Introduction to Machine Learning in R

UC3M/IC3JM Graduate Workshop, Fall 2024

Patrick Kraft

2024-12-03

Introduction

Learning outcomes

- **Strategy**: From the simple to the complex, slowly diving into the logic of machine learning using building blocks that we already know

- **Goals**: By the end of the course participants will:
  - understand key concepts underlying machine learning.
  - be able to interpret and evaluate machine learning models.
  - be able to critically assess model performance on different dimensions of quality.
  - be able to use various machine learning models for prediction and classification.
  - have learned how to use the tidymodels framework for machine learning in R.
  - have learned how to evaluate and visualize model performance using ggplot2.

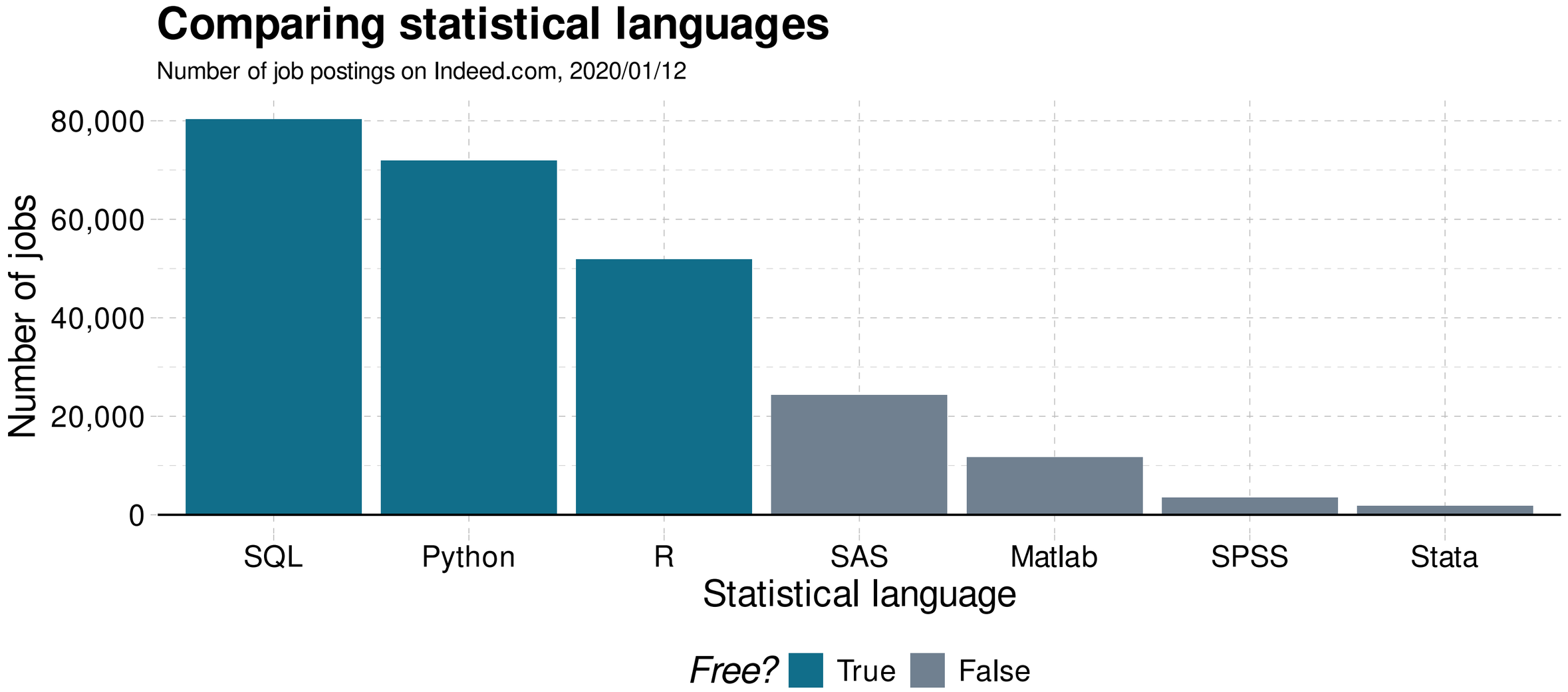
Recommended Readings

- **Important**: The workshop/seminar does not require any prior reading.

- However, we recommend the following textbooks for further reading (see additional references on syllabus):
  - Kuhn, M., & Silge, J. (2022). Tidy modeling with R. O'Reilly Media, Inc. https://www.tmwr.org/
  - James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013, corrected printing June 2023). An introduction to statistical learning (Vol. 112, p. 18). New York: springer. https://www.statlearning.com/
  - Mehrabi, N., Morstatter, F., Saxena, N., Lerman, K., & Galstyan, A. (2021). A survey on bias and fairness in machine learning. ACM computing surveys (CSUR), 54(6), 1-35. https://arxiv.org/abs/1908.09635
  - Chollet, F., & Allaire, J. J. (2018). Deep learning with R, Second Edition. https://livebook.manning.com/book/deep-learning-with-r-second-edition/

- **Acknowledgements:** This lecture draws on material developed by Paul C. Bauer, Katrien Antonio, and others.

Why R and RStudio?



Why R and RStudio? (cont.)

Data science positivism

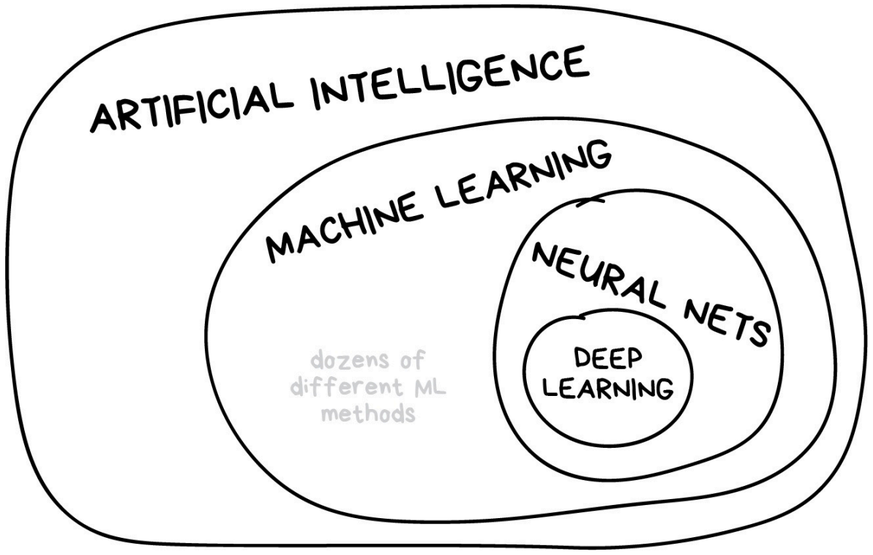
- Next to Python, R has become the *de facto* language for data science, with a cutting edge *machine learning toolbox*.
- See: The Popularity of Data Science Software
- R is open-source with a very active community of users spanning academia and industry.
- R does not try to be everything to everyone. The RStudio IDE and ecosystem allow for further, seamless integration (with e.g. python, keras, tensorflow or C).

Disclaimer + Read more

- It's also the language that we know best.
- If you want to read more:
  R-vs-Python, when to use Python or R or Hadley Wickham on the future of R

Artificial Intelligence and Machine Learning

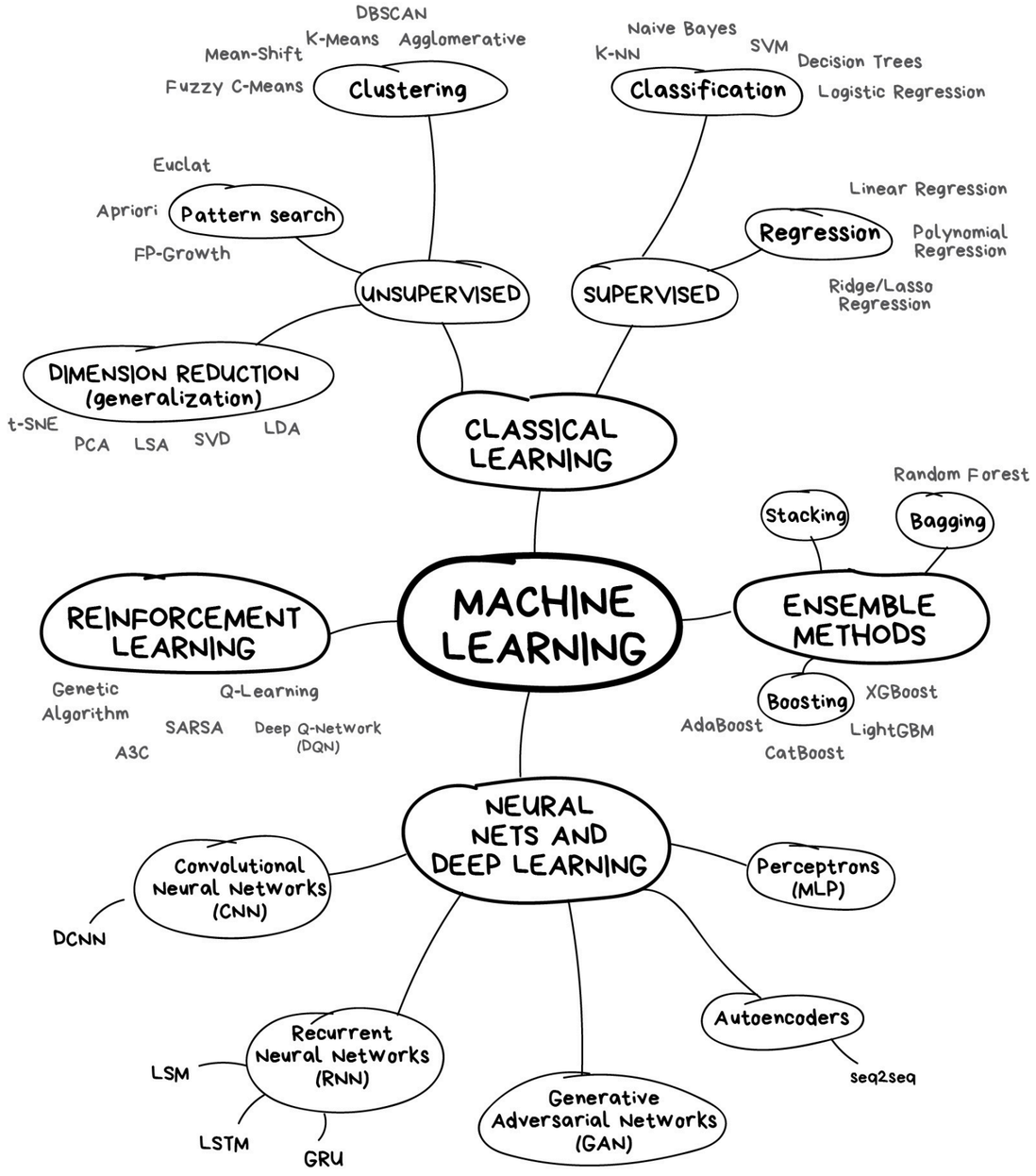
Definitions: What is AI?



Source: Machine Learning for Everyone. In simple words. With real-world examples. Yes, again.

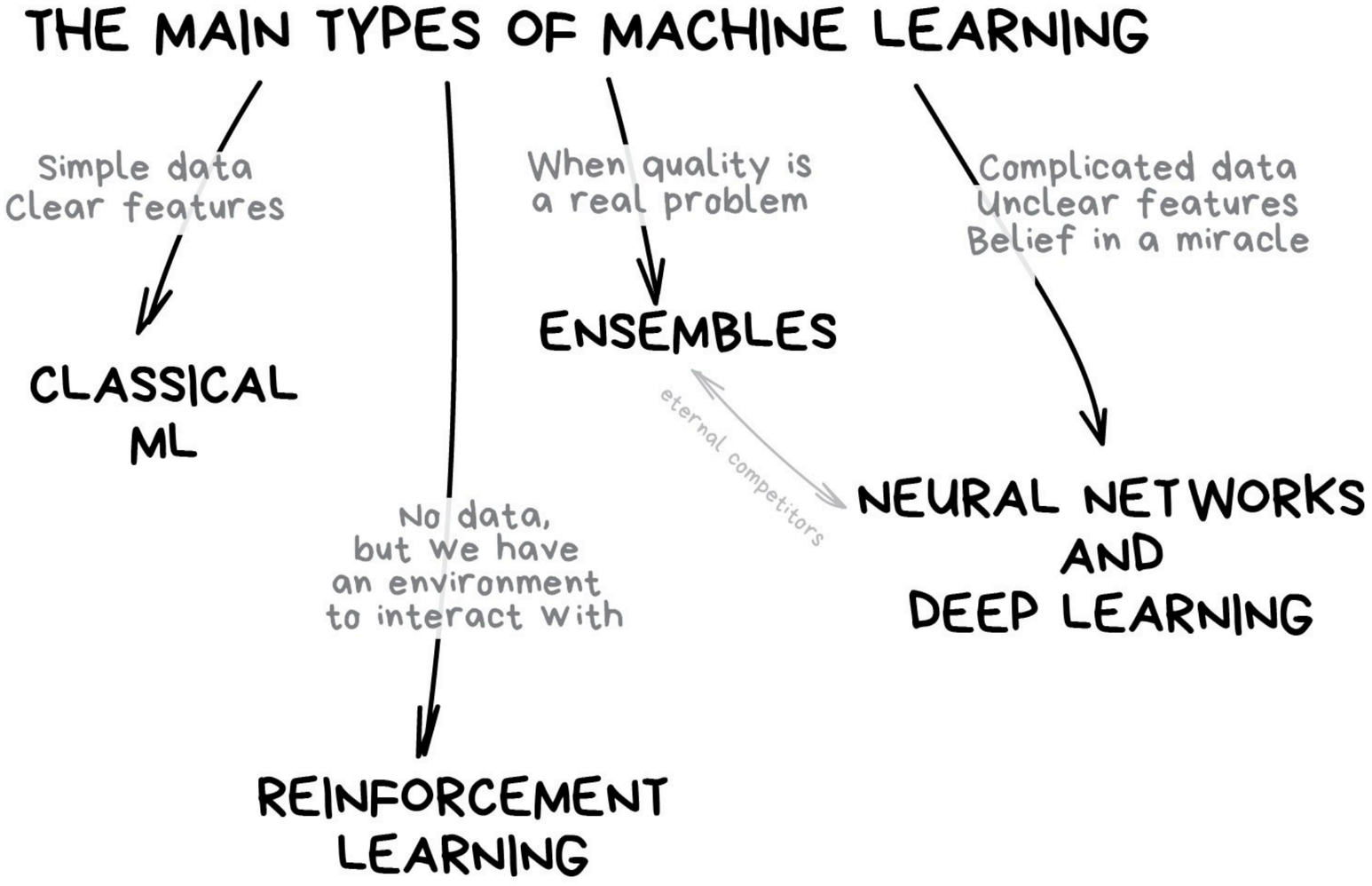
- "Artificial intelligence (AI) is **intelligence - perceiving, synthesizing, and inferring information - demonstrated by machines**, as opposed to intelligence displayed by non-human animals and humans. **Example tasks** in which this is done include **speech recognition, computer vision, translation between (natural) languages**, as well as other mappings of inputs." (Wikipedia)
- "the effort to automate intellectual tasks normally performed humans" (Chollet and Allaire 2018:2) (includes chess computers!)

Definitions: What is Machine Learning?



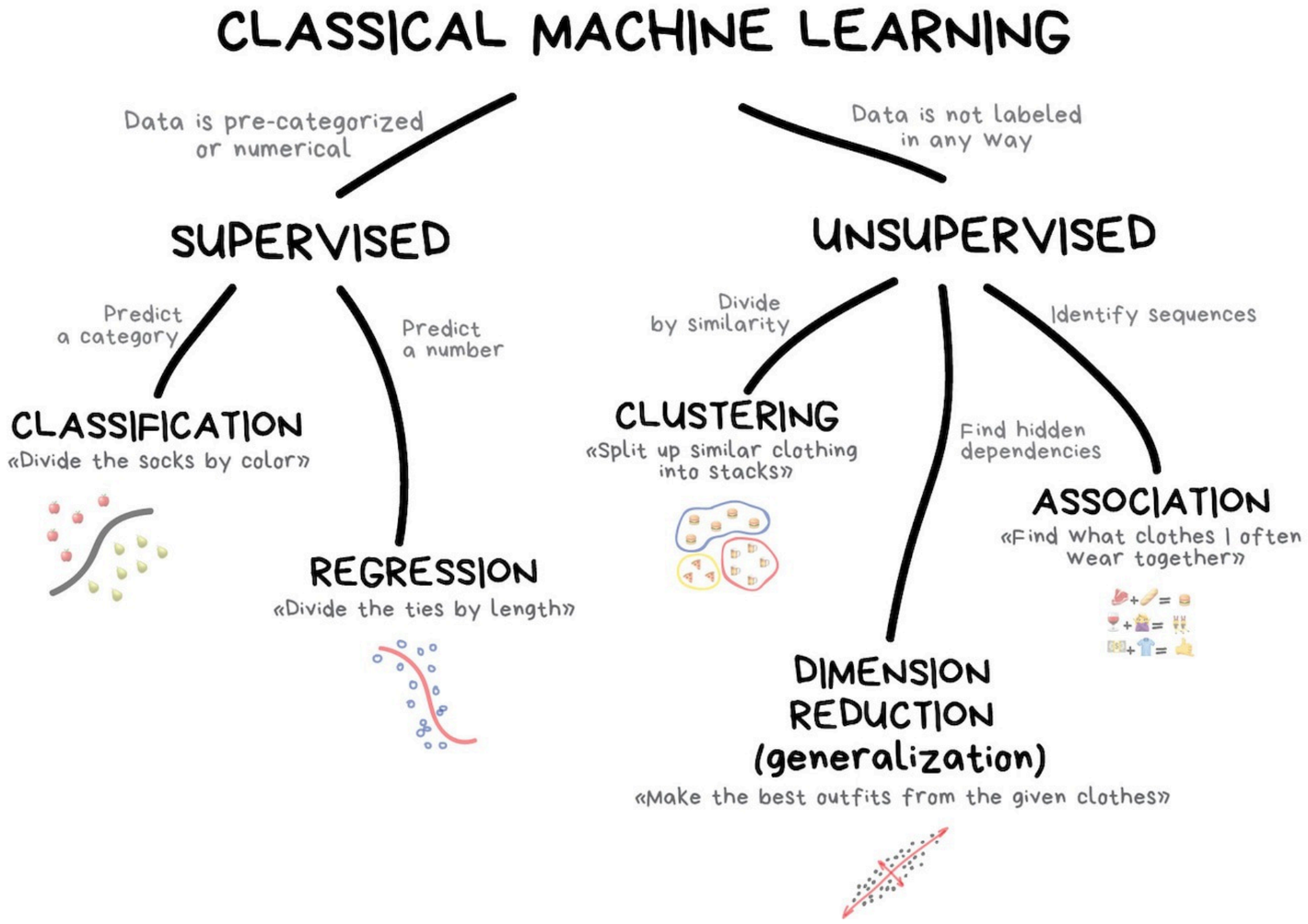
- "Machine learning (ML) is a **field of inquiry devoted to understanding and building methods that "learn"** [...] It is seen as a **part of artificial intelligence**." (Wikipedia)
  - "Machine learning is **programming computers** to **optimize a performance criterion** using example data or past experience." (Alpaydin 2014:3)
  - "Machine learning is a specific **subfield of AI** that aims at automatically developing **programs** (called **models**) purely from exposure to **training data**. This process of turning models data into a program is called **learning**." (Chollet and Allaire 2018:307)

Definitions: Types of Machine Learning



Source: Machine Learning for Everyone

Definitions: Classic Machine Learning



Source: Machine Learning for Everyone

Stats on AI: Capabilities

Stats on AI: Academia vs. Industry

Machine Learning in the Social Sciences

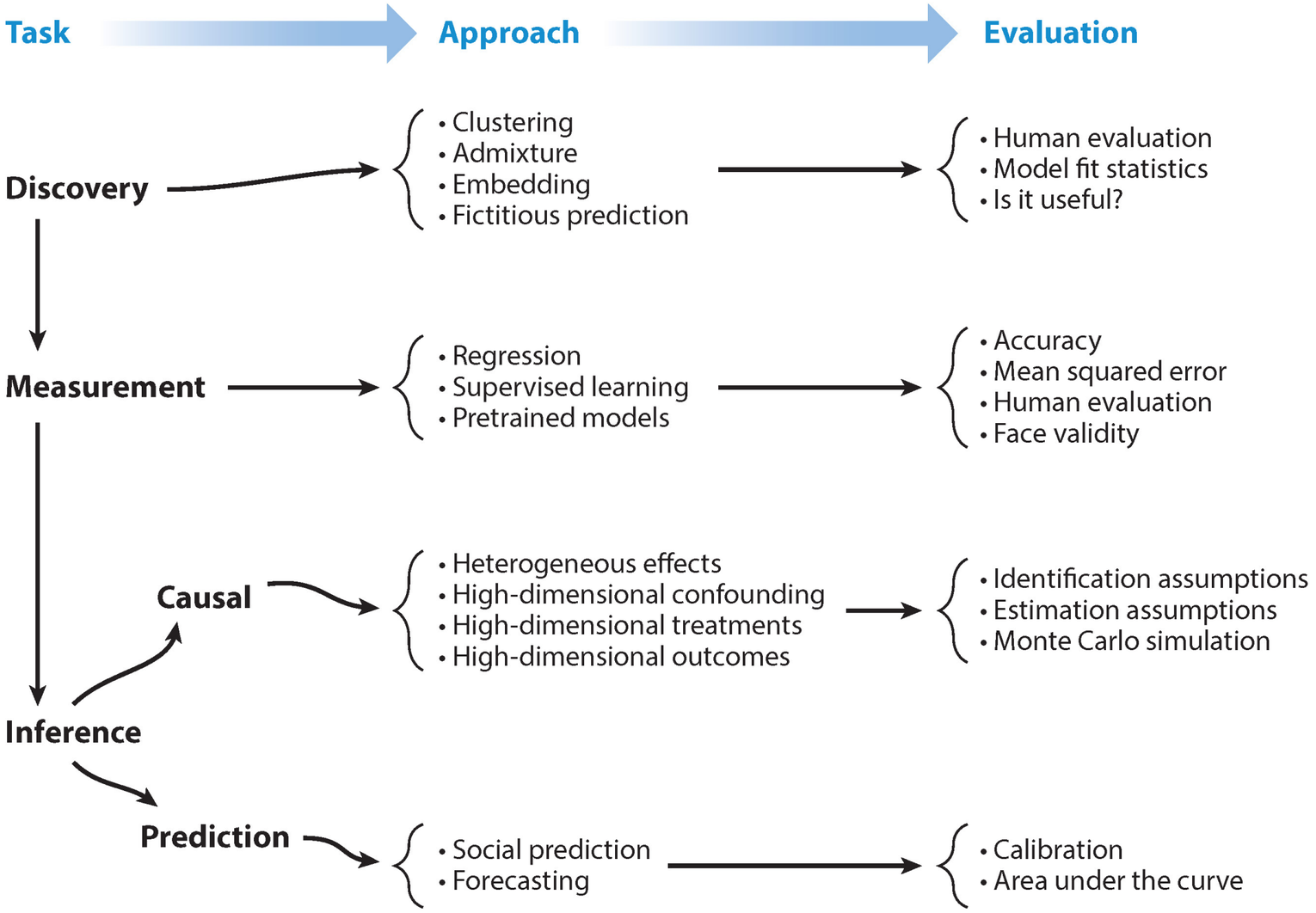


Figure 1

Our approach to machine learning in the social sciences. We reframe the tasks to ones relevant to social science: discovery, measurement, causal inference, and prediction.

Statistics vs. Machine Learning: Cultures and Goals

Two cultures of statistical analysis (Breiman 2001; Molina and Garip 2019:29)

Generative modeling
(Objective: Inference / Explanation)

- Goal: understand how an outcome is related to inputs
- Analyst proposes a stochastic model that could have generated the data, and estimates the parameters of the model from the data
- Leads to simple and interpretable models BUT often ignores model uncertainty and out-of-sample performance

Predictive modeling
(Objective: Prediction)

- Goal: prediction, i.e., forecast the outcome for unseen or future observations
- Analyst treats underlying generative model for data as unknown and primarily considers the predictive accuracy of alternative models on new data
- Leads to complex models that perform well out of sample BUT can produce black-box results that offer little insight on the mechanism linking the inputs to the output (but Interpretable ML)

Statistics vs. Machine Learning: Differences in Terminology

Well-established older labels vs. "new" terminology

- Sample used to estimate the parameters vs. training sample
- Model is estimated vs. Model is trained/fitted
- Regressors, covariates, predictors vs. features (or inputs)
- Dependent variable/outcome vs. output
- Regression parameters (coefficients) vs. weights

Machine learning as programming paradigm

- Machine learning arises from this question: *could a computer [...] learn on its own how to perform a specified task? [...] Rather than programmers crafting data-processing rules by hand, could a computer automatically learn these rules by looking at data?* (Chollet and Allaire 2018:1.1.2)

Debates around AI/ML

- (How well) Does AI really work?
  - "Rebooting AI" (Marcus and Davis 2019), etc.
  - How should AI be tested? (e.g., competitions etc.)
- Interpretable machine learning (e.g. Molnar 2022; Rudin 2019)
- Bias of ML algorithms (training data, personnel, minorities) (e.g., Metz 2022, Ch. 15)
  - "I call it a sea of dudes" (Margaret Mitchell, member of Microsoft's "cognition" group)
  - Garbage in, garbage out; reproduction of inequalities (rental/loan markets)
- Weaponization (e.g., Metz 2022, Ch. 16)
  - General adversarial networks (GANS) + adversarial attacks (e.g., Metz 2022, Ch. 13)
  - Book: "Weapons of Math Destruction" by Cathy O'Neill

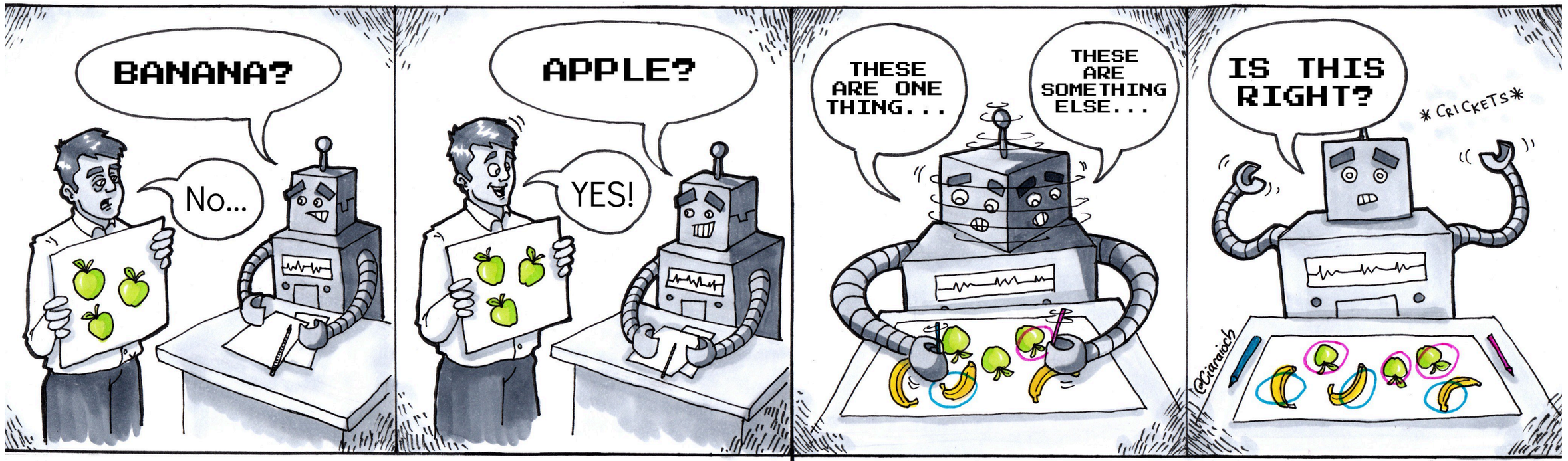
Exercise: Examples of Machine Learning Applications

Q: What do the following techniques predict? What is the input/what is the output? How could we use those ML models for research in our disciplines? (Discuss 2!)

1. Image recognition (Clarify) Image recognition (Google)
   - ▶ Answer
2. Speech recognition
   - ▶ Answer
3. Translation
   - ▶ Answer
4. Text analysis/Natural language processing (NLP)
   - ▶ Answer
5. Pose estimation (2018!)
   - ▶ Answer

Fundamental Concepts in Machine Learning

Supervised vs. Unsupervised Learning



Supervised Learning

Unsupervised Learning

Supervised Learning: Constructing predictive models

- **Goal:** Build ("learn") a model f (*the Signal*) such that the outcome or target Y can be written as

$$Y = f(x_1, \ldots, x_p) + \epsilon$$

  with features $x_1, \ldots, x_p$ and error term ϵ (*the Noise*).
- **Methods & models**: Linear/logistic regression, Penalized regression, classification and regression trees, nearest neighbor, neural networks/deep learning

Regression Problems

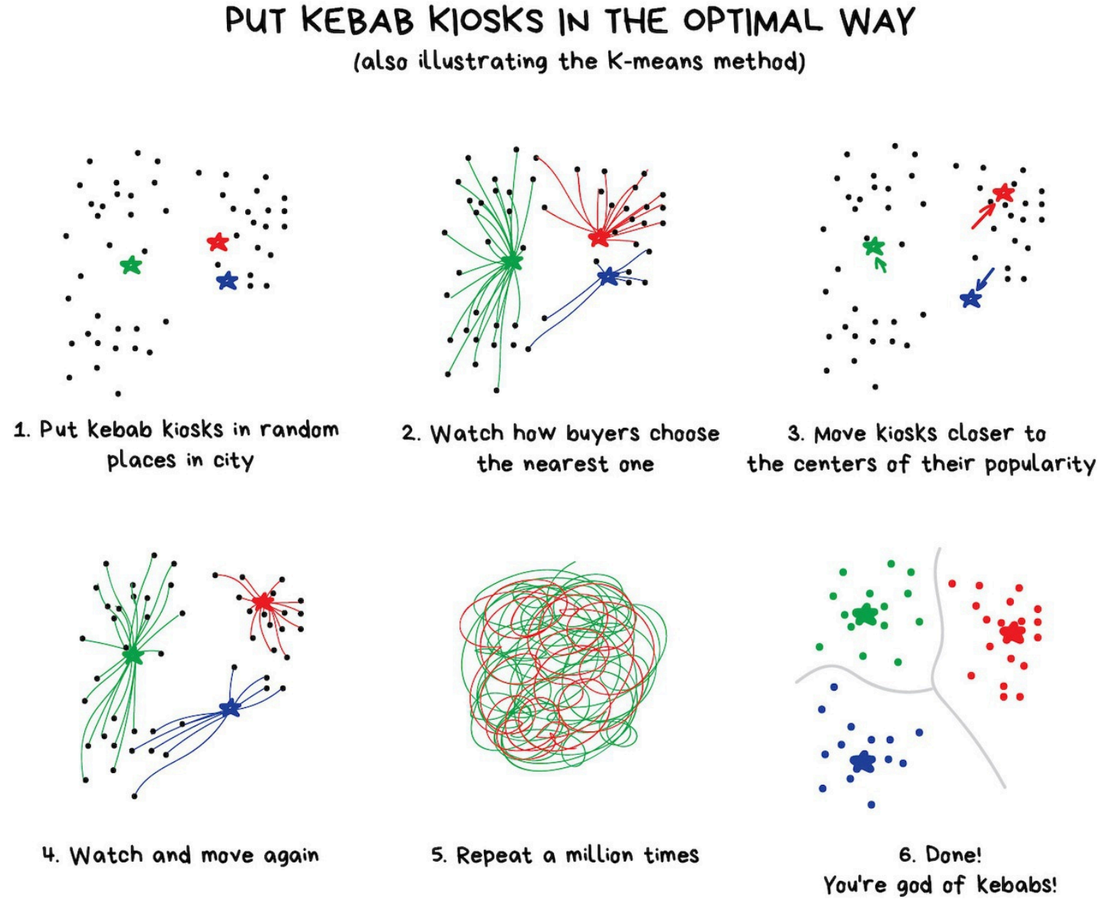
- Predicting **quantitative** responses
- Numerical values, e.g., person's age, height, or income, …

Classification Problems

- Predicting **qualitative** responses
- Values in one of K different classes or categories, e.g., a person's gender, party identification, …

Unsupervised Learning: Finding patterns in the data

- **Goal:** Construct descriptive models, without any *supervising* output, letting the data "speak for itself".
- With unsupervised learning there is **NO** outcome or target Y, only the feature vector $x = (x_1, \ldots, x_p)$. Our goal is to cluster observations that are similar in terms of their underlying features.
- x are often based on texts, images, audio snippets, videos
- **Methods & models**: Principal component, factor- , cluster-, latent class and sequence analysis; topic modelling; community detection



Source: Machine Learning for Everyone

Predictive modeling

How to use the observed data to learn or to estimate the unknown $f(.)$?

$$y = f(x_1, x_2, \ldots, x_p) + \epsilon.$$

How do I **estimate** $f(.)$?

Traditional Statistics **Machine Learning**

White-box modelling
simpler computation, emphasis on introspection, form, causal effects and processes, finding a 'correct' model

Black-box modelling
high computational complexity, emphasis on speed and quality of prediction, finding a 'performant' model

Model accuracy

We assess **model** or **predictive accuracy** by evaluating how well predictions actually match observed data.

Use **loss functions**, i.e. metrics that compare predicted values to actual values.

Regression

Use e.g. the **Mean Squared Error (MSE)**

$$\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{f}(x_i))^2,$$

Objective: **minimize**!

Classification

Use e.g. the **cross-entropy** or **log loss**

$$-\frac{1}{n}\sum_{i=1}^{n}\left(y_i \cdot \log(p_i) + (1-y_i)\cdot \log(1-p_i)\right)$$

Objective: **minimize**!

Trade-Off(s): Prediction Accuracy vs. Model Interpretability

- Some ML methods are more some are less flexible (shape of $\hat{f}$), e.g., linear model
  - James et al. (2013:25), Fig. 2.7. provides an overview
- Q: Why would we ever choose to use a more restrictive method (less flexible) model instead of a very flexible approach?

▶ Answer

Data splitting

We fit our model on past data $\{(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)\}$ and get $\hat{f}$.

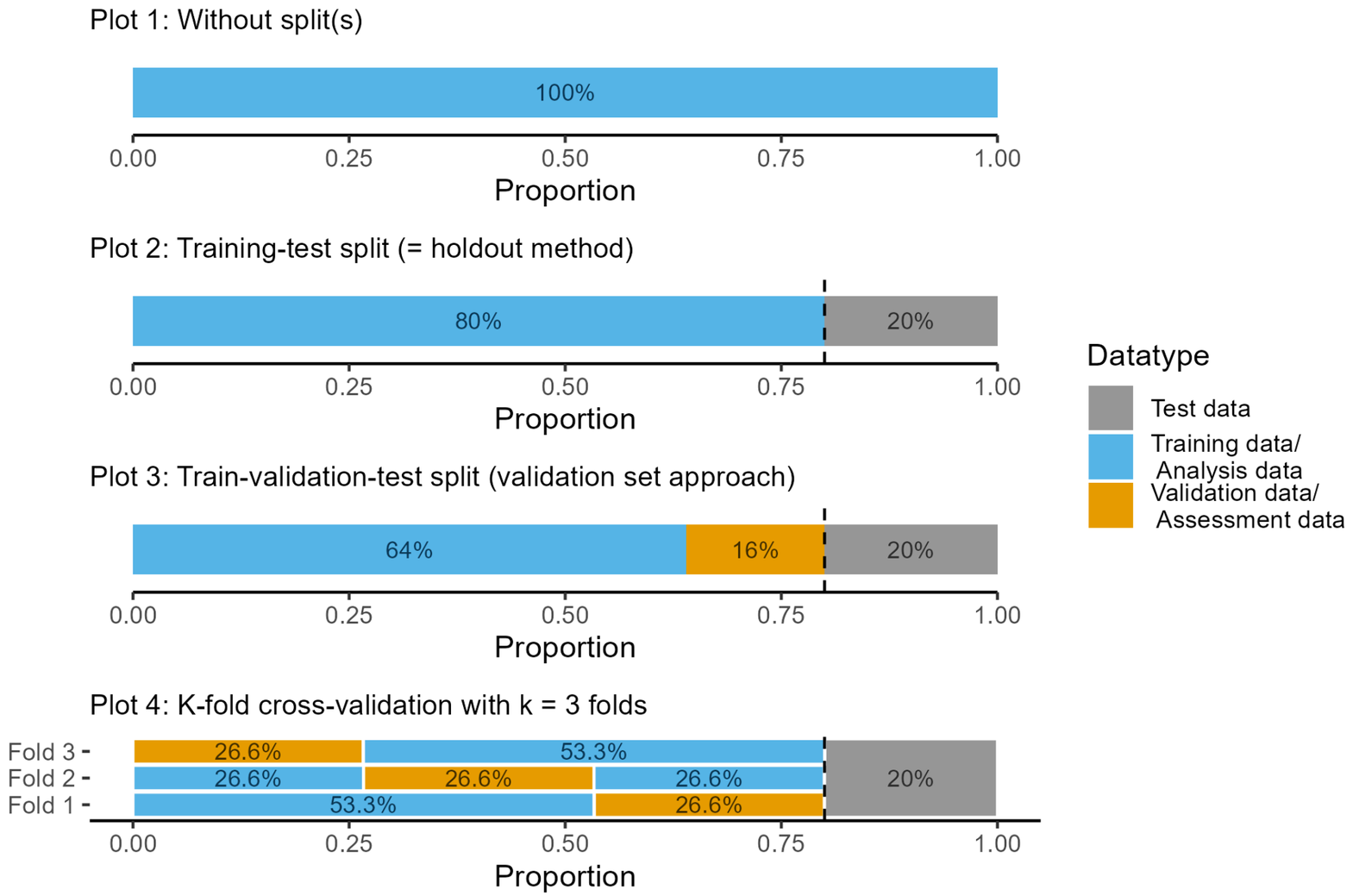
How does our model **generalize** to new, unseen data (x_0, y_0), or: is $\hat{f}(x_0)$ close to y_0?

Training set

- to develop, to train, to tune, to compare different settings, …

Test set

- to obtain unbiased estimate of final model's performance.



Note: Graph visualizes different evaluation protocols and what percentage of the data is used for training, validation and testing; Plot 3: Assigns 20% to the test data, whereby the remaining 80% are split 20/80 into validation data (16% of all data) and training data (64% of all data); Plot 4: 3-fold cross-validation randomly splits the full training data (here 80% of all data) 3 times, always keeping one third of the data for validation; © Paul C. Bauer

Figure 1

Size of training, validation and test sets

- Size of datasets: Usually 80/20 splits but depends..
  - Q: What could be a problem if training and/or test dataset is too small? (uncertainty, representativeness)

▶ Answer

Trade-Off(s): Bias vs. Variance

- **Variance:** error from sensitivity to small fluctuations in the training set
  - High variance may result from an algorithm modeling the random noise in the training data (**overfitting**)

- **Bias** error from approximating a (potentially complicated) real-life problem through a much simpler model (=$\hat{f}$)
  - High bias can cause an algorithm to miss relevant relations between *features* and *target outputs* (**underfitting**)

- **Bias-variance trade-off**: Reducing bias often implies increasing variance of an estimator (and vice verse). In general, variance will increase and bias will decrease with more flexible methods/models.

Exercise 1

Adapted from James et al. (2013, Exercise 2.4.1): Thinking of classification problems in the social sciences, indicate whether we would generally expect the performance of a flexible statistical learning method to be better or worse than an inflexible method. Justify your answer.

a. The sample size n is extremely large, and the number of predictors p is small.
   - ▶ Answer

b. The number of predictors p is extremely large, and the number of observations n is small.
   - ▶ Answer

c. The relationship between the predictors and response is highly non-linear.
   - ▶ Answer

d. The variance of the error terms, i.e. $\sigma^2 = Var(\epsilon)$, is extremely high.
   - ▶ Answer

Exercise 2

James et al. (2013, Exercise 2.4.2): Explain whether each scenario is a classification or regression problem, and indicate whether we are most interested in inference or prediction. Finally, provide n and p.

a. We collect a set of data on the top 500 firms in the US. For each firm we record profit, number of employees, industry and the CEO salary. We are interested in understanding which factors affect CEO salary.
   - ▶ Answer

b. We are considering launching a new product and wish to know whether it will be a success or a failure. We collect data on 20 similar products that were previously launched. For each product we have recorded whether it was a success or failure, price charged for the product, marketing budget, competition price, and ten other variables.
   - ▶ Answer

c. We are interesting in predicting the % change in the US dollar in relation to the weekly changes in the world stock markets. Hence we collect weekly data for all of 2012. For each week we record the % change in the dollar, the % change in the US market, the % change in the British market, and the % change in the German market.
   - ▶ Answer

Machine Learning in R using tidymodels

TIDYMODELS

The tidymodels framework is a collection of packages for modeling and machine learning using tidyverse principles.

Install tidymodels with:

```
install.packages("tidymodels")
```

- See github website.

A tidy machine learning workflow

Data resampling & feature engineering

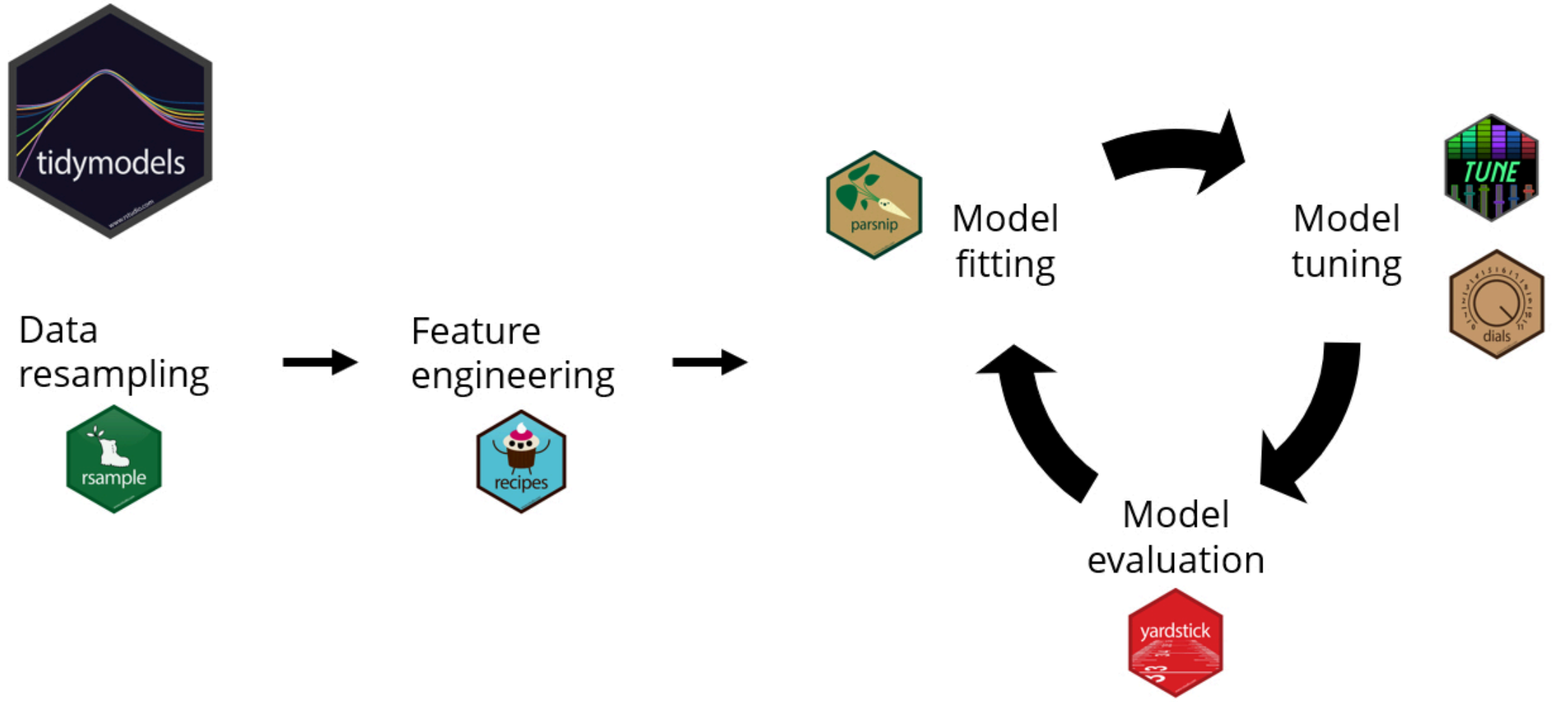
- rsample
- recipes

Model fitting & tuning

- tune
- parsnip
- dials

Model evaluation

- yardstick



Overview of `tidymodels` packages

`rsample`: for sample splitting (e.g. train/test or cross-validation)

Overview of `tidymodels` packages (cont.)

`recipes`: for pre-processing

- Use dplyr-like pipeable sequences of feature engineering steps to get your data ready for modeling.

`parsnip`: specifying the model namely model type, engine and mode

- **Goal**: provide a tidy, unified interface to access models from different packages
- `model type`-argument: e.g, linear or logistic regression
- `engine`-argument: R packages that contain these models
- `mode`-argument: either regression or classification

Overview of `tidymodels` packages (cont.)

`tune`: for model tuning

- Goal: facilitate hyperparameter tuning. It relies heavily on `recipes`, `parsnip`, and `dials`
- `dials`: contains infrastructure to create and manage values of tuning parameters

`yardstick`: evaluate model accuracy

- **Goal**: estimate how well models are working using tidy data principles
- `conf_mat()`: calculates cross-tabulation of observed and predicted classes
- `metrics()`: estimates 1+ performance metrics

Overview of `tidymodels` packages (cont.)

`workflowsets`:

- **Goal**: allow users to create and easily fit a large number of different models.
- Use `workflowsets` to create a `workflow set` that holds multiple `workflow objects`
  - These objects can be created by crossing all combinations of preprocessors (e.g., formula, recipe, etc) and model specifications. This set can be tuned or resampled using a set of specific functions.

Question: How to choose out of many methods & models?

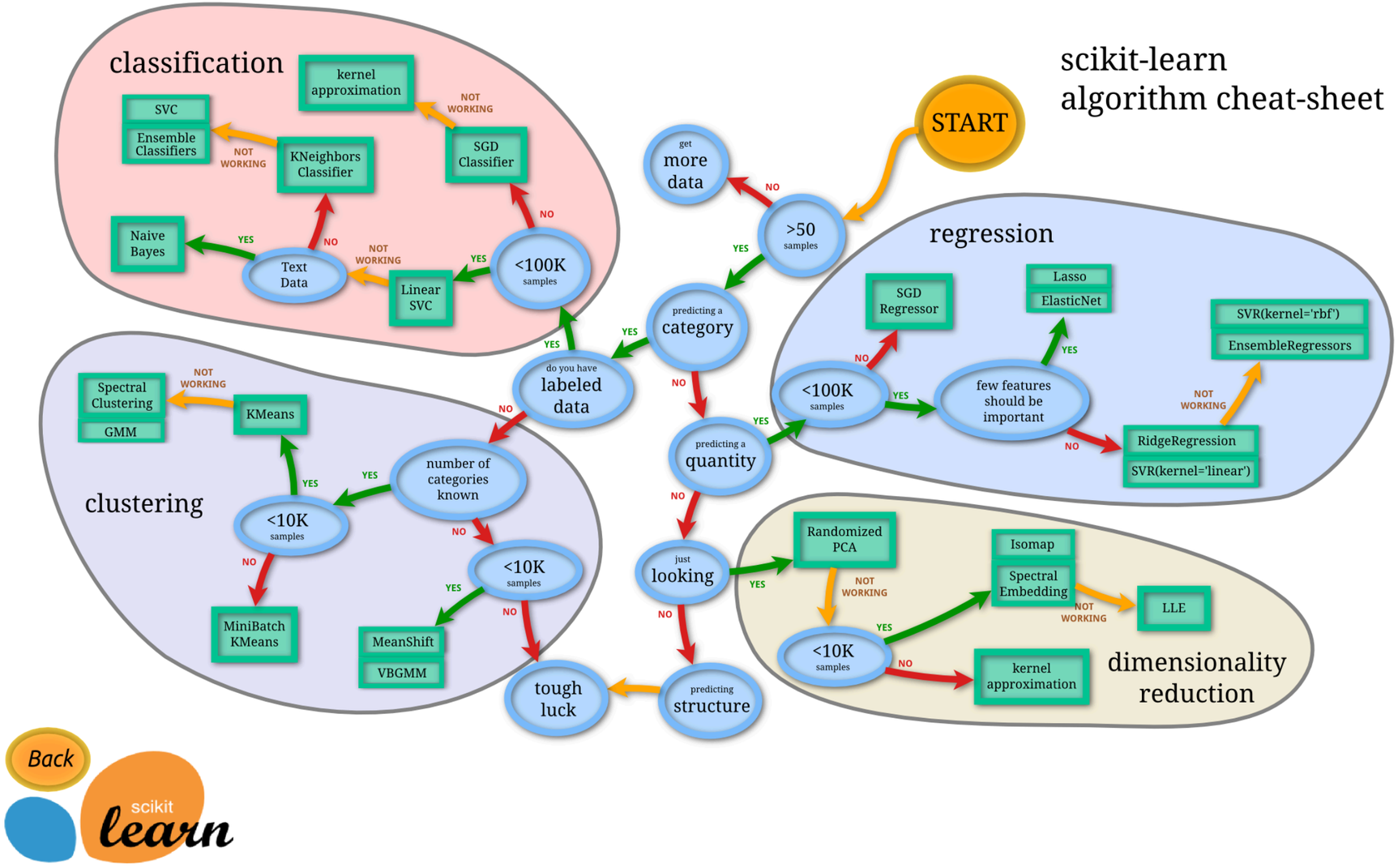


Figure 2: Source: https://scikit-learn.org/stable/tutorial/machine_learning_map/index.html

R Session

Regression example

European Social Survey (ESS)

- Round 10 - 2020. Democracy, Digital social contacts
- Outcome: Life satisfaction (0-10)
- The ESS contains different outcomes amenable to both classification and regression as well as a lot of variables that could be used as features (~580 variables).
- Research: Shen, Yin, and Jiao (2023); Collins et al. (2015); Kaiser, Otterbach, and Sousa-Poza (2022); Pan and Cutumisu (2023); Prati (2022)

▶ Overview of ESS variables

Classification example

COMPAS

- Outcome: Recidvism (0,1)
- We will be using the dataset published by propublic (Github) that is described by Angwin et al. (2016; James et al. 2013; and Lee, Du, and Guerzhoy 2020)
- The data is based on the COMPAS risk assessment tools (RAT). RATs are increasingly being used to assess a criminal defendant's probability of re-offending.

▶ Overview of COMPAS variables

References

Alpaydin, Ethem. 2014. *Introduction to Machine Learning*. MIT Press.

Angwin, Julia, Jeff Larson, Lauren Kirchner, and Surya Mattu. 2016. "Machine Bias."

Breiman, Leo. 2001. "Statistical Modeling: The Two Cultures (with Comments and a Rejoinder by the Author)." *SSO Schweiz. Monatsschr. Zahnheilkd.* 16(3):199–231.

Chollet, Francois, and J. J. Allaire. 2018. *Deep Learning with R*. 1st ed. Manning Publications.

Collins, Susan, Yizhou Sun, Michal Kosinski, David Stillwell, and Natasha Markuzon. 2015. "Are You Satisfied with Life?: Predicting Satisfaction with Life from Facebook." Pp. 24–33 in *Social computing, Behavioral-Cultural modeling, and prediction*. Springer International Publishing.

Grimmer, Justin, Margaret E. Roberts, and Brandon M. Stewart. 2021. "Machine Learning for Social Science: An Agnostic Approach." *Annu. Rev. Polit. Sci.* 24(1):395–419.

James, Gareth, Daniela Witten, Trevor Hastie, and Robert Tibshirani. 2013. *An Introduction to Statistical Learning: With Applications in R*. Springer.

Kaiser, Micha, Steffen Otterbach, and Alfonso Sousa-Poza. 2022. "Using Machine Learning to Uncover the Relation Between Age and Life Satisfaction." *Sci. Rep.* 12(1):5263.

Kuhn, Max, and Kjell Johnson. 2019. *Feature Engineering and Selection: A Practical Approach for Predictive Models*. CRC press (Taylor & Francis).

Lee, Claire S., Jeremy Du, and Michael Guerzhoy. 2020. "Auditing the COMPAS Recidivism Risk Assessment Tool: Predictive Modelling and Algorithmic Fairness in CS1." Pp. 535–36 in *Proceedings of the 2020 ACM conference on innovation and technology in computer science education, ITiCSE '20*. New York, NY, USA: Association for Computing Machinery.

Marcus, Gary, and Ernest Davis. 2019. *Rebooting AI: Building Artificial Intelligence We Can Trust*. Knopf Doubleday Publishing Group.

Metz, Cade. 2022. *Genius Makers: The Mavericks Who Brought AI to Google, Facebook, and the World*. United Kingdom: Penguine Random House.

Molina, Mario, and Filiz Garip. 2019. “Machine Learning for Sociology.” *Annu. Rev. Sociol.*
Molnar, Christoph. 2022. “Interpretable Machine Learning.”
Pan, Zexuan, and Maria Cutumisu. 2023. “Using Machine Learning to Predict UK and Japanese Secondary Students’ Life Satisfaction in PISA 2018.” *Br. J. Educ. Psychol.*
Prati, Gabriele. 2022. “Correlates of Quality of Life, Happiness and Life Satisfaction Among European Adults Older Than 50 Years: A Machine-learning Approach.” *Arch. Gerontol. Geriatr.* 103:104791.
Rudin, Cynthia. 2019. “Stop Explaining Black Box Machine Learning Models for High Stakes Decisions and Use Interpretable Models Instead.” *Nat Mach Intell* 1(5):206–15.
Shen, Xiaofang, Fei Yin, and Can Jiao. 2023. “Predictive Models of Life Satisfaction in Older People: A Machine Learning Approach.” *Int. J. Environ. Res. Public Health* 20(3).