PUCRS online

APRESENTAÇÃO DE APOIO

INTRODUÇÃO À CIÊNCIA DE DADOS E À INTELIGÊNCIA ARTIFICIAL

Por Renan Xavier Cortes - Aula 01

Pós-Graduação em

Ciência de Dados e Inteligência Artificial

Ementa da disciplina

Fundamentos de Ciência de Dados: um pouco de história; as disciplinas; data x big data. Estado da arte: academia e indústria. Estudo de conceitos relacionados à Ciência de Dados: mentalidade orientada a dados, inferência estatística, análise de dados exploratória, o processo da ciência de dados, os processos de negócios x ciência de dados. Visão geral sobre algoritmos: regressão, similaridade, vizinhos, agrupamentos.

Pensamento analítico: o que são bons modelos, visualização. Aplicações. Fundamentos de Inteligência Artificial: um pouco de história; as disciplinas; a noção de inteligência. Visão Geral das Áreas da IA: Representação do Conhecimento, Raciocínio e Planejamento; Agentes e Sistemas Multiagentes; Robótica; Machine Learning. Conceitos de Processamento da Língua Natural. Conceitos de Visão Computacional.

Professores

RENAN XAVIER CORTES
Professor convidado

Trabalha como Data Science Specialist no banco Agibank e possui Pós-Doutorado pelo Center for Geospatial Sciences na Universidade da Califórnia. Como profissional, tem atuado no time de Data Science Corporativo do banco em projetos multidisciplinares auxiliando diversos departamentos internos abrangendo Fraudes, Jurídico, Recursos Humanos, Customer Relationship Management (CRM), Crédito, etc. Atua principalmente na exploração de bases de dados para construção e implantação de modelos preditivos, automatização de processos, construção e gestão de dashboards. Seus principais interesses abrangem análise e visualização de dados, modelagem estatística, open source (R, Python, Shiny, RStudio, Jupyter, etc.), desenvolvimento de software e Big Data. No Pós-Doutorado, trabalhou sob a supervisão de Serge Rey desenvolvendo novos métodos de inferência para segregação não espacial e espacial, interpolação espacial e visualização de dados espaciais, tornando-o core developer do Python Spatial Analysis Library (PySAL). Possui doutorado em Economia pela PUCRS, mestrado em Estatística pela UFMG e graduação em estatística pela UFRGS.

MICHAEL MORA
Professor PUCRS

Possui graduação em Ciência da Computação pela Universidade Federal do Rio Grande do Sul (1991), mestrado em Computação pela Universidade Federal do Rio Grande do Sul (1993) e doutorado em Ciência da Computação pela Universidade Federal do Rio Grande do Sul (2000). Atualmente é professor adjunto do Instituto de Informática. Tem experiência na área de Ciência da Computação, com ênfase em Inteligência Artificial, atuando principalmente nos seguintes temas: Inteligência Artificial, Aprendizagem de Máquina, Agentes inteligentes e Sistemas Multiagentes, Engenharia de Software e Desenvolvimento de Sistemas, Ensino de Programação e de Ciência da Computação.

Professores

SILVIA MORAES

Professora PUCRS

Professora da Faculdade de Informática (FACIN) da Pontifícia Universidade Católica do Rio Grande do Sul (PUCRS), desde agosto de 1997. Ela também é um dos professores colaboradores do Grupo de Pesquisa em Linguagem Natural da PUCRS. Obteve título de doutora em Ciência da Computação PPGCC / PUCRS com a tese "Construção de Estruturas Ontológicas a partir de Textos: um Estudo Baseado no método Formal Concept Analysis e em Papéis Semânticos", em 2012. Anteriormente, ela cursou mestrado em Ciência da Computação no PPGC / UFRGS de 1994 a 1997. Obteve o título de bacharel em Informática também na FACIN / PUCRS em 1992. Seus principais interesses de pesquisa estão relacionados ao processamento de linguagem natural: mineração de texto, a categorização de texto, agrupamento de texto, aprendizagem automática, aprendizagem de ontologias, extração de conceitos, análise de sentimentos, agentes conversacionais, etc.

Encontros e resumo da disciplina

AULA 1

Um cientista de dados não necessariamente vai trabalhar com somente com dados que estão estruturados.

O nosso cérebro está muito pronto para ser surpreendido por visualizações.

É interessante olhar os dados sob as multifacetas possíveis e estar ciente das limitações dessas visualizações.

RENAN XAVIER CORTES
Professor convidado

AULA 2

A inteligência artificial nasce junto com a computação.

Podemos usar a inteligência artificial para resolver problemas complexos de uma forma mais interessante.

Uma boa heurística é indispensável neste universo tão complexo.

MICHAEL MORA
Professor PUCRS

AULA 3

Os paradigmas de aprendizado dependem muito da natureza e do problema que se trabalha.

O aprendizado supervisionado é uma tarefa preditiva.

A deep learning foi a principal responsável por avanços de linguagem natural.

SILVIA MORAES
Professora PUCRS

Introdução à Ciência de Dados e Inteligência Artificial

Renan Xavier Cortes

Uma aula introdutória focando mais em Ciência de Dados

PUCRS online

UOL edtech
TECNOLOGIA PARA EDUCAÇÃO

Antes de Começar... Quem é esse Renan?

Academia

B.S.

M.S.

Ph.D.

Pós-Doutorado

Experiências

Interesses

Cronograma

- Fundamentos de Ciência de Dados
  - Contexto Histórico
  - Características e Principais Ferramentas de um Cientista de Dados
  - Data vs. Big Data
- Os processos de Ciência de Dados
  - Processo de Negócio
  - Processo Operacional (*qual o fluxo operacional de um Data Scientist?*)
- Alguns Casos de Uso

Cronograma

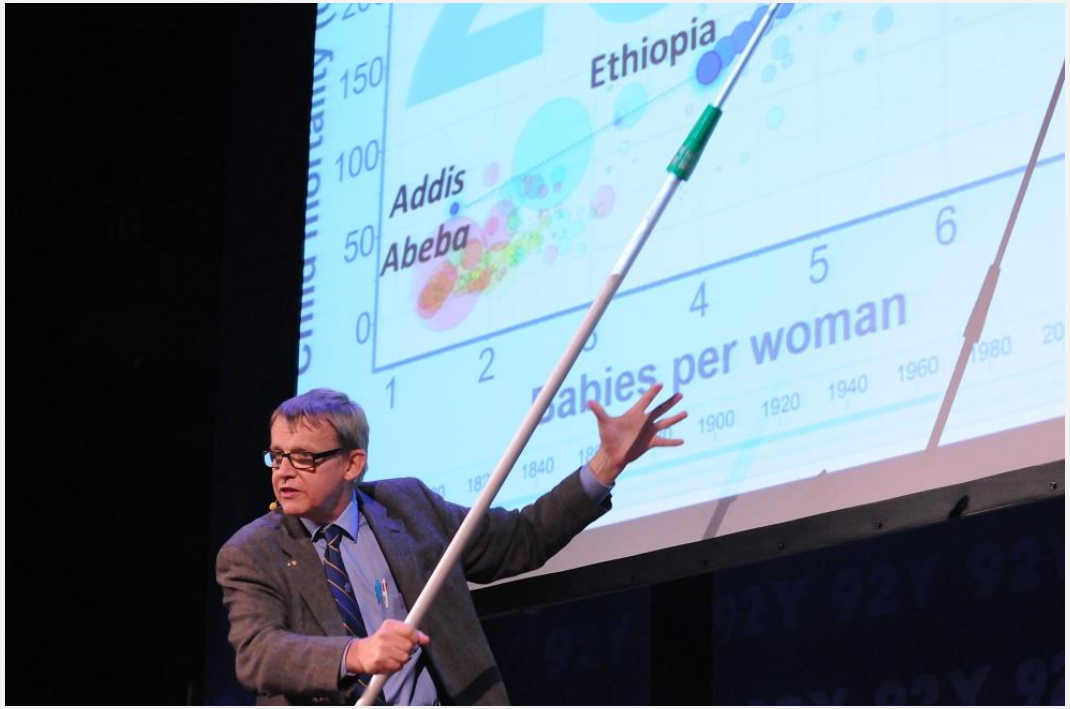
- Visão Geral sobre algoritmos
  - Algoritmos Supervisionados e Não-Supervisionados
  - Alguns dos Principais Algoritmos
  - Underfitting e Overfitting
    - Motivação
    - Estratégias de Validação

- Hora de sujar um pouco as mãos...
  - Exemplo prático com Análise Exploratória e Visualização de Dados
  - Estado da Arte em Machine Learning (ML): AutoML

Contexto Histórico

"the sexiest job in the next 10 years will be statisticians." **– Varian, Hal (2009). Economista-chefe do Google de 2002 até o presente.**

Contexto Histórico



"*Statistics is now the sexisest job around*" – **Rosling, Hans. Documentário da BBC em 2011.**

Contexto Histórico

Outubro de 2012

Thomas H. Davenport e D.J. Patil

Data Scientist: The Sexiest Job of the 21st Century

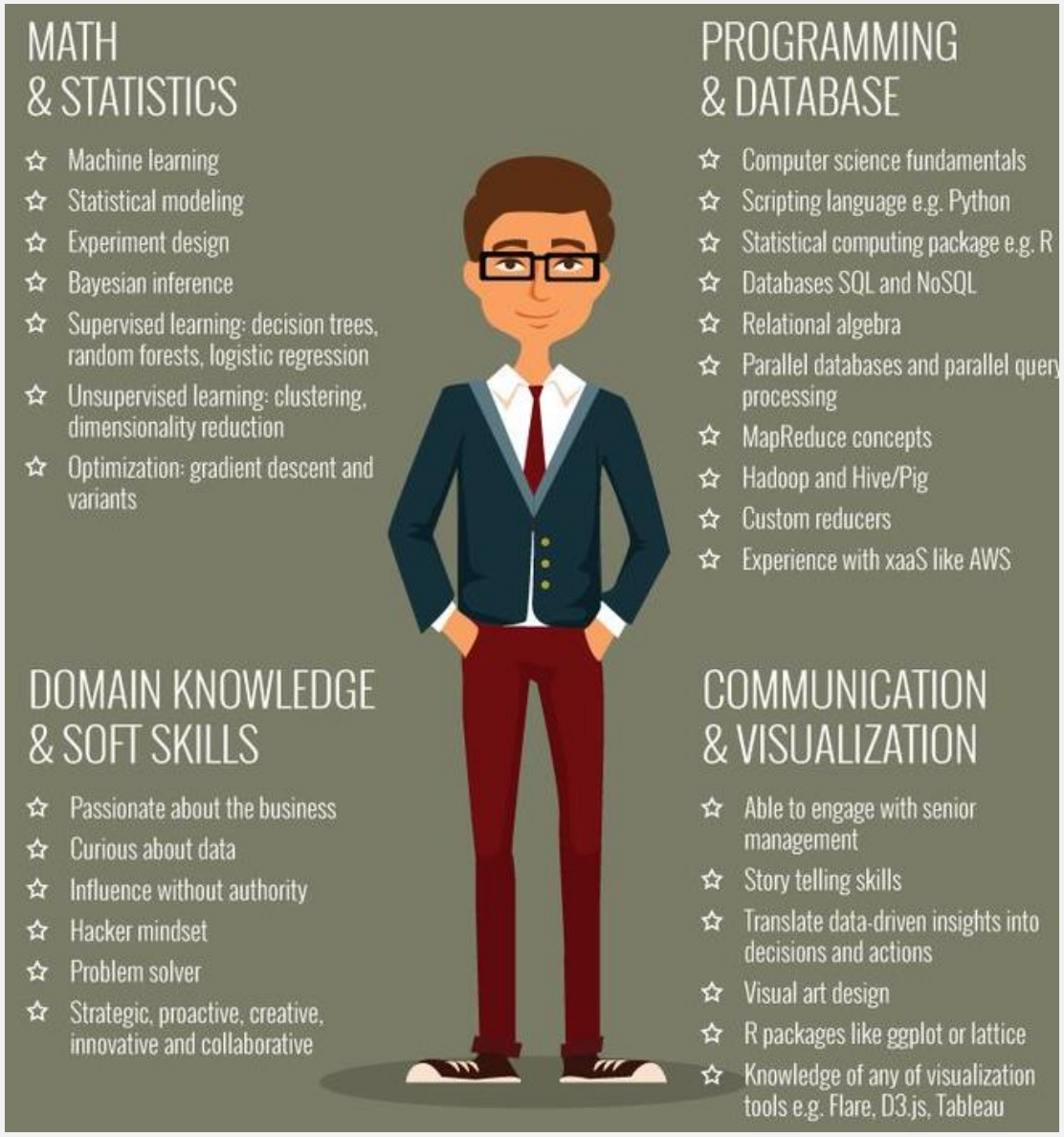
"Cientistas de Dados hoje são como os quants de Wall Street nas décadas de 80 e 90."

Contexto Histórico

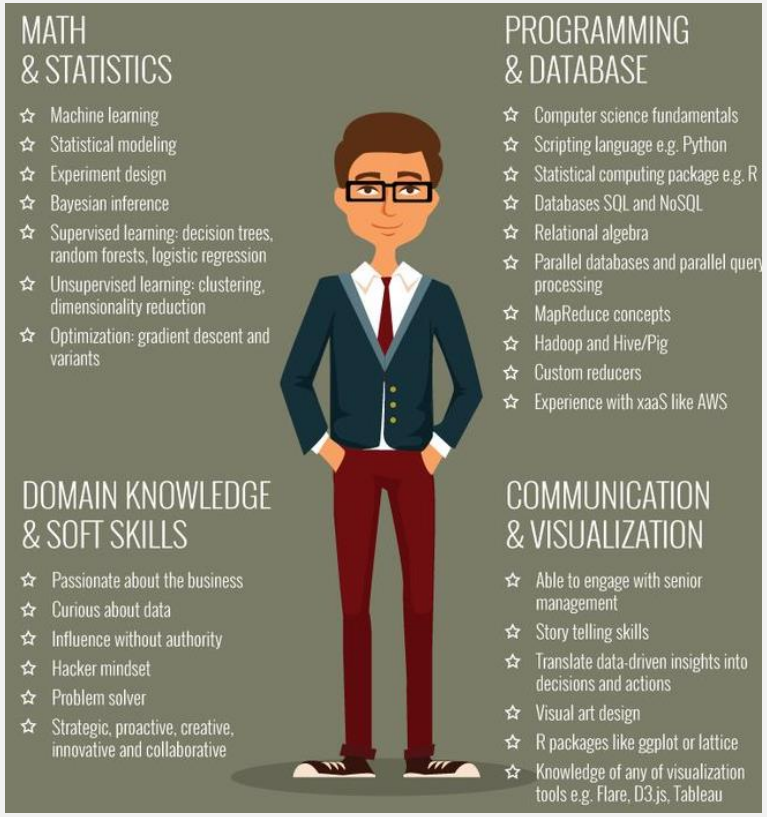
D.J. Patil

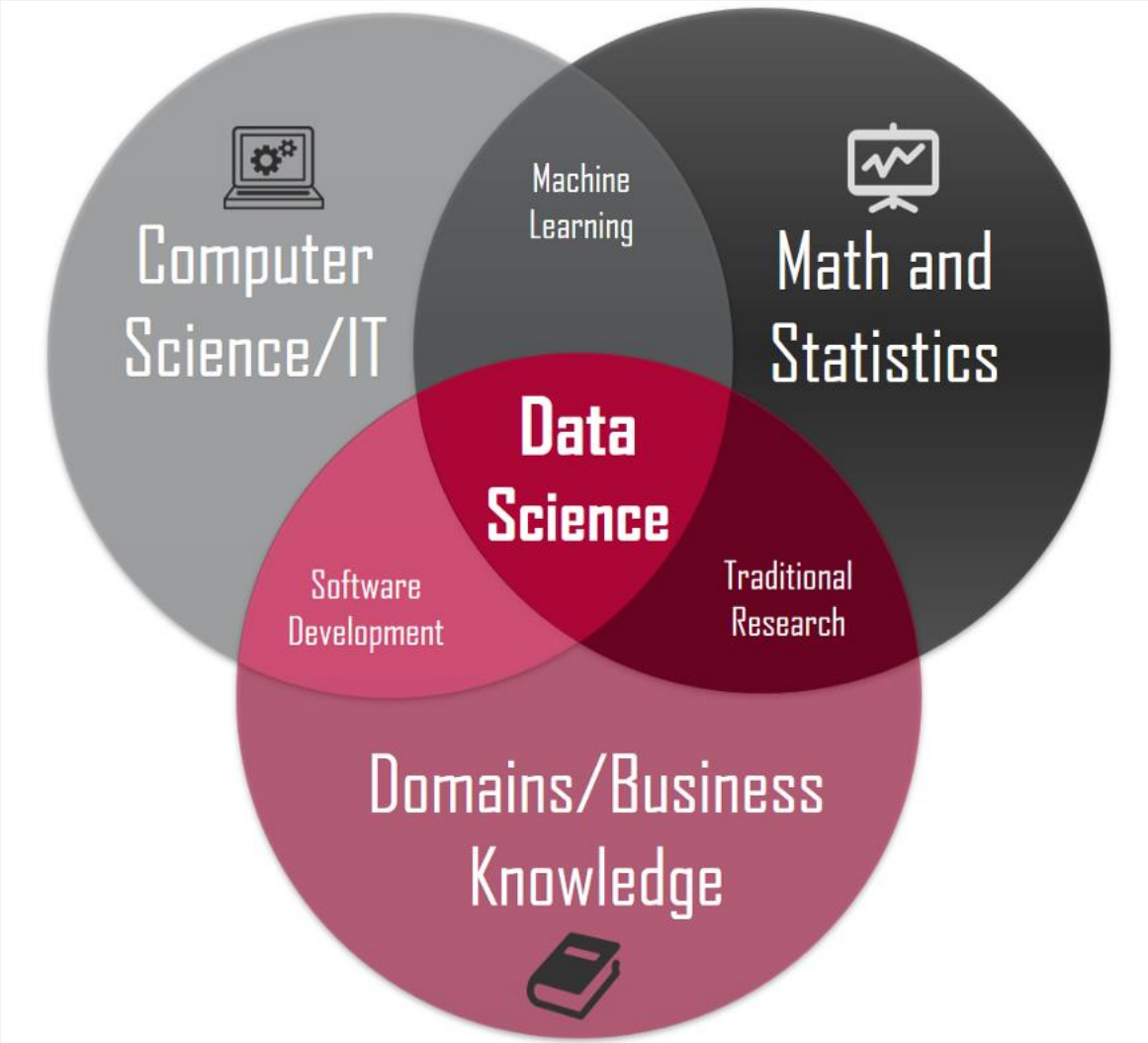
Former U.S. Chief Data Scientist

Características de um Data Scientist



Características de um Data Scientist



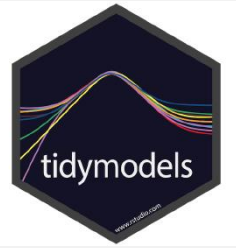


Exemplos de ferramentas de um Data Scientist

Extração de Dados e Manipulação de Dados

Visualização de Dados

Modelagem



Comunicação

Mas... Existe esse profissional no Mercado?

"Não existe espaço para heroísmo em Data Science" **– Crepalde, Neylson (2020). Data Scientist e Head de MLOps da A3Data.**

E... todas as empresas sabem como procurar?

Sebastián Ramírez
@tiangolo

I saw a job post the other day. 📪

It required 4+ years of experience in FastAPI. 💪

I couldn't apply as I only have 1.5+ years of experience since I created that thing. 😅

Maybe it's time to re-evaluate that "years of experience = skill level". ♻️

Traduzir Tweet

10:40 AM · 11 de jul de 2020 · Twitter Web App

45,3 mil Retweets **3 mil** Tweets de comentário **178,2 mil** Curtidas

Data vs. Big Data

- **Big Data: "*estudo e aplicações de dados que são muito complexos para software de processamento de dados tradicionais*" (Wikipedia).**

- **Será que precisamos de novas ferramentas para lidar com esse novo cenário?**


SCIENCE \ US & WORLD \ TECH

Excel spreadsheet error blamed for UK's 16,000 missing coronavirus cases

The case went missing after the spreadsheet hit its filesize limit

By James Vincent | Oct 5, 2020, 9:41am EDT

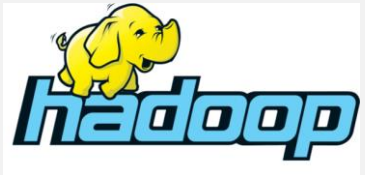
Fonte: https://www.theverge.com/2020/10/5/21502141/uk-missing-coronavirus-cases-excel-spreadsheet-error

Data vs. Big Data

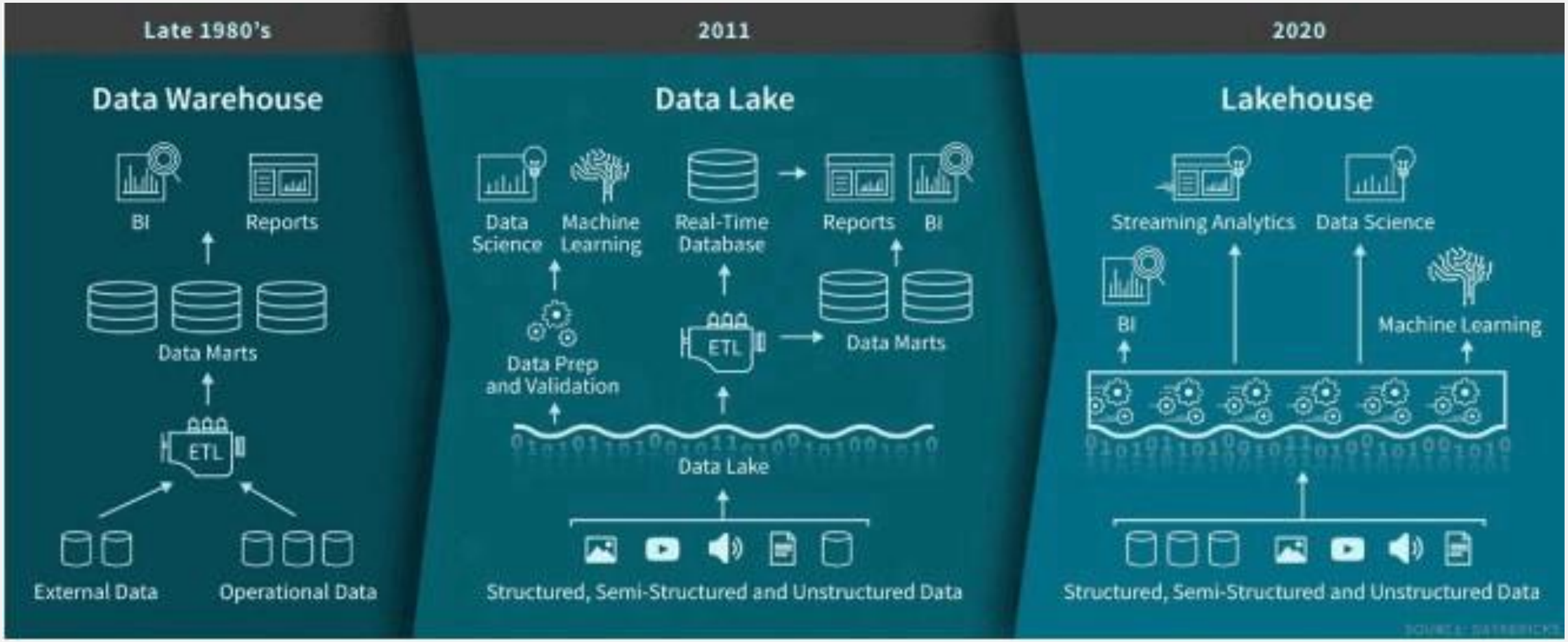
- **Os 3 V's (ou 5) do Big Data:**
  - **Volume**
  - **Velocidade**
  - **Variedade**
  - **Valor**
  - **Veracidade**

- **Qual a infraestrutura para lidar com essa quantidade de dados?**

- **Novo paradoxo:**



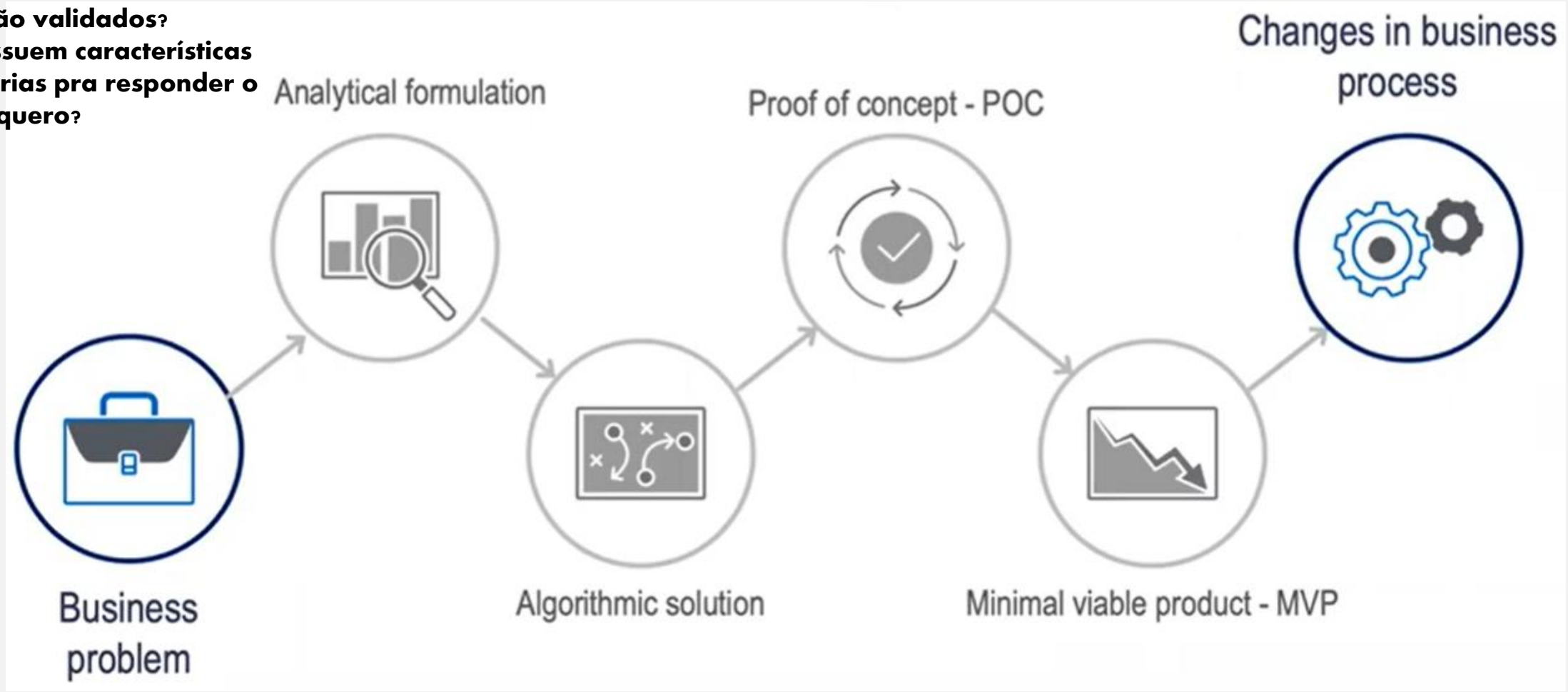
Data vs. Big Data



- **Lembre-se de que "*não existe heroísmo em Data Science*".**

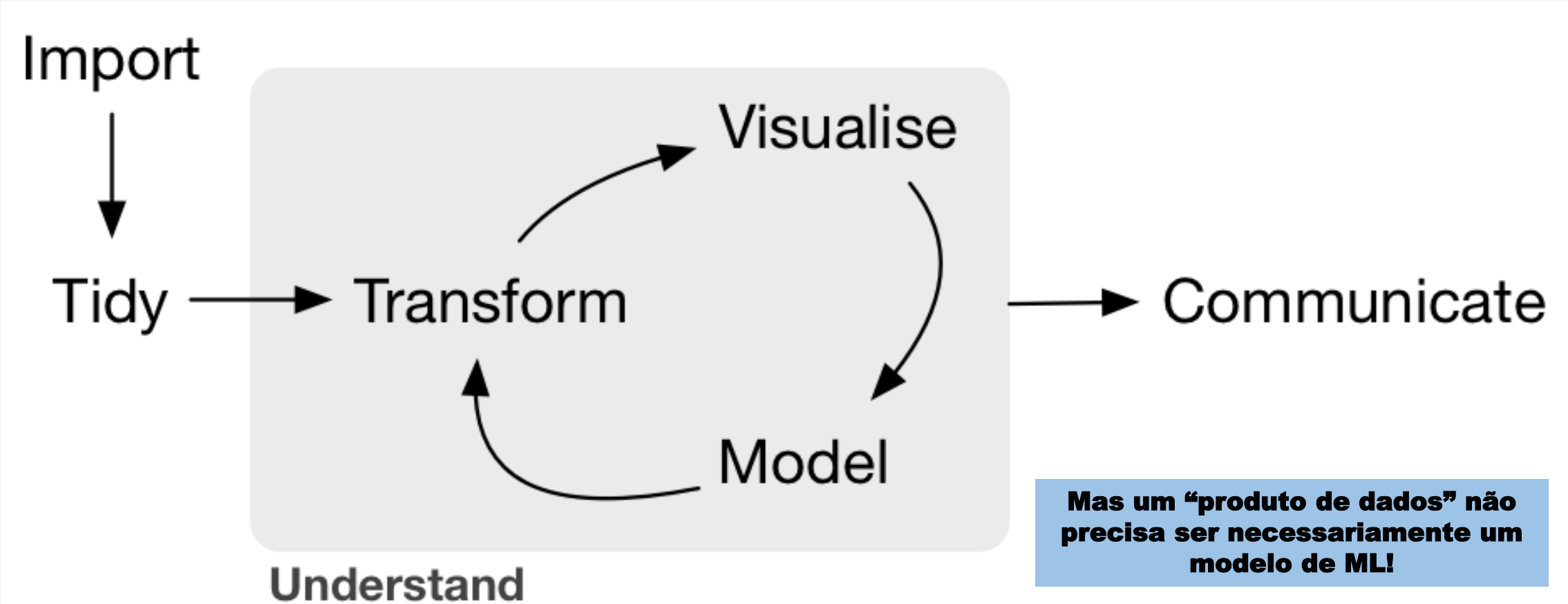
Processo de Negócio em Data Science

Os dados são confiáveis?
Eles estão validados?
Eles possuem características necessárias pra responder o que eu quero?



Fonte: *https://www.youtube.com/watch?v=oseHCwDBfSE&list=PLriUvS7IljvlcLnrvYUyNc9nXhiM9kWjq*

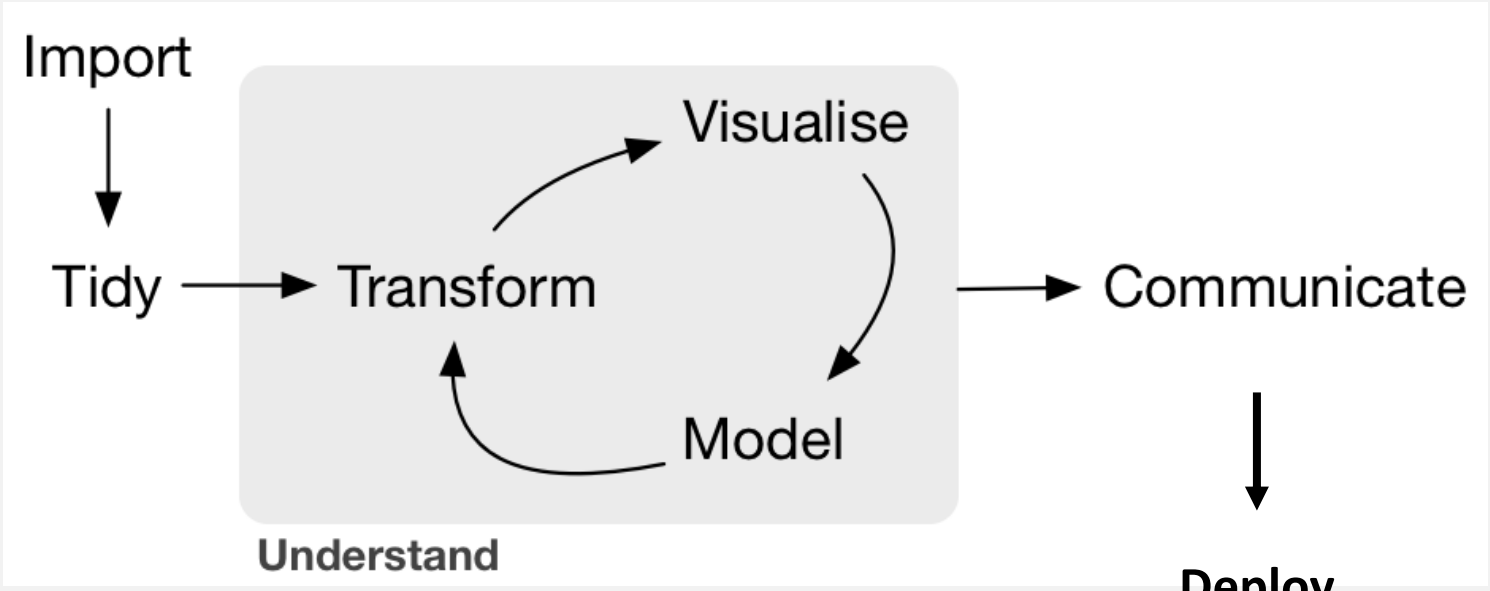
... E o processo operacional?



Mas um "produto de dados" não precisa ser necessariamente um modelo de ML!

Fonte: **Wickham, Hadley, and Garrett Grolemund.** *R for data science: import, tidy, transform, visualize, and model data. " O'Reilly Media, Inc.", 2016.*

... E o processo operacional?



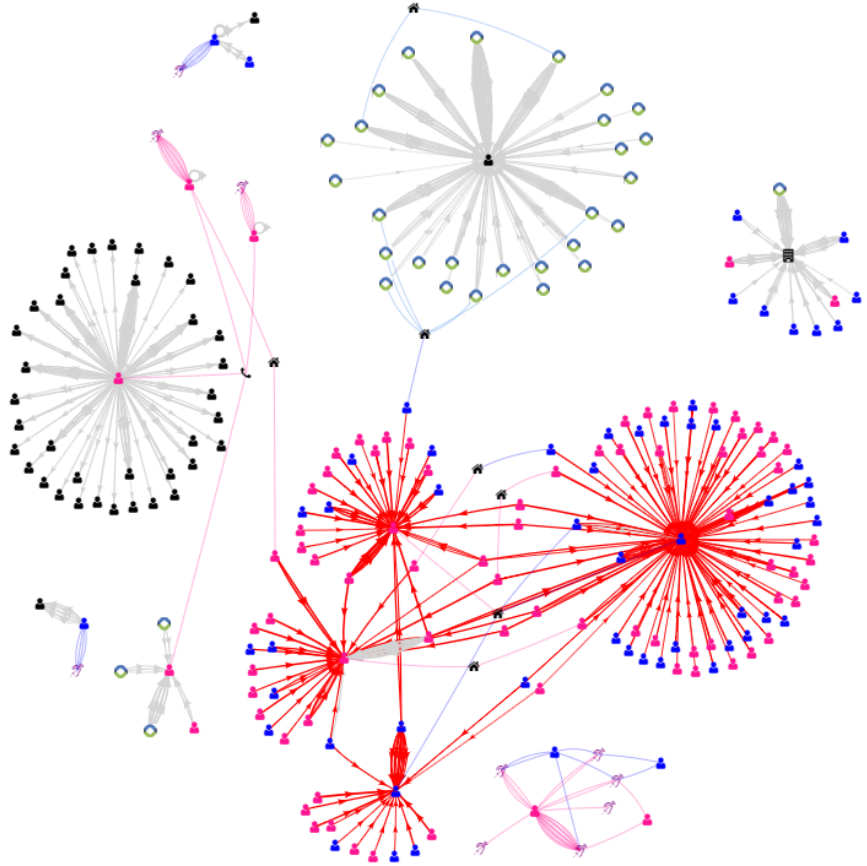
"Visualizations surprises you, but do not scale. Model scales but do not surprises you" - **Wickham, Hadley**

PUCRS online

UOL edtech
TECNOLOGIA PARA EDUCAÇÃO

Alguns Casos de Uso: Redes de Relacionamento de Fraudes

- **Motivação: identificar padrões de transferências entre pessoas/organizações.**
- **Solução: Ferramenta de Exploração Rápida das redes.**
- **Desafios: conexão de múltiplas tabelas de um Data Lake e volumetria alta de transações.**

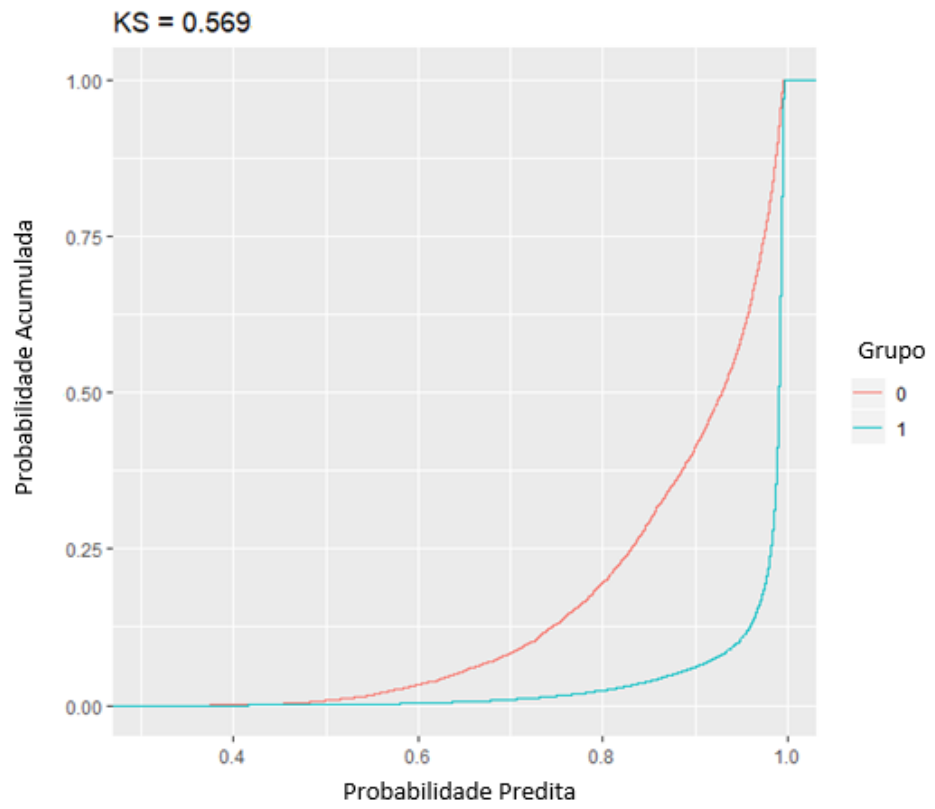


Alguns Casos de Uso: Modelo de Cobrança

- **Motivação: identificar clientes mais propensos a negociar e pagar dívidas.**

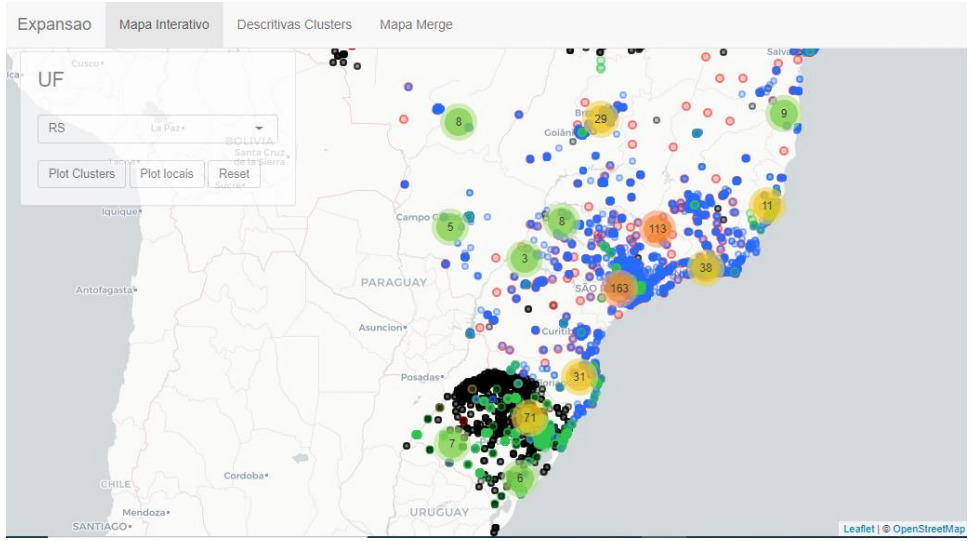
- **Solução: modelo de Machine Learning que prevê a probabilidade de um cliente pagar.**

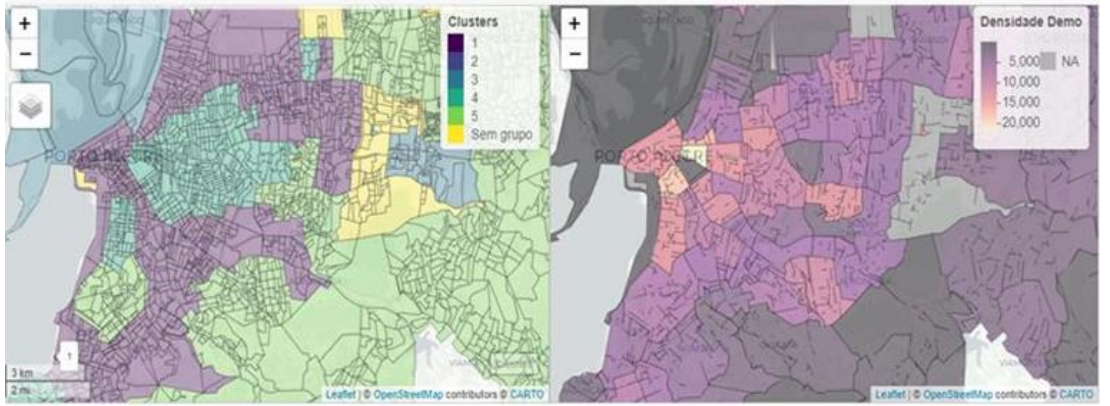
- **Desafios: conexão de múltiplas tabelas e tratamento de datas de passado e futuro para evitar "*data leakage*".**



Alguns Casos de Uso: Mapas de Expansão e Clusters

- **Motivação: auxiliar o processo de expansão de lojas e qualificação de divulgações.**
- **Solução: ferramenta que plota informações geolocalizadas junto com clusters regionais.**
- **Desafios: harmonização de dados do IBGE com arquivos espaciais (*shapefiles*).**



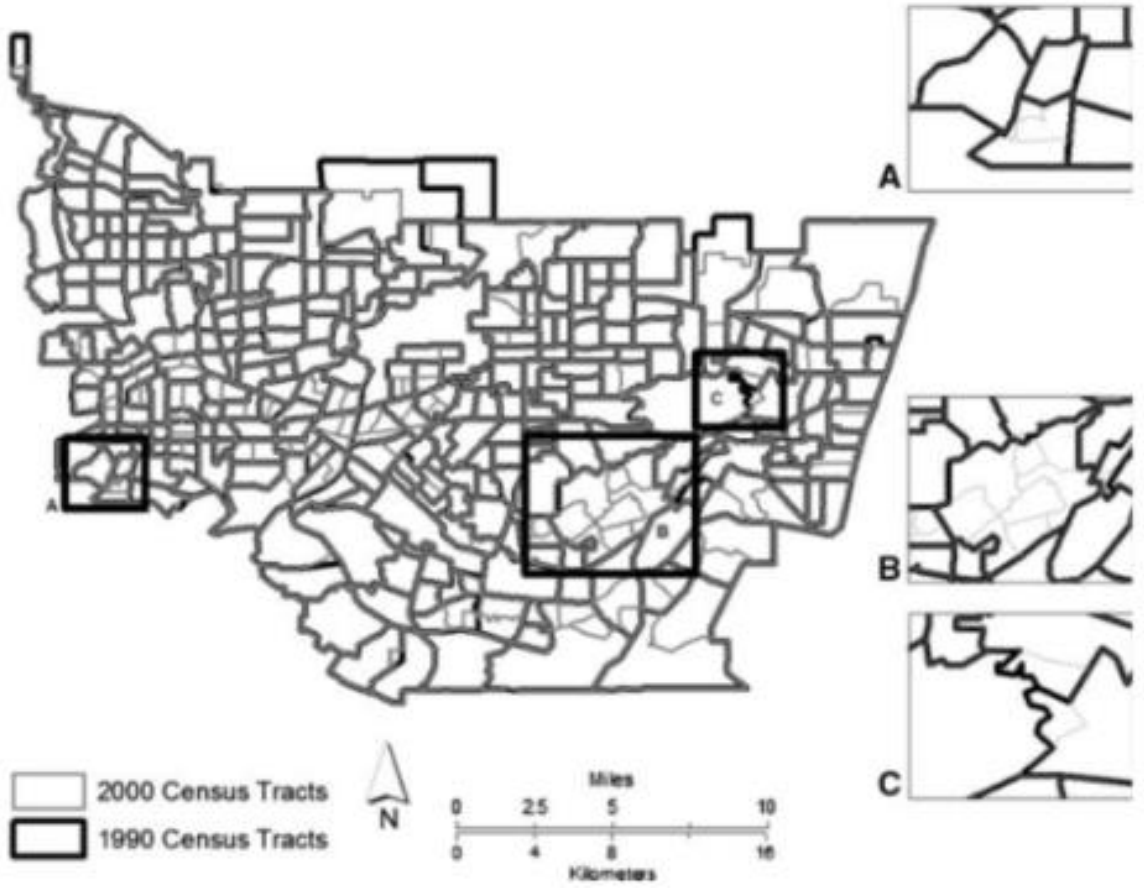


PUCRS online

UOL edtech
TECNOLOGIA PARA EDUCAÇÃO

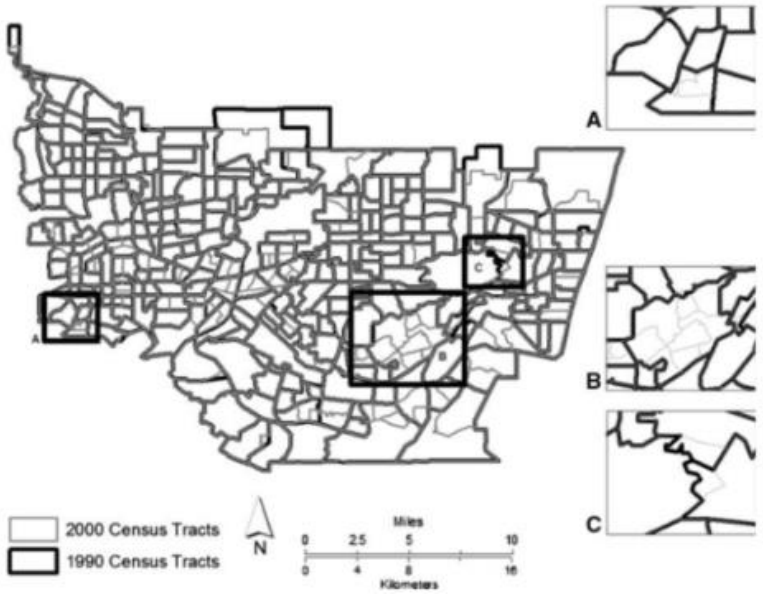
Alguns Casos de Uso: Harmonização/Interpolação Espacial

- **Motivação: limites não-harmonizados entre diferentes anos**



Fonte: *"Areal interpolation of population counts using pre-classified land cover data." Population Research and Policy Review 26.5-6 (2007): 619-633.*

Alguns Casos de Uso: Harmonização/Interpolação Espacial



➢ **Possível Solução:**

Areal Weights

- W_i = areal weight for intersected feature i
- A_i = area of intersected feature i
- A_j = total area of source feature j

$$W_i = \frac{A_i}{A_j}$$

- E_i = estimated value for intersected feature i
- V_j = population value for source feature j

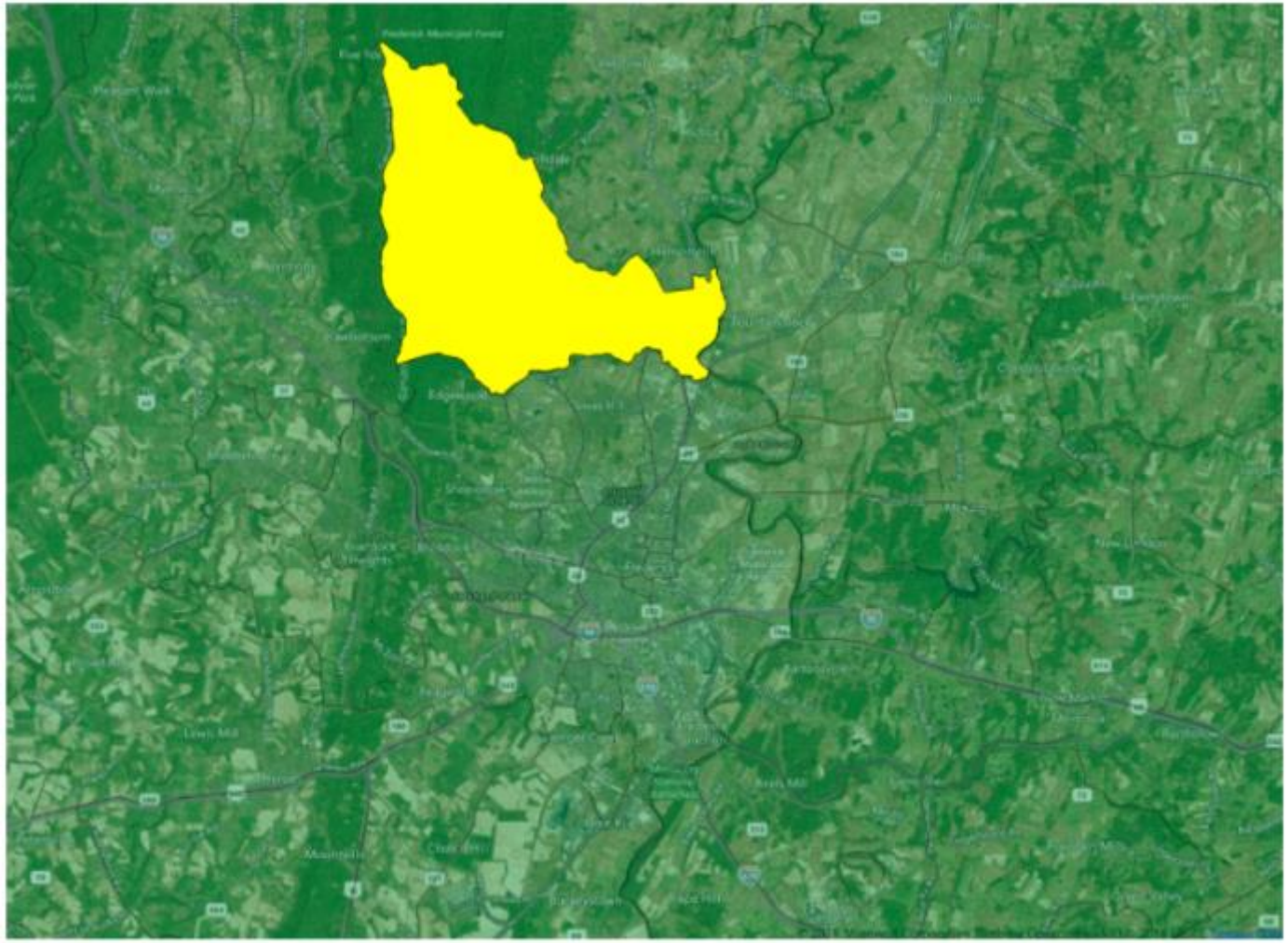
$$E_i = V_j * W_i$$

Alguns Casos de Uso: Harmonização/Interpolação Espacial

Limitação:

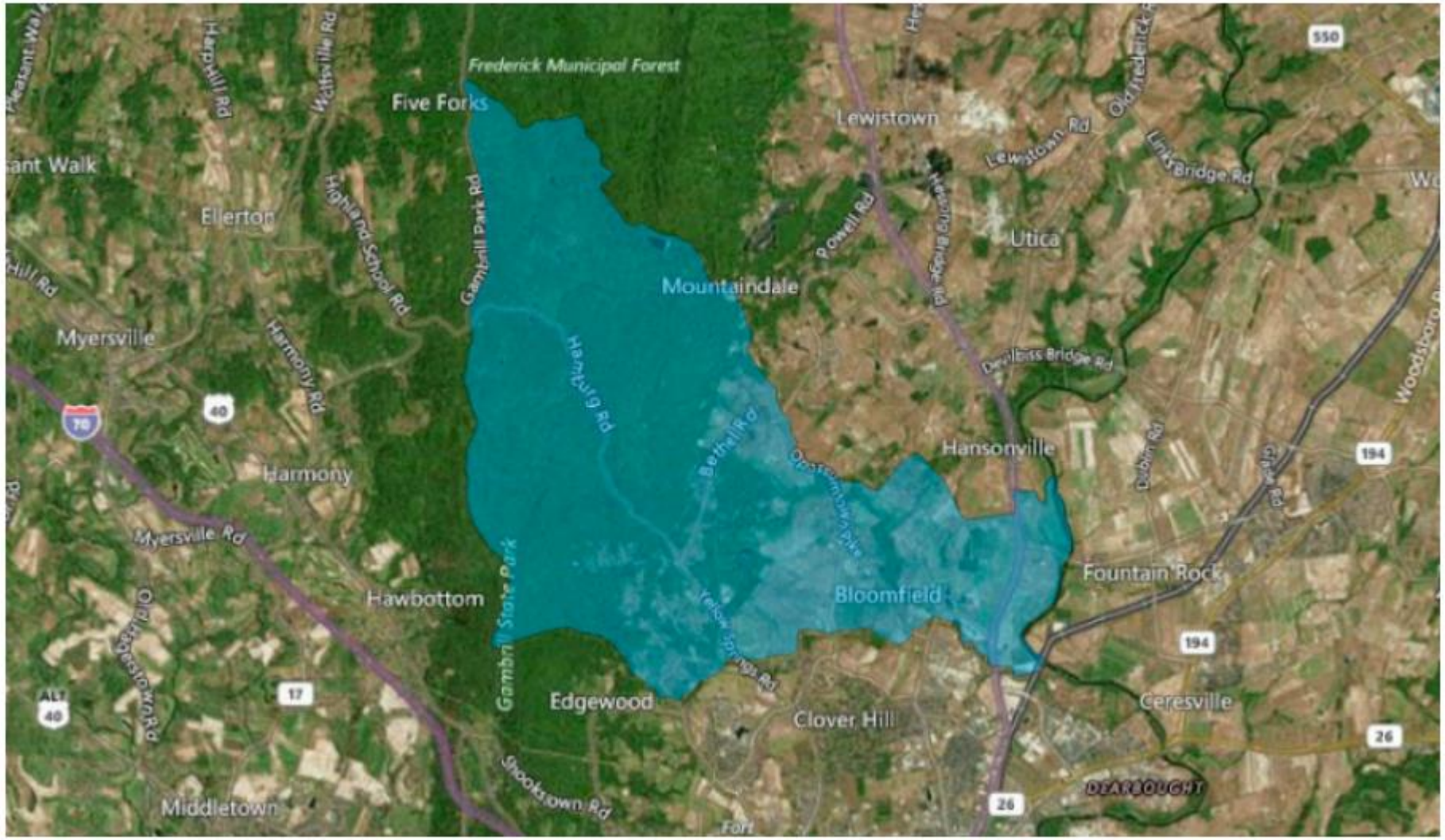
Alguns Casos de Uso: Harmonização/Interpolação Espacial

Limitação:



Alguns Casos de Uso: Harmonização/Interpolação Espacial

Limitação:

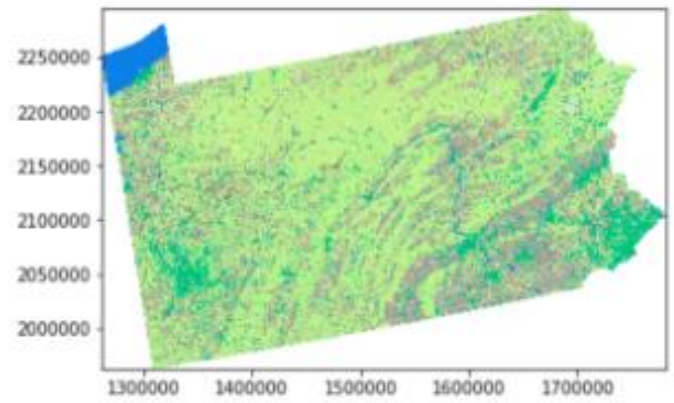


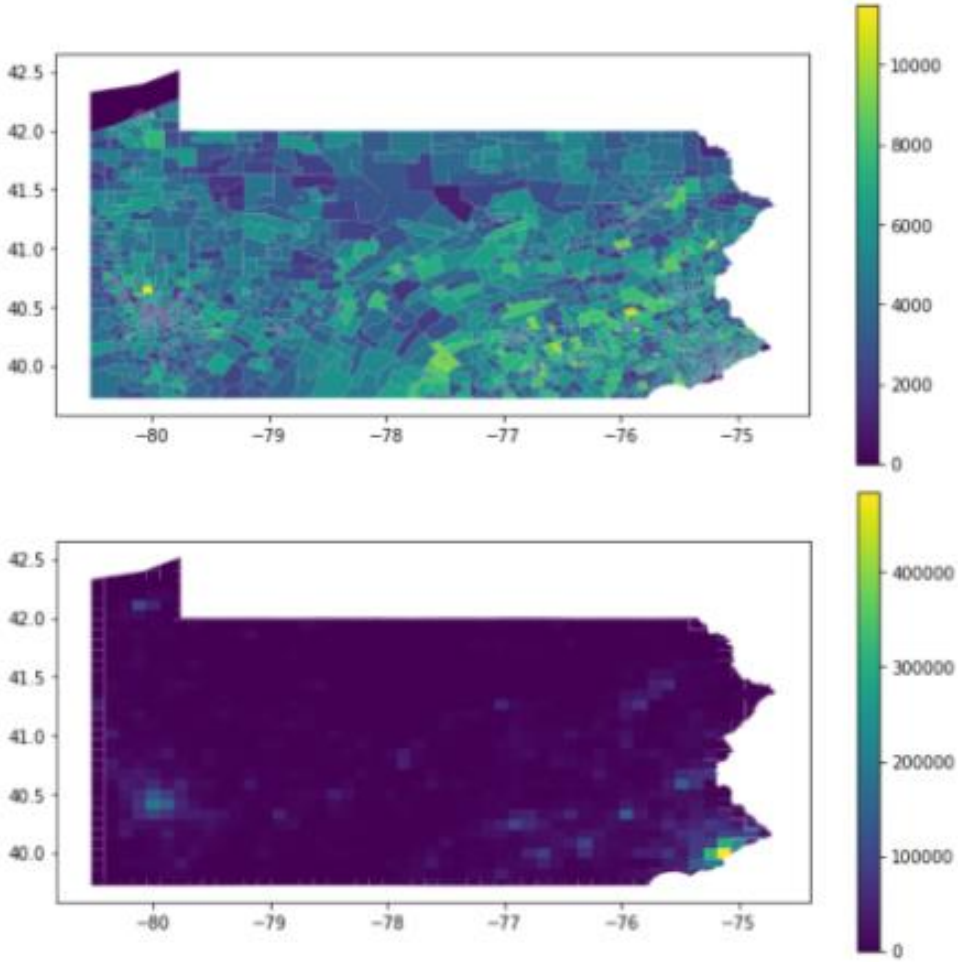
Alguns Casos de Uso: Harmonização/Interpolação Espacial

Solução:

Incorporar imagens de satélite na modelagem e disponibilizar uma ferramenta open-source.

```
penn_raster = rasterio.open(filepath)
show(penn_raster, cmap='terrain')
```



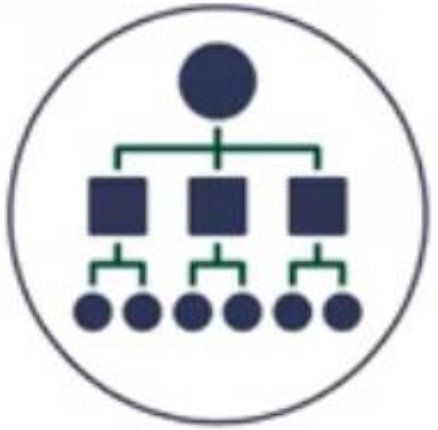


Fonte: *https://github.com/pysal/tobler*

Aprendizado Supervisionado e Não-Supervisionado

Supervised Learning

Aim to predict or model a known target



Unsupervised Learning

Aim to discover structure: no target variable known

Fonte: *https://www.youtube.com/watch?v=oseHCwDBfSE&list=PLriUvS7IljvlcLnrvYUyNc9nXhiM9kWjq*

Alguns dos Principais algoritmos

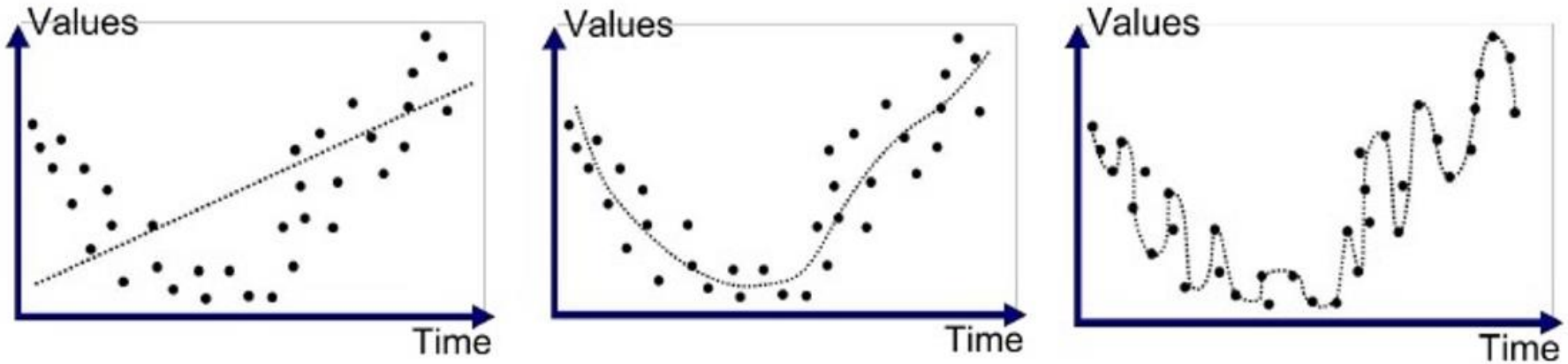
Supervisionado

- Classificação
  - Regressão Logística
  - Support Vector Machines
- Regressão
  - Regressão Linear
  - Regressão Ridge/Lasso
- Modelos Baseados em Árvore
- *K- Nearest Neighbors*

Não-Supervisionado

- Análise de Componentes Principais
- Análise de Cluster
  - K-Means
  - Hierárquico
- Escalonamento Multidimensional

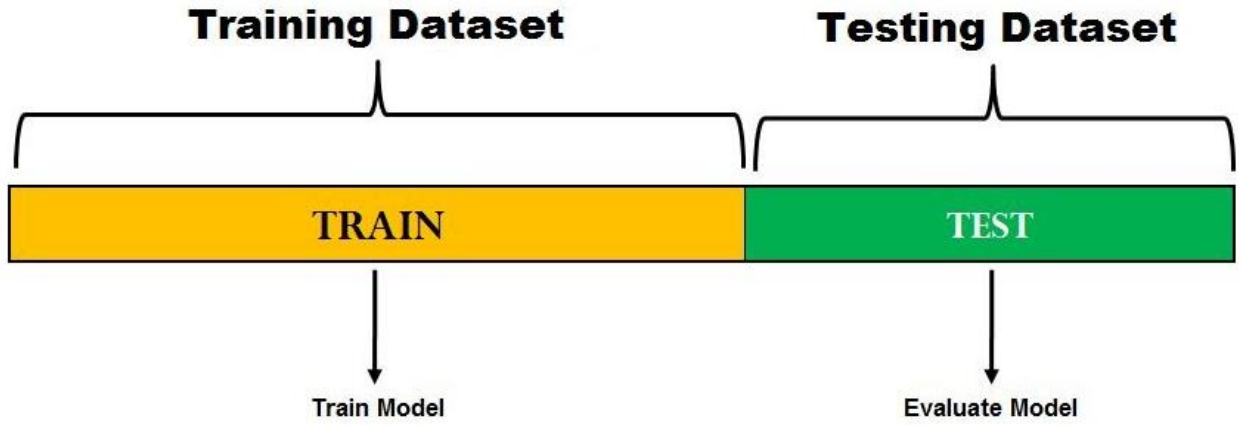
Overfitting e Underfitting: Motivação



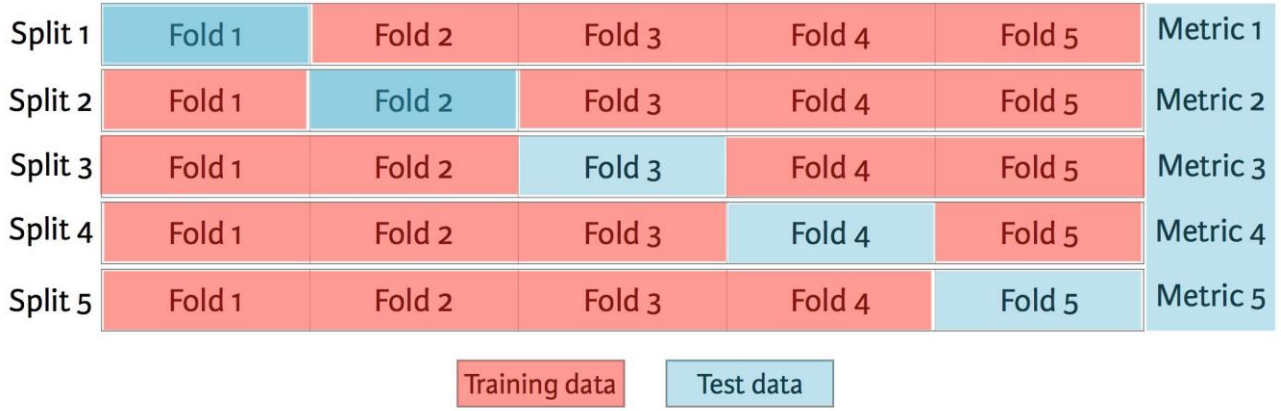
Underfitted | Good Fit/Robust | Overfitted

Overfitting e Underfitting: Estratégias de Validação

- **Holdout set simples**



- **K-Fold Cross-Validation**



Overfitting e Underfitting: Estratégias de Validação

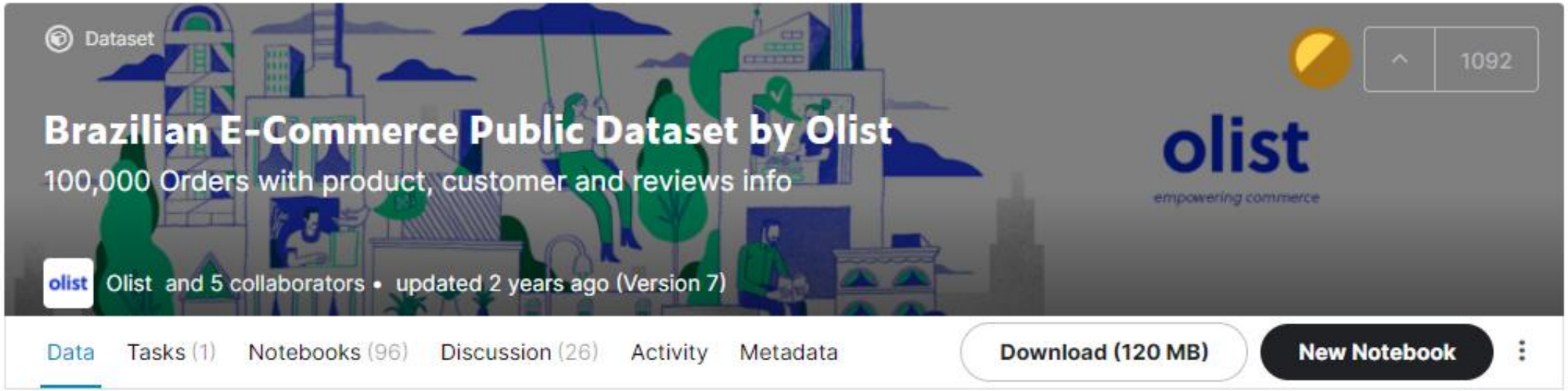
- **Se estamos em um cenário com muitos dados. Podemos "nos dar ao luxo" de criar também uma amostra de validação apartada da de teste.**

- **Discussões Interessantes: https://machinelearningmastery.com/train-final-machine-learning-model/ e https://stats.stackexchange.com/questions/52274/how-to-choose-a-predictive-model-after-k-fold-cross-validation**

Fonte: *Friedman, Jerome, Trevor Hastie, and Robert Tibshirani. The elements of statistical learning. Vol. 1. No. 10. New York: Springer series in statistics, 2001.*

Ok, hora de sujar um pouco as mãos...

Estudo de satisfação do cliente: como prever?



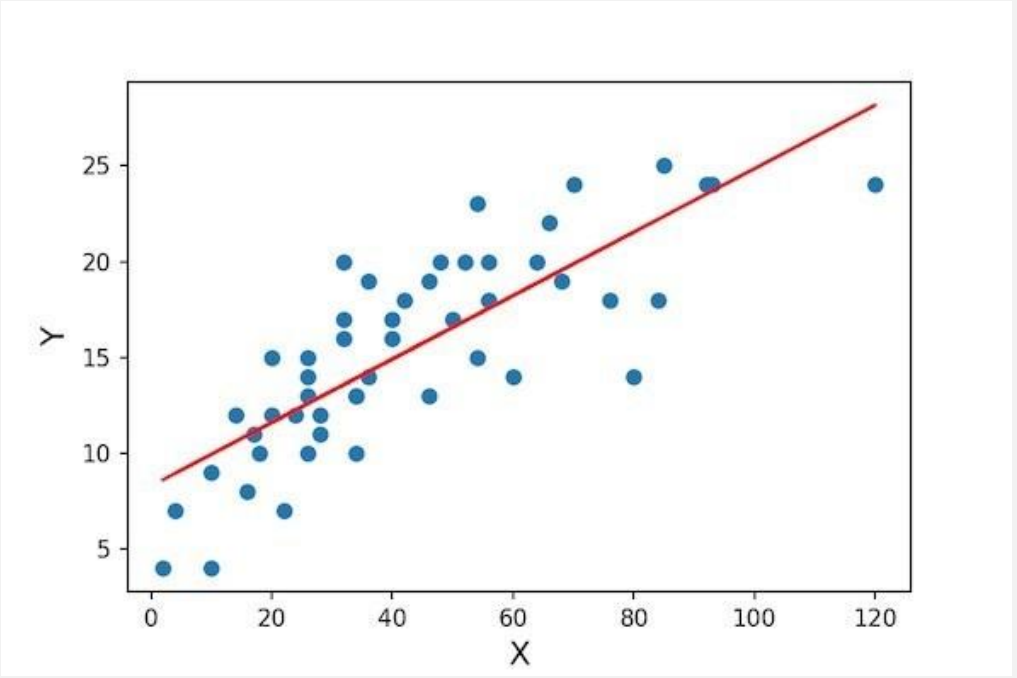
Fonte: https://www.kaggle.com/olistbr/brazilian-ecommerce

Explicando um pouco os modelos utilizados no exemplo

Regressão Linear Simples

$$y_i = \beta_0 + \beta_1 \times x_i + \epsilon_i$$

$$\epsilon_i \sim Normal(0, \sigma^2)$$



Explicando um pouco os modelos utilizados no exemplo

Estruturas Hierárquicas de Representação

Regressão Linear Simples

$$y_i \sim Normal(\eta_i, \sigma^2)$$

$$\eta_i = \beta_0 + \beta_1 \times x_i$$

Regressão Logística

$$y_i \sim Bernoulli(\theta_i)$$

$$\log\left(\frac{\theta_i}{1-\theta_i}\right) = \beta_0 + \beta_1 \times x_i$$

PUCRS online UOL edtech_