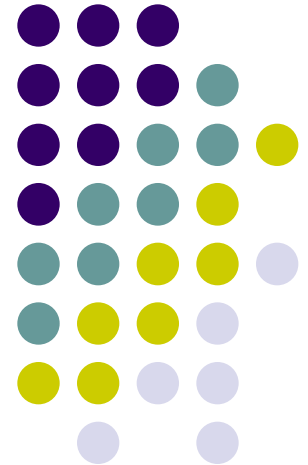


Applied Scaling & Classification Techniques in Political Science

Lecture 1 (first part)
An introduction to textual analysis
methods

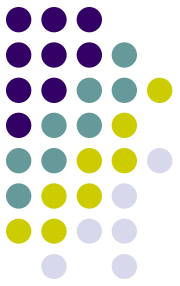




References

- ✓ Grimmer, Justin, and Stewart, Brandon M. (2013). Text as Data: The Promise and Pitfalls of Automatic Content Analysis Methods for Political Texts. *Political Analysis*, 21(3): 267-297
- ✓ Benoit, Kenneth (2020). Text as data: An overview. In Luigi Curini and Robert Franzese (eds.), *SAGE Handbook of Research Methods in Political Science & International Relations*, London, Sage, chapter 26

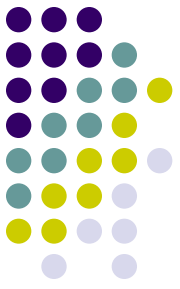
It all began with...



Language is the **medium** for politics and political conflict!

Some examples:

- *Candidates* **debate and state** policy positions during a campaign
- *Representatives* **debate** legislation
- *Nations* negotiate and sign **treaties**, with language that signals the motivations and relative power of the countries involved
- *Terrorist groups* reveal their preferences and goals through **public statements**
- *Parties and citizens* discuss about politics **on-line**



It all began with...

It is no therefore exaggeration to consider text as “the most pervasive - and certainly the most persistent artifact of political behavior”

As a result, to understand what politics is about we need (**quite often**) to know what political actors (but also citizens) are *saying and writing*

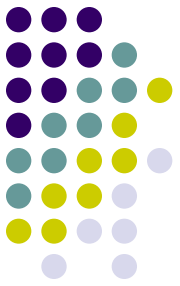
Recognizing that language is central to the study of politics is **not new**...

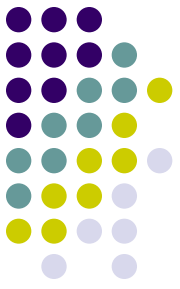
...however scholars have struggled when using texts to make inferences about politics!

It all began with...

Why? **Volume matters!** There are simply too many political texts out there!

Rarely scholars are able (time/resources constrain!) to manually read all the texts



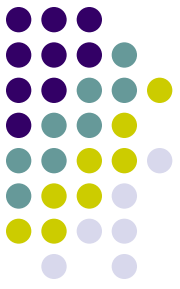


It all began with...

Recent methods have made progress by breaking from traditional (human) content analysis to treat text:

- not as an *object for subjective interpretation*, but...
- ...as *objective data from which information about the author can be estimated...i.e., **treating words as data!***

What do we mean by that?



It all began with...

Text is often referred to as “**unstructured data**”, because it is structured not for the purposes of serving as any form of data but rather structured according to the rules of language

Because “data” means, in its simplest form, information collected for use, **text starts to become data** when we record it for reference or analysis, and this process always involves imposing **some abstraction or structure that exist outside the text itself**

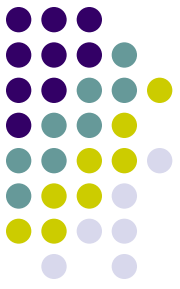


It all began with...

Absent the **imposition of this structure**, the text remains **informative** - we can read it and understand (on some form) what it means - but it does not provide a **form of information**

That is, **treating texts-as-data** means arranging texts for the purpose of analysis, using a structure that probably was not part of the process that generated the data itself, to make texts amenable to the familiar tools of data-analysis

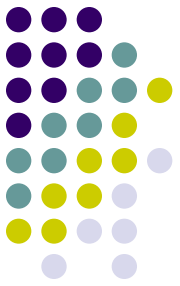
It all began with...



Treating texts-as-data enables therefore the use of **statistical methods**, allowing inferences to be drawn about observable (and unobservable) underlying characteristics of a text's author and through that...

...it can make possible the previously impossible in political science: **the systematic analysis** of large-scale text collections that facilitates substantively important inferences about politics

It all began with...

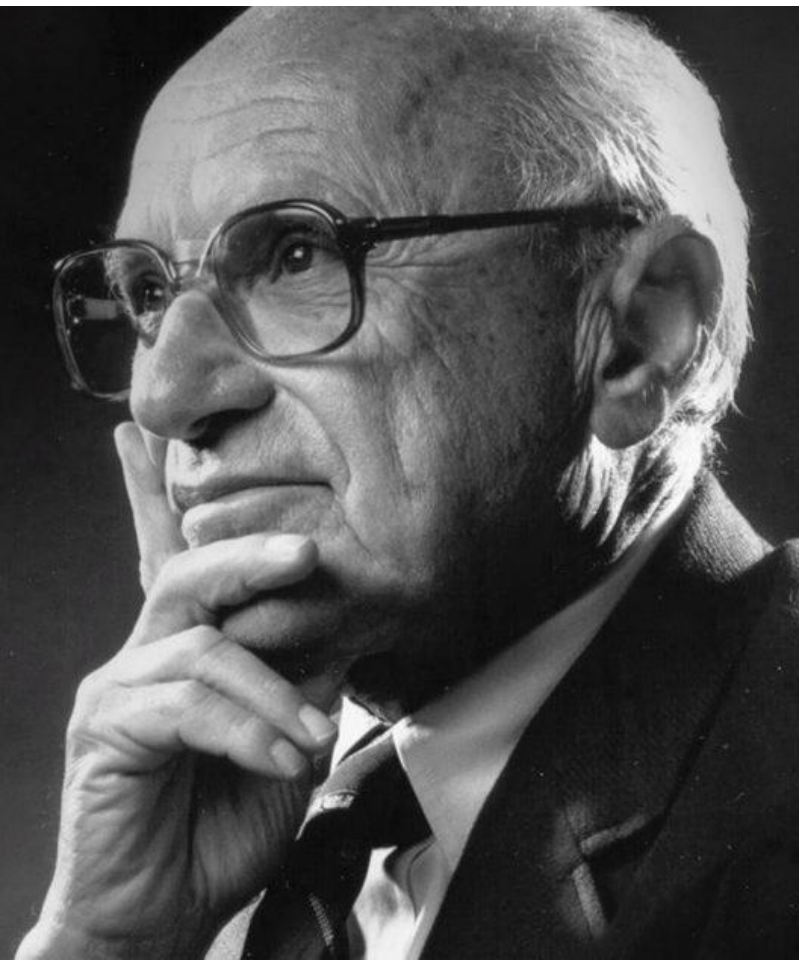


OK OK you convinced me! But how to that?!? Can we now finally move to the beef?!?

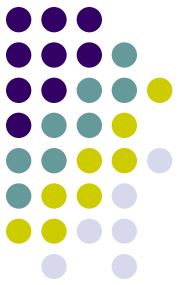
Yes...but before doing that...beware...

A Milton Friedman favorite political aphorism:

“There’s no
such thing
as a free lunch.”



Just as no body escapes Newton’s laws, no technique can escape the following fundamental principles of text analysis



Four principles of Automated Text Analysis to keep in mind (as social scientists!)

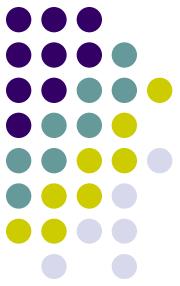


1) All quantitative models of language are wrong – but some are useful



Colourbox

The first principle



Data generation process for any text is a **mystery**

If a sentence has complicated structure, its meaning could change drastically with the inclusion of new words (or punctuation...)

The first principle

The Sibyl

“ibis, redibis, non morieris in bello”

vs.

“ibis, redibis non, morieris in bello”

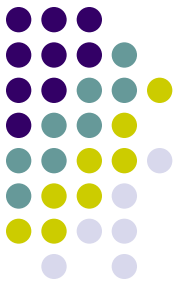


The first principle

The **complexity of language** implies that all methods necessarily fail to provide an **accurate account of the data-generating process** used to produce texts

That all automated methods are based on **incorrect models of language** implies that the models **should be evaluated** based on their ability to perform some useful social scientific task

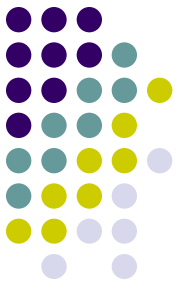




2) Quantitative methods amplify humans, not replace them



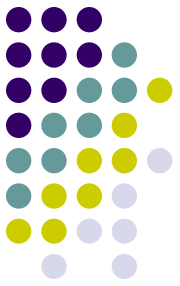
The second principle



The **complexity of language** implies that automated content analysis methods will never replace careful and close reading of texts

Rather, such methods are best thought of as **amplifying careful reading and thoughtful analysis**

Researchers still guide the process, make modeling decisions, and interpret the output of the models



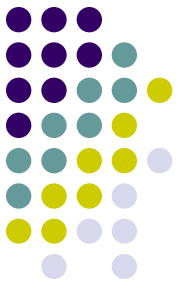
«The best technology is **human-empowered** and **computer-assisted**»
(Gary King, Harvard University)



3) There is no a best method for automated text analysis



The third principle



Different datasets and different research questions often lead to different quantities of interest. This is particularly true with text models!

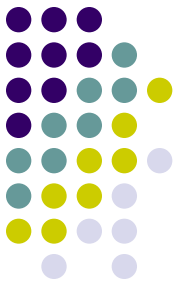
We should simply acknowledging that there are **different research questions** and designs that imply **different types of models**



4) Validate,
validate, validate



The fourth principle



As already told, the **complexity of language** implies that automated content methods are **incorrect models** of language

This means that the performance of any one method on a new data set cannot be guaranteed, and therefore **validation** is essential when applying automated content methods

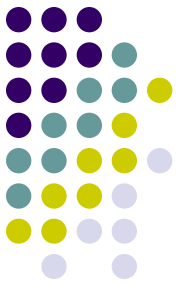
We will discuss about validation a lot!

What should be avoided, then, is the **blind use of any method** without a validation step

Having that in mind...

As social scientists, we have two options in front of us...





Having that in mind...

Either we do like the ostrich...i.e., we can simply ignore the amount of digital textual data currently available for our research!





Having that in mind...

Or...we can do like Galileo with his telescope: finding **new** patterns in **new** data, with **new** methods (telescope?) available, and, in the best scenario, developing **new** theories thanks to that! After all, the telescope came before Galileo's astronomical theories...

