Class 19 Frontiers of Marketing Analytics

Dr Wei Miao

UCL School of Management

December 6, 2023

Section 1

Causal Machine Learning

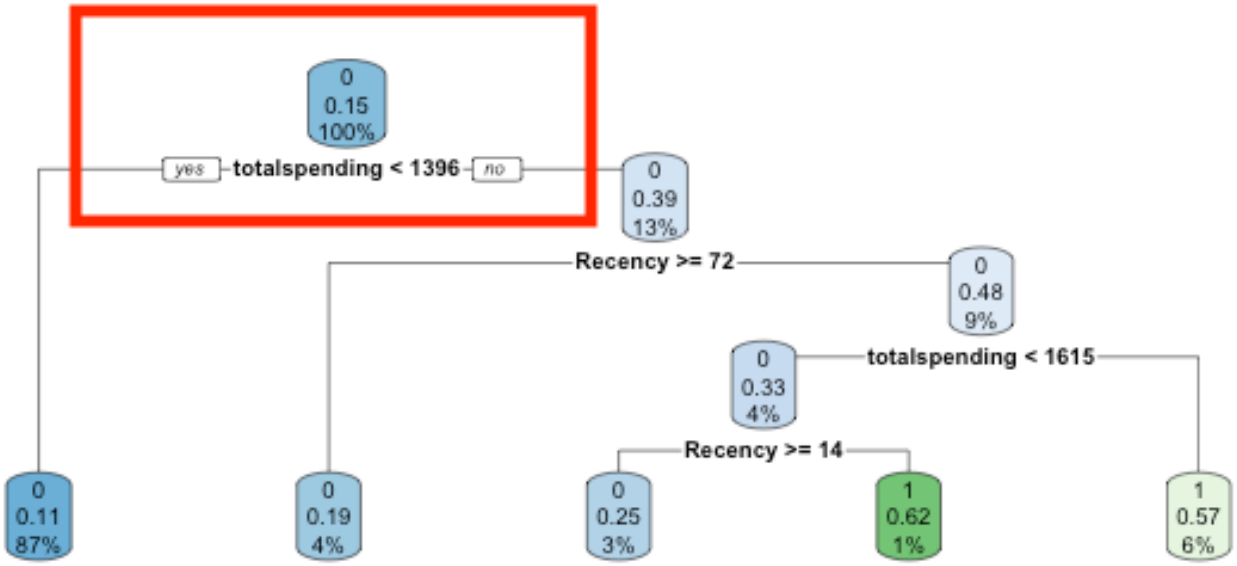
When Machine Learning Meets Causal Inference

- **Causal Machine Learning** represents the state-of-the-art development in the field of data science
  - While conventional machine learning excels at finding patterns and making predictions, it often falls short in understanding causation.
  - While conventional causal inference techniques (instrumental variable, DiD, RDD) can estimate average treatment effects, they are not good at estimating heterogeneous treatment effects.
- This is where CML steps in, aiming to uncover these causal relationships borrowing the predictive power of machine learning tools.

Causal Forest

- Causal Forest, a part of the CML toolkit, is particularly noteworthy. It is an extension of the Random Forest. The core idea of Causal Forest is to estimate the causal effect of a treatment using recursive binary splitting similar to decision trees in random forest.
- It does so by building a large number of *causal trees*, each based on a subset of data and features.

Visual Illustration of Causal Forest



- Unlike standard decision trees which aim at predicting outcomes, trees in a Causal Forest predict the *effect* of the intervention on each leaf.
- Each causal tree uses binary splitting at each possible value of all features X, and try to make the predicted treatment effects as differentiated as possible.

Implementation and Additional Reading Materials

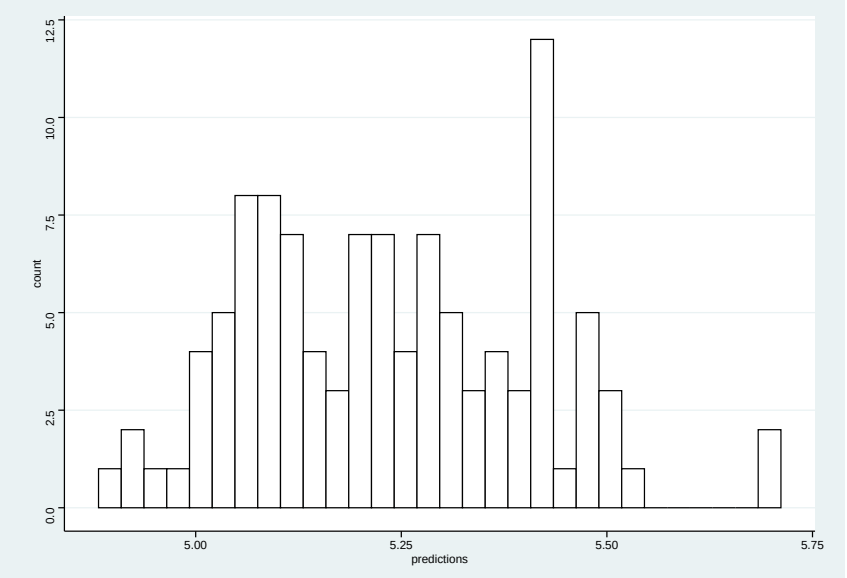
- grf is the R package that can implement causal forest.
  - Microsoft Research team has developed Python version of grf, named EconML.
  - Stanford YT channel also provides comprehensive tutorial videos by Prof Susan Athey et al.
- Data cleaning: each row stands for an individual

```
pacman::p_load(grf,fixest,dplyr,ggplot2,ggthemes)
data("base_did")
data_Y <- base_did %>%
  mutate(Post = ifelse(period >=6,1,0))%>%
  group_by(id,Post)%>%
  summarise(avg_outcome = mean(y)) %>%
  group_by(id) %>%
  summarise(first_diff = avg_outcome[2] - avg_outcome[1] )%>%
  ungroup()

data_W <- base_did %>%
  select(id, treat) %>%
  unique()

data_X <- base_did %>%
  filter(period <6) %>%
  group_by(id) %>%
  summarise(avg_x = mean(x1)) %>%
  ungroup()
```

Run Causal Forest

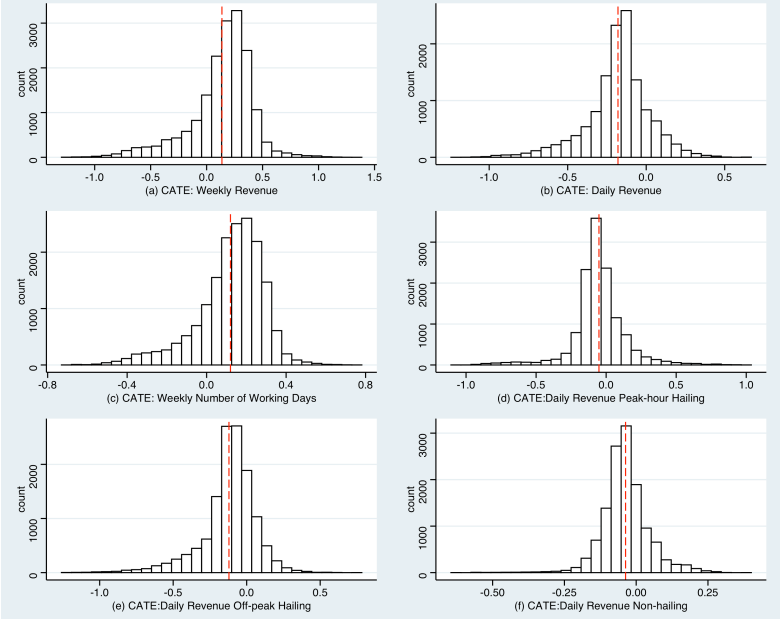
- We can use causal forest to estimate the treatment effects for each individual and plot the histogram.

```
cf <- causal_forest(X = data.matrix(data_X$avg_x),
                    Y = data.matrix(data_Y$first_diff),
                    W = data_W$treat)

predicted_CATE <- predict(cf)

ggplot()+
  geom_histogram(data = predicted_CATE,
                         aes(x = predictions),
                  color = 'black',fill = 'white')+
  theme_stata()
```

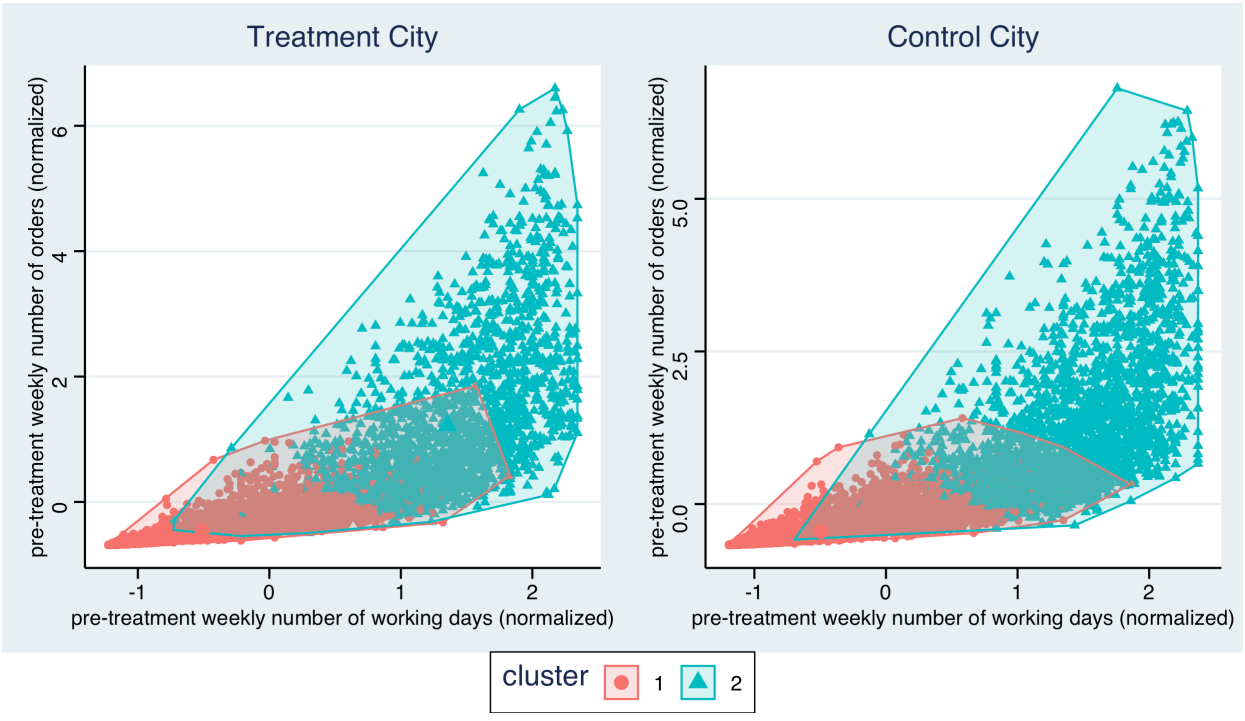


Application: Heterogeneous Causal Effect of Surge Pricing on Uber Drivers

- Miao et al. (2022) study the causal effects of surge pricing on driver labor supply decisions. The ridesharing company introduced surge pricing in one city but not the other, such that we have a nice difference-in-differences setup:
- Using causal forest method, we are able to compute the treatment effect of surge pricing for each individual driver, and plot the distribution.



Clustering for Heterogeneity Analyses



- We use K-means clustering to segment out 2 clusters of drivers: full-time and part-time drivers.
  - Full-time drivers have decreased weekly revenue due to capacity constraint.
  - Part-time drivers flooded into the market and have increased weekly revenues by working more days.
- Although surge pricing enlarged the total pie for the company, the benefit was unevenly distributed across full-time and part-time drivers.

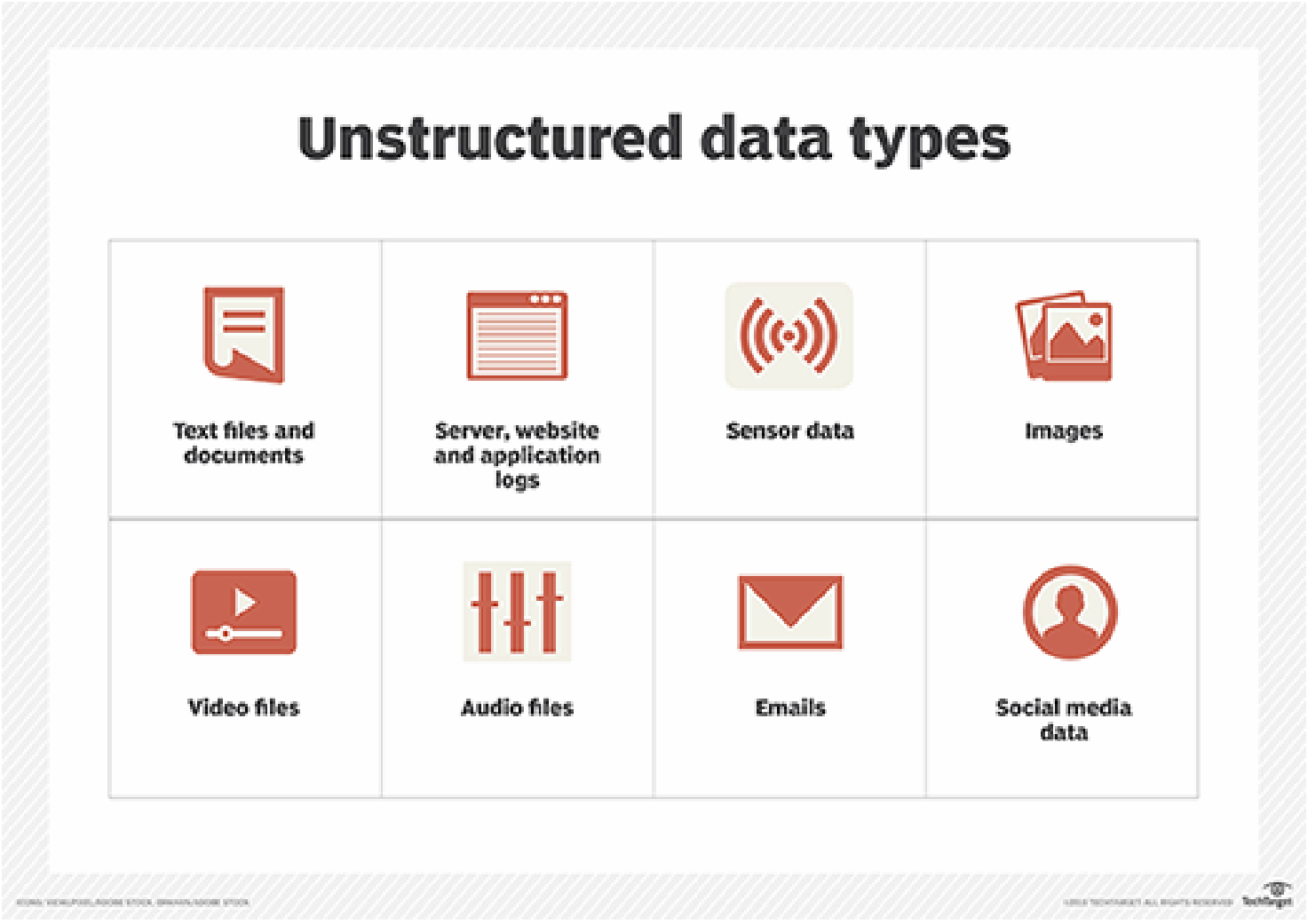
Tips for Term 3 Dissertation Using Causal Forest

- Help the company to analyze
  - A/B testings they have run; investigate heterogeneous treatment effects
  - natural experiment: some policies are introduced to some markets first
- Focus on how treatment effects vary with individuals of different characteristics

Section 2

Unstructured Data

Unstructured Data



Sentiment Analysis

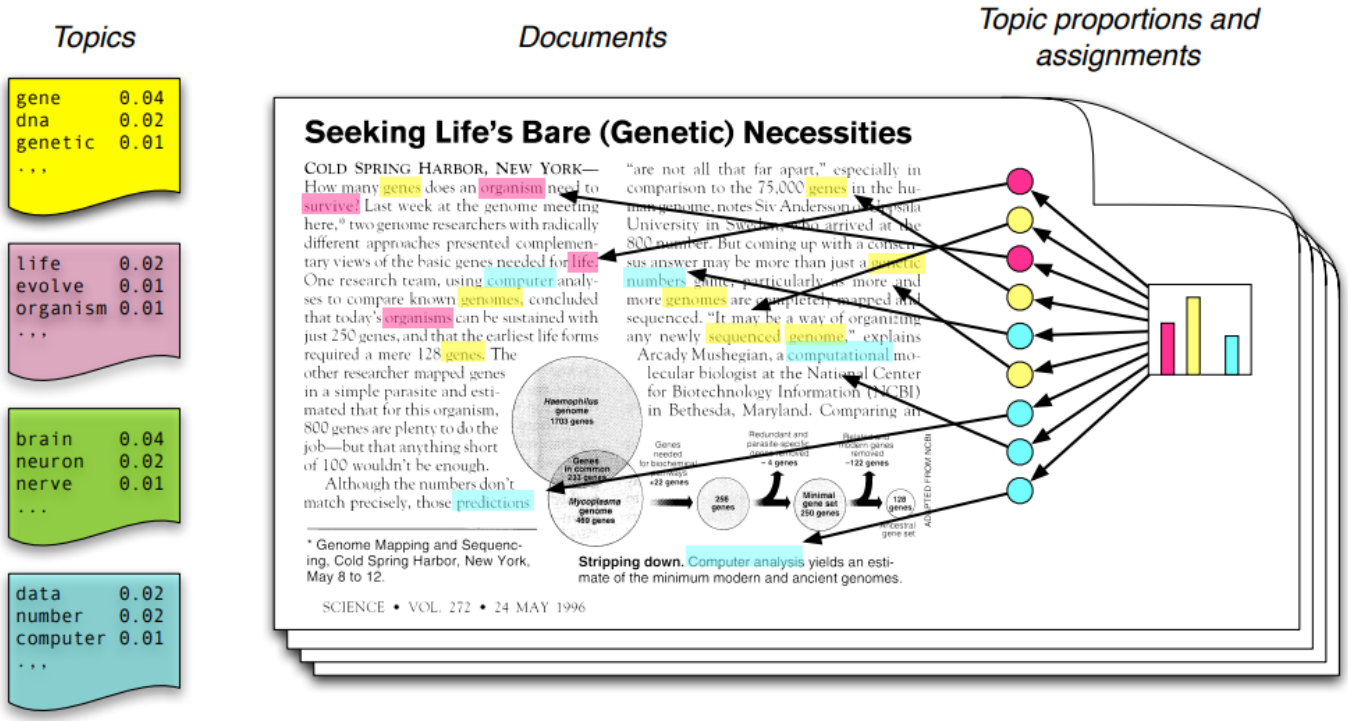
- Sentiment Analysis leverages the power of natural language processing (NLP) and machine learning to understand customer emotions and opinions.
- This analytical technique processes vast amounts of unstructured text data from sources like social media posts, reviews, forum discussions, and customer feedback. By evaluating the tone and context of these texts, sentiment analysis classifies them into categories such as positive, negative, or neutral. This classification helps businesses gauge overall customer sentiment, monitor brand reputation, and understand consumer needs and preferences.

Implementation of Sentiment Analysis

- Implementation: Sentiment Analysis in R

Topic Modeling

- Topic Modeling is a natural language processing (NLP) technique used to automatically identify and extract underlying topics from large volumes of text data.



- Each **topic** is a distribution over words
- Each **document** is a mixture of corpus-wide topics
- Each **word** is drawn from one of those topics

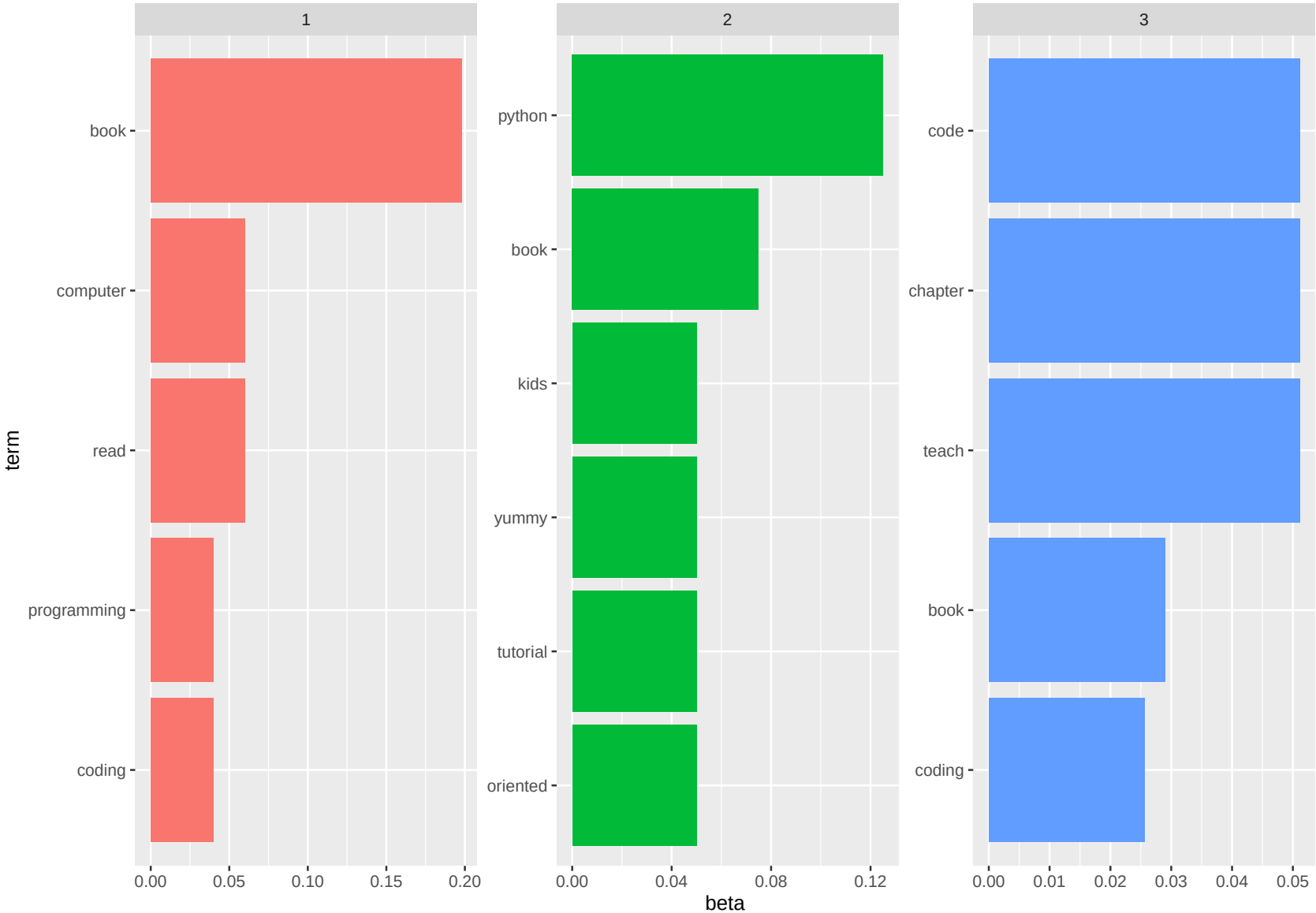
Implementation of Topic Modeling

- The intuition behind topic modeling is that documents comprise mixtures of topics, where a topic is characterized by a cluster of words with high probability of appearing together. Algorithms like Latent Dirichlet Allocation (LDA) are commonly used; they assume each document is a mixture of a small number of topics and that each word's presence is attributable to one of the document's topics. This probabilistic approach enables the algorithm to categorize and group words into topics without any prior labeling or training, making it an unsupervised machine learning technique.
- Implementation: Topic modeling in R

Application in Marketing

- Topic modeling enables marketers to uncover prevailing subjects in customer feedback or online discussions, thus providing insights into consumer behavior and preferences. This can inform targeted marketing strategies, product development, and content creation.
- For instance, by analyzing customer reviews, a company can identify common themes in customer satisfaction or dissatisfaction, guiding product improvements or highlighting areas for enhanced customer service.

Example of NLP: Guess the Book Name



Tips for Term 3 Dissertation Projects Using Unstructured Data

- Collect text data from Google review, online forums, TrustPilot, Twitter using data crawlers.
- Conduct sentiment analysis and topic modeling for the text data by each month.
- Investigate the evolving trend of sentiments and topic, and how company strategies dynamically affect customer sentiment/topics.