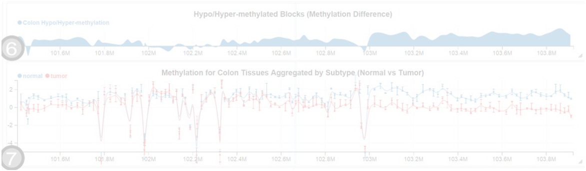


CMSC320 Introduction to Data Science: Course Introduction and Overview

Héctor Corrada Bravo

University of Maryland, College Park, USA

CMSC320: 2020-01-27

Business First

Course Webpage: http://bit.ly/hcb-ids

What is Data Science?

Data science encapsulates the interdisciplinary activities required to create data-centric artifacts and applications that address specific scientific, socio-political, business, or other questions.

Data

Observable units of information measured or captured from activity of people, places and things.

Data

Observable units of information measured or captured from activity of people, places and things.

Specific Questions

Seeking to understand a phenomenon, natural, social or other

Data

Observable units of information measured or captured from activity of people, places and things.

Specific Questions

Seeking to understand a phenomenon, natural, social or other

Can we formulate specific questions for which an answer posed in terms of patterns observed, tested and or modeled in data is appropriate.

Interdisciplinary Activities

- Formulating a question, assessing the appropriateness of the data and findings used to find an answer require understanding of the specific subject area.

Interdisciplinary Activities

- Formulating a question, assessing the appropriateness of the data and findings used to find an answer require understanding of the specific subject area.
- Deciding on the appropriateness of models and inferences made from models based on the data at hand requires understanding of statistical and computational methods.

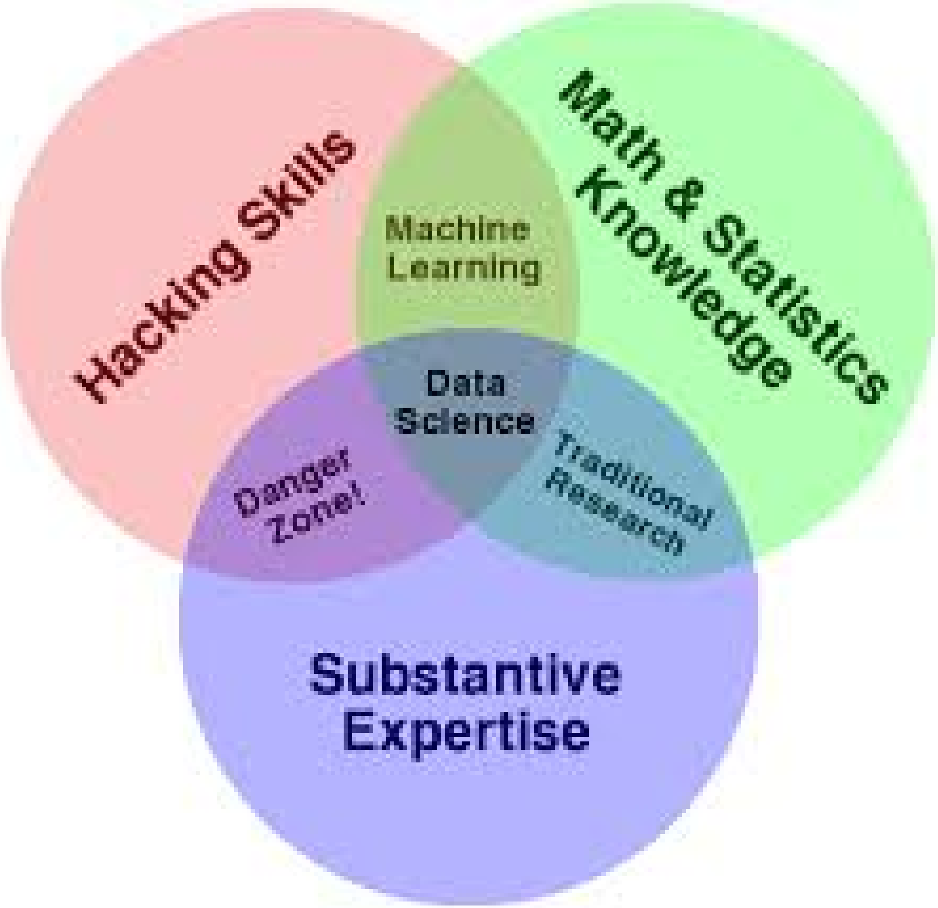
Data-centric artifacts and applications

- Answers to questions derived from data are usually shared and published in meaningful, succint but sufficient, reproducible artifacts (papers, books, movies, comics).

Data-centric artifacts and applications

- Answers to questions derived from data are usually shared and published in meaningful, succint but sufficient, reproducible artifacts (papers, books, movies, comics).
- Going a step further, interactive applications that let others explore data, models and inferences are great.

Data Science



Why Data Science?

The granularity, size and accessibility data, comprising both physical, social, commercial and political spheres has exploded in the last decade or more.

> I keep saying that the sexy job in the next 10 years will be statisticians”
>
> Hal Varian, Chief Economist at Google (http://www.nytimes.com/2009/08/06/technology/06stats.html?_r=0)

Why Data Science?

> “The ability to take data—to be able to understand it, to process it, to extract value from it, to visualize it, to communicate it—that’s going to be a hugely important skill in the next decades, not only at the professional level but even at the educational level for elementary school kids, for high school kids, for college kids.”
>
> Hal Varian (http://www.mckinsey.com/insights/innovation/hal_varian_on_how_the_web_

Why Data Science?

> “Because now we really do have essentially free and ubiquitous data. So the complimentary scarce factor is the ability to understand that data and extract value from it.”
>
> Hal Varian (http://www.mckinsey.com/insights/innovation/hal_varian_on_how_the_web_

Data Science in Society

Large amounts of data produced across many spheres of human activity,

Data Science in Society

Large amounts of data produced across many spheres of human activity,

Many societal questions may be addressed by characterizing patterns in data.

Data Science in Society

This can range from unproblematic questions:

- how to dissect a large creative corpora, say music, literature, based on raw characteristics of those works, text, sound and image.

Data Science in Society

This can range from unproblematic questions:

- how to dissect a large creative corpora, say music, literature, based on raw characteristics of those works, text, sound and image.

To more problematic questions

- analysis of intent, understanding, appreciation and valuation of these creative corpora.

Data Science in Society

Issues of fairness and transparency in the current era of big data are especially problematic.

- Is data collected representative of population for which inferences are drawn?

Data Science in Society

Issues of fairness and transparency in the current era of big data are especially problematic.

- Is data collected representative of population for which inferences are drawn?

- Are methods employed learning latent unfair factors from ostensibly fair data?

Data Science in Society

Issues of fairness and transparency in the current era of big data are especially problematic.

- Is data collected representative of population for which inferences are drawn?
- Are methods employed learning latent unfair factors from ostensibly fair data?
- These are issues that the research community is now starting to address.

Data Science in Society

In all settings, issues of ethical collection of data, application of models, and deployment of data-centric artifacts are essential to grapple with.

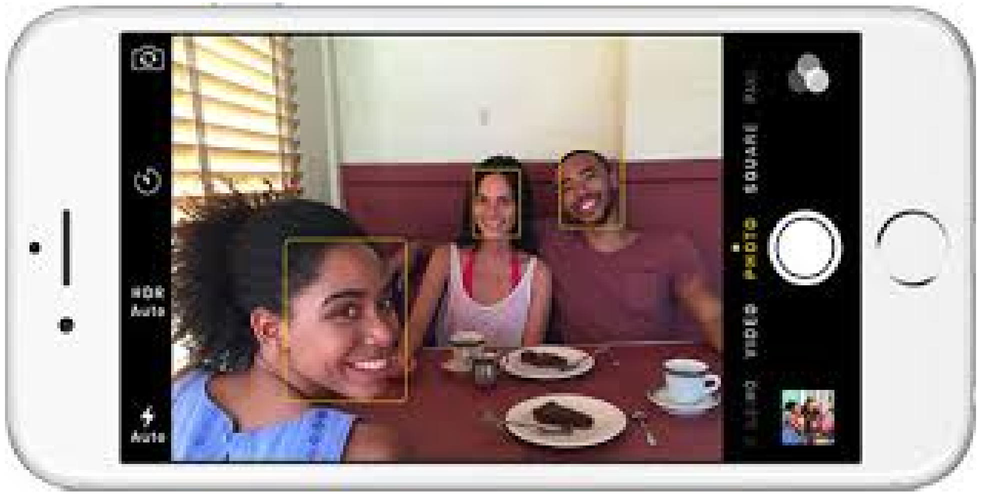
Issues of privacy are equally important.

Data Science in Society: Machine Learning

Self driving cars make use of ML models for sensor processing.

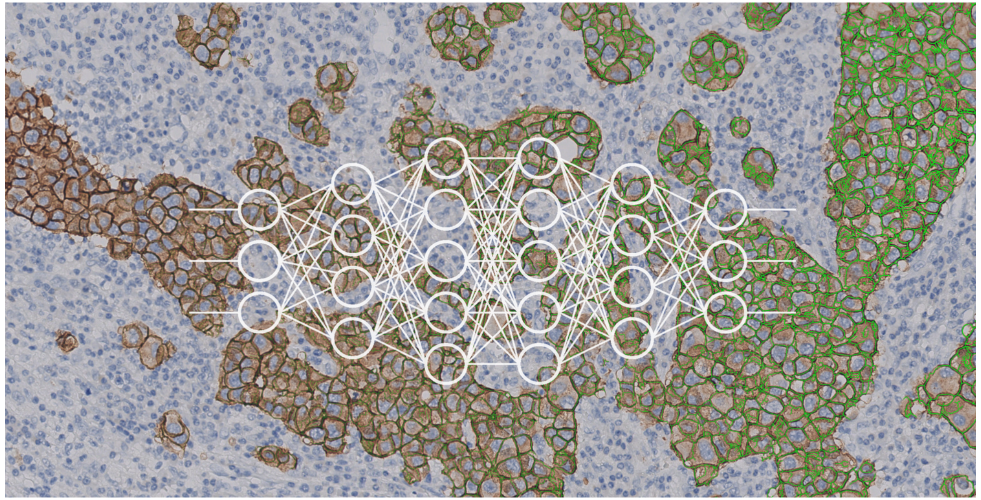
Data Science in Society: Machine Learning

Image recognition software uses ML to identify individuals in photos.



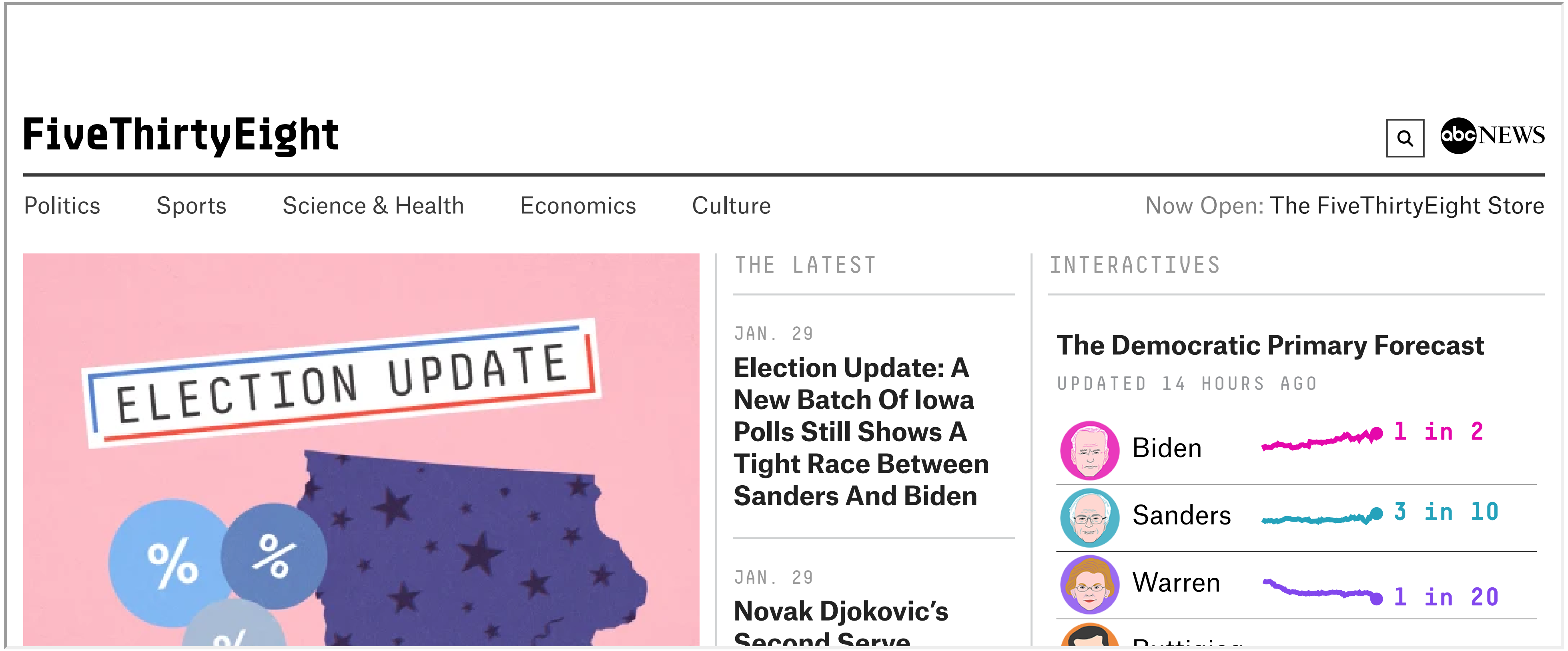
Data Science in Society: Machine Learning

ML models have been applied to medical imaging to yield expert-level prognosis.



Data Science in Society: Data Journalism

http:://fivethirtyeight.com



THE LATEST

JAN. 29

Election Update: A New Batch Of Iowa Polls Still Shows A Tight Race Between Sanders And Biden

JAN. 29

Novak Djokovic's Second Serve

INTERACTIVES

The Democratic Primary Forecast

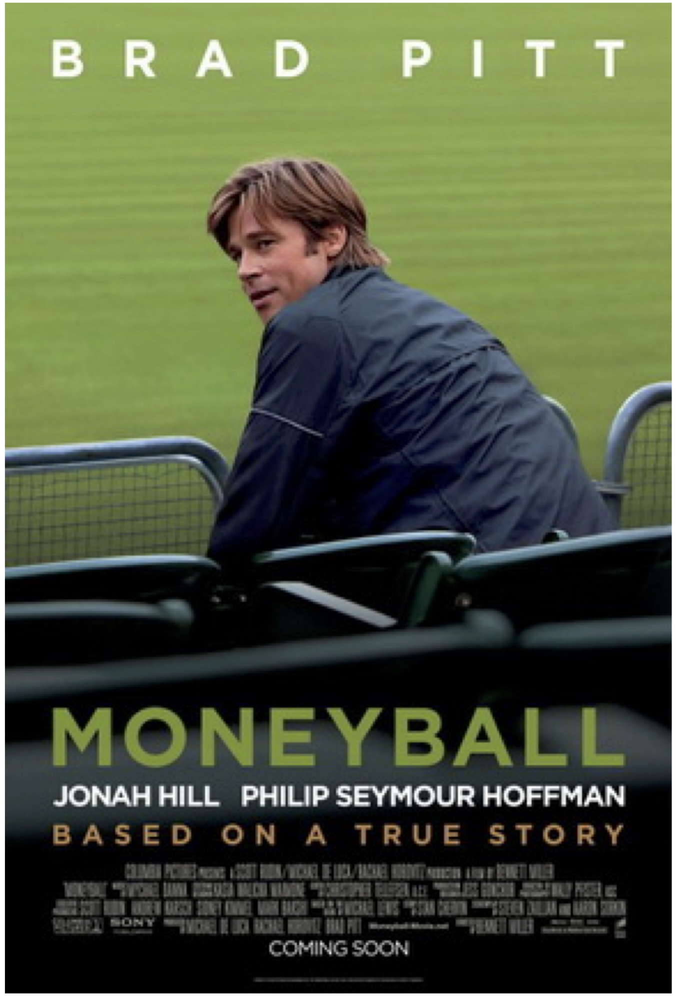
UPDATED 14 HOURS AGO

Biden	1 in 2
Sanders	3 in 10
Warren	1 in 20
Buttigieg	

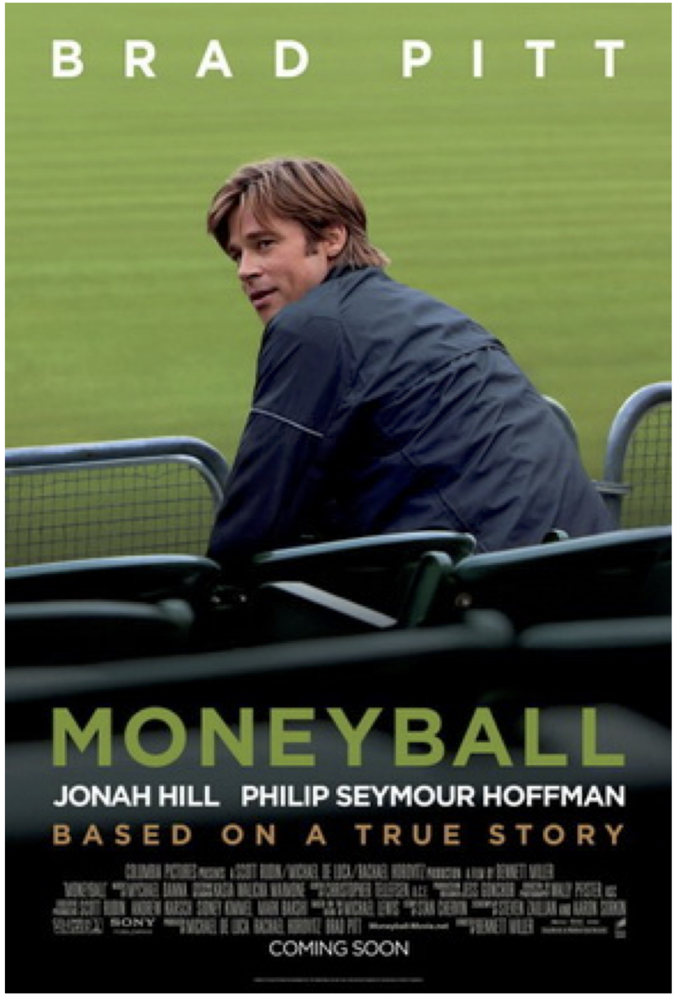
Data Science in Society: Data Journalism

http://www.nytimes.com/section/upshot

Data Science in Society: Business



Data Science in Society: Business

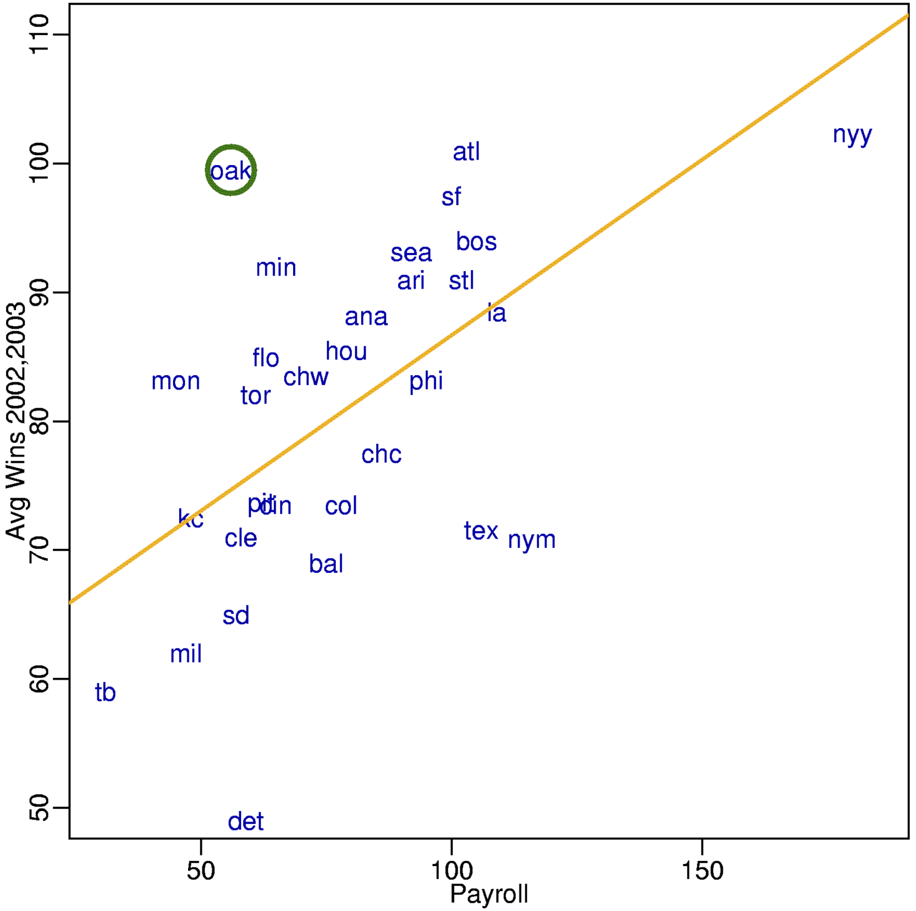


Actual

Hollywood

Data Science in Society: Business

In the early 2000's the Oakland A's were winning as much as teams with much bigger payrolls by evaluating players using data differently than other teams.



Data Science in Society: Entertainment

The story of the Netflix Prize

In October 2006 Netflix announced a prize around their movie recommendation engine.

Data Science in Society: Entertainment

The story of the Netflix Prize

In October 2006 Netflix announced a prize around their movie recommendation engine.

Supervised Machine Learning (ML) task:

- Dataset of users and their ratings, (1,2,3,4 or 5 stars), of movies they have rated.
- Build an ML model that given predicts a specific user's rating to a movie they have not rated.

Data Science in Society: Entertainment

The story of the Netflix Prize

In October 2006 Netflix announced a prize around their movie recommendation engine.

Supervised Machine Learning (ML) task:

- Dataset of users and their ratings, (1,2,3,4 or 5 stars), of movies they have rated.
- Build an ML model that given predicts a specific user's rating to a movie they have not rated.

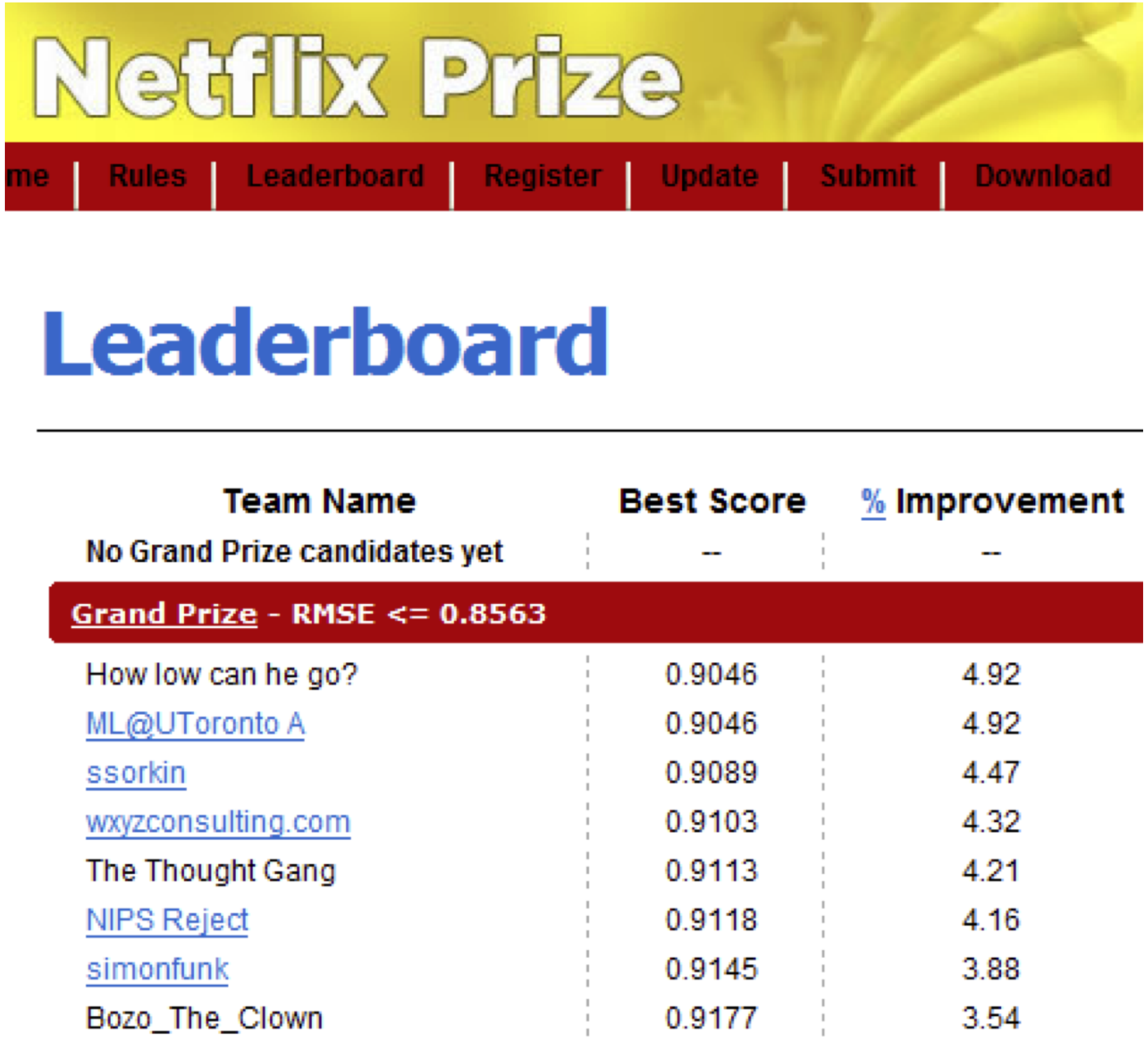
They can recommend movies to users if they predict high rating.

Data Science in Society: Entertainment

Netflix would award $1M for the first ML system that provided a 10% improvement to their existing system

Data Science in Society: Entertainment

Existing system had a 0.9514 mean squared error


Netflix Prize
me | Rules | Leaderboard | Register | Update | Submit | Download

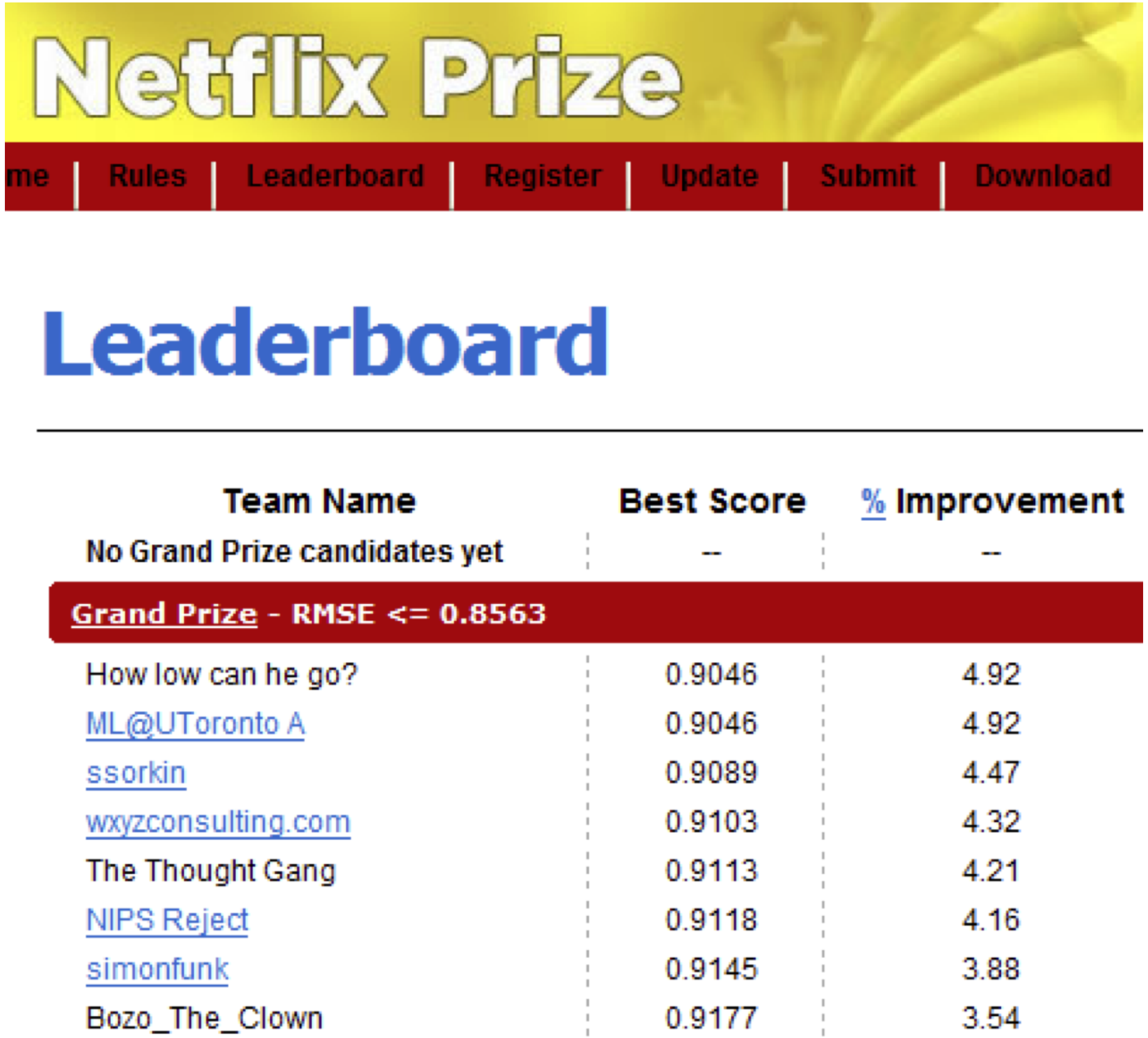
Leaderboard

Team Name	Best Score	% Improvement
No Grand Prize candidates yet	--	--
Grand Prize - RMSE <= 0.8563		
How low can he go?	0.9046	4.92
ML@UToronto A	0.9046	4.92
ssorkin	0.9089	4.47
wxyzconsulting.com	0.9103	4.32
The Thought Gang	0.9113	4.21
NIPS Reject	0.9118	4.16
simonfunk	0.9145	3.88
Bozo_The_Clown	0.9177	3.54

Data Science in Society: Entertainment

Within three weeks, at least 40 teams had improved upon the existing Netflix system.

The top teams were showing improvement over 5%.


Netflix Prize
me | Rules | Leaderboard | Register | Update | Submit | Download

Leaderboard

Team Name	Best Score	% Improvement
No Grand Prize candidates yet	--	--
Grand Prize - RMSE <= 0.8563		
How low can he go?	0.9046	4.92
ML@UToronto A	0.9046	4.92
ssorkin	0.9089	4.47
wxyzconsulting.com	0.9103	4.32
The Thought Gang	0.9113	4.21
NIPS Reject	0.9118	4.16
simonfunk	0.9145	3.88
Bozo_The_Clown	0.9177	3.54

Course organization

This course will cover basics of how to represent, model and communicate about data and data analyses using the R and/or Python environments for Data Science

Course organization

This course will cover basics of how to represent, model and communicate about data and data analyses using the R and/or Python environments for Data Science

- Area 0: Tools and skills

Course organization

This course will cover basics of how to represent, model and communicate about data and data analyses using the R and/or Python environments for Data Science

- Area 0: Tools and skills
- Area 1: Data types and operations

Course organization

This course will cover basics of how to represent, model and communicate about data and data analyses using the R and/or Python environments for Data Science

- Area 0: Tools and skills
- Area 1: Data types and operations
- Area 2: Data wrangling

Course organization

This course will cover basics of how to represent, model and communicate about data and data analyses using the R and/or Python environments for Data Science

- Area 0: Tools and skills
- Area 1: Data types and operations
- Area 2: Data wrangling
- Area 3: Modeling

Course organization

This course will cover basics of how to represent, model and communicate about data and data analyses using the R and/or Python environments for Data Science

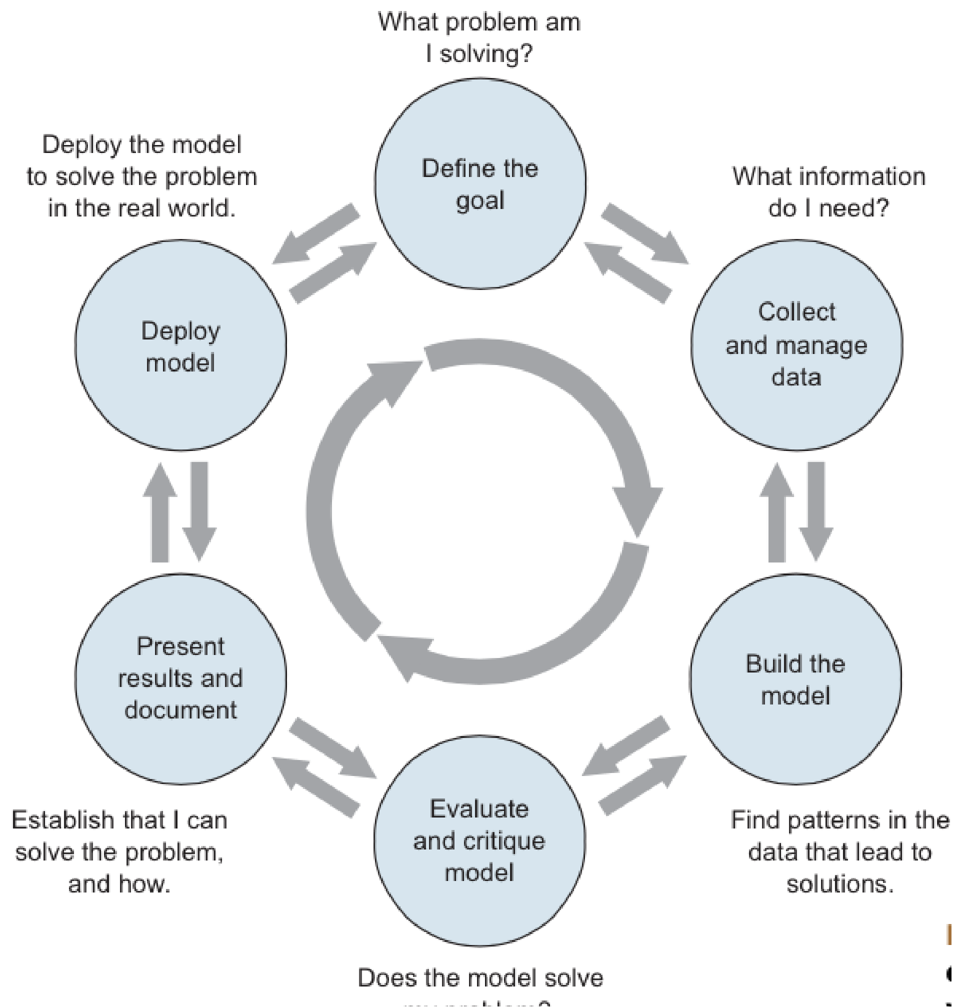
- Area 0: Tools and skills
- Area 1: Data types and operations
- Area 2: Data wrangling
- Area 3: Modeling
- Area 4: Applications

Course organization

This course will cover basics of how to represent, model and communicate about data and data analyses using the R and/or Python environments for Data Science

- Area 0: Tools and skills
- Area 1: Data types and operations
- Area 2: Data wrangling
- Area 3: Modeling
- Area 4: Applications
- Area 5: Communication

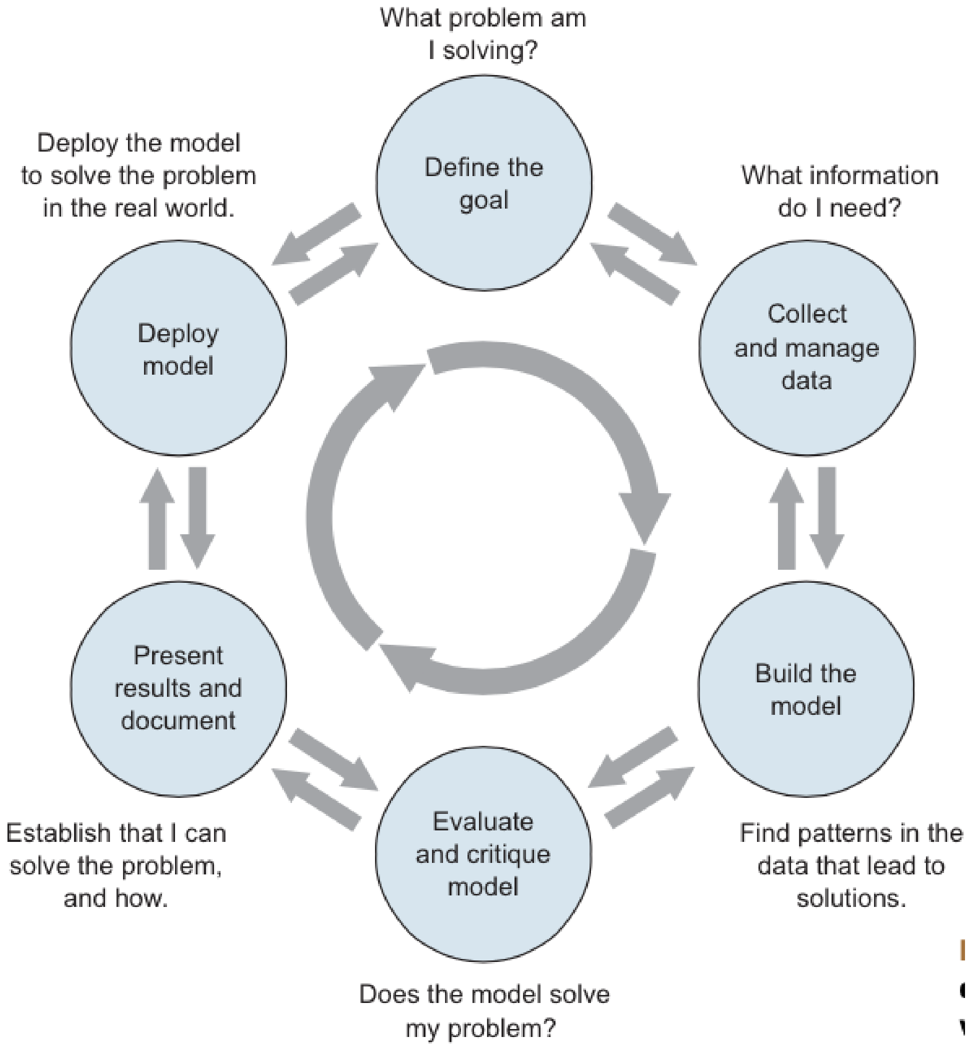
General Workflow


What problem am I solving?
Define the goal
What information do I need?
Collect and manage data
Build the model
Find patterns in the data that lead to solutions.
Evaluate and critique model
Does the model solve
Present results and document
Establish that I can solve the problem, and how.
Deploy model
Deploy the model to solve the problem in the real world.

Zumel and Mount

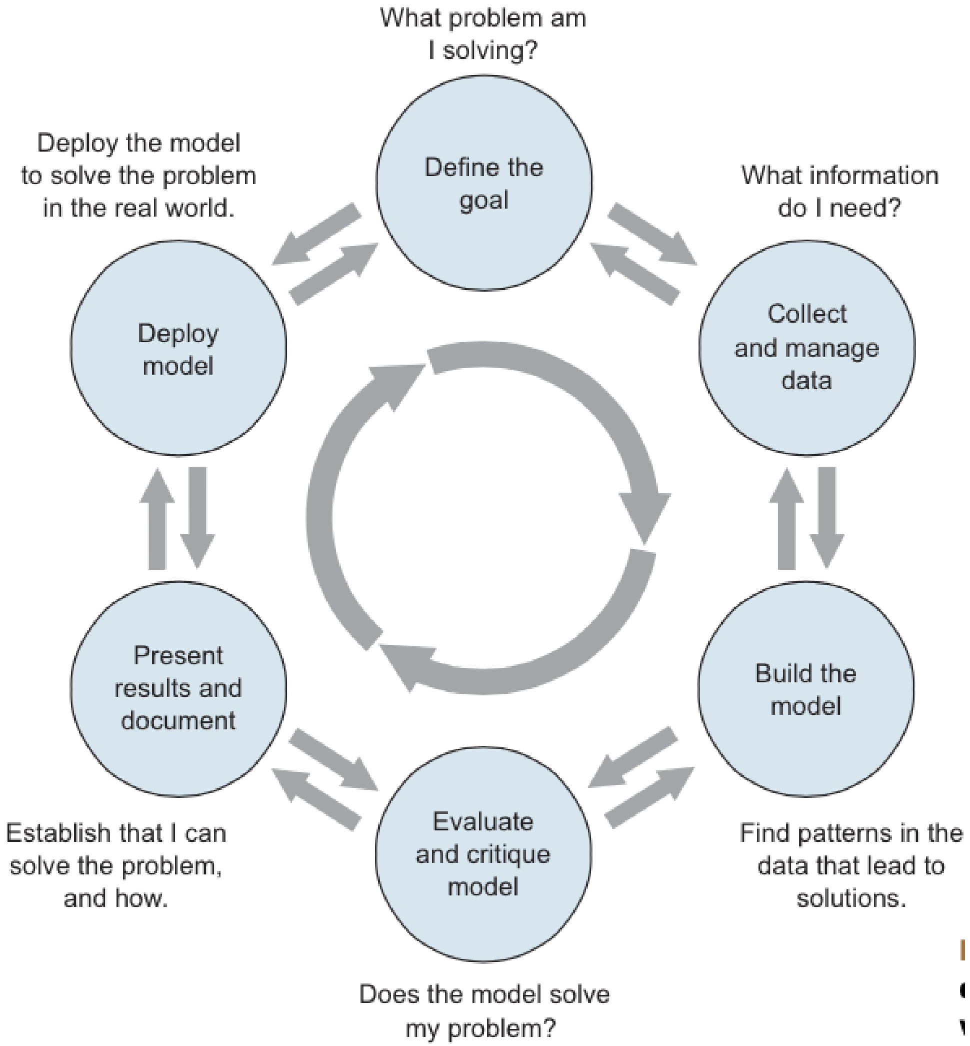
Defining the goal

- What is the question/problem?
- Who wants to answer/solve it?
- What do they know/do now?
- How well can we expect to answer/solve it?
- How well do they want us to answer/solve it?


What problem am I solving?
Define the goal
What information do I need?
Collect and manage data
Build the model
Find patterns in the data that lead to solutions.
Evaluate and critique model
Does the model solve my problem?
Present results and document
Establish that I can solve the problem, and how.
Deploy model
Deploy the model to solve the problem in the real world.

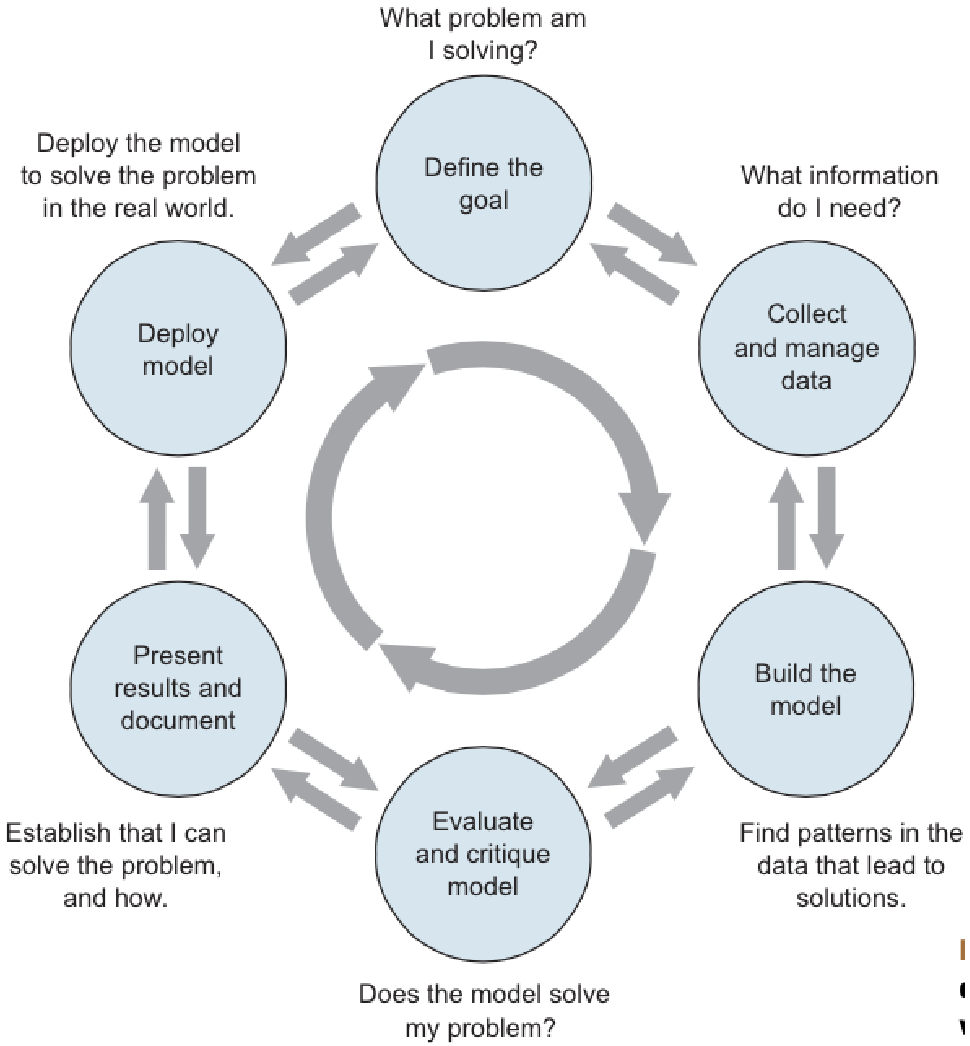
Data collection and Management

- What data is available?
- Is it good enough?
- Is it enough?
- What are sensible measurements to derive from this data? Units, transformations, rates, ratios, etc.


What problem am I solving?
Define the goal
What information do I need?
Collect and manage data
Build the model
Find patterns in the data that lead to solutions.
Evaluate and critique model
Does the model solve my problem?
Present results and document
Establish that I can solve the problem, and how.
Deploy model
Deploy the model to solve the problem in the real world.

Modeling

- What kind of problem is it? E.g., classification, clustering, regression, etc.
- What kind of model should I use?
- Do I have enough data for it?
- Does it really answer the question?



What problem am I solving?

Define the goal

What information do I need?

Collect and manage data

Build the model

Find patterns in the data that lead to solutions.

Evaluate and critique model

Does the model solve my problem?

Present results and document

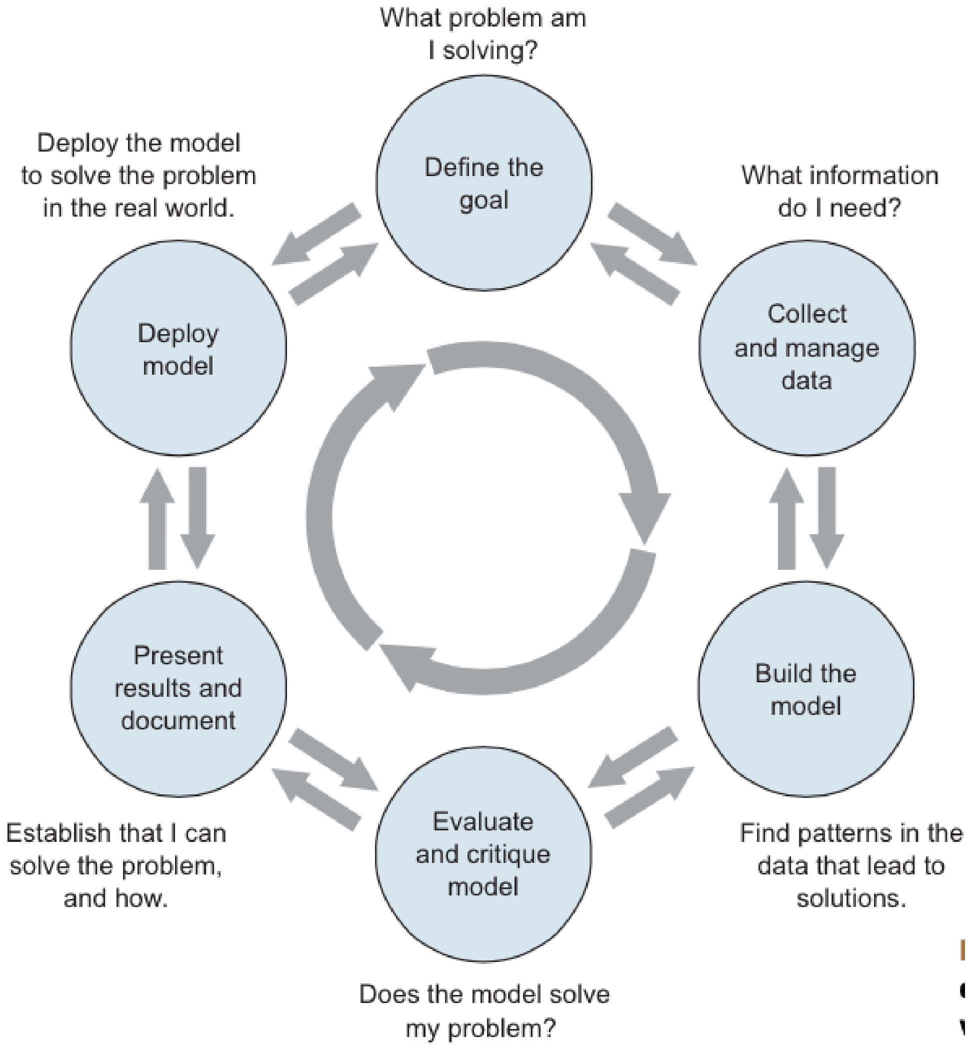
Establish that I can solve the problem, and how.

Deploy model

Deploy the model to solve the problem in the real world.

Model evaluation

- Did it work? How well?
- Can I interpret the model?
- What have I learned?



What problem am I solving?

Define the goal

What information do I need?

Collect and manage data

Build the model

Find patterns in the data that lead to solutions.

Evaluate and critique model

Does the model solve my problem?

Present results and document

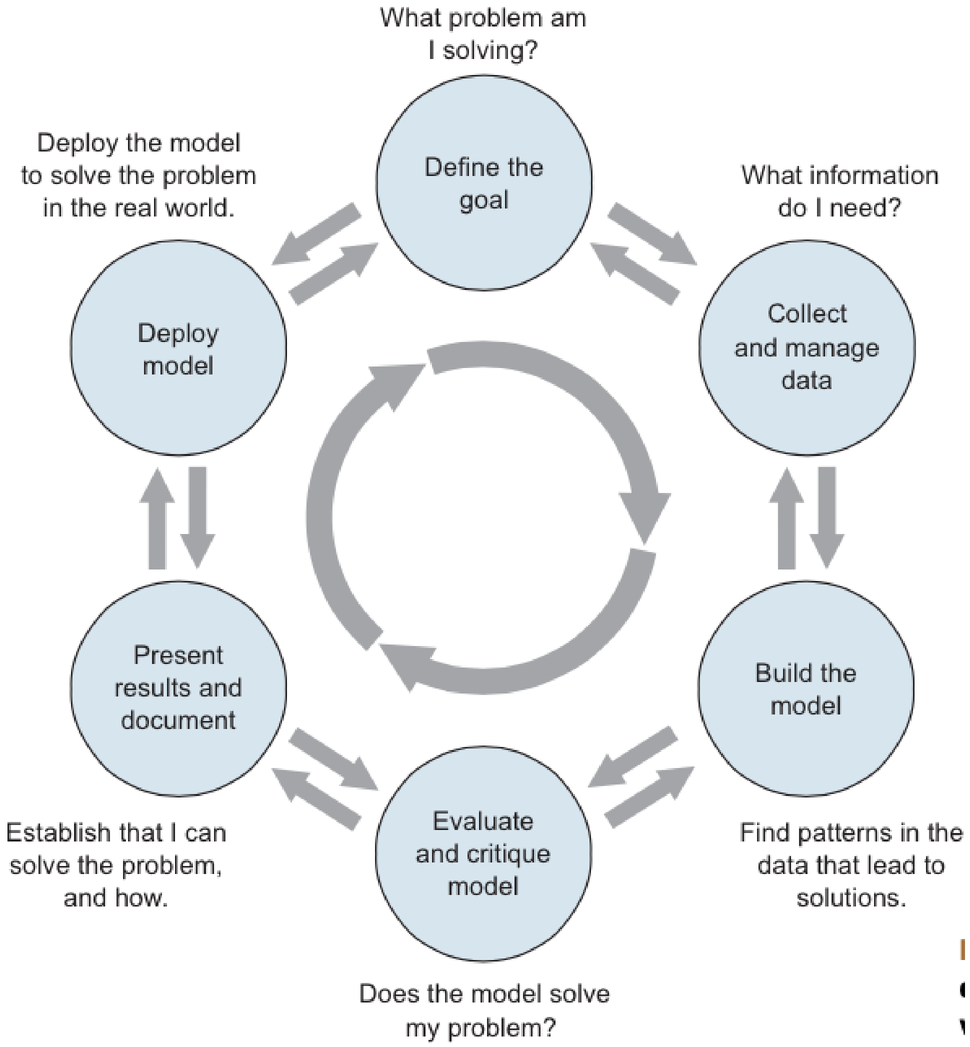
Establish that I can solve the problem, and how.

Deploy model

Deploy the model to solve the problem in the real world.

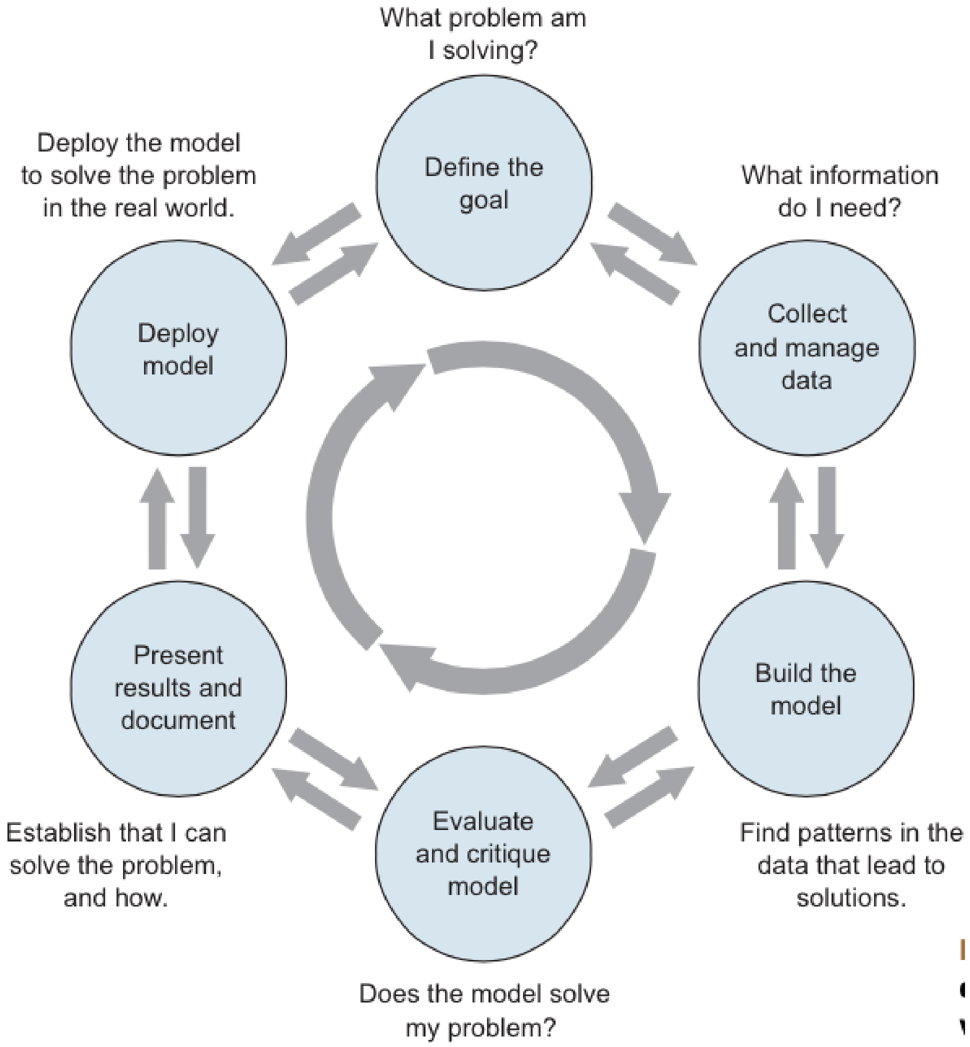
Presentation

- Again, what are the measurements that tell the real story?
- How can I describe and visualize them effectively?


What problem am I solving?
Define the goal
What information do I need?
Collect and manage data
Build the model
Find patterns in the data that lead to solutions.
Evaluate and critique model
Does the model solve my problem?
Present results and document
Establish that I can solve the problem, and how.
Deploy model
Deploy the model to solve the problem in the real world.

Deployment

- Where will it be hosted?
- Who will use it?
- Who will maintain it?


What problem am I solving?
Define the goal
What information do I need?
Collect and manage data
Build the model
Find patterns in the data that lead to solutions.
Evaluate and critique model
Does the model solve my problem?
Present results and document
Establish that I can solve the problem, and how.
Deploy model
Deploy the model to solve the problem in the real world.

An Illustrative Analysis

http://fivethirtyeight.com has a clever series of articles on the types of movies different actors make in their careers: https://fivethirtyeight.com/tag/hollywood-taxonomy/

I'd like to do a similar analysis. Let's do this in order:

1) Let's do this analysis for Diego Luna

2) Let's use a clustering algorithm to determine the different types of movies they make

3) Then, let's write an application that performs this analysis for any actor and test it with Gael García Bernal

Gathering data

Movie ratings

For this analysis we need to get the movies Diego Luna was in, along with their Rotten Tomatoes ratings. For that we scrape this webpage: https://www.rottentomatoes.com/celebrity/diego_luna.

Rating	Title	Credit	BoxOffice	Year
11	Berlin, I Love You	Drag Queen	—	2019
95	If Beale Street Could Talk	Pedrocito	—	2019
60	A Rainy Day in New York	Actor	—	2019
4	Flatliners	Ray	$16.9M	2017

Movie budgets and revenue

For the movie budgets and revenue data we scrape this webpage: http://www.the-numbers.com/movie/budgets/all

This is part of what we have for that table after scraping and cleaning up:

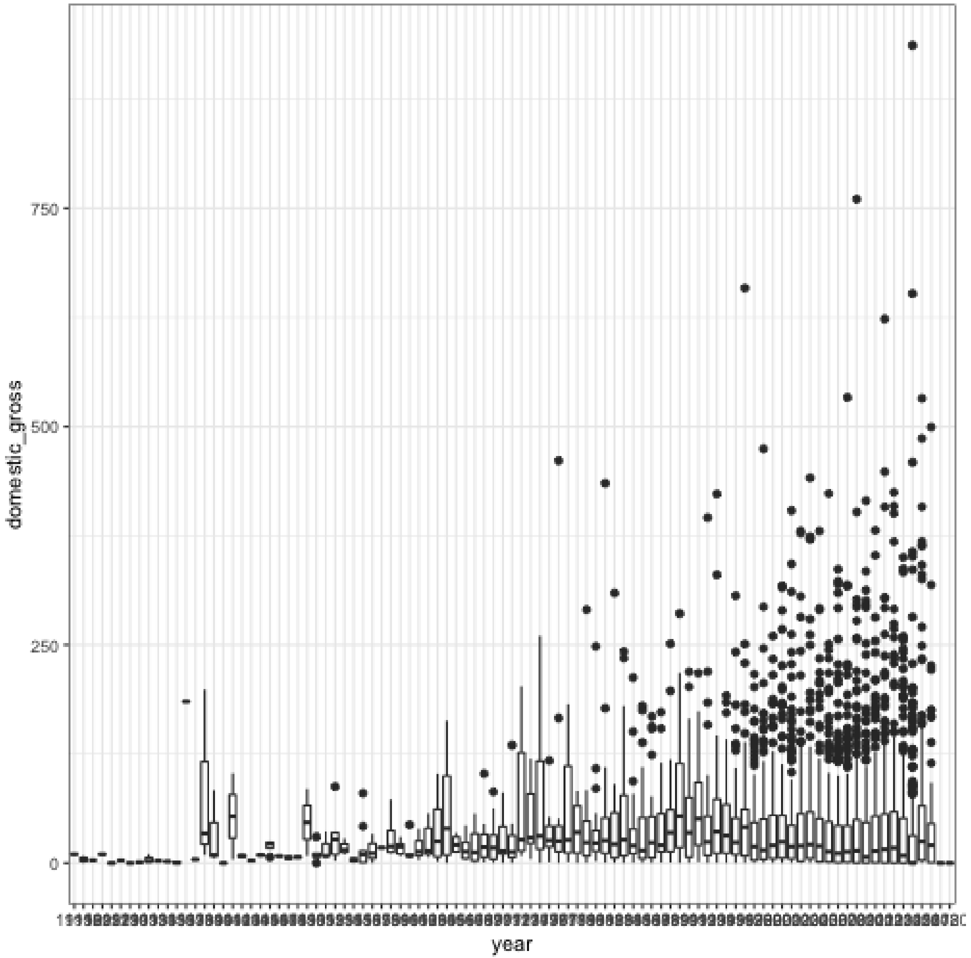
release_date	movie	production_budget	domestic_gross	worldwide_gr
2009-12-18	Avatar	425	760.50762	2783.9
2015-12-18	Star Wars Ep. VII: The Force Awakens	306	936.66223	2058.6
	Pirates of			

Movie budgets and revenue

Now we have data for 5358 movies, including its release date, title, production budget, US domestic and worlwide gross earnings. The latter three are in millions of U.S. dollars.

Movie budgets and revenue

One thing we might want to check is if the budget and gross entries in this table are inflation adjusted or not.



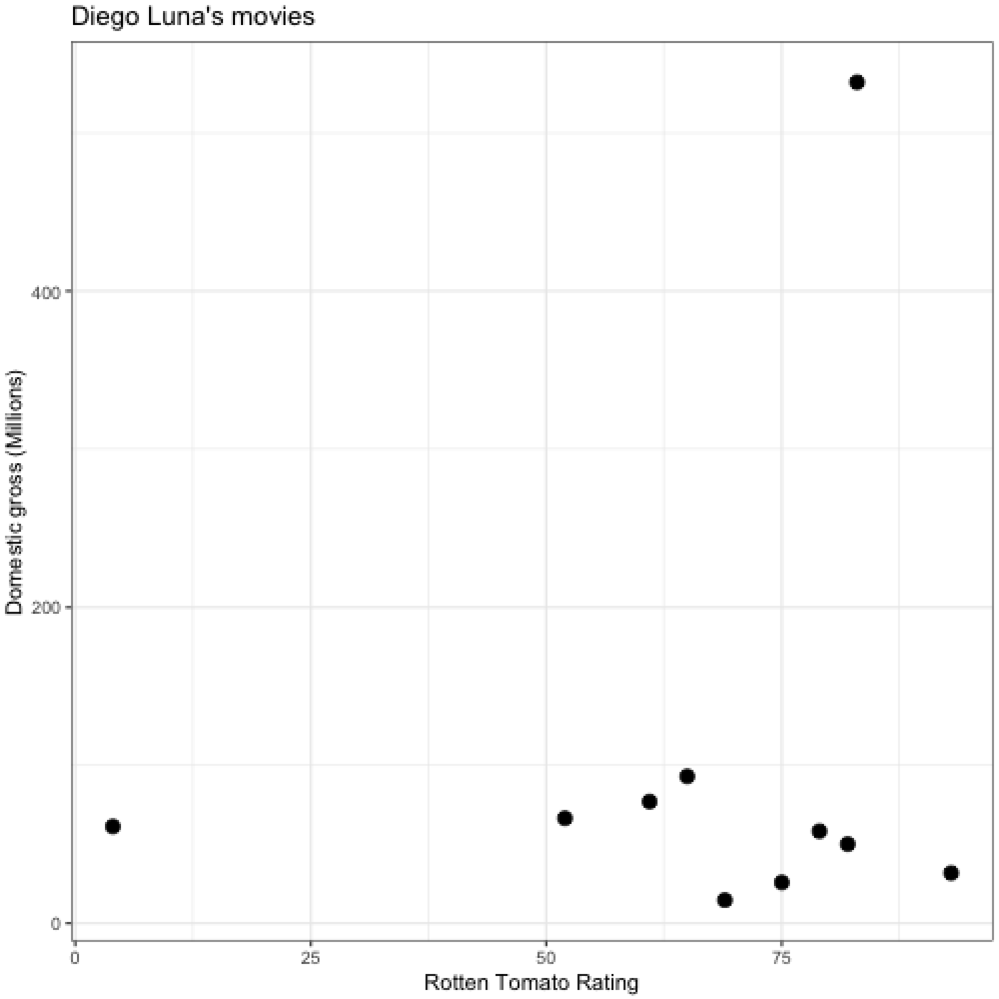
Manipulating the data

Next, we combine the datasets we obtained to get closer to the data we need to make the plot we want.

We combine the two datasets using the movie title, so that the end result has the information in both tables for each movie.

Rating	Title	Credit	BoxOffice	Year	release_date	production_budg
4	Flatliners	Ray	$16.9M	2017	1990-08-10	26
83	Rogue One: A Star Wars Story	Captain Cassian Andor	$532.2M	2016	2016-12-16	200
	The Book					

Visualizing the data

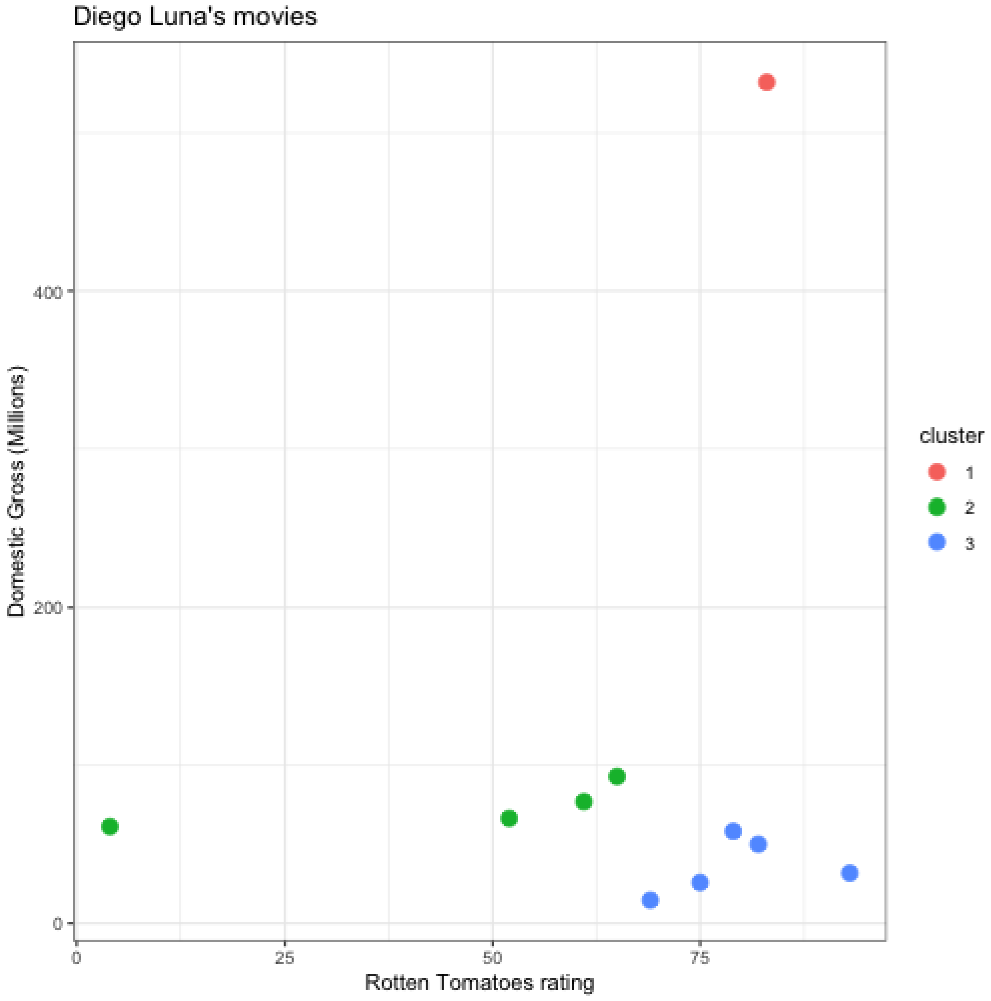


Modeling data

Use a clustering algorithm to partition Diego Luna's movies based on rating and domestic gross.

Title	Rating	domestic_gross	cluster
Rogue One: A Star Wars Story	83	532.17732	1
Flatliners	4	61.30815	2
Elysium	65	93.05012	2
Contraband	52	66.52800	2
The Terminal	61	77.07396	2
The Book of Life	82	50.15154	3

Visualizing model result



Visualizing model result

To make the plot and clustering more interpretable, let's annotate the graph with some movie titles.

- In the k-means algorithm, each group of movies is represented by an average rating and an average domestic gross.

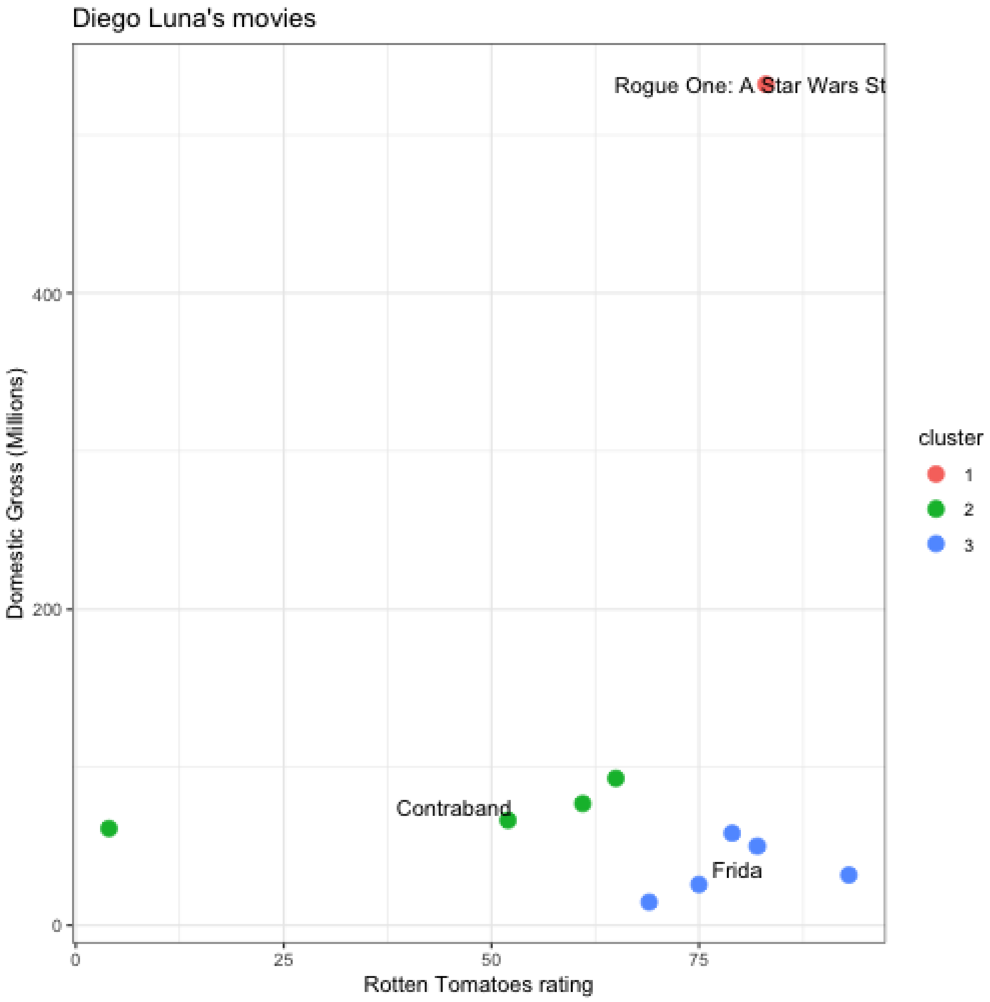
Visualizing model result

To make the plot and clustering more interpretable, let's annotate the graph with some movie titles.

- In the k-means algorithm, each group of movies is represented by an average rating and an average domestic gross.

- Find the movie in each group that is closest to the average and use that movie title to annotate each group in the plot.

Visualizing model result



Abstracting the analysis

While not a tremendous success, we decide we want to carry on with this analysis. We would like to do this for other actors' movies.

One of the big advantages of using R and Python is that we can write a piece of code as functions that takes an actor's name as input, and reproduces the steps of this analysis for that actor.

Abstracting the analysis

For our analysis, this function must do the following:

1. Scrape movie ratings from Rotten Tomatoes
2. Clean up the scraped data
3. Join with the budget data we downloaded previously
4. Perform the clustering algorithm
5. Make the final plot

With this in mind, we can write functions for each of these steps, and then make one final function that puts all of these together.

Abstracting the analysis

For instance, let's write the scraping function. It will take an actor's name and output the scraped data.

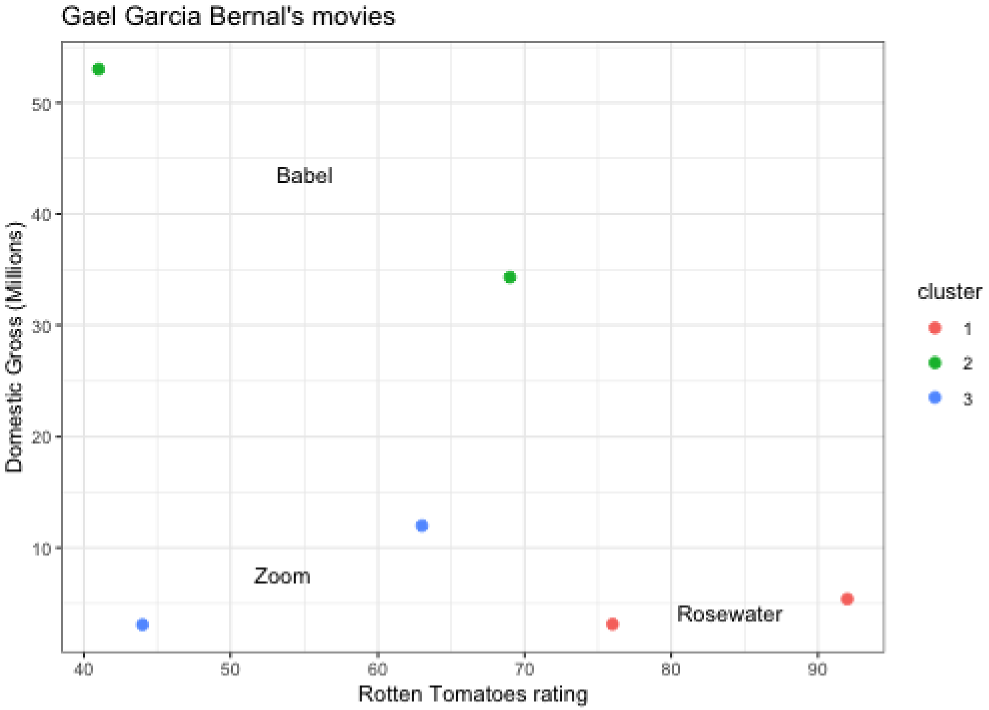
Let's test it with Gael García Bernal:

Rating	Title	Credit	BoxOffice	Year
No Score Yet	It Must Be Heaven	Actor	—	2019
No Score Yet	Lorena, Light-Footed Woman (Lorena, la de pies ligeros)	Executive Producer	—	2019

Abstracting the analysis

We can then write functions for each of the steps we did with Diego Luna before.

```
analyze_actor("Gael Garcia Bernal")
```



Making analyses accessible

Now that we have written a function to analyze an actor's movies, we can make these analyses easier to produce by creating an interactive application that wraps our new function. The `shiny` R package makes creating this type of application easy.

https://hcorrada.shinyapps.io/movie_app/

Summary

In this analysis we saw examples of the common steps and operations in a data analysis:

1) Data ingestion: we scraped and cleaned data from publicly accessible sites

2) Data manipulation: we integrated data from multiple sources to prepare our analysis

Summary

3) Data visualization: we made plots to explore patterns in our data

4) Data modeling: we made a model to capture the grouping patterns in data automatically, using visualization to explore the results of this modeling

5) Publishing: we abstracted our analysis into an application that allows us and others to perform this analysis over more datasets and explore the result of modeling using a variety of parameters