Content Analysis

Frederik Hjorth
fh@ifs.ku.dk
fghjorth.github.io

Logic of Quantitative Research in Political Science

September 25th, 2015

Thanks to Rasmus!

- What is content analysis?
- The uses of content analysis
- Doing content analysis
- 1. Research question
- 2. Unstructured data
- 3. Coding
- 4. Reliability
- 5. Analysis
- Automated content analysis

- What is content analysis?
  - What isn't?
  - Definitions
  - Structured vs. unstructured data
- The uses of content analysis
- Doing content analysis
- 1. Research question
- 2. Unstructured data
- 3. Coding
- 4. Reliability
- 5. Analysis
- Automated content analysis

- What is content analysis?
  - What isn't?
  - Definitions
  - Structured vs. unstructured data
- The uses of content analysis
- Doing content analysis
- 1. Research question
- 2. Unstructured data
- 3. Coding
- 4. Reliability
- 5. Analysis
- Automated content analysis

What *isn't*? (Neuendorf)

What *isn't*? (Neuendorf)

- Rhetorical analysis

What *isn't*? (Neuendorf)

- Rhetorical analysis
- Narrative analysis

What *isn't*? (Neuendorf)

- Rhetorical analysis
- Narrative analysis
- Discourse analysis

What *isn't*? (Neuendorf)

- Rhetorical analysis
- Narrative analysis
- Discourse analysis
- Structuralist/semiotic analysis

What *isn't*? (Neuendorf)

- Rhetorical analysis
- Narrative analysis
- Discourse analysis
- Structuralist/semiotic analysis
- Interpretive analysis

What *isn't*? (Neuendorf)

- Rhetorical analysis
- Narrative analysis
- Discourse analysis
- Structuralist/semiotic analysis
- Interpretive analysis
- Conversation analysis

What *isn't*? (Neuendorf)

- Rhetorical analysis
- Narrative analysis
- Discourse analysis
- Structuralist/semiotic analysis
- Interpretive analysis
- Conversation analysis
- Critical analysis

What *isn't*? (Neuendorf)

- Rhetorical analysis
- Narrative analysis
- Discourse analysis
- Structuralist/semiotic analysis
- Interpretive analysis
- Conversation analysis
- Critical analysis
- Normative analysis

What *isn't*? (Neuendorf)

- Rhetorical analysis
- Narrative analysis
- Discourse analysis
- Structuralist/semiotic analysis
- Interpretive analysis
- Conversation analysis
- Critical analysis
- Normative analysis

By convention, 'content analysis' $\approx$ manual, quantitative

What *isn't*? (Neuendorf)

- Rhetorical analysis
- Narrative analysis
- Discourse analysis
- Structuralist/semiotic analysis
- Interpretive analysis
- Conversation analysis
- Critical analysis
- Normative analysis

By convention, 'content analysis' $\approx$ manual, quantitative $\neq$ 'qualitative', 'automated'

- What is content analysis?
  - What isn't?
  - Definitions
  - Structured vs. unstructured data
- The uses of content analysis
- Doing content analysis
- 1. Research question
- 2. Unstructured data
- 3. Coding
- 4. Reliability
- 5. Analysis
- Automated content analysis

»Content analysis is a research technique for the objective, systematic and quantitative description of the manifest content of communication« (Berelson 1952)

»Content analysis is a research technique for the objective, systematic and quantitative description of the manifest content of communication« (Berelson 1952)

»Quantitative content analysis is the systematic and replicable examination of symbols of communication,

»Content analysis is a research technique for the objective, systematic and quantitative description of the manifest content of communication« (Berelson 1952)

»Quantitative content analysis is the systematic and replicable examination of symbols of communication, which have been assigned numerical values according to valid measurement rules,

»Content analysis is a research technique for the objective, systematic and quantitative description of the manifest content of communication« (Berelson 1952)

»Quantitative content analysis is the systematic and replicable examination of symbols of communication, which have been assigned numerical values according to valid measurement rules, and the analysis of relationships involving those values using statistical methods

»Content analysis is a research technique for the objective, systematic and quantitative description of the manifest content of communication« (Berelson 1952)

»Quantitative content analysis is the systematic and replicable examination of symbols of communication, which have been assigned numerical values according to valid measurement rules, and the analysis of relationships involving those values using statistical methods in order to describe the communication,

»Content analysis is a research technique for the objective, systematic and quantitative description of the manifest content of communication« (Berelson 1952)

»Quantitative content analysis is the systematic and replicable examination of symbols of communication, which have been assigned numerical values according to valid measurement rules, and the analysis of relationships involving those values using statistical methods in order to describe the communication, draw inferences about its meaning,

»Content analysis is a research technique for the objective, systematic and quantitative description of the manifest content of communication« (Berelson 1952)

»Quantitative content analysis is the systematic and replicable examination of symbols of communication, which have been assigned numerical values according to valid measurement rules, and the analysis of relationships involving those values using statistical methods in order to describe the communication, draw inferences about its meaning, or infer from the communication to its context,

»Content analysis is a research technique for the objective, systematic and quantitative description of the manifest content of communication« (Berelson 1952)

»Quantitative content analysis is the systematic and replicable examination of symbols of communication, which have been assigned numerical values according to valid measurement rules, and the analysis of relationships involving those values using statistical methods in order to describe the communication, draw inferences about its meaning, or infer from the communication to its context, both of production and consumption« (Riffe, Lacy & Fico 1998)

»Content analysis is a research technique for the objective, systematic and quantitative description of the manifest content of communication« (Berelson 1952)

»Quantitative content analysis is the systematic and replicable examination of symbols of communication, which have been assigned numerical values according to valid measurement rules, and the analysis of relationships involving those values using statistical methods in order to describe the communication, draw inferences about its meaning, or infer from the communication to its context, both of production and consumption« (Riffe, Lacy & Fico 1998)

»Content analysis is a research technique for making replicable and valid inferences from texts (or other meaningful matter) to the contexts of their use« (Krippendorff p. 18)

»Content analysis may be briefly defined as the systematic, objective, quantitative analysis of message characteristics« (Neuendorf p. 1)

»Content analysis may be briefly defined as the systematic, objective, quantitative analysis of message characteristics« (Neuendorf p. 1)

»Content analysis is a summarizing, quantitative analysis of messages that relies on the scientific method

»Content analysis may be briefly defined as the systematic, objective, quantitative analysis of message characteristics« (Neuendorf p. 1)

»Content analysis is a summarizing, quantitative analysis of messages that relies on the scientific method (including attention to objectivity-intersubjectivity, a priori design, reliability, validity, generalizability, replicability, and hypothesis testing)

»Content analysis may be briefly defined as the systematic, objective, quantitative analysis of message characteristics« (Neuendorf p. 1)

»Content analysis is a summarizing, quantitative analysis of messages that relies on the scientific method (including attention to objectivity-intersubjectivity, a priori design, reliability, validity, generalizability, replicability, and hypothesis testing) and is not limited as to the types of variables that may be measured or the contexts in which the messages are created or presented.« (Neuendorf p. 10)

»Content analysis may be briefly defined as the systematic, objective, quantitative analysis of message characteristics« (Neuendorf p. 1)

»Content analysis is a summarizing, quantitative analysis of messages that relies on the scientific method (including attention to objectivity-intersubjectivity, a priori design, reliability, validity, generalizability, replicability, and hypothesis testing) and is not limited as to the types of variables that may be measured or the contexts in which the messages are created or presented.« (Neuendorf p. 10)

- What is content analysis?
  - What isn't?
  - Definitions
  - Structured vs. unstructured data
- The uses of content analysis
- Doing content analysis
- 1. Research question
- 2. Unstructured data
- 3. Coding
- 4. Reliability
- 5. Analysis
- Automated content analysis

Unstructured data

Unstructured data $\rightarrow$ *Structured data*

Structured data

Structured data

TABLE of the binary Combinations of Oxygen with simple Substances.

	Names of the simple substances.	First degree of oxygenation.		Second degree of oxygenation.		Third degree of oxygenation.		Fourth degree of oxygenation.	
		New Names.	Ancient Names.	New Names.	Ancient Names.	New Names.	Ancient Names.	New Names.	Ancient Names
Combinations of oxygen with simple non-metallic substances.	Caloric	Oxygen gas	Vital or dephlogisticated air						
	Hydrogen	Water *							
	Azote	Nitrous oxyd, or base of nitrous gas	Nitrous gas or air	Nitrous acid	Smoaking nitrous acid	Nitric acid	Pale, or not smoaking nitrous acid	Oxygenated nitric acid	Unknown
	Charcoal	Oxyd of charcoal, or carbonic oxyd	Unknown	Carbonous acid	Unknown	Carbonic acid	Fixed air	Oxygenated carbonic acid	Unknown
	Sulphur	Oxyd of sulphur	Soft sulphur	Sulphurous acid	Sulphureous acid	Sulphuric acid	Vitriolic acid	Oxygenated sulphuric acid	Unknown
	Phosphorus	Oxyd of phosphorus	Refiduum from the combuſtion of phoſphorus	Phoſphorous acid	Volatile acid of phoſphorus	Phoſphoric acid	Phoſphoric acid	Oxygenated phoſphoric acid	Unknown
	Muriatic radical	Muriatic oxyd	Unknown	Muriatous acid	Unknown	Muriatic acid	Marine acid	Oxygenated muriatic acid	Dephlogiticated marine acid
	Fluoric radical	Fluoric oxyd	Unknown	Fluorous acid	Unknown	Fluoric acid	Unknown till lately		
	Boracic radical	Boracic oxyd	Unknown	Boracous acid	Unknown	Boracic acid	Homberg's sedative ſalt		
Combinations of oxygen with the ſimple metallic ſubſtances.	Antimony	Grey oxyd of antimony	Grey calx of antimony	White oxyd of antimony	White calx of antimony, diaphoretic antimony	Antimonic acid			
	Silver	Oxyd of ſilver	Calx of ſilver			Argentic acid			
	Arſenic	Grey oxyd of arſenic	Grey calx of arſenic	White oxyd of arſenic	White calx of arſenic	Arſeniac acid	Acid of arſenic	Oxygenated arſeniac acid	Unknown
	Biſmuth	Grey oxyd of biſmuth	Grey calx of biſmuth	White oxyd of biſmuth	White calx of biſmuth	Biſmuthic acid			
	Cobalt	Grey oxyd of cobalt	Grey calx of cobalt			Cobaltic acid			
	Copper	Brown oxyd of copper	Brown calx of copper	Blue and green oxyds of copper	Blue and green calces of copper	Cupric acid			
	Tin	Grey oxyd of tin	Grey calx of tin	White oxyd of tin	White calx of tin, or putty of tin	Stannic acid			
	Iron	Black oxyd of iron	Martial ethiops	Yellow and red oxyds of iron	Ochre and ruſt of iron	Ferric acid			
	Manganeſe	Black oxyd of manganeſe	Black calx of manganeſe	White oxyd of manganeſe	White calx of manganeſe	Manganeſic acid			
	Mercury	Black oxyd of mercury	Ethiops mineral †	Yellow and red oxyds of mercury	Turbith mineral, red precipitate, calcined mercury, precipitate *per ſe*	Mercuric acid			
	Molybdena	Oxyd of molybdena	Calx of molybdena			Molybdic acid	Acid of molybdena	Oxygenated molybdic acid	Unknown
	Nickel	Oxyd of nickel	Calx of nickel			Nickelic acid			
	Gold	Yellow oxyd of gold	Yellow calx of gold	Red oxyd of gold	Red calx of gold, purple precipitate of caſſius	Auric acid			
	Platina	Yellow oxyd of platina	Yellow calx of platina			Platinic acid			
	Lead	Grey oxyd of lead	Grey calx of lead	Yellow and red oxyds of lead	Maſſicot and minium	Plumbic acid			
	Tungſtein	Oxyd of Tungſtein	Calx of Tungſtein			Tungſtic acid	Acid of Tungſtein	Oxygenated Tungſtic acid	Unknown
	Zinc	Grey oxyd of zinc	Grey calx of zinc	White oxyd of zinc	White calx of zinc, pompholix	Zincic acid			

* Only one degree of oxygenation of hydrogen is hitherto known.—A

† Ethiops mineral is the ſulphuret of mercury; this ſhould have been called black precipitate of mercury.—E.

Structured data

year	s	rv	k	cd	rfb	sf	dkp	df	fk	lc	kd	sp	u	v	vs	fp	el	la	alt	other	total
1953	74	14	30		6		8					1	0	42							175
1957	70	14	30		9		6					1	0	45						0	175
1960	76	11	32		0	11	0					1	6	38						0	175
1964	76	10	36		0	10	0					0	5	38						0	175
1966	69	13	34		0	20	0			4			0	35						0	175
1968	62	27	37		0	11	0			0		0	0	34	4					0	175
1971	70	27	31		0	17	0				0	0		30	0					0	175
1973	46	20	16	14	5	11	6				7			22	0	28				0	175

Unstructured data

Unstructured data

De anførte eksempler giver grund til bekymring, hvad enten den oplevede frygt hviler på et falsk grundlag eller ej. Faktum er, at den findes, og at den fører til selvcensur. Der sker en intimidering af det offentlige rum. Kunstnere, forfatt tegnere, oversættere og teaterfolk går derfor i en stor bue uden om vor tids vigtige kulturmøde, det mellem islam og de sekulære, vestlige samfund med rod i kristendommen.

Det moderne, sekulære samfund afvises af nogle muslimer. De gør krav på en særstil når de insisterer på særlig hensyntagen til egne religiøse følelser. Det er uforer med et verdsligt demokrati og ytringsfrihed, hvor man må være rede til at finde sig i hån, spot og latterliggørelse. Det er bestemt ikke altid lige sympatisk og pænt at se på, og det betyder ikke, at religiøse følelser for enhver pris skal gøres til grin, men det er underordnet i sammenhængen.

Unstructured data

De anførte eksempler giver grund til bekymring, hvad enten den oplevede frygt hviler på et falsk grundlag eller ej. Faktum er, at den findes, og at den fører til selvcensur. Der sker en intimidering af det offentlige rum. Kunstnere, forfatt tegnere, oversættere og teaterfolk går derfor i en stor bue uden om vor tids vigtige kulturmøde, det mellem islam og de sekulære, vestlige samfund med rod i kristendommen.

Det moderne, sekulære samfund afvises af nogle muslimer. De gør krav på en særstil når de insisterer på særlig hensyntagen til egne religiøse følelser. Det er uforer med et verdsligt demokrati og ytringsfrihed, hvor man må være rede til at finde sig i hån, spot og latterliggørelse. Det er bestemt ikke altid lige sympatisk og pænt at se på, og det betyder ikke, at religiøse følelser for enhver pris skal gøres til grin, men det er underordnet i sammenhængen.

- »Muhammeds ansigt«, *Jyllands-Posten*, September 30th 2005

Unstructured data

Unstructured data

Unstructured data

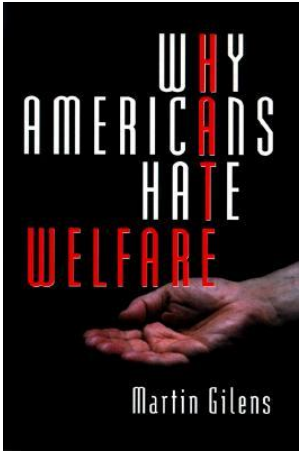
	Analog	Digital
Structured	data table in a book	.csv file
Unstructured	text in a book	online text

- What is content analysis?
- The uses of content analysis
  - Motivating examples
  - Pros and cons of content analysis
- Doing content analysis
- 1. Research question
- 2. Unstructured data
- 3. Coding
- 4. Reliability
- 5. Analysis
- Automated content analysis

- What is content analysis?
- The uses of content analysis
  - Motivating examples
  - Pros and cons of content analysis
- Doing content analysis
- 1. Research question
- 2. Unstructured data
- 3. Coding
- 4. Reliability
- 5. Analysis
- Automated content analysis

How do American news media portray people on welfare?

How do American news media portray people on welfare?



How do American news media portray people on welfare?

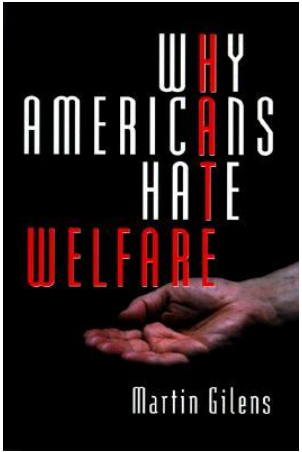


Table 4. Percent African Americans in Pictures of the Poor by Topic of Story

Topic	Number of Stories	Number of Poor People Pictured[a]	Percent African American
Underclass	6	36	100
Poor	33	147	69
Housing/homelessness[b]	96	195	66
Education for the poor[c]	4	17	65
Poor children[d]	24	70	60
Public welfare	25	97	57
Employment programs for the poor[e]	9	52	40
Medicaid	7	6	17
Miscellaneous others[f]	14	13	43
Total	182	560	62

NOTE.—Column entries exceed totals shown because stories may be indexed under more than one topic.

[a] Excludes 75 people for whom race could not be determined.

[b] Includes Housing [city/state], U.S.; Housing projects; Housing, federal aid; Housing vouchers; Department of H.U.D.; Homeless; Poor, housing; Welfare hotels; Habitat for Humanity; Covenant House.

[c] Includes Head Start; Poor, education.

[d] Includes Child welfare; Children, homeless; Runaways; Socially handicapped children.

[e] Includes Workfare; Job Corps; American Conservation Corps.

[f] Includes MadCAPP; LIFE program; I Have a Dream Foundation; Refugees; Economic assistance, domestic; Legal aid; Relief work; Unemployment insurance; Street News; Entitlement spending.

Are political campaigns in Denmark negative?

Are political campaigns in Denmark negative?

»It is my impression that some Danish politicians, in the media, use an increasingly rude tone towards their opponents.«

Are political campaigns in Denmark negative?

»It is my impression that some Danish politicians, in the media, use an increasingly rude tone towards their opponents.« - Associate Professor in the Danish language, Randi Benedikte Brodersen, »The language of politicians has become nastier«, *Politiken*, October 16th 2001

Are political campaigns in Denmark negative? (Hansen & Pedersen, 2008)

Are political campaigns in Denmark negative? (Hansen & Pedersen, 2008)

Table 1. Party Advertisements Coded for Tone (Number of Advertisements)

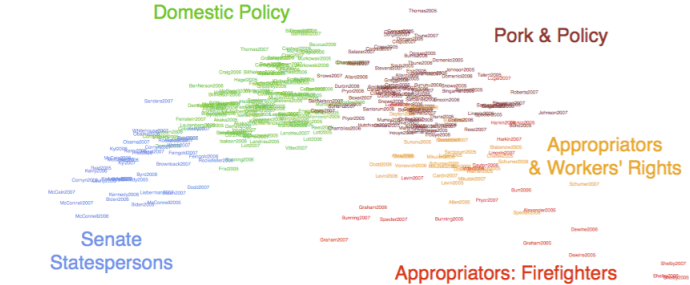
	Positive	Primarily positive	Balanced	Primarily negative	Negative	Total	Percentage negative or primarily negative of total number of party advertisements
Danish Red-Green Alliance	17	0	2	2	2	23	17
Socialist People's Party	74	8	19	15	0	116	13
Minority Party	2	0	0	0	0	2	0
Social Democrats	86	12	45	0	28	171	39
Social Liberals	21	0	3	6	0	30	20
Centrum-Demokraterne	33	0	0	0	0	33	0
Christian Democrats	3	1	0	0	0	4	0
Danish People's Party	122	0	4	0	4	130	3
Venstre	178	0	0	6	0	184	3
Conservatives	64	0	7	0	0	71	0
Government & support party	367	1	11	6	4	389	4*
Opposition	233	20	69	23	30	375	14
Frontrunners	354	8	31	29	2	424	7
Runners-up	246	13	49	0	32	340	9
Total	600	21	80	29	34	764	8
Percentage of total	78	3	10	4	5	100	

Are political campaigns in Denmark negative? (Elmelund-Præstekær & Svensson, 2014)

Are political campaigns in Denmark negative? (Elmelund-Præstekær & Svensson, 2014)

How do American elected officials talk to their constituents? (Grimmer, 2013)

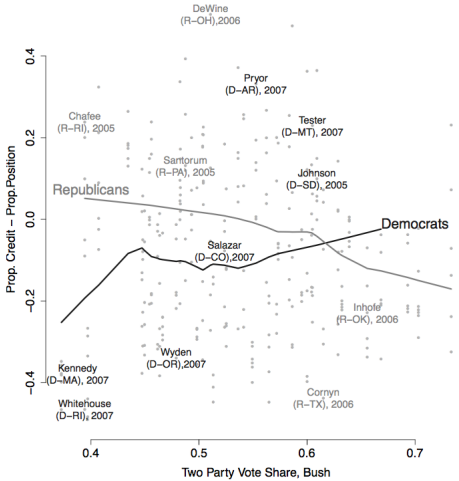
How do American elected officials talk to their constituents? (Grimmer, 2013)



Senate Position Taker	Domestic Policy	Pork & Policy	Appropriators
- Iraq War	- Environment	- WRDA grants	- Fire Grants
- Intelligence	- Gas prices	- Farming	- Airport Grants
- Intl. Relations	- DHS	- Health Care	- University Money
- Budget	- Consumer Safety	- Education Policy	- Police Grants

How do American elected officials talk to their constituents? (Grimmer, 2013)

How do American elected officials talk to their constituents? (Grimmer, 2013)



Are cartoons about Islam more negative than cartoons about Christianity? (Kaylor, 2012)

Are cartoons about Islam more negative than cartoons about Christianity? (Kaylor, 2012)

TABLE 1
Tone of Cartoons by Religion

	Positive	*Negative*	*Neutral*	*Total*
Christian	14 (7.2%)	148 (76.3%)	32 (16.5%)	194 (100%)
Muslim	1 (1.9%)	45 (84.9%)	7 (13.2%)	53 (100%)
Other religions	0 (0%)	3 (100%)	0 (0%)	3 (100%)
Atheist/Agnostic	0 (0%)	4 (80%)	1 (20%)	5 (100%)
All religions	0 (0%)	10 (100%)	0 (0%)	10 (100%)
Total	15 (5.7%)	210 (79.2%)	40 (15.1%)	265 (100%)

Do liberals and conservatives have different living spaces? (Carney et al., 2008)

Do liberals and conservatives have different living spaces? (Carney et al., 2008)

Table 4. Relations between Political Conservatism of Occupant and Room Cues in Bedrooms and Office Spaces (Study 3)

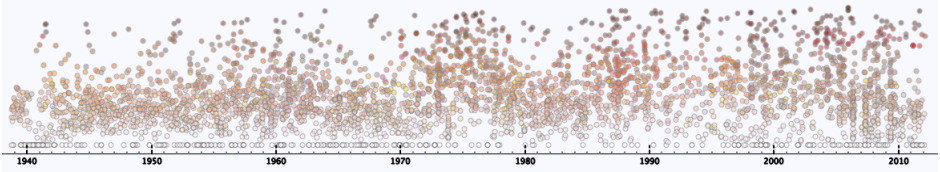
	Relation with liberalism-conservatism			
	Bedrooms		Offices	
	β	*b* (*SE*)	β	*b* (*SE*)
Sports-related décor (posters, paintings, photos)	.34**	.23 (.07)	n/a	n/a
Event calendar	.31**	.27 (.10)	n/a	n/a
Postage stamps	.30**	.29 (.11)	n/a	n/a
Presence of string/thread	.29*	.33 (.12)	n/a	n/a
Iron and/or ironing board	.28*	.20 (.08)	n/a	n/a
Laundry basket	.25*	.11 (.05)	n/a	n/a

Table 4. (cont.)

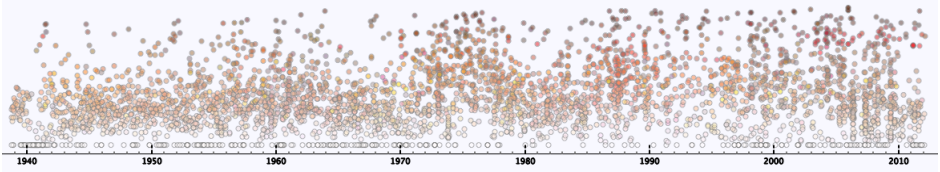
	Relation with liberalism-conservatism			
	Bedrooms		Offices	
	β	*b* (*SE*)	β	*b* (*SE*)
Many (vs. few) items of stationery	−.26*	−.27 (.12)	−.18	−.10 (.07)
World music CDs	−.26*	−.13 (.05)	n/a	n/a
Art supplies	−.27*	−.12 (.05)	n/a	n/a
Variety of music	−.27*	−.34 (.14)	n/a	n/a
Varied (vs. homogenous) books	−.34**	−.40 (.13)	−.29⁺	−.09 (.05)

Have TIME Magazine covers become more racially diverse over time? (Conway, 2012)

Have TIME Magazine covers become more racially diverse over time? (Conway, 2012)



Have TIME Magazine covers become more racially diverse over time? (Conway, 2012)



Drew Conway, The Shades of TIME

- What is content analysis?
- The uses of content analysis
  - Motivating examples
  - Pros and cons of content analysis
- Doing content analysis
- 1. Research question
- 2. Unstructured data
- 3. Coding
- 4. Reliability
- 5. Analysis
- Automated content analysis

Cons:

Cons:

- Often requires a lot of work (of the boring kind)

Cons:

- Often requires a lot of work (of the boring kind)
- Results are often 'merely descriptive' (though see Gerring, 2012; Grimmer, 2015)

Cons:

- Often requires a lot of work (of the boring kind)
- Results are often 'merely descriptive' (though see Gerring, 2012; Grimmer, 2015)
- Data often (seemingly) idiosyncratic

Cons:

- Often requires a lot of work (of the boring kind)
- Results are often 'merely descriptive' (though see Gerring, 2012; Grimmer, 2015)
- Data often (seemingly) idiosyncratic
- Coding typically requires rich contextual knowledge

Cons:

- Often requires a lot of work (of the boring kind)
- Results are often 'merely descriptive' (though see Gerring, 2012; Grimmer, 2015)
- Data often (seemingly) idiosyncratic
- Coding typically requires rich contextual knowledge (*is that a con?*)

Cons:

- Often requires a lot of work (of the boring kind)
- Results are often 'merely descriptive' (though see Gerring, 2012; Grimmer, 2015)
- Data often (seemingly) idiosyncratic
- Coding typically requires rich contextual knowledge (*is that a con?*)

Pros:

Cons:

- Often requires a lot of work (of the boring kind)
- Results are often 'merely descriptive' (though see Gerring, 2012; Grimmer, 2015)
- Data often (seemingly) idiosyncratic
- Coding typically requires rich contextual knowledge (*is that a con?*)

Pros:

- Easier path to originality

Cons:

- Often requires a lot of work (of the boring kind)
- Results are often 'merely descriptive' (though see Gerring, 2012; Grimmer, 2015)
- Data often (seemingly) idiosyncratic
- Coding typically requires rich contextual knowledge (*is that a con?*)

Pros:

- Easier path to originality
- Research questions often more intuitively motivating

Cons:

- Often requires a lot of work (of the boring kind)
- Results are often 'merely descriptive' (though see Gerring, 2012; Grimmer, 2015)
- Data often (seemingly) idiosyncratic
- Coding typically requires rich contextual knowledge (*is that a con?*)

Pros:

- Easier path to originality
- Research questions often more intuitively motivating
- Methodologically simple (ctr. Neuendorf)

Cons:

- Often requires a lot of work (of the boring kind)
- Results are often 'merely descriptive' (though see Gerring, 2012; Grimmer, 2015)
- Data often (seemingly) idiosyncratic
- Coding typically requires rich contextual knowledge (*is that a con?*)

Pros:

- Easier path to originality
- Research questions often more intuitively motivating
- Methodologically simple (ctr. Neuendorf)
  $\rightarrow$ great for student work

Content analysis as 'mere description'

Content analysis as 'mere description'

»Political scientists prioritize causal inference and theory building, often pejoratively dismissing measurement—inferences characterizing and measuring conditions as they are in the world—as 'mere description' or 'induction.'

Content analysis as 'mere description'

»Political scientists prioritize causal inference and theory building, often pejoratively dismissing measurement—inferences characterizing and measuring conditions as they are in the world—as 'mere description' or 'induction.' (...) The dismissal of description is ironic because much of the empirical work of political scientists and theories that they construct are a direct product of description.

Content analysis as 'mere description'

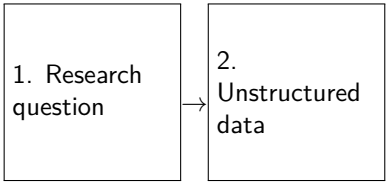
»Political scientists prioritize causal inference and theory building, often pejoratively dismissing measurement—inferences characterizing and measuring conditions as they are in the world—as 'mere description' or 'induction.' (...) The dismissal of description is ironic because much of the empirical work of political scientists and theories that they construct are a direct product of description. Indeed, political scientists have developed a wide range of strategies for carefully measuring quantities of interest from data, validating those measures, and distributing them for subsequent articles.«

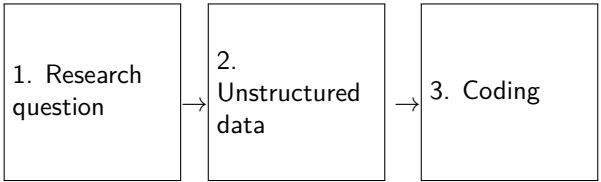
Content analysis as 'mere description'

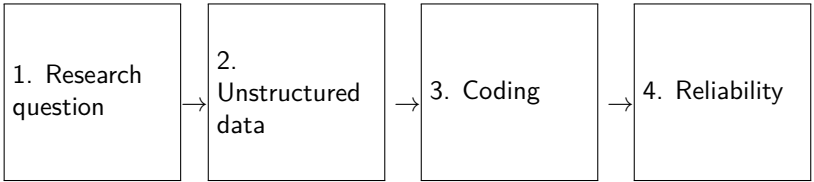
»Political scientists prioritize causal inference and theory building, often pejoratively dismissing measurement—inferences characterizing and measuring conditions as they are in the world—as 'mere description' or 'induction.' (...) The dismissal of description is ironic because much of the empirical work of political scientists and theories that they construct are a direct product of description. Indeed, political scientists have developed a wide range of strategies for carefully measuring quantities of interest from data, validating those measures, and distributing them for subsequent articles.« - Grimmer, »We Are All Social Scientists Now: How Big Data, Machine Learning, and Causal Inference Work Together«, 2015

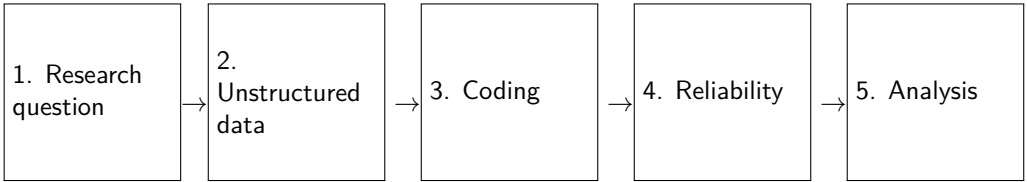
- What is content analysis?
- The uses of content analysis
- Doing content analysis
- 1. Research question
- 2. Unstructured data
- 3. Coding
- 4. Reliability
- 5. Analysis
- Automated content analysis

1. Research question


1. Research question
2. Unstructured data


1. Research question
2. Unstructured data
3. Coding


1. Research question
2. Unstructured data
3. Coding
4. Reliability


1. Research question
2. Unstructured data
3. Coding
4. Reliability
5. Analysis

- What is content analysis?
- The uses of content analysis
- Doing content analysis
- 1. Research question
- 2. Unstructured data
- 3. Coding
- 4. Reliability
- 5. Analysis
- Automated content analysis

Exercise 1

Exercise 1

What would be a research question (relevant to your project) answerable using unstructured data?

- What is content analysis?
- The uses of content analysis
- Doing content analysis
- 1. Research question
- 2. Unstructured data
  - Sampling strategy
  - Sample size
  - Data sources
- 3. Coding
- 4. Reliability
- 5. Analysis
- Automated content analysis

- What is content analysis?
- The uses of content analysis
- Doing content analysis
- 1. Research question
- 2. Unstructured data
  - Sampling strategy
  - Sample size
  - Data sources
- 3. Coding
- 4. Reliability
- 5. Analysis
- Automated content analysis

Census

Census

- Entire content universe of interest

Census

- Entire content universe of interest
- E.g.: State of the Union speeches

Census

- Entire content universe of interest
- E.g.: State of the Union speeches

Sample

Census

- Entire content universe of interest
- E.g.: State of the Union speeches

Sample

- Subset of content universe of interest

Census

- Entire content universe of interest
- E.g.: State of the Union speeches

Sample

- Subset of content universe of interest
- E.g.: Newspaper articles about the US President

Census

- Entire content universe of interest
- E.g.: State of the Union speeches

Sample

- Subset of content universe of interest
- E.g.: Newspaper articles about the US President
- Crucial issue: representativeness

»A sample is said to be representative of a population if studying leads to results that are approximately the same as those that one would reach by studying the entire population.« (Krippendorff, p. 112)

»A sample is said to be representative of a population if studying leads to results that are approximately the same as those that one would reach by studying the entire population.« (Krippendorff, p. 112)

- Ideal: simple random sampling

»A sample is said to be representative of a population if studying leads to results that are approximately the same as those that one would reach by studying the entire population.« (Krippendorff, p. 112)

- Ideal: simple random sampling
- Special problem in content analysis: exhaustive sampling frame $(u_1, u_2, \ldots, u_i, \ldots, u_N)$ rarely available

Typical approach: sampling from 'most important' cluster(s) of units

Typical approach: sampling from 'most important' cluster(s) of units

Benoit (2005): »we decided to sample a single newspaper, the New York Times. This paper is considered by many to be the national paper of record.

Typical approach: sampling from 'most important' cluster(s) of units

Benoit (2005): »we decided to sample a single newspaper, the New York Times. This paper is considered by many to be the national paper of record. Its coverage, therefore, may not be typical of other newspapers; however, the news coverage in the New York Times is arguably the most influential during this time period.«

Typical approach: sampling from 'most important' cluster(s) of units

Benoit (2005): »we decided to sample a single newspaper, the New York Times. This paper is considered by many to be the national paper of record. Its coverage, therefore, may not be typical of other newspapers; however, the news coverage in the New York Times is arguably the most influential during this time period.«

Strömbäck & van Aelst (2010): »In both countries, the aim was to include the most important newspapers and TV news sources, in essence, the newspapers and TV news shows that have the largest audiences and a national reach.«

Typical approach: sampling from 'most important' cluster(s) of units

Benoit (2005): »we decided to sample a single newspaper, the New York Times. This paper is considered by many to be the national paper of record. Its coverage, therefore, may not be typical of other newspapers; however, the news coverage in the New York Times is arguably the most influential during this time period.«

Strömbäck & van Aelst (2010): »In both countries, the aim was to include the most important newspapers and TV news sources, in essence, the newspapers and TV news shows that have the largest audiences and a national reach.«

$\rightarrow$ is this a reasonable criterion?

When sampling frame is not indexed: *systematic random sampling*

When sampling frame is not indexed: *systematic random sampling*

- select random starting position and random sampling interval m, sample every mth unit N times

When sampling frame is not indexed: *systematic random sampling*

- select random starting position and random sampling interval m, sample every mth unit N times
- potential issue: periodicity

When sampling frame is not indexed: *systematic random sampling*

- select random starting position and random sampling interval m, sample every mth unit N times
- potential issue: periodicity
- (in)famous example: Hatch & Hatch sample 413 June wedding announcements from the 1932-42 NYT

When sampling frame is not indexed: *systematic random sampling*

- select random starting position and random sampling interval m, sample every mth unit N times
- potential issue: periodicity
- (in)famous example: Hatch & Hatch sample 413 June wedding announcements from the 1932-42 NYT $\rightarrow$ absence of Jewish weddings as sign of low social status

When sampling frame is not indexed: *systematic random sampling*

- select random starting position and random sampling interval m, sample every mth unit N times
- potential issue: periodicity
- (in)famous example: Hatch & Hatch sample 413 June wedding announcements from the 1932-42 NYT $\rightarrow$ absence of Jewish weddings as sign of low social status

Cahnman (1948): »Jewish weddings are not performed in the seven weeks between Passover and the Feast of Weeks and in the three weeks preceding the day of mourning for the destruction of the Holy Temple in Jerusalem.

When sampling frame is not indexed: *systematic random sampling*

- select random starting position and random sampling interval m, sample every mth unit N times
- potential issue: periodicity
- (in)famous example: Hatch & Hatch sample 413 June wedding announcements from the 1932-42 NYT $\rightarrow$ absence of Jewish weddings as sign of low social status

Cahnman (1948): »Jewish weddings are not performed in the seven weeks between Passover and the Feast of Weeks and in the three weeks preceding the day of mourning for the destruction of the Holy Temple in Jerusalem. Almost invariably, June falls into the one or the other period.«

Contemporary systematic random sampling

Contemporary systematic random sampling

MEDIER 22. SEP. 2015 KL. 12.00

Dansk taleradio bliver mere ensformigt

For at nå bredere ud, har P1 siden 2007 skåret ned på diversiteten.

Contemporary systematic random sampling

MEDIER 22. SEP. 2015 KL. 12.00

Dansk taleradio bliver mere ensformigt

For at nå bredere ud, har P1 siden 2007 skåret ned på diversiteten.

»A major part of the thesis is an analysis of P1's shows in one week in 2007 compared with P1 and Radio 24syv in the same week in 2015. (...) Her data suggest a clear narrowing in the ways radio is made. (...) Though the analysis is not fully representative, it suggests some general tendencies in Danish talk radio.«

Contemporary systematic random sampling

MEDIER 22. SEP. 2015 KL. 12.00

Dansk taleradio bliver mere ensformigt

For at nå bredere ud, har P1 siden 2007 skåret ned på diversiteten.

»A major part of the thesis is an analysis of P1's shows in one week in 2007 compared with P1 and Radio 24syv in the same week in 2015. (...) Her data suggest a clear narrowing in the ways radio is made. (...) Though the analysis is not fully representative, it suggests some general tendencies in Danish talk radio.«

→ periodicity problem here?

Stratified random sampling

Stratified random sampling

Example: Coverage of 'Political Scandal' in two newspapers

Stratified random sampling

Example: Coverage of 'Political Scandal' in two newspapers

- Newspaper 1: 150 articles

Stratified random sampling

Example: Coverage of 'Political Scandal' in two newspapers

- Newspaper 1: 150 articles
- Newspaper 2: 3000 articles

Stratified random sampling

Example: Coverage of 'Political Scandal' in two newspapers

- Newspaper 1: 150 articles
- Newspaper 2: 3000 articles

Random sampling (1/10)

Stratified random sampling

Example: Coverage of 'Political Scandal' in two newspapers

- Newspaper 1: 150 articles
- Newspaper 2: 3000 articles

Random sampling (1/10)

- Newspaper 1: 15 articles

Stratified random sampling

Example: Coverage of 'Political Scandal' in two newspapers

- Newspaper 1: 150 articles
- Newspaper 2: 3000 articles

Random sampling (1/10)

- Newspaper 1: 15 articles
- Newspaper 2: 300 articles

Stratified random sampling

Example: Coverage of 'Political Scandal' in two newspapers

- Newspaper 1: 150 articles
- Newspaper 2: 3000 articles

Random sampling (1/10)

- Newspaper 1: 15 articles
- Newspaper 2: 300 articles

Stratified random sampling

Stratified random sampling

Example: Coverage of 'Political Scandal' in two newspapers

- Newspaper 1: 150 articles
- Newspaper 2: 3000 articles

Random sampling (1/10)

- Newspaper 1: 15 articles
- Newspaper 2: 300 articles

Stratified random sampling

- Newspaper 1: 150 articles

Stratified random sampling

Example: Coverage of 'Political Scandal' in two newspapers

- Newspaper 1: 150 articles
- Newspaper 2: 3000 articles

Random sampling (1/10)

- Newspaper 1: 15 articles
- Newspaper 2: 300 articles

Stratified random sampling

- Newspaper 1: 150 articles
- Newspaper 2: 150 articles

- What is content analysis?
- The uses of content analysis
- Doing content analysis
- 1. Research question
- 2. Unstructured data
  - Sampling strategy
  - Sample size
  - Data sources
- 3. Coding
- 4. Reliability
- 5. Analysis
- Automated content analysis

How many units to sample?

How many units to sample?

»Unfortunately, there is no universally accepted set of criteria for selecting the size of sample« Neuendorf, p. 88

How many units to sample?

»Unfortunately, there is no universally accepted set of criteria for selecting the size of sample« Neuendorf, p. 88

→ Not helpful!

Better approach: power analysis

Better approach: power analysis

$$\beta = \Phi\left(\frac{|\mu_t - \mu_c|\sqrt{N}}{2\sigma} - \Phi^{-1}(1 - \frac{\alpha}{2})\right) \quad (1)$$

Better approach: power analysis

$$\beta = \Phi\left(\frac{|\mu_t - \mu_c|\sqrt{N}}{2\sigma} - \Phi^{-1}(1 - \frac{\alpha}{2})\right) \quad (1)$$

β= prob. of observing significant result at level α, sample size N, true effect size $\frac{|\mu_t - \mu_c|}{\sigma}$

Better approach: power analysis

$$\beta = \Phi\left(\frac{|\mu_t - \mu_c|\sqrt{N}}{2\sigma} - \Phi^{-1}(1 - \frac{\alpha}{2})\right) \quad (1)$$

β= prob. of observing significant result at level α, sample size N, true effect size $\frac{|\mu_t - \mu_c|}{\sigma}$

In applied psychology, estimated avg. power$\approx$.52 (Mone et al., 1996)

Better approach: power analysis

$$\beta = \Phi\left(\frac{|\mu_t - \mu_c|\sqrt{N}}{2\sigma} - \Phi^{-1}(1 - \frac{\alpha}{2})\right) \quad (1)$$

β= prob. of observing significant result at level α, sample size N, true effect size $\frac{|\mu_t - \mu_c|}{\sigma}$

In applied psychology, estimated avg. power$\approx$.52 (Mone et al., 1996); in neuroscience$\approx$.21 (Button et al., 2013).

In R, package pwr function pwr.2p.test (and similar):

In R, package pwr function pwr.2p.test (and similar):

```
> pwr.2p.test(h=0.3,sig.level=0.05,power=.90,alternative="two.sided")
```

In R, package pwr function pwr.2p.test (and similar):

```
> pwr.2p.test(h=0.3,sig.level=0.05,power=.90,alternative="two.sided")
h = 0.3
n = 233.4982
sig.level = 0.05
power = 0.9
alternative = greater
```

Assumptions about effect size are consequential:

Assumptions about effect size are consequential:

```
> pwr.2p.test(h=0.1,sig.level=0.05,power=.90,alternative="two.sided")
```

Assumptions about effect size are consequential:

```
> pwr.2p.test(h=0.1,sig.level=0.05,power=.90,alternative="two.sided")
h = 0.1
n = 2101.484
sig.level = 0.05
power = 0.9
alternative = two.sided
```

Assumptions about effect size are consequential:

```
> pwr.2p.test(h=0.1,sig.level=0.05,power=.90,alternative="two.sided")
h = 0.1
n = 2101.484
sig.level = 0.05
power = 0.9
alternative = two.sided
```

Countervailing concern: *cost*

- What is content analysis?
- The uses of content analysis
- Doing content analysis
- 1. Research question
- 2. Unstructured data
  - Sampling strategy
  - Sample size
  - Data sources
- 3. Coding
- 4. Reliability
- 5. Analysis
- Automated content analysis

Where to find unstructured data?

Where to find unstructured data?

Canonical sources:

Where to find unstructured data?

Canonical sources:

- Infomedia (Danish news media)

Where to find unstructured data?

Canonical sources:

- Infomedia (Danish news media)
- LexisNexis (US news media)

Where to find unstructured data?

Canonical sources:

- Infomedia (Danish news media)
- LexisNexis (US news media)
- Comparative Manifesto Project (European party manifestos)

Exercise 2

Exercise 2

What type of data would you need for the RQ from Ex. 1? How would you gather it?

- What is content analysis?
- The uses of content analysis
- Doing content analysis
- 1. Research question
- 2. Unstructured data
- 3. Coding
  - Principles
  - Examples
- 4. Reliability
- 5. Analysis
- Automated content analysis

- What is content analysis?
- The uses of content analysis
- Doing content analysis
- 1. Research question
- 2. Unstructured data
- 3. Coding
  - Principles
  - Examples
- 4. Reliability
- 5. Analysis
- Automated content analysis

Unit of sampling $\neq$ unit of analysis

Unit of sampling $\neq$ unit of analysis

- Contrast w. survey research

Unit of sampling $\neq$ unit of analysis

- Contrast w. survey research
- Ex. 1: sampling articles ctr. analyzing paragraphs

Unit of sampling $\neq$ unit of analysis

- Contrast w. survey research
- Ex. 1: sampling articles ctr. analyzing paragraphs
- Ex. 2: Comparative Manifesto Project

Unit of sampling $\neq$ unit of analysis

- Contrast w. survey research
- Ex. 1: sampling articles ctr. analyzing paragraphs
- Ex. 2: Comparative Manifesto Project
- Sampling units typically $>$ analysis units, but not always (e.g. King et al.)

»[Coding instructions] must delineate the phenomena of interest

»[Coding instructions] must delineate the phenomena of interest and define the recording units to be described in analyzable terms,

»[Coding instructions] must delineate the phenomena of interest and define the recording units to be described in analyzable terms, the categories relevant to the research project,

»[Coding instructions] must delineate the phenomena of interest and define the recording units to be described in analyzable terms, the categories relevant to the research project, and their organization into a system of separate variables—also called a data language.« (Krippendorff, p. 351)

»[Coding instructions] must delineate the phenomena of interest and define the recording units to be described in analyzable terms, the categories relevant to the research project, and their organization into a system of separate variables—also called a data language.« (Krippendorff, p. 351)

Codebooks serve three purposes:

»[Coding instructions] must delineate the phenomena of interest and define the recording units to be described in analyzable terms, the categories relevant to the research project, and their organization into a system of separate variables—also called a data language.« (Krippendorff, p. 351)

Codebooks serve three purposes:

a. Instruct coders

»[Coding instructions] must delineate the phenomena of interest and define the recording units to be described in analyzable terms, the categories relevant to the research project, and their organization into a system of separate variables—also called a data language.« (Krippendorff, p. 351)

Codebooks serve three purposes:

a. Instruct coders
b. Link structured and unstructured data

»[Coding instructions] must delineate the phenomena of interest and define the recording units to be described in analyzable terms, the categories relevant to the research project, and their organization into a system of separate variables—also called a data language.« (Krippendorff, p. 351)

Codebooks serve three purposes:

a. Instruct coders
b. Link structured and unstructured data
c. Document the research process

Exercise 3

Exercise 3

What would be your coding strategy for the data gathered in Ex. 2? How would you assess validity and reliability?

Labor

Labor

»Three words describe good coder preparation:

Labor

»Three words describe good coder preparation: train

Labor

> »Three words describe good coder preparation: train, train

Labor

»Three words describe good coder preparation: train, train, train

Labor

»Three words describe good coder preparation: train, train, train (Neuendorf, p. 133)«

Labor

»Three words describe good coder preparation: train, train, train (Neuendorf, p. 133)«

»Content analyst have reported spending months in training sessions with coders, during which time they refined categories, altered instructions [...]« (Krippendorff, p. 129)

Blind coding

Blind coding

»Ideally, the individuals who take part in the development of the recording instructions should not be the ones who apply them, for they will have acquired an implicit consensus that new coders cannot have and that other scholars who may wish to use the instructions cannot replicate.« (Krippendorff, p. 131)

Blind coding

»Ideally, the individuals who take part in the development of the recording instructions should not be the ones who apply them, for they will have acquired an implicit consensus that new coders cannot have and that other scholars who may wish to use the instructions cannot replicate.« (Krippendorff, p. 131)

»Blind coding, in which the coders do not know the purporse of the study, is desirable, to reduce bias that compromises validity« (Neuendorf, p. 133)

Blind coding

»Ideally, the individuals who take part in the development of the recording instructions should not be the ones who apply them, for they will have acquired an implicit consensus that new coders cannot have and that other scholars who may wish to use the instructions cannot replicate.« (Krippendorff, p. 131)

»Blind coding, in which the coders do not know the purporse of the study, is desirable, to reduce bias that compromises validity« (Neuendorf, p. 133)

→ potential tradeoff btw. training and blindness

- What is content analysis?
- The uses of content analysis
- Doing content analysis
- 1. Research question
- 2. Unstructured data
- 3. Coding
  - Principles
  - Examples
- 4. Reliability
- 5. Analysis
- Automated content analysis

Example: open-ended responses about reasons for voting for/against euro adoption

Example: open-ended responses about reasons for voting for/against euro adoption

	KATEGORI	FORKLARING	EKSEMPEL
1	Generel	Generel, uspecificeret modstand mod EU/euro	"Jeg bryder mig ikke om euroen"
2	Suverænitet	Landets suverænitet/selvbestemmelse	"National selvstændighed"
3	Afstand	Afstand til Bruxelles/centralisme/demokrati	"Demokrati i EU"
4	Usikker	Generel usikkerhed/foretrækker at vente	"Vi ved ikke hvad vi får"
5	Priser	Frygt for højere prisniveau	"Priser"
6	Proces	Utilfredshed med en udemokratisk proces	"Politikerne snyder befolkningen"
7	Identitet	Vil ikke miste national identitet, dansk/svenskhed	"Vil forblive dansk/svensk"
8	Kronen	Specifikt ønske om at bevare kronen	"Vil bevare kronen"
9	Andet	Residualkategori - andre svar	-

Example: open-ended responses about reasons for voting for/against euro adoption

	KATEGORI	FORKLARING	EKSEMPEL
1	Generel	Generel, uspecificeret modstand mod EU/euro	"Jeg bryder mig ikke om euroen"
2	Suverænitet	Landets suverænitet/selvbestemmelse	"National selvstændighed"
3	Afstand	Afstand til Bruxelles/centralisme/demokrati	"Demokrati i EU"
4	Usikker	Generel usikkerhed/foretrækker at vente	"Vi ved ikke hvad vi får"
5	Priser	Frygt for højere prisniveau	"Priser"
6	Proces	Utilfredshed med en udemokratisk proces	"Politikerne snyder befolkningen"
7	Identitet	Vil ikke miste national identitet, dansk/svenskhed	"Vil forblive dansk/svensk"
8	Kronen	Specifikt ønske om at bevare kronen	"Vil bevare kronen"
9	Andet	Residualkategori - andre svar	-

Reliability: $\alpha \approx .4$

Example: open-ended responses about reasons for voting for/against euro adoption

	KATEGORI	FORKLARING	EKSEMPEL
1	Generel	Generel, uspecificeret modstand mod EU/euro	"Jeg bryder mig ikke om euroen"
2	Suverænitet	Landets suverænitet/selvbestemmelse	"National selvstændighed"
3	Afstand	Afstand til Bruxelles/centralisme/demokrati	"Demokrati i EU"
4	Usikker	Generel usikkerhed/foretrækker at vente	"Vi ved ikke hvad vi får"
5	Priser	Frygt for højere prisniveau	"Priser"
6	Proces	Utilfredshed med en udemokratisk proces	"Politikerne snyder befolkningen"
7	Identitet	Vil ikke miste national identitet, dansk/svenskhed	"Vil forblive dansk/svensk"
8	Kronen	Specifikt ønske om at bevare kronen	"Vil bevare kronen"
9	Andet	Residualkategori - andre svar	-

Reliability: $\alpha \approx .4 \rightarrow$ what was the problem?

	A	B	C	D	E	F	G	H	I
1	*Artikel-ID*	*Outlet*	*Date*	*Relevance*	*Frame*	*Winningorloosing*	*Strategy*	*Polconse*	*Coalition building*
2	e0bff6d4	24timer	13.11.2007	1	99	2	2	2	2
3	e0bff529	24timer	13.11.2007	1	99	2	2	2	2
4	e0bfc5c9	24timer	12.11.2007	1	3	1	2	1	1
5	e0bfc573	24timer	12.11.2007	1	2	2	1	1	1
6	e0bfc3bb	24timer	12.11.2007	1	3	2	1	1	1
7	e0bfc39c	24timer	12.11.2007	1	1	2	1	1	1
8	e0bfc387	24timer	12.11.2007	1	3	2	1	2	1
9	e0bf1f96	24timer	09.11.2007	1	1		2	2	2
10	e0bf1f95	24timer	09.11.2007	1	1	2	2	2	2
11	e0bf1ea9	24timer	09.11.2007	1	3	2	1	2	2
12	e0bf1e97	24timer	09.11.2007	1	1	2	1	1	2
13	e0bffc5e	BT	13.11.2007	99					
14	e0bffbbd	BT	13.11.2007	1	99	2	2	2	2
15	e0bffbb8	BT	13.11.2007	1	3	2	1	1	1
16	e0bfc7f3	BT	12.11.2007	1	99	2	2	2	2
17	e0bcf7f1	BT	12.11.2007	1	3	2	2	1	1
18	e0bfc7b8	BT	12.11.2007	1	99	2	2	2	2
19	e0cd1b29	BT	11.11.2007	1	99	2	2	2	2
20	e0bfa912	BT	11.11.2007	1	99	2	1	2	2
21	e0bfa8eb	BT	11.11.2007	1	3	2	1	2	2

Ark4 Ark2 Ark1 Ark3

Exercise 3

Exercise 3

How would you design a coding strategy for the data gathered in Ex. 2? What would be important considerations?

- What is content analysis?
- The uses of content analysis
- Doing content analysis
- 1. Research question
- 2. Unstructured data
- 3. Coding
- 4. Reliability
  - Defining
  - Assessing
  - Krippendorff's alpha
  - Examples
- 5. Analysis
- Automated content analysis

- What is content analysis?
- The uses of content analysis
- Doing content analysis
- 1. Research question
- 2. Unstructured data
- 3. Coding
- 4. Reliability
  - Defining
  - Assessing
  - Krippendorff's alpha
  - Examples
- 5. Analysis
- Automated content analysis

»Reliability can be defined as *the extent to which a measuring procedure yields the same results on repeated trials*« (Neuendorf p. 141, emhp. added)

»Reliability can be defined as *the extent to which a measuring procedure yields the same results on repeated trials*« (Neuendorf p. 141, emhp. added)

Conceptually, potential tradeoff btw. reliability and validity:

»Reliability can be defined as *the extent to which a measuring procedure yields the same results on repeated trials*« (Neuendorf p. 141, emhp. added)

Conceptually, potential tradeoff btw. reliability and validity:

- e.g. measuring 'readability'

»Reliability can be defined as *the extent to which a measuring procedure yields the same results on repeated trials*« (Neuendorf p. 141, emhp. added)

Conceptually, potential tradeoff btw. reliability and validity:

- e.g. measuring 'readability'
  - subjectively assessed

»Reliability can be defined as *the extent to which a measuring procedure yields the same results on repeated trials*« (Neuendorf p. 141, emhp. added)

Conceptually, potential tradeoff btw. reliability and validity:

- e.g. measuring 'readability'
  - subjectively assessed
  - Flesch-Kincaid: $206.835 - 1.015 \left(\frac{\text{total words}}{\text{total sentences}}\right) - 84.6 \left(\frac{\text{total syllables}}{\text{total words}}\right)$

»Reliability can be defined as *the extent to which a measuring procedure yields the same results on repeated trials*« (Neuendorf p. 141, emhp. added)

Conceptually, potential tradeoff btw. reliability and validity:

- e.g. measuring 'readability'
  - subjectively assessed
  - Flesch-Kincaid: $206.835 - 1.015 \left(\frac{\text{total words}}{\text{total sentences}}\right) - 84.6 \left(\frac{\text{total syllables}}{\text{total words}}\right)$
- but: badly conceptualized coding scheme $\rightarrow$ low reliability *and* validity

»Reliability can be defined as *the extent to which a measuring procedure yields the same results on repeated trials*« (Neuendorf p. 141, emhp. added)

Conceptually, potential tradeoff btw. reliability and validity:

- e.g. measuring 'readability'
  - subjectively assessed
  - Flesch-Kincaid: $206.835 - 1.015 \left(\frac{\text{total words}}{\text{total sentences}}\right) - 84.6 \left(\frac{\text{total syllables}}{\text{total words}}\right)$
- but: badly conceptualized coding scheme $\rightarrow$ low reliability *and* validity

»Vad du ej klart kan säga, vet du ej:
med tanken ordet föds på mannens läppar:
det dunkelt sagda är det dunkelt tänkta. «

- Esaias Tegner (1820)

- What is content analysis?
- The uses of content analysis
- Doing content analysis
- 1. Research question
- 2. Unstructured data
- 3. Coding
- 4. Reliability
  - Defining
  - Assessing
  - Krippendorff's alpha
  - Examples
- 5. Analysis
- Automated content analysis

The naive approach: percent agreement

The naive approach: percent agreement

Example: dichotomous coding 0/1, 2 coders

The naive approach: percent agreement

Example: dichotomous coding 0/1, 2 coders

	1	2	3	4	5	6	7
Lars	0	0	1	0	0	0	0

The naive approach: percent agreement

Example: dichotomous coding 0/1, 2 coders

	1	2	3	4	5	6	7
Lars	0	0	1	0	0	0	0
Solrun	0	0	0	0	1	0	0

The naive approach: percent agreement

Example: dichotomous coding 0/1, 2 coders

	1	2	3	4	5	6	7
Lars	0	0	1	0	0	0	0
Solrun	0	0	0	0	1	0	0

Percent agreement $= \frac{5}{7} = 71$ percent

The naive approach: percent agreement

Example: dichotomous coding 0/1, 2 coders

	1	2	3	4	5	6	7
Lars	0	0	1	0	0	0	0
Solrun	0	0	0	0	1	0	0

Percent agreement $= \frac{5}{7} = 71$ percent

- problem:

The naive approach: percent agreement

Example: dichotomous coding 0/1, 2 coders

	1	2	3	4	5	6	7
Lars	0	0	1	0	0	0	0
Solrun	0	0	0	0	1	0	0

Percent agreement $= \frac{5}{7} = 71$ percent

- problem: does not correct for chance

The naive approach: percent agreement

Example: dichotomous coding 0/1, 2 coders

	1	2	3	4	5	6	7
Lars	0	0	1	0	0	0	0
Solrun	0	0	0	0	1	0	0

Percent agreement $= \frac{5}{7} = 71$ percent

- problem: does not correct for chance
- most severe with presence of high-frequency, theoretically unimportant categories

- What is content analysis?
- The uses of content analysis
- Doing content analysis
- 1. Research question
- 2. Unstructured data
- 3. Coding
- 4. Reliability
  - Defining
  - Assessing
  - Krippendorff's alpha
  - Examples
- 5. Analysis
- Automated content analysis

Krippendorff's α:

Krippendorff's α:

$$\alpha = 1 - \frac{D_{\textit{within units = in error}}}{D_{\textit{within and between units = in total}}} = 1 - \frac{D_o}{D_e} \tag{2}$$

Krippendorff's α:

$$\alpha = 1 - \frac{D_{within\ units\ =\ in\ error}}{D_{within\ and\ between\ units\ =\ in\ total}} = 1 - \frac{D_o}{D_e} \tag{2}$$

where

$$\alpha_{metric} = 1 - \frac{D_o}{D_e} = 1 - \frac{\frac{1}{n}\sum_c \sum_k o_{ck}\delta_{metric}(c,k)}{\frac{1}{n(n-1)}\sum_c \sum_k n_c n_k\ \delta_{metric}(c,k)} \tag{3}$$

$$\alpha_{metric} = 1 - \frac{D_o}{D_e} = 1 - \frac{\frac{1}{n}\sum_c \sum_k o_{ck}\delta_{metric}(c,k)}{\frac{1}{n(n-1)}\sum_c \sum_k n_c n_k \ \delta_{metric}(c,k)}$$

$$\alpha_{metric} = 1 - \frac{D_o}{D_e} = 1 - \frac{\frac{1}{n}\sum_c \sum_k o_{ck}\delta_{metric}(c,k)}{\frac{1}{n(n-1)}\sum_c \sum_k n_c n_k \ \delta_{metric}(c,k)}$$

o_{ck} and n_c/n_k are cell counts and row/column sums in a *coincidence matrix*:

	1	.	k	.	.
1	o_{11}	.	o_{1k}	.	n_1
.	.	.	.	.	.
c	o_{c1}	.	o_{ck}	.	n_c
.	.	.	.	.	.
	n_1	.	n_k	.	n

$$\alpha_{metric} = 1 - \frac{D_o}{D_e} = 1 - \frac{\frac{1}{n}\sum_c \sum_k o_{ck}\delta_{metric}(c,k)}{\frac{1}{n(n-1)}\sum_c \sum_k n_c n_k \ \delta_{metric}(c,k)}$$

o_{ck} and n_c/n_k are cell counts and row/column sums in a *coincidence matrix*:

	1	.	k	.	.
1	o_{11}	.	o_{1k}	.	n_1
.	.	.	.	.	.
c	o_{c1}	.	o_{ck}	.	n_c
.	.	.	.	.	.
	n_1	.	n_k	.	n

o_{ck}: count of codings assigned values c and k

$$\alpha_{metric} = 1 - \frac{D_o}{D_e} = 1 - \frac{\frac{1}{n}\sum_c \sum_k o_{ck}\delta_{metric}(c,k)}{\frac{1}{n(n-1)}\sum_c \sum_k n_c n_k \ \delta_{metric}(c,k)}$$

$$\alpha_{metric} = 1 - \frac{D_o}{D_e} = 1 - \frac{\frac{1}{n}\sum_c \sum_k o_{ck}\delta_{metric}(c,k)}{\frac{1}{n(n-1)}\sum_c \sum_k n_c n_k \ \delta_{metric}(c,k)}$$

$\delta_{metric}(c,k)$ depends on level of measurement:

$$\alpha_{metric} = 1 - \frac{D_o}{D_e} = 1 - \frac{\frac{1}{n}\sum_c \sum_k o_{ck}\delta_{metric}(c,k)}{\frac{1}{n(n-1)}\sum_c \sum_k n_c n_k \ \delta_{metric}(c,k)}$$

$\delta_{metric}(c,k)$ depends on level of measurement:

- Nominal data: $\delta_{nominal}(c,k) = \begin{cases} 0 & \text{iff } c = k \\ 1 & \text{iff } c \neq k \end{cases}$

$$\alpha_{metric} = 1 - \frac{D_o}{D_e} = 1 - \frac{\frac{1}{n}\sum_c \sum_k o_{ck}\delta_{metric}(c,k)}{\frac{1}{n(n-1)}\sum_c \sum_k n_c n_k \ \delta_{metric}(c,k)}$$

$\delta_{metric}(c,k)$ depends on level of measurement:

- Nominal data: $\delta_{nominal}(c,k) = \begin{cases} 0 & \text{iff } c = k \\ 1 & \text{iff } c \neq k \end{cases}$
- Ordinal data: $\delta_{ordinal}(c,k) = \left(\sum_{g=c}^{g=k} n_g - \frac{n_c + n_k}{2}\right)^2$

$$\alpha_{metric} = 1 - \frac{D_o}{D_e} = 1 - \frac{\frac{1}{n}\sum_c \sum_k o_{ck}\delta_{metric}(c,k)}{\frac{1}{n(n-1)}\sum_c \sum_k n_c n_k \ \delta_{metric}(c,k)}$$

$\delta_{metric}(c,k)$ depends on level of measurement:

- Nominal data: $\delta_{nominal}(c,k) = \begin{cases} 0 & \text{iff } c = k \\ 1 & \text{iff } c \neq k \end{cases}$
- Ordinal data: $\delta_{ordinal}(c,k) = \left(\sum_{g=c}^{g=k} n_g - \frac{n_c + n_k}{2}\right)^2$
- Interval data: $\delta_{interval}(c,k) = (c-k)^2$

$$\alpha_{metric} = 1 - \frac{D_o}{D_e} = 1 - \frac{\frac{1}{n}\sum_c \sum_k o_{ck}\delta_{metric}(c,k)}{\frac{1}{n(n-1)}\sum_c \sum_k n_c n_k \ \delta_{metric}(c,k)}$$

$\delta_{metric}(c,k)$ depends on level of measurement:

- Nominal data: $\delta_{nominal}(c,k) = \begin{cases} 0 & \text{iff } c = k \\ 1 & \text{iff } c \neq k \end{cases}$
- Ordinal data: $\delta_{ordinal}(c,k) = \left(\sum_{g=c}^{g=k} n_g - \frac{n_c + n_k}{2}\right)^2$
- Interval data: $\delta_{interval}(c,k) = (c-k)^2$

Note: for nominal data, $\sum_c \sum_k o_{ck}\delta_{metric}(c,k)$ reduces to sum of off-diagonal cells!

- What is content analysis?
- The uses of content analysis
- Doing content analysis
- 1. Research question
- 2. Unstructured data
- 3. Coding
- 4. Reliability
  - Defining
  - Assessing
  - Krippendorff's alpha
  - Examples
- 5. Analysis
- Automated content analysis

	1	2	3	4	5	6	7
Lars	0	0	1	0	0	0	0
Solrun	0	0	0	0	1	0	0

	1	2	3	4	5	6	7
Lars	0	0	1	0	0	0	0
Solrun	0	0	0	0	1	0	0

$$\alpha_{nominal} = 1 - \frac{D_o}{D_e} = 1 - \frac{13}{12} = -.083$$

Coding on 1-7 scale, 2 coders

Coding on 1-7 scale, 2 coders

	1	2	3	4	5	6	7
Margrethe	1	2	3	5	7	3	2

Coding on 1-7 scale, 2 coders

	1	2	3	4	5	6	7
Margrethe	1	2	3	5	7	3	2
Henrik	1	2	2	5	6	4	3

Coding on 1-7 scale, 2 coders

	1	2	3	4	5	6	7
Margrethe	1	2	3	5	7	3	2
Henrik	1	2	2	5	6	4	3

Coincidence matrix:

Coding on 1-7 scale, 2 coders

	1	2	3	4	5	6	7
Margrethe	1	2	3	5	7	3	2
Henrik	1	2	2	5	6	4	3

Coincidence matrix:

	1	2	3	4	5	6	7
1	2						
2		2	2				
3		2		1			
4			1				
5					2		
6							1
7						1	

In R, package `irr` function `kripp.alpha`:

In R, package irr function kripp.alpha:

```
> kripp.alpha(comat,method="nominal")
```

In R, package irr function kripp.alpha:

```
> kripp.alpha(comat,method="nominal")
Krippendorff's alpha
Subjects = 7
Raters = 2
alpha = 0.35
```

When assigning higher levels of measurement:

When assigning higher levels of measurement:

```
> kripp.alpha(comat,method="ordinal")
```

When assigning higher levels of measurement:

```
> kripp.alpha(comat,method="ordinal")
Krippendorff's alpha
Subjects = 7
Raters = 2
alpha = 0.875
```

When assigning higher levels of measurement:

```
> kripp.alpha(comat,method="ordinal")
Krippendorff's alpha
Subjects = 7
Raters = 2
alpha = 0.875
> kripp.alpha(comat,method="interval")
```

When assigning higher levels of measurement:

```
> kripp.alpha(comat,method="ordinal")
Krippendorff's alpha
Subjects = 7
Raters = 2
alpha = 0.875
> kripp.alpha(comat,method="interval")
Krippendorff's alpha
Subjects = 7
Raters = 2
alpha = 0.917
```

- What is content analysis?
- The uses of content analysis
- Doing content analysis
- 1. Research question
- 2. Unstructured data
- 3. Coding
- 4. Reliability
- 5. Analysis
  - Rasmus' paper
- Automated content analysis

Exercise 4

Exercise 4

Assume reasonable reliability for the data coded in Ex. 3. How would you analyze it? What would be your testable hypothesis?

- What is content analysis?
- The uses of content analysis
- Doing content analysis
- 1. Research question
- 2. Unstructured data
- 3. Coding
- 4. Reliability
- 5. Analysis
  - Rasmus' paper
- Automated content analysis

Rasmus' paper

Rasmus' paper

»At a specific level [the project] is about how DEMA's communicative prevention initiatives are communicated and legitimized in the media and interpreted and assessed by media users.« (p. 3)

Rasmus' paper

»At a specific level [the project] is about how DEMA's communicative prevention initiatives are communicated and legitimized in the media and interpreted and assessed by media users.« (p. 3)

	Quantitative methods	Qualitative methods	Data material
Content analysis: How prevention is *disseminated*	Content analysis: Focus on e.g. reach, media type and sources	Text and argument analysis: Focus on e.g. legitimacy and the classical forms of appeal	Media texts about prevention within several areas, e.g. emergency, traffic, health and environment
Reception analysis: How prevention is *interpreted*	Questionnaire among broad target group on e.g. knowledge, relevance, credibility and attitude	Interviews and focus groups among specific target groups on e.g. knowledge, relevance, attitude and credibility	Media texts about prevention solely within the field of emergency

- What is content analysis?
- The uses of content analysis
- Doing content analysis
- 1. Research question
- 2. Unstructured data
- 3. Coding
- 4. Reliability
- 5. Analysis
- Automated content analysis
  - General principles
  - King et al.

- What is content analysis?
- The uses of content analysis
- Doing content analysis
- 1. Research question
- 2. Unstructured data
- 3. Coding
- 4. Reliability
- 5. Analysis
- Automated content analysis
  - General principles
  - King et al.

Four principles for automated text analysis, from Grimmer & Stewart (2013)

Four principles for automated text analysis, from Grimmer & Stewart (2013)

1. All quantitative models of language are wrong—but some are useful.

Four principles for automated text analysis, from Grimmer & Stewart (2013)

1. All quantitative models of language are wrong—but some are useful.
2. Quantitative methods augment humans, they do not replace them.

Four principles for automated text analysis, from Grimmer & Stewart (2013)

1. All quantitative models of language are wrong—but some are useful.
2. Quantitative methods augment humans, they do not replace them.
3. There is no globally best method for automated text analysis.

Four principles for automated text analysis, from Grimmer & Stewart (2013)

1. All quantitative models of language are wrong—but some are useful.
2. Quantitative methods augment humans, they do not replace them.
3. There is no globally best method for automated text analysis.
4. Validate, validate, validate.

Typical tasks:

Typical tasks:

- Assign texts positions on a (left-right) scale: Wordscores/Wordfish

Typical tasks:

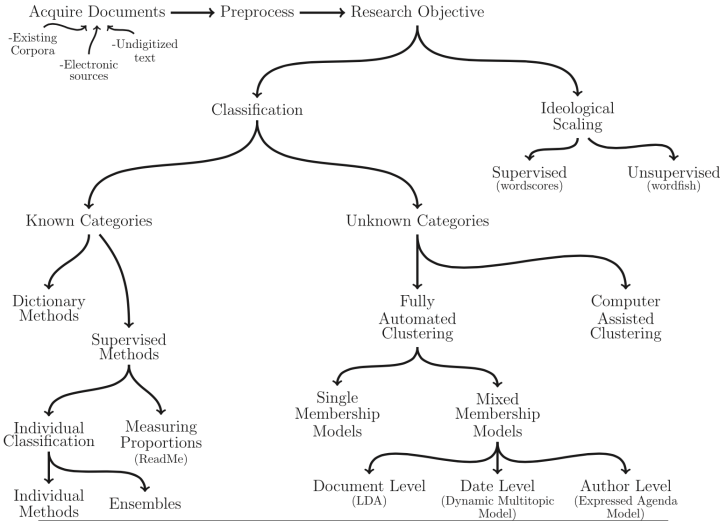
- Assign texts positions on a (left-right) scale: Wordscores/Wordfish
- Assign texts values on a variable: dictionary approaches (e.g., Lexicoder)

Typical tasks:

- Assign texts positions on a (left-right) scale: Wordscores/Wordfish
- Assign texts values on a variable: dictionary approaches (e.g., Lexicoder)
- Characterize the distribution of categories across texts: ReadMe

Typical tasks:

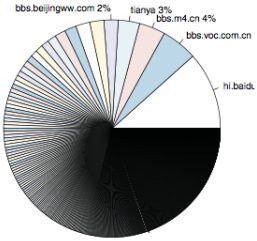
- Assign texts positions on a (left-right) scale: Wordscores/Wordfish
- Assign texts values on a variable: dictionary approaches (e.g., Lexicoder)
- Characterize the distribution of categories across texts: ReadMe
- Characterize the unknown distribution of topics across and within texts: topic models


Acquire Documents
Preprocess
Research Objective
-Existing Corpora
-Electronic sources
-Undigitized text
Classification
Ideological Scaling
Supervised (wordscores)
Unsupervised (wordfish)
Known Categories
Unknown Categories
Dictionary Methods
Supervised Methods
Fully Automated Clustering
Computer Assisted Clustering
Individual Classification
Measuring Proportions (ReadMe)
Single Membership Models
Mixed Membership Models
Individual Methods
Ensembles
Document Level (LDA)
Date Level (Dynamic Multitopic Model)
Author Level (Expressed Agenda Model)

- What is content analysis?
- The uses of content analysis
- Doing content analysis
- 1. Research question
- 2. Unstructured data
- 3. Coding
- 4. Reliability
- 5. Analysis
- Automated content analysis
  - General principles
  - King et al.

Figure 1. The Fractured Structure of the Chinese Social Media Landscape

(a) Sample of Sites



(b) All Sites excluding Sina

Figure 6. Low Censorship on News and Policy Events (in 2011)

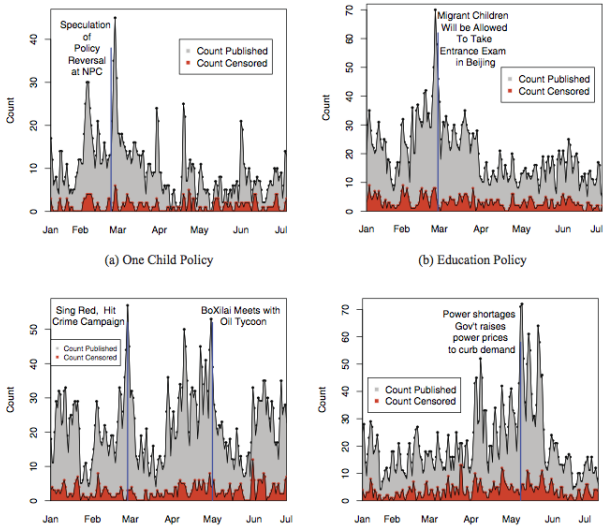
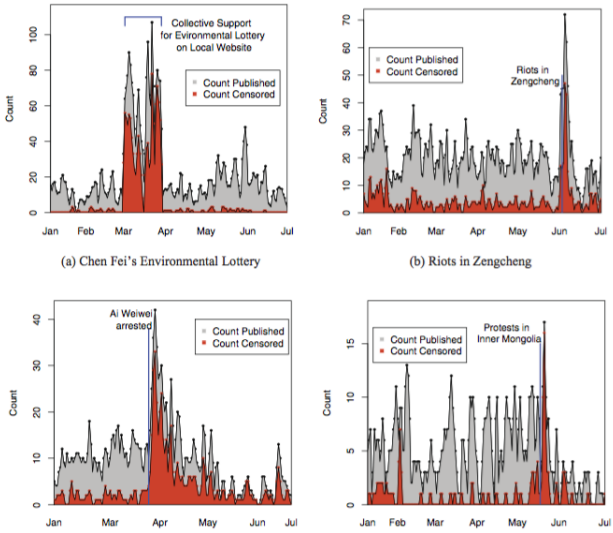


Figure 5. High Censorship During Collective Action Events (in 2011)



(a) Chen Fei's Environmental Lottery

(b) Riots in Zengcheng

Reliability of event coding?

Reliability of event coding?

»we conducted a study to verify the reliability of our event coding rules.

Reliability of event coding?

»we conducted a study to verify the reliability of our event coding rules. To do this, we gave our rules above to two people familiar with Chinese politics and asked them to code each of the eighty-seven events (each associated with a volume burst) into one of the five categories.

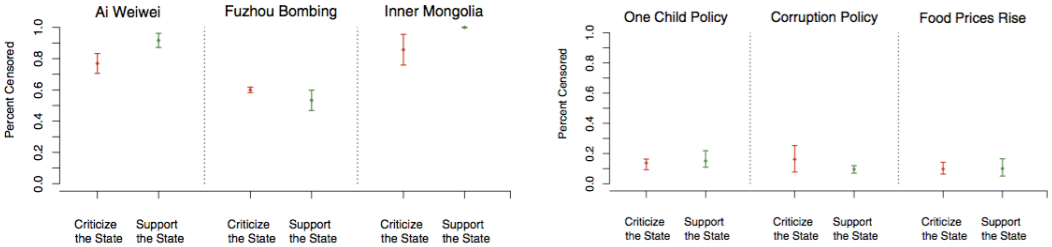
Reliability of event coding?

»we conducted a study to verify the reliability of our event coding rules. To do this, we gave our rules above to two people familiar with Chinese politics and asked them to code each of the eighty-seven events (each associated with a volume burst) into one of the five categories. The coders worked independently and classified each of the events on their own.

Reliability of event coding?

»we conducted a study to verify the reliability of our event coding rules. To do this, we gave our rules above to two people familiar with Chinese politics and asked them to code each of the eighty-seven events (each associated with a volume burst) into one of the five categories. The coders worked independently and classified each of the events on their own. Decisions by the two coders agreed in 98.9% (i.e., eighty-six of eighty-seven) of the events.« - King et al. (2013)

Figure 8. Content of Censored Posts by Topic Area



The end