Aula 6: Validação de Modelos

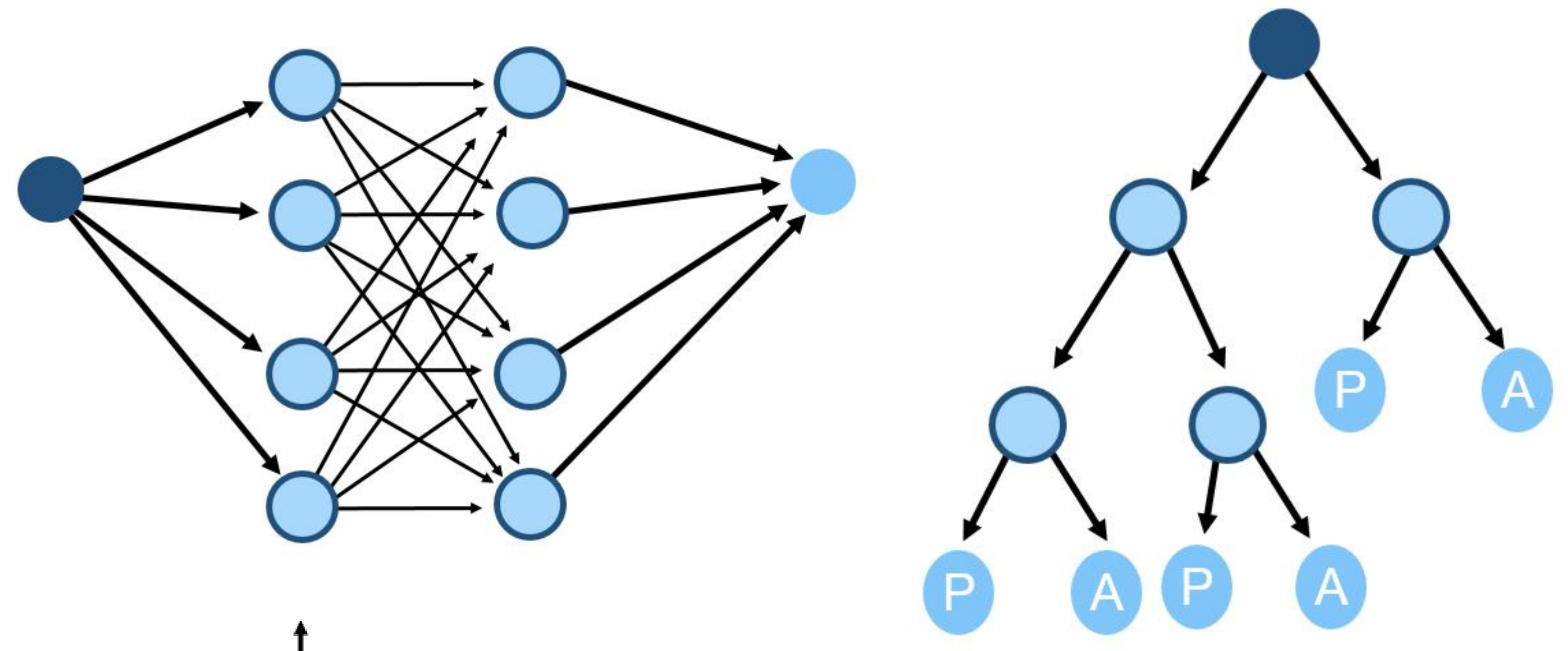
Luíz Fernando Esser

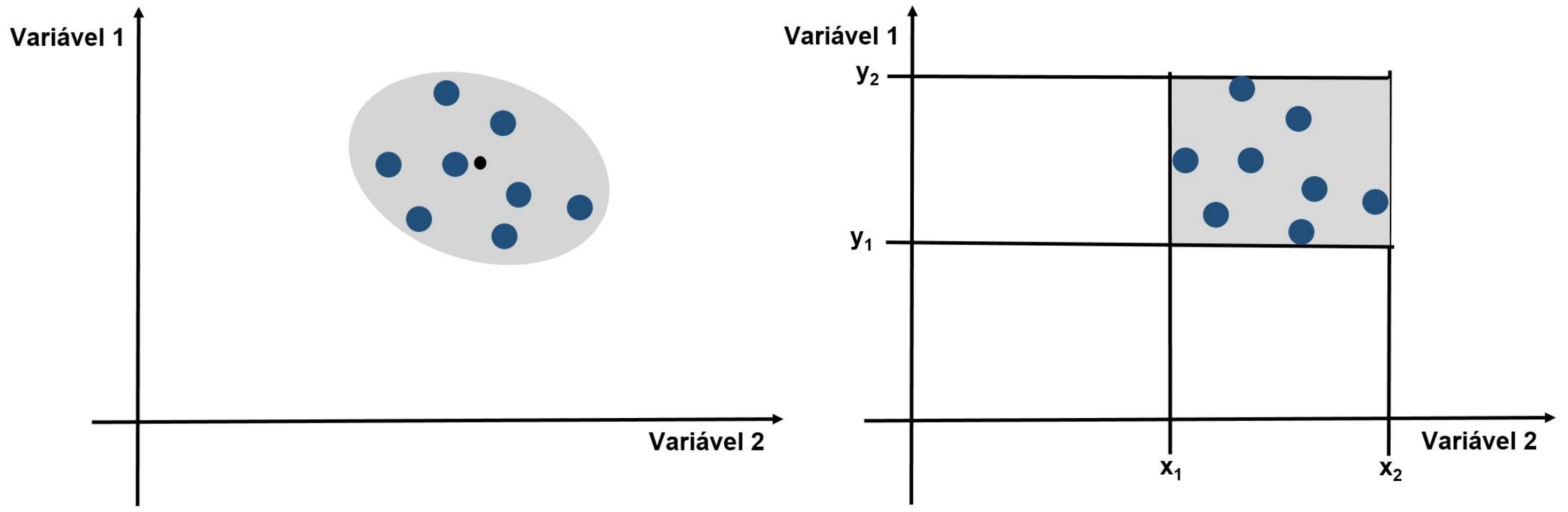
Fundamentos de Modelagem de Distribuição de Espécies no R

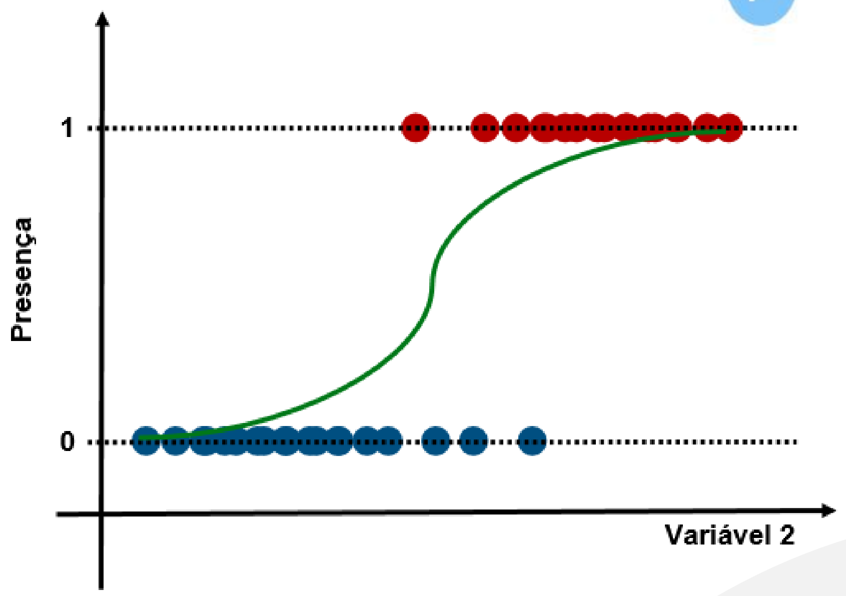
Na última aula...

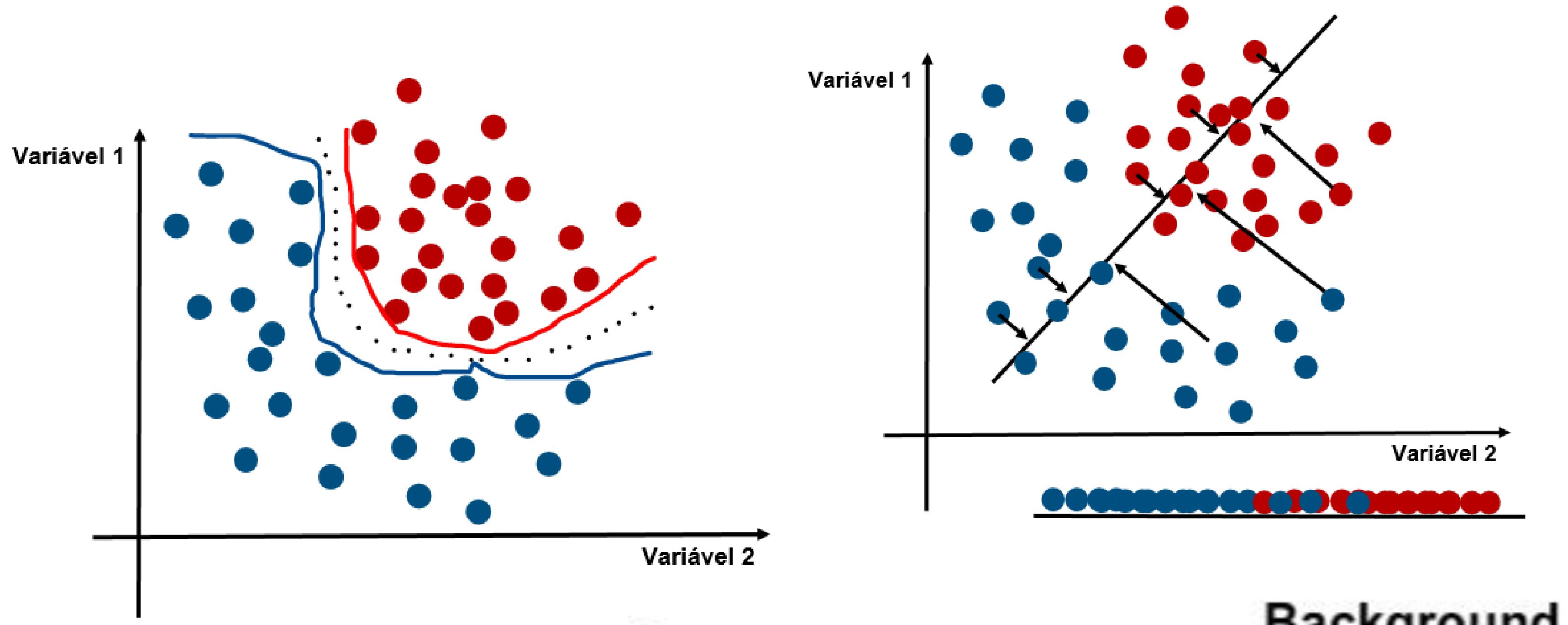
- ✓ Distâncias
- ✓ Envelopes
- ✓ Regressões
- ✓ Redes Neurais
- ✓ SVM
- ✓ CART
- ✓ Discriminantes
- ✓ MaxEnt

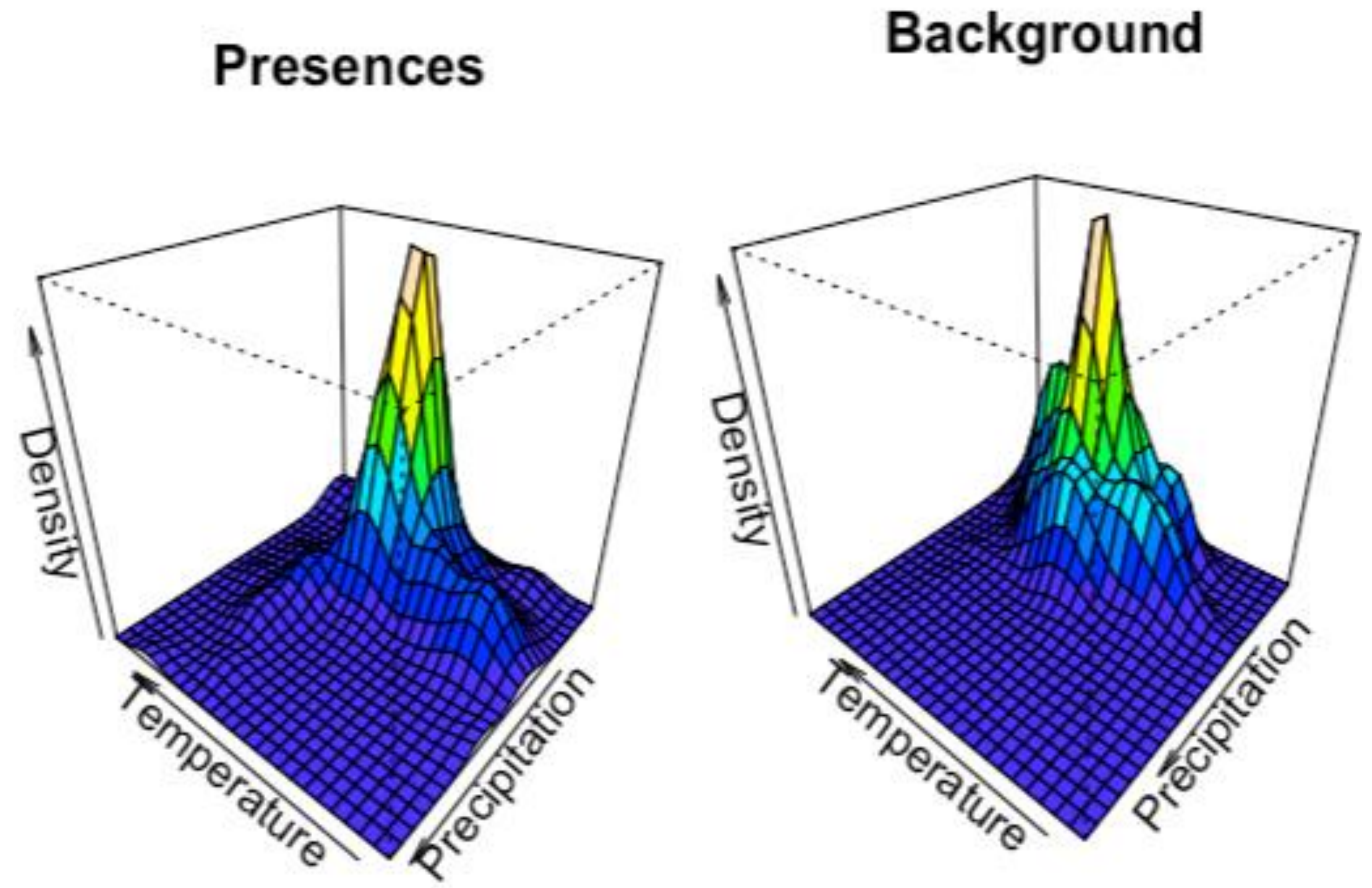
Garbage in, garbage out.











Dados Geográficos

↓

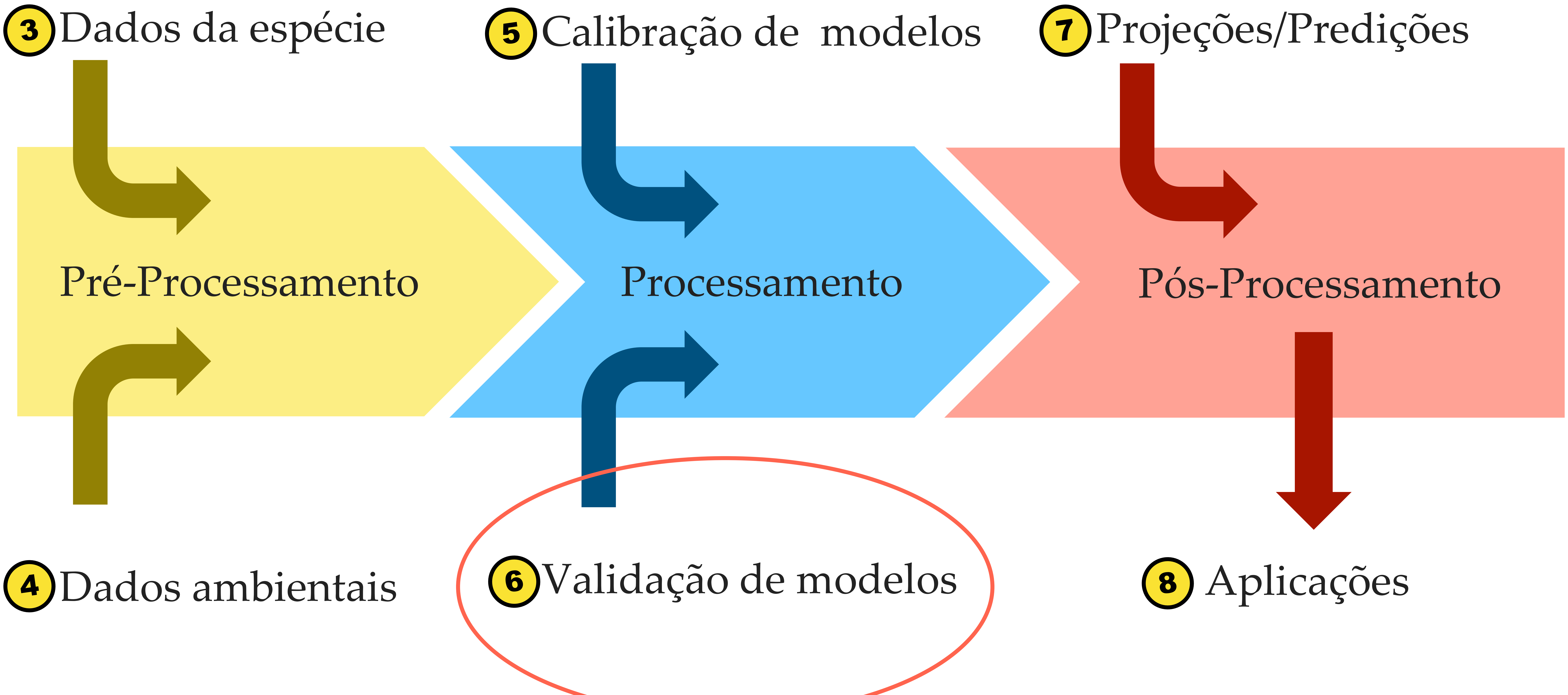
Informação Ambiental

O que vamos fazer hoje?

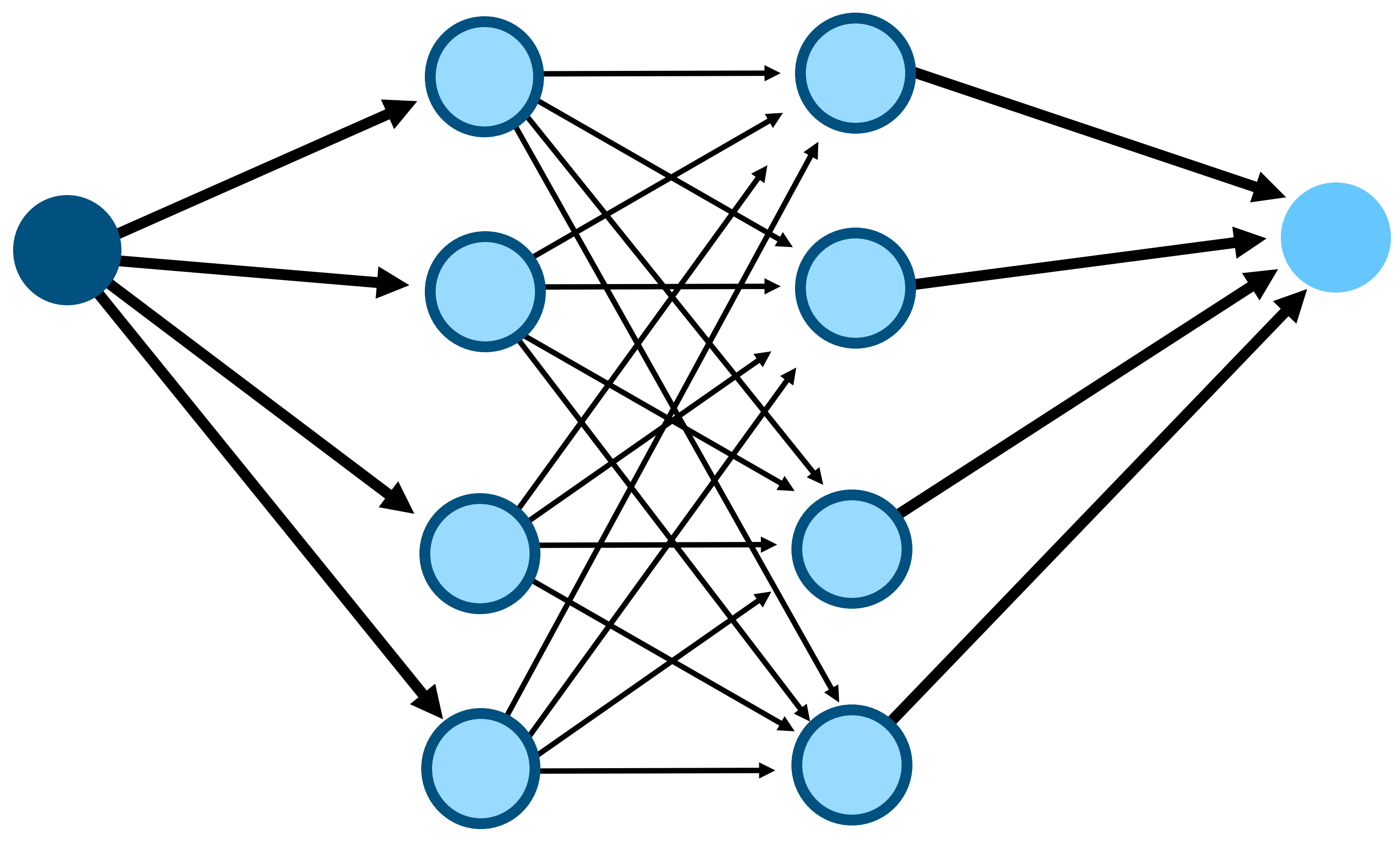
Teórica (13:30 ~ 15:10):

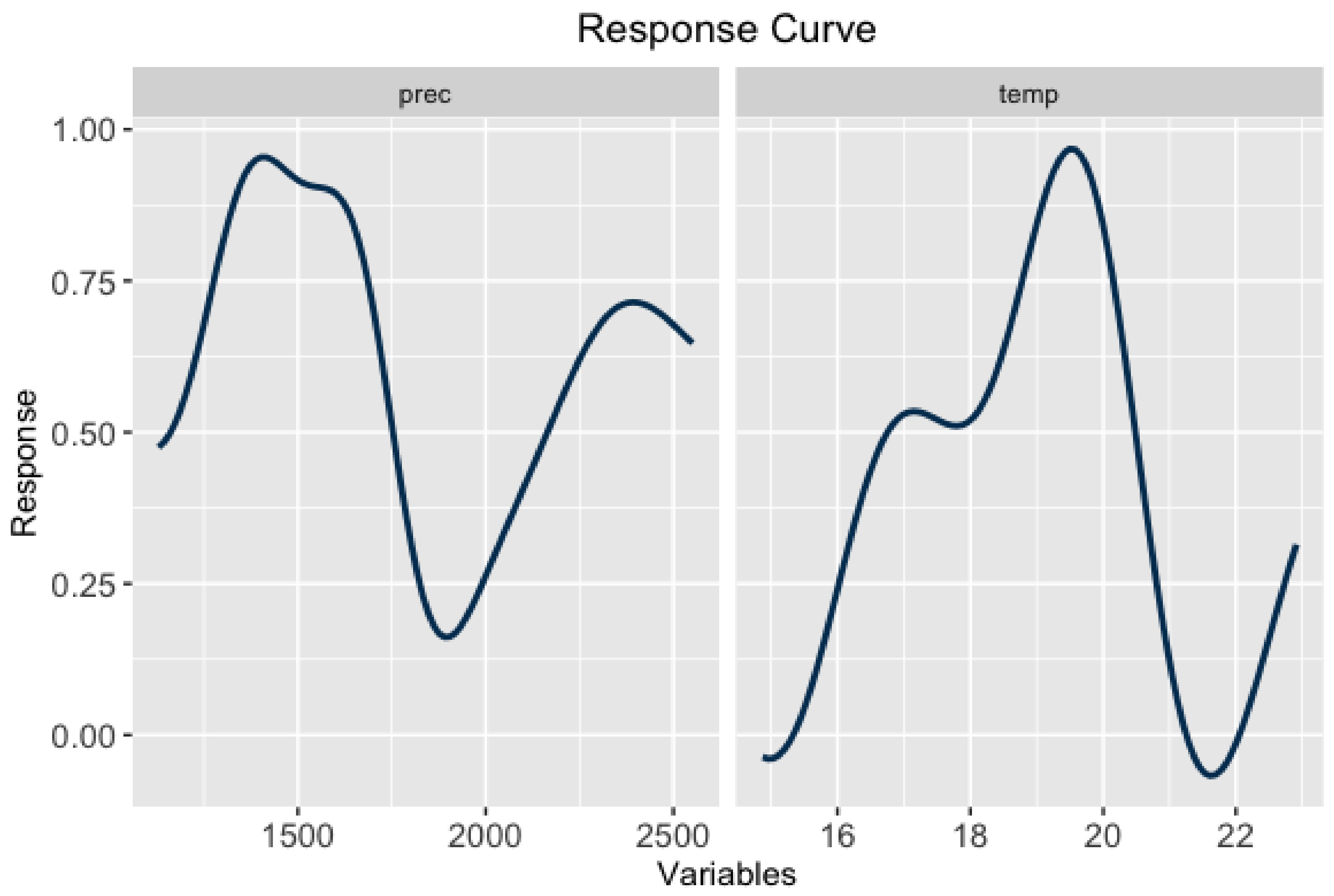
- ❑Problemas com os modelos
- ❑Subsampling
- ❑Validação Cruzada
- ❑Bootstraping
- ❑Matriz de Confusão
- ❑Binarização
- ❑Métrica sem binarização

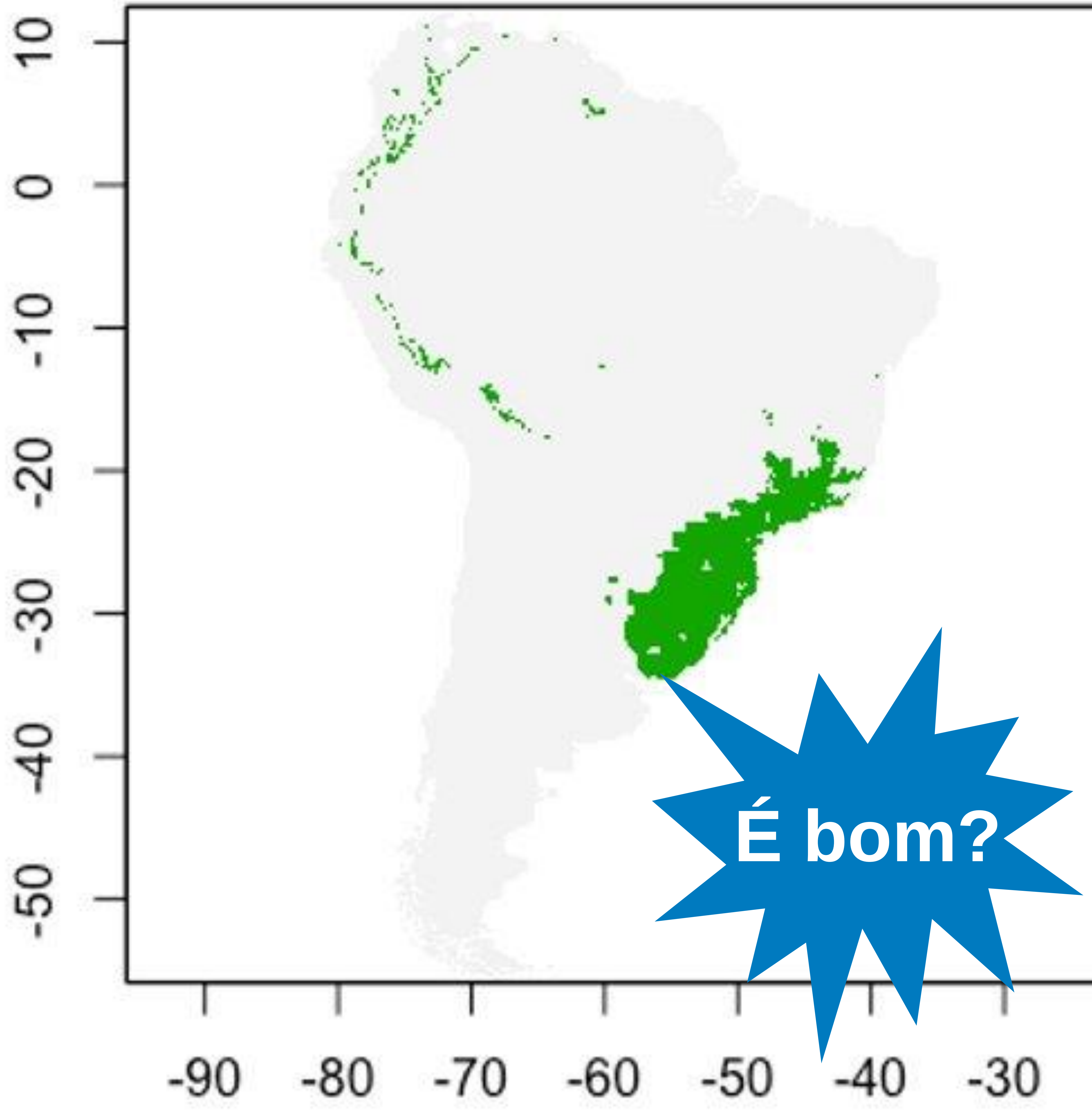
Framework - SDM



Imaginemos o seguinte cenário...





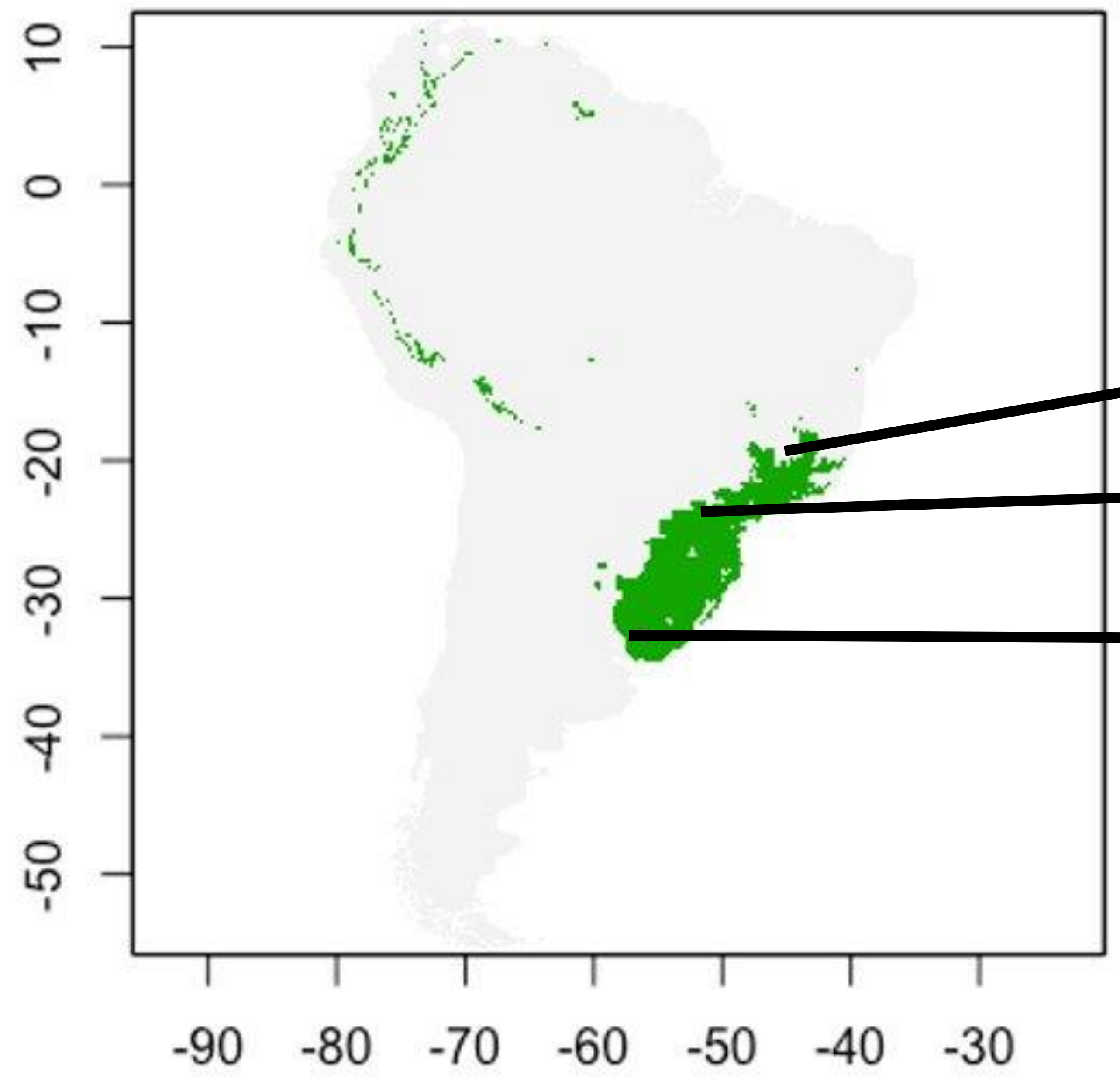


Como saber se meu modelo condiz com a realidade?

Esse modelo é melhor do que o modelo que eu gerei com outro algoritmo (SVM, MaxEnt, …) ?

Algumas ideias...

Voltar a campo para coletar mais dados...



66% de acerto

Acertou

Acertou

Errou

Montar novos modelos

Mudou muito?

Novos dados
são diferentes!

Mudou pouco?

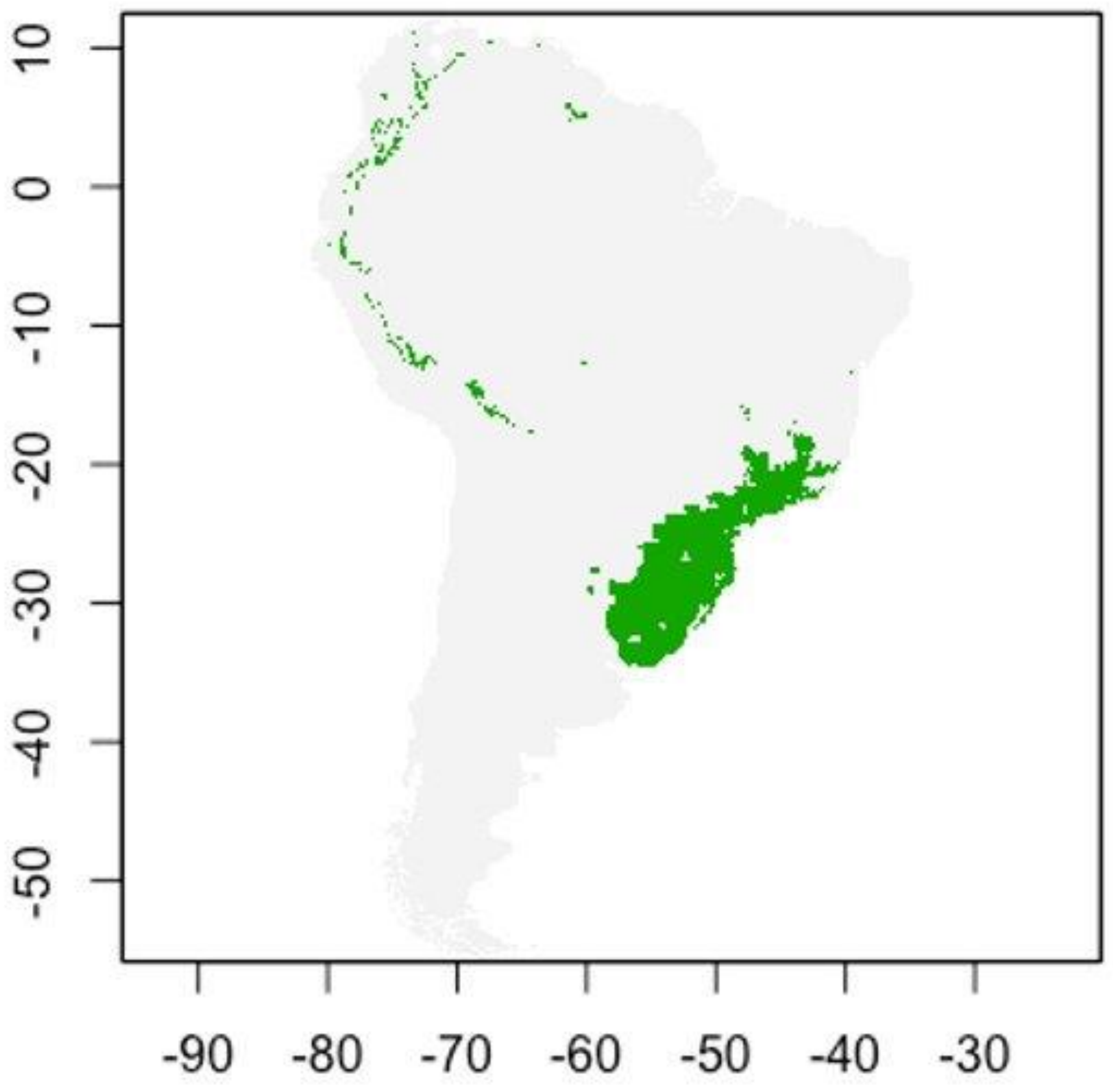
Novos dados
são parecidos!

Escondendo dados:

Subsampling

spp	Long	Lat
Sp1	-49.14	-27.62
Sp1	-49.53	-26.81
Sp1	-49.41	-26.97
Sp1	-49.07	-26.22
Sp1	-48.90	-26.05
Sp1	-49.10	-25.33
Sp1	-48.69	-25.61
Sp1	-48.91	-25.39
Sp1	-48.85	-25.27
Sp1	-48.17	-25.18
Sp1	-49.13	-24.96
Sp1	-48.70	-24.86
Sp1	-48.76	-24.48
Sp1	-49.05	-24.17
Sp1	-45.69	-23.66

Treino



Teste

spp	Long	Lat	
Sp1	-46.61	-23.41	✓
Sp1	-45.30	-23.39	✓
Sp1	-44.81	-23.23	✓
Sp1	-46.96	-23.25	✓
Sp1	-45.02	-23.26	✖

80% de acerto

Escondendo dados:

Validação Cruzada / Crossvalidation

spp	Long	Lat	Grupo
Sp1	-49.14	-27.62	1
Sp1	-49.53	-26.81	1
Sp1	-49.41	-26.97	1
Sp1	-49.07	-26.22	1
Sp1	-48.90	-26.05	1
Sp1	-49.10	-25.33	2
Sp1	-48.69	-25.61	2
Sp1	-48.91	-25.39	2
Sp1	-48.85	-25.27	2
Sp1	-48.17	-25.18	2
Sp1	-49.13	-24.96	3
Sp1	-48.70	-24.86	3
Sp1	-48.76	-24.48	3
Sp1	-49.05	-24.17	3
Sp1	-45.69	-23.66	3
Sp1	-46.61	-23.41	4
Sp1	-45.30	-23.39	4
Sp1	-44.81	-23.23	4
Sp1	-46.96	-23.25	4
Sp1	-45.02	-23.26	4

4-fold crossvalidation

→ **Treino: 1 2 3 Teste: 4**

→ **Treino: 1 2 4 Teste: 3**

→ **Treino: 1 3 4 Teste: 2**

→ **Treino: 2 3 4 Teste: 1**

Repetir *n* vezes

Jackknife:
20-fold crossvalidation

→ **Leave-one-out:**
Deixe um de fora.

Escondendo dados:

Bootstrap

Código	spp	Long	Lat
1	Sp1	-49.14	-27.62
2	Sp1	-49.53	-26.81
3	Sp1	-49.41	-26.97
4	Sp1	-49.07	-26.22
5	Sp1	-48.90	-26.05
6	Sp1	-49.10	-25.33
7	Sp1	-48.69	-25.61
8	Sp1	-48.91	-25.39
9	Sp1	-48.85	-25.27
10	Sp1	-48.17	-25.18
11	Sp1	-49.13	-24.96
12	Sp1	-48.70	-24.86
13	Sp1	-48.76	-24.48
14	Sp1	-49.05	-24.17
15	Sp1	-45.69	-23.66
16	Sp1	-46.61	-23.41
17	Sp1	-45.30	-23.39
18	Sp1	-44.81	-23.23
19	Sp1	-46.96	-23.25
20	Sp1	-45.02	-23.26

Reamostragem com reposição

Treino: 75%

4
5
7
8
9
9
10
11
12
12
14
14
17
18
19

Teste: 25%

3
3
9
11
15

Pontos sorteados para o grupo treino também podem cair no grupo teste!

Qual dos métodos se saiu melhor?

spp	Long	Lat
Sp1	-46.61	-23.41
Sp1	-45.30	-23.39
Sp1	-44.81	-23.23
Sp1	-46.96	-23.25
Sp1	-45.02	-23.26

80% de acerto → **Acurácia (proporção de acertos)**

Matriz de Confusão

Matriz de Confusão

Dados Modelados (Observados)

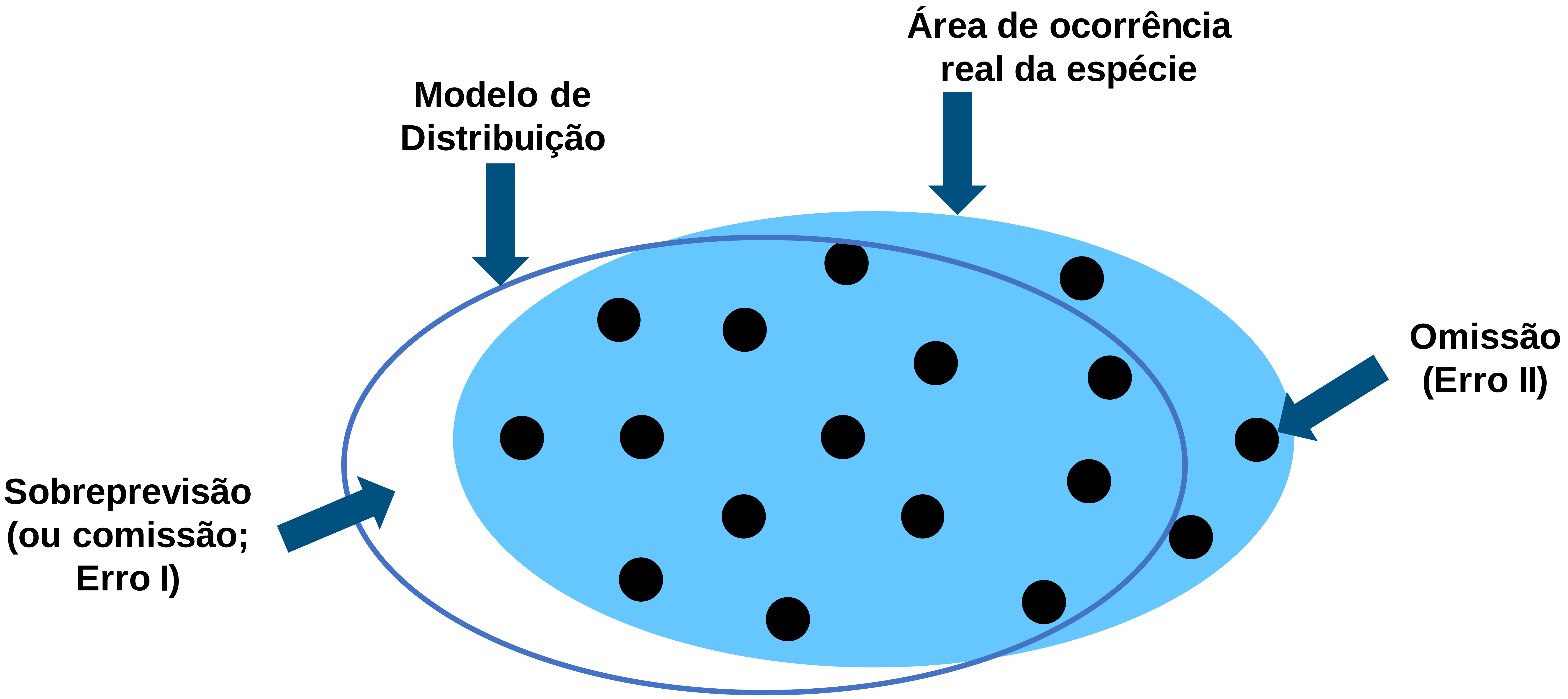
Dados Coletados (Esperados)	Presenças	Ausências
Presenças	Presença verdadeira!	Erro tipo II (falso negativo)
Ausências	Erro tipo I (falso positivo)	Ausência verdadeira!

Os algoritmos retornam probabilidade de presença!

Como conseguir dados de presença e ausência dos modelos?

Binarização

Mais intuitiva: Threshold (limiar) = 0.5
Probabilidade de ocorrência > 0.5 = Presença
Probabilidade de ocorrência < 0.5 = Ausência



tarizM ed Cãonuosf

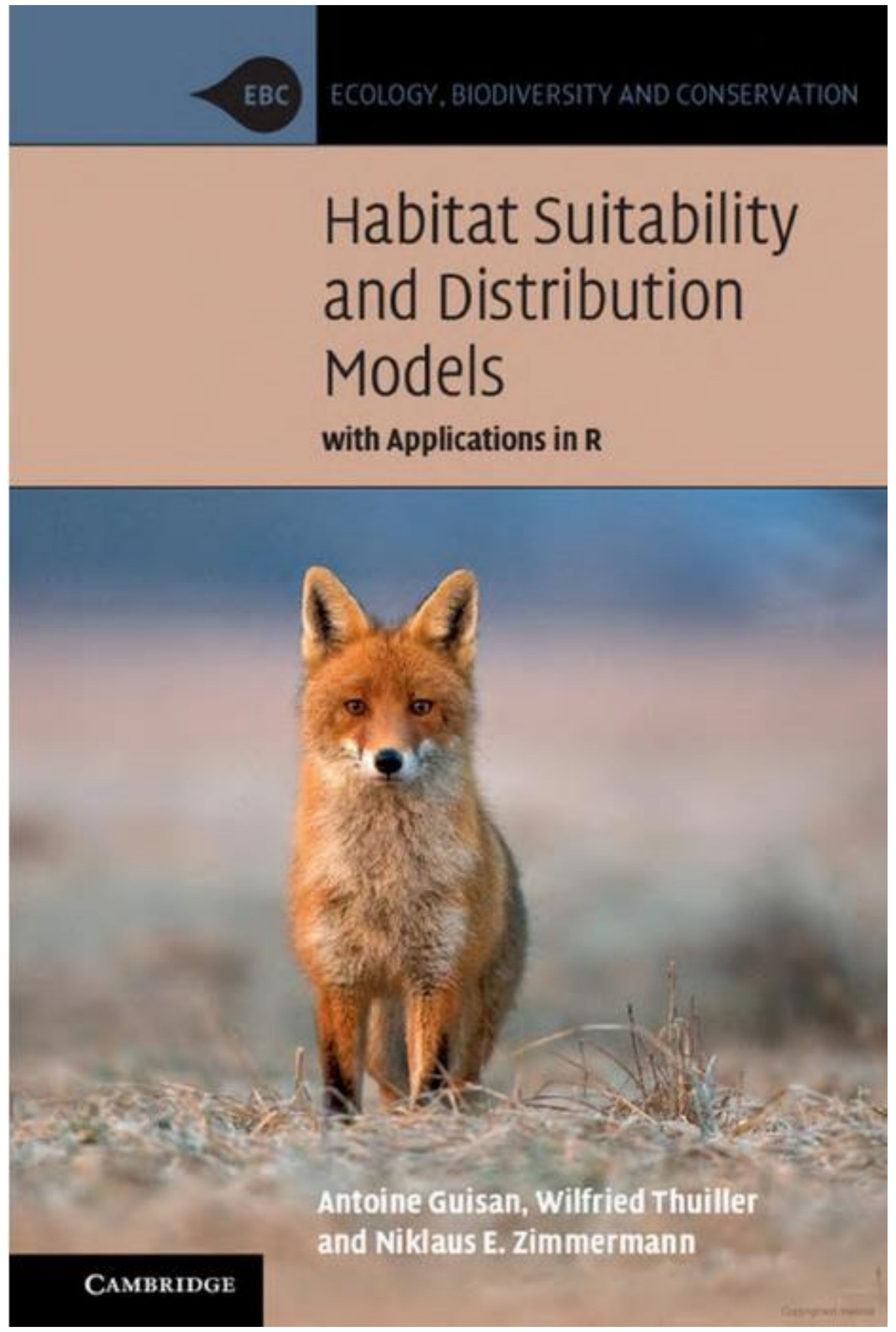
https://en.wikipedia.org/wiki/Confusion_matrix

		Predicted condition		Sources: [13][14][15][16][17][18][19][20]	view·talk·edit
	Total population $= P + N$	Predicted condition **positive** (PP)	Predicted condition **negative** (PN)	Informedness, bookmaker informedness (BM) $= TPR + TNR - 1$	Prevalence threshold (PT) $= \frac{\sqrt{TPR \cdot FPR} - FPR}{TPR - FPR}$
Actual condition	Actual condition positive (P)	**True positive (TP), hit**	**False negative (FN), Type II error, miss, underestimation**	True positive rate (TPR), recall, sensitivity (SEN), probability of detection, hit rate, power $= \frac{TP}{P} = 1 - FNR$	False negative rate (FNR), miss rate $= \frac{FN}{P} = 1 - TPR$
	Actual condition negative (N)	**False positive (FP), Type I error, false alarm, overestimation**	**True negative (TN), correct rejection**	False positive rate (FPR), probability of false alarm, fall-out $= \frac{FP}{N} = 1 - TNR$	True negative rate (TNR), specificity (SPC), selectivity $= \frac{TN}{N} = 1 - FPR$
	Prevalence $= \frac{P}{P + N}$	Positive predictive value (PPV), precision $= \frac{TP}{PP} = 1 - FDR$	False omission rate (FOR) $= \frac{FN}{PN} = 1 - NPV$	Positive likelihood ratio (LR+) $= \frac{TPR}{FPR}$	Negative likelihood ratio (LR−) $= \frac{FNR}{TNR}$
	Accuracy (ACC) $= \frac{TP + TN}{P + N}$	False discovery rate (FDR) $= \frac{FP}{PP} = 1 - PPV$	Negative predictive value (NPV) $= \frac{TN}{PN} = 1 - FOR$	Markedness (MK), deltaP (Δp) $= PPV + NPV - 1$	Diagnostic odds ratio (DOR) $= \frac{LR+}{LR-}$
	Balanced accuracy (BA) $= \frac{TPR + TNR}{2}$	F_1 score $= \frac{2 \cdot PPV \cdot TPR}{PPV + TPR} = \frac{2TP}{2TP + FP + FN}$	Fowlkes–Mallows index (FM) $= \sqrt{PPV \cdot TPR}$	Matthews correlation coefficient (MCC) $= \sqrt{TPR \cdot TNR \cdot PPV \cdot NPV} - \sqrt{FNR \cdot FPR \cdot FOR \cdot FDR}$	Threat score (TS), critical success index (CSI) $= \frac{TP}{TP + FN + FP}$

Max (Sensitivity + Specificity)

TSS (True Skill Statistics) = Sensitivity + Specificity -1

Métricas para validação



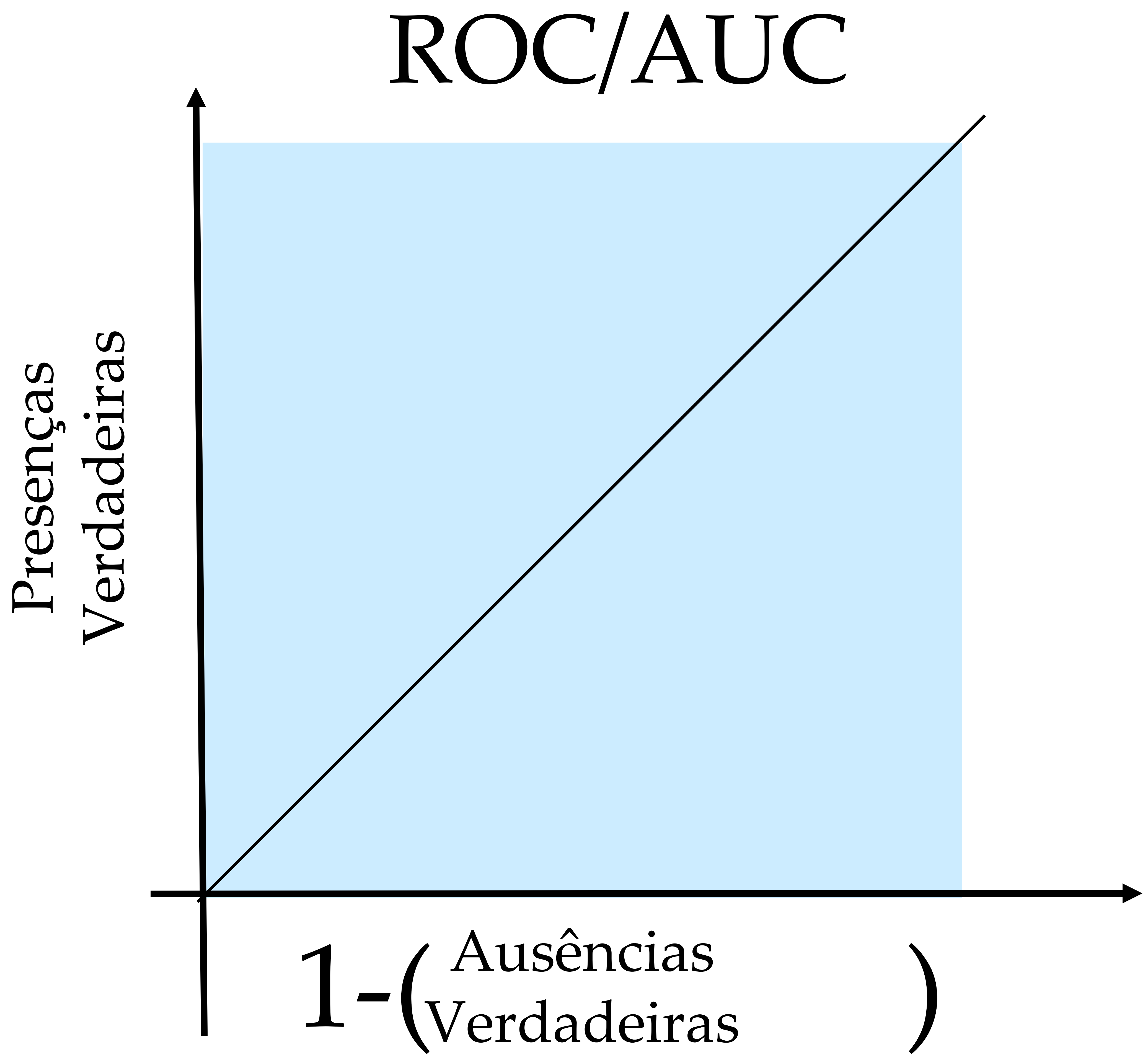
Capítulo 15
Table 15.3

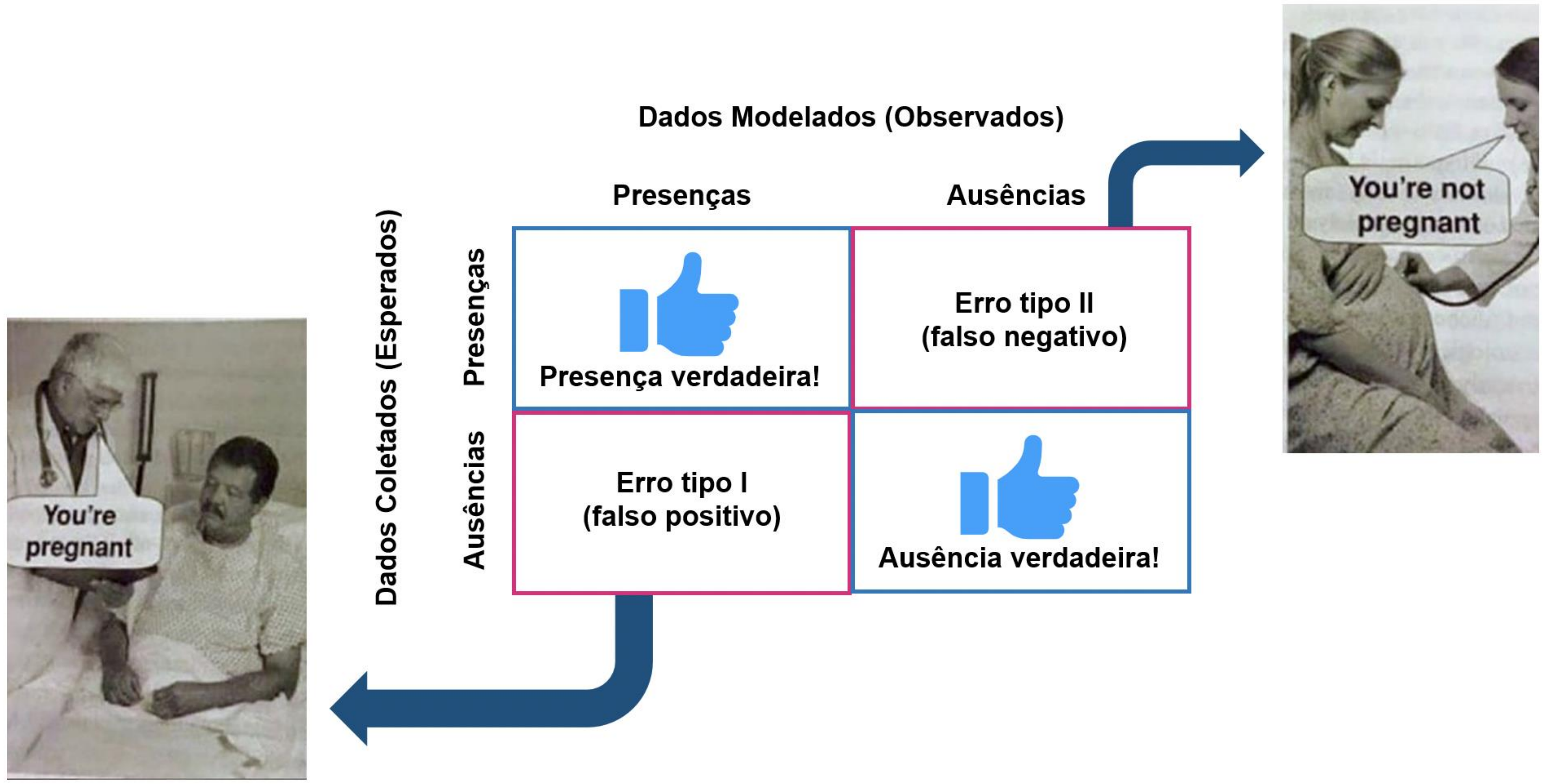
Table 15.3 *The most commonly used metrics that can be derived from a two-way contingency table comparing presence–absence observations to binary predictions. These metrics are therefore threshold-dependent. TP = true presence, FP = false presence, FA = false absence, TA = true absence, N = TP + FP + FA + TA; see Table 15.2. See also Liu et al. (2011) for additional measures.*

Type	Metric	Abbreviation	Description	Range	Formula
Data properties	Sample size	N	Total number of observations	[1: inf]	TP + FP + FA + TA
	Prevalence	PRE	Proportion of presences in the dataset	[0: 1]	(TP + FA)/N
	Overall diagnostic power	ODP	Proportion of absences in the dataset	[0: 1]	(FP + TA)/N = 1 – PREV
Optimist's view(no difference between types of errors)	Correct classification rate	CCR	Percentage of correct predictions (presences and absences)	[0: 1]	(TP + TA)/N
	Misclassification rate	MR	Percentage of false predictions (presences and absences)	[0: 1]	(FP + FA)/N
Observer's view(by column in Table 15.2)	Sensitivity (=true positive rate)	SE	Percentage of presences correctly predicted	[0: 1]	TP/(TP + FA)
	False absence rate (=false negative rate)	FAR	Percent of presences falsely predicted	[0: 1]	FA/(TP + FA) = 1 – SE
	Specificity (=true negative rate)	SP	Percentage of absences correctly predicted	[0: 1]	TA/(TA + FP)
	False presence rate (=false positive rate)	FPR	Percentage of absences falsely predicted	[0: 1]	FP/(FP + TA) = 1 – SP
Modeler's view (by row in Table 15.2)	Presence predictive power (=positive predictive power)	PPP	Percentage of all positive predictions being presences	[0: 1]	TP/(TP + FP)
	Absence predictive power (=negative predictive power)	APP	Percentage of all negative predictions being absences	[0: 1]	TA/(FA + TA)
Balanced view (full use of the confusion matrix,i.e. Table 15.2)	Normalized mutual information	NMI	See Forbes (1995); non-monotonic when excessive error rates		[-TP * ln(TP) – FP * ln(FP) – FN * ln(FA)–TN * ln(TA) + (TP+FP) * ln(TP + FP) + (FA + TA)* ln(FA + TA)]/[N * lnN – ((TP + FA) * ln(TP + FA) + (FP + TA) * ln(FP + TA))]
	Kappa	K	See Cohen (1960); sensitive to sample size and prevalence	[-1: 1]	[(TP + TA)–(((TP + FA) * (TP + FP) + (FP + TA) * (FA + TA))/N)]/[N – (((TP + FA) * (TP + FP) + (FP + TA) * (FA + TA))/N)]
	Weighted Kappa	WK	See Cohen (1968); same as K but with weights assigned to TP, FP, FA and TA	[-1: 1]	Above formula weighted for TP, FP, FA and TA; see Cohen (1968)
	Odds Ratio	OR	Infinite when either b or c are 0; i.e. same value when the algorithm is perfect or lacks one type of error	[0: inf]	(TP * TA)/(FP * FA)
	True skill statistic (or Hanssen-Kuiper skill score)	TSS (or HKSS)	See Hanssen and Kuipers (1965); tends to converge to the prevalence for rare events (i.e. when TA is very large)	[-1: 1]	[(TP * TA) - (FP * FA)]/ [(TP + FA) * (FP + TA)] = SE + SP -1

E se não quiser binarizar?

Area Under the Receiver Operating Characteristic Curve





Alinhavando...

- ✓ Problemas com os modelos
- ✓ Subsampling
- ✓ Validação Cruzada
- ✓ Bootstraping
- ✓ Matriz de Confusão
- ✓ Binarização
- ✓ Métrica sem binarização