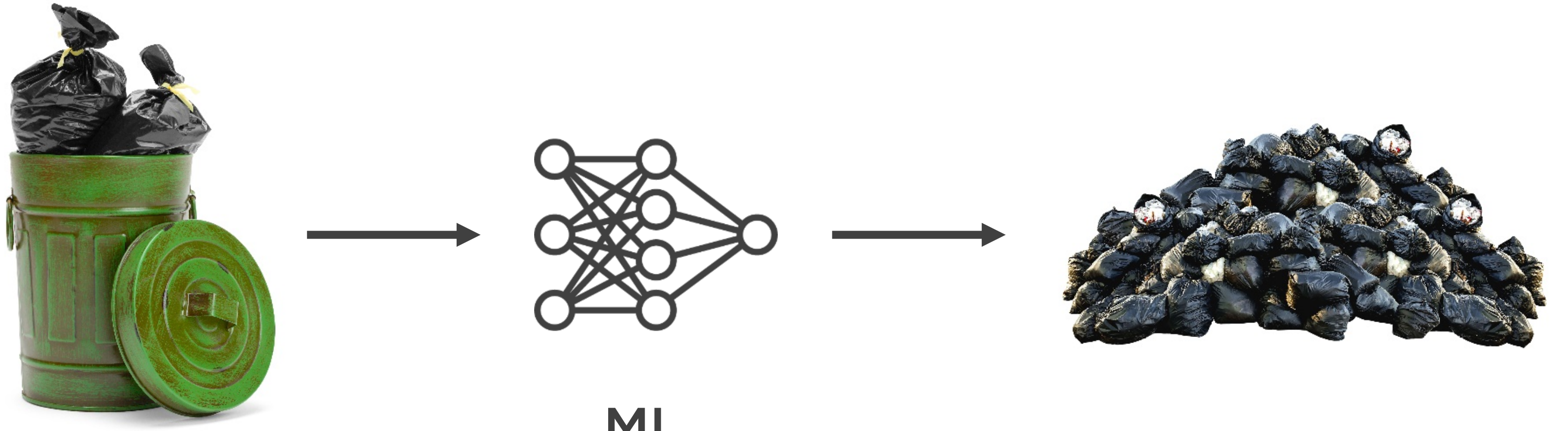
Preparing Input Data for Machine Learning Models

Ravikiran Srinivasulu
SOFTWARE CONSULTANT

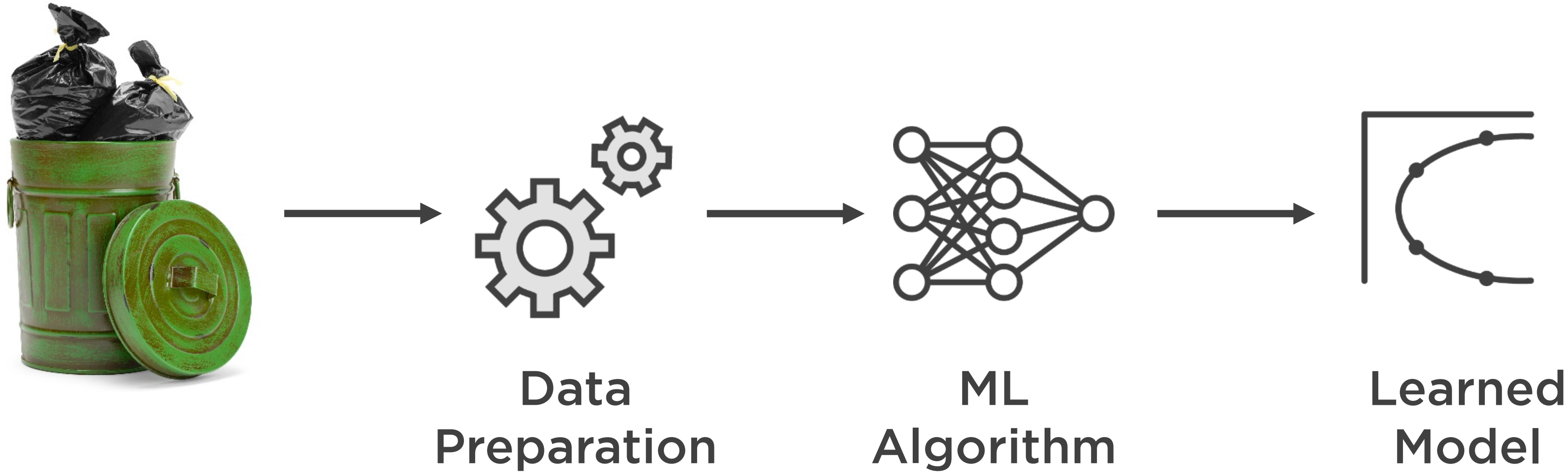
ravikirans.com | ravikirans.com/YouTube

Garbage In, Garbage Out



ML
Algorithm

Garbage In, Garbage Out



Data Preparation

ML Algorithm

Learned Model

Agenda

Exploratory Data Analysis (EDA)

Uncover data issues

- Erroneous data
- Outliers
- Duplicate records
- ...

Clean dataset ready for ML

Data Preprocessing Methods

Data Preprocessing Methods

Data Cleaning

Missing values,

Noisy data,

Outliers

Data Transformation

Normalization

Data Discretization

Binning Methods

Data Reduction

Sampling

Demo

Run Exploratory Data Analysis (EDA)

Demo

Cleaning erroneous data

Demo

Handling Outliers in dataset

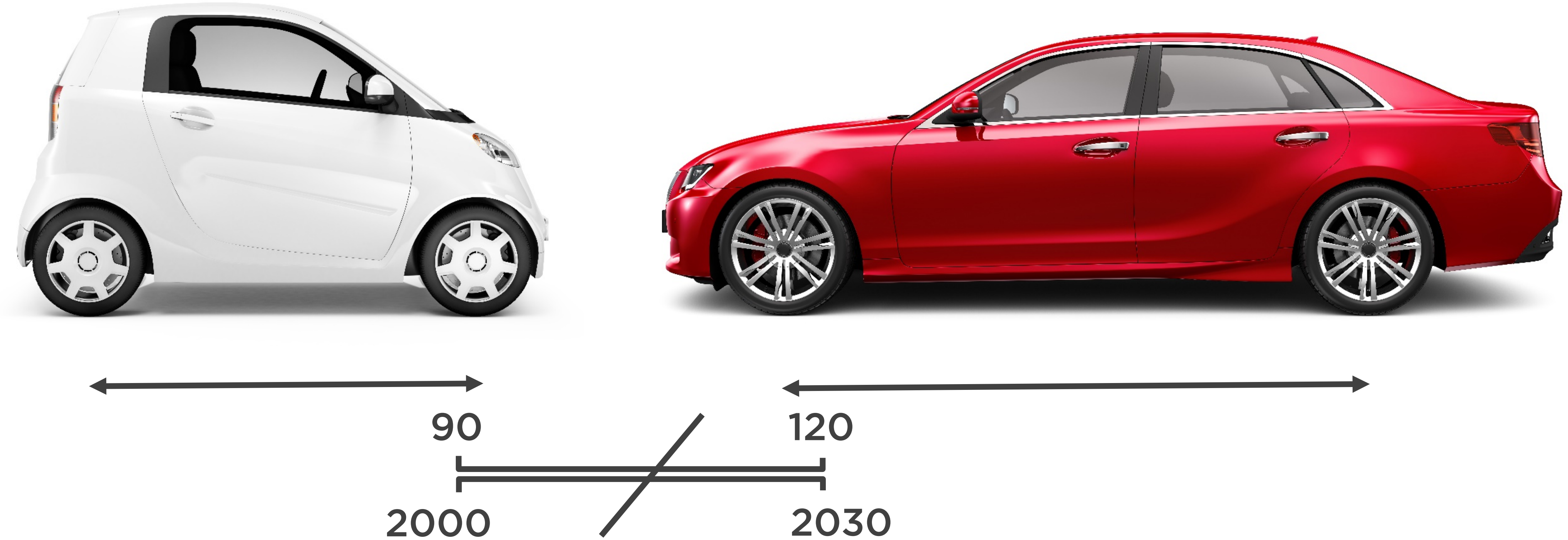
Demo

Remove duplicate records

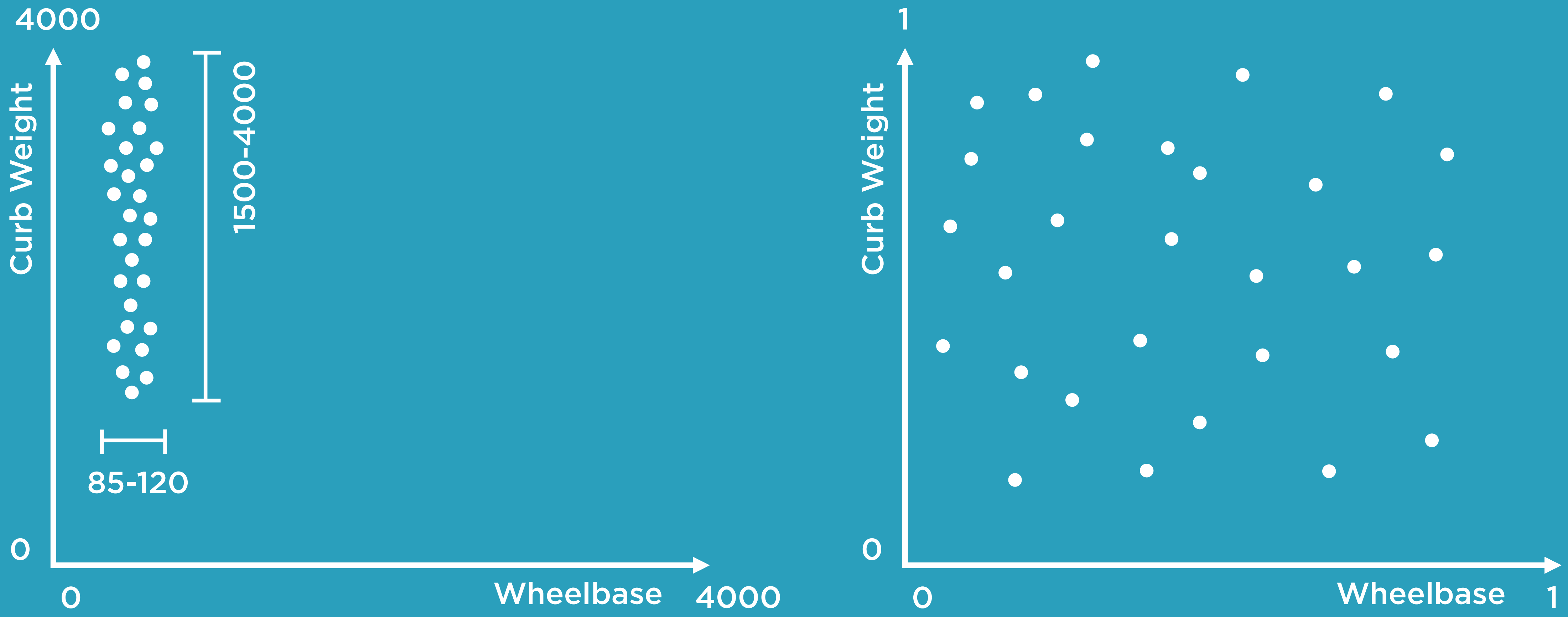
Demo

Data Transformation - Normalization

Comparison of Features

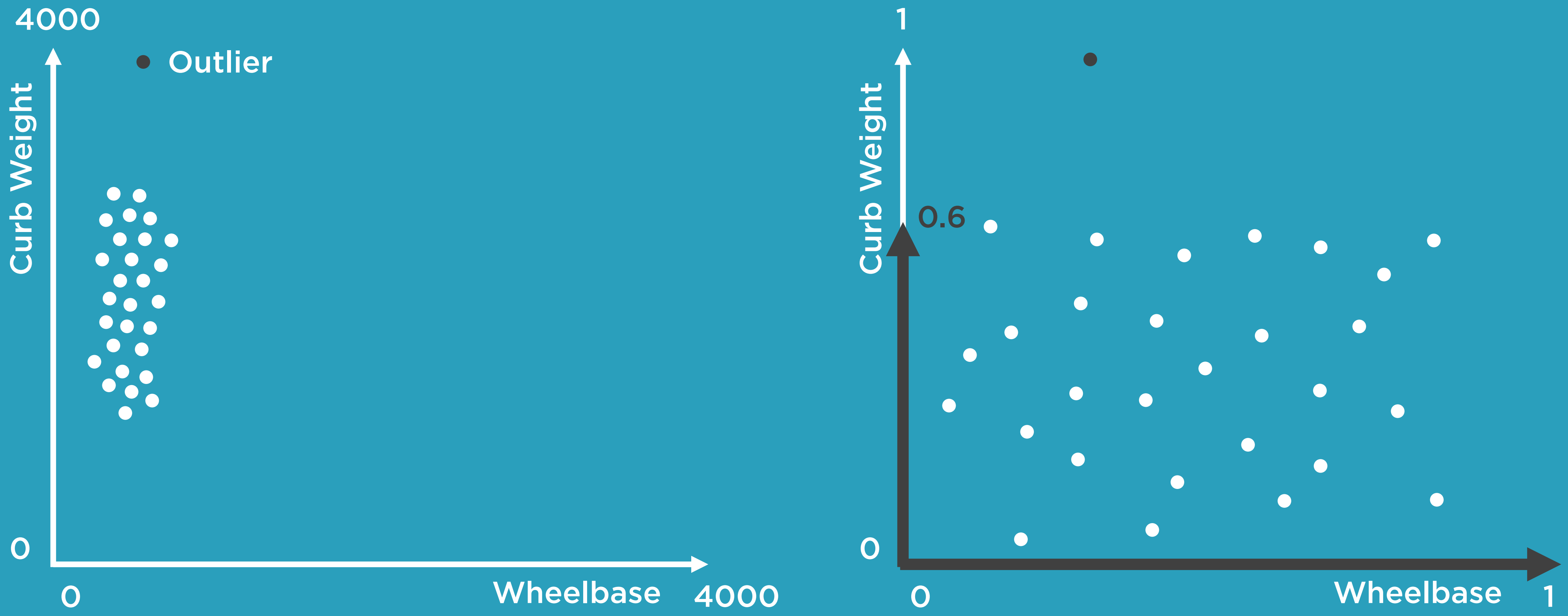


Min-Max Normalization



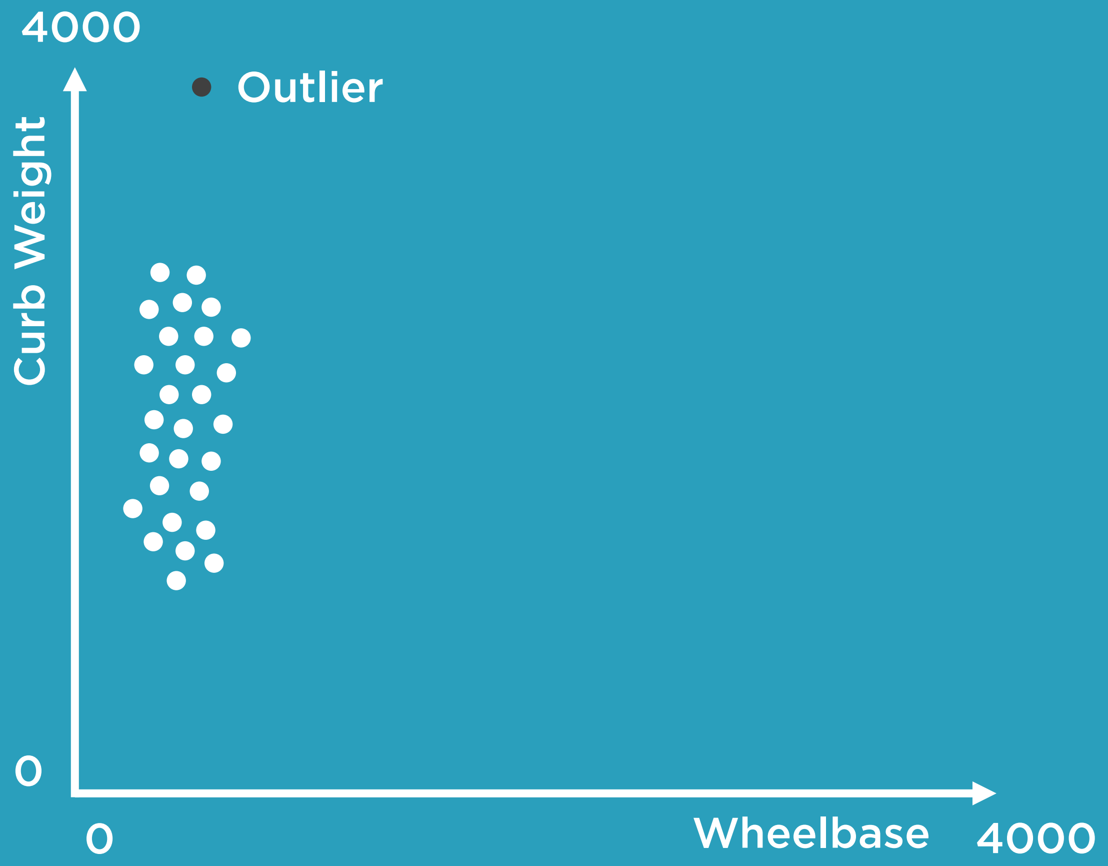
Not to scale

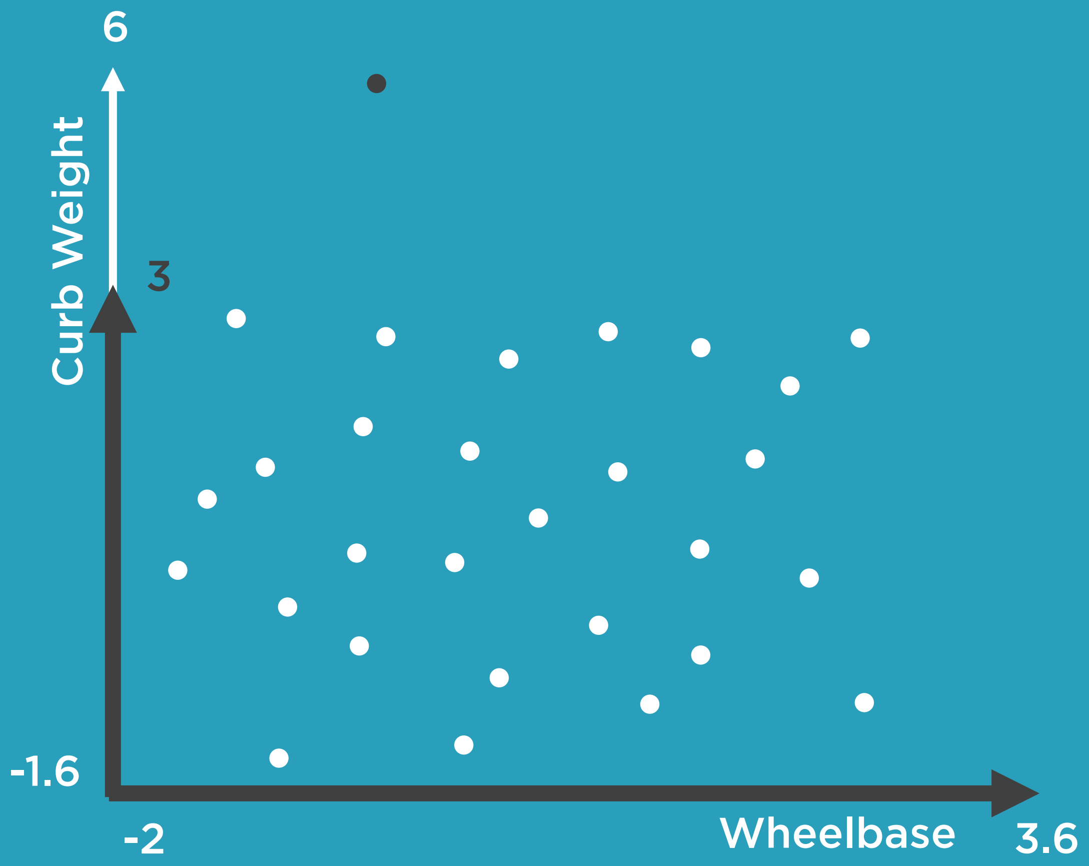
Min-Max Normalization



Not to scale

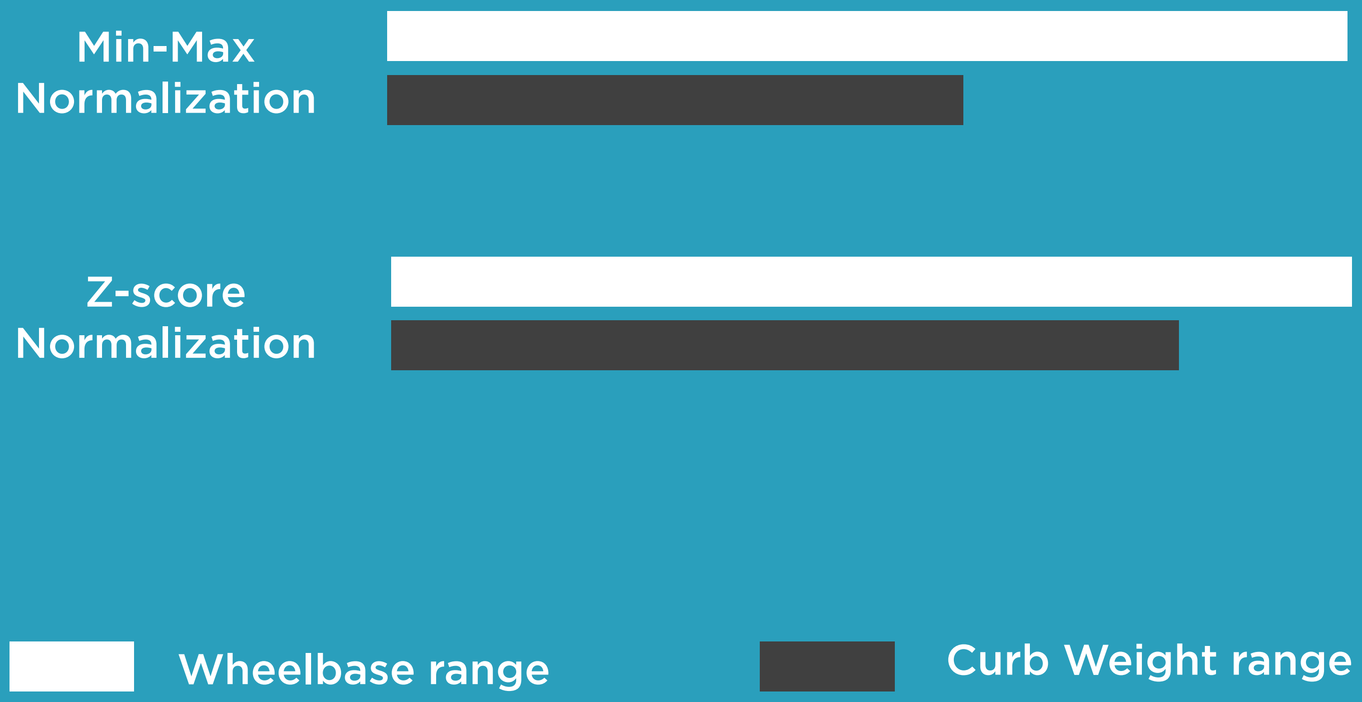
Z-Score Normalization





Not to scale

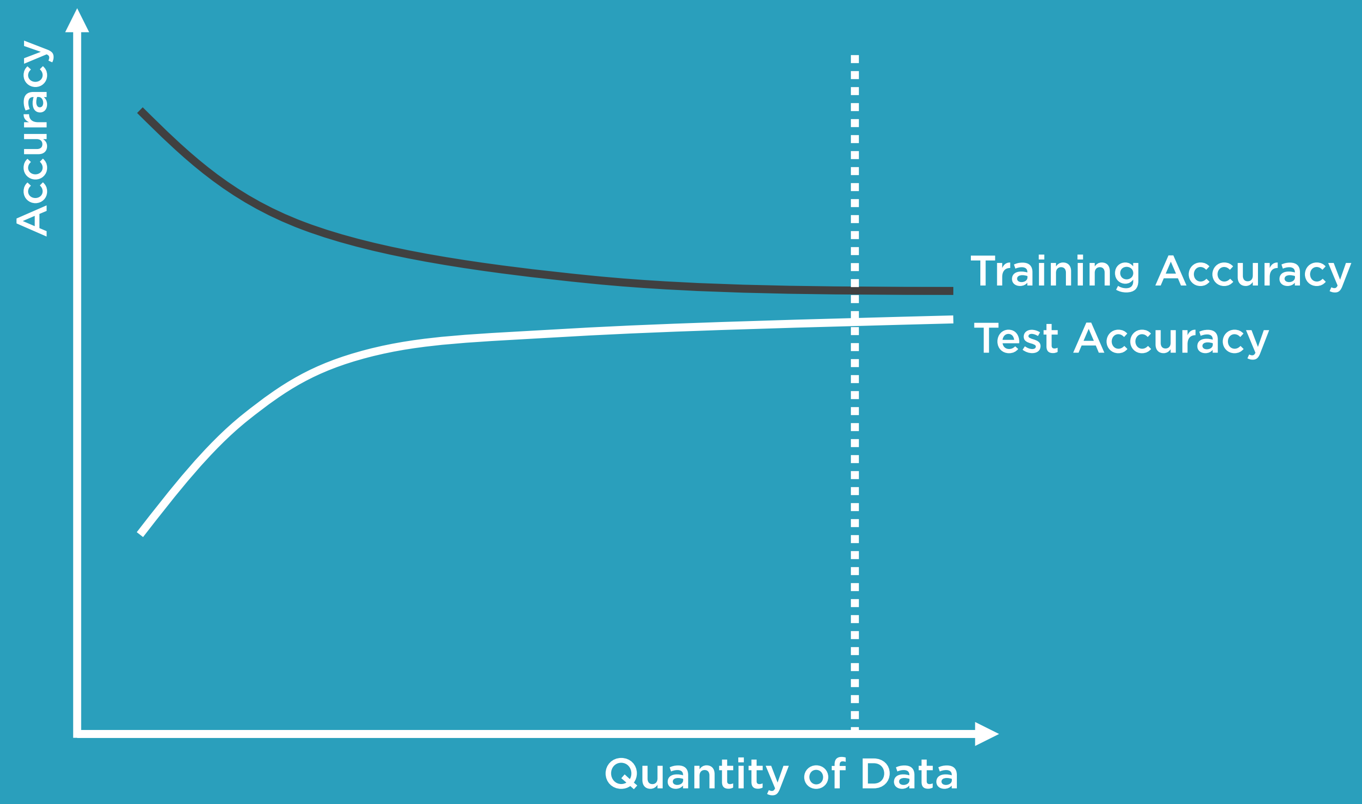
Min-Max vs. Z-Score Normalization



Demo

Let’s Sample records

How ML Algorithms Perform with Data?



Demo

Let’s select relevant columns in dataset

Demo

Data Discretization - Binning

Entropy-based Discretization

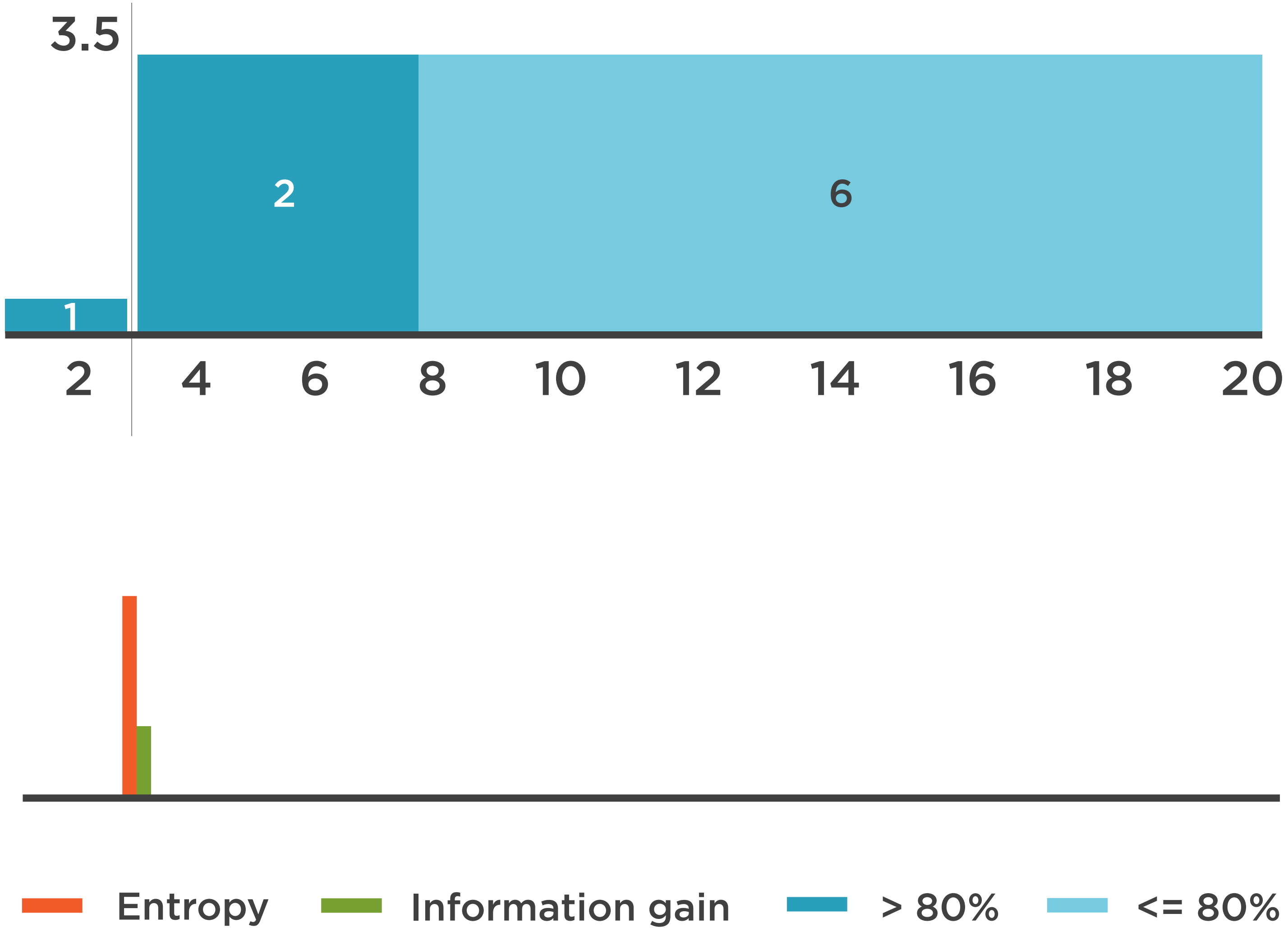
Entropy

A measure of randomness in the data

Entropy-based Discretization

Facebook hours	> 80%
3	1
4	1
5	0
7	0
10	0
11	0
14	1
15	0
19	0

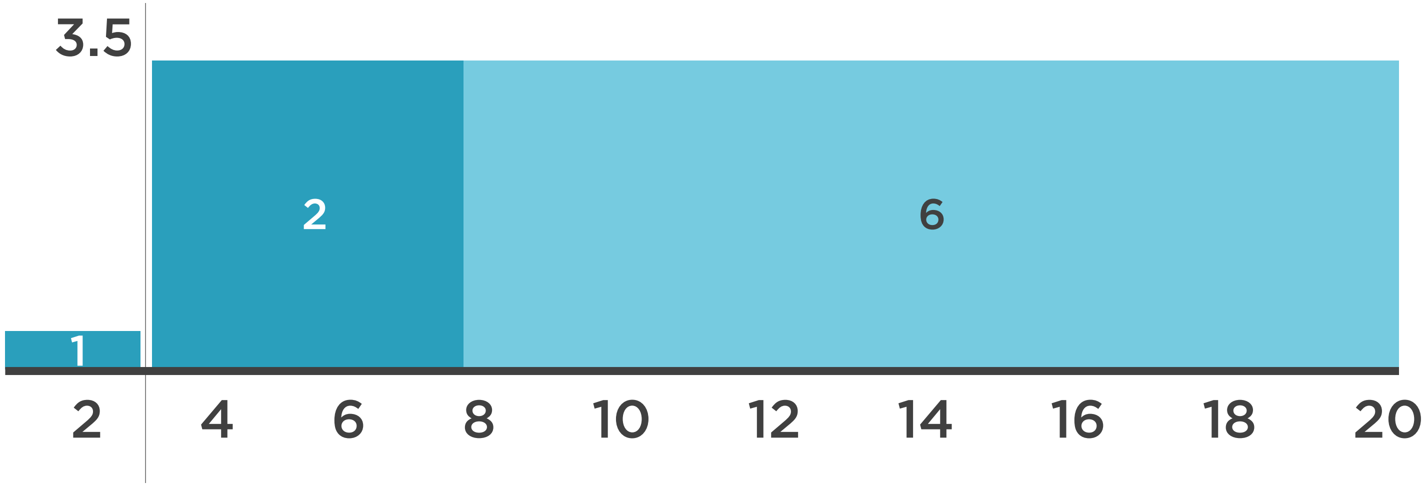
**per week*

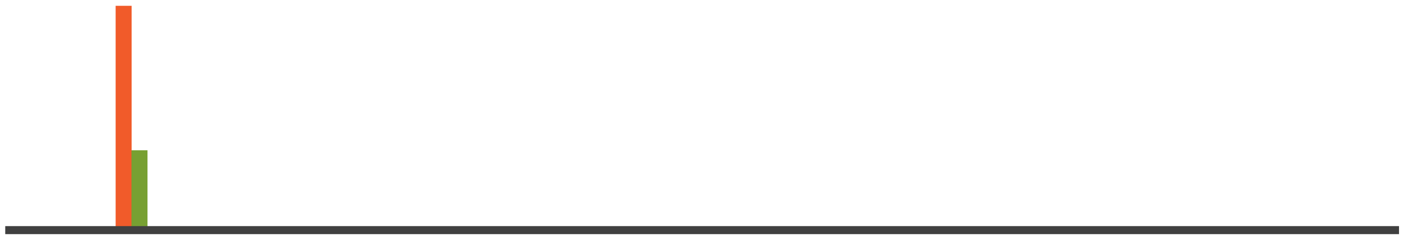


Entropy-based Discretization

Facebook hours	> 80%
3	1
4	1
5	0
7	0
10	0
11	0
14	1
15	0
19	0

*per week



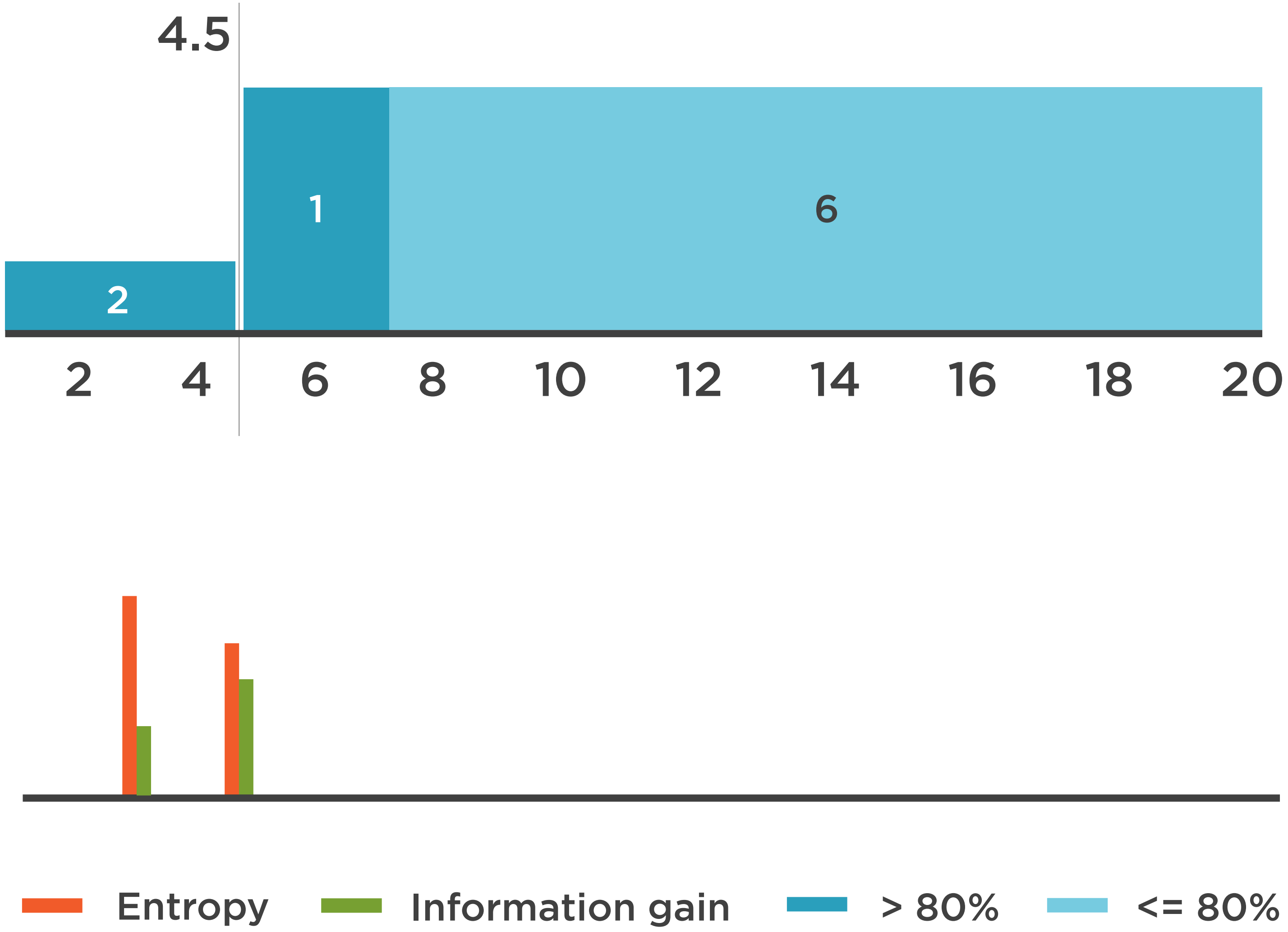


Entropy is inversely proportional to Information gain

Entropy-based Discretization

Facebook hours	> 80%
3	1
4	1
5	0
7	0
10	0
11	0
14	1
15	0
19	0

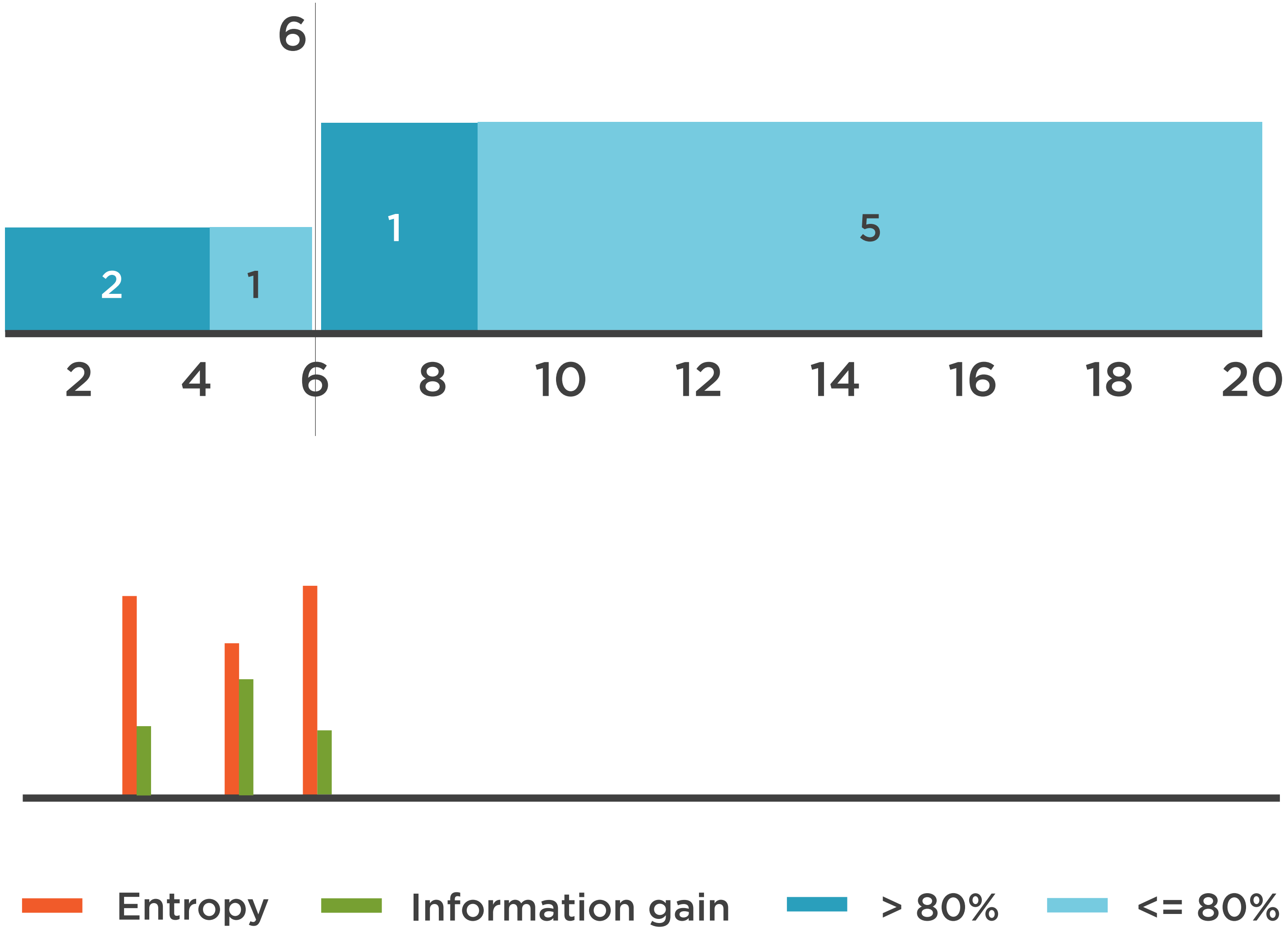
*per week



Entropy-based Discretization

Facebook hours	> 80%
3	1
4	1
5	0
7	0
10	0
11	0
14	1
15	0
19	0

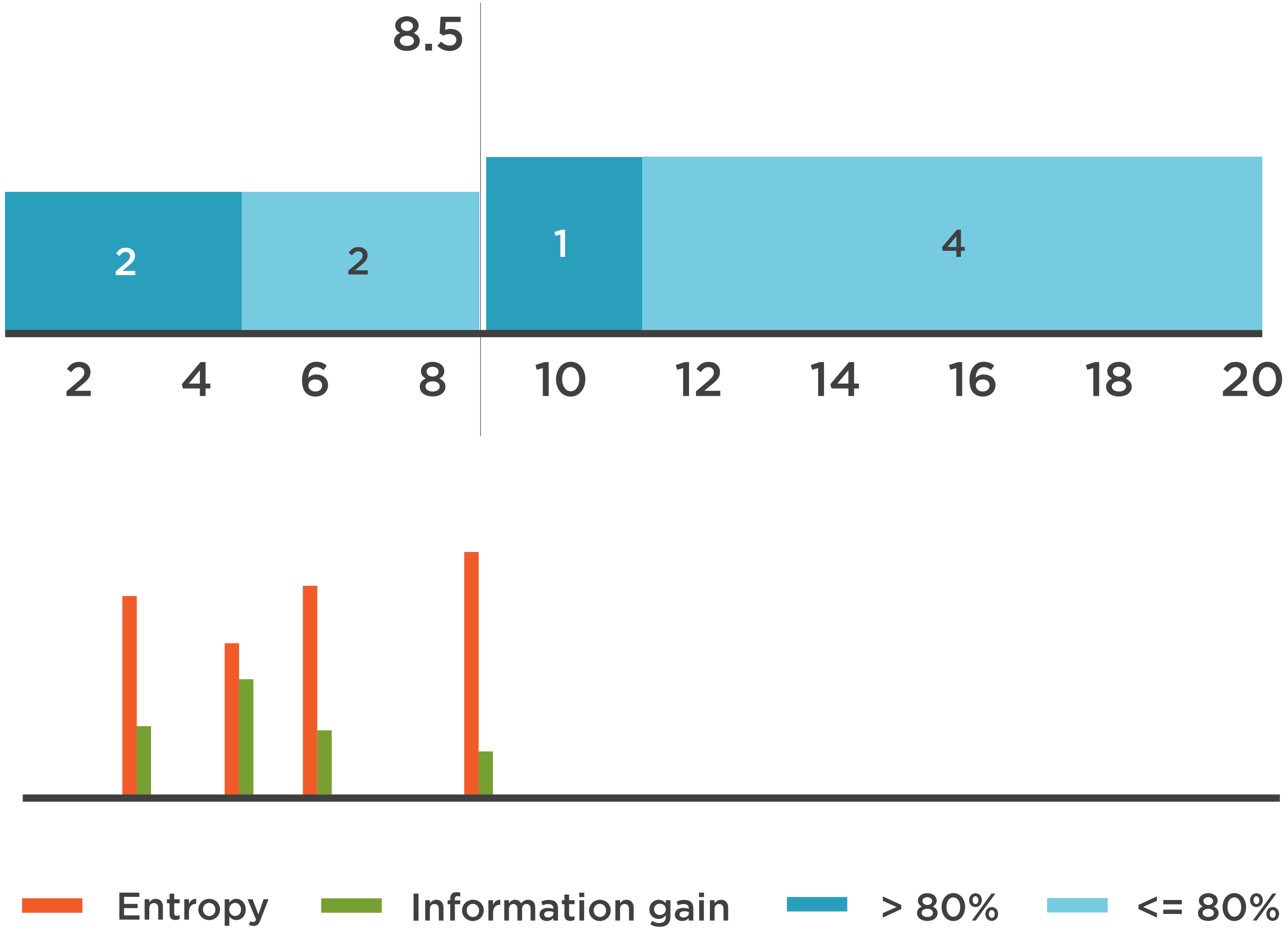
*per week



Entropy-based Discretization

Facebook hours	> 80%
3	1
4	1
5	0
7	0
10	0
11	0
14	1
15	0
19	0

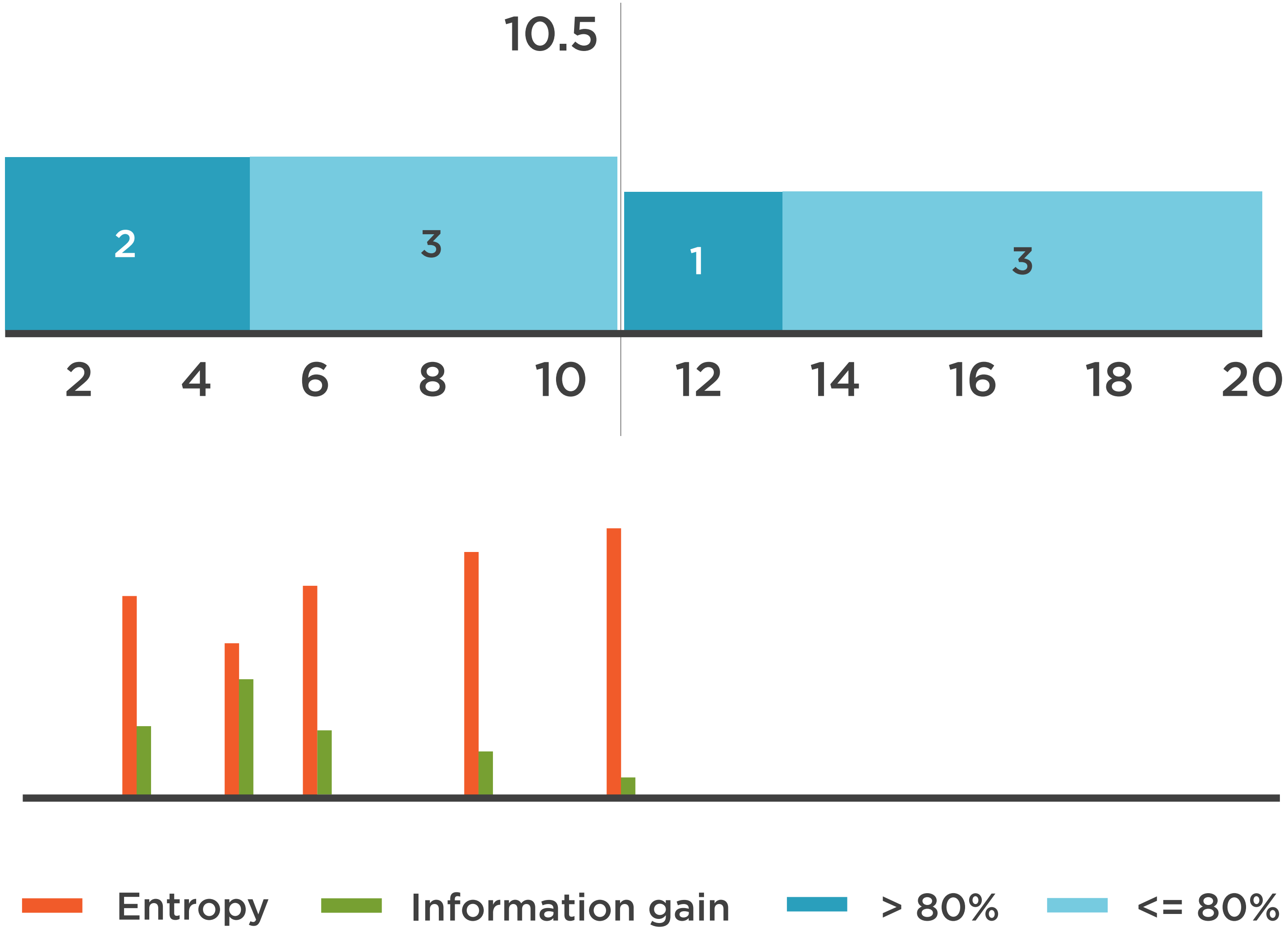
**per week*



Entropy-based Discretization

Facebook hours	> 80%
3	1
4	1
5	0
7	0
10	0
11	0
14	1
15	0
19	0

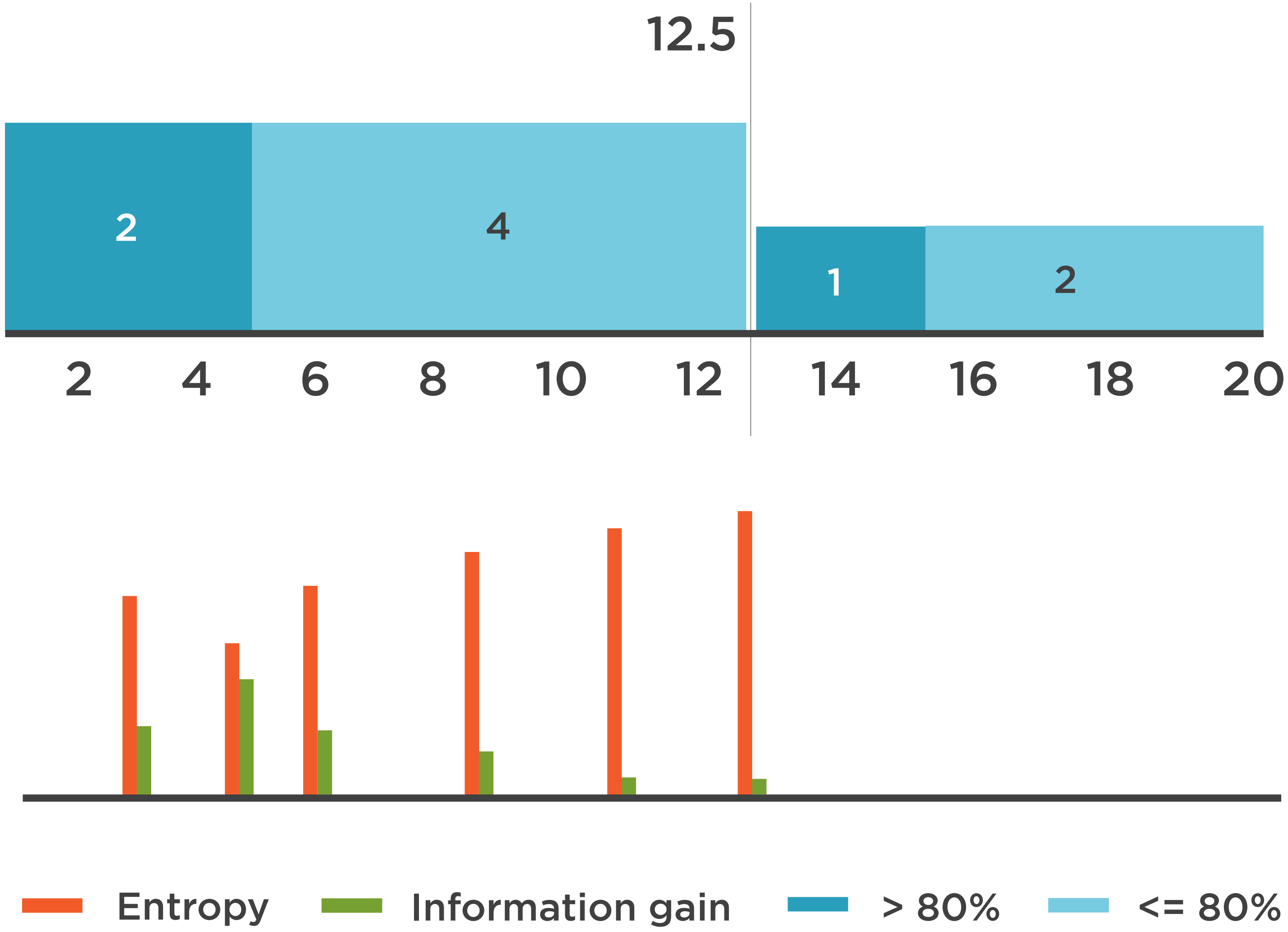
*per week



Entropy-based Discretization

Facebook hours	> 80%
3	1
4	1
5	0
7	0
10	0
11	0
14	1
15	0
19	0

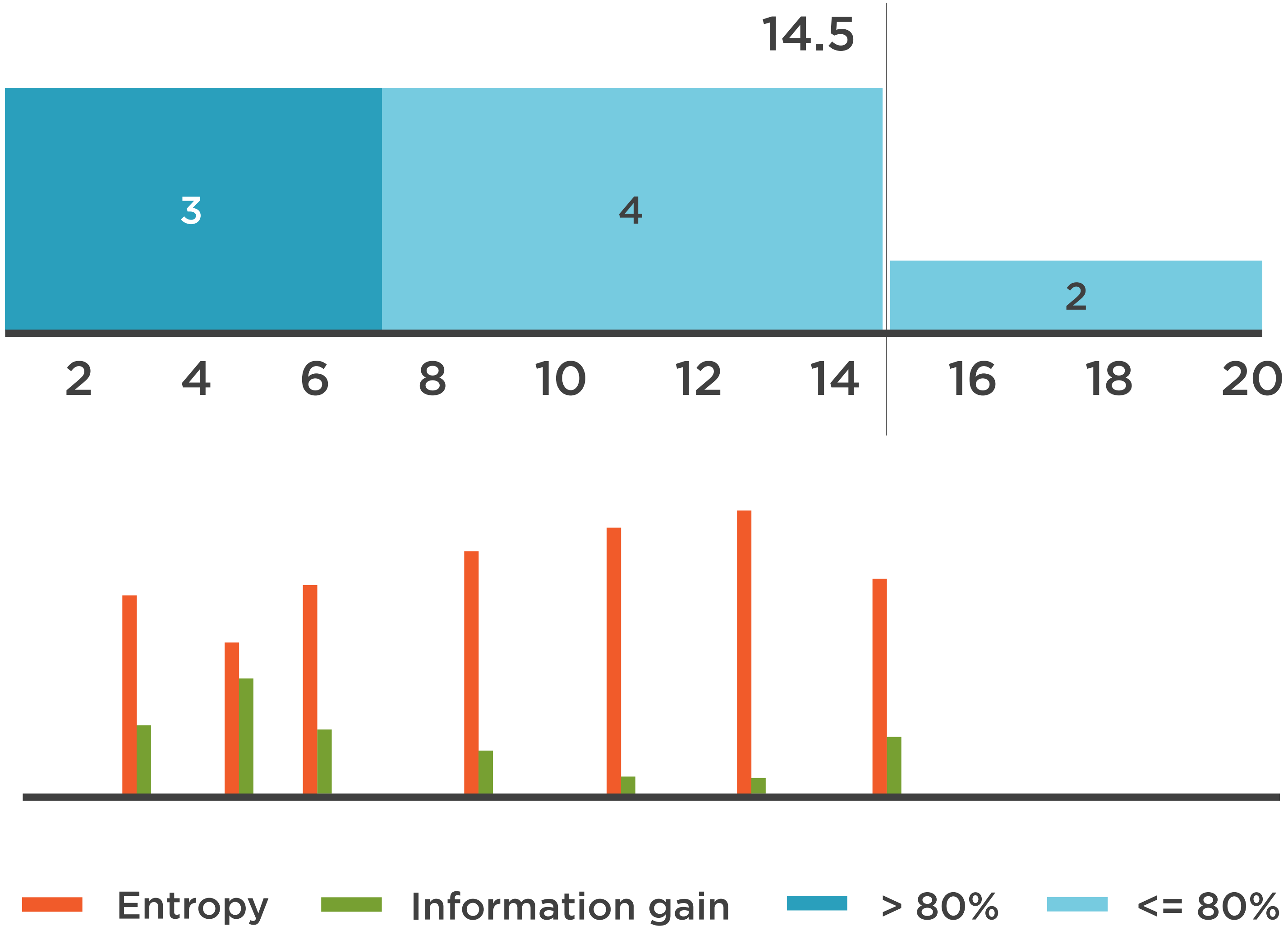
*per week



Entropy-based Discretization

Facebook hours	> 80%
3	1
4	1
5	0
7	0
10	0
11	0
14	1
15	0
19	0

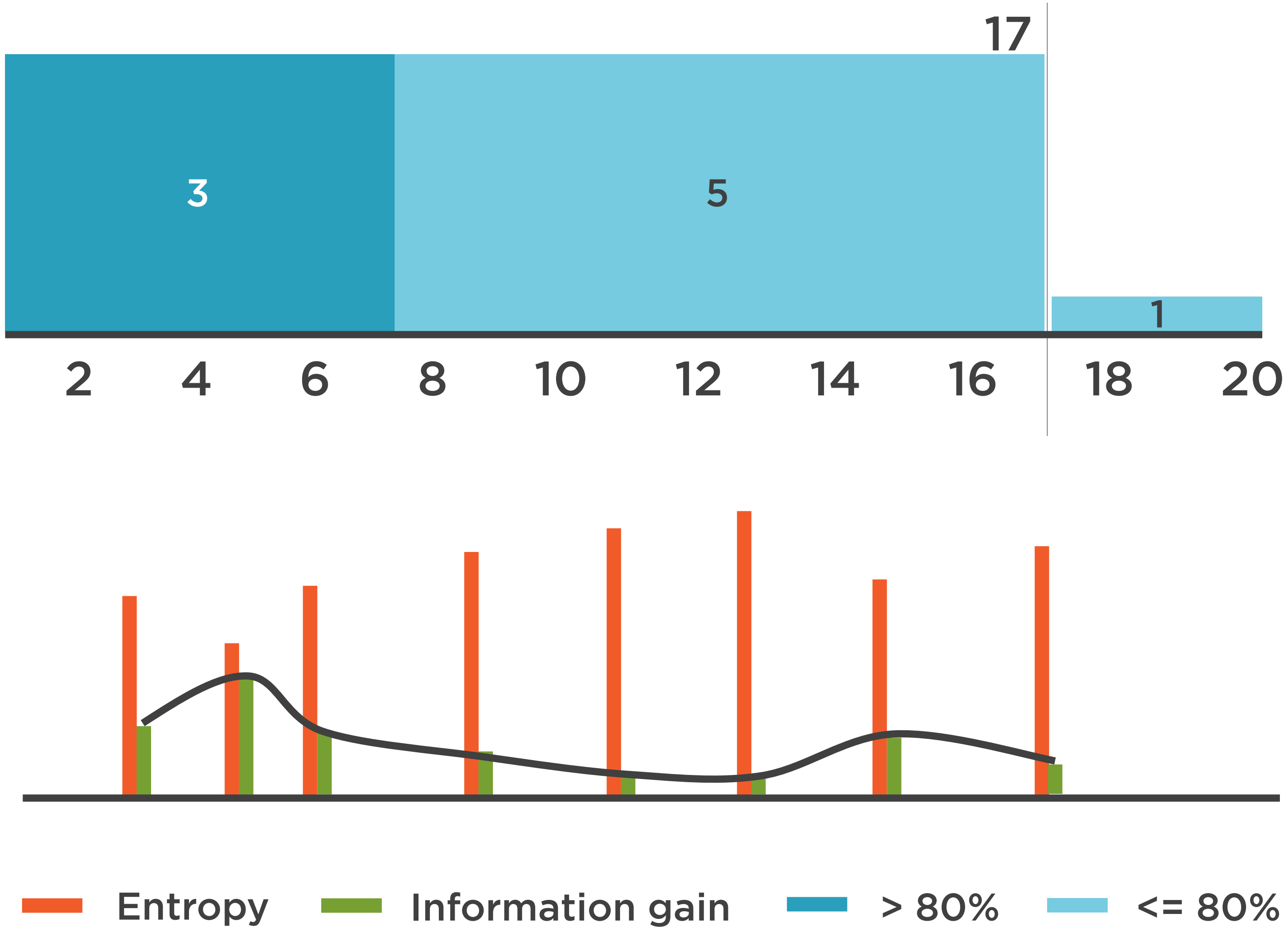
*per week



Entropy-based Discretization

Facebook hours	> 80%
3	1
4	1
5	0
7	0
10	0
11	0
14	1
15	0
19	0

per week



Demo

Entropy MDL

Summary

EDA helps us to understand data better

Data Preprocessing transforms the data suitable for ML

Algorithms “somehow” cannot find patterns. We help it to do so!

Not all tasks are required for every problem