The two most disruptive political events of the last few years are undoubtedly the Brexit referendum to leave the European Union and the election of Donald Trump. Both are commonly associated with the political consulting firm *Cambridge Analytica* and a technique known as *Microtargeting*.

If you want to understand the *data science* behind the *Cambridge Analytica/Facebook data scandal* and Microtargeting (i.e. *LASSO regression*) by building a toy example in R read on!

The following post is mainly based on the excellent case study [“Cambridge Analytica: Rise and Fall”](https://www.thecasecentre.org/main/products/view?id=156216) by my colleague Professor Oliver Gassmann (who was so kind as to email it to me) and Raphael Boemelburg, both from the University of St. Gallen, Switzerland (where I did my PhD), and [“Weapons of Micro Destruction: How Our ‘Likes’ Hijacked Democracy”](https://towardsdatascience.com/weapons-of-micro-destruction-how-our-likes-hijacked-democracy-c9ab6fcd3d02) by Data Scientist Dave Smith (the data for the toy example is also from that article).

Also well worth a watch is the Netflix documentary “The Great Hack”. I encourage all of my colleagues from academia and all teachers to consider screening it to their students/pupils. Not many know that Netflix is kind enough to provide a license to do this legally: [Neflix grant of permission for educational screenings](https://media.netflix.com/en/only-on-netflix/446957). In this documentary, it becomes clear that microtargeting can be considered as some form of *psychological warfare* and *information warfare*, i.e. as a military-grade *weapon* with the potential to *destroy our democratic system*:

So, how does it actually work?

Basically, *Microtargeting* is the prediction of psychological profiles on the basis of social media activity and using that knowledge to address different personality types with customized ads. Microtargeting is not only used in the political arena but of course also in *Marketing* and *Customer Relationship Management (CRM)*.

A well-known psychological model is the so-called [OCEAN model](https://en.wikipedia.org/wiki/Big_Five_personality_traits):



Source: Wikimedia

The five personality traits are:

* *Openness to experience* (inventive/curious vs. consistent/cautious)
* *Conscientiousness* (efficient/organized vs. easy-going/careless)
* *Extraversion* (outgoing/energetic vs. solitary/reserved)
* *Agreeableness* (friendly/compassionate vs. challenging/detached)
* *Neuroticism* (sensitive/nervous vs. secure/confident)

You can find out about your own personality by taking this free, anonymous test: [The Big Five Project Personality Test](https://www.outofservice.com/bigfive/).

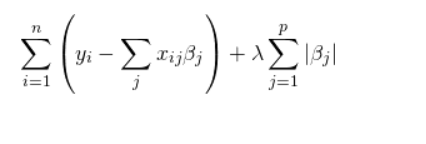
If you had the psychological profiles of many individuals you could use this together with modern advertising technology (which lets you show different ads to different people) to cater to their individual needs with the aim to manipulate them very efficiently:

Source: Cambridge Analytica

So far, so standard… the difficult part is predicting psychological traits with high accuracy! And here comes data science into play, namely a technique called *LASSO regression* (for *Least Absolute Shrinkage and Selection Operator*), or simply *the LASSO*.

If you are not familiar with *Classical Linear Regression (CLR)* please read my post [Learning Data Science: Modelling Basics](http://blog.ephorie.de/learning-data-science-modelling-basics) first. The difference between CLR and the LASSO is that with the latter you simultaneously minimize the error of the regression and the sum of the coefficients so that some coefficients will become zero. Thereby you only retain the important variables, so LASSO regression provides automatic *variable selection*! The other effect is that by effectively reducing the complexity of the model (also called *shrinkage* in this context, some form of *regularization*) you prevent *overfitting*.

Mathematically you minimize the following expression:



The first summand is the error term of classical linear regression (the difference between the real values and the predicted values), the second is the regularization term. (lambda) is a *hyperparameter* which controls how big the shrinkage of the coefficients get (the bigger the smaller the coefficients).

Enough of the theory, let’s get to a toy example! We use the psychological factor “Openness” as the trait to predict. We have seven individuals with their openness score and their individual likes on certain Facebook posts (you can find the data for this example here: [cambridge-analytica.csv](http://blog.ephorie.de/wp-content/uploads/2020/03/cambridge-analytica.csv)). We use five for the *training set* and two for the *test set*:

data <- read.csv("data/cambridge-analytica.csv")

data

## Openness Person The\_Colbert\_Report TED George\_Takei Meditation

## 1 1.85 Adam 1 1 1 1

## 2 1.60 Bob 1 1 1 1

## 3 -0.26 Cathy 0 1 1 0

## 4 -2.00 Donald 0 1 0 0

## 5 -2.50 Erin 0 0 0 0

## 6 1.77 Hilary 1 1 1 1

## 7 -2.20 Mike 0 0 0 0

## Bass\_Pro\_Shops NFL\_Network The\_Bachelor

## 1 0 0 0

## 2 0 0 0

## 3 0 0 1

## 4 1 1 0

## 5 1 1 1

## 6 0 0 0

## 7 1 1 0

## Ok\_If\_we\_get\_caught\_heres\_the\_story

## 1 0

## 2 1

## 3 1

## 4 1

## 5 1

## 6 1

## 7 1

x\_train <- as.matrix(data[1:5, 3:10])

x\_test <- as.matrix(data[6:7, 3:10])

y\_train <- as.matrix(data[1:5, 1])

y\_test <- as.matrix(data[6:7, 1])

Now we build the actual model with the excellent glmnet package (on CRAN), which has been co-developed by one of the discoverers of the LASSO, my colleague Professor Robert Tibshirani from Standford University, and after that plot how the coefficients get smaller the bigger gets:

library(glmnet)

## Loading required package: Matrix

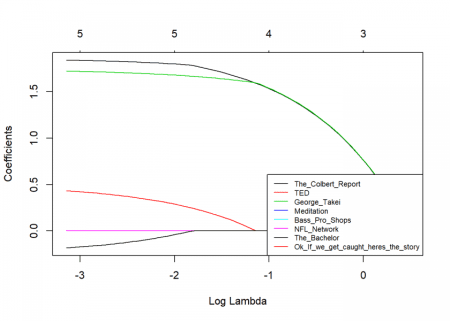
## Loading required package: foreach

## Loaded glmnet 2.0-18

LASSO <- glmnet(x\_train, y\_train, alpha = 1) # alpha = 1 for LASSO regression

plot(LASSO, xvar = "lambda")

legend("bottomright", lwd = 1, col = 1:6, legend = colnames(x\_train), cex = .7)



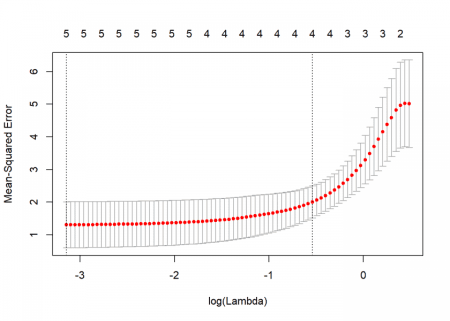
In the plot, you can see how growing lets the coefficients shrink. To find a good value for we use a technique called *cross-validation*. What it basically does is building a lot of different training- and test-sets automatically and averaging the error over all of them for different values of . After that, we plot the resulting errors with upper and lower standard-deviations:

cv\_LASSO <- cv.glmnet(x\_train, y\_train)

## Warning: Option grouped=FALSE enforced in cv.glmnet, since < 3 observations

## per fold

plot(cv\_LASSO)



In this case, we see that the minimal gives the smallest error, so we choose it for the prediction of the openness score of our test set:

round(predict(LASSO, x\_test, s = cv\_LASSO$lambda.min), 2)

## 1

## 6 1.58

## 7 -2.41

y\_test

## [,1]

## [1,] 1.77

## [2,] -2.20

Not too bad! In reality, with only 70 likes the algorithm can assess personality better than a friend of the person would be able to, with 150 likes it is better than the parents, and with 300 likes, it is even better than the spouse! Creepy, isn’t it!

Another big advantage is the *interpretability* of LASSO regression. It is easily discernible that “The Colbert Report” and “George Takei” (a former Star Trek actor who became a gay rights and left-leaning political activist) are the biggest drivers here:

round(coef(LASSO, s = cv\_LASSO$lambda.min), 2)

## 9 x 1 sparse Matrix of class "dgCMatrix"

## 1

## (Intercept) -2.23

## The\_Colbert\_Report 1.84

## TED 0.43

## George\_Takei 1.72

## Meditation 0.00

## Bass\_Pro\_Shops .

## NFL\_Network .

## The\_Bachelor .

## Ok\_If\_we\_get\_caught\_heres\_the\_story -0.18

It is no overstatement to say that with those new possibilities of microtargeting we have entered a new era of potential (and real!) manipulation. I hope that you now understand the data science behind it better.

In my opinion that knowledge is important to be part of the necessary conversation about the consequences for our society. This conversation has only just begun… looking forward to your comments on this topic!