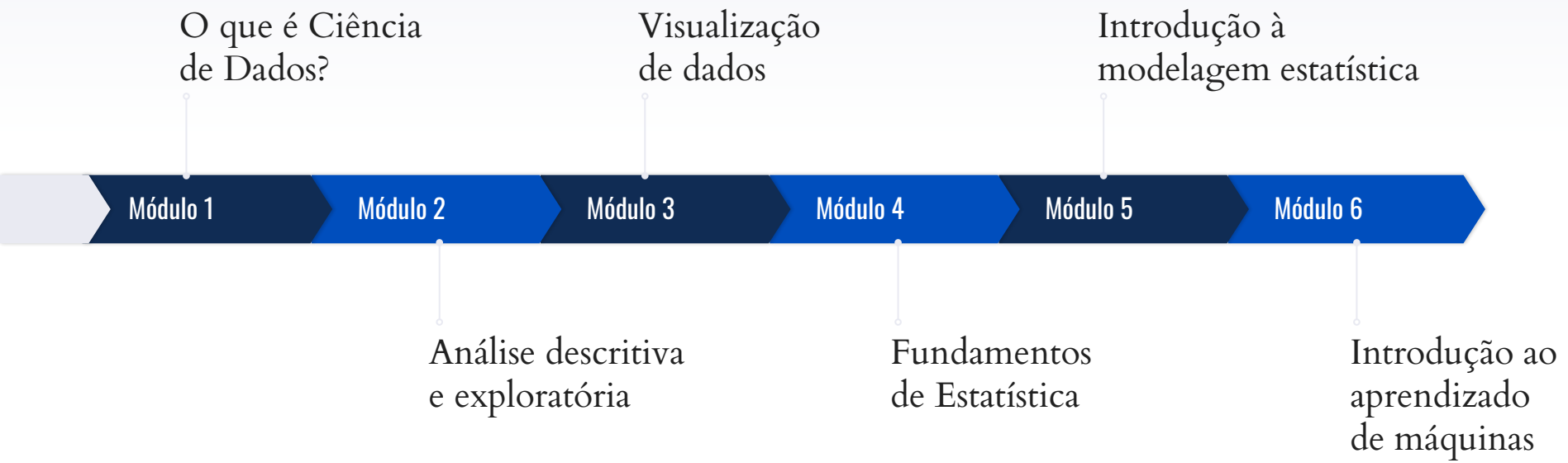
INTRODUÇÃO À CIÊNCIA DE DADOS

Sua jornada começa aqui

omegadatascience.com.br

Estrutura do curso



MÓDULO 4
FUNDAMENTOS DE ESTATÍSTICA

Objetivo

Dar uma visão geral do pensamento estatístico, seus pilares e principais conceitos.

O pensamento analítico

Compreenda como a estatística modela fenômenos aleatórios e permite tomar melhores decisões baseadas em dados.

Problemas de negócios

Entenda como a estatística é usada para resolver problemas e criar novos negócios.

Inferência estatística

Aprenda como a ciência estatística é construída por meio de um projeto de dados.

MÓDULO 4
FUNDAMENTOS DE ESTATÍSTICA

1. O problema de negócio.
2. O pensamento estatístico.
3. Tomando decisões na presença da incerteza.
4. O desenvolvimento de um produto baseado em dados.
5. Distribuição amostral e a inferência estatística.
6. Separando fontes de variação.
7. Projeto IV: Identificando oportunidades.

Vamos começar!

Telegram: t.me/omegadatascience

Instagram: @omegadatascience

Twitter: @omegadatascienc

YouTube: /OmegaDataScience

Ômega Data Science

PLATAFORMA DE CURSOS ONLINE

omegadatascience.com.br

Ômega Fly

- **Área de negócio:** Venda de passagens áreas para empresas.
- **Objetivos:**
  - Aumentar o portfólio de produtos.
  - Aumentar a satisfação e lealdade dos clientes.
- **Problema:** Voo atrasado.
- **Oportunidade:** Criar um produto que ajude os colaboradores a planejar melhor suas viagens.

Relato do time de produto

Ao conversar com as equipes de aproximadamente 200 empresas multinacionais o time de negócios da Ômega Fly detectou que uma dor de cabeça comum é saber qual o melhor horário para reservar um voo para uma reunião/missão importante de negócios. O atraso dos voos é frequente e causa desconforto nos colaboradores, uma vez que:

1. A reunião/missão pode atrasar ou mesmo não acontecer devido ao atraso.
2. Custos adicionais com mudança no voo de volta e acomodação.
3. Retrabalho para buscar realocação de voo.

O seu time

José: Diretor de produto
Formação: ADM + Pós em Gestão de Projetos

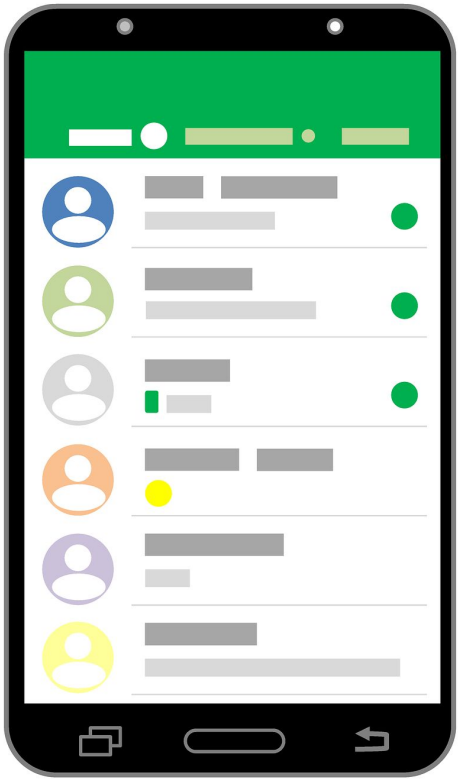
Maria: Gerente de vendas
Formação: Negócios internacionais

João: Desenvolvedor Full Stack
Formação: Vida

Você: Cientista de dados
Formação: Ômega

Produto

- **Smart travel**
  - Cliente reporta cidade de origem e de destino e qual o horário gostaria de chegar.
  - Sistema busca qual é o voo mais indicado de modo antes do horário desejado.
- **Se o voo atrasar a Ômega Fly cobre:**
  - Custos adicionais de remarcação.
  - Acomodação + estadia no local de destino.



Por que te contrataram?

- Como devemos fazer a procura por voos de forma que o cliente não se atrase?

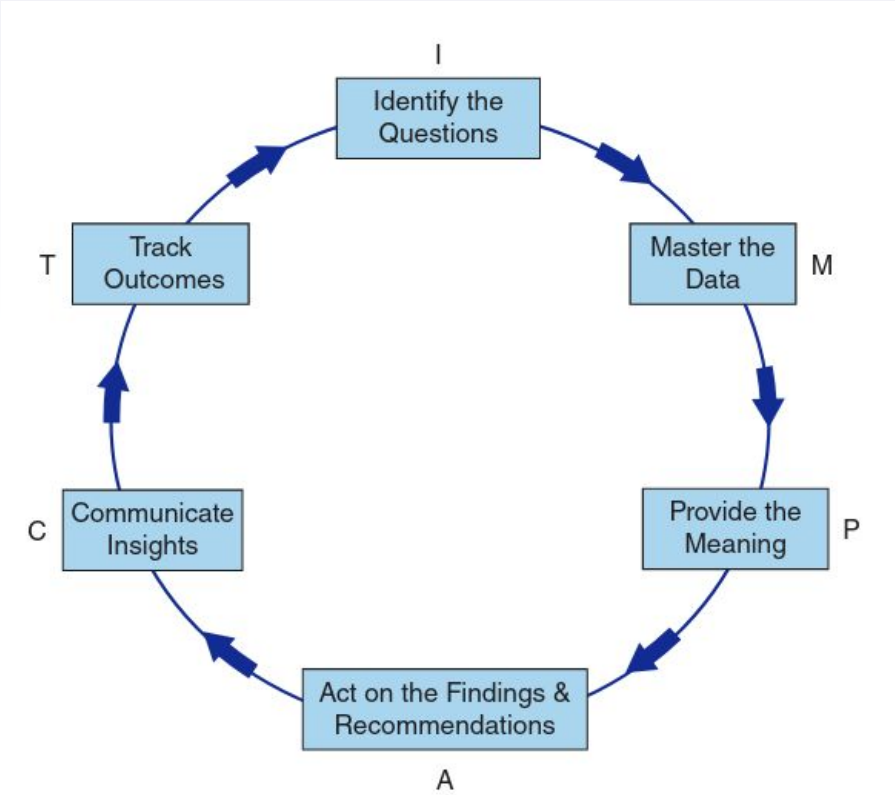
Por que te contrataram?

- Como devemos fazer a procura por voos de forma que o cliente não se atrase para a reunião?

É impossível!

Se existe uma chance de dar errado, vai dar errado!

IMPACT



Fonte: Isson, J. P.; Harriot, J. S. People Analytics in the Era of Big Data, Wiley, 2016. Pag 56.

O que é importante pro negócio?

Incerteza é inevitável	Se viável	Aspectos chave
O produto é viável?	Como monetizar?	Entenda tudo de atraso!
Qual o risco associado?	Quanto cobrar?	Fonte de informações.
Como criar garantias?	Será que o cliente quer pagar o quanto eu preciso cobrar?	Entenda o negócio.

Entenda o negócio e comece pequeno!

Você: Qual é a nossa principal rota de vendas?

Maria: É Nova York até São Francisco.

Você: Qual o tempo aproximado de voo?

Maria: Aproximadamente 3:40.

Busque dados

Você: Quais dados temos sobre os voos entre Nova York e São Francisco?

João: Temos um banco de dados relacional em MySQL. Guardamos diversas informações sobre todos os voos reportados pela agência nacional de aviação.

Você: Você pode me dar acesso?

João: Hum… é complicado o banco está em produção e todos os nossos produtos dependem dele! Mas posso extrair pra você.

Hora do código!

Tipos de fenômenos

Fenômenos determinísticos

Dizemos que um fenômeno é determinístico quando repetido inúmeras vezes, sob as mesmas condições, conduz a resultados essencialmente idênticos:

- Aceleração da gravidade.
- Leis da física (mecânica clássica) e da Química.

Fenômenos aleatórios

Dizemos que um fenômeno é aleatório quando repetido inúmeras vezes, em condições, conduz a resultados diferentes:

- Lançamento de uma moeda.
- Resultado de um evento esportivo.
- Atrasos em voos.
- Praticamente toda a natureza!

Teoria das probabilidades

O que é a teoria das probabilidades?

Ramo da matemática que desenvolve e avalia modelos para descrever fenômenos aleatórios.

Qual o objetivo da teoria das probabilidades?

Construir um arcabouço matemático adequado para descrever fenômenos aleatórios.

O que precisamos para começar?

Descrever o conjunto de resultados possíveis (espaço amostral **Ω**).

Atribuir pesos a cada possível resultado, refletindo suas **chances** de ocorrência.

O que é probabilidade?

Probabilidade é uma função $P(\cdot)$ que atribui valores numéricos aos eventos do espaço amostral de tal forma que

1. $0 \leq P(A) \leq 1, \quad \forall A \in \Omega$;
2. $P(\Omega) = 1$;
3. $P(\bigcup_{j=1}^{n} A_j) = \sum_{j=1}^{n} P(A_j)$, com os A_j's disjuntos.

Como atribuir probabilidades?

1. Definição clássica: baseia-se nas características teóricas da realização do fenômeno.

2. Frequentista: baseia-se nas frequências (relativas) de ocorrência do fenômeno.

3. Subjetiva: baseia-se no julgamento pessoal ou experiência própria sobre a plausibilidade/chance de algo ocorrer.

Distribuição de probabilidade

- A **função de probabilidade** (fp) da v.a. discreta Y, que assume os valores $y_1, y_2, \ldots, y_n$, é a função que atribui probabilidades a cada um dos possíveis valores: $\{y_i, \mathrm{p}(y_i)\}, i = 1, 2, \ldots$, ou seja,

$$\mathrm{P}(Y = y_i) = \mathrm{p}(y_i) = \mathrm{p}_i, \quad i = 1, 2, \ldots$$

om as seguintes propriedades:

- A probabilidade de cada valor deve estar entre 0 e 1,

$$0 \leq \mathrm{p}(y_i) \leq 1, \quad \forall\, i = 1, 2, \ldots.$$

- A soma de todas as probabilidades é igual a 1

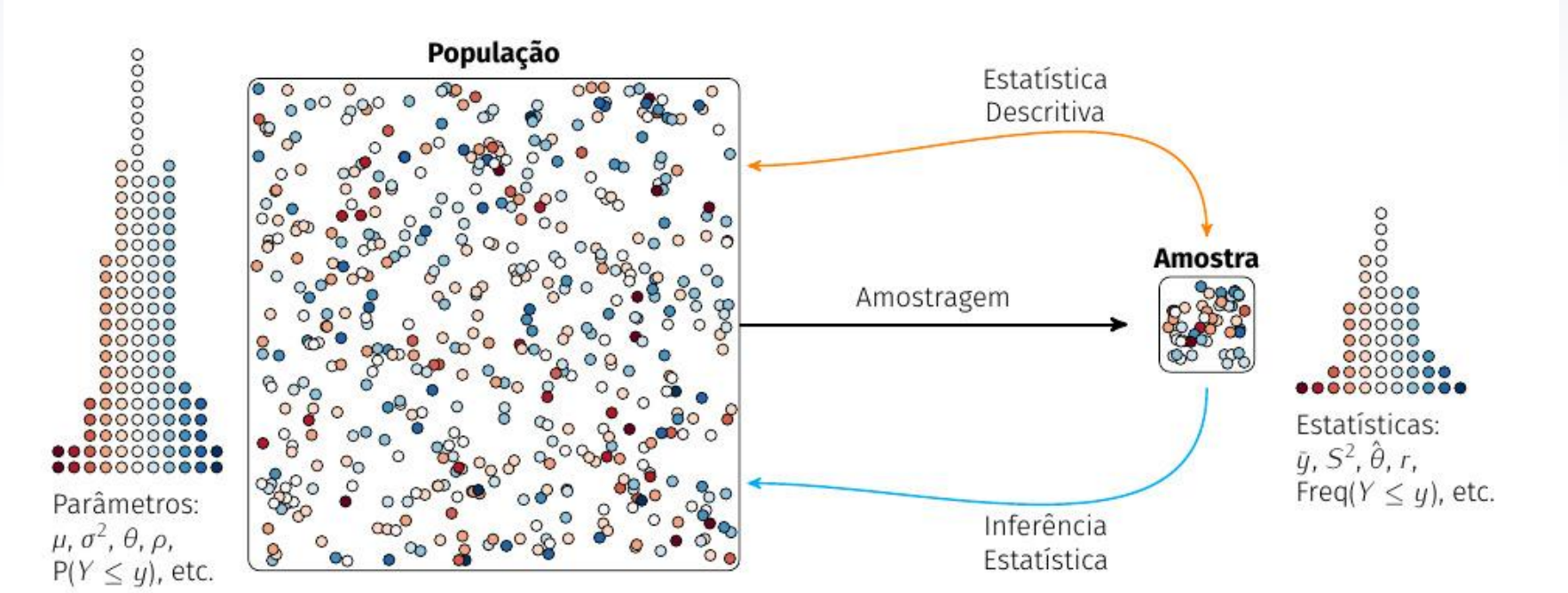
$$\sum_i \mathrm{p}(y_i) = 1.$$

Esperança matemática

- Seja Y uma v.a. com distribuição de probabilidade $f(y)$.
- A **média** ou **valor esperado** de Y é dado por:
  - Caso discreto: $\mu = \mathrm{E}(Y) = \sum_y yf(y)$;
  - Caso contínuo: $\mu = \mathrm{E}(Y) = \int_{-\infty}^{\infty} yf(y)dy$.

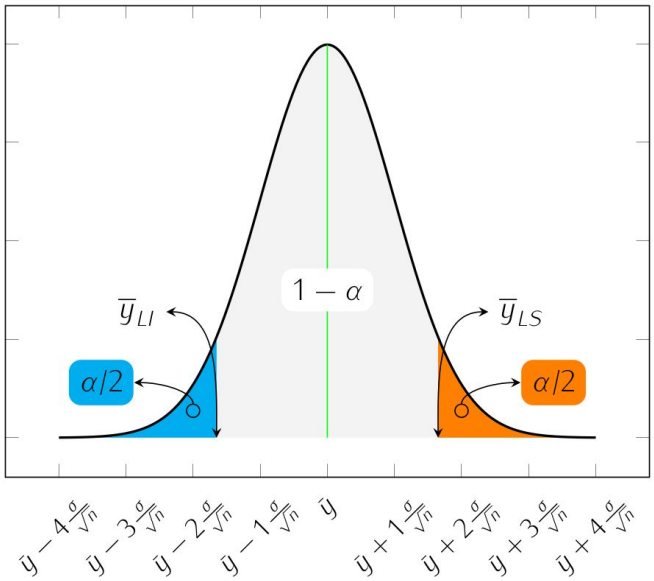
Figura 1. Foto de Clement Eastwood no Pexels.

Inferência estatística



Distribuição amostral

1. É o foco da abordagem frequentista.
2. Fracamente falando: O que acontece se eu repetir o experimento muitas vezes?
3. Estimativa pontual.
4. Intervalos de confiança.
5. Testes de hipóteses.
6. p-valor.



Como se entrega resultados?

Insights acionáveis

Prescrever o que deve ser feito baseado no conhecimento adquirido.

Informação

Entender o que aconteceu no passado.

Conhecimento

Entender o que está acontecendo agora e porque.

Inteligência

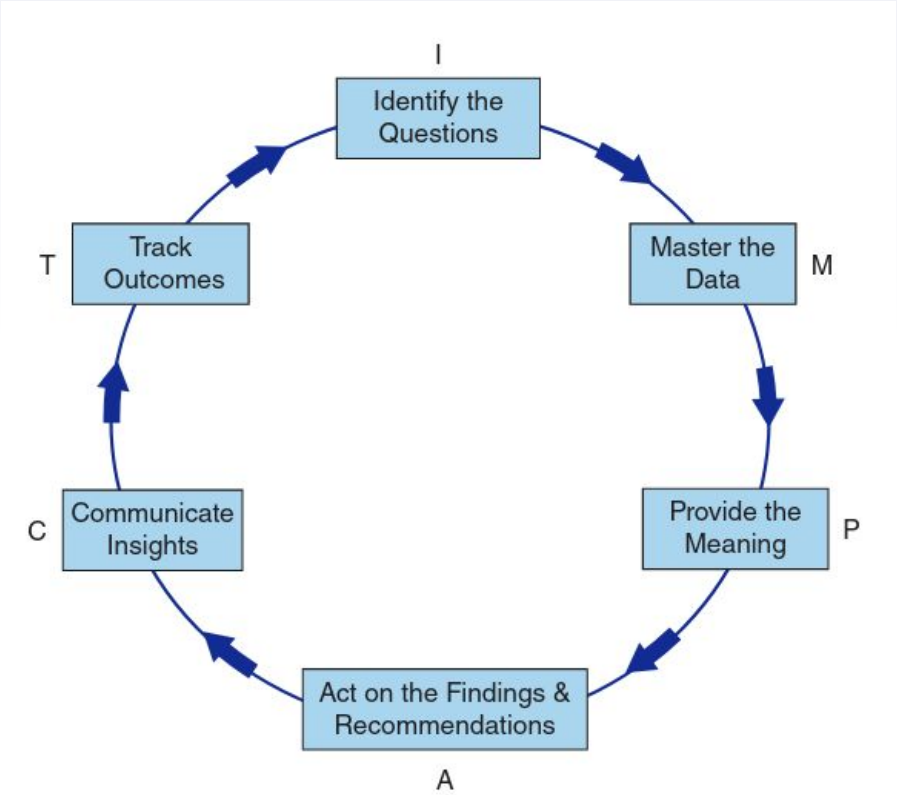
Antecipar o que vai acontecer no futuro.

Resumo executivo

- **Suposições**
  - Venderemos 10000 tickets no ano.
  - Condições do passado serão mantidas aproximadamente constantes.
  - Atraso é independente entre os clientes.
  - Política trivial: 60 minutos antes para todos.

- **Resultados**
  - Qual o percentual de tickets teremos que ressarcir?
  - 6,94% (6,47% | 7,46%)
  - Qual o custo esperado total do produto?
  - 312.383,70 (291.150,00 | 335.700,00)
  - Quanto devemos acrescer em cada passagem para em média cobrir os custos do produto?
  - 31,23 (29,11 | 33,57).
  - Quanto devemos cobrar para não ter prejuízo em 95% das vezes?
  - R$ 33,57.

IMPACT



Fonte: Isson, J. P.; Harriot, J. S. People Analytics in the Era of Big Data, Wiley, 2016. Pag 56.

Algumas questões

- Produto ficou muito caro!
- Recomendação muito simples?
- Pode ter baixa adesão pelo preço e baixa complexidade.
- Será que esta entregando algum valor para o cliente?
- Risco é alto.
- Onde podemos melhorar?
- **Precisamos entender tudo de atraso!**

Por que um voo atrasa?

- Aeroporto de origem.
- Operadora do voo.
- Horário previsto para a saída.
- Dia da semana ou do mês.
- Mês do ano.

Testes de hipóteses

Hipótese

É uma afirmativa sobre uma propriedade da população.

Hipótese nula

É uma afirmativa de que o valor de um parâmetro populacional é **igual** a algum valor especificado.

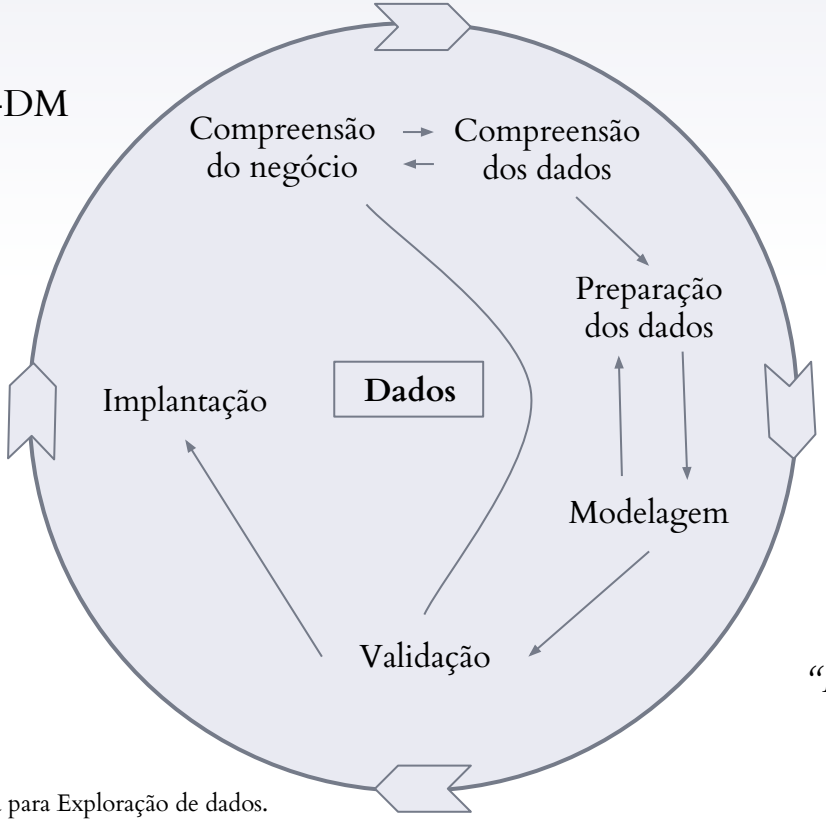
Nula -> Nenhuma mudança.

Hipótese alternativa

É uma afirmativa de que o parâmetro tem um valor que, de alguma forma, **difere** da hipótese nula.

Projetos de Ciência de Dados

Metodologia · CRISP-DM



Processo investigativo.
Não apenas construtivo.

"Falhar rápido para ter sucesso mais cedo!"

Processo Padrão de Indústria Cruzada para Exploração de dados.

Onde estamos?

- Avaliamos o risco de uma estratégia.
- E quanto a outras?
- Podemos automatizar a avaliação do risco?
- Como entregamos essa automação?

Oportunidades de melhorias

Como podemos diminuir o preço do produto?

1. Diminuir a probabilidade de pagar!
2. Encontrar subgrupos onde é mais vantajoso ofertar o produto.
3. Aumentar o tempo de tolerância.
4. Encontrar outras rotas que sejam mais rentáveis.
5. Diluir os custos entre diferentes rotas.
6. Precisamos de uma forma mais geral de controlar as causas de variação -> Modelo estatístico.

Identificando oportunidades

1. Construa uma narrativa para justificar a escolha de política de preços diferentes entre operadoras (carrier).
2. Se tivesse que escolher apenas uma operadora para cobrar mais caro, qual você escolheria?
3. Como poderíamos criar políticas de preços melhores?
4. O que você faria para melhorar a rentabilidade deste produto?
5. Como você modificaria o produto para torná-lo mais interessante para o cliente?

Revisando

1. O problema de negócio.
2. O pensamento estatístico.
3. Tomando decisões na presença da incerteza.
4. O desenvolvimento de um produto baseado em dados.
5. Distribuição amostral e a inferência estatística.
6. Separando fontes de variação.