**Week7**

**###**

**set.seed(123)#设置随机数**

**data<-rnorm(10000,5,3)#生成一个均值为5，标准差为3的随机正太分布**

**data**

**n<-1000**

**reserve<-data[1:1000]**

**#check**

**data[1000:1005]**

**reserve[999:1000]**

**set.seed(123)#设置随机数**

**runif(1)#生成一个在[0,1]区间上均匀分布的随机数**

**1000/1001#计算1000/1001的值**

**runif(1)<=1000/1001#生成一个随机数检查是否小于等于1000/1001的值**

**runif(10)<=1000/1001#生成10个随机数检查是否都小于等于**

**runif(10)<=0.5#生成10个随机数，检查是否小于等于0.5**

**set.seed(123)**

**table(runif(1000)<=0.9)#生成1000个在[0,1]区间上均匀分布的随机数，并检查它们是否都小于等于0.9**

**set.seed(123)**

**table(runif(1)<=1000/1001)**

**###**

**set.seed(123)**

**data<-rnorm(10000,5,3)#均值为5，标准差为3的正太分布**

**sample(1:1000,1)#roll a number从1到1000中随机选择一个数**

**reserve<-data[1001]#将第1001个元素存在reserve中**

**reserve**

**data[1002]#显示data的第1002个元素的值**

**reserve[597]<-data[1002]#将1002个元素的值赋值给597的数**

**reserve[597]**

**table(runif(sample(1:1000,1))<=1000/1001)#生成一个随机数，并检查它是否小于等于1000/1001**

**###Reservoir Sampling的算法来从一个大数据集中随机选择样本**

**set.seed(123)**

**for(i in 1001:10000){**

**if(runif(1)<=(1000/i)){**

**id<-sample(1:1000,1)**

**reserve[id]<-data[i]**

**}**

**}**

**# 对于数据集中的每一个元素**

**# 如果一个在[0,1]区间上均匀分布的随机数小于等于1000/i**

**# 从1到1000中随机选择一个数**

**# 将data的第i个元素的值赋给reserve的第id个元素**

**set.seed(123)**

**data<-rnorm(10000,5,3)**

**reservesample<-function(data,n){**

**reserve<-data[1:n]**

**N<-length(data)#获取数据的长度**

**for(i in (n+1):N){**

**if(runif(1)<=(n/i)){**

**id<-sample(1:n,1)**

**reserve[id]<-data[i]**

**}**

**}**

**return(reserve)**

**}**

**reserve**

**set.seed(123)**

**reservesample(data,1000)#从data中随机选择1000个样本**

**###**

**reservesample<-function(data,n){**

**reserve<-data[1:n]**

**N<-length(data)**

**i<-n+1**

**while(i<=N){**

**if(runif(1)<=(n/i)){#如果一个在[0,1]区间上均匀分布的随机数小于等于n/i**

**id<-sample(1:n,1)## 从1到n中随机选择一个数**

**reserve[id]<-data[i]## 将data的第i个元素的值赋给reserve的第id个元素**

**}**

**i<-i+1**

**}**

**return(reserve)**

**}**

**###reservoir sampling**

**reservesample<-function(data,n){**

**reserve<-data[1:n]**

**N<-length(data)**

**for(t in (n+1):N){**

**u<-sample.int(t,1)**

**if(u<=n){**

**reserve[u]<-data[t]**

**return(reserve)**

**}**

**}**

**return(reserve)**

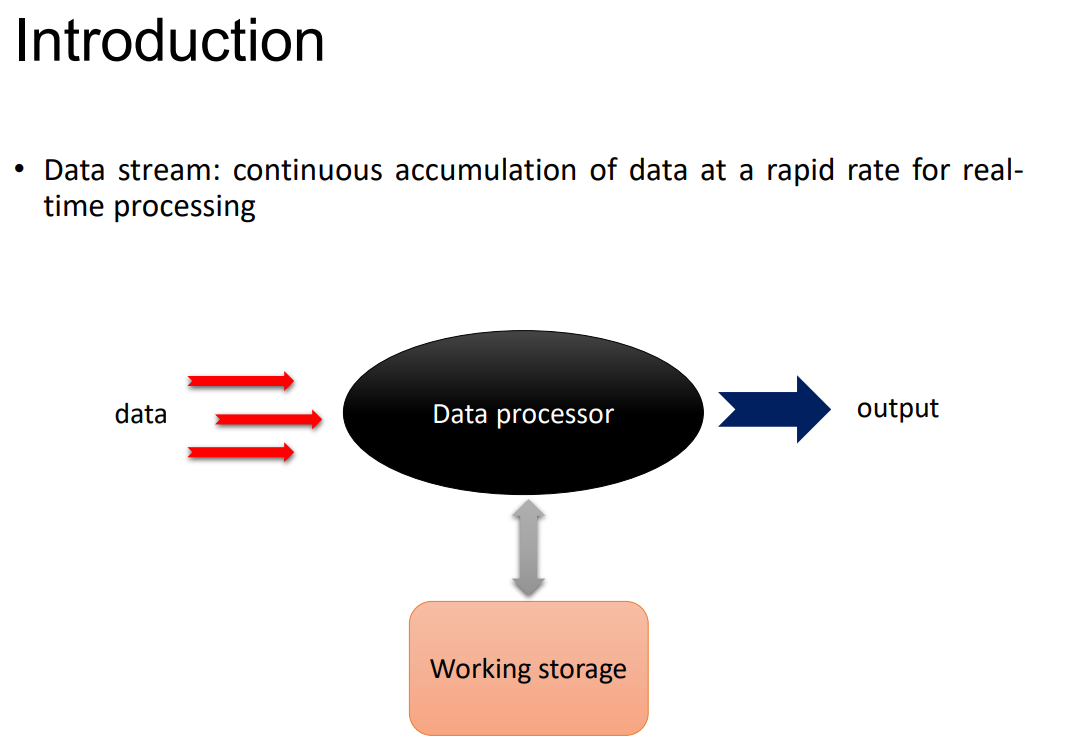
**}**

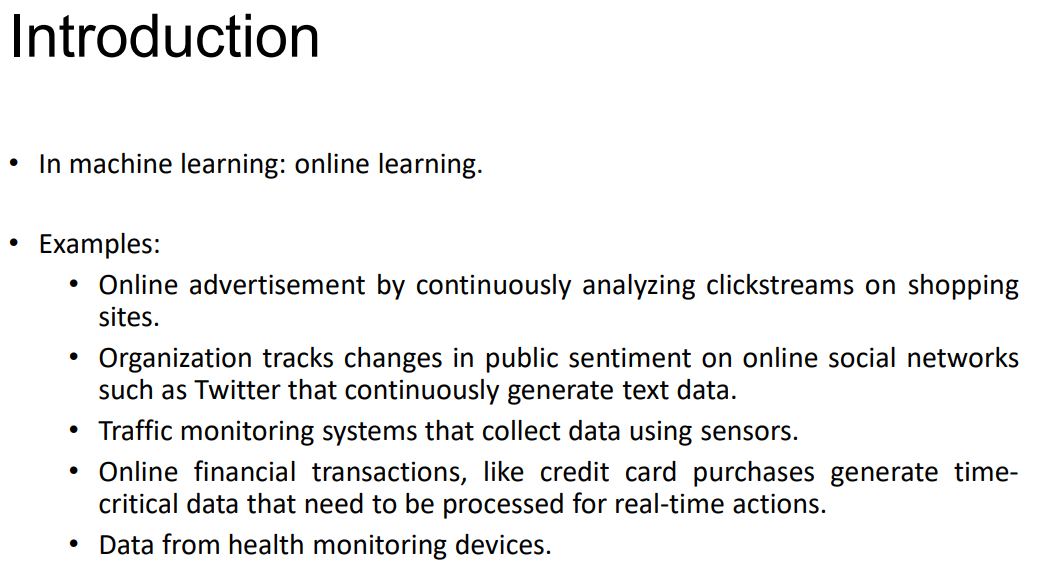
**reservesample(data,1000)**

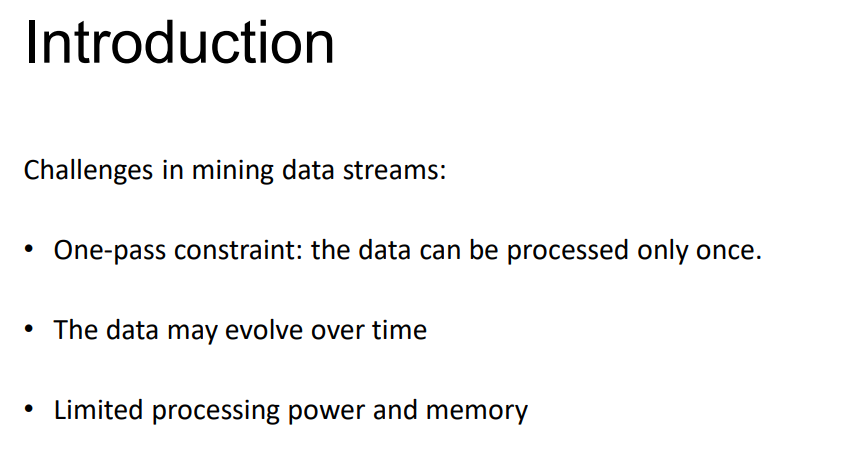
**x<-reservesample(data,1000)**

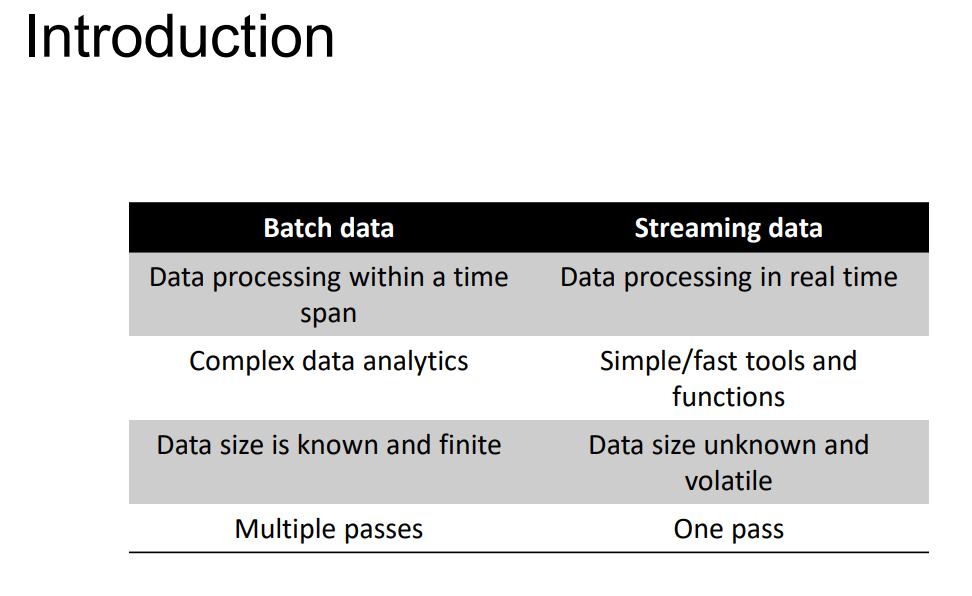
**mean(x)#计算样本的均值**

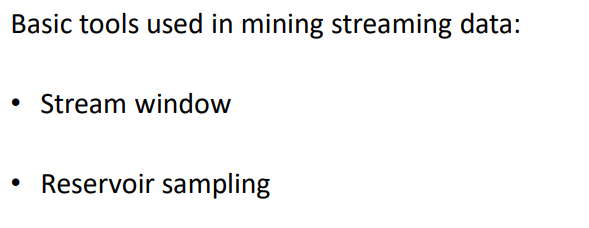
**sd(x)# 计算样本的标准差**

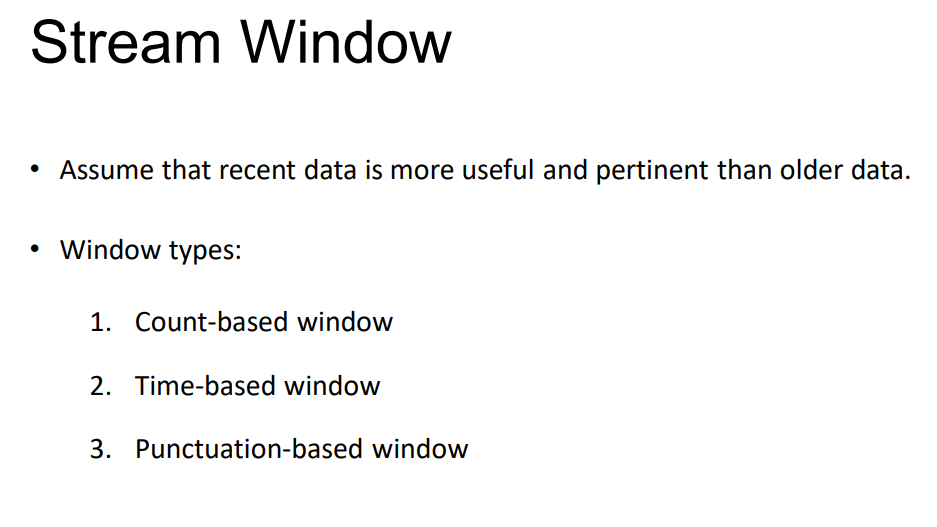
****

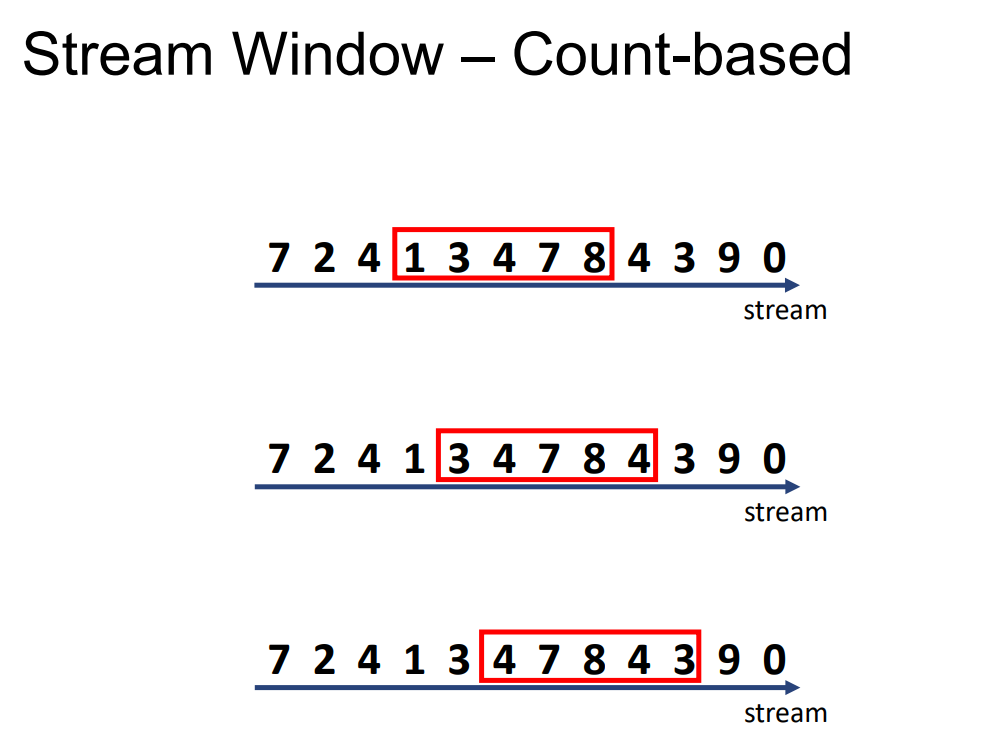
****

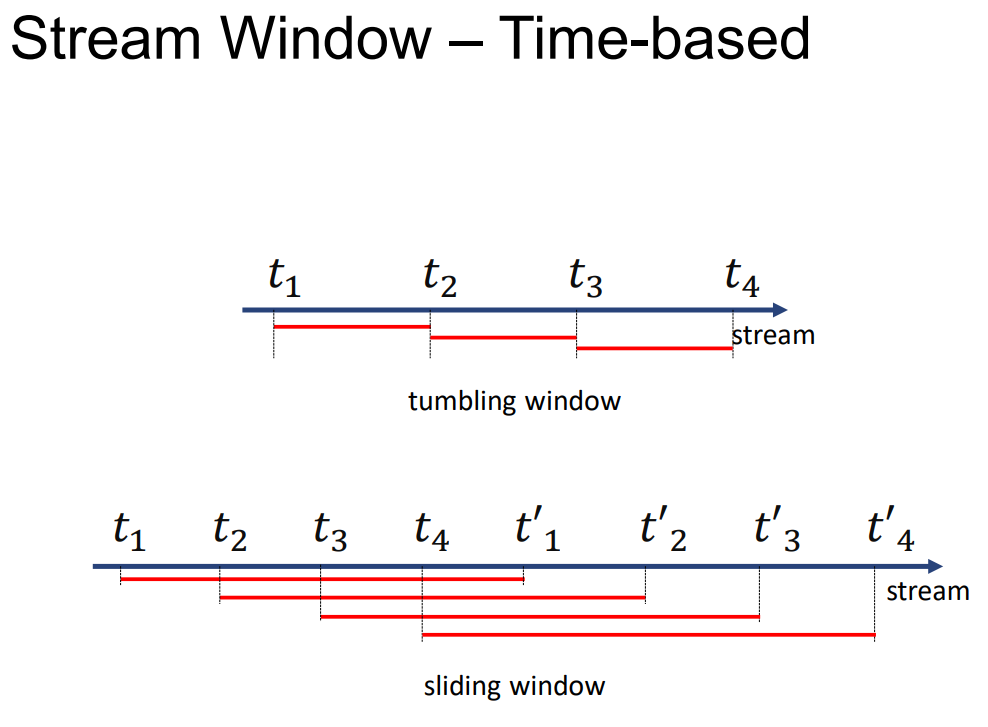
****

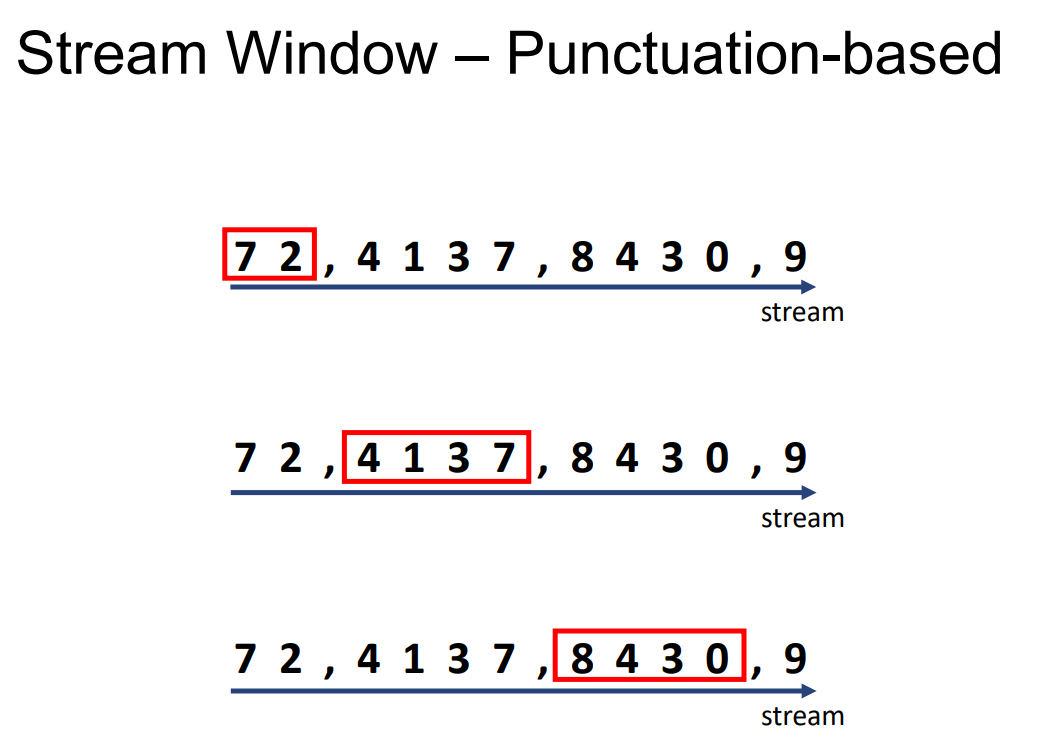
****

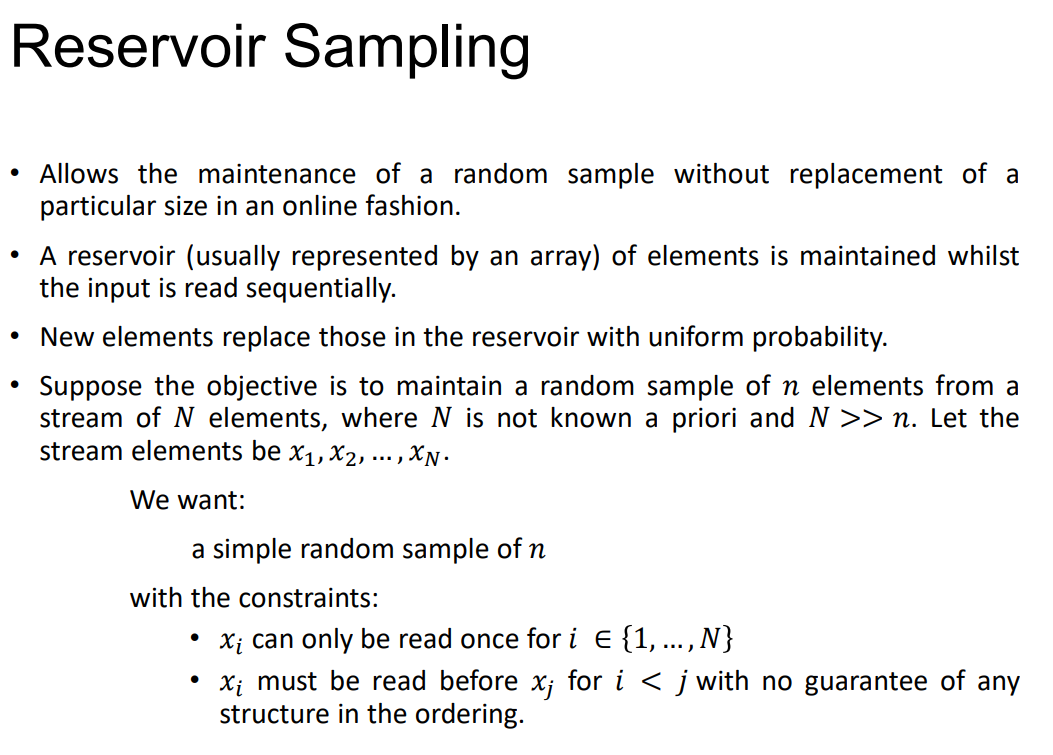
****

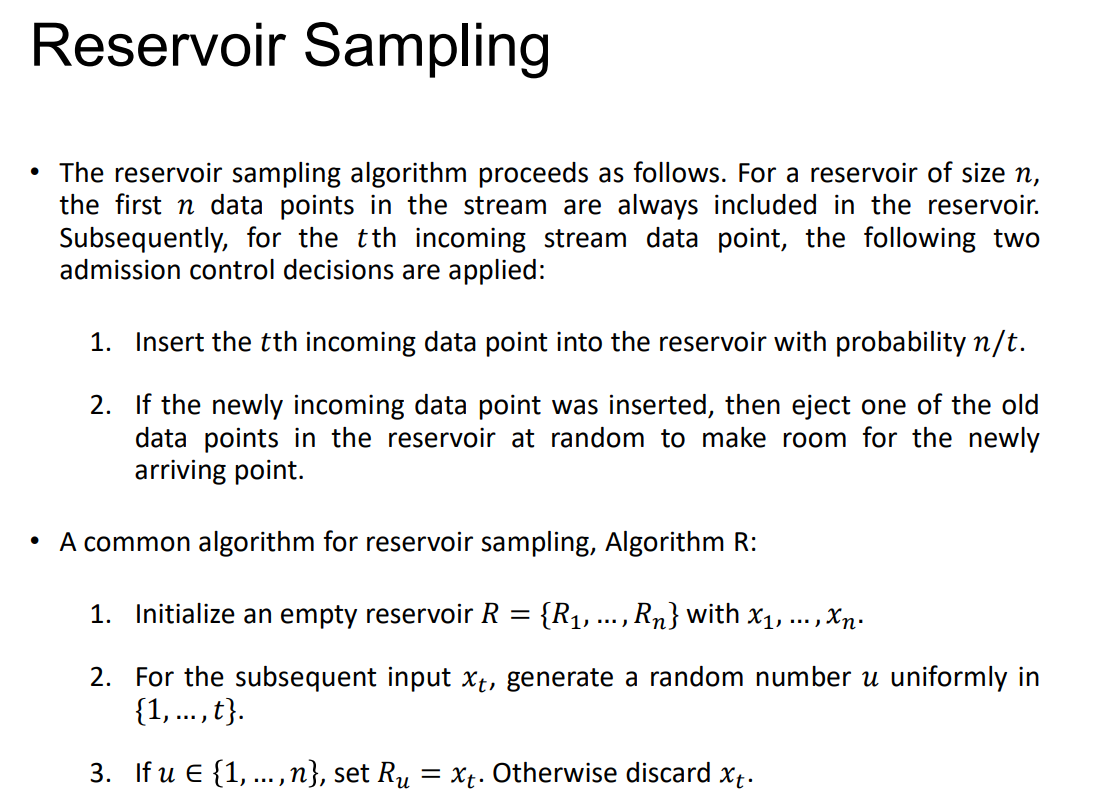
****

****

****

****

****

****