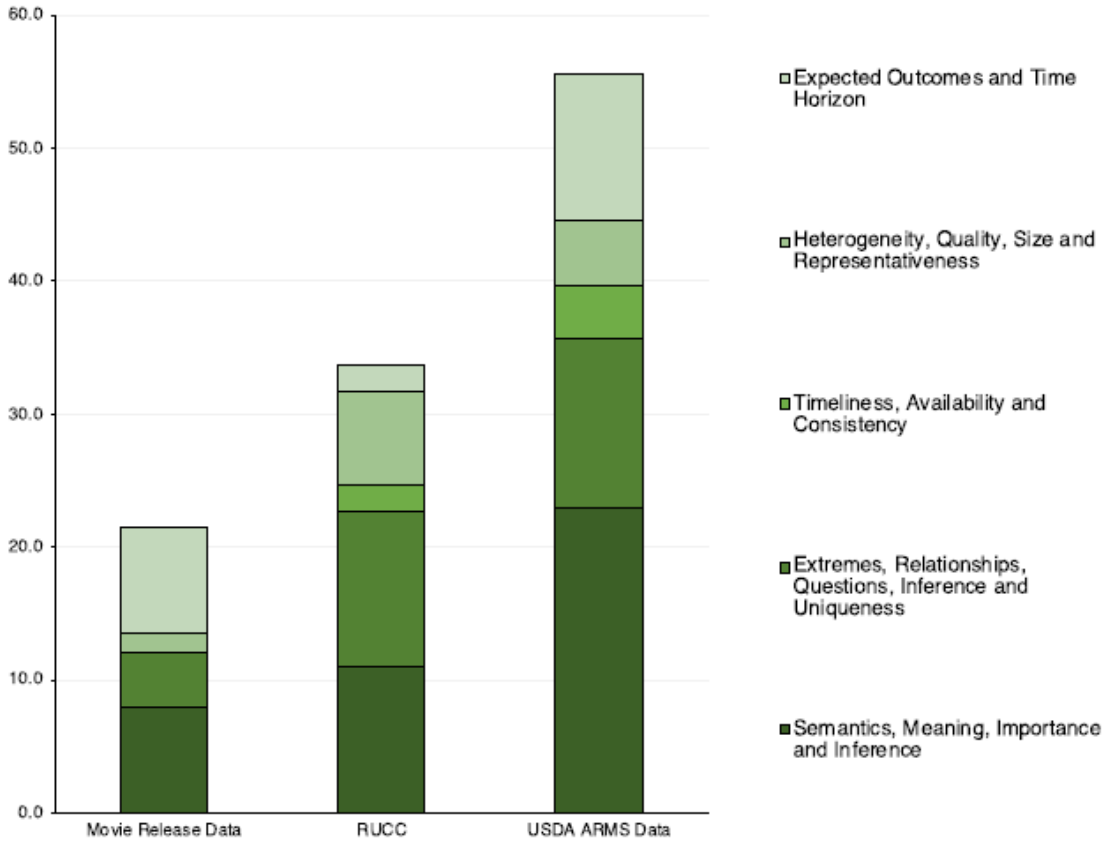
Quality and Value Blocks



Attributes of Data Quality: Basic Block 2



Attributes of Potential Value: Basic Block 3



Value of Access to Data

- The value of access to the data is also relative.
- But the value to access the data is positive only if the value of the dataset itself is positive.
- Practically, the potential value of the data can only be realized if we have access to that data.
- Therefore, if that access is free (publicly available) for all, and straightforward, then the full potential of the data can be materialized.
- **Note:** That logic implies that the potential value of data proposed here also includes access to all. Otherwise, the full potential of the data may not be realized.

Notes on Monetary Value

- Can we convert relative value into monetary value?
- Assume, that we were able to answer all of the potential questions that a dataset can answer (a subset of Block 3 attributes); The complete potential of the data materialized.
- With the above, we can calculate the potential benefit (in monetary units) of these answered questions to society. Is society's welfare improved?
- Do the data lead to a Pareto improvement?

- **Note:** This is not trivial and problem/dataset specific.

A Partial List of Open Questions

- How can AI contribute to the potential value of data? Value is independent of inference. Can AI impact value without inference?
- Can we improve how attributes regarding meaning and semantics are defined and evaluated?
- Are there other attributes that contribute to value? Looking for the minimal set.
- Should the attributes be independent and mutually exclusive of each other?
- Relative value reliant on scaling. Is there a better way?
- Can a set of axioms be developed?
- What is the impact of data aggregation on value? Is it problem specific, or is there a general ‘law’ to follow?

Comments & Questions

[Potential Value of Data and Free Access to Data | HCEO (uchicago.edu)]