INTRO TO MACHINE LEARNING FOR SOCIAL SCIENCES

Dr. Iulia Cioroianu

Institute for Policy Research

University of Bath

Why machine learning?

Can we predict armed conflict by using newspaper text, reducing vast quantities of newspaper articles to interpretable topics? (Mueller and Rauh, 2017, APSR)

Can we predict local opinion surveys and polls using large volumes of social media data? (Beauchamp, 2017, AJPS)

Can we infer the ideological positions of legislators and other relevant actors from millions of recorded campaign contributions? (Bonica, 2018, AJPS)

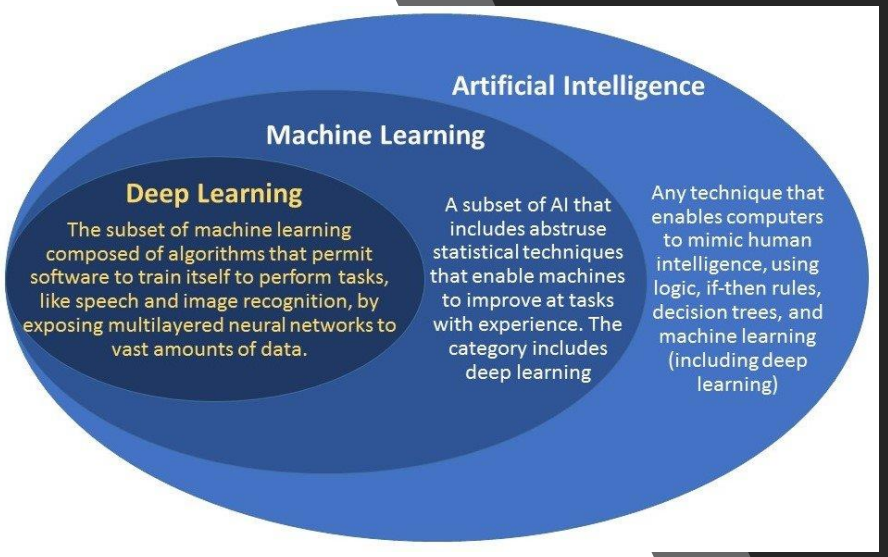
Can we understand censorship in China based on millions of social media posts? (King, Pan and Roberts, 2013, APSR)

Increasingly, research impossible to conduct in the absence of computational methods for:

Data collection	Data processing	Data analysis

Important part – machine learning.

What is machine learning?



https://www.geospatialworld.net/blogs/difference-between-ai%EF%BB%BF-machine-learning-and-deep-learning/

- Commonly held preconception: “computers can't do anything that they're not explicitly programmed to do”.
- Arthur Samuel (1959) game of checkers – computer learns to become better than the human programming it by repeatedly playing against itself.
- Machine learning: “the field of study that gives computers the ability to learn without being explicitly programmed”.
- Learning from experience.
- Tom Mitchell (1997) “A computer program is set to learn from an experience E with respect to some task T and some performance measure P if its performance on T as measured by P improves with experience E.”
- Highly interdisciplinary, impact across multiple fields and applications.

Types of data on which ML can be applied in social science research

- ML often used with working with “big data”.
- Characteristics (Lazer and Radford, 2017):
  - Large, heterogeneous, complex, constantly changing.
  - Hard to process with existing software.
  - New tools needed to collect, process and analyse it.
  - Intersection between computer science and social sciences.
- Examples:
  - Social media posts, newspaper articles, party manifestos, party electoral materials, images, videos, blog posts, digital books, location and space, networks, streaming, campaign contributions, CDR (call detail record), sensor data.

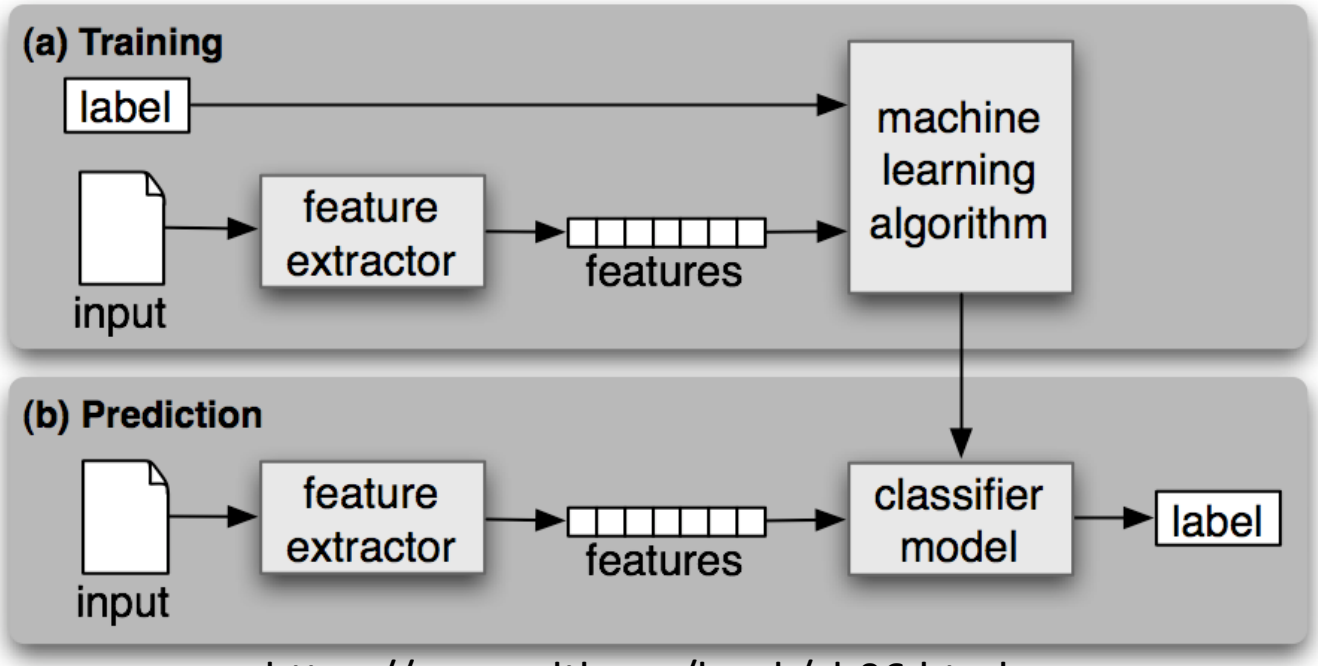
Supervised vs. unsupervised learning

Supervised learning

- Dataset in which the output is known.
- Pre-coded (pre-labelled) data - expert coding of a subset of data.
- The algorithms learns the relation between the input and the output.
- Use it to predict labels for new data.
- Two broad types of problems:
  - Regression
  - Classification

Unsupervised Learning

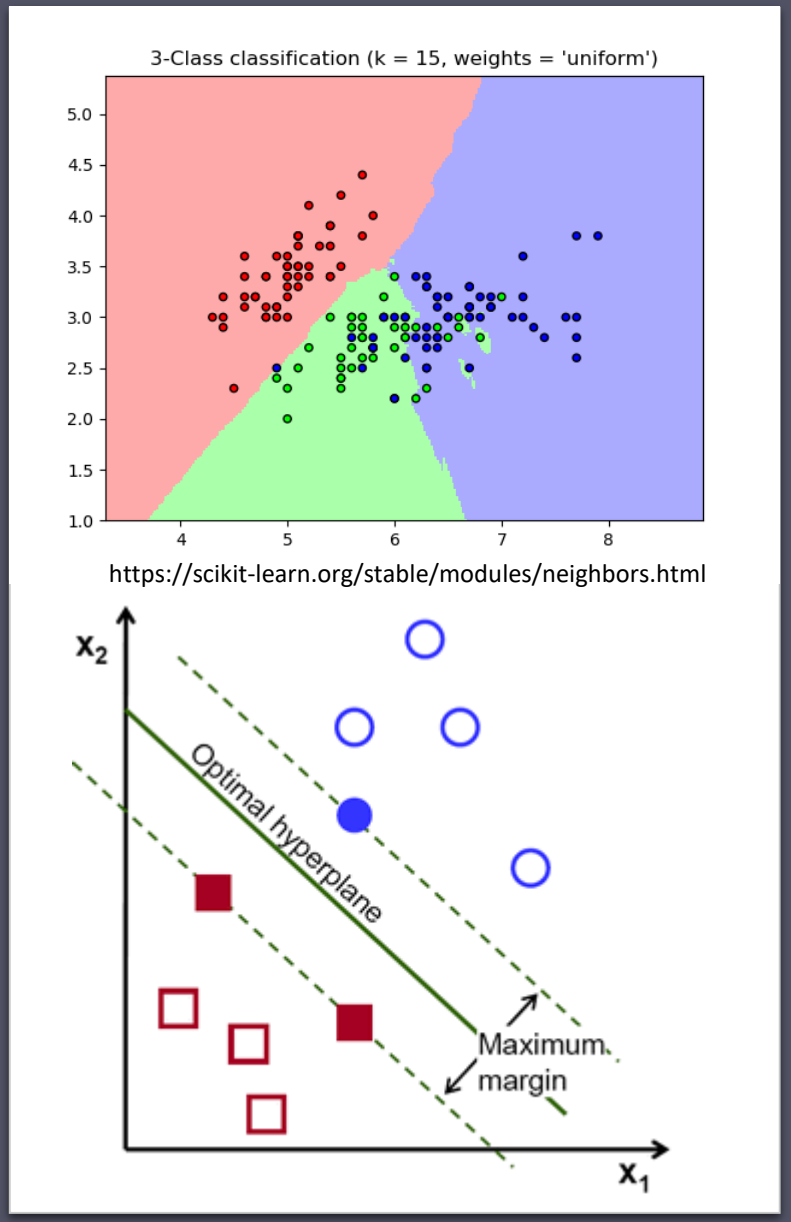
- Dataset in which the output is not known.
- The algorithms learn from the data, deriving the structure from the relations between the input factors (variables).
- Use it to learn data patterns and groupings.
- Types of problems:
  - Clustering
  - Scaling and dimensionality reduction



https://www.nltk.org/book/ch06.html

- Two processes:
  - Training
    - Using labelled data
    - Training a classifier to learn the relations
  - Prediction
    - Using the classifier to predict labels for new data
- At each stage, important decisions before training the classifier:
  - Data pre-processing
  - Feature extraction
  - Feature selection

Supervised classification



https://scikit-learn.org/stable/modules/neighbors.html

https://towardsdatascience.com/support-vector-machine-introduction-to-machine-learning-algorithms-934a444fca47

Examples of supervised classification algorithms

- Naïve Bayes
  - probabilistic classifier based on Bayes theorem
  - determine the probability of the features occurring in each class, return the most likely class
  - strong assumptions regarding independence $\text{posterior} = \frac{\text{prior} \times \text{likelihood}}{\text{evidence}}$
  - performs surprisingly well given its simplicity
- K-Nearest neighbours
  - find a predefined number of training samples closest in distance to the new point, and predict the label from these.
  - use any metric measure for distance
- Support vector machines
  - plot each data item as a point in n-dimensional space where the coordinate values correspond to the value of each feature
  - find the hyper-plane that best differentiate the two classes

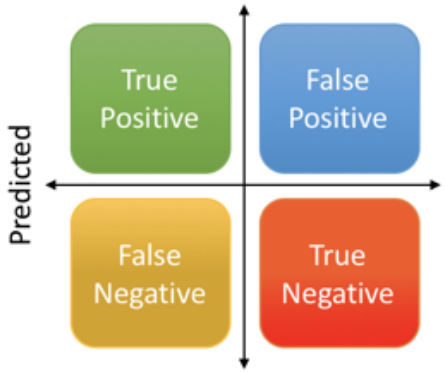
Evaluation

- Spam filter example.
  - Goal: label emails as spam of not.
- Precision:
  - Of those classified as spam, what proportion were actually spam?
  - True positive / (True positive + False positive)
- Recall:
  - Of those that were indeed spam, what proportion were classified that way?
  - True positive / (True positive + False negative)
- Better models have higher values for precision and recall.
  - But there is a tradeoff between precision and recall
- Accuracy
  - What is the proportion of correctly predicted out of total?
- Cross-validation - out-of-sample testing

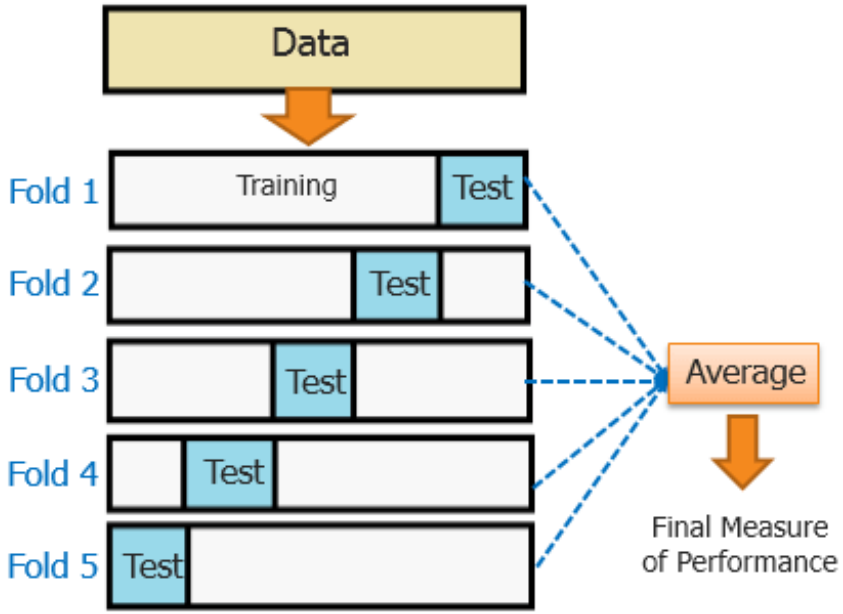
$$\textbf{Precision} = \frac{\text{True Positive}}{\text{Actual Results}} \quad \text{or} \quad \frac{\text{True Positive}}{\text{True Positive + False Positive}}$$

$$\textbf{Recall} = \frac{\text{True Positive}}{\text{Predicted Results}} \quad \text{or} \quad \frac{\text{True Positive}}{\text{True Positive + False Negative}}$$

$$\textbf{Accuracy} = \frac{\text{True Positive + True Negative}}{\text{Total}}$$



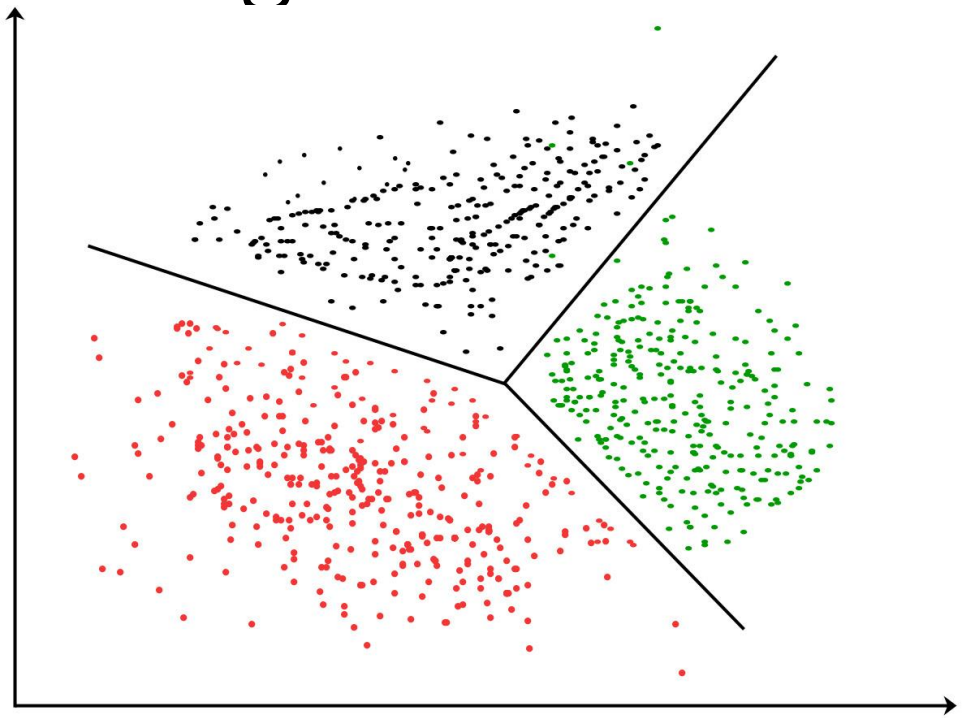
https://towardsdatascience.com/precision-vs-recall-386cf9f89488



https://blog.contactsunny.com/data-science/different-types-of-validations-in-machine-learning-cross-validation

Unsupervised learning - Clustering

- Group observations into a clusters such that:
  - points within each cluster are similar to each other
  - points from different clusters are dissimilar
- High--dimensional space
- Similarity is defined using a distance measure
- Hierarchical
  - repeatedly combine the two “nearest” clusters into one.
- Point assignment:
  - Start with a set of clusters.
  - Place observations into nearest cluster.

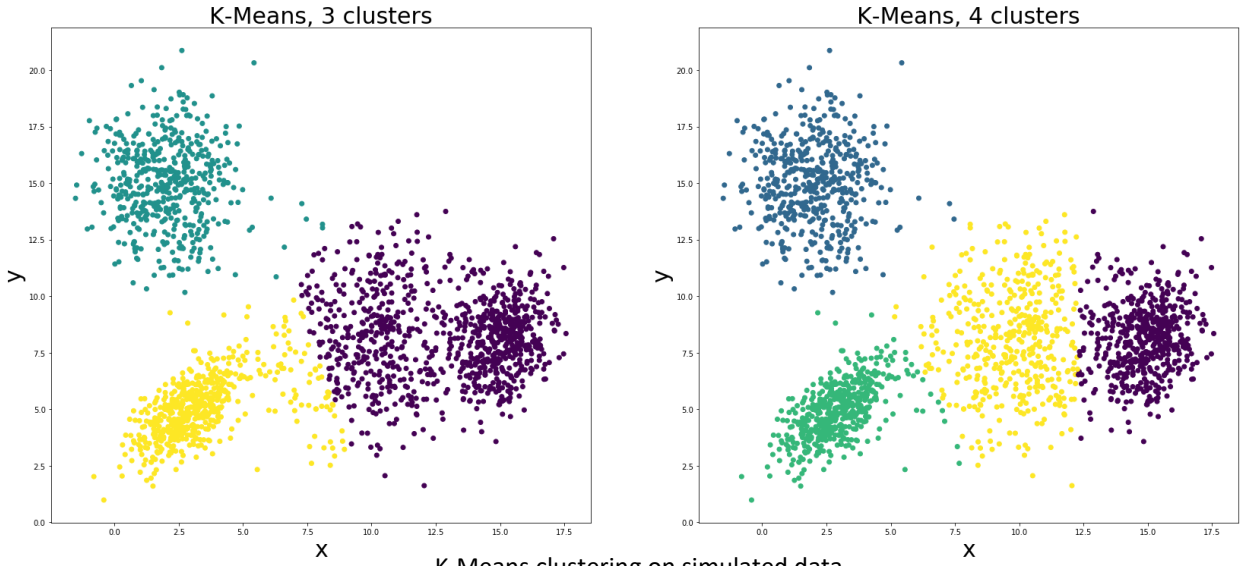


https://www.geeksforgeeks.org/clustering-in-machine-learning/

- Examples: K-means, Mean-shift, Density-based clustering, Expectation–Maximization (EM) Clustering using Gaussian Mixture Models (GMM), Agglomerative.
- Stopping criteria based on cohesion. Homogeneity, completeness and V-measure.
- Evaluation:
  - Rand index, mutual information criteria, Jaccard index.

K-Means example

- Select a number of classes
- Randomly initialize centroids.
- Compute the distance between the point and each group center.
- Classify the point to be in the group whose center is closest to it.
- Recompute the group center by taking the mean of all the vectors in the group.
- Repeat for a set number of iterations or until the group centers don’t change a lot between iterations.
- Procedure can be repeated a number of times.



K-Means clustering on simulated data.

Dimensionality reduction

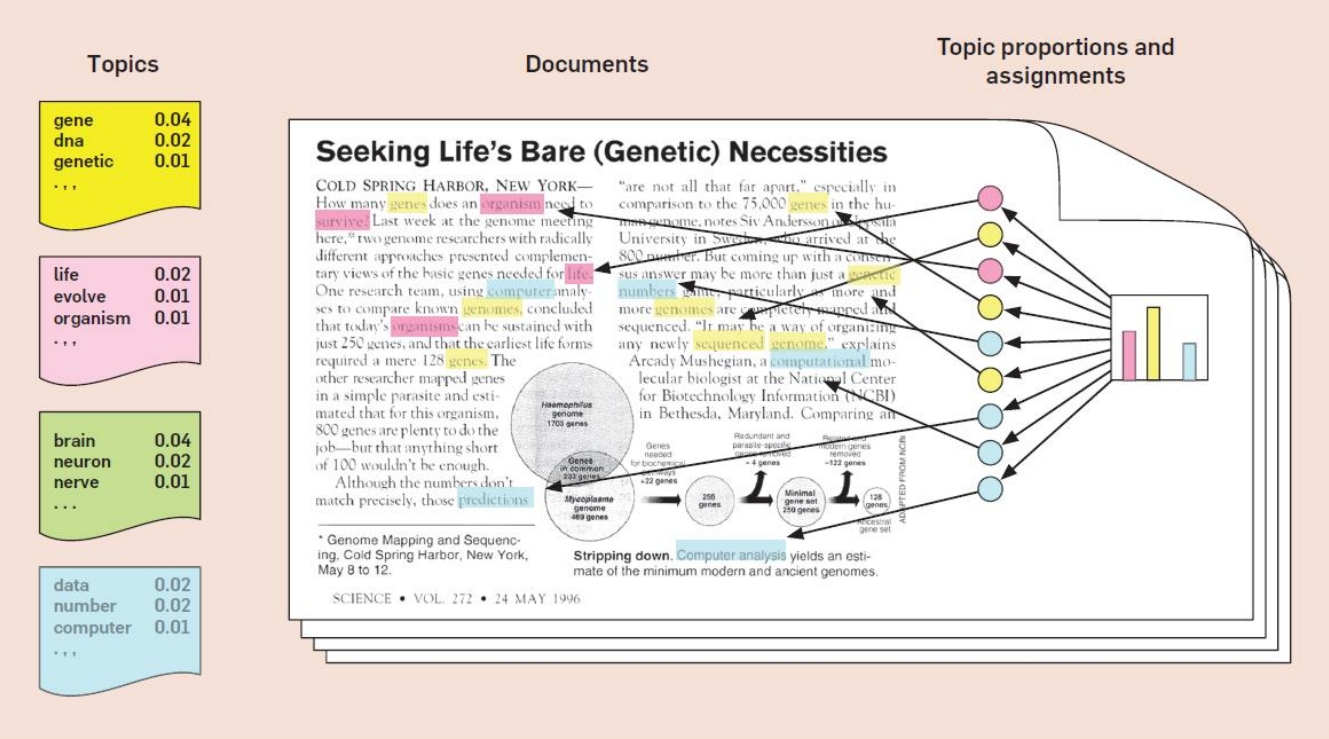
Goal is to summarize data using less information, simplify or visualize highly dimensional data

Principal Component Analysis (PCA)

Multidimensional Scaling (MDS)

Linear Discriminant Analysis (LDA)

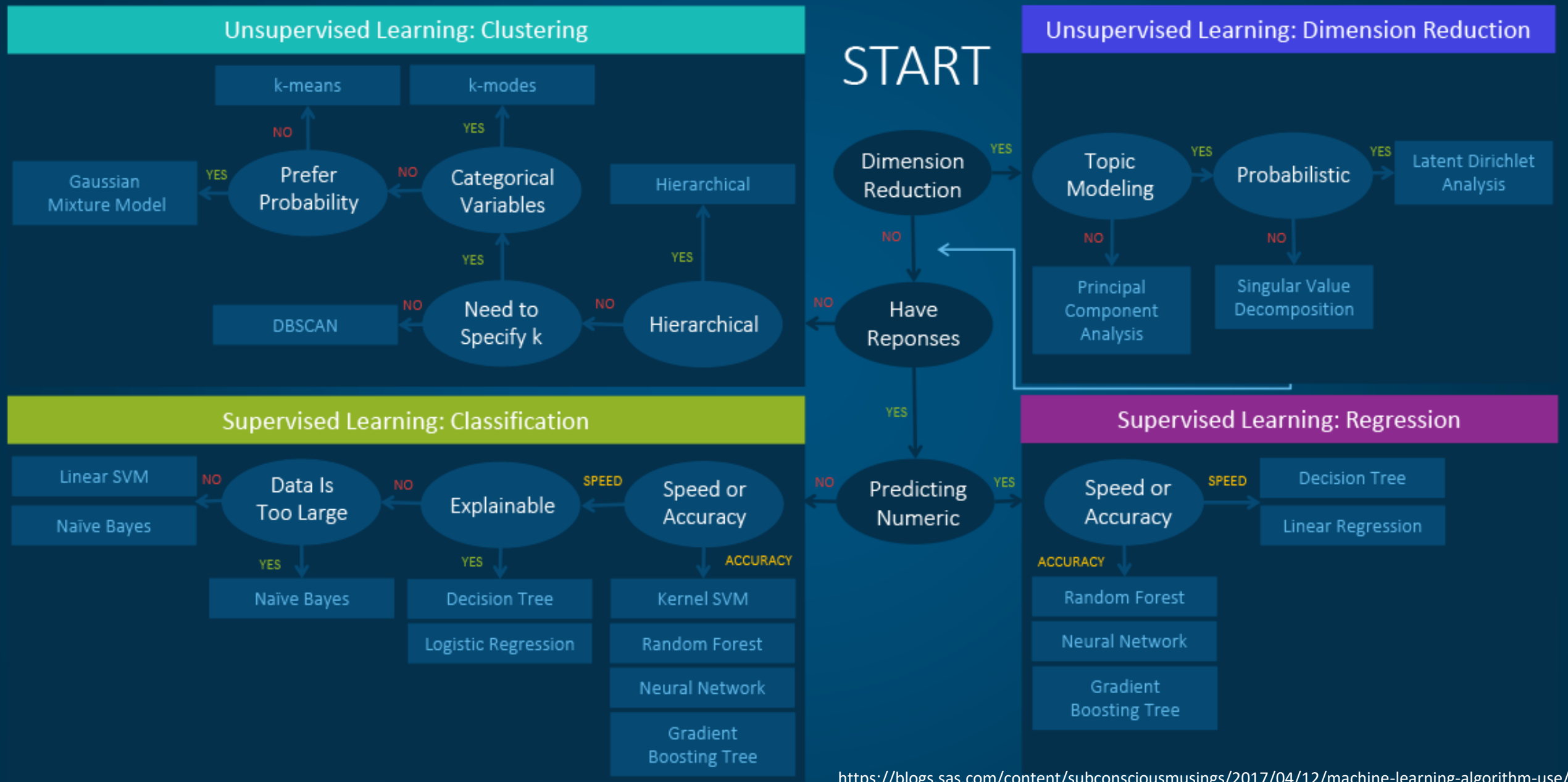
Ideology scaling – DW-Nominate, Wordscores, Wordfish



- Latent Dirichlet Allocation (LDA)
  - Documents are a mixture of topics.
  - Topics generate words based on their probability distribution.
  - Algorithm:
  - Determine number of words in document.
    - Determine mixture of topics in document.
    - Based on the topics' multinomial distribution, assign words to documents.
- Mallet, Python Gensim, R Quanteda.

Topic models (Blei, 2012)

Machine Learning Algorithms Cheat Sheet



https://blogs.sas.com/content/subconsciousmusings/2017/04/12/machine-learning-algorithm-use/

Your ML project (text analysis example)

- Level of existing skills and/or willingness to develop new skills
- What kind of data?
- Is ML needed?
- How often are you planning on using the methods?

Storing ‘big’ data

- As raw files
  - More flexible, but can be slower as you need to write all functions for querying and processing the data
  - Or need to be able to read in large volumes of data
- In a database
  - Relational databases
    - Spreadsheets
    - MySQL – fast, efficient, less storage, but less flexible
  - Non-relational databases
    - NoSQL - MongoDB –a document database very flexible, scalable, but also querying and indexing features that make if easier to work with than raw json.
    - Visualizing with RoboMongo.

Choice of software

R and Python most commonly used by social science researchers.

Python

- better text processing and analysis (Textbolob, Stanford Core NLP, Gensim)
- NLTK – most popular - includes graphical demonstrations and sample data; accompanied by a book that explains the underlying concepts behind the language processing tasks supported by the toolkit, plus a cookbook
- Better machine learning libraries - Scikit-learn https://scikit-learn.org/

R – better statistical analysis packages.

- R has the newly developed Quanteda for political scientists
- Multiple packages for machine learning: https://cran.r-project.org/web/views/MachineLearning.html

Both catching up fast, less differences.

No/less programming: Google AutoML; Orange; Knime; DLNC; Data Robot.

Jupyter Notebook and R Notebook

- Both environments that allows you to create and share documents that contain code, results, graphs, explanatory text and equations.
- Jupyter which was initially called IPython notebook is the standard choice for Python users
- R Notebook is an R Markdown document with chunks that can be executed independently and interactively, with output visible immediately beneath the input.
- Using R Notebook in an RStudio environment is preferred by R users.
- Both supports both types of programming languages, and others
- Both allow you to have the code and results on the screen inline.
- Both integrate well with Github.
- Rmarkdown slightly better for version control.
- Jupyter slight advantage in number of languages implemented.

First step - processing and cleaning data, dealing with encodings

Combination of Pandas (great encoding functions – deal with strange characters or foreign – common on social media or other types of online data) and NLTK in Python

Common tasks include removing most common words (stopwords and, to for, from), tokenizing the corpus.

Corpus - Body of text, multiple text are called corpora.

Token – A single "entity“ in text, based on what rule you use for splitting the text. Most commonly words, but you can also tokenized the sentences out of a paragraph.

Different tokenizers functions

Other useful functions

- Stemming algorithms work by cutting off the end or the beginning of the word, taking into account a list of common prefixes and suffixes that can be found in an inflected word.
- Studies - studi, studying – study

- Lemmatization, on the other hand, is based on detailed dictionaries which the algorithm can look through to link the form back to its lemma.
- Both studies and studying turned to study.
- Both take time, lemmatization more complex.

Counts and searches

Count Word Frequency – FreqDist

Searches with:
Keywords
Regular expressions

Search for keywords and show where they appear in the corpus.

Use more complex searcher based on regular expressions

You often need to do this to identify specific search patterns and capturing what you want and not more than you want

In this example count how many times reference to the UK appear in the text, but making sure we don’t capture Ukraine as well.

Measuring similarity between documents

- Cosine similarity.
- Evaluate how similar texts are based on the words included, regardless of word order.
- We tokenize the texts, take the vector of counts and calculate the cosine of the angle between them.
- So for any two pairs of texts we can calculate the cosine similarity and then identify in a large collection the documents that are most similar.

Machine learning methods for the analysis of social media data include

- Supervised methods - Such as supervised classifiers which allow us to label social media posts and put them into categories.
  - For example, we have posts that are related to Brexit and posts which are not. We can core a sample of our data on these two categories. We then train a supervised classifier to predict if the rest of the posts are about Brexit or not.
- Unsupervised methods - Such as clustering algorithms, which aims to group together similar social media posts based on the language, but without human coding.

What does the text of the posts talk about?

- Relative word frequencies
- Plot the using word clouds, although not the most informative method.
- If we already know the categories we can used supervised classification.
- If not, clustering.
- Or topic modelling.

Resources: Textbooks

- Hastie, Tibshirani and Friedman. The Elements of Statistical Learning: Data Mining, Inference, and Prediction.
- James, Witten, Hastie and Tibshirani. An Introduction to Statistical Learning with Applications in R
- Bishop. Pattern Recognition and Machine Learning.
- Murphy. Machine Learning: A Probabilistic Perspective.
- Guido and Mueller. Introduction to Machine Learning in Python.
- Raschka and Mirjalili. Python Machine Learning.

Resources: Online tutorials

- Intro to programming and data science
- Python: https://github.com/iuliacioroianu/python_workshop
- R: https://github.com/iuliacioroianu/Intro-to-R

- Quantitative text analysis:
- Python: https://github.com/iuliacioroianu/NLTK_examples
- R: https://kenbenoit.net/TAUR-Hugo/readme/

- Machine learning examples
- Python: https://scikit-learn.org/stable/auto_examples/index.html
- R: https://lgatto.github.io/IntroMachineLearningWithR/index.html

- Other resources and tutorials freely available online. Github search is a good starting point.

Resources: Applications and examples

- Baturo, A., Dasandi, N., & Mikhaylov, S. J. (2017). Understanding state preferences with text as data: Introducing the UN General Debate corpus. *Research & Politics*, *4*(2), 2053168017712821.
- Baumgartner, F. R., Jones, B. D., & Wilkerson, J. (2011). Comparative studies of policy dynamics. *Comparative Political Studies*, *44*(8), 947–972.
- Beauchamp, N. (2017). Predicting and interpolating state-level polls using Twitter textual data. *American Journal of Political Science*, *61*(2), 490–503.
- Benoit, K., Conway, D., Lauderdale, B. E., Laver, M., & Mikhaylov, S. (2016). Crowd-sourced text analysis: Reproducible and agile production of political data. *American Political Science Review*, *110*(2), 278–295.
- Bonica, A. (2018). Inferring Roll-Call Scores from Campaign Contributions Using Supervised Machine Learning. *American Journal of Political Science*, *62*(4), 830–848.
- Burscher, B., Vliegenthart, R., & De Vreese, C. H. (2015). Using supervised machine learning to code policy issues: Can classifiers generalize across contexts? *The ANNALS of the American Academy of Political and Social Science*, *659*(1), 122–131.
- Colleoni, E., Rozza, A., & Arvidsson, A. (2014). Echo chamber or public sphere? Predicting political orientation and measuring political homophily in Twitter using big data. *Journal of Communication*, *64*(2), 317–332.
- Desmarais, B. A., Harden, J. J., & Boehmke, F. J. (2015). Persistent policy pathways: Inferring diffusion networks in the American states. *American Political Science Review*, *109*(2), 392–406.
- Gentzkow, M., Kelly, B. T., & Taddy, M. (2017). *Text as data*. National Bureau of Economic Research.
- Grimmer, J. (2015). We are all social scientists now: How big data, machine learning, and causal inference work together. *PS: Political Science & Politics*, *48*(1), 80–83.
- Grimmer, J., & Stewart, B. M. (2013). Text as data: The promise and pitfalls of automatic content analysis methods for political texts. *Political Analysis*, *21*(3), 267–297.
- Huff, C., & Kertzer, J. D. (2018). How the public defines terrorism. *American Journal of Political Science*, *62*(1), 55–71.
- King, G., Pan, J., & Roberts, M. E. (2013). How censorship in China allows government criticism but silences collective expression. *American Political Science Review*, *107*(2), 326–343.
- Mueller, H., & Rauh, C. (2018). Reading between the lines: Prediction of political violence using newspaper text. *American Political Science Review*, *112*(2), 358–375.
- Wilkerson, J., Smith, D., & Stramp, N. (2015). Tracing the flow of policy ideas in legislatures: A text reuse approach. *American Journal of Political Science*, *59*(4), 943–956.