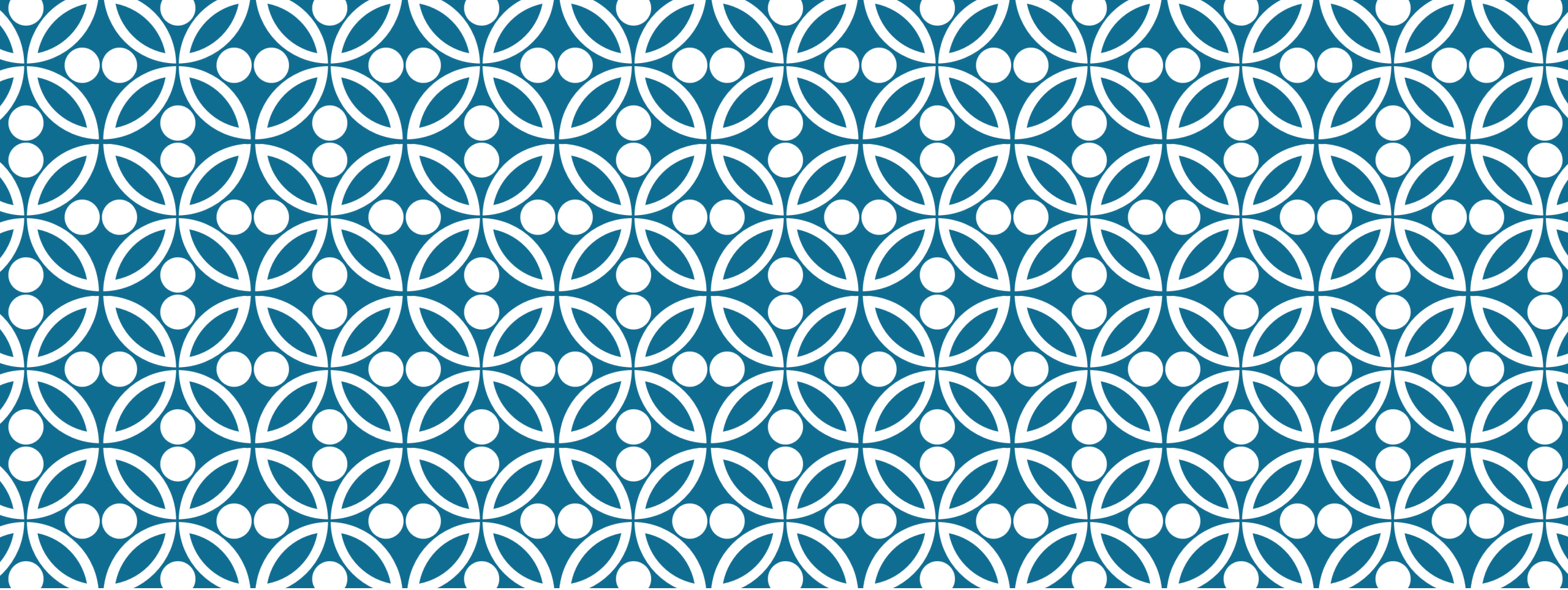
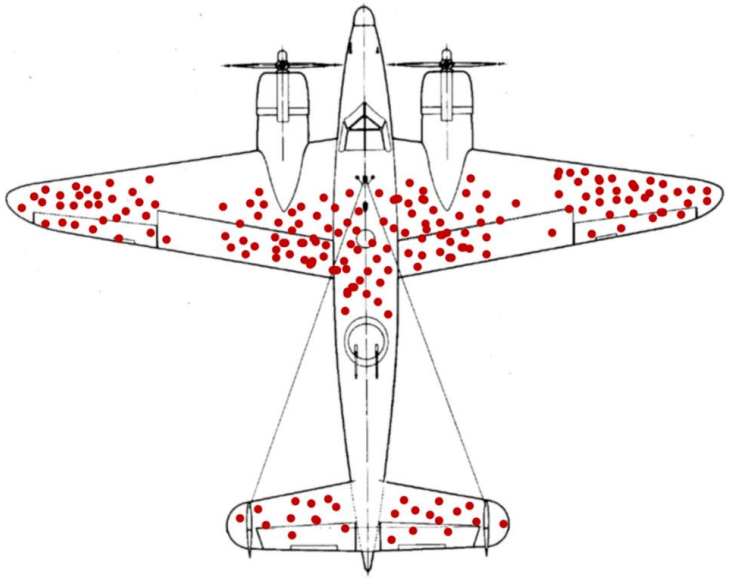


應用機器學習

Brian Chan 陳醒凡

SURVIVORSHIP BIAS 倖存者偏差

Survivorship bias or survival bias is the logical error of concentrating on the people or things that made it past some selection process and overlooking those that did not, typically because of their lack of visibility.



This can lead to false conclusions in several different ways. It is a form of selection bias.

https://en.wikipedia.org/wiki/Survivorship_bias

SURVIVORSHIP BIAS 倖存者偏差

- Steve Jobs, Bill Gates, and Mark Zuckerberg dropped out of college and became millionaires, so will I.

- I'll calculate annual recurring revenue (*ARR) based on our current customers.*

SAMPLING BIAS

Money laundering

Data snooping – p-value

MOTTO

All models are wrong, but some are useful.

George E. P. Box

TRAIN-TEST DIAGNOSIS

	Low *Training* Error	High *Training* Error
Low *Testing* Error	The model is learning!	Probably some error in your code. Or you've created a *psychic* AI.
High *Testing* Error	OVERFITTING	The model is not learning.

PERFORMANCE COMPARISON

Table 2: The Average RMSEs in Cross-sectional and Longitudinal Test Samples

Overall Performance			Overall Performance (Cross-sectional)		Overall Performance (Longitudinal)	
Methods	Average RMSE	Rank	Average RMSE	Rank	Average RMSE	Rank
Linear Regression	3.433	9	5.177	9	1.688	8
Ridge Regression	1.818	7	3.066	7	0.570	5
Lasso Regression	0.775	6	0.982	6	0.567	4
Support Vector Regression	0.701	5	0.689	5	0.713	7
Neural Network	3.167	8	3.284	8	3.05	9
Regression Tree	0.554	4	0.537	4	0.571	6
Random Forest	0.454	3	0.458	3	0.450	3
Bagging	0.413	2	0.434	2	0.391	1
Gradient Boosting	0.397	1	0.401	1	0.393	2

POOLING TEST

https://www.nature.com/articles/d41586-020-02053-6

https://members.loria.fr/ADeleforge/the-maths-of-pool-testing-mixing-samples-to-speed-up-covid-19-detection/