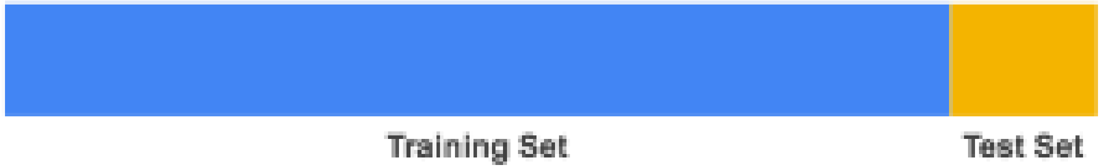
DATA PROCESSING

Narges Norouzi

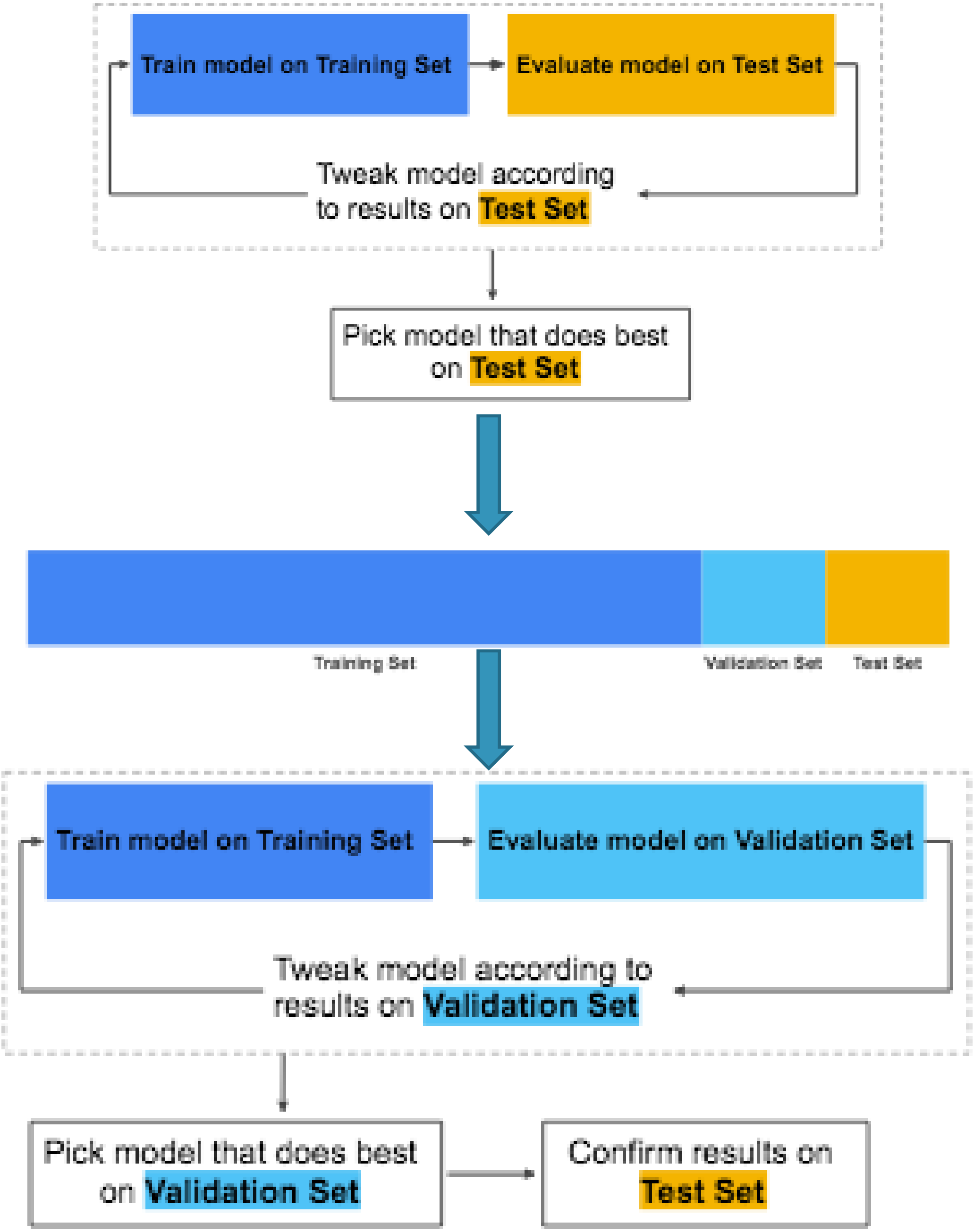
WHAT TO DO WITH ONLY ONE DATASET?

- Divide the data into two sets:
  - Test data
  - Training data
    - Training data will then be spitted into a training set and a validation set
  - Make sure to randomize the data before splitting
- DO NOT TRAIN ON THE TEST DATA
  - Getting surprisingly low loss?



ANOTHER PARTITION

- Note about test data:
  - Should be large enough to yield statistically significant results
  - Should be representative of the data as a whole



Train model on Training Set

Evaluate model on Test Set

Tweak model according to results on Test Set

Pick model that does best on Test Set

Training Set

Validation Set

Test Set

Train model on Training Set

Evaluate model on Validation Set

Tweak model according to results on Validation Set

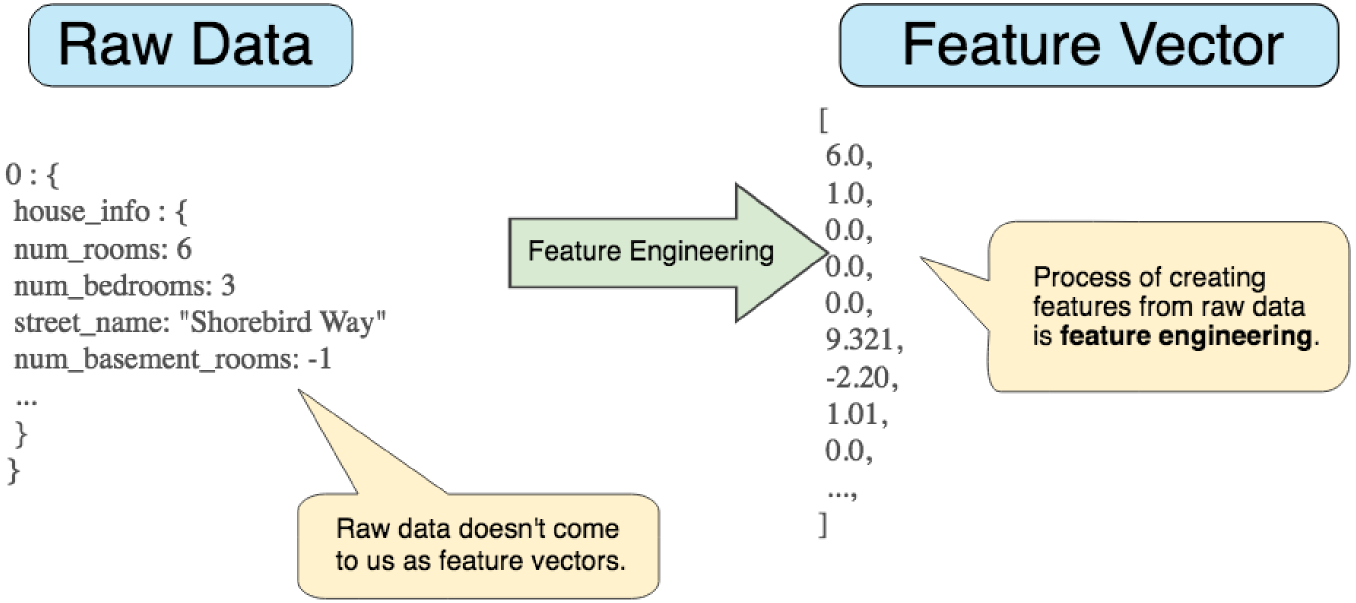
Pick model that does best on Validation Set

Confirm results on Test Set

MAJOR TASKS IN DATA PREPARATION

- Data cleaning
  - Fill in missing values, smooth noisy data, identify and remove outliers
- Data integration
  - Combining multiple sources of data
- Data transformation
  - Normalization or aggregation – min-max normalization or z-score normalization
- Data reduction
  - Obtains reduced representation in volume but produces the same or similar analytical results

MAPPING RAW DATA TO FEATURES



Raw Data

```
0 : {
 house_info : {
 num_rooms: 6
 num_bedrooms: 3
 street_name: "Shorebird Way"
 num_basement_rooms: -1
 ...
 }
}
```

Raw data doesn't come to us as feature vectors.

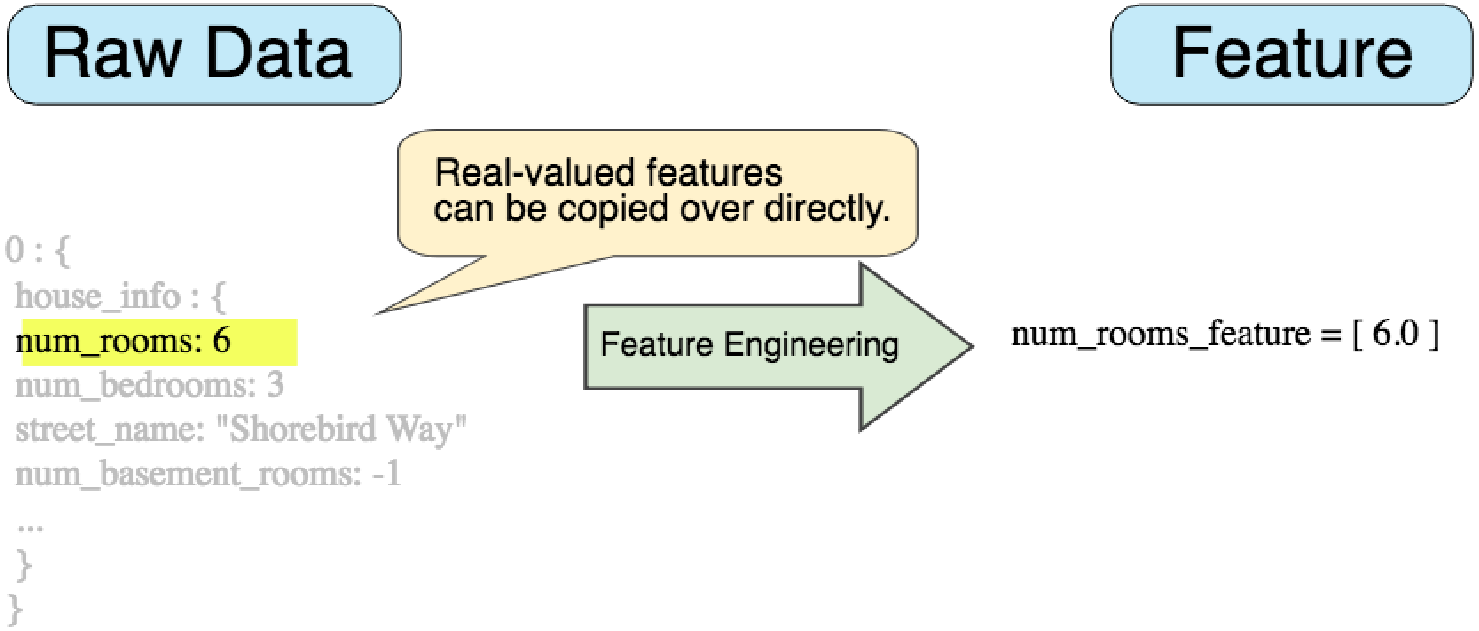
Feature Engineering

Feature Vector

```
[
 6.0,
 1.0,
 0.0,
 0.0,
 0.0,
 9.321,
 -2.20,
 1.01,
 0.0,
 ...,
]
```

Process of creating features from raw data is **feature engineering**.

MAPPING NUMERIC VALUES



Raw Data

Feature

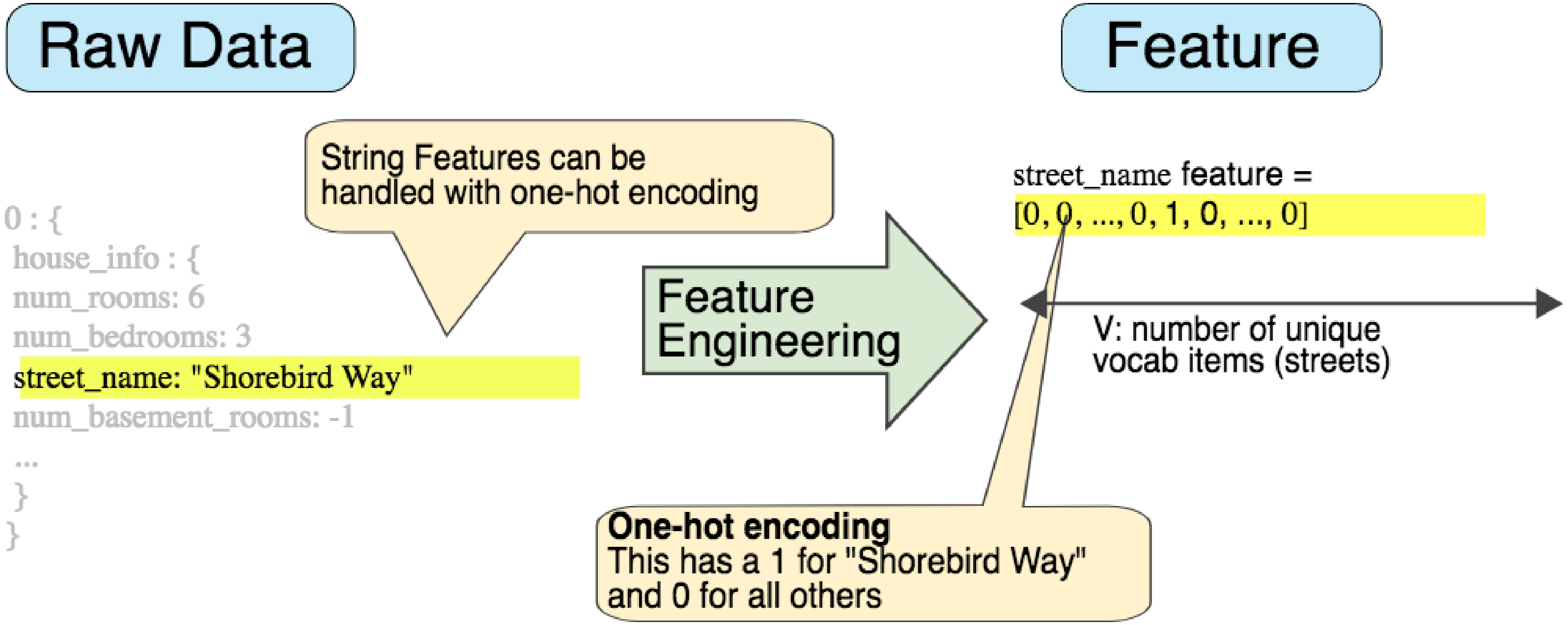
Real-valued features can be copied over directly.

```
0 : {
 house_info : {
 num_rooms: 6
 num_bedrooms: 3
 street_name: "Shorebird Way"
 num_basement_rooms: -1
 ...
 }
}
```

Feature Engineering

```
num_rooms_feature = [ 6.0 ]
```

MAPPING CATEGORICAL VALUES



Raw Data

Feature

String Features can be handled with one-hot encoding

```
0 : {
 house_info : {
 num_rooms: 6
 num_bedrooms: 3
 street_name: "Shorebird Way"
 num_basement_rooms: -1
 ...
 }
}
```

street_name feature =
[0, 0, ..., 0, 1, 0, ..., 0]

V: number of unique vocab items (streets)

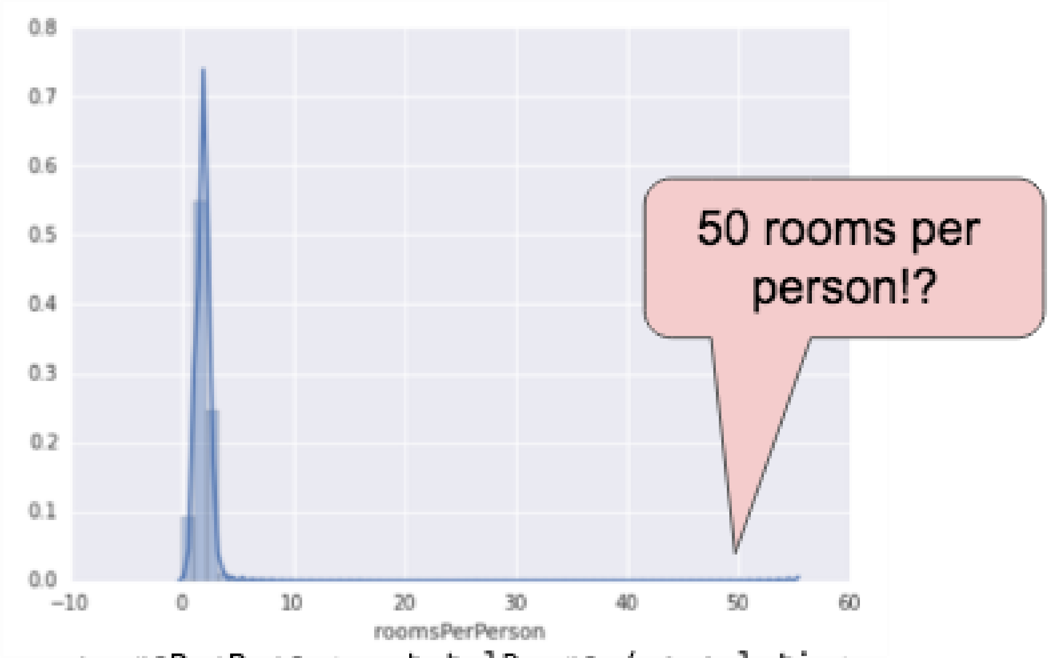
One-hot encoding
This has a 1 for "Shorebird Way" and 0 for all others

NOTES ABOUT A GOOD FEATURE

- Avoid rarely-used discrete values
  - House type vs. unique house id
- Prefer clear and obvious meanings
  - User age: 23 or 1234556
- Don’t mix magic values with actual data
  - Watch time: -1
  - Use indicator value to account for undefined values
- Shouldn’t change over time (stationarity)
  - Happens when we connect multiple models with different definitions

QUALITIES OF A GOOD FEATURE

- Should not have extreme outliers

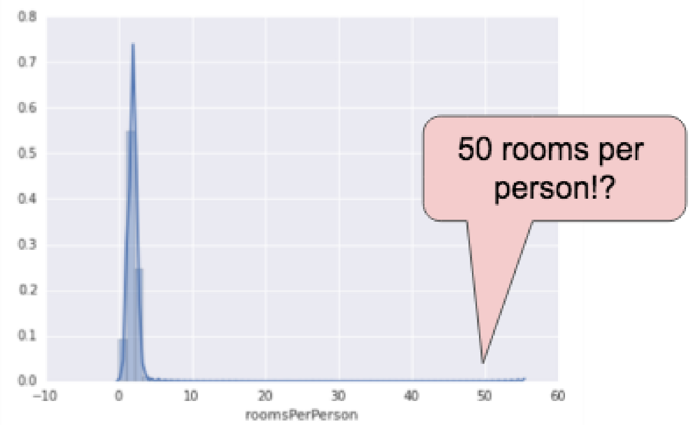


50 rooms per person!?

```
roomsPerPerson = totalRooms / population
```

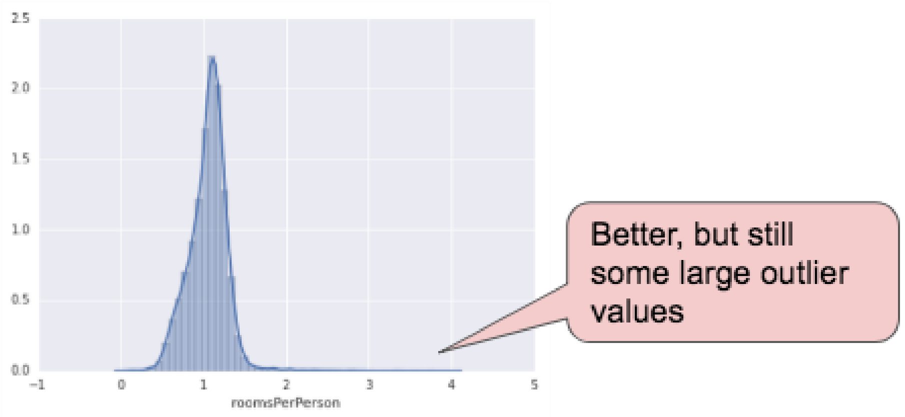
CLEANING DATA (1)

- Handling extreme outliers
  - Changing the scale – log or exponential
  - Capping or clipping the data



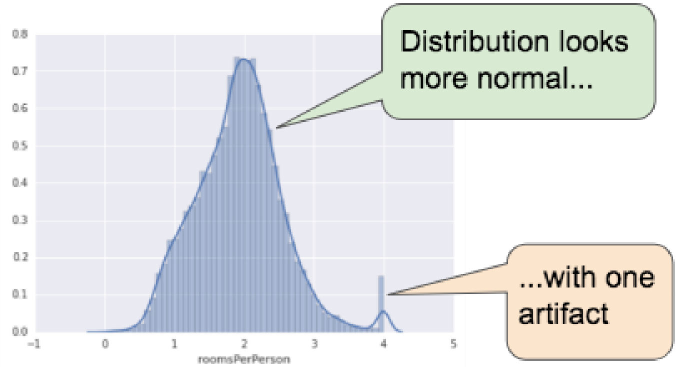
50 rooms per person!?

`roomsPerPerson = totalRooms / population`



Better, but still some large outlier values

`roomsPerPerson = log((totalRooms / population) + 1)`



Distribution looks more normal...

...with one artifact

`roomsPerPerson = min(totalRooms / population, 4)`

CLEANING DATA (2)

- Scrubbing
  - Omitted values: For instance, a person forgot to enter a value for a house's age.
  - Duplicate examples: For example, a server mistakenly uploaded the same logs twice.
  - Bad labels: For instance, a person mislabeled a picture of an oak tree as a maple.
  - Bad feature values: For example, someone typed in an extra digit, or a thermometer was left out in the sun.

CLASS EXERCISE

bit.ly/ce-3

ENCODING NON-LINEARITY: FEATURE CROSSES

- Feature cross is a synthetic feature that encodes non-linearity
- Kinds of feature crosses:
  - $[A \times B]$: a feature cross formed by multiplying the values of two features.
  - $[A \times B \times C \times D \times E]$: a feature cross formed by multiplying the values of five features.
  - $[A \times A]$: a feature cross formed by squaring a single feature.

PLAYGROUND EXERCISE

bit.ly/featurecross

QUESTIONS?