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?

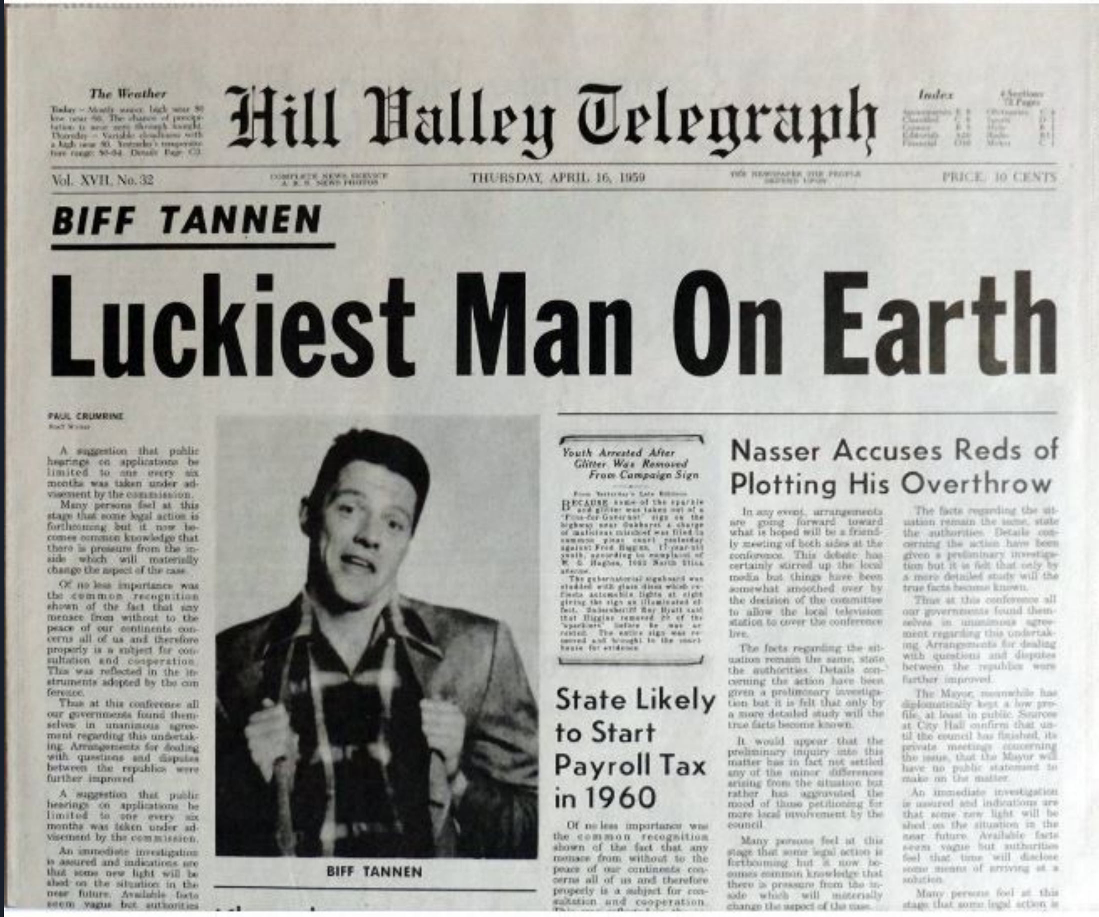
- R, RStudio, and R packages
- Cross-validation
- Machine Learning Life Cycle
- Target Leakage

Back to the Future Part II

GRAYS
ORTS
ANAC
1950–2000
SALES
SERVICE
INSTALLATION
LLECT

Hilldale

ALMANAC
1950–2000



The Weather

Hill Valley Telegraph

Index

Vol. XVII, No. 32

THURSDAY, APRIL 16, 1959

PRICE: 10 CENTS

BIFF TANNEN

Luckiest Man On Earth

PAUL CRUMRINE

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

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.

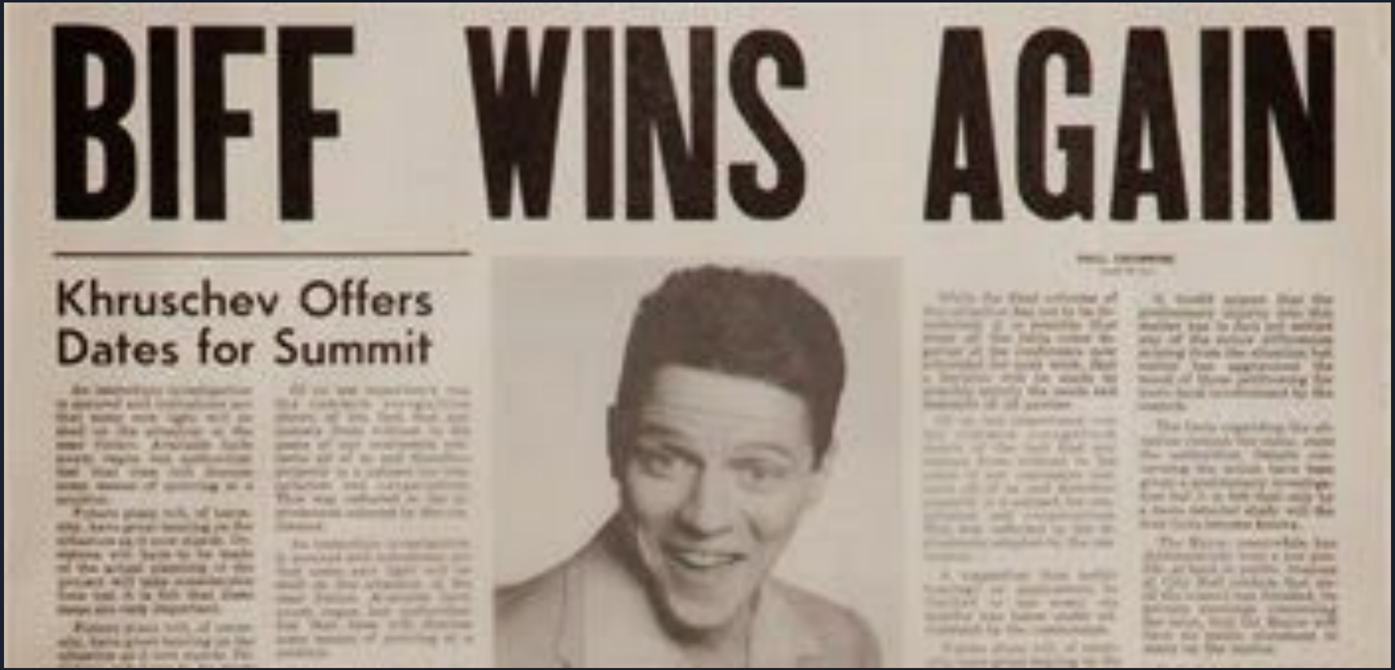
The facts regarding the situation remain the same, state the authorities. Details concerning the action have been given a preliminary investigation but it is felt that only by a more detailed study will the true facts become known.

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

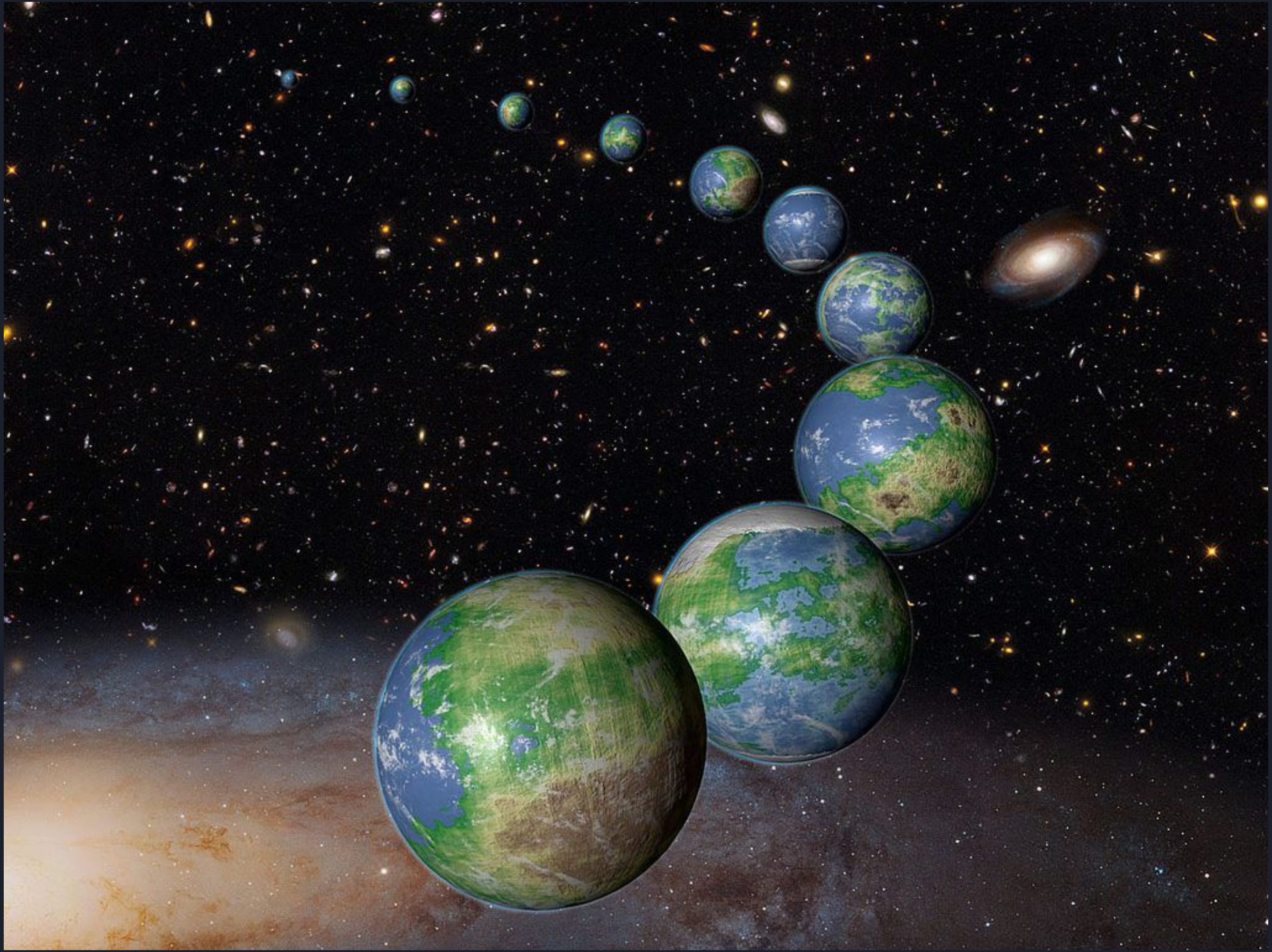


BIFF WINS AGAIN

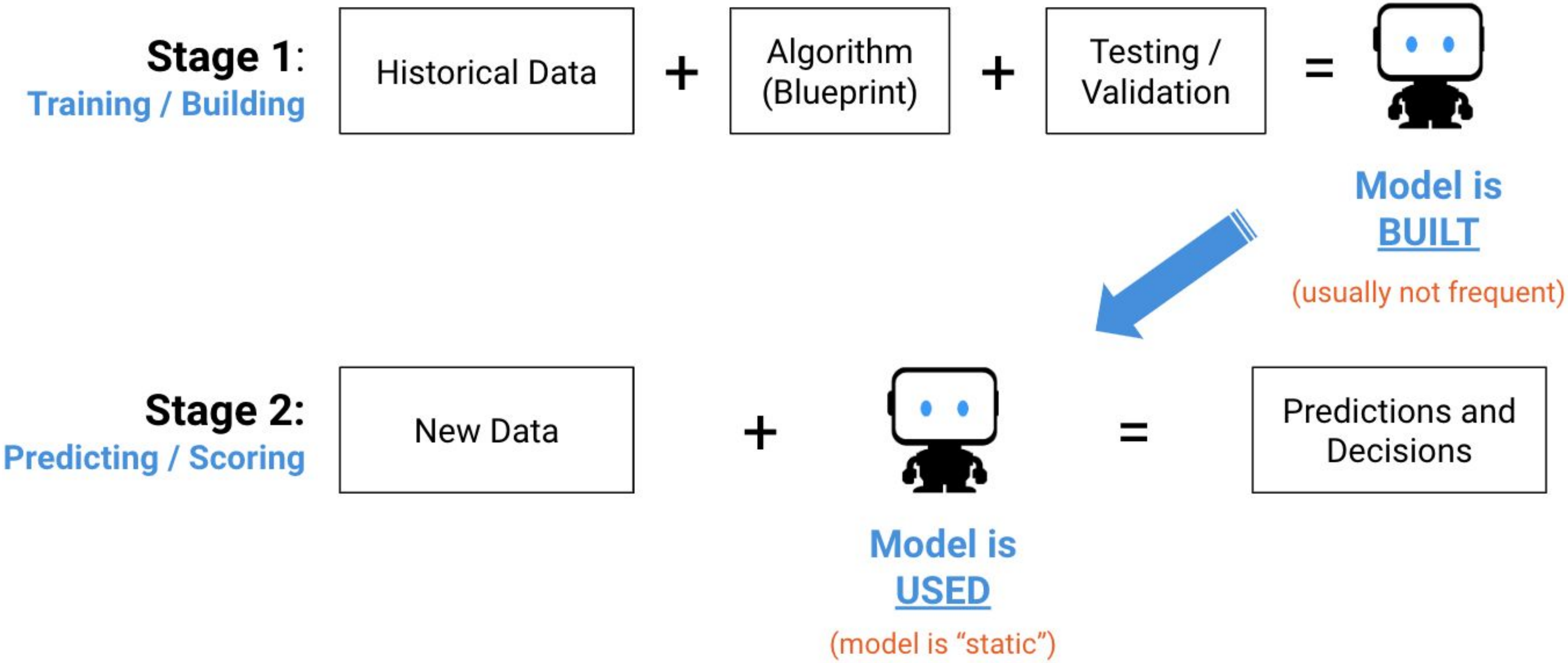
Khruschev Offers Dates for Summit

Biff's
Biff Tannen's
PLEASURE PARADISE
CASINO
HOTEL

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



Two-Stage Modeling Process



Stage 1:
Training / Building

Historical Data + Algorithm (Blueprint) + Testing / Validation =

Model is BUILT

(usually not frequent)

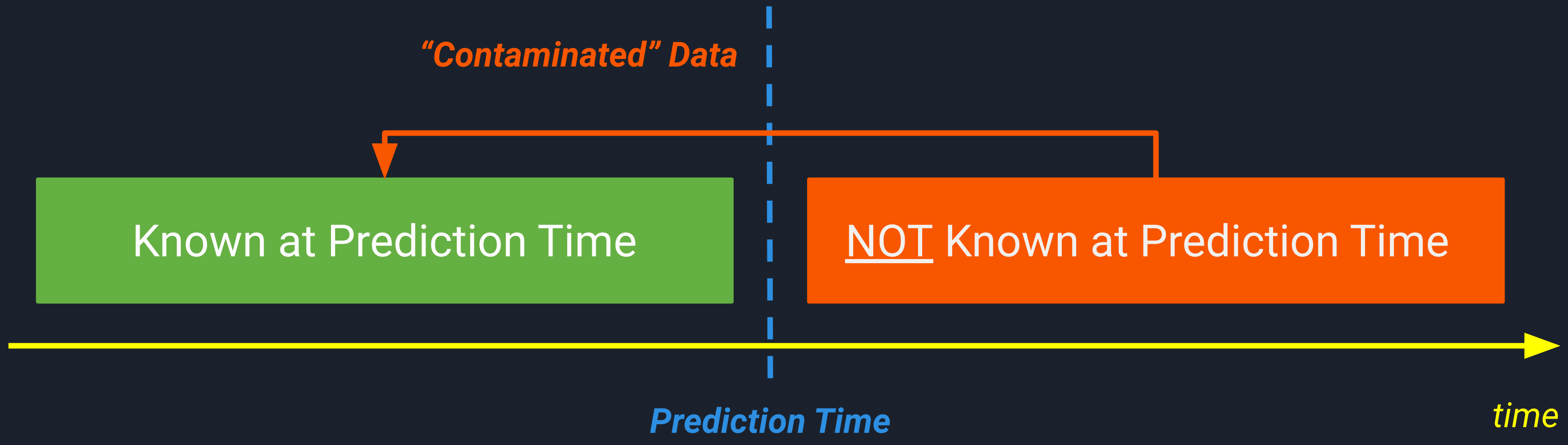
Stage 2:
Predicting / Scoring

New Data + [model] = Predictions and Decisions

Model is USED

(model is “static”)

Target Leakage



"Contaminated" Data

Known at Prediction Time

NOT Known at Prediction Time

Prediction Time

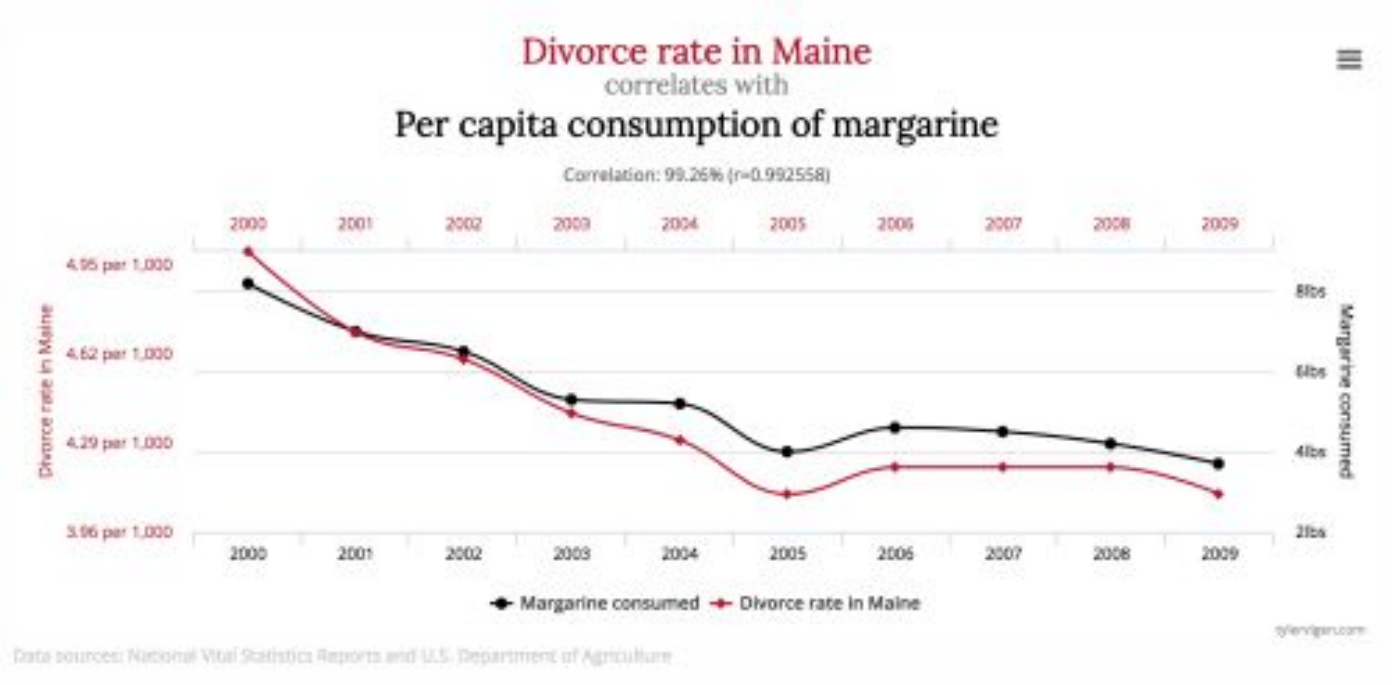
time

Predicting Recidivism

Preventing Target Leakage

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

Preventing Target Leakage



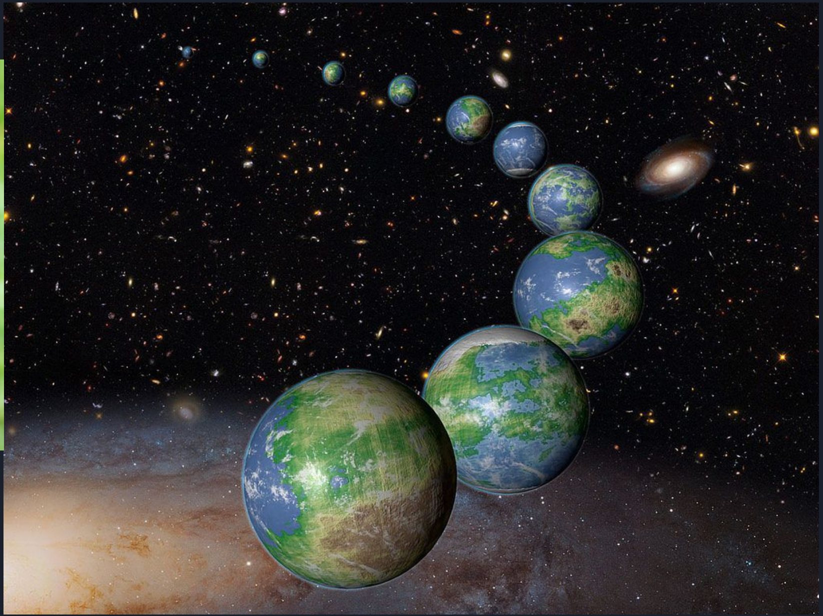
Data sources: National Vital Statistics Reports and U.S. Department of Agriculture

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

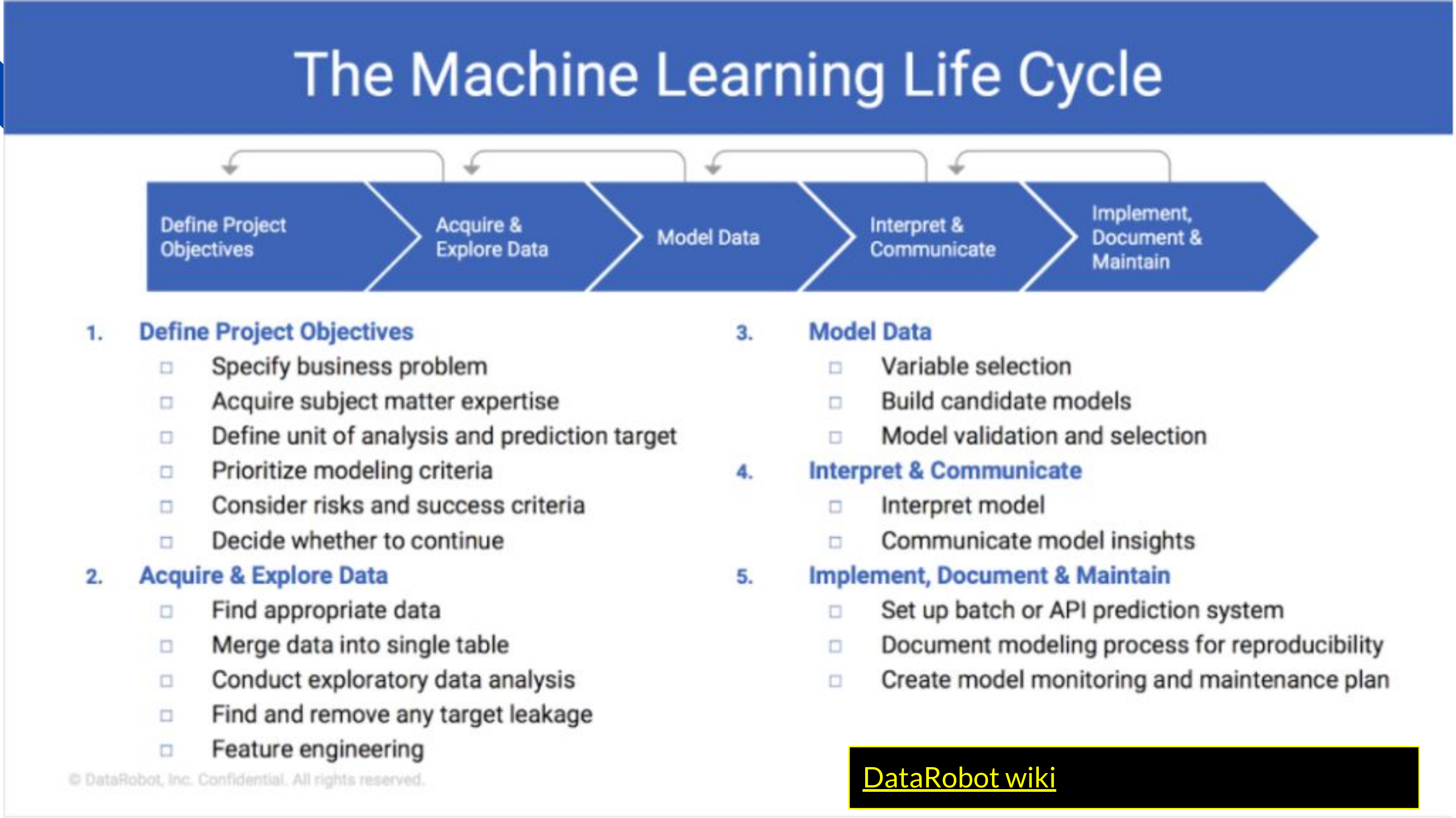


Prevent Target Leakage to Prevent Bad Parallel Universes

Time Travel from *Back to the Future Part II*

→ Target Leakage

The Machine Learning Life Cycle

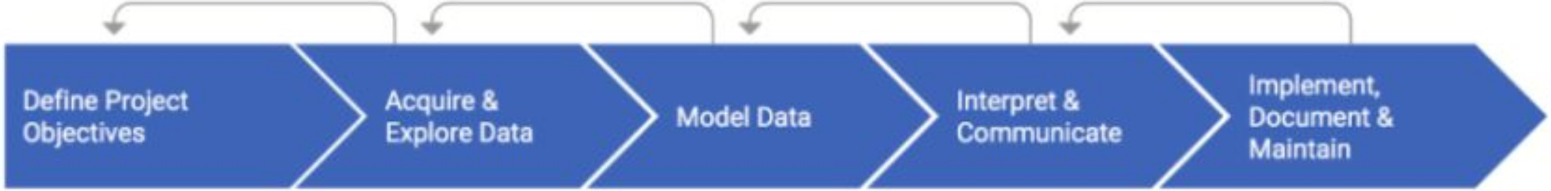

Define Project Objectives
Acquire & Explore Data
Model Data
Interpret & Communicate
Implement, Document & Maintain

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, Inc. Confidential. All rights reserved.

DataRobot wiki

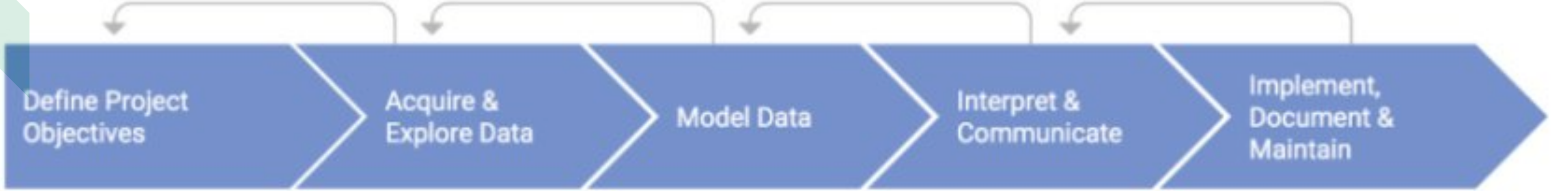
The Machine Learning Life Cycle


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, Inc. Confidential. All rights reserved.

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

ls

selection

insights

ntain

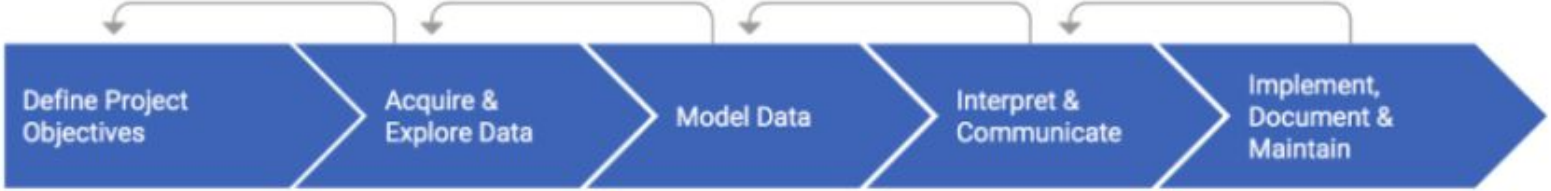
ediction system

rocess for reproducibility

ing and maintenance plan

© DataRobot, Inc. Confidential. All rights reserved.

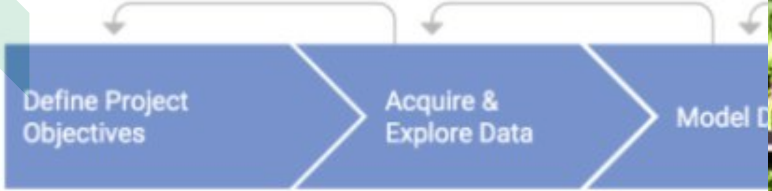
The Machine Learning Life Cycle


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, Inc. Confidential. All rights reserved.

The Machine Learning Life Cycle



Define Project Objectives

Acquire & Explore Data

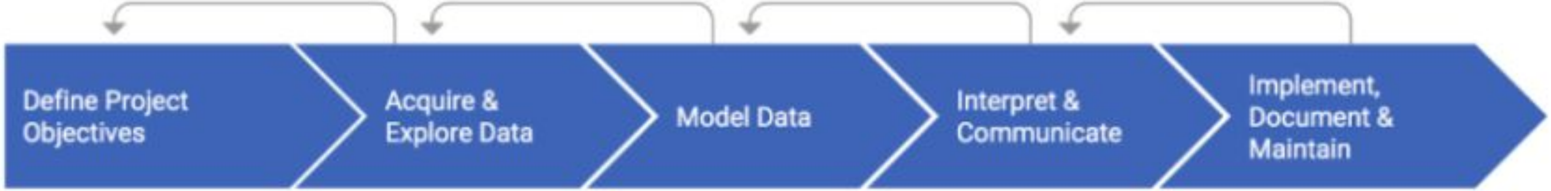
Model D

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

© DataRobot, Inc. Confidential. All rights reserved.

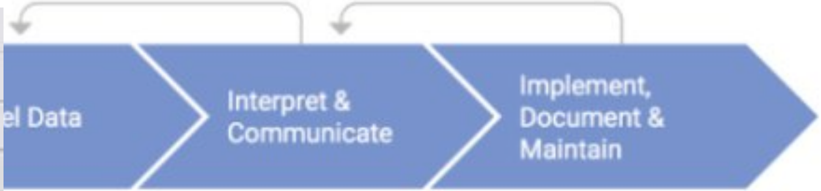
The Machine Learning Life Cycle


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, Inc. Confidential. All rights reserved.

The Machine Learning Life Cycle



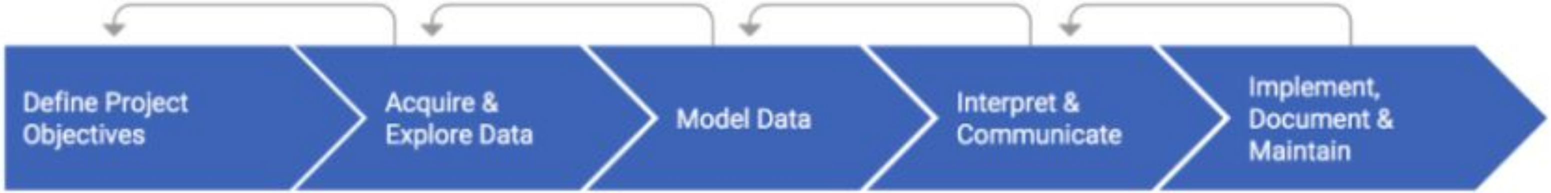
- Decide whether to continue

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

© DataRobot, Inc. Confidential. All rights reserved.

The Machine Learning Life Cycle

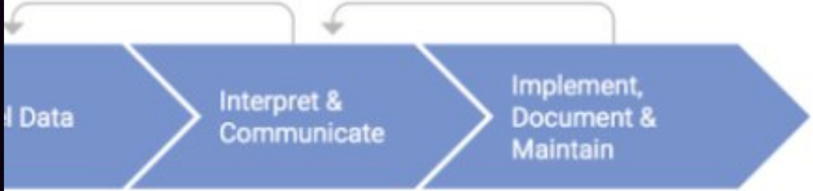

Define Project Objectives
Acquire & Explore Data
Model Data
Interpret & Communicate
Implement, Document & Maintain

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, Inc. Confidential. All rights reserved.

The Machine Learning Life Cycle

GARFIELD
BOOMERANG

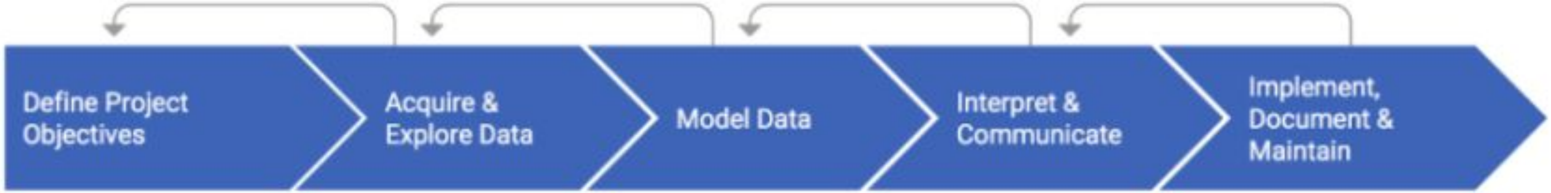

l Data
Interpret & Communicate
Implement, Document & Maintain

- 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, Inc. Confidential. All rights reserved.

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, Inc. Confidential. All rights reserved.

The Machine Learning Life Cycle

BASICS
MAIN DISHES
DESSERTS
family
FAVORITE
RECIPES

el Data
Interpret & Communicate
Implement, Document & Maintain

- 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

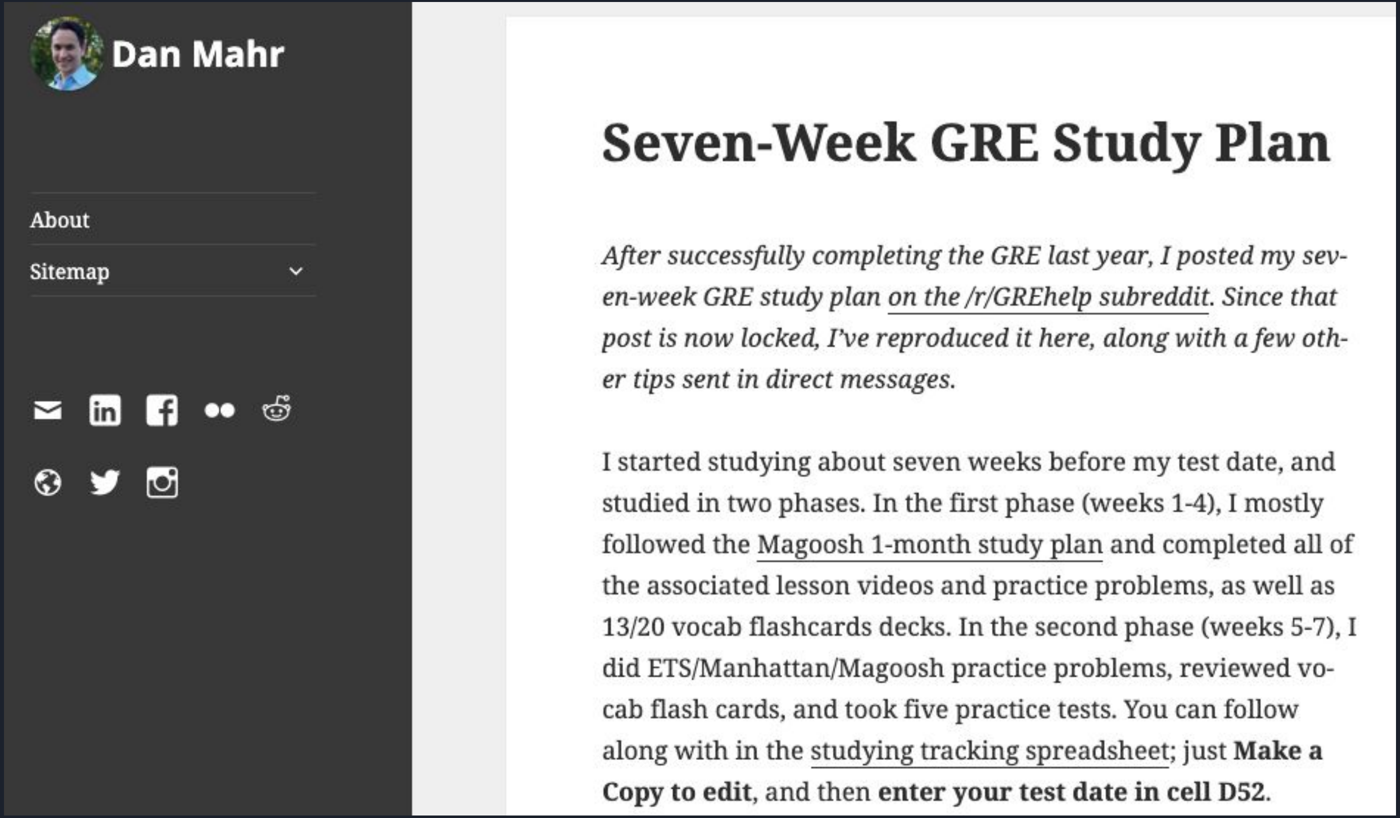
© DataRobot, Inc. Confidential. All rights reserved.

Cooking or Baking a Meal

→ Machine Learning Life Cycle

Studying and taking exams

One successful strategy



Seven-Week GRE Study Plan

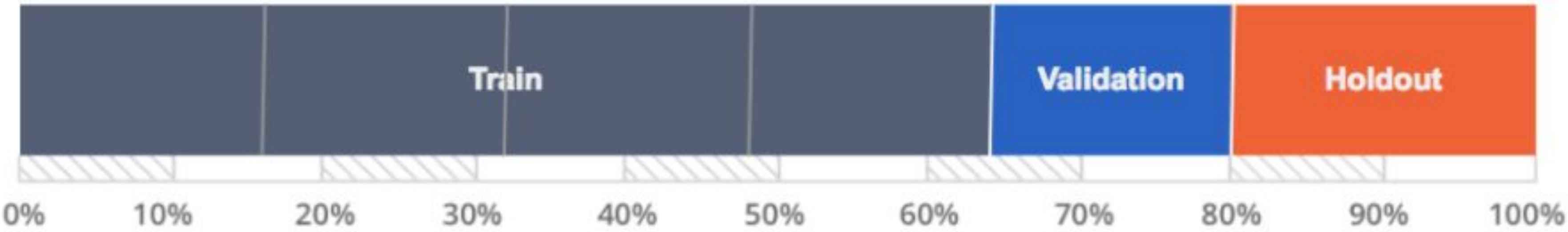
After successfully completing the GRE last year, I posted my seven-week GRE study plan on the /r/GREhelp subreddit. Since that post is now locked, I've reproduced it here, along with a few other tips sent in direct messages.

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?

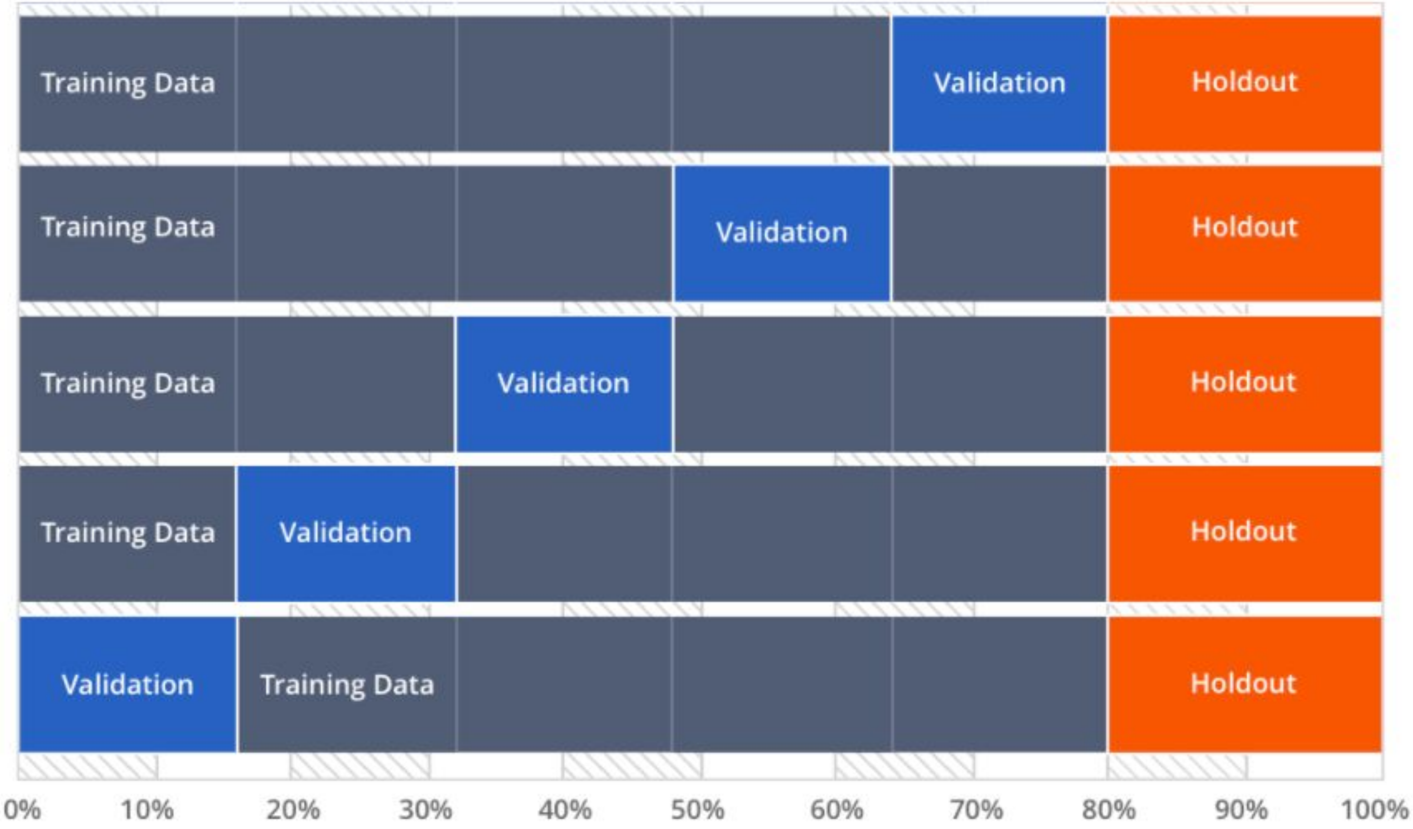
OVERFITTING

Training / Validation / Holdout


Train
Validation
Holdout
0%
10%
20%
30%
40%
50%
60%
70%
80%
90%
100%

DataRobot wiki

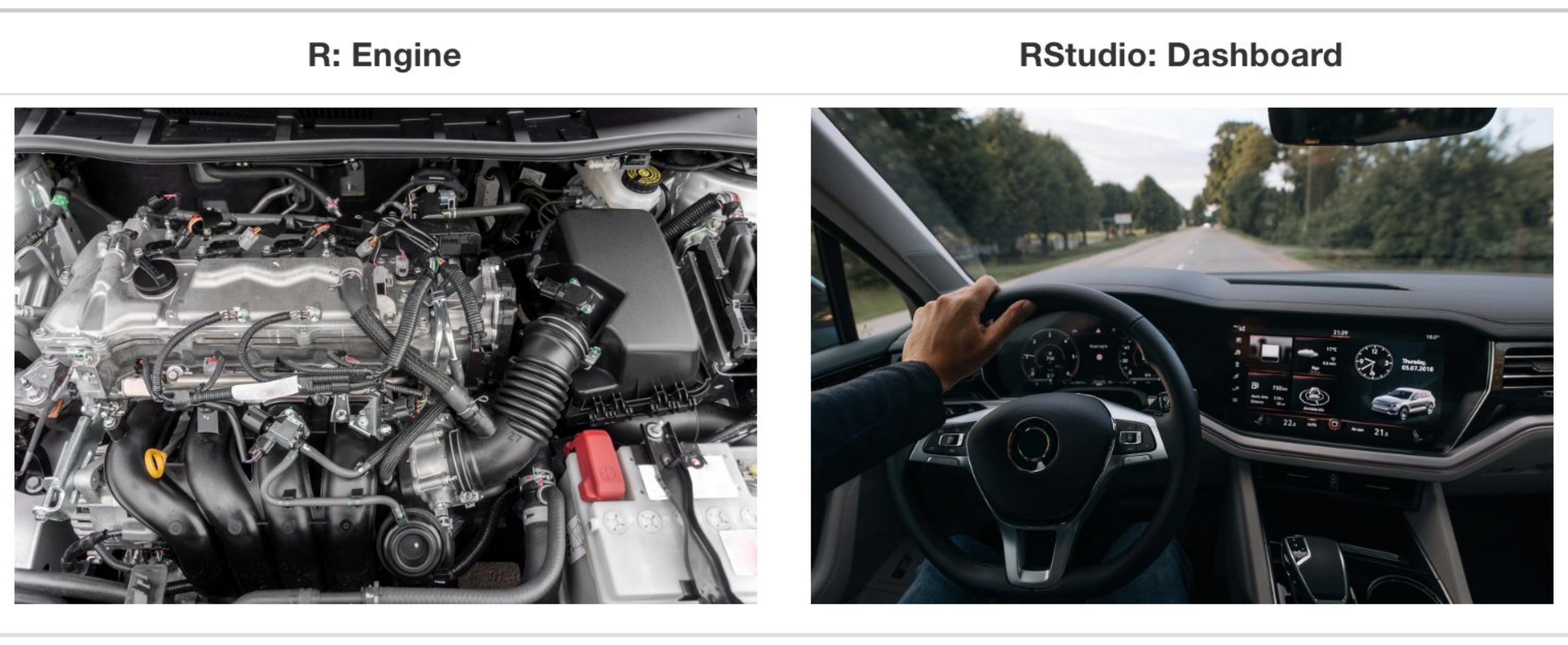
Cross-validation


Training Data
Validation
Holdout
Training Data
Validation
Holdout
Training Data
Validation
Holdout
Training Data
Validation
Holdout
Validation
Training Data
Holdout
0%
10%
20%
30%
40%
50%
60%
70%
80%
90%
100%

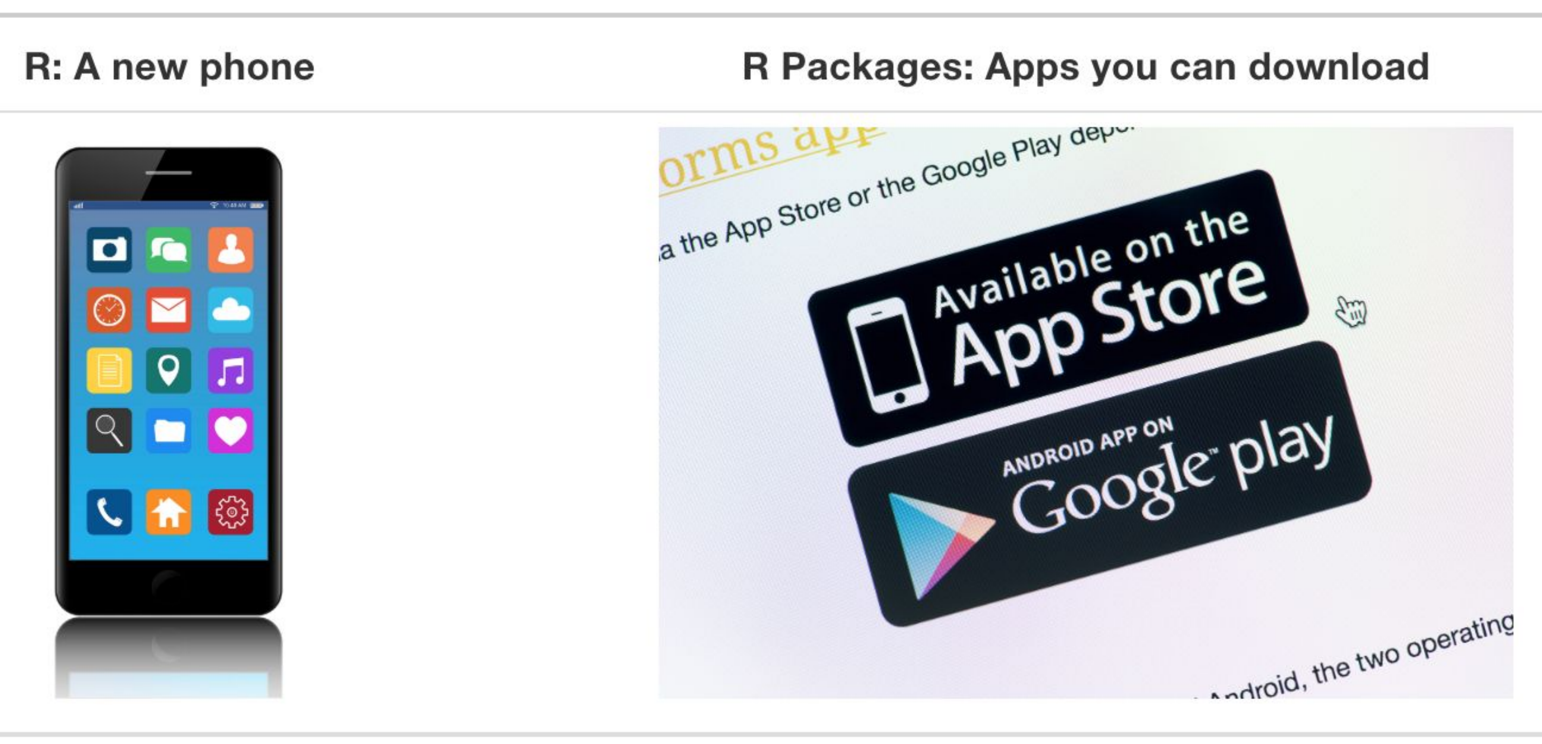
Studying for an Exam

→ Cross-validation

R and RStudio


R: Engine
RStudio: Dashboard

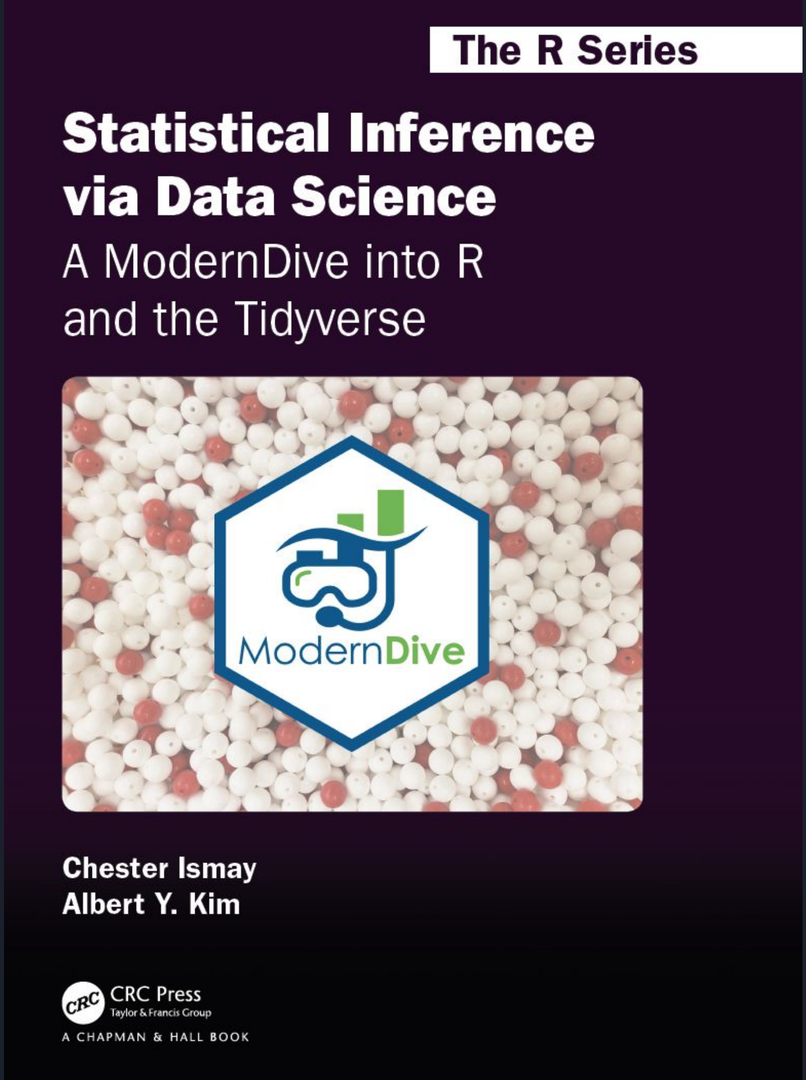
R packages



R: A new phone	R Packages: Apps you can download

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

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!

Any questions?
NBC

Dr. Chester Ismay

chester.ismay@datarobot.com

https://chester.rbind.io

Slides available at

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