Data Quality, Labeling, and Validation

Jared Edgerton

Why Data Quality Is Central

Data quality determines:

- What can be learned
- Which populations are visible
- How uncertainty propagates

Poor data quality cannot be fixed downstream.

Validity Threats in Big Social Data

Common threats include:

- Coverage bias
- Measurement error
- Selection effects
- Temporal instability

Scale amplifies, rather than removes, these problems.

Bias Enters at Collection

Data collection processes shape:

- Who is observed
- What is recorded
- Which events are missing

Bias is often structural, not accidental.

Media-Sourced Event Data

Event data based on news reporting reflect:

- Journalist locations
- Editorial priorities
- Press freedom and resources

Events in well-covered areas are overrepresented.

Geographic Reporting Bias

Consequences include:

- Urban bias
- Underreporting in rural or conflict zones
- Systematic cross-national differences

Absence of reports does not imply absence of events.

Social Media as Data

Social media data represent:

- Platform-specific user populations
- Non-random participation
- Algorithmic amplification

Users are not representative of populations.

Opt-In and Platform Studies

Opt-in data involve:

- Self-selection
- Differential attrition
- Strategic participation

Inference requires strong assumptions.

Satellite and Remote Sensing Bias

Remote sensing data face:

- Cloud cover and atmospheric interference
- Sensor resolution limits
- Strategic behavior to avoid detection

Measurement error varies across space and time.

Labeling and Annotation

Labels are often produced by:

- Human coders
- Crowdsourcing platforms
- Machine-assisted workflows

Labeling introduces additional uncertainty.

Inter-Coder Reliability

Reliability assesses:

- Agreement across coders
- Consistency of interpretation
- Ambiguity in categories

High agreement does not guarantee validity.

Reliability Example (Conceptual)

```
# Compute agreement statistics (e.g., Cohen's kappa)
```

```
# Compute agreement statistics across annotators
```

Validation Beyond Labels

Validation includes:

- Spot checks
- External benchmarks
- Temporal consistency
- Sensitivity analyses

No single validation step is sufficient.

Monitoring and Drift

Over time:

- Data-generating processes change
- Platforms update rules
- User behavior shifts

Static validation is inadequate.

Documenting Limitations

Good practice requires:

- Explicit statements of bias
- Known blind spots
- Assumptions required for inference

Transparency strengthens credibility.

Data Quality as Design

Quality should be:

- Anticipated at collection
- Monitored over time
- Reported alongside results

Validation is part of methodology.

What We Emphasize in Practice

- Treat bias as structural
- Validate at multiple stages
- Separate reliability from validity
- Document uncertainty explicitly

Discussion

- Which biases are hardest to detect?
- When is validation infeasible?
- How should uncertainty be communicated?