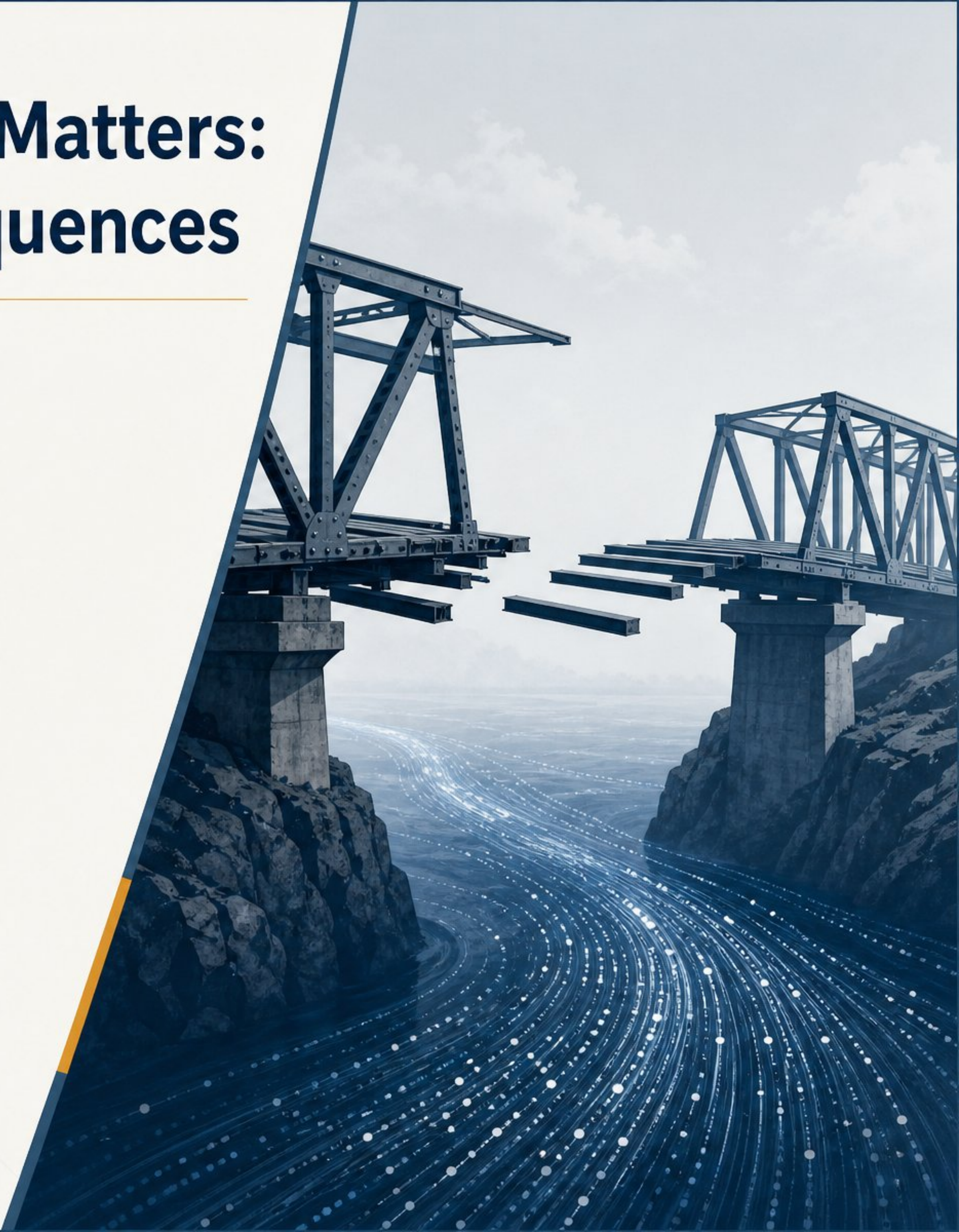
Understanding Missing Data:

Absence, Codes, and Limits

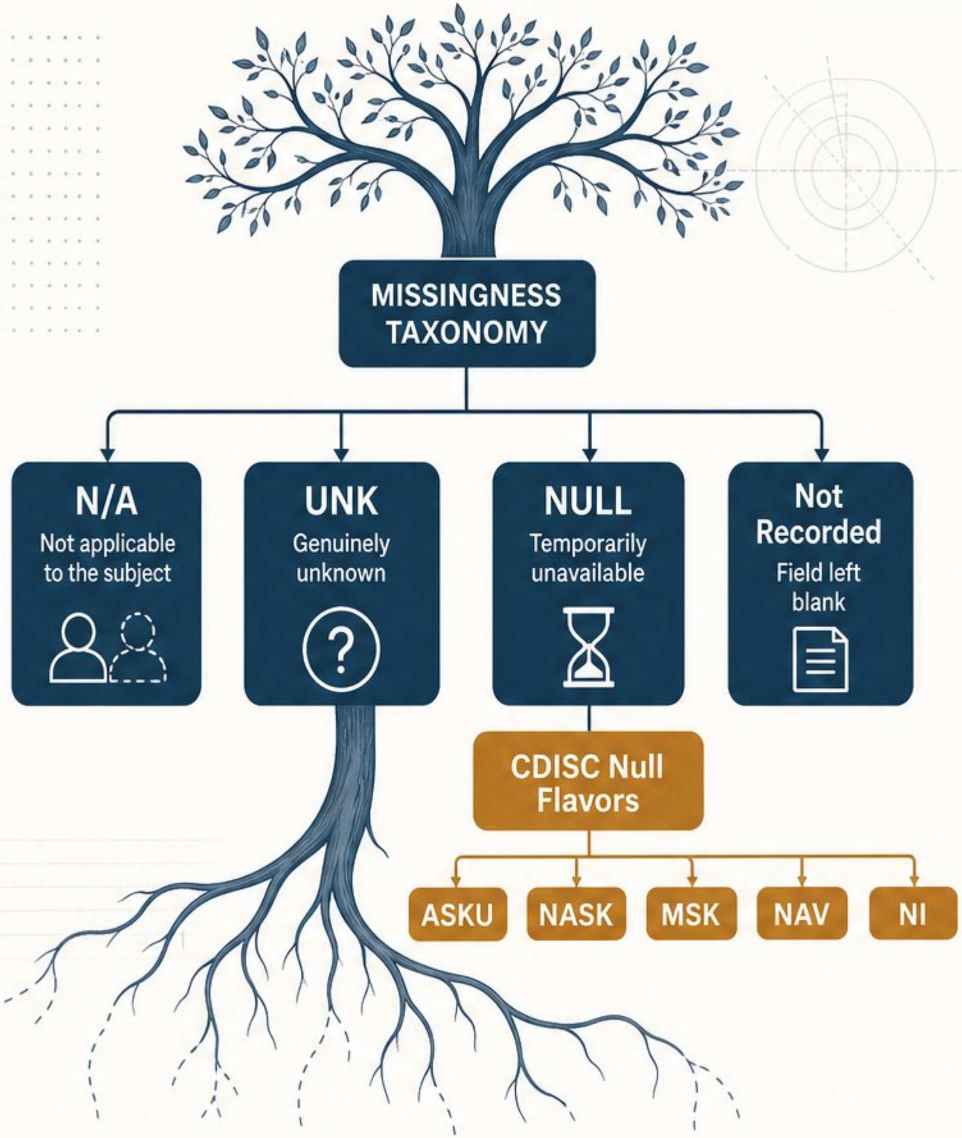
- - 'Missing data' encompasses more than empty cells; it's a signal with meaning.
- - Four key categories: Blank, Unknown, Not Applicable, and Not Recorded.
- - Treating all missingness equally causes interpretation errors and hidden bias.
- - Each type imposes distinct limits on analysis, statistical validity, and decisions.

Why Missing Data Matters: Real-World Consequences

- Mishandled data distorted AI/ML models, masking COVID-19 viral coinfections until MICE imputation used
- OHCA study: imputation choice flipped statistical significance of bystander CPR associations
- NSCLC trials: 70.2% reported missing PRO data, but only 12 of 52 primary-endpoint trials ran sensitivity analyses
- Complete-case analysis introduces selection bias under MNAR; ignoring missingness inflates false positives
- Missingness mechanisms (MCAR/MAR/MNAR) dictate valid handling strategies and interpretability limits



A Shared Language: Standard Definitions for Missingness



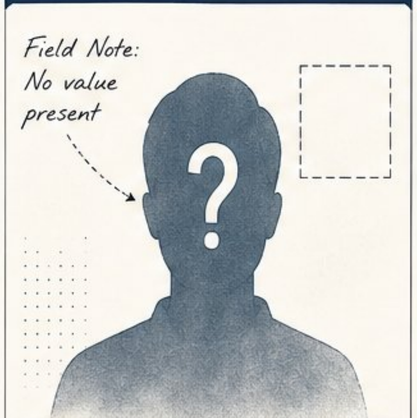
- **N/A:** Not applicable to the subject (e.g., pregnancy for males)
- **UNK:** Genuinely unknown and never recoverable (lost log, absent info)
- **NULL:** Temporarily unavailable; a placeholder for expected values
- **Not Recorded:** Data should exist but field was left blank

---

- **CDISC Null Flavors** refine: ASKU, NASK, MSK, NAV, NI
- **HL7 DataAbsentReason** codes align clinical system expectations
- **INSDC** uses "missing:" + reason for biosample metadata

Not All Blanks Are Equal: Unpacking the Four Categories

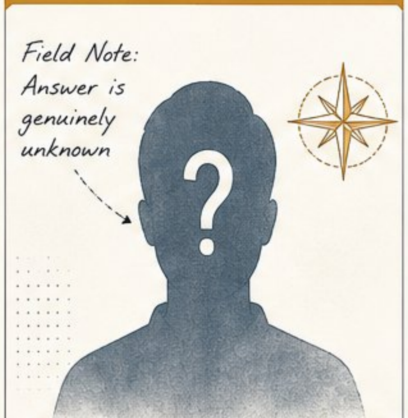
BLANK



SPECIMEN: MISSING-001

- Blank: may be truly missing, intentionally skipped, or structurally absent (implicit missingness)

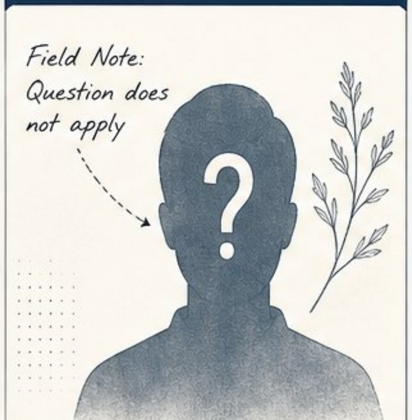
UNKNOWN



SPECIMEN: MISSING-002

- Unknown: respondent or system genuinely lacks the answer, will never be known (UNK / ASKU)

NOT APPLICABLE



SPECIMEN: MISSING-003

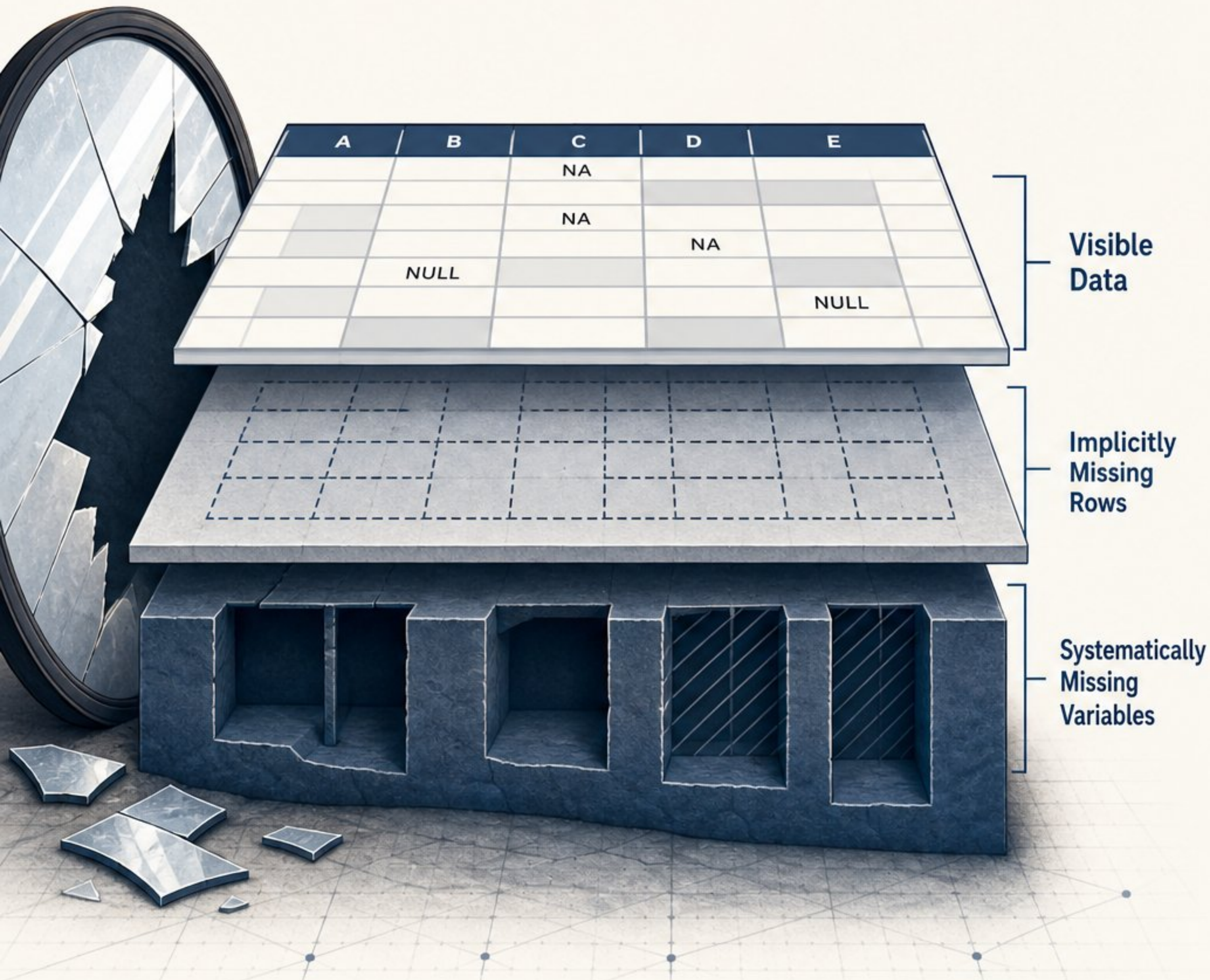
- Not Applicable: question does not apply (e.g., pregnancy status for males)

NOT RECORDED

SPECIMEN: MISSING-004

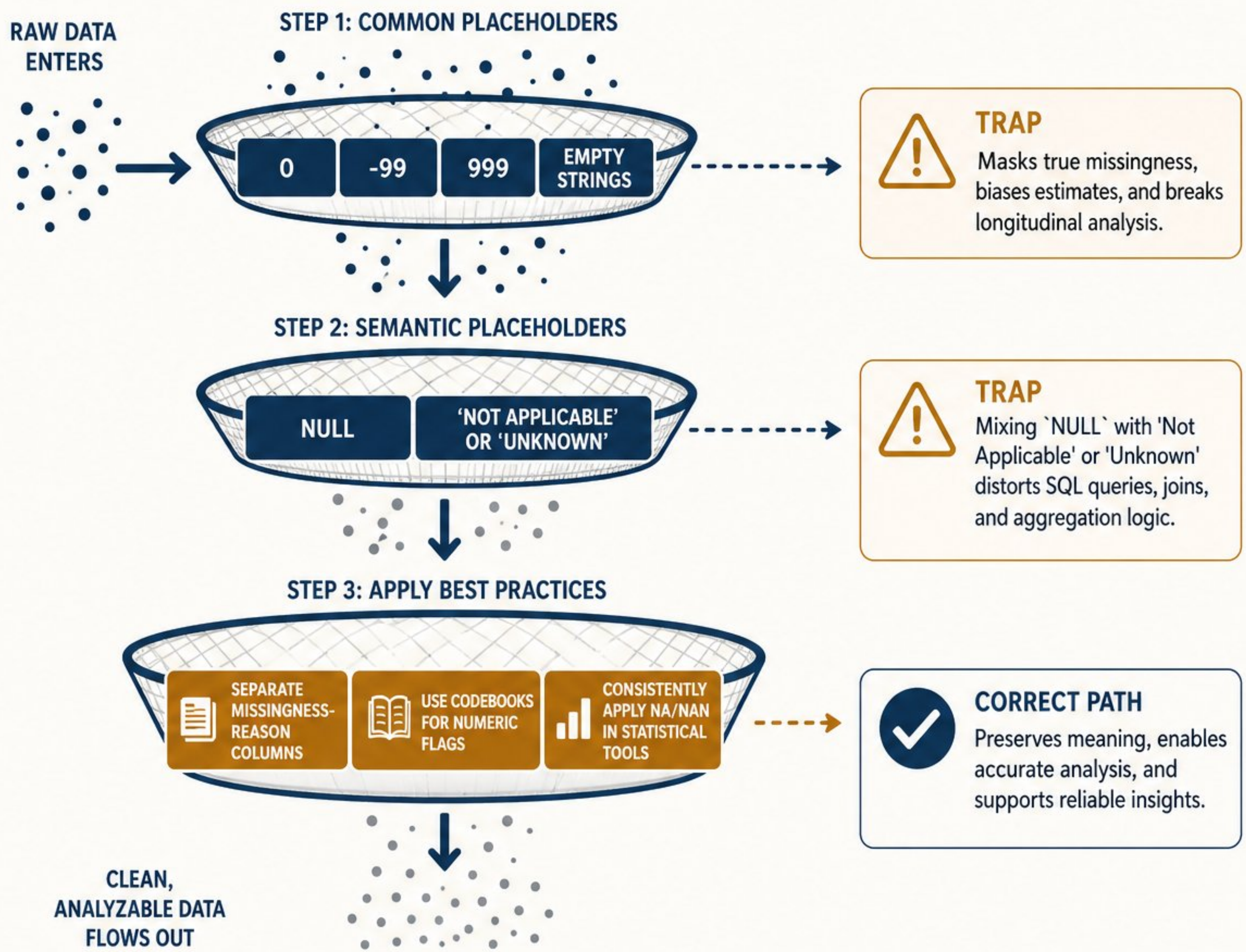
- Not Recorded: data collection intended but did not occur; temporary NULL or workflow failure

Structural vs. Sporadic Missingness: Patterns That Mislead



- **Implicitly missing rows:** combinations expected but entirely absent
- **Explicitly missing cells:** NA, NULL visible in existing rows
- **Systematically missing:** entire variables absent in some studies
- **Sporadically missing:** scattered gaps within a dataset
- **Structural gaps** (time points not applicable, unadministered modules) are often mistaken for errors

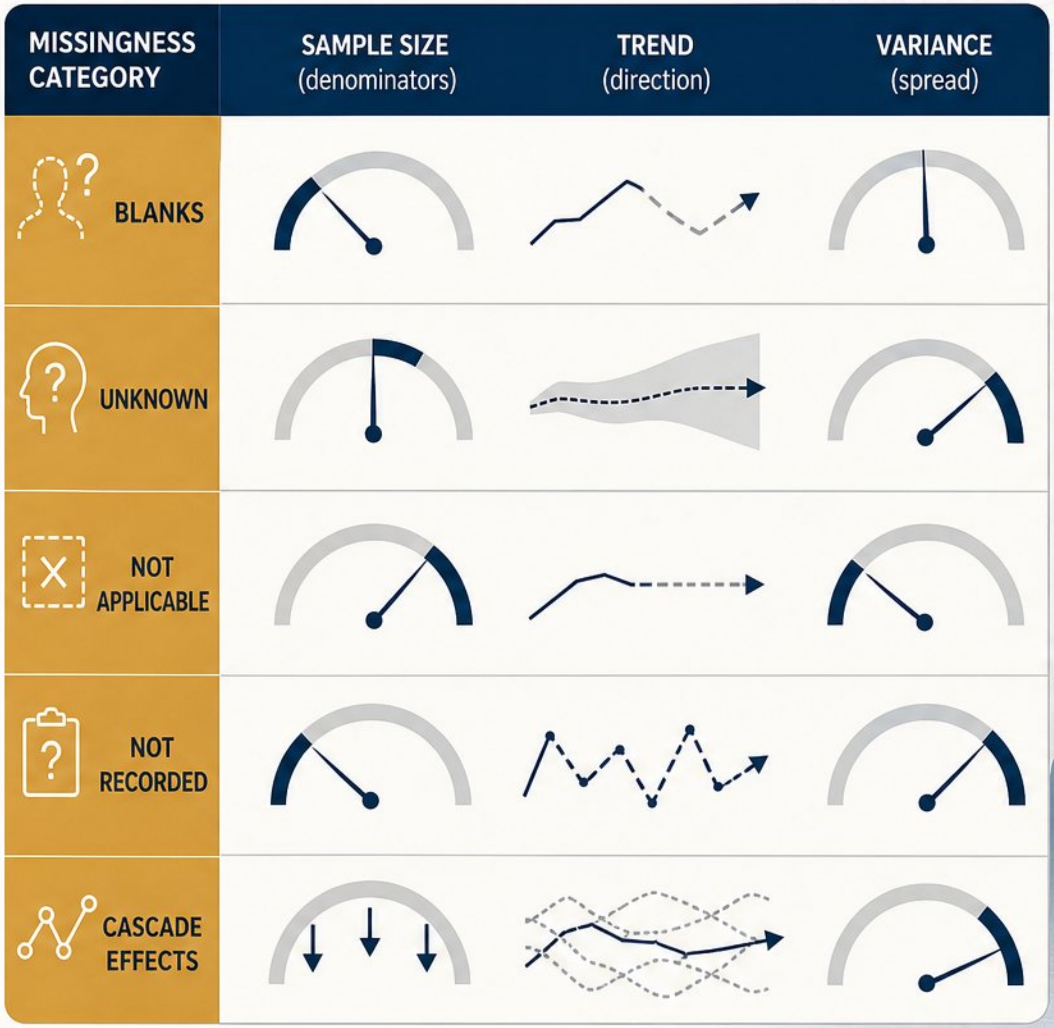
Coding Pitfalls: 0, 99, NULL, and Other Dangerous Placeholders



- 0, -99, 999, empty strings, and sentinel dates like 1900-01-01 often mask true missingness, causing biased estimates and broken longitudinal analysis.
- Mixing `NULL` with 'Not Applicable' or 'Unknown' distorts SQL queries, joins, and aggregation logic.
- Best practice: separate missingness-reason columns, use codebooks for numeric flags, and consistently apply NA/NaN in statistical tools.

Limits on Interpretation: How Each Category Changes the Story

- Blanks mask true sample size, inflating or deflating apparent denominators
- Unknown introduces irreducible uncertainty—imputation requires strong assumptions
- Not Applicable shifts denominators and invalidates subgroup comparisons
- Not Recorded hides systematic data collection failures
- Consequences cascade through averages, trends, correlations, and model performance



Spotting the Difference: From Data Profiling to Diagnosis

- Use a checklist to differentiate Blank, Unknown, N/A, and Not Recorded during EDA
- Catch red flags: unexpected blank rates, inconsistent codes, metadata mismatches
- Leverage data dictionaries, CRF skip logic, and survey routing to infer reasons
- Profile missingness patterns via frequency, cross-tabs, and gap visualization

Prevention by Design: Building Better Data Collection

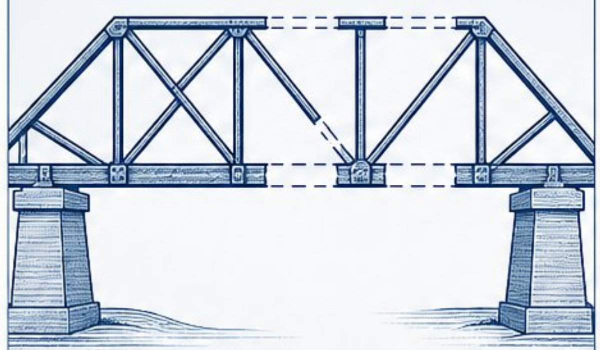
- Design forms that force meaningful absence coding: 'Not Applicable' vs 'Unknown' radio buttons.
- Implement early validation to distinguish skipped questions from non-applicable branches.
- Use split-questionnaire designs to reduce burden while maintaining analytic power.
- Track missingness rates over time with dashboards and pattern alerts.
- Standardize missing data codes across your dataset—don't leave blanks ambiguous.

Communicating Limits with Transparency

- ▶ - Adopt reporting guidelines (STROBE, PCORI) for missing data disclosure.
- ▶ - Use plain language to describe missingness mechanisms and proportions.
- ▶ - Separate true missing values from structural absences in tables.
- ▶ - Tailor summaries: executive highlights vs. analyst/ regulator detail.
- ▶ - Report sensitivity analyses to demonstrate result robustness.

TRANSPARENT DISCLOSURE

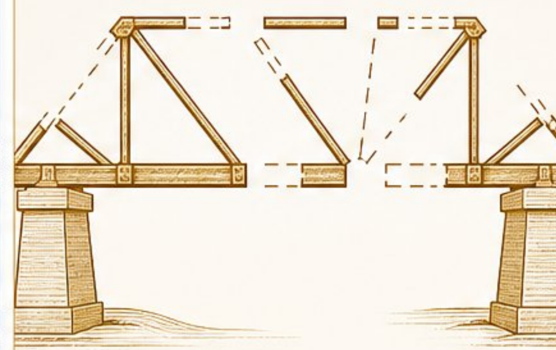
Clear about what is missing and why



- ✓ Discloses missingness mechanisms and proportions in plain language.
- ✓ Distinguishes true missing values from structural absences.
- ✓ Provides tailored summaries for different audiences.
- ✓ Includes sensitivity analyses to support robustness.
- ✓ Builds trust through clarity and completeness.

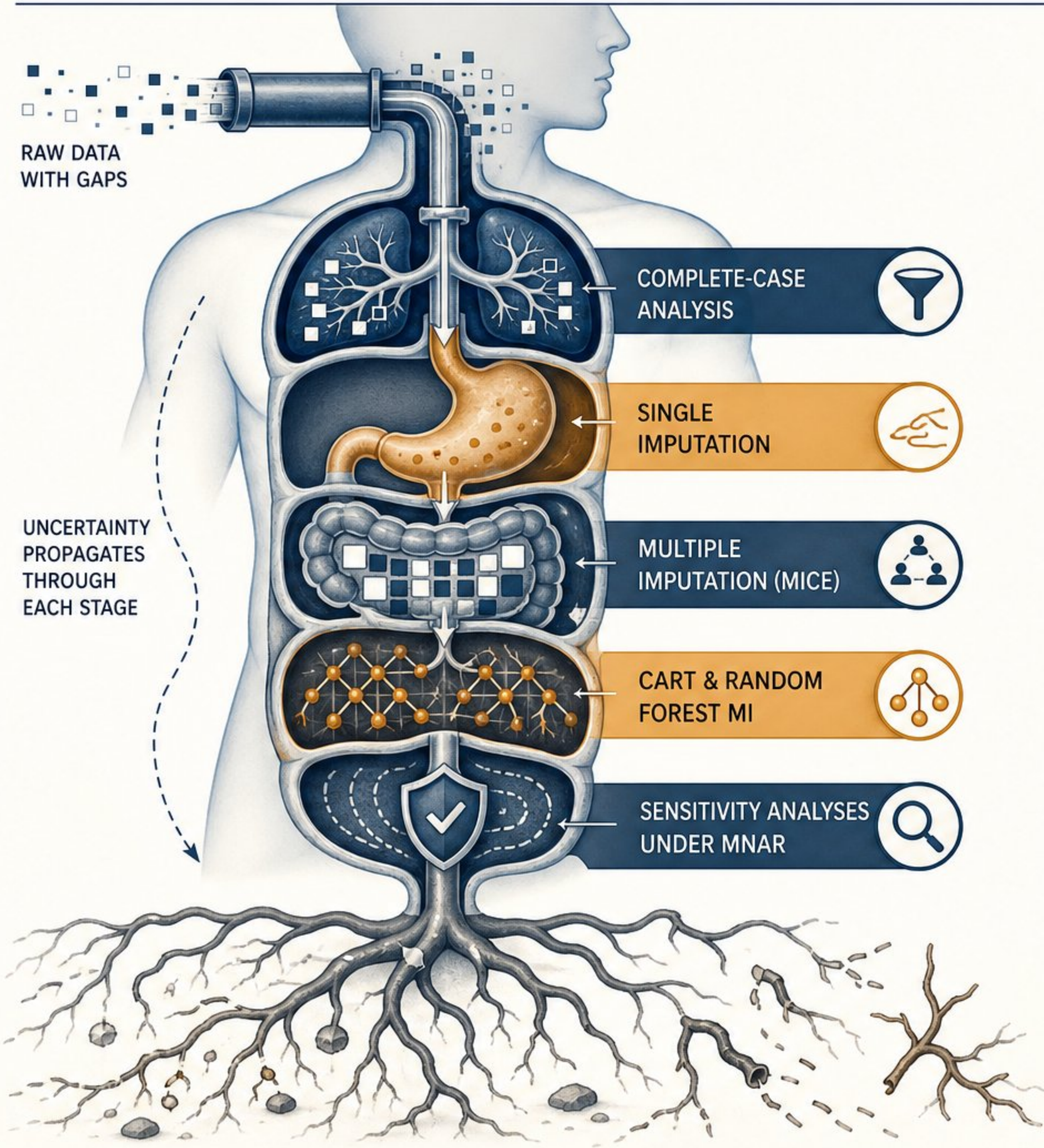
OPAQUE DISCLOSURE

Vague, incomplete, and hard to interpret.



- ✗ Uses jargon; unclear about what is missing and why.
- ✗ Does not separate true missing from structural absences.
- ✗ One-size-fits-all reporting hides important detail.
- ✗ No sensitivity analyses reported.
- ✗ Undermines confidence and accountability.

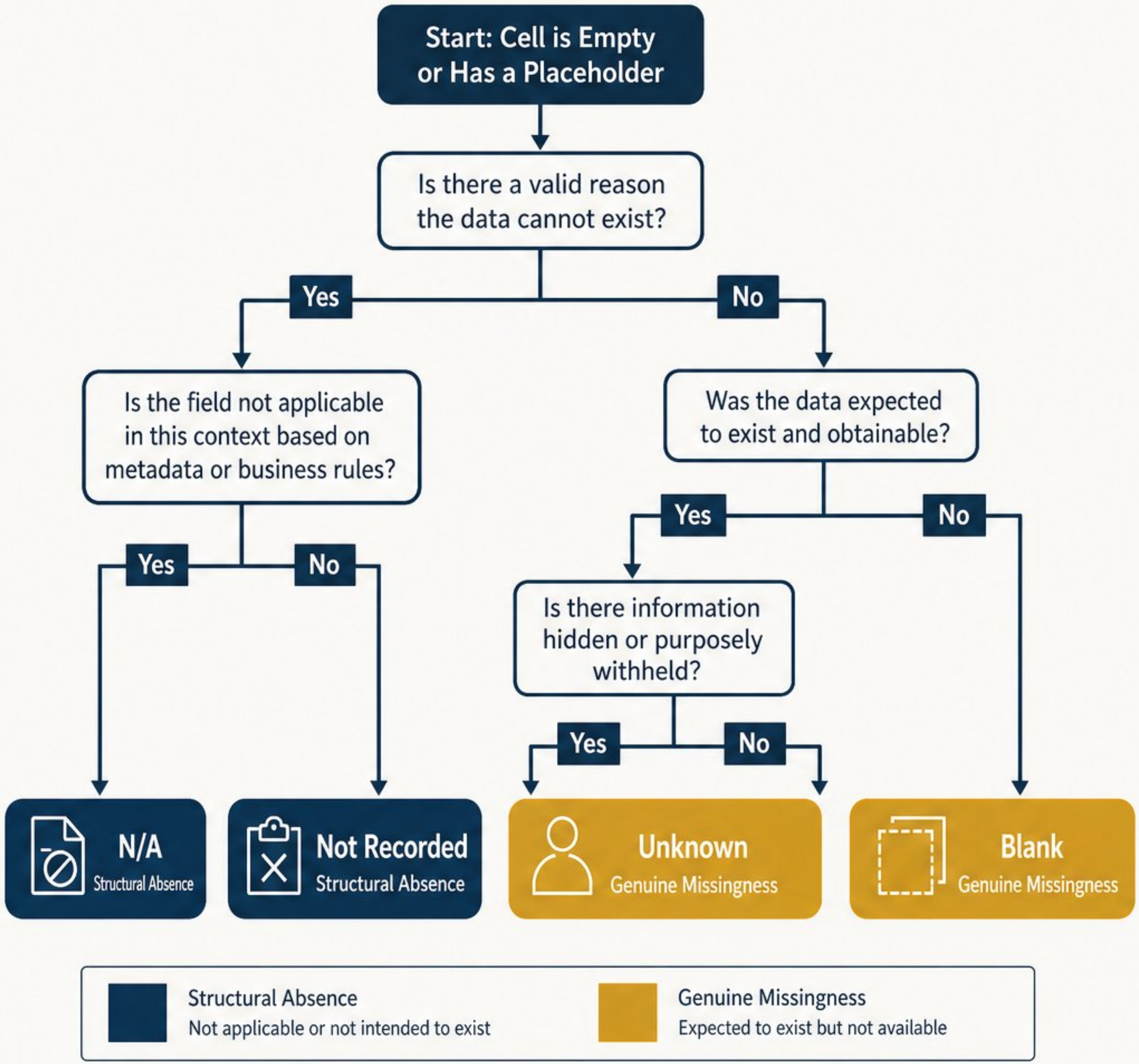
Handling Missing Data in Analysis: Overview of Strategies



- Complete-case analysis is biased unless data are MCAR
- Single imputation (mean, KNN, LOCF) underestimates uncertainty
- Multiple imputation (MICE) is preferred under MAR
- CART and random forest MI offer robust, stable performance
- Always add sensitivity analyses under MNAR assumptions

Exercise: Classifying Missingness in a Realistic Dataset

- Classify blanks, 'Unknown', 'N/A', and 'Not Recorded' using metadata context
- Apply a decision tree to distinguish structural absence from genuine missingness
- Discuss how each classification shifts denominators and usable records
- Identify which missingness types are hardest to distinguish and why



Key Takeaways and Your Next Steps

- Blank, Unknown, Not Applicable, and Not Recorded each impose distinct interpretation limits
- Missingness is a design, documentation, and communication challenge, not just a data problem
- Use the one-page checklist: profiling questions, coding standards, reporting templates, prevention tips
- Tomorrow: audit one dataset, propose missing-data standards, and advocate for a missing-data protocol