Intro in ML. Part 2

Prof E. Burnaev

A. Zaytsev

Skoltech

Level of data analytics for client care

Descriptive analytics:
what has happened?

- What types of customers did buy?
- How much money did we make?
- Why did they buy?

Predictive analytics:
what will happen?

- Which customers will buy?
- How much money will we make?
- Why would they buy?

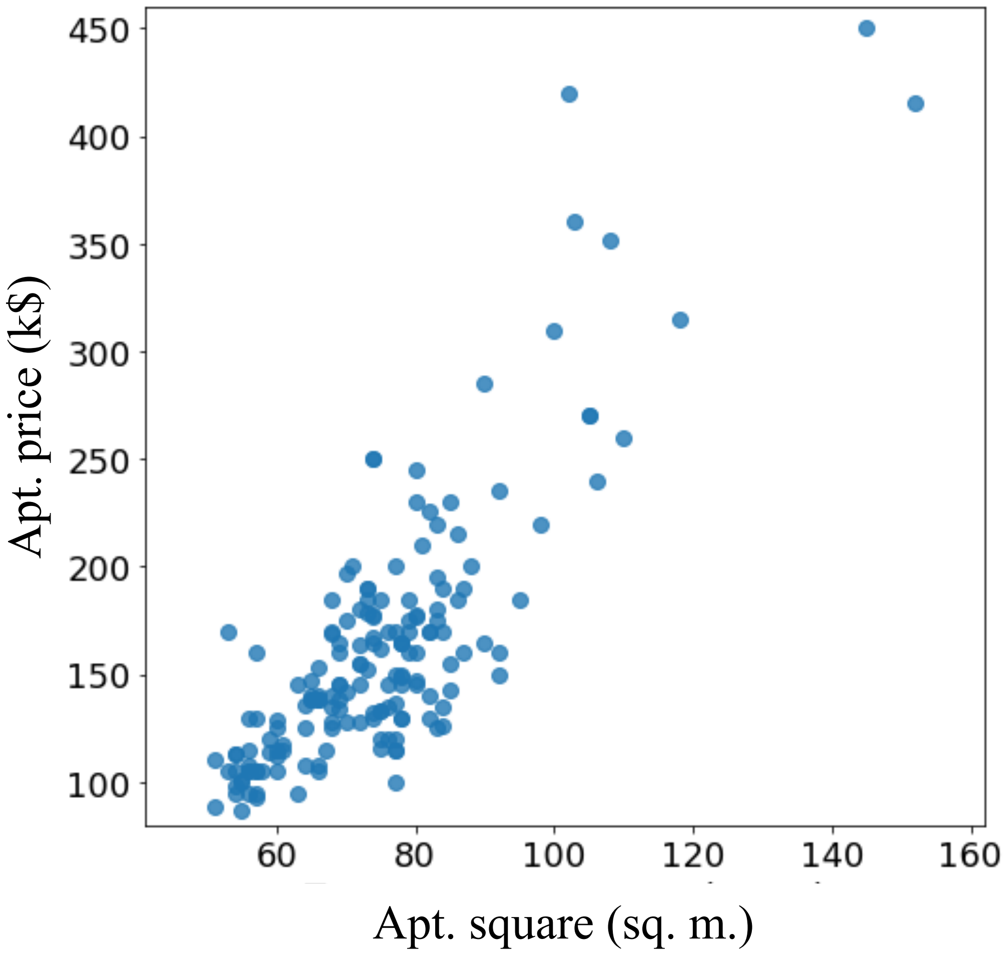
Prescriptive analytics:
what should we do about it?

- How to influence the customer to buy?
- When does it make sense to influence in terms of resource optimization?

Machine Learning – a tool to solve real-world problems

- Credit scoring: understand whether a person will be a reliable customer based on the available data
- Detection of accidents on the rig
- Video analytics: use the video to understand whether the staff is working in helmets
- Prediction of well debit by hydraulic fracturing parameters

Important parts of a machine learning problem statement

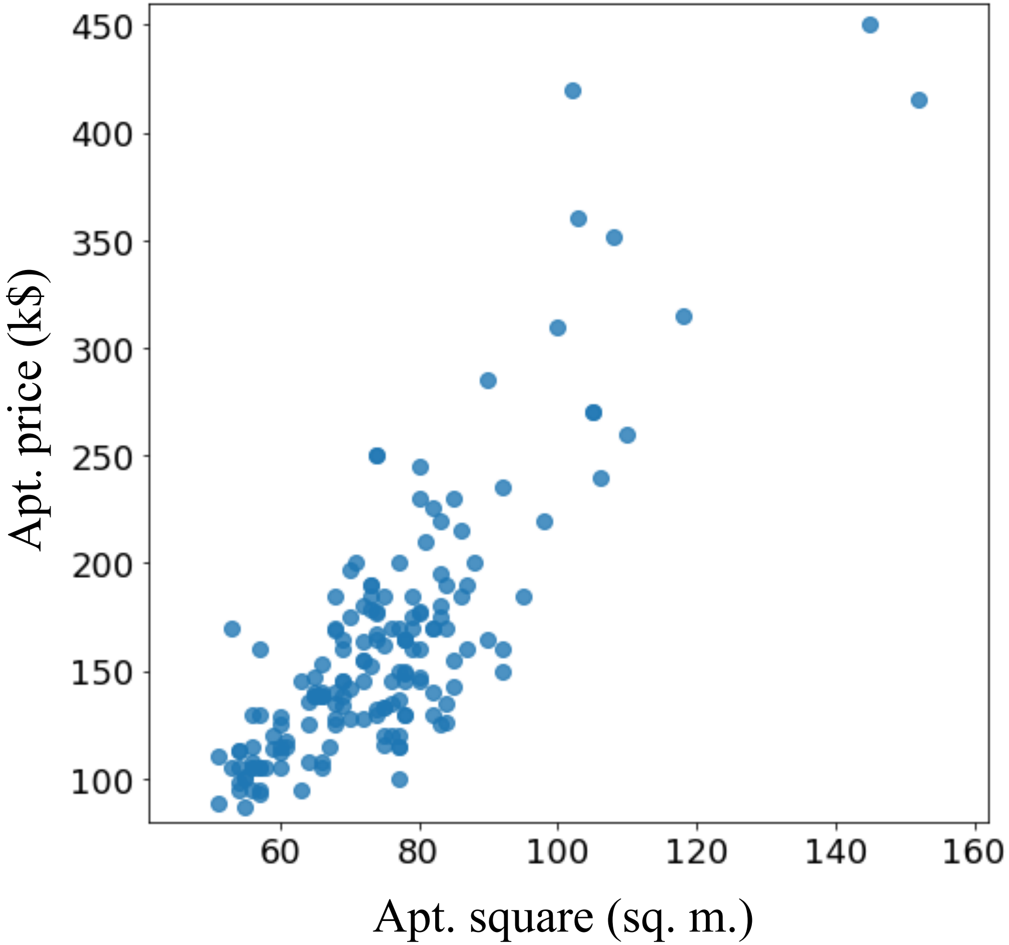


Standard ML problem (supervised learning): construct a model of how «output» depends on «input»

Input: **x**, apartment square

Output: **y(x)**, apartment price

Each point on the plot –
one object in a learning sample



Skoltech
Skolkovo Institute of Science and Technology

We have observations, table data

Input: **x,** apt. square

Output: **y(x)**, apt. price

Data:
this was the data from ads published the last year

Features and target variable

Objects

x		y
Square sq.m.	Distance to a downtown, km	Price thousands $
77	9	115
79	9	175
84	11	170
65	8	140

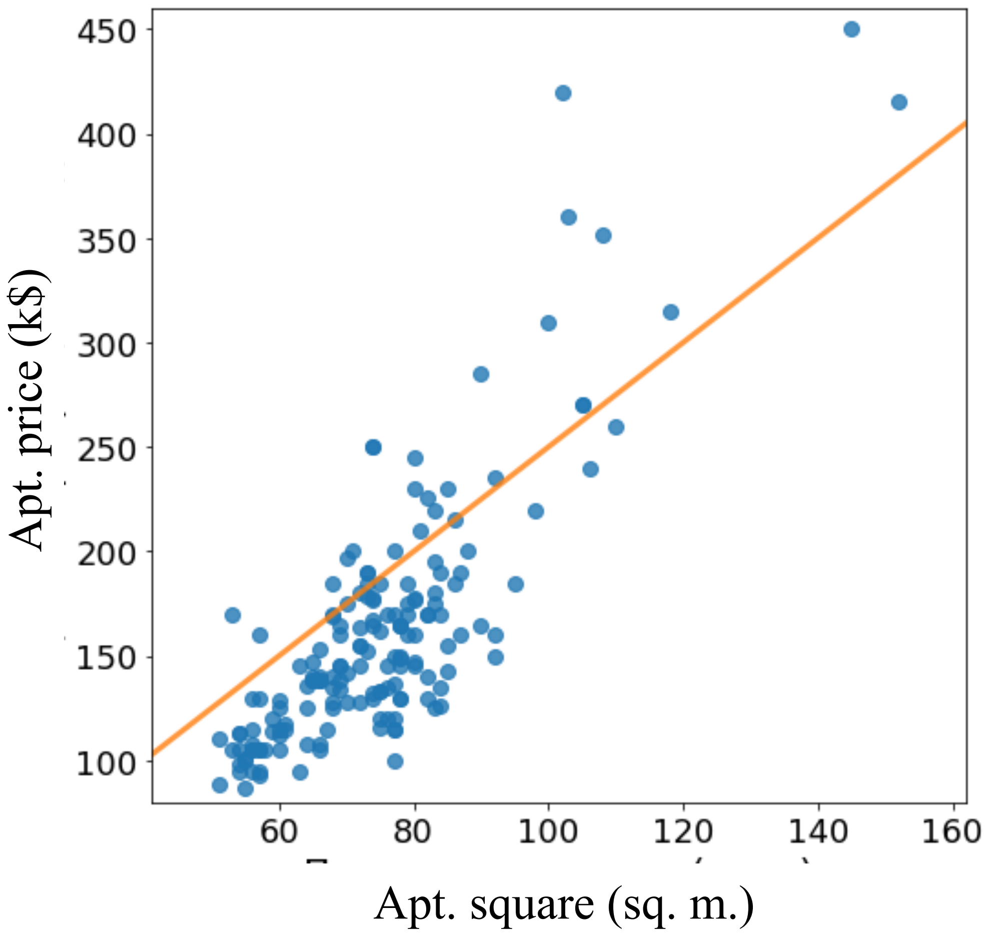
Table data:
«Excel-table»

Construct a model of how «output» depends on «input»

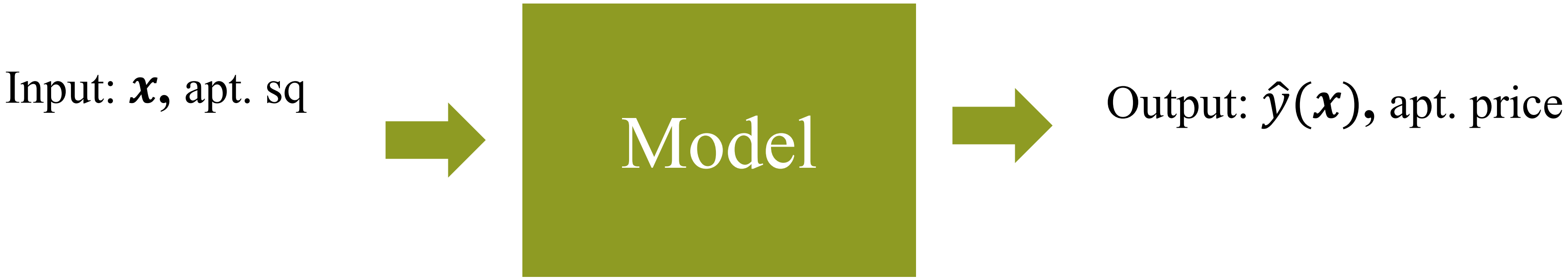
Input: **x,** apt. sq

Output: **y(x)**, apt. price

Model $\hat{y}$**(x)**:
If apt. sq is 100 sq.m, then its price is 250 k$



Problem solution: regression model



Input: $\boldsymbol{x}$, apt. sq

Model

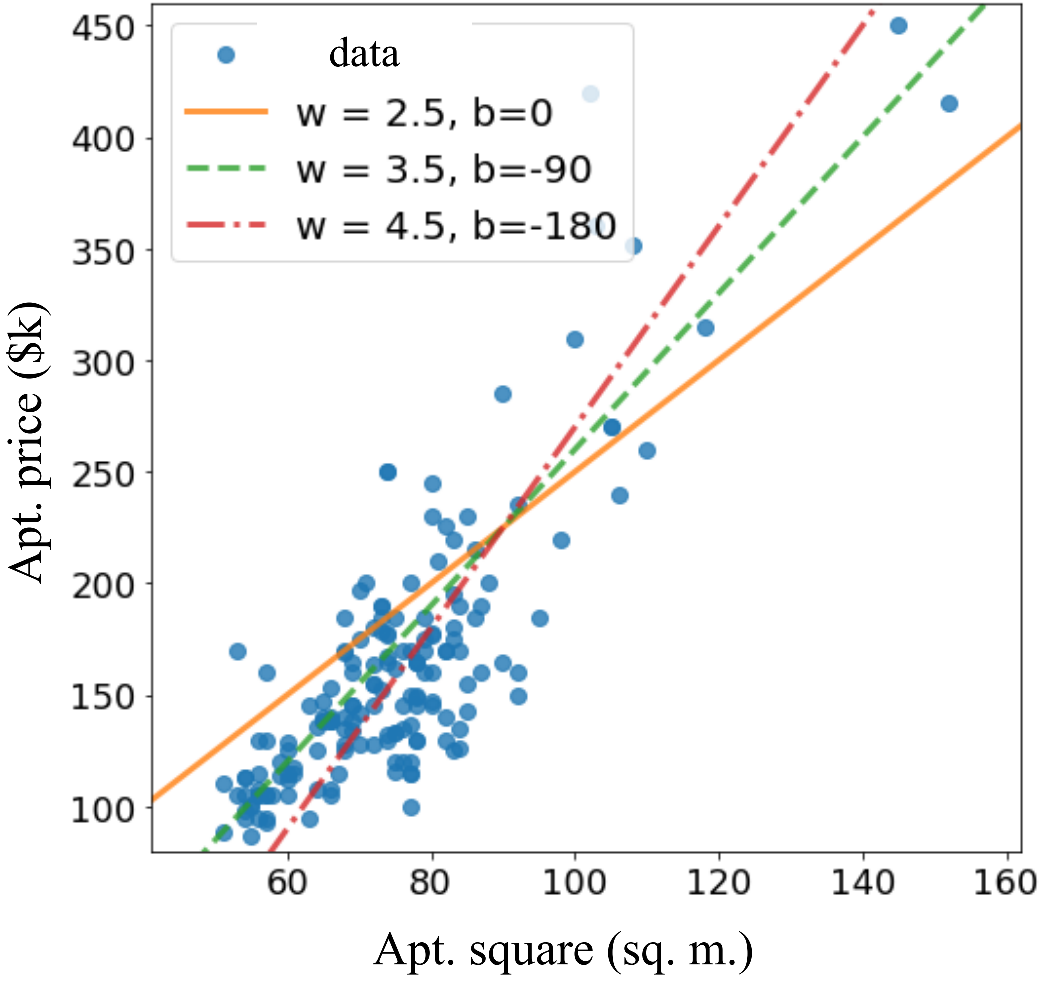
Output: $\hat{y}(\boldsymbol{x})$, apt. price

Problem solution: linear model

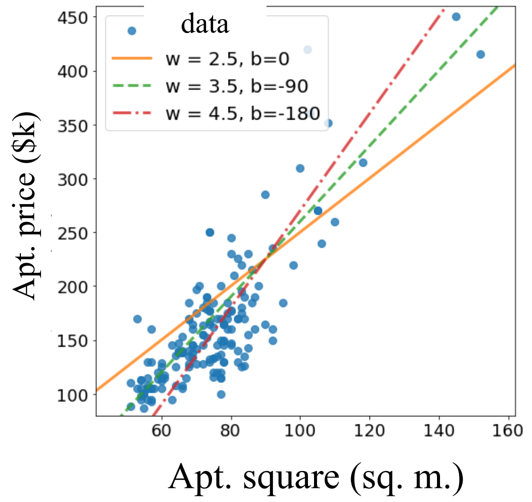
Example of a model – a linear regression model

$$\hat{y}(x) = w\,x + b$$

- y – real apt. price
- $\hat{y}(x)$ – model prediction
  x – apt. sq.
- w, b – coefficients (parameters) of a linear regression model



Examples of predictions by two different regression models



$$\hat{y}(x) = w\,x + b$$

y – apt. price
x – apt. sq.
w, b – coefficients (parameters) of a linear regression model

Machine learning is how to select such parameters that provides the best prediction accuracy

x	y	$\hat{y}(x)$, w = 2.5, b = 0	$\hat{y}(x)$, w = 3.5, b = -90
77	115	192.5	179.5
79	175	197.5	186.5
84	170	210.0	204.0
65	140	162.5	137.5

Model accuracy?!

We would like that predictions $\hat{y}(\boldsymbol{x})$
are similar to real values $y(\boldsymbol{x})$

Discrepancy between the prediction and the real value - squared error

$$\mathrm{SE(x)} = (\hat{y}(\boldsymbol{x}) - y)^2$$

Mean squared error

$$\mathrm{MSE} = \frac{1}{n}\sum_{i=1}^{n}(\hat{y}(\boldsymbol{x}_i) - y_i)^2$$

$\boldsymbol{x}_i$	y_i	$\hat{y}(x)$	SE
77	115	122.5	56.25
79	175	177.5	6.25
84	170	173.0	9
65	140	145.0	25

MSE = 24.275

We can use different loss functions

The bigger errors the model $\hat{y}(x)$ makes on specific objects, the worse it is

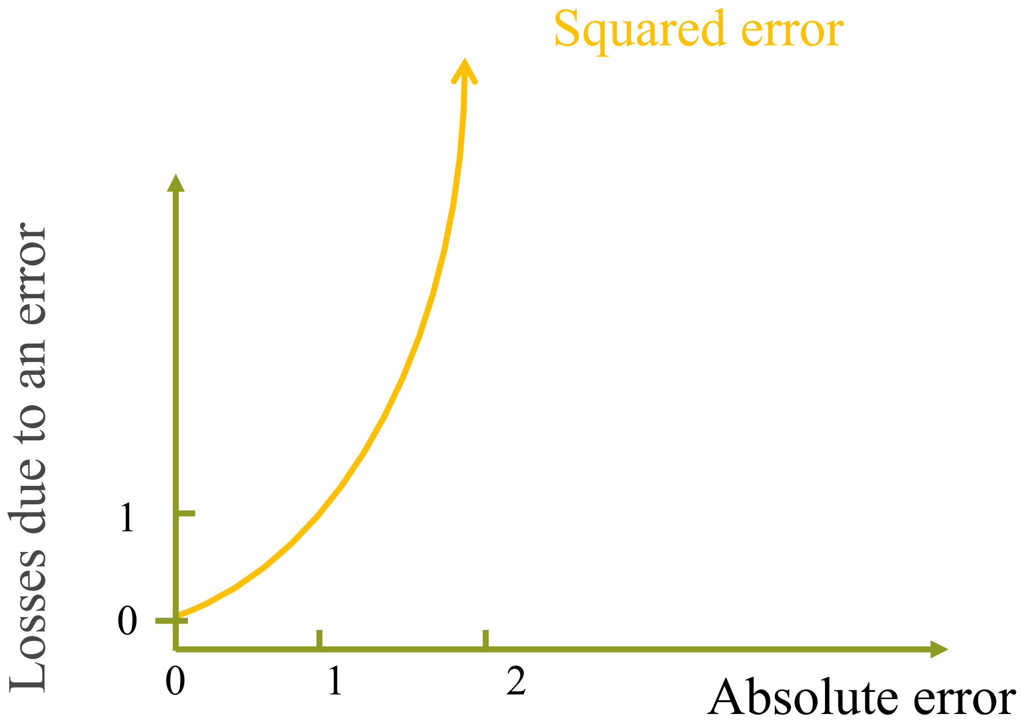
Absolute error

$$AE = |\hat{y}(x) - y(x)|$$

Squared error

$$SE = \left(\hat{y}(x) - y(x)\right)^2$$

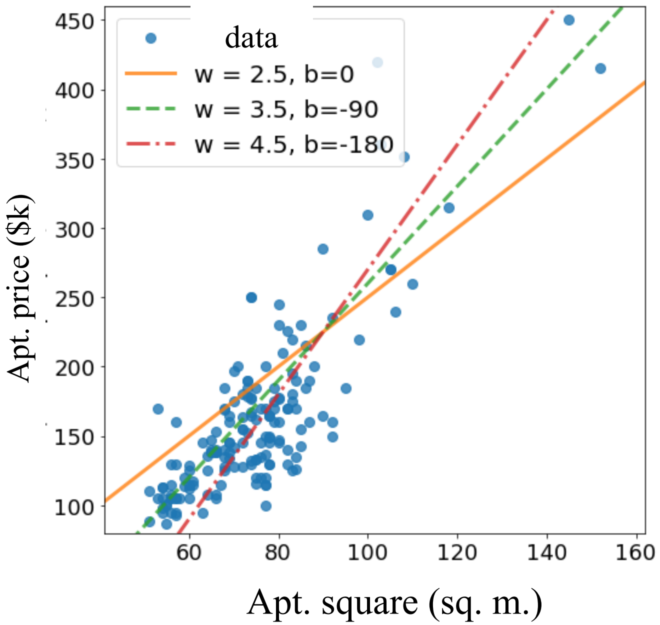
Due to mathematical convenience and general adequacy, the mean-squared error is usually used



Let us calculate the loss function for the regression

$$\hat{y}(x) = w\,x + b$$

y – apt. price
x – apt. sq.



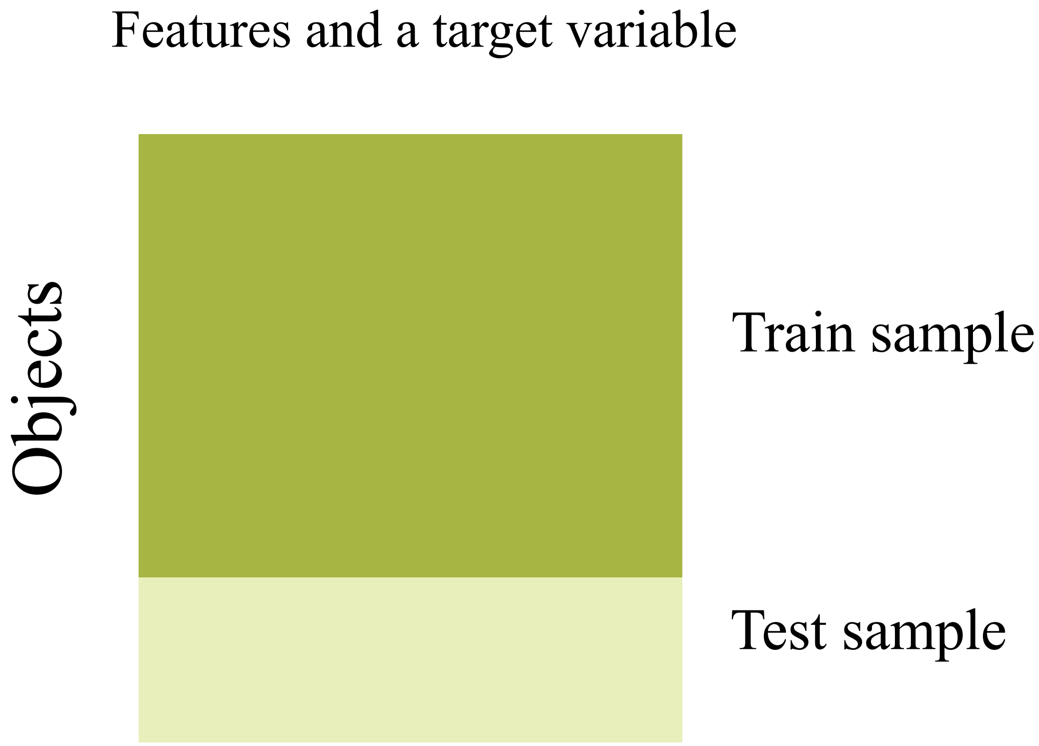
x	y	$\hat{y}(x)$	Threshold loss function	Squared error
77	115	192.5	1	6006.25
79	175	197.5	1	506.25
84	170	210.0	1	1600
65	140	162.5	1	???

Skoltech
Skolkovo Institute of Science and Technology

We can't train a model and test it on the same data, so we use an independent test sample

- We use independent test sample
- We calculate mean error using this sample

- The loss function is not a business metric of solution quality



Main parts of the data analysis task problem statement

1. What do we want to predict? What is the input and what is the output?	1. We predict the price of an apartment given its sq.
2. What data is available?	2. We have data for the last year
3. Which model class do we use?	3. We construct a linear model
4. How to estimate quality of the solution?	4. We would like to minimize a squared loss function

Main parts of the data analysis task problem statement

1. What do we want to predict? What is the input and what is the output?
2. What data is available?
3. Which model class do we use?
4. How to estimate quality of the solution?