Welcome to GBUS 738

David Svancer

Adjunct Professor of Business Analytics

George Mason University

School of Information Systems and Operations Management

GEORGE MASON UNIVERSITY
School of Business

GBUS 738

Skills You Will Develop In This Course

Fundamentals of Programming with R	The basics of R programming
Data Analysis with the *Tidyverse*	Data analysis and visualization techniques using the popular *tidyverse* R package
Machine Learning with *tidymodels*	Training machine learning models with the *tidymodels* R framework
Managing Analytics Projects and Communicating Business Value	Data analysis and machine learning projects from start to finish using R

Course Goals

Computer Programming Fundamentals

Data types and structures in R

- Vectors, Matrices, Lists and Data frames

Writing Custom Functions for Data Analysis Tasks

```r
my_data <- data.frame(gender = c("M","F","F"),
                      test_1_grade = c(82, 93, 87),
                      hw_1_grade = c(92, 89, 98),
                      session = c("7 AM", "7 PM", "7 AM"))

# View the data
my_data
```

```
  gender test_1_grade hw_1_grade session
1      M           82         92    7 AM
2      F           93         89    7 PM
3      F           87         98    7 AM
```

```r
mean_dev_3 <- function(x) {
                mean_x <- mean(x) # calculate average
                dev_vec <- x - mean_x # calculate deviation vector

                return(list(mean_value = mean_x,
                            dev_vector = dev_vec))
}

my_result <- mean_dev_3(data)
```

School of Business

Course Goals

Data Analysis with the *Tidyverse*

tidyverse.org

R packages for data science

The tidyverse is an opinionated collection of R packages designed for data science. All packages share an underlying design philosophy, grammar, and data structures.

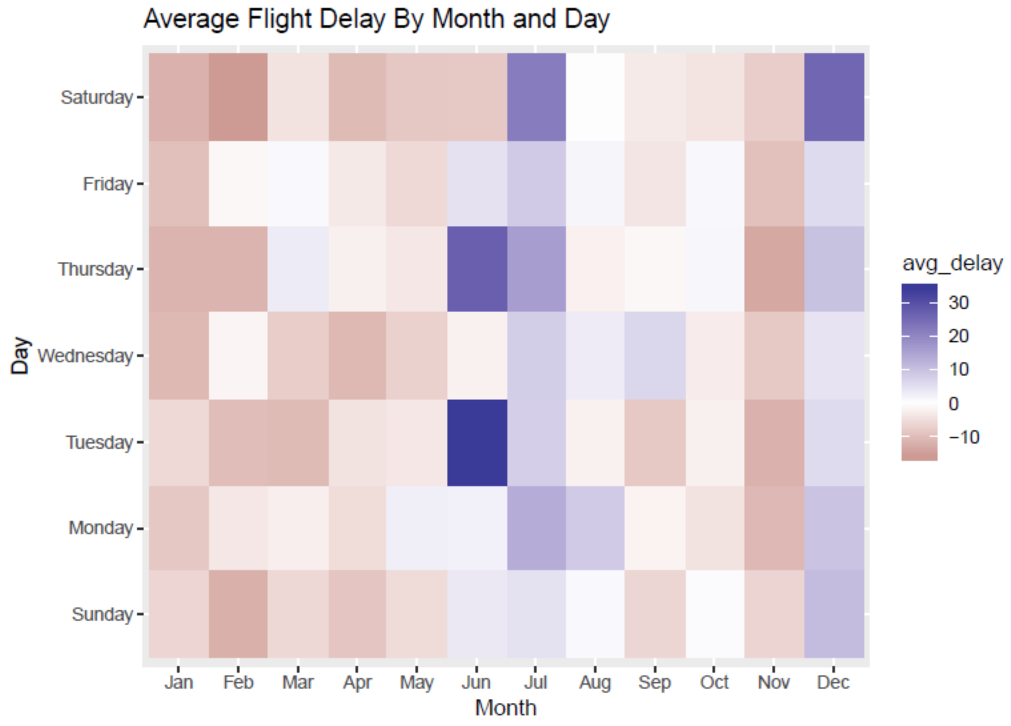
Data Manipulation

```
heart %>% group_by(ChestPain, HeartDisease) %>%
          summarise(patients_n = n(),
                    avg_chol = mean(Cholesterol),
                    sd_chol = sd(Cholesterol))
```

```
# A tibble: 8 x 5
# Groups:   ChestPain [4]
  ChestPain    HeartDisease patients_n avg_chol sd_chol
  <chr>        <chr>             <int>    <dbl>   <dbl>
1 asymptomatic No                   39     245.    48.9
2 asymptomatic Yes                 103     253.    52.9
3 nonanginal   No                   65     247.    64.7
4 nonanginal   Yes                  18     239     43.8
5 nontypical   No                   40     241.    45.3
```

Data Visualization

```
ggplot(data = average_delays, mapping = aes(x = month_text, y = day_text,
                                            fill = avg_delay)) +
  geom_tile() +
  scale_fill_gradient2() +
  labs(title = "Average Flight Delay By Month and Day",
       x = "Month", y = "Day")
```



Data Wrangling and Reshaping

Country	1999	2000
Afghanistan	745	2,666
Brazil	37,737	80,488
China	212,258	213,766

→

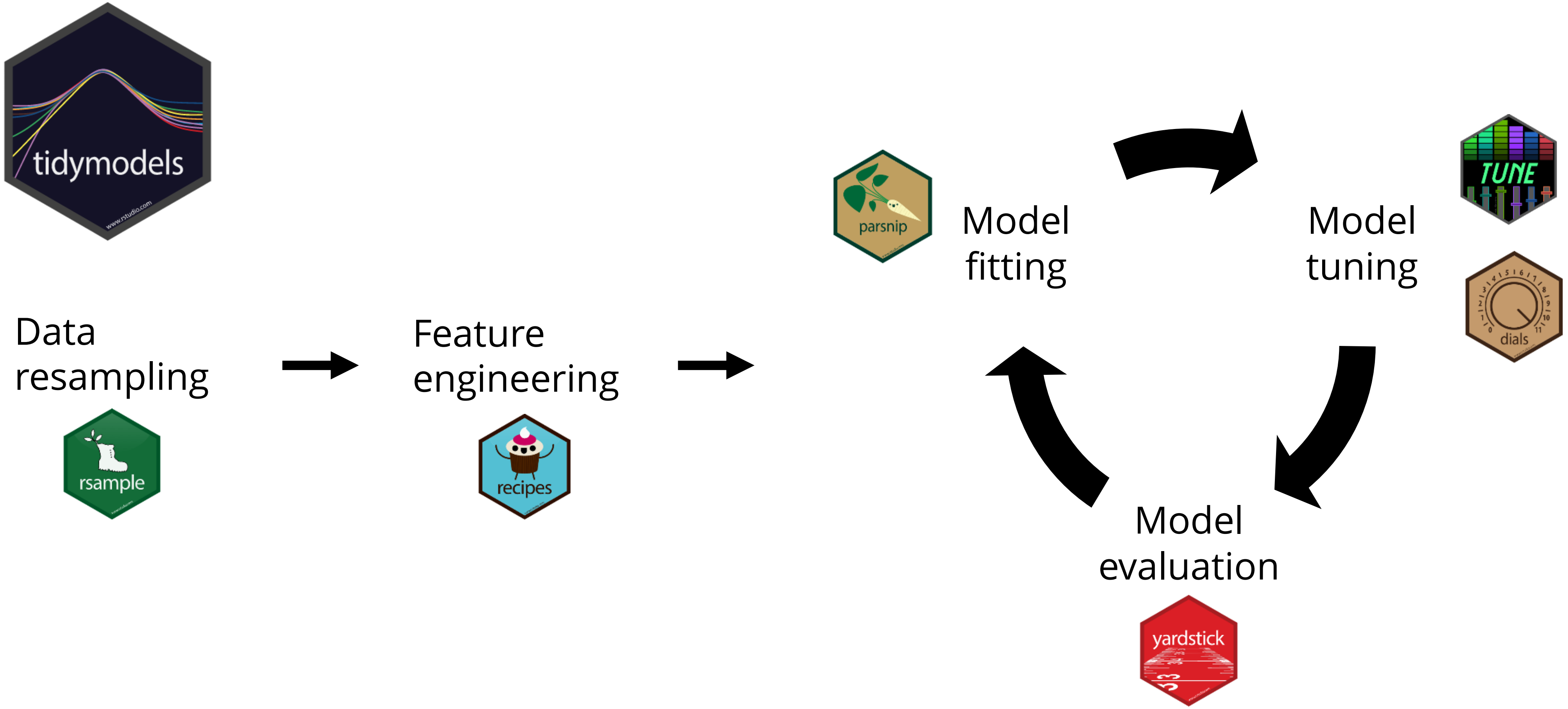
Country	Year	Count
Afghanistan	1999	745
Brazil	1999	37,737
China	1999	212,258
Afghanistan	2000	2,666
Brazil	2000	80,488
China	2000	213,766

School of Business

Course Goals

Machine Learning Framework in R - *tidymodels*

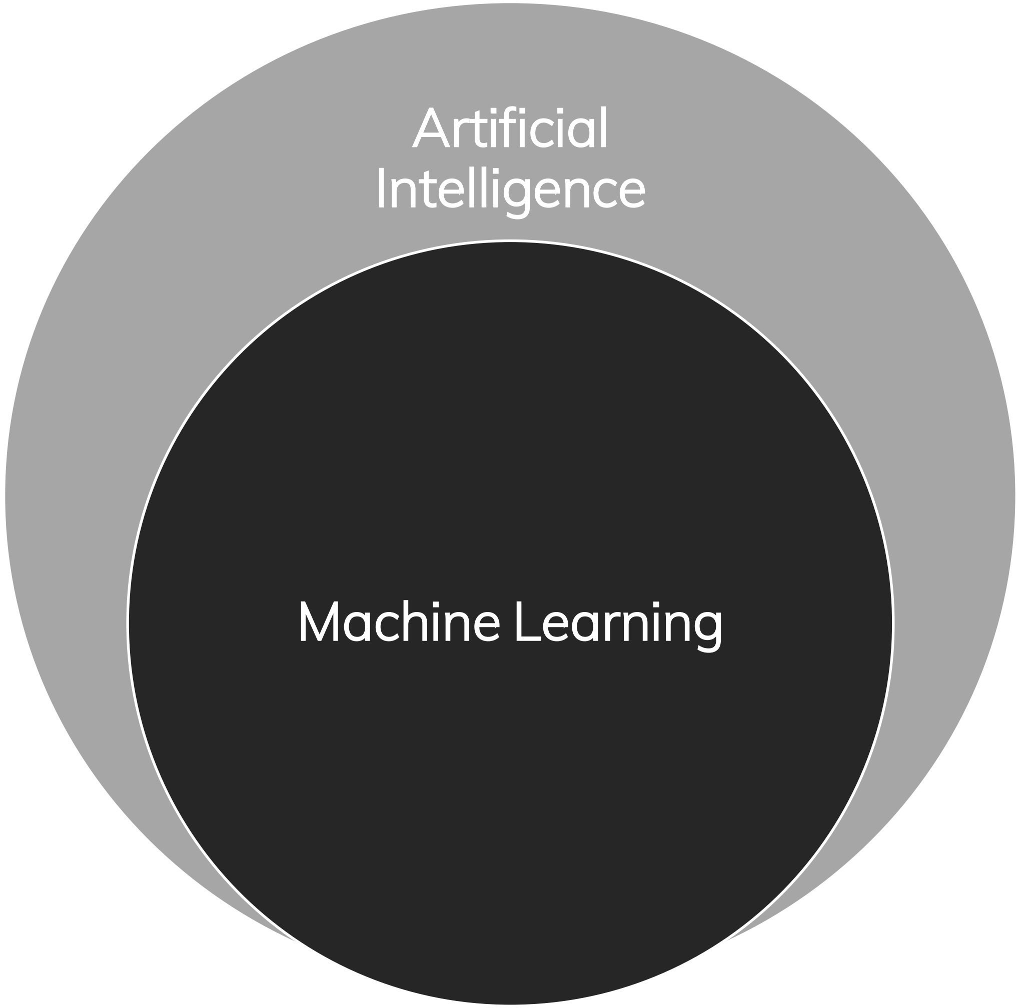
Tidymodels



School of Business

Machine Learning
What is Machine Learning?

A subset of *Artificial Intelligence* that gives computers the capability to learn without being *explicitly programmed*



Machine Learning
A New Programming Paradigm

Before ML

Computers were *explicitly programmed* to achieve desired results

Explicit Program

If input number is even → return "Yes"

If input number is odd → return "No"

Program Execution

Input		Output
6 →	*Program Logic Executed* →	"Yes"

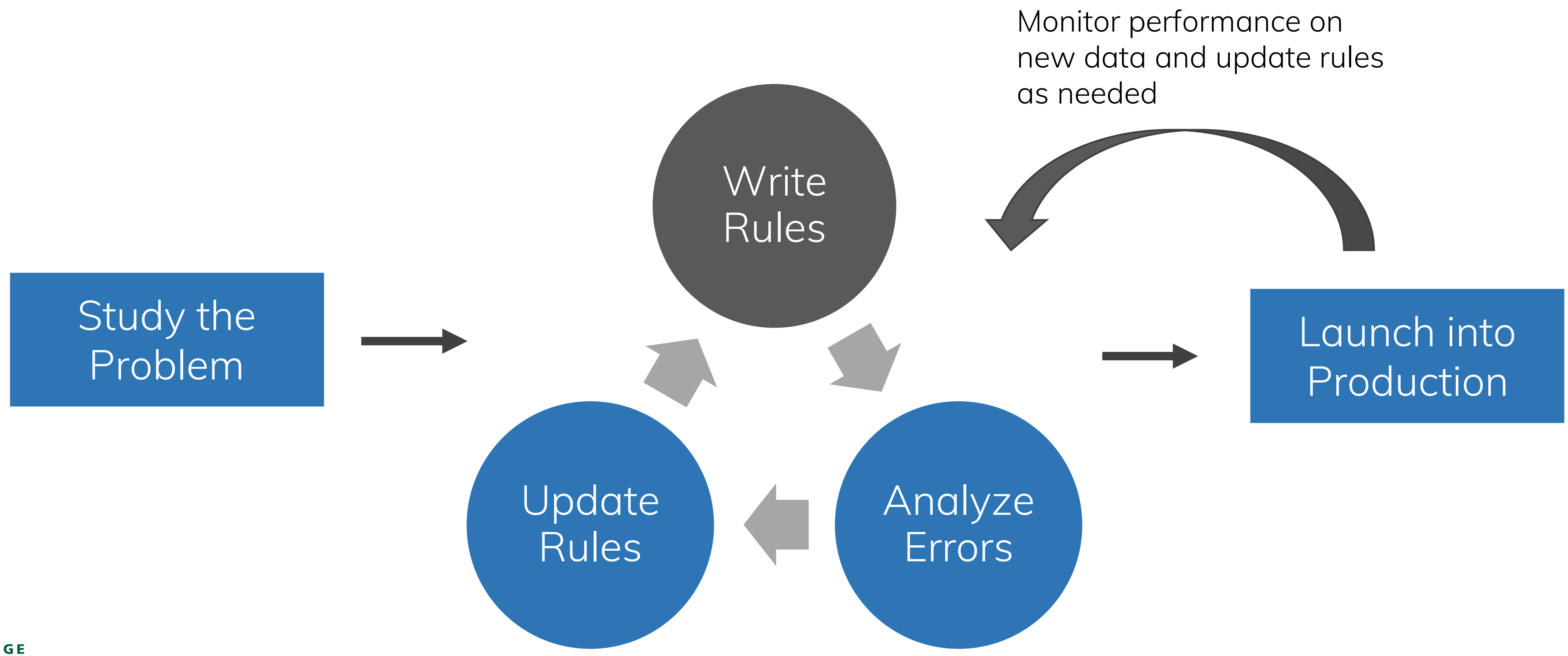
Benefit

Correct output on every execution

Challenge

All rules to accomplish task must be *known in advance*

Machine Learning
Explicit Programming Workflow



Machine Learning
Learning From Data

Today
ML algorithms use vast amounts of data to discover patterns and relationships without relying on a *predetermined* equations or set of rules as a model

ML Program

Label	Data Value
Yes	2
Yes	12
No	3
Yes	4
No	5
No	39
…	…

→ *Learned* prediction function

ML Prediction Function Execution

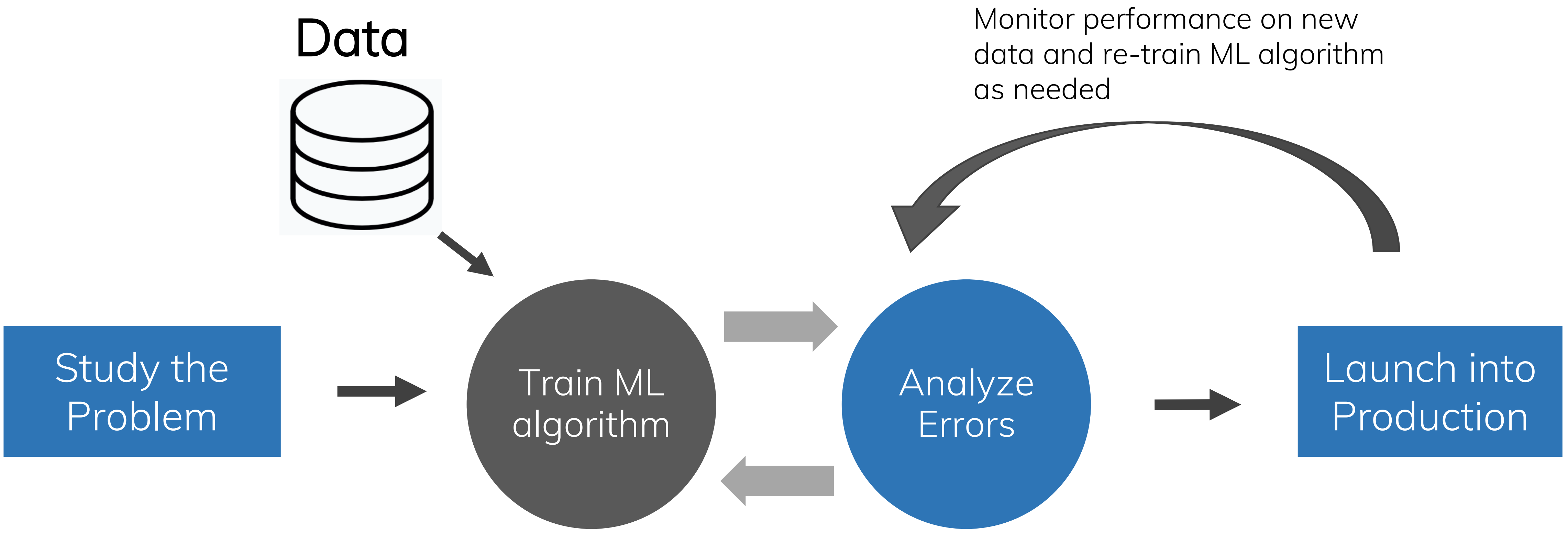
Input				Output
6	→	prediction function	→	"Yes"
4	→	prediction function	→	"No"
3	→	prediction function	→	"No"

Benefit
All steps/rules to accomplish ***do not*** have to be known or programmed explicitly

Challenge
Prediction error

Machine Learning

Machine Learning Workflow



School of Business

Machine Learning
Example - Image Recognition

Task

Identify handwritten digits

For a Human

Easy

For a Computer

Extremely difficult

MNIST Database of Handwritten Digits

Machine Learning
Without ML – Explicit Program

Explicit Program to Identify Digits

Imagine having to develop explicit instructions for a program to correctly identify handwritten digits

- You must identify ***every possible variation*** of how digits appear and instruct a computer to label them correctly
- Practically impossible – your program would be millions of lines long!

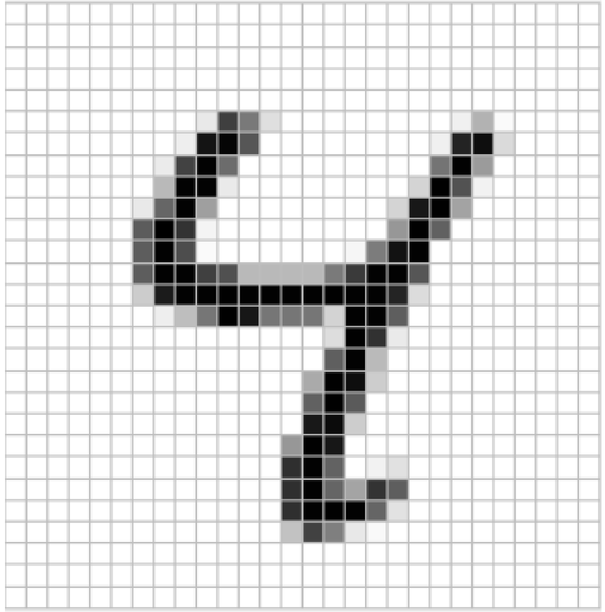
MNIST Database of Handwritten Digits

Machine Learning

A Machine Learning Approach

Encode Color Intensities and Apply ML Algorithms to Learn Patterns

28 x 28 image grid



Color intensities (0 – 255)

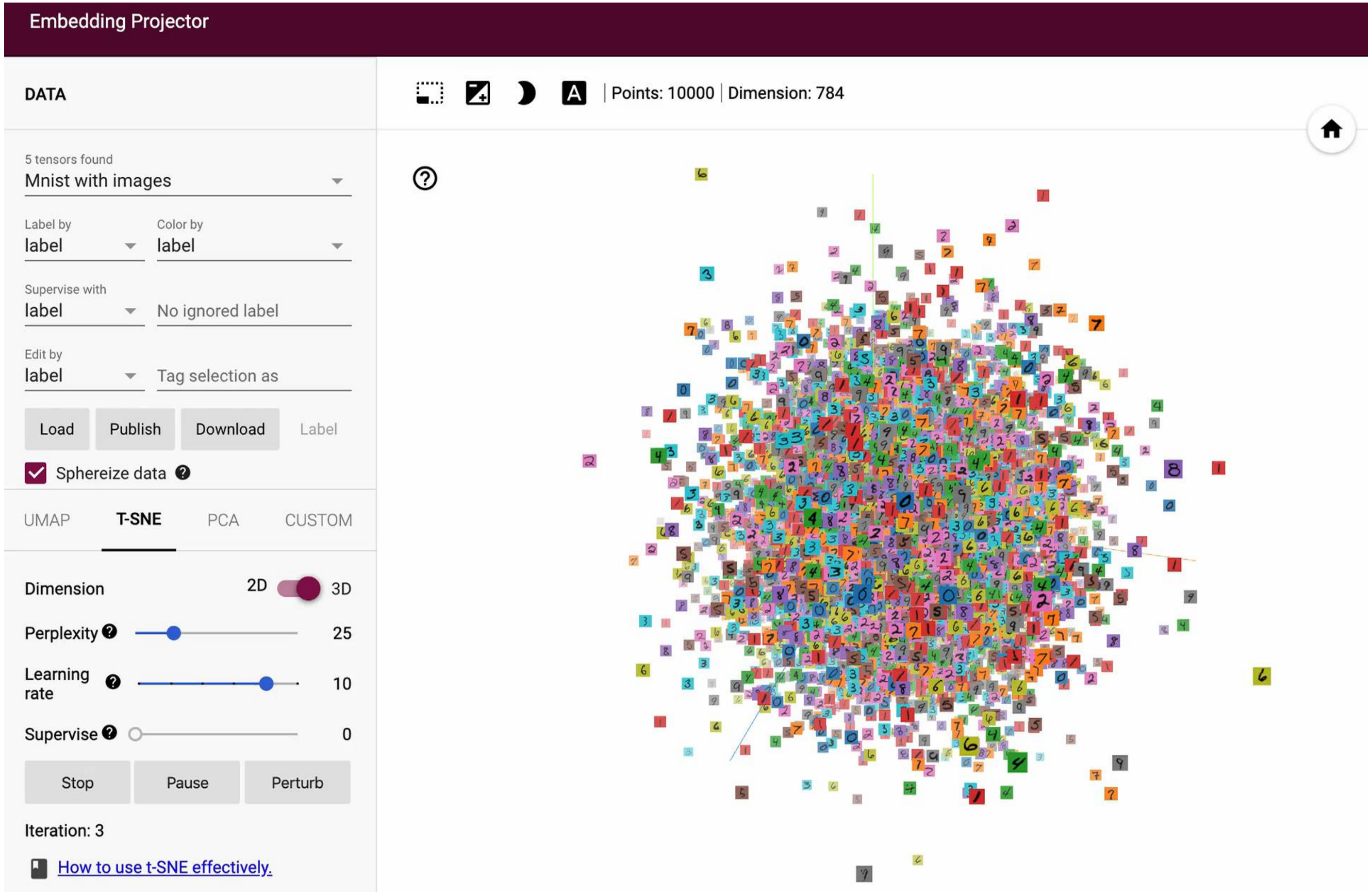
Number	Region_1	...	Region_467	Region_468	...	Region_783	Region_784
4	0		158	242		0	0
5	85		0	63		16	66
1	32		0	92		0	93
9	10		95	0		55	73
3	0		60	25		92	139
...	...	...	...	...	...	...	...

Machine Learning
Demonstration of ML Algorithm

TensorFlow projection tool
https://projector.tensorflow.org/

- *Goal* – find the optimal way to compress digit image data to 3 dimensions so that the same digits are grouped together

- Once this model is discovered, we can use it to predict new images based on where they fall in this 3-dimensional space



Machine Learning Methods
Supervised Learning

Supervised learning algorithms learn prediction functions from ***labeled training data***.

Labeled data set from a hospital

- Each row represents a patient who eventually did or did not develop heart disease (*the outcome variable* – ***Heart Disease***)
- Our goal might be to predict whether a new patient will develop heart disease using the predictor variables
  - For each set of predictor values, **we have a known outcome**
  - We also have a set of predictor values for each known outcome

Outcome (Target, Response, Dependent) **variable**

Heart Disease	Age	Chest Pain	Resting BP	Cholesterol
No	63	typical	145	233
Yes	67	asymptomatic	160	286
Yes	67	asymptomatic	120	229
No	37	nonanginal	130	250
No	41	nontypical	130	204

Predictor (Feature, Independent) **variables**

GEORGE MASON UNIVERSITY
School of Business

Machine Learning Methods

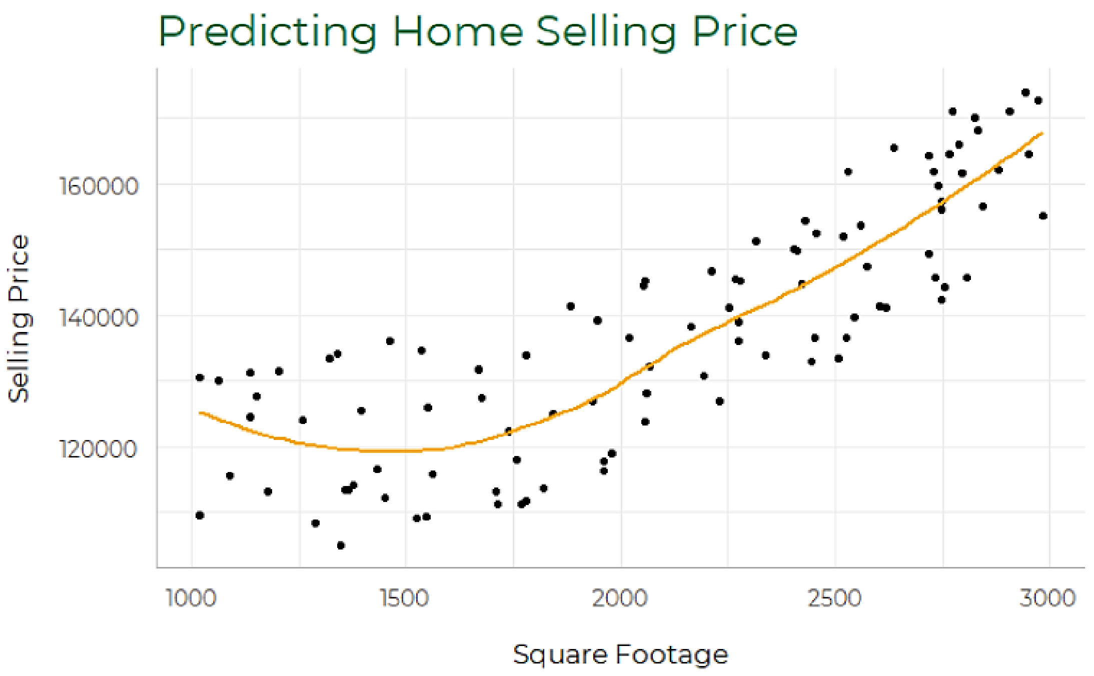
Supervised Learning - Regression

Regression

- Supervised learning methods are used to predict ***quantitative*** outcome variables
- Example
  - Predict the selling price of homes using features such as square footage, age, location

Outcome | Predictor

Selling Price	Square Footage
$105,667	1,100
$118,659	1,490
$134,268	1,850
$165,000	2,300



Machine Learning Methods

Supervised Learning - Classification

Classification

Supervised learning methods used to predict ***categorical*** outcome variables

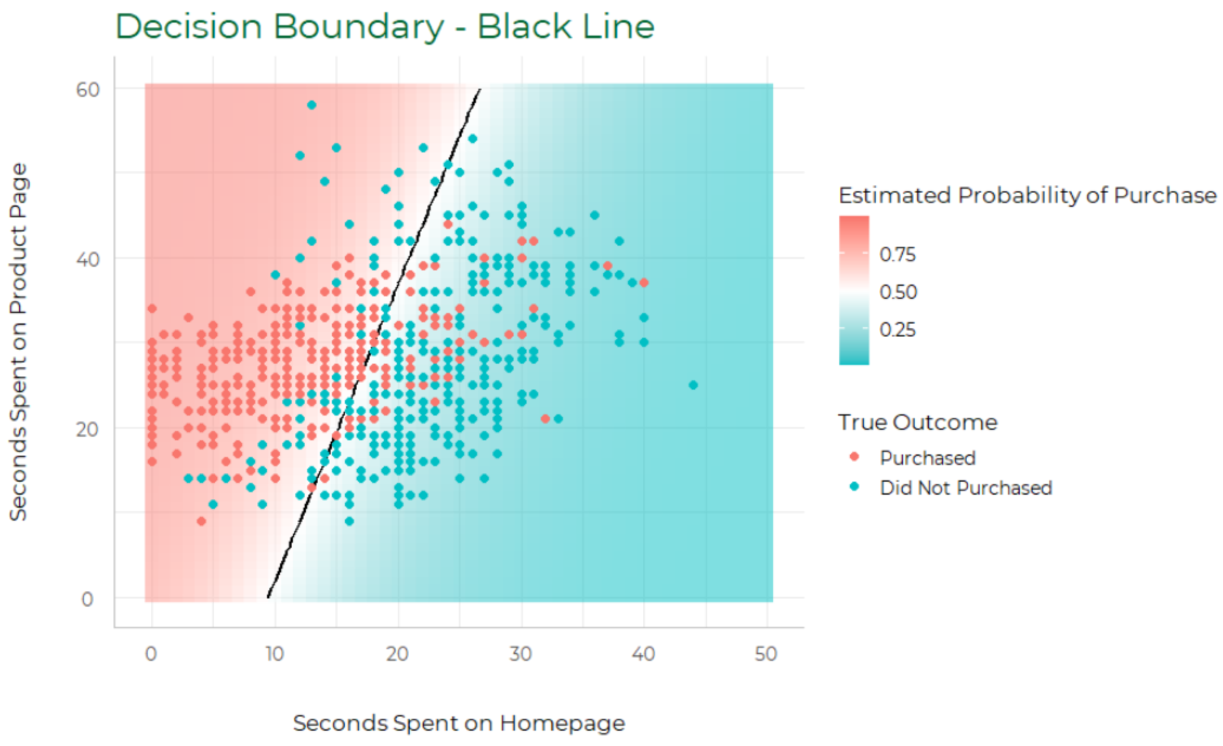
Segmenting the predictor values into distinct, non-overlapping regions to predict a category

Example

- Predict whether a customer will purchase a product based on the seconds they have spent browsing a company's homepage and product page

Outcome | Predictors

Purchase	Seconds Homepage	Seconds Product Page
Did Not Purchase	4	30
Purchased	32	43
Did Not Purchase	2	22
Purchased	24	36



Machine Learning Methods
Unsupervised Learning

In **unsupervised learning**, there are *feature* or *input* variables, but no labeled outcome variable

- no "**correct**" prediction

In this setting, it is typically of interest to learn the **structure** and **relationships** present in the unlabeled input data

- Methods include Clustering and Principal Components (PCA)

Marketing Example: Are there *customer segments* based on purchasing behavior?

Are there different types or species of plants present in the data below?

```
# A tibble: 150 x 4
   Sepal.Length Sepal.Width Petal.Length Petal.Width
          <dbl>       <dbl>        <dbl>       <dbl>
 1          5.1         3.5          1.4         0.2
 2          4.9         3            1.4         0.2
 3          4.7         3.2          1.3         0.2
 4          4.6         3.1          1.5         0.2
 5          5           3.6          1.4         0.2
 6          5.4         3.9          1.7         0.4
 7          4.6         3.4          1.4         0.3
 8          5           3.4          1.5         0.2
 9          4.4         2.9          1.4         0.2
10          4.9         3.1          1.5         0.1
# ... with 140 more rows
```

K-means Clustering

Finding observations that group together based on their proximity in the input data space