Data Processing:

cross validation, bias control, missing data, limited size, and data heterogeneity

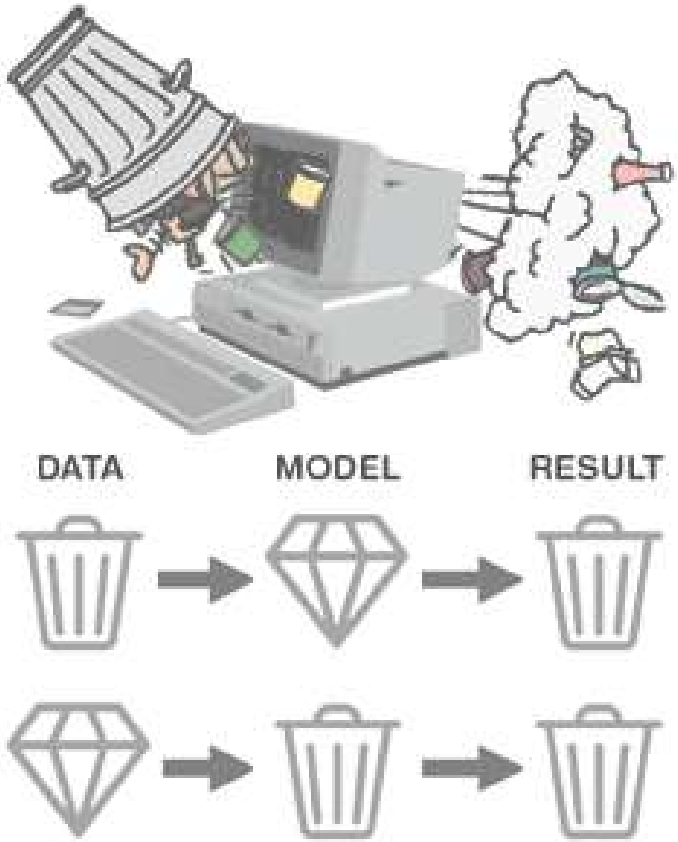
Or: what machine learning doesn't want you to know!

John Kang

PGY4 radiation oncology resident

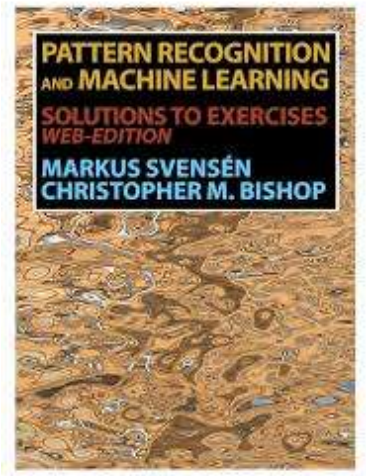
University of Rochester

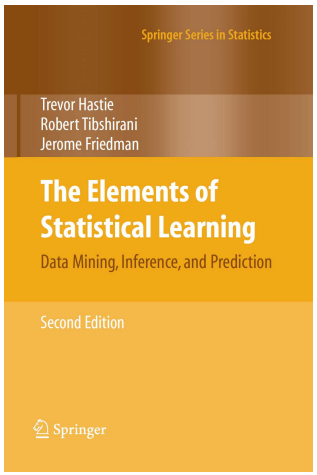
Why care about data preprocessing?

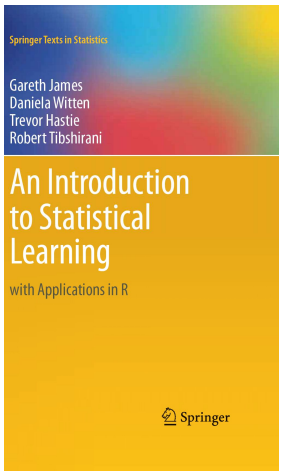


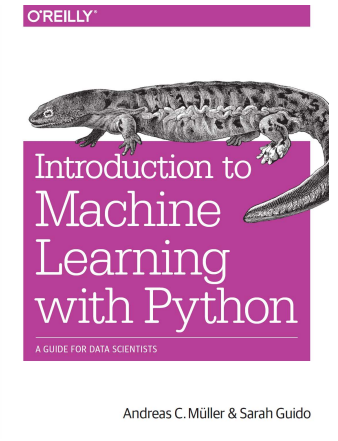

THE BIG
SC
SPURS
21
FUNDAMENTAL

Despite not data preprocessing is important, it is covered in ML text books







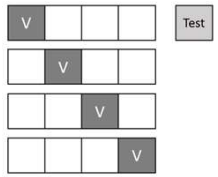


- Kaggle tutorials
- Stack Exchange

Outline

RNG

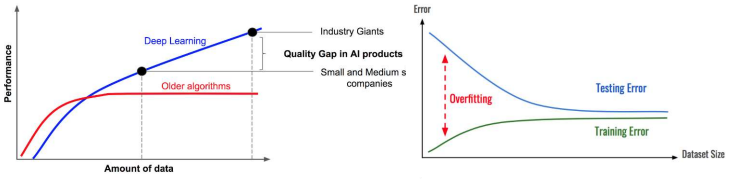
- Cross validation



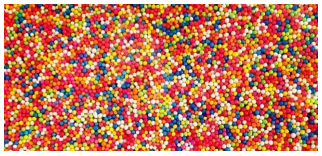
- Bias control

- Missing data

- Limited size



- Data heterogeneity



Which order of feature selection & model selection?

Option 1:

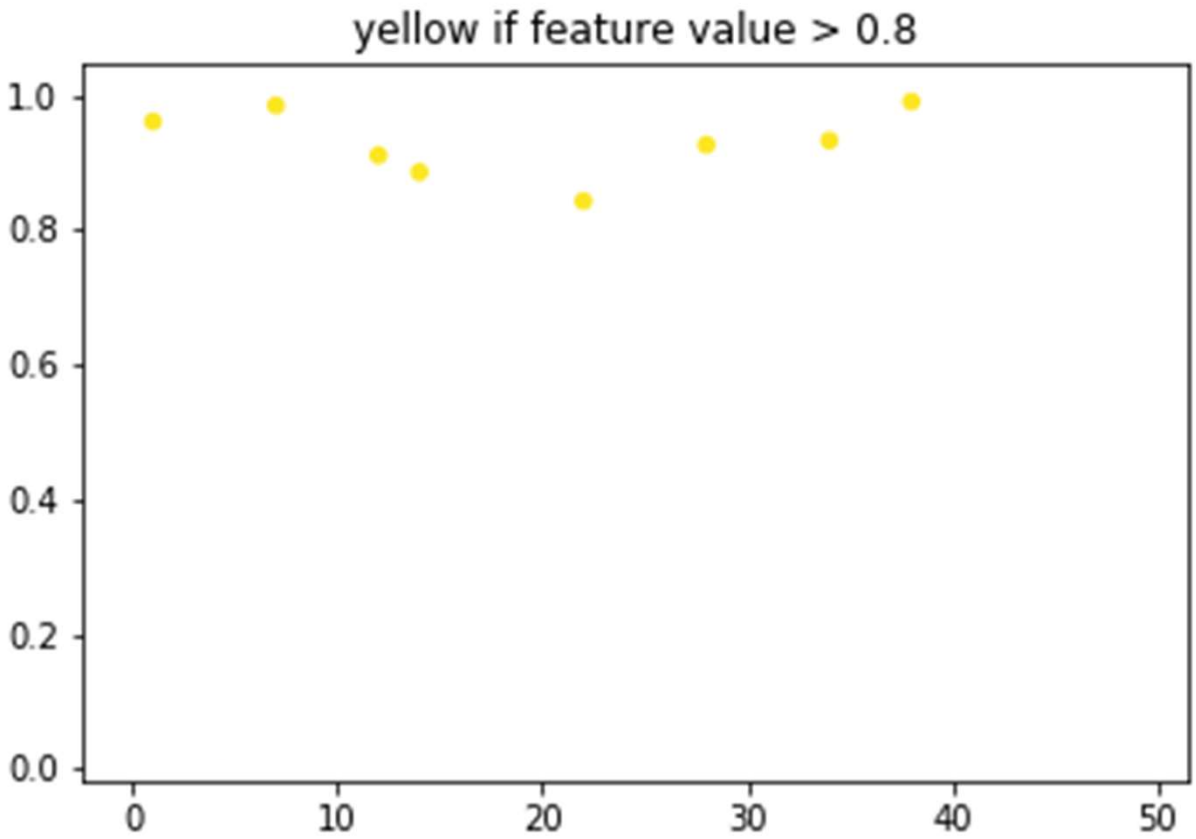
1) **Feature selection**
   X' = Run_filter(X, y)

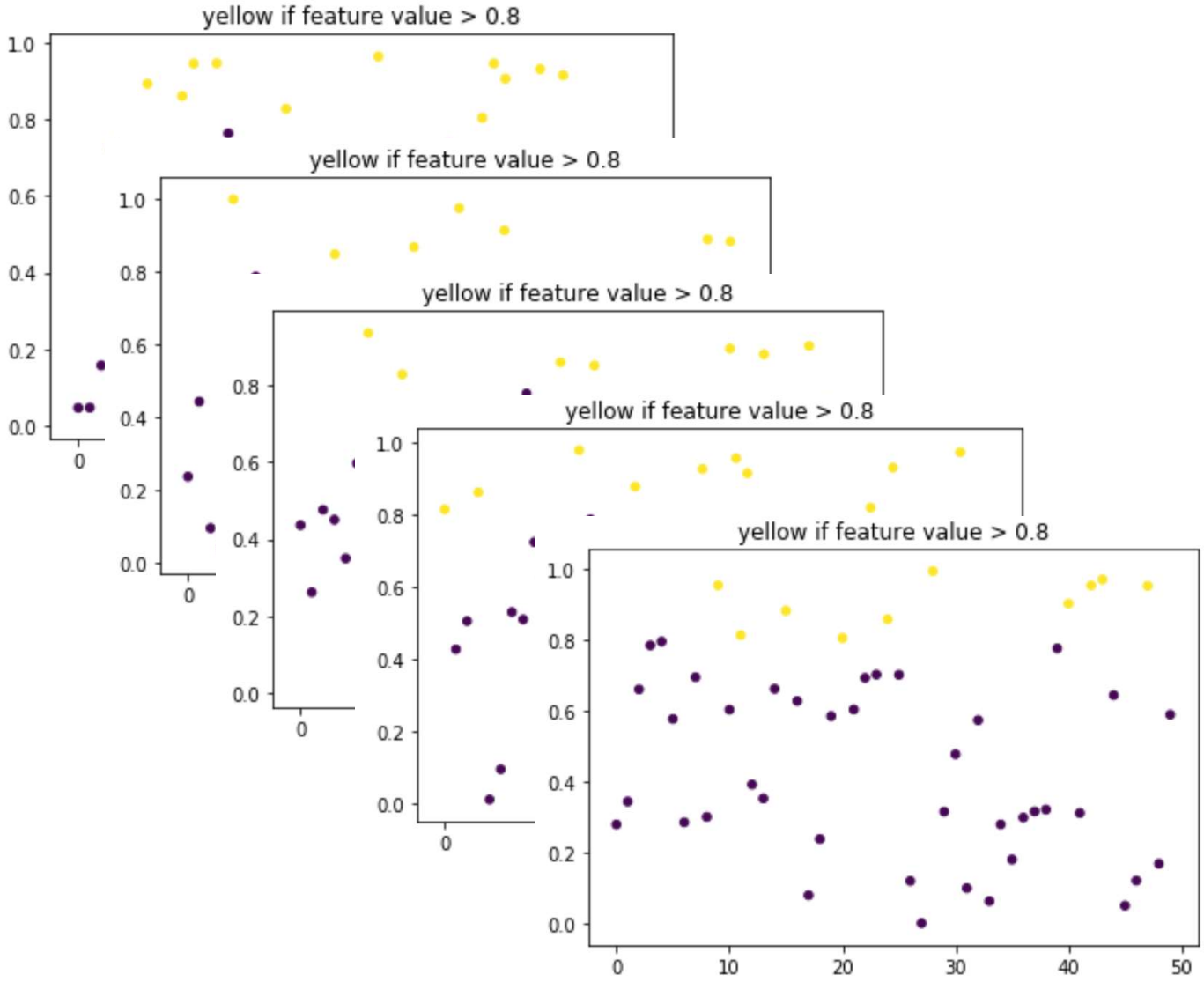
2) **Cross validation (Model selection)**
   err = Mdl_select(X',y)

Option 2:

1) **CV fold 1**
   a) **Feature selection**
   …
2) **CV fold n**
   a) **Feature selection**

Filtering features




yellow if feature value > 0.8
yellow if feature value > 0.8
yellow if feature value > 0.8
yellow if feature value > 0.8
yellow if feature value > 0.8
0.0
0.2
0.4
0.6
0.8
1.0
0
10
20
30
40
50

Issues of sequential model selection are not well described in textbooks

- Subsequently, discussed in several Stack Exchange posts
- Feature selection and cross-validation
- Tune parameters with nested SVM in MATLAB
- Nested cross validation for model selection

DAVID A. FREEDMAN*

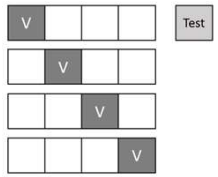
A Note on Screening Regression Equations

Consider developing a regression model in a context where substantive theory is weak. To focus on an extreme case, suppose that in fact there is no relationship between the dependent variable and the explanatory variables. Even so, if there are many explanatory variables, the R^2 will be high. If explanatory variables with small t statistics are dropped and the equation refitted, the R^2 will stay high and the overall F will become highly significant. This is demonstrated by simulation and by asymptotic calculation.

© *The American Statistician, May 1983, Vol. 37, No. 2*

Outline

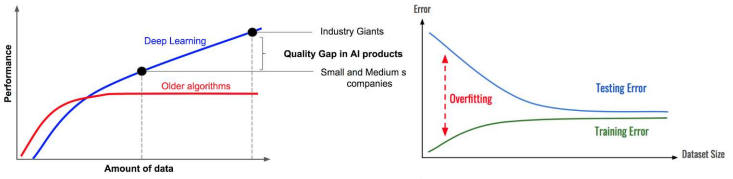
- Cross validation



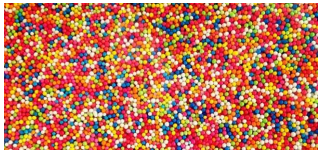
- **Bias control**

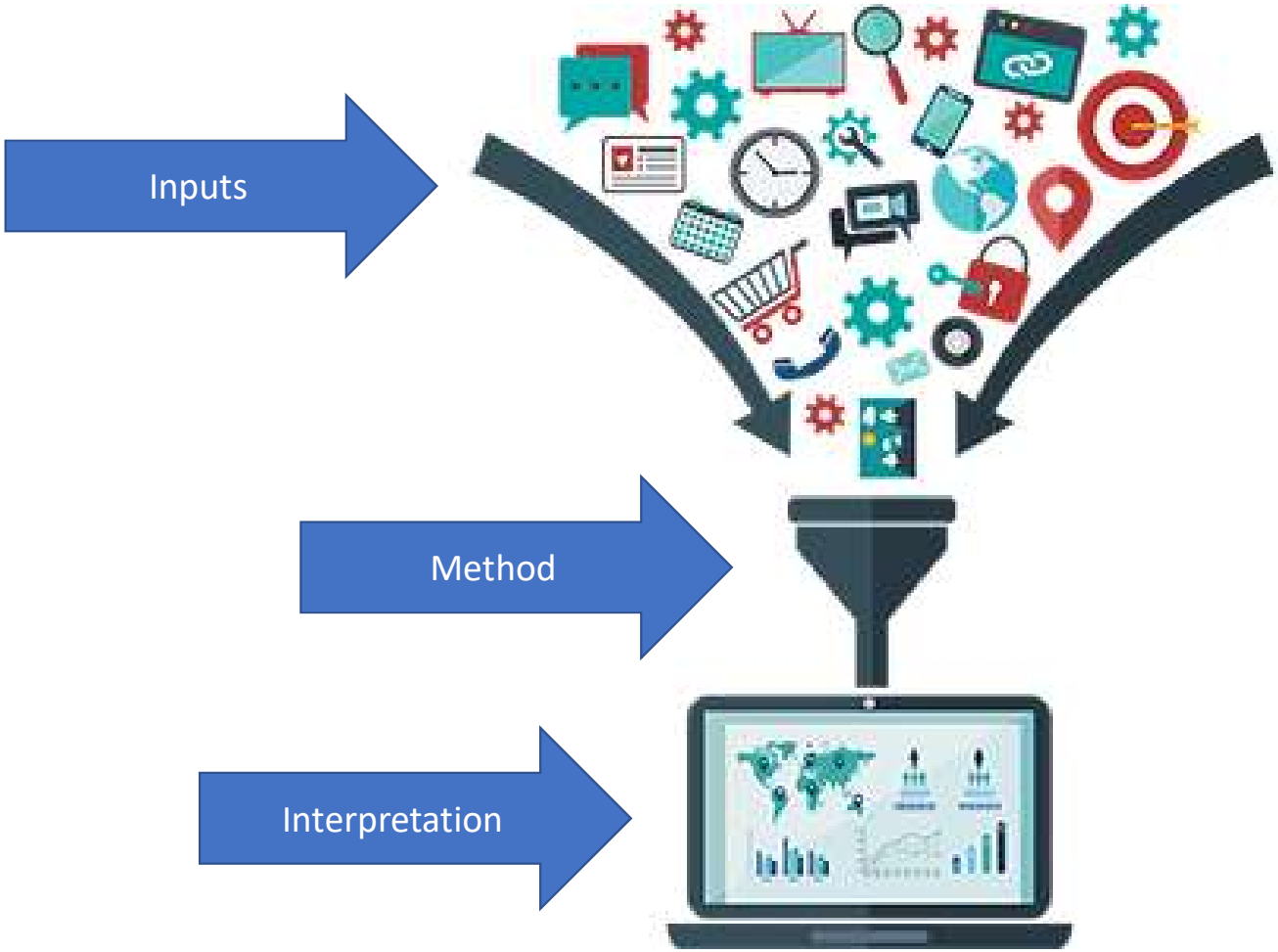
- Missing data

- Limited size



- Data heterogeneity





Source: www.supplychainquarterly.com

Lancet 2002; **359:** 572–77

Use of proteomic patterns in serum to identify ovarian cancer

Emanuel F Petricoin III, Ali M Ardekani, Ben A Hitt, Peter J Levine, Vincent A Fusaro, Seth M Steinberg, Gordon B Mills, Charles Simone, David A Fishman, Elise C Kohn, Lance A Liotta

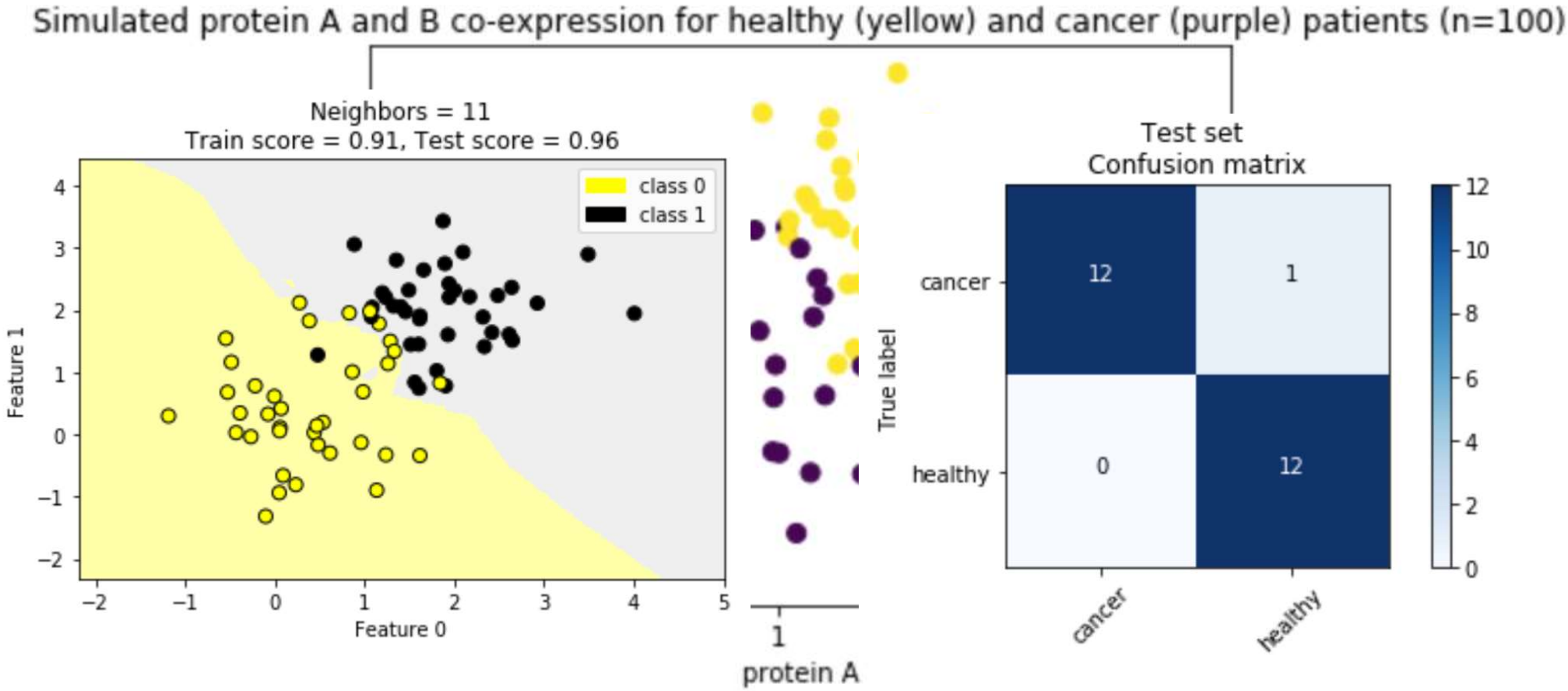
Background New technologies for the detection of early-stage ovarian cancer are urgently needed. Pathological changes within an organ might be reflected in proteomic patterns in serum. We developed a bioinformatics tool and used it to identify proteomic patterns in serum that distinguish neoplastic from non-neoplastic disease within the ovary.

Methods Proteomic spectra were generated by mass spectroscopy (surface-enhanced laser desorption and ionisation). A preliminary "training" set of spectra derived from analysis of serum from 50 unaffected women and 50 patients with ovarian cancer were analysed by an iterative searching algorithm that identified a proteomic pattern that completely discriminated cancer from non-cancer. The discovered pattern was then used to classify an independent set of 116 masked serum samples: 50 from women with ovarian cancer, and 66 from unaffected women or those with non-malignant disorders.

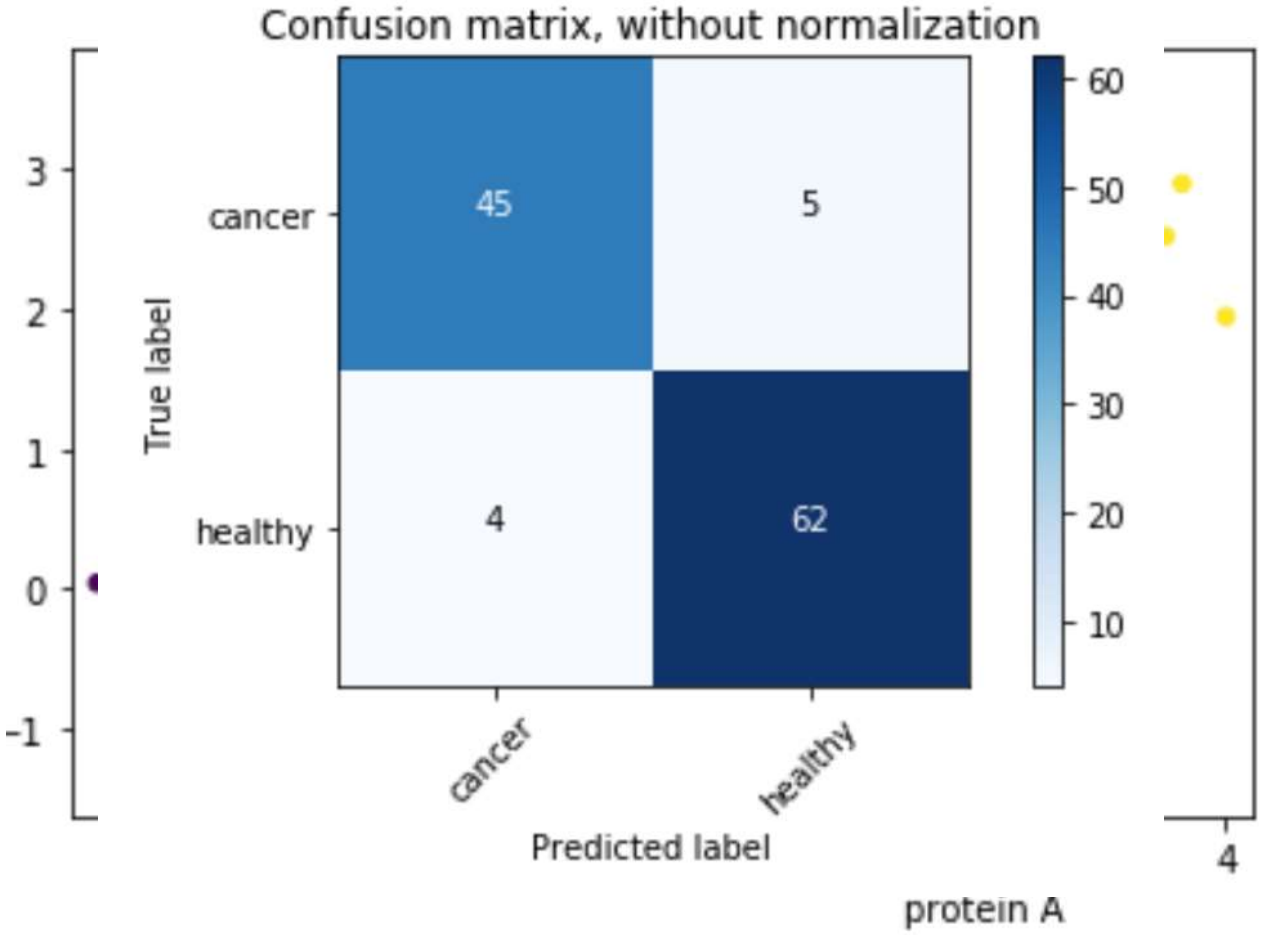
Findings The algorithm identified a cluster pattern that, in the training set, completely segregated cancer from non-cancer. The discriminatory pattern correctly identified all 50 ovarian cancer cases in the masked set, including all 18 stage I cases. Of the 66 cases of non-malignant disease, 63 were recognised as not cancer. This result yielded a sensitivity of 100% (95% CI 93–100), specificity of 95% (87–99), and positive predictive value of 94% (84–99).

Interpretation These findings justify a prospective population-based assessment of proteomic pattern technology as a screening tool for all stages of ovarian cancer in high-risk and general populations.

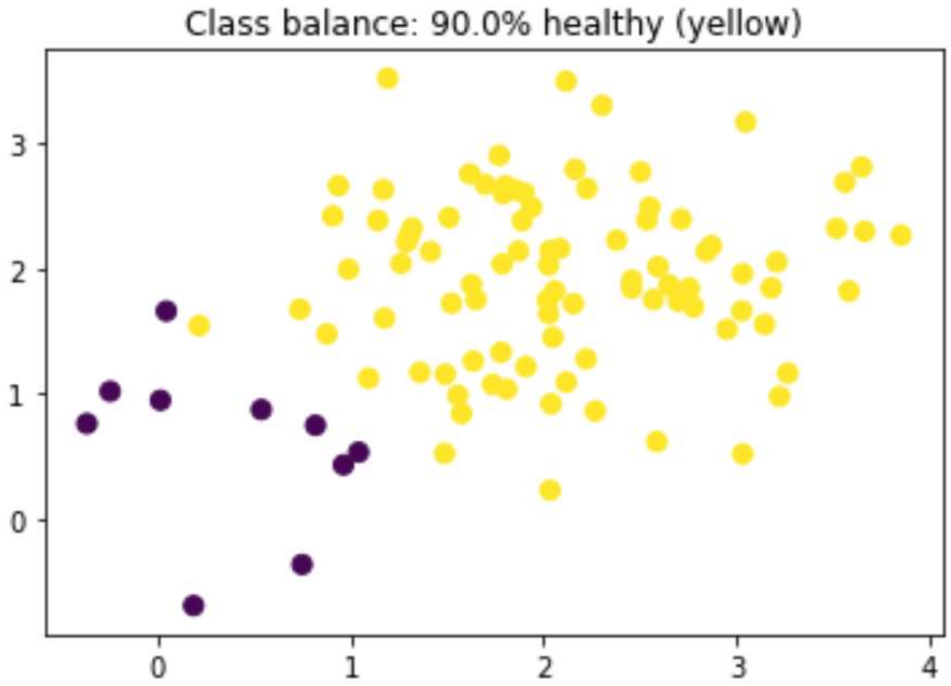
http://www.thelancet.com/pdfs/journals/lancet/PIIS0140-6736(02)09388-1.pdf

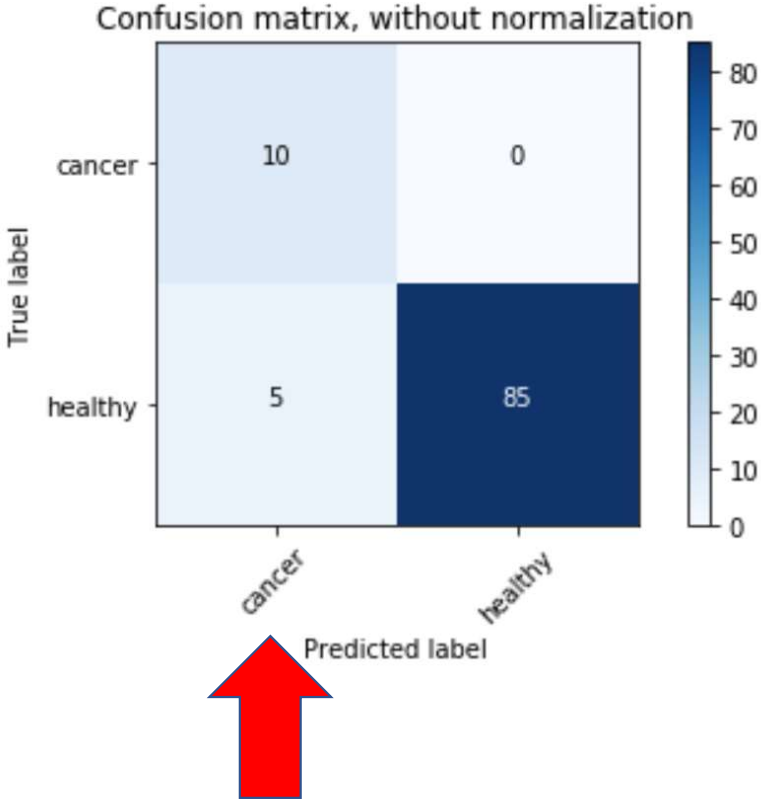

Simulated protein A and B co-expression for healthy (yellow) and cancer (purple) patients (n=100)
Neighbors = 11
Train score = 0.91, Test score = 0.96
class 0
class 1
Feature 1
Feature 0
1
protein A
Test set
Confusion matrix
True label
cancer
healthy
12
1
0
12
cancer
healthy

LEFT: new validation data (n=116)
RIGHT: training data (n=100)



What about unbalanced data?





“What’s the problem you’re trying to solve?”

-Andre Dekker, June 6, 2019

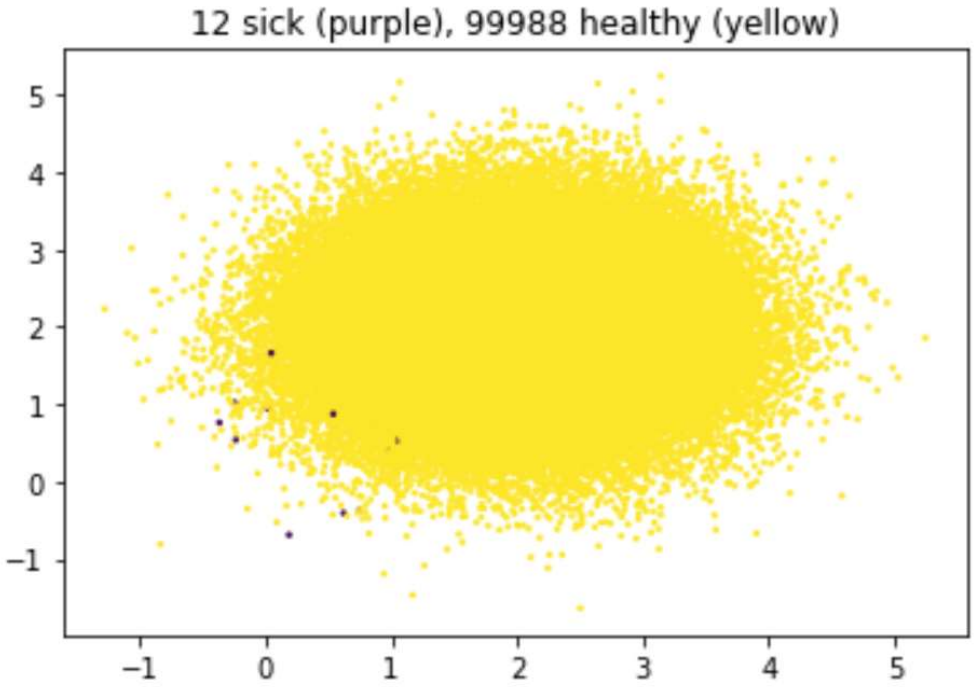
NATIONAL CANCER INSTITUTE

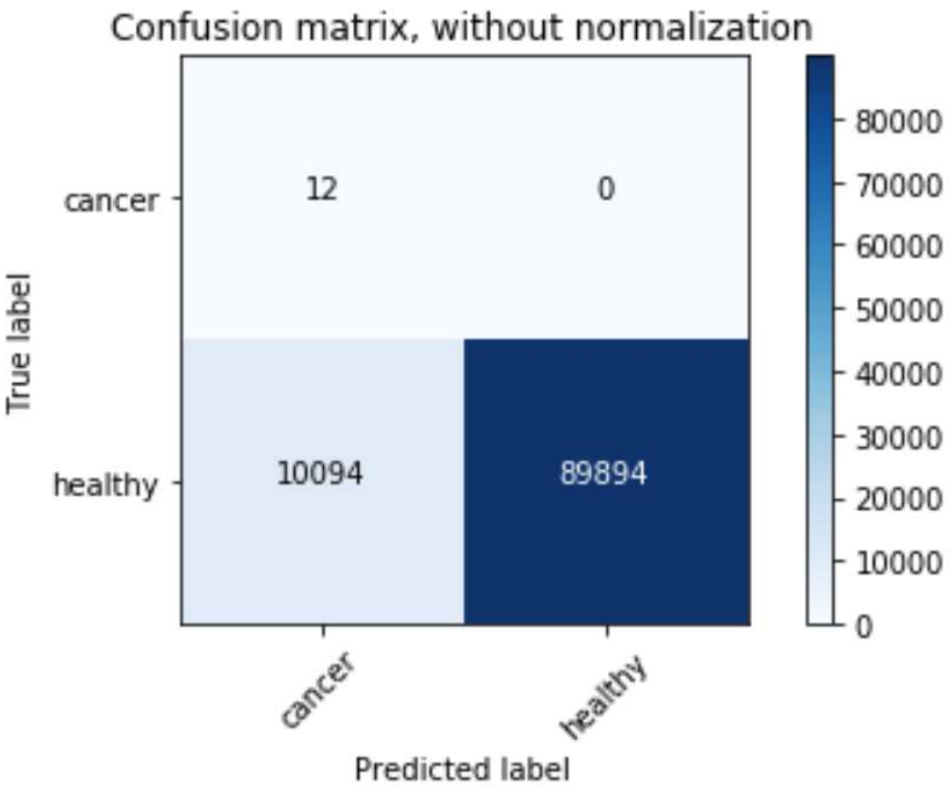
Surveillance, Epidemiology, and End Results Program

Number of New Cases and Deaths per 100,000: The number of new cases of ovarian cancer was 11.4 per 100,000 women per year. The number of deaths was 7.0 per 100,000 women per year. These rates are age-adjusted and based on 2012-2016 cases and deaths.

This is the problem that needs to be solved

If we generated a ROC curve, it would have looked very good! Also accuracy = 90%





Do the results justify the authors' interpretation?

Interpretation These findings justify a prospective population-based assessment of proteomic pattern technology as a screening tool for all stages of ovarian cancer in high-risk and general populations.

	Prevalence					
	50%	43%	10%	1%	0·1%	One per 2500
Specificity (%)						
90	91	88	53	9	1	0·4
95	95	94*	69	17	2	0·8†
99	99	99	92	50	9	4
99·9	99·9	99·9	99	91	50	29

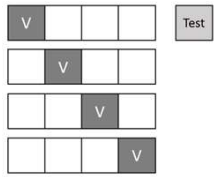
Assumes a constant sensitivity of 100%. *Value reported by Petricoin and colleagues.[1] †Corresponding value for low-risk clinical setting.

Positive predictive value at specificity and prevalence

Outline

RNG

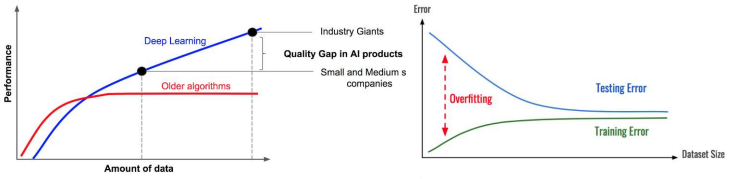
- Cross validation



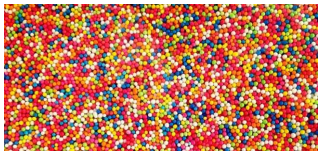
- Bias control

- Missing data

- Limited size

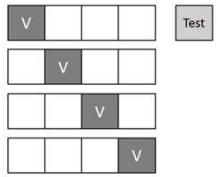


- Data heterogeneity



Outline

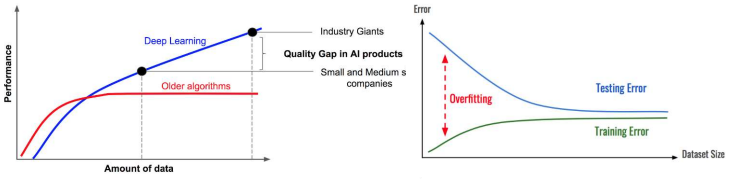
- Cross validation



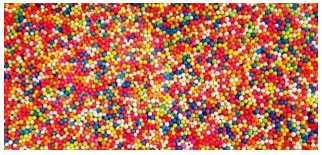
- Bias control

- **Missing data**

- Limited size



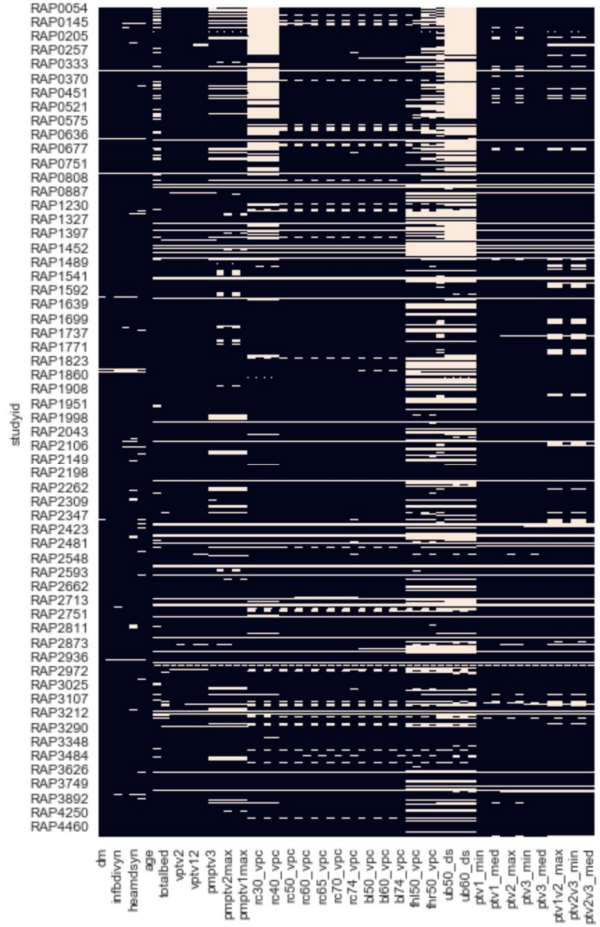
- Data heterogeneity



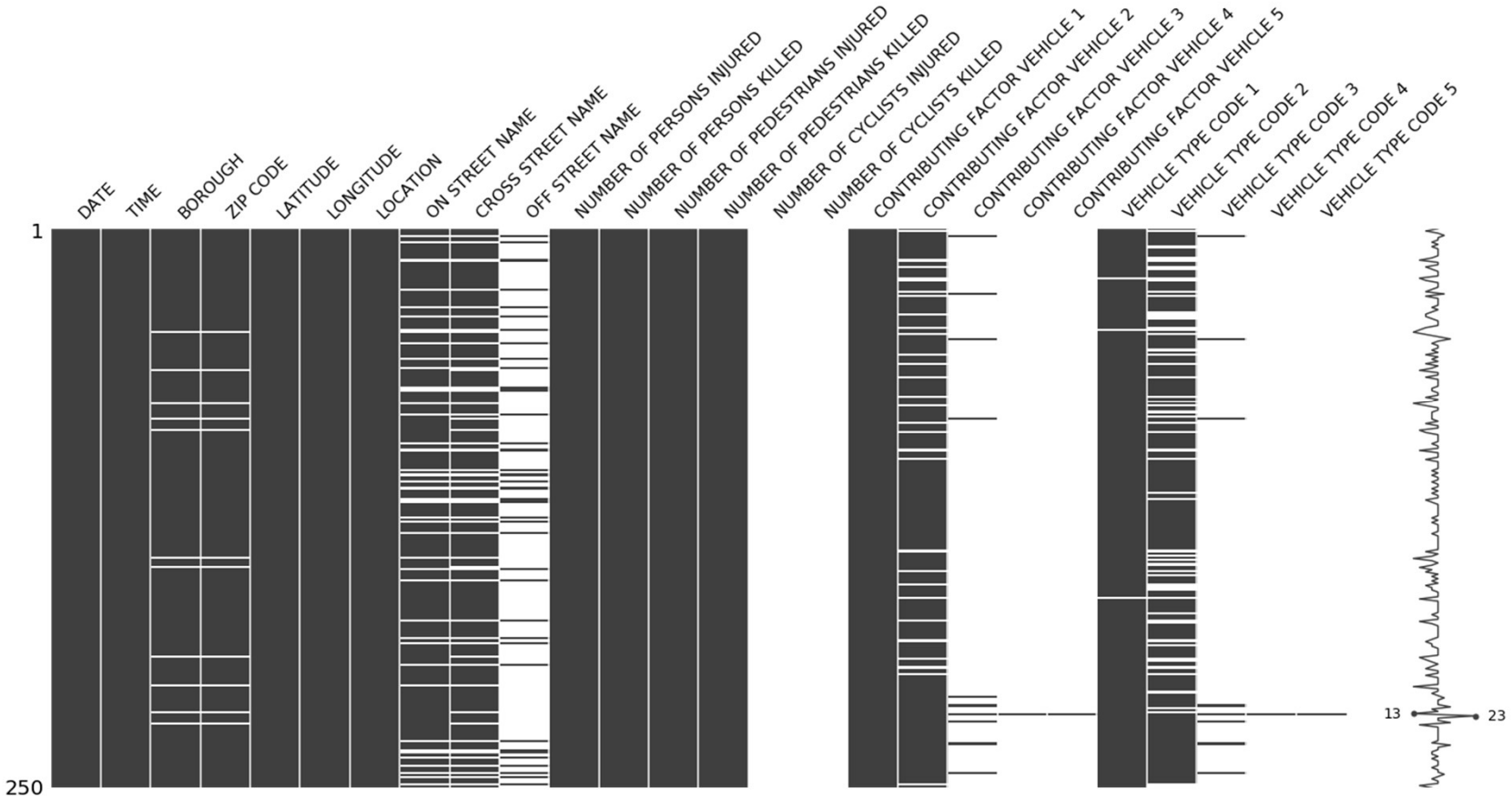
“Handling missing data is more of an art than a science”

-Robert Strawderman, Chair of Biostatistics, Univ. of Rochester

- Imputation: adds in missing information (adds bias)
- Remove data: columns OR samples (adds bias)
- WHY is the data missing
  - Often data “exists” but wasn’t captured
  - If missing at random, then easier to handle
  - Sometimes the missinginess contains information
  - If systematic, then can introduce bias (follow up missing because patient was admitted)
- Imputation solutions https://pypi.org/project/missingpy/
  - median
  - kNN
  - Random Forest



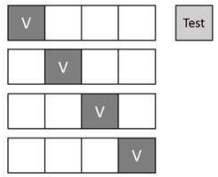
Visualizing missing data



https://github.com/ResidentMario/missingno

Outline

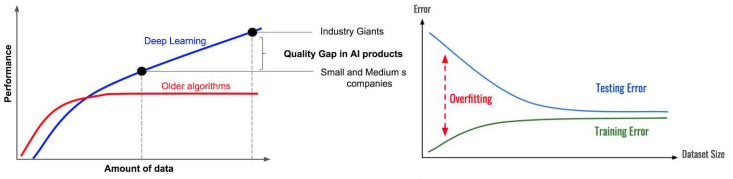
- Cross validation



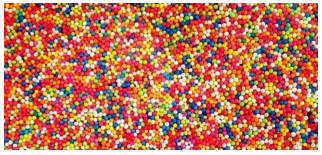
- Bias control

- Missing data

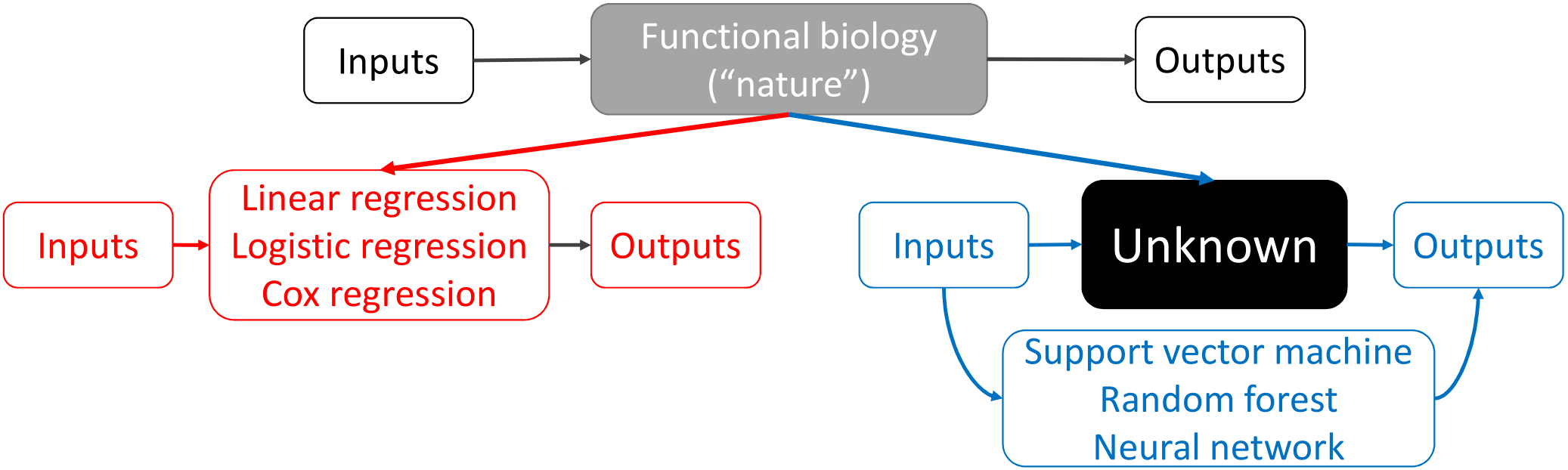
- **Limited size**



- Data heterogeneity

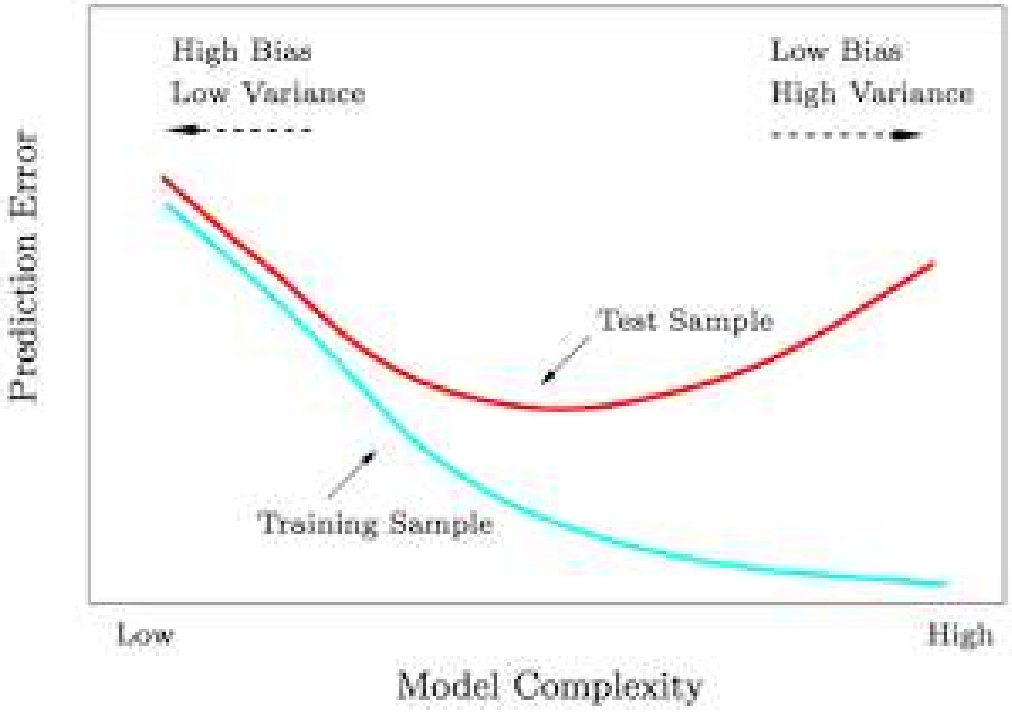


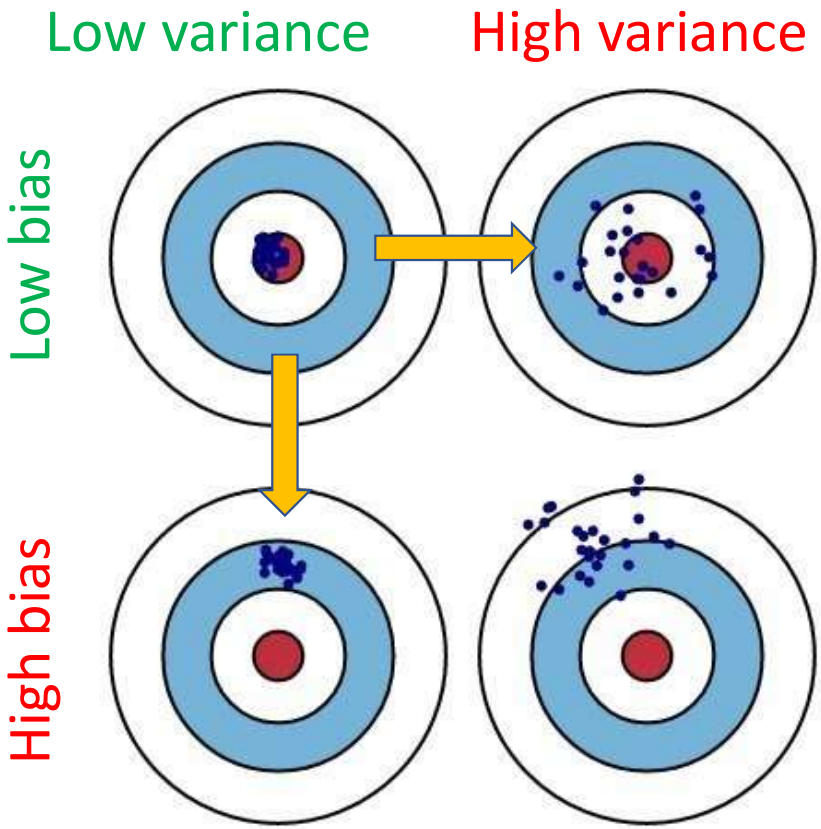
How to model nature?



Statistics: **Assume** model, **test** how well data fits, **infer** model properties

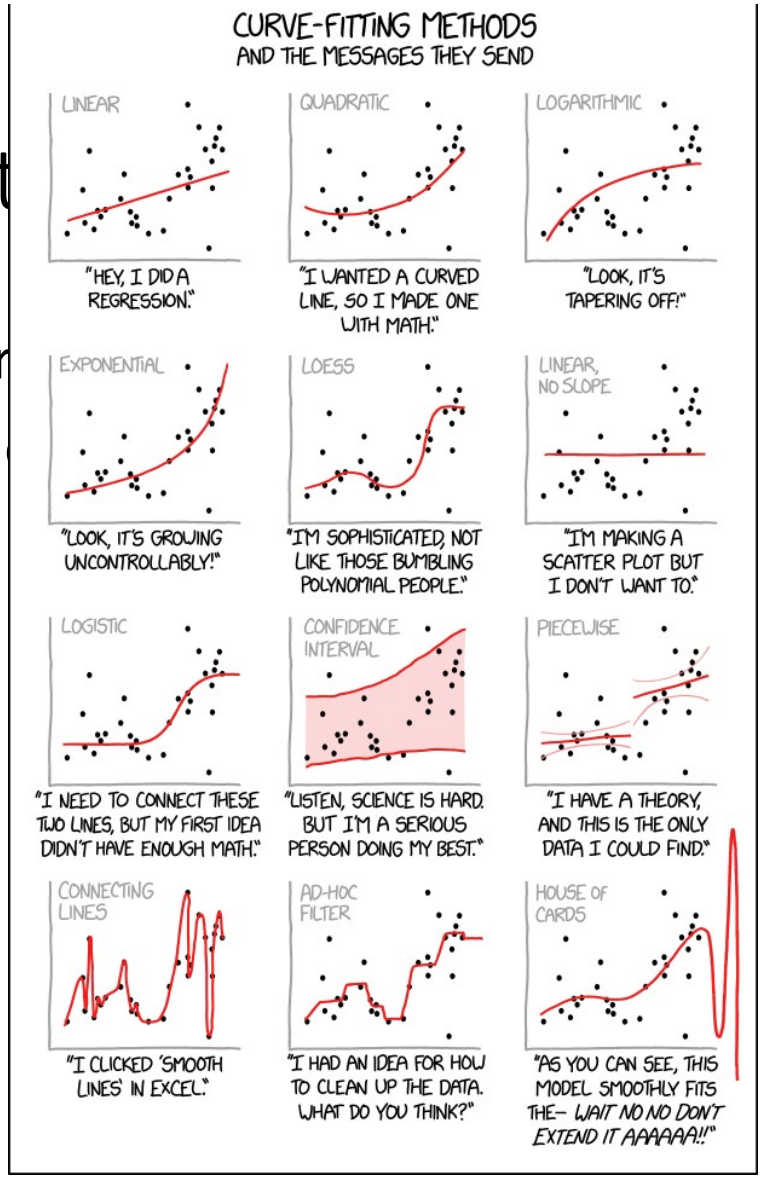
ML: **Ignore** black box and find model(s) **with best performance**


High Bias
Low Variance
Low Bias
High Variance
Test Sample
Training Sample
Prediction Error
Low
High
Model Complexity


Low variance
High variance
Low bias
High bias

Taking a stat

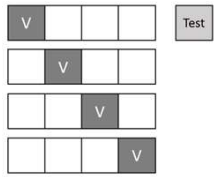
- If assumptions ar
- May be able to d
- But if not….



https://xkcd.com/2048/

Outline

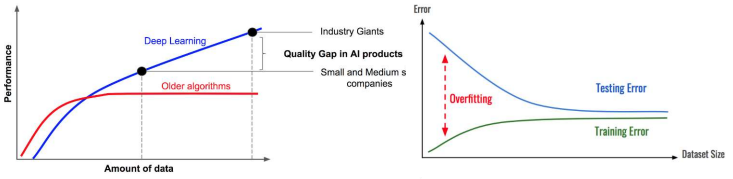
- Cross validation



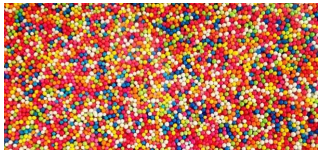
- Bias control

- Missing data

- Limited size



- **Data heterogeneity**



Data heterogeneity: feature engineering

- Domain expertise to determine which variables to keep or drop
- Scaling or normalization
  - Support vector machine: min/max scale to [-1, 1]
  - Decision tree: do not need to normalize/scale
- Unsupervised learning to combine (reduce) similar variables
- Unstructured data can be appealing but also challenging