

SERVIÇO GEOLÓGICO DO BRASIL – CPRM

Machine Learning para geração de mapas litológicos preditivos Módulo 1

MSc. Iago Costa

Pesquisador em Geociências - Geofísico

Coordenador Executivo da Diretoria de Geologia e Recursos Minerais - DGM

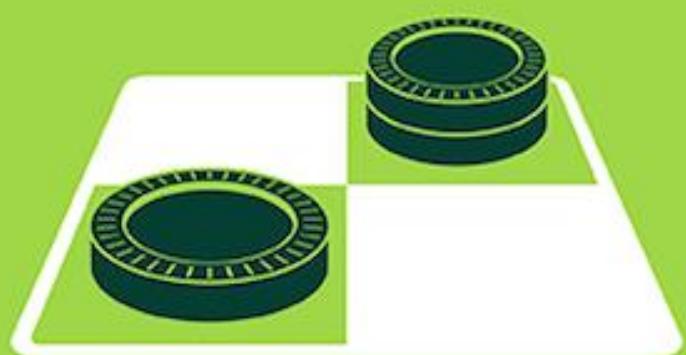
Divisão de Sensoriamento Remoto e Geofísica – DISEGE



SERVIÇO GEOLÓGICO DO BRASIL
CPRM



ARTIFICIAL INTELLIGENCE



1950's

1960's

1970's

1980's

1990's

2000's

2010's

MACHINE LEARNING



DEEP LEARNING



Aplicações do Machine Learning

Filtragem de Spam

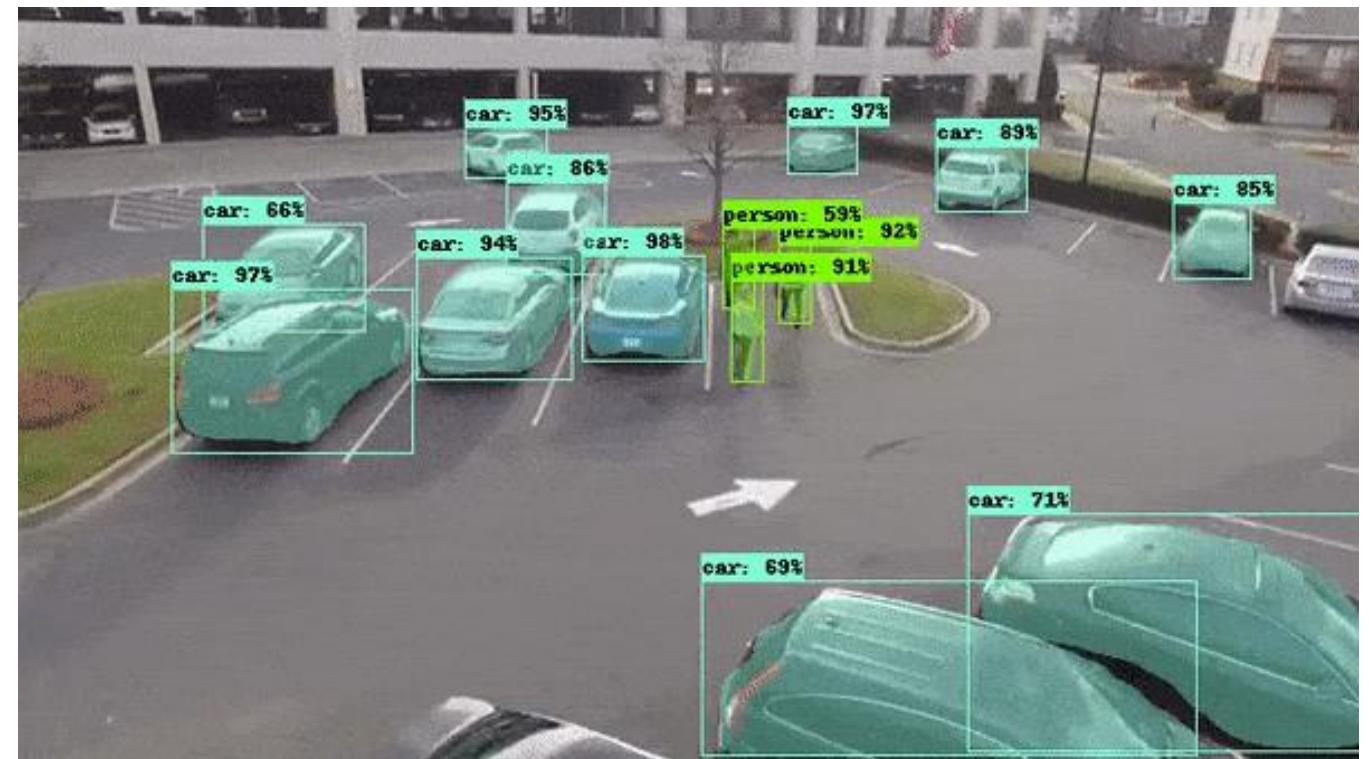
Detecção de Fraudes em Cartões de Crédito

Reconhecimento Facial

Sistemas de Recomendação

Sistemas de Pesquisa

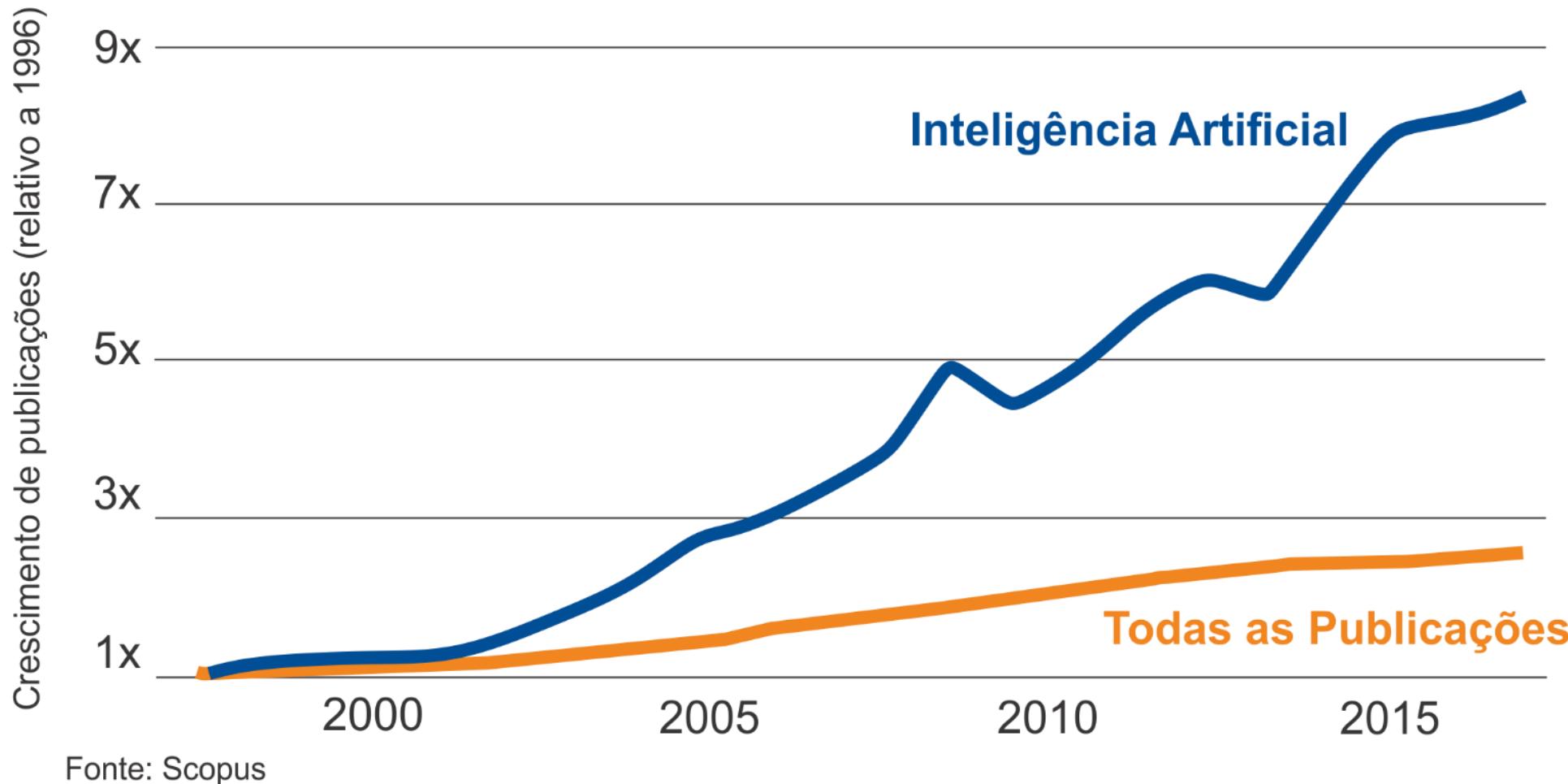
Carros Autônomos



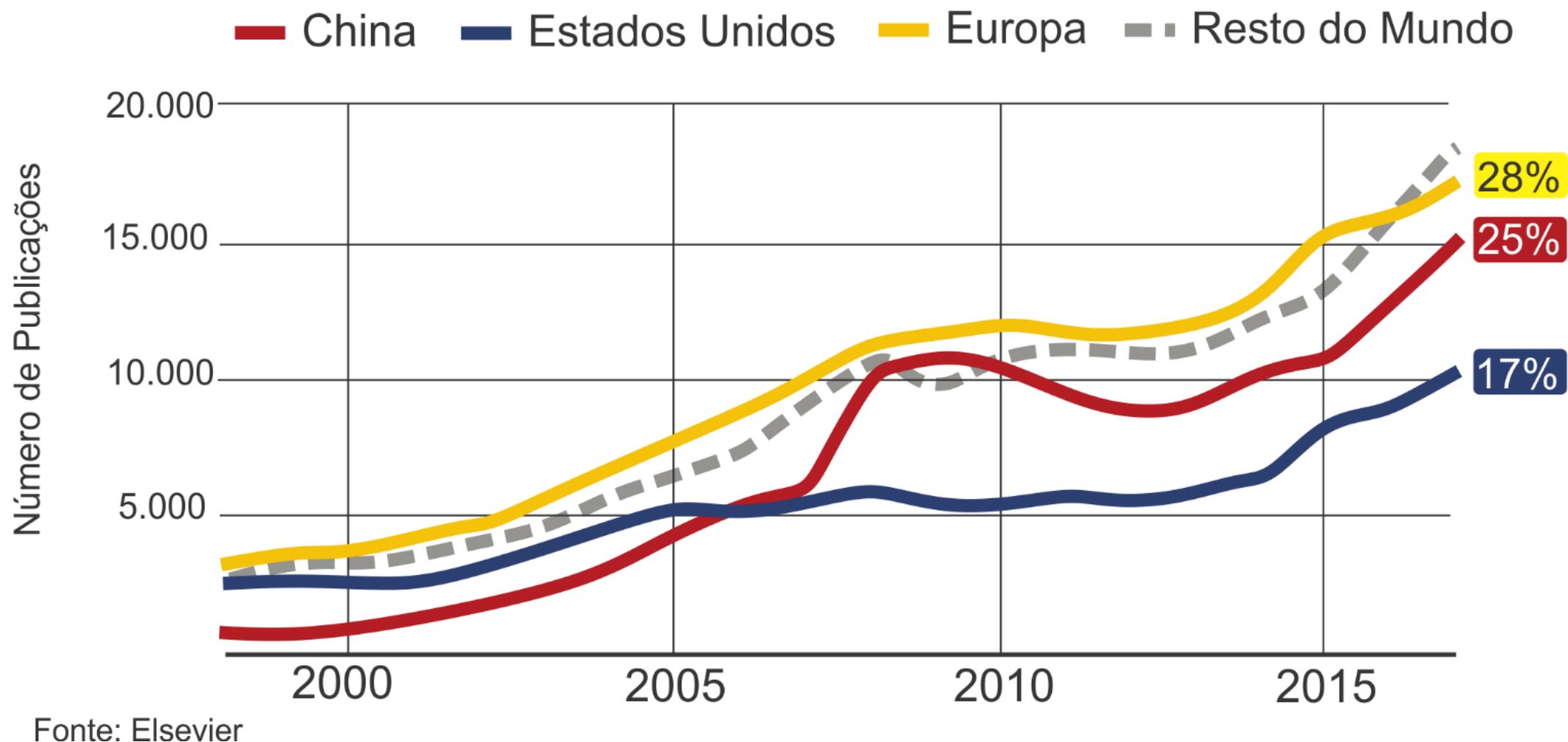
Qual o cenário atual da Inteligência Artificial no mundo?



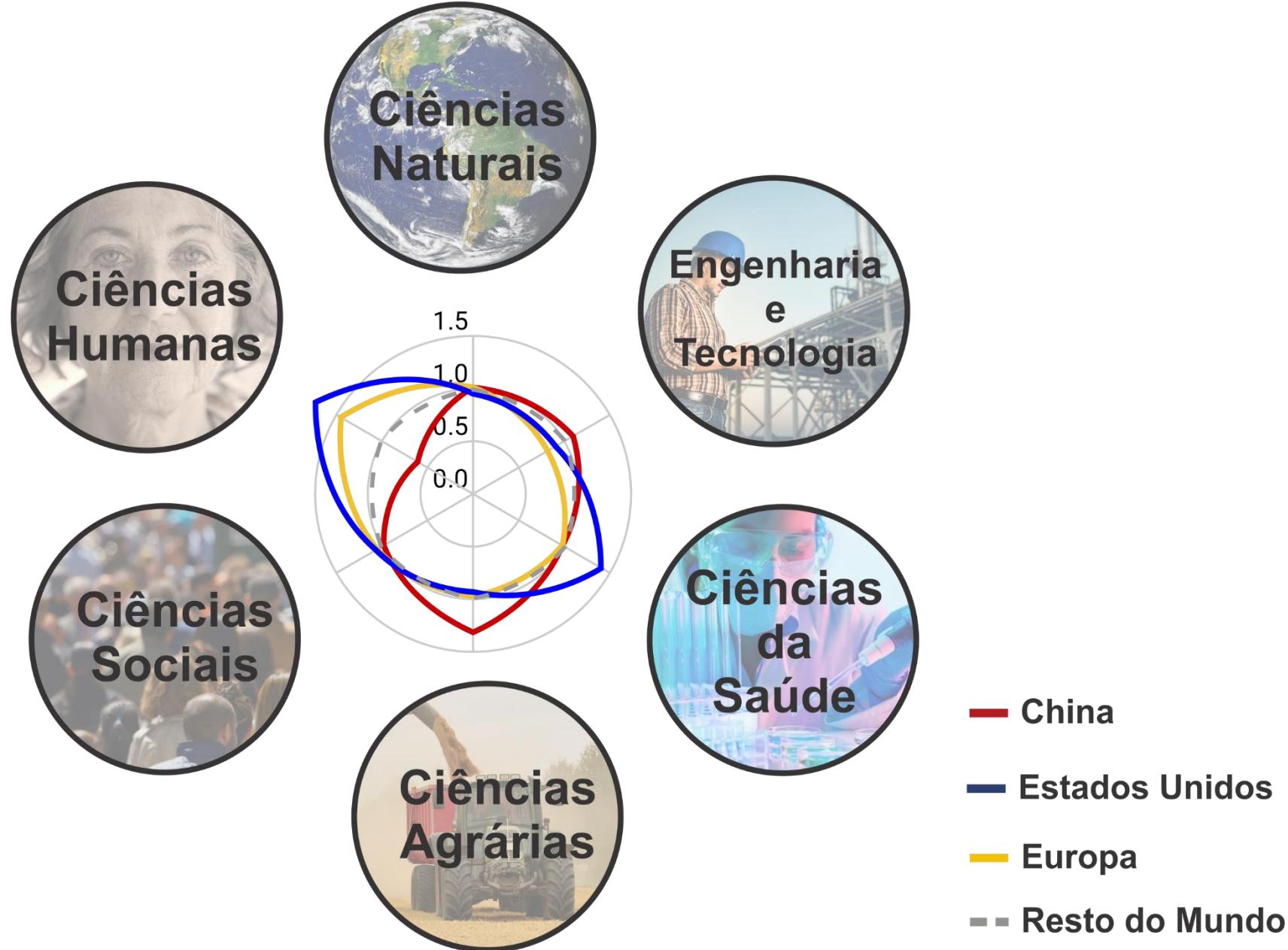
Crescimento anual de publicações (1996 - 2017)



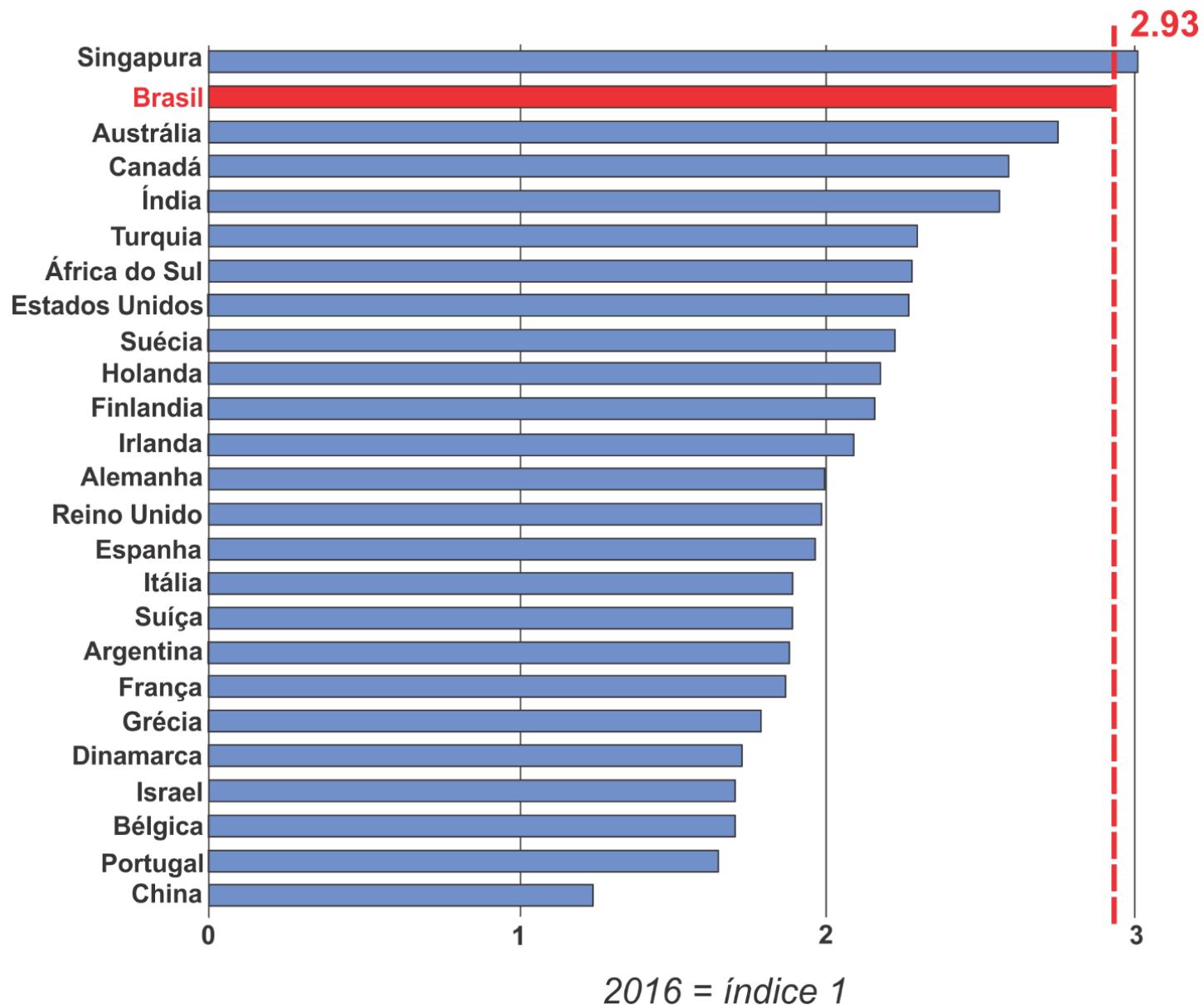
Publicações sobre Inteligência Artificial por Região



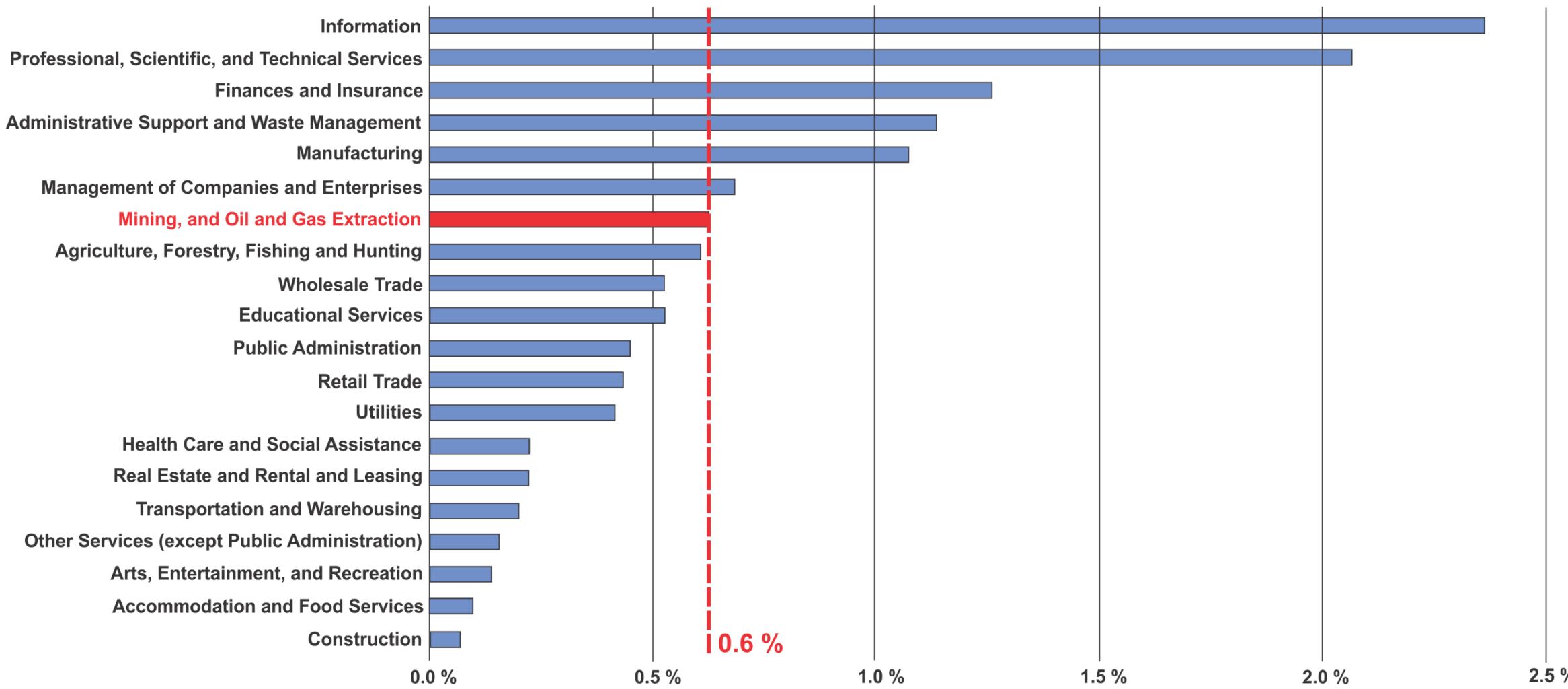
Foco das publicações sobre Inteligência Artificial por Região (2017)



Índice de contratações de Profissionais de IA por País (2019)



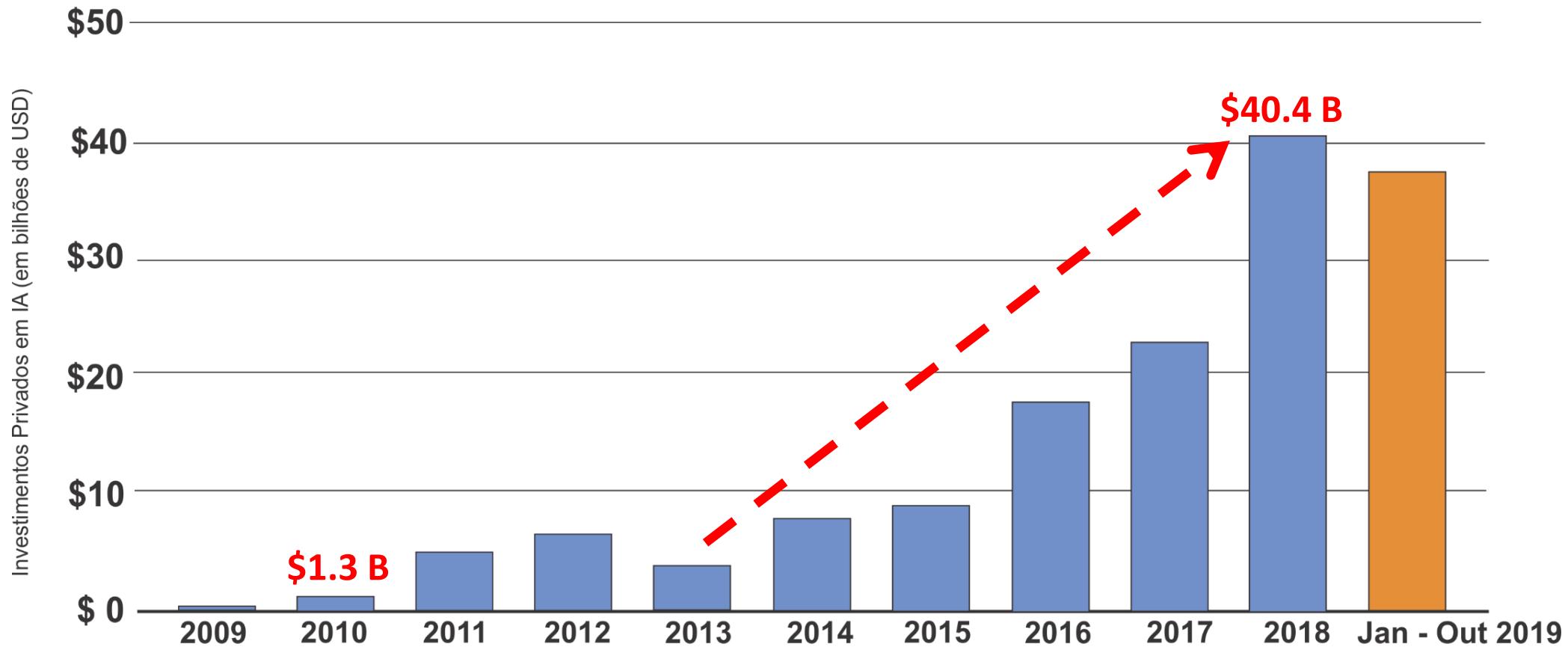
Porcentagem de vagas de trabalhos para IA por setor no US (2019)



Source: BurningGlass

Investimento Privado em IA no Mundo

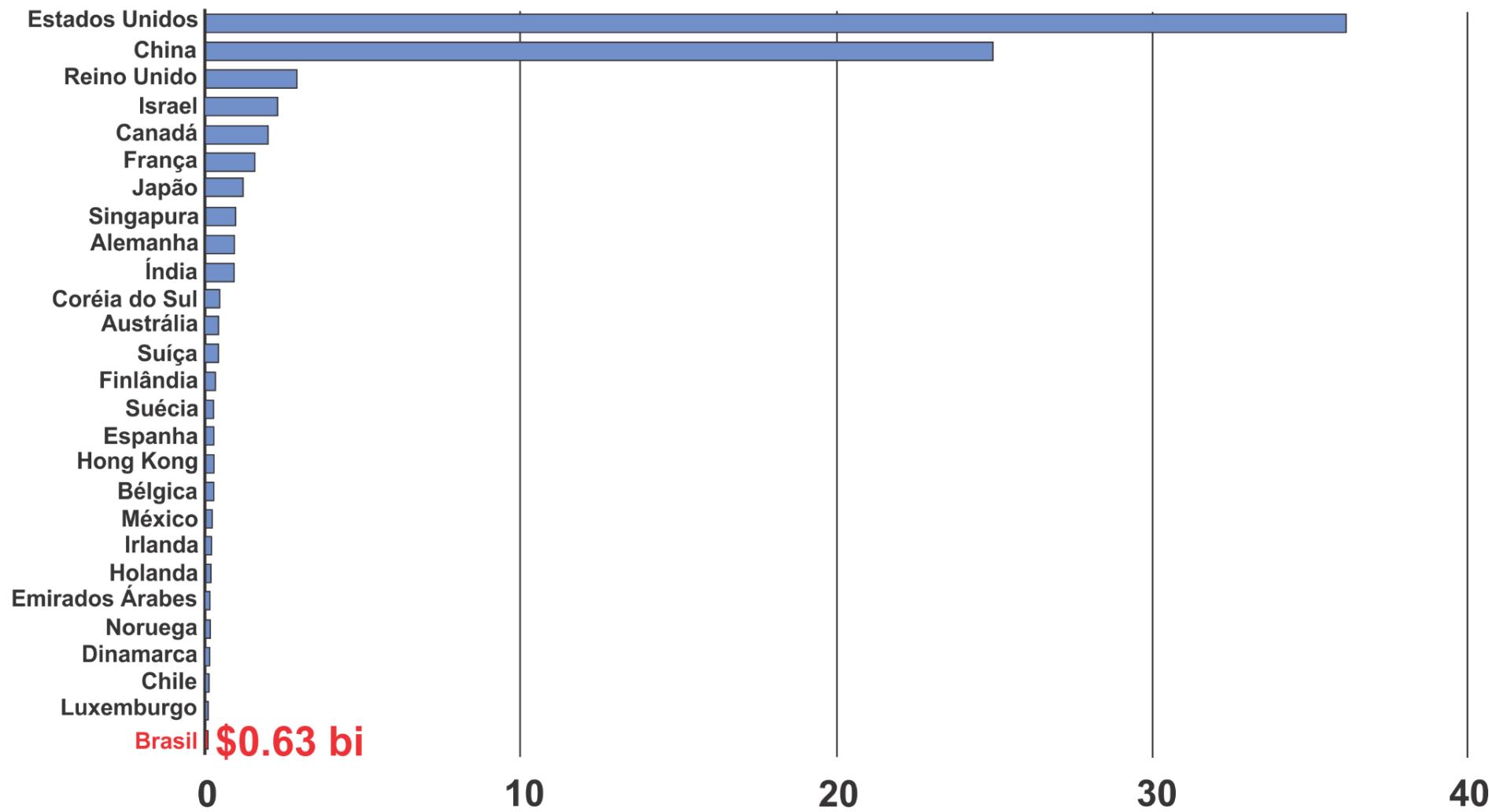
Investimentos Privados em AI (em bilhões de USD)



Fonte: CAPIQ, Crunchbase, Quid, 2019

Investimentos Privados em IA por País

Janeiro 2018 - Outubro 2019



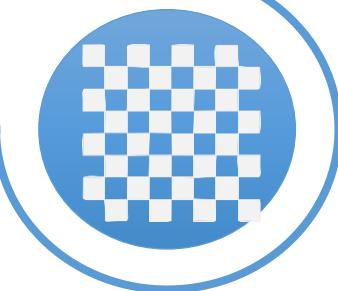
Fonte: CAPIQ, Crunchbase, Quid, 2019

Marcos de desempenho da Inteligência Artificial a Nível Humano

1994

Chinook

O Programa **Chinnok** derrota o Campeão Mundial de Damas *Marion Tinsley*



1997



IBM DeepBlue

O Sistema **IBM's DeepBlue** Derrota o Campeão Mundial de Xadrez Gary Kasparov.

Google DeepMind

Um time do **Google DeepMind** treinou um sistema para aprender como jogar 49 jogos de Atari. O Sistema conseguiu jogar em performance humana a maioria deles



2015

Marcos de desempenho da Inteligência Artificial a Nível Humano

2016

ImageNet

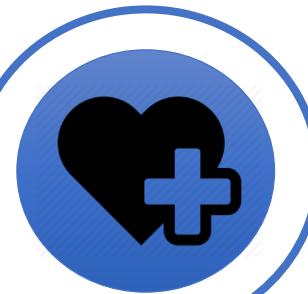
O Erro de classificação automática de imagens do ImageNet caiu de 28% em 2010 para menos de 3% em 2016. Performance humana é ~5%.



2017

Reconhecimento de Fala

Microsoft e IBM alcançam performances de reconhecimento de fala semelhante ao de humanos



Classificação de câncer de pele

Esteva et al. (2017) treinaram um sistema com ~129 mil imagens clínicas de 2.032 tipos de doenças de pele. O Sistema treinado foi capaz de identificar imagens com câncer de Pele com competência comparável a dermatologistas.

Marcos de desempenho da Inteligência Artificial a Nível Humano

2017

Libratus

O Sistema **Libratus**, criado nos Estados Unidos, derrotou os 4 melhores jogadores de um torneio de Poker que teve 120.000 partidas



2018

OpenAI Five

Um time de 5 redes neurais, chamado de **OpenAI Five**, venceu jogos de Dota2 contra jogadores amadores. O **OpenAI Five** aprendeu jogando partidas contra ele mesmo, equivalente a 180 anos de partidas por dia



Tradução Chinês - Inglês

A Microsoft desenvolveu um Sistema de tradução Chinês – Inglês com performance similar a humana

Marcos de desempenho da Inteligência Artificial a Nível Humano

2018

Google

A Google desenvolveu um sistema de detecção de câncer de próstata com acurácia de 70%. A média de precisão alcançada por patologistas gerais certificados pelos EUA era de 61%.



2018

Retinopatia diabética

Pesquisadores desenvolveram um sistema de *Deep Learning* capaz de detectar **Retinopatia Diabética** com acurácia comparada a especialistas

2019



AlphaFold

A organização *DeepMind* desenvolveu o sistema **AlphaFold**, que utilizou dados de sequências geométricas para predizer a estruturas 3D de proteínas

E na Geociências?



SHARE**REVIEW**

Machine learning for data-driven discovery in solid Earth geoscience



[Karianne J. Bergen](#)^{1,2}, [Paul A. Johnson](#)³, [Maarten V. de Hoop](#)⁴, [Gregory C. Beroza](#)^{5,*}



[+ See all authors and affiliations](#)



Science 22 Mar 2019;
Vol. 363, Issue 6433, eaau0323
DOI: 10.1126/science.aau0323

Karianne Bergen



Professora visitante na Universidade de Brown

Pós-doc em *Data Science* em Harvard

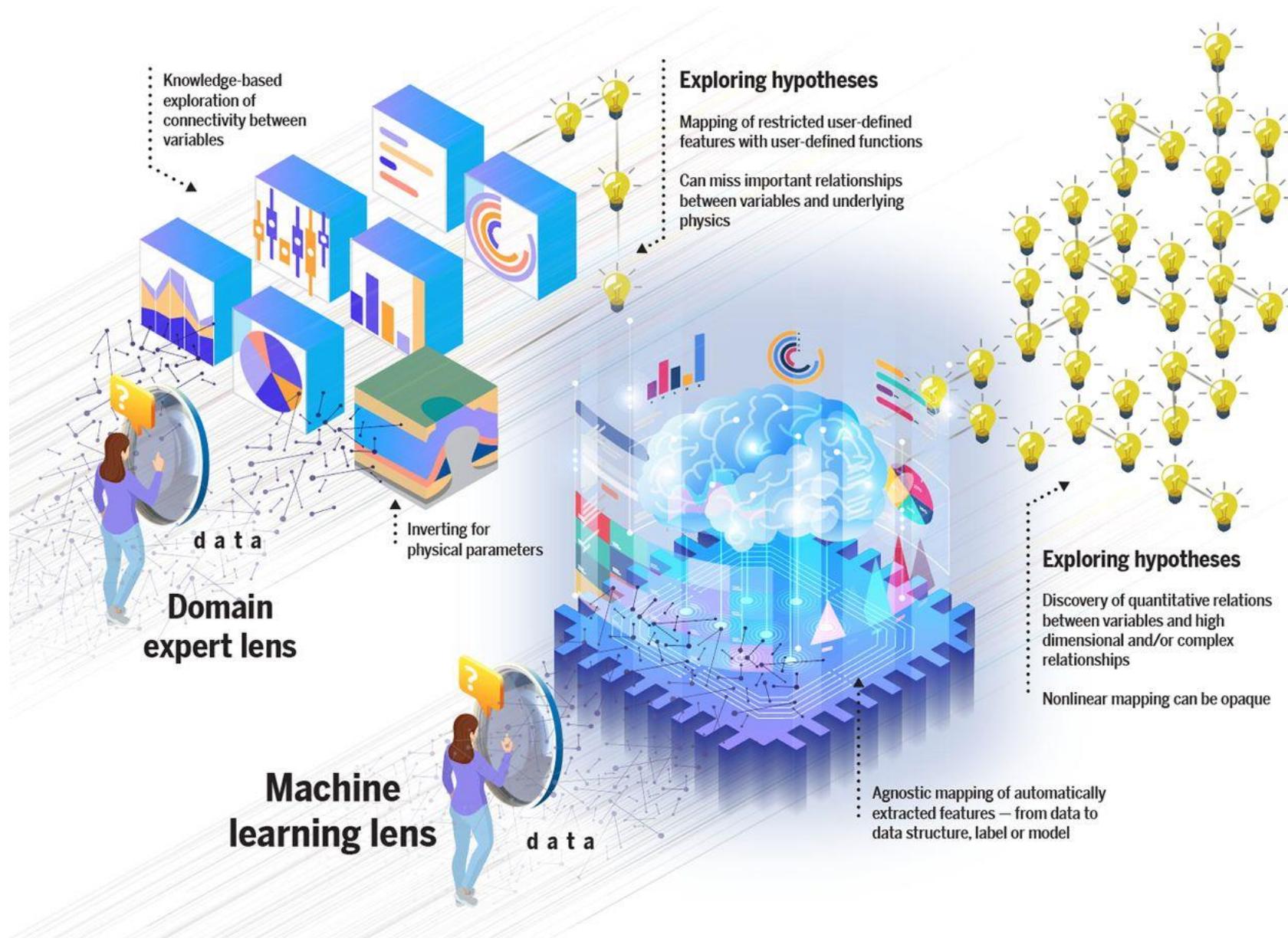
Ph.D em *Engenharia Matemática e Computacional* em Stanford

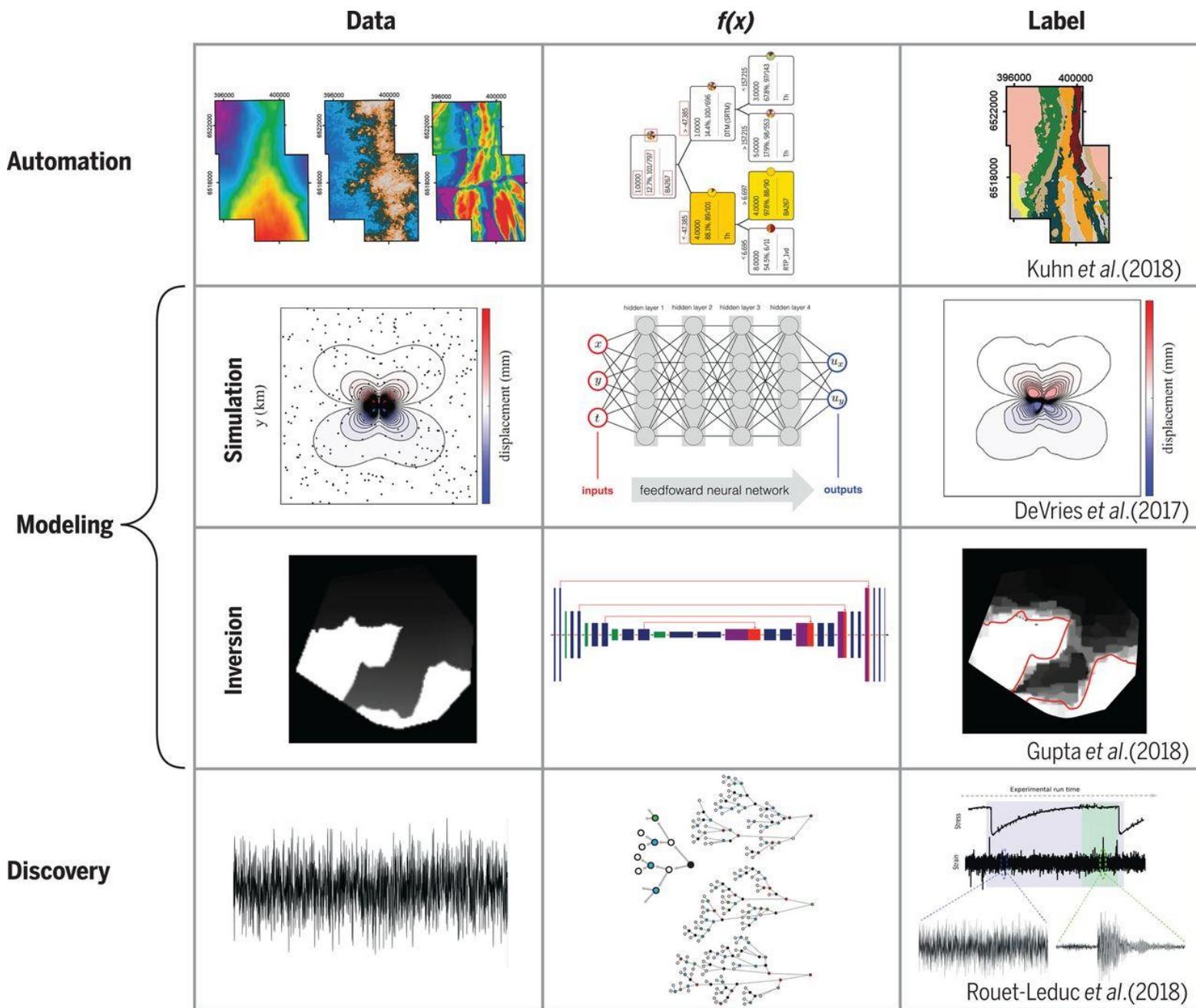


“Na última década, a quantidade de dados disponíveis para os geocientistas cresceu em uma velocidade extraordinária.”

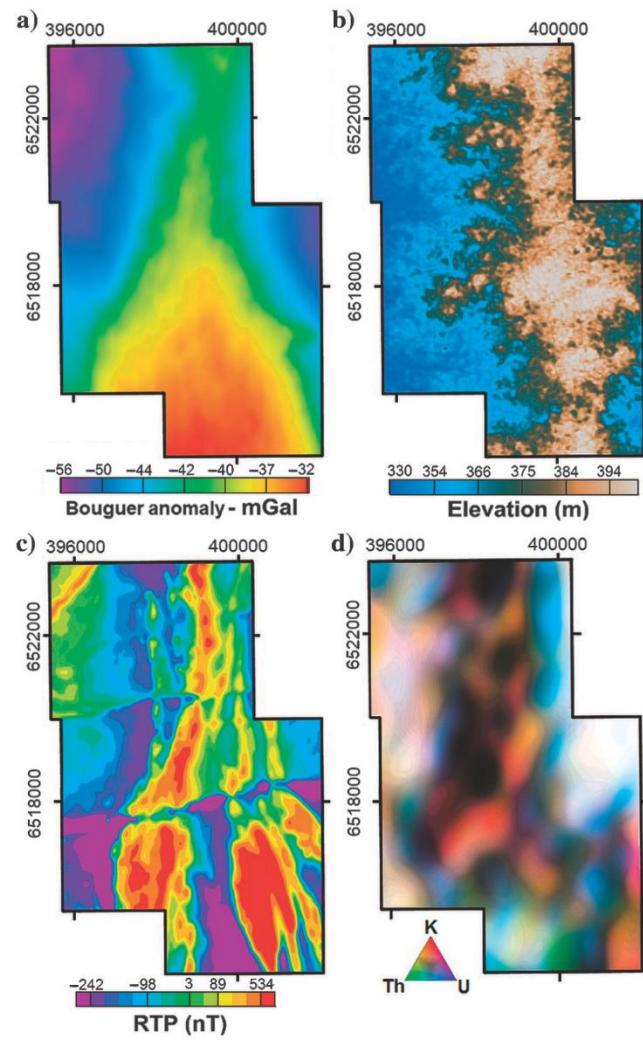
“Da mesma maneira, simulações computacionais que predizem o comportamento da Terra também estão evoluindo rapidamente.”

“O Machine Learning ajudará os geocientistas a desenvolverem novos insights a partir da integração destes dados adquiridos e de novas simulações”

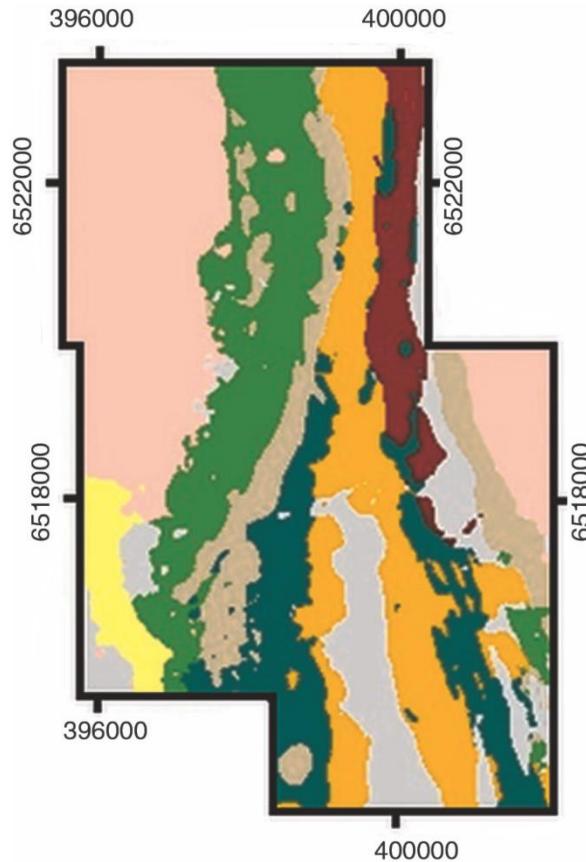




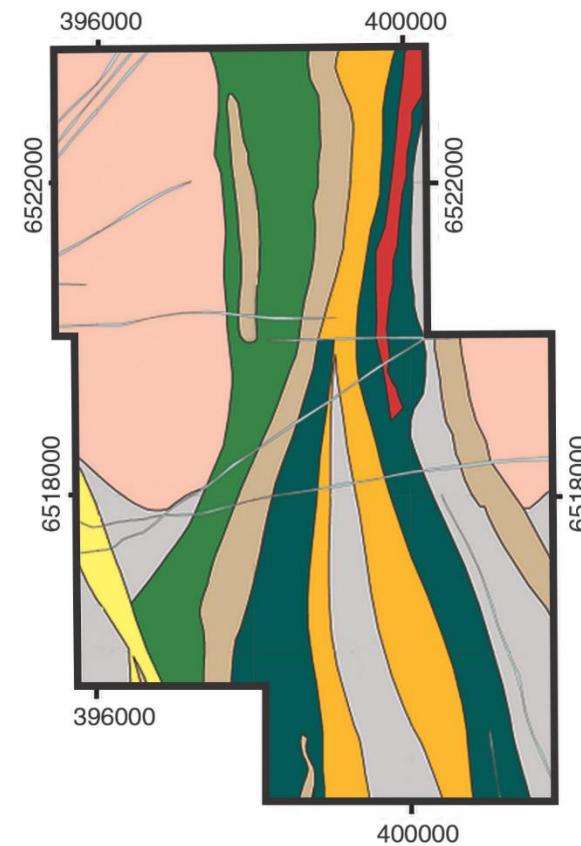
Dados Remotos



Mapa Litológico Preditivo por Machine Learning



Mapa Litológico Atual



Lithological unit

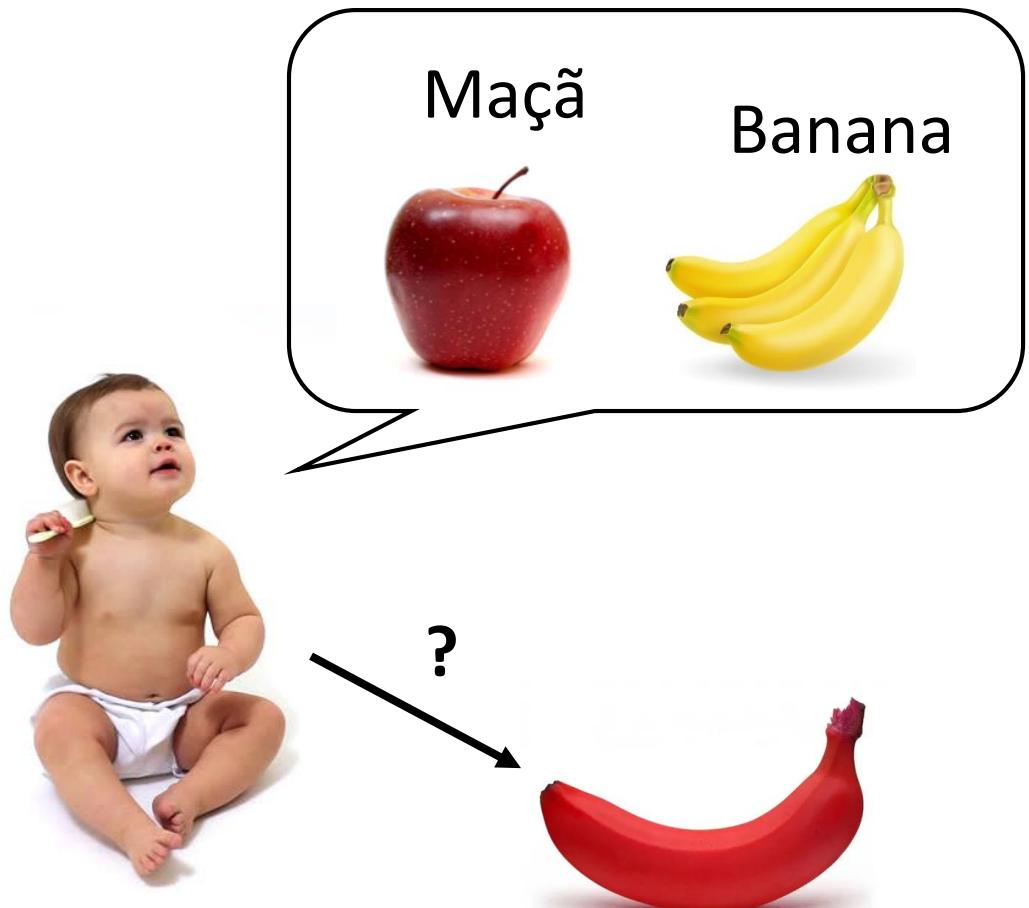
Volcanogenic sediments	High MgO Basalt
Tripod Hill Komatiite	Basalt
	Paringa Basalt
	Dolerite1
	Dolerite2
	Proterozoic dykes

Mas como a *Machine* realmente *Learning*?

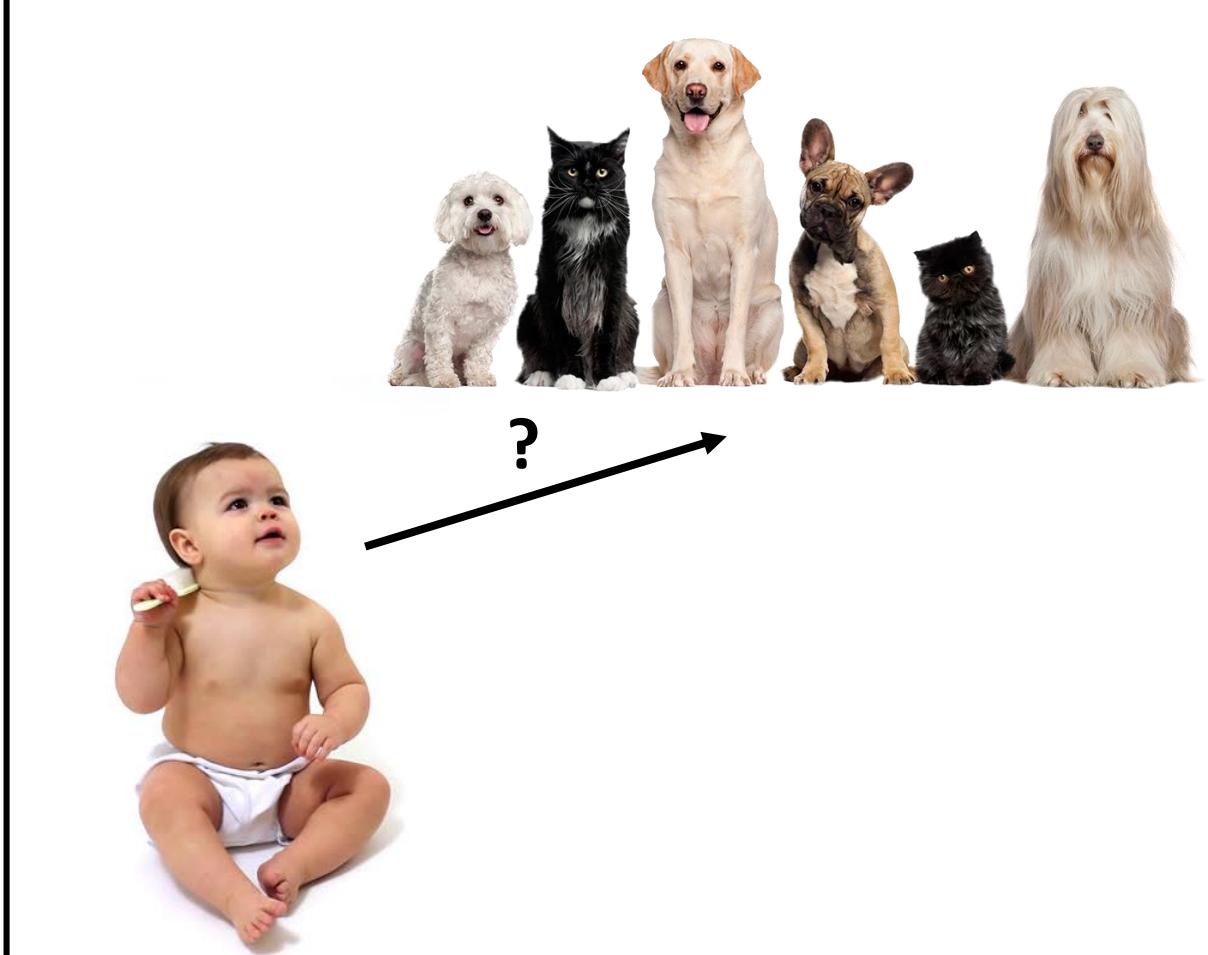


Aprendizagem supervisionada vs não-supervisionada

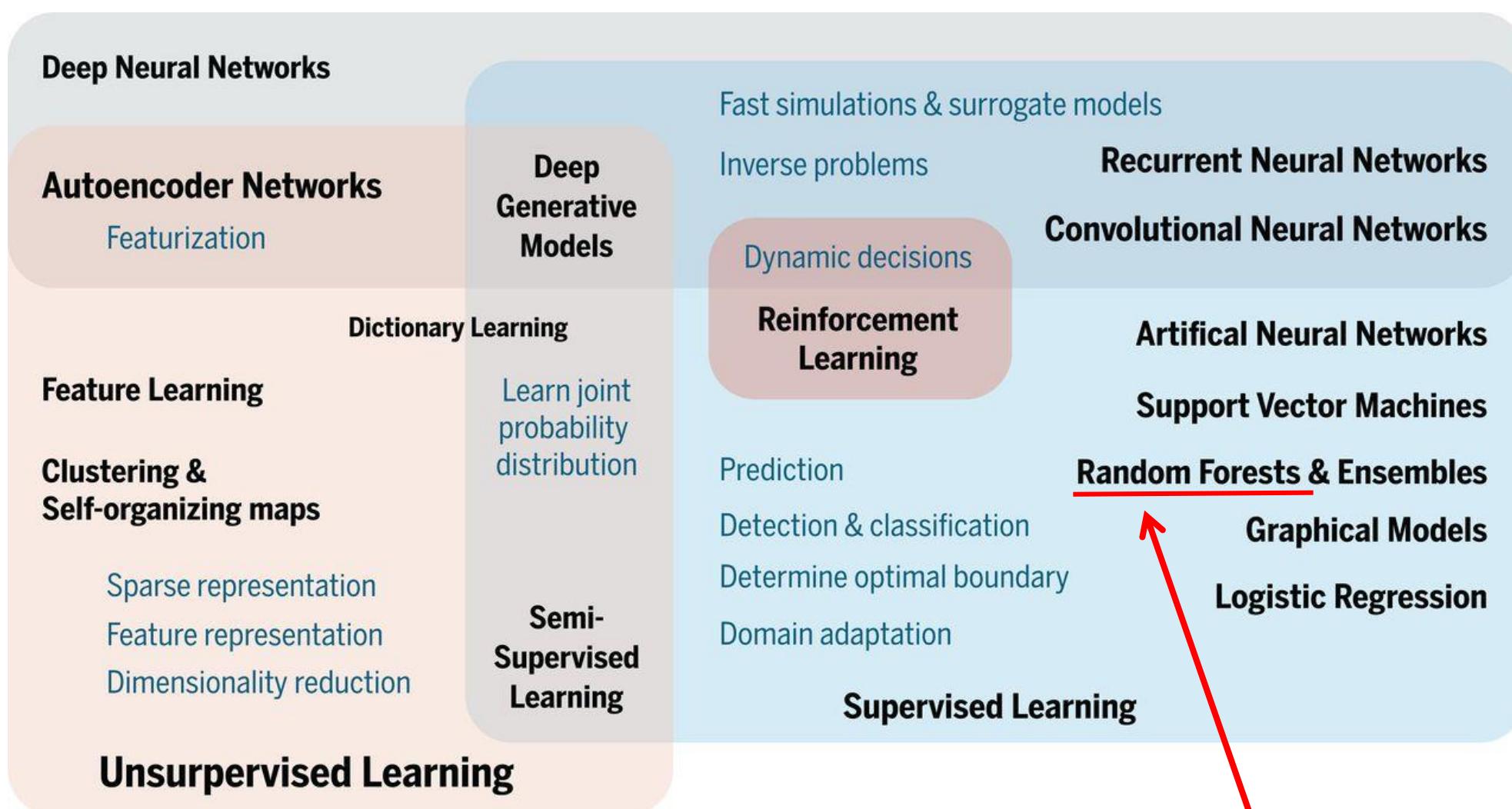
supervisionada



Não-supervisionada



Algoritmos de Machine Learning



Random Forest

- **Random Forest (RF)** é um algoritmo de *machine learning* para classificação e regressão de dados, que combina múltiplas **árvores de decisão** para realizar previsões.
- **RF** é robusto a ruídos na base de dados de treinamento
- **RF** é rápido e efetivo para processar grandes bases de dados

Dados

Em um processo de *Machine Learning*, os dados (D) são frequentemente representados na forma de uma matriz $n \times d$, com n linhas e d colunas

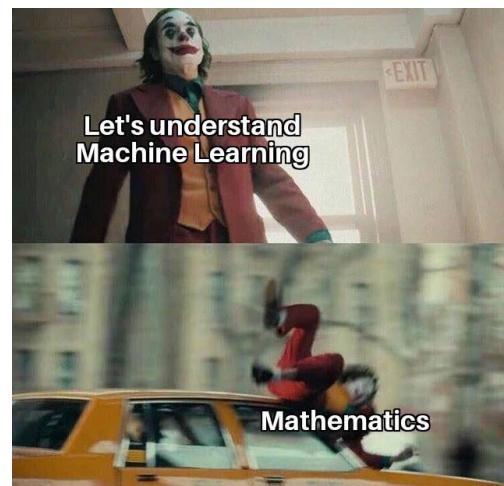
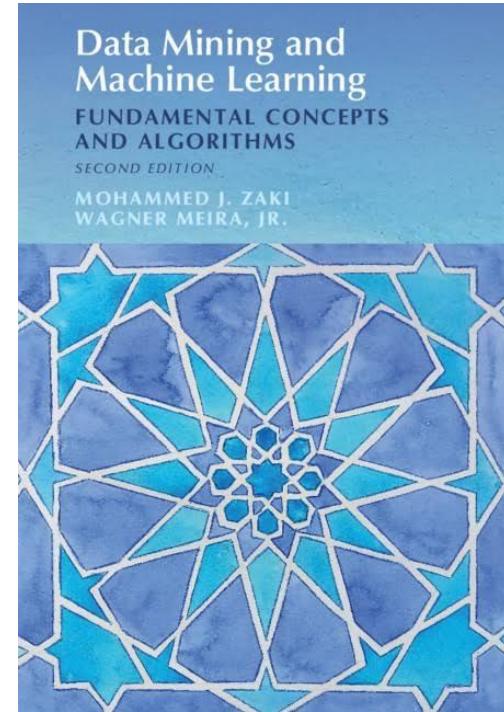
$$\mathbf{D} = \left(\begin{array}{c|cccc} & X_1 & X_2 & \dots & X_d \\ \hline \mathbf{x}_1 & x_{11} & x_{12} & \dots & x_{1d} \\ \mathbf{x}_2 & x_{21} & x_{22} & \dots & x_{2d} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \mathbf{x}_n & x_{n1} & x_{n2} & \dots & x_{nd} \end{array} \right)$$

Onde \mathbf{x}_i denota a i -ésima linha, representada por uma d -tupla na forma:

$$\mathbf{x}_i = (x_{i1}, x_{i2}, \dots, x_{id})$$

e X_j denota a j -ésima coluna, representada por uma n -tupla na forma:

$$X_j = (x_{1j}, x_{2j}, \dots, x_{nj})$$



Exemplo

$$\mathbf{D} = \left(\begin{array}{c|cccc} & X_1 & X_2 & \cdots & X_d \\ \hline \mathbf{x}_1 & x_{11} & x_{12} & \cdots & x_{1d} \\ \mathbf{x}_2 & x_{21} & x_{22} & \cdots & x_{2d} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \mathbf{x}_n & x_{n1} & x_{n2} & \cdots & x_{nd} \end{array} \right)$$

Classe	Nome	Vitamina A (%)	Vitamina C (%)	Cálcio (%)	Ferro (%)	Magnésio (%)	Calorias (Kcal)	Potássio (mg)	Proteína (g)	Fibra (g)
Fruta	Maçã	1.0	7.0	0.0	0.0	1.0	52	107	0.3	2.5
Fruta	Banana	1.0	14.0	0.0	1.0	6.0	89.0	358	1.1	2.6
Vegetal	Beterraba	0.0	8.0	1.0	1.0	5.0	43.0	325	1.6	2.8
Fruta	Abacate	2.0	16.0	1.0	3.0	7.0	160.0	485.0	2.0	7.0
Vegetal	Cenoura	334.0	9.0	3.0	1.0	3.0	41.0	320.0	0.9	2.8

Exemplo

$$\mathbf{D} = \left(\begin{array}{c|cccc} & X_1 & X_2 & \cdots & X_d \\ \mathbf{x}_1 & x_{11} & x_{12} & \cdots & x_{1d} \\ \mathbf{x}_2 & x_{21} & x_{22} & \cdots & x_{2d} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \mathbf{x}_n & x_{n1} & x_{n2} & \cdots & x_{nd} \end{array} \right)$$

Classe	Nome	Vitamina A (%)	Vitamina C (%)	Cálcio (%)	Ferro (%)	Magnésio (%)	Calorias (Kcal)	Potássio (mg)	Proteína (g)	Fibra (g)
Fruta	Maçã	1.0	7.0	0.0	0.0	1.0	52	107	0.3	2.5
Fruta	Banana	1.0	14.0	0.0	1.0	6.0	89.0	358	1.1	2.6
Vegetal	Beterraba	0.0	8.0	1.0	1.0	5.0	43.0	325	1.6	2.8
Fruta	Abacate	2.0	16.0	1.0	3.0	7.0	160.0	485.0	2.0	7.0
Vegetal	Cenoura	334.0	9.0	3.0	1.0	3.0	41.0	320.0	0.9	2.8

Instance

Records
Objects
Points

Feature-vectors
Tuples

Exemplo

Feature 
*Attribute
Proprietie
Dimension
Variable
Field*

$$\mathbf{D} = \left(\begin{array}{c|cccc} & X_1 & X_2 & \cdots & X_d \\ \mathbf{x}_1 & x_{11} & x_{12} & \cdots & x_{1d} \\ \mathbf{x}_2 & x_{21} & x_{22} & \cdots & x_{2d} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \mathbf{x}_n & x_{n1} & x_{n2} & \cdots & x_{nd} \end{array} \right)$$

Classe	Nome	Vitamina A (%)	Vitamina C (%)	Cálcio (%)	Ferro (%)	Magnésio (%)	Calorias (Kcal)	Potássio (mg)	Proteína (g)	Fibra (g)
Fruta	Maçã	1.0	7.0	0.0	0.0	1.0	52	107	0.3	2.5
Fruta	Banana	1.0	14.0	0.0	1.0	6.0	89.0	358	1.1	2.6
Vegetal	Beterraba	0.0	8.0	1.0	1.0	5.0	43.0	325	1.6	2.8
Fruta	Abacate	2.0	16.0	1.0	3.0	7.0	160.0	485.0	2.0	7.0
Vegetal	Cenoura	334.0	9.0	3.0	1.0	3.0	41.0	320.0	0.9	2.8

Instance

*Records
Objects
Points*

*Feature-vectors
Tuples*

Exemplo

$$\mathbf{D} = \left(\begin{array}{c|cccc} & X_1 & X_2 & \cdots & X_d \\ \mathbf{x}_1 & x_{11} & x_{12} & \cdots & x_{1d} \\ \mathbf{x}_2 & x_{21} & x_{22} & \cdots & x_{2d} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \mathbf{x}_n & x_{n1} & x_{n2} & \cdots & x_{nd} \end{array} \right)$$

Features

Classe	Nome	Vitamina A (%)	Vitamina C (%)	Cálcio (%)	Ferro (%)	Magnésio (%)	Calorias (Kcal)	Potássio (mg)	Proteína (g)	Fibra (g)
Instance	Maçã	1.0	7.0	0.0	0.0	1.0	52	107	0.3	2.5
	Banana	1.0	14.0	0.0	1.0	6.0	89.0	358	1.1	2.6
	Beterraba	0.0	8.0	1.0	1.0	5.0	43.0	325	1.6	2.8
	Abacate	2.0	16.0	1.0	3.0	7.0	160.0	485.0	2.0	7.0
	Cenoura	334.0	9.0	3.0	1.0	3.0	41.0	320.0	0.9	2.8

Fruta ou Vegetal?

Exemplo

$$\mathbf{D} = \left(\begin{array}{c|cccc} & X_1 & X_2 & \cdots & X_d \\ \mathbf{x}_1 & x_{11} & x_{12} & \cdots & x_{1d} \\ \mathbf{x}_2 & x_{21} & x_{22} & \cdots & x_{2d} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \mathbf{x}_n & x_{n1} & x_{n2} & \cdots & x_{nd} \end{array} \right)$$

Classe
ou
Target

Features

Instance

Classe	Nome	Vitamina A (%)	Vitamina C (%)	Cálcio (%)	Ferro (%)	Magnésio (%)	Calorias (Kcal)	Potássio (mg)	Proteína (g)	Fibra (g)
Fruta	Maçã	1.0	7.0	0.0	0.0	1.0	52	107	0.3	2.5
Fruta	Banana	1.0	14.0	0.0	1.0	6.0	89.0	358	1.1	2.6
Vegetal	Beterraba	0.0	8.0	1.0	1.0	5.0	43.0	325	1.6	2.8
Fruta	Abacate	2.0	16.0	1.0	3.0	7.0	160.0	485.0	2.0	7.0
Vegetal	Cenoura	334.0	9.0	3.0	1.0	3.0	41.0	320.0	0.9	2.8

Fruta ou Vegetal?

Exemplo

$$\mathbf{D} = \left(\begin{array}{c|cccc} & X_1 & X_2 & \cdots & X_d \\ \mathbf{x}_1 & x_{11} & x_{12} & \cdots & x_{1d} \\ \mathbf{x}_2 & x_{21} & x_{22} & \cdots & x_{2d} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \mathbf{x}_n & x_{n1} & x_{n2} & \cdots & x_{nd} \end{array} \right)$$

Classe
ou
Target

Features

Instance

Classe	Nome	Vitamina A (%)	Vitamina C (%)	Cálcio (%)	Ferro (%)	Magnésio (%)	Calorias (Kcal)	Potássio (mg)	Proteína (g)	Fibra (g)
Fruta	Maçã	1.0	7.0	0.0	0.0	1.0	52	107	0.3	2.5
Fruta	Banana	1.0	14.0	0.0	1.0	6.0	89.0	358	1.1	2.6
Vegetal	Beterraba	0.0	8.0	1.0	1.0	5.0	43.0	325	1.6	2.8
Fruta	Abacate	2.0	16.0	1.0	3.0	7.0	160.0	485.0	2.0	7.0
Vegetal	Cenoura	334.0	9.0	3.0	1.0	3.0	41.0	320.0	0.9	2.8

Metadata

Fruta ou Vegetal?

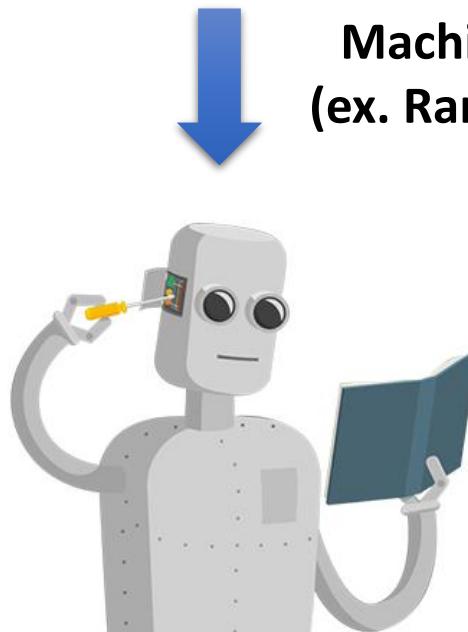
**Classe
ou
Target**

Features

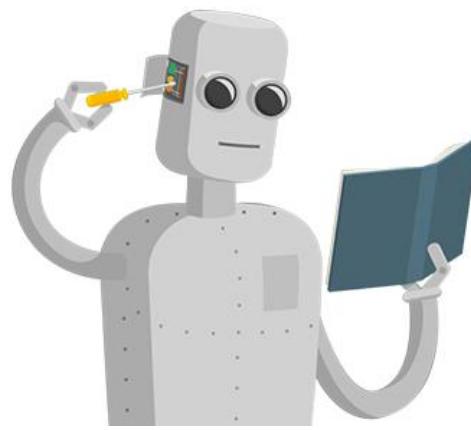
Classe	Nome	Vitamina A (%)	Vitamina C (%)	Cálcio (%)	Ferro (%)	Magnésio (%)	Calorias (Kcal)	Potássio (mg)	Proteína (g)	Fibra (g)
Fruta	Maçã	1.0	7.0	0.0	0.0	1.0	52	107	0.3	2.5
Fruta	Banana	1.0	14.0	0.0	1.0	6.0	89.0	358	1.1	2.6
Vegetal	Beterraba	0.0	8.0	1.0	1.0	5.0	43.0	325	1.6	2.8
Fruta	Abacate	2.0	16.0	1.0	3.0	7.0	160.0	485.0	2.0	7.0
Vegetal	Cenoura	334.0	9.0	3.0	1.0	3.0	41.0	320.0	0.9	2.8

Metadata

**Machine Learning
(ex. Random Forest)**



Fruta ou Vegetal?



Classe
?

Features

Classe	Nome	Vitamina A (%)	Vitamina C (%)	Cálcio (%)	Ferro (%)	Magnésio (%)	Calorias (Kcal)	Potássio (mg)	Proteína (g)	Fibra (g)
?	?	1.0	154.0	3.0	1.0	4.0	61.0	213.0	1.1	3.0
?	?	15.0	300.0	2.0	11.0	3.0	20.0	202.0	2.2	2.1
?	?	0.0	43.0	2.0	3.0	5.0	53.0	151.0	1.1	7.0

Metadata

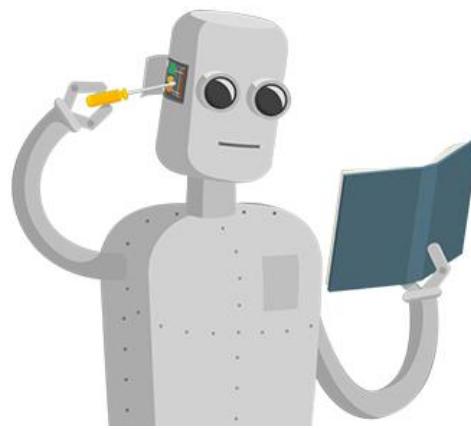
?



1 – 78 % Fruta

2 – 71 % Vegetal

3 - 90 % Fruta



Classe
?

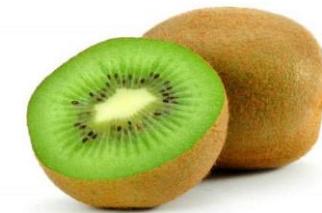
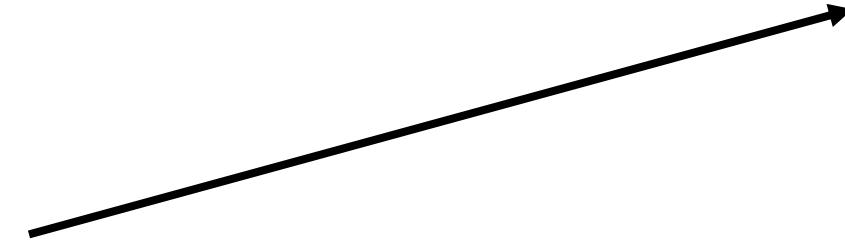
Features

Classe	Nome	Vitamina A (%)	Vitamina C (%)	Cálcio (%)	Ferro (%)	Magnésio (%)	Calorias (Kcal)	Potássio (mg)	Proteína (g)	Fibra (g)
?	?	1.0	154.0	3.0	1.0	4.0	61.0	213.0	1.1	3.0
?	?	15.0	300.0	2.0	11.0	3.0	20.0	202.0	2.2	2.1
?	?	0.0	43.0	2.0	3.0	5.0	53.0	151.0	1.1	7.0

Metadata
?



1 – 78 % Fruta



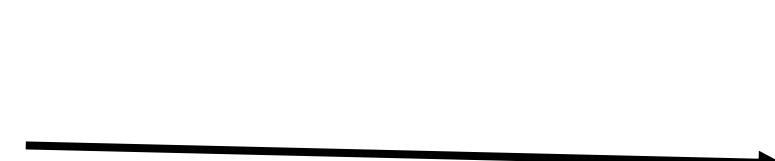
Kiwi

2 – 71 % Vegetal



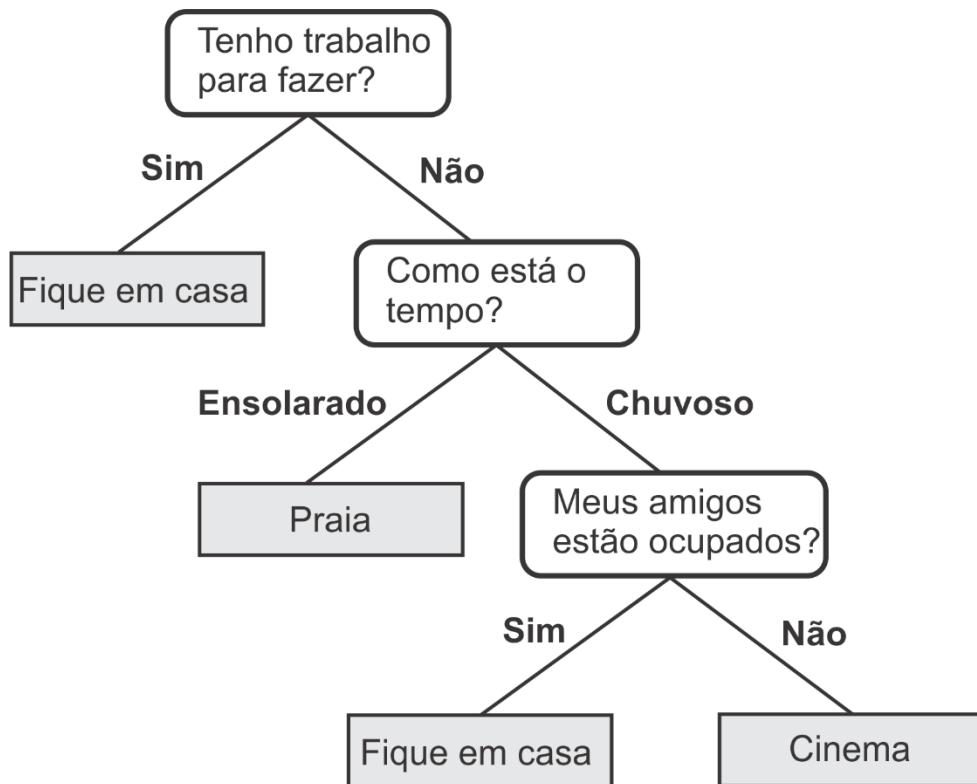
Aspargos

3 - 90 % Fruta



Framboesa

Mas como o Random Forest cria as Árvores de Decisão a partir dos dados?



Random Forest

Features?

Target

??	??	??	??	??	??	Litologia
**	**	**	**	**	**	A4_mu_c
**	**	**	**	**	**	A3xi
**	**	**	**	**	**	A3xi
**	**	**	**	**	**	A4_mu_c

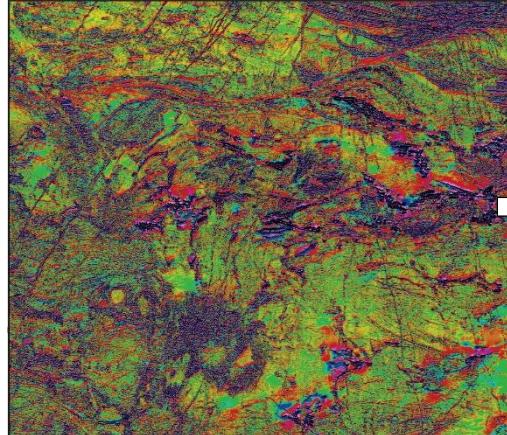
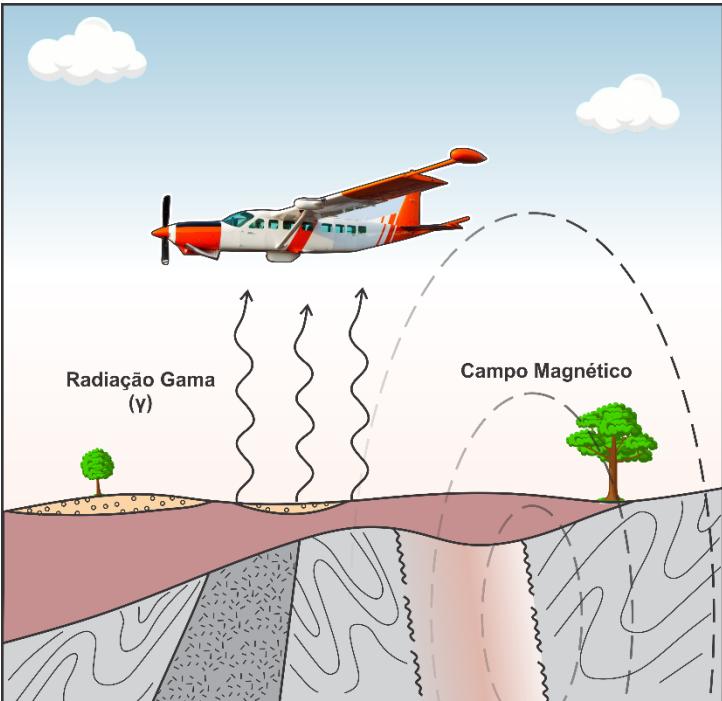
Qual a litologia?

Random Forest

Features?

Target

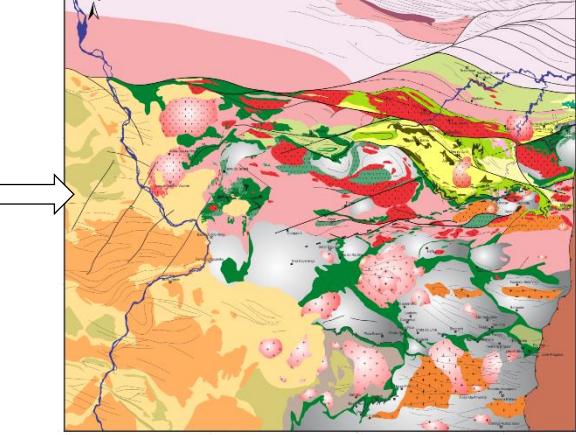
??	??	??	??	??	??	Litologia
**	**	**	**	**	**	A4_mu_c
**	**	**	**	**	**	A3xi
**	**	**	**	**	**	A3xi
**	**	**	**	**	**	A4_mu_c



Dado Magnetométrico



Dado Gamaespectrométrico



Mapa Geológico

Sensoriamento Remoto
Geoquímica

Levantamentos Aerogeofísicos – SGB/CPRM

2000
CPRM
1 CPRM/CODEMIG/ANP
2 Itagimirim - Medeiros Neto (BA)
3 Senhor do Bonfim
4 Ibitiara - Rio de Contas (BA)
5 Parima - Urarcoera

2001
CPRM
1 Área 1 (MG)
2 Área 2 (MG)
3 Área 3 (MG)
4 Área 4 (MG)
5 Área 5 (MG)
6 Área 6 (MG)
7 Riacho Seco - Andorinha (BA)

2003
CPRM
1 Anapu-Tueré

2004
CPRM
1 Anapu-Tueré
2 Trombetas
3 Rio Araguari

Convênio CPRM/Estados
4 Arco Magmático de Arêncópolis:
Sequência Juscelândia
5 Arco Magmático de Mara Rosa

2005
CPRM
1 Amapá
2 Sudeste de Rondônia
Convênio CPRM/ANP
3 Novo Oriente
4 Tocantins
Convênio CPRM/Estados
5 Área 7 (MG)
6 Área 8 (MG)
7 Área 9 (MG)
8 Campo Alegre de Lourdes - Mortugaba(BA)
9 Faixa Brasília Sul(GO)
10 Oeste do Arco Magmático de Mara Rosa(GO)

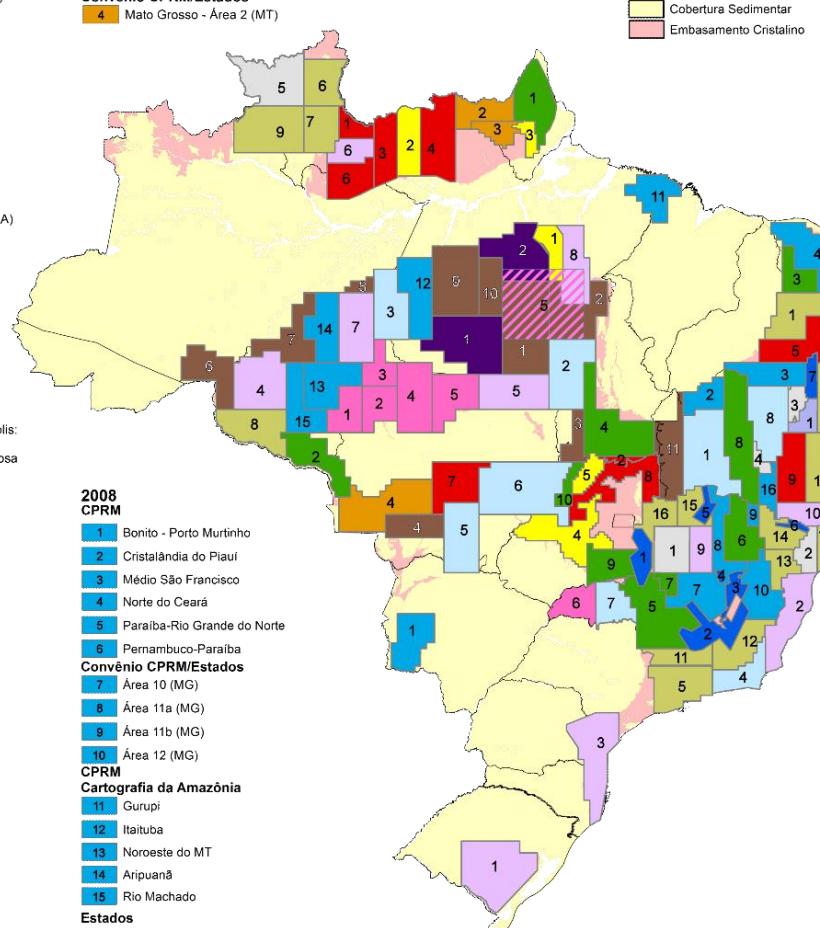
2006
CPRM
1 Anauá
2 Complemento de Tocantins
3 Mapuera
4 Paru do Oeste
5 Pernambuco-Piauí
6 Pitinga

Convênio CPRM/Estados
7 Mato Grosso - Área 1(MT)
Paleo-Neoproterózico do Nordeste de Goiás(GO)

Estados
8 Ruy Barbosa - Vitoria da Conquista (BA)

2007
CPRM
1 Borda Leste do Planalto da Borborema
2 Tumucumaque
3 Complemento da Renca
Convênio CPRM/Estados
4 Mato Grosso - Área 2 (MT)

LEVANTAMENTOS AEROGEOFÍSICOS Projetos em Alta Resolução (Magnetometria e Gamaespectrometria) 2000-2014



2009
CPRM-PAC
1 Escudo do Rio Grande do Sul
2 Espírito Santo
3 Paraná-Santa Catarina
4 Rondônia Central
5 Nordeste do Mato Grosso
CPRM
Cartografia da Amazônia
6 Carará-Jatapu
7 Sucunduri
8 Tucuruí
Convênio CPRM/Estados/PAC
9 Área 13 - Ubai - Pirapora - Joaquim Felício (MG)

Estados
10 Cândido Sales - Mascote (BA)

2011
CPRM PAC
1 Bambuí-Bahia
2 Conceição do Araguaia
3 Província Aurífera dos Tapajós - Fase 2
4 Rio de Janeiro
5 Rondonópolis Dom. Aquino
6 Sudeste do Mato Grosso
Convênio CPRM/Estados/PAC
7 Área 20 Minas Gerais
8 Centro Norte da Bahia

2012
CPRM PAC
1 Japuira
2 Rio Juruena
3 Serra dos Apicás
4 Serra dos Calabás
5 Norte do Mato Grosso
Convênio CPRM/Estados/PAC
6 Área 21 Minas Gerais

2013
CPRM PAC
1 Oeste de Carajás
2 Rio Maria
3 Rio Formoso
4 Cuiabá
5 Carajás (Gravimetria)

CPRM
PAC/Cartografia da Amazônia
6 Rio Madeira-Ituxi
7 Branco-Machadinho
8 Complemento do Sucunduri
9 Rio Curuá
10 Rio Iiri
Convênio CPRM/Estados/PAC
11 Extremo Oeste da Bahia

2014
CPRM PAC
1 Cachoeira do Curuá
CPRM
PAC/Cartografia da Amazônia
2 Rio Bacajá

2004 - 2014

3.726.364 km²

**43,76 % do Território
Brasileiro**

**93% do Embasamento
Cristalino**

Random Forest

Dados de Treinamento

GT (nT/m)	K (%)	eTh (ppm)	eU (ppm)	CT (μ R/h)	Elevação (m)	Litologia
0,219	0,11	5,14	1,62	2,70	306,83	A4_mu_c
0,094	0,69	3,73	2,25	3,60	285,20	A3xi
0,068	0,89	3,30	2,44	3,90	282,77	A3xi
0,18	0,13	5,42	2,20	3,19	287,05	A4_mu_c

Features **Target**



Random Forest

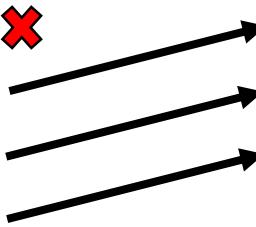
1º Selecionar aleatoriamente 2/3 dos dados (ou instances) colocar em uma “bag”

Dados de Treinamento

GT (nT/m)	K (%)	eTh (ppm)	eU (ppm)	CT (μ R/h)	Elevação (m)	Litologia
0,219	0,11	5,14	1,62	2,70	306,83	A4_mu_c
0,094	0,69	3,73	2,25	3,60	285,20	A3xi
0,068	0,89	3,30	2,44	3,90	282,77	A3xi
0,18	0,13	5,42	2,20	3,19	287,05	A4_mu_c

Bootstrapped Data (~2/3)

GT (nT/m)	K (%)	eTh (ppm)	eU (ppm)	CT (μ R/h)	Elevação (m)	Litologia
0,094	0,69	3,73	2,25	3,60	285,20	A3xi
0,068	0,89	3,30	2,44	3,90	282,77	A3xi
0,18	0,13	5,42	2,20	3,19	287,05	A4_mu_c
**	**	**	**	**	**	**



Random Forest

2º Repetir instâncias aleatórias dentro da “bag” para ficar do mesmo tamanho do dado original

Bootstrapped Data (~2/3)



Bootstrapped Data (~2/3) + (~1/3) replacement

GT (nT/m)	K (%)	eTh (ppm)	eU (ppm)	CT (μ R/h)	Elevação (m)	Litologia
0,094	0,69	3,73	2,25	3,60	285,20	A3xi
0,068	0,89	3,30	2,44	3,90	282,77	A3xi
0,18	0,13	5,42	2,20	3,19	287,05	A4_mu_c
0,18	0,13	5,42	2,20	3,19	287,05	A4_mu_c

Dado repetido (replacement)

Random Forest

3º Selecionar features aleatórias (mtry)

Bootstrapped Data ($\sim 2/3$) + ($\sim 1/3$) replacement

GT (nT/m)	K (%)	eTh (ppm)	eU (ppm)	CT (μ R/h)	Elevação (m)	Litologia
0,094	0,69	3,73	2,25	3,60	285,20	A3xi
0,068	0,89	3,30	2,44	3,90	282,77	A3xi
0,18	0,13	5,42	2,20	3,19	287,05	A4_mu_c
0,18	0,13	5,42	2,20	3,19	287,05	A4_mu_c

$mtry \leq \sqrt{numero\ de\ features}$

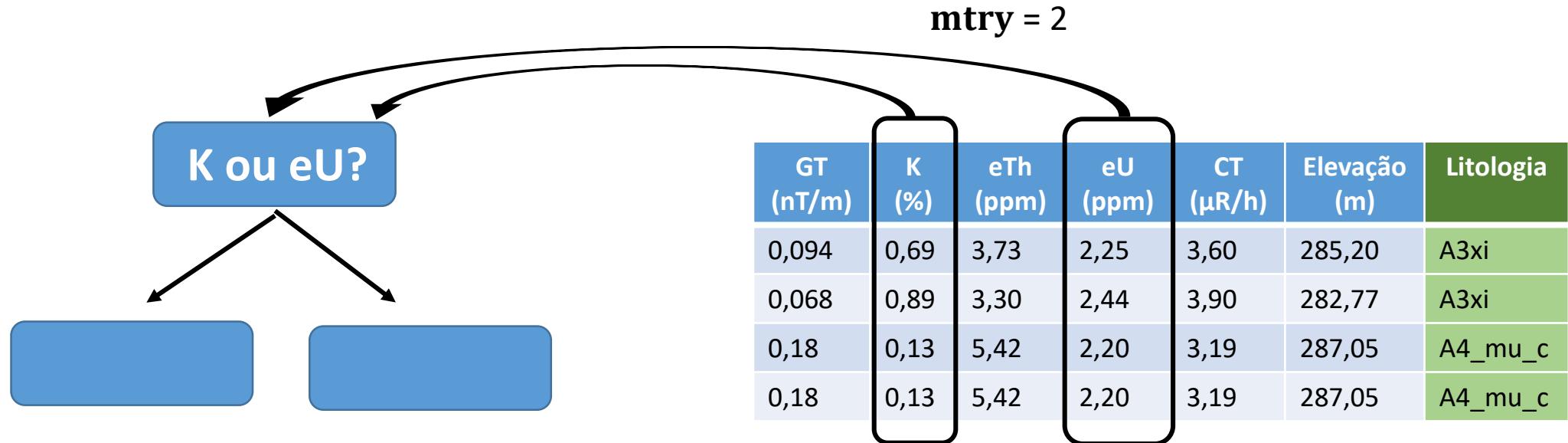
$mtry \leq \sqrt{6}$

$mtry \leq \sim 2.45$

Seleciona aleatoriamente 2 variáveis!

Random Forest

3º Selecionar features aleatórias (mtry)

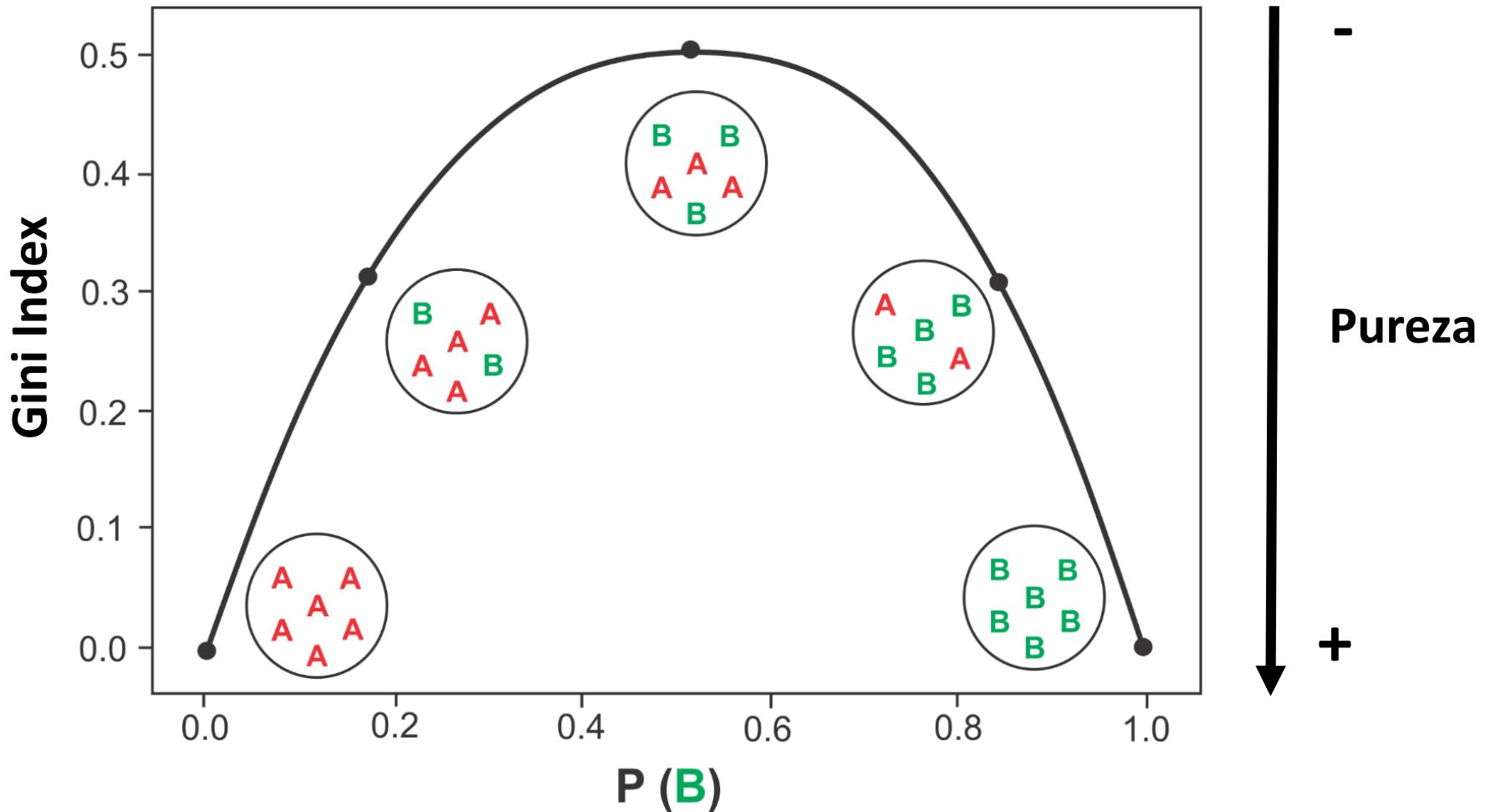


1) Qual *feature* separa melhor os dados? Potássio (K) ou Urânio (eU)?

2) Qual o valor da feature escolhida que melhor separa os dados?

Pureza!

Gini Index



Gini Impurity

Potássio (K)

Litologia	Potássio (K)
A	0
A	0.1
A	0.2
B	0.3
B	0.4
B	0.5
B	0.6
B	0.7

$$Gini = 0.2$$

Urânio (eU)

Litologia	Urânio (eU)
A	0
B	2
A	4
B	5
B	7
B	9
B	11
B	12

$$Gini = 0.3$$



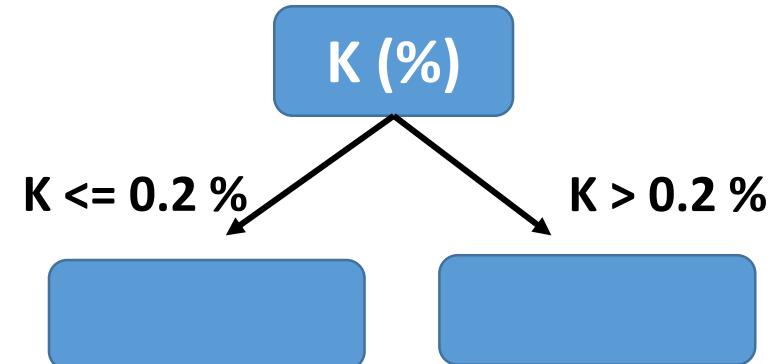
Potássio (K) é a *feature* que melhor separa os dados!
E o melhor valor de separação é 0.2!

Gini Impurity

Potássio (K)

Litologia	Potássio (K)
A	0
A	0.1
A	0.2
B	0.3
B	0.4
B	0.5
B	0.6
B	0.7

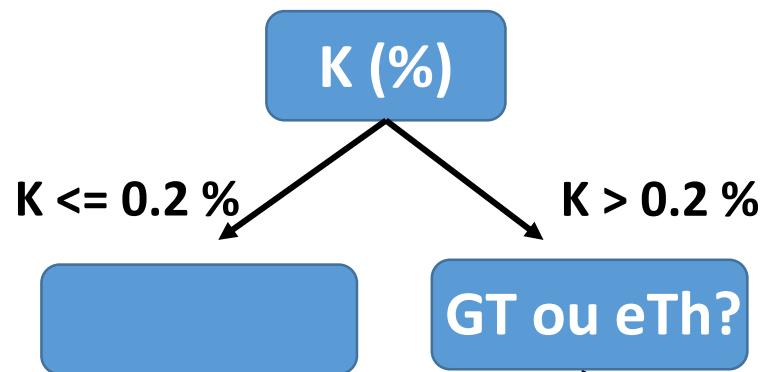
$$Gini = 0.2$$



Potássio (K) é a *feature* que melhor separa os dados!
E o melhor valor de separação é 0.2!

Random Forest

4º Construir a árvore de decisão minimizando a impureza nos nós subsequentes

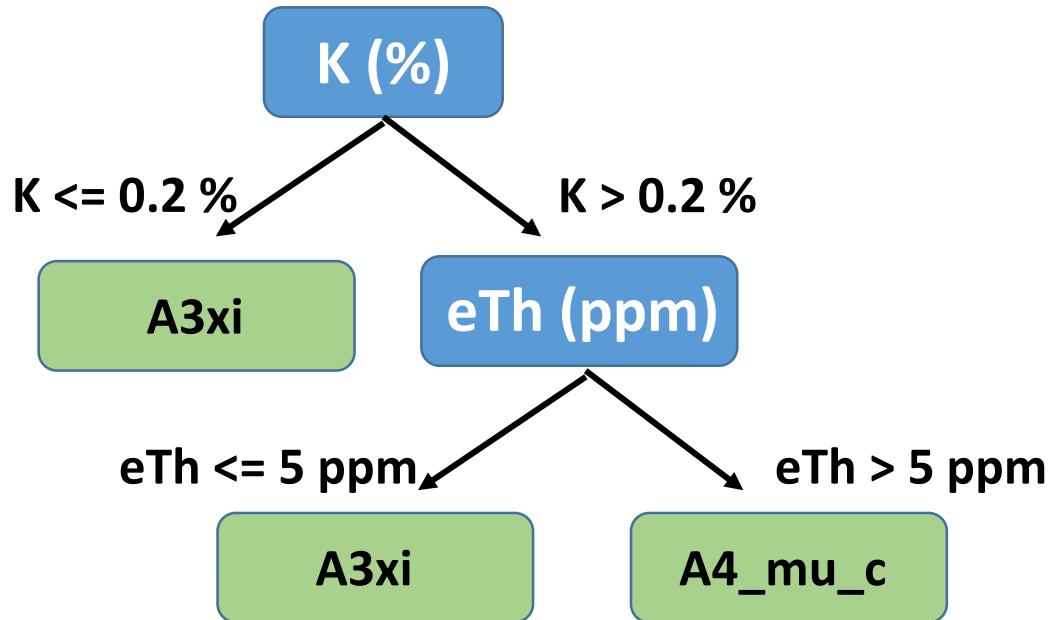


Qual variável
separa melhor?
GT ou eTh?

Conjunto de dados com $K > 0.2 \%$

GT (nT/m)	K (%)	eTh (ppm)	eU (ppm)	CT (μ R/h)	Elevação (m)	Litologia
0,094	0,69	3,73	2,25	3,60	285,20	A3xi
0,068	0,89	3,30	2,44	3,90	282,77	A3xi
0,18	0,23	5,42	2,20	3,19	287,05	A4_mu_c
0,18	0,23	5,42	2,20	3,19	287,05	A4_mu_c

Random Forest



1ª Árvore de decisão!



Random Forest

Repetir esse procedimento para diferentes conjuntos de dados aleatórios (Bootstrapped Data)

1º

Dados de Treinamento

GT (nT/m)	K (%)	eTh (ppm)	eU (ppm)	CT (μ R/h)	Elevação (m)	Litologia
0,219	0,11	5,14	1,62	2,70	306,83	A4_mu_c
0,094	0,69	3,73	2,25	3,60	285,20	A3xi
0,068	0,89	3,30	2,44	3,90	282,77	A3xi
0,18	0,13	5,42	2,20	3,19	287,05	A4_mu_c

Bootstrapped Data (~2/3) + (~1/3) replacement

GT (nT/m)	K (%)	eTh (ppm)	eU (ppm)	CT (μ R/h)	Elevação (m)	Litologia
0,094	0,69	3,73	2,25	3,60	285,20	A3xi
0,068	0,89	3,30	2,44	3,90	282,77	A3xi
0,18	0,13	5,42	2,20	3,19	287,05	A4_mu_c
0,18	0,13	5,42	2,20	3,19	287,05	A4_mu_c

GT (nT/m)	K (%)	eTh (ppm)	eU (ppm)	CT (μ R/h)	Elevação (m)	Litologia
0,219	0,11	5,14	1,62	2,70	306,83	A4_mu_c
0,094	0,69	3,73	2,25	3,60	285,20	A3xi
0,068	0,89	3,30	2,44	3,90	282,77	A3xi
0,18	0,13	5,42	2,20	3,19	287,05	A4_mu_c

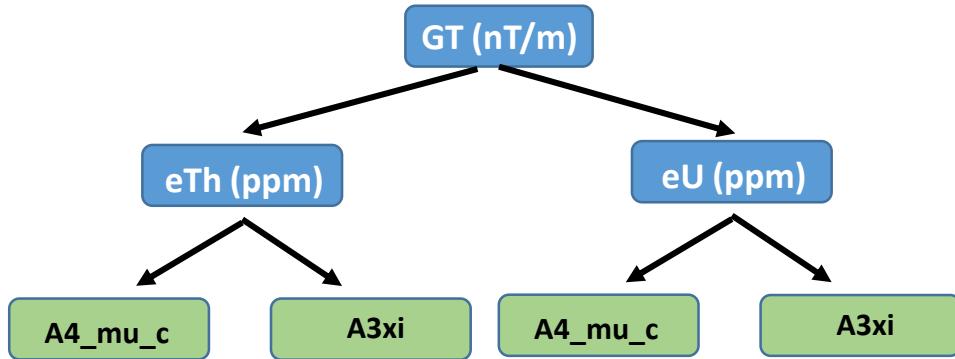
GT (nT/m)	K (%)	eTh (ppm)	eU (ppm)	CT (μ R/h)	Elevação (m)	Litologia
0,219	0,11	5,14	1,62	2,70	306,83	A4_mu_c
0,219	0,11	5,14	1,62	2,70	306,83	A4_mu_c
0,094	0,69	3,73	2,25	3,60	285,20	A3xi
0,068	0,89	3,30	2,44	3,90	282,77	A3xi

*Dados repetidos (replacement)

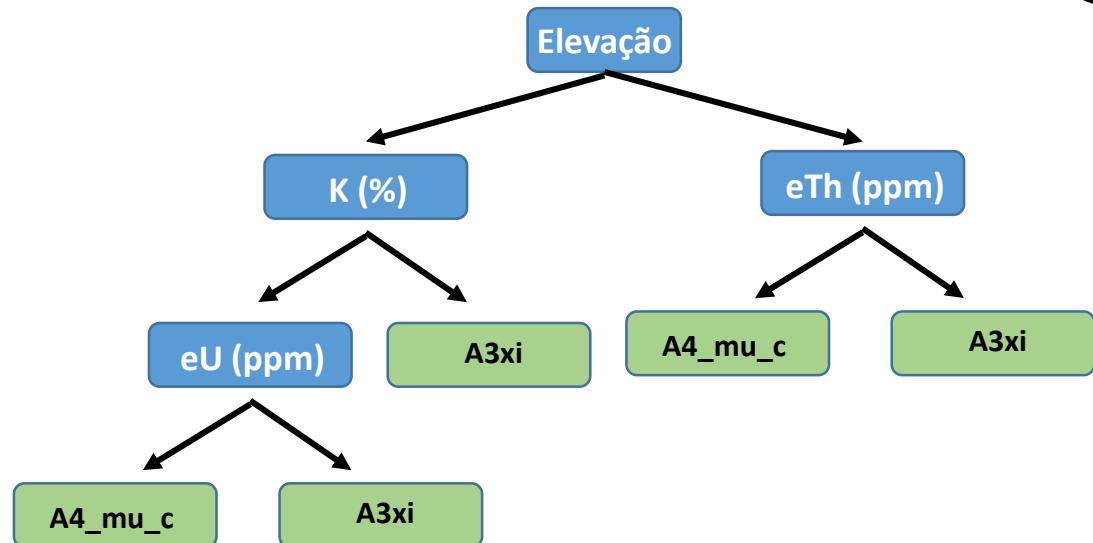
2º

Random Forest

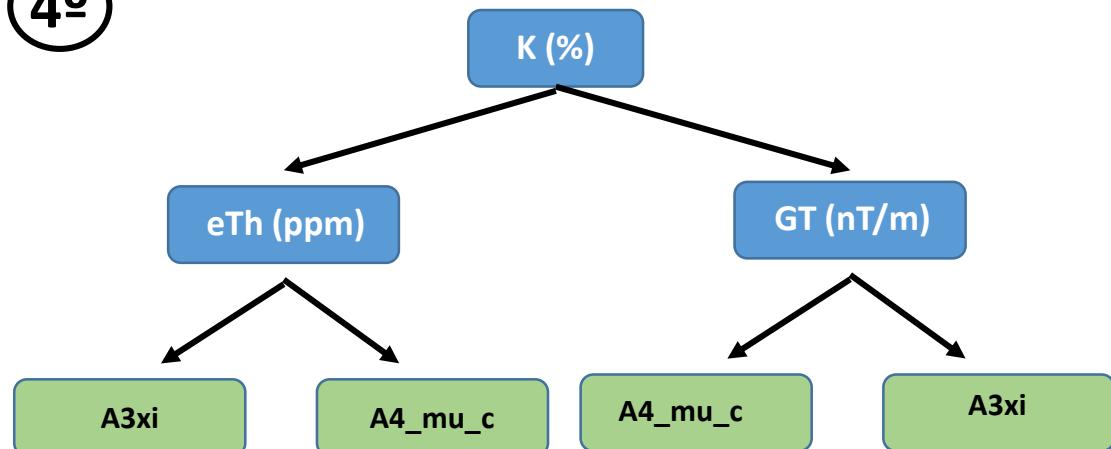
2º



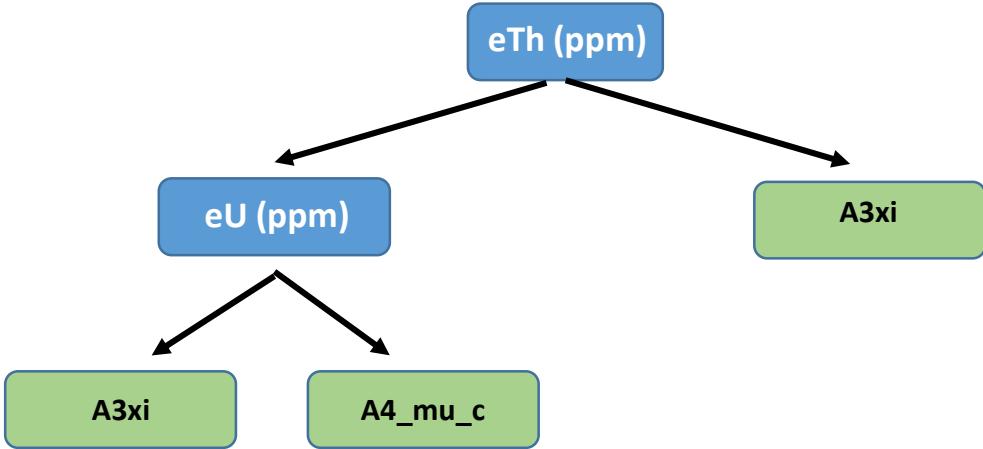
3º



4º



5º

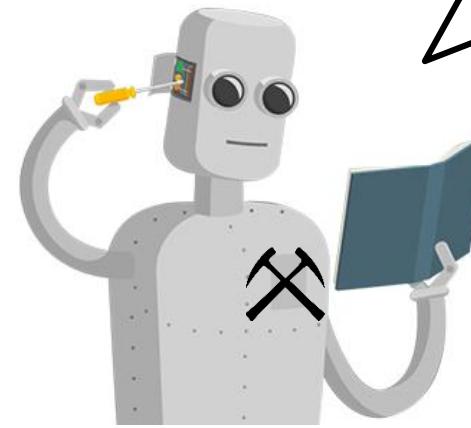


...

Random Forest

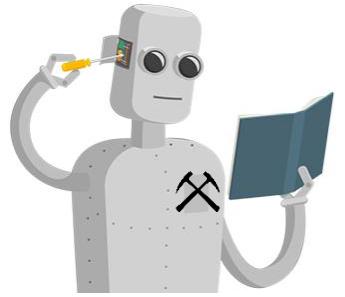
Litologia desconhecida

GT (nT/m)	K (%)	eTh (ppm)	eU (ppm)	CT (μ R/h)	Elevação (m)	Litologia
0,166	0,25	4,79	2,34	3,56	285,27	???



Agora que
aprendi o
comportamento
de várias
litologias, vou
descobrir que
litologia é essa!

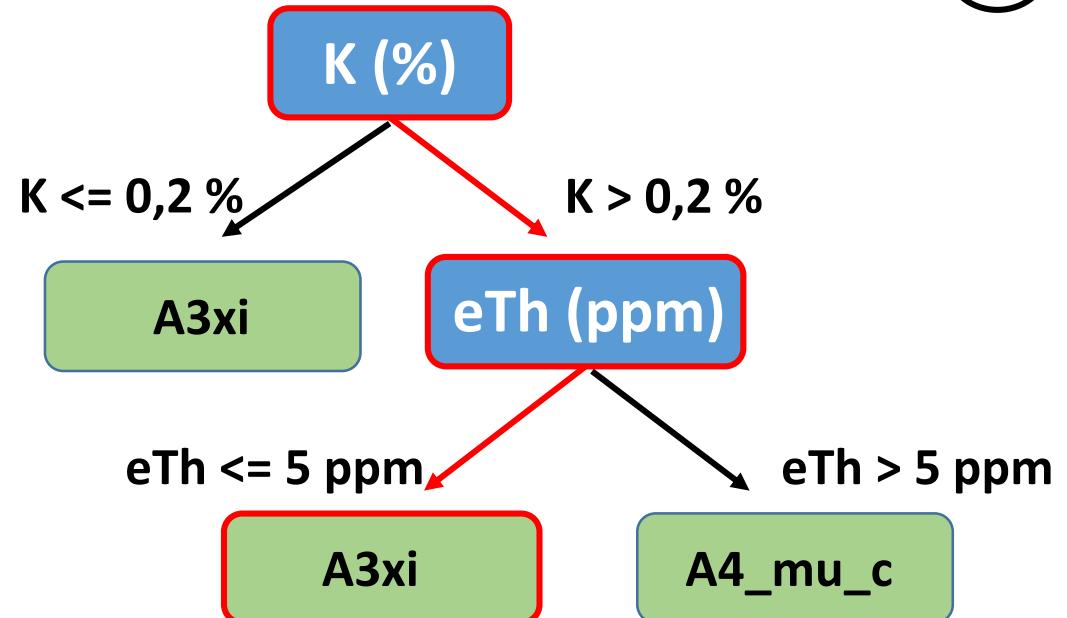
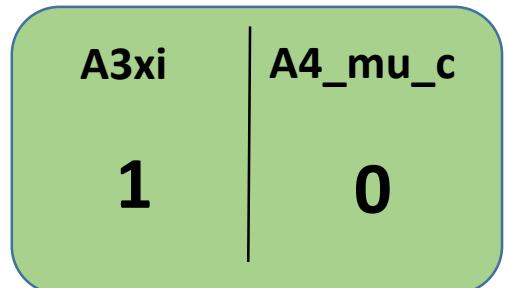
Random Forest



Litologia desconhecida

GT (nT/m)	K (%)	eTh (ppm)	eU (ppm)	CT (μ R/h)	Elevação (m)	Litologia
0,166	0,25	4,79	2,34	3,56	285,27	???

1º

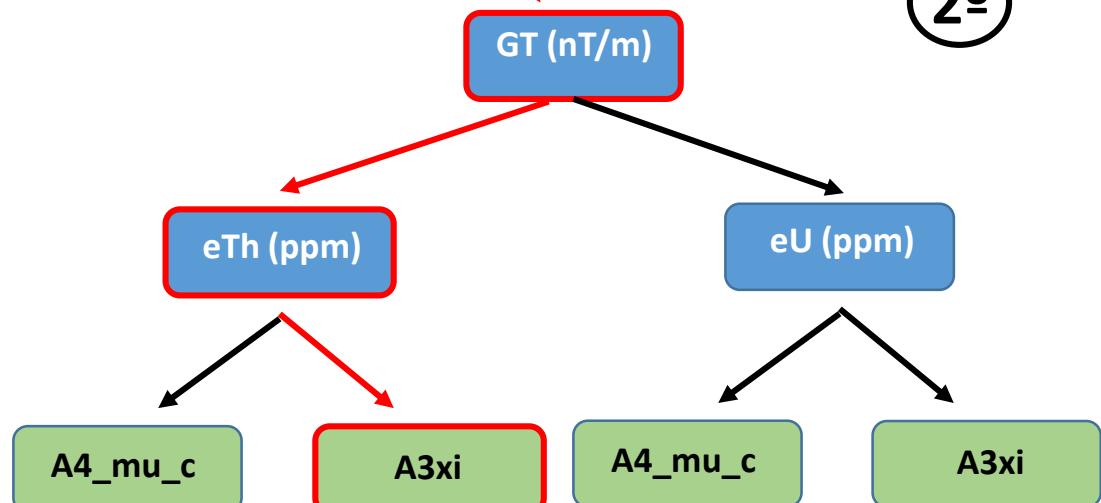
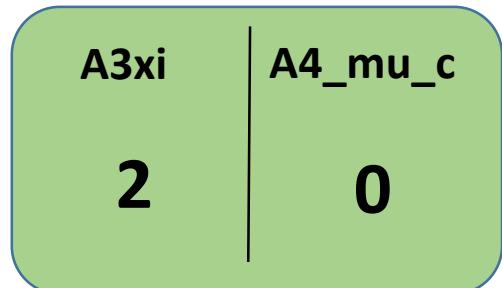


Random Forest

Litologia desconhecida

GT (nT/m)	K (%)	eTh (ppm)	eU (ppm)	CT (μ R/h)	Elevação (m)	Litologia
0,166	0,25	4,79	2,34	3,56	285,27	???

2º



And so on...

Random Forest

GT (nT/m)	K (%)	eTh (ppm)	eU (ppm)	CT (μ R/h)	Elevação (m)	Litologia
0,166	0,25	4,79	2,34	3,56	285,27	A4_mu_c



Mas como definir a quantidade de árvores de decisão?



Random Forest

1º

Dados de Treinamento

GT (nT/m)	K (%)	eTh (ppm)	eU (ppm)	CT (μ R/h)	Elevação (m)	Litologia
0,219	0,21	4,95	1,62	2,70	306,83	A4_mu_c
0,094	0,69	3,73	2,25	3,60	285,20	A3xi
0,068	0,89	3,30	2,44	3,90	282,77	A3xi
0,18	0,13	5,42	2,20	3,19	287,05	A4_mu_c

Bootstrapped Data (~2/3)

GT (nT/m)	K (%)	eTh (ppm)	eU (ppm)	CT (μ R/h)	Elevação (m)	Litologia
0,094	0,69	3,73	2,25	3,60	285,20	A3xi
0,068	0,89	3,30	2,44	3,90	282,77	A3xi
0,18	0,13	5,42	2,20	3,19	287,05	A4_mu_c
0,18	0,13	5,42	2,20	3,19	287,05	A4_mu_c

“Out-Of-Bag Data”

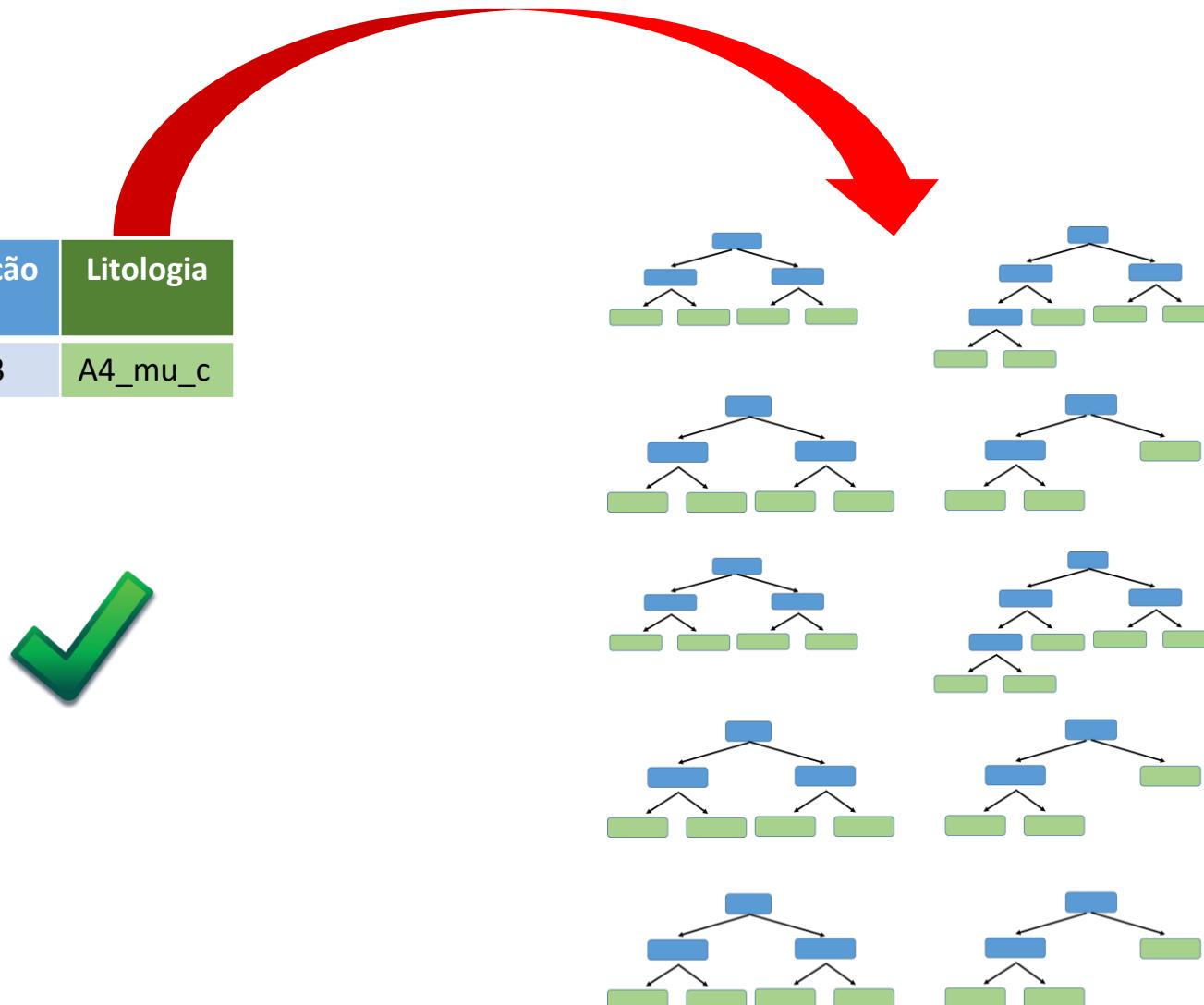
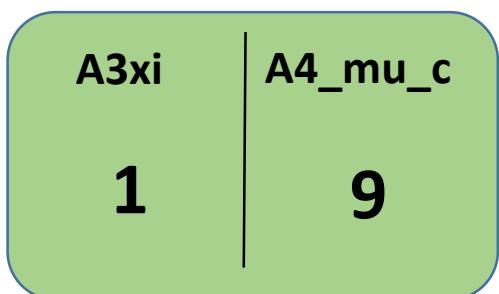
GT (nT/m)	K (%)	eTh (ppm)	eU (ppm)	CT (μ R/h)	Elevação (m)	Litologia
0,219	0,21	4,95	1,62	2,70	306,83	A4_mu_c



Random Forest

“Out-Of-Bag Data”

GT (nT/m)	K (%)	eTh (ppm)	eU (ppm)	CT (μ R/h)	Elevação (m)	Litologia
0,219	0,21	4,95	1,62	2,70	306,83	A4_mu_c



Random Forest

GT (nT/m)	K (%)	eTh (ppm)	eU (ppm)	CT (μ R/h)	Elevação (m)	Litologia
0,219	0,21	5,14	1,62	2,70	306,83	A4_mu_c

A3xi 1	A4_mu_c 9
------------------	---------------------



GT (nT/m)	K (%)	eTh (ppm)	eU (ppm)	CT (μ R/h)	Elevação (m)	Litologia
0,094	0,69	3,73	2,25	3,60	285,20	A3xi

A3xi 8	A4_mu_c 2
------------------	---------------------



“Out-of-Bag Error”

GT (nT/m)	K (%)	eTh (ppm)	eU (ppm)	CT (μ R/h)	Elevação (m)	Litologia
0,068	0,89	3,30	2,44	3,90	282,77	A3xi

A3xi 8	A4_mu_c 2
------------------	---------------------

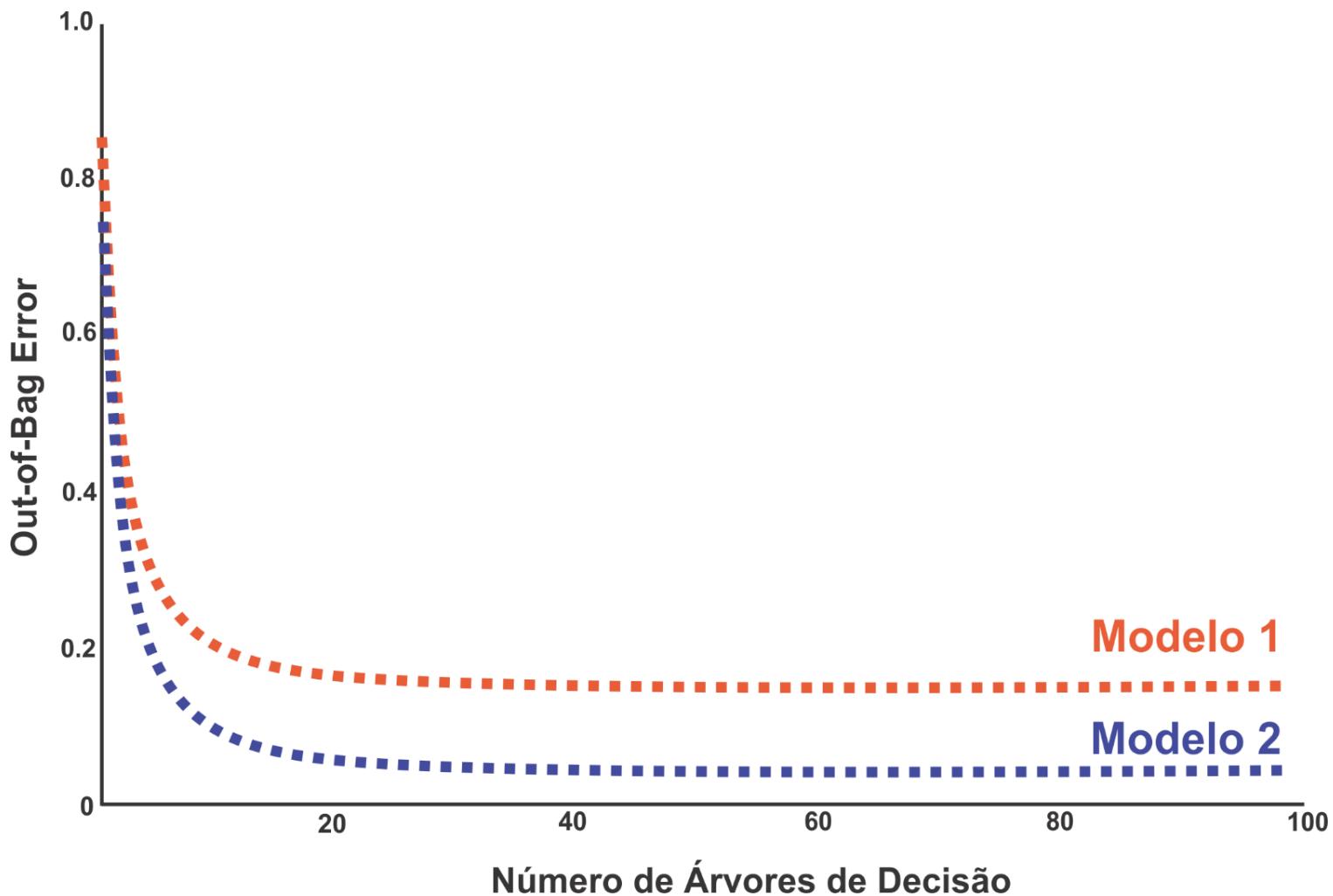


GT (nT/m)	K (%)	eTh (ppm)	eU (ppm)	CT (μ R/h)	Elevação (m)	Litologia
0,18	0,13	5,42	2,20	3,19	287,05	A4_mu_c

A3xi 6	A4_mu_c 4
------------------	---------------------



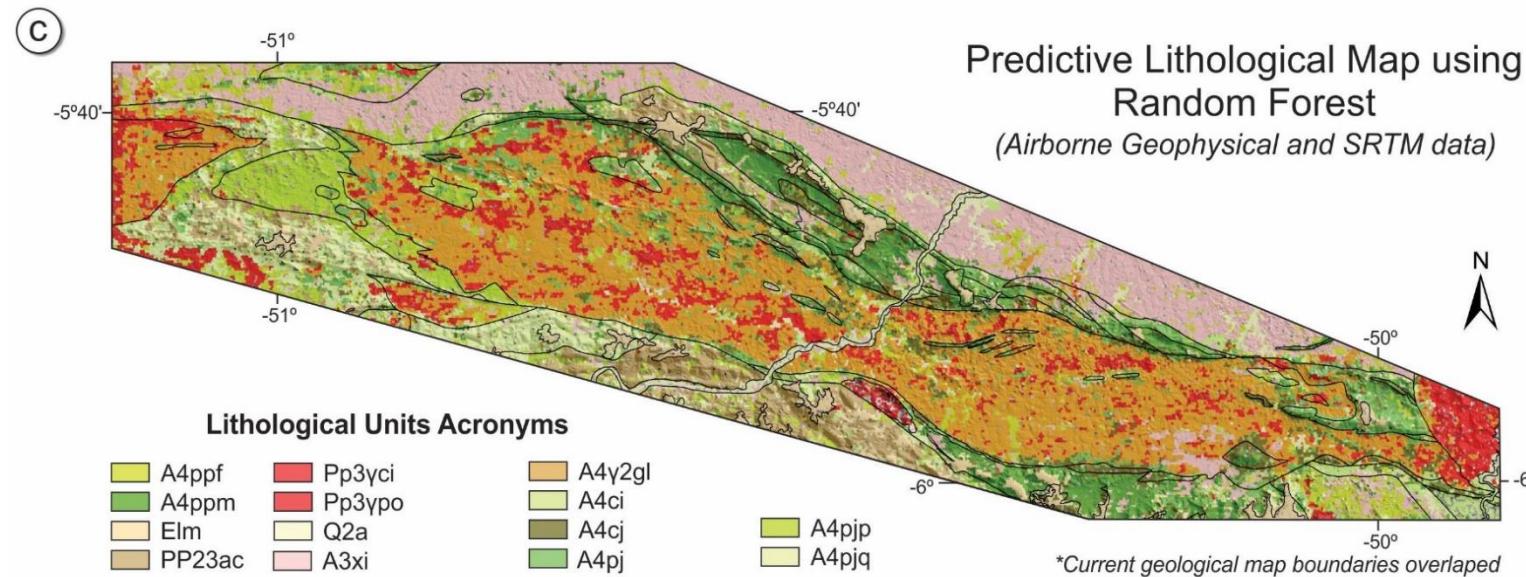
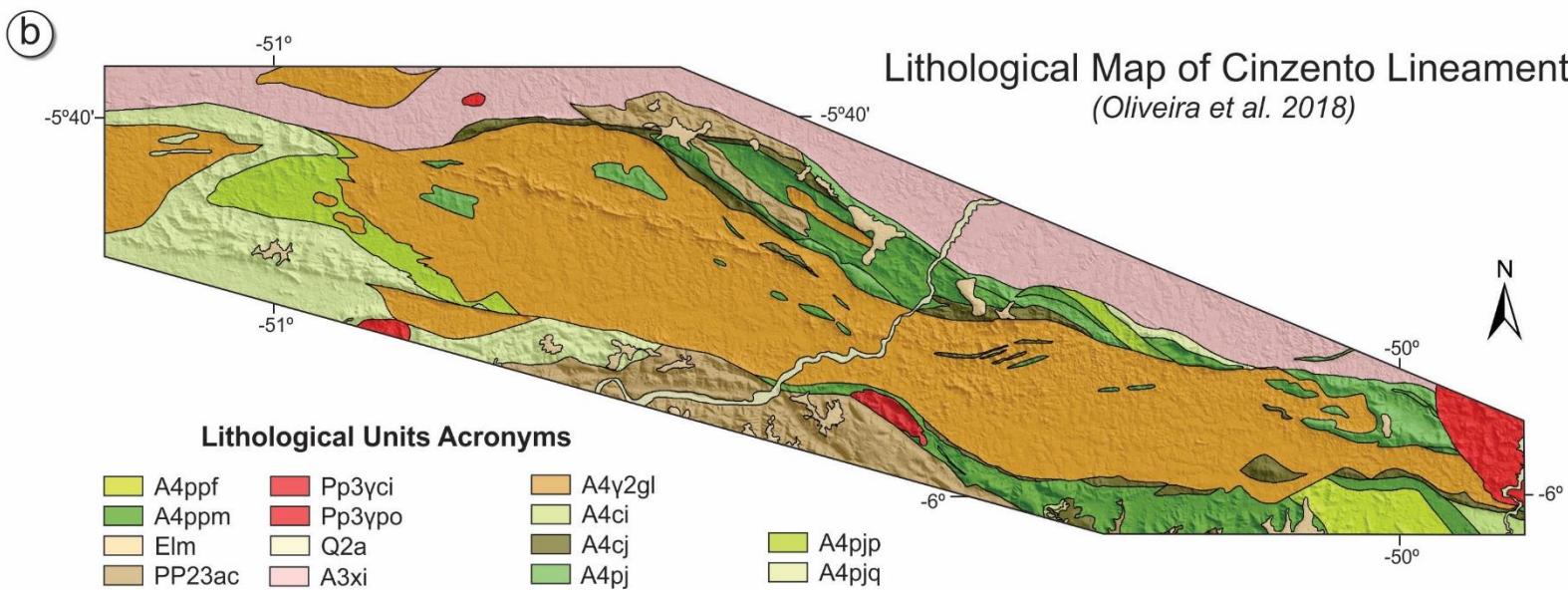
Random Forest



Exemplos de aplicações no Serviço Geológico do Brasil



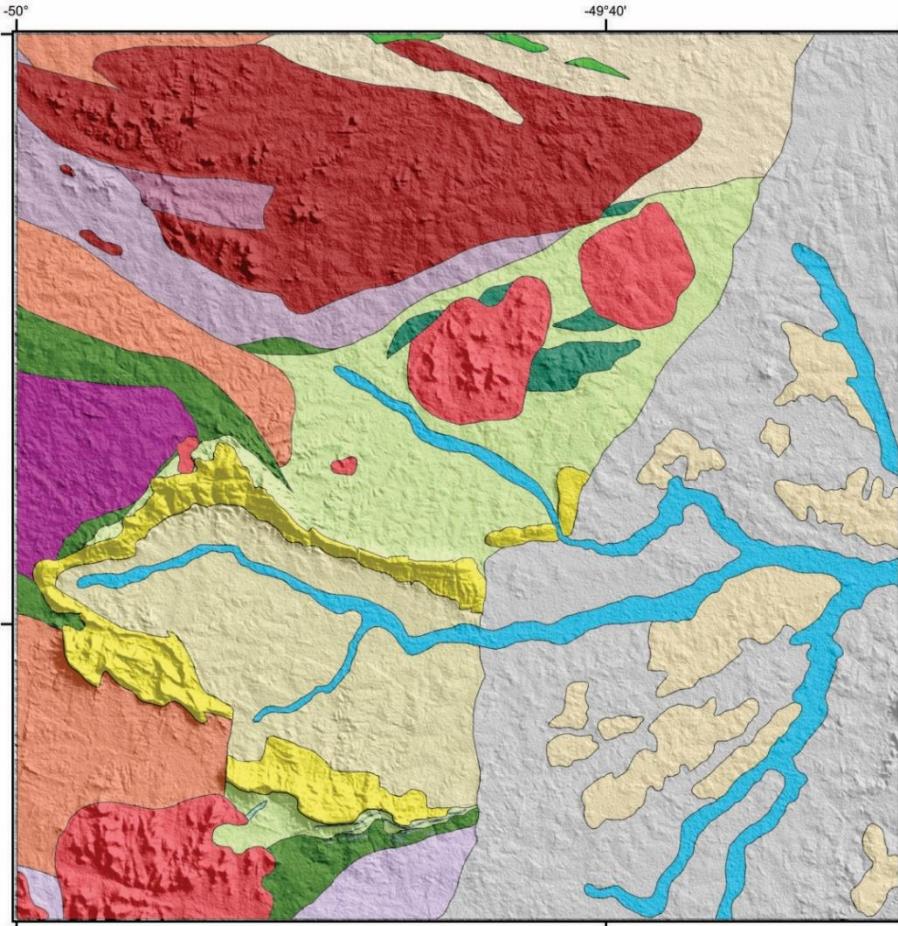
Exemplos



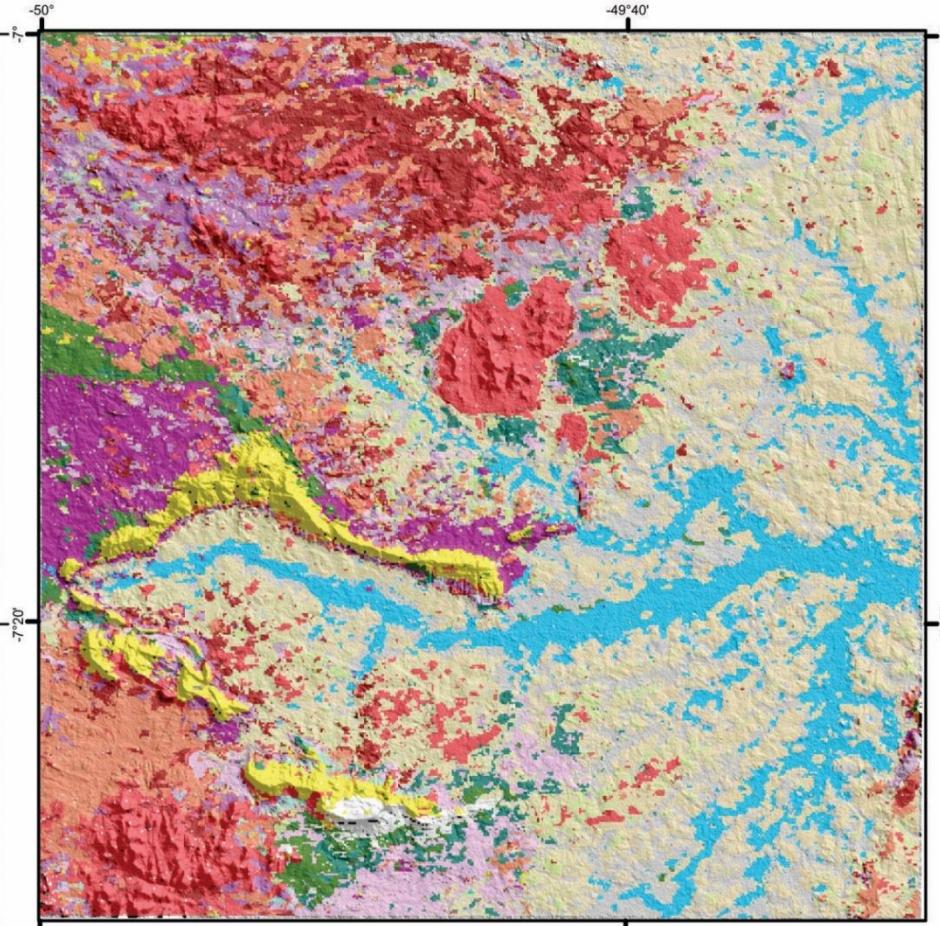
Exemplos

Domínio Rio Maria

Mapa Geológico Atual

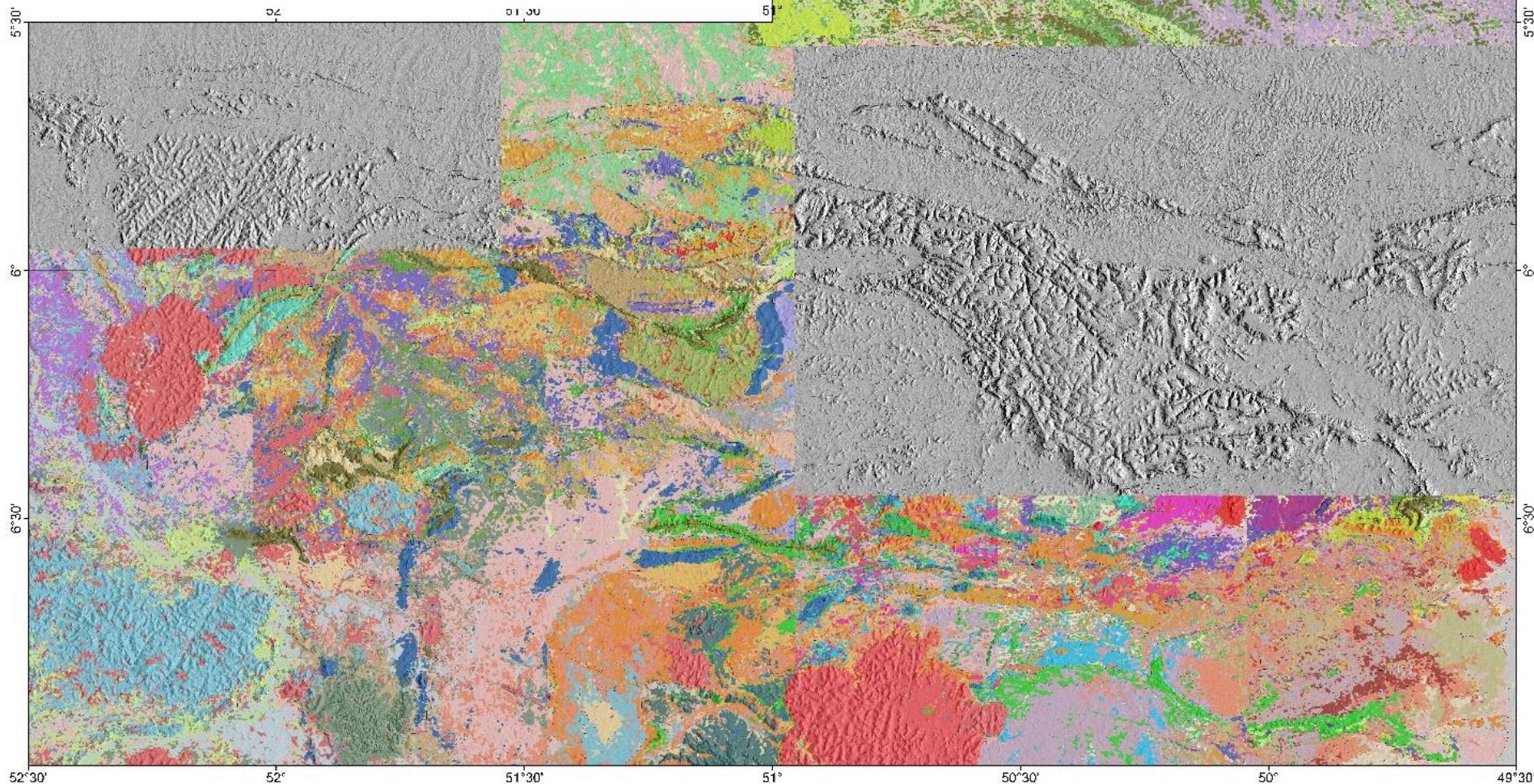


Mapa Geológico Preditivo
(Machine Learning)



Exemplos

MAPA LITOLÓGICO PREDITIVO PROJETO CARAJÁS



Costa, I.S.L.. 2019. Mapa Litológico Preditivo do Projeto Carajás 2020-2023.

É possível gerar modelos de *Machine Learning* sem saber programação?



Sim! Mas...

- A organização dos dados e construção lógica dos modelos pode se tornar complexa caso não seja desenvolvida uma rotina automatizada
- O *Machine Learning* possibilita uma infinidade de possibilidades, que talvez possam ser limitadas em softwares existentes

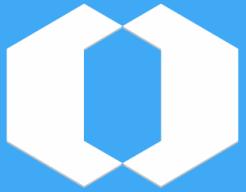
Sugestão!



- “Trate o resultado do **Machine Learning** não como verdade absoluta, mas como uma sugestão inteligente que pode ser útil para entender profundamente o seu problema”



Leo Breiman



SERVIÇO GEOLÓGICO DO BRASIL
CPRM



SECRETARIA DE
GEOLOGIA, MINERAÇÃO
E TRANSFORMAÇÃO MINERAL

MINISTÉRIO DE
MINAS E ENERGIA



Iago Costa

Pesquisador em Geociências – Geofísico

Coordenador Executivo da Diretoria de Geologia e Recursos Minerais

Serviço Geológico do Brasil – CPRM

Brasília - Sede

E-mail: iago.costa@cprm.gov.br

Telefone: 61 2108-8413

www.cprm.gov.br

Módulo 2

02/09



orange
DATA MINING



ArcMap