Estatística para Data Science

Paulo Henrique Sales Guimarães

Departamento de Estatística - UFLA

Conteúdo da palestra

Conteúdo da palestra

- Dados, informação e conhecimento

Conteúdo da palestra

- Dados, informação e conhecimento
- Data Science

Conteúdo da palestra

- Dados, informação e conhecimento
- Data Science
- Big data

Conteúdo da palestra

- Dados, informação e conhecimento
- Data Science
- Big data
- Estatística para data science

Conteúdo da palestra

- Dados, informação e conhecimento
- Data Science
- Big data
- Estatística para data science
- Machine learning

Dados, informação e conhecimento

Dados, informação e conhecimento

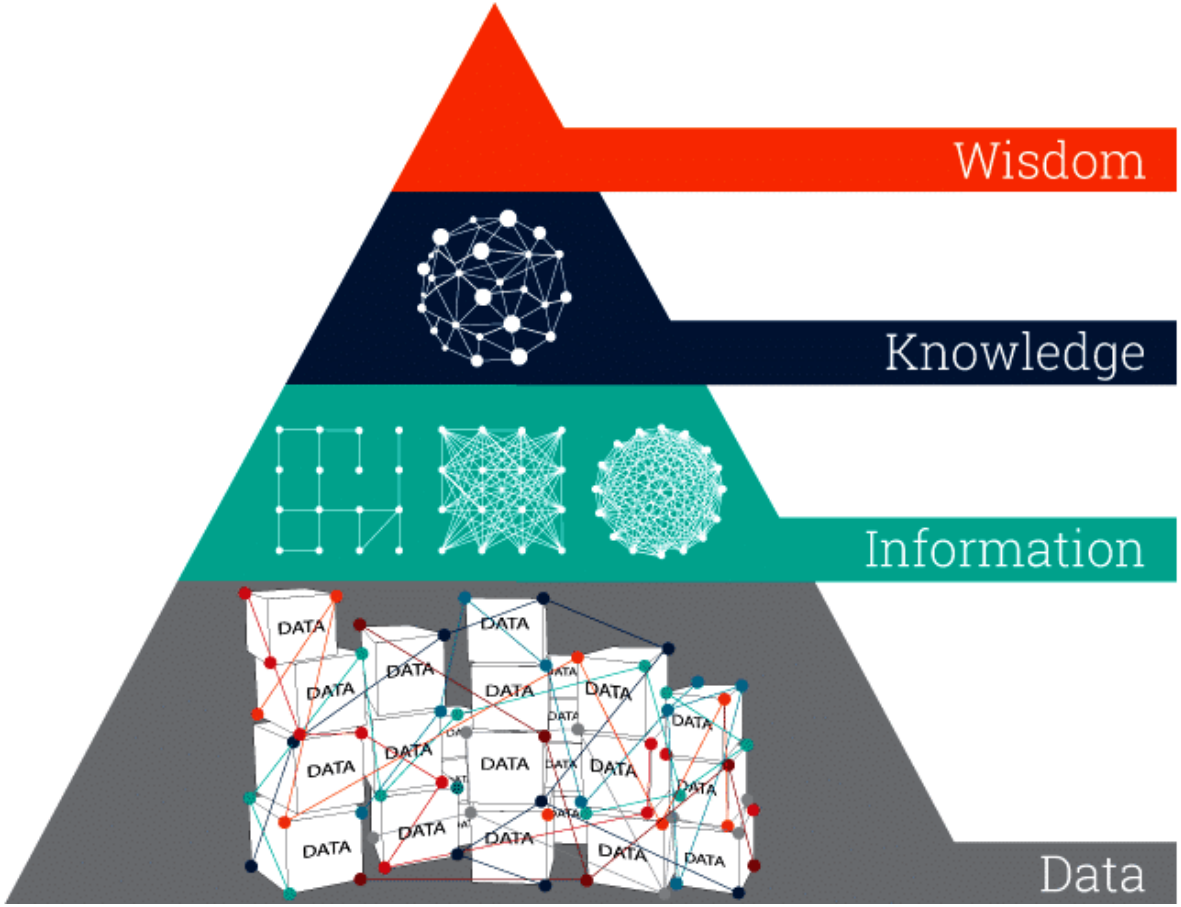

Figure: Pirâmede DIKW - Disponível em: *.

Dados, informação e conhecimento

Figure: Dados no mundo: *.

Dados, informação e conhecimento

Alguns dados (referência de 01/04/2019 - fonte)

Dados, informação e conhecimento

Alguns dados (referência de 01/04/2019 - fonte)

- 3,9 bilhões de pessoas conectadas na internet no mundo (51% da população mundial);

Dados, informação e conhecimento

Alguns dados (referência de 01/04/2019 - fonte)

- 3,9 bilhões de pessoas conectadas na internet no mundo (51% da população mundial);
- O Google realiza 3,8 milhões de buscas por minutos;

Dados, informação e conhecimento

Alguns dados (referência de 01/04/2019 - fonte)

- 3,9 bilhões de pessoas conectadas na internet no mundo (51% da população mundial);
- O Google realiza 3,8 milhões de buscas por minutos;
- 4,5 milhões de vídeos carregados no Youtube em um minuto;

Dados, informação e conhecimento

Alguns dados (referência de 01/04/2019 - fonte)

- 3,9 bilhões de pessoas conectadas na internet no mundo (51% da população mundial);
- O Google realiza 3,8 milhões de buscas por minutos;
- 4,5 milhões de vídeos carregados no Youtube em um minuto;
- 41,6 milhões de mensagens enviadas no Facebook Messenger e no WhatsApp em um minuto;

Dados, informação e conhecimento

Alguns dados (referência de 01/04/2019 - fonte)

- 3,9 bilhões de pessoas conectadas na internet no mundo (51% da população mundial);
- O Google realiza 3,8 milhões de buscas por minutos;
- 4,5 milhões de vídeos carregados no Youtube em um minuto;
- 41,6 milhões de mensagens enviadas no Facebook Messenger e no WhatsApp em um minuto;
- 87 mil tweets por minuto;

Dados, informação e conhecimento

Alguns dados (referência de 01/04/2019 - fonte)

- 3,9 bilhões de pessoas conectadas na internet no mundo (51% da população mundial);
- O Google realiza 3,8 milhões de buscas por minutos;
- 4,5 milhões de vídeos carregados no Youtube em um minuto;
- 41,6 milhões de mensagens enviadas no Facebook Messenger e no WhatsApp em um minuto;
- 87 mil tweets por minuto;
- O volume de dados criado nos últimos dois anos é maior do que a quantidade produzida em toda a história.

Dados, informação e conhecimento

Alguns dados (dezembro de 2018) - fonte:

Dados, informação e conhecimento

Alguns dados (dezembro de 2018) - fonte:

- 44 Zettabytes (ou 44 trilhões de Gigabytes) já este ano;

Dados, informação e conhecimento

Alguns dados (dezembro de 2018) - fonte:

- 44 Zettabytes (ou 44 trilhões de Gigabytes) já este ano;
- Com o 5G, viveremos no limite?;

Dados, informação e conhecimento

Alguns dados (dezembro de 2018) - fonte:

- 44 Zettabytes (ou 44 trilhões de Gigabytes) já este ano;
- Com o 5G, viveremos no limite?;
- Estratégia *multi-cloud*;

Dados, informação e conhecimento

Alguns dados (dezembro de 2018) - fonte:

- 44 Zettabytes (ou 44 trilhões de Gigabytes) já este ano;
- Com o 5G, viveremos no limite?;
- Estratégia *multi-cloud*;
- Geração Z já se inserindo no mercado de trabalho;

Dados, informação e conhecimento

Alguns dados (dezembro de 2018) - fonte:

- 44 Zettabytes (ou 44 trilhões de Gigabytes) já este ano;
- Com o 5G, viveremos no limite?;
- Estratégia *multi-cloud*;
- Geração Z já se inserindo no mercado de trabalho;
- Maior reestruturação da rede de suprimentos;

Dados, informação e conhecimento

Alguns dados (dezembro de 2018) - fonte:

- 44 Zettabytes (ou 44 trilhões de Gigabytes) já este ano;
- Com o 5G, viveremos no limite?;
- Estratégia *multi-cloud*;
- Geração Z já se inserindo no mercado de trabalho;
- Maior reestruturação da rede de suprimentos;
- *Internet of Things* (*IoT*);

Dados, informação e conhecimento

Alguns dados (dezembro de 2018) - fonte:

- 44 Zettabytes (ou 44 trilhões de Gigabytes) já este ano;
- Com o 5G, viveremos no limite?;
- Estratégia *multi-cloud*;
- Geração Z já se inserindo no mercado de trabalho;
- Maior reestruturação da rede de suprimentos;
- *Internet of Things* (*IoT*);
- etc... .

Dados, informação e conhecimento

Dados:

Dados, informação e conhecimento

Dados:

- Dado é o registro do atributo de um ente, objeto ou fenômeno [1];

Dados, informação e conhecimento

Dados:

- Dado é o registro do atributo de um ente, objeto ou fenômeno [1];
- Dados brutos - unidade básica de valor;

Dados, informação e conhecimento

Dados:

- Dado é o registro do atributo de um ente, objeto ou fenômeno [1];
- Dados brutos - unidade básica de valor;
- Dados estruturados e não estruturados;

Dados, informação e conhecimento

Dados:

- Dado é o registro do atributo de um ente, objeto ou fenômeno [1];
- Dados brutos - unidade básica de valor;
- Dados estruturados e não estruturados;
- Facilmente obtidos por máquinas;

Dados, informação e conhecimento

Dados:

- Dado é o registro do atributo de um ente, objeto ou fenômeno [1];
- Dados brutos - unidade básica de valor;
- Dados estruturados e não estruturados;
- Facilmente obtidos por máquinas;
- Frequentemente quantificados;

Dados, informação e conhecimento

Dados:

- Dado é o registro do atributo de um ente, objeto ou fenômeno [1];
- Dados brutos - unidade básica de valor;
- Dados estruturados e não estruturados;
- Facilmente obtidos por máquinas;
- Frequentemente quantificados;
- Facilmente transferíveis.

Dados, informação e conhecimento

Informação:

Dados, informação e conhecimento

Informação:

- Dados organizados que possuem algum sentido;

Dados, informação e conhecimento

Informação:

- Dados organizados que possuem algum sentido;
- Dados dotados de relevância e propósito;

Dados, informação e conhecimento

Informação:

- Dados organizados que possuem algum sentido;
- Dados dotados de relevância e propósito;
- Exige consenso em relação ao significado;

Dados, informação e conhecimento

Informação:

- Dados organizados que possuem algum sentido;
- Dados dotados de relevância e propósito;
- Exige consenso em relação ao significado;
- Exige necessariamente a mediação humana;

Dados, informação e conhecimento

Informação:

- Dados organizados que possuem algum sentido;
- Dados dotados de relevância e propósito;
- Exige consenso em relação ao significado;
- Exige necessariamente a mediação humana;
- Requer unidade de análise.

Dados, informação e conhecimento

Conhecimento:

- Várias informações organizadas de forma lógica;

Dados, informação e conhecimento

Conhecimento:

- Várias informações organizadas de forma lógica;
- Produto de reflexão e síntese;

Dados, informação e conhecimento

Conhecimento:

- Várias informações organizadas de forma lógica;
- Produto de reflexão e síntese;
- De difícil estruturação;

Dados, informação e conhecimento

Conhecimento:

- Várias informações organizadas de forma lógica;
- Produto de reflexão e síntese;
- De difícil estruturação;
- De difícil captura em máquinas;

Dados, informação e conhecimento

Conhecimento:

- Várias informações organizadas de forma lógica;
- Produto de reflexão e síntese;
- De difícil estruturação;
- De difícil captura em máquinas;
- Geralmente tácito;

Dados, informação e conhecimento

Conhecimento:

- Várias informações organizadas de forma lógica;
- Produto de reflexão e síntese;
- De difícil estruturação;
- De difícil captura em máquinas;
- Geralmente tácito;
- De difícil transferência.

Data Science

Data Science

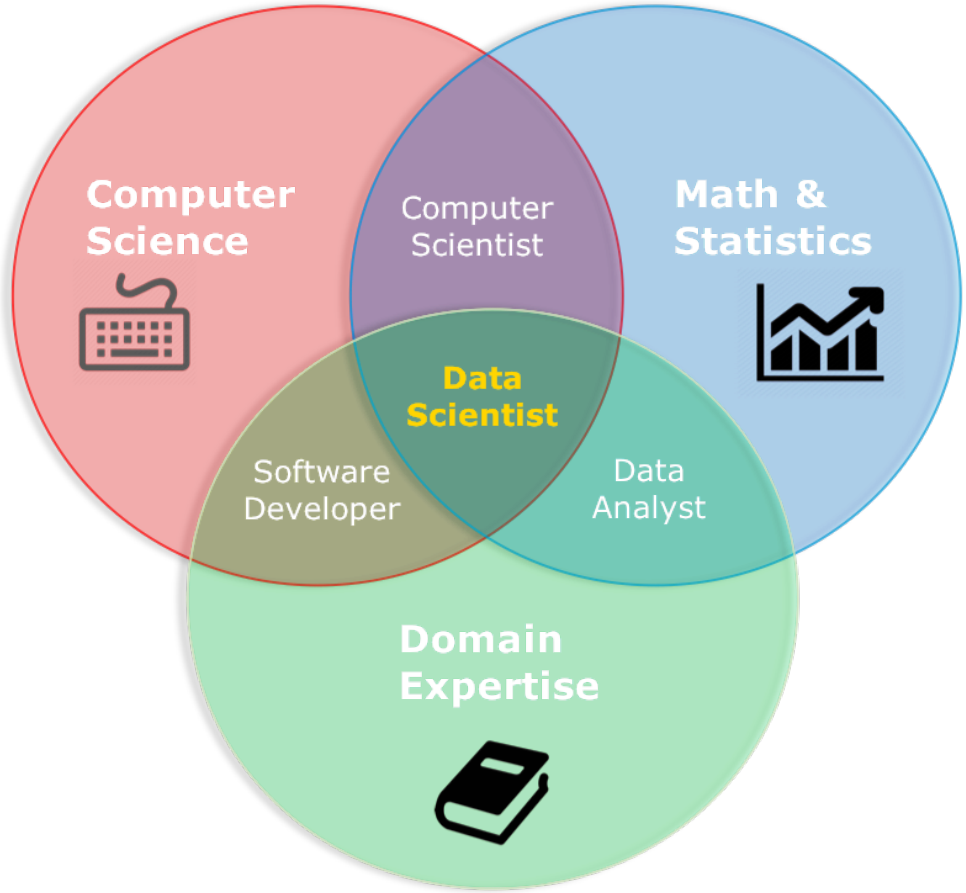

Figure: O que é Data Science? - Disponível em:*

Data Science

Um cientista de dados seria alguém que sabe mais sobre estatística do que um cientista da computação e mais sobre computação do que um estatístico?

Data Science

Um cientista de dados seria alguém que sabe mais sobre estatística do que um cientista da computação e mais sobre computação do que um estatístico?

Figure: Data scientist? - Disponível em:*.

Data Science

Podemos entender **Data Science** como sendo uma área que estuda como extrair conhecimento dos dados.

Data Science

Podemos entender **Data Science** como sendo uma área que estuda como extrair conhecimento dos dados.

Data science é a coleta de dados de diversas fontes para analisar e subsidiar a tomada de decisões, de forma preditiva, em grandes quantidades e gerando *insights*.

Figure: Data Science - *insights* - Disponível em:*.

Data Science

Data Science

- Área multidisciplinar (*Unicorn Data Scientist*);

Data Science

- Área multidisciplinar (*Unicorn Data Scientist*);
- Múltiplas áreas de aplicação;

Data Science

- Área multidisciplinar (*Unicorn Data Scientist*);
- Múltiplas áreas de aplicação;
- Banco de Dados;

Data Science

- Área multidisciplinar (*Unicorn Data Scientist*);
- Múltiplas áreas de aplicação;
- Banco de Dados;
- Computação de Alto Desempenho;

Data Science

- Área multidisciplinar (*Unicorn Data Scientist*);
- Múltiplas áreas de aplicação;
- Banco de Dados;
- Computação de Alto Desempenho;
- Machine Learning;

Data Science

- Área multidisciplinar (*Unicorn Data Scientist*);
- Múltiplas áreas de aplicação;
- Banco de Dados;
- Computação de Alto Desempenho;
- Machine Learning;
- Estatística;

Data Science

- Área multidisciplinar (*Unicorn Data Scientist*);
- Múltiplas áreas de aplicação;
- Banco de Dados;
- Computação de Alto Desempenho;
- Machine Learning;
- Estatística;
- etc ... .

Data Science

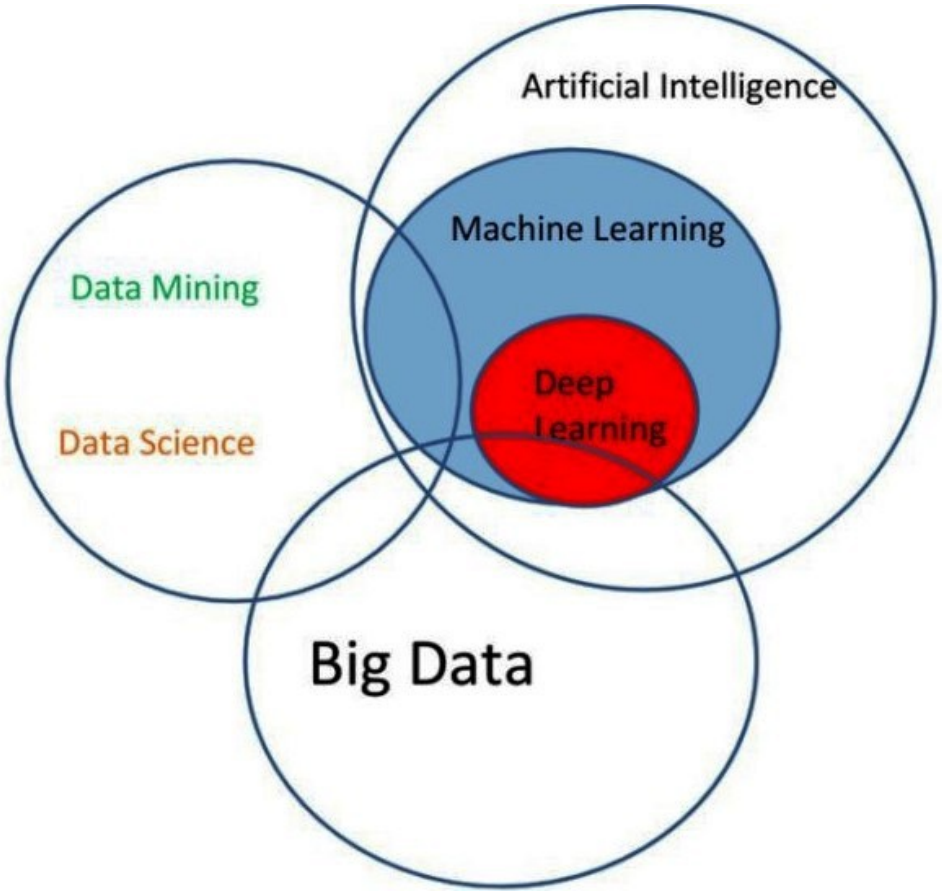

Figure: Áreas Data Science - Disponível em:*.

Big Data

Big Data

- Volume - trata-se de volumes imensos de dados e voláteis;

Big Data

- Volume - trata-se de volumes imensos de dados e voláteis;

- Velocidade - dados coletados bem próximos do momento em acontecem;

Big Data

- Volume - trata-se de volumes imensos de dados e voláteis;

- Velocidade - dados coletados bem próximos do momento em acontecem;

- Variedade - os dados não estão necessariamente preparados e podem ser obtidos de fontes completamente distintas;

Big Data

- Volume - trata-se de volumes imensos de dados e voláteis;

- Velocidade - dados coletados bem próximos do momento em acontecem;

- Variedade - os dados não estão necessariamente preparados e podem ser obtidos de fontes completamente distintas;

- Veracidade - diz respeito sobre as fontes e a qualidade dos dados, pois eles devem ser confiáveis;

Big Data

- Volume - trata-se de volumes imensos de dados e voláteis;

- Velocidade - dados coletados bem próximos do momento em acontecem;

- Variedade - os dados não estão necessariamente preparados e podem ser obtidos de fontes completamente distintas;

- Veracidade - diz respeito sobre as fontes e a qualidade dos dados, pois eles devem ser confiáveis;

- Valor - diz respeito aos benefícios que as soluções de Big Data vão trazer para uma empresa, por exemplo.

Business Intelligence e Data Analytics

Business Intelligence (BI): analisa fatos que já tenha ocorrido em um determinado momento se fundamentando em dados exatos que já existam, não se importando tanto quanto em Data Science em realizar predições em prazos longínquos.

Business Intelligence e Data Analytics

Business Intelligence (BI): analisa fatos que já tenha ocorrido em um determinado momento se fundamentando em dados exatos que já existam, não se importando tanto quanto em Data Science em realizar predições em prazos longínquos.

Data Analytics: É o processo pelo qual se procura inspecionar, limpar, transformar e modelar dados.

Data Mining

Data Mining

Serve para descobrirmos padrões em grandes conjuntos de dados, gerando conhecimento de alto valor competitivo.

Data Mining

Serve para descobrirmos padrões em grandes conjuntos de dados, gerando conhecimento de alto valor competitivo.

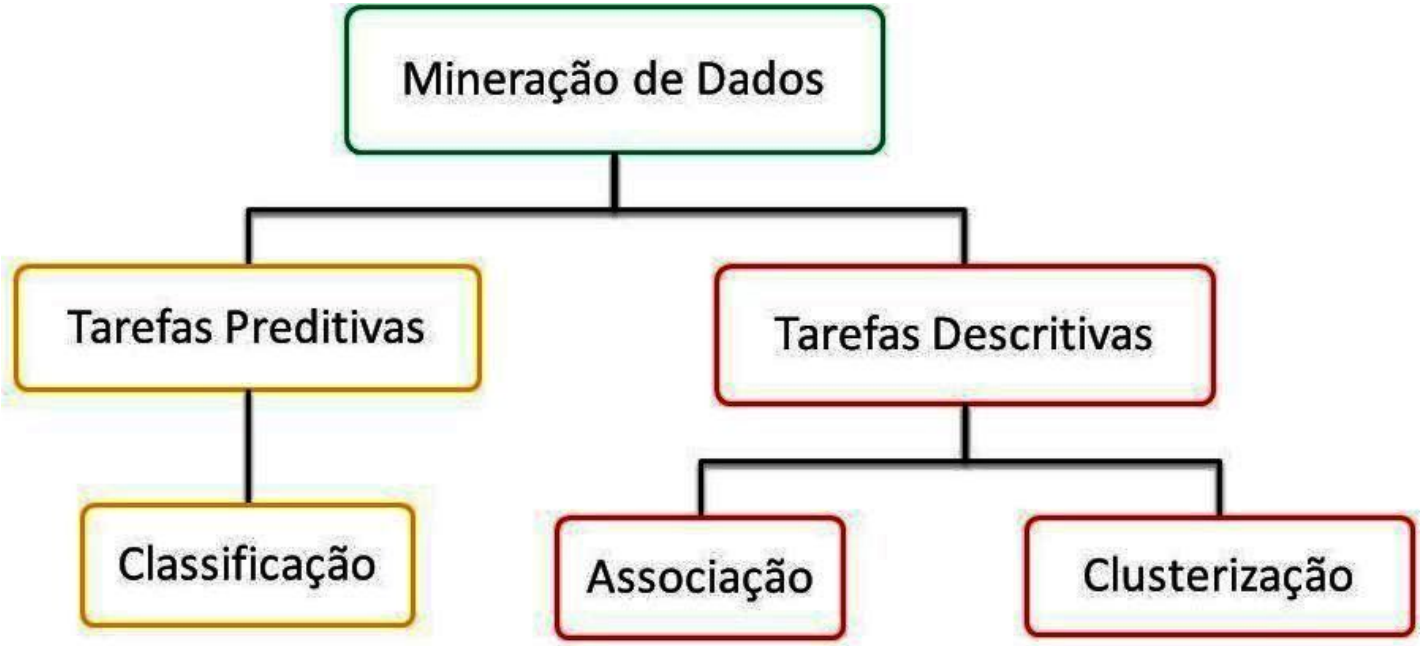

Figure: O que é Data Mining?

Machine Learning

Machine Learning

- Subárea (subcategoria) da Inteligência Artificial;

Machine Learning

- Subárea (subcategoria) da Inteligência Artificial;

- Estuda a construção de algoritmos que podem aprender de seus erros e fazer previsões sobre dados;

Machine Learning

- Subárea (subcategoria) da Inteligência Artificial;

- Estuda a construção de algoritmos que podem aprender de seus erros e fazer previsões sobre dados;

- Aprendizado supervisionado;

Machine Learning

- Subárea (subcategoria) da Inteligência Artificial;

- Estuda a construção de algoritmos que podem aprender de seus erros e fazer previsões sobre dados;

- Aprendizado supervisionado;

- Aprendizado não supervisionado;

Machine Learning

- Subárea (subcategoria) da Inteligência Artificial;

- Estuda a construção de algoritmos que podem aprender de seus erros e fazer previsões sobre dados;

- Aprendizado supervisionado;

- Aprendizado não supervisionado;

- Utilizado em problemas de classificação, regressão, agrupamento e regras de associação.

Machine Learning

Alguns principais algoritmos:

Machine Learning

Alguns principais algoritmos:

- Árvores de Decisão;

Machine Learning

Alguns principais algoritmos:

- Árvores de Decisão;
- Regressão linear;

Machine Learning

Alguns principais algoritmos:

- Árvores de Decisão;
- Regressão linear;
- Regressão logística;

Machine Learning

Alguns principais algoritmos:

- Árvores de Decisão;
- Regressão linear;
- Regressão logística;
- SVM (*Support Vector Machine*);

Machine Learning

Alguns principais algoritmos:

- Árvores de Decisão;
- Regressão linear;
- Regressão logística;
- SVM (*Support Vector Machine*);
- Naïve Bayes;

Machine Learning

Alguns principais algoritmos:

- Árvores de Decisão;
- Regressão linear;
- Regressão logística;
- SVM (*Support Vector Machine*);
- Naïve Bayes;
- etc... .

Machine Learning

Exemplos de aplicação:

Machine Learning

Exemplos de aplicação:

- Detecção de fraudes;

Machine Learning

Exemplos de aplicação:

- Detecção de fraudes;
- Previsão de falhas em equipamentos;

Machine Learning

Exemplos de aplicação:

- Detecção de fraudes;
- Previsão de falhas em equipamentos;
- Reconhecimento de determinados padrões e imagens;

Machine Learning

Exemplos de aplicação:

- Detecção de fraudes;
- Previsão de falhas em equipamentos;
- Reconhecimento de determinados padrões e imagens;
- Filtragem de spams em e-mail;

Machine Learning

Exemplos de aplicação:

- Detecção de fraudes;
- Previsão de falhas em equipamentos;
- Reconhecimento de determinados padrões e imagens;
- Filtragem de spams em e-mail;
- Anúncios em tempo real, tanto em páginas da *web* como em dispositivos móveis;

Machine Learning

Exemplos de aplicação:

- Detecção de fraudes;
- Previsão de falhas em equipamentos;
- Reconhecimento de determinados padrões e imagens;
- Filtragem de spams em e-mail;
- Anúncios em tempo real, tanto em páginas da *web* como em dispositivos móveis;
- etc... .

Deep Learning

Deep Learning

Uma subcategoria de aprendizado de máquina que diz respeito a oportunidades de aprendizagem profundas com o uso de redes neurais para melhorar resultados, tais como reconhecimento de fala, visão computacional e processamento de linguagem natural.

Deep Learning

Uma subcategoria de aprendizado de máquina que diz respeito a oportunidades de aprendizagem profundas com o uso de redes neurais para melhorar resultados, tais como reconhecimento de fala, visão computacional e processamento de linguagem natural.

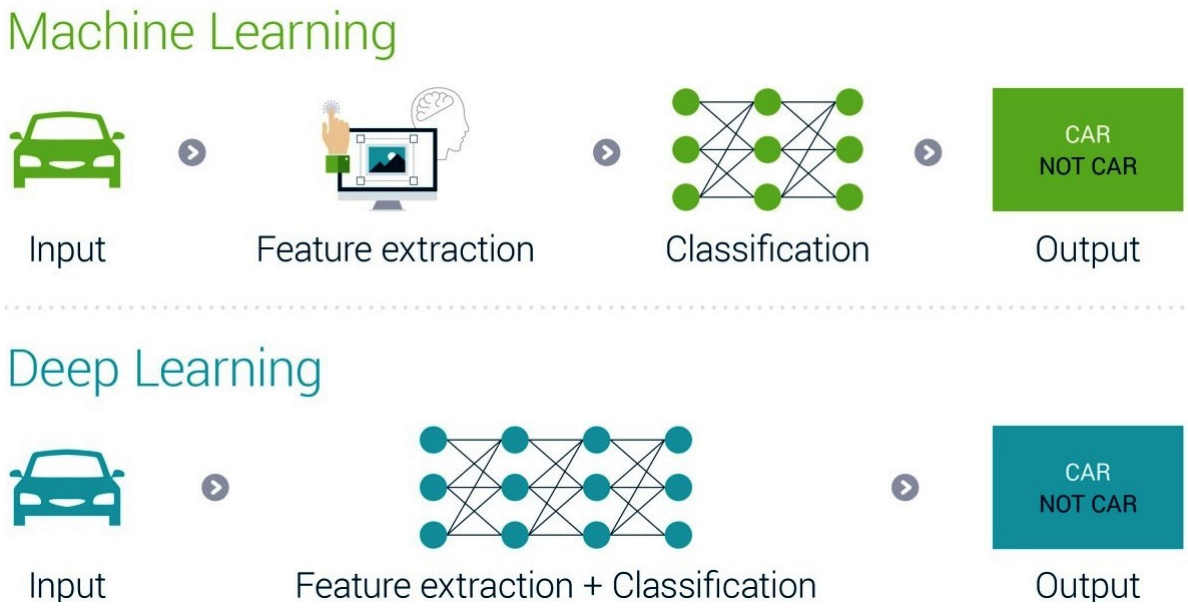

Figure: Deep Learning versus Machine Learning.

Deep Learning

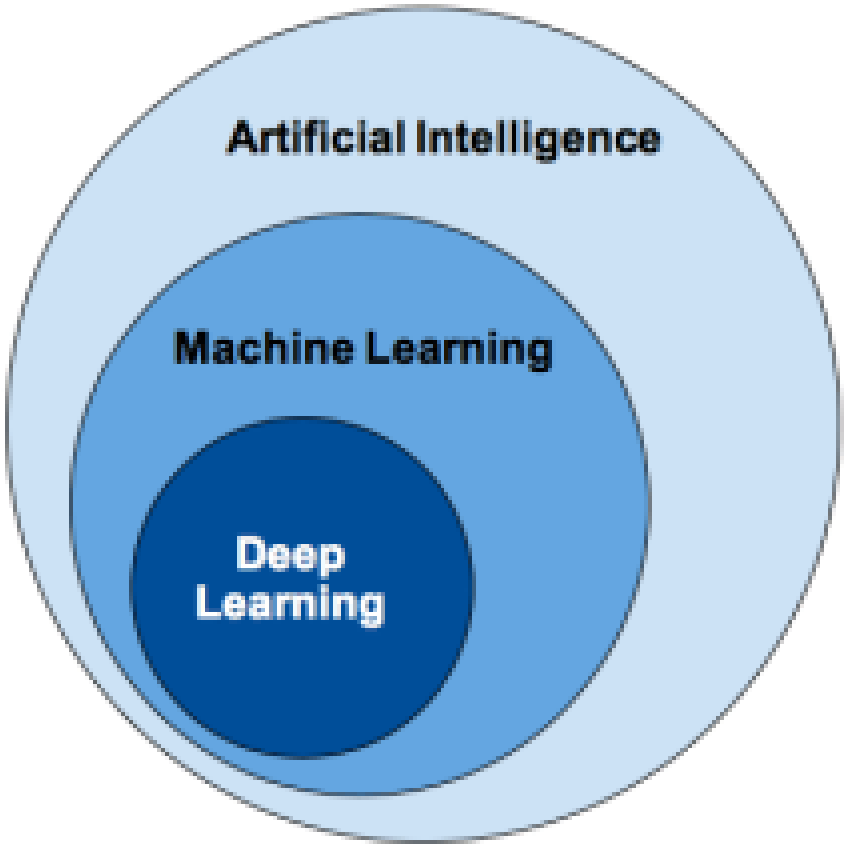

Figure: ML, DL & IA.

Estatística para DS

Estatística para DS

Estatística e DS são a mesma?

Estatística para DS

Estatística e DS são a mesma?

Figure: Estatística e DS.

Estatística para DS

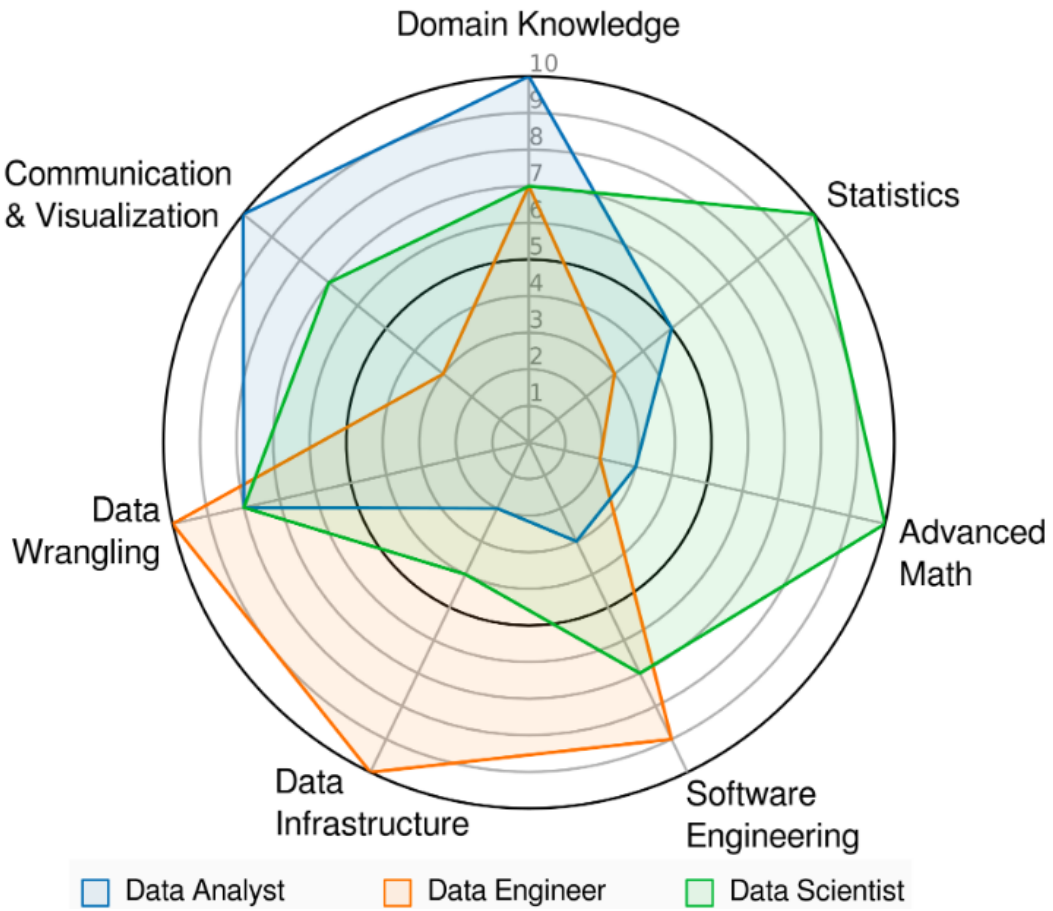

Figure: Estatística em DS.

Estatística para DS

Estatística para DS

- Estatística Descritiva (AED);

Estatística para DS

- Estatística Descritiva (AED);
- Probabilidade;

Estatística para DS

- Estatística Descritiva (AED);
- Probabilidade;
- Estatística Inferencial;

Estatística para DS

- Estatística Descritiva (AED);
- Probabilidade;
- Estatística Inferencial;
- *Data Analytics* (*Data Wrangling*);

Estatística para DS

- Estatística Descritiva (AED);
- Probabilidade;
- Estatística Inferencial;
- *Data Analytics* (*Data Wrangling*);
- Estatística Bayesiana;

Estatística para DS

- Estatística Descritiva (AED);
- Probabilidade;
- Estatística Inferencial;
- *Data Analytics* (*Data Wrangling*);
- Estatística Bayesiana;
- etc... .

Estatística para DS

Estatística para DS

Em outubro de 2015, a *American Statistical Association* (ASA) divulgou uma declaração sobre o papel da Estatística na Ciência de Dados, no qual o presidente **David Morganstein** falou:

Estatística para DS

Em outubro de 2015, a *American Statistical Association* (ASA) divulgou uma declaração sobre o papel da Estatística na Ciência de Dados, no qual o presidente **David Morganstein** falou:

"Através desta declaração, a ASA e seus membros reconhecem que a ciência dados abrange mais do que estatísticas, mas ao mesmo tempo também reconhece que a ciência estatística desempenha um papel fundamental no rápido crescimento deste campo. É nossa esperança que esta declaração possa reforçar a relação de estatísticas para a ciência de dados e ainda fomentar relacionamentos mútuos de colaboração entre todos os contribuintes na ciência de dados" [6].

Estatística para DS

Estatística é fundamental para DS.

Estatística para DS

Estatística é fundamental para DS.

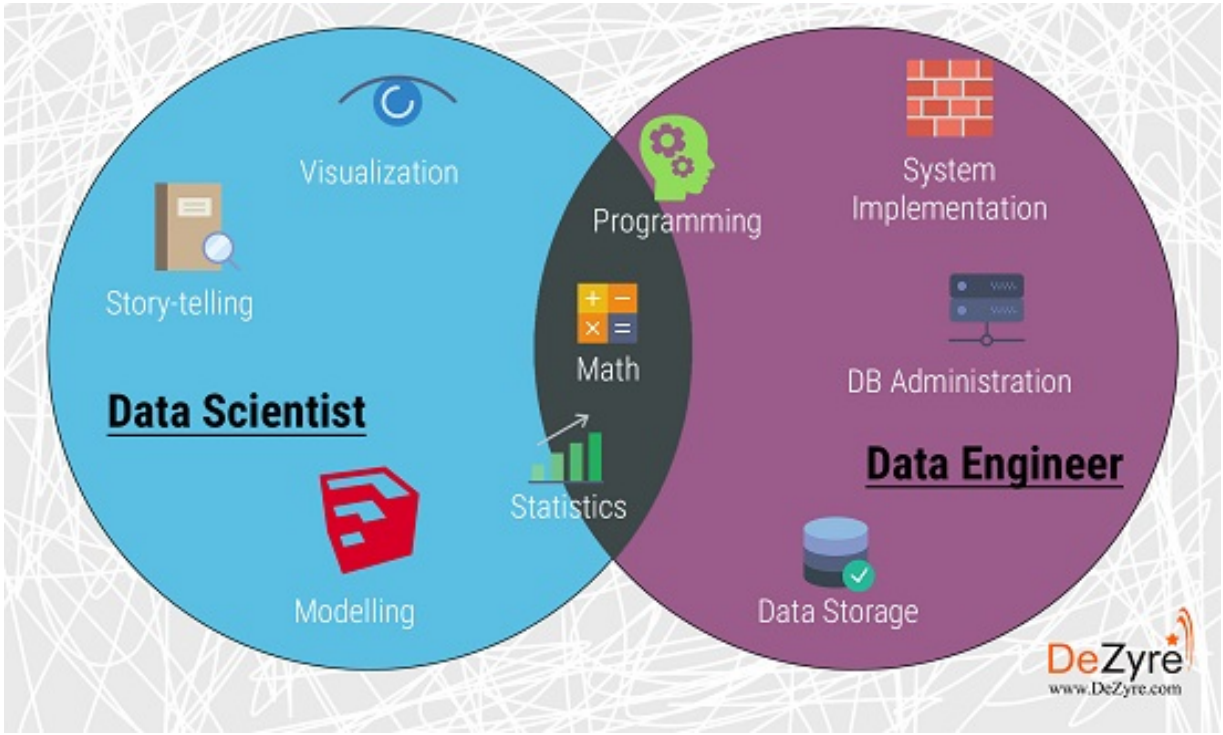

Figure: Importância da Estatística em DS.

Estatística e ML

Estatística e ML

Técnica	Aprendizado de Máquina	Estatística
Manipulação de dados	Trabalha com big data na forma de redes e gráficos; dados brutos de sensores ou texto da web são divididos em dados de treinamento e teste.	Modelos são usados para criar poder preditivo em pequenas amostras.
Entrada de dados	Os dados são amostrados, randomizados e transformados para maximizar a pontuação da precisão na predição de exemplos fora da amostra (ou completamente novos).	Parâmetros interpretam fenômenos do mundo real e fornecem uma ênfase em relação à magnitude.
Resultado	A probabilidade é levada em conta para comparar qual poderia ser a melhor conjectura ou decisão.	A saída captura a variabilidade e a incerteza dos parâmetros.
Suposições	O cientista aprende a partir dos dados.	O cientista presume certa saída e tenta prová-la.
Distribuição	A distribuição é desconhecida ou ignorada antes de aprender dos dados.	O cientista presume uma distribuição bem definida.
Ajuste	O cientista cria o modelo mais adequado, mas generalizável.	O resultado é adequado à distribuição de dados presente.

Figure: Comparação entre Estatística e ML.

Estatística e ML

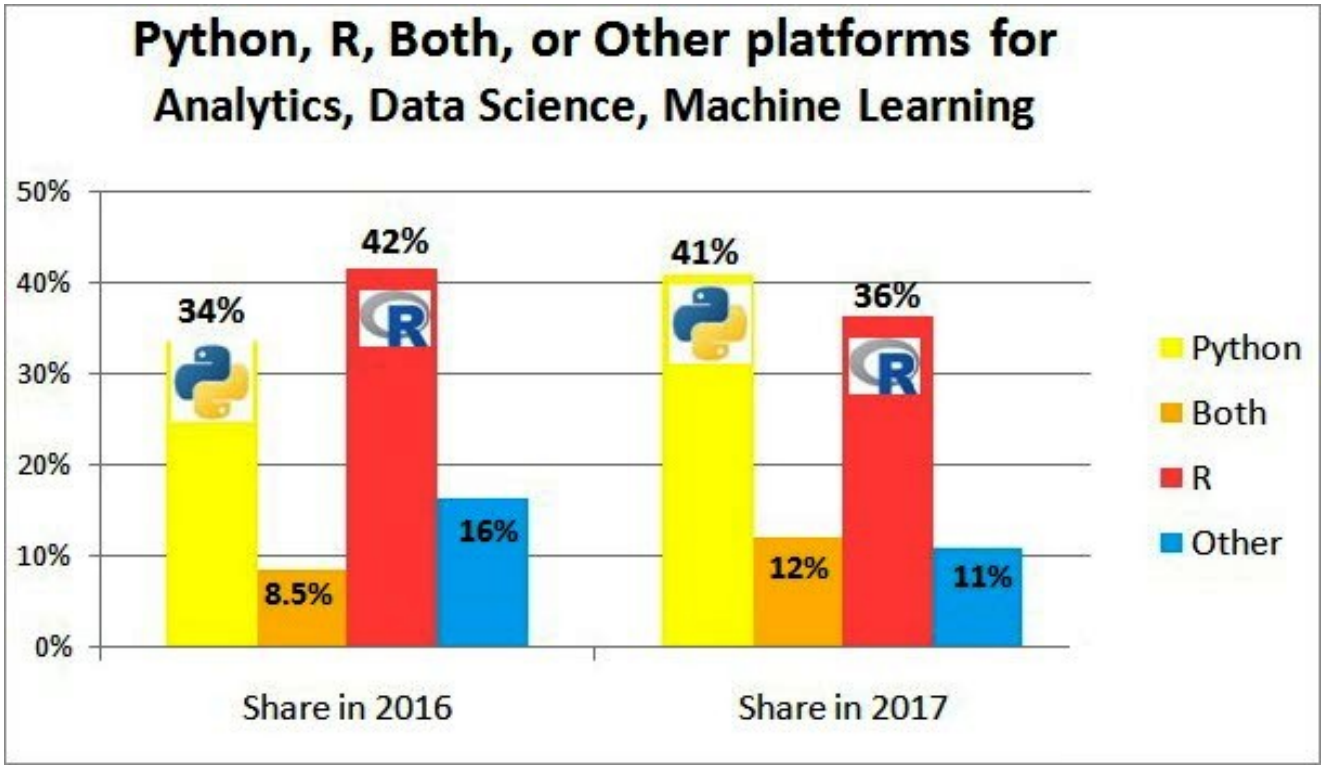

Figure: Principais *softwares* em Estatística e ML - https://www.softwaretestinghelp.com/data-science-tools/

Considerações Finais

Considerações Finais

- Era Big Data;

Considerações Finais

- Era Big Data;
- Novos desafios;

Considerações Finais

- Era Big Data;
- Novos desafios;
- Data Science ainda está evoluindo;

Considerações Finais

- Era Big Data;
- Novos desafios;
- Data Science ainda está evoluindo;
- Profissões - cientistas de dados (não é um emprego da moda);

Considerações Finais

- Era Big Data;
- Novos desafios;
- Data Science ainda está evoluindo;
- Profissões - cientistas de dados (não é um emprego da moda);
- etc... .

Referências

1) Gomes; Pimenta; Schneider. Data Mining in Information Science Research: Challenges And Opportunities:
Disponível em: https://zenodo.org/record/3521038#.XjHYMehKgVg.

2) Francesco Corea. An Introduction to Data: Everything You Need to Know About AI, Big Data and Data Science. Springer International Publishing, 2019.

3) John W. Tukey. The Future of Data Analysis. 1962.
Disponível em: https://projecteuclid.org/euclid.aoms/1177704711

4) Mueller, J.P.; Massaron, L. Aprendizado de Máquina para leigos. Alta Books, 2019.

Referências

- **Sites:**
1) https://www.ibm.com/br-pt/analytics/machine-learning
2) https://www.cetax.com.br/blog/machine-learning/
3) https://transformacaodigital.com/o-que-e-machine-learning-e-como-funciona/
4) https://towardsdatascience.com/how-to-become-a-data-scientist-2a02ed565336
5) https://medium.com/beacon-insight/ci%C3%AAncia-de-dados-e-o-cientista-de-dados-72634fcc1a4c

Referências

6) https://community.amstat.org/blogs/ronald-wasserstein/2015/10/01/t
role-of-statistics-in-data-science-an-asa-statement
7) http://www.cienciaedados.com/o-papel-da-estatistica-na-ciencia-de-dados/
8) https://storm.cis.fordham.edu/ gweiss/selected-papers/data-mining-and-statistics-hand.pdf
9) https://arxiv.org/ftp/arxiv/papers/1509/1509.02900.pdf
10) https://projecteuclid.org/euclid.aoms/1177704711

Perguntas???