From Traditional Statistical Models to Machine Learning: Choosing the Approach to Fit the Research Question

JAMIE E COLLINS, PHD
VERITY/BRIGHAM COURSE IN RHEUMATOLOGY CLINICAL RESEARCH

APRIL 4, 2024

What is the difference between machine learning and statistical modeling?

“*The short answer is: None. They are both concerned with the same question: how do we learn from data?*”

– Dr. Larry Wasserman, Professor of Statistics and Data Science in the Department of Statistics and Data Science and in the Machine Learning Department at Carnegie Melon

Outline

1. **Background: Methods for Learning from Data**

   unsupervised, semi-supervised, supervised

2. **To Explain or to Predict?**

   What is the question?

3. **Principles of Risk Prediction**

   Best practices

4. **Methodology**

   Statistical Modeling to Machine Learning to Artificial Intelligence

Methods for Learning from Data

Supervised

Labeled outcomes or classes

Unsupervised

No labels/ annotations

Semi-supervised

Some labels/outcomes

Methods for Learning from Data: Supervised Methods

- Labeled outcomes or classes
- Focus may be on best prediction algorithm, on which variables (features) are most closely associated with outcome, or on assessing whether outcomes differ between exposure groups

Predicting who is likely to achieve remission among patients with rheumatoid arthritis starting tocilizumab monotherapy

Predicting who is likely to need total joint replacement among patients with osteoarthritis

Assessing associations between environmental exposures and care fragmentation

- Example methods: linear regression, logistic regression, random forest, support vector machines

Methods for Learning from Data: Unsupervised Methods

- No labels/annotations
- Goal is to uncover hidden structure/patterns in the dataset

Assessing medication adherences trend over time in patients with Systemic Lupus Erythematosus

Investigating osteoarthritis endotypes through clustering of biochemical marker data

Describing patterns of pain sensitization among patients with knee osteoarthritis

- Data reduction: principal component analysis, factor analysis
- Clustering: model-based cluster analysis, K-means

Methods for Learning from Data: Semi-supervised Methods

- Combination of Supervised and Unsupervised approaches
- Outcomes/classes are labeled for some part of the dataset
- Analysis usually done in steps with supervised followed by unsupervised or vice versa
- Examples: often used in natural language processing

Methods for Learning from Data

Supervised

Labeled outcomes or classes

Unsupervised

No labels/ annotations

Semi-supervised

Some labels/outcomes

What is the Question?

Supervised Methods
To explain, or to predict?

TO EXPLAIN

- We use a mathematical model to formalize the relationship between variables.
- We focus on obtaining unbiased estimates of the associations between our independent and dependent variables.
- Goal may be causal inference: does our predictor have a causal effect on outcome?

TO PREDICT

- We use a mathematical model to make predictions about the dependent variable.
- We focus on obtaining the optimal prediction based on a combination of available variables.
- Goal may be to reliability predict outcomes for individuals.

Shmueli G. To explain or to predict?. Statistical science. 2010.
https://statisticalhorizons.com/prediction-vs-causation-in-regression-analysis

Example: To Explain

RHEUMATOLOGY

Rheumatology 2022;61:1430–1439
doi:10.1093/rheumatology/keab535
Advance Access publication 10 July 2021

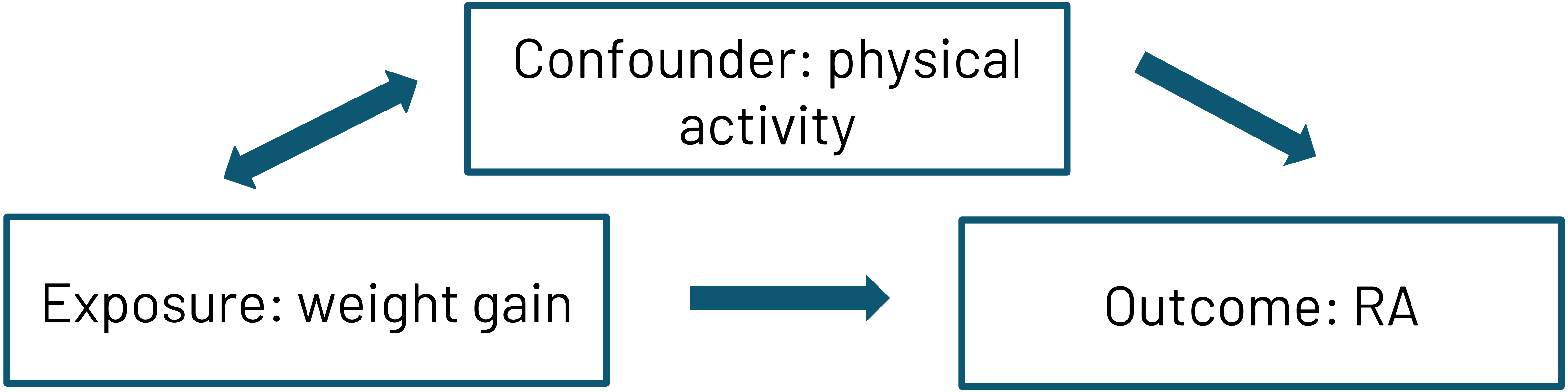
Original article

Long-term weight changes and risk of rheumatoid arthritis among women in a prospective cohort: a marginal structural model approach

Nathalie E. Marchand [1], Jeffrey A. Sparks [1], Susan Malspeis[1], Kazuki Yoshida[1], Lauren Prisco[1], Xuehong Zhang[2,3], Karen Costenbader[1], Frank Hu[2,3,4], Elizabeth W. Karlson[1] and Bing Lu[1]

Objective: To examine the association of long-term weight change with RA risk in a large prospective cohort study.

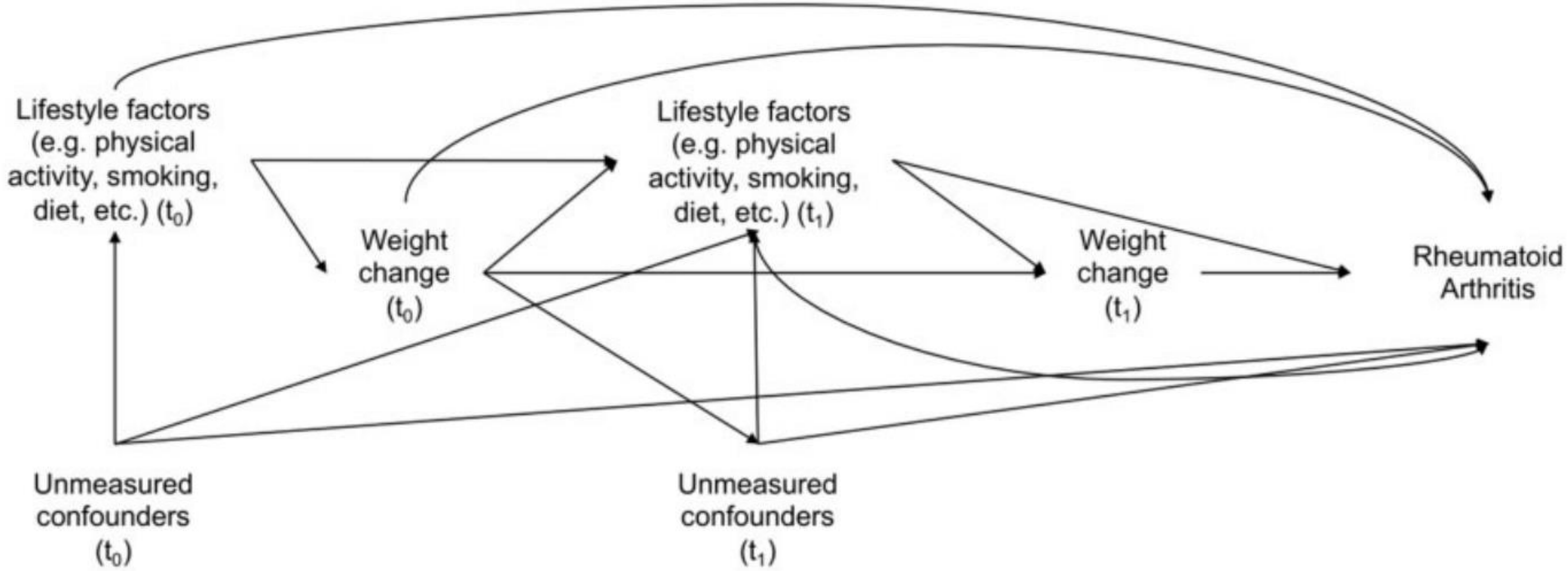
Example: To Explain



"An analysis of weight change and RA risk in prospective cohort studies may be limited by time-varying confounders, which may themselves be affected by previous weight change, that lie on the causal pathway between weight change and RA."

Example: To Explain

Fig. 1 Directed acyclic graph showing the relationships between weight change and rheumatoid arthritis in the presence of time-varying confounding



Example: To Explain

- *Using an MSM approach in our analyses allowed us to deal with the time-varying confounding. In addition, by conducting our analyses in the 'pseudo-population' we were able to statistically approximate the study conditions of a randomized controlled trial (RCT) in which we could specify* ***hypothetical weight-change interventions*** *of interest.*

Rheumatology key messages

- Long-term weight gain during adult life may nearly quadruple rheumatoid arthritis risk in women.
- Rheumatoid arthritis risk increased starting with a weight gain of 2–10 kilograms from study baseline.

Marchand NE, et al. "Long-term Weight Changes and Risk of Rheumatoid Arthritis Among Women in a Prospective Cohort: A Marginal Structural Model Approach." *Rheumatology* (2020).

Example: To Predict

Vodencarevic *et al. Arthritis Research & Therapy* (2021) 23:67
https://doi.org/10.1186/s13075-021-02439-5

Arthritis Research & Therapy

RESEARCH ARTICLE Open Access

Advanced machine learning for predicting individual risk of flares in rheumatoid arthritis patients tapering biologic drugs

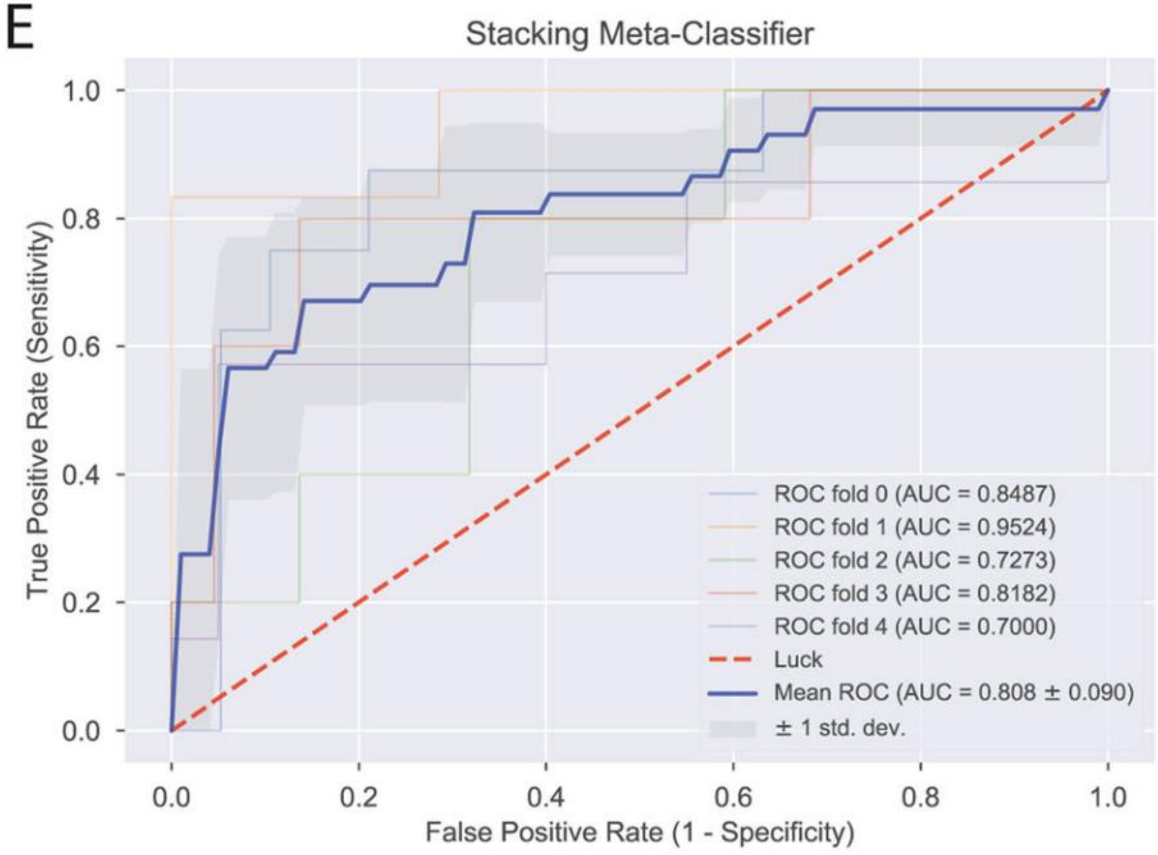
Asmir Vodencarevic[1†], Koray Tascilar[2,3†], Fabian Hartmann[2,3], Michaela Reiser[2,3], Axel J. Hueber[2,3,4], Judith Haschka[2,3,5], Sara Bayat[2,3], Timo Meinderink[2,3], Johannes Knitza[2,3], Larissa Mendez[2,3], Melanie Hagen[2,3], Gerhard Krönke[2,3], Jürgen Rech[2,3], Bernhard Manger[2,3], Arnd Kleyer[2,3], Marcus Zimmermann-Rittereiser[1], Georg Schett[2,3], David Simon[2,3*] and on behalf of the RETRO study group

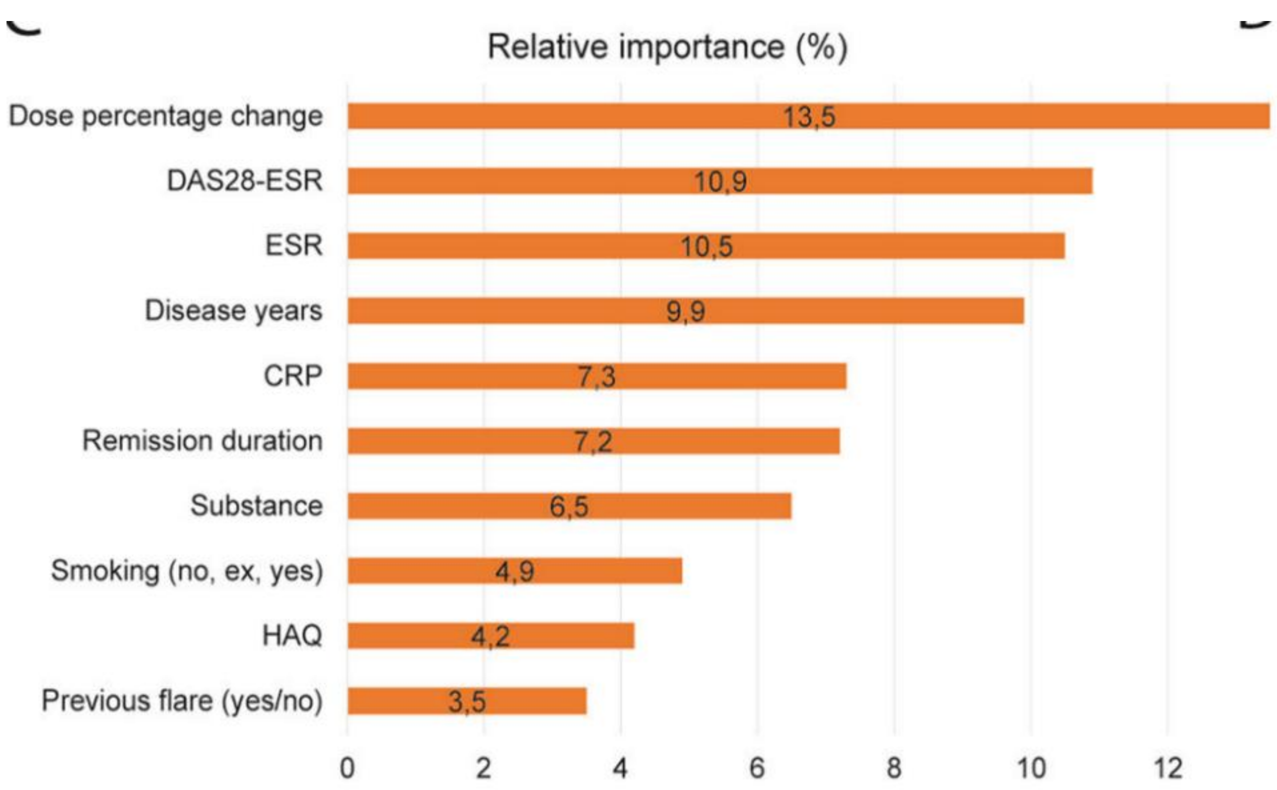
Objective: To assess the feasibility of building a model to estimate the individual flare probability in RA patients tapering bDMARDs.

Example: To Predict

- Used data from the REduction of Therapy in patients with rheumatoid arthritis in ongoing remission (RETRO) study. 135 visits from 41 patients
- Outcome: a binary indicator of whether a patient suffered a flare within 14 weeks after a given visit. 31 total flares.
- Predictors: patient characteristics, disease characteristics, medication data, laboratory data (n > 30)
- Analytic approach: ensemble machine learning model

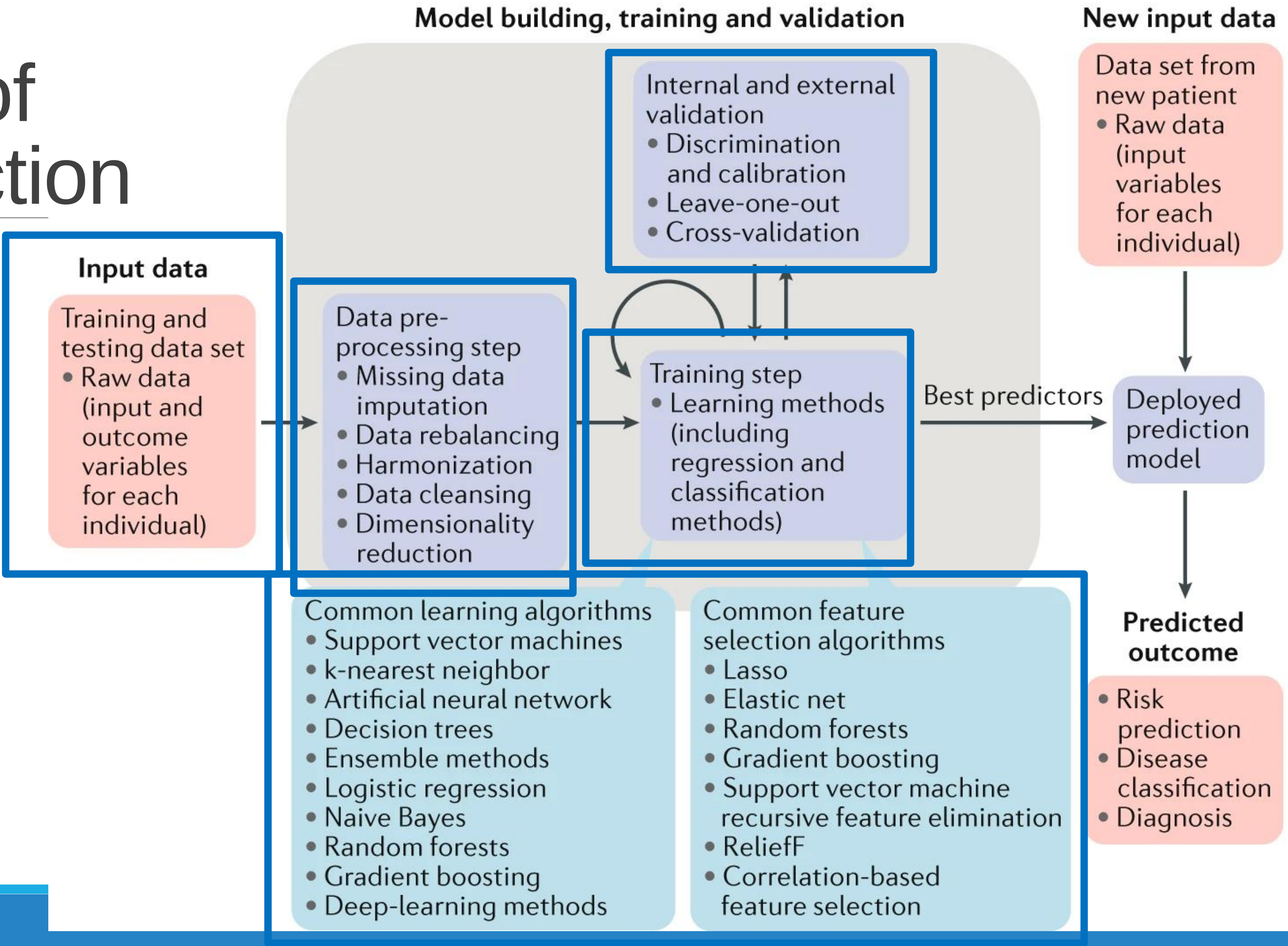
Example: To Predict





Principles of Risk Prediction

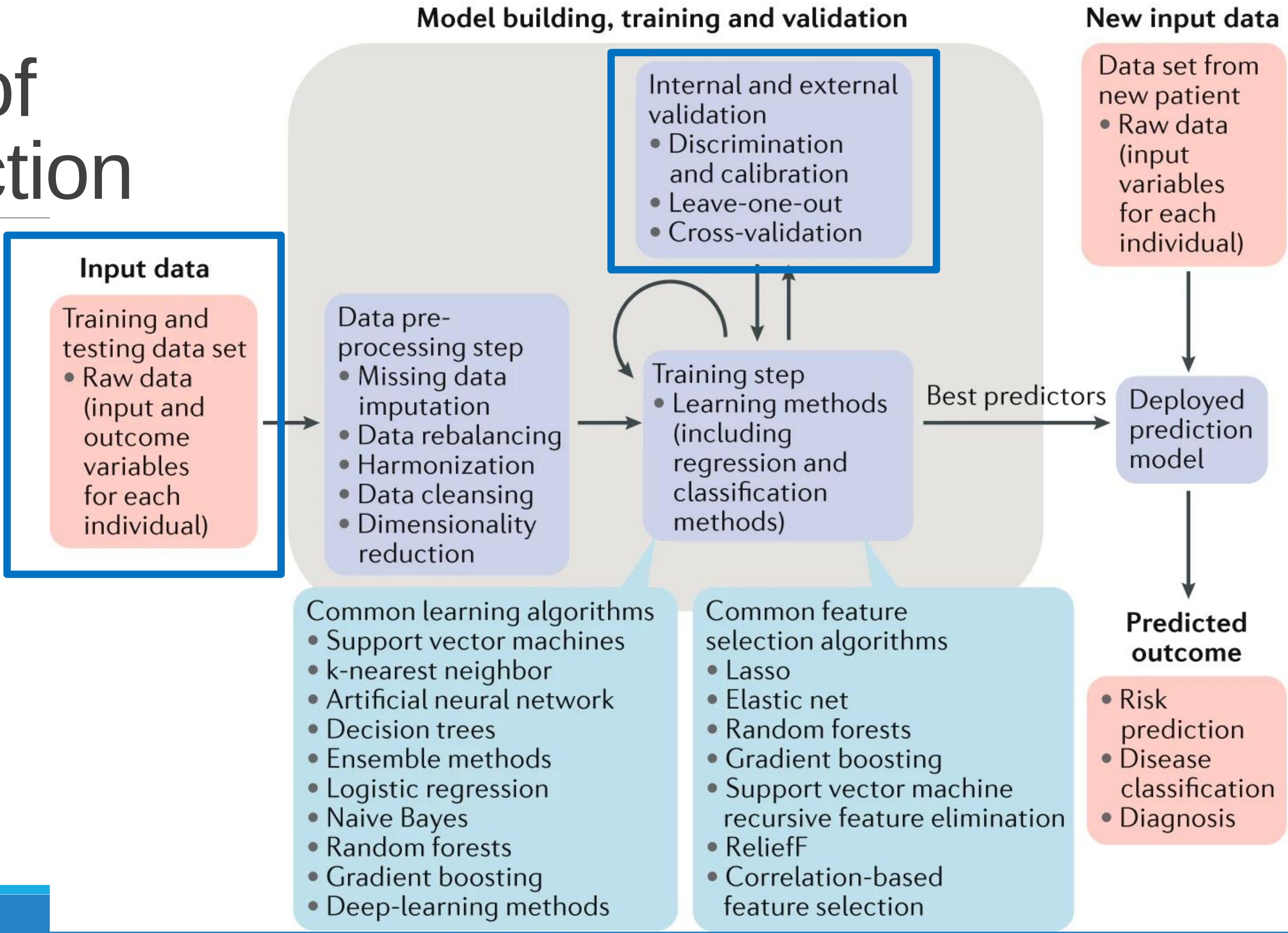
Principles of Risk Prediction



Jamshidi A, et al. Machine-learning-based patient-specific prediction models for knee osteoarthritis. *Nat Rev Rheumatol* (2019).

Principles of Risk Prediction

Training, testing, and validation



Jamshidi A, et al. Machine-learning-based patient-specific prediction models for knee osteoarthritis. *Nat Rev Rheumatol* (2019).

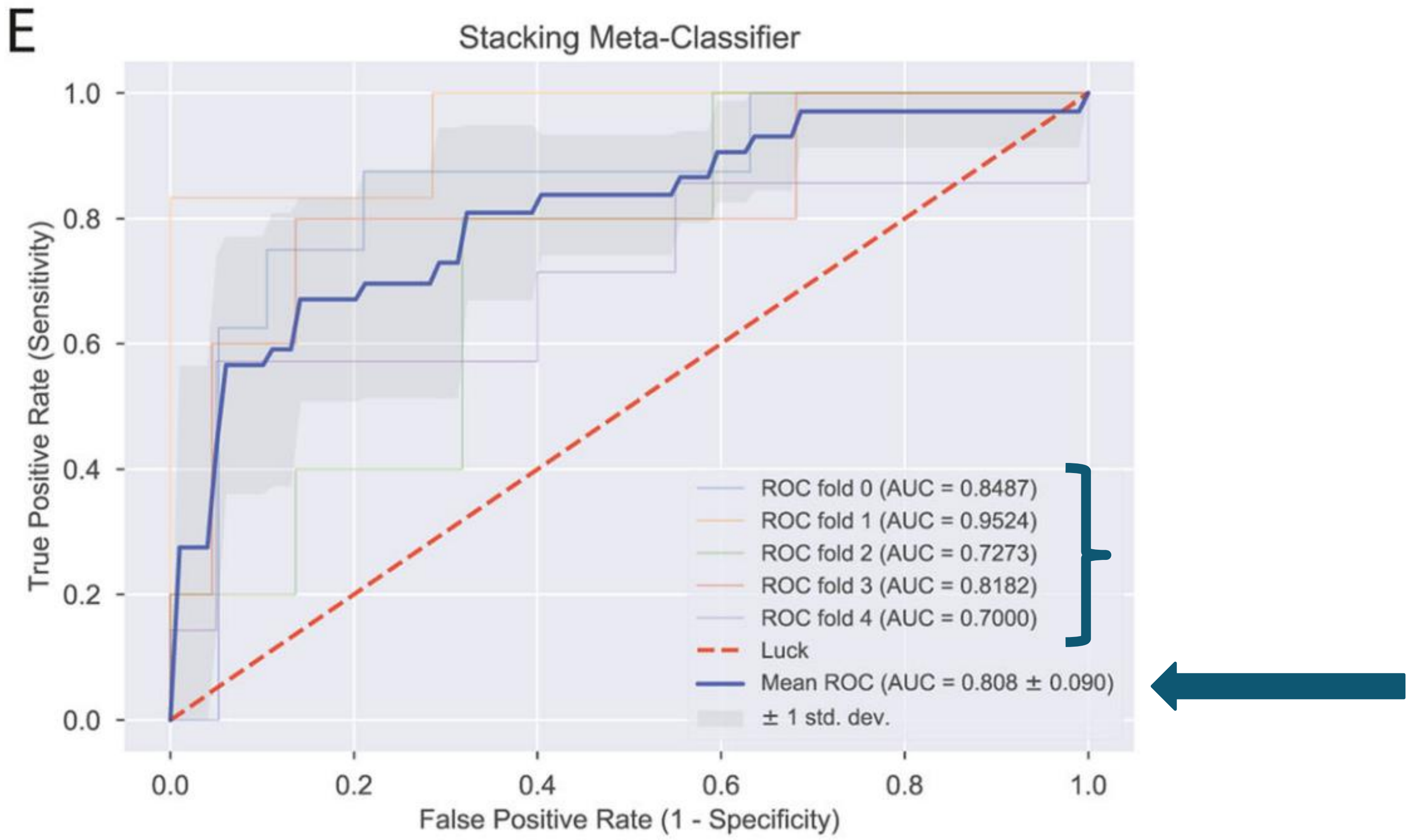
Principles of Risk Prediction
Input Data: Training and Testing

- The model should generalize to populations that were not included in the derivation sample. Overfitting is when the model captures random variation in the data.
- If number of predictors is greater than the number of observations, we can get perfect prediction ($p > n$).
- Model fits well in the dataset used to create the model, but how will it perform on “new” data?

Principles of Risk Prediction
Input Data: Training and Testing

- Training and testing datasets: hold out part of sample when model building
- Cross validation: Partition data into subsets, and hold out one subset for testing. Repeat until all subsets have been hold out and average over all subsets.
- Resampling procedures (e.g., bootstrap): resample (with replacement) from original dataset to compute optimism adjusted measures of predictive performance.
- External validation: test predictions in new dataset.

Principles of Risk Prediction
Input Data: Training and Testing

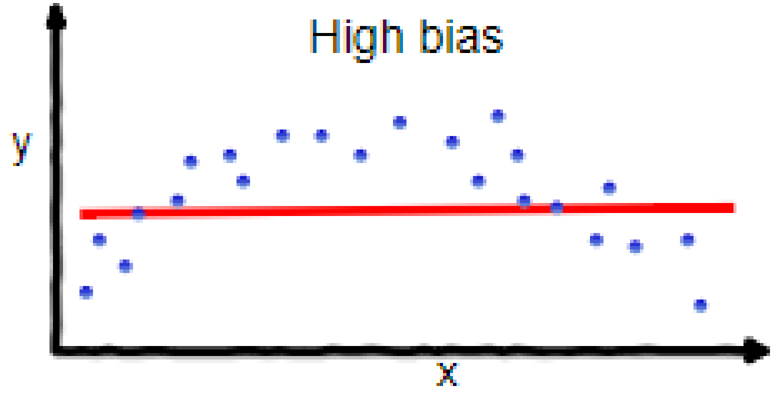


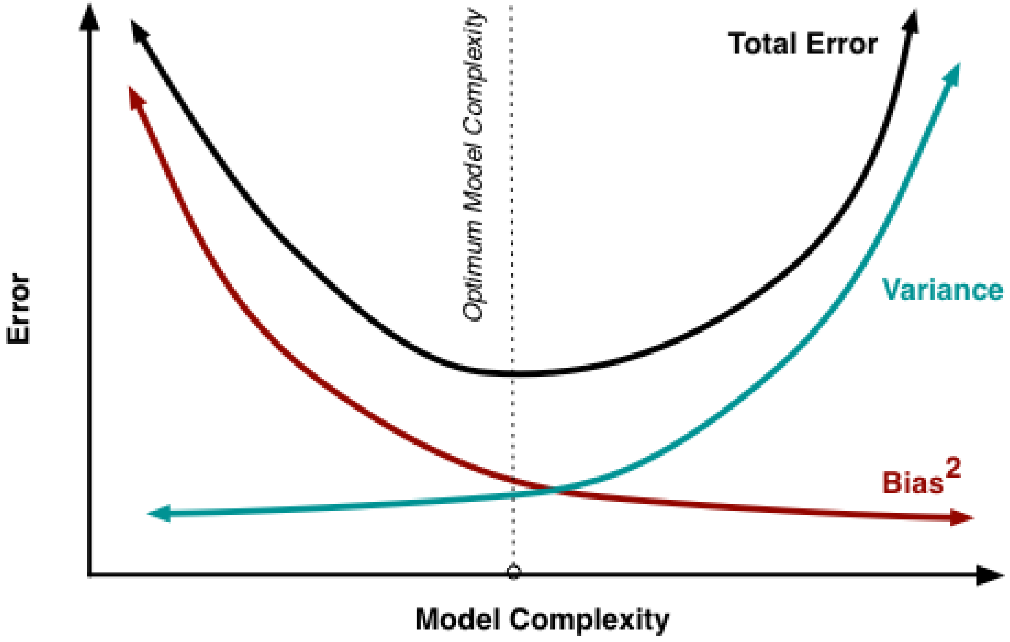
Vodencarevic A, et al. "Advanced machine learning for predicting individual risk of flares in rheumatoid arthritis patients tapering biologic drugs." *Arthritis research & therapy* (2021).

Principles of Risk Prediction
Bias-Variance Tradeoff

Bias

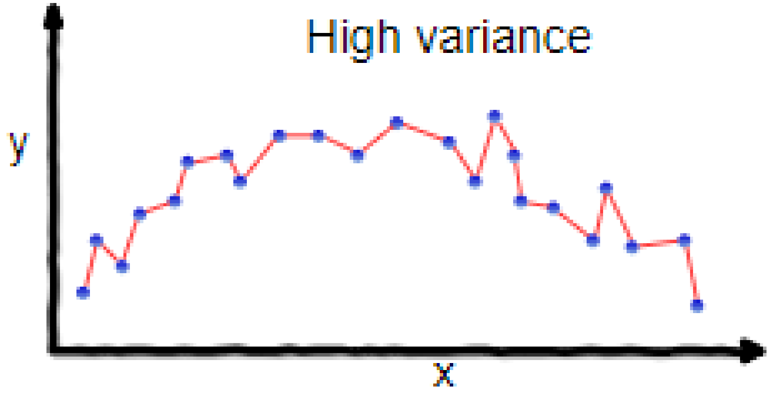
how close is our model to the true underlying model?





Variance

how do our predictions do on a new dataset?



Supervised Learning Algorithms

Supervised Learning Algorithms Logistic Regression

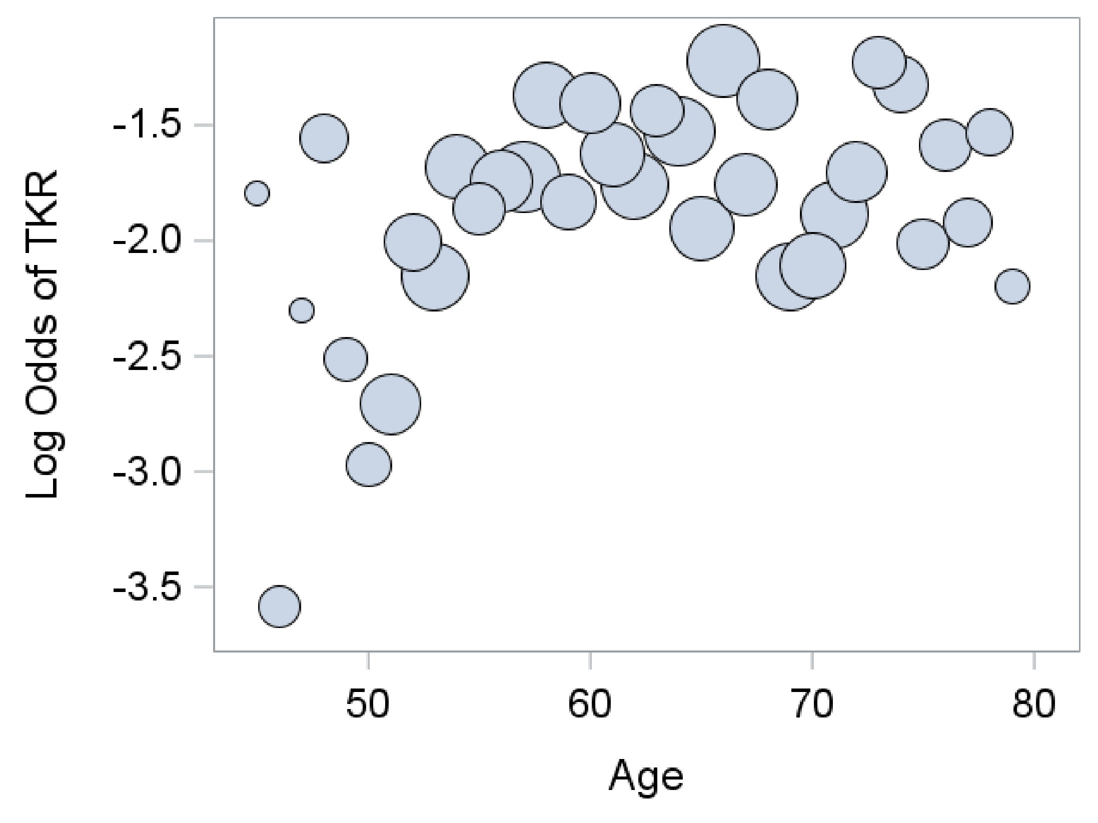
- Parametric regression model
  - Parametric: assume a form for the model
- log(Odds of Outcome) = $\beta 0 + \beta 1$*covariate
  - Log odds of outcome (logit) is a linear function of covariates
- The odds ratio quantifies association between predictor and outcome
- C-Statistic/AUC (Area under the ROC Curve) is a measure of model discrimination
  - 0.5 = coin flip, 1 = perfect prediction

Supervised Learning Algorithms Logistic Regression

- Among patients with knee osteoarthritis, is age associated with total knee replacement?

- log(Odds of Outcome) = β0 + β1*covariate
- log(Odds of TKR) = β0 + β1*age
- Recall: odds = p / (1-p)

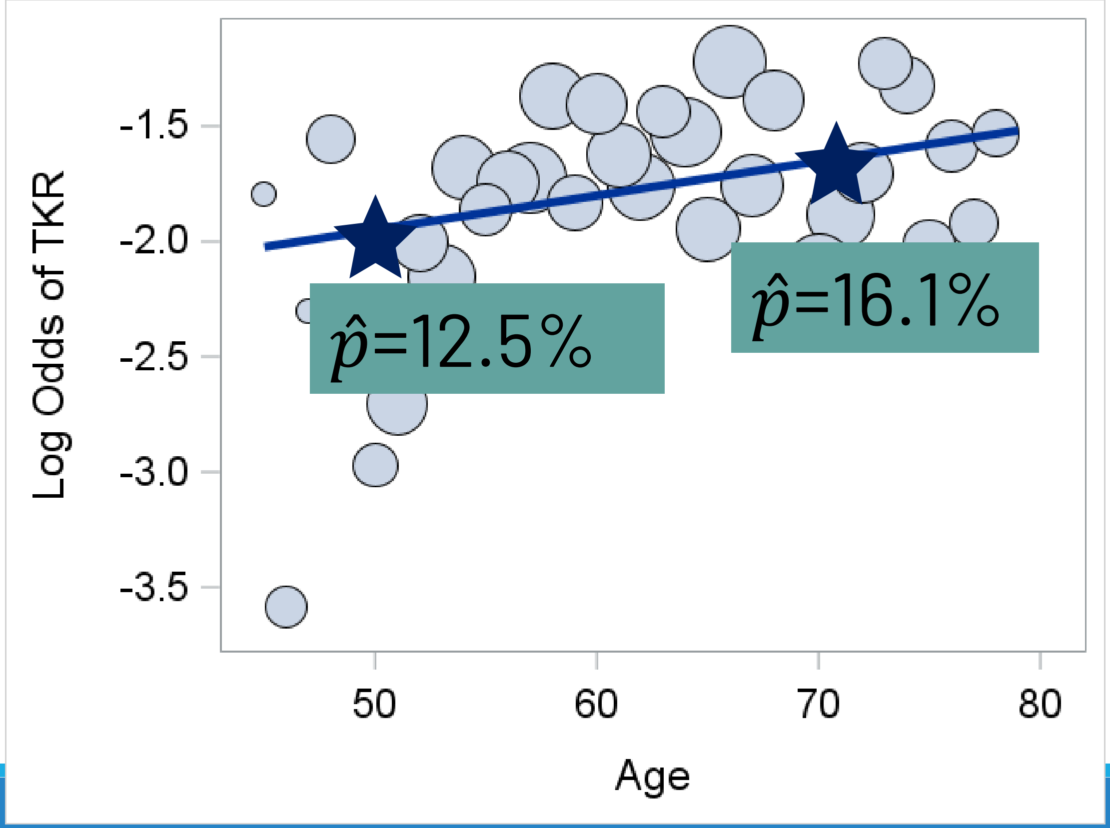
Log (odds)	probability
-3.5	3%
-1.5	18%
0	50%



Supervised Learning Algorithms
Logistic Regression

- Among patients with knee osteoarthritis, is age associated with total knee replacement?

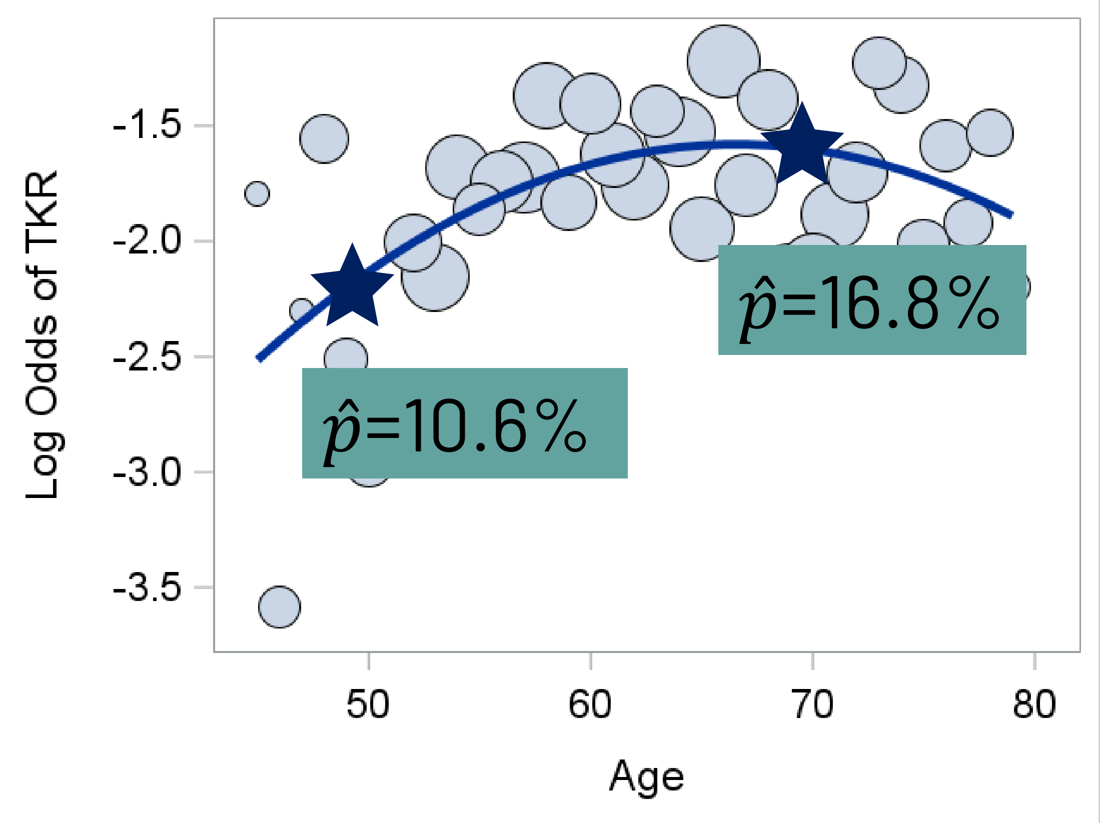
- log(Odds of Outcome) = β0 + β1*covariate
- log(Odds of TKR) = β0 + β1*age
- log(Odds of TKR) = -2.7 + 0.015*age
  - OR=1.015



Supervised Learning Algorithms
Logistic Regression

- Among patients with knee osteoarthritis, is age associated with total knee replacement?

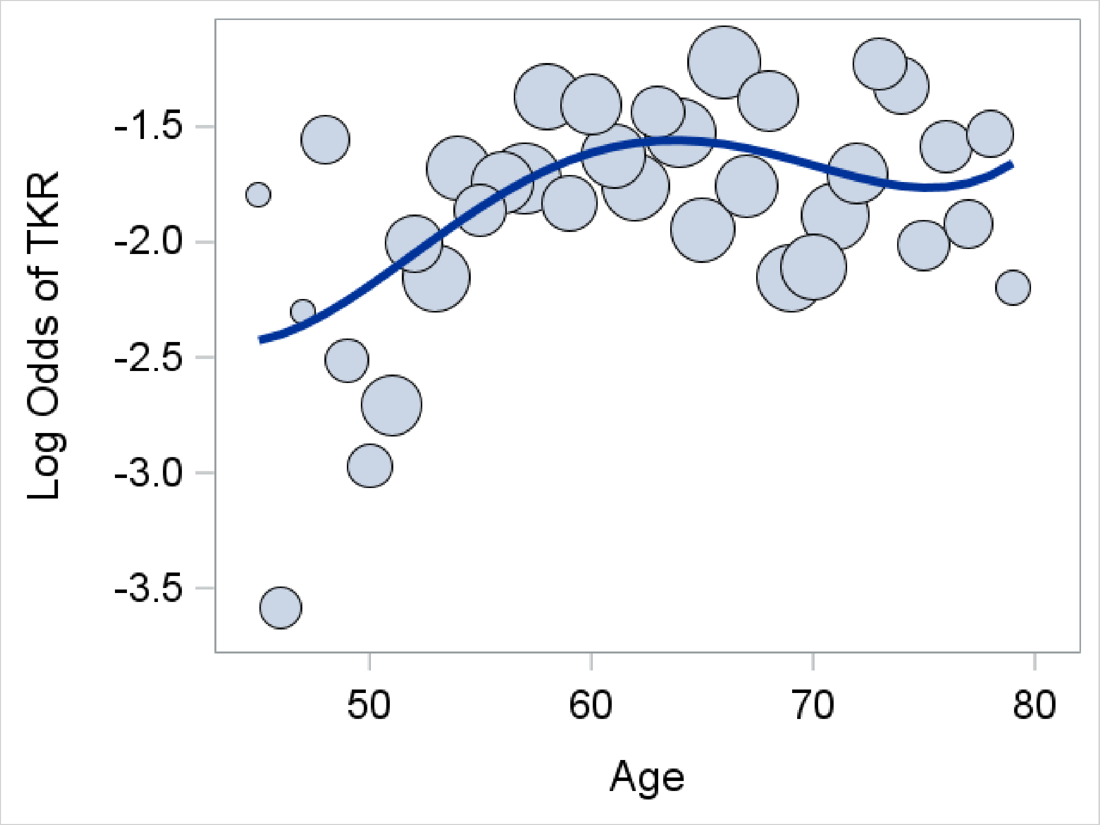
- log(Odds of Outcome) = β0 + β1*covariate + β2*covariate
- log(Odds of TKR) = β0 + β1*age + β2*age^2
- log(Odds of TKR) = -10.5 + 0.26*age – 0.002*age^2



Supervised Learning Algorithms
Logistic Regression

- Among patients with knee osteoarthritis, is age associated with total knee replacement?

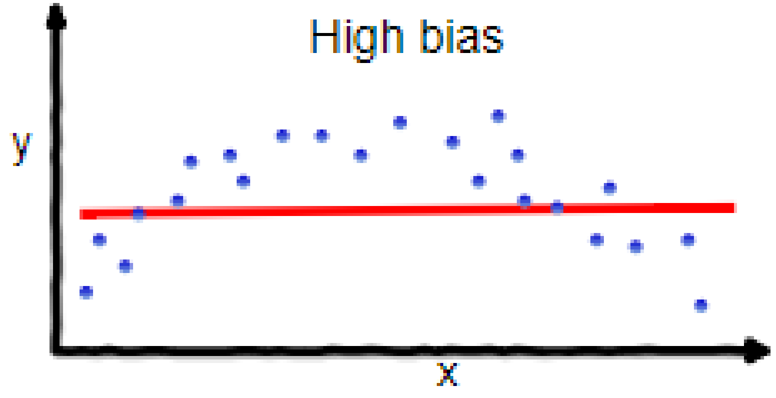
- log(Odds of Outcome) = β0 + β1*covariate + β2*covariate + β3*covariate + ….



Principles of Risk Prediction
Bias-Variance Tradeoff

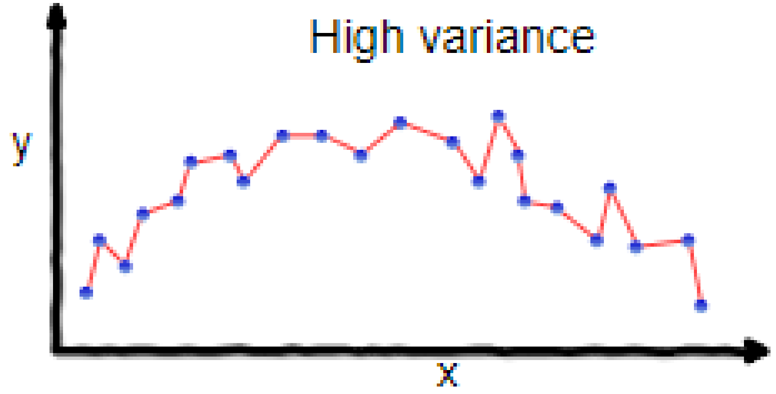
Bias

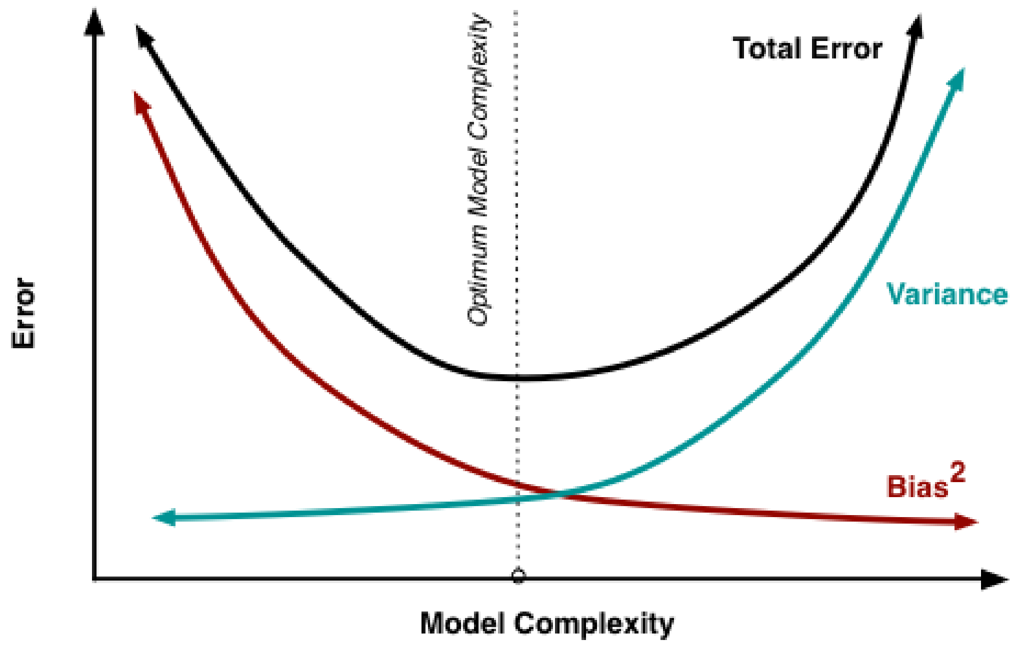
how close is our model to the true underlying model?



Variance

how do our predictions do on a new dataset?



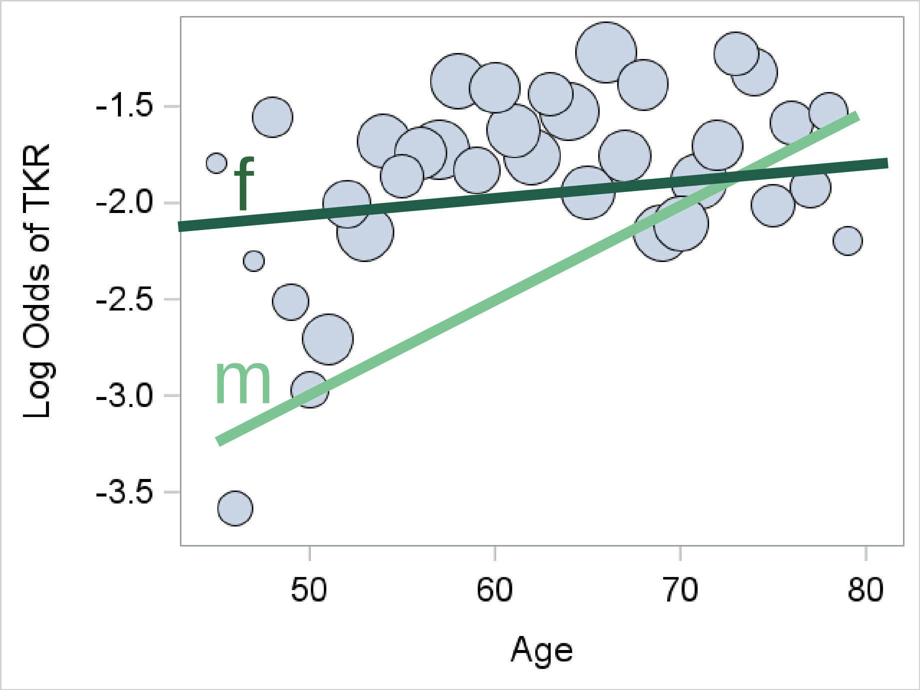


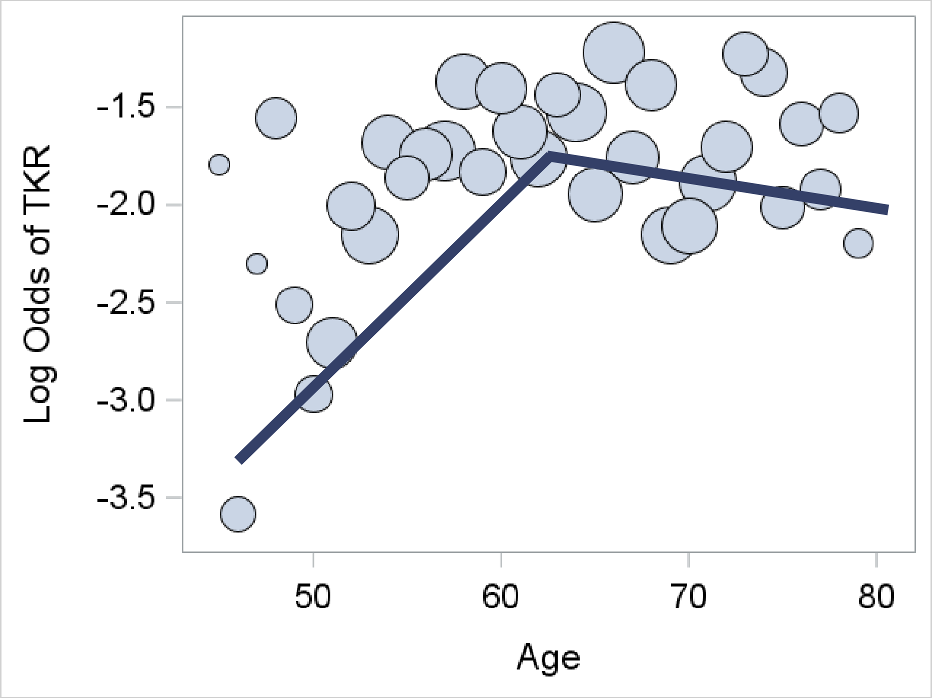
Supervised Learning Algorithms Parametric Models

- Parametric: assume a form for the model
- A regression equation describes the association between each parameter and the outcome
  - Y = intercept + beta*covariate
  - Log(odds of Y) = intercept + beta*covariate

Supervised Learning Alg Parametric Models

- Additional terms can be added to capture non-linear associations (splines, polynomials) or interactions between variables
  - Stratify by sex; Model association separately for age < 65 vs. age 65+
- With a large number of predictors it would be impossible to try all possible combinations, including interactions and non-linear associations





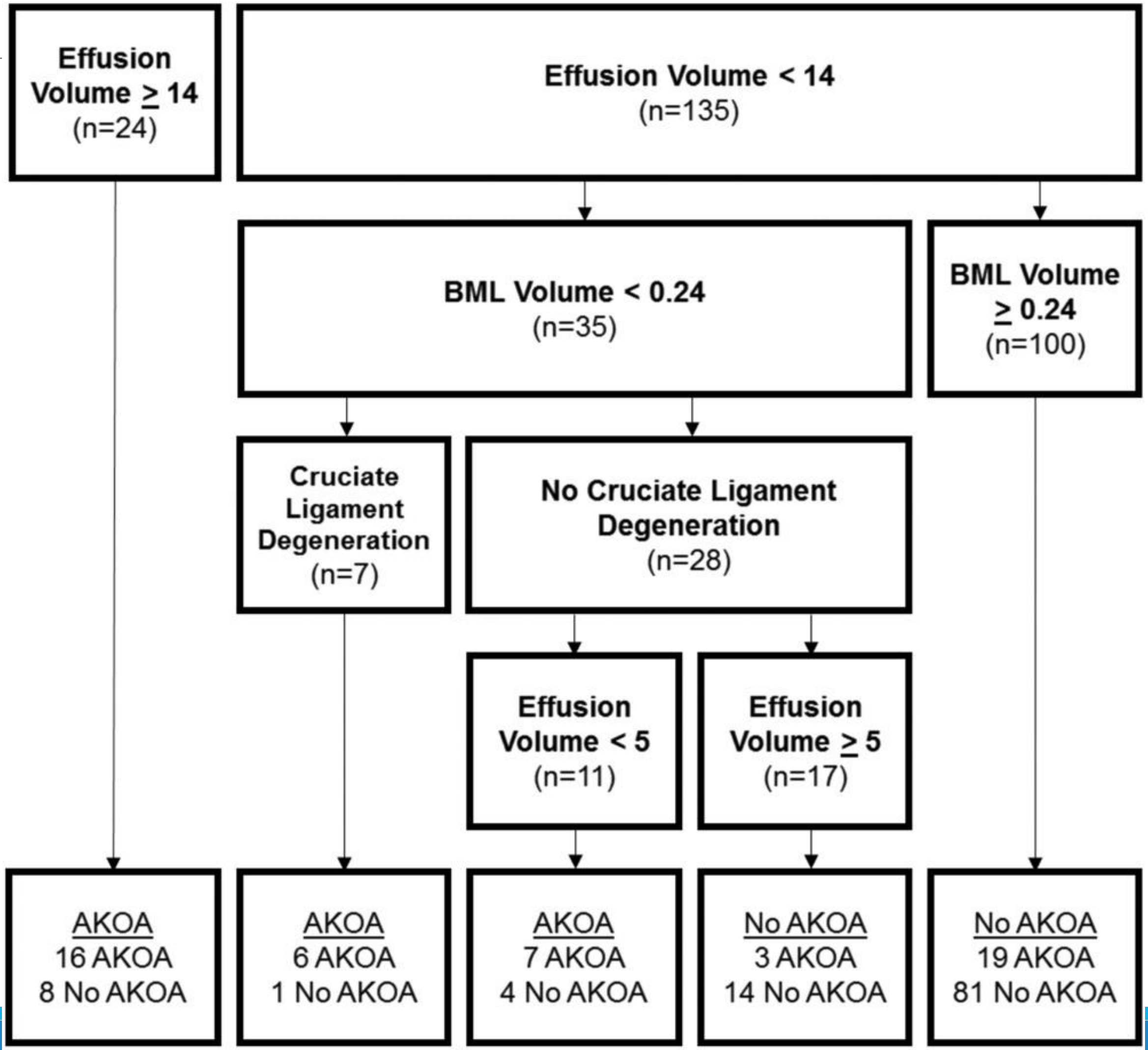
Supervised Learning Algorithms
CART

- Classification and Regression Trees (CART): Recursive partitioning: the data are partitioned into subsets – there is no regression equation (non-parametric)

  Explicitly models interactions between variables (effect of variable b depends on level of variable a)

  Results are intuitive and clinically interpretable – clear rules

- Example: Price et al. attempted to predict development of accelerated knee osteoarthritis from imaging data

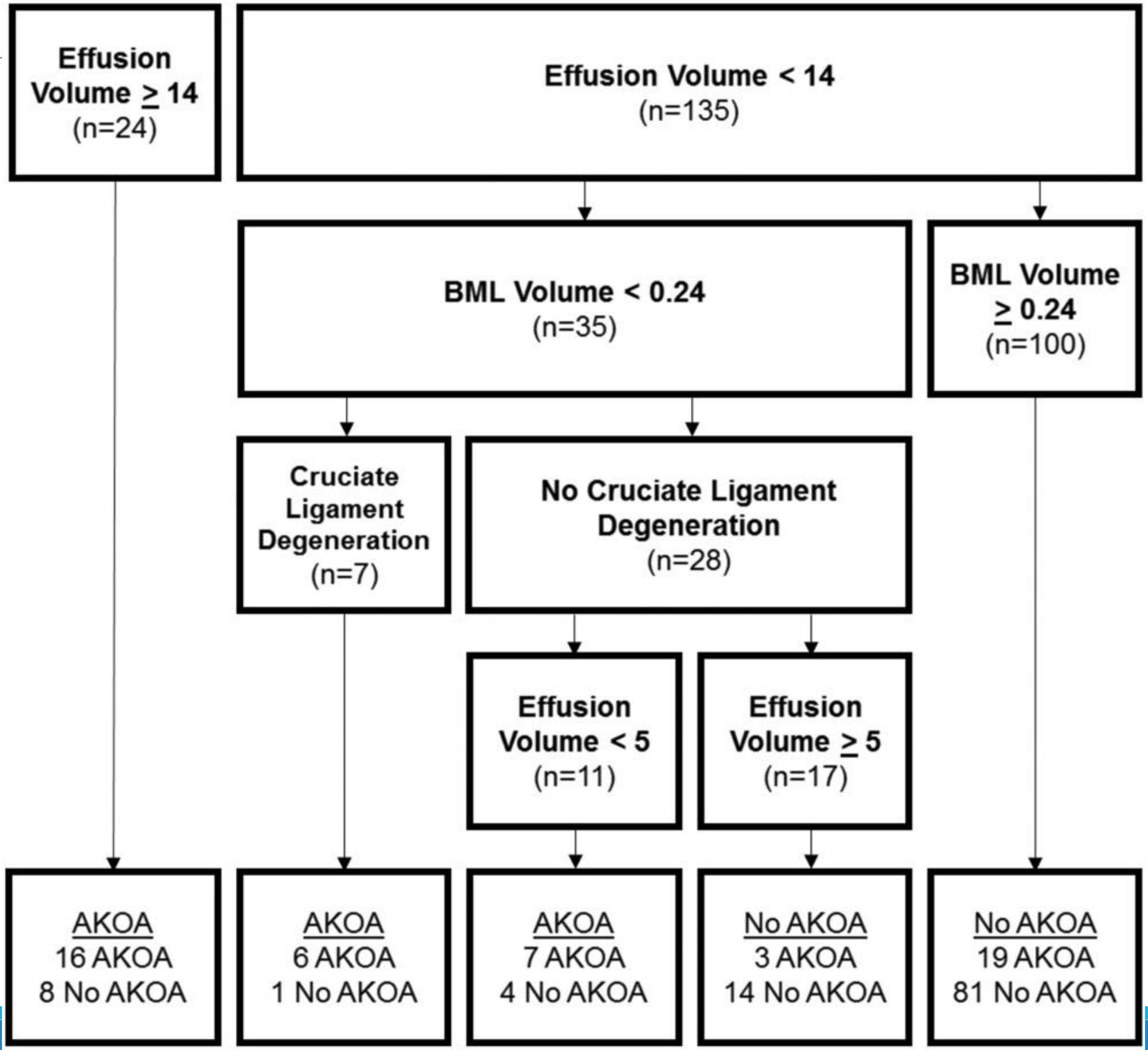


Price LL et al. "Role of Magnetic Resonance Imaging in Classifying Individuals Who Will Develop Accelerated Radiographic Knee Osteoarthritis." 2019.

Supervised Learning Algorithms
CART

Concerns/criticisms:

- Greedy approach can lead to over-fitting
- Highly dependent on input data → small changes to input data can lead to different trees
  - ↑variance – tends to overfit

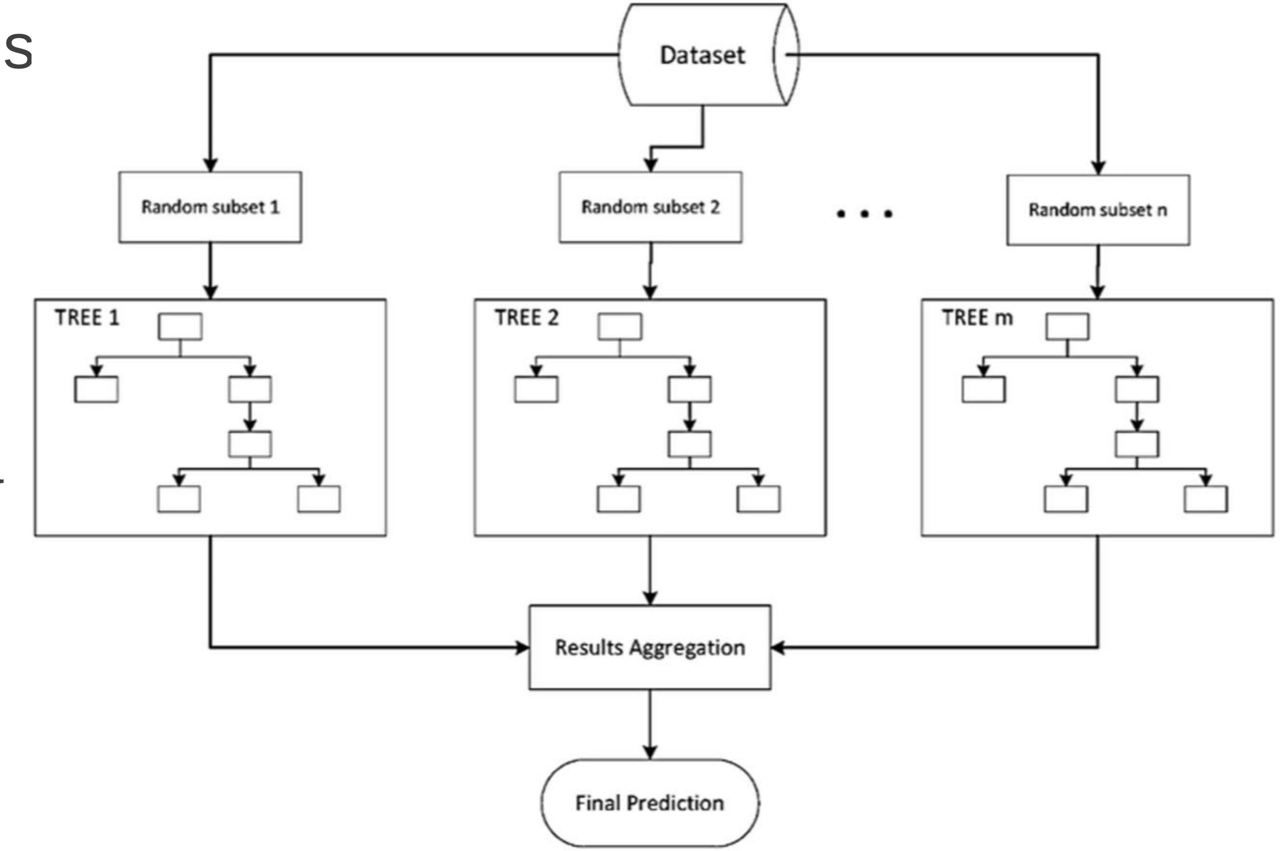


Price LL et al. "Role of Magnetic Resonance Imaging in Classifying Individuals Who Will Develop Accelerated Radiographic Knee Osteoarthritis." 2019.

Supervised Learning Algorithms
Ensemble Methods

- Combine information from multiple models improve model performance
  - *Develop* many prediction models
  - *Combine* to form a composite predictor

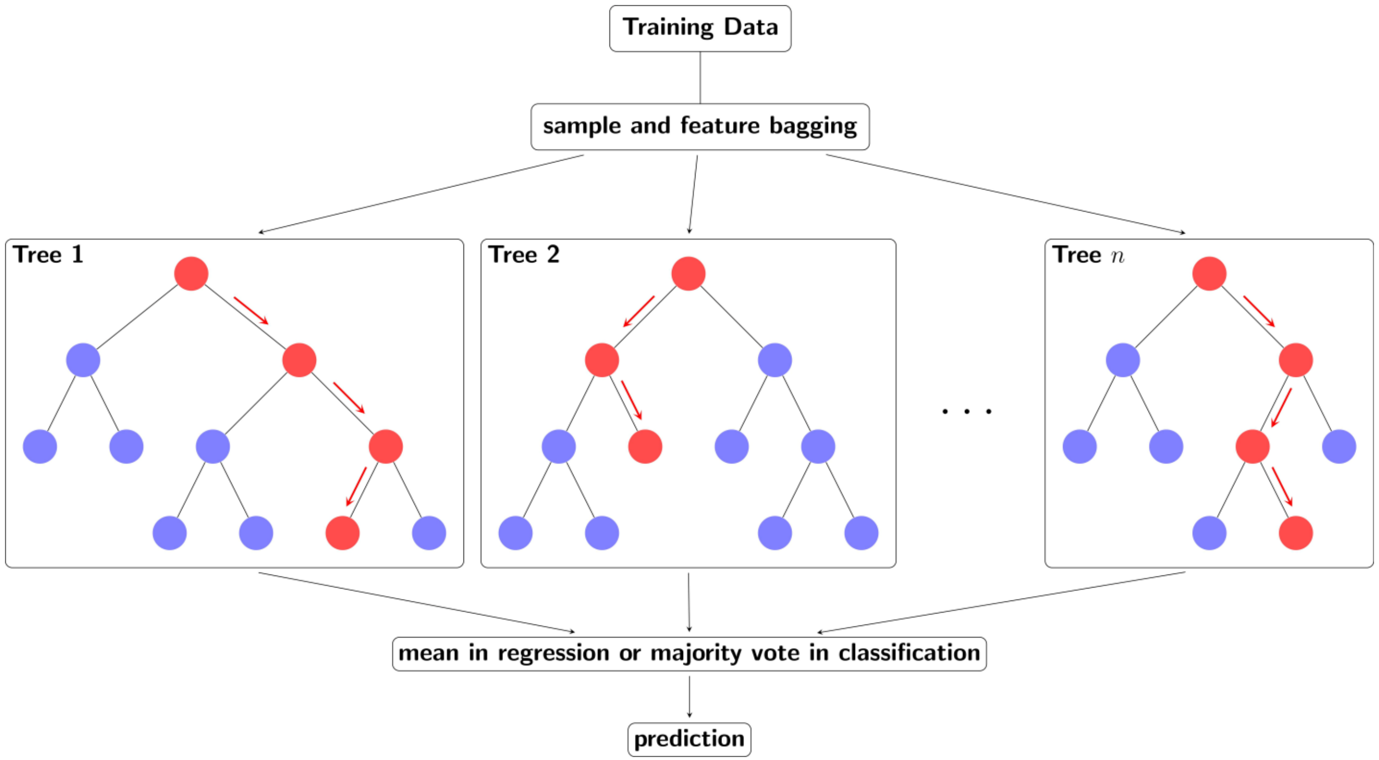
- Bagging (Bootstrap Aggregation):
  - Draw a bootstrap sample from the data (i.e., with replacement)
  - Fit a model to this sample
  - Get a prediction
  - Repeat
  - Average predicted values across all bootstrapped samples.



https://towardsdatascience.com/ensemble-methods-in-machine-learning-what-are-they-and-why-use-them-68ec3f9fef5f

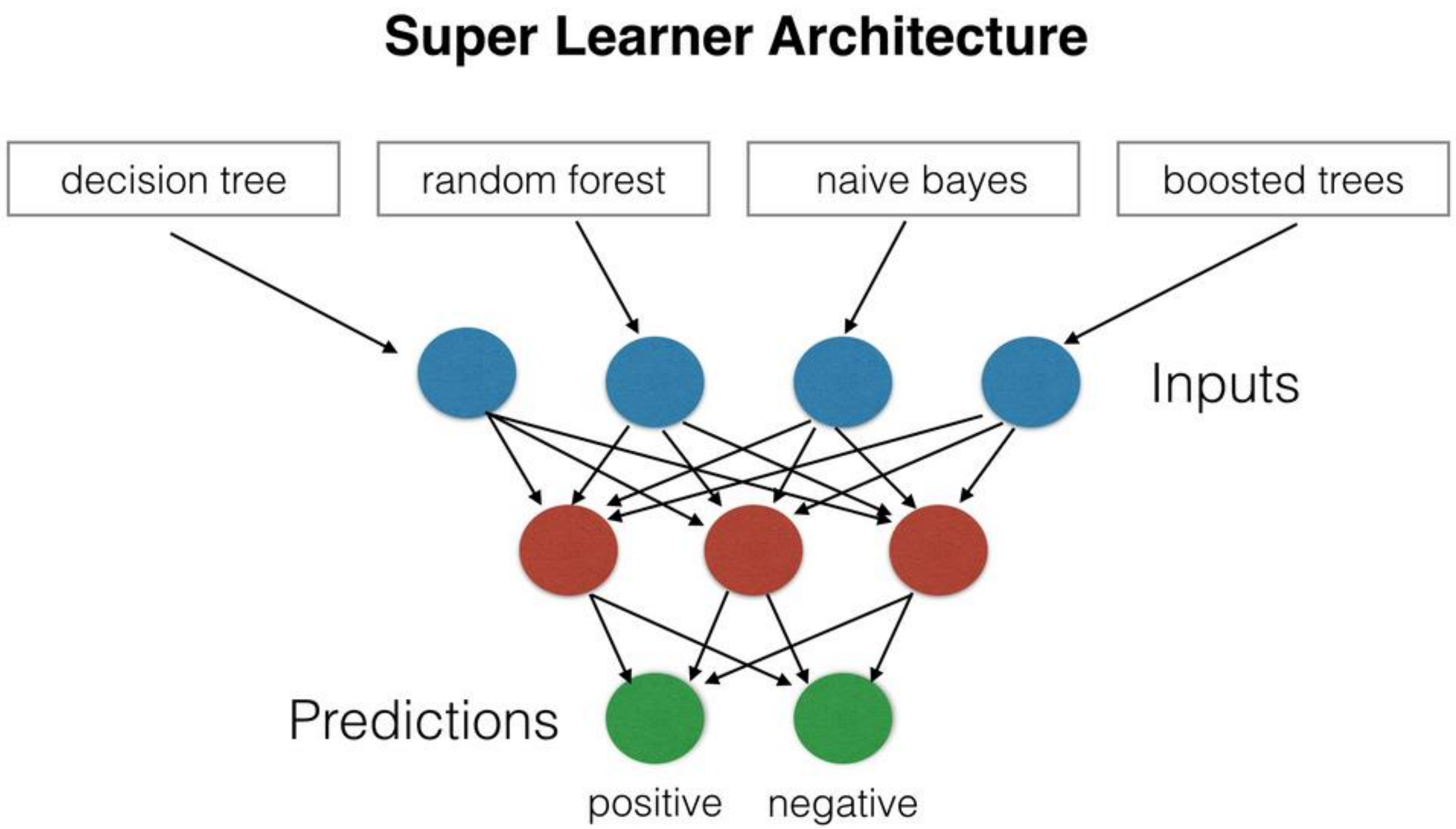
Supervised Learning Algorithms Random Forest

- Tree-based approach (like CART)
- Draw a random sample of subjects and a random sample of predictors and then create decision tree
- Average across trees
- Pros: improved prediction, more stable than CART
- Cons: interpretability – No clear measure to assess the association between predictors and outcome (e.g., OR), no final tree



https://tex.stackexchange.com/questions/503883/illustrating-the-random-forest-algorithm-in-tikz

Supervised Learning Algorithms
Super Learner



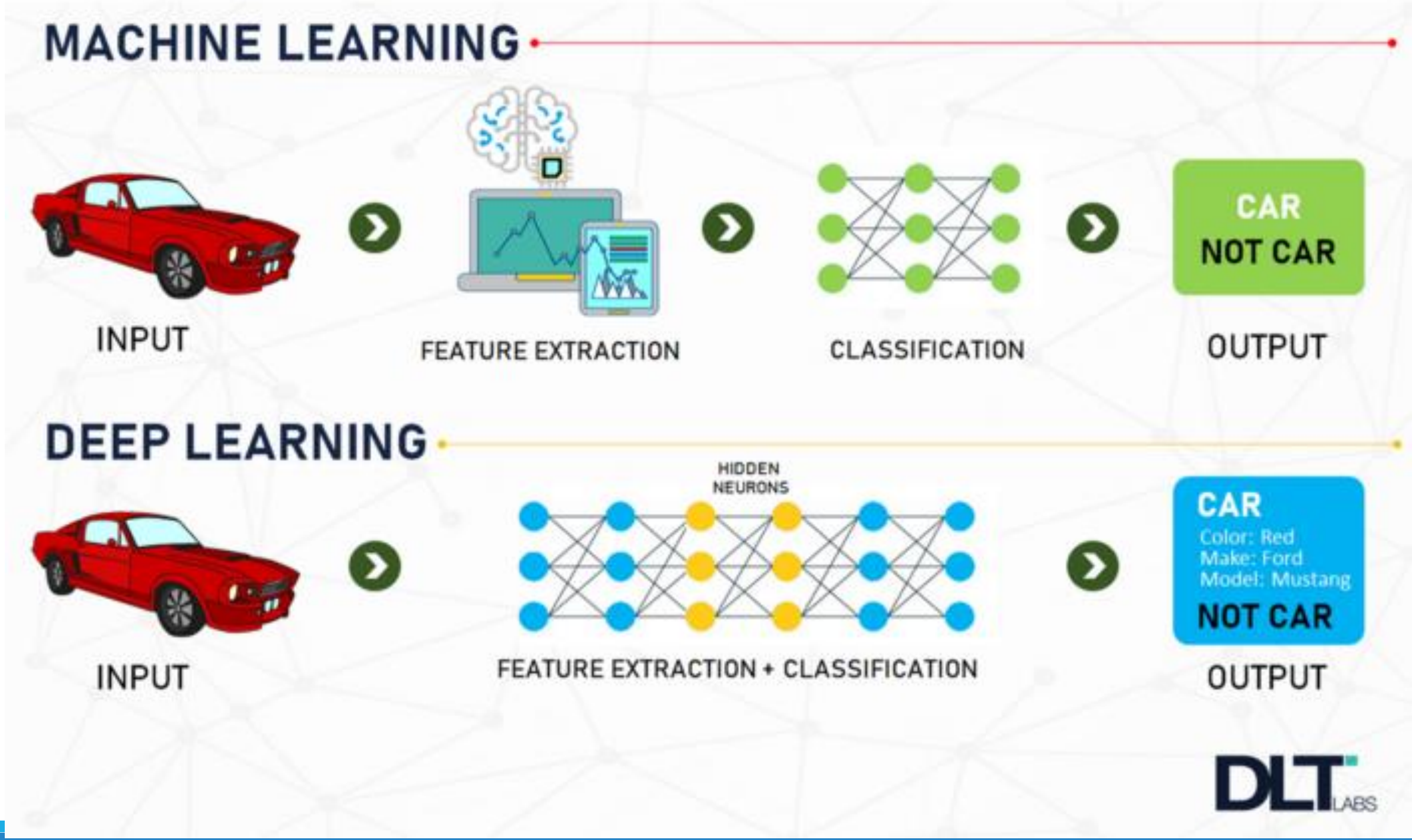
https://blog.jaysinha.me/train-your-first-super-learner-ensemble-model-for-classification/

Supervised Learning Algorithms
From ML to AI

Wheels	Doors	Motor	Steering wheel	Headlights	**Outcome: car**
4	4	Yes	Yes	Yes	**Yes**
2	0	Yes	No	Yes	**No**
4	2	No	No	No	**No**

https://dltlabs.medium.com/understanding-machine-learning-deep-learning-f5aa95264d61

Supervised Learning Algorithms From ML to AI



https://dltlabs.medium.com/understanding-machine-learning-deep-learning-f5aa95264d61

Supervised Learning Algorithms
From ML to AI

Panda

Jacques

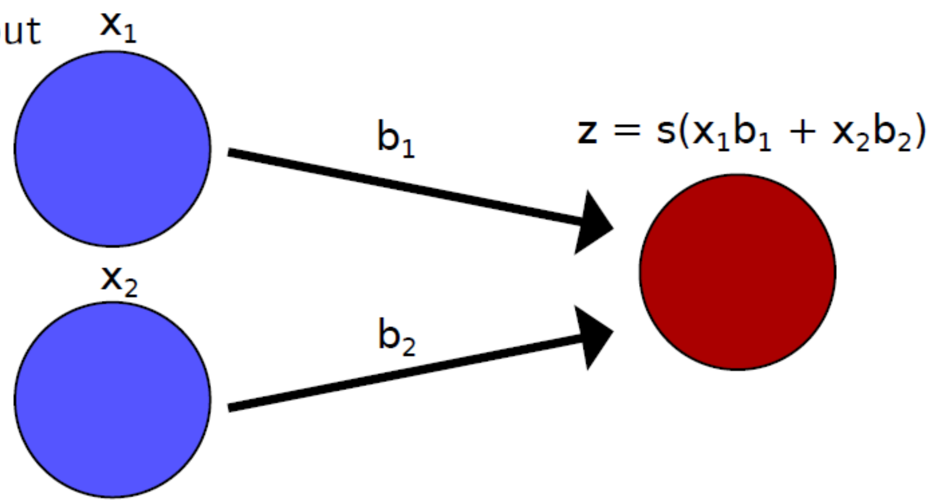
Learning Algorithms: Deep Learning
Simple Neural Net

input layer (features or predictors)

output layer (outcome)

use a simple linear map from input to output

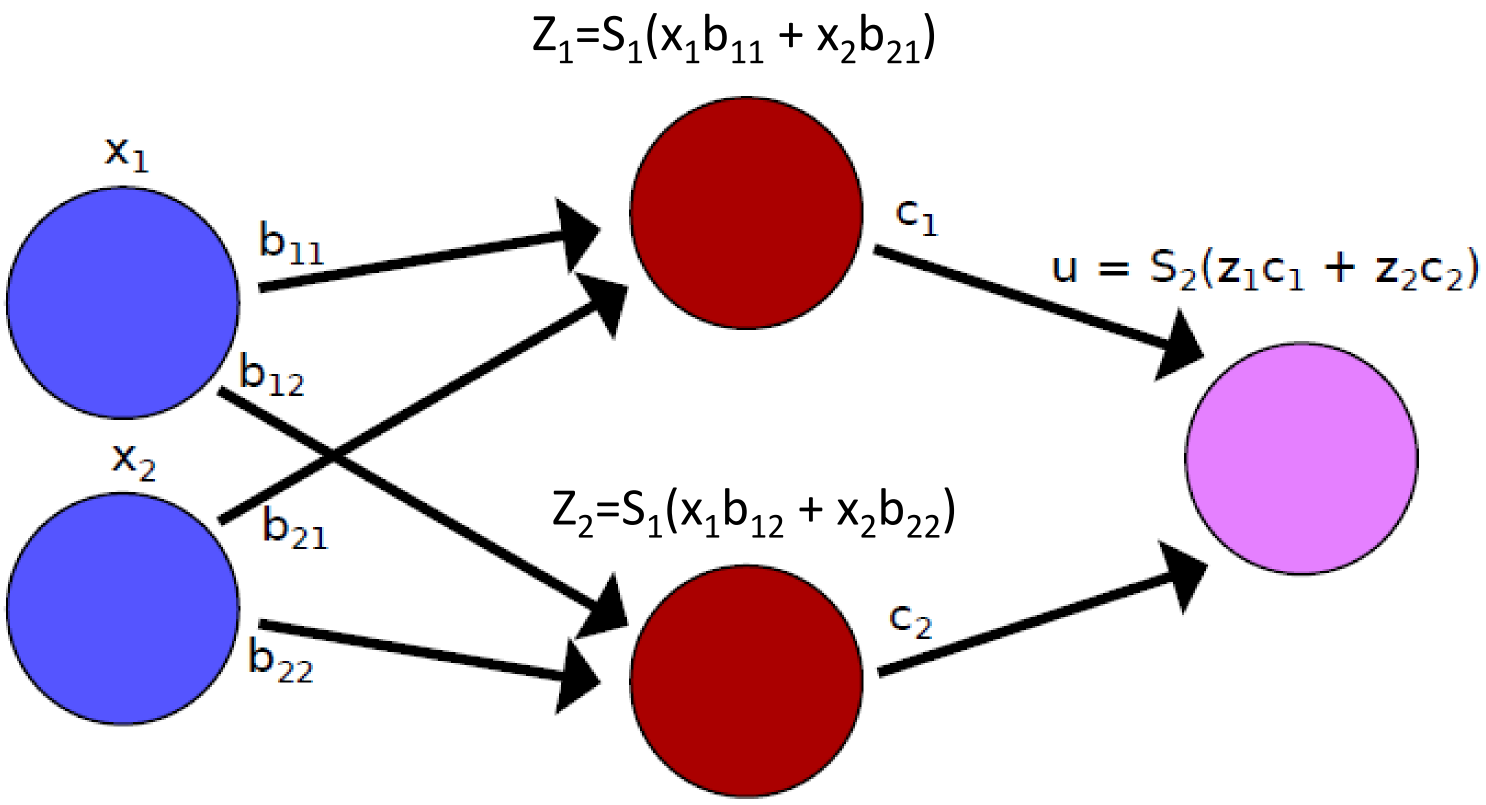
- x_1, x_2 - inputs
- z - output



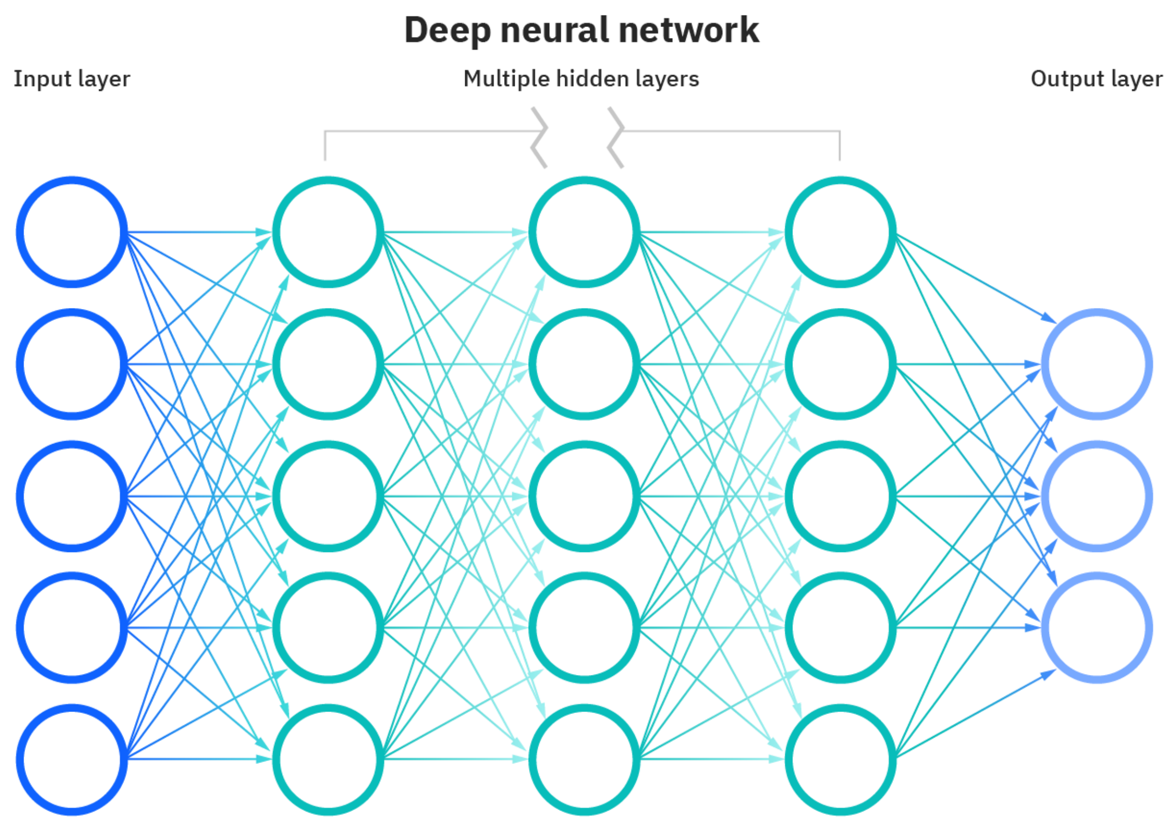
$$f_{b_1,b_2}(x_1, x_2) = S(x_1b_1 + x_2b_2)$$

Simon & Shojaie. Supervised Learning: Neural Networks and Deep Learning. Summer Institute in Statistics for Big Data. University of Washington. 2021

Learning Algorithms: Deep Learning
Neural Net with 1 Hidden Layer



Learning Algorithms: Deep Learning Neural Net with Many Hidden Layers



https://www.ibm.com/cloud/learn/neural-networks

Machine Learning and MI Reviewer Resources

Radiology

COMMUNICATIONS

Assessing Radiology Research on Artificial Intelligence: A Brief Guide for Authors, Reviewers, and Readers—From the *Radiology* Editorial Board

David A. Bluemke, MD, PhD • Linda Moy, MD • Miriam A. Bredella, MD • Birgit B. Ertl-Wagner, MD, MHBA • Kathryn J. Fowler, MD • Vicky J. Goh, MBBCh • Elkan F. Halpern, PhD • Christopher P. Hess, MD • Mark L. Schiebler, MD • Clifford R. Weiss, MD

Radiology 2020; 294:487–489 • https://doi.org/10.1148/radiol.2019192515 • ©RSNA, 2019

Key Considerations for Authors, Reviewers, and Readers of AI/ML Manuscripts in *Radiology*

Key Considerations
Are all three image sets (training, validation, and test sets) defined?
Is an *external* test set used for final statistical reporting?
Have multivendor images been used to evaluate the AI algorithm?
Are the sizes of the training, validation, and test sets justified?
Was the AI algorithm trained using a standard of reference that is widely accepted in our field?
Was preparation of images for the AI algorithm adequately described?
Were the results of the AI algorithm compared with radiology experts and/or pathology?
Was the manner in which the AI algorithm makes decisions demonstrated?
Is the AI algorithm publicly available?

Note.—AI = artificial intelligence, ML = machine learning.

Machine Learning and MI Reviewer Resources

Radiology:Artificial Intelligence

Checklist for Artificial Intelligence in Medical Imaging (CLAIM): A Guide for Authors and Reviewers

John Mongan, MD, PhD • *Linda Moy, MD* • *Charles E. Kahn, Jr, MD, MS*

Radiology: Artificial Intelligence 2020; 2(2):e200029 • https://doi.org/10.1148/ryai.2020200029 • Content codes: IN AI • ©RSNA, 2020

Ground Truth	14	Definition of ground truth reference standard, in sufficient detail to allow replication
	15	Rationale for choosing the reference standard (if alternatives exist)
	16	Source of ground truth annotations; qualifications and preparation of annotators
	17	Annotation tools
	18	Measurement of inter- and intrarater variability; methods to mitigate variability and/or resolve discrepancies
Data Partitions	19	Intended sample size and how it was determined
	20	How data were assigned to partitions; specify proportions
	21	Level at which partitions are disjoint (eg, image, study, patient, institution)
Model	22	Detailed description of model, including inputs, outputs, all intermediate layers and connections
	23	Software libraries, frameworks, and packages
	24	Initialization of model parameters (eg, randomization, transfer learning)
Training	25	Details of training approach, including data augmentation, hyperparameters, number of models trained
	26	Method of selecting the final model
	27	Ensembling techniques, if applicable

Machine Learning and MI Reviewer Resources

Open access | Protocol

BMJ Open

Protocol for development of a reporting guideline (TRIPOD-AI) and risk of bias tool (PROBAST-AI) for diagnostic and prognostic prediction model studies based on artificial intelligence

Gary S Collins,[1,2] Paula Dhiman,[1,2] Constanza L Andaur Navarro,[3] Jie Ma,[1] Lotty Hooft,[3,4] Johannes B Reitsma,[3] Patricia Logullo,[1,2] Andrew L Beam,[5,6] Lily Peng,[7] Ben Van Calster,[8,9,10] Maarten van Smeden,[3] Richard D Riley,[11] Karel GM Moons[3,4]

Key References

Collins S., et al. *Transparent Reporting of a Multivariable Prediction Model for Individual Prognosis or Diagnosis (TRIPOD) The TRIPOD Statement*. Circulation 131.2 (2015): 211-219.

James G, Witten D, Hastie T, Tibshirani R. *An introduction to statistical learning*. New York: Springer; 2013.

Harrell Jr, Frank E. *Regression modeling strategies: with applications to linear models, logistic and ordinal regression, and survival analysis*. Springer, 2015.

Steyerberg, Ewout W. *Clinical prediction models*. Springer International Publishing, 2019.

van der Laan MJ, Rose S. *Targeted Learning: Causal Inference for Observational and Experimental Data*. Springer, 2011.

Blakely T, Lynch J, Simons K, Bentley R, Rose S. *Reflection on modern methods: when worlds collide—prediction, machine learning and causal inference*. International journal of epidemiology. 2020 Dec 1;49(6):2058-64.

Guthridge JM, Wagner CA, James JA. *The promise of precision medicine in rheumatology*. Nature medicine. 2022 Jul;28(7):1363-71.

Thank You!

Jamie E. Collins, PhD

Orthopaedic and Arthritis Center for Outcomes Research, BWH

Department of Orthopaedic Surgery, HMS

oracore.bwh.harvard.edu

JCollins13@bwh.harvard.edu

Funding

NIH NIAMS K01 AR075879

NIH NIAMS P30 AR072577