**Assignment - Various Sampling Methods**

Shaun Pritchard

Rasmussen College

QMB4000

Benjamin Tasker

December 15, 2020

**Assignment - Various Sampling Methods**

**Assignment - Various Sampling Methods**

Netflix is one of the biggest data-centric companies on the planet. Netflix uses many different models, data, and approaches to lead innovation and research in data analytics, AI, Machine Learning, and Big data to drive business performance and user engagement.

* Netflix has more than 150 Million active subscribers across the planet.
* Data is at the core of Netflix's business strategies and models.
* Netflix video streaming consumes 37% of downstream bandwidth across the entire internet on the planet.

**Netflix Big Data**

Netflix collects customer interaction, engagements, and response data in correlation to data features it collects from users on each TV shows and movie that it streams. They use data analytics to analyze to better understand users based on *(Netflix Research, 2020)*:

* What each customer watches
* How much time they spend watching
* What content they watch
* Location of users
* What times content is watched
* How many times the content is paused
* Which content is fast forward
* Which content is repeated

They even use CV and machine learning to collect users’ facial expressions and engagements like on each TV show or movie they watch. Simply collect to enhance user experience and provide the best-personalized experience of TV shows and movies to their users.

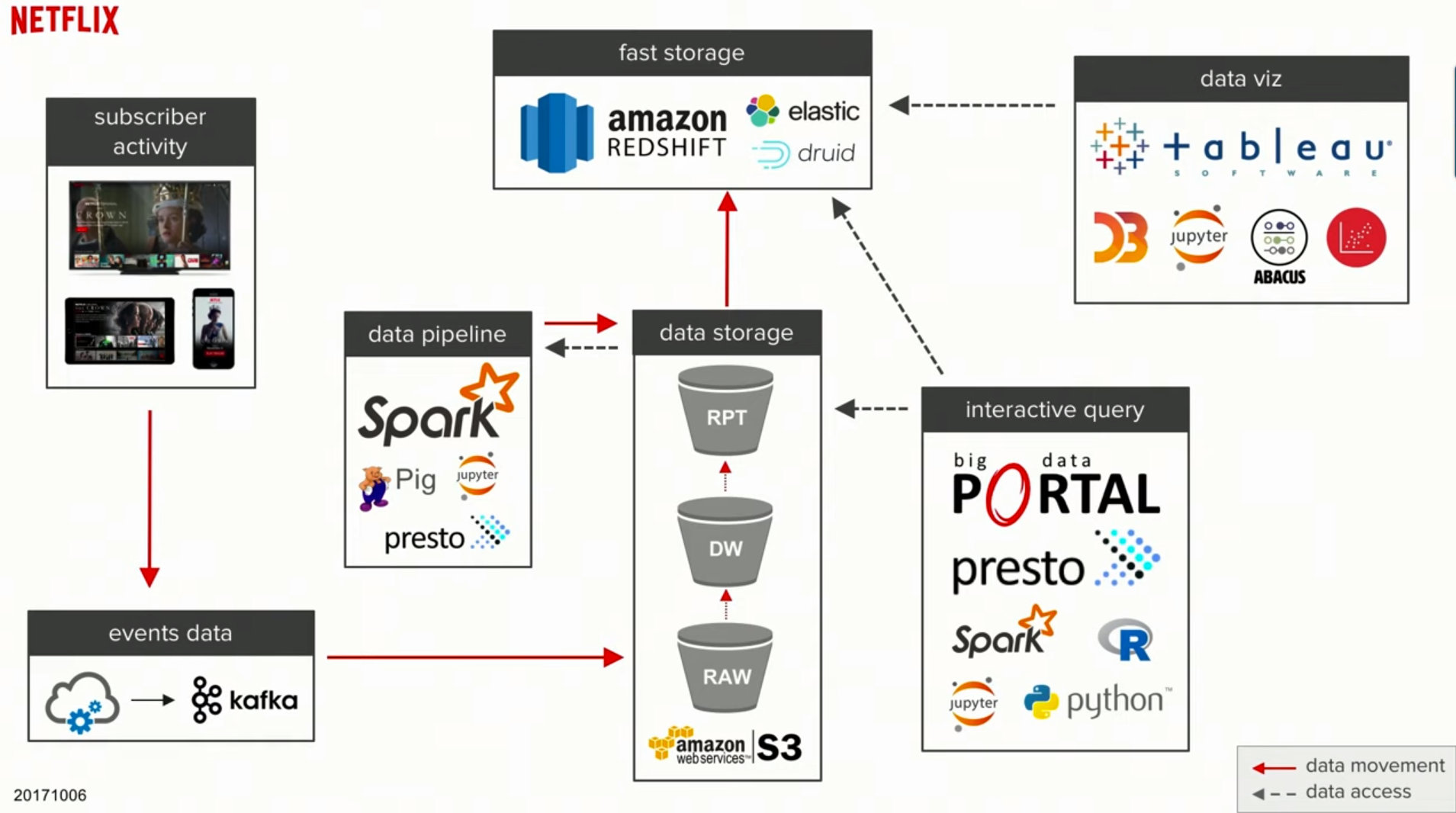
Netflix even has screenshots of scenes people might have viewed repeatedly, the rating content is given, the number of searches, and what is searched for. With this data, Netflix can create a detailed profile of its users. To collect all this data and harness it into meaningful business information. According to Netflix, they earn over a billion dollars in customer retention because the recommendation system accounts for over 80% of the content streamed on the platform.

**Netflix Recommendation Data Analysis**

Netflix uses big data and analytics to conduct custom marketing, for example as they report, to promote ‘House of Cards’ Netflix cut over ten different versions of a trailer and image placeholders to promote the show. If you watched lots of TV shows centered on women, you get a trailer focused on the female characters. However, if you watched a lot of content directed by David Finch, you would have gotten a trailer that focused the trailer on him and so on. The use of Powerful analytics models to process terabytes of information and turn it into useful business data and logic.

**Netflix infrastructure**

Netflix has a real-time data infrastructure based on in-house technologies that have developed themselves and incorporated into platforms such as HADOOP, Python AWS, HIVE, PIG, HBASE, Cassandra, Presto, Spark, S3, Kafka, Amazon web services, and Genie as shown below *(Netflix Research, 2020).*



**Data Process Analytics:**

Netflix has several areas in which employs its operations to process and analyze data

1. Analytics
2. Content studio
3. Experimentation R&D
4. Machine learning
5. Personalization search
6. User recommendations.

Each of these departments has access to pull and sample data that it can use for their day to day operation to produce results and make business decisions for their platform.

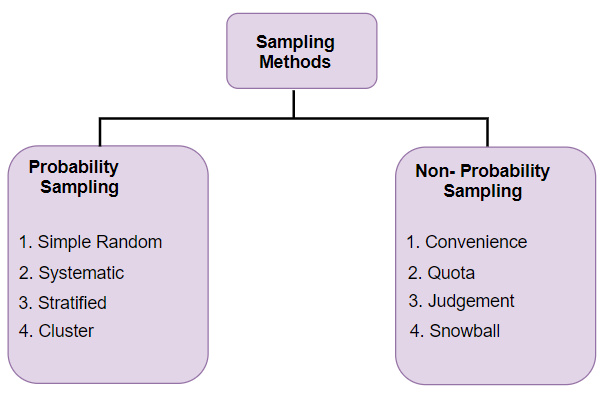
**Why Sampling Big Data is important:**

Sampling is done to conclude populations from samples, and it enables us to determine a population’s characteristics by directly observing only a portion (or sample) of the population.

* Selecting a sample requires less time than selecting every item in a population
* Sample selection is a cost-efficient method
* Analysis of the sample is less cumbersome and more practical than an analysis of the entire population.

**Sampling Techniques**

Netflix use a wide variety of sampling techniques derived from 2 base classifications of sampling methods for different business operations.



Within each of these categories of techniques lies subcategories of techniques that can be to derive the data conclusions based on the output and object(i.e. Analytics, modeling, Machine learning, etc.) *(GANGWAL, 2019)*

* **Probability Sampling:** In probability sampling, every element of the population has an equal chance of being selected. Probability sampling gives us the best chance to create a sample that is truly representative of the population.
* **Non-Probability Sampling**: In non-probability sampling, all elements do not have an equal chance of being selected. Consequently, there is a significant risk of ending up with a non-representative sample that does not produce generalizable results.

**How Netflix uses Sampling**

Since there are so many operations that use data This example will cover the basis of sampling methods Netflix uses for its recommendations systems and analytics departments. I would like to note that Netflix has thousands of clusters across a broad open-source network. Allowing them to operate on a massive scale while having a platform that actively employs techniques and resources to remove bottlenecks. Using an open-source platform called Open Connect *(Netflix Research, 2020)*.

**Netflix Sampling**

* Netflix uses implements fast storage which is precultured data ready to be used by data visualization and interactive queries.
* Pre-clustering observations by user groups/apps to help remove noise from user behaviors with main access to fast and direct data storage .
* Real-time data analytics push or pull data infrastructure
* Classification - classifier models are used on both real-time and historical data to determine the probability of access to a given cluster by a user based on features.
* Dynamic threshold -Each classifier is assigned to a threshold determined programmatically to capture the bulk of legitimate accesses.
* They use a schedular too on amazon S3 that mirrors there data allowing them to spin up a new cluster even if the original cluster failed.

**Pre-clustering and Cluster Sampling:**

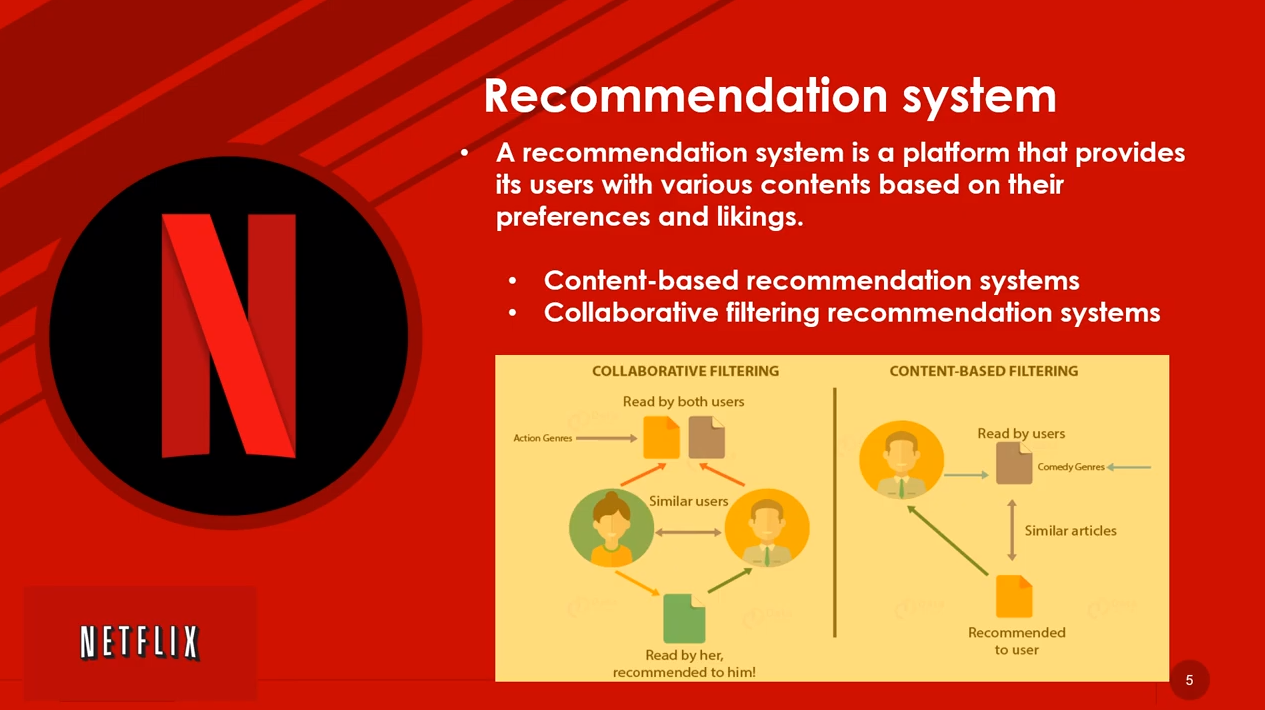
Cluster sampling divides a population into smaller groups known as clusters. They then randomly select among these clusters to form a sample. Cluster sampling is a method of probability sampling that is often used to study large populations, particularly those that are widely geographically dispersed, such as Netflix shows and movies.

There are several types of clustering methods *(Blog, 2018):*

* **Centroid Clustering**- In centroid cluster analysis you choose the number of clusters that you want to classify.
* **Density Clustering** - Density clustering groups data points by how densely populated they are.
* **Distribution Clustering** - Distribution clustering identifies the probability that a point
* **Connectivity Clustering** - Unlike the other three techniques of clustering analysis reviewed above, connectivity clustering initially recognizes each data point as its cluster.
* **Multi-Stage Clustering**  - multi-stage clustering, rather than collect data from every single unit in the selected clusters, you randomly select individual units from within the cluster to use as your sample.

Pre-clustering is an applied analytics method that is implemented to a clustering method *(Reed Hastings, 2009).*

**Sampling user Data for Recommendations**



Netflix devoted a lot of resources and an open-source challenge to develop its solutions for user recommendations of movies and TV shows on its platform. In 2006 they held a $1 million competition to solve this problem .

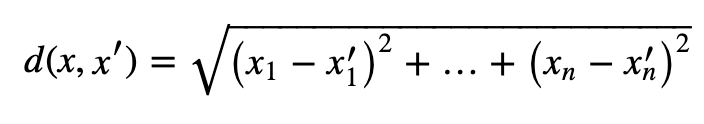
Sampling and analytics techniques that implement the solution were based on a pre-clustering data set comprised of 100 movie ratings and associated data. The techniques and algorithms used to pre cluster their data for recommendations are as follows *(Koren, 2009).*

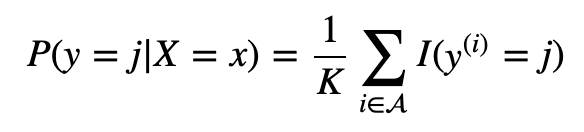
* **K- nearest neighbors’**algorithms for post-processing data(Pre-clustering)
* **Singular value decomposition (SVD)Restricted Boltzmann Machine (RBM)**

**k-nearest neighbors (KNN)**

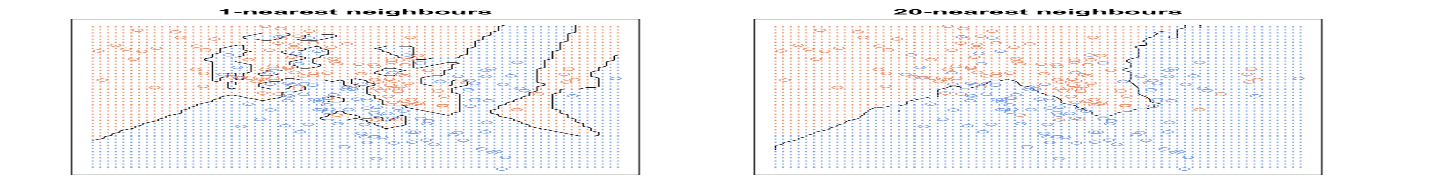
The k-nearest neighbors (KNN) algorithm is a simple, easy-to-implement supervised machine learning algorithm. The KNN algorithm assumes that similar things exist nearby. In other words, similar things are near to each other. This means that we have a dataset with labels training measurements (x, y) and would want to find the link between x and y. CNN's main disadvantage is of becoming significantly slower as the volume of data increases

For distance metrics, we will use the Euclidean metric.



Finally, the input x gets assigned to the class with the largest probability. 

We can use regresions to then determie the best fit line of the classifer based off a hyperparameter to determine a given region and K-averages *(EduPristine, 2015).*



**Singular value decomposition (SVD)**

Netflix implemented SVD with Latent Semantic Indexing(LSI) to reduce the dimensionality of the clustered dataset. Latent semantic indexing (LSI) is a concept used by search engines such as Netflix to discover how a term and content work together to mean the same thing, even if they do not share keywords or synonyms. SVD was used in correlation to LSI for rank reduction of the matrix while capturing the latent relationships of the data *(Mabey, 2008).*

SVD- The math behind LSI
   Singular Value Decomposition

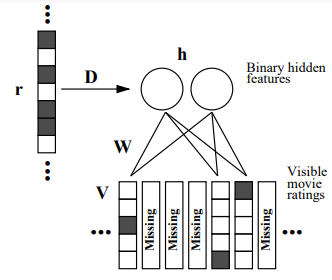
      For any M x N matrix A of rank r, it can
      decomposed ...

**Restricted Boltzmann Machine (RBM)**

Restricted Boltzmann Machine (RBM) is a two-layer bipartite graph i.e. it consists of one layer of hidden units (h) and another layer of visible units (v) all interconnected except in the same layer. this was used for collaborative filtering. Collaborative filtering models try to capture the interactions between users and items that produce different rating values.

Boltzmann machines are stochastic and generative neural networks capable of learning internal representations and can represent and (given sufficient time) solve difficult combinatoric problems.

Boltzmann machines are non-deterministic (or stochastic) generative Deep Learning models with only two types of nodes — hidden and visible Netflix used a restricted Boltzmann machine with binary hidden units and variables called SoftMax visible units. where used for each user in the data set based on the movies the user had rated. M movies, N users, and integer rating values from 1 to K (Using K-nearest neighbors cluster values) (Koren, 2009; Salakhutdinov, 2020)*.*

Essentially distributed samples are clusters and preprocessed through a data set filtered through SVD algorithm and K-nearest regression while implementing Restricted Boltzmann Machine (RBM) deep learning models to correlate the recommendation data to provide the most accurate result to give back user recommendations per user.

Sampling gives data scientists and analytics professionals the tools to create an accurate solution on only a segment representation of a data set. Allowing us to have features and user experiences we have all come to know and love from platforms like Netflix**.**

# References

Blog, T. D. (2018, June 19). *4 Types of Cluster Analysis Techniques Used in Data Science .* Retrieved from http://blog.decisivedata.net/: http://blog.decisivedata.net/blog/4-types-of-cluster-analysis-techniques-used-in-data-science

EduPristine. (2015, July 28). *Beyond the k-Means – Prepping the Data.* Retrieved from https://www.edupristine.com: https://www.edupristine.com/blog/k-means-algorithm

GANGWAL, R. (2019, Sept 11). *A Data Scientist’s Guide to 8 Types of Sampling Techniques.* Retrieved from https://www.analyticsvidhya.com: https://www.analyticsvidhya.com/blog/2019/09/data-scientists-guide-8-types-of-sampling-techniques/

Koren, Y. (2009, Aug). *The BellKor Solution to the Netflix Grand Prize.* Retrieved from https://www.asc.ohio-state.edu/: https://www.asc.ohio-state.edu/statistics/dmsl/GrandPrize2009\_BPC\_BellKor.pdf

Mabey, B. (2008, Sept 24). *SVD and the Netflix Dataset.* Retrieved from https://www.slideshare.ne: https://www.slideshare.net/bmabey/svd-and-the-netflix-dataset-presentation

Netflix Research. (2020). *Netflix Research.* Retrieved from https://research.netflix.com: https://research.netflix.com/articles?q=Analytics

Reed Hastings, C. N. (2009, Aug 1). *Culture.* Retrieved from https://www.slideshare.net: https://www.slideshare.net/reed2001/culture-1798664/12-12ImpactYou\_accomplish\_amazingamounts\_of\_important

Salakhutdinov, R. (2020). *Restricted Boltzmann Machines.* Retrieved from https://www.cs.toronto.edu/: https://www.cs.toronto.edu/~hinton/absps/netflix.pdf