Data Preprocessing Pipeline for Innovation Clustering
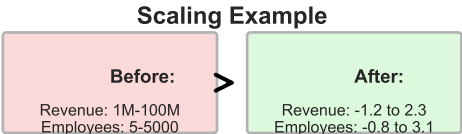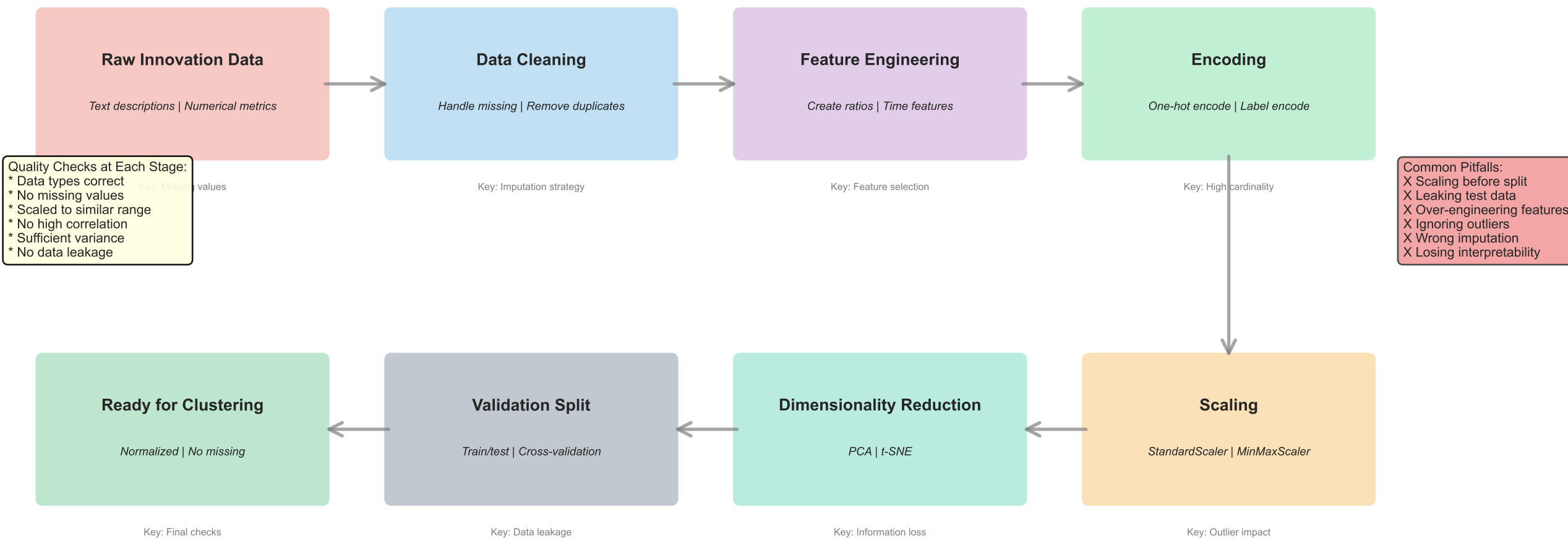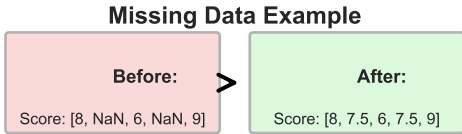
From Raw Data to Clustering-Ready Features
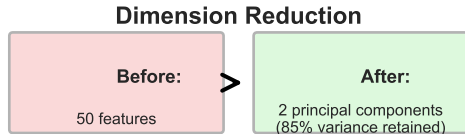
Raw Innovation Data

Text descriptions | Numerical metrics

Key: Missing values

Data Cleaning

Handle missing | Remove duplicates

Key: Imputation strategy

Feature Engineering

Create ratios | Time features

Key: Feature selection

Encoding

One-hot encode | Label encode

Key: High cardinality

Quality Checks at Each Stage:
* Data types correct
* No missing values
* Scaled to similar range
* No high correlation
* Sufficient variance
* No data leakage

Common Pitfalls:
X Scaling before split
X Leaking test data
X Over-engineering features
X Ignoring outliers
X Wrong imputation
X Losing interpretability

Ready for Clustering

Normalized | No missing

Key: Final checks

Validation Split

Train/test | Cross-validation

Key: Data leakage

Dimensionality Reduction

PCA | t-SNE

Key: Information loss

Scaling

StandardScaler | MinMaxScaler

Key: Outlier impact

Scaling Example

Before:

Revenue: 1M-100M
Employees: 5-5000

After:

Revenue: -1.2 to 2.3
Employees: -0.8 to 3.1

(StandardScaler)

Encoding Example

Before:

Type: [Product, Service, Platform]

After:

Type_Product: [1,0,0]
Type_Service: [0,1,0]

(One-Hot Encoding)

Missing Data Example

Before:

Score: [8, NaN, 6, NaN, 9]

After:

Score: [8, 7.5, 6, 7.5, 9]

(Mean Imputation)

Dimension Reduction

Before:

50 features

After:

2 principal components
(85% variance retained)

(PCA)

```
from sklearn.pipeline import Pipeline
from sklearn.preprocessing import StandardScaler
from sklearn.decomposition import PCA

pipeline = Pipeline([
    ('scaler', StandardScaler()),
    ('pca', PCA(n_components=2)),
    ('clustering', KMeans(n_clusters=5))
])
```

Typical Time: 60-80% of project