**Why do we use randomization in sample selection?**

Recently, I had a chance to read a legal opinion in a document classification case and a particular comment made by the judge caught my attention. Specifically the Judge commented "The point of random sampling is to eliminate irrelevant documents from the group identified by a computerized search and focus the parties' search on relevant documents only."[[1]](#footnote-1) This quote struck me as a bit odd in that it seemed to miss the whole point of randomization from a statistical perspective or, at very least, was extremely loose with the application of statistical principles to a rather knotty and expensive question. The purpose of this post, thus, is to take a look at the application of randomization in large scale document classification cases and point out why it is so necessary.

As a motivation, say you have a friend that owns a battery company and some of his clients, maybe 1000 clients, are making the crack pot claim that you friend's batteries somehow inexplicitly explode injuring the user. To add insult to injury, the complaining clients contend that your friend and some of his workers actually knew about the problem and talked about it in company emails. The company collected all 3,141,459 emails and put them on a thumb drive. The clients file suit and northern California because they really like the rain in that part of the country and they want to look at all of your friend's email. Your friend wants no part of this knowing that there is a secret Katie Perry cult in the organization and disclosure of this would result in a complete ban of his company from the annual battery convention in Vegas. The judge, let's call her Judge Donna, splits the baby. The judge tells your friend that he has to produce all "relevant" emails, and he asks you to help him sort them out.

1. IMO Lithium Batteries Antitrust Litigation, Dkt: 13-MD-02420, Slip Op. Page 3 (N.D.C. Filed 2/24/15). [↑](#footnote-ref-1)