Ciência de Dados e Multidisciplinaridade

II Seminário de Aplicações em Ciência de Dados

Marcus Nunes

7 de Dezembro de 2018

Universidade Federal do Rio Grande do Norte

Quem sou Eu?

Quem sou Eu?

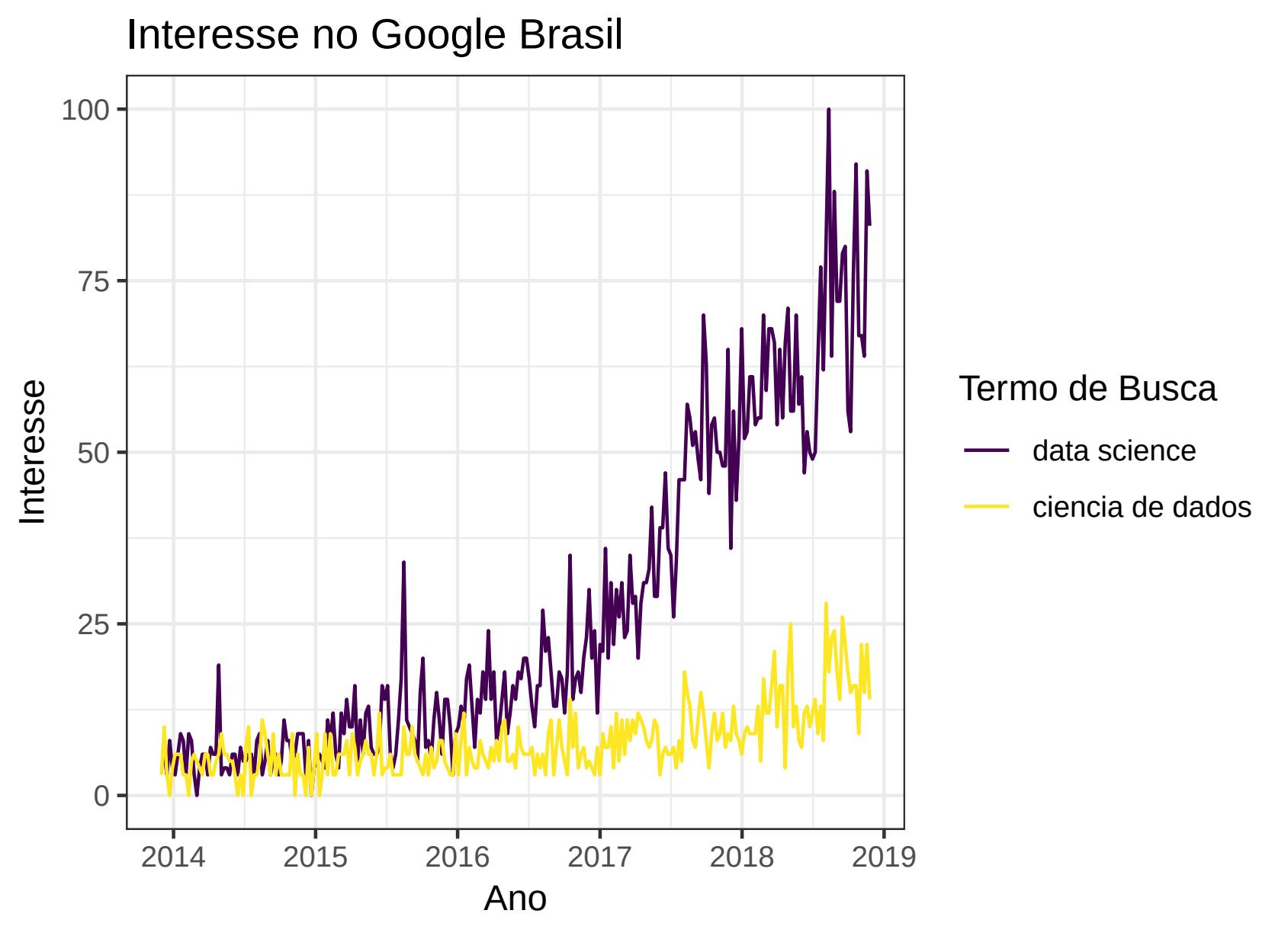
- Marcus Nunes, professor do Departamento de Estatística da UFRN
- Coordenador do Laboratório de Estatística Aplicada - http://www.estatistica.ccet.ufrn.br/?page_id=156
- 1ª Competição de Ciência de Dados do Departamento de Estatística - http://www.estatistica.ccet.ufrn.br/?page_id=608
- https://marcusnunes.me/

O que é Ciência de Dados?

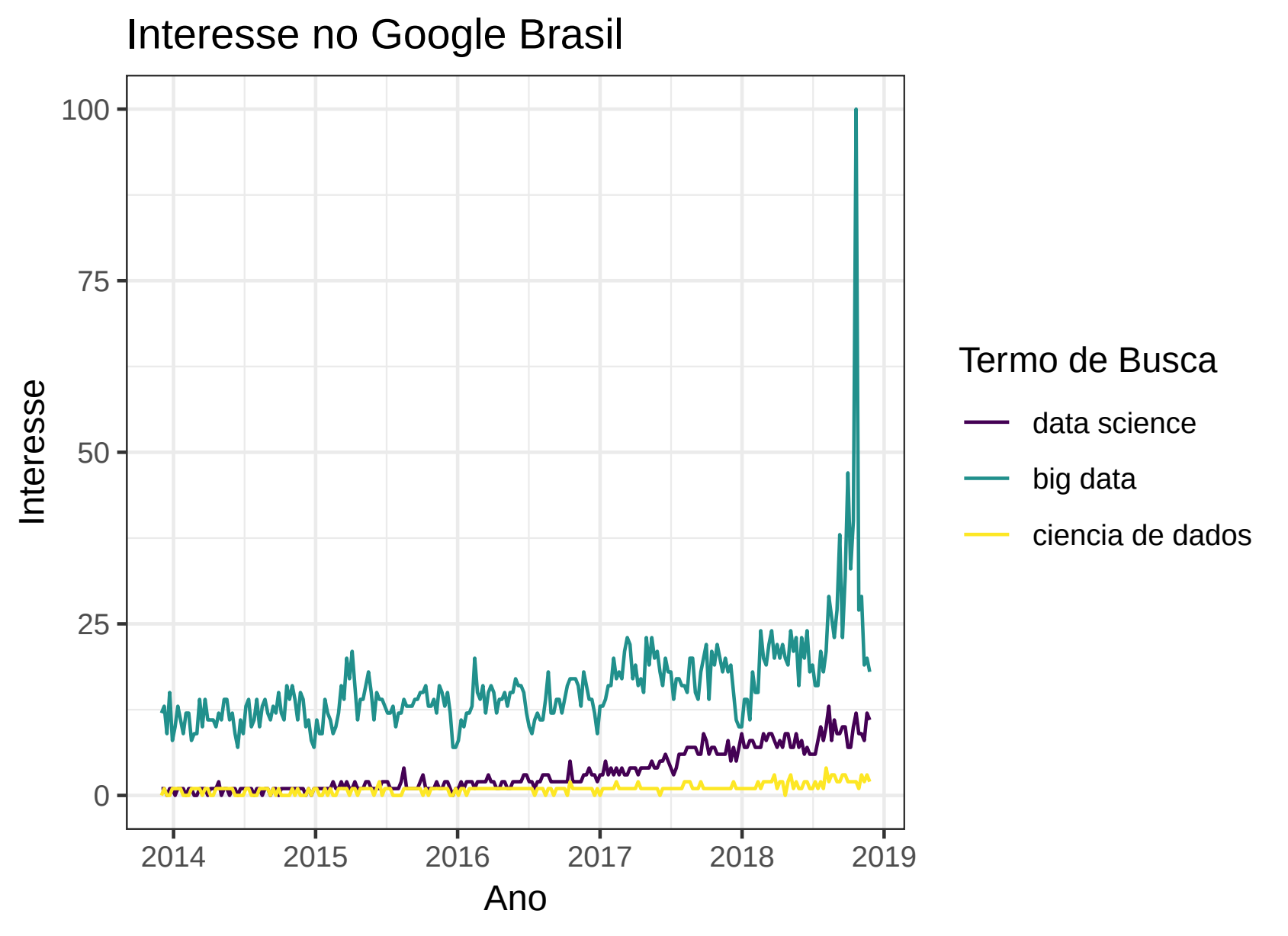
O que é Ciência de Dados?

- *Buzzword* muito utilizada atualmente
- Juntamente com *big data* e *data science*, o termo ganhou muita força nos últimos cinco anos

O que é Ciência de Dados?



O que é Ciência de Dados?



O que é Ciência de Dados?

- Alguém tem alguma definição?

Multidisciplinaridade

- Uma equipe multidisciplinar é formada por especialistas em várias áreas do conhecimento
- É através da combinação de ideias e do conflito ente elas que surgirão soluções para os problemas

Multidisciplinaridade

- Competências de um profissional 100% capacitado para trabalhar com Ciência de Dados:
  - Estatística
  - Programação
  - Negócios
  - Conhecer bem a área de atuação (internet, varejo, finanças etc)

Multidisciplinaridade

- Que tipo de profissionais temos no momento?
  - Bons estatísticos e matemáticos que escrevem códigos sem otimização
  - Bons cientistas da computação que entendem um pouco de estatística e matemática
  - Bons cientistas da computação que entendem um pouco de negócios, depois de muita experiência na área
  - Especialistas em alguma área de atuação
  - Gerentes que sabem fazer estas pessoas trabalharem juntas

Quem Trabalha com Ciência de Dados?

- Estatísticos
- Programadores
- Físicos
- Cientistas de Dados

Um Problema Real

Como solucionar o limite 53 categorias do randomForest do R?

Faça uma pergunta

6

No R, utilizando a library `randomForest`, quando executo `randomForest()` recebo a seguente mensagem de erro:

```
Error in randomForest.default(m, y, ...) :
  Can not handle categorical predictors with more than 53 categories.
```

O fator em questão tem 57 categorias. Como posso mudar este limite ou contornar este problema?

r randomforest

compartilhar editar
fechar sinalizar

editada 5/11 às 16:36

4.259 ● 2 ● 4 ● 28

perguntada 5/11 às 16:22

541 ● 2 ● 23

Um Problema Real

- A construção da árvore envolve os seguintes passos:
    1. A seleção das divisões
    2. As decisões de quando declarar um nó como folha ou continuar dividindo-o
    3. A alocação de cada nó folha a uma classe

Um Problema Real

- Em particular, precisamos decidir o seguinte:
  1. Um conjunto $\mathcal{Q}$ de decisões binárias do tipo $\{X \in A\}, A \subset \mathcal{X}$
  2. Um critério de divisão $\phi(s, t)$ que pode ser calculado para qualquer divisão s de qualquer nó t
  3. Uma regra para finalizar as divisões
  4. Uma regra para alocar cada nó folha a uma classe

Um Problema Real

- Se X é categórica com valores, digamos, $\{1, 2, \cdots, M\}$, então $\mathcal{Q}$ contém todas as questões da forma

$$\{X_j \in A?\},$$

  onde A varia sob todos os subconjuntos de $\{1, 2, \cdots, M\}$
- As divisões para todas as p variáveis constituem o conjunto padrão de questões

Um Problema Real

- Como são 57 categorias na variável, são $2^{57} - 2 \approx 1{,}441 \times 10^{17}$ separações diferentes que devem ser feitas
- Por melhor que seja o seu hardware, algo da ordem de 10^{17} ainda é muito grande
- Desta forma, não basta apenas trabalhar mais: é preciso trabalhar de forma inteligente

Um Problema Real

- Ela seria uma variável numérica que foi lida incorretamente?
- Caso a variável seja categórica, é possível tratá-la como variável ordinal?
- Caso a variável seja categórica nominal, é possível simplificá-la em menos categorias? Por exemplo, caso sejam países do mundo, é possível criar uma nova variável chamada continente que vai ter apenas 6 níveis?
- Caso a variável seja categórica nominal, é possível simplificá-la em menos categorias? Todos os níveis são representativos? Seria possível combinar os níveis de menor frequência em um novo nível?

O que é um Cientista de Dados?

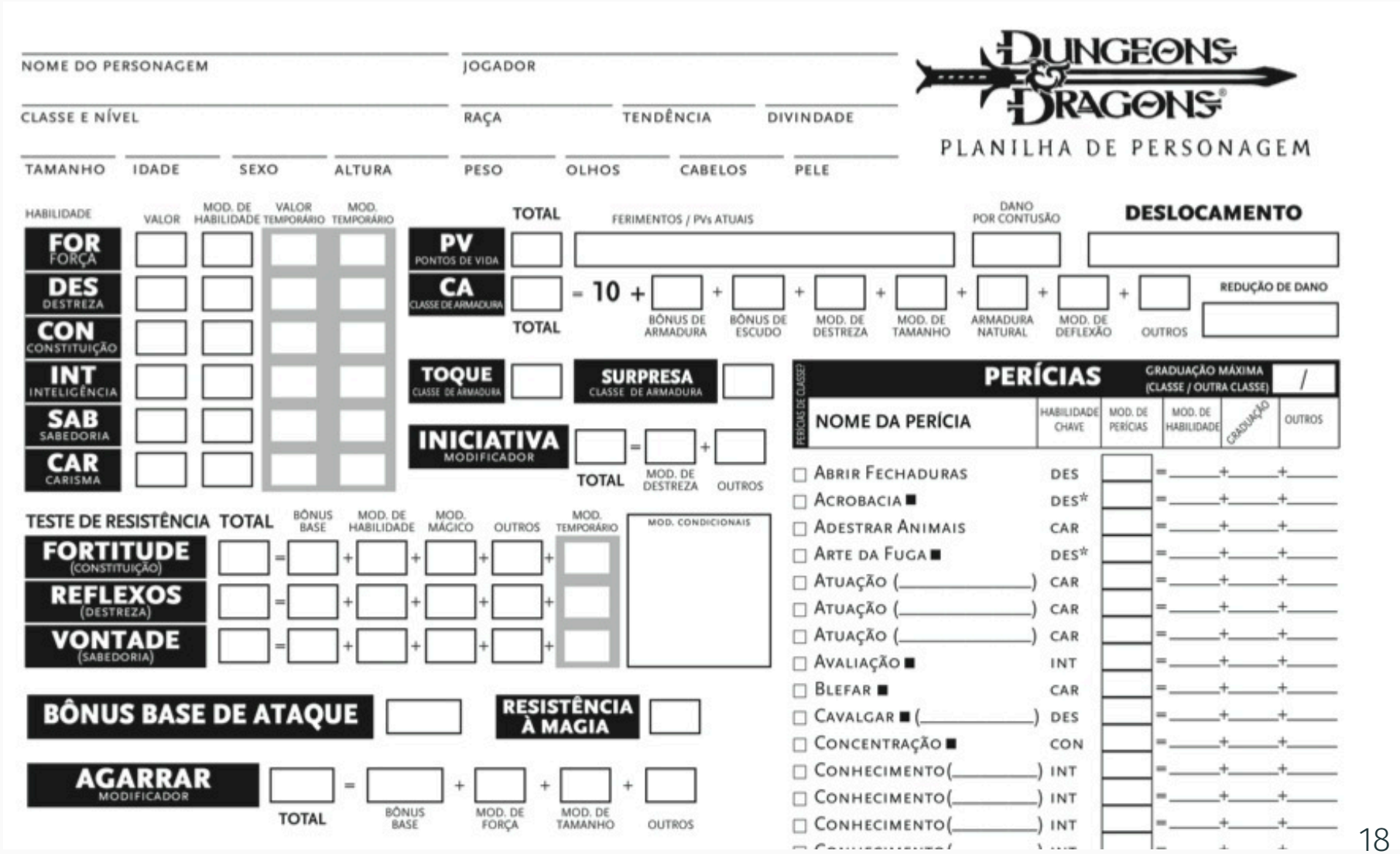
- Cientista de Dados (*Data Scientist*) é o novo nome para Estatístico
- Alguns dizem que o Cientista de Dados é um Estatístico que mora em São Francisco e usa um Mac
- No fundo, ambos são a mesma coisa, embora uma destas profissões trabalhe melhor seu marketing pessoal

O que é um Cientista de Dados?

- É alguém que entende mais de programação do que um Estatístico tradicional
- É alguém que entende mais de estatística do que um Cientista da Computação tradicional
- E, principalmente, é alguém que consegue encontrar soluções para problemas juntando estas duas áreas do conhecimento

Quem já jogou RPG?

Quem já jogou RPG?



NOME DO PERSONAGEM | JOGADOR

CLASSE E NÍVEL | RAÇA | TENDÊNCIA | DIVINDADE

TAMANHO | IDADE | SEXO | ALTURA | PESO | OLHOS | CABELOS | PELE

DUNGEONS & DRAGONS®

PLANILHA DE PERSONAGEM

HABILIDADE | VALOR | MOD. DE HABILIDADE | VALOR TEMPORÁRIO | MOD. TEMPORÁRIO

FOR FORÇA

DES DESTREZA

CON CONSTITUIÇÃO

INT INTELIGÊNCIA

SAB SABEDORIA

CAR CARISMA

PV PONTOS DE VIDA — TOTAL — FERIMENTOS / PVs ATUAIS — DANO POR CONTUSÃO — DESLOCAMENTO

CA CLASSE DE ARMADURA — TOTAL = 10 + BÔNUS DE ARMADURA + BÔNUS DE ESCUDO + MOD. DE DESTREZA + MOD. DE TAMANHO + ARMADURA NATURAL + MOD. DE DEFLEXÃO + OUTROS — REDUÇÃO DE DANO

TOQUE CLASSE DE ARMADURA

SURPRESA CLASSE DE ARMADURA

INICIATIVA MODIFICADOR — TOTAL = MOD. DE DESTREZA + OUTROS

TESTE DE RESISTÊNCIA | TOTAL | BÔNUS BASE | MOD. DE HABILIDADE | MOD. MÁGICO | OUTROS | MOD. TEMPORÁRIO | MOD. CONDICIONAIS

FORTITUDE (CONSTITUIÇÃO)

REFLEXOS (DESTREZA)

VONTADE (SABEDORIA)

BÔNUS BASE DE ATAQUE

RESISTÊNCIA À MAGIA

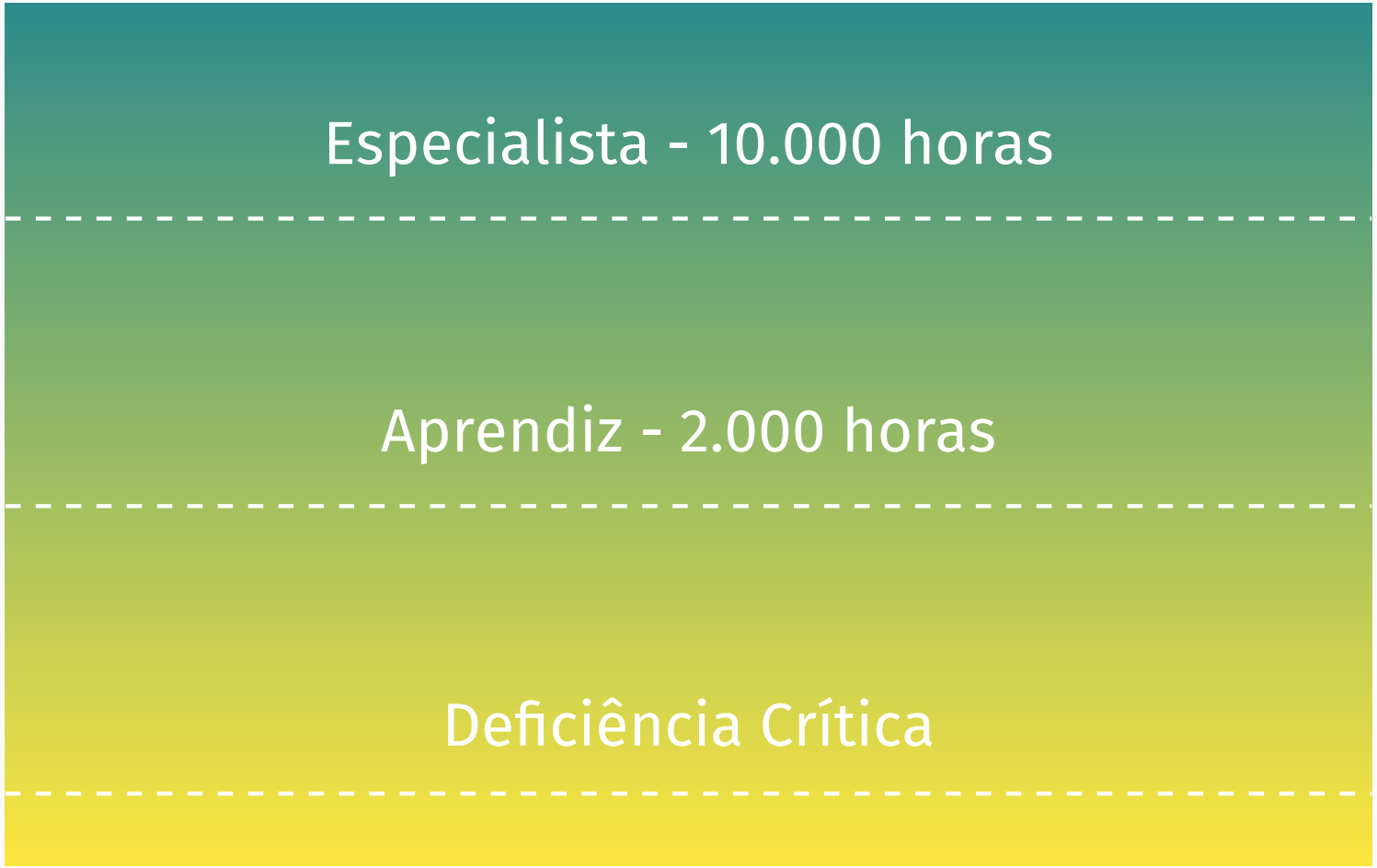
AGARRAR MODIFICADOR — TOTAL = BÔNUS BASE + MOD. DE FORÇA + MOD. DE TAMANHO + OUTROS

PERÍCIAS — GRADUAÇÃO MÁXIMA (CLASSE / OUTRA CLASSE) /

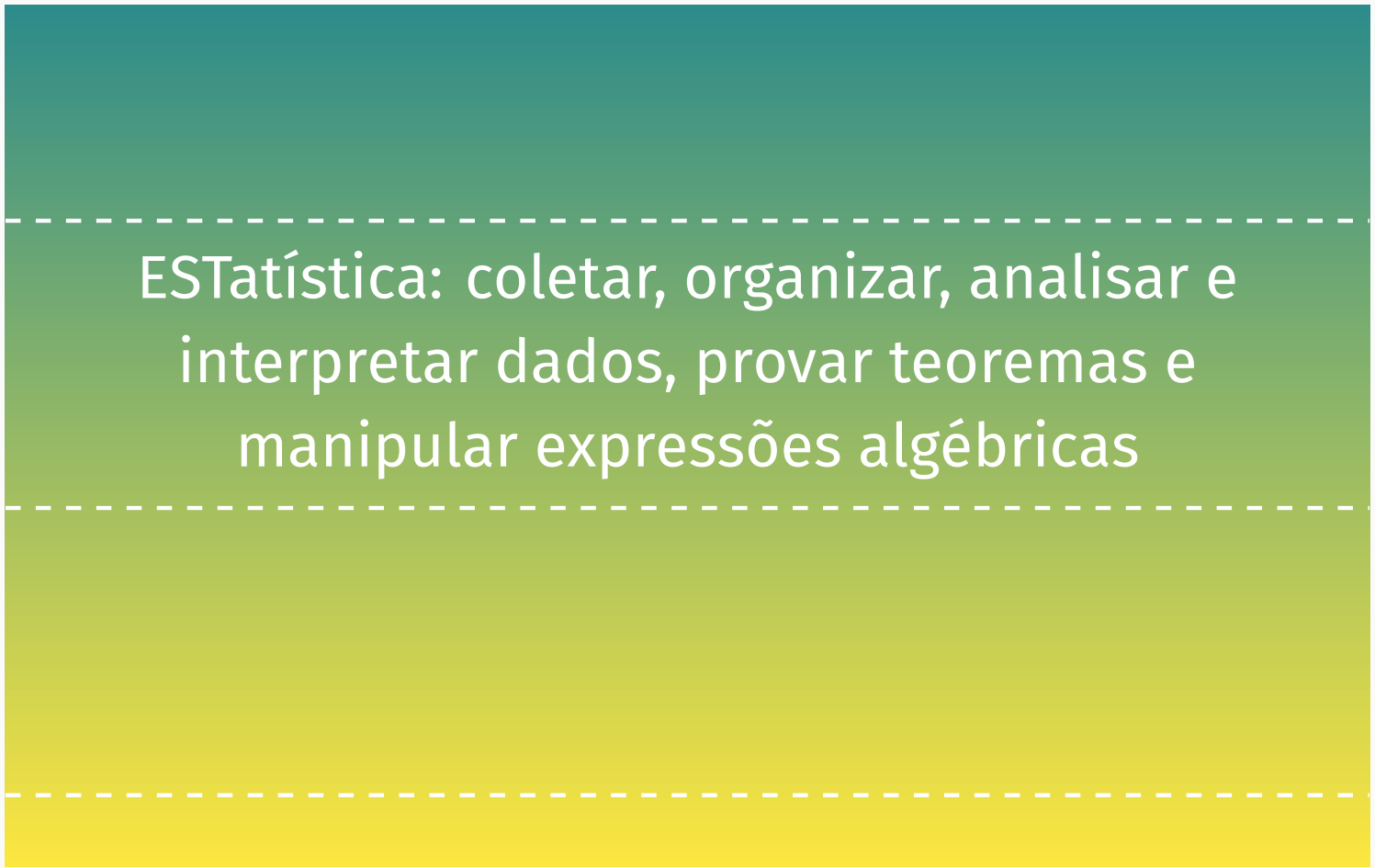
PERÍCIAS DE CLASSE?

NOME DA PERÍCIA	HABILIDADE CHAVE	MOD. DE PERÍCIAS	MOD. DE HABILIDADE	GRADUAÇÃO	OUTROS
☐ Abrir Fechaduras	DES				
☐ Acrobacia ■	DES*				
☐ Adestrar Animais	CAR				
☐ Arte da Fuga ■	DES*				
☐ Atuação (_____)	CAR				
☐ Atuação (_____)	CAR				
☐ Atuação (_____)	CAR				
☐ Avaliação ■	INT				
☐ Blefar ■	CAR				
☐ Cavalgar ■ (_____)	DES				
☐ Concentração ■	CON				
☐ Conhecimento (_____)	INT				
☐ Conhecimento (_____)	INT				
☐ Conhecimento (_____)	INT				

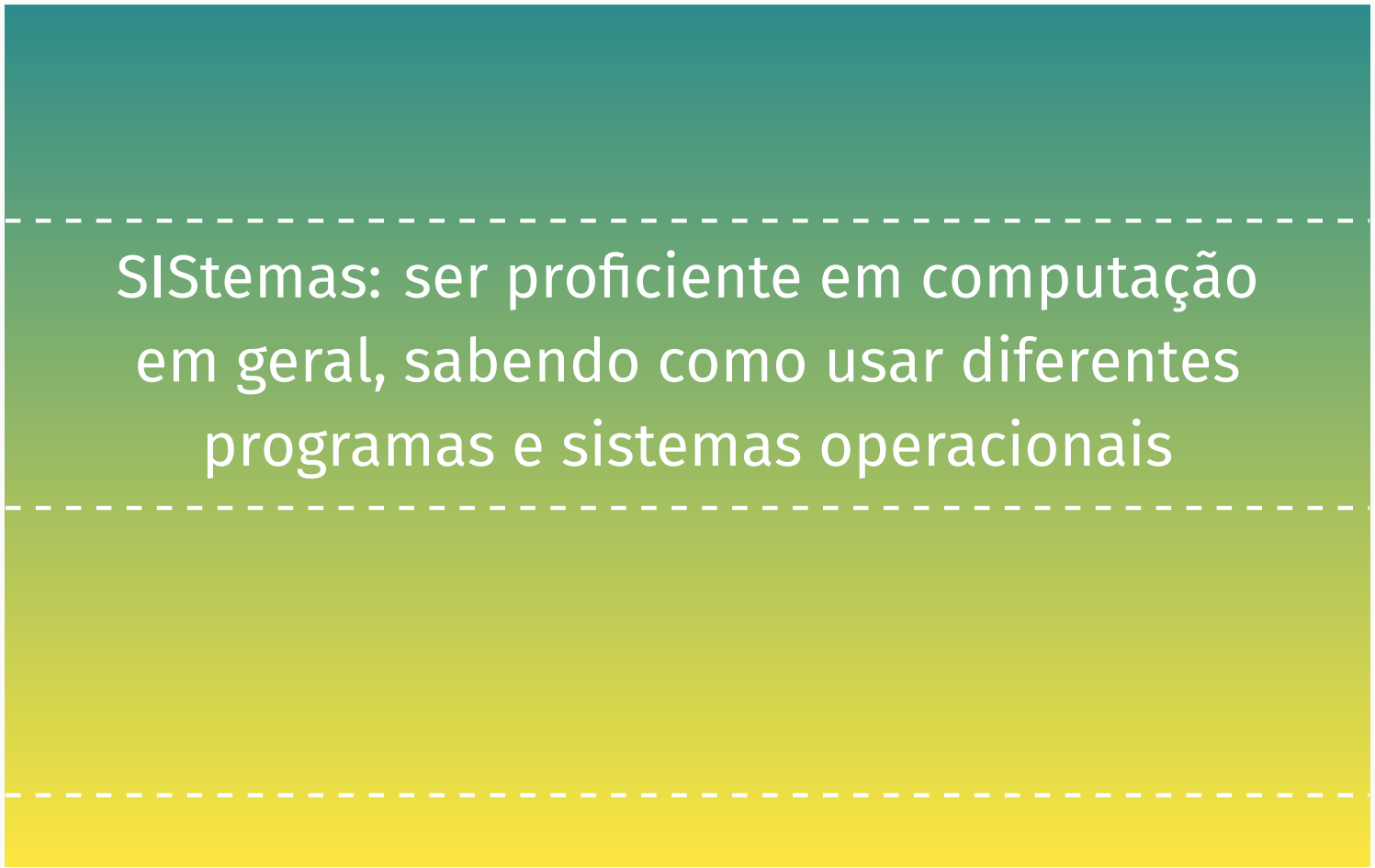
Malcolm Gladwell - Outliers



O que um Cientista de Dados Precisa Saber?

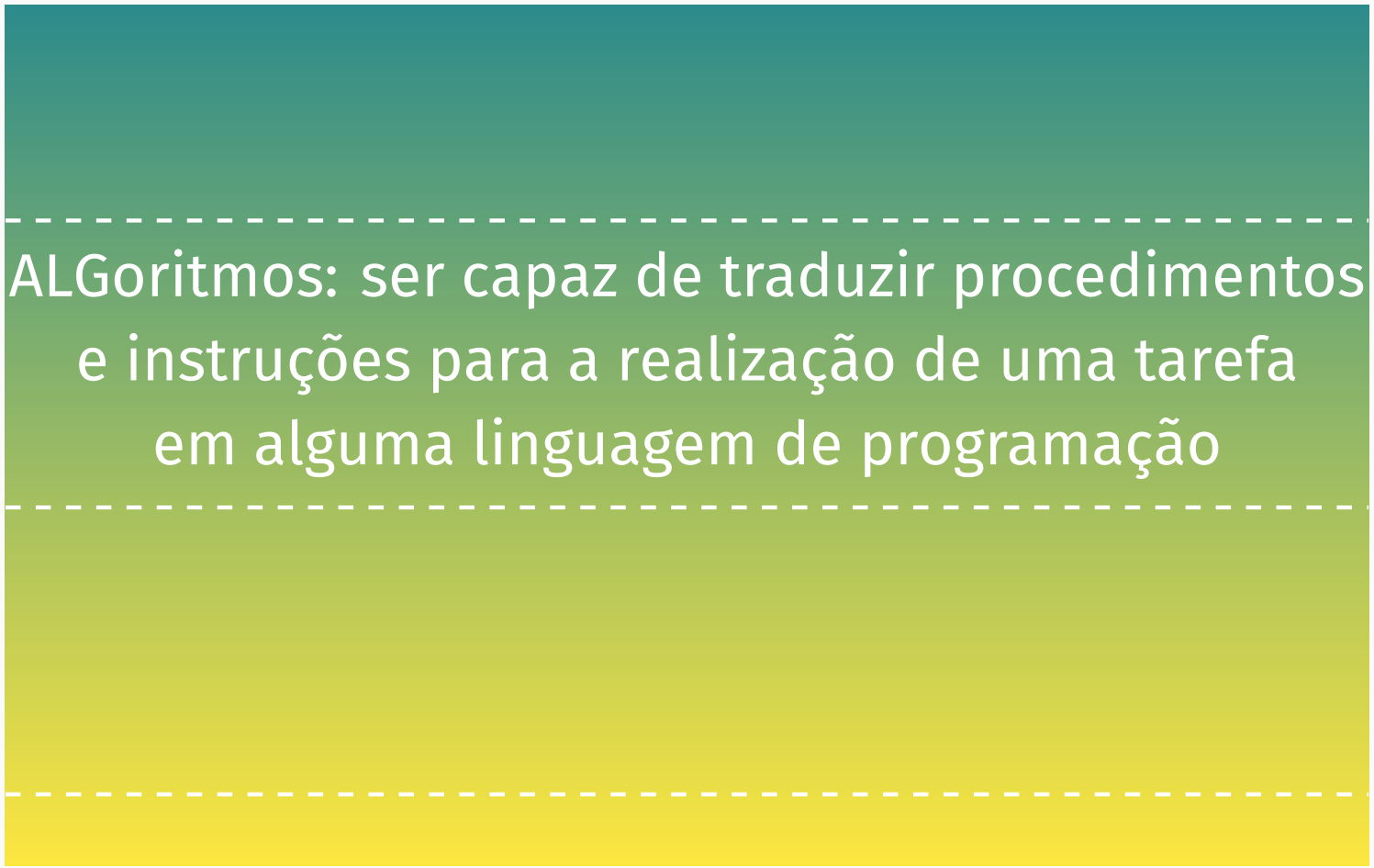


O que um Cientista de Dados Precisa Saber?



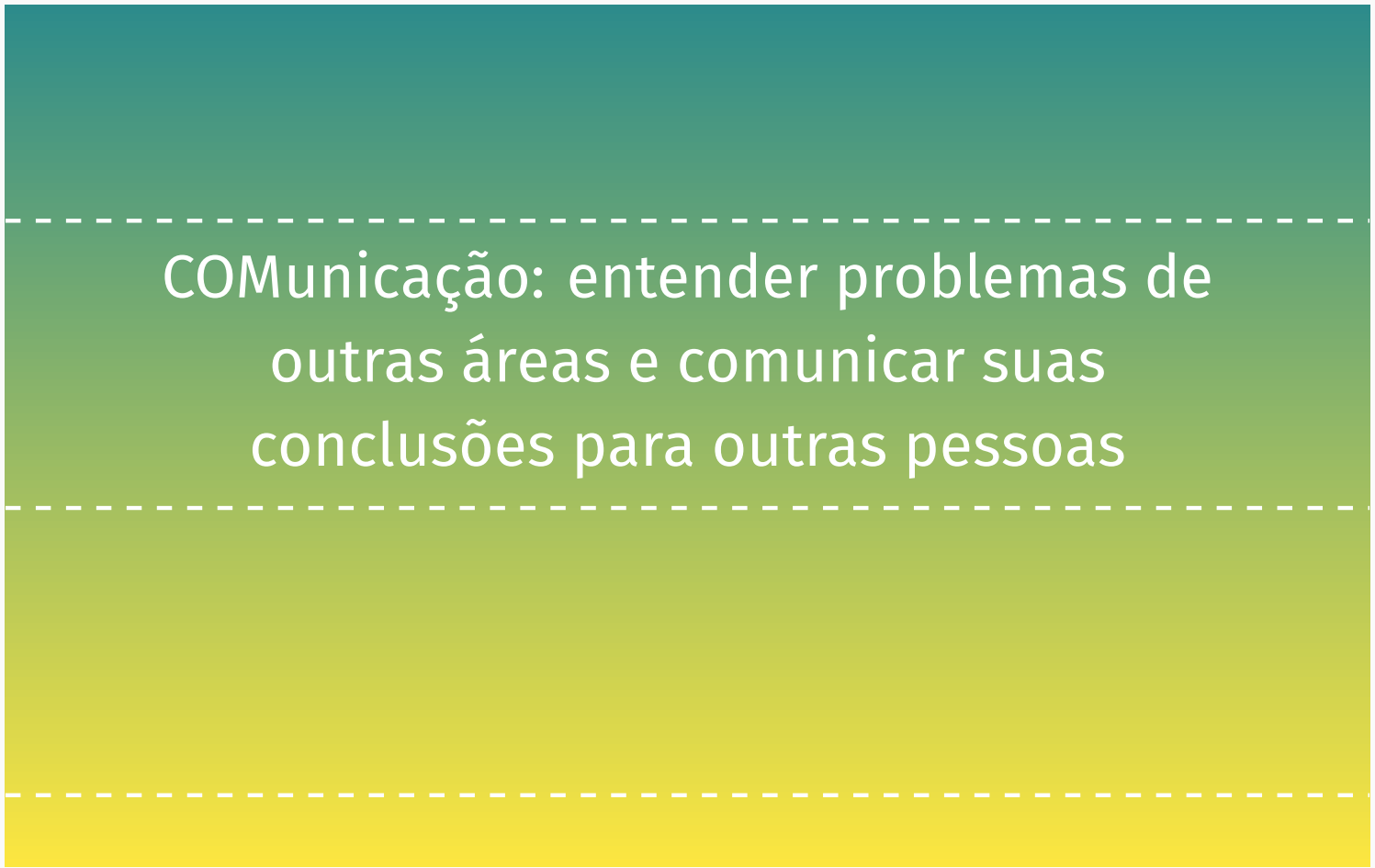
EST SIS

O que um Cientista de Dados Precisa Saber?



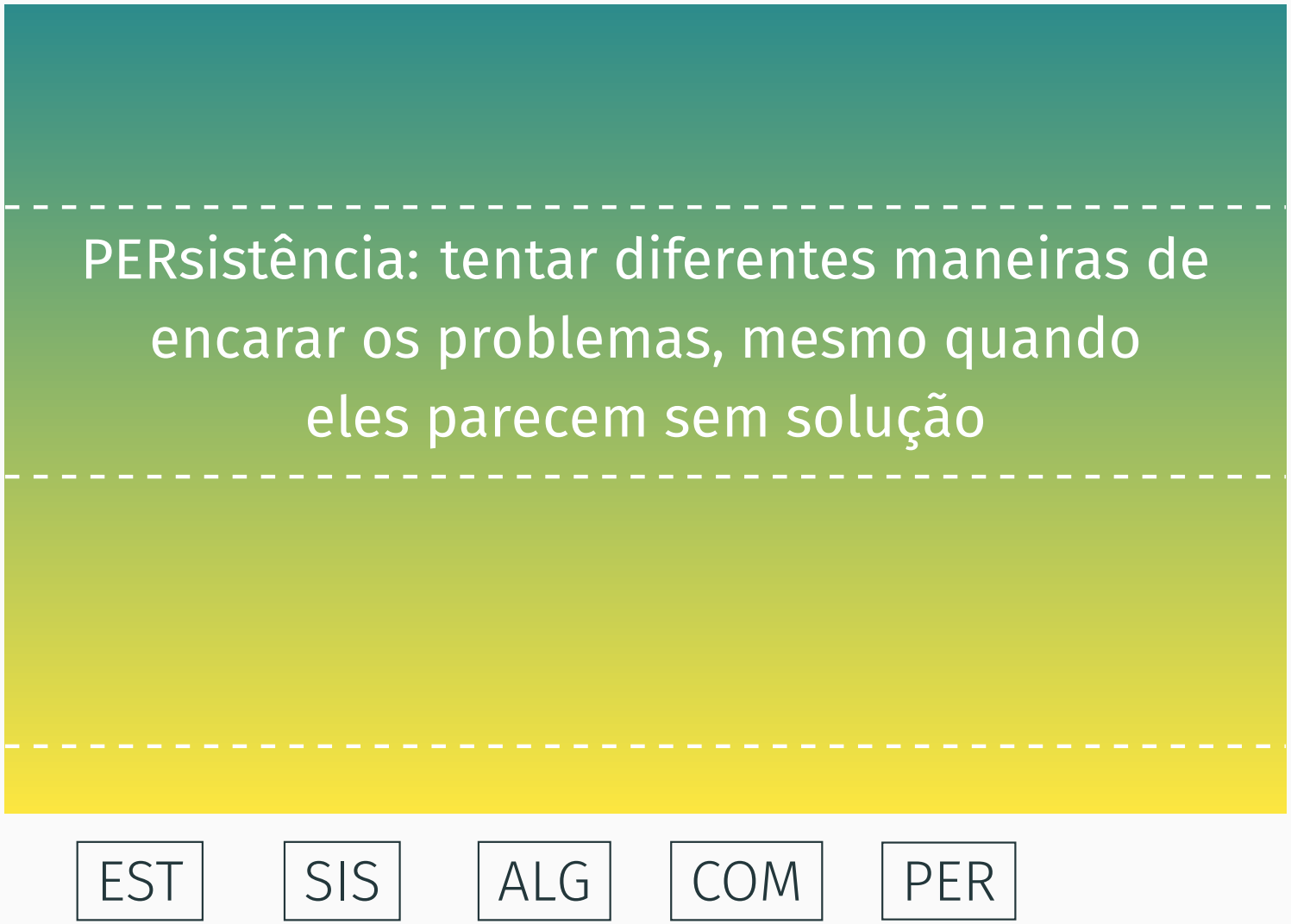
EST SIS ALG

O que um Cientista de Dados Precisa Saber?

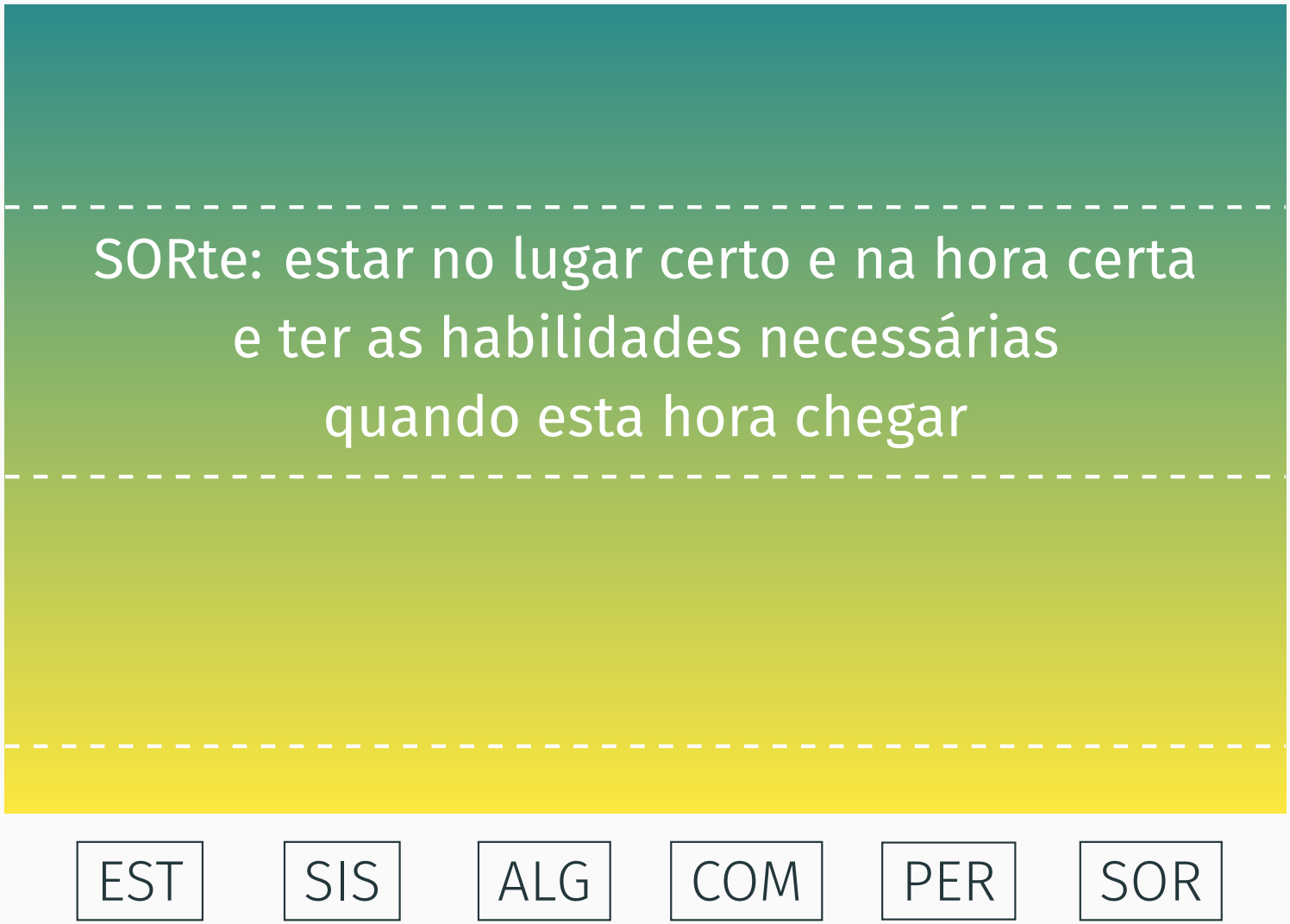


EST SIS ALG COM

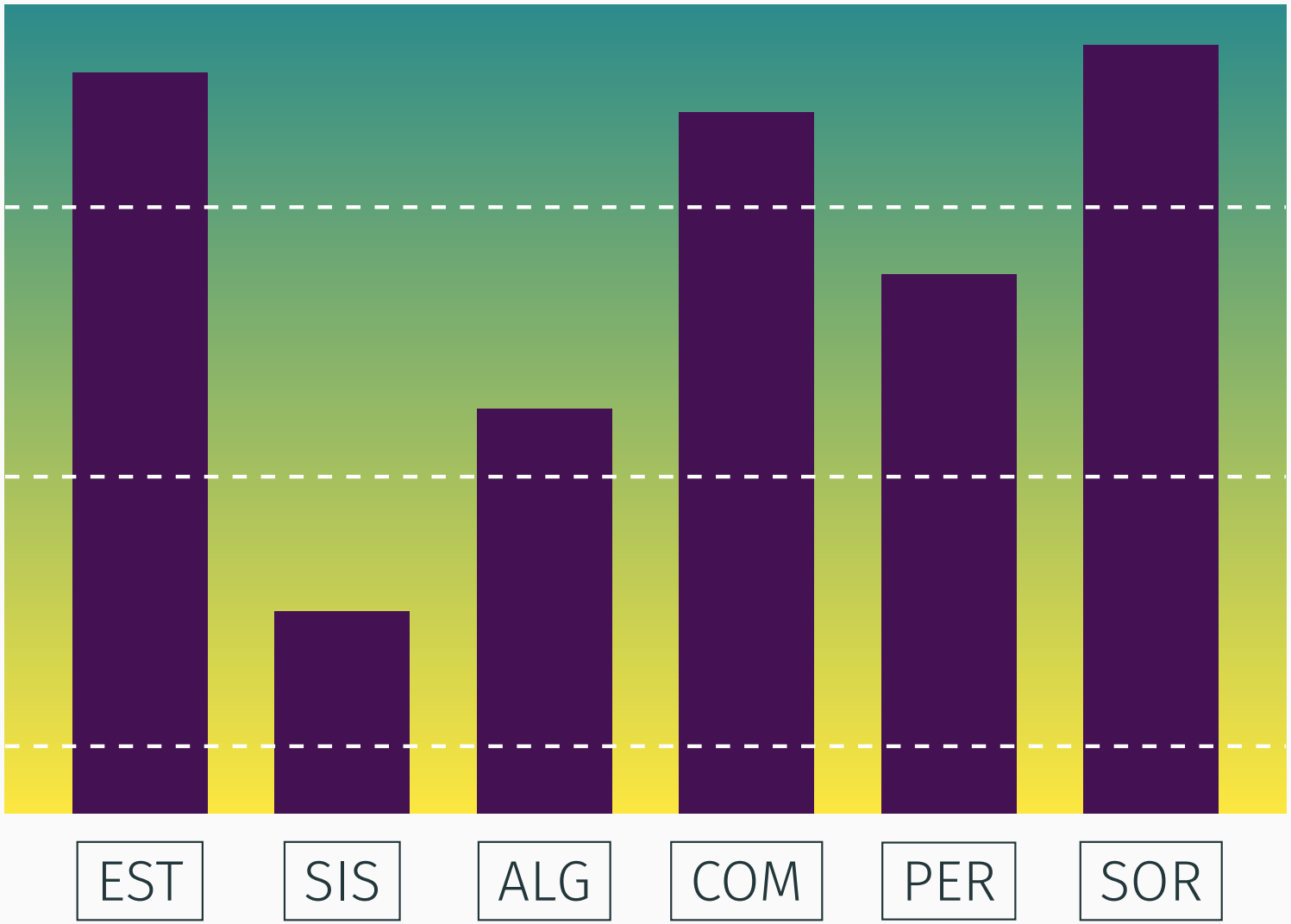
O que um Cientista de Dados Precisa Saber?



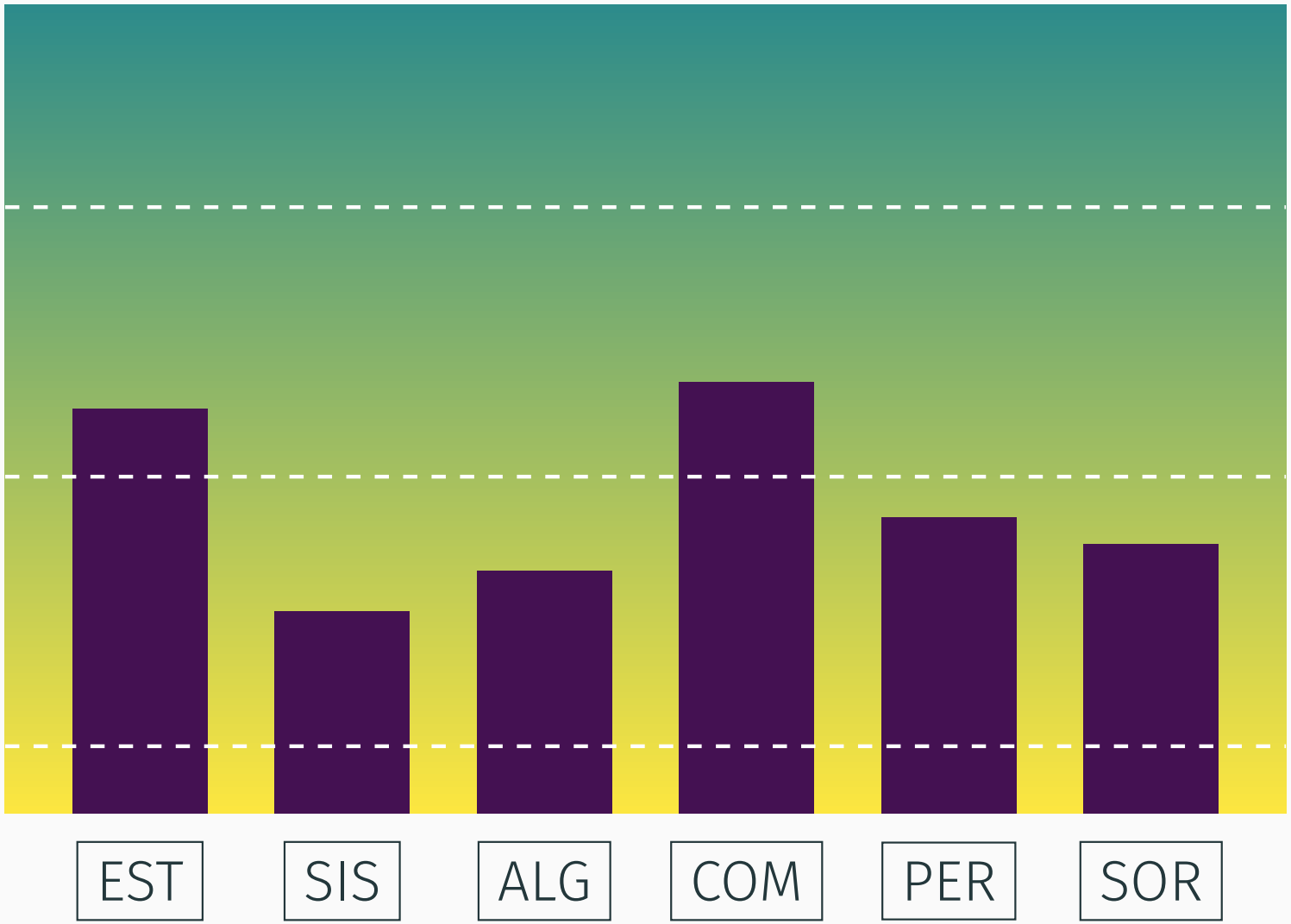
O que um Cientista de Dados Precisa Saber?



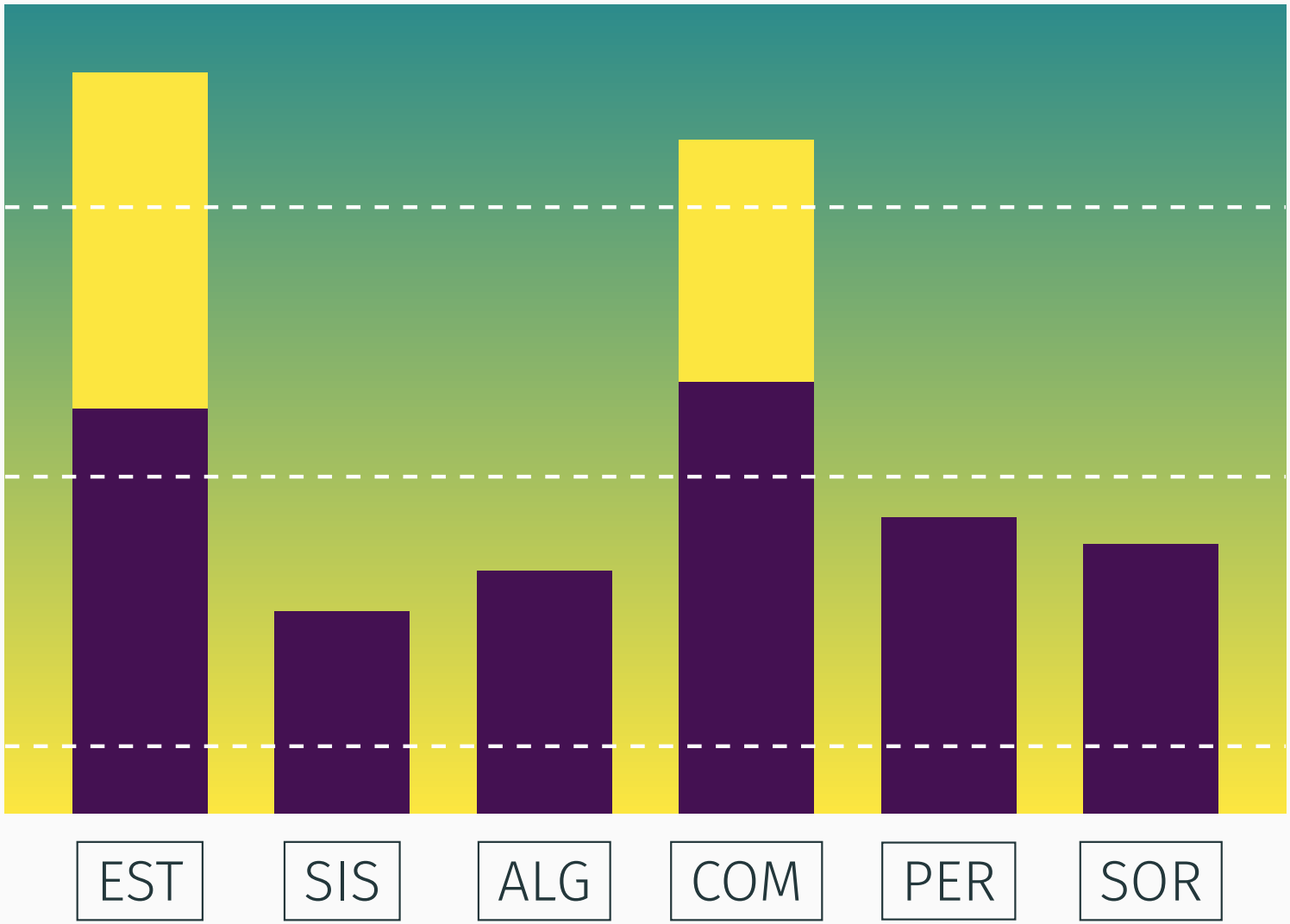
O que Eu Sei



Não é Bom Estar na Média



Seja Muito Bom em Algumas Áreas



Considerações Finais

Considerações Finais

- Dá pra fazer Ciência de Dados sozinho? Sim
- É melhor fazer em equipe? Evidentemente que sim
- Afinal, diferentes habilidades trarão conhecimentos diferentes para quem for trabalhar na área

Ciência de Dados e Multidisciplinaridade

II Seminário de Aplicações em Ciência de Dados

Marcus Nunes

7 de Dezembro de 2018

Universidade Federal do Rio Grande do Norte