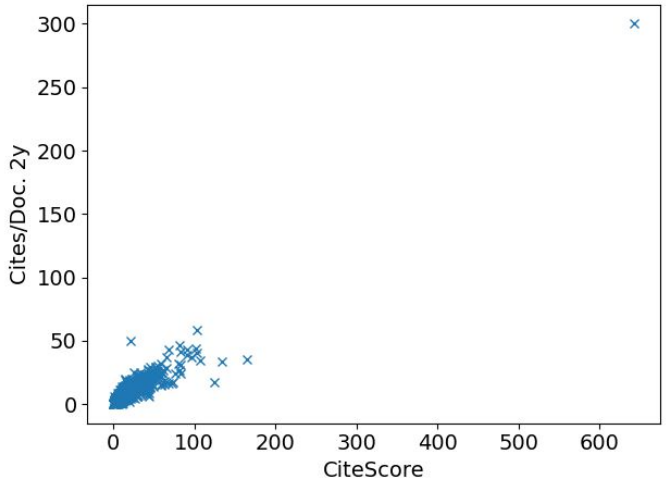
3. Research Data Quality

Collecting data

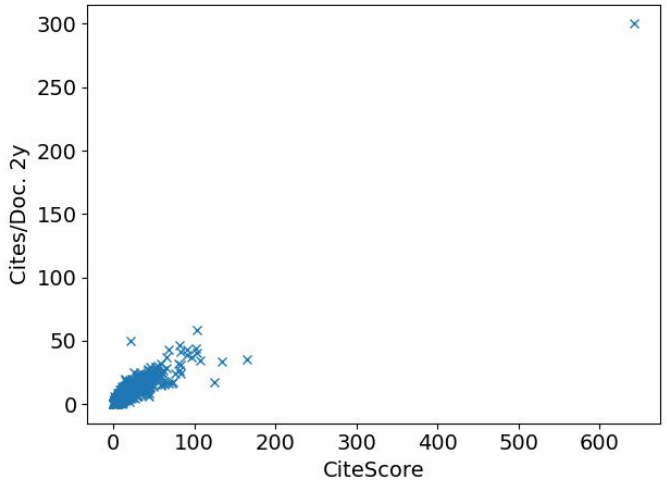
- Data must contain all ranges of the condition that is to be sampled
  - For example: To predict the impact of temperature on reactivity, all temperatures that are of interest need to be sampled (predictions only interpolate between data points but cannot extrapolate).
  - For example: CiteScore (Scopus citation index) vs. citations over all documents from last 2 years, for scientific journals.



Dataset: journal ranking dataset https://www.kaggle.com/datasets/xabirhasan/journal-ranking-dataset

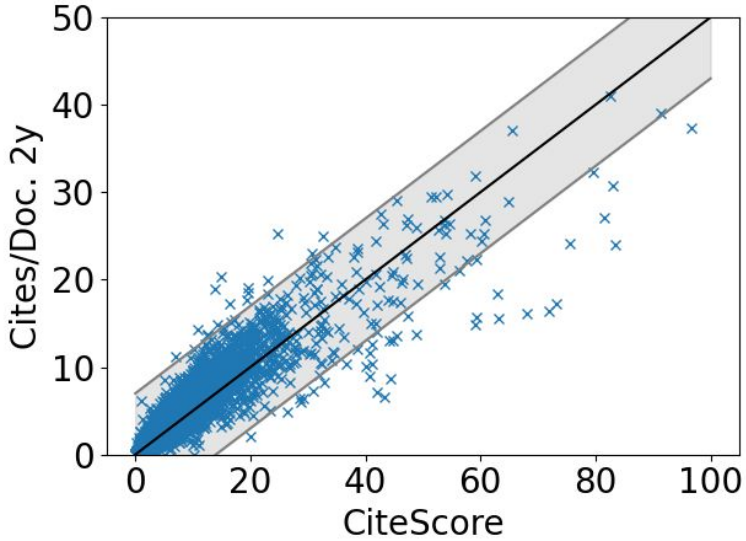
Collecting data

- Data must be homogeneous throughout feature space
  - For example: If temperature and pressure are both sampled, all combinations of features must be recorded for a homogeneous distribution of data points.
  - For example: CiteScore (Scopus citation index) vs. citations over all documents from last 2 years, for scientific journals.



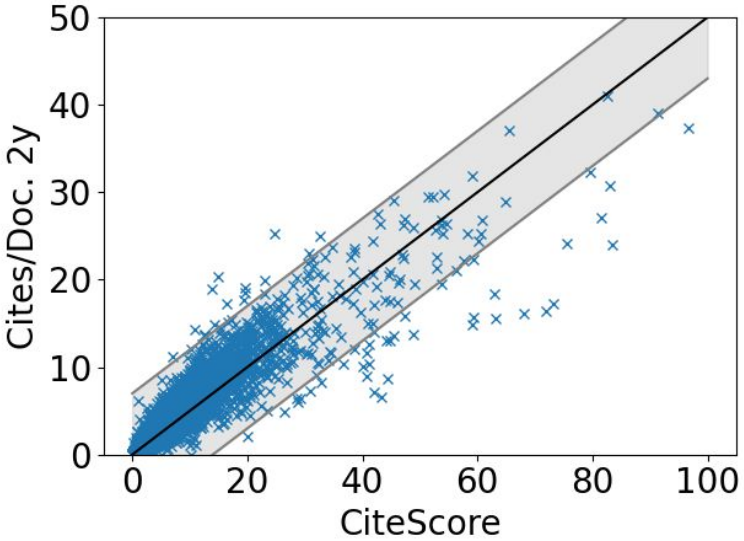
Collecting data

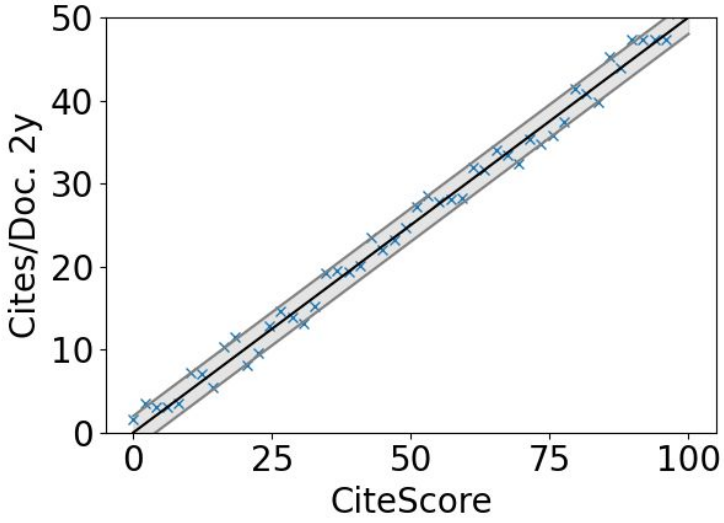
- Data must be of good quality
  - Whether it is real or synthetic data, the model can only make accurate predictions if the data itself is accurate.
  - For example: CiteScore (Scopus citation index) vs. citations over all documents from last 2 years, for scientific journals



Collecting data

- Data volume must be sufficient
  - Only with enough data can a model be trained to make accurate predictions.
  - For example: Complex data - more data points required; simpler data - fewer data points required





Collecting data

- Depending on the type of learning, data must be labeled and labeled correctly
  - Incorrect labelling interferes with the learning process.

Photo by nishizuka:
https://www.pexels.com/photo/brown-chihuahua-485294/

Photo by Maksim Goncharenok:
https://www.pexels.com/photo/a-chocolate-muffin-on-blue-surface-5994864/

Data preparation

- Make sure data is clean.
  - Correct typos, misidentified data types

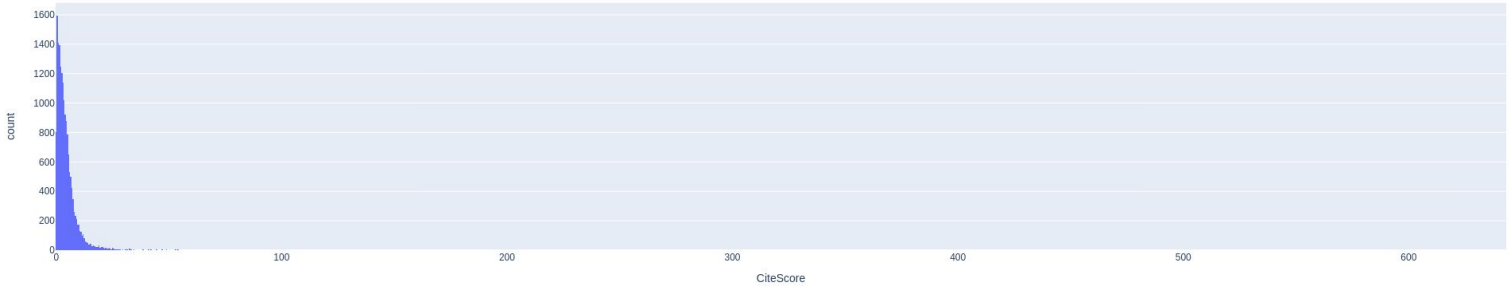
Chihuahuah →Chihuahua

Photo by nishizuka:
https://www.pexels.com/photo/brown-chihuahua-485294/

“26-04-24” →2024-04-26

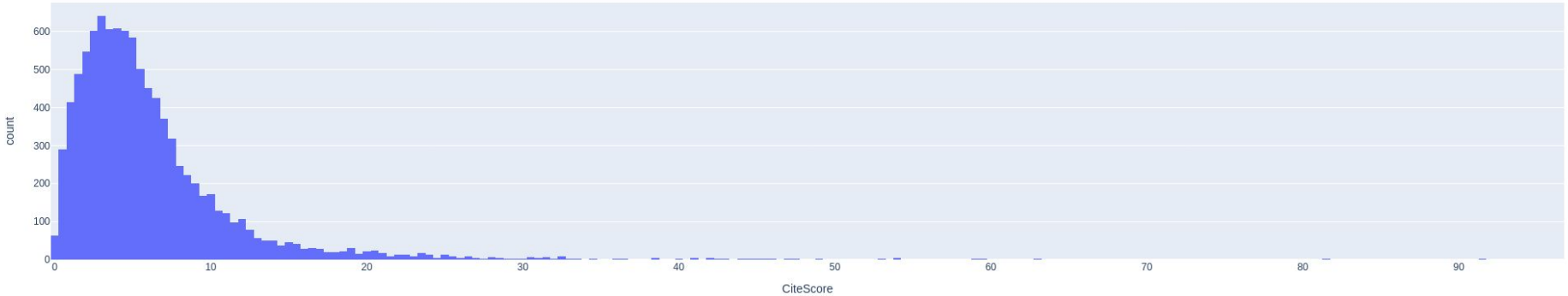
Data preparation

- Make sure data is homogeneous.
  - Visualize the data and use clustering analysis to identify outliers.
  - Use `df.describe()` and `plotly.express` to better understand your data



Data preparation

- Remove duplicates.
  - Duplicates introduce bias.
  - Use df.drop_duplicates()



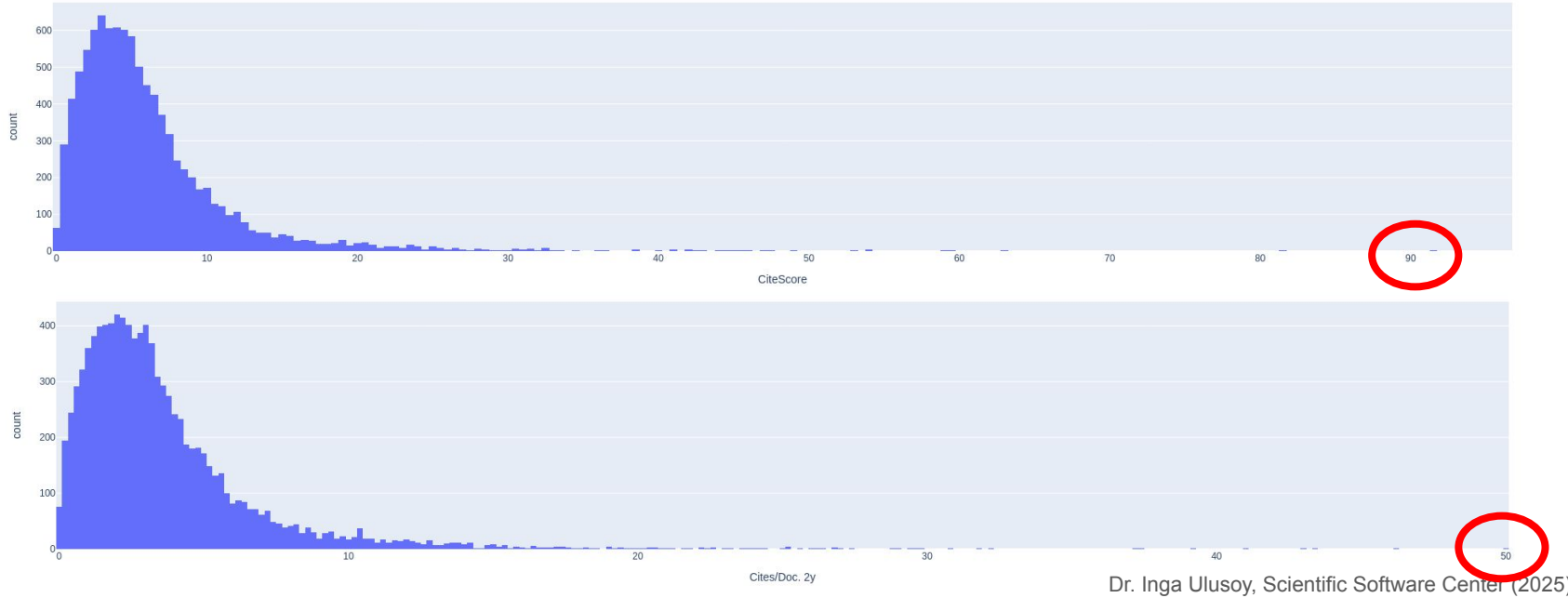
Data preparation

- Feature Engineering: Select influential features, remove unnecessary ones.
  - Unimportant features increase the complexity and reduce robustness.
  - For example: only choose features that are clearly correlated

	Rank	OA	SJR-index	CiteScore	H-index	Best Subject Rank	Total Docs.	Total Docs. 3y	Total Refs.	Total Cites 3y	Citable Docs. 3y	Cites/Doc. 2y
Rank	1.000000	0.111300	-0.503617	-0.485568	-0.625403	0.558208	-0.192069	-0.196795	-0.196338	-0.243070	-0.185484	-0.560625
OA	0.111300	1.000000	-0.069304	-0.056997	-0.178146	0.114037	0.061870	0.046403	0.058683	0.024084	0.048485	-0.045120
SJR-index	-0.503617	-0.069304	1.000000	0.878000	0.565015	-0.281225	0.091092	0.102424	0.094227	0.270083	0.081086	0.828618
CiteScore	-0.485568	-0.056997	0.878000	1.000000	0.527957	-0.279983	0.112000	0.127705	0.122350	0.285965	0.110357	0.943584
H-index	-0.625403	-0.178146	0.565015	0.527957	1.000000	-0.362788	0.331053	0.393130	0.313698	0.505095	0.362266	0.512423
Best Subject Rank	0.558208	0.114037	-0.281225	-0.279983	-0.362788	1.000000	-0.114754	-0.117089	-0.132615	-0.150247	-0.118463	-0.334142
Total Docs.	-0.192069	0.061870	0.091092	0.112000	0.331053	-0.114754	1.000000	0.934468	0.968011	0.806830	0.932626	0.150987
Total Docs. 3y	-0.196795	0.046403	0.102424	0.127705	0.393130	-0.117089	0.934468	1.000000	0.887417	0.854647	0.995085	0.148272
Total Refs.	-0.196338	0.058683	0.094227	0.122350	0.313698	-0.132615	0.968011	0.887417	1.000000	0.802696	0.893789	0.173401
Total Cites 3y	-0.243070	0.024084	0.270083	0.285965	0.505095	-0.150247	0.806830	0.854647	0.802696	1.000000	0.844114	0.308644
Citable Docs. 3y	-0.185484	0.048485	0.081086	0.110357	0.362266	-0.118463	0.932626	0.995085	0.893789	0.844114	1.000000	0.139525
Cites/Doc. 2y	-0.560625	-0.045120	0.828618	0.943584	0.512423	-0.334142	0.150987	0.148272	0.173401	0.308644	0.139525	1.000000
Refs./Doc.	-0.390894	-0.064572	0.267383	0.306913	0.247259	-0.299281	0.032381	0.019822	0.109949	0.076826	0.030626	0.382891
Life Sciences	-0.166150	0.073645	0.071380	0.125088	0.210414	-0.183625	0.044939	0.050866	0.068018	0.051387	0.051948	0.114416

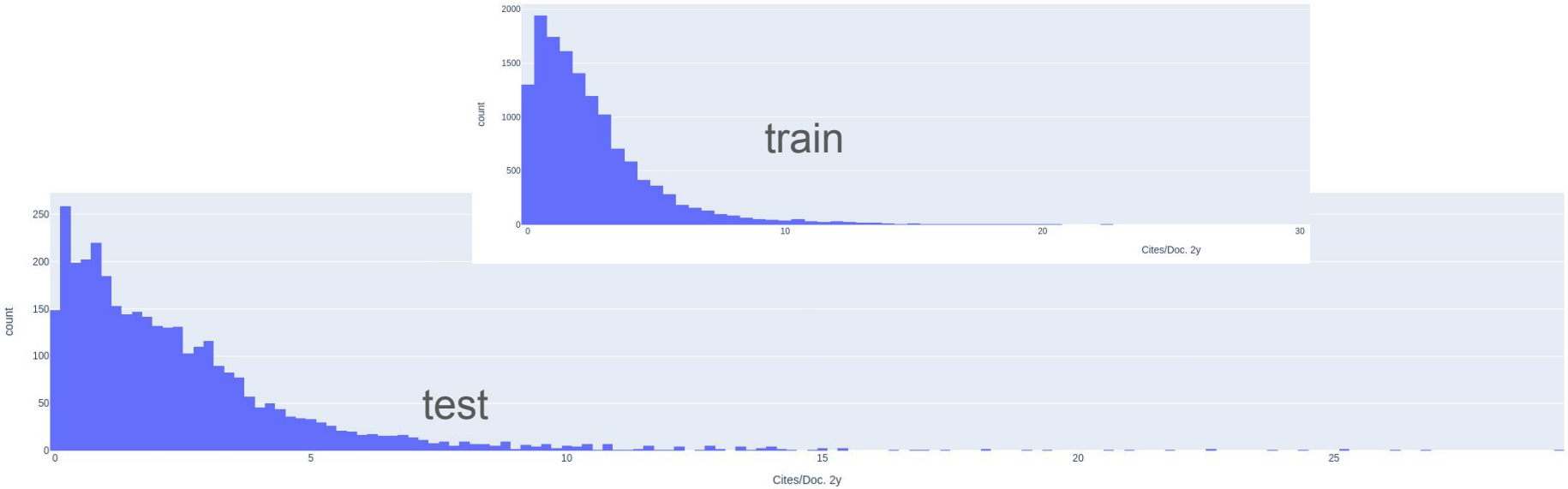
Data preparation

- Feature Engineering: Normalize features.
  - Features should have similar data ranges for the weights to be in similar ranges, and improved model robustness and faster training.



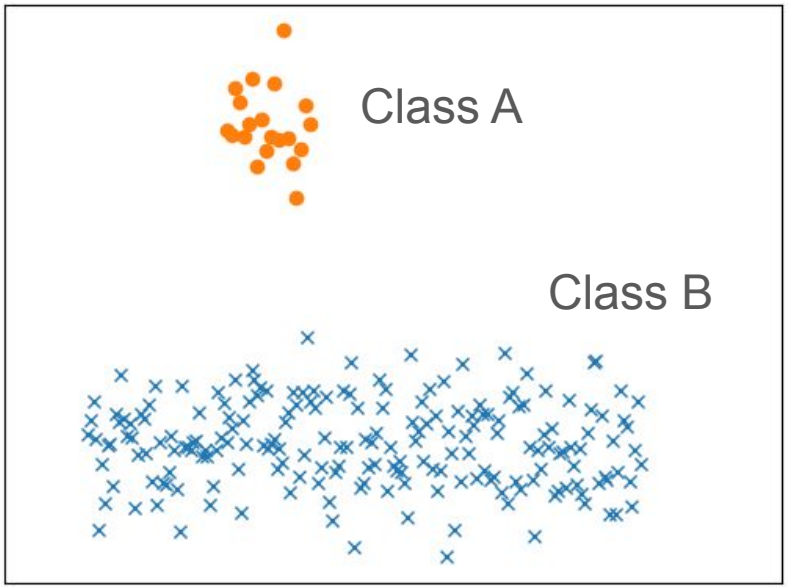
Data preparation

- Make sure to randomize your data.
  - Otherwise, your train and test data could contain more/less data of a certain kind (inhomogeneous data)



Data preparation

- Feature engineering: Make sure your dataset is balanced.
  - For classification tasks, all classes should have comparable sizes (similar numbers of examples).



Data preparation

- Feature engineering: Pick the right scale.
  - Visualize your data to see if you need to transform ie. onto a log scale.