Predictive Models in Health Research

Trevor Hastie
Department of Statistics
Department of Biomedical Data Science
Stanford University



41st Meeting of the International Society for Clinical Biostatistics
ISCB, Krakow, Poland, August 23-27, 2020

Some Take Home Messages

This talk is about supervised learning: building models from data that predict an outcome using a collection of input features.

- There are some powerful and exciting tools for making predictions from data.
- They are not magic! You should be skeptical. They require good data and proper internal validation.
- Human judgement and ingenuity are essential for their success.
- They may be overkill — if simpler and more traditional methods do as well, they are preferable (Occam's razor).

Some Definitions

Machine Learning constructs algorithms that can learn from data.

Statistical Learning is a branch of applied statistics that emerged in response to machine learning, emphasizing statistical models and assessment of uncertainty.

Data Science is the extraction of knowledge from data, using ideas from mathematics, statistics, machine learning, computer science, engineering, ...

All of these are very similar — with different emphases.

Internal Model Validation

- IMPORTANT! Don't trust me or anyone who says they have a wonderful machine learning algorithm, unless you see the results of a careful internal validation.
- Eg: divide data into two parts A and B. Run algorithm on part A and then test it on part B.
  Algorithm must not have seen any of the data in part B.
- If it works in part B, you have (some) confidence in it

Internal Model Validation

- IMPORTANT! Don't trust me or anyone who says they have a wonderful machine learning algorithm, unless you see the results of a careful internal validation.
- Eg: divide data into two parts A and B. Run algorithm on part A and then test it on part B.
  Algorithm must not have seen any of the data in part B.
- If it works in part B, you have (some) confidence in it

  Simple? Yes

Internal Model Validation

- IMPORTANT! Don't trust me or anyone who says they have a wonderful machine learning algorithm, unless you see the results of a careful internal validation.
- Eg: divide data into two parts A and B. Run algorithm on part A and then test it on part B.
  Algorithm must not have seen any of the data in part B.
- If it works in part B, you have (some) confidence in it

  Simple? Yes
  Done properly in practice? Rarely

Internal Model Validation

- IMPORTANT! Don't trust me or anyone who says they have a wonderful machine learning algorithm, unless you see the results of a careful internal validation.
- Eg: divide data into two parts A and B. Run algorithm on part A and then test it on part B.
  Algorithm must not have seen any of the data in part B.
- If it works in part B, you have (some) confidence in it

Simple? Yes
Done properly in practice? Rarely

In God we trust. All others bring data.

Internal Model Validation

- IMPORTANT! Don't trust me or anyone who says they have a wonderful machine learning algorithm, unless you see the results of a careful internal validation.
- Eg: divide data into two parts A and B. Run algorithm on part A and then test it on part B.
  Algorithm must not have seen any of the data in part B.
- If it works in part B, you have (some) confidence in it

Simple? Yes
Done properly in practice? Rarely

*In God we trust. All others bring data.**

* Statistical "proverb" sometimes attributed to W. Edwards Deming.

Big data vary in *shape*. These call for different approaches.

Wide Data



Thousands / Millions of Variables

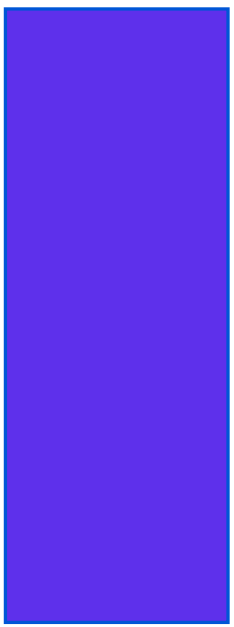
Hundreds of Samples

Screening and fdr,
Lasso, SVM, Stepwise

We have too many variables; prone to overfitting.

Need to remove variables, or regularize, or both.

Tall Data



Tens / Hundreds of Variables

Thousands / Millions of Samples

GLM, Random Forests,
Boosting, Deep Learning

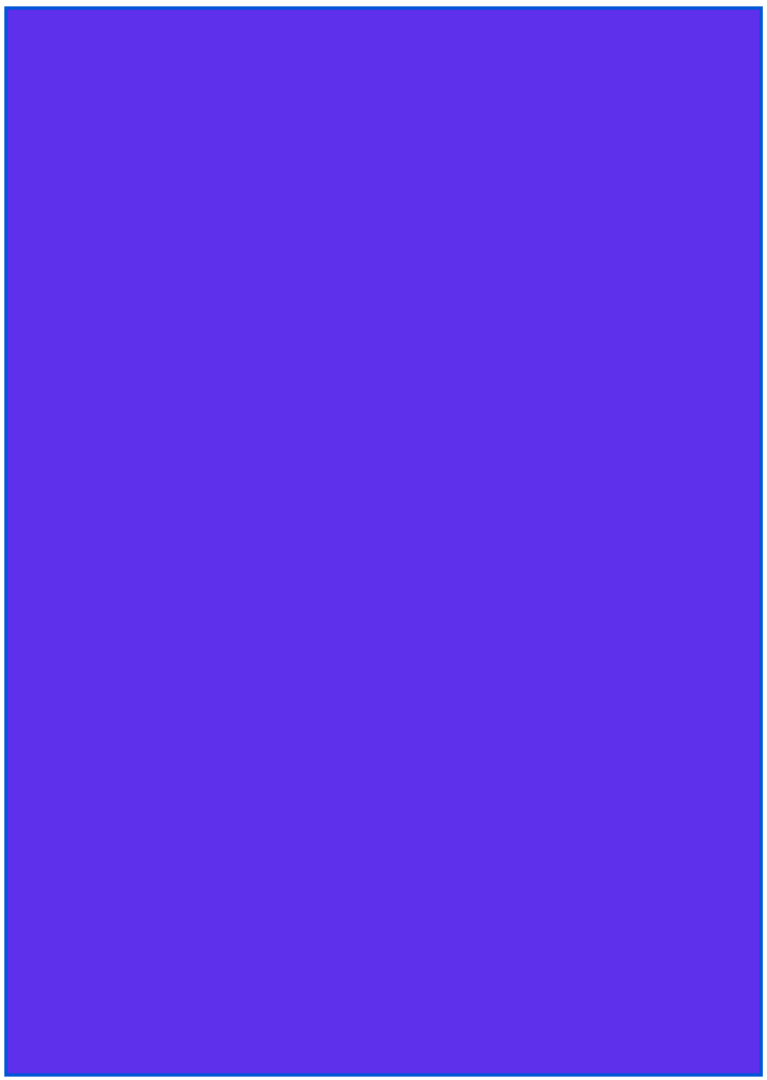
Sometimes simple models (linear) don't suffice.

We have enough samples to fit nonlinear models with many interactions, and not too many variables.

Good automatic methods for doing this.

Big data vary in *shape*. These call for different approaches.

Tall and Wide Data



Thousands / Millions of Variables

Millions to Billions of Samples

Tricks of the Trade

Exploit sparsity

Random projections / hashing

Variable screening

Subsample rows

Divide and recombine

Case/ control sampling

MapReduce

ADMM (divide and conquer)

.
.
.

Outline of remainder of talk

- Large scale lasso models using `glmnet` (tall and wide)
- Random Forest used as ensemble learner (tall)
- Deep learning used for image classification (wide)
- Machine learning overkill (very wide)

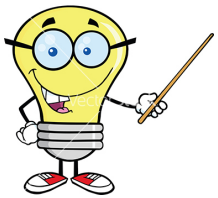
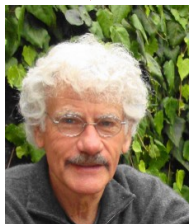
glmnet

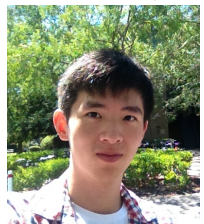
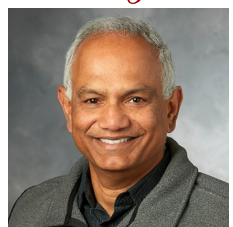
Fit regularization paths for a variety of GLMs with lasso and elastic net penalties; e.g. logistic regression

$$\log \frac{\Pr(Y = 1 \mid X = x)}{\Pr(Y = 0 \mid X = x)} = \beta_0 + \sum_{j=1}^{p} x_j \beta_j$$

- Lasso penalty [Tibshirani, 1996] induces *sparsity* in coefficients: $\sum_{j=1}^{p} |\beta_j| \leq s$. It shrinks them toward zero, and sets many to zero.
- Fit efficiently using coordinate descent. Handles sparse X naturally, and exploits sparsity of solutions, warms starts, variable screening, and includes methods for model selection using cross-validation.

glmnet team: TH, Jerome Friedman, Rob Tibshirani, Noah Simon, Junyang Qian, Balasubramanian Narasimhan, Ken Tay.


Cold Spring Harbor Laboratory
bioRxiv
THE PREPRINT SERVER FOR BIOLOGY

HOME | ABOUT | SUBMIT | NEWS & NOTES | ALERTS / RSS | CHANNE

Search

Advanced Sea

bioRxiv is receiving many new papers on coronavirus SARS-CoV-2. A reminder: these are preliminary reports that have not been peer-reviewed. They should not be regarded as conclusive practice/health-related behavior, or be reported in news media as established information.

New Results

Comments (6)

A Fast and Scalable Framework for Large-scale and Ultrahigh-dimensional Sparse Regression with Application to the UK Biobank

Junyang Qian, Yosuke Tanigawa, Wenfei Du, Matthew Aguirre, Chris Chang, Robert Tibshirani, Manuel A. Rivas, Trevor Hastie

doi: https://doi.org/10.1101/630079

This article is a preprint and has not been certified by peer review [what does this mean?].

Abstract | Full Text | Info/History | Metrics | Preview PDF

Abstract

The UK Biobank (Bycroft et al., 2018) is a very large, prospective population-based cohort study across the United Kingdom. It provides unprecedented opportunities for researchers to investigate the relationship between genotypic information and phenotypes of interest. Multiple regression methods, compared with GWAS, have already been showed to greatly improve the prediction performance for a variety of phenotypes. In the high-dimensional settings, the lasso (Tibshirani, 1996), since its

Previous

Posted May 31, 2020.

Download PDF
XML
Revision Summary

Email
Share
Citation Tool

Tweet Like 0

COVID-19 SARS-CoV-2 preprints from medRxiv and bioRxiv

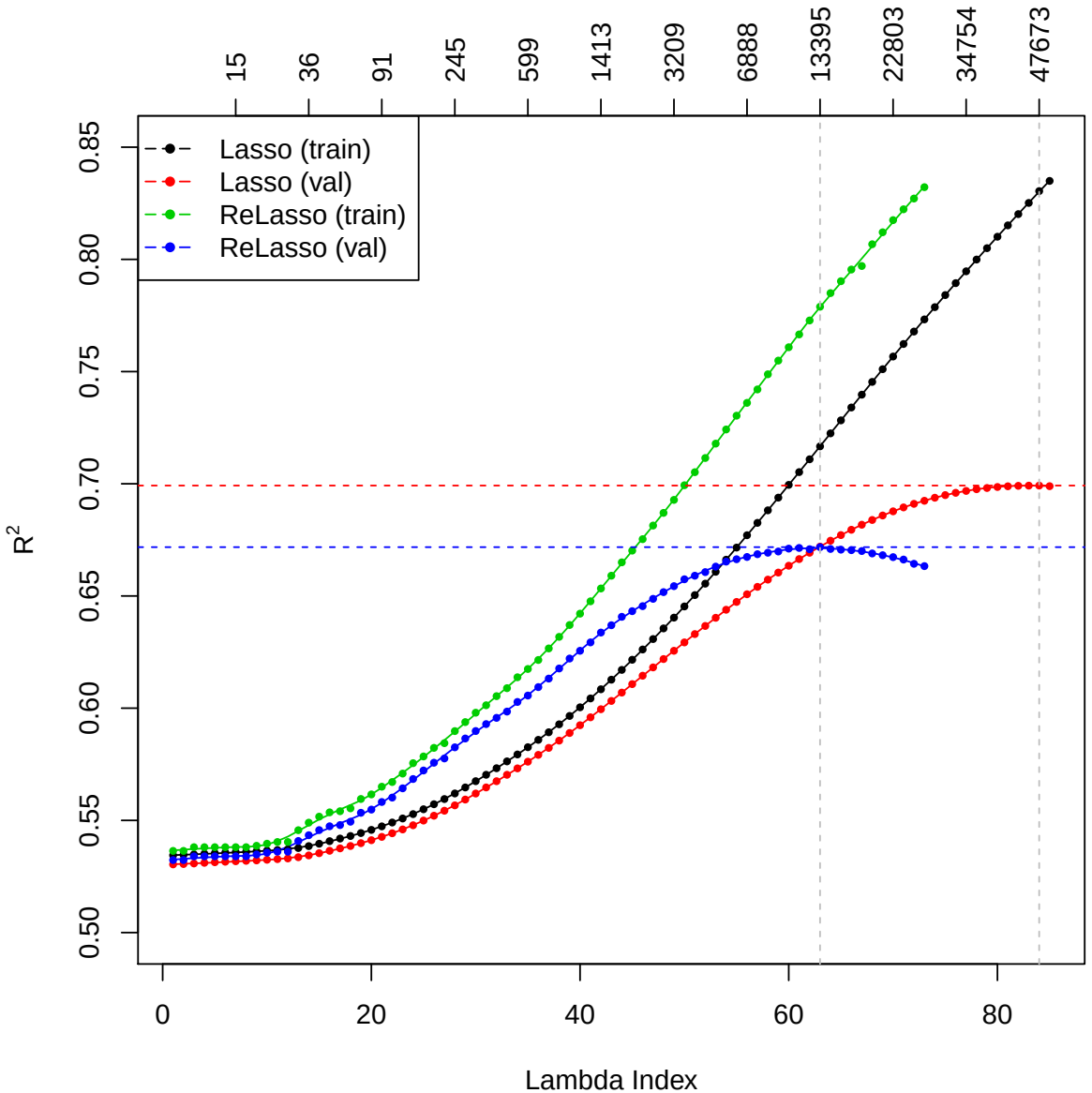
Subject Area

Bioinformatics

Subject Areas

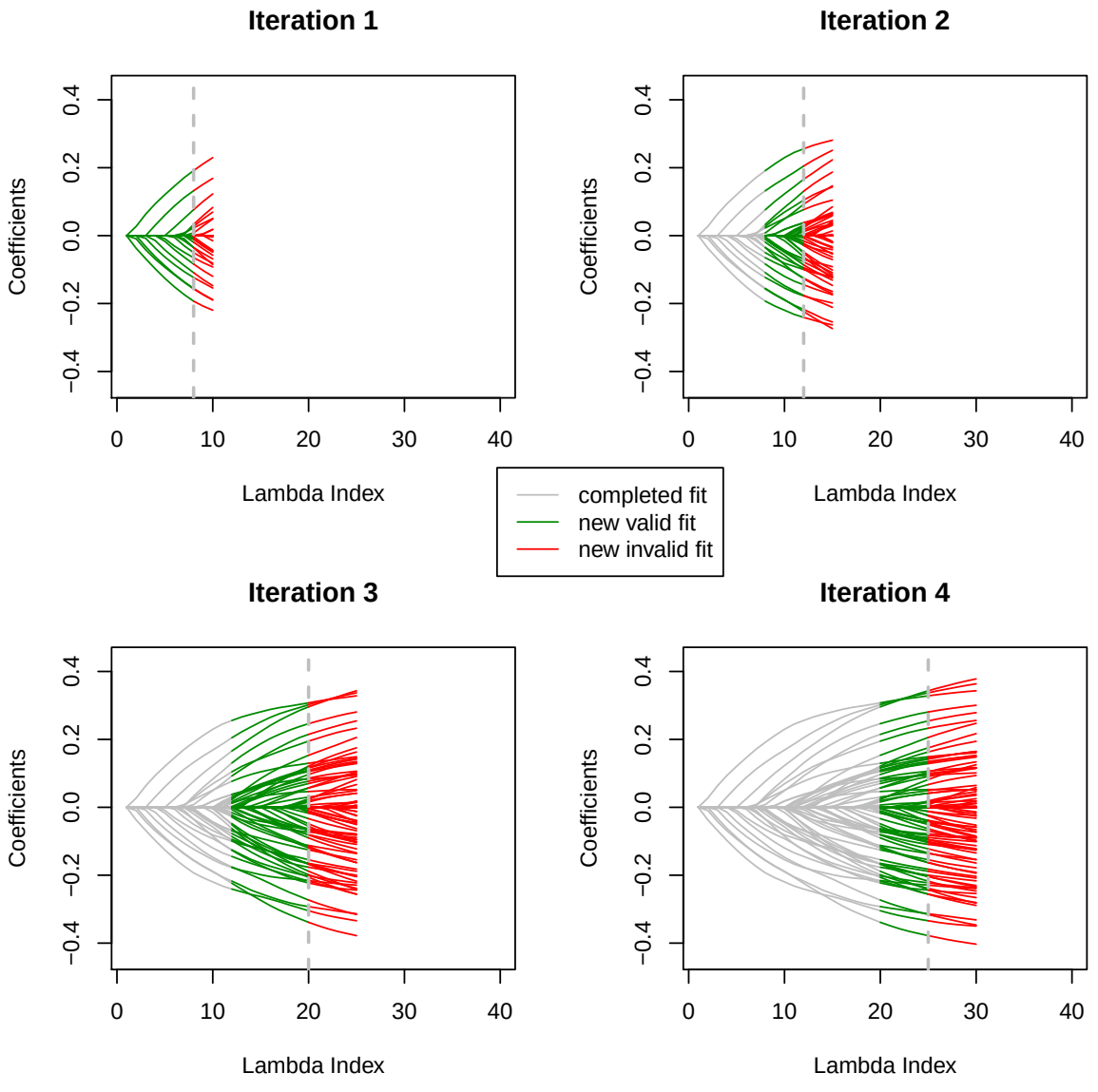
All Articles

UK Biobank Genomic Data



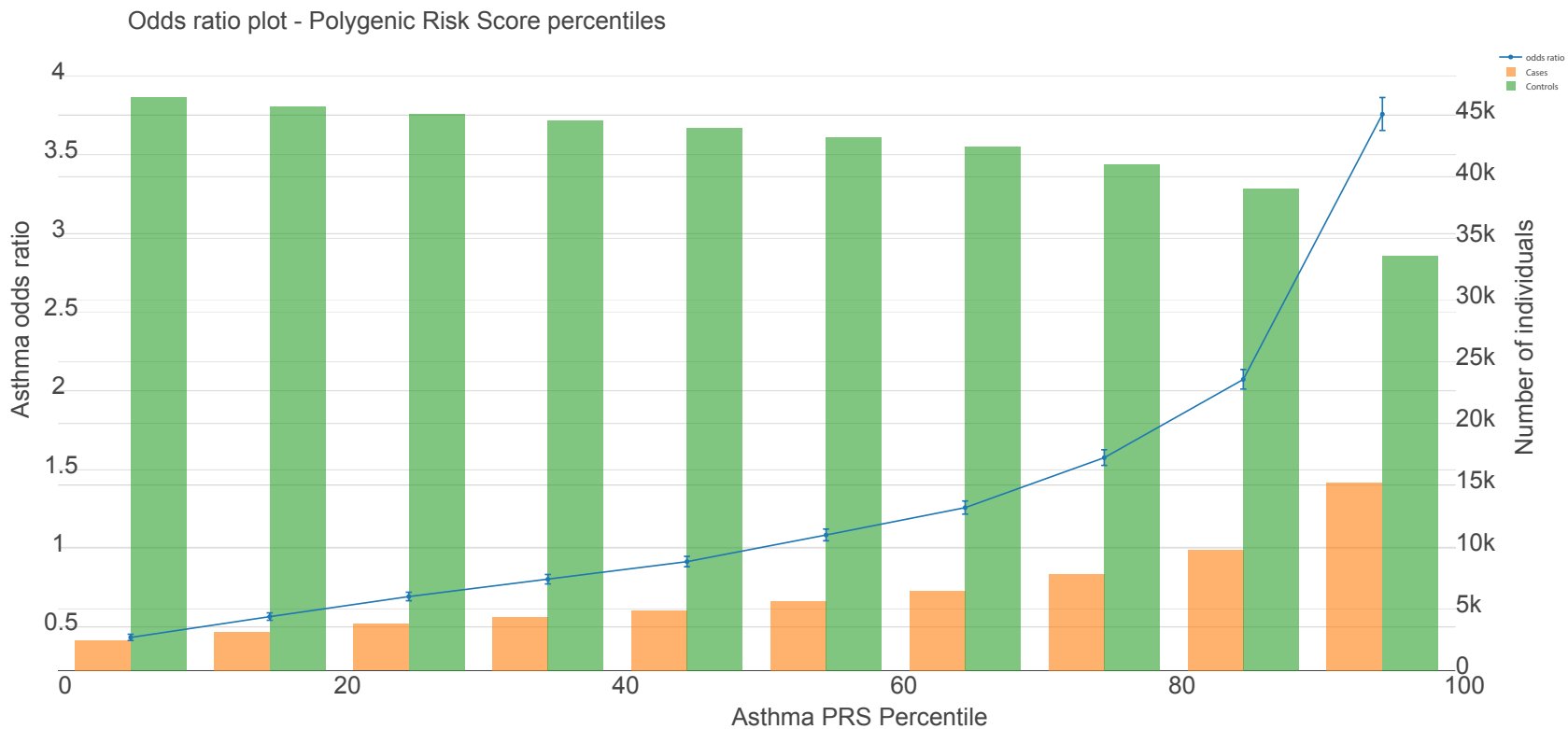
Build phenotype prediction models using genotypes at 800K loci (SNPs) for 500K individuals.
Here we predict Standing Height using lasso model via `glmnet`, and achieve $R^2 = 70\%$.

Algorithmic details in pictures

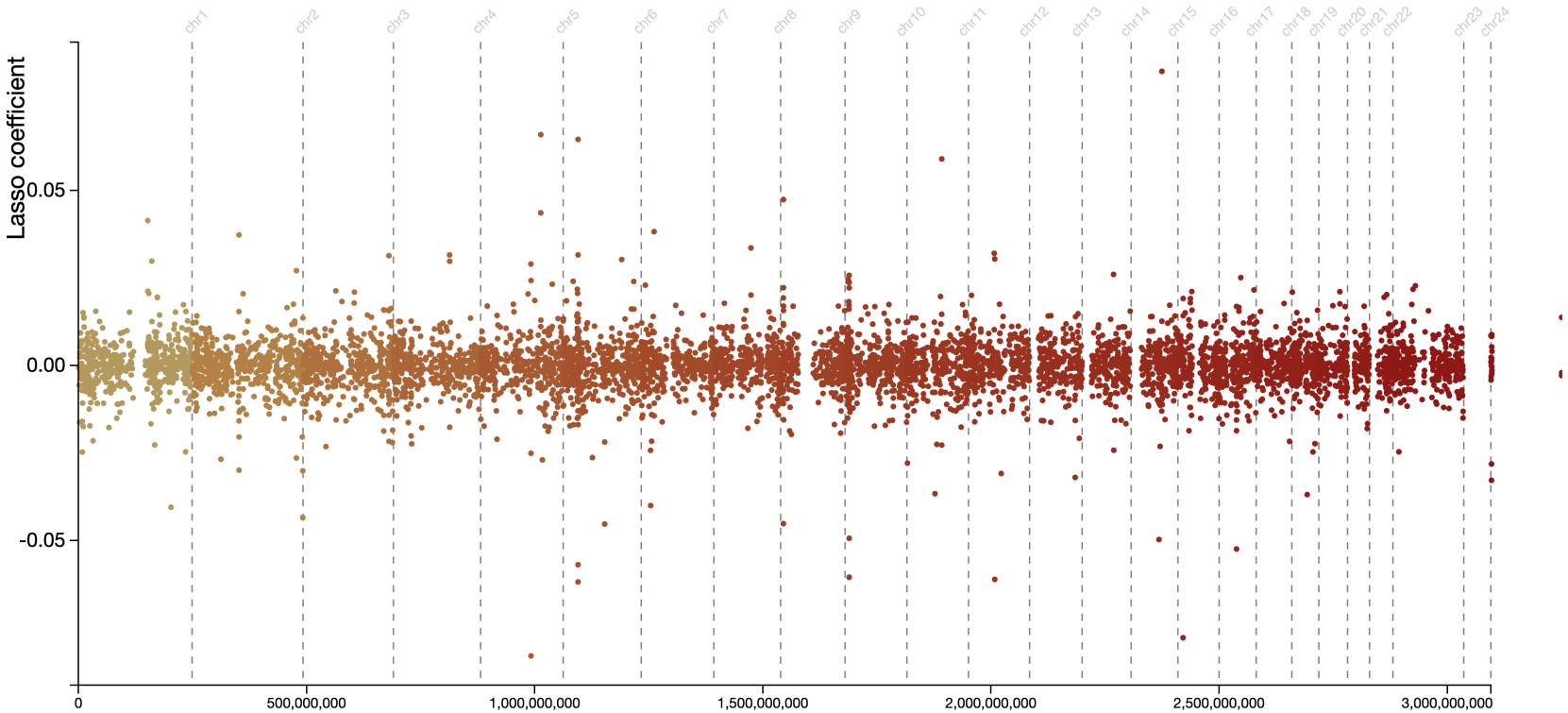


UK Biobank — Asthma

Build polygenic risk scores for disease phenotypes

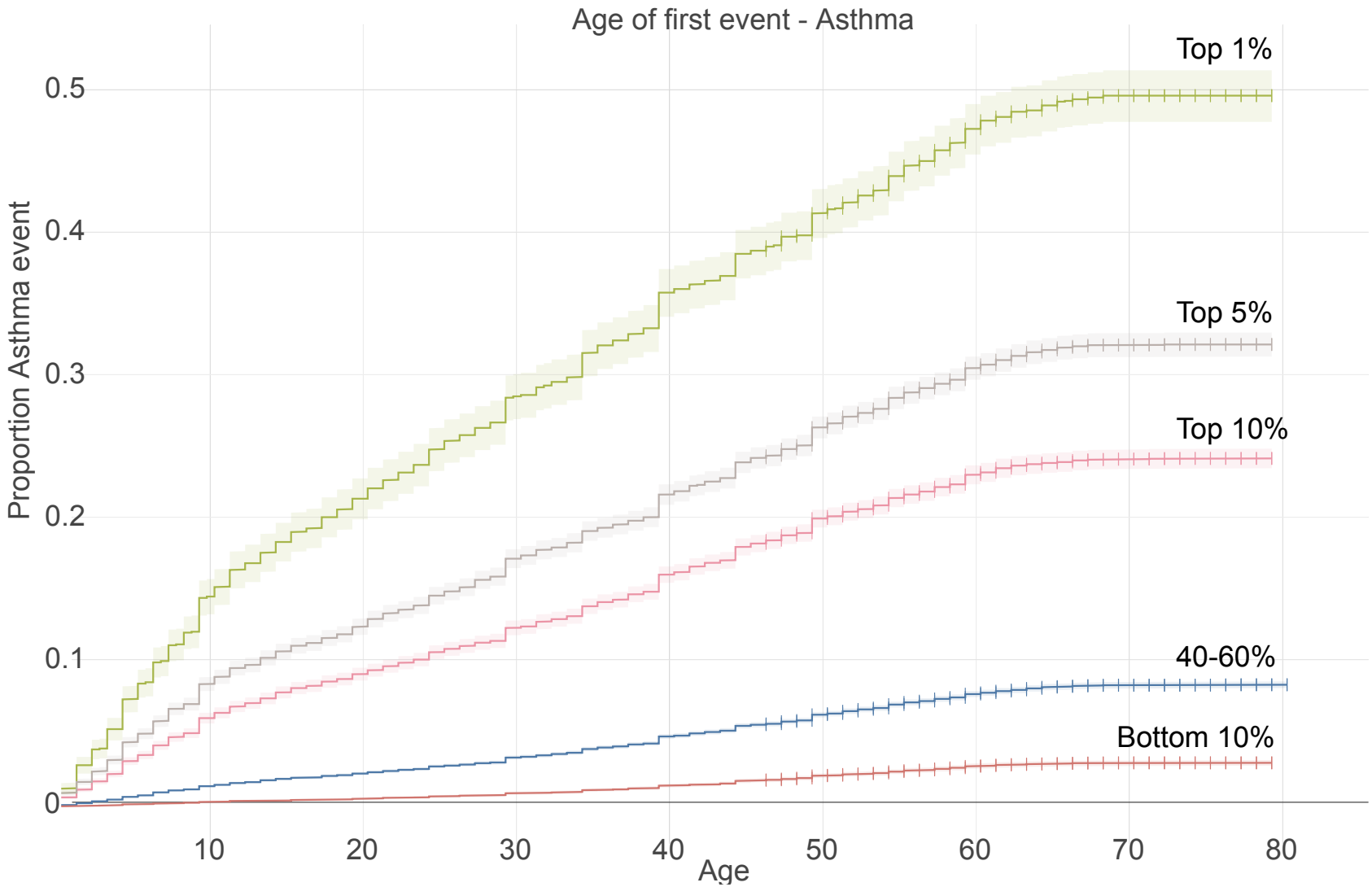


UK Biobank — Asthma



Note: only 6347 variants in active path for Lasso are included in th coefficient plot.

UK Biobank — Asthma



UK Biobank — Asthma

Variable	Level	N	Events	Hazard Ratio	*P* value
Polygenic risk score	40-60%	62139	5017	Reference	
	Bottom 10%	31036	896	0.36 (0.34, 0.39)	<0.001
	Top 1%	3085	1494	7.31 (6.90, 7.74)	<0.001
	Top 10%	15480	3636	3.06 (2.93, 3.20)	<0.001
	Top 5%	12403	3894	4.27 (4.09, 4.45)	<0.001
Eosinophil threshold	below .18	73655	6392	Reference	
	above .18	50488	8545	1.61 (1.56, 1.67)	<0.001
Sex	Female	66292	8330	Reference	
	Male	57851	6607	0.87 (0.85, 0.90)	<0.001
Body Mass Index	Below 25	40969	4414	Reference	
	25-30	53273	6220	1.04 (1.00, 1.08)	0.06
	30-35	21600	2880	1.12 (1.07, 1.17)	<0.001
	35-40	6010	954	1.26 (1.17, 1.35)	<0.001
	Above 40	2291	469	1.59 (1.44, 1.75)	<0.001

More on `glmnet`

Current version 4.02. Check out the website at `glmnet.stanford.edu`, and find the four vignettes there.

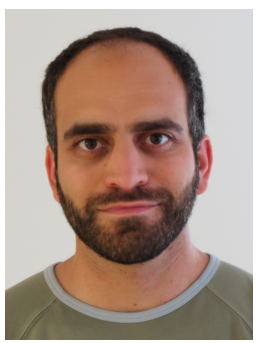
- All GLM `family()` choices are accommodated, along with the original hardwired families. e.g. `family=binomial(link="probit")`, `family=negative.binomial(theta=5)`
- New `relax=TRUE` argument which allows for refitting the active set without regularization.
- C index for survival models
- New functions for assessing fits
- Convenient functions for fitting big GLMs
- Progress bar via `trace=TRUE`

Predicting the Pathogenicity of Missense Variants

Goal: prioritize list of candidate genes for prostate cancer

Joint work with Epidemiology colleagues Weiva Sieh, Nilah Monnier Ioannidis, Joe Rothstein, Alice Whittemore, $\cdots$






REVEL — rare exome variant ensemble learner

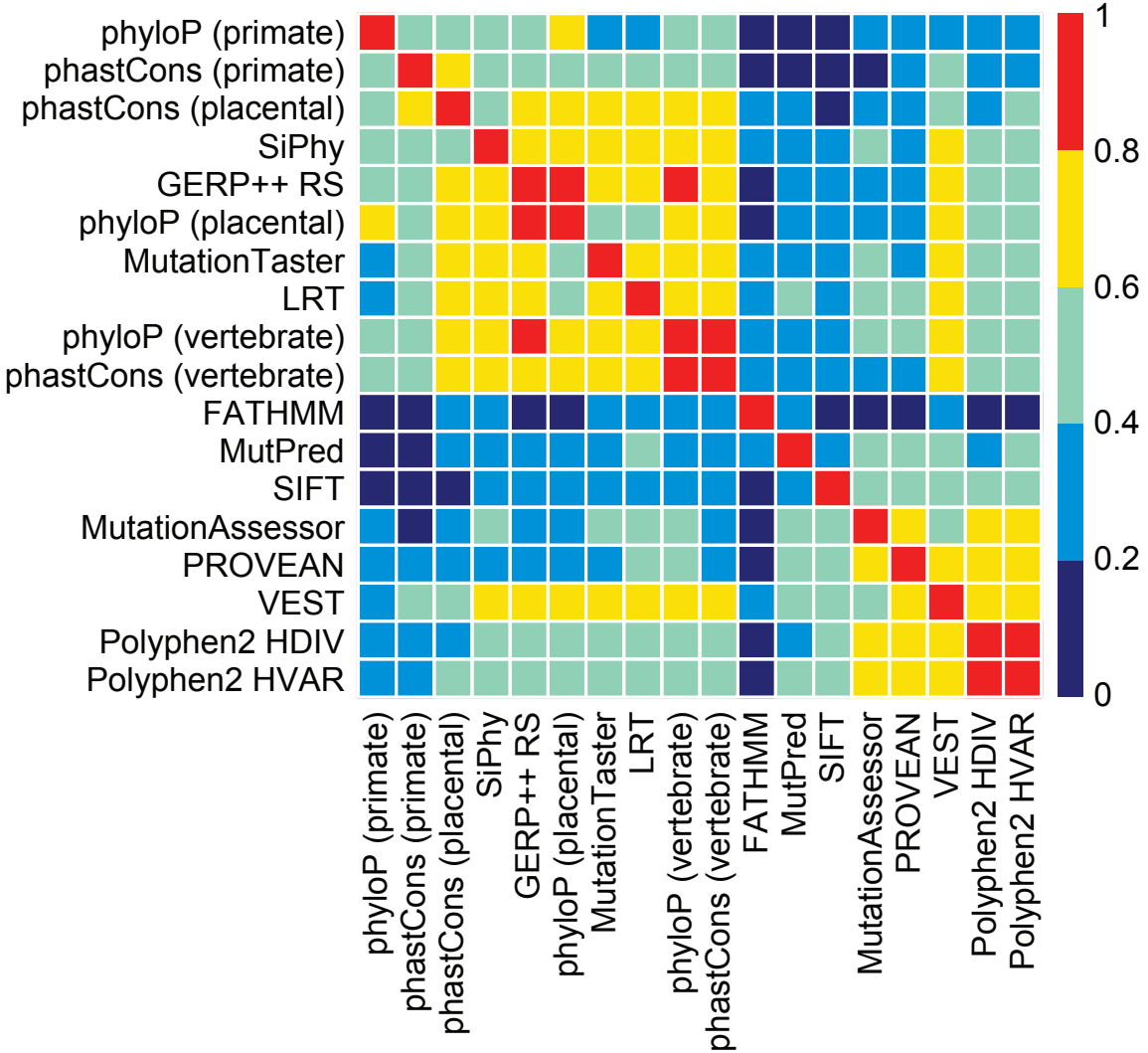
Ioannidis, N., $\cdots$, Sieh, W. (Oct 2016) Amer. J. Human Genetics

Approach

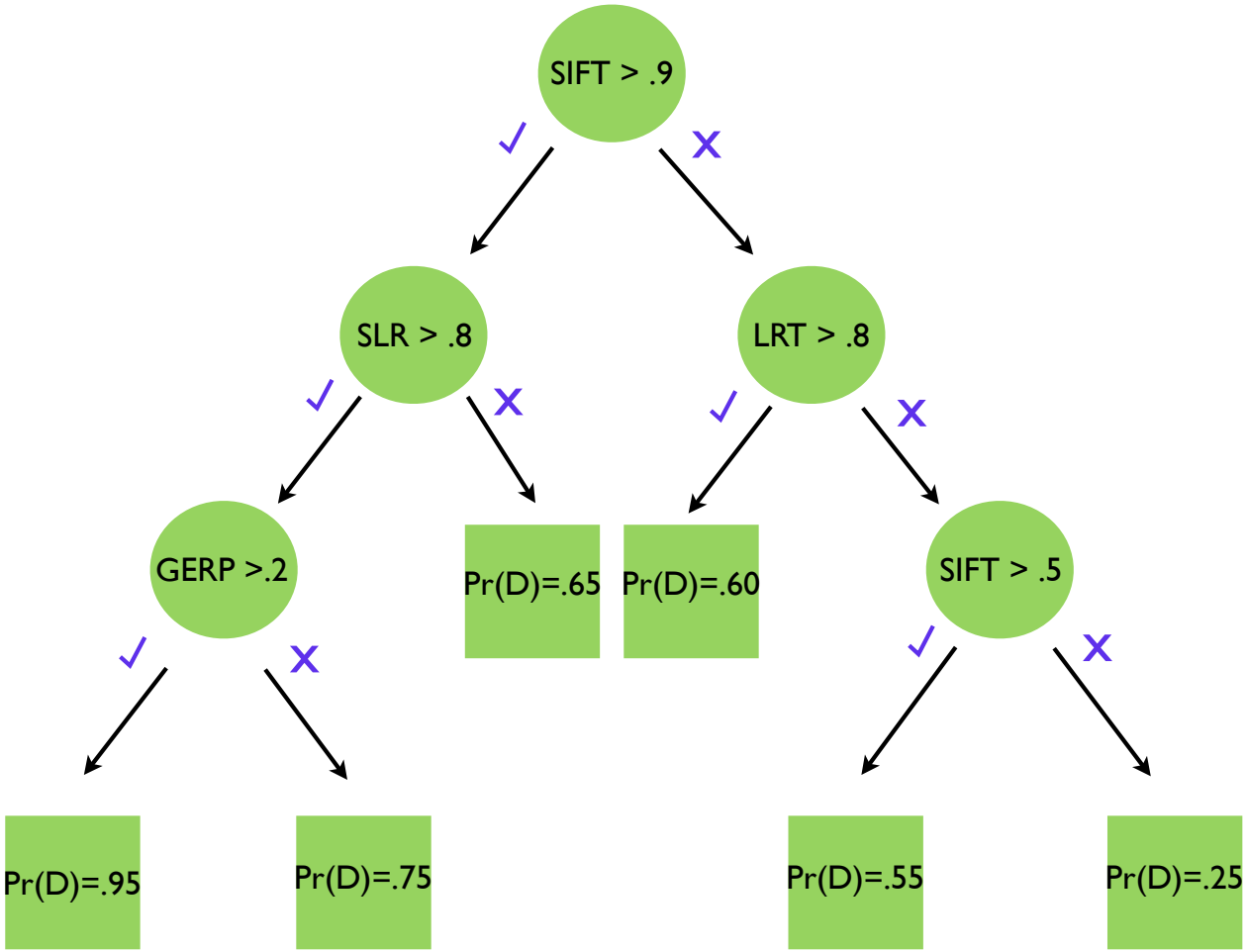
- A number of existing scores for disease status do not always agree (e.g SIFT, MutPred).
- Idea is to use a Random Forest algorithm to integrate these scores into a single consensus score for predicting disease.
- We will use existing functional prediction scores, conservation scores, etc as features — 12 features in all.
- Data acquired through Human Gene Mutation Database, SwissVar and ClinVar.

	Neutral	Disease
Train	123,706	6,182
Test	2,406	1,953

Correlation of Features



Decision Trees



Trees use the features to create subgroups in the data to refine the estimate of disease. Shallow trees are too coarse/inaccurate.

Random Forests

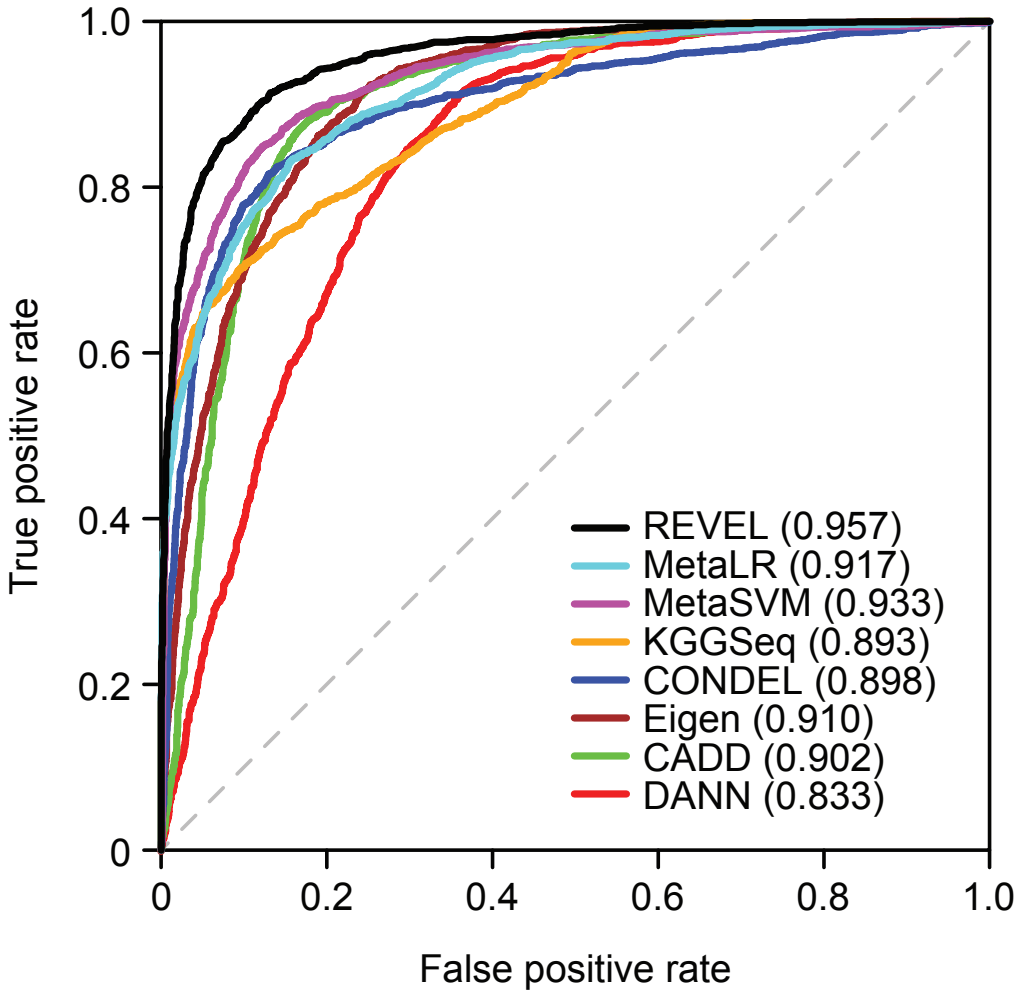
Leo Breiman (1928–2005)



- Deep trees (fine subgroups) are more accurate, but very noisy.
- Idea: fit many (1000s) different and very-deep trees, and average their predictions to reduce the noise.
- How to get different trees?
  - Grow trees to bootstrap subsampled versions of the data.
  - Randomly ignore variables as candidates for splits.

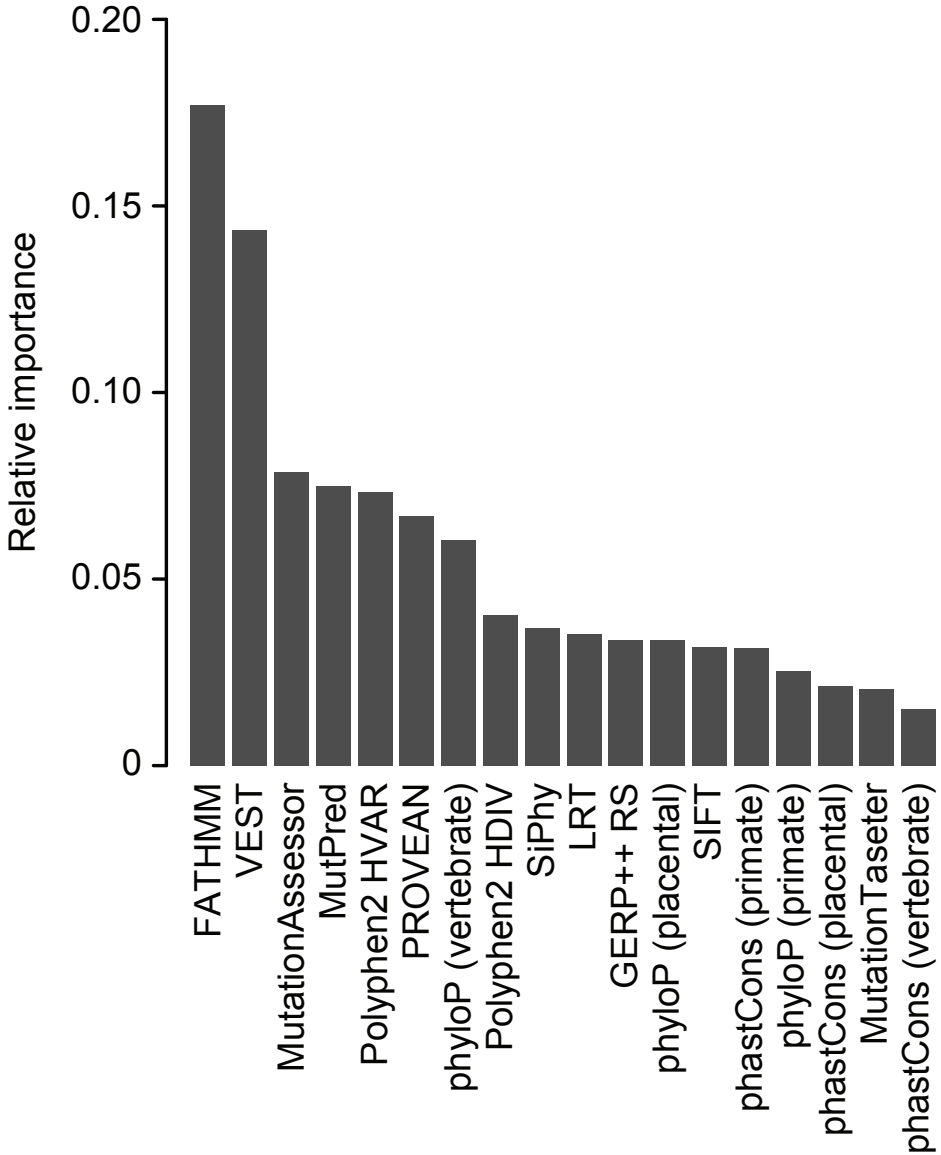
Random Forests are very effective and give accurate predictions. They are automatic, and give good CV estimates of prediction error (for free!). R package `RandomForest`.

Results for REVEL

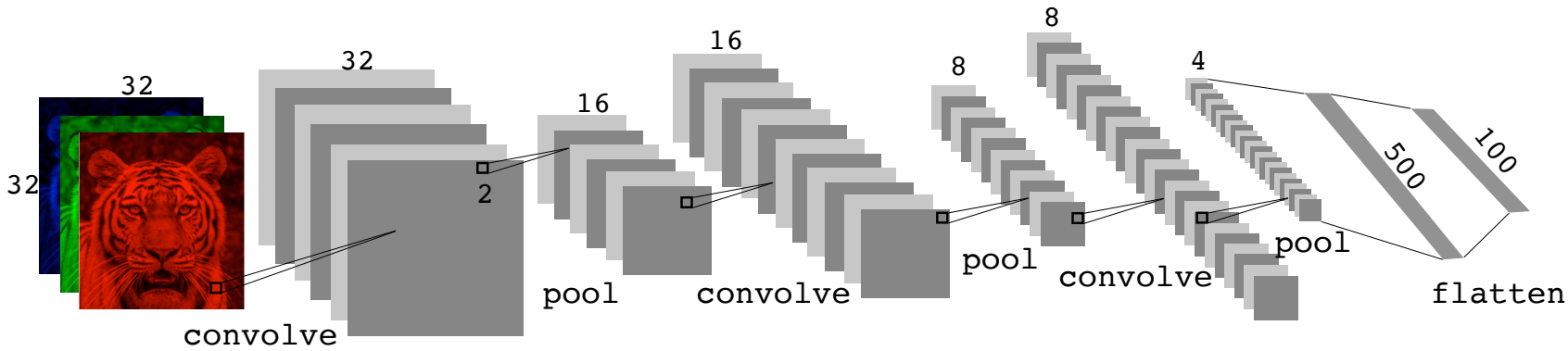


Performance evaluated on independent test set, and REVEL compared with 7 other ensemble competitors.

Feature Importance



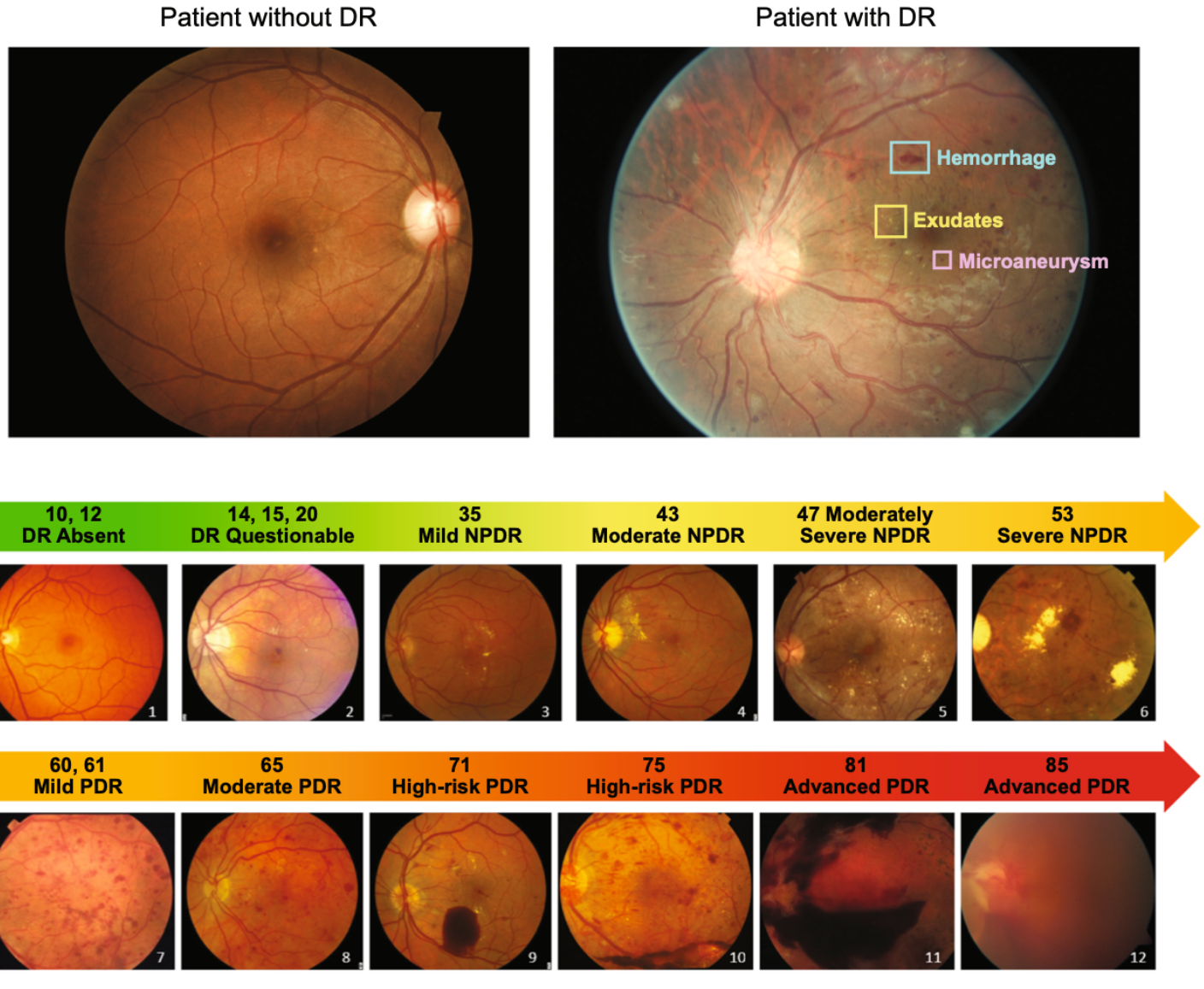
Deep Learning



Convolutional neural network (CNN) for image modeling.

- Big successes with image classification. Revolutionary!
- Many digital images in healthcare: mammograms, MRI scans, X-rays, optical scans, skin photographs, ...
- Modern software tools accessible to all (e.g. `keras` and `tensorflow` in R).

Diabetic Retinopathy



Two recent articles (2019-20)

npj | Digital Medicine

www.nature.com/npjdigitalmed

ARTICLE OPEN

Deep learning algorithm predicts diabetic retinopathy progression in individual patients

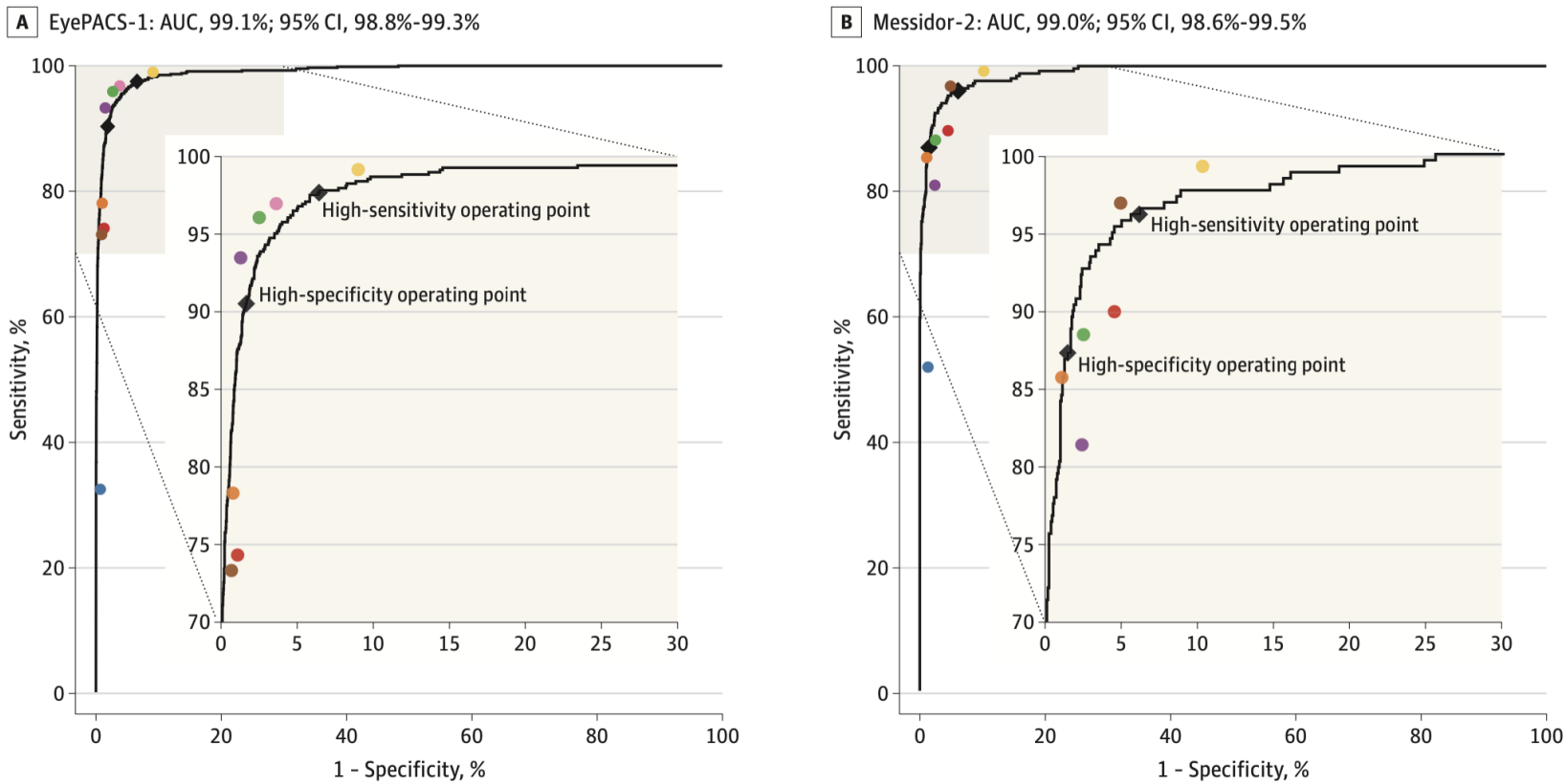
Filippo Arcadu[1,2], Fethallah Benmansour[1,2], Andreas Maunz[1,2], Jeff Willis[3,4], Zdenka Haskova[3,4,7*] and Marco Prunotto[2,5,6,7*]

JAMA | Original Investigation | INNOVATIONS IN HEALTH CARE DELIVERY

Development and Validation of a Deep Learning Algorithm for Detection of Diabetic Retinopathy in Retinal Fundus Photographs

Varun Gulshan, PhD; Lily Peng, MD, PhD; Marc Coram, PhD; Martin C. Stumpe, PhD; Derek Wu, BS; Arunachalam Narayanaswamy, PhD; Subhashini Venugopalan, MS; Kasumi Widner, MS; Tom Madams, MEng; Jorge Cuadros, OD, PhD; Ramasamy Kim, OD, DNB; Rajiv Raman, MS, DNB; Philip C. Nelson, BS; Jessica L. Mega, MD, MPH; Dale R. Webster, PhD

Validation Set Performance for Referable Diabetic Retinopathy



[A] 8788 images, 8 experts [B] 1745 images, 7 experts

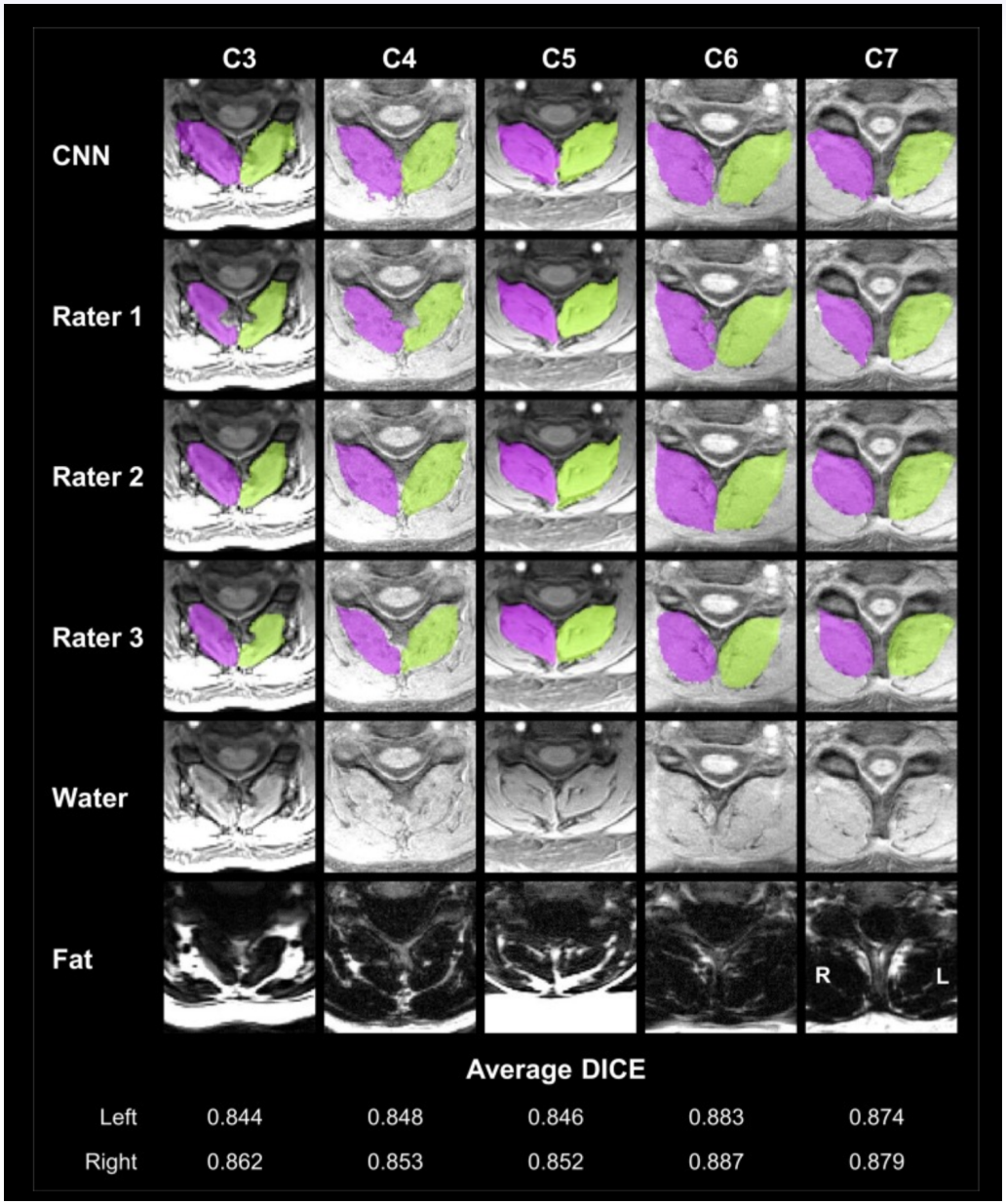


Image Segmentation

Deep Learning CNN on MRI scans

Identify deep cervical spine extensors and measure muscle fat infiltration after whiplash injury

Adapt pretrained network

Compare with expert raters

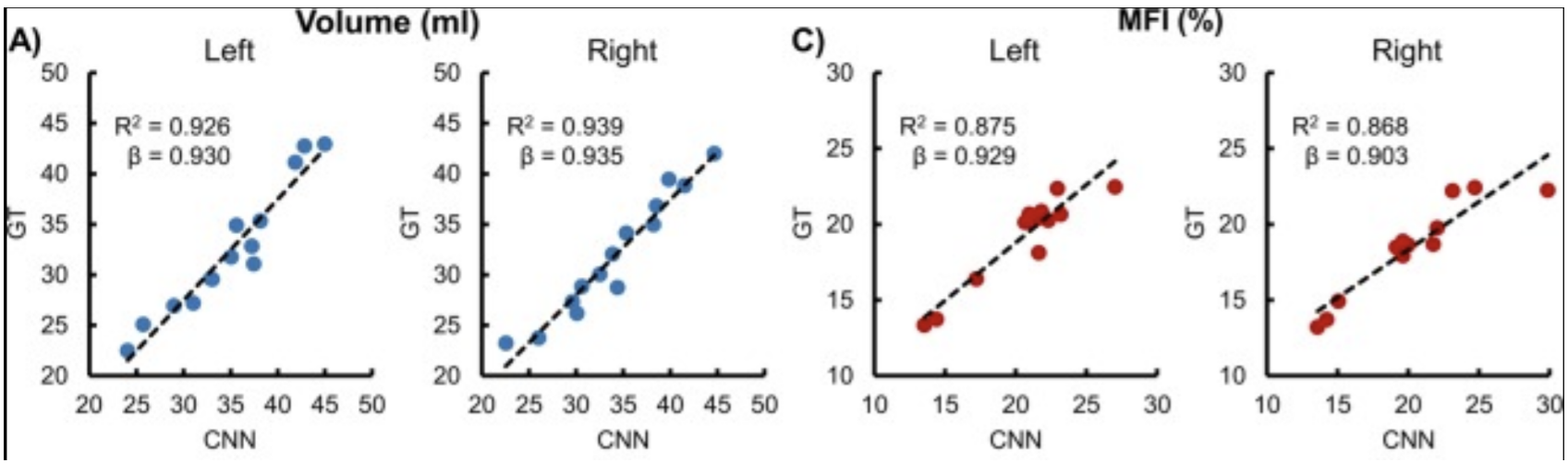
› Sci Rep. 2019 May 28;9(1):7973. doi: 10.1038/s41598-019-44416-8.

Deep Learning Convolutional Neural Networks for the Automatic Quantification of Muscle Fat Infiltration Following Whiplash Injury

Kenneth A Weber [1], Andrew C Smith [2], Marie Wasielewski [3], Kamran Eghtesad [4], Pranav A Upadhyayula [4], Max Wintermark [5], Trevor J Hastie [6], Todd B Parrish [7], Sean Mackey [4], James M Elliott [3 8 9]

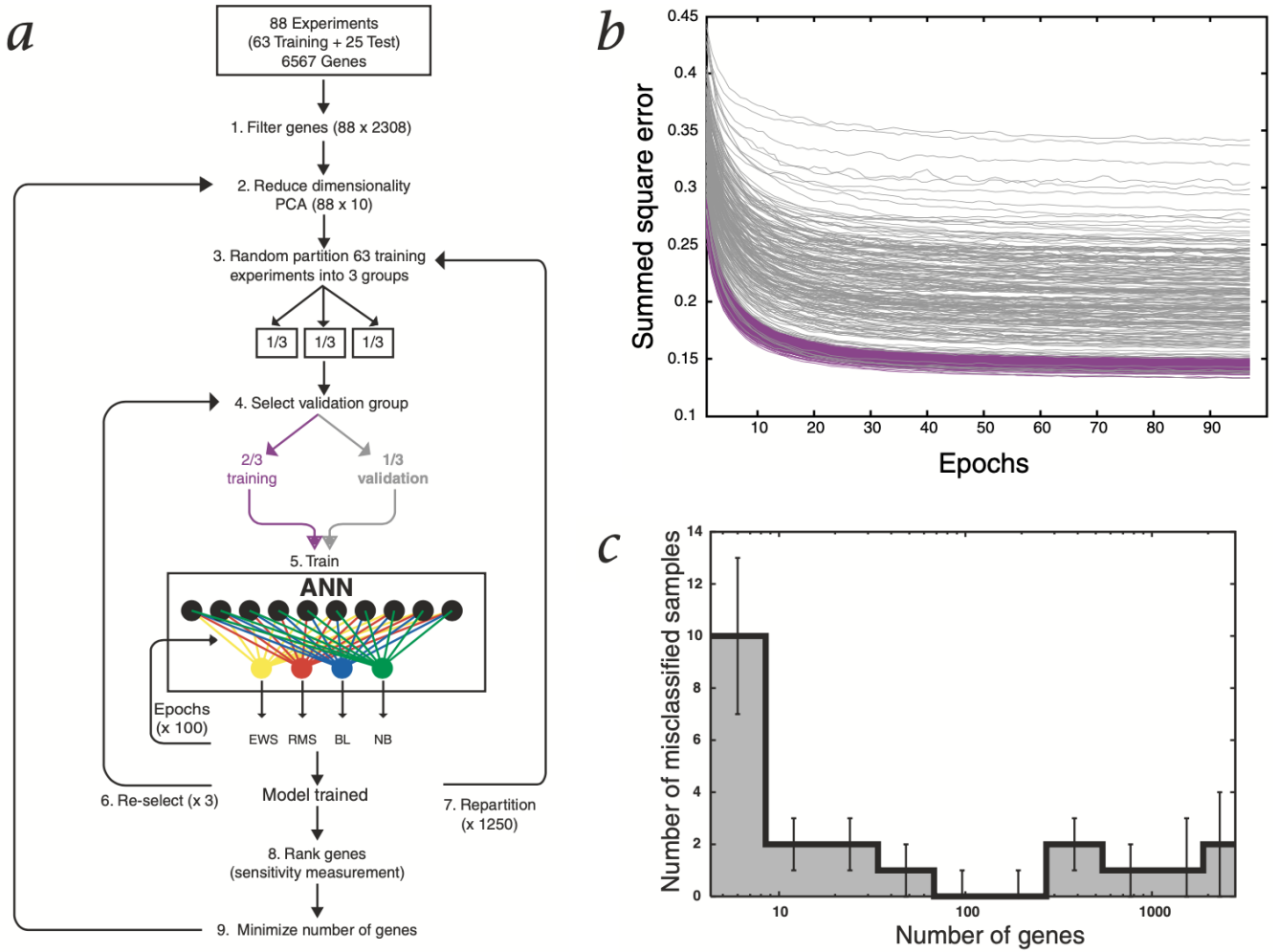
Affiliations + expand

PMID: 31138878 PMCID: PMC6538618 DOI: 10.1038/s41598-019-44416-8



GT = Ground Truth (average of three raters)
CNN = Convolutional Neural Network

Overkill - article in Nature Medicine 2001



Simple linear method does as well, and does gene selection

Elements of Statistical Learning, Section 18.2

Summary

- We saw examples of Lasso, Random Forest, and Deep Learning.
- We saw an example where a Neural Networks was overkill, and a simpler model was sufficient.

Summary

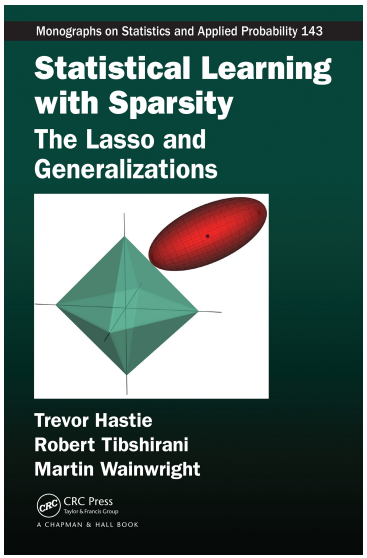
- We saw examples of Lasso, Random Forest, and Deep Learning.
- We saw an example where a Neural Networks was overkill, and a simpler model was sufficient.

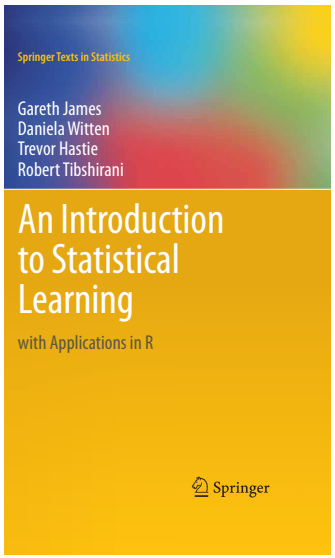
$\cdots$ and now for some cheap marketing $\ldots$

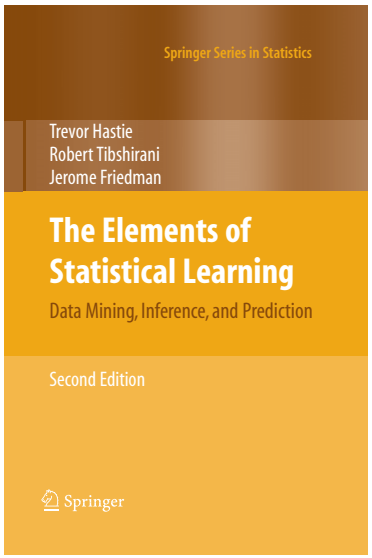
Summary

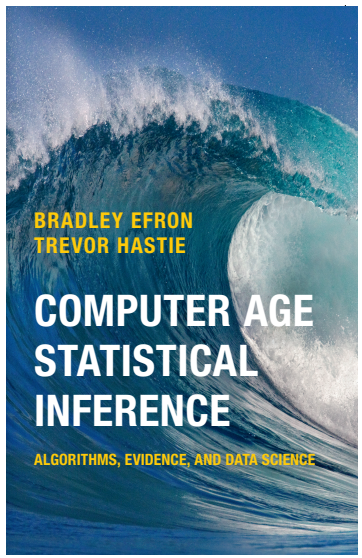
- We saw examples of Lasso, Random Forest, and Deep Learning.
- We saw an example where a Neural Networks was overkill, and a simpler model was sufficient.

$\cdots$ and now for some cheap marketing $\ldots$





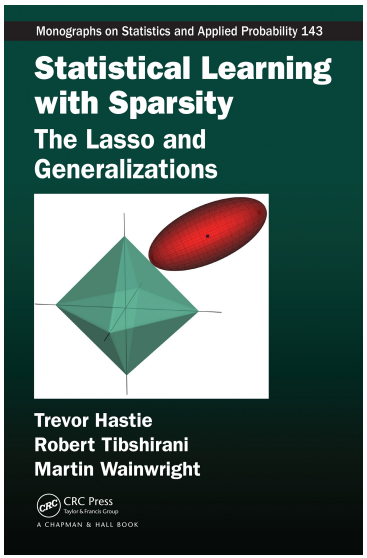


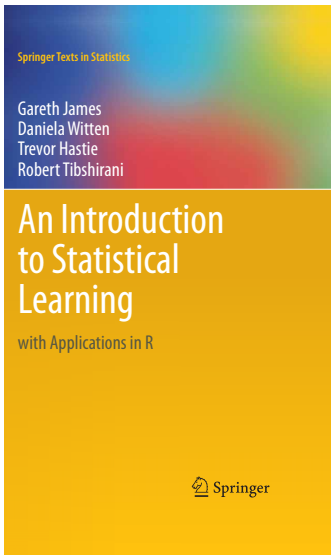


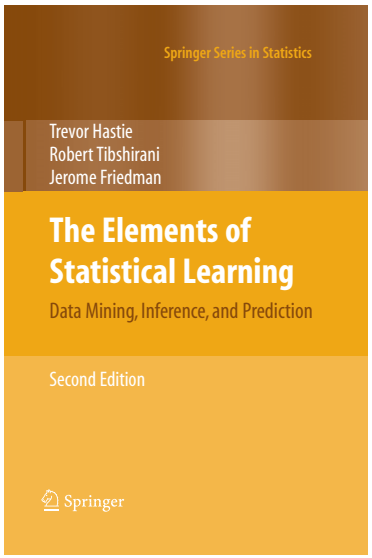
Summary

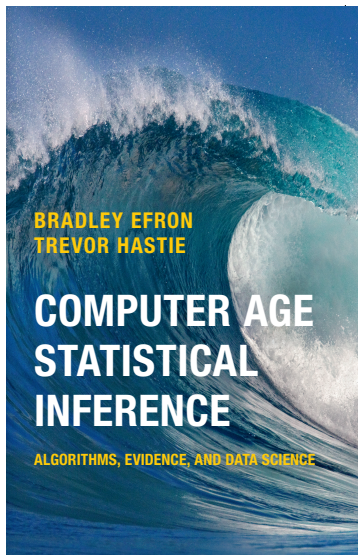
- We saw examples of Lasso, Random Forest, and Deep Learning.
- We saw an example where a Neural Networks was overkill, and a simpler model was sufficient.

$\cdots$ and now for some cheap marketing $\ldots$









Thank you!