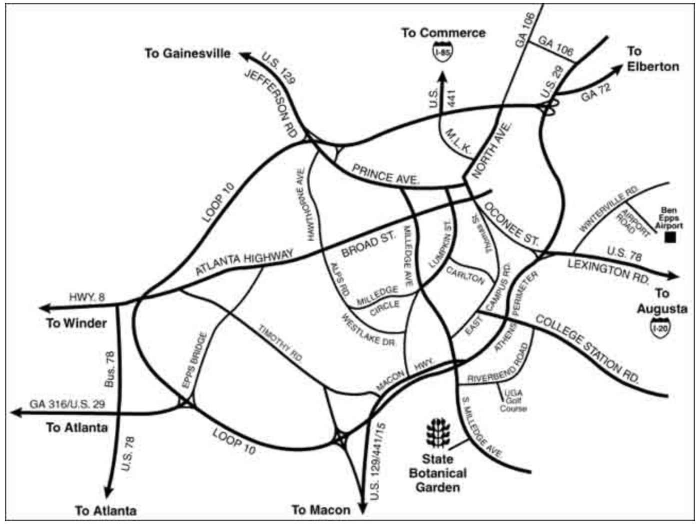
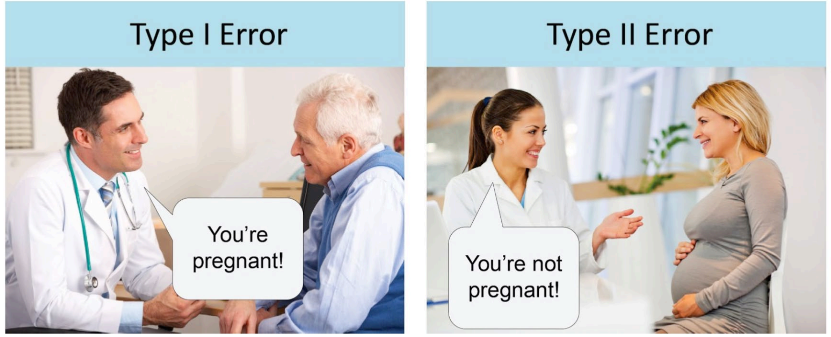
# Model Basics

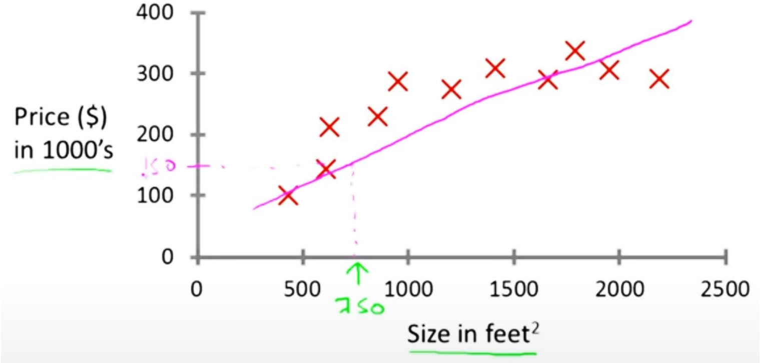
John Rios

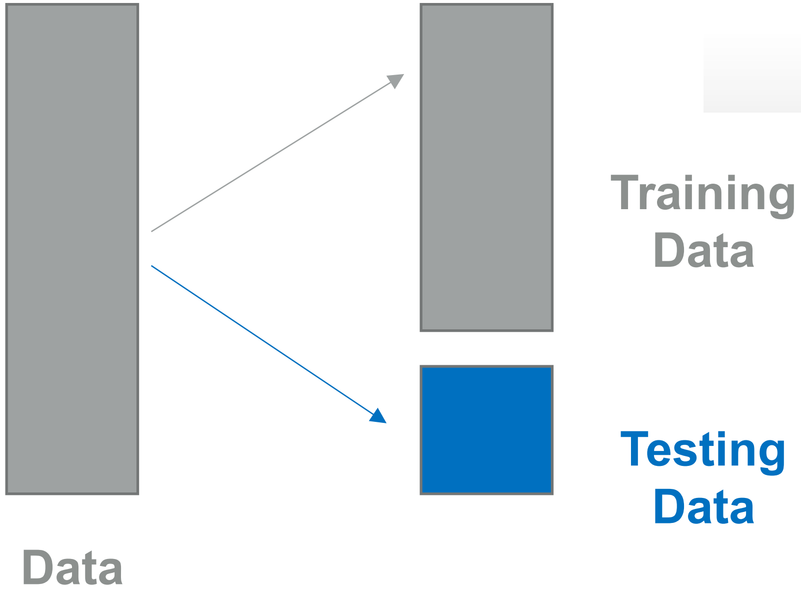
*Business Intelligence and Analytics*

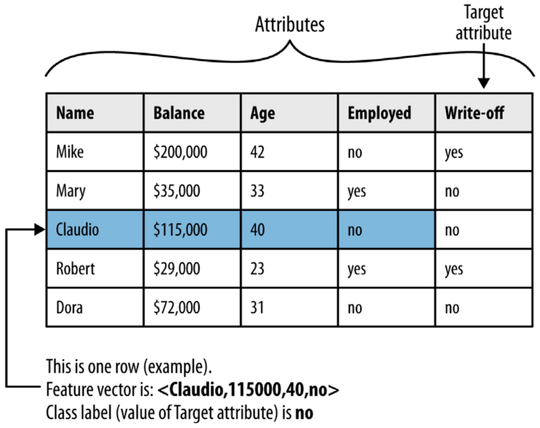
Terry College of Business
UNIVERSITY OF GEORGIA







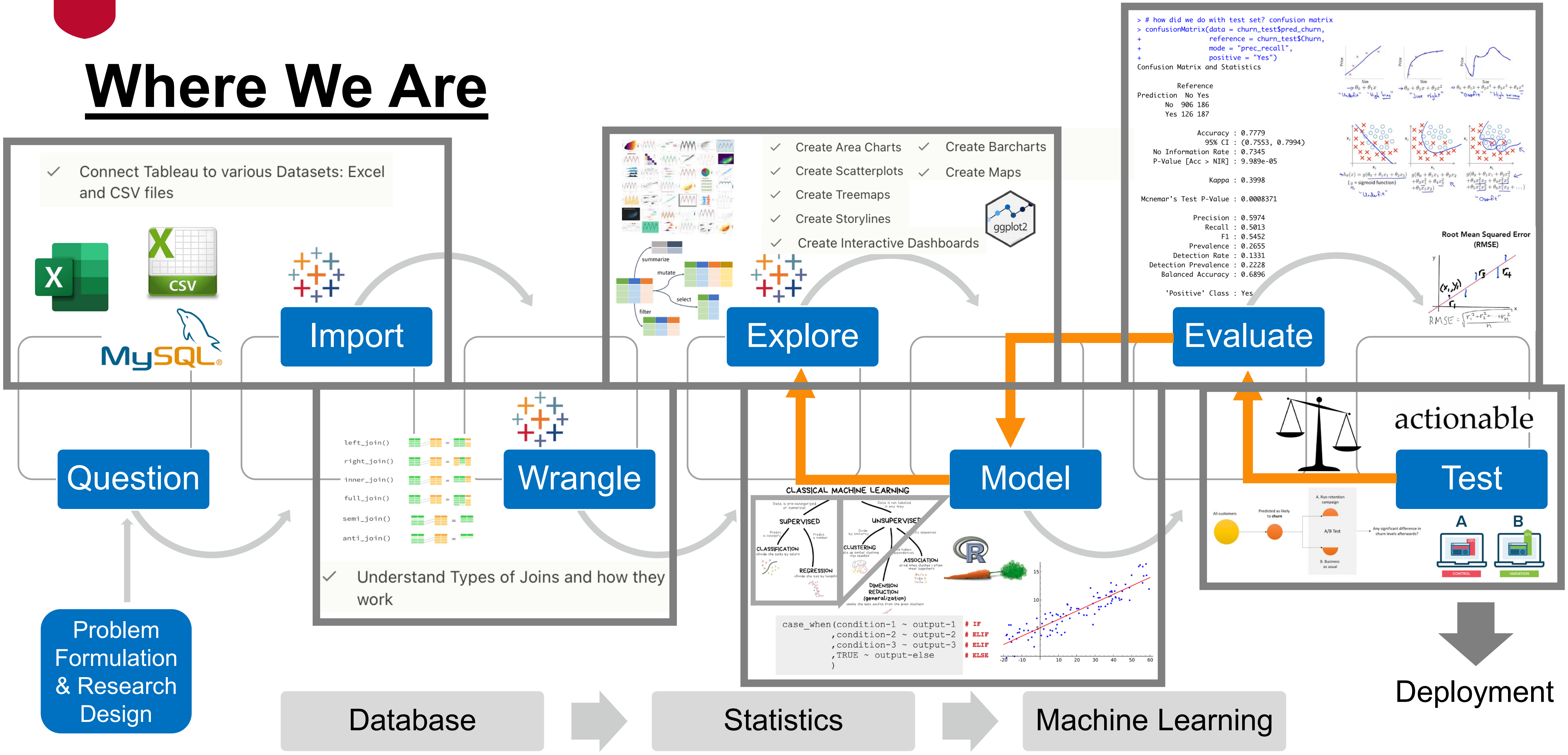



Attributes
Target attribute

| Name | Balance | Age | Employed | Write-off |
|---|---|---|---|---|
| Mike | $200,000 | 42 | no | yes |
| Mary | $35,000 | 33 | yes | no |
| Claudio | $115,000 | 40 | no | no |
| Robert | $29,000 | 23 | yes | yes |
| Dora | $72,000 | 31 | no | no |

This is one row (example).
Feature vector is: **<Claudio,115000,40,no>**
Class label (value of Target attribute) is **no**

# Where We Are

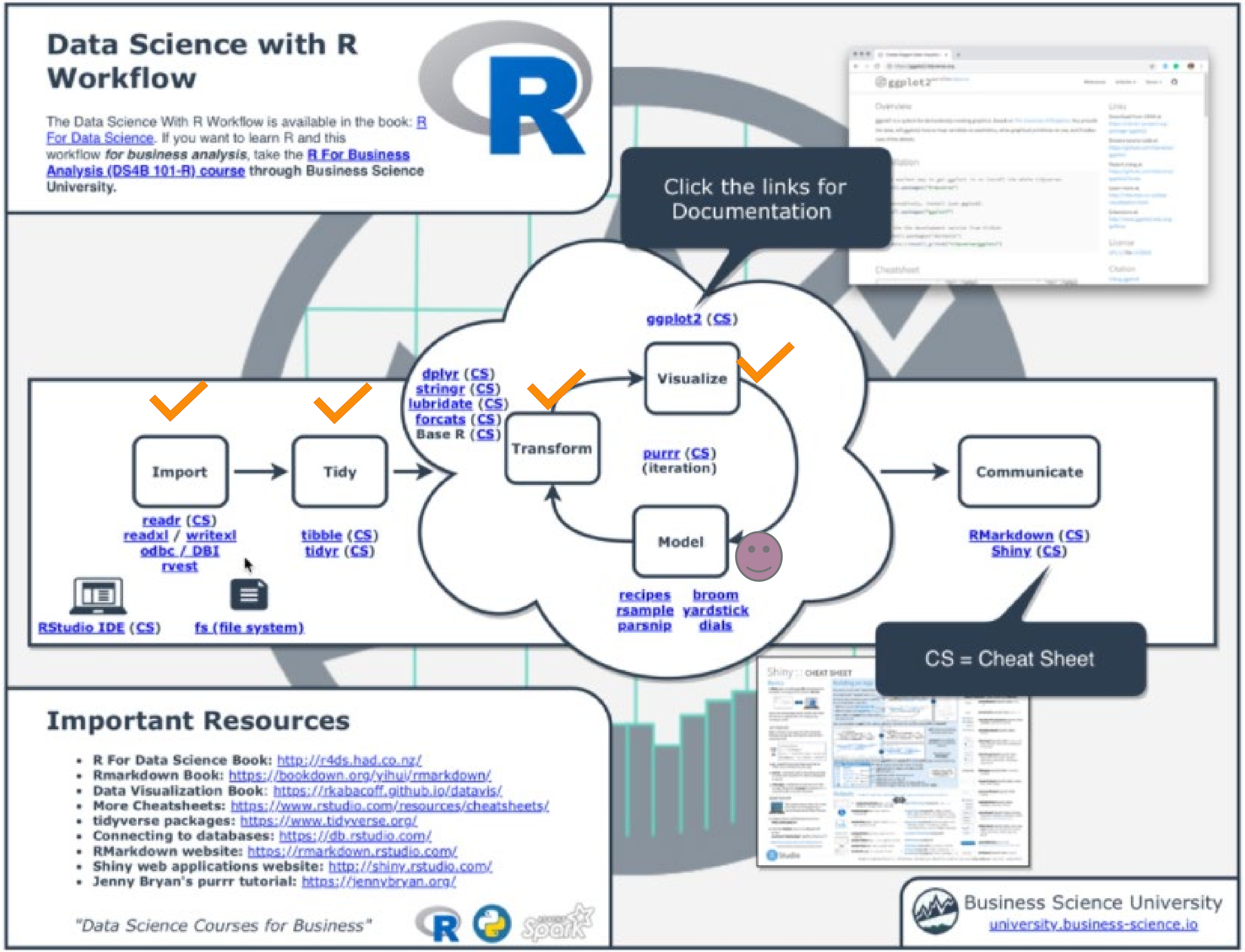


- ✓ Connect Tableau to various Datasets: Excel and CSV files
- ✓ Create Area Charts
- ✓ Create Scatterplots
- ✓ Create Treemaps
- ✓ Create Storylines
- ✓ Create Interactive Dashboards
- ✓ Create Barcharts
- ✓ Create Maps
- ✓ Understand Types of Joins and how they work

```
> # how did we do with test set? confusion matrix
> confusionMatrix(data = churn_test$pred_churn,
+                 reference = churn_test$Churn,
+                 mode = "prec_recall",
+                 positive = "Yes")
Confusion Matrix and Statistics

          Reference
Prediction  No Yes
       No  906 186
       Yes 126 187

               Accuracy : 0.7779
                 95% CI : (0.7553, 0.7994)
    No Information Rate : 0.7345
    P-Value [Acc > NIR] : 9.989e-05

                  Kappa : 0.3998

 Mcnemar's Test P-Value : 0.0008371

              Precision : 0.5974
                 Recall : 0.5013
                     F1 : 0.5452
             Prevalence : 0.2655
         Detection Rate : 0.1331
   Detection Prevalence : 0.2228
      Balanced Accuracy : 0.6896

       'Positive' Class : Yes
```

```
left_join()
right_join()
inner_join()
full_join()
semi_join()
anti_join()
```

```
case_when(condition-1 ~ output-1   # IF
         ,condition-2 ~ output-2   # ELIF
         ,condition-3 ~ output-3   # ELIF
         ,TRUE ~ output-else       # ELSE
         )
```



# Data Science with R Workflow

The Data Science With R Workflow is available in the book: R For Data Science. If you want to learn R and this workflow ***for business analysis***, take the **R For Business Analysis (DS4B 101-R) course** through **Business Science University.**

Click the links for Documentation

Import → Tidy → Transform → Visualize → Model → Communicate

Import: readr (CS), readxl / writexl, odbc / DBI, rvest

Tidy: tibble (CS), tidyr (CS)

Transform: dplyr (CS), stringr (CS), lubridate (CS), forcats (CS), Base R (CS)

Visualize: ggplot2 (CS)

purrr (CS) (iteration)

Model: recipes, rsample, parsnip, broom, yardstick, dials

Communicate: RMarkdown (CS), Shiny (CS)

RStudio IDE (CS)

fs (file system)

CS = Cheat Sheet

## Important Resources

- R For Data Science Book: http://r4ds.had.co.nz/
- Rmarkdown Book: https://bookdown.org/yihui/rmarkdown/
- Data Visualization Book: https://rkabacoff.github.io/datavis/
- More Cheatsheets: https://www.rstudio.com/resources/cheatsheets/
- tidyverse packages: https://www.tidyverse.org/
- Connecting to databases: https://db.rstudio.com/
- RMarkdown website: https://rmarkdown.rstudio.com/
- Shiny web applications website: http://shiny.rstudio.com/
- Jenny Bryan's purrr tutorial: https://jennybryan.org/

*"Data Science Courses for Business"*

Business Science University
university.business-science.io

https://www.business-science.io/r-cheatsheet.html

# Model Defined

A **simplified* representation** of reality created for a **specific purpose**

- ***based on some assumptions** about what is and is not important, or sometimes based on constraints on information or tractability

Provost, F. & Fawcett, T. (2013). Data Science for Business.

# Model Goal

Not to uncover truth, but to discover a simple approximation that is still useful

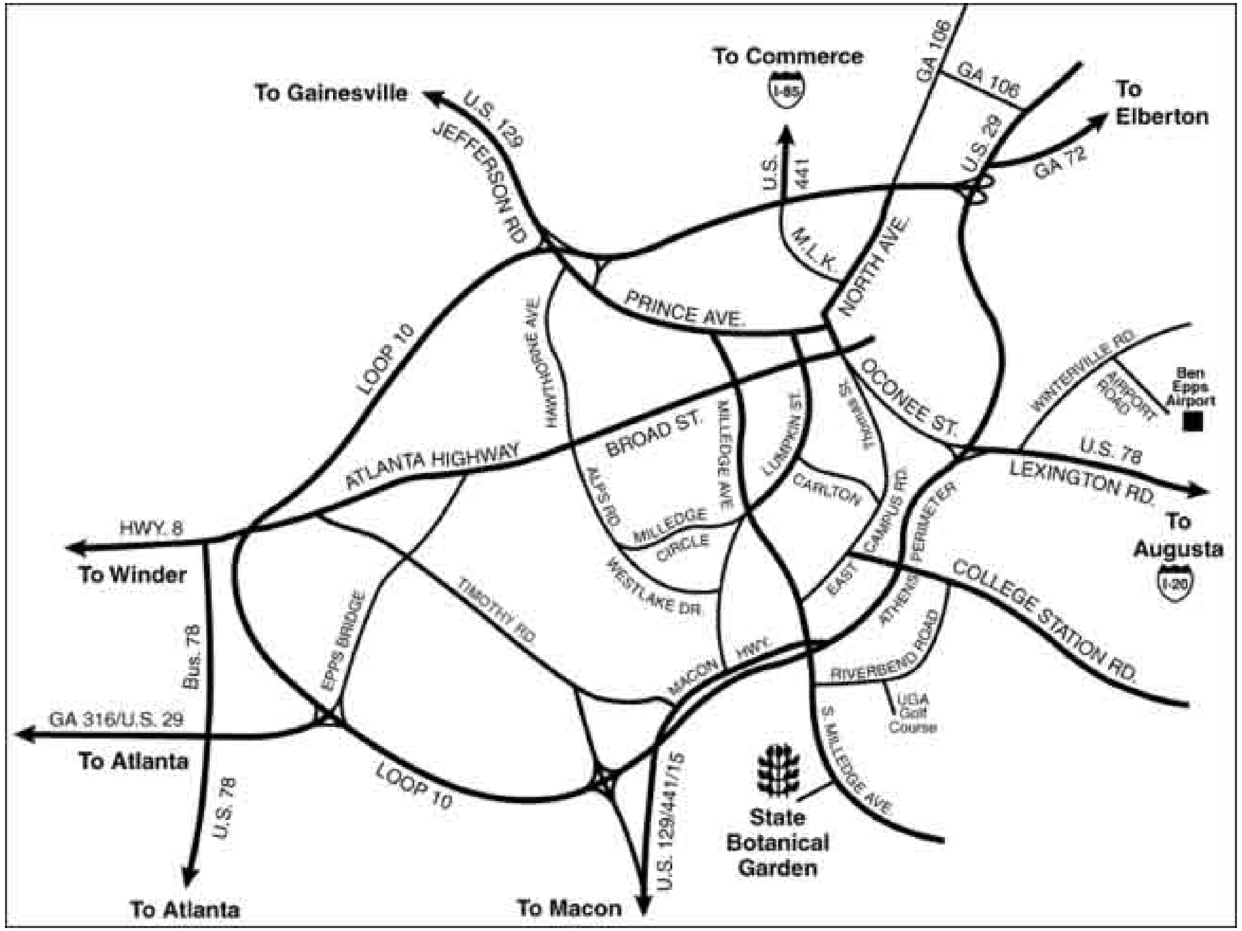
- **i.e., capture true “signals”** (or patterns generated by the phenomenon of interest) **and ignore “noise”** (or random variation that you’re not interested in)

***“All models are wrong, but some are useful”***

George Box

Provost, F. & Fawcett, T. (2013). Data Science for Business.

# Model Example



# Predictive (or Supervised Learning) Model

A formula for estimating the unknown value of interest:

- **The target!**
    - The formula could be mathematical or a logical statement, such as a rule. Often, it is a hybrid

Provost, F. & Fawcett, T. (2013). Data Science for Business.

# More Terminology

**Instance / example** = a fact or data point described by a set of attributes (also known as variables, columns, or features)

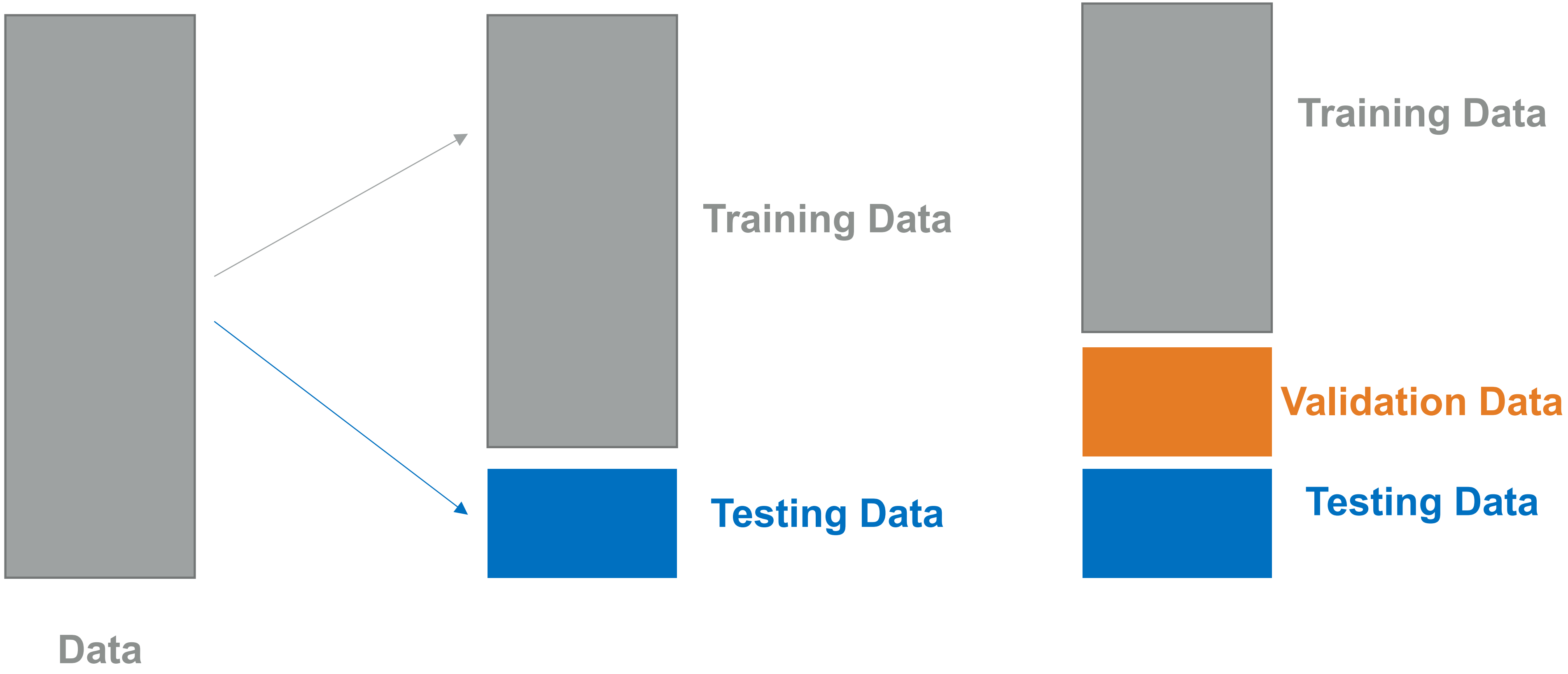
**Model induction** = the creation of models from data

**Training data** = the input data used for model induction

**Testing data** = the input data used for model testing

Provost, F. & Fawcett, T. (2013). Data Science for Business.

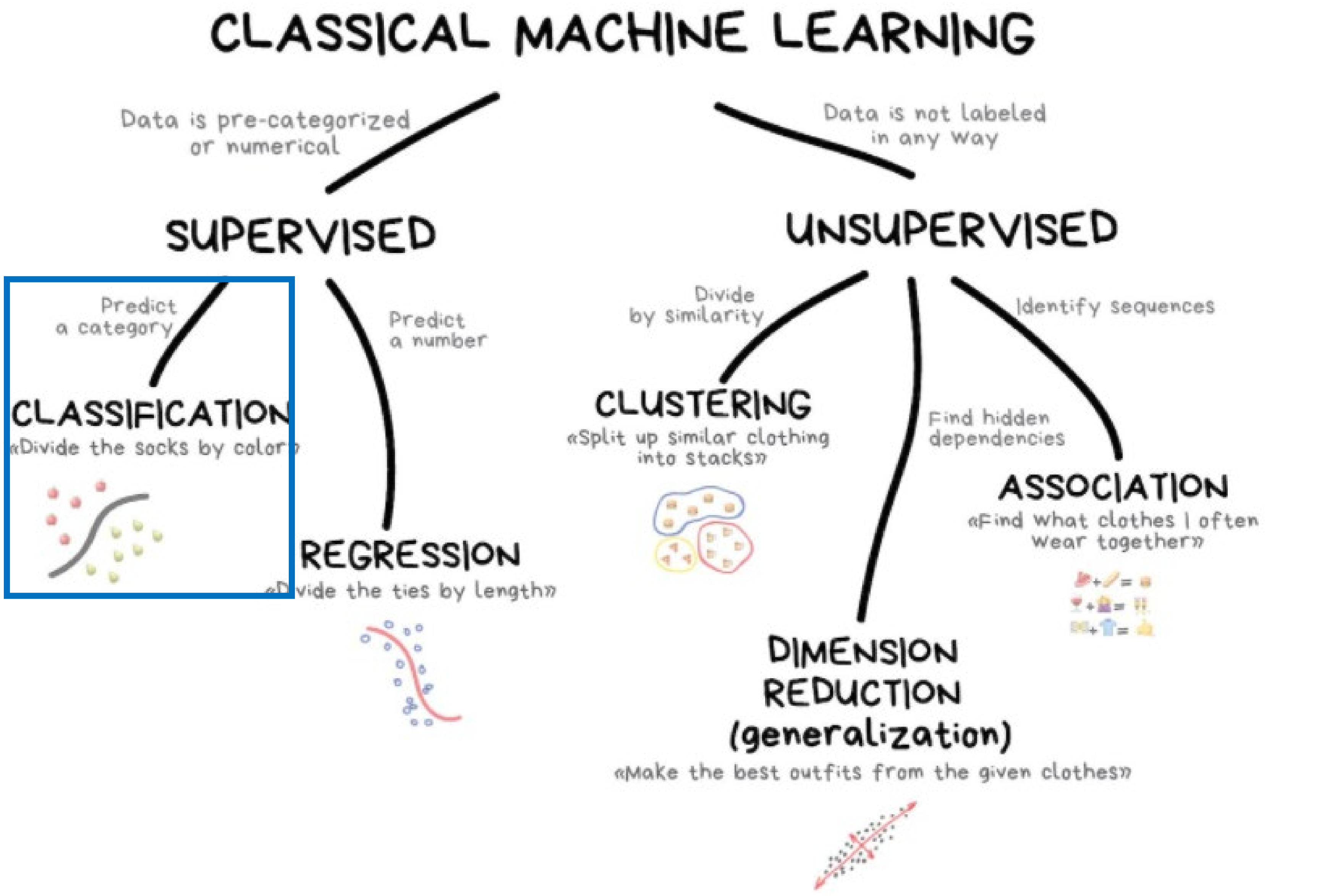
# Predictive Model: Data Splitting



# Classification or Regression?

- Is this email spam?
- What will Google’s stock price be tomorrow?
- Will Google’s stock price go up tomorrow?
- Is this X/Twitter account a bot?
- What will the total electricity demand be tomorrow in Athens, GA?

Provost, F. & Fawcett, T. (2013). Data Science for Business.

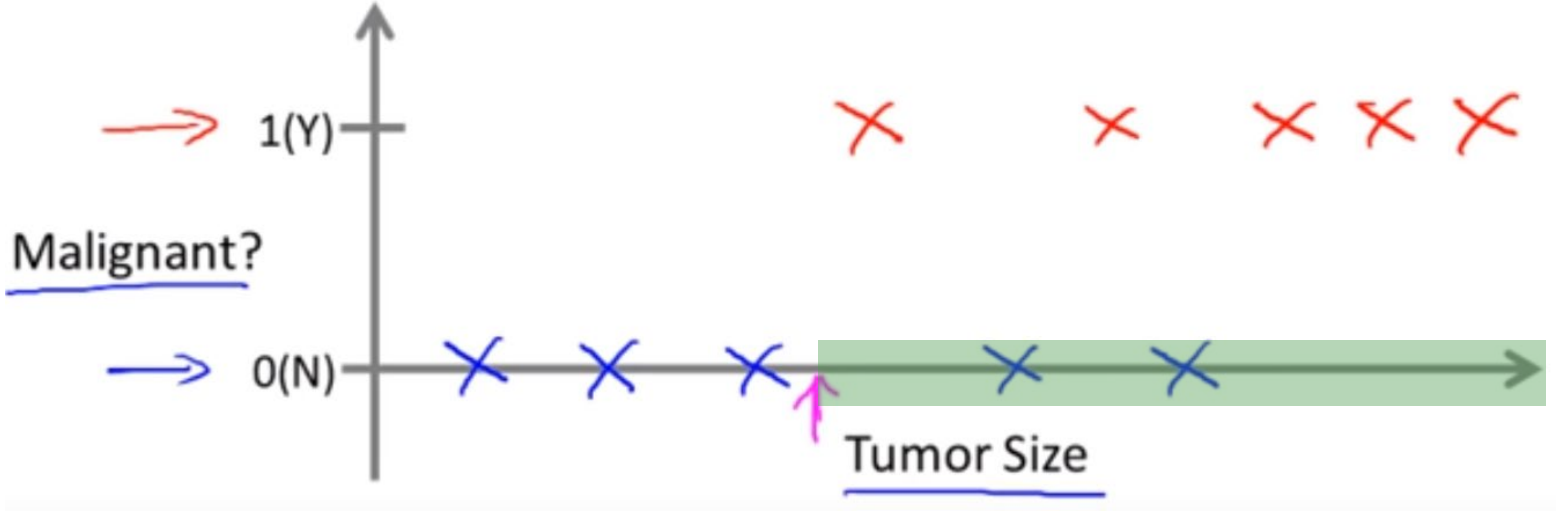

CLASSICAL MACHINE LEARNING
Data is pre-categorized or numerical
Data is not labeled in any way
SUPERVISED
UNSUPERVISED
Predict a category
Predict a number
CLASSIFICATION
«Divide the socks by color»
REGRESSION
«Divide the ties by length»
Divide by similarity
Identify sequences
CLUSTERING
«Split up similar clothing into stacks»
Find hidden dependencies
ASSOCIATION
«Find what clothes I often wear together»
DIMENSION REDUCTION (generalization)
«Make the best outfits from the given clothes»

https://vas3k.com/blog/machine_learning/

# Predictive Model: Classification

Suppose you want to predict whether someone's breast cancer is malignant

**Estimate class probability (e.g., with a logistic regression)**



https://www.youtube.com/watch?v=bQI5uDxrFfA

# Predictive Model: Classification Performance

| | | Actual | |
|---|---|---|---|
| | | Positive | Negative |
| Predicted | Positive | True Positives | False Positives |
| | Negative | False Negatives | True Negatives |

- An email sent from your friend identified as spam (+: spam)
- A fraud successfully caught by the system monitor (+: fraud)
- An internet intrusion passed as a normal activity (+: intrusion)
- A recovered patient approved to be discharged (+: still sick)

# Predictive Model: Classification Performance

## Additional metrics

|  |  | Actual |  |
|---|---|---|---|
|  |  | **Positive** | **Negative** |
| **Predicted** | **Positive** | **True Positives** | **False Positives** |
|  | **Negative** | **False Negatives** | **True Negatives** |

- Precision = true positives / (true positives + false positives)
- Recall = true positives / (true positives + false negatives)

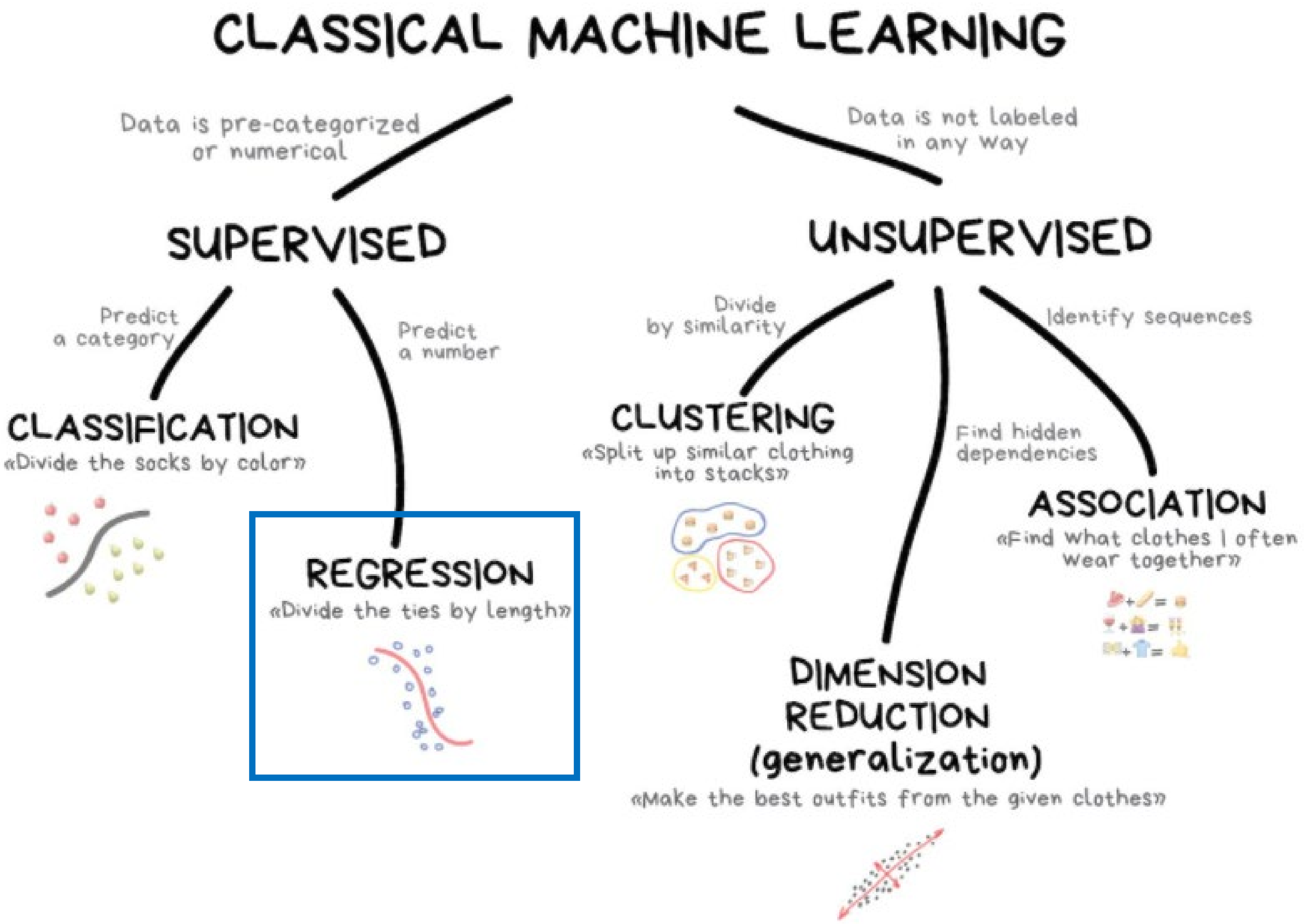
Provost, F. & Fawcett, T. (2013). Data Science for Business.

# Exercise

Model X is developed to help doctors diagnose breast cancer. Its performance was evaluated over a test data set of 1000 patients.

- What is the accuracy of Model X?
- What is the precision and recall?

| | | Actual | |
|---|---|---|---|
| | | + | - |
| Predicted | + | 16 | 4 |
| | - | 3 | 977 |

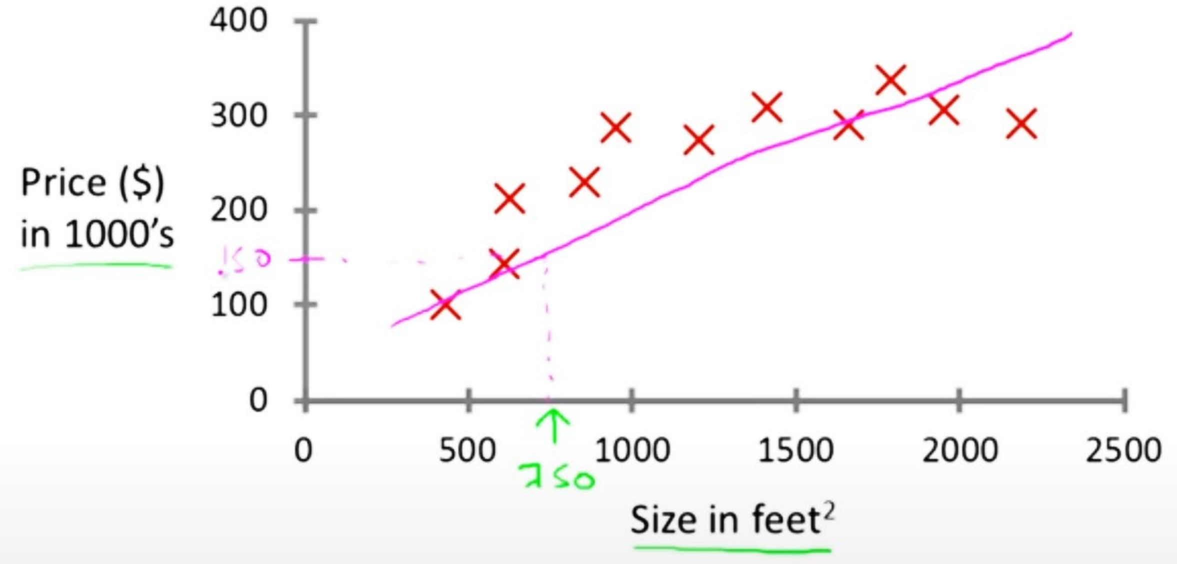

CLASSICAL MACHINE LEARNING
Data is pre-categorized or numerical
Data is not labeled in any way
SUPERVISED
UNSUPERVISED
Predict a category
Predict a number
Divide by similarity
Identify sequences
CLASSIFICATION
«Divide the socks by color»
REGRESSION
«Divide the ties by length»
CLUSTERING
«Split up similar clothing into stacks»
Find hidden dependencies
ASSOCIATION
«Find what clothes I often wear together»
DIMENSION REDUCTION (generalization)
«Make the best outfits from the given clothes»

https://vas3k.com/blog/machine_learning/

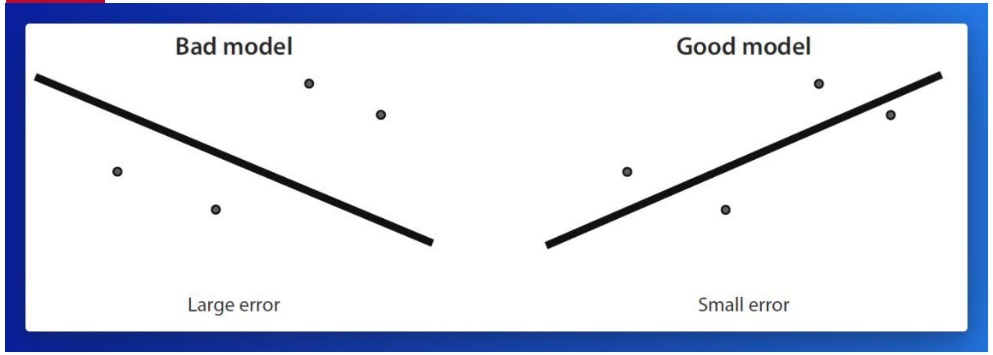
# Predictive Model: Regression

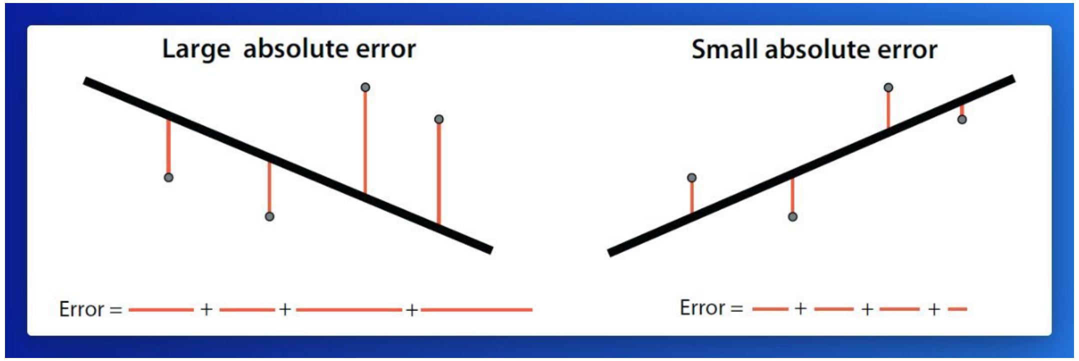
Suppose you want to predict house prices, and you have some data about the price of a house (in thousands of $) over size (sqft)

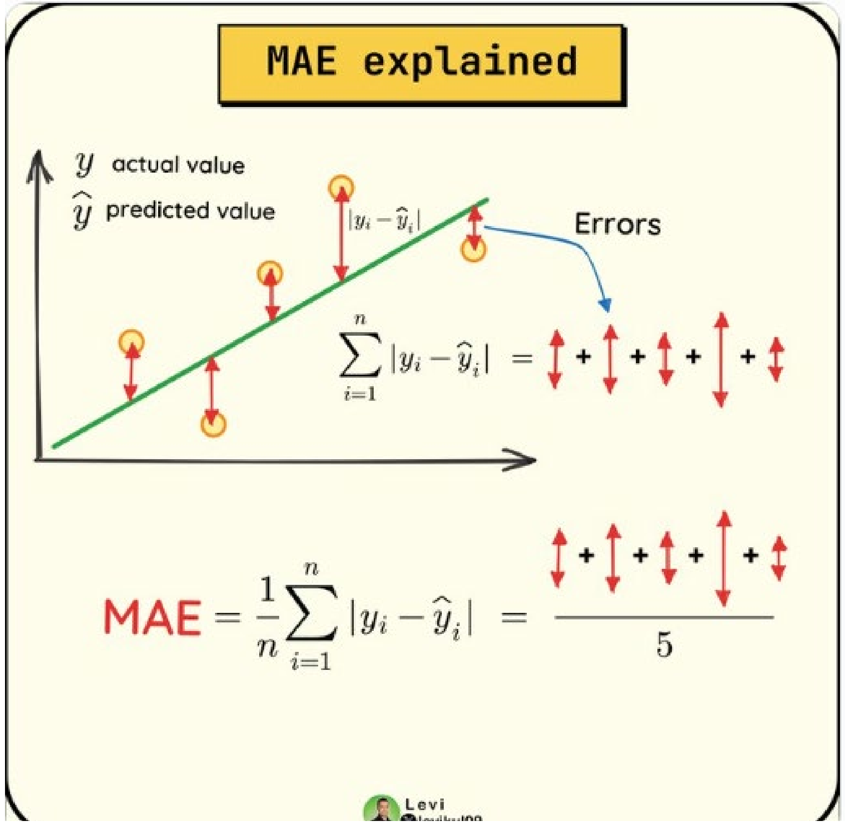
**Estimate numeric value (e.g., with a linear regression)**

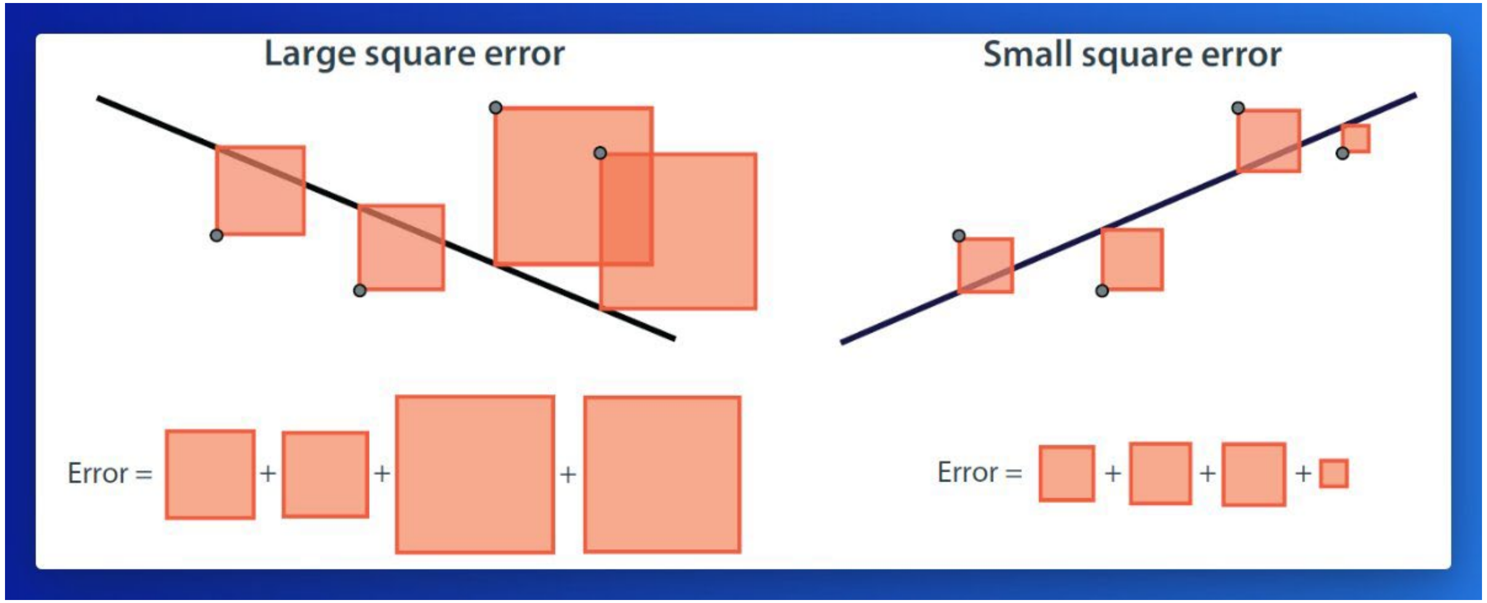


https://www.youtube.com/watch?v=bQI5uDxrFfA






MAE explained
$y$ actual value
$\hat{y}$ predicted value
$|y_i - \hat{y}_i|$
Errors
$\sum_{i=1}^{n} |y_i - \hat{y}_i| =$
$\text{MAE} = \frac{1}{n}\sum_{i=1}^{n} |y_i - \hat{y}_i| = \frac{\phantom{x}}{5}$
Levi



https://threadreaderapp.com/thread/1730168597040013608.html

# Model Evaluation

Error for data record = predicted (p) minus actual (a)

**RMSE: Root Mean Squared Error:** $\sqrt{\frac{1}{n}\sum_{1}^{n}\left(Y_i - \hat{Y}_i\right)^2}$

MAE: Mean Absolute Error: $\frac{1}{n}\sum_{1}^{n}\left|\left(Y_i - \hat{Y}_i\right)\right|$

MAPE: Mean Absolute Percentage Error: $\frac{100}{n}\sum_{1}^{n}\left|\frac{Y_i - \hat{Y}_i}{Y_i}\right|$

Total SSE: Total Sum of Squared Errors: $\sum_{1}^{n}\left(Y_i - \hat{Y}_i\right)^2$

# RMSE

Error for data record = predicted (p) minus actual (a)

RMSE = how much the *p*'s diverge from the *a*'s, on average

Assume the regression equation is y = 1.74x. What is the root mean squared error for the sample dataset?

| x | a | p | (p – a)^2 |
|---|---|---|---|
| 1 | 2 | 1.74 | 0.0676 |
| 2 | 5 | 3.48 | 2.3104 |
| -1 | -2 | -1.74 | 0.0676 |

**RMSE =**

$$\sqrt{(0.0676 + 2.3104 + 0.0676) / 3}$$

$$= \sqrt{0.8152}$$

$$= .903$$

# Things to Consider

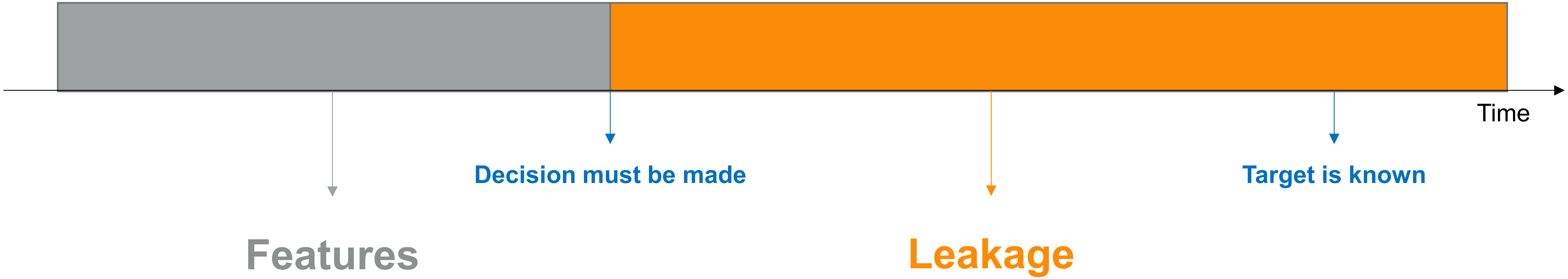
- Is there a specific, quantifiable target that you are interested in predicting?
    - If yes, is it a **class or a number**?
        - Think about the decision

- Do you have data on the target?
    - Do you have **enough** data?
        - If the target is a class, a min of ~500 for each class type is needed

Provost, F. & Fawcett, T. (2013). Data Science for Business.

# Another Thing to Consider

- Do you have relevant data **prior** to the decision?
  - Think about the timing of decision and action leading up to it



Provost, F. & Fawcett, T. (2013). Data Science for Business.

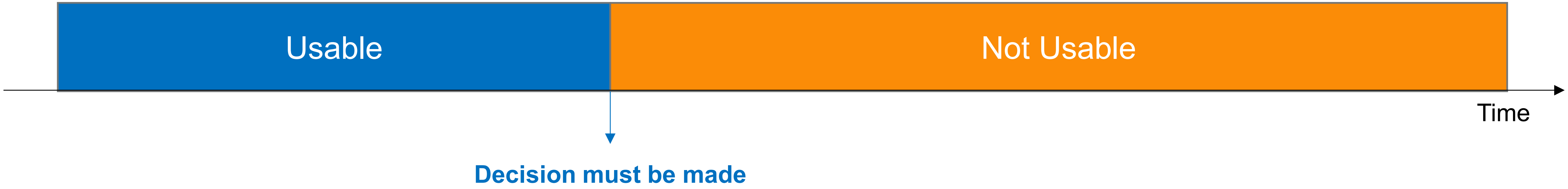
# Predicting Loan Default

What information about its customers can a bank use to predict loan default?

-
-
-

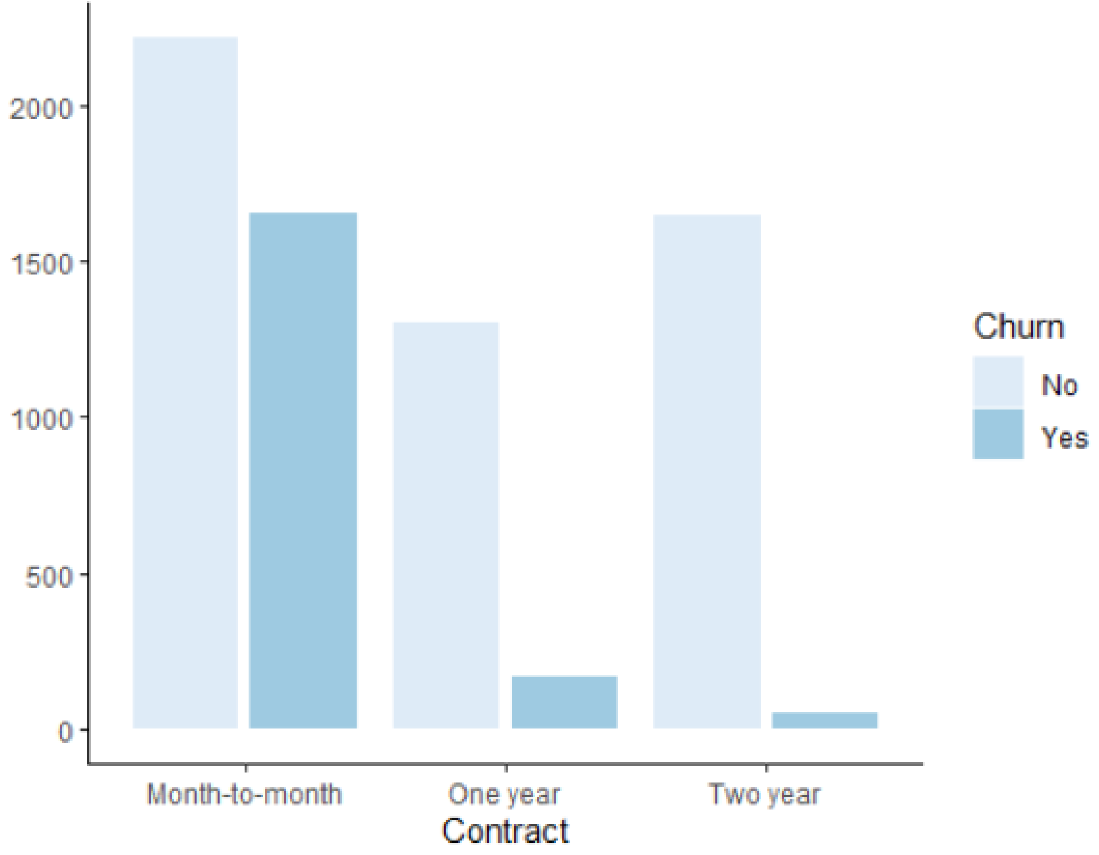
# Avoiding Leakage

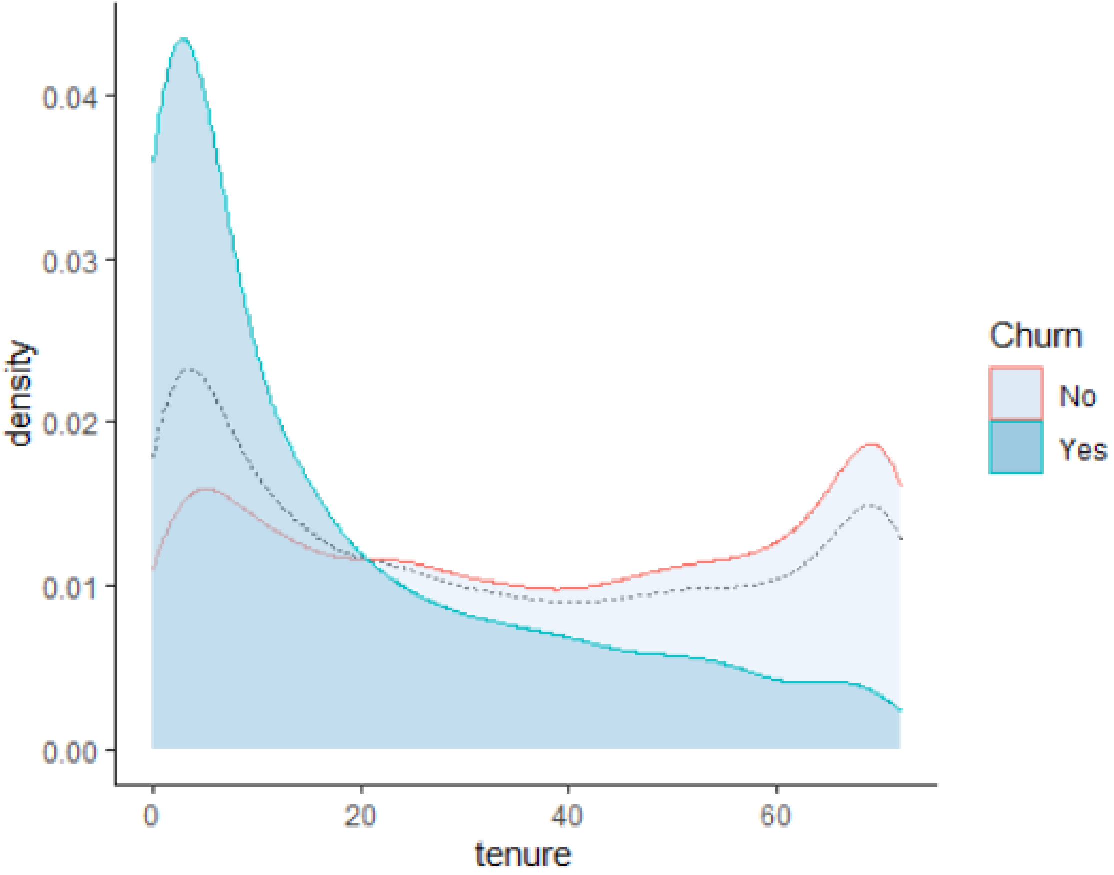
- Do you have relevant data **prior** to the decision?
  - Think about the timing of decision and action leading up to it



Provost, F. & Fawcett, T. (2013). Data Science for Business.

# Exploratory analysis





# Class Activity 1

An online clothing store wants to create a supervised model that will offer personalized clothing recommendations to customers. This means that the model will recommend certain clothes to Janeth and different clothes to Joe. The model will use past purchasing behavior to generate training data. Mention five attributes that are useful for the model and justify your answer.

# Class Activity 1

An online clothing store wants to create a supervised model that will offer personalized clothing recommendations to customers. This means that the model will recommend certain clothes to Janeth and different clothes to Joe. The model will use past purchasing behavior to generate training data. Mention five attributes that are useful for the model and justify your answer.

- Size
- User clicks on the product description
- Clothes beauty