DATA ISSUES

Messy Data
Noisy data
Biased Data
Imbalanced Data
Correlated Data

Merged data can result in these ↗

- Mixed languages, incorrect spelling
- Different scales / units (length: miles, km mixed)

Mixed types - different modalities mixed into the same columns → separate features

- Classification problem can have response var imbalanced, ~~to avoid~~ skews the data set
  - sample bias
  - outliers

- Highly correlated features can cause collinearity problems and numerical instability