



# Introdução à Ciência de Dados

Prof. Dr. Francisco Carlos Souza  
Prof. Dr. Anderson Carniel

# O Surgimento

- Apesar de parecer uma área nova, o termo ciência de dados **surgiu em 1960**
- Mas somente se **popularizou** nas **últimas décadas** devido alguns fatores:
  - **Maior abundância** de dados
  - **Velocidade** em que a informação é **distribuída**
  - **Aumento** de dados **não estruturados** disponíveis
    - Dados que necessitam de pré-processamento para se tornar uma informação ou conhecimento
    - Esse grande volume de dados **não estruturados**, também é conhecido como **Big Data**

# O que é Ciência de Dados

- Para entender com mais precisão essa área, é necessário compreender três conceitos essenciais.

**Dados**



**Informação**



**Big Data**



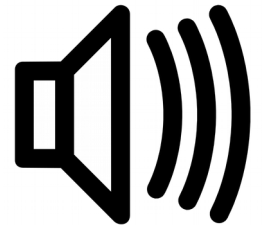
# Dados

- Atualmente os dados são provenientes de diferentes artefatos como, **textos, documentos, áudios, vídeos, imagens, geolocalização** e principalmente mídias sociais que englobam todos eles.

Although data immersion is nothing new, you may have noticed that the phenomenon is accelerating. Lakes, puddles, and rivers of data have turned to floods and veritable tsunamis of structured, semistructured, and unstructured data that's streaming from almost every activity that takes place in both the digital and physical worlds. **Welcome to the world of big data!**

Lillian Pierson

# Dados

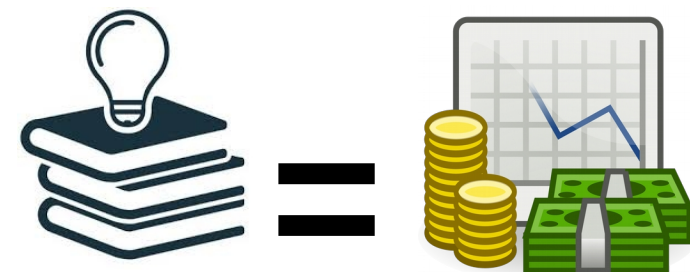


# Big Data

- Big data é um termo que **surgiu na década de 90**
- Ele se refere ao **armazenamento e ao tratamento de dados estruturados e não estruturados**
- O Big data se **baseia em cinco características**, definida como os 5V's
  - Valor, volume, velocidade, variedade e veracidade

# Informações

- Por décadas toda essa **grande** quantidade de **dados** estava sendo **acumulada**
- **Poucos dados** eram **convertidos** em informações para **gerar lucro**
- **Converter** dados em informações para ser utilizadas como **estratégias empresárias** se mostrou uma das **fontes mais lucrativas** nos últimos anos



# Informações




“A partir da análise desses dados é possível provar que o Presidente será investigado por crimes de responsabilidade”



# O que é Ciência de Dados

Data science involves principles, processes, and techniques for understanding phenomena via the (automated) analysis of data.

Foster Provost - Data Science for Bussines



... data science represents the optimization of processes and resources. Data science produces data insights — actionable, data-informed conclusions or predictions that you can use to understand and improve your business, your investments, your health, and even your lifestyle and social life.

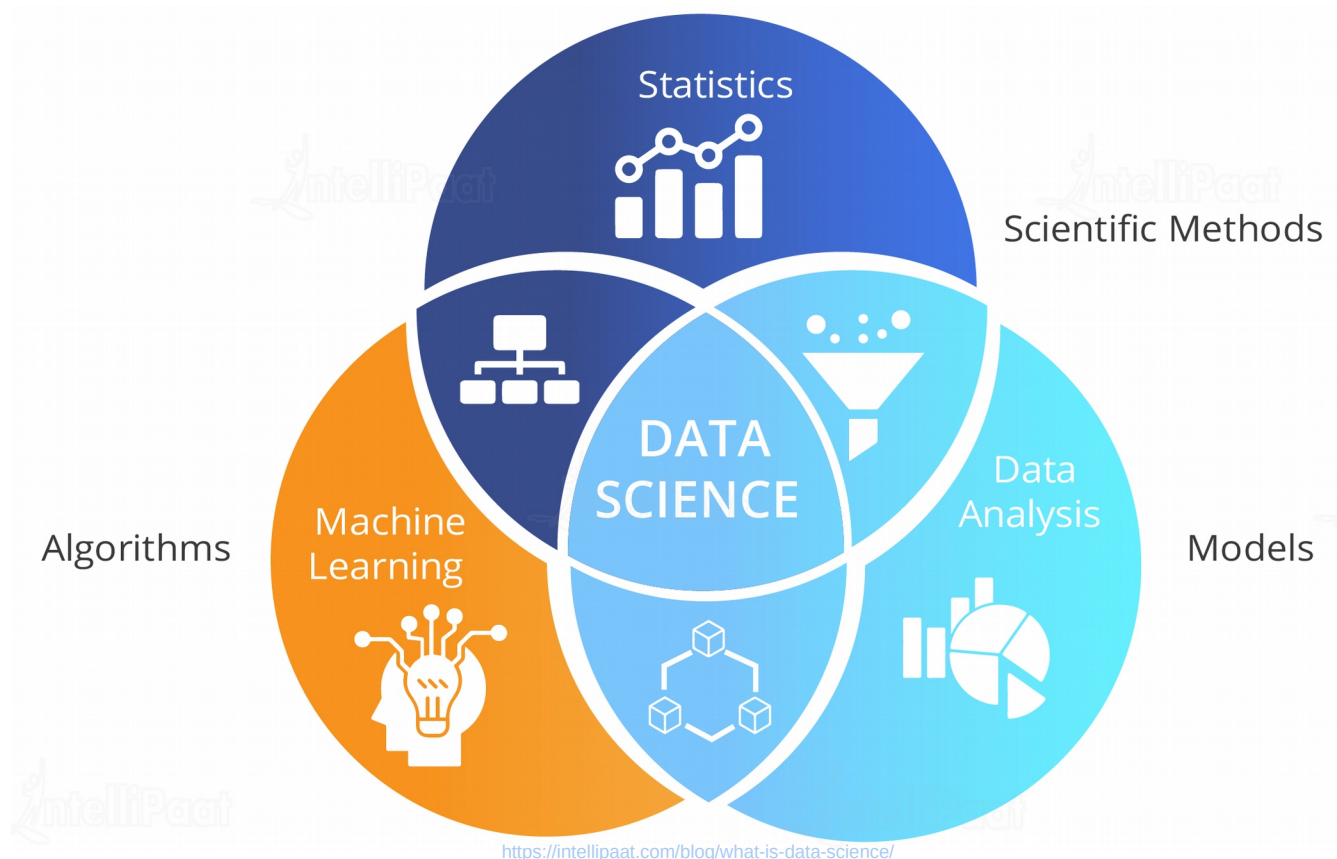
Lillian Pierson – Data Science for Dummies

# O que é Ciência de Dados

- É um domínio **interdisciplinar** que utiliza métodos, processos, técnicas e algoritmos para **extrair informação e conhecimento** de um conjunto dados
- A ciência de dados inclui áreas como:
  - Ciência da computação
    - Algoritmos e Programação
    - Inteligência Artificial
    - Banco de Dados, etc.
  - Matemática
  - Estatística
  - Conhecimento em Negócio

# O que é Ciência de Dados

- Representação da Interdisciplinaridade




# O que é Ciência de Dados

- **Data analysis:** são métodos que auxiliam na análise em big data e traduzir os dados em informações úteis para tomada de decisão.
- Os principais tipos de análises são:
  - **Preditiva:** previsão de cenários futuros
  - **Prescritiva:** auxilia na tomada de medidas
  - **Descritiva:** compreensão eventos em tempo real
  - **Diagnóstica:** compreensão de causas de eventos

# O que é Ciência de Dados

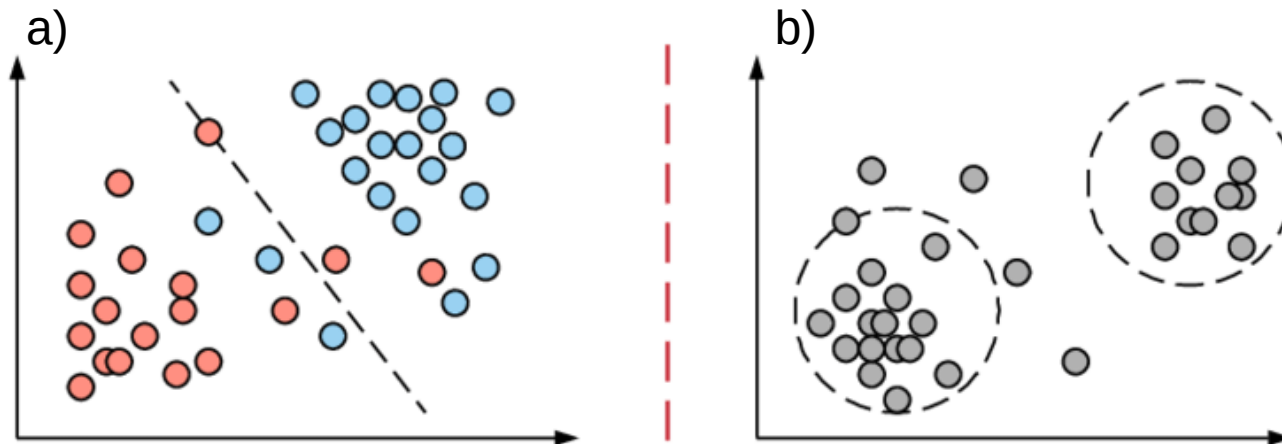
- **Statistics:** usadas para analisar variáveis e conjuntos de dados do nosso cotidiano, como apresentado na tabela abaixo.



Previsão de vendas e lucros
Custo de construções
Níveis de satisfação de clientes
Climas
Resultados de eleições
Número de matrículas
Médias de notas
Taxa de juros
Câmbios

# O que é Ciência de Dados

- **Machine Learning:** são algoritmos e técnicas utilizados para dar **habilidades de aprendizado** para máquinas.
- Esse processo ocorre por meio de **aprendizado por meio de exemplos (a)** ou **reconhecimento de padrões através de similaridades** entre dados (b).



# O que é Ciência de Dados

- Trabalha com três tipos de dados:
  - **Estruturados:** representam **dados** que são **armazenados, processados e manipulados** em sistemas **tradicionais** de bancos de dados relacionais.
  - **Não Estruturados:** descrevem dados que são **produzidos** a partir de **atividades humanas** e não se encaixam em um formato de banco de dados tradicional.
  - **Semi Estruturados:** representa uma estrutura flexível, são estruturados por *tags* que são úteis para criar ordem e hierarquia nos dados.

# Ciência de Dados x Inteligência de Negócios

- É comum confundir seus conceitos em função de suas **similaridades**.
- Ambos trabalham com dados para alcançar os mesmos objetivos.
- A inteligência de negócios consiste em **converter dados brutos em *insights*** de negócios para **auxílio à tomada decisões**.



# Ciência de Dados x Inteligência de Negócios

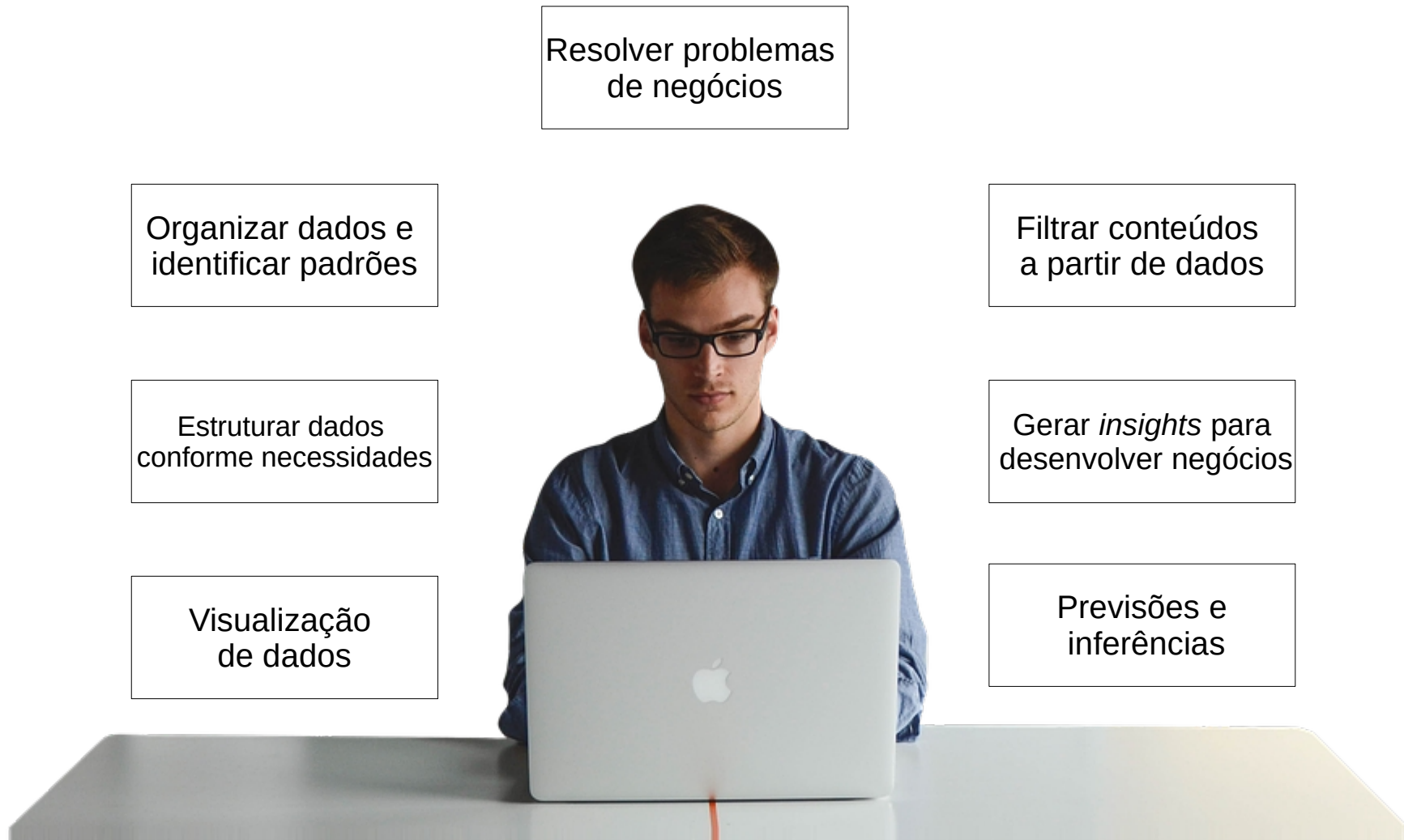
- A principal diferença é a cerca das **tecnologias e métodos científicos** utilizados entre eles, como por exemplo:

	Ciência de dados	Inteligência de Negócios
<b>Entradas</b>	Dados coletados dentro da organização e dados externos	Dados coletados dentro da organização
<b>Tecnologias e Ferramentas</b>	Aprendizado de máquina, estatística, Python, R	OLAP, Data Mart, ETL
<b>Saídas</b>	Analisar padrões e gerar previsões de grandes quantidades de dados	Inferências a partir de dados históricos ou atuais.

# O que faz um Cientista de dados?

- Com o avanço da tecnologia e da globalização as **organizações** tendem a **gerar uma grande quantidade de dados**.
- Criando uma **oportunidade** de negócio para esses **dados serem analisados** para gerar **valor**.
- Essa **atividade cabe ao Cientista de dados**, uma **carreira** que é vista por especialista como uma das mais **promissoras** da atualidade.

# O que faz um Cientista de dados?



# Perfil do Profissional

Business

Tecnologia

Estatística

Algoritmos e  
Programação

Matemática

Capacidade de  
aprendizagem

Criatividade



# Glossário do Cientista de dados

- 1. Insight:** é um termo utilizado quando após a análise de dados se consegue encontrar uma solução/conclusão de um problema ou se identifica algum padrão por meio da observação e dedução.
- 2. Algoritmo:** consiste em um serie de instruções que serão executadas por uma máquina para alcançar um objetivo.
- 3. Dataset:** representa um conjunto de dados

# Glossário do Cientista de dados

**4. Reconhecimento de Padrões:** é o ato de analisar automaticamente dados brutos e realizar uma ação com base na categoria de um padrão.

**5. Machine Learning:** Aprendizado de máquina são algoritmos da ciência da computação capazes de ensinar máquinas.

**6. Features:** Um expressão do aprendizado de máquina para se referir a uma característica que pode representar um dado

# Glossário do Cientista de dados

**7. Mineração de dados:** é um processo para se identificar tendências, correlacionar dados e segregar dados em big data.

**8. Análise preditiva:** modelos para tentar prever situações que podem ocorrer no futuro.

**9. Clustering:** técnicas para coletar e categorizar dados em grupos que sejam suficientemente similares.

**10. R statistical:** uma linguagem de programação open source e multiplataforma

# Ciência de Dados

- Por que a ciência de dados está causando uma corrida por informações?





# Ciência de Dados

- Como essas coisas podem se conectar? Restaurantes, mercados, viagens, qualificação profissional, satisfação, etc...

**Tudo está conectado de alguma forma, os dados estão em toda parte**



# Ciência de Dados

- O papel da ciência de dados é **encontrar** essa **relação** e como podem ser **convertida em estratégias**, como:
  - Marketing
  - Vendas
  - Novos segmentos
  - Otimização de lucros
  - Minimização de perdas
  - Satisfação do cliente
  - Gostos do cliente
  - Tipo do cliente
  - Parceiros em potencias
  - Etc.