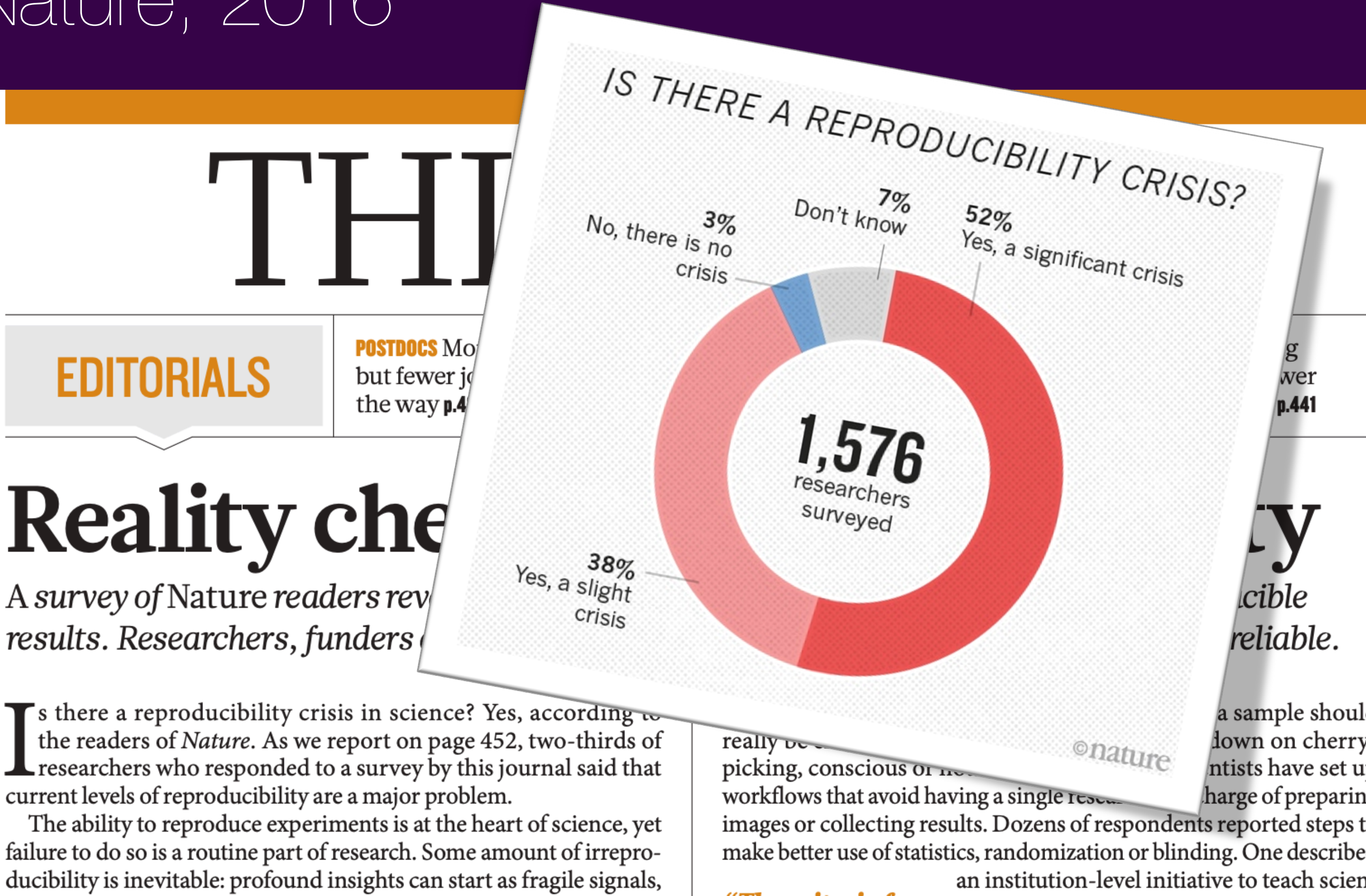
On the Stability and **Reproducibility** of Data Science Pipelines

Professor Gavin Brown
University of Manchester, UK

Nature, 2016



THI

EDITORIALS

POSTDOCS Mo
but fewer jo
the way p.4

g
wer
p.441

Reality che... ty

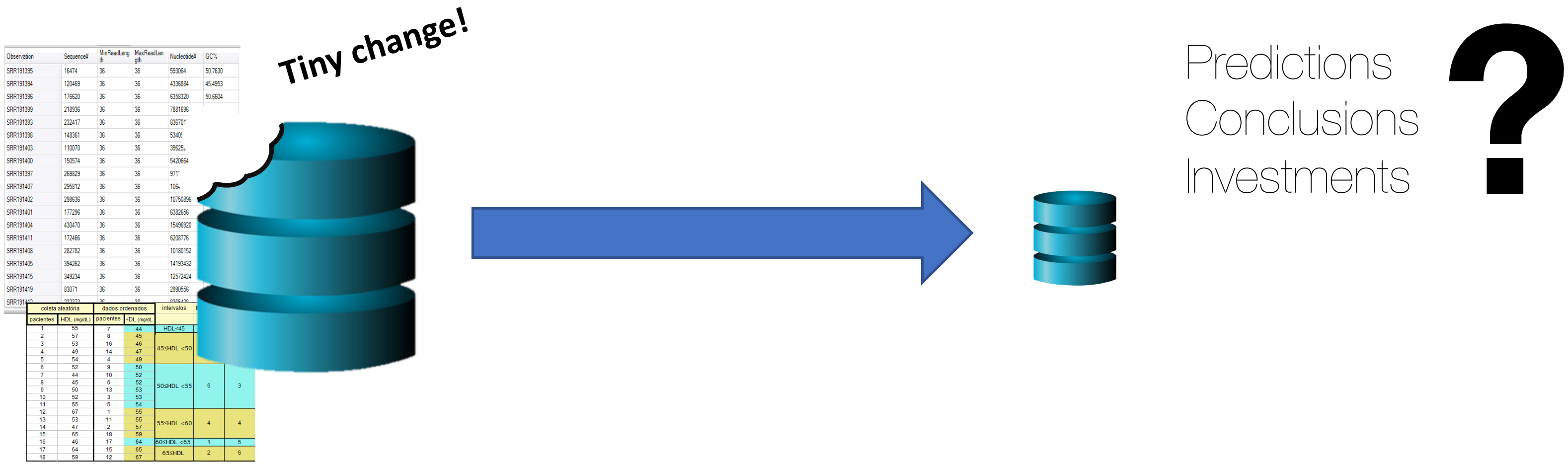
A survey of Nature *readers rev... cible results. Researchers, funders ... reliable.*

Is there a reproducibility crisis in science? Yes, according to the readers of *Nature*. As we report on page 452, two-thirds of researchers who responded to a survey by this journal said that current levels of reproducibility are a major problem.

The ability to reproduce experiments is at the heart of science, yet failure to do so is a routine part of research. Some amount of irreproducibility is inevitable: profound insights can start as fragile signals,

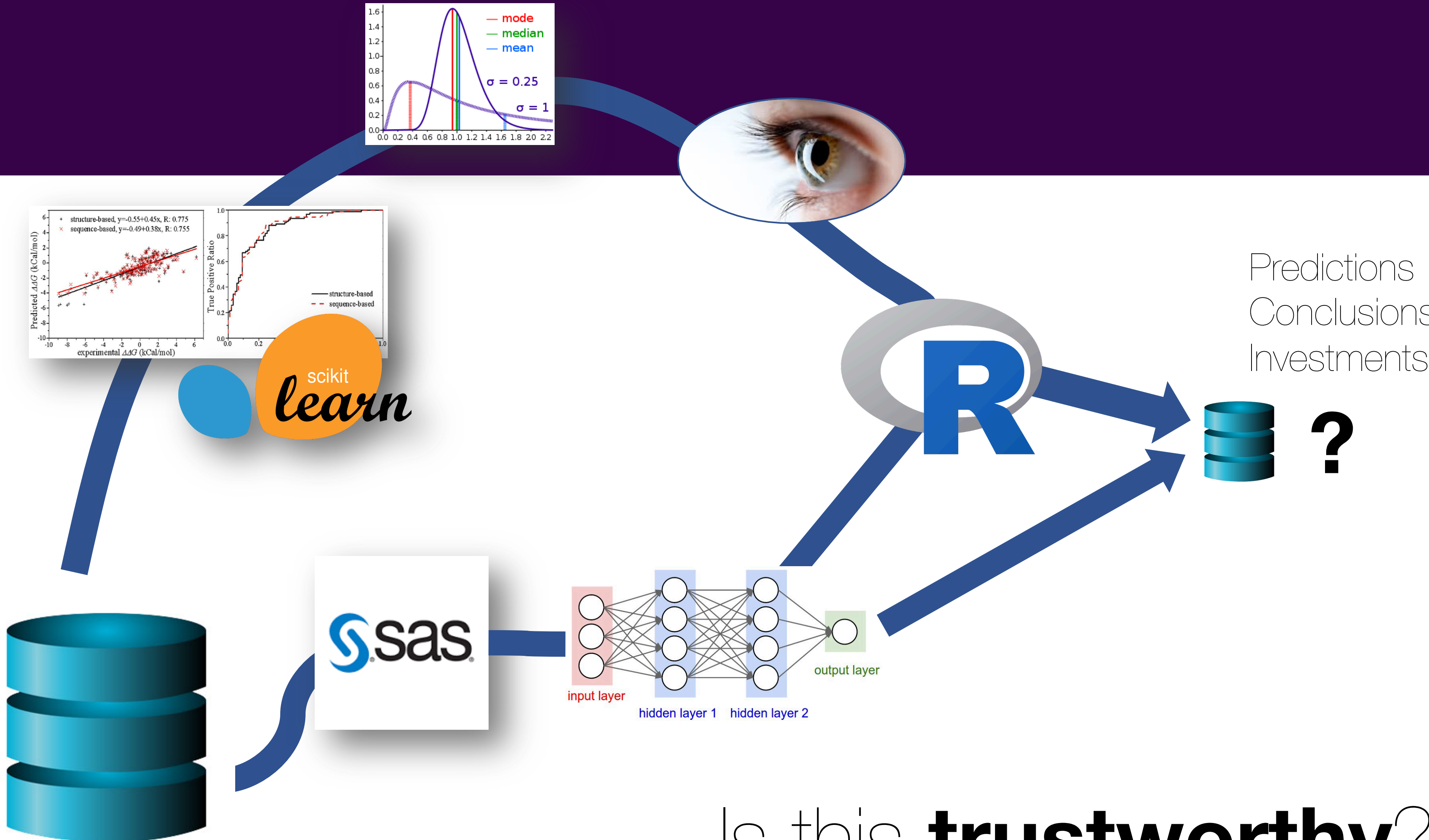
... a sample should really be ... down on cherry-picking, conscious or ... ntists have set up workflows that avoid having a single resea... harge of preparing images or collecting results. Dozens of respondents reported steps to make better use of statistics, randomization or blinding. One described an institution-level initiative to teach scien-

This is **your** data science pipeline.



“How much do you **trust** your data pipeline?”

“How **reproducible** is your result?”



Is this **trustworthy**?

Reproducibility = ***Trust***

Reproducibility is **not a "yes/no"** question.

Conjecture: ***We can measure reproducibility.***

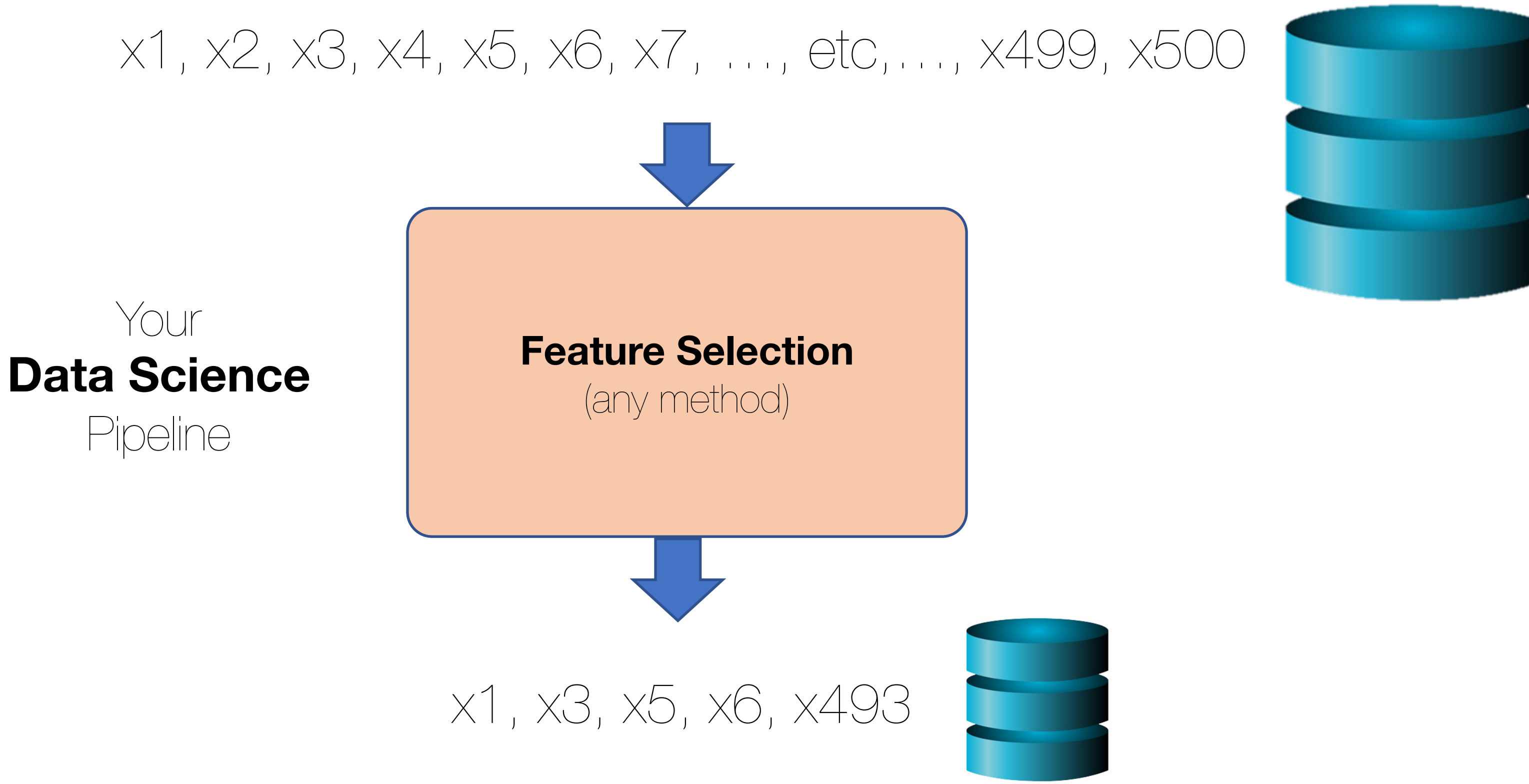
Let's take something specific.

Data-driven biomarker selection

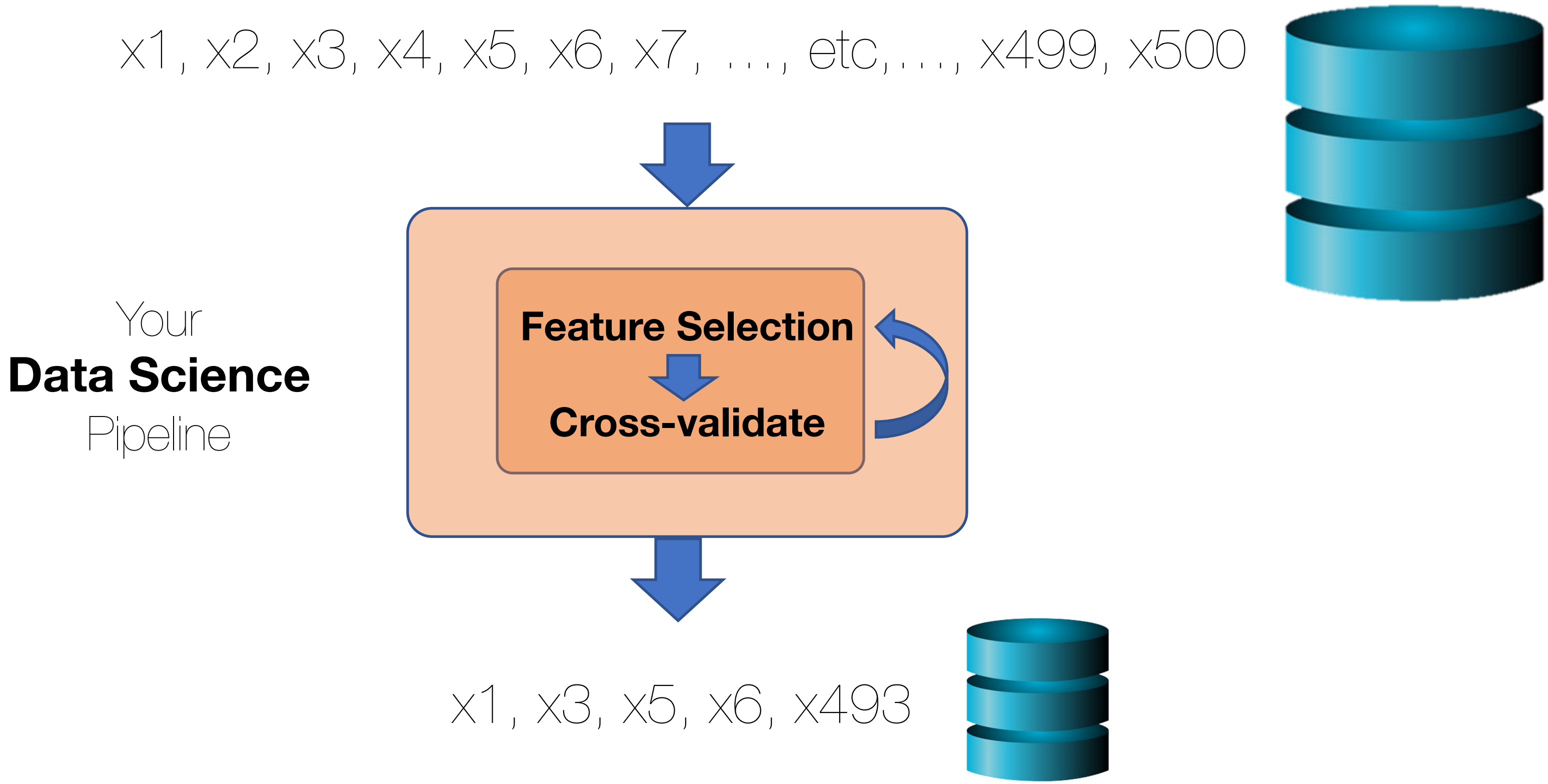
Only a <u>subset</u> of features actually influence the phenotype.

AstraZeneca

How much do you **trust your choices**?



How much do you **trust your choices**?



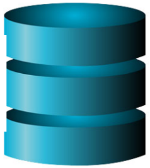
How much do you **trust your choices**?

x1 , x2, x3, x4, x5, x6, x7, . . . , etc, . . . , x499, x500

Drop a **random 1%** of examples

Will not make a difference
. . .or will it?

x2 , x3, x5, x6, **x491**



“**Stability**”

“**Stability**”

A specific instance of **reproducibility.**

For the task of data-driven **biomarker selection.**

But… how to measure it?

Estimating **Stability**

[**x1**, x2, **x3**, x4, **x5**, **x6** . . . , **x493**, . . . ,x499, x500]

[x1 , x3, x5, x6, x493]

In

[x2, x4, x7, . . . , x492, x494 . . . x500]

Out

Estimating **Stability**

Set intersection? (i.e. features in common)

$$\phi(s_i, s_j) = 3$$

My selected biomarkers.
When using **ALL** data.

[x1, x3, x5, x6, x493 **]**

[x2, x3, x5, x6, x491 **]**

Small change if I drop a random **1%** of data

Estimating **Stability…** which set measure?

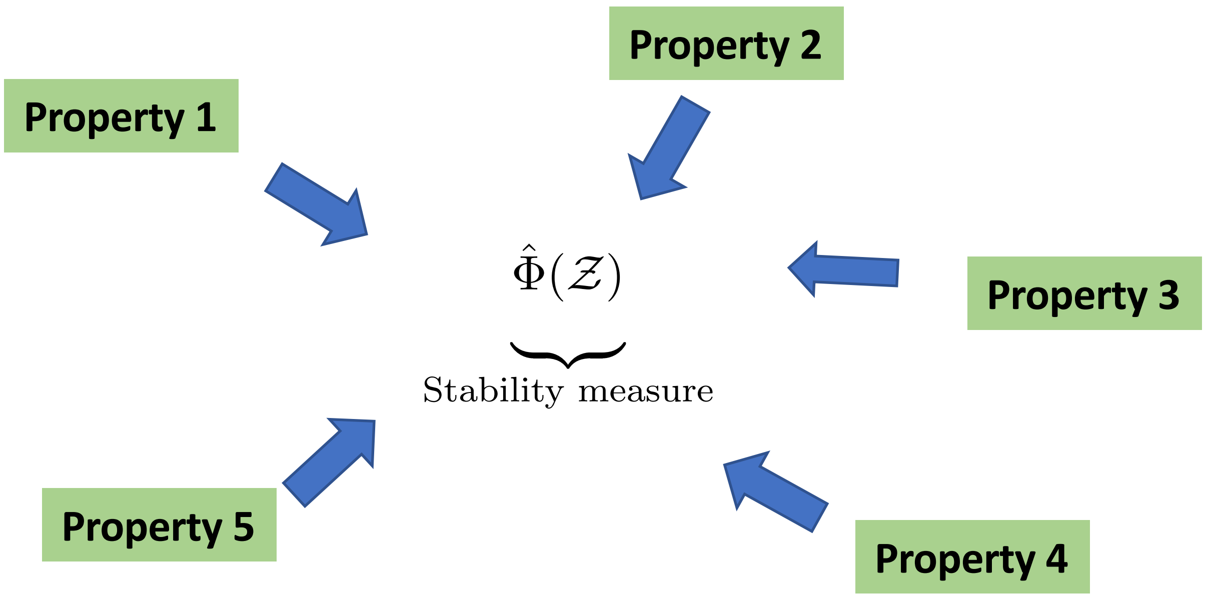
Dunne et al. (2002)	Hamming	$1 - \frac{\lvert s_i \setminus s_j \rvert + \lvert s_i \setminus s_j \rvert}{d}$
Kalousis et al. (2005)	Jaccard	$\frac{\lvert s_i \cap s_j \rvert}{\lvert s_i \cup s_j \rvert}$
Yu et al. (2008)	Dice-Sørenson	$\frac{2\lvert s_i \cap s_j \rvert}{\lvert s_i \rvert + \lvert s_j \rvert}$
Goh and Wong (2016)	Ochiai	$\frac{\lvert s_i \cap s_j \rvert}{\sqrt{\lvert s_i \rvert \lvert s_j \rvert}}$
Shi et al (2006)	POG	$\frac{\lvert s_i \cap s_j \rvert}{\lvert s_i \rvert}$
Kuncheva (2007)	Consistency	$\frac{r_{i,j} - \frac{k^2}{d}}{k - \frac{k^2}{d}}$
Lustgarten et al. (2009)	Lustgarten	$\frac{r_{i,j} - \frac{k_i k_j}{d}}{min(k_i, k_j) - max(0, k_i + k_j - d)}$
Wald et al. (2013)	Wald	$\frac{r_{i,j} - \frac{k_i k_j}{d}}{min(k_i, k_j) - \frac{k_i k_j}{d}}$

Set theoretic measures 1957-2017.
Many definitions.
(about 20)

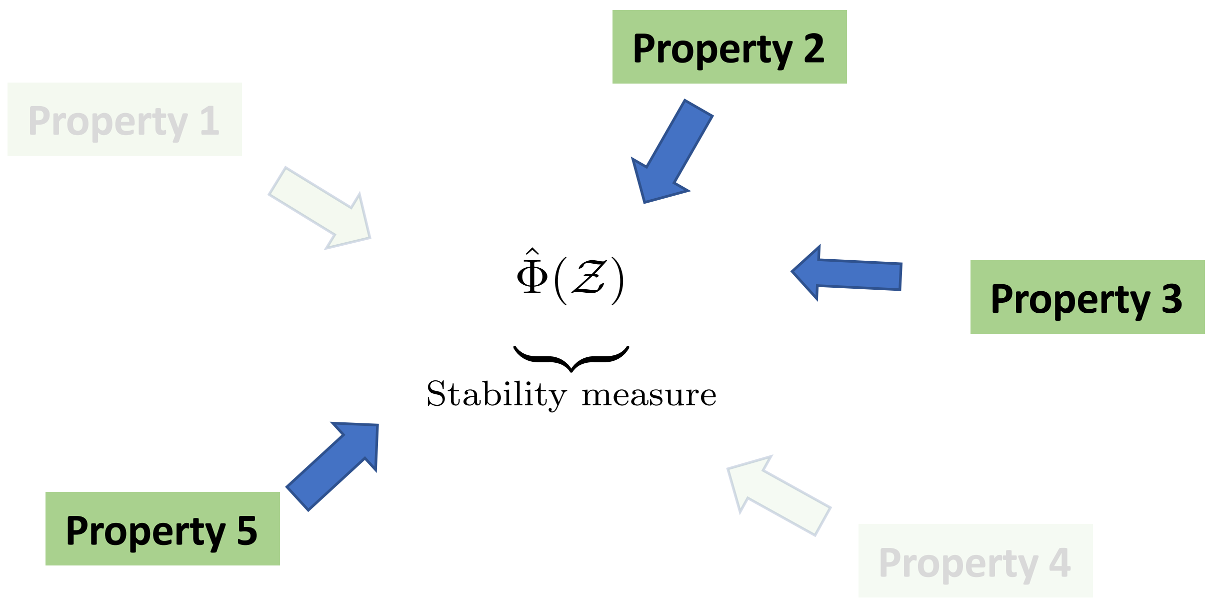
Mostly heuristic.
Conflicting opinions.

No principled way to choose between them.

What properties do we want?



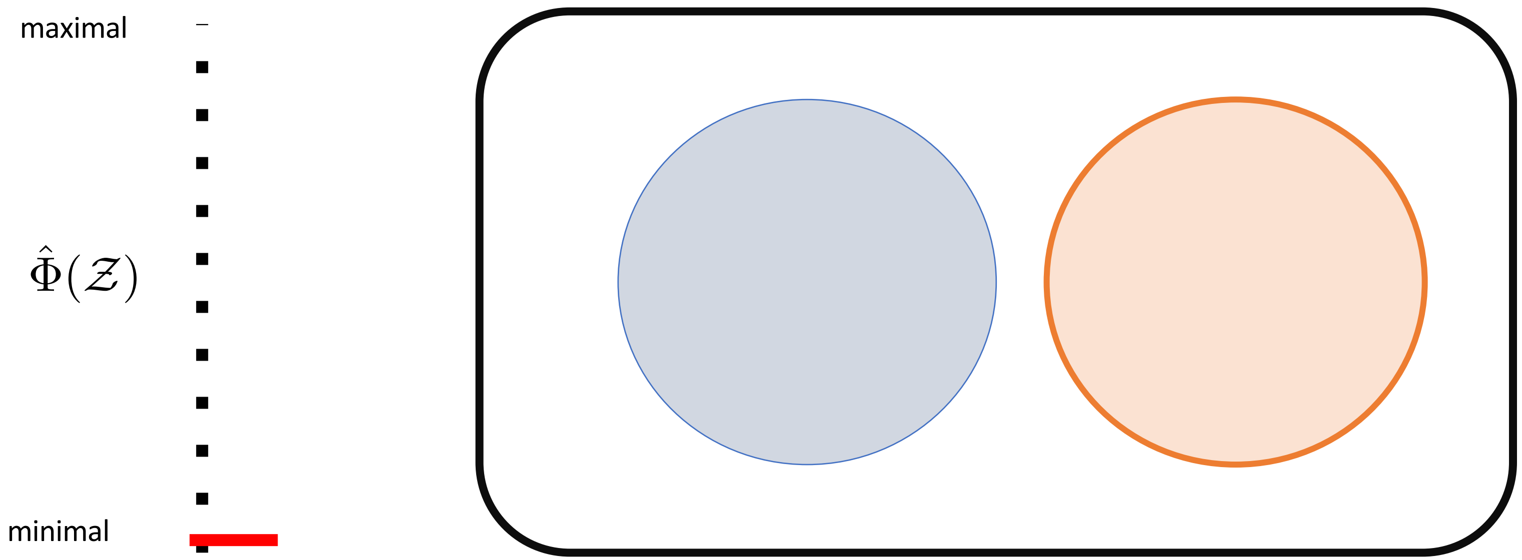
What properties do we want?



Desirable property 2: *Strict Monotonicity*

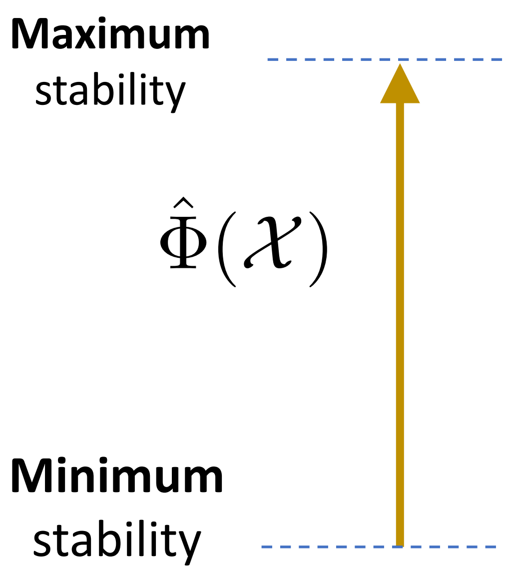
Property 2

. . .as the sets **overlap** more, the measure should **increase**.



Desirable property 3: *Known upper/lower Bounds*

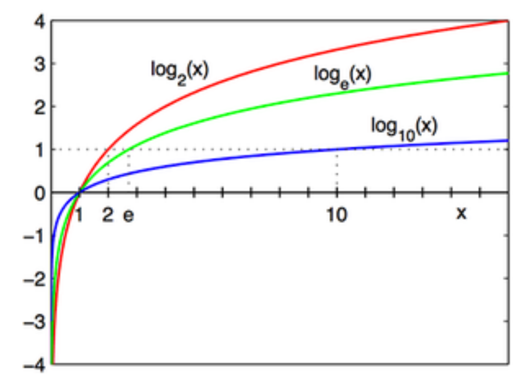
Property 3



For **interpretability** and **comparison** across problems/algorithms,

… it should have **known, finite upper/lower bounds,**

No logarithms!



Desirable property 5: *Correction for chance*

Property 5

Selecting 2 features from 200…

Trial 1	0 0 0 0 0 0 0 0 0 **1** 0 0 **1** 0 0 0 0 0 0 0…
Trial 2	0 0 0 0 0 0 0 0 0 0 0 0 0 **1** 0 0 0 0 0 **1**…

Is very different to selecting 2 from 5…

Trial 1	**1 1** 0 0 0
Trial 2	0 **1 1** 0 0

High chance of intersection, even if random!

Dunne et al. (2002)	Hamming	$1 - \frac{\lvert s_i \setminus s_j \rvert + \lvert s_i \setminus s_j \rvert}{d}$
Kalousis et al. (2005)	Jaccard	$\frac{\lvert s_i \cap s_j \rvert}{\lvert s_i \cup s_j \rvert}$
Yu et al. (2008)	Dice-Sørenson	$\frac{2\lvert s_i \cap s_j \rvert}{\lvert s_i \rvert + \lvert s_j \rvert}$
Goh and Wong (2016)	Ochiai	$\frac{\lvert s_i \cap s_j \rvert}{\sqrt{\lvert s_i \rvert \lvert s_j \rvert}}$
Shi et al (2006)	POG	$\frac{\lvert s_i \cap s_j \rvert}{\lvert s_i \rvert}$
Kuncheva (2007)	Consistency	$\frac{r_{i,j} - \frac{k^2}{d}}{k - \frac{k^2}{d}}$
Lustgarten et al. (2009)	Lustgarten	$\frac{r_{i,j} - \frac{k_i k_j}{d}}{min(k_i, k_j) - max(0, k_i + k_j - d)}$
Wald et al. (2013)	Wald	$\frac{r_{i,j} - \frac{k_i k_j}{d}}{min(k_i, k_j) - \frac{k_i k_j}{d}}$

Remember this?

Results…

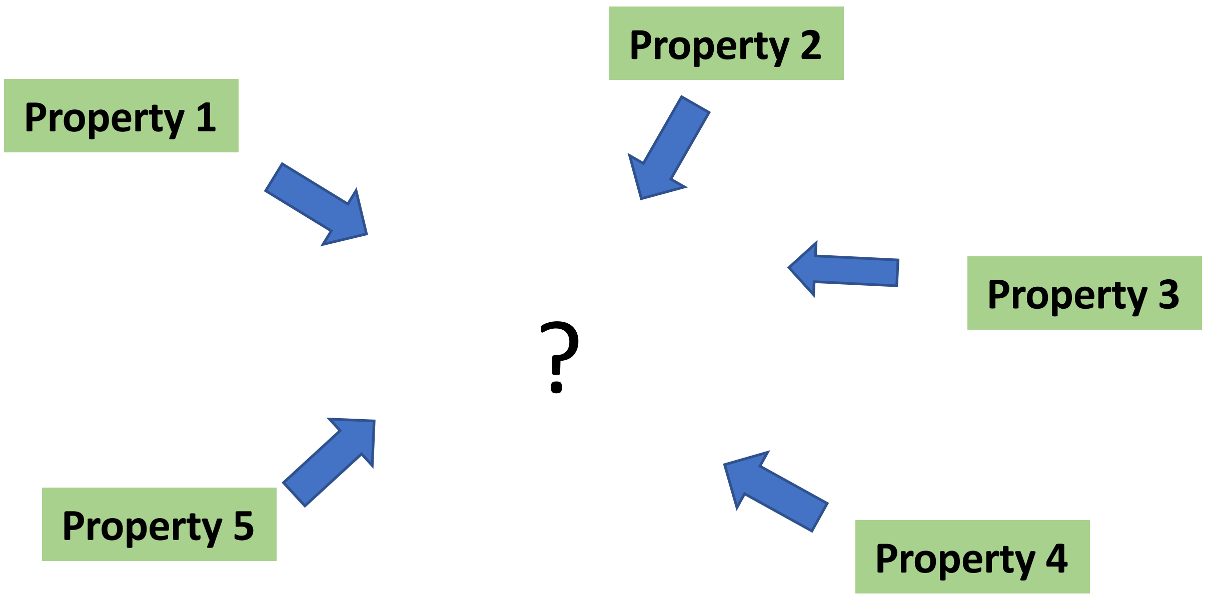
See paper for the 75 proofs!

Name	Fully defined (1)	Monotonicity (2)	Bounds (3)	Maximum (4)	Correction (5)
Hamming					
Jaccard					
Dice					
Ochiai					
POG					
Kuncheva					
Lustgarten					
Wald					
nPOG					
Goh					
Davis					
Krízek					
Guzmán					
CW_{rel}					
Lausser					

?

So where do these properties point?

Nogueira et al, "On The Stability of Feature Selection"
Journal of Machine Learning Research 2018



Definition 2 (Effective Stability for Pairwise Feature Redundancy). *Given a matrix* $\mathbb{C}$ *specifying feature relationships, the effective stability is*

$$\widehat{\Phi}_{\mathbb{C}}(\mathcal{Z}) = 1 - \frac{\sum_{f=1}^{d}\sum_{f'=1}^{d} c_{f,f'}\widehat{\mathrm{cov}}(Z_f, Z_{f'})}{\sum_{f=1}^{d}\sum_{f'=1}^{d} c_{f,f'}\mathrm{cov}(Z_f, Z_{f'}|H_0)} = 1 - \frac{\mathrm{tr}(\mathbb{C}\mathbf{S})}{\mathrm{tr}(\mathbb{C}\mathbf{\Sigma^0})}, \quad (5)$$

where $\mathbf{S}$ *is an unbiased estimator of the variance-covariance matrix of* $\mathcal{Z}$*, i.e.* $\mathbf{S}_{f,f'} = \widehat{\mathrm{Cov}}(Z_f, Z_{f'}) = \frac{M}{M-1}(\hat{p}_{f,f'} - \hat{p}_f\hat{p}_{f'}),\ \forall\ f, f' \in \{1...d\}$*, while* $\mathbf{\Sigma^0}$ *is the covariance matrix of* $\mathcal{Z}$ *under the null model of feature selection*

1. All 5 **desirable properties**, as discussed.

2. Clean **statistical** interpretation

   ….Confidence intervals and hypothesis tests come for free

3. Computable in **closed form**, as opposed to quadratic

Journal off Machine Machine Learning Research, 2018

Journal of Machine Learning Research 18 (2018) 1-54 Submitted 9/17; Revised 2/18; Published 4/18

On the Stability of Feature Selection Algorithms

Sarah Nogueira SARAH.NOGUEIRA@MANCHESTER.AC.UK
Konstantinos Sechidis KONSTANTINOS.SECHIDIS@MANCHESTER.AC.UK
Gavin Brown GAVIN.BROWN@MANCHESTER.AC.UK
School of Computer Science
University of Manchester
Manchester M13 9PL, UK

Editor: Isabelle Guyon

Abstract

Feature Selection is central to modern data science, from exploratory data analysis to predictive model-building. The "stability" of a feature selection algorithm refers to the *robustness* of its feature preferences, with respect to data sampling and to its stochastic nature. An algorithm is 'unstable' if a *small* change in data leads to *large* changes in the chosen feature subset. Whilst the idea is simple, *quantifying* this has proven more challenging—we note numerous proposals in the literature, each with different motivation

Case Study: Non-Small Cell Lung Cancer

Efficacy of **gefitinib** vs **chemotherapy** for lung cancer.

2 competing biomarker sets. Which do we trust?

Rank	GBM	CMIM
1	EGFR expression (X_4)	EGFR mutation (X_2)
2	Disease stage (X_{10})	Serum ALP(X_{13})
3	WHO perform. status (X_1)	Blood leukocytes (X_{21})
4	Serum ALT(X_{12})	Serum ALT (X_{12})

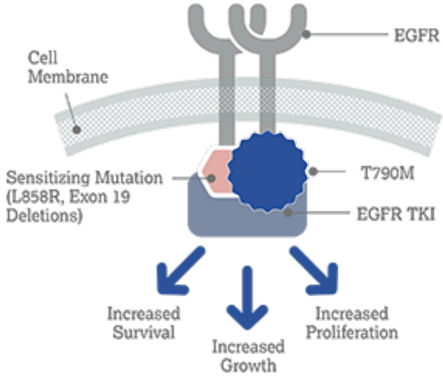
AstraZeneca

Case Study: Non-Small Cell Lung Cancer

	GBM		CMIM
Stability $\widehat{\Phi}(\mathcal{Z})$	**0.87**	>	0.68
- within **Group A**	0.96		0.45
- within **Group B**	0.82		0.80
- within **Group C**	0.14		0.43
Effective stability $\widehat{\Phi}_{\mathbb{C}}(\mathcal{Z})$	0.87	<	**0.91**

All **EGFR** gene **mutations**

(known to play a role in NSCLC)



Measure within-group stability

to see what's happening…

Changes our view

of the "best" algorithm to invest in.

AstraZeneca

The Take-Home Message

Reproducibility is not a yes /no question.

Reproducibility = ***<u>Trust</u>***

The industry needs methods to **<u>quantify reproducibility.</u>**

AMLD EPFL

MANCHESTER 1824
The University of Manchester

On the Stability and **Reproducibility** of Data Science Pipelines

Professor Gavin Brown
University of Manchester, UK