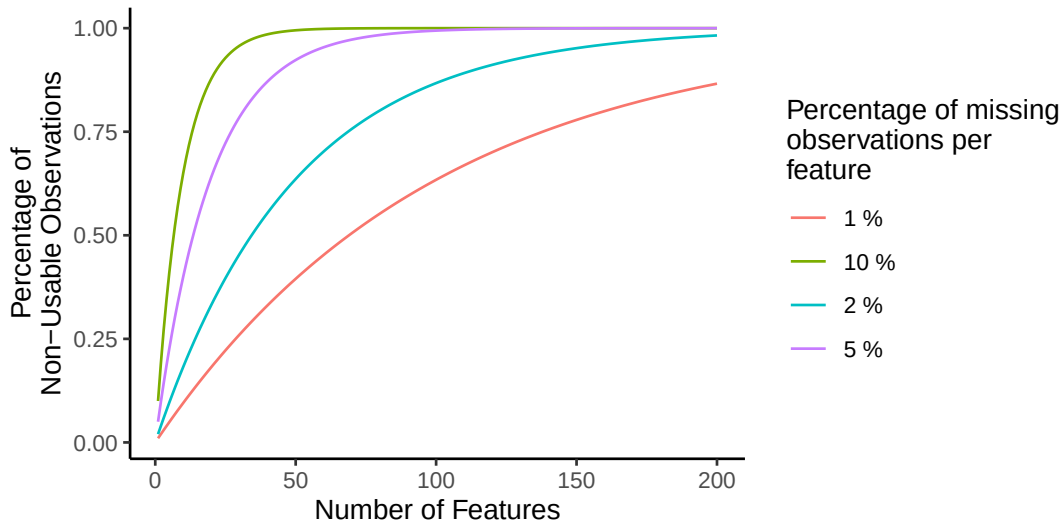
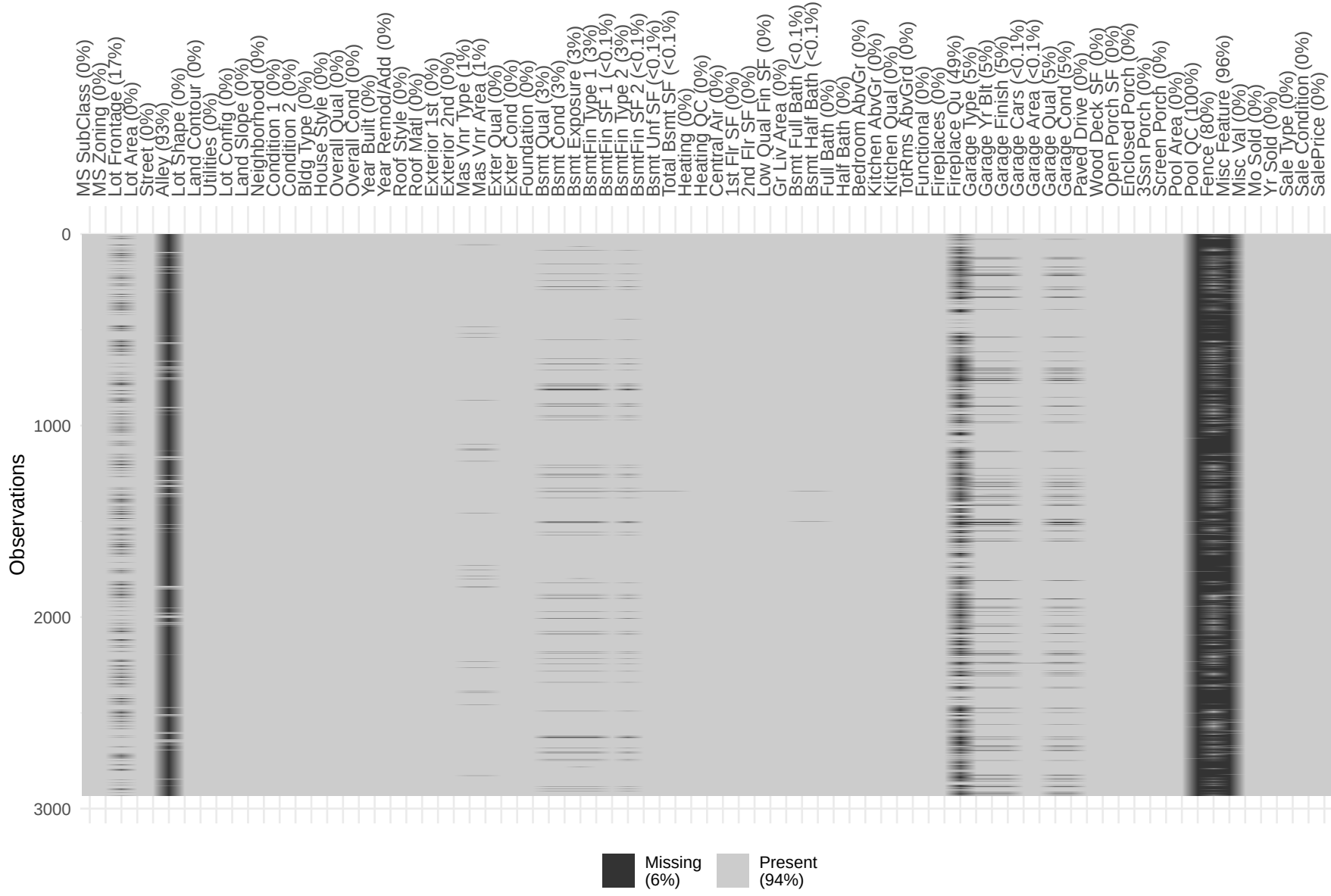
MOTIVATING EXAMPLE

- Assume each feature in your dataset has 2 % missing values.
- The missing values are randomly distributed over the observations.
- How many rows can be used if all observations that contain at least a missing value is dropped?



With 100 features and 2 % missing values only 13 % of our data can be used.

VISUALIZING MISSING VALUES



POSSIBLE WAYS TO DEAL WITH MISSING VALUES

- Remove observations that contain missing values.
  But: Could lead to a very small dataset.
- Remove features that contain mostly missing values.
  But: Can lose (important) information.
- Use models that can handle missing values, e.g., (most) tree-based methods
  But: Restriction in model choice.
- **Imputation**
  $\rightarrow$ Replace missing values with *plausible* values.