Exploring Approaches to Examine NCI's Cancer Health Disparities Portfolio

NCAB Ad Hoc Subcommittee on Population Sciences, Epidemiology, and Disparities

Christine Burgess, Ph.D. and L. Michelle Bennett, Ph.D.
Center for Research Strategy

NIH NATIONAL CANCER INSTITUTE

September 1, 2020

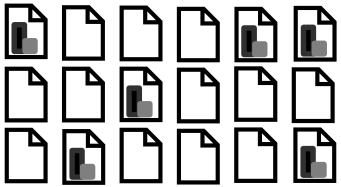
Overview

- Identifying relevant research: a critical first step for portfolio analysis
- Machine assisted methods
- Benefits and challenges
- Initial steps

Note: Information discussed here pertains only to portfolio analysis and not official reporting

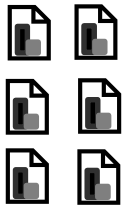
Identifying relevant research: a critical first step in portfolio analysis

Universe of Research



Identify Relevant Research

Research Portfolio



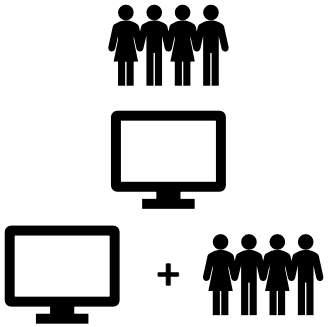
Analyze

Portfolio Analysis

Multiple methods available to identify research relevant to a particular topic

Methods include

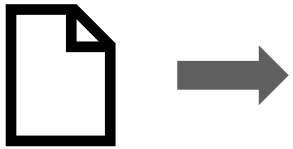
- Manual identification by subject matter experts
- Machine assisted approaches such as machine learning
- Hybrid approaches



- Each method has pros and cons
- Final use case and desired definition of research portfolio are key factors in selection of best methods for a particular question or goal

Machine assisted methods

- Convert documents into features based on available information
- Features are used to classify document
- Broad range of machine assisted methods
  - Machine Learning
    - Support Vector Machines (SVM)
    - Random Forest
  - Thesaurus Based (such as RCDC)
  - Business Rules



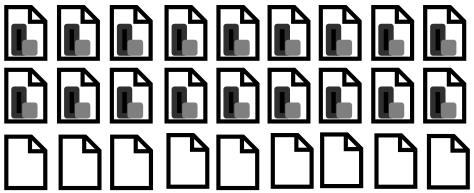
Feature X
Feature Y
Feature Z
Feature ….

Key information for NIH grants that has already been processed and is available for machine assisted methods:

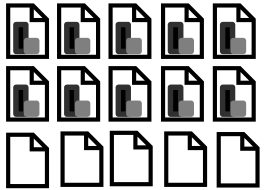
- Title
- Abstract
- Specific Aims
- PI
- Institution

Example of model development process

1. Develop Gold Standard Set

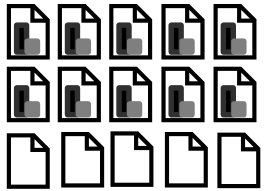


- *Gold standard set includes both positive and negative examples*
- *Requires input from SMEs*



Training Set

2. Model Development



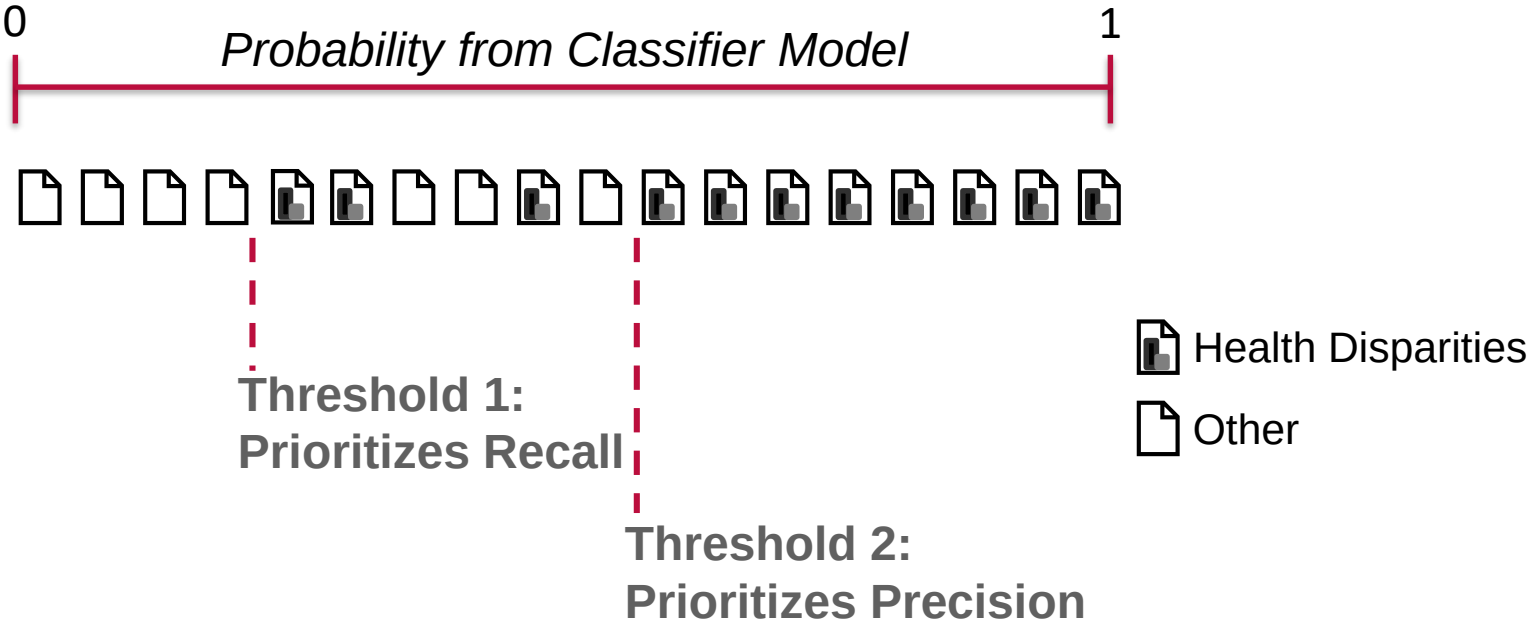
Test Set

3. Test Model

Quality of final results is dependent on SME input in first step

Machine assisted methods – performance measures

- **Recall** – Fraction of positives identified by classifier
- **Precision** – Fraction of detections that are true positives



Benefits and challenges of machine based methods

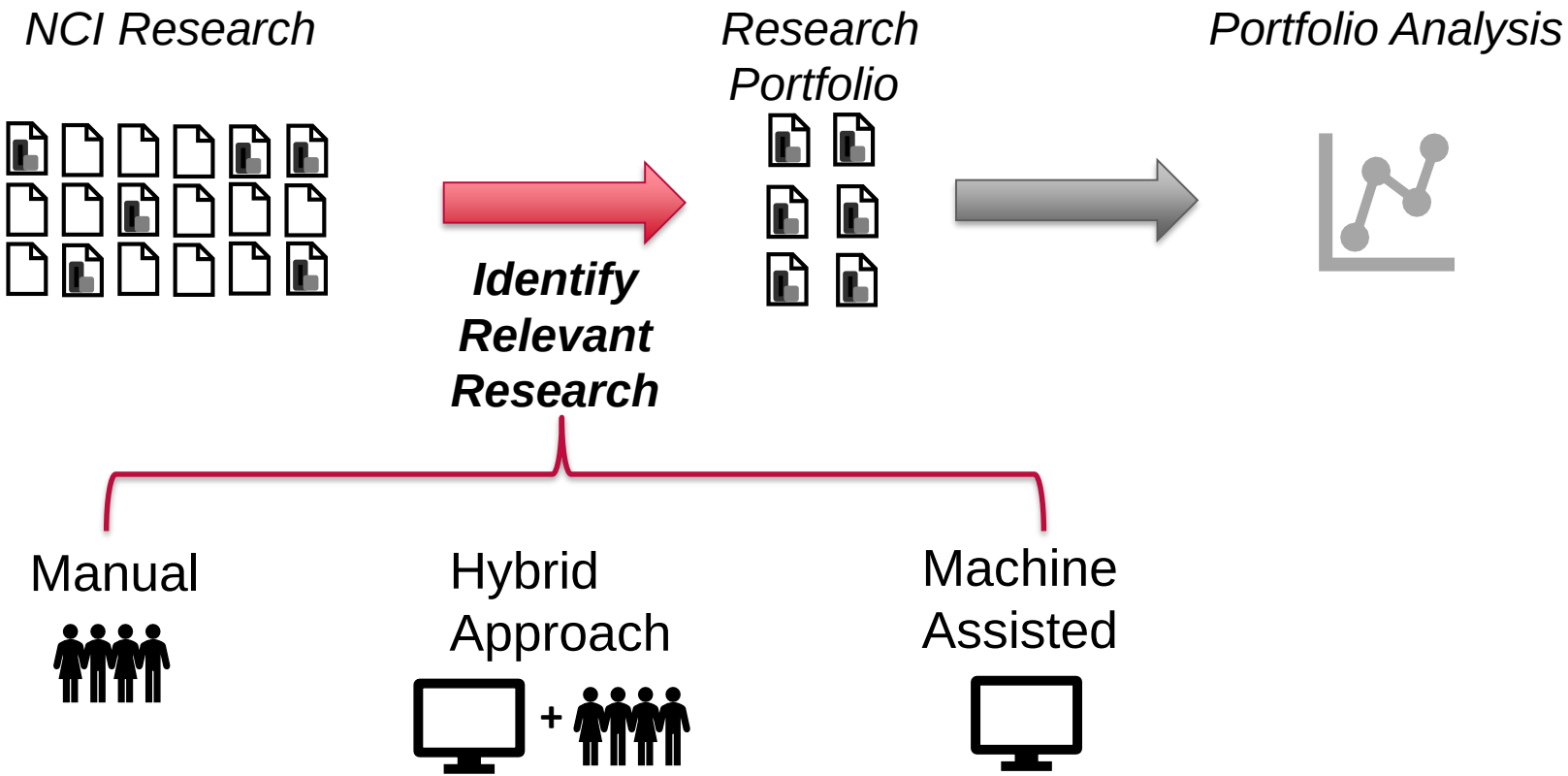
Benefits

- Consistent
- Once method is developed, fast classification at scale
- Works very well for clearly defined topics such as cancer type

Challenges

- Difficult to implement for nuanced topics
- Variations in language used to discuss a topic
- Dependent on quality of input data and available document features
- Method development can be time and resource heavy

Summary


NCI Research
Research Portfolio
Portfolio Analysis
Identify Relevant Research
Manual
Hybrid Approach
Machine Assisted

Initial steps and considerations

- Defining goal and desired outcomes of portfolio analysis
- Successful analysis would require a large amount of input from working group throughout the process
- Desired timeline and effort level
  - Intended to be a one-time analysis or the start of a tracking effort?
- Are data already collected as part of NCI Report on Minority Health and Health Disparities sufficient for goals of analysis?

www.cancer.gov

www.cancer.gov/espanol