HOGESCHOOL UTRECHT

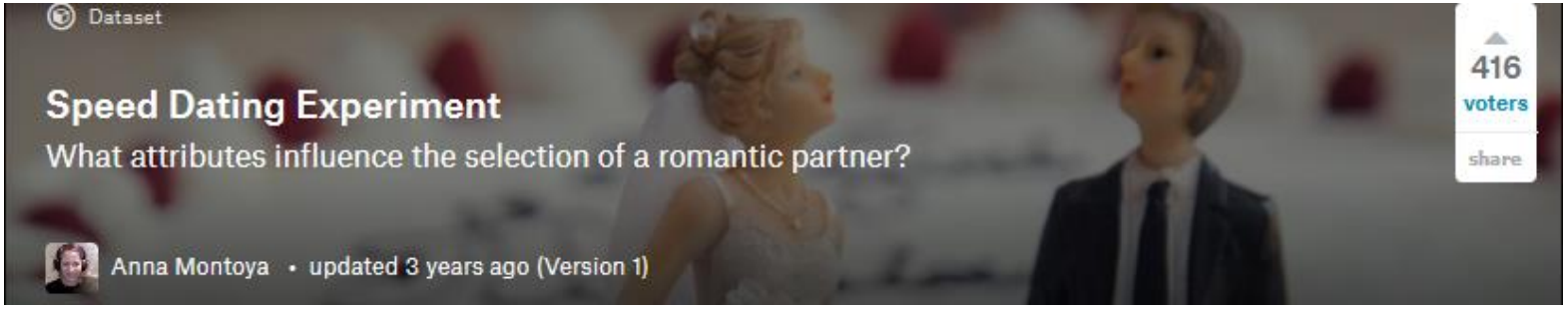
Fundamentals of Machine Learning

Week 4: recap linear regression

Introduction machine learning

Jonas Moons

Intro


Dataset
Speed Dating Experiment
What attributes influence the selection of a romantic partner?
Anna Montoya • updated 3 years ago (Version 1)
416
voters
share

Feedback assignment #2

- Keep on documenting your analysis:
  - Use Markdown
  - Headings
  - Introductions
  - Intermediate steps / processing
  - Conclusions

- Don't use .fillna(0). Use .dropna() instead!

Research project with AI group

- Building on a project to predict fake news tweets
- Lecture and Q&A from Stefan Leijnen, professor of the research group. 8 December 11:00-11:45
- Work individually but also as a team
- More (advanced) work on feature engineering & model building
- Some more supervision from me
- Present your work to AI research group at the end
- Details on Canvas tomorrow. You can indicate if you want to join in your proposal.

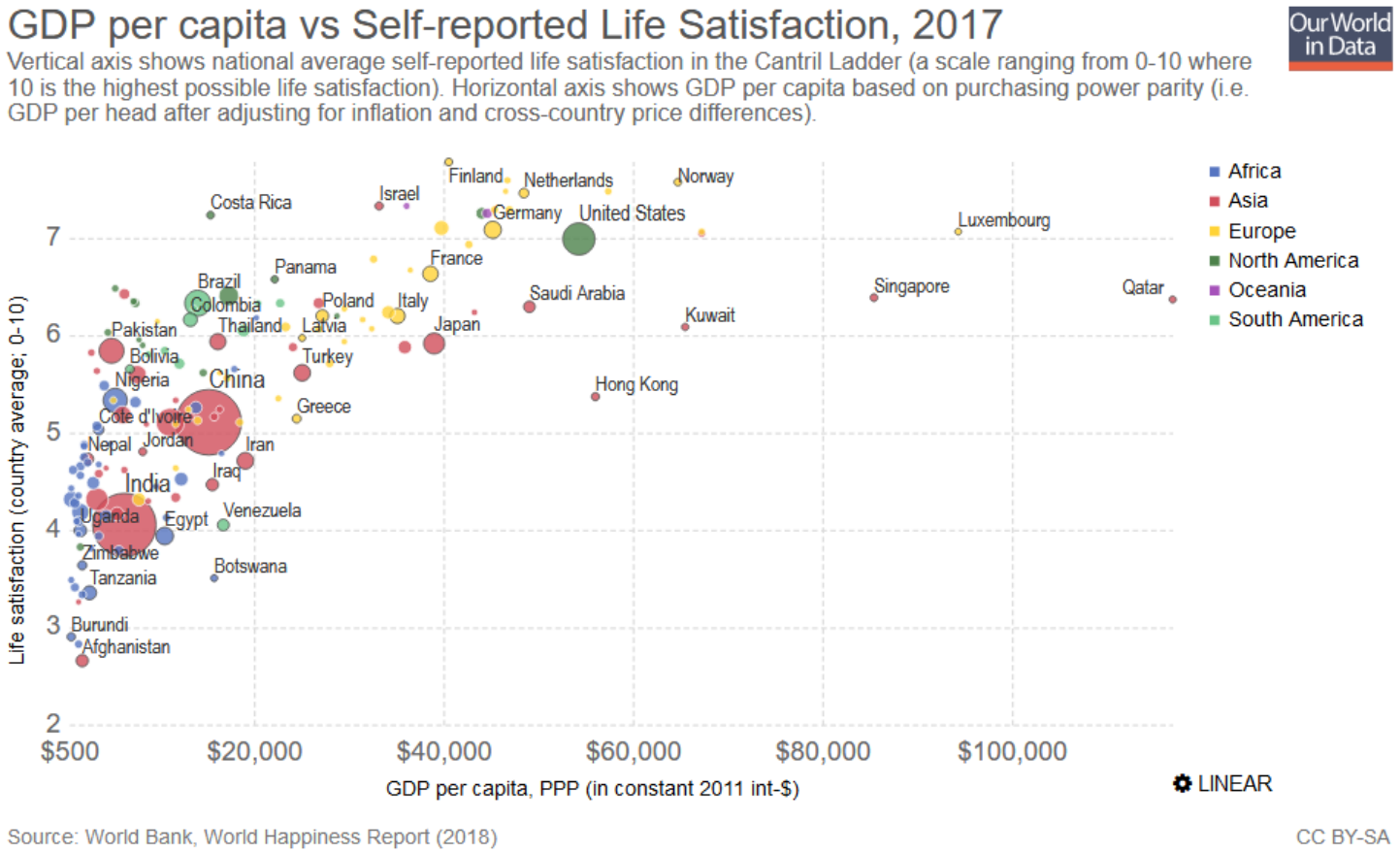
Topics

- Variable transformations
- Recap and recap exercise
- Machine learning
- *k*-nearest neighbor algorithm

Variable transformations

- Variable transformations are a kind of **feature engineering**
- You transform information the algorithm can't use very well into something it can
- Example: in a linear regression, how would you predict traffic jams (km/year) in a country from:
  - Population (millions of inhabitants)
  - Land area (km^2)

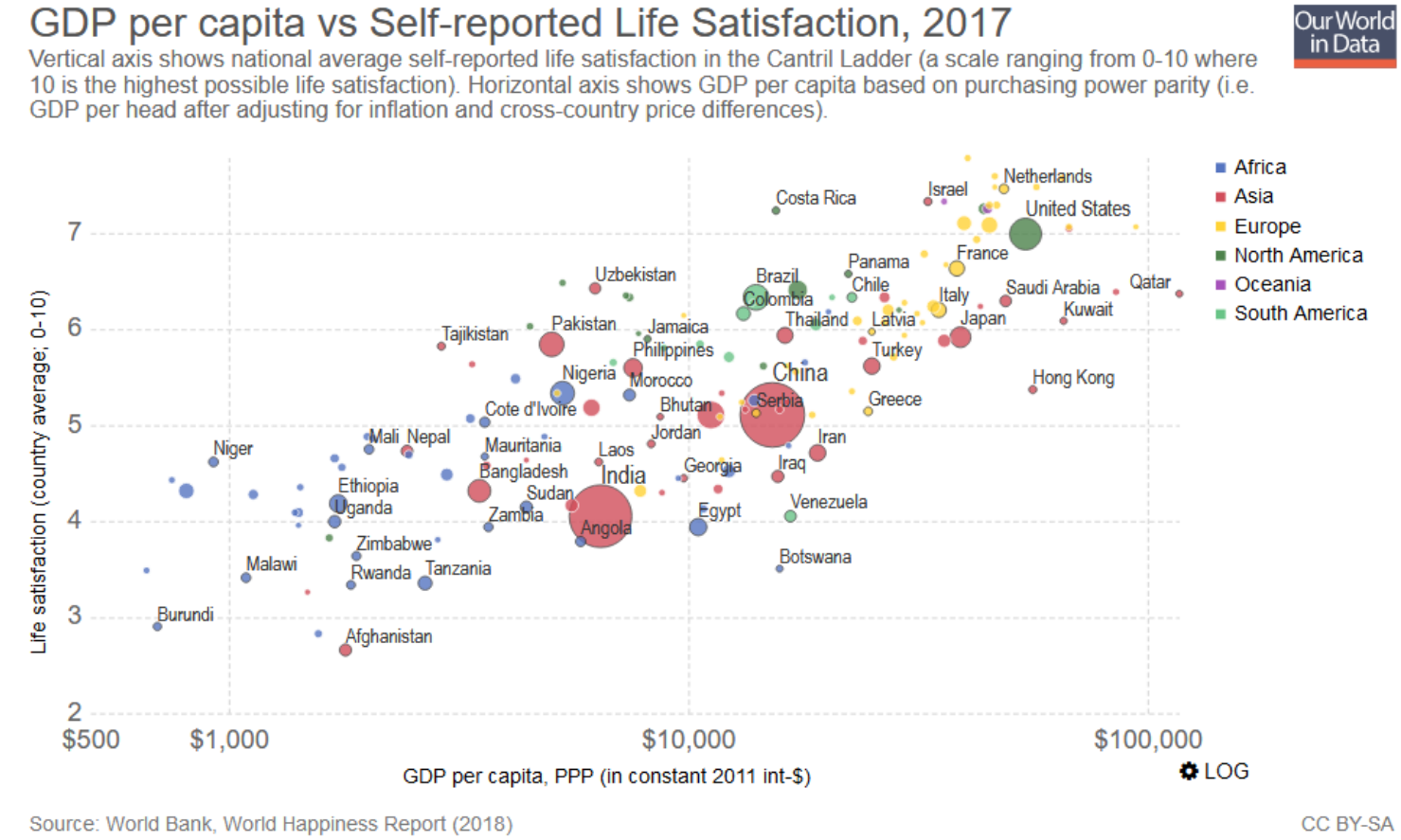
Log transformation



GDP/capita and self-reported life satisfaction don't show a linear relation – instead, logarithmic

Ourworldindata.org, Max Roser (CC-BY-SA)

Log transformation



If we take the logarithm of GDP/capita, we get a linear relationship

Ourworldindata.org, Max Roser (CC-BY-SA)

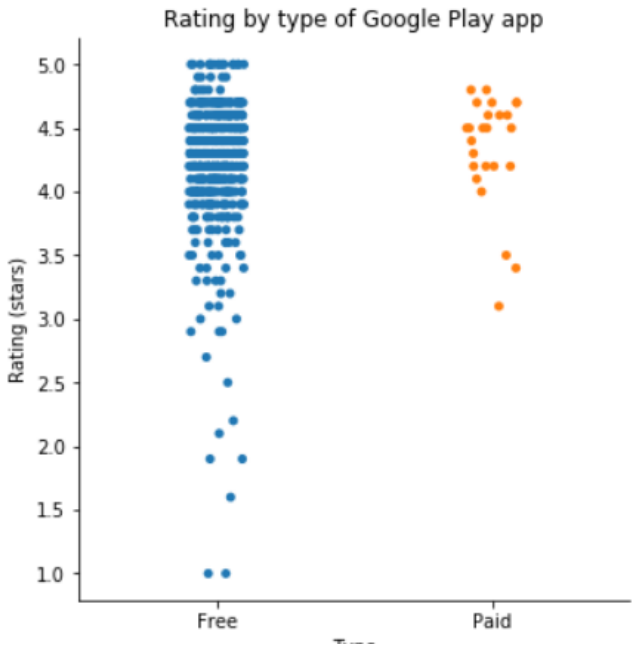
Topics

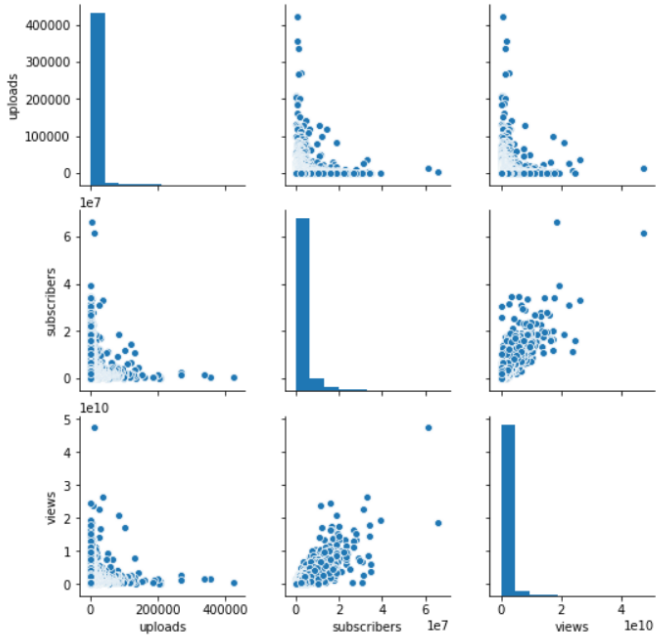
- Variable transformations
- Recap and recap exercise
- Machine learning
- *k*-nearest neighbor algorithm

As a data scientist, you…

- Clean and explore data
- Look for patterns in data, investigate relations, e.g.
  - Find types of users that use your website differently
- Build statistical models to make predictions to add value to your organization, e.g.
  - Predict the success of a new video ad
  - Approach the users most likely to make a purchase

Data exploration





Plot the distributions of qualitative variables to understand the differences

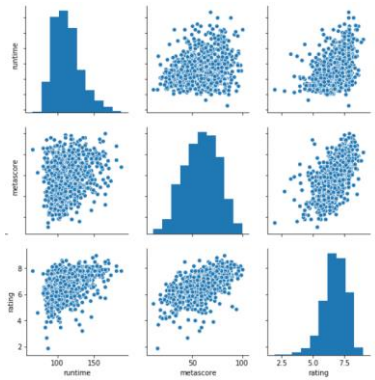
Scatterplots and scatterplot matrices are a great way to understand the relations between quantitative variables

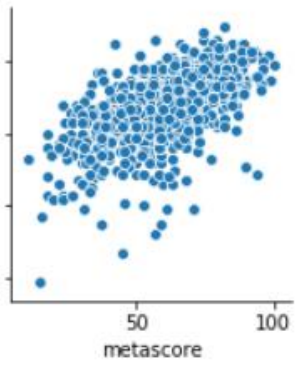
Linear model

$$Y = b_0 + b_1 X_1 + b_2 X_2 + \cdots + e$$

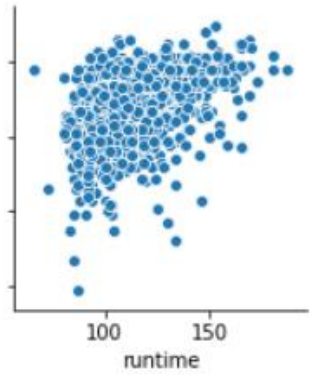
Build

- Select variables
- Make dummy variables for qualitative variables
- Check linearity, consider transformations





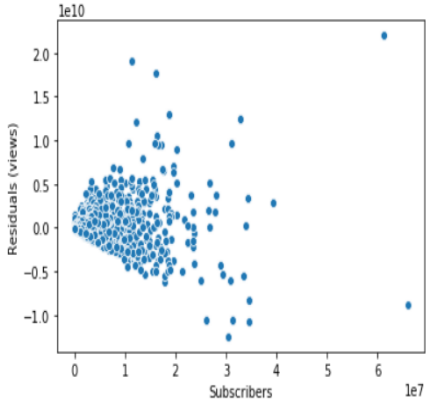
Linear! 👆



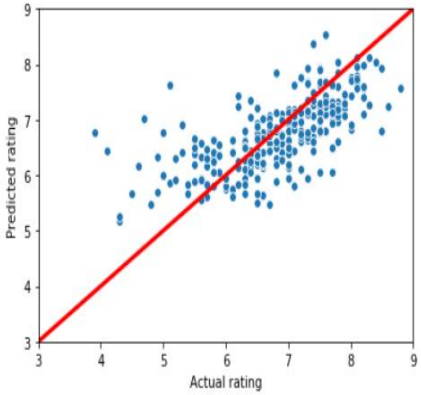
Consider log?

Evaluate

- Create train and test data set. **Evaluate on the test set.**
- Calculate:
  - R^2: proportion of variance explained
  - RMSE: typical error in prediction
- Plots:



Residual plots



Y-Y' (actual vs. predicted)

Exercise 1: recap exercise

This is a recap exercise to get you working on your own as a data scientist.

Use the example Notebooks from previous weeks and/or the *Cookbook* Notebook in the main folder.

A Notebook to start with can be found in *exercises/recap_linear_regression.*
The data can be found in the same folder. The data are artificial, but some relations have been built in. The data are on frequent bol.com shoppers.

Your task is to create a linear model that predicts the money spent on bol.com per year. Try and find the optimal model.

Then, predict the variable *spent_bol* for the hold-out test in the folder. You are no longer allowed to change your model, so do this at the end!

Some tips:

- Use the example Notebooks and *Cookbook* to get snippets of code that you need
- Use graphs to explore the data
- Think of possible transformations
- Calculate the performance of your model with RMSE
- Beware of overfitting…
- Best performance wins a prize!

Topics

- Variable transformations
- Recap and recap exercise
- Machine learning
- *k*-nearest neighbor algorithm

Machine learning

- ‘the study of algorithms and statistical models that computer systems use to progressively improve their performance on a specific task.’ (Wikipedia)

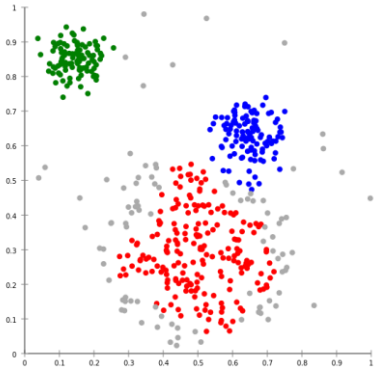
- Like traditional statistics, but…
  - More variables and cases
  - Complicated and expensive (in terms of computing power) models/algorithms
  - Fitting every ‘nook and cranny’ of the data

Supervised vs. unsupervised learning

- Supervised: use known patterns to predict new cases
  - Handwriting recognition

MNIST data set

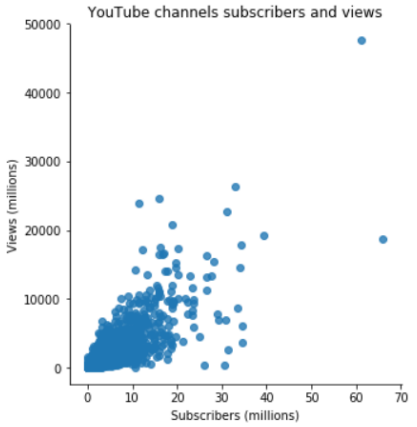
- Unsupervised: you let the algorithm discover patterns/clusters on its own
  - Spotify Radio / Discover Weekly



Spotify®

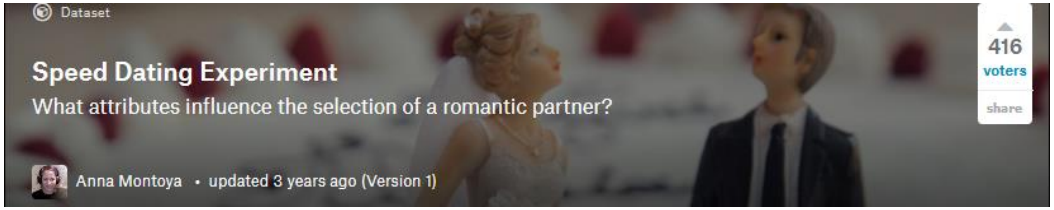
Classification vs. regression

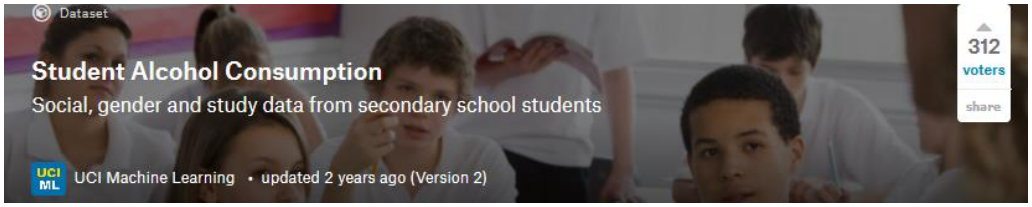
- Within supervised learning, there are two types:
- Classification: categorical dependent
- Regression: numerical dependent

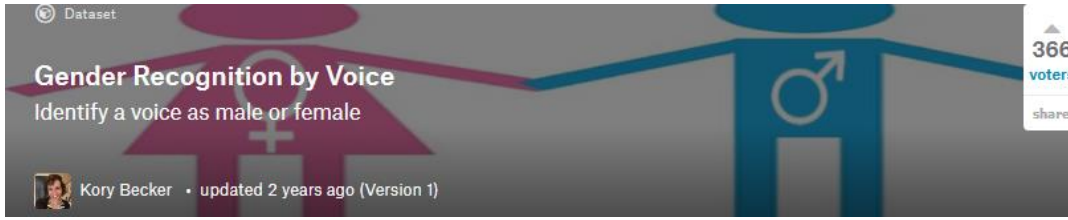


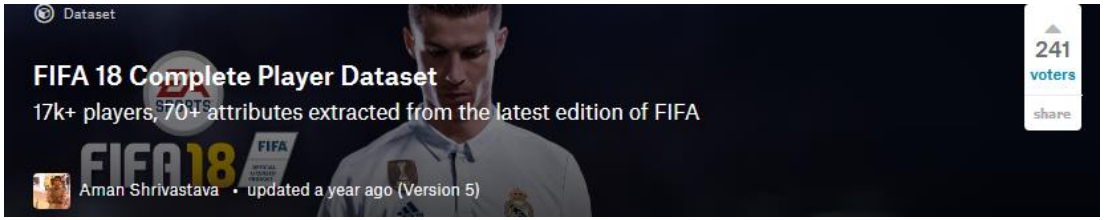
Weekly assignment

kaggle









Excercise 2: Kaggle

Go to www.kaggle.com:

1. Have a look at the possible data sets for the weekly assignment

2. Discuss with your neighbor which data set you would like to try.

Topics

- Variable transformations
- Recap and recap exercise
- Machine learning
- *k*-nearest neighbor algorithm

k-nearest neighbor: intuition

Who is your favorite?

Cristiano Ronaldo or **Lionel Messi**

k-nearest neighbor algorithm

- *k*-nearest neighbor is one of the simplest algorithms in machine learning

- From the data set, pick the *k* nearest neighbors (*k* = 3, 5, 7, etc.) of the individual you want to predict for
  - Classification: pick the most frequent answer (e.g., Ronaldo or Messi)
  - Regression: take the mean of the neighbors (e.g., apps downloaded last year)

- What is the obvious ‘problem’ with this algorithm?

Distance

- How do you calculate a ‘distance’ between individuals?
- Take ‘Euclidean distance’
  - A ‘distance’ between two individuals can be calculated for any number of dimensions/variables
- Treat all variables the same
  - *Normalize* all variables so that they are all on the same scale (mean = 0, sd = 1)

Id	nr phones	nr tablets	Apps donwloaded
1	2	2	**50**
2	1	2	**42**
3	1	1	**23**

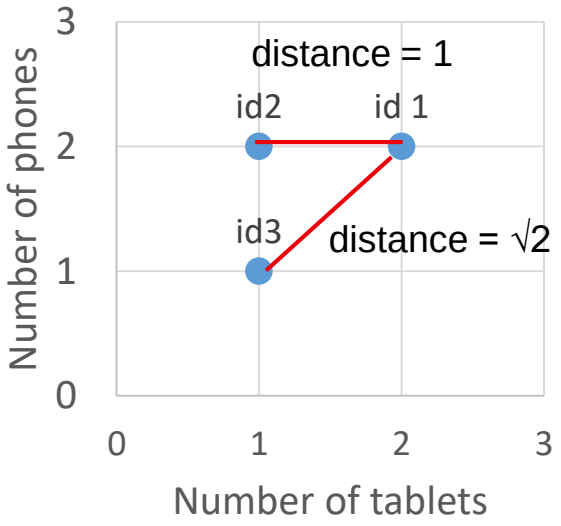


Image credit

- GDP vs. happiness (data: Worldbank, visualization by Max Roser, CC-BY-SA)
- MNIST data set by Josef Steppan (CC-BY-SA)
- Spam by Etonic (CC-BY-SA)
- Clusters by Chire (CC-BY-SA)
- Cristiano Ronaldo by Ruben Ortega (CC-BY-SA)
- Lionel Messi by Кирилл Венедиктов (GNU license)