Student Lab
Bioinformatics Munich

Technische
Universität
München

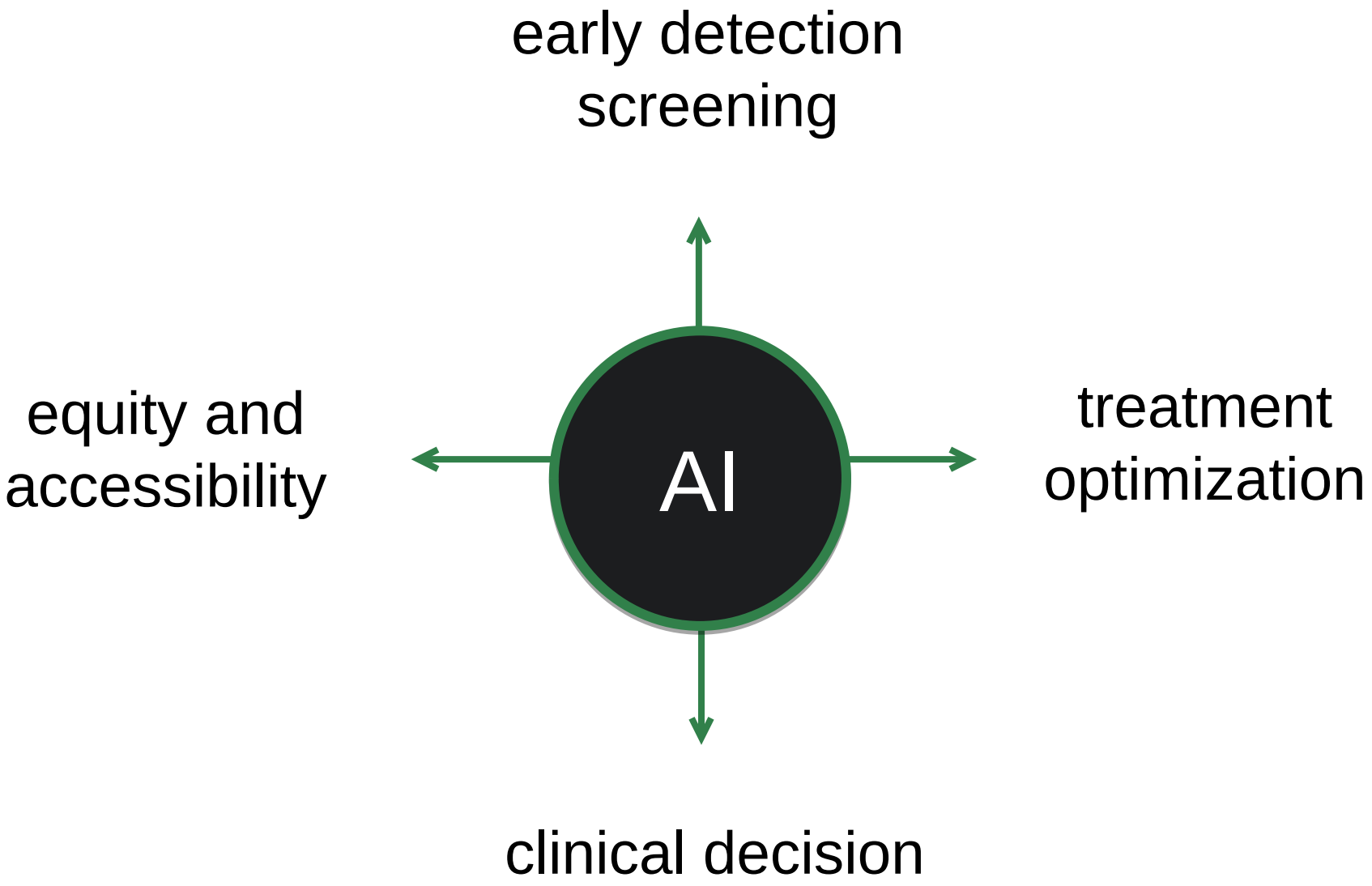
Machine Learning in R

Tolga Tabanli

BMSL, 25.10.2025

Packages to install

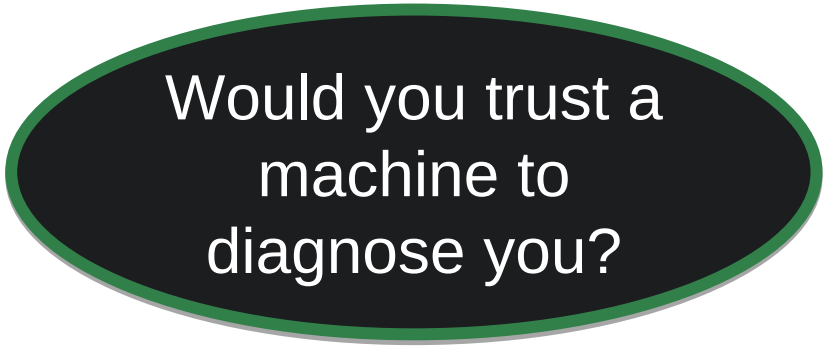
- ❑ tidyverse
- ❑ tidymodels
- ❑ vip
- ❑ corrplot
- ❑ naniar
- ❑ GGally
- ❑ factoextra



European Commission: Directorate-General for Health and Food Safety, EEIG, Open Evidence & PwC. (2025). Study on the deployment of AI in healthcare : final report. Publications Office of the European Union. https://data.europa.eu/doi/10.2875/2169577.

early detection
screening

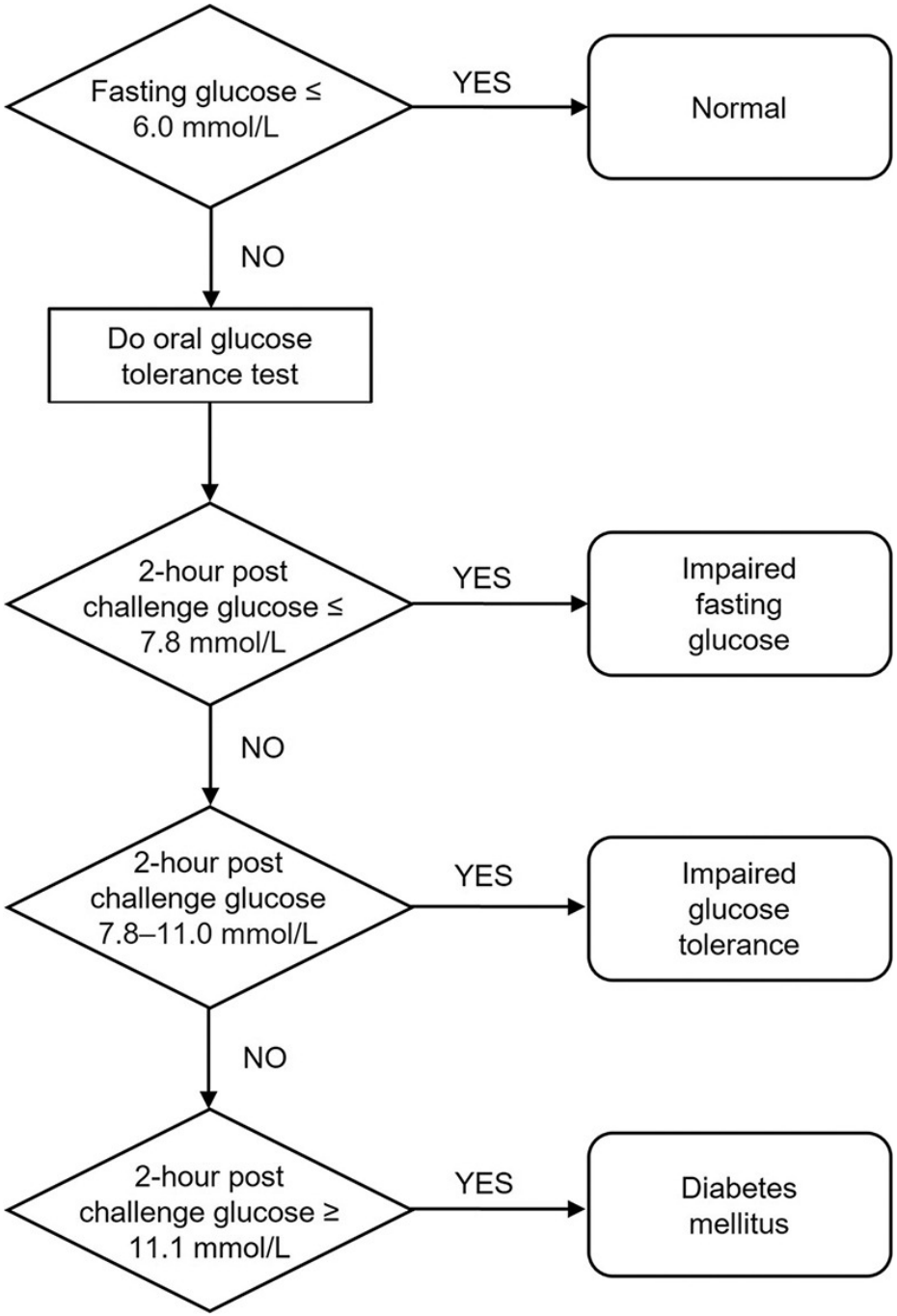
equity and
accessibility



treatment
optimization

clinical decision

European Commission: Directorate-General for Health and Food Safety, EEIG, Open Evidence & PwC. (2025). Study on the deployment of AI in healthcare : final report. Publications Office of the European Union. https://data.europa.eu/doi/10.2875/2169577.



Ting Sim, J. Z., Fong, Q. W., Huang, W., & Tan, C. H. (2023). Machine learning in medicine: what clinicians should know. Singapore medical journal, 64(2), 91–97. https://doi.org/10.11622/smedj.2021054

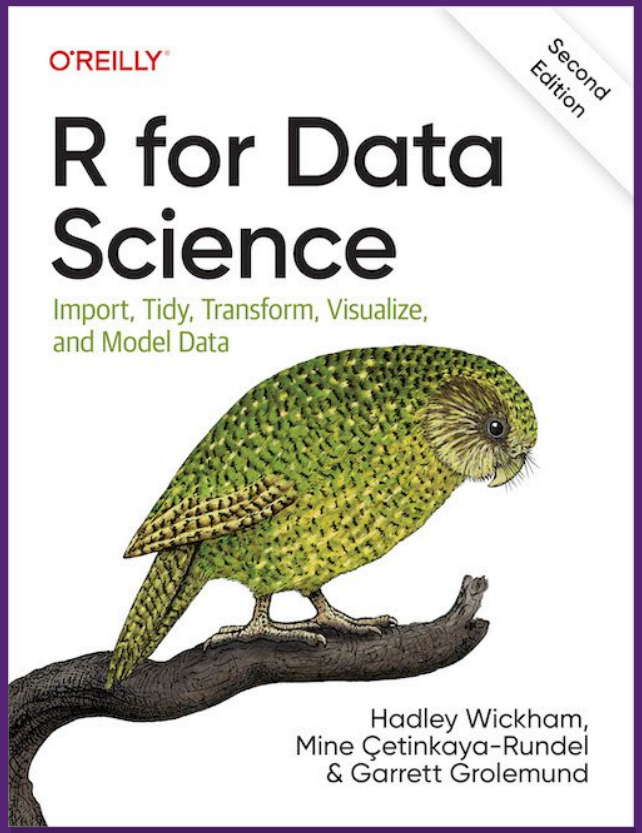
Agenda

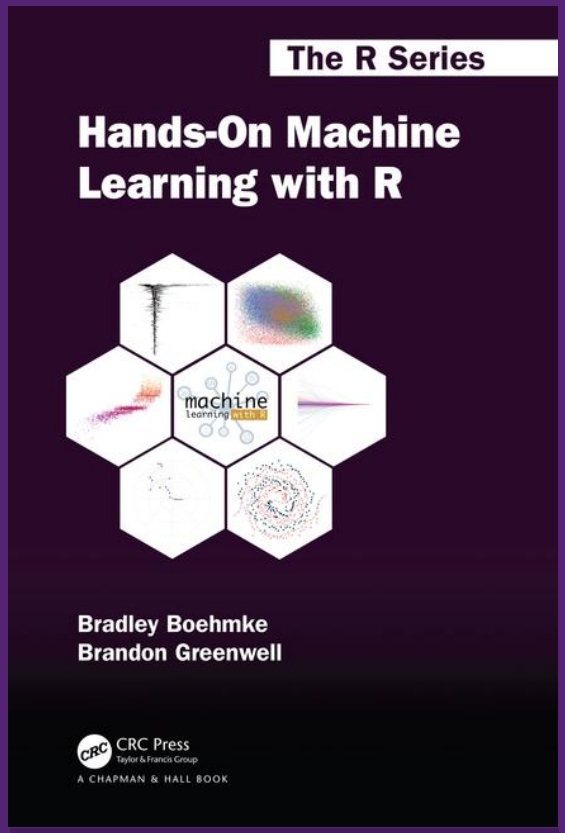
Theory

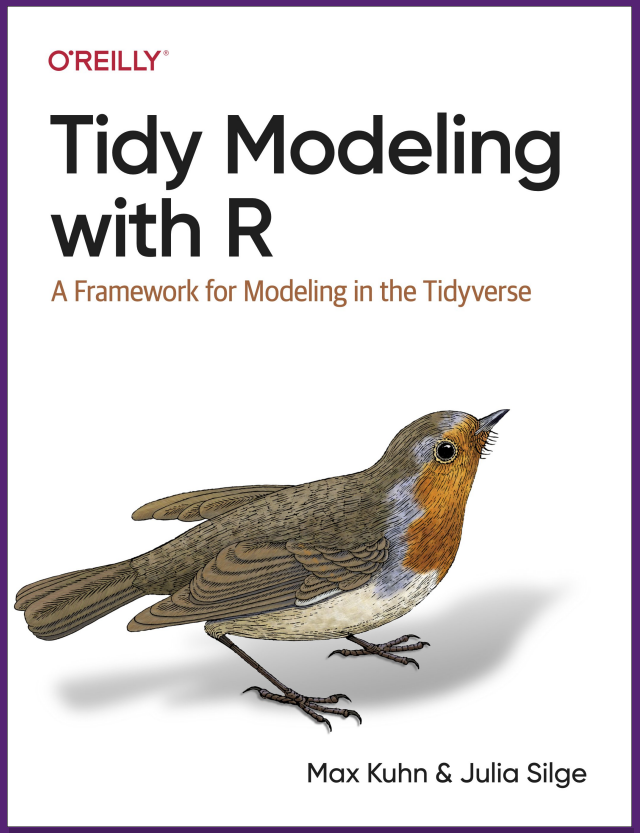
- ❑ Machine Learning Concept and Workflow
- ❑ Algorithms
- ❑ tidymodels

Tasks

- ❑ Regression
- ❑ Classification
- ❑ Clustering


O'REILLY
Second Edition
R for Data Science
Import, Tidy, Transform, Visualize, and Model Data
Hadley Wickham, Mine Çetinkaya-Rundel & Garrett Grolemund


The R Series
Hands-On Machine Learning with R
machine learning with R
Bradley Boehmke
Brandon Greenwell
CRC Press
Taylor & Francis Group
A CHAPMAN & HALL BOOK


O'REILLY
Tidy Modeling with R
A Framework for Modeling in the Tidyverse
Max Kuhn & Julia Silge

tidyverse Refresher

Data manipulation

- **read_csv()**, **read_tsv**()
- **%>%** pipe output to next fxn
- **filter**() rows
- **select**() columns
- **arrange**() rows based on col
- **group_by**() a col‘s values
- **summarise**() wrt. group
- **mutate**() new/existing cols
- **pivot_longer**() cols to rows
- **pivot_wider**() rows to cols

Visualization

- **ggplot**(data, **aes**())
- **aes**(x, y, color, fill...) value mapping
- **geom_point**(), **geom_boxplot**(), **geom_histogram**()
- **labs**(title, x, y) labels

Important data types: **integer**, **double**, **character**, **factor**
Important data structures: **vector**, **list**, **data.frame/tibble/data.table**, **matrix**

```
abalone_raw %>%
    pivot_longer(cols = -c("id", "sex"),
                 names_to ="variable",
                 values_to = "value")
```

abalone.data

sex <chr>	length <dbl>	diameter <dbl>	height <dbl>	whole_weight <dbl>	shucked_weig <dbl>
M	0.455	0.365	0.095	0.5140	0.22
M	0.350	0.265	0.090	0.2255	0.09
F	0.530	0.420	0.135	0.6770	0.25
M	0.440	0.365	0.125	0.5160	0.21

sex <chr>	variable <chr>	value <dbl>
M	length	0.4550
M	diameter	0.3650
M	height	0.0950
M	whole_weight	0.5140
M	shucked_weight	0.2245
M	viscera_weight	0.1010

tidyverse Refresher

Data manipulation

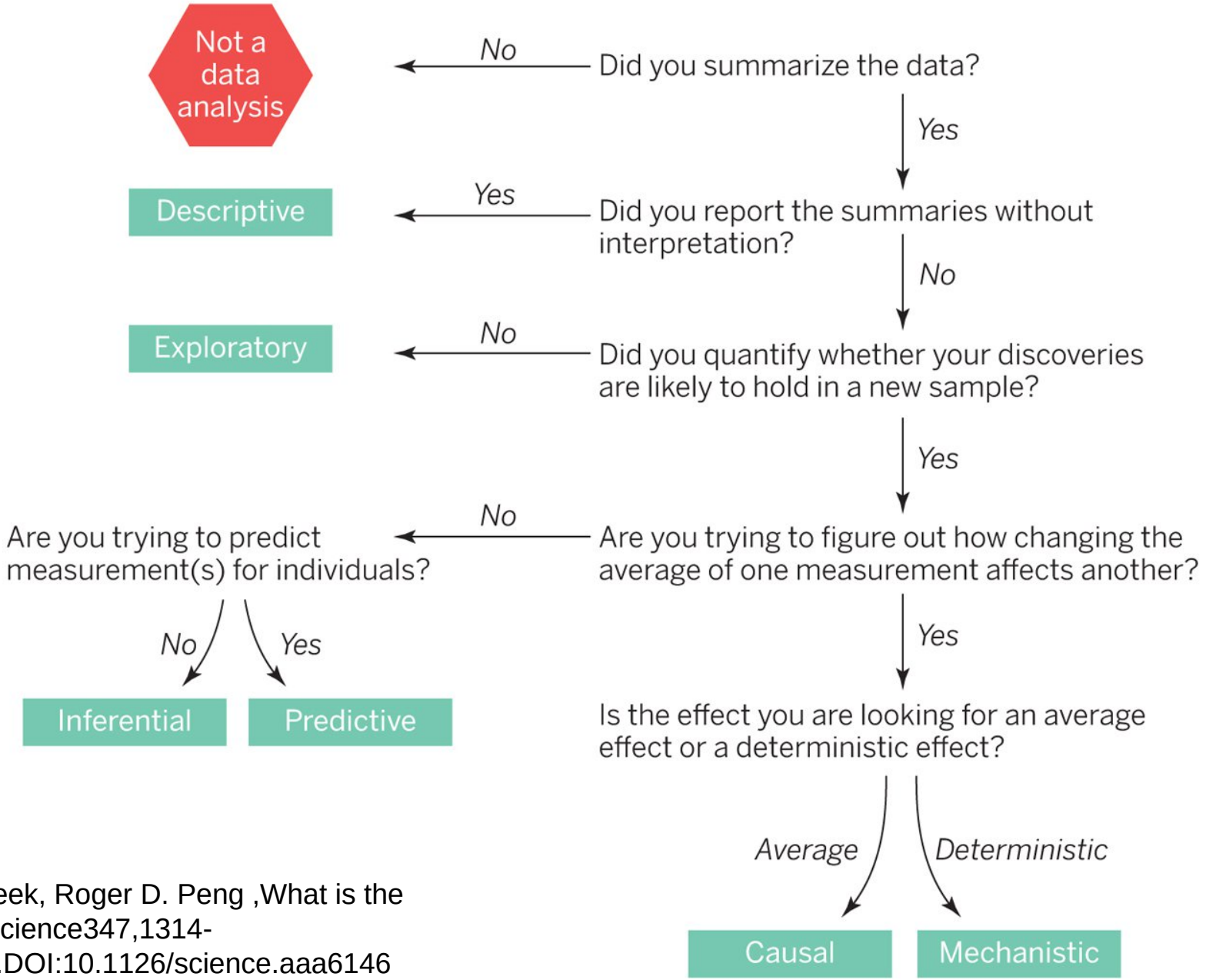
- **read_csv()**, **read_tsv**()
- **%>%** pipe output to next fxn
- **filter**() rows
- **select**() columns
- **arrange**() rows based on col
- **group_by**() a col‘s values
- **summarise**() wrt. group
- **mutate**() new/existing cols
- **pivot_longer**() cols to rows
- **pivot_wider**() rows to cols

Visualization

- **ggplot**(data, **aes**())
- **aes**(x, y, color, fill...) value mapping
- **geom_point**(), **geom_boxplot**(), **geom_histogram**()
- **labs**(title, x, y) labels

Important data types: **integer**, **double**, **character**, **factor**
Important data structures: **vector**, **list**, **data.frame/tibble/data.table**, **matrix**

Data analysis flowchart

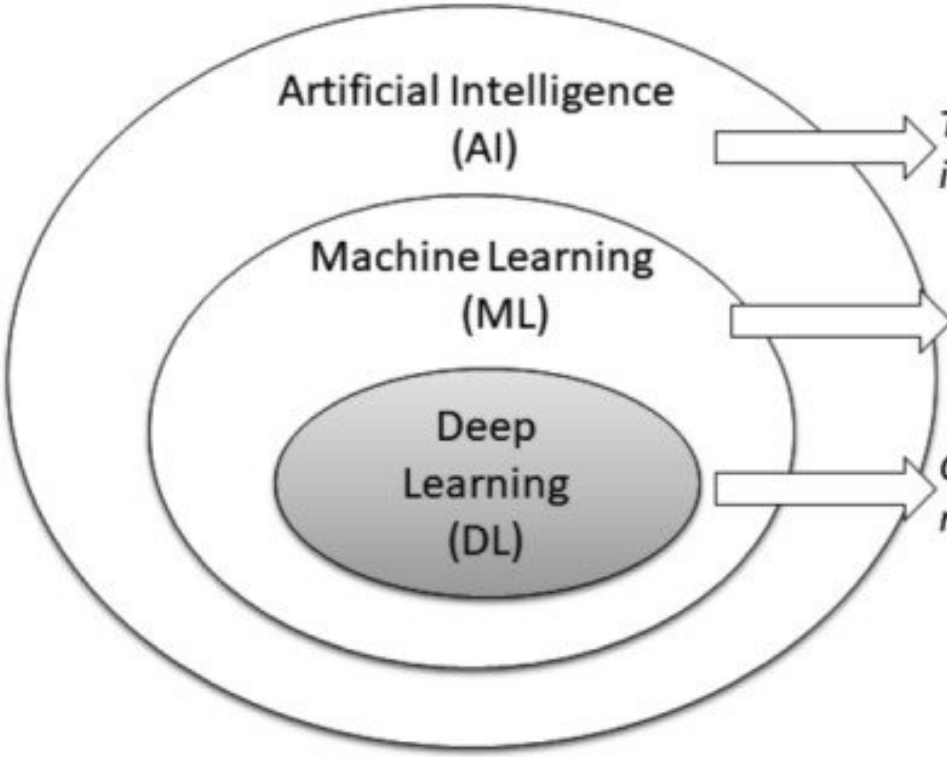


Jeffery T. Leek, Roger D. Peng ,What is the question?.Science347,1314-1315(2015).DOI:10.1126/science.aaa6146

Machine Learning

The goal is not interpretability, but accurate information.
- Bellman, 2001

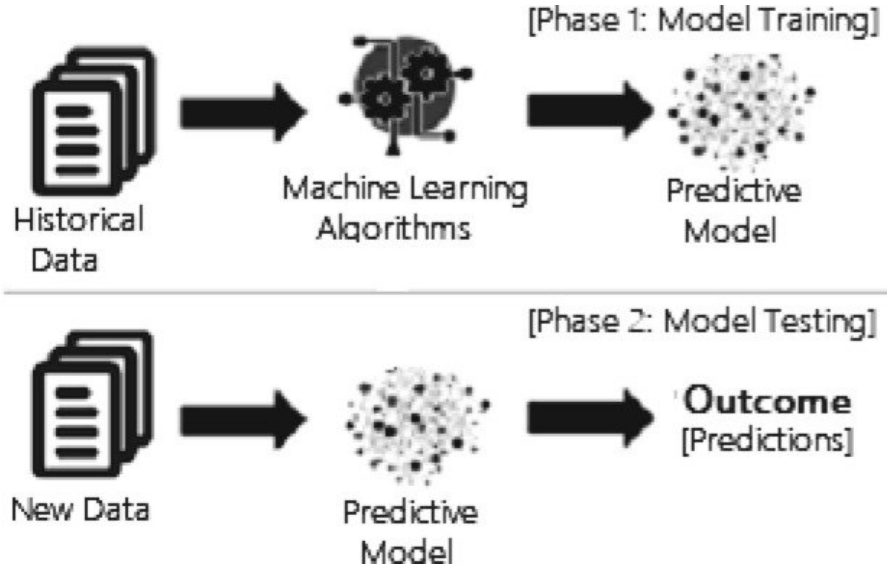
Purpose **Algorithms** **Data Budget** **Workflow**

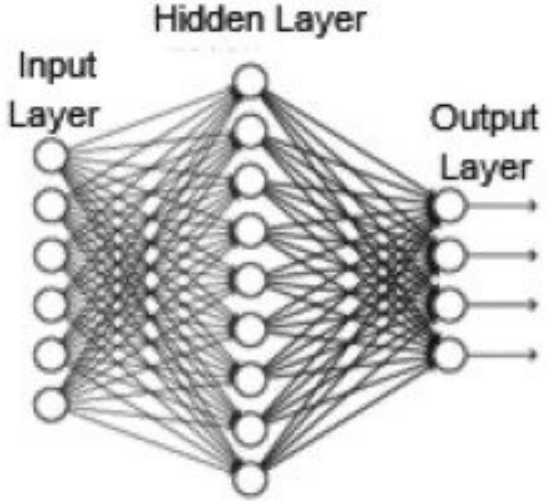


To incorporate human behavior and intelligence to machine or systems.

Methods to learn from data or past experience, which automates analytical model building.

Computation through multi-layer neural networks and processing.





Machine Learning

Supervised Learning

X1	X2	...	Outcome
0.24	0.3	12	15
1.4	0.4	6	7
1.55	4.0	3	9
...	...	...	...

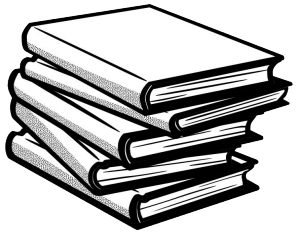
Predict the **outcome** given the **predictors**

Unsupervised Learning

X1	X2	X3	...
0.24	0.3	12	15
1.4	0.4	6	7
1.55	4.0	3	9
...	...	...	...

Find **patterns** in the data given all the variables

Training set

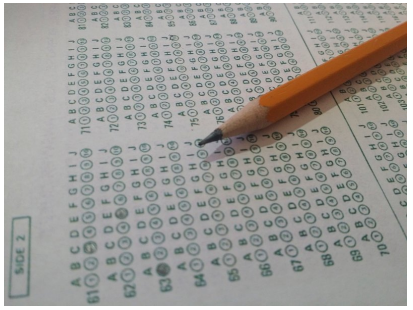


Algorithms

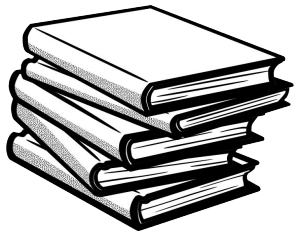
Model

Performance Metric

Test set



Training set



Algorithms

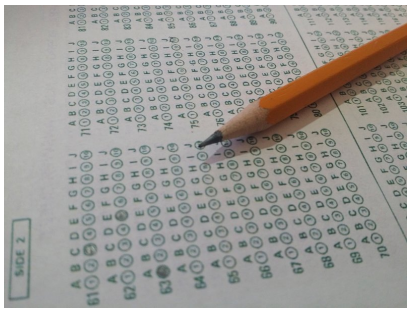
Hyperparameters

- ❑ Repetition frequency, duration
- ❑ language
- ❑ Time of day
- ❑ Place
- ❑ ...

Model

Performance Metric

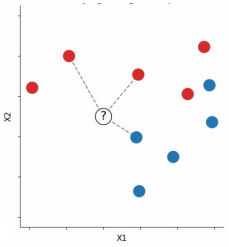
Test set

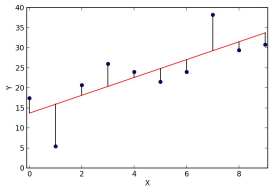


Training set

```
id,full_name,age,gender,smoking_status,bmi,blood_pressure,glucose_levels,condition
1,User0001,,male,Non-Smoker,,,,Pneumonia
2,User0002,38.0,male,Non-Smoker,,105.31586426419374,,Diabetic
3,User0003,18.0,male,Non-Smoker,35.61248565817603,,,Pneumonia
4,User0004,,male,Non-Smoker,,99.11982937713174,,Pneumonia
5,User0005,76.0,male,Non-Smoker,,,,Diabetic
6,User0006,40.0,male,Non-Smoker,33.840722848624225,,,Diabetic
7,User0007,49.0,male,Smoker,,,153.15112560650817,Cancer
8,User0008,47.0,male,Non-Smoker,,,115.82632188486876,199.3396990399767,Diabetic
9,User0009,,male,Non-Smoker,39.64967943512448,,,Diabetic
10,User0010,65.0,male,Smoker,,,,Diabetic
11,User0011,,female,Non-Smoker,,,187.6337512793357,Pneumonia
12,User0012,44.0,male,Non-Smoker,34.44127516695784,,158.37503363765 95,Diabetic
13,User0013,,male,Non-Smoker,20.698671533255396,,,Diabetic
14,User0014,72.0,male,Non-Smoker,,,165.9699843191511,,Diabetic
15,User0015,,male,Non-Smoker,,,117.46478965803582,,Cancer
16,User0016,,male,Non-Smoker,38.57157641296467 4,,,Pneumonia
17,User0017,67.0,male,Non-Smoker,,,109.2593616125518,,Cancer
18,User0018,,male,Non-Smoker,,,135.3309109180692,Diabetic
19,User0019,64.0,male,Non-Smoker,,,175.36588372639818,104.22672364376942,Cancer
20,User0020,,male,Non-Smoker,,,,Diabetic
```

Algorithms





Hyperparameters

- Number of trees in random forest
- Penalty in regression
- Explained variance in PCA
- ...

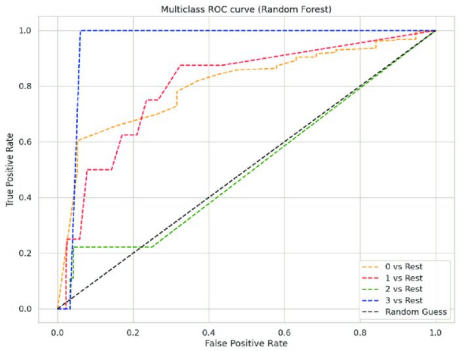
Model

$f(x)$

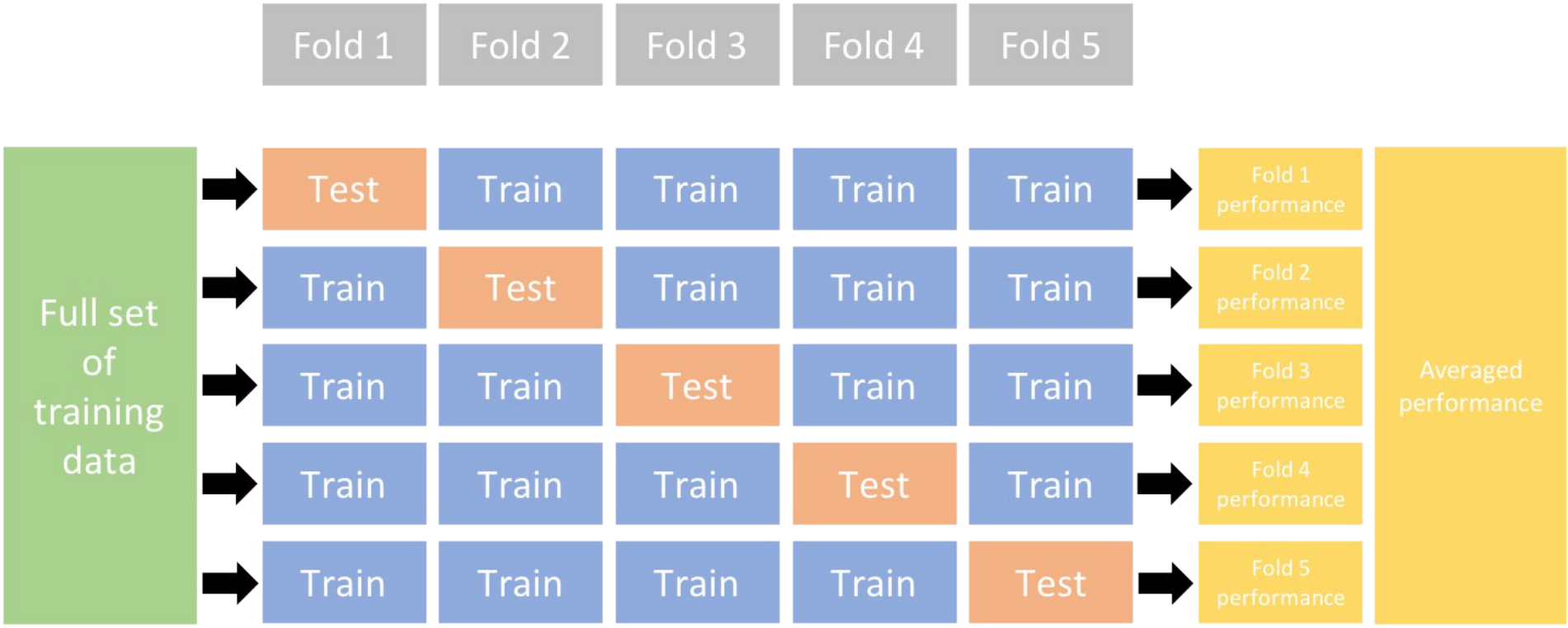
Test Set

```
9994,User9994,,male,Non-Smoker,,127.16529250726067,,Pneumonia
9995,User9995,34.0,male,Non-Smoker,,,181.1528922343641,Diabetic
9996,User9996,,male,Non-Smoker,25.029002450964644,152.54035471406985,137.55145136435425,Pneumonia
9997,User9997,,male,Non-Smoker,27.01748742489308,,,Diabetic
9998,User9998,23.0,male,Smoker,,148.83332145235516,173.93148045105488,Pneumonia
9999,User9999,,female,Non-Smoker,,,,Pneumonia
10000,User10000,27.0,male,Non-Smoker,25.45489062552681,,196.08326727804257,Diabetic
```

Performance Metric

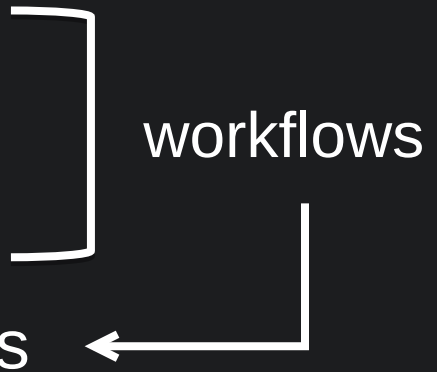


Cross-Validation

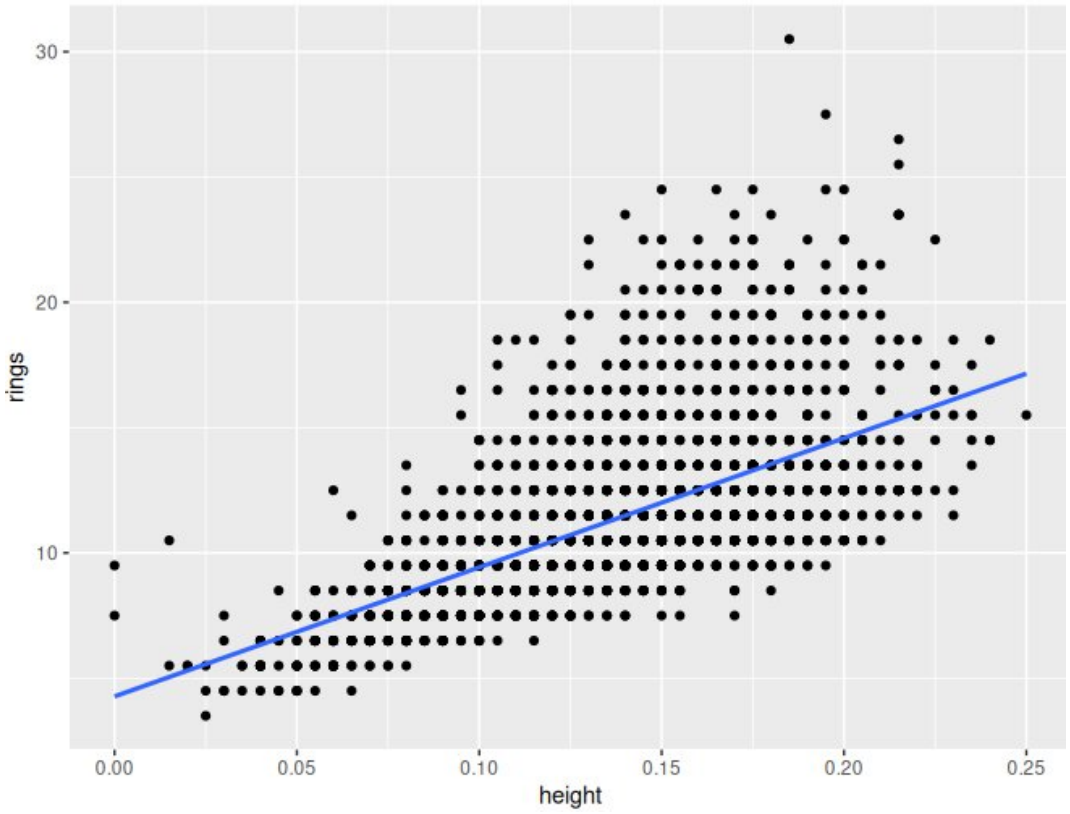


General steps	tidymodels
1. Explore your data	(tidyverse)
— Data split —	rsample
2. Preprocessing	recipes
3. Training	parsnip
4. Model tune - selection	tune, dials
5. Evaluation	yardstick

General steps	tidymodels
1. Explore your data	(tidyverse)
— Data split —	rsample
2. Preprocessing	recipes
3. Training	parsnip
4. Model tune - selection	tune, dials
5. Evaluation	yardstick



Regression



Linear relationship

Target: numeric
Predictor: numeric

Objective: Minimize

$$L(w) = \sum_{i=1} (y_i - wx_i)^2$$

Example Metric: Mean absolute error

Regularized regression

Ridge regression

$$L(w) = \sum_{i=1} (y_i - wx_i)^2 + \lambda \sum_{j=0}^{d} w_j^2$$

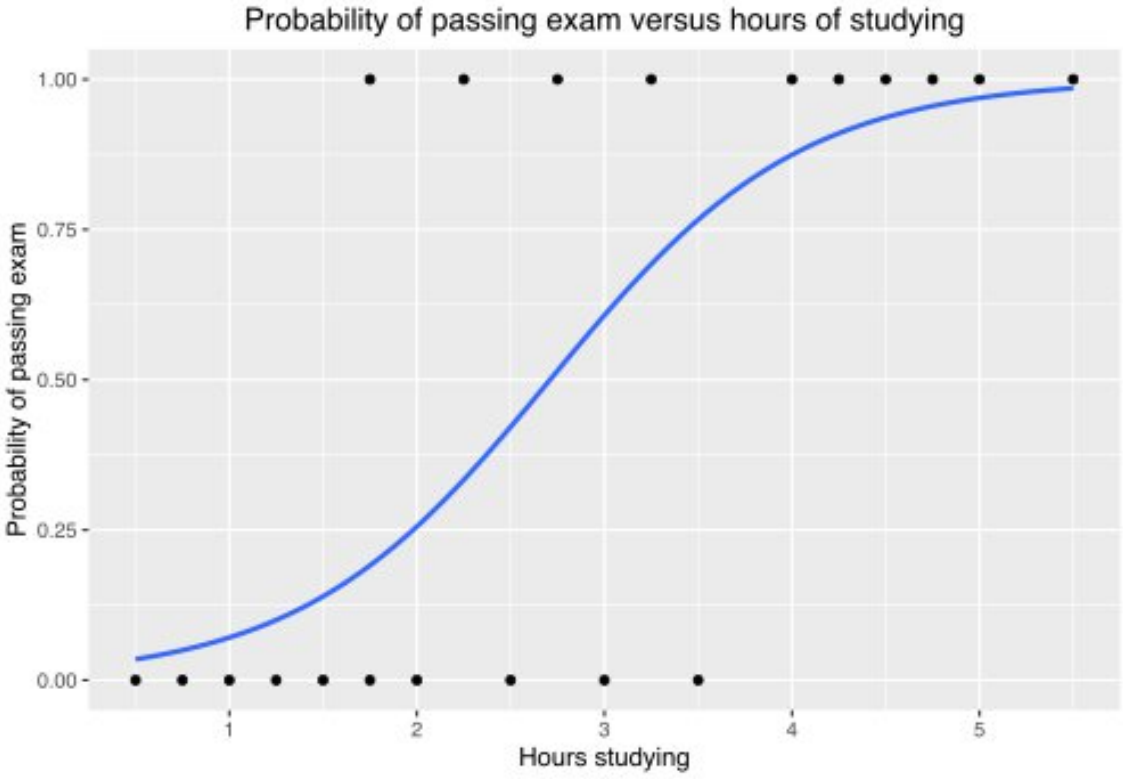
- **small** sample size
- **sparse** data

Lasso regression

$$L(w) = \sum_{i=1} (y_i - wx_i)^2 + \lambda \sum_{j=0}^{d} |w_j|$$

- if **few** predictors with real effect
- can **eliminate** predictors
- more **robust**

Logistic Regression



Linear relationship

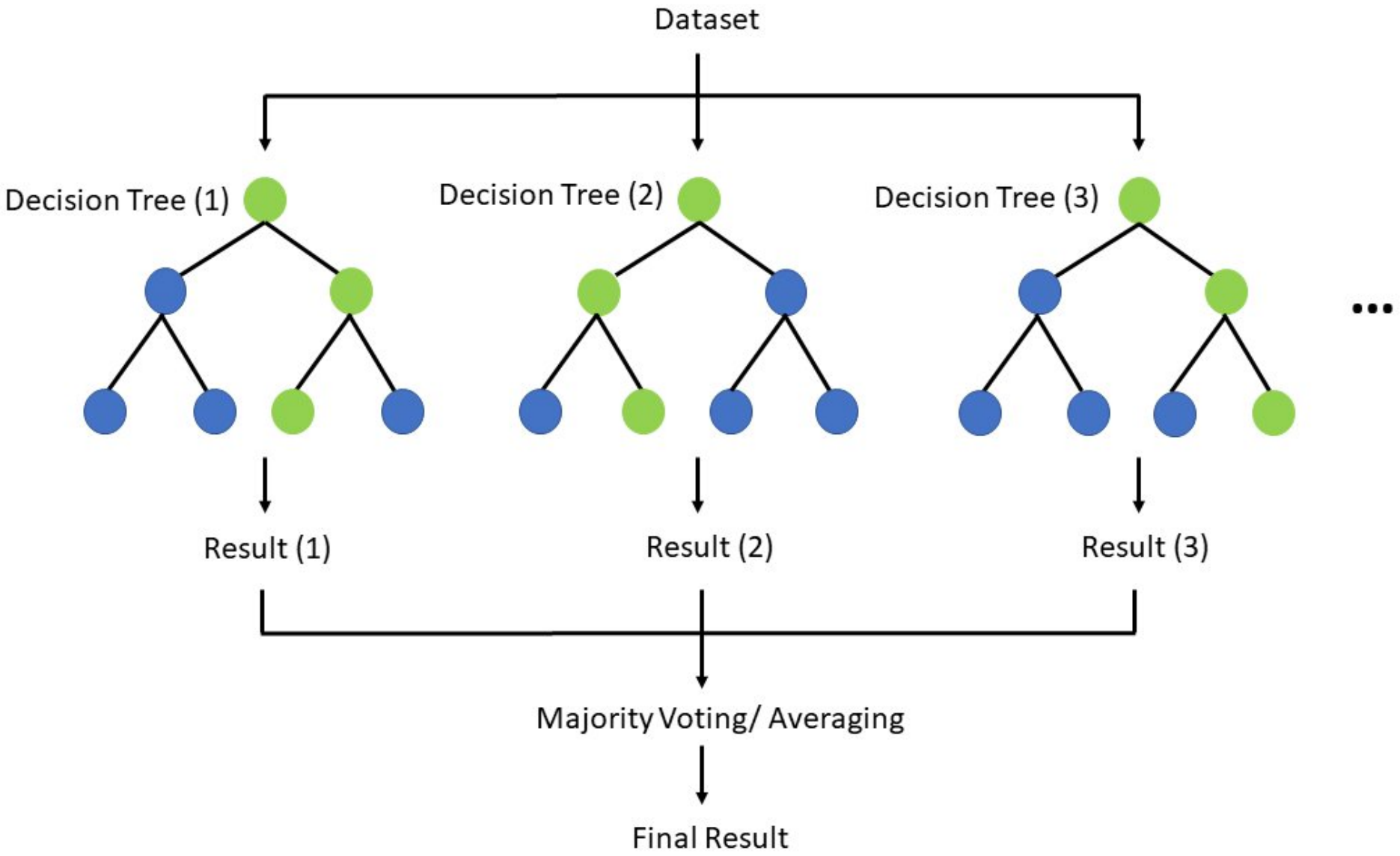
Target: nominal
Predictor: numeric

$$\ln\left(\frac{groupA}{non.groupA}\right) = \beta_0 + \beta_1 x_1 + \ldots$$

Objective: Maximize

$$L = \prod_{k:\, y_k=1} p_k \prod_{k:y_k=0} (1-p_k)$$

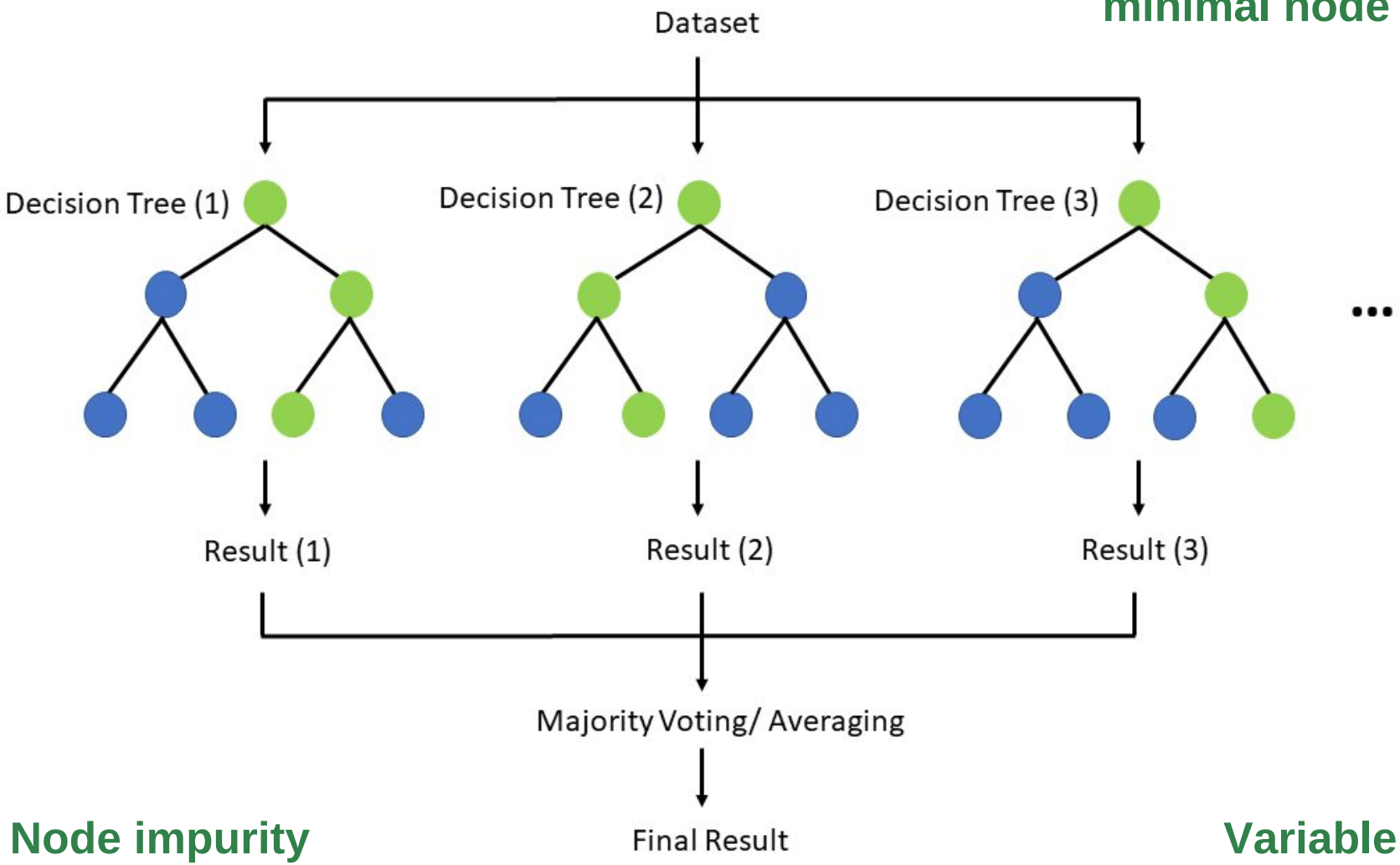
Random forest



Random forest

Regression
&
Classification

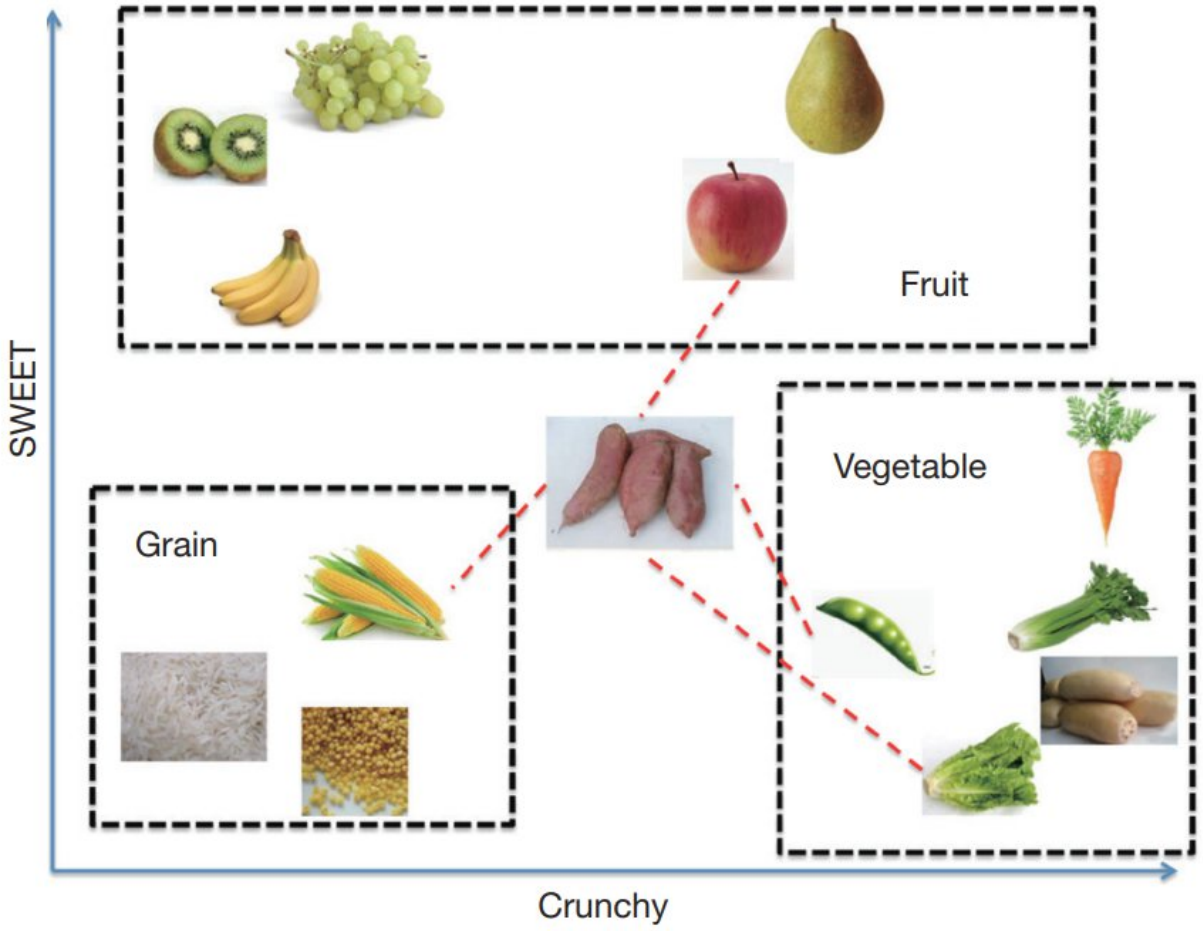
Variable selection
#trees
minimal node size



Node impurity
Bagging
Out-of-bag error

Variable
importance

K-Nearest Neighbour



Distance

How many voters?

k

Handling the data

- **Train & test** as separate data frames
- Test → reality

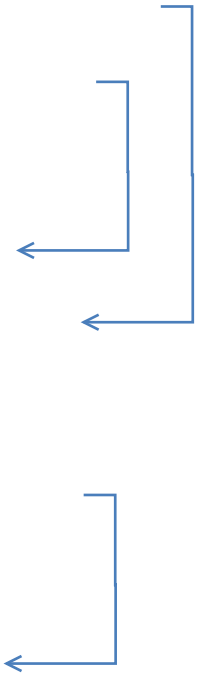
  No oversampling or imbalance corrections !
- **Data leakage**

  Data-driven steps: tuning, CV, imputations...

```
#1 split data - rsample
initial_split(data, prop = 0.8) %>% training()
                                    testing()

#2 preprocess - recipes
recipe <- recipe(y ~ x, data = data) %>%
    step_...() %>%
    step_...() ...

#3 model choose 1) model, 2) engine, 3) mode
linear_reg() or rand_forest() %>%
    set_engine() %>%
    set_mode()
```

Mode	Backend package	Notes
regression	**stats::lm**	The base R linear model (lm()), ordinary least squares regression.
regression	**stats::glm**	Generalized linear model, allows different families (e.g. Gaussian, binomial).
regression	**glmnet::glmnet**	Regularized regression (LASSO, ridge, elastic net). Efficient for high-dimensional data.

```
#4 tuning
grid_regular()
vfold_cv(training_set)
tune_grid(object = workflow_object,
          resamples =  ,
          grid =  ,
          metrics = metric_set())
select_best()
finalize_workflow(workflow_object,
                  parameters = )

#5 fitting
fit(model) # uses formula
fit_xy(model, x = , y = )
last_fit()
```

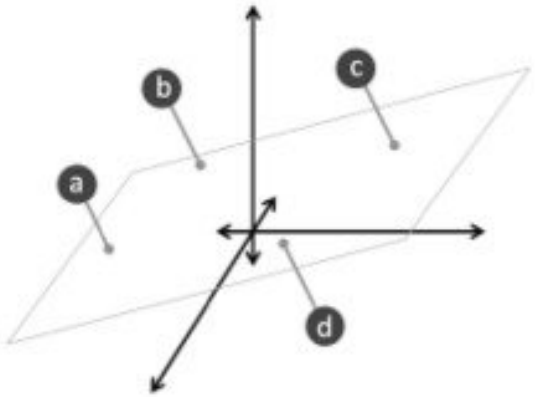


```r
#4 tuning
grid <- grid_regular(penalty(range = c(-3, 3)))
folds <- vfold_cv(training_set, v = 10)
tuning_results <- tune_grid(object = workflow_object
                            resamples = folds,
                            grid = grid,
                            metrics = metric_set(mae))
best_pars <- select_best(tuning_results, metric = "mae")
final_wf <- finalize_workflow(workflow_object, best_pars)


#5 fitting
fit(model, y ~ x, data)
fit_xy(model, x = data[pred1, pred2...], y = data[outcome])
last_fit(final_wf, split = data_split, metrics = metric_set(mae))
```

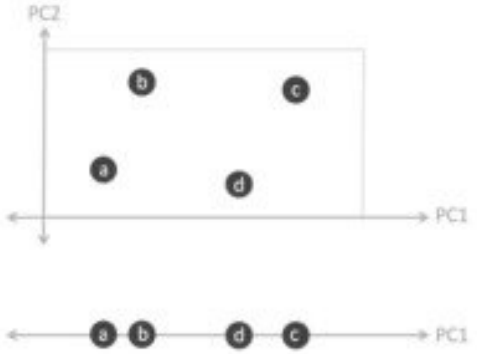
Q: Data leakage

1) Scale features with zero mean and unit variance
2) Dimension reduction, explaining 90% variance
3) Fit ridge regression
4) Tune λ using CV

Dimensionality reduction



(a) An example of the original features in a 3D space.



(b) Principal components in 2D and 1D space.

https://www.tidymodels.org/learn/statistics/k-means/kmeans.gif

Dataset	Type	Source
abalone.data	Regression	UCI ML Repo
agaricus-lepiota.data	Classification	UCI ML Repo
hepatitis	Classification	UCI ML Repo
Breast Cancer Wisconsin (Original)	Classification	UCI ML Repo

HANDS-ON

REGRESSION

Abalone Dataset

Predict age of abalone from physical measurements.
The age of abalone is determined by cutting the shell through the cone, staining it, and counting the number of rings through a microscope -- a boring and time-consuming task. Other measurements, which are easier to obtain, are used to predict the age.

From the original data examples with missing values were removed (the majority having the predicted value missing), and the ranges of the continuous values have been scaled for use with an ANN (by dividing by 200).

Has Missing Values? No

Abalone Age Classifier

Abalone Dataset

Variable Name	Role	Type	Description	Units
Sex	Feature	Categorical	M, F, and I (infant)	
Length	Feature	Continuous	Longest shell measurement	mm
Diameter	Feature	Continuous	perpendicular to length	mm
Height	Feature	Continuous	with meat in shell	mm
Whole_weight	Feature	Continuous	whole abalone	grams
Shucked_weight	Feature	Continuous	weight of meat	grams
Viscera_weight	Feature	Continuous	gut weight (after bleeding)	grams
Shell_weight	Feature	Continuous	after being dried	grams
Rings	Target	Integer	+1.5 gives the age in years	

Is this mushroom edible?

Attribute	
cap-shape	stalk-surface-above-ring
cap-surface	stalk-surface-below-ring
cap-color	stalk-color-above-ring
bruises	stalk-color-below-ring
odor	veil-type
gill-attachment	veil-color
gill-spacing	ring-number
gill-size	ring-type
gill-color	spore-print-color
stalk-shape	population
stalk-root	habitat

Hepatitis Survival Prediction

Attribute	Attribute
Class: DIE, LIVE	ALBUMIN: 2.1, 3.0, 3.8, 4.5, 5.0, 6.0
AGE: 10, 20, 30, 40, 50, 60, 70, 80	PROTIME: 10, 20, 30, 40, 50, 60, 70, 80, 90
SEX: male, female	HISTOLOGY: no, yes
STEROID: no, yes	BILIRUBIN: 0.39, 0.80, 1.20, 2.00, 3.00, 4.00
ANTIVIRALS: no, yes	ALK PHOSPHATE: 33, 80, 120, 160, 200, 250
FATIGUE: no, yes	SGOT: 13, 100, 200, 300, 400, 500
MALAISE: no, yes	SPIDERS: no, yes
ANOREXIA: no, yes	ASCITES: no, yes
LIVER BIG: no, yes	VARICES: no, yes
LIVER FIRM: no, yes	SPLEEN PALPABLE: no, yes

Breast Tumor, Benign or Malignant?

Attribute	**Domain**
Sample code number	id number
Clump Thickness	1 - 10
Uniformity of Cell Size	1 - 10
Uniformity of Cell Shape	1 - 10
Marginal Adhesion	1 - 10
Single Epithelial Cell Size	1 - 10
Bare Nuclei	1 - 10
Bland Chromatin	1 - 10
Normal Nucleoli	1 - 10
Mitoses	1 - 10
Class:	(2 for benign, 4 for malignant)

Conflicts and Dilemmas

- Multiplicity of good models

  In my experiments with trees, if the training set is perturbed only slightly, say by removing a random 2–3% of the data, I can get a tree quite different from the original but with almost the same test set error. – Breiman, 20

- Small n – large p
- Simplicity vs accuracy
- Curse of dimensionality

  Complete info vs scarcity

„In treating processes of high dimension, involving large quantities of data, complete information is as much of a handicap as a scarcity of information.“
Bellman, 1959

Further Reading

- building R modelling package - video https://canal.uned.es/video/5dd25b9f5578f275e407dd88
- Machine Learning for Biostatistics Module 1 - Bookdown with Video interludes - https://bookdown.org/content/30d75162-d57a-42d1-b26f-77d5c56b20a6/
- Random forests and variable importance - https://proceedings.neurips.cc/paper_files/paper/2013/file/e3796ae838835da0b6f6ea37bcf8bcb7-Paper.pdf
- Project report on predicting abalone age - https://www.researchgate.net/publication/377565848_Predicting_the_age_of_Abalones
- For more on ridge and lasso regression - https://bookdown.org/ssjackson300/Machine-Learning-Lecture-Notes/the-lasso.html
- Tidy Modeling with R Book Club - https://r4ds.github.io/bookclub-tmwr/
- Leo Breiman "Statistical Modeling: The Two Cultures (with comments and a rejoinder by the author)," Statistical Science, Statist. Sci. 16(3), 199-231, (August 2001)

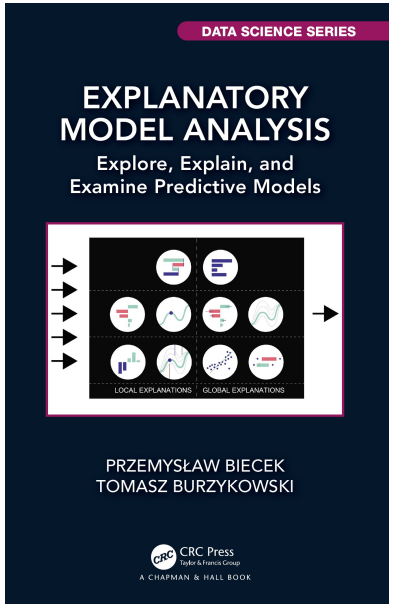


Image sources

- Nested AI/ML/DL & Neural network – Sarker I. H. (2021). Deep Learning: A Comprehensive Overview on Techniques, Taxonomy, Applications and Research Directions. SN computer science, 2(6), 420. https://doi.org/10.1007/s42979-021-00815-1
- ML Phases, Dim. Reduction – Sarker I. H. (2021). Machine Learning: Algorithms, Real-World Applications and Research Directions. SN computer science, 2(3), 160. https://doi.org/10.1007/s42979-021-00592-x
- Random forest – TseKiChun @wikimediaCommons, CC BY-SA 4.0
- KNN – Zhang, Zhongheng. (2016). Introduction to machine learning: K-nearest neighbors. Annals of Translational Medicine. 4. 218-218. 10.21037/atm.2016.03.37.