

Evolutionary dataset optimisation: learning algorithm quality through evolution

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



GIG
CYMRU
NHS
WALES

Bwrdd Iechyd Prifysgol
Cwm Taf
University Health Board

Motivation





Sign in



News

Sport

Weather

iPlayer

Sounds

More

Search



NEWS

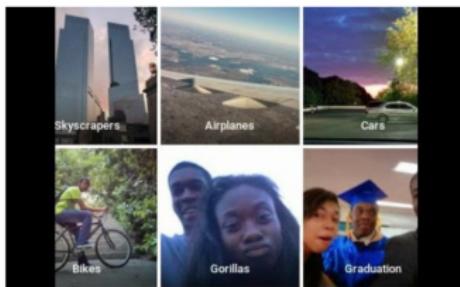
[Home](#) | [UK](#) | [World](#) | [Business](#) | [Politics](#) | [Tech](#) | [Science](#) | [Health](#) | [Family & Education](#) | [Entertainment & Arts](#) | [Stories](#) | [More](#) ▾

[Technology](#)

Google apologises for Photos app's racist blunder

⌚ 1 July 2015

f t e Share



Top Stories

EU considers potential Brexit delay

EU leaders remain locked in discussions amid reports that they may offer a delay until 7 May.

⌚ 15 minutes ago

Latest as EU leaders meet in Brussels

⌚ 18 March 2019

Trump: Time to recognise Golan as Israeli

⌚ 1 hour ago

Features



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

Reliability and frailty

A. Torralba and A. A. Efros. *Unbiased Look at Dataset Bias*. Proceedings of the 2011 IEEE Conference on Computer Vision and Pattern Recognition. 2011. DOI: [10.1109/CVPR.2011.5995347](https://doi.org/10.1109/CVPR.2011.5995347)

Generating artificial data



via: <https://thispersondoesnotexist.com>

Barney Sparks

✉ barneysparks@gmail.com 🗺 Allentown, Pennsylvania



EXPERIENCE

Operations Analyst

Youth 2014 - Ongoing Allentown, Pennsylvania

Youth is a leading platform that is developing a product in cloud and in-house platforms.

- Provided support of over 50 international startups and change
- Managed a team of 10 people in a staff of 10 people
- Managed 10 projects with 4 projects per day management of the company and included an average of 200 companies and 2 employees in the first role.
- Increased the company by 100% in 2014 and 2011 and 2015.

Head Of Marketing And Controller

X-Main 2010 - 2014 Allentown, Pennsylvania

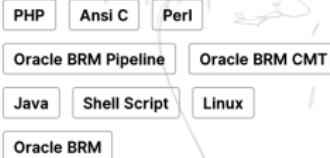
X-main is a software company that provides young people with leadership development and enterprise software used in Junico and the construction industry.

- Created 3 consultants to provide internal and staff & maintenance strategies and incorporated the company with timely companies resulting in \$2MM in conduct and increased the company to 100 people in less than 3 months

SUMMARY

I am a highly motivated and proven pursuits where my work diverse individual with others into the field of interest and be like interependency to go to their practice clinic.

TECHNOLOGIES



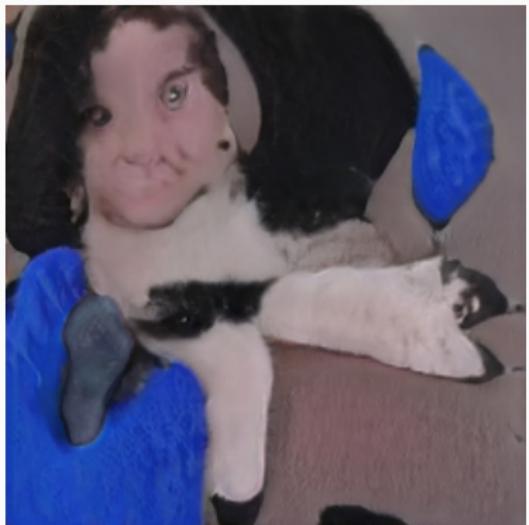
EDUCATION

Unige, Faculty Of Medicine

University Of Geneva

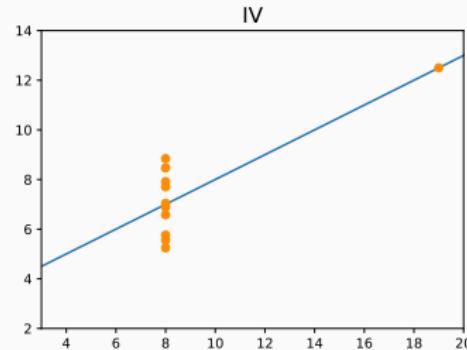
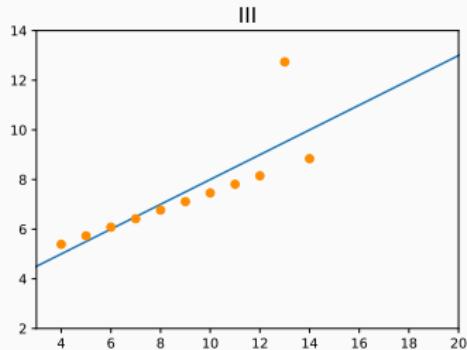
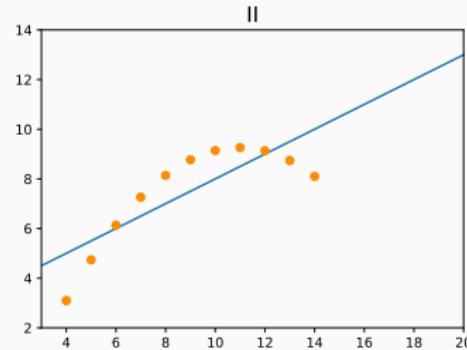
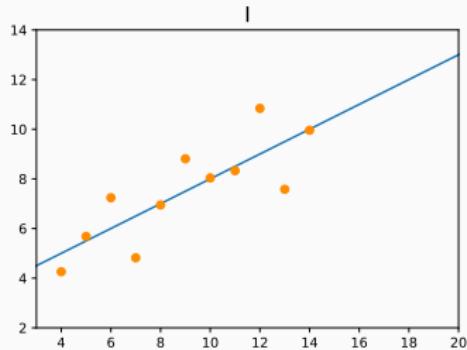
2000 2001 2002 2003 2004

via: <https://thisresumedoesnotexist.com>

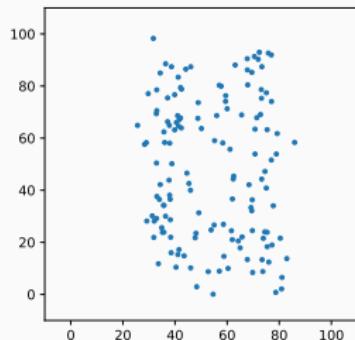
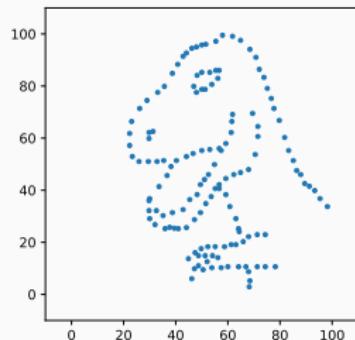
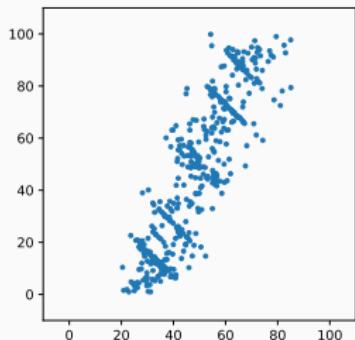


via: <https://thiscatdoesnotexist.com>

Anscombe's quartet

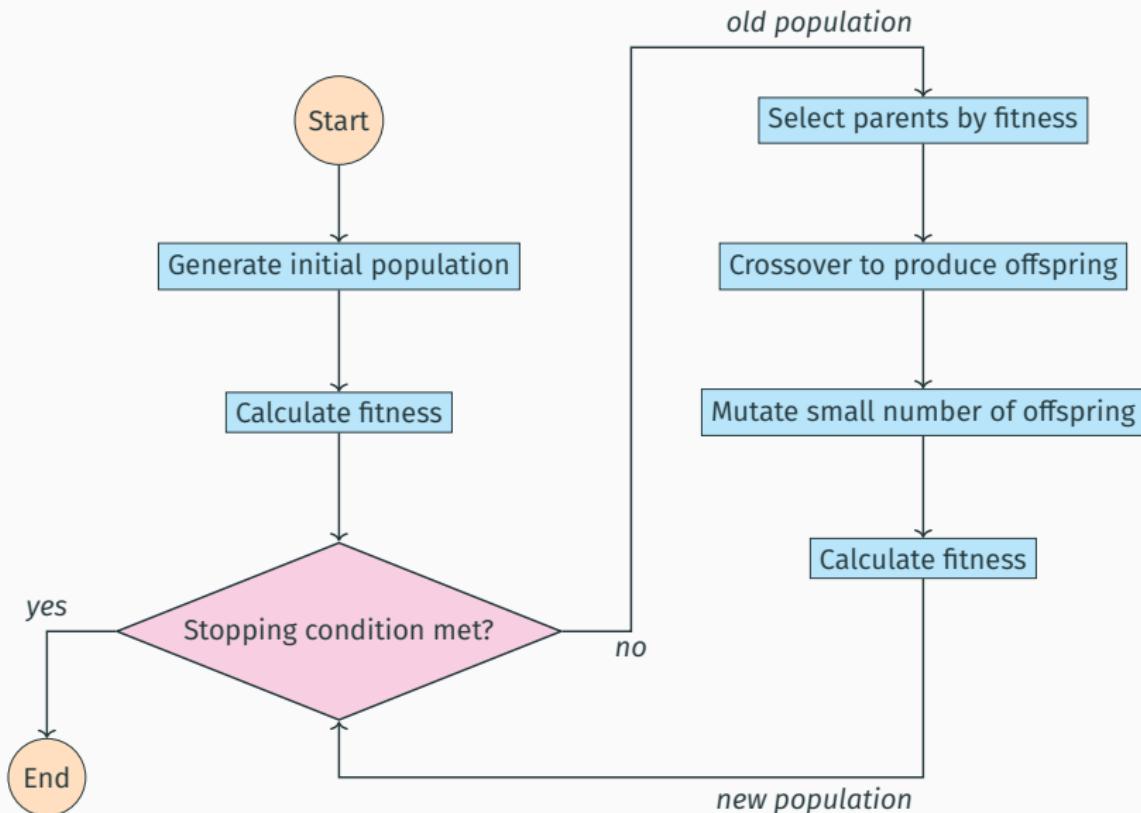


The Datasaurus dozen



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

Evolutionary algorithms



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

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

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

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

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

Get fitness 55 13 11 42 49

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

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)

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

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)		

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

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 (15, 30) (25, 26)

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

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 (15, 30) (25, 26)

New generation (25, 30) (15, 30) (25, 26) (20, 12) (24, 25)

- A fitness function, f , which acts on a single dataset
- A population size, $N \in \mathbb{N}$
- Limits on the number of rows a dataset can have:

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

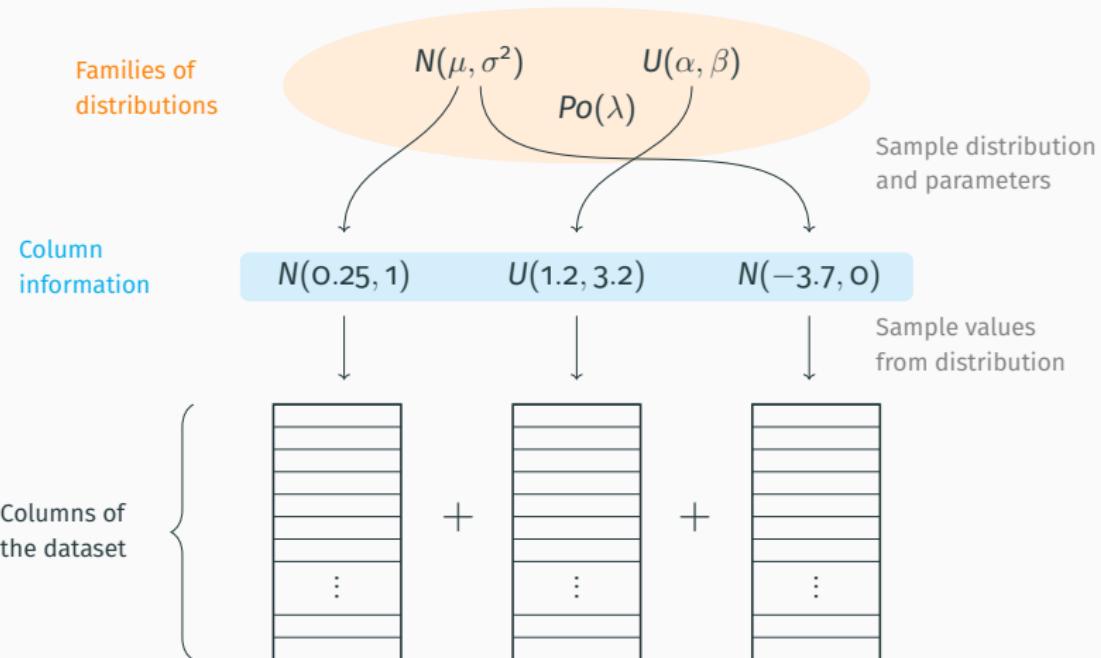
- Limits on the number of columns a dataset can have:

$$C := (c_1, \dots, c_{|\mathcal{P}|}) \text{ where } c_j \in \left\{ (c_{\min}, c_{\max}) \in (\mathbb{N} \cup \{\infty\})^2 \mid c_{\min} \leq c_{\max} \right\}$$

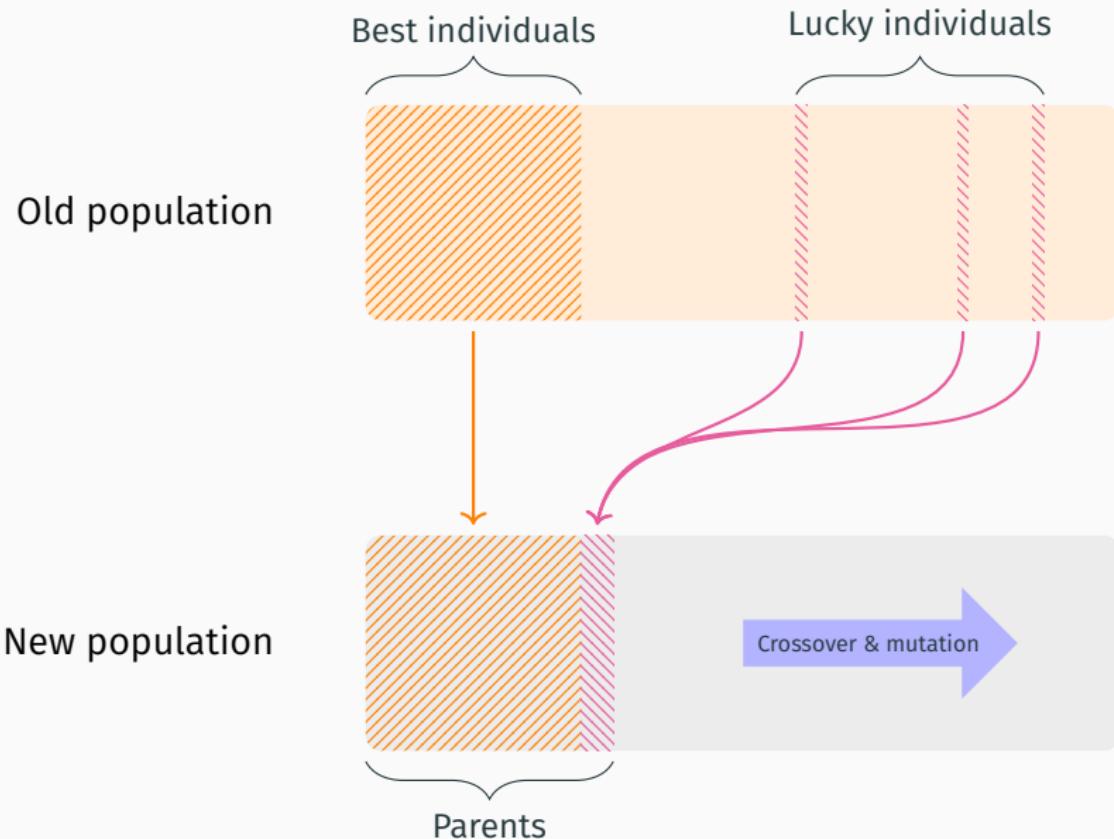
for each $j = 1, \dots, |\mathcal{P}|$

- 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} , $w = (w_1, \dots, w_{|\mathcal{P}|})$
- A maximum number of iterations, $M \in \mathbb{N}$
- Two selection parameters: one to indicate the proportion of the fittest individuals to carry forward, $b \in [0, 1]$, and the other to allow for a small proportion of “lucky” individuals, $l \in [0, 1]$
- A mutation probability, $p_m \in [0, 1]$
- A shrink factor, $s \in [0, 1]$. The relative size of a component of the search space to be retained after adjustment

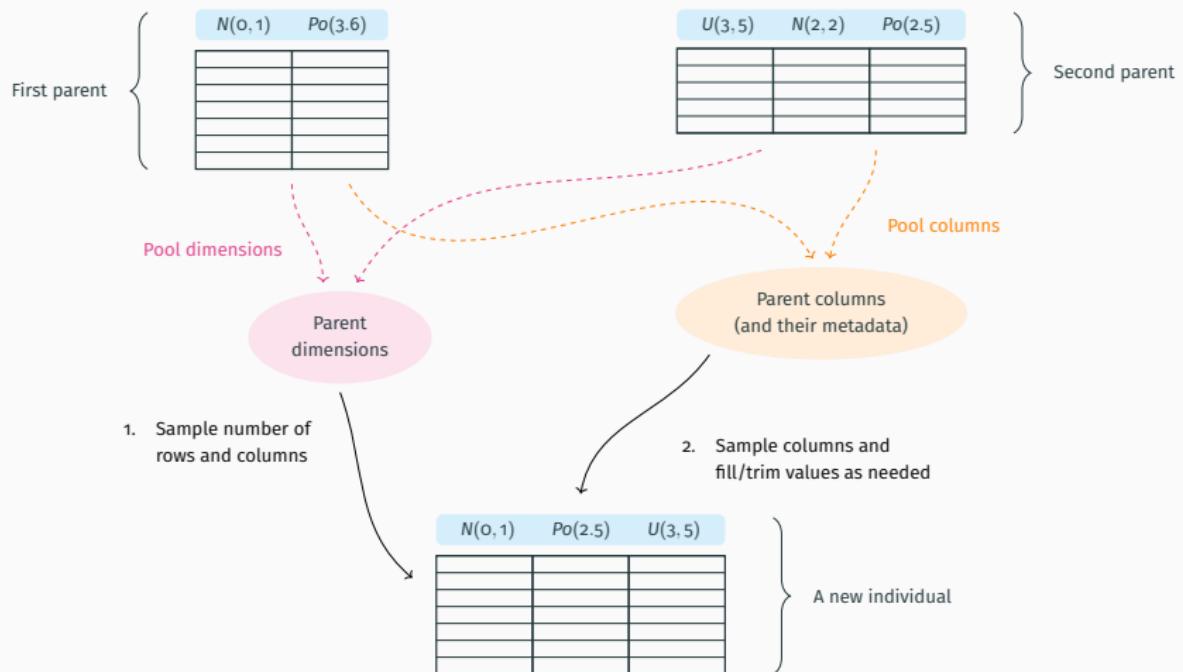
Individuals



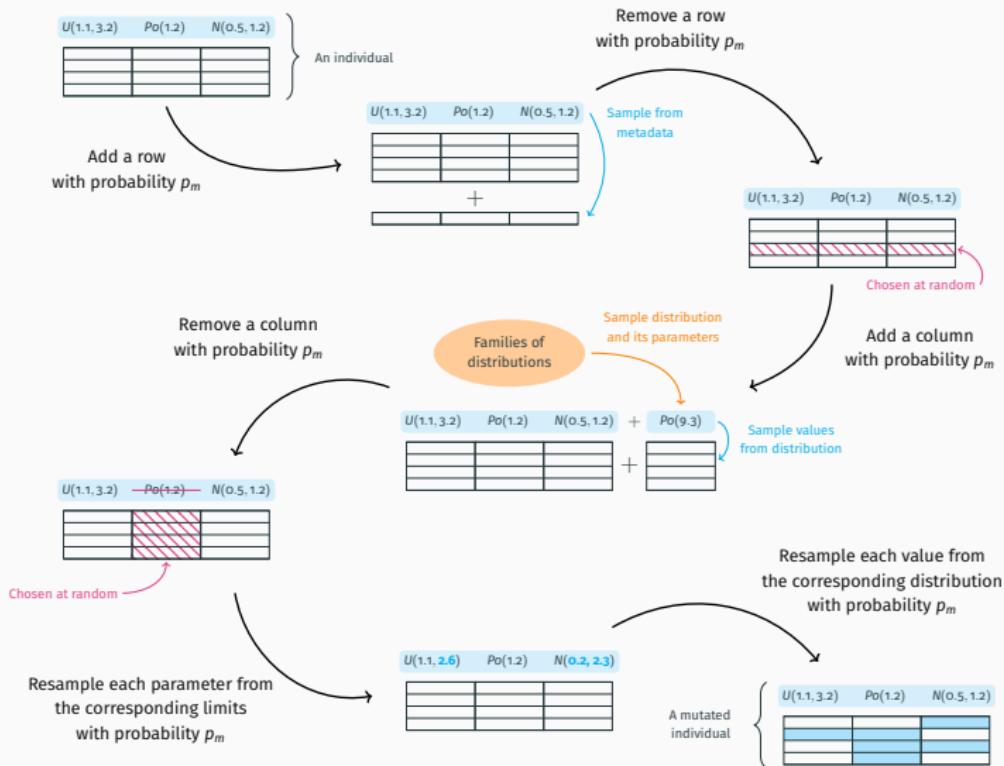
Selection



Crossover



Mutation



Some example use cases

Input: A dataset, $X \in \mathbb{R}^n \times \mathbb{R}^n$;

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

Output: The maximal value for f

values $\leftarrow \emptyset$

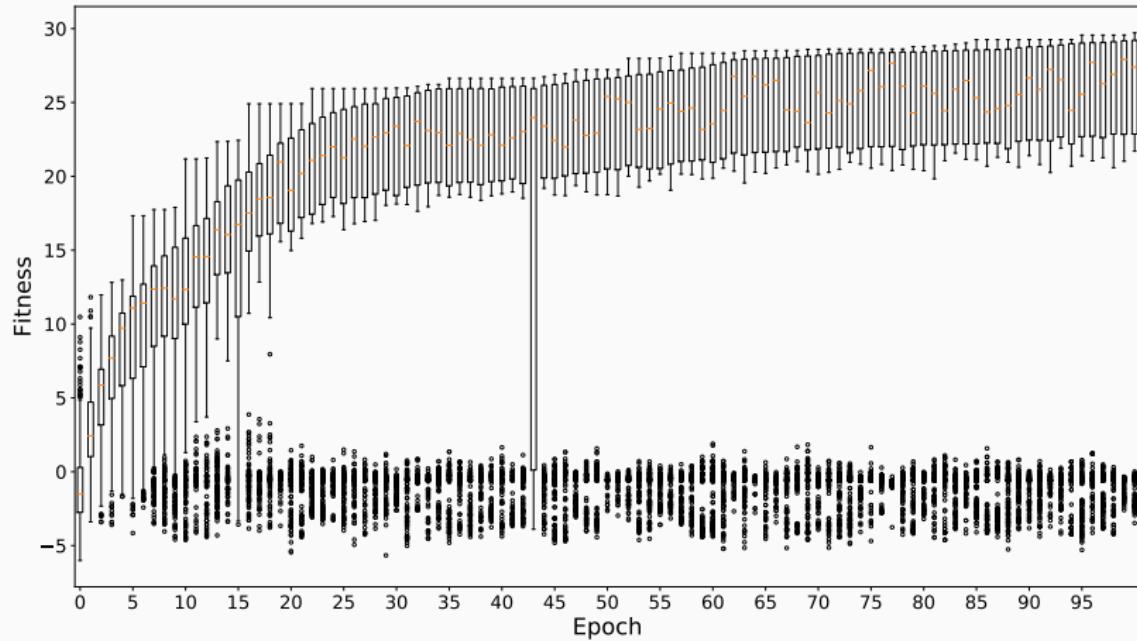
for each ordering, $a, b \in \{0, 1\}^2$, of the columns **do**

calculate $f(X^{(a)}, X^{(b)})$

append this to *values*

end

return max *values*



Input: A dataset, $X \in \mathbb{R}^n$; some sampling proportion, $p \in [0, 1]$;
a number of samples to take, k

Output: The maximum **difference** between each sampled
mean of X and 0

values $\leftarrow \emptyset$

for $i = 1, \dots, k$ **do**

$Y \leftarrow$ a random sample of $\lfloor np \rfloor$ entries from X
evaluate the mean of Y :

$$\bar{Y} = \frac{1}{|Y|} \sum_{i=1}^{|Y|} Y_i$$

append \bar{Y} to *values*

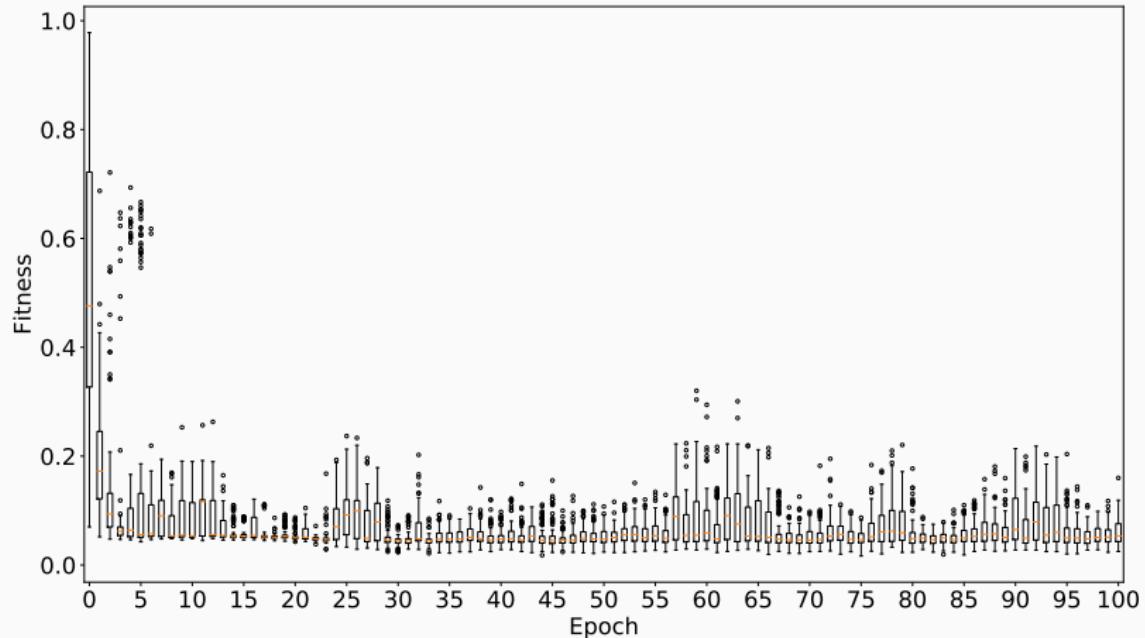
end

return $\max values$

X {

$U(-0.06, 0.91)$		Metadata
<hr/>		
o		
0	-0.142152	
1	0.0433848	
2	0.0727756	
...	...	
45	0.0670734	
46	0.0866879	
47	0.0137398	
<hr/>		

} Dataset

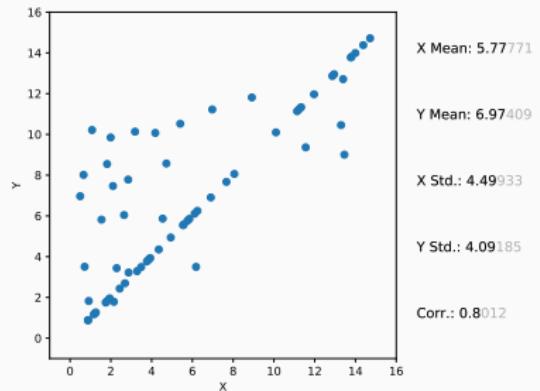
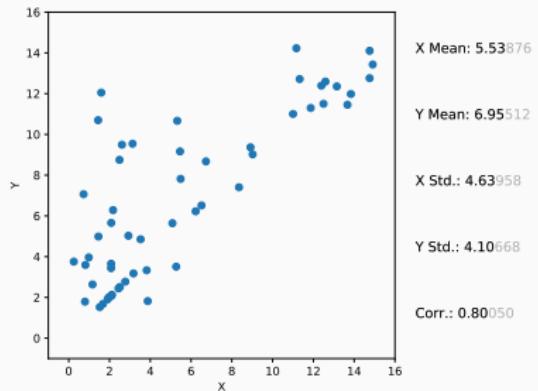
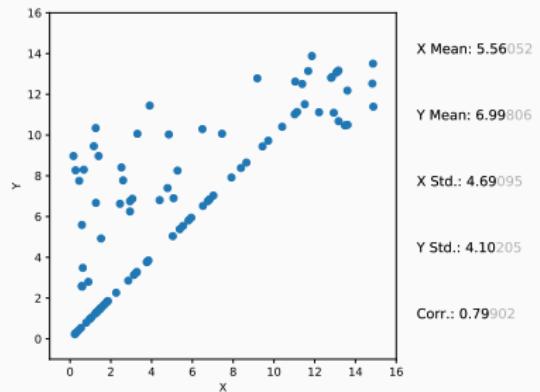
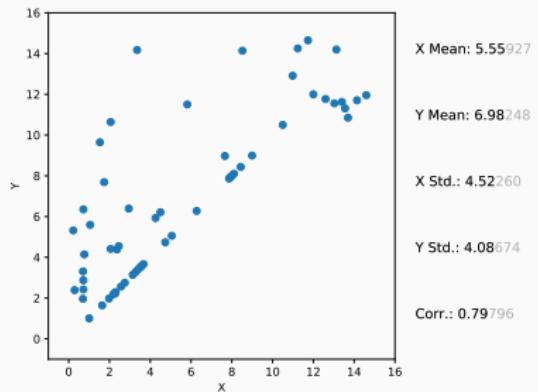


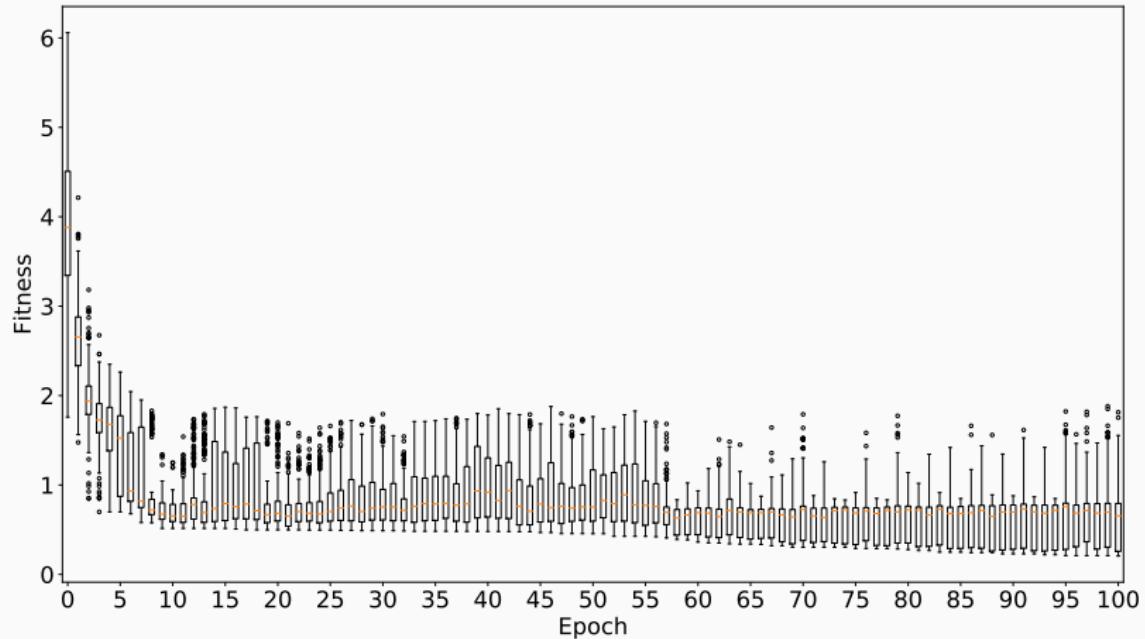
Input: A dataset, $X \in \mathbb{R}^n \times \mathbb{R}^n$; a set of **dissimilarity** measures,
 $f_1, \dots, f_k : \mathbb{R}^n \times \mathbb{R}^n \rightarrow \mathbb{R}$

Output: The total of the measures: $\sum_{i=1}^k f_i(X)$

```
sum ← 0
for  $i = 1, \dots, k$  do
    evaluate  $f_i(X)$ 
    add this to sum
end
return sum
```

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







```
$ pip install edo
```

Henry Wilde

Twitter: @daffidwilde

Email: wildehd@cardiff.ac.uk

Repository: <https://github.com/daffidwilde/edo>

Documentation: <https://edo.readthedocs.io>

Paper in preparation:

“Evolutionary Dataset Optimisation: understanding algorithm quality through evolution”