How to Interpret (or not) a Machine Learning model

Tiana_Athriel
dhuppenkothen

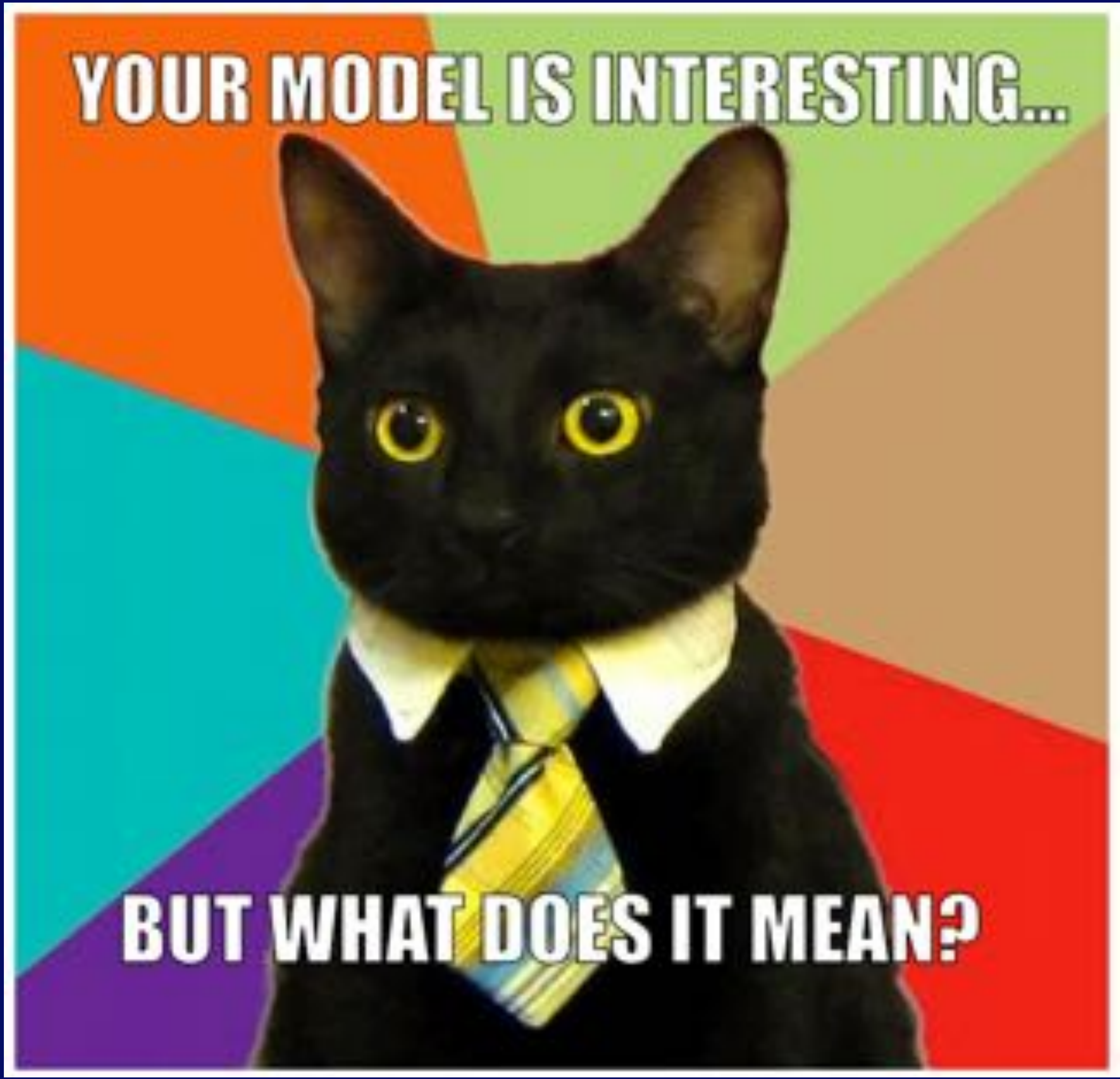
Daniela Huppenkothen
NYU Center for Cosmology and Particle Physics
NYU Center for Data Science

2 topics

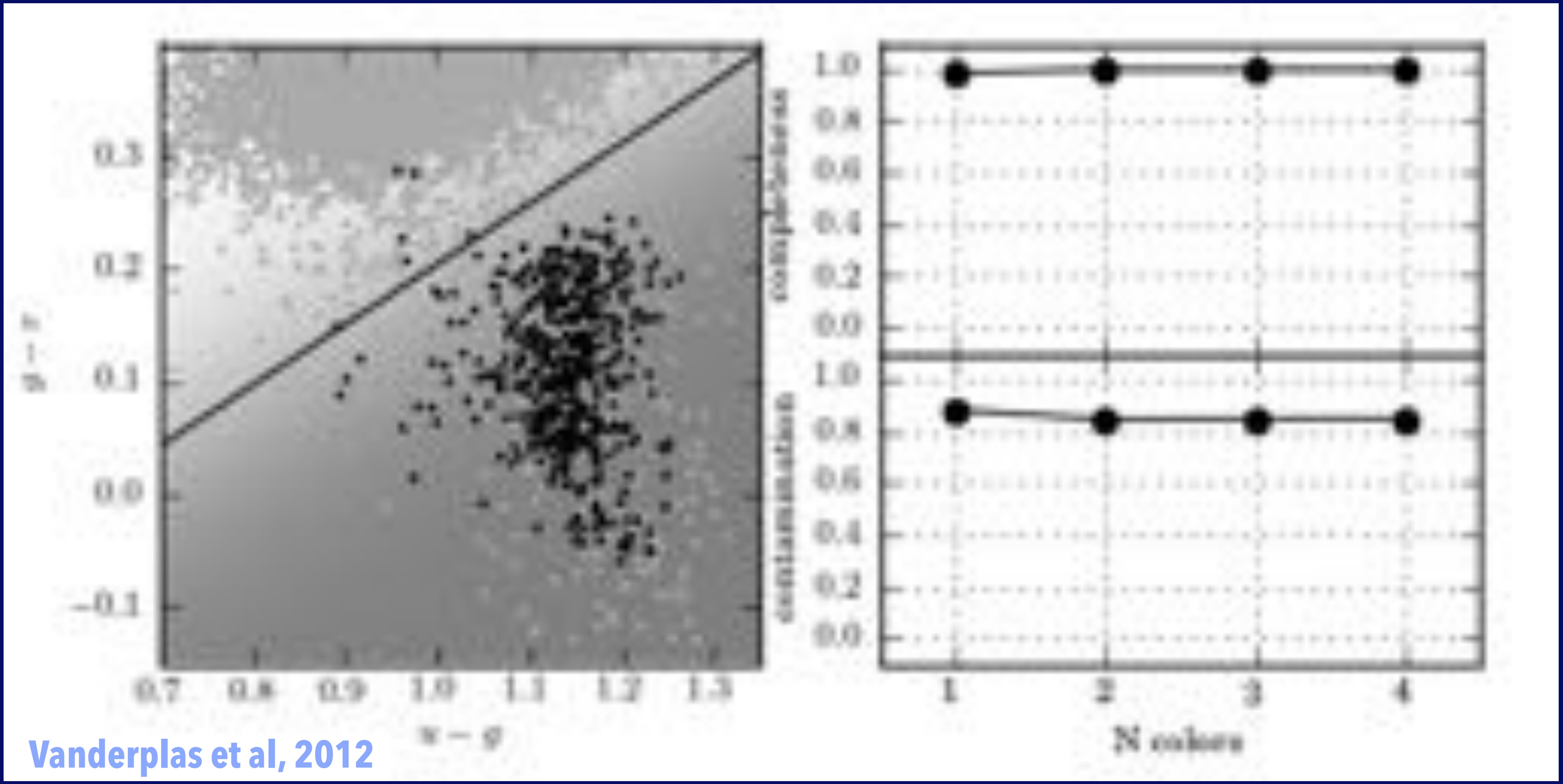
Interpretability

Model Selection

Interpretability


YOUR MODEL IS INTERESTING...
BUT WHAT DOES IT MEAN?

Logistic Regression of RR Lyrae Stars



Vanderplas et al, 2012

... now what?

What does interpretability mean to you?

2 main goals:
inference versus prediction

2 main goals:

inference versus prediction

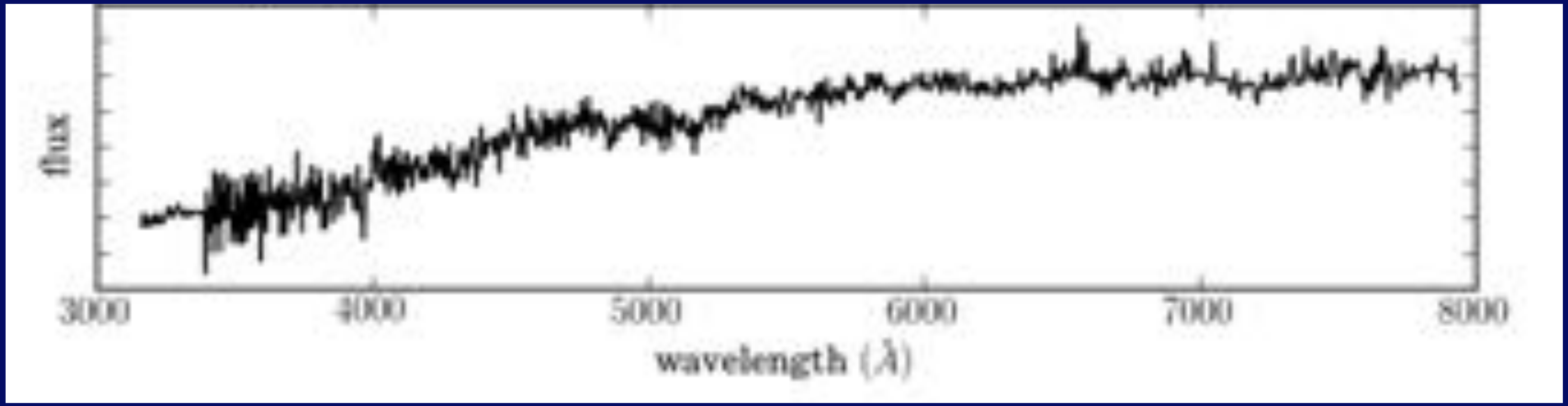
statistics! — **machine learning**

Inference

“a conclusion reached on the basis of evidence and reasoning”

Inference

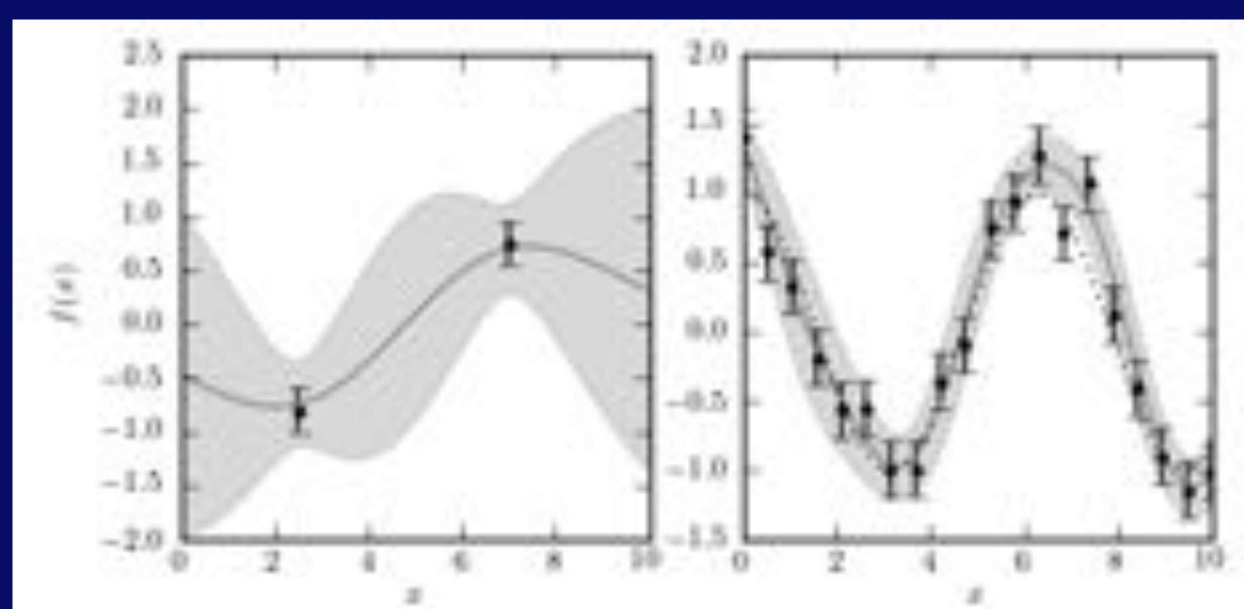
“Why do stars have different colours from galaxies?”



Vanderplas et al, 2012

Prediction

"Given my data points X and outcomes y, what outcome will I predict for a new data point x?"



Vanderplas et al, 2012

Z. Lipton: The Mythos of Model Interpretability

https://arxiv.org/abs/1606.03490

Scientific goal: uncover causal relationship

≠

ML goal: minimize prediction error

Motives

Trust

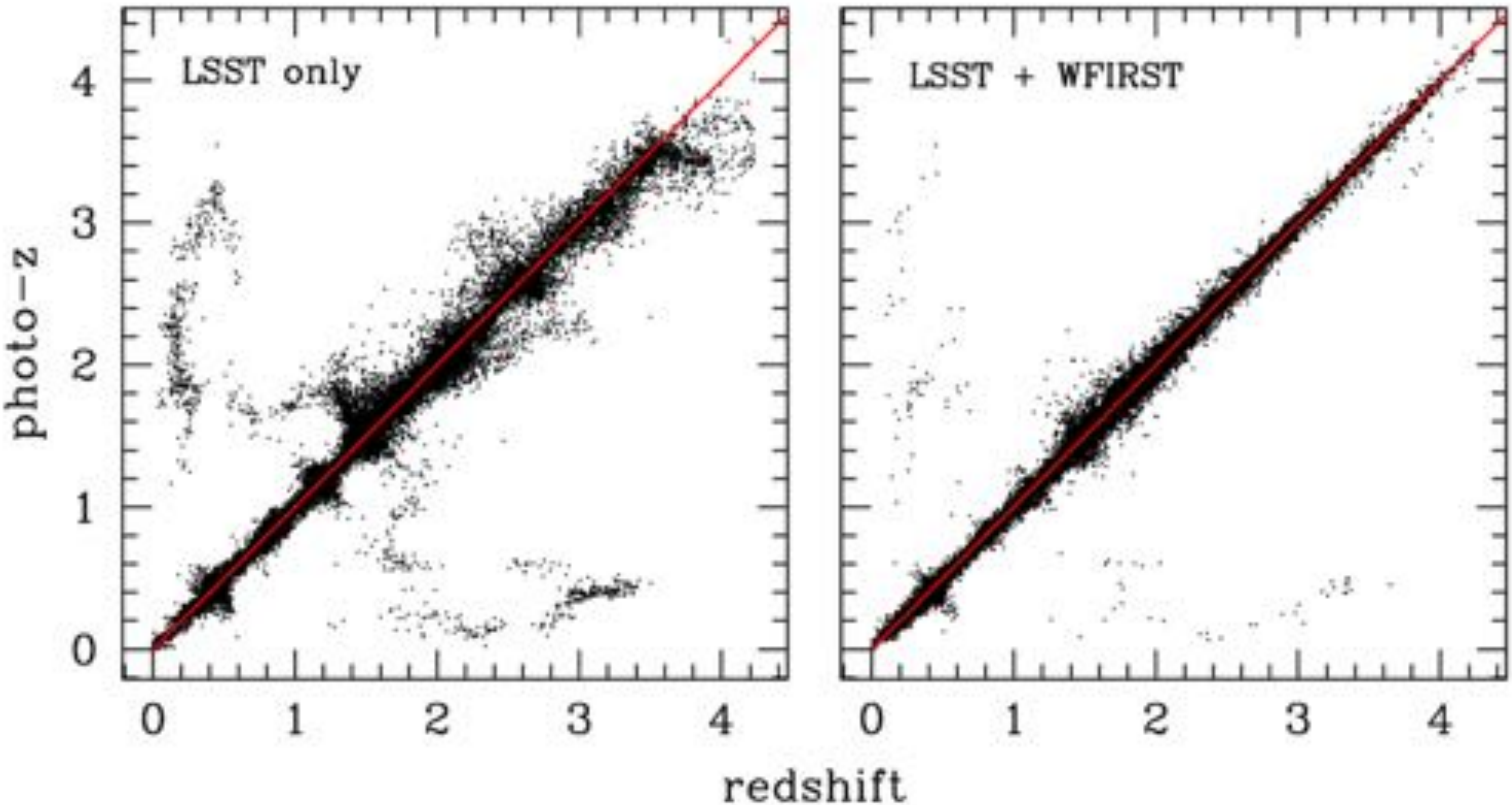
Understandability? Of features? Parameters? Models? Algorithms?

Low test error?

Does training data match deployment environment?

Jain et al, 2015

Trust



Jain et al, 2015

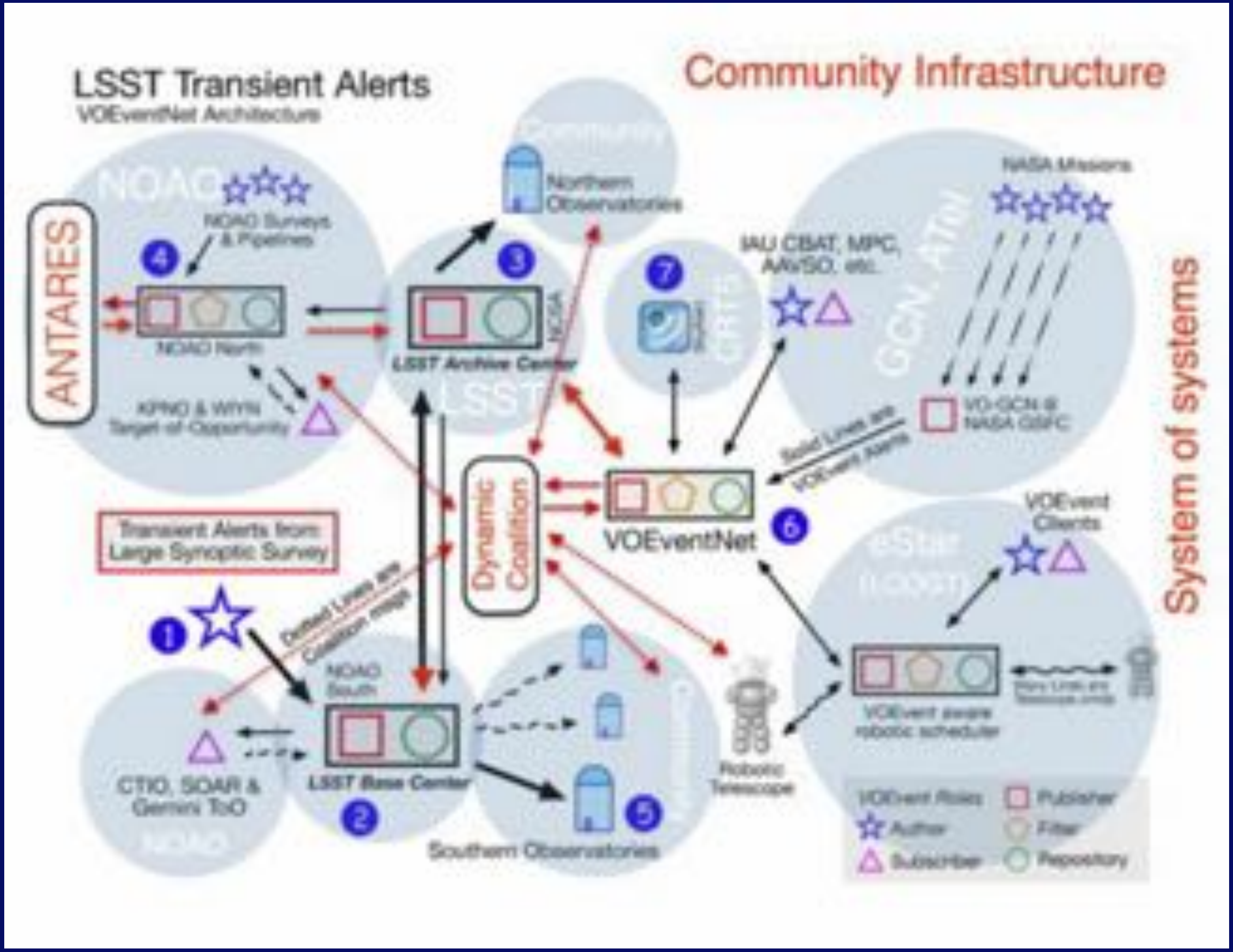
Causality

Transferability

http://emedicine.medscape.com/article/360090-overview

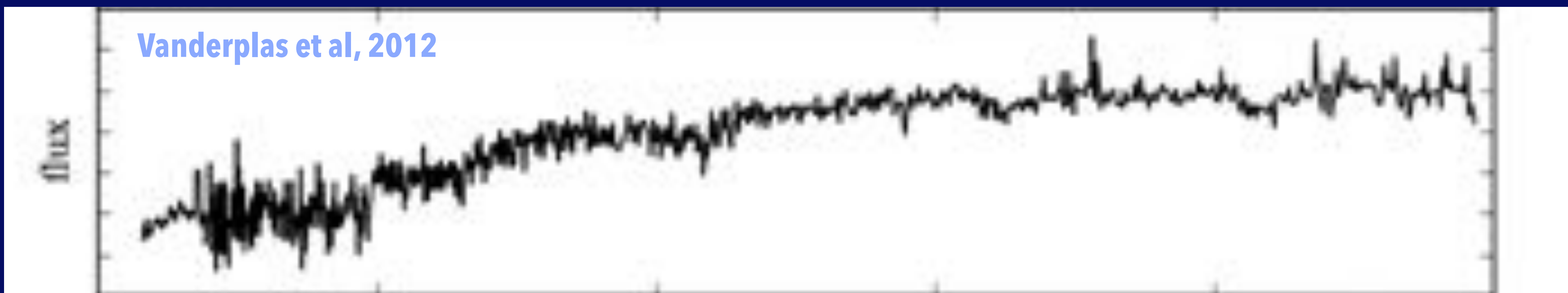
See also: Caruana et al, 2015

Transferability

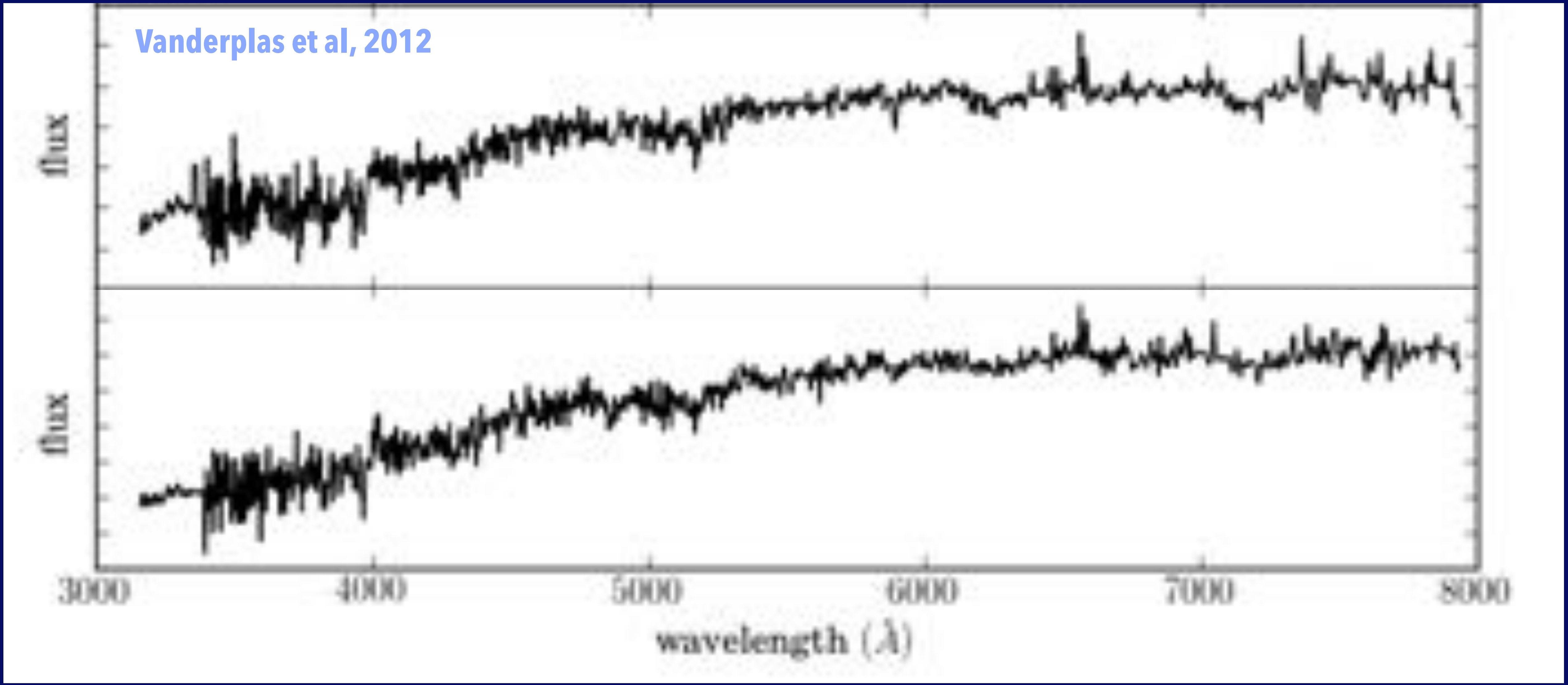


Seaman et al, 2014

Informativeness

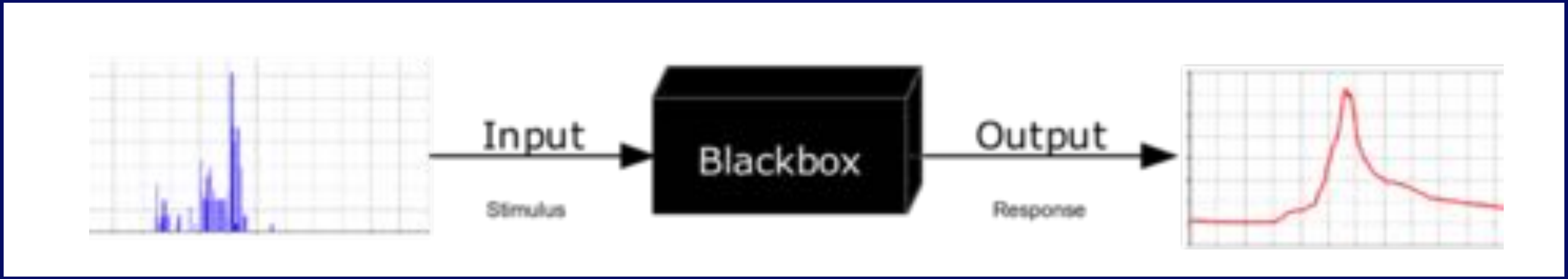


Informativeness

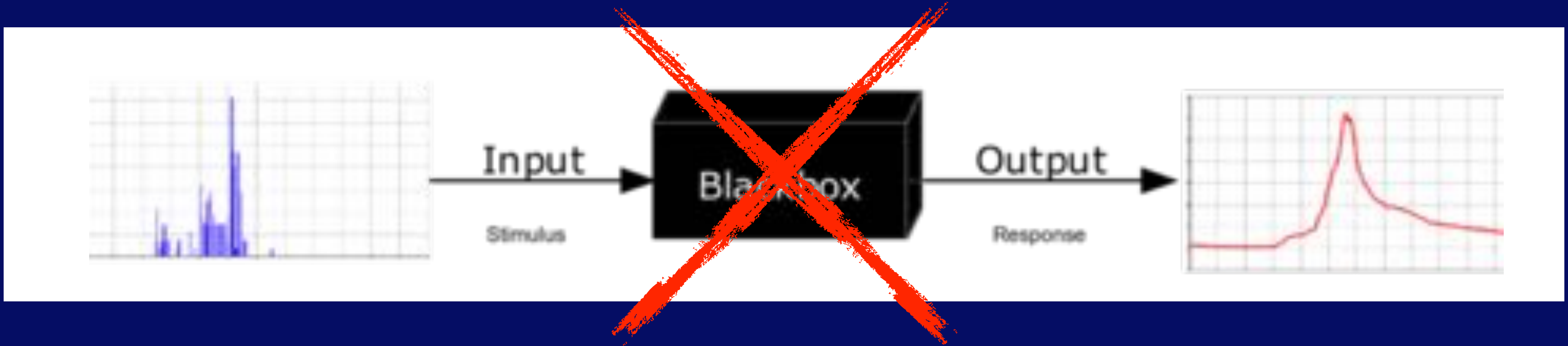


Properties of an interpretable model

1) Transparency



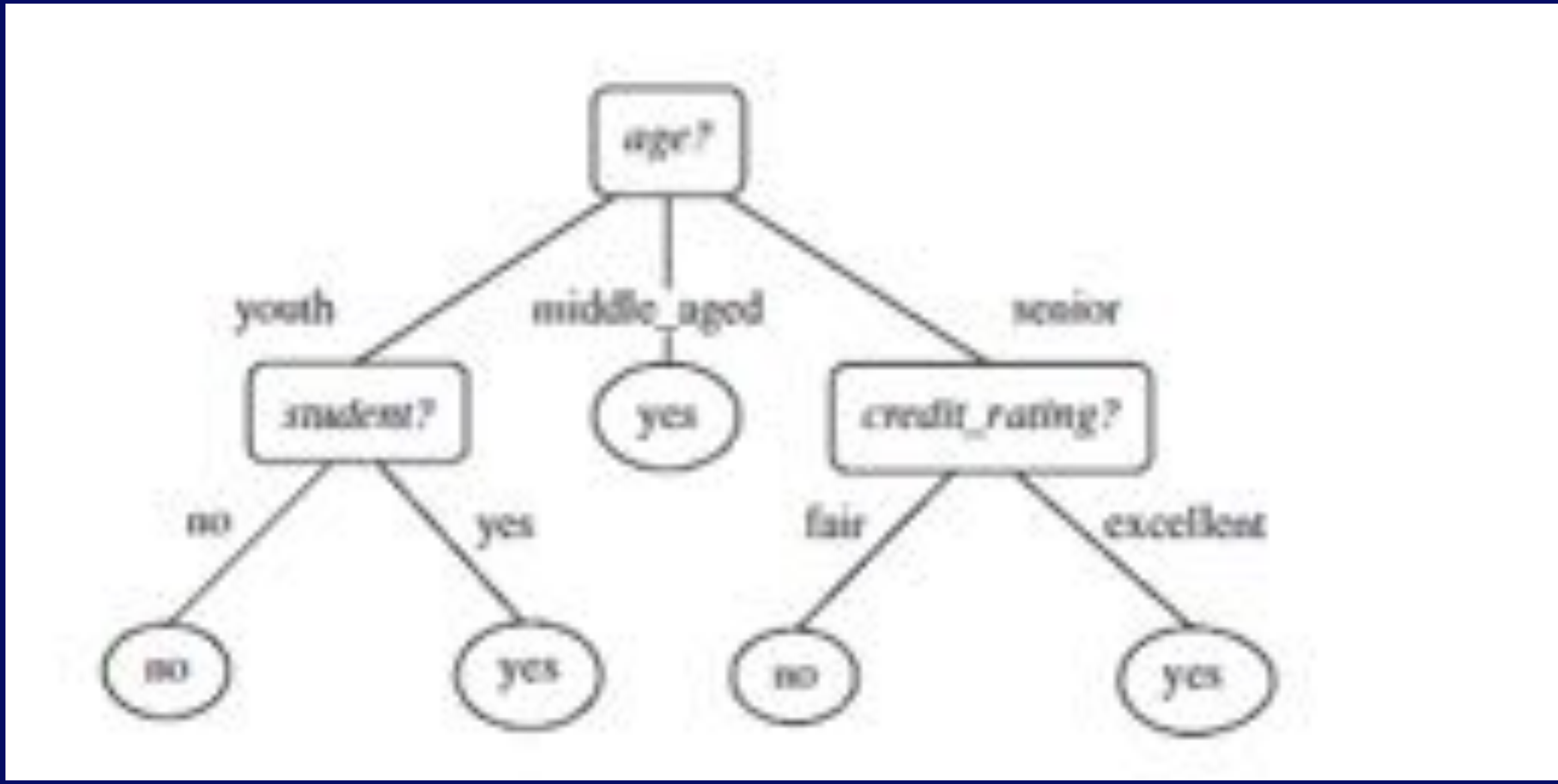
1) Transparency



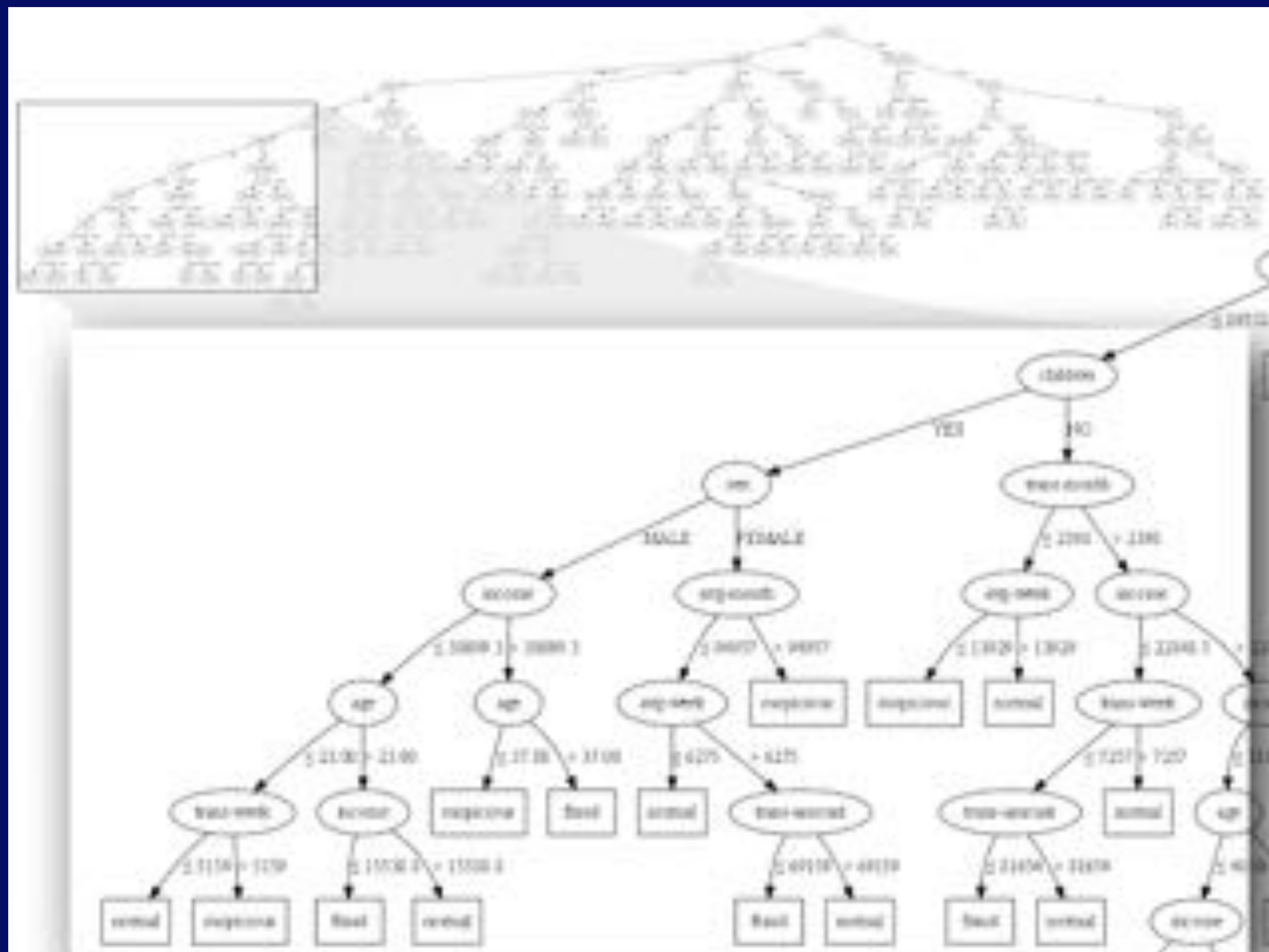
Simulatability

= ability to understand model in your head

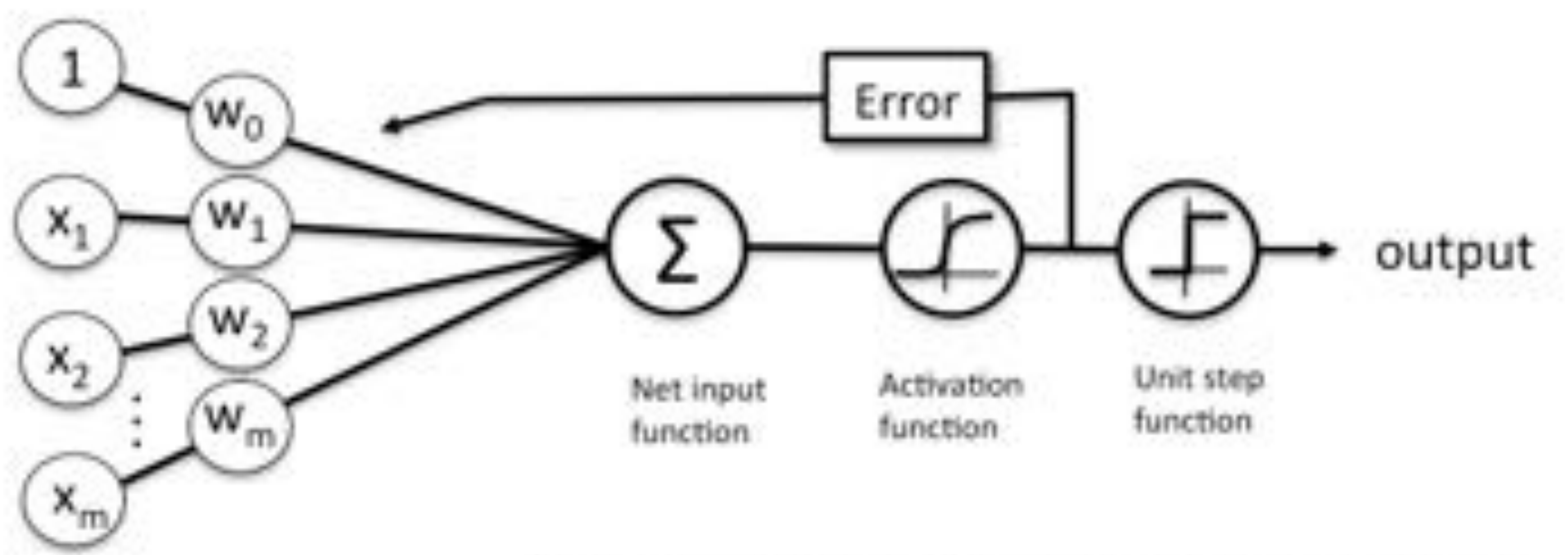
Example: Decision Trees



Example: Decision Trees



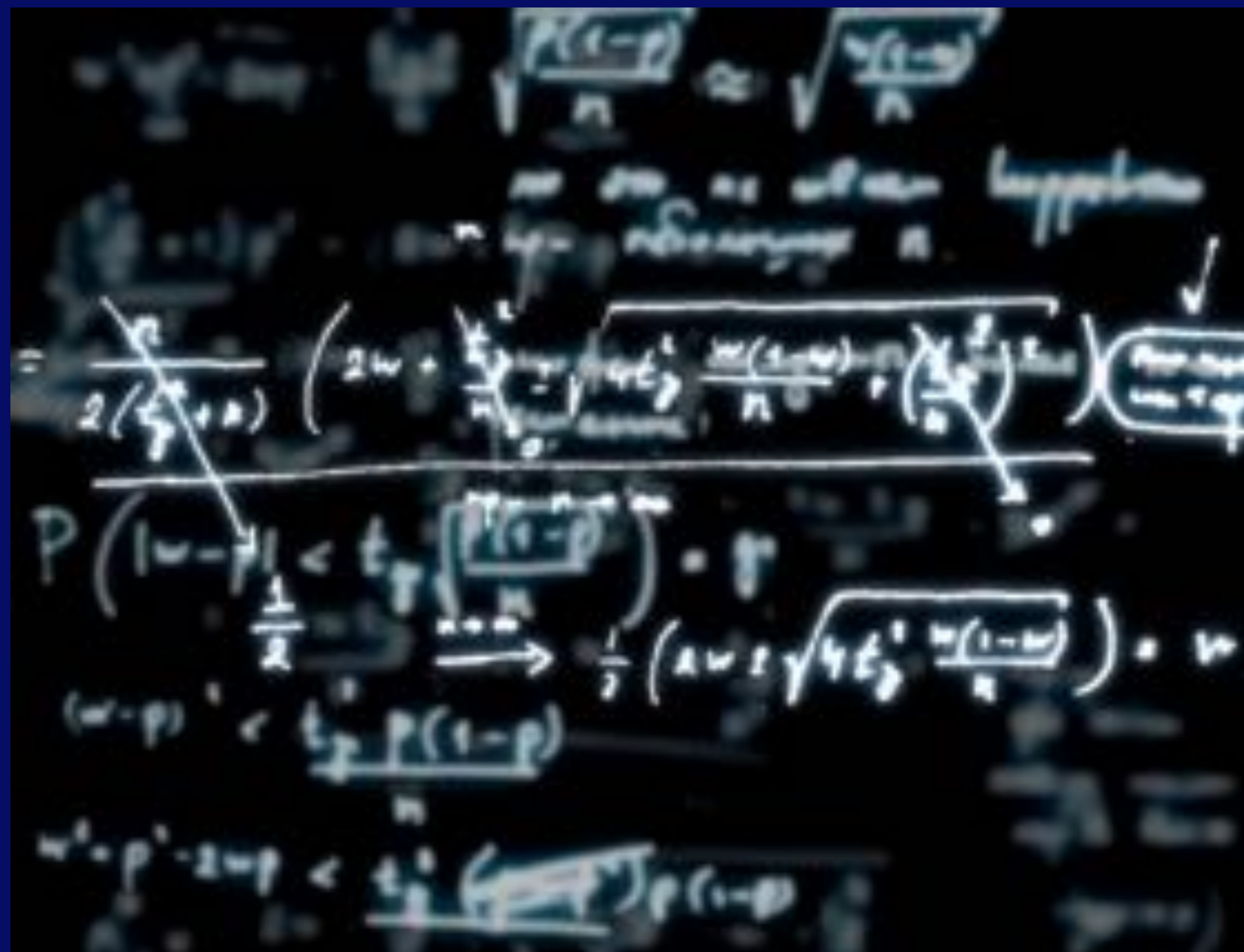
Decomposability



Schematic of a logistic regression classifier.

= ability to understand model components

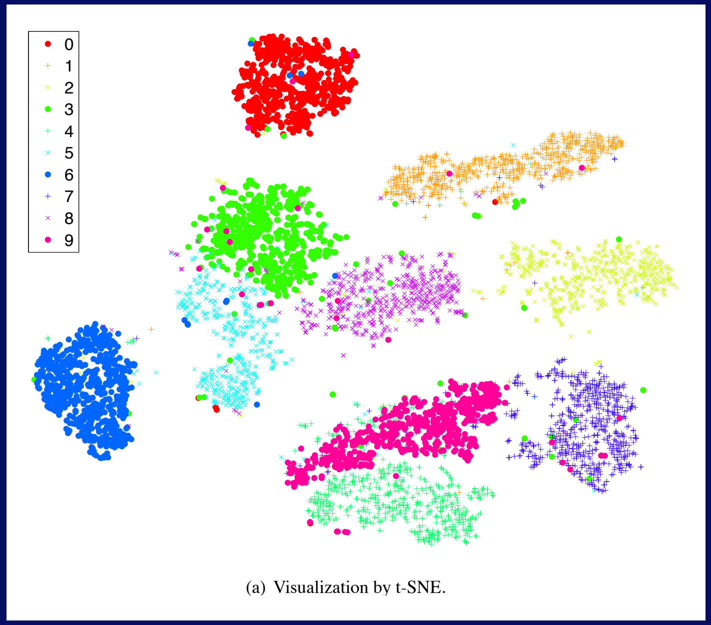
Algorithmic Transparency*



***note: humans have no algorithmic transparency whatsoever!**

2) Post-Hoc Interpretability

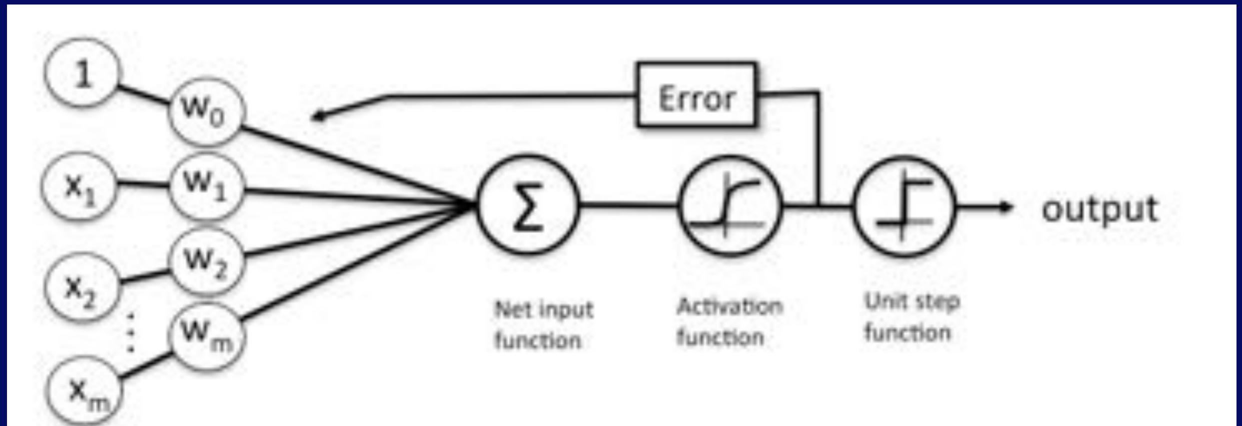
natural language explanations



visualization

learning by example

Black Box Benchmarking



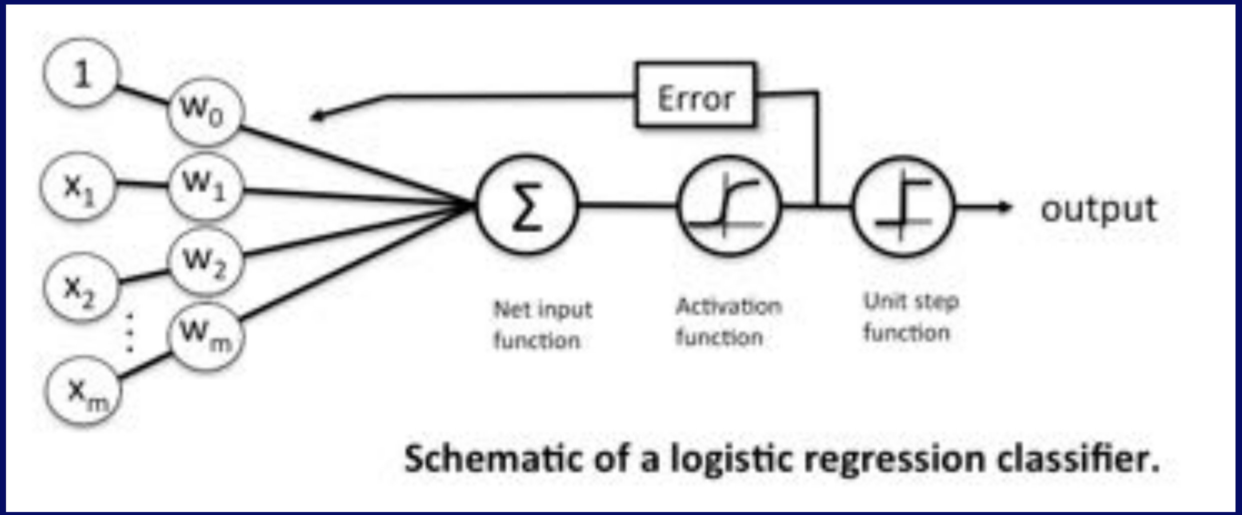
Schematic of a logistic regression classifier.

1) train linear model

2) train blackbox model

compare!

Surrogate Models

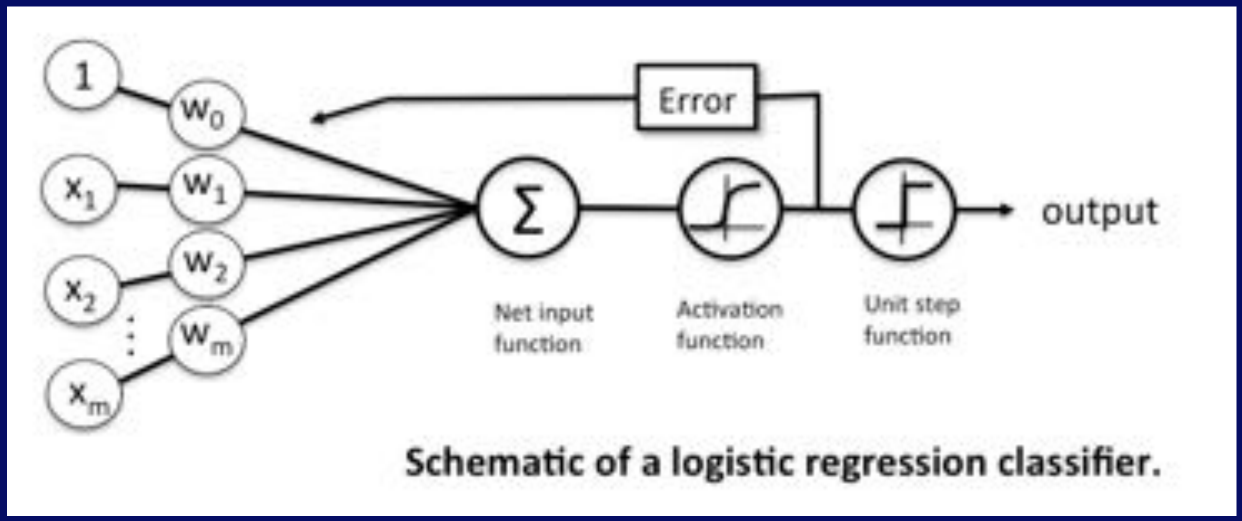


1) train blackbox model

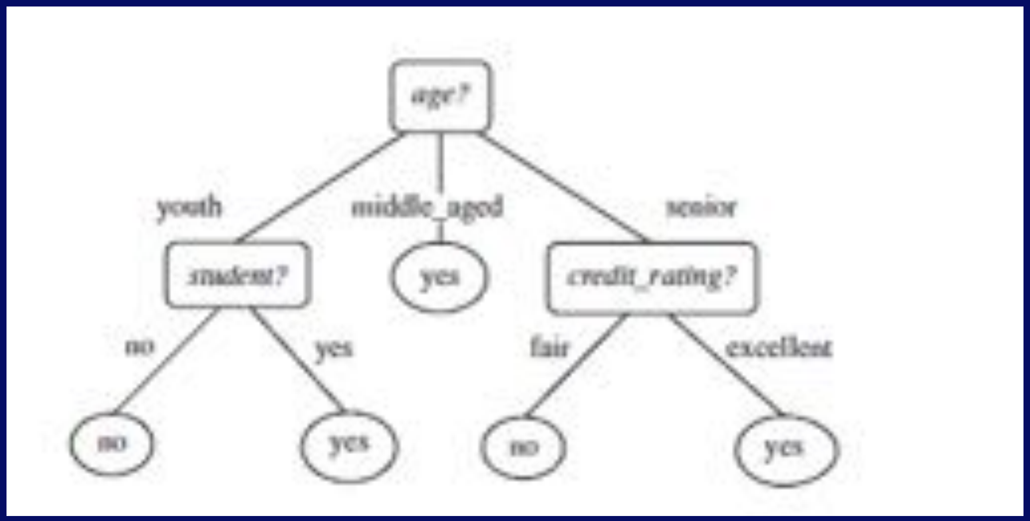
2) train interpretable model on predictors

interpret

Ensemble of Models



Schematic of a logistic regression classifier.

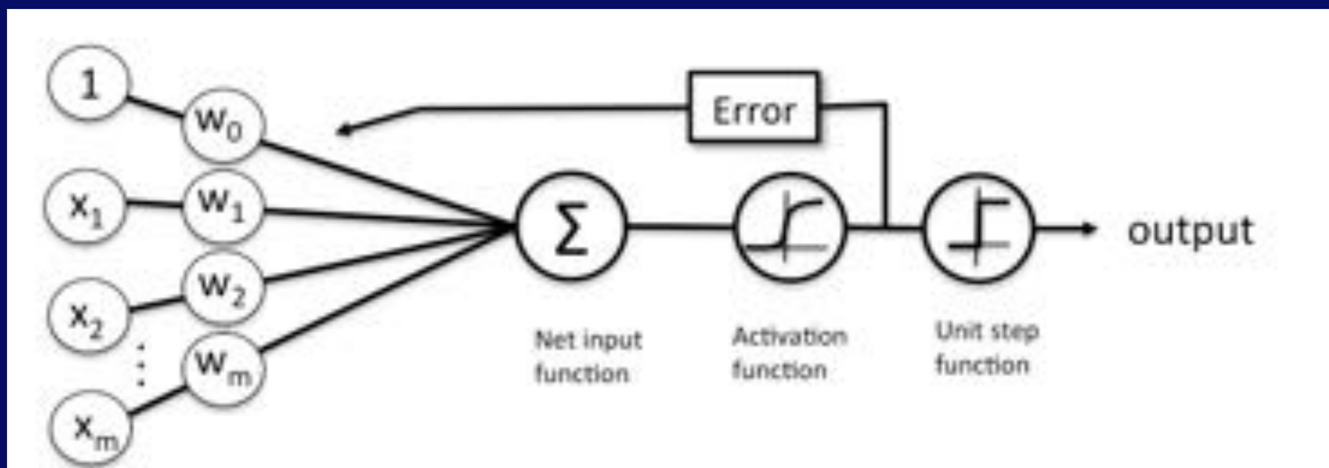


1) train interpretable linear model

2) train interpretable decision tree

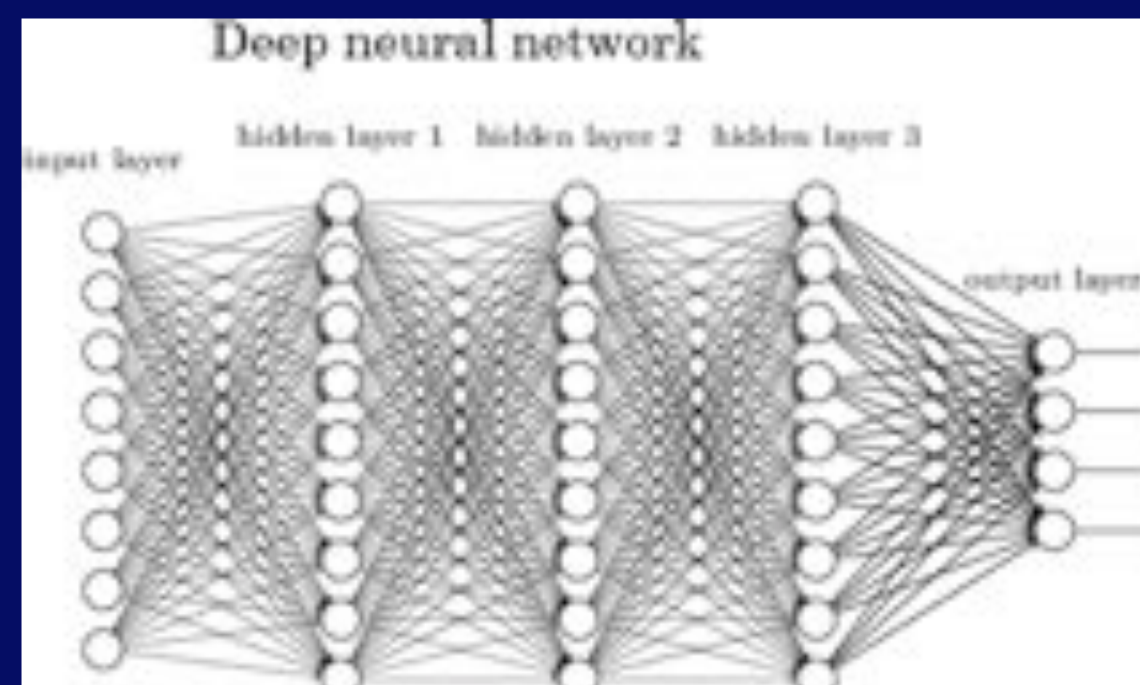
compare

Which is easier to interpret?



Schematic of a logistic regression classifier.

or



Answer: it depends!

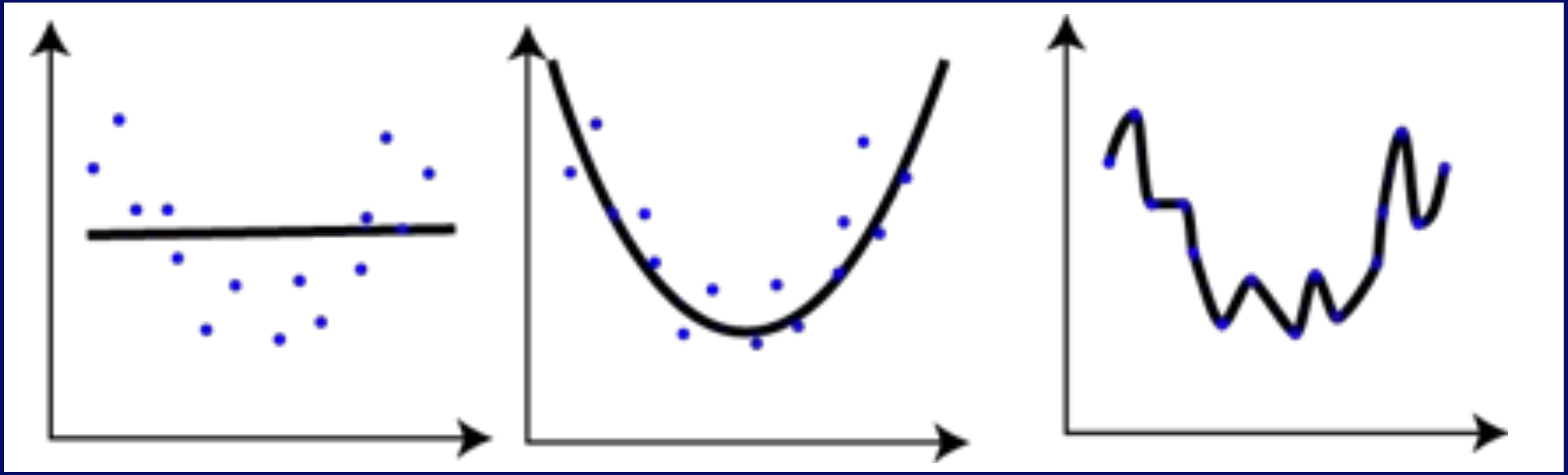
A linear model with highly engineered features and high-dimensional variables may not be very interpretable

but: linear models have a better track record for modelling the natural world and identifying weaknesses in the training data

Think carefully about your goals, your features, and your feature engineering!

Model Selection

Model Selection



1) avoid overfitting (prediction)
2) decide between (physical) models (inference)

Possible models?

1) algorithms
2) algorithm-specific parameters
3) regularization parameters
4) feature selection

Cross-validation

1) hold-out cross-validation
2) k-fold cross-validation
3) leave-one-out (LOO) cross validation
4) random subset cross validation

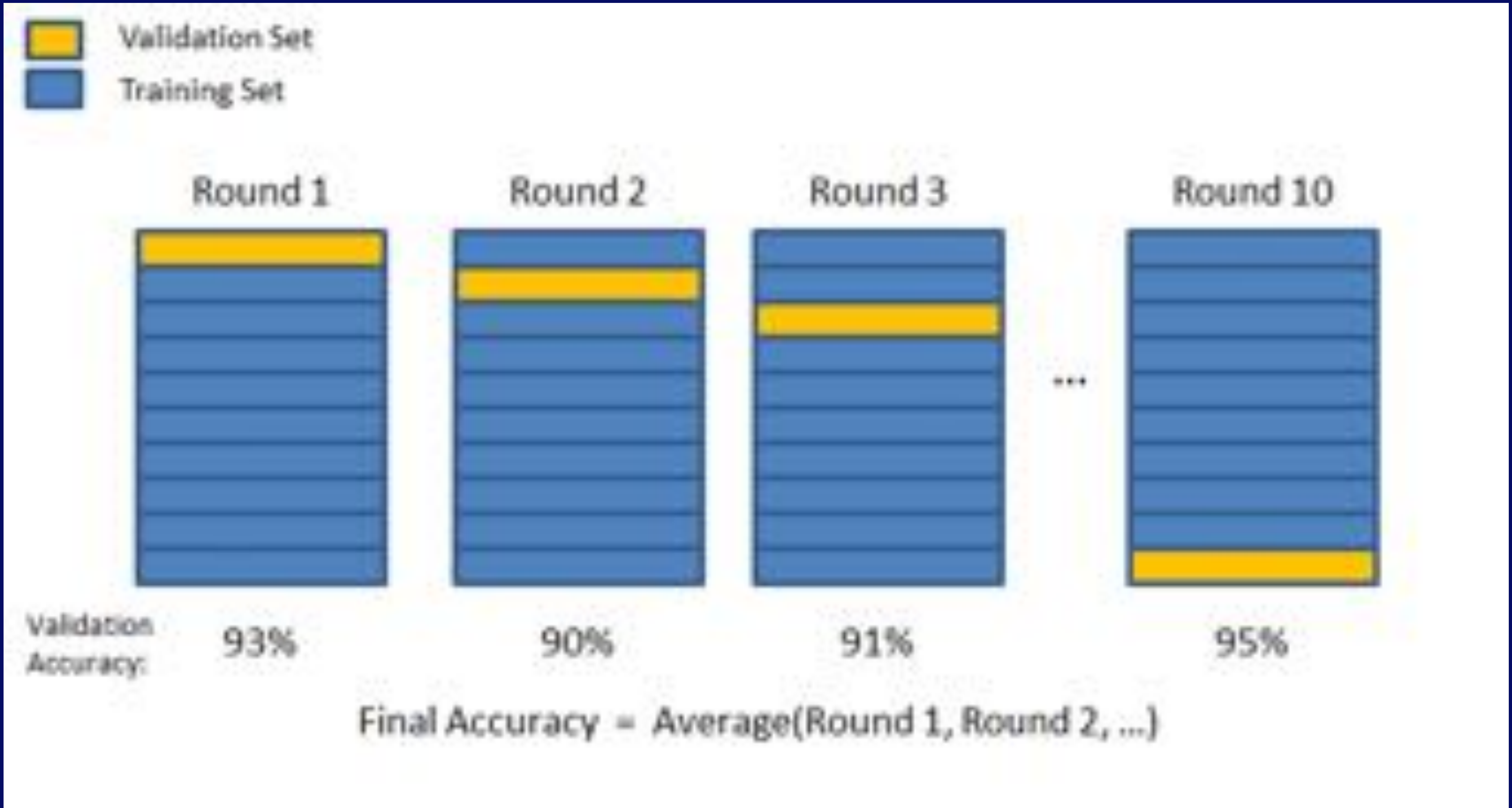
Nomenclature

training set: a data set to train your algorithm on

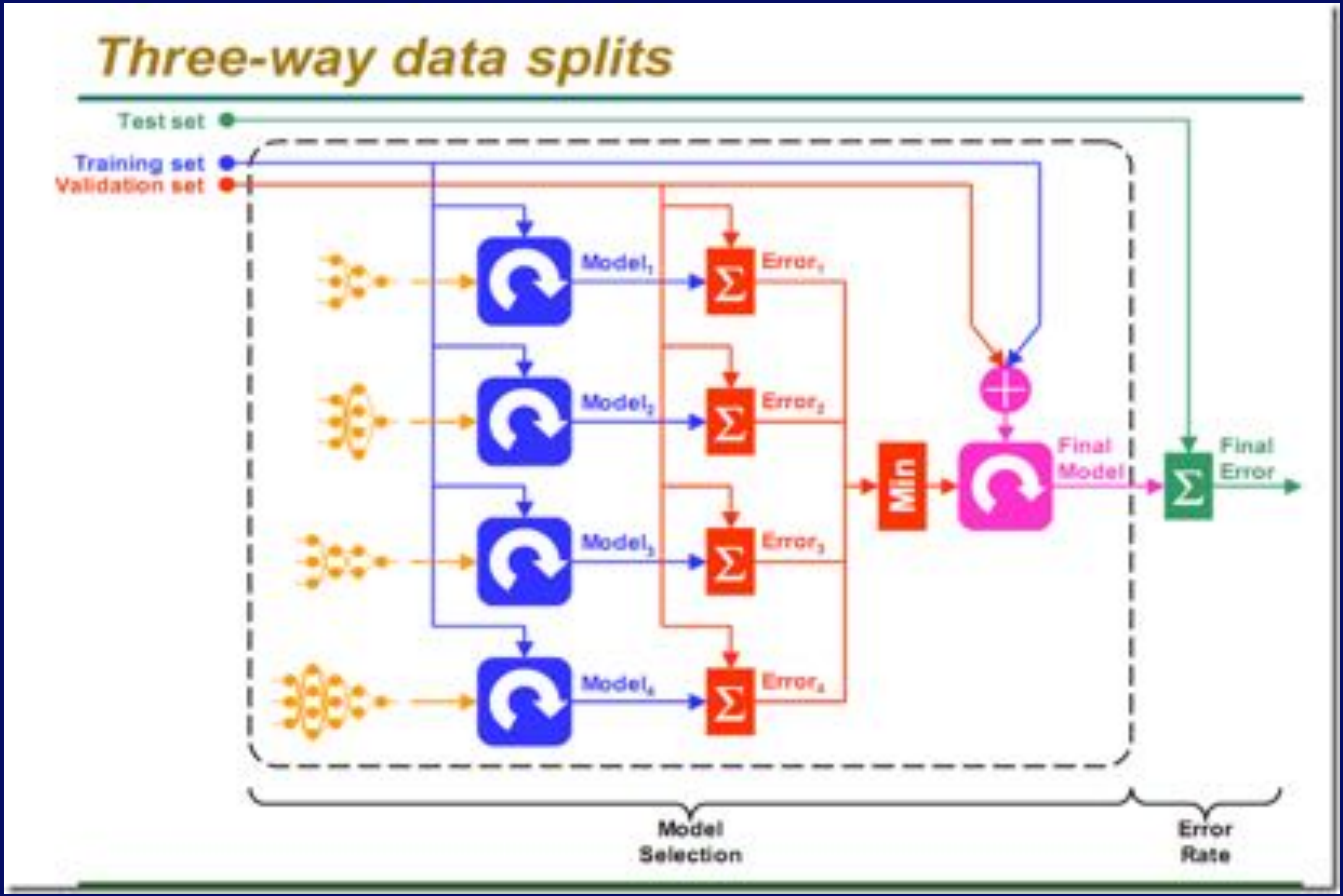
validation set: a data set to use for comparing the performance of different models

test set: a data set reserved to compute the error estimate of the final chosen model

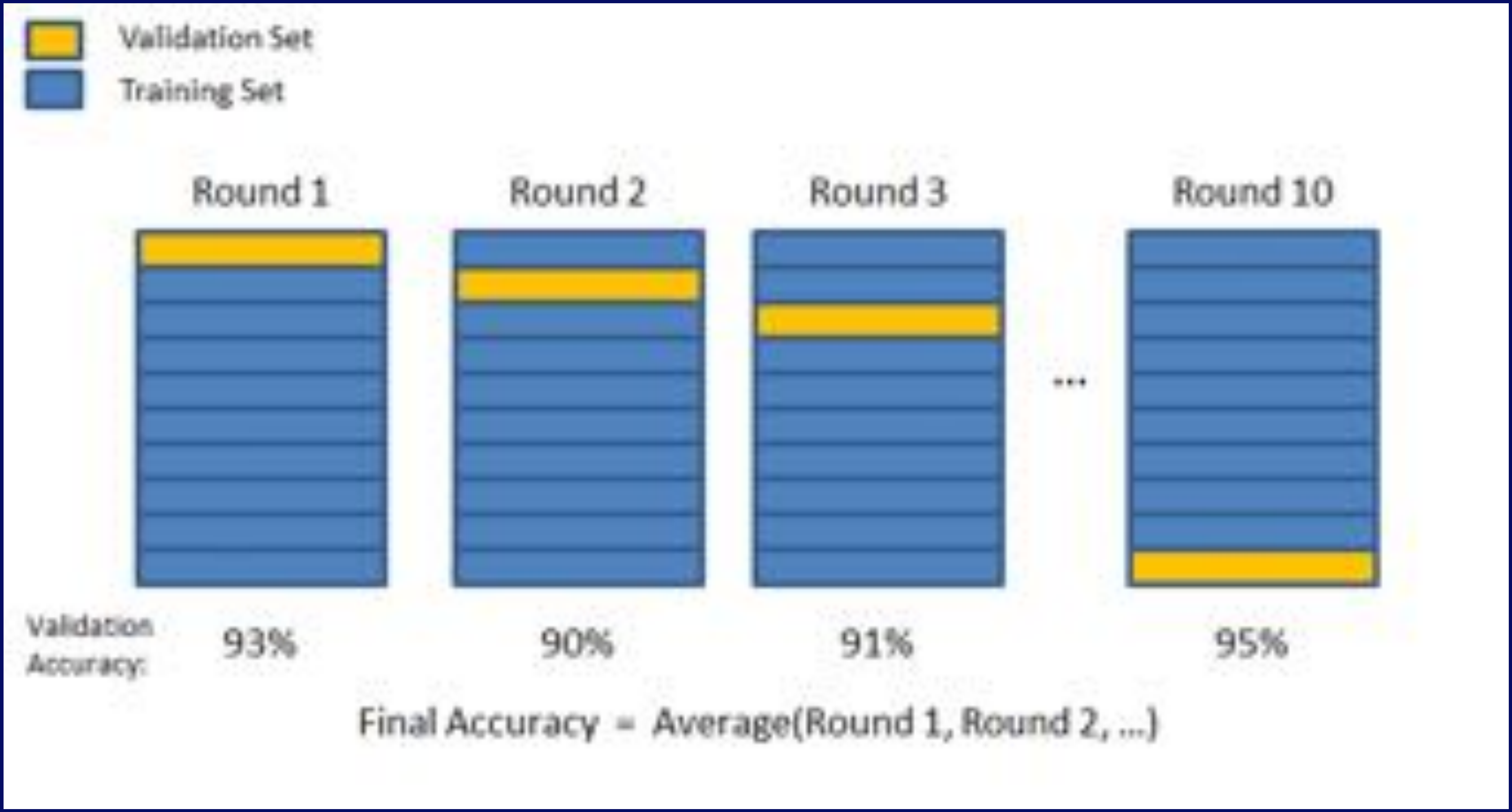
Hold-out + k-fold cross-validation



Hold-out + k-fold cross-validation



Leave one out cross validation



special case: $k = N$

Random subset cross validation

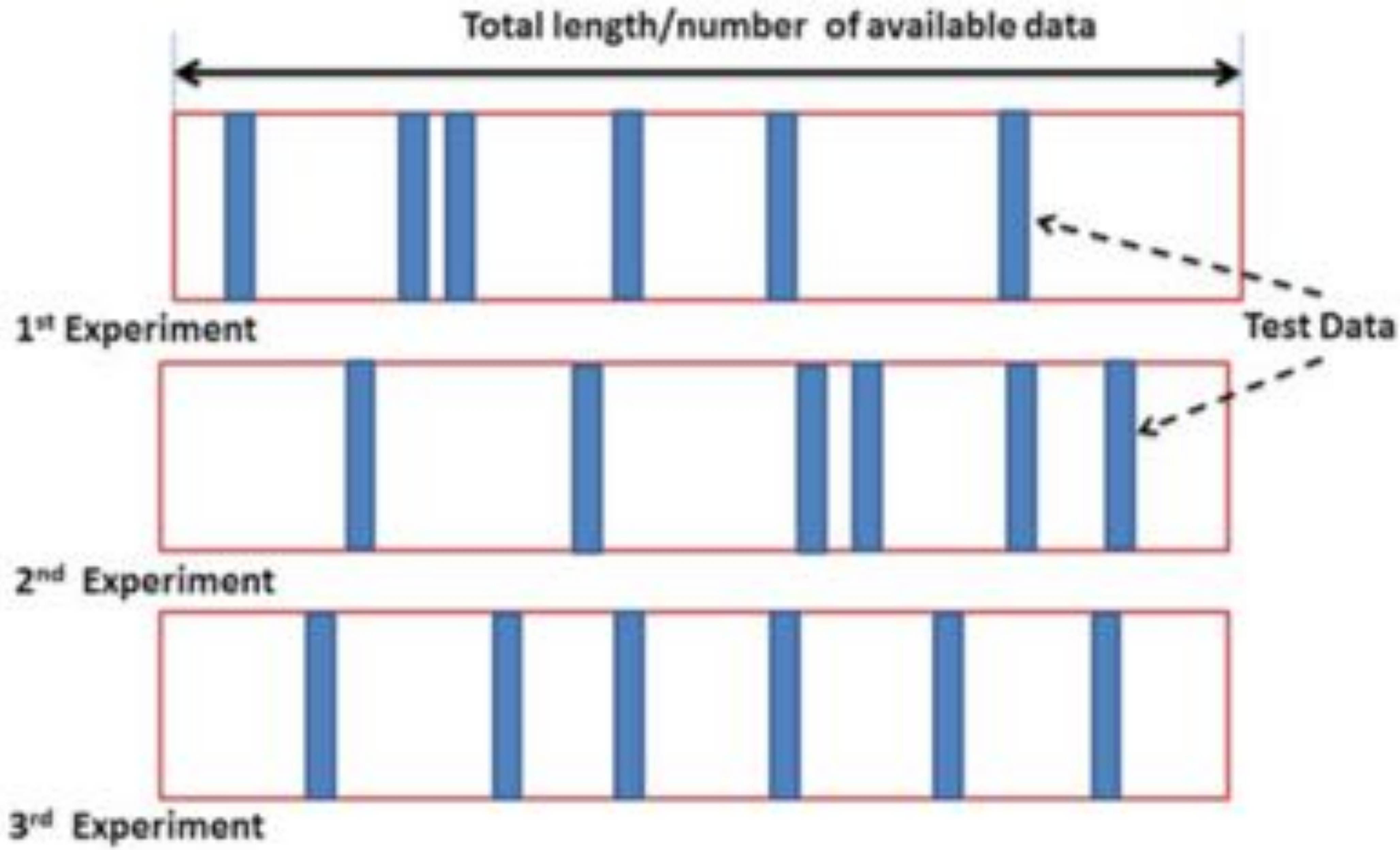
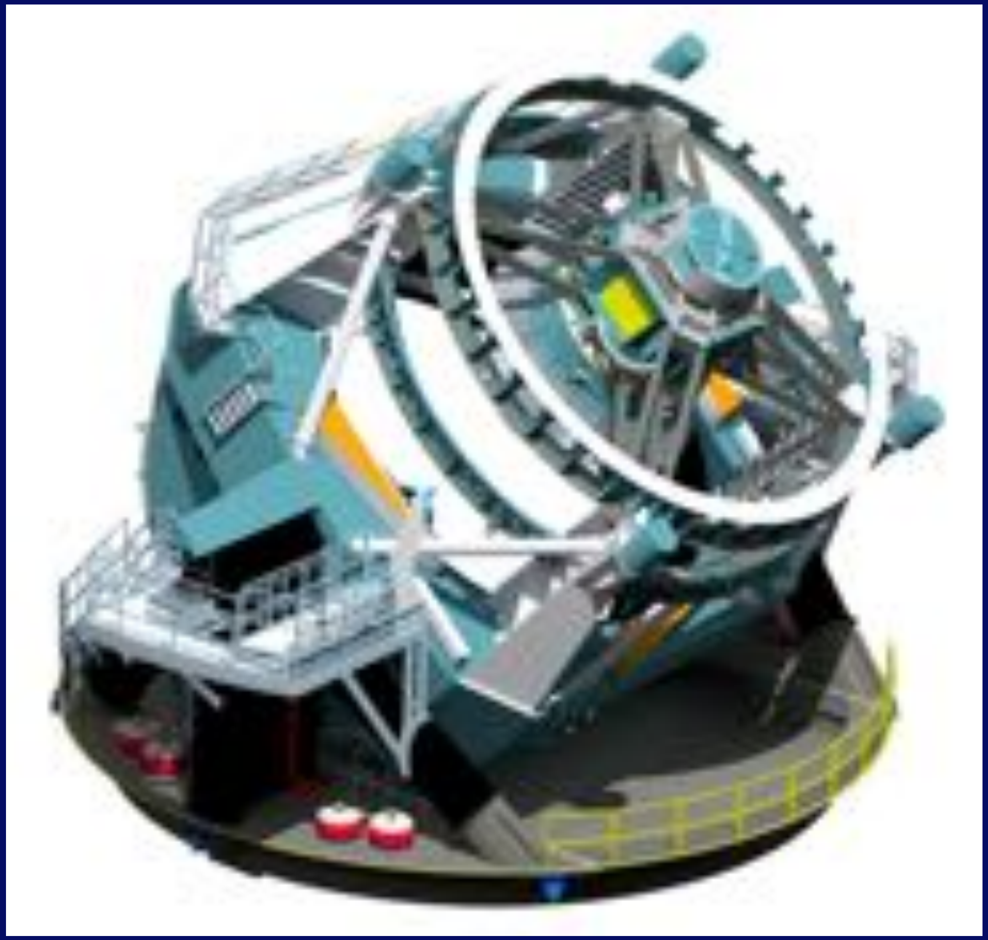


Fig. 3.7 Data splitting in the random sub-sampling approach

What do you compare during cross validation?

Example: LSST alerts!



10 million alerts per night

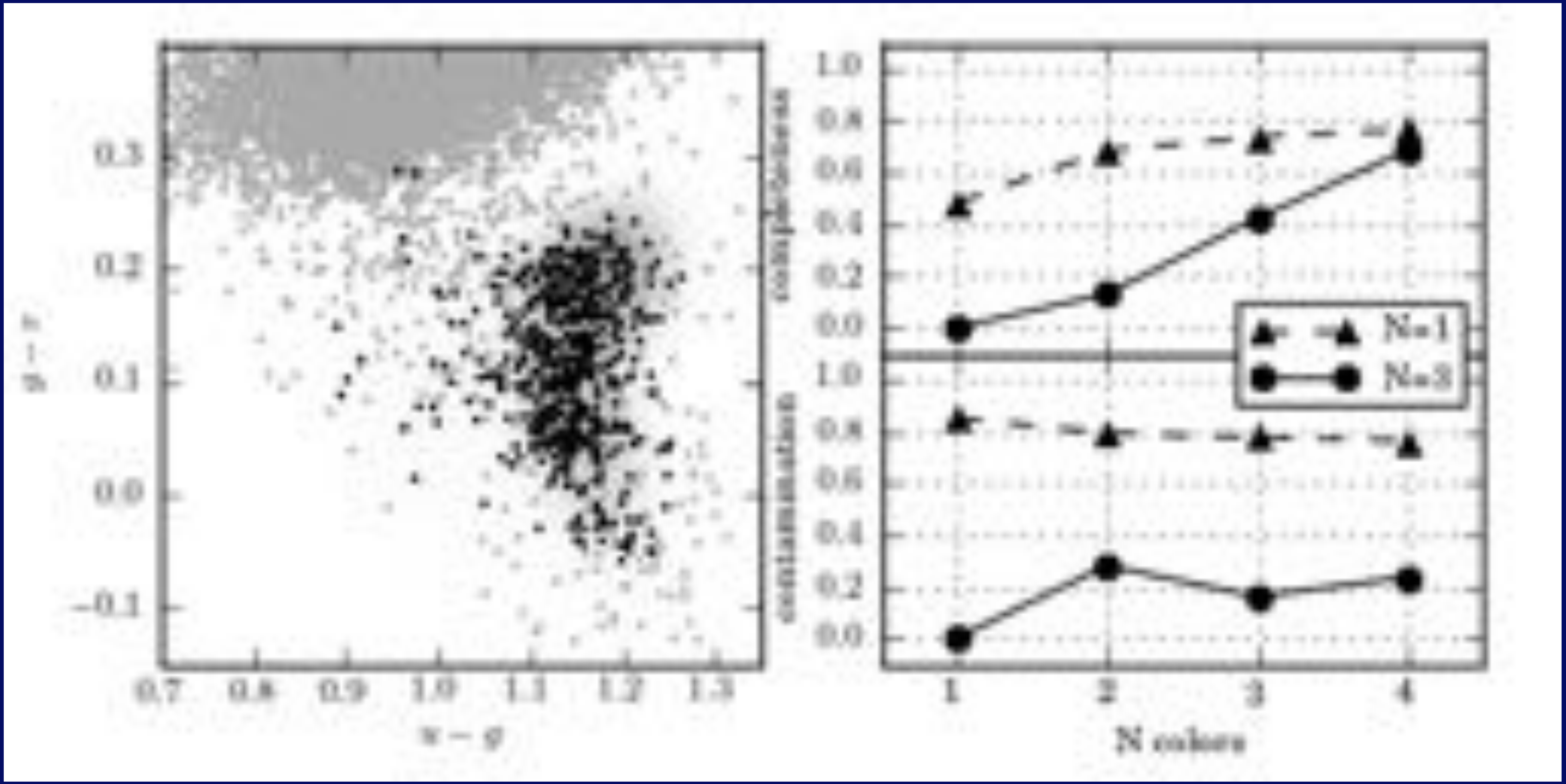
0.1% interesting

		Predicted	
		+	-
Actual	+	5000	0
	-	200	1

Accuracy Paradox

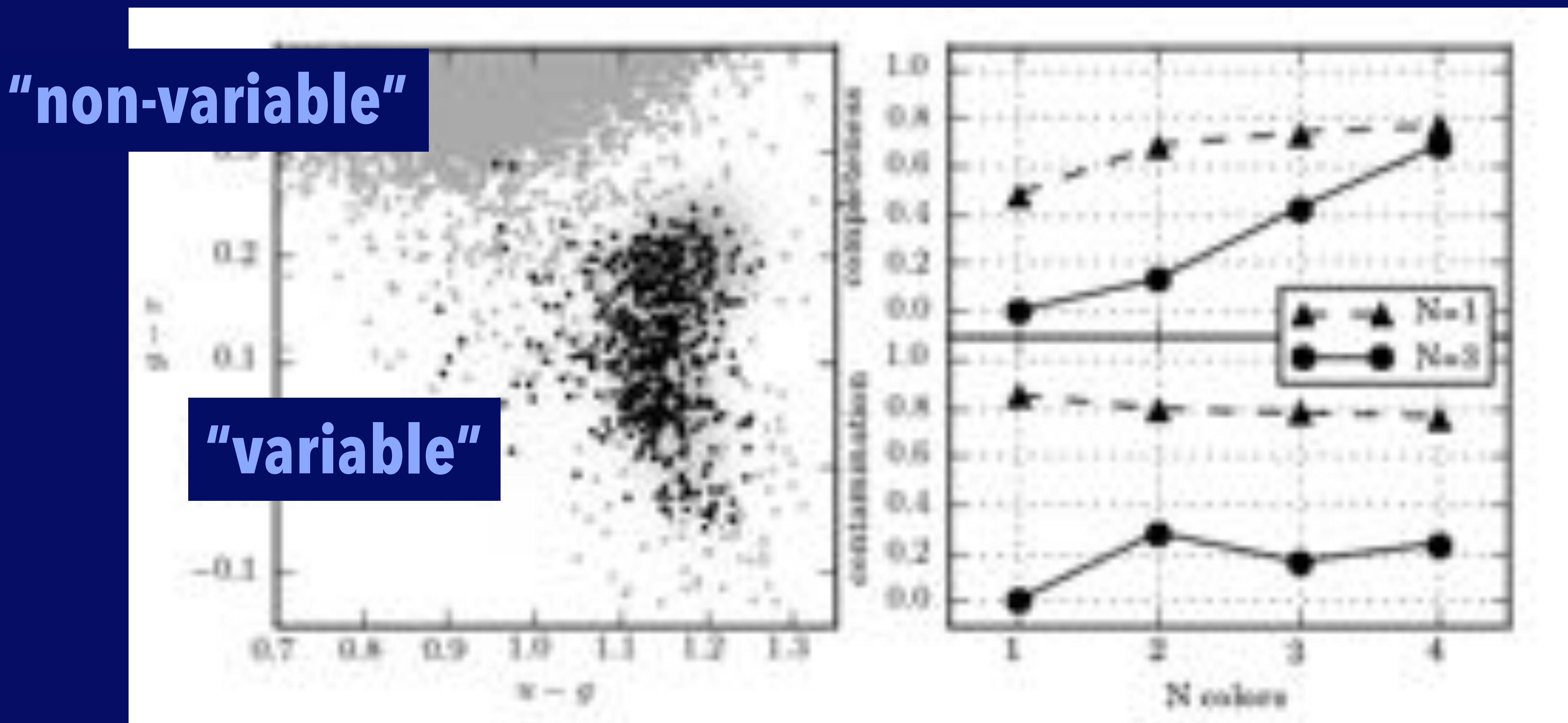
different metrics are useful for different use cases!

Unsupervised classification



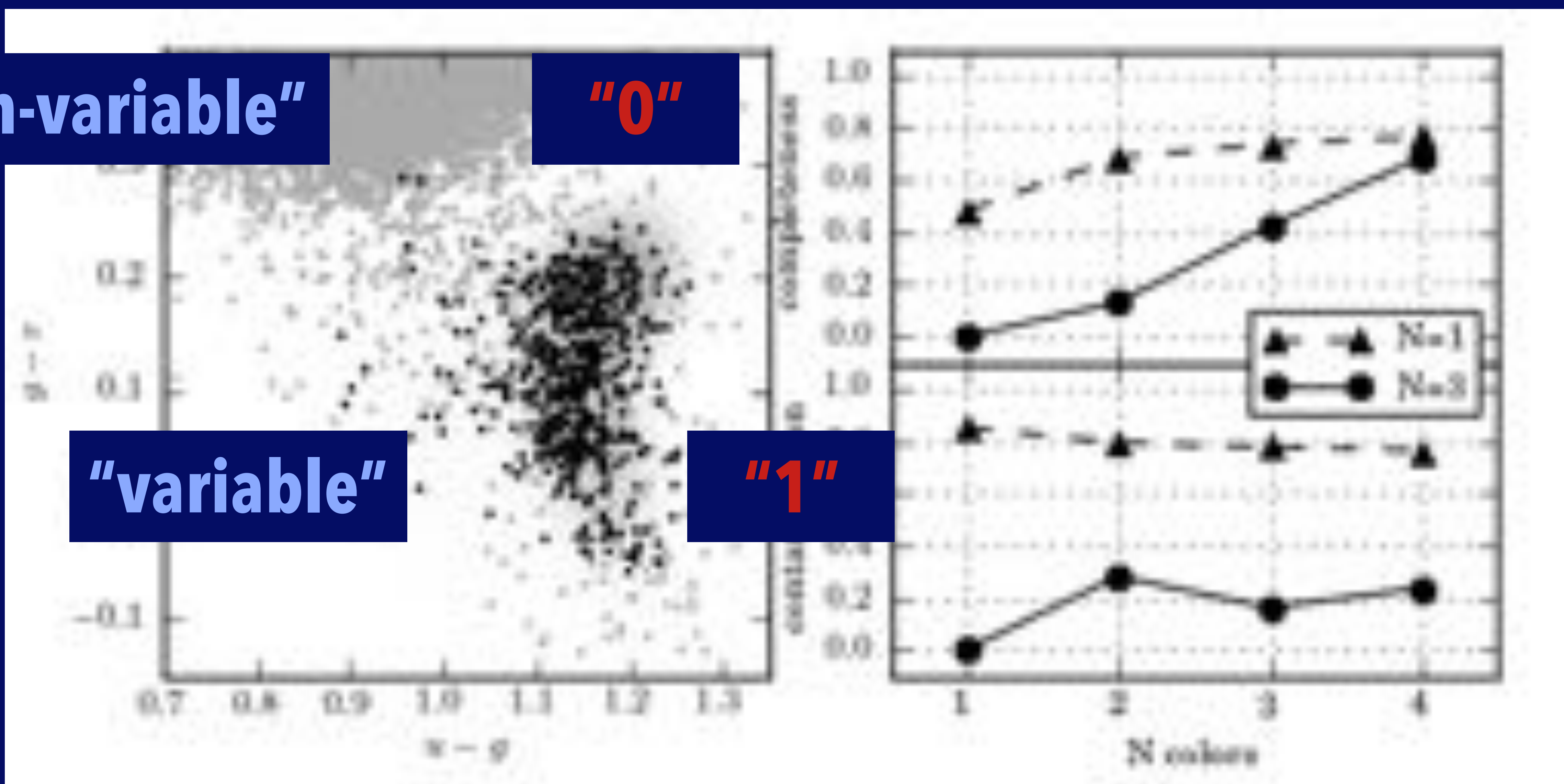
Unsupervised classification

Human:



Unsupervised classification

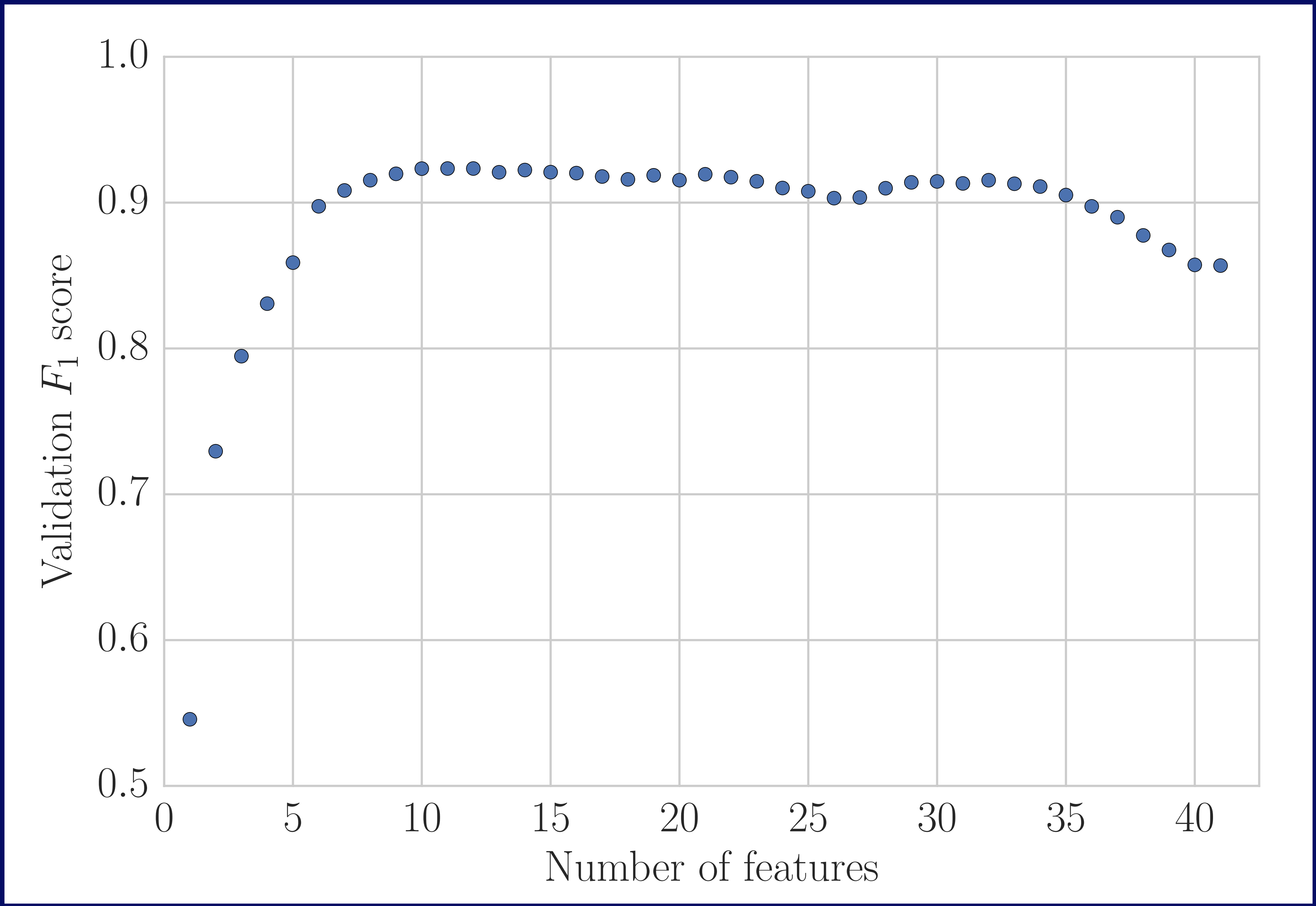
Human: **Computer:**



Unsupervised classification

1) adjusted Rand index (ARI)
2) adjusted mutual information score
3) Silhouette coefficient
4) Information criteria

Feature Selection



Maybe only a subset of available features is predictive!

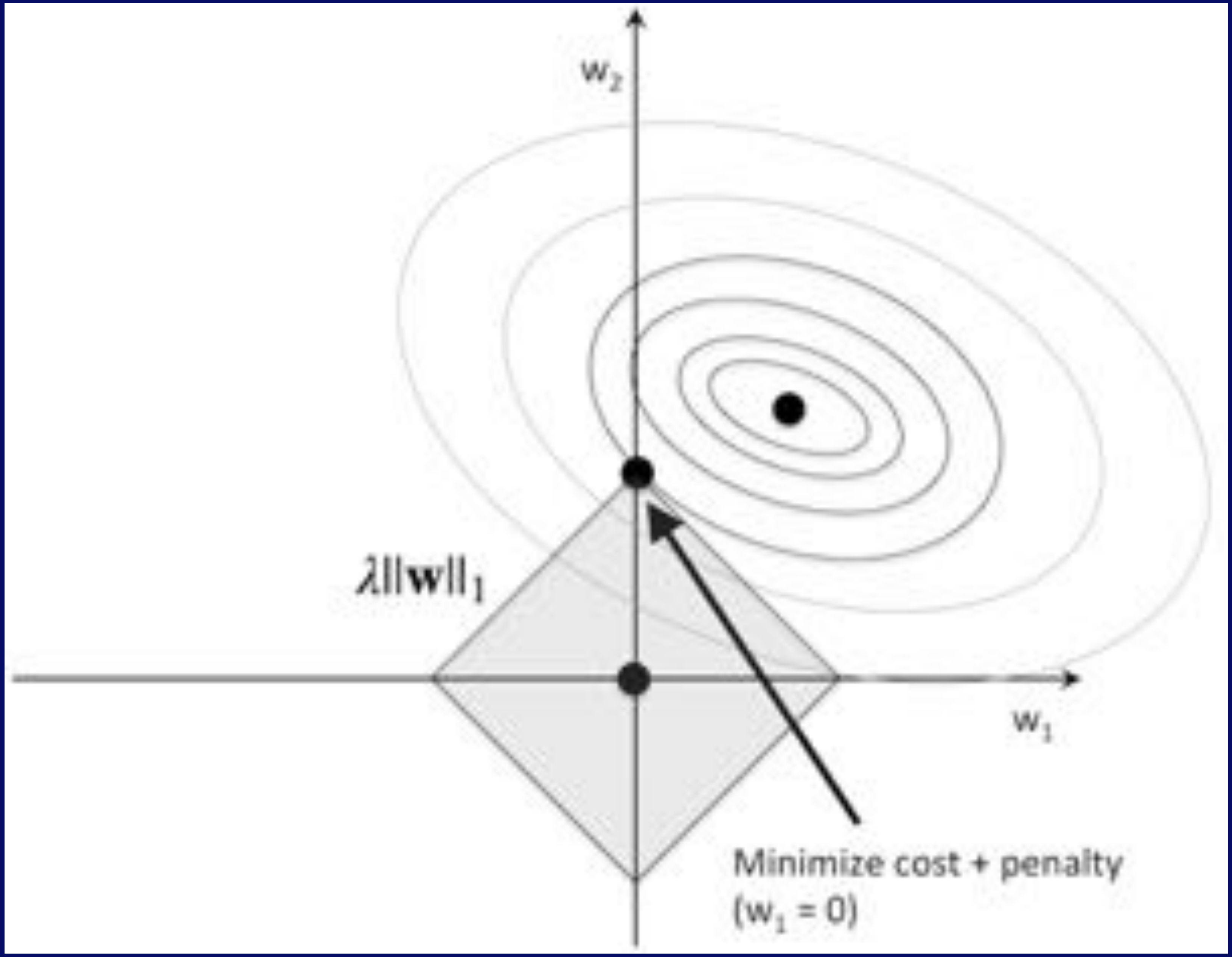
Exhaustive search:

$>2^n$ model evaluations

Linear models:
L1 regularization

$$L1 : \lambda \,\|\mathbf{w}\|_1 = \lambda \sum_{j=1}^{m} |w_j|$$

$$SSE = \sum_{i=1}^{n} \left(\text{target}^{(i)} - \text{output}^{(i)}\right)^2 + L1$$



Forward or backward search:

$>n^2$ model evaluations

search procedure is called **forward search**:

1. Initialize $\mathcal{F} = \emptyset$.

2. Repeat {

   (a) For $i = 1, \ldots, n$ if $i \notin \mathcal{F}$, let $\mathcal{F}_i = \mathcal{F} \cup \{i\}$, and use some version of cross validation to evaluate features $\mathcal{F}_i$. (I.e., train your learning algorithm using only the features in $\mathcal{F}_i$, and estimate its generalization error.)

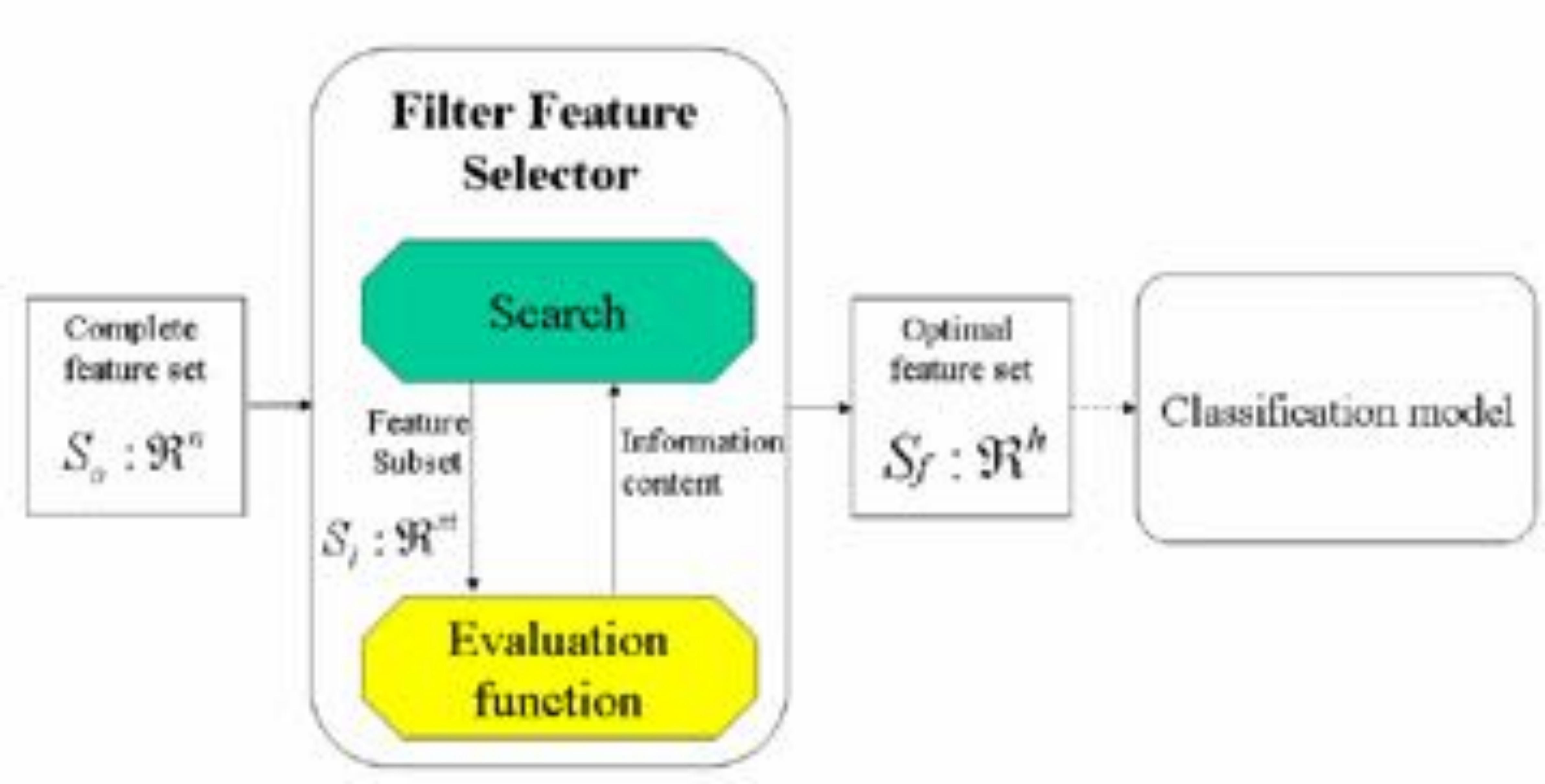
   (b) Set $\mathcal{F}$ to be the best feature subset found on step (a).

   }

3. Select and output the best feature subset that was evaluated during the entire search procedure.

Filter feature selection:

>n model evaluations



e.g. correlation between features and labels, Kullback-Leibler divergence, …

The Bayesian Perspective

“I have a generative model and a likelihood”

+ usually easier to interpret and reason about

– usually much more computationally expensive

approximation of the Bayes factor

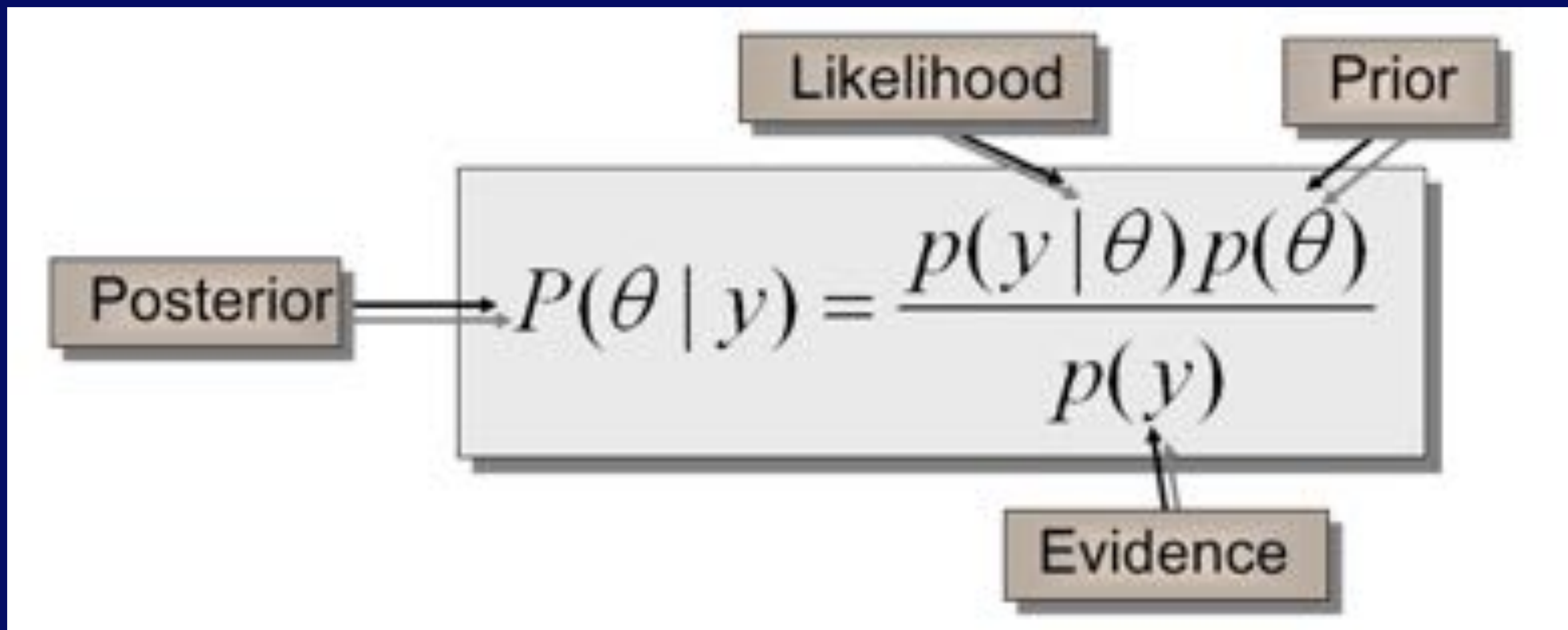
↓

Information Criteria

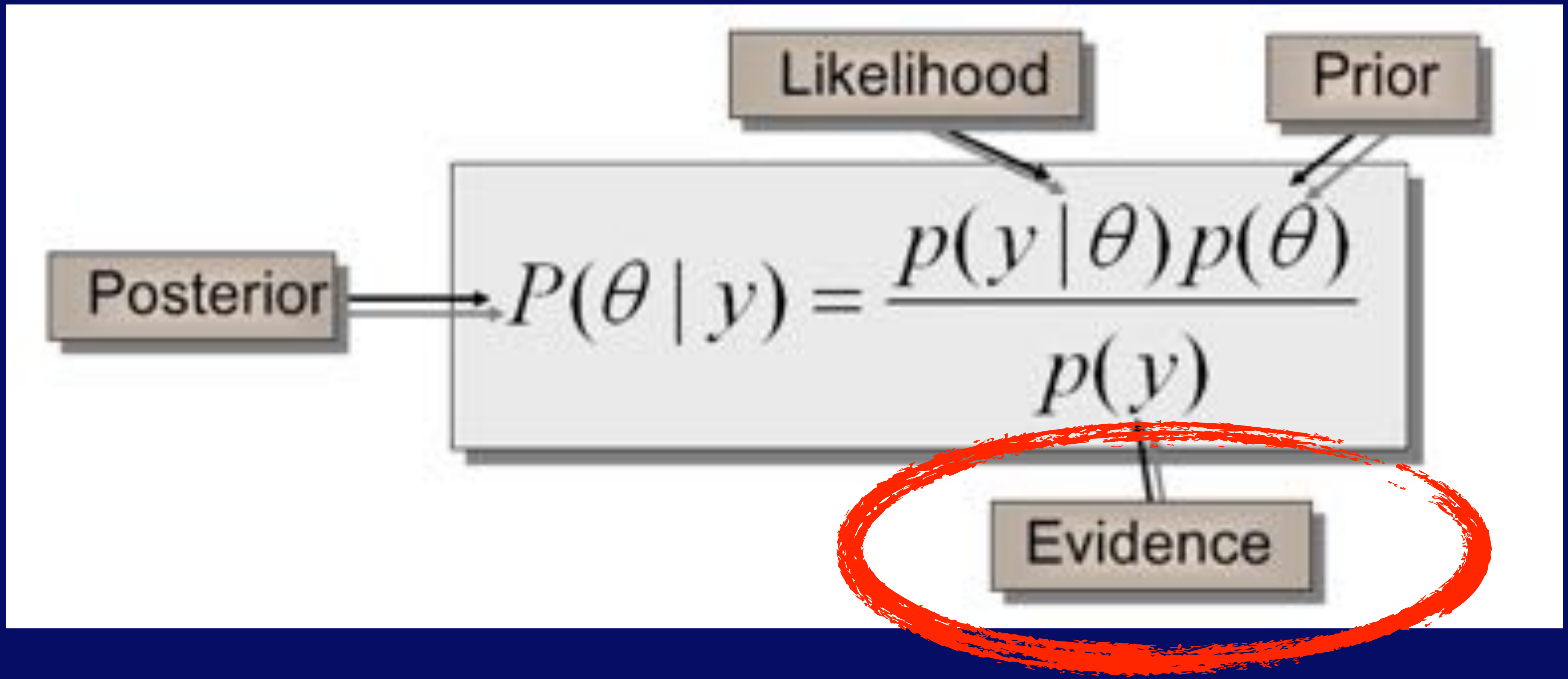
↑

approximation of Bayesian cross validation

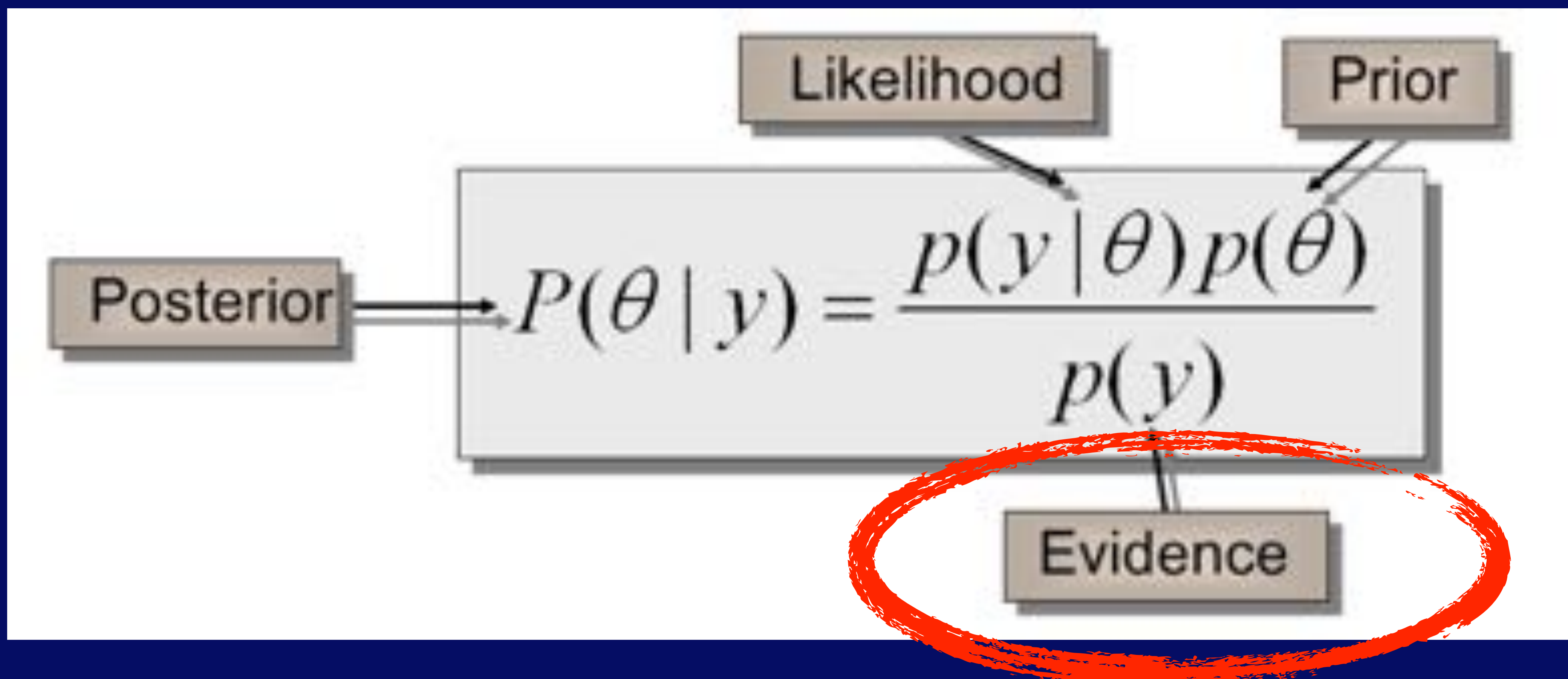
approximation of the Bayes factor



approximation of the Bayes factor



approximation of the Bayes factor



$$P(m|\mathbf{y}) = \frac{P(\mathbf{y}|m)P(m)}{P(\mathbf{y})}$$

Bayesian Information Criterion

\# data points   # parameters   likelihood

$$\mathrm{BIC} = \ln(n)k - 2\ln(\hat{L}).$$

- rough approximation of the Bayes factor (for unit uniform prior)
- conservative estimate
- useful as a baseline

approximation of Bayesian CV

Akaike Information Criterion (AIC)*

Deviance Information Criterion (DIC)

Widely Applicable Information Criterion (WAIC)

are all approximation to leave-one-out cross-validation in Bayesian models (Gelman et al, 2013)

Akaike Information Criterion

$$\mathrm{AIC} = 2k - 2\ln(\hat{L})$$

only works for linear models with flat priors or

models with a normally distributed posterior

Deviance Information Criterion

$$\mathrm{AIC} = 2k - 2\ln(\hat{L})$$

replace with data-based bias correction

replace with posterior mean

$$\text{computed } p_{\mathrm{DIC}} = 2\left(\log p(y|\hat{\theta}_{\mathrm{Bayes}}) - \frac{1}{S}\sum_{s=1}^{S}\log p(y|\theta^s)\right).$$

Resources:

- http://www.stat.columbia.edu/~gelman/research/published/waic_understand3.pdf
- https://github.com/marcotcr/lime
- https://arxiv.org/abs/1606.03490
- https://www.stat.washington.edu/raftery/Research/PDF/kass1995.pdf
- Gelman et al, Bayesian Data Analysis, 2004
- Bishop, Pattern Recognition + Machine Learning
- http://users.isr.ist.utl.pt/~wurmd/Livros/school/Bishop%20-%20Pattern%20Recognition%20And%20Machine%20Learning%20-%20Springer%20%202006.pdf
- https://www.stat.washington.edu/research/reports/1999/tr347.pdf