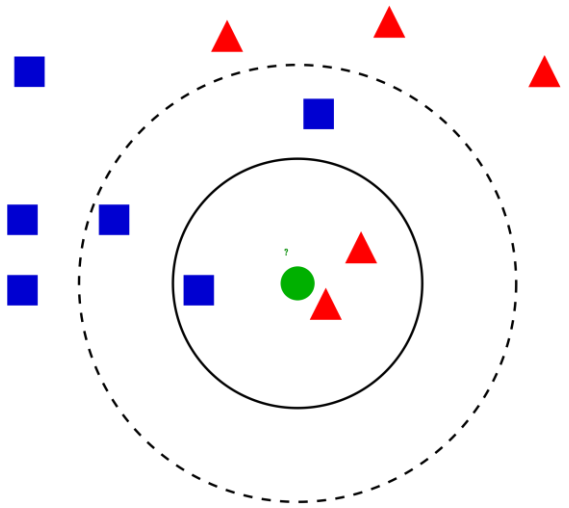
kNN – k-nearest neighbors

$$h(\mathrm{x}; D) = \arg\max_{y \in \mathrm{Y}} \sum_{\mathrm{x}_i \in D} [y_i = y] w(\mathrm{x}_i, \mathrm{x})$$

$$w(\mathrm{x}_i, \mathrm{x}) = 1, \text{if } \mathrm{x}_i - \text{one of the k nearest neighbors of x}$$

$$w(\mathrm{x}_i, \mathrm{x}) = 1, \text{if distance } \rho(\mathrm{x_i}, \mathrm{x}) < \mathrm{R} \text{ (Radius Neighbors)}$$



Cell features

label	1	2	3	4	5	6	7	8	9	10
M	17.99	10.38	122.8	1001	0.1184	0.2776	0.3001	0.1471	0.2419	0.07871
M	20.57	17.77	132.9	1326	0.08474	0.07864	0.0869	0.07017	0.1812	0.05667
M	19.69	21.25	130	1203	0.1096	0.1599	0.1974	0.1279	0.2069	0.05999
M	11.42	20.38	77.58	386.1	0.1425	0.2839	0.2414	0.1052	0.2597	0.09744
M	20.29	14.34	135.1	1297	0.1003	0.1328	0.198	0.1043	0.1809	0.05883
M	12.45	15.7	82.57	477.1	0.1278	0.17	0.1578	0.08089	0.2087	0.07613
M	18.25	19.98	119.6	1040	0.09463	0.109	0.1127	0.074	0.1794	0.05742
M	13.71	20.83	90.2	577.9	0.1189	0.1645	0.09366	0.05985	0.2196	0.07451
M	13	21.82	87.5	519.8	0.1273	0.1932	0.1859	0.09353	0.235	0.07389
M	12.46	24.04	83.97	475.9	0.1186	0.2396	0.2273	0.08543	0.203	0.08243
M	16.02	23.24	102.7	797.8	0.08206	0.06669	0.03299	0.03323	0.1528	0.05697
M	15.78	17.89	103.6	781	0.0971	0.1292	0.09954	0.06606	0.1842	0.06082
M	19.17	24.8	132.4	1123	0.0974	0.2458	0.2065	0.1118	0.2397	0.078
M	15.85	23.95	103.7	782.7	0.08401	0.1002	0.09938	0.05364	0.1847	0.05338
M	13.73	22.61	93.6	578.3	0.1131	0.2293	0.2128	0.08025	0.2069	0.07682
M	14.54	27.54	96.73	658.8	0.1139	0.1595	0.1639	0.07364	0.2303	0.07077
M	14.68	20.13	94.74	684.5	0.09867	0.072	0.07395	0.05259	0.1586	0.05922
M	16.13	20.68	108.1	798.8	0.117	0.2022	0.1722	0.1028	0.2164	0.07356
M	19.81	22.15	130	1260	0.09831	0.1027	0.1479	0.09498	0.1582	0.05395
B	13.54	14.36	87.46	566.3	0.09779	0.08129	0.06664	0.04781	0.1885	0.05766
B	13.08	15.71	85.63	520	0.1075	0.127	0.04568	0.0311	0.1967	0.06811
B	9.504	12.44	60.34	273.9	0.1024	0.06492	0.02956	0.02076	0.1815	0.06905
M	15.34	14.26	102.5	704.4	0.1073	0.2135	0.2077	0.09756	0.2521	0.07032
M	21.16	23.04	137.2	1404	0.09428	0.1022	0.1097	0.08632	0.1769	0.05278
M	16.65	21.38	110	904.6	0.1121	0.1457	0.1525	0.0917	0.1995	0.0633

Categorical features
One Hot Encoding

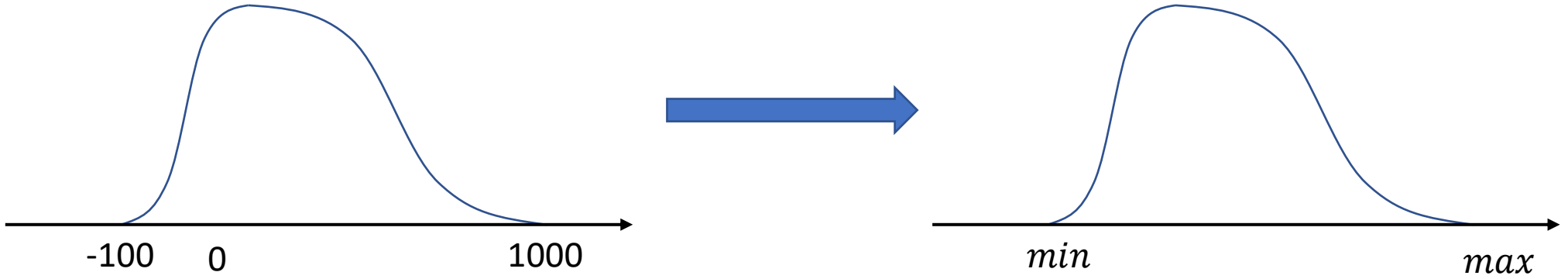
color	color_index
red	0
green	1
blue	2
red	0

Categorical features
One Hot Encoding

color	color_red	color_blue	color_green
red	1	0	0
green	0	0	1
blue	0	1	0
red	1	0	0

Scalers

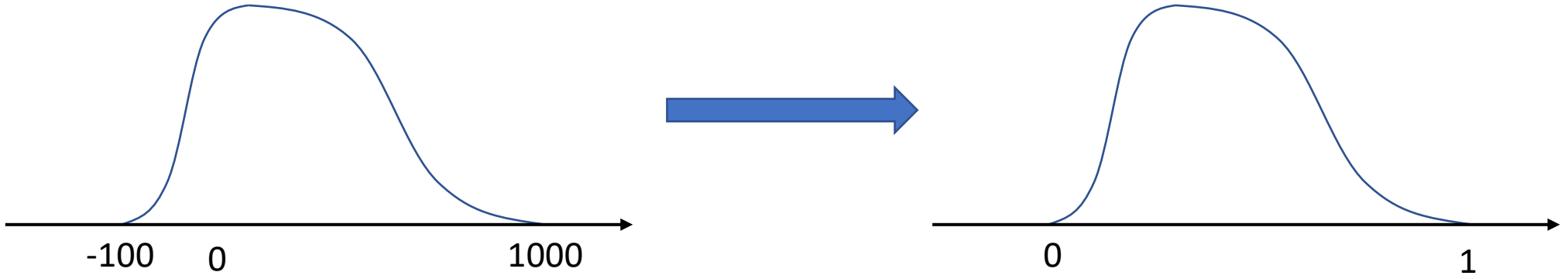
MinMax Scaler



$$x_{scaled} = \frac{x - \min(x)}{\max(x) - \min(x)} * (max - min) + min$$

Scalers

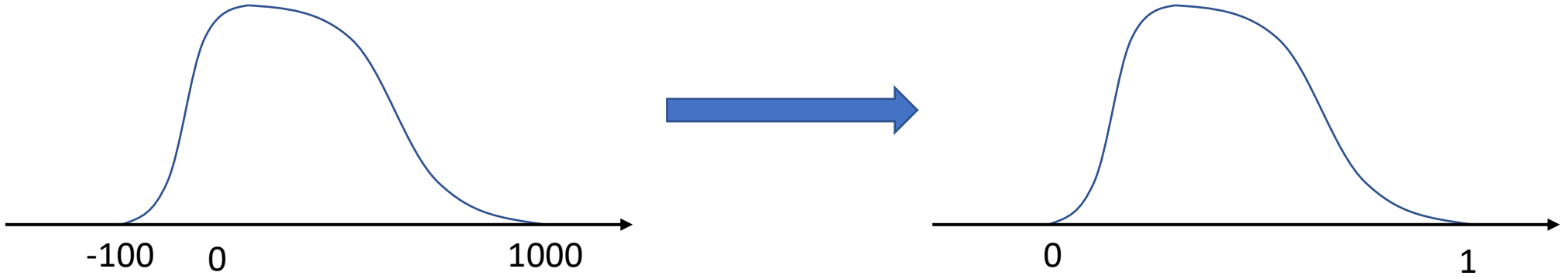
MinMax Scaler



$$x_{scaled} = \frac{x - \min(x)}{\max(x) - \min(x)}$$

Scalers

MinMax Scaler



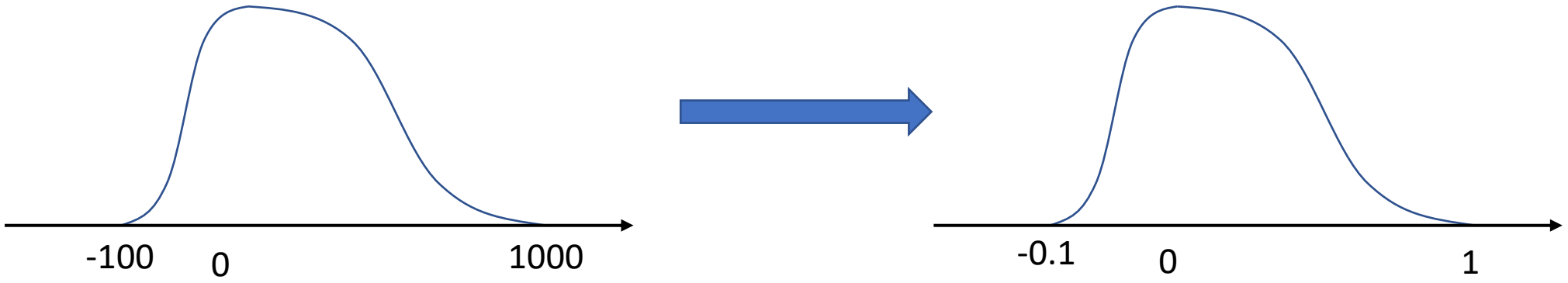
$$x_{scaled}^{train} = \frac{x^{train} - \min(x^{train})}{\max(x^{train}) - \min(x^{train})}$$

$$x_{scaled}^{\boldsymbol{val}} = \frac{x^{\boldsymbol{val}} - \min(x^{train})}{\max(x^{train}) - \min(x^{train})}$$

$$x_{scaled}^{\boldsymbol{test}} = \frac{x^{\boldsymbol{test}} - \min(x^{train})}{\max(x^{train}) - \min(x^{train})}$$

Scalers

MaxAbs Scaler

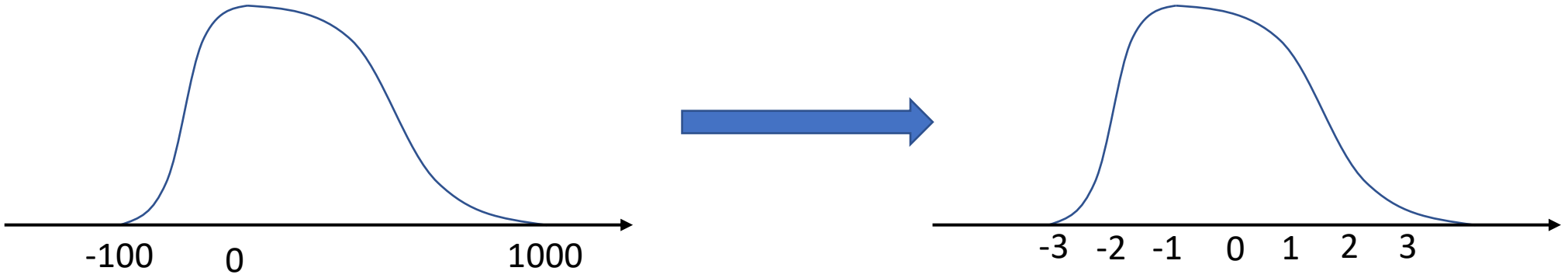


$$x_{scaled} = \frac{x}{\max(|x|)}$$

Keep the sparsity of the data!

Scalers

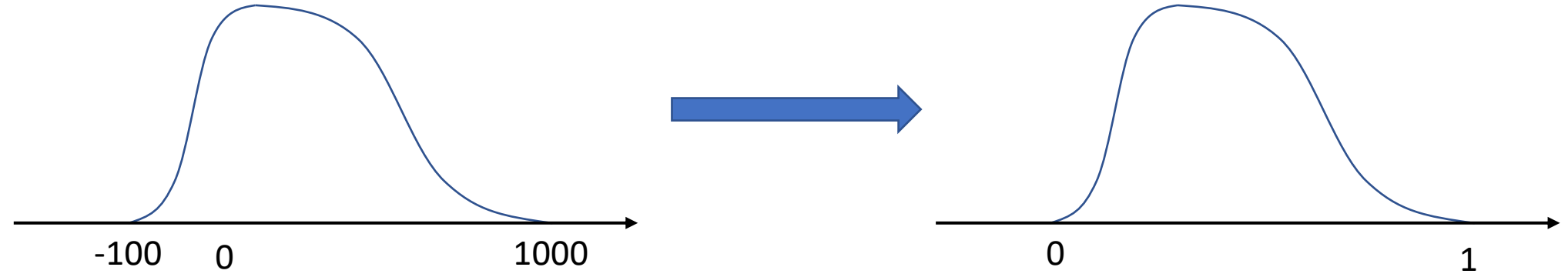
Standard Scaler



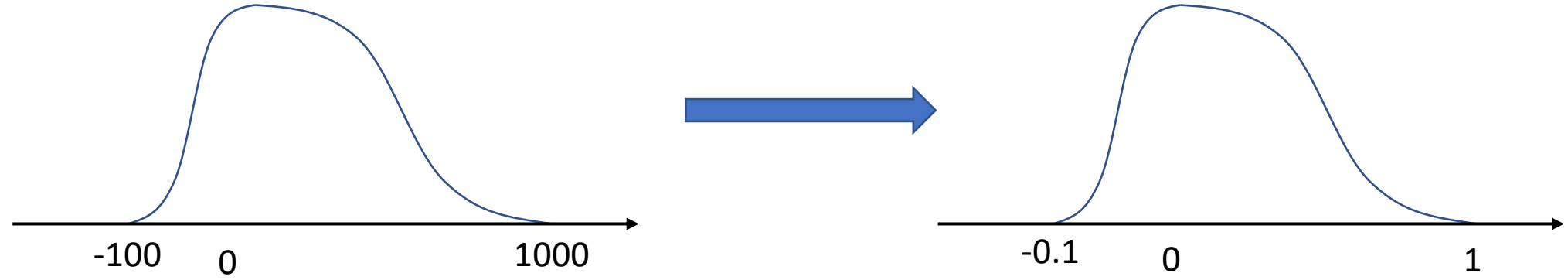
$$x_{scaled} = \frac{x - \text{mean}(x)}{\text{std}(x)}$$

Scalers

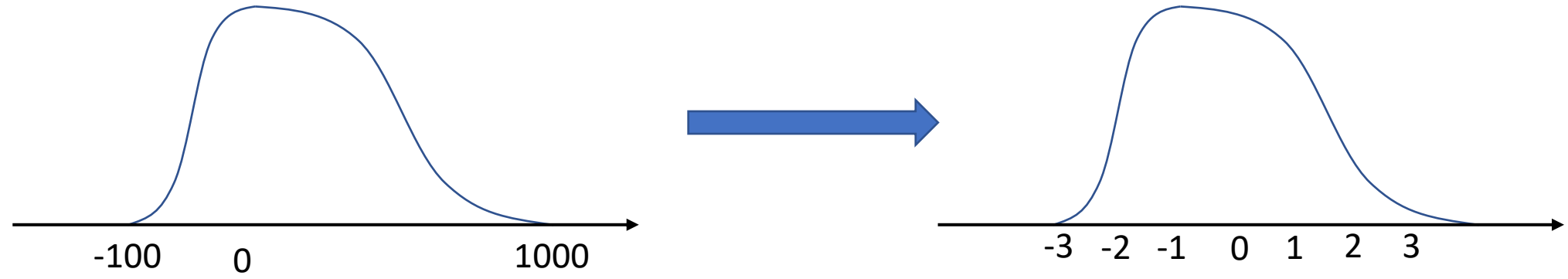
MinMax Scaler



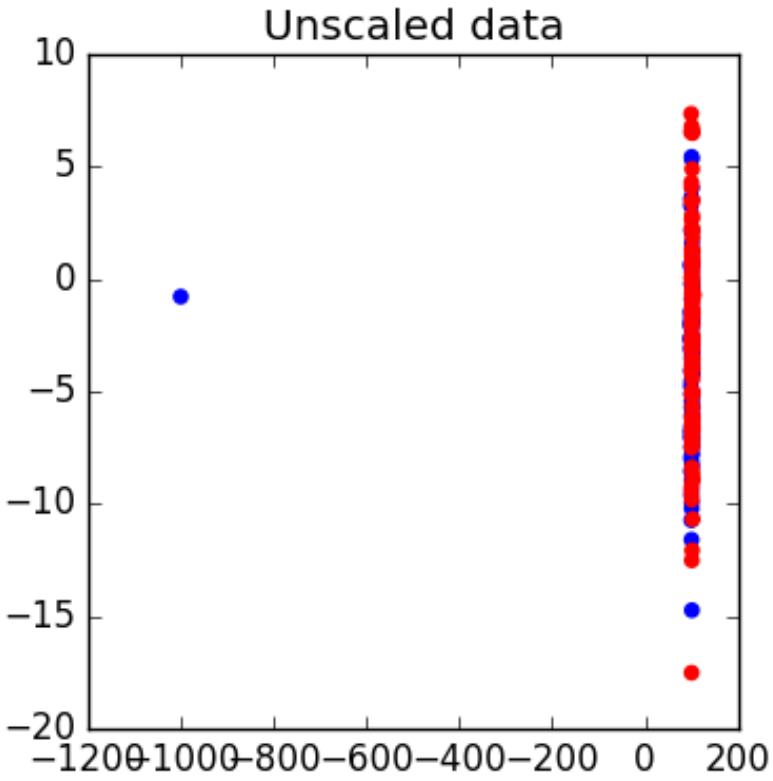
MaxAbs Scaler

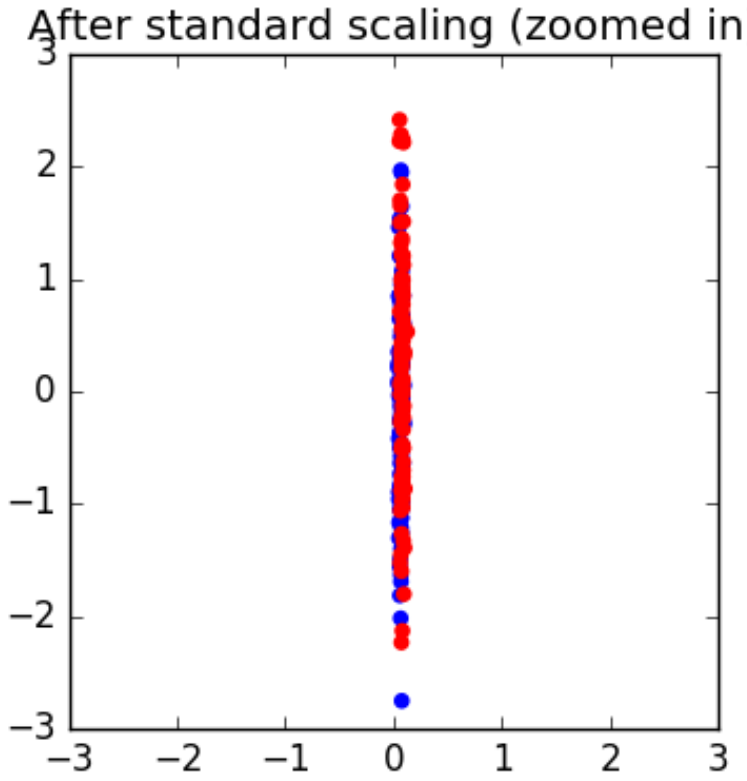


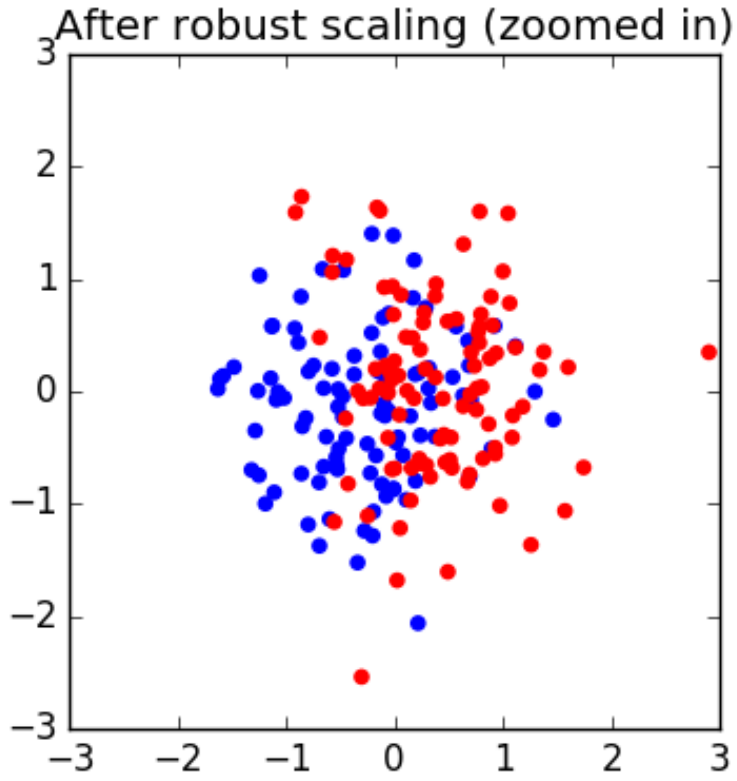
Standard Scaler



Robust scaler

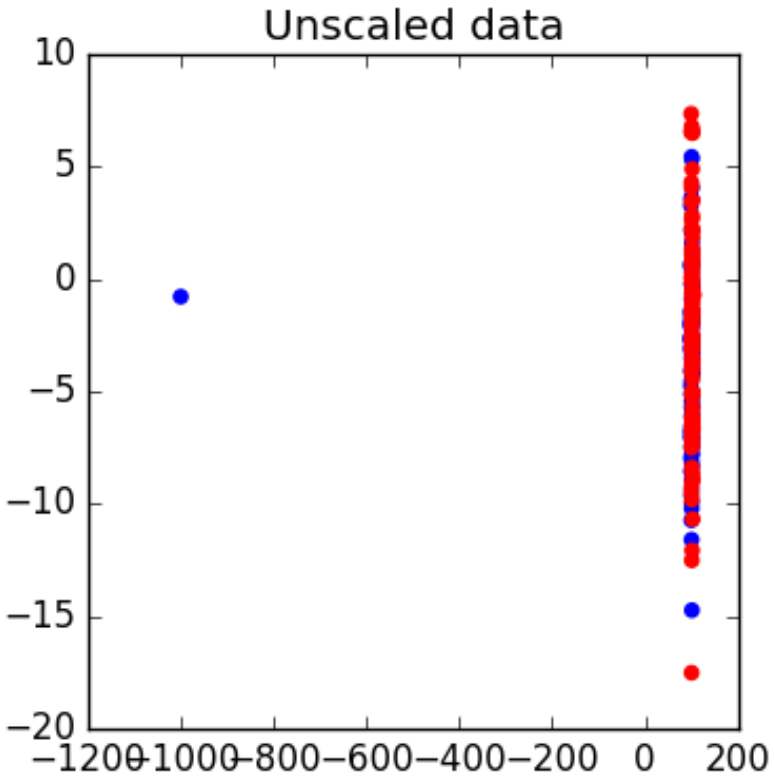


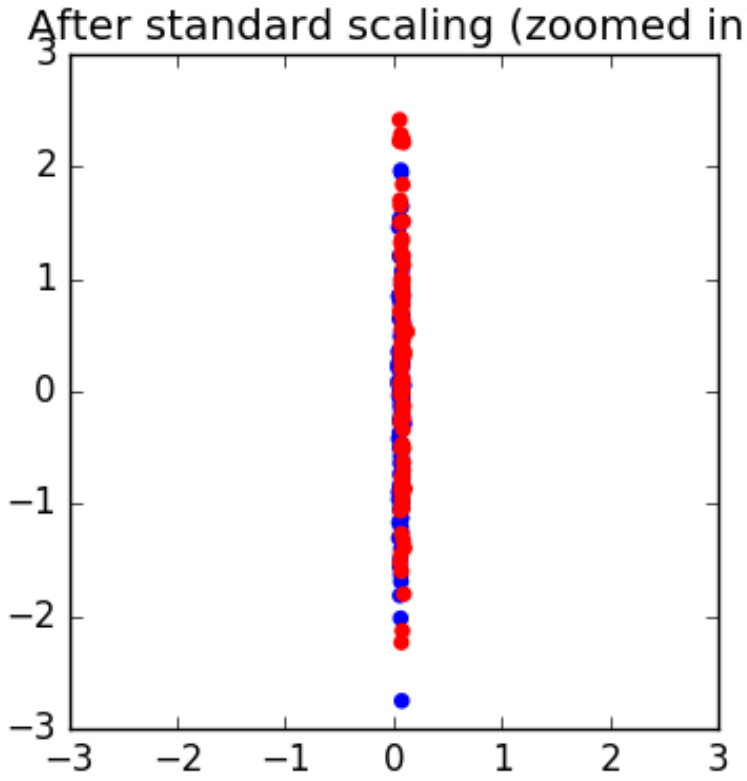


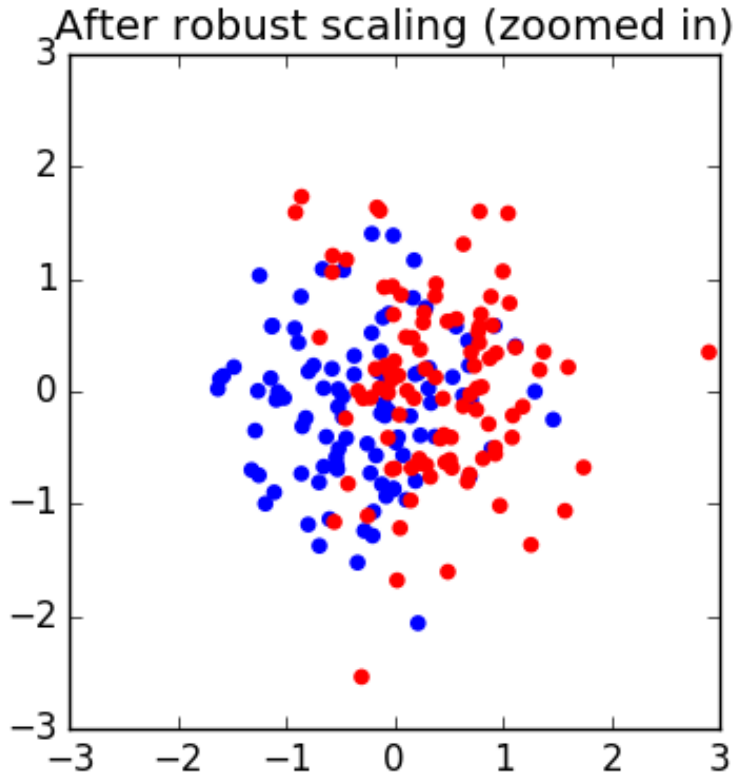


$$x_{scaled} = \frac{x - \text{median}(x)}{\text{percentile}_{max}(x) - \text{percentile}_{min}(x)}$$

Robust scaler



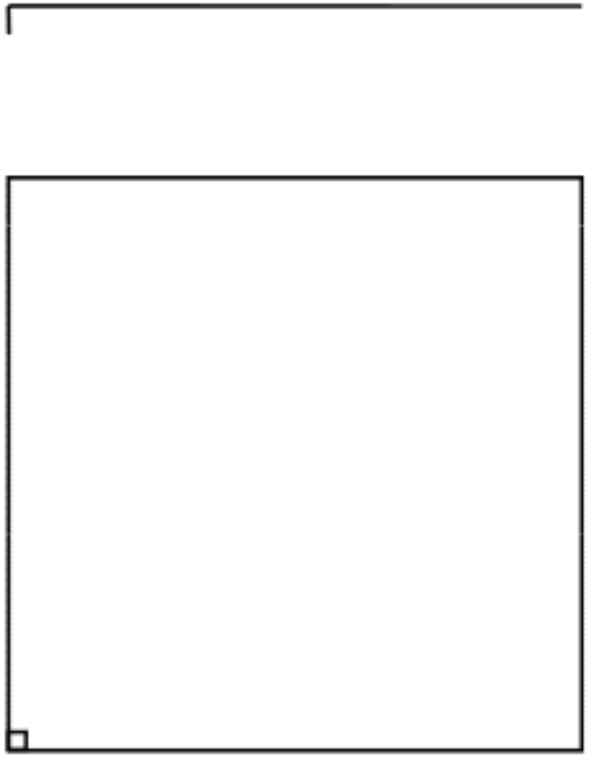




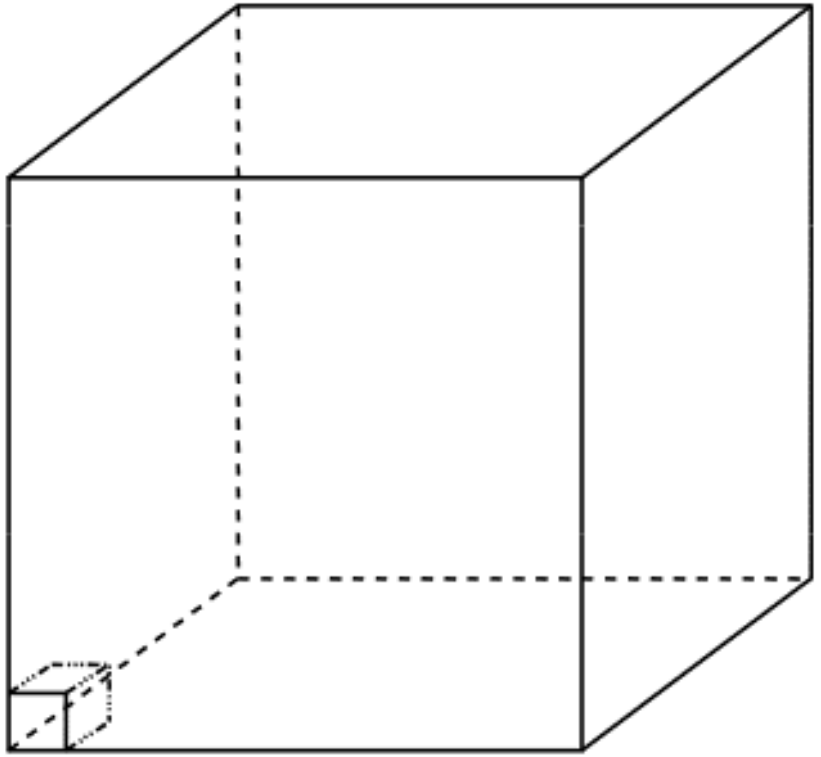
$$x_{scaled} = \frac{x - \text{median}(x)}{\text{percentile}_{0.75}(x) - \text{percentile}_{0.25}(x)}$$

The curse of dimensionality

5000 evenly distributed points, 5 nearest neighbors



0.033 x 0.033



0.1 x 0.1 x 0.1

100 dimensions – (0.93 x 0.93 x 0.93 x 0.93…..)

Step-wise kNN

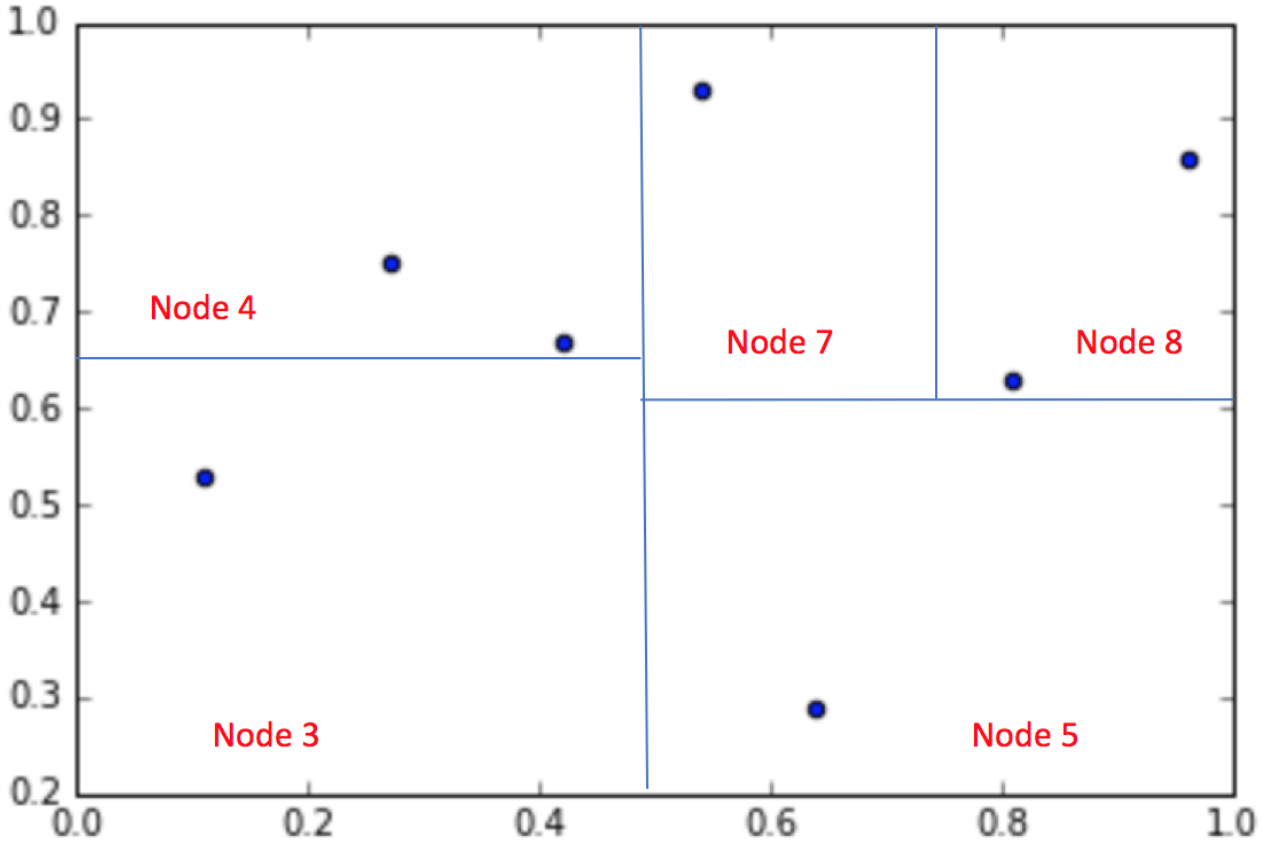
1. Select the best feature l: $\rho_l(\mathrm{x}', \mathrm{x}) = |x'_l - x_l|$

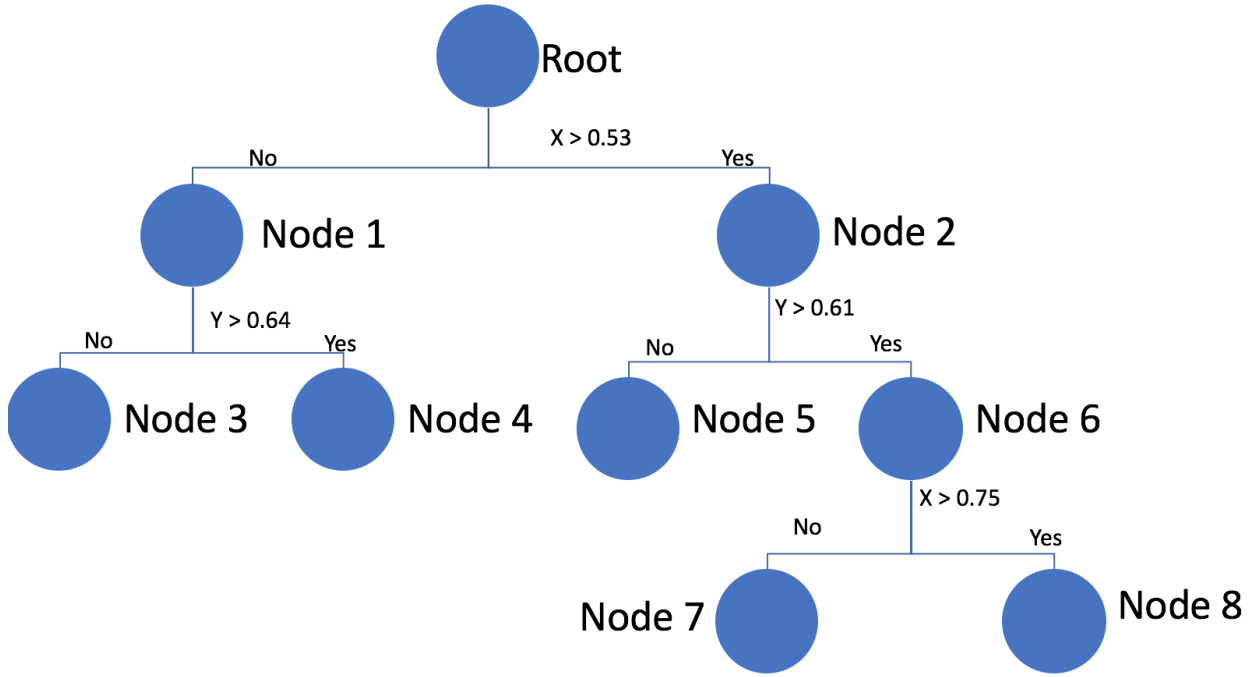
2. Find the best feature l *and the weight*:

$$\rho(\mathrm{x}', \mathrm{x}) = \rho(x', x) + w_{l'}|x'_{l'} - x_{l'}|$$

3. Repeat (2) while the LOO or the validation error is decreasing (or the accuracy is increasing).

Fast nearest neighbor search
k-d tree

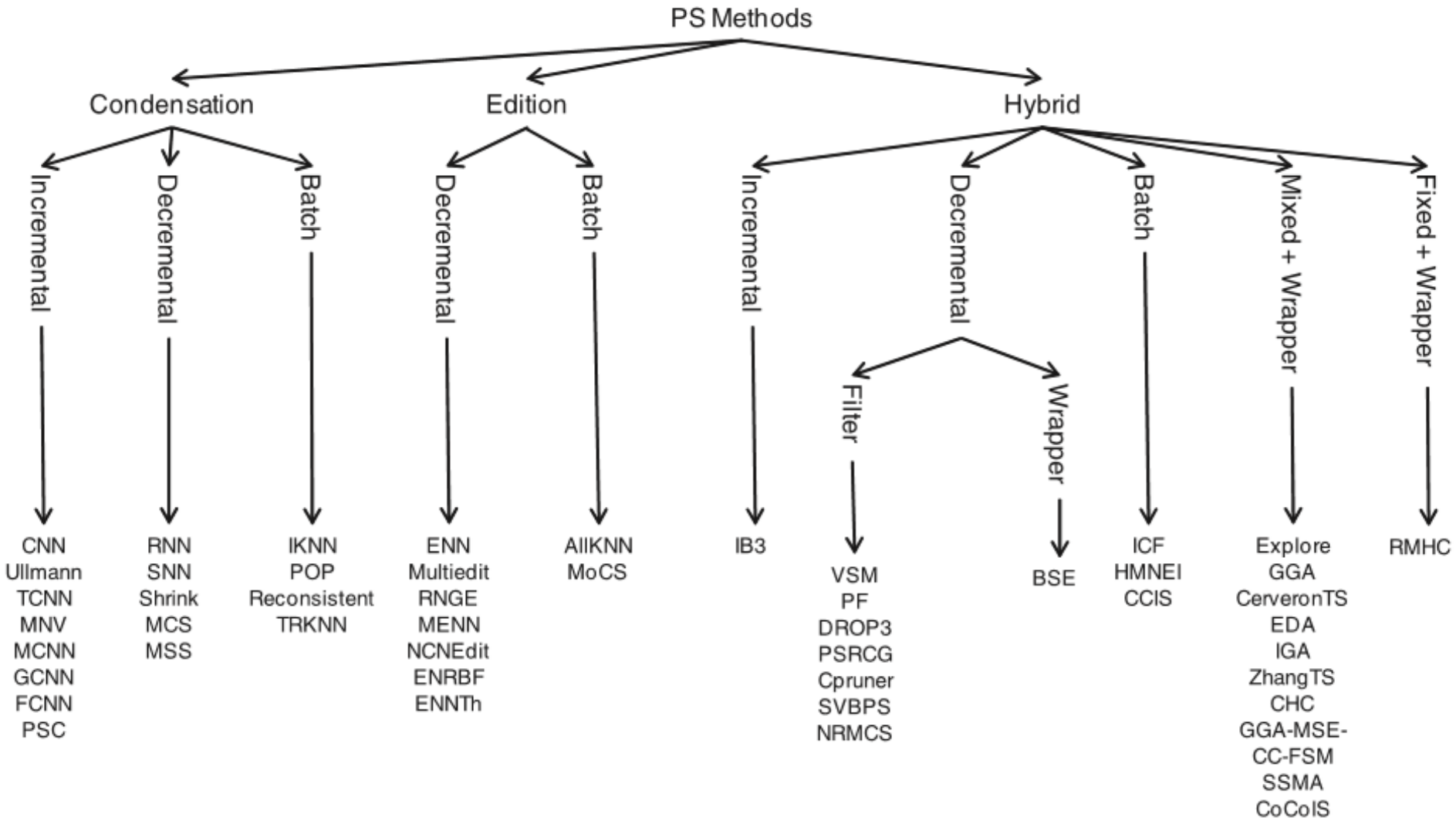




Prototype selection

$$h(\mathrm{x};\Omega) = \arg\max_{y \in Y} \sum_{\mathrm{x}_i \in \Omega} [y_i = y]\, w(\mathrm{x}_i, \mathrm{x})$$

Prototype selection methods taxonomy



DROP5
(*Decremental Reduction Optimization Procedure)*

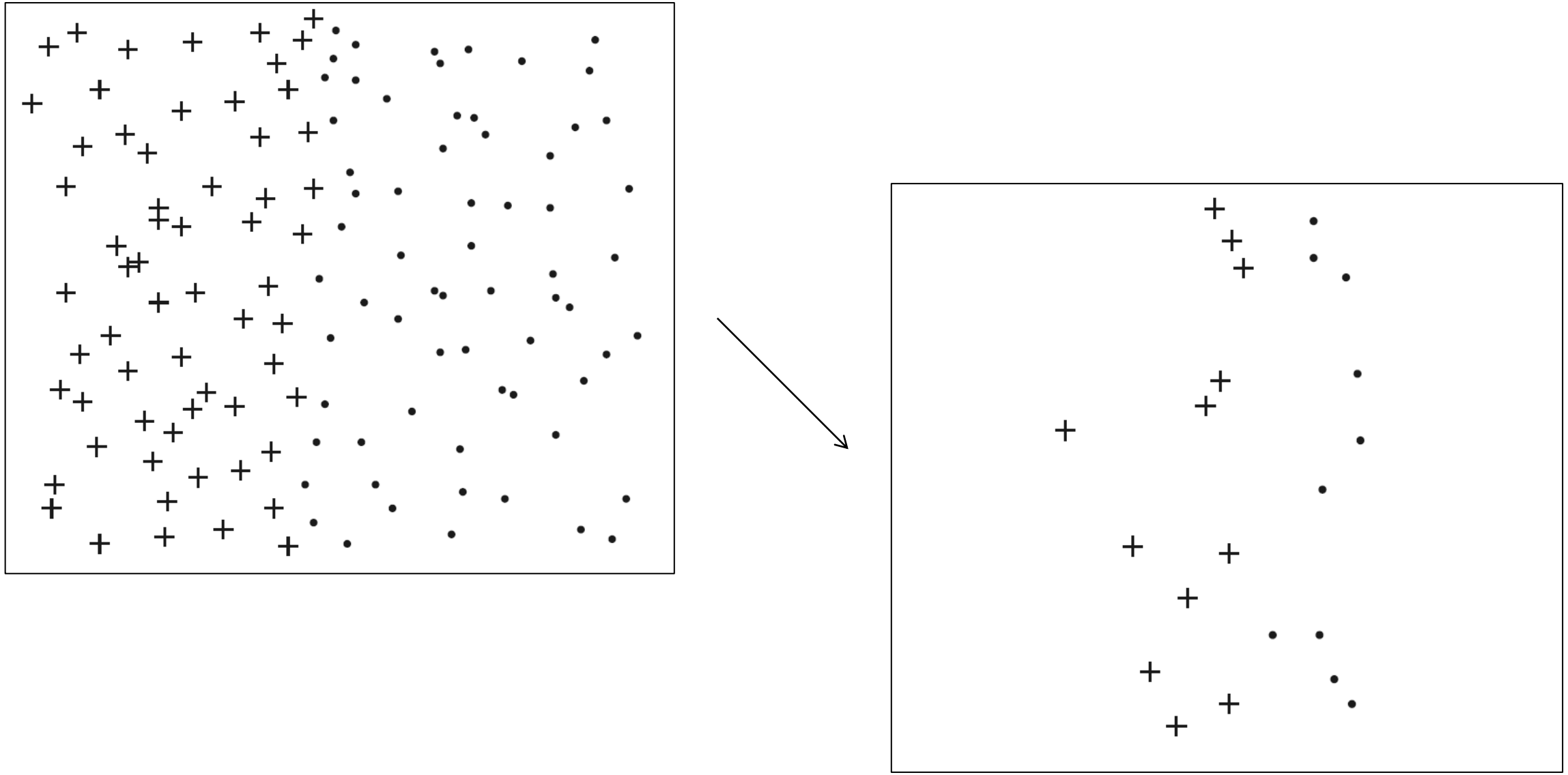
Start with the full dataset.

Sort data points by the affinity to the closest incorrect class.

Go in the ascending order.

Delete point **x**, if that does not increase the LOO error for the points that consider **x** one of their closest neighbors.

DROP5



What is machine learning?

”it is a field of study that gives the ability to the computer to self-learn without being explicitly programmed”, - Arthur Samuel

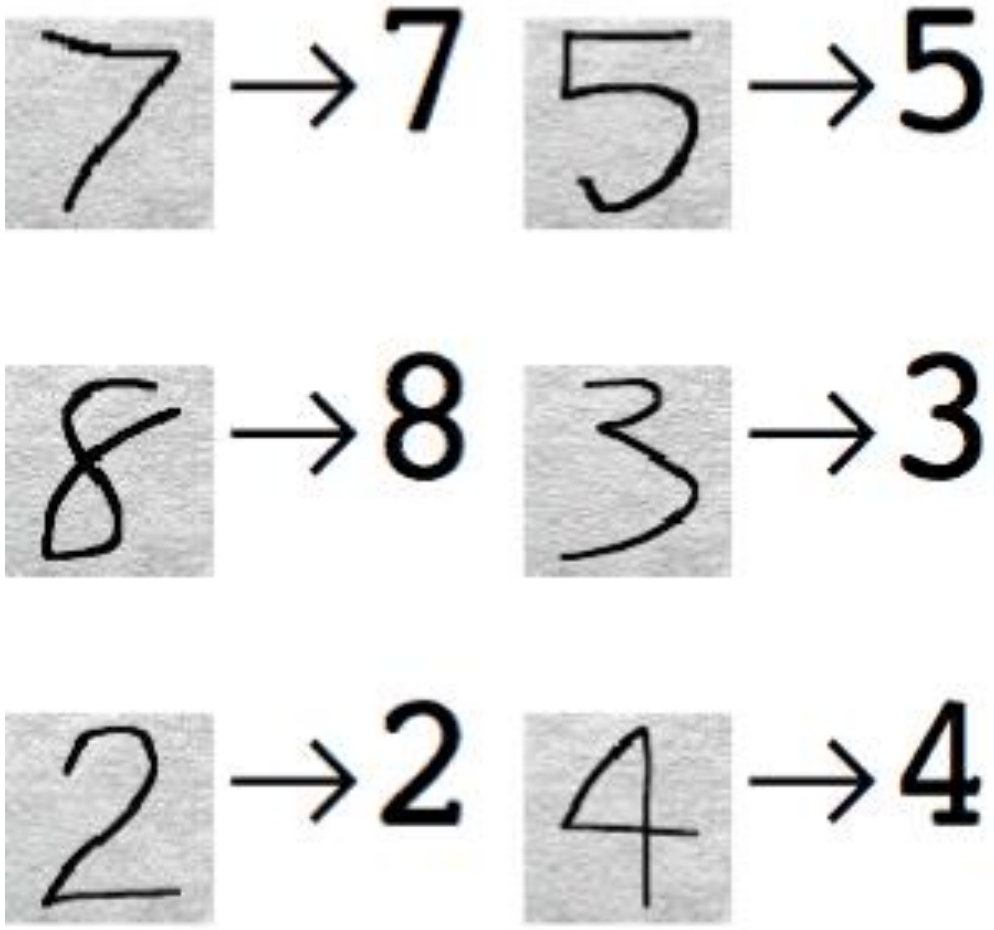
Some classification of machine learning situations

	Small data	**Big data**
Panel data	kNN, SVM Linear regression	Boosted decision trees
Images, sound, text	Deep learning, but with tricks	Deep learning
Cluster analysis	What?	Clustering methods
Optimization	Bayesian optimization	Hill climb, annealing, GA
Agent systems	Q-learning	Deep RL

Types of machine learning

- Supervised learning

- Input: $\mathbf{x}$
- Output (label): $\boldsymbol{y}$
- Target function: $\boldsymbol{f}: \boldsymbol{X} \rightarrow \boldsymbol{Y}$
- Data: $(\mathbf{x_1}, \boldsymbol{y_1}), (\mathbf{x_2}, \boldsymbol{y_2}), .., (\mathbf{x_N}, \boldsymbol{y_N})$
- Hypothesis: $\boldsymbol{h}: \boldsymbol{X} \rightarrow \boldsymbol{Y}$



Common notation for datapoints

$$\mathbf{D} = [\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n]$$

Data

Datapoint (vector)

$$\mathbf{x} = [x_1, x_2, \ldots, x_k]$$

Datapoint

Feature (value)

Types of machine learning

- Unsupervised learning

- Input: $\mathbf{x}$
- Data: $(\mathbf{x}_1), (\mathbf{x}_2), .., \quad (\mathbf{x}_N)$
- Goals:
  - Information extraction.
  - Dependencies extraction.
  - Reducing data size.
  - ....

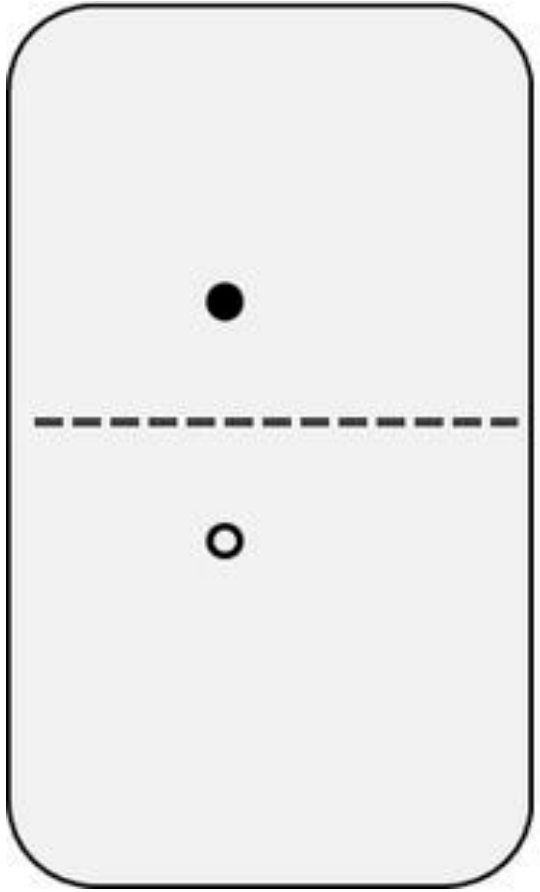
Types of machine learning

- Semi-supervised learning

We utilize unlabeled data for better performance.

- Active learning

We can ask for more labels but do it on a budget.



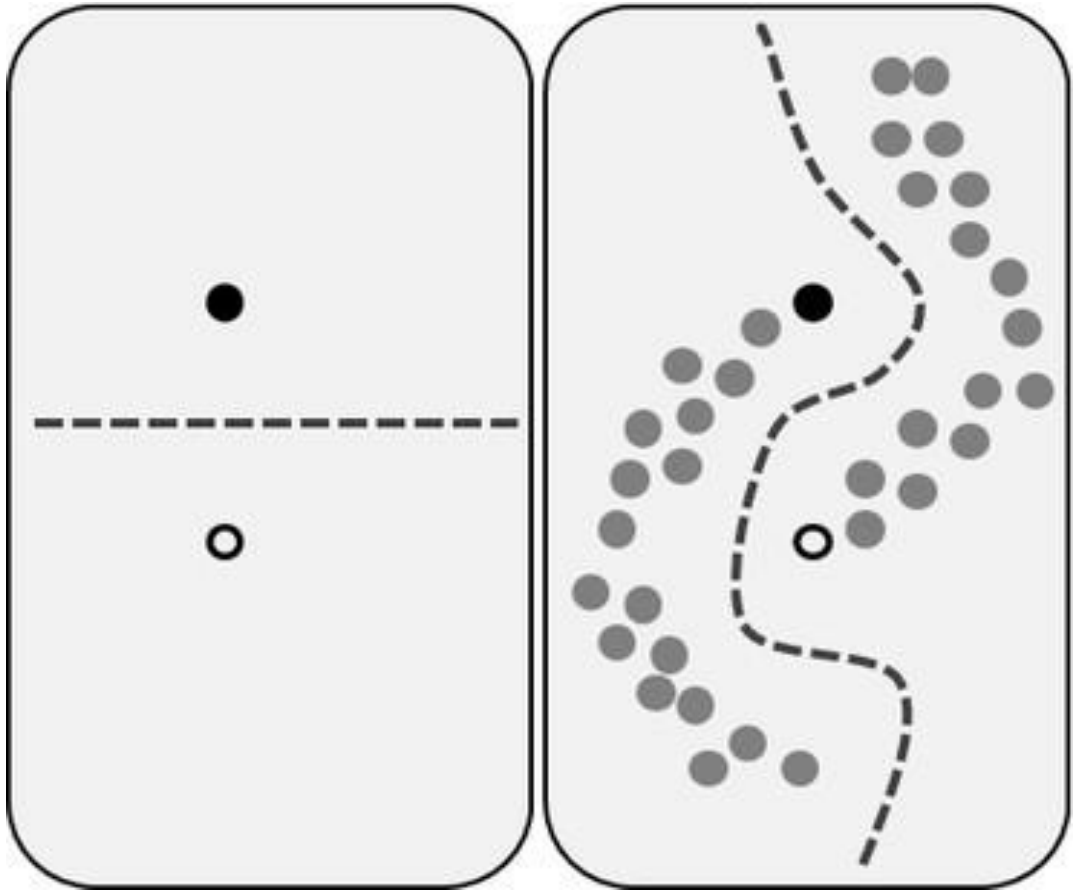
Types of machine learning

- Semi-supervised learning

We utilize unlabeled data for better performance.

- Active learning

We can ask for more labels but do it on a budget.

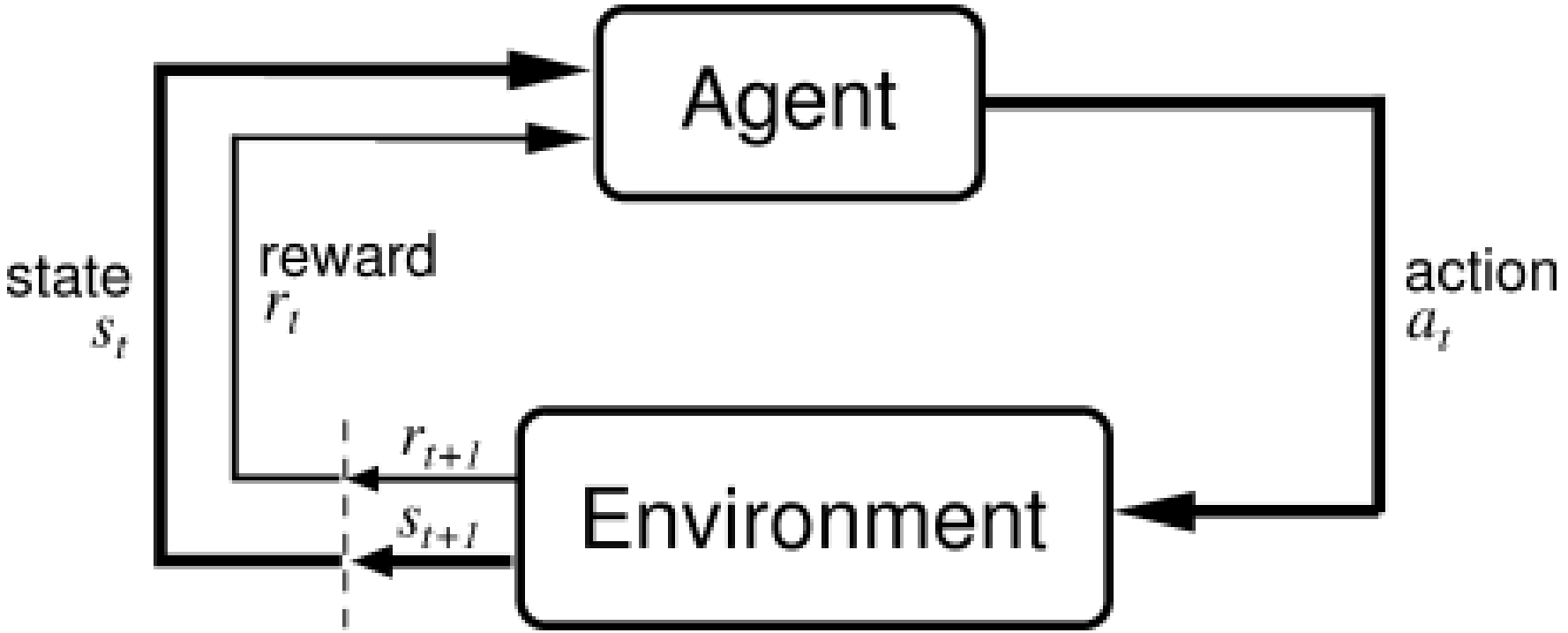


Types of machine learning

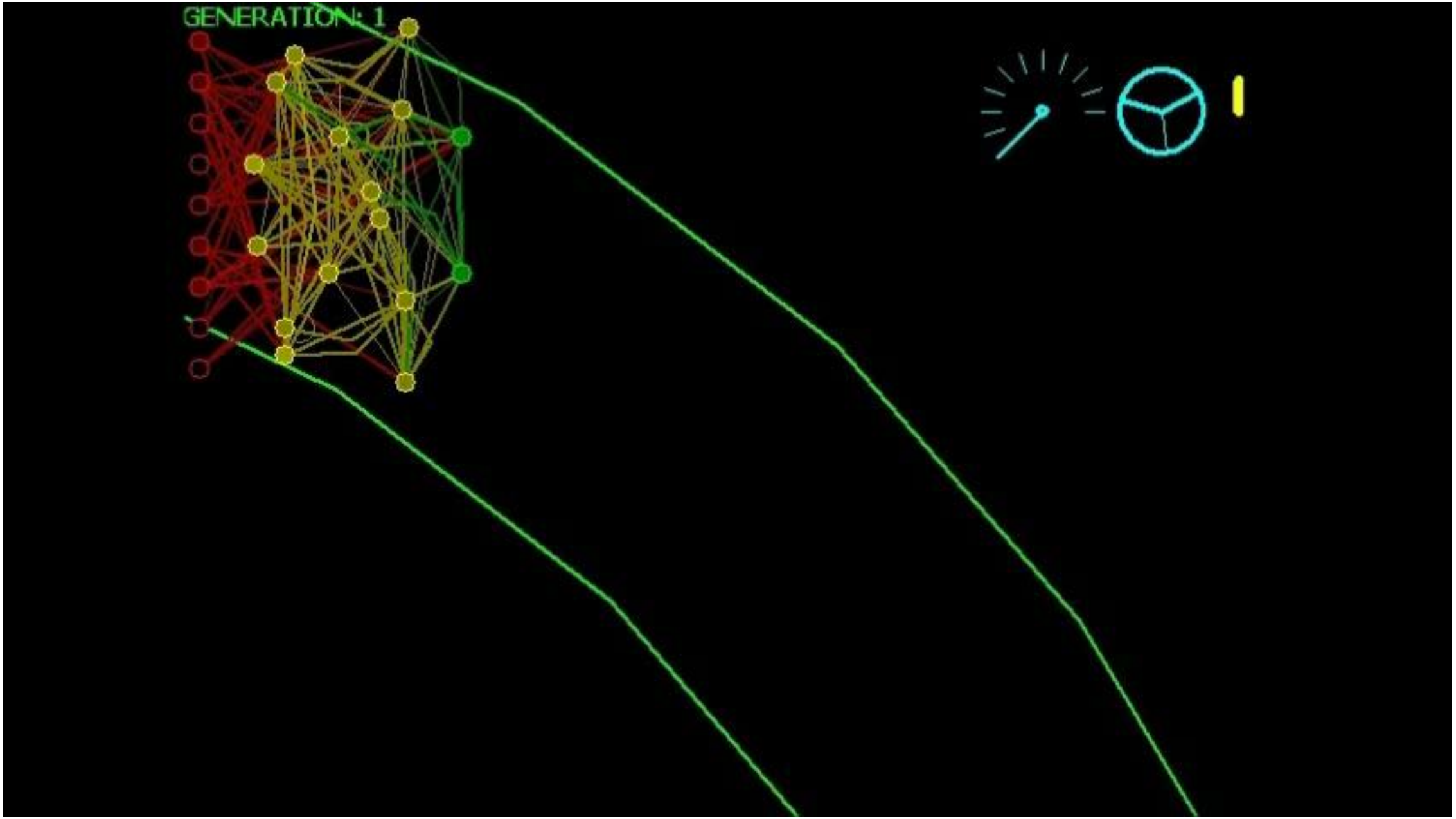
- Reinforcement learning

Types of machine learning

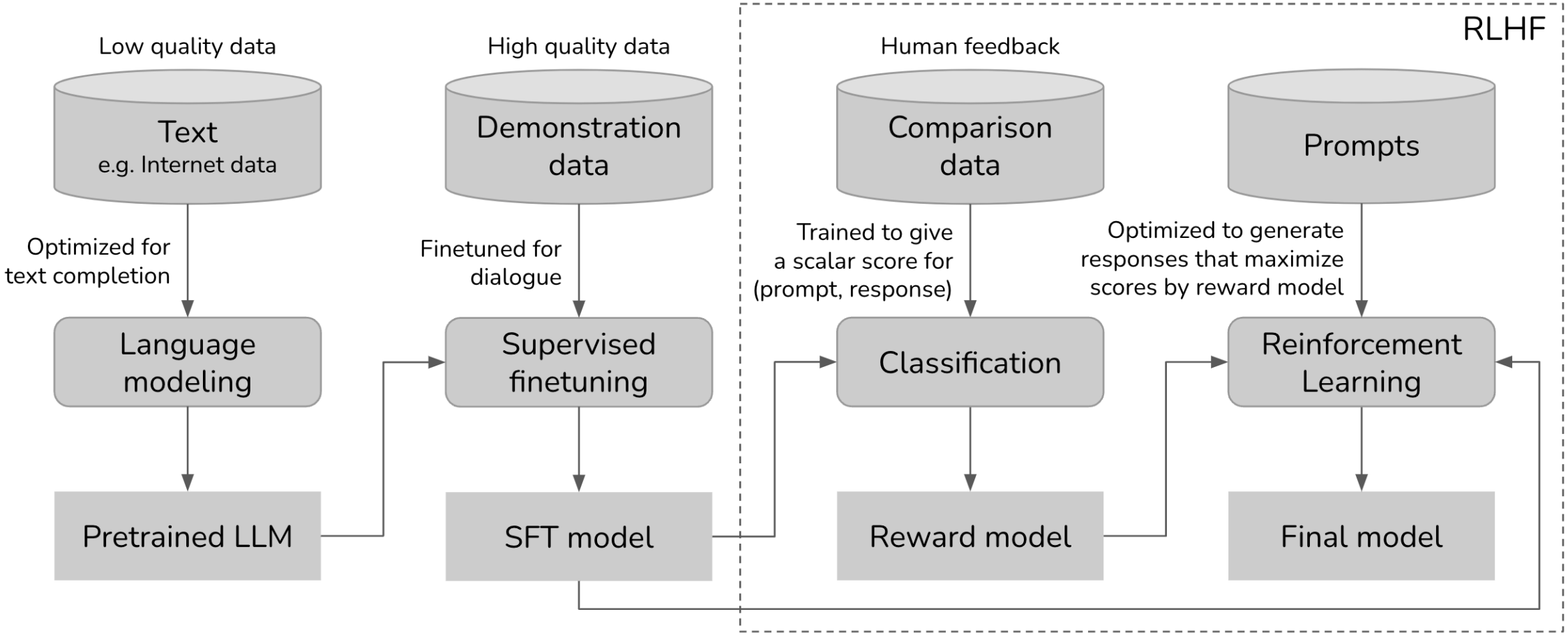
- Reinforcement learning




Reinforcement Learning
First trial...


GENERATION: 1

ChatGPT



Scale
May '23

>1 trillion
tokens

10K - 100K
(prompt, response)

100K - 1M comparisons
(prompt, winning_response, losing_response)

10K - 100K
prompts

Examples
Bolded: open sourced

GPT-x, Gopher, **Falcon**, LLaMa, **Pythia**, **Bloom**, **StableLM**

Dolly-v2, Falcon-Instruct

InstructGPT, ChatGPT, Claude, **StableVicuna**