UNIVERSITÄT
LUZERN

Big Data Analytics

Tutor	Luigi Curini is Professor of Political Science at University of Milan (Italy) and Visiting Professor at Waseda University (Tokyo). His research interests include party competition, legislative behavior, and text analytics. He is one of the founders of VOICES from the Blogs, a spin-off company of University of Milan dealing with Big Data Analytics. He has been the CEO of that company for 4 years. He is the co-editor of the SAGE Handbook of Research Methods in Political Science & International Relation (2020).
Organization	Digital Skills, University of Lucerne
Language	English
ECTS-Points	4
Contact	Nadia.buehler@unilu.ch
Content	Big data are those labeled, for strange reasons, with the capitalized “Big”. Nevertheless, they are still “data”, altough with some specific characteristics: large volume, high frequency and, most notably, unpredictability - data come in the many different forms, they are raw, messy, unstructured, not ready for processing, and so on. Still, these data convey a lot of information to social scientists and good statistical techniques are required in order to extract meaningful results from them. In this workshop we will focus on a specific type of Big data, namely digital texts, both from social media as well as other sources (such as legislative speeches or electoral programs). The aim is to provide an introductory guide to this exciting new area of research, while also offering guidelines on how to effectively use statistical methods on texts for social scientific research by discussing the advantages, but also the limits, of each approach. The attention will be devoted to three main areas: 1) scaling methods that allow to estimate the location

FROHBURGSTRASSE 3
P.O.BOX 4466
6002 LUCERNE

T

	of actors in some policy space; 2) supervised classification methods that allow to organize texts into a set of pre-defined categories; 3) unsupervised classification that allow to discover new ways of organizing texts into a set of unknown categories. Time permitting, beyond the Bag-of-Word approach we will also briefly cover the word-embedding approach. Day one, Content: supervised and unsupervised scaling methods (Wordscores and Wordfish) Day two, Content: unsupervised and semi-supervised classification methods (LDA, Structural Topic Models, keyATM) Day three, Content: an introduction to supervised classification methods & on how to extract texts from social media source Day four, Content: supervised classification methods (machine learning algorithms)
Prerequisites/Materials	An elementary knowledge of R (having attended any of the introductory workshops offered by Campus Lucerne usually satisfies this requirement), plus a curiosity towards applied statistics, are good prerequisites for the lab sessions. Participants will familiarize with quanteda, one of the most well-known and better-developed text-mining R package. On top of that, in our lab examples we will employ several other packages, in particular when discussing about classification methods (for example: topicmodels, stm, keyATM, naivebayes, e1071, randomForest). We will also use some R packages to extract texts from social media source, such as rtweet. All the datasets, replication files of the lab sessions and reference texts will be made available at a dedicated URL before the beginning of the workshop. Workshop participants should bring their own laptop with R, RStudio and the relevant packages previously installed and functioning (instructions will be circulated beforehand).