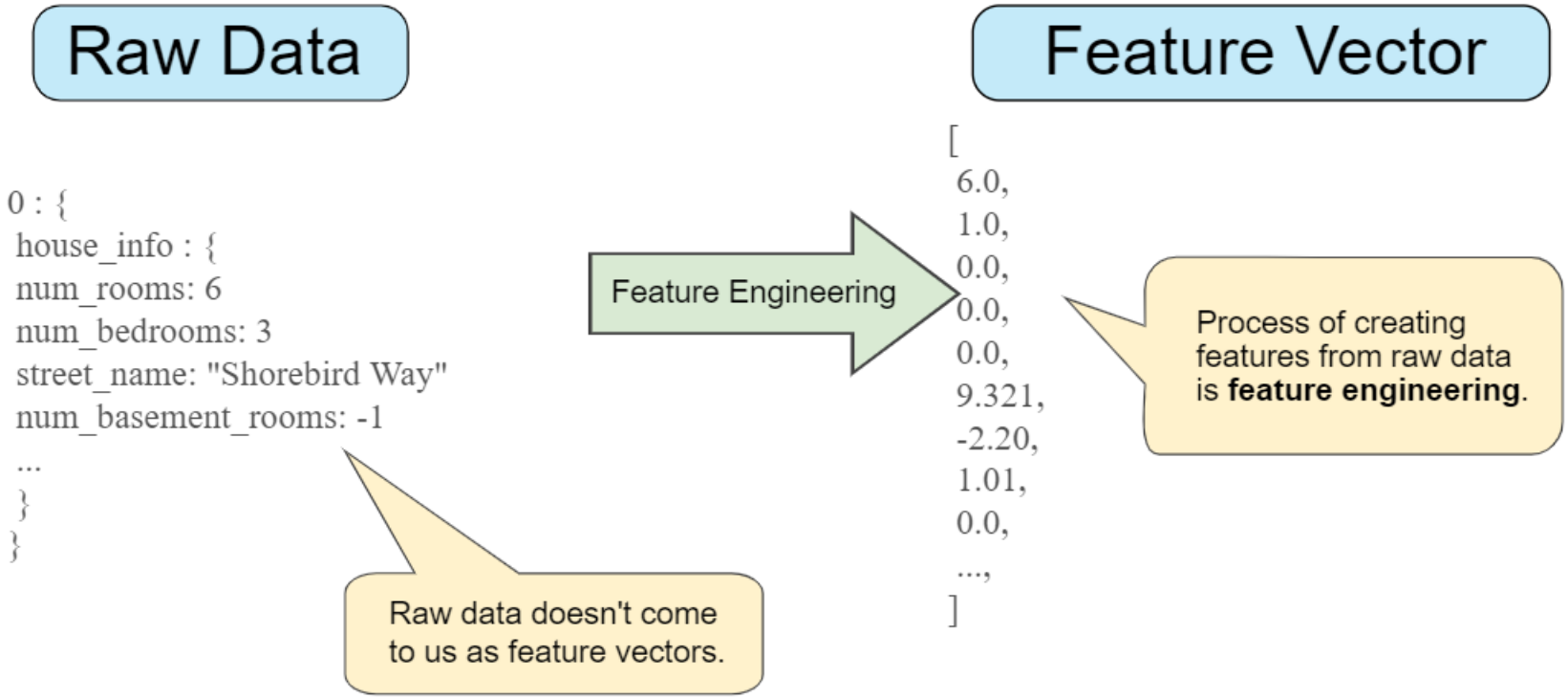
Representation

How to get feature vectors out of raw data

(Source: https://developers.google.com/machine-learning/crash-course/representation/qualities-of-good-features

Feature Engineering: Mapping Raw Data to Features



Mapping Categorical Values

Categorical features have a ***discrete*** (finite) set of possible values. For example, there might be a feature called `faculty` with options that include:

Simple Solution:

{‘Betriebswirtschaft’, ‘Soziale Arbeit’, ‘Informatik’, ‘Maschinenbau’, ‘E-Technik’}

{ 0 , 1 , 2 , 3 , 4 }

Raw Data

Feature

faculty: „Betriebswirtschaft”

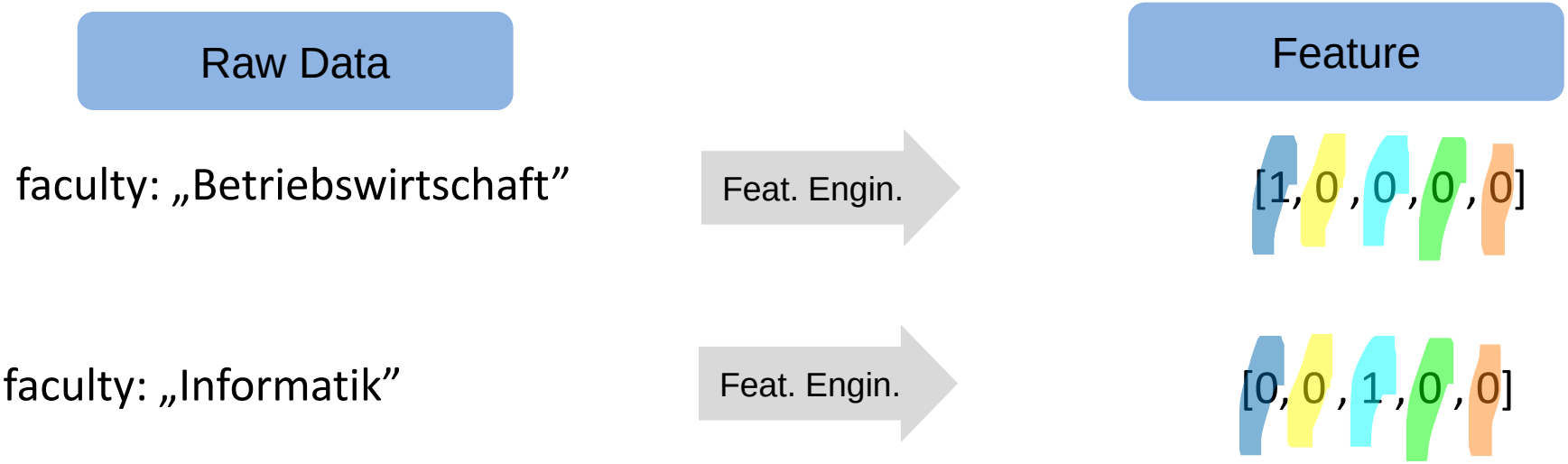
Feat. Engin.

0

Problems

- How to deal with multiple values?
- Impact of values in a (pseudo)-linear Model?

Mapping Categorical Values: **One Hot Encoding**



{‘Betriebswirtschaft’, ‘Soziale Arbeit’, ‘Informatik’, ‘Maschinenbau’, ‘E-Technik’}

Tips for good features

- Avoid rarely used discrete feature values
- Clear and obvious meanings
- Don't mix "magic" values with actual data
- Account for upstream instability

Cleaning feature values: Know your data

- *Cleaning („Scrubbing“)*, i.e.
  - handle missing values
  - remove duplicate examples
  - remove false labels
- *Outlier Handling* (Logarithmic Scaling)
- *Binning* (if there is obviosuly no linear relationship)
- *Scaling*