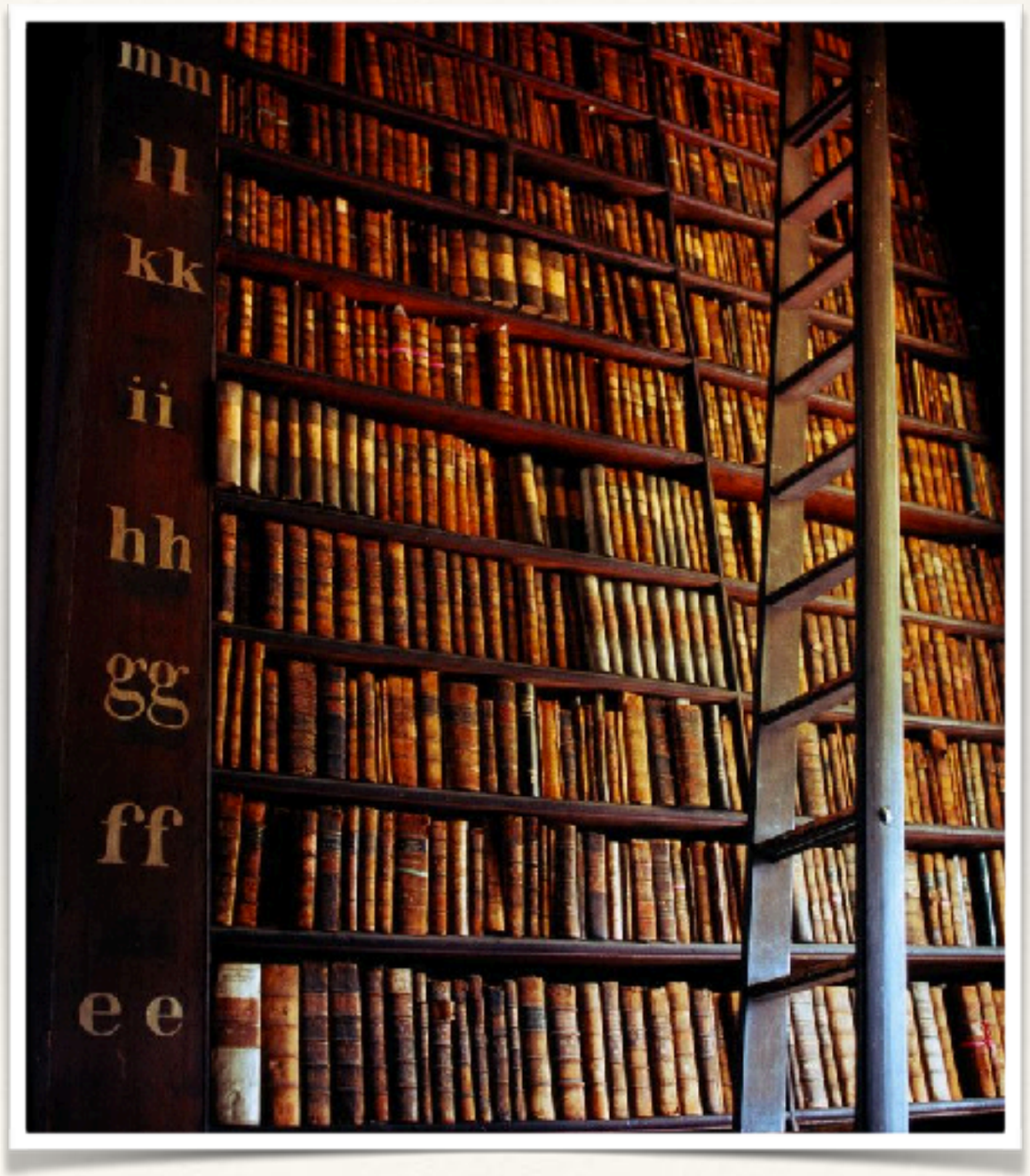
Humanities Data

Creating Data I

our object of study?

- Determine our object of study.
- Identify the necessary data.
- Unit of analysis

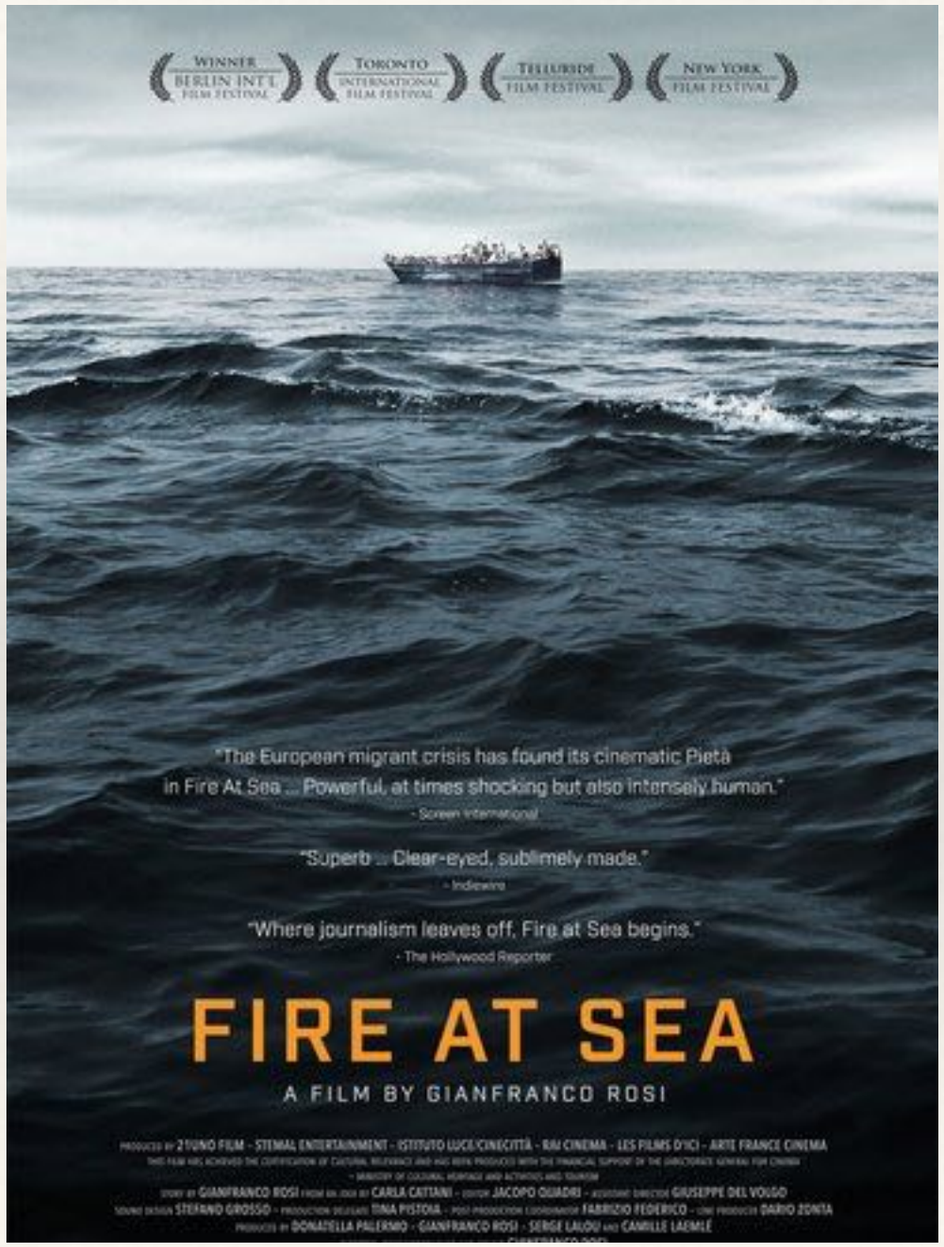


object of study

- #OscarsSoWhite brought renewed attention to how the Academy of Motion Picture Arts & Sciences awards filmmakers.
- Are there trends in nominations for Best Documentary?
- If so, what are they and what can they tell us about the culture and politics of Hollywood?

build data - movies

- Goal: Create a data set from a set of movies.
- https://tinyurl.com/dhmoviedata
- Group 1: 2017
- Group 2: 2016
- Group 3: 2015



data

- Movie Title
- Director
- Director Place of Birth
- Country of Origin
- Language
- Date Released
- Run Time
- Budget
- B&W or Color
- Topic

build data - movies

- Let's compare. What similarities or differences do we see?

normalizing data

- Each variable holds only one piece of information
- Each row represents one specific example of the unit of analysis
- If we need multiple units of analysis, store each of these in a different table

variable names

- Each variable holds only one piece of information.
- Depending on the object of study, the columns will change.
- Name conventions: prefer lower case with no spaces such as “movie_title”

- Movie Title
- Director
- Director Place of Birth
- Country of Origin
- Language
- Date Released
- Run Time
- Budget
- B&W or Color
- Topic

data types

- Variables have a“type”: numbers, strings, dates …
- Want a consistent format for each variable
- Need to take particular care with strings, which have two different flavors:
  - variable characters: in theory, can be any combination of characters, numbers, and other symbols
  - categorical: set vocabularies; use established standards when possible

building a schema

- step through each data element and figure out how to represent it consistently in a dataset
- discuss best practices
- consistency!

movie title

- data type: string, variable character
- things to consider:
  - language of the title (english? native language?)
  - stylized titles (ex. *SE7EN* or Seven)
  - subtitles (own field, part of title, or ignore?)
  - punctuation (keep or discard?)
- no italics or quotes

date released

- data type: date or number
- YYYY-MM-DD (ISO 8601 standard)
- things to consider:
  - missing or imprecise data?
  - what do we mean by released? country of origin? just use one country as a benchmark?

country of origin

- data type: string, categorical
- ISO -3166-1
  - ex. Germany is either DE, DEU, or 276
- things to consider:
  - multiple countries?

language

- data type: string, categorical
- ISO - 639
  - ex. English is “en” or Italian is “it”
- things to consider:
  - multiple languages?

movie gross

- data type: number
- things to consider:
  - currency
  - use punctuation? (Don’t!)

run time

- data type: number
- things to consider:
  - time frame (hours? minutes? seconds?)
- ex: PBCore : http://pbcore.org/pbcoreinstantiation/instantiationduration/

b&w or color

- data type: string, categorical
- things to consider:
  - how to type out black and white? just be consistent!

director

- data type: string, variable character
- things to consider:
  - how to represent the name?
    - Fullname (Stanley Kubrick)
    - Firstname, Lastname (Stanley | Kubrick)

director place of birth

- data type: string, variable character
- things to consider:
  - needs to be another table
  - what other data might you want to include?

build your data II

- Using the strategies discussed, revise your data.