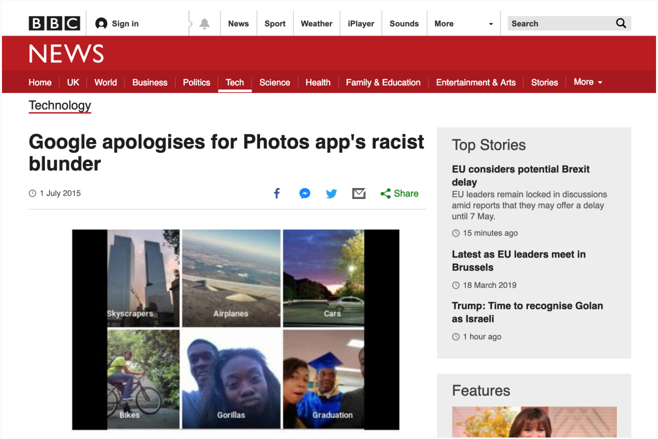
Evolutionary dataset optimisation: learning algorithm quality through evolution

Henry Wilde, Dr. Jonathan Gillard, Dr. Vincent Knight



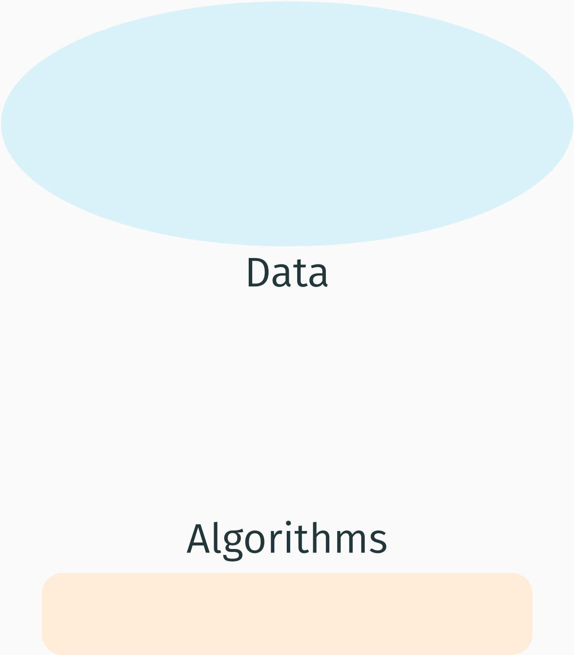
via: BBC News (`https://www.bbc.co.uk/news/technology-33347866`)

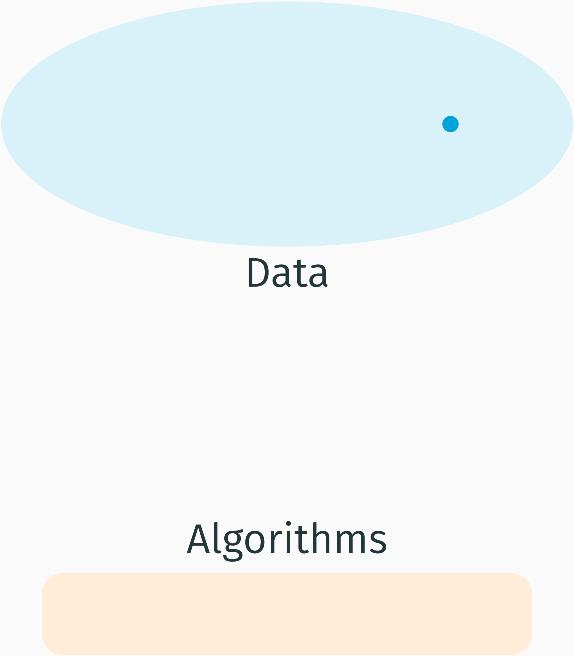
Reliability

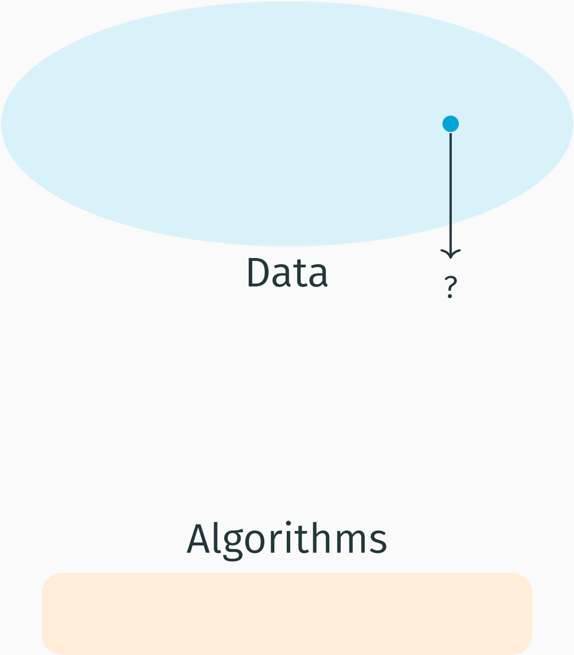
R. Hyndman. *Prediction competitions*. 2014. URL: `https://robjhyndman.com/hyndsight/prediction-competitions/`

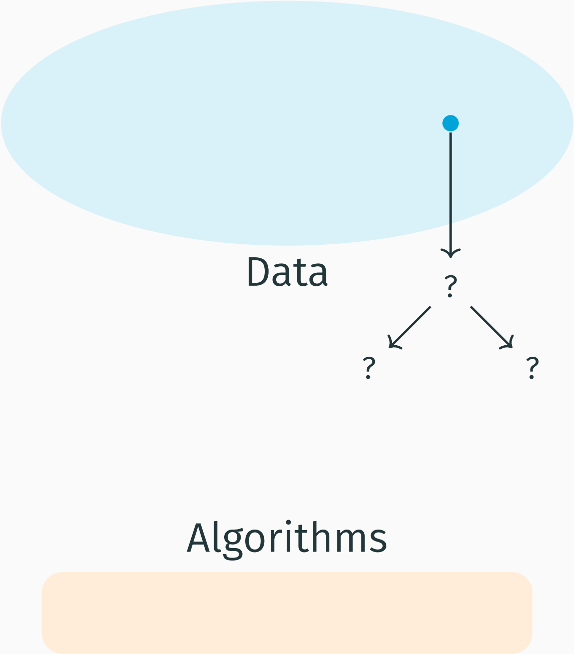
Frailty

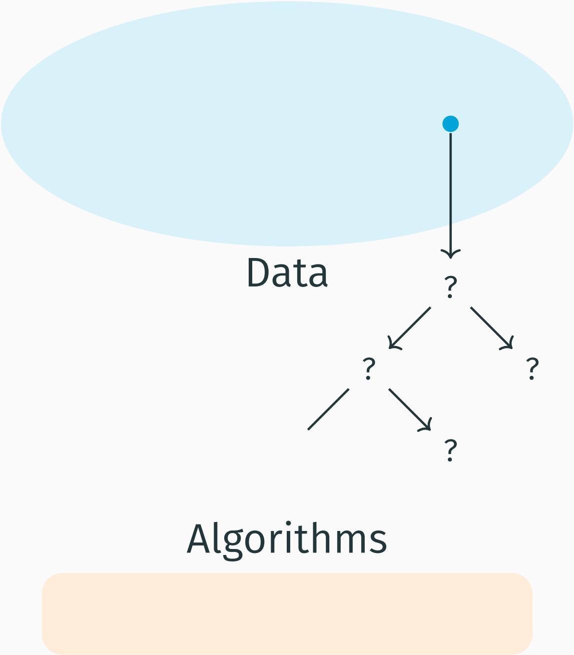
A. Torralba and A. A. Efros. “Unbiased Look at Dataset Bias”. In: *Proceedings of the 2011 IEEE Conference on Computer Vision and Pattern Recognition*. 2011, pp. 1521–1528. DOI: `10.1109/CVPR.2011.5995347`

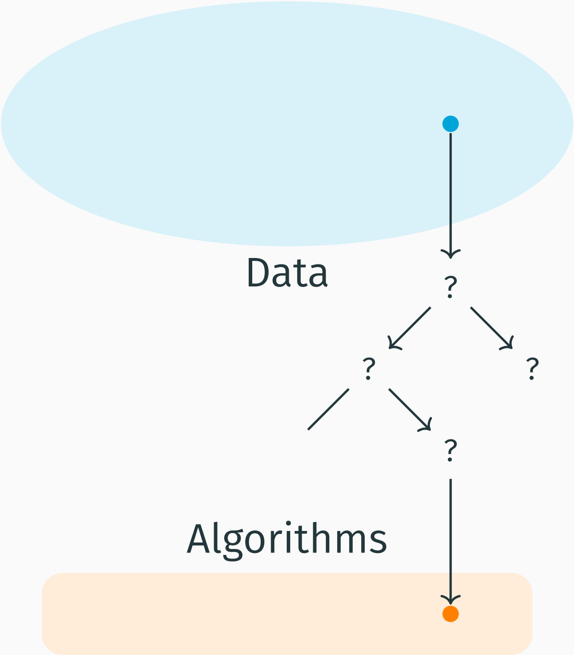

Data
Algorithms

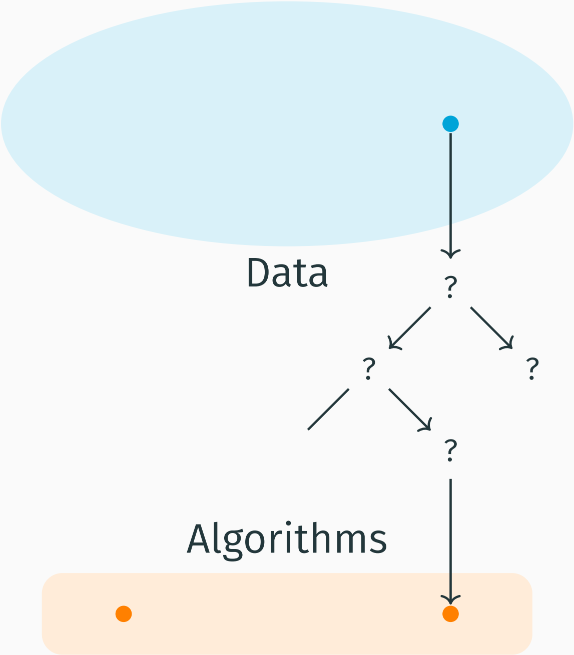

Data
Algorithms

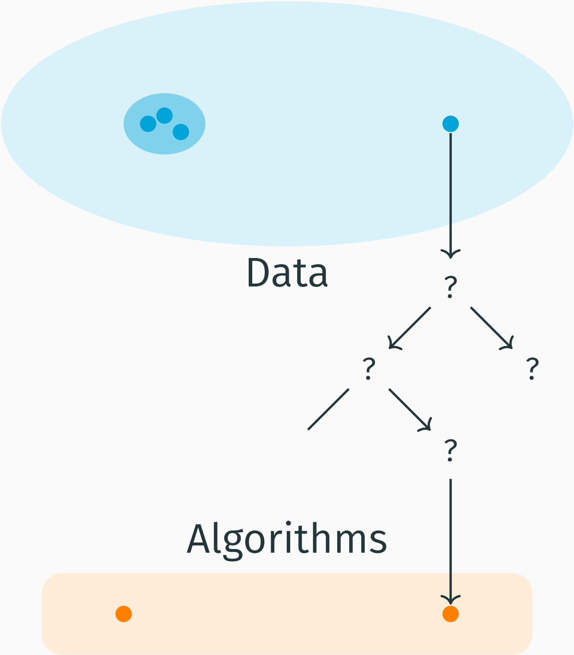

Data
?
Algorithms

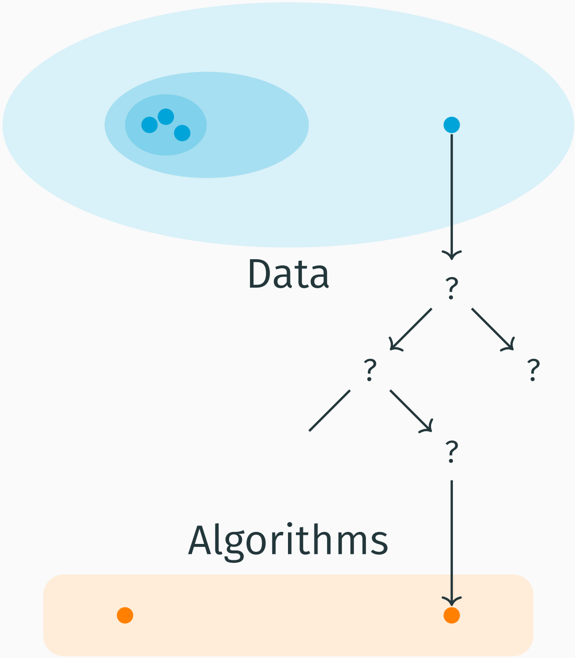

Data
?
?
?
Algorithms

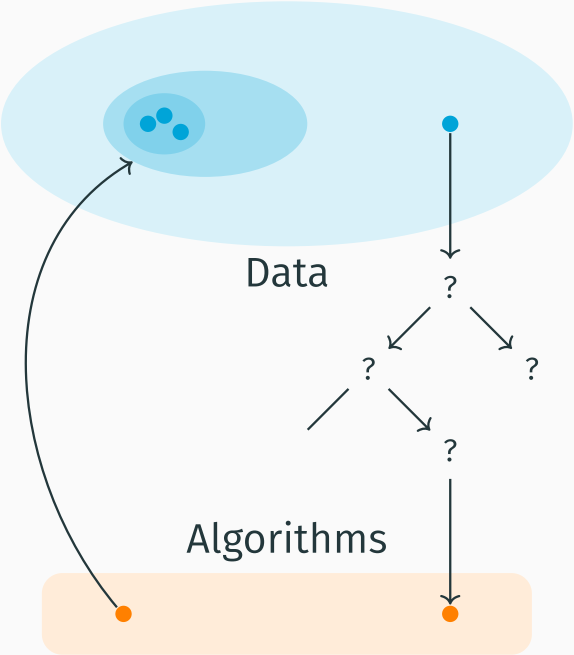

Data
?
?
?
?
Algorithms


Data
?
?
?
?
Algorithms


Data
?
?
?
?
Algorithms


Data
?
?
?
?
Algorithms


Data
?
?
?
?
Algorithms

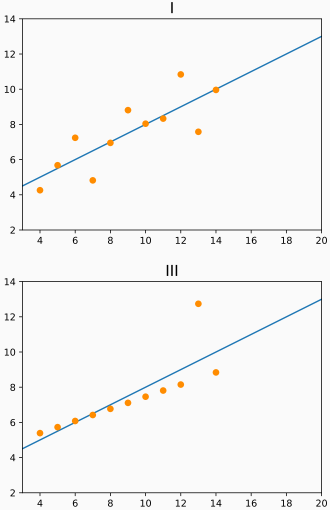

Data
?
?
?
?
Algorithms

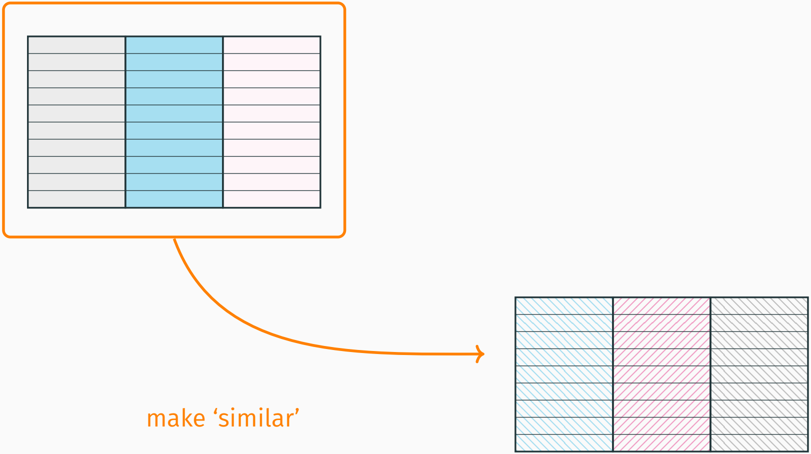
Generating artificial data



via: `https://thispersondoesnotexist.com`

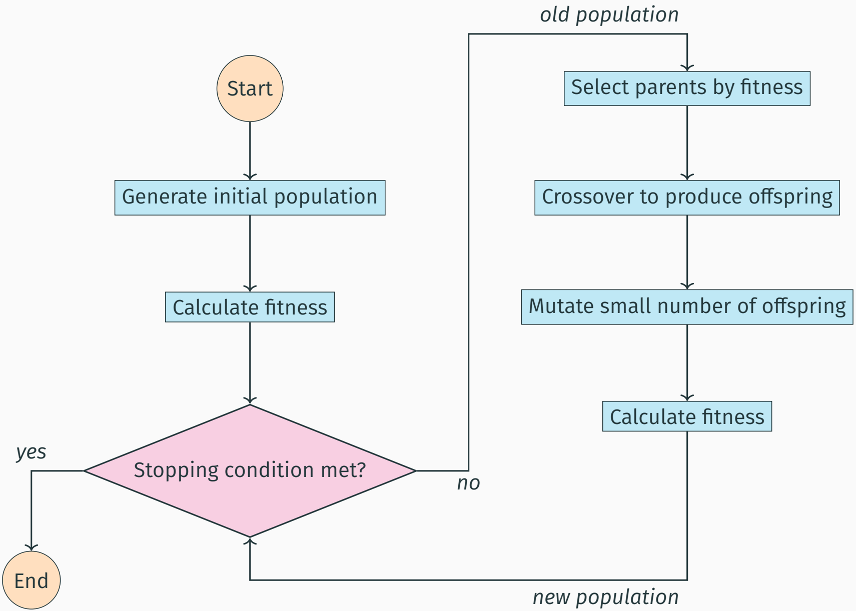
Anscombe's quartet




make ‘similar’

Given an algorithm, how can one find sets of data for which it performs well?

Evolutionary algorithms



Start
Generate initial population
Calculate fitness
Stopping condition met?
yes
no
End
old population
Select parents by fitness
Crossover to produce offspring
Mutate small number of offspring
Calculate fitness
new population

$$\max \quad f : \mathbb{N}^2 \to \mathbb{N}; \qquad f(x_1, x_2) = x_1 + x_2$$

$$\max \quad f : \mathbb{N}^2 \to \mathbb{N}; \qquad f(x_1, x_2) = x_1 + x_2$$

Population (25, 30) (12, 1) (11, 0) (20, 12) (24, 25)

$$\max \quad f : \mathbb{N}^2 \to \mathbb{N}; \qquad f(x_1, x_2) = x_1 + x_2$$

Population	(25, 30)	(12, 1)	(11, 0)	(20, 12)	(24, 25)
Get fitness	55	13	11	42	49

$$\max \quad f : \mathbb{N}^2 \to \mathbb{N}; \qquad f(x_1, x_2) = x_1 + x_2$$

Population	(25, 30)	(12, 1)	(11, 0)	(20, 12)	(24, 25)
Get fitness	55	13	11	42	49
Select parents	(25, 30)			(20, 12)	(24, 25)

$$\max \quad f : \mathbb{N}^2 \to \mathbb{N}; \qquad f(x_1, x_2) = x_1 + x_2$$

Population	(25, 30)	(12, 1)	(11, 0)	(20, 12)	(24, 25)
Get fitness	55	13	11	42	49
Select parents	(25, 30)			(20, 12)	(24, 25)
Create offspring		(24, 30)	(25, 12)		

$$\max \quad f : \mathbb{N}^2 \to \mathbb{N}; \qquad f(x_1, x_2) = x_1 + x_2$$

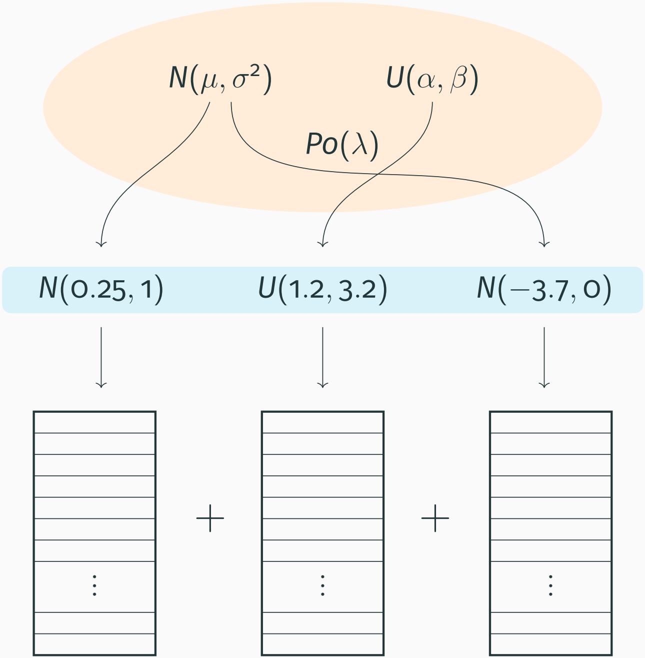
Population	(25, 30)	(12, 1)	(11, 0)	(20, 12)	(24, 25)
Get fitness	55	13	11	42	49
Select parents	(25, 30)			(20, 12)	(24, 25)
Create offspring		(24, 30)	(25, 12)		
Mutate offspring		(24, 30)	(25, 13)		

$$\max \quad f : \mathbb{N}^2 \to \mathbb{N}; \qquad f(x_1, x_2) = x_1 + x_2$$

Population	(25, 30)	(12, 1)	(11, 0)	(20, 12)	(24, 25)
Get fitness	55	13	11	42	49
Select parents	(25, 30)			(20, 12)	(24, 25)
Create offspring		(24, 30)	(25, 12)		
Mutate offspring		(24, 30)	(25, 13)		
New generation	(25, 30)	(24, 30)	(25, 13)	(20, 12)	(24, 25)

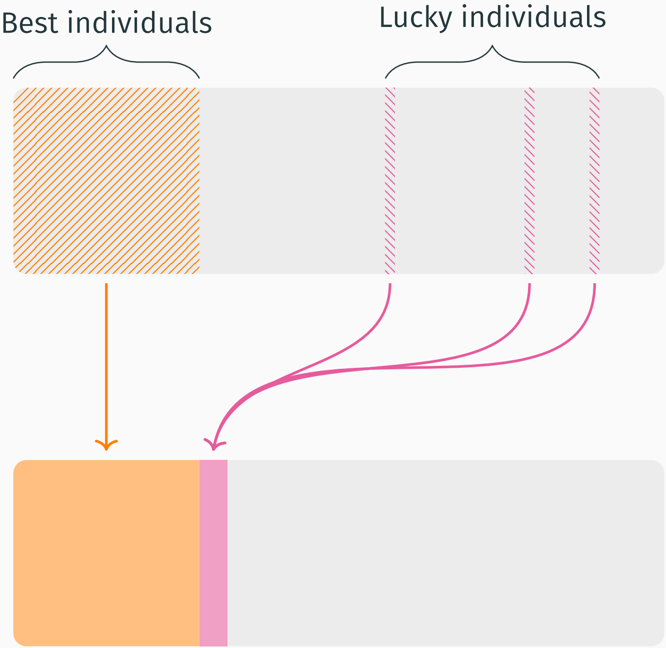
$$\max \quad f : \mathbb{N}^2 \to \mathbb{N}; \qquad f(x_1, x_2) = x_1 + x_2$$

Population	(25, 30)	(12, 1)	(11, 0)	(20, 12)	(24, 25)
Get fitness	55	13	11	42	49
Select parents	(25, 30)			(20, 12)	(24, 25)
Create offspring		(24, 30)	(25, 12)		
Mutate offspring		(24, 30)	(25, 13)		
New generation	(25, 30)	(24, 30)	(25, 13)	(20, 12)	(24, 25)


$N(\mu, \sigma^2)$
$U(\alpha, \beta)$
$Po(\lambda)$
$N(0.25, 1)$
$U(1.2, 3.2)$
$N(-3.7, 0)$
$+$
$+$

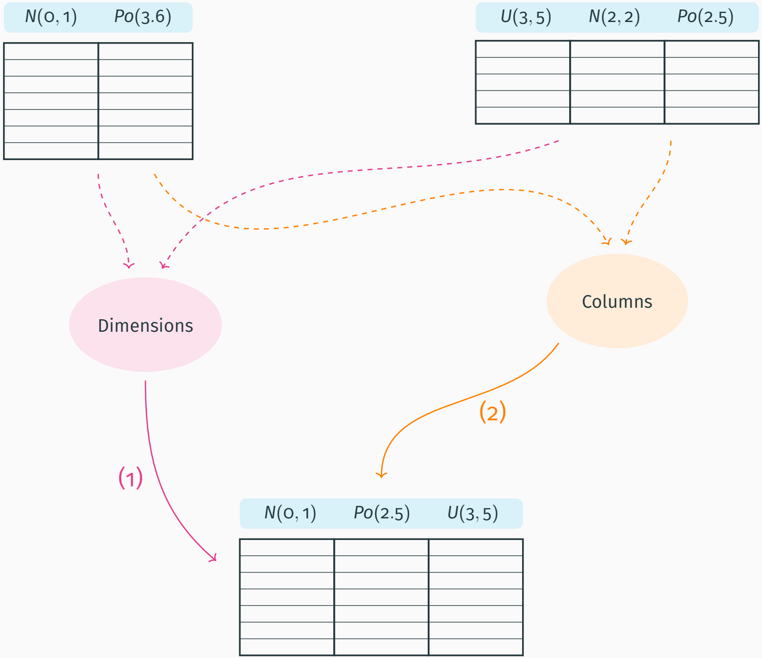
$$\max \quad f : \mathbb{N}^2 \to \mathbb{N}; \qquad f(x_1, x_2) = x_1 + x_2$$

Population	(25, 30)	(12, 1)	(11, 0)	(20, 12)	(24, 25)
Get fitness	55	13	11	42	49
Select parents	(25, 30)			(20, 12)	(24, 25)
Create offspring		(24, 30)	(25, 12)		
Mutate offspring		(24, 30)	(25, 13)		
New generation	(25, 30)	(24, 30)	(25, 13)	(20, 12)	(24, 25)


Best individuals
Lucky individuals

$$\max \quad f : \mathbb{N}^2 \to \mathbb{N}; \qquad f(x_1, x_2) = x_1 + x_2$$

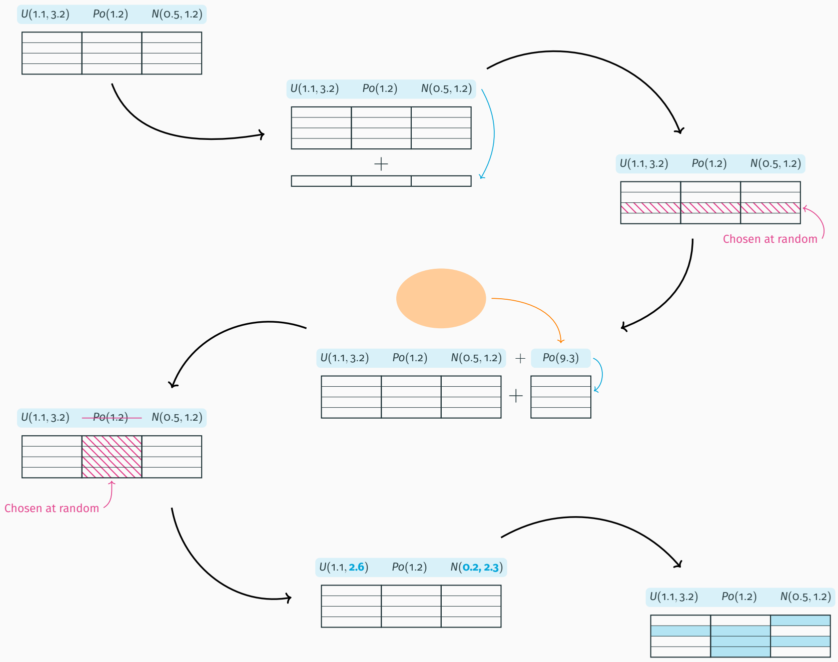
Population	(25, 30)	(12, 1)	(11, 0)	(20, 12)	(24, 25)
Get fitness	55	13	11	42	49
Select parents	(25, 30)			(20, 12)	(24, 25)
Create offspring		(24, 30)	(25, 12)		
Mutate offspring		(24, 30)	(25, 13)		
New generation	(25, 30)	(24, 30)	(25, 13)	(20, 12)	(24, 25)



$N(0,1)$ $Po(3.6)$
$U(3,5)$ $N(2,2)$ $Po(2.5)$
Dimensions
Columns
(1)
(2)
$N(0,1)$ $Po(2.5)$ $U(3,5)$

$$\max \quad f : \mathbb{N}^2 \to \mathbb{N}; \qquad f(x_1, x_2) = x_1 + x_2$$

Population	(25, 30)	(12, 1)	(11, 0)	(20, 12)	(24, 25)
Get fitness	55	13	11	42	49
Select parents	(25, 30)			(20, 12)	(24, 25)
Create offspring		(24, 30)	(25, 12)		
Mutate offspring		(24, 30)	(25, 13)		
New generation	(25, 30)	(24, 30)	(25, 13)	(20, 12)	(24, 25)

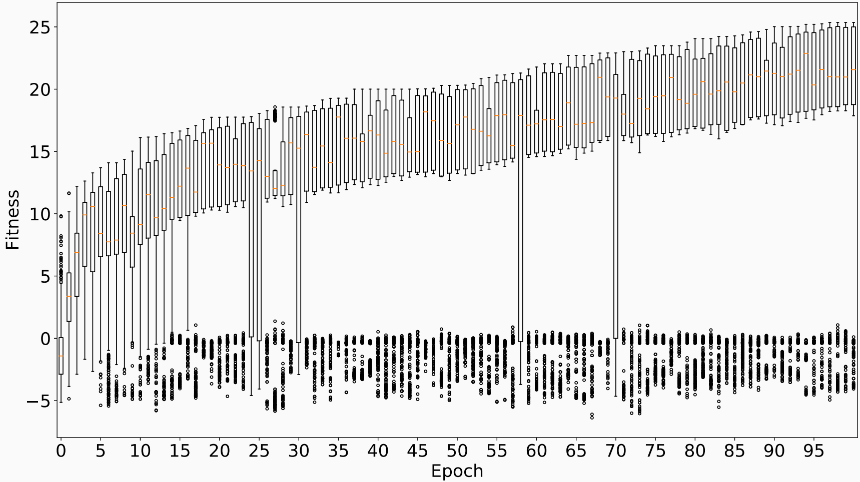


U(1.1, 3.2) Po(1.2) N(0.5, 1.2)
Chosen at random
Po(9.3)
Chosen at random
U(1.1, 2.6) Po(1.2) N(0.2, 2.3)

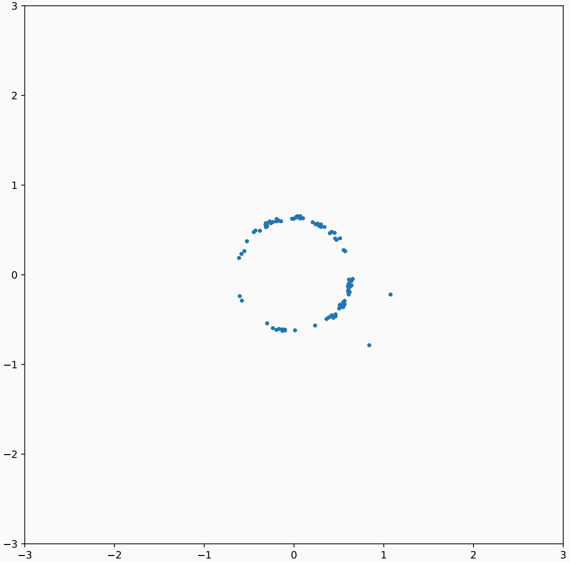
Some example use cases

Maximise

$$f : \mathbb{R}^n \times \mathbb{R}^n \to \mathbb{R}, \qquad f(A, B) = Var(A) - \max_i |B_i - 1|$$



Fitness
Epoch
25
20
15
10
5
0
−5
0
5
10
15
20
25
30
35
40
45
50
55
60
65
70
75
80
85
90
95

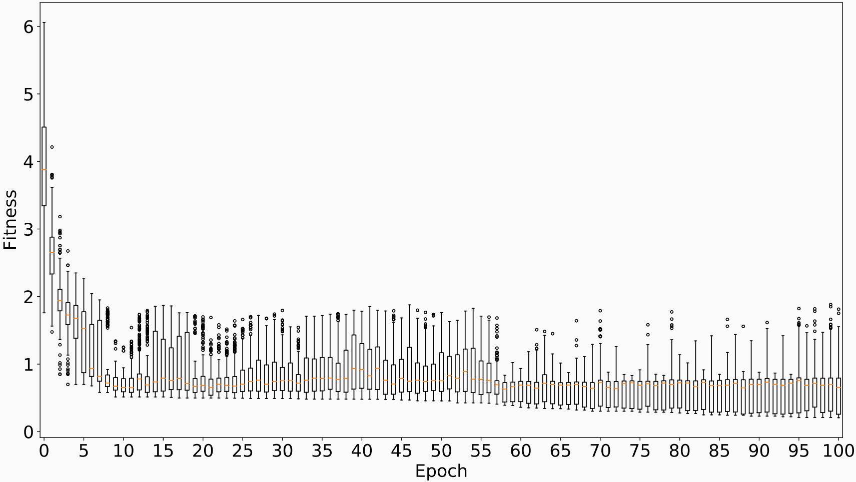


3
2
1
0
−1
−2
−3
−3
−2
−1
0
1
2
3

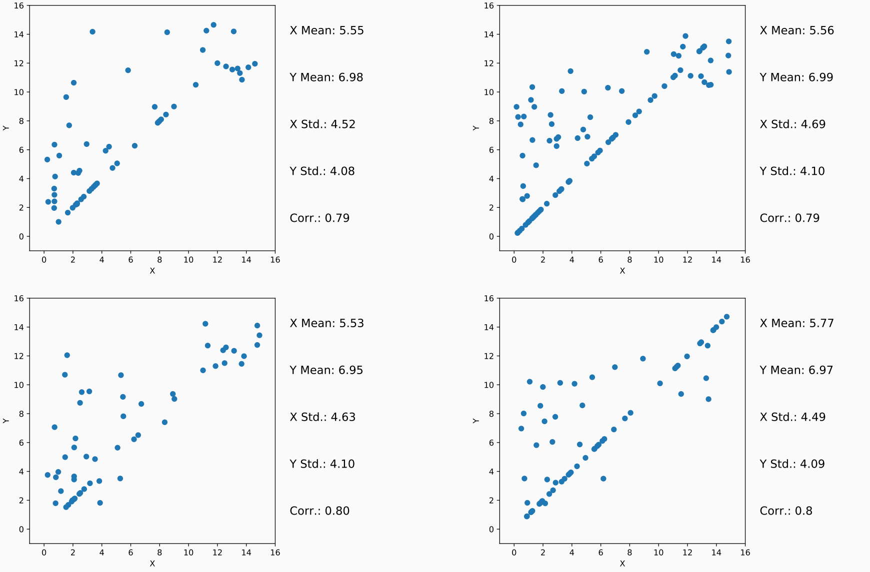
Given a set of k dissimilarity measures:

$$f_1, \ldots, f_k : \mathbb{R}^n \times \mathbb{R}^n \to \mathbb{R}$$

Minimise their sum


Fitness
0
1
2
3
4
5
6
Epoch
0
5
10
15
20
25
30
35
40
45
50
55
60
65
70
75
80
85
90
95
100

X Mean: 5 *Y* Mean: 7 *X* Std.: 4.7 *Y* Std.: 4.1 Corr.: 0.8



Henry Wilde

Twitter: @daffidwilde

Email: wildehd@cardiff.ac.uk

Repository: `github.com/daffidwilde/edo`

Documentation: `edo.readthedocs.io`

Paper in preparation:

"Evolutionary Dataset Optimisation: understanding algorithm quality through evolution"

- A fitness function, f, which acts on a single dataset
- A population size, $N \in \mathbb{N}$
- A maximum number of iterations, $M \in \mathbb{N}$
- A selection parameter to detail the proportion of the fittest individuals to carry forward, $b \in [0, 1]$
- A mutation probability, $p_m \in [0, 1]$

- Limits on the number of rows a dataset can have:

$$R \in \left\{(r_{\min}, r_{\max}) \in \mathbb{N}^2 \mid r_{\min} \leq r_{\max}\right\}$$

- Limits on the number of columns a dataset can have:

$$C := \left(C_1, \dots, C_{|\mathcal{P}|}\right) \text{ where } C_j \in \left\{(c_{\min}, c_{\max}) \in (\mathbb{N} \cup \{\infty\})^2 \mid c_{\min} \leq c_{\max}\right\}$$

  for each $j = 1, \dots, |\mathcal{P}|$
- A set of probability distribution families, $\mathcal{P}$. Each family in this set has some parameter limits which form a part of the overall search space
- A probability vector to sample distributions from $\mathcal{P}$, $w = \left(w_1, \dots, w_{|\mathcal{P}|}\right)$
- A second selection parameter, $l \in [0, 1]$, to allow for a small proportion of "lucky" individuals to be carried forward
- A shrink factor, $s \in [0, 1]$. The relative size of a component of the search space to be retained after adjustment