

# MBA em Inteligência Artificial

## Big Data e Visualização de Dados

Prof. Diego Nogare



Faculdade de Computação e Informática  
**Mackenzie**

# Diego Nogare

---



Former MVP Artificial Intelligence  
Microsoft Regional Director  
Diretor no PASS.org  
Chief Data Officer @ Lambda3  
Membro notável na I2AI  
Mestre em IA

[www.diegonogare.net](http://www.diegonogare.net)  
[www.livrosdonogare.com.br](http://www.livrosdonogare.com.br)



---



# COMO INICIAR O PROJETO DE BIGDATA / IA / ML / DS

---



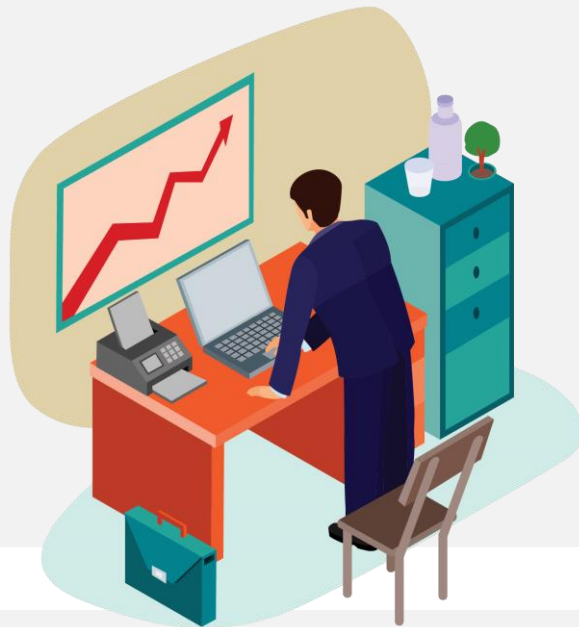
# DATA DRIVEN PROCESS

Processo de tomar decisão baseado em dados

# ■ Análise de Dados – Um pilar do método científico

---

É um processo que busca encontrar padrões e extrair valor a partir de dados





The diagram illustrates a cyclical process for data science. It features six rounded rectangular boxes arranged in a circle, connected by a light pink circular arrow pointing clockwise. The boxes are: 'Natureza do problema' (dark red, top), 'O que / como medir' (brownish-orange, top-right), 'Coleta' (orange, bottom-right), 'Exploratória' (yellowish-brown, bottom), 'Formal' (greenish-yellow, bottom-left), and 'Interpretação' (green, top-left). A large, faint watermark 'MNO' is visible in the background.

**Natureza do problema**

**O que / como medir**

**Coleta**

**Exploratória**

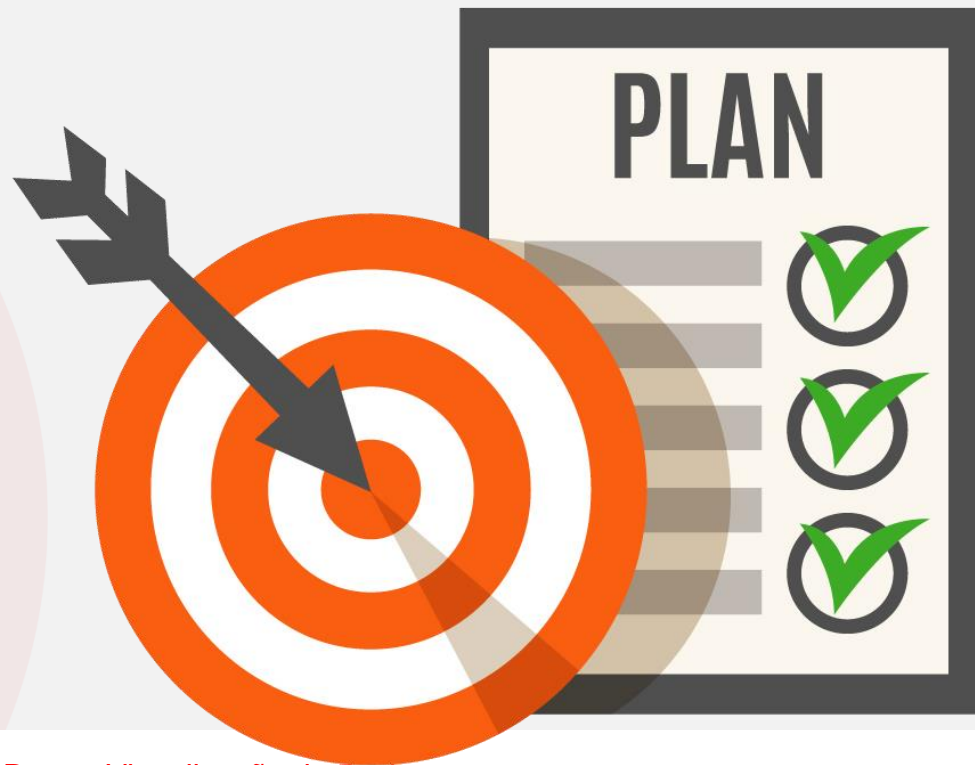
**Formal**

**Interpretação**

# ■ Natureza do problema

---

Qual é o  
meu  
objetivo?



# ■ Natureza do Problema

---

Para uma análise sólida, é importante **entender o problema central da pesquisa**. Quais objetivos esperamos responder. Um gestor de dados pode ajudar no direcionamento da pesquisa quando isso envolver um estudo multicêntrico.



# ■ Decidindo o que e como medir

---

Discutir os objetivos do trabalho e quais serão as métricas a serem analisadas.

Quais serão os benefícios para a empresa?

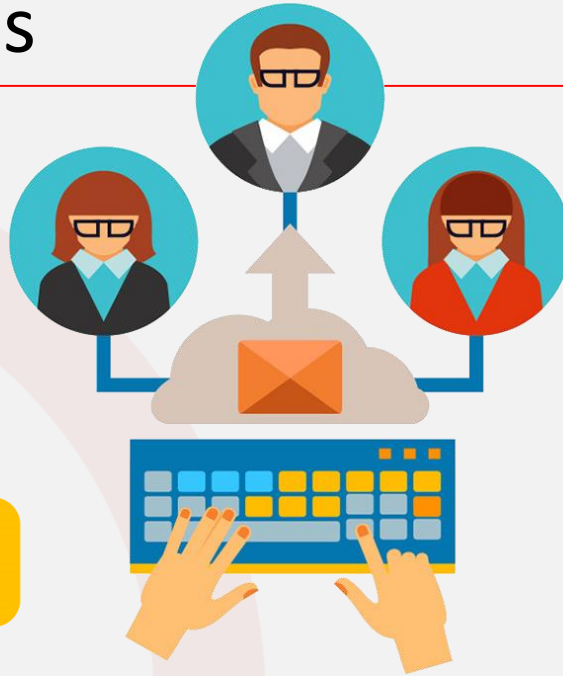


# ■ Decidindo o que e como medir

---

A partir da definição do problema, é **criado o planejamento** para levantamento dos dados que ajudarão a responder o que é necessário. Muitas vezes a resposta pode parecer óbvia, mas mesmo assim, é importante **deixar as “crenças” de lado**. Associações espúrias nos ajudam a enxergar isso de forma palpável.

# ■ Coleta de dados



# ■ Coleta dos dados

---

Algumas perguntas sobre a coleta de dados ajuda o nosso trabalho. Questionamentos como:

- A fonte de dados é suficiente?
- Se não for, precisamos coletar mais dados?
- Temos todos dados dentro de casa, ou precisamos ir para o mercado?
- Qual metodologia foi usada para coletar os dados?

As conclusões dependem da análise inicial, que por sua vez, depende da qualidade dos dados e de como foram coletados!

# ■ Análise exploratória descritiva

Com elementos de estatística, são descobertos padrões, distribuições e comportamento dos dados.



# ■ Análise Exploratória descritiva

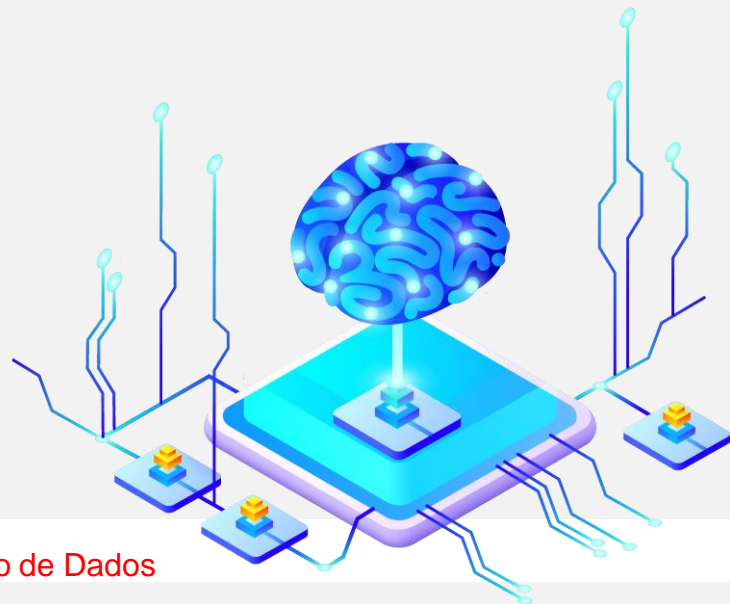
---

O passo seguinte à coleta é realizar a **Análise Exploratória**. Processos de agrupamentos, tabulações, extrações e gráficos são artifícios para entender os dados e as características para aplicar as técnicas de análise mais sofisticadas. Este processo é chamado de **Estatística Descritiva**, que permite um próximo passo, a **Estatística Inferencial**. Este último consiste em responder novas hipóteses a partir de dados históricos e calcular os métodos de confiança destas inferências.

# ■ Análise formal e preditiva

---

Ensinaamos o computador e criamos elementos de Inteligência Artificial para nos ajudar a reconhecer determinados padrões.



# ■ Análise formal e preditiva

---

Após entender o domínio dos dados na Análise Exploratória, o cientista pode aplicar métodos estatísticos mais avançados como **Aprendizagem de Máquina**, seja com método **supervisionado**, **semi-supervisionado**, **por reforço** ou então **não-supervisionado**. Estas técnicas baseiam-se no tipo de problema a ser resolvido.



# ■ Interpretação e discussão dos resultados

---

Reunimos o grupo novamente e discutimos o que aprendemos com o resultado e quais decisões podem ser tomadas



# ■ Interpretação e discussão dos resultados

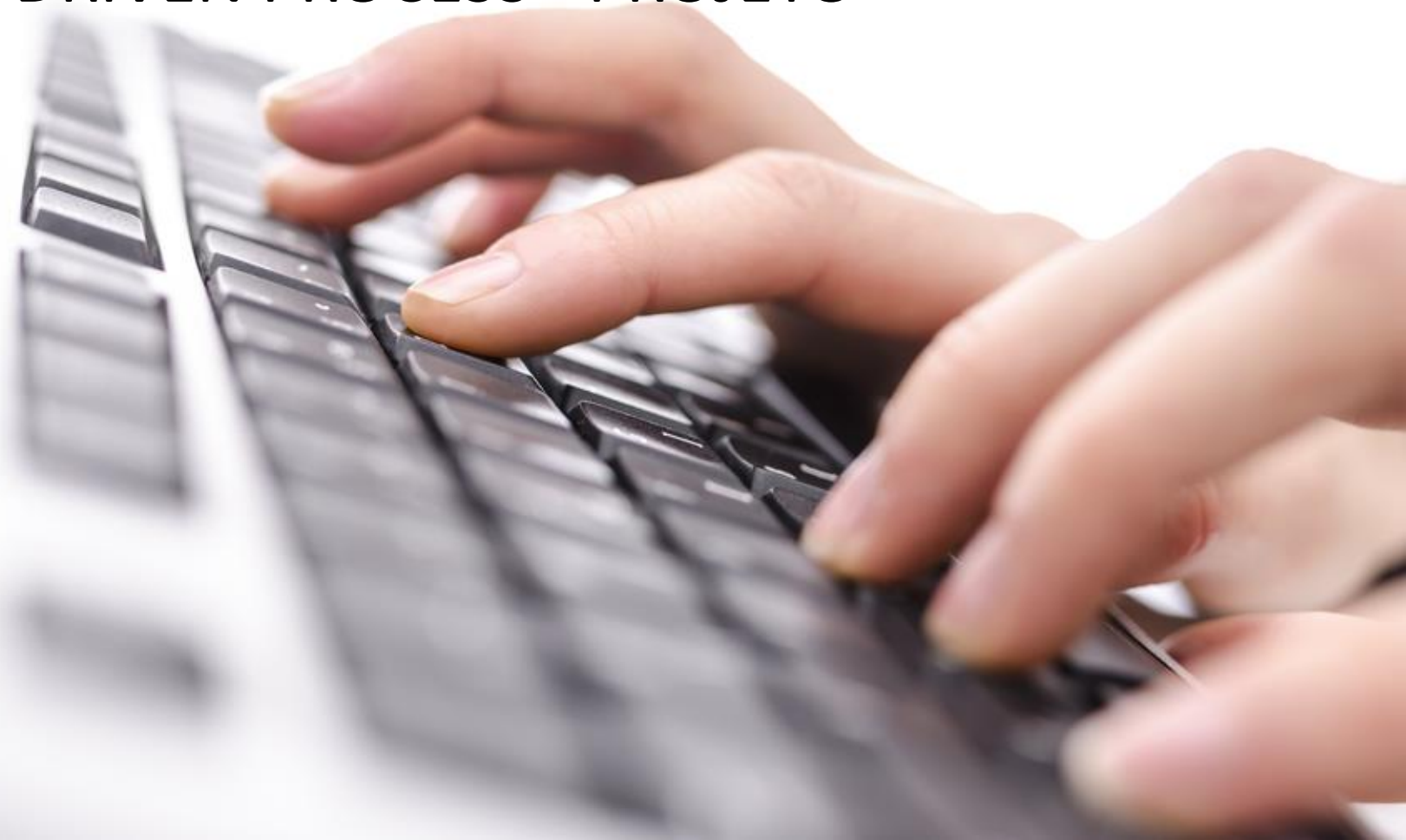
---

No fim do ciclo, é esperado que as **respostas** planejadas lá no começo do processo **sejam respondidas**. Porém, algumas outras que são pertinentes à esta etapa, também devem ser respondidas. Questões como:

- O que aprendemos com o resultado?
- Quais decisões tomar?

Como o processo é cíclico, voltamos a parte de **Natureza do Problema** e geramos novas perguntas a partir dos insight que surgiram por agora.

# DATA DRIVEN PROCESS - PROJETO













---



# MACHINE LEARNING CANVAS

# Objetivo

<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.				

o e Informática











# ■ Objetivo

---

Geralmente se começa pela área da **Proposta de Valor**, que fica localizada no centro do Canvas. Nesta região, procurem responder o **What+Why+Who**. Estes três elementos são importantes para definir:

- O que estamos tentando fazer;
- Por que é importante, e
- Quem vai usar e ser impactado (ou seja, quem é o usuário final)!

# Aprendizado

<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. 			

de Informática

# ■ Aprendizado

---

**Fontes de Dados.** Neste quadro é respondido quais são as fontes de dados brutas que vão ser trabalhadas para ajudar a resolver o problema.

**Coleta de Dados,** é respondido como podemos obter novos dados para aprender (entradas e saídas)?



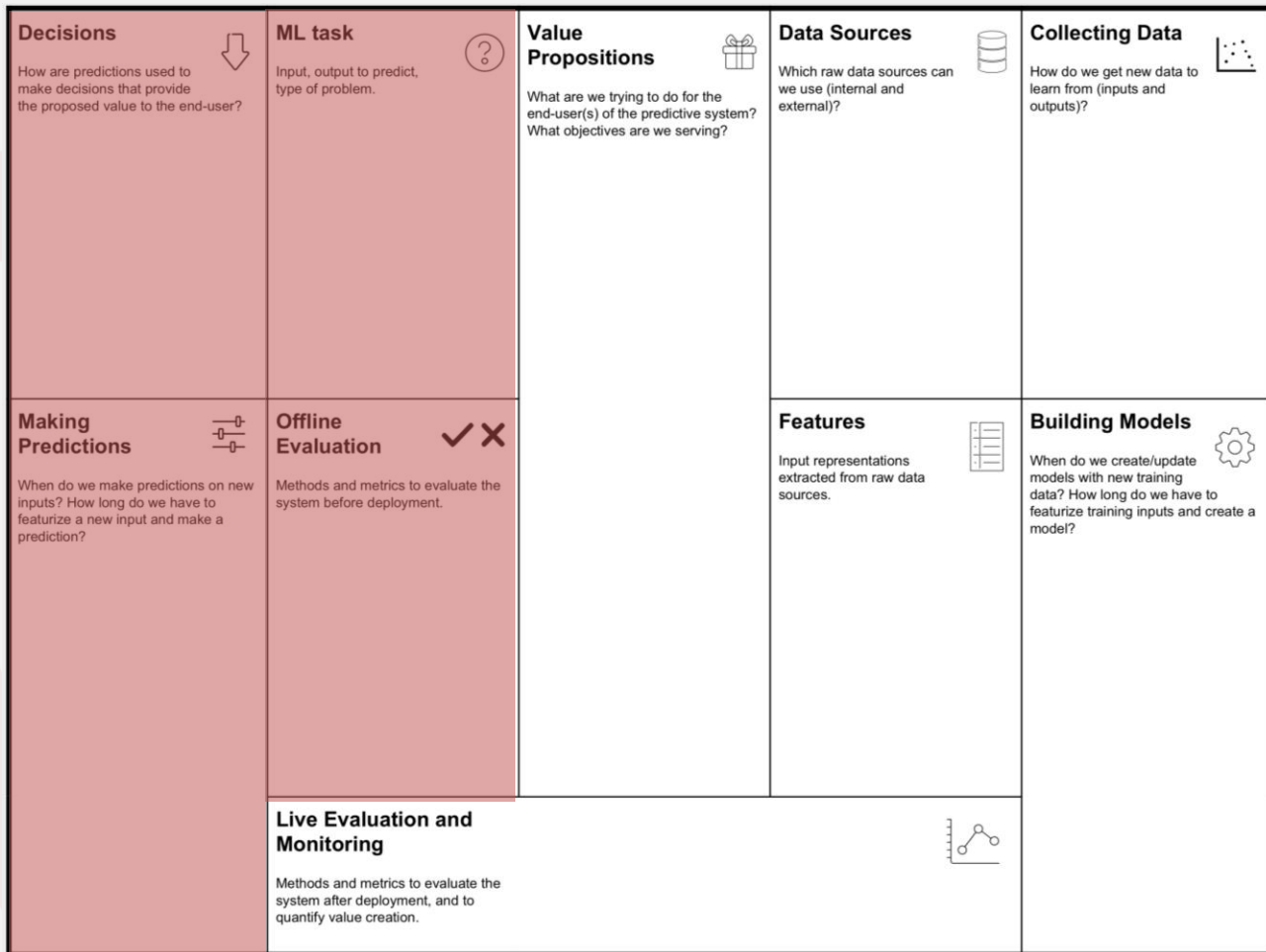
# ■ Aprendizado

---

**Características** é fundamental para definir quais itens das variáveis existentes nos dados brutos serão utilizadas para criar o modelo. Olhando para as fontes de dados escolhidas, é preciso escolher quais serão as características que se aplicam lá para resolver o problema.

**Construção de Modelos**, é importante definir como serão feitas as extrações dos dados da base de origem, e quais dados vão ser utilizados para a construção ou atualização do modelo, como também quanto tempo temos de janela para fazer esse trabalho.

# Predição



Informática

# ■ Predição

---

**Tarefa de Machine Learning** é esperado que seja definida família de algoritmos que será utilizada para resolver o problema, também quais serão as entradas e qual é o resultado esperado. Esta visão ajuda a direcionar a escolha das técnicas para elaboração do modelo de predição.

# Predição

---

**Decisões** condiz com as conversas que a área de negócio trás como problemas para nós. Aqui é esperado deixar claro como as previsões que retornam do modelo vão ser aproveitadas pelo usuário final, lembrando em responder o que temos lá na Proposta de Valor.

# ■ Predição

---

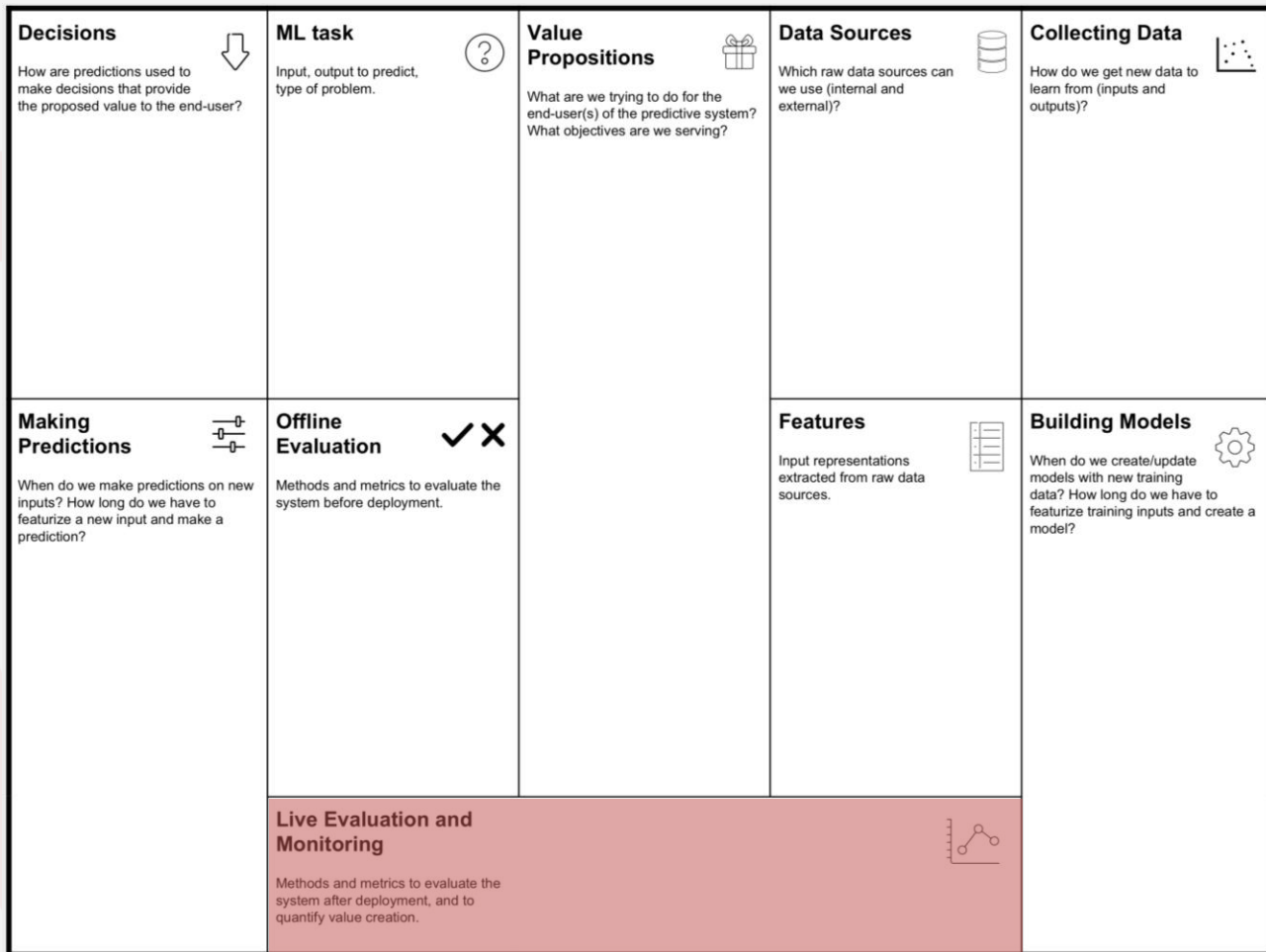
**Fazer Previsões** é discutido o momento no qual o modelo será acionado para responder às chamadas. Aqui pode ser um processo em tempo real ou então em lote (batch). Cada necessidade de negócio possui sua característica e necessidade, podendo haver situações que precisam ser analisados em tempo real, e outras em lote. E quanto tempo temos disponível para tomar uma ação com base na resposta.

# Predição

---

**Avaliação Offline** é descrito quais métodos e métricas serão utilizados para avaliar a maneira de como as previsões são feitas e utilizadas, isso antes de ser implantado

# Avaliação



Informática

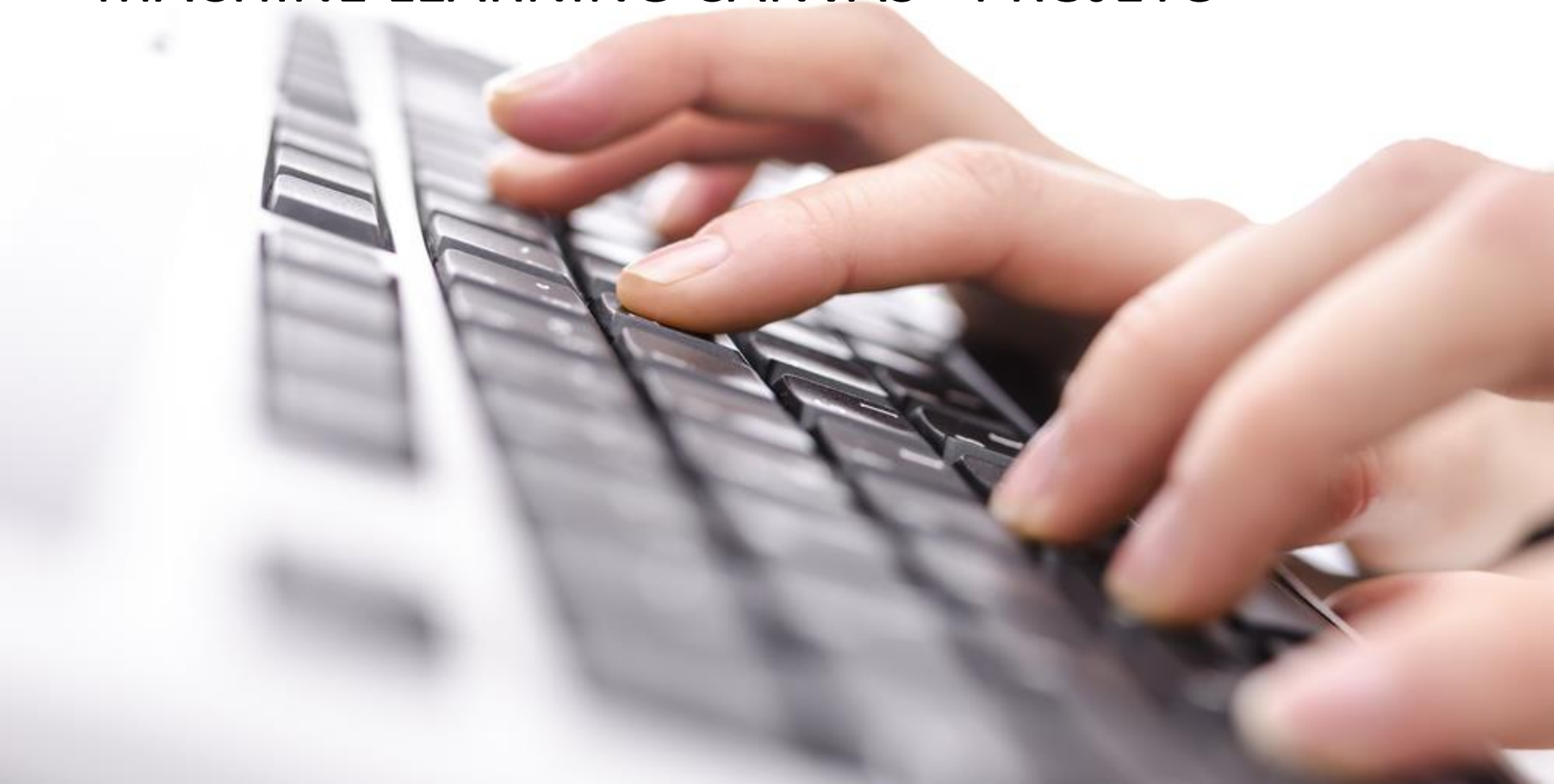
# ■ Avaliação

---

**Avaliação e Monitoramento ao Vivo** é dedicada a medir o funcionamento do modelo e monitorar se o valor de acertos continua aceitável.



# MACHINE LEARNING CANVAS - PROJETO



# Referências

---

Machine Learning Canvas:

<https://www.louisdorard.com/machine-learning-canvas>

Introdução à Machine Learning

<https://towardsdatascience.com/machine-learning-an-introduction-23b84d51e6d0>





**Faculdade de Computação e Informática**  
**Mackenzie**