DRAFT

to be presented at F. Lane Hardy seminar post-pandemic

Visualizing human diversity in data with hierarchical clustering analysis

Dr. Julie C. La Corte
Georgia State University

First draft: May 3, 2020
Last revised: Dec. 1, 2020

Declarative vs. behavioral data

"When we ask, and when we measure"

Declarative data is self-reported information about individuals' characterizations of themselves, their beliefs, and their opinions.

Behavioral data is information about individuals' observed behaviors.

Declarative vs. behavioral data

Declarative data collection

Declarative data is self-reported information about individuals' characterizations of themselves, their beliefs, and their opinions.

- Respondents give consent, then willingly volunteer data about themselves when asked
- The quality of the data depends on...
  - the respondent's honesty and understanding of themselves
  - the researcher's instrument and data collection methods *(e.g., survey design and sample selection)*

Declarative vs. behavioral data

Declarative data collection

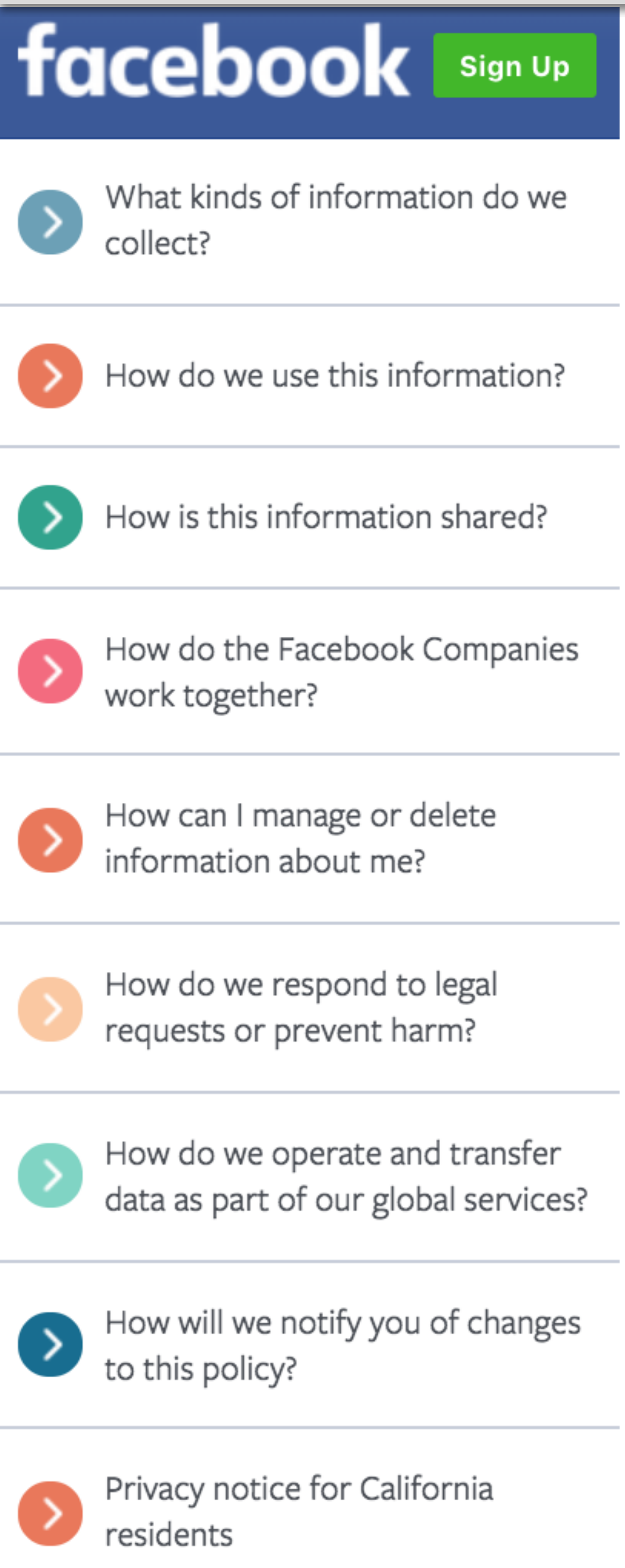
Declarative data is self-reported information about individuals' characterizations of themselves, their beliefs, and their opinions.

- Respondents give consent, then willingly volunteer data about themselves when asked

- The quality of the data depends on...
  - the respondent's honesty and understanding of themselves
  - the researcher's instrument and data collection methods *(e.g., survey design and sample selection)*

Declarative vs. behavioral data

Behavioral data collection



What kinds of information do we collect?

How do we use this information?

How is this information shared?

How do the Facebook Companies work together?

How can I manage or delete information about me?

How do we respond to legal requests or prevent harm?

How do we operate and transfer data as part of our global services?

How will we notify you of changes to this policy?

Privacy notice for California residents

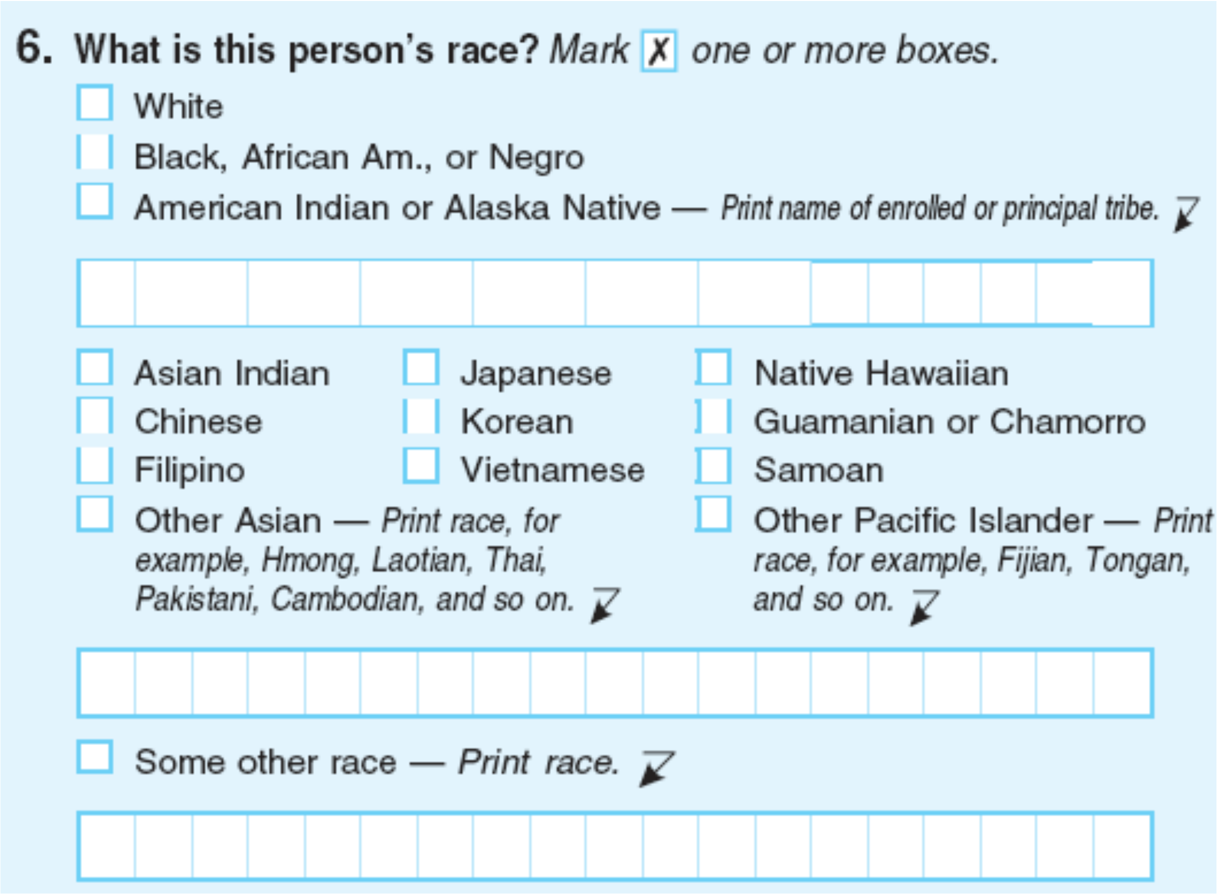
Behavioral data is information about individuals' observed behaviors.

- The researcher collects observations of subjects who may not be cognizant that they're being monitored
  - Do *you* understand Facebook's Terms of Service?
  - What does Target know about you?
- Privacy concerns
  - Individual responses should be anonymized
  - Aggregates of individuals may still be stigmatized

Declarative vs. behavioral data

When is “behavior” not behavioral data?

Behavioral data is information about individuals’ observed behaviors.



6. What is this person’s race? *Mark ☒ one or more boxes.*

□ White
□ Black, African Am., or Negro
□ American Indian or Alaska Native — *Print name of enrolled or principal tribe.*

□ Asian Indian
□ Chinese
□ Filipino
□ Other Asian — *Print race, for example, Hmong, Laotian, Thai, Pakistani, Cambodian, and so on.*

□ Japanese
□ Korean
□ Vietnamese

□ Native Hawaiian
□ Guamanian or Chamorro
□ Samoan
□ Other Pacific Islander — *Print race, for example, Fijian, Tongan, and so on.*

□ Some other race — *Print race.*

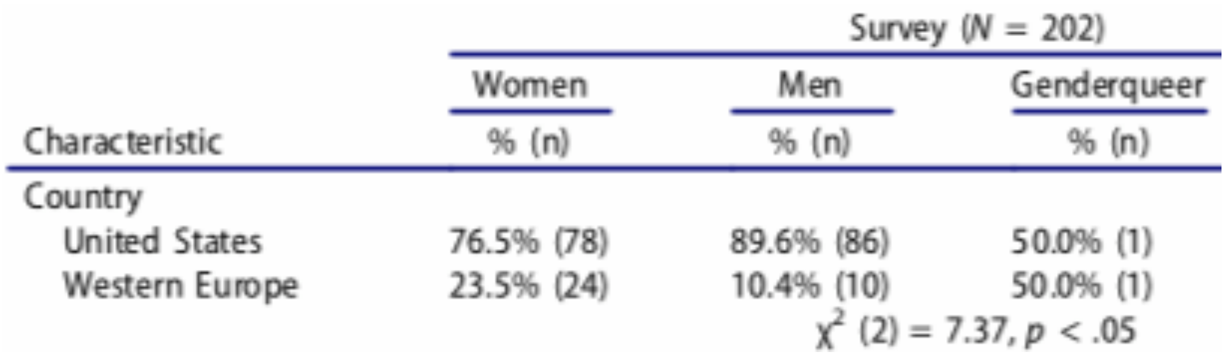
From the 2010 U.S. Census

- Inappropriate for certain types of data about individuals
  - Who has the right to “observe” what race or gender you identify as?
  - Identity may or may not be a behavior—but *identifying* surely is.

Declarative vs. behavioral data

When is “behavior” not behavioral data?

Behavioral data is information about individuals’ observed behaviors.



	Survey (N = 202)		
	Women	Men	Genderqueer
Characteristic	% (n)	% (n)	% (n)
Country			
United States	76.5% (78)	89.6% (86)	50.0% (1)
Western Europe	23.5% (24)	10.4% (10)	50.0% (1)
		χ^2 (2) = 7.37, $p < .05$	

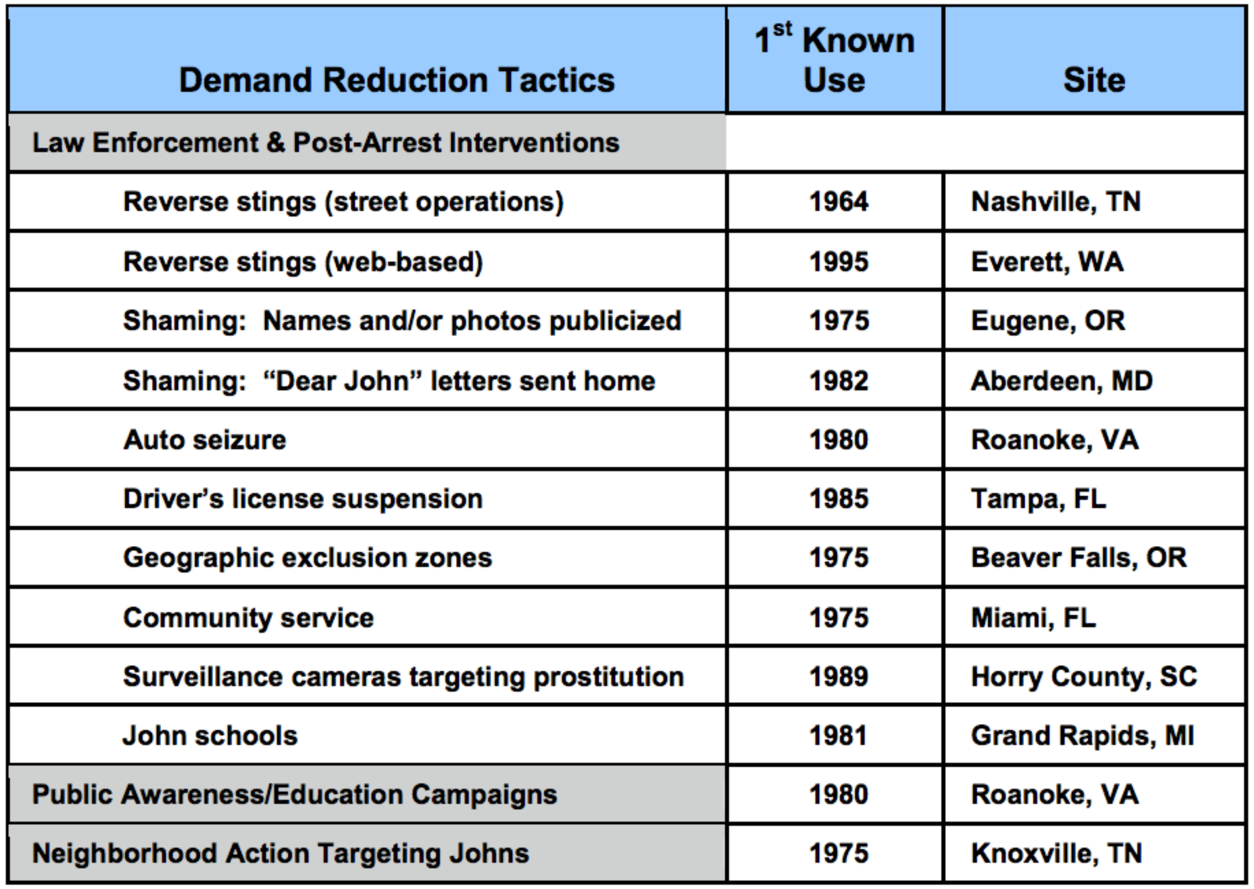
K. Martinez, “BDSM Role Fluidity: A Mixed-Methods Approach to Investigating Switches within Dominant/Submissive Binaries” *J. Homosexuality* 65(10) (2018)

- Inappropriate for “intrinsic” data about individuals
  - Who has the right to “observe” what race or gender you identify with?
  - Identity may or may not be a behavior—but *identifying* surely is.

Declarative vs. behavioral data

When is "behavior" not behavioral data?

Behavioral data is information about individuals' observed behaviors.



Demand Reduction Tactics	1st Known Use	Site
Law Enforcement & Post-Arrest Interventions		
Reverse stings (street operations)	1964	Nashville, TN
Reverse stings (web-based)	1995	Everett, WA
Shaming: Names and/or photos publicized	1975	Eugene, OR
Shaming: "Dear John" letters sent home	1982	Aberdeen, MD
Auto seizure	1980	Roanoke, VA
Driver's license suspension	1985	Tampa, FL
Geographic exclusion zones	1975	Beaver Falls, OR
Community service	1975	Miami, FL
Surveillance cameras targeting prostitution	1989	Horry County, SC
John schools	1981	Grand Rapids, MI
Public Awareness/Education Campaigns	1980	Roanoke, VA
Neighborhood Action Targeting Johns	1975	Knoxville, TN

M. Shively, et al., "A National Overview of Prostitution and Sex Trafficking Demand Reduction Efforts, Final Report" (2012)

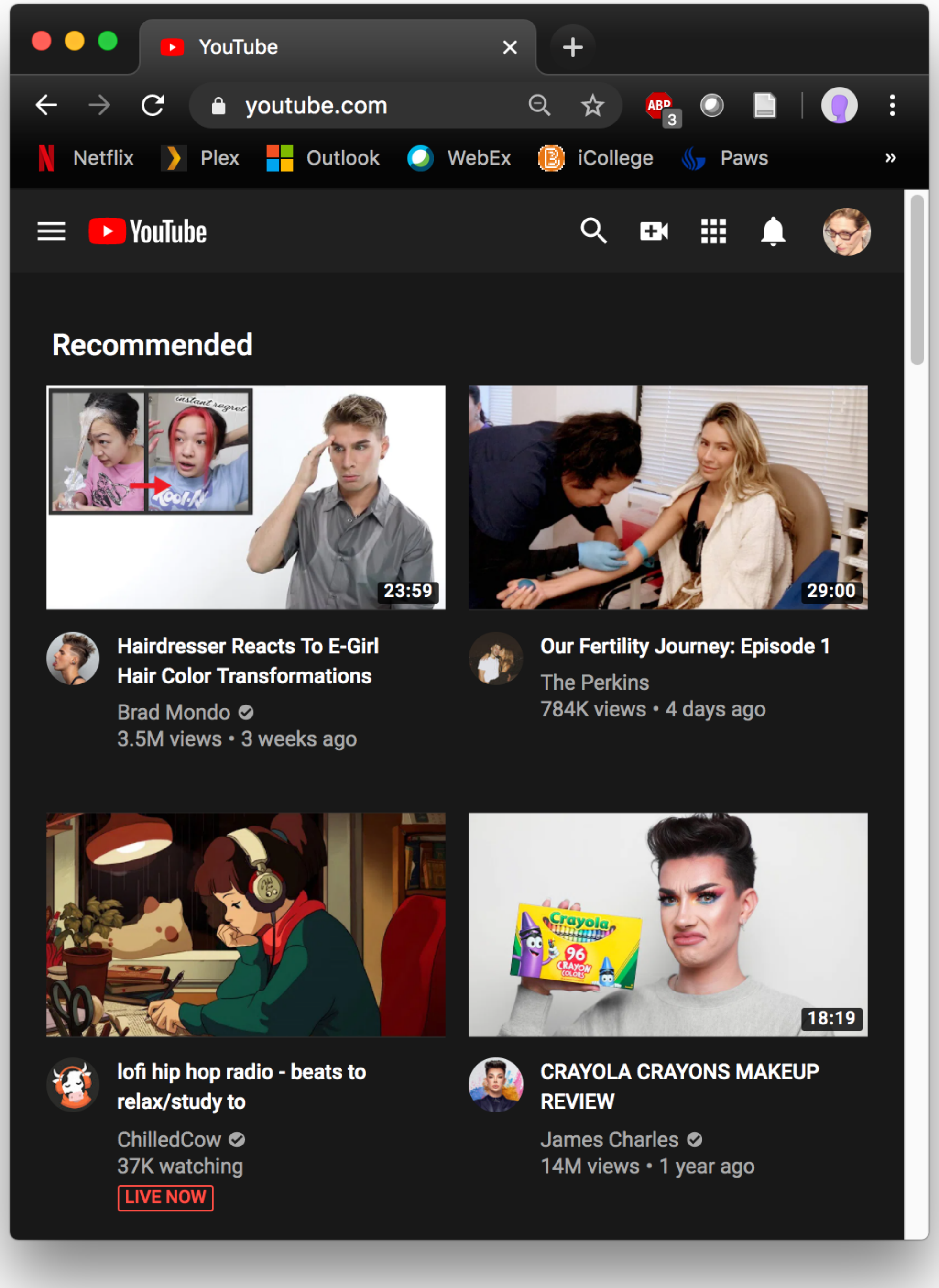
- Difficult or impossible to collect an unbiased sample for "invisible" behaviors
  - Non-normative sexual behavior
  - Unreported criminal activity
  - Reading books

Declarative vs. behavioral data

Are preferences declared or observed?

Into which category does "preference" data fall?

- Behavioral data
  - What you bought on Amazon
  - What you watched on YouTube
  - What you retweeted on Twitter
- Declarative data
  - What you say you tend to buy
  - What you say you like
  - What you say you'd never do



Recommended

Hairdresser Reacts To E-Girl Hair Color Transformations
Brad Mondo
3.5M views • 3 weeks ago
23:59

Our Fertility Journey: Episode 1
The Perkins
784K views • 4 days ago
29:00

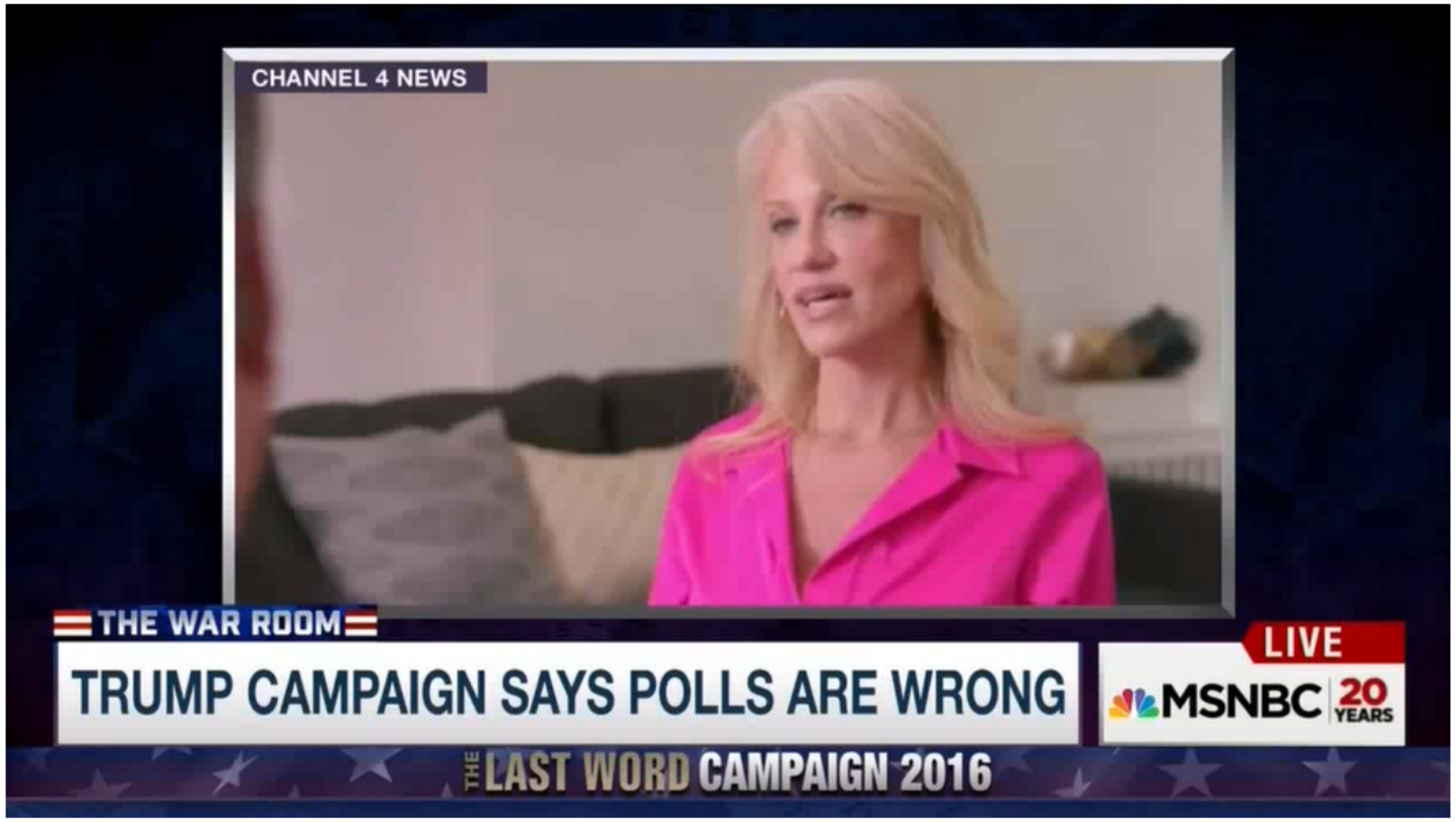
lofi hip hop radio - beats to relax/study to
ChilledCow
37K watching
LIVE NOW

CRAYOLA CRAYONS MAKEUP REVIEW
James Charles
14M views • 1 year ago
18:19

The Bradley and "shy Trump" effects

'A huge challenge for Obama, insiders say, is simply determining how much skin color will matter in November. Race is nearly impossible to poll—no one ever says "I'm a racist"—and no campaign wants it revealed they are even asking questions on the issue. "It's the uncertainty that kills me—we know it's going to be a factor, but how big a factor?" asks a Democratic operative with ties to the Obama camp. "How do you even measure such a thing?"'

G. Thrush, "7 worrisome signs for Obama," *Politico* (Aug. 11, 2008)
https://www.politico.com/story/2008/08/7-worrisome-signs-for-obama-012433


CHANNEL 4 NEWS
THE WAR ROOM
TRUMP CAMPAIGN SAYS POLLS ARE WRONG
LIVE
MSNBC 20 YEARS
THE LAST WORD CAMPAIGN 2016

attribution: MSNBC

Team Trump is counting on "hidden voters" too ashamed to declare their support in public.

'[P]eople will often hide behind the label "independent" to avoid publicly associating with either of the two parties—particularly when those parties are stigmatized. The culprit is social desirability bias. To avoid "looking bad," some people avoid answering survey questions or, even worse, outright lie.'

G. Thrush, "There may have been shy Trump supporters after all," *Washington Post* (Nov. 12, 2016)
https://www.washingtonpost.com/news/monkey-cage/wp/2016/11/12/there-may-have-been-shy-trump-supporters-after-all/

Traditional statistical analyses of declarative data

By which we mean...

- One-point estimators and confidence intervals
    - For population proportion, mean, etc.
    - For conditional probabilities
- Hypothesis testing
- Regression
    - Bivariate or multivariate
    - Parametric or nonparametric
    - Linear or nonlinear
- Analysis of variance

Traditional statistical analyses of declarative data

What can they tell us about social phenomena?

When applied to preferences, behaviors, and social identities, the aforementioned methods characterize:

- Prevalence
- Correspondence
- Longitudinal trends
- Aggregate experiences of individuals

Traditional statistical analyses of declarative data

What kinds of conclusions do we draw from them?

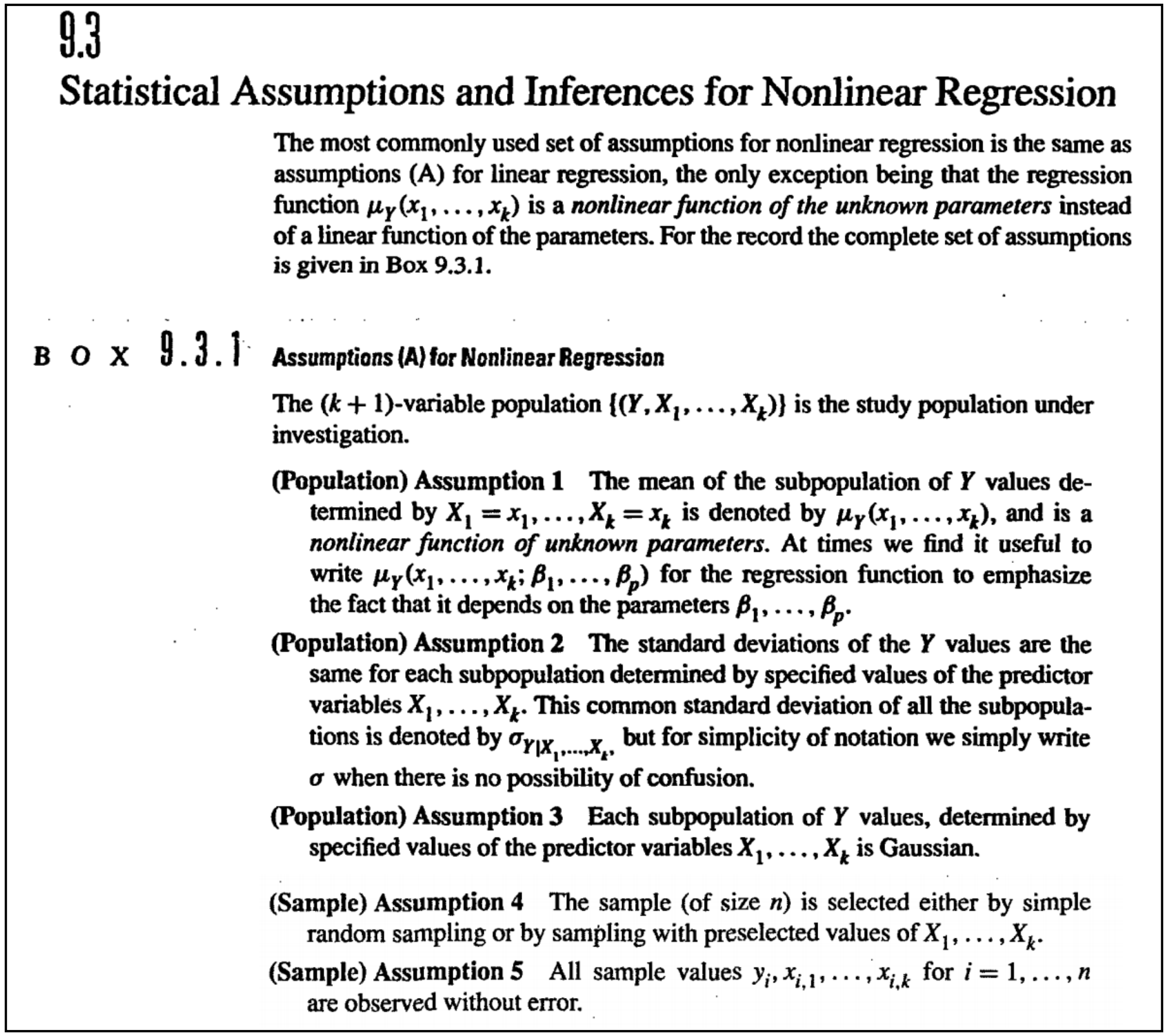
They provide evidence for or against hypotheses about:

- The effectiveness of institutional policies
- Patterns of collective behavior
- The impact of historical change
- Attitudes toward subsets of the population

Traditional statistical analyses of declarative data

Their limitations include...

Assumptions about how the population is distributed



9.3

Statistical Assumptions and Inferences for Nonlinear Regression

The most commonly used set of assumptions for nonlinear regression is the same as assumptions (A) for linear regression, the only exception being that the regression function $\mu_Y(x_1, \ldots, x_k)$ is a *nonlinear function of the unknown parameters* instead of a linear function of the parameters. For the record the complete set of assumptions is given in Box 9.3.1.

BOX 9.3.1 **Assumptions (A) for Nonlinear Regression**

The $(k+1)$-variable population $\{(Y, X_1, \ldots, X_k)\}$ is the study population under investigation.

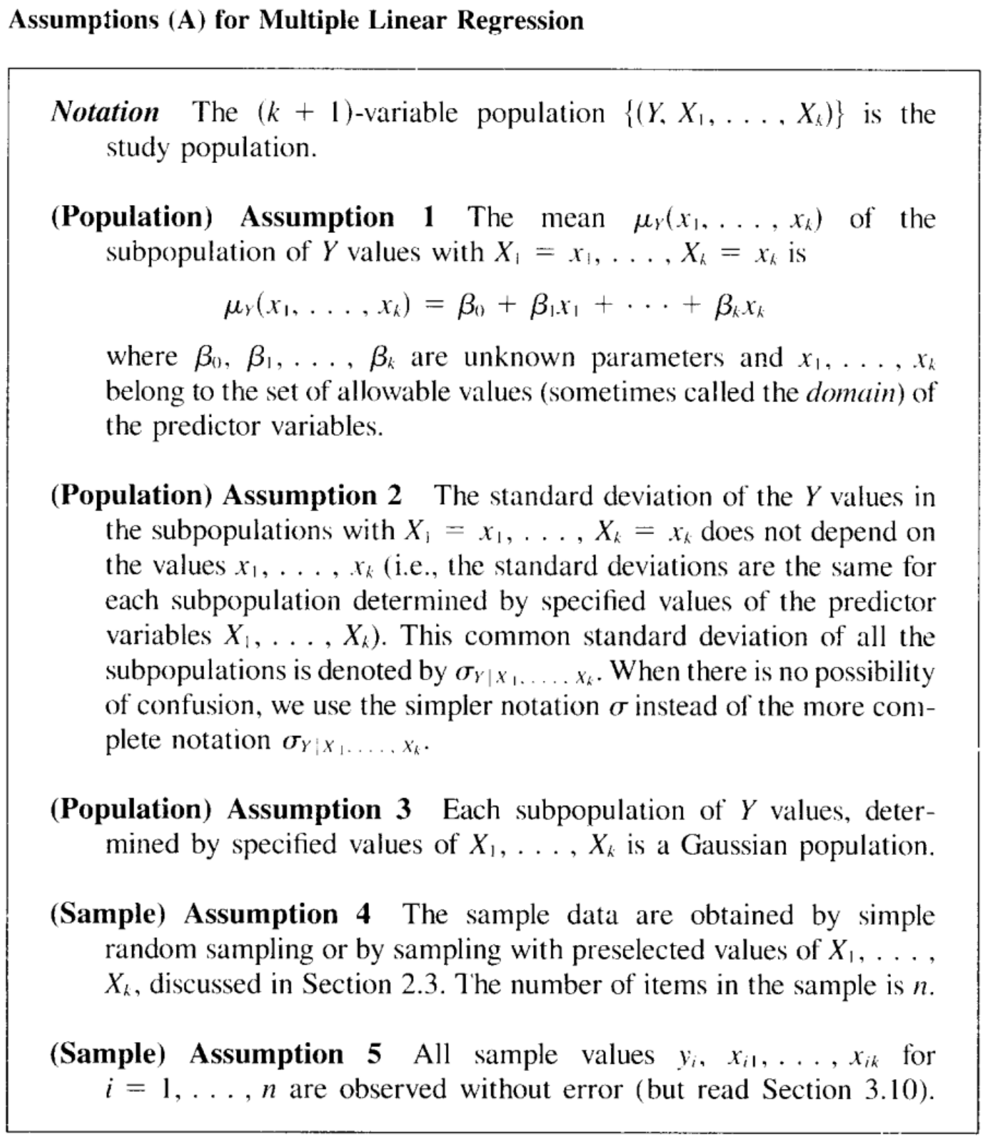
(Population) Assumption 1 The mean of the subpopulation of Y values determined by $X_1 = x_1, \ldots, X_k = x_k$ is denoted by $\mu_Y(x_1, \ldots, x_k)$, and is a *nonlinear function of unknown parameters*. At times we find it useful to write $\mu_Y(x_1, \ldots, x_k; \beta_1, \ldots, \beta_p)$ for the regression function to emphasize the fact that it depends on the parameters $\beta_1, \ldots, \beta_p$.

(Population) Assumption 2 The standard deviations of the Y values are the same for each subpopulation determined by specified values of the predictor variables $X_1, \ldots, X_k$. This common standard deviation of all the subpopulations is denoted by $\sigma_{Y|X_1,\ldots,X_k}$ but for simplicity of notation we simply write σ when there is no possibility of confusion.

(Population) Assumption 3 Each subpopulation of Y values, determined by specified values of the predictor variables $X_1, \ldots, X_k$ is Gaussian.

(Sample) Assumption 4 The sample (of size n) is selected either by simple random sampling or by sampling with preselected values of $X_1, \ldots, X_k$.

(Sample) Assumption 5 All sample values $y_i, x_{i,1}, \ldots, x_{i,k}$ for $i = 1, \ldots, n$ are observed without error.



Assumptions (A) for Multiple Linear Regression

Notation The $(k + 1)$-variable population $\{(Y, X_1, \ldots, X_k)\}$ is the study population.

(Population) Assumption 1 The mean $\mu_Y(x_1, \ldots, x_k)$ of the subpopulation of Y values with $X_1 = x_1, \ldots, X_k = x_k$ is

$$\mu_Y(x_1, \ldots, x_k) = \beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k$$

where $\beta_0, \beta_1, \ldots, \beta_k$ are unknown parameters and $x_1, \ldots, x_k$ belong to the set of allowable values (sometimes called the *domain*) of the predictor variables.

(Population) Assumption 2 The standard deviation of the Y values in the subpopulations with $X_1 = x_1, \ldots, X_k = x_k$ does not depend on the values $x_1, \ldots, x_k$ (i.e., the standard deviations are the same for each subpopulation determined by specified values of the predictor variables $X_1, \ldots, X_k$). This common standard deviation of all the subpopulations is denoted by $\sigma_{Y|X_1,\ldots,X_k}$. When there is no possibility of confusion, we use the simpler notation σ instead of the more complete notation $\sigma_{Y|X_1,\ldots,X_k}$.

(Population) Assumption 3 Each subpopulation of Y values, determined by specified values of $X_1, \ldots, X_k$ is a Gaussian population.

(Sample) Assumption 4 The sample data are obtained by simple random sampling or by sampling with preselected values of $X_1, \ldots, X_k$, discussed in Section 2.3. The number of items in the sample is n.

(Sample) Assumption 5 All sample values $y_i, x_{i1}, \ldots, x_{ik}$ for $i = 1, \ldots, n$ are observed without error (but read Section 3.10).

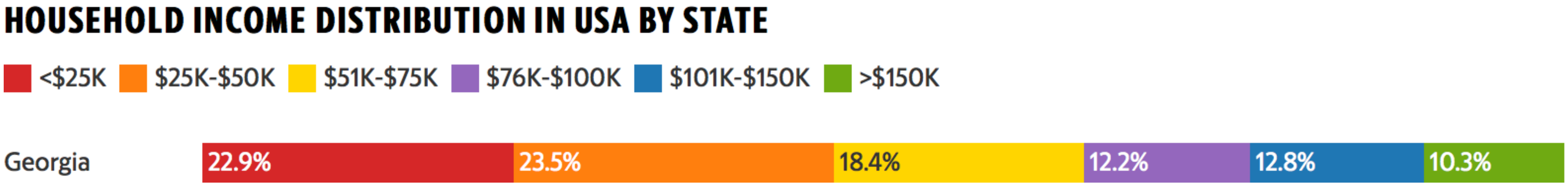
Graybill and Iyer, *Regression Analysis: Concepts and Applications* (1994)
https://www.stat.colostate.edu/regression_book/

Traditional statistical analyses of declarative data

Their limitations include...

Arbitrary assumptions implicit in the definition of categories

- How would *you* define each of these sets?
  - Wealthy U.S. citizens
  - People of color
  - People whose genders are nonbinary
  - Academically successful students
  - Young people
  - Healthy adults
  - Likely voters
  - U.S. residents



J. Desjardins, "Visualizing Household Income Distribution in the U.S. by State" (Nov. 6, 2017)
https://www.visualcapitalist.com/household-income-distribution-u-s-state/

Traditional statistical analyses of declarative data

Their limitations include...

Reduction of mixed and nonconforming individuals to “noise”

- Some techniques that obscure diversity:
  - Lump all individuals who fall into more than one “primary” category into a catch-all category

    ontological category error
  - Segregate individuals who do not conform to traditional definitions into relatively tiny subsets of the sample

    third-gendering
  - Sample sizes too small to represent “insignificant” subpopulations

    marginalization
  - Decide beforehand whether the study will be “mainstream” (majoritarian) or “special interest”

    balkanization

Traditional statistical analyses of declarative data

Their limitations include...

Reliance on self-reporting of identity and behavior

“Facebook introduced dozens of options for users to identify their gender today—and although the social media giant said it would not be releasing a comprehensive list, ABC News has found at least 58 so far. [...] The following are the 58 gender options identified by ABC News:”

- Agender
- Androgyne
- Androgynous
- Bigender
- Cis
- Cisgender
- Cis Female
- Cis Male
- Cis Man
- Cis Woman
- Cisgender Female
- Cisgender Male
- Cisgender Man
- Cisgender Woman
- Female to Male
- FTM
- Gender Fluid
- Gender Nonconforming
- Gender Questioning
- Gender Variant
- Genderqueer
- Intersex
- Male to Female
- MTF
- Gender Fluid
- Gender Nonconforming
- Gender Questioning
- Gender Variant
- Genderqueer
- Intersex
- Male to Female
- MTF
- Neither
- Neutrois
- Non-binary
- Other
- Pangender
- Trans
- Trans*
- Trans Female
- Trans* Female
- Trans Male
- Trans* Male
- Trans Man
- Trans* Man
- Trans Person
- Trans* Person
- Trans Woman
- Trans* Woman
- Transfeminine
- Transgender
- Transgender Female
- Transgender Male
- Transgender Man
- Transgender Person
- Transgender Woman
- Transmasculine
- Transsexual
- Transsexual Female
- Transsexual Male
- Transsexual Man
- Transsexual Person
- Transsexual Woman
- Two-Spirit

R. Goldman, “Here's a List of 58 Gender Options for Facebook Users” (Feb. 13, 2014)
https://www.visualcapitalist.com/household-income-distribution-u-s-state/

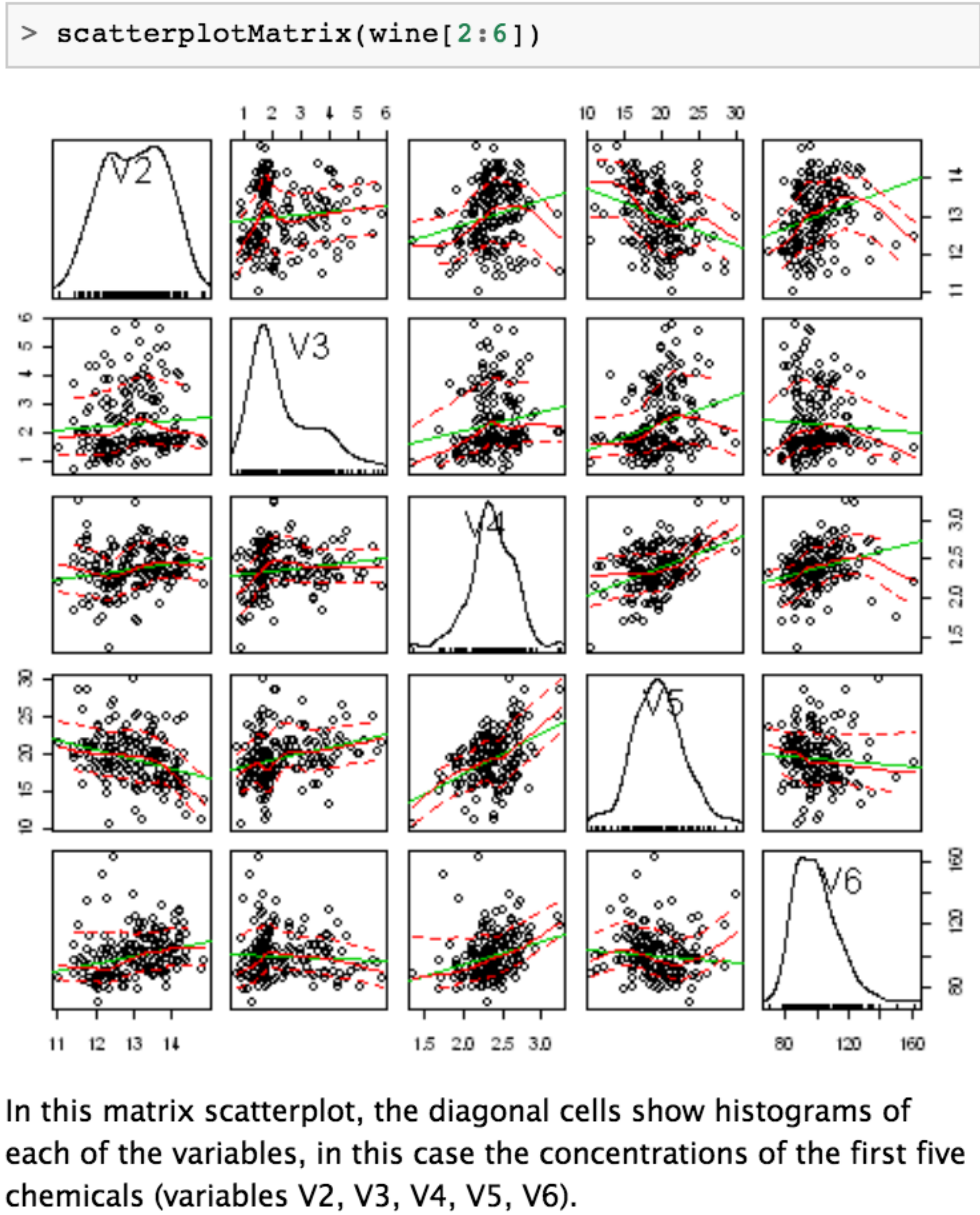
“Now, it appears those options are gone as the company has started rolling out a new version of its custom gender field. This version still populates suggestions as you type, but also allows users to type in any word they wish to represent themselves with across Facebook. [...] It works the same way as the "infinite" gender options Google+ rolled out back in December [...]”

S. O’Kane, “Facebook stops defining gender for its users” (Feb. 26, 2015)
https://www.visualcapitalist.com/household-income-distribution-u-s-state/

Traditional statistical analyses of declarative data

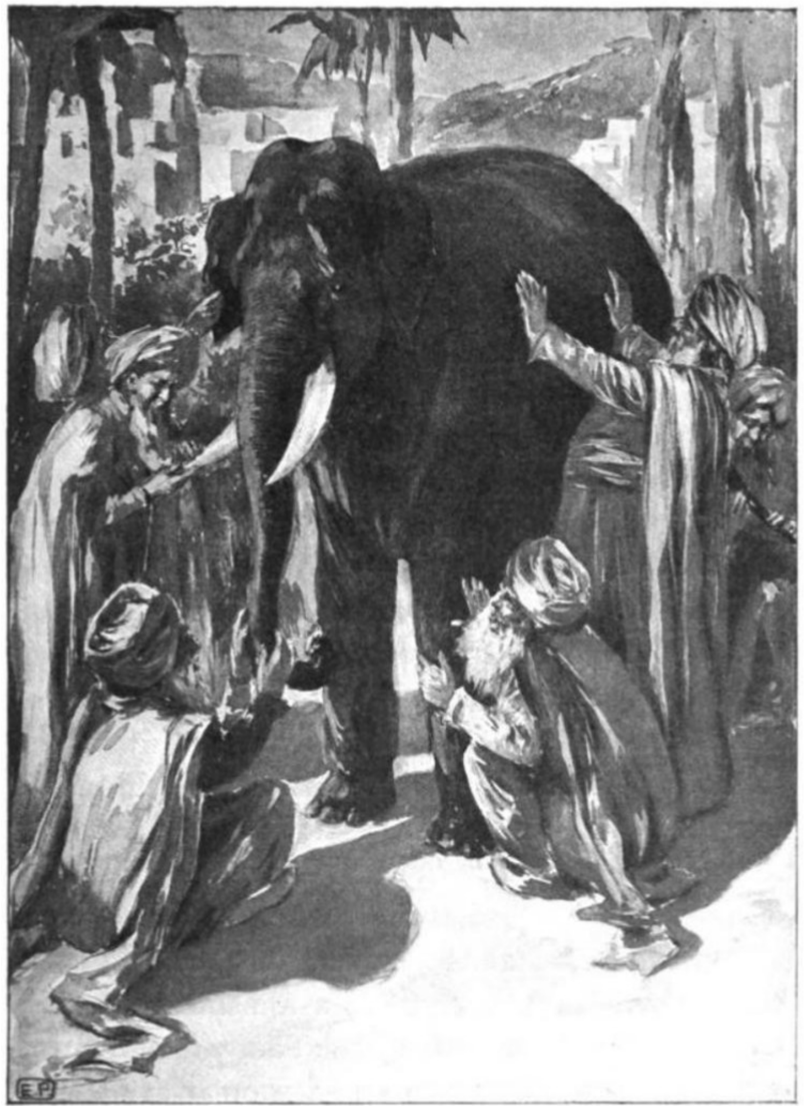
Their limitations include...

Restriction of attention to one or two variables at a time



```
> scatterplotMatrix(wine[2:6])
```

In this matrix scatterplot, the diagonal cells show histograms of each of the variables, in this case the concentrations of the first five chemicals (variables V2, V3, V4, V5, V6).

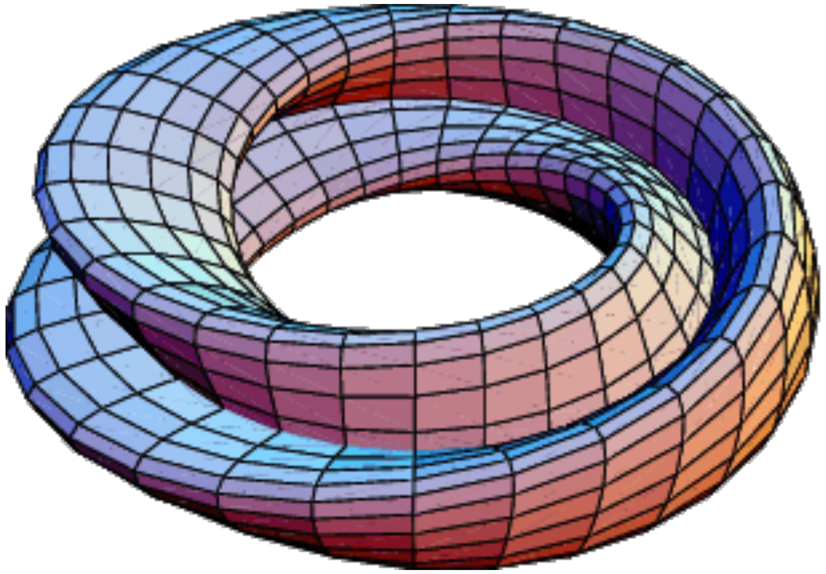


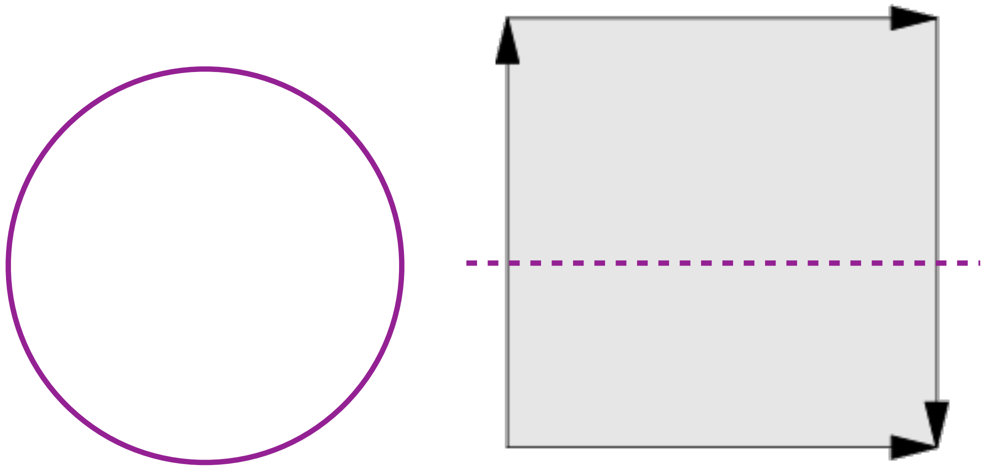
Scatterplots and text: "Using R for Multivariate Analysis"
https://little-book-of-r-for-multivariate-analysis.readthedocs.io/en/latest/src/multivariateanalysis.html

Traditional statistical analyses of declarative data

Their limitations include...

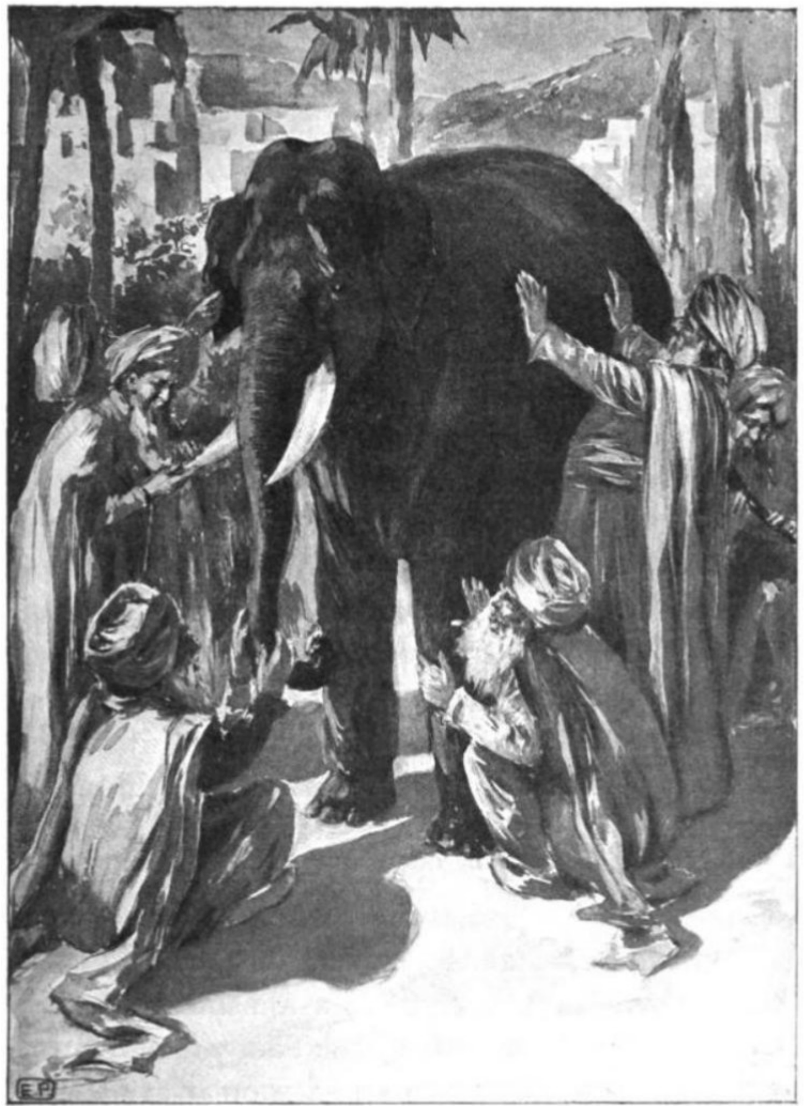
Restriction of attention to one or two variables at a time





Klein bottle

$$\text{level set} = \frac{[0,1]}{\{0,1\}} = S^1$$



Klein bottle rendering and gluing diagram:
https://mathworld.wolfram.com/KleinBottle.html

Traditional statistical analyses of declarative data

Their limitations include...

Dependence on a predetermined mathematical model

Exercise.
Try to write a mathematical equation that models the relationship between...

(a) ...daily nutritional intake, income level, and proximity to a supermarket.
(b) ...femininity and aptitude for leadership.
(c) ...masculinity and willingness to ask for help.
(d) ...nonbinary gender and marital status.
(e) ...race and academic success.

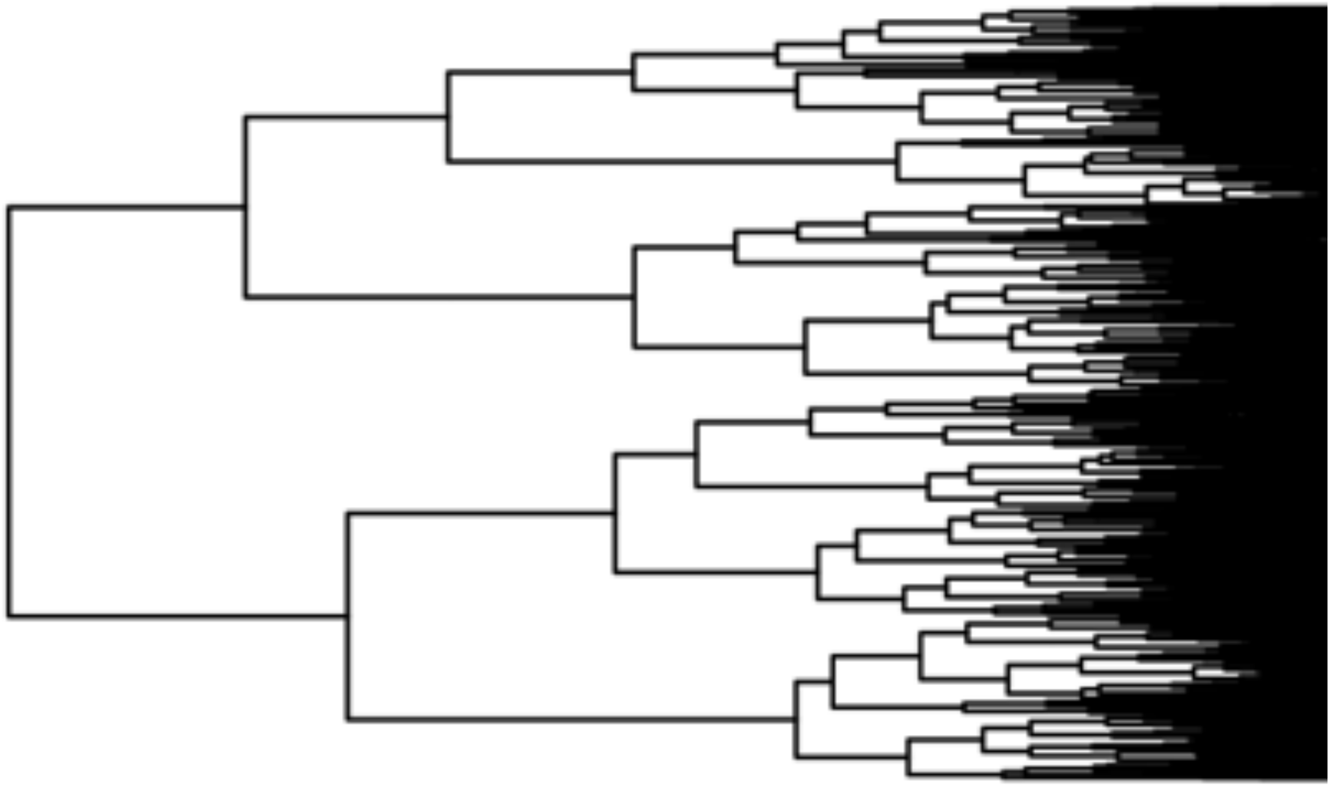
Once you're done laughing at how pointless this exercise is...

- Identify which of the above relationships you've heard discussed at a college-wide meeting or academic conference.
- Explain why or why not in each case.

Hierarchical agglomerative clustering

Observing diversity without categorical assumptions

Clustering analysis subdivides data into subsets (**clusters**) of similar responses.

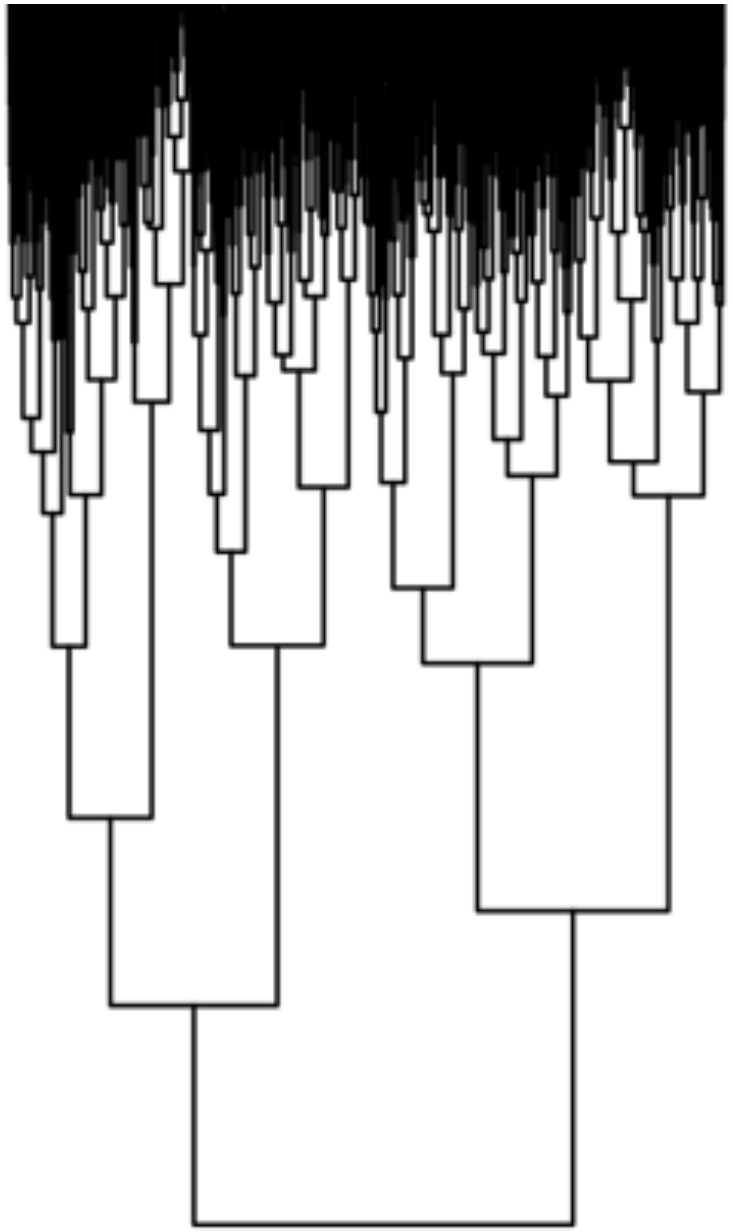


Dendrogram for a sample of over 200,000 test results.
Each test result was represented by a 19-tuple of numbers ranging from 0% to 100%.

Hierarchical agglomerative clustering

How it works

- A distance function is specified
- Each response is agglomerated into a cluster with the nearest response(s), where “nearness” can be measured in terms of any number of variables
- Clusters are agglomerated by progressively relaxing the nearness threshold
- The maximal cluster is the set of all responses



Distance function: L_1 metric on ▯^{19}

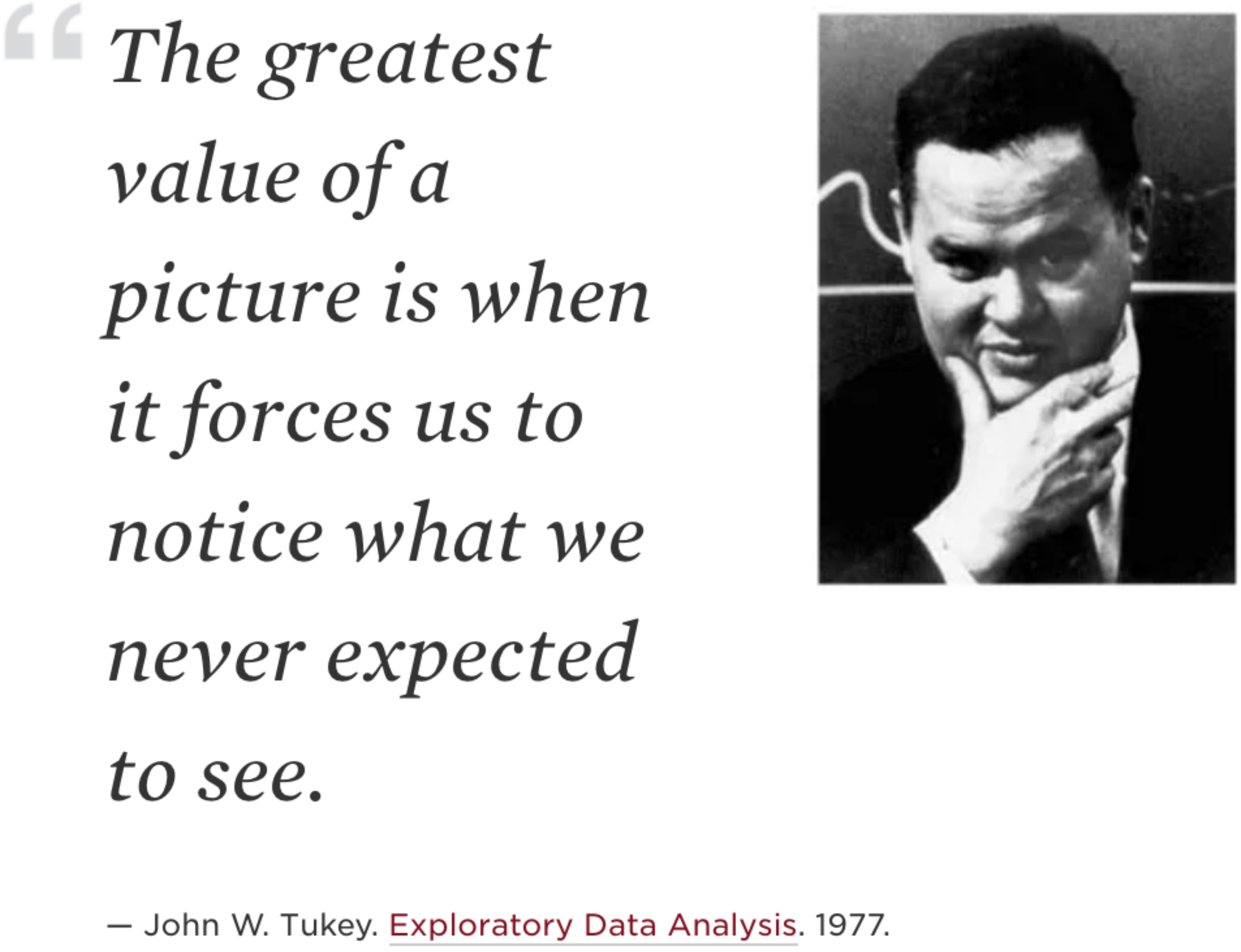
Hierarchical agglomerative clustering

How it differs from classical parametric statistical analysis

- It's nonparametric
- It's exploratory, not confirmatory
- Its output is a partition of the data, not an estimate of a population parameter
- It works well with very large datasets
  - Whereas *p*-values don't.

Hierarchical agglomerative clustering

How it differs from classical parametric statistical analysis



> “The greatest value of a picture is when it forces us to notice what we never expected to see.

— John W. Tukey. Exploratory Data Analysis. 1977.

Hierarchical agglomerative clustering

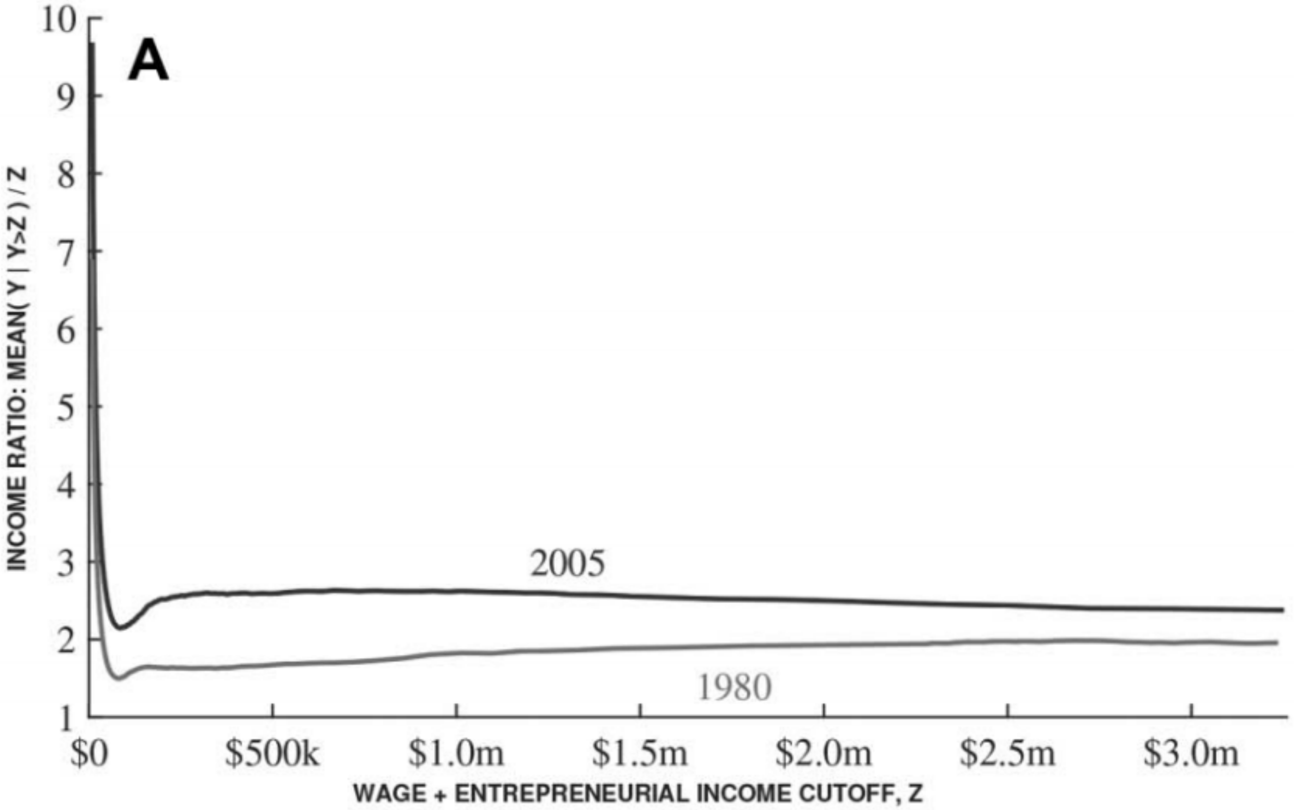
How it differs from classical parametric statistical analysis

- It's nonparametric
- It's exploratory, not confirmatory
- Its output is a partition of the data, not an estimate of a population parameter
- It works well with very large datasets
    - Whereas p-values don't.

The rich get richer...

‘It is the purpose of this paper to analyse a class of distribution functions that appears in a wide range of empirical data—particularly data describing sociological, biological and economic phenomena. [...] The empirical distributions to which we shall refer specifically are (A) distributions of words in prose samples by their frequency of occurrence, (B) distributions of scientists by number of papers published, (C) distributions of cities by population, (D) distributions of income by size, and (E) distributions of biological genera by number of species.’

H. A. Simon, “On a class of skew distribution functions” (1955)
https://academic.oup.com/biomet/article-abstract/42/3-4/425/296312



C. I. Jones and J. Kim, "A Schumpeterian Model of Top Income Inequality", *J. Political Economy*, 126(5) (2018)
https://www.nber.org/papers/w20637

Hierarchical agglomerative clustering

How it differs from other types of statistical analysis

- It's nonparametric
- It's exploratory, not confirmatory
- Its output is a partition of the data, not an estimate of a population parameter
- It works well with large datasets
  - Whereas *p*-values don't.

The PEAR lab

Graph showing psychokinesis scores lying well outside the mean

‘The most common type of statistical testing—the one adopted by the PEAR group—relies on a quantity called a “*p*-value,” which is an indication of “significance.” [...] The smaller the *p*-value, the less likely the data are to be compatible with the chance hypothesis. [...] The problem is that *p*-values have a well known and somewhat fatal problem [...] if the sample size—the number of trials, in this case—is very, very large (and in the PEAR case, it was huge), one is guaranteed to find artificially low *p*-values, indicating a statistically significant result, even though there is nothing going on other than small biases in the experimental apparatus.’

M. Pigliucci, *Nonsense on Stilts: How to Tell Science from Bunk* (2010)
https://books.google.com/books?id=aC8Baky2qTcC&pg=PA79

‘Nearly a hundred volunteers have conducted 212 million REG trials during the 15 years of the lab's existence, and the research shows a tiny but statistically significant result that is not attributable to chance.’

R. Van Bakel, "Mind over Matter" (Apr. 1, 1995)
https://www.wired.com/1995/04/pear/

Hierarchical agglomerative clustering

What's it good for?

- It avoids a priori characterizations of subpopulations that traditional methods erase and/or marginalize
  - Combinations of learning styles
  - Mixed race
  - Intersex, transgender, nonbinary
  - Bisexual, pansexual, asexual, demisexual
- It enables us to formulate hypotheses for confirmatory testing based on observed data, not a priori assumptions
  - *Traditional*: How do the specified individuals diverge from the population as a whole?
  - *Clustering*: What internally coherent groups do we find within the population?

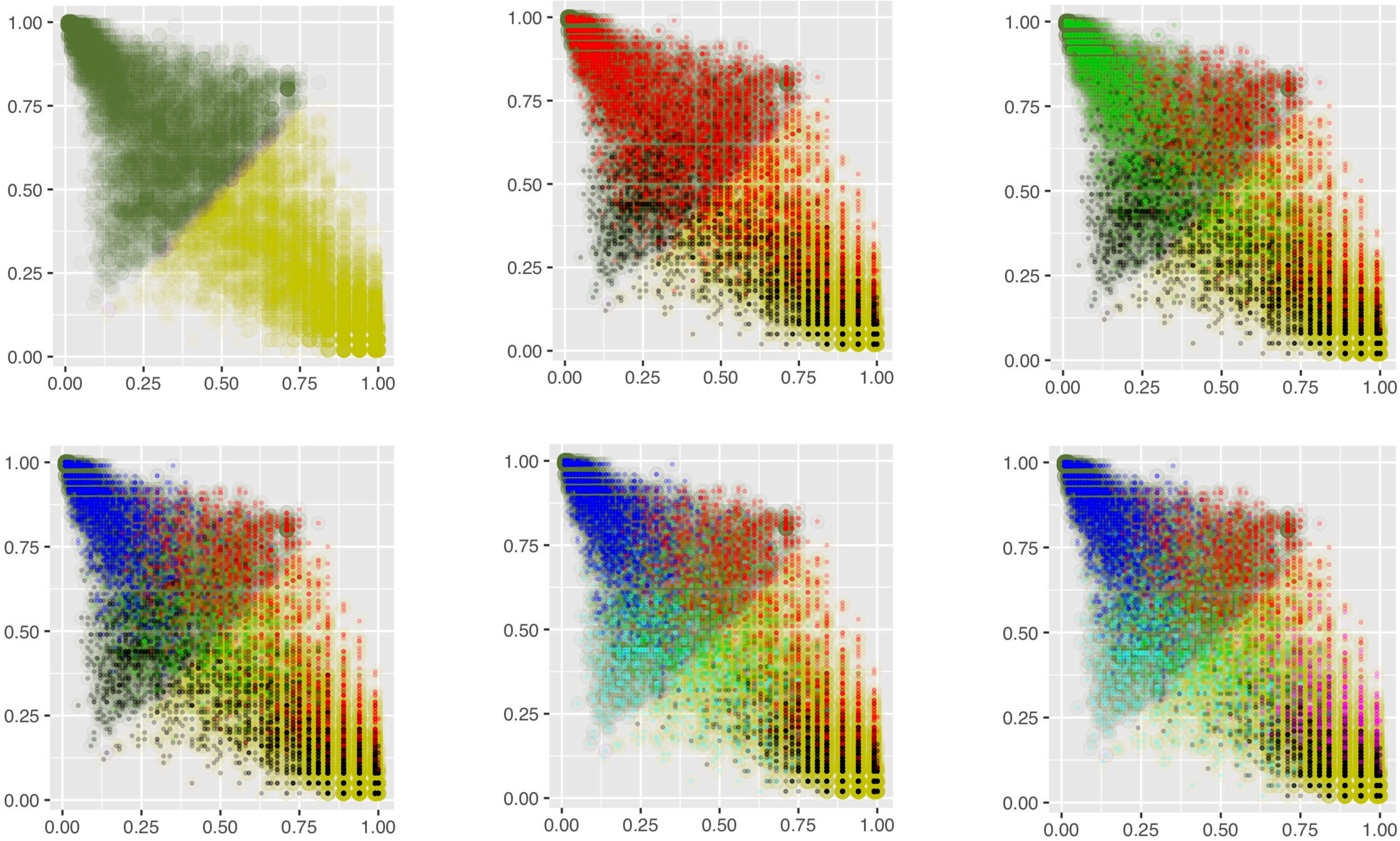
Hierarchical agglomerative clustering

What are some of its limitations?

- The optimal number of clusters is determined heuristically, not algorithmically
- It's not obvious how to quantify how likely it is that sample proportions match population proportions
  - But clustering analysis is an exploratory, not a confirmatory method.
  - If you want p-values and confidence intervals, formulate a hypothesis.
- It doesn't allow for overlapping clusters
  - But it *can* account for overlapping categories, if the *strength* of the respondent's identification with each category is separately solicited.
- It doesn't make predictions or provide explanations

Hierarchical agglomerative clustering

How many clusters is “optimal”?



Hierarchical agglomerative clustering

What are some of its limitations?

- The optimal number of clusters is determined heuristically, not algorithmically

- It's not obvious how to quantify how likely it is that sample proportions match population proportions
  - But clustering analysis is an exploratory, not a confirmatory method.
  - If you want *p*-values and confidence intervals, formulate a hypothesis.

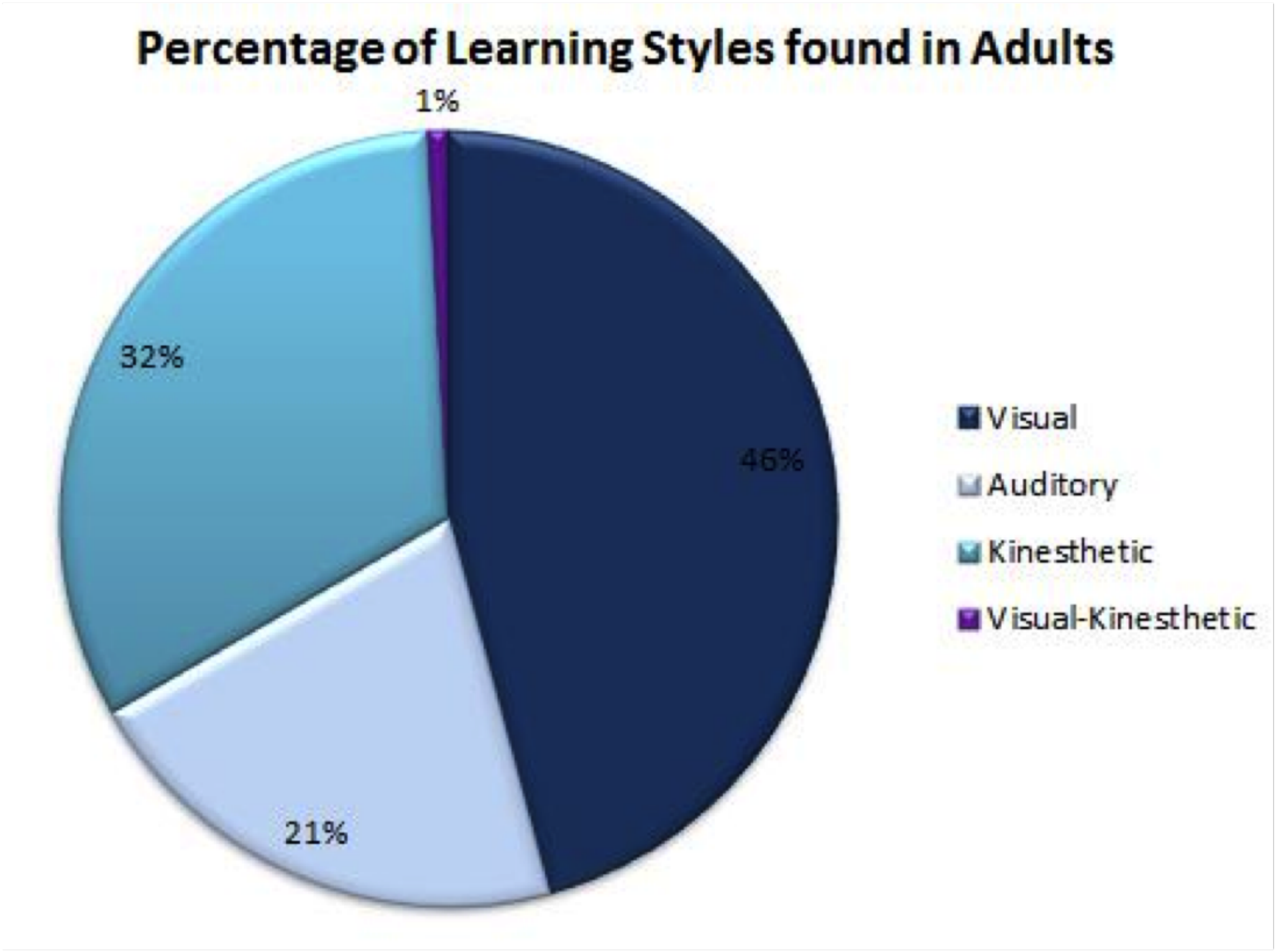
- It doesn't allow for overlapping clusters
  - But it *can* account for overlapping *categories*, if the strength of the respondent's identification with each category is separately solicited.

- It doesn't make predictions or provide explanations

Hierarchical agglomerative clustering

Do we really *want* to predict or explain?

- Suppose we set out to determine which combinations of teaching methods will engage a significant number of students.
    - A priori, we can *explain* a student's affinity for a specified teaching method based on their preferred learning style.
    - We do not seek to *predict* a student's affinities based on personal information.



http://hfuart.blogspot.com/2014/07/percentage-breakdown-of-learning-styles.html

Comparison of clustering methods

There are many to choose from...

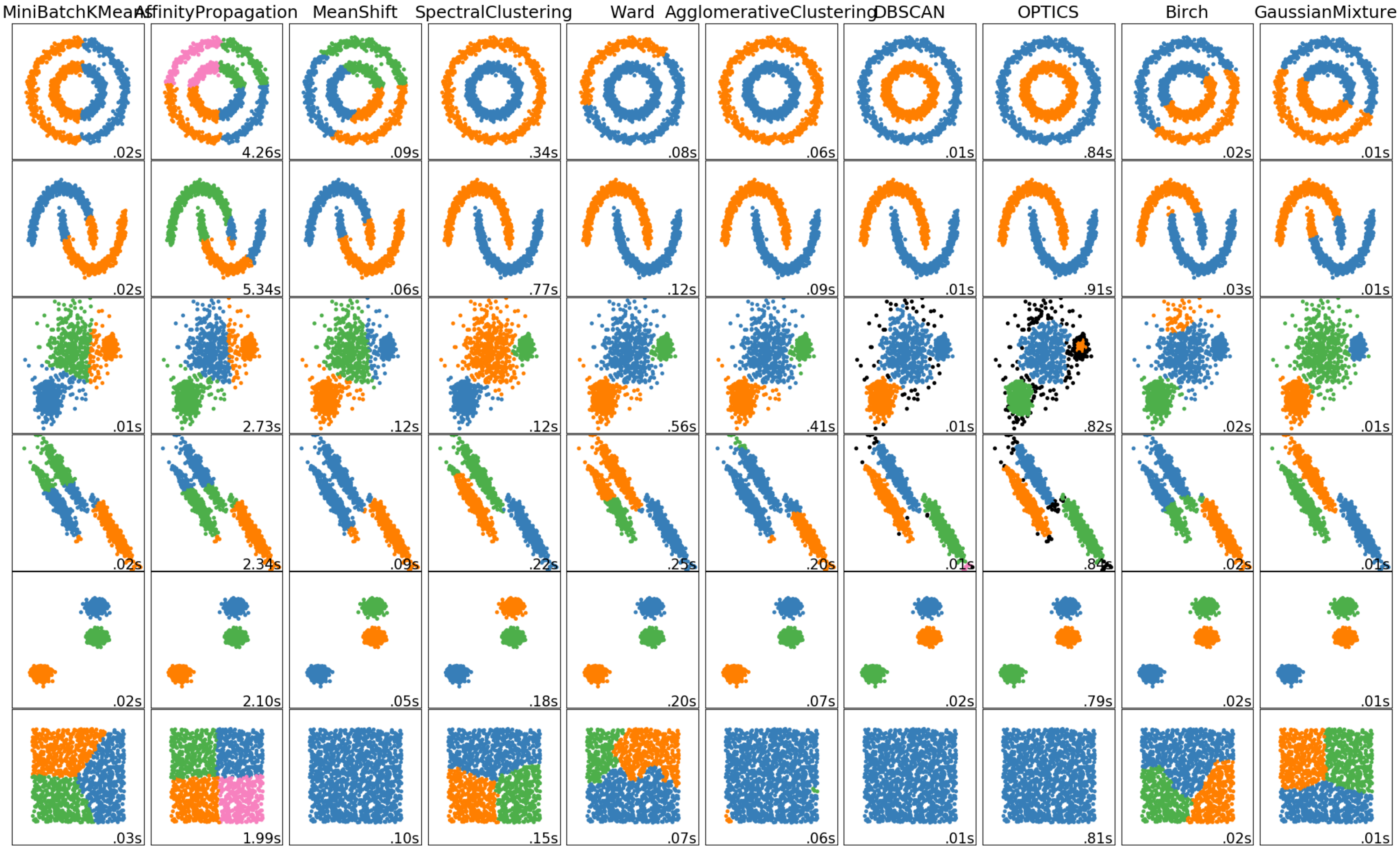
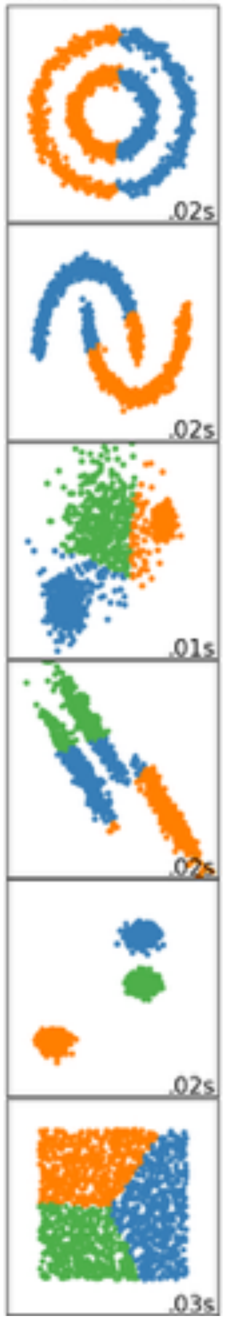


Image from Scikit Learn
https://scikit-learn.org/stable/modules/clustering.html

Comparison of clustering methods

k-means clustering



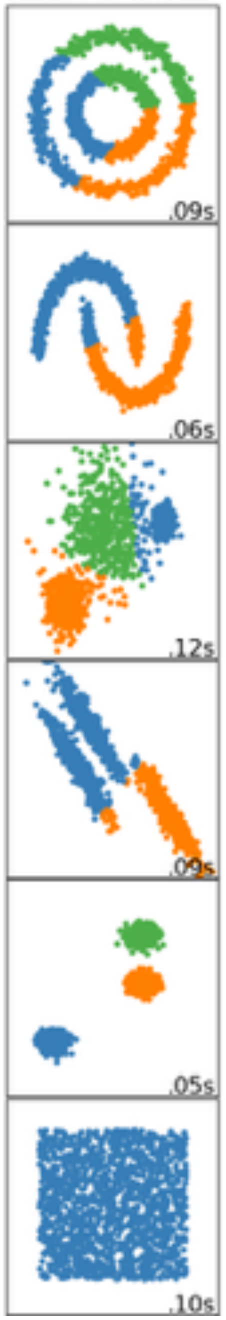
- **Pros**
  - Easy to understand
  - Available in MATLAB, Mathematica, Python, R, Stata, SPSS
  - Fast: $\mathcal{O}(n)$
- **Cons**
  - Number of clusters must be predetermined
  - No support for categorical variables or nonlinear data
  - Performs poorly for non-spherical clusters
  - Dependent on randomly chosen seeds of each cluster
  - Sensitive to noise, outliers, near-duplicates, and choice of metric
  - Not invariant under nonlinear transformations
  - Tends not to recover hierarchical structures in underlying data

A. Singh, et al., "K-means with Three different Distance Metrics," *Int. J. Comput. Appl.* (2013)
https://pdfs.semanticscholar.org/a630/316f9c98839098747007753a9bb6d05f752e.pdf

G. Seif, "The 5 Clustering Algorithms Data Scientists Need to Know," *Medium* (Feb. 5, 2018)
https://towardsdatascience.com/the-5-clustering-algorithms-data-scientists-need-to-know-a36d136ef68

Comparison of clustering methods

Mean-shift clustering



- **Pros**
  - Number of clusters algorithmically determined
  - Clusters need not have any predefined shape
  - Clusters converge toward points of maximum density
  - Discards near-duplicates
  - No proof of convergence for high-dimensional spaces
  - Available in Mathematica, Python, R
- **Cons**
  - Sensitive to choice of kernel and window size ε
  - May cluster disparate outliers together
  - Tends not to recover hierarchical structures in underlying data

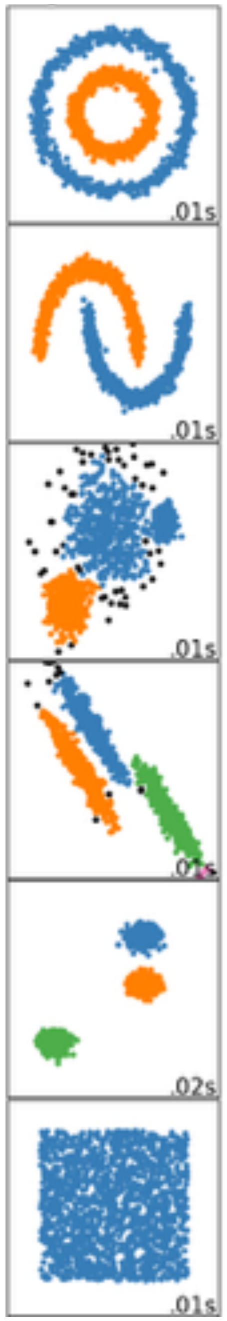
Y. Ghassabeh, "A sufficient condition for the convergence of the mean shift algorithm with Gaussian kernel," *J. Multivar. Anal.*, 67(10) (2013)
https://www.sciencedirect.com/science/article/pii/S0047259X14002644

G. Seif, "The 5 Clustering Algorithms Data Scientists Need to Know," *Medium* (Feb. 5, 2018)
https://towardsdatascience.com/the-5-clustering-algorithms-data-scientists-need-to-know-a36d136ef68

Comparison of clustering methods

DBSCAN

(Density-Based Spatial Clustering of Applications with Noise)



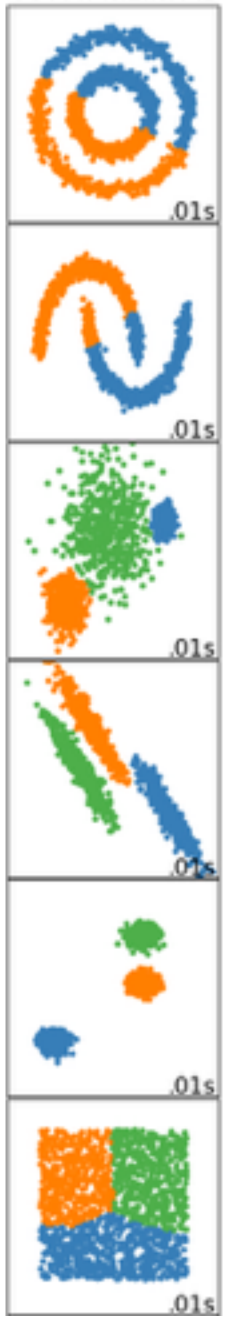
- **Pros**
  - Number of clusters algorithmically determined
  - Finds clusters of any shape or size
  - Not sensitive to outliers
  - Available in Mathematica, Python
  - Not too slow: $\mathcal{O}(n^2)$
- **Cons**
  - Performs poorly with clusters of variable density
  - Sensitive to choice of metric and neighborhood radius ε
  - Excludes outliers from all clusters
  - Tends not to recover hierarchical structures in underlying data

J. Jang and H. Jiang, "DBSCAN++: Towards fast and scalable density clustering," *ArXiv* (2019)
https://arxiv.org/pdf/1810.13105.pdf

G. Seif, "The 5 Clustering Algorithms Data Scientists Need to Know," *Medium* (Feb. 5, 2018)
https://towardsdatascience.com/the-5-clustering-algorithms-data-scientists-need-to-know-a36d136ef68

Comparison of clustering methods

Gaussian Mixture Models (GMMs)

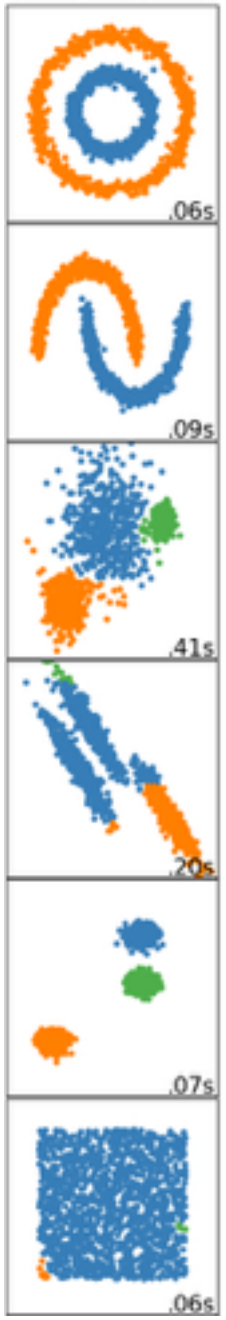


- **Pros**
  - Permits "mixed membership"
  - Performs well with ellipse-shaped clusters
  - Available in MATLAB, Mathematica, Python, R, SPSS
  - Fast: $\mathcal{O}(n)$
- **Cons**
  - Assumes a Gaussian distribution in each cluster
  - Tends not to recover hierarchical structures in underlying data

G. Seif, "The 5 Clustering Algorithms Data Scientists Need to Know," *Medium* (Feb. 5, 2018)
https://towardsdatascience.com/the-5-clustering-algorithms-data-scientists-need-to-know-a36d136ef68

Comparison of clustering methods

Single-linkage hierarchical agglomerative clustering



- **Pros**
  - Clusters may have any shape, and variable sizes
  - Robust under different choices of metric
  - Tends to recover hierarchical structures in underlying data
  - Can handle text and non-numeric data
  - Available in Mathematica, MATLAB, Python, R, SPSS, Stata
  - Not too slow: $\mathcal{O}(n^2)$

- **Cons**
  - Requires visual inspection to determine optimal number of clusters
  - Single-linkage may yield long, stringy clusters
  - Large memory overhead

"Clustering," *SciKit* (BSD license)
https://scikit-learn.org/stable/modules/clustering.html

R. Sibson, "SLINK: An optimally efficient algorithm for the single-link cluster method," *Comput. J.* 16(1) (1972)
https://grid.cs.gsu.edu/~wkim/index_files/papers/sibson.pdf

G. Seif, "The 5 Clustering Algorithms Data Scientists Need to Know," *Medium* (Feb. 5, 2018)
https://towardsdatascience.com/the-5-clustering-algorithms-data-scientists-need-to-know-a36d136ef68

Outline of a study based on clustering analysis

Steps

- **Disaggregation**
  - Divide the sample into clusters of “similar” individuals while minimizing a priori assumptions.
- **Characterization of clusters**
  - What proportion of the sample does each cluster comprise, and what properties appear to be shared (or *not* shared) within each subset?
- **Formulation of a model**
  - Formulate a model that can be clearly articulated and meaningfully tested.

An unruly dataset

Mathematical properties of the KRA dataset

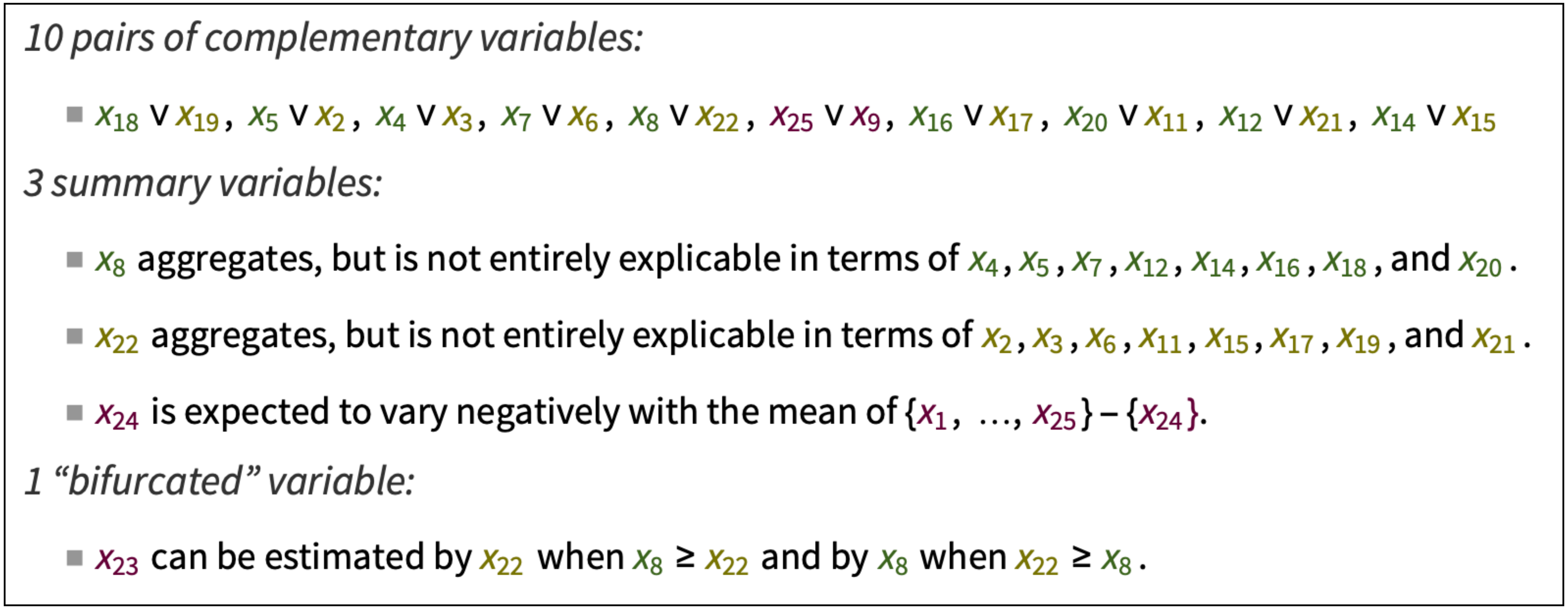
The KRA dataset (La Corte, 2020) was collected from publicly available web survey responses.

- **Large, high-dimensional, and bounded**
  - $n = 236{,}353$ survey results
  - Each survey result consists of 25 scores in the range 0% to 100%
- **Non-normally distributed**
  - All but one of the univariate distributions is flat-tailed
  - No bivariate distribution was found suitable by R for linear regression
- **Interdependent variables**
  - Strong bivariate and multivariate correlations were expected

An unruly dataset

Interdependencies

- **Interdependent variables**
  - Strong bivariate and multivariate correlations were expected



10 pairs of complementary variables:

- $x_{18} \vee x_{19}$, $x_5 \vee x_2$, $x_4 \vee x_3$, $x_7 \vee x_6$, $x_8 \vee x_{22}$, $x_{25} \vee x_9$, $x_{16} \vee x_{17}$, $x_{20} \vee x_{11}$, $x_{12} \vee x_{21}$, $x_{14} \vee x_{15}$

3 summary variables:

- x_8 aggregates, but is not entirely explicable in terms of x_4, x_5, x_7, x_{12}, x_{14}, x_{16}, x_{18}, and x_{20}.
- x_{22} aggregates, but is not entirely explicable in terms of x_2, x_3, x_6, x_{11}, x_{15}, x_{17}, x_{19}, and x_{21}.
- x_{24} is expected to vary negatively with the mean of $\{x_1, \ldots, x_{25}\} - \{x_{24}\}$.

1 "bifurcated" variable:

- x_{23} can be estimated by x_{22} when $x_8 \geq x_{22}$ and by x_8 when $x_{22} \geq x_8$.

An unruly dataset

Most expected interdependencies appeared in the clustering

Hierarchical clustering of variables:

- $x_8, x_4, x_5, x_7, x_{12}, x_{14}, x_{16}, x_{18}, x_{20}, x_{23}$
- $x_{22}, x_2, x_3, x_6, x_{11}, x_{15}, x_{17}, x_{19}, x_{21}$
- $x_1, x_9, x_{10}, x_{13}, x_{25}$
- x_{24}

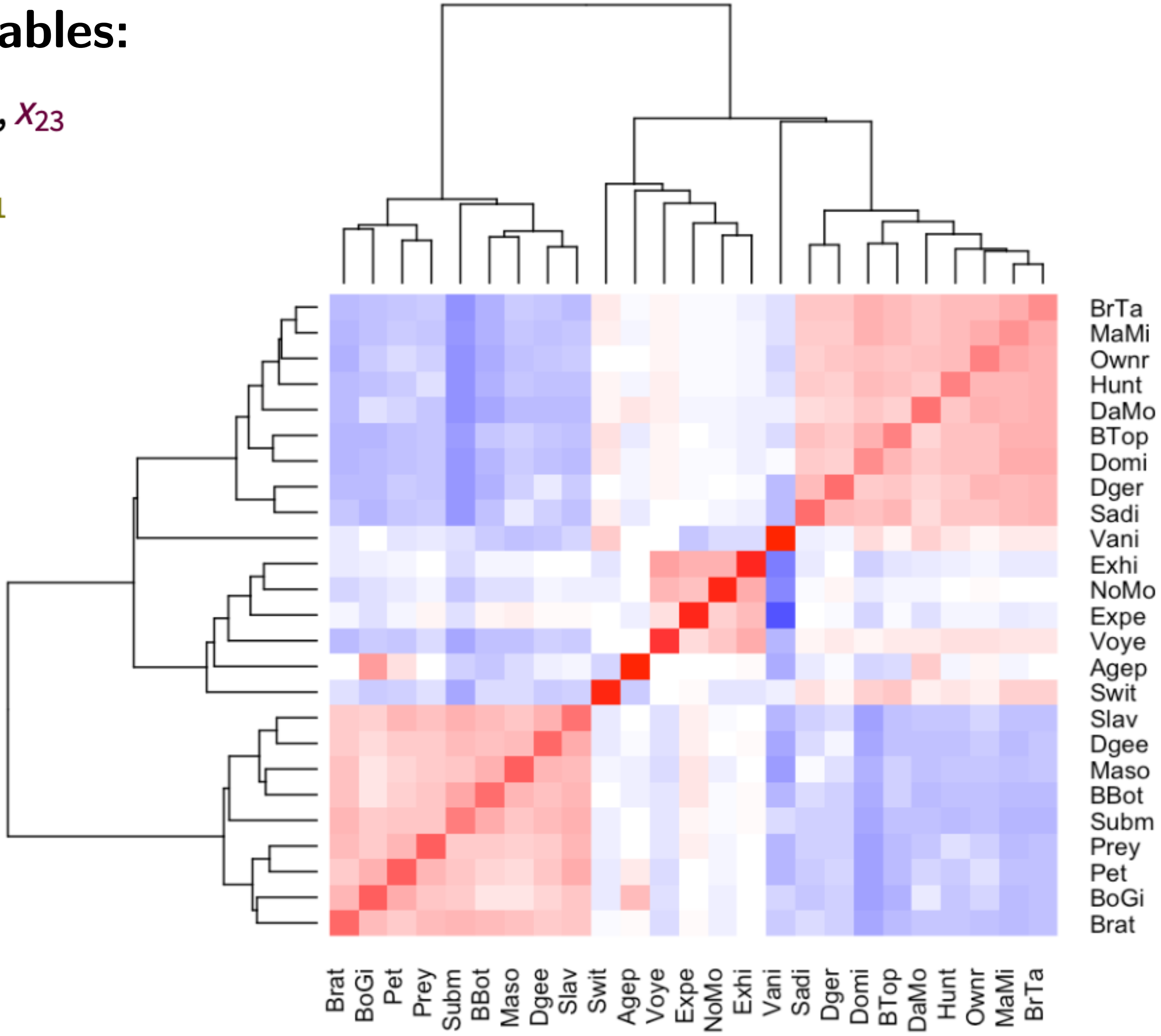
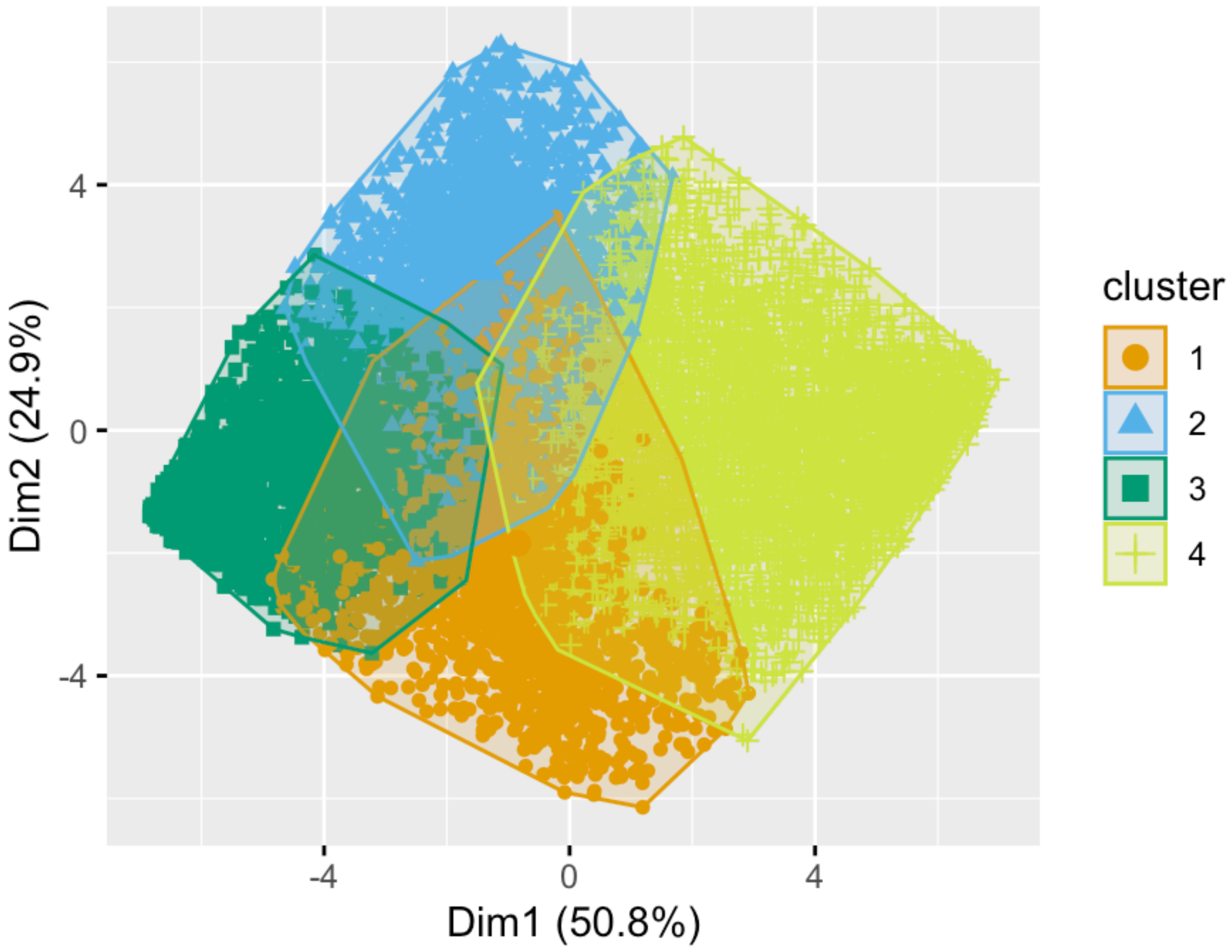


Fig.: Dendrogram for the 25 variables, with a heat map indicating strength and direction of pairwise correlations.

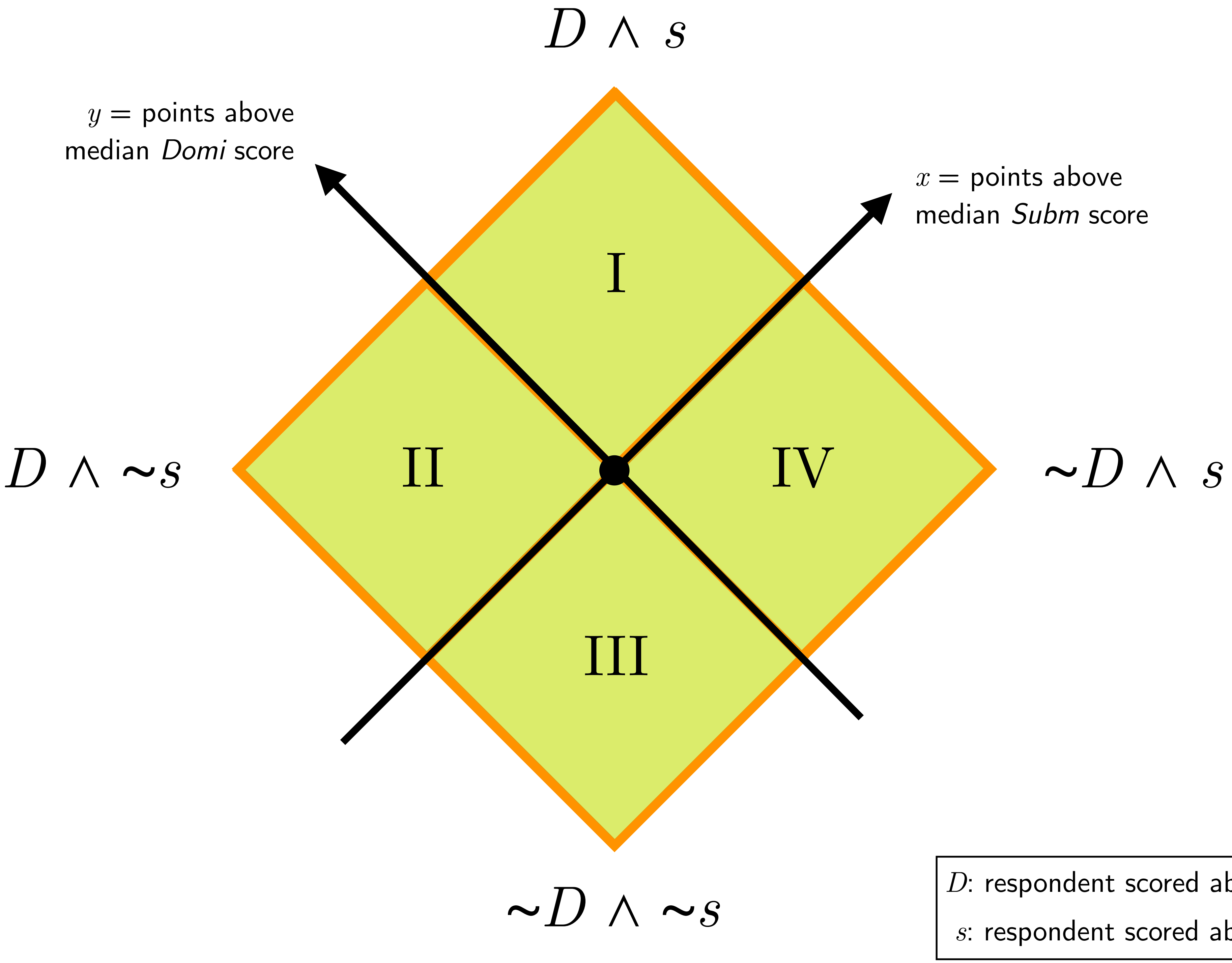
An unruly dataset

Cluster plot for survey responses



An unruly dataset

Formal model for “types” of individual



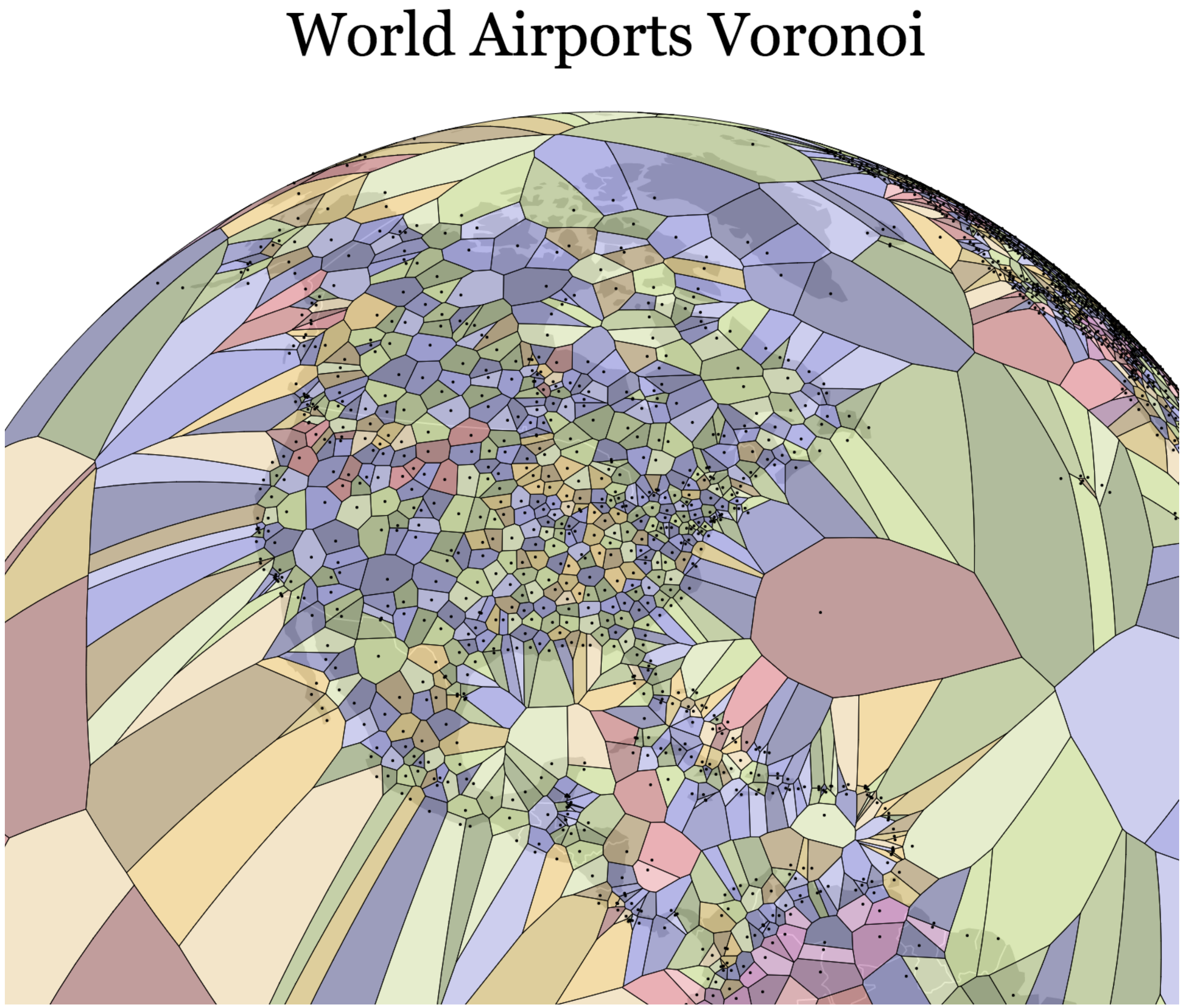
D: respondent scored above median in *Domi*

s: respondent scored above median in *Subm*

Visualizing diversity in a dataset

Some visualization techniques

- **Graphs**
  - Dendrograms
- **Charts**
  - Rank-rank summary plots
  - Proportional Venn diagrams
- **Geometric models**
  - Voronoi diagrams
- **Topological models**
  - Vietoris-Rips complexes
  - Persistence homology



https://www.jasondavies.com/maps/voronoi/airports/

Linked from American Geographical Society:
https://ubique.americangeo.org/map-of-the-week/map-of-the-week-voronoi-diagrams-in-geography/