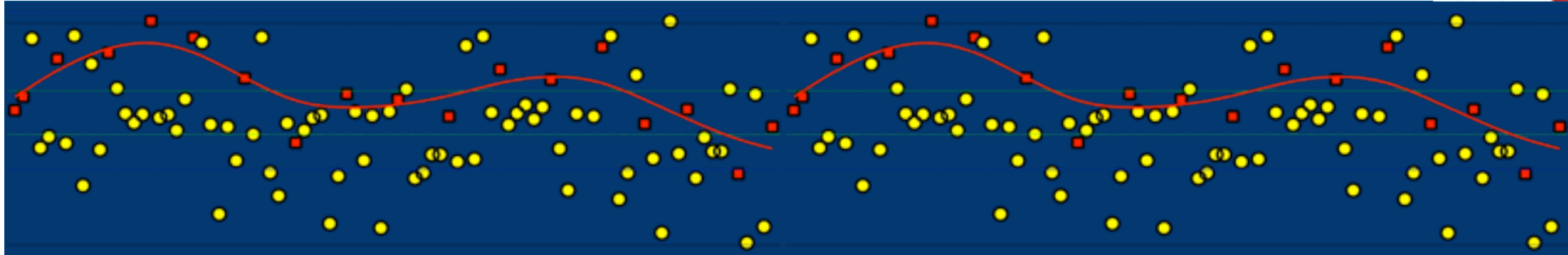




Centre for Integrative Metabolomics
& Computational Biology



Is Metabolomics ready for the return of Artificial Neural Networks?

Prof. David Broadhurst

Director, Centre for Integrative Metabolomics & Computational Biology

Professor of Computational Systems Biology | School of Science

Edith Cowan University

Background: Goodacre & Kell 1992-1998

1992

Neural networks and olive oil

See — Virgin olive oil is the oil extracted by purely mechanical means from sound, ripe fruits of the olive tree (*Olea europaea* L.). Such oils with a free fatty acid content below 1% are termed 'extra virgin', whereas those with good flavour but greater acidity may be graded as 'fine' or 'semi-fine'. Lower grades, including those that have been subjected to refining, are called 'lampante' or 'pure'. Olive oil is considered to contribute significantly to the nutritional and health benefits of Mediterranean-type diets and, uniquely among vegetable oils, the flavour of olive oil is best enjoyed without refining. Olive oil therefore commands a higher price than other vegetable oils, and these and other properties mean that there is a great temptation to adulterate olive oils with other seed oils¹. Although various methods have been proposed for the detection of olive oil adulteration², none has found widespread usage. We wish to report here that a combination of Curie-point pyrolysis mass spectrometry

We trained an artificial neural network consisting of an input layer of the 150 normalized ion intensities with mass charge in the range 51–200 and one hidden layer of eight nodes, using the standard back-propagation algorithm³ as implemented in the NeuralDesk package (Neural Computer Sciences, Totton, Southampton), coding virgin oils as 1, non-virgin as 0. When the network had trained (r.m.s. error <0.001), we tested it on the unknown. When the code was broken, it transpired that the network had correctly assessed each oil. In a typical run, the virgins were assessed with a code of 0.99976 ± 0.00046 (range 0.99954 – 1.00016) and the non-virgins with a code of 0.00379 ± 0.00268 (range 0.00026 – 0.01099). We conclude that the combination of Curie-point PyMS and artificial neural networks constitutes a powerful approach to the assessment of olive oil adulteration.

Royston Goodacre
Douglas B. Kell¹
Department of Biological Sciences,
University of Wales,
Aberystwyth, Dyfed SY23 3DA, UK
Giorgio Bianchi
Istituto Sperimentale per la Elastotecnica
Contrada "Fonte Umana" n.27,
65023 Città S. Angelo,
Pescara, Italy

¹To whom correspondence should be addressed.

1. Kikvidze, A. & Moshkova, P. In *Essential Oils and Flavours* (London, W. F. & Jackson, J. F. 1–20) (Springer, Heidelberg, 1992).
2. Moshkova, P. S. C. *Heuristics* 1, 1–10 (1992).
3. Rumelhart, D. E., Hinton, G. E. & Williams, R. J. *Nat. Phys.* 1, 318–320 (1986).
4. Rumelhart, D. E. et al. *Artificial Distributed Processing* (MIT, Cambridge, 1986).

NATURE • VOL 359 • 15 OCTOBER 1992

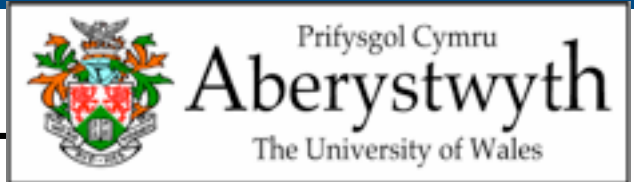
1996

Zbl. Bakt. 284, 516–539 (1996)
© Gustav Fischer Verlag, Stuttgart · Jena · New York

Quantitative Analysis of Multivariate Data Using Artificial Neural Networks: A Tutorial Review and Applications to the Deconvolution of Pyrolysis Mass Spectra

ROYSTON GOODACRE, MARK J. NEAL, and DOUGLAS B. KELL

Institute of Biological Sciences, University of Wales, Aberystwyth, Dyfed, UK



28 ANN papers

1998

Rapid identification of urinary tract infection bacteria using hyperspectral whole-organism fingerprinting and artificial neural networks

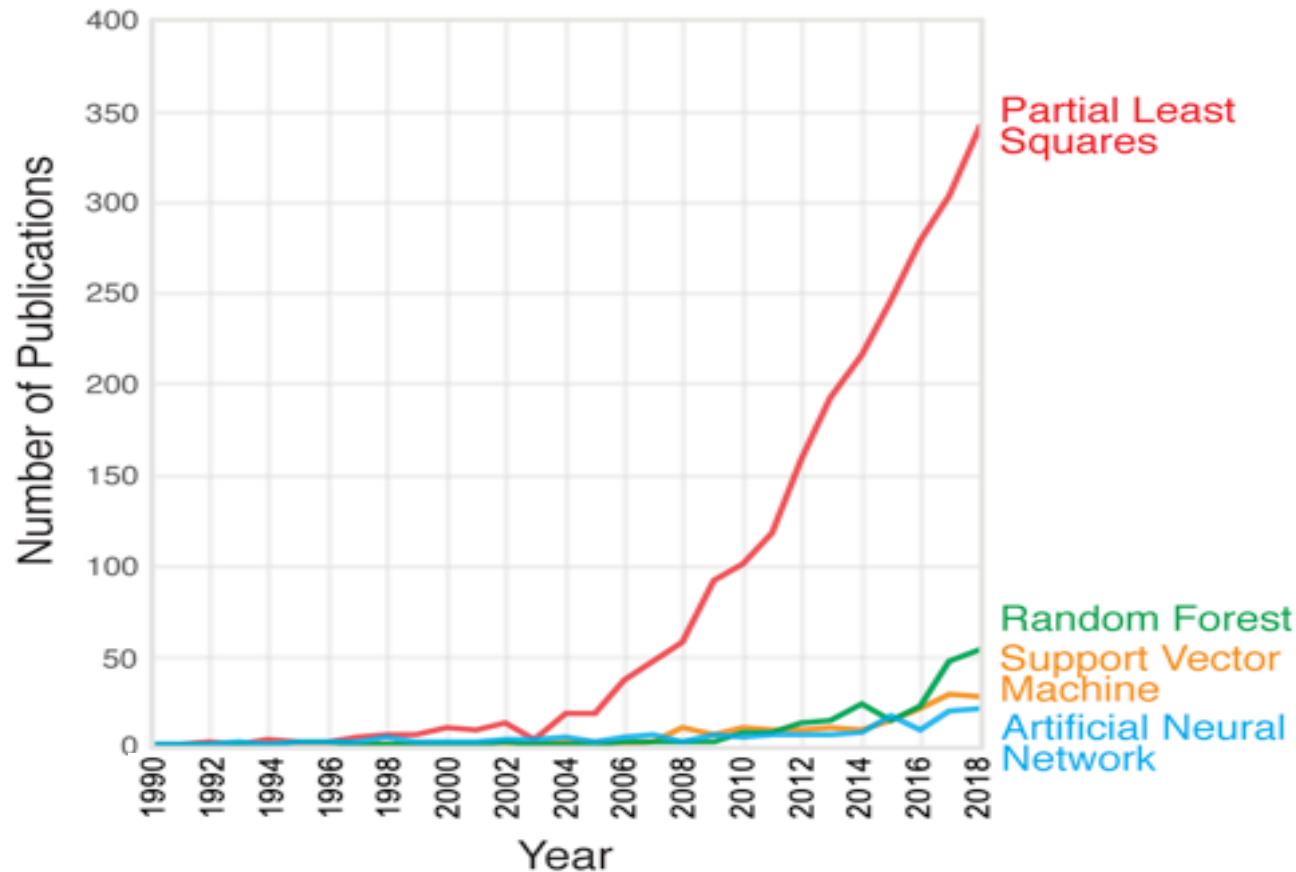
Royston Goodacre,¹ Éadaoin M. Timmins,¹ Rebecca Burton,¹ Naheed Kaderbhai,¹ Andrew M. Woodward,¹ Douglas B. Kell¹ and Paul J. Rooney²

Microbiology (1998), 144, 1157–1170

1994-1998

'Genetic Algorithms, Artificial Neural Networks and their Application to Chemometrics'.

Rise of PLS-DA

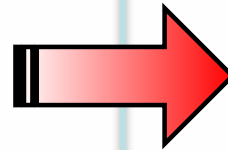


Number of publications per year
(Web of Science) including the
key term **metabolite***, **metabolom***
or **metabonom***



Fall & Rise of ANN

1998

- Black Box
- Computer power
- Access to software
- Overtraining?
- Suitable data sets?



2019

- Societal acceptance
- Cloud Computing
- Free Code  Keras 
- Better Understanding?
- Suitable data sets?



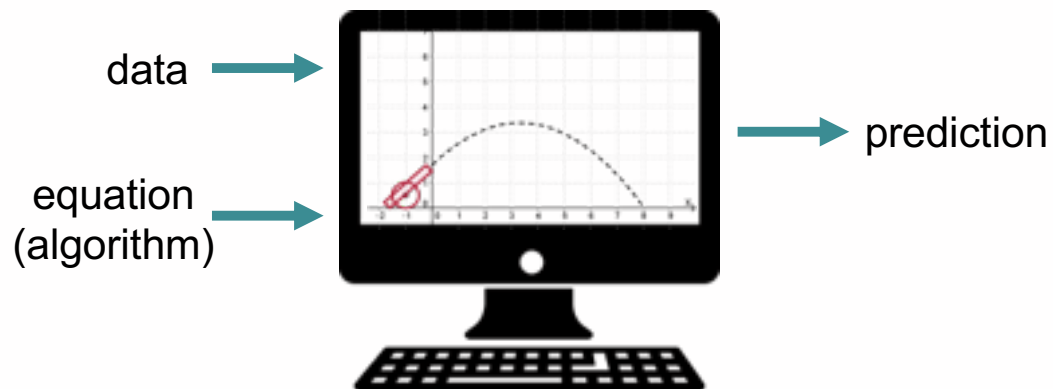
Centre for Integrative Metabolomics
& Computational Biology



Background Theory

Machine Learning: Data Driven

Traditional Programming



e.g.
$$y = x \tan \theta - \frac{x^2 g}{2 (v_i \tan \theta)^2}$$

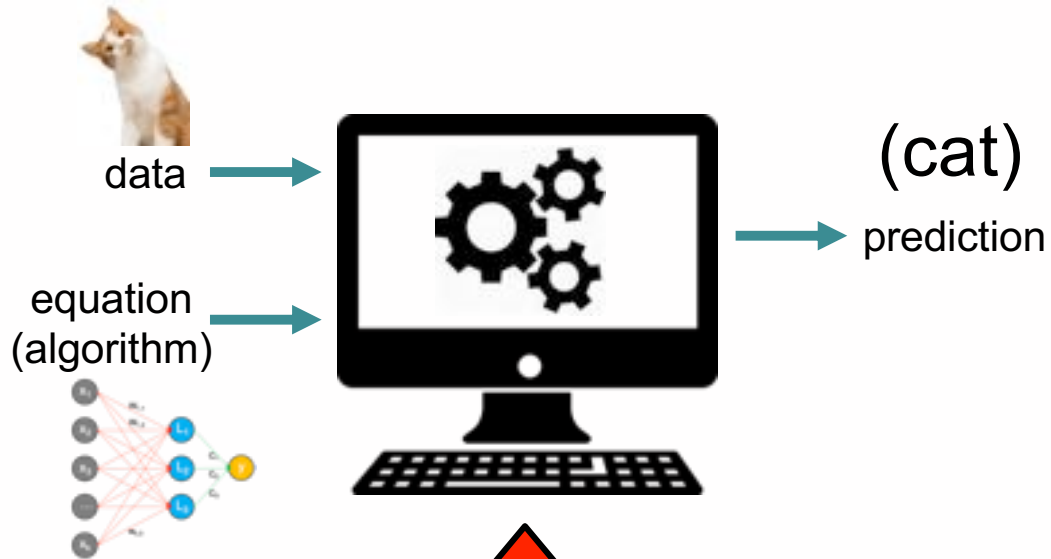
Machine Learning



e.g. Image Classification

Machine Learning: Data Driven

Traditional Programming



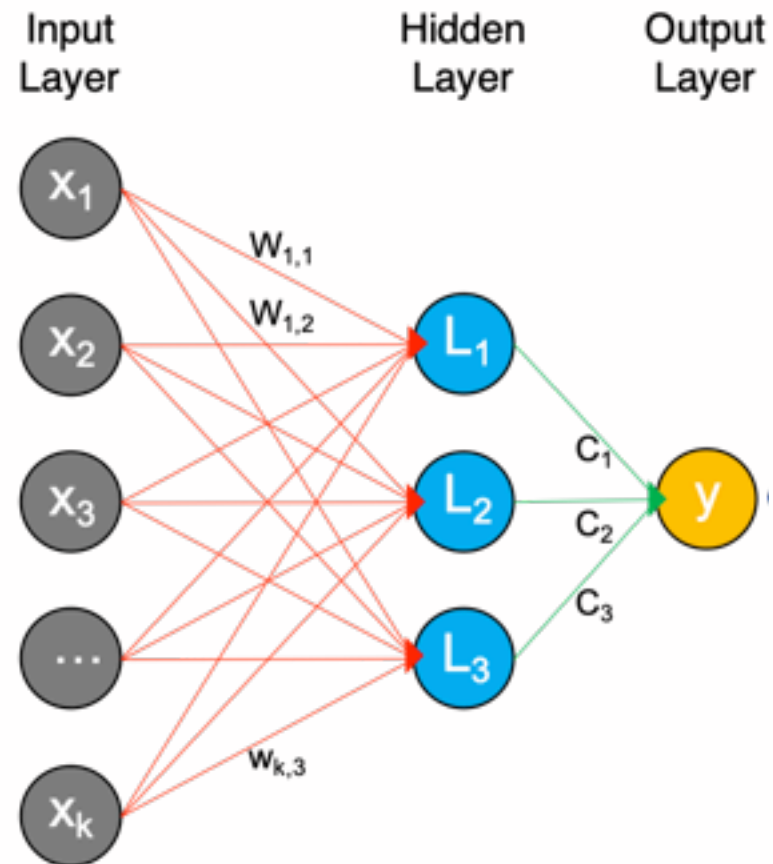
Machine Learning



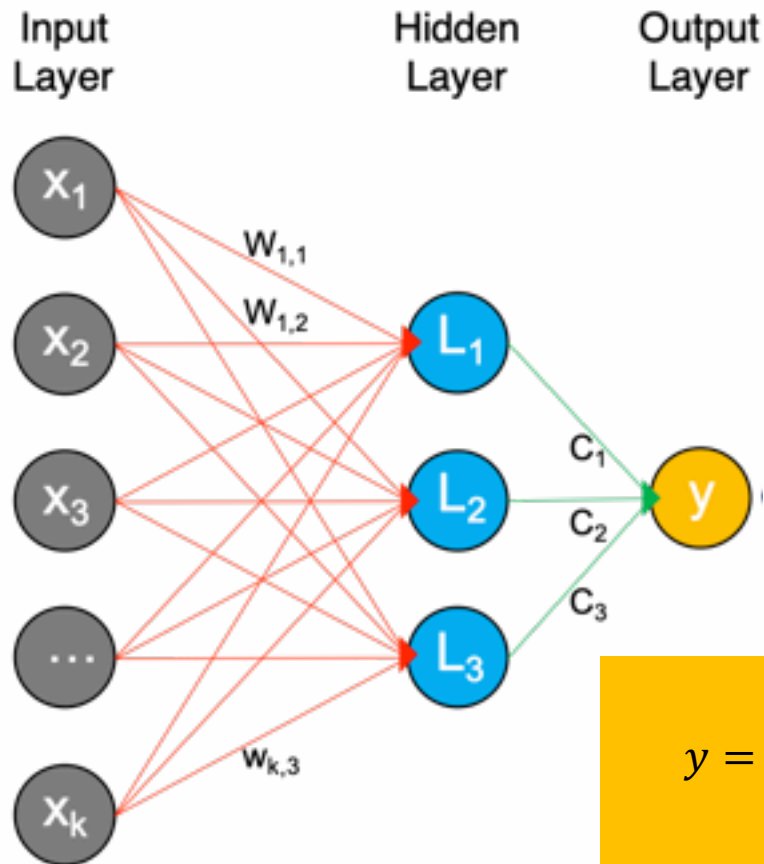
deployment



Artificial Neural Networks



Artificial Neural Networks

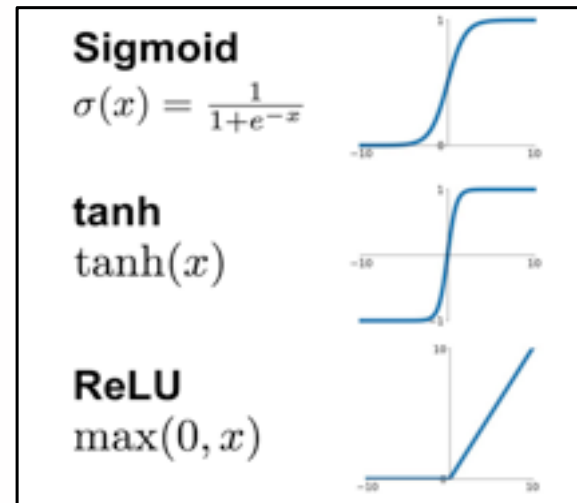
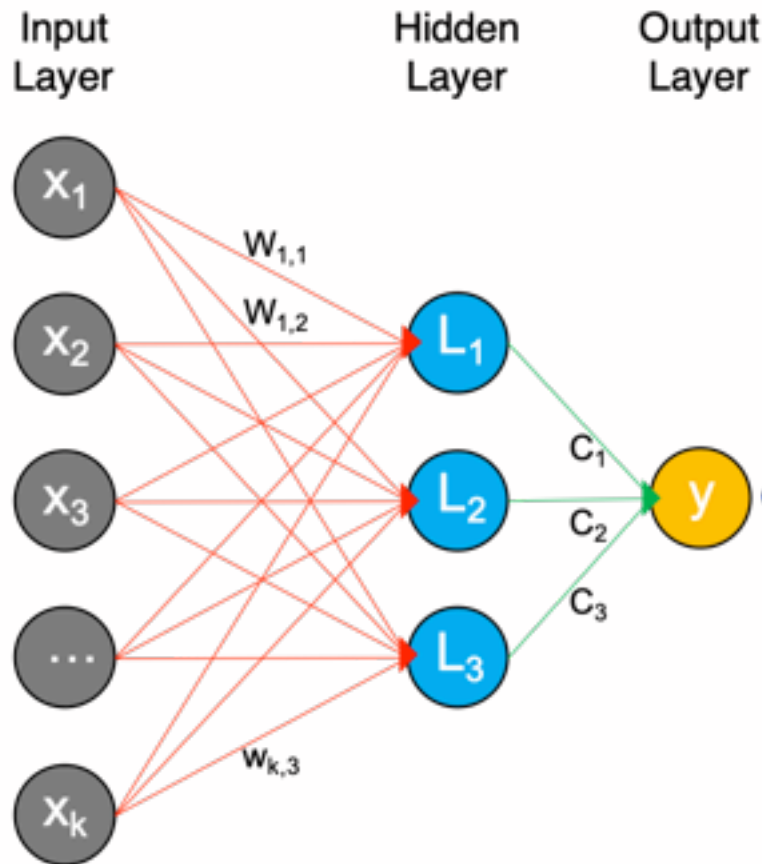


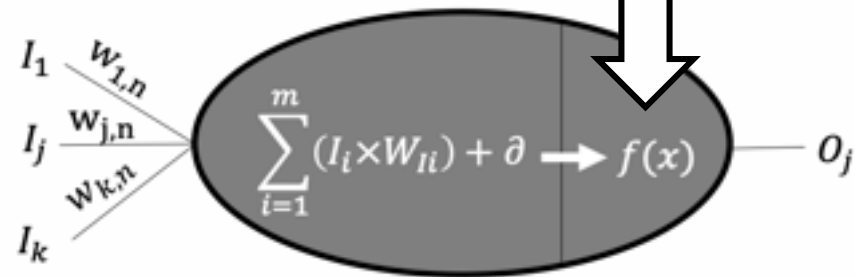
$$y = f_o \left(\sum_{j=1}^3 c_j \times L_j \right)$$

$$L_j = f_j \left(\sum_{i=1}^k w_{i,j} \times x_i \right)$$

$$y = f_o \left(c_1 f_1 \left(\sum_{i=1}^k w_{i,1} \times x_i \right) + c_2 f_2 \left(\sum_{i=1}^k w_{i,2} \times x_i \right) + c_3 f_3 \left(\sum_{i=1}^k w_{i,3} \times x_i \right) \right)$$

Artificial Neural Networks

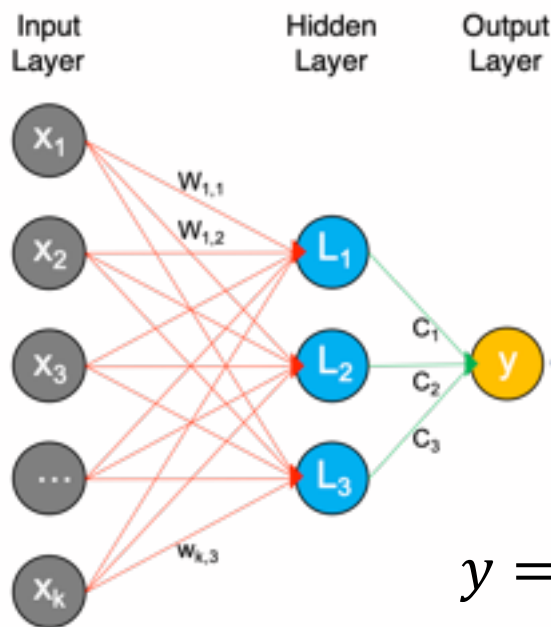




The diagram shows a single neuron model where multiple inputs I_1, I_j, I_k are weighted by $w_{1,n}, w_{j,n}, w_{k,n}$ and summed to produce the output O_j .

$$\sum_{i=1}^m (I_i \times W_{in}) + \theta \rightarrow f(x) \rightarrow O_j$$

Artificial Neural Networks

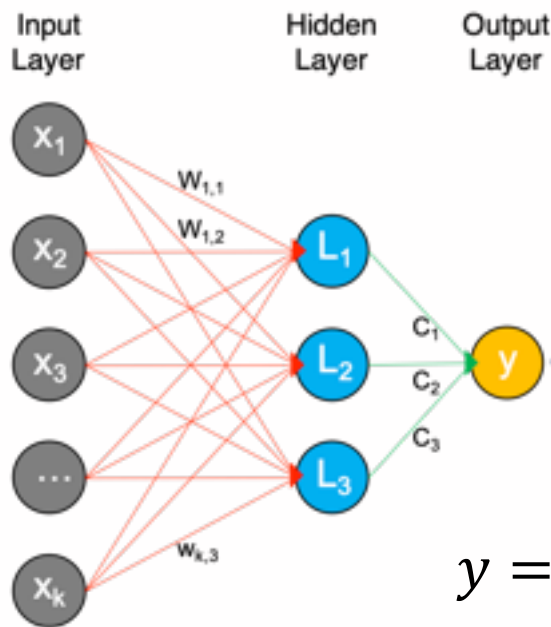


$$y = f_o \left(c_1 f_1 \left(\sum_{i=1}^k w_{i,1} \times x_i \right) + c_2 f_2 \left(\sum_{i=1}^k w_{i,2} \times x_i \right) + c_3 f_3 \left(\sum_{i=1}^k w_{i,3} \times x_i \right) \right)$$

$$f_o = \frac{1}{1 + e^{-(z)}} \quad f_j = \frac{1}{1 + e^{-(z)}}$$

$$y = \frac{1}{1 + e^{-\left(\frac{c_1}{1 + e^{-(\sum_{i=1}^k w_{i,1} \times x_i)}} + \frac{c_2}{1 + e^{-(\sum_{i=1}^k w_{i,2} \times x_i)}} + \frac{c_3}{1 + e^{-(\sum_{i=1}^k w_{i,3} \times x_i)}} \right)}}$$

Artificial Neural Networks



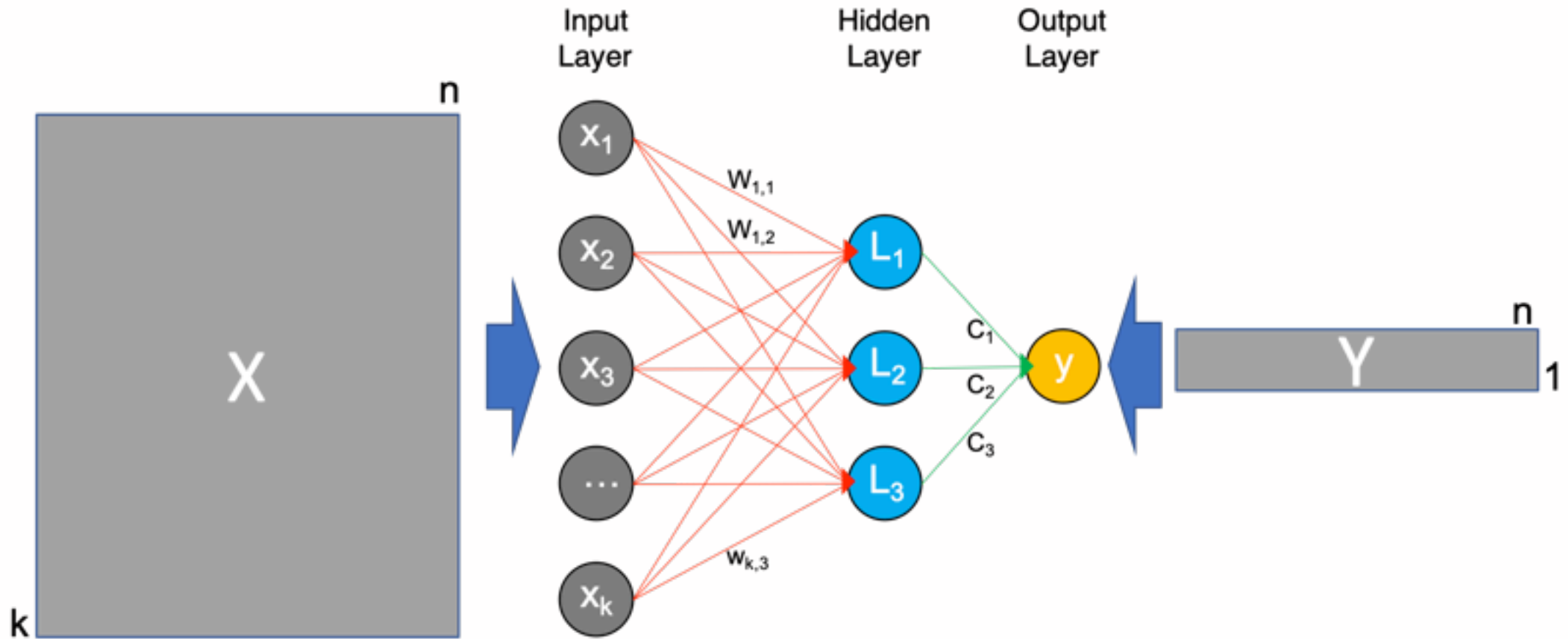
$y =$

$$y = f_o \left(f_1 \left(\sum_{i=1}^k w_{i,1} \times x_i \right) + f_2 \left(\sum_{i=1}^k w_{i,2} \times x_i \right) + f_3 \left(\sum_{i=1}^k w_{i,3} \times x_i \right) \right)$$

$X \rightarrow$ complicated nonlinear equation $\rightarrow y$

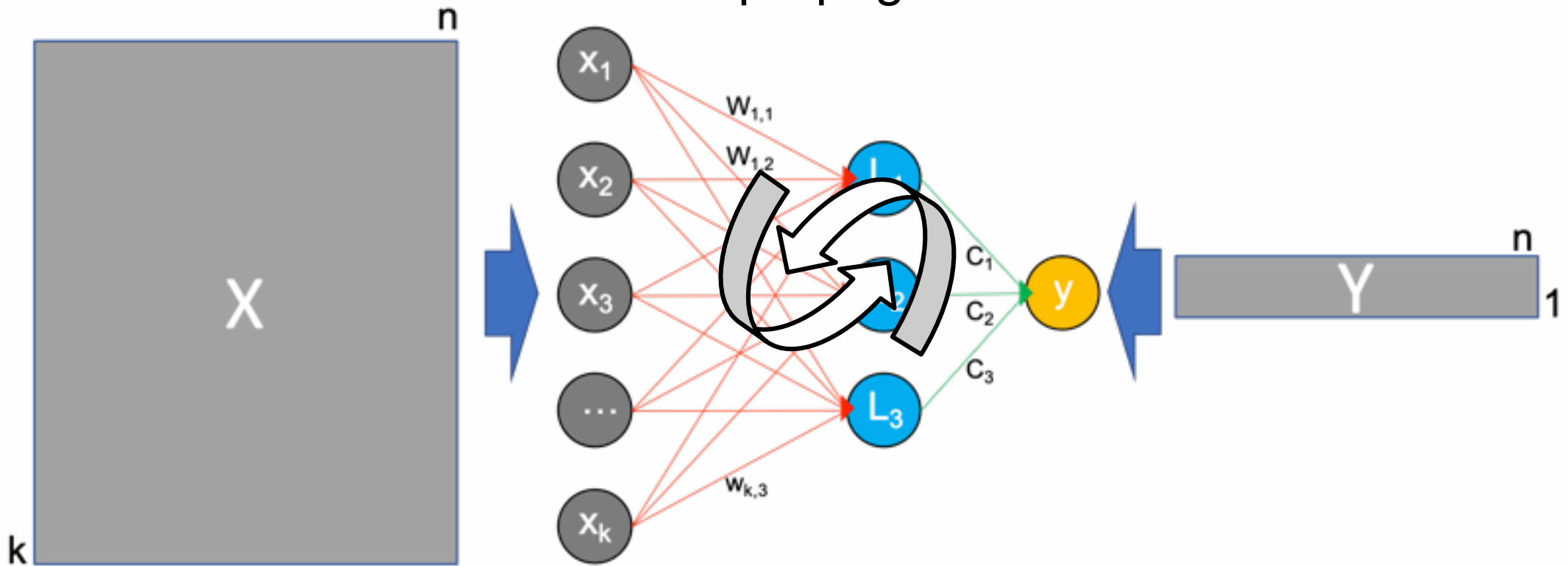
$$y = \frac{f_1 \left(\sum_{i=1}^k w_{i,1} \times x_i \right) + f_2 \left(\sum_{i=1}^k w_{i,2} \times x_i \right) + f_3 \left(\sum_{i=1}^k w_{i,3} \times x_i \right)}{1 + e^{-\left(f_1 \left(\sum_{i=1}^k w_{i,1} \times x_i \right) + f_2 \left(\sum_{i=1}^k w_{i,2} \times x_i \right) + f_3 \left(\sum_{i=1}^k w_{i,3} \times x_i \right) \right)}}$$

Artificial Neural Networks



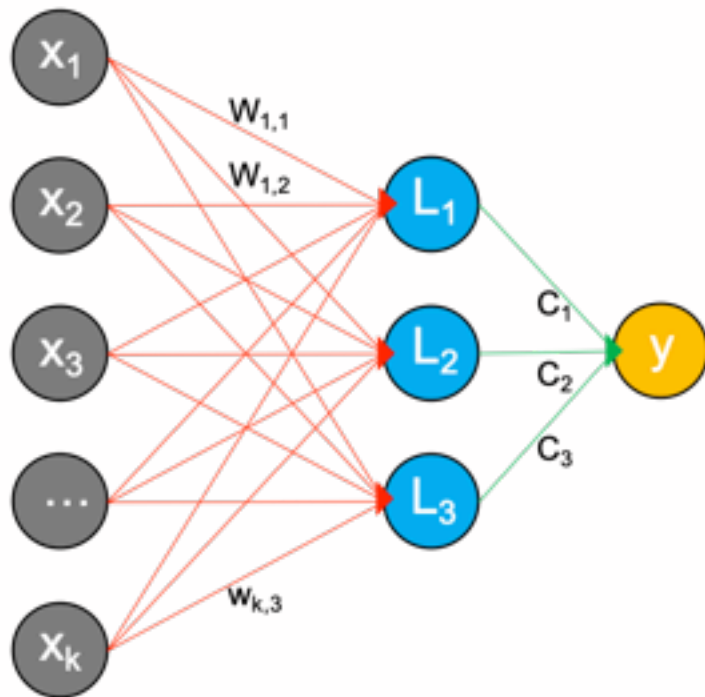
Artificial Neural Networks

Backpropagation

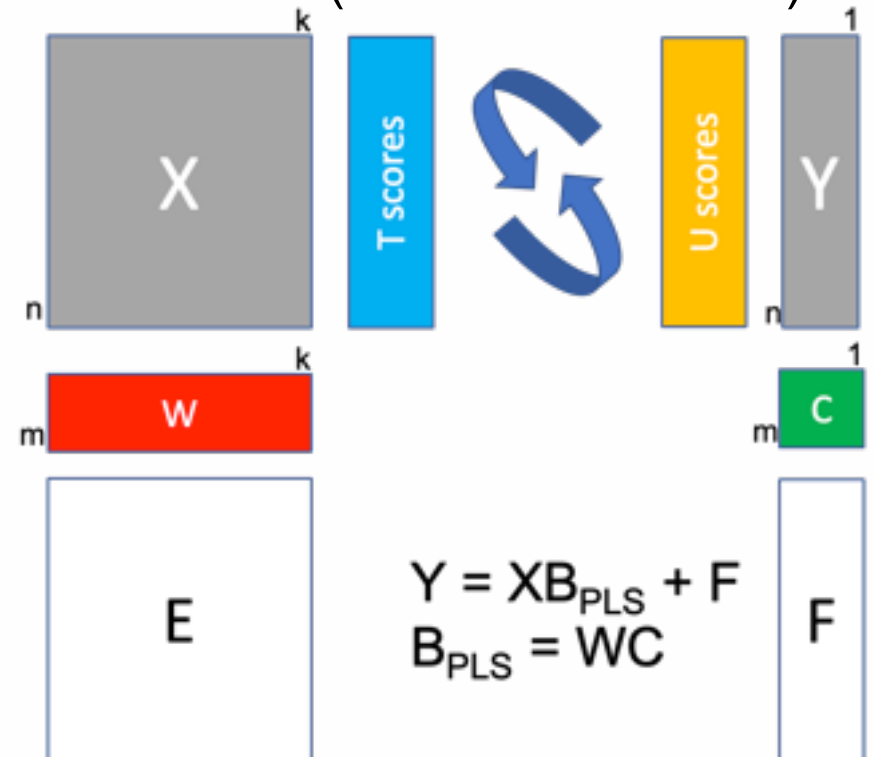


Structural Equivalence

FF-Lin-ANN (Backprop)



PLS (NIPALS / SIMPLS)





Centre for Integrative Metabolomics
& Computational Biology

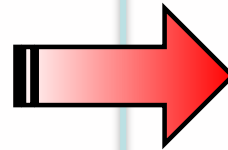


Current Issues


Fall & Rise of ANN

1998

- Black Box
- Computer power
- Access to software
- Overtraining?
- Suitable data sets?



2019

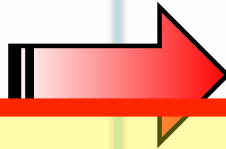
- Societal acceptance
- Cloud Computing
- Free Code  Keras
- Better Understanding?
- Suitable data sets?



Fall & Rise of ANN

1998

- Black Box
- Computer power
- Access to software



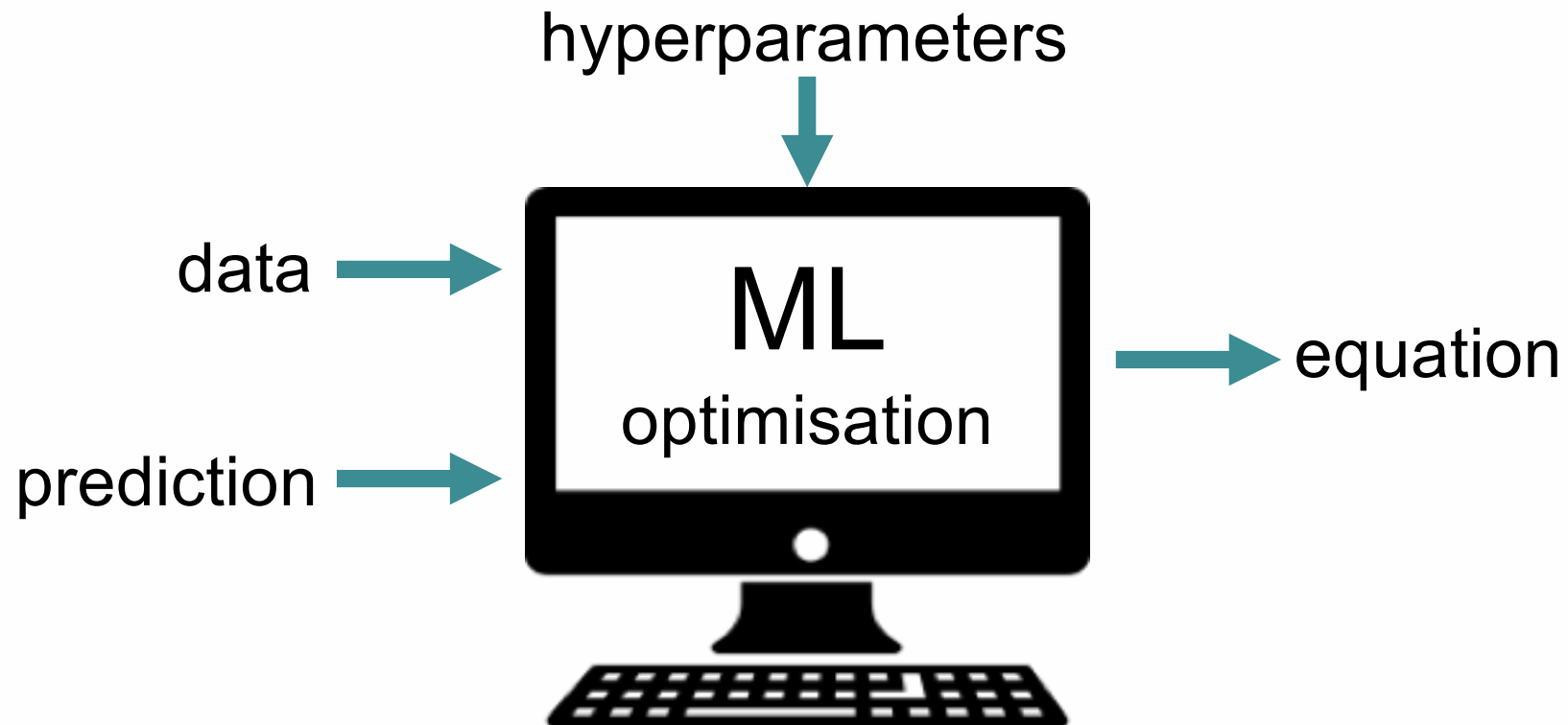
2019

- Societal acceptance
- Cloud Computing
- Free Code  Keras 

- Overtraining?
- Suitable data sets?

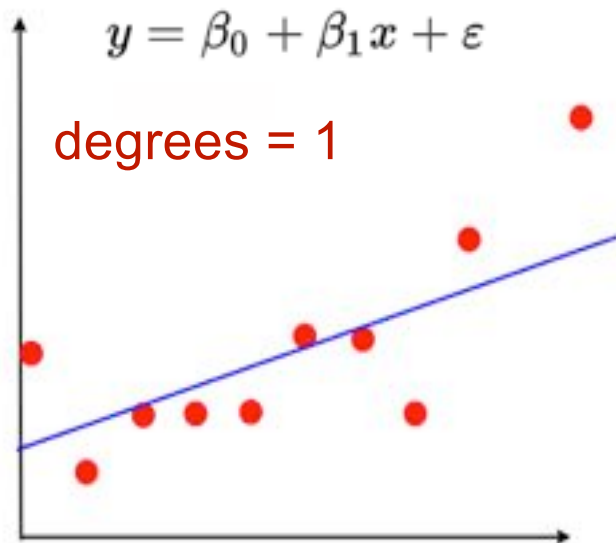
- Better Understanding?
- Suitable data sets?

Hyperparameters (structural / optimisation)

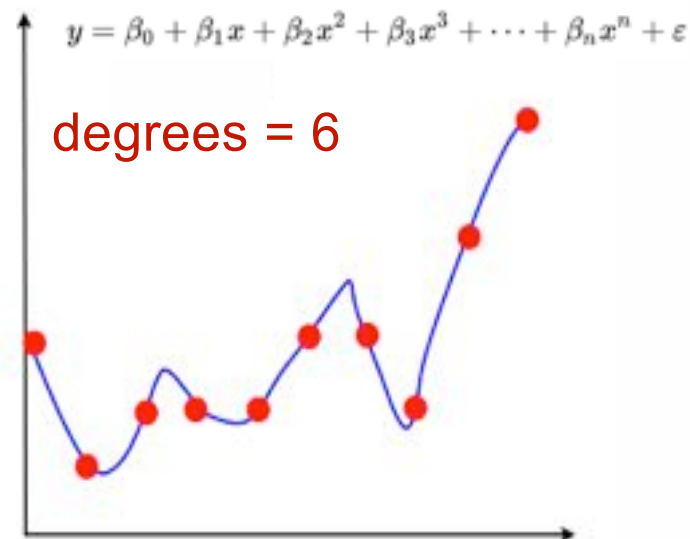


Hyperparameters

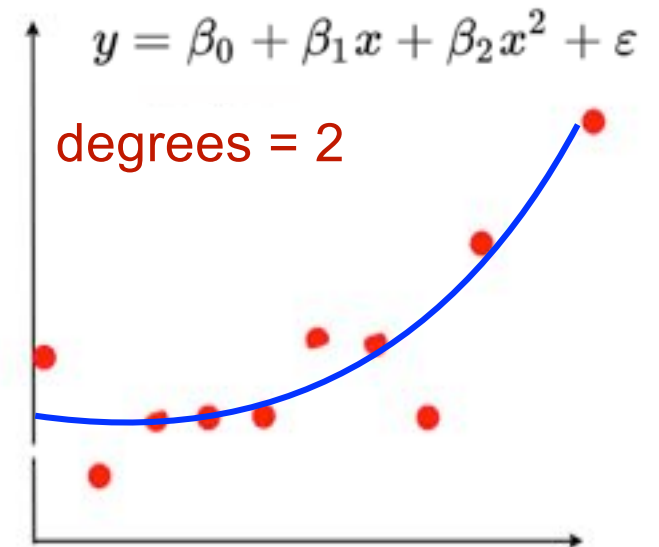
Polynomial regression: hyperparameter = number of degrees



Low Bias
High Variance



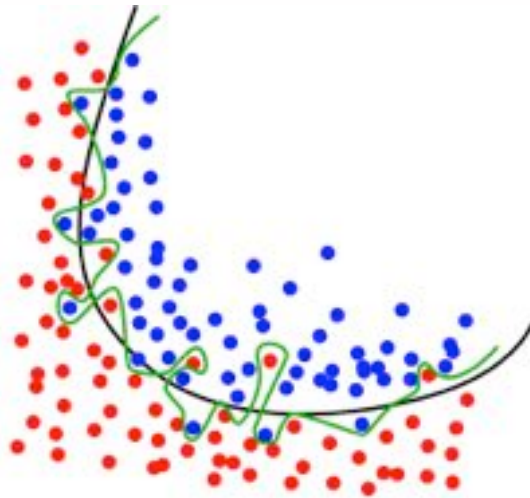
High Bias
Low Variance



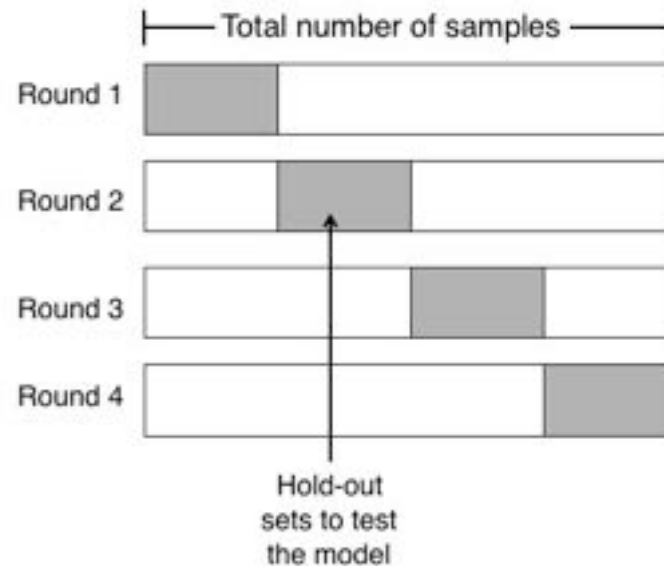
Optimal Bias
Optimal Variance

K-fold Cross Validation

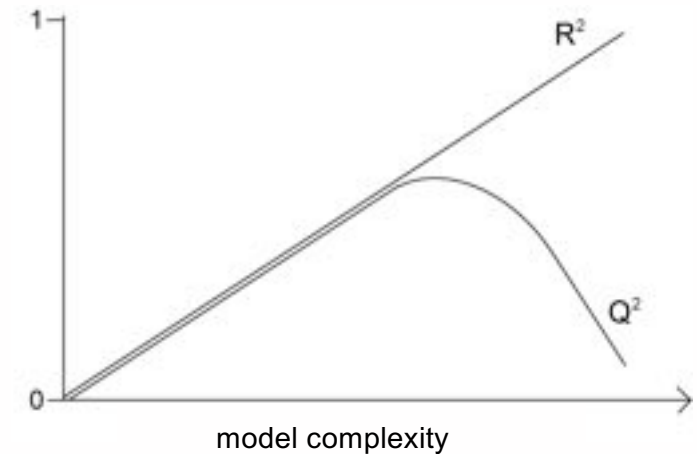
Avoid
Overfitting



K-fold CV



R^2/Q^2



$$R^2 = 1 - RSS/TSS$$

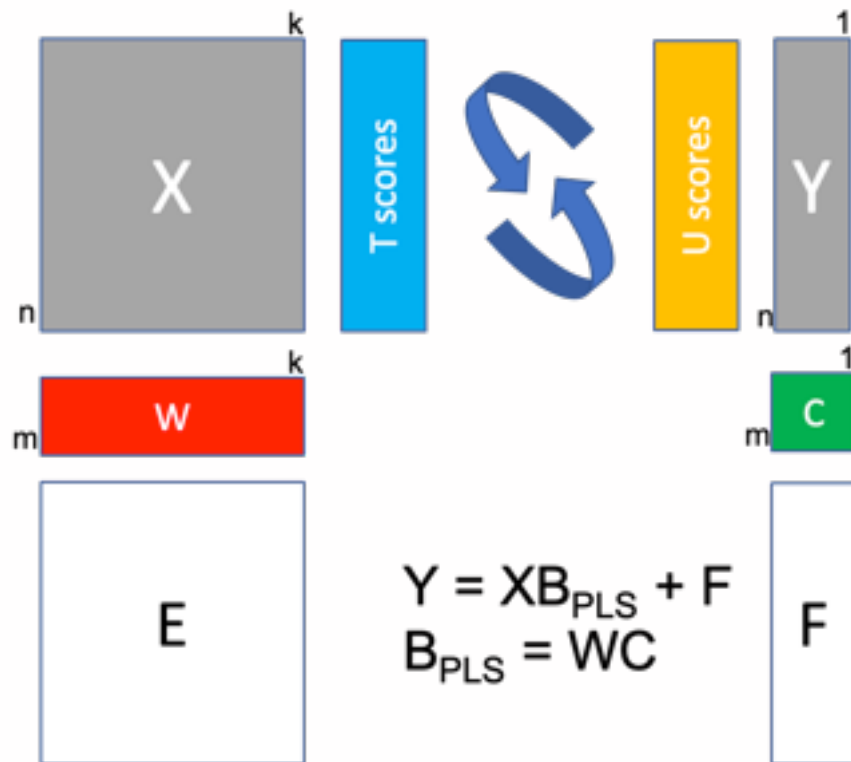
$$RSS = \sum (y - \hat{y})^2$$

$$TSS = \sum (y - \bar{y})^2$$

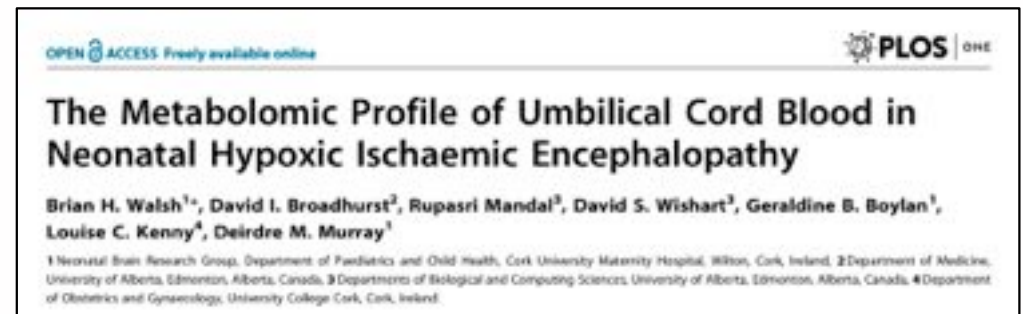
$$Q^2 = 1 - PRESS/TSS$$

$$PRESS = \sum (y - \hat{y})^2$$

PLS-DA (one hyperparameter)



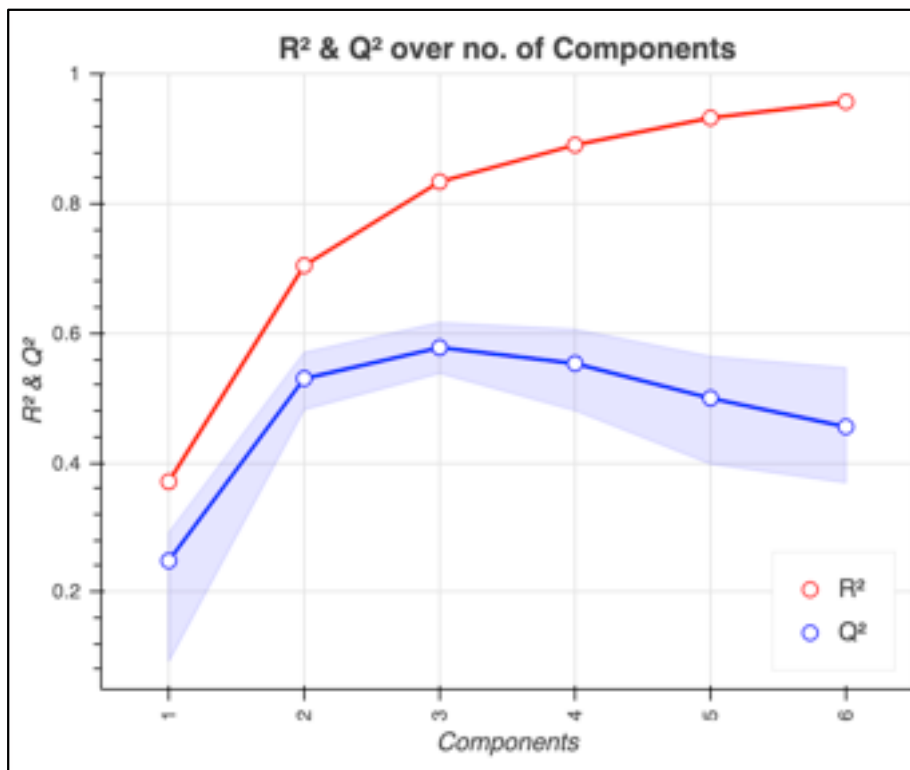
Example



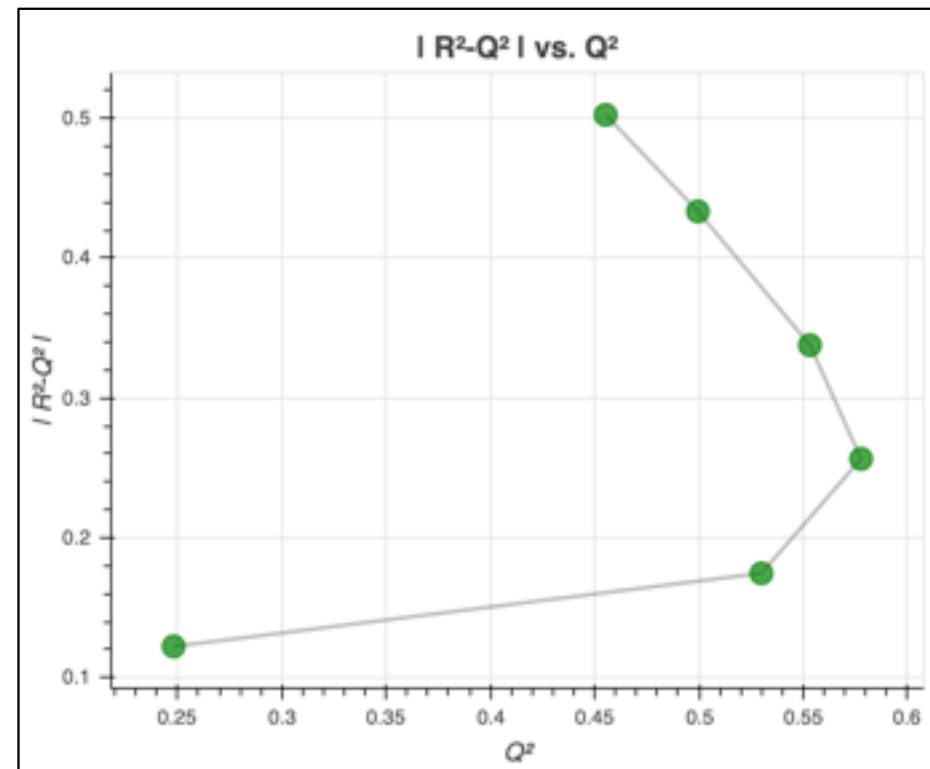
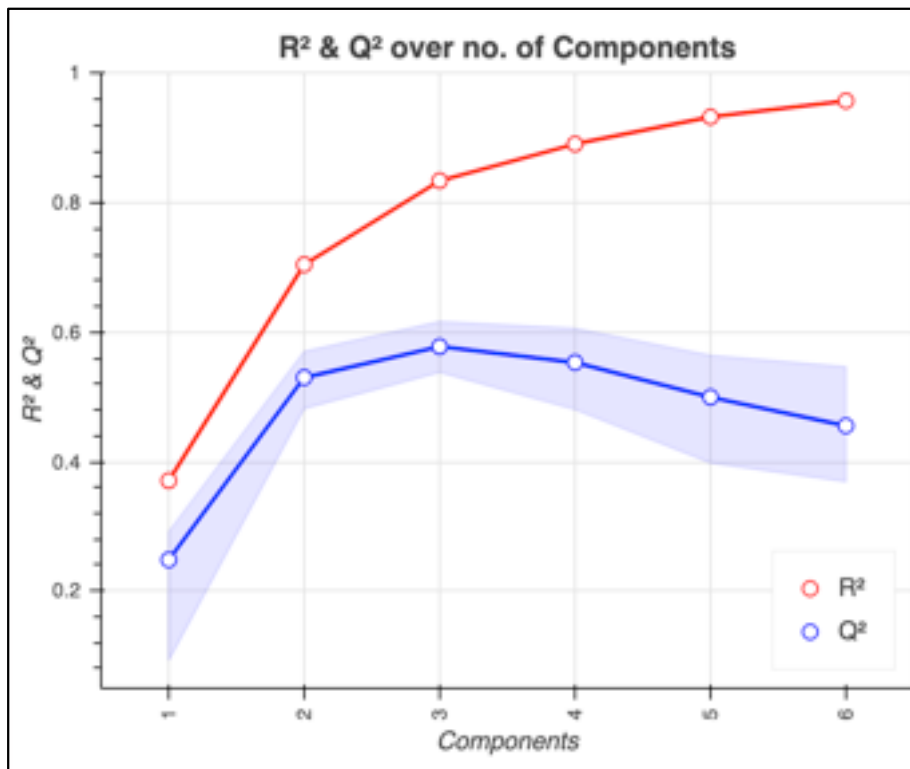
147 metabolites (25 case 25 control)



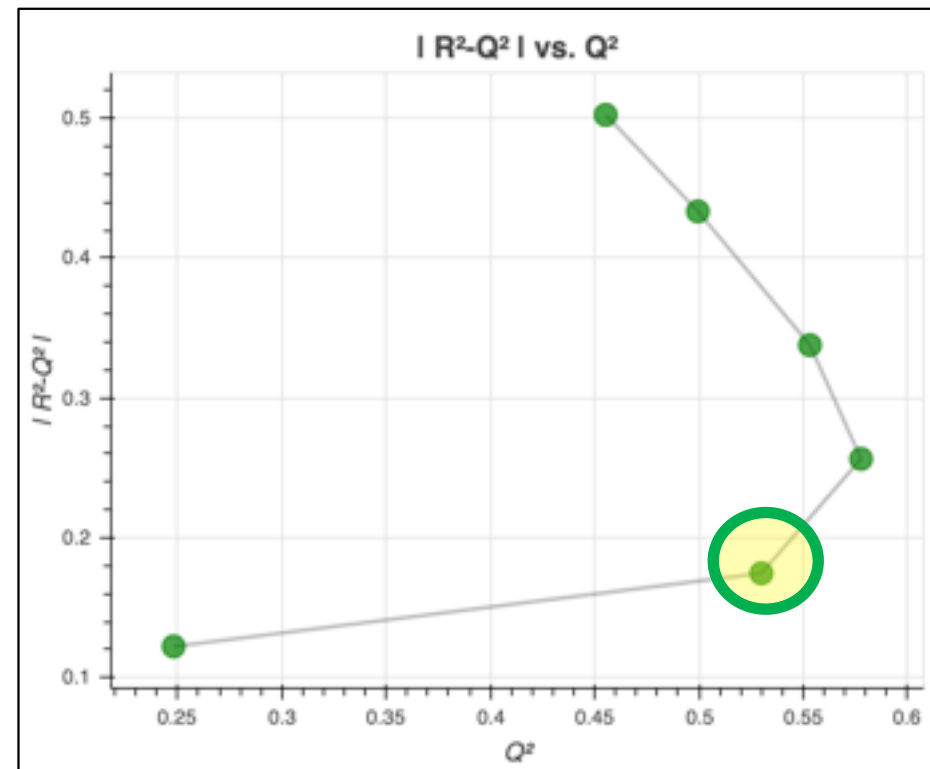
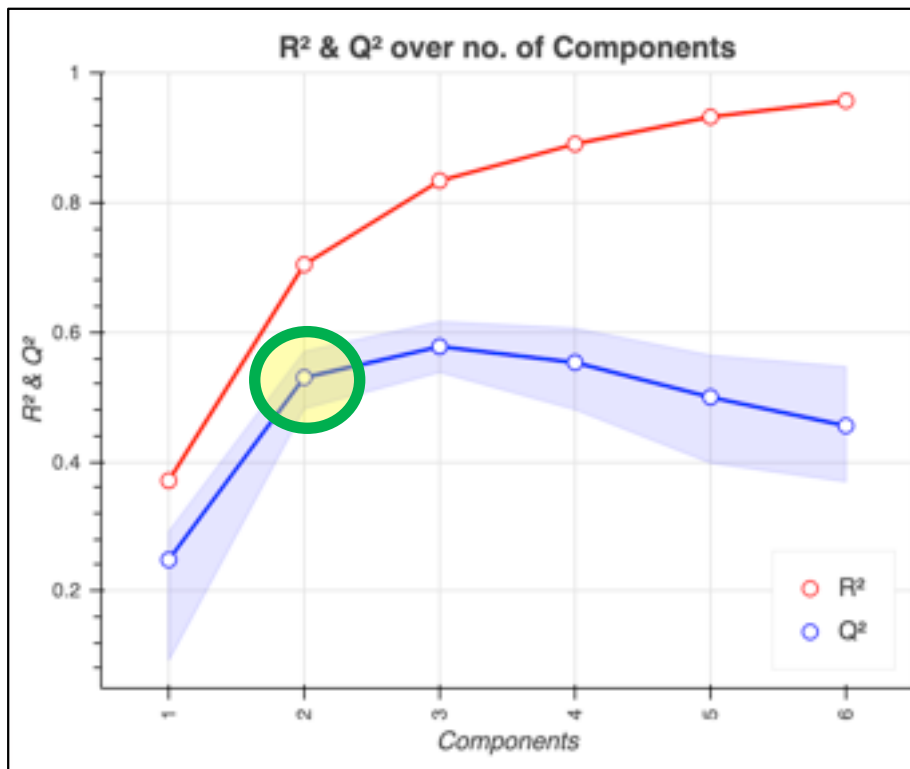
PLS-DA Example.



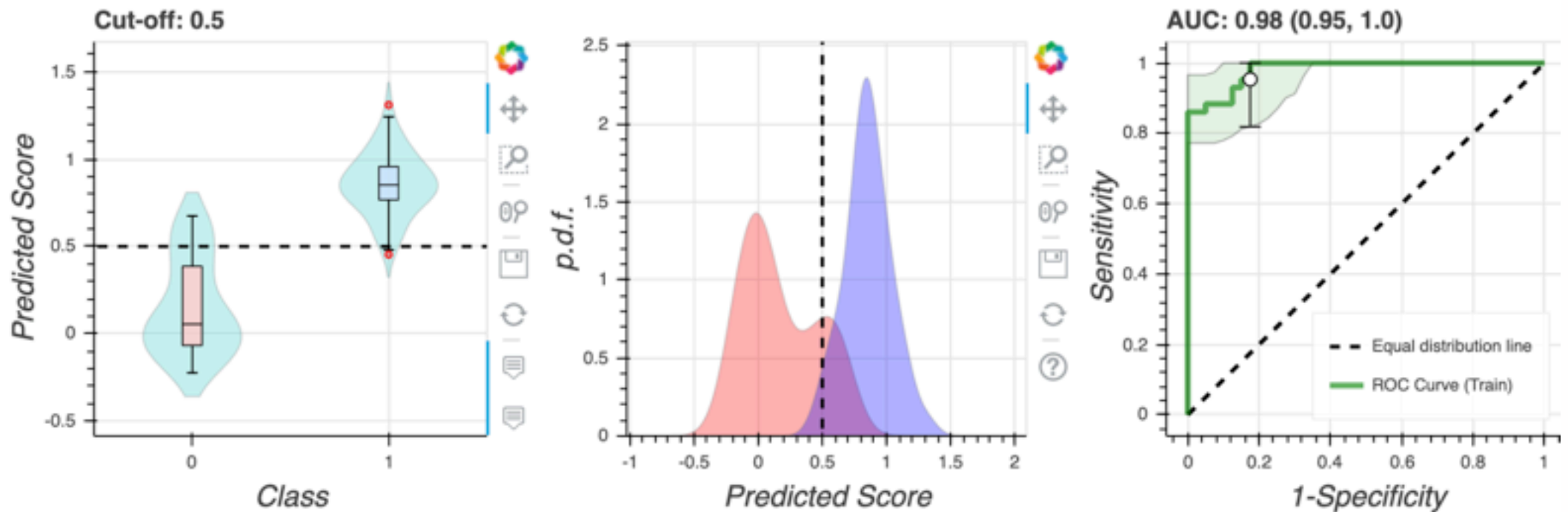
PLS-DA Example: Pareto Front



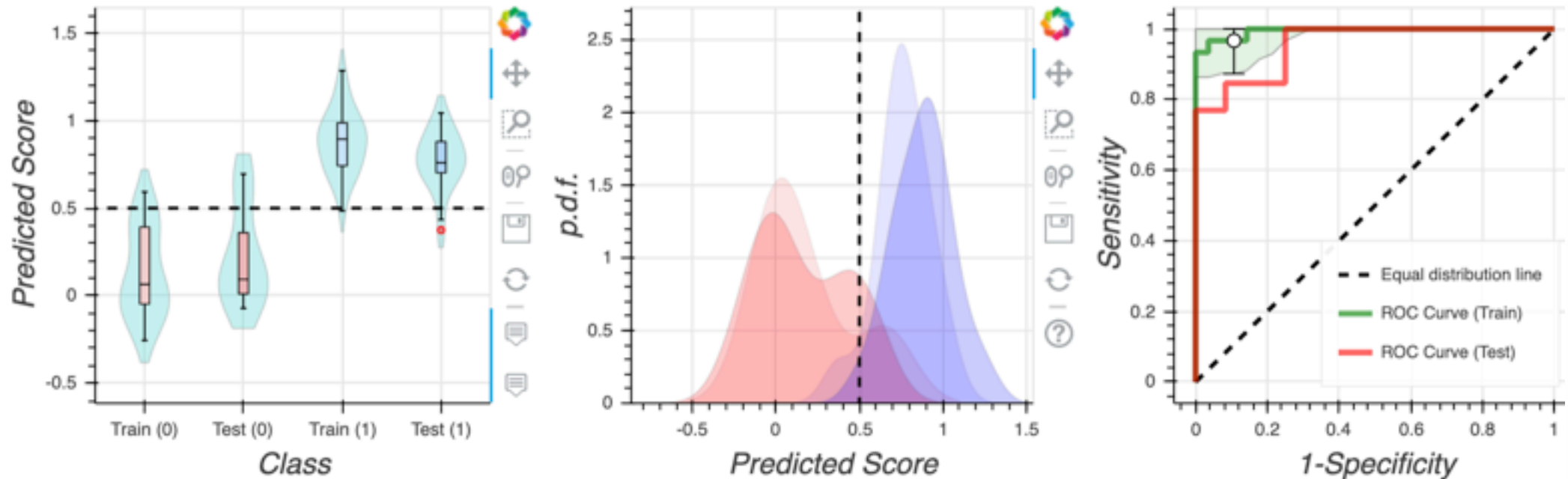
PLS-DA Example: Pareto Front



Train & Evaluate Predictive Ability

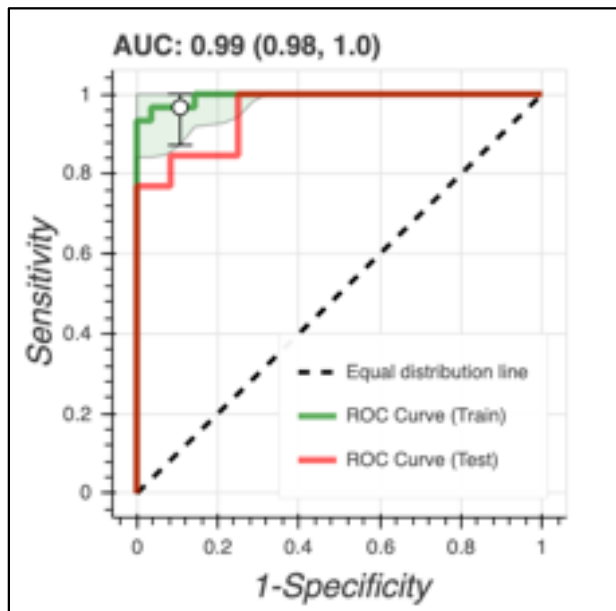


Holdout independent validation (1/3rd)

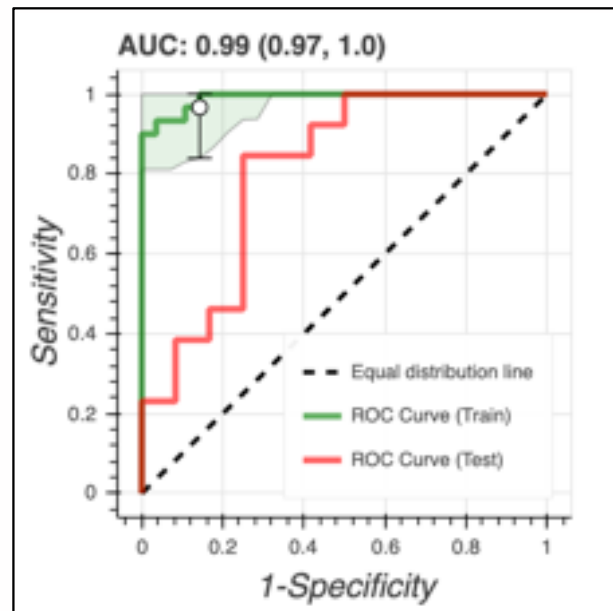


#	Evaluate	MW-U Pvalue	R2	Accuracy	Precision	F1score	Sensitivity	Specificity
0	Train	1.12e-10	0.75 (0.64, 0.83)	0.93 (0.88, 0.97)	0.91 (0.86, 0.95)	0.94 (0.88, 0.97)	0.97 (0.87, 1.0)	0.89
1	Test	1.26e-04	0.6	0.84	0.8	0.86	0.92	0.75

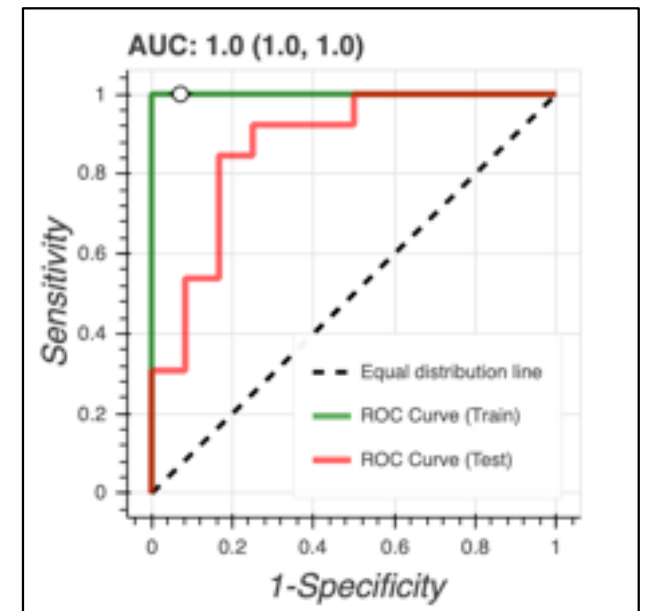
Holdout independent validation (1/3rd)



$AUC_{\text{holdout}}=0.96$

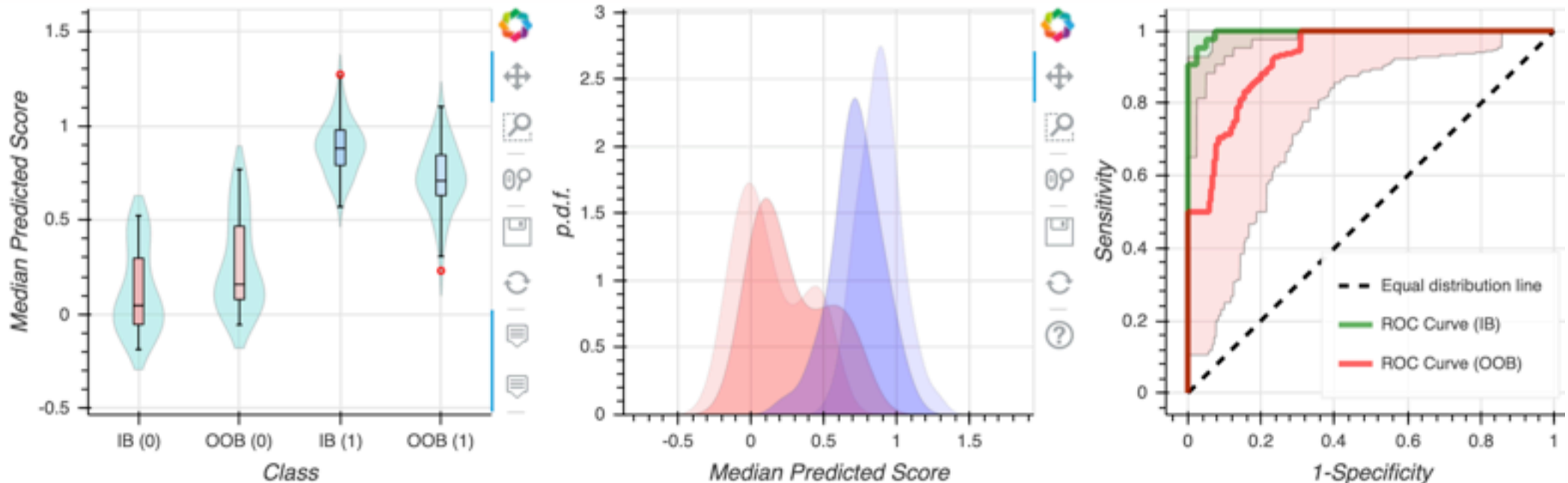


$AUC_{\text{holdout}}=0.81$



$AUC_{\text{holdout}}=0.87$

Bootstrap Validation (n=100)



#	Evaluate	ManW P-Value	R2	AUC
0	In Bag	8.98e-15	0.77 (0.69, 0.83)	1.0 (0.98, 1.0)
1	Out of Bag	1.06e-04	0.51 (0.25, 0.68)	0.93 (0.8, 0.99)

Hyperparameters for ANN

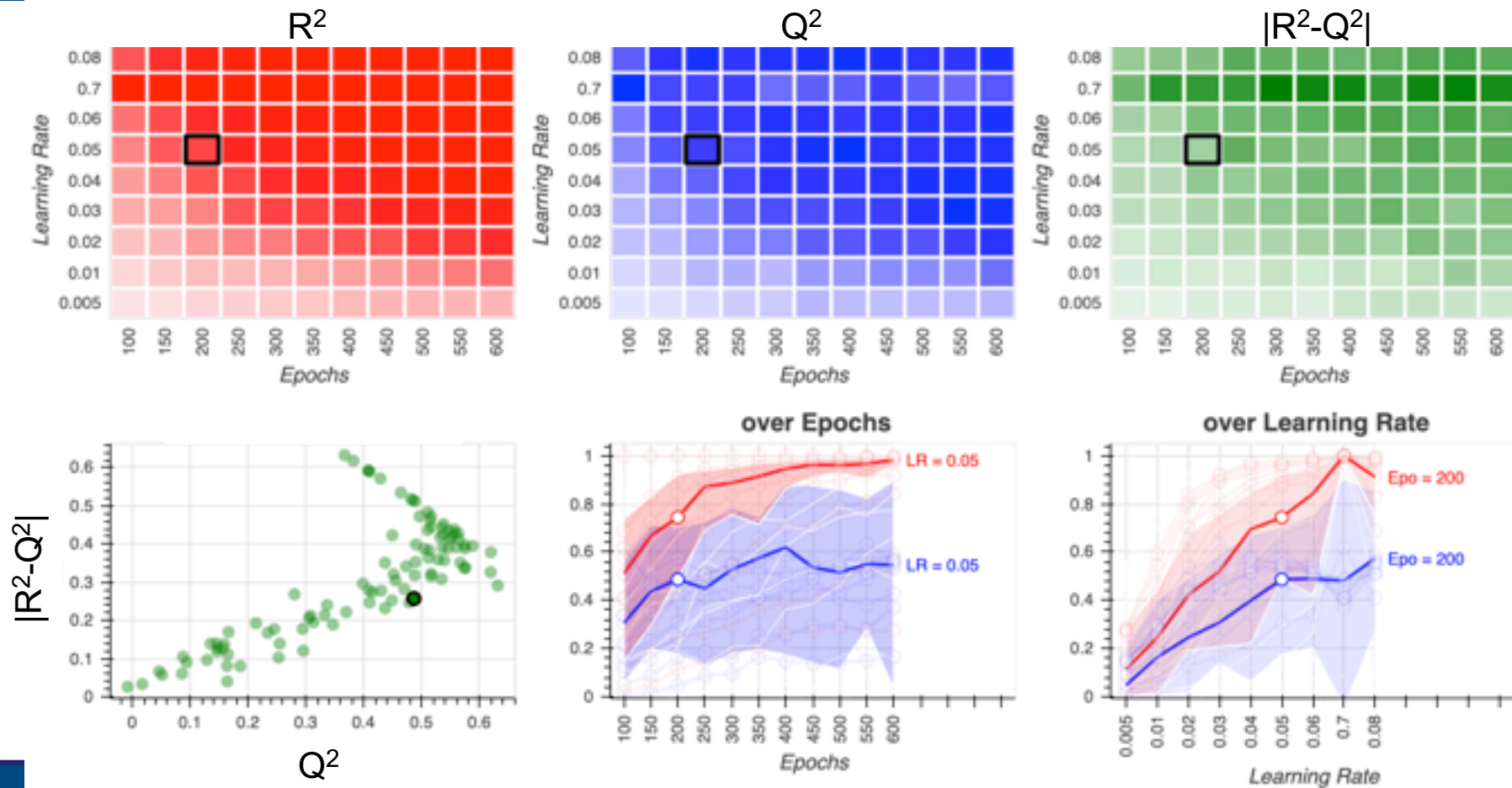
PLS

ANN (3-layer)

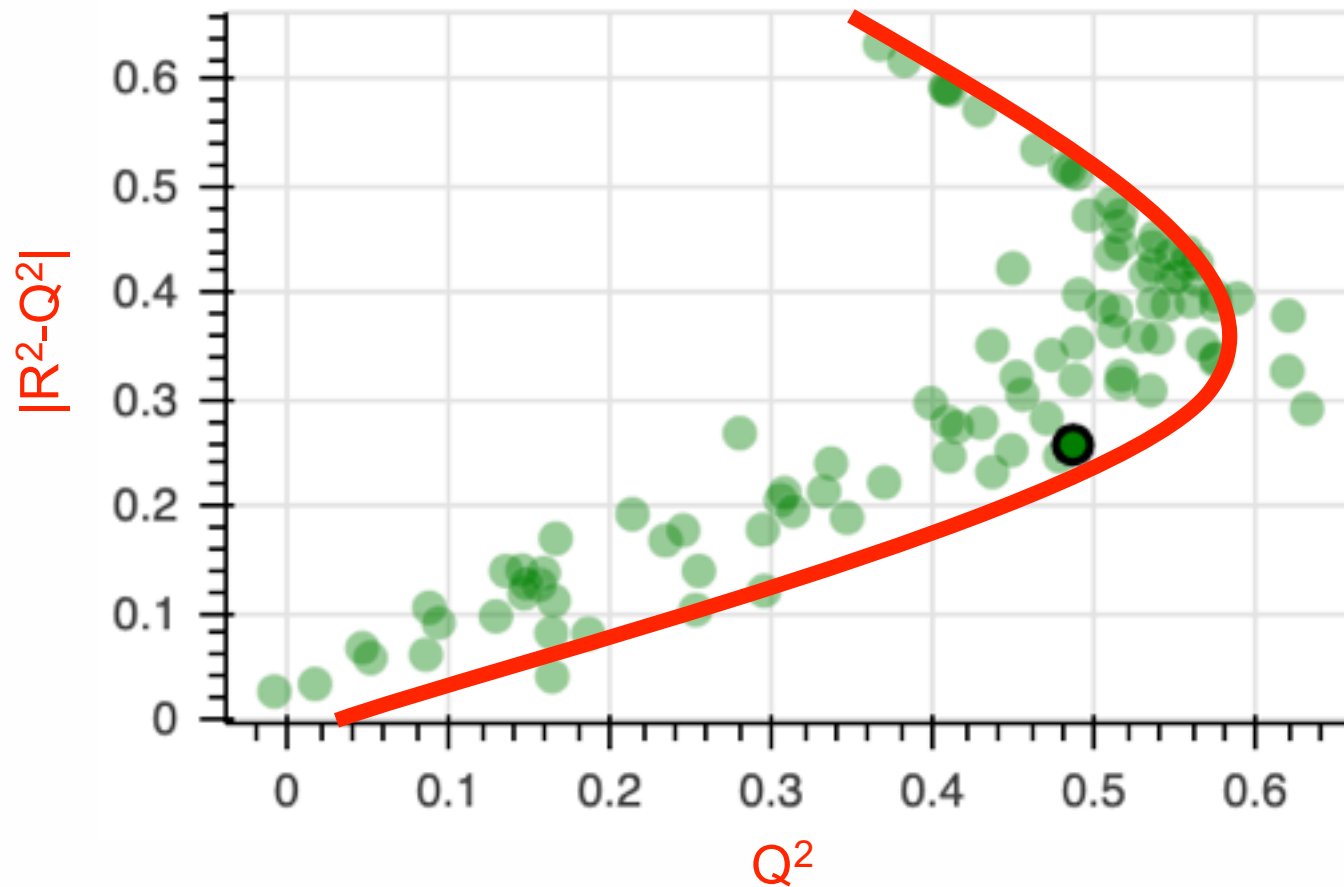
1. No. of Latent variables

1. No. Latent Neurons
2. No. of training epochs
3. Learning rate
4. Learning momentum
5. Learning decay rate

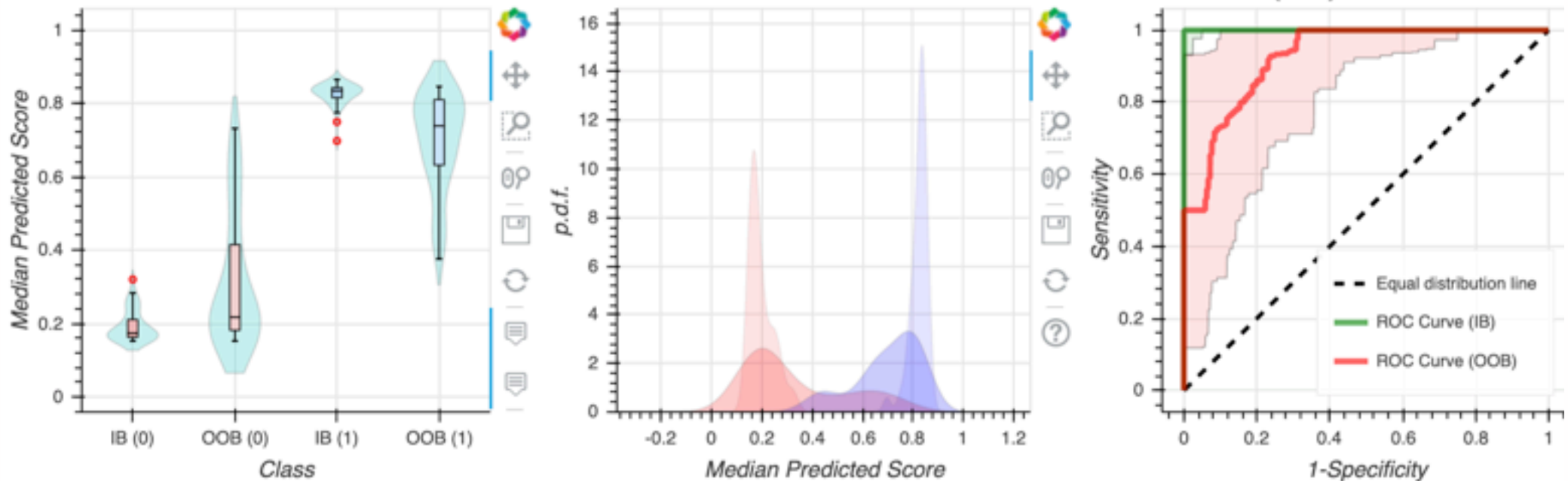
Neurons=2; momentum = 0.5; decay rate = 0



Pareto Front

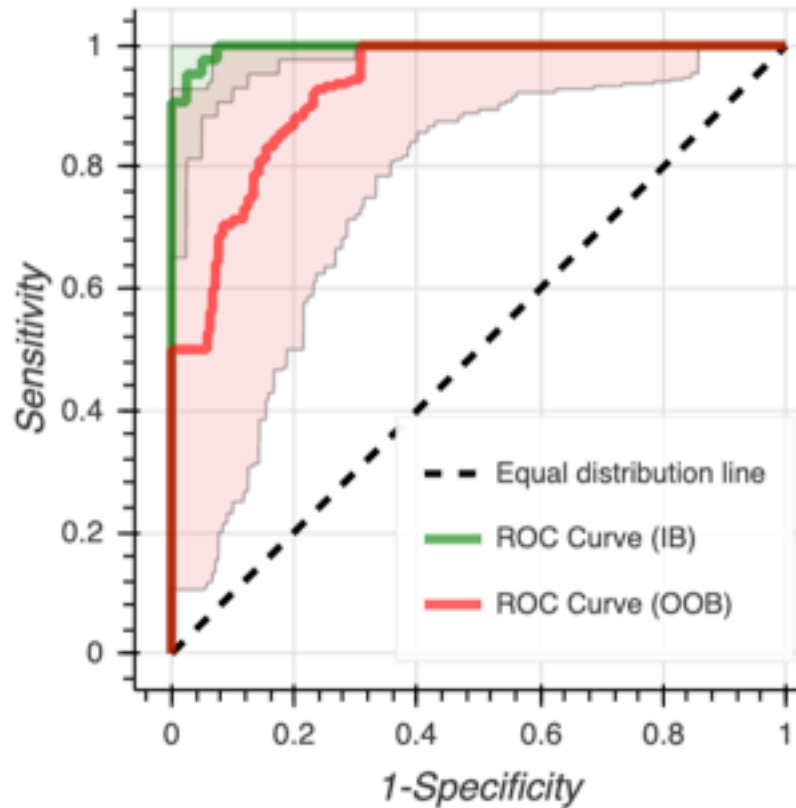


Bootstrap Validation (n=100)

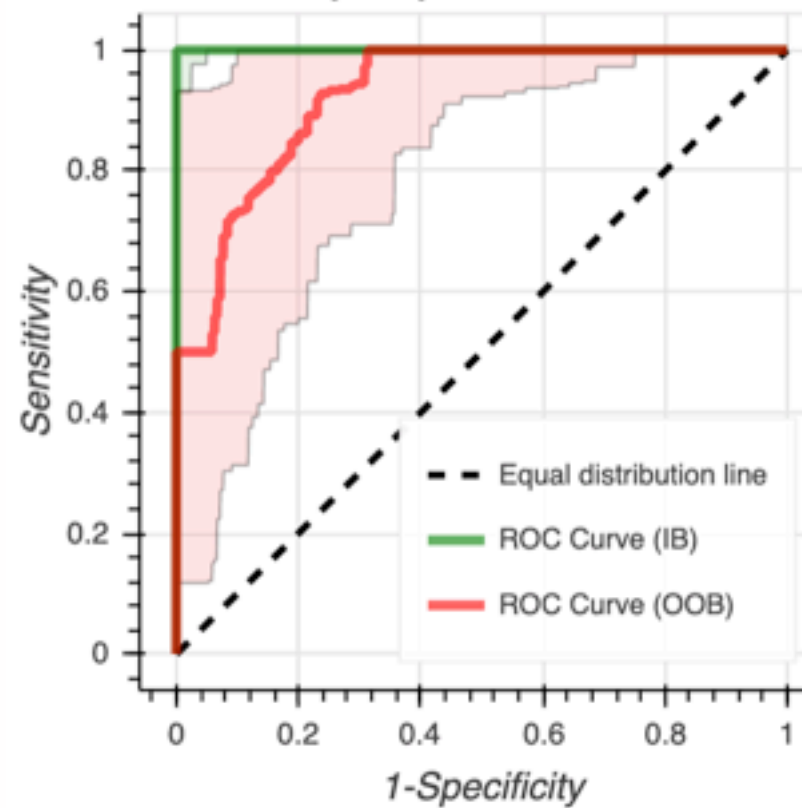


#	Evaluate	ManW P-Value	R2	AUC
0	In Bag	4.69e-15	0.85 (0.66, 0.94)	1.0 (1.0, 1.0)
1	Out of Bag	1.35e-04	0.47 (0.27, 0.65)	0.93 (0.84, 0.99)

PLS-DA



ANN



Example 2: MTBLS93

OPEN ACCESS PEER-REVIEWED

RESEARCH ARTICLE



Large-scale Metabolomic Profiling Identifies Novel Biomarkers for Incident Coronary Heart Disease

Andrea Ganna, Samira Salihovic, Johan Sundström, Corey D. Broeckling, Åsa K. Hedman, Patrik K. E. Magnusson, Nancy L. Pedersen, Anders Larsson, Agneta Siegbahn, Mihkel Zilmer, Jessica Prenni, Johan Ärnlöv, Lars Lind, Tove Fall, Erik Ingelsson 

Published: December 11, 2014 • <https://doi.org/10.1371/journal.pgen.1004801>



LC-MS: 202 metabolites
2,139 subjects
Male vs. Female.

Twin Res Hum Genet, 2013 Feb;16(1):317-29. doi: 10.1017/thg.2012.104. Epub 2012 Nov 9.

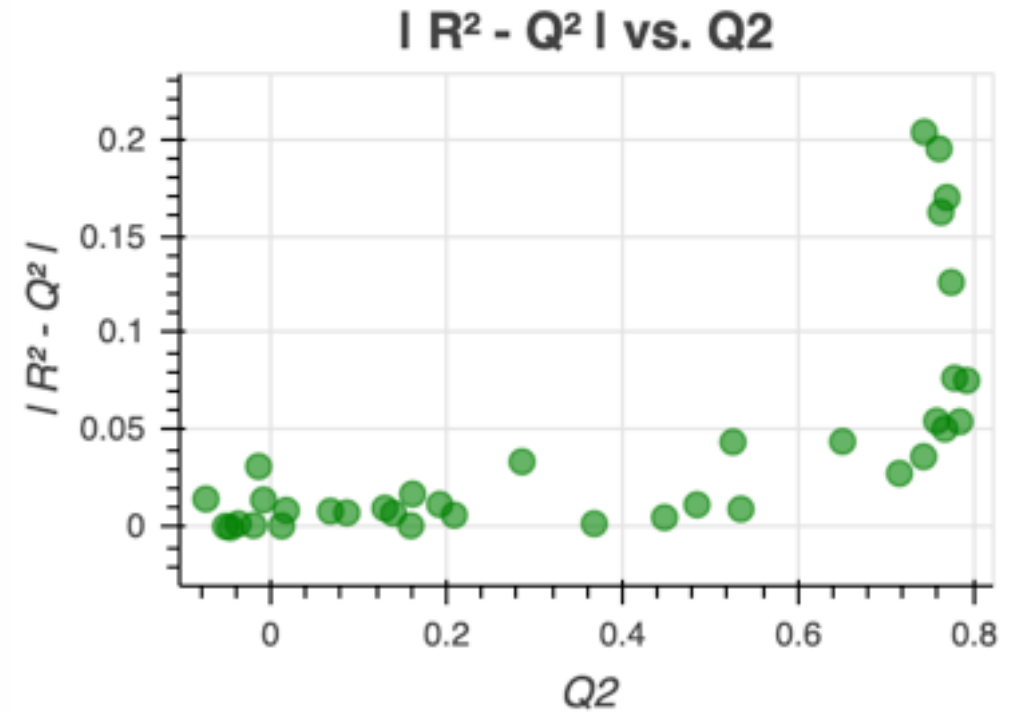
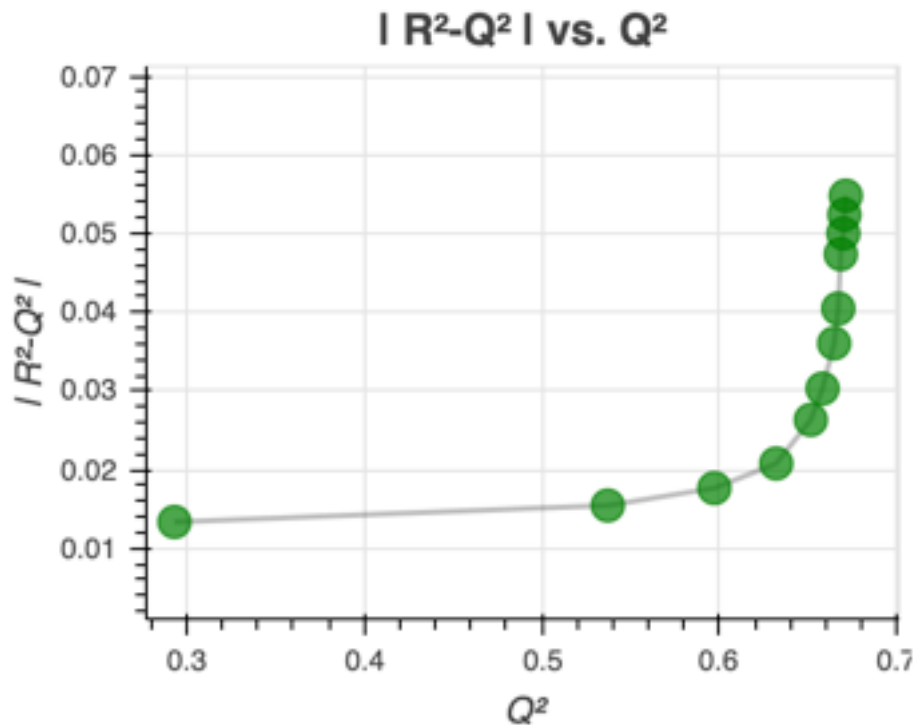
The Swedish Twin Registry: establishment of a biobank and other recent developments.

Magnusson PK¹, Almqvist C, Rahman I, Ganna A, Viktorin A, Walum H, Halldner L, Lundström S, Ullén F, Långström N, Larsson H, Nyman A, Gumpert CH, Råstam M, Anckarsäter H, Cnattingius S, Johannesson M, Ingelsson E, Klareskog L, de Faire U, Pedersen NL, Lichtenstein P.

Example 2: MTBLS93

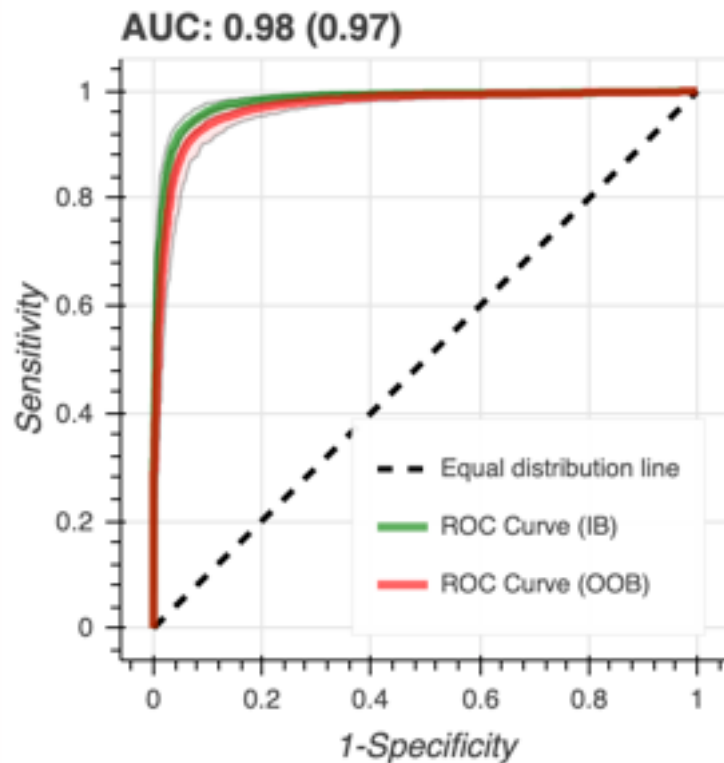
PLS-DA

ANN

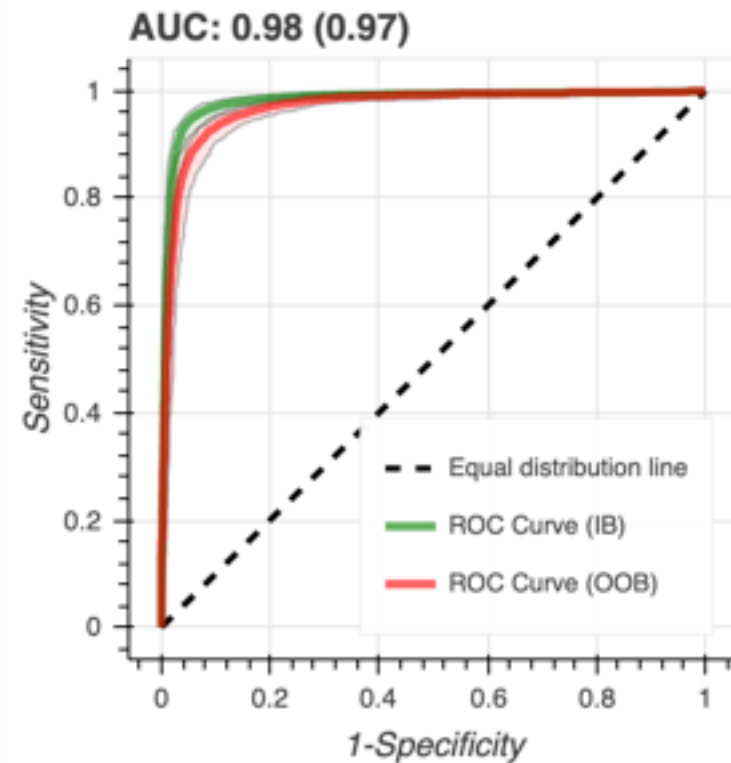


Example 2: MTBLS93

PLS-DA



ANN





Comparison



	Datasets	Platform	No. of Samples	No. of Peaks	PLS-DA	ANN-SS
	ST001047	NMR	140	149	0.93 (0.82, 0.99)	0.93 (0.82, 0.98)
	MTBLS90	LC	968	189	0.82 (0.80, 0.85)	0.83 (0.80, 0.87)
	MTBLS93	LC	2139	202	0.97 (0.96, 0.98)	0.97 (0.96, 0.98)
	MTBLS92	LC	447	240	0.70 (0.64, 0.76)	0.70 (0.62, 0.77)
	MTBLS24	NMR	106	701	0.76 (0.61, 0.89)	0.78 (0.63, 0.91)
	ST000496	GC	100	69	0.96 (0.88, 1.00)	0.96 (0.87, 1.00)
	ST000369	GC	163	180	0.77 (0.63, 0.86)	0.83 (0.73, 0.90)
	MTBLS161U	NMR	58	30	0.72 (0.51, 0.91)	0.80 (0.57, 0.93)
	MTBLS161S	NMR	58	30	0.88 (0.70, 0.99)	0.85 (0.64, 0.96)
	MTBLS547	LC	118	42	0.93 (0.82, 0.99)	0.98 (0.87, 1.00)
	MTBLS404	LC	184	120	0.94 (0.88, 0.98)	0.95 (0.88, 0.98)

Comparison

Poster Number 256

Datasets	Platform	No. of Samples	No. of Peaks	PLS-DA	PCR	PCLR	RBF-SVM	RF	ANN-LS	ANN-SS
ST001047	NMR	140	149	0.93 (0.82, 0.99)	0.89 (0.73, 0.97)	0.79 (0.62, 0.93)	0.92 (0.84, 0.96)	0.89 (0.67, 0.99)	0.78 (0.43, 0.93)	0.93 (0.82, 0.98)
MTBLS90	LC	968	189	0.82 (0.80, 0.85)	0.81 (0.78, 0.85)	0.79 (0.76, 0.83)	0.84 (0.81, 0.88)	0.81 (0.76, 0.85)	0.82 (0.79, 0.85)	0.83 (0.80, 0.87)
MTBLS93	LC	2139	202	0.97 (0.96, 0.98)	0.96 (0.95, 0.97)	0.90 (0.88, 0.91)	0.97 (0.96, 0.98)	0.92 (0.89, 0.94)	0.97 (0.96, 0.98)	0.97 (0.96, 0.98)
MTBLS92	LC	447	240	0.70 (0.64, 0.76)	0.68 (0.61, 0.74)	0.65 (0.60, 0.70)	0.72 (0.65, 0.78)	0.73 (0.67, 0.80)	0.70 (0.63, 0.78)	0.70 (0.62, 0.77)
MTBLS24	NMR	106	701	0.76 (0.61, 0.89)	0.76 (0.58, 0.87)	0.70 (0.57, 0.84)	0.83 (0.72, 0.94)	0.85 (0.73, 0.93)	0.80 (0.67, 0.89)	0.78 (0.63, 0.91)
ST000496	GC	100	69	0.96 (0.88, 1.00)	0.89 (0.70, 0.98)	0.89 (0.77, 0.97)	0.96 (0.88, 1.00)	0.79 (0.62, 0.94)	0.93 (0.78, 0.99)	0.96 (0.87, 1.00)
ST000369	GC	163	180	0.77 (0.63, 0.86)	0.70 (0.54, 0.80)	0.65 (0.55, 0.74)	0.82 (0.71, 0.90)	0.73 (0.59, 0.84)	0.72 (0.59, 0.81)	0.83 (0.73, 0.90)
MTBLS161U	NMR	58	30	0.72 (0.51, 0.91)	0.76 (0.55, 0.94)	0.73 (0.48, 0.89)	0.85 (0.66, 0.96)	0.70 (0.46, 0.84)	0.76 (0.56, 0.91)	0.80 (0.57, 0.93)
MTBLS161S	NMR	58	30	0.88 (0.70, 0.99)	0.75 (0.54, 0.90)	0.69 (0.54, 0.87)	0.85 (0.68, 0.97)	0.78 (0.60, 0.94)	0.77 (0.53, 0.94)	0.85 (0.64, 0.96)
MTBLS547	LC	118	42	0.93 (0.82, 0.99)	0.92 (0.80, 0.99)	0.85 (0.74, 0.94)	0.97 (0.89, 1.00)	0.92 (0.77, 0.98)	0.93 (0.80, 0.99)	0.98 (0.87, 1.00)
MTBLS404	LC	184	120	0.94 (0.88, 0.98)	0.91 (0.84, 0.96)	0.81 (0.69, 0.89)	0.95 (0.88, 0.99)	0.81 (0.68, 0.91)	0.88 (0.72, 0.95)	0.95 (0.88, 0.98)



Kevin
Mendez

Summary

1. All methods prone to overtrain (validation is important)
2. Increased number of hyperparameters and the more complex the model the less robust the model (large confidence intervals)
3. Only the big data sets produce models with low bias.

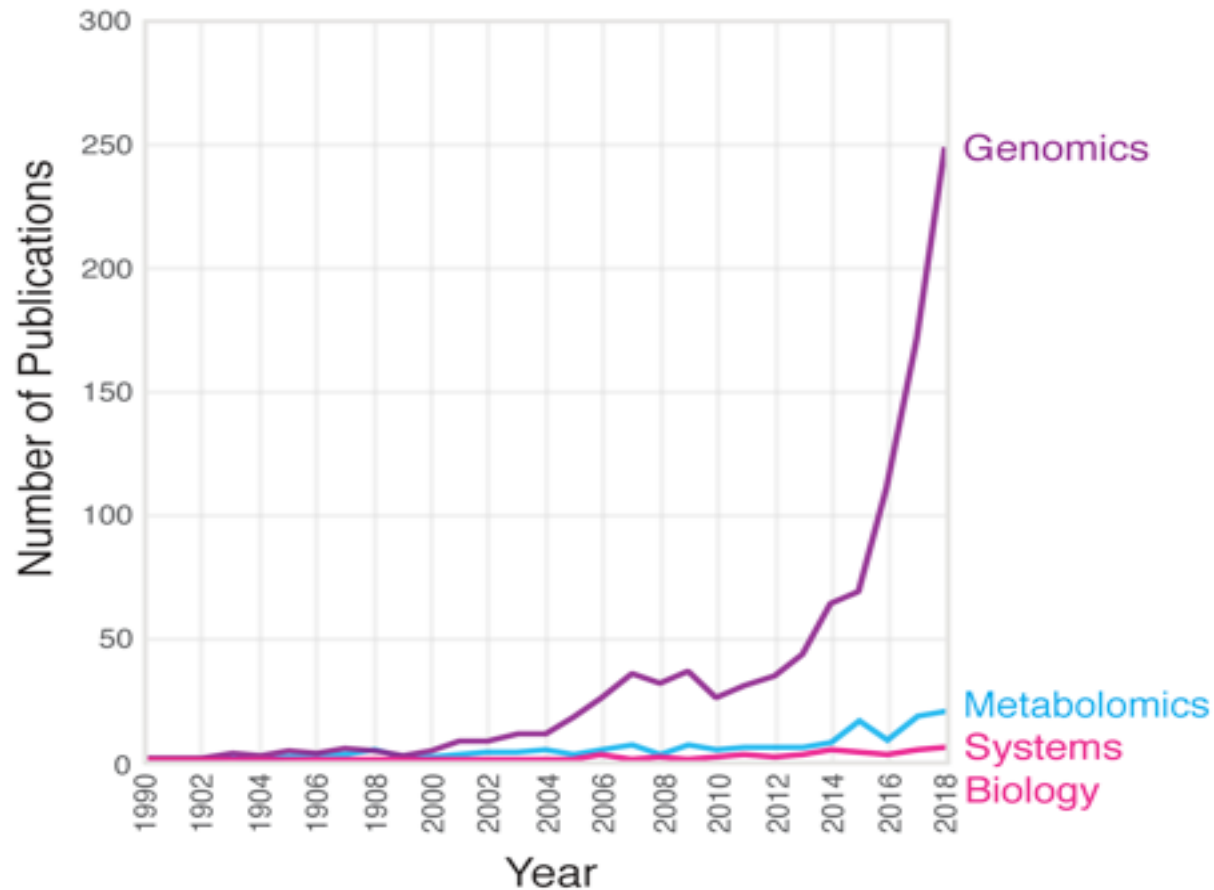


Centre for Integrative Metabolomics
& Computational Biology

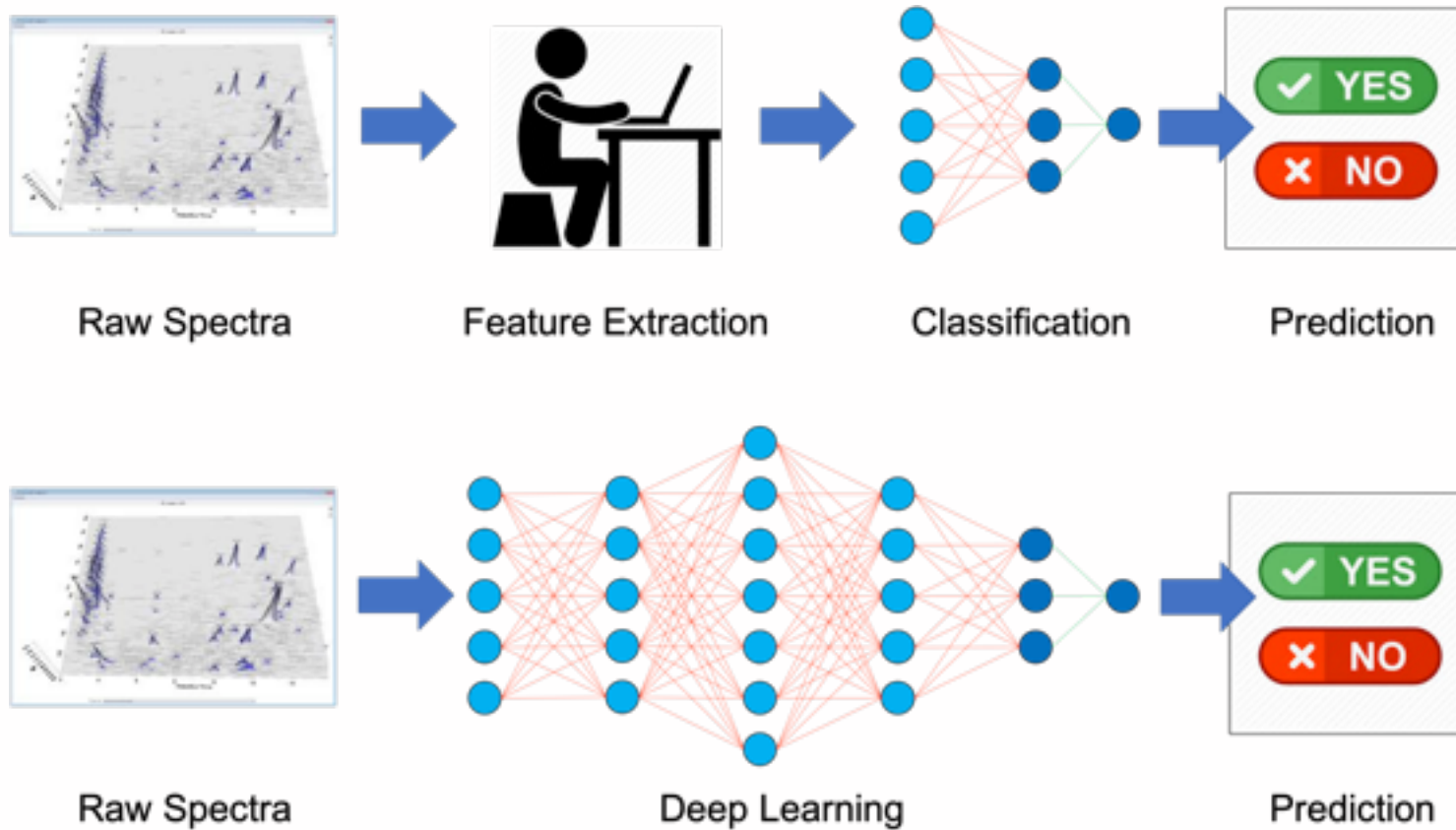


Future Perspectives

Rise of ANNs & Deep Learning



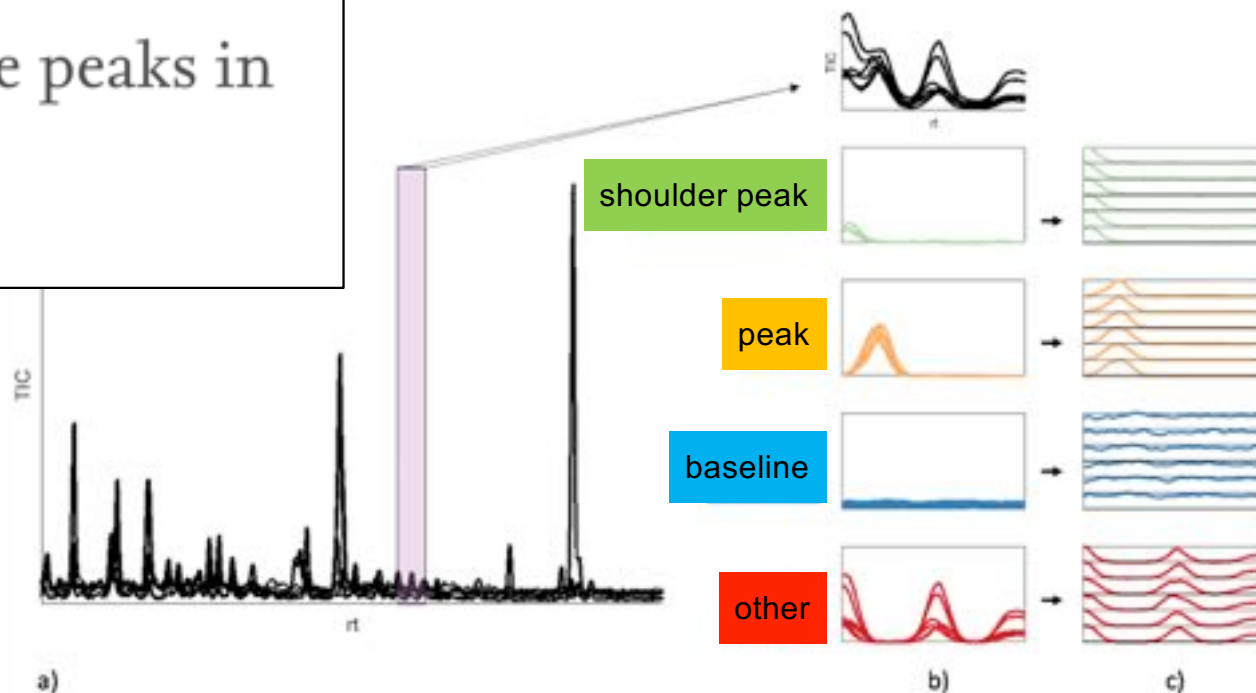
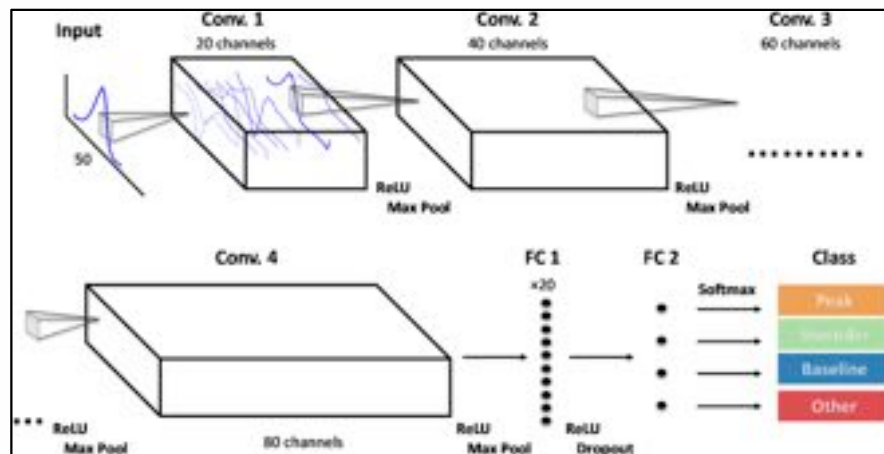
Deep Learning



Deep Learning

Using deep learning to evaluate peaks in chromatographic data

Anne Bech Risum, Rasmus Bro  



The models are built on a training set with PARAFAC2 resolved components from eight different aroma related GC-MS runs with a total of over 70,000 elution profile samples, and validated using another, independent, GC-MS dataset.

Deep Learning

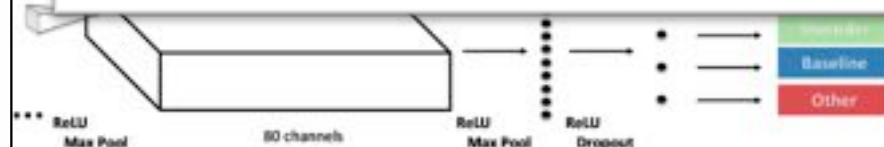
Using deep learning to evaluate peaks in

P-528 Metabolite identification from LC-MS/MS spectra using deep learning

PRESENTING AUTHOR: Svetlana Kutuzova, *Technical University of Denmark, Denmark*

CO-AUTHORS: Douglas McCloskey, Christian Igel

Mass spectrometry is a powerful high-throughput technology for chemical composition assessment. However the data processing of the resulting spectra is a major bottleneck for large studies, and in particular the metabolite identification from the mass spectra. The joint community effort of collecting and maintaining metabolomics spectral databases provides the opportunity to approach the metabolite identification problem with powerful but data-hungry algorithms including deep learning. We present a novel deep learning based algorithm for compound identification that makes a prediction of a structural chemical fingerprint based on a LC-MS/MS spectrum of a compound. Both raw spectra and fragmentation tree predicted by SIRIUS software are used as an input. A Tree-LSTM network is used to process the fragmentation trees alongside with a feed forward neural network that captures patterns in the spectral data. Our method is validated on the CASMI 2017 challenge dataset. While the method does not yet outperform the state-of-the-art approach it is shown to be a proof of concept and a solid base for future developments. The future work would include learning fragmentation rules from the spectrum itself enabling a complete end-to-end spectrum analysis.



a)

b)

c)

The models are built on a training set with PARAFAC2 resolved components from eight different aroma related GC-MS runs with a total of over 70,000 elution profile samples, and validated using another, independent, GC-MS dataset.

Multi-omic Integration

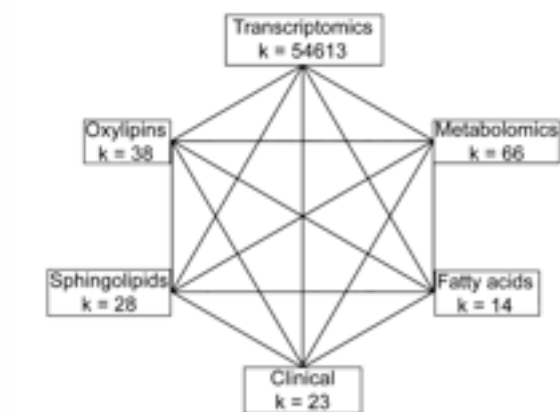
analytical
chemistry

Cite This: Anal. Chem. 2018, 90, 13400–13408

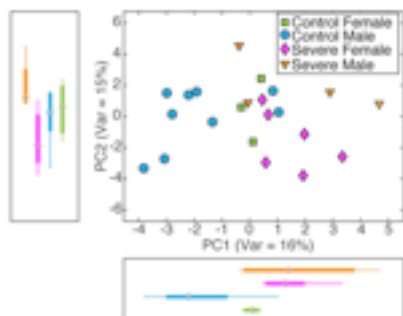
pubs.acs.org/doi

OnPLS-Based Multi-Block Data Integration: A Multivariate Approach to Interrogating Biological Interactions in Asthma

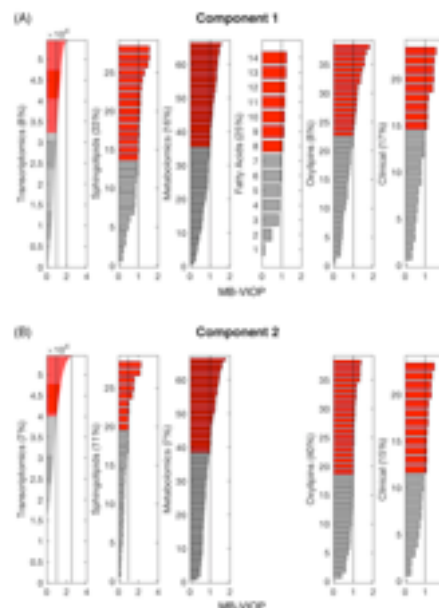
Stacey N. Reinke,^{*,1,2} Beatriz Galindo-Prieto,^{3,4,5} Tomas Skotare,⁶ David I. Broadhurst,² Akul Singhania,^{*,7} Daniel Horowitz,⁸ Ratko Djukanovic,^{*,9} Timothy S.C. Hinks,^{*,10,11} Paul Geladi,¹² Johan Trygg,¹³ and Craig E. Wheelock^{*,1,14}



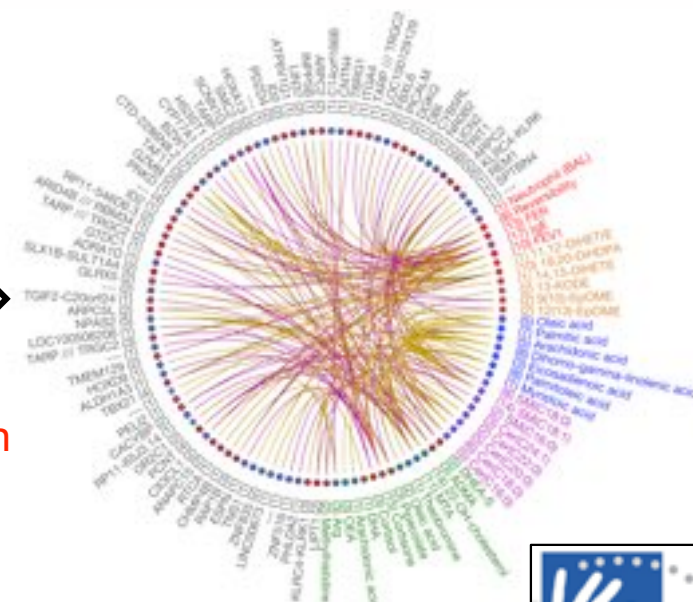
multi-block
modelling



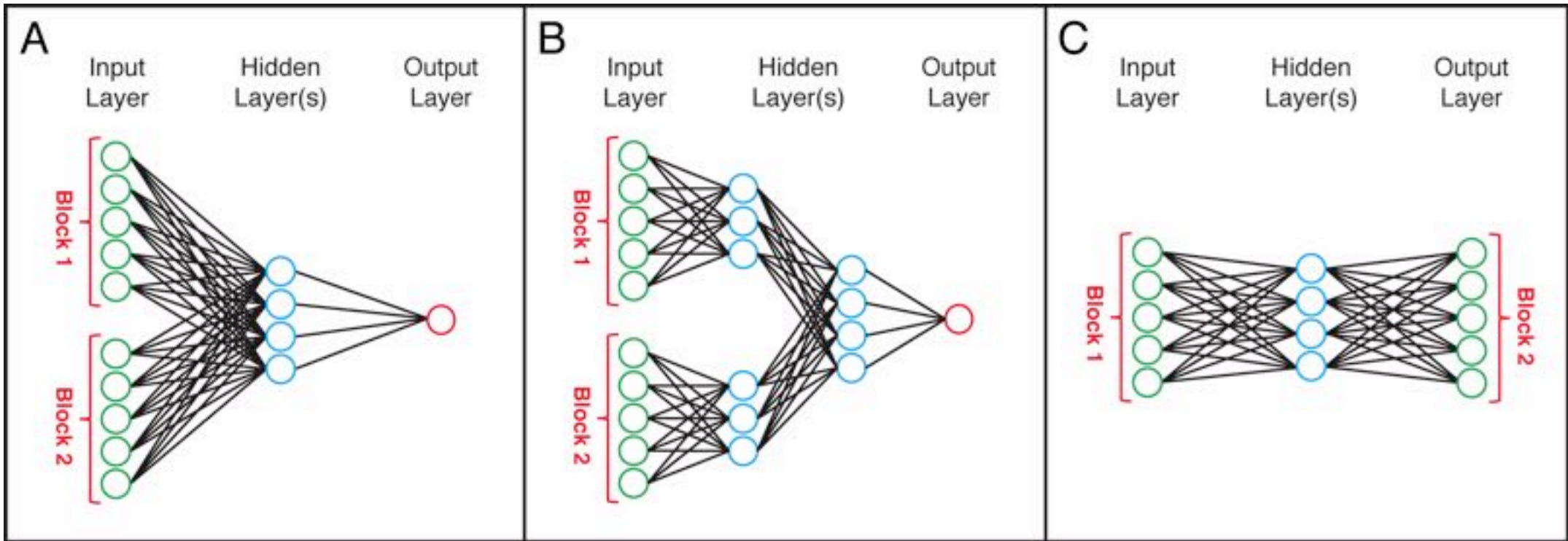
feature
selection



biological
association



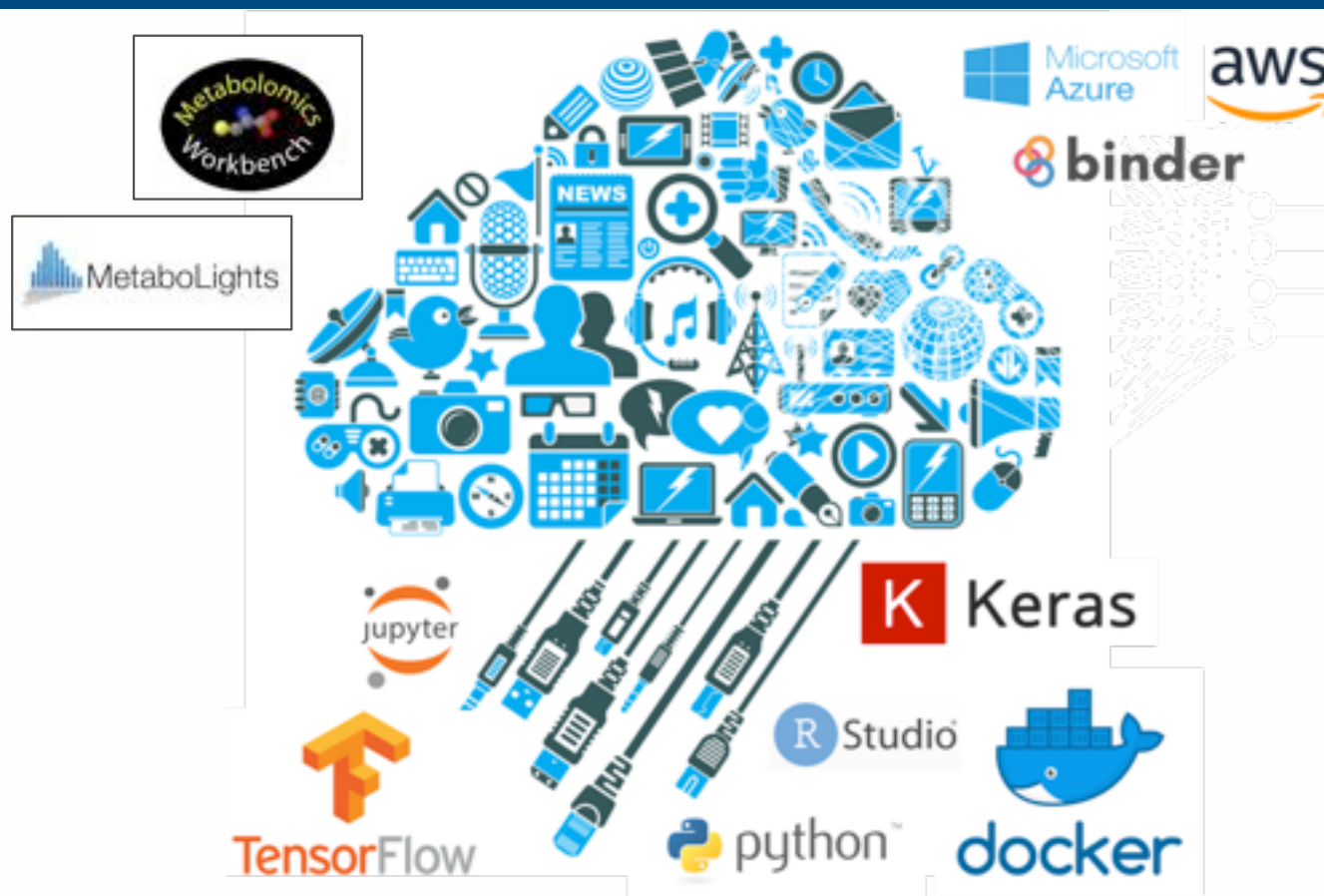
Multi-block ANN





Centre for Integrative Metabolomics
& Computational Biology

The time is now!





Centre for Integrative Metabolomics
& Computational Biology

Tidy Data (not Raw)

+



+



+



Microsoft
Azure



Anyone
can
learn
to
code!



Conclusions

- Be wary of the HYPE! (often simple is better)
- Large data is required (Tidy Data)
- Black Box +++
- Huge Potential
- It is not going away - so get educated
- Learn to code!

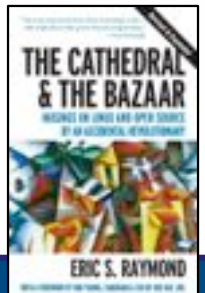


python™



MATLAB

STATA





Centre for Integrative Metabolomics
& Computational Biology



Acknowledgements



Kevin Mendez (MSc Student)



Dr. Stacey Reinke



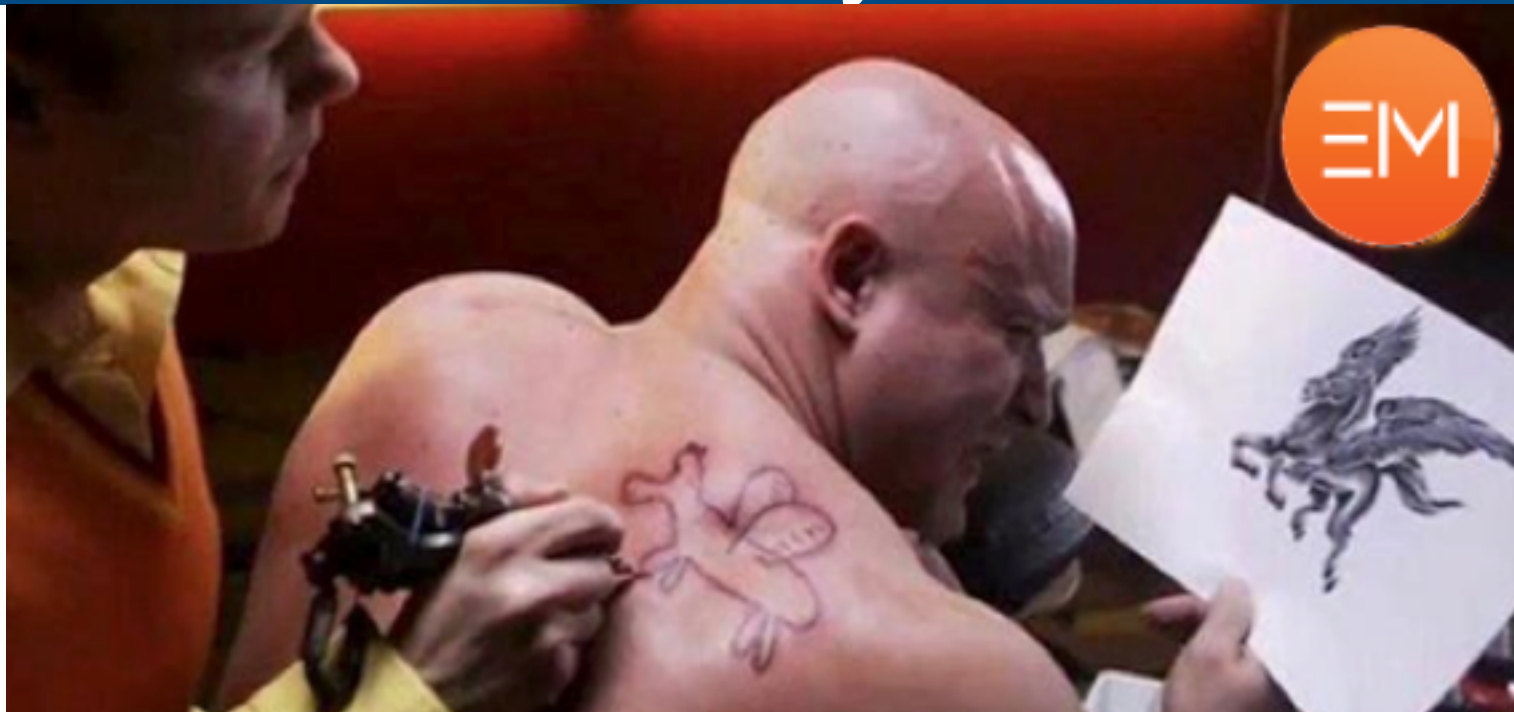
Dr. Leighton Pritchard



Australian Government
Australian Research Council



Thank you!



**THERE WILL ALWAYS BE SOMEONE WHO
SAYS THAT THEY CAN DO IT CHEAPER...
BUT AT WHAT COST?**



Centre for Integrative Metabolomics
& Computational Biology

