

Aula 01: Introdução à Ciência de Dados

PPGGCO – MINTER – INDAIAL

PROF. DR. IGOR DA PENHA NATAL

E-MAIL: IGOR.NATAL@UNICESUMAR.EDU.BR

Passado

POUCOS DADOS



Presente

INUNDAÇÃO DE DADOS



Inundação de dados

- Sem perceber, as pessoas geram dados a todo momento.
- Como?
 - Utilizam cartões de fidelidade (empresa aérea, supermercado, etc.).
 - Compras com cartões.
 - Navegam na internet.
 - Vão a um posto de saúde.
 - Utilizam smartphones.

Inundação de dados

- Dados gerados por empresas:
 - Em um passado próximo, empresas usavam apenas uma pequena parcela dos dados que produzem e armazenam.
 - Qual a importância dos dados coletados?
- Além dos dados internos, há e haverá um grande aumento da quantidade de dados externos.

Dados externos

- São gerados por:
 - Outras empresas;
 - Órgãos públicos;
 - ONGs;
 - Mídias;
 - E muito mais....
- Vocês sabem quais dados estão sendo coletados agora mesmo no seu smartphone?

Causas da inundação de dados

- Avanços recentes nas tecnologias para:
 - Aquisição;
 - Armazenamento;
 - Transmissão;
 - Processamento.
- Hoje é possível processar uma maior quantidade de dados, com mais rapidez e menor custo.

BIG DATA

O que é Big Data?

- Várias definições:
 - Dados que são grandes demais para sistemas tradicionais de processamento de dados.
 - Dados que precisam de novas técnicas para serem processados.
 - Dados que são muito complexos.
 - Dados que são importantes.
 - Desafios e oportunidades decorrentes da disponibilidade de dados sobre tudo.

O que é Big Data?

- Várias definições:
 - Dados que são grandes demais para sistemas tradicionais de processamento de dados.
 - Dados que precisam de novas técnicas para serem processados.
 - Dados que são muito complexos.
 - Dados que são importantes.
 - Desafios e oportunidades decorrentes da disponibilidade de dados sobre tudo.

Características de Big Data

- Grande volume de dados, gerados a uma grande velocidade e com uma grande variedade (3 Vs):
 - Volume: tanto de dados estruturados quanto de não estruturados.
 - Variedade: vindos de fontes diferentes e que precisam ser integrados.
 - Velocidade: gerados em fluxos cada vez mais rápidos.
- Dados gerados geralmente contêm informações relevantes:
 - Uma vez analisados, podem trazer vários benefícios: Sociais, políticos e econômicos.
 - Crescente interesse na análise de dados.
 - Interesse em análise de dados não é uma atividade recente.







Previsões

- Pessoas podem errar previsões, principalmente baseadas na intuição.
- “Os americanos precisam de telefones, nós não, pois temos muitos mensageiros”, William Preece (1878), Engenheiro chefe do serviço postal britânico.
- “Nenhuma base de dados online substituirá meu jornal diário”, Clifford Stroll (1995), famoso astrônomo Americano.
- “É impossível que o iPhone tenha mercado”, Steve Ballmer (2007) presidente da Microsoft.

Dado é dinheiro

TOP 10 RANKING CHANGES SIGNIFICANTLY...

Only three brands that appeared in the BrandZ™ Global Top 10 in 2006—Google, Microsoft, and IBM—remain in the Top 10 in 2017.

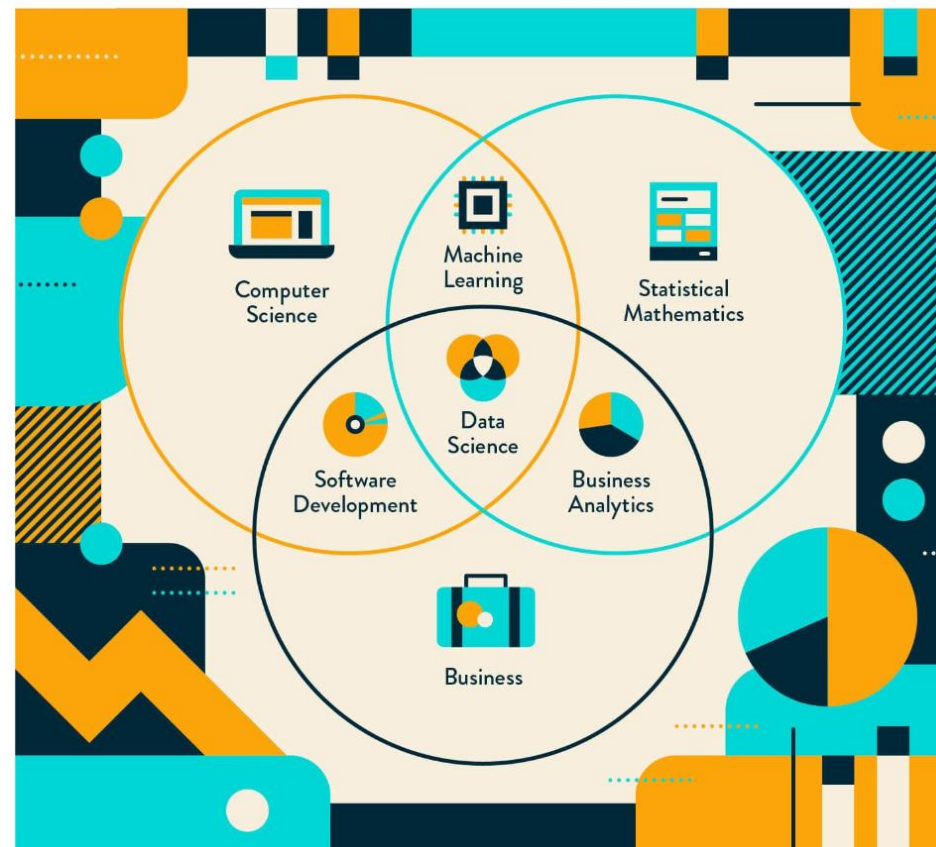
	2006	Brand Value 2006 \$Mil.	2017	Brand Value 2017 \$Mil.
1	 Microsoft	62,039	 Google	245,581
2		55,834		234,671
3		41,406	 Microsoft	143,222
4	 中国移动 China Mobile	39,168	 amazon	139,286
5		38,510	 facebook	129,800
6	 Walmart	37,567	 AT&T	115,112
7	 Google	37,445	 VISA	110,999
8		36,084	 Tencent 腾讯	108,292
9		31,028		102,088
10	 TOYOTA	30,201	 McDonald's	97,723

Source: Kantar Millward Brown / BrandZ™ (including data from Bloomberg)

Dado é dinheiro

- Valor dos dados de 2 bilhões de perfis de usuários do facebook:
 - Estimado em US\$ 32 bilhões em 2012 e US\$ 368 bilhões em 2016.
- Valor global de vendas relacionadas a aplicações de Big Data:
 - US\$ 7 bilhões em 2012 e US\$ 122 bilhões em 2015.
 - US\$ 190 bilhões em 2019.
- Em 2016, quase todo crescimento de arrecadação em propaganda nos EUA foi para Google e facebook.
- Em 2017, 43% de todo o gasto em comércio eletrônico nos EUA foi para a Amazon.
- Até março de 2017, Alphabet, Amazon, Apple, facebook e Microsoft lucraram mais de US\$ 25 bilhões.

Ciência de Dados



Ciência de Dados

- É o estudo disciplinado dos dados e informações inerentes ao negócio e todas as visões que possam cercar um determinado assunto.
- É uma ciência que estuda as informações, seu processo de captura, transformação, geração e, posteriormente, análise de dados.
- A ciência de dados envolve diversas disciplinas:
 - Computação;
 - Estatística;
 - Matemática;
 - Conhecimento do negócio;
 - E mais.....

Ciência de Dados

- Mercado de Trabalho
 - Estudos recentes produzidos nos EUA mostra que Big Data Analytics vai gerar 10 milhões de oportunidades de trabalho em todo o mundo na próxima década.
- Exemplos de carreiras na área:
 - Engenheiros de Dados;
 - Engenheiros de Big Data;
 - Arquiteto de Soluções de Big Data;
 - Cientista de Machine Learning;
 - Engenheiro de Machine Learning;
 - Gerente de Analytics;
 - Cientista de Dados.

Ciência de Dados

- O que estudar para dominar a área?
 - Estatística e Matemática.
 - Algoritmos de Machine Learning.
 - Linguagens de programação com bom suporte na área de análise de dados (Python e R).

Ciência de Dados x Big Data

- Frequentemente usados como sinônimos:
 - Big Data lida com tecnologias para coletar, gerenciar e processar (Big) dados.
 - Ciência de Dados lida com a criação de soluções para modelagem de dados.
 - Capazes de extrair conhecimento de dados reais.

PROCESSAR X DESCOBRIR

Ciência de Dados



Importância da Ciência de Dados

- Tomada de decisões:
 - Decisões mais assertivas e estratégicas, baseadas em evidências e insights extraídos de dados.
- Inovação e Competitividade:
 - Novos produtos e serviços.
 - Otimizar processos e se manter à frente da concorrência.
- Soluções para a Sociedade:
 - Soluções para problemas sociais, saúde, educação, meio ambiente e segurança.

Técnicas de Ciência de Dados

- Análise dos dados por seres humanos:
 - Falta de especialistas.
 - Custo elevado.
 - Subjetividade.
 - Grande volume.
- Técnicas tradicionais de análise:
 - Planilhas.
 - Sistemas de Gerenciamento de Bancos de Dados.

Técnicas de Ciência de Dados

- Técnicas tradicionais de análise de dados permitem apenas consultas simples.
 - Quantos itens de um produto em particular foram vendidos em um dado dia?
 - Não conseguem responder consultas do tipo:
 - Que novo filme eu gostaria de assistir?
 - Dado o que estou sentindo, posso estar doente?
 - O que significa esse texto em chinês?
- Por isso utilizamos o Aprendizado de Máquina.

Aprendizado de Máquina

- Investiga técnicas computacionais capazes de adquirir automaticamente:
 - Novas habilidades, conhecimentos e formas de organizar o conhecimento existente.
- Definições:
 - Área de pesquisa que dá aos computadores a habilidade de aprender sem ser explicitamente programado (Arthur Samuel, 1959).
 - Técnicas capazes de melhorar seu desempenho em uma dada tarefa utilizando experiências prévias (Mitchell, 1997).

Aprendizado de Máquina

- Programas baseados em AM têm sido bem sucedidos para:
 - Análise de redes sociais;
 - Análise de dados biológicos;
 - Detecção de fraudes;
 - Diagnóstico médico;
 - Biometria;
 - Recomendação de filmes e séries;
 - E muito mais...

Aprendizado de Máquina

- Programas baseados em AM têm sido bem sucedidos para:
 - Análise de redes sociais;
 - Análise de dados biológicos;
 - Detecção de fraudes;
 - Diagnóstico médico;
 - Biometria;
 - Recomendação de filmes e séries;
 - E muito mais...