Machine Learning with R using Qualitative Research Codes

Marla Stuart, PhD, MSW
Fellow, Berkeley Institute for Data Science
Fellow, Guizhou Berkeley Big Data Innovation Research Center

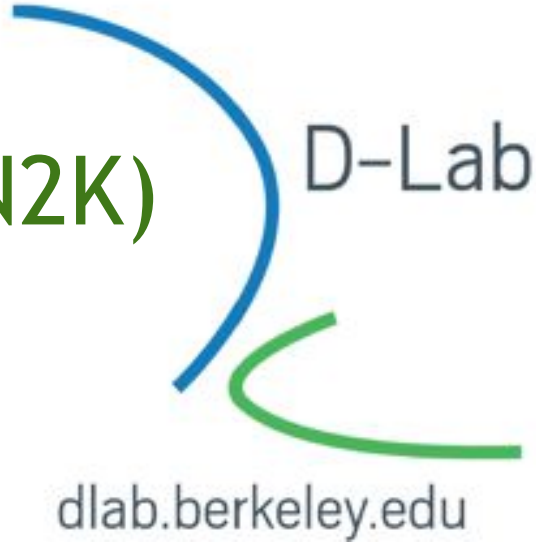
Josué Meléndez Rodríguez, MA, MSW
Qualitative Research Lead & Senior Data Science Fellow, D-Lab

Berkeley
UNIVERSITY OF CALIFORNIA

BIDS
BERKELEY INSTITUTE
FOR DATA SCIENCE

D-Lab
dlab.berkeley.edu

- It’s Okay Not to Know (IOKN2K)
- 280 workshops
- 1,100 consultations
- working groups
- special research projects
- approx. 6,000 scholars served per year



Agenda

- Introductions
- Qualitative Research with QDA Software (Josué)
  - Qualitative Coding & Analysis
  - Data in Sample Project
  - Qualitative Data Analysis Software
- Qualitative Research with Machine Learning (Marla)
  - The case for using machine learning
  - What is Machine Learning?
  - Unsupervised ML for theming
  - An Example
- Questions and Discussion

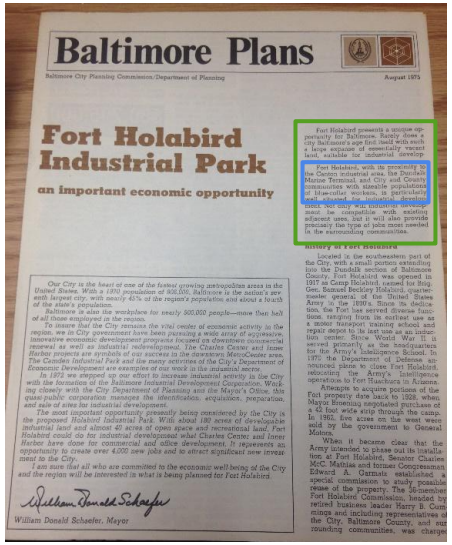
Qualitative Research with QDA Software

Qualitative Research...

What are Codes? What is Coding?

Coding is a way of organizing the data around some common idea, concept, or category ACROSS sources.

A



Baltimore Plans

Baltimore City Planning Commission/Department of Planning — August 1975

Fort Holabird Industrial Park

an important economic opportunity

Fort Holabird presents a unique opportunity for Baltimore. Rarely does a city Baltimore's age find itself with such a large expanse of essentially vacant land, suitable for industrial develop-

Fort Holabird, with its proximity to the Canton industrial area, the Dundalk Marine Terminal and City and County communities with sizable populations of blue-collar workers, is particularly well situated for industrial develop-

ment. Not only will industrial develop-ment be compatible with existing adjacent uses, but it will also provide precisely the type of jobs most needed in the surrounding communities.

History of Fort Holabird

Located in the southeastern part of the City, with a small portion extending into the Dundalk section of Baltimore County, Fort Holabird was opened in 1917 as Camp Holabird, named for Brig. Gen. Samuel Beckley Holabird, quarter-master general of the United States Army in the 1890's. Since its dedica-tion, the Fort has served diverse func-tions, ranging from its earliest use as a motor transport training school and repair depot to its last use as an induc-tion center. Since World War II, it served primarily as the headquarters for the Army's Intelligence School. In 1970 the Department of Defense an-nounced plans to close Fort Holabird, relocating the Army's intelligence operations to Fort Huachuca in Arizona.

Attempts to acquire portions of the Fort property date back to 1928, when Mayor Broening negotiated purchase of a 45 foot wide strip through the camp. In 1962, five acres on the west were sold by the government to General Motors.

When it became clear that the Army intended to phase out its installa-tion at Fort Holabird, Senator Charles McC. Mathias and former Congressman Edward A. Garmatz established a special commission to study possible re-lease of the property. The 36-member Fort Holabird Commission, headed by retired business leader Harry B. Cum-mings and including representatives of the City, Baltimore County, and sur-rounding communities, was charged

Our City is the heart of one of the fastest growing metropolitan areas in the United States. With a 1970 population of 905,300, Baltimore is the nation's sev-enth largest city, with nearly 40% of the region's population and about a fourth of the state's population.

Baltimore is also the workplace for nearly 500,000 people—more than half of all those employed in the region.

To insure that the City remains the vital center of economic activity in the region, we in City government have been pursuing a wide array of aggressive, innovative economic development programs focused on downtown commercial renewal as well as industrial redevelopment. The Charles Center and Inner Harbor projects are symbols of our success in the downtown MetroCenter area. The Camden Industrial Park and the many activities of the City's Department of Economic Development are examples of our work in the industrial sector.

In 1973 we stepped up our effort to increase industrial activity in the City with the formation of the Baltimore Industrial Development Corporation. Work-ing closely with the City Department of Planning and the Mayor's Office, this quasi-public corporation manages the identification, acquisition, preparation, and sale of sites for industrial development.

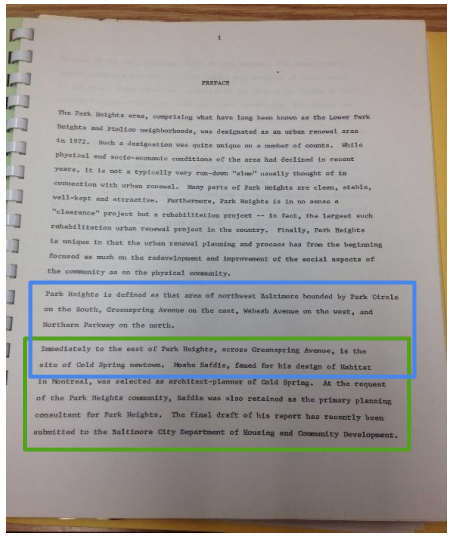
The most important opportunity presently being considered by the City is the proposed Holabird Industrial Park. With about 180 acres of developable industrial land and almost 40 acres of open space and recreational land, Fort Holabird could do for industrial development what Charles Center and Inner Harbor have done for commercial and office development. It represents an opportunity to create over 4,000 new jobs and to attract significant new invest-ment to the City.

I am sure that all who are committed to the economic well being of the City and the region will be interested in what is being planned for Fort Holabird.

William Donald Schaefer

William Donald Schaefer, Mayor

B



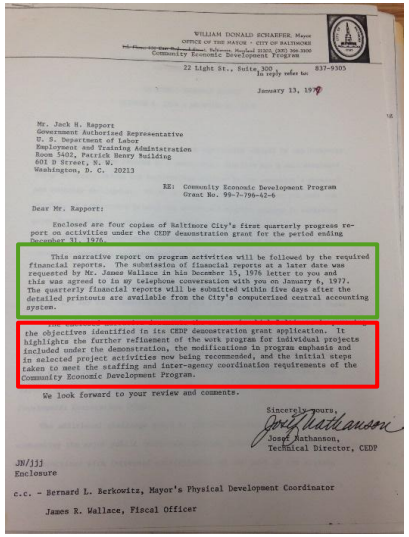
PREFACE

The Park Heights area, comprising what have long been known as the Lower Park Heights and Pimlico neighborhoods, was designated as an urban renewal area in 1972. Such a designation was quite unique on a number of counts. While physical and socio-economic conditions of the area had declined in recent years, it is not a typically very run-down "slum" usually thought of in connection with urban renewal. Many parts of Park Heights are clean, stable, well-kept and attractive. Furthermore, Park Heights is in no sense a "clearance" project but a rehabilitation project -- in fact, the largest such rehabilitation urban renewal project in the country. Finally, Park Heights is unique in that the urban renewal planning and process has from the beginning focused as much on the redevelopment and improvement of the social aspects of the community as on the physical community.

Park Heights is defined as that area of northwest Baltimore bounded by Park Circle on the South, Greenspring Avenue on the east, Wabash Avenue on the west, and Northern Parkway on the north.

Immediately to the east of Park Heights, across Greenspring Avenue, is the site of Cold Spring newtown. Moshe Safdie, famed for his design of Habitat in Montreal, was selected as architect-planner of Cold Spring. At the request of the Park Heights community, Safdie was also retained as the primary planning consultant for Park Heights. The final draft of his report has recently been submitted to the Baltimore City Department of Housing and Community Development.

C



WILLIAM DONALD SCHAEFER, Mayor
OFFICE OF THE MAYOR - CITY OF BALTIMORE
Community Economic Development Programs

22 Light St., Suite 300, 837-9305
In reply refer to:

January 13, 1977

Mr. Jack H. Rapport
Government Authorized Representative
U. S. Department of Labor
Employment and Training Administration
Room 5402, Patrick Henry Building
601 D Street, N.W.
Washington, D. C. 20213

RE: Community Economic Development Program
Grant No. 99-7-796-42-6

Dear Mr. Rapport:

Enclosed are four copies of Baltimore City's first quarterly progress re-port on activities under the CEDP demonstration grant for the period ending December 31, 1976.

This narrative report on program activities will be followed by the required financial reports. The submission of financial reports at a later date was requested by Mr. James Wallace in his December 15, 1976 letter to you and this was agreed to in my telephone conversation with you on January 6, 1977. The quarterly financial reports will be submitted within five days after the detailed printouts are available from the City's computerized central accounting system.

the objectives identified in its CEDP demonstration grant application. It highlights the further refinement of the work program for individual projects included under the demonstration, the modifications in program emphasis and in selected project activities now being recommended, and the initial steps taken to meet the staffing and inter-agency coordination requirements of the Community Economic Development Program.

We look forward to your review and comments.

Sincerely yours,

Josef Nathanson
Technical Director, CEDP

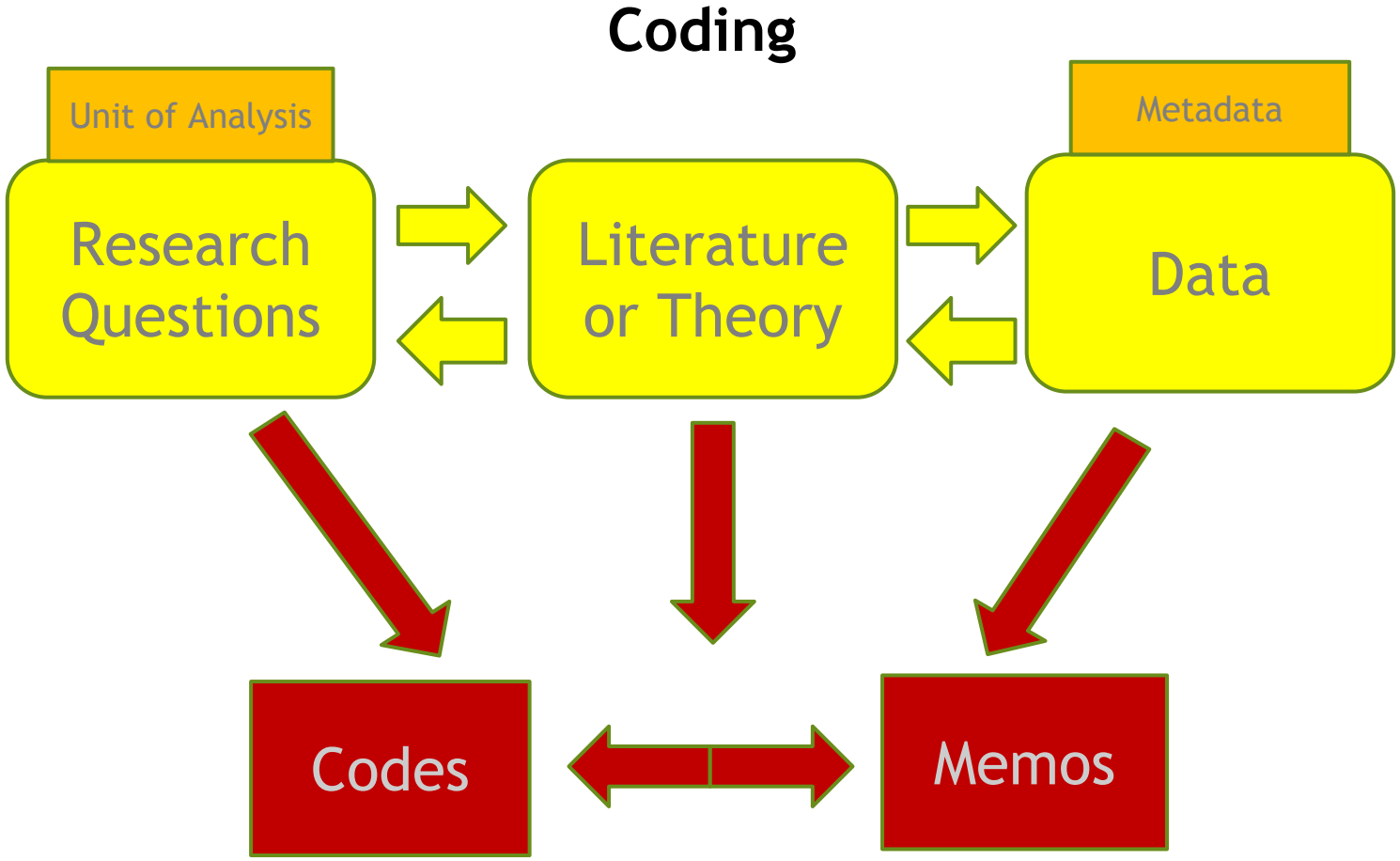
JN/jjj
Enclosure

c.c. - Bernard L. Berkowitz, Mayor's Physical Development Coordinator
James R. Wallace, Fiscal Officer

The code of "financial planning" is applied to the selected text from documents A, B, and C, because they all discuss this topic.

Qualitative Research...

Coding as a Multistep Nonlinear Process



Qualitative Research...

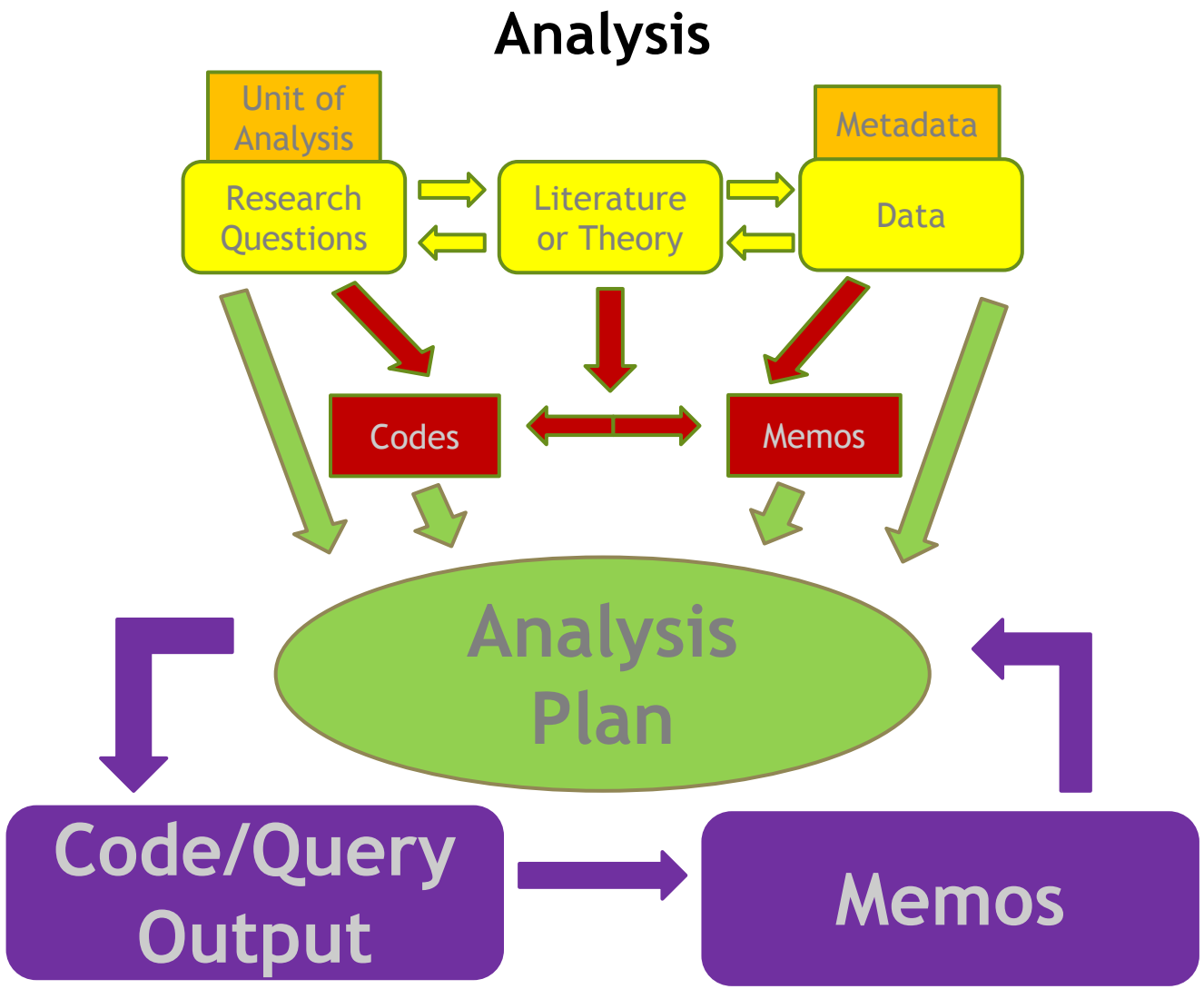
What is Analysis?

The process of identifying themes related to your research findings. This is different than identifying ideas/concepts/topics that come up throughout your data set. It's "bigger picture" stuff…

- Overarching Themes
  - What is happening in your data overall?
- Subgroup Themes
  - What is happening in your data for specific subgroups?
- Typology Themes
  - What is happening in your data by specific dimensions of coded data?

Qualitative Research...

Analysis as a Multistep Nonlinear Process



Qualitative Research...

Data Used in Sample Project

Data Sources

journal articles focusing on intersectionality and social work practice

- Intersectionality as a Useful Tool: Anti-Oppressive SW and Critical Reflection
- Children's Health Disparities as Embodiment of Social Class
- Recognizing Students of Color as Holders and Creators of Knowledge
- Undocumented Mexican Mothers in the Current Policy Context
- Identity, Oppression, and Power: Feminisms and Intersectionality Theory
- African American Women's Perceptions of Intersection of DV & HIV/AIDS

Codes

- Intersectionality
- Practice Level
  - Micro
  - Macro
- Privilege and Oppression
  - Sexism
  - Racism
  - Classism
  - Heteronormativity
  - Cisgenderism
  - Ableism
  - ...
- Practice Recommendations

Qualitative Research...

QDA Software as a Tool for Coding & Analysis

What It Does

- Structure & Organize
- Code & Retrieve
- Memo
- Explore
- Query
- Visualize

What It Does Not

- Error-Free Auto Coding
- Analytic Thinking
- Eliminate Bias
- Advanced Quantitative Analysis
  - MAXQDA engages in some quantitative analysis

Potential Benefits

- Frees Time to Focus on Analysis
- Can Deal with Large Data Sets
- Improves Auditability
  (among some audiences)
- Improves Credibility
  (among some audiences)

Potential Drawbacks

- Requires Learning the Software
- May Force Production of Meaningless Findings
- May Create Pressure to Engage Excessive Features

Qualitative Research...

Overview of Relevant QDA Software Programs

[QDA Miner]

[HyperResearch]

[ANSWR]

[Transana]

[Aquad] (open source)

[Quirkos] (visual exports)

[Saturate] (app-based)

Qualitative Research and Machine Learning

The case for machine assisted text analysis

The Qualitative Report

Volume 16 | Number 3 Article 6

5-1-2011

Compatibility between Text Mining and Qualitative Research in the Perspectives of Grounded Theory, Content Analysis, and Reliability

Chong Ho Yu
Arizona State University, chonghoyu@gmail.com

Angel Jannasch-Pennell
Arizona State University, angel@asu.edu

Samuel DiGangi
Arizona State University, sam@asu.edu

99 citations

Benefits of Machine Learning

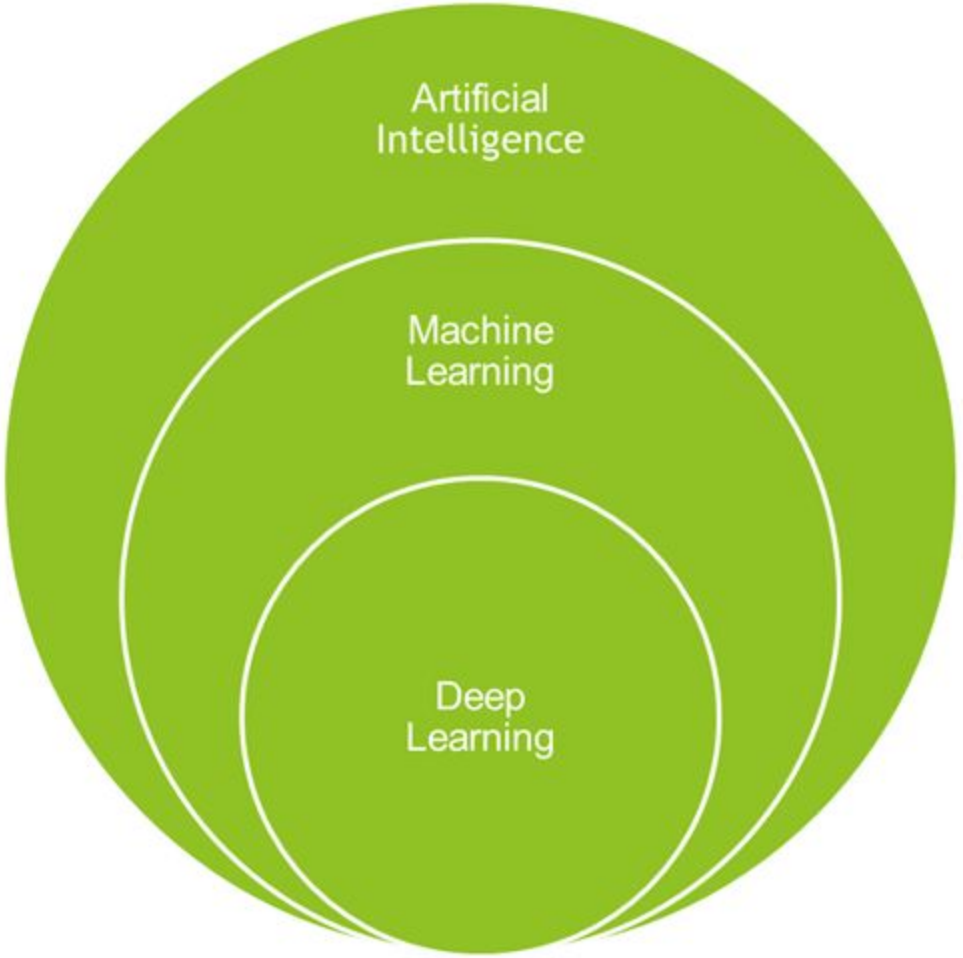
Natural extension of Content Analysis and Grounded Theory (data-informed)

Facilitates use of more text

Replicable
Transparent
Fit statistics
More reliable

Efficient

What is Machine Learning?



AI: Techniques that enables machines to mimic human logic, make predictions, and act on those predictions.

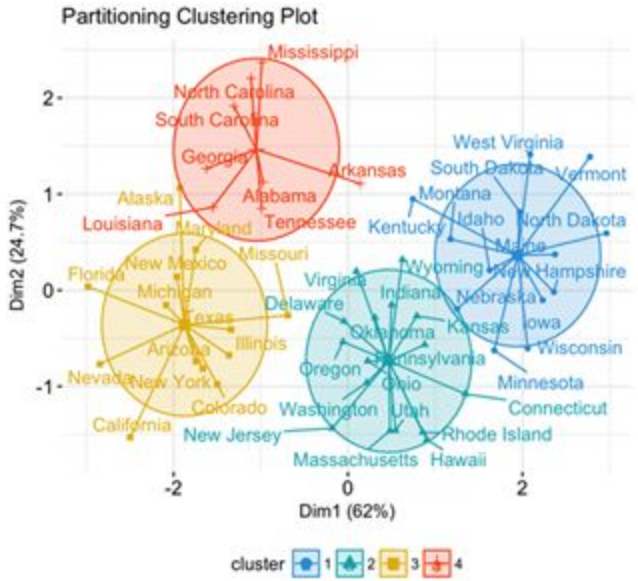
ML: Techniques by which a computer deploys user-defined logical sequences and statistical methods to predict an observed or unobserved class. If predictions are inaccurate, adaptations to the rules are user-defined.

DL: Techniques by which a computer deploys user-defined logical sequences and statistical methods to predict an observed or unobserved class. If predictions are inaccurate, adaptations to the rules are computer generated.

Equations and Algorithms

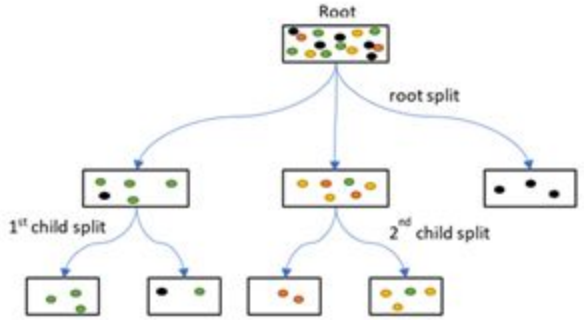
	Equations	Algorithms
	Column-based Relationships among variables relative to all observations Measures of uncertainty (model fit) well developed Statistics	Row-based Relationships among observations relative to all variables Measures of uncertainty less well developed Computer Science/Engineering
Predicting observed outcomes	Regressions	Supevised Machine Learning
Predicting unobserved outcomes	Factor Analysis Latent Class Analysis Structural Equation Modeling	Unsupervised Machine Learning

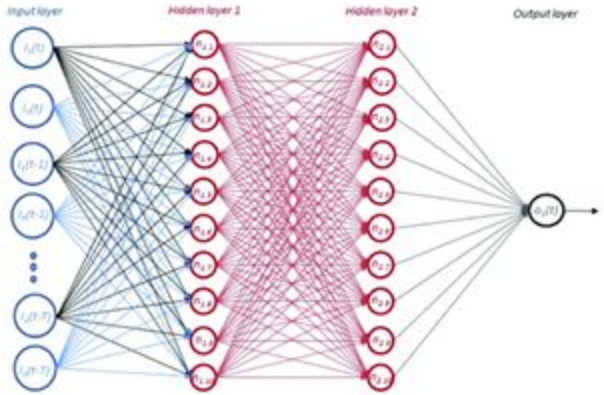
Common Machine Learning Methods



Unsupervised Models

Predicting an unobserved outcome





Supervised Models

Predicting an observed outcome

Unsupervised Machine Learning for qualitative code analysis / theming

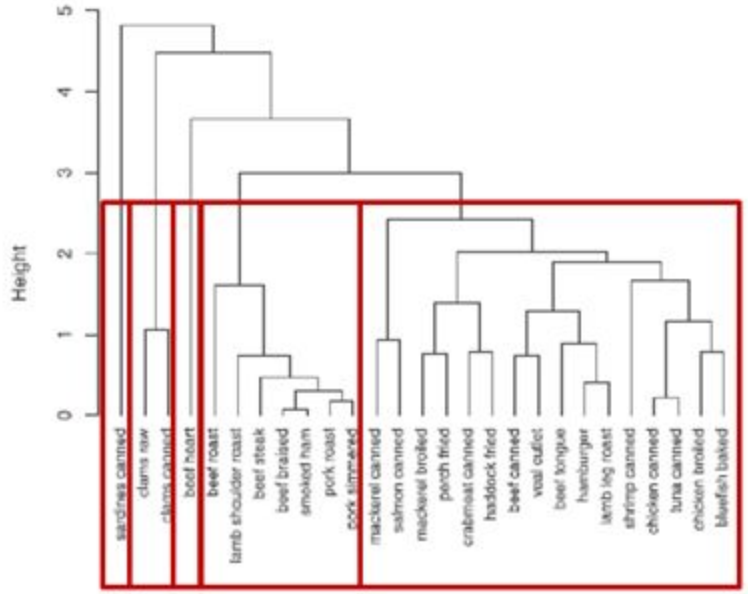
clustering: distance measure

$$d_{ji} = \sqrt{\sum_{p-1}^{p} (1x_{ip} - x_{jp})^2}$$

observation	energy	protein	fat	calcium	iron
BEEF BRAISED	340	20	28	9	2.6
HAMBURGER	245	21	17	9	2.7
BEEF ROAST	420	15	39	7	2.0
BEEF STEAK	375	19	32	9	2.6

$$d = \sqrt{(340-245)^2 + (20-21)^2 + (28-17)^2 + (9-9)^2 + (26-27)^2} = 95.64$$

hierarchical



cluster	energy	protein	fat	calcium	iron
1	340	19	29	9	2.50
2	170	20	8	13	1.45
3	160	26	5	14	5.9
4	57	9	1	78	5.70
5	180	22	9	367	2.5

Logic

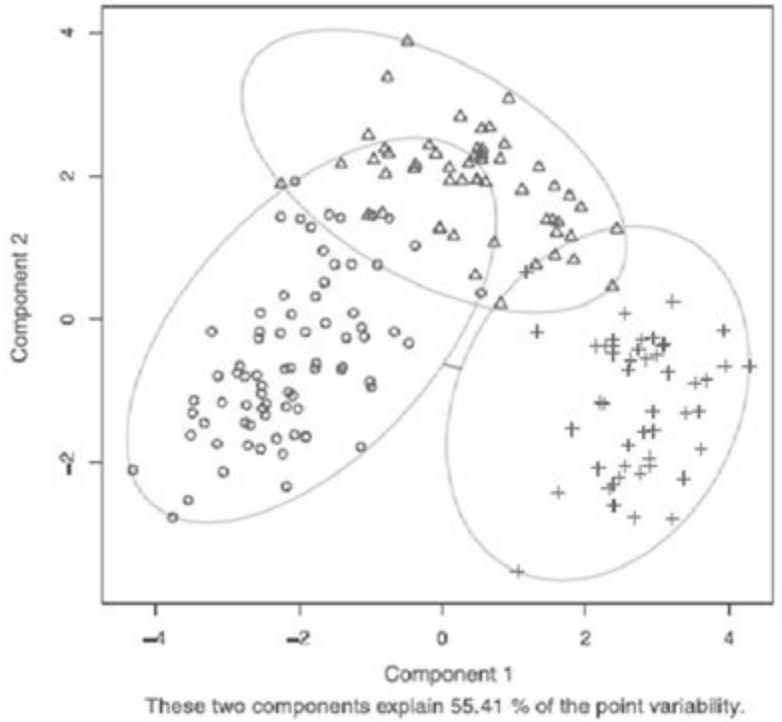
1. Each case or observation starts as its own cluster.
2. Clusters are combined two at a time until all clusters are merged into a single cluster.

Algorithm

1. Define each observation (row, case) as a cluster.
2. Calculate the distances between every cluster and every other cluster.
3. Combine the two clusters that have the smallest distance. This reduces the number of clusters by one.
4. Repeat steps 2 and 3 until all clusters have been merged into a single cluster containing all observations.

Kabacoff, R. I., (2015). Chapter 16: Cluster analysis. In R in Action: Data analysis and graphics with R (2nd Edition). Manning, Shelter Island.

partitioning: k-means and PAM



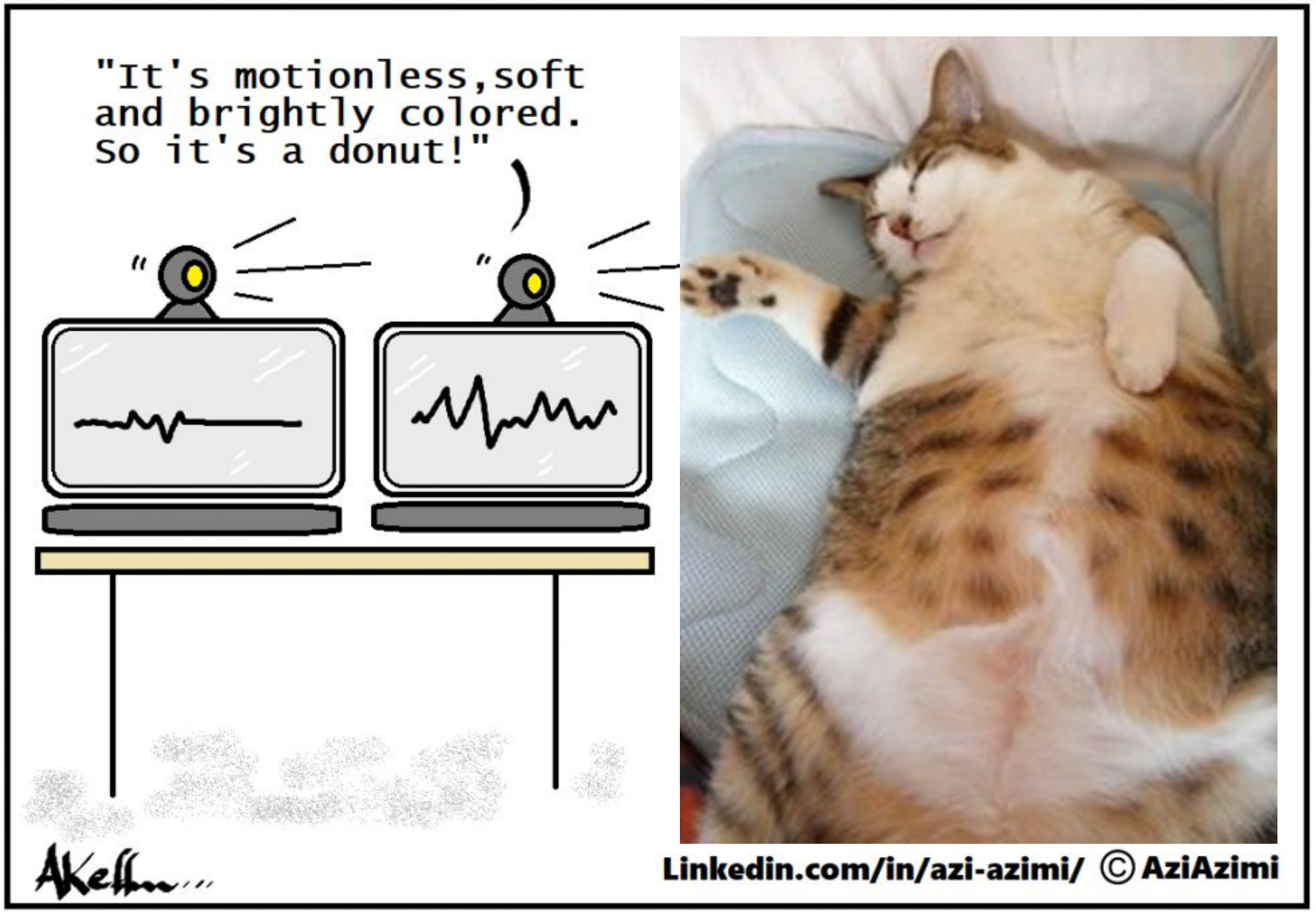
Logic

1. Observations divided into *K* groups.
2. Observations are reshuffled to form the most cohesive clusters possible.

Algorithm

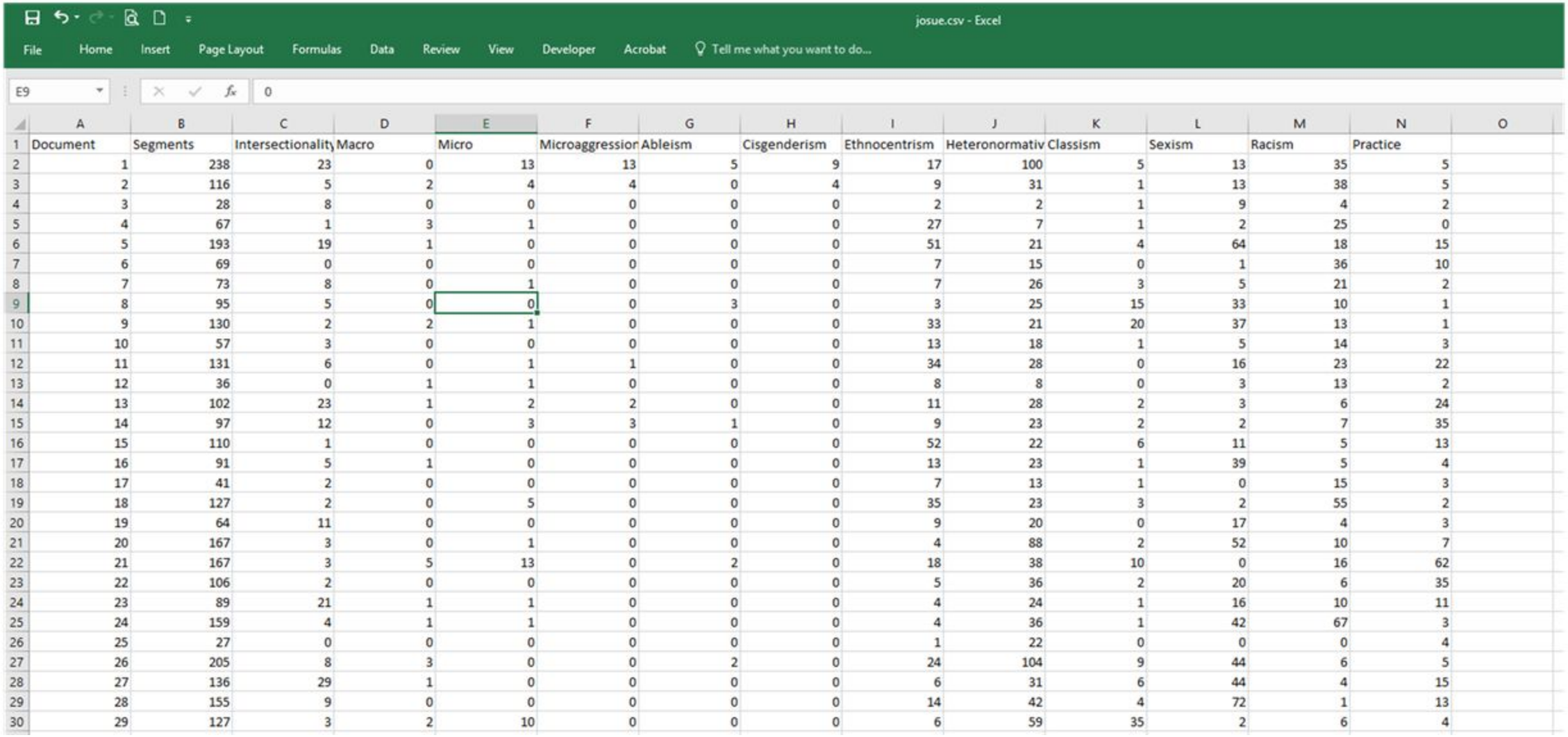
1. Select *K* centroids (*K* rows chosen at random).
2. Assign each data point to its closest centroid.
3. Recalculate the centroids as the average of all data points in a cluster.
4. Assign data points to their closest centroids.
5. Continue steps 3 and 4 until observations aren't reassigned or the maximum number of iterations is reached.

Kabacoff, R. I., (2015). Chapter 16: Cluster analysis. In R in Action: Data analysis and graphics with R (2nd Edition). Manning, Shelter Island.


"It's motionless,soft and brightly colored. So it's a donut!"
Akeem...
Linkedin.com/in/azi-azimi/ © AziAzimi

An Application to Josué's Sample Study

Data (codes)

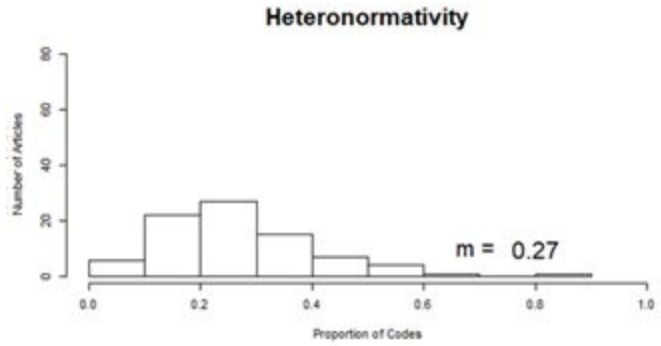

josue.csv - Excel

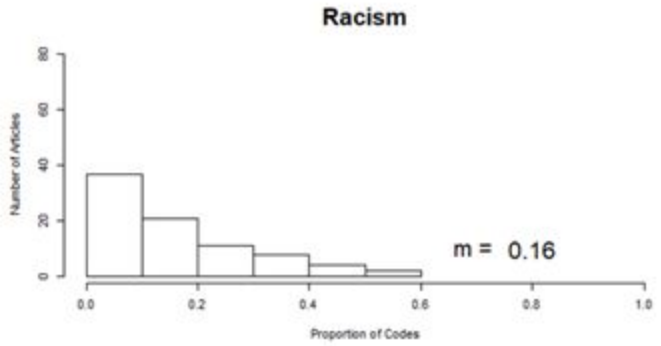
File Home Insert Page Layout Formulas Data Review View Developer Acrobat Tell me what you want to do...

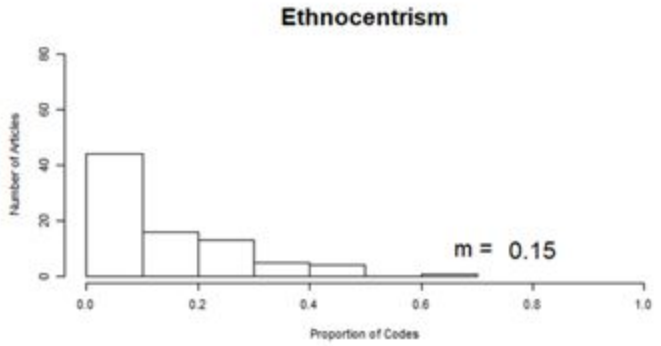
E9 0

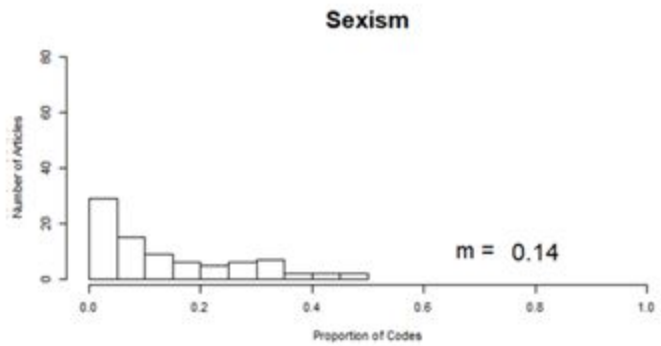
	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O
1	Document	Segments	Intersectionality	Macro	Micro	Microaggression	Ableism	Cisgenderism	Ethnocentrism	Heteronormativ	Classism	Sexism	Racism	Practice	
2	1	238	23	0	13	13	5	9	17	100	5	13	35	5	
3	2	116	5	2	4	4	0	4	9	31	1	13	38	5	
4	3	28	8	0	0	0	0	0	2	2	1	9	4	2	
5	4	67	1	3	1	0	0	0	27	7	1	2	25	0	
6	5	193	19	1	0	0	0	0	51	21	4	64	18	15	
7	6	69	0	0	0	0	0	0	7	15	0	1	36	10	
8	7	73	8	0	1	0	0	0	7	26	3	5	21	2	
9	8	95	5	0	0	0	3	0	3	25	15	33	10	1	
10	9	130	2	2	1	0	0	0	33	21	20	37	13	1	
11	10	57	3	0	0	0	0	0	13	18	1	5	14	3	
12	11	131	6	0	1	1	0	0	34	28	0	16	23	22	
13	12	36	0	1	1	0	0	0	8	8	0	3	13	2	
14	13	102	23	1	2	2	0	0	11	28	2	3	6	24	
15	14	97	12	0	3	3	1	0	9	23	2	2	7	35	
16	15	110	1	0	0	0	0	0	52	22	6	11	5	13	
17	16	91	5	1	0	0	0	0	13	23	1	39	5	4	
18	17	41	2	0	0	0	0	0	7	13	1	0	15	3	
19	18	127	2	0	5	0	0	0	35	23	3	2	55	2	
20	19	64	11	0	0	0	0	0	9	20	0	17	4	3	
21	20	167	3	0	1	0	0	0	4	88	2	52	10	7	
22	21	167	3	5	13	0	2	0	18	38	10	0	16	62	
23	22	106	2	0	0	0	0	0	5	36	2	20	6	35	
24	23	89	21	1	1	0	0	0	4	24	1	16	10	11	
25	24	159	4	1	1	0	0	0	4	36	1	42	67	3	
26	25	27	0	0	0	0	0	0	1	22	0	0	0	4	
27	26	205	8	3	0	0	2	0	24	104	9	44	6	5	
28	27	136	29	1	0	0	0	0	6	31	6	44	4	15	
29	28	155	9	0	0	0	0	0	14	42	4	72	1	13	
30	29	127	3	2	10	0	0	0	6	59	35	2	6	4	

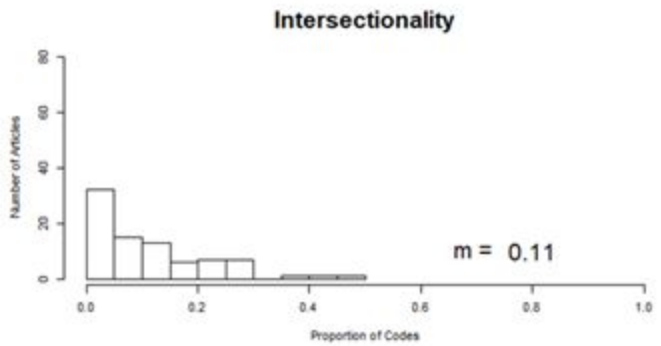
Codes distribution

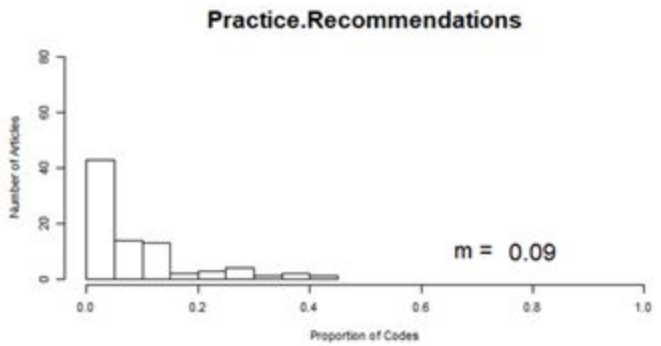


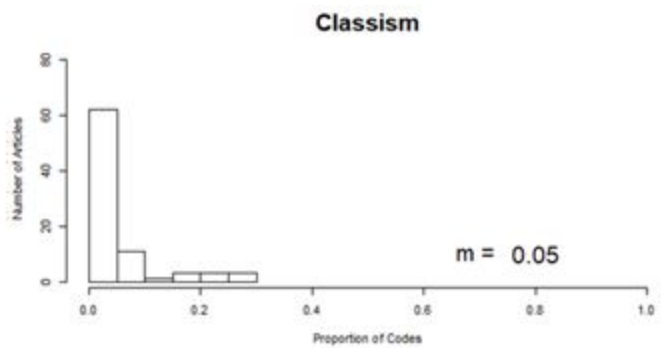


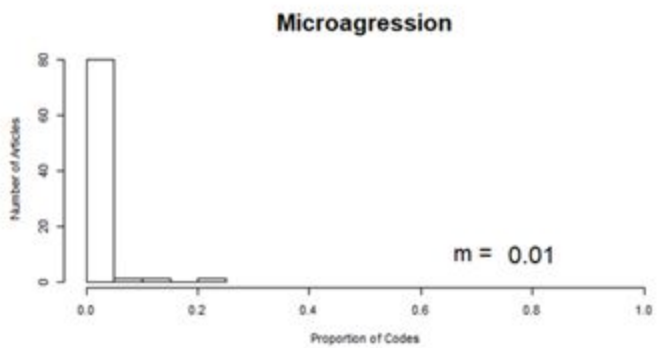


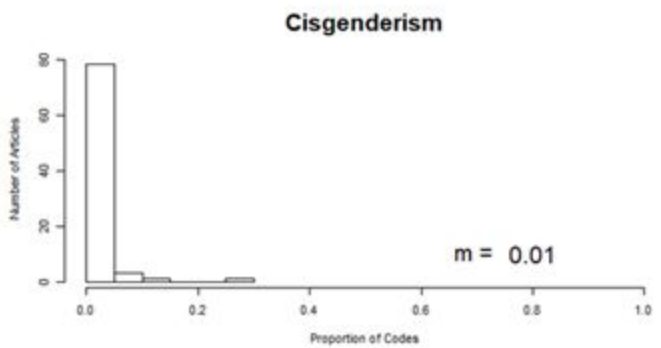








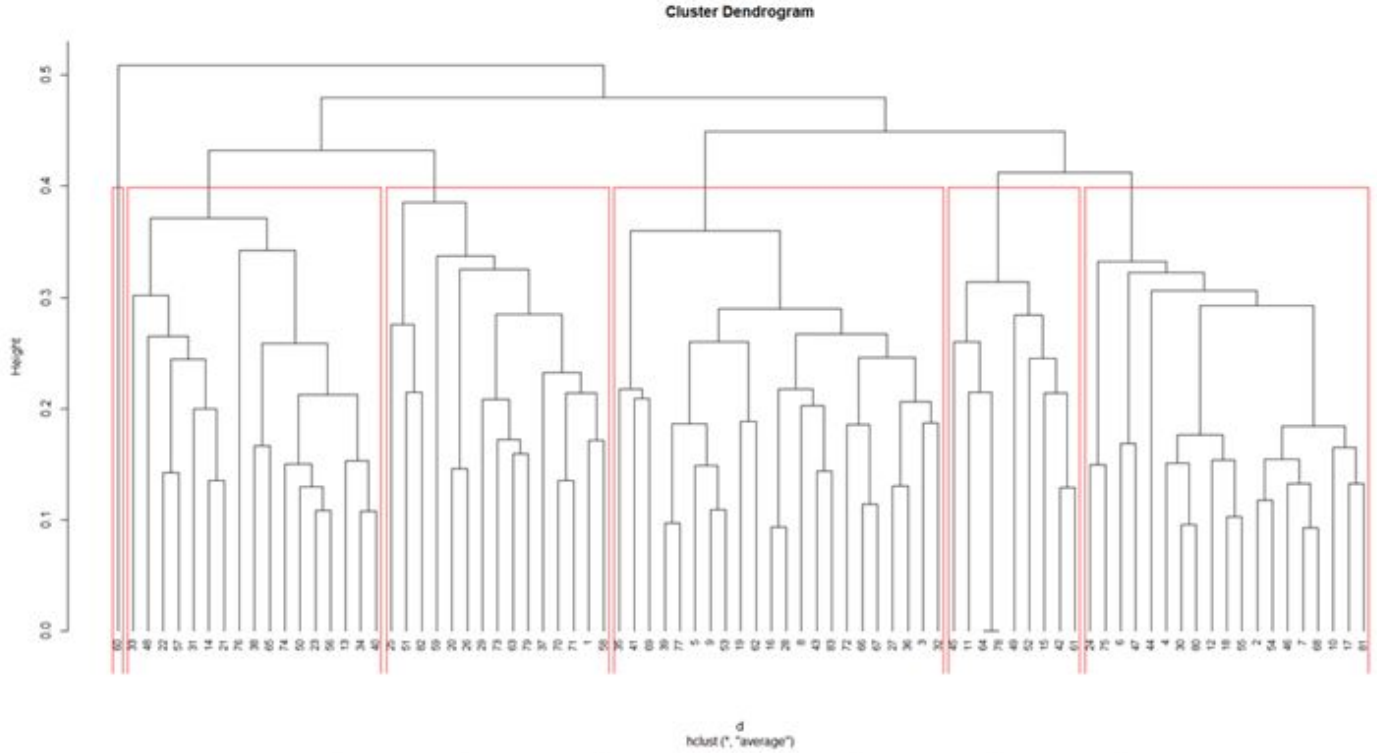




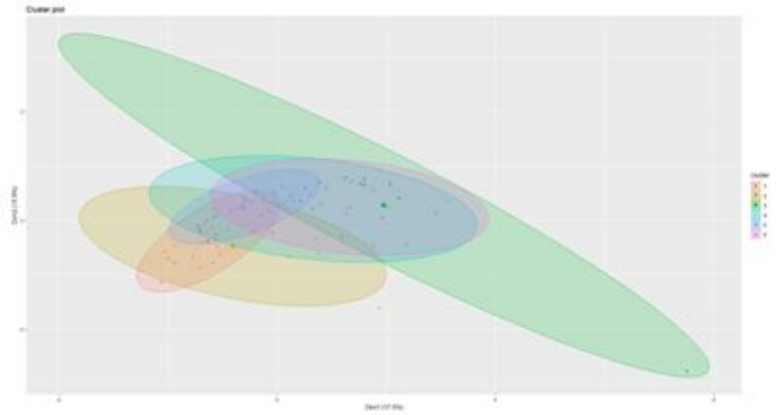
How many clusters?

```
*****************************************************************
* Among all indices:
* 3 proposed 2 as the best number of clusters
* 4 proposed 3 as the best number of clusters
* 1 proposed 4 as the best number of clusters
* 1 proposed 5 as the best number of clusters
* 6 proposed 6 as the best number of clusters
* 1 proposed 7 as the best number of clusters
* 3 proposed 9 as the best number of clusters
* 1 proposed 13 as the best number of clusters
* 3 proposed 20 as the best number of clusters

***** Conclusion *****

* According to the majority rule, the best number of clusters is  6

*****************************************************************
```

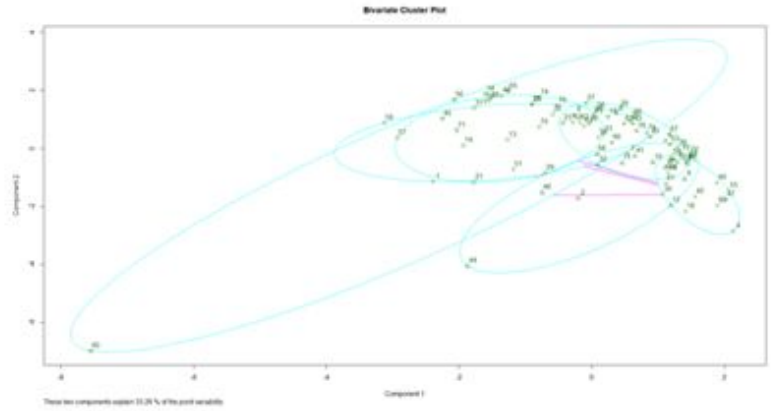
Comparing Models



1	2	3	4	5	6
15	19	22	9	17	1



1	2	3	4	5	6
14	16	10	13	19	11



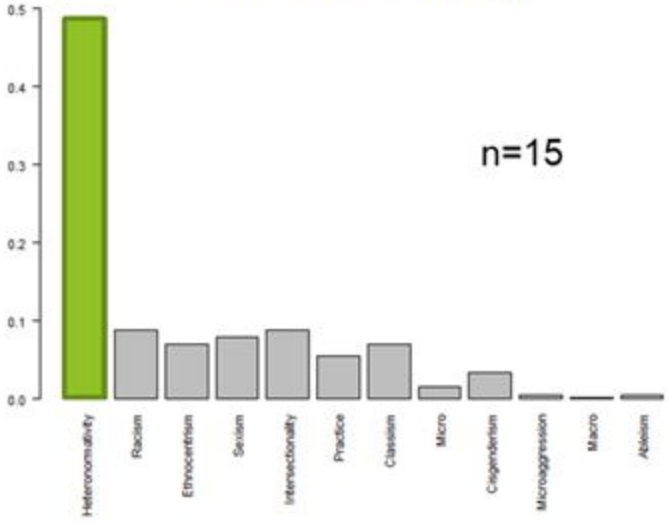
1	2	3	4	5	6
16	17	15	13	8	14

Confirming Model Choice

Codes	Hierarchical Model p	KMeans Model p	PAM Model p
Intersectionality	<0.001	<0.001	<0.001
Macro	0.374	0.870	0.397
Micro	<0.001	0.346	0.692
Microaggression	<0.001	0.430	0.605
Ableism	0.007	0.002	0.020
Cisgenderism	0.003	0.085	0.106
Ethnocentrism	<0.001	<0.001	<0.001
Heteronormativity	<0.001	<0.001	<0.001
Classism	0.187	0.103	0.277
Sexism	<0.001	<0.001	<0.001
Racism	<0.001	<0.001	<0.001
Practice	<0.001	<0.001	<0.001

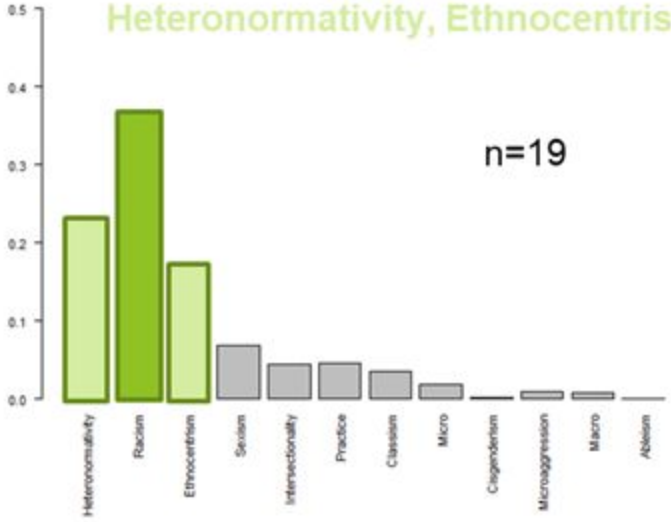
Hierarchical Clusters

Heteronormativity



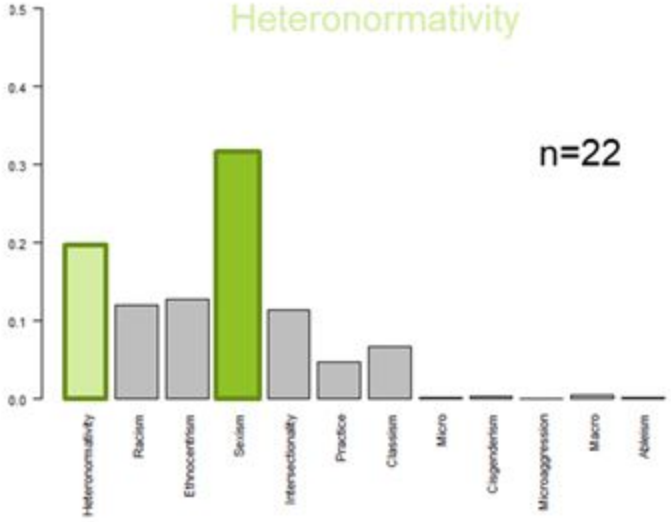
Racism

Heteronormativity, Ethnocentrism



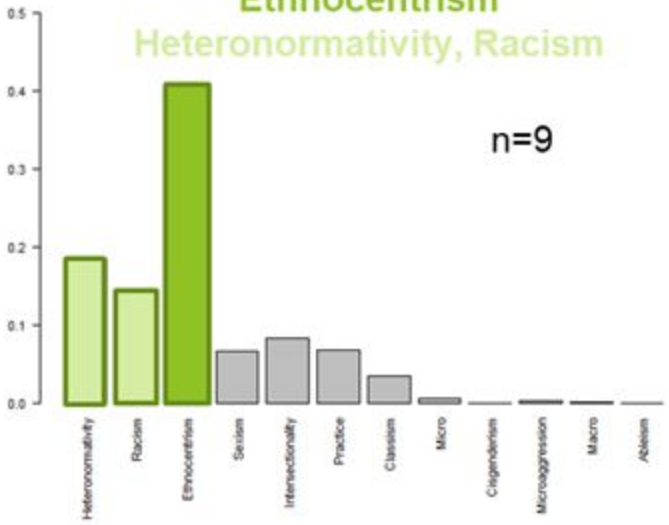
Sexism

Heteronormativity

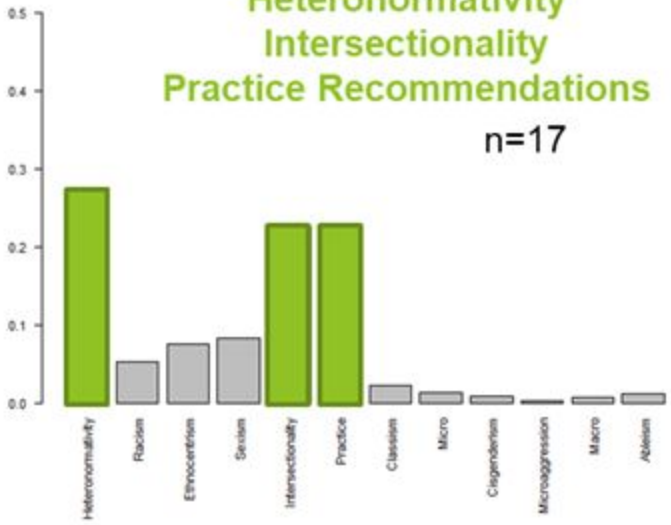


Ethnocentrism

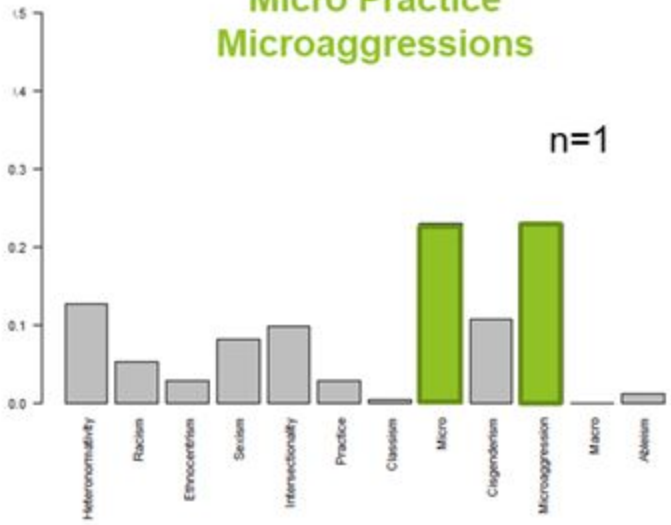
Heteronormativity, Racism



Heteronormativity Intersectionality Practice Recommendations



Micro Practice Microaggressions



Hierarchical Clusters: Sample Articles

1.
Heteronormativity

Exploring Biculturality and Beauty Standards in Sexual Minority Women, *Psychology of Sexual Orientation and Gender Diversity* (2015)

2.
Racism
Heteronormativity, Ethnocentrism

"Antiblackness in Education Policy and Discourse", *Theory Into Practice* (2016)

3.
Sexism
Heteronormativity

"Capitalism and Welfare Reform: Who Really Benefits?", *Race, Gender and Class* (2008)

4.
Ethnocentrism
Heteronormativity, Racism

"Discrimination and Substance Use Disorders among Latinos", *American Journal of Public Health* (2014)

5.
Heteronormativity
Intersectionality
Practice Recommendations

Intersectionality and Social Work: Omissions of Race, Class, and Sexuality in Graduate School Education, *Journal of Social Work Education* (2016)

6.
Micro Practice
Microaggressions

Sexual Orientation and Gender Identity Microaggression, *Journal of Ethnic and Cultural Diversity in Social Work* (2017)

Questions / Discussion