ML/AI Classification

Classification

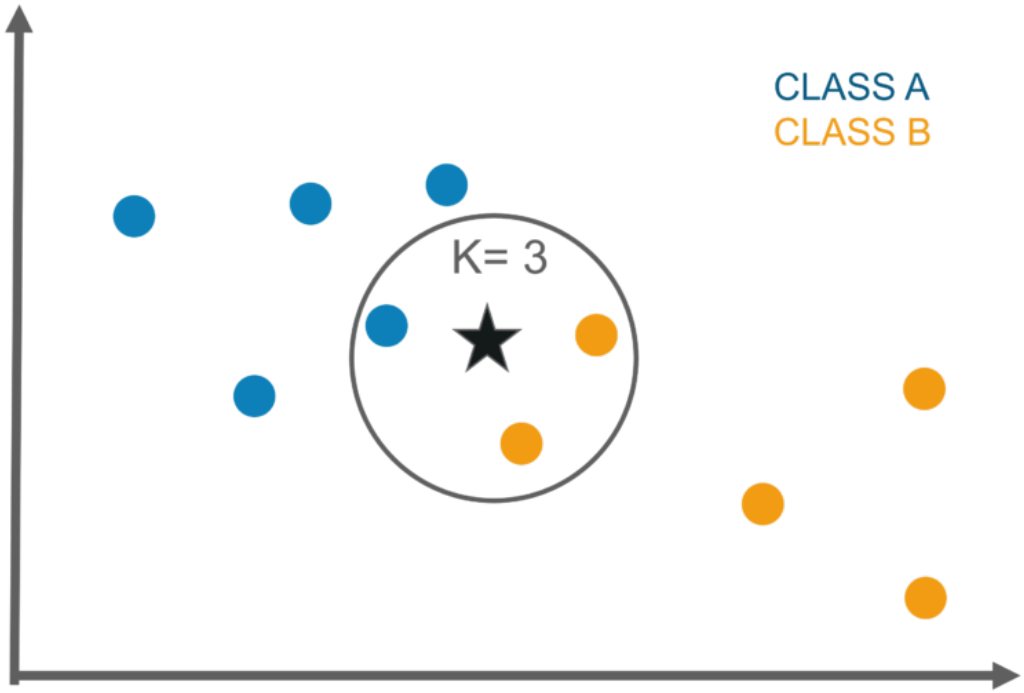
- the basic problem is given a data set, with features and classess for some set of objects build a classifier that uses the features to make predictions for new data points that arrive, where we only know the features, but not the classes
- provide reasonable estimates of the misclassifcation rate

Supervised Machine Learning

- this problem is also referred to as supervised machine learning
- no free lunch theorem (Wolport and Macready)

any two optimization algorithms are equivalent when their performance is averaged across all possible problems

Classification



some important thought

https://www.edureka.co/blog/k-nearest-neighbors-algorithm/

ML

- Model Assessment
- Bias – Variance and Model complexity
- in general low complexity goes with higher error rates (as the model cannot adapt enough), increasing complexity decreases the error rate on the training set, but the test set error starts high, goes low and then rebounds

ML

- Model selection: estimating the performance of different models to choose the best one
- Model assessment: having chosen a final model, what is its error rate on new data
- We need to partition the data into three mutually exclusive sets Train, Validation and Test.
  - the Test set will be used to answer the Model assessment question and should be totally sequestered and then processed as you would some new data that was not part of the original experiment

Cross-Validation

- usually we do not have enough data to have both a training and a validation set
  - when these are too small the error estimates are highly variable
- in those cases we can use cross-validation (usually 5 or 10 fold, depending)
  - in 5 fold, we split our data into 5 mutually exclusive sets, and then we leave one of those out, build the model in the remaining 4, compute the error rate on the held out set and repeat
  - the CV error rate is then the average of these
  - you should also report the sd of your estimate….

From Elements of Statistical Learning

- wrong way to do CV

  1) screen predictors: find a subset of good predictors in some way

  2) using this subset of predictors build a classifier

  3) use CV to estimate the unknown tuning parameters in your classifier and estimate the prediction error of the final classifier

Problem: you have used all the data in step 1 and that will yield an under estimate of the prediction error.

Example Golub leukemia from:

Selection bias in gene extraction on the basis of microarray gene-expression data

Christophe Ambroise† and Geoffrey J. McLachlan‡§

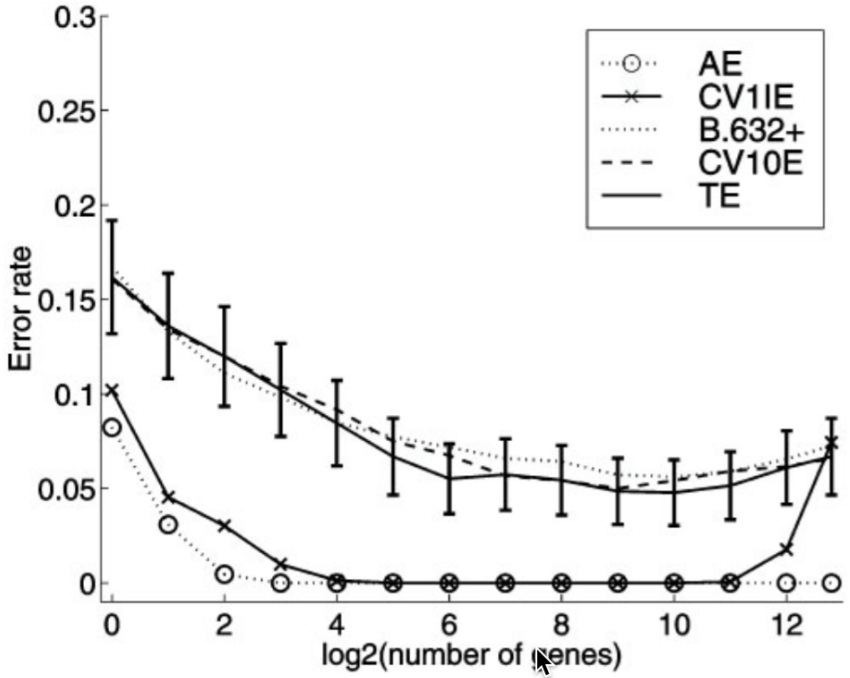


Fig. 2. Error rates of the SVM rule with RFE procedure averaged over 50 random splits of the 72 leukemia tissue samples into training and test subsets of 38 and 34 samples, respectively. TE, test error.

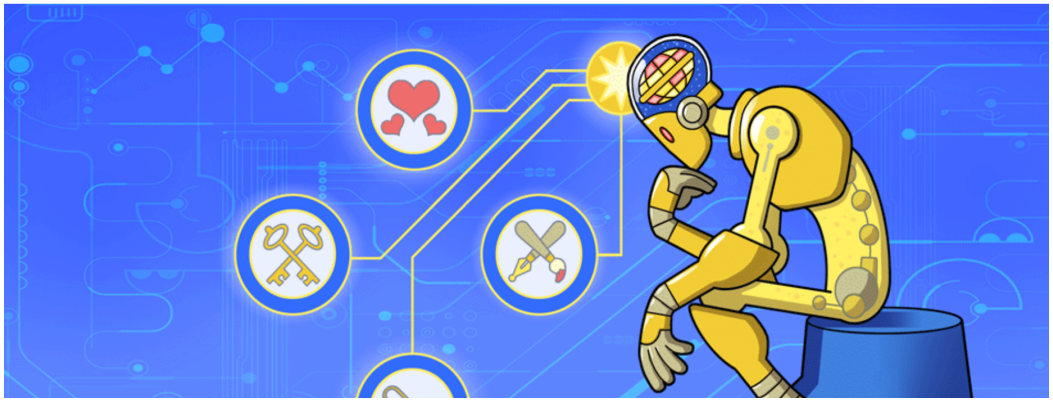
Model Selection

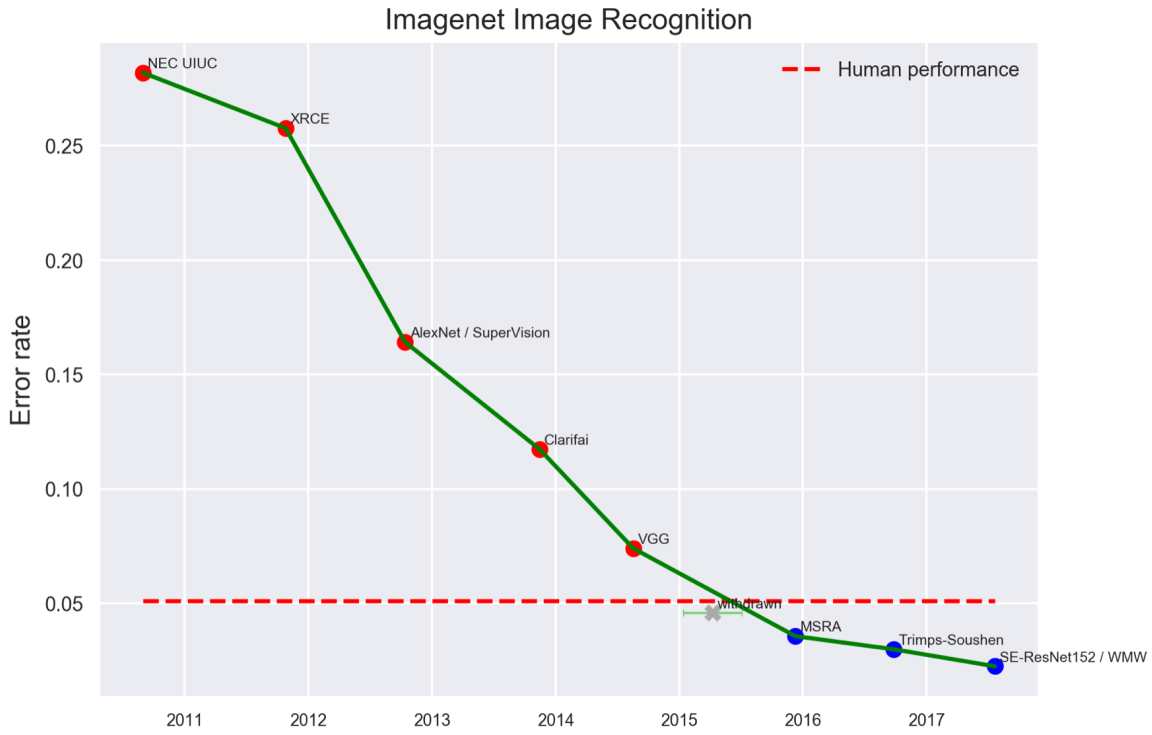
- one of the most important decisions is what *loss function* to use
- the loss function determines the cost of a misclassification
  - classification loss – 0 if correct class predicted, 1 if incorrect
  - regression/continuous loss: squared error between true value and predicted value
  - not all losses are equal – if one class is rare you will have to penalize mistakes for it more than for the other class

Machine Learning

- improvements in the approach over the past 10 years have yielded very good results
  - detecting specific artifacts in images
    - yield better phenotyping from slides and other images
    - digital pathology
  - vector representations of text (word embeddings)
    - yield an ability to search the literature in more interesting ways
  - classifications: hearing, vision, disease risk, etc
  - chatbots: companies like Lark, much of customer service

AI Progress Measurement





ML/AI applications in genetics/medicine

- many different applications and lots of papers outlining them
- trait prediction
- drug/vaccine/antibody design
- protein folding
- risk prediction
- covariate imputation – eg impute likely smoking status
- genetic imputation – also a ML/AI problem
- phasing

Joint Trait Prediction

- we wanted to study some traits that were highly related and unlikely to have a strong gene by environment interaction
- we chose skin color, eye color and hair color
  - these are known to be associated
  - they are known to have shared underlying genetic associations

Joint Trait Prediction

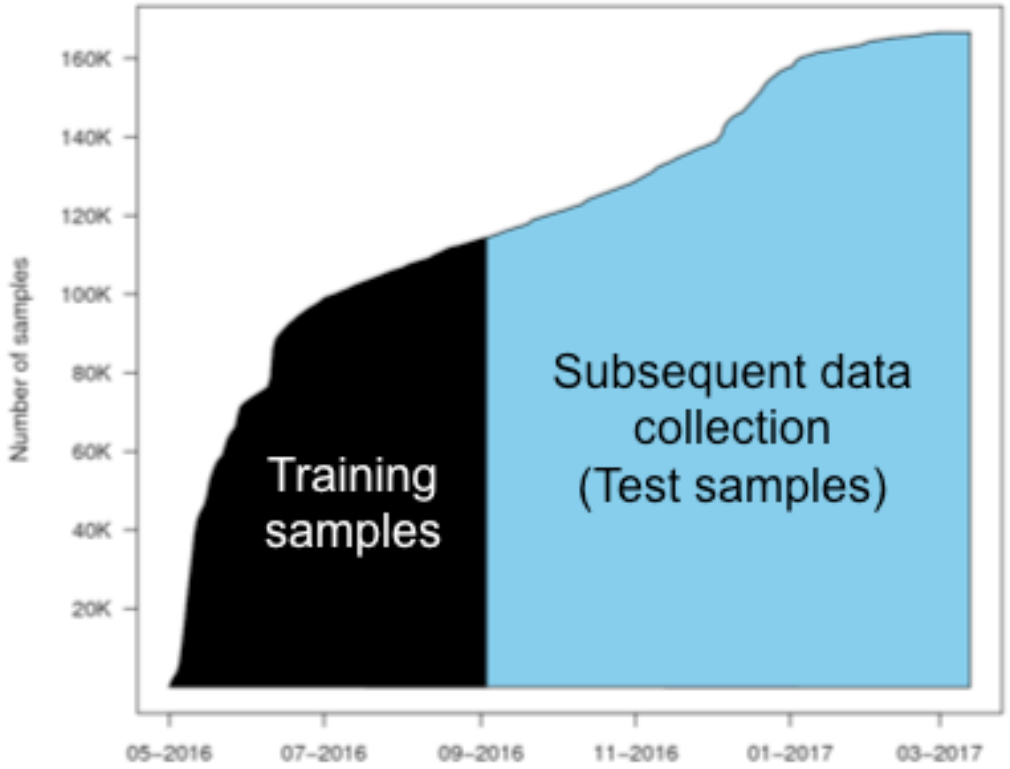


Figure 1: Data acquisition as a function of date. To train our pigmentation model, we selected data collected prior to September 2016, with subsequently collected data reserved for model validation purposes. All plots in this poster are made from the validation data

Our ML

- We adapted the “specialist-generalist” idea proposed in Warde-Farley, D. *et al.* (2014) Self-informed neural network structure learning. *arXiv*:1412.6563
- non-genetic covariates were the first 5 PCs from our genotypes, Age and Sex
- The final model was an ensemble of 10 models learned on random 80/20 splits of the training data (20% of data was used for model validation).

Our ML

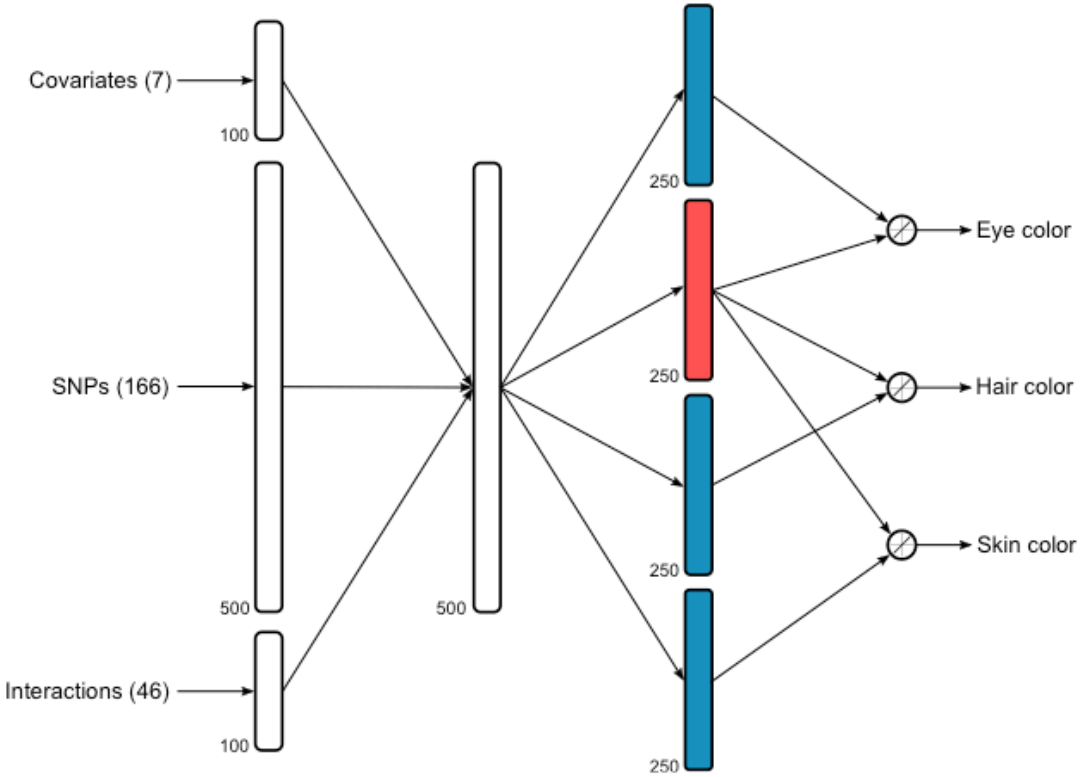
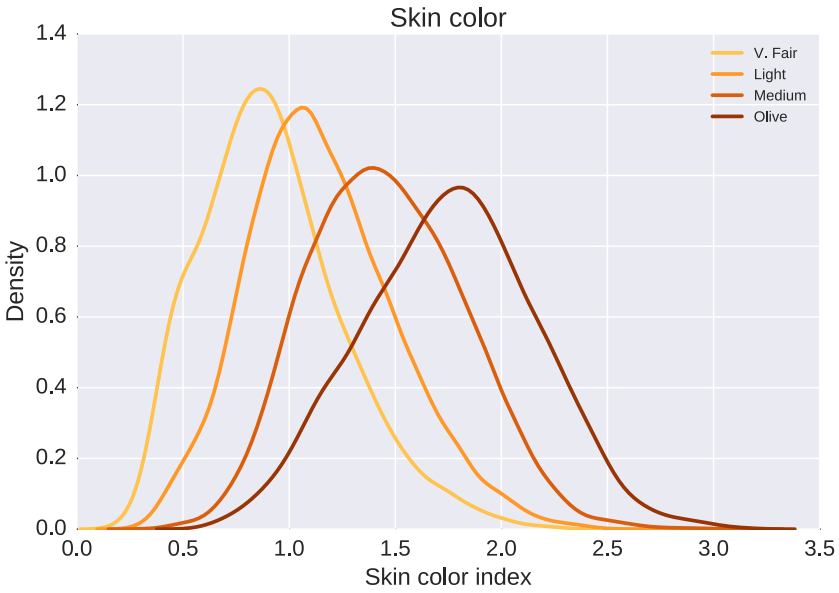


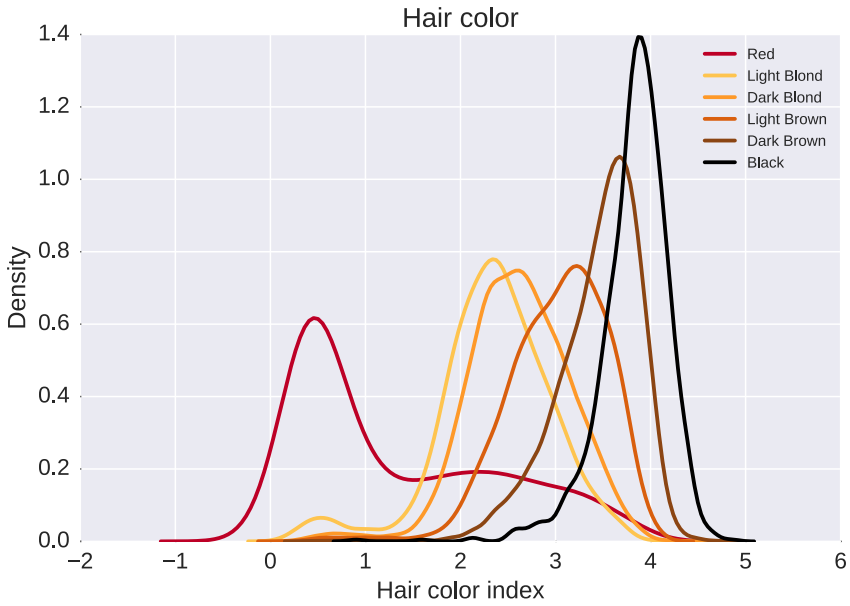
Figure 2: The neural network architecture used for learning the three pigmentation phenotypes. The number of nodes in each hidden layer is shown at the bottom left of that layer. The specialist hidden layers are shown in blue and the generalist hidden layer is shown in red. All hidden layer nodes have ReLU activations and the output nodes are linear. The numbers in parenthesis next to inputs denotes their dimensionality.

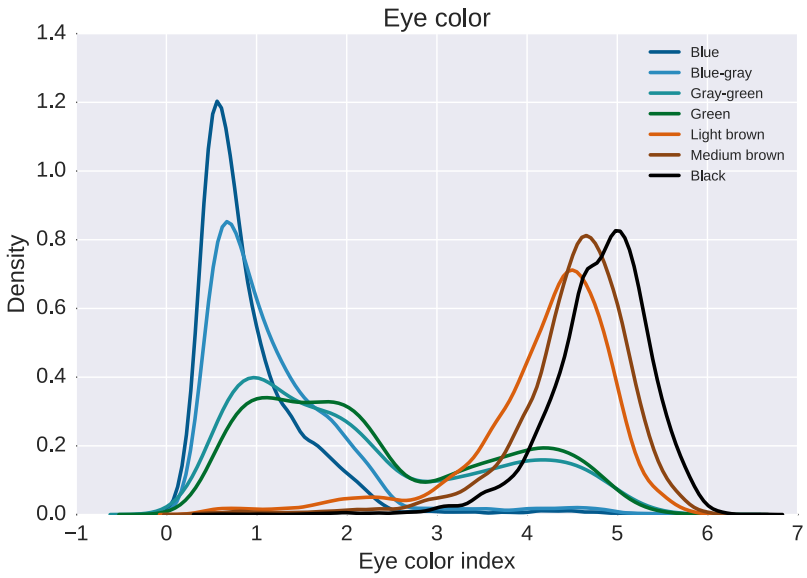
Results

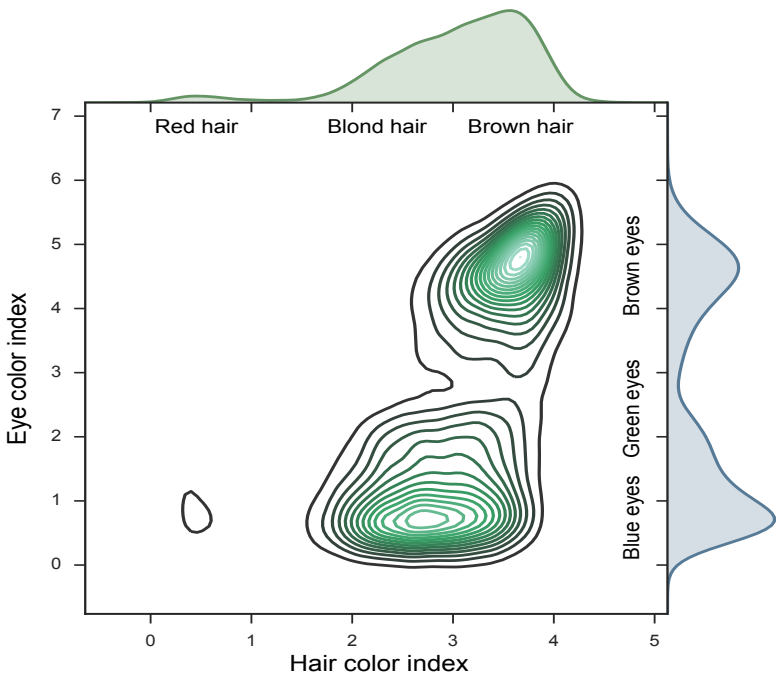
- We note that in spite of assuming arbitrary uniform spacing between phenotypes levels, the model puts the modes of the level distributions in an intuitively-meaningful order.

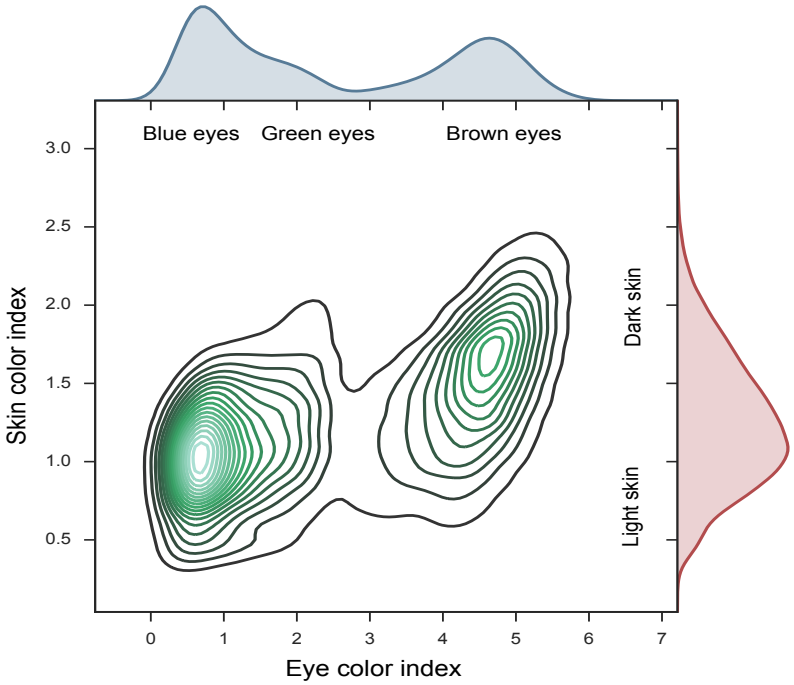
Distributions











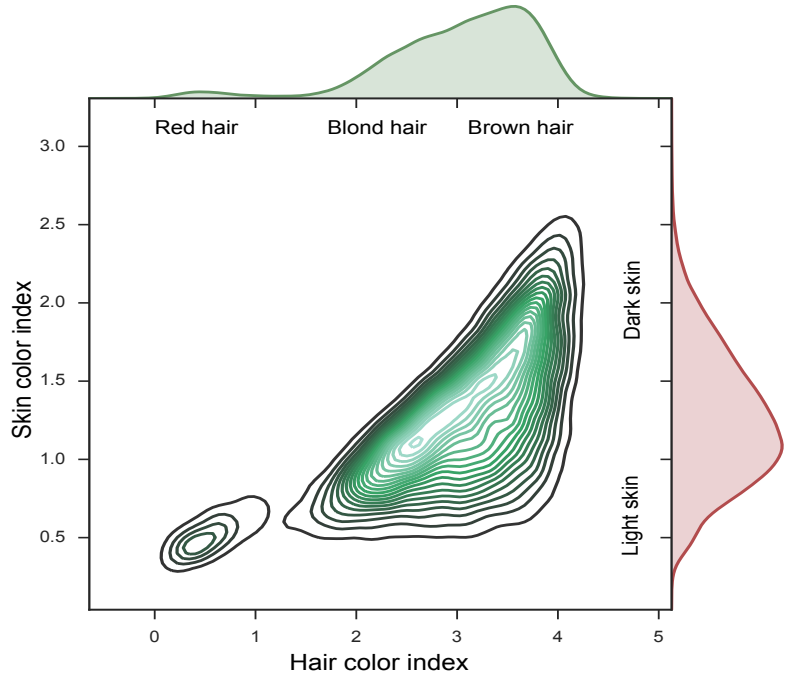
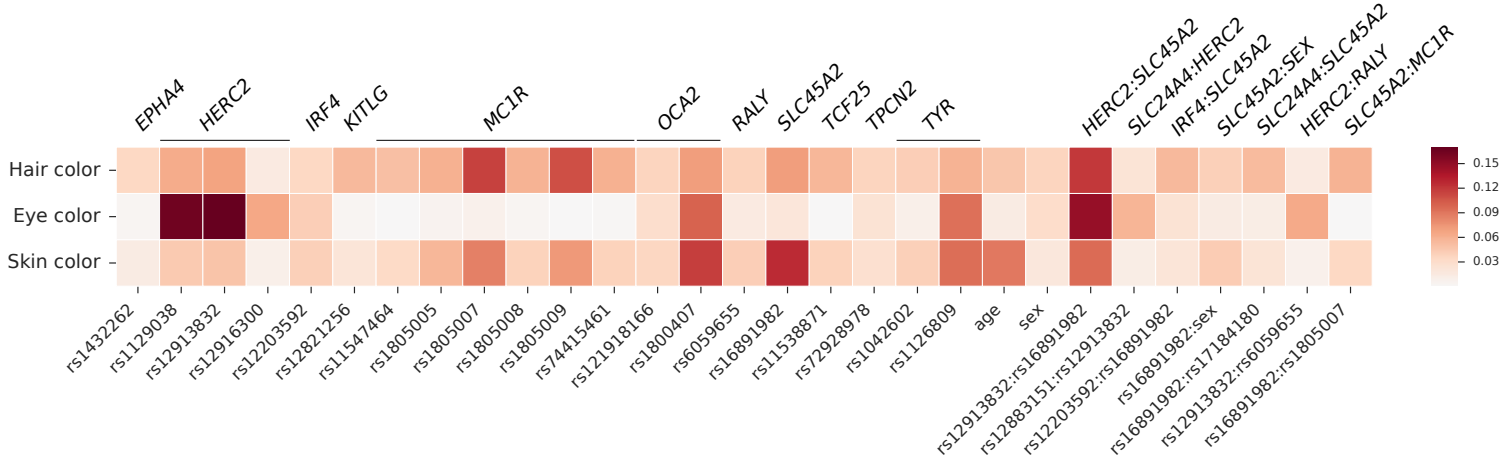


Figure 4: Estimated joint distribution of the predicted pigmentation phenotype pairs. The predicted phenotypes capture the correlation structure between phenotypes.

Gradients


EPHA4
HERC2
IRF4
KITLG
MC1R
OCA2
RALY
SLC45A2
TCF25
TPCN2
TYR
HERC2:SLC45A2
SLC24A4:HERC2
IRF4:SLC45A2
SLC45A2:SEX
SLC24A4:SLC45A2
HERC2:RALY
SLC45A2:MC1R
Hair color
Eye color
Skin color
rs1432262
rs1129038
rs12913832
rs12916300
rs12203592
rs12821256
rs11547464
rs1805005
rs1805007
rs1805008
rs1805009
rs74415461
rs121918166
rs1800407
rs6059655
rs16891982
rs11538871
rs72928978
rs1042602
rs1126809
age
sex
rs12913832:rs16891982
rs12883151:rs12913832
rs12203592:rs16891982
rs16891982:sex
rs16891982:rs17184180
rs12913832:rs6059655
rs16891982:rs1805007
0.15
0.12
0.09
0.06
0.03

Results

To assess the performance of Pigmentor, for each phenotype, we computed pairwise AUCs between all pairs of levels and then computed the aggregate AUC as follows:

$$\text{Aggregate AUC} = \frac{\sum_{i,j} \pi_i \pi_j \text{AUC}_{i,j}}{\sum_{i,j} \pi_i \pi_j}$$

Phenotype	Aggregate AUC
Eye color	85.42%
Hair color	80.77%
Skin color	76.12%

Genetics in Medicine

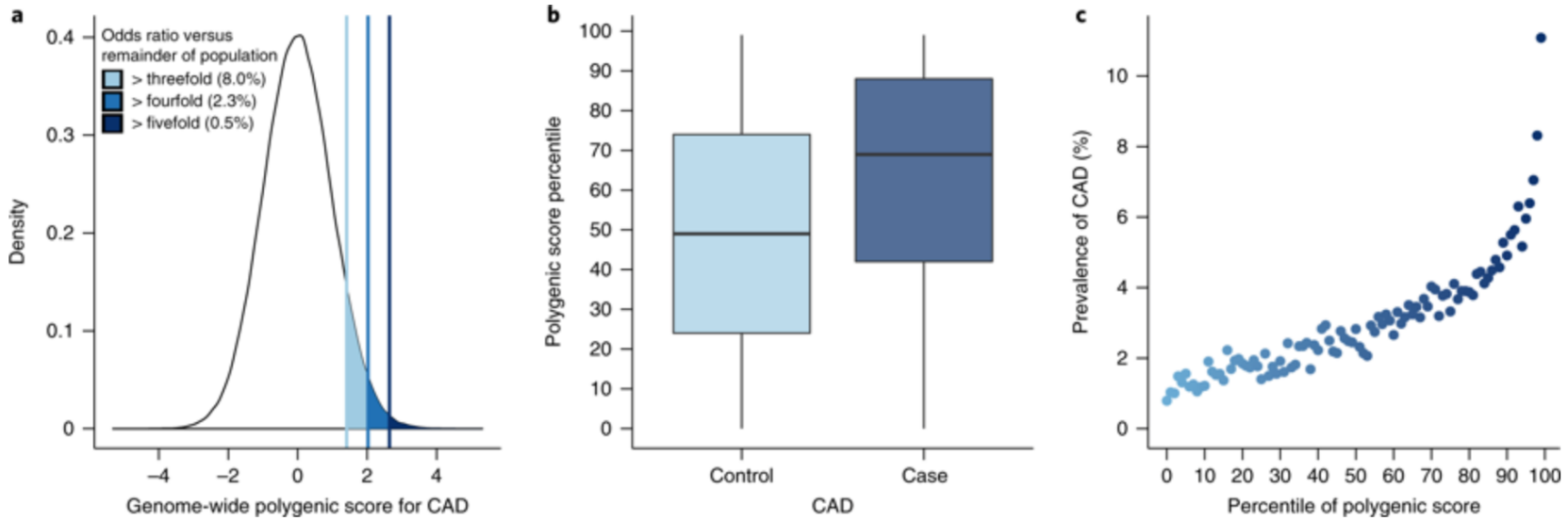
- high risk alleles for largely monogenic diseases
  - tend to be loss of function (LoF)
  - these tend to be quite rare
- examples
  - BRCA1 and BRCA2, LoF variants are well established risk factor for breast, ovarian and other cancers
  - NOD2 LoF variants are associated with increased risk of IBD
  - CFTR variants with risk for cystic fibrosis
- but most SNPs detected by GWAS are not LoF

What about polygenic diseases?

- many diseases and traits are polygenic with hundreds or thousands of variants
  - height, weight, type 2 diabetes, NASH/NAFLD
- for drug discovery we pick out some that are interesting and look like they might be druggable
  - to modulate disease progression you don't need to fix everything
- more recently use of highly polygenic risk scores are leading to interesting applications

Highly Polygenic Risk Scores

From: Genome-wide polygenic scores for common diseases identify individuals with risk equivalent to monogenic mutations



They report:

- CAD polygenic predictors derived from a GWAS involving 184,305 participants
- evaluated on UK Biobank CAD diagnosis
- AUCs ranging from 0.79–0.81 in the validation set
- best predictor (GPSCAD) used 6,630,150 variants