Inria

Inserm
Institut national
de la santé et de la recherche médicale

COMPO
COMPUTATIONAL ONCOLOGY

Introduction to Machine Learning

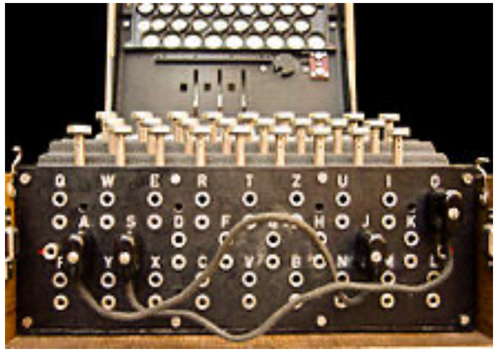
S.Benzekry

Inria – Inserm team COMPO

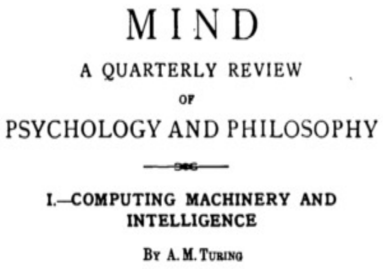
Ok Google: What is Machine Learning (ML)??

Definition: “Machine learning is the field of study that gives the computer the ability to learn *without being explicitly programmed* “ Arthur Samuel, Computer Scientist, 1959

- Exists since decades



Enigma



MIND

A QUARTERLY REVIEW

OF

PSYCHOLOGY AND PHILOSOPHY

I.—COMPUTING MACHINERY AND INTELLIGENCE

By A. M. Turing

Turing, Mind, 1950

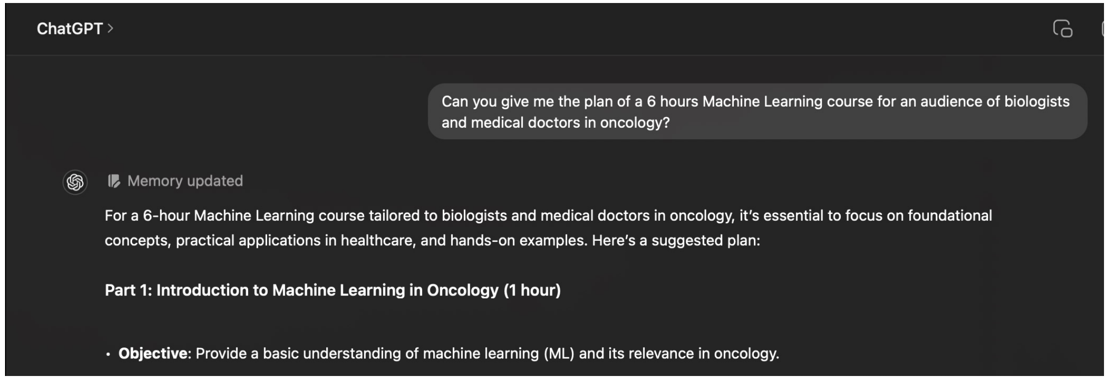
- New « hype » since ~ 2011 mostly thanks to :
  - Computing power
  - Big data

Deep Learning

G. Hinton, Nobel Prize, 2024

SARS-CoV-2 is evolving fast. What will it do next? p. 844
Effects on surface water from hydraulic fracturing pp. 853 & 896
Baby bats babble like human babies p. 923
Science
$15 20 AUGUST 2021 sciencemag.org
AAAS
PREDICTING STRUCTURES
Deep learning accurately folds proteins p. 871

Alphafold, 2021
Hassabis, Nobel Prize, 2024



Can you give me the plan of a 6 hours Machine Learning course for an audience of biologists and medical doctors in oncology?

Memory updated

For a 6-hour Machine Learning course tailored to biologists and medical doctors in oncology, it's essential to focus on foundational concepts, practical applications in healthcare, and hands-on examples. Here's a suggested plan:

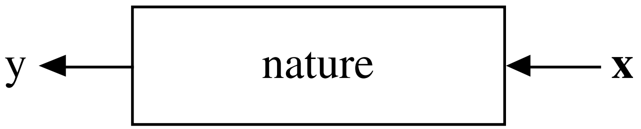
Part 1: Introduction to Machine Learning in Oncology (1 hour)

- **Objective**: Provide a basic understanding of machine learning (ML) and its relevance in oncology.

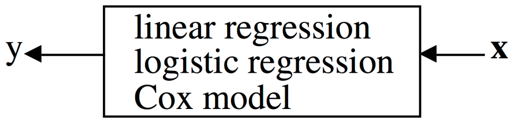
Statistical Science
2001, Vol. 16, No. 3, 199–231

Statistical Modeling: The Two Cultures

Leo Breiman

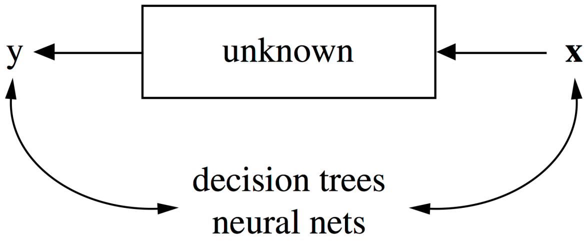


The data modeling culture

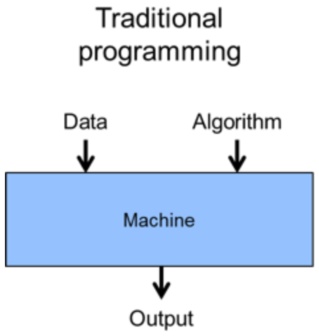


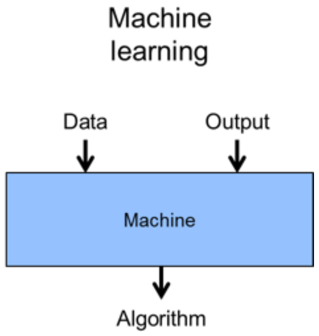
Model validation. Yes–no using goodness-of-fit tests and residual examination.
Estimated culture population. 98% of all statisticians.

The algorithmic modeling culture

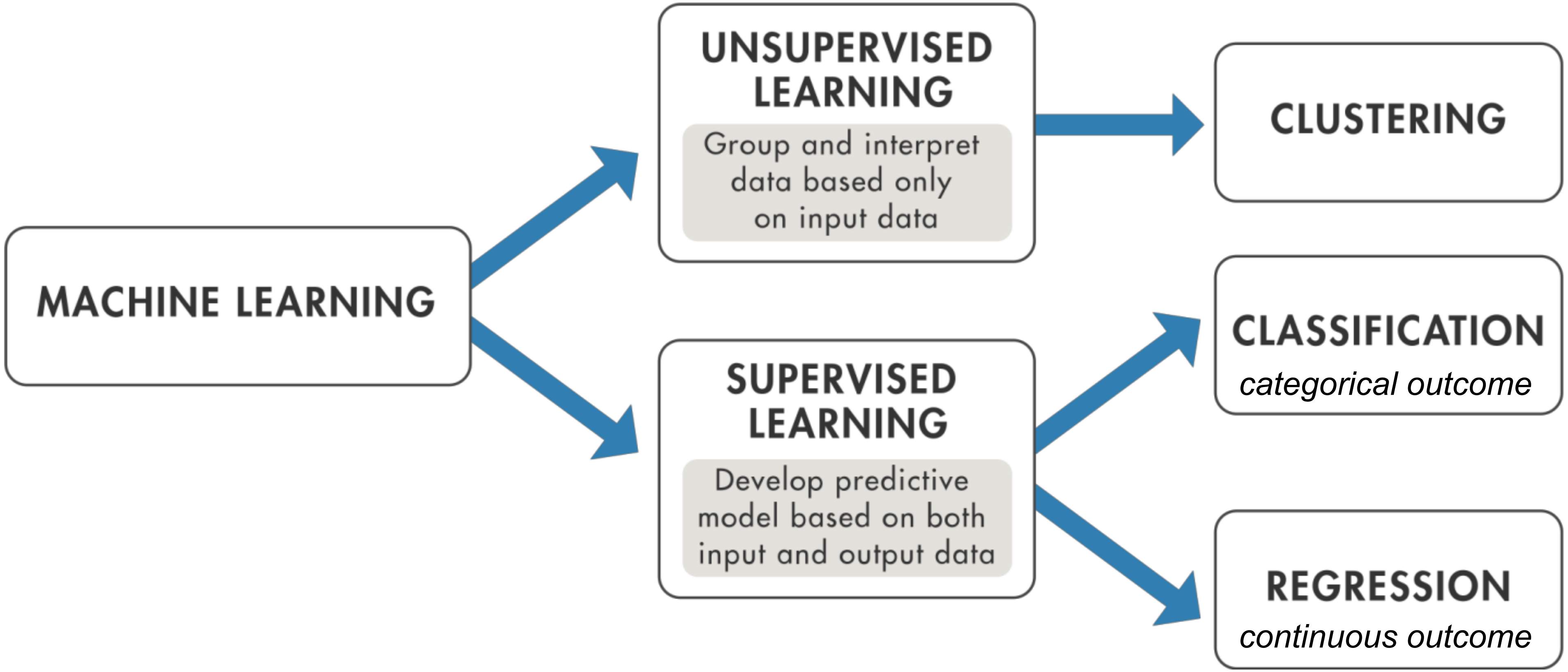


Model validation. Measured by predictive accuracy.
Estimated culture population. 2% of statisticians, many in other fields.





Unsupervised VS supervised ML



MACHINE LEARNING

UNSUPERVISED LEARNING

Group and interpret data based only on input data

CLUSTERING

SUPERVISED LEARNING

Develop predictive model based on both input and output data

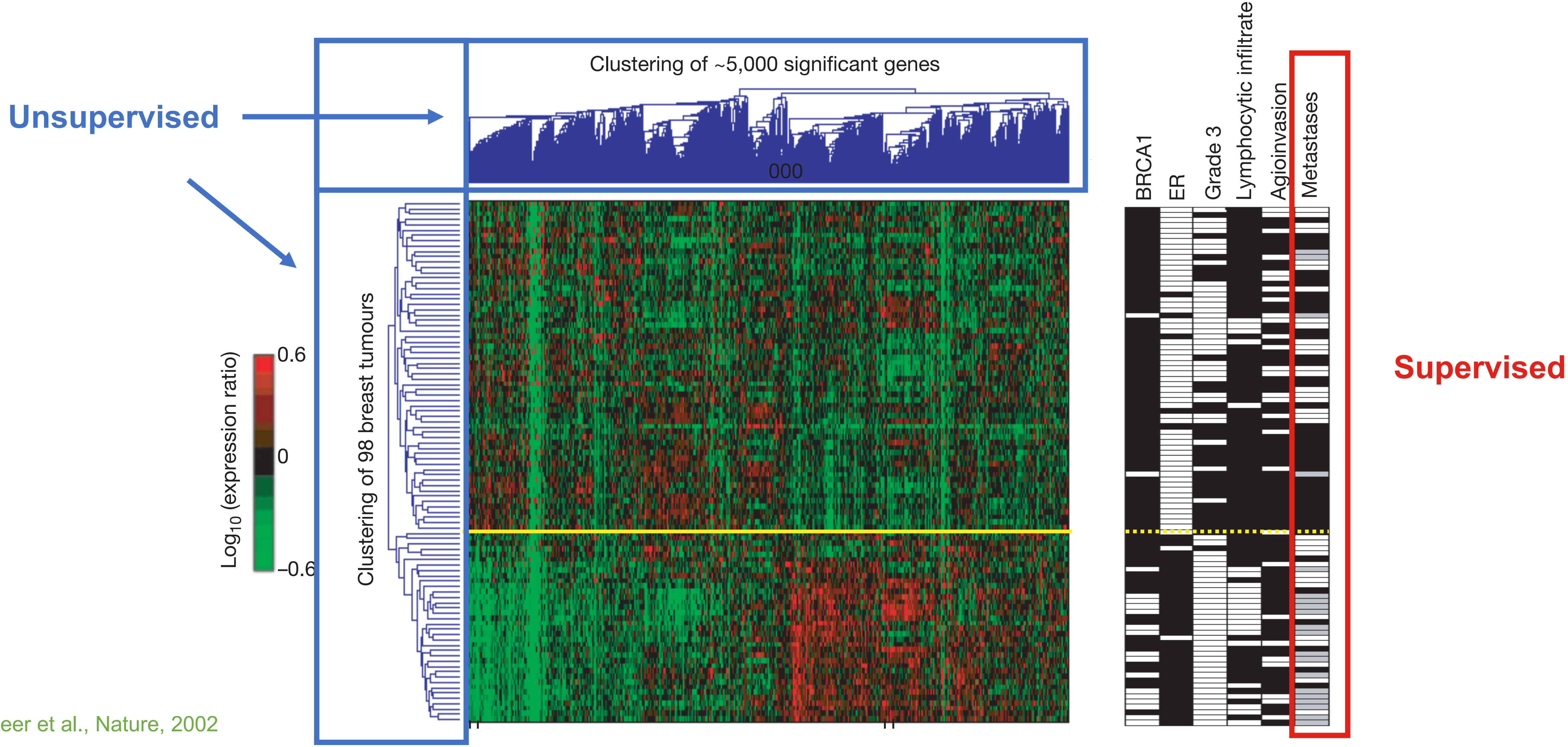
CLASSIFICATION

categorical outcome

REGRESSION

continuous outcome

Example: gene expression and metastatic relapse in breast cancer



van't veer et al., Nature, 2002

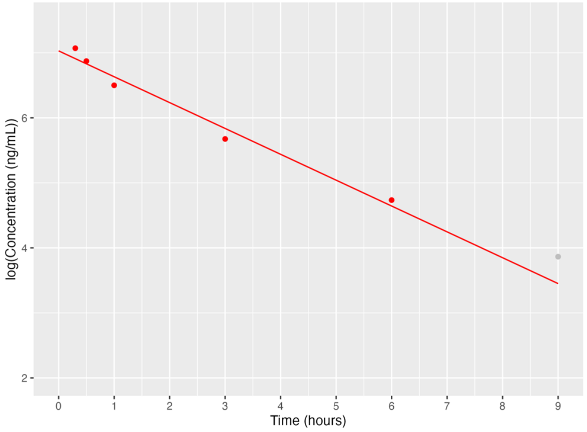
Supervised learning: classification vs regression

Regression: continuous outcome

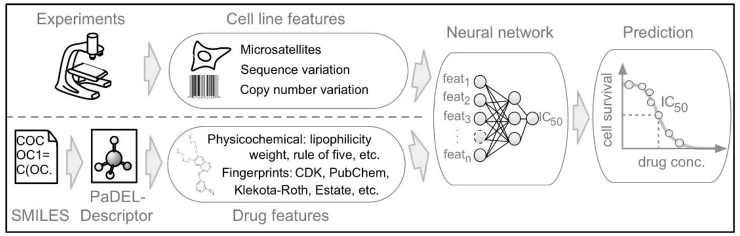
- Predict drug concentration

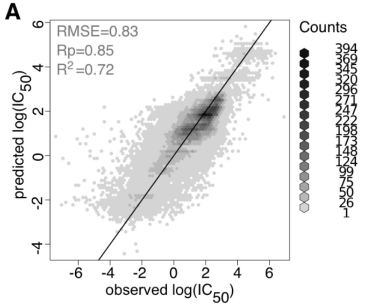
$$x = \{(t_1, C_1), \cdots, (t_k, C_k), t_K\}$$

$$y = C_K$$



- Predict drug IC_{50} from genomic (138) + chemical (689) features

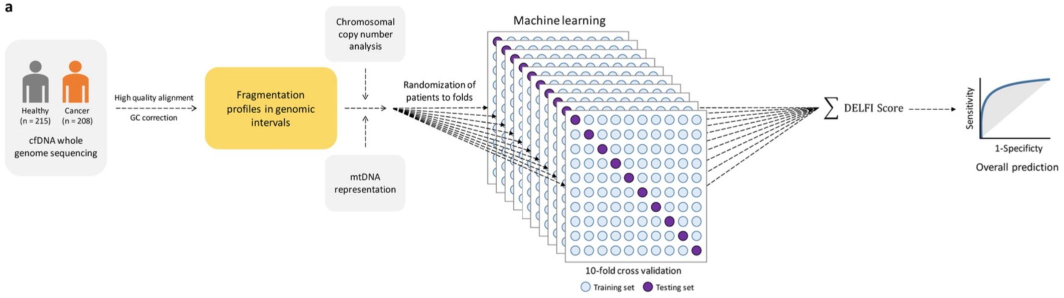




Menden et al., PLoS One, 2013

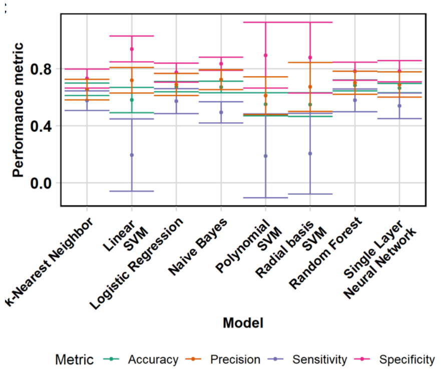
Classification: Categorical outcome

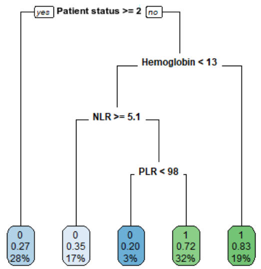
- Cancer vs non-cancer from cfDNA fragmentomics



Cristino et al., Nature, 2019

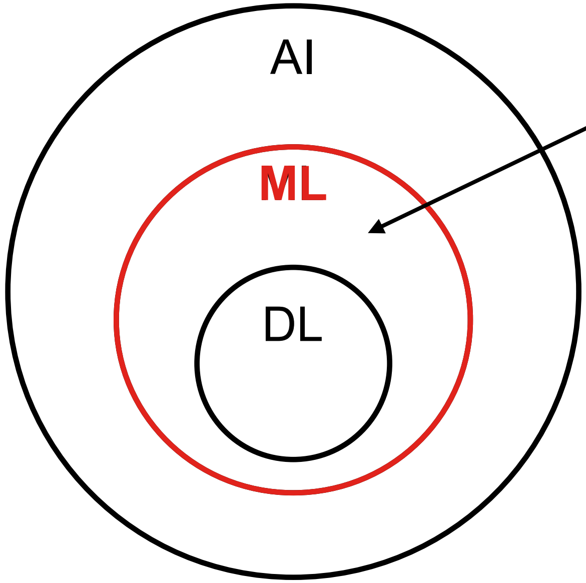
- Response to immunotherapy from blood markers





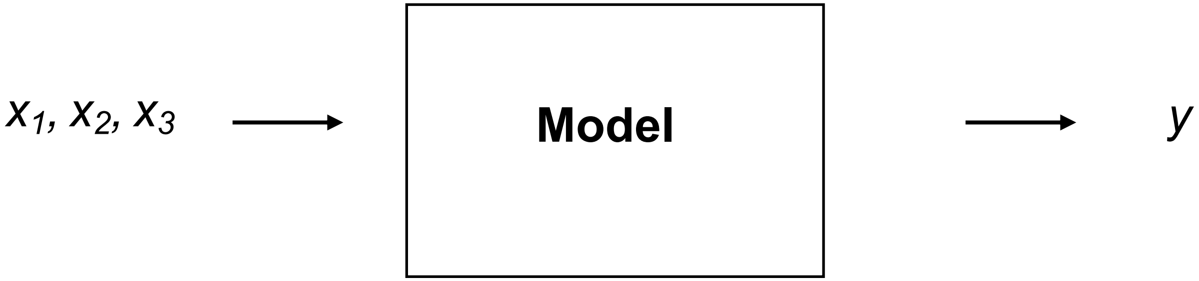
Benzekry et al., Cancers, 2021

Artificial Intelligence, Machine Learning and Deep Learning

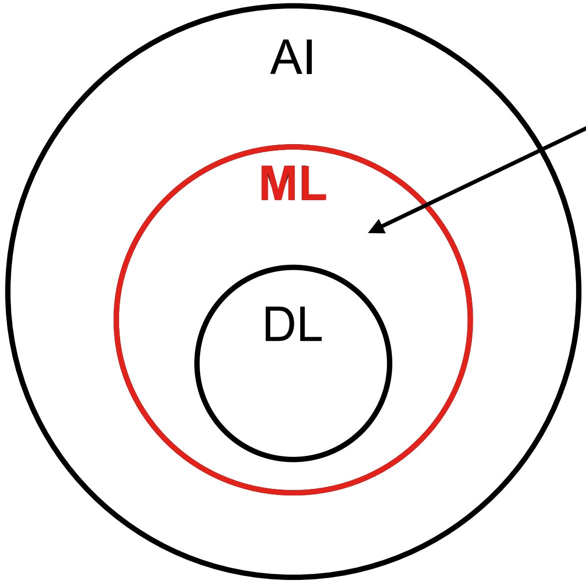


ML = machine (automatic) learning

Goal = predict outcome y as a function of input / features $x_1,\ldots,x_n$



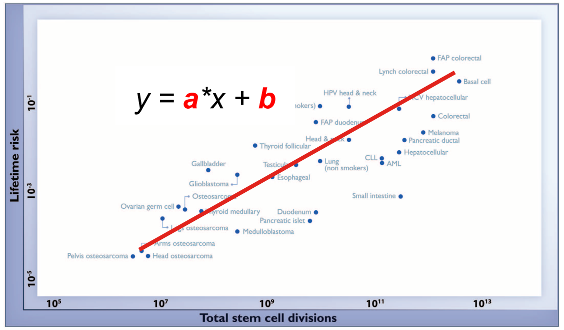
Artificial Intelligence, Machine Learning and Deep Learning



ML = machine (automatic) learning

Goal = predict outcome y as a function of input / features $x_1, \ldots, x_n$

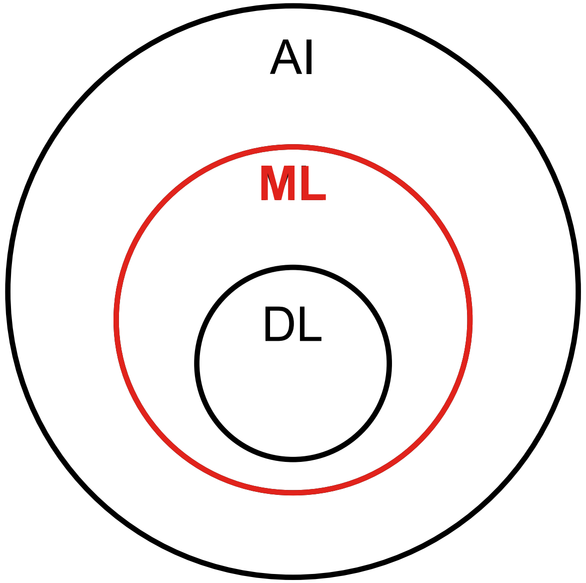
x_1, x_2, x_3 →



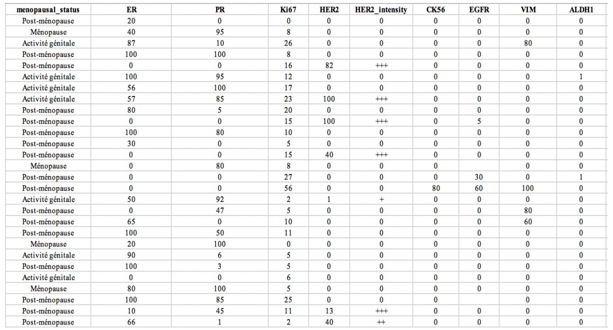
→ y

Artificial Intelligence, Machine Learning and Deep Learning

Supervised machine learning

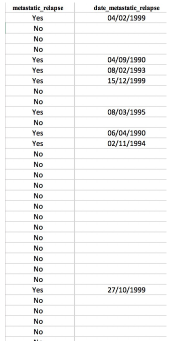


Features

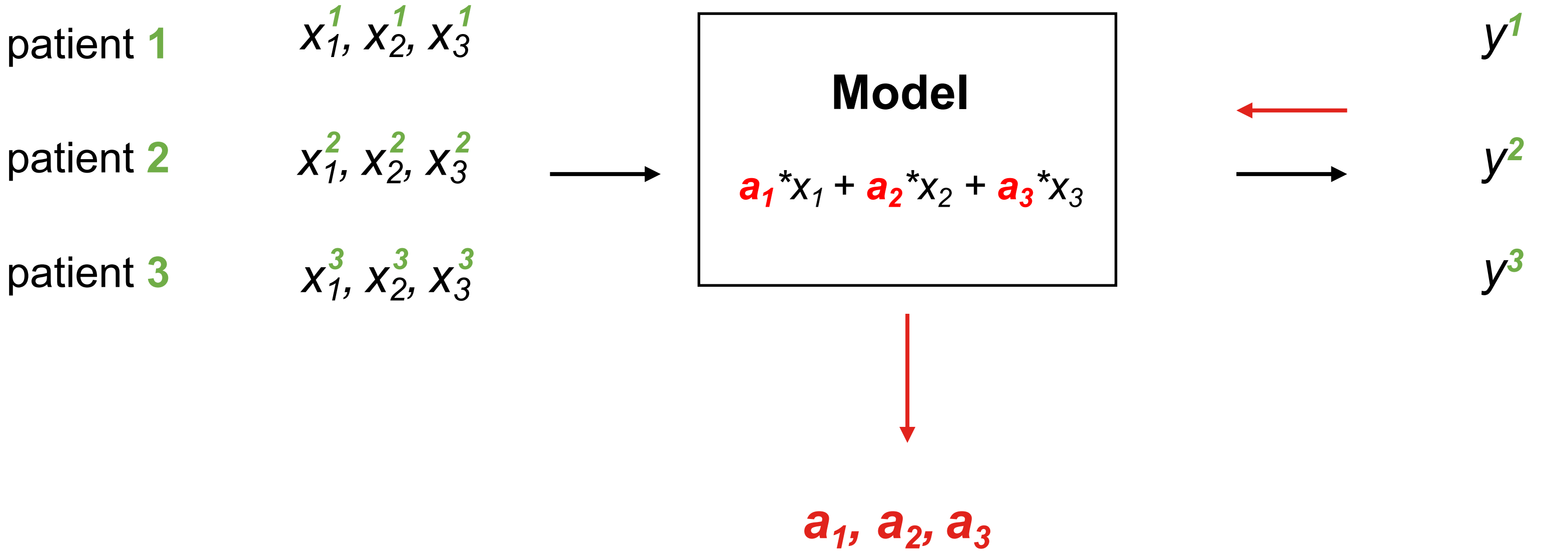


menopausal_status	ER	PR	Ki67	HER2	HER2_intensity	CK56	EGFR	VIM	ALDH1
Post-ménopause	20	0	0	0	0	0	0	0	0
Ménopause	40	95	8	0	0	0	0	0	0
Activité génitale	87	10	26	0	0	0	0	80	0
Post-ménopause	100	100	8	0	0	0	0	0	0
Post-ménopause	0	0	16	82	+++	0	0	0	0
Activité génitale	100	95	12	0	0	0	0	0	1
Activité génitale	56	100	17	0	0	0	0	0	0
Activité génitale	57	85	23	100	+++	0	0	0	0
Post-ménopause	80	5	20	0	0	0	0	0	0
Post-ménopause	0	0	15	100	+++	0	5	0	0
Post-ménopause	100	80	10	0	0	0	0	0	0
Post-ménopause	30	0	5	0	0	0	0	0	0
Post-ménopause	0	0	15	40	+++	0	0	0	0
Ménopause	0	80	8	0	0	0		0	0
Post-ménopause	0	0	27	0	0	0	30	0	1
Post-ménopause	0	0	56	0	0	80	60	100	0
Activité génitale	50	92	2	1	+	0	0	0	0
Post-ménopause	0	47	5	0	0	0	0	80	0
Post-ménopause	65	0	10	0	0	0	0	60	0
Post-ménopause	100	50	11	0	0	0	0	0	0
Ménopause	20	100	0	0	0	0	0	0	0
Activité génitale	90	6	5	0	0	0	0	0	0
Post-ménopause	100	3	5	0	0	0	0	0	0
Activité génitale	0	0	6	0	0	0	0	0	0
Ménopause	80	100	5	0	0	0	0	0	0
Post-ménopause	100	85	25	0	0	0		0	0
Post-ménopause	10	45	11	13	+++	0	0	0	0
Post-ménopause	66	1	2	40	++	0	0	0	0

Outcome



metastatic_relapse	date_metastatic_relapse
Yes	04/02/1999
No	
No	
No	
Yes	04/09/1990
Yes	08/02/1993
Yes	15/12/1999
No	
No	
Yes	08/03/1995
No	
Yes	06/04/1990
Yes	02/11/1994
No	
No	
No	
No	
No	
No	
No	
No	
No	
No	
No	
No	
No	
Yes	27/10/1999
No	
No	
No	
No	



patient 1 x_1^1, x_2^1, x_3^1 y^1

patient 2 x_1^2, x_2^2, x_3^2 y^2

patient 3 x_1^3, x_2^3, x_3^3 y^3

Model

$a_1 * x_1 + a_2 * x_2 + a_3 * x_3$

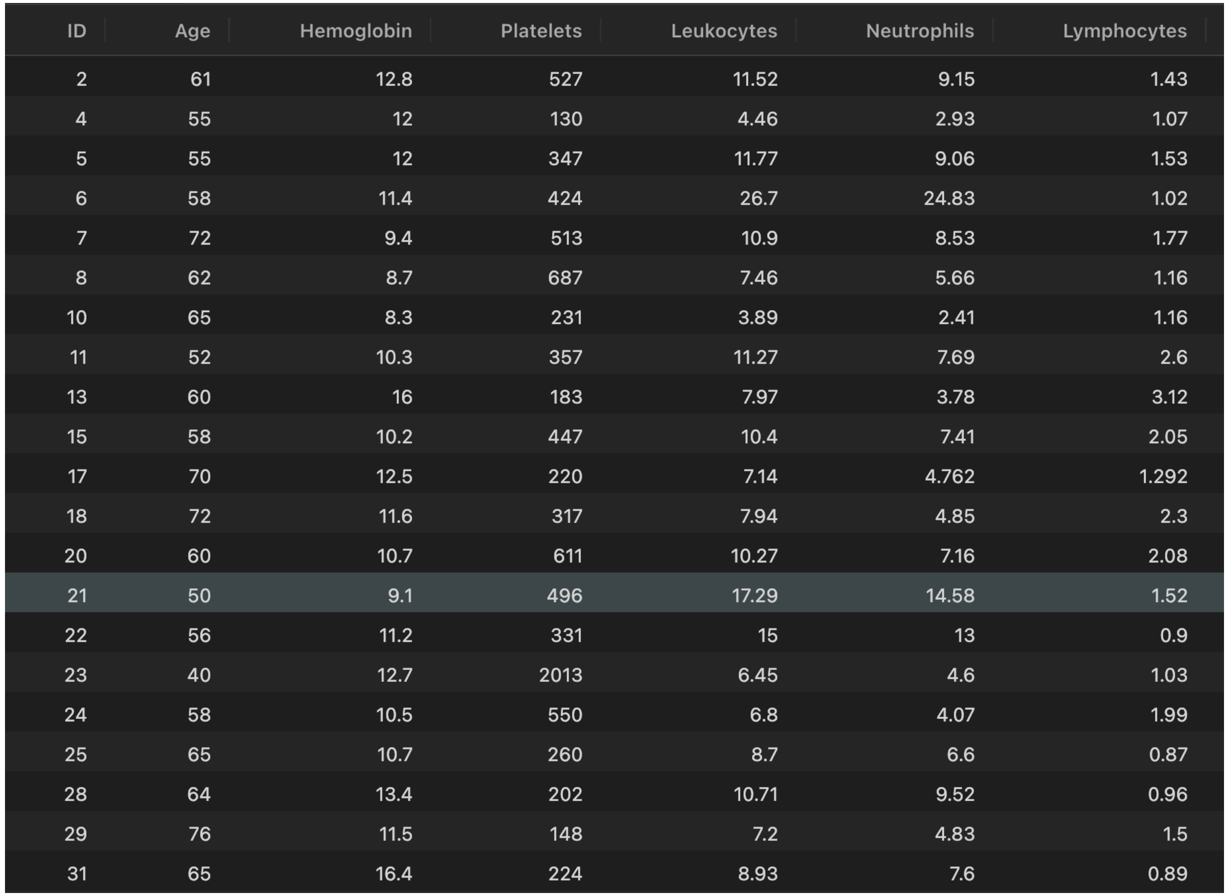
a_1, a_2, a_3

Example: predicting respone to immunotherapy in non-small cell lung cancer

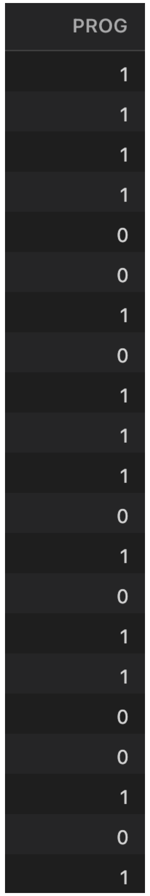
p = 10 features = $(x_1, \ldots, x_{10})$

y = response

n =298 patients



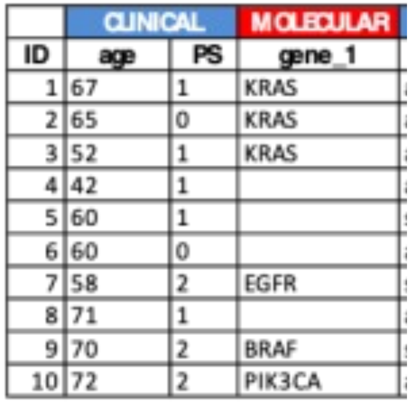
ID	Age	Hemoglobin	Platelets	Leukocytes	Neutrophils	Lymphocytes
2	61	12.8	527	11.52	9.15	1.43
4	55	12	130	4.46	2.93	1.07
5	55	12	347	11.77	9.06	1.53
6	58	11.4	424	26.7	24.83	1.02
7	72	9.4	513	10.9	8.53	1.77
8	62	8.7	687	7.46	5.66	1.16
10	65	8.3	231	3.89	2.41	1.16
11	52	10.3	357	11.27	7.69	2.6
13	60	16	183	7.97	3.78	3.12
15	58	10.2	447	10.4	7.41	2.05
17	70	12.5	220	7.14	4.762	1.292
18	72	11.6	317	7.94	4.85	2.3
20	60	10.7	611	10.27	7.16	2.08
21	50	9.1	496	17.29	14.58	1.52
22	56	11.2	331	15	13	0.9
23	40	12.7	2013	6.45	4.6	1.03
24	58	10.5	550	6.8	4.07	1.99
25	65	10.7	260	8.7	6.6	0.87
28	64	13.4	202	10.71	9.52	0.96
29	76	11.5	148	7.2	4.83	1.5
31	65	16.4	224	8.93	7.6	0.89



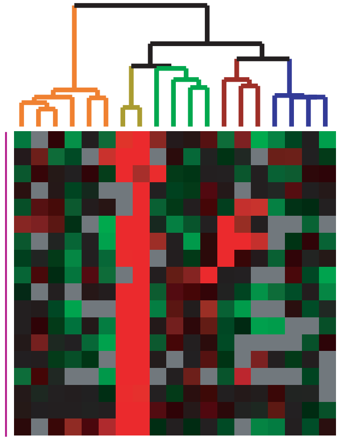
PROG
1
1
1
1
0
0
1
0
1
1
1
0
1
0
1
1
0
0
1
0
1

Types of data

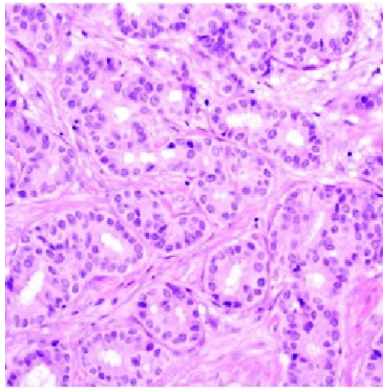
Tabular



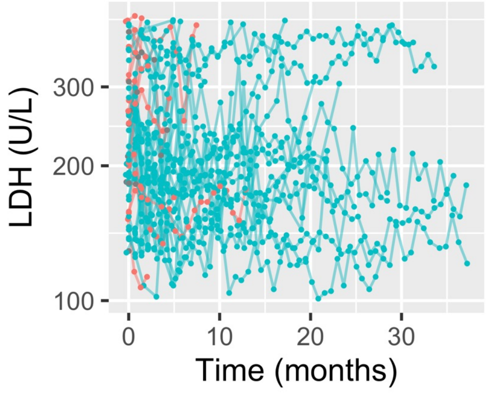
	CLINICAL		MOLECULAR
ID	age	PS	gene_1
1	67	1	KRAS
2	65	0	KRAS
3	52	1	KRAS
4	42	1	
5	60	1	
6	60	0	
7	58	2	EGFR
8	71	1	
9	70	2	BRAF
10	72	2	PIK3CA



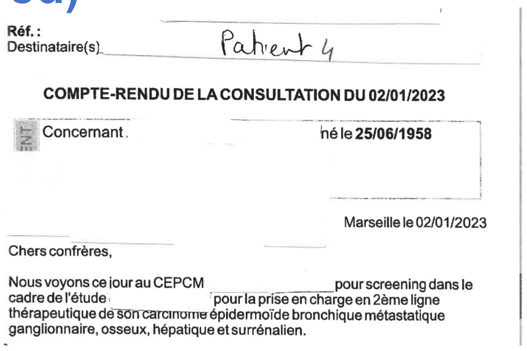
Imaging



Longitudinal



Text (unstructured)



Réf. :
Destinataire(s) Patient 4

COMPTE-RENDU DE LA CONSULTATION DU 02/01/2023

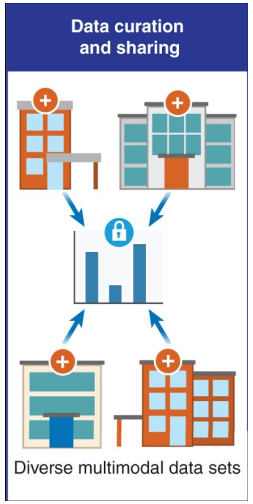
Concernant né le 25/06/1958

Marseille le 02/01/2023

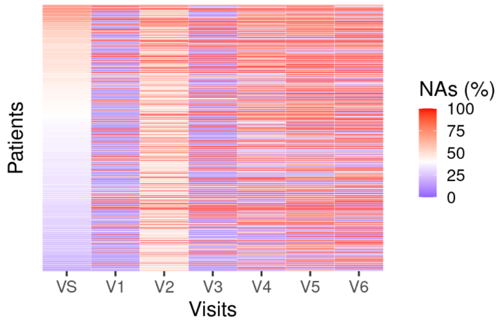
Chers confrères,

Nous voyons ce jour au CEPCM pour screening dans le cadre de l'étude pour la prise en charge en 2ème ligne thérapeutique de son carcinome épidermoïde bronchique métastatique ganglionnaire, osseux, hépatique et surrénalien.

Preprocessing



- Load data and possibly merge different sources / types
- Document the data : dictionary + types (categorical / numeric)
- Clean the data (outliers? aberrant values? units errors? exclusion criteria?)
- Define features of interest (e.g., BMI) and feature sets (e.g., monotherapy patients)
- Dummify categorical variables, transform numerics (e.g., log)
- **Missing values** (not covered in this course but ++)
- Scaling



⇒ **First, look at the data and perform exploratory data analysis**

Garbage in = garbage out

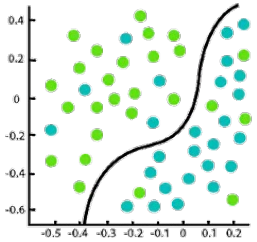
Formalism

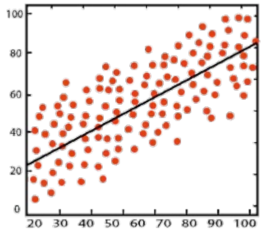
Machine (Statistical) (supervised) Learning

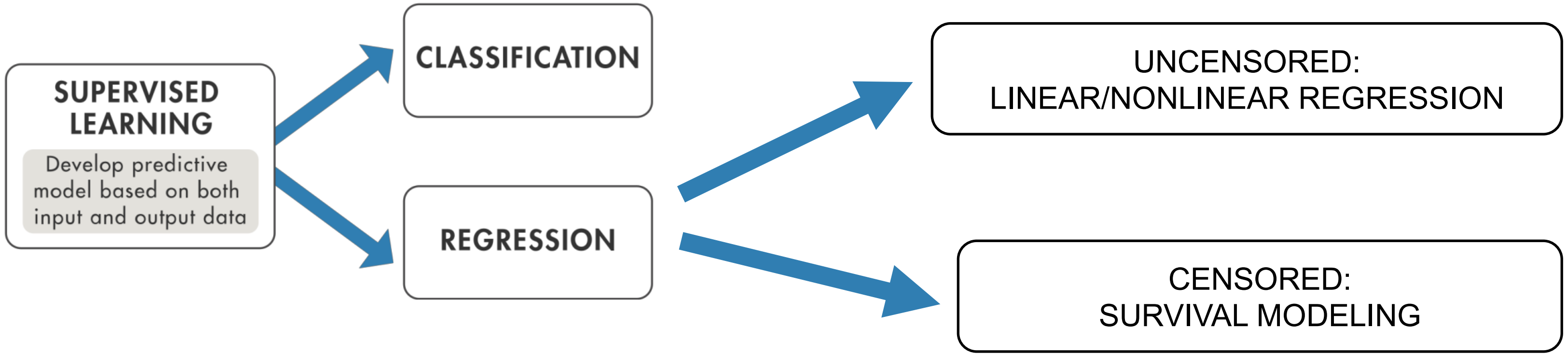
$$y = f(x) + \varepsilon$$

ε = irreducible error

- $x = x_1, x_2, \dots, x_p$ set of variables / features / predictors (e.g., biomarkers)
- Goal = predict y from x = **learn** $\hat{f}$ that is “close” to f → prediction $\hat{y} = \hat{f}(x)$
- $y \in \{Y_1, Y_2\}$ qualitative/categorical ⇒ classification
- $y \in \mathbb{R}$ quantitative/continuous ⇒ regression







SUPERVISED LEARNING

Develop predictive model based on both input and output data

CLASSIFICATION

REGRESSION

UNCENSORED:
LINEAR/NONLINEAR REGRESSION

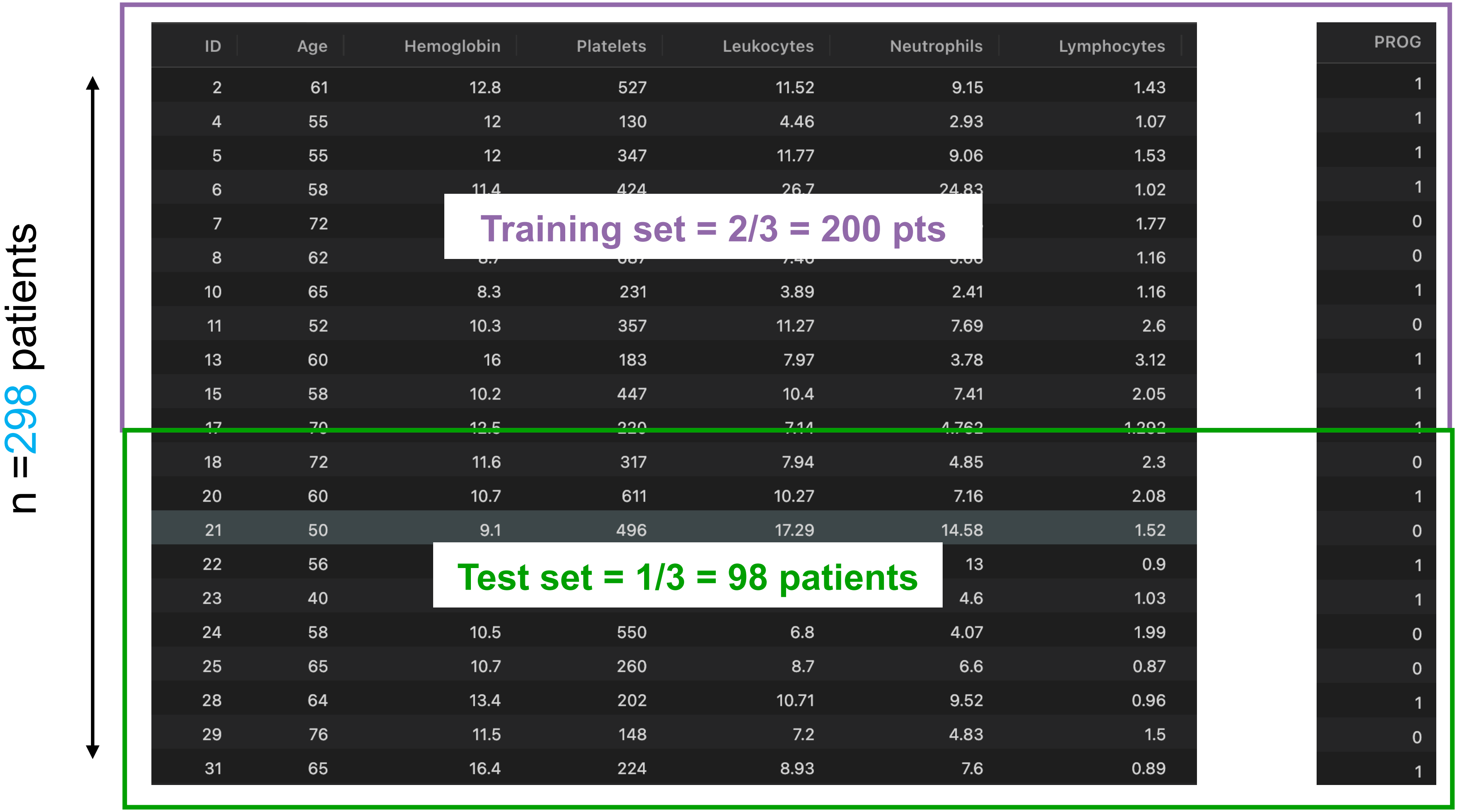
CENSORED:
SURVIVAL MODELING

Training / test split

- How to evaluate the predictive performance of $\hat{f}$?
- It is trivial to find a model that perfectly predicts the data it has seen (the training data)
- We want to test the performances of $\hat{f}$ on *unseen* data
- Best solution: have an external validation set (e.g., from a different study / hospital)
- If not: randomly split the data between a training (usually 2/3 or 3/4) and a test set
- Warning! from the moment you see the test data and the model performances, if you further change anything, you cheat! (there is leakage)

😱

Training / test split



n =298 patients

Training set = 2/3 = 200 pts

Test set = 1/3 = 98 patients

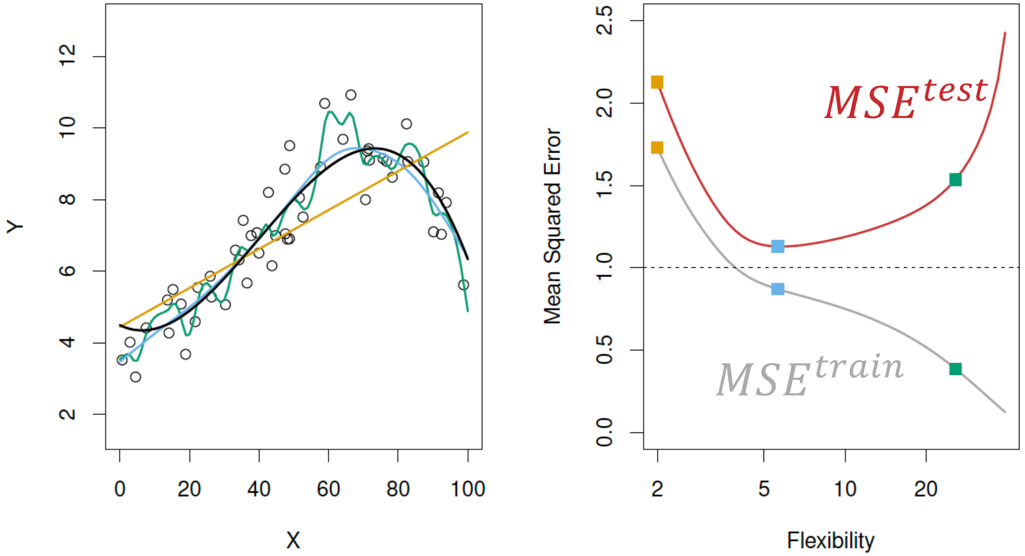
ID	Age	Hemoglobin	Platelets	Leukocytes	Neutrophils	Lymphocytes	PROG
2	61	12.8	527	11.52	9.15	1.43	1
4	55	12	130	4.46	2.93	1.07	1
5	55	12	347	11.77	9.06	1.53	1
6	58	11.4	424	26.7	24.83	1.02	1
7	72	[illegible]	[illegible]	[illegible]	[illegible]	1.77	0
8	62	[illegible]	[illegible]	[illegible]	[illegible]	1.16	0
10	65	8.3	231	3.89	2.41	1.16	1
11	52	10.3	357	11.27	7.69	2.6	0
13	60	16	183	7.97	3.78	3.12	1
15	58	10.2	447	10.4	7.41	2.05	1
17	70	12.5	220	7.14	4.762	1.292	1
18	72	11.6	317	7.94	4.85	2.3	0
20	60	10.7	611	10.27	7.16	2.08	1
21	50	9.1	496	17.29	14.58	1.52	0
22	56	[illegible]	[illegible]	[illegible]	13	0.9	1
23	40	[illegible]	[illegible]	[illegible]	4.6	1.03	1
24	58	10.5	550	6.8	4.07	1.99	0
25	65	10.7	260	8.7	6.6	0.87	0
28	64	13.4	202	10.71	9.52	0.96	1
29	76	11.5	148	7.2	4.83	1.5	0
31	65	16.4	224	8.93	7.6	0.89	1

Evaluating performances: regression

- Let $x^t = x^{t_1}, \ldots, x^{t_T}$ the test set variables and $y^t = y^{t_1}, \ldots, y^{t_T}$ the associated test outcomes

Mean squared error = $MSE^{train} = Ave\left(y - \hat{f}(x)\right)^2, \quad MSE^{test} = Ave\left(y^t - \hat{f}(x^t)\right)^2$

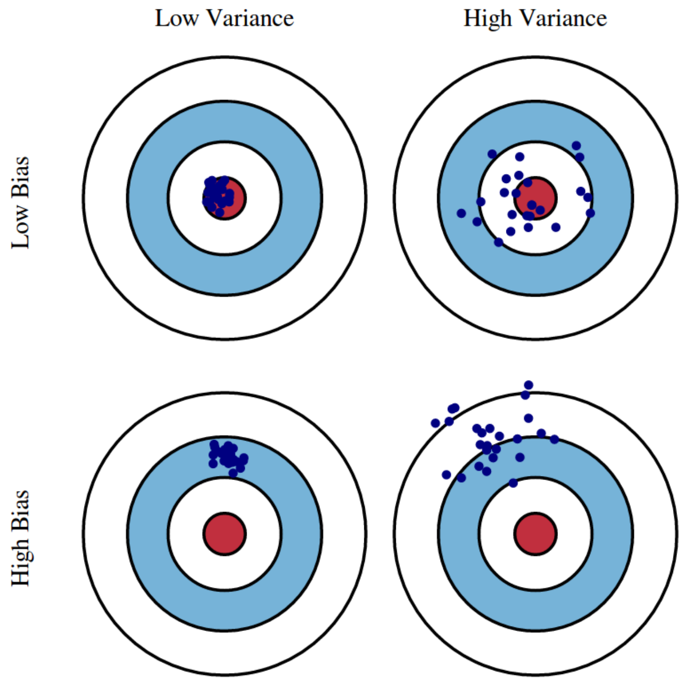
Should we minimize the MSE^{train} ?



Bias and variance

- Bias = how accurate is the prediction, *in average*

$$E\left[f(x) - \hat{f}(x)\right]$$



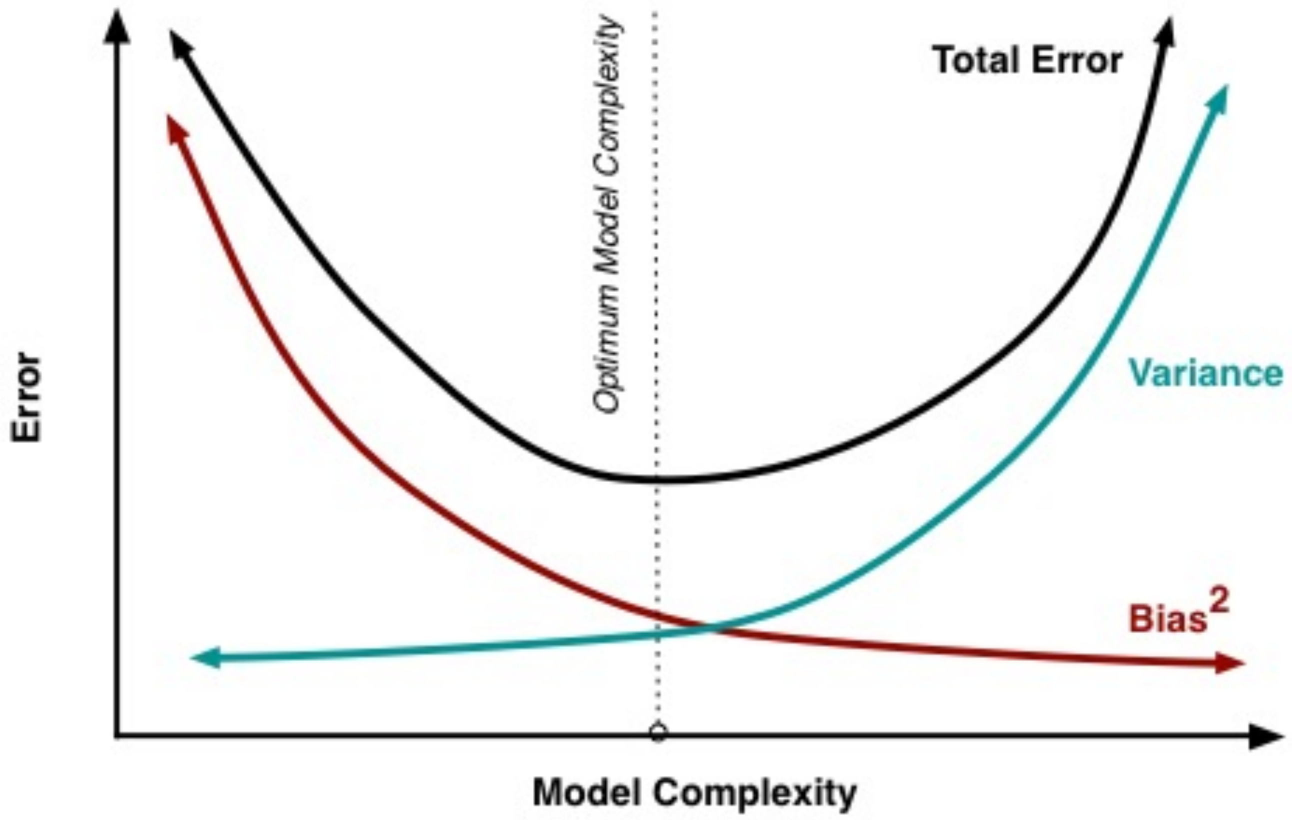
- Variance = how variable is the prediction, *in average*

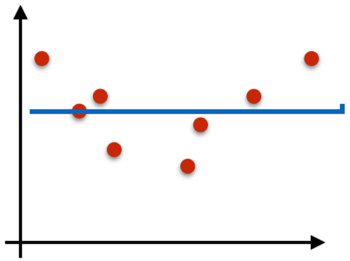
$$E\left[\left(\hat{f}(x) - E\left[\hat{f}(x)\right]\right)^2\right]$$

where the average is to be understood as if we repeatedly estimated f using a large number of training sets

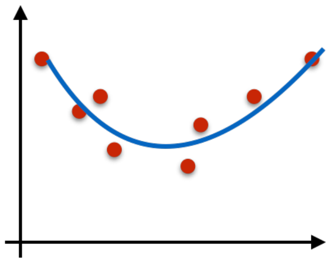
Bias versus variance trade-off

Theorem: $$E\left[\left(y^t - \hat{f}(x^t)\right)^2\right] = Var\left(\hat{f}(x^t)\right) + Bias\left(\hat{f}(x^t)\right)^2 + Var(\varepsilon)$$

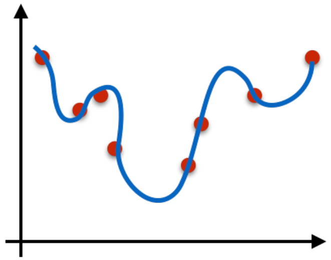




Underfitting



Correct fitting



Overfitting

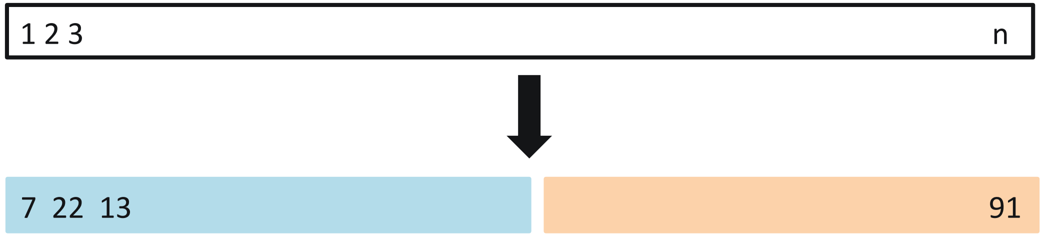
Resampling methods

Resampling method = drawing samples from a training set
and refitting a model of interest

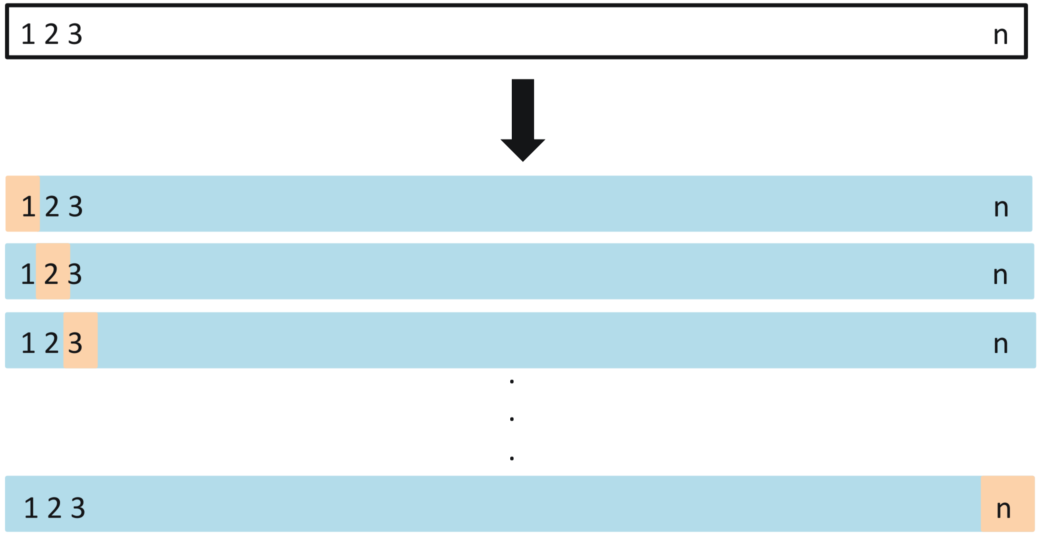
- No external test set available
- Gives information about the variability and sensitivity of the model (model assessment)
- Select a model among candidates (model selection)
- Tune the hyperparameters (e.g., tree depth or minimal number of samples in each leaf)
- Two main resampling methods: cross-validation and bootstrap

Cross validation

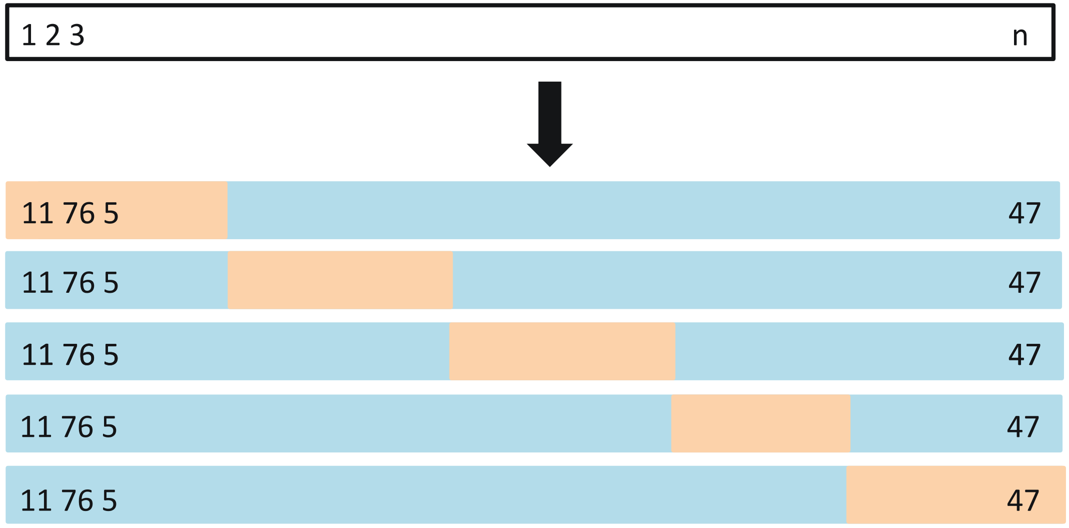
Train/test



Leave-one-out cross-validation (LOOCV)



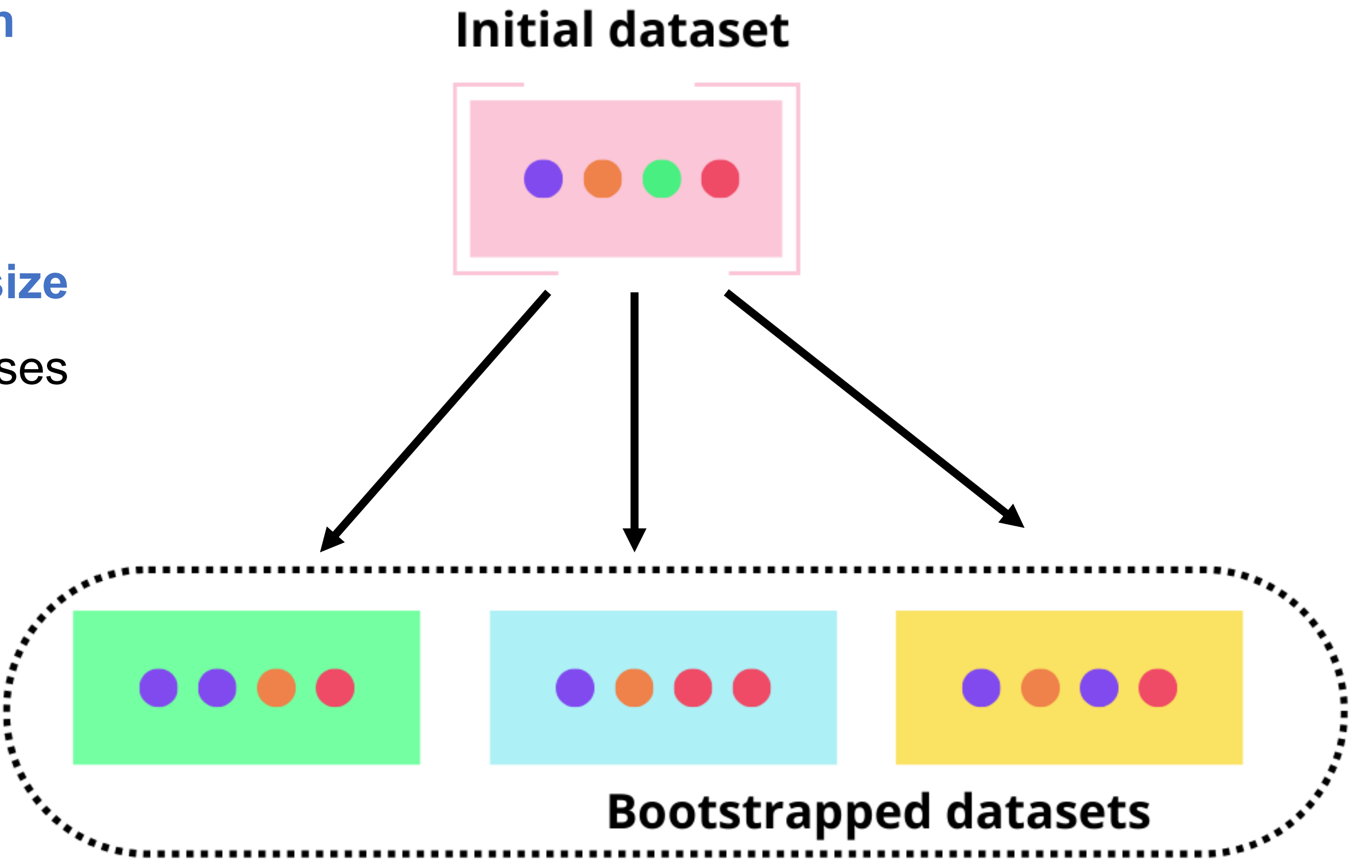
k-fold cross-validation (k = 5)



Bootstrap

- Randomly select *n* times a subject, **with replacement**

- A bootstrapped dataset has the **same size** but contains only 63.2% of the initial cases



Even less data (because of splitting)

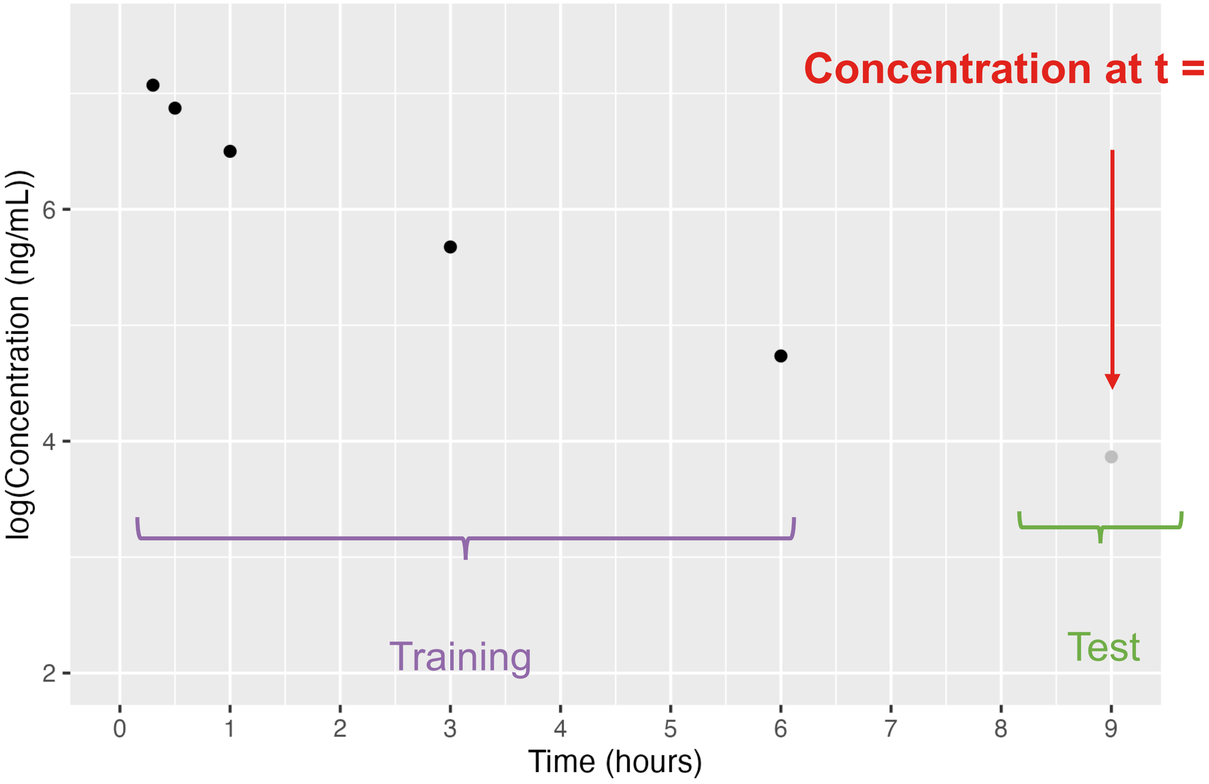
t (50)

Effecti

t (50)

Linear regression

Example: concentration of a drug (sunitinib in rats) over time



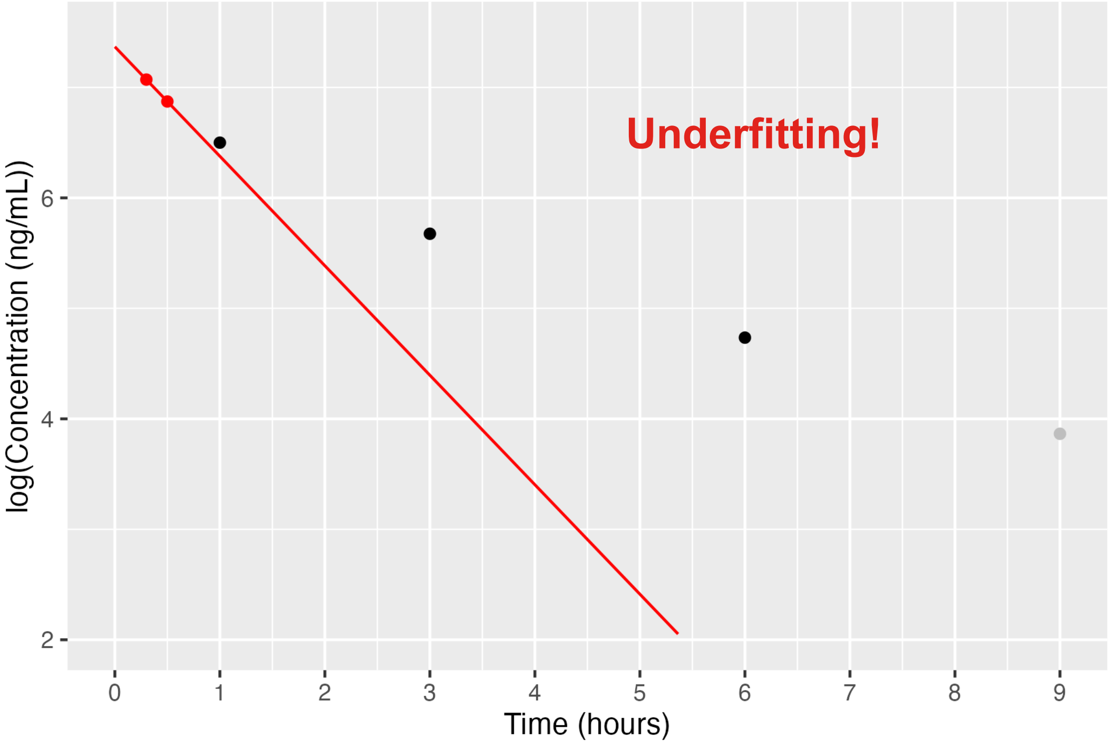
y = log(concentration), x = time

y = f(x) ?

Example: concentration of a drug (sunitinib in rats) over time

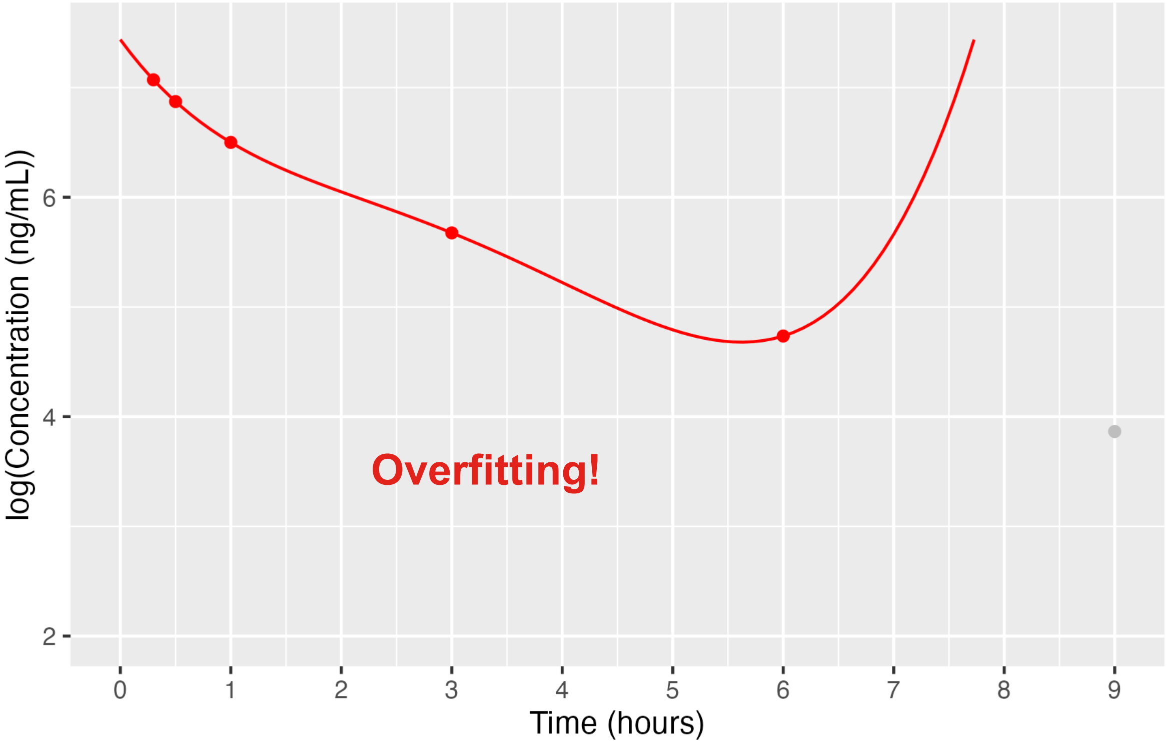
Linear

$$y = \theta_0 + \theta_1 x$$


Underfitting!

Polynomial

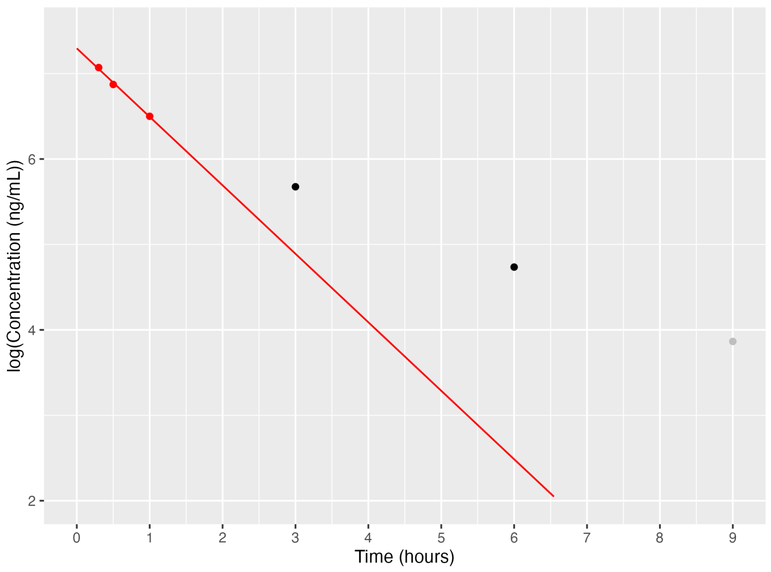
$$y = \theta_0 + \theta_1 x^2 + \theta_2 x^3 + \theta_3\ x^4 + \theta_4 x^5$$


Overfitting!

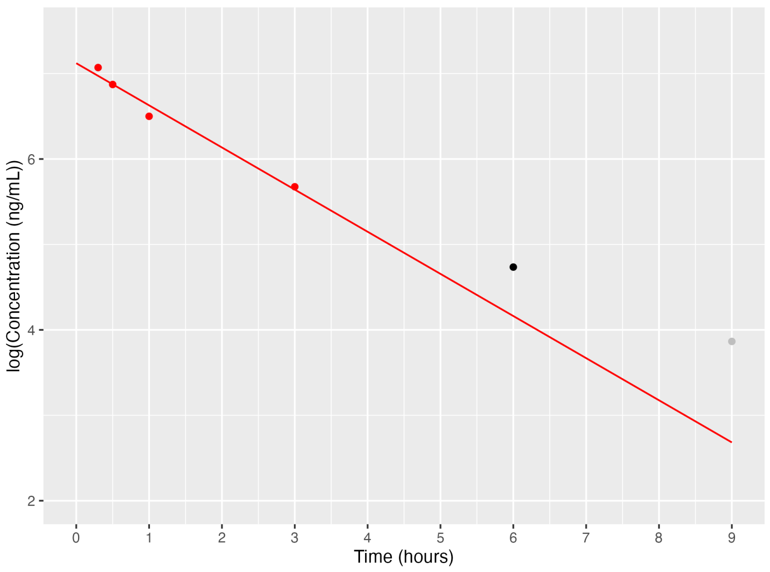
Linear regression

$$y = \theta_0 + \theta_1 x + \varepsilon$$

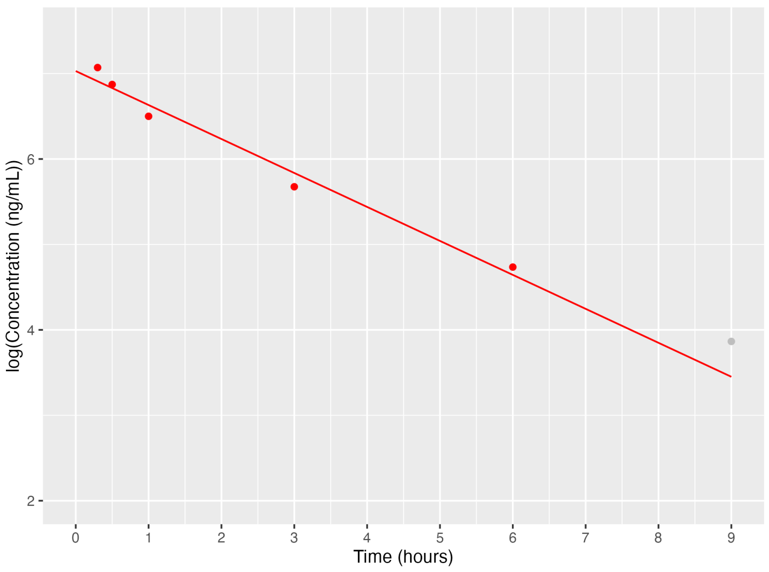
3 data points



4 data points



5 data points



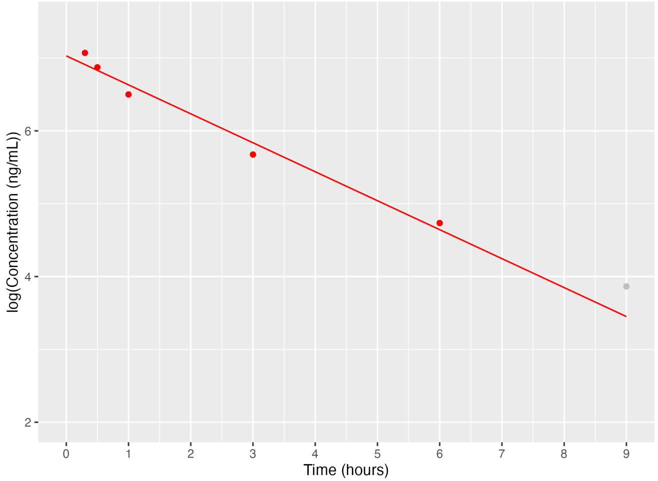
Linear regression: under the hood

$$y = \beta_0 + \beta_1 x + \varepsilon$$

How to find $\widehat{\beta_0} \approx \beta_0$ and $\widehat{\beta_1} \approx \beta_1$?

- $\hat{\beta} = (\widehat{\beta_0}, \widehat{\beta_1})$ is the value that **minimizes** the **sum of squared residuals**

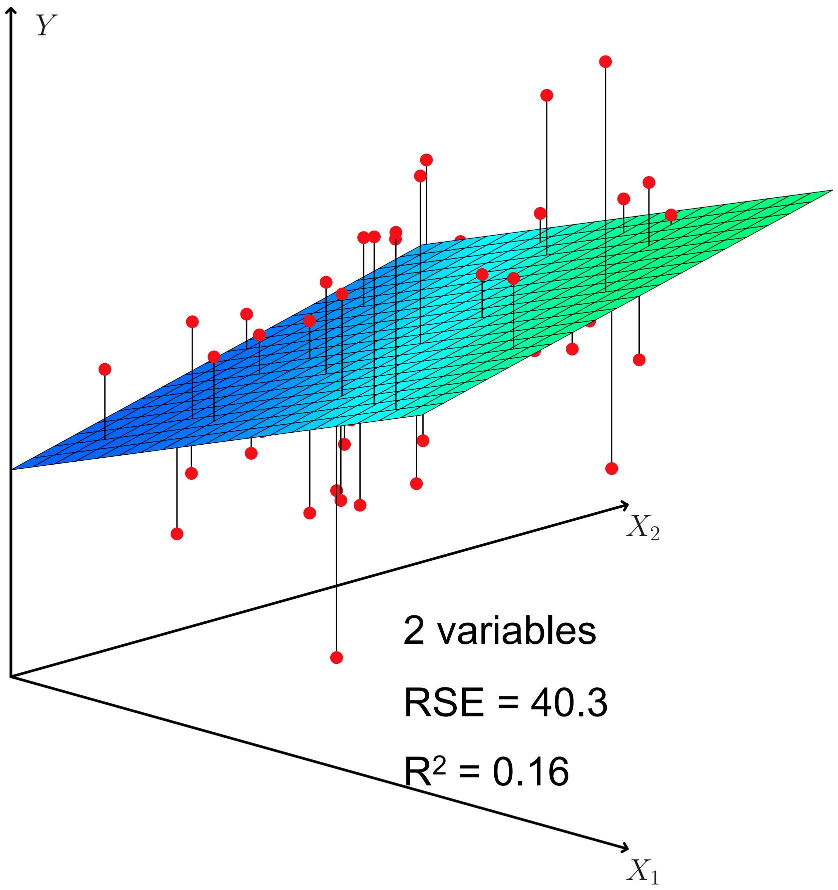
$$SS = \sum_{i=1}^{n} \left(y_i - (\beta_0 + \beta_1 t_i)\right)^2$$



ML training ⇔ **Optimization** of an objective function (also called "loss")

Multiple linear regression

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \varepsilon$$

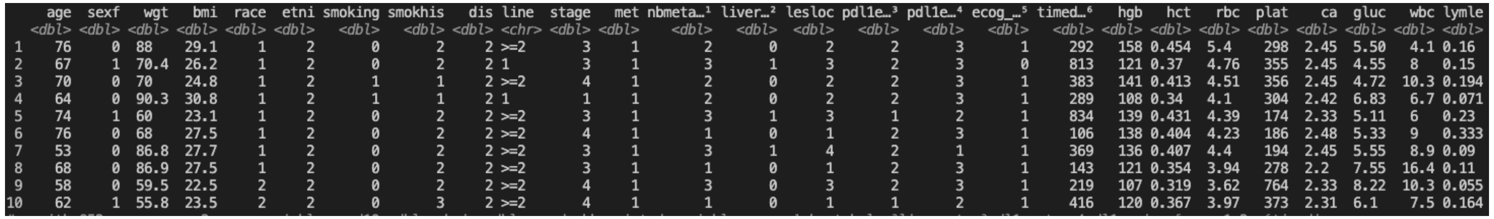


2 variables

RSE = 40.3

R^2 = 0.16

Predict tumor size (SLD) from 59 variables



```
     age  sexf  wgt    bmi  race  etni smoking smokhis   dis line  stage   met nbmeta…¹ liver…² lesloc pdl1e…³ pdl1e…⁴ ecog_…⁵ timed…⁶   hgb   hct   rbc  plat    ca  gluc   wbc lymle
   <dbl> <dbl> <dbl> <dbl> <dbl> <dbl>   <dbl>   <dbl> <dbl> <chr> <dbl> <dbl>    <dbl>   <dbl>  <dbl>   <dbl>   <dbl>   <dbl>   <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl>
 1    76     0 88     29.1     1     2       0       2     2 >=2       3     1        2       0      2       2       3       1     292   158 0.454 5.4     298  2.45 5.50    4.1 0.16
 2    67     1 70.4   26.2     1     2       0       2     2 1         3     1        3       1      3       2       3       0     813   121 0.37  4.76    355  2.45 4.55  8     0.15
 3    70     0 70     24.8     1     2       1       1     2 >=2       4     1        2       0      2       2       3       1     383   141 0.413 4.51    356  2.45 4.72   10.3 0.194
 4    64     0 90.3   30.8     1     2       1       1     2 1         1     1        2       0      2       2       3       1     289   108 0.34  4.1     304  2.42 6.83    6.7 0.071
 5    74     1 60     23.1     1     2       0       2     2 >=2       3     1        3       1      3       1       2       1     834   139 0.431 4.39    174  2.33 5.11  6     0.23
 6    76     0 68     27.5     1     2       0       2     2 >=2       4     1        1       0      1       2       3       1     106   138 0.404 4.23    186  2.48 5.33  9     0.333
 7    53     0 86.8   27.7     1     2       0       2     2 >=2       3     1        3       1      4       2       1       1     369   136 0.407 4.4     194  2.45 5.55    8.9 0.09
 8    68     0 86.9   27.5     1     2       0       2     2 >=2       3     1        1       0      1       2       3       1     143   121 0.354 3.94    278  2.2  7.55   16.4 0.11
 9    58     0 59.5   22.5     2     2       0       2     2 >=2       4     1        3       0      3       2       3       1     219   107 0.319 3.62    764  2.33 8.22   10.3 0.055
10    62     1 55.8   23.5     2     2       0       3     2 >=2       4     1        1       0      1       1       2       1     416   120 0.367 3.97    373  2.31 6.1     7.5 0.164
```

…

$$y = \beta_0 + \beta_1 x_1 + \cdots + \beta_{59} x_{59} + \varepsilon$$

RSE = 36.0

R^2 = 0.44

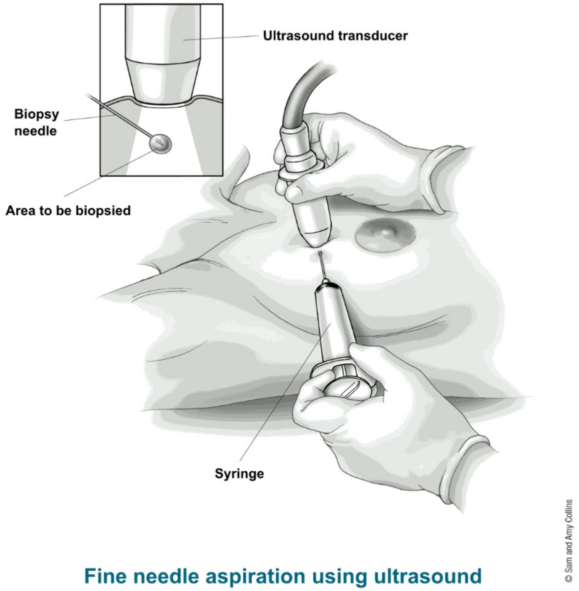
Categorical variables? → dummified (= one-hot-encoding)

- SEX = M, F → SEX = {0, 1}
- NB_META = {0, 1, 2, ≥ 3} → NB_META_1, NB_META_2 and NB_META_≥3

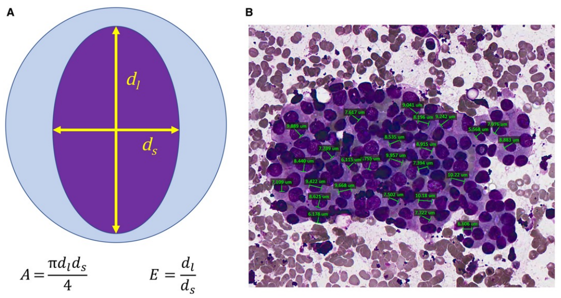
⚠️ Variables need to be **scaled**

Linear classification: logistic regression

Example: breast cancer diagnosis



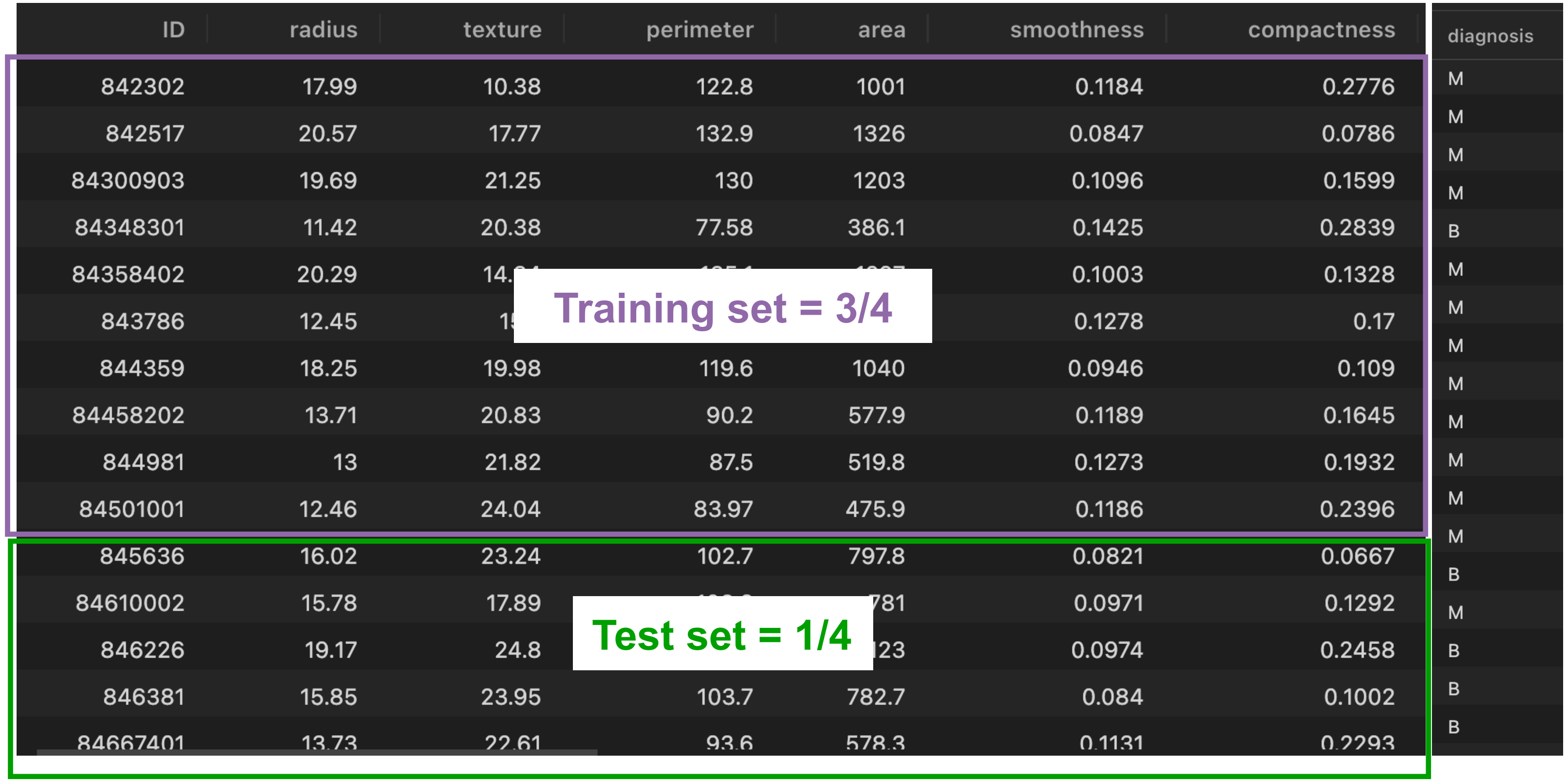
Fine needle aspiration using ultrasound



n =569 subjects

p = 32 features = $(x_1, \ldots, x_{32})$

y



Training set = 3/4

Test set = 1/4

ID	radius	texture	perimeter	area	smoothness	compactness
842302	17.99	10.38	122.8	1001	0.1184	0.2776
842517	20.57	17.77	132.9	1326	0.0847	0.0786
84300903	19.69	21.25	130	1203	0.1096	0.1599
84348301	11.42	20.38	77.58	386.1	0.1425	0.2839
84358402	20.29	14.[illegible]	[illegible]	[illegible]	0.1003	0.1328
843786	12.45	15[illegible]	[illegible]	[illegible]	0.1278	0.17
844359	18.25	19.98	119.6	1040	0.0946	0.109
84458202	13.71	20.83	90.2	577.9	0.1189	0.1645
844981	13	21.82	87.5	519.8	0.1273	0.1932
84501001	12.46	24.04	83.97	475.9	0.1186	0.2396
845636	16.02	23.24	102.7	797.8	0.0821	0.0667
84610002	15.78	17.89	[illegible]	781	0.0971	0.1292
846226	19.17	24.8	[illegible]	[illegible]123	0.0974	0.2458
846381	15.85	23.95	103.7	782.7	0.084	0.1002
84667401	13.73	22.61	93.6	578.3	0.1131	0.2293

diagnosis
M
M
M
M
B
M
M
M
M
M
M
M
M
B
M
B
B
B

Logistic regression

$$p = \mathbb{P}(Y=1) \in (0,1) \xrightarrow{?} \mathbb{R}$$

$$\frac{p}{1-p} \in (0,+\infty) = \frac{\mathbb{P}(Y=1)}{\mathbb{P}(Y=0)} = \text{odds} \simeq \text{chance}$$

$$\ln\left(\frac{p}{1-p}\right) = \beta_0 + \beta_1 x_1 + \cdots + \beta_L x_L$$

Why not linear regression?

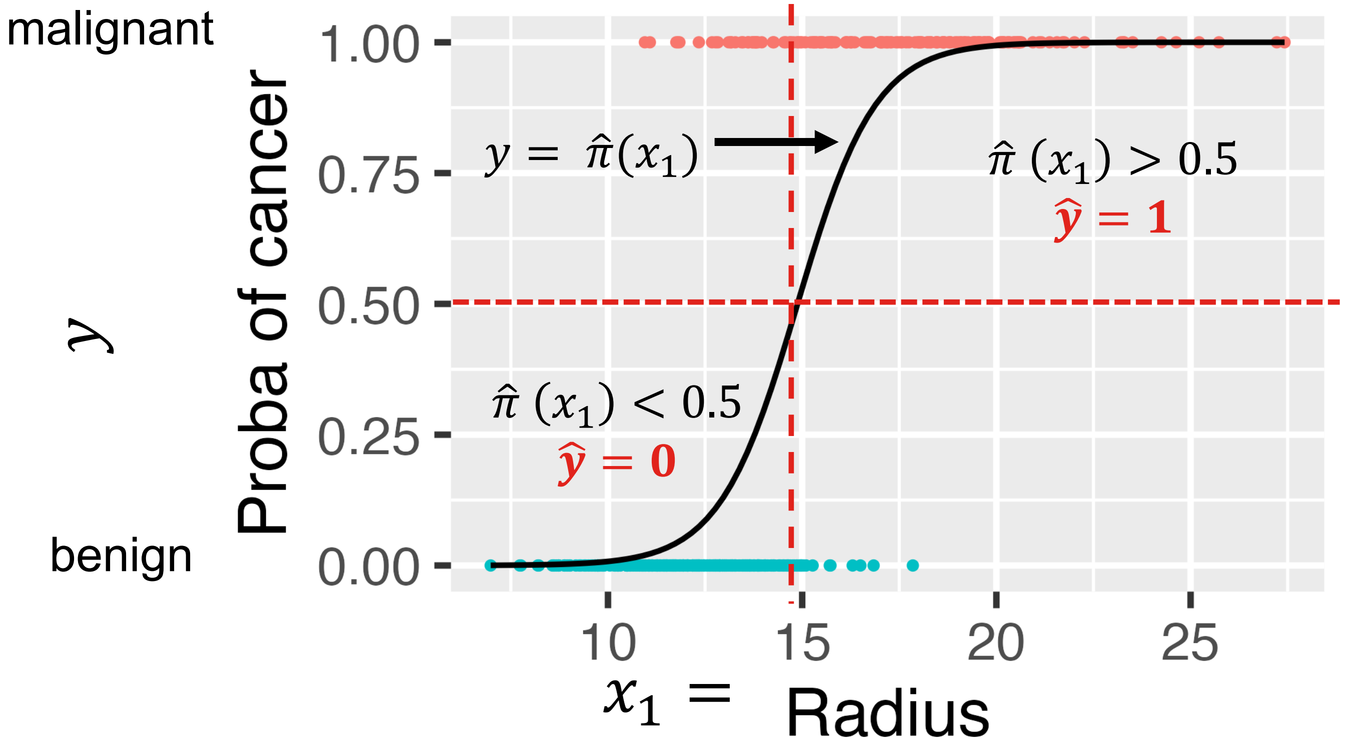
$$\Leftrightarrow p = \frac{e^{\beta_0+\beta_1 x_1+\cdots+\beta_L x_L}}{1+e^{\beta_0+\beta_1 x_1+\cdots+\beta_L x_L}} = \pi(x)$$

Estimation: likelihood maximization $\longrightarrow (\widehat{\beta_k})$

Interpretation: for one variable x, $\text{odds}(x) = e^{\widehat{\beta_0}+\widehat{\beta_1}x}$

$$\Rightarrow e^{\widehat{\beta}} = \frac{\text{odds}(x+1)}{\text{odds}(x)} = \text{odds ratio} = OR$$

$$\hat\pi(x_1) = 0.5$$
$$\Leftrightarrow x_1 = -\frac{\widehat{\beta_0}}{\widehat{\beta_1}}$$



Training set

if $OR = 1.5$ there is a 50% increase of chance of having $Y = 1$ for an increase of x of one unit

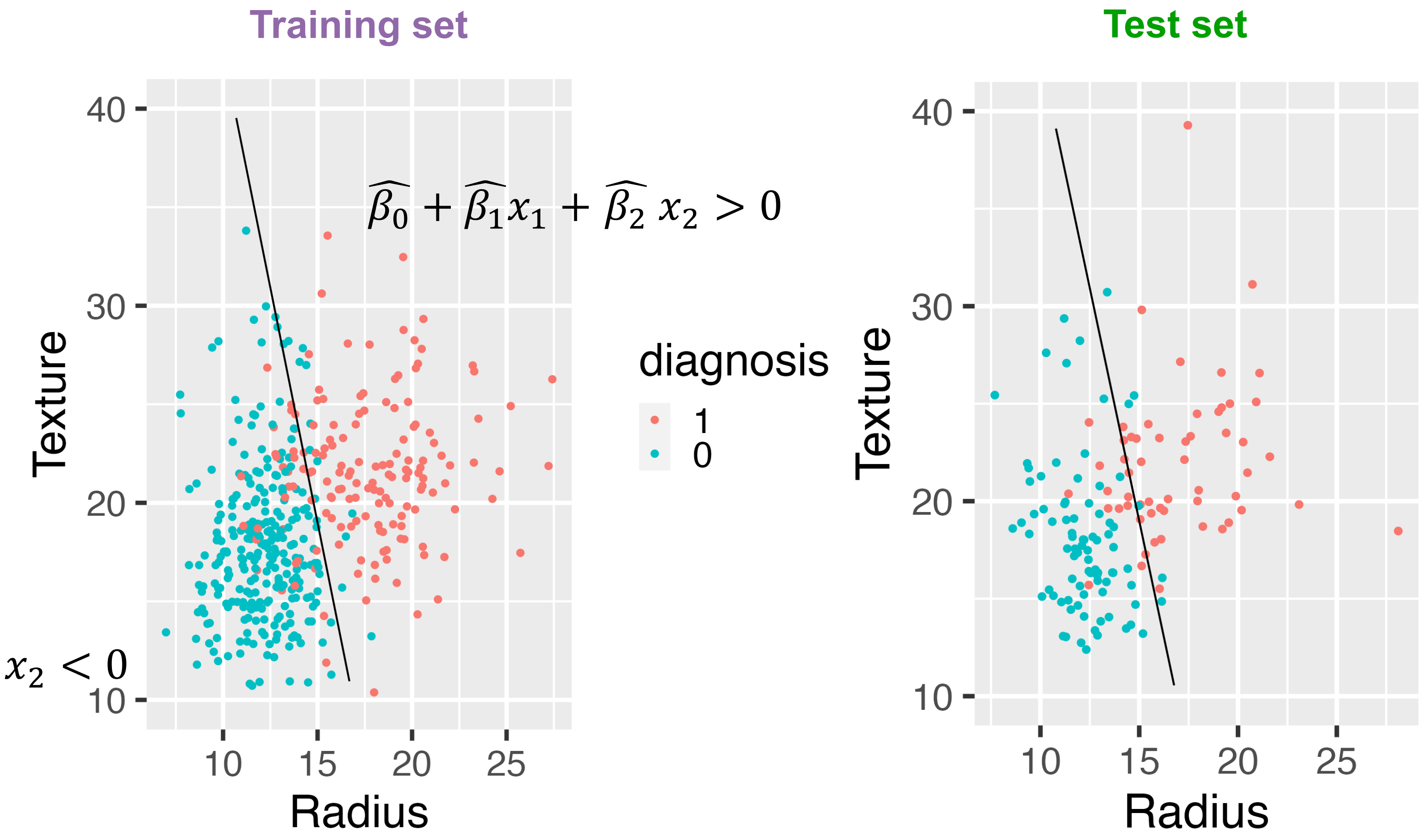
Logistic regression = linear classification

- 2 features: radius and texture

$$\ln\left(\frac{p}{1-p}\right) = \beta_0 + \beta_1 x_1 + \beta_2 x_2$$

Fit on training set
(Likelihood maximization)

$$\widehat{\beta_0}, \widehat{\beta_1}, \widehat{\beta_2}$$



Training set

Test set

$$\widehat{\beta_0} + \widehat{\beta_1} x_1 + \widehat{\beta_2}\, x_2 > 0$$

$$\widehat{\beta_0} + \widehat{\beta_1} x_1 + \widehat{\beta_2}\, x_2 < 0$$

Classification: additional prediction metrics

Performance evaluation: Confusion matrix

$$\text{Data}\begin{pmatrix} x^1 \\ \vdots \\ x^N \end{pmatrix} \longrightarrow \text{Predictions}\begin{pmatrix} \hat{y}^1 \\ \vdots \\ \hat{y}^N \end{pmatrix} = \begin{pmatrix} \widehat{M}(x^1) \\ \vdots \\ \widehat{M}\,(x^N) \end{pmatrix} \quad \text{vs reality}\begin{pmatrix} y^1 \\ \vdots \\ y^N \end{pmatrix}$$

		Actual	
		1	0
Model	+	TP (Sensitivity)	FP
	-	FN	TN (Specificity)

Accuracy = $\frac{TP+TN}{TP+TN+FP+TN}$

Sensitivity = $SE = \mathbb{P}\,(+|1) = TPR = \frac{TP}{TP+FN}$

$\beta = \mathbb{P}(-|1) = FNR = 1 - SE =$ proba of type II error

(classify as benign what is cancer)

Specificity = $SP = \mathbb{P}\,(-|0) = TNR = \frac{TN}{FP+TN}$

$\alpha = \mathbb{P}(+|0) = FPR = 1 - SP =$ proba of type I error

(classify as tumor what is benign)

Performances

Radius | **Radius + texture** | **All**

Training set

	1	0
+	122	15
-	33	256

Accuracy = 0.887

	1	0
+	124	13
-	31	258

Accuracy = 0.897

	1	0
+	155	0
-	0	271

Accuracy = 1

Test set

	1	0
+	42	4
-	15	82

Accuracy = 0.867

	1	0
+	44	6
-	13	80

Accuracy = 0.867

	1	0
+	52	7
-	5	79

Accuracy = 0.916

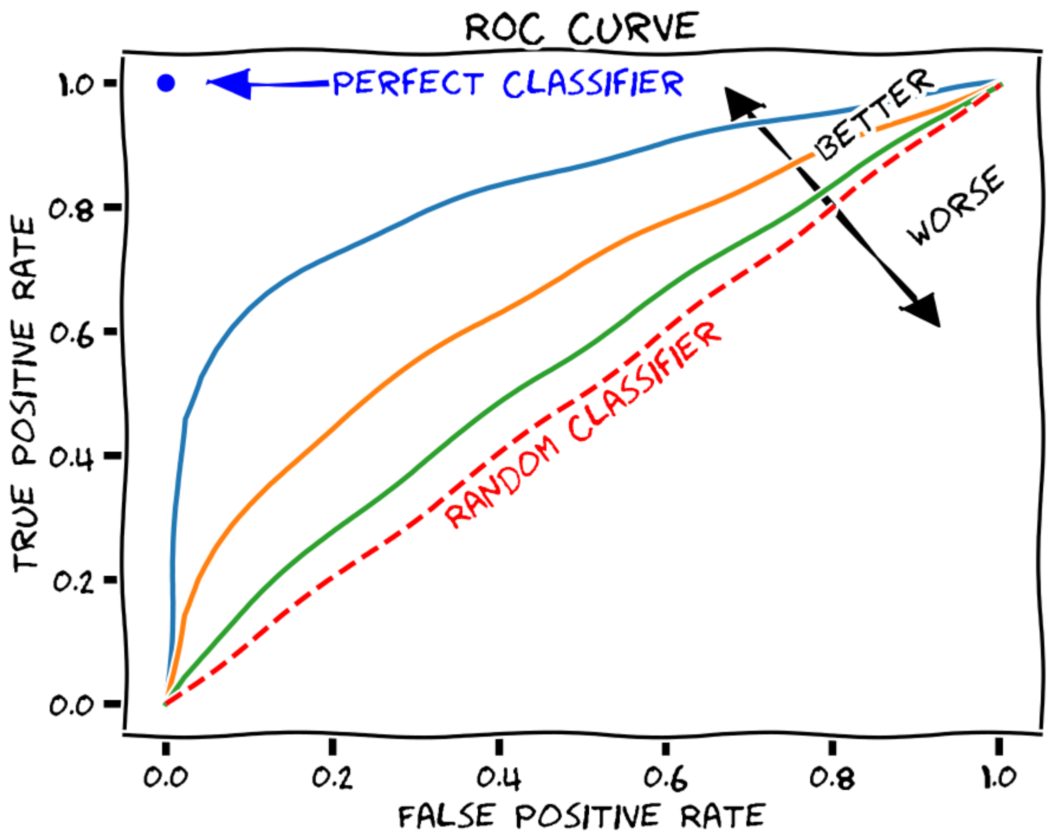
ROC curve analysis

- In practical cases a classification model often assigns a score (e.g. proba)

- For each value of a threshold, one SE and one SP value

- Global quantification of performances = area under the curve (AUC)

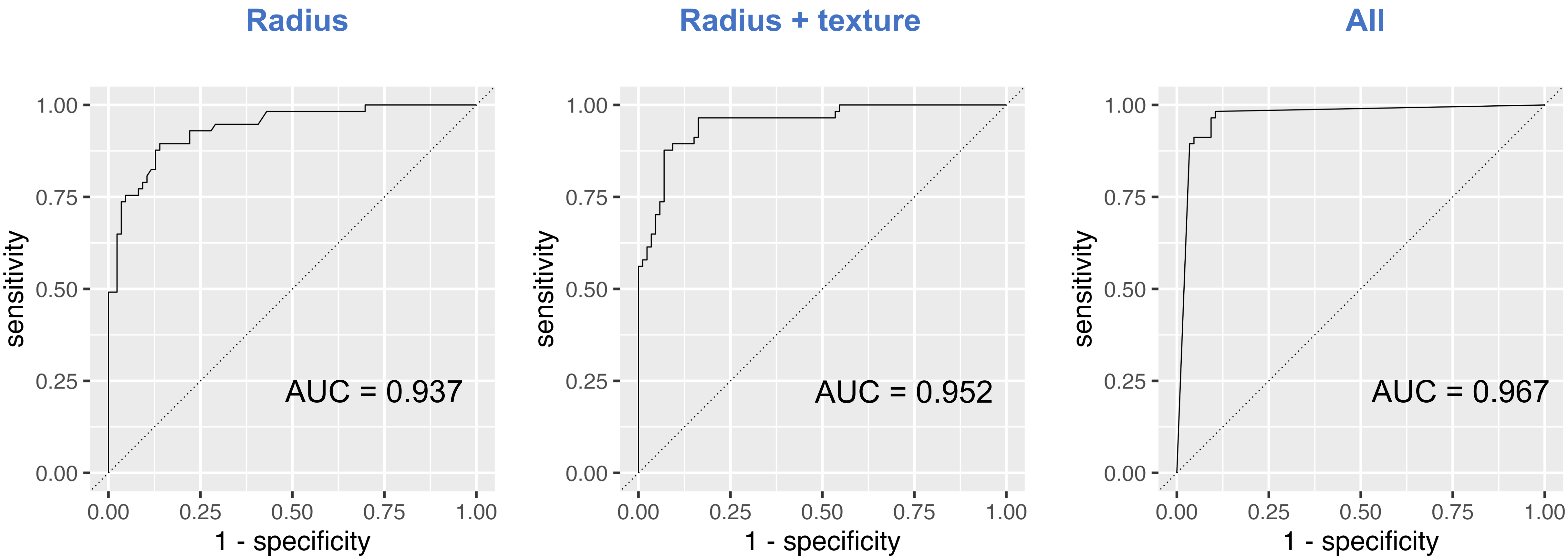
- In practice, one threshold needs to be defined **from the train set**



Sensitivity

1 - Specificity

AUCs of logistic regression (test set)



Radius

Radius + texture

All

Interpretation of AUC

AUC = probability that a random pair of predictions $(\hat{y}^1, \hat{y}^2)$ is concordant with the observations i.e that the score of $\hat{y}^1$ is larger than the score of $\hat{y}^2$ if $y^1 > y^2$.

- S_1 = score in class we want to classify as positive (say, malignant), density f_1
- S_0 = score in other class (say, healthy/benign), density f_0
- T = threshold

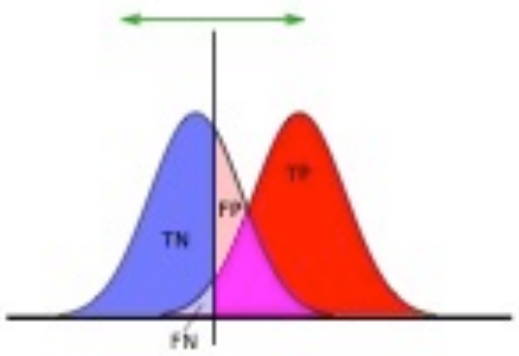
$$AUC = \int_{T_{max}}^{T_{min}} SE(T)d(FPR(T))$$

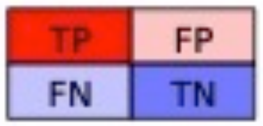
$$SE(T) = \mathbb{P}(S \geq T|1) = \int_{T}^{T_{max}} f_1(x)dx$$

$$FPR(T) = \mathbb{P}(S \geq T|0) = \int_{T}^{T_{max}} f_0(x)dx$$

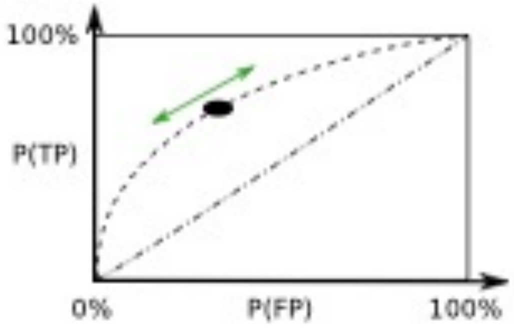
$$AUC = \int_{T_{min}}^{T_{max}} \int_{T}^{T_{max}} f_1(x)f_0(T)dT$$

$$= \mathbb{P}(S_1 \geq S_0)$$





TP	FP
FN	TN



Positive and negative predictive value

- Accuracy, sensitivity and specificity are not sufficient to assess a model

- We are often more interested in $\mathbb{P}(1|+)$ (= positive predictive value, PPV) and $\mathbb{P}(0|-)$ (= negative predictive value, NPV)

- From Bayes

p prevalence

$$PPV = \mathbb{P}(1|+) = \frac{\mathbb{P}(+|1)\mathbb{P}(1)}{\mathbb{P}(+)}$$

$$\begin{aligned}\mathbb{P}(+) &= \mathbb{P}(+|0)\mathbb{P}(0) + \mathbb{P}(+|1)\mathbb{P}(1) = \big(1-\mathbb{P}(-|0)\big)\big(1-\mathbb{P}(1)\big) + SE\cdot\mathbb{P}(1)\\ &= (1-SP)\cdot(1-p) + SE\cdot p\end{aligned}$$

$$PPV = \frac{SE\cdot p}{(1-SP)\cdot(1-p) + SE\cdot p}$$

- Other metrics: $F1$ = harmonic mean of PPV (precision) and sensitivity (recall) = $2(PPV^{-1} + SE^{-1})^{-1}$

Example: Lung cancer and smoking status

- Percentage of smokers among lung cancer patients = 90%, i.e. SE of a model based on smoking status is 0.9
- Approx. 30% of population is composed of smokers ⇒ $SP(= TNR$, i.e. proportion of people who don’t smoke and don’t have cancer) is 70%.
- Assuming a lifetime risk of having lung cancer of 7.19% (= prevalence)

$PPV = \mathbb{P}$(lung cancer during lifetime |smoker)= 18.9%

Nonlinear methods: decision trees

Classification and regression trees (CART)

🔴 = Progression

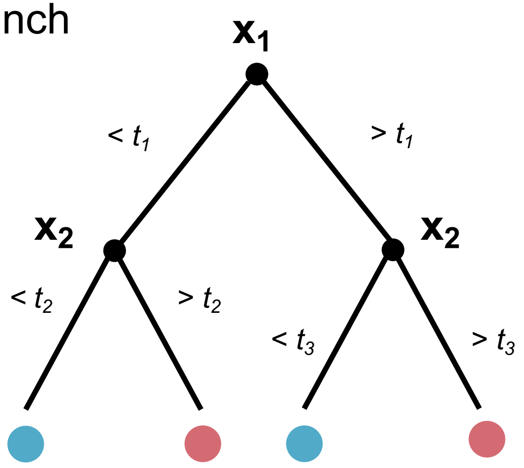
🔵 = Response

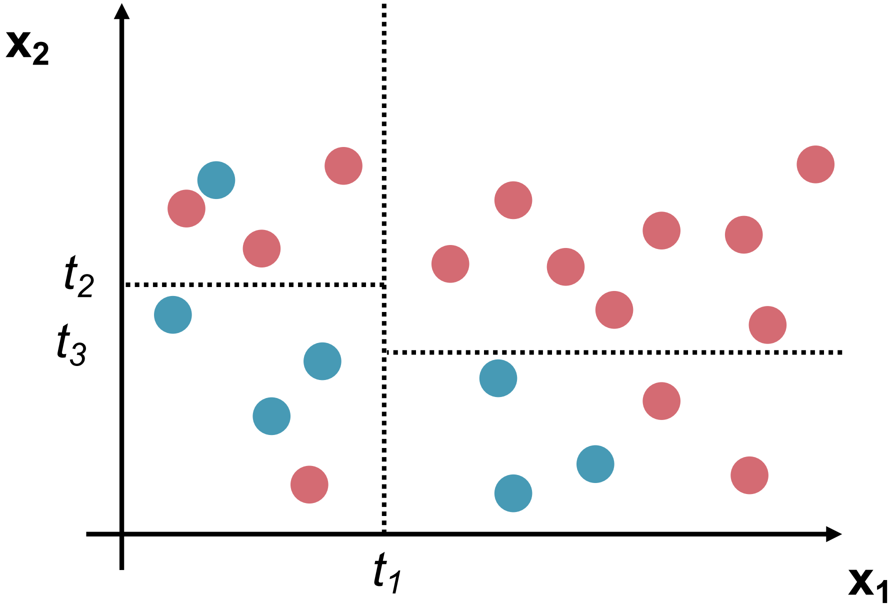
- Stratifying or segmenting the predictor/variable space into simple regions

- Classification tree: vote in each branch

- Regression tree: average in each branch

- Hyperparameters?
    - Tree depth
    - Minimal node size
    - Cost-complexity





Here, no need to scale ☺

Breiman et al., CART, 1984

Node splitting

Regression

$$R_1(j,s) = \{X | X_j < s\} \text{ and } R_2(j,s) = \{X | X_j \geq s\}$$

left child node based on variable *j* and cutoff *s*

- For each node, recursively find value of *j* and *s* that minimize

$$\sum_{i:\, x_i \in R_1(j,s)} (y_i - \hat{y}_{R_1})^2 + \sum_{i:\, x_i \in R_2(j,s)} (y_i - \hat{y}_{R_2})^2,$$

Classification

- Minimize the Gini index = total variance across the *K* classes = purity index

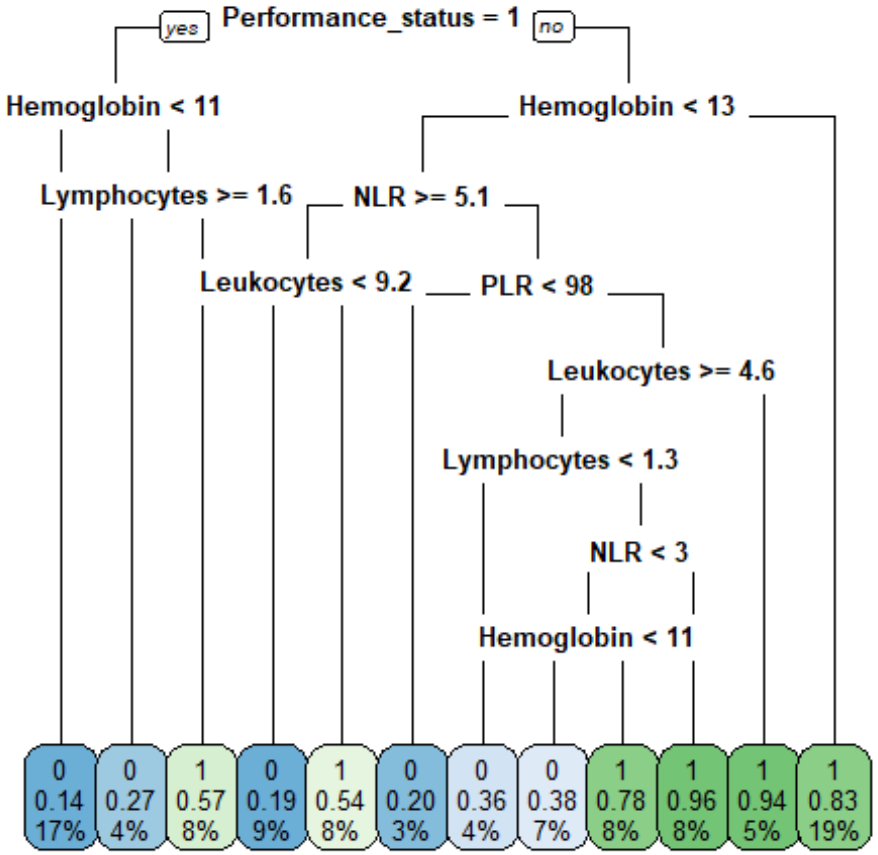
$$G = \sum_{k=1}^{K} \hat{p}_{mk}(1 - \hat{p}_{mk})$$

where $\hat{p}_{mk}$, = proportion of the *k*-th class in node *m.*

For each potential split (i.e., variable x_p and cutoff *s*)

- Calculate *G* in the two child nodes
- Calculate the difference between parent and childs
- Choose the split with maximum difference

Pruning



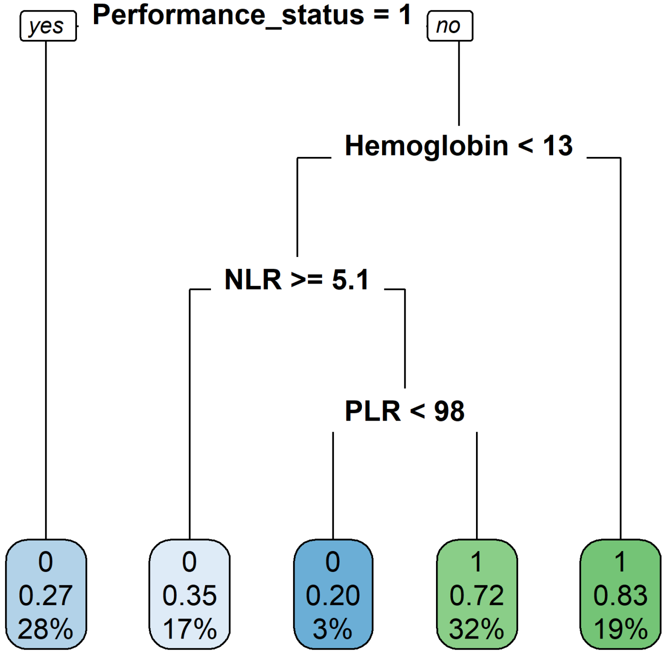
Cost-complexity α = 0

Pruning

$$R_\alpha (T) = R(T) + \alpha|T|$$

Misclassificati
on error

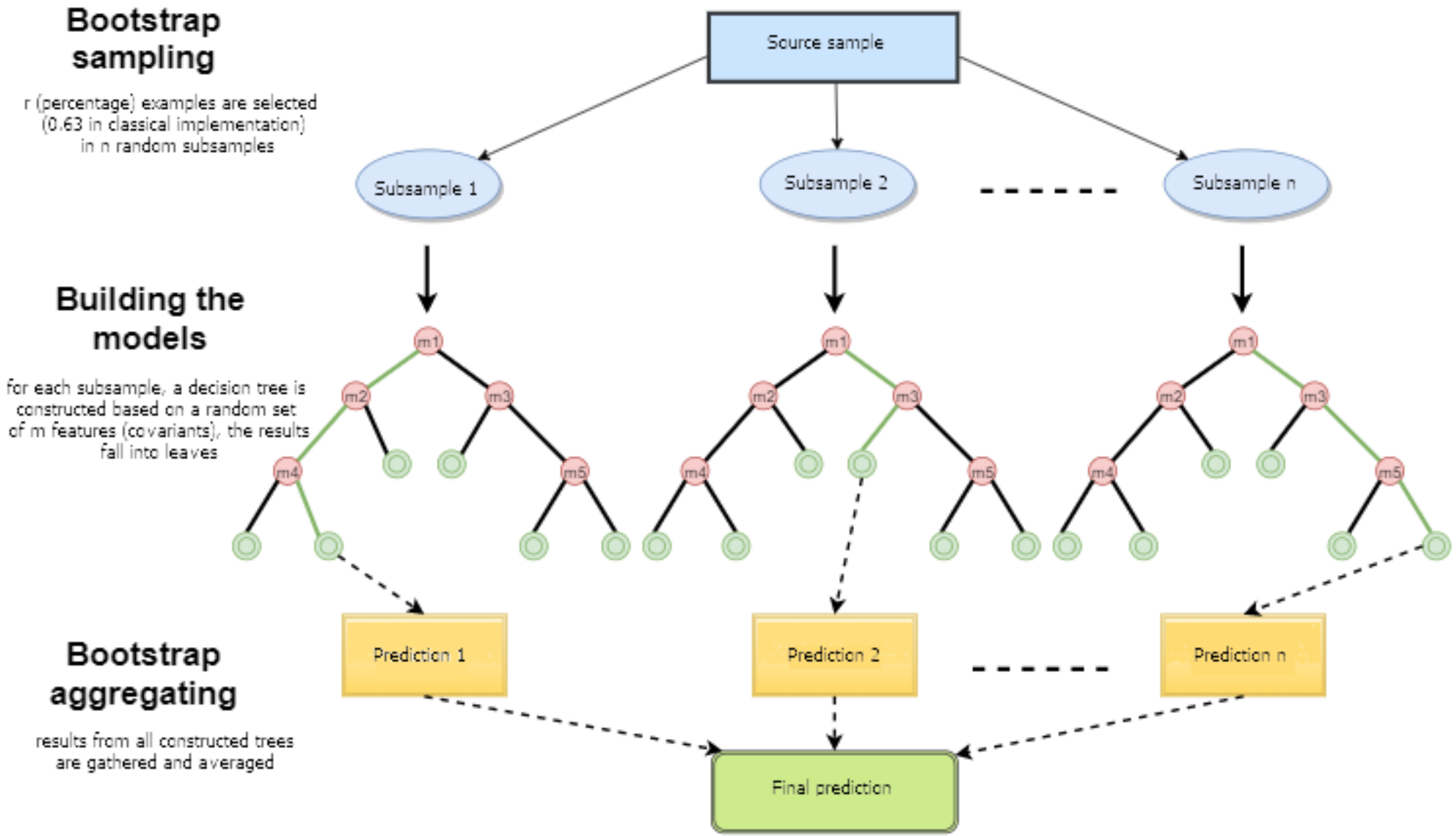
nb of
terminal nodes



Cost-complexity α = 0.02

Main issue = overfitting

Ensemble method: random forest



Bootstrap sampling

r (percentage) examples are selected
(0.63 in classical implementation)
in n random subsamples

Building the models

for each subsample, a decision tree is
constructed based on a random set
of m features (covariants), the results
fall into leaves

Bootstrap aggregating

results from all constructed trees
are gathered and averaged

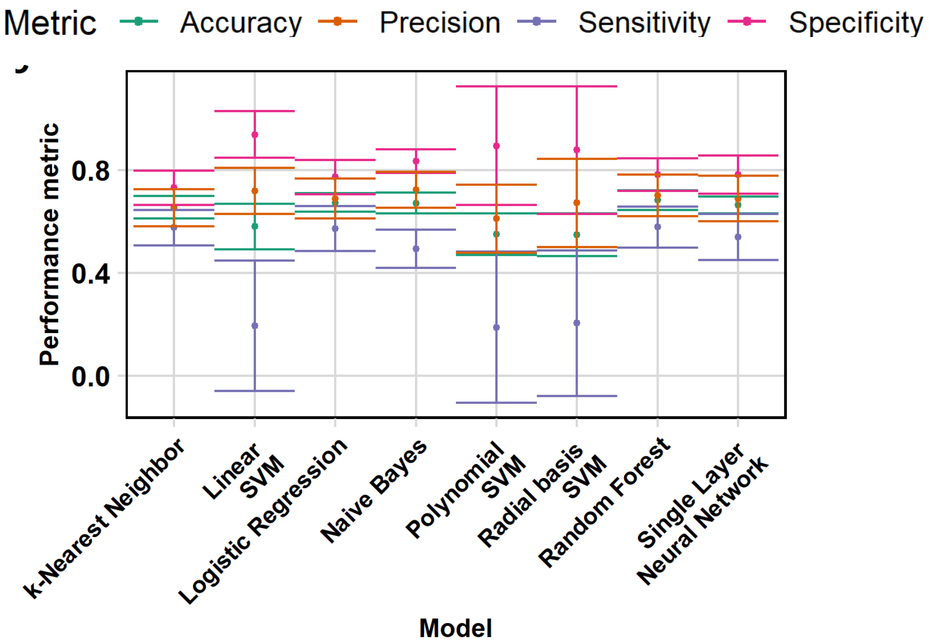
Breiman, Random forests, Machine Learning, 2001

Random forest: hyperparameters

- Number of trees : *trees* [R ranger, default 500], *n_estimators* [python sklearn, default 100]
- Number of variables randomly selected to split each node: *mtry* [R], *max_features* [sklearn], default = $\sqrt{p}$
- Minimal node size *min_n* [R, default 10], *min_samples_leaf* [sklearn, default 1]
- Additional parameters in sklearn: *criterion* (default: Gini), *max_depth, min_impurity_decrease,...*

Example on predict NSCLC response to ICI

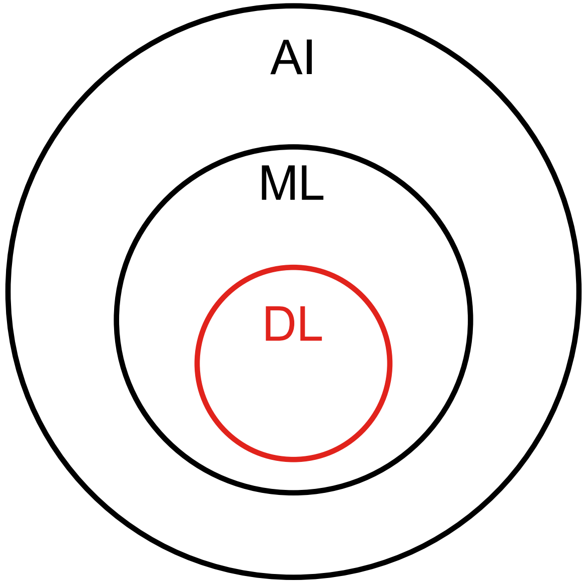
Model	Accuracy	ROC AUC	PPV	NPV	Sensitivity	Specificity
Random Forest	**0.68 ± 0.04**	**0.74 ± 0.03**	0.70 ± 0.08	**0.68 ± 0.06**	**0.58 ± 0.08**	0.78 ± 0.06
Logistic Regression	0.67 ± 0.04	0.73 ± 0.03	0.69 ± 0.08	0.67 ± 0.06	0.57 ± 0.09	0.77 ± 0.07
Naive Bayes	0.67 ± 0.04	0.73 ± 0.03	**0.72 ± 0.07**	0.65 ± 0.06	0.49 ± 0.07	0.83 ± 0.05
Single Layer Neural Network	0.66 ± 0.03	0.72 ± 0.03	0.69 ± 0.09	0.66 ± 0.06	0.54 ± 0.09	0.78 ± 0.07
k-Nearest Neighbour	0.66 ± 0.04	0.69 ± 0.04	0.65 ± 0.07	0.66 ± 0.06	0.58 ± 0.07	0.73 ± 0.07
Linear SVM	0.58 ± 0.09	0.73 ± 0.03	0.72 ± 0.09	0.58 ± 0.10	0.19 ± 0.25	**0.94 ± 0.09**
Polynomial SVM	0.55 ± 0.08	0.73 ± 0.03	0.61 ± 0.13	0.58 ± 0.13	0.19 ± 0.29	0.89 ± 0.23
Radial basis SVM	0.55 ± 0.08	0.73 ± 0.03	0.67 ± 0.17	0.56 ± 0.06	0.20 ± 0.28	0.88 ± 0.25

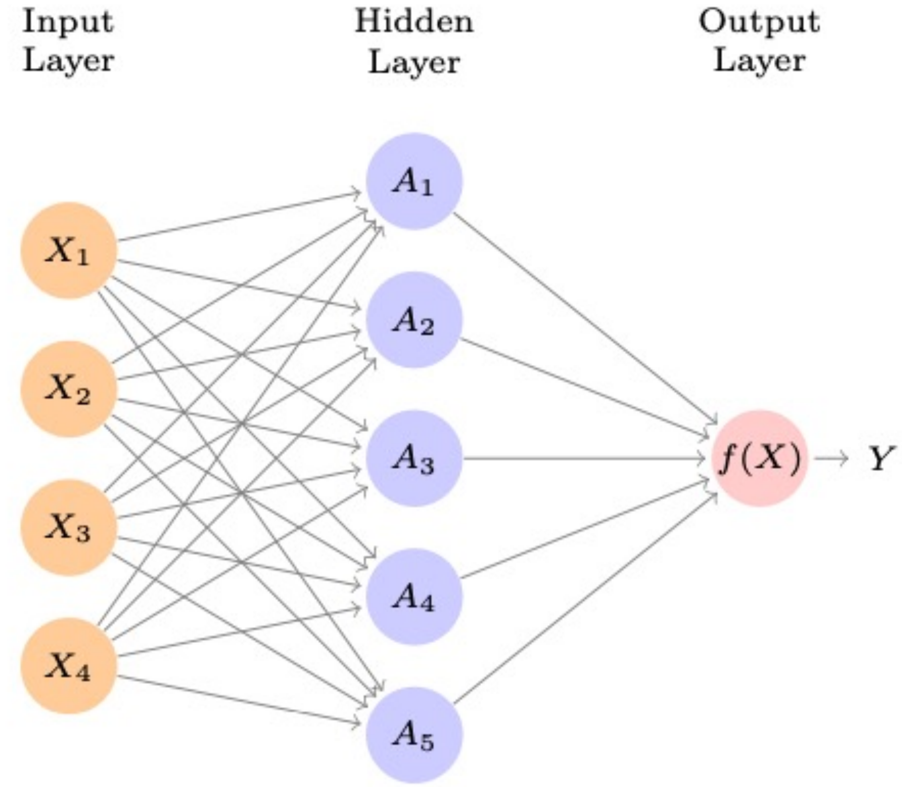


Benzekry et al., Cancers, 2021

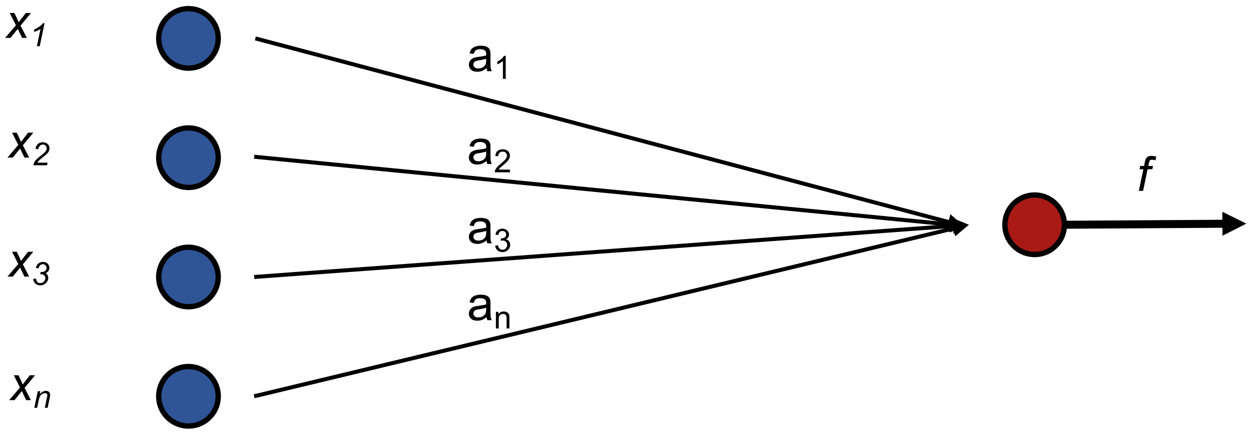
Artificial neural networks

Artificial neural networks





Perceptron

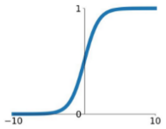


$y = f(b + a_1x_1 + \ldots + a_nx_n)$

Activation Functions

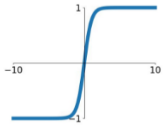
Sigmoid

$\sigma(x) = \frac{1}{1+e^{-x}}$



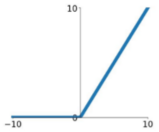
tanh

$\tanh(x)$

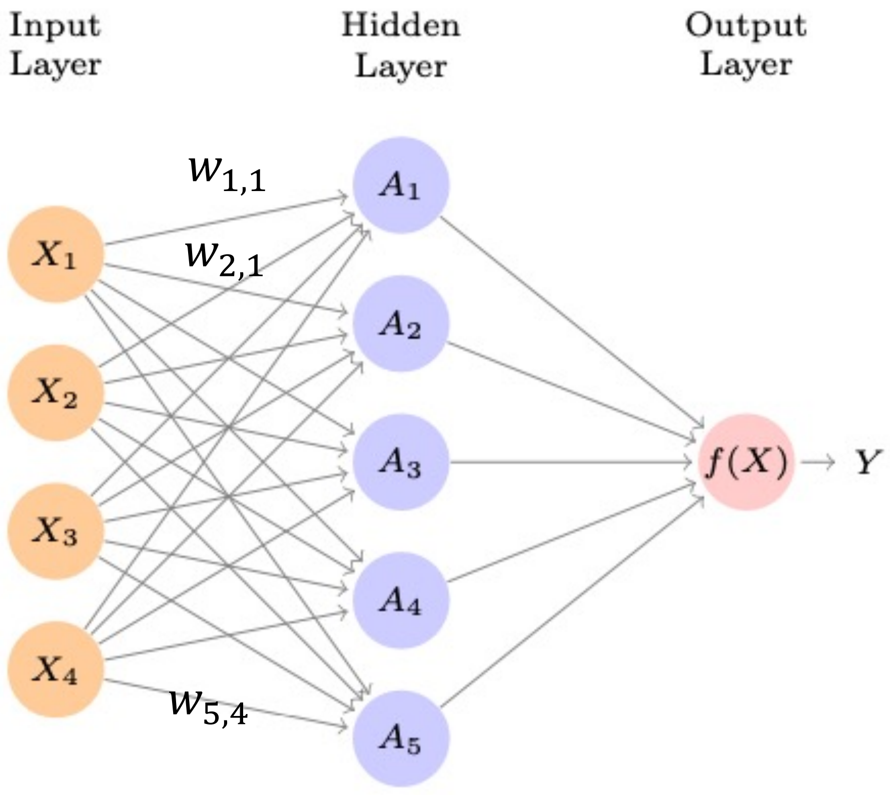


ReLU

$\max(0, x)$



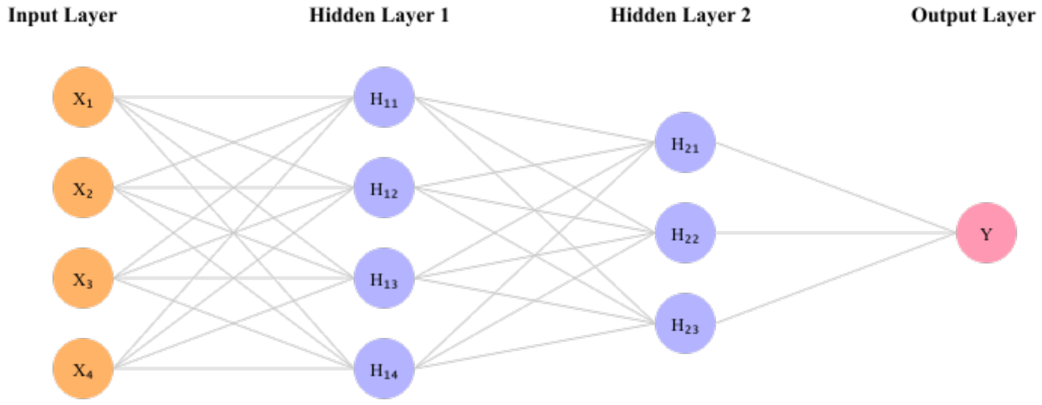
Feed-forward neural network



$$W \in \mathbb{R}^{5,4}$$

$$A = W \cdot X$$
$$Y = f(W \cdot X)$$

Multiple layers



$$Y = f_2(W_2 \cdot f_1(W_1 \cdot X))$$

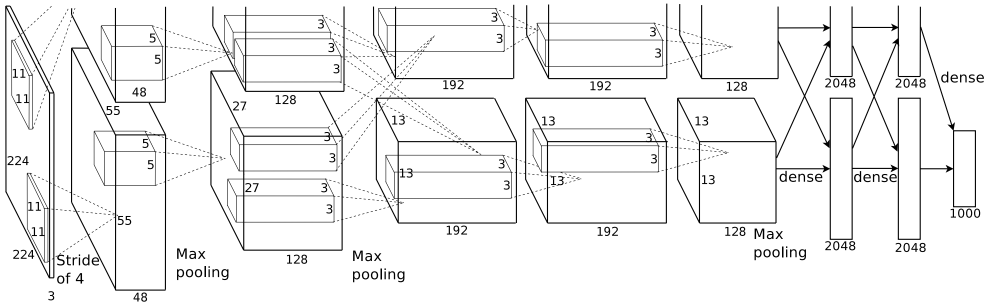
Training = minimize loss ⟶ **gradient descent**

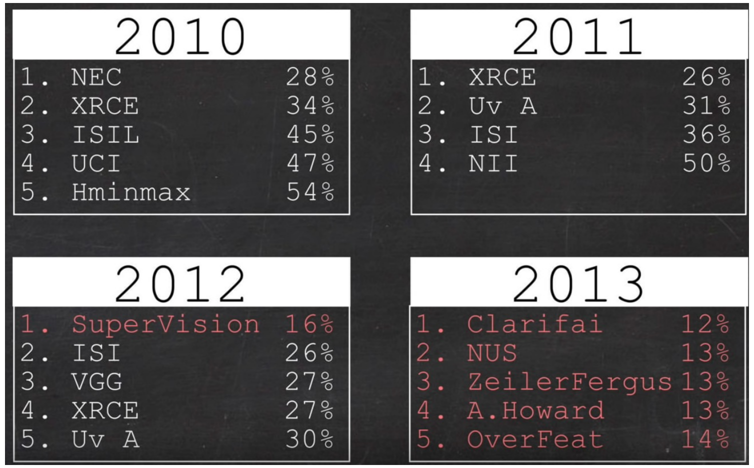
Backpropagation uses the chain rule and matrix products

Rumelhart, D. E., Hinton, G. E. & Williams, R. J. Learning representations by back-propagating errors. Nature ***323****, 533–536 (1986).*

Success example of DL: computer vision

- 1.2 million images (ImageNet, Stanford) used to train a deep convolutional neural network





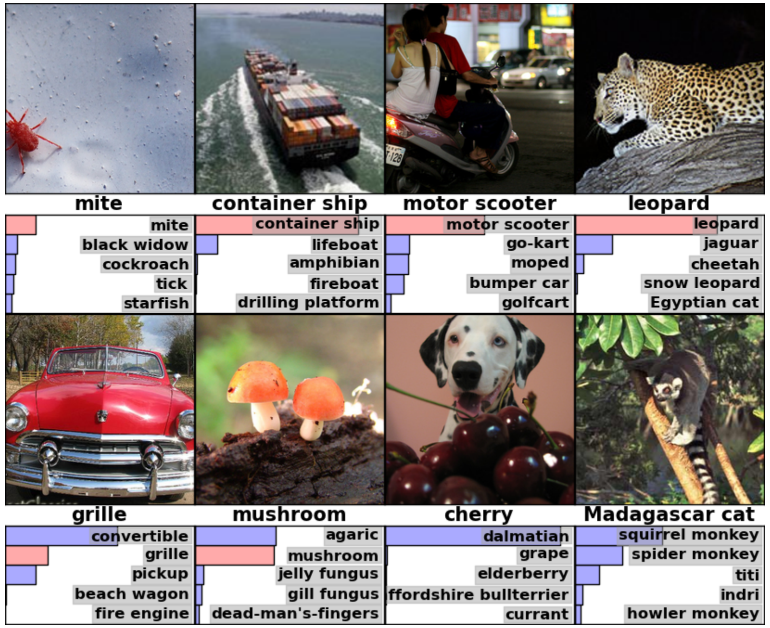
2010	
1. NEC	28%
2. XRCE	34%
3. ISIL	45%
4. UCI	47%
5. Hminmax	54%

2011	
1. XRCE	26%
2. Uv A	31%
3. ISI	36%
4. NII	50%

2012	
1. SuperVision	16%
2. ISI	26%
3. VGG	27%
4. XRCE	27%
5. Uv A	30%

2013	
1. Clarifai	12%
2. NUS	13%
3. ZeilerFergus	13%
4. A.Howard	13%
5. OverFeat	14%

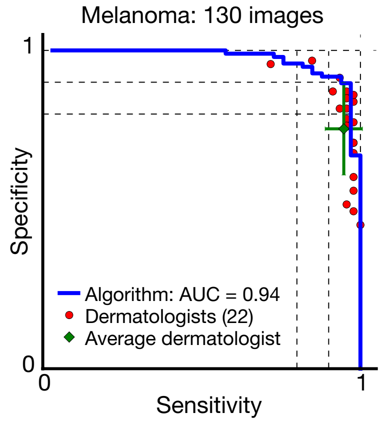
©Science Etonnante

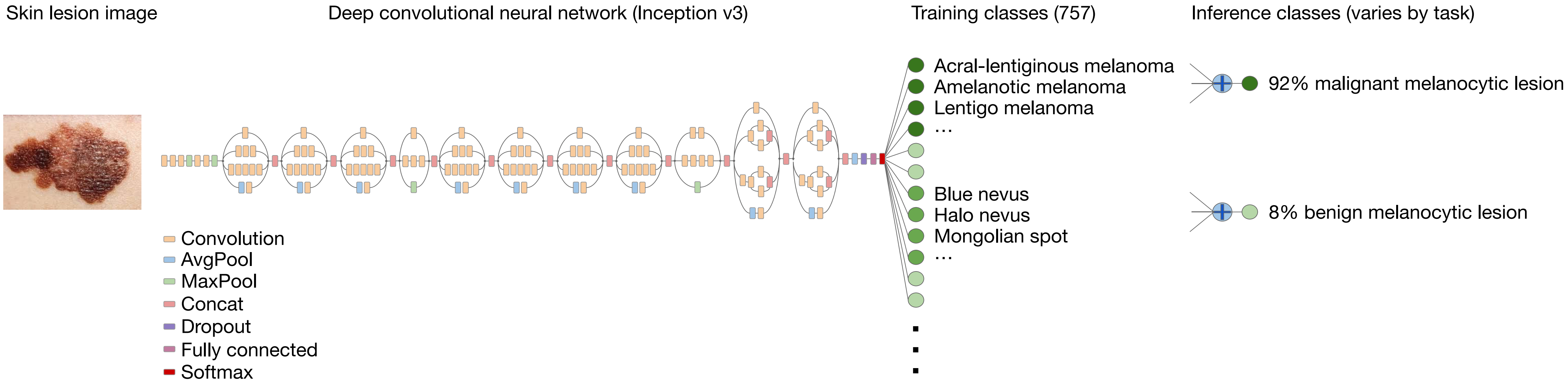


Krizhevsky, Sutskever, Hinton, ImageNet classification with deep convolutional neural networks, NIPS, 2012 (cited 135 158)

Classification of skin lesions

- 129 450 anntotated images
- Task = prediction benign/malignant
- Similar performances as dermatologists





Skin lesion image

Deep convolutional neural network (Inception v3)

Training classes (757)

Inference classes (varies by task)

- Acral-lentiginous melanoma
- Amelanotic melanoma
- Lentigo melanoma
- …
- Blue nevus
- Halo nevus
- Mongolian spot
- …

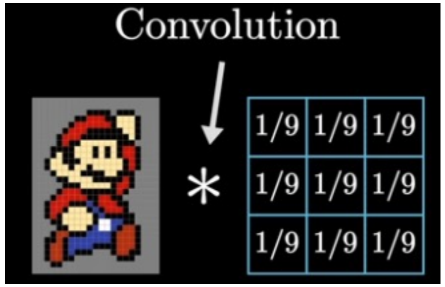
92% malignant melanocytic lesion

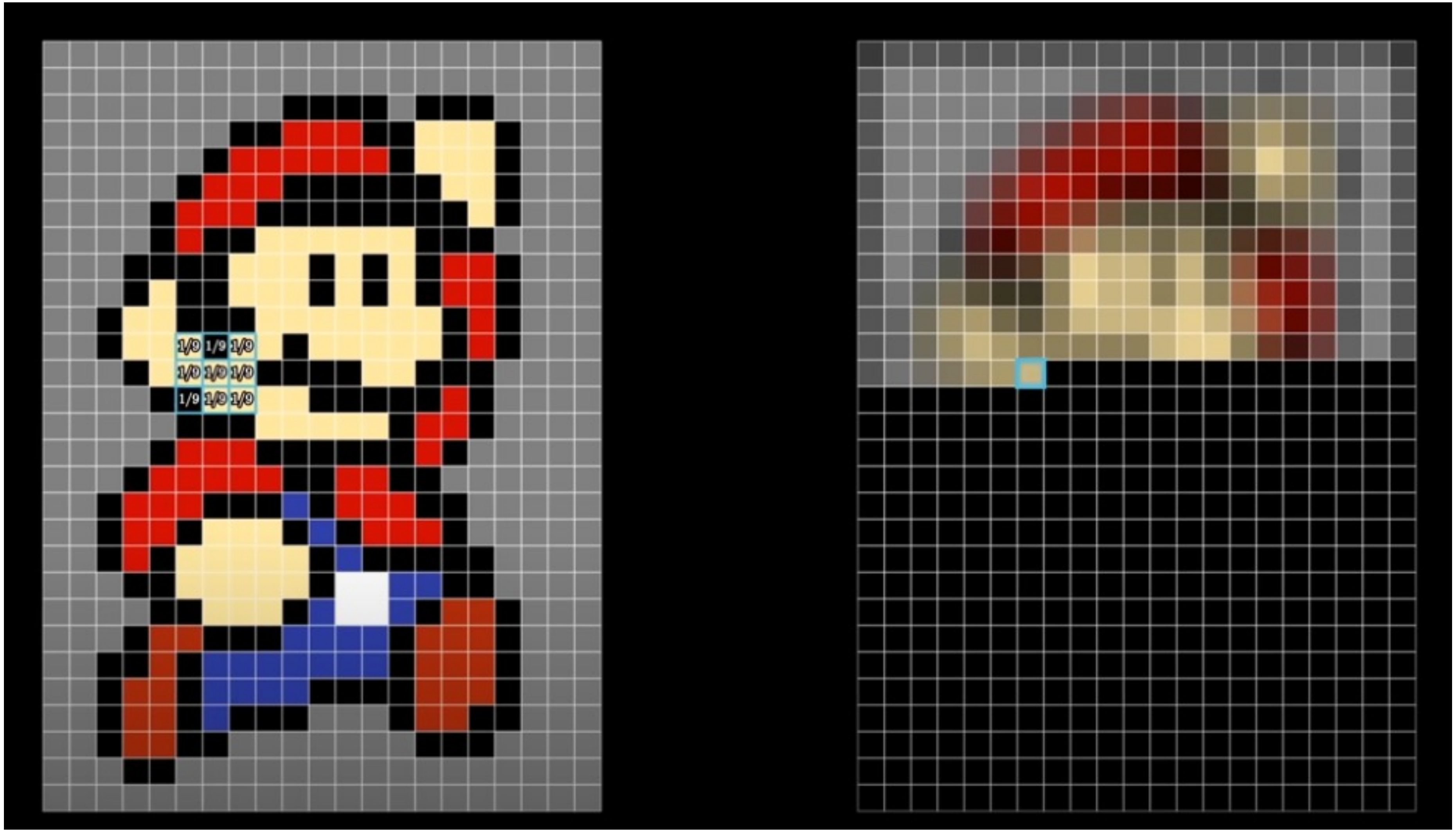
8% benign melanocytic lesion

- Convolution
- AvgPool
- MaxPool
- Concat
- Dropout
- Fully connected
- Softmax

Esteva et al. (Stanford), Dermatologist-level classification of skin cancer with deep neural networks, Nature, 2017

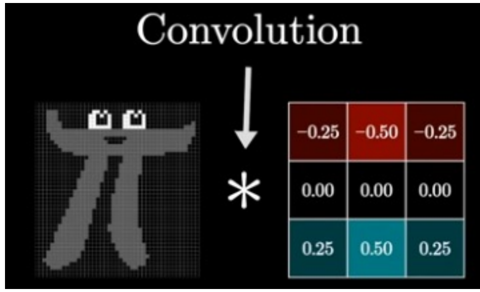
Convolutional neural network



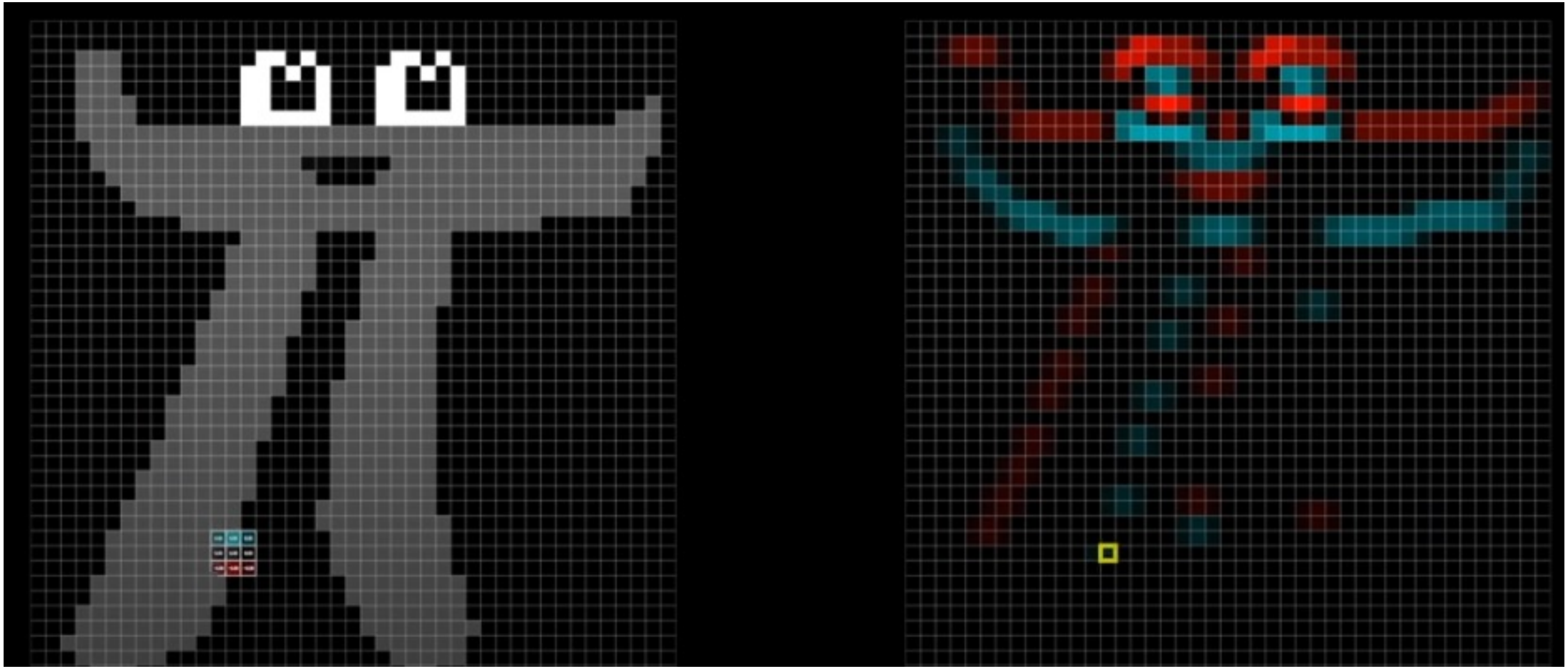


©3Blue1Brown

Convolutional neural network



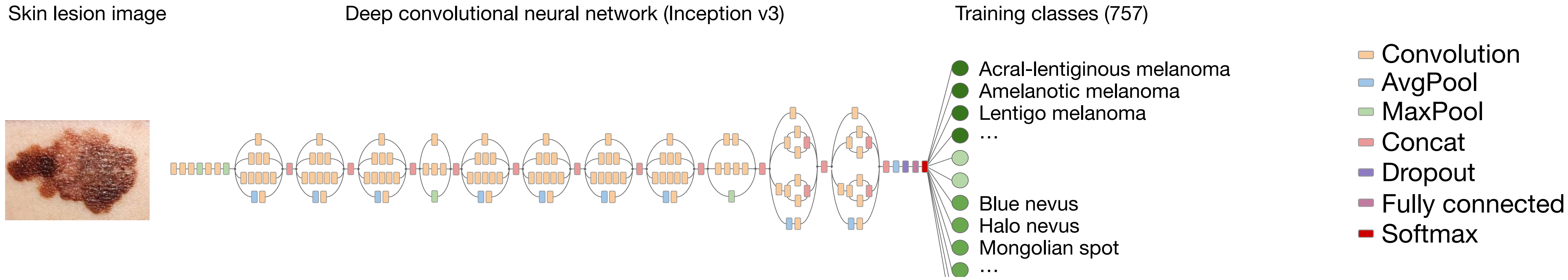
−0.25	−0.50	−0.25
0.00	0.00	0.00
0.25	0.50	0.25



©3Blue1Brown

Other NNs used

- Avg/MaxPool = reduce the image dimension by subdividing and taking the average/max in each region
- Concat = concatenates the outputs
- Dropout = randomly drops a subset of neurons during a training iteration (disabled during testing)
- Fully connected
- Softmax = generalization of logistic to *K* classes



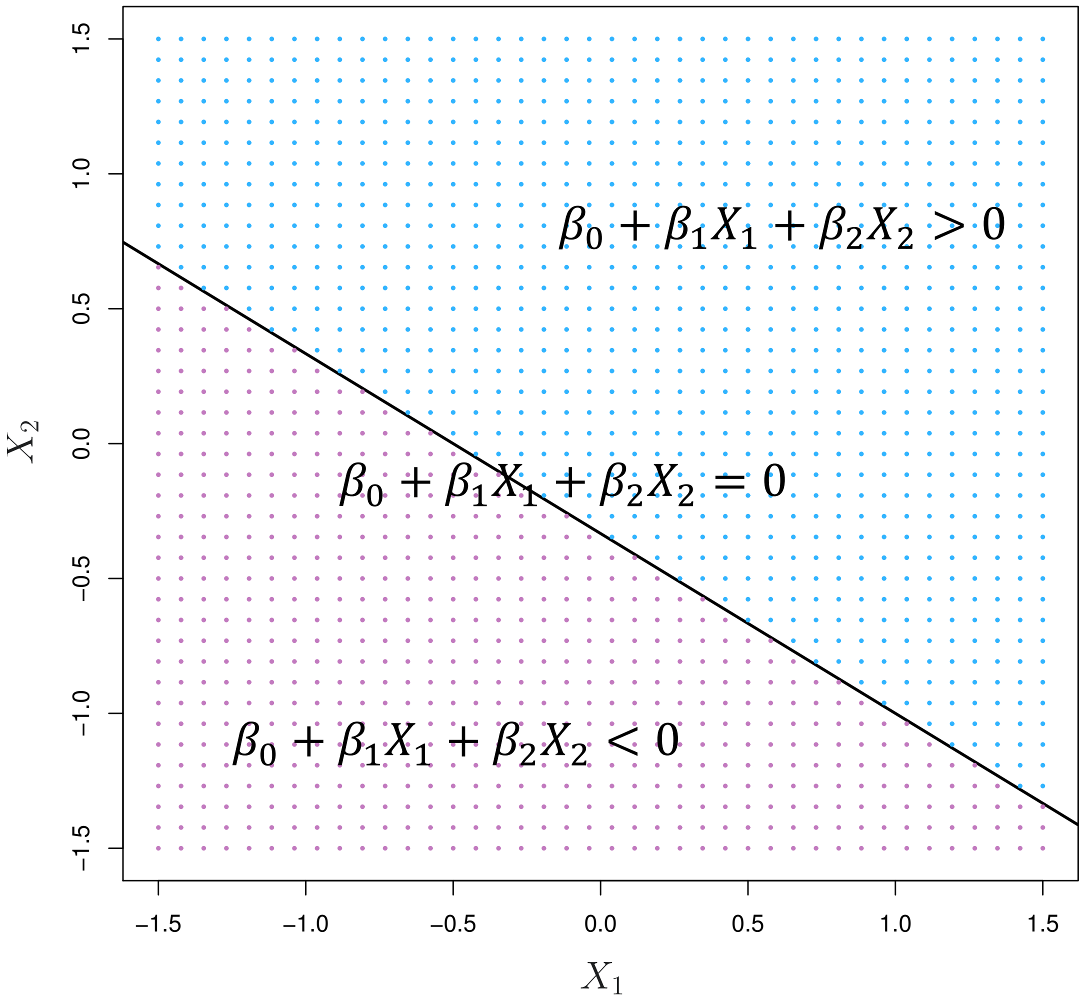
Support vector machines

Support vector machines

- Developed in the computer science community in the 1990s

- Considered one of the best “out of the box” classifiers

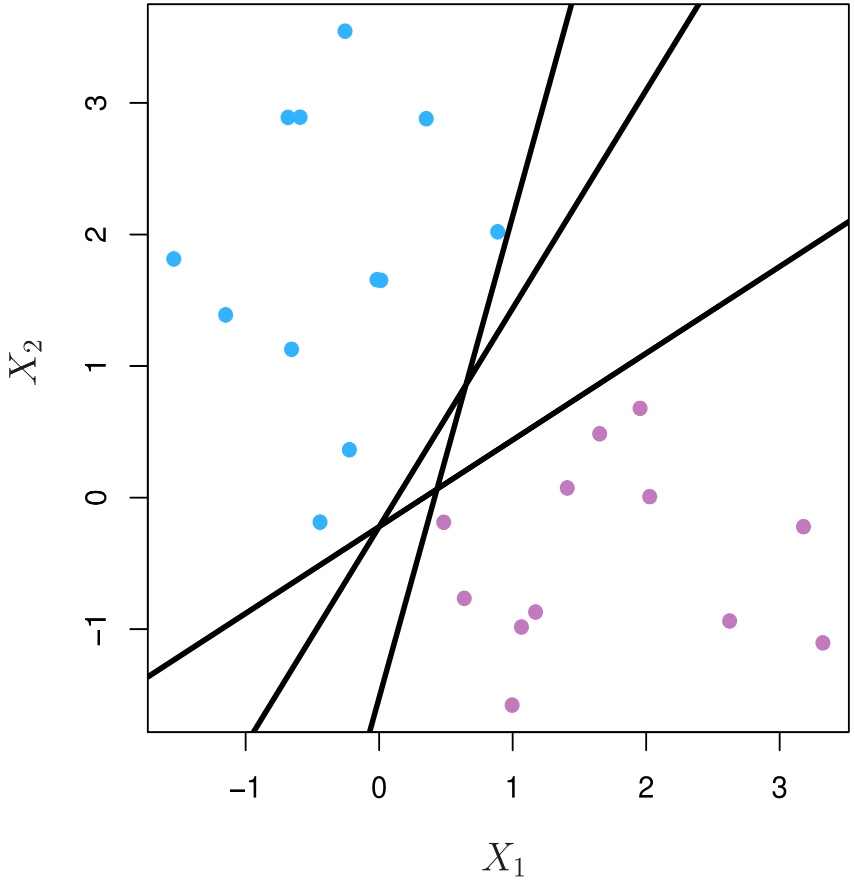
Hyperplane


$\beta_0 + \beta_1 X_1 + \beta_2 X_2 > 0$
$\beta_0 + \beta_1 X_1 + \beta_2 X_2 = 0$
$\beta_0 + \beta_1 X_1 + \beta_2 X_2 < 0$
X_2
1.5
1.0
0.5
0.0
−0.5
−1.0
−1.5
−1.5
−1.0
−0.5
0.0
0.5
1.0
1.5
X_1

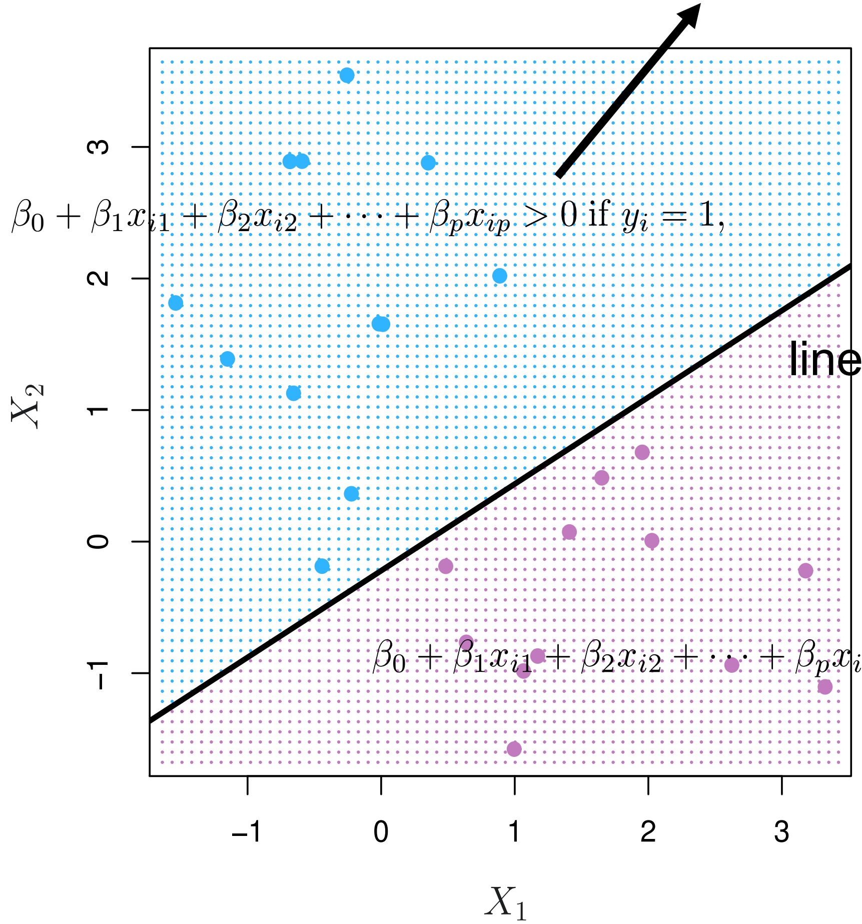
Separating hyperplanes

Assume two classes: y = 1 or y = -1

3 separating hyperplanes among many possibles



$$y_i(\beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_p x_{ip}) > 0$$



$\beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_p x_{ip} > 0$ if $y_i = 1$,

linear decision boundary

$\beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_p x_{ip} < 0$ if $y_i = -1$

Maximal margin classifier

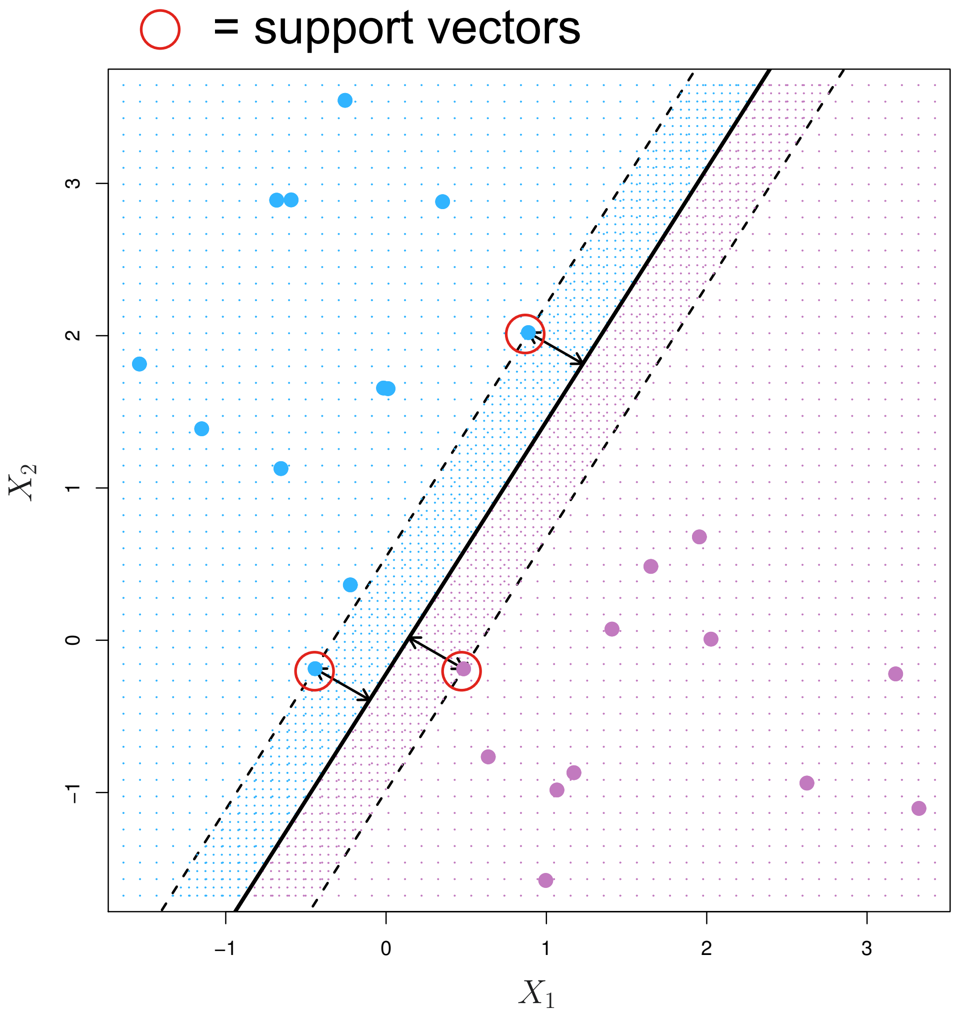
Which of the infinite possible separating hyperplanes to use?

→ Maximal margin hyperplane = separating hyperplane that is the farthest from train observations

→ Maximal margin classifier

- It depends strongly on the support vectors but not on the other observations

→ robust to the behavior of observations far from hyperplane (outliers)



How to find the maximal margin classifier?

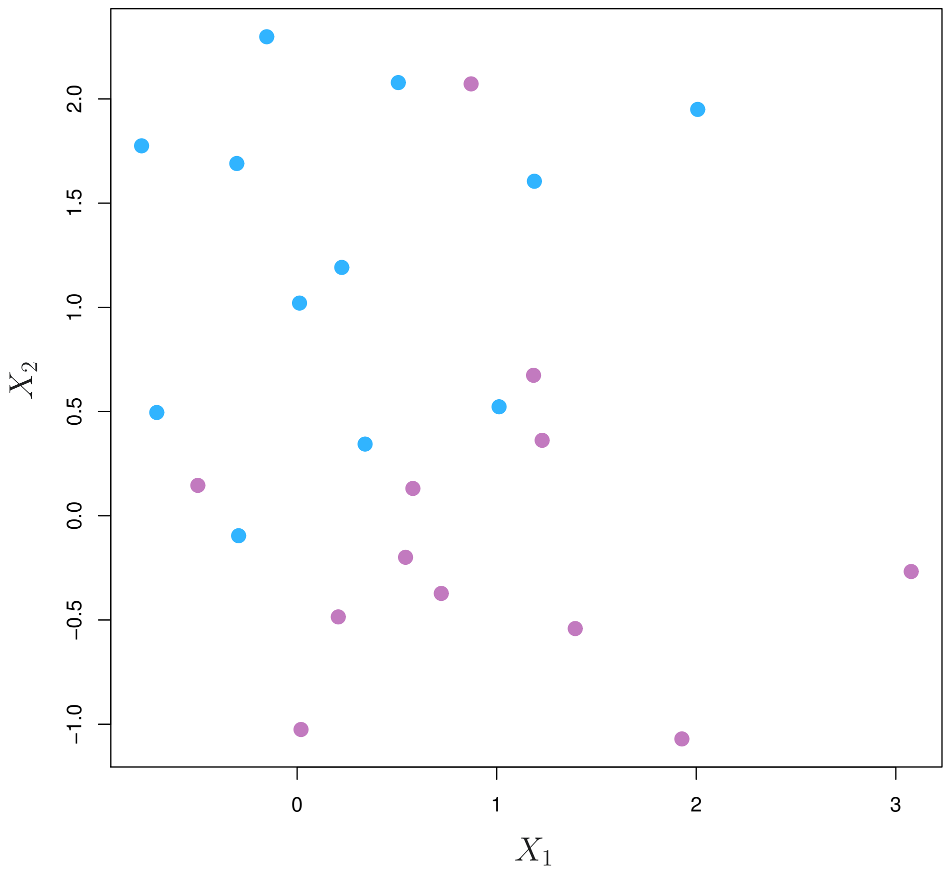
Optimization problem!

$$\underset{\beta_0,\beta_1,\ldots,\beta_p,M}{\text{maximize}} \quad M$$ ← find $\beta_0, \beta_1, \ldots, \beta_p$ that maximize the margin *M*

$$\text{subject to } \sum_{j=1}^{p} \beta_j^2 = 1,$$ ← ensures that distance is given by $y_i(\beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_p x_{ip})$

$$y_i(\beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_p x_{ip}) \geq M \quad \forall\, i = 1, \ldots, n$$ ← correct side of hyperplane
distance ≥ *M*

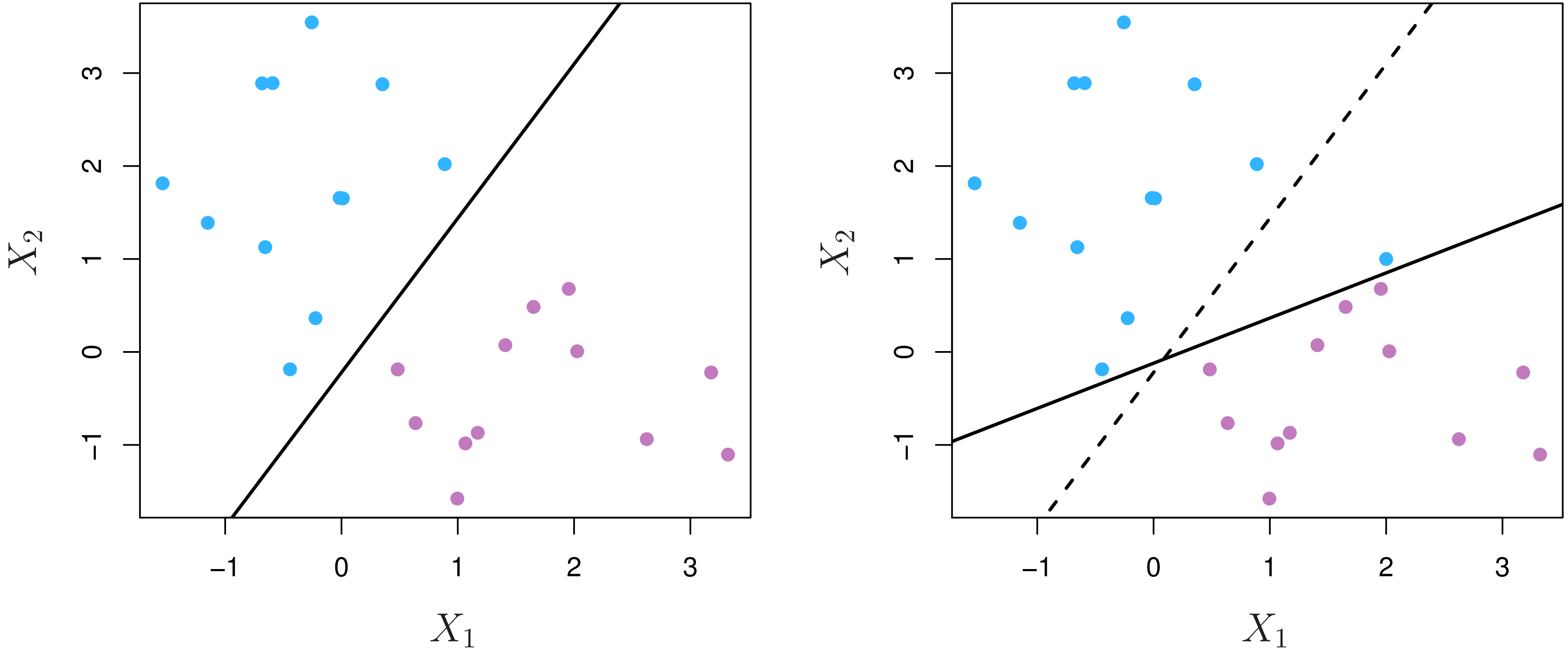
Non-separable case



No solution exist to the optimization problem!

→ extend the concept to a hyperplane that *almost* separates the classes, using a *soft-margin*

Even when separable



→ the maximal margin classifier has high variance! (linked to overfit)

→ solution = allow for some observations to be misclassified

Support vector classifier

Separate most of the training observations, but allow some misclassification

Optimization problem

$$\underset{\beta_0,\beta_1,\ldots,\beta_p,\epsilon_1,\ldots,\epsilon_n,\,M}{\text{maximize}} \quad M$$

find $\beta_0, \beta_1, \ldots, \beta_p$ that maximize the margin *M*

$$\text{subject to } \sum_{j=1}^{p} \beta_j^2 = 1,$$

ensures that distance is given by $y_i(\beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_p x_{ip})$

$$y_i(\beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_p x_{ip}) \geq M(1 - \epsilon_i),$$

distance can be smaller than *M*

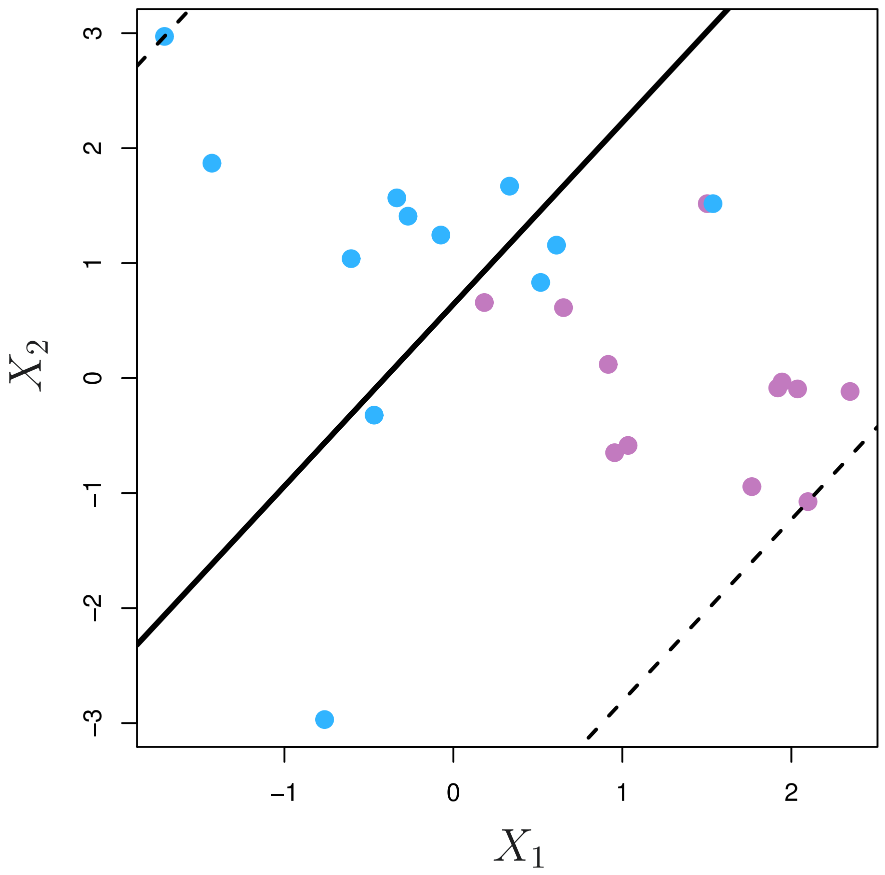
= 0 → correct side of margin
> 0 → wrong side of margin
> 1 → wrong side of hyperplane

$$\epsilon_i \geq 0, \quad \sum_{i=1}^{n} \epsilon_i \leq C,$$

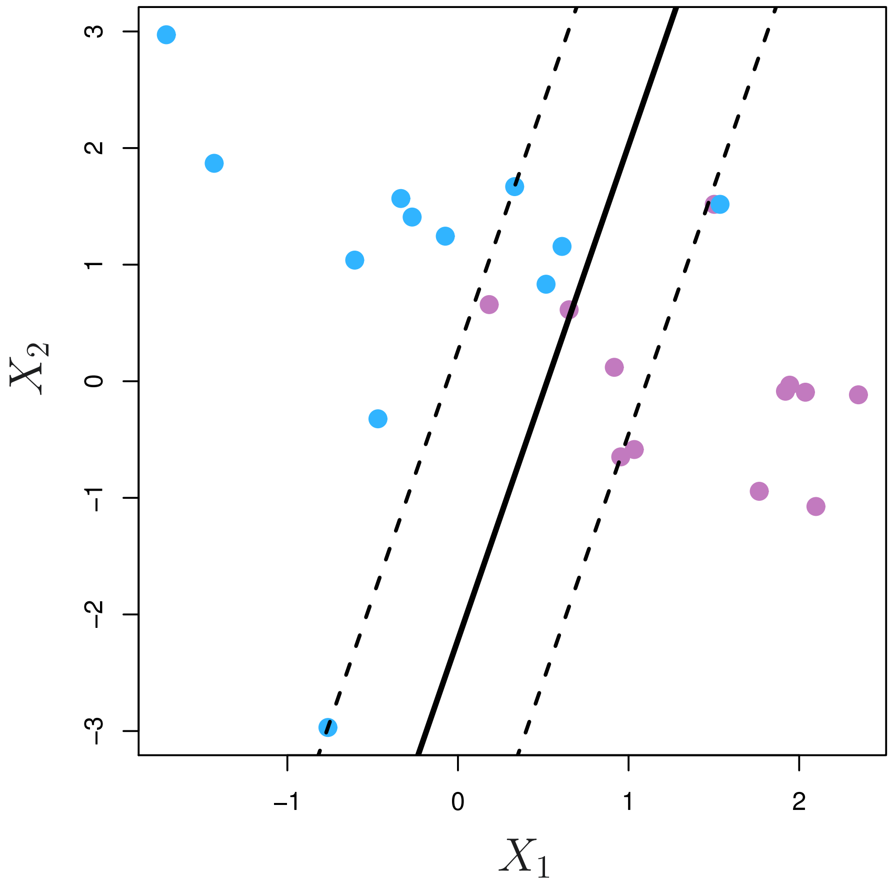
tuning hyperparameter
number and severity of violations of the margin we tolerate

Examples

C large, high bias, low variance



C small, low bias, high variance



C = tuning hyperparameter, choosen by cross-validation, determines bias-variance tradeoff

Large dimension and variable selection

Linearity in large dimension

$$y = f(x) + \varepsilon \approx \beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p + \varepsilon$$

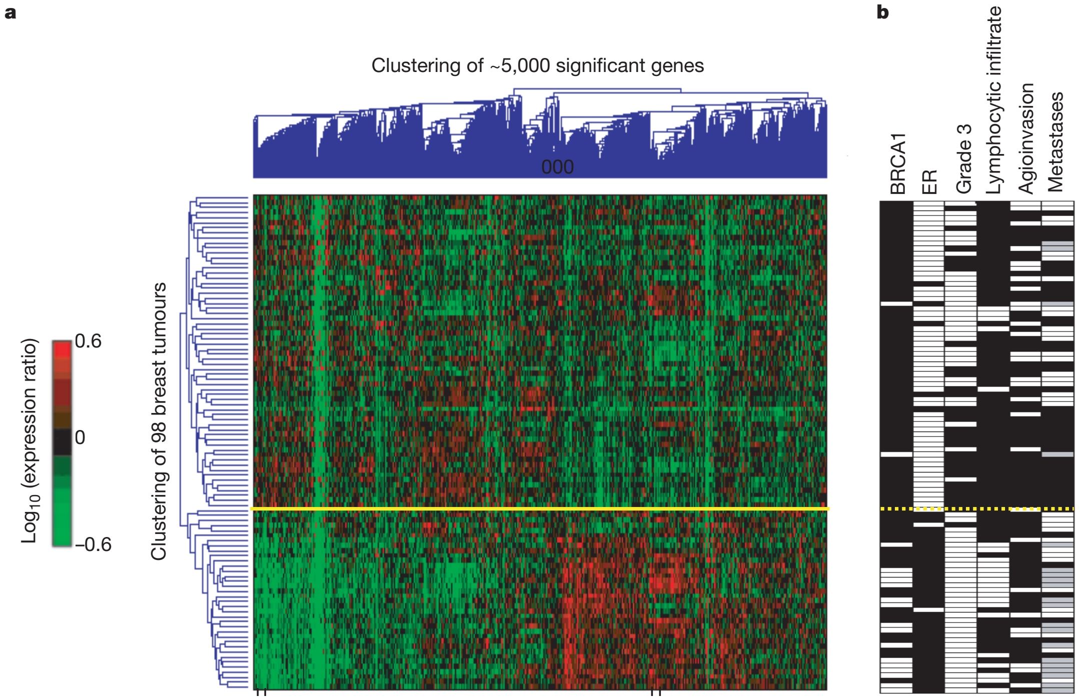
n = number of observations: $y = (y^1, \ldots, y^n)$
p = number of variables

- If f close to linear → low bias
- If n >> p → low variance
- If n ~ p → high variance
- If n << p → infinite variance (no unique least-squares estimate)

→ constraining (or shrinking) the coefficients (β_k) can substantially reduce variance at moderate bias cost

→ variable selection

→ improved accuracy



→ in addition, a lot of variables might be irrelevant, setting $\beta_k = 0$ for them improves interpretability and reduces complexity

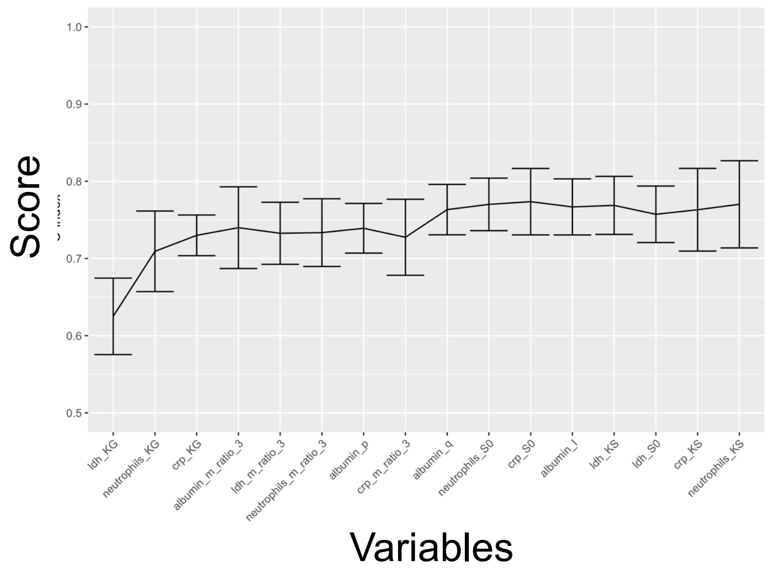
Elementary variable selection

- Rule of thumb: $n = 10 * p$

- Best subset selection: perform all models based on all possible subsets of variables, select best using cross-validation error
  - Costly (2^p possibilities, $2^{15} = 1.13 \times 10^{15}$) !!

- Stepwise selection
  - Forward: start with no variable, add variables one-at-time by selecting the one leading to greatest improvement of fit until all, select best by CV

  - Backward: same but starting by all and removing each on-at-a-time



- However, such methods are usually not advised by the statistical community (usually, due to overfitting)

Three classes of variable selection methods

1. **Filters**: Select features based on statistical properties of data, independent of any specific machine learning algorithm.

   + Fast and computationally efficient.

   - Does not capture feature interactions

   - Examples: Variance, t-tests or chi-square.

2. **Wrappers** : Select features based on a ML model performance by iteratively adding or removing features.

   + Can capture feature interactions.

   + Often provides high accuracy for selected features.

   - Computationally expensive, especially with large feature sets.

   - Examples: Forward/backward selection, recursive feature elimination (RFE).

3. **Embedded** : Feature selection occurs within the training process of the model.

   + Efficient and often provides high accuracy.

   + Integrates selection into model training.

   + Examples: Lasso (L1 regularization), decision tree feature importance, Elastic Net.

Ridge regression

$$\mathrm{RSS} = \sum_{i=1}^{n}\left(y_i - \beta_0 - \sum_{j=1}^{p}\beta_j x_{ij}\right)^2$$

$$\Big\downarrow$$

$$\sum_{i=1}^{n}\left(y_i - \beta_0 - \sum_{j=1}^{p}\beta_j x_{ij}\right)^2 + \lambda\sum_{j=1}^{p}\beta_j^2 = \mathrm{RSS} + \lambda\sum_{j=1}^{p}\beta_j^2$$

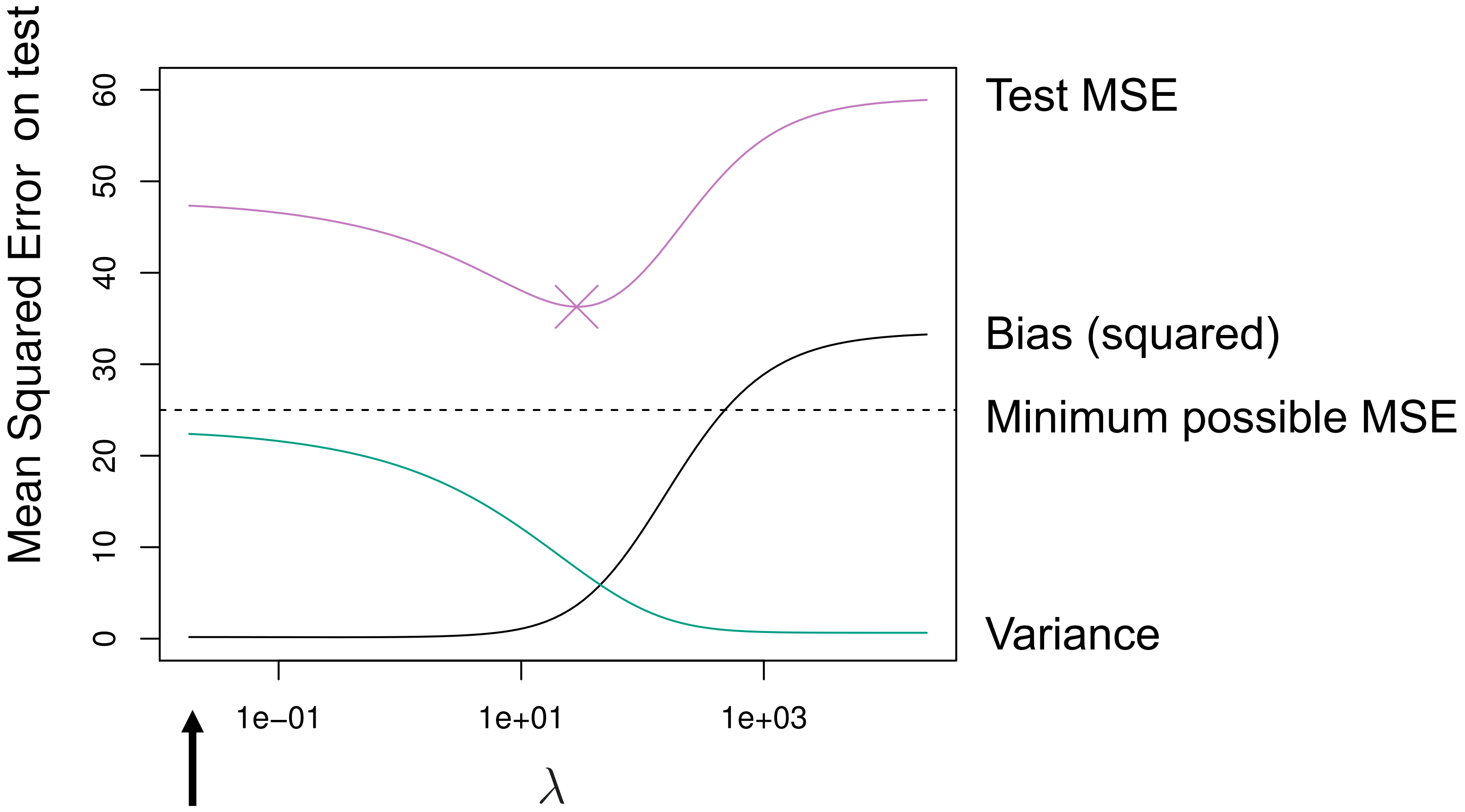
note this does not contain β_0 = mean value with no variables

- different set of coefficient estimates $\hat{\beta}$ for each value of λ
- λ increases → increased bias, decreased variance
- $\lambda =$ tuning parameter, to be determined separately, by cross-validation
- Computational advantage over best subset selection (2^p)

Example

n = 50

p = 45



$\lambda = 0$: least squares

- ridge regression works best in situations where the least squares estimates have **high variance**
- disadvantage = includes all *p* variables

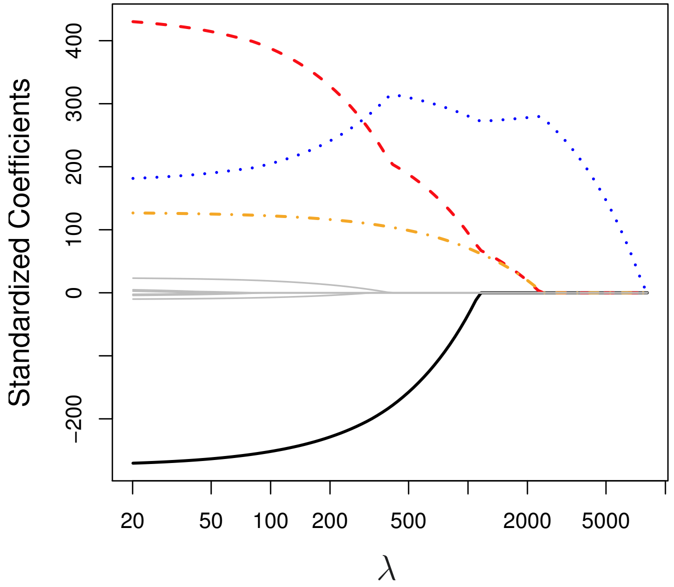
Least absolute shrinkage and selection operator (LASSO)

$$\sum_{i=1}^{n}\left(y_i-\beta_0-\sum_{j=1}^{p}\beta_j x_{ij}\right)^2+\lambda\sum_{j=1}^{p}|\beta_j| = \mathrm{RSS}+\lambda\sum_{j=1}^{p}|\beta_j|$$

- Difference with ridge = ℓ_1 penalization versus ℓ_2
- Forces some coefficients to be zero

→ variable selection

→ better interpretability

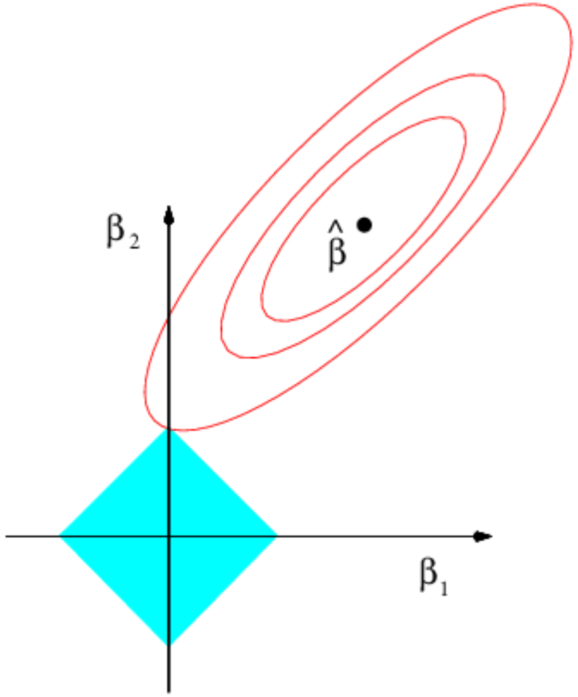


LASSO and ridge

LASSO

$\Updownarrow$

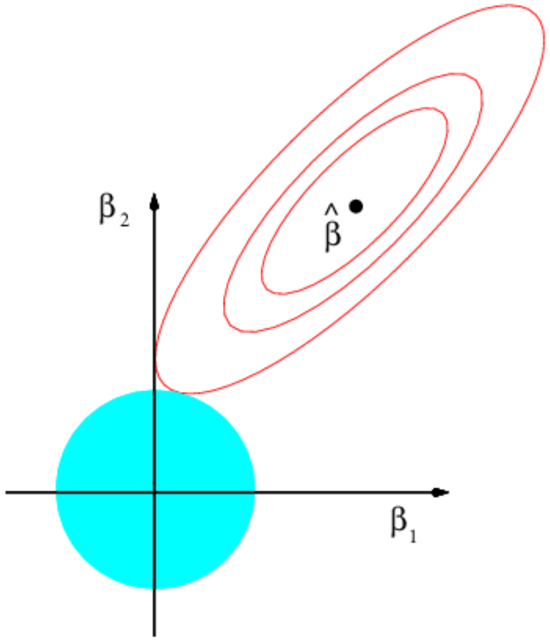
$$\underset{\beta}{\text{minimize}} \left\{ \sum_{i=1}^{n} \left(y_i - \beta_0 - \sum_{j=1}^{p} \beta_j x_{ij} \right)^2 \right\} \quad \text{subject to} \quad \sum_{j=1}^{p} |\beta_j| \leq s$$



Ridge

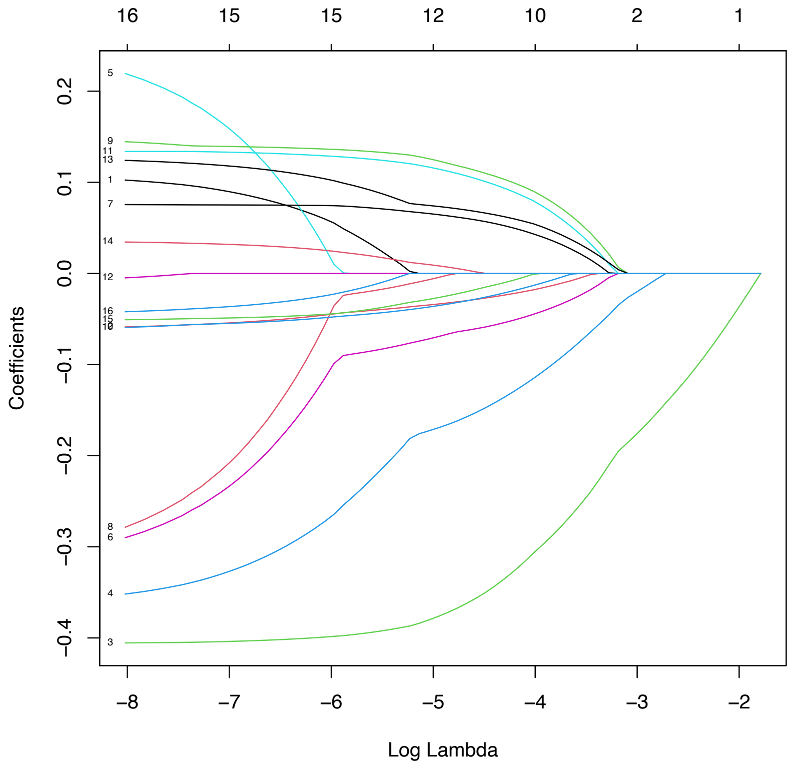
$\Updownarrow$

$$\underset{\beta}{\text{minimize}} \left\{ \sum_{i=1}^{n} \left(y_i - \beta_0 - \sum_{j=1}^{p} \beta_j x_{ij} \right)^2 \right\} \quad \text{subject to} \quad \sum_{j=1}^{p} \beta_j^2 \leq s.$$

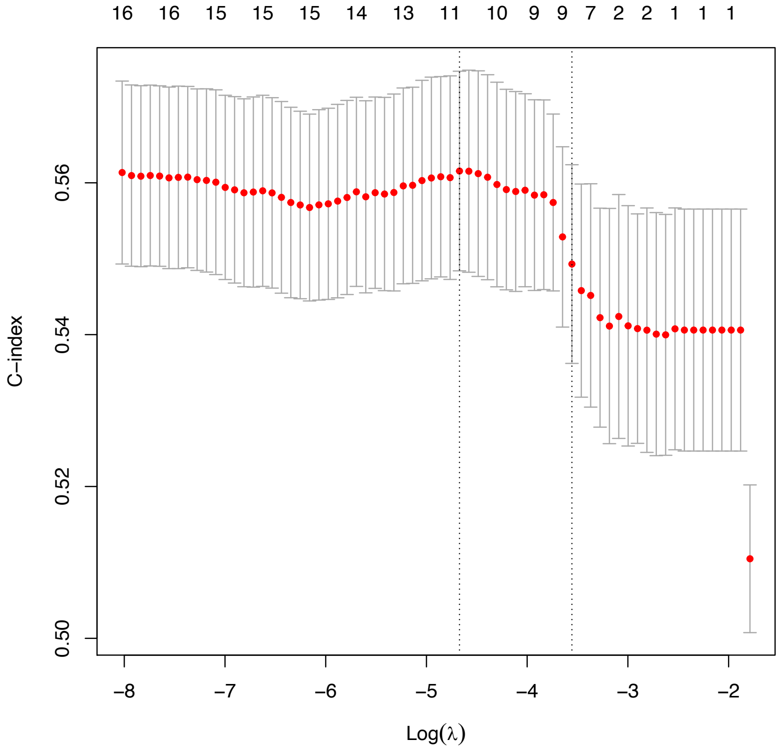


Selecting the tuning parameter λ

Values of the estimated coefficients as λ decreases



Prediction score as λ decreases



Unsupervised learning

Challenge of unsupervised learning

- For supervised learning, we have ways to **assess the performances**

- In unsupervised learning, there is **no truth** to refer to

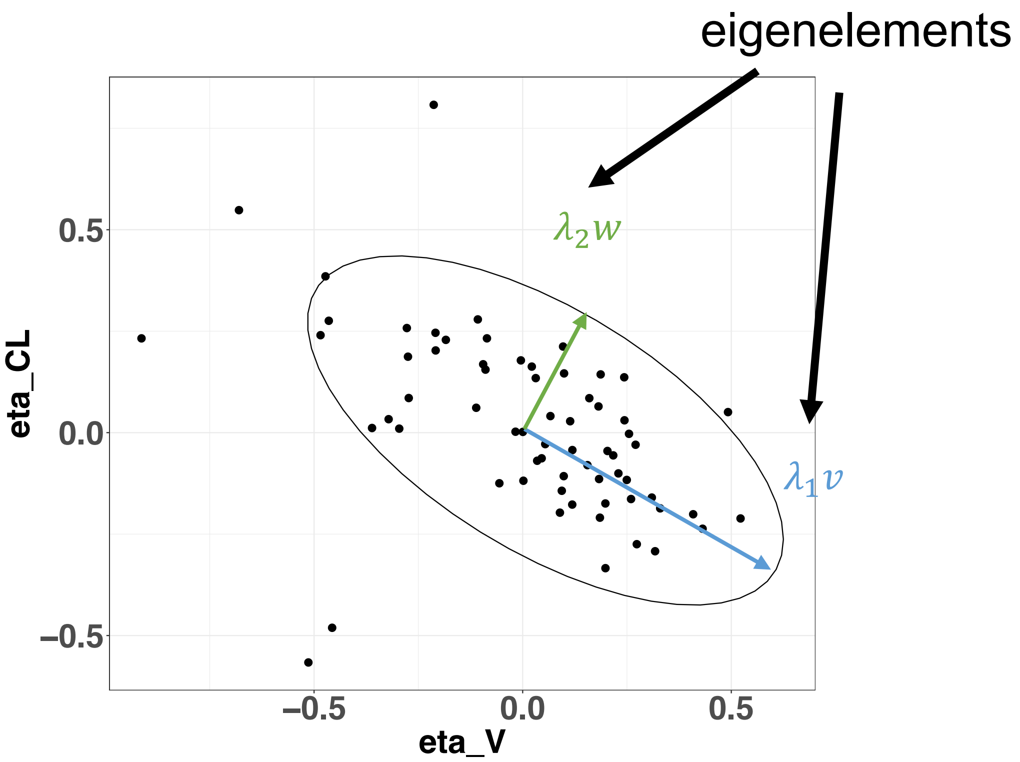
Dimensionality reduction: Principal Component Analysis

Transforms (correlated) variables into a set of uncorrelated (orthogonal) components

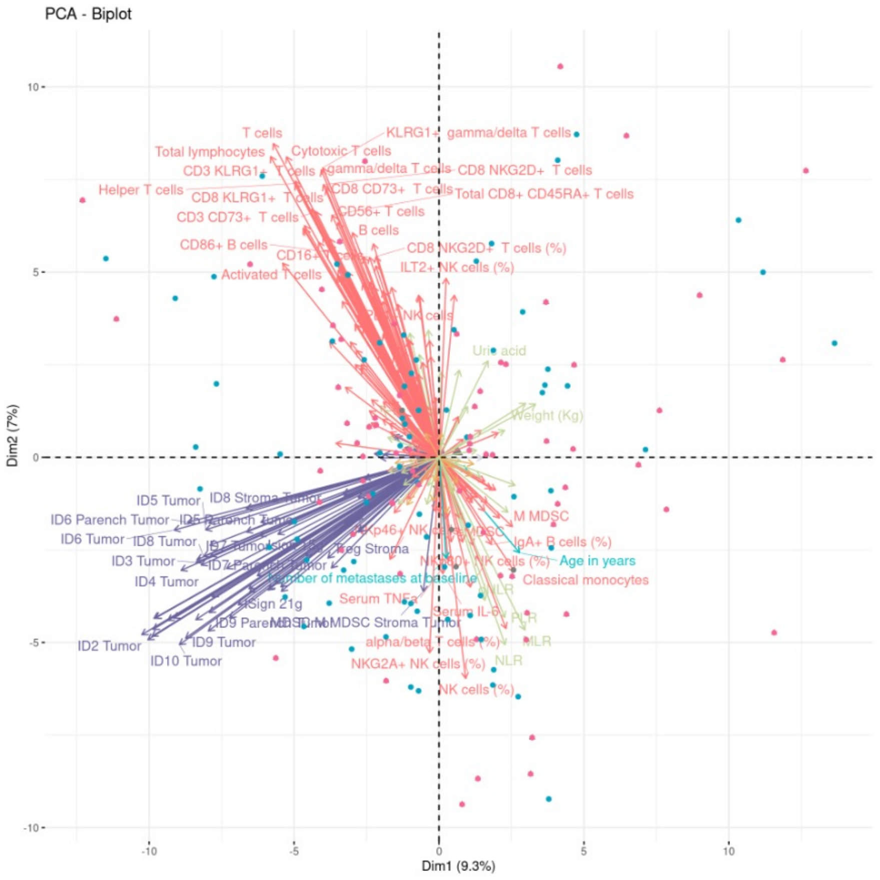
+ Reduces the number of features while retaining as much variance (information) as possible.

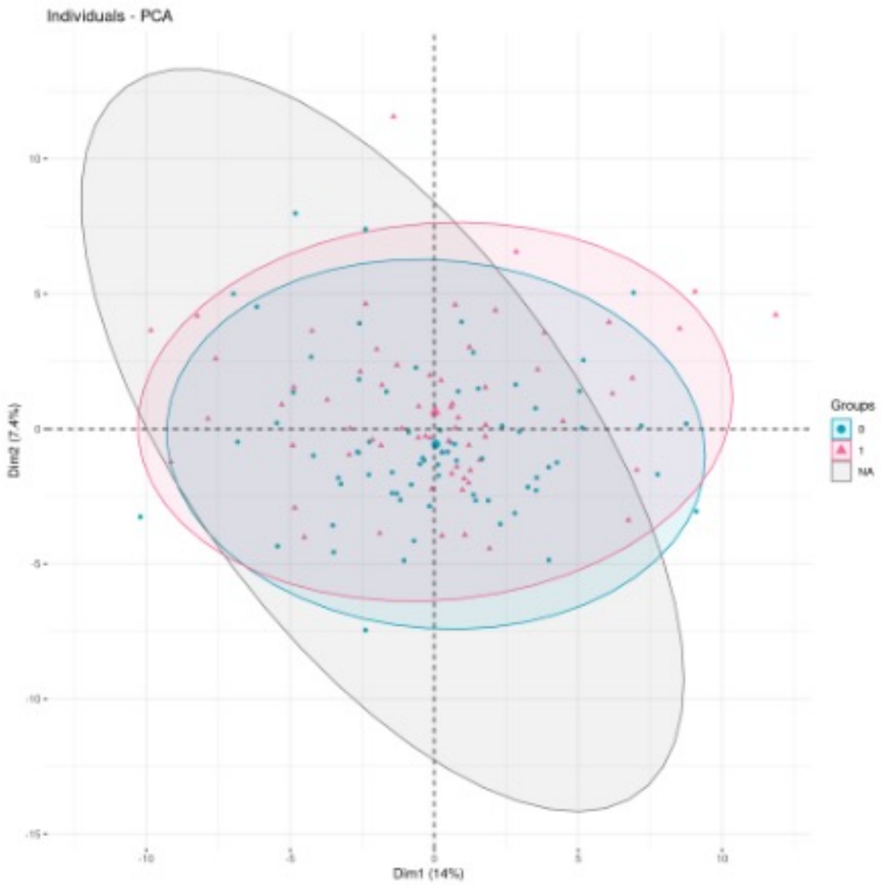
- The new variables are not interpretable anymore

- first eigenvector = direction of the data of maximal variance
- first eigenvalue = variance of the data in this direction



Dimensionality reduction: Principal Component Analysis



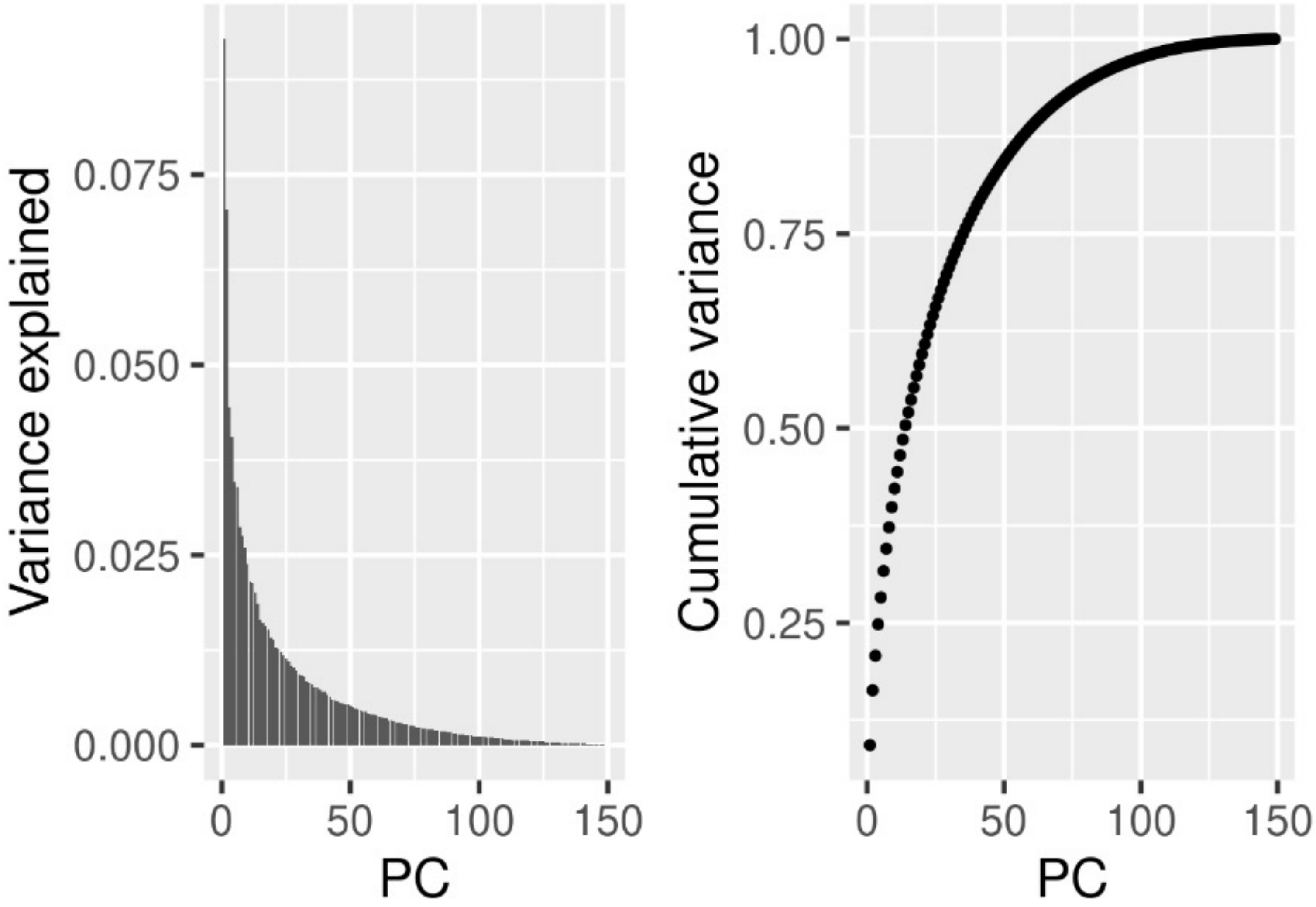


N = 149
p = 315

Clinical follow up
Tumor biomarker
Vasculo
Routine blood tests
Pathology
Circulating immune cells
Soluble immune markers

FactoMineR package

Dimensionality reduction: Principal Component Analysis



N = 149
p = 315

FactoMineR package

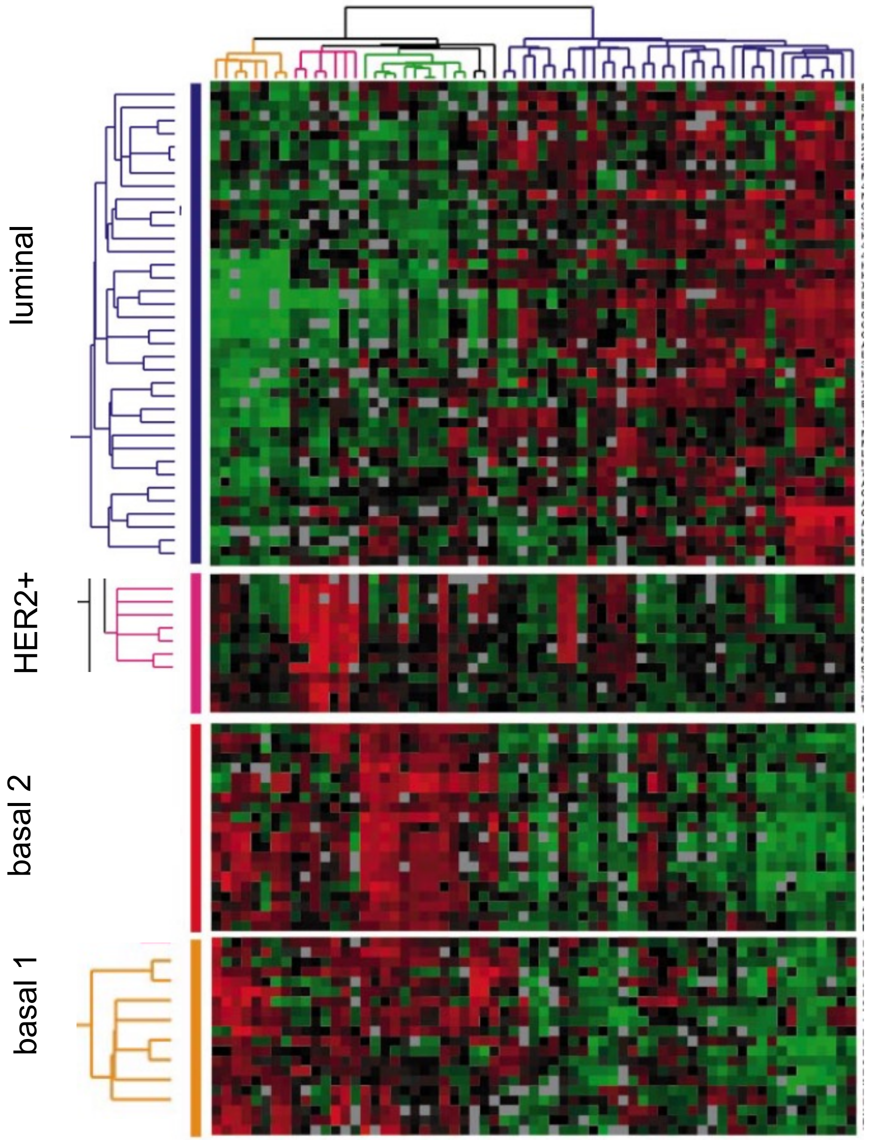
Clustering

- Finding subgroups (or clusters) in the data

- Ex: (unknown!) subgroups classifying different breast cancers

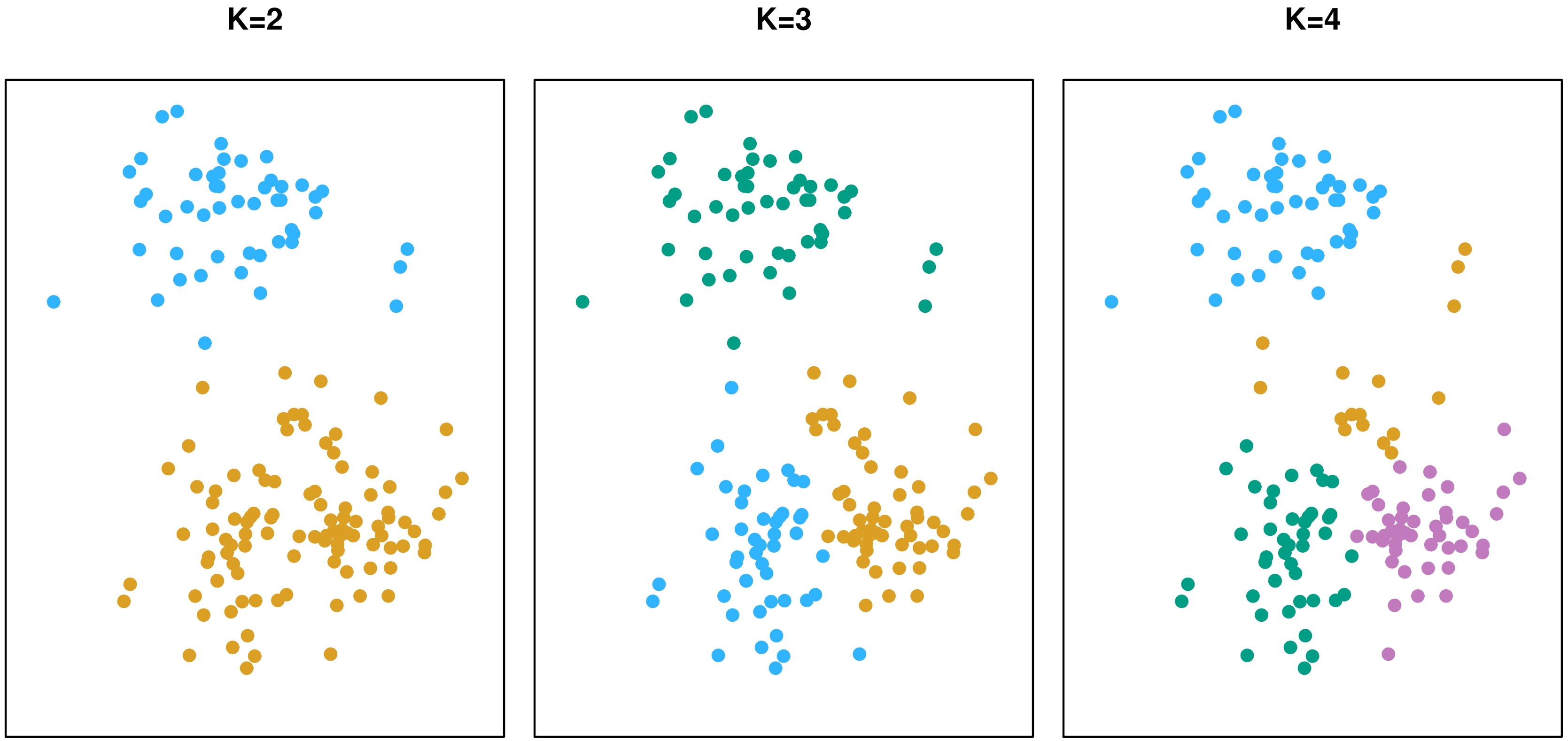
- Observations that are “similar” or “different”

- Problem = Define “similar” and “different”



K-means clustering

- Partitioning the data into K distinct, non-overlapping clusters
- K is chosen



K-means formalism

- Let *C₁*, ..., *C_K* be the *K* clusters

1. $C_1 \cup C_2 \cup \ldots \cup C_K = \{1, \ldots, n\}$ each observation belongs to at least one of the K clusters
2. $C_k \cap C_{k'} = \emptyset \text{ for all } k \neq k'$ the clusters are non- overlapping: no observation belongs to more than one cluster

- Idea: good clustering = within-cluster variation is as small as possible

$$\underset{C_1,\ldots,C_K}{\text{minimize}} \left\{ \sum_{k=1}^{K} W(C_k) \right\}$$

Squared Euclidian distance

$$W(C_k) = \frac{1}{|C_k|} \sum_{i,i' \in C_k} \sum_{j=1}^{p} (x_{ij} - x_{i'j})^2$$

K-means algorithm

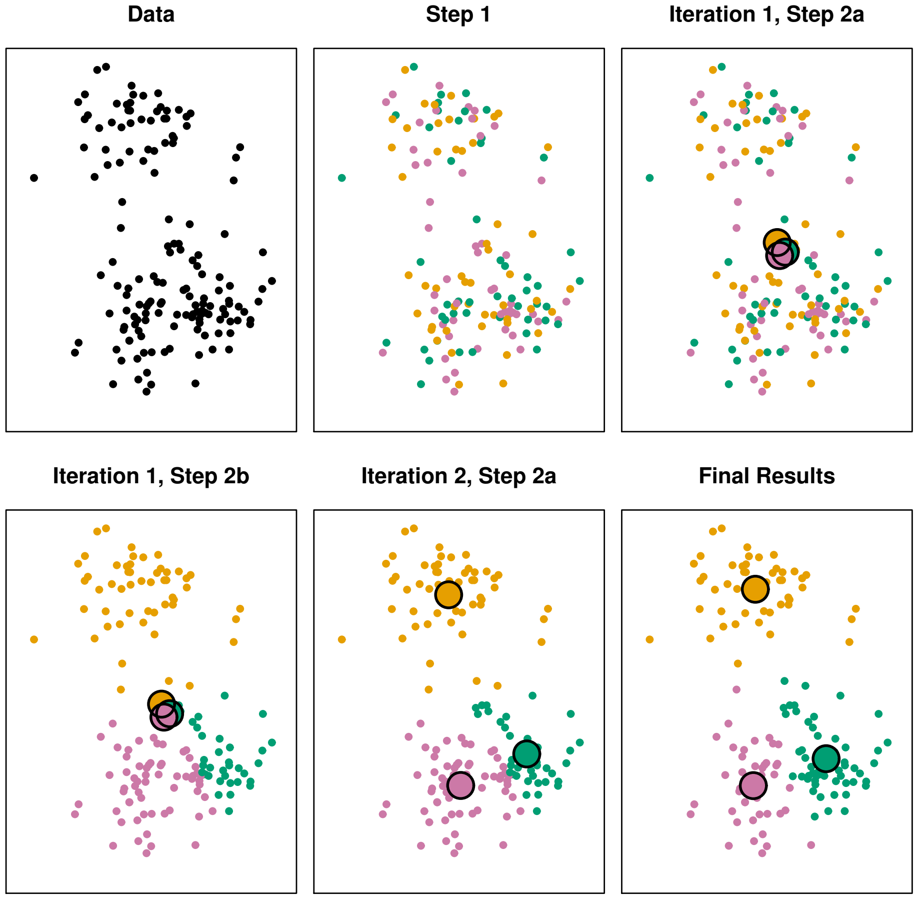
Algorithm 12.2 *K-Means Clustering*

1. Randomly assign a number, from 1 to K, to each of the observations. These serve as initial cluster assignments for the observations.

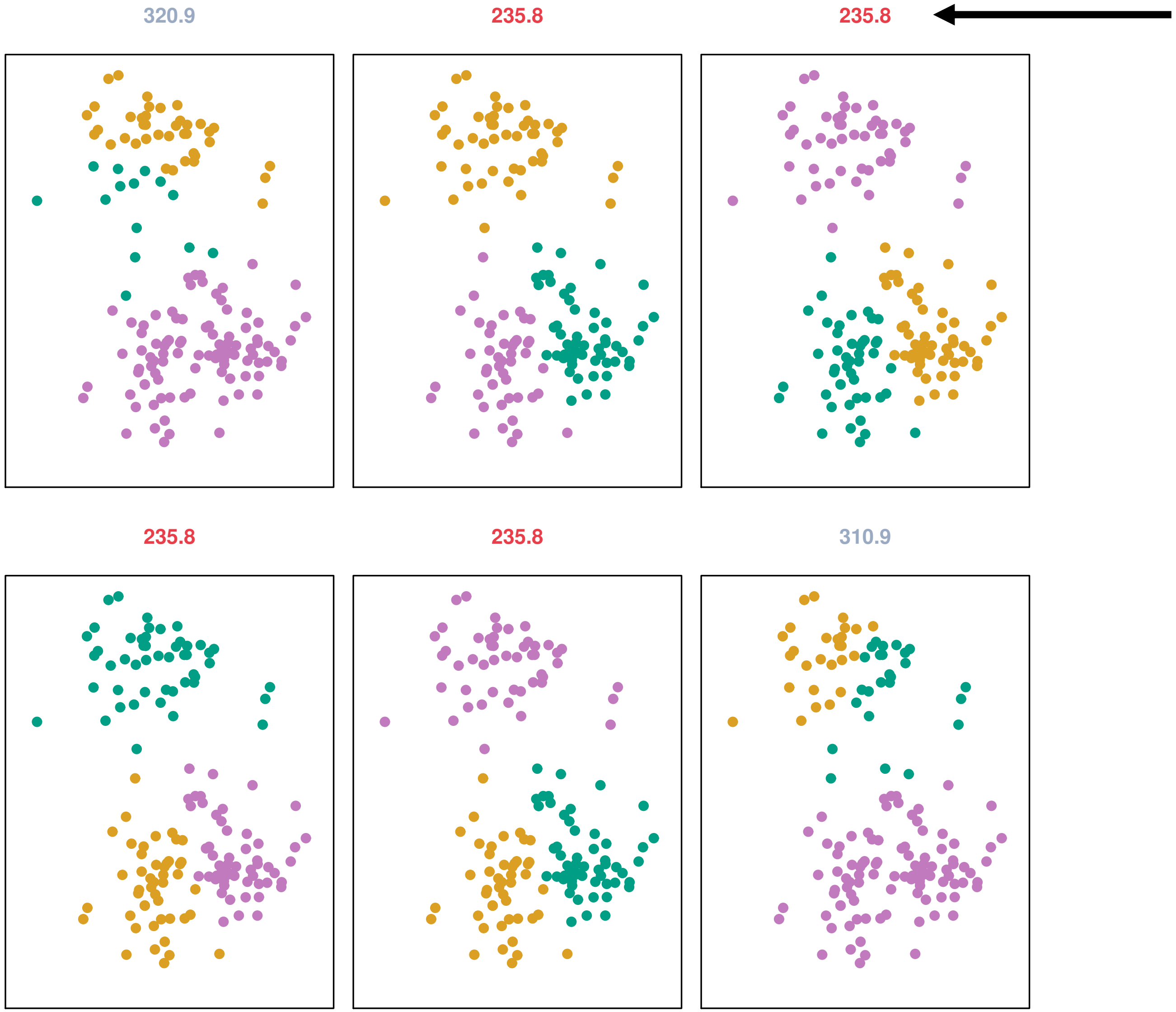
2. Iterate until the cluster assignments stop changing:

   (a) For each of the K clusters, compute the cluster *centroid*. The kth cluster centroid is the vector of the p feature means for the observations in the kth cluster.

   (b) Assign each observation to the cluster whose centroid is closest (where *closest* is defined using Euclidean distance).



Varying the initial random assignment



final value of the objective function

Hierarchical clustering

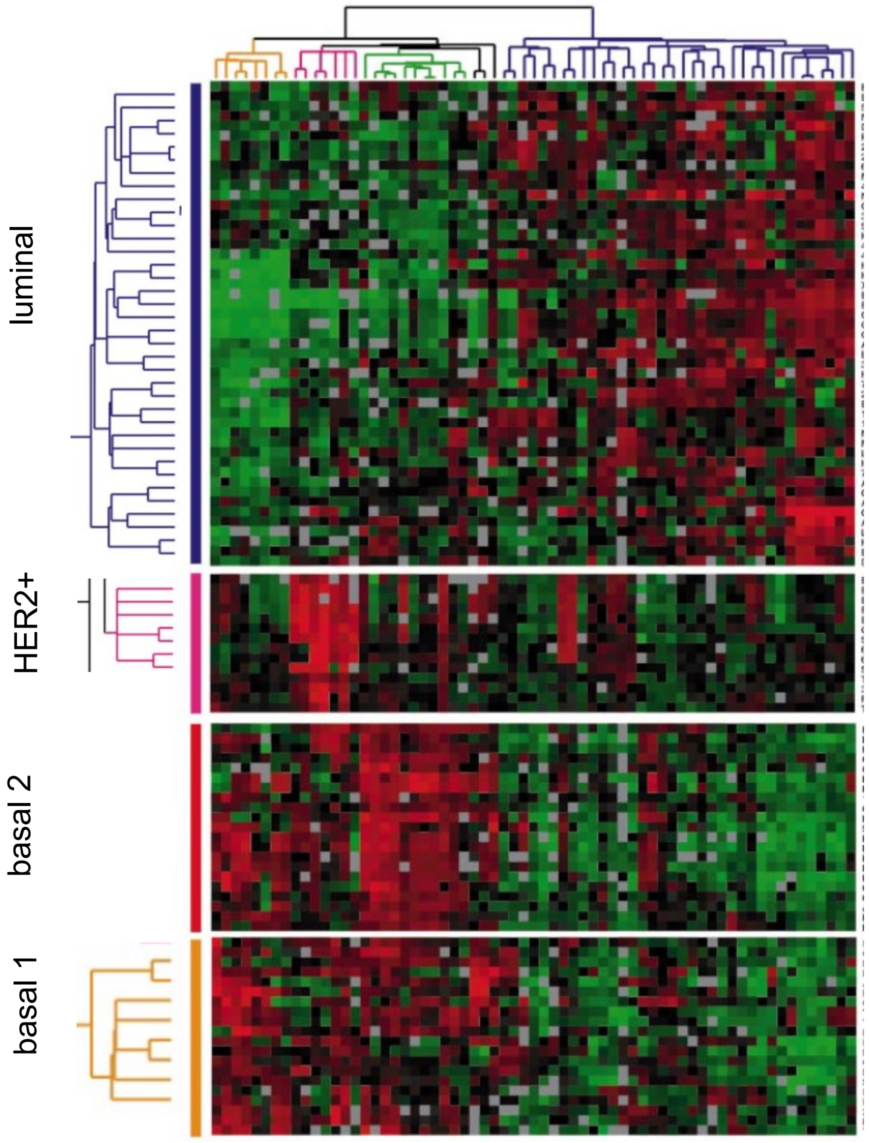
- Disadvantage of K-means : need to specify K

- Hierarchical clustering gives an interpretrable tree-based output: a dendrogram

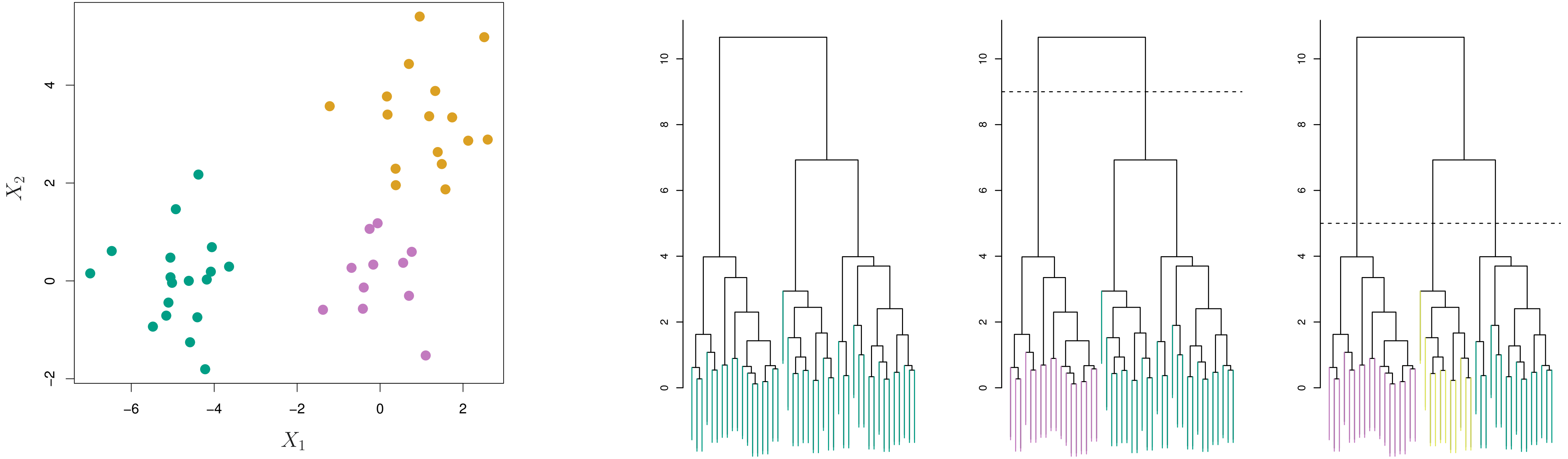
- Bottom-up = starting from the leaves

vs

- Top-down = starting from all data



Example: dendrograms



Hierarchical clustering algorithm

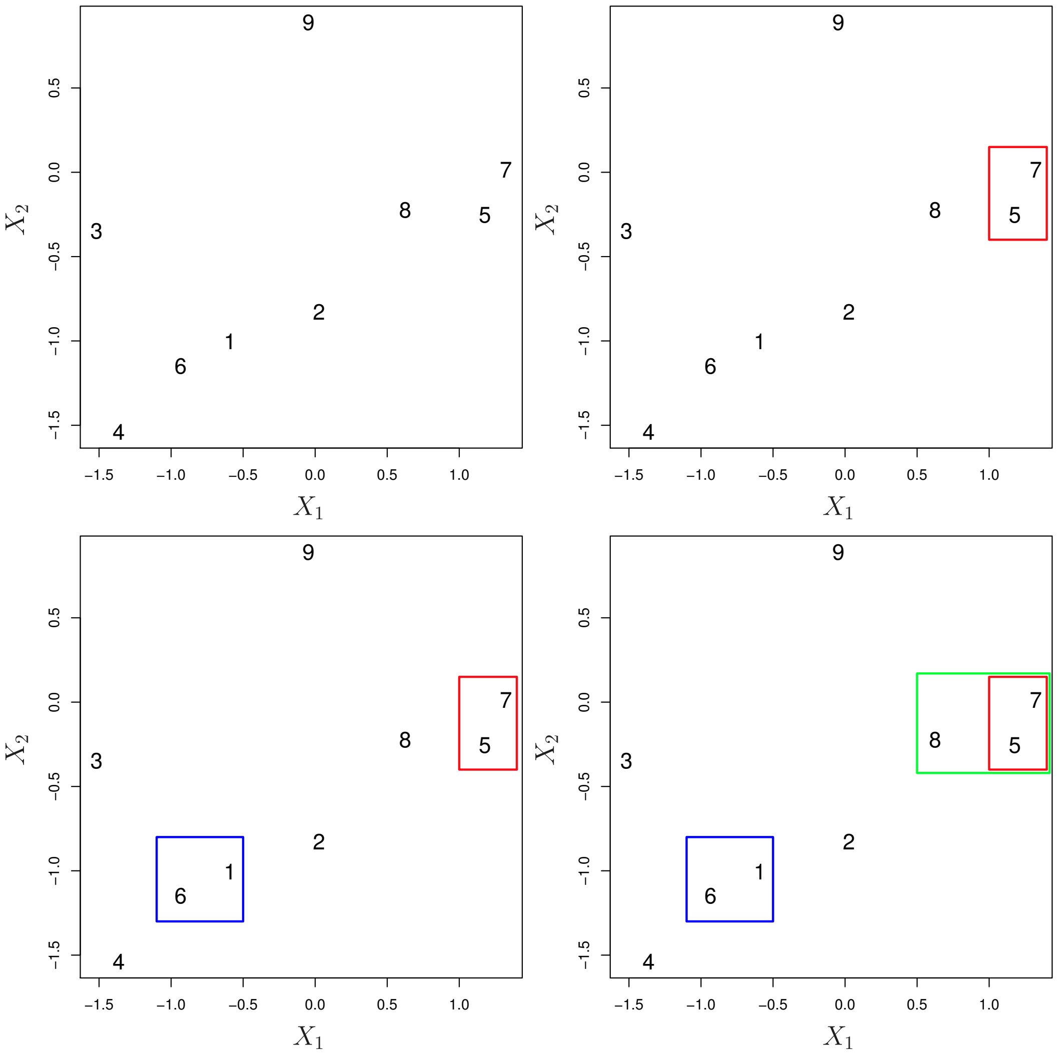
Algorithm 12.3 *Hierarchical Clustering*

1. Begin with n observations and a measure (such as Euclidean distance) of all the $\binom{n}{2} = n(n-1)/2$ pairwise dissimilarities. Treat each observation as its own cluster.

2. For $i = n, n-1, \ldots, 2$:

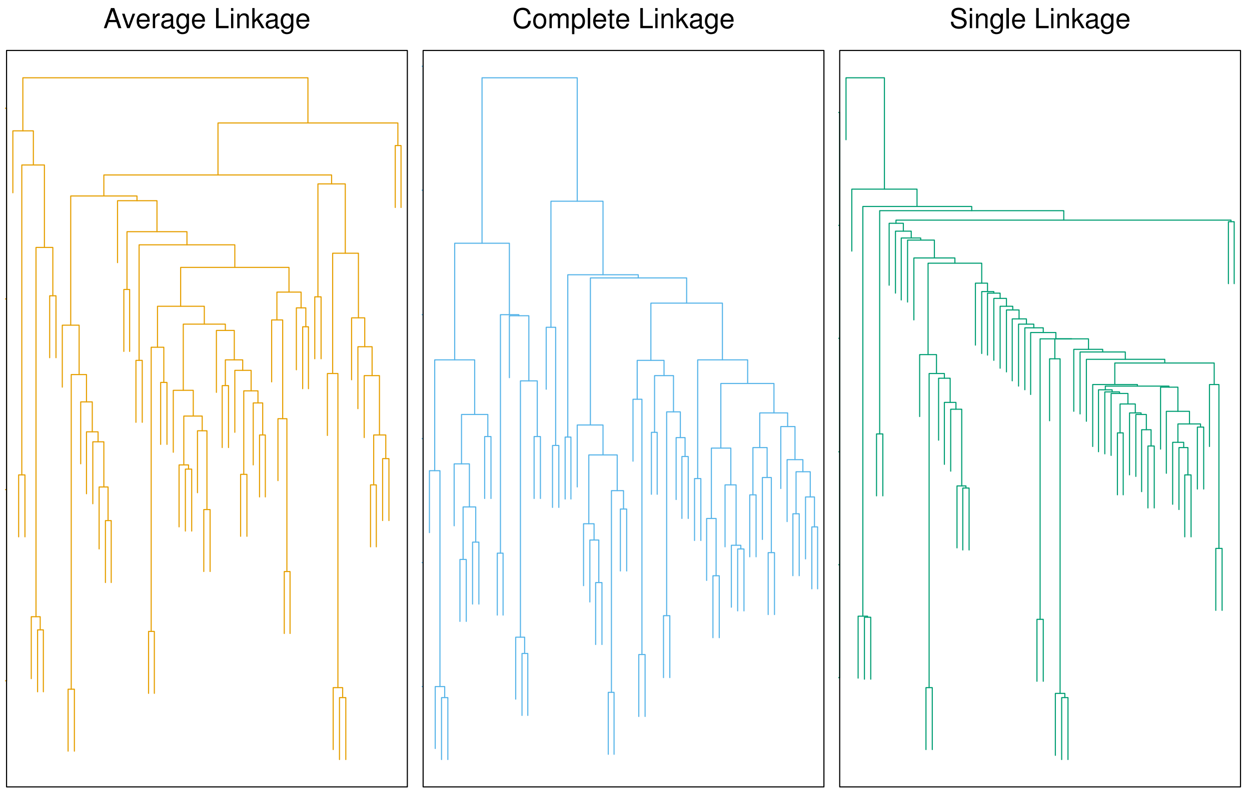
   (a) Examine all pairwise inter-cluster dissimilarities among the i clusters and identify the pair of clusters that are least dissimilar (that is, most similar). Fuse these two clusters. The dissimilarity between these two clusters indicates the height in the dendrogram at which the fusion should be placed.

   (b) Compute the new pairwise inter-cluster dissimilarities among the $i-1$ remaining clusters.



Possible linkages (=dissimilarities between groups)

Linkage	*Description*
Complete	Maximal intercluster dissimilarity. Compute all pairwise dissimilarities between the observations in cluster A and the observations in cluster B, and record the *largest* of these dissimilarities.
Single	Minimal intercluster dissimilarity. Compute all pairwise dissimilarities between the observations in cluster A and the observations in cluster B, and record the *smallest* of these dissimilarities. Single linkage can result in extended, trailing clusters in which single observations are fused one-at-a-time.
Average	Mean intercluster dissimilarity. Compute all pairwise dissimilarities between the observations in cluster A and the observations in cluster B, and record the *average* of these dissimilarities.
Centroid	Dissimilarity between the centroid for cluster A (a mean vector of length p) and the centroid for cluster B. Centroid linkage can result in undesirable *inversions*.



Average Linkage

Complete Linkage

Single Linkage

References

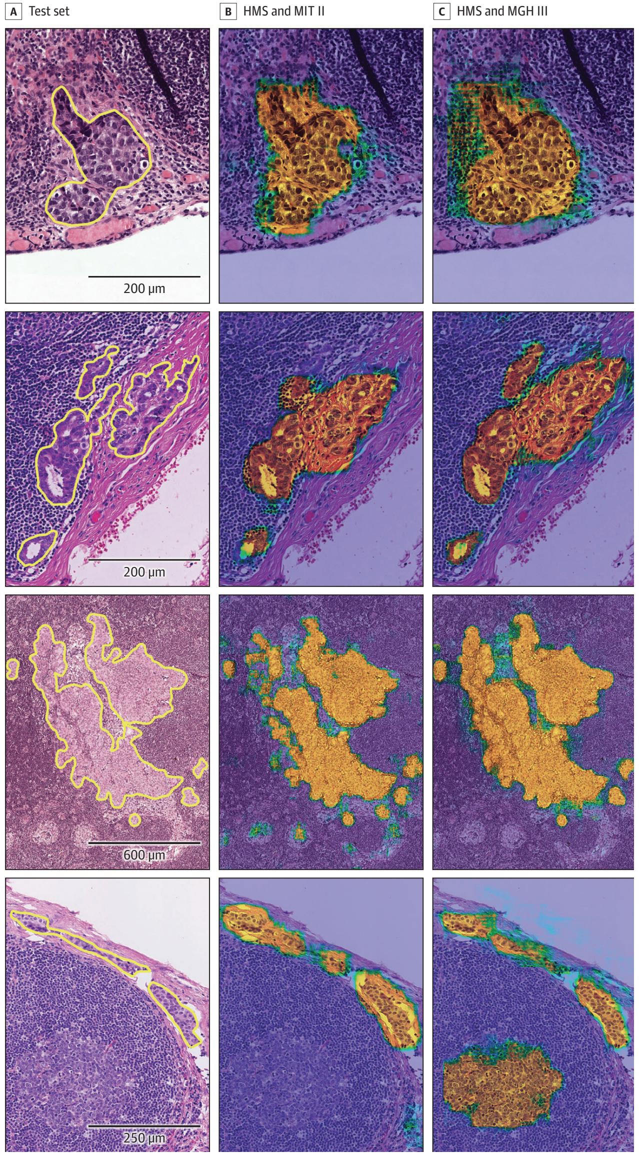
- **Textbooks and theory**
  - *An introduction to statistical learning*. James, Witten, Hastie, Tibshirani
    Multiple illustrations were taken from this book.
    https://www.statlearning.com/
  - *Applied predictive modeling*. Kuhn and Johnson
    http://appliedpredictivemodeling.com/
  - *Scikit-learn's user guide*
    https://scikit-learn.org/stable/user_guide.html

- **Programming**
  - R (basic, no ML): *R for data science.* H. Wickham
    https://r4ds.hadley.nz/
  - R ML: *Tidy modeling with R.* M. Kuhn and J. Silge
    https://www.tmwr.org/
  - python: scikit-learn
    https://scikit-learn.org/stable/

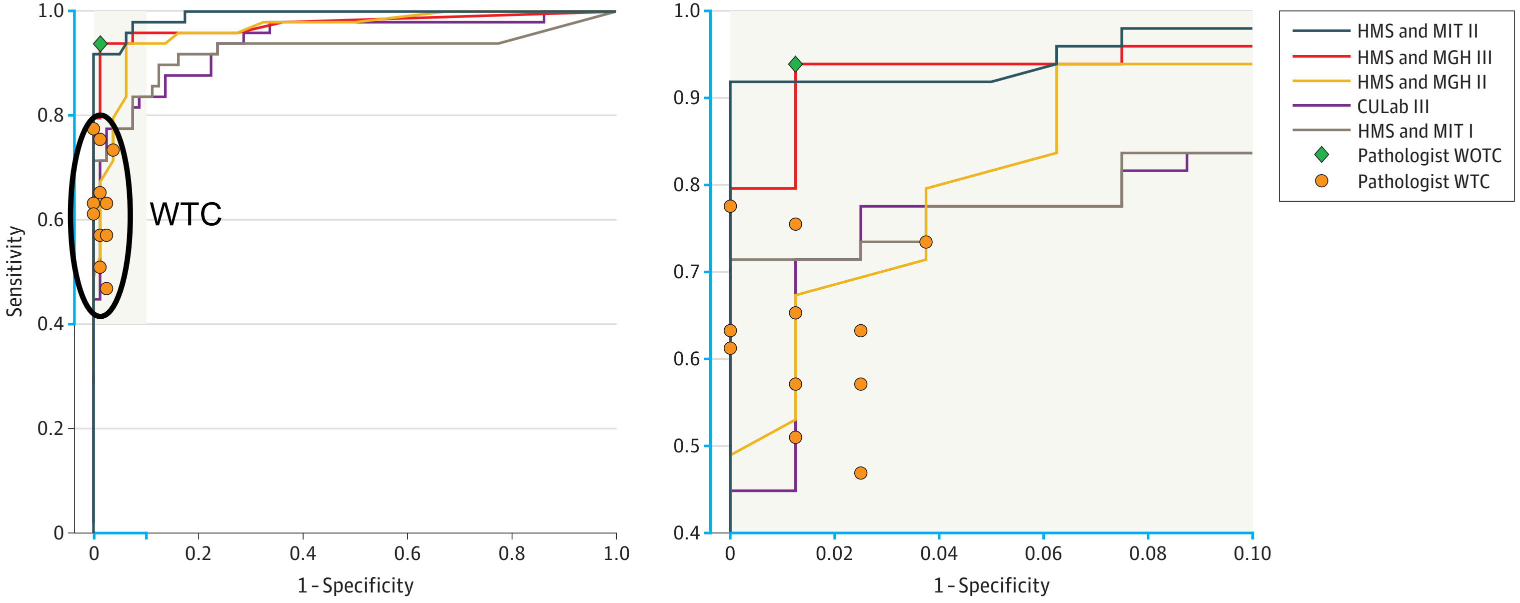
- **Youtube**
  - Science étonnante
  - 3 Blue 1 brown

Additional

Detection of lymph node metastases from histological images

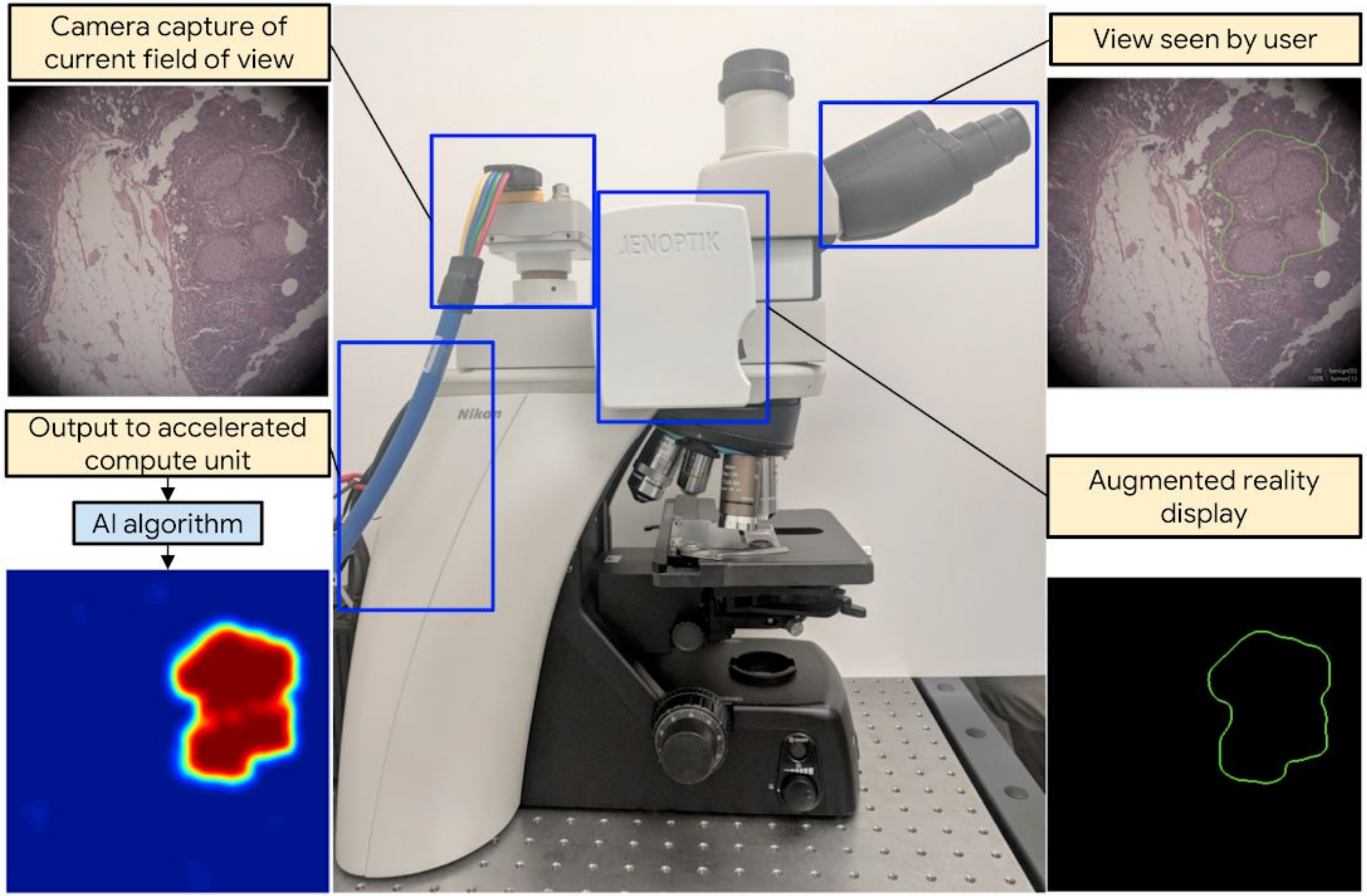


- One pathology slide = several gigapixels
- Best algorithms of the challenge = Deep Learning
- Same performances as pathologists without time constraint, but significatively better than 11 pathologists with constraint (WTC)



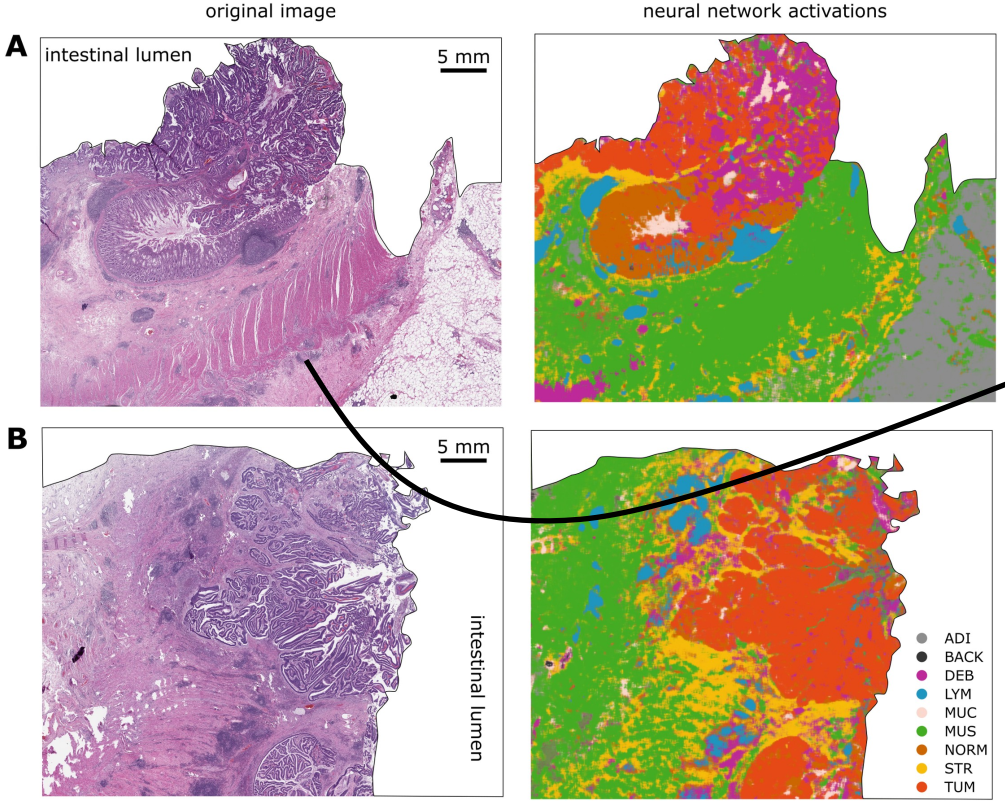
Bejnordi et al., Diagnostic Assessment of Deep Learning Algorithms for Detection of Lymph Node Metastases in Women With Breast Cancer, JAMA, 2017

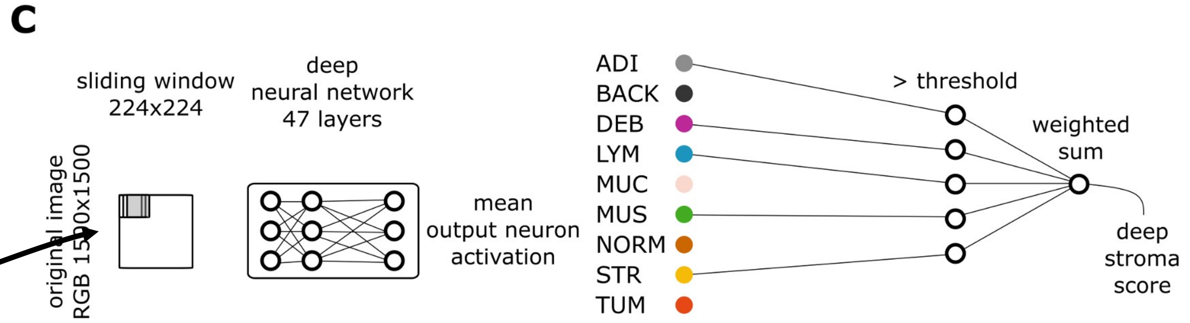
Microscope 2.0


Camera capture of current field of view
Output to accelerated compute unit
AI algorithm
JENOPTIK
Nikon
View seen by user
Augmented reality display

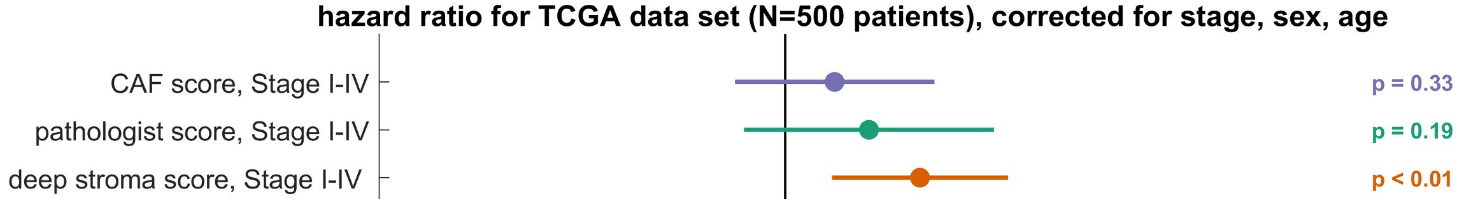
Chen et al. (Google AI Healthcare), Microscope 2.0: An Augmented Reality Microscope with Real-time Artificial Intelligence Integration, arXiv, 2018

Quantitative analysis of histopathological slides in CRC





- 100,000 patches of histological slides
- Stroma
- 94% classification accuracy on test data set



- « Deep stroma score » is a predictive factor of survival independent of TNM stage (current state of the art)

Kather et al., Predicting survival from colorectal cancer histology slides using deep learning: A retrospective multicenter study, PLoS Med, 2019

Prediction of response to immune-checkpoint inhibition

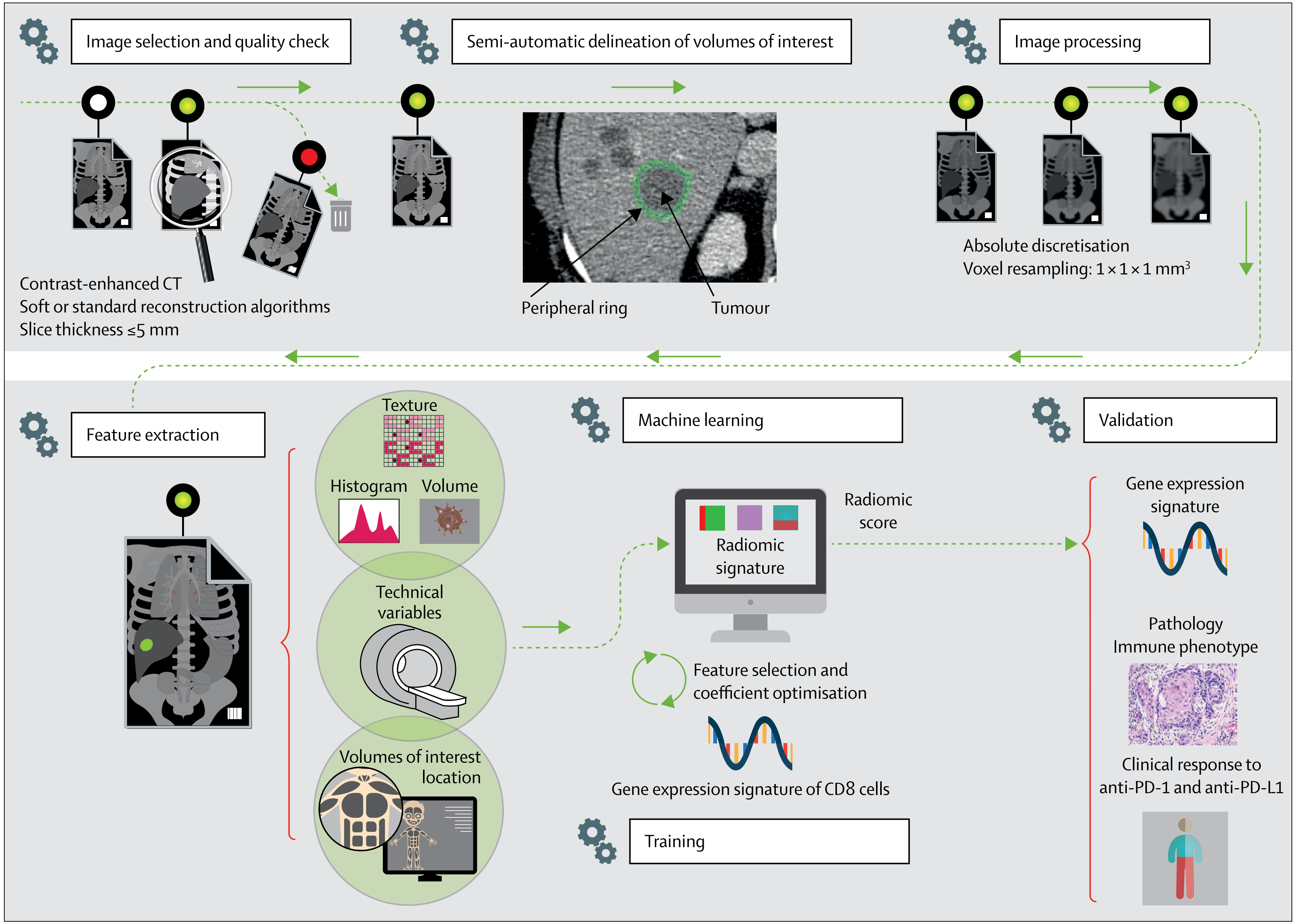


Image selection and quality check
Semi-automatic delineation of volumes of interest
Image processing
Contrast-enhanced CT
Soft or standard reconstruction algorithms
Slice thickness ≤5 mm
Peripheral ring
Tumour
Absolute discretisation
Voxel resampling: 1 × 1 × 1 mm³
Feature extraction
Texture
Histogram
Volume
Technical variables
Volumes of interest location
Machine learning
Radiomic signature
Radiomic score
Feature selection and coefficient optimisation
Gene expression signature of CD8 cells
Training
Validation
Gene expression signature
Pathology
Immune phenotype
Clinical response to anti-PD-1 and anti-PD-L1

Sun et al., Lancet Oncol, 2018

Prediction of response to immune-checkpoint inhibition

but…..

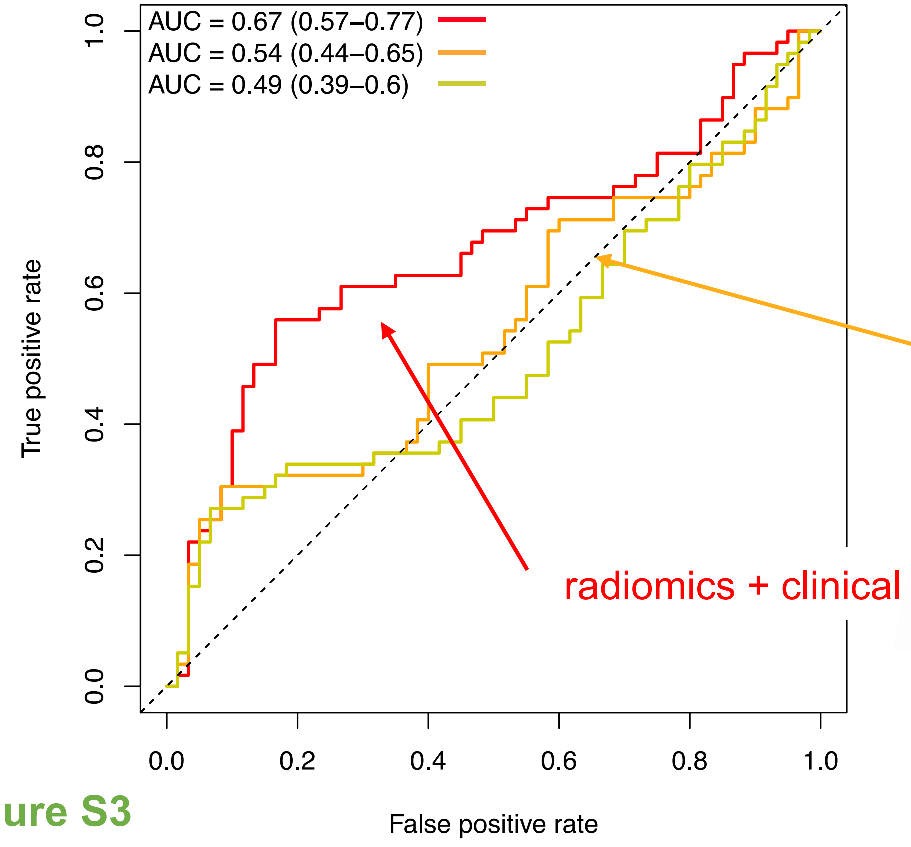
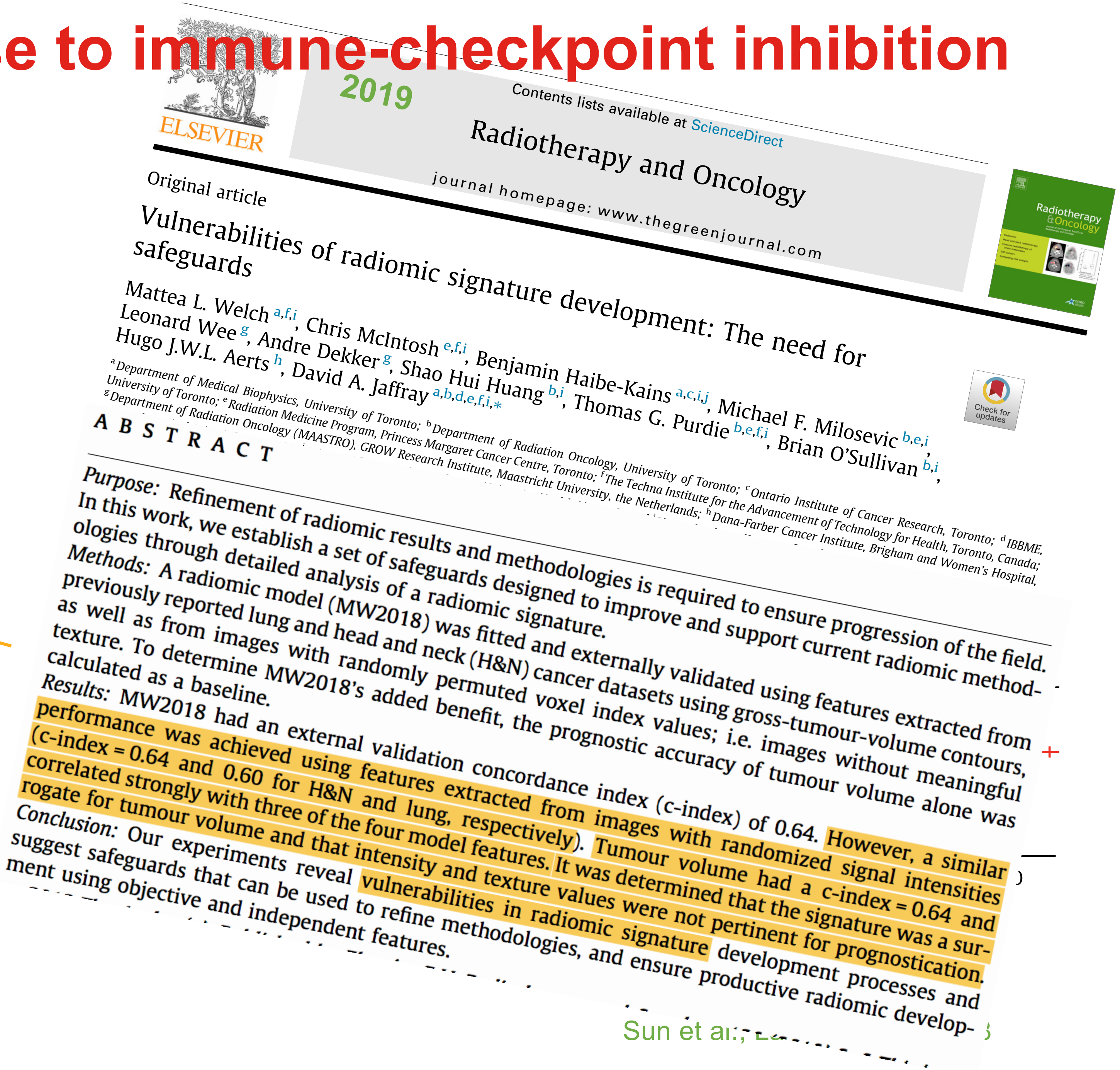


Figure S3



2019

ELSEVIER

Contents lists available at ScienceDirect

Radiotherapy and Oncology

journal homepage: www.thegreenjournal.com

Original article

Vulnerabilities of radiomic signature development: The need for safeguards

Mattea L. Welch [a,f,i], Chris McIntosh [e,f,i], Benjamin Haibe-Kains [a,c,i,j], Michael F. Milosevic [b,e,i], Leonard Wee [g], Andre Dekker [g], Shao Hui Huang [b,i], Thomas G. Purdie [b,e,f,i], Brian O'Sullivan [b,i], Hugo J.W.L. Aerts [h], David A. Jaffray [a,b,d,e,f,i,*]

[a] Department of Medical Biophysics, University of Toronto; [b] Department of Radiation Oncology, University of Toronto; [c] Ontario Institute of Cancer Research, Toronto; [d] IBBME, University of Toronto; [e] Radiation Medicine Program, Princess Margaret Cancer Centre, Toronto; [f] The Techna Institute for the Advancement of Technology for Health, Toronto, Canada; [g] Department of Radiation Oncology (MAASTRO), GROW Research Institute, Maastricht University, the Netherlands; [h] Dana-Farber Cancer Institute, Brigham and Women's Hospital,

Check for updates

A B S T R A C T

Purpose: Refinement of radiomic results and methodologies is required to ensure progression of the field. In this work, we establish a set of safeguards designed to improve and support current radiomic methodologies through detailed analysis of a radiomic signature.

Methods: A radiomic model (MW2018) was fitted and externally validated using features extracted from previously reported lung and head and neck (H&N) cancer datasets using gross-tumour-volume contours, as well as from images with randomly permuted voxel index values; i.e. images without meaningful texture. To determine MW2018's added benefit, the prognostic accuracy of tumour volume alone was calculated as a baseline.

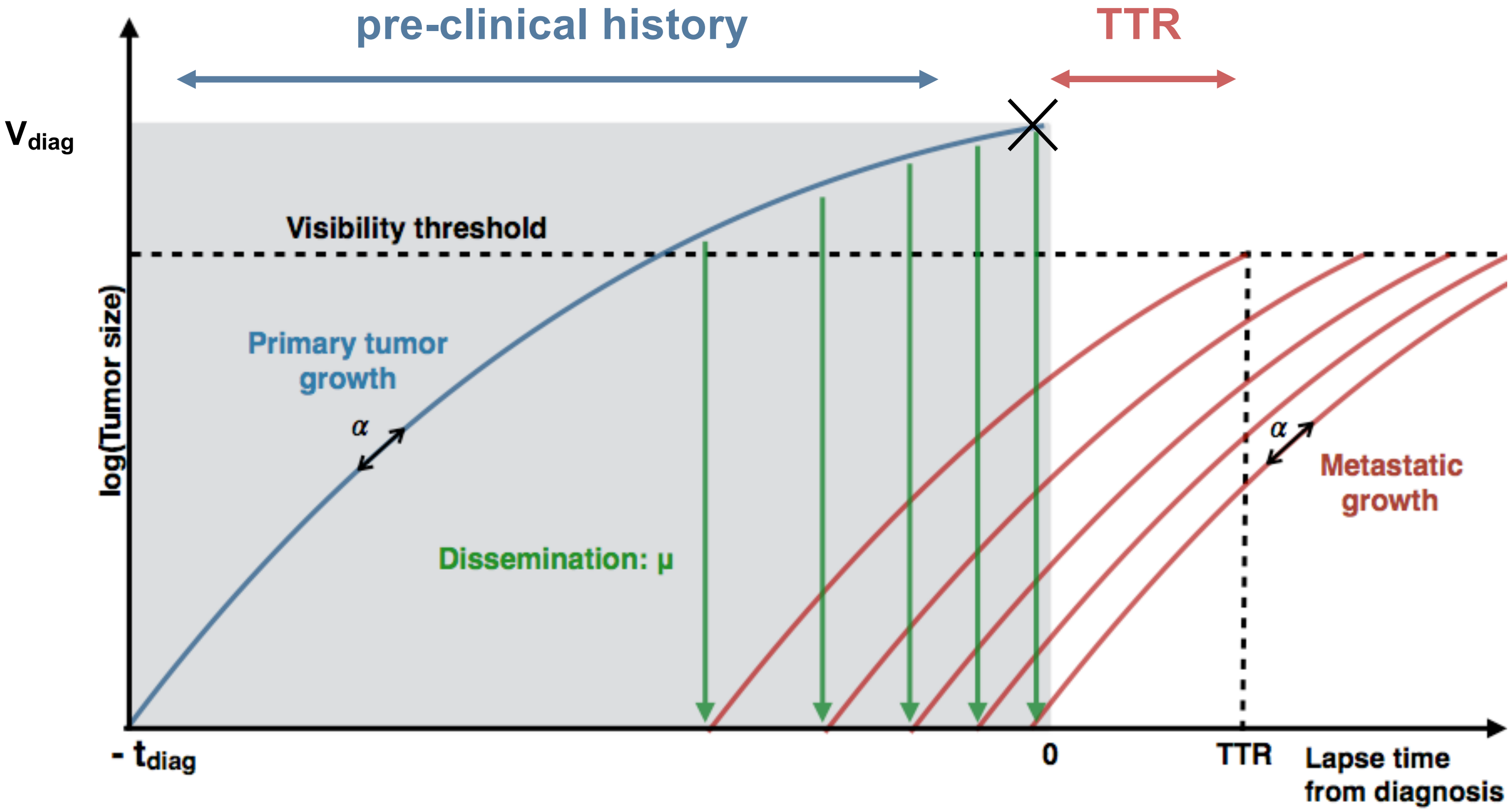
Results: MW2018 had an external validation concordance index (c-index) of 0.64. However, a similar performance was achieved using features extracted from images with randomized signal intensities (c-index = 0.64 and 0.60 for H&N and lung, respectively). Tumour volume had a c-index = 0.64 and correlated strongly with three of the four model features. It was determined that the signature was a surrogate for tumour volume and that intensity and texture values were not pertinent for prognostication.

Conclusion: Our experiments reveal vulnerabilities in radiomic signature development processes and suggest safeguards that can be used to refine methodologies, and ensure productive radiomic development using objective and independent features.

Sun et al.;

Mechanistic modeling of metastatic relapse

Mechanistic modeling of time to relapse



- Number of metastases with size larger than the visible size V_{vis}

$$N_{vis}(t) = \int_{V_{vis}}^{+\infty} \rho(t,v)dv$$
$$= \int_0^{t-\tau_{vis}} d(V_p(t))dt$$

τ_{vis} = time to reach V_{vis}

- Time to relapse (TTR) = time elapsed from diagnosis to the appearance of a first visible metastasis

$$TTR = \inf\left\{t > 0 : N_{vis}(t_{diag} + t) \geq 1\right\}$$

- Parameter β fixed such that $V_\infty = e^{\frac{\alpha}{\beta}} = 10^{12}$ cells

Mixed-effects statistical model

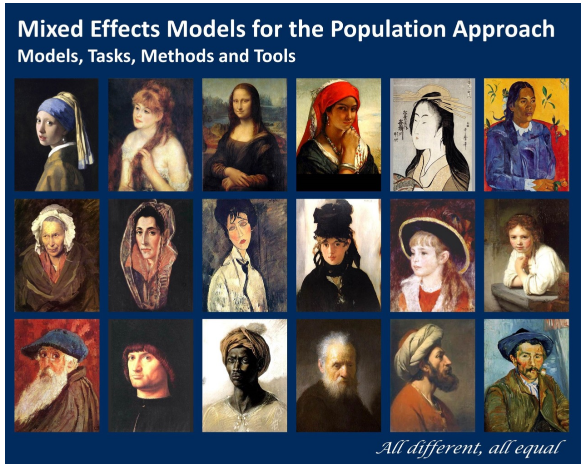
$$\ln\left(T^i\right) = \ln\left(TTR\left(V^i_{diag};\alpha^i,\mu^i\right)\right) + \varepsilon^i, \quad \varepsilon^i \sim \mathcal{N}(0,\sigma^2)$$

(Observation model)

$$S\left(t|\alpha^i,\mu^i\right) = \mathbb{P}\left(T^i > t|\alpha^i,\mu^i\right)$$

Survival function to account for censoring in the likelihood

$$\boxed{\begin{aligned} \ln\left(\alpha^i\right) &= \ln\left(\alpha_{pop}\right) + \eta^i_\alpha, \quad \eta^i_\alpha \sim \mathcal{N}(0,\omega^2_\alpha) \\ \ln\left(\mu^i\right) &= \ln\left(\mu_{pop}\right) + \eta^i_\mu, \quad \eta^i_\mu \sim \mathcal{N}(0,\omega^2_\mu) \end{aligned}}$$

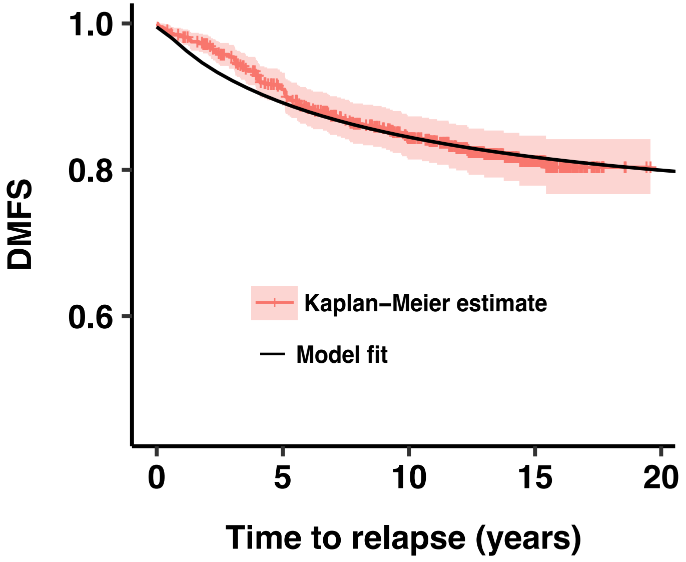

Mixed Effects Models for the Population Approach
Models, Tasks, Methods and Tools
All different, all equal

Lavielle, CRC press, 2014

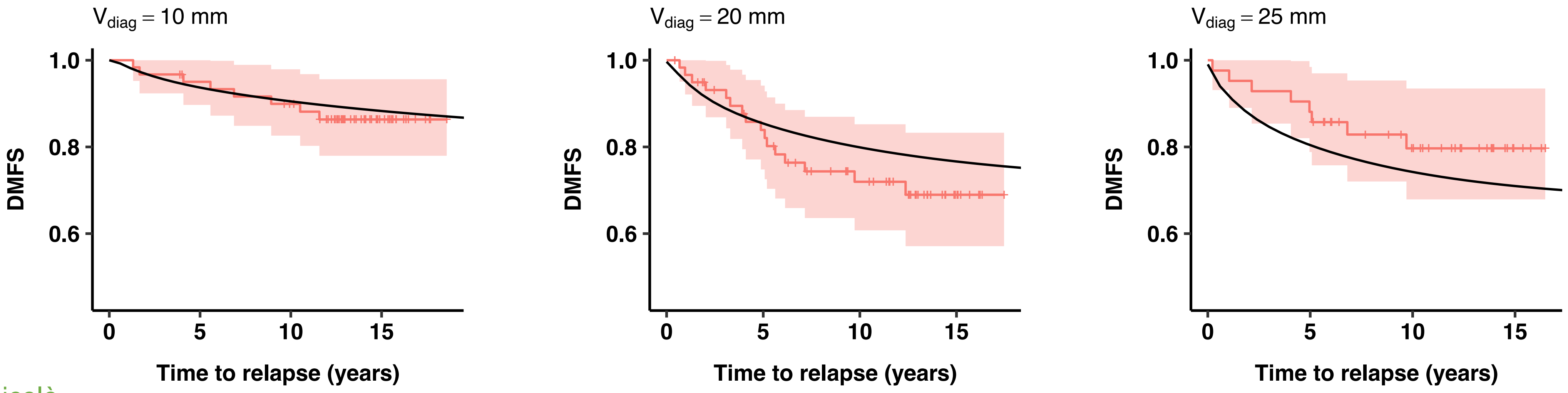
Likelihood maximization performed using the SAEM algorithm implemented in the *saemix* R package

Comets, Lavenu, Lavielle, J Stat Softw, 2017

Descriptive power: fit to the data



Parameter	Estimate	r.s.e. (%)
$\log \alpha_{pop}$	-6.34	12.6
$\log \mu_{pop}$	-26.8	3.68
σ	0.542	28.4
ω_α	3.37	36.4
ω_μ	3.78	15.9



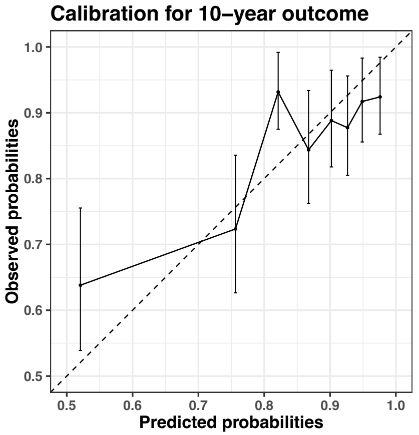
PhD of Chiara Nicolò

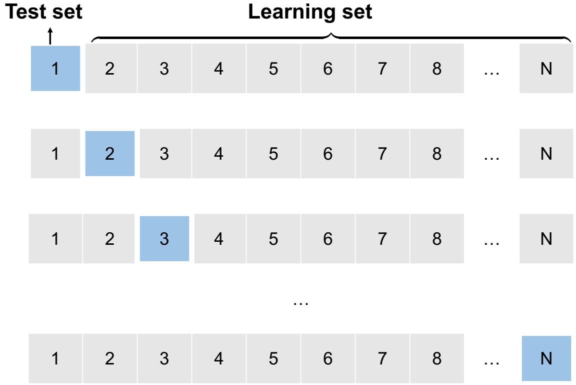
Predictive power: covariates

$$\ln\left(\mu^i\right) = \ln\left(\mu_{pop}\right) + \beta_\mu^T \mathbf{x}_\mu^\mathbf{i} + \eta_\mu^i, \quad \eta_\mu^i \sim \mathcal{N}(0, \omega_\mu^2)$$

$$\ln\left(\alpha^i\right) = \ln\left(\alpha_{pop}\right) + \beta_\alpha^T \mathbf{x}_\alpha^\mathbf{i} + \eta_\alpha^i, \quad \eta_\alpha^i \sim \mathcal{N}(0, \omega_\alpha^2)$$

Parameter	Estimate	r.s.e. (%)	p-value
$\log \alpha_{pop}$	-8.883	10.151	
$\beta_{\text{Ki67},\alpha}$	0.086	27.376	$2.59 \cdot 10^{-4}$
$\beta_{\text{HER2},\alpha}$	0.029	42.833	0.020
$\beta_{\text{CD44},\alpha}$	0.011	60.816	0.1
$\beta_{\text{TRIO},\alpha}$	0.016	58.119	0.085
$\log \mu_{pop}$	-26.342	3.696	
$\beta_{\text{EGFR},\mu}$	0.039	47.527	0.035
σ	0.606	23.104	
ω_α	2.062	22.715	
ω_μ	3.563	16.759	





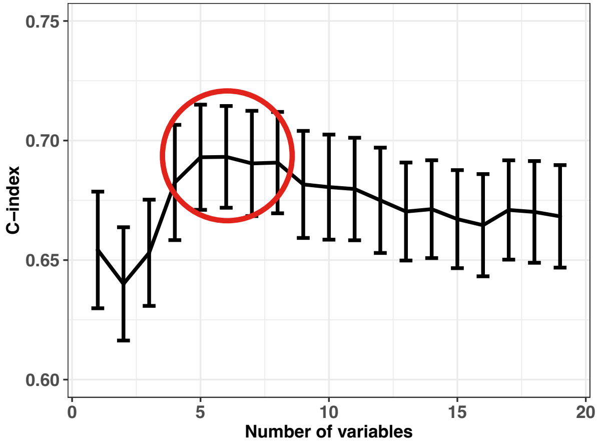
c-index = 0.67
(10-folds cross-validation)

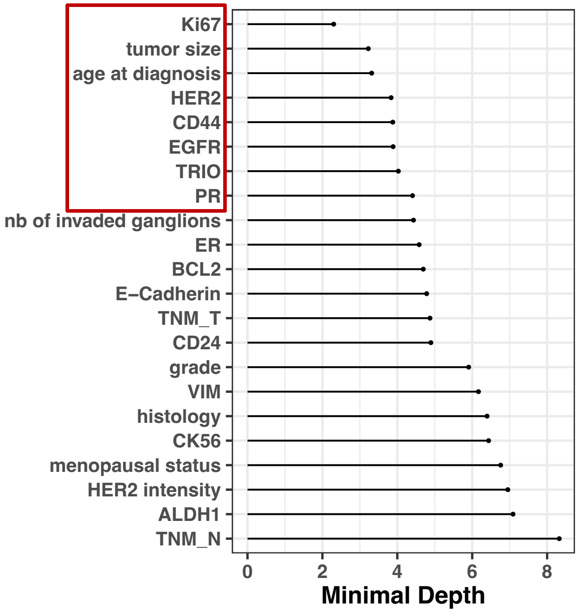
Patient ID	Tumor size (mm)	Ki67	HER2	CD44	TRIO	EGFR	Observed TTR (cens)	Predicted TTR	Prediction error (days)
255	25	1	60	90	60	0	1812 (1)	1609	203
47	20	32	100	0	0	50	739 (1)	447	292
143	18	60	0	50	0	0	2798 (1)	434	2364
12	10	20	0	23	0	0	5970 (0)	$+\infty$	-

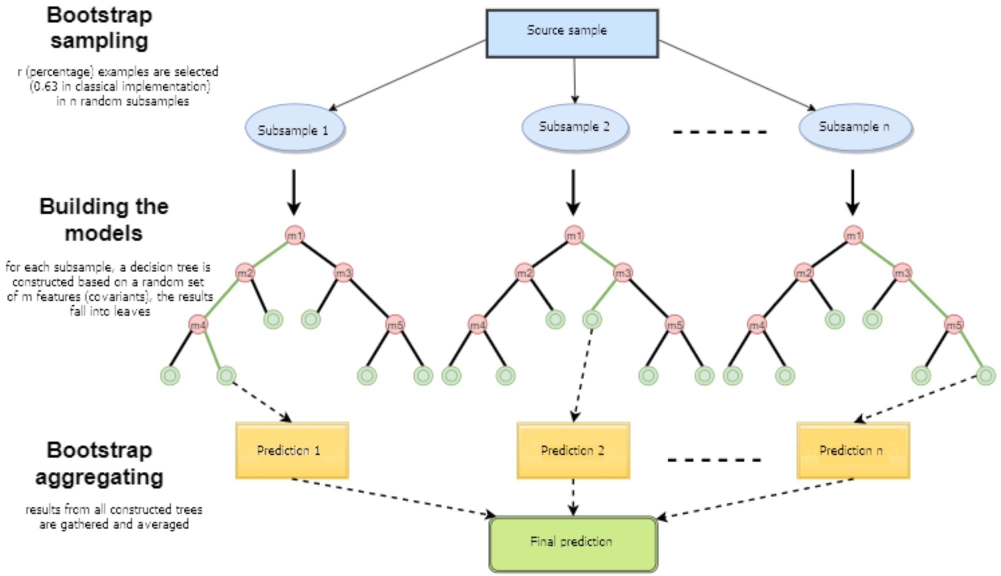
PhD of Chiara Nicolò

Random survival forests

c-index = 0.69
(cross-validation)






Bootstrap sampling

r (percentage) examples are selected (0.63 in classical implementation) in n random subsamples

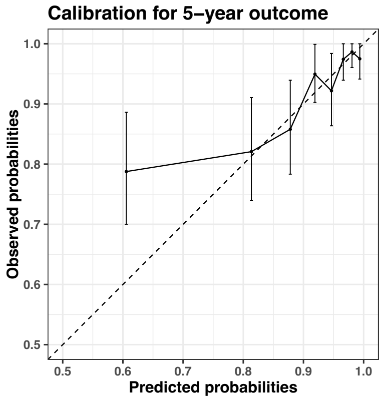
Building the models

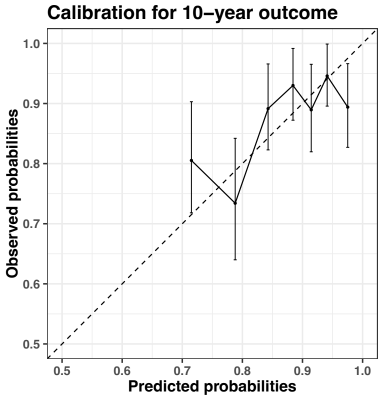
for each subsample, a decision tree is constructed based on a random set of m features (covariants), the results fall into leaves

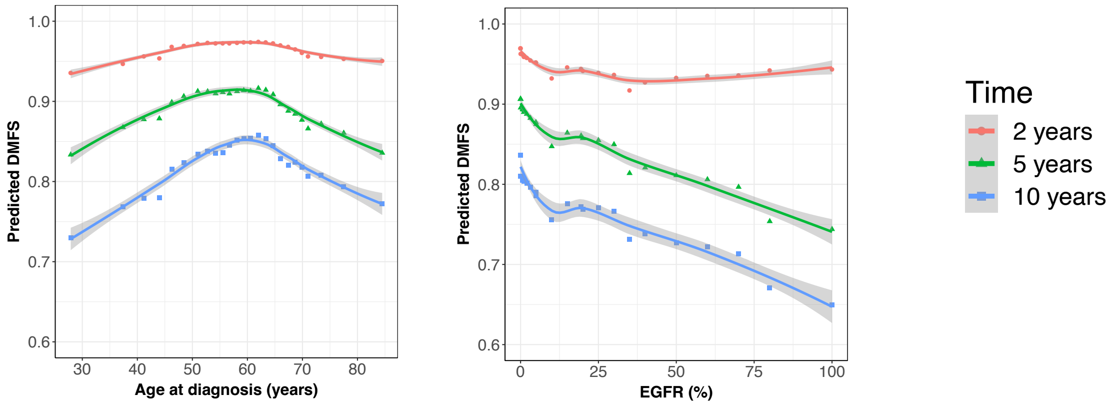
Bootstrap aggregating

results from all constructed trees are gathered and averaged

M. Dmitrievsky, https://www.mql5.com/en/articles/3856







⇒ nonlinear effect of covariates and non proportional hazard

Ishwaran et al., Ann Appl Stat, 2008

PhD of Chiara Nicolò

Comparison of predictive metrics

5 years metastatic-free survival

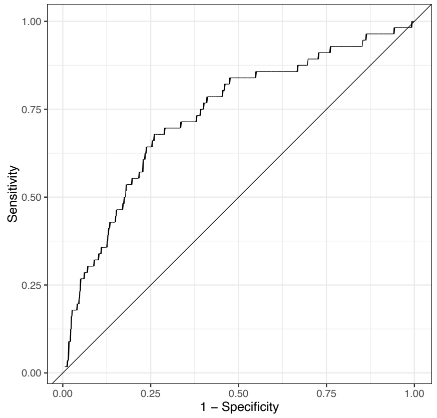
	AUROC	Accuracy	PPV	NPV
RSF	0.75	0.90	0.71	0.71
Mechanistic model	0.73	0.90	0.72	0.70
Cox	0.75	0.91	0.77	0.71

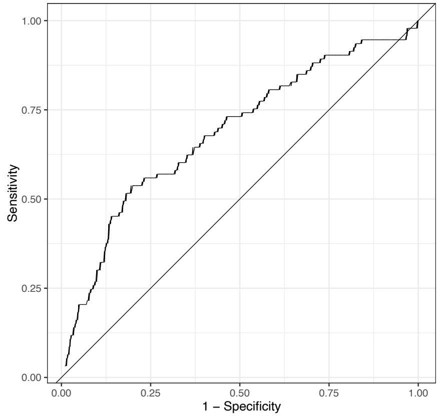
10 years metastatic-free survival

	AUROC	Accuracy	PPV	NPV
RSF	0.69	0.82	0.68	0.66
Mechanistic model	0.69	0.81	0.71	0.64
Cox	0.71	0.82	0.70	0.68

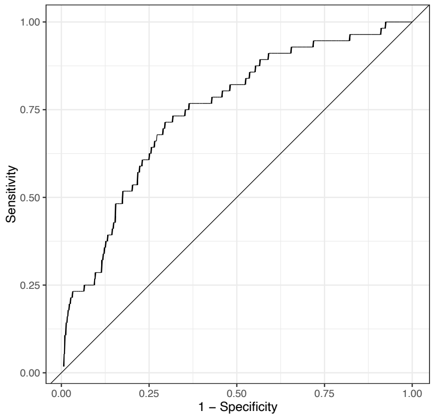
other tested ML models (support vector machine, k-nearest neighbors, gradient boosting) had similar or worse performances

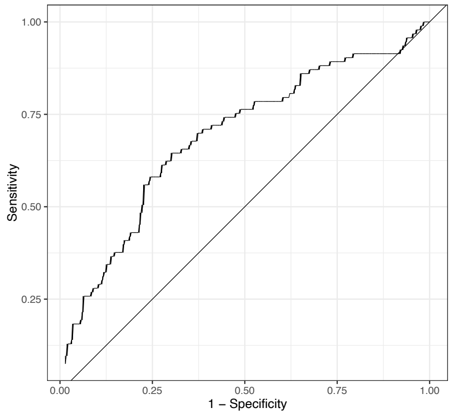
Mechanistic

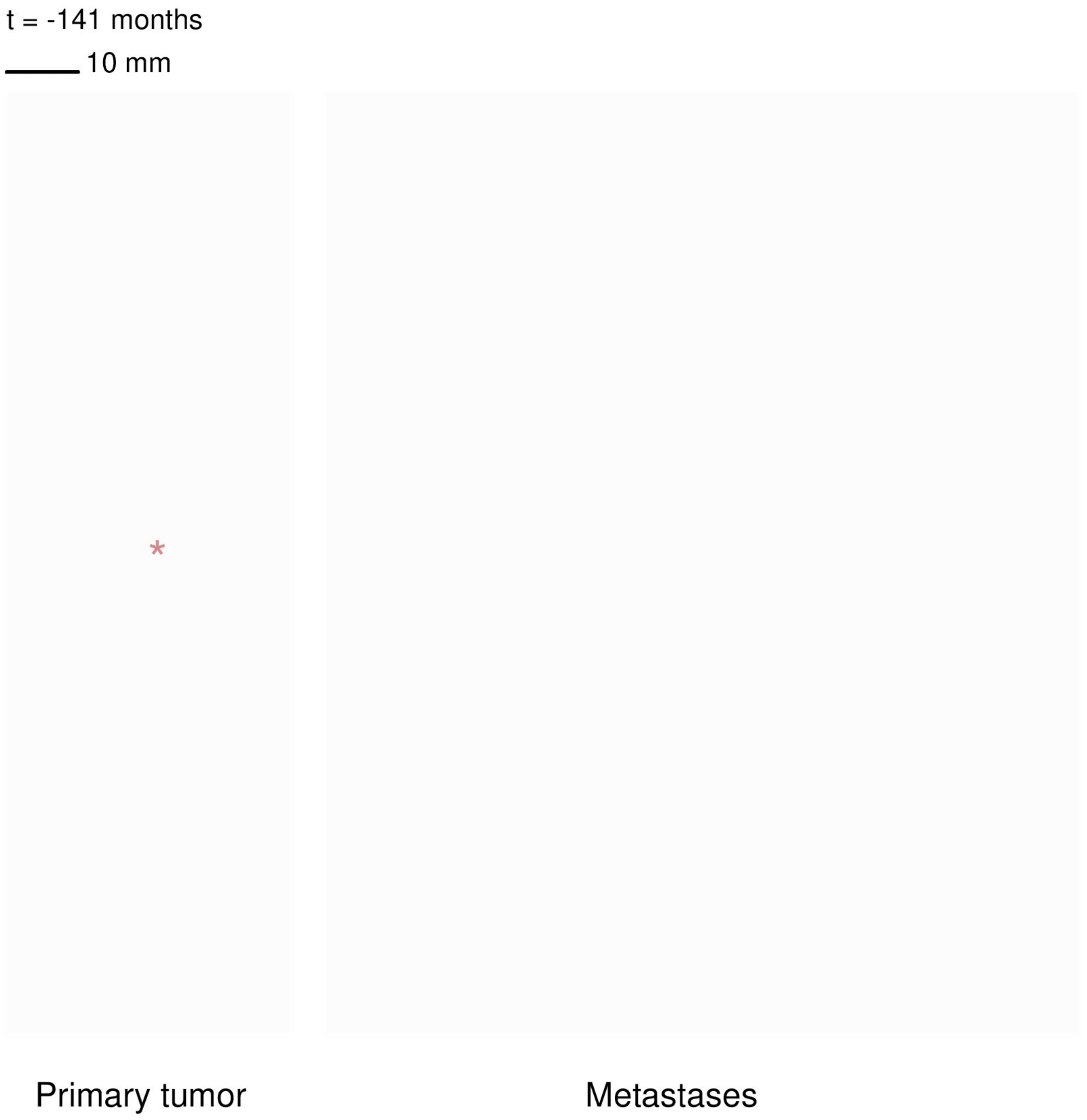


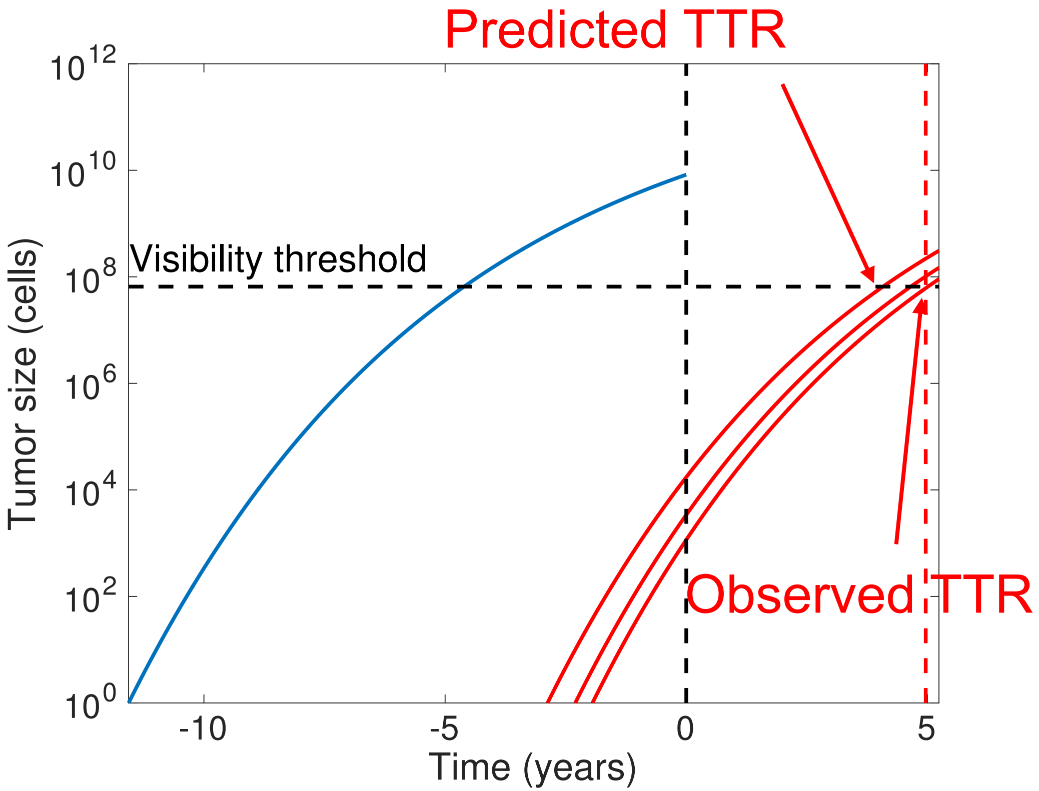


RSF






t = -141 months
10 mm
Primary tumor
Metastases


Predicted TTR
Observed TTR
Visibility threshold
Tumor size (cells)
Time (years)
10^12
10^10
10^8
10^6
10^4
10^2
10^0
-10
-5
0
5

Conclusions and perspectives

- Similar predictive performances of Cox regression (c-index 0.67 - 0.72), random survival forest (c-index 0.67-0.71) and a novel mechanistic model (c-index 0.63 - 0.70) for pure prediction

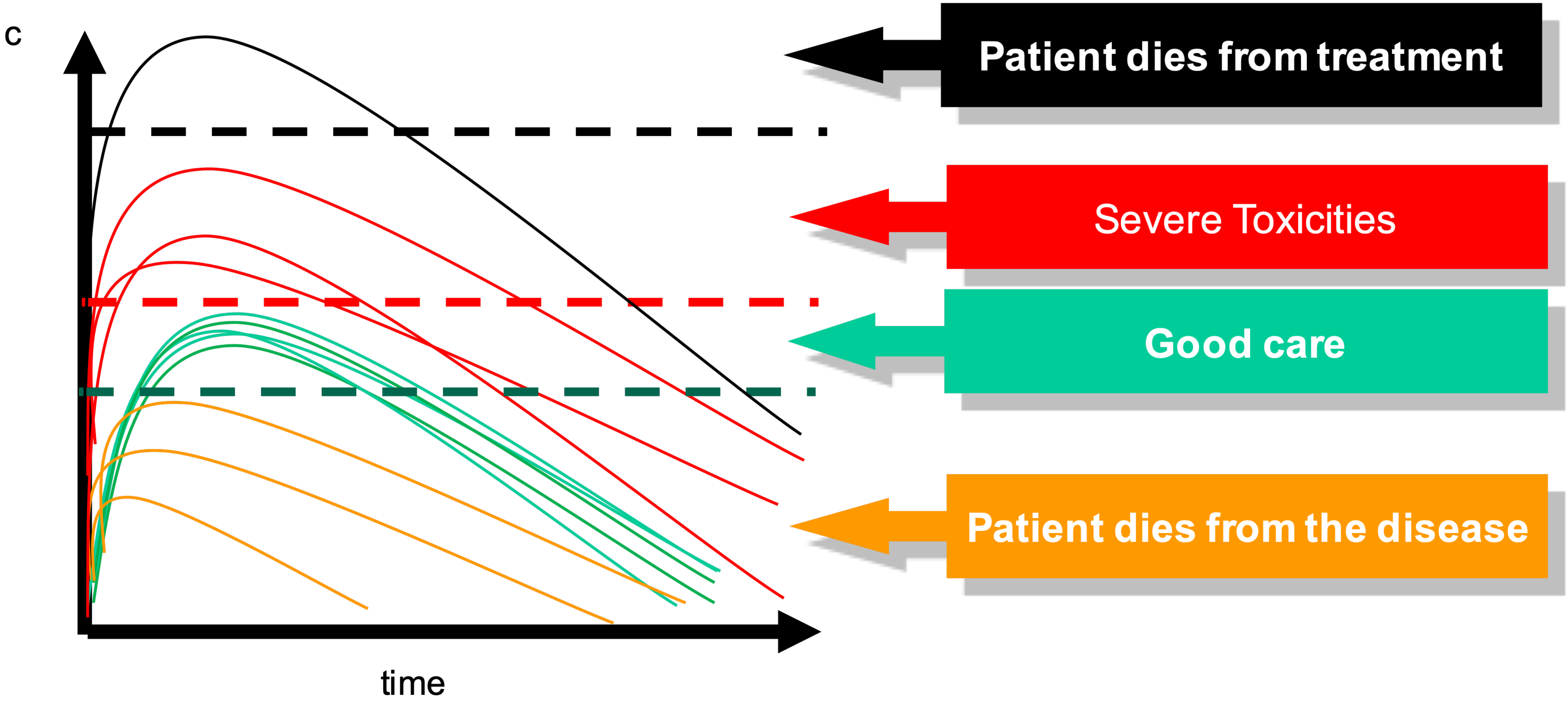
- Other machine learning algorithms tested for classification of 5-years relapse (logistic regression, support vector machine, random forests, k-nearest neighbors and gradient boosting)) gave similar results

- Mechanistic modeling provides biological and clinical insights that ML does not:
  - Ki67 correlates with proliferation rate α (expected but reassurring)
  - HER2 correlates with α, EGFR with μ (metastatic potential)
  - prediction of the **invisible metastatic state** at diagnosis ⇒ potential for **personalized adjuvant therapy**

- This is a first attempt of a mechanistic, individual-level, predictive metastatic model. A lot remains to be done:
  - Refinement to well-established breast cancer molecular subtypes
  - Further investigations to refine the modeling (dormancy, etc…)
  - Predictive power to be confirmed in external data sets

Pharmacometrics and precision dosing

Inter-individual variability



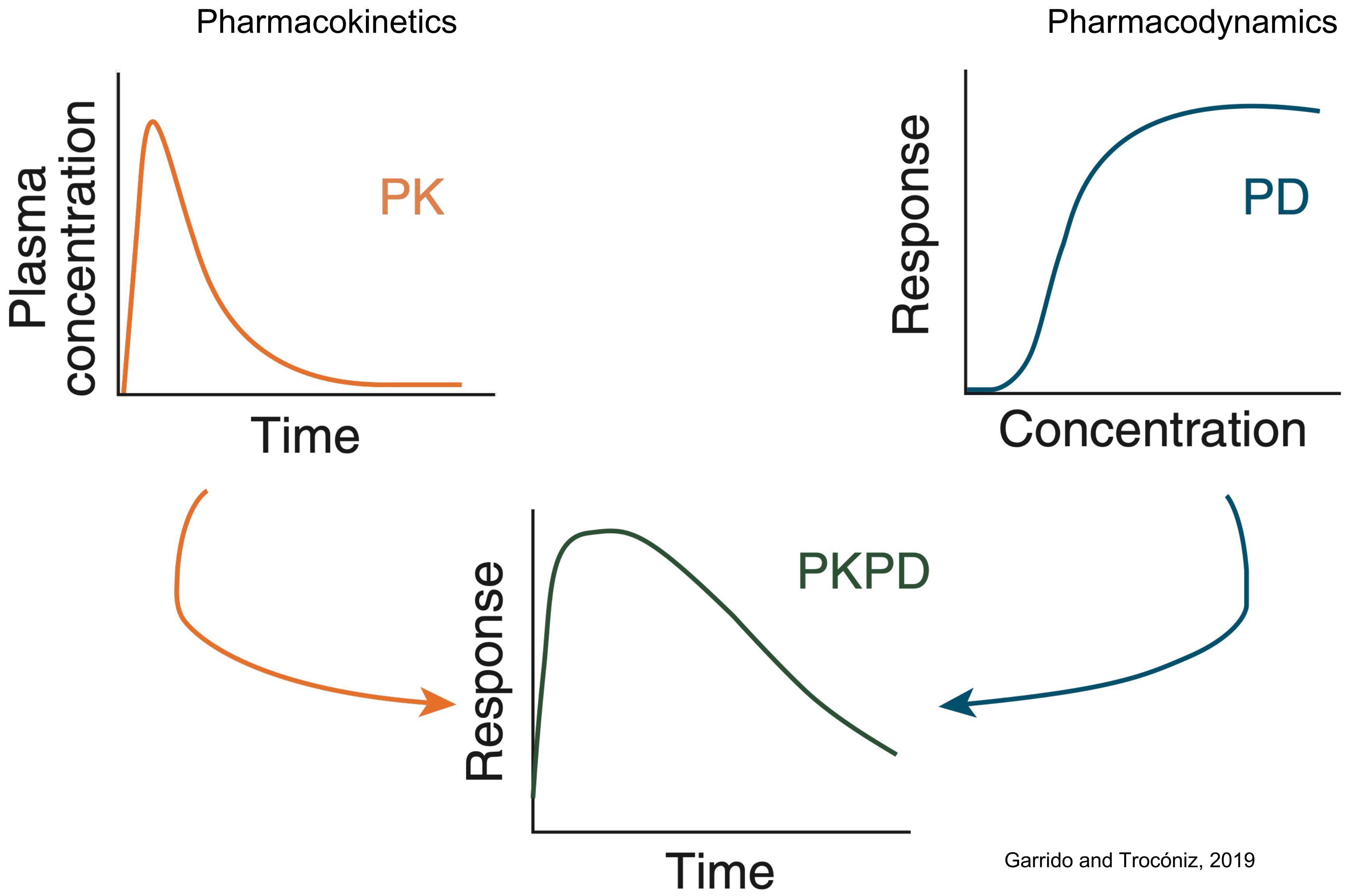
Patient dies from treatment

Severe Toxicities

Good care

Patient dies from the disease

Pharmacometrics = the science of quantitative pharmacology



Garrido and Trocóniz, 2019

Historical overview of PMX in oncology

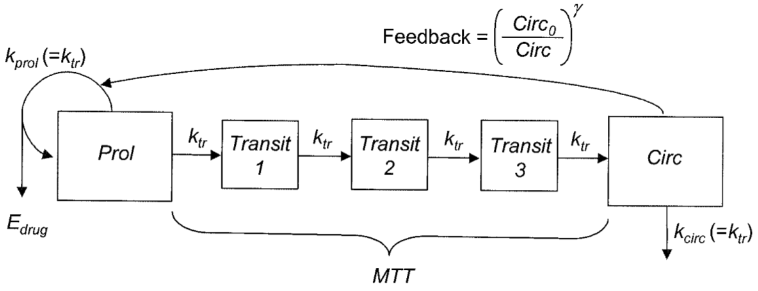
- 1980’s: Principles of population PK modeling by Lewis Sheiner and Stuart Beal
- 1990’s: pop PK models of cytotoxics
- 2000’s: models of hematopoietic toxicity
- 2010’s: tumor growth inhibition models

COMPUTERS AND BIOMEDICAL RESEARCH **5**, 441–459 (1972)

Modelling of Individual Pharmacokinetics for Computer-Aided Drug Dosage*

LEWIS B. SHEINER, BARR ROSENBERG,† AND KENNETH L. MELMON

Departments of Medicine and Pharmacology, Division of Clinical Pharmacology, University of California San Francisco Medical Center, San Francisco, California 94122



Friberg et al., J Clin Oncol, 2002

VOLUME 27 · NUMBER 25 · SEPTEMBER 1 2009

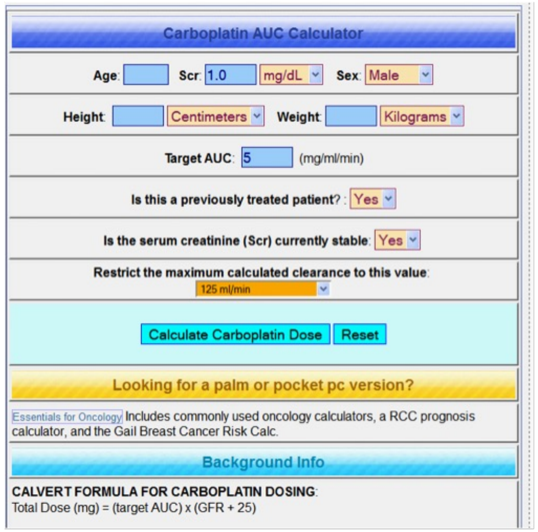
JOURNAL OF CLINICAL ONCOLOGY | ORIGINAL REPORT

Model-Based Prediction of Phase III Overall Survival in Colorectal Cancer on the Basis of Phase II Tumor Dynamics

Laurent Claret, Pascal Girard, Paulo M. Hoff, Eric Van Cutsem, Klaas P. Zuideveld, Karin Jorga, Jan Fagerberg, and René Bruno

How can standard dosing be part of personalized medicine?

- Most anticancer agents are given as:
  - mg/m²
  - mg/kg
  - mg (flat-dose)



Carboplatin AUC Calculator

Age: Scr: 1.0 mg/dL Sex: Male

Height: Centimeters Weight: Kilograms

Target AUC: 5 (mg/ml/min)

Is this a previously treated patient? : Yes

Is the serum creatinine (Scr) currently stable: Yes

Restrict the maximum calculated clearance to this value: 125 ml/min

Looking for a palm or pocket pc version?

Essentials for Oncology Includes commonly used oncology calculators, a RCC prognosis calculator, and the Gail Breast Cancer Risk Calc.

Background Info

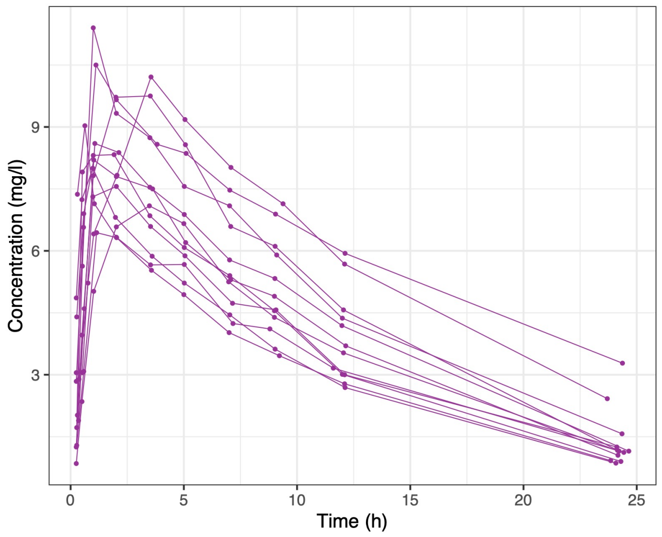
CALVERT FORMULA FOR CARBOPLATIN DOSING:
Total Dose (mg) = (target AUC) x (GFR + 25)

- Only carboplatin is given in a tailored fashion (i.e., AUC5 or AUC6 dosing).

- **« One dose fits all » (standard dosing)**

Mixed-effects modeling

Population data

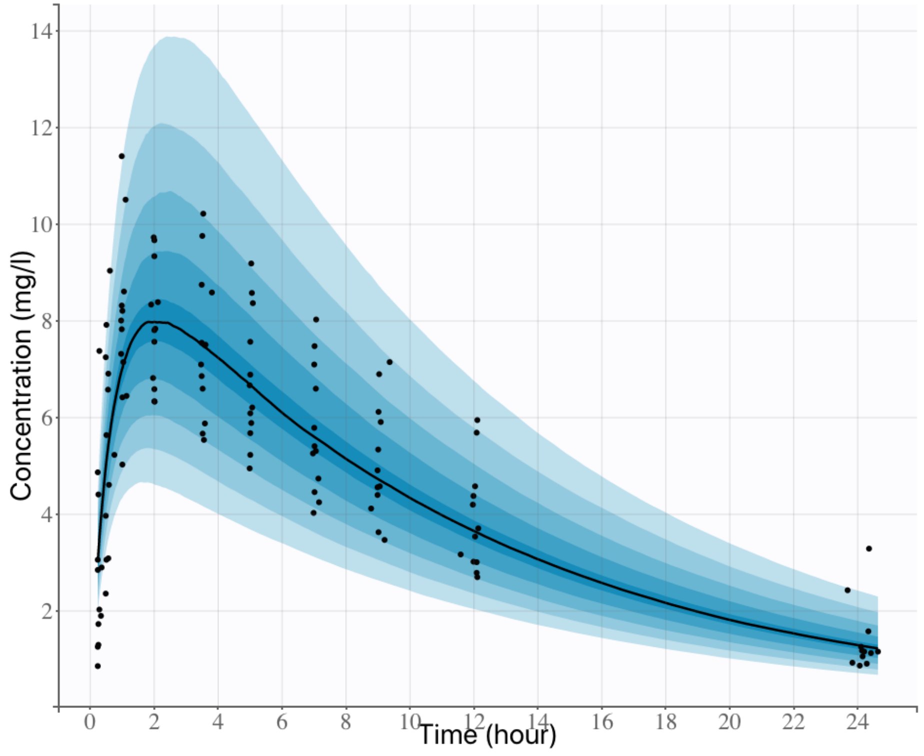


$$\psi^i = \psi_{pop} + \eta^i, \eta^i \sim \mathcal{N}(0, \Omega)$$

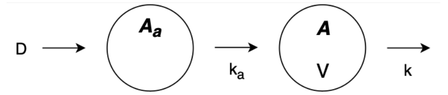
fixed effects

random effects

Population fit (MLE)

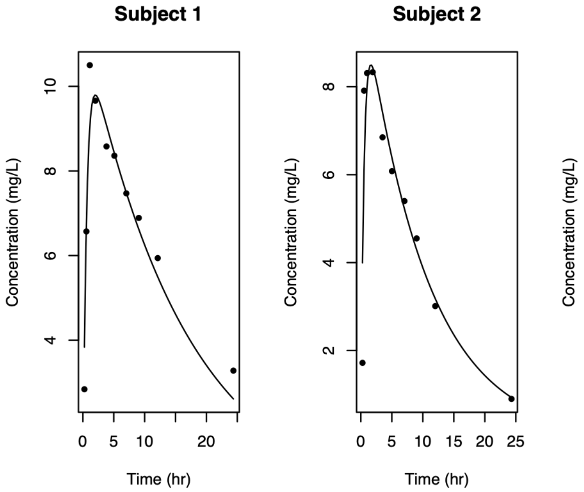


Individual structural model



$$\begin{cases} \frac{dA_a}{dt} = -k_a A_a \\ \frac{dA}{dt} = k_a A_a - kA \\ A_a(t=0) = D, \quad (t=0) = 0 \end{cases} \qquad C(t) = \frac{A(t)}{V}.$$

Individual fit

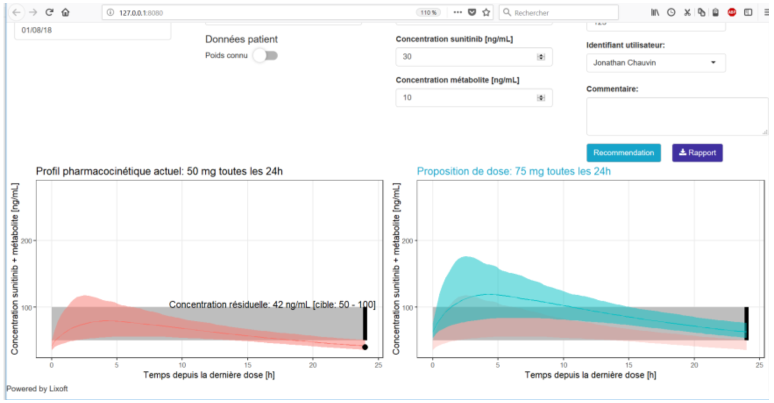

Subject 1

Subject 2

Médecine de précision et bioguidage des ITK

Suivi Thérapeutique Pharmacologique des ITKs (imatinib, sunitinib, dasatinib, cabozantinib, sorafenib, ibrutinib…).

Monolix



127.0.0.1:8080

01/08/18

Données patient

Poids connu

Concentration sunitinib [ng/mL]: 30

Concentration métabolite [ng/mL]: 10

Identifiant utilisateur: Jonathan Chauvin

Commentaire:

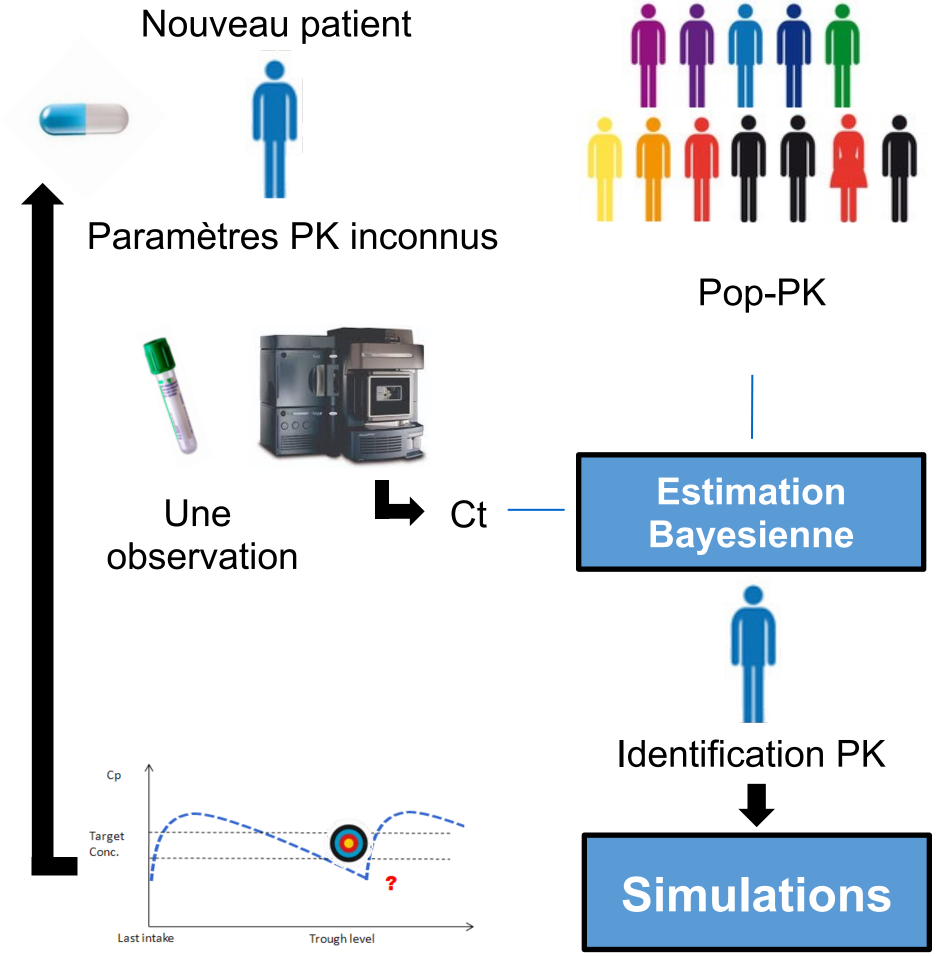
Profil pharmacocinétique actuel: 50 mg toutes les 24h

Concentration résiduelle: 42 ng/mL [cible: 50 - 100]

Proposition de dose: 75 mg toutes les 24h

Powered by Lixoft

Rechercher
110 %
Recommendation
Rapport
Concentration sunitinib + métabolite [ng/mL]
100
200
Temps depuis la dernière dose [h]
0
5
10
15
20
25



Nouveau patient

Paramètres PK inconnus

Pop-PK

Une observation

Ct

Estimation Bayesienne

Identification PK

Simulations

Assistance Publique Hôpitaux de Marseille

SMARTc
SIMULATION MODELING ADAPTIVE RESPONSE FOR THERAPEUTICS IN CANCER

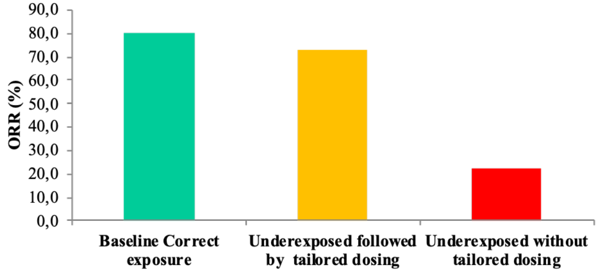
Sunitinib in metastatic kidney cancer

Patient #	Starting Dose (mg)	otal Su + met (ng/ml)	Sampling Time	Simulated Trough Level (ng/ml)	Proposed Dose (mg)	% change
1	50	195	5H30	161	25	-50
2	50	55	23H00	56	62,5	25
3	50	37,4	24H15	40	87,5	75
4	50	40	23h45	42	75	50
5	50	166	22H20	158	25	-50
6	50	161	4H45	136	25	-50
7	50	70	24H00	73	50	no change
8	50	161	4h45	136	25	-50
9	50	17,1	24H00	18	100	**100**
10	50	170	12H30	149	25	-50
11	50	90	24H00	90	37,5	-25
12	50	44,3	24H00	47	75	50
13	50	88	2H15	76	50	no change
14	50	106	19H00	100	37,5	-25
15	50	54,2	6H00	42	87,5	75
16	50	141	1H30	81	37,5	-25
17	50	128	24H00	106	37,5	-25
18	50	118,9	1H00	81	50	no change
19	50	145	19H00	115	37,5	-25
20	50	87	9H30	72	50	no change
21	50	104	3H20	90	37,5	-25
22	50	125	24h00	112	37,5	-25
23	50	62	19H00	58	62,5	25
24	50	246	24H00	231	12,5	**-75**
25	50	150	24H00	143	25	-50
26	50	83	12h00	71	50	no change
27	50	216	24h00	204	12,5	-75
28	50	197	24h00	192	25	-50
29	50	116	8H30	97	37,5	-25
30	50	78	24H00	71	50	no change

Standard dose: 50 mg

80% of AP-HM patient have dose modification of Sutent® 12.5 <>100 mg (-75% ⇨ + 100%!)

Evaluation response as a function of drug exposure (AUC) and consideration of subsequent dose modifications (n=25)

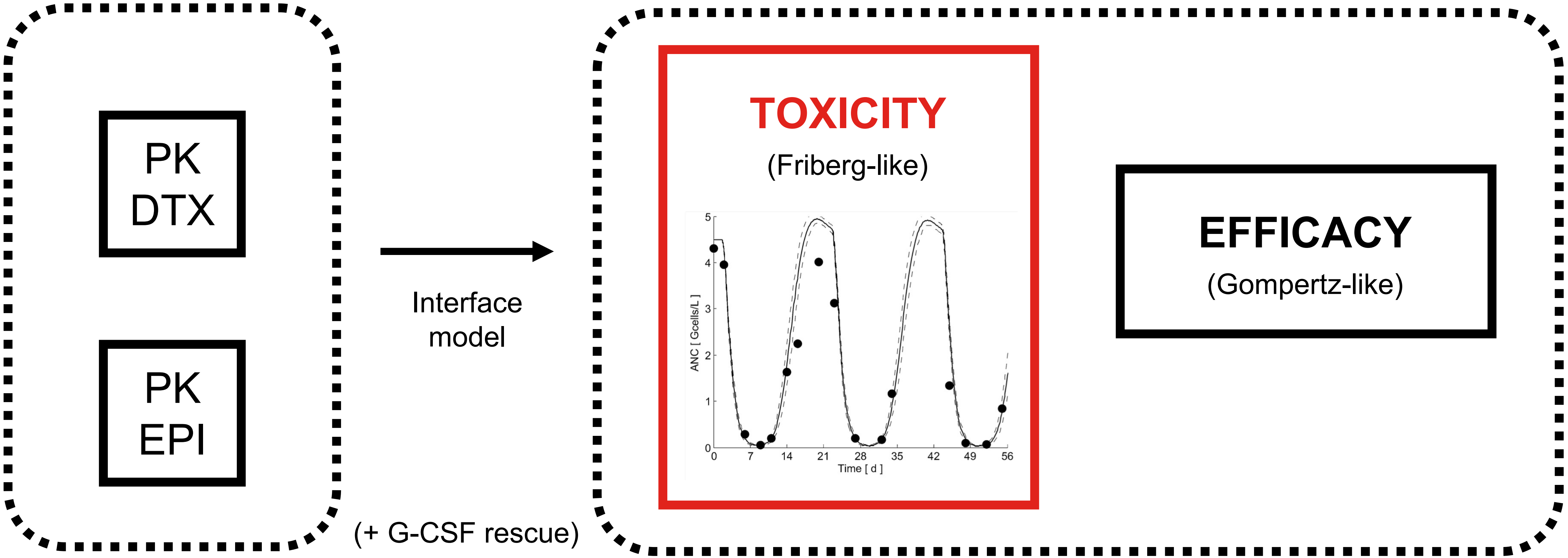


Unpublished data - do not post

J. Ciccolini

Model-based dosing regimen for a phase I/II clinical trial

Goal: safe densification of docetaxel (DTX) + epirubicin (EPI) in metastatic breast cancer



PK models

PK DTX

PK EPI

Interface model

(+ G-CSF rescue)

PD models

TOXICITY

(Friberg-like)

EFFICACY

(Gompertz-like)

Meille et al. (Iliadis), Clin Pharmacokinet, 2016

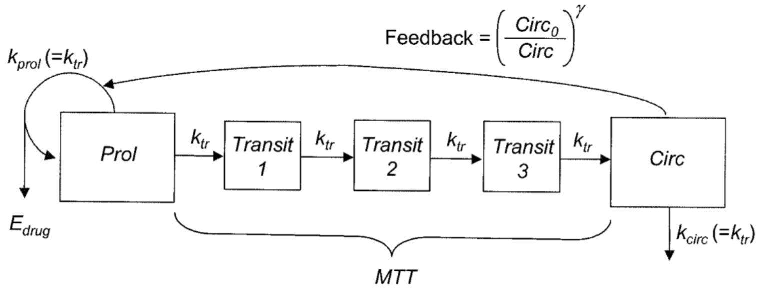
Model equations

Neutrophils kinetics

drug effect

$$
\begin{aligned}
\dot{w}_1(t) &= \gamma \cdot \lambda/\omega \cdot w_0 \cdot \Phi[w(t), \varphi] - \gamma \cdot w_1(t) - N[y(t), v] \cdot w_1(t) \qquad & w_1(0) &= \lambda/\omega \cdot w_0 \\
\dot{w}_2(t) &= \gamma \cdot [w_1(t) - w_2(t)] - N[y(t), v] \cdot w_2(t) & w_2(0) &= \lambda/\omega \cdot w_0 \\
\dot{w}_3(t) &= \omega \cdot w_2(t - \tau) - \lambda \cdot w_3(t) & w_3(-\tau \le t \le 0) &= w_0 \\
\dot{w}(t) &= \lambda \cdot \{\, M[z(t), \mu] \cdot w_3(t) - w(t) \,\} & w(0) &= w_0
\end{aligned}
$$

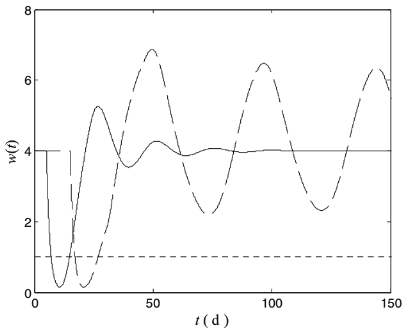
G-CSF



Constraints

$$w(t) \ge W_{\mathrm{D}}$$

$$t_{\mathrm{U}}[w(t) \le W_{\mathrm{U}}] \le T_{\mathrm{U}}$$



Tumor kinetics

$$\frac{\mathrm{d}n(t)}{\mathrm{d}t} = \rho \cdot n(t) \cdot \ln[\theta/n(t)] - \kappa \cdot f\left(c_L^{(D)},\, c_L^{(E)}\right) \cdot n(t) \qquad n(0) = n_0$$

Optimization

$$\underline{d}^* = \arg\min \left[\frac{1}{T} \int_0^T n(t, \underline{d}, \underline{t}^*) \cdot dt \right]$$

under toxicity constraints

Scheduling optimization

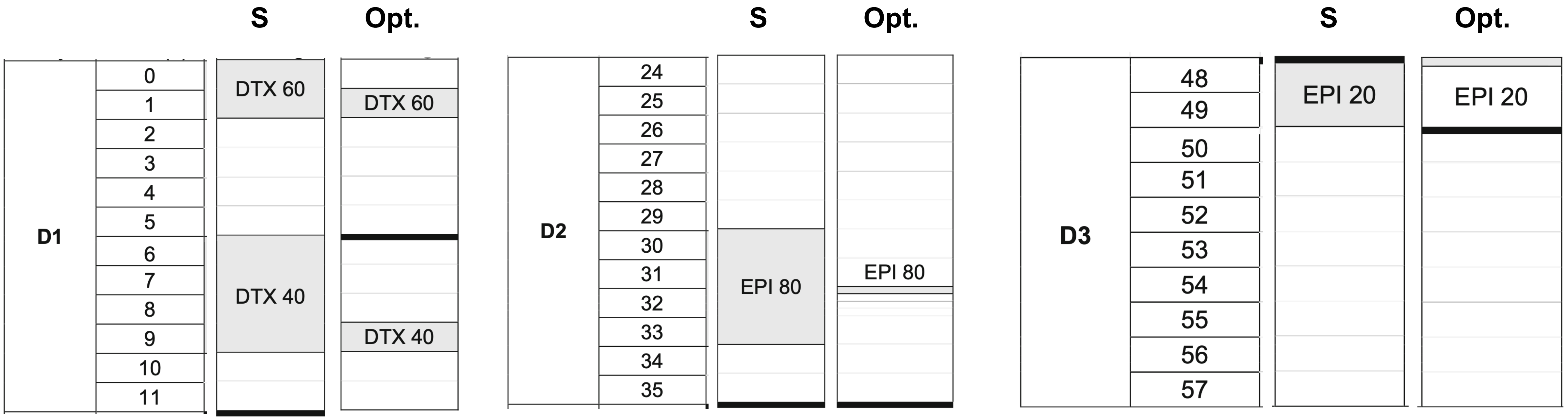
Parameter estimation

- PK: popPK previous studies
- PD toxicity: estimated from previous phase I study
- PD efficacy: *in vitro* cytotoxicity + fit to previously published clinical studies

Optimization

$$\underline{d}^* = \arg\min\left[\frac{1}{T}\int_0^T n(t,\underline{d},\underline{t}^*)\cdot dt\right]$$

under toxicity constraints



S = standard, Opt = optimized

Meille et al. (Iliadis), Clin Pharmacokinet, 2016

MODEL1 clinical results

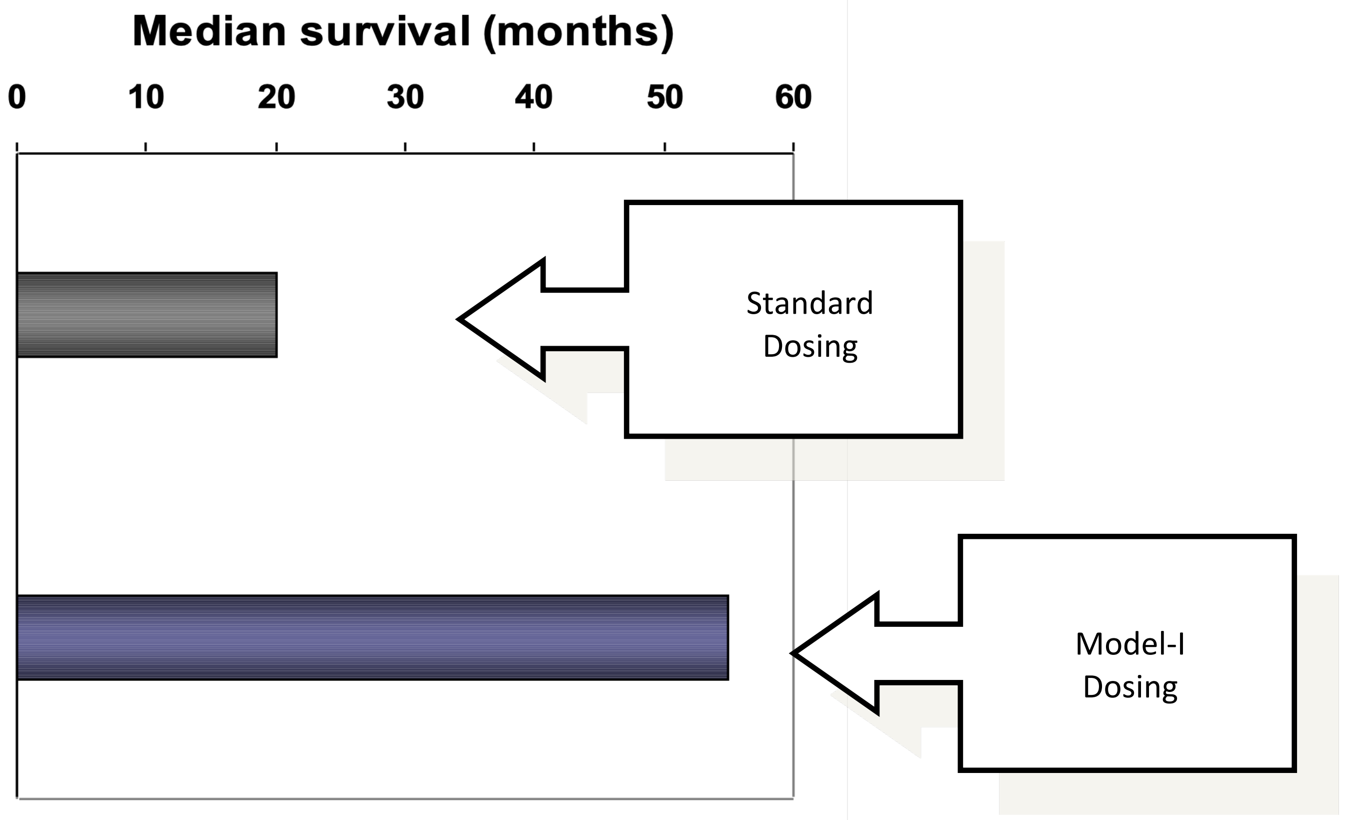
Previously: life-threatening toxicities

- 100% grade ≥ 3 neutropenia
- **1 death**

Viens et al., J Clin Oncol, 2001

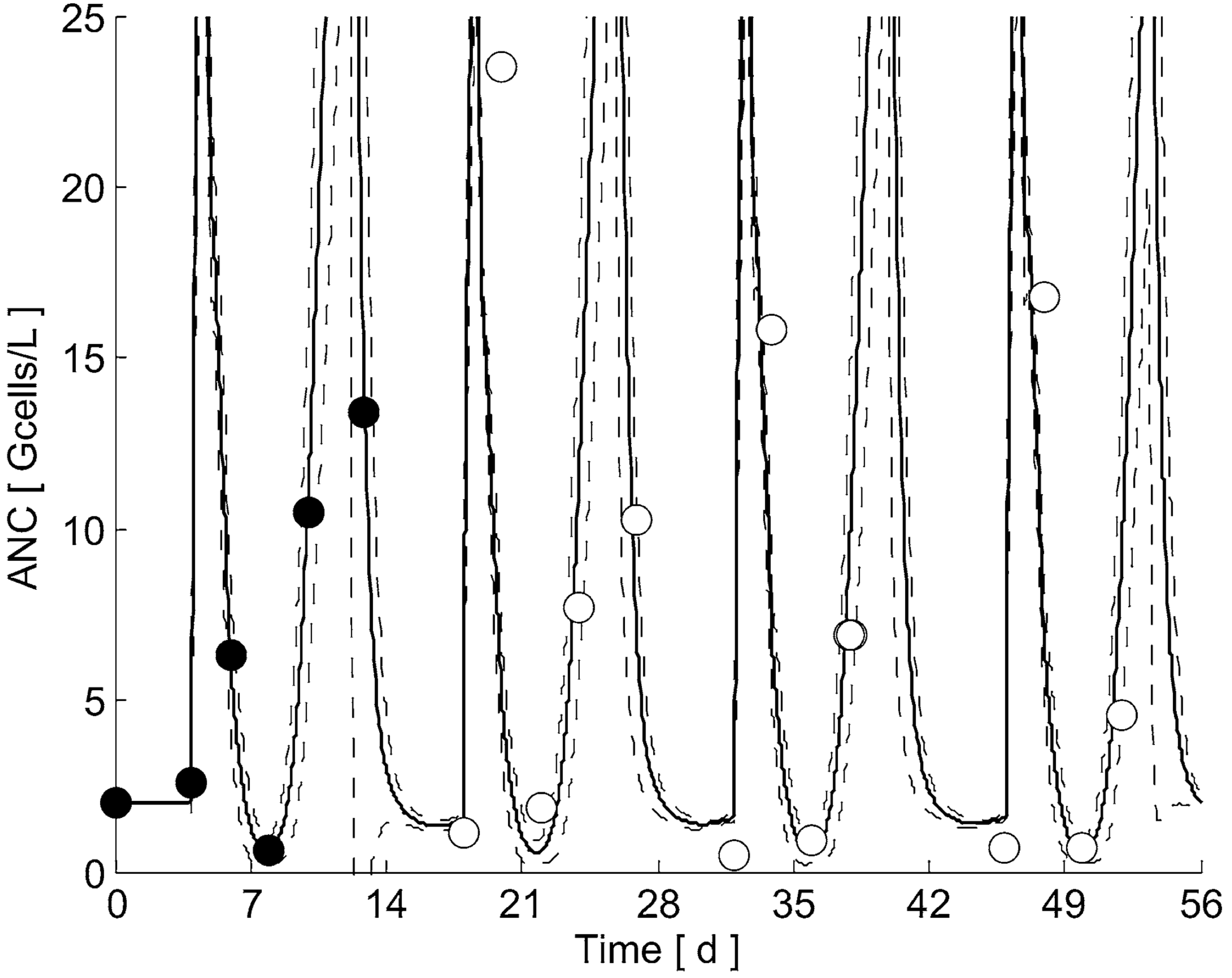
MODEL1: no lethal toxicities

- **0% grade ≥ 3** neutropenia

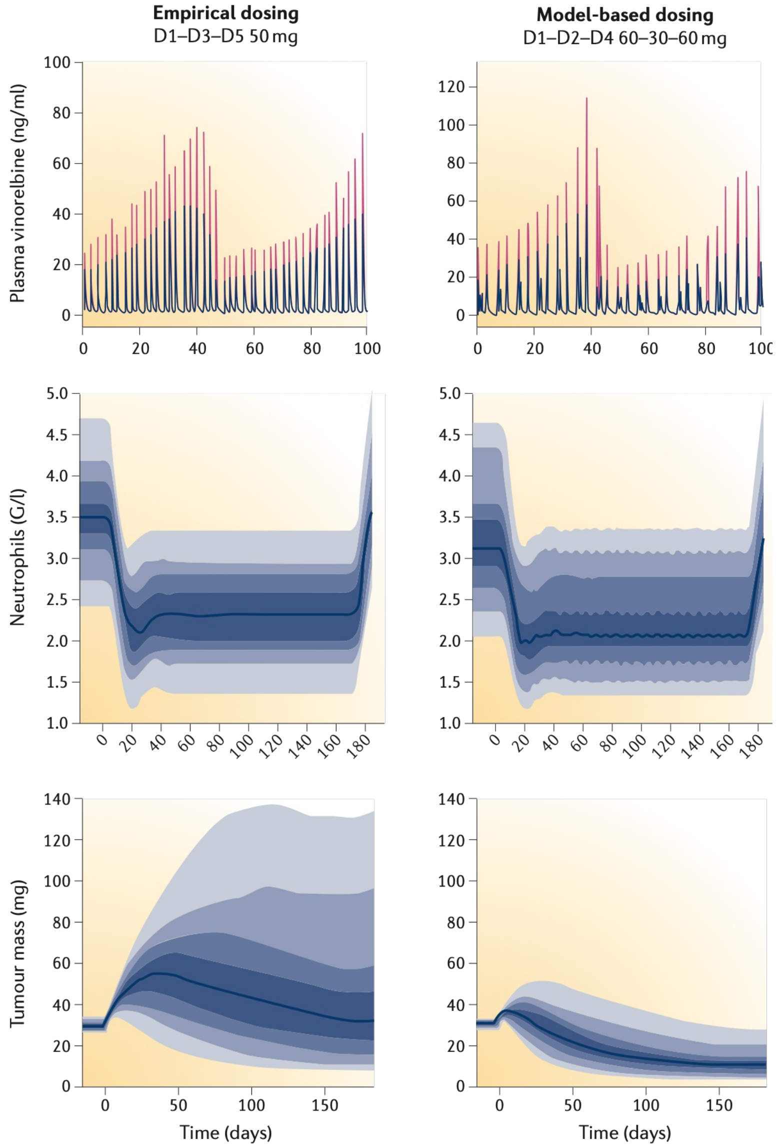


Hénin et al. (Iliadis, Freyer), Breast Cancer Res Treat, 2016

Individualization of parameter estimates

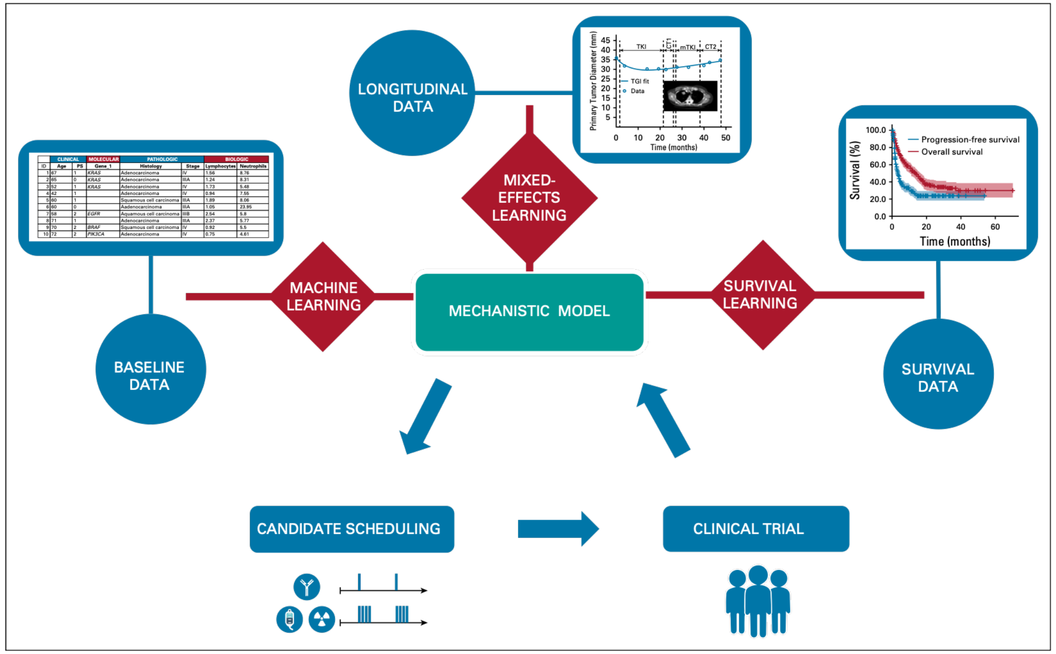


Other model-based trials



- Metronomic vinorelbine in NSCLC (NCT02555007)

- Combination of radiotherapy and immune-checkpoint inhibition (NCT03509584)

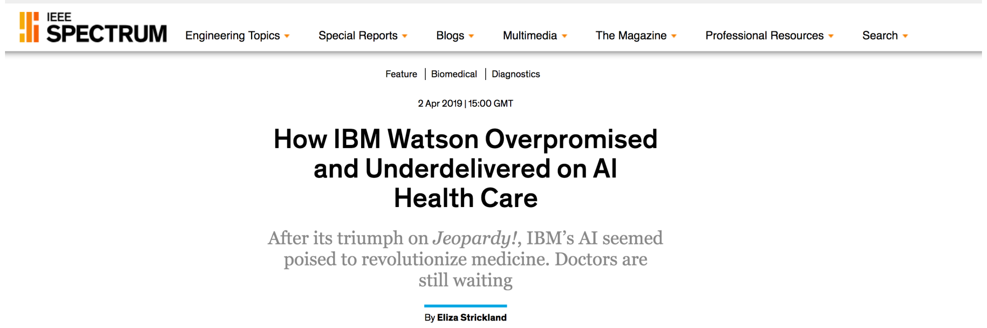


Barbolosi et al., Nat Rev Clin Oncol, 2016
Ciccolini et al. (Benzekry), J Clin Oncol: Precision Oncology, 2020

Conclusion

Conclusion

- Great success of machine learning methods when there are a lot of features and annotated data:
  - genetic sequencing data
  - imaging (pathology, imaging)
- So far, almost no study validated prospectively
- Very few studies using ML/DL in clinical oncology. Almost none in pharmacometrics.
- IBM Watson. Tried to « learn » how oncologists are treating their patients and to digest literature. So far, failed.
- AI will not replace radiologist/pathologist but will become a supplementary tool for daily medical practice



Feature | Biomedical | Diagnostics

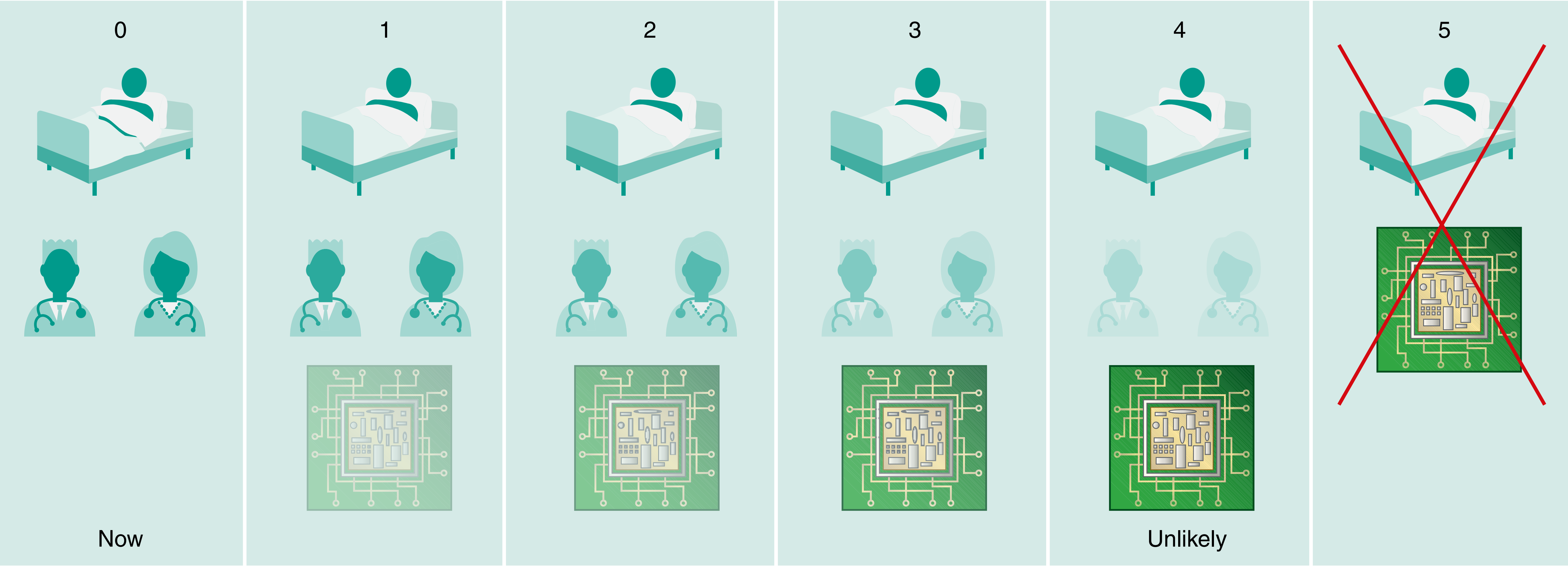
2 Apr 2019 | 15:00 GMT

How IBM Watson Overpromised and Underdelivered on AI Health Care

After its triumph on *Jeopardy!*, IBM's AI seemed poised to revolutionize medicine. Doctors are still waiting

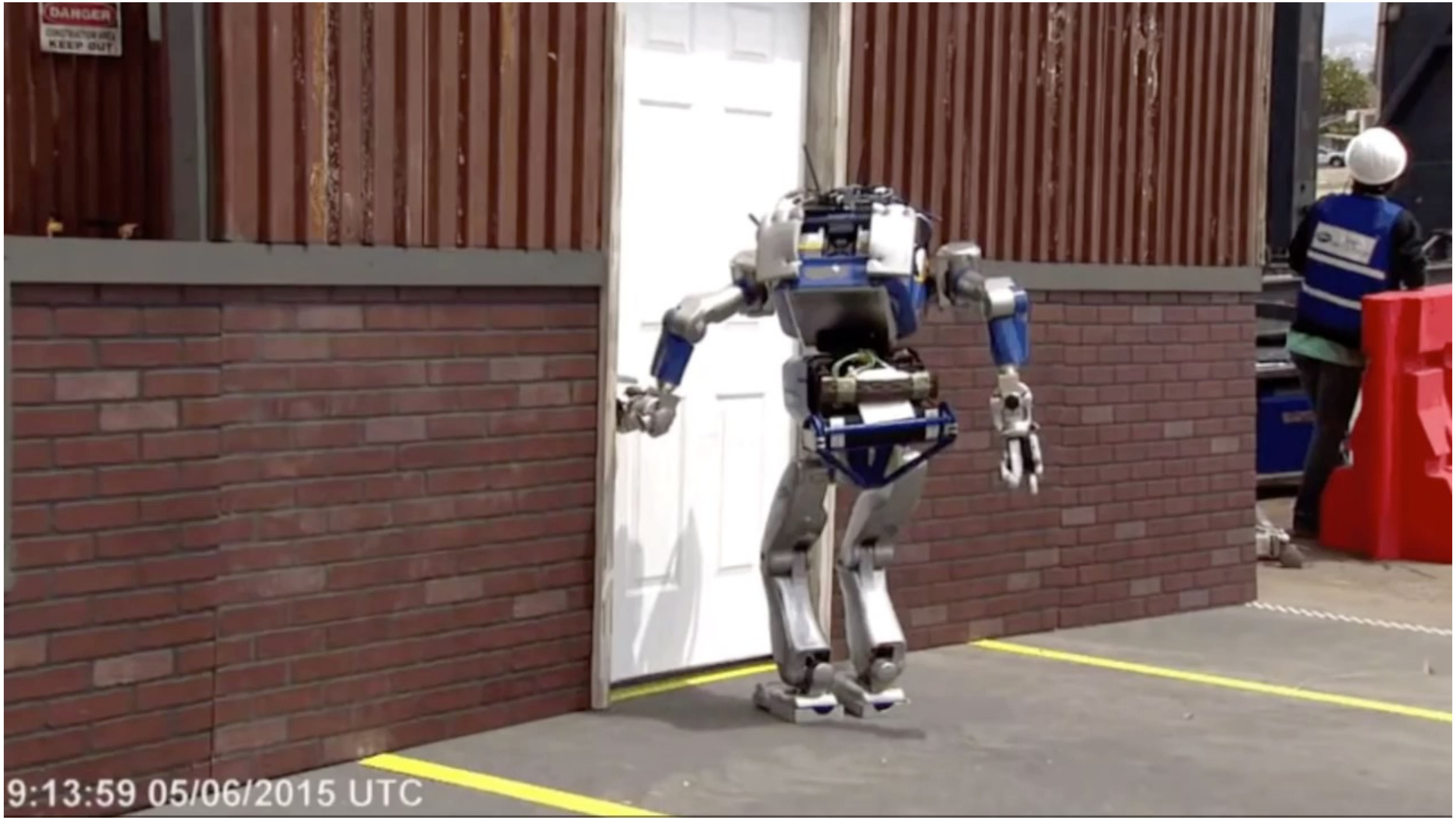
By **Eliza Strickland**

Humans and machine doctors



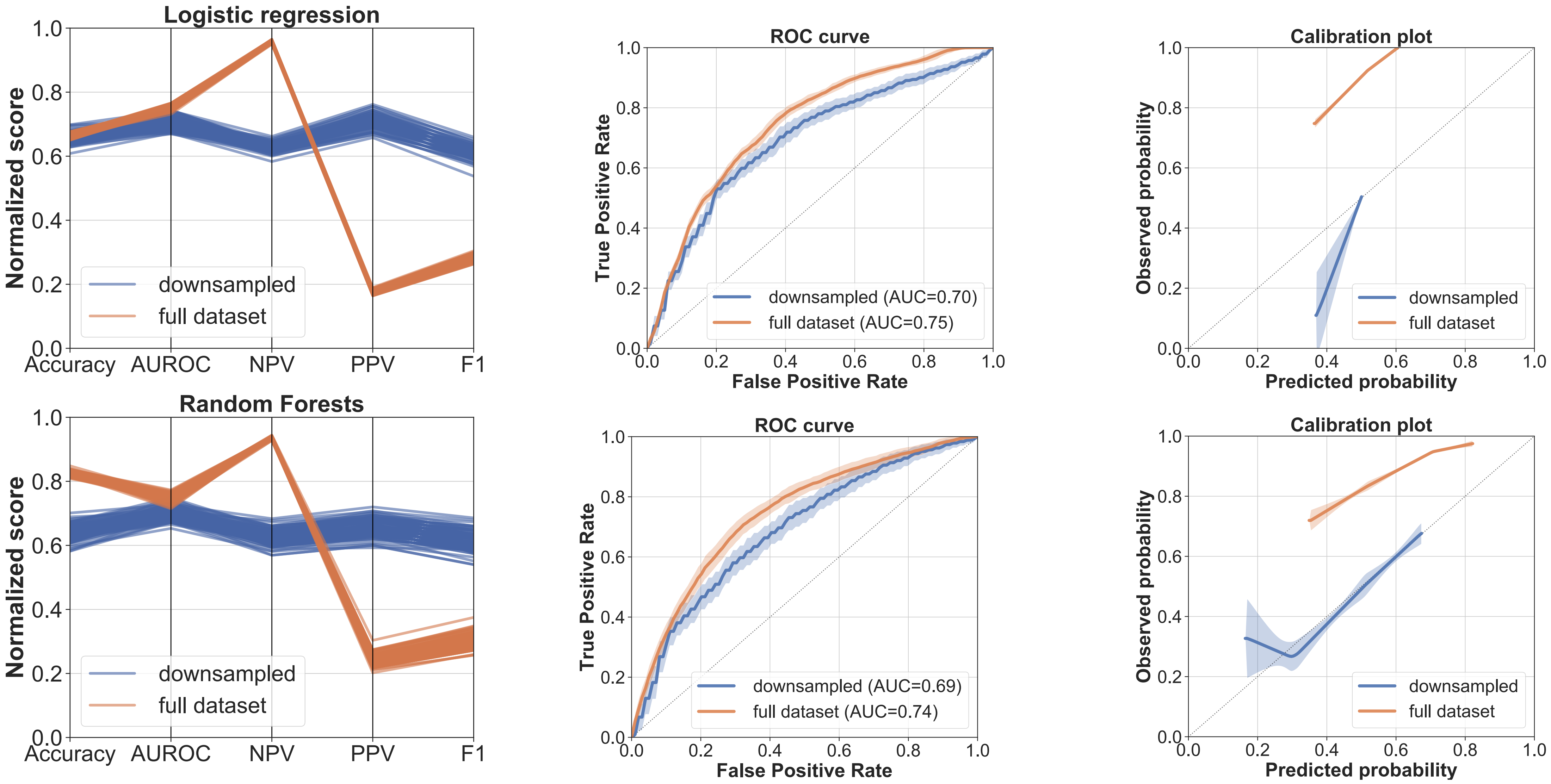
Topol, High-performance medicine: the convergence of human and artificial intelligence, Nat Med, 2019

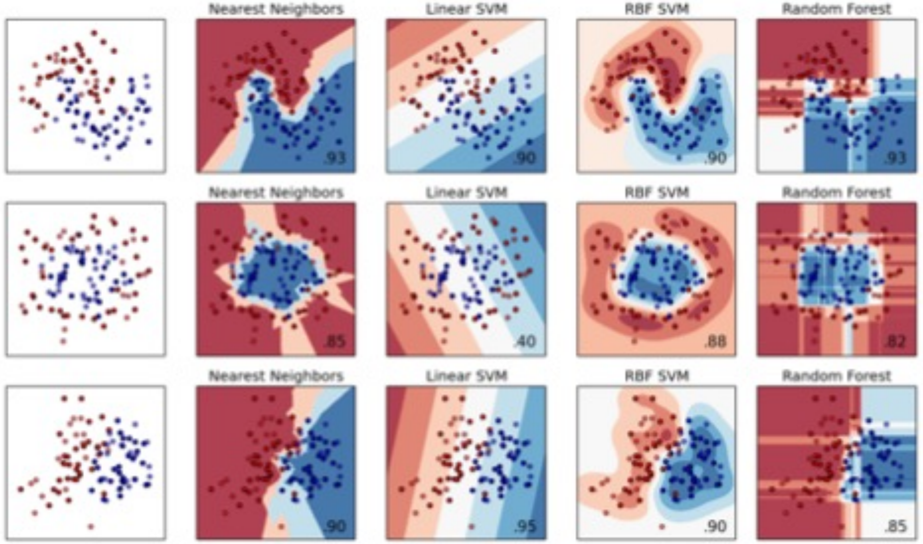
Thank you for your attention!



Additional 2

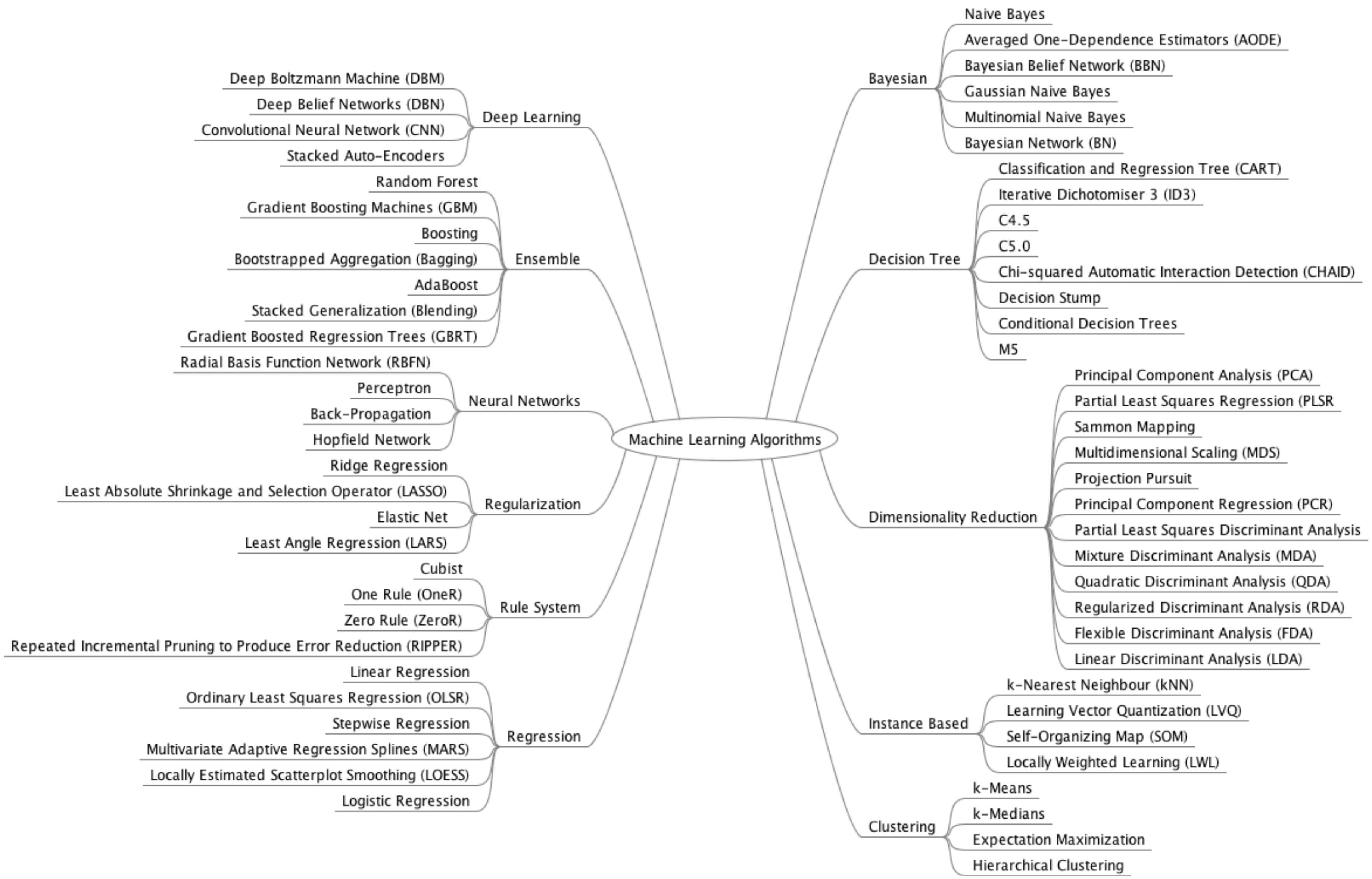
Prediction results





source : scikit-learn

Title



- Machine Learning Algorithms
  - Deep Learning
    - Deep Boltzmann Machine (DBM)
    - Deep Belief Networks (DBN)
    - Convolutional Neural Network (CNN)
    - Stacked Auto-Encoders
  - Ensemble
    - Random Forest
    - Gradient Boosting Machines (GBM)
    - Boosting
    - Bootstrapped Aggregation (Bagging)
    - AdaBoost
    - Stacked Generalization (Blending)
    - Gradient Boosted Regression Trees (GBRT)
  - Neural Networks
    - Radial Basis Function Network (RBFN)
    - Perceptron
    - Back-Propagation
    - Hopfield Network
  - Regularization
    - Ridge Regression
    - Least Absolute Shrinkage and Selection Operator (LASSO)
    - Elastic Net
    - Least Angle Regression (LARS)
  - Rule System
    - Cubist
    - One Rule (OneR)
    - Zero Rule (ZeroR)
    - Repeated Incremental Pruning to Produce Error Reduction (RIPPER)
  - Regression
    - Linear Regression
    - Ordinary Least Squares Regression (OLSR)
    - Stepwise Regression
    - Multivariate Adaptive Regression Splines (MARS)
    - Locally Estimated Scatterplot Smoothing (LOESS)
    - Logistic Regression
  - Bayesian
    - Naive Bayes
    - Averaged One-Dependence Estimators (AODE)
    - Bayesian Belief Network (BBN)
    - Gaussian Naive Bayes
    - Multinomial Naive Bayes
    - Bayesian Network (BN)
  - Decision Tree
    - Classification and Regression Tree (CART)
    - Iterative Dichotomiser 3 (ID3)
    - C4.5
    - C5.0
    - Chi-squared Automatic Interaction Detection (CHAID)
    - Decision Stump
    - Conditional Decision Trees
    - M5
  - Dimensionality Reduction
    - Principal Component Analysis (PCA)
    - Partial Least Squares Regression (PLSR
    - Sammon Mapping
    - Multidimensional Scaling (MDS)
    - Projection Pursuit
    - Principal Component Regression (PCR)
    - Partial Least Squares Discriminant Analysis
    - Mixture Discriminant Analysis (MDA)
    - Quadratic Discriminant Analysis (QDA)
    - Regularized Discriminant Analysis (RDA)
    - Flexible Discriminant Analysis (FDA)
    - Linear Discriminant Analysis (LDA)
  - Instance Based
    - k-Nearest Neighbour (kNN)
    - Learning Vector Quantization (LVQ)
    - Self-Organizing Map (SOM)
    - Locally Weighted Learning (LWL)
  - Clustering
    - k-Means
    - k-Medians
    - Expectation Maximization
    - Hierarchical Clustering

Title

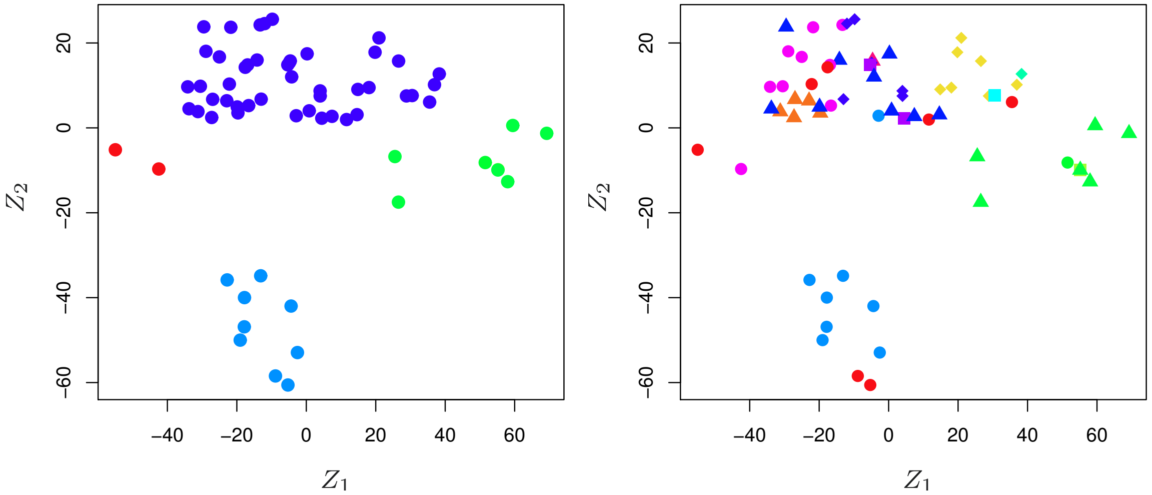


FIGURE 1.4. Left: *Representation of the* NCI60 *gene expression data set in a two-dimensional space, Z_1 and Z_2. Each point corresponds to one of the* 64 *cell lines. There appear to be four groups of cell lines, which we have represented using different colors.* Right: *Same as left panel except that we have represented each of the* 14 *different types of cancer using a different colored symbol. Cell lines corresponding to the same cancer type tend to be nearby in the two-dimensional space.*