Background: Over 500 films are released in US theaters each year, and many more are made and released on other platforms (broadcast, streaming, etc.). While many of those movies fail commercially (A.K.A. “flop”), many meet or exceed expectations. A film is the sum of many parts, from the subject matter, release timing, director, script, actors/performers and below-the-line crew. What determines the success of a film? The decision to make (“greenlight”) a film is a complicated process involving artistic considerations, production ambition and budget, actor availability, and studio relationships. Historically, studio executives have relied on artistic/commercial instincts and other heuristics like analyzing the performance of comparable films. Nowadays, these decisions are, at least partially, data-driven, and companies like Cinelytic offer proprietary products that make sophisticated predictions about the likely performance of a film based on the aforementioned parts.

Setting: Assume we are working in the acquisitions and/or productions department of a movie studio, internet streamer or small production company. Directors/writers/producers come to pitch us their movie ideas. How can we leverage historical data on movie performance to decide which projects to invest in?

Data (not limited to):

* <https://www.kaggle.com/datasets/rounakbanik/the-movies-dataset/data>
* <https://about.netflix.com/en/news/what-we-watched-a-netflix-engagement-report>

Objectives:

* Design models of increasing complexity to predict the box office performance of a film. Models will likely include logistic regression, KNN regression, hidden Markov models, etc. Let’s brainstorm!
* Estimate the expected contribution to commercial success (A.K.A. “bankability”) of cast, crew and production members.