Clustering

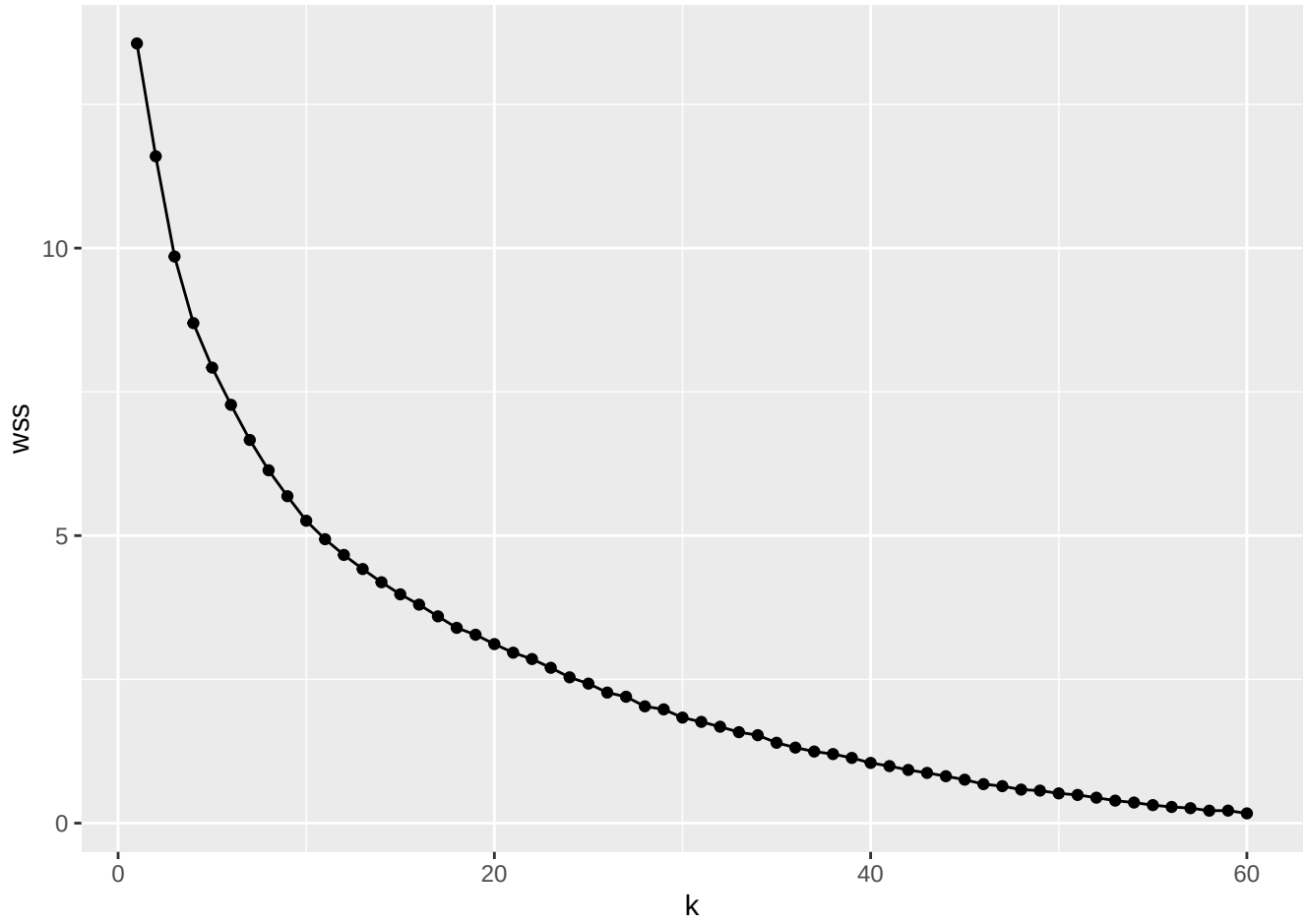
High-Dimensional Data Analysis and Machine Learning

Camille Mondon

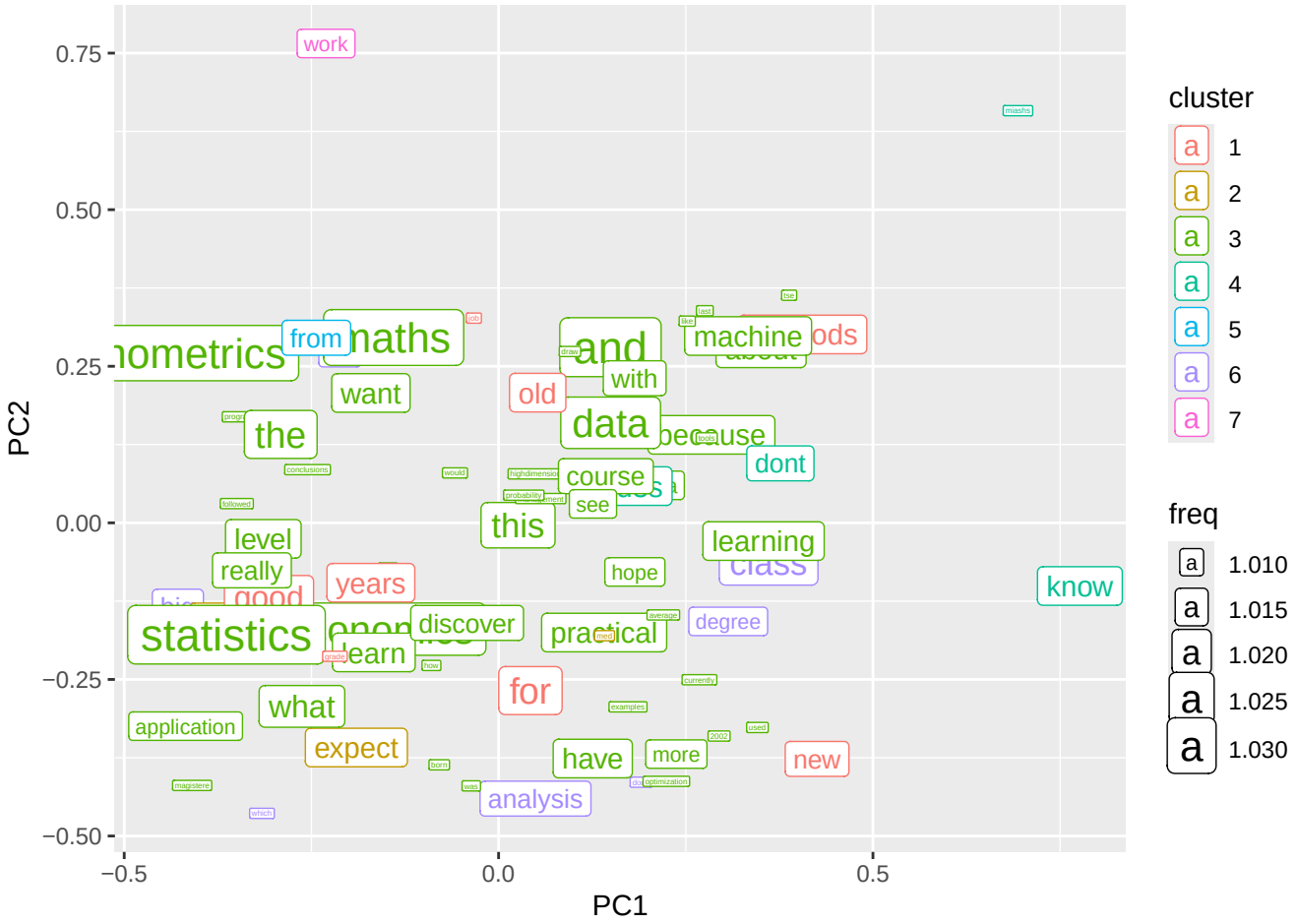
1 k-means

https://www.tidymodels.org/learn/statistics/k-means/

```
library(tm)
library(tidyverse)

postits_corpus <- Corpus(DirSource("postits"),
  readerControl = list(
    reader = readPlain,
    language = "en",
    load = TRUE
  )
) |>
  tm_map(removePunctuation)

postits_freq <- TermDocumentMatrix(postits_corpus,
  control = list(bounds = list(global = c(2, Inf)))
) |>
  as.matrix() |>
  rowSums()

postits <- TermDocumentMatrix(postits_corpus,
  control = list(bounds = list(global = c(2, Inf)), weighting = weightTfIdf)
)

tibble(k = 1:60) |>
  mutate(wss = map_dbl(k, ~ kmeans(postits, .x, nstart = 100)$tot.withinss)) |>
  ggplot(aes(k, wss)) +
  geom_point() +
  geom_line()
```




```
postits_kmeans <- kmeans(postits, centers = 7, nstart = 100)

prcomp(postits)$x |>
  as.data.frame() |>
  rownames_to_column("name") |>
  mutate(
    cluster = as.factor(postits_kmeans$cluster),
    freq = postits_freq^0.01
  ) |>
  ggplot() +
  geom_label(aes(PC1, PC2, label = name, color = cluster, size = freq),
    position = position_jitter(width = 0.4, height = 0.4)
  )
```



The between-total ratio is 50.85%.

2 Hierarchical clustering