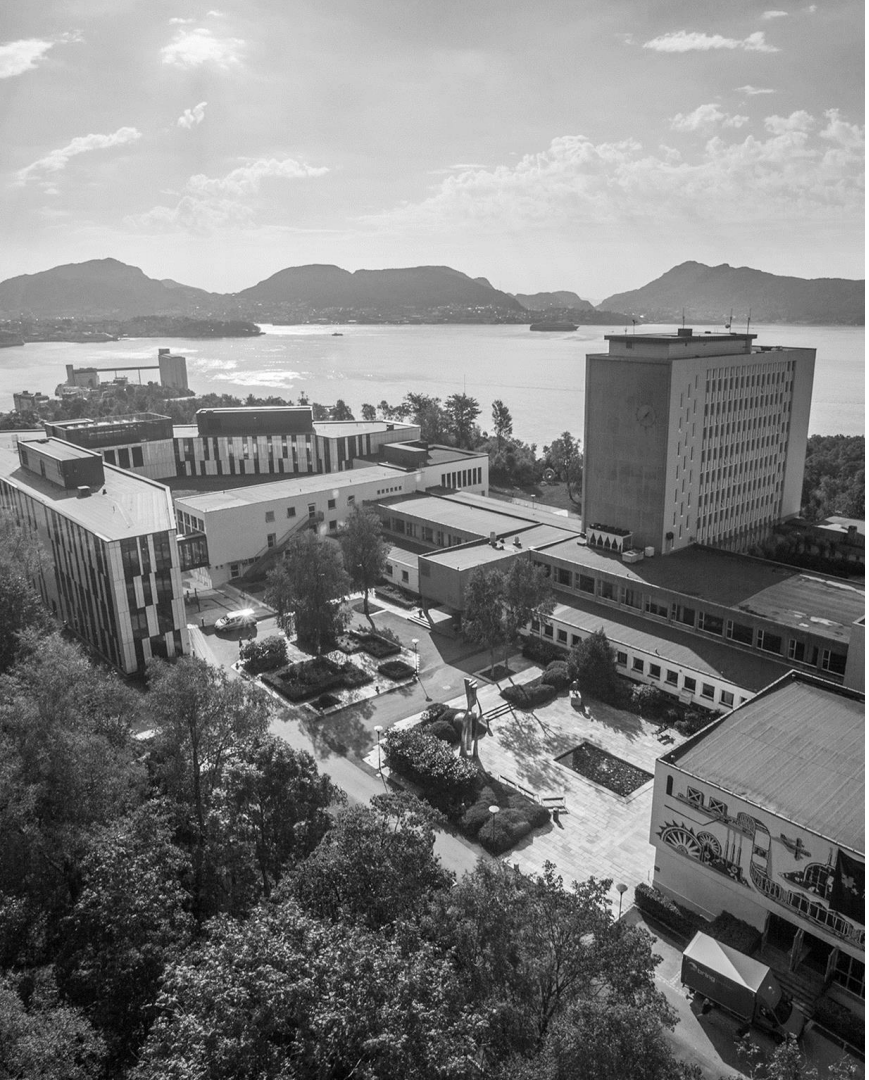


Guest Lecture BAN400

05. October, 2021

1. Introduction
2. Why R?
3. Project example: Valuation of Holiday Houses
4. Code along and exercises
5. Closing remarks

pwc

1

Introduction

Thomas Hansen
Direktør

+47 952 60 254
thomas.hansen@pwc.com

Line Melby
Senior Associate

+47 948 21 449
line.melby@pwc.com

Nora Hansen
Manager

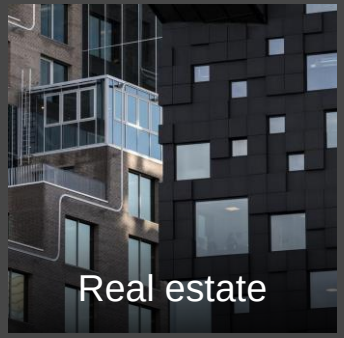
+47 418 58 060
nora.hansen@pwc.com

Fredrik Angelvik
Senior Associate

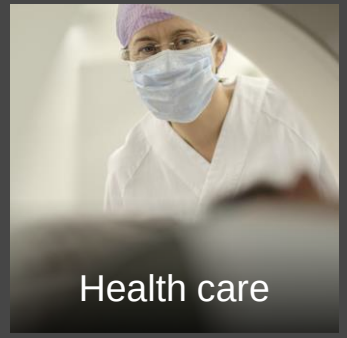
+47 480 48 017
fredrik.angelvik@pwc.com

PwC are working with changing industries

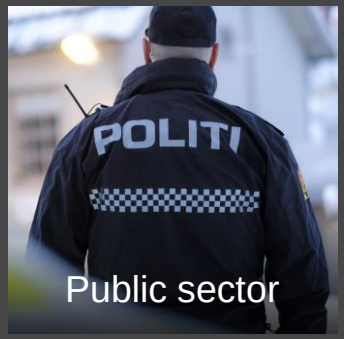
Finance and banking


Real estate


Energy


Health care

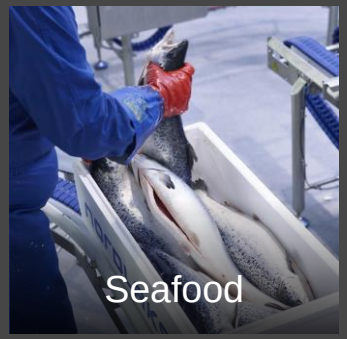

Retail

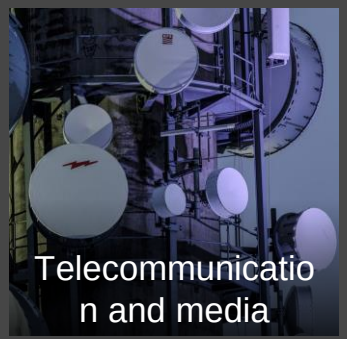


Public sector

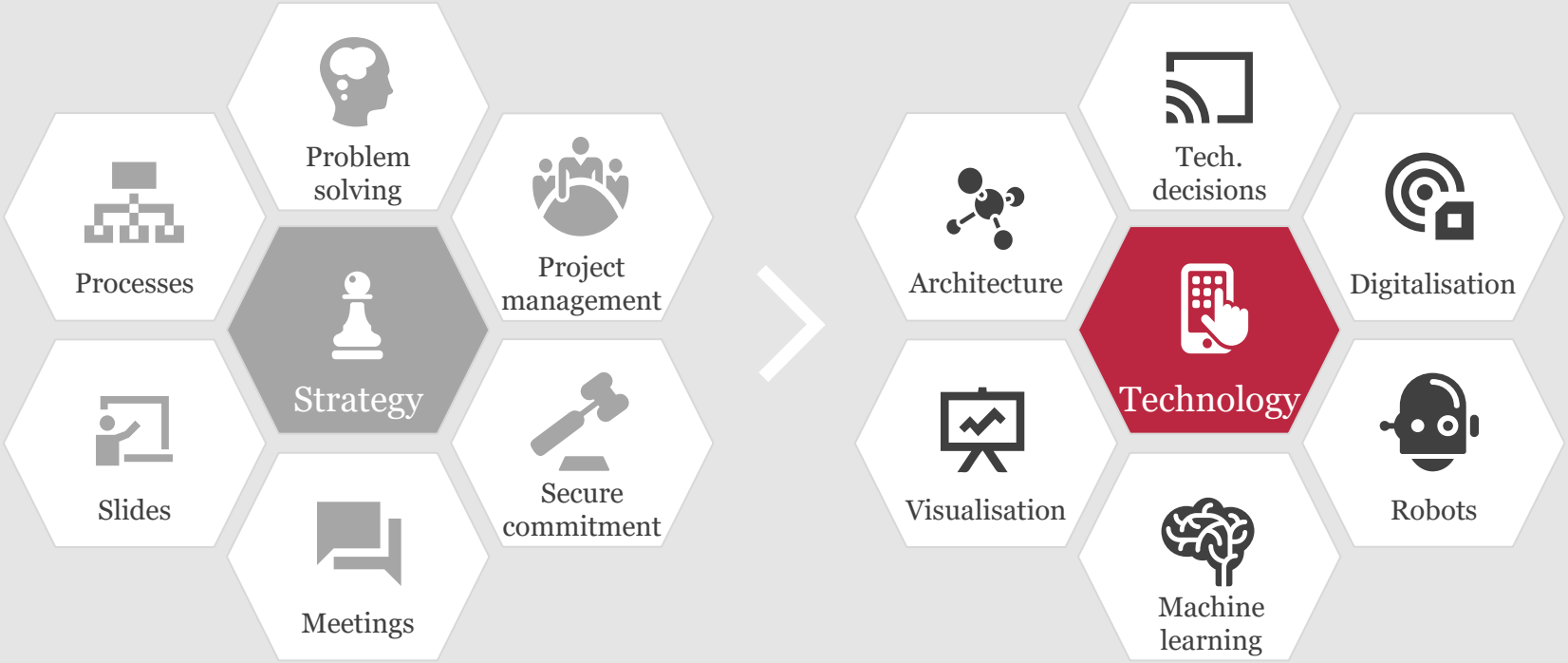

Oil and gas


Shipping


Seafood


Telecommunication and media

Technological development creates needs and expectations. Tomorrow's consultants are dependent on mastering the technological landscape



2

Why R?

Why we love R
(We use Python as well)

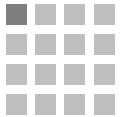
1. A common syntax which is easy to write – but also understand

Calculate average square meter price for 3-room apartments over 60 square meters for each district in Oslo. Sort results by district with lowest square meter price.

```
data <- data_raw %>%
  filter(fylke == "Oslo", sqm > 60, bedrooms = 2) %>%
  group_by(bydel) %>%
  summarise(price_sqm = avg(price_sqm)) %>%
  arrange(price_sqm)
```

2. Functionality covering 99.9 % of the use cases for a Data Scientist with the same coherent syntax

Database



Big data

Visualisation

Text analysis

Mail/Push

API

Notebooks

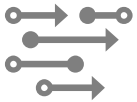
Maps

Transformation

Machine learning

Computer Vision

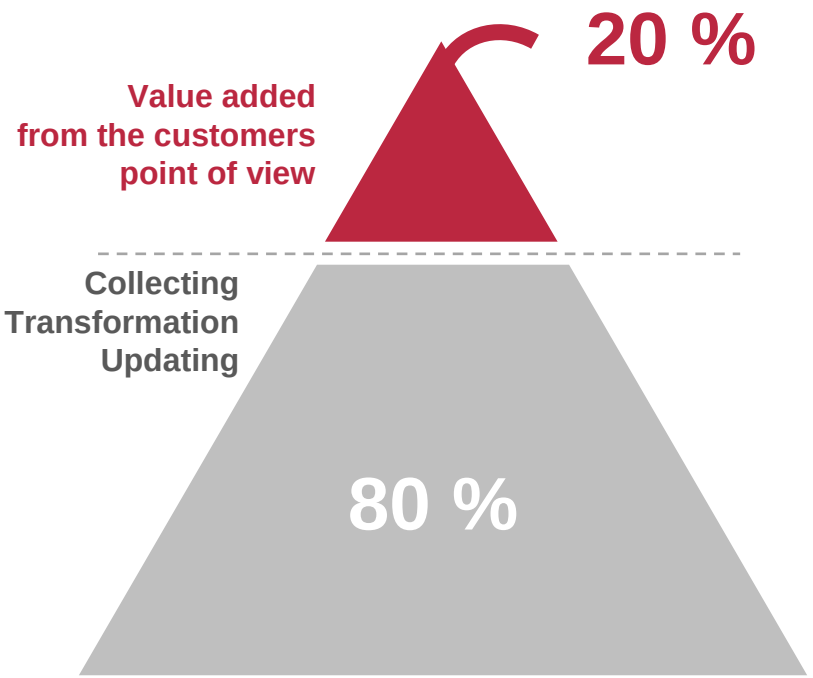
Simulation



Optimisation

Python

3. It’s the best tool for data wrangling and visualization (Opinionated)



atomic_explosion 20 points · 6 days ago

I agree with u/baconshoplifter, it's the libraries that matter not the language.

I use Python to clean and wrangle my data as libraries like Pandas and Numpy make it super easy especially with the ton of CSV files I work with. I use R to analyze and plot my data as I love ggplot and the ton of statistical packages in R.

I also use Python to build APIs, using Flask, and run them on AWS. So it really depends on what you are trying to do but I find it useful knowing both.

Reply Share Report Save

foxhollow 47 points · 6 days ago

I use python for many tasks and generally like it, but IMO, if you're comfortable in R, you're crazy to use pandas for data wrangling instead of dplyr. dplyr is so consistent and beautiful, while pandas is a design disaster.

Reply Share Report Save

ACrispWinterDay 11 points · 6 days ago

Yeah I was going to say the same thing. RStudio with dplyr and the tidyverse in general is heaven when it comes to data wrangling

Reply Share Report Save

novokaoi 7 points · 6 days ago

I had that exact thought the other day when I went back to using python for a project after using R for 2 years. How did I ever cope with the mess that is the Pandas API?

Reply Share Report Save

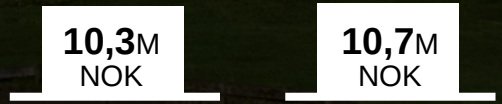
3

Project example:
Valuation of Holiday Houses

Market Value

VS

Asset Value



The current model for valuation of primary homes only accounts for four variables: Area, age, home type and location (price zone)

Today's model
Four variables

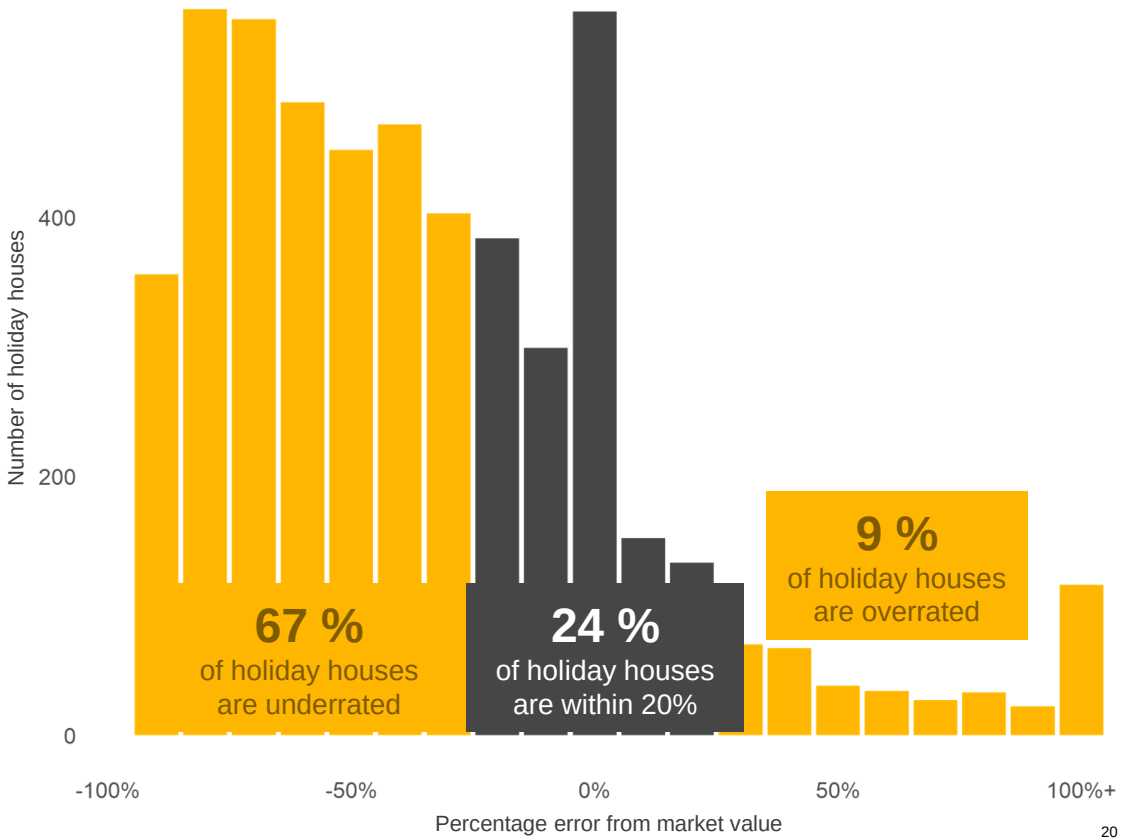
- Use area
- Built year
- Type of housing
- Price zone

Does not take into account:

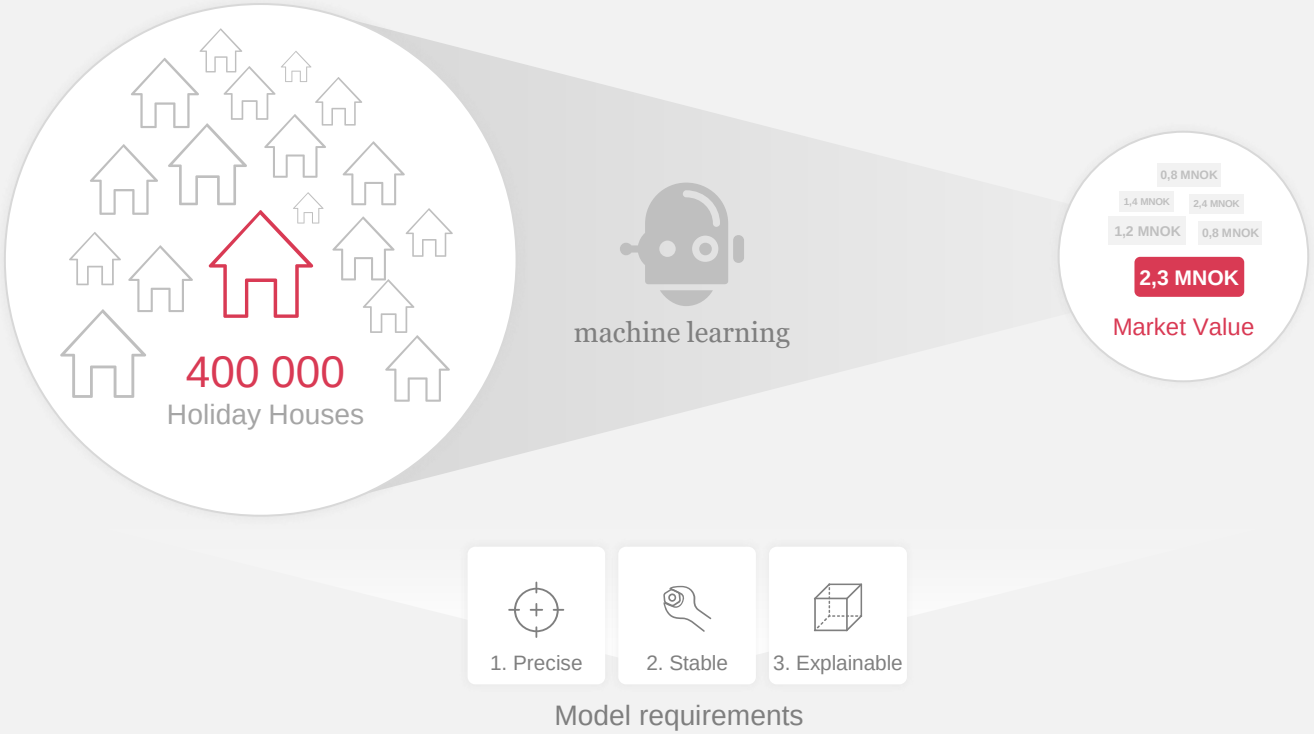
- Standard
- Bedroom
- Stories
- View
- Plot
- Garage
- Public transportation
- Distance to school
- Fireplace, pool, etc.

In 2019, the asset values for holiday homes sold during the year had a mean absolute percentage error of 48 percent measured against the real market value

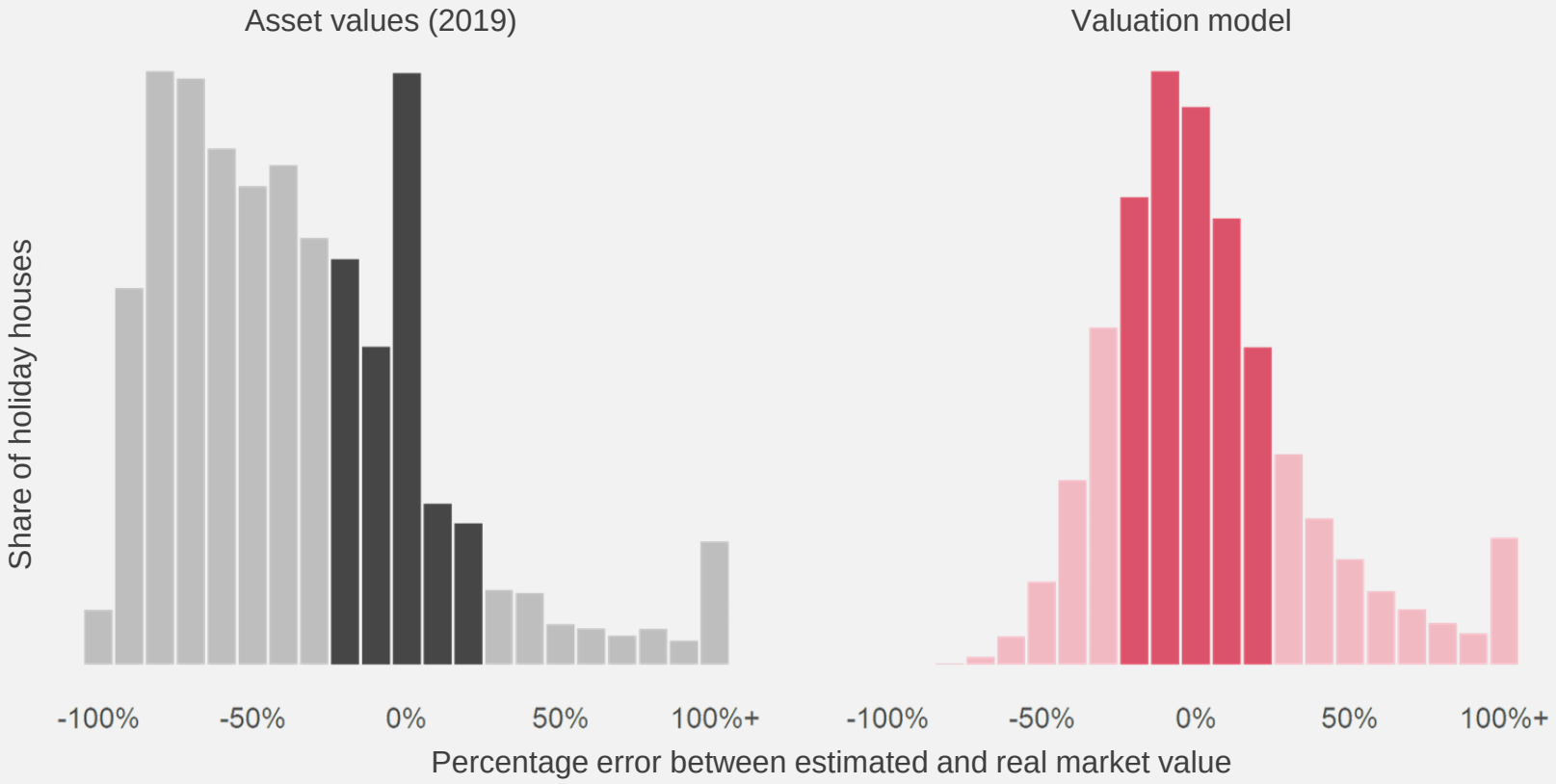
REDACTED



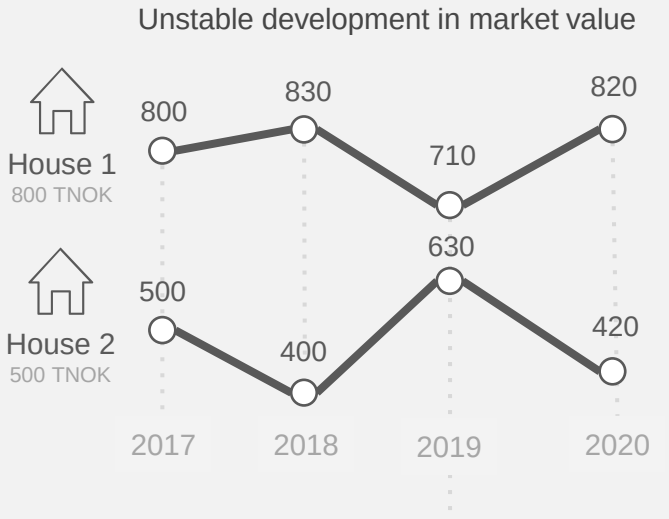
The purpose of the project is to estimate the market value of all Norwegian holiday homes in a precise, stable and explainable manner using current available data



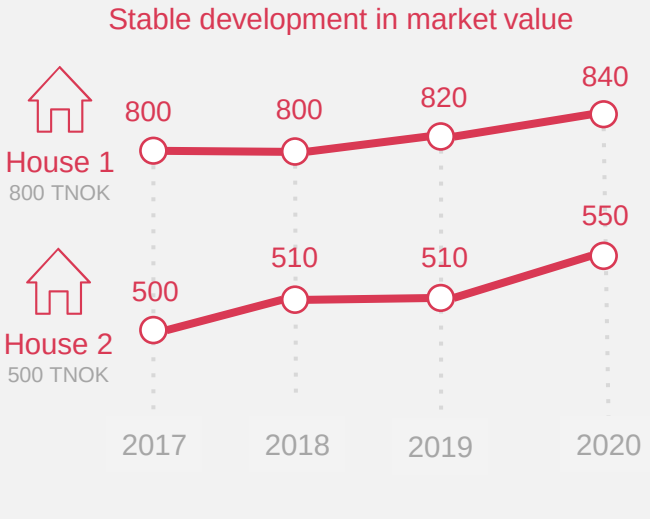
1. Precision: The model should give more precise and fair asset values



2. Stable: The valuations should be stable over time



Annual market development					COV
House 1	800	830	710	820	7%
House 2	500	400	630	420	21%
Average coefficient of variation					14%



Annual market development					COV
House 1	800	800	820	840	2%
House 2	500	510	510	540	4%
Average coefficient of variation					3%

2. Stable: The valuations should be stable within the same area

Unstable market value development in municipalities

	Annual market development			Std.dev
Municipality 1	8%	-10%	2%	9%
Municipality 2	5%	-20%	-12%	13%
Average standard deviation				11%

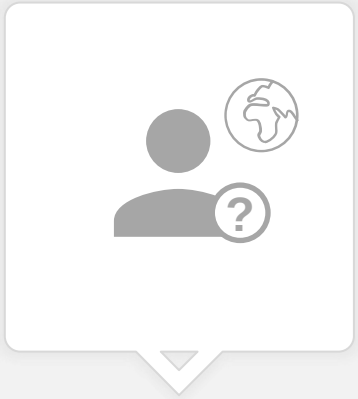
Stable market value development in municipalities

	Annual market development			Std.dev
Municipality 1	2%	-1%	0%	2%
Municipality 2	2%	0%	2%	1%
Average standard deviation				1,5%

3. Explainable: Each valuation must be explainable in a simple manner

Location	Trysil
Area	130 m²
Altitude	898 m
Constr. year	1976
Distance to ski trail	6.2 km
Distance to road	18 m
Distance to sea	175 km
Distance to lake	10 km
Neighbours	776 in 1000 m
Closest neighbour	31 m
Population	76 in 5 km

Today’s valuation models are too crude and there is a great potential for utilizing more granular data

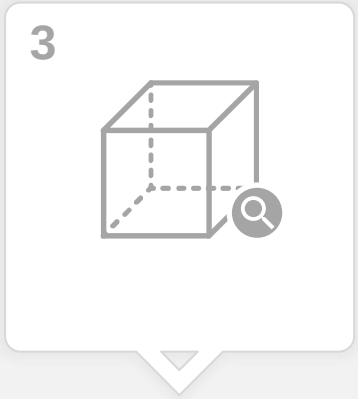


New data sources for valuation

2

Machine learning for valuation

3



Explainability of machine learning

Today’s valuation models are too crude and there is a great potential for utilizing more granular data

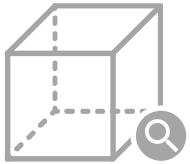
1

New data sources for valuation

2

Machine learning for valuation

3



Explainability of machine learning

The Norwegian property register currently lacks use area and year of construction for 30% and 60% of the holiday homes, respectively

DATA

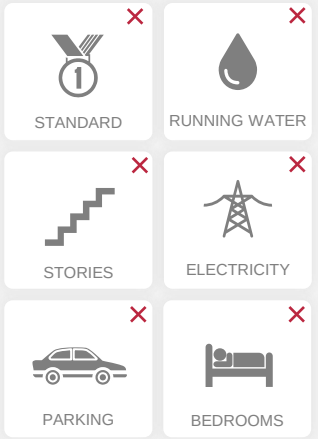
MACHINE LEARNING

EXPLAINABILITY

30 % of holiday homes lacks registered use area

60 % of holiday homes lacks registered year of construction

... and further information is seldom registered



DATA

MACHINE LEARNING

EXPLAINABILITY

Property register

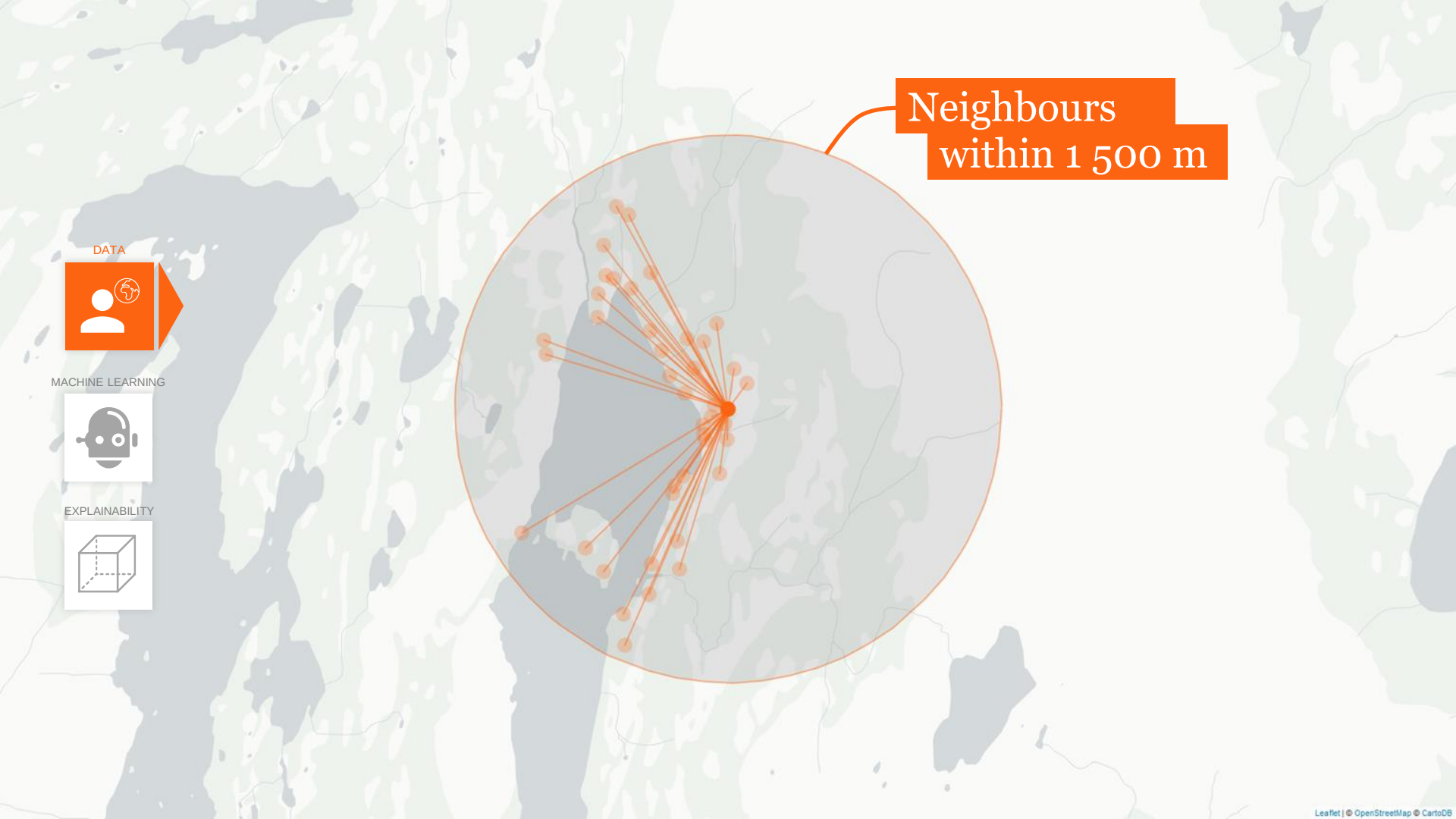
Use area (Building)	-

Finn.no

Use area	287

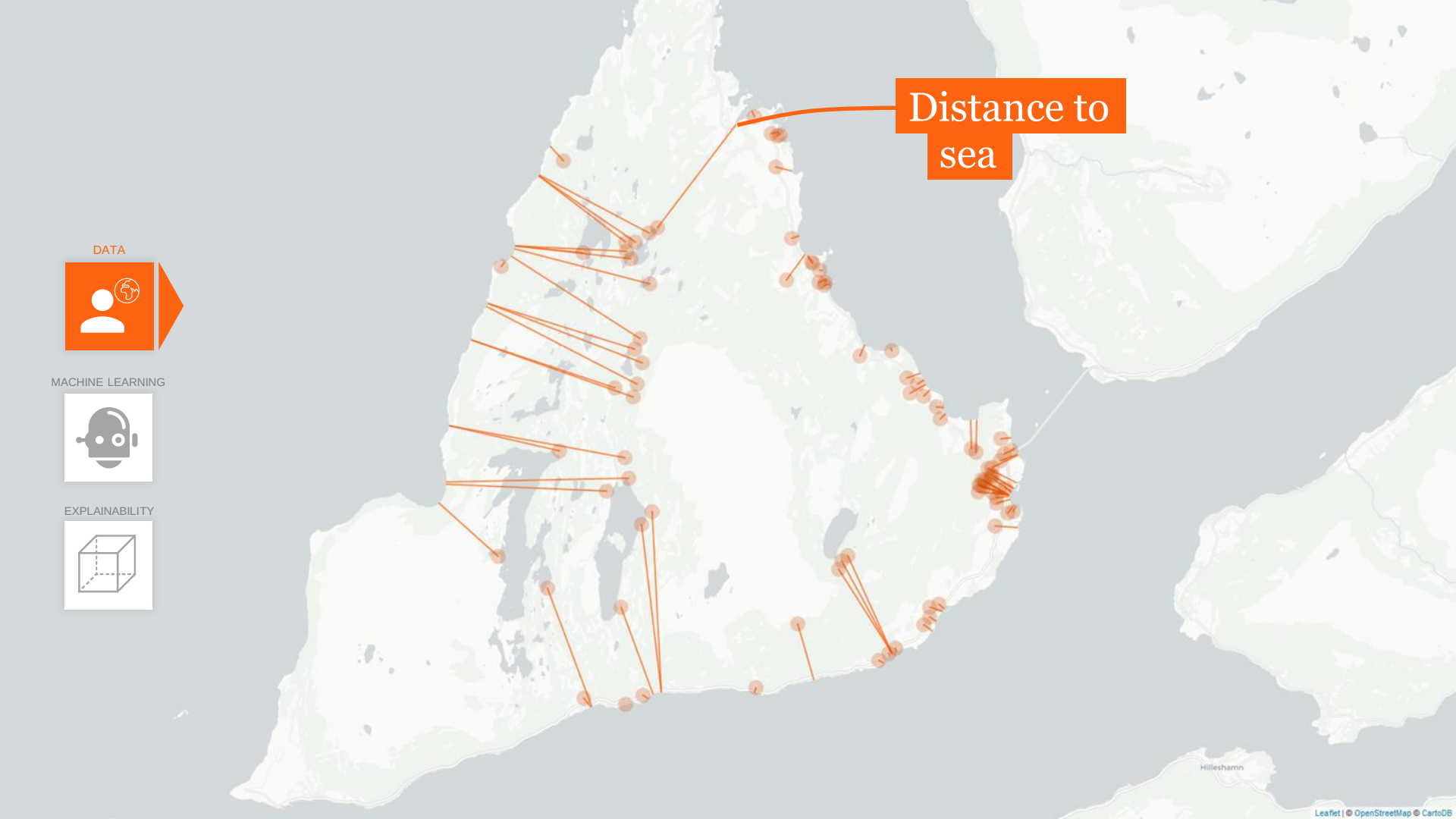
Map

Estimated area	385


Neighbours
within 1 500 m
DATA
MACHINE LEARNING
EXPLAINABILITY
Leaflet | © OpenStreetMap © CartoDB


Distance to road
DATA
MACHINE LEARNING
EXPLAINABILITY
Leaflet | © OpenStreetMap © CartoDB


Distance to lake
DATA
MACHINE LEARNING
EXPLAINABILITY
Leaflet | © OpenStreetMap © CartoDB

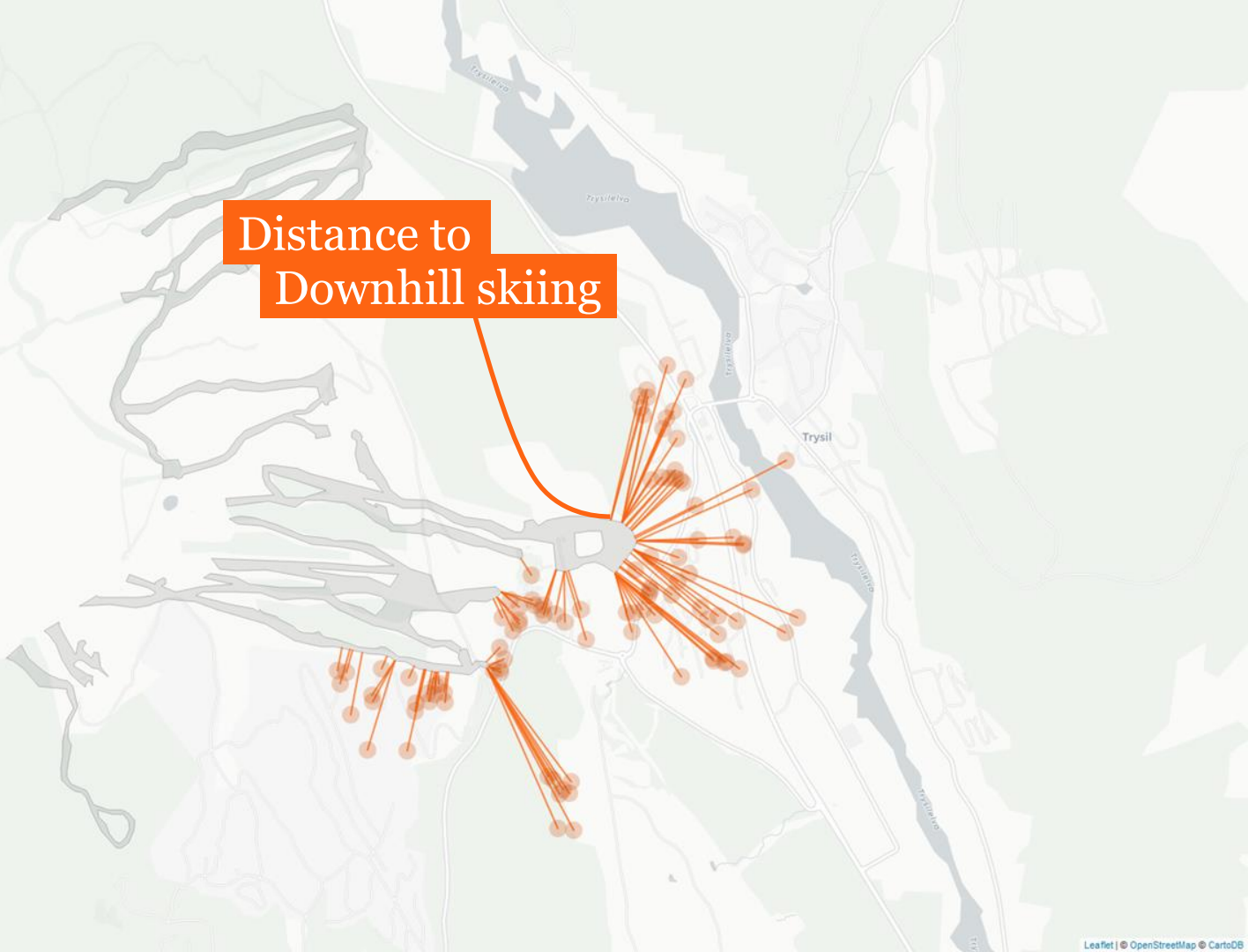

Distance to sea
DATA
MACHINE LEARNING
EXPLAINABILITY
Hilleshamn
Leaflet | © OpenStreetMap © CartoDB

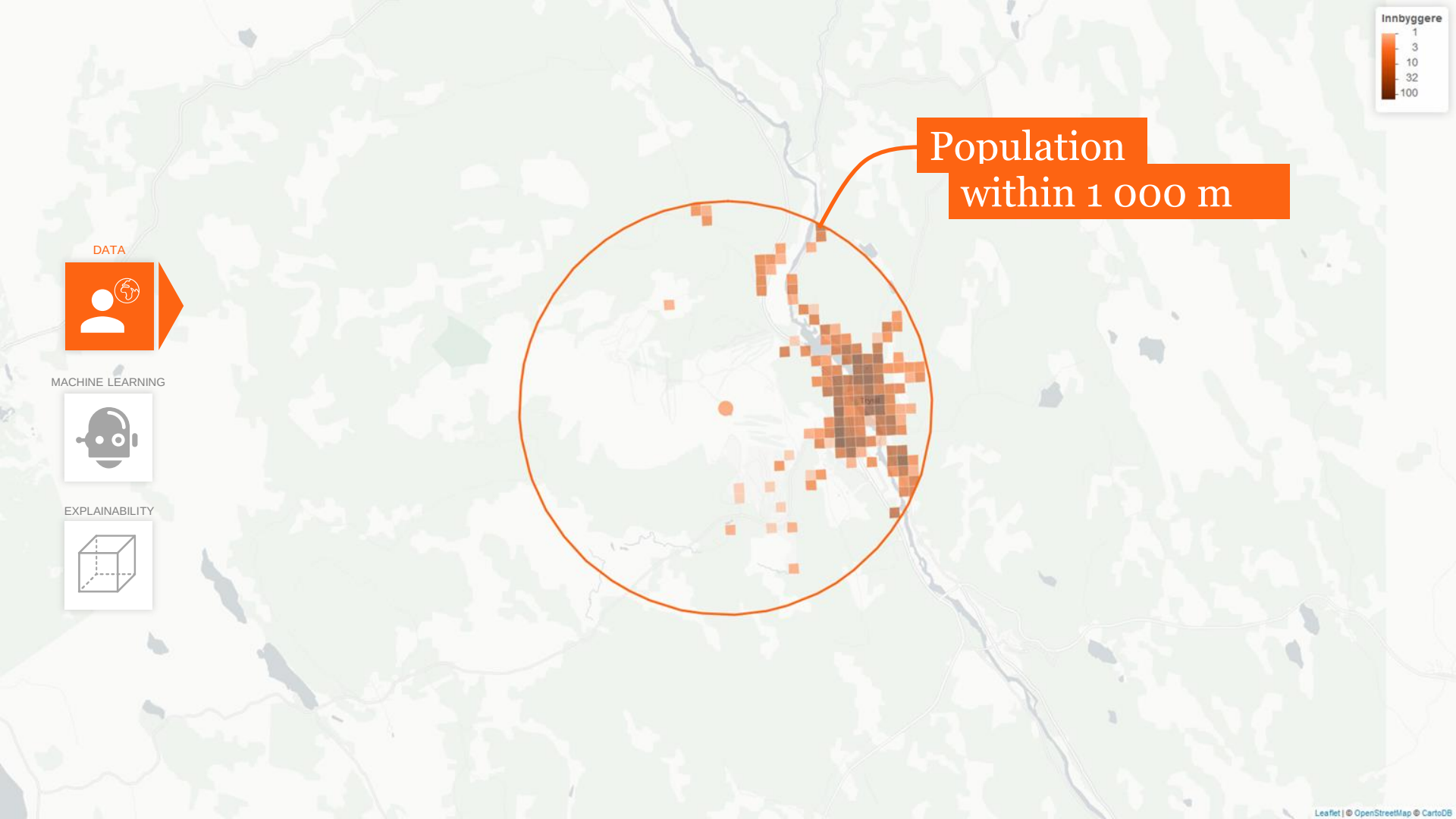
DATA

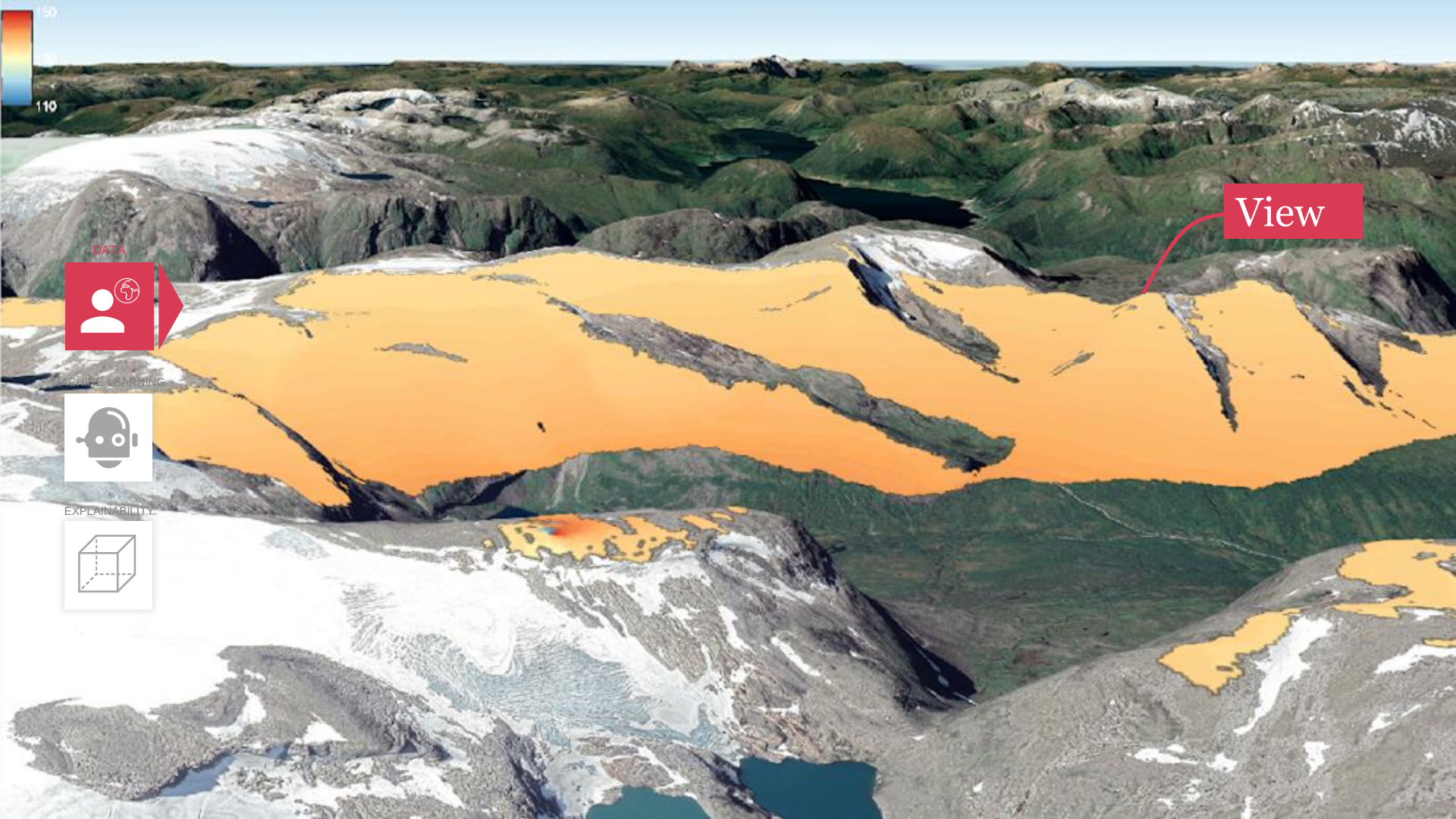
MACHINE LEARNING

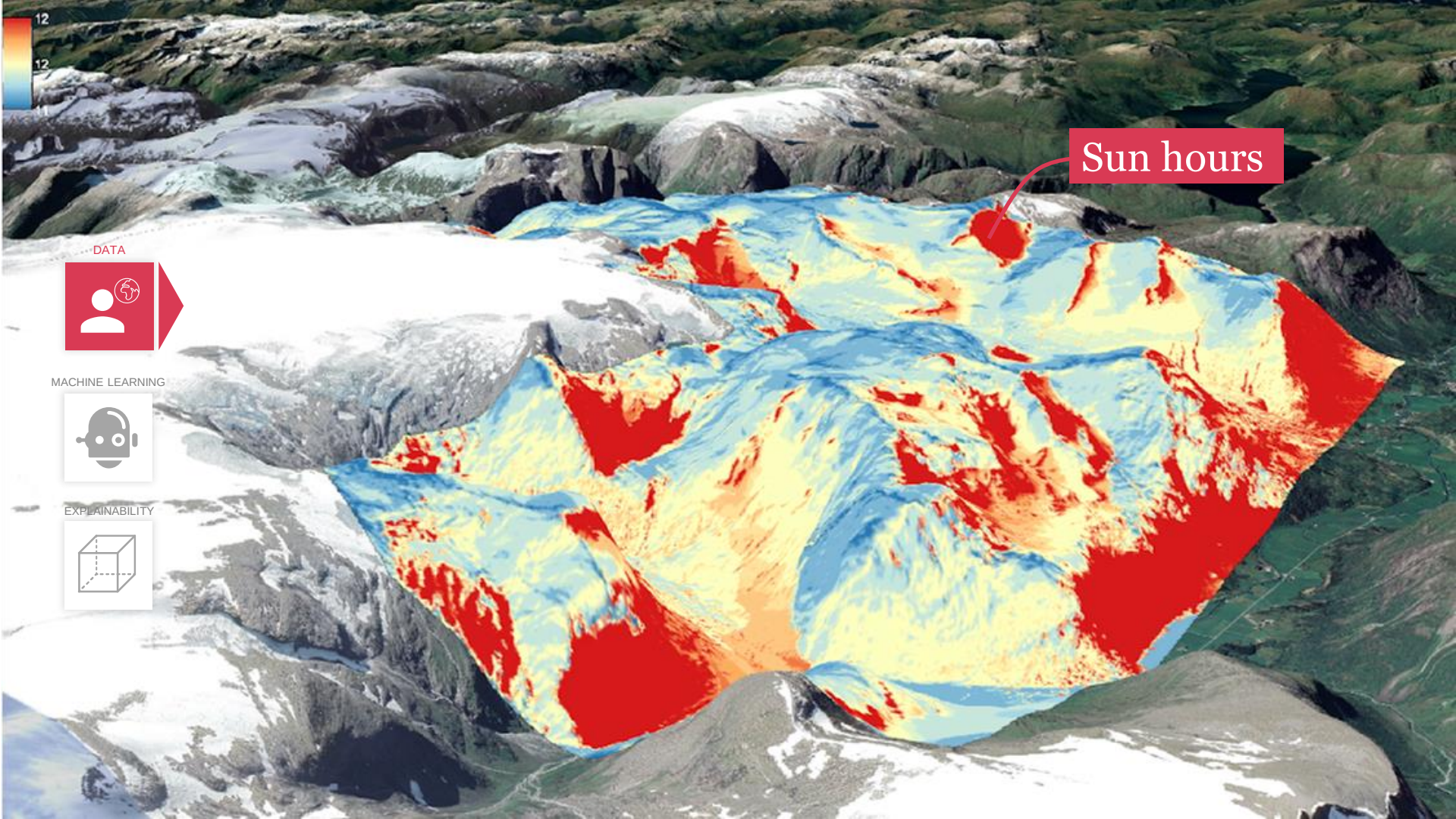
EXPLAINABILITY

Distance to Downhill skiing




Innbyggere
1
3
10
32
100
Population
within 1 000 m
DATA
MACHINE LEARNING
EXPLAINABILITY
Leaflet | © OpenStreetMap © CartoDB


150
110
DATA
MACHINE LEARNING
EXPLAINABILITY
View

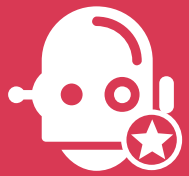

12
12
Sun hours
DATA
MACHINE LEARNING
EXPLAINABILITY

Traditional linear models does not capture the complex relationships that determine the market value of holiday homes

1

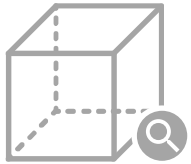
New data sources for valuation

2



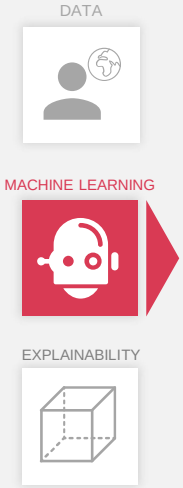
Machine learning for valuation

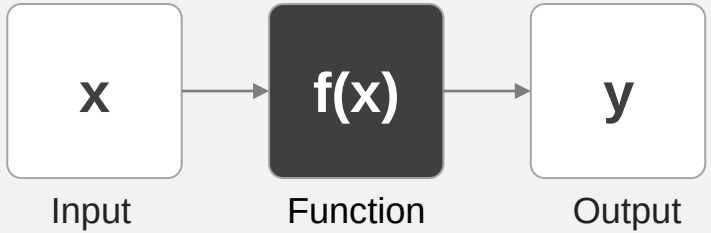
3



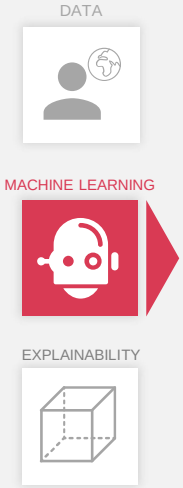
Explainability of machine learning

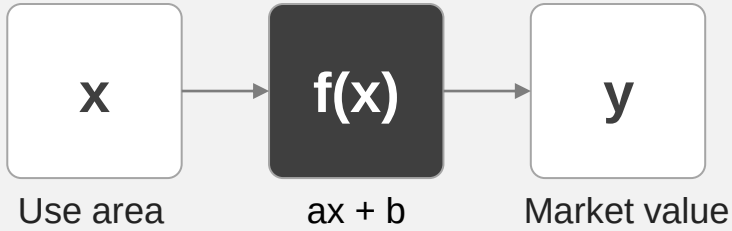
In traditional analyzes we use functions to calculate results (output) based on variables (input)



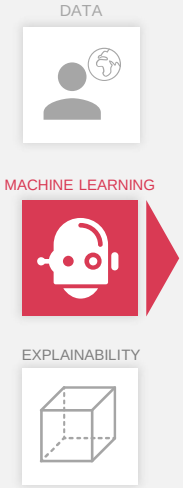


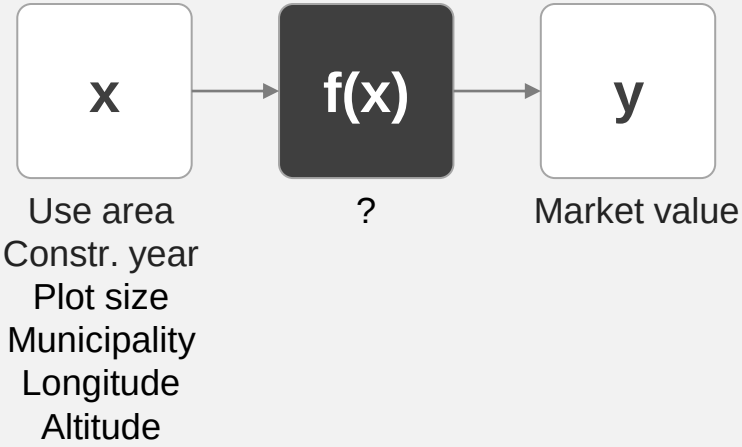
In traditional analyzes we use functions to calculate results (output) based on variables (input)



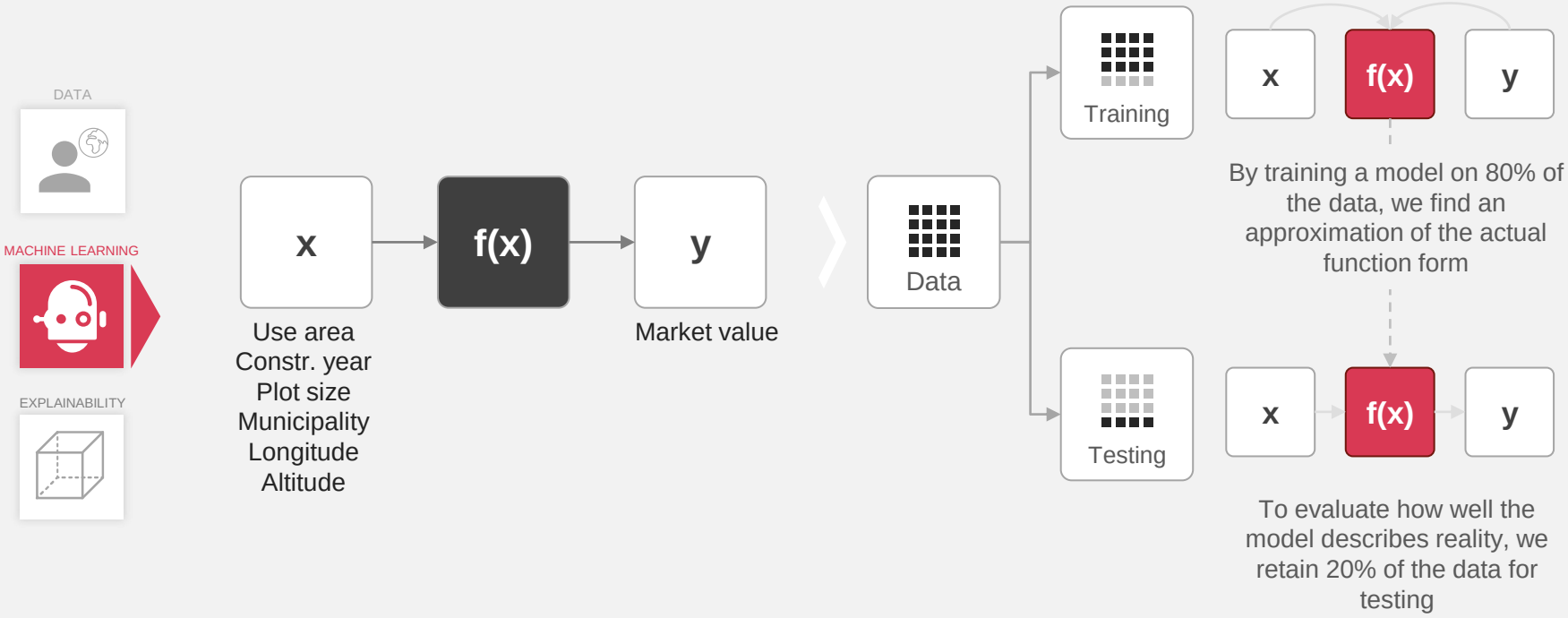


Machine learning involves using data to find a suitable function

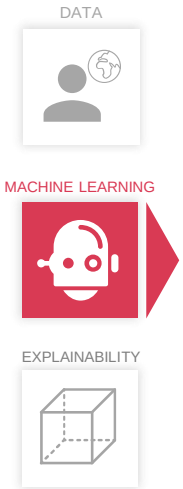


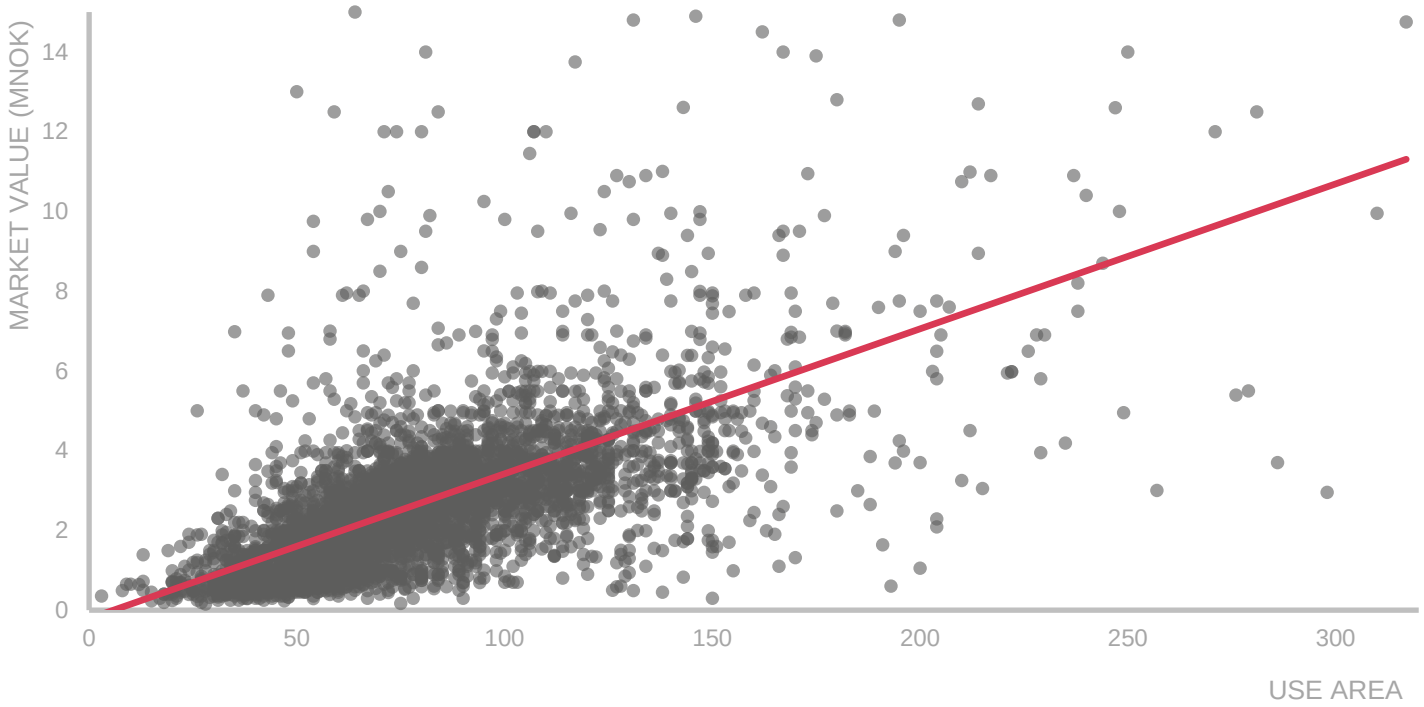


In a typical case, we divide the data in two: We use most (80%) of the data to train the model to find a suitable function. The remaining data (20%) is used for validation

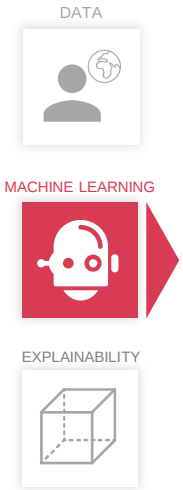


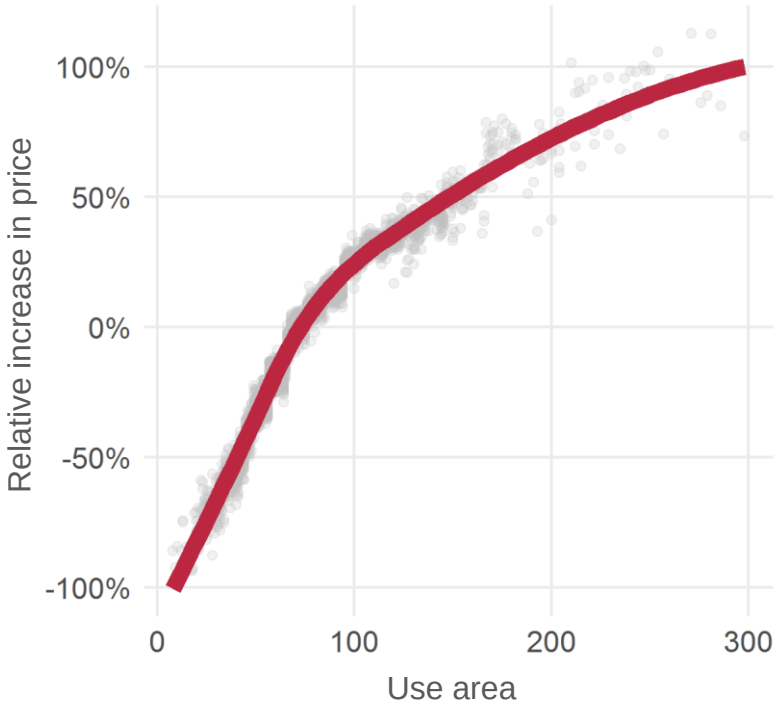
Machine learning can be used to find the linear relationships between input and output that constitute the best linear regression model

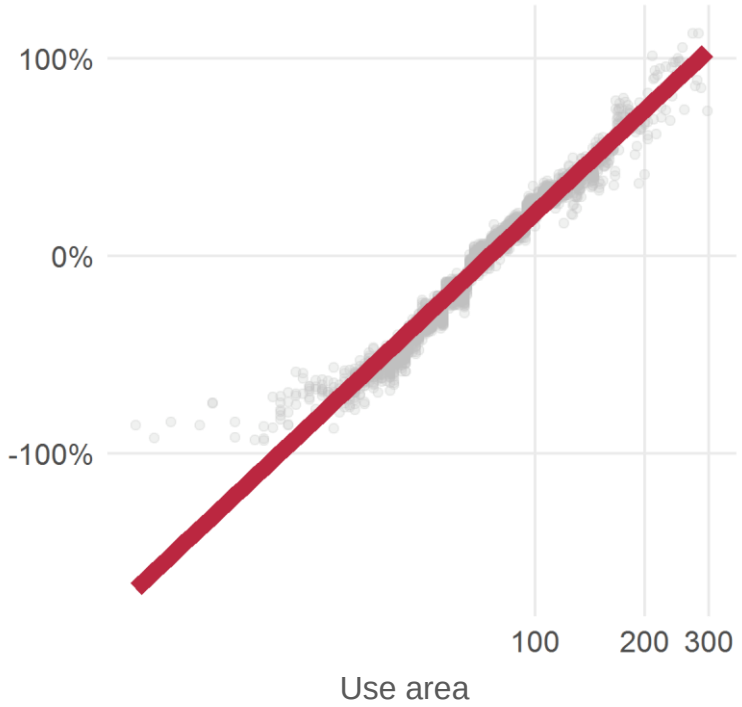




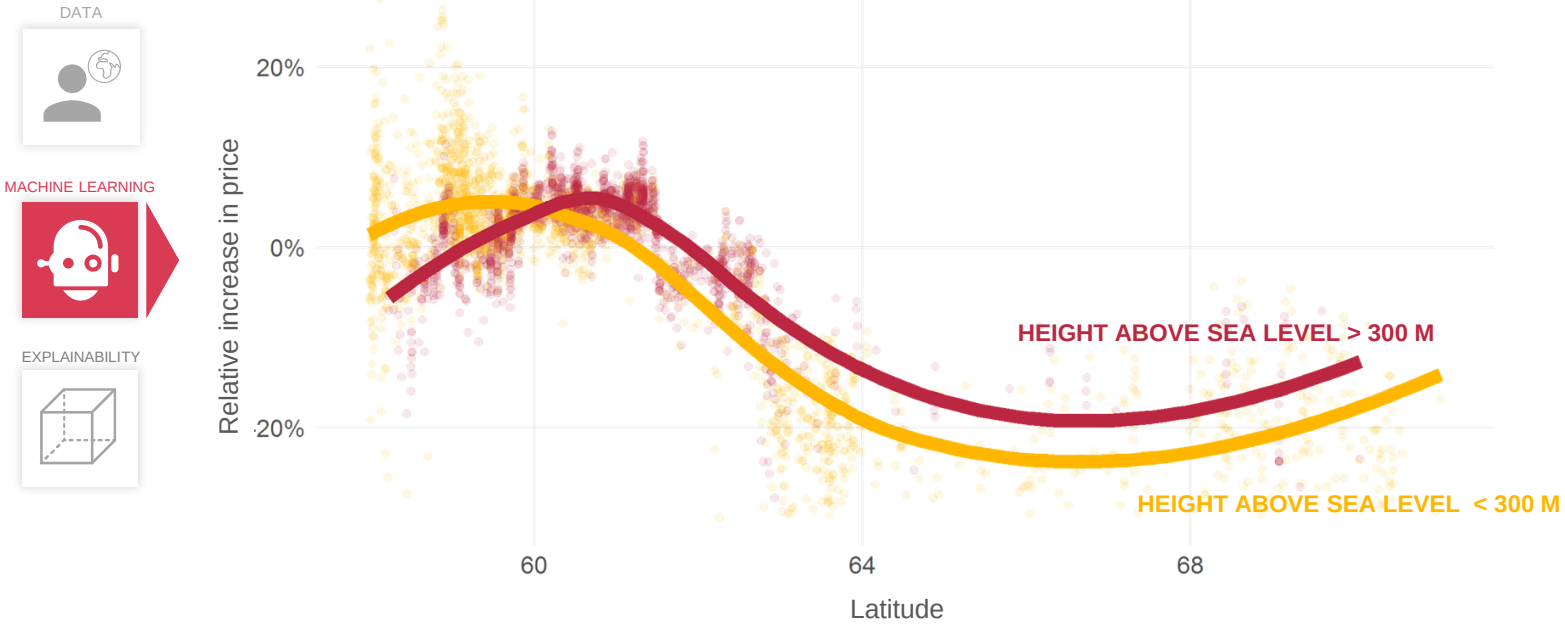
Real world relations are usually non-linear. To capture these using linear models, we need to transform the data in advance







Machine learning can be used to develop models that capture complex relationships and underlying patterns in the data

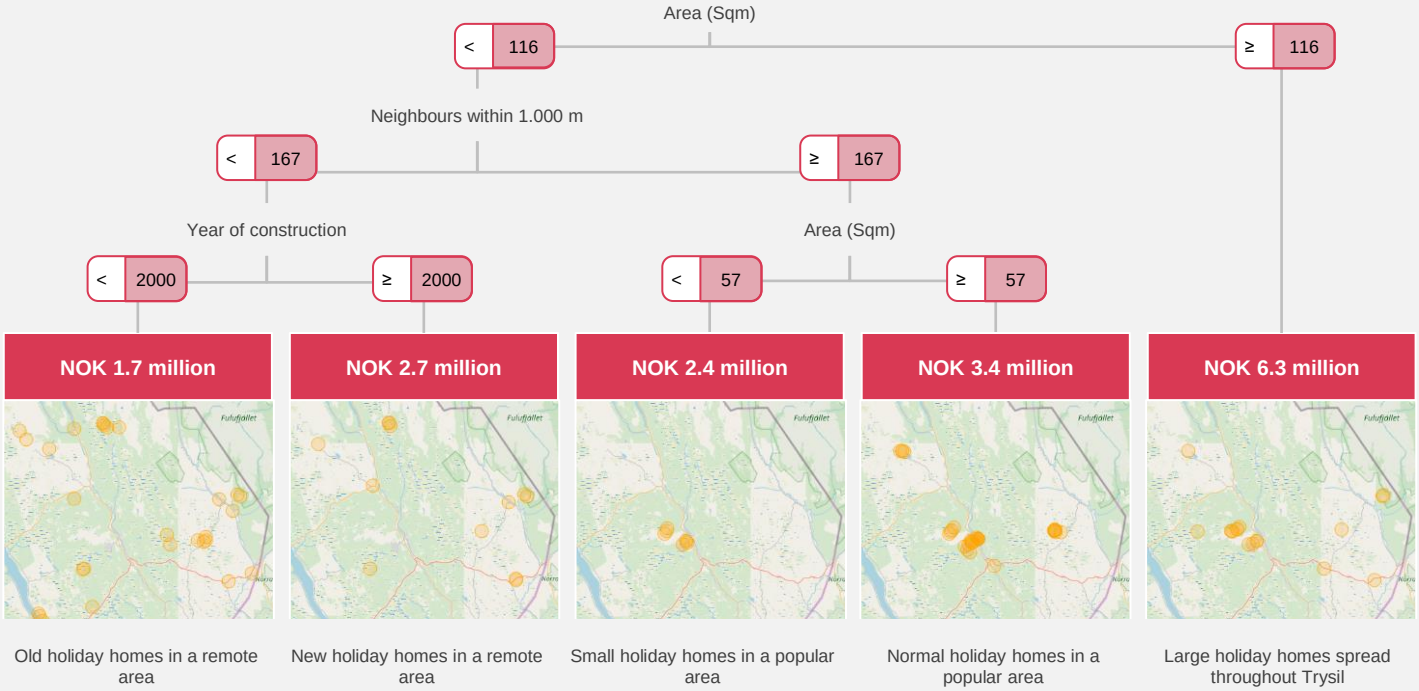


Machine learning can be used to create decision trees, which capture complex interactions and underlying relationships better than linear models

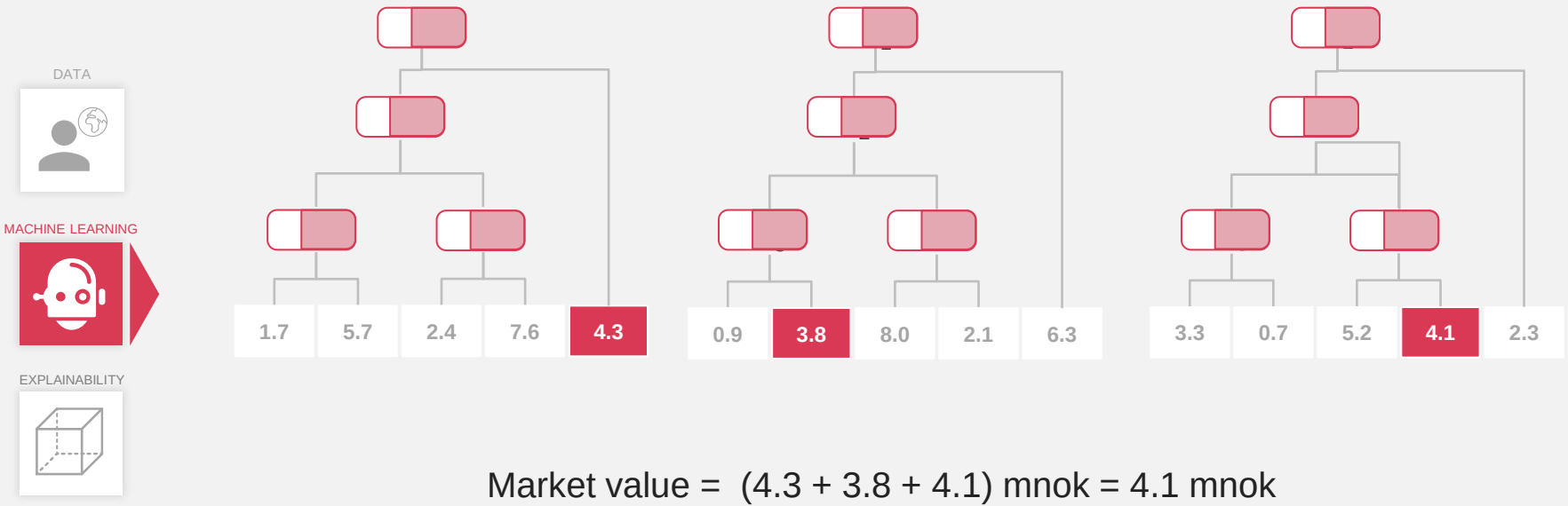
DATA

MACHINE LEARNING

EXPLAINABILITY

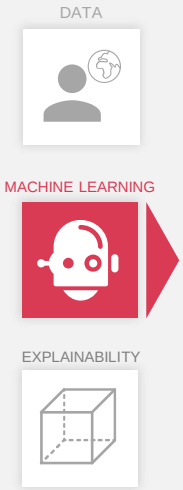


To increase precision, we can create multiple decision trees and use the average of all predictions as market value

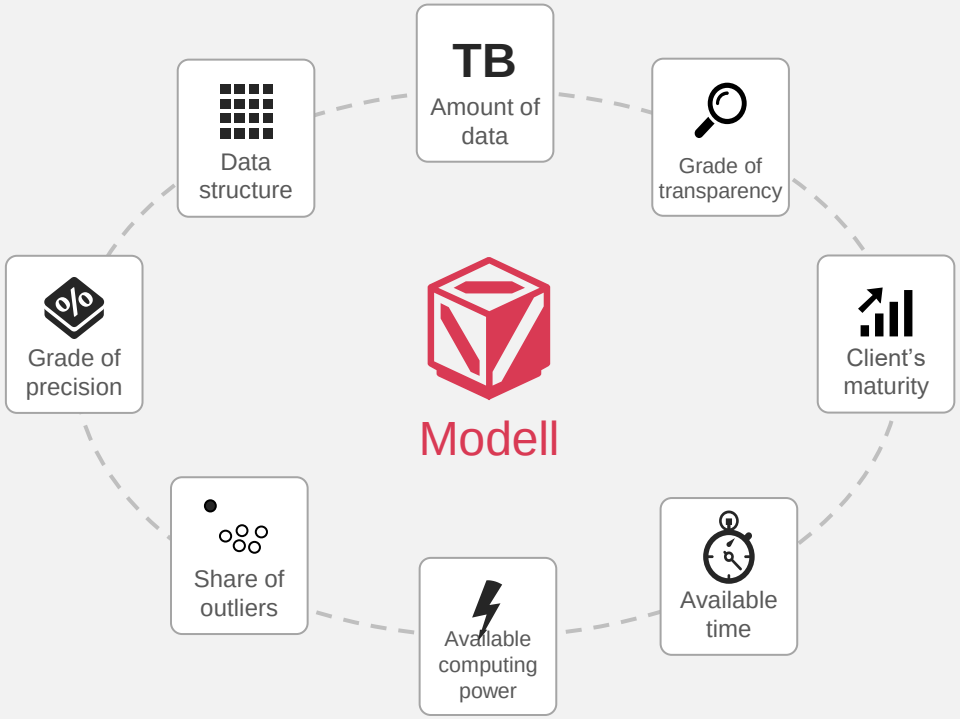


$$\text{Market value} = \frac{(4.3 + 3.8 + 4.1)\ \text{mnok}}{3} = 4.1\ \text{mnok}$$

There are many different machine learning models. The right model choice depends on both the problem to be solved and the customer's maturity.



- Linear Regression
- Logistic Regression
- Naive Bayes
- Support Vector Machines
- Decision Trees
- Random Forest
- Boosted Trees
- K-Nearest Neighbour
- Neural Networks
- +++



Using machine learning, the project have developed a model that is more precise than the current asset values and where the deviations are more evenly distributed

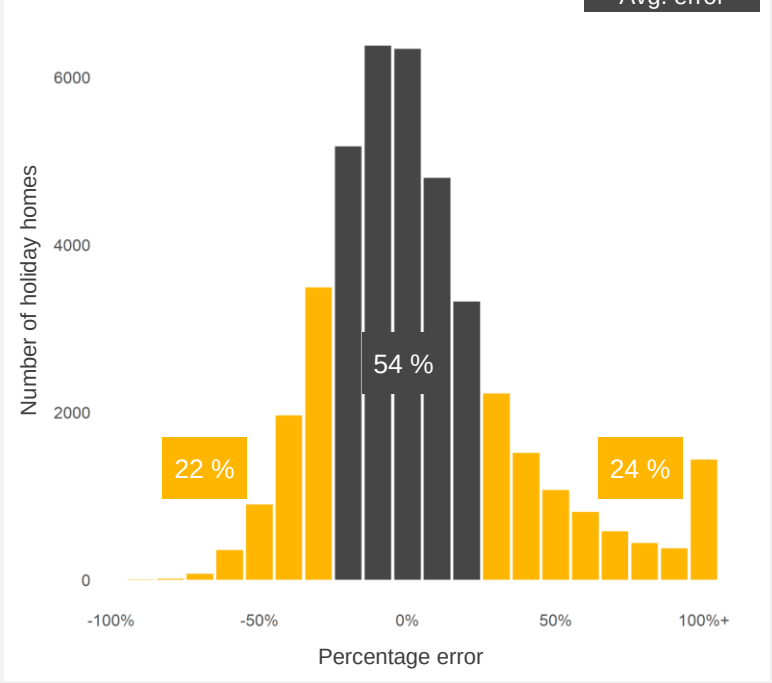
DATA

MACHINE LEARNING

EXPLAINABILITY

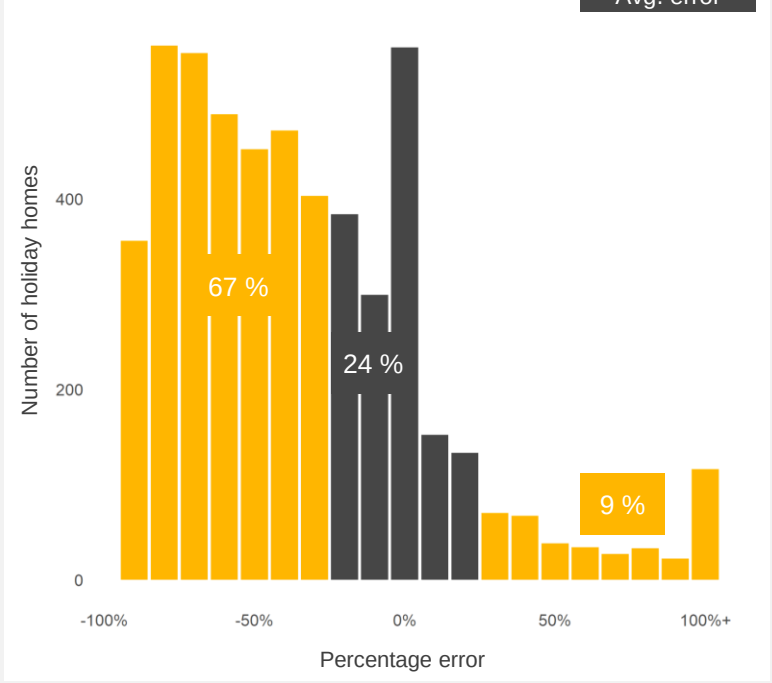
Machine learning model

26%
Avg. error



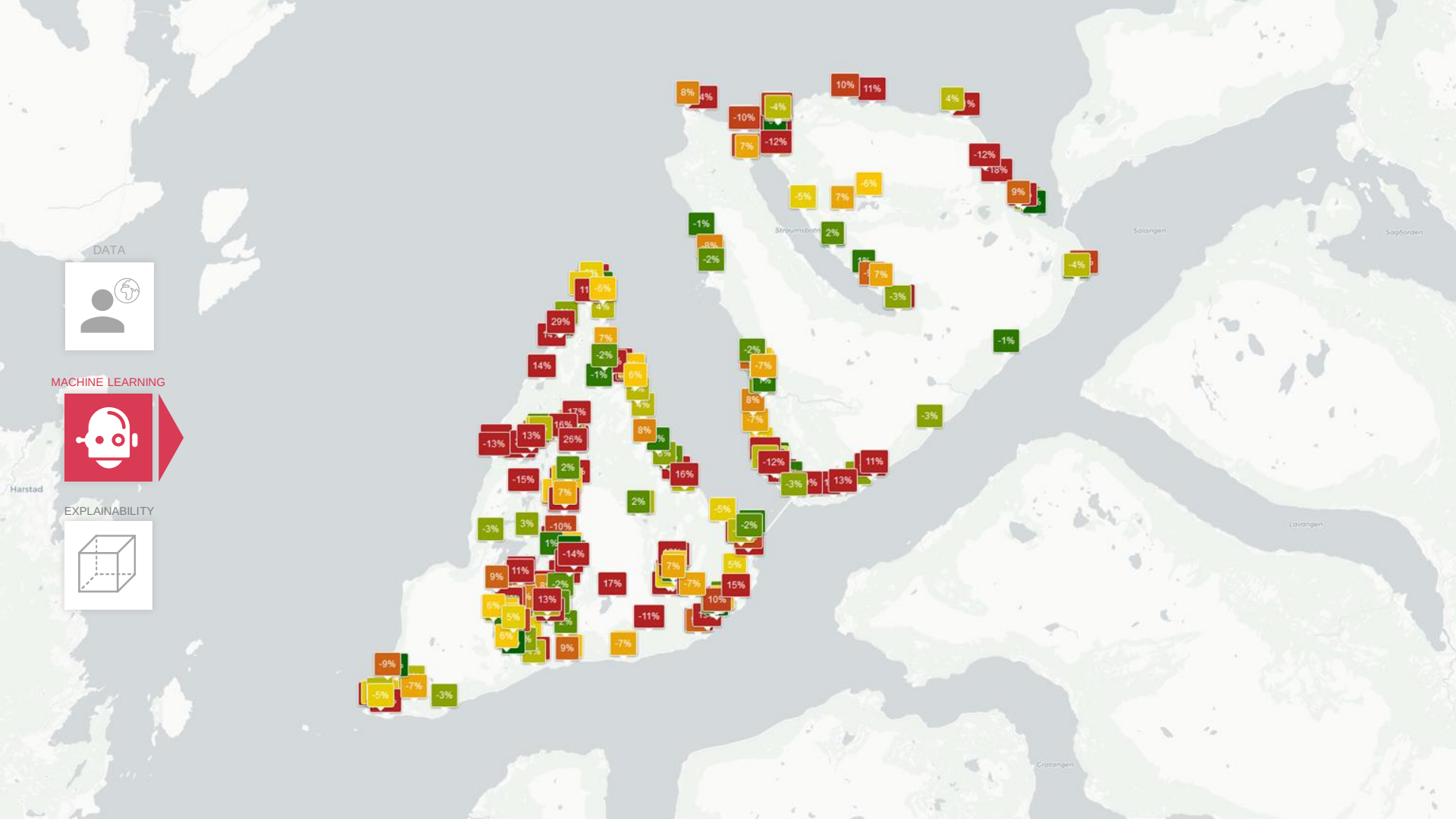
Asset values 2019

48%
Avg. error



REDACTED

REDACTED

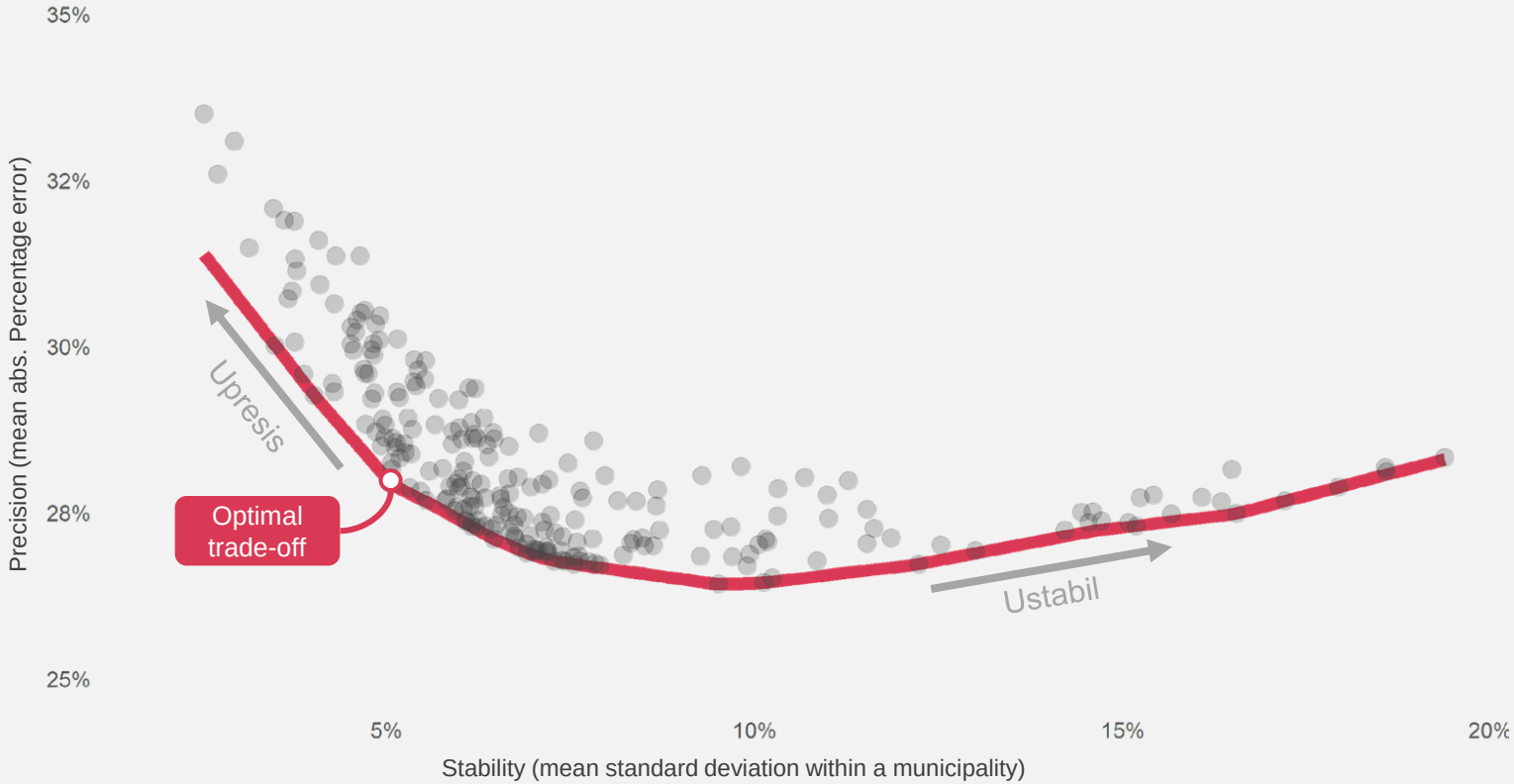

DATA
MACHINE LEARNING
EXPLAINABILITY
Harstad
Straumsbotn
Salangen
Sagfjorden
Lavangen
Gratangen

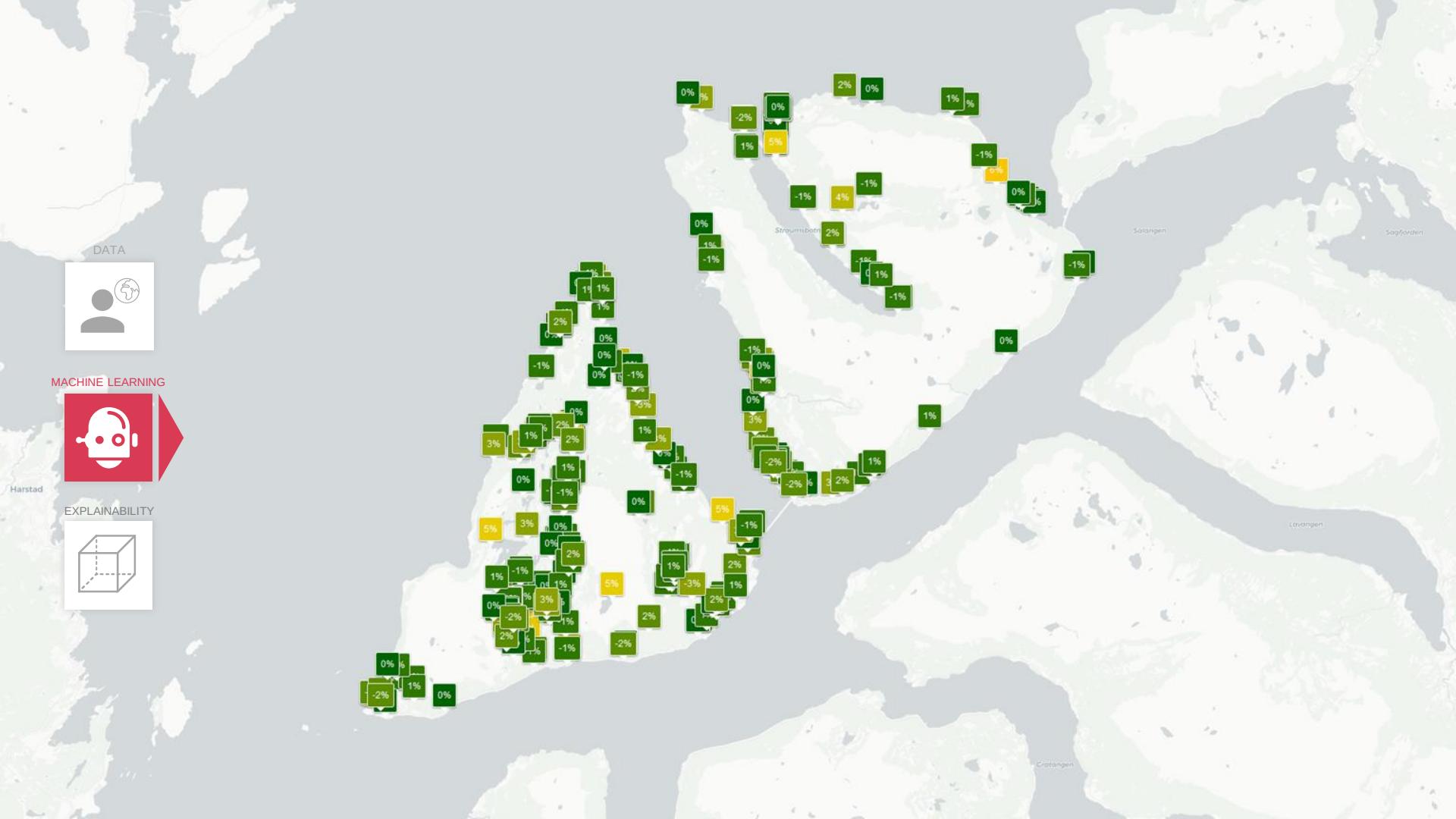
We used hyperparameter tuning to find the optimal trade-off between precision and stability

DATA

MACHINE LEARNING

EXPLAINABILITY




DATA
MACHINE LEARNING
EXPLAINABILITY
Harstad
Straumsbotn
Salangen
Sagfjorden
Lavangen
Gratangen

Valuation models must be explainable to taxpayers

1

New data sources for valuation

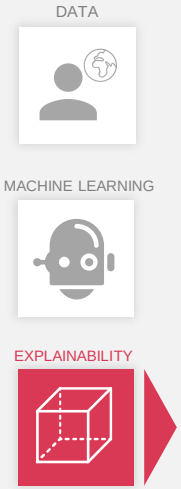
2

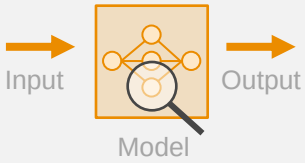
Machine learning for valuation

3

Explainability of machine learning

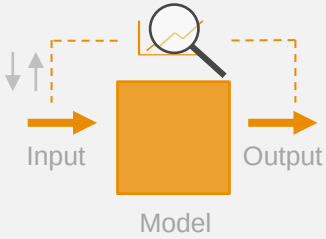
Three different methods to explain models (Molnar, 2020)





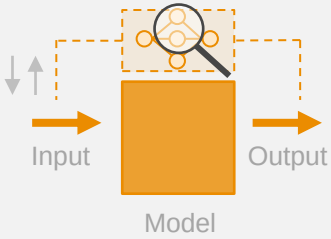
Methods analysing model components

Linear models and decision trees are examples of naturally explainable models that can be explained by analysing components of the models, ie. model structure and parameters



Methods analysing model sensitivity

The methods are typically model agnostic, where the model is considered a closed system. The methods analyse how the estimates are affected by changes in the variables values.



Methods based on surrogate models

Surrogate models are naturally explainable models designed to copy the behaviour of unexplainable models. The methods are based on the explanations of the surrogate model


DATA
MACHINE LEARNING
EXPLAINABILITY

SHAP values makes it possible to explain how the various price drivers affect the final prediction for each individual holiday home

DATA

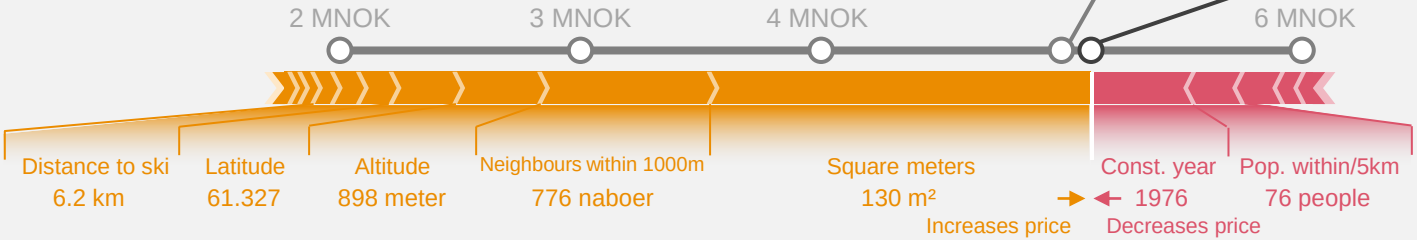
MACHINE LEARNING

EXPLAINABILITY

Location	Trysil
Area	130 m²
Altitude	898 m
Constr. year	1976
Distance to ski trail	6.2 km
Distance to road	18 m
Distance to sea	175 km
Distance to lake	10 km
Neighbours	776 in 1000 m
Closest neighbour	31 m
Population	76 in 5 km

5.0 MNOK
SALES PRICE

5.2 MNOK
VALUATION

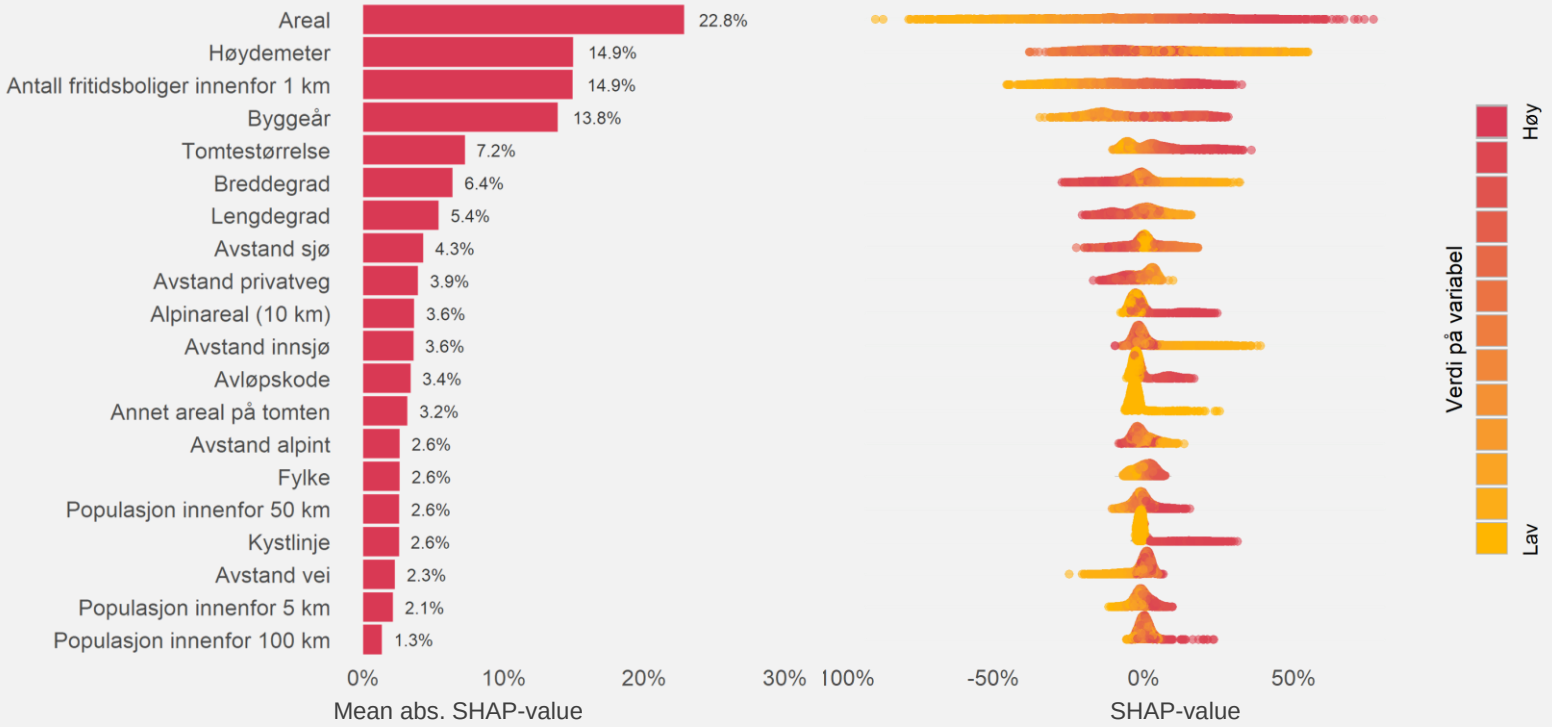


By aggregating the SHAP values, the most important variables can be derived. Structural variables such as area, build year and plot size explains less than 50 % of the valuation on average

DATA

MACHINE LEARNING

EXPLAINABILITY



By analyzing how complex machine learning algorithms emphasise different variables, it is possible to gain insight on the underlying patterns in the data

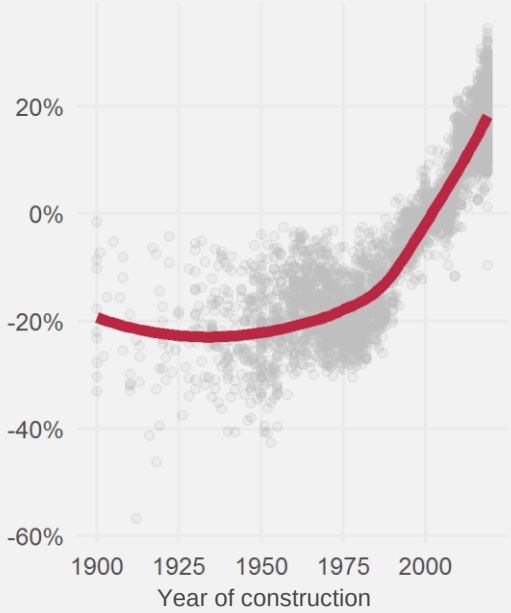
DATA

MACHINE LEARNING

EXPLAINABILITY

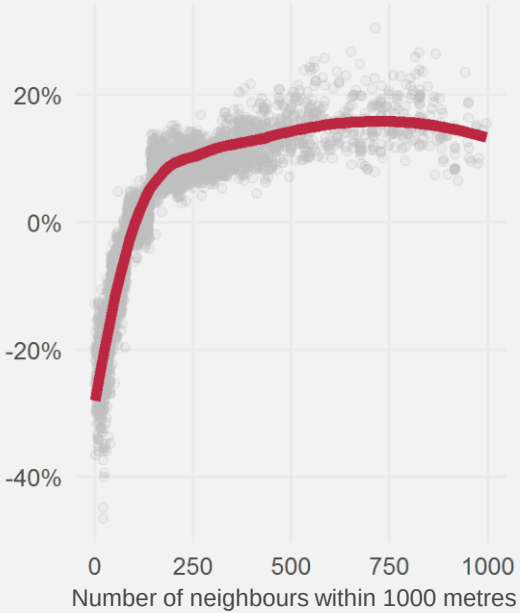
Year of construction

If the house is **older than 50 years old** its actual age matters less



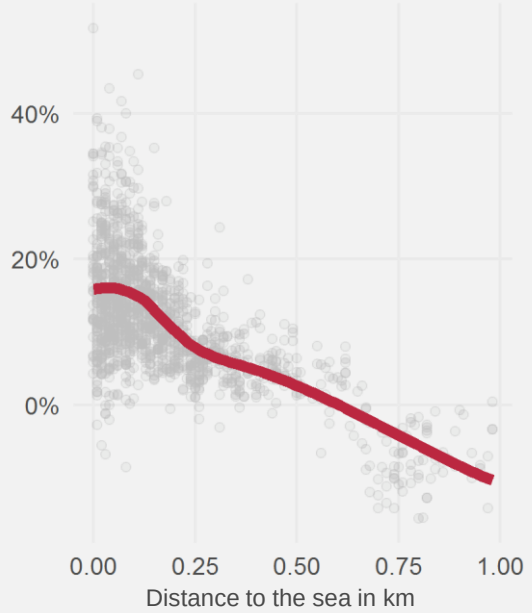
Number of neighbours

Several **neighbours within 1000 m** gives a sharp increase in value that flattens out at 250



Distance to the sea

Proximity to the sea can for some holiday homes constitute **an increase in value of up to 50 %**



03
Code along and exercises

Case:
Buying a house

Apply for relevant positions at pwc.no/karriere

Full-time position

Relevant for	5th grade
When	August 2022
Where	Bergen
Deadline	October 10th

Data & Analytics Internship

Relevant for	4th grade
When	January 2022
Where	Bergen
Deadline	TBA

63