Feature Selection in Machine Learning for BioMedical Data

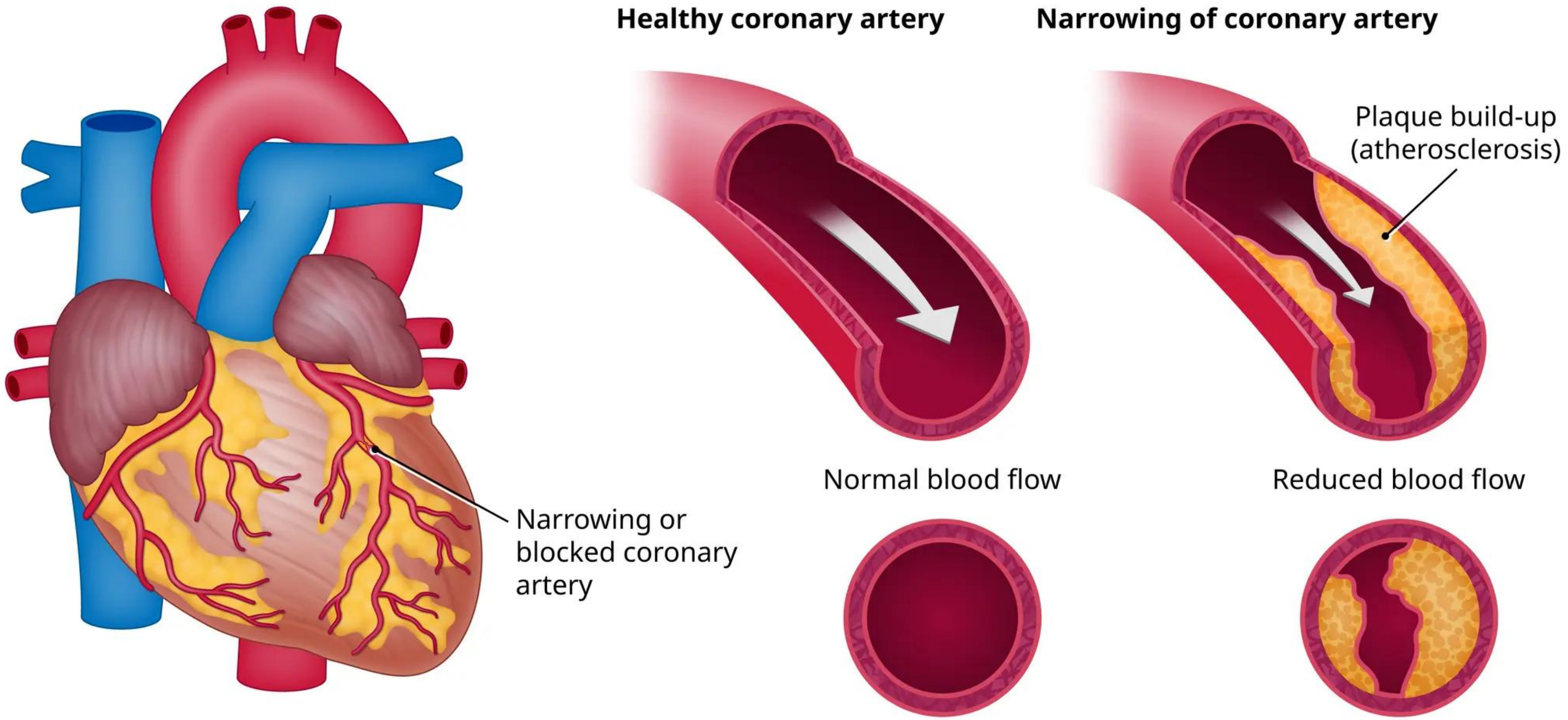
Nov 07 2025

Giảng viên: TS. Lưu Phúc Lợi
Email: Luu.p.loi@googlemail.com
Zalo: 0901802182

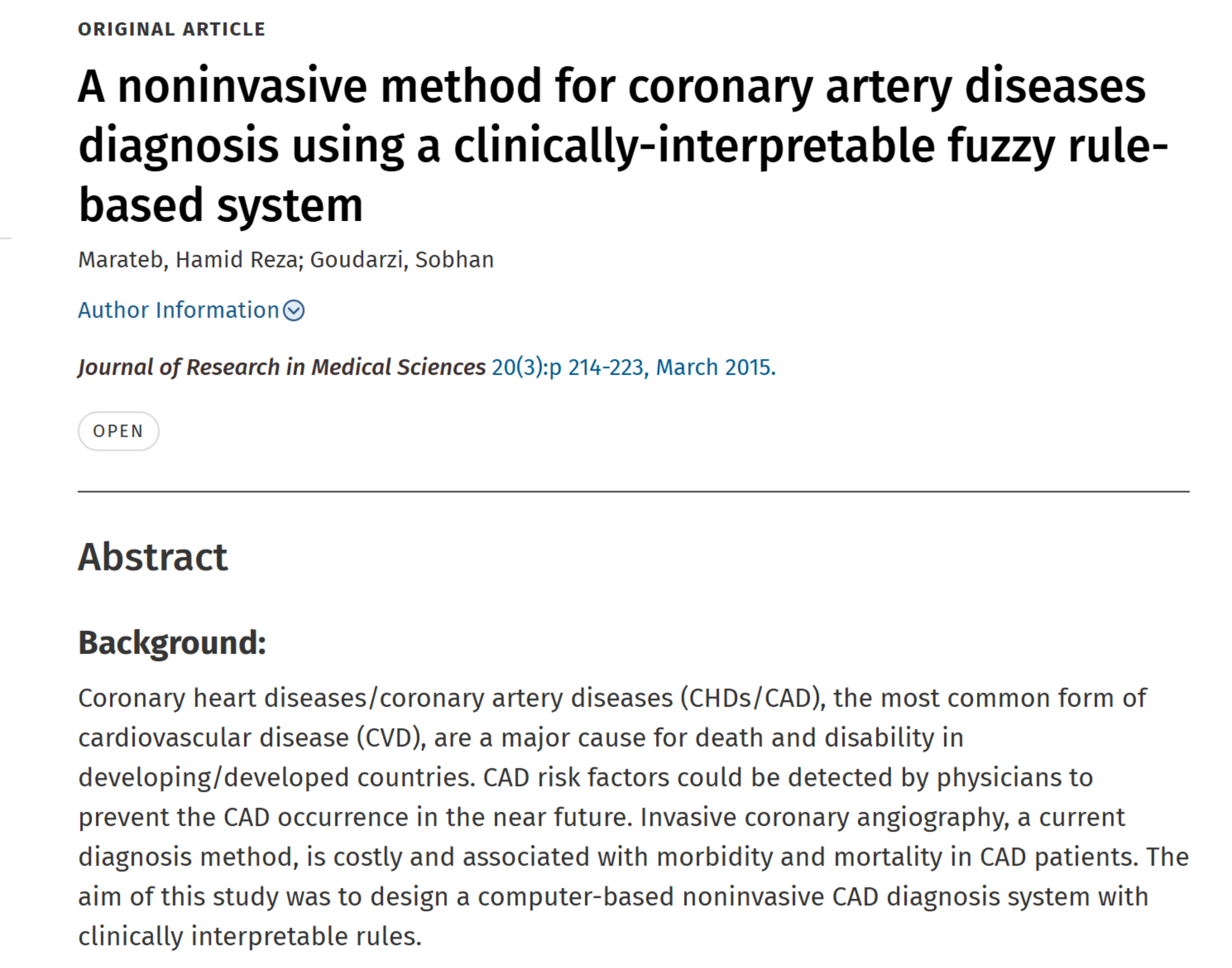
Content

1. Introduction to Feature Selection
2. Filter method
3. Wrapper methods
4. Embedded method
5. Hydrid method

Coronary Artery Disease


Healthy coronary artery
Narrowing of coronary artery
Plaque build-up (atherosclerosis)
Normal blood flow
Reduced blood flow
Narrowing or blocked coronary artery

© IHH Healthcare Singapore



ORIGINAL ARTICLE

A noninvasive method for coronary artery diseases diagnosis using a clinically-interpretable fuzzy rule-based system

Marateb, Hamid Reza; Goudarzi, Sobhan

Author Information

Journal of Research in Medical Sciences 20(3):p 214-223, March 2015.

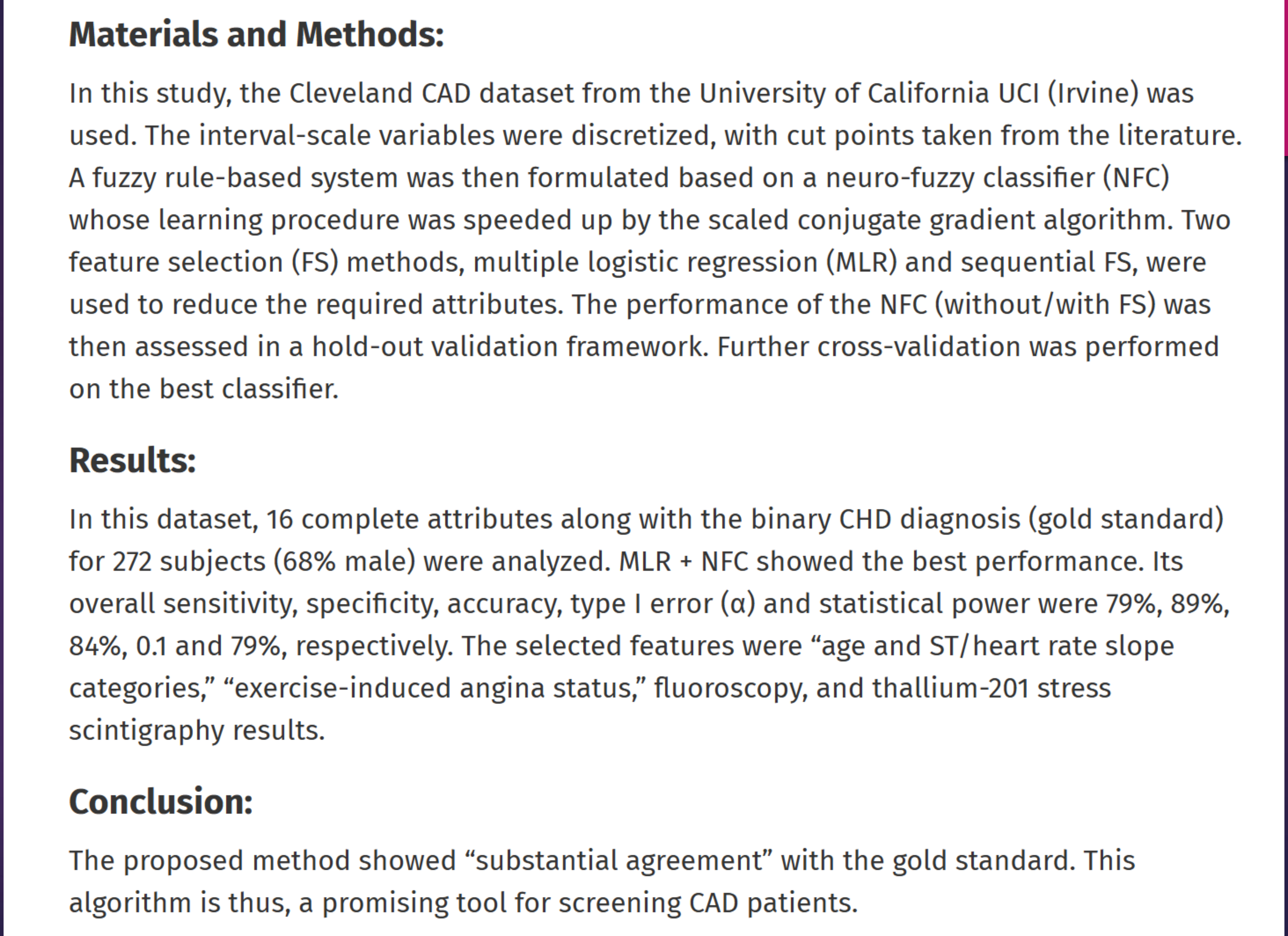
OPEN

Abstract

Background:

Coronary heart diseases/coronary artery diseases (CHDs/CAD), the most common form of cardiovascular disease (CVD), are a major cause for death and disability in developing/developed countries. CAD risk factors could be detected by physicians to prevent the CAD occurrence in the near future. Invasive coronary angiography, a current diagnosis method, is costly and associated with morbidity and mortality in CAD patients. The aim of this study was to design a computer-based noninvasive CAD diagnosis system with clinically interpretable rules.

https://journals.lww.com/jrms/fulltext/2015/20030/a_noninvasive_method_for_coronary_artery_diseases.2.aspx



Materials and Methods:

In this study, the Cleveland CAD dataset from the University of California UCI (Irvine) was used. The interval-scale variables were discretized, with cut points taken from the literature. A fuzzy rule-based system was then formulated based on a neuro-fuzzy classifier (NFC) whose learning procedure was speeded up by the scaled conjugate gradient algorithm. Two feature selection (FS) methods, multiple logistic regression (MLR) and sequential FS, were used to reduce the required attributes. The performance of the NFC (without/with FS) was then assessed in a hold-out validation framework. Further cross-validation was performed on the best classifier.

Results:

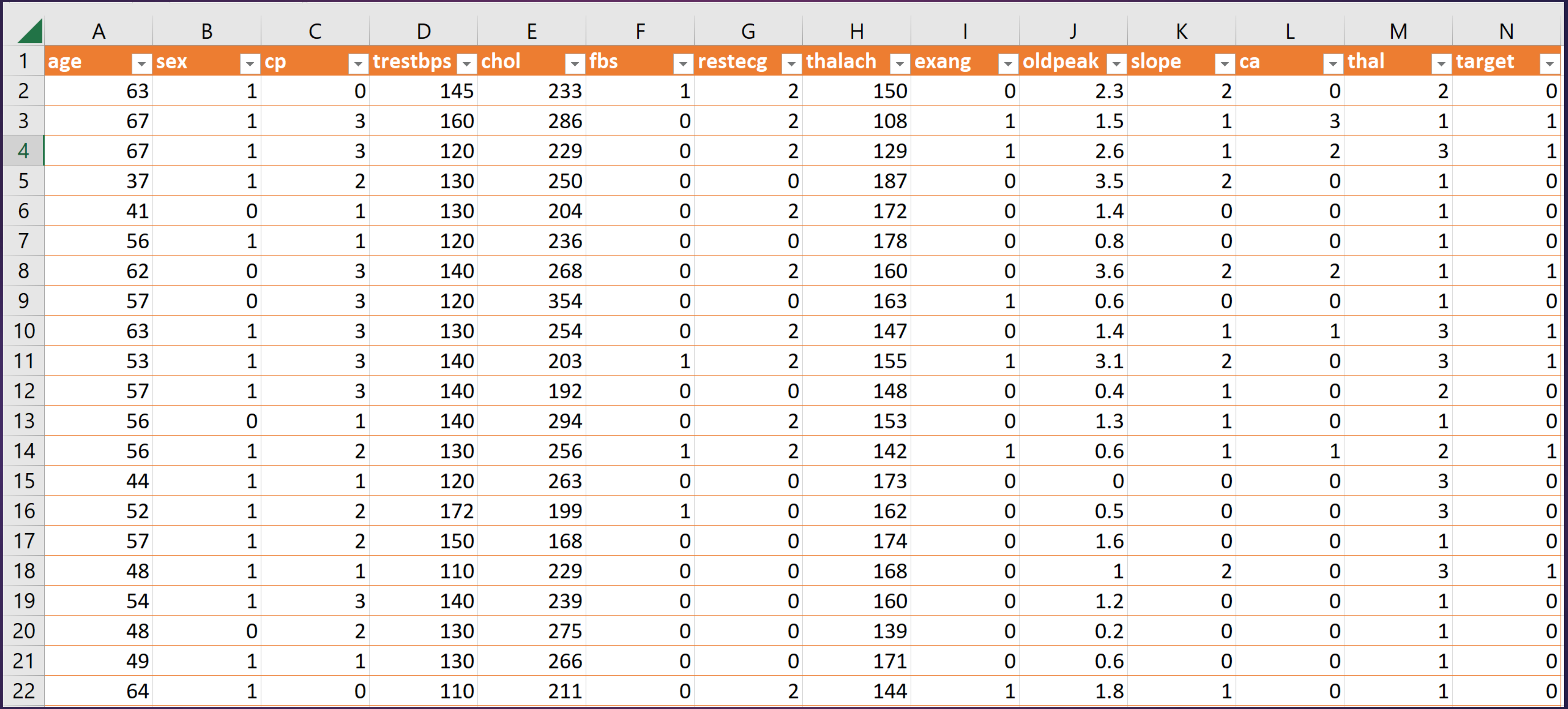
In this dataset, 16 complete attributes along with the binary CHD diagnosis (gold standard) for 272 subjects (68% male) were analyzed. MLR + NFC showed the best performance. Its overall sensitivity, specificity, accuracy, type I error (α) and statistical power were 79%, 89%, 84%, 0.1 and 79%, respectively. The selected features were “age and ST/heart rate slope categories,” “exercise-induced angina status,” fluoroscopy, and thallium-201 stress scintigraphy results.

Conclusion:

The proposed method showed “substantial agreement” with the gold standard. This algorithm is thus, a promising tool for screening CAD patients.

Problem statement

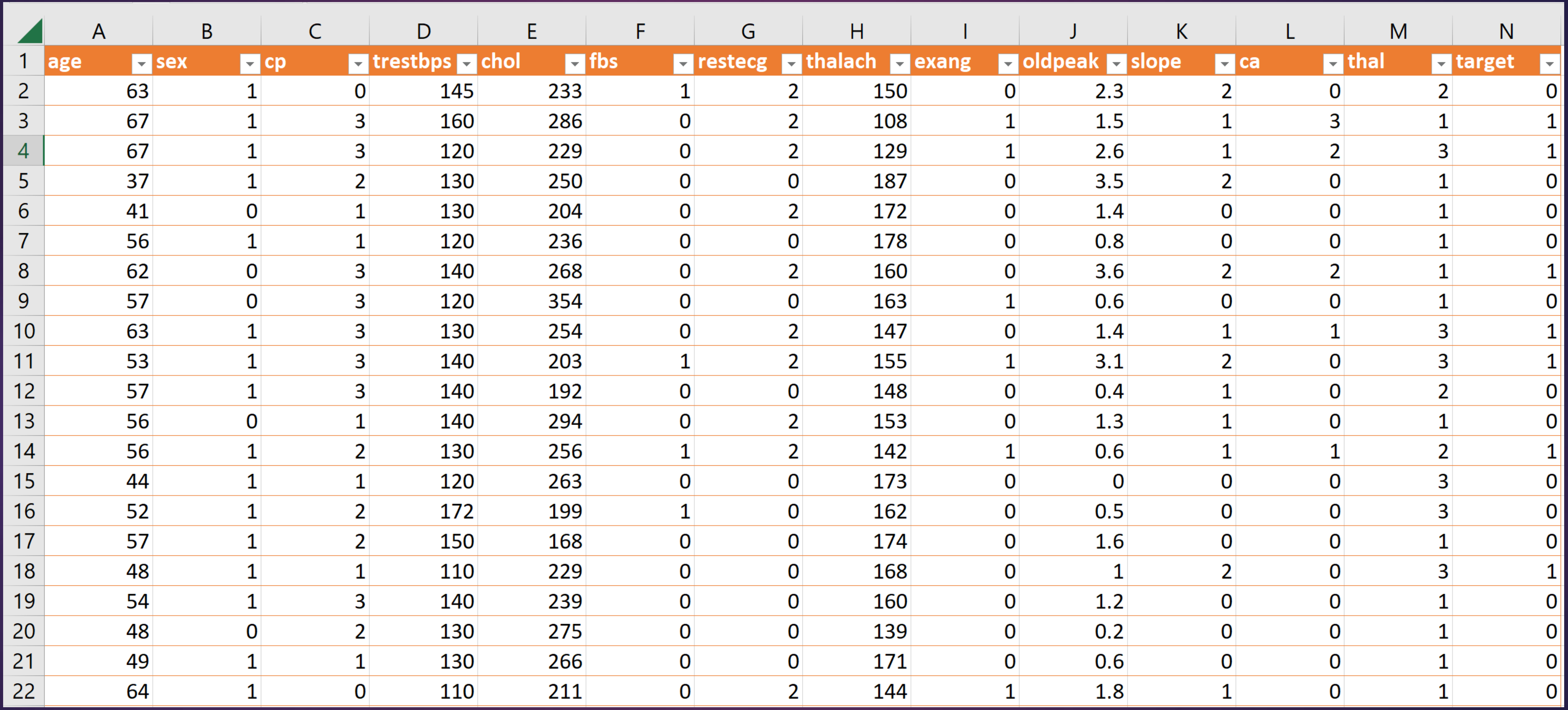
Predicting Heart Disease from the Cleveland Heart Disease Dataset (303 samples x 76 features)



	A	B	C	D	E	F	G	H	I	J	K	L	M	N
1	age	sex	cp	trestbps	chol	fbs	restecg	thalach	exang	oldpeak	slope	ca	thal	target
2	63	1	0	145	233	1	2	150	0	2.3	2	0	2	0
3	67	1	3	160	286	0	2	108	1	1.5	1	3	1	1
4	67	1	3	120	229	0	2	129	1	2.6	1	2	3	1
5	37	1	2	130	250	0	0	187	0	3.5	2	0	1	0
6	41	0	1	130	204	0	2	172	0	1.4	0	0	1	0
7	56	1	1	120	236	0	0	178	0	0.8	0	0	1	0
8	62	0	3	140	268	0	2	160	0	3.6	2	2	1	1
9	57	0	3	120	354	0	0	163	1	0.6	0	0	1	0
10	63	1	3	130	254	0	2	147	0	1.4	1	1	3	1
11	53	1	3	140	203	1	2	155	1	3.1	2	0	3	1
12	57	1	3	140	192	0	0	148	0	0.4	1	0	2	0
13	56	0	1	140	294	0	2	153	0	1.3	1	0	1	0
14	56	1	2	130	256	1	2	142	1	0.6	1	1	2	1
15	44	1	1	120	263	0	0	173	0	0	0	0	3	0
16	52	1	2	172	199	1	0	162	0	0.5	0	0	3	0
17	57	1	2	150	168	0	0	174	0	1.6	0	0	1	0
18	48	1	1	110	229	0	0	168	0	1	2	0	3	1
19	54	1	3	140	239	0	0	160	0	1.2	0	0	1	0
20	48	0	2	130	275	0	0	139	0	0.2	0	0	1	0
21	49	1	1	130	266	0	0	171	0	0.6	0	0	1	0
22	64	1	0	110	211	0	2	144	1	1.8	1	0	1	0

https://www.kaggle.com/datasets/ritwikb3/heart-disease-cleveland

Predicting Heart Disease from the Cleveland Heart Disease Dataset (303 samples x 13 features)



	A	B	C	D	E	F	G	H	I	J	K	L	M	N
1	age	sex	cp	trestbps	chol	fbs	restecg	thalach	exang	oldpeak	slope	ca	thal	target
2	63	1	0	145	233	1	2	150	0	2.3	2	0	2	0
3	67	1	3	160	286	0	2	108	1	1.5	1	3	1	1
4	67	1	3	120	229	0	2	129	1	2.6	1	2	3	1
5	37	1	2	130	250	0	0	187	0	3.5	2	0	1	0
6	41	0	1	130	204	0	2	172	0	1.4	0	0	1	0
7	56	1	1	120	236	0	0	178	0	0.8	0	0	1	0
8	62	0	3	140	268	0	2	160	0	3.6	2	2	1	1
9	57	0	3	120	354	0	0	163	1	0.6	0	0	1	0
10	63	1	3	130	254	0	2	147	0	1.4	1	1	3	1
11	53	1	3	140	203	1	2	155	1	3.1	2	0	3	1
12	57	1	3	140	192	0	0	148	0	0.4	1	0	2	0
13	56	0	1	140	294	0	2	153	0	1.3	1	0	1	0
14	56	1	2	130	256	1	2	142	1	0.6	1	1	2	1
15	44	1	1	120	263	0	0	173	0	0	0	0	3	0
16	52	1	2	172	199	1	0	162	0	0.5	0	0	3	0
17	57	1	2	150	168	0	0	174	0	1.6	0	0	1	0
18	48	1	1	110	229	0	0	168	0	1	2	0	3	1
19	54	1	3	140	239	0	0	160	0	1.2	0	0	1	0
20	48	0	2	130	275	0	0	139	0	0.2	0	0	1	0
21	49	1	1	130	266	0	0	171	0	0.6	0	0	1	0
22	64	1	0	110	211	0	2	144	1	1.8	1	0	1	0

https://www.kaggle.com/datasets/ritwikb3/heart-disease-cleveland

Cleveland Heart Disease Dataset

This database contains 13 attributes and a target variable. It has 8 nominal values and 5 numeric values. The detailed description of all these features are as follows:

1. Age: Patients Age in years (Numeric)

2. Sex: Gender (Male : 1; Female : 0) (Nominal)

3. cp: Type of chest pain experienced by patient. This term categorized into 4 category. 0 typical angina, 1 atypical angina, 2 non- anginal pain, 3 asymptomatic (Nominal)

4. trestbps: patient's level of blood pressure at resting mode in mm/HG (Numerical)

5. chol: Serum cholesterol in mg/dl (Numeric)

6. fbs: Blood sugar levels on fasting > 120 mg/dl represents as 1 in case of true and 0 as false (Nominal)

7. restecg: Result of electrocardiogram while at rest are represented in 3 distinct values 0 : Normal 1: having ST-T wave abnormality (T wave inversions and/or ST elevation or depression of > 0.05 mV) 2: showing probable or definite left ventricular hypertrophyby Estes' criteria (Nominal)

8. thalach: Maximum heart rate achieved (Numeric)

Cleveland Heart Disease Dataset

This database contains 13 attributes and a target variable. It has 8 nominal values and 5 numeric values. The detailed description of all these features are as follows:

9. exang: Angina induced by exercise 0 depicting NO 1 depicting Yes (Nominal)

10. oldpeak: Exercise induced ST-depression in relative with the state of rest (Numeric)

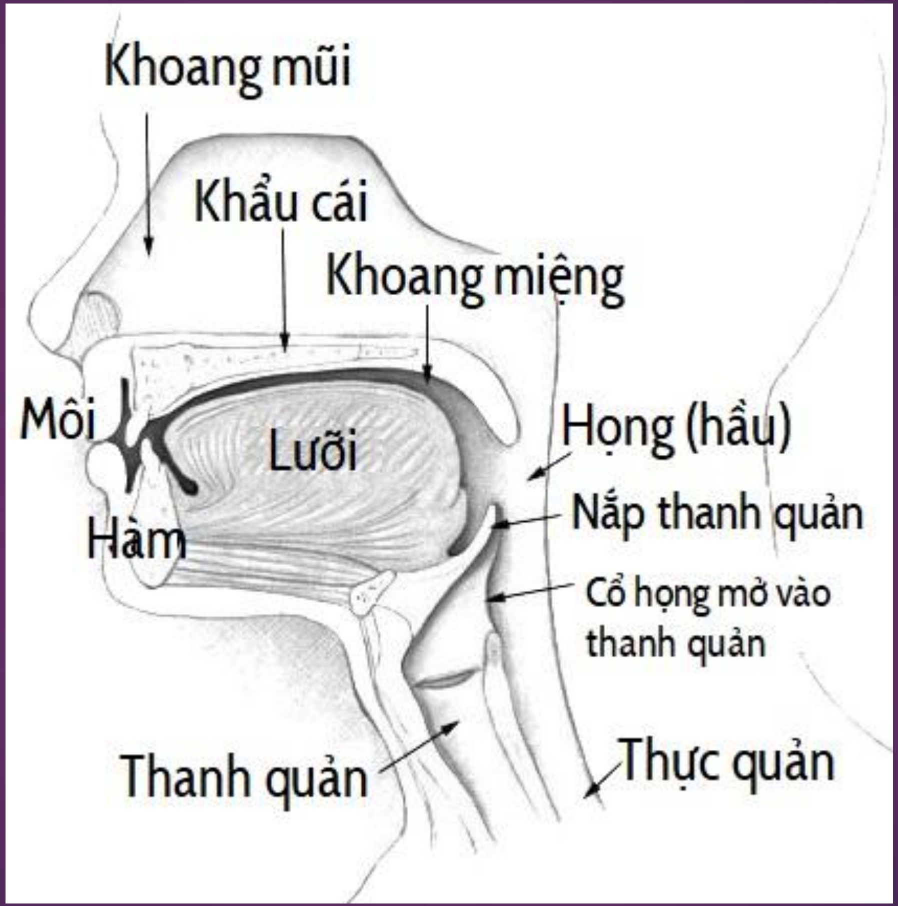
11. slope: ST segment measured in terms of slope during peak exercise
0: up sloping; 1: flat; 2: down sloping(Nominal)

12. ca: The number of major vessels (0–3)(nominal)

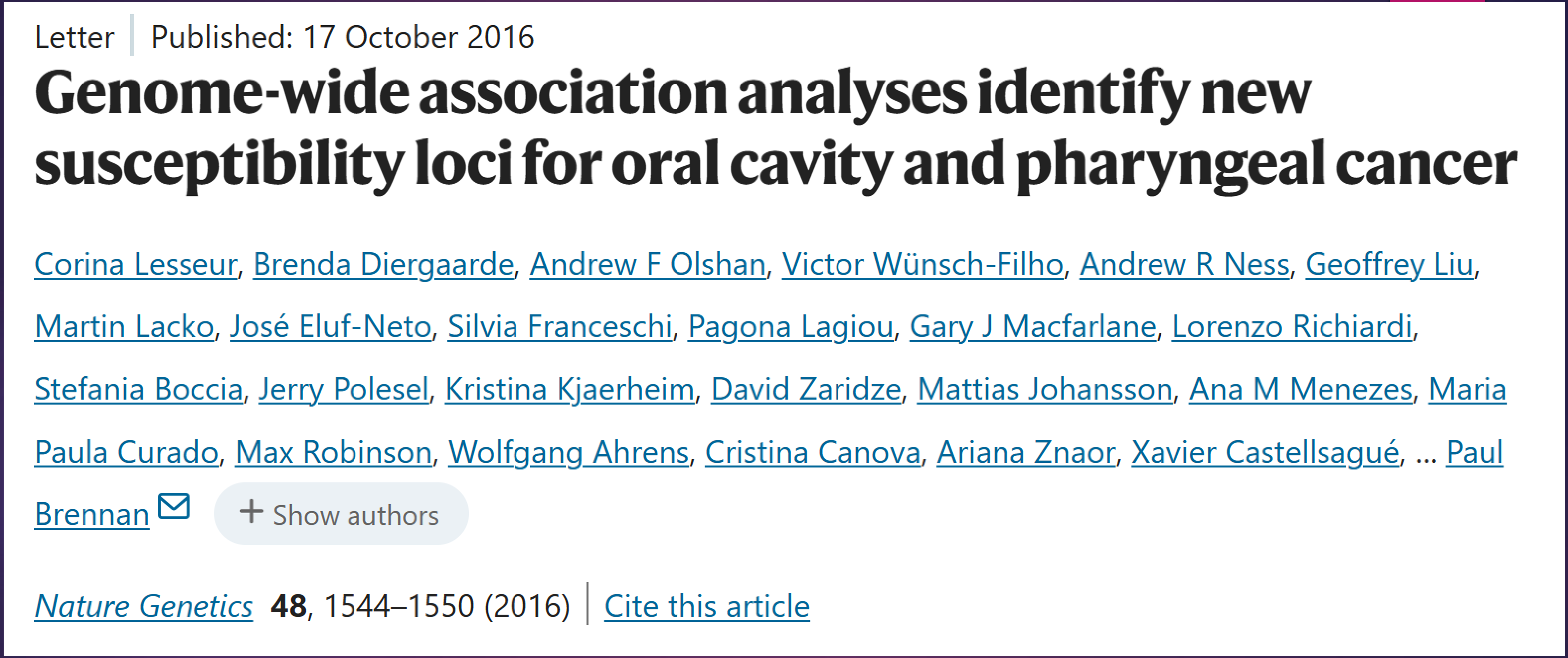
13. thal: A blood disorder called thalassemia
0: NULL 1: normal blood flow 2: fixed defect (no blood flow in some part of the heart) 3: reversible defect (a blood flow is observed but it is not normal(nominal)

14. target: It is the target variable which we have to predict 1 means patient is suffering from heart disease and 0 means patient is normal.
Variable to be predicted
Absence (1) or presence (2) of heart disease

Oral Cavity and Pharyngeal Cancer



https://vi.wikipedia.org/wiki/C%E1%BB%95_h%E1%BB%8Dng#/media/T%E1%BA%ADp_tin:Illu01_head_neck_vi.jpg



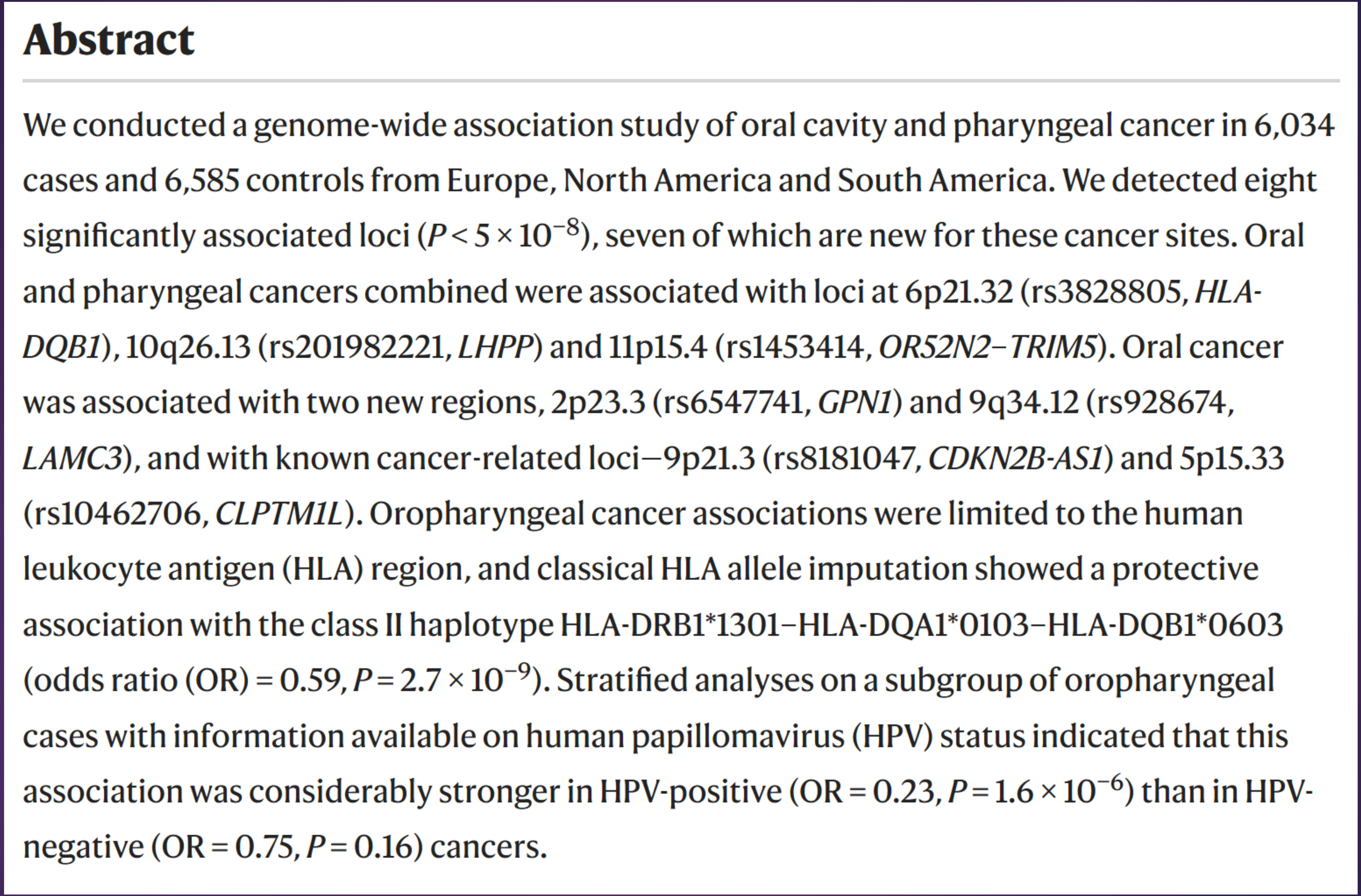
Letter | Published: 17 October 2016

Genome-wide association analyses identify new susceptibility loci for oral cavity and pharyngeal cancer

Corina Lesseur, Brenda Diergaarde, Andrew F Olshan, Victor Wünsch-Filho, Andrew R Ness, Geoffrey Liu, Martin Lacko, José Eluf-Neto, Silvia Franceschi, Pagona Lagiou, Gary J Macfarlane, Lorenzo Richiardi, Stefania Boccia, Jerry Polesel, Kristina Kjaerheim, David Zaridze, Mattias Johansson, Ana M Menezes, Maria Paula Curado, Max Robinson, Wolfgang Ahrens, Cristina Canova, Ariana Znaor, Xavier Castellsagué, … Paul Brennan ✉

Nature Genetics **48**, 1544–1550 (2016) | Cite this article

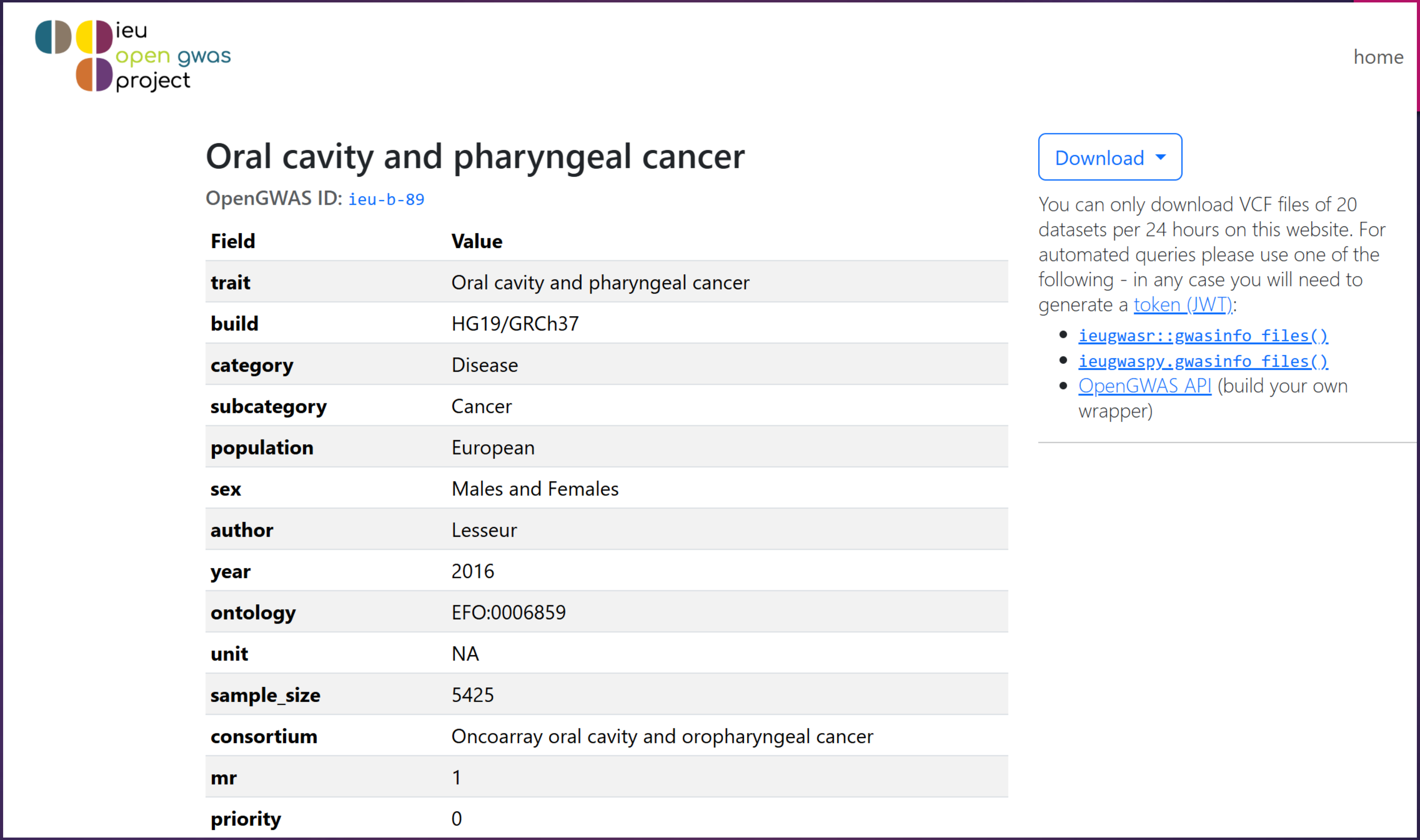
https://www.nature.com/articles/ng.3685



Abstract

We conducted a genome-wide association study of oral cavity and pharyngeal cancer in 6,034 cases and 6,585 controls from Europe, North America and South America. We detected eight significantly associated loci ($P < 5 \times 10^{-8}$), seven of which are new for these cancer sites. Oral and pharyngeal cancers combined were associated with loci at 6p21.32 (rs3828805, *HLA-DQB1*), 10q26.13 (rs201982221, *LHPP*) and 11p15.4 (rs1453414, *OR52N2–TRIM5*). Oral cancer was associated with two new regions, 2p23.3 (rs6547741, *GPN1*) and 9q34.12 (rs928674, *LAMC3*), and with known cancer-related loci—9p21.3 (rs8181047, *CDKN2B-AS1*) and 5p15.33 (rs10462706, *CLPTM1L*). Oropharyngeal cancer associations were limited to the human leukocyte antigen (HLA) region, and classical HLA allele imputation showed a protective association with the class II haplotype HLA-DRB1*1301–HLA-DQA1*0103–HLA-DQB1*0603 (odds ratio (OR) = 0.59, $P = 2.7 \times 10^{-9}$). Stratified analyses on a subgroup of oropharyngeal cases with information available on human papillomavirus (HPV) status indicated that this association was considerably stronger in HPV-positive (OR = 0.23, $P = 1.6 \times 10^{-6}$) than in HPV-negative (OR = 0.75, $P = 0.16$) cancers.

https://www.nature.com/articles/ng.3685



ieu open gwas project

home

Oral cavity and pharyngeal cancer

OpenGWAS ID: `ieu-b-89`

Field	Value
trait	Oral cavity and pharyngeal cancer
build	HG19/GRCh37
category	Disease
subcategory	Cancer
population	European
sex	Males and Females
author	Lesseur
year	2016
ontology	EFO:0006859
unit	NA
sample_size	5425
consortium	Oncoarray oral cavity and oropharyngeal cancer
mr	1
priority	0

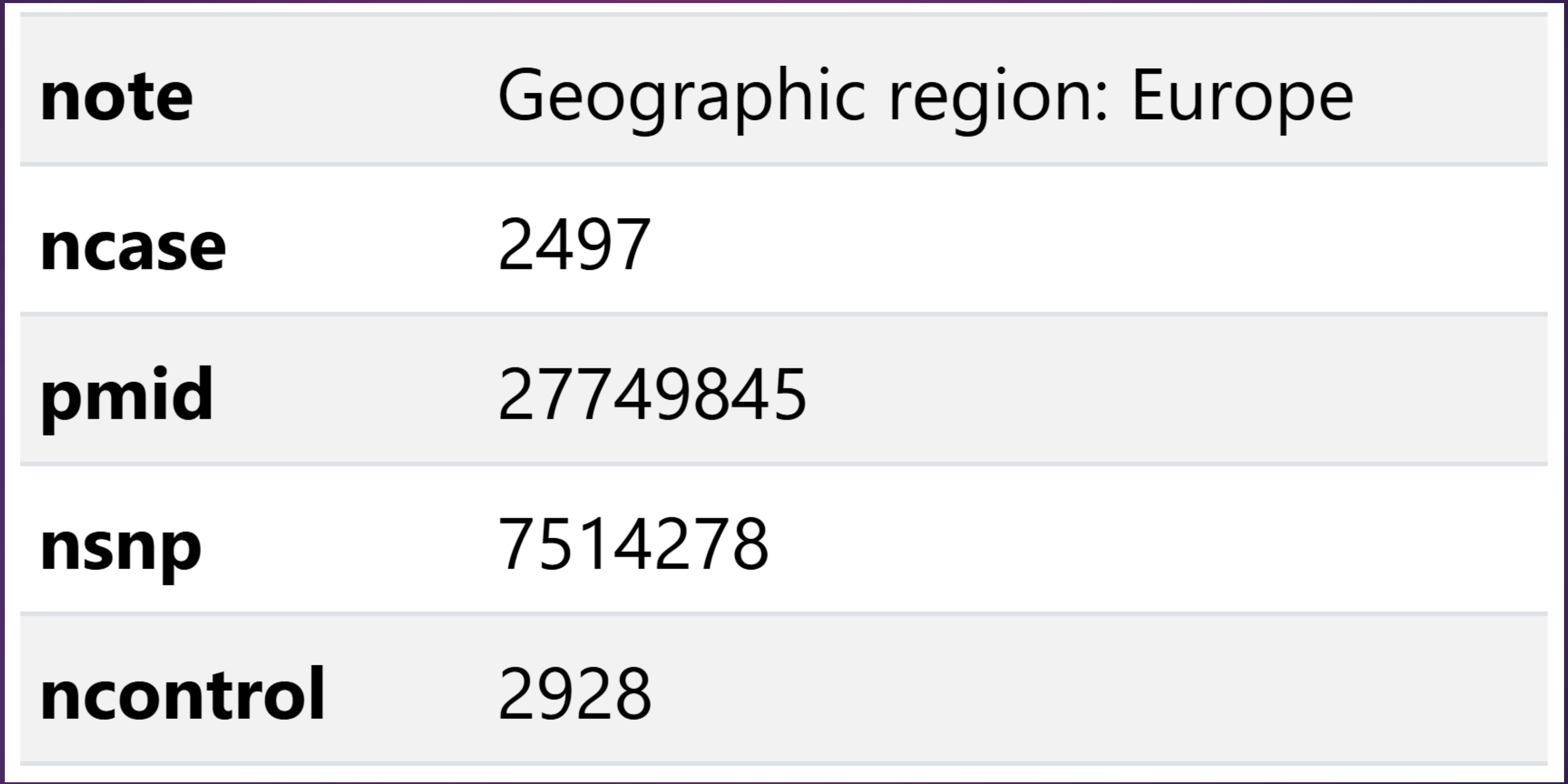
Download ▾

You can only download VCF files of 20 datasets per 24 hours on this website. For automated queries please use one of the following - in any case you will need to generate a token (JWT):

- `ieugwasr::gwasinfo_files()`
- `ieugwaspy.gwasinfo_files()`
- OpenGWAS API (build your own wrapper)

https://opengwas.io/datasets/ieu-b-89

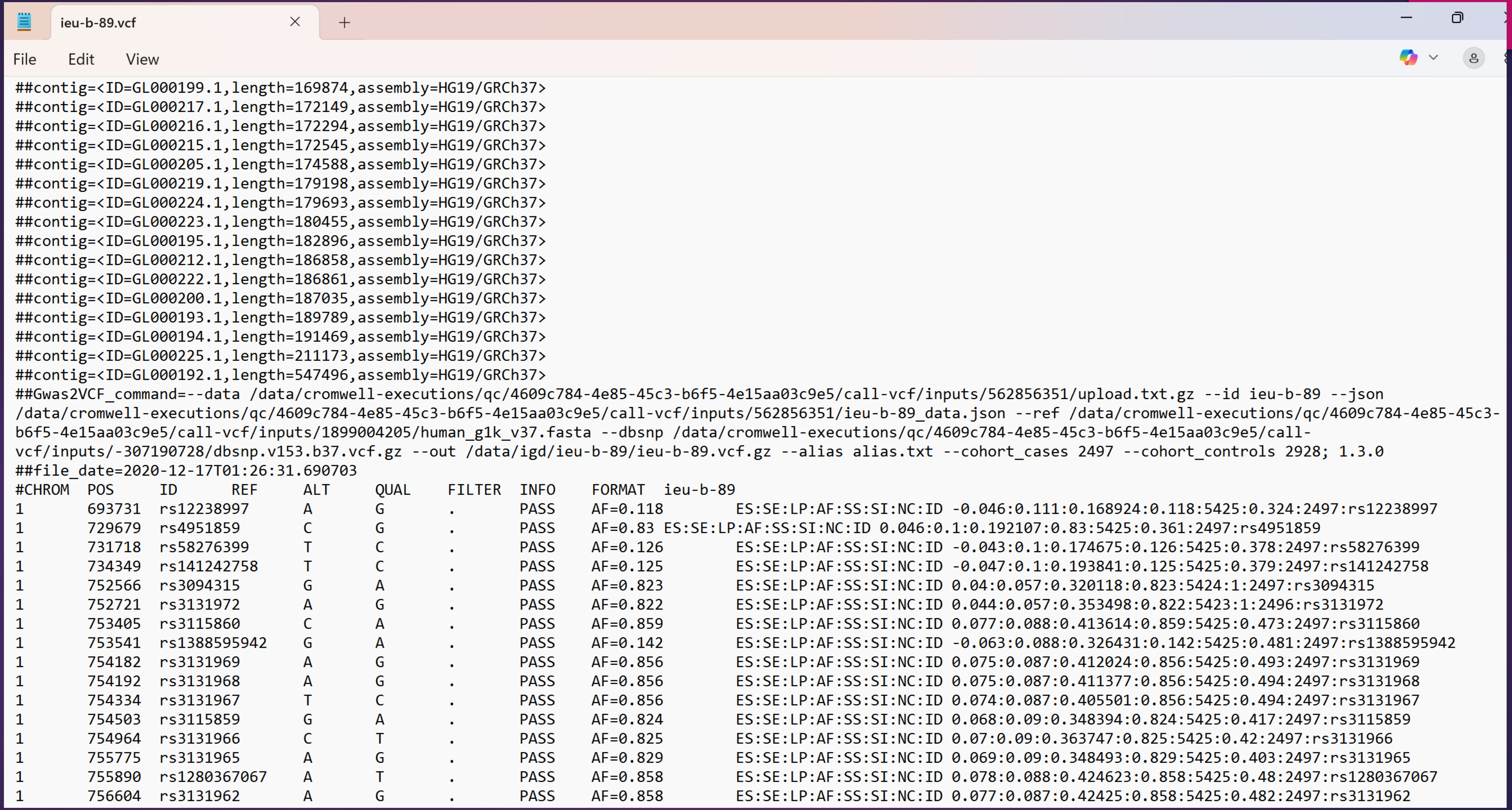
SNP - Oral Cavity and Pharyngeal Cancer Dataset



note	Geographic region: Europe
ncase	2497
pmid	27749845
nsnp	7514278
ncontrol	2928

https://opengwas.io/datasets/ieu-b-89

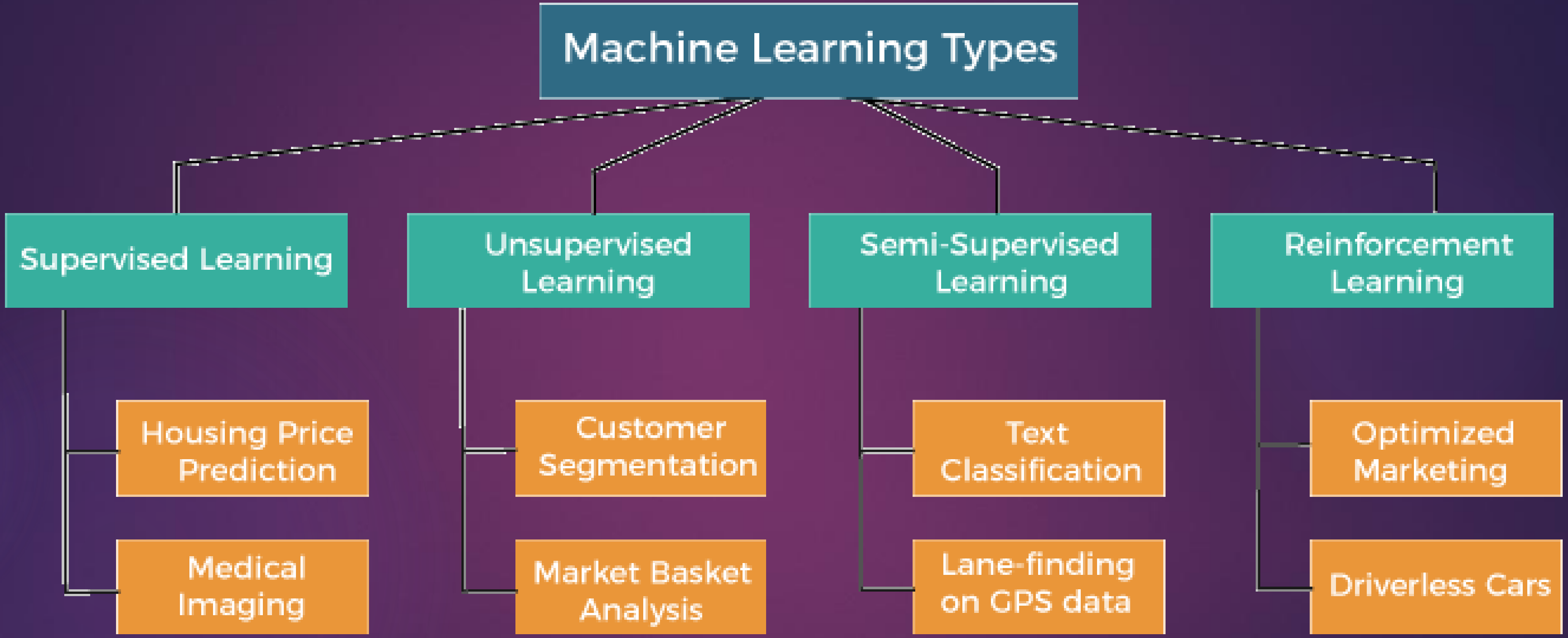
SNP - Oral Cavity and Pharyngeal Cancer Dataset



```
##contig=<ID=GL000199.1,length=169874,assembly=HG19/GRCh37>
##contig=<ID=GL000217.1,length=172149,assembly=HG19/GRCh37>
##contig=<ID=GL000216.1,length=172294,assembly=HG19/GRCh37>
##contig=<ID=GL000215.1,length=172545,assembly=HG19/GRCh37>
##contig=<ID=GL000205.1,length=174588,assembly=HG19/GRCh37>
##contig=<ID=GL000219.1,length=179198,assembly=HG19/GRCh37>
##contig=<ID=GL000224.1,length=179693,assembly=HG19/GRCh37>
##contig=<ID=GL000223.1,length=180455,assembly=HG19/GRCh37>
##contig=<ID=GL000195.1,length=182896,assembly=HG19/GRCh37>
##contig=<ID=GL000212.1,length=186858,assembly=HG19/GRCh37>
##contig=<ID=GL000222.1,length=186861,assembly=HG19/GRCh37>
##contig=<ID=GL000200.1,length=187035,assembly=HG19/GRCh37>
##contig=<ID=GL000193.1,length=189789,assembly=HG19/GRCh37>
##contig=<ID=GL000194.1,length=191469,assembly=HG19/GRCh37>
##contig=<ID=GL000225.1,length=211173,assembly=HG19/GRCh37>
##contig=<ID=GL000192.1,length=547496,assembly=HG19/GRCh37>
##Gwas2VCF_command=--data /data/cromwell-executions/qc/4609c784-4e85-45c3-b6f5-4e15aa03c9e5/call-vcf/inputs/562856351/upload.txt.gz --id ieu-b-89 --json
/data/cromwell-executions/qc/4609c784-4e85-45c3-b6f5-4e15aa03c9e5/call-vcf/inputs/562856351/ieu-b-89_data.json --ref /data/cromwell-executions/qc/4609c784-4e85-45c3-
b6f5-4e15aa03c9e5/call-vcf/inputs/1899004205/human_g1k_v37.fasta --dbsnp /data/cromwell-executions/qc/4609c784-4e85-45c3-b6f5-4e15aa03c9e5/call-
vcf/inputs/-307190728/dbsnp.v153.b37.vcf.gz --out /data/igd/ieu-b-89/ieu-b-89.vcf.gz --alias alias.txt --cohort_cases 2497 --cohort_controls 2928; 1.3.0
##file_date=2020-12-17T01:26:31.690703
#CHROM  POS     ID      REF     ALT     QUAL    FILTER  INFO    FORMAT  ieu-b-89
1       693731  rs12238997      A       G       .       PASS    AF=0.118        ES:SE:LP:AF:SS:SI:NC:ID -0.046:0.111:0.168924:0.118:5425:0.324:2497:rs12238997
1       729679  rs4951859       C       G       .       PASS    AF=0.83 ES:SE:LP:AF:SS:SI:NC:ID 0.046:0.1:0.192107:0.83:5425:0.361:2497:rs4951859
1       731718  rs58276399      T       C       .       PASS    AF=0.126        ES:SE:LP:AF:SS:SI:NC:ID -0.043:0.1:0.174675:0.126:5425:0.378:2497:rs58276399
1       734349  rs141242758     T       C       .       PASS    AF=0.125        ES:SE:LP:AF:SS:SI:NC:ID -0.047:0.1:0.193841:0.125:5425:0.379:2497:rs141242758
1       752566  rs3094315       G       A       .       PASS    AF=0.823        ES:SE:LP:AF:SS:SI:NC:ID 0.04:0.057:0.320118:0.823:5424:1:2497:rs3094315
1       752721  rs3131972       A       G       .       PASS    AF=0.822        ES:SE:LP:AF:SS:SI:NC:ID 0.044:0.057:0.353498:0.822:5423:1:2496:rs3131972
1       753405  rs3115860       C       A       .       PASS    AF=0.859        ES:SE:LP:AF:SS:SI:NC:ID 0.077:0.088:0.413614:0.859:5425:0.473:2497:rs3115860
1       753541  rs1388595942    G       A       .       PASS    AF=0.142        ES:SE:LP:AF:SS:SI:NC:ID -0.063:0.088:0.326431:0.142:5425:0.481:2497:rs1388595942
1       754182  rs3131969       A       G       .       PASS    AF=0.856        ES:SE:LP:AF:SS:SI:NC:ID 0.075:0.087:0.412024:0.856:5425:0.493:2497:rs3131969
1       754192  rs3131968       A       G       .       PASS    AF=0.856        ES:SE:LP:AF:SS:SI:NC:ID 0.075:0.087:0.411377:0.856:5425:0.494:2497:rs3131968
1       754334  rs3131967       T       C       .       PASS    AF=0.856        ES:SE:LP:AF:SS:SI:NC:ID 0.074:0.087:0.405501:0.856:5425:0.494:2497:rs3131967
1       754503  rs3115859       G       A       .       PASS    AF=0.824        ES:SE:LP:AF:SS:SI:NC:ID 0.068:0.09:0.348394:0.824:5425:0.417:2497:rs3115859
1       754964  rs3131966       C       T       .       PASS    AF=0.825        ES:SE:LP:AF:SS:SI:NC:ID 0.07:0.09:0.363747:0.825:5425:0.42:2497:rs3131966
1       755775  rs3131965       A       G       .       PASS    AF=0.829        ES:SE:LP:AF:SS:SI:NC:ID 0.069:0.09:0.348493:0.829:5425:0.403:2497:rs3131965
1       755890  rs1280367067    A       T       .       PASS    AF=0.858        ES:SE:LP:AF:SS:SI:NC:ID 0.078:0.088:0.424623:0.858:5425:0.48:2497:rs1280367067
1       756604  rs3131962       A       G       .       PASS    AF=0.858        ES:SE:LP:AF:SS:SI:NC:ID 0.077:0.087:0.42425:0.858:5425:0.482:2497:rs3131962
```

https://opengwas.io/datasets/ieu-b-89#

Introduction to Feature Selection



Machine Learning Types

Supervised Learning

- Housing Price Prediction
- Medical Imaging

Unsupervised Learning

- Customer Segmentation
- Market Basket Analysis

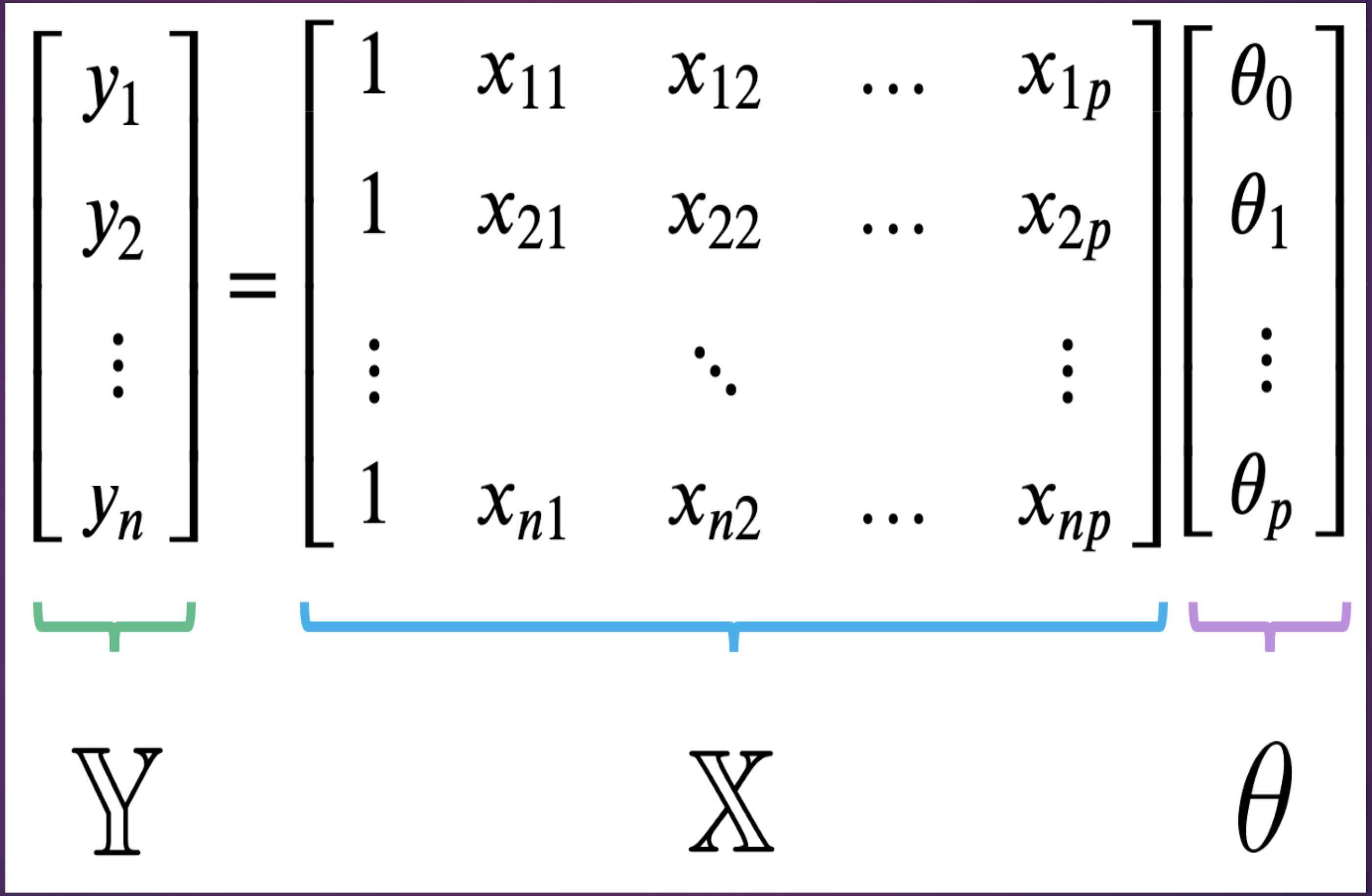
Semi-Supervised Learning

- Text Classification
- Lane-finding on GPS data

Reinforcement Learning

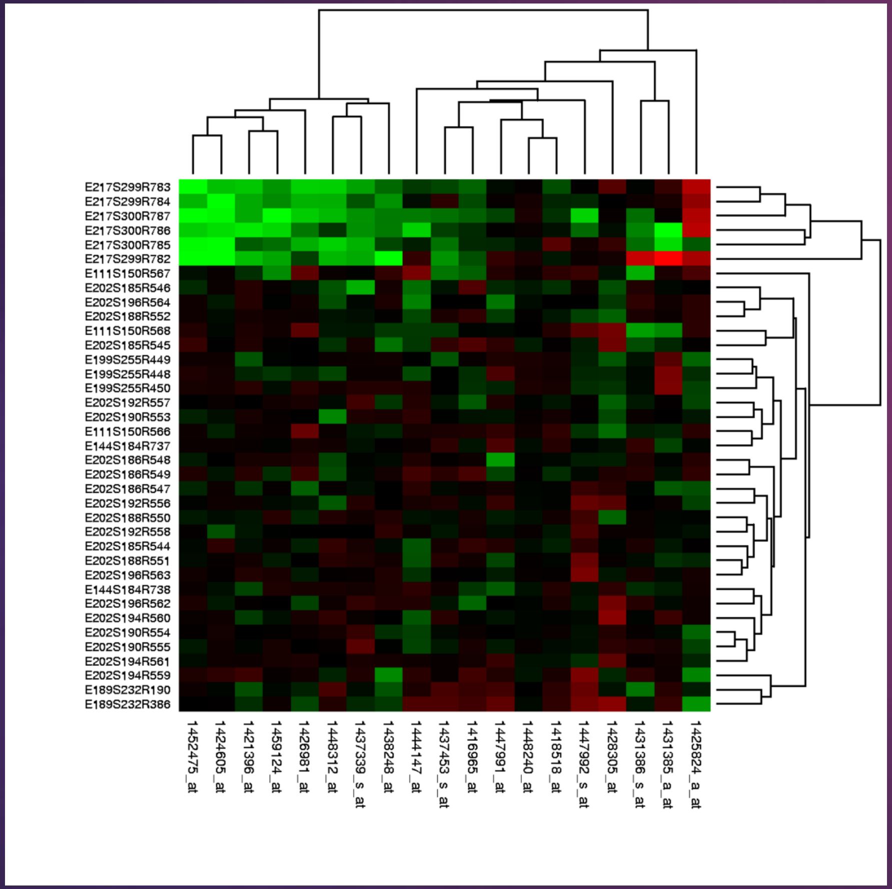
- Optimized Marketing
- Driverless Cars

Supervise Learning: regression or classification



$$\underbrace{\begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix}}_{\mathbb{Y}} = \underbrace{\begin{bmatrix} 1 & x_{11} & x_{12} & \dots & x_{1p} \\ 1 & x_{21} & x_{22} & \dots & x_{2p} \\ \vdots & & \ddots & & \vdots \\ 1 & x_{n1} & x_{n2} & \dots & x_{np} \end{bmatrix}}_{\mathbb{X}} \underbrace{\begin{bmatrix} \theta_0 \\ \theta_1 \\ \vdots \\ \theta_p \end{bmatrix}}_{\theta}$$

BioMedical data: gene expression with p >> n

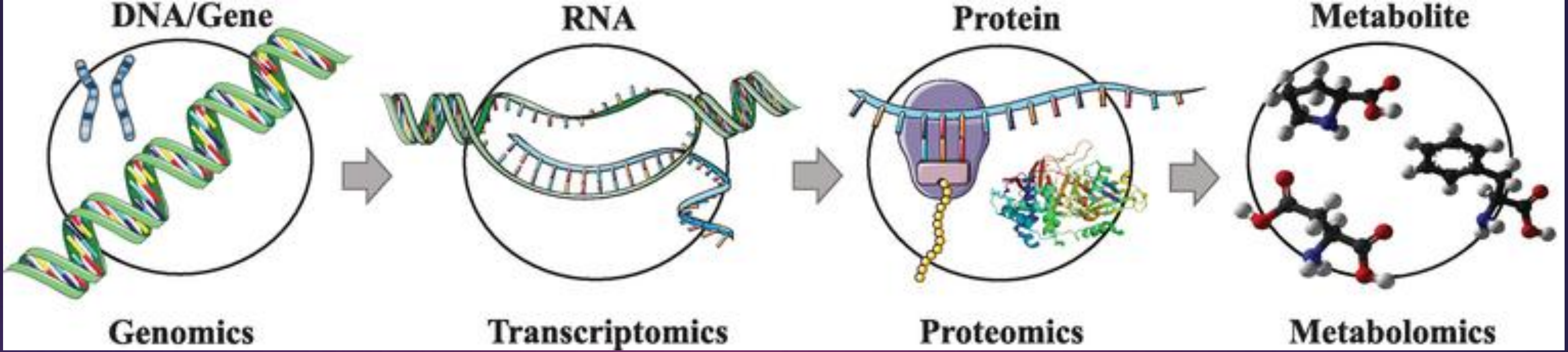


n = number of samples 6, 10, 100, 1k

P = number of genes 20k

https://en.wikipedia.org/wiki/Gene_expression_profiling

Mối liên kết: Biến thể gen và bệnh di truyền



PAH gene

Ref ...ATCGAT...

P1 ...AACGAT...

NM_000277.3(PAH):c.971T>A

PAH mRNA

Ref ...AUCGAU...

P1 ...AACGAU...

NM_000277.3(PAH):c.971T>A

PAH protein

Ref ...Ile-Asp...

P1 ...Asn-Asp...

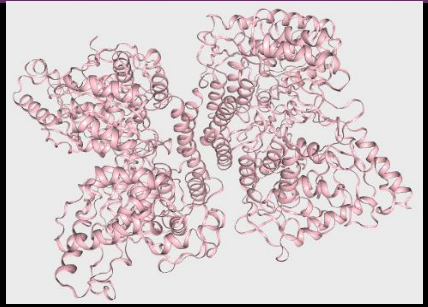
NM_000277.3(PAH):p.Ile324Asn

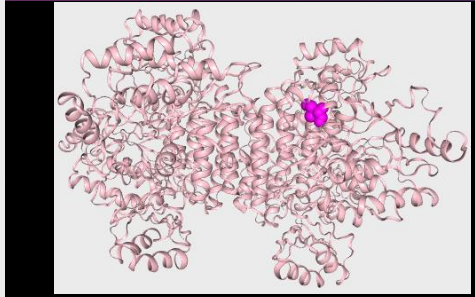
PAH

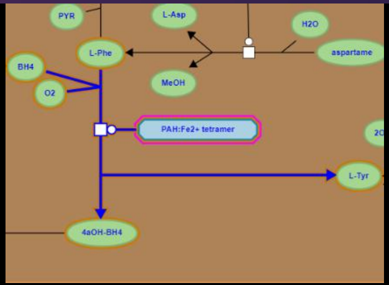
Ref Phe → Tyr

PAH

P1 Phe $\cancel{\rightarrow}$ Tyr

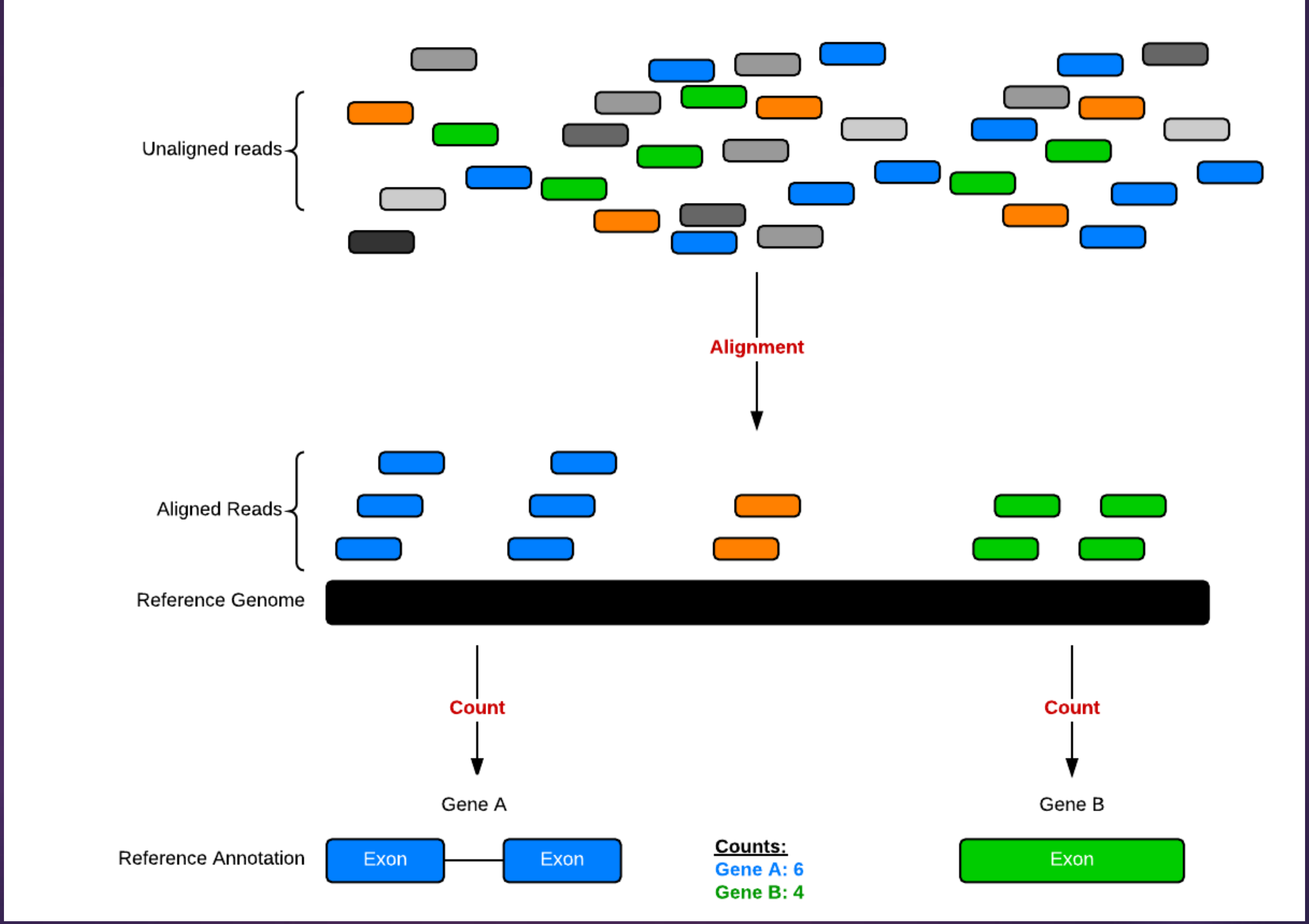




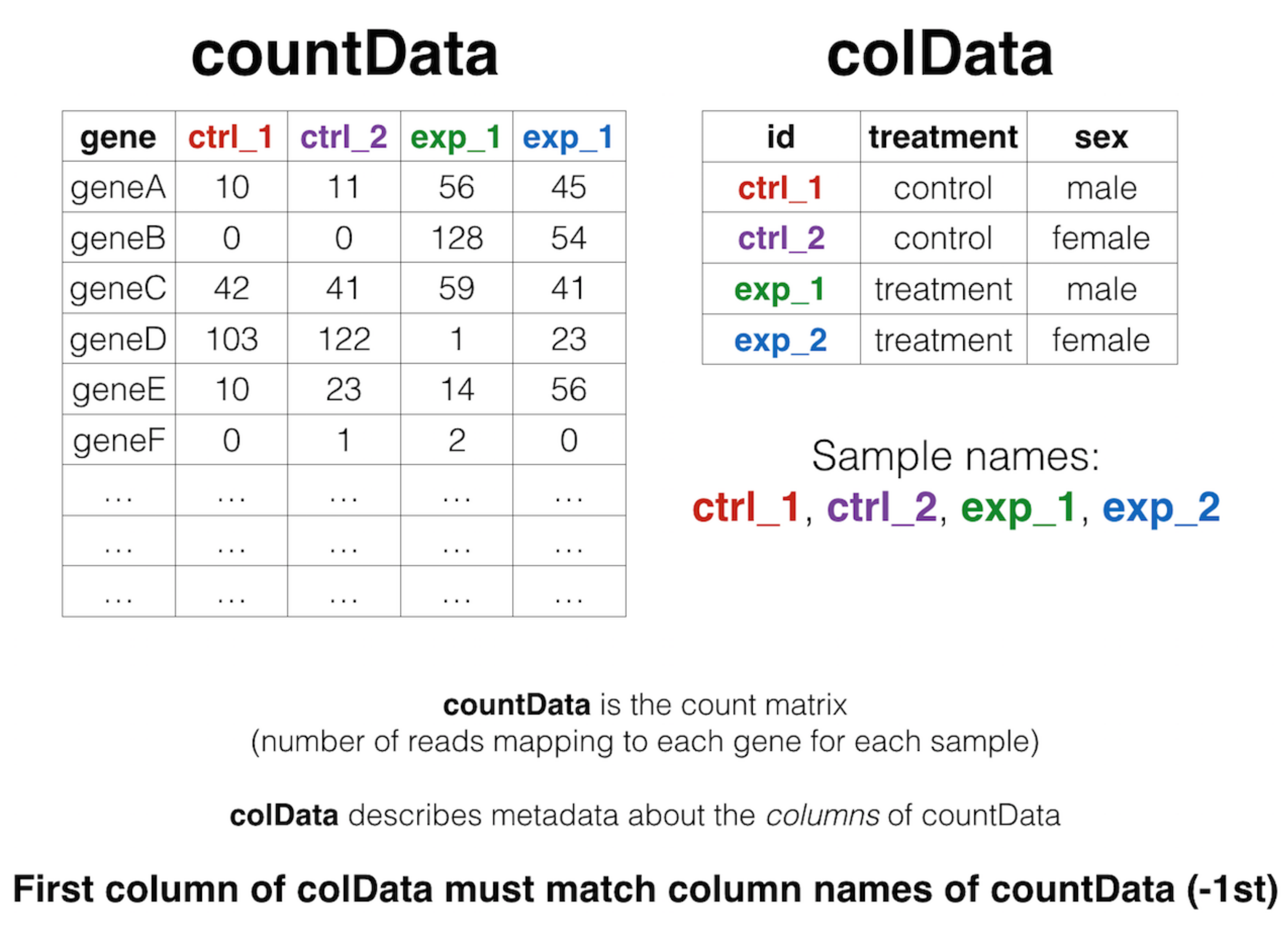


10.5772/intechopen.71769

How to generate genomic data: RNA-seq



RNA-seq count table



countData

gene	ctrl_1	ctrl_2	exp_1	exp_1
geneA	10	11	56	45
geneB	0	0	128	54
geneC	42	41	59	41
geneD	103	122	1	23
geneE	10	23	14	56
geneF	0	1	2	0
…	…	…	…	…
…	…	…	…	…
…	…	…	…	…

colData

id	treatment	sex
ctrl_1	control	male
ctrl_2	control	female
exp_1	treatment	male
exp_2	treatment	female

Sample names:
ctrl_1, **ctrl_2**, **exp_1**, **exp_2**

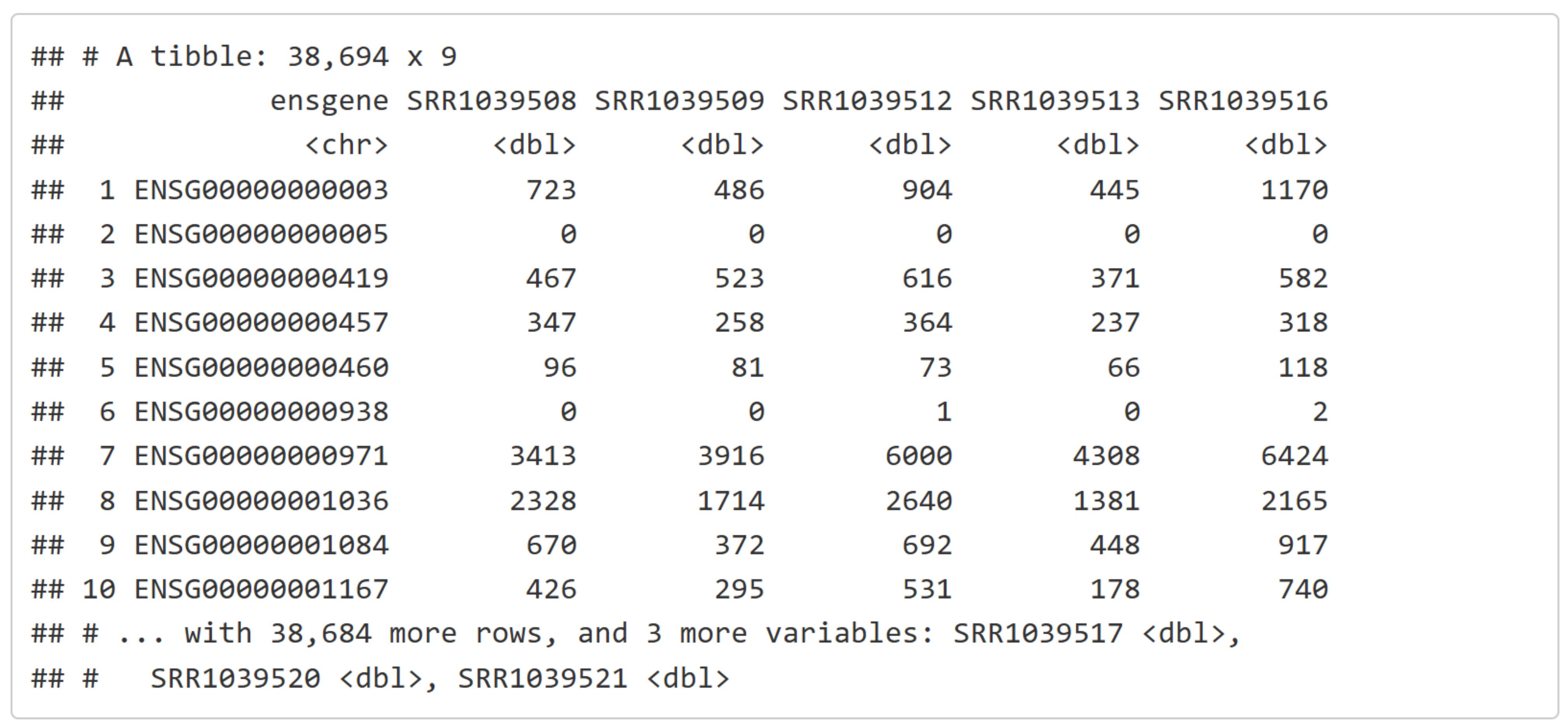
countData is the count matrix
(number of reads mapping to each gene for each sample)

colData describes metadata about the *columns* of countData

First column of colData must match column names of countData (-1st)

https://bioconnector.github.io/workshops/r-rnaseq-airway.html

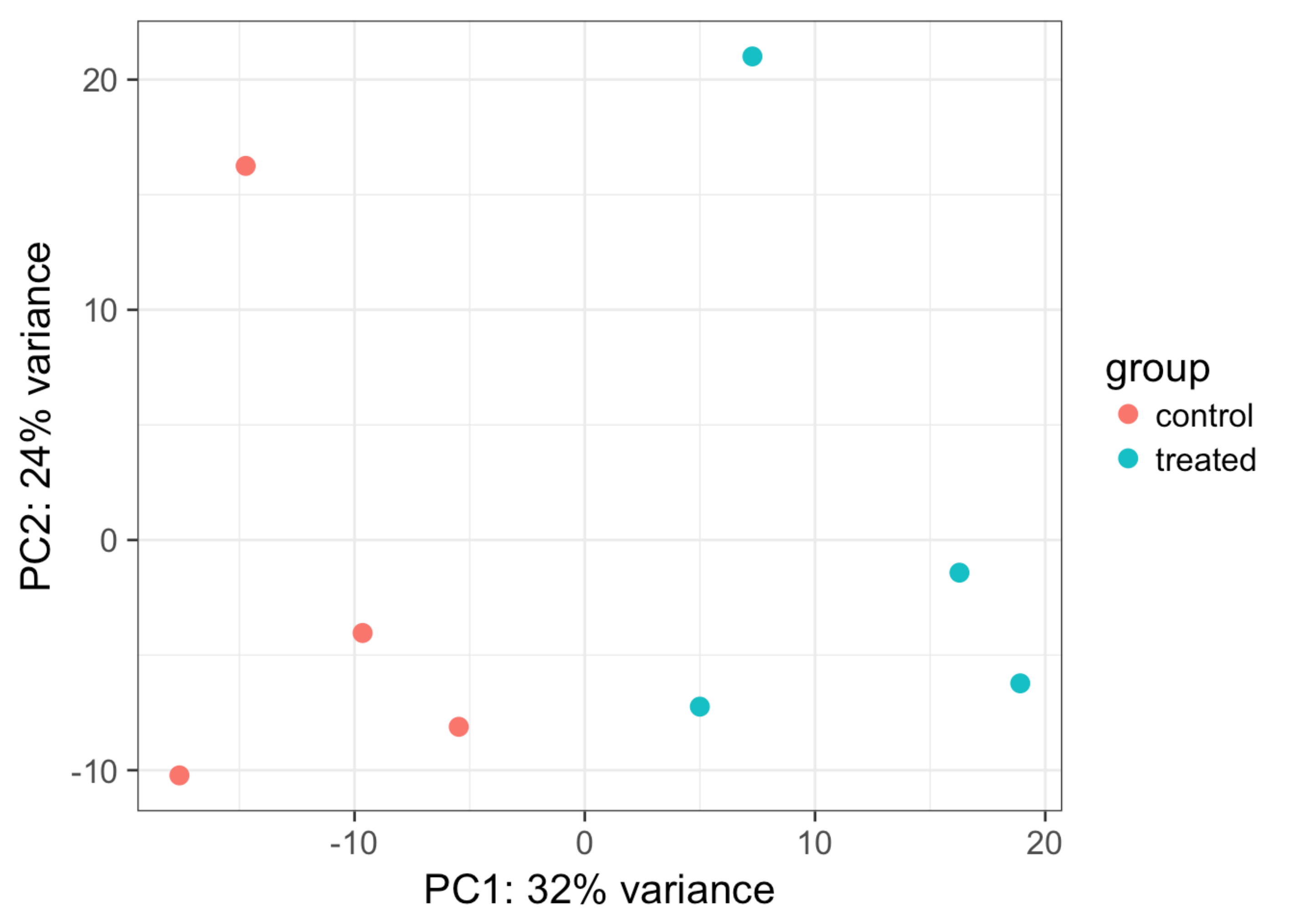
RNA-seq count table



```
## # A tibble: 38,694 x 9
##            ensgene SRR1039508 SRR1039509 SRR1039512 SRR1039513 SRR1039516
##              <chr>      <dbl>      <dbl>      <dbl>      <dbl>      <dbl>
##  1 ENSG00000000003        723        486        904        445       1170
##  2 ENSG00000000005          0          0          0          0          0
##  3 ENSG00000000419        467        523        616        371        582
##  4 ENSG00000000457        347        258        364        237        318
##  5 ENSG00000000460         96         81         73         66        118
##  6 ENSG00000000938          0          0          1          0          2
##  7 ENSG00000000971       3413       3916       6000       4308       6424
##  8 ENSG00000001036       2328       1714       2640       1381       2165
##  9 ENSG00000001084        670        372        692        448        917
## 10 ENSG00000001167        426        295        531        178        740
## # ... with 38,684 more rows, and 3 more variables: SRR1039517 <dbl>,
## #   SRR1039520 <dbl>, SRR1039521 <dbl>
```

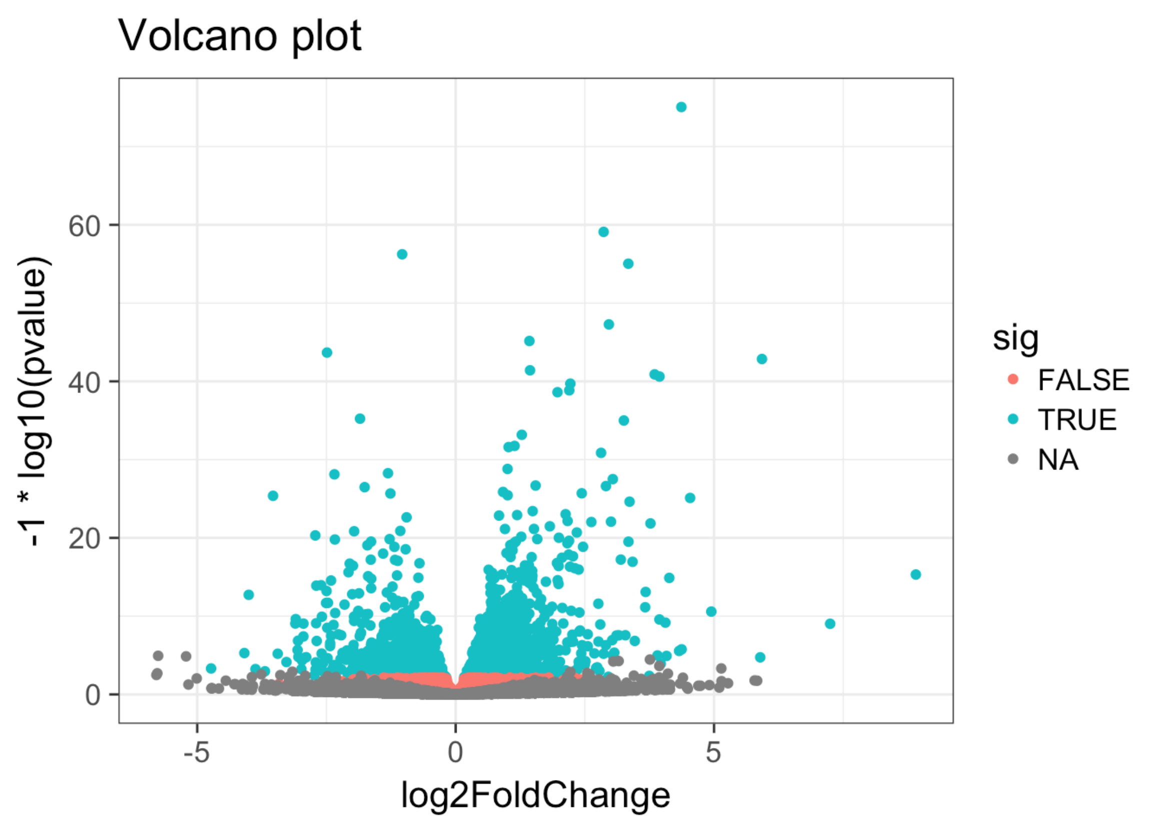
https://bioconnector.github.io/workshops/r-rnaseq-airway.html

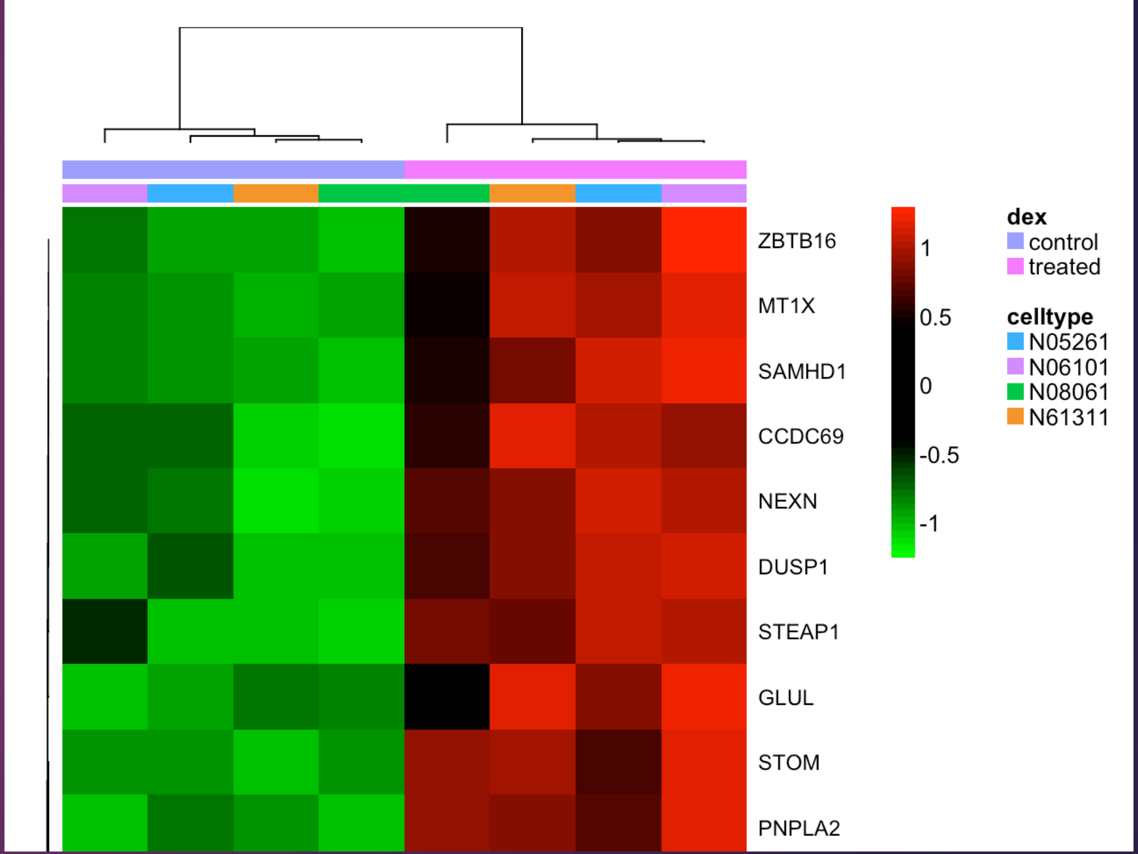
RNA-seq Downstream Analysis



https://bioconnector.github.io/workshops/r-rnaseq-airway.html

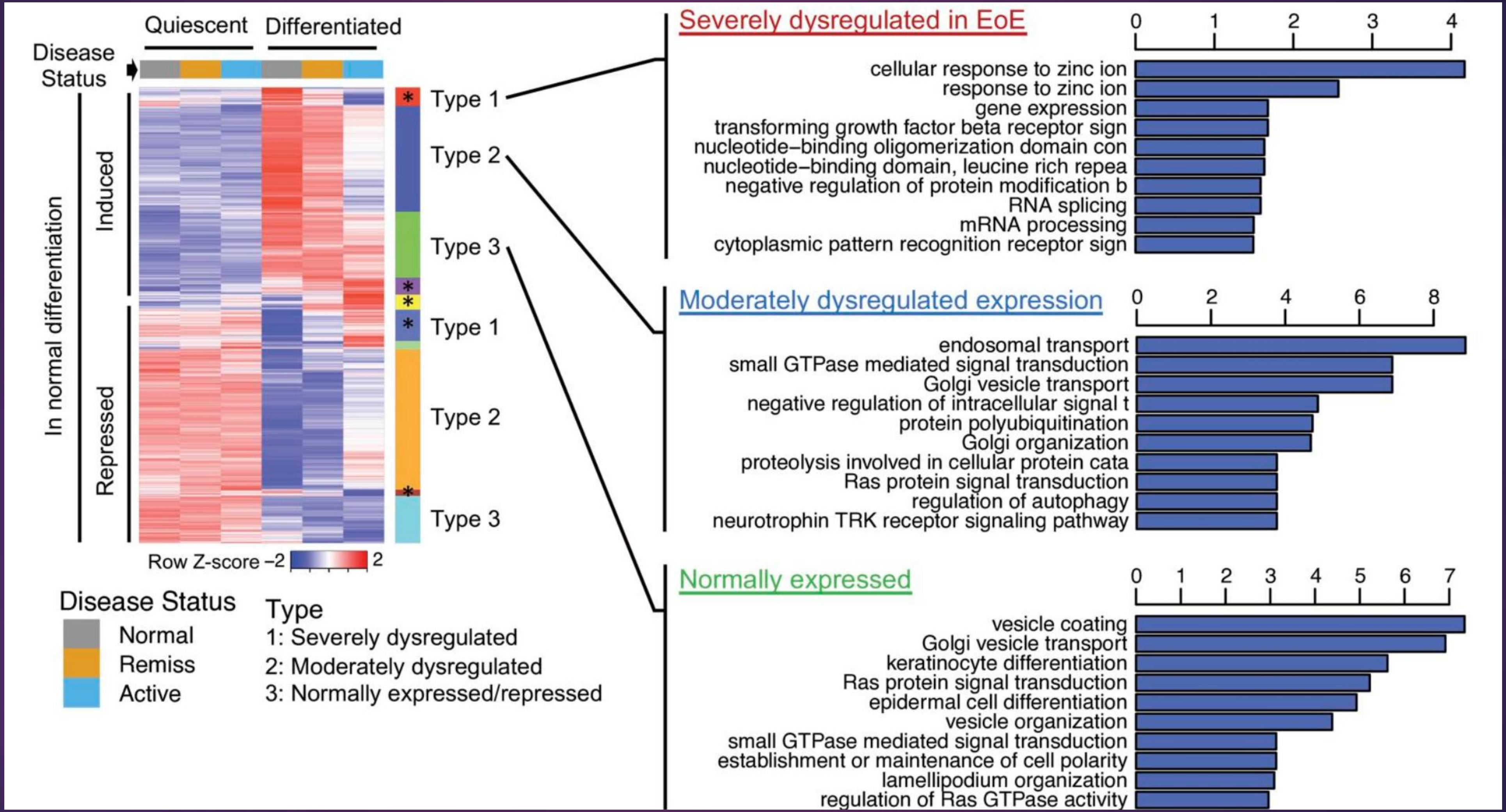
RNA-seq Downstream Analysis





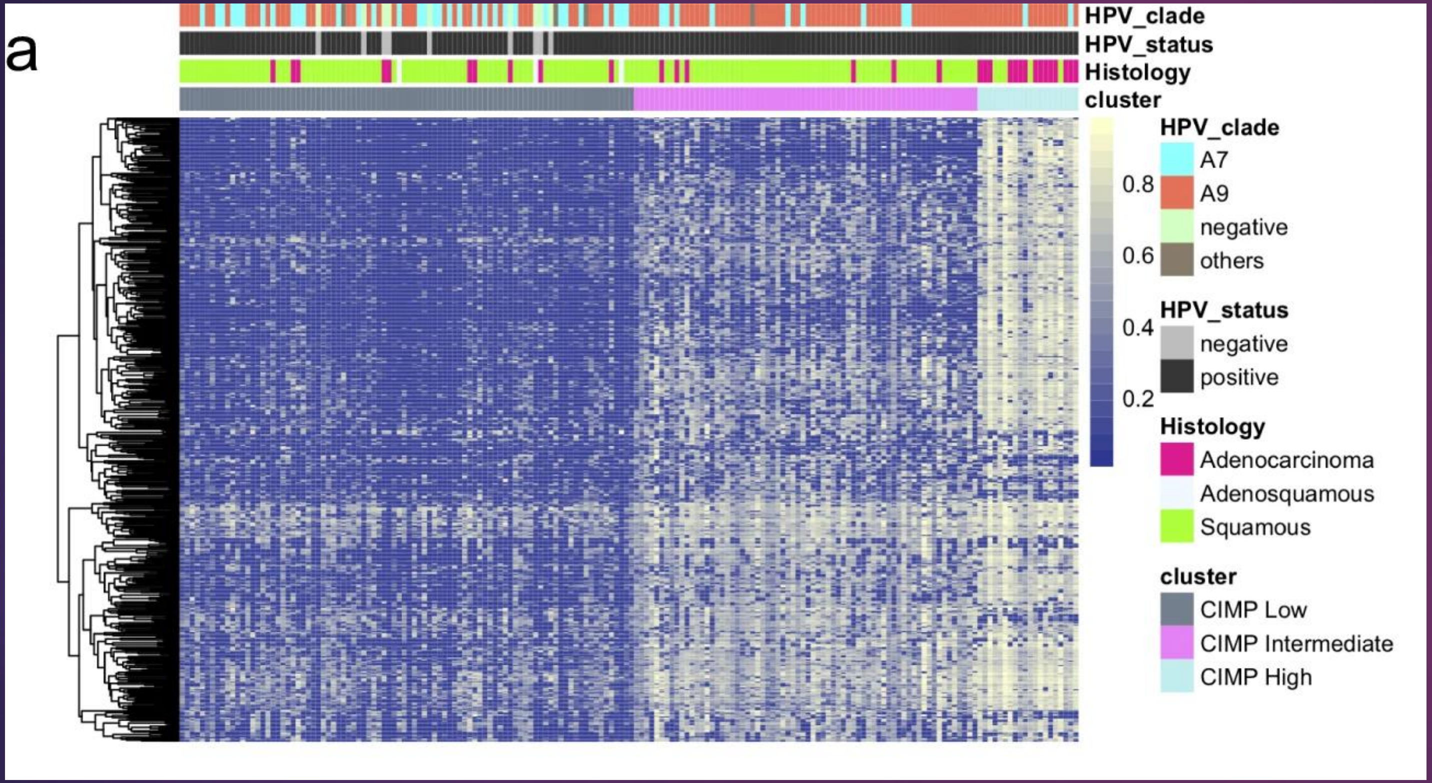
https://bioconnector.github.io/workshops/r-rnaseq-airway.html

RNA-seq Downstream Analysis



https://insight.jci.org/articles/view/159093/figure/5

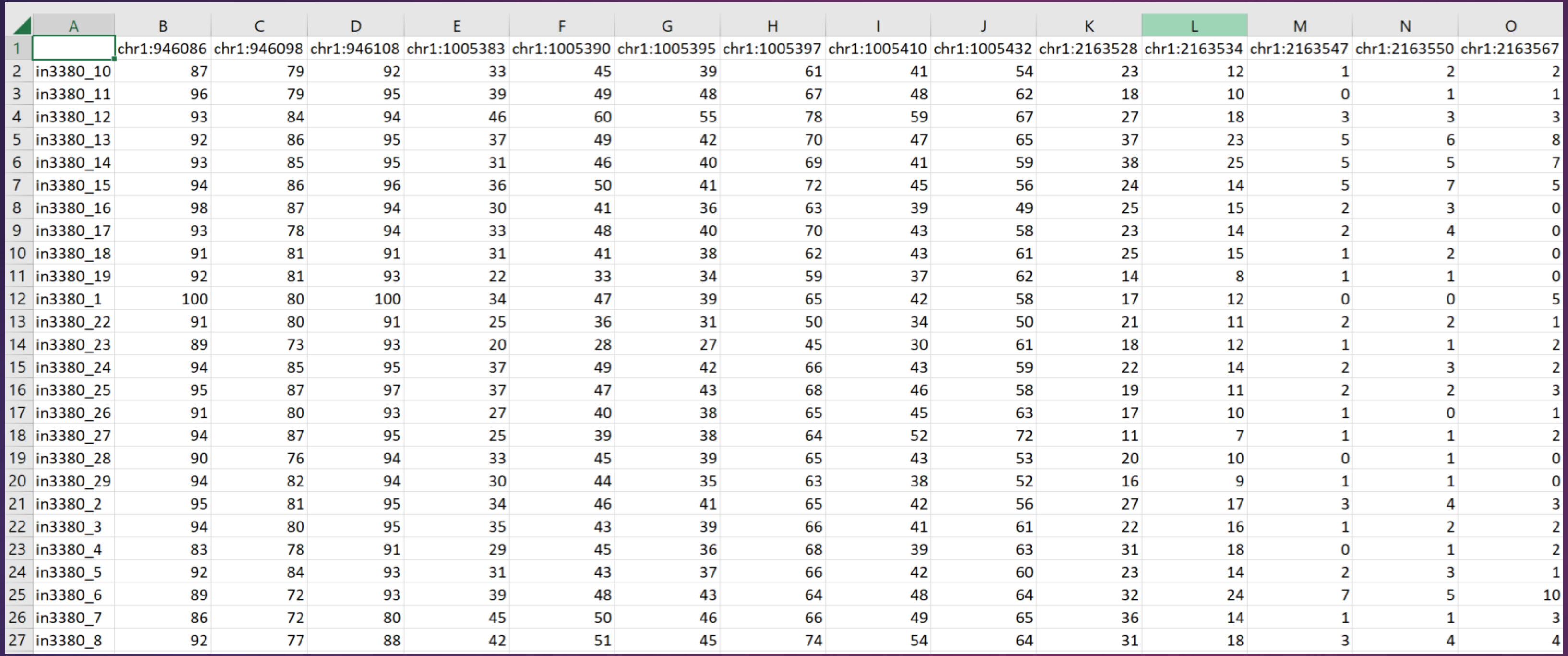
BioMedical data: DNA methylation (p >> n)



n = number of samples 6, 10, 100, 1k
P = number of CpG 28M

https://www.nature.com/articles/s41392-019-0081-6/figures/1

BioMedical data: DNA methylation (p >> n)



	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O
1		chr1:946086	chr1:946098	chr1:946108	chr1:1005383	chr1:1005390	chr1:1005395	chr1:1005397	chr1:1005410	chr1:1005432	chr1:2163528	chr1:2163534	chr1:2163547	chr1:2163550	chr1:2163567
2	in3380_10	87	79	92	33	45	39	61	41	54	23	12	1	2	2
3	in3380_11	96	79	95	39	49	48	67	48	62	18	10	0	1	1
4	in3380_12	93	84	94	46	60	55	78	59	67	27	18	3	3	3
5	in3380_13	92	86	95	37	49	42	70	47	65	37	23	5	6	8
6	in3380_14	93	85	95	31	46	40	69	41	59	38	25	5	5	7
7	in3380_15	94	86	96	36	50	41	72	45	56	24	14	5	7	5
8	in3380_16	98	87	94	30	41	36	63	39	49	25	15	2	3	0
9	in3380_17	93	78	94	33	48	40	70	43	58	23	14	2	4	0
10	in3380_18	91	81	91	31	41	38	62	43	61	25	15	1	2	0
11	in3380_19	92	81	93	22	33	34	59	37	62	14	8	1	1	0
12	in3380_1	100	80	100	34	47	39	65	42	58	17	12	0	0	5
13	in3380_22	91	80	91	25	36	31	50	34	50	21	11	2	2	1
14	in3380_23	89	73	93	20	28	27	45	30	61	18	12	1	1	2
15	in3380_24	94	85	95	37	49	42	66	43	59	22	14	2	3	2
16	in3380_25	95	87	97	37	47	43	68	46	58	19	11	2	2	3
17	in3380_26	91	80	93	27	40	38	65	45	63	17	10	1	0	1
18	in3380_27	94	87	95	25	39	38	64	52	72	11	7	1	1	2
19	in3380_28	90	76	94	33	45	39	65	43	53	20	10	0	1	0
20	in3380_29	94	82	94	30	44	35	63	38	52	16	9	1	1	0
21	in3380_2	95	81	95	34	46	41	65	42	56	27	17	3	4	3
22	in3380_3	94	80	95	35	43	39	66	41	61	22	16	1	2	2
23	in3380_4	83	78	91	29	45	36	68	39	63	31	18	0	1	2
24	in3380_5	92	84	93	31	43	37	66	42	60	23	14	2	3	1
25	in3380_6	89	72	93	39	48	43	64	48	64	32	24	7	5	10
26	in3380_7	86	72	80	45	50	46	66	49	65	36	14	1	1	3
27	in3380_8	92	77	88	42	51	45	74	54	64	31	18	3	4	4

Columns chr___: 1254 CpG sites

Columns y:

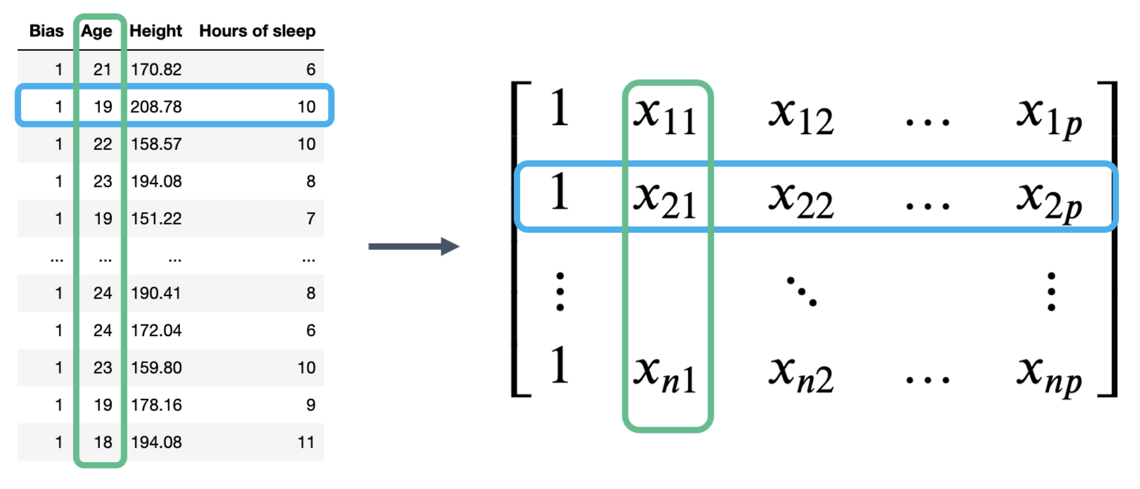
0: normal sample,
1: insignificant tumor,
2: significant tumor

Rows: 334 samples

What is feature/variable selection?

- Find the features (variables/columns) in X which are important for predicting, and remove the features that are not
- Give:

$$\underbrace{\begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix}}_{\mathbb{Y}} = \underbrace{\begin{bmatrix} 1 & x_{11} & x_{12} & \dots & x_{1p} \\ 1 & x_{21} & x_{22} & \dots & x_{2p} \\ \vdots & & \ddots & & \vdots \\ 1 & x_{n1} & x_{n2} & \dots & x_{np} \end{bmatrix}}_{\mathbb{X}} \underbrace{\begin{bmatrix} \theta_0 \\ \theta_1 \\ \vdots \\ \theta_p \end{bmatrix}}_{\theta}$$



Bias	Age	Height	Hours of sleep
1	21	170.82	6
1	19	208.78	10
1	22	158.57	10
1	23	194.08	8
1	19	151.22	7
...	...	...	...
1	24	190.41	8
1	24	172.04	6
1	23	159.80	10
1	19	178.16	9
1	18	194.08	11

$$\rightarrow \begin{bmatrix} 1 & x_{11} & x_{12} & \dots & x_{1p} \\ 1 & x_{21} & x_{22} & \dots & x_{2p} \\ \vdots & & \ddots & & \vdots \\ 1 & x_{n1} & x_{n2} & \dots & x_{np} \end{bmatrix}$$

- Find the columns in X which are important for predicting y

https://ds100.org/course-notes/ols/ols.html

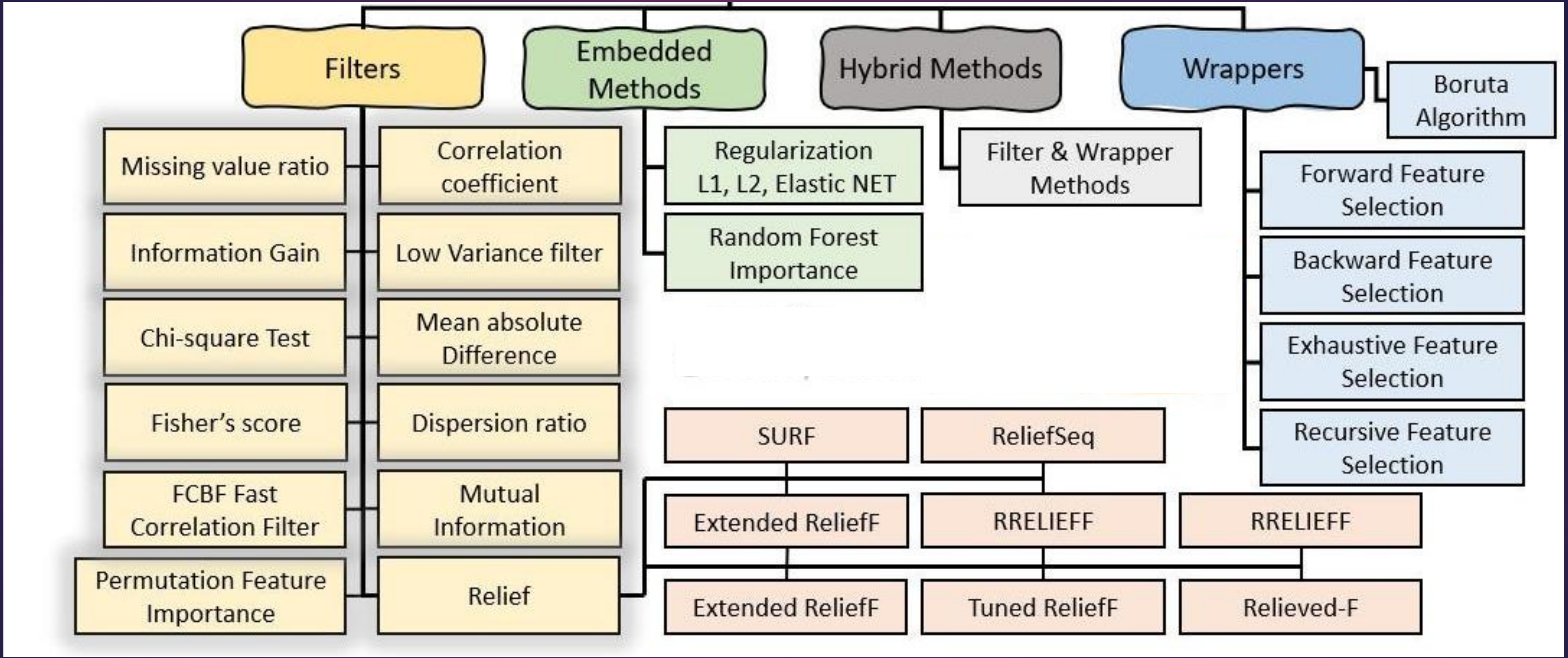
https://ubc-cs.github.io/cpsc330-2023W1/lectures/13_feature-engineering-selection.html

Why feature selection?

- Interpretability: Models are more interpretable with fewer features.
- If you get the same performance with 10 features instead of 500 features, why not use the model with smaller number of features?
- Computation: Models fit/predict faster with fewer columns.
- Data collection: What type of new data should I collect? It may be cheaper to collect fewer columns.
- Fundamental tradeoff: Can I reduce overfitting by removing useless features?
- Feature selection can often result in better performing (less overfit), easier to understand, and faster model.

https://ubc-cs.github.io/cpsc330-2023W1/lectures/13_feature-engineering-selection.html

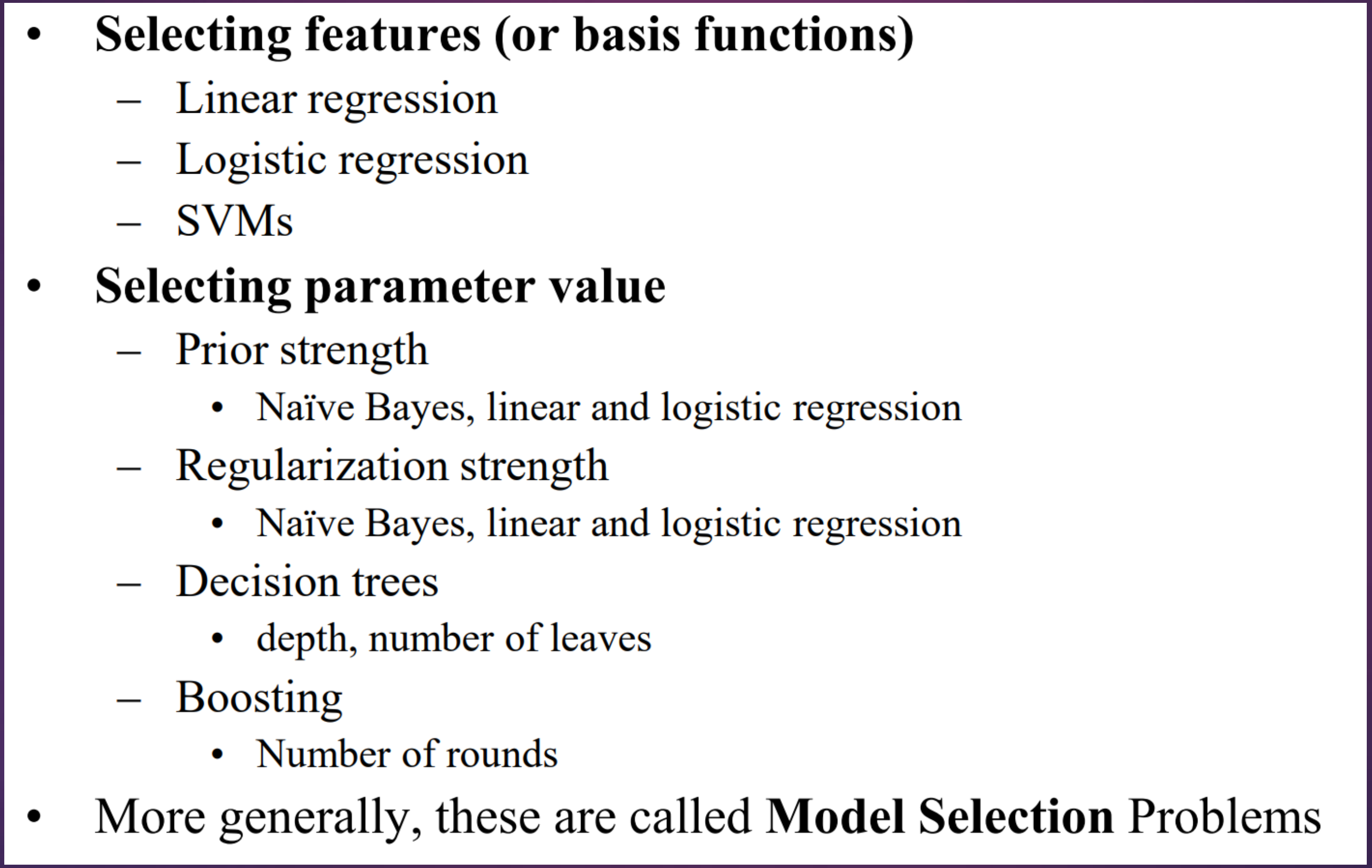
How do we carry out feature selection?
Supervised Feature Selection



https://medium.com/analytics-vidhya/feature-selection-extended-overview-b58f1d524c1c

Model Selection vs Feature Selection

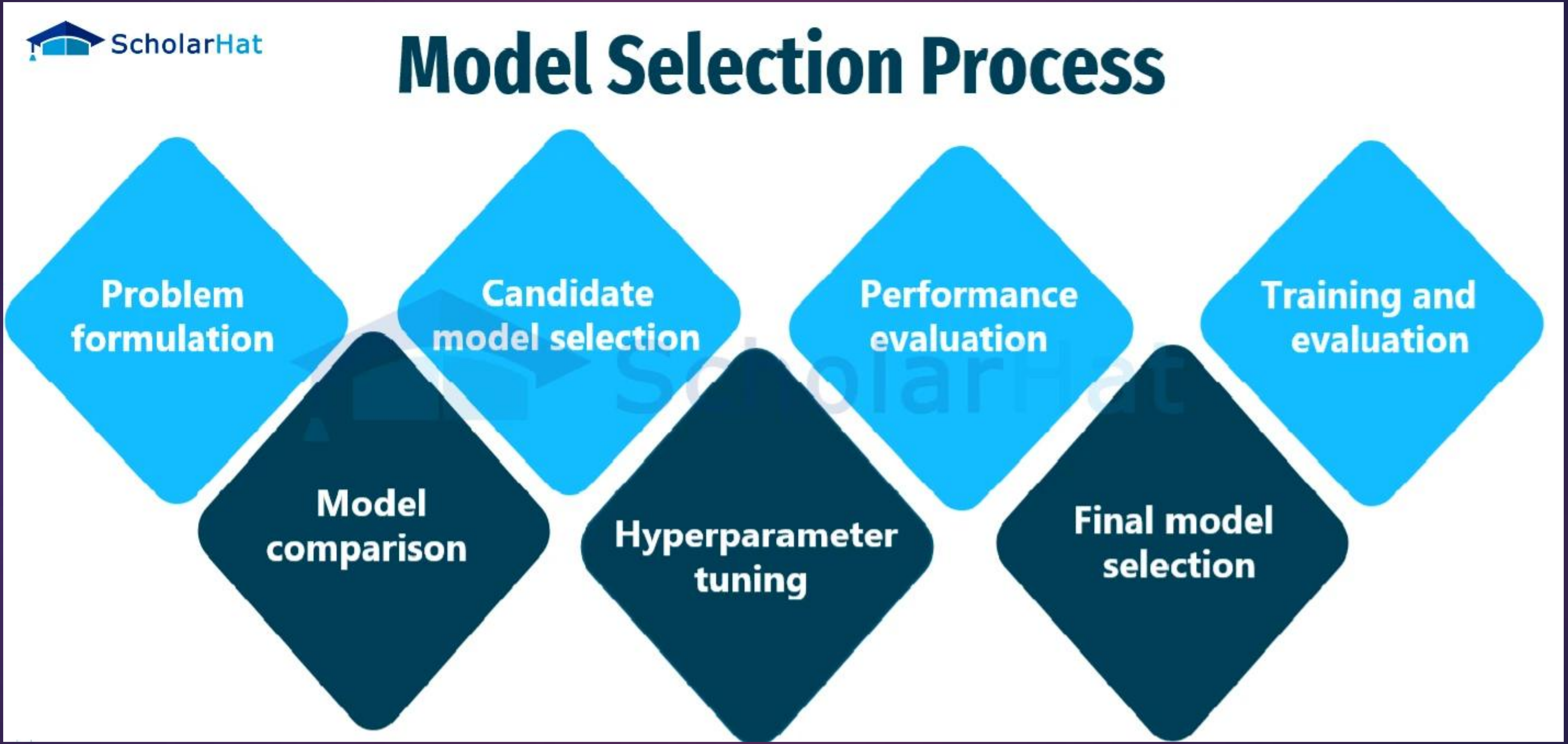
- Feature Selection is a part of Model Selection



- **Selecting features (or basis functions)**
  - Linear regression
  - Logistic regression
  - SVMs
- **Selecting parameter value**
  - Prior strength
    - Naïve Bayes, linear and logistic regression
  - Regularization strength
    - Naïve Bayes, linear and logistic regression
  - Decision trees
    - depth, number of leaves
  - Boosting
    - Number of rounds
- More generally, these are called **Model Selection** Problems

https://www.cs.cmu.edu/afs/cs/academic/class/10601-f10/lecture/lec16.pdf

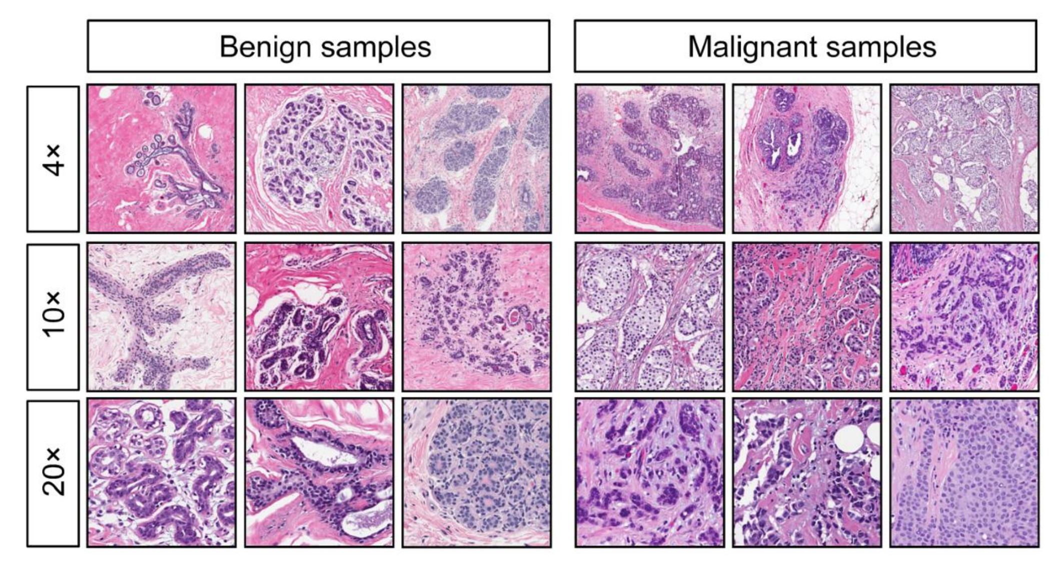
Model selection: steps



https://www.scholarhat.com/tutorial/machinelearning/model-selection-for-machine-learning

Model selection: steps

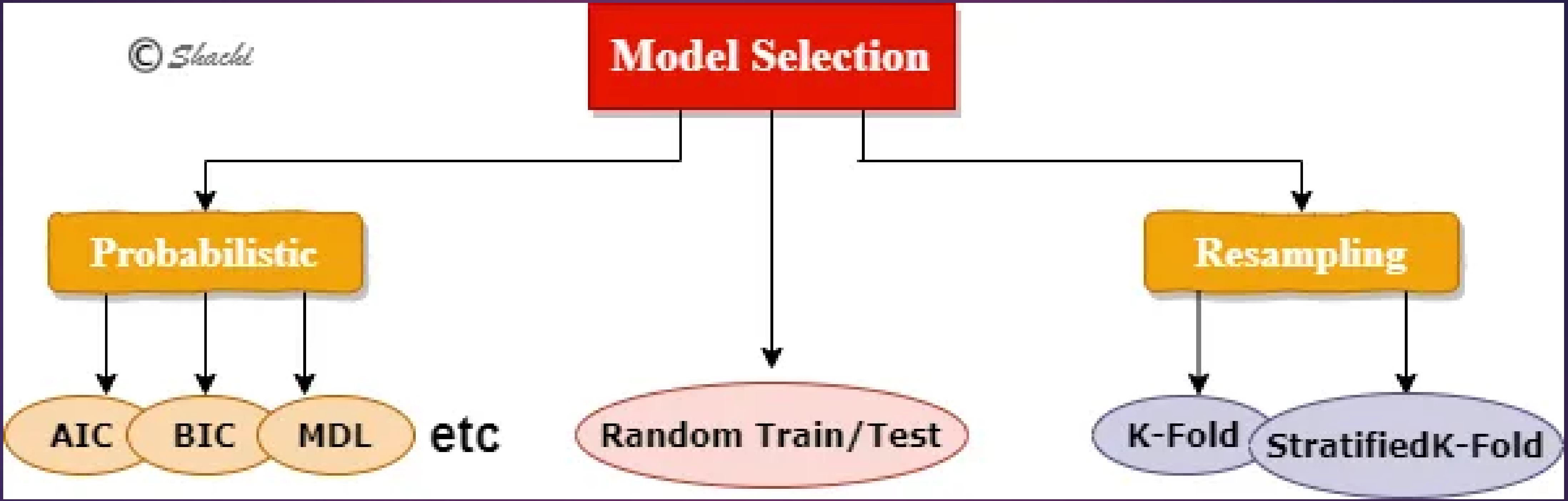
Stage 1: Selecting the regression model forms
Stage 2: Selecting the regression model and the independent variables
Stage 3: Fitting the model
Stage 4: Examining or validation of the applied model



https://www.shiksha.com/online-courses/articles/model-selection-in-machine-learning-regression/

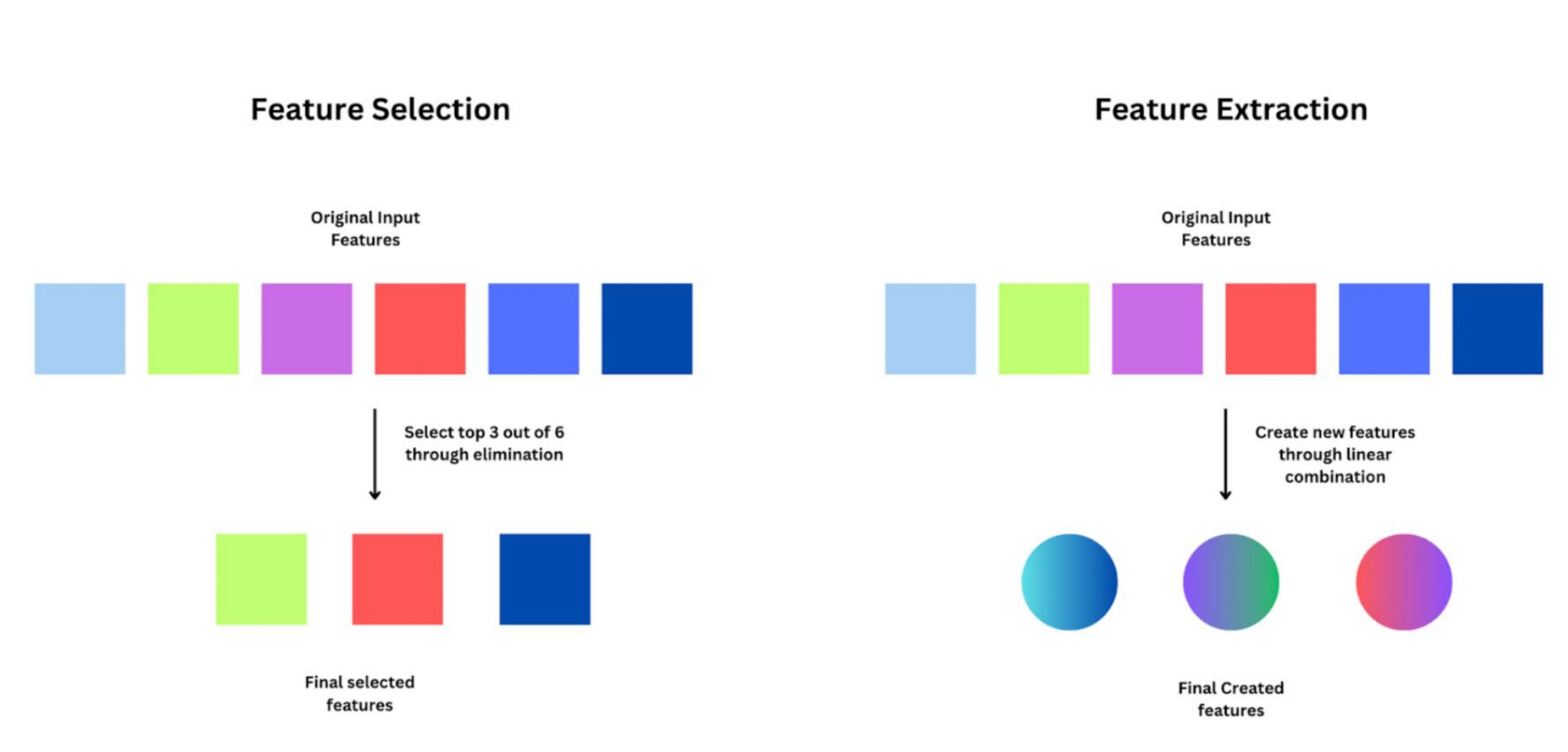
https://medium.com/analytics-vidhya/breast-cancer-diagnostic-dataset-eda-fa0de80f15bd

Model selection: methods



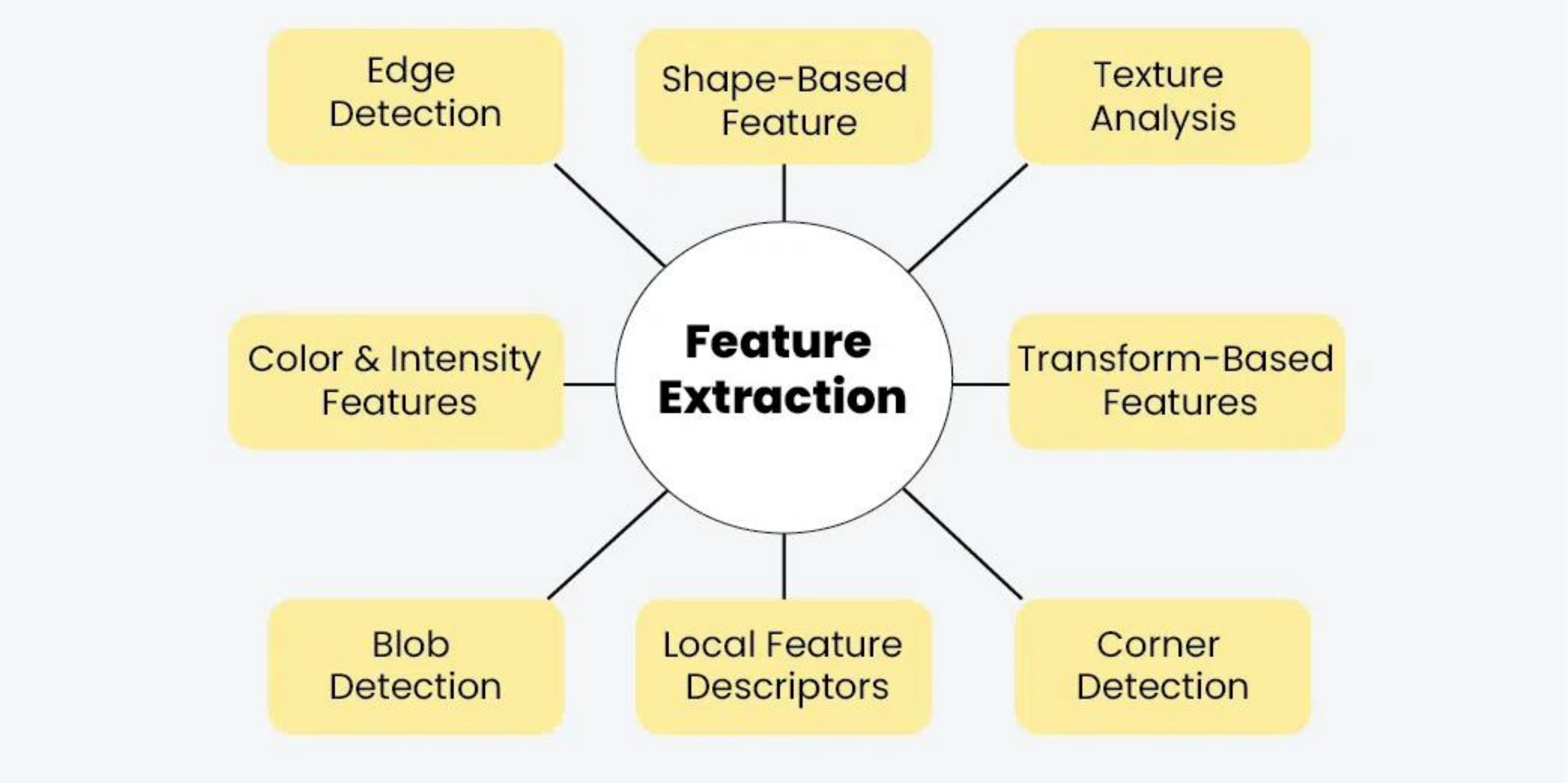
https://medium.com/analytics-vidhya/model-selection-techniques-in-ml-ai-with-python-fdf308d9fa10

Feature Selection vs Feature Extraction/Engineering



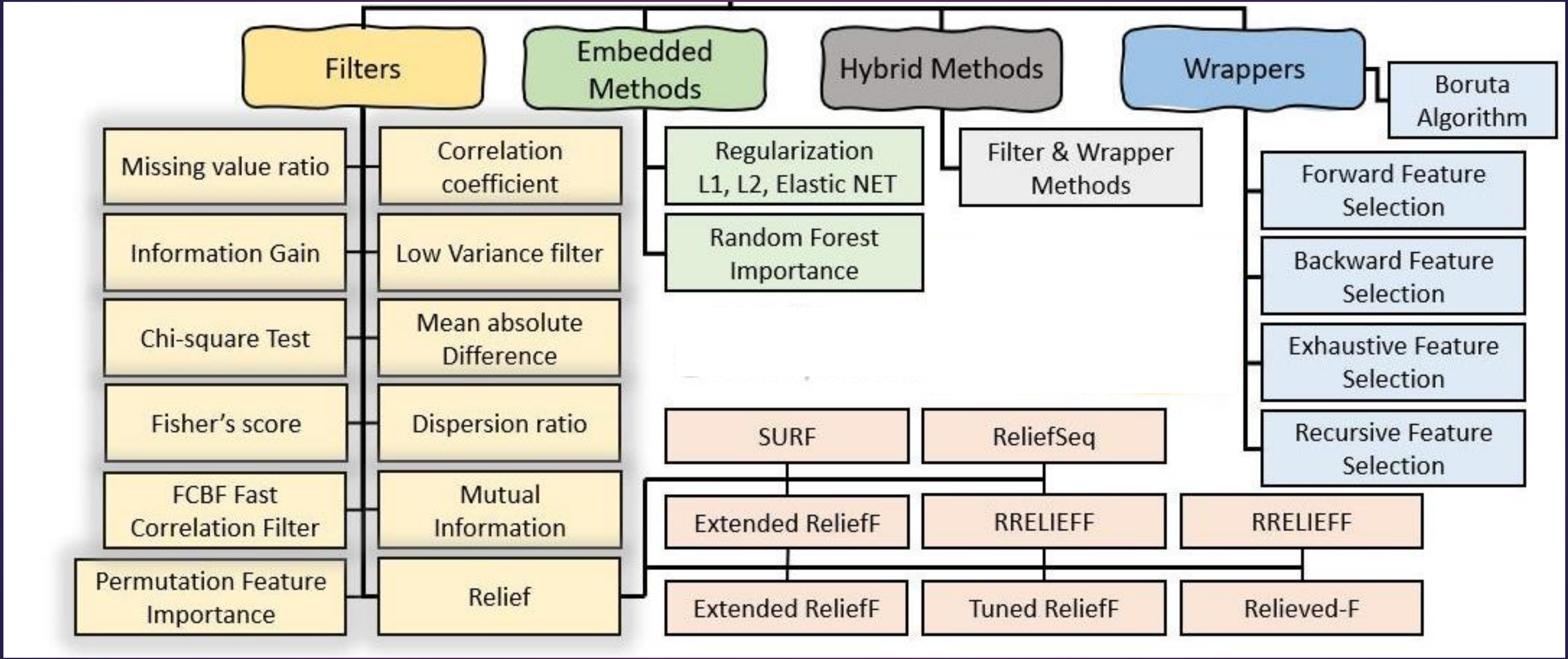
https://viso.ai/deep-learning/feature-extraction-in-python/

Feature Extraction/Engineering



https://www.geeksforgeeks.org/feature-extraction-in-image-processing-techniques-and-applications/

How do we carry out feature selection?
Supervised Feature Selection



- Filters
  - Missing value ratio
  - Information Gain
  - Chi-square Test
  - Fisher’s score
  - FCBF Fast Correlation Filter
  - Permutation Feature Importance
  - Correlation coefficient
  - Low Variance filter
  - Mean absolute Difference
  - Dispersion ratio
  - Mutual Information
  - Relief
    - SURF
    - ReliefSeq
    - Extended ReliefF
    - RRELIEFF
    - RRELIEFF
    - Extended ReliefF
    - Tuned ReliefF
    - Relieved-F
- Embedded Methods
  - Regularization L1, L2, Elastic NET
  - Random Forest Importance
- Hybrid Methods
  - Filter & Wrapper Methods
- Wrappers
  - Boruta Algorithm
  - Forward Feature Selection
  - Backward Feature Selection
  - Exhaustive Feature Selection
  - Recursive Feature Selection

https://medium.com/analytics-vidhya/feature-selection-extended-overview-b58f1d524c1c

Xin chân thành cảm ơn!

LUU PHUC LOI, PHD

ZALO: 0901802182

LUU.P.LOI@GOOGLEMAIL.COM