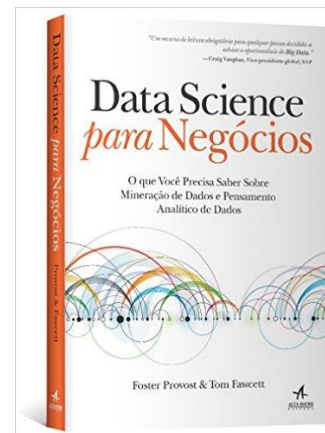

Introdução a Ciência de Dados

Valkson Saraiva

Dados são o novo ouro

A Ciência de Dados é o estudo que tem como objetivo a exploração de dados estruturados e não estruturados para identificar padrões que podem ajudar uma organização no controle de custos, aumento de eficiência, reconhecimento e descoberta de novos mercados, oportunidades e aumento de vantagem competitiva.

Indicação:





A quarta Revolução Industrial teve início neste século — e é a revolução digital. Para Schwab, nesta fase, inúmeras tecnologias, que vão do sequenciamento genético à nanotecnologia, das energias renováveis à inteligência artificial, vão mudar completamente a vida das pessoas num ritmo e com uma abrangência inéditos. “A questão para as empresas não é se elas serão afetadas pela disrupção. E sim quando a disrupção chegar, qual será a forma que terá?”, afirma.



Numa entrevista exclusiva a EXAME, Jeffrey Immelt, presidente do conselho de administração da companhia, conta como a GE passou a investir numa mudança radical de sua estrutura e de sua estratégia (*pág. 56*). A meta é deixar de ser uma empresa que só produz equipamentos industriais e se tornar capaz de fornecer dados aos clientes por meio de sensores conectados às suas máquinas. Longe de ser uma mudança trivial, é algo que está mexendo profundamente nas bases da companhia.



Seguindo a máxima de que a análise de dados é na economia digital o que o aço foi na era industrial, os executivos da Netflix decidiram colocar um caminhão de dinheiro no projeto porque tinham informações das quais os canais de TV não dispunham. Sabiam que o ator Kevin Spacey e o diretor David Fincher, ambos já parte do projeto, eram muito admirados. Os assinantes da Netflix não costumavam fazer buscas por filmes com Spacey, mas, quando viam um, acabavam assistindo a vários.

Tecnologias que recebem mais investimentos
hoje e a projeção para 2020

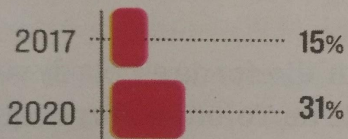
Internet das coisas



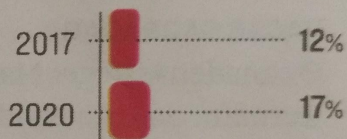
Inteligência artificial



Robótica



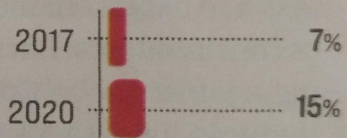
Impressão 3D



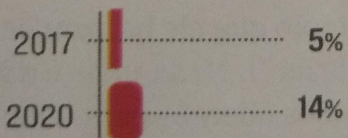
Realidade aumentada



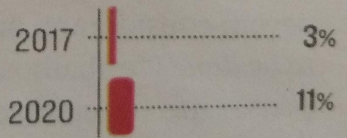
Realidade virtual



Drones



Blockchain



Fonte: PwC (pesquisa com 2 216 entrevistados)

Internet da Coisas e Inteligência Artificial no topo de investimentos

Qual a diferença?

Engenheiro/Arquiteto de Dados

Big Data

Grande volume
de dados

Analista de BI

**Business
Intelligence**

Visão de passado

Cientista de Dados

Data Science

Visão de presente
e futuro

● data science
Termo de pesquisa

● business intelligen...
Termo de pesquisa

● big data
Termo de pesquisa

+ Adicionar comparação

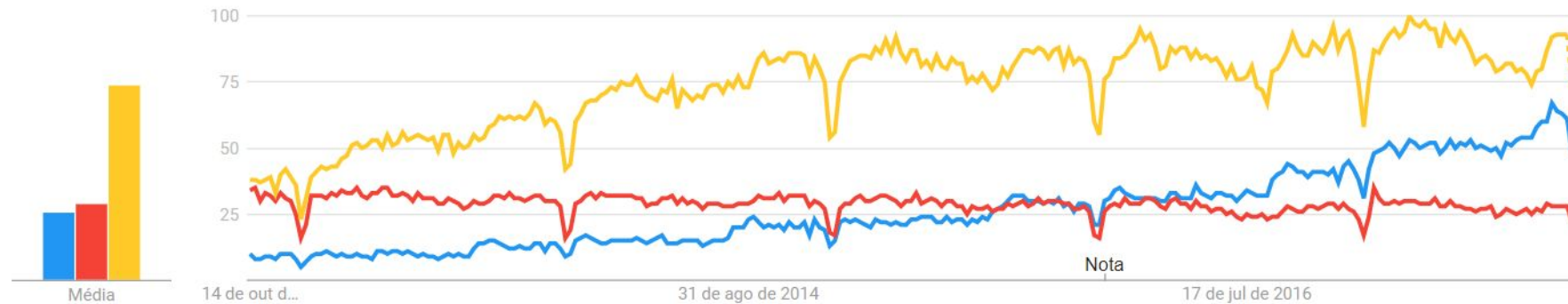
Todo o mundo ▾

Nos últimos 5 anos ▾

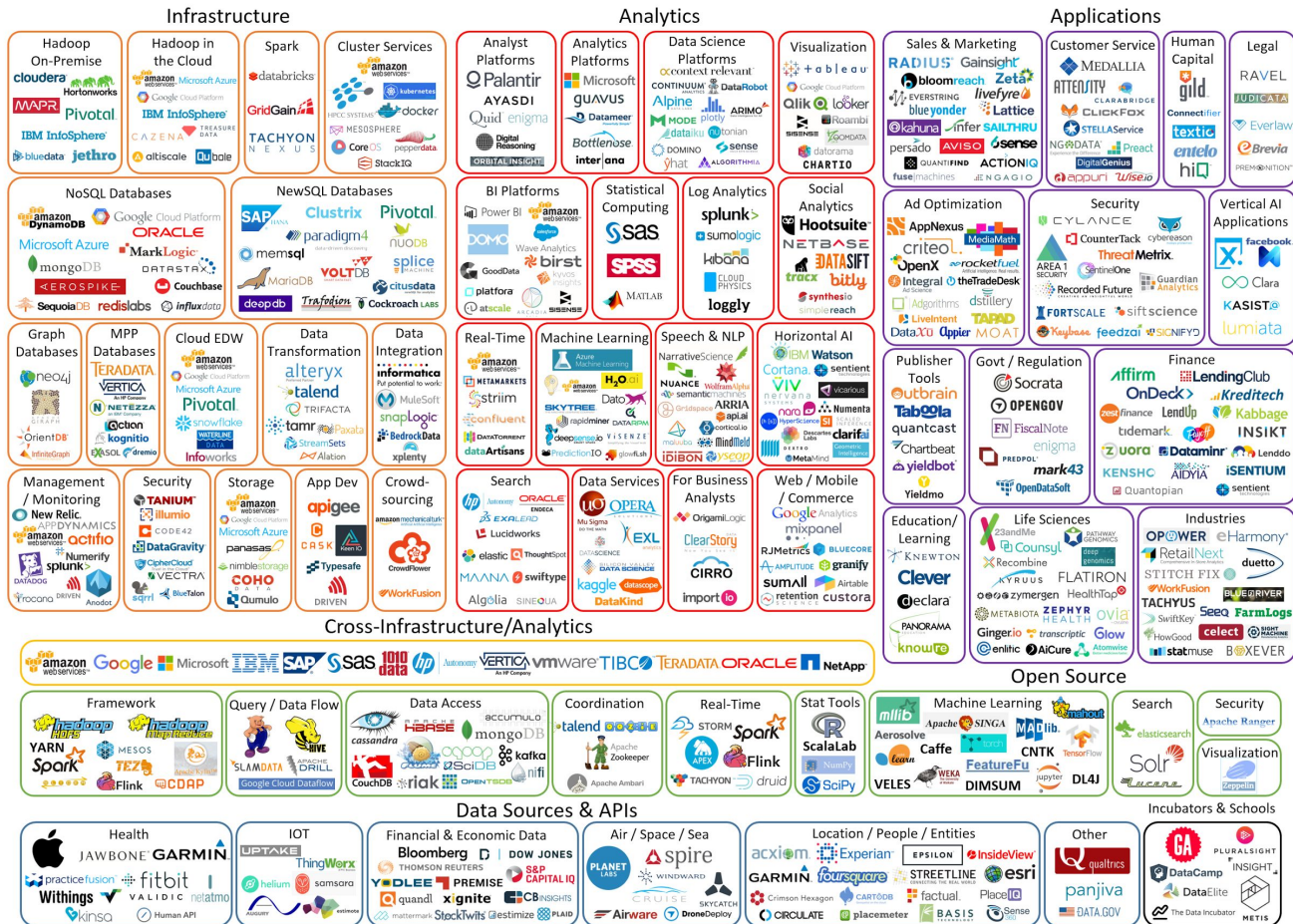
Todas as categorias ▾

Pesquisa na Web ▾

Interesse ao longo do tempo ⓘ



Big Data Landscape 2016 (Version 3.0)



Last Updated 3/23/2016

© Matt Turck (@mattturck), Jim Hao (@jimhao), & FirstMark Capital (@firstmarkcap)

FIRST MARK



Fontes Internas de Dados

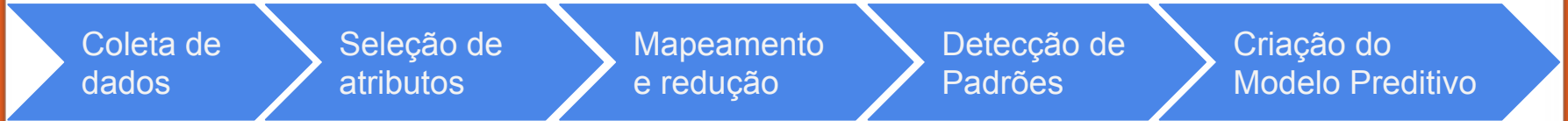
- PRODUTOS DIGITAIS
- CRM
- ERP
- PDV
- SISTEMAS DE PROJETOS
- PROCESSOS
- PLANILHAS
- DOCUMENTOS

- ENEM
- SENSO ESCOLAR
- SITE DE CONCORRENTES
- RECLAME AQUI
- REDES SOCIAIS
- IMAGENS
- TRANSPARÊNCIA

Fontes Externas de Dados



Etapas



Classificação

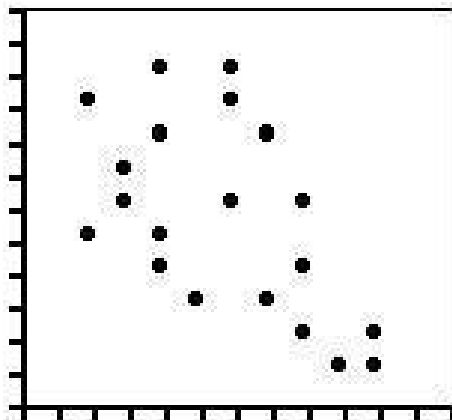
- Sistemas de classificação são usados geralmente quando as previsões são de natureza distinta, ou seja, um simples “sim ou não”. Ex: O cliente irá comprar determinado produto?

Regressão

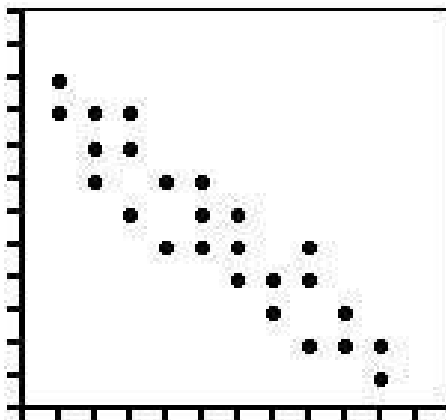
- Sistemas de regressão poderiam ser usados, por exemplo, para responder às perguntas: “Quanto o cliente comprará?” ou “Quando o cliente comprará?”.

Correlação

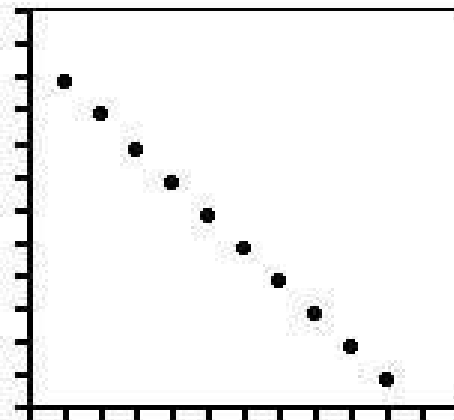
Diagramas de dispersão que mostram correlação negativa entre as variáveis



Correlação fraca



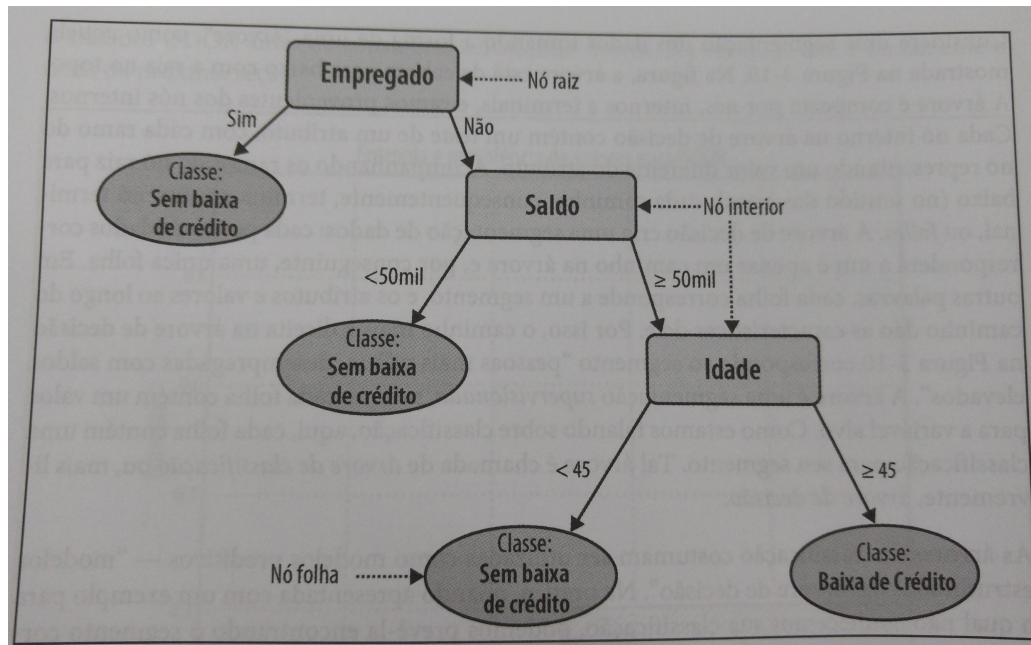
Correlação forte



Correlação perfeita

Técnicas

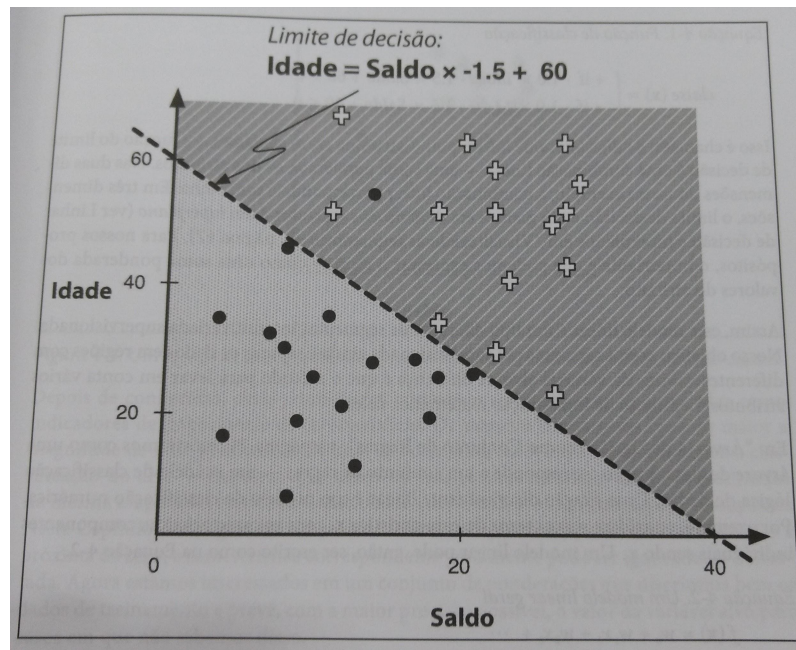
Segmentação Supervisionada - Árvore de Decisão



Fonte: Livro Ciência de Dados para Negócios

Técnicas

Classificação por Funções Matemáticas



Fonte: Livro Ciência de Dados para Negócios



Técnicas

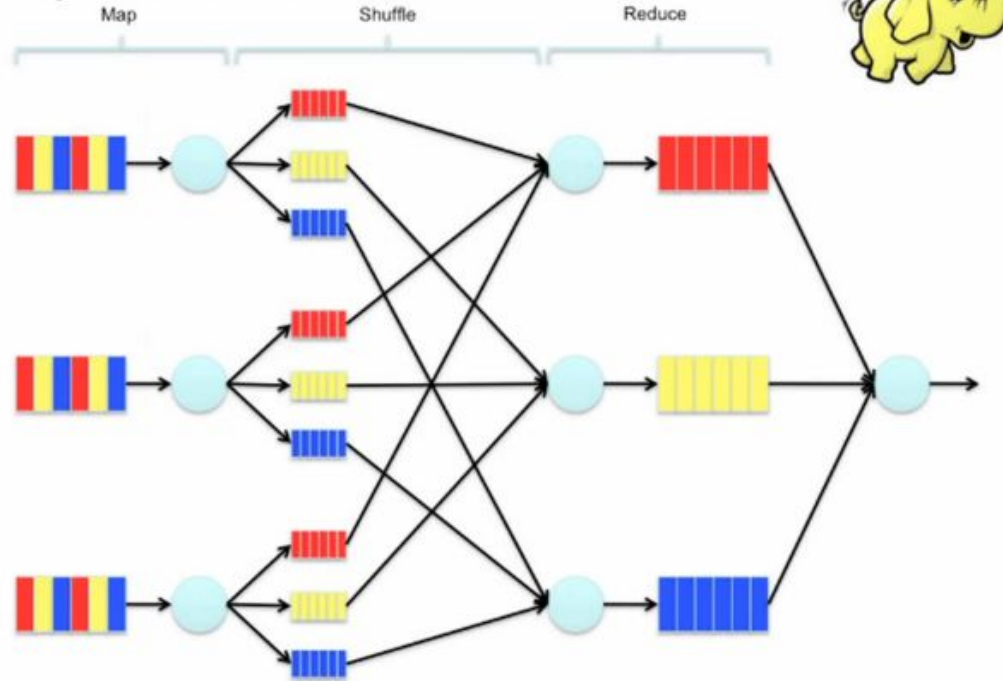
Sobreajuste

Lei dos Grandes Números

- De acordo com a LGN, a média aritmética dos resultados da realização da mesma experiência repetidas vezes tende a se aproximar do valor esperado à medida que mais tentativas se sucederem. Em outras palavras, quanto mais tentativas são realizadas, maior a probabilidade da média aritmética dos resultados observados se aproximar da probabilidade real.

Técnicas

MapReduce Overview



Fonte: Google

Dúvidas?



Obrigado!