Fundamentos de Data Stream Mining

Fabrício Enembreck PhD

Encontro 1

Introdução a DSM

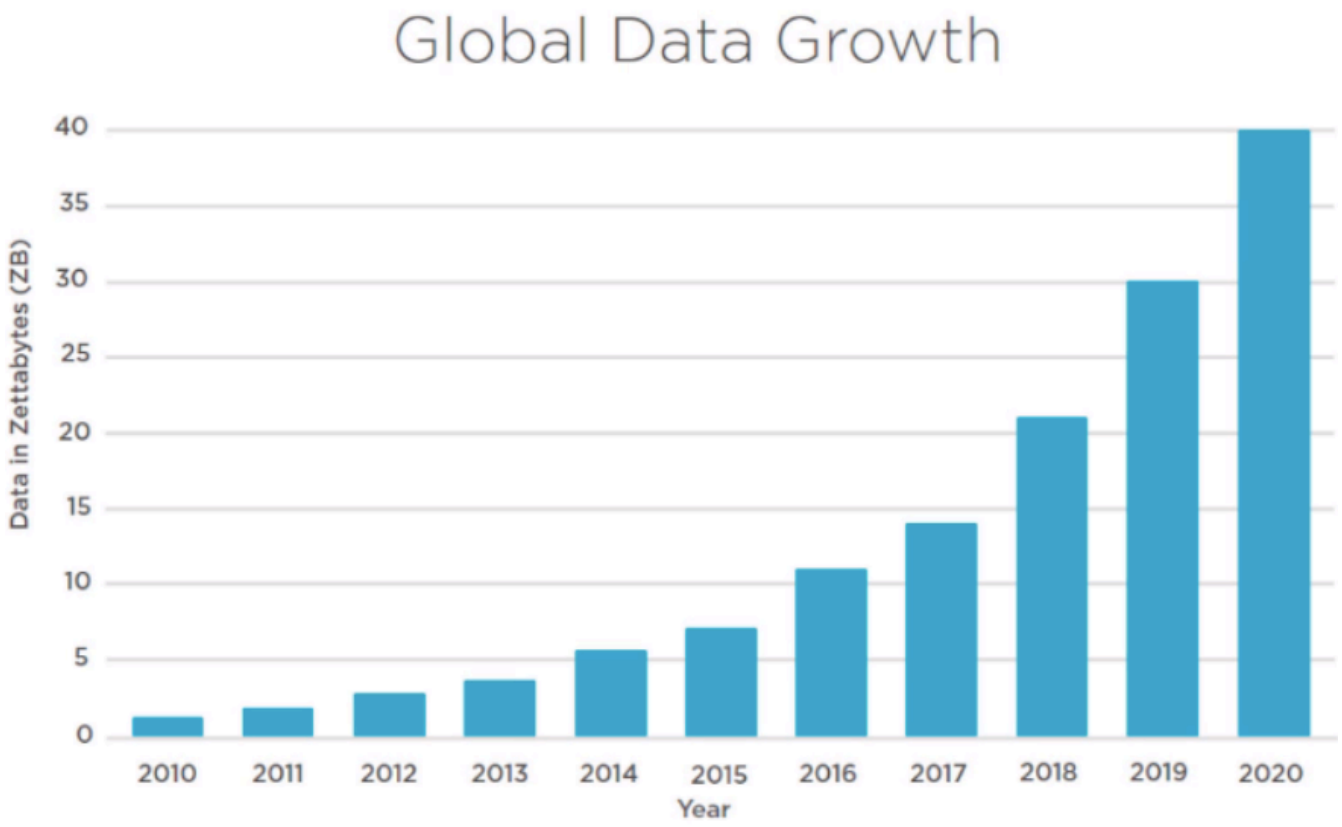


Figura 6—Tendência de crescimento do volume de dados durante os anos (Fonte: UNECE Statistics wikis)

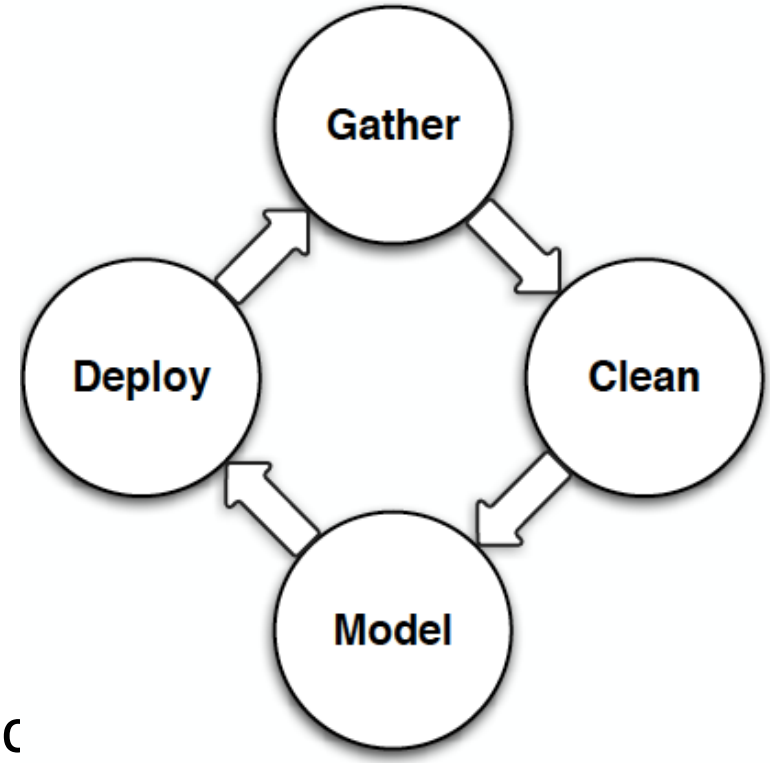
Prefixo do SI		
Nome	**Símbolo**	**Múltiplo**
byte	B	10^{0}
kilobyte	kB	10^{3}
megabyte	MB	10^{6}
gigabyte	GB	10^{9}
terabyte	TB	10^{12}
petabyte	PB	10^{15}
exabyte	EB	10^{18}
zettabyte	ZB	10^{21}
yottabyte	YB	10^{24}
undecabyte	UB	10^{33}

Fonte: Wikipedia

Dados x Recursos

Ciclo da Mineração de Dados

- Processo
  - Selecionar Dados
  - Preparar Dados
  - Gerar Modelo/Avaliar
  - Implantar



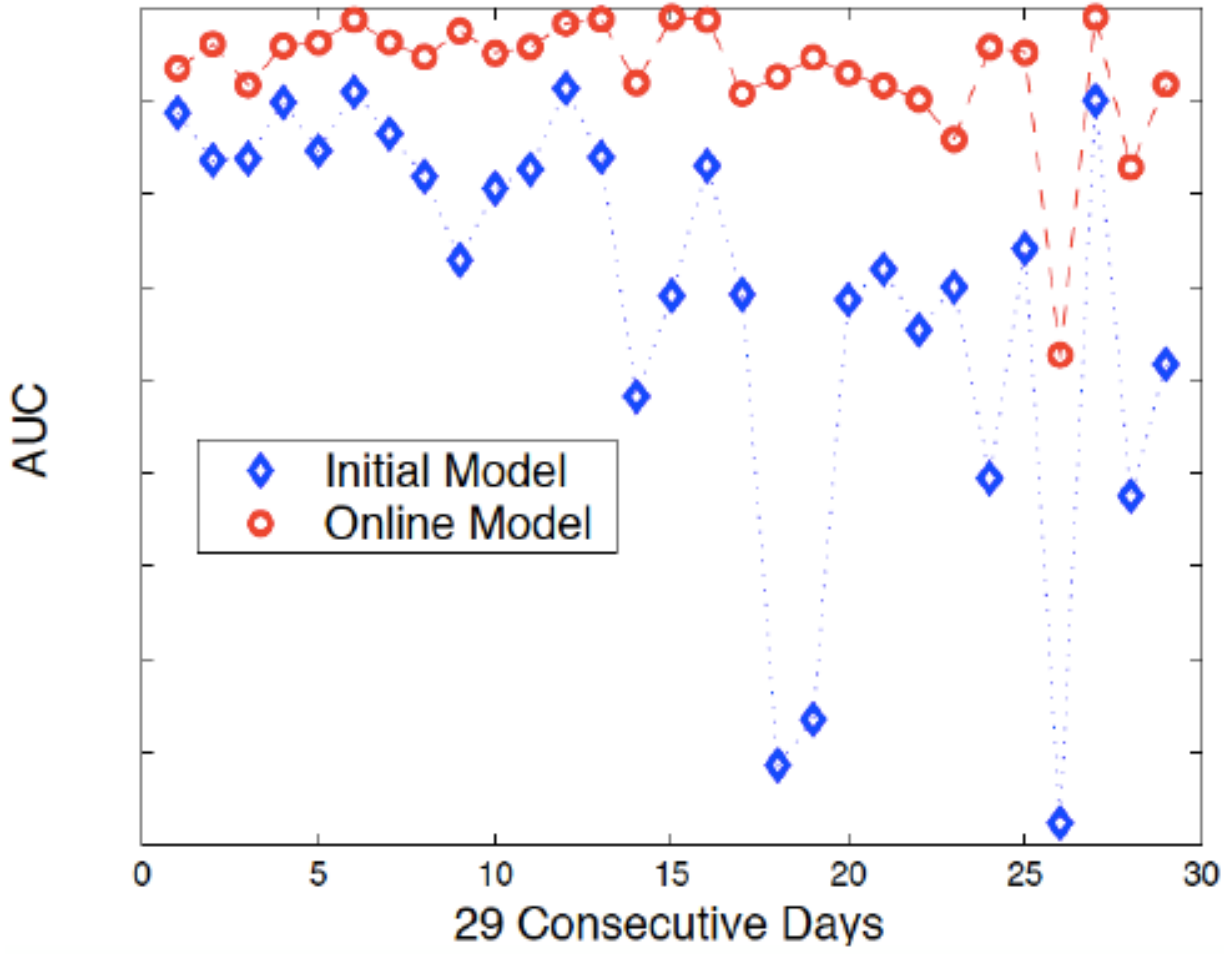
- Características
  - Forte dependência de espec
  - Quantidade de dados é finita
  - Modelo em produção não muda (estático)
  - Projeto de modelagem pode durar meses!

Ciclo da Mineração de Dados (cont.)

- Modelo degrada
- Modelo precisa ser retreinado
- Dados precisam ser armazenados, mesmo sem ser usados
- Porque
  - Coisas mudam
  - Processo de geração de dados muda
  - Usuários mudam
  - Perfis de compra mudam

Ciclo da Mineração de Dados (cont.)

- Visualmente



Ciclo da Mineração de Dados (cont.)

- Retreinar é fácil?
  - Por que modelo degradou?
  - Quando ocorreu(ram) a(s) mudança(s)?
  - Quais variáveis foram afetadas?
  - Qual porção de dados reflete a mudança?
  - Quais dados selecionar?
    - 6 meses? 12 meses?
    - Junta com dados antigos ou treina apenas com mais recentes? Mas afinal, o que significa “mais recentes”?
- Restrições computacionais e algoritmicas

Mineração de Fluxo de Dados

- Modelos online
  - Treina com dados mesmo em produção
  - Quantidade de dados pode ser infinita
  - Mudanças devem ser detectadas automaticamente
  - Alterações no modelo também são automáticas
  - Modelos tornam-se dinâmicos
  - Ciclo de vida praticamente infinito

Característica de uma Stream de Dados

- Volume potencialmente infinito
- Dados podem ser voláteis
- Não pode ser armazenada inteiramente em memória
- Pode ter elevada dimensionalidade
- Podem ser fontes variadas
  - ecommerce, transações bancárias, sensores, cliques de usuários, fluxos de rede, mensagens de redes sociais, etc.
- Pode possuir dependência temporal

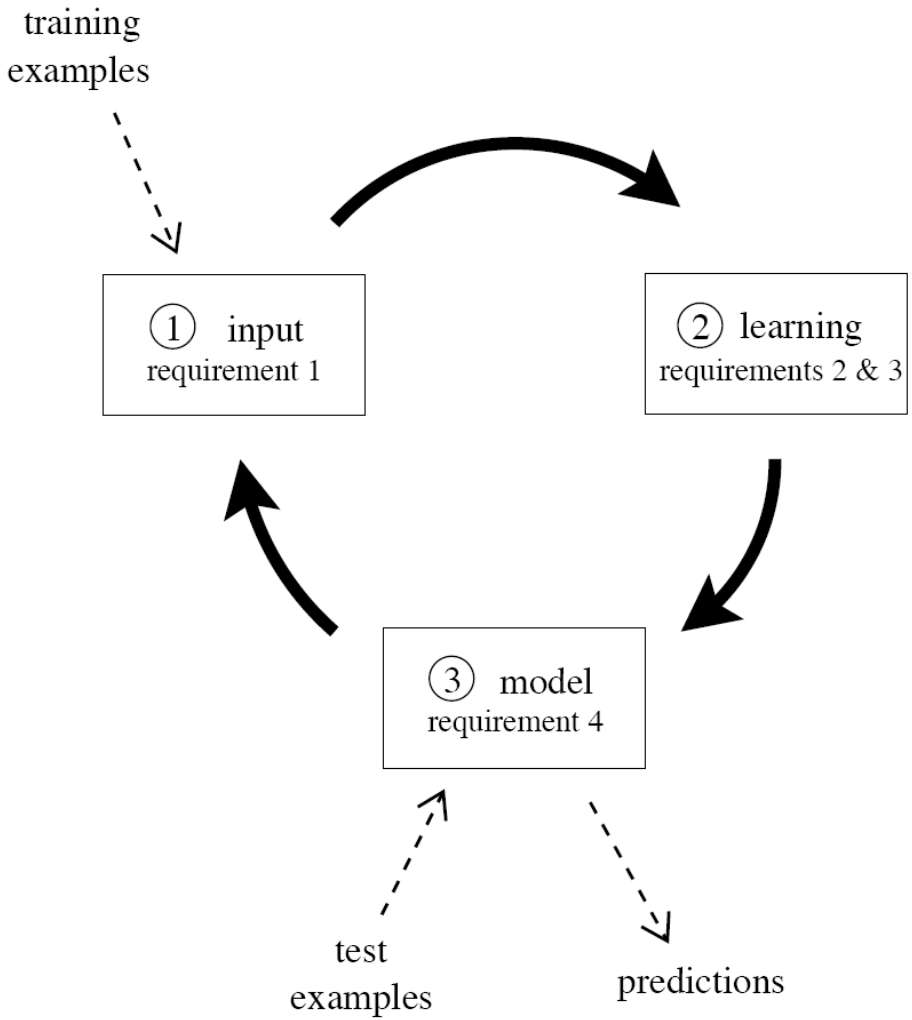
Características de um Algoritmo de DSM

- Eficiente
  - Usa pouca memória
  - Usa pouco processamento
- Explora paralelismo (threads, GPU, big data frameworks)
- Precisa detectar mudanças
  - Explicitamente
  - Implicitamente
- Precisa adaptar o modelo gerado às mudanças

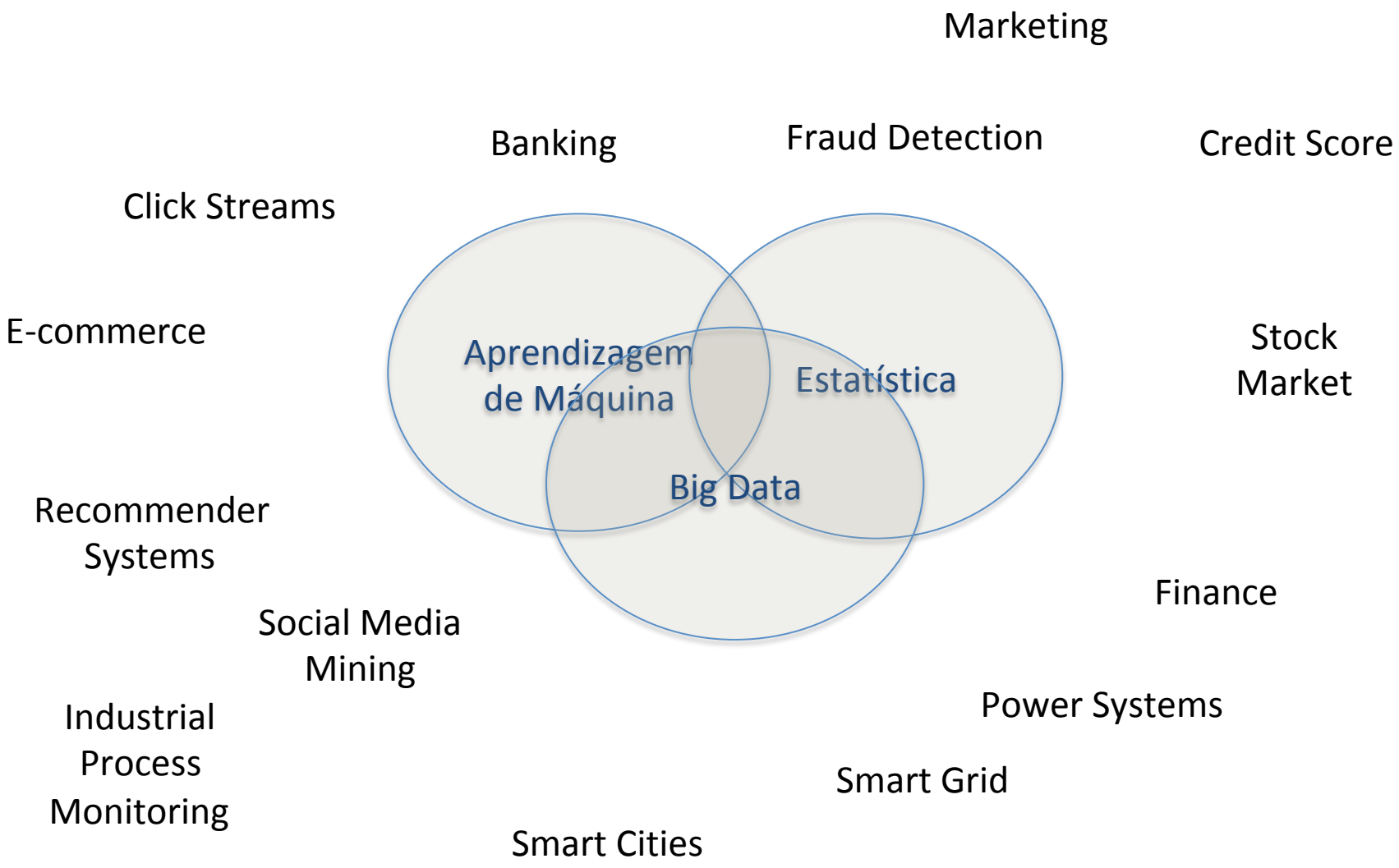
Características do Processo de DSM

- Cada instância é usada apenas uma vez e descartada
- Processamento deve ser rápido para evitar bufferização/perda de dados
- Precisa ser monitorado
- Reduz a dependência do especialista
- Ciclo de vida longo
- Modelo anytime

Características do Processo de DSM



Áreas Relacionadas a DSM



Técnicas Incrementais Elementares

- Calculando a média de uma variável x
  - Some x_i em s para cada novo valor informado em um instante i
  - Quando precisar calcular a média, calcule s/n, sendo n a quantidade de observações

Técnicas Incrementais Elementares

- Calculando o desvio padrão de uma variável x
  - Some x_i em s para cada novo valor informado em um instante i
  - *Some* x_i^2 em s' para cada novo valor informado em um instante i
  - Quando precisar calcular o desvio no instante n, calcule:

$$d_{x\_n} = \text{sqrt}((s' - (s^2/n)) / (n - 1))$$

Técnicas Incrementais Elementares

- Calculando a correlação entre duas variáveis x e y
  - Some x_i em x*s* e y_i em *ys* para cada novo valor informado de x e y em um instante *i*
  - *Some* x_i^2 em xs' e y_i^2 *em ys'* para cada novo valor informado de x e y em um instante *i*
  - Some ($x_i * y_i$) em p para ter o produto vetorial

$$corr(x,y) = (p - (xs*yx) / n) /$$

$$sqrt(xs' - xs'/n)*sqrt(ys' - ys'/n)$$

Técnicas Incrementais Elementares

- Média, desvio e correlação, mesmo sendo calculadas de forma incremental, são exatas
- Ajudam a descrever o comportamento de variáveis mesmo com quantidade infinita de observações

Técnicas Incrementais Elementares

- Intervalo aproximado de uma variável
  - Como não é possível armazenar todos os valores para uma variável, muitas vezes é necessário estimar os seus limites (mínimo e máximo) a partir da média atual

Absolute approximation: $\overline{X} - \epsilon \leq \mu \leq \overline{X} + \epsilon$, where ϵ is the absolute error;

Relative approximation: $(1-\delta)\overline{X} \leq \mu \leq (1+\delta)\overline{X}$, where δ is the relative error.

Técnicas Incrementais Elementares

- Intervalo aproximado de uma variável
- Hoeffding Bound

Theorem 2.2.3 (Hoeffding Bound) *Let $X_1, X_2, \ldots, X_n$ be independent random variables. Assume that each x_i is bounded, that is $P(X_i \in R = [a_i, b_i]) = 1$. Let $S = 1/n \sum_{i=1}^{n} X_i$, whose expected value is $E[S]$. Then, for any $\epsilon > 0$,*

$$P[S - E[S] > \epsilon] \leq e^{-\frac{2n^2\epsilon^2}{R^2}} \tag{2.3}$$

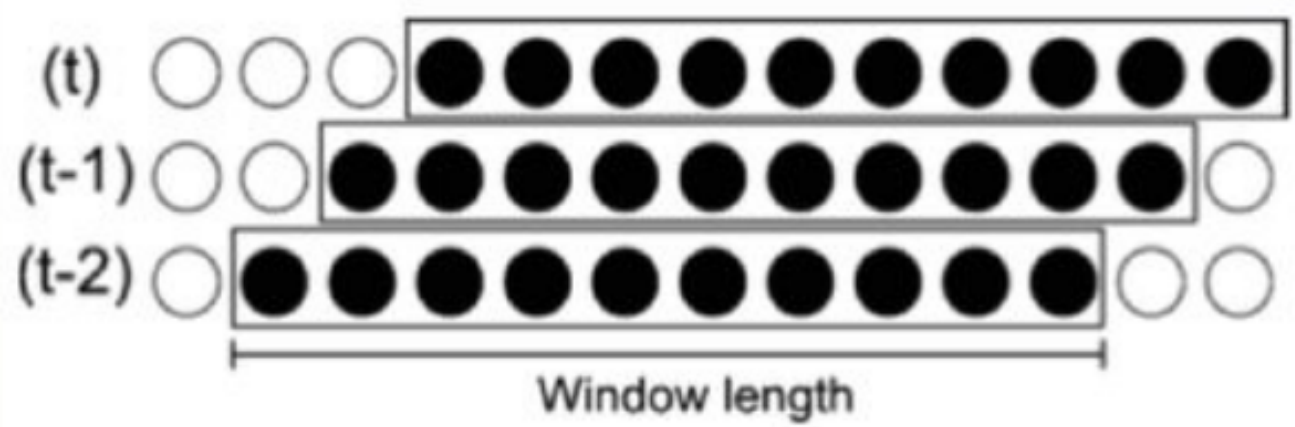
From this theorem, we can derive the absolute error (Motwani and Raghavan, 1997):

$$\epsilon \leq \sqrt{\frac{R^2 ln(2/\delta)}{2n}} \tag{2.4}$$

- Sendo R o range da variável e delta o fator de confiança esperado
- Isso permite dizer a média esperada é >= (r − e) com um fator de confiança, independente da distribuição da variável, para a média r observada

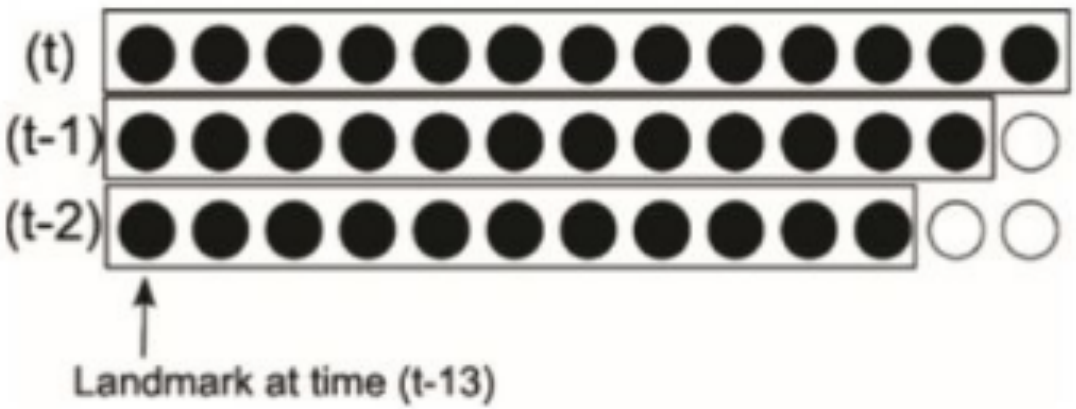
Técnicas de Janelamento

- Sliding Windows
  - Tamanho fixo ou variável (mensal, diário, etc)
  - Processamento FIFO (Fila)
  - Estatísticas dependem do conteúdo da janela



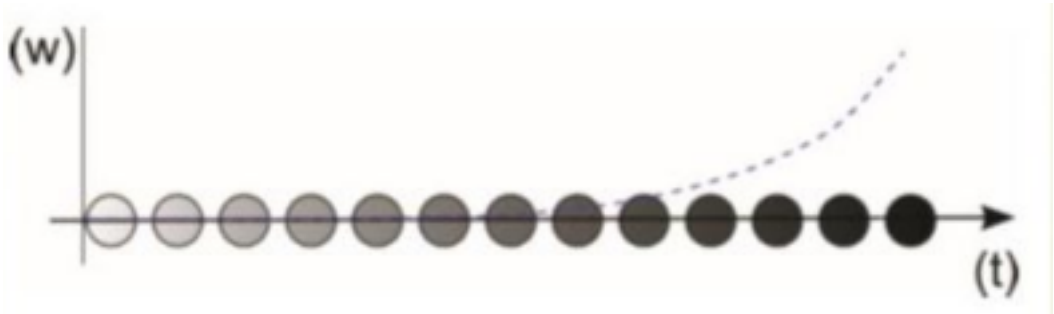
Técnicas de Janelamento

- Landmark Windows
  - Tamanho fixo ou variável (mensal, diário, etc)
  - Tamanho depende de um ponto inicial (landmark)
  - Chunks disjuntos de dados. Um novo landmark é usado para exluir elementos da janela anterior e iniciar uma nova



Técnicas de Janelamento

- Damped Windows
  - Instâncias são ponderadas (mais recentes possuem maior peso, cor preta)
  - Função de decaimento exponencial.
    - Ex: $2^{(-\alpha*(t - t0))}$ para $\alpha > 0$
  - Janela possui instâncias com peso > 0



Atividade em Grupos

- Formar grupos de até 4 pessoas
- Entregar por email (uma única mensagem por equipe) até o final da aula
- Realizar Atividade 1 - Blackboard