

Projetos de

# Ciência de dados

Jéssika Ribeiro

mentorama.

mentorama.

# Introdução

# Nem tudo em DS é ML!

ML é algoritmo, DL é algoritmos...

DS não é sobre *algoritmos* é sobre *resolver problemas*!



Análise



Otimização, Aha-Moment, Market basket analysis, Análise causal.



Modelos



Predição de Churn, CLTV, Recomendações.

# Cultura Data-driven

métodos  
eficientes  
de coleta e  
análise de  
dados.

Cultura orientada a dados consiste na prática de colocar os dados no centro das decisões, ou seja, decisões são fortemente baseadas em informação proveniente dos dados e não apenas do *feeling*.

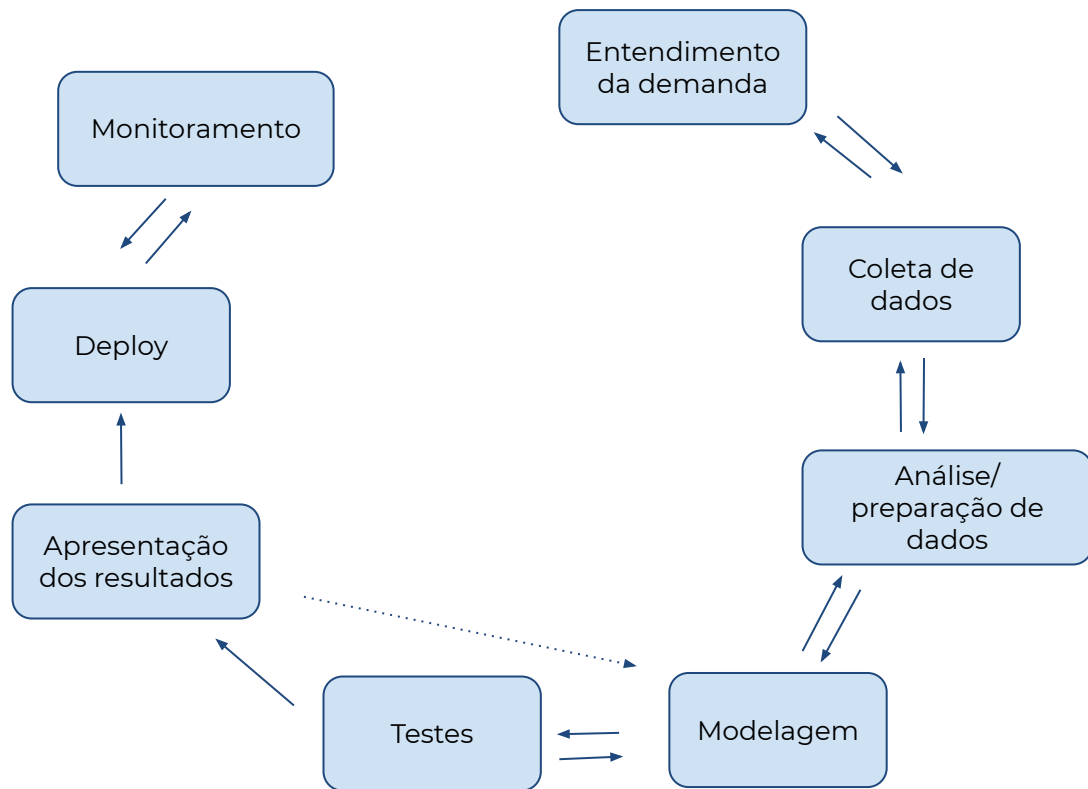
Times de dados levantam  
insumo de dados para  
testar hipóteses, análises e  
modelos que auxiliem  
negócio a tomar decisões  
mais assertivas

Intuição de  
pessoas que  
conhecem  
do business

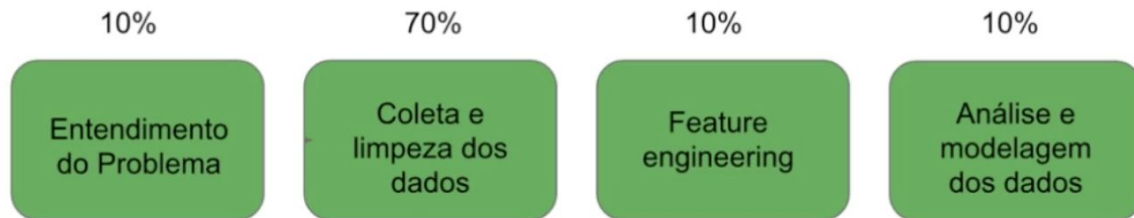
**mentorama.**

# Ciclo de vida de um projeto

# Ciclo de vida de um projeto de DS



## Tempo em cada uma das fases



# Ideação



# Entendimento da demanda

Entendimento da **pergunta/dor** de negócio. Momento onde deve ficar claro **o que** será resolvido, **porque** será resolvido, **como** a área de negócio usará a solução desenhada, quais são as **premissas/hipóteses** de negócio, entre outros. Nessa etapa é possível levantar insumos para definir a viabilidade técnica do projeto.



# Entendimento da demanda



"Estamos perdendo muito clientes. Temos iniciativas para mantê-los engajados, porém somos sempre passivos. Precisaríamos saber de antemão que eles vão sair."



"Certo. Hoje vocês tem uma definição do que é perder um cliente, porém só tem essa informação quando ele já saiu correto?"



"Exato. Hoje entendemos que perdemos um cliente quando ele fica 1 mês sem fazer transações."



"E qual o cenário de desligamentos hoje? Porque vocês vieram até nós? quero dizer, qual a importância de se resolver isso?"



Saber disso só após ele já ter saído não nos dá a oportunidade de tentar reverter a situação. Dessa forma, atualmente perdemos em média 20% dos clientes por mês, gerando um impacto de -500k de receita mensal.



"Entendi! Entendo que se nós fizéssemos um modelo para prever a chance de um cliente dar churn no próximo mês pudesse ajudar vocês!"



Sim! É exatamente isso que nós queremos, vocês entenderam bem :)



Ótimo! E como vocês usariam o nosso modelo no dia a dia? vocês teriam uma estratégia?



Entendo que poderíamos inserir essa probabilidade no nosso sistema de promoções e disparar promoções toda vez que entendermos que o usuário está muito propenso a sair.











# 5W2H - Business

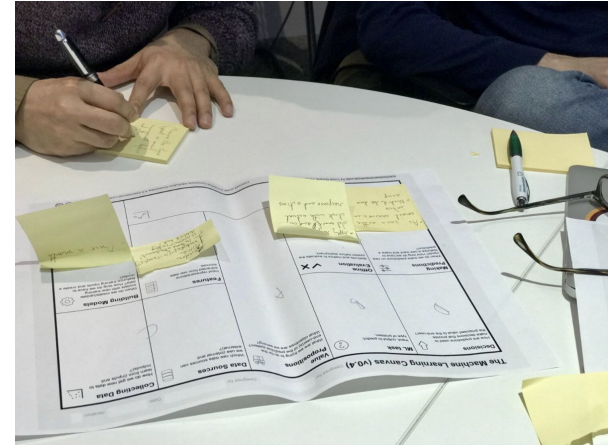
- **What?** O que será feito? Qual é a pergunta de negócio? Chrun? Recomendação? Chatbots?
- **Why?** Por que isso será feito? Qual é a motivação para resolver esse problema? como ele tem impactado a empresa?
- **Who?** Quem serão as pessoas-chave para o auxílio de conhecimento de negócio (keyusers)?;
- **When?** Existe uma data limite para a entrega?
- **Where?** Onde será implementado? No app? no site? sistema interno? para o cliente final?
- **How?** Como será feita a solução que está sendo pedida? É online? É batch? é no app? é no nosso chatbot?
- **How Much?** Estimativa de impacto? Quanto esperamos resolver do problema quando o modelo estiver pronto?

## 5W2H - Técnico

- **What?** O que será entregue? uma API? uma análise? Qual a variável resposta?
- **Who?** Público alvo da análise? Clientes? Funcionários? Clientes que contrataram determinado produto?
- **When?** Período considerado no estudo. Vamos prever chrun no próximo mês? Temos quanto de histórico?
- **Where?** Onde desenvolver? Jupyter? IDE? python? Spark?
- **How?** Como será feita a solução? Desenho da solução técnica: Regressão, classificação...
- **How Much?** Quanto de esforço/ tempo para cada atividade (planning)

# Canvas

<b>Decisions</b>  How are predictions used to make decisions that provide the proposed value to the end-user?	<b>ML task</b>  Input, output to predict, type of problem.	<b>Value Propositions</b>  What are we trying to do for the end-user(s) of the predictive system? What objectives are we serving?	<b>Data Sources</b>  Which raw data sources can we use (internal and external)?	<b>Collecting Data</b>  How do we get new data to learn from (inputs and outputs)?
<b>Making Predictions</b>  When do we make predictions on new inputs? How long do we have to featurize a new input and make a prediction?	<b>Offline Evaluation</b>  Methods and metrics to evaluate the system before deployment.		<b>Features</b>  Input representations extracted from raw data sources.	<b>Building Models</b>  When do we create/update models with new training data? How long do we have to featurize training inputs and create a model?
<b>Live Evaluation and Monitoring</b>  Methods and metrics to evaluate the system after deployment, and to quantify value creation.				



# Os dados

# Identificação

Esse é o momento de entender, junto com os *stakeholders*, com time de engenharia, governança, quais dados temos para explicar o evento que estamos querendo prever.



# Coleta

- Dados em várias fontes: logs de servidores Web, informações de mídia social, bases internas de transações, de clientes, etc.
- Dados no lake: necessárias habilidades técnicas para captura dos dados (SQL)

## Perguntas que auxiliam na fase de coleta:

- *Já temos os dados disponíveis?*
- *Podemos ter acesso aos dados?*
- *Precisa comprar ? Gerar ?*
- *Qual o histórico de dados que temos?*
- *Os dados são tratados? são de qualidade?*





# Preparação e limpeza de dados (Data prep)

- Inconsistência de dados;
- Categorias descontinuadas ou inexistentes na documentação;
- Registros duplicados;
- Formatados não-convencional (ex.: campos de data);
- Inconsistências de cadastros (ex.: idade negativa)
- Valores faltantes

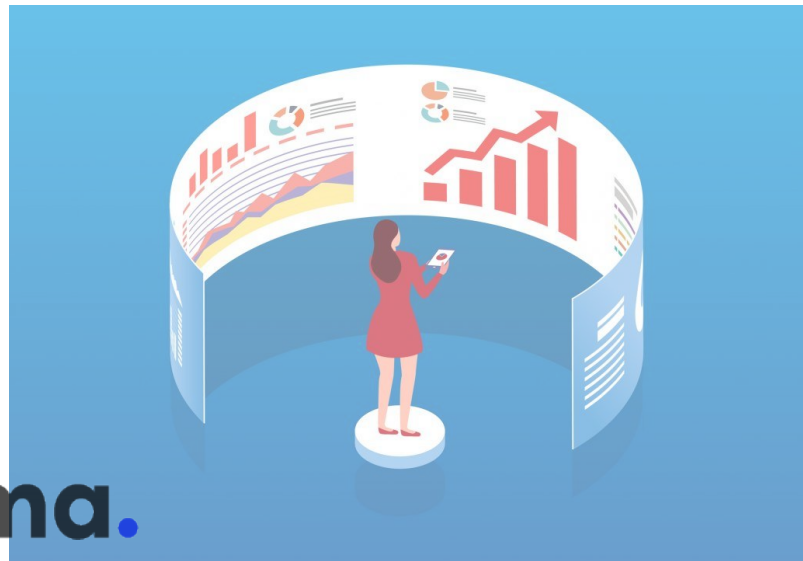


"This is not what I meant when I said 'we need better data cleansing!'"

# Análise

- Compreensão das variáveis a serem usadas;
- Estatísticas descritivas para melhor conhecimento da base: distribuições, quantidade de valores distintos (no caso das categóricas), relações entre variáveis, etc.
- Apresentação dos padrões.

**mentorama.**



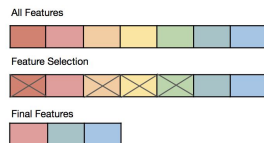
# Modelagem e testes

# Modelagem

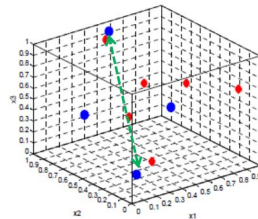
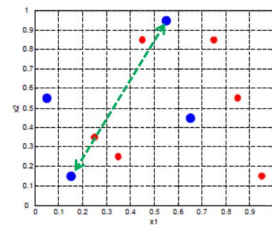
Feature engineering



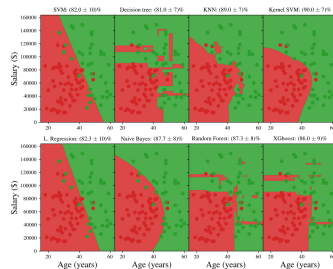
Seleção de variáveis



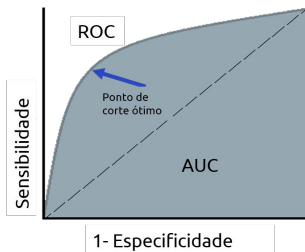
Redução de dimensão



Treinamento/ Identificação de padrão (análise)



Escolha de métricas



Avaliação



mentorama.

# Testes

- Análise de performance na base de teste;
- Análise em dados reais antes de produtização dos modelos;
- Teste A/B (**Recomendação**)
- Retreino ou deploy;



# Apresentação dos resultados

# Resultados

Apresentação dos resultados obtidos através de métricas e visualizações

entender o público

Visualizações

Resultado em termos de  
negócio

Uso do modelo

**mentorama.**

# Deploy/ Monitoramento

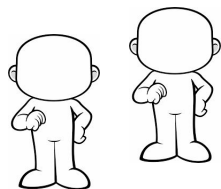
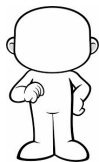
mentorama.

mentorama.



# Deploy

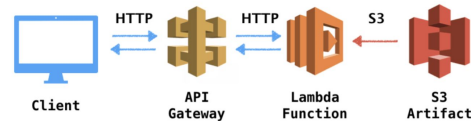
Ciência de dados



Engenharia /Engenharia de ML



Modelo construído e testado



Deploy (API)

# Monitoramento



Monitoramento do input	.....>	Distribuição das variáveis, categorias, tipos de dados...
Monitoramento do output	.....>	Distribuição dos scores, proporção das classes...
Monitoramento online	.....>	Fraude, crédito...
Monitoramento de métricas	.....>	Métricas pós aplicação do modelo com intuito de retreino, por exemplo.

# Papéis e responsabilidades

# Papéis e responsabilidades

01	Sponsor	<ul style="list-style-type: none"><li>• Pessoa mais interessada que o projeto alcance os objetivos de negócio;</li><li>• Tem a autoridade total pelo rumo do projeto;</li><li>• Mantê-los informados e envolvidos(stakeholders);</li></ul>
02	Analista de negócio	<ul style="list-style-type: none"><li>• Tem a função de apresentar a visão do usuário final;</li><li>• Possui conhecimento dos processos, métricas de negócio e dados;</li></ul>
03	Cientista de dados	<ul style="list-style-type: none"><li>• Responsável pelo entendimento da demanda;</li><li>• Coleta, tratamento e limpeza de dados</li><li>• Responsável pela modelagem;</li><li>• Responsável pelos testes e avaliação do modelo;</li></ul>
04	Engenheiro de dados/ML	<ul style="list-style-type: none"><li>• Responsável pela disponibilização de dados confiáveis;</li><li>• Responsável pela produtização dos modelos</li></ul>
05	Governança	<ul style="list-style-type: none"><li>• Conhecedora dos dados que estão disponíveis: onde encontrá-los e como interpretá-los;</li><li>• Responsável pela organização de documentações de modelos;</li></ul>

# Pontos críticos em um projeto de DS

# Fatores que impactam o sucesso de um projeto

1. Começar com as perguntas erradas;
2. Focar no problema errado;
3. Qualidade e quantidade dos dados;
4. Falta de comunicação;
5. Falta de todos os profissionais necessários;
6. Falta de entrega contínua;



**mentorama.**

# Boas práticas

## Boas práticas

1. Ter um foco ou meta e mantê-lo;
2. Procurar a homogeneidade de ferramentas;
3. Notebook é estudo, não código produtivo!
4. Salvar .pkl de modelos;
5. Versionamento de código
6. Compartilhar de conhecimento;
7. Ambiente do projeto limpo e organizado;





# Cookiecutter

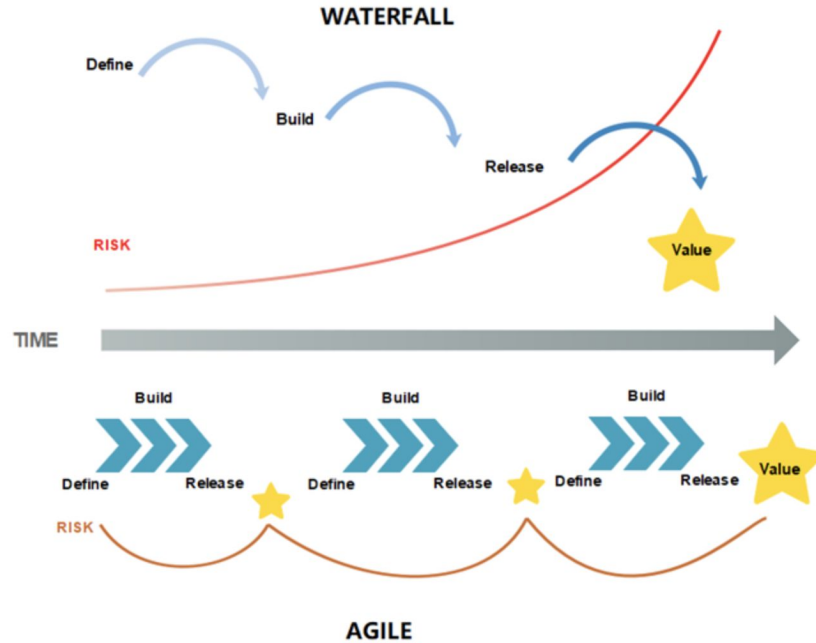
```
— LICENSE
— Makefile      <- Makefile with commands like `make data` or `make train`
— README.md     <- The top-level README for developers using this project.
— data
  |— external   <- Data from third party sources.
  |— interim    <- Intermediate data that has been transformed.
  |— processed  <- The final, canonical data sets for modeling.
  |— raw        <- The original, immutable data dump.
— docs          <- A default Sphinx project; see sphinx-doc.org for details
— models        <- Trained and serialized models, model predictions, or model summaries
— notebooks     <- Jupyter notebooks. Naming convention is a number (for ordering),
                  the creator's initials, and a short '-' delimited description, e.g.
                  `1.0-jqp-initial-data-exploration`.
— references    <- Data dictionaries, manuals, and all other explanatory materials.
— reports
  |— figures    <- Generated graphics and figures to be used in reporting
— requirements.txt <- The requirements file for reproducing the analysis environment, e.g.
                  generated with `pip freeze > requirements.txt`
— setup.py      <- Make this project pip installable with `pip install -e`
— src
  |— __init__.py <- Makes src a Python module
  |— data        <- Scripts to download or generate data
  |   |— make_dataset.py
  |— features    <- Scripts to turn raw data into features for modeling
  |   |— build_features.py
  |— models      <- Scripts to train models and then use trained models to make
  |   |           predictions
  |   |— predict_model.py
  |   |— train_model.py
  |— visualization <- Scripts to create exploratory and results oriented visualizations
  |   |— visualize.py
— tox.ini       <- tox file with settings for running tox; see tox.testrun.org
```

# Gestão de projetos

mentorama.

mentorama.

# Agile vs Waterfall



# Agile vs Waterfall

## Principais diferenças

	Waterfall	Agile
Análise de Viabilidade	- Geralmente demorado. - Processo detalhado para evitar retrabalho.	A ideia é demorar o menor tempo possível.
Planejamento	- Mais rígido. - Sem alterações ao longo do projeto.	- Flexível - Feito no início das sprints; - pode ser alterado no decorrer do processo.
Papéis e responsabilidades	Bem definidos desde o início do projeto	Times se auto-organizam a fim de cumprir uma tarefa.
Documentações	- bem detalhadas; - presentes em várias etapas;	Não há necessidade de documentação.

## Características comuns:

1. Software de alta qualidade;
2. Atividades similares: Entendimento, planejamento, desenvolvimento, teste e deploy;

# Metodologia para ciência de dados

Além de programação, estatística, matemática, Data Science requer criatividade!



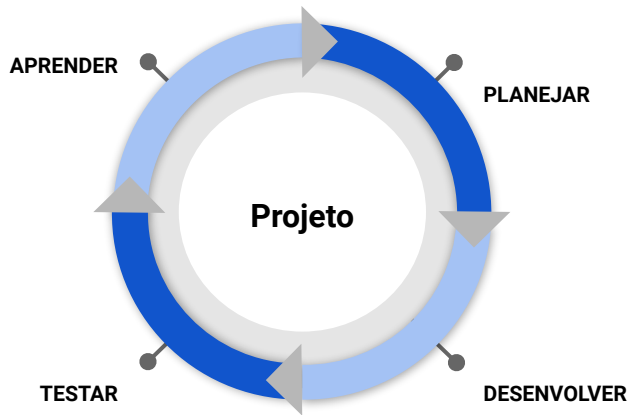
Ciclo de vida dos projetos é não-linear, de intensa pesquisa e experimentação



O que gera grandes incertezas sobre o processo

# Agile em Data Science

É uma metodologia que propõe uma forma de **acelerar as entregas** durante o desenvolvimento de um projeto. O entregável final é **fracionado em entregas incrementais**. Os times geralmente são **multidisciplinares** e trabalham para atingir uma meta estabelecida a cada **fase (sprint)**.



Princípios relevantes em DS:

- Iteração;
- Iterações em períodos curtos de tempo;
- Feedback em cada iteração.