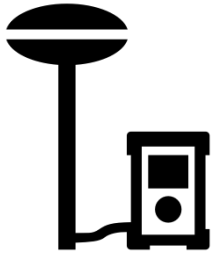
Machine Learning

Parte 1

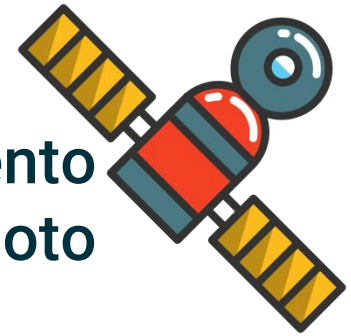
Camila Duelis Viana



Instrumentação
Geofísica

Campo
Amostras
Análises

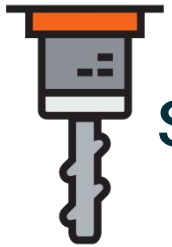
Sensoriamento
remoto



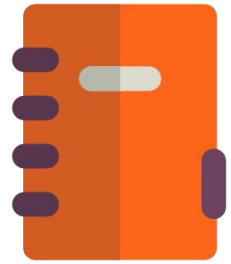
Aplicativos
Colaborativo

Mapas
Cartas

Internet
Repositórios
Nuvem

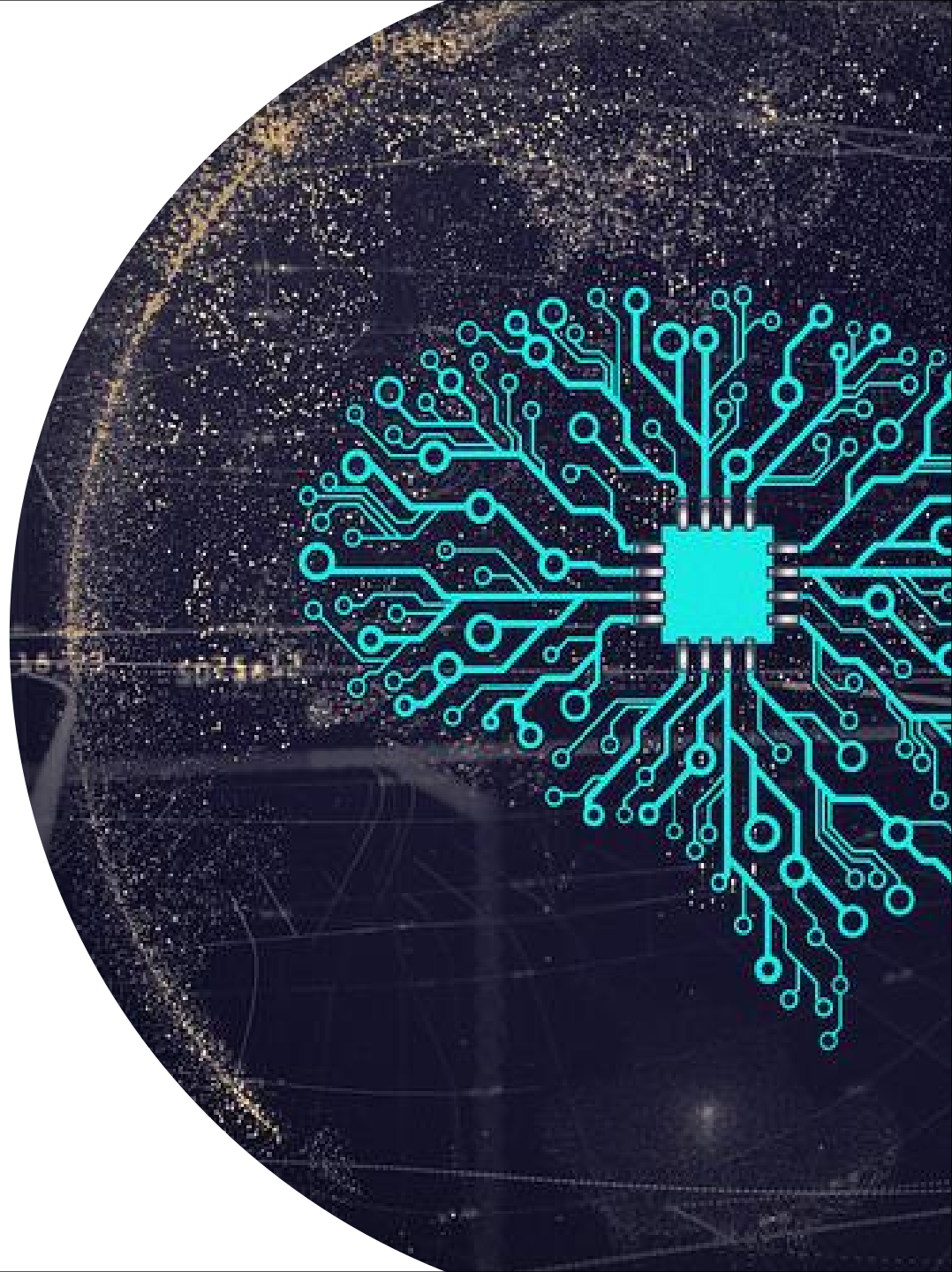


Sondagens



Descrições
Relatórios

O que fazer com tantos dados?



4° paradigma da ciência

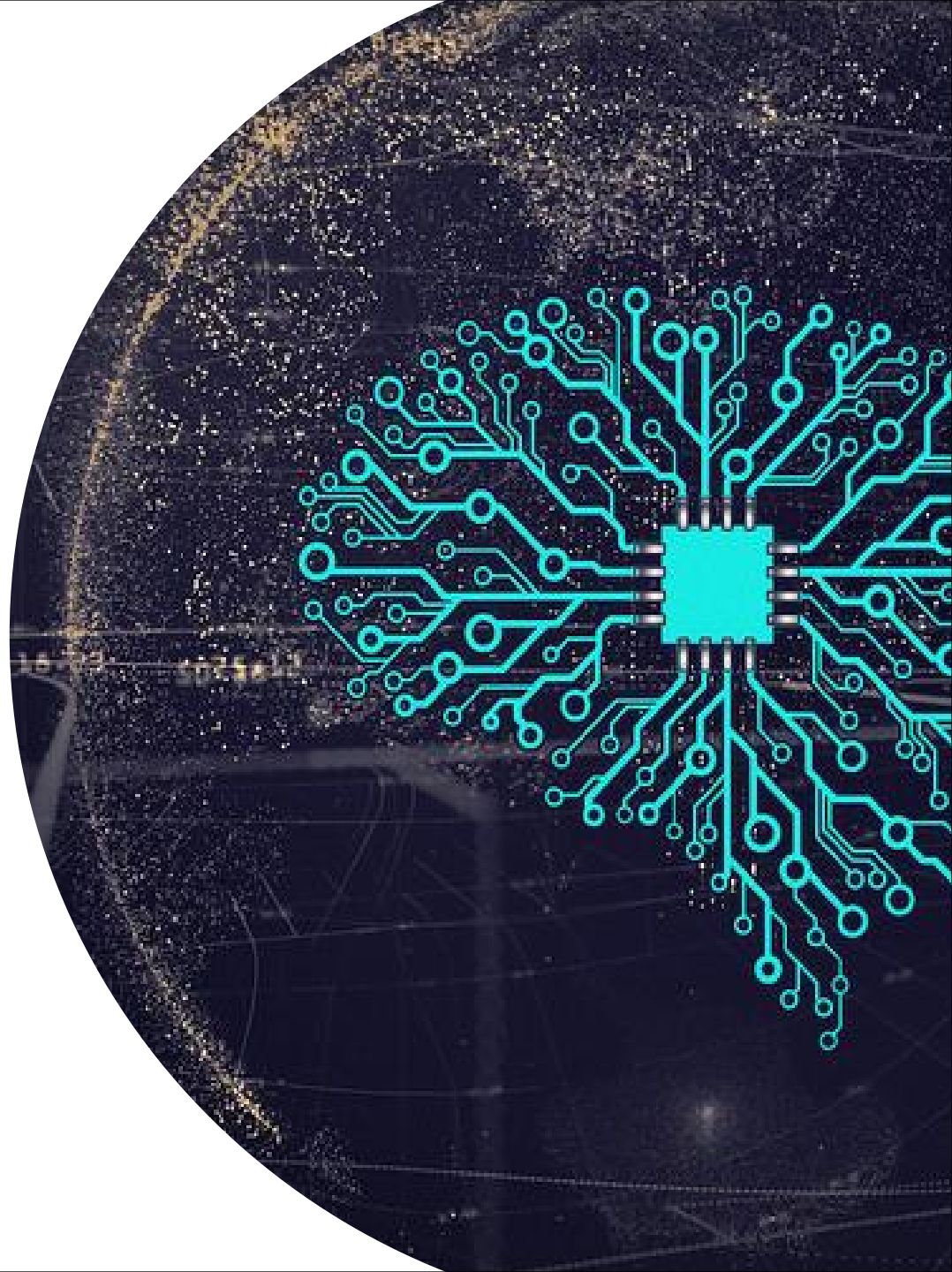
Ciência é centrada nos dados, sejam observados ou simulados, unificando teoria, experimentos e simulações - eScience

- Eficiência
- Digitalização
- Novas visões
- Mais poderosas em conjunto

(GEO)BIG DATA

(GEO)DATA DRIVEN

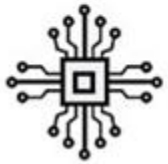
Como extrair informação de tantos dados?

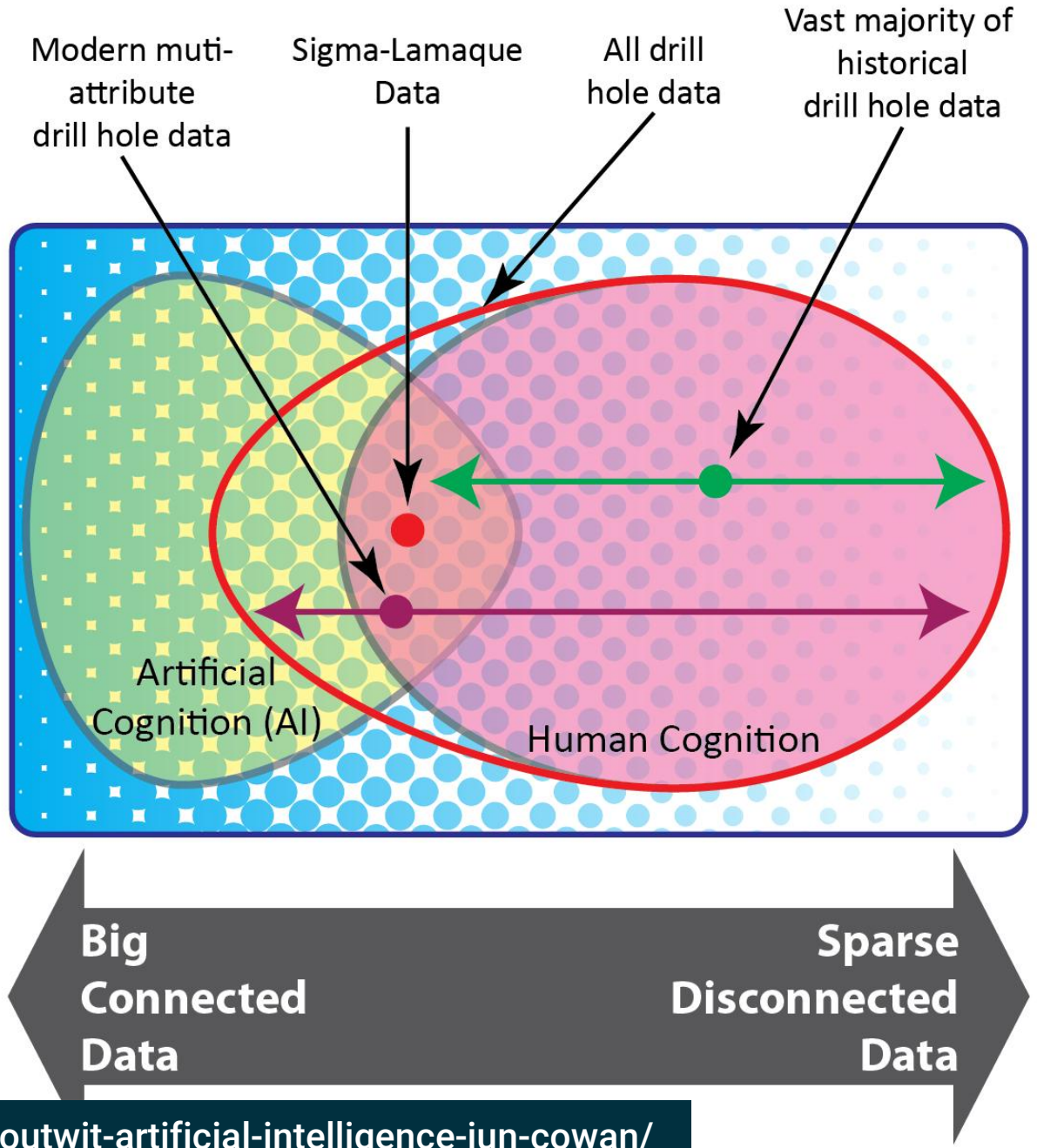


Big Data

- Volume, Velocidade, Variedade
- Exige processamento diferenciado

Humanos vs. Máquinas

	Weight	Space	Processor Speed	Energy Efficiency
	3 pounds (1.4 kg)	1/6 basketball (80 cubic inches or 1,300 cm^3)	Up to 1,000,000 trillion operations per second	20 watts
	150 tons	Basketball court (cabinets over 4,350 square feet, or 400 m^2)	93,000 trillion operations per second	10 million watts



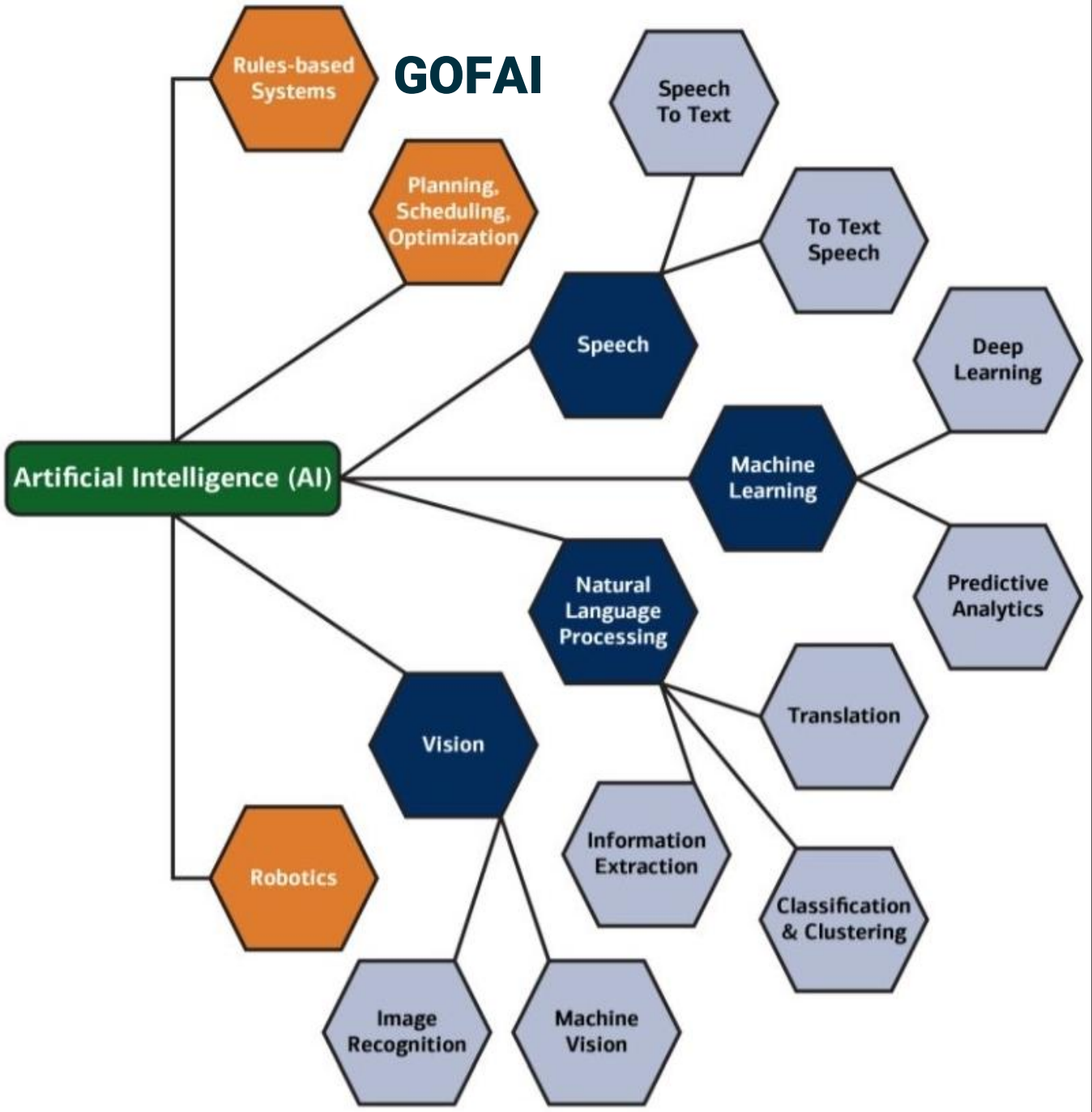
Saiba mais: https://www.linkedin.com/pulse/how-geologists-can-outwit-artificial-intelligence-jun-cowan/

Inteligência Artificial

- 1950s
- Inteligência humana exibida por máquinas
- Termo geral para fazer com que computadores realizem atividades humanas

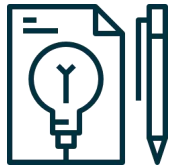
IA Limitada

Sistema pode fazer apenas uma (ou poucas) coisas definidas tão bem ou melhor que humanos



Machine Learning

- 1980s
- Uma abordagem para alcançar IA através de sistemas que podem aprender por experiência para encontrar padrões em um conjunto de dados
- Reúso de código
- Tarefas:
  - Automação
  - Modelagem e problemas inversos
  - Descoberta (padrões, estruturas, relacionamentos)



Problema Inverso
Inverse problem

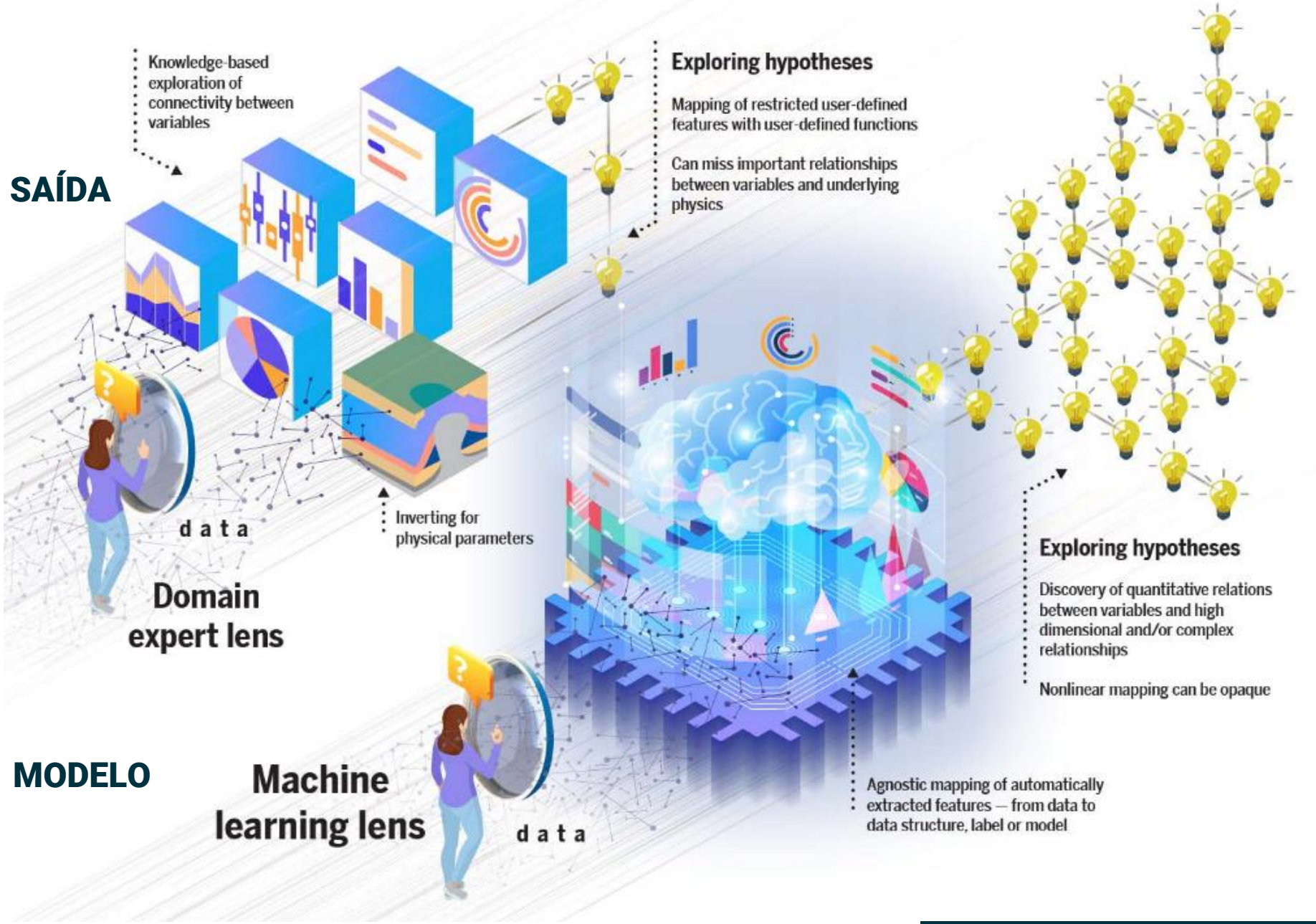
Converter medidas observadas em informações sobre um objeto ou sistema.

Regressão Prever valores numéricos	**Classificação** Valores categóricos
Agrupamento Outros exemplos mais similares	**Previsão de sequência** O que vem depois?

DADOS
MODELO → Programação tradicional → **SAÍDA**

Ensinar o computador a reconhecer padrões através de exemplos, ao invés de programar regras específicas

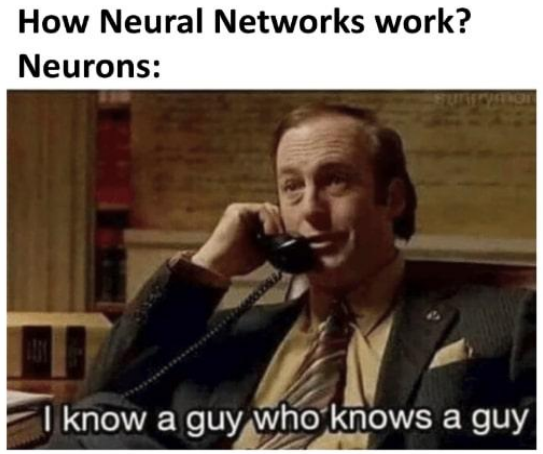
DADOS
SAÍDA → Aprendizado de máquina → **MODELO**

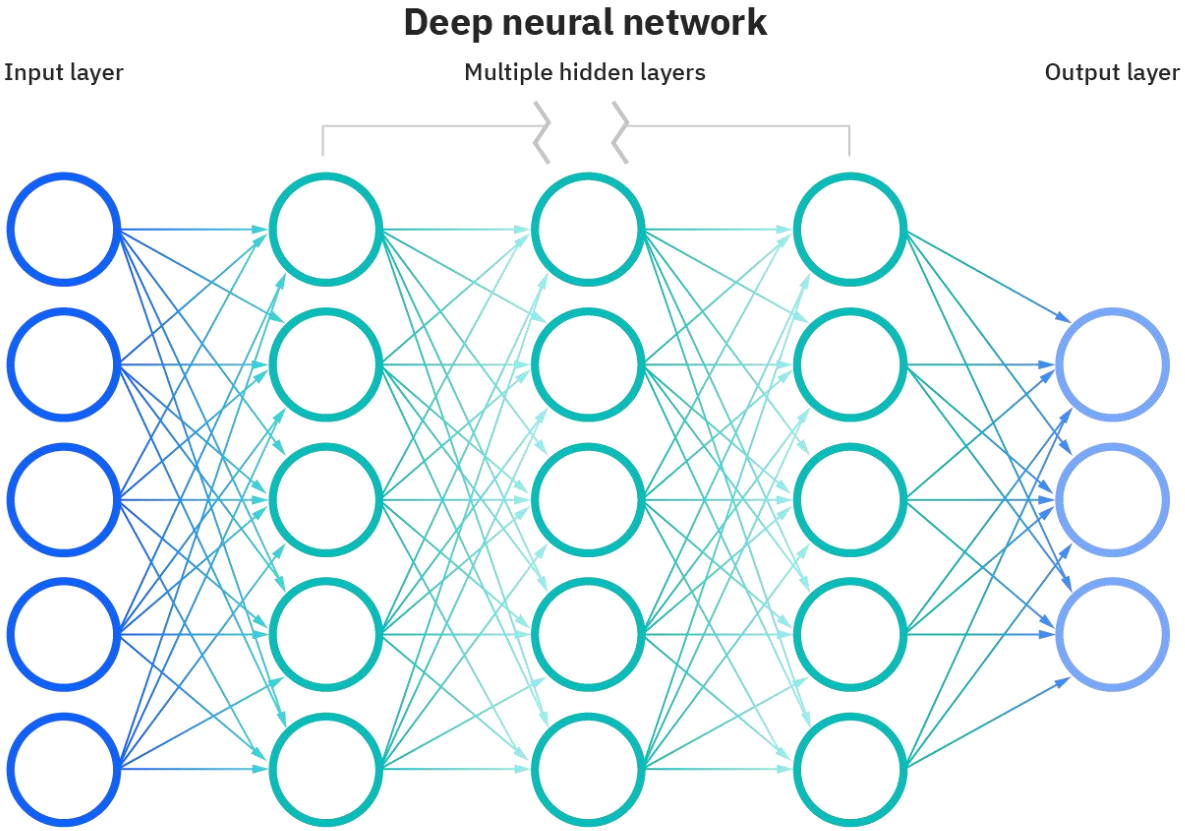


Saiba mais: Bergen et al., 2019

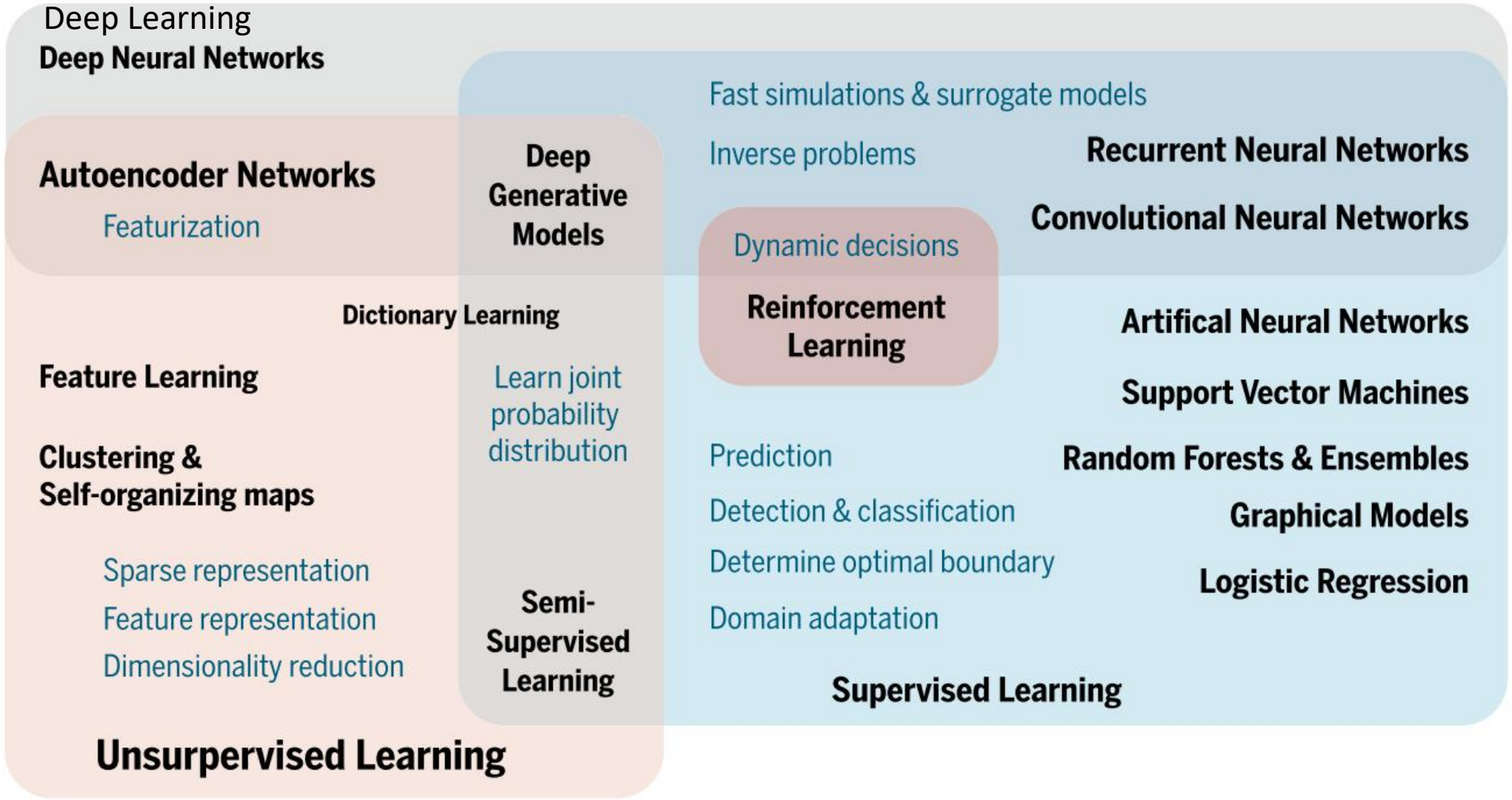
Deep Learning

- 2010
- Uma técnica para implementar ML
- Redes Neurais Profundas (DNNs)
- Estruturas de código em camadas que imitam vagamente o cérebro humano
- Habilidade em extrair automaticamente representações de alto nível de dados complexos

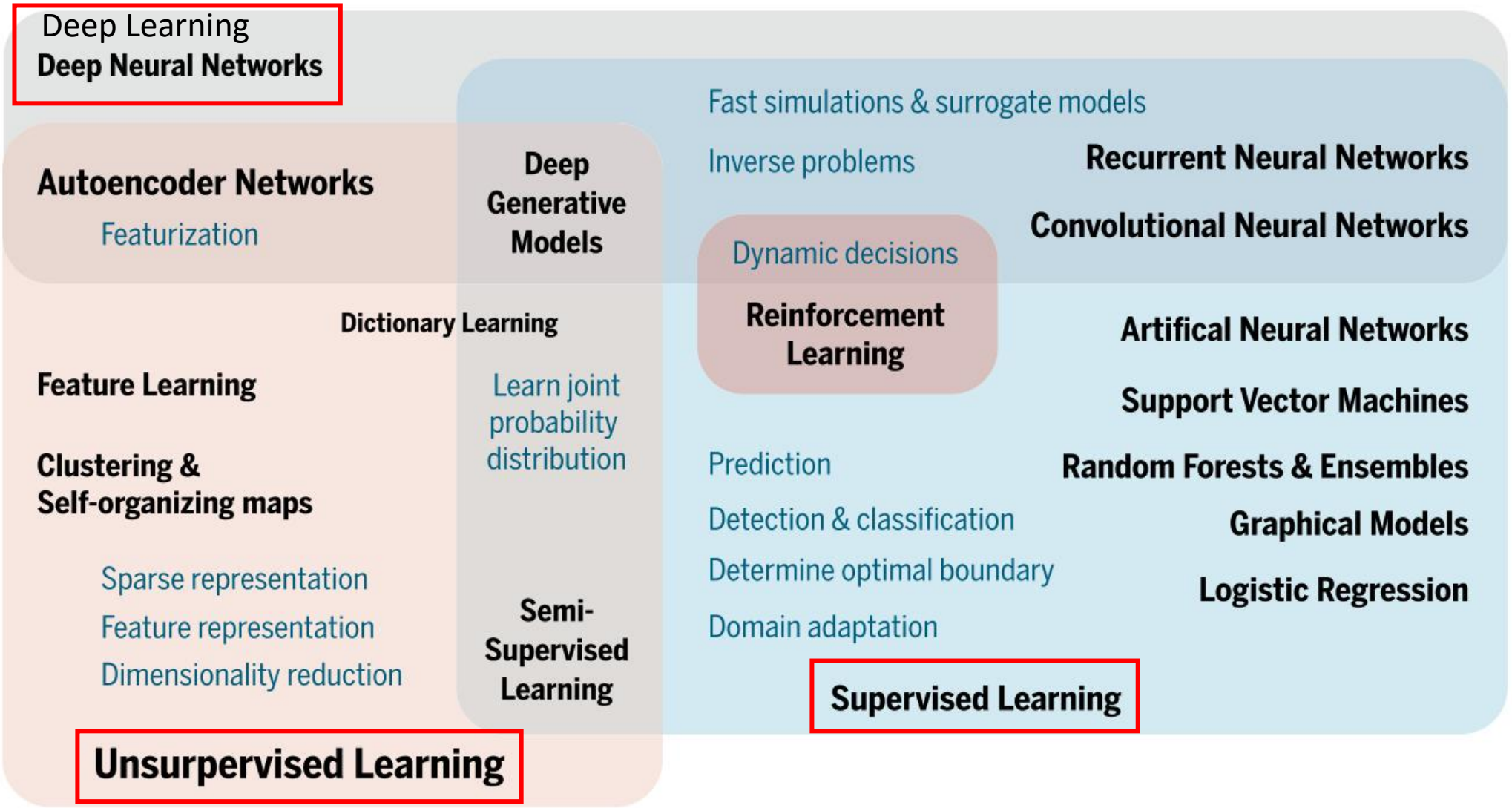




Saiba mais: https://playground.tensorflow.org/

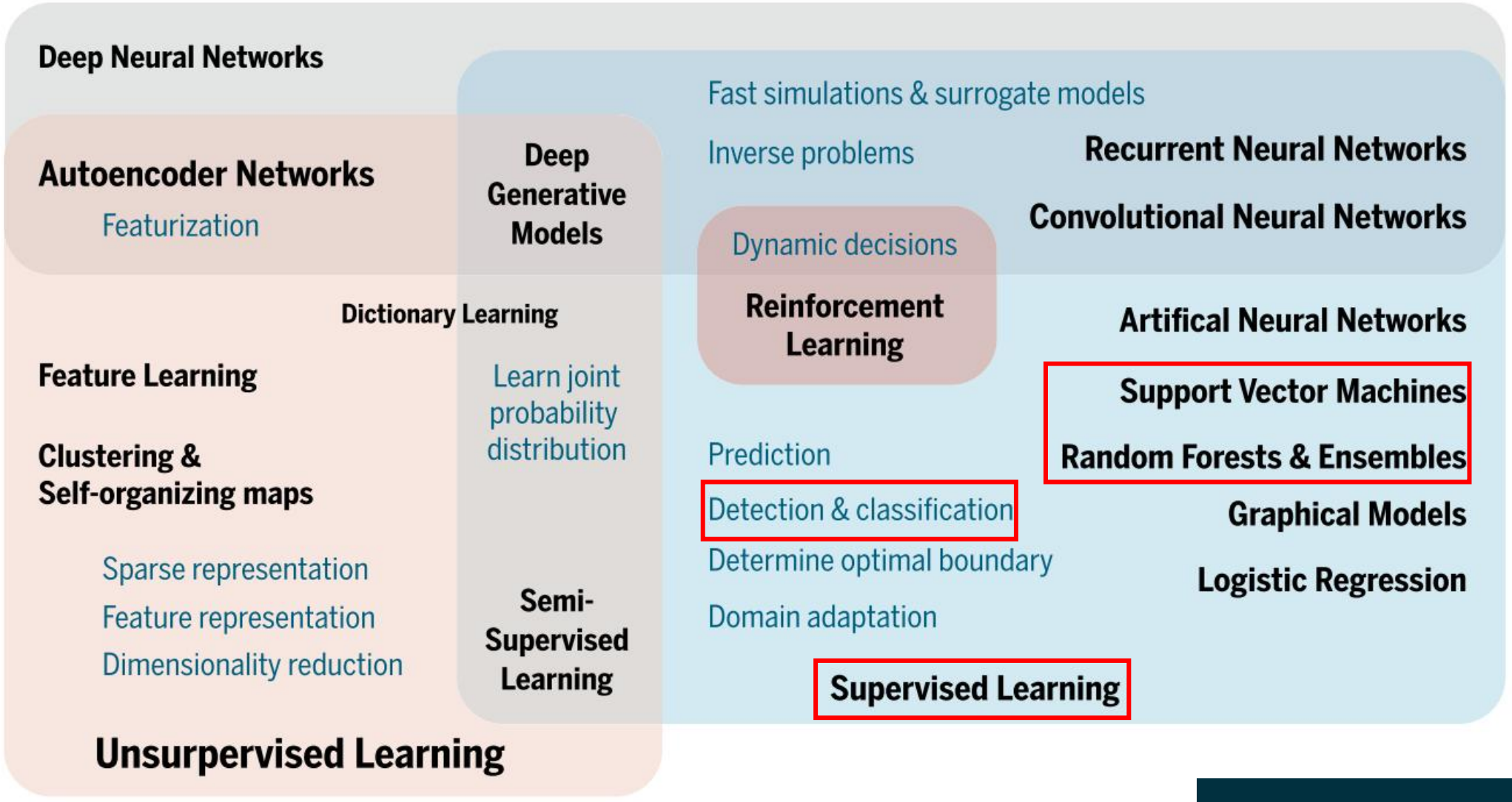


Saiba mais: Bergen et al., 2019

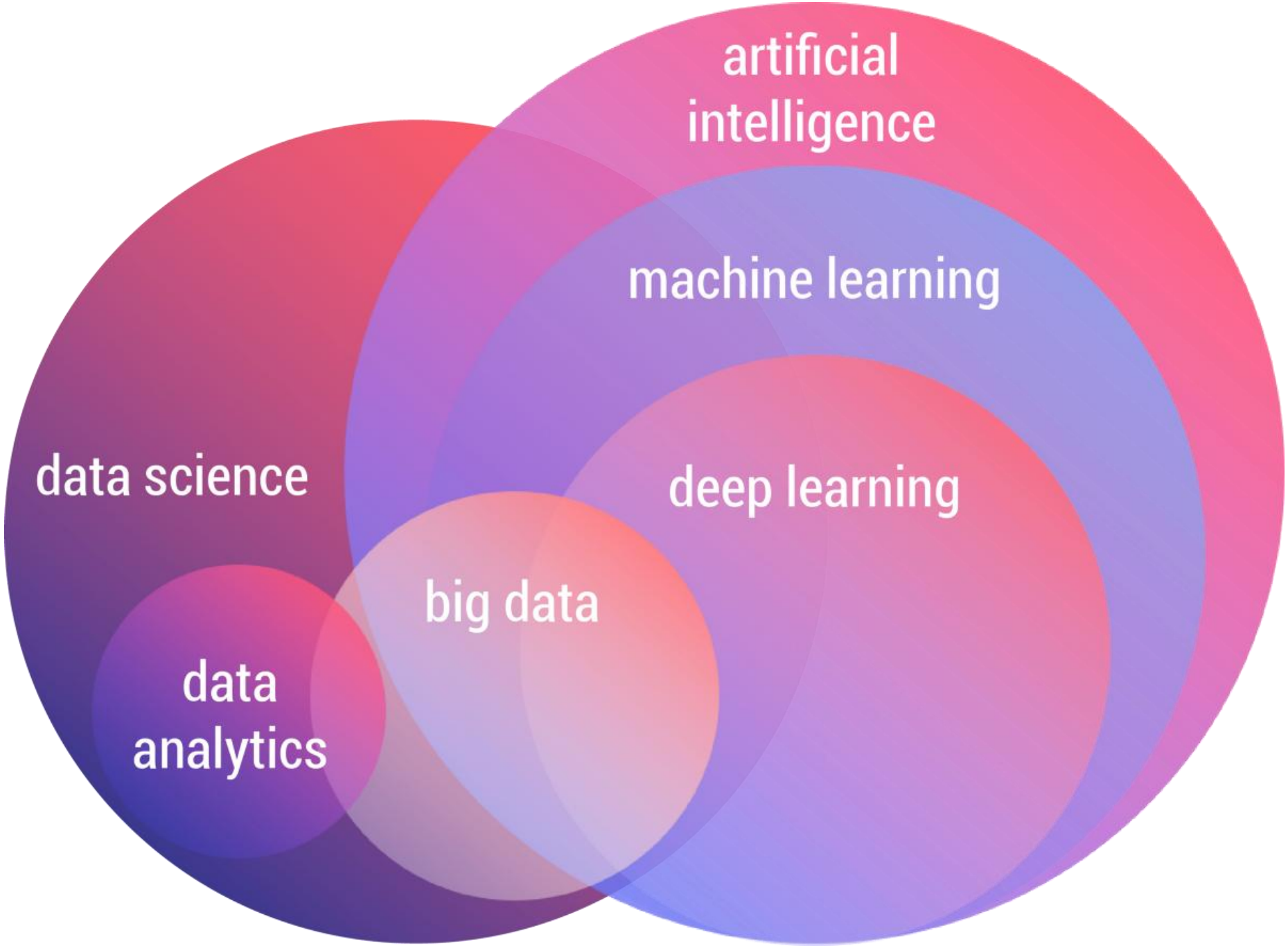


Saiba mais: Bergen et al., 2019

Exercício de hoje



Saiba mais: Bergen et al., 2019


artificial
intelligence
machine learning
data science
deep learning
big data
data
analytics

Os desafios

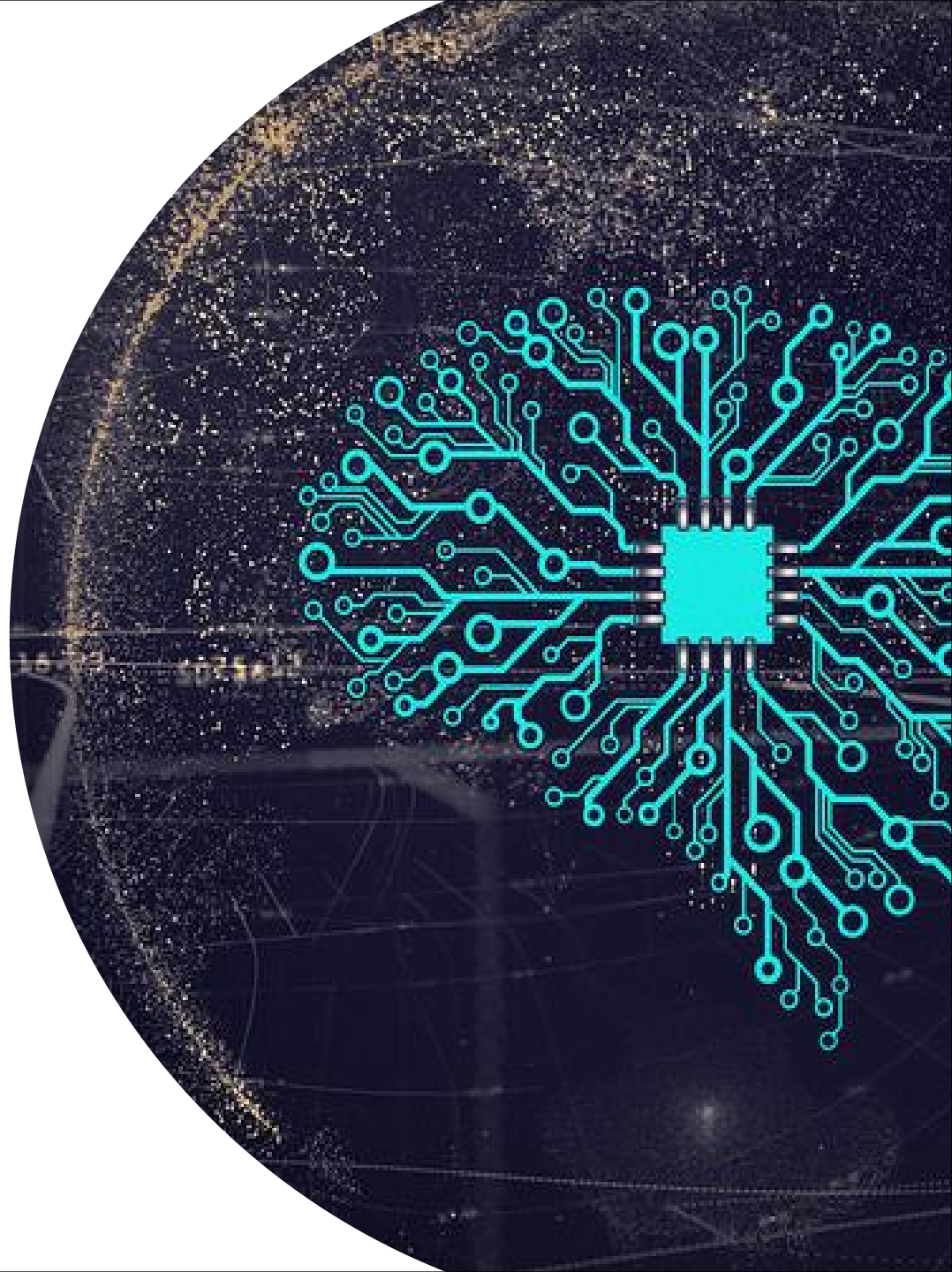
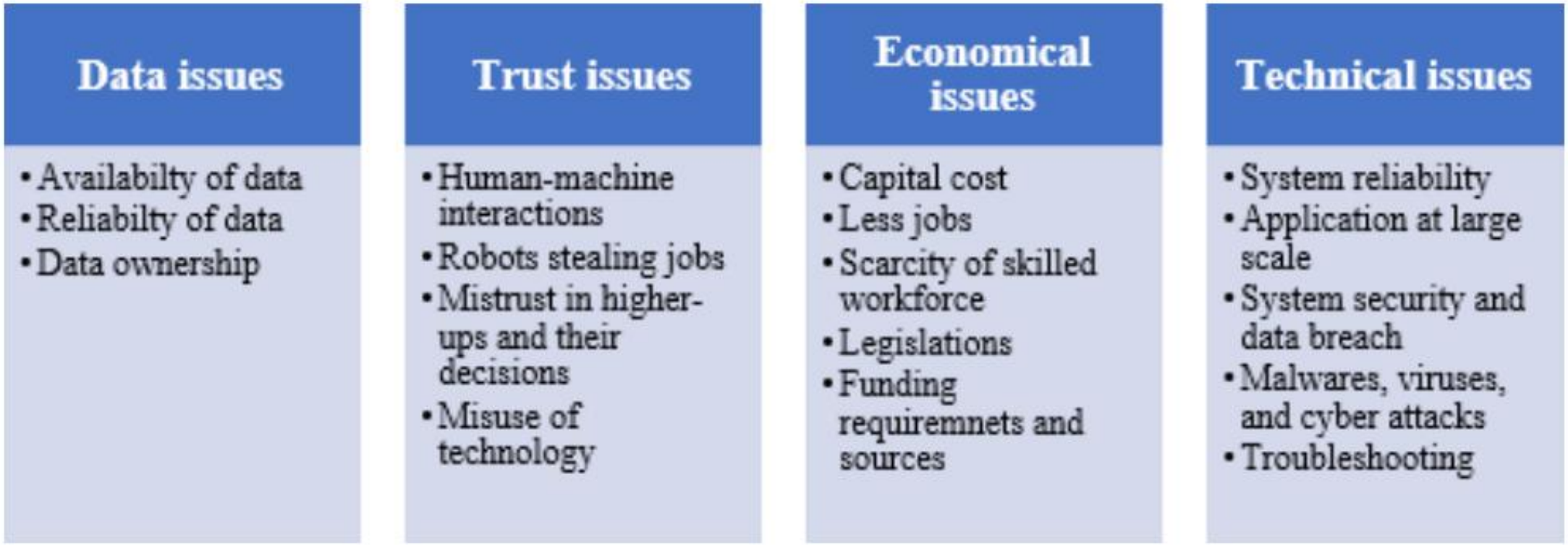


Figure 4. Some of the major challenges in implementing AI and machine learning in mining



20 entrevistas

Saiba mais: Hyder et al., 2019

Os desafios dos dados geológicos

NATUREZA

- Dados em sequência - espacial / temporal
- Multivariados
- Não-lineares
- Não-estacionários
- Múltiplas escalas espaciais e temporais
- Interconexões complexas
- Eventos raros

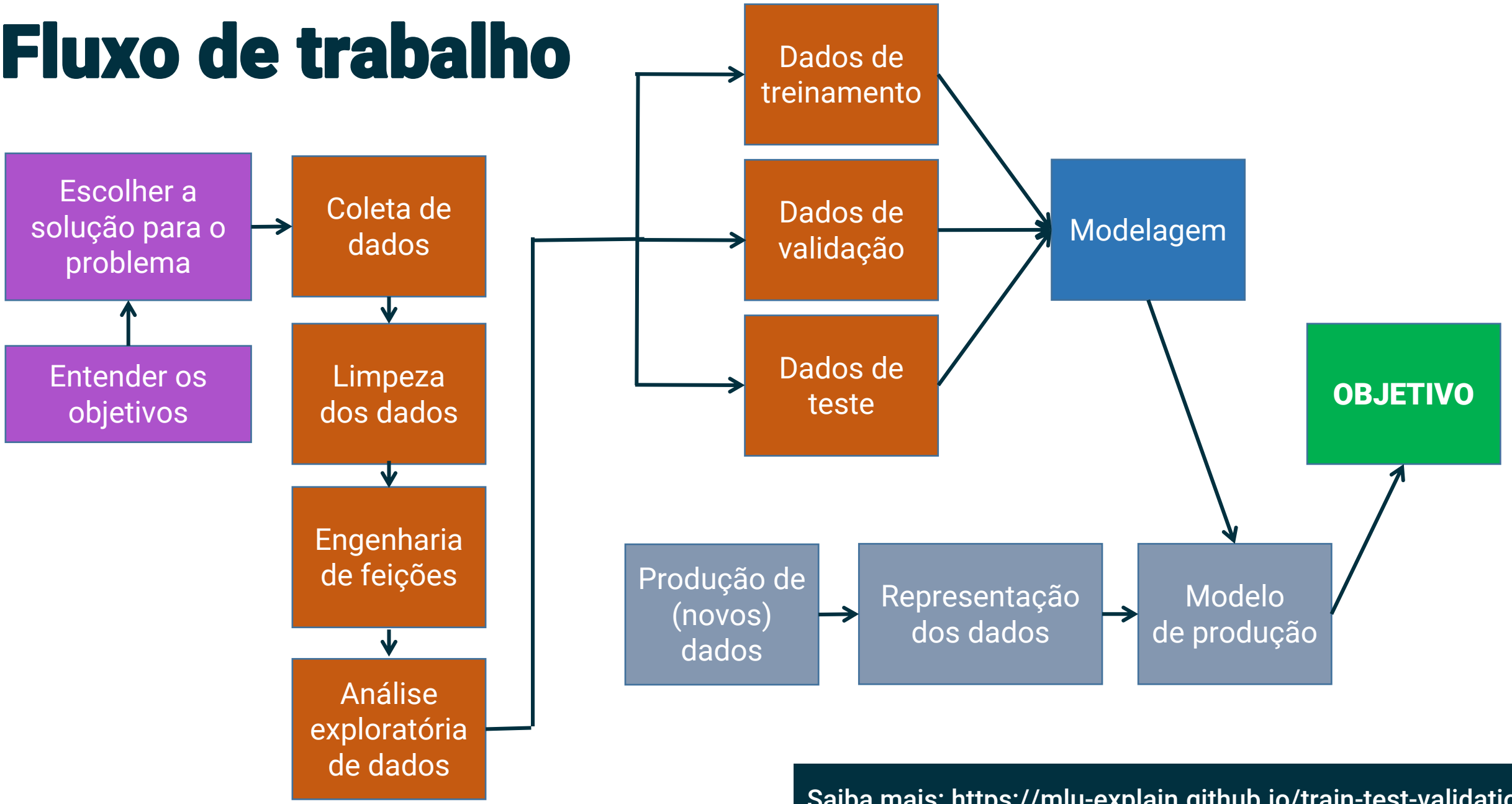
COLETA

- Várias escalas espaciais e temporais
- Várias fontes de erros e ruídos
- Incompletos
- Incertos

ARMAZENAMENTO

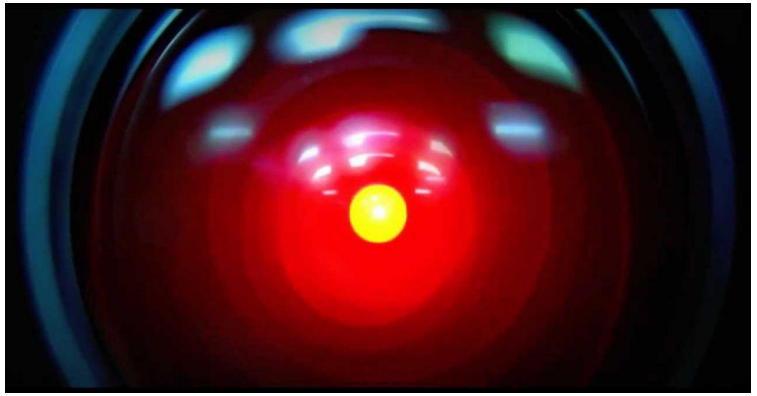
- Poucos registros adequados
- Falta de padrão ouro de referência

Fluxo de trabalho

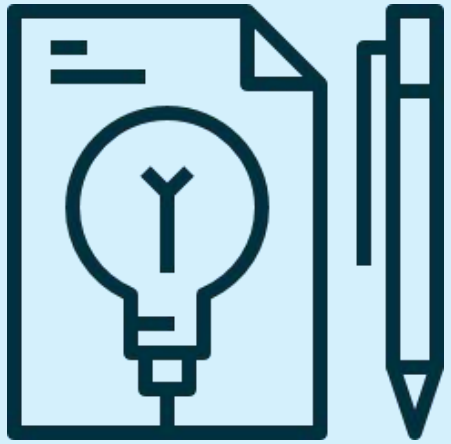


Saiba mais: https://mlu-explain.github.io/train-test-validation/

Fluxo de trabalho



IA não sabe os problemas que pode resolver!



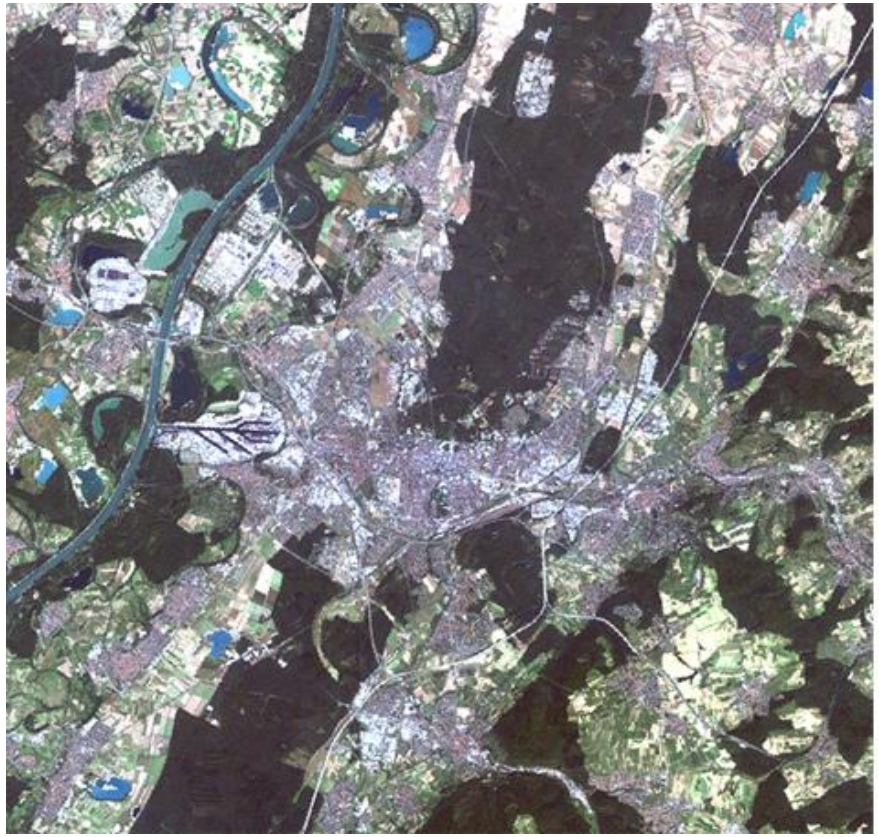
Classificação

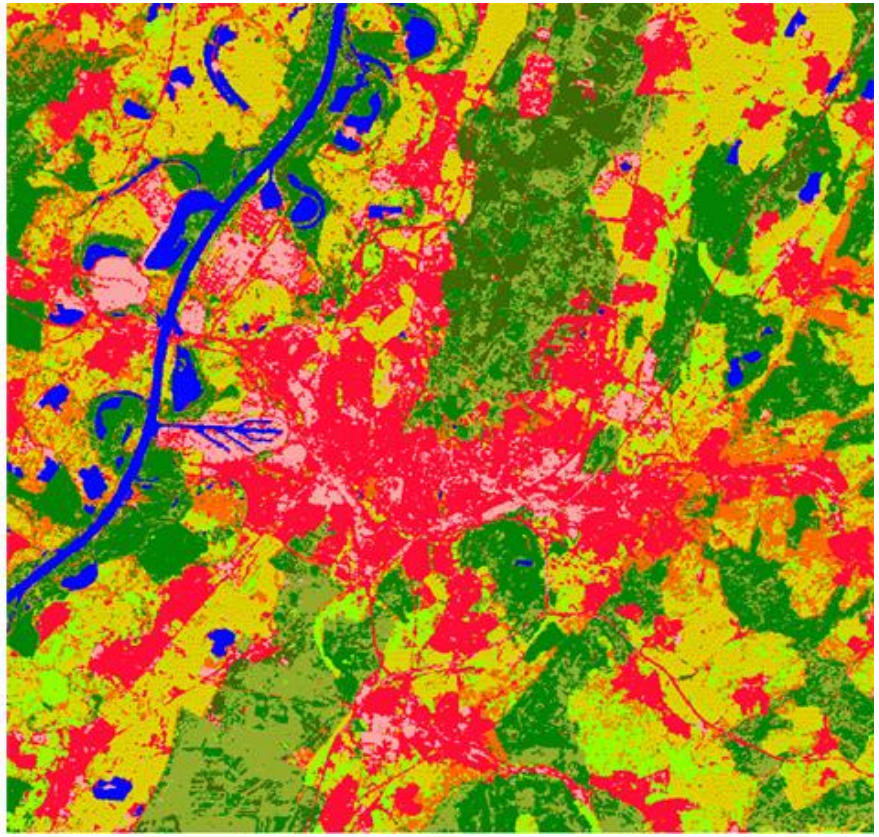
Classification

Os dados pertencem a classes, e o objetivo é encontrar padrões que permitam determinar se um dado pertence a uma ou outra(s) classe(s).

Classificação

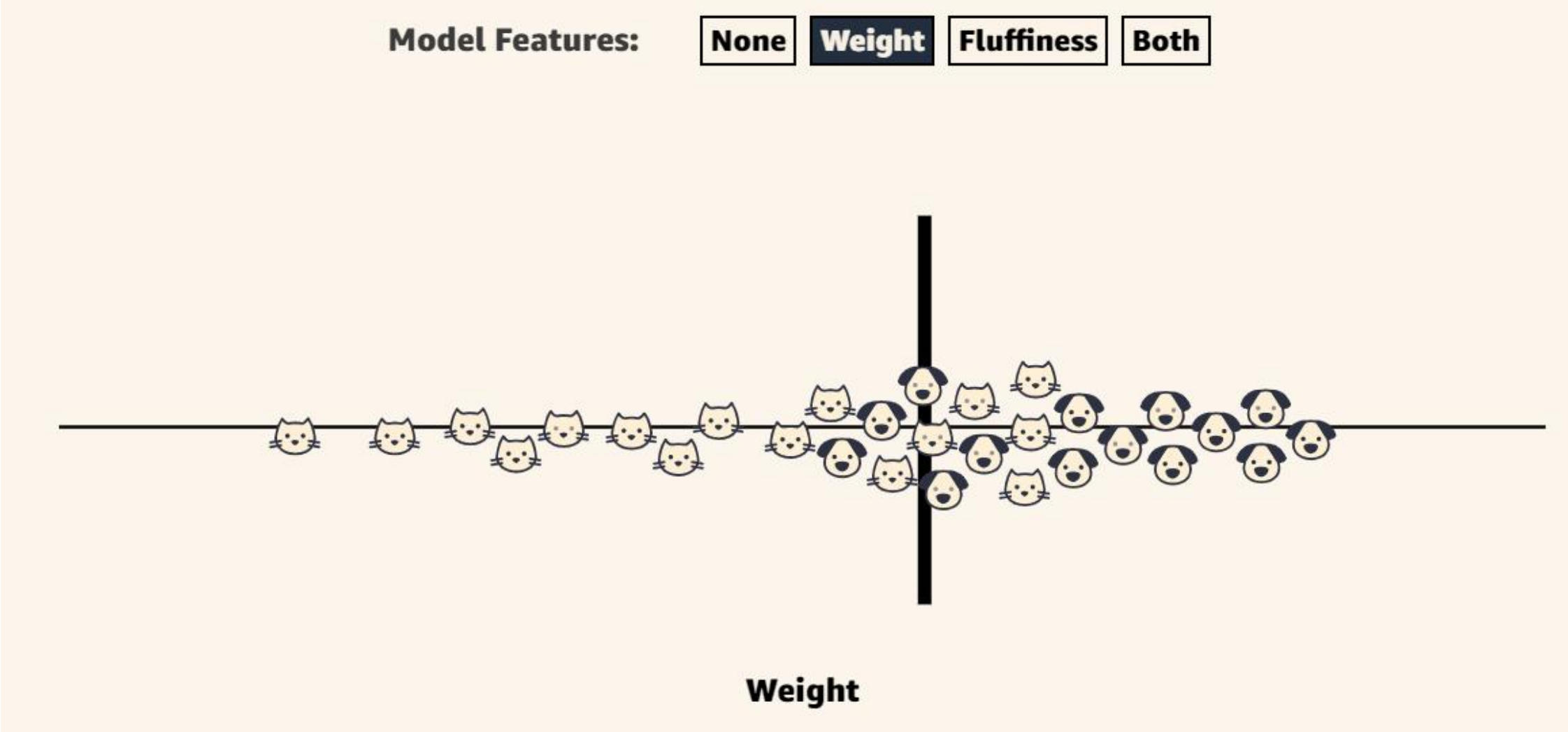
- Predição de dados categóricos
- O resultado da classificação digital de imagens de Sensoriamento Remoto é um mapa temático





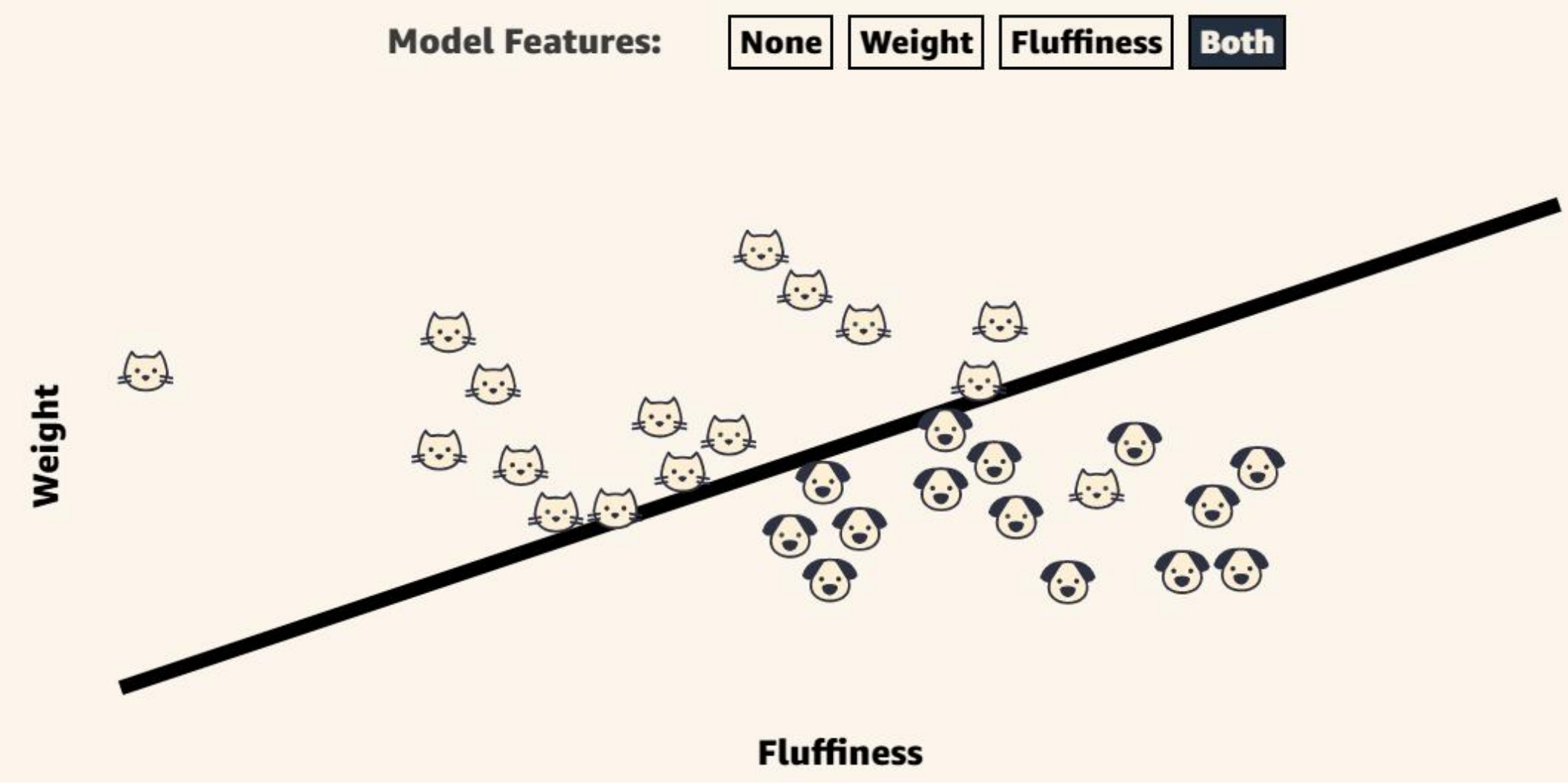
Explore: https://mlu-explain.github.io/train-test-validation/

Técnicas e métodos



Explore: https://mlu-explain.github.io/train-test-validation/

Técnicas e métodos



Técnicas e métodos

- Dados 0D:
  - Sem dimensões espaciais ou temporais
  - Ex.: geoquímicos ou petrofísicos
  - Redes Neurais Artificiais (ANN) e/ou variações
- Dados 1D:
  - Uma dimensão espacial ou temporal
  - Ex.: log de furo de sondagem; sísmica (série temporal)
  - ANN e variantes; ANN, *Support Vector Machines (SVM)*, *Self-organizing maps*
- Dados 2D:
  - Duas dimensões espaciais
  - Ex.: processamento de imagens
  - ANN, SVM

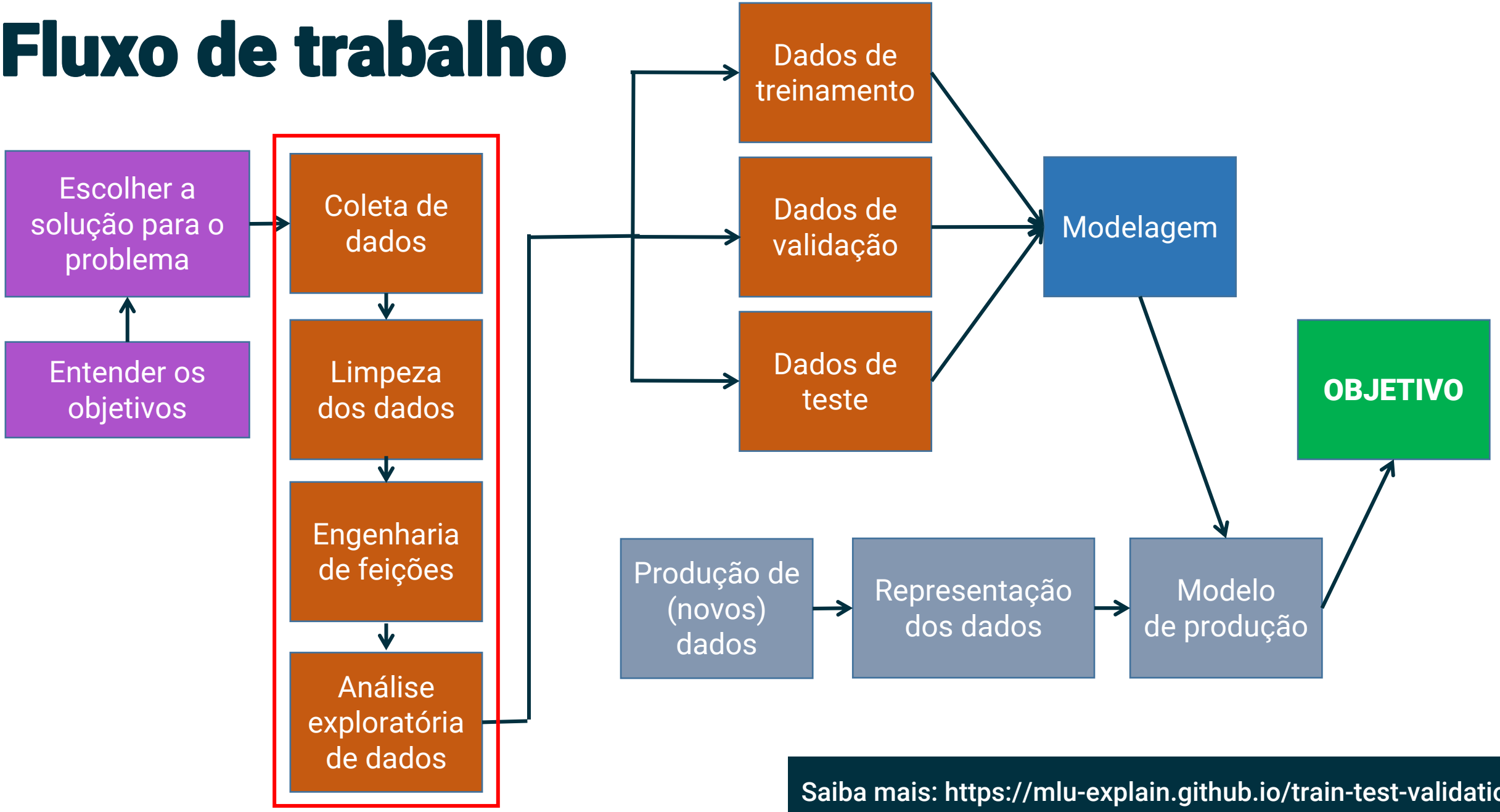
Saiba mais: Cracknell (2014), pág. 45

Técnicas e métodos

Classification	Minimum Distance	AVIRIS, WorldView-3	Minerals	General exploration	Kruse and Perry (2013)
		Hyperion, ASTER, Landsat 8	Lithological units	General exploration	Pan et al. (2019)
		ASTER	Lithological units	Cr	Othman and Gloaguen (2014)
		ASTER	Alteration types, minerals	Au	Xu et al. (2019)
		Hyperion	Alteration types, minerals	Au	Wang and Zheng (2010)
	Support Vector Machines	Sentinel-2	Lithological units, alteration types, minerals	Cu, Au	Abdolmaleki et al. (2020)
		Sentinel-2	Lithological units	Li	Cardoso-Fernandes et al. (2020b)
		UAV	Lithological units, alteration types	General exploration	Lorenz et al. (2021)
		Landsat 5	Minerals	Au	Rigol-Sanchez et al. (2003)
	Simple Neural Networks	Hyperion, Landat 5	Lithological units	General exploration	Leverington (2010)
		Landsat 7	Alteration types, minerals	Mo, Pb, Zn, Ag	Wang et al. (2010)
		Landsat 7	Structures	General exploration	Borisova et al. (2014)
	Random Forest	Sentinel-2	Lithological units	Li	Cardoso-Fernandes et al. (2019)
		Landsat 5	Lithological units	Au	Kuhn et al. (2018)
		Landsat 7	Lithological units	General exploration	Cracknell and Reading (2014)
		ASTER, Sentinel-2	Lithological units	Rare metals	Wang et al. (2020b)
		Sentinel-2, PALSAR	Lithological units	General exploration	Bachri et al. (2020)
		TOPSAR	Lithological units	General exploration	Radford et al. (2018)

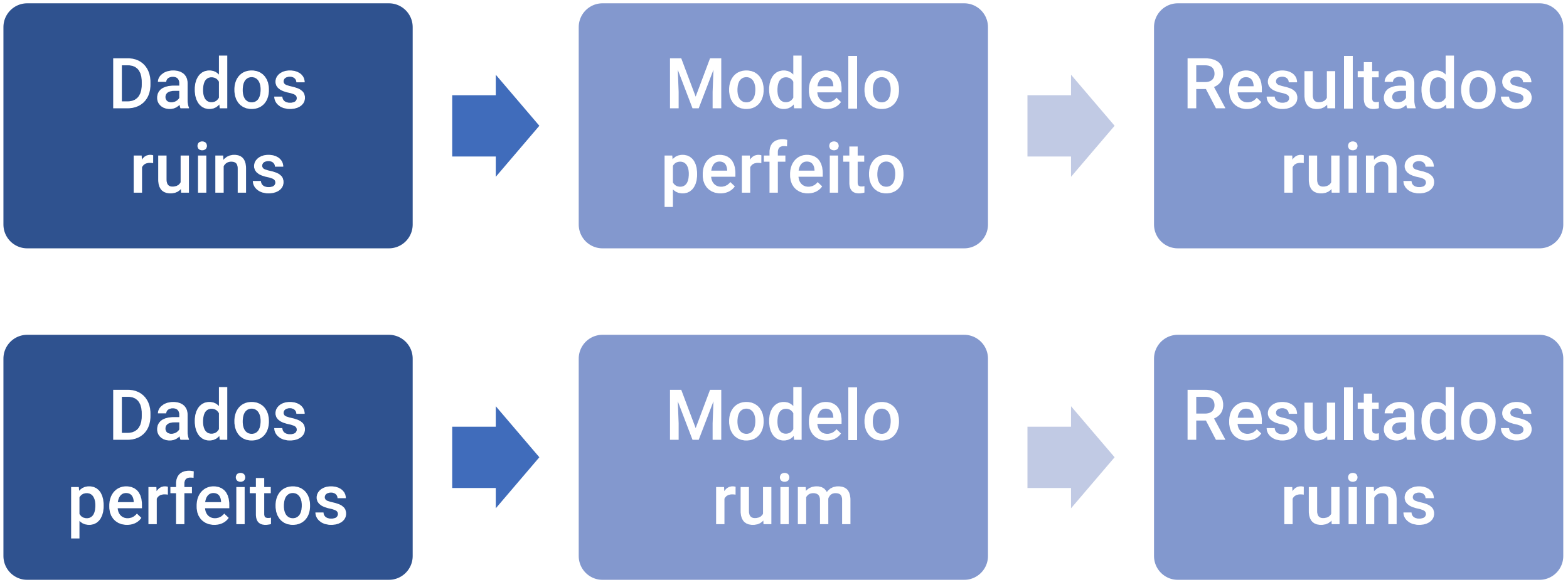
Saiba mais: Shirmard et al. (2022)

Fluxo de trabalho



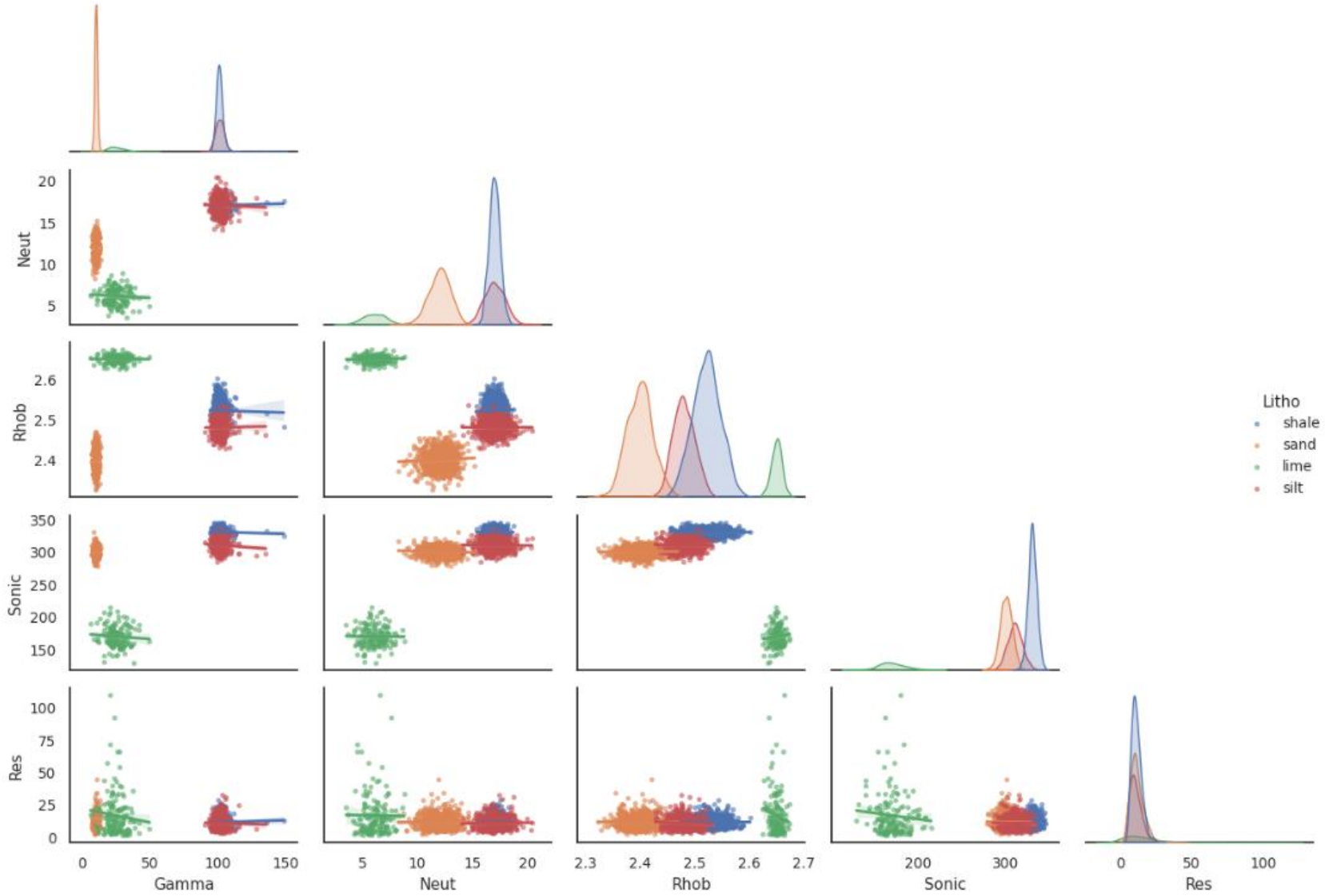
Saiba mais: https://mlu-explain.github.io/train-test-validation/

Garbage in, garbage out

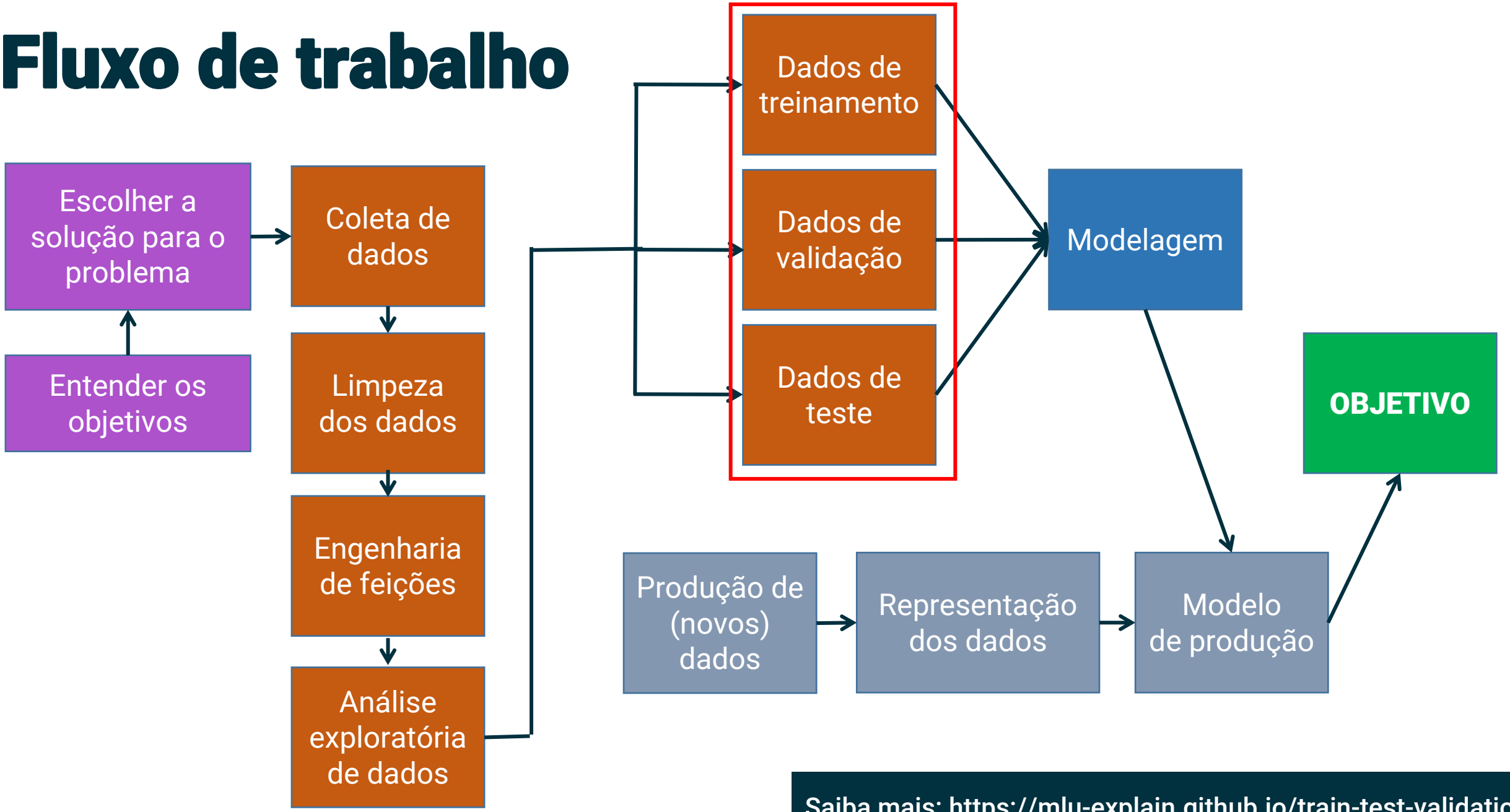


	Litho	Gamma	Neut	Rhob	Sonic	Res
0	shale	103.493505	16.638491	2.559117	339.089517	21.514267
1	shale	100.570423	16.654216	2.477209	342.483419	6.959325
2	shale	105.908270	17.000423	2.520837	331.018479	15.673613
3	shale	100.734513	17.069938	2.519240	336.926799	9.013816
4	shale	106.816393	17.082515	2.529335	328.066325	6.436128
...	...	...	...	...	...	...
2468	silt	101.923845	17.933285	2.482602	302.882393	15.693331
2469	silt	106.556193	16.962746	2.480323	311.355489	13.305511
2470	silt	106.245301	15.379808	2.464955	324.200124	13.257587
2471	silt	101.664183	17.831292	2.464230	321.297745	20.291701
2472	silt	102.830445	15.718543	2.452138	298.667458	16.973809

2473 rows × 6 columns


Litho
shale
sand
lime
silt
Gamma
Neut
Rhob
Sonic
Res

Fluxo de trabalho



Saiba mais: https://mlu-explain.github.io/train-test-validation/

Dados de treinamento, validação e teste

- Treinamento: usado para treinar o modelo. O conjunto de dados que alimentamos nosso modelo para aprender possíveis padrões e relacionamentos subjacentes.
- Validação: usado para entender o desempenho de diferentes tipos de modelos e escolhas de hiperparâmetros.
- Teste: usado para comprovar que aquele modelo realmente funciona. São dados ignorados no treinamento e no processo de escolha de hiperparâmetros. O conjunto de dados que usamos para aproximar a precisão imparcial do nosso modelo.

Saiba mais: https://mlu-explain.github.io/train-test-validation/

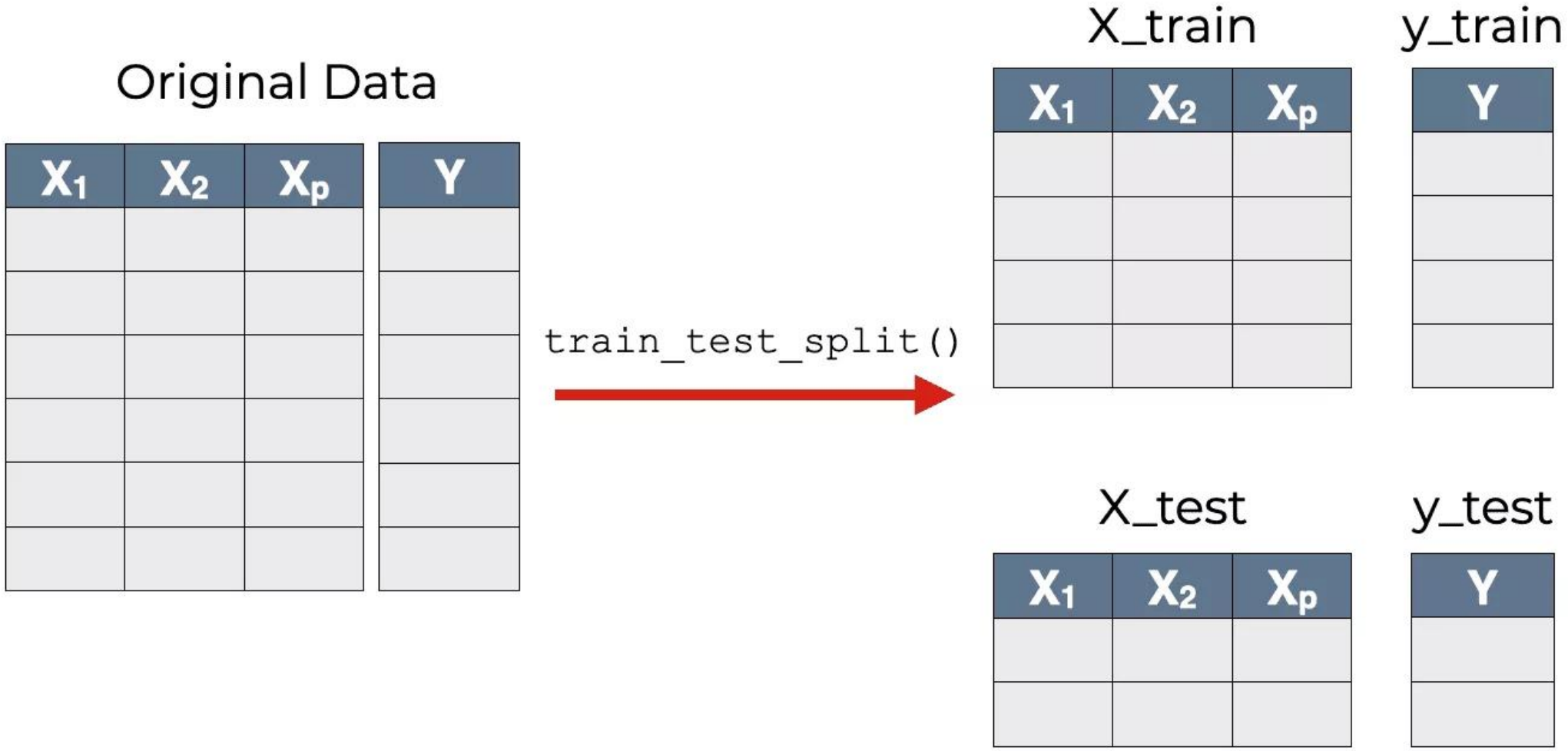
variável alvo

	Litho	Gamma	Neut	Rhob	Sonic	Res
0	shale	103.493505	16.638491	2.559117	339.089517	21.514267
1	shale	100.570423	16.654216	2.477209	342.483419	6.959325
2	shale	105.908270	17.000423	2.520837	331.018479	15.673613
3	shale	100.734513	17.069938	2.519240	336.926799	9.013816
4	shale	106.816393	17.082515	2.529335	328.066325	6.436128
...	...	...	...	...	...	...
2468	silt	101.923845	17.933285	2.482602	302.882393	15.693331
2469	silt	106.556193	16.962746	2.480323	311.355489	13.305511
2470	silt	106.245301	15.379808	2.464955	324.200124	13.257587
2471	silt	101.664183	17.831292	2.464230	321.297745	20.291701
2472	silt	102.830445	15.718543	2.452138	298.667458	16.973809

2473 rows × 6 columns

Padronizar escalas

	Gamma	Neut	Rhob	Sonic	Res
0	0.684067	0.775722	0.674410	0.972804	0.188691
1	0.663547	0.776651	0.440358	0.988555	0.054524
2	0.701019	0.797103	0.565026	0.935345	0.134852
3	0.664699	0.801209	0.560462	0.962767	0.073462
4	0.707394	0.801952	0.589309	0.921644	0.049701



X_train, X_test, y_train, y_test = train_test_split(df.data, df.target, test_size=0.3,random_state=130)

Data leakage

- Informações compartilhadas entre o conjunto de treino e teste
- Modelo vai ser “contaminado” por informações que não deveriam estar ali --> performance do seu conjunto de testes funcione bem, mas não se mantenha no ambiente de produção

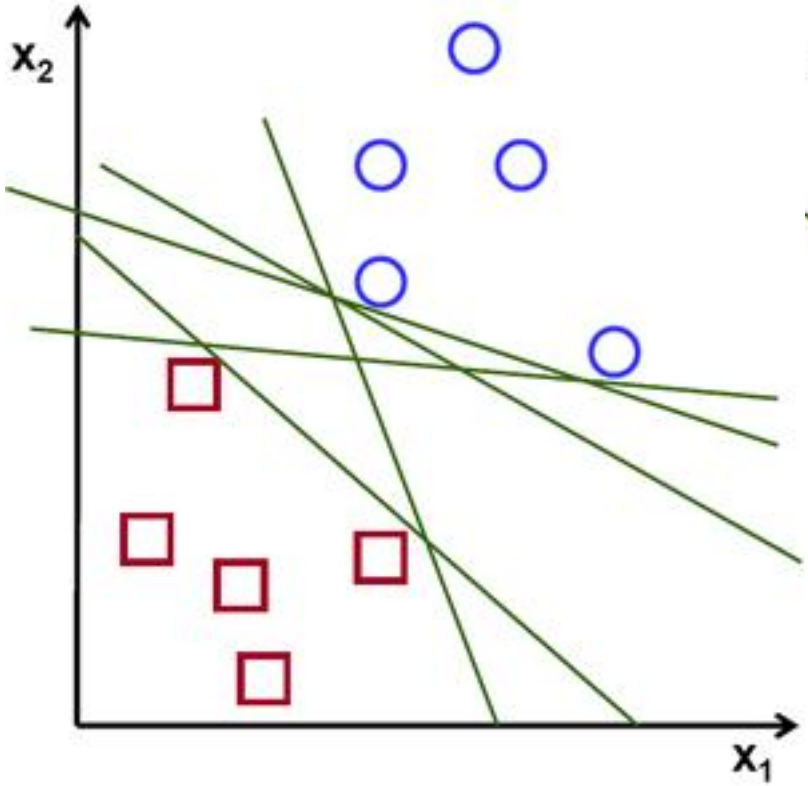
Saiba mais: https://towardsdatascience.com/data-leakage-in-machine-learning-how-it-can-be-detected-and-minimize-the-risk-8ef4e3a97562

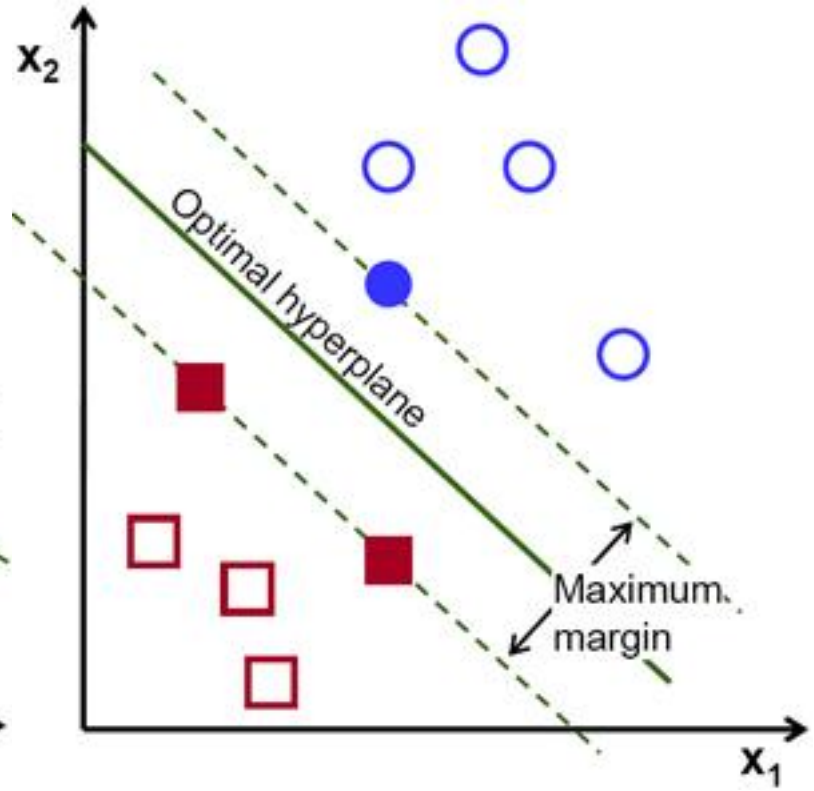
Treinar os modelos - GMM

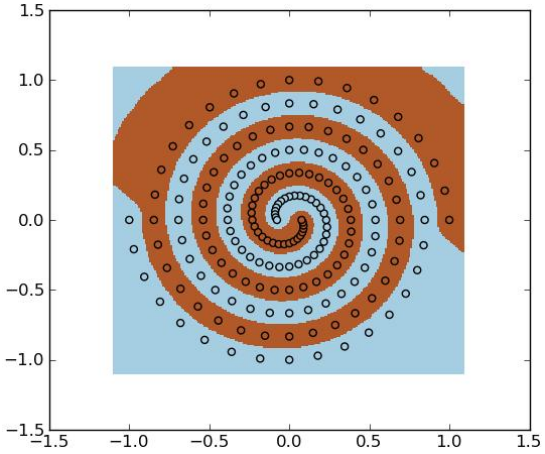
```
# Definir o GMM com quatro componentes
GMM = GaussianMixture(n_components=4)

# Treinar o modelo usando os conjuntos de treino
GMM.fit(X_train, y_train)

# Prever a resposta para o conjunto de teste
y_pred = GMM.predict(X_test)

# Plotar dados previstos e de teste
#plt.plot(y_test,y_pred,'ro',alpha=0.01)
```

Support Vector Machines - SVM

- O objetivo do algoritmo SVM é encontrar um hiperplano em um espaço N-dimensional (N — o número de recursos) que classifique distintamente os pontos de dados
- Regressão e classificação



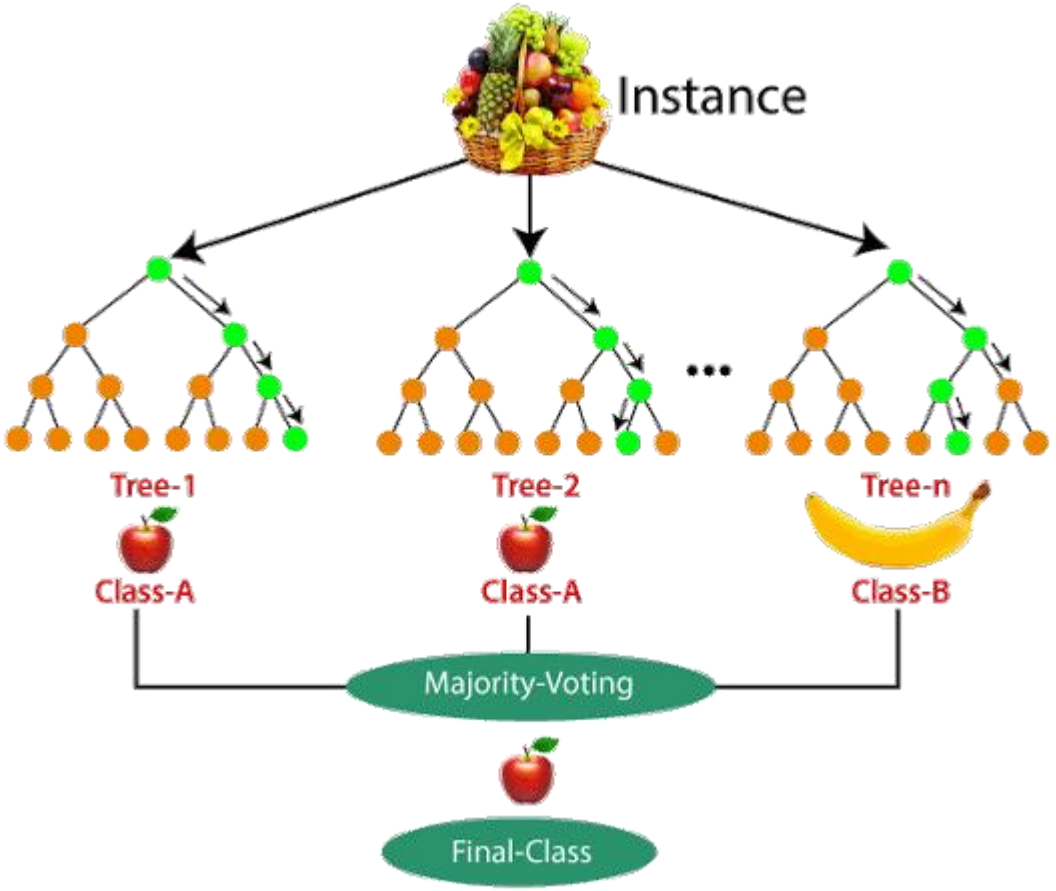




Saiba mais: https://towardsdatascience.com/support-vector-machine-introduction-to-machine-learning-algorithms-934a444fca47

Random Forest - RF

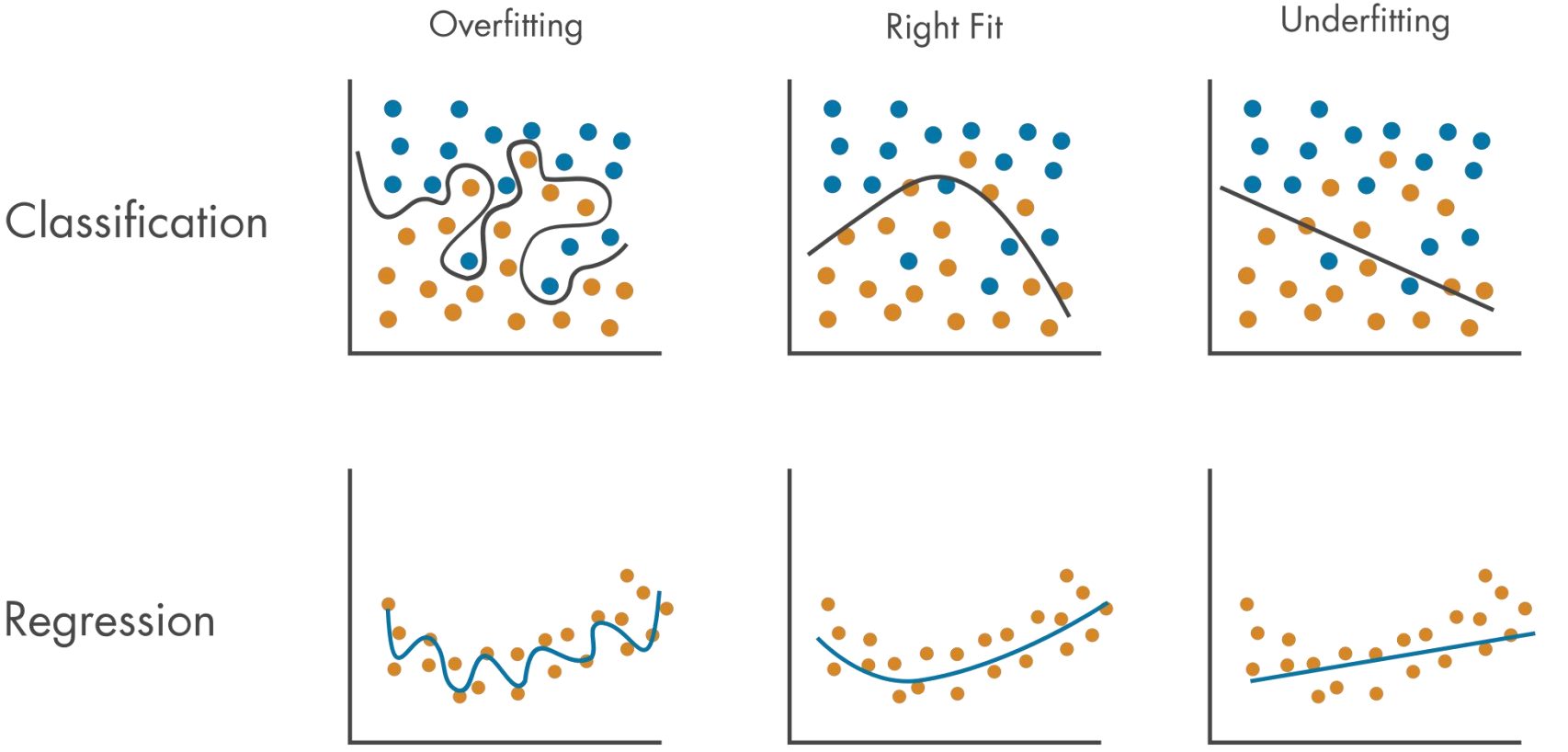
- Combina saída de múltiplas árvores de decisão
- *Ensemble learning*: métodos feitos por um conjunto de classificadores —e.g. árvores de decisão — e suas previsões são agregadas para identificar o resultado mais popular
- Regressão, classificação e outras tarefas



Saiba mais: https://towardsdatascience.com/support-vector-machine-introduction-to-machine-learning-algorithms-934a444fca47

Overfitting e Underfitting

- Sub-ajustado: modelo não aprendeu o suficiente sobre os dados
- Sobre-ajustado: performance cai no conjunto de teste



- Algoritmo muito complexo para os dados
- Poucos dados de treinamento
- Ruído nos dados de treinamento

- Algoritmo pouco poderoso para os dados
- Características não representativas
- Modelo muito restrito

Saiba mais: https://abracd.org/overfitting-e-underfitting-em-machine-learning/

Métricas de avaliação de desempenho

- Accuracy $\frac{(TP+TN)}{total}$
  - Com que frequência está correto
- Precision $\frac{TP}{predictedyes}$
  - Quando prevê classe, com que frequência está correto
- Recall $\frac{TP}{actualyes}$
  - Quantos dos casos de classe nos dados foram corretamente identificados
- F1 $2 * \frac{Precision * Recall}{Precision + Recall}$
  - Métrica combinada

n=165	Predicted: NO	Predicted: YES	
Actual: NO	TN = 50	FP = 10	60
Actual: YES	FN = 5	TP = 100	105
	55	110	

Saiba mais: https://towardsdatascience.com/understanding-confusion-matrix-a9ad42dcfd62

Métricas de avaliação de desempenho

GMM

col_0 / Litho	lime	sand	shale	silt
lime	41	0	0	0
sand	0	216	0	0
shale	0	0	290	12
silt	0	0	5	178

GMM
Accuracy: 0.977088948787062

	Litho	Recall	Precision	F1
0	lime	1.000000	1.000000	1.000000
1	sand	1.000000	1.000000	1.000000
2	shale	0.960265	0.983051	0.971524
3	silt	0.972678	0.936842	0.954424

Support Vector Machine

col_0 / Litho	lime	sand	shale	silt
lime	41	0	0	0
sand	0	216	0	0
shale	0	0	295	7
silt	0	0	7	176

Accuracy: 0.9811320754716981

	Litho	Recall	Precision	F1
0	lime	1.000000	1.000000	1.000000
1	sand	1.000000	1.000000	1.000000
2	shale	0.976821	0.976821	0.976821
3	silt	0.961749	0.961749	0.961749

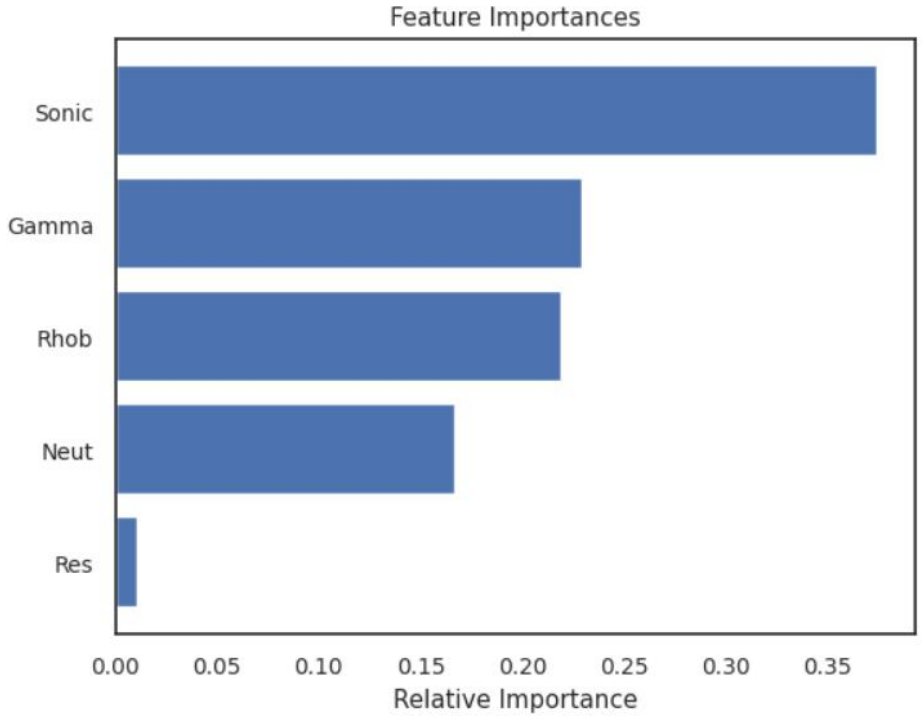
Random Forest

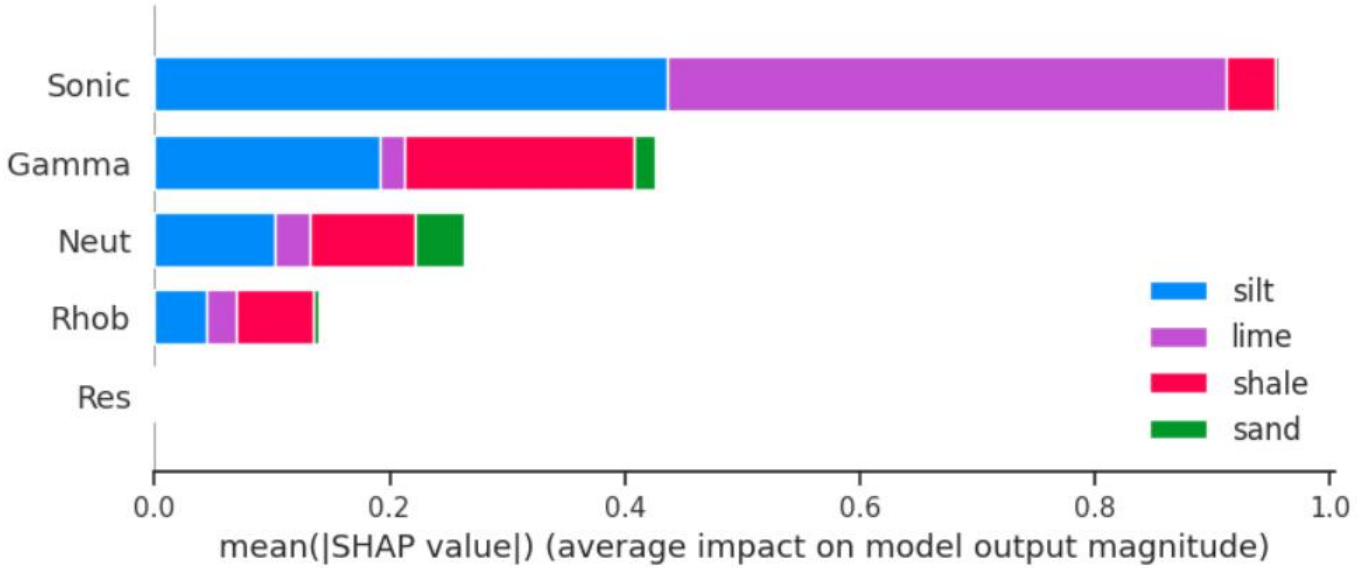
col_0 / Litho	lime	sand	shale	silt
lime	41	0	0	0
sand	0	216	0	0
shale	0	0	292	10
silt	0	0	11	172

Accuracy: 0.9716981132075472

	Litho	Recall	Precision	F1
0	lime	1.000000	1.000000	1.000000
1	sand	1.000000	1.000000	1.000000
2	shale	0.966887	0.963696	0.965289
3	silt	0.939891	0.945055	0.942466

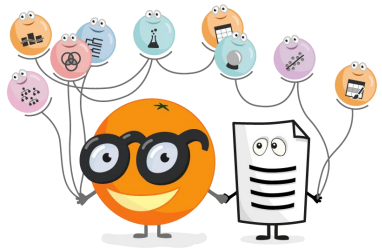
Explainable ML



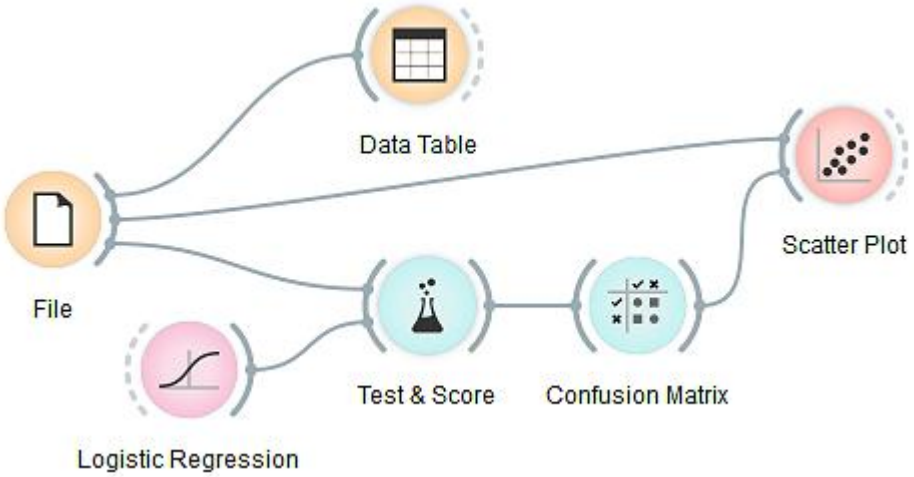


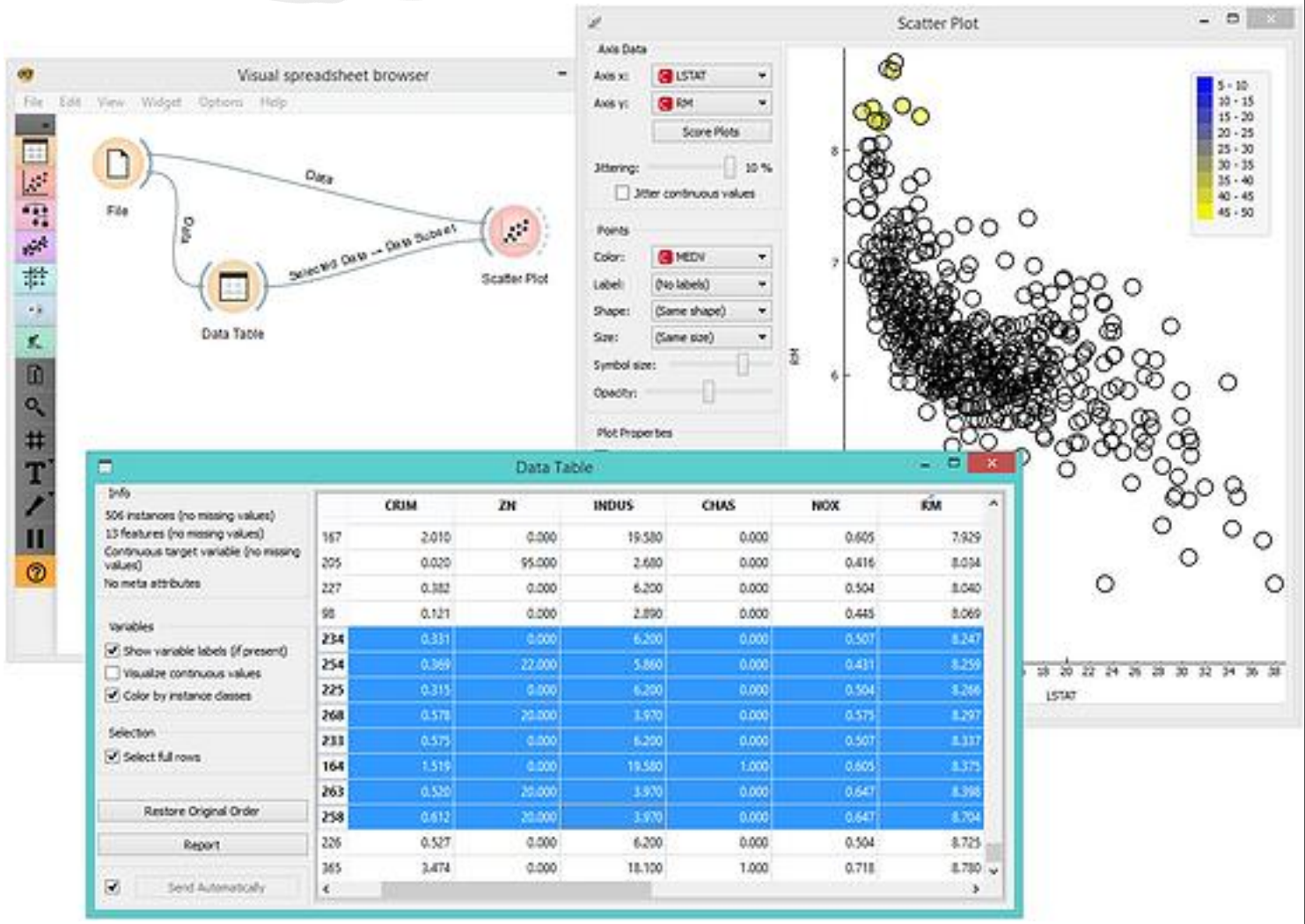
Explainable ML

Opção mais amigável



- Opção open source para aprendizado de máquina e visualização de dados

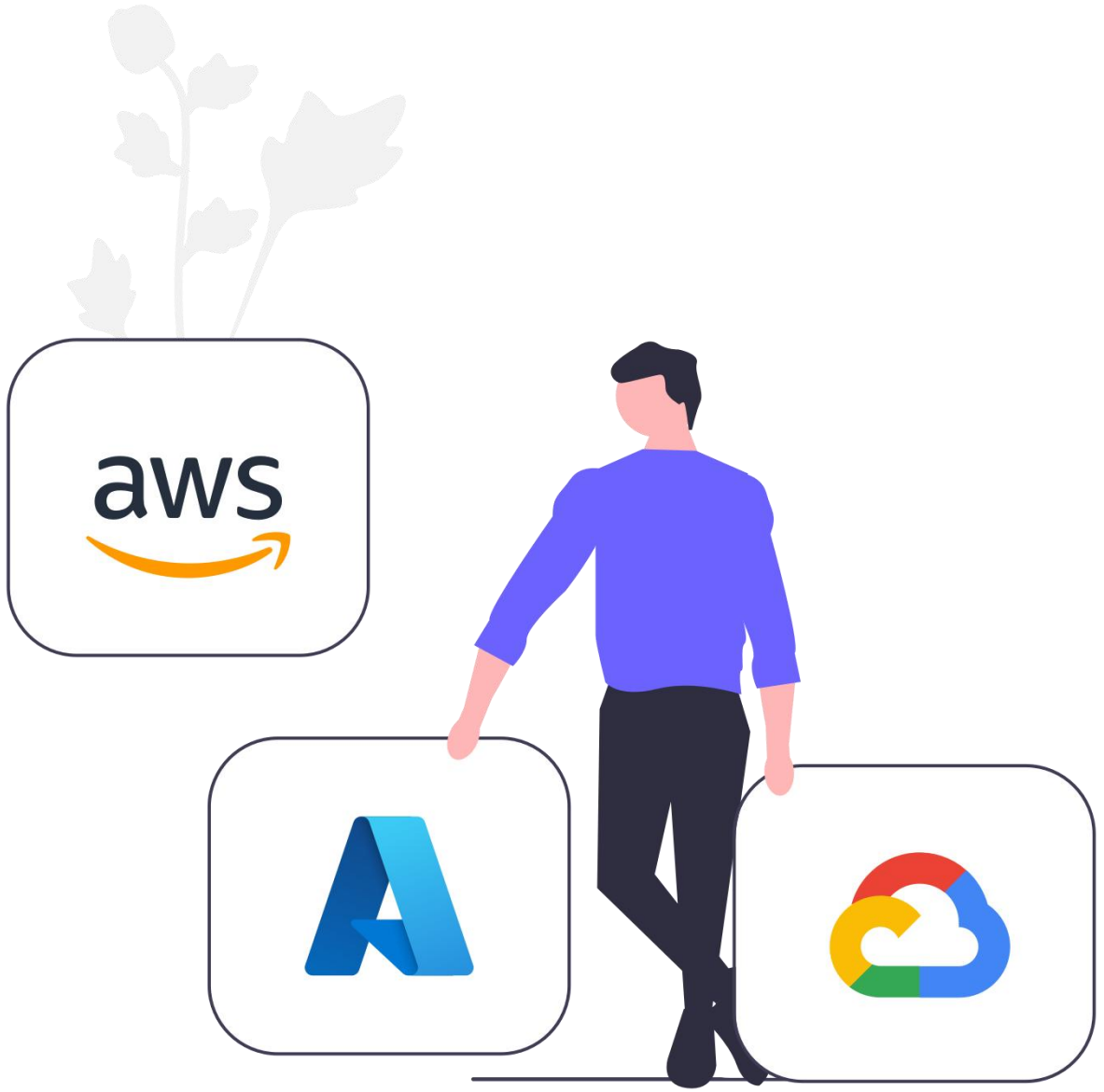




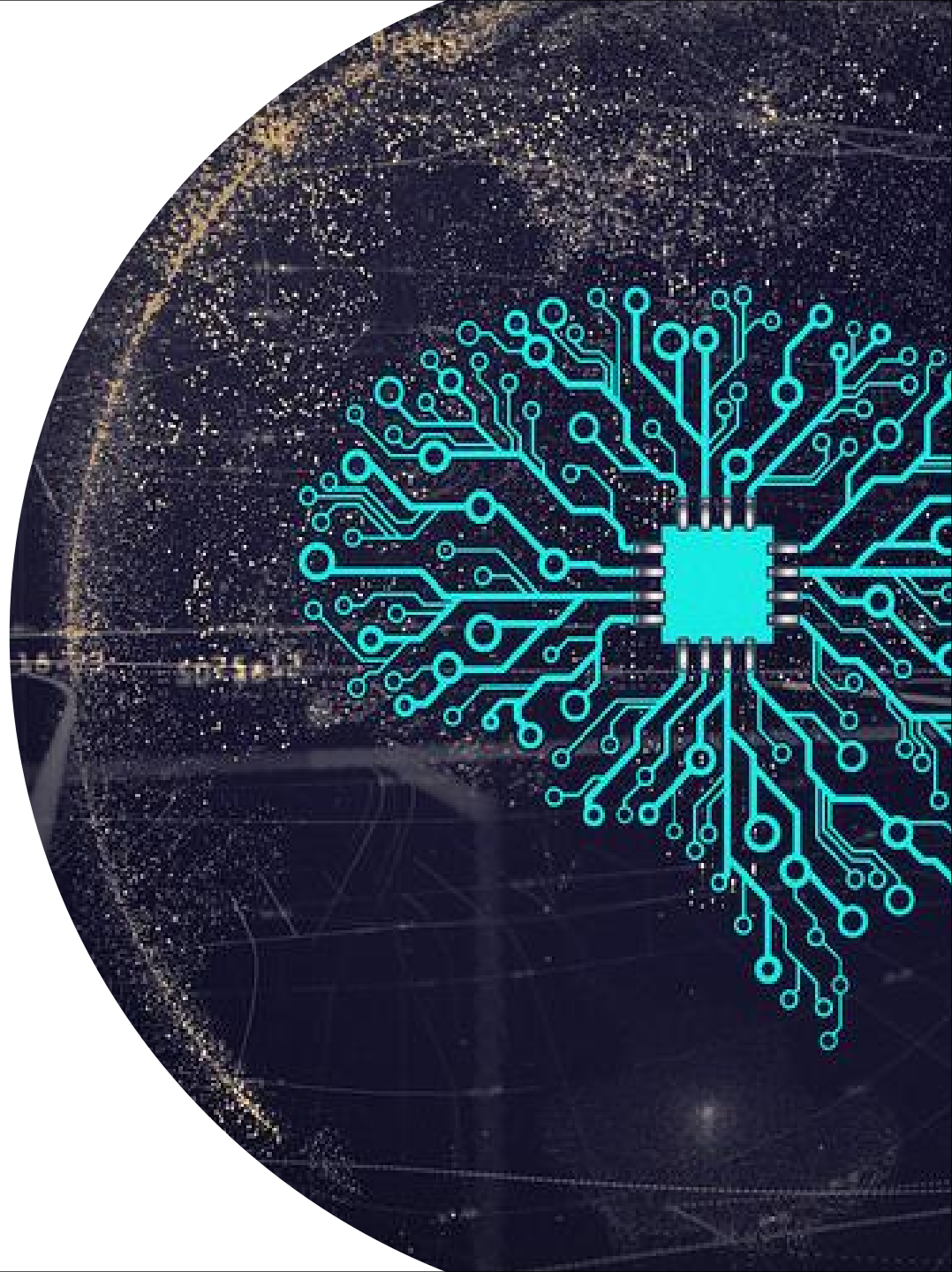
Saiba mais: https://orangedatamining.com/

Escalabilidade

- Complexidade computacional: tempo de processamento e/ou espaço de armazenamento
- Algumas soluções são proibitivas



O futuro: Uma IA geral?



~~4º~~ 5° paradigma da ciência

Era em que os sistemas artificiais (ou seja, as máquinas) poderão gerar conhecimento, com pouca ou nenhuma intervenção humana

Approaching the Fifth Paradigm of Cognitive Applications

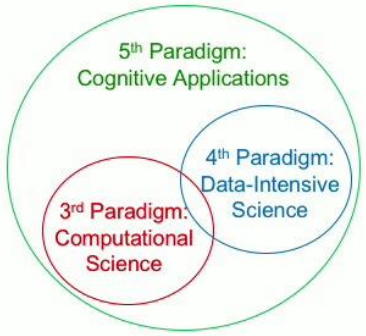


Figure 1: The Fifth Paradigm

The consolidation of HPC and Big Data machine learning technologies represents the prerequisite for developing the next paradigm of cognitive applications

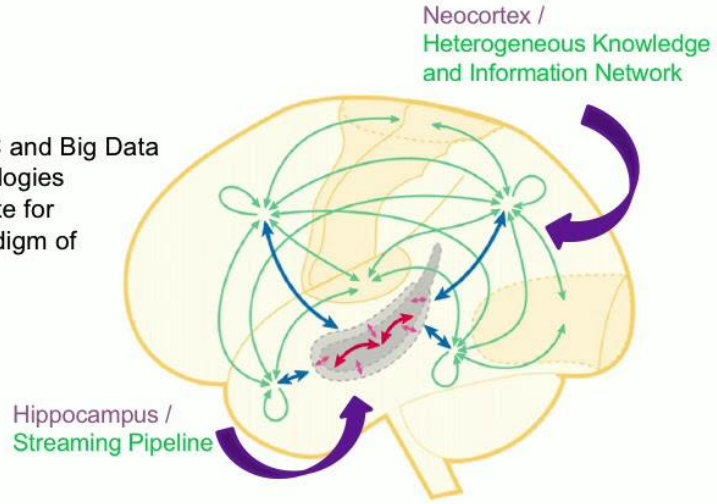
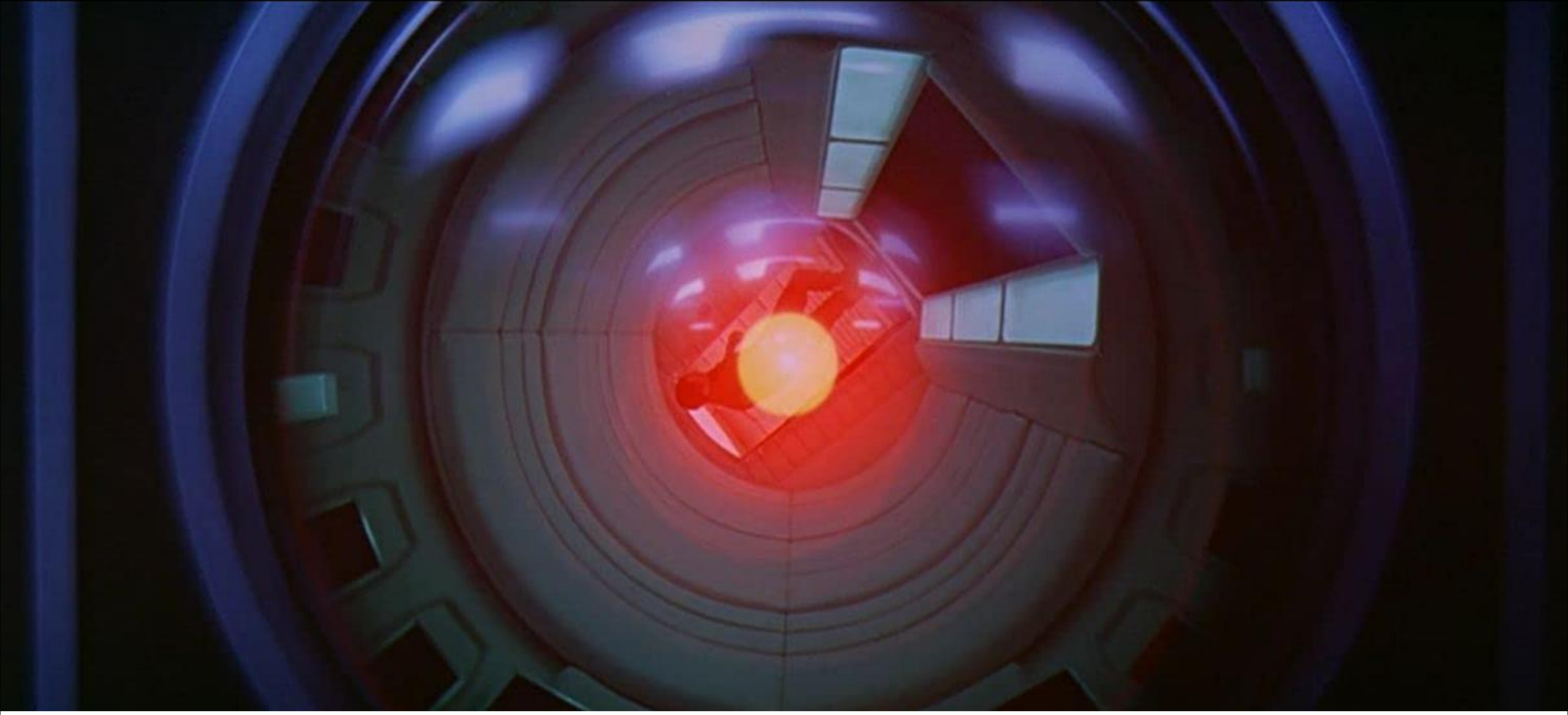


Figure 2: Complementary Learning Systems*

*Dharshan Kumaran, Demis Hassabis, and James L. McClelland, What Learning Systems do Intelligent Agents Need? Complementary Learning Systems, Trends in Cognitive Sciences, 2016

Saiba mais: https://www.microsoft.com/en-us/research/blog/ai4science-to-empower-the-fifth-paradigm-of-scientific-discovery/?OCID=msr_blog_AI4Science_hero



HAL : I am putting myself to the fullest possible use, which is all I think that any conscious entity can ever hope to do.

2001: Uma Odisséia no Espaço (1968)

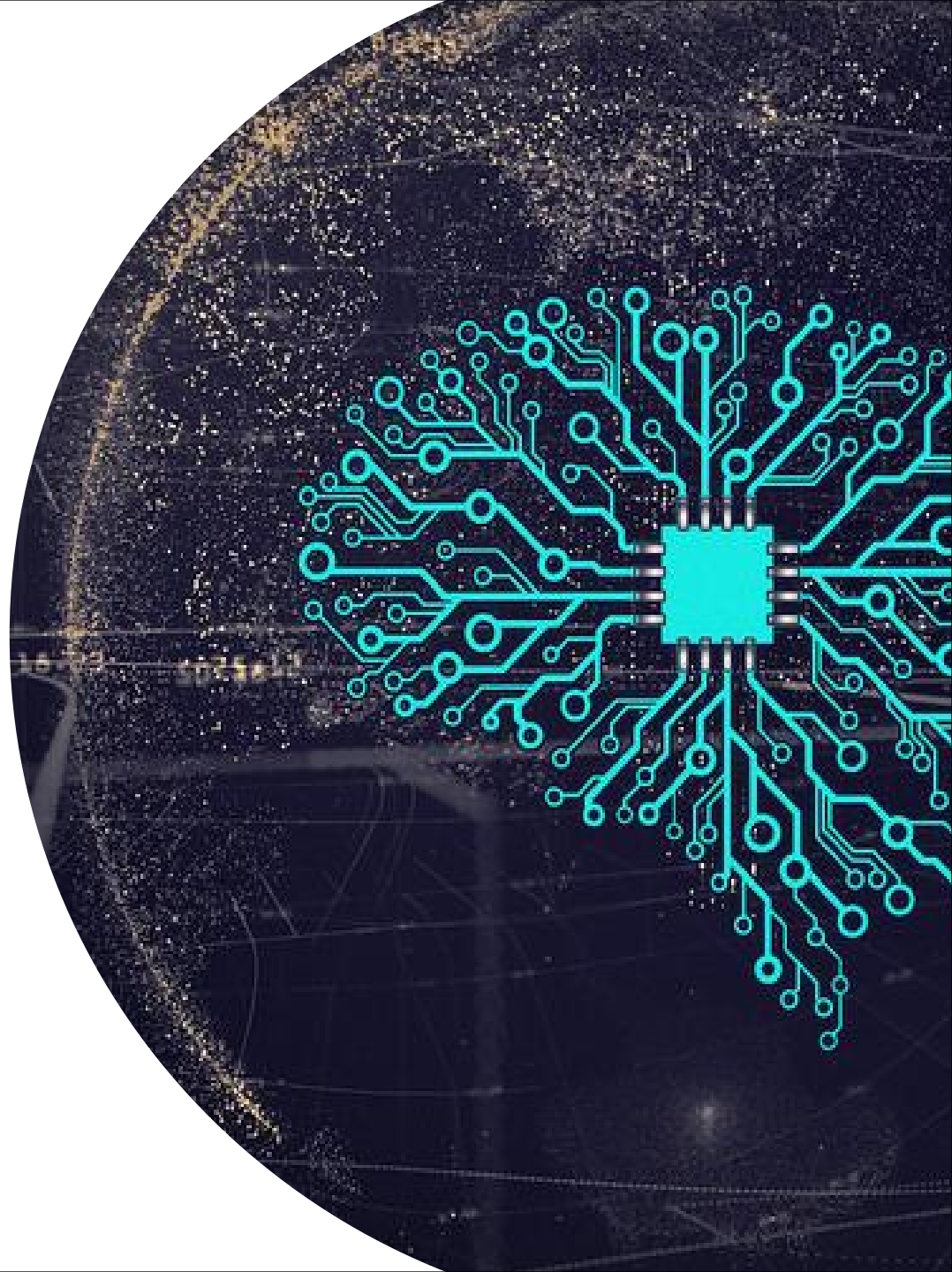
55% das companhias aceleraram seus planos de IA em 2020

86% dos negócios consideram IA uma tecnologia *mainstream*

Metade dos especialistas acredita que IAG acontecerá antes de **2060**

Saiba mais: Forbes, 2022 (shorturl.at/dFGT1)

Concluindo...



Mensagens principais

- Dados em geologia são desafiadores (natureza, coleta, armazenamento, distribuição...)
- Avanço tecnológico permite novos meios de coleta e processamento de dados, que podem se aliar ou substituir técnicas consagradas
- Estamos caminhando para uma nova fase: sistemas artificiais gerando conhecimento com pouca ou nenhuma intervenção humana
- Aplicar *Machine Learning* em geologia é desafiador e ainda exige muita intervenção humana (cuidado com os dados, seleção de técnicas, criação de novas abordagens...)
- Tecnologias que se desenvolvem em outras áreas podem ser aplicadas em geologia

MUITO TRABALHO !

Let's see who you really are machine learning

Para saber mais

- Bergen, K.J., Johnson, P.A., De Hoop, M. V., and Beroza, G.C., 2019, Machine learning for data-driven discovery in solid Earth geoscience: Science, v. 363, doi:10.1126/science.aau0323.
- Hyder, Z., Siau, K., and Nah, F., 2019, Artificial intelligence, machine learning, and autonomous technologies in mining industry: Journal of Database Management, v. 30, p. 67–79, doi:10.4018/JDM.2019040104.
- Karpatne, A., Ebert-Uphoff, I., Ravela, S., Babaie, H.A., and Kumar, V., 2019, Machine Learning for the Geosciences: Challenges and Opportunities: IEEE Transactions on Knowledge and Data Engineering, v. 31, p. 1544–1554, doi:10.1109/TKDE.2018.2861006.
- Ma, X., 2021, Data Science for Geoscience: Recent Progress and Future Trends from the Perspective of a Data Life Cycle. EarthArXiv, https://doi.org/10.31223/X55S4D .

Imagens

- https://www.industrytap.com/wp-content/uploads/2018/12/42271822770_6d2a1d533f_z.jpg
- https://www.mdpi.com/forests/forests-10-00001/article_deploy/html/images/forests-10-00001-g001.png
- https://spamlab.github.io/img/tapagem/diabo_captura_nuvem_super.jpg
- https://amyingramblog.files.wordpress.com/2016/03/screen-shot-2016-02-29-at-9-59-43-pm.png
- https://www.qgis.org/pt_BR/_images/qfield.jpg
- https://is2-ssl.mzstatic.com/image/thumb/Purple123/v4/98/0e/d8/980ed8c7-f225-da6a-1eb0-3a257a262e20/pr_source.png/643x0w.jpg
- https://img.aeroexpo.online/pt/images_ar/photo-mg/185786-14807687.jpg
- https://aerocorner.com/wp-content/uploads/2020/02/IAI-Heron-TP-static-display-at-ILA-2018.jpg
- https://www.google.com/url?sa=i&url=https%3A%2F%2Fwww.cs.cornell.edu%2F~snavely%2Fbundler%2F&psig=AOvVaw2kEYZuQOHOp30RdIzaGbTU&ust=1600965711341000&source=images&cd=vfe&ved=0CAIQjRxqFwoTCKjWyKDc_-sCFQAAAAAdAAAAABAW
- https://iq.opengenus.org/content/images/2022/05/data-science-related-domains.png
- http://thumbs.web.sapo.io/?W=1200&H=627&crop=center&delay_optim=1&epic=YzYwfA6TLWCQaZMKDTXlcDla3sSo1f1pSuSM9NuWxaonNmwGbWkpLfXYjZCKAIJCzyijcKFxJtFo32Wa9q5QNLLNGEKgfDbcX6IgW1WIW2O/SGQ=
- https://www.ga.gov.au/__data/assets/image/0008/22859/13-7402-1-sml1.jpg
- https://m.media-amazon.com/images/M/MV5BOGUwZTMyN2MtMGNkOC00Y2FiLThhY2UtN2M2Yjg0MmI1MWFmXkEyXkFqcGdeQXVyNDIyNjA2MTk@._V1_SY100_CR59,0,100,100_AL_.jpg
- https://1.cms.s81c.com/sites/default/files/2021-04-15/ICLH_Diagram_Batch_01_03-DeepNeuralNetwork-WHITEBG.png
- https://64.media.tumblr.com/f7d89d41d96e2dd8fe6eb8e9802143c4/tumblr_old5i0qHce1usfud0o1_1280.jpg
- https://i.ytimg.com/vi/8qmYQYwslgo/maxresdefault.jpg
- https://cdn.educba.com/academy/wp-content/uploads/2018/06/comparison-of-brain-computer.png
- https://miro.medium.com/max/828/1*OYalBcMGoyYc5lN0ywSXiQ.png
- https://global-uploads.webflow.com/5ef788f07804fb7d78a4127a/6268e0e4dc89f3d071a4c0b9_How%20neural%20networks%20work.jpeg
- https://media.makeameme.org/created/hard-work-happens-591479.jpg
- 10.13140/RG.2.2.31838.43844