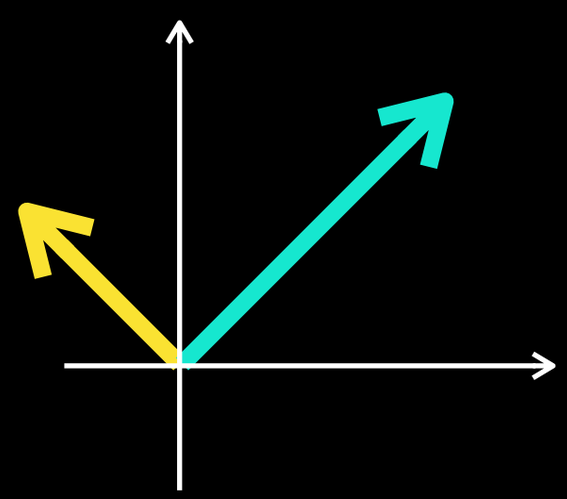
Data normalization

Linear Algebra Essentials



California Housing Prices

https://www.kaggle.com/camnugent/california-housing-prices

```
df = pd.read_csv("housing.csv")
df.head()
```

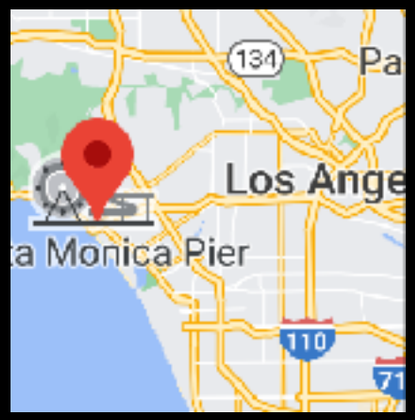
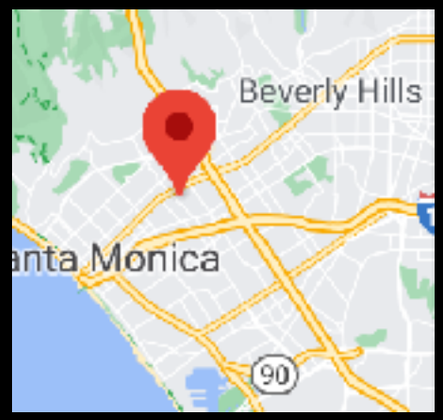
	longitude	latitude	housing_median_age	total_rooms	total_bedrooms	population	households	median_income	median_house_value	ocean_proximity
0	-122.23	37.88	41.0	880.0	129.0	322.0	126.0	8.3252	452600.0	NEAR BAY
1	-122.22	37.86	21.0	7099.0	1106.0	2401.0	1138.0	8.3014	358500.0	NEAR BAY
2	-122.24	37.85	52.0	1467.0	190.0	496.0	177.0	7.2574	352100.0	NEAR BAY
3	-122.25	37.85	52.0	1274.0	235.0	558.0	219.0	5.6431	341300.0	NEAR BAY
4	-122.25	37.85	52.0	1627.0	280.0	565.0	259.0	3.8462	342200.0	NEAR BAY

California housing prices

Santa Monica: (34.0218555,-118.5158609)
House age: 20 years
Total area: 3000 sq ft
bedrooms area: 1500 sq ft
population: 1000
households: 500
income: 8 K$
house value: 360,000 $

```
(3079.852424674151,
 array([-1.2220e+02,  3.7820e+01,  3.9000e+01,  3.7700e+03,  5.3400e+02,
         1.2650e+03,  5.0000e+02,  6.3302e+00,  3.6280e+05])),
(4186.368601776515,
 array([-1.2194e+02,  3.7540e+01,  3.1000e+01,  2.5370e+03,  3.8200e+02,
         1.0670e+03,  4.1000e+02,  6.7599e+00,  3.5600e+05])),
(4650.349236146062,
 array([-1.2222e+02,  3.7860e+01,  2.1000e+01,  7.0990e+03,  1.1060e+03,
         2.4010e+03,  1.1380e+03,  8.3014e+00,  3.5850e+05])),
```

Closest vectors

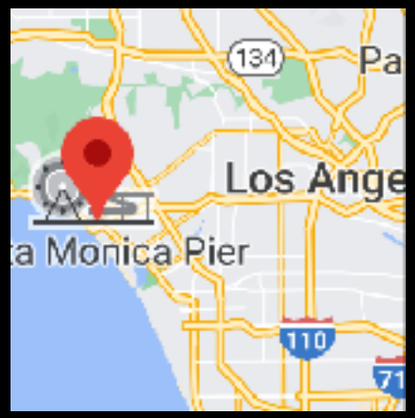
	query	best 1	best 2	best 3
Location				
House age:	20 years	25	21	52
Total area:	3000 sq ft	2768	2819	2680
bedrooms area:	1500 sq ft	850	648	740
population:	1000	1558	1435	1587
households:	500	784	593	713
income:	8.0	3.7	3.9	2.6
house value:	360,000 $	360,000	360,200	359,600

Custom norm

$$\|v\| = \sum \alpha_i |x_i|$$

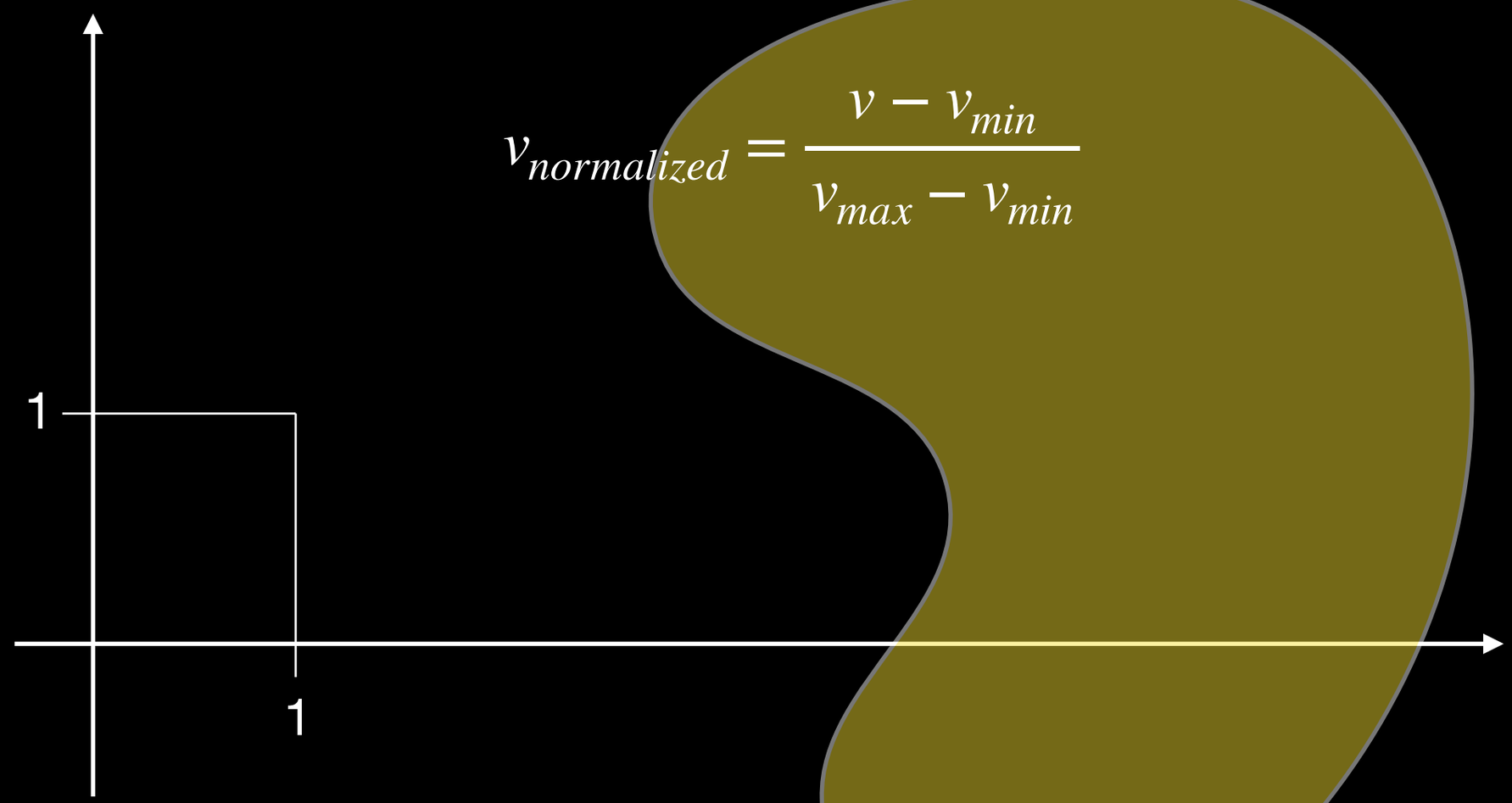
$$\alpha = (10^4, 10^4, 10^{-1}, 10^{-3}, 10^{-3}, 10^{-2}, 10^{-2}, 10^0, 10^{-4})$$

Closest vectors (custom norm)

	query	best 1	best 2	best 3
Location				
House age:	20 years	35	24	24
Total area:	3000 sq ft	2914	7418	2924
bedrooms area:	1500 sq ft	934	1755	1013
population:	1000	1334	2713	1492
households:	500	870	1577	943
income:	8.0 K$	2.99	5.09	2.8
house value:	360,000 $	350,000	500,000	291,700

Data normalization

$$v_{normalized} = \frac{v - v_{min}}{v_{max} - v_{min}}$$



```python
mi = data.min(axis=0)
mi
```

```
array([-1.2435e+02,  3.2540e+01,  1.0000e+00,  2.0000e+00,  1.0000e+00,
        3.0000e+00,  1.0000e+00,  4.9990e-01,  1.4999e+04])
```

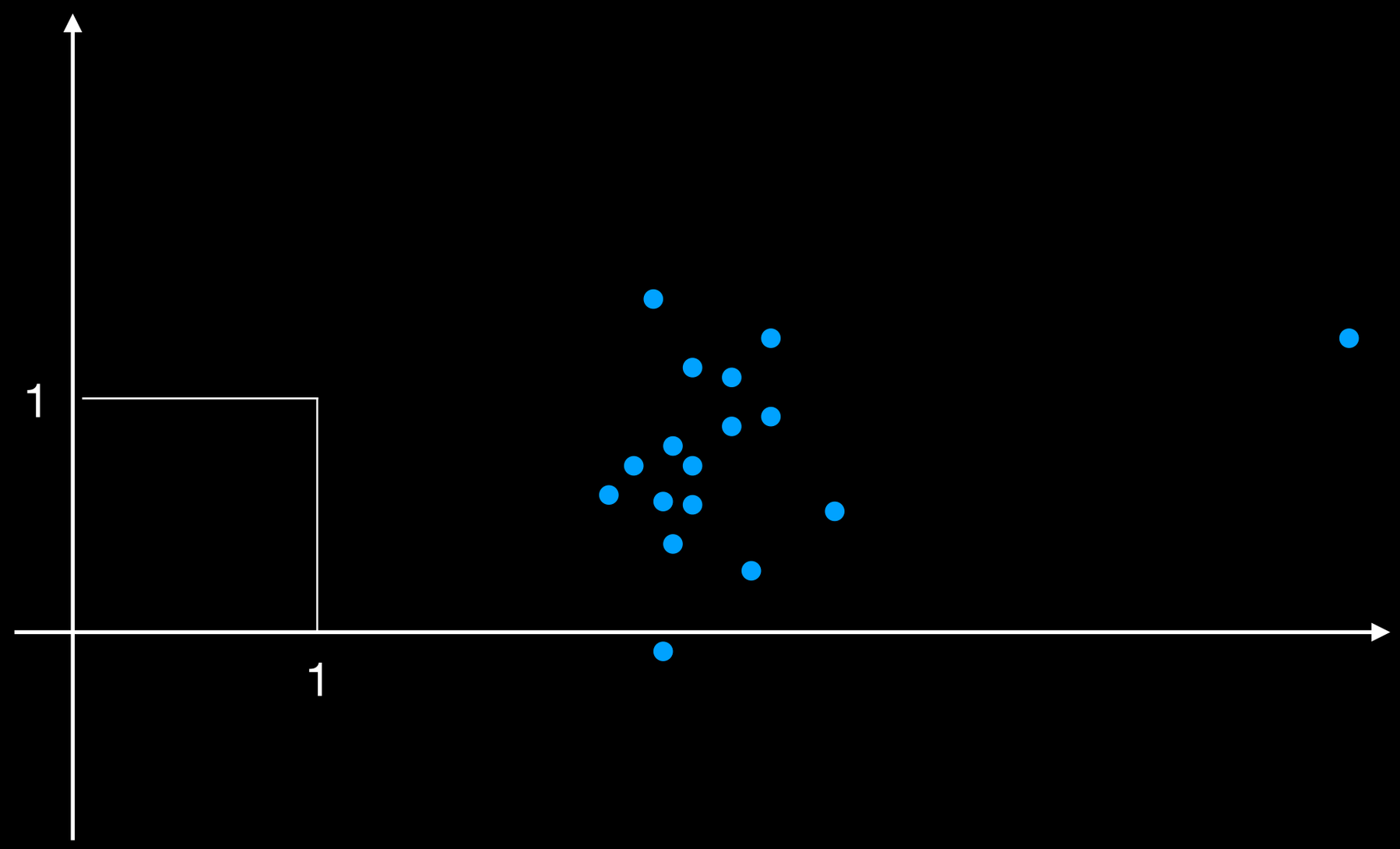
```python
ma = data.max(axis=0)
ma
```

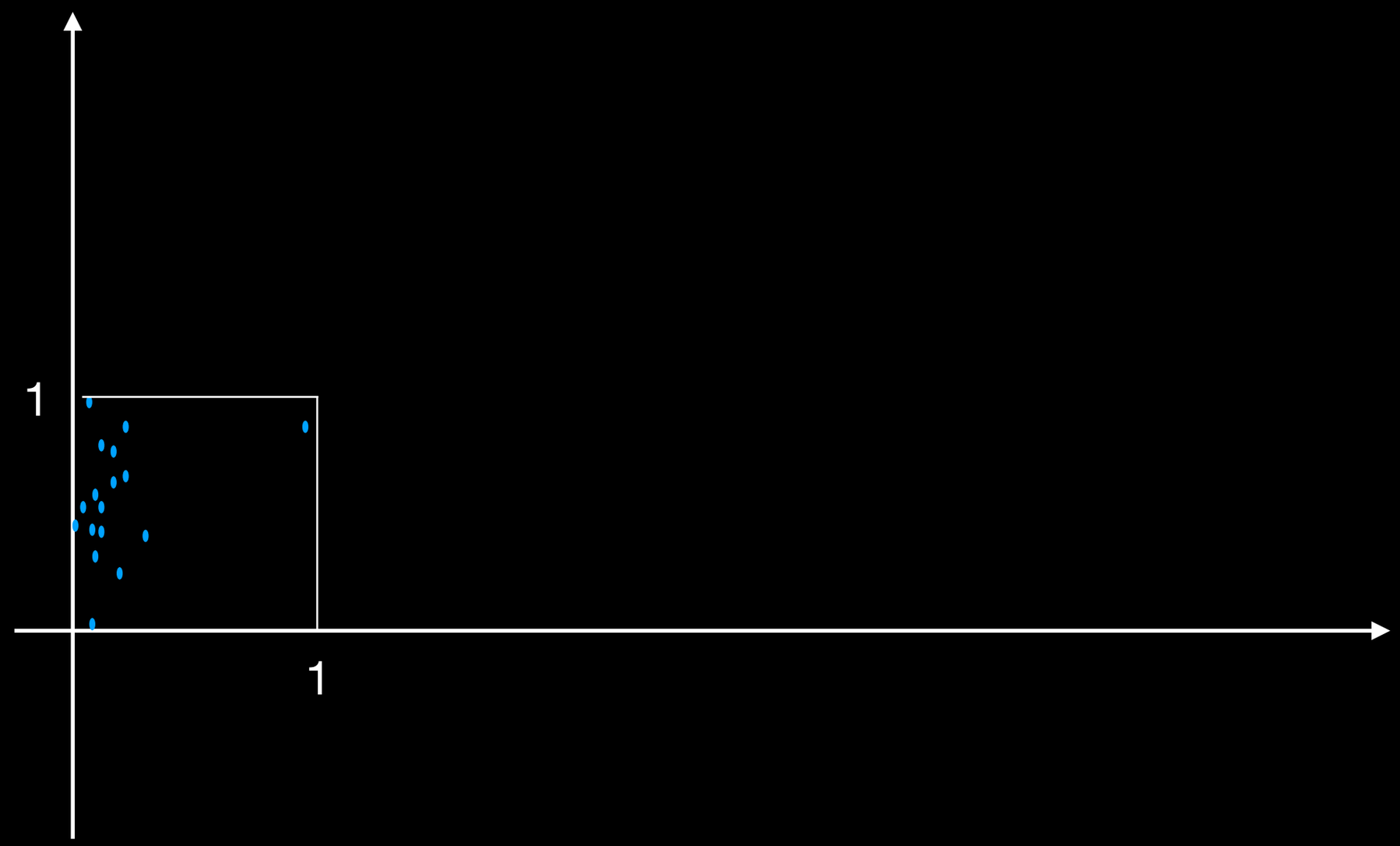
```
array([-1.14310e+02,  4.19500e+01,  5.20000e+01,  3.93200e+04,
        6.44500e+03,  3.56820e+04,  6.08200e+03,  1.50001e+01,
        5.00001e+05])
```

```python
1/(ma-mi)
```

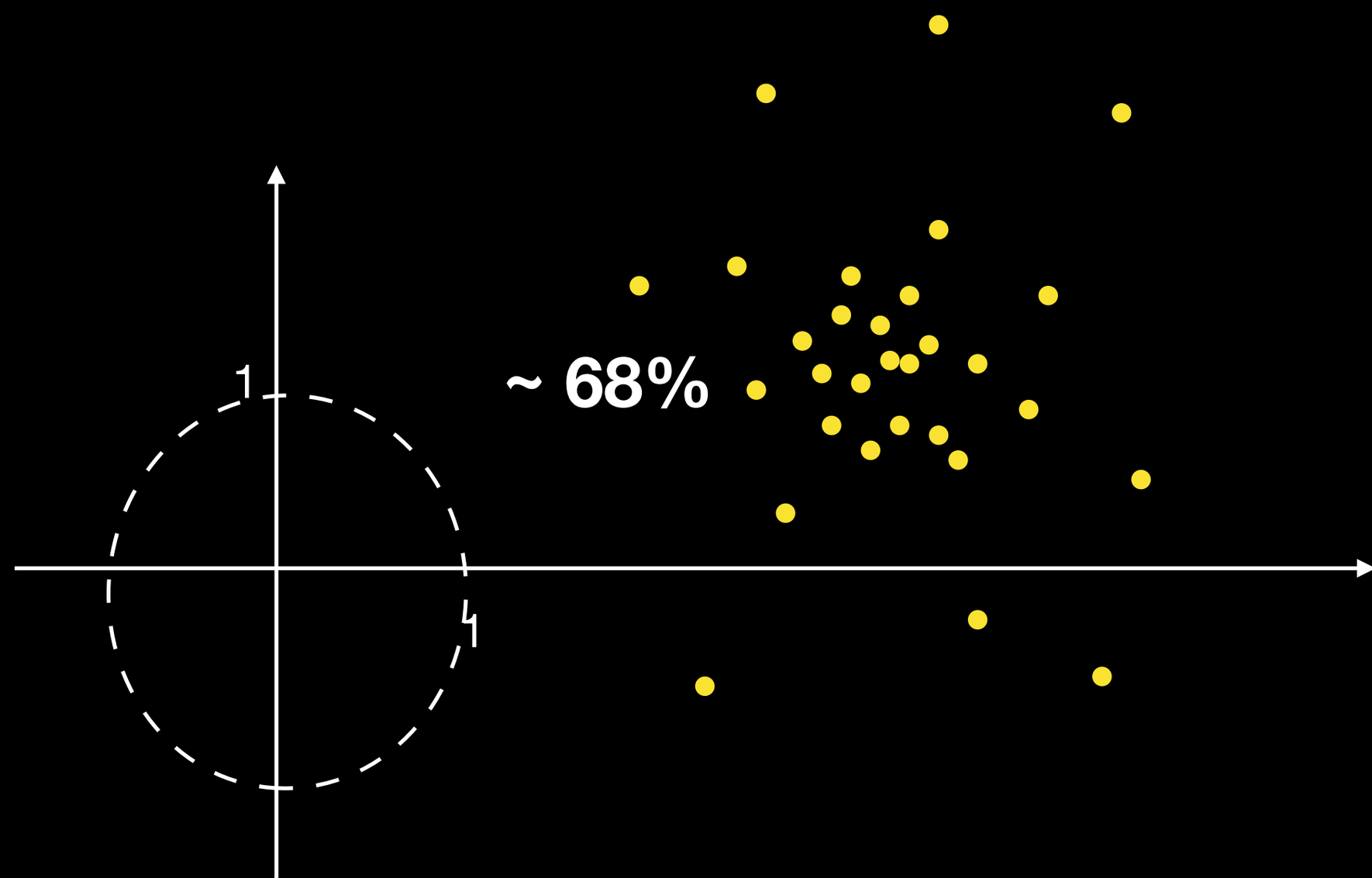
```
array([9.96015936e-02, 1.06269926e-01, 1.96078431e-02, 2.54336436e-05,
       1.55183116e-04, 2.80276914e-05, 1.64446637e-04, 6.89645660e-02,
       2.06184717e-06])
```

$$\alpha = (10^4,\ 10^4,\ 10^{-1},\ 10^{-3},\ 10^{-3},\ 10^{-2},\ 10^{-2},\ 10^0,\ 10^{-4})$$


1
1


1
1

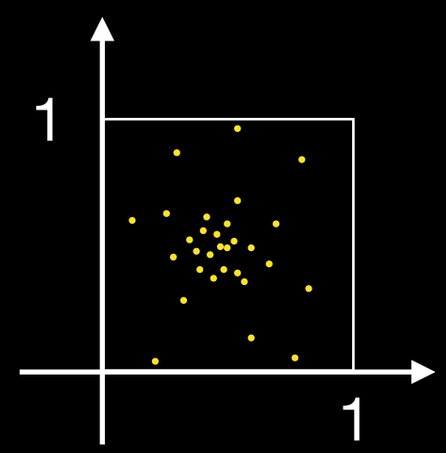
Data standardization



Summary

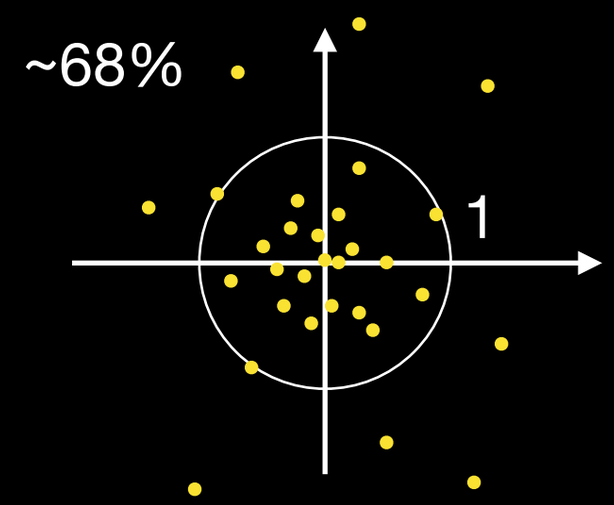
Normalization

data distributed uniformly



Standardization

data distributed normally



Common metrics can be used

Custom norm

Fine-tuned way of distance measuring