The Surprising Power of Extreme Case Sampling:

Qualitative Analysis of Extreme Case Samples as an Alternative to Data Mining Methods like Random Forests for Exploratory Research

David Cochrane
Cosmic Patterns Software
1. The first step in research is to have a research question.

2. The second step is to select a research procedure.

3. In the social sciences a qualitative study of extreme cases is often the best research method, but it is often overlooked.
A research question:

What helps students to excel academically and as human beings?

An excellent research method:

Interview winners of the national spelling bee who also have a positive outlook or who are now positive contributing members of society.
Rule #1 for extreme case sampling studies:

We wish to know if there is a clear relationship of variables.

(a) The ultimate goal is inferential statistics, not descriptive statistics, and (b) exploratory research to help identify potential important predictor variables would be helpful.

Example: Ethnographic studies that provide descriptive information without interested in inferences for possible predictors of behavior are not likely to benefit from qualitative analysis of extreme cases.
Rule #2 for extreme case sampling studies (helpful but not required):

The variable of interest can be operationally defined and other variables may be unknown or unclear.

Qualitative studies of extreme cases are usually a form of exploratory research. The variable of interest may be a predictor variable (e.g. length of school day) or an outcome variable (e.g. winner of spelling bee).
Note: Qualitative studies of extreme cases are an alternative to quantitative data mining methods.

In my research I found qualitative extreme case studies more helpful as the starting point for exploratory research than Random Forests, Support Vector Machines, and decision trees.
Ruler #3 (Not required, but important to note): Fairly large measurement errors in determining extreme cases typically is not a big problem.

Example: winner of the spelling bee was lucky that the words happened to be ones that he/she knew. The student is still exceptionally good at spelling.
Ruler #4 (Not required, but important to note): If the variable of interest is a latent variable and construct validity is low, extreme case sampling can still provide valuable information!

Example: Even if winners of the spelling bee are not the best educated or most desirable students, we have still learned something about a characteristic related to our interest.
Rule #5: for extreme case sampling studies (helpful but not required):

The relationship among variables may be very complex, and there may be complex interactions among the variables.

As noted earlier, qualitative research of extreme cases is an alternative to more commonly recognized data mining methods.
Rule #6 for extreme case sampling studies (very important!):

You have a clear and objective way to select extreme cases.

Select the extreme cases using objective criteria so there is no selection bias.

Example: winners of a spelling bee.
Rule #7 for extreme case sampling studies:

Extreme case sampling of a sample, which is a kind of “double-sampling” procedure is OK.

Note that if the sample size is large, then biases in the sample are often NOT a big problem!

(Biases in the way you select extreme cases from the sample, however, are a problem).
Rule #8 for extreme case sampling studies:

Select people who are extreme but not separated from other scores by a large gap. Avoid these extreme outliers.

In the next slide I give a visualization of this.
A hypothetical study on self-confidence and physical height for men in the USA:

We study the men who are around 4'10” to 5'1” or around 6'3” to 6'4” tall rather than those who are 3'7” or 6'10” tall.
A term that comes to mind for extreme cases is “outliers”. We are suspicious of outliers.

However, in some research scenarios these extreme cases provide the best answers to our research questions!!!
Rule #9: Be comfortable with guessing! All major discoveries start with guessing!!!

We don't know if these cases are qualitatively different so guess!

Richard Feynman on how to discover a new scientific law:

"Now I am going to discuss how we would look for a new law. In general we look for a new law by the following process. First we guess it, (laughter). But don't laugh, that is really true. And then we compute the consequences of the guess. . . "

https://youtu.be/EYPapE-3FRw
Rule #10 for extreme case sampling studies:

Analyze all extreme cases in the range of scores that you select. For example, you might decide to interview 6 people.

If you skip cases, there is selection bias!!!!!
Rule #11 for extreme case sampling studies:

If you guess that the relationship of continuous variables to the variable of interest is gradually increasing or gradually decreasing, then you can proceed with the extreme case sampling study.

This does not need to be completely true but only mostly true.
Rule #11 restated: The continuous predictor variables have roughly a linear or non-linear continuously increasing or decreasing relationship to the outcome variable.

The data shown above is not ideal but the extremes in the lower left and upper right make this still reasonably good for extreme case sampling.
The data shown above is NOT very good for an extreme case sampling study.

This is the key to the power of extreme case sampling!!!!!
Qualitative variables are nicely detected by extreme case sampling studies.

Note: Interactions of variables are not a big problem for extreme case sampling studies!
Rule #12 for extreme case sampling studies:

The researcher(s) have content expertise.

The importance of this rule can hardly be exaggerated. This is qualitative research and the ability of the researcher to notice and to discover relevant consistent characteristics among the extreme cases is almost purely a function of the researcher's content expertise.
Rules #13 and #14 for extreme case sampling studies:

Extreme case sampling can be more helpful than data mining when
(a) the effect size of predictor variables is small, and/or
(b) the predictor variables might be variables not currently considered important enough to be put in the model.
Rule #13: Effect size is small

We do exploratory research because we do not know what the predictor variables are. If the signal-to-noise ratio is small, i.e., there are predictors but the effect size is small, the signal can get drowned out by a large amount of random variation. Huge sample sizes may be needed to detect the signal (i.e., small effect size). A person with content expertise might notice the relevant predictor variable because the expert has prior knowledge (Bayesian perspective). This prior knowledge can be difficult to articulate and often are hunches that arise when studying a specific case. This is why extreme case sampling worked better for me in early exploratory research than using data mining methods like Random Forests and Support Vector Machines.
Rule #14: Variables not currently considered

While analyzing a chart for the time a criminal committed a crime, I explored a huge number of different possible astrological predictors. I noticed something that I did not even have a name for, which we now refer to as a “mixed midpoint structure”. I then looked at another time of an accident and noticed a powerful mixed midpoint structure. A person with content expertise can build upon knowledge and get creative new insights. These insights lead to the identification of possible new predictor variables that can be put into a data mining analysis or some other statistical analysis or visualization of data. My research would have proceeded much more slowly if I did not spend a tremendous amount of time analyzing extreme cases. This analysis provided insight into new predictor variables that became the basis of the model that was developed.
Rule #15: The minimum sample size that the extreme cases are selected from is about a hundred. A sample size of a few thousand gives stronger inferences. A sample size of tends of thousands is not necessary.

For example: qualitative analysis of the top 5 spellers among 100 students can provide insight into predictor variables for excellent spelling performance. The top 5 spellers among 1,000 students are even more exceptional. For most practical purposes we have useful and actionable information from this study.
A reminder about extreme case sampling:

* On predictor variables (independent variables)
* On outcome variables (dependent variables)

In education:

Predictor variables: curriculum, number of school hours, etc.

Outcome variables: test scores, winners of spelling bee, etc.
3 Concluding comments:

1. Qualitative studies of extreme cases are often overlooked and undervalued as a powerful exploratory research methodology!
3 Concluding comments:

2. In my case, this simple research method proved to be more useful than very sophisticated data mining methods primarily for these 3 inter-related reasons:

(a) The variables found were not included in the data mining studies.

(b) Some of the results are qualitative and tend to get overlooked by data mining methods, and

(c) The variables functioned in specific and complex ways that were easier to detect using content expertise.
3. Concluding comments:

3. The title of this presentation is “The Surprising Power of Extreme Case Sampling”.

“Power” is a term used in quantitative research that means the ability of a statistical test to detect an existing relationship.

Here I use the term in a slightly different way to mean the ability of qualitative research to detect relationships. Although we cannot put a number on it, extreme case sampling often provides an extraordinarily efficient way to detect relationships between variables.