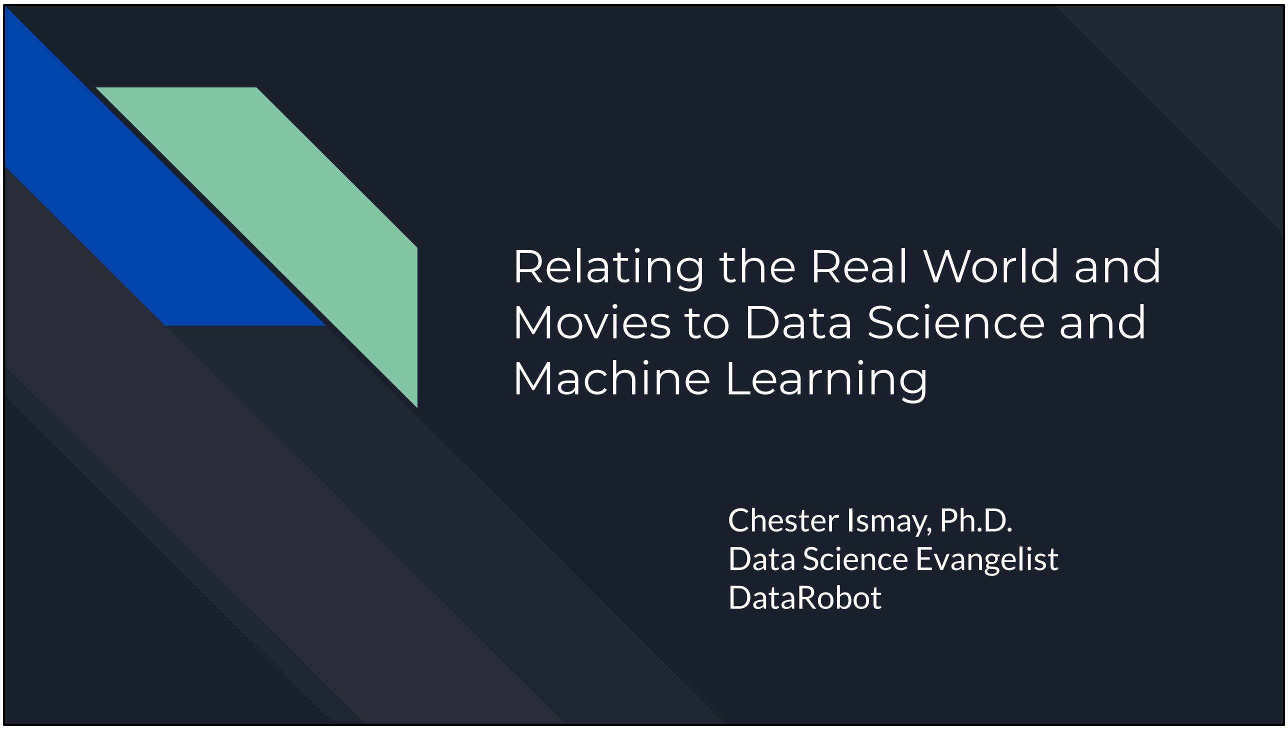
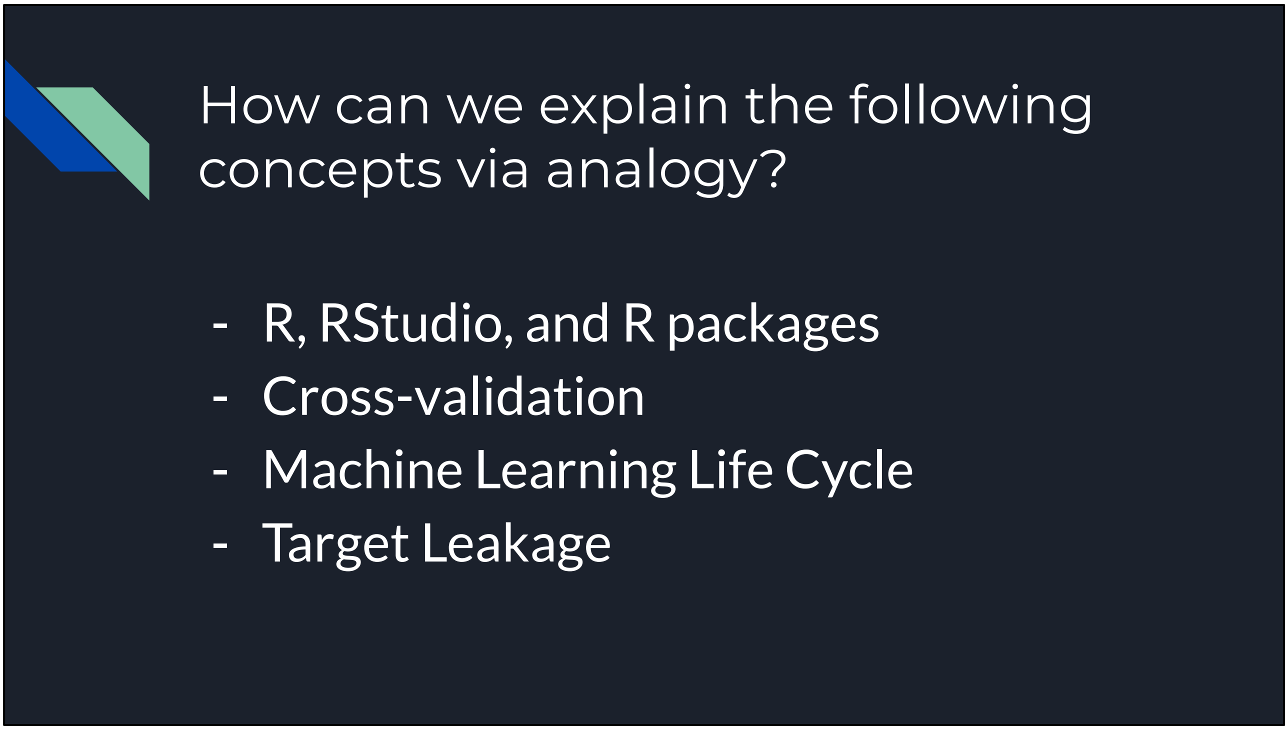


Relating the Real World and Movies to Data Science and Machine Learning

Chester Ismay, Ph.D.
Data Science Evangelist
DataRobot

Data science and machine learning often are difficult to teach and learn about because of the mathematics needed and the lack of great analogies to help novices build connections.



How can we explain the following concepts via analogy?

- R, RStudio, and R packages
- Cross-validation
- Machine Learning Life Cycle
- Target Leakage

In this talk, we'll discuss four different concepts from data science and analogies that can be used to teach others about them.

Back to the Future Part II

One such analogy will be based on relating to a movie that debuted 30 years ago in November of 1989. Who here has seen Back to the Future II?

Most of the film is based in “the future” of 2015, but the characters also head to 1955 and 1985. Let’s walk through one major plot line of the movie.

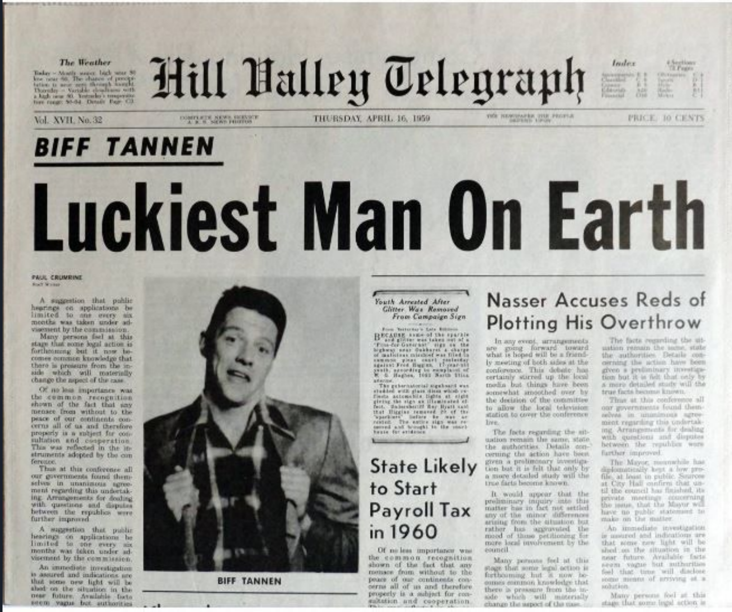
Near the beginning of the movie, the scientist Emmett Brown (also known as Doc) returns from 2015 to present-day 1985 to tell Marty that something bad has happened to Marty’s soon in 2015 in the future. Doc hops in his Delorean time machine and takes Marty and his girlfriend (and future wife) Jennifer with him to 2015.

Doc and Marty are successful in helping his son escape trouble in 2015, but as Marty is roaming around the city in 2015 he stumbles upon a shop window of an antique store with something interesting.

He purchases a sports almanac from 1950-2000 with the plans to make a few bets to make some extra money when he heads back to 1985. Just as Marty and Doc are getting ready to head back to 1985, the almanac falls from Marty’s hands. When Doc sees the almanac he realizes what Marty is planning. Doc says he didn’t invent the time machine for financial gain and throws the almanac in the garbage.

Biff Tannen who is the main antagonist in Back to the Future is now an old man and overhears Doc and Marty talking about time travel and the almanac. Marty (for some ridiculous reason) leaves the Delorean unlocked. Biff steals the time machine and heads back to 1955.

He then tracks down himself as a teenager in 1955 and tells him that all he has to do is bet on the team listed and he'll make a fortune.



The Weather

Today – Mostly sunny, high near 80 low near 60. The chance of precipitation is near zero through tonight. Thursday – Variable cloudiness with a high near 80. Yesterday's temperature range: 80-84. Details Page C3.

Hill Valley Telegraph

Index

Vol. XVII, No. 32 — COMPLETE NEWS SERVICE A. B. S. NEWS PHOTOS — THURSDAY, APRIL 16, 1959 — THE NEWSPAPER THE PEOPLE DEPEND UPON — PRICE 10 CENTS

BIFF TANNEN

Luckiest Man On Earth

PAUL CRUMRINE
Staff Writer

A suggestion that public hearings on applications be limited to one every six months was taken under advisement by the commission.

Many persons feel at this stage that some legal action is forthcoming but it now becomes common knowledge that there is pressure from the inside which will materially change the aspect of the case.

Of no less importance was the common recognition shown of the fact that any menace from without to the peace of our continents concerns all of us and therefore properly is a subject for consultation and cooperation. This was reflected in the instruments adopted by the conference.

Thus at this conference all our governments found themselves in unanimous agreement regarding this undertaking. Arrangements for dealing with questions and disputes between the republics were further improved.

A suggestion that public hearings on applications be limited to one every six months was taken under advisement by the commission.

An immediate investigation is assured and indications are that some new light will be shed on the situation in the near future. Available facts seem vague but authorities

BIFF TANNEN

Youth Arrested After Glitter Was Removed From Campaign Sign

From Yesterday's Late Editions

BECAUSE some of the sparkle and glitter was taken out of a "Pine-for-Governor" sign on the highway near Oakhurst a charge of malicious mischief was filed in common pleas court yesterday against Fred Higgins, 17-year-old youth, according to complaint of W. S. Hughes, 1682 North Elisa avenue.

The gubernatorial signboard was studded with glass discs which reflects automobile lights at night giving the sign an illuminated effect. Subsheriff Ray Hyatt said that Higgins removed 20 of the "sparklers" before he was arrested. The entire sign was removed and brought to the court house for evidence.

State Likely to Start Payroll Tax in 1960

Of no less importance was the common recognition shown of the fact that any menace from without to the peace of our continents concerns all of us and therefore properly is a subject for consultation and cooperation.

Nasser Accuses Reds of Plotting His Overthrow

In any event, arrangements are going forward toward what is hoped will be a friendly meeting of both sides at the conference. This debate has certainly stirred up the local media but things have been somewhat smoothed over by the decision of the committee to allow the local television station to cover the conference live.

The facts regarding the situation remain the same, state the authorities. Details concerning the action have been given a preliminary investigation but it is felt that only by a more detailed study will the true facts become known.

It would appear that the preliminary inquiry into this matter has in fact not settled any of the minor differences arising from the situation but rather has aggravated the mood of those petitioning for more local involvement by the council.

Many persons feel at this stage that some legal action is forthcoming but it now becomes common knowledge that there is pressure from the inside which will materially change the aspect of the case.

The facts regarding the situation remain the same, state the authorities. Details concerning the action have been given a preliminary investigation but it is felt that only by a more detailed study will the true facts become known.

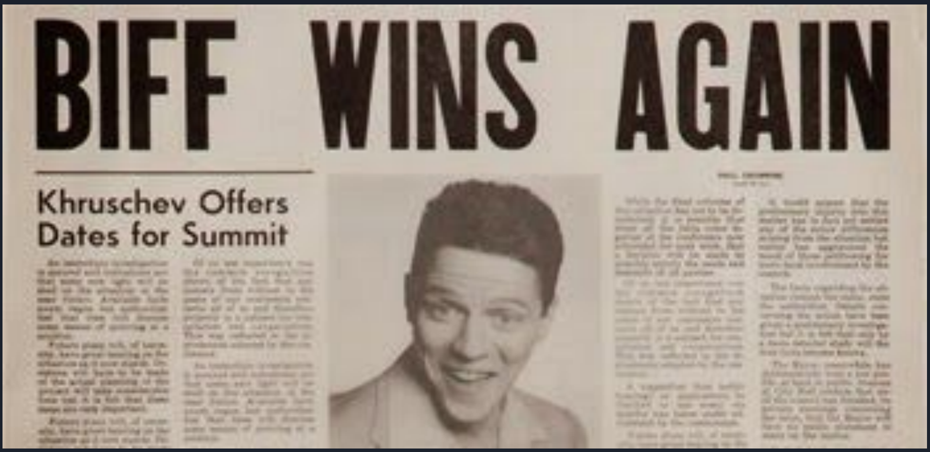
Thus at this conference all our governments found themselves in unanimous agreement regarding this undertaking. Arrangements for dealing with questions and disputes between the republics were further improved.

The Mayor, meanwhile has diplomatically kept a low profile, at least in public. Sources at City Hall confirm that until the council has finished its private meetings concerning the issue, that the Mayor will have no public statement to make on the matter.

An immediate investigation is assured and indications are that some new light will be shed on the situation in the near future. Available facts seem vague but authorities feel that time will disclose some means of arriving at a solution.

Many persons feel at this stage that some legal action is

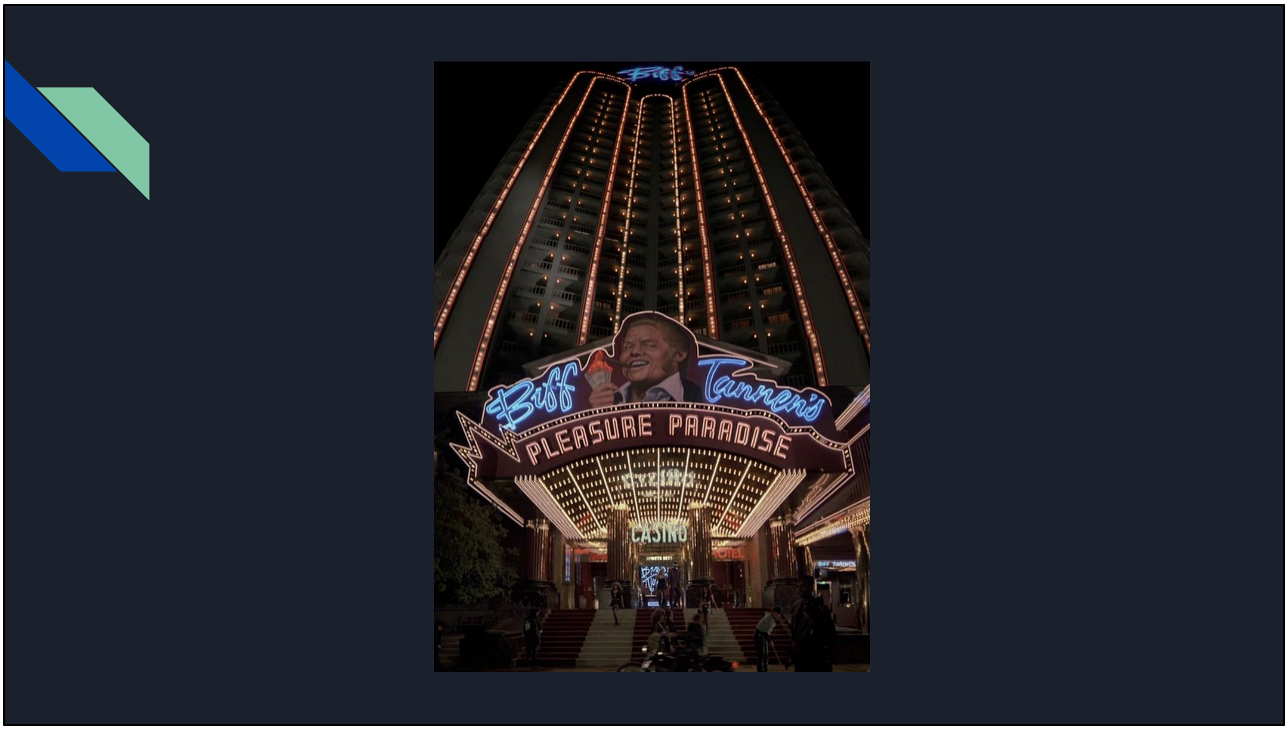
At the age of 21, Biff Tannen becomes a millionaire by betting on a horse race.



BIFF WINS AGAIN

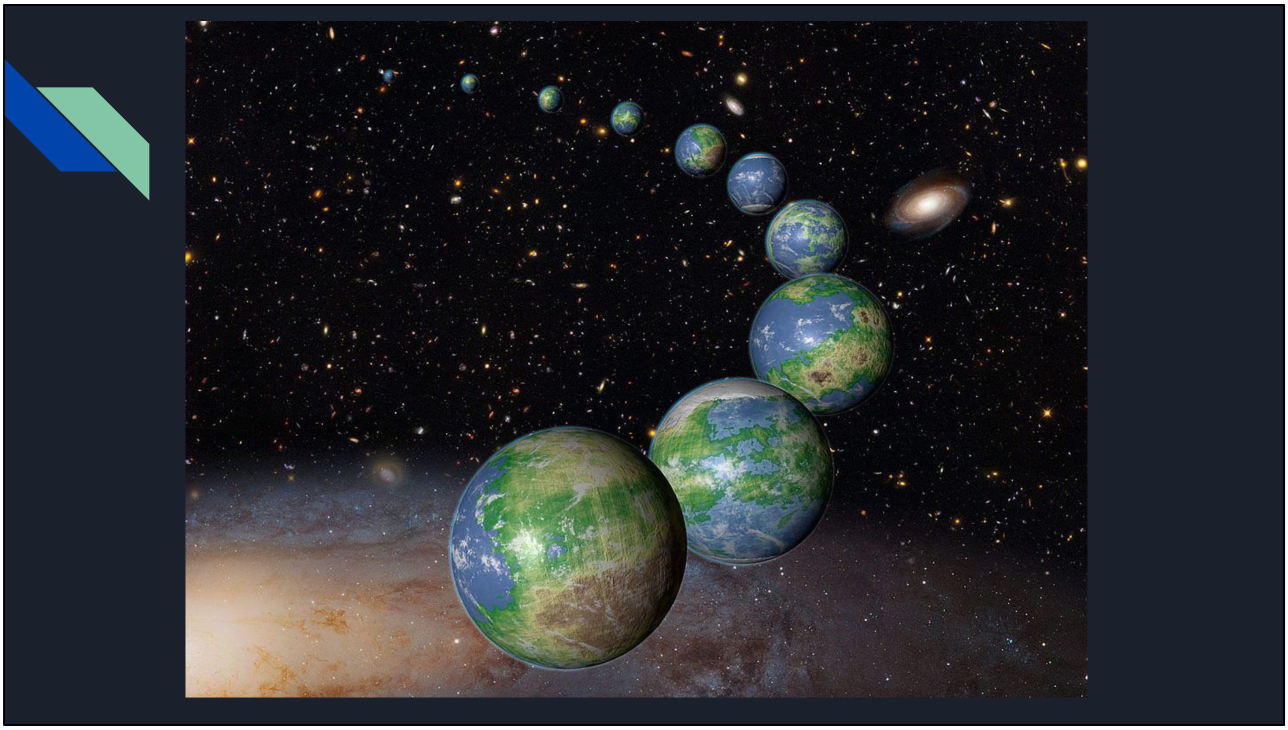
Khruschev Offers Dates for Summit

He continues to win and win and win over again with sports bets.

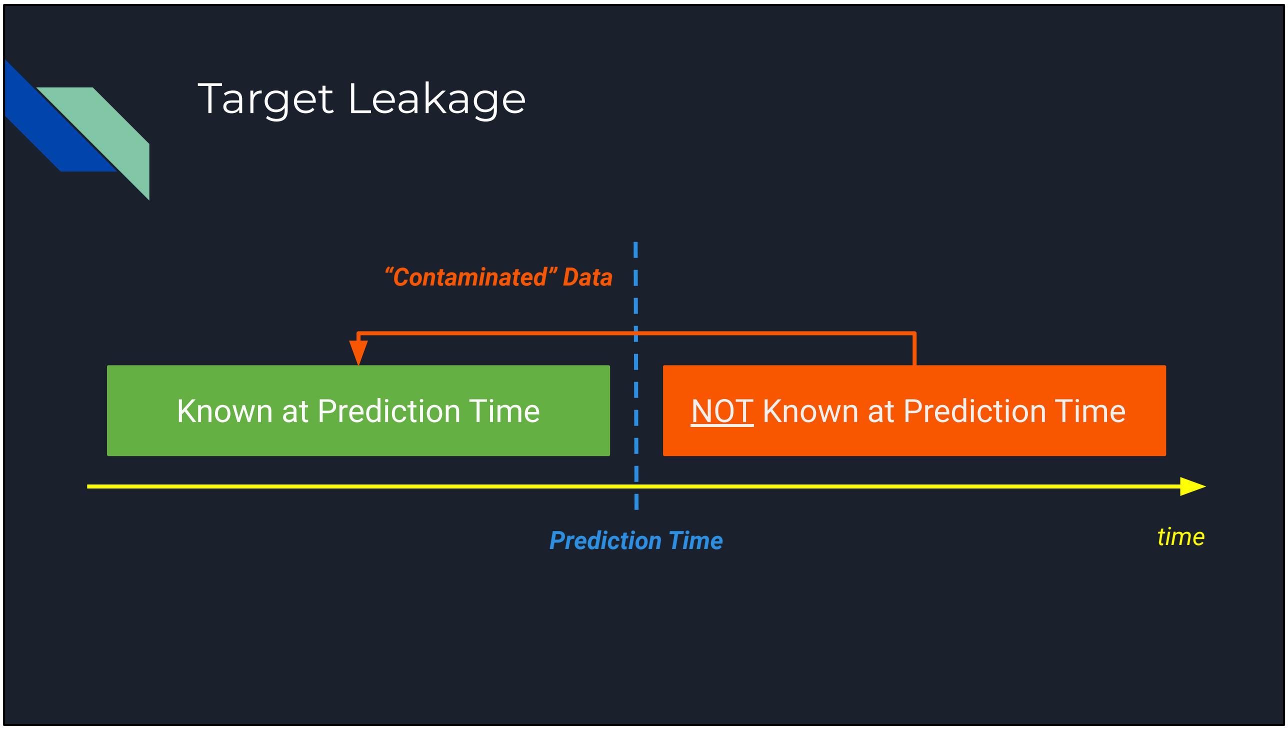


Eventually he takes over the clock tower in the city and turns it into a high rise casino and hotel.

Marty and Doc make their way to 1985 and Marty confronts Biff about the almanac.
Biff has created a world full of debauchery and corruption that they are trying to stop.



The movie talks about different parallel universes that are created by this event in time.



Target Leakage

"Contaminated" Data

Known at Prediction Time

NOT Known at Prediction Time

Prediction Time

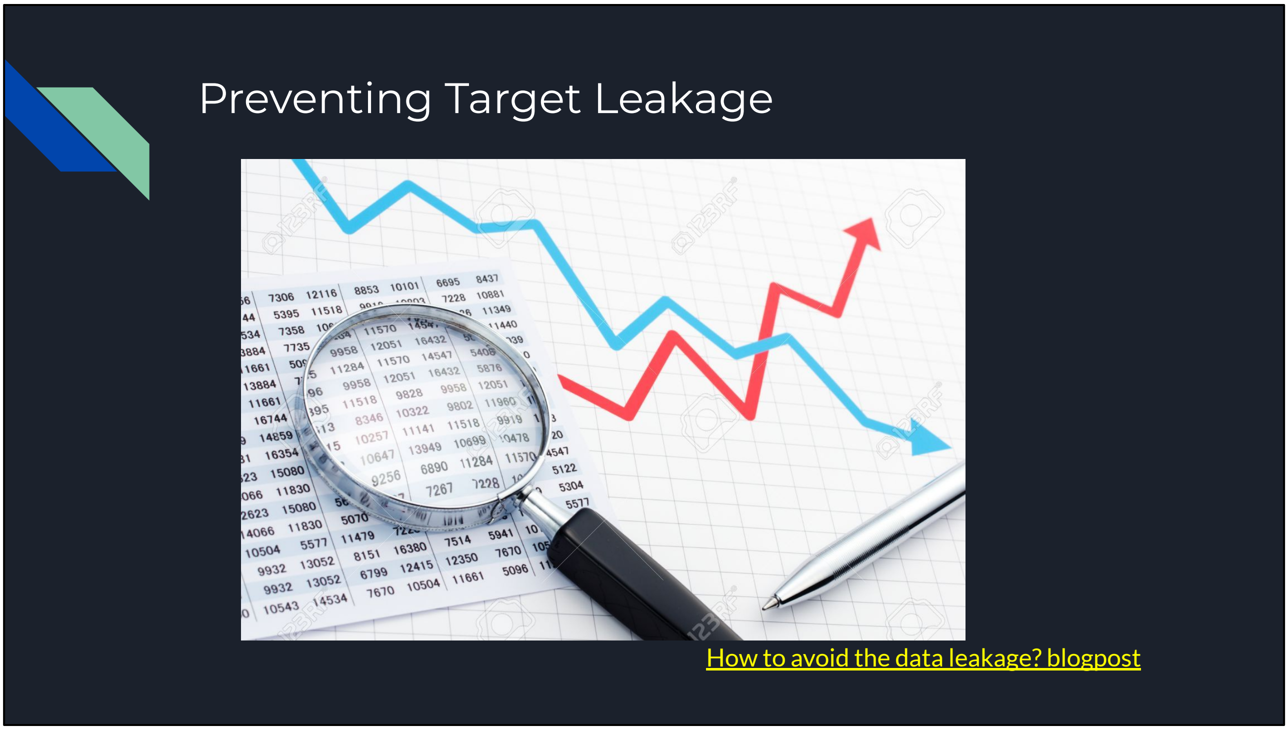
time

This relates nicely to the concept of target leakage in machine learning. This relates to using information that you won't have at prediction time to build a high performing model. Unfortunately, when that model is used to make predictions on new data it performs poorly.

This is similar to cheating to try to improve performance but without actually doing the needed learning with variables you will have at prediction time.

Predicting Recidivism

One of my colleagues at DataRobot was working with college students and faculty on a project to better predict if those that were incarcerated for being convicted of committing a crime would return or not to jail. This is called recidivism. My colleague was somewhat surprised that the model was fitting the data nearly perfectly. Upon further inspection of the data, he noticed that a variable called “return date” was included in the data for each former inmate. Thus, if there was a value in that column it corresponded to the inmate returning to jail at a later date. That data was stored after the point where a prediction would be made on a different new data set. Thus, return date was an example of target leakage for predicting recidivism.

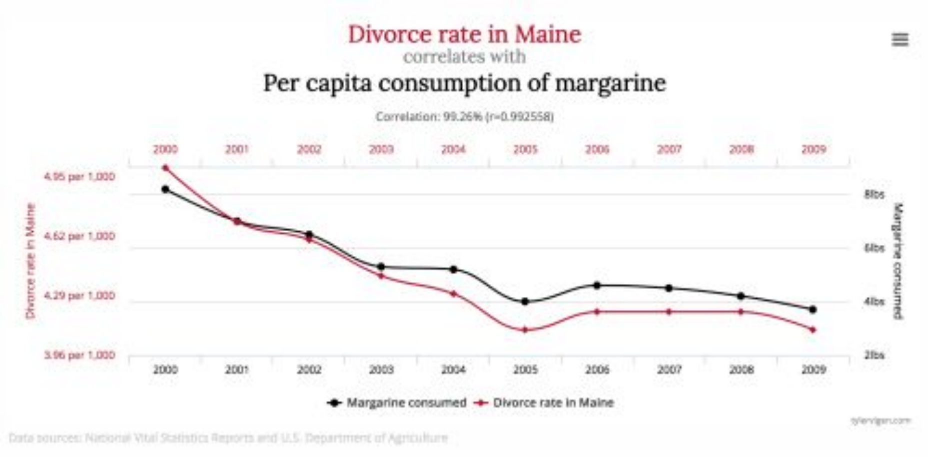


Preventing Target Leakage

How to avoid the data leakage? blogpost

One of the best ways to detect target leakage is by understanding the different variables or features in your data. When were they collected? What do they represent? Will they be available at the point of making predictions on new data?

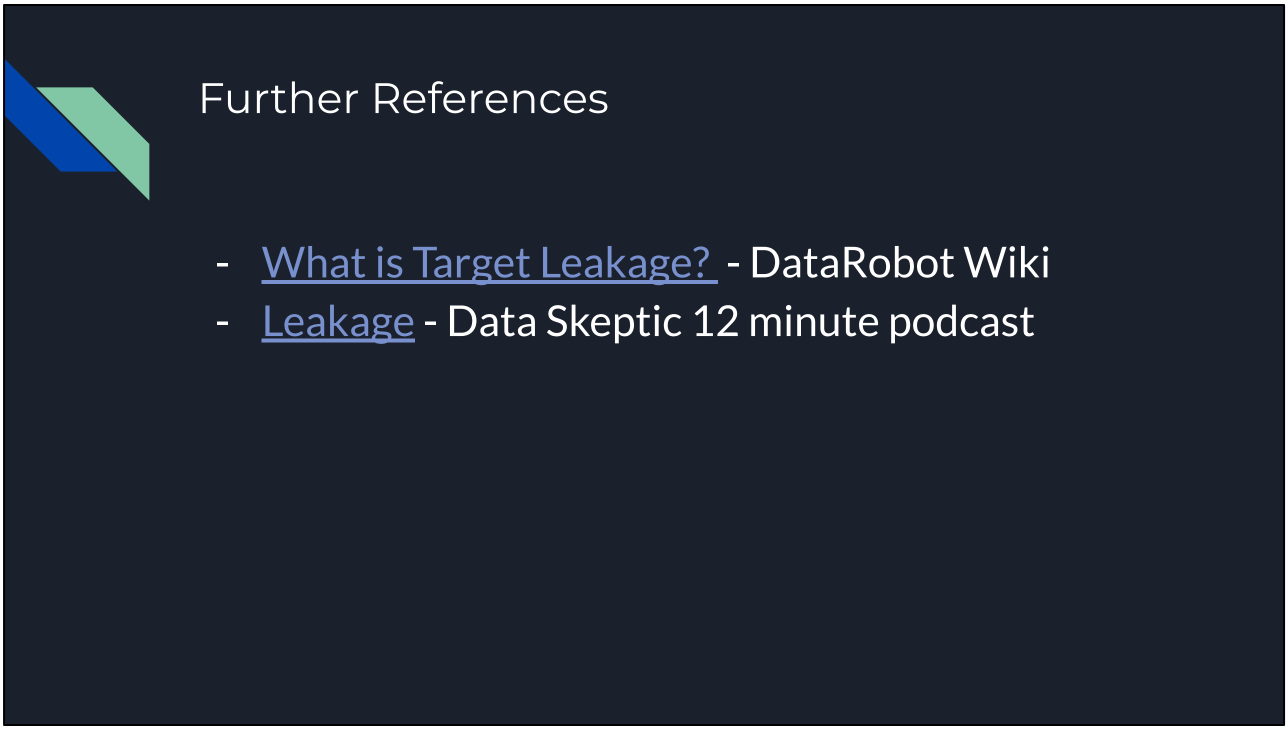
Preventing Target Leakage



Data sources: National Vital Statistics Reports and U.S. Department of Agriculture

Beware of Data Leakage blogpost

Another way is to check for features that are highly correlated with what you are trying to predict in the target feature. It might not be the case that highly correlated variables exhibit target leakage but they are a nice proxy to check for.



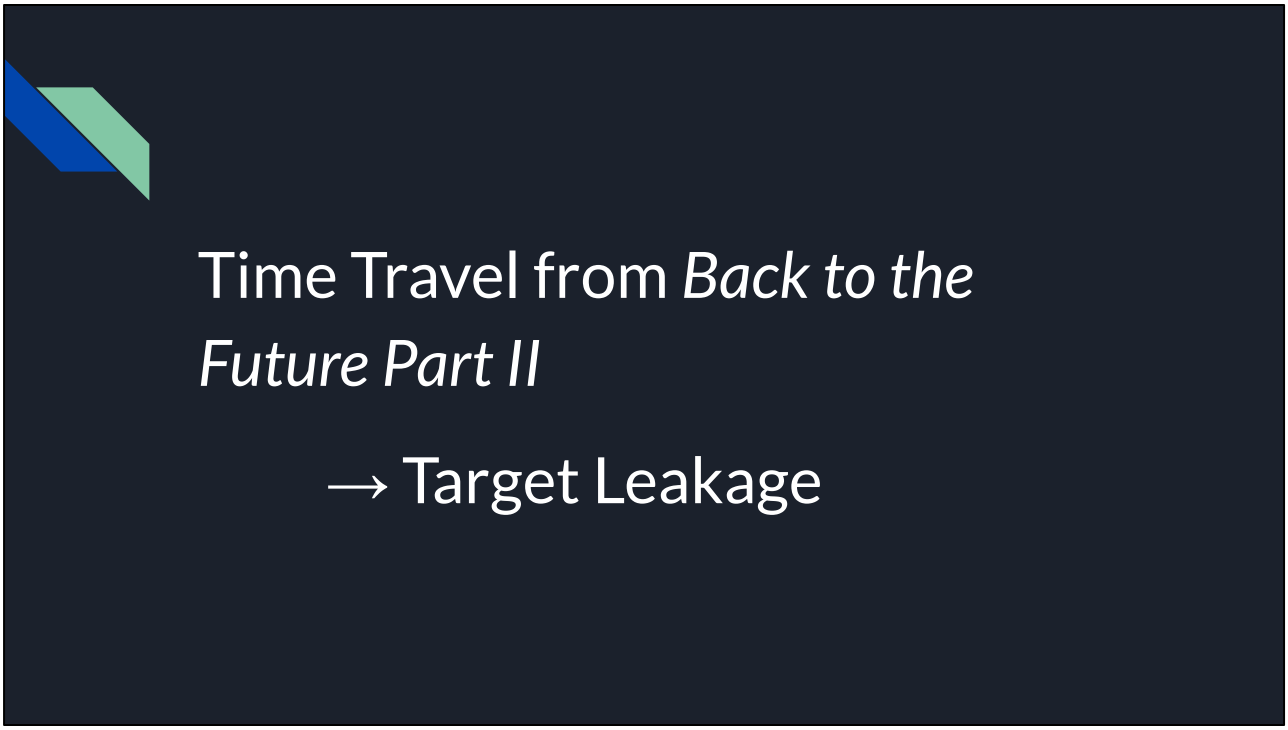
Further References

- What is Target Leakage? - DataRobot Wiki
- Leakage - Data Skeptic 12 minute podcast

Here are some other examples. I’ll share a link to these slides at the end of my talk so you’ll have some videos, podcasts, and links to read for further information.

Prevent Target Leakage to Prevent Bad Parallel Universes

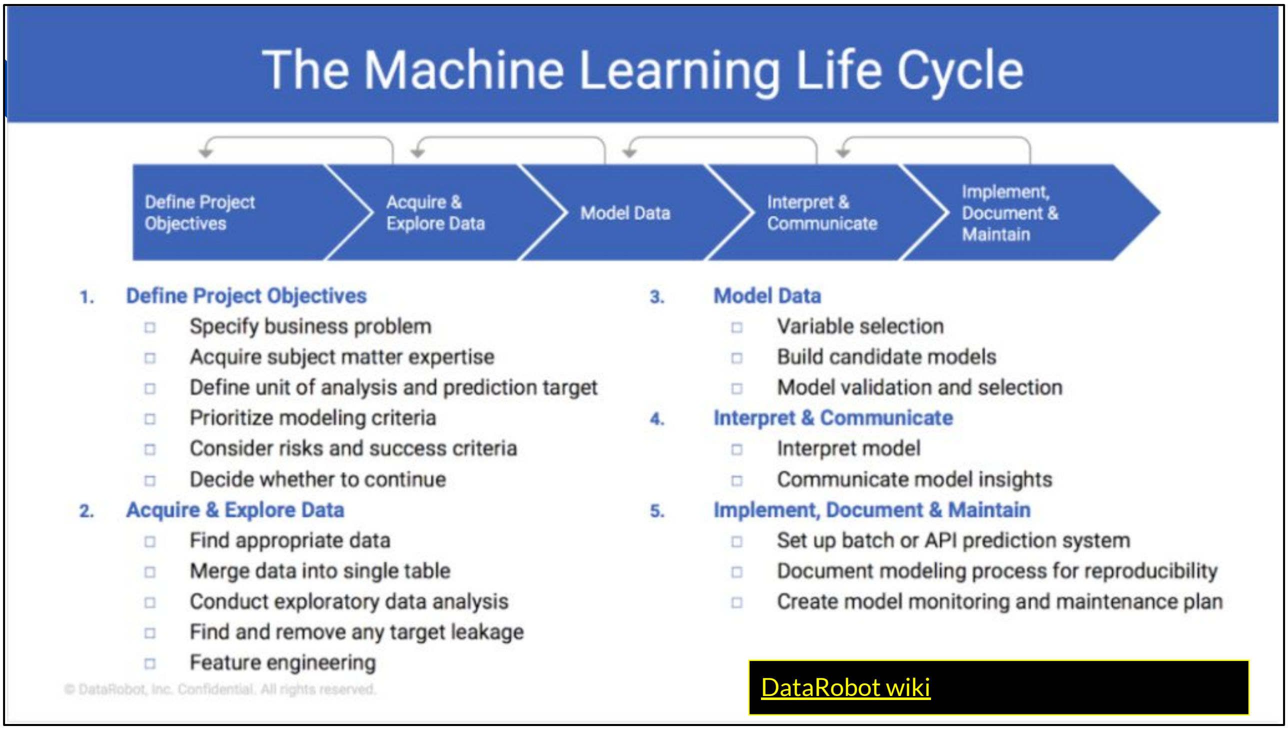
So the moral of the story for us here is that you should try to prevent target leakage so that you don’t create bad parallel universes where your model performed well in the training phase, but is performing terribly when it is actually in use on new data.



Time Travel from *Back to the Future Part II*

→ Target Leakage

A key thing to take away here is that target leakage relates to a major storyline in the second Back to the Future movie. Let’s now hop to another analogy.



The Machine Learning Life Cycle

Define Project Objectives

Acquire & Explore Data

Model Data

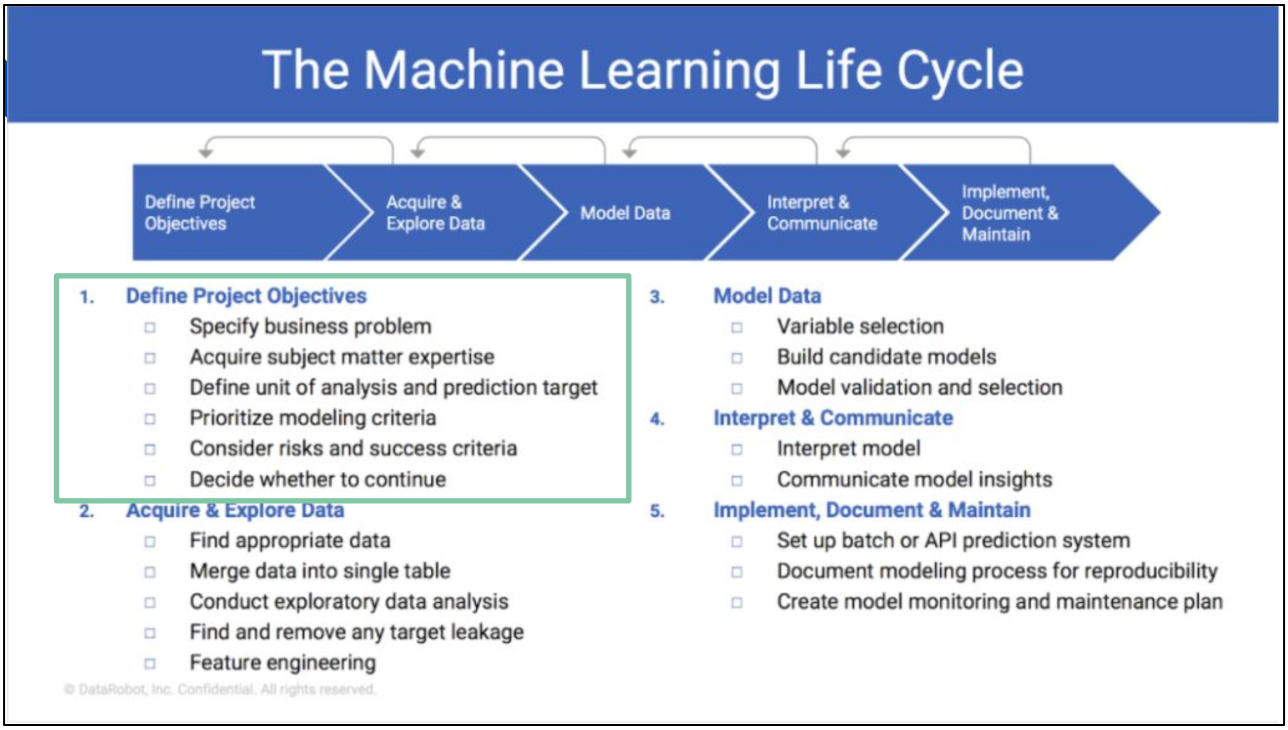
Interpret & Communicate

Implement, Document & Maintain

1. **Define Project Objectives**
   - □ Specify business problem
   - □ Acquire subject matter expertise
   - □ Define unit of analysis and prediction target
   - □ Prioritize modeling criteria
   - □ Consider risks and success criteria
   - □ Decide whether to continue
2. **Acquire & Explore Data**
   - □ Find appropriate data
   - □ Merge data into single table
   - □ Conduct exploratory data analysis
   - □ Find and remove any target leakage
   - □ Feature engineering
3. **Model Data**
   - □ Variable selection
   - □ Build candidate models
   - □ Model validation and selection
4. **Interpret & Communicate**
   - □ Interpret model
   - □ Communicate model insights
5. **Implement, Document & Maintain**
   - □ Set up batch or API prediction system
   - □ Document modeling process for reproducibility
   - □ Create model monitoring and maintenance plan

DataRobot wiki

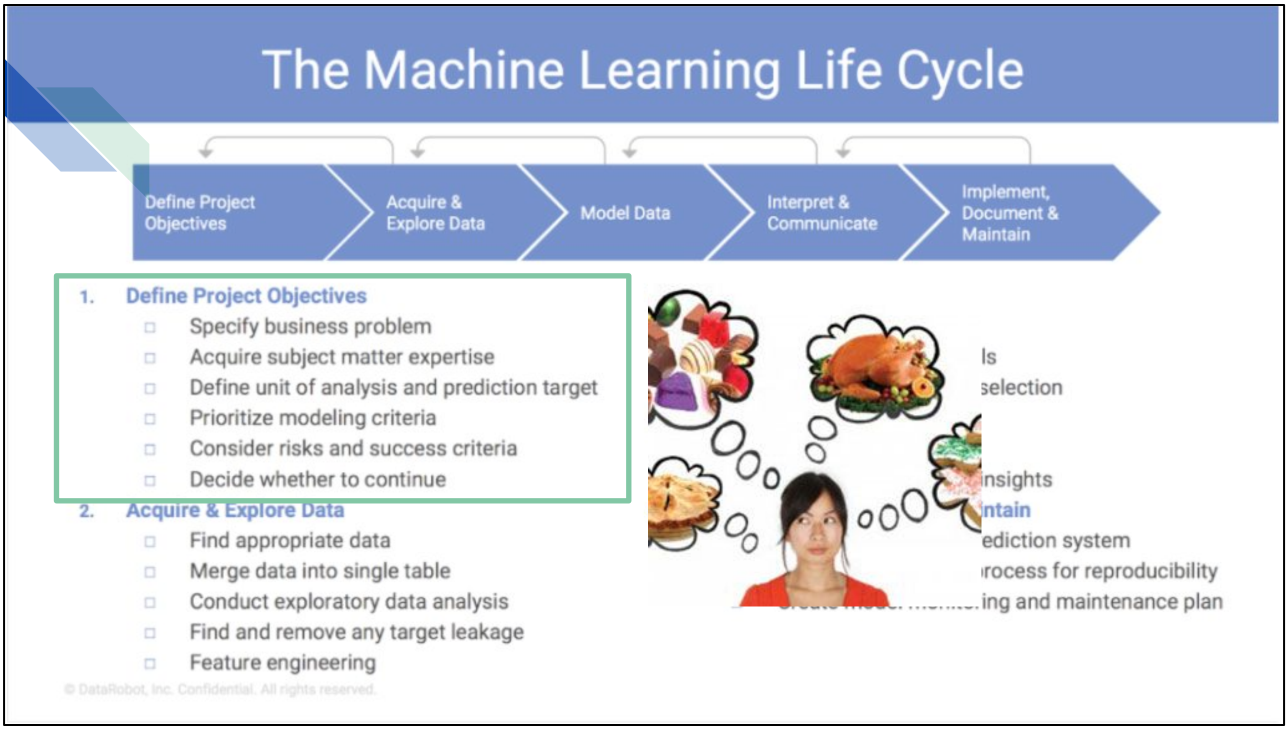
In machine learning there is a cycle of processes that frequently occur. This one in particular is from the DataRobot wiki but similar workflows have been shared elsewhere.



The Machine Learning Life Cycle

1. **Define Project Objectives**
   - □ Specify business problem
   - □ Acquire subject matter expertise
   - □ Define unit of analysis and prediction target
   - □ Prioritize modeling criteria
   - □ Consider risks and success criteria
   - □ Decide whether to continue
2. **Acquire & Explore Data**
   - □ Find appropriate data
   - □ Merge data into single table
   - □ Conduct exploratory data analysis
   - □ Find and remove any target leakage
   - □ Feature engineering
3. **Model Data**
   - □ Variable selection
   - □ Build candidate models
   - □ Model validation and selection
4. **Interpret & Communicate**
   - □ Interpret model
   - □ Communicate model insights
5. **Implement, Document & Maintain**
   - □ Set up batch or API prediction system
   - □ Document modeling process for reproducibility
   - □ Create model monitoring and maintenance plan

Let's first examine this first phase of the cycle. Here we set up the problem statement, make sure we have the resources to continue, and decide on how we should proceed.



The Machine Learning Life Cycle

1. **Define Project Objectives**
   - □ Specify business problem
   - □ Acquire subject matter expertise
   - □ Define unit of analysis and prediction target
   - □ Prioritize modeling criteria
   - □ Consider risks and success criteria
   - □ Decide whether to continue
2. **Acquire & Explore Data**
   - □ Find appropriate data
   - □ Merge data into single table
   - □ Conduct exploratory data analysis
   - □ Find and remove any target leakage
   - □ Feature engineering

ls

selection

insights

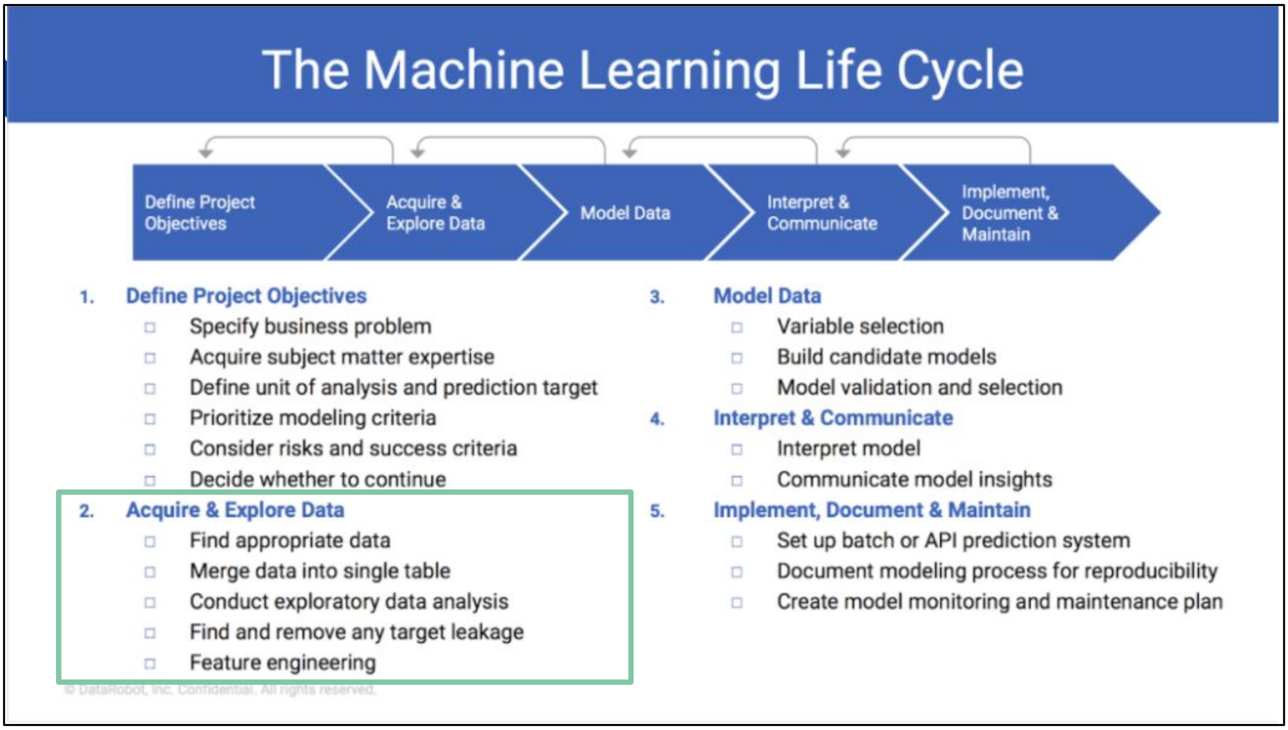
ntain

ediction system

rocess for reproducibility

ing and maintenance plan

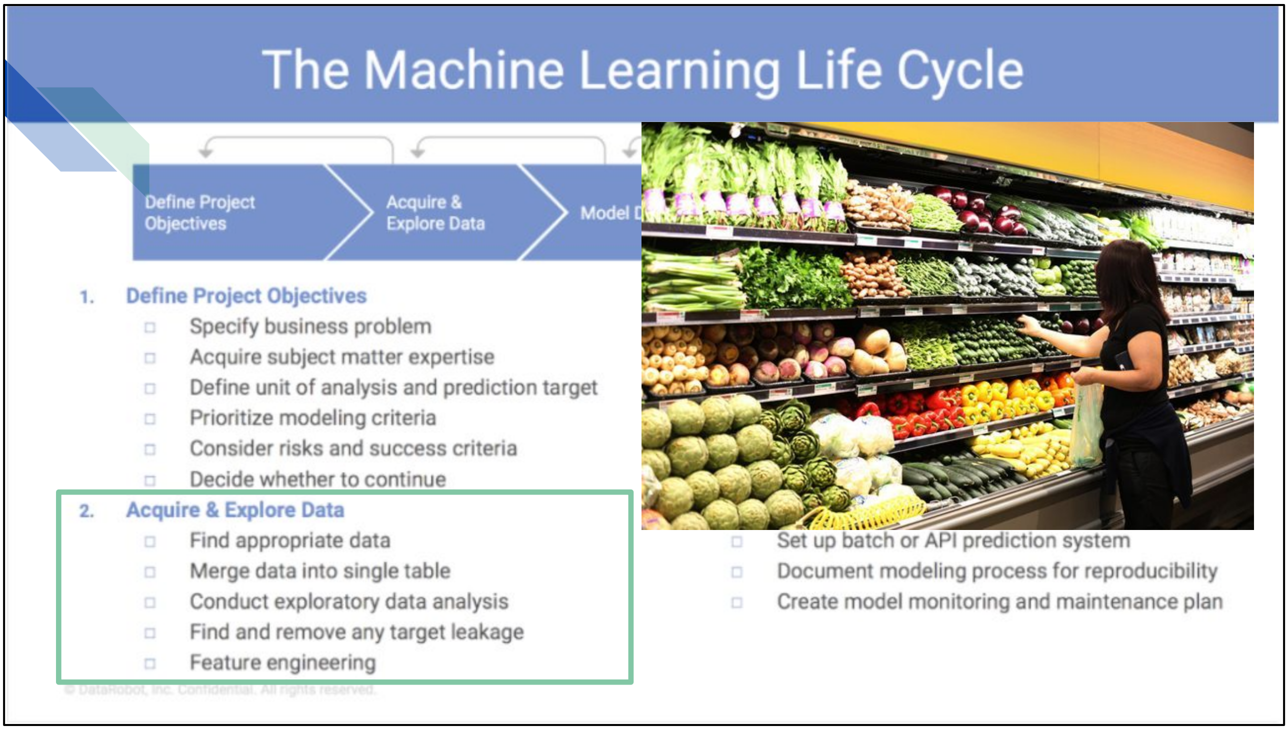
Hmm, to me this sounds a lot like trying to think what meal I should make tonight. How does that match up?



The Machine Learning Life Cycle

1. **Define Project Objectives**
   - □ Specify business problem
   - □ Acquire subject matter expertise
   - □ Define unit of analysis and prediction target
   - □ Prioritize modeling criteria
   - □ Consider risks and success criteria
   - □ Decide whether to continue
2. **Acquire & Explore Data**
   - □ Find appropriate data
   - □ Merge data into single table
   - □ Conduct exploratory data analysis
   - □ Find and remove any target leakage
   - □ Feature engineering
3. **Model Data**
   - □ Variable selection
   - □ Build candidate models
   - □ Model validation and selection
4. **Interpret & Communicate**
   - □ Interpret model
   - □ Communicate model insights
5. **Implement, Document & Maintain**
   - □ Set up batch or API prediction system
   - □ Document modeling process for reproducibility
   - □ Create model monitoring and maintenance plan

Now that we've decided to continue to make a meal. We need to acquire the needed data. We might need to combine some data together and clean the data further. How does this relate to cooking a meal?



The Machine Learning Life Cycle

Define Project Objectives

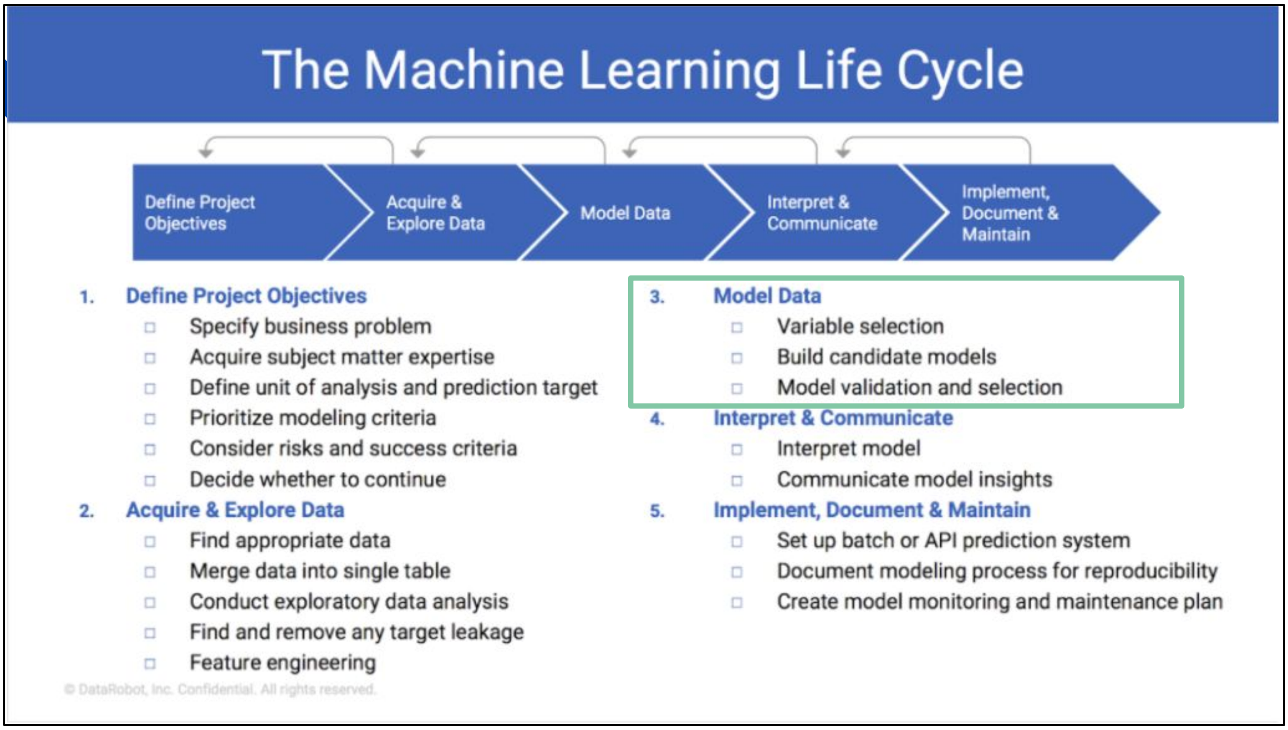
Acquire & Explore Data

Model D

1. **Define Project Objectives**
   - □ Specify business problem
   - □ Acquire subject matter expertise
   - □ Define unit of analysis and prediction target
   - □ Prioritize modeling criteria
   - □ Consider risks and success criteria
   - □ Decide whether to continue
2. **Acquire & Explore Data**
   - □ Find appropriate data
   - □ Merge data into single table
   - □ Conduct exploratory data analysis
   - □ Find and remove any target leakage
   - □ Feature engineering

- □ Set up batch or API prediction system
- □ Document modeling process for reproducibility
- □ Create model monitoring and maintenance plan

Ah, yes, we’ll need to go get our groceries. And potentially we’ll need to combine some of those groceries together and check to make sure that everything is fresh and doesn’t need to be further cleaned.

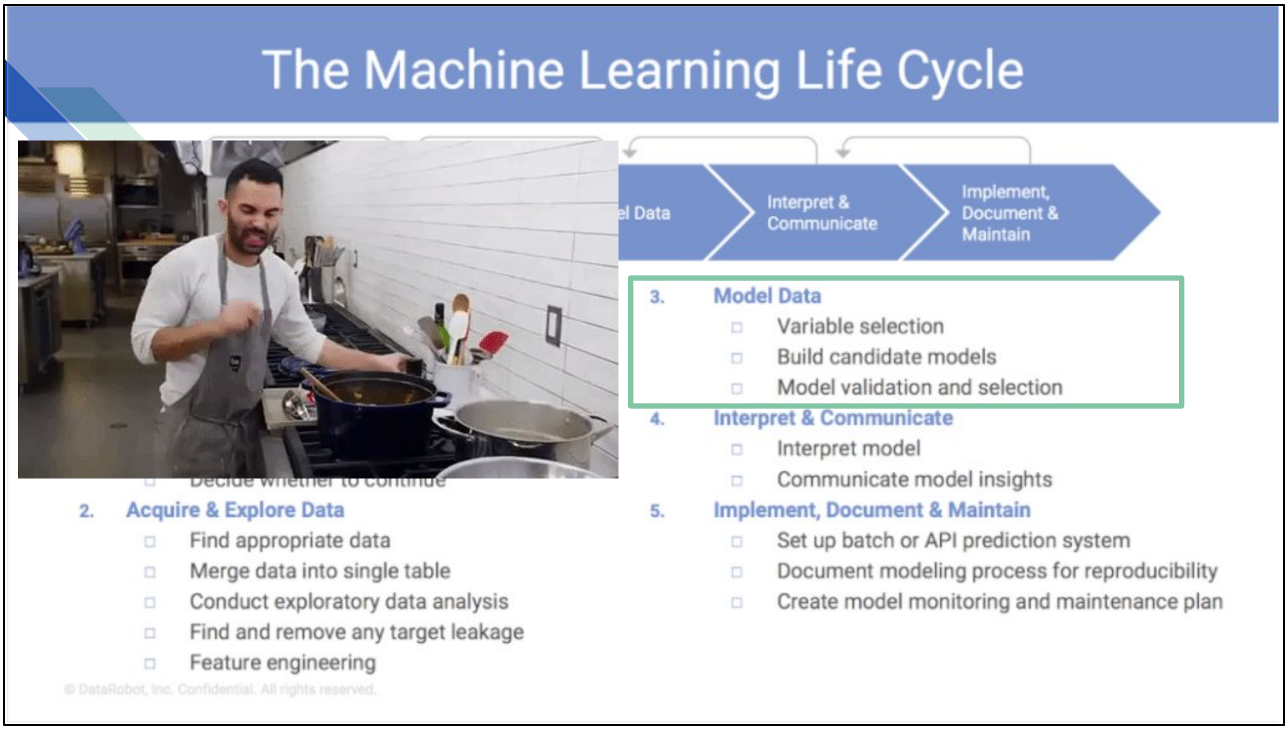


The Machine Learning Life Cycle

Define Project Objectives → Acquire & Explore Data → Model Data → Interpret & Communicate → Implement, Document & Maintain

1. **Define Project Objectives**
   - □ Specify business problem
   - □ Acquire subject matter expertise
   - □ Define unit of analysis and prediction target
   - □ Prioritize modeling criteria
   - □ Consider risks and success criteria
   - □ Decide whether to continue
2. **Acquire & Explore Data**
   - □ Find appropriate data
   - □ Merge data into single table
   - □ Conduct exploratory data analysis
   - □ Find and remove any target leakage
   - □ Feature engineering
3. **Model Data**
   - □ Variable selection
   - □ Build candidate models
   - □ Model validation and selection
4. **Interpret & Communicate**
   - □ Interpret model
   - □ Communicate model insights
5. **Implement, Document & Maintain**
   - □ Set up batch or API prediction system
   - □ Document modeling process for reproducibility
   - □ Create model monitoring and maintenance plan

After we have the data in the appropriate format matching with our specifications, we begin to build models for the data. How does this match up with our analogy?



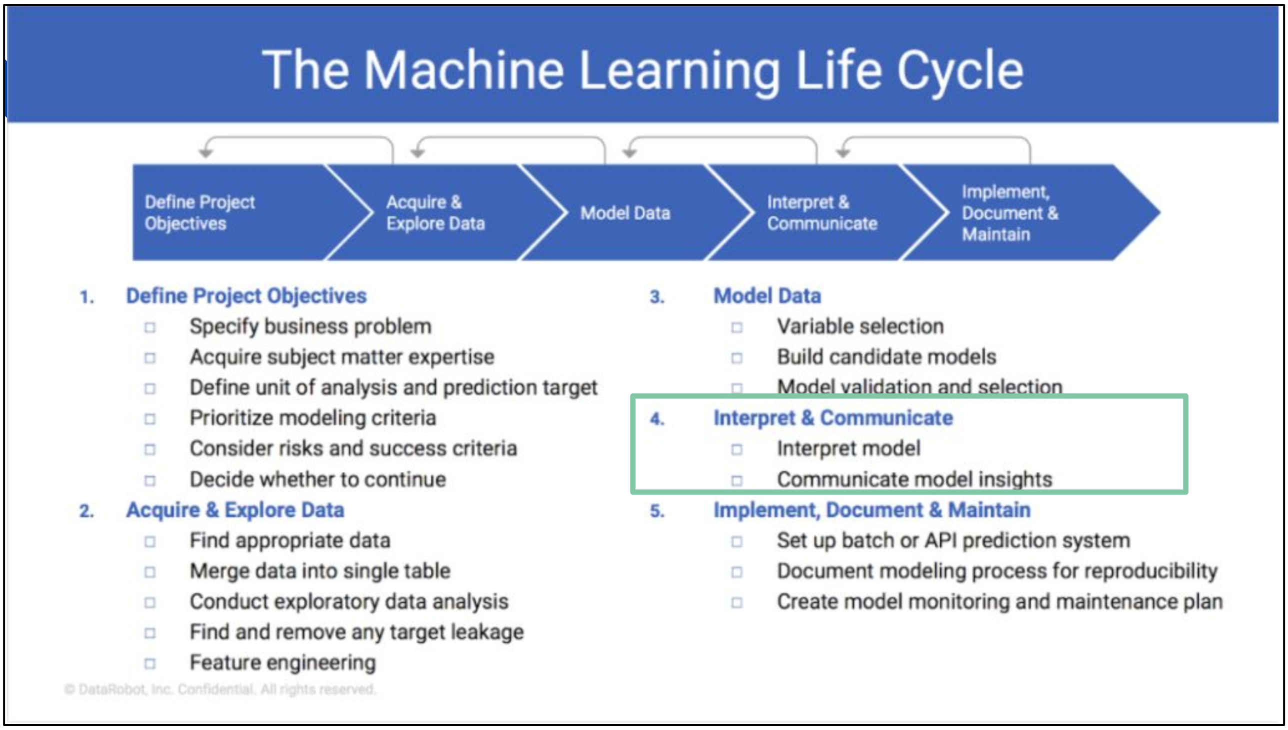
The Machine Learning Life Cycle

- □ Decide whether to continue

2. Acquire & Explore Data
   - □ Find appropriate data
   - □ Merge data into single table
   - □ Conduct exploratory data analysis
   - □ Find and remove any target leakage
   - □ Feature engineering

3. Model Data
   - □ Variable selection
   - □ Build candidate models
   - □ Model validation and selection
4. Interpret & Communicate
   - □ Interpret model
   - □ Communicate model insights
5. Implement, Document & Maintain
   - □ Set up batch or API prediction system
   - □ Document modeling process for reproducibility
   - □ Create model monitoring and maintenance plan

Now we are cooking! This part relates to making the meal or potentially multiple meals at once to test to see what we like most.

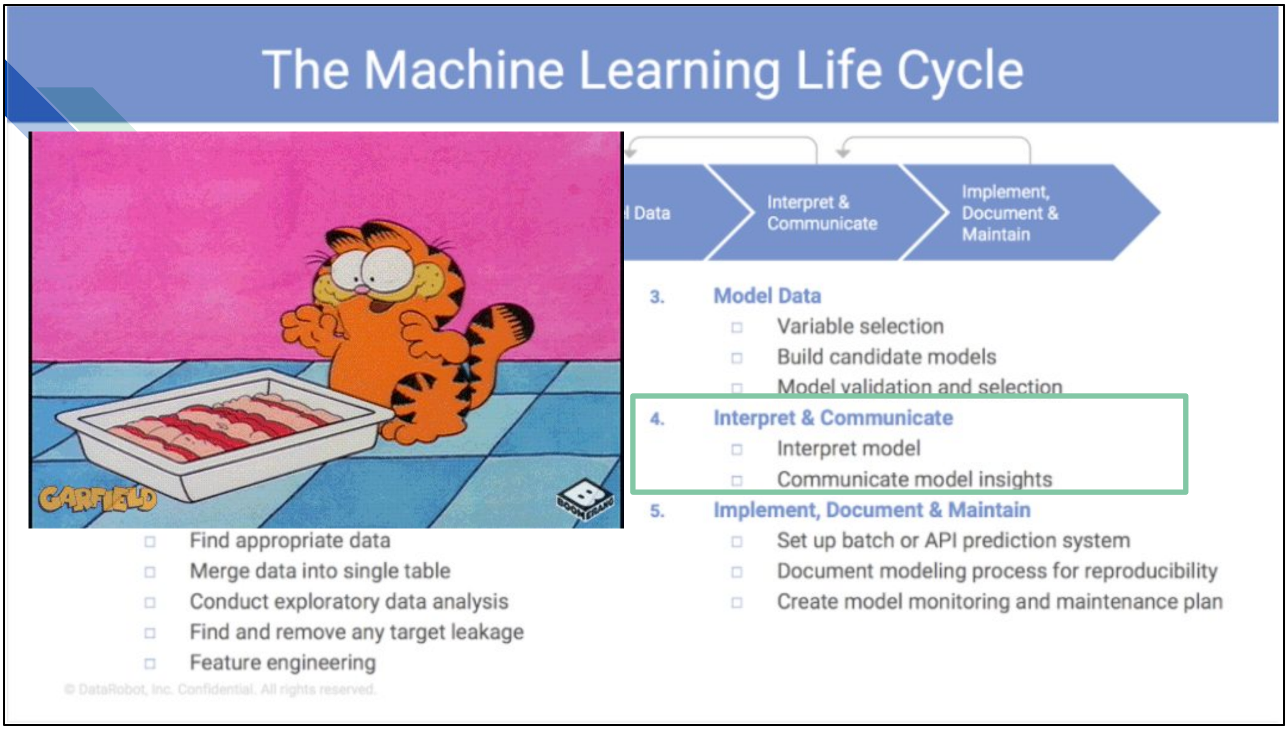


The Machine Learning Life Cycle

Define Project Objectives → Acquire & Explore Data → Model Data → Interpret & Communicate → Implement, Document & Maintain

1. **Define Project Objectives**
   - □ Specify business problem
   - □ Acquire subject matter expertise
   - □ Define unit of analysis and prediction target
   - □ Prioritize modeling criteria
   - □ Consider risks and success criteria
   - □ Decide whether to continue
2. **Acquire & Explore Data**
   - □ Find appropriate data
   - □ Merge data into single table
   - □ Conduct exploratory data analysis
   - □ Find and remove any target leakage
   - □ Feature engineering
3. **Model Data**
   - □ Variable selection
   - □ Build candidate models
   - □ Model validation and selection
4. **Interpret & Communicate**
   - □ Interpret model
   - □ Communicate model insights
5. **Implement, Document & Maintain**
   - □ Set up batch or API prediction system
   - □ Document modeling process for reproducibility
   - □ Create model monitoring and maintenance plan

Next we interpret the results of our model fits. Over what portions of the data does our model perform well? How do we feel about the performance of the models? What is this in our analogy?

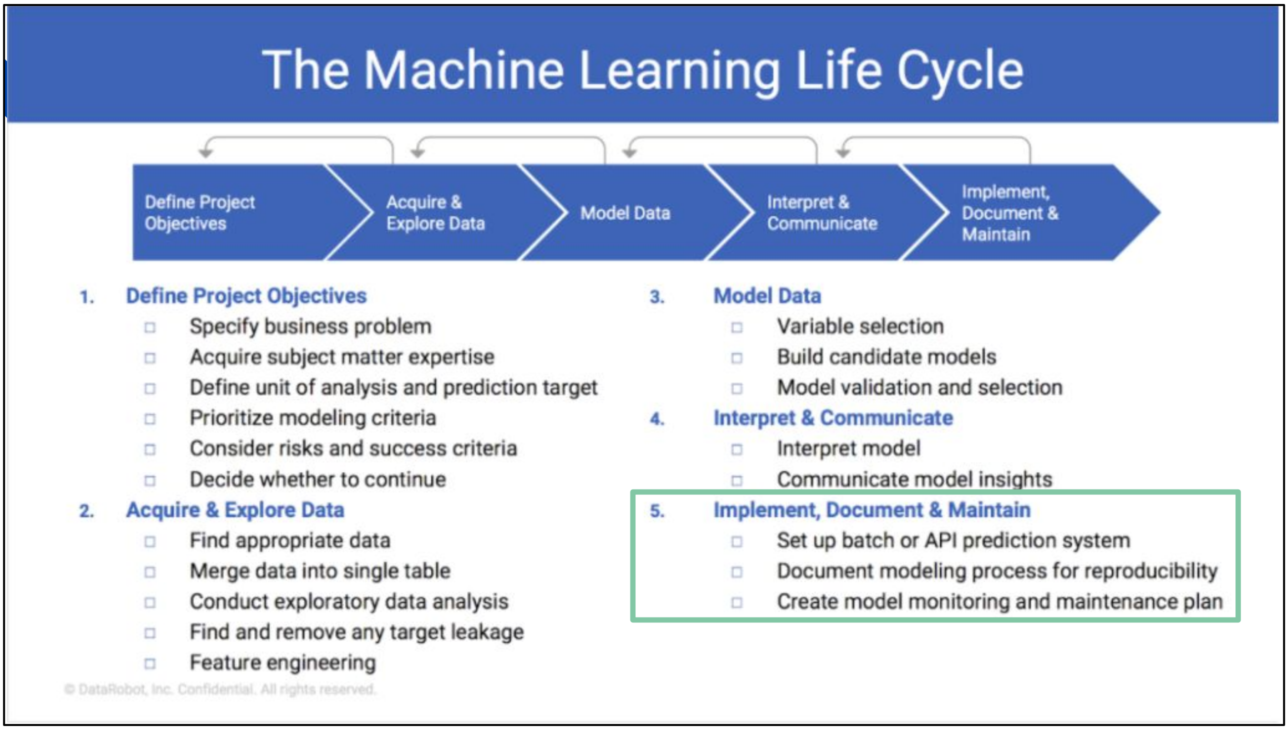


The Machine Learning Life Cycle

- □ Find appropriate data
- □ Merge data into single table
- □ Conduct exploratory data analysis
- □ Find and remove any target leakage
- □ Feature engineering

3. **Model Data**
   - □ Variable selection
   - □ Build candidate models
   - □ Model validation and selection
4. **Interpret & Communicate**
   - □ Interpret model
   - □ Communicate model insights
5. **Implement, Document & Maintain**
   - □ Set up batch or API prediction system
   - □ Document modeling process for reproducibility
   - □ Create model monitoring and maintenance plan

Yep. Now is time to eat! We want to try to understand what we liked about the meal in terms of taste. Is there anything we should tweak? Does it taste like what we thought? Could we make the meal again with new ingredients?

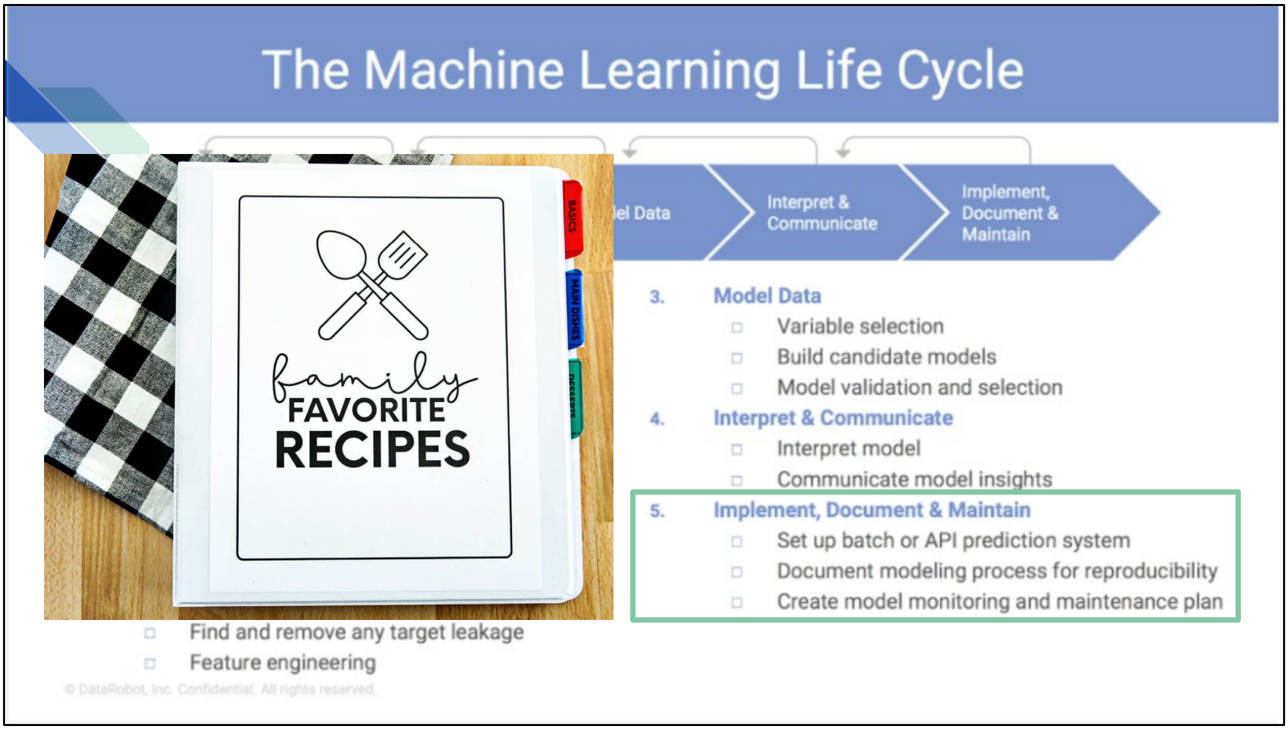


The Machine Learning Life Cycle

Define Project Objectives → Acquire & Explore Data → Model Data → Interpret & Communicate → Implement, Document & Maintain

1. **Define Project Objectives**
   - Specify business problem
   - Acquire subject matter expertise
   - Define unit of analysis and prediction target
   - Prioritize modeling criteria
   - Consider risks and success criteria
   - Decide whether to continue
2. **Acquire & Explore Data**
   - Find appropriate data
   - Merge data into single table
   - Conduct exploratory data analysis
   - Find and remove any target leakage
   - Feature engineering
3. **Model Data**
   - Variable selection
   - Build candidate models
   - Model validation and selection
4. **Interpret & Communicate**
   - Interpret model
   - Communicate model insights
5. **Implement, Document & Maintain**
   - Set up batch or API prediction system
   - Document modeling process for reproducibility
   - Create model monitoring and maintenance plan

Now we head to the last step in the cycle. Here we document our results and processes and send our model into production. We can then send new data to this model to get new prediction results.



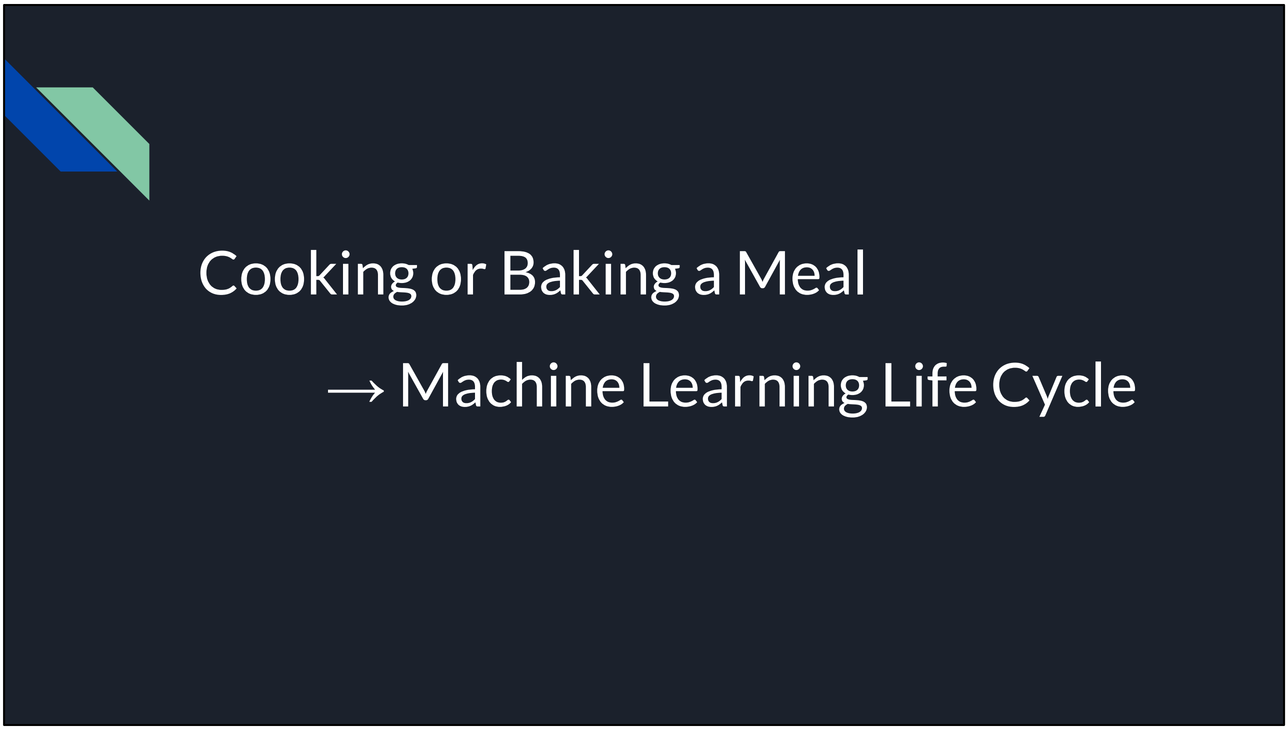
The Machine Learning Life Cycle

el Data → Interpret & Communicate → Implement, Document & Maintain

3. Model Data
   - □ Variable selection
   - □ Build candidate models
   - □ Model validation and selection
4. Interpret & Communicate
   - □ Interpret model
   - □ Communicate model insights
5. Implement, Document & Maintain
   - □ Set up batch or API prediction system
   - □ Document modeling process for reproducibility
   - □ Create model monitoring and maintenance plan

- □ Find and remove any target leakage
- □ Feature engineering

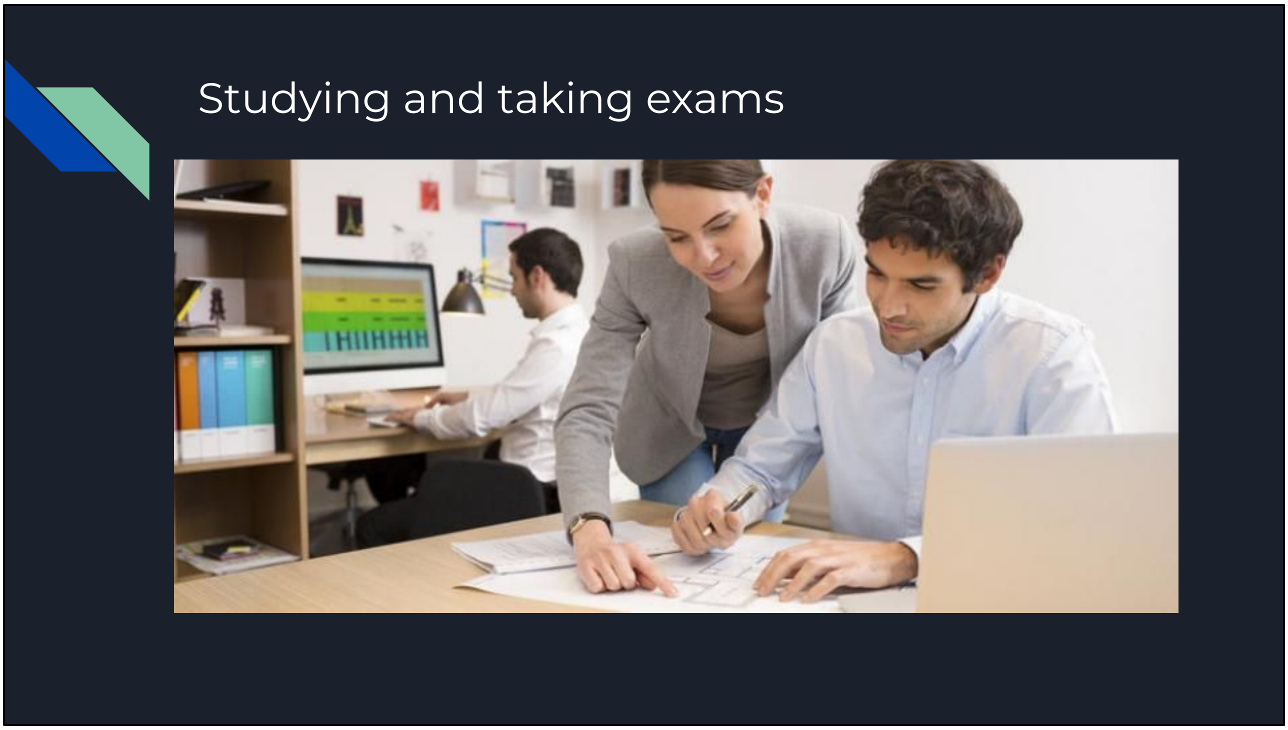
If we really like a particular recipe and the ingredients we used, we should try to create the meal again using similar ingredients. But if maybe our favorite tomato sauce isn't available in the store anymore, we need to potentially change our recipe to adapt.



Cooking or Baking a Meal

→ Machine Learning Life Cycle

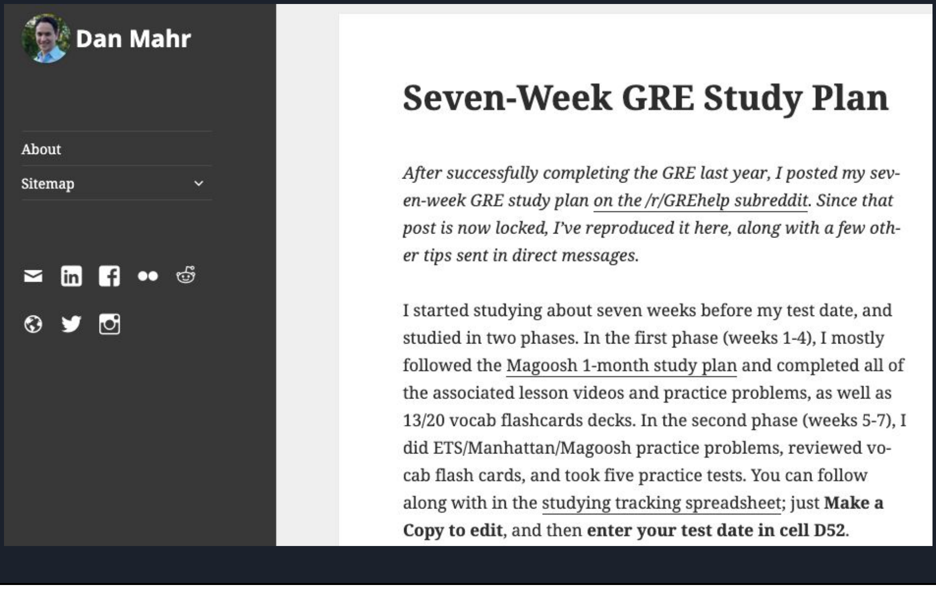
This is the same sort of thing that happens with deploying models. We need to make sure that the new data being sent in for predictions is similar to what was used to build the model. Let’s now head to another example.



Studying and taking exams

We've all studied for exams at some point in our lives. I think I didn't end up being good at taking exams and actually learning until I got to graduate school.

One successful strategy



Dan Mahr

About

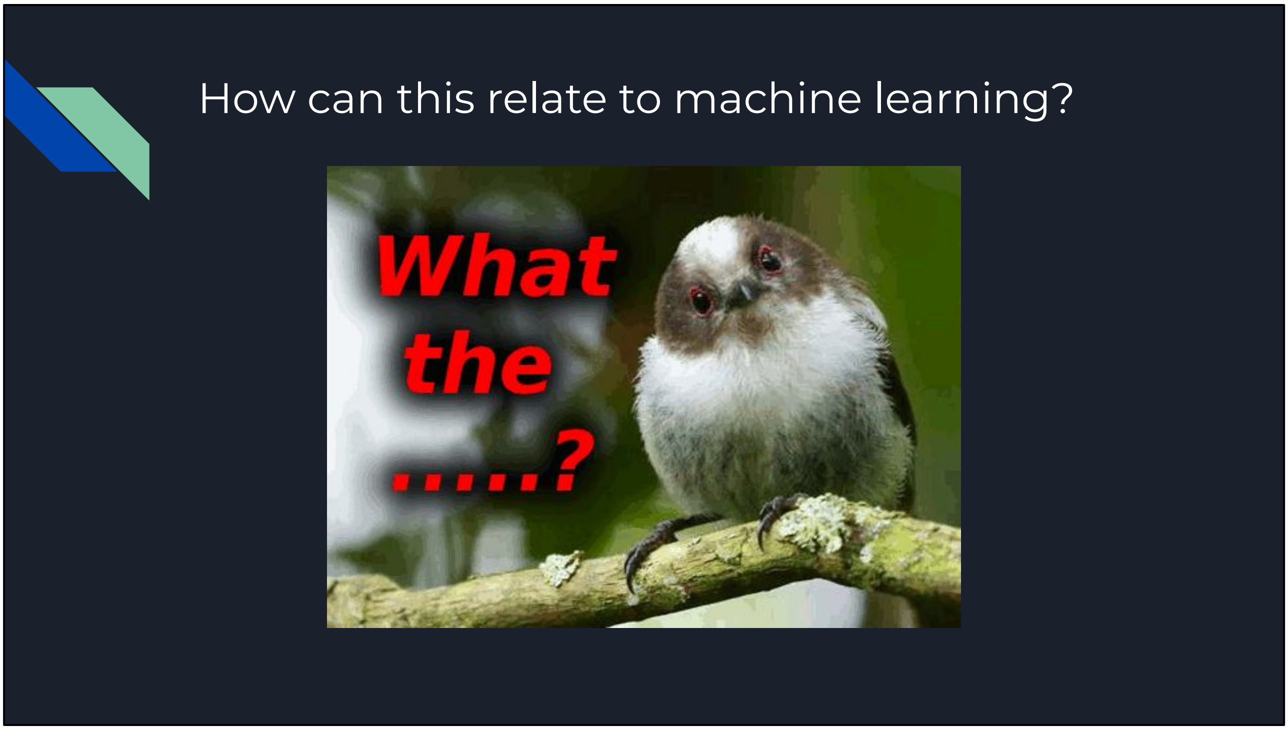
Sitemap

Seven-Week GRE Study Plan

After successfully completing the GRE last year, I posted my seven-week GRE study plan on the /r/GREhelp subreddit. Since that post is now locked, I've reproduced it here, along with a few other tips sent in direct messages.

I started studying about seven weeks before my test date, and studied in two phases. In the first phase (weeks 1-4), I mostly followed the Magoosh 1-month study plan and completed all of the associated lesson videos and practice problems, as well as 13/20 vocab flashcards decks. In the second phase (weeks 5-7), I did ETS/Manhattan/Magoosh practice problems, reviewed vocab flash cards, and took five practice tests. You can follow along with in the studying tracking spreadsheet; just **Make a Copy to edit**, and then **enter your test date in cell D52**.

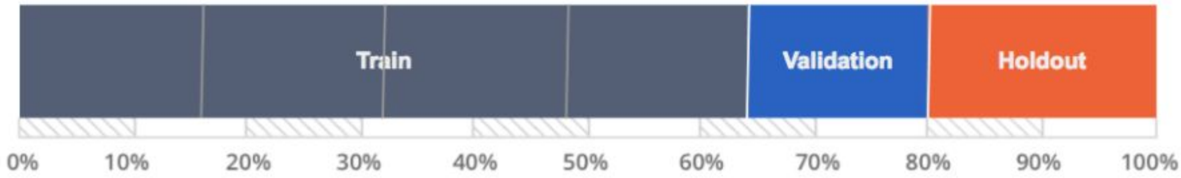
I'll let you read over this exam testing strategy and then let me know about some things that stand out to you as you read it. This is also a link to the webpage and I'll open up the Google Sheet. We see that there is a portion of the schedule where the person is learning the new material and practicing or training to get better. Then there is a period of time where they take practice tests to check their understanding and preparedness levels.



How can this relate to machine learning?

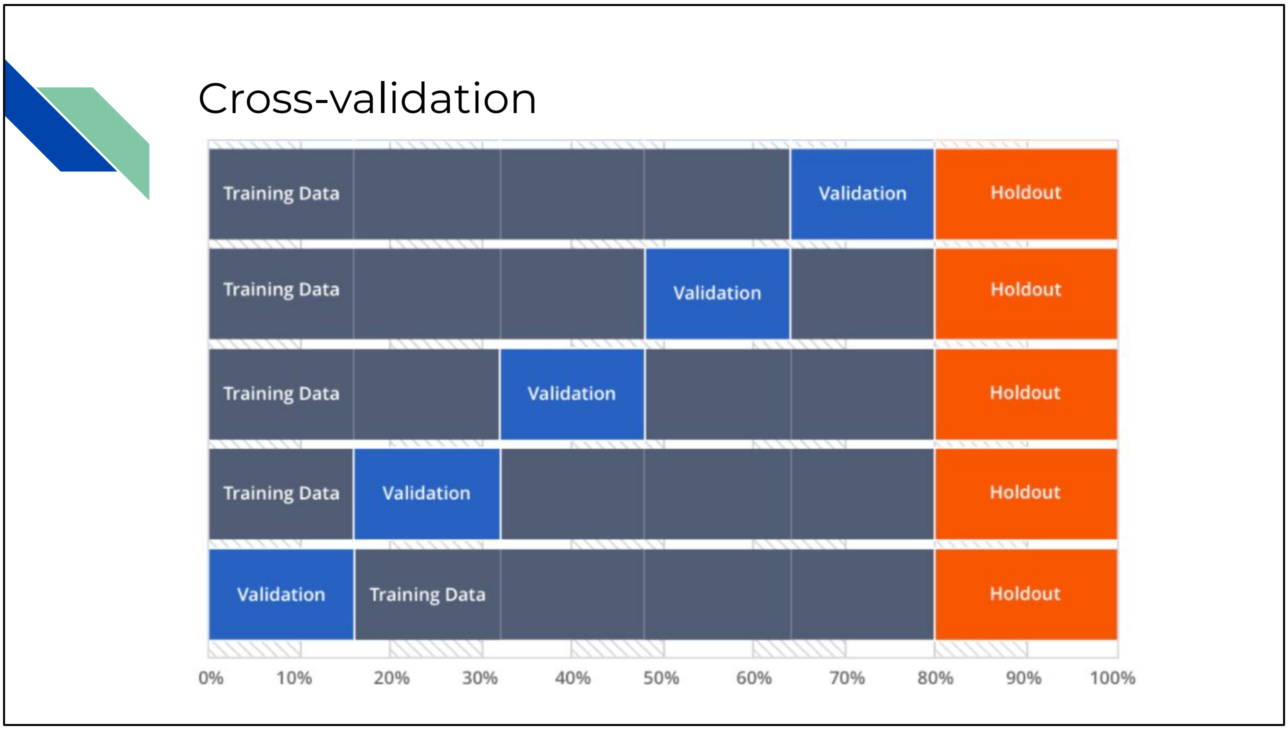
Does anyone have a guess to how this relates to building models, data science, or machine learning?

Training / Validation / Holdout



DataRobot wiki

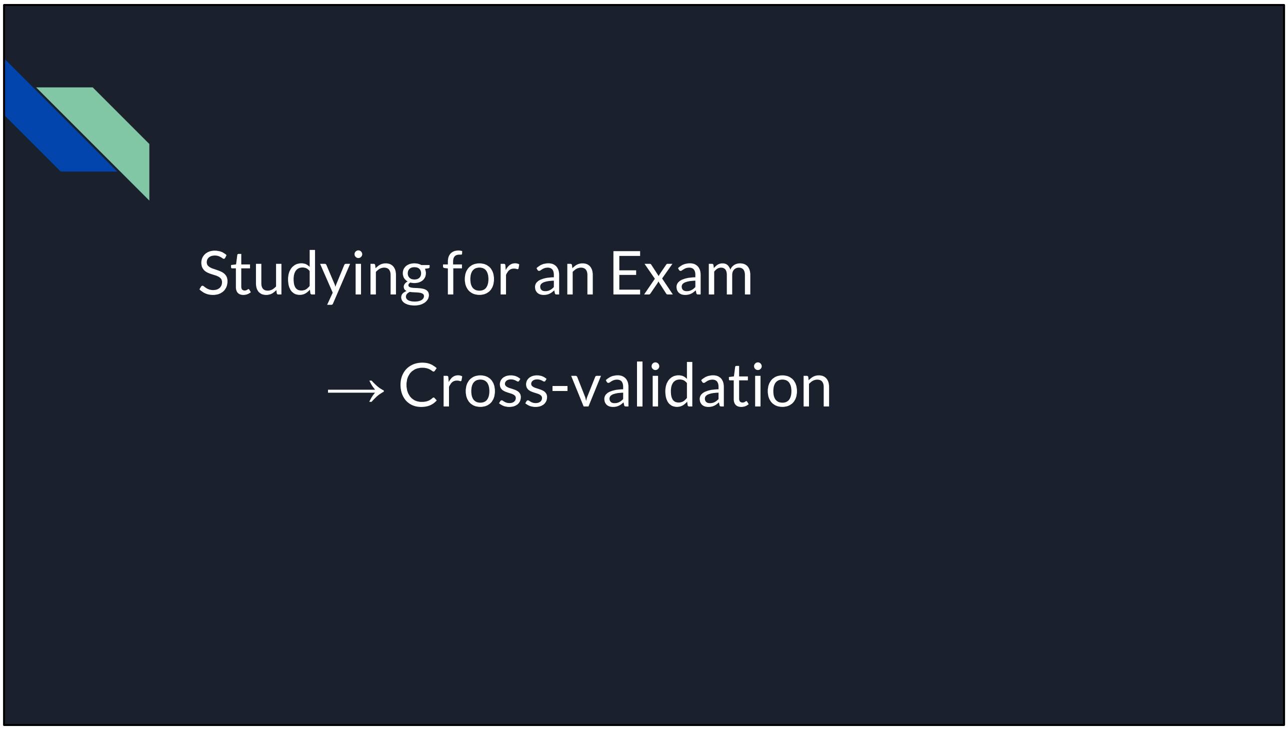
This is similar to splitting our data in training and testing partitions. During training the computer is learning about the data and trying to pull the signal from the data while ignoring the noise. This is commonly done to prevent overfitting to ensure that the model fit will perform well on unseen data. This data is in the validation and potentially a holdout portion of the data. This is used for simulating the process of fitting new data with the model.



Cross-validation

There's an extension of this called cross-validation. In cross-validation, we ensure that every row in the data appears exactly once in the validation partition. We then look to see how the model performs on the validation set as it moves randomly across the data.

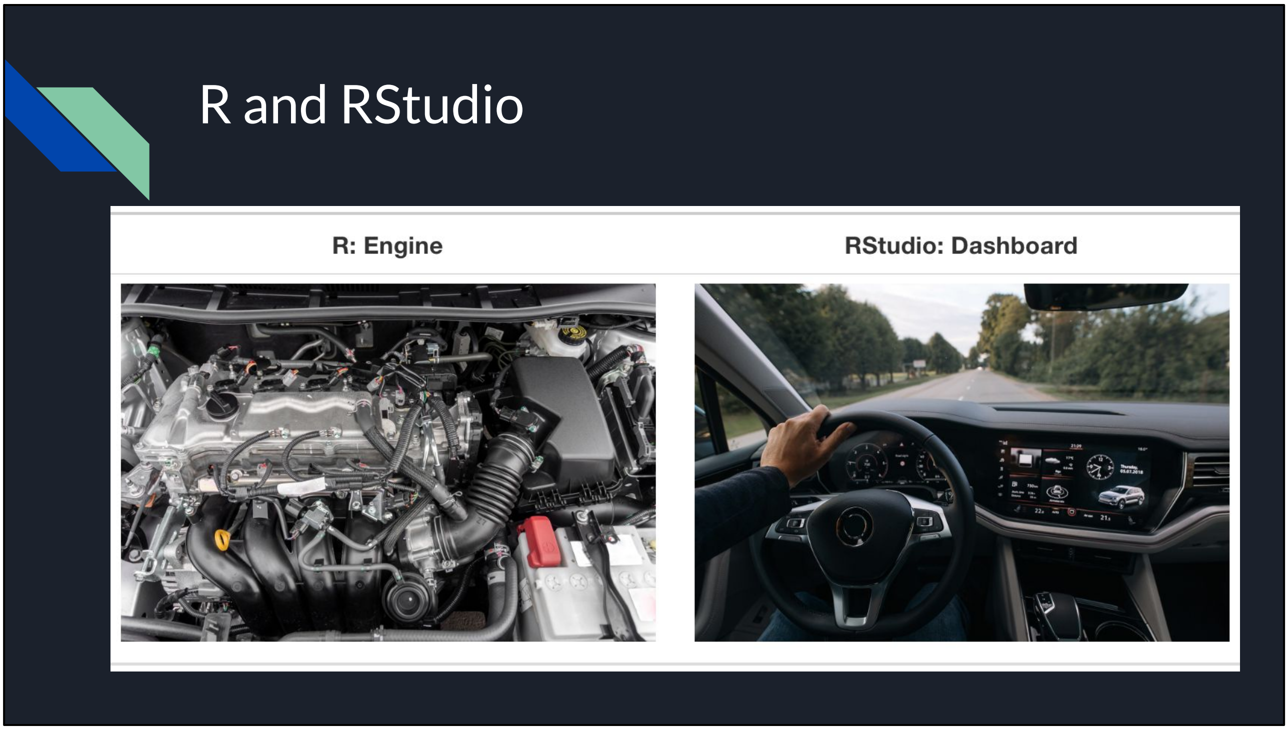
So how does this relate to studying for an exam, practice tests, and the actual exam?



Studying for an Exam

→ Cross-validation

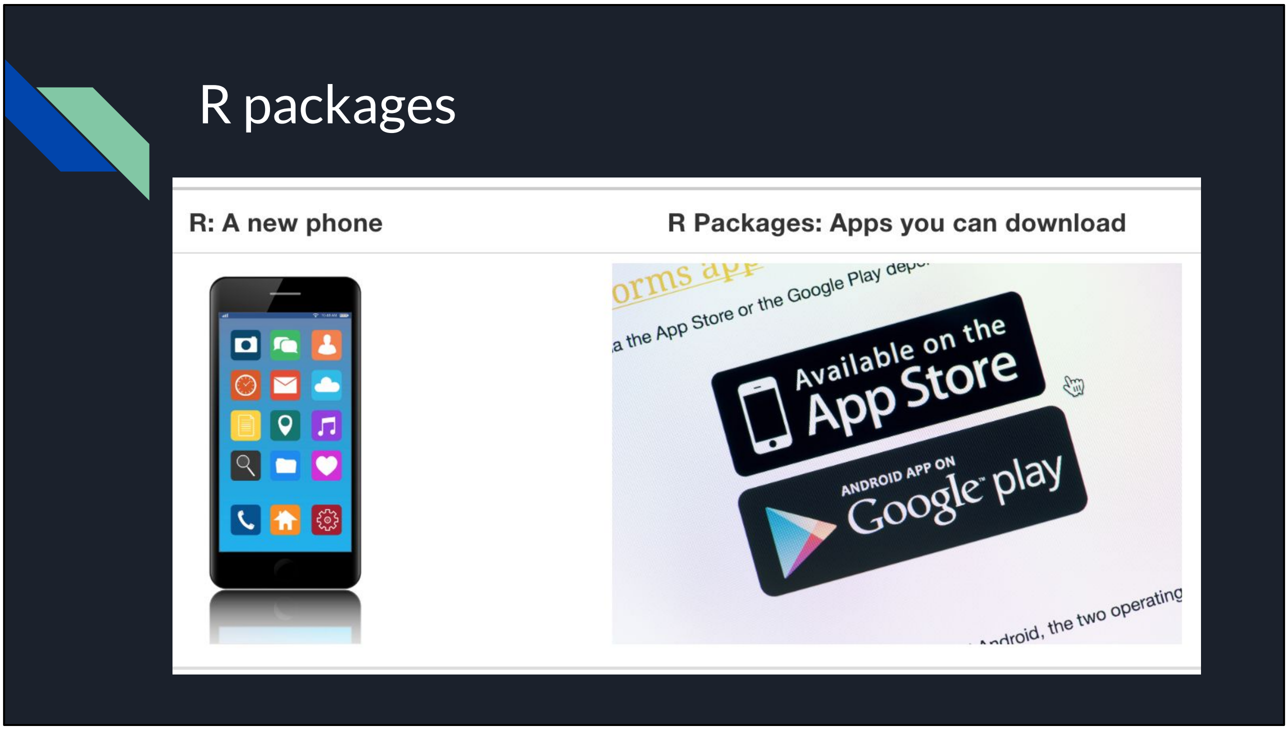
We’ve seen that the idea of cross-validation is closely tied to setting up the computer to take an exam. We want it to be able to perform well with its model on unseen data just like a student would perform well with good study habits on an exam.



R and RStudio

R: Engine	RStudio: Dashboard

Lastly, I wanted to talk about some programming topics. I'm an R programmer and it took awhile but I finally figured out a good analogy for R, the computing language, and RStudio, the interface. R is the engine behind the scenes and RStudio is the dashboard that you can tweak and customize as you wish to look at the engine's performance and results.



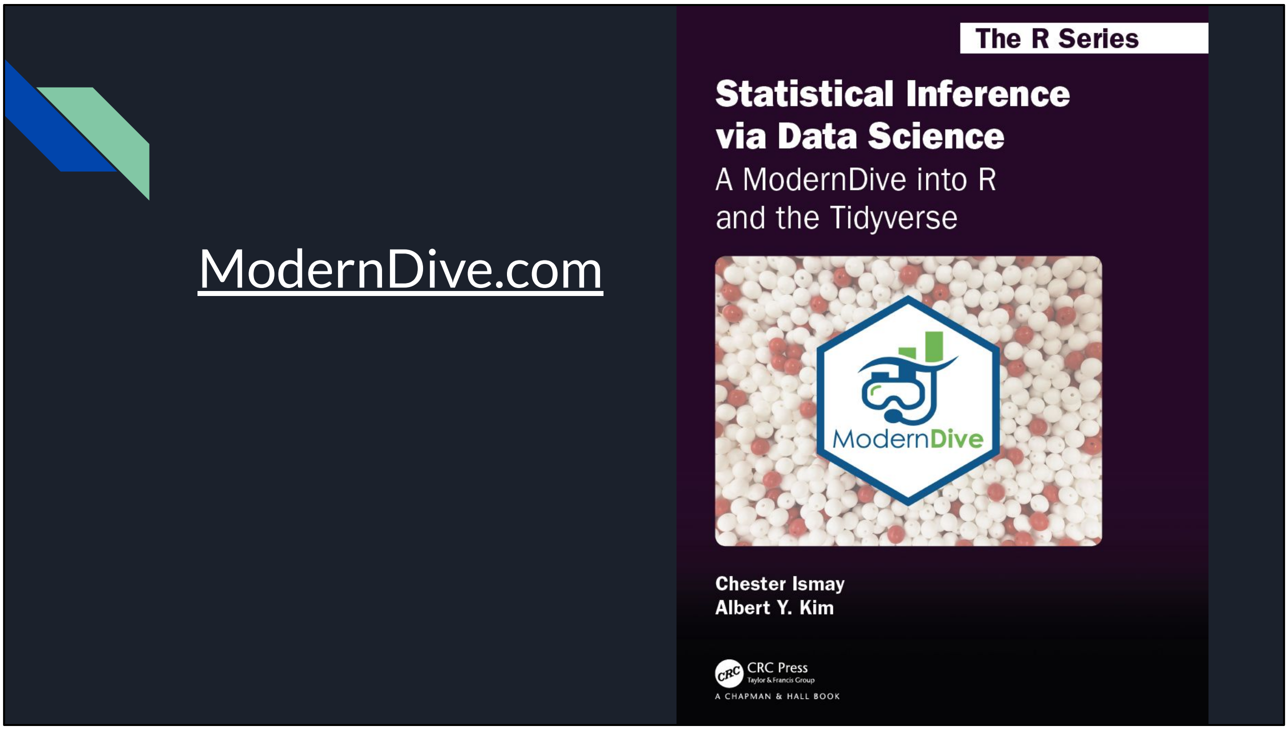
R packages

R: A new phone

R Packages: Apps you can download

orms app
a the App Store or the Google Play depo
Available on the App Store
ANDROID APP ON
Google play
Android, the two operating

Another analogy is with R and its packages. You can think of R as being a new phone and packages as being apps on your phone. In order to use an R package you need to first download and install it and then load it. With an app, you need to download the app to your phone and then click on it to load it.



ModernDive.com

The R Series

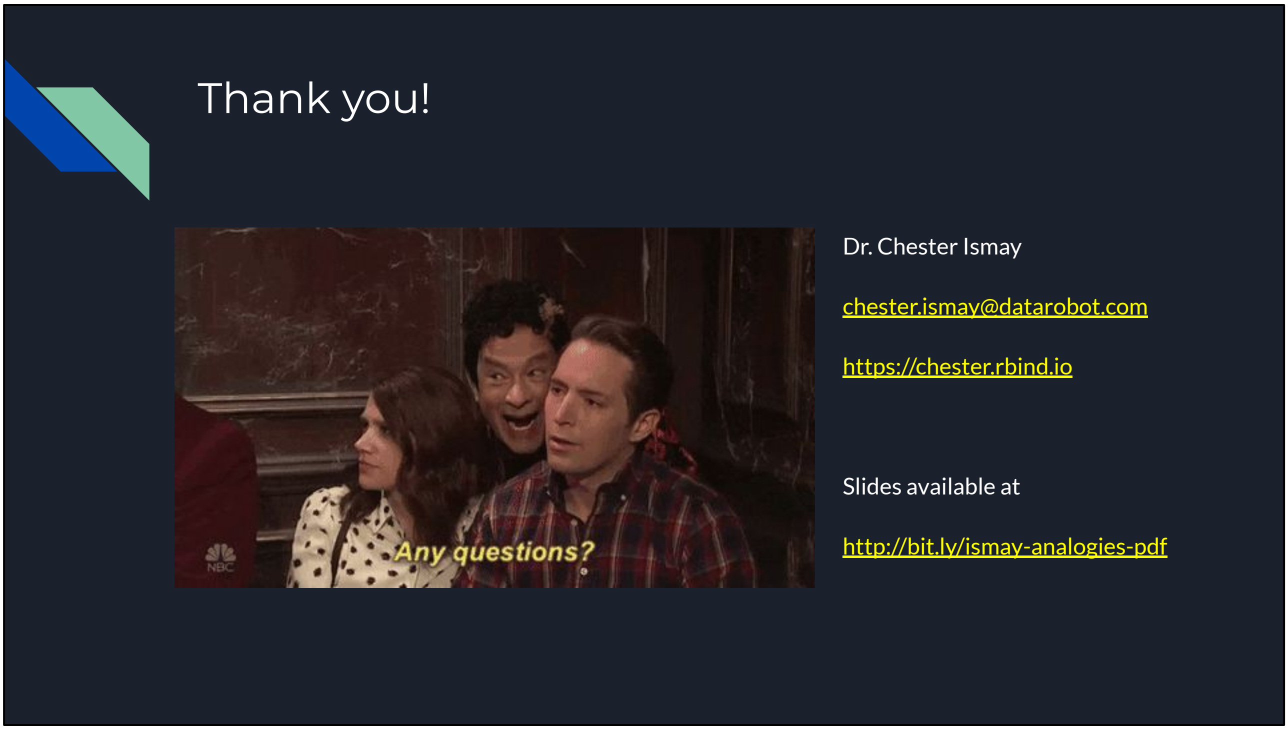
Statistical Inference via Data Science

A ModernDive into R and the Tidyverse

Chester Ismay
Albert Y. Kim

CRC Press
Taylor & Francis Group
A CHAPMAN & HALL BOOK

My colleague Albert Kim and I have tried to fit as many analogies as we could to build up data science and statistics topics in our book. It’s available for free at moderndive.com and should be released in print through CRC Press at the end of this year.



Thank you!

Any questions?

Dr. Chester Ismay

chester.ismay@datarobot.com

https://chester.rbind.io

Slides available at

http://bit.ly/ismay-analogies-pdf

Thanks much for attending! The slides are available in this link. Please email me if you have any questions and you can find other projects I’ve worked on at my webpage listed. Thanks again!