**Ten Simple Rules for Digital Data Storage:**

**Rule 8 References**

**30** - “Famously, New York City officials shared what they thought was anonymized data on cab drivers and over 173 million cab rides. However, it was quickly recognized that the city anonymized the data with a simple MD5 hashing scheme and all 20 GB of data were de-anonymized in a matter of hours”

<https://arstechnica.com/tech-policy/2014/06/poorly-anonymized-logs-reveal-nyc-cab-drivers-detailed-whereabouts/>

**Title: Poorly anonymised logs reveal NYC cab driver’s detailed whereabouts**

* New York City officials have inadvertently revealed the detailed comings and goings of individual taxi drivers over more than 173 million trips.
* Data released in response to a public records request.
* Drivers’ hack license numbers and medallion numbers were obscured by one-way cryptographic hashes using MD5 algorithm.
* Instead of medallion number 9Y99 or hack number 5296319, they were converted to 71b9c3f3ee5efb81ca05e9b90c91c88f  and 98c2b1aeb8d40ff826c6f1580a600853, respectively.
* One-way hashes can’t be mathematically converted back into their original values.
* Flaw: the medallion and hack license numbers are structured in predictable patterns making it trivial to run all possible iterations through the same MD5 algorithm and compare the output data to data contained in the 20GB file.
* Software developer Vijay Pandurangan cracked it in less than 2 hours, “… simply using hash functions is an ineffective way to anonymise data”, he said.

**31** – “In more problematic cases, the data itself allows identifiability: this is the case with human genomic data that map directly onto a subject’s identity”

https://journals.plos.org/plosgenetics/article?id=10.1371/journal.pgen.1000167

# Title: Resolving Individuals Contributing Trace Amounts of DNA to Highly Complex Mixtures Using High-Density SNP Genotyping Microarrays

**Abstract:** We use high-density single nucleotide polymorphism (SNP) genotyping microarrays to demonstrate the ability to accurately and robustly determine whether individuals are in a complex genomic DNA mixture. We first develop a theoretical framework for detecting an individual's presence within a mixture, then show, through simulations, the limits associated with our method, and finally demonstrate experimentally the identification of the presence of genomic DNA of specific individuals within a series of highly complex genomic mixtures, including mixtures where an individual contributes less than 0.1% of the total genomic DNA. These findings shift the perceived utility of SNPs for identifying individual trace contributors within a forensics mixture, and suggest future research efforts into assessing the viability of previously sub-optimal DNA sources due to sample contamination. These findings also suggest that composite statistics across cohorts, such as allele frequency or genotype counts, do not mask identity within genome-wide association studies. The implications of these findings are discussed.

**32** – “Methods for dealing with these complex issues at the intersection of data storage and privacy are rapidly evolving and include storing changes against a reference genome to help with privacy and reduce overall data volumes”

<https://science.sciencemag.org/content/331/6018/728>

**Title: On the Future of Genomic Data**

* Challenges in genomics derive from the informatics needed to store and analyse the raw sequencing data.
* The need to process terabytes of information has become difficult for many labs.

**33** – “Methods for dealing with these complex issues at the intersection of data storage and privacy are rapidly evolving and include storing changes against a reference genome to help with privacy and reduce overall data volumes”

<https://www.eurekaselect.com/122208/article>

**Title: Trends in Genome Compression**

* Technological advancemens have led to a tremendous increase in the amount of genomic data produced.
* Issues with large amounts of data:
  + Storing.
  + Managing.
  + Analysis.
* One possible remedy is to compress the data.

34 – “Bringing computation to data storage facilities instead of vice versa”

<https://academic.oup.com/ije/article/43/6/1929/707730>

**Title: DataSHIELD: taking the analysis to the data, not the data to the analysis**

* Large sample sizes needed for biomedicine and social science.
* DataSHIELD provides a novel technological solution to facilitate access of resarchers and other healthcare professionals to individual-level data.\