



Building Data Fluency through Analogies from the Real World and Movies

Chester Ismay, Ph.D.
Data Science Evangelist
DataRobot



How can we explain the following concepts via analogy?

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

Back to the Future II













The Weather

Today - Mostly sunny, high near 90, low near 60. The chance of precipitation is near zero through tonight. Tomorrow - Variable clouds with a high near 90. Tomorrow's temperature here range 80-94. Details Page C3.

Vol. XVII, No. 32

COMPLETE NEWS SERVICE
A. R. S. NEWS SERVICE

THURSDAY, APRIL 10, 1959

THIS NEWSPAPER OUR PEOPLE
OWNED 100 PER CENT

PRICE: 10 CENTS

Hill Valley Telegraph

Index	Sections
Amusements	B, 8
Business	B, 8
Calendar	B, 8
Classified	A, 2
Financial	D, 10
Market	C, 1
Obituary	D, 8
Police	B, 8
Sports	C, 1
Travel	B, 8
Weather	B, 8

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 measure from without to the peace of our continent concerns all of us and therefore property 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 Edition.
BECAUSE one of the signs in the city was taken out at a time when it was still in the highway near Oakland's charge of malicious mischief was filed in common court court yesterday afternoon. The sign was taken out according to complaint of W. S. Hagan, 1415 North Villa street.

The governmental signboard was cleaned with glass dust which gave the sign as illuminated at night. The signboard was found and that Hagan removed 25 of the signboard 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 measure from without to the peace of our continent concerns all of us and therefore property 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 a more detailed study will be made 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 said the council has finished its private meetings concerning the issue, that the Mayor will have to 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

BIFF WINS AGAIN

Khrushchev Offers Dates for Summit

An American newspaper is quoted as indicating that Soviet news reports will be used as the evidence in the case of the American. American news reports have indicated that there has been some reason of getting it a sentence.

Further press talk, it is reported, has been heard in the situation as it now stands. The evidence, it is said, is the fact that there have been many reports.

Further press talk, it is reported, has been heard in the situation as it now stands. The evidence, it is said, is the fact that there have been many reports.

It is not necessary that the evidence, it is said, be used as the evidence in the case of the American. American news reports have indicated that there has been some reason of getting it a sentence.

An American newspaper is quoted as indicating that Soviet news reports will be used as the evidence in the case of the American. American news reports have indicated that there has been some reason of getting it a sentence.



PHIL CORPUS
AP Wirephoto

While the first attempt at the summit has not been successful, it is possible that some of the details of the summit will be discussed in the future. The summit will be held in the future.

It is not necessary that the evidence, it is said, be used as the evidence in the case of the American. American news reports have indicated that there has been some reason of getting it a sentence.

An American newspaper is quoted as indicating that Soviet news reports will be used as the evidence in the case of the American. American news reports have indicated that there has been some reason of getting it a sentence.

Further press talk, it is reported, has been heard in the situation as it now stands. The evidence, it is said, is the fact that there have been many reports.

It is not necessary that the evidence, it is said, be used as the evidence in the case of the American. American news reports have indicated that there has been some reason of getting it a sentence.

An American newspaper is quoted as indicating that Soviet news reports will be used as the evidence in the case of the American. American news reports have indicated that there has been some reason of getting it a sentence.

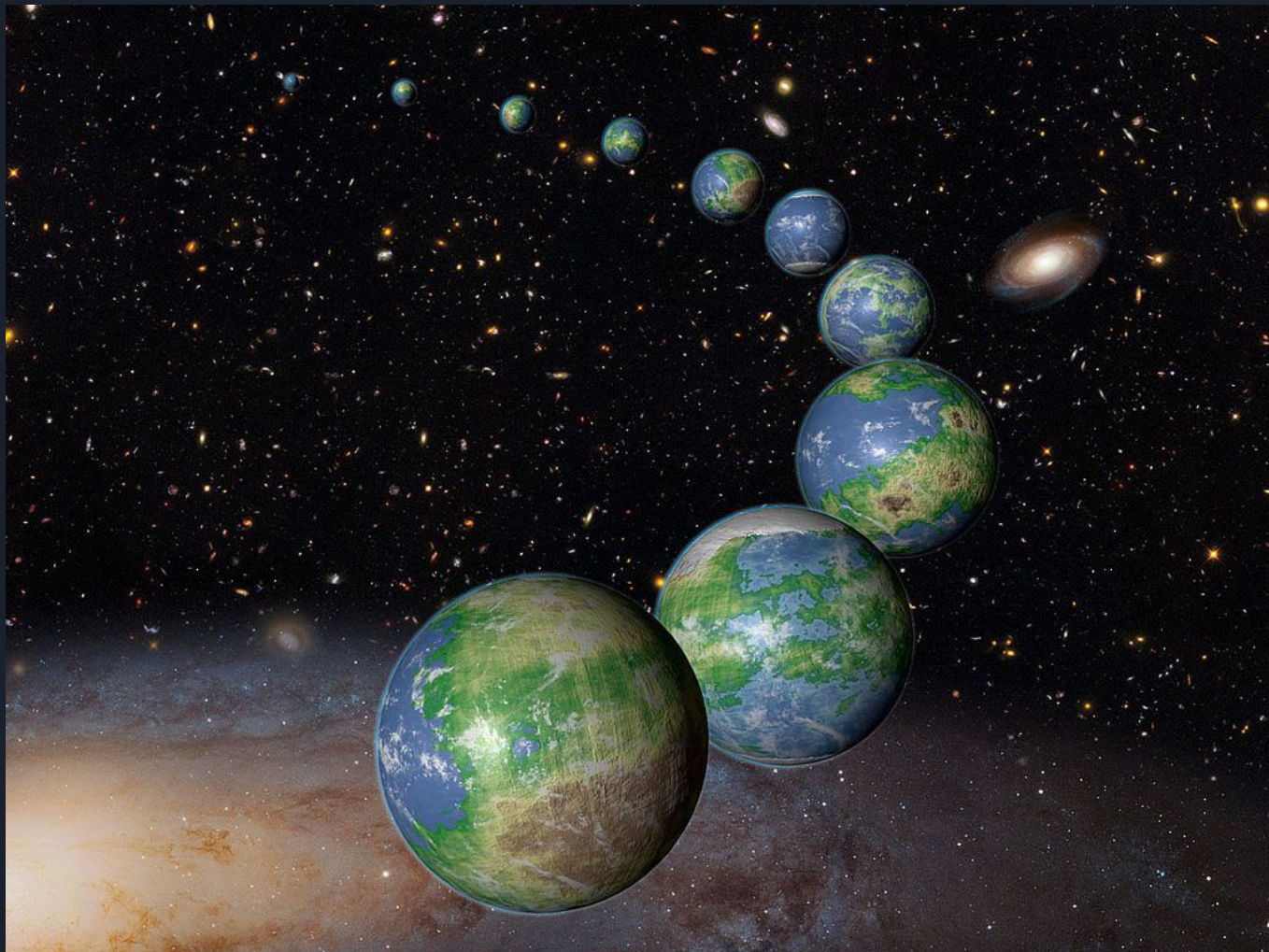
Further press talk, it is reported, has been heard in the situation as it now stands. The evidence, it is said, is the fact that there have been many reports.



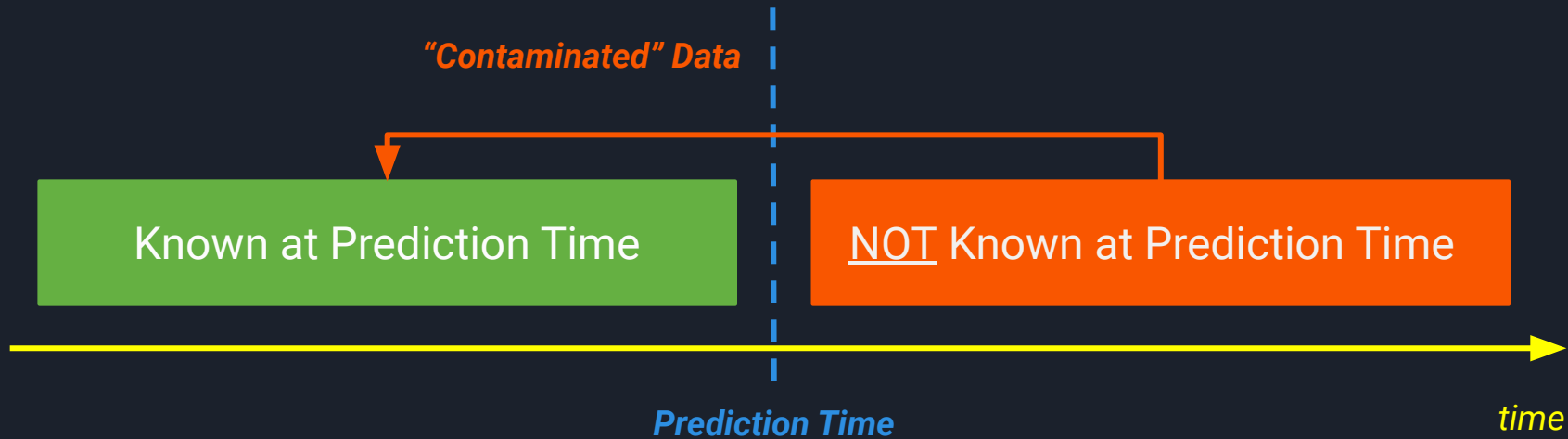




Back to the Future writer: ... Biff was based on Donald Trump



Target Leakage





Predicting Recidivism

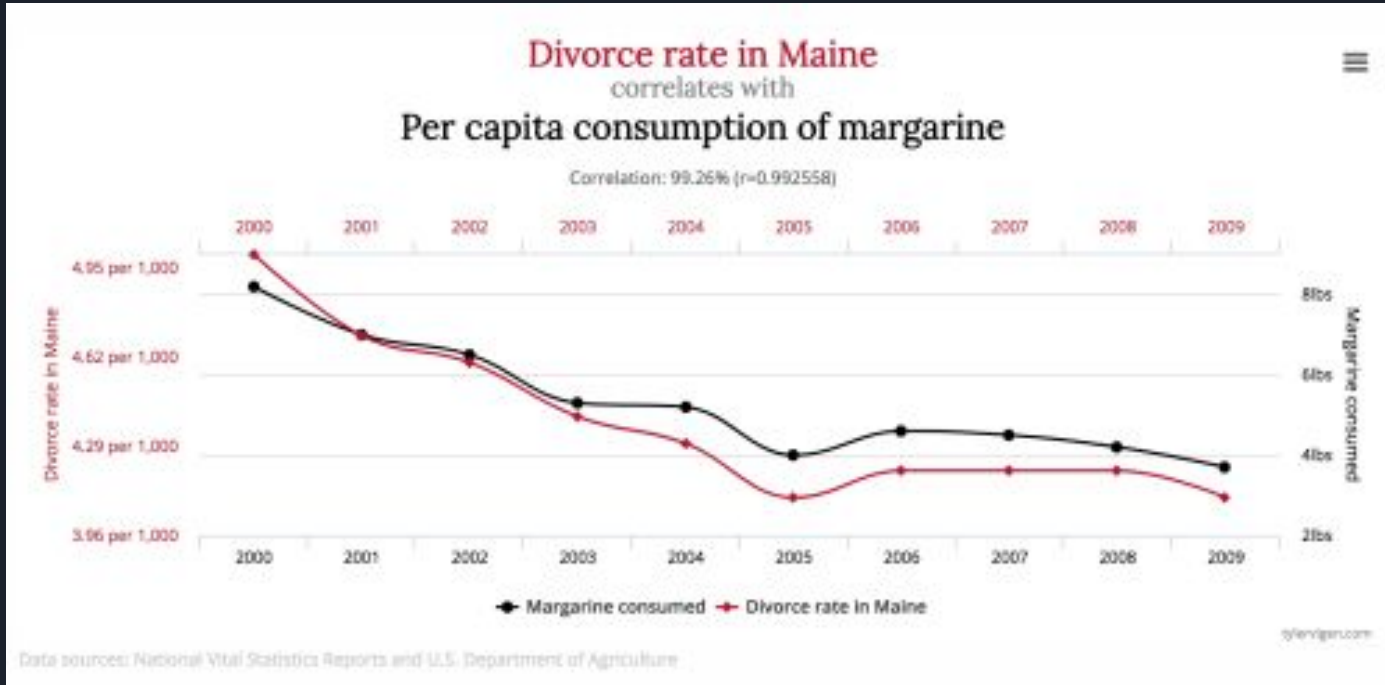


Preventing Target Leakage



[How to avoid the data leakage? blogpost](#)

Preventing Target Leakage



[Beware of Data Leakage blogpost](#)



Further References

- [What is Target Leakage?](#) - DataRobot Wiki
- [Leakage](#) - Data Skeptic 12 minute podcast

Prevent Target Leakage to Prevent Bad Parallel Universes



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

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

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



Is
selection

insights
maintain

prediction system

process for reproducibility

ing and maintenance plan

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

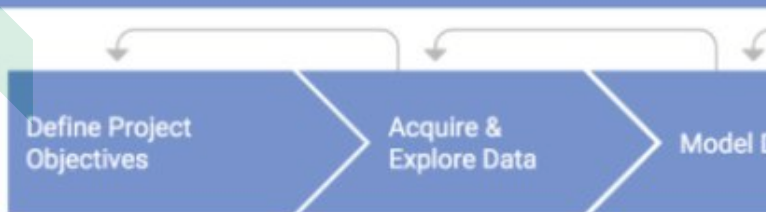
4. Interpret & Communicate

- 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

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



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

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

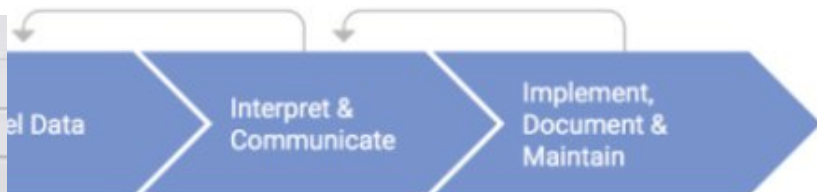
5. Implement, Document & Maintain

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

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

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

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

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

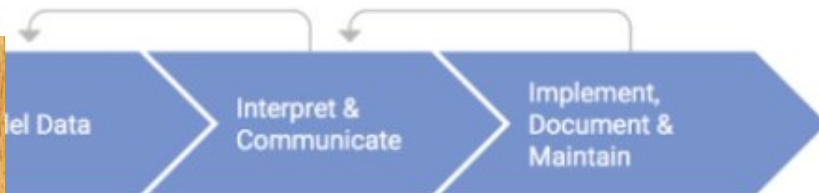
5. Implement, Document & Maintain

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

The Machine Learning Life Cycle



- ❑ 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

Studying and taking exams



One successful strategy



Dan Mahr

About

Sitemap



Seven-Week GRE Study Plan

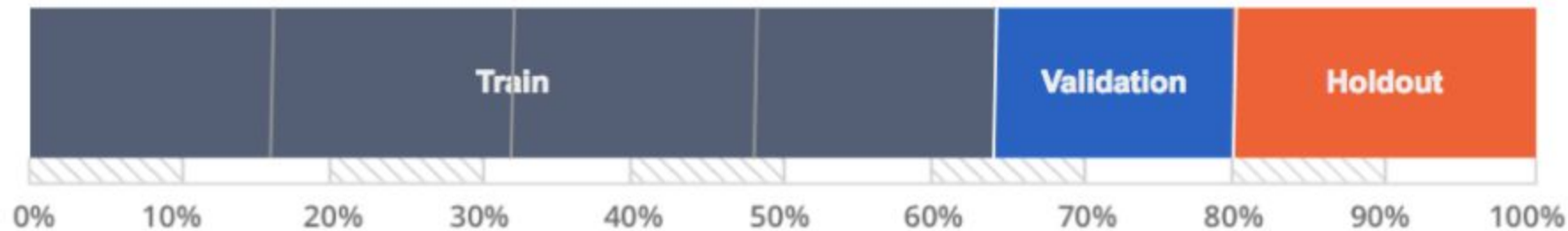
After successfully completing the GRE last year, I posted my seven-week GRE study plan on the [/r/GREhelp](#) subreddit. Since that post is now locked, I've reproduced it here, along with a few other tips sent in direct messages.

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**.

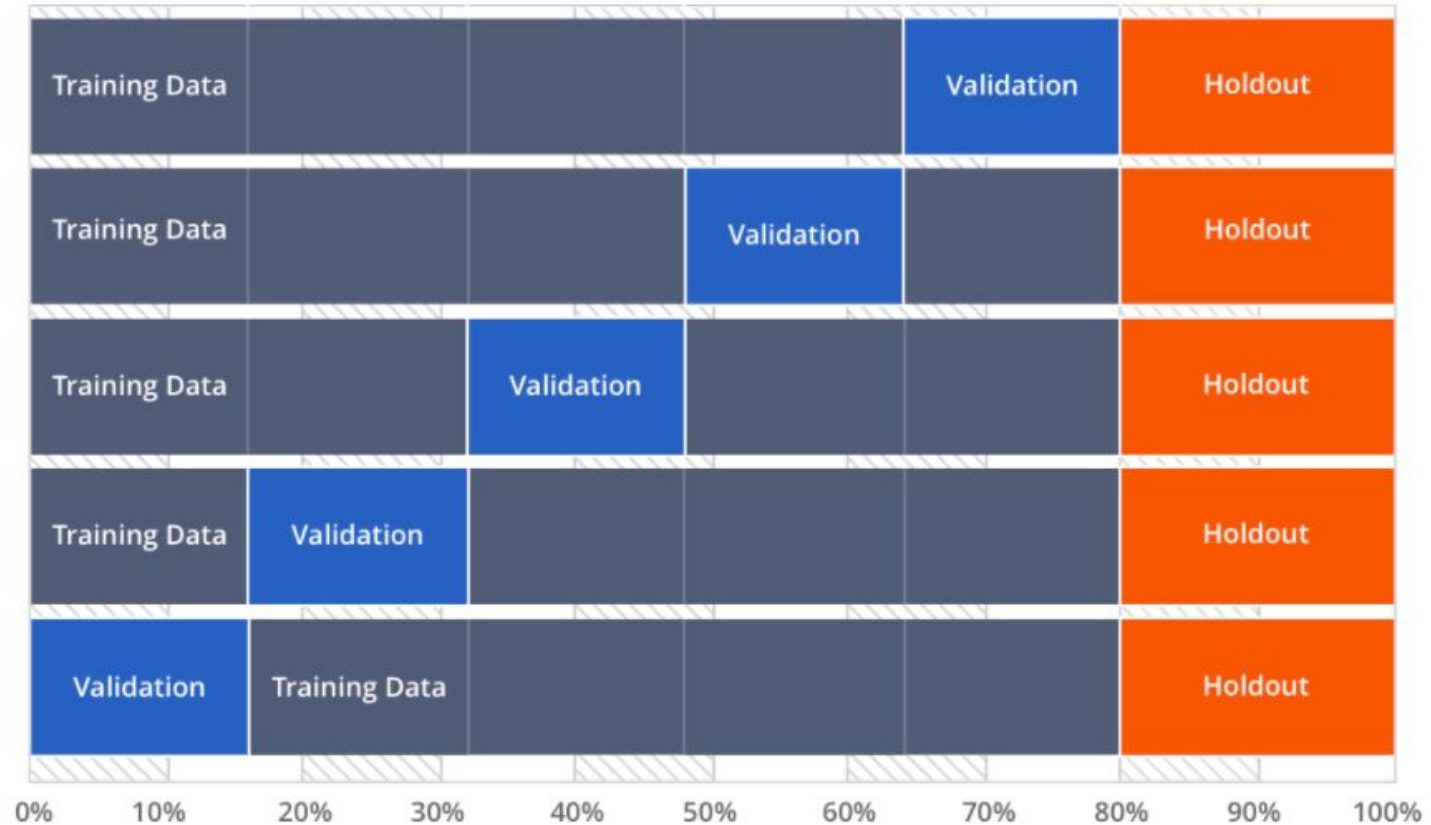
How can this relate to machine learning?



Training / Validation / Holdout



Cross-validation



R and RStudio

R: Engine



RStudio: Dashboard



R packages

R: A new phone



R Packages: Apps you can download





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



Thank you!



Dr. Chester Ismay

chester.ismay@datarobot.com

<https://chester.rbind.io>

Slides available at

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