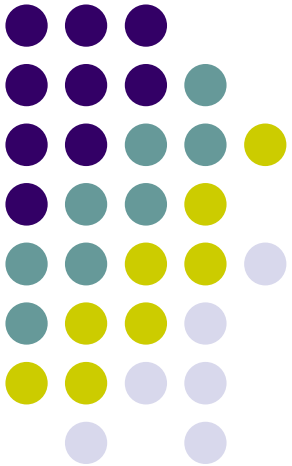
Big Data Analytics

Lecture 1 (first part)
An introduction to text analytics



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 is Political Science & International Relations*, London, Sage, chapter 26
- ✓ Grossman, Jonathan, and Pedahzur Ami (2020). Political Science and Big Data: Structured Data, Unstructured Data, and How to Use Them, *Political Science Quarterly*, 135(2): 225-257

Big Data

The amount of data generated as a by-product in society is growing fast including data from satellites, sensors, transactions, social media and smartphones, just to name a few

"*Every day 2.5 quintillions* of bytes are being created…so much that the 90% of today's available data have been created in the last 2 years*" IBM, January 2012

*=1,000,000,000,000,000,000

Big Data

Such data are often referred to as ***Big Data***

While the information overload is not a recent phenomenon, the term Big Data is relatively new

It first emerged in the information technology industry in the mid-1990s and made its academic debut in a 1998 computer science paper. In the two decades that followed, it gained popularity rapidly

Big Data

Big Data pose several interesting and new challenges to those who want to extract information from data

And indeed, we use the expression *Big Data* and not *Big Information*, because there is a lot of work for analysts before information can be gained from “auxiliary traces of some process that is going on in society”

But what are Big Data from a more formal point of view?

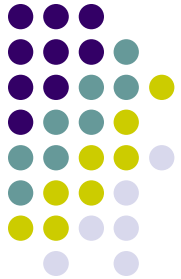
Big Data

What Big Data <u>are not</u>

Big Data are not **<u>just</u>** a data collection with a very large-N

That is, a very *large survey of citizen participation* cross-nationally is **not**, strictly speaking, Big Data

Big Data



What Big Data are: 3 main attributes should be present (at the same time!)

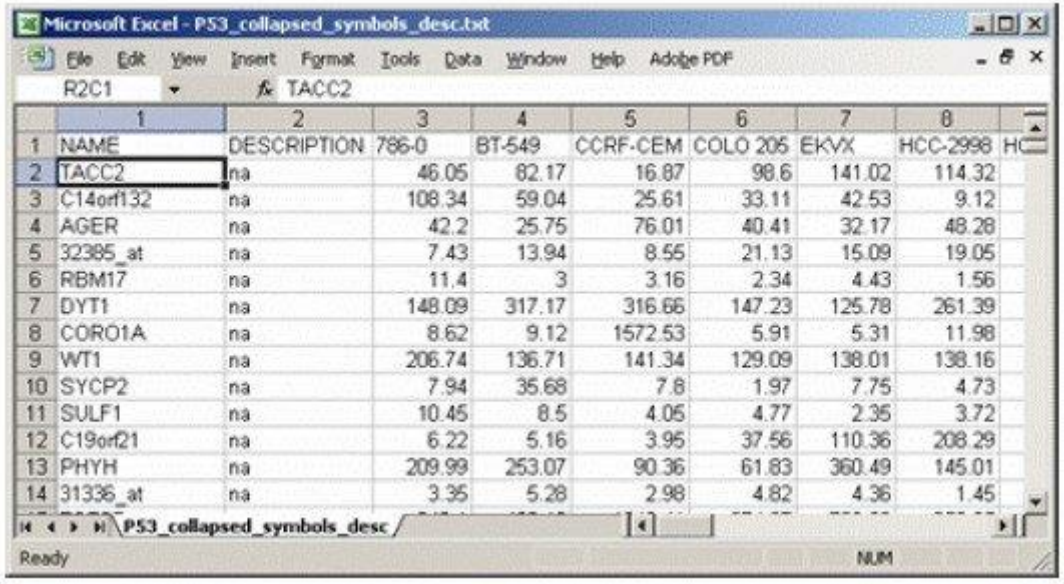
The most common definitions are in fact based on the "**three Vs**" framework

1. *volume* (the sheer size of the data set is large)
2. *velocity* (data are produced in or almost in real time, i.e., size per unit of time matters)
3. *variety* (data come in different types and formats and may be structured or, more often, unstructured)

This last property is a crucial one in terms of the challanges it provides

Big Data

For most people, the word “data” means this…



	1	2	3	4	5	6	7	8	
1	NAME	DESCRIPTION	786-0	BT-549	CCRF-CEM	COLO 205	EKVX	HCC-2998	HC
2	TACC2	na	46.05	82.17	16.87	98.6	141.02	114.32	
3	C14orf132	na	108.34	59.04	25.61	33.11	42.53	9.12	
4	AGER	na	42.2	25.75	76.01	40.41	32.17	48.28	
5	32385_at	na	7.43	13.94	8.55	21.13	15.09	19.05	
6	RBM17	na	11.4	3	3.16	2.34	4.43	1.56	
7	DYT1	na	148.09	317.17	316.66	147.23	125.78	261.39	
8	CORO1A	na	8.62	9.12	1572.53	5.91	5.31	11.98	
9	WT1	na	206.74	136.71	141.34	129.09	138.01	138.16	
10	SYCP2	na	7.94	35.68	7.8	1.97	7.75	4.73	
11	SULF1	na	10.45	8.5	4.05	4.77	2.35	3.72	
12	C19orf21	na	6.22	5.16	3.95	37.56	110.36	208.29	
13	PHYH	na	209.99	253.07	90.36	61.83	360.49	145.01	
14	31336_at	na	3.35	5.28	2.98	4.82	4.36	1.45	

These are structured data - data that can fit squarely into a table, where every row is an observation, every column a variable, and the cells at the intersection of rows and columns contain values

Big Data

Data of this kind, however, are but a fraction of the total amount of data in the world. According to different estimates, 80 to 95 percent of existing data are unstructured data, that is, data that cannot fit easily snugly into rows and columns

Unstructured data may take the form of text, audio, video, or any other observable manifestation (usually in a digital format)

The content of a political speech, the video recording of that speech, the blog post commenting on the video, and the academic article analyzing the post are all unstructured

Big Data

Because of the messy and eclectic nature of unstructured data, attempts to investigate them with conventional statistical methods would often be futile

To analyze such unruly data in a quantitative way, one needs to impose a **structure** upon them

"*In analytics there is no such thing as unstructured data, just data that structure has not yet been applied to*"

Big Data

Texts are probably the most common source of Big Data in political and social science

And it is precisely on them (that is on text-as-data approach) that we are going to focus our course

Big Data

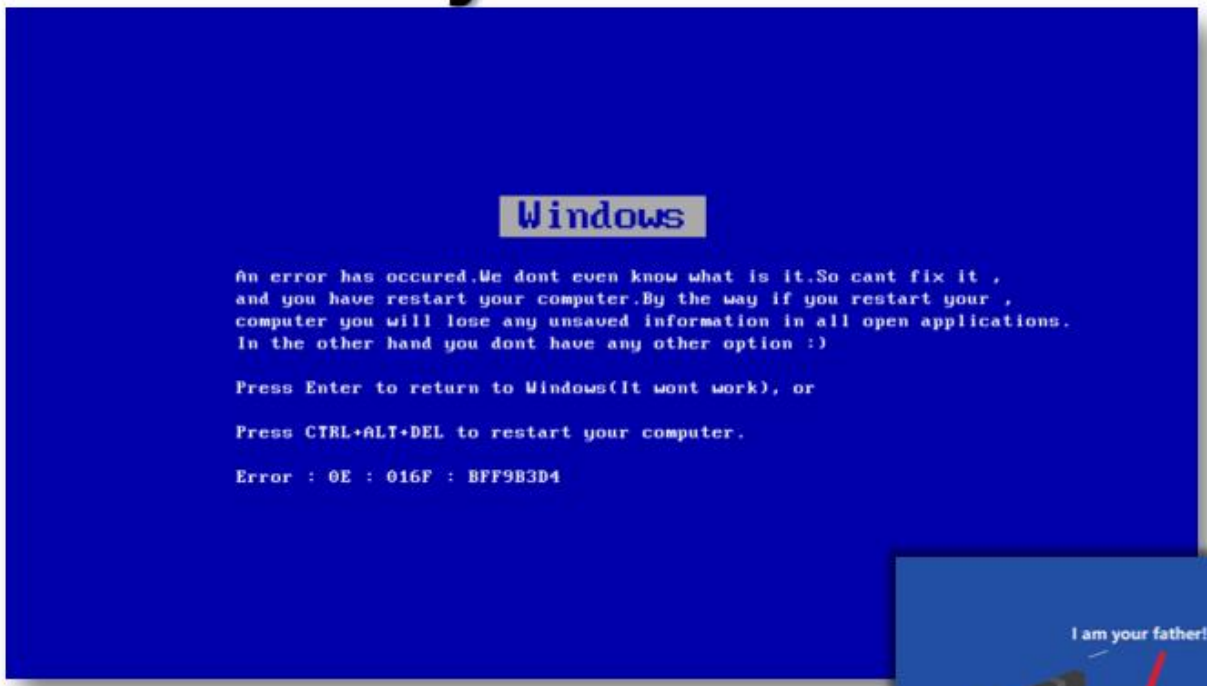
Summing up

The incredible quantity of Big Data tests scholars' aims to conduct research with measurable goals. Not only is the volume of data overwhelming, but data can also be misleading and onerous

Finding the evidence that we need in an ocean of information is a constant challenge. Even if we encounter some details that seem pertinent, we still have to verify their accuracy

Big Data

Do they work at all?



The blue side of Big data

Big Data

In this respect, the Big Data revolution is a blessing and a curse, as it magnifies and exacerbates reliability problems that have always been part of any inquiry

Having that in mind…either we do like the ostrich…i.e., we can simply ignore the amount of new data currently available for our research!

Big Data

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…

Big Data

So let's keep this in mind! And let's start our journey in text-analytics!

It all began with…

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

Some examples:

- *Candidates* **state** policy positions during a campaign
- *Representatives* **debate** legislation
- *Parties* and *citizens* **discuss** about politics **on-line**
- *Terrorist groups* reveal their preferences and goals through **public statements**

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**…

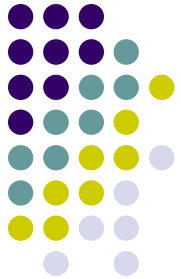
…however scholars have struggled when using texts to make inferences about politics!

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…

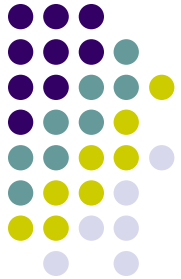


Recent methods have made progress by breaking from traditional (human) content analysis to treat text:

- 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**" (does this term ring you a bell?!?), 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 what it means - but it does not provide a **form of information**

That is, **treating texts-as-data** means:

1. arranging texts for the purpose of analysis, using a structure that probably was not part of the process that generated the data itself
2. through that, making texts amenable to the familiar tools of data-analysis

This make possible what was previously impossible: **the systematic analysis** of large-scale text collections that facilitates substantively important inferences about politics & society

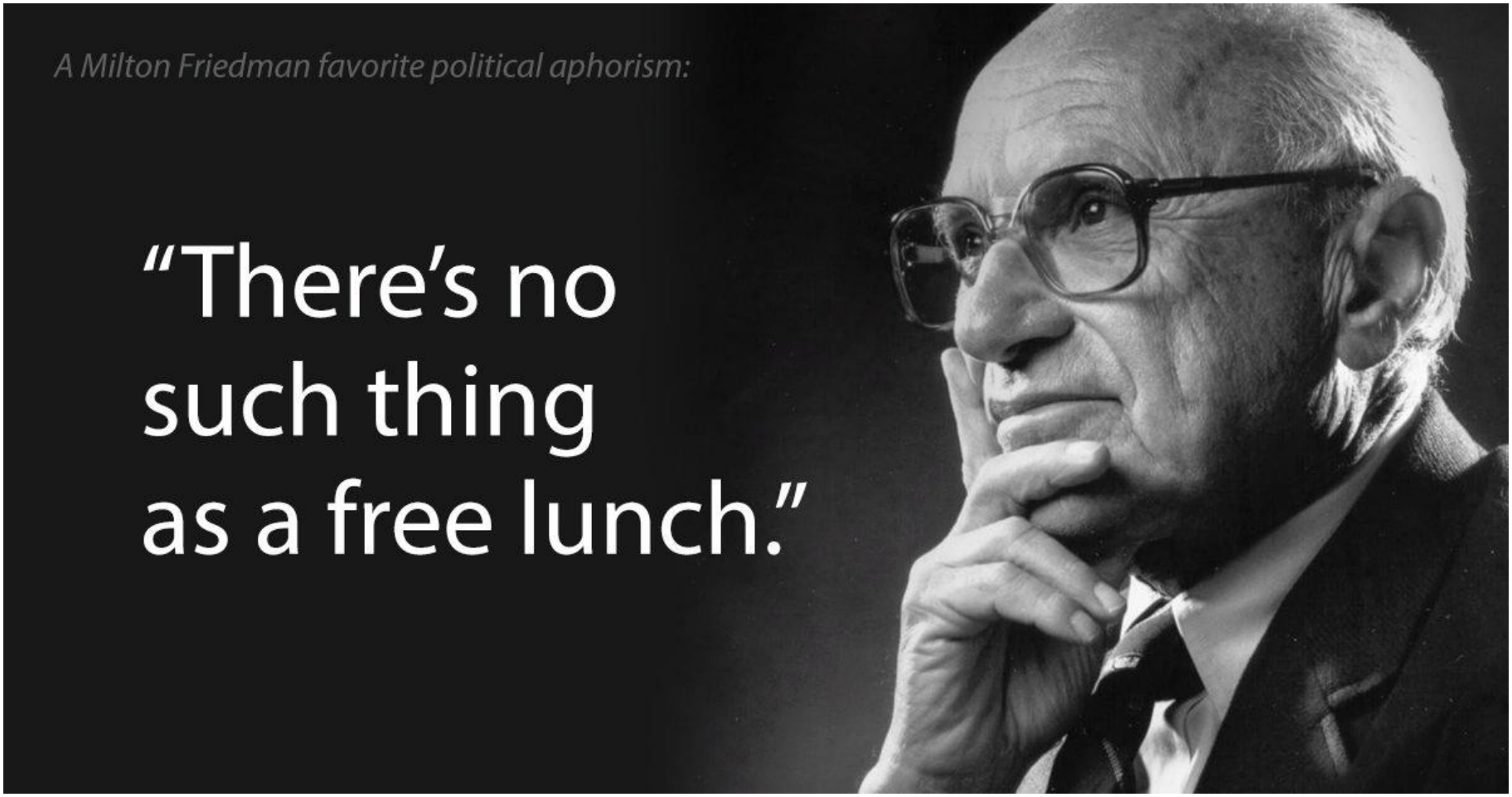
It all began with…

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

Yes...but before doing that…**beware!!!**

The opportunities afforded by vast electronic text archives and algorithms for text analysis are in a real sense unlimited

Yet in a rush to take advantage of the opportunities, it is easy to overlook some important questions and to underappreciate the consequences of some decisions



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

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

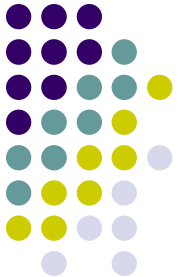
The Sibyl

"*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** *therefore* 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

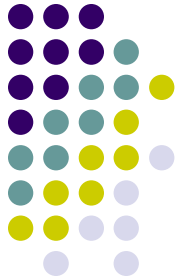
Different datasets and different research questions often lead to different quantities of interest. This is particularly true with text models!

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

As a result, every research question and every text-as-data enterprise is **unique**. Analysts should do their own testing to determine how the decisions they are making affect the substance of their conclusions, and be mindful and transparent at all stages in the process

4) Validate, validate, validate

The fourth principle



As 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

For analysts using text as data, there are decisions at every turn, and even the ones we assume are benign may have meaningful downstream consequences!!!