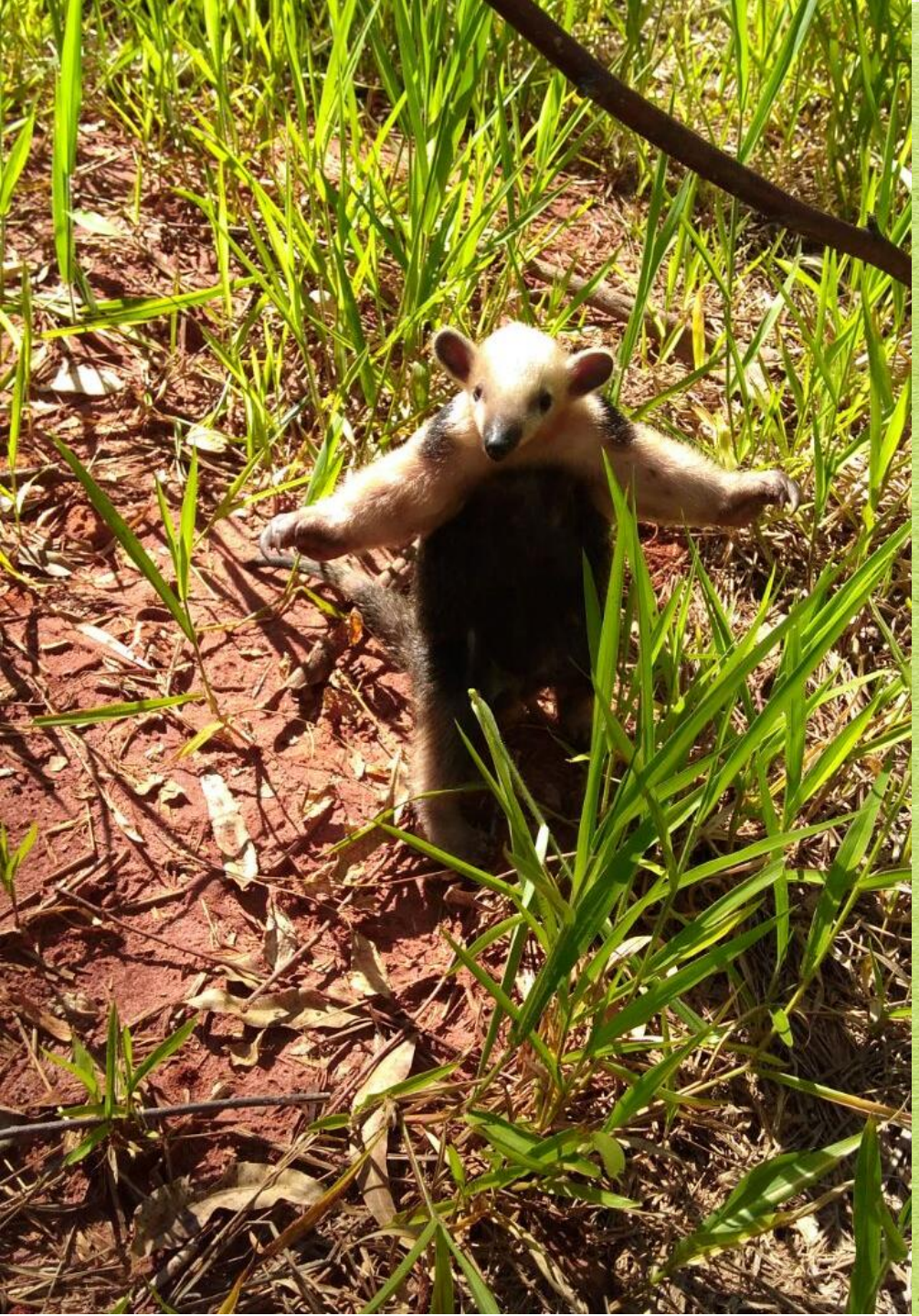
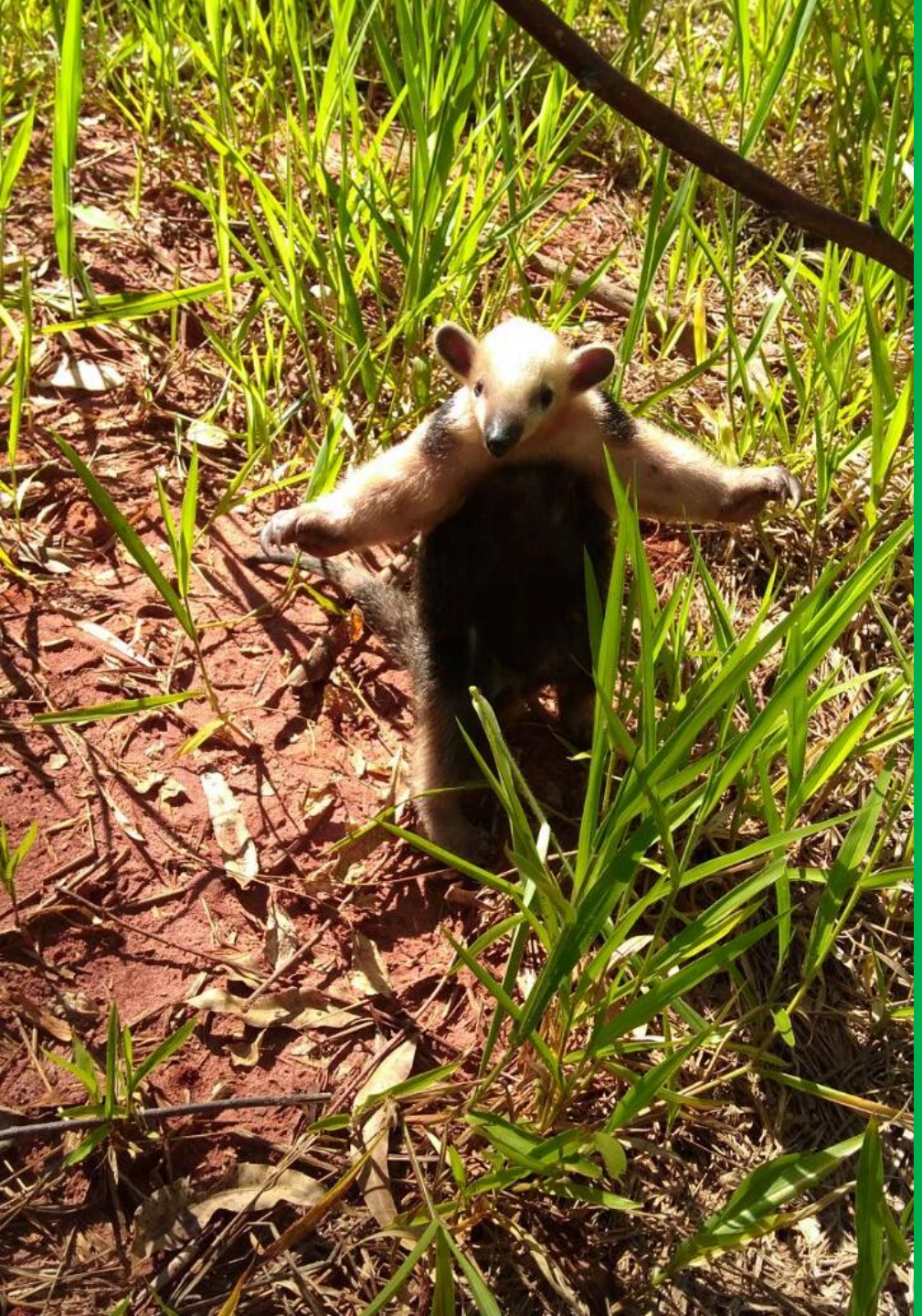


Estatística Aplicada e Análise de Dados

Eduardo Moré de Mattos

Piracicaba, 05 de março de 2022



Roteiro

O quê? (Definição)

Contextualização (Aplicações)

Como analiso meus dados?

- Definição do problema
- Respostas desejadas
- **Análise exploratória!!!**
- Modelagem
- Inferência

Prática

Como pensar sua análise

Descrição dos dados *(Fontes, Amostragem, etc.)*

Formular as hipóteses *(Quais perguntas queremos responder?)*

Que tipo de análise melhor se encaixa ao meu banco de dados para responder estas perguntas? *(NUNCA O CONTRÁRIO!)*

Análise exploratória *(Namorando o banco de dados)*

Manipulação | Data wrangling *(Padronização)*

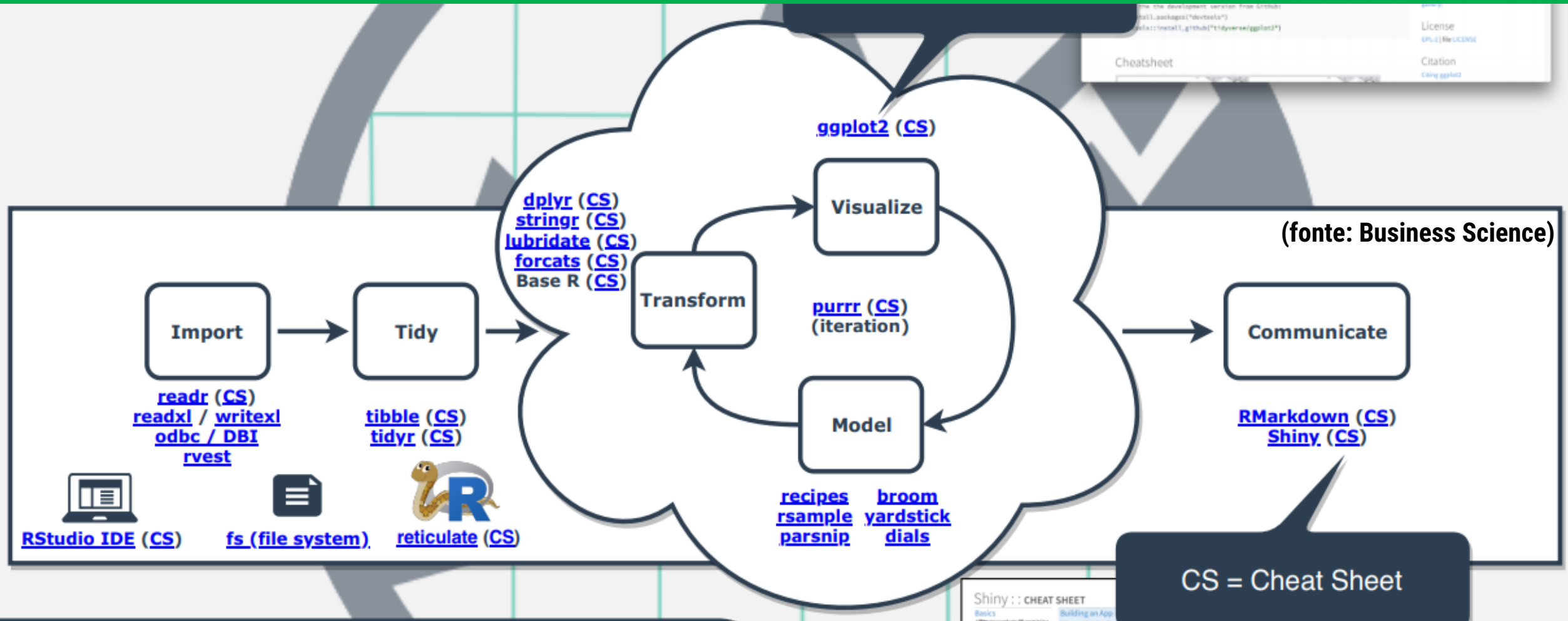
Inferindo sobre o banco de dados *(Modelagem)*

Reportando os resultados *(Reports, Visualizações, BI)*

Click baby click

https://www.youtube.com/watch?v=N1ltwg2nTK4

Workflow



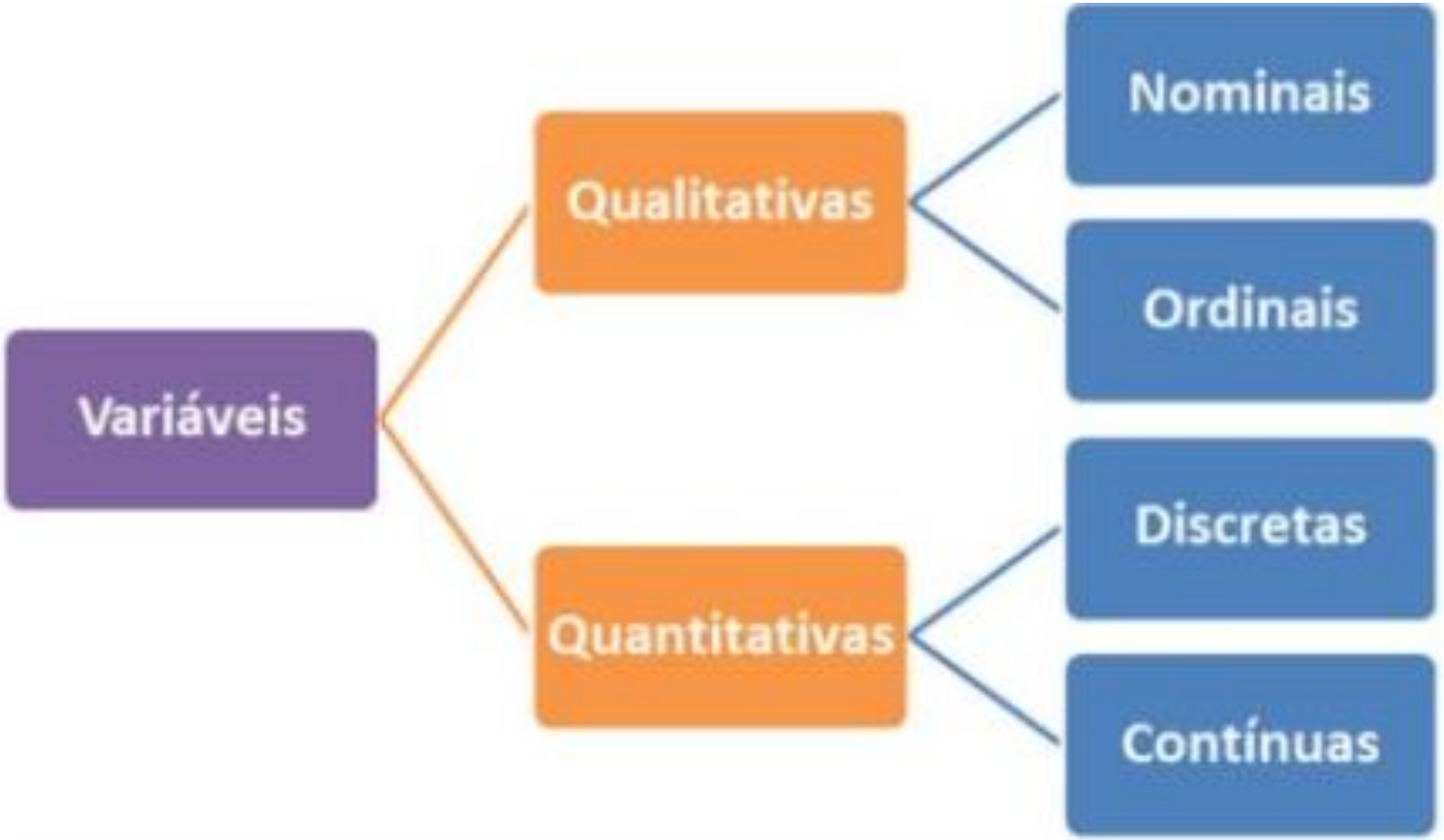
Important Resources

- **R For Data Science Book:** http://r4ds.had.co.nz/
- **Rmarkdown Book:** https://bookdown.org/yihui/rmarkdown/
- **Data Visualization Book:** https://r-graphics.org/

Um pouco de estatística...

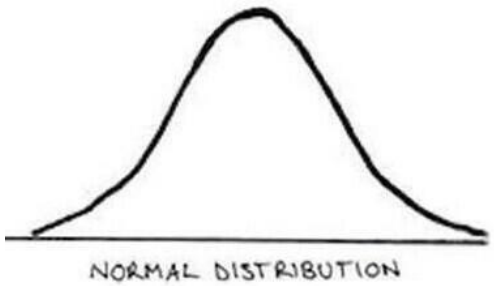
- **População e amostra**
- **Tipos de variáveis** (Contínuas, Discretas, etc.)
- **Medidas de posição e dispersão**
- **Distribuições**
- **Propriedades da distribuição normal**
- **Correlação**
- **Teste-t (uma amostras, duas amostras, pareado)**
- **Análise exploratória de dados**
    - *Sumarização (Estatísticas resumo)*
    - *Tabela de contingência*
    - *Análise gráfica (Boxplots, Histogramas, dispersão)*
    - *Gráficos em painéis*

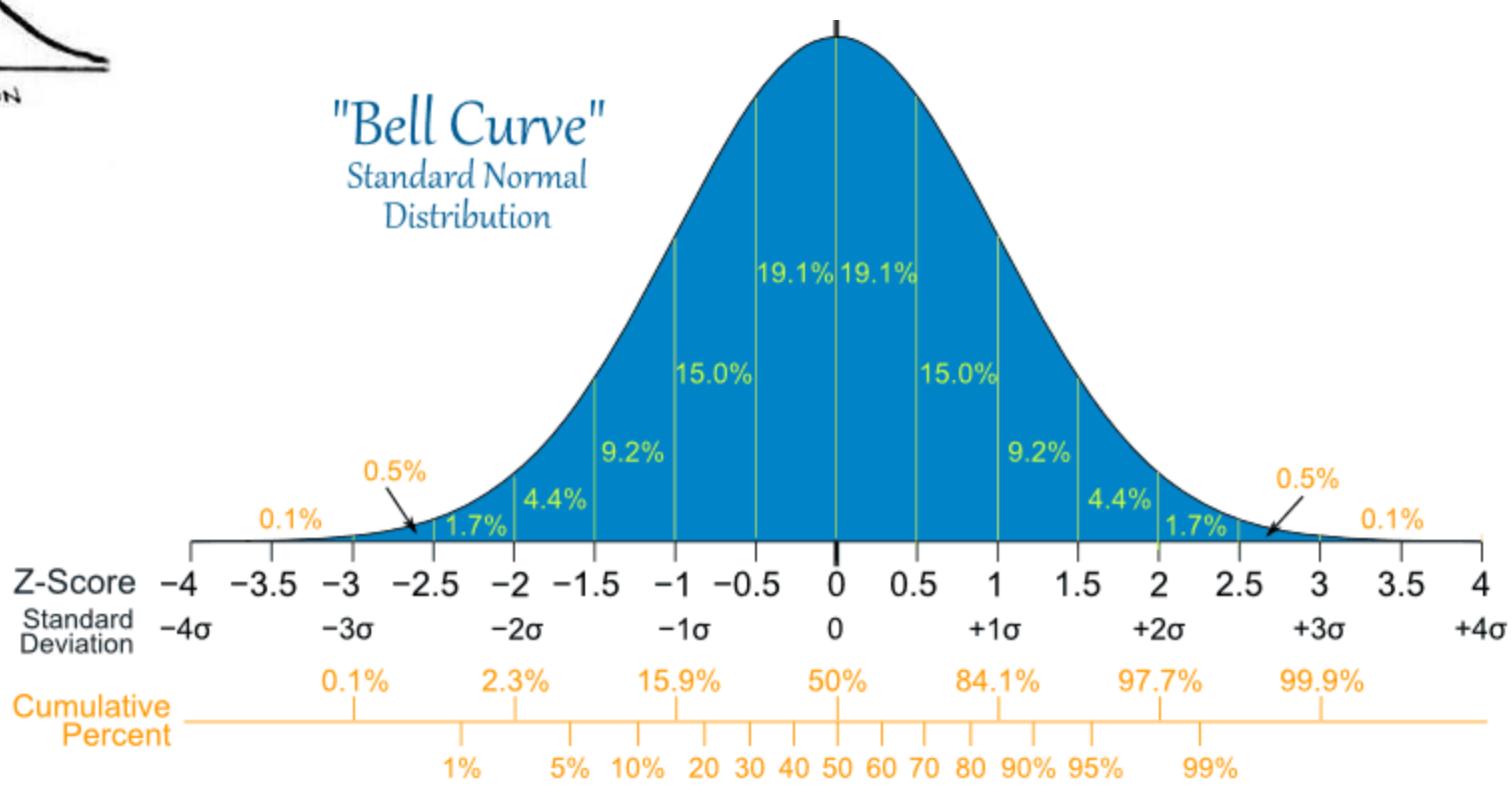
O mais importante...



Análise Exploratória | **Descritiva**

Distribuições | Histogramas

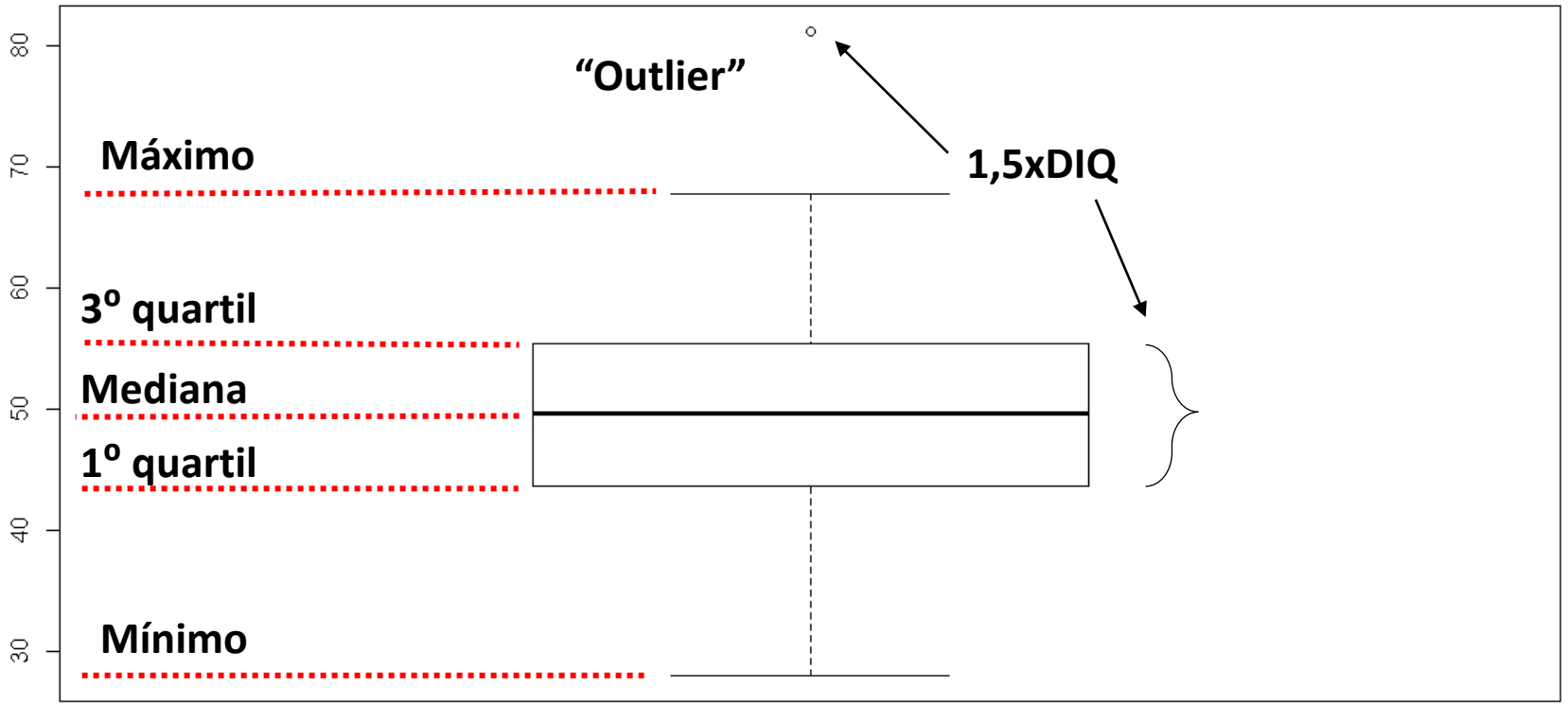




https://www.mathsisfun.com/data/standard-normal-distribution.html

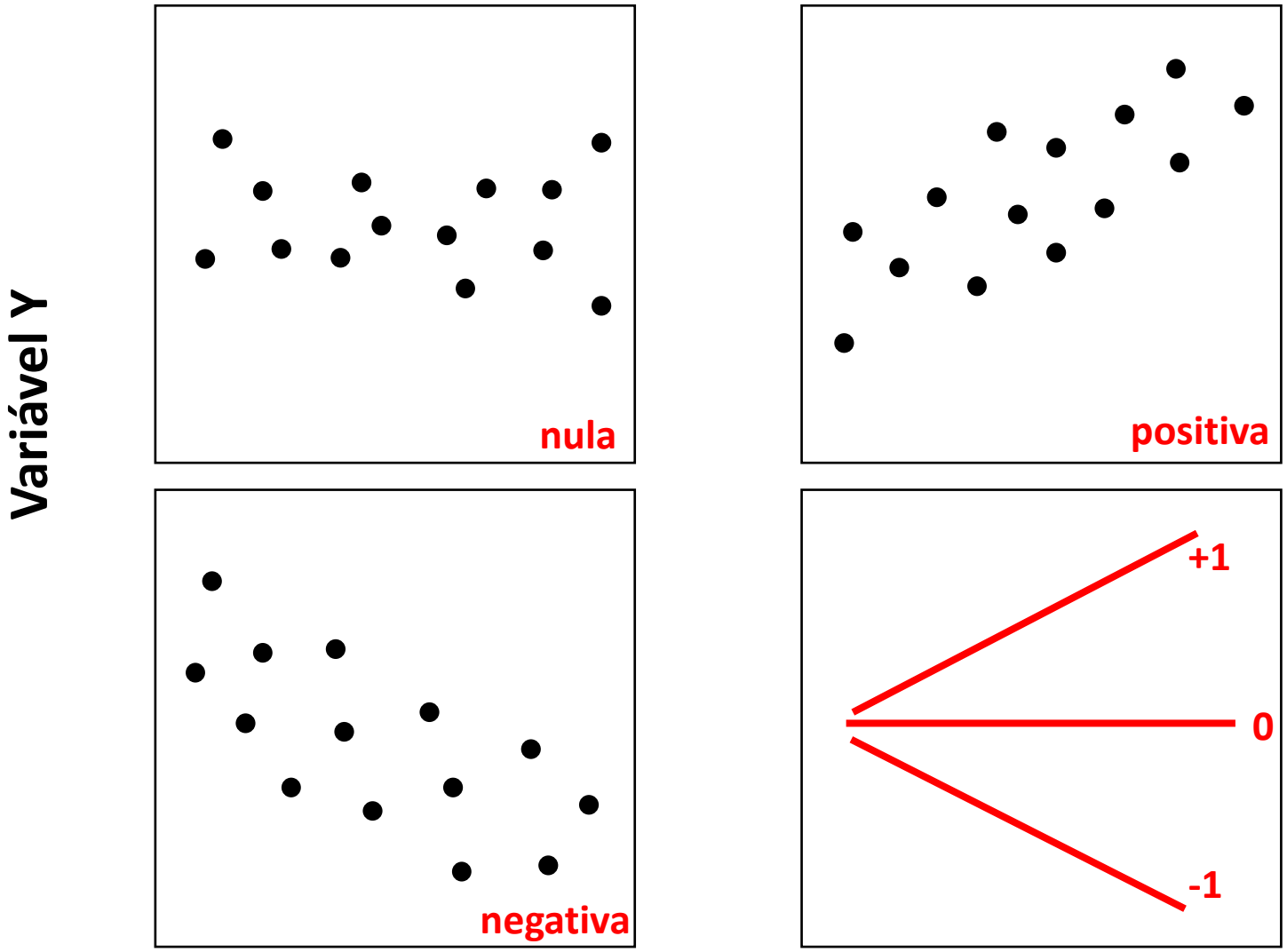
Análise Exploratória | **Descritiva**

Distribuições | Boxplots



Análise Exploratória | Descritiva

Correlação



Modelagem | Estatística e ML

Geplant
Tecnologia Florestal

Modelos Estatísticos no Mundo Real

05 - Análise Discriminante

1. Possui objetivo muito parecido com a análise de clusters. A diferença é que atua como técnica de classificação, pois os grupos (ou "clusters") são definidos "a priori", diferentemente da análise de clusters, em que os grupos são definidos "a posteriori"

06 - Modelos log-lineares

1. Bastante usados quando a resposta reflete contagens. Nesse caso, assume como um modelo de Regressão Poisson, pela natureza da distribuição da variável resposta (contagens).

07 - Regressão *Jacknife*

1. Trabalha de forma mais robusta com variáveis altamente correlacionadas, bastante usada como técnica para redução de variáveis. Ideal como técnica preditiva "caixa preta", por conta da dificuldade de interpretação dos parâmetros.

08 - Regressão Quantílica

1. Alternativa mais robusta à presença de outliers ou quando são desejadas estimativas mais precisas de diferentes quantis da variável resposta e, não apenas, da média.

01 - Regressão Linear

1. Adequada quando os dados podem ser ajustados por uma reta, resposta é quantitativa (intervalar) ou para realizar interpolações.
2. Possui algumas limitações por conta da rigidez de seus pressupostos e da instabilidade do modelo quando variáveis independentes são altamente correlacionadas.

02 - Regressão Logística

1. Bastante adequada para problemas em que a variável resposta é categórica (binária ou multinomial) e flexível para incorporar variáveis explicativas quantitativas ou categóricas.

03 - Regressão *Ridge*

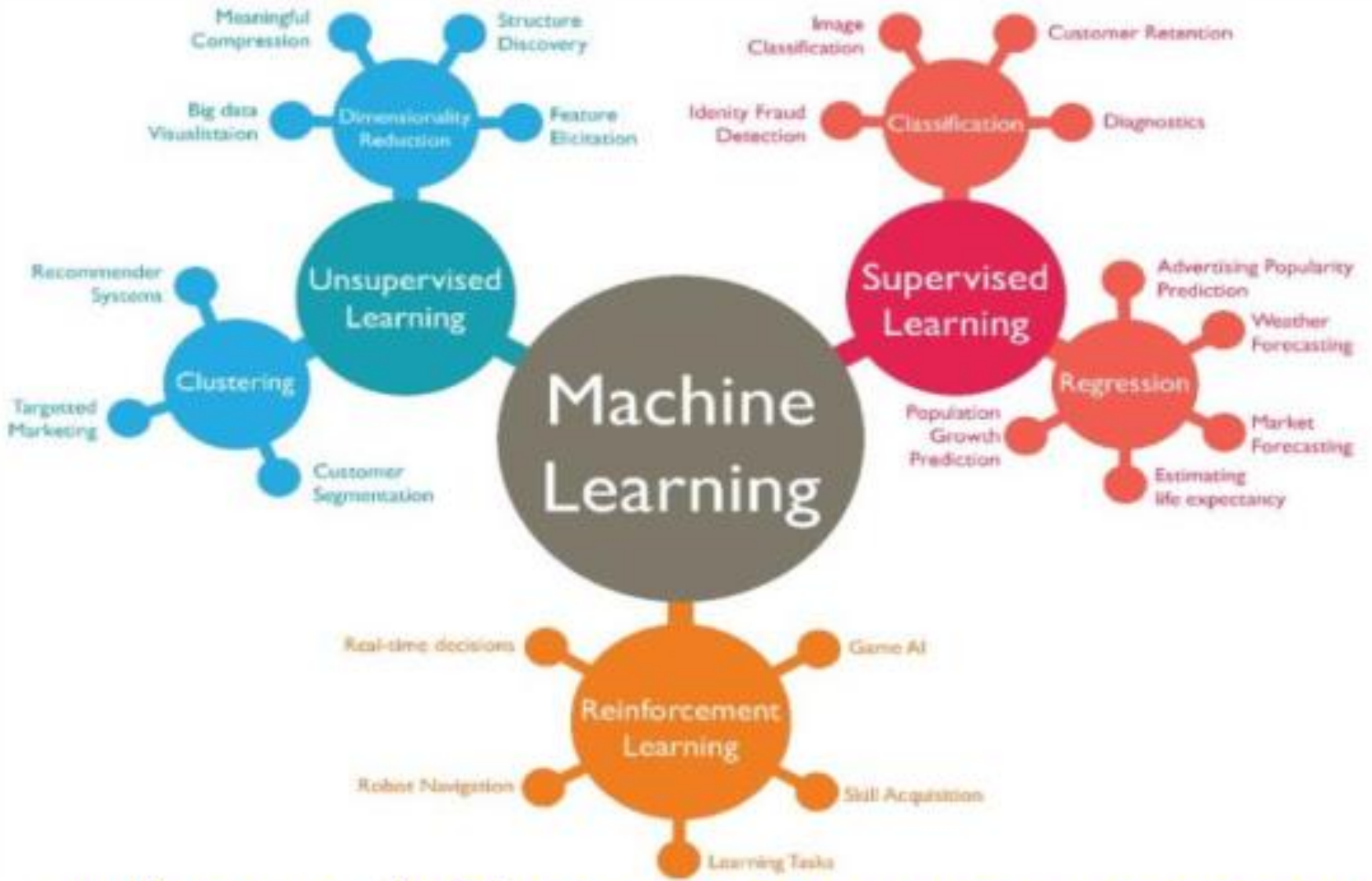
1. Versão mais robusta da regressão linear, menos sujeita a "*overfitting*", com parâmetros que sofrem restrições e são mais fáceis de serem interpretados.

04 - Regressão Bayesiana

1. É um tipo de regressão *ridge* (com restrições nos parâmetros) e cujo conhecimento sobre os coeficientes de regressão são definidos "a priori", mas que, na prática, é baseado em suposições artificiais.

fonte: Professor Marcelo Fernandes

Modelagem | Estatística e ML



https://www.kdnuggets.com/2019/04/poll-data-science-machine-learning-methods-algorithms-use-2018-2019.html

Para a vida...

SEMPRE haverá algo que pode ser feito.

Analisar dados não é submeter um conjunto de observações a um procedimento estatístico ou computacional.

A pergunta motiva o método a ser empregado.

Violar pressuposições exige conhecimento de causa.

A análise estatística é uma das ferramentas para a TOMADA DE DECISÃO, porém o uso criterioso de métodos adequados leva a melhores decisões.

Exercício 1

Banco de dados sobre qualidade de vinhos

Qualidade (0 = ruim | 1 = bom)

Manipulação

- Calcular o percentual de vinhos considerados bons na base
- Calcular o percentual de vinhos bons por tipo de vinho (Branco e Tinto)
- Comparar os grupos de vinhos (bons e ruins) com métricas de posição e dispersão
- Comparar a média do teor alcoólico entre os grupos (vinhos bons e vinhos ruins)

Visualização

- Elaborar um gráfico de barras com media do teor alcoólico por grupo (bons e ruins)
- Construir um histograma para avaliar a distribuião do teor de açúcar dos vinhos
- Avaliar a correlação entre teor de álcool e teor de açúcar para os vinhos, considerando:

- Todos os vinhos, grupos (bons e ruins), grupos (tintos e brancos) e interações

Exercício 2

Banco de dados meteorológicos da ESALQ

Médias mensais de 1917 a 2017

Manipulação

- Importar o banco de dados para o R
- Verificar e ajustar o banco de dados
- Criar uma coluna para identificar os anos que não estão completos
- Sumarizar a precipitação por ano

Visualização

- Avaliar a flutuação das temperaturas em Piracicaba
- Construir um histograma da distribuição de chuvas
- Criar um gráfico mostrando a variação anual da precipitação –

Destacar no gráfico a linha da média anual

Obrigado!

Eduardo Moré de Mattos

eduardo@geplant.com.br