ISA 419: Data-Driven Security

16: Machine Learning Regression Models

Fadel M. Megahed, PhD

Endres Associate Professor
Farmer School of Business
Miami University

@FadelMegahed
fmegahed
fmegahed@miamioh.edu
Automated Scheduler for Office Hours

Spring 2024

Quick Refresher of Last Class

✅ Describe the basic concepts of regression analysis, including the roles of independent and dependent variables.

✅ Assess regression models using metrics like R-squared and Mean Squared Error and interpret the results.

✅ Describe the two modeling mindsets: explanatory and predictive.

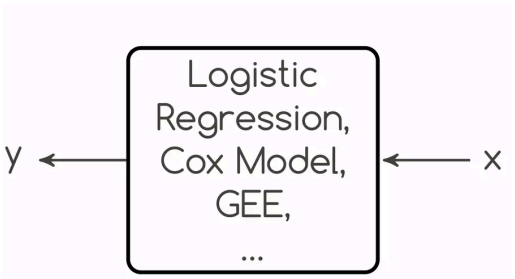
Learning Objectives for Today's Class

- Recognize the differences and similarities between traditional regression and machine learning regression models
- Describe advanced techniques like Ridge and Lasso
- Explore non-linear models like Decision Trees and Random Forests for regression tasks
- Apply machine learning regression models to cybersecurity datasets

Differences Between Traditional (Statistical) and Algorithmic (Machine Learning) Modeling Cultures

Statistical Modeling Culture

Find a stochastic model of the data-generating process in the form of: $Y = f(X, parameters, \epsilon)$



Assumptions

- Specific stochastic model of the data-generating process
- Distributional assumptions about the errors
- Linearities
- Manual specification of interactions

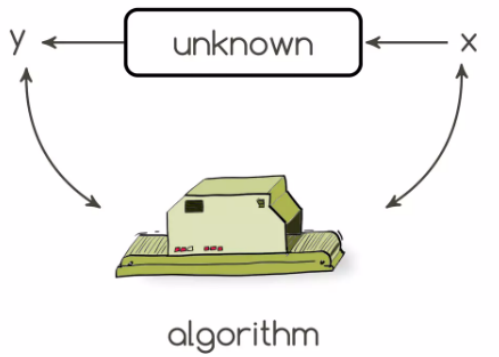
Problems

- Conclusions about model, and not about "nature"
- Assumptions often violated
- Often no model evaluation (no focus on prediction)

Image Source: Chris Molnar. (2014). Statistical Modeling: The Two Cultures - Based on Leo Breiman's paper

Algorithmic (Machine Learning) Modeling Culture

Find a function $f(x)$ that minimizes the loss $L(Y, f(X))$



Assumptions

- No assumptions about the data-generating process (true mechanism of generating the data is **unknown** and not of direct interest).
- No assumptions about the distribution of the errors.

Problems

- Some models are "black boxes".
- Stability of the model is not guaranteed (and honestly this is also a problem in statistical modeling).

Image Source: Chris Molnar. (2014). Statistical Modeling: The Two Cultures - Based on Leo Breiman's paper

Statistical Modeling: The Two Cultures

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

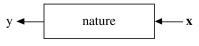
Statistical Modeling: The Two Cultures

Leo Breiman

Abstract. There are two cultures in the use of statistical modeling to reach conclusions from data. One assumes that the data are generated by a given stochastic data model. The other uses algorithmic models and treats the data mechanism as unknown. The statistical community has been committed to the almost exclusive use of data models. This commitment has led to irrelevant theory, questionable conclusions, and has kept statisticians from working on a large range of interesting current problems. Algorithmic modeling, both in theory and practice, has developed rapidly in fields outside statistics. It can be used both on large complex data sets and as a more accurate and informative alternative to data modeling on smaller data sets. If our goal as a field is to use data to solve problems, then we need to move away from exclusive dependence on data models and adopt a more diverse set of tools.

1. INTRODUCTION

Statistics starts with data. Think of the data as being generated by a black box in which a vector of input variables **x** (independent variables) go in one side, and on the other side the response variables **y** come out. Inside the black box, nature functions to associate the predictor variables with the response variables, so the picture is like this:



There are two goals in analyzing the data:

Prediction. To be able to predict what the responses are going to be to future input variables;
Information. To extract some information about how nature is associating the response variables to the input variables.

There are two different approaches toward these goals:

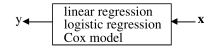
The Data Modeling Culture

The analysis in this culture starts with assuming a stochastic data model for the inside of the black box. For example, a common data model is that data are generated by independent draws from

response variables = f(predictor variables, random noise, parameters)

Leo Breiman is Professor, Department of Statistics, University of California, Berkeley, California 94720-4735 (e-mail: leo@stat.berkeley.edu).

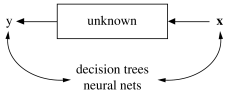
The values of the parameters are estimated from the data and the model then used for information and/or prediction. Thus the black box is filled in like this:



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

The Algorithmic Modeling Culture

The analysis in this culture considers the inside of the box complex and unknown. Their approach is to find a function $f(\mathbf{x})$—an algorithm that operates on **x** to predict the responses **y**. Their black box looks like this:



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

In this paper I will argue that the focus in the statistical community on data models has:

- Led to irrelevant theory and questionable scientific conclusions;

Note: You are highly encouraged to read Leo Breiman's paper on the two cultures of statistical modeling. It's a classic and a must-read for anyone interested in data science and machine learning. See Breiman 2001.

Advanced Techniques for Regression

Ridge Regression

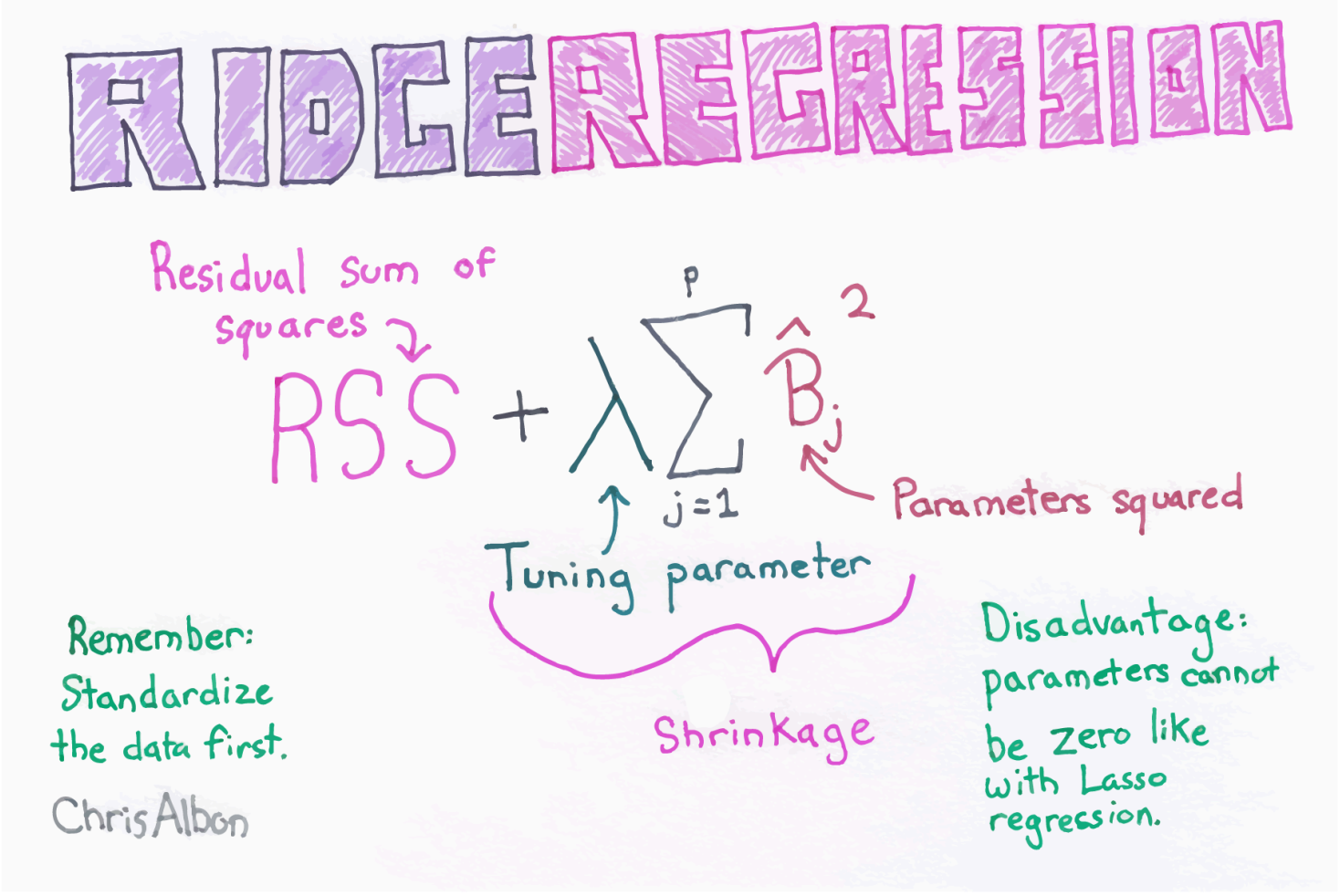


Image Source: Chris Albon. (2017). Machine Learning Flashcards. Flashcards were purchased by the author and are not available for free.

Lasso Regression

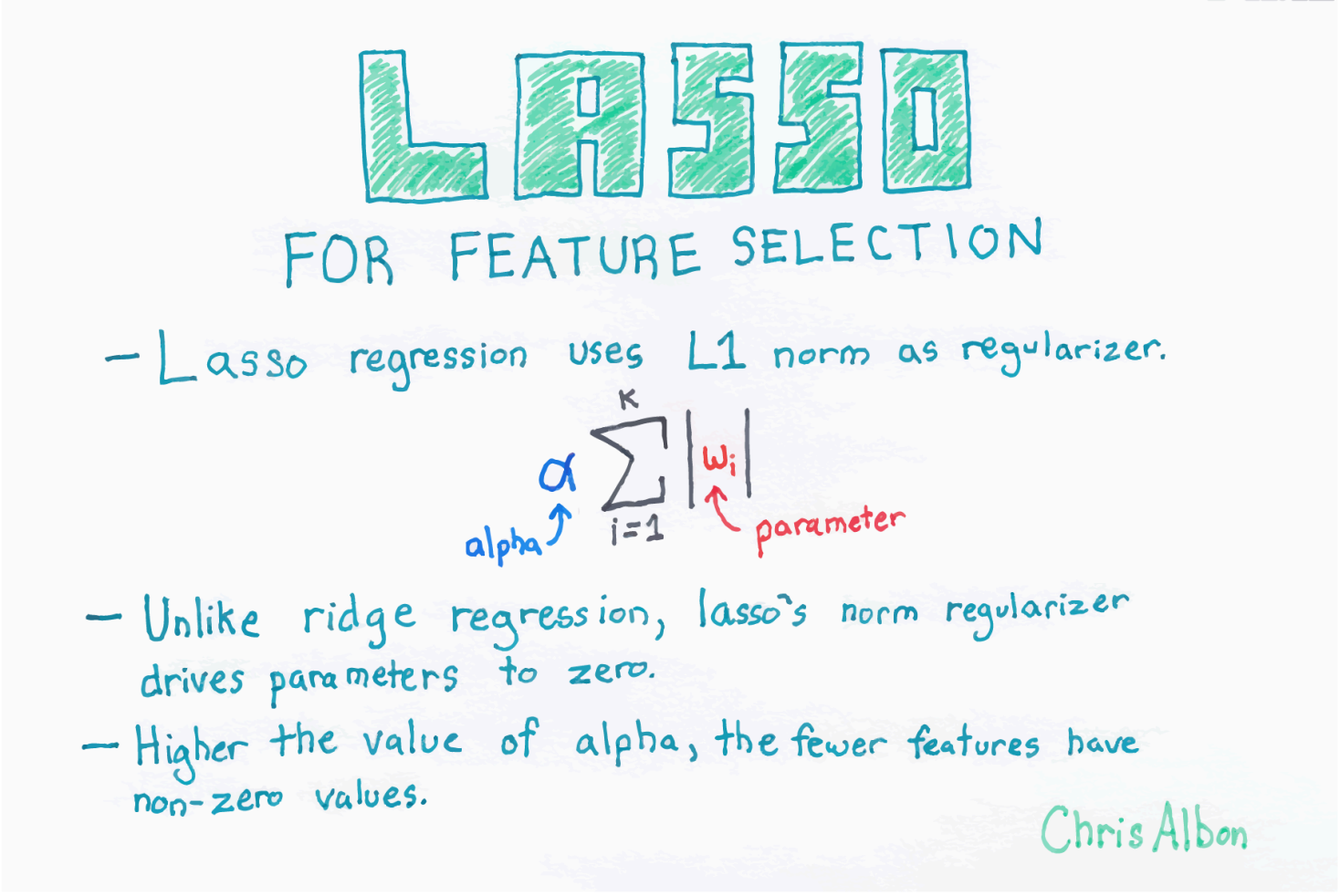


Image Source: Chris Albon. (2017). Machine Learning Flashcards. Flashcards were purchased by the author and are not available for free.

Non-Linear Models for Regression

Decision Trees (Classification and Regression Trees)

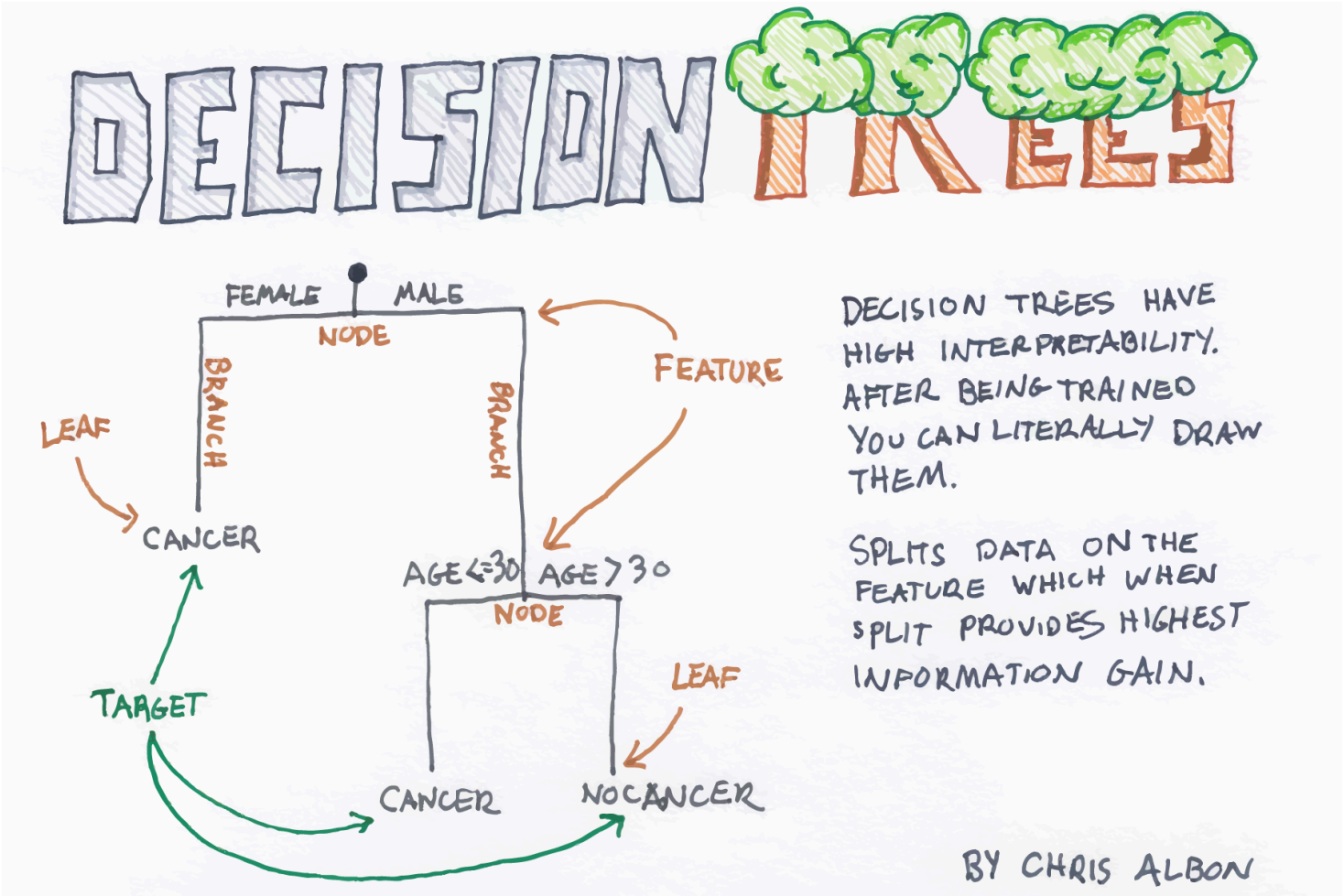


Image Source: Chris Albon. (2017). Machine Learning Flashcards. Flashcards were purchased by the author and are not available for free.

Decision Trees (Classification and Regression Trees)

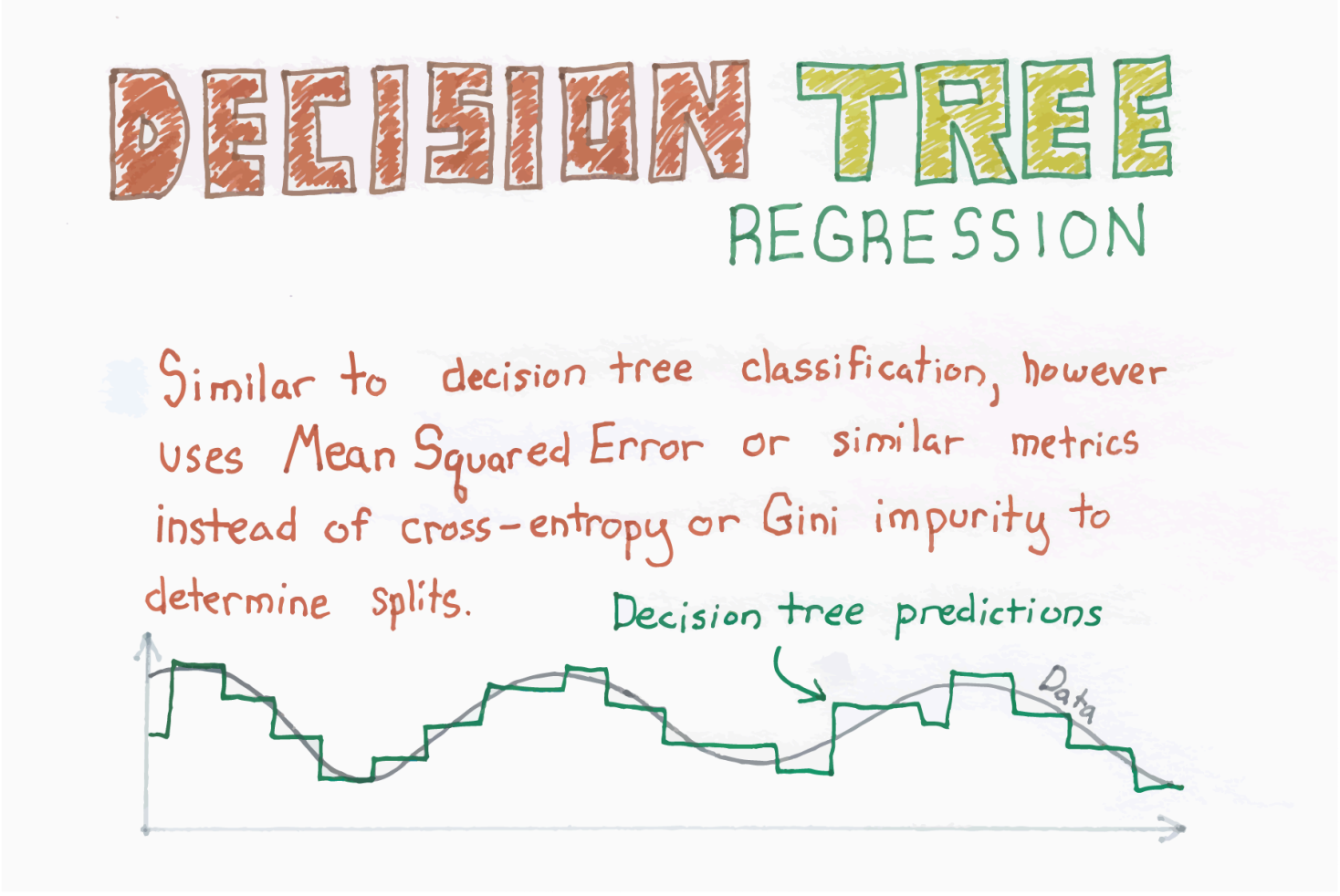


Image Source: Chris Albon. (2017). Machine Learning Flashcards. Flashcards were purchased by the author and are not available for free.

Random Forests

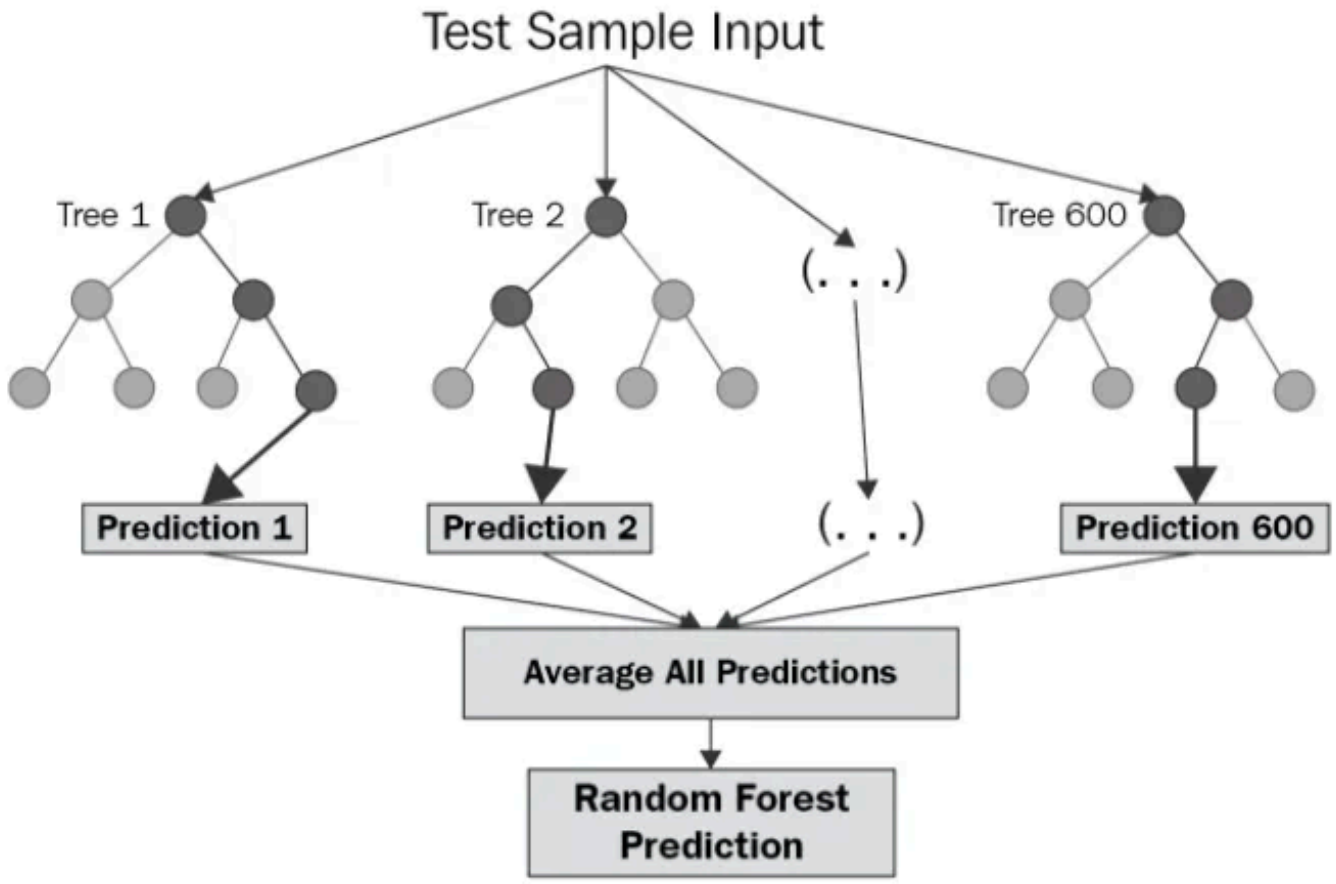


Image Source: Chaya. (2020). Random Forest Regression.

Random Forests

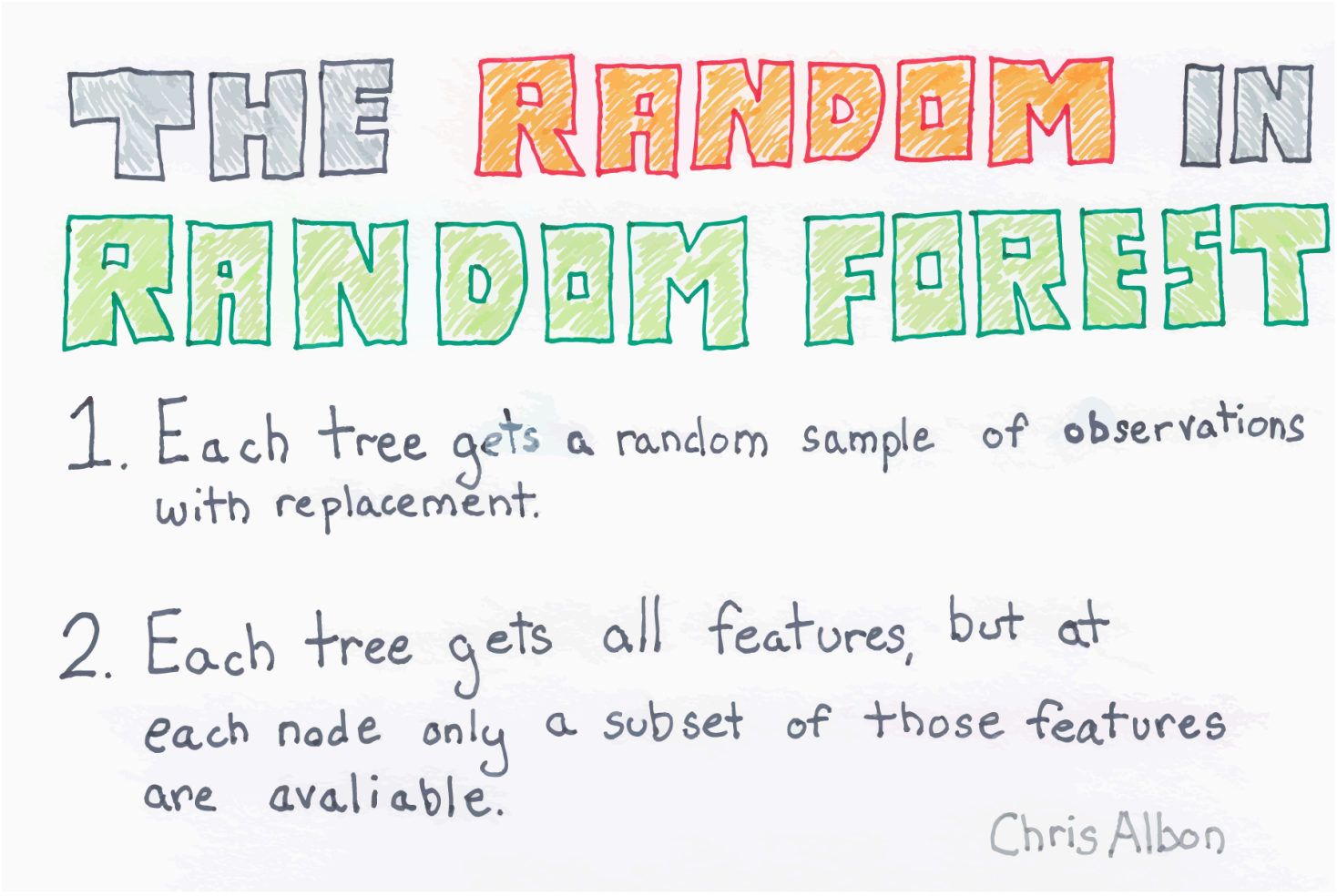


Image Source: Chris Albon. (2017). Machine Learning Flashcards. Flashcards were purchased by the author and are not available for free.

Applications to Cybersecurity Datasets

Example: Predicting Flow Duration in IoT Networks

```
from ucimlrepo import fetch_ucirepo

# fetch dataset
rt_iot2022 = fetch_ucirepo(id=942)

# data (as pandas dataframes)
X = rt_iot2022.data.features
y = rt_iot2022.data.targets
df = pd.concat([X, y], axis=1)

# subset the features and dataset
df = (
  df
  .sample(frac=0.1, random_state=2024)
  .reset_index(drop=True)
  .loc[:, ['proto', 'service', 'fwd_pkts_payload.avg', 'bwd_pkts_payload.avg', 'fwd_subflow_
  )
```

Note: The dataset is from the UCI Machine Learning Repository. It contains features of network traffic in IoT networks and the target variable is the flow duration. The goal is to predict the flow duration using the features. For more information, see UCI RT-IoT2022.

Recap

Summary of Main Points

By now, you should be able to do the following:

- Recognize the differences and similarities between traditional regression and machine learning regression models
- Describe advanced techniques like Ridge and Lasso
- Explore non-linear models like Decision Trees and Random Forests for regression tasks
- Apply machine learning regression models to cybersecurity datasets

 Review and Clarification 

- **Class Notes**: Take some time to revisit your class notes for key insights and concepts.
- **Zoom Recording**: The recording of today's class will be made available on Canvas approximately 3-4 hours after the end of class.
- **Questions**: Please don't hesitate to ask for clarification on any topics discussed in class. It's crucial not to let questions accumulate.