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


1950–2000


Hilldale


ALMANAC
1950 - 2000

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: 50-64. Details Page C2.

Hill Valley Telegraph

Index
Obituaries C 8
Classified C 3
Comics B 5
Editorials A10
Financial C10
Sports D 1
Movies B 1

Vol. XVII, No. 32 | COMPLETE NEWS SERVICE A. P. N. S. NEWS PHOTOS | THURSDAY, APRIL 16, 1959 | THE NEWSPAPER THAT SERVES THE PEOPLE | 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 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 Editions

BECAUSE some of the sparkle and glitter was taken out of a "Vote for Governor" sign on the highway near Oakhurst a charge of malicious mischief was filed in common pleas court yesterday against Fred Biggins, 17-year-old youth, according to complaint of W. L. Hughes, 1042 North Ellis avenue.

The gubernatorial signboard was studded with glass gems which reflects automobile lights at night giving the sign an illuminated effect. Sheriff Ray Hyatt said that Biggins removed 50 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 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 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.

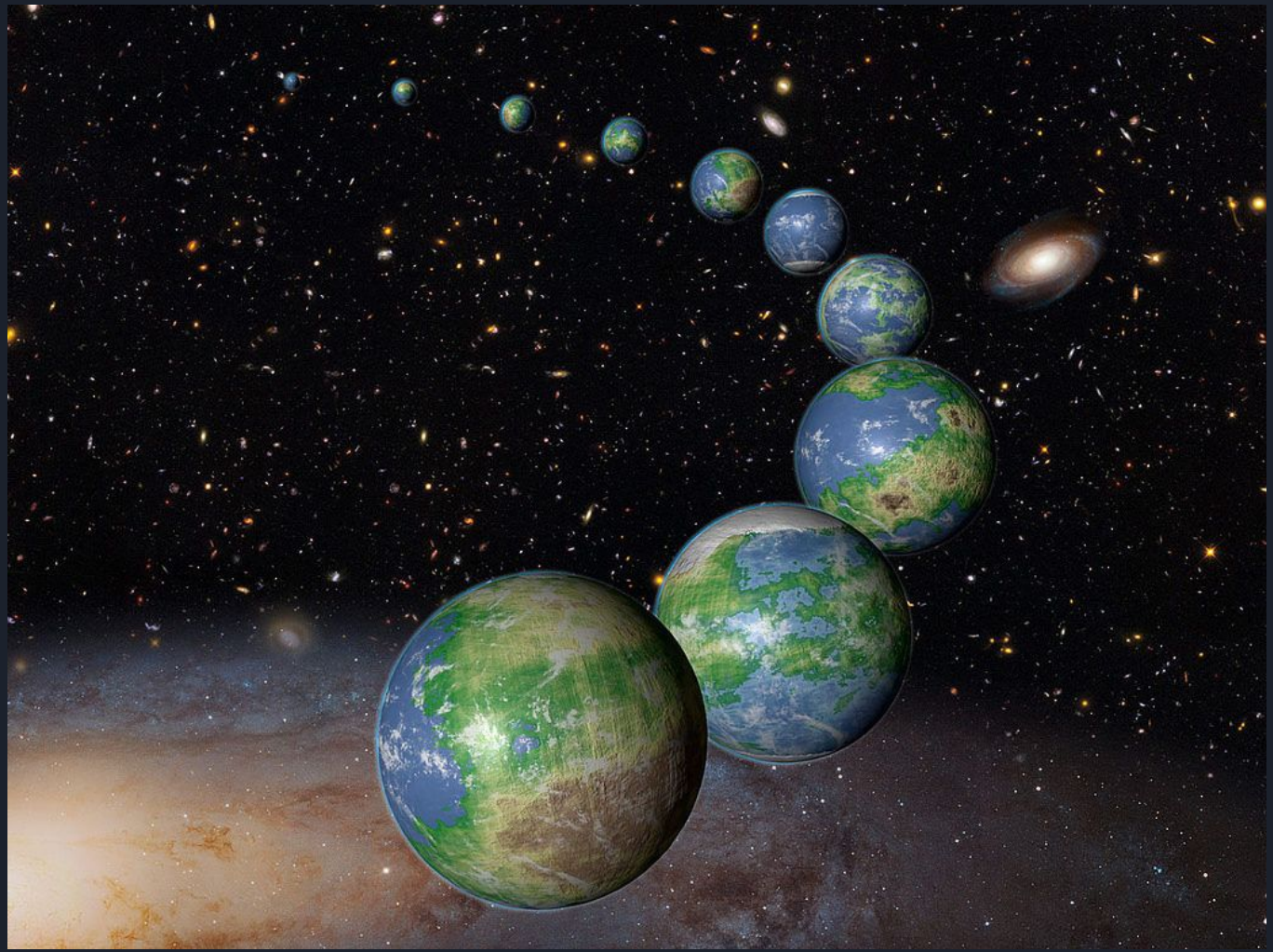
Many persons feel at this stage that some legal action is

BIFF WINS AGAIN

Khruschev Offers Dates for Summit


Buff Tanner's
PLEASURE PARADISE
CASINO

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

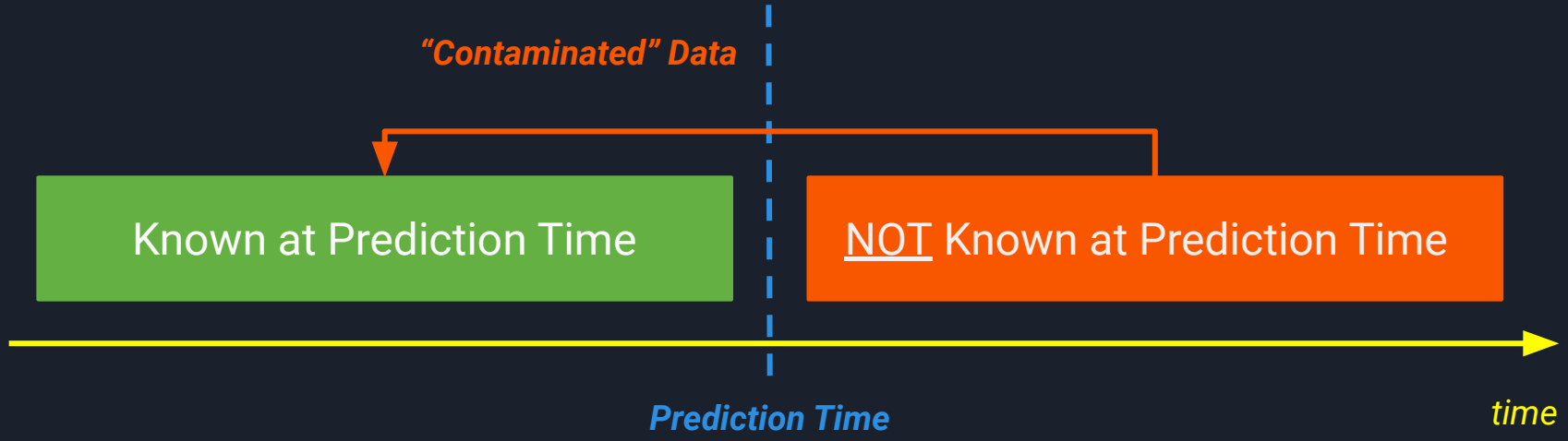
New Data +

= Predictions and Decisions

Model is USED

(model is “static”)

Target Leakage

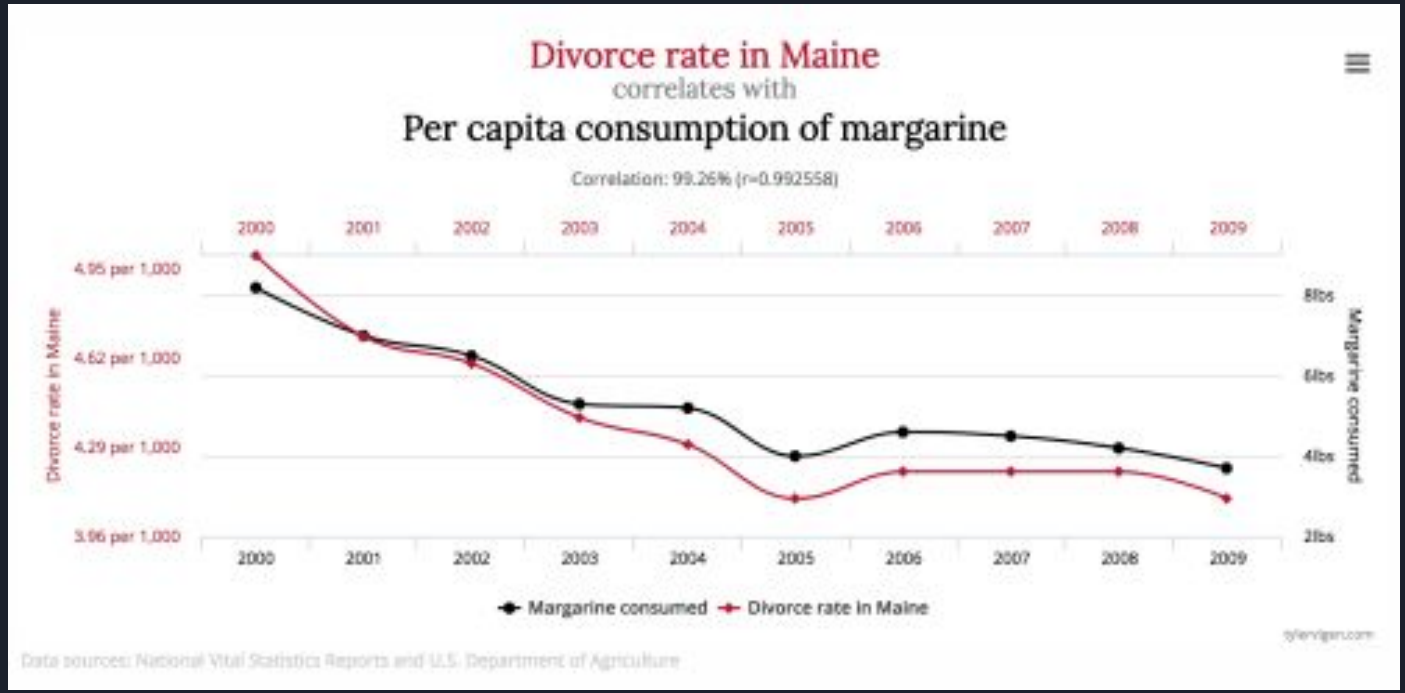


Predicting Recidivism

Preventing Target Leakage

[How to avoid the data leakage? blogpost]

Preventing Target Leakage

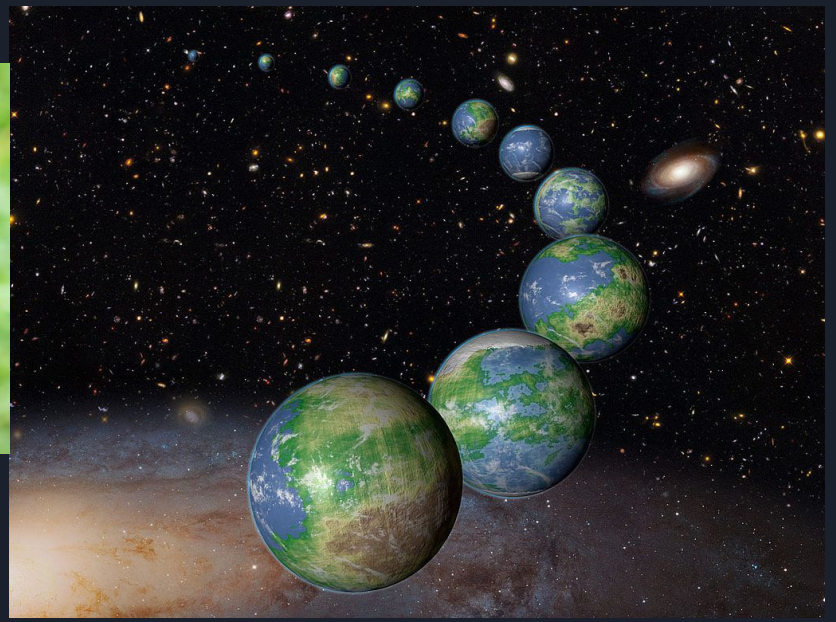


Beware of Data Leakage blogpost

Further References

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

Prevent Target Leakage to Prevent Bad Parallel Universes



Prevent Target Leakage to Prevent Bad Parallel Universes

Time Travel from *Back to the Future Part II*

→ Target Leakage

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

DataRobot wiki

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

ls
selection
nsights
ntain
ediction system
rocess 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

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

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

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

- 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

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

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

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

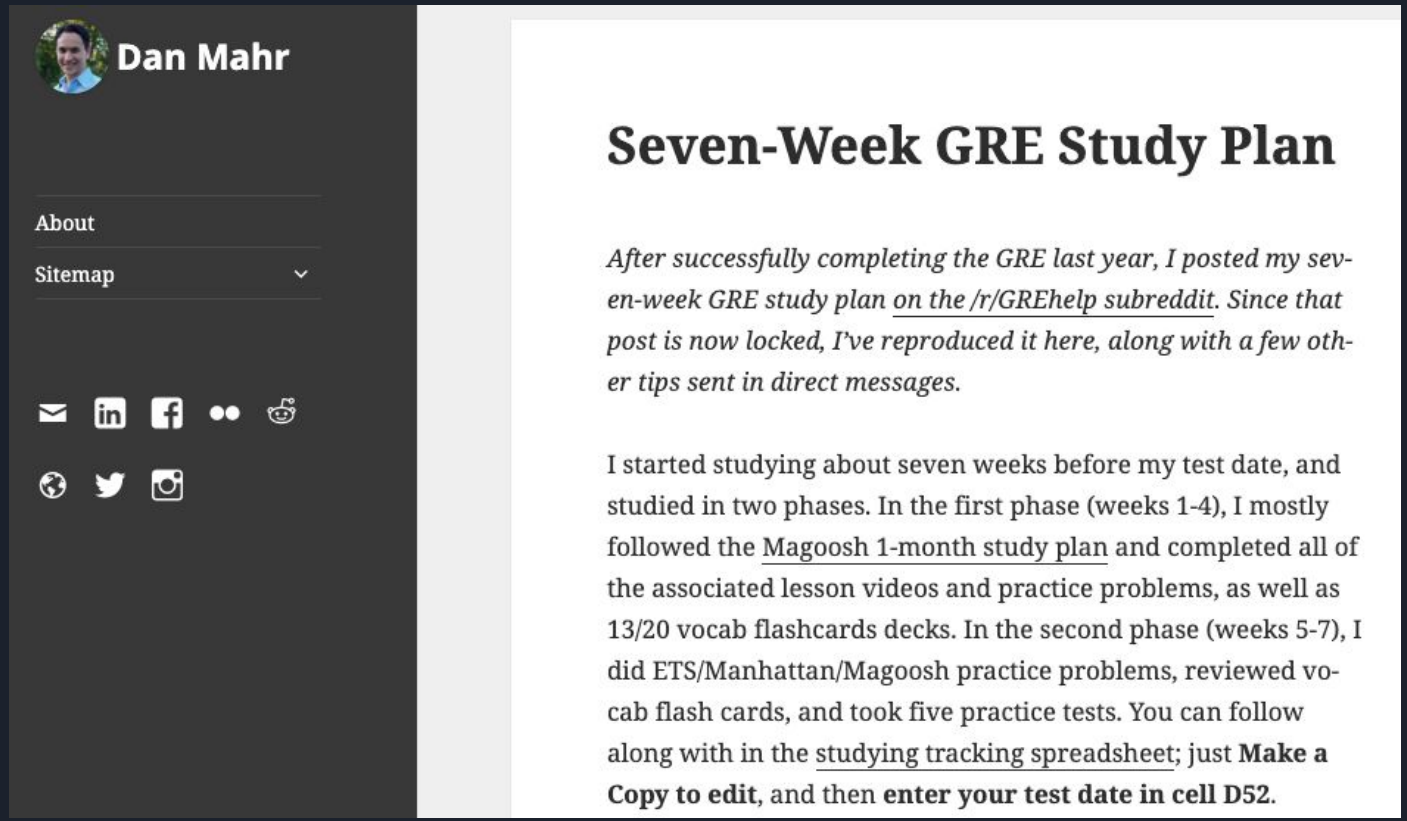
- ☐ Find and remove any target leakage
- ☐ Feature engineering

Cooking or Baking a Meal

→ Machine Learning Life Cycle

Studying and taking exams

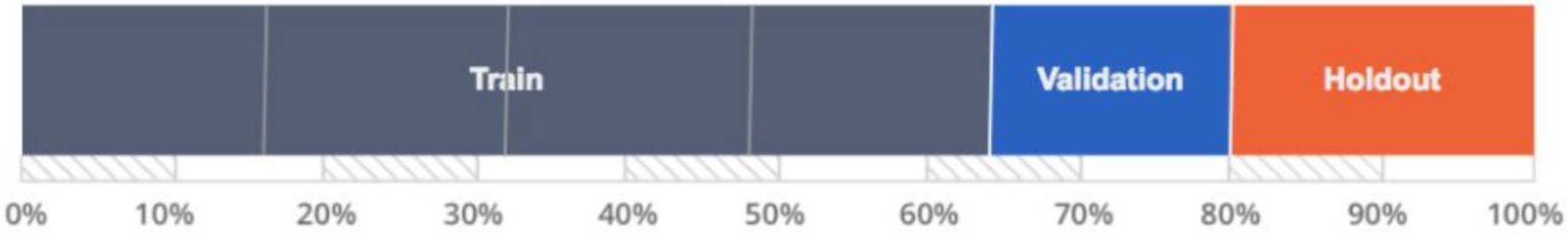
One successful strategy



How can this relate to machine learning?

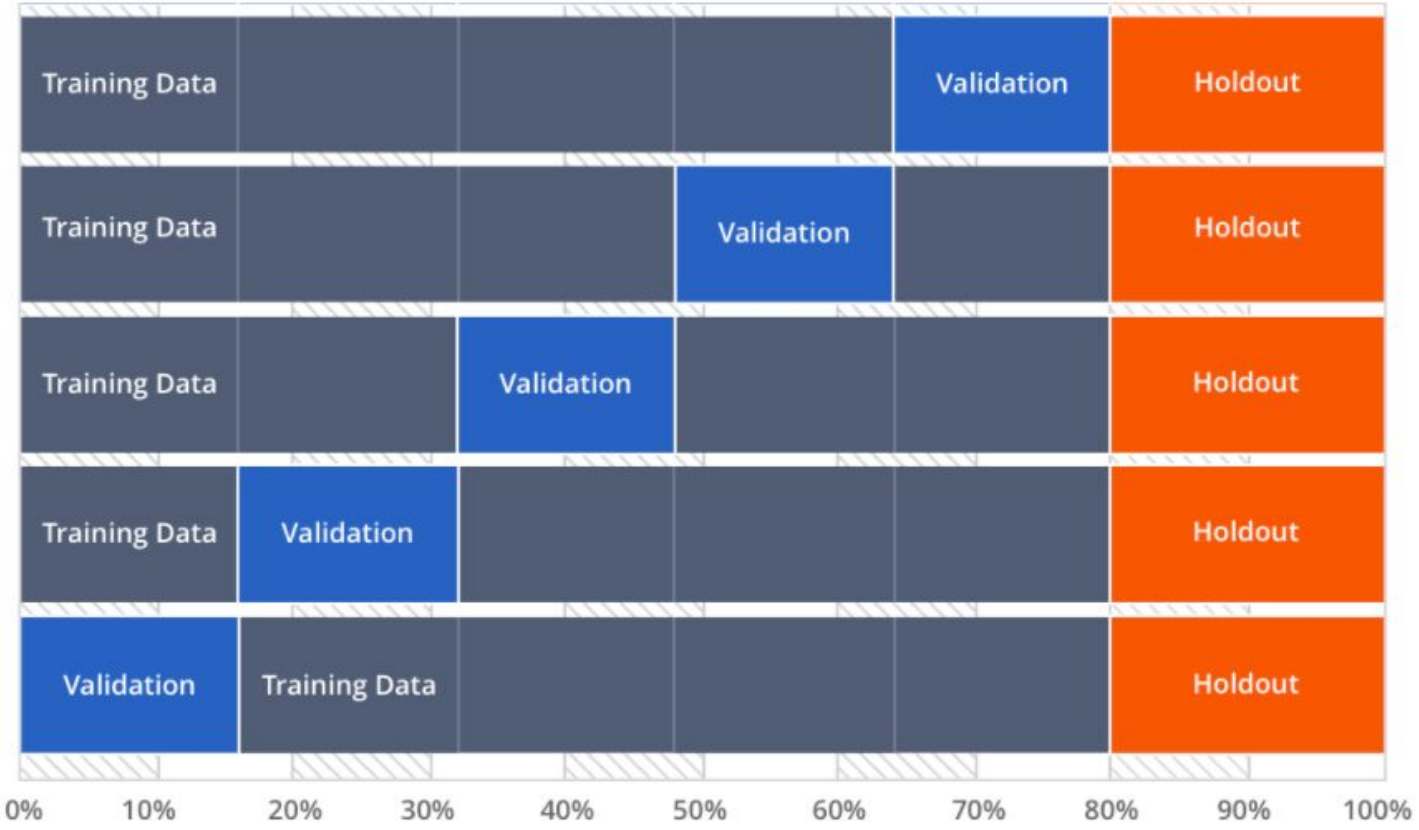

OVERFITTING

Training / Validation / Holdout



DataRobot wiki

Cross-validation



Studying for an Exam

→ Cross-validation

R and RStudio

R: Engine

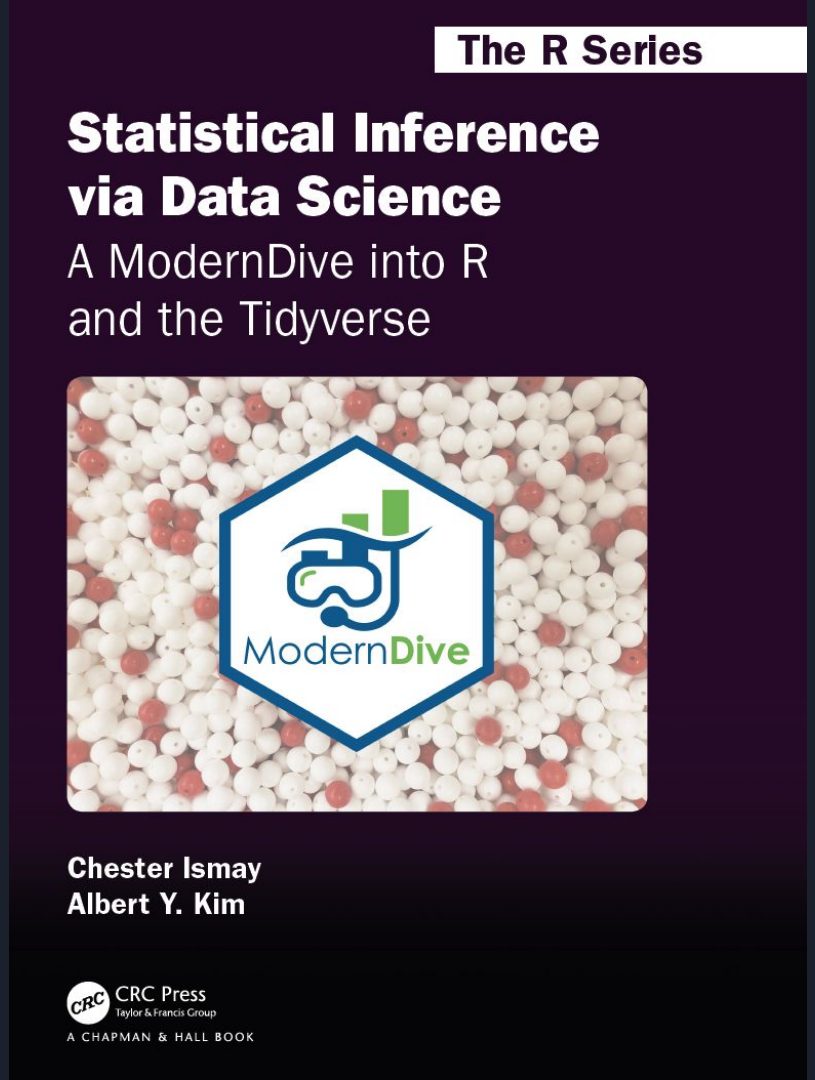
RStudio: Dashboard

R packages

R: A new phone

R Packages: Apps you can download

ModernDive.com



Thank you!

Dr. Chester Ismay

chester.ismay@datarobot.com

https://chester.rbind.io

Slides available at

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