Texto como data

Fernanda Sobrino

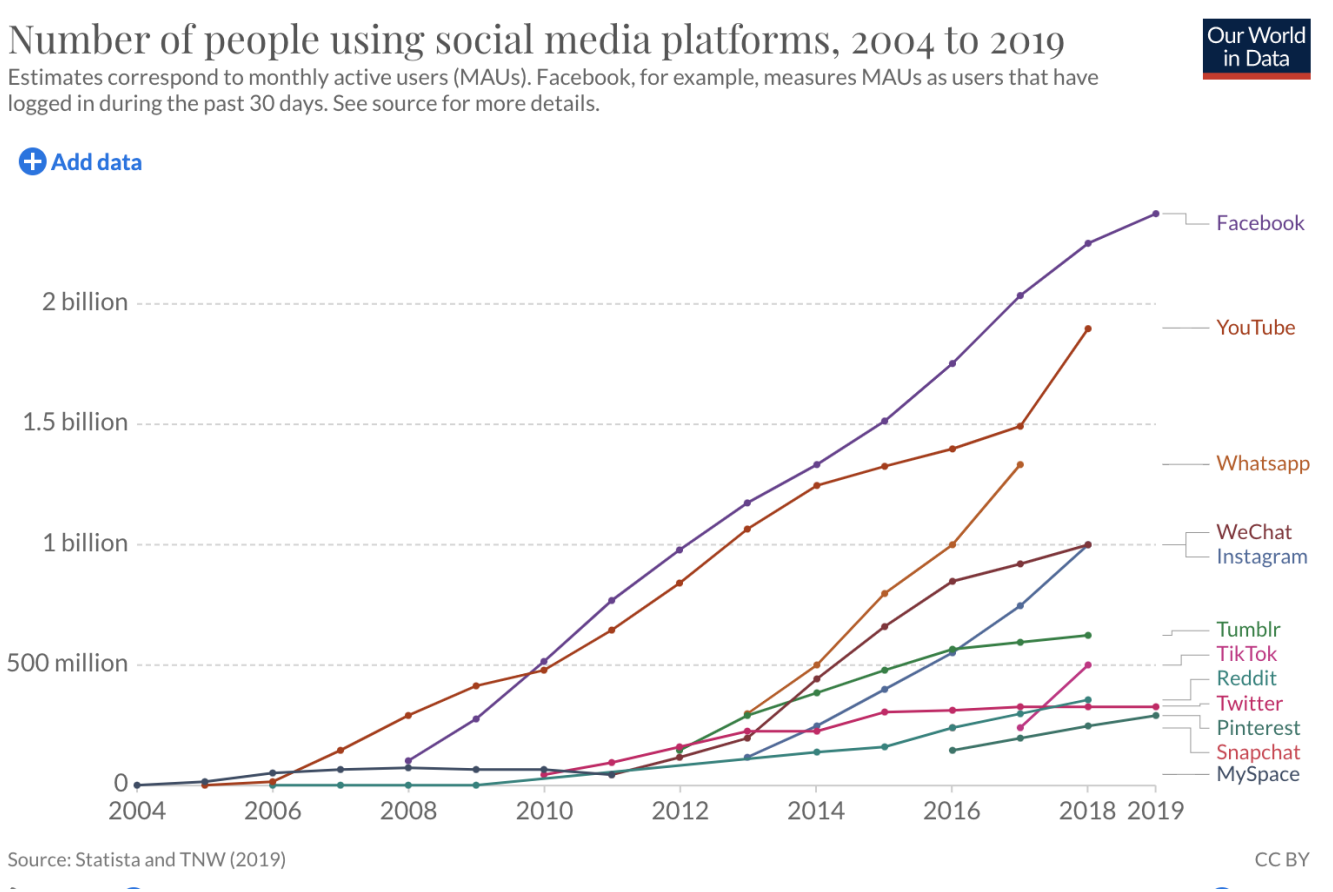
6/19/2021

Introducción

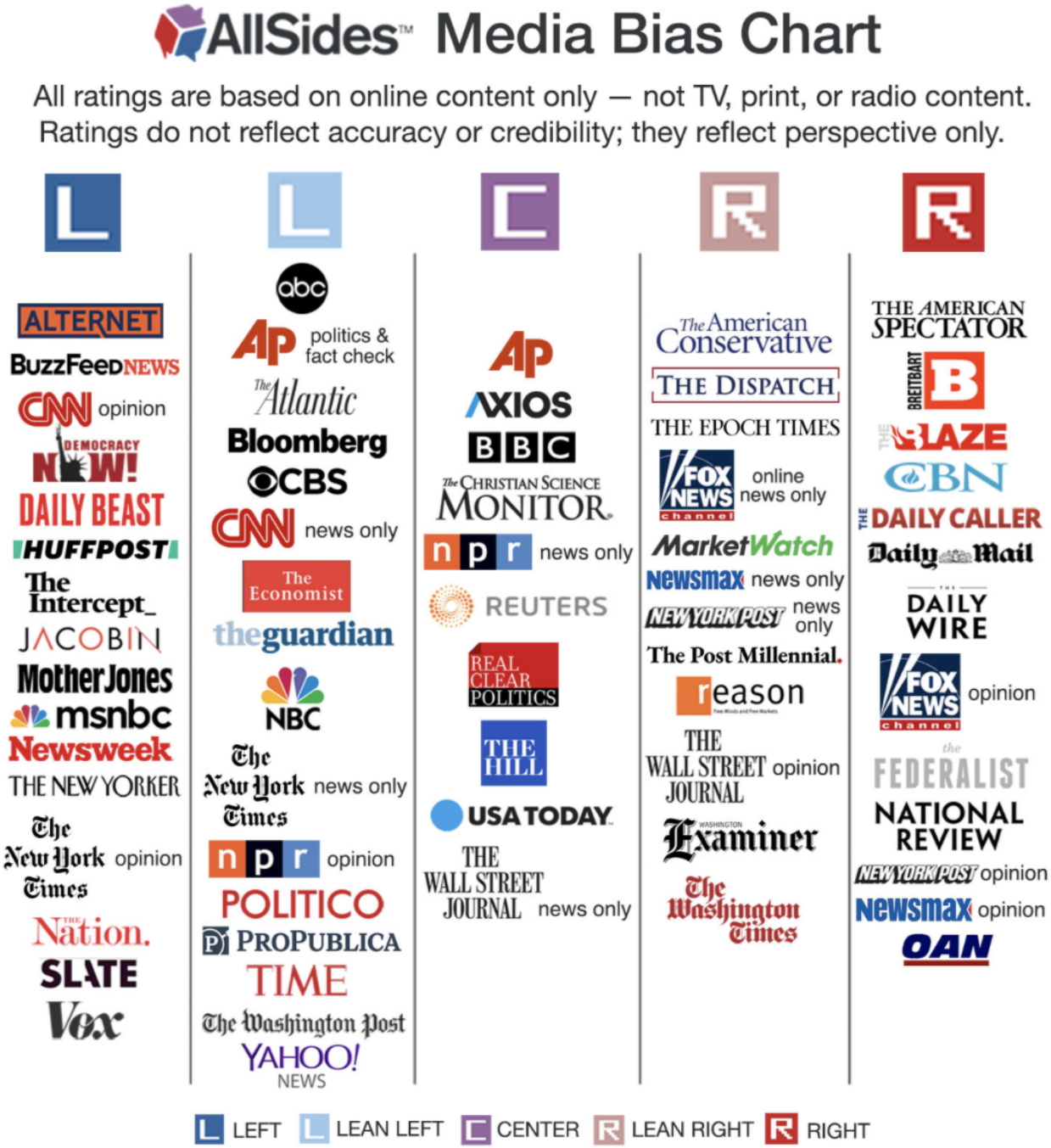
Historicamente (pre 2000)

- las interacciones sociales se dan por medio de texto
- las ciencias sociales han evitado estudiar este tipo de datos
- Por qué?
  - difíciles de conseguir
  - pérdida de tiempo
  - no generalizables (formatos distintos, etc)
  - difícil de guardar y alcanzar (archivos históricos)
  - intensivos computacionalmente

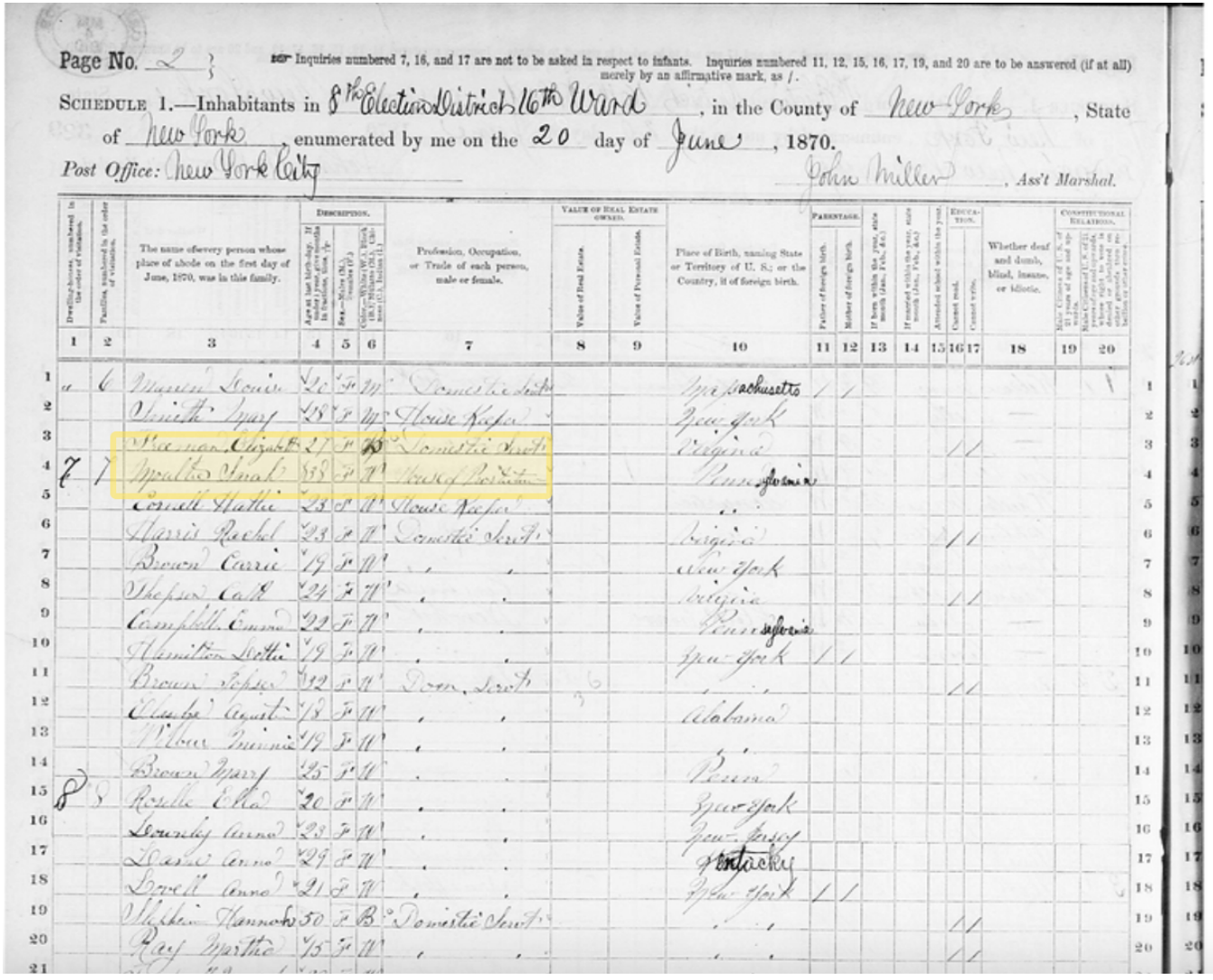
Post 2000: Redes Sociales



Post 2000: Medios tradicionales



Post 2000: Archivos históricos



Page No. 2

Inquiries numbered 7, 16, and 17 are not to be asked in respect to infants. Inquiries numbered 11, 12, 15, 16, 17, 19, and 20 are to be answered (if at all) merely by an affirmative mark, as /.

SCHEDULE 1.—Inhabitants in 8th Election District 16th Ward, in the County of New York, State of New York, enumerated by me on the 20 day of June, 1870.

Post Office: New York City

John Miller, Ass't Marshal.

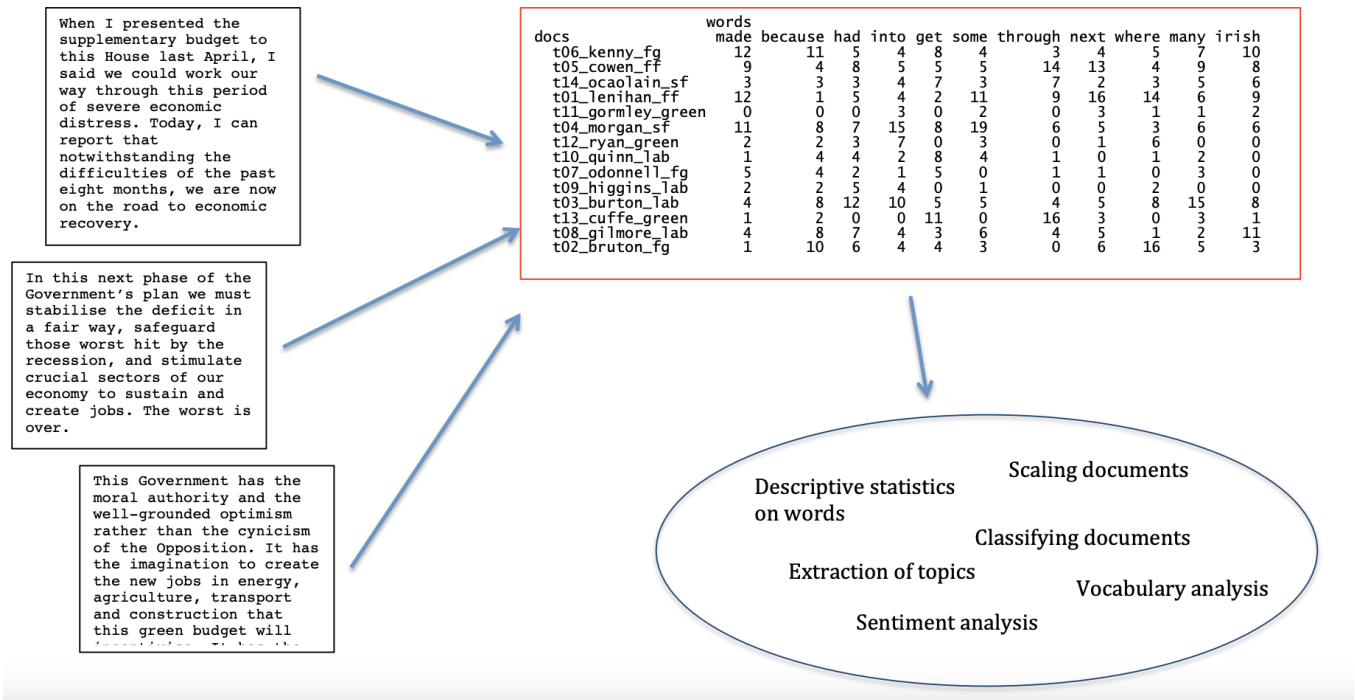
Dwelling-houses, numbered in the order of visitation.	Families, numbered in the order of visitation.	The name of every person whose place of abode on the first day of June, 1870, was in this family.	Description: Age at last birth-day. If under 1 year, give months	Description: Sex — Male (M.) Female (F.)	Description: Color — White (W.), Black (B.), Mulatto (M.), Chinese (C.), Indian (I.)	Profession, Occupation, or Trade of each person, male or female.	Value of Real Estate owned.	Value of Personal Estate.	Place of Birth, naming State or Territory of U. S.; or the Country, if of foreign birth.	Parentage: Father of foreign birth.	Parentage: Mother of foreign birth.	If born within the year, give months (Jan., Feb., &c.)	If married within the year, give month (Jan., Feb., &c.)	Attended school within the year.	Education: Cannot read.	Education: Cannot write.	Whether deaf and dumb, blind, insane, or idiotic.	Constitutional Relations: Male Citizens of U. S. of 21 years of age and upwards.	Constitutional Relations: Male Citizens of U. S. of 21 years of age and upwards, whose right to vote is denied or abridged on other grounds than participation in rebellion or other crime.
1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
1	a	6	Mayer Louis	26	M	W	Domestic Servt		Massachusetts	/	/								
2			Smith Mary	28	F	W	House Keeper		New York										
3			Keenan Elizabeth	27	F	W	Domestic Servt		Virginia						/	/			
4	7	1	Wallis Sarah	58	F	W	Keeping Boarders		Pennsylvania										
5			Cornell Hattie	23	F	W	House Keeper		"										
6			Harris Rachel	23	F	W	Domestic Servt		Virginia						/	/			
7			Brown Carrie	19	F	W	" "		New York										
8			Shepherd Sarah	24	F	W	" "		Virginia						/	/			
9			Campbell Emma	22	F	W	" "		Pennsylvania										
10			Hamilton Hattie	19	F	W	" "		New York	/	/								
11			Brown Sophia	32	F	W	Dom. Servt		" "						/	/			
12			Clarke Agnes	18	F	W	" "		Alabama										
13			Wilcox Minnie	19	F	W	" "		"										
14			Brown Mary	25	F	W	" "		Penn										
15	8	2	Roche Ella	20	F	W	" "		New York										
16			Grundy Anna	23	F	W	" "		New Jersey										
17			Dare Anna	29	F	W	" "		Kentucky										
18			Lovell Anna	21	F	W	" "		New York	/	/								
19			Wilkin Hannah	50	F	B	Domestic Servt		" "						/	/			
20			Ray Martha	15	F	W	" "		" "						/	/			
21																			

Por qué paso esto?

- incremento en la disponibilidad de texto no estructurado
- reducción en los costos de almacenamiento 1956 $\$10,000$ por megabyte, 2014 menos de 0.0001 por magabyte
- explosión en métodos, lenguajes de programación y librerias capaces de lidear con texto
  - generalizable: un mismo método puede ser usado entre bases de datos
  - sistemático: parámetros y estadísticas comparables entre datos
  - barato: R,Python,Julia, etc pueden lidear con estos datos y cualquiera puede hacerlo desde su casa

Analísis cuantitativo de texto

Procesamiento del texto a datos estructurados



Métafora del pajar de Grimmer

El análisis automatizado mejoran la lectura

- analizar una pajita de heno $\equiv$ entender el significado de una oración
  - Humanos: Muy buenos
  - Máquina: Bastante malas
- organizar todo el pajar $\equiv$ describir, clasificar y escalar textos
  - Humanos: Bastante malos
  - Máquina: Muy buenas

Principios del analisis cuantitativo de texto (Grimmer & Stuart 2013)

1. Todos los modelos están mal pero algunos son útiles
   - proceso de generación del texto desconocido
   - complejidad del lenguaje: temporalidad, sinónimos, ironía, sarcasmo, etc.
   - los modelos fallan al intentar capturar el lenguaje pero son buenos para tareas específicas
2. Los métodos cuantitativos aumentan las capacidades humanas pero no remplazan al humano
   - algoritmo: organizan, dirigen, sugieren
   - humano: lee e interpreta

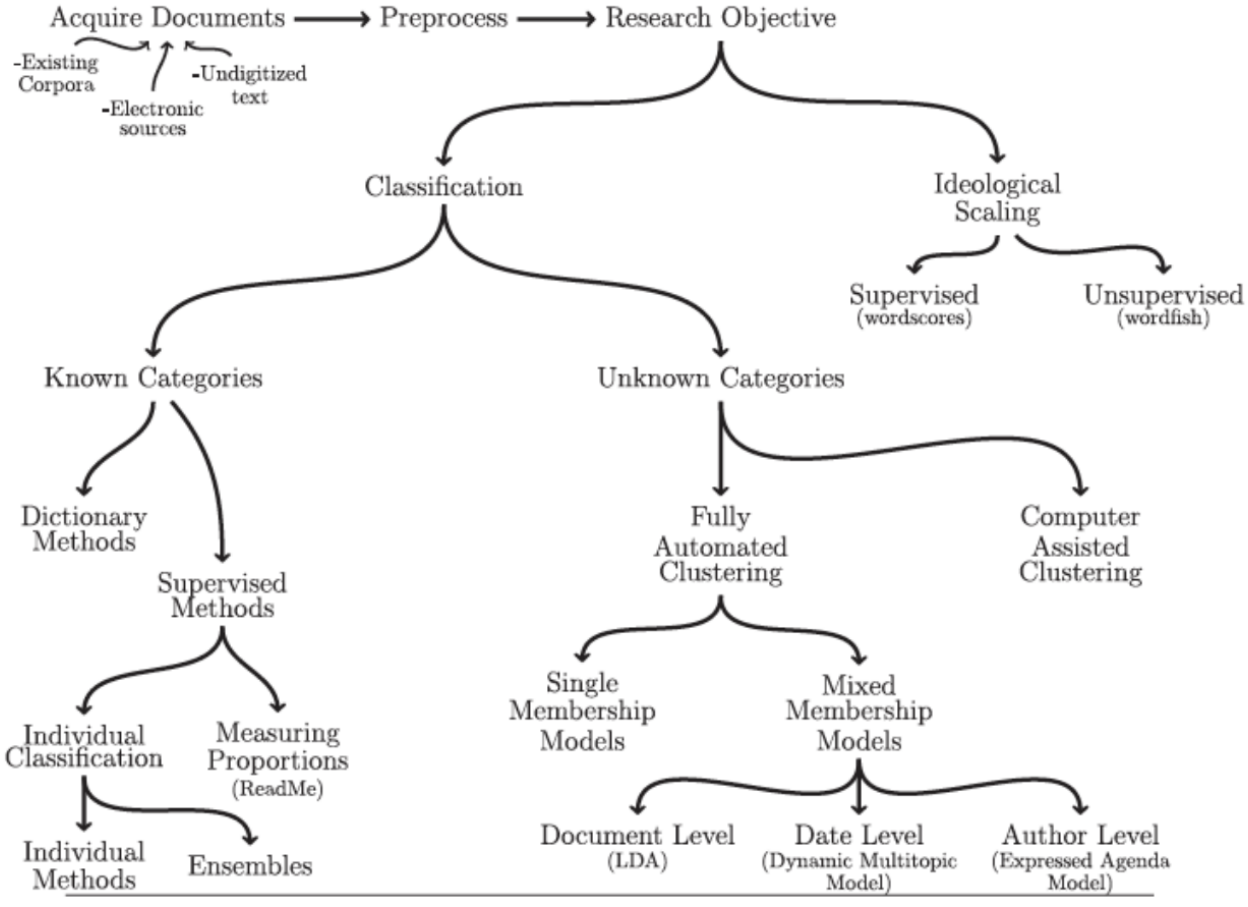
Principios del analisis cuantitativo de texto (Grimmer & Stuart 2013)

3. No existe un un método mejor que otro para automatizar texto
   - aprendizaje supervisado $\rightarrow$ categorías pre determinadas
   - aprendizaje no supervisado $\rightarrow$ categorías por clasificar
4. $\Longrightarrow$ validar, validar, validar
   - distinto rendimiento dependiendo de la tarea en cuestión
   - poca teoría que sustente que método escoger
   - Evitar aplicar métodos ciegamente

Supuestos del analisis cuantitivo del texto

1. El texto representa alguna característica de interés
   - un atributo del autor
   - sentimiento/emoción
   - relevancia de alguna cuestión política, etc
2. El texto se puede representar extrayendo sus características
   - bolsa de palabras (bag of words)
   - word embeddings, etc
3. Una matriz de características del documento puede analizarse usando métodos cuantitativos para producir estimados validos y significativos de la característica de interés

Métodos de análisis de texto



Proceso de análisis de texto

Proceso de análisis de texto

1. Seleccionar el texto: definir el corpus
2. Convertir el texto a algo que la máquina pueda leer
3. Definir los documentos es decir la unidad de análisis
4. Definir las características: tokens, diccionarios, lenguaje, etc
5. Convertir las características a una matriz
6. Escoger un método cuantitativo para el análisis de las matrices
7. Presentar los resultados obtenidos

Ejemplo: Hoberg and Phillips 2016

- nueva manera de clasificar industrias usando la descripción de sus productos
- esto permite clasificaciones de industria mas flexibles y capaces de evolucionar con el tiempo
- por qué queremos saber esto? analizar mejor los efectos de la competencia y la oferta de distintos productos

Ejemplo: Hoberg and Phillips 2016

Cómo lo hacen?

- todas las empresas que cotizan en bolsa tienen reportes 10-K
- estos reportes describen los productos que cada empresa provee
- para cada empresa i cuentan cuantas veces aparece cada palabra y lo representan como c_i
- calculan que tan diferentes son dos productos usando el angulo entre los vectores c_i y c_j

Ejemplo: Hoberg and Phillips 2016

- definen las industrias agrupando las empresas con las c's mas cercanas
- 300 grupos distintos para que coincida con la clasificación de industrias pre existentes
- asignan una industria a cada una de las empresas por año $\hat{v}_{it}$
- examinan el efecto de shocks a la industria militar y de software en la competencia y la oferta de productos

Ejemplo: Hoberg and Phillips 2016

SubMarket 1 Entertainment (Sample Focal Firm: Wanderlust Interactive)

43 rivals: Maxis, Piranha Interactive Publishing, Brilliant Digital Entertainment, Midway Games, Take Two Interactive Software, THQ, 3DO, New Frontier Media, ...

SIC codes of rivals: computer programming and data processing [sic3=737] (24 rivals), motion picture production and allied services [sic3=781] (4 rivals), misc other (13 rivals)

Core words: entertainment (42), video (42), television (38), royalties (35), internet (34), content (33), creative (31), promotional (31), copyright (31), game (30), sound (29), publishing (29), ...

SubMarket 2: Medical services (Sample Focal Firm: Quadramed Corp)

66 rivals: IDX Systems, Medicus Systems, Hpr, Simione Central Holdings, National Wireless Holdings, HCIA, Apache Medical Systems, ...

SIC codes of rivals: computer programming and data processing [sic3=737] (45 rivals), insurance agents, brokers, and service [sic3=641] (5 rivals), miscellaneous health services [sic3=809] (4 rivals), management and public relations services [sic3=874] (3 rivals), misc other (9 rivals)

Core words: client (59), database (54), solution (49), patient (47), copyright (47), secret (47), physician (47), hospital (46), healthcare (46), server (45), resource (44), functionality (44), billing (44), ...

SubMarket 3: Information Transmission (Sample Focal Firm: FAXSAV)

259 rivals: Omtool Ltd, Concentric Network, Premiere Technologies, International Telecommunication Data Systems, IDT Corp, Axent Technologies, Solopoint, Precision Systems, Netrix Corp, ...

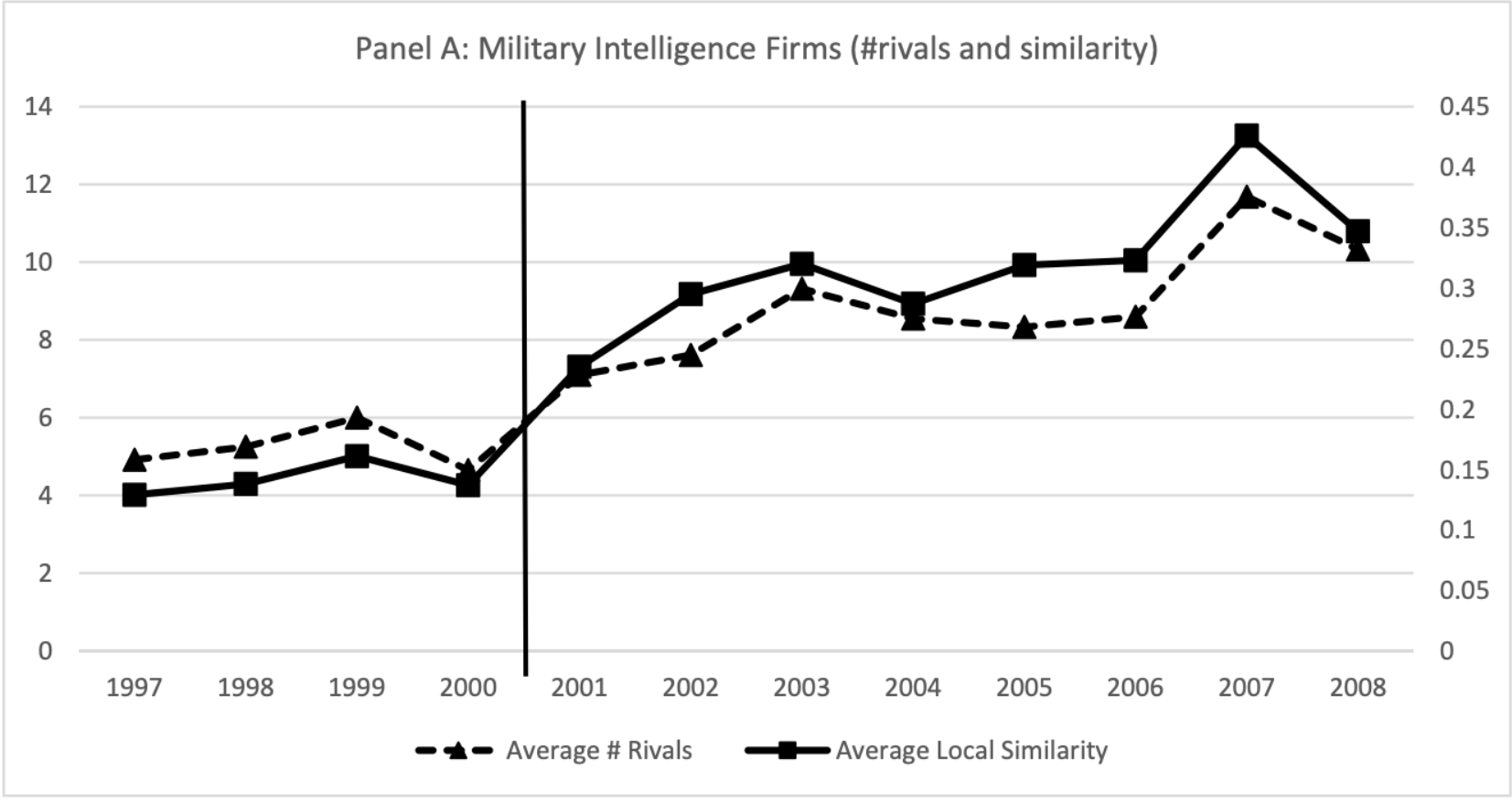
SIC codes of rivals: computer programming and data processing [sic3=737] (112 rivals), communications equipment [sic3=366] (45 rivals), telephone communications [sic3=481] (38 rivals), computer and office equipment [sic3=357] (29 rivals), communications services, other [sic3=489] (7 rivals), miscellaneous business services [sic3=738] (7 rivals), misc other (15 rivals)

Core words: internet (236), telecommunications (211), interface (194), communication (188), solution (187), platform (184), architecture (182), call (177), infrastructure (173), voice (173), functionality (173), server (173), ...

Ejemplo: Hoberg and Phillips 2016

Resultado: después del 11 de Septiembre del 2001 hay entrada de empresas al sector militar de inteligencia

Military Intelligence Firms: Competitor Changes



Características de un texto

Algunas definiciones

- corpus: conjunto amplio y estructurado de texto para analizar
- documento: cada una de las unidades del corpus
- tokens: partes del documento pueden ser palabras, frases, oraciones, etc.
- stems: reducir una palabra a su raíz por ejemplo hablar $\rightarrow$ habl
- lemmas: toma en consideración el análisis morfológico de la palabra niñas/niños $\rightarrow$ niño
- stop words: palabras que por lo general se excluyen por no aportar nada al análisis, depende de la aplicación si las dejas o no. Ejemplos: un, una, es, en, por, pero