The `default` dataset from the book “An Introduction to Statistical Learning”

```
> head(df)
  default student   balance    income
1      No      No  729.5265 44361.625
2      No     Yes  817.1804 12106.135
3      No      No 1073.5492 31767.139
4      No      No  529.2506 35704.494
5      No      No  785.6559 38463.496
6      No     Yes  919.5885  7491.559
> dim(df)
[1] 10000     4
>
```

The `default` dataset from the book “An Introduction to Statistical Learning”

```
> head(df)
  default student   balance    income
1      No      No  729.5265 44361.625
2      No     Yes  817.1804 12106.135
3      No      No 1073.5492 31767.139
4      No      No  529.2506 35704.494
5      No      No  785.6559 38463.496
6      No     Yes  919.5885  7491.559
> dim(df)
[1] 10000     4
>
```

4 predictors

The `default` dataset from the book “An Introduction to Statistical Learning”

Label

```
> head(df)
  default student   balance    income
1      No      No  729.5265 44361.625
2      No     Yes  817.1804 12106.135
3      No      No 1073.5492 31767.139
4      No      No  529.2506 35704.494
5      No      No  785.6559 38463.496
6      No     Yes  919.5885  7491.559
> dim(df)
[1] 10000     4
>
```

We fit a logistic regression classifier to the data and check its predictions

```
logreg <- glm(default ~ ., data=df_train, family=binomial)
y_pred <- predict.glm(logreg, newdata=df_test, type="response")
```

We fit a logistic regression classifier to the data and check its predictions

```
logreg <- glm(default ~ ., data=df_train, family=binomial)
y_pred <- predict.glm(logreg, newdata=df_test, type="response")
```

The accuracy of the classifier after a 10-fold cross validation is…

# 97.1%

Looks impressive, right?

BUT

The dataset is highly **unbalanced** in its classes…

3.3%

`default` = “YES”

96.7%

`default` = “NO”

The dataset is highly **unbalanced** in its classes…

3.3%

`default` = “YES”

96.7%

`default` = “NO”

So a dummy classifier giving “NO” to every data point will achieve…

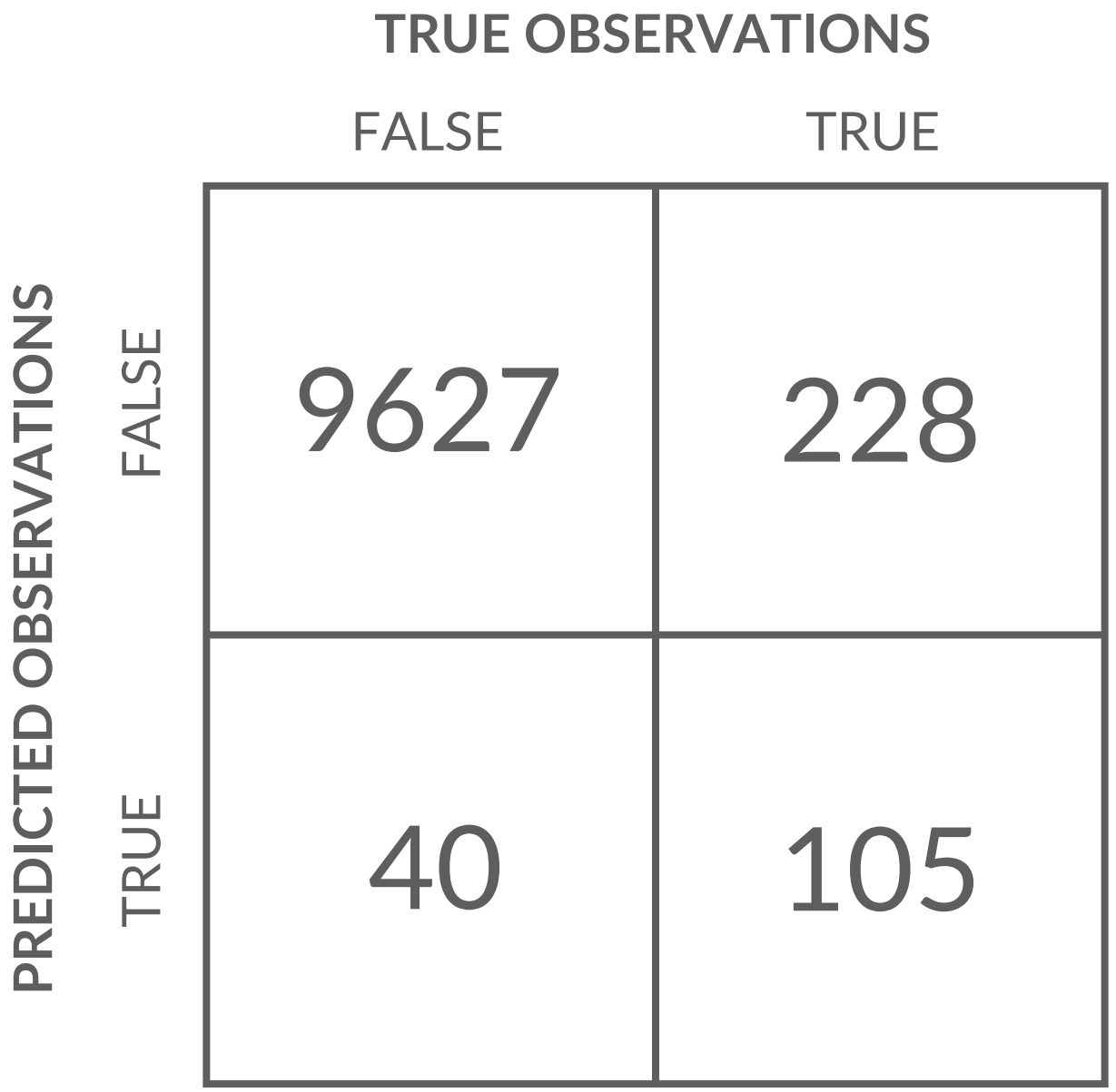
96.7%

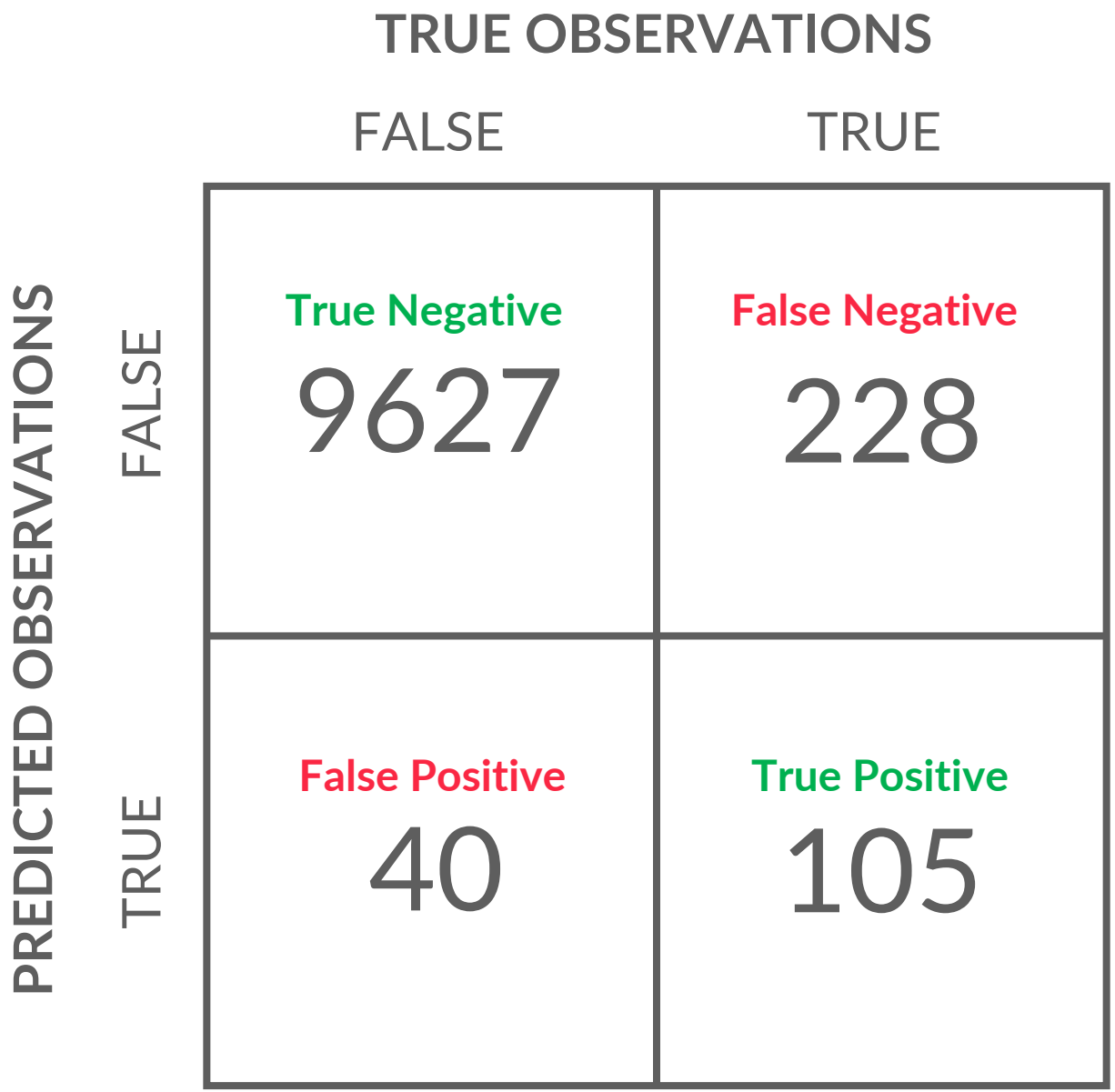
Dummy classifier

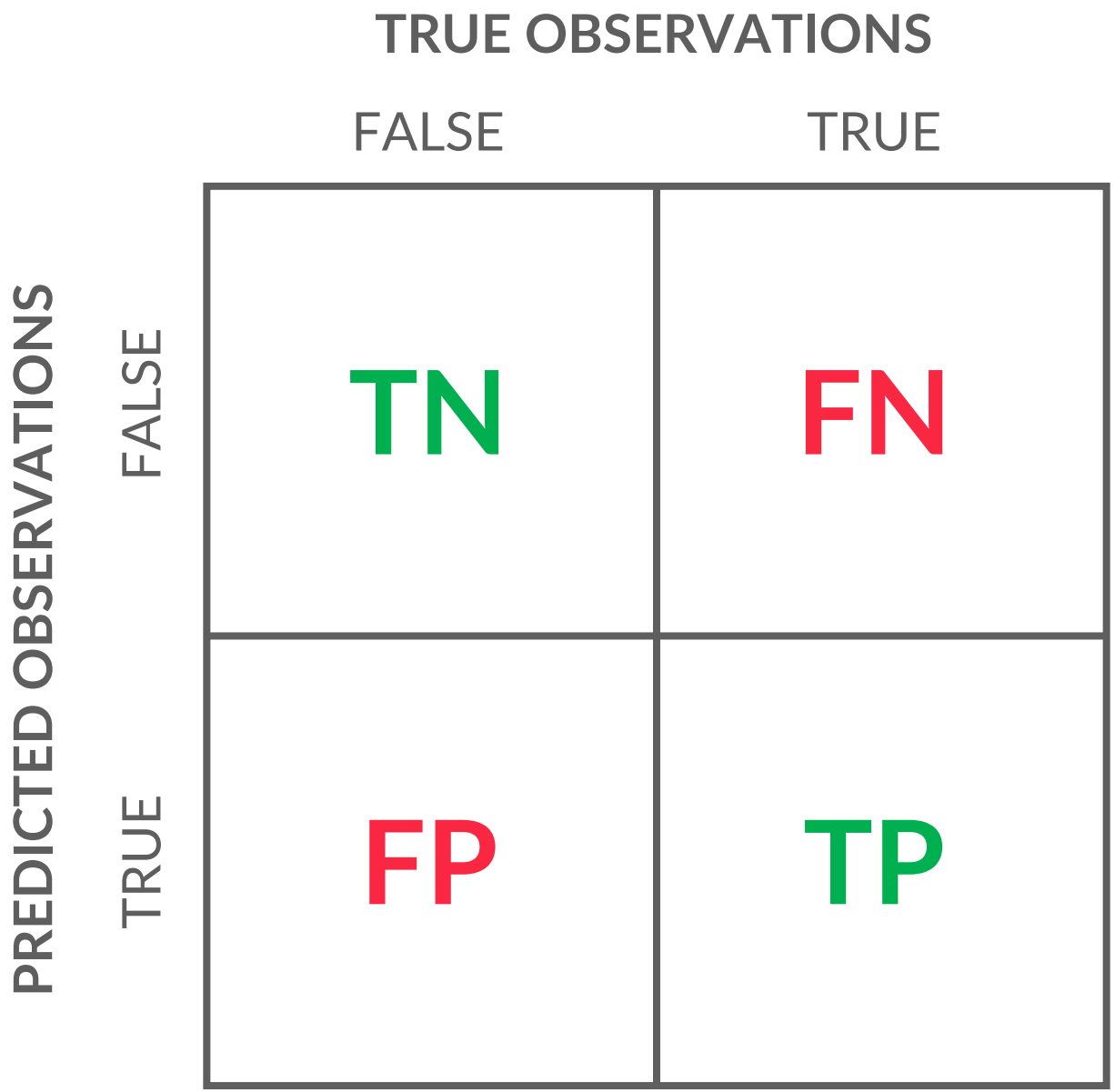
97.1%

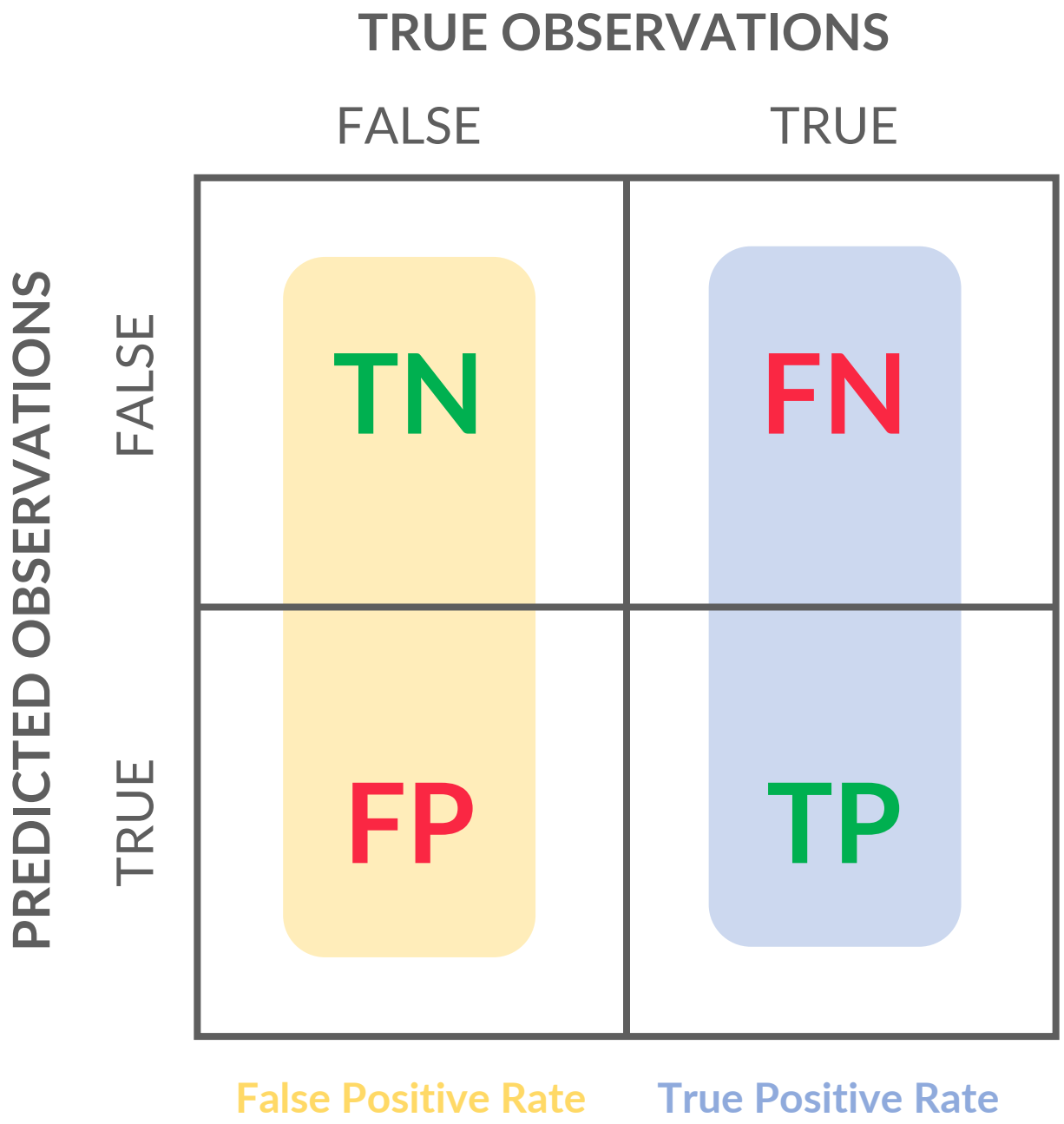
Logistic regression

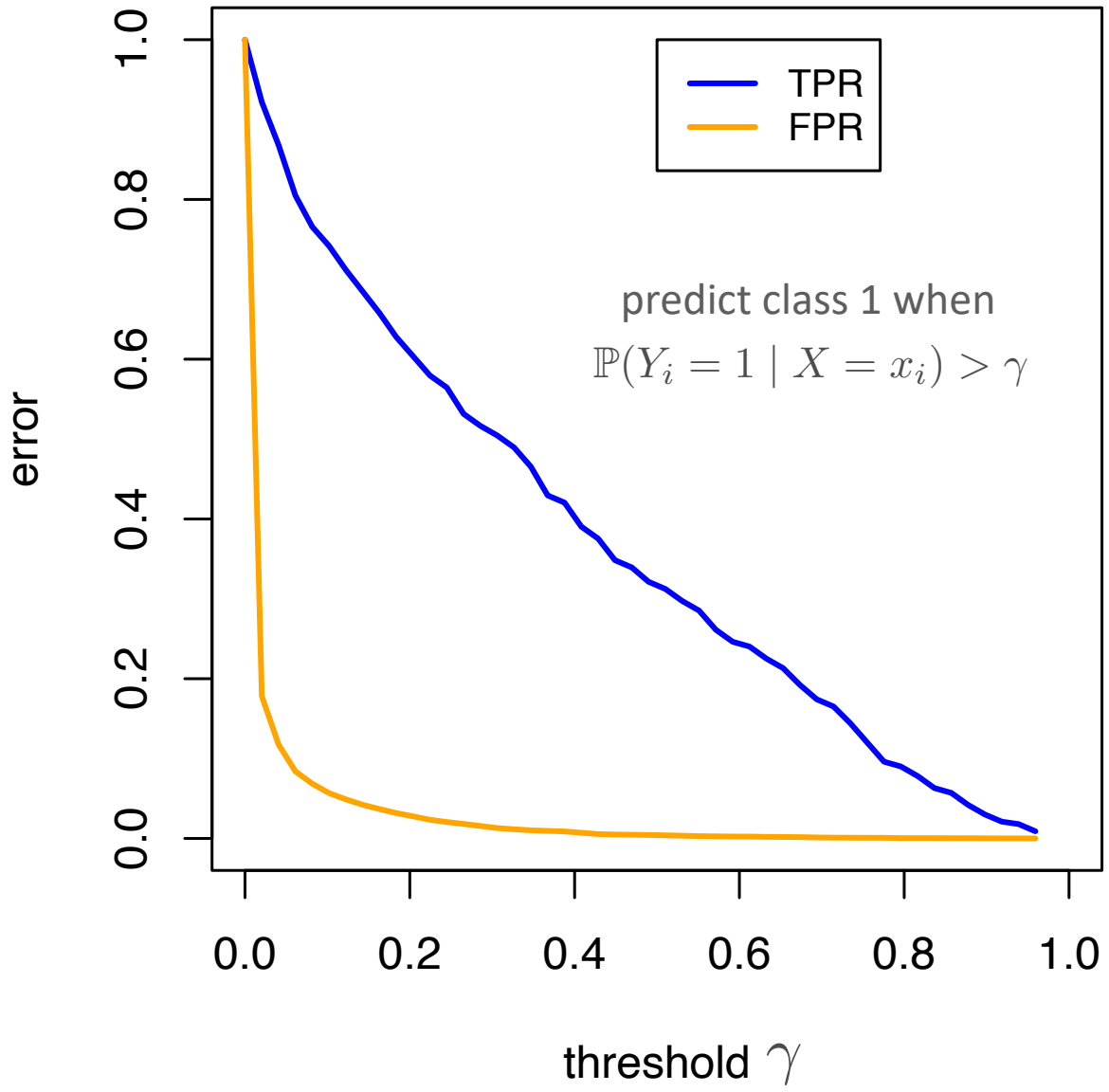
Much less impressive, right?

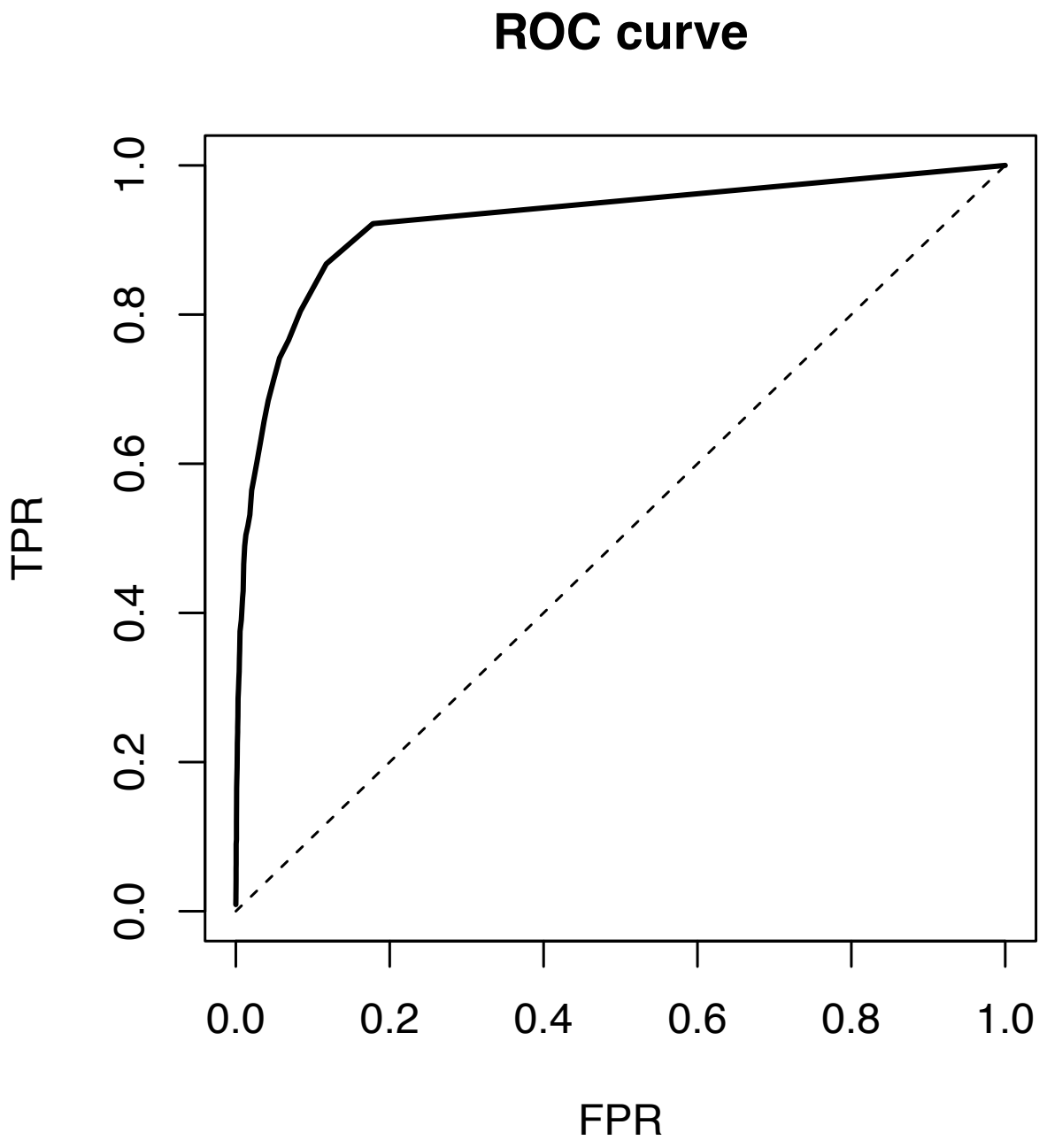

TRUE OBSERVATIONS
FALSE
TRUE
PREDICTED OBSERVATIONS
FALSE
TRUE
9627
228
40
105


TRUE OBSERVATIONS
FALSE
TRUE
PREDICTED OBSERVATIONS
FALSE
TRUE
True Negative
9627
False Negative
228
False Positive
40
True Positive
105


TRUE OBSERVATIONS
FALSE
TRUE
PREDICTED OBSERVATIONS
FALSE
TRUE
TN
FN
FP
TP


TRUE OBSERVATIONS
FALSE
TRUE
PREDICTED OBSERVATIONS
FALSE
TRUE
TN
FN
FP
TP
False Positive Rate
True Positive Rate


TPR
FPR
predict class 1 when
$\mathbb{P}(Y_i = 1 \mid X = x_i) > \gamma$
error
threshold $\gamma$
0.0
0.2
0.4
0.6
0.8
1.0


ROC curve
TPR
FPR
0.0
0.2
0.4
0.6
0.8
1.0

## CARACTÉRISTIQUES DE PERFORMANCE

### Sensibilité, spécificité et précision clinique

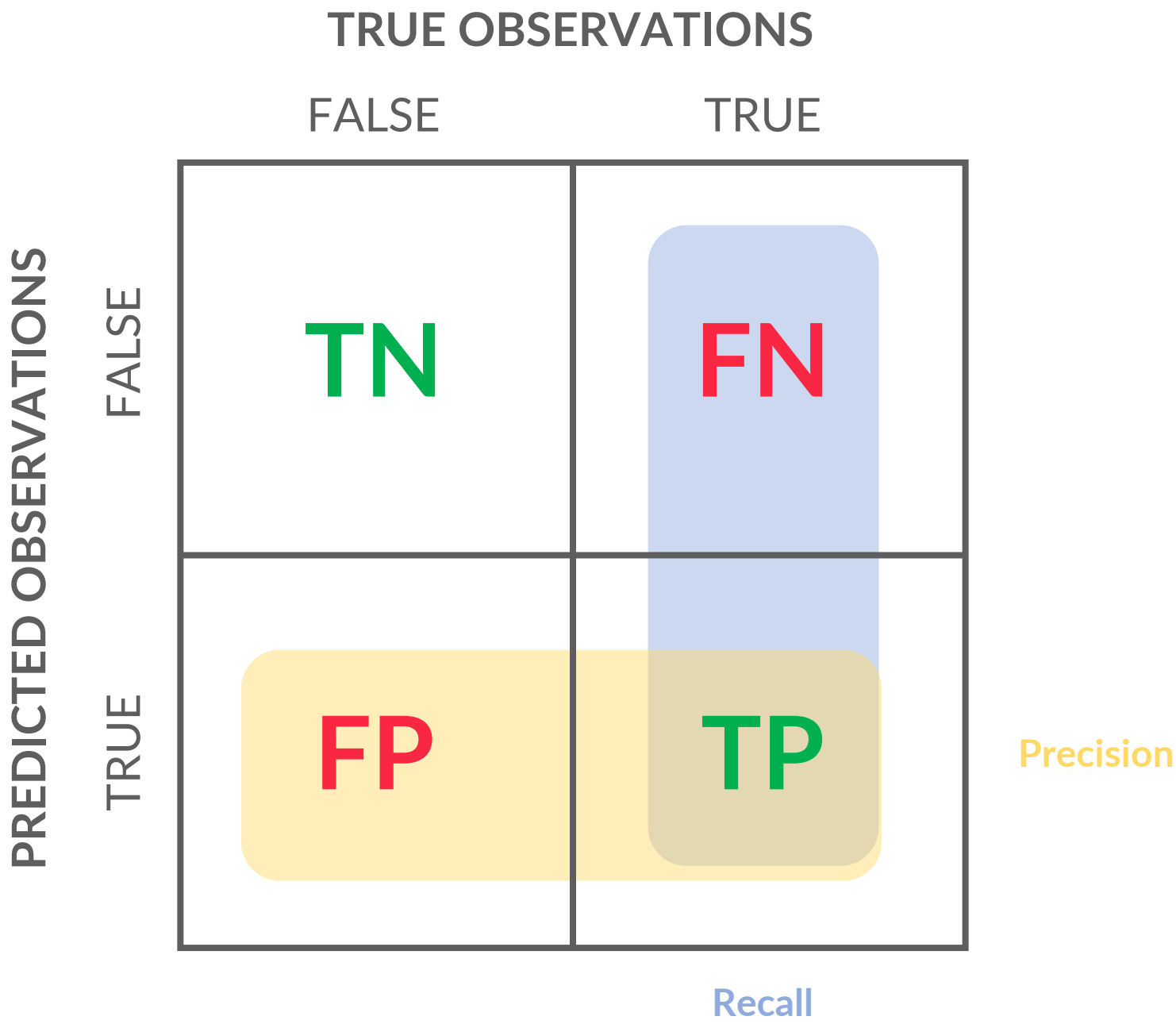
Les performances du test rapide de détection de l'antigène du SRAS-CoV-2 ont été établies à partir de 605 écouvillons nasaux prélevés chez des personnes symptomatiques suspectes de COVID-19.
Les résultats indiquent que la sensibilité relative et la spécificité relative sont les suivantes :

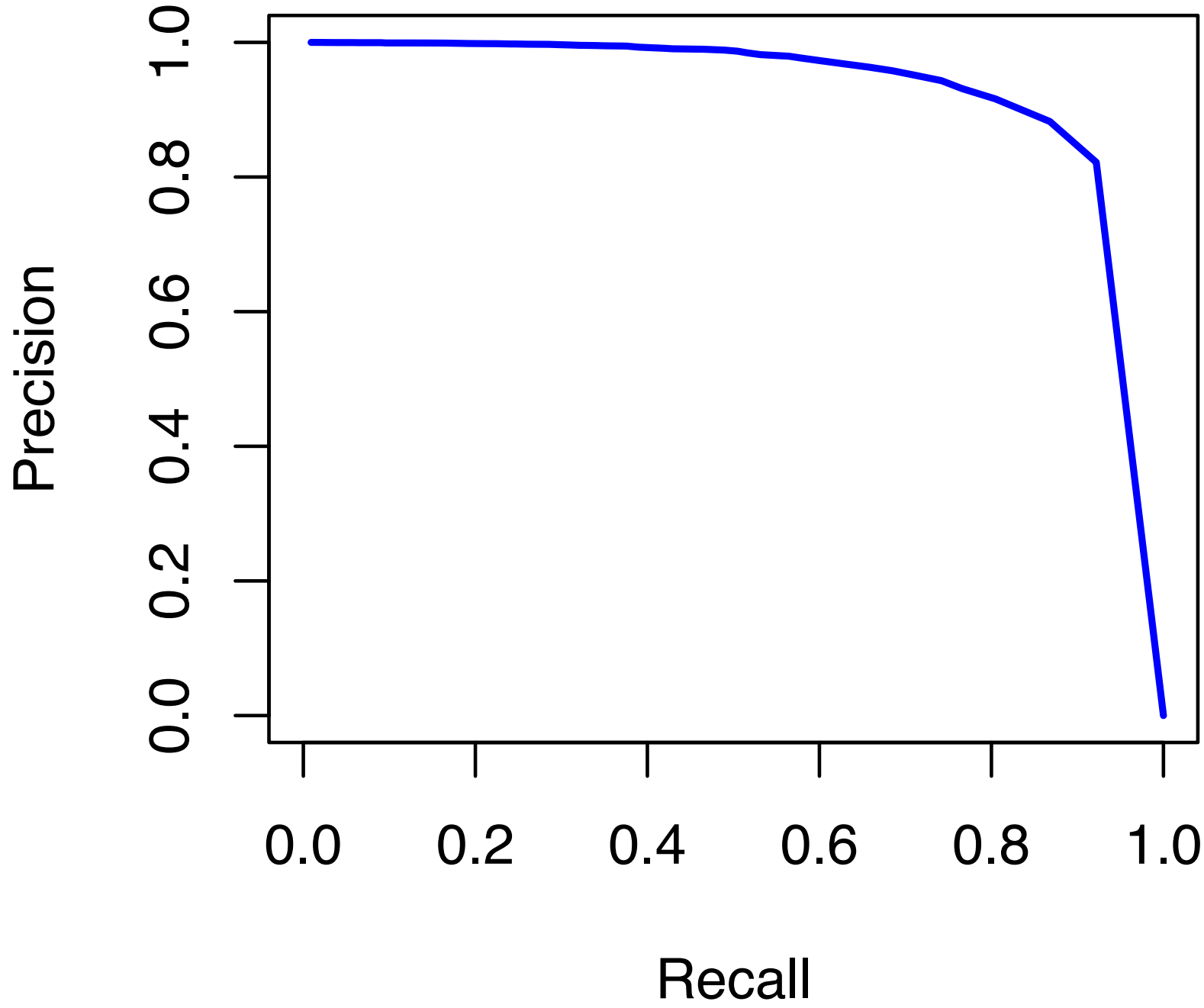
**Performance clinique du test rapide de détection de l'antigène SRAS-CoV-2**

| **Méthode** | | **RT-PCR** | | **Résultats totaux** |
|---|---|---|---|---|
| Test rapide de détection de l'antigène SARS-CoV-2 | **Résultats** | Négatif | Positif | |
| | Négatif | 433 | 5 | 438 |
| | Positif | 2 | 165 | 167 |
| **Résultats totaux** | | 435 | 170 | 605 |

Sensibilité relative : 97,1 % (93,1 %-98,9 %)*
Spécificité relative : 99,5% (98,2 %-99,9 %)*
Précision : 98,8 % (97,6 %-99,5 %)*
*95 % Intervalles de confiance

La stratification des échantillons positifs après l'apparition des symptômes entre 0 et 3 jours a un pourcentage de concordance positive (PPA) de 98,8 % (n=81) et entre 4 et 7 jours a un PPA de 96,8 % (n=62).
Les échantillons positifs avec une valeur Ct ≤33 ont un pourcentage de concordance positive (PPA) plus élevé de 98,7 % (n=153).


TRUE OBSERVATIONS
FALSE
TRUE
PREDICTED OBSERVATIONS
FALSE
TRUE
TN
FN
FP
TP
Precision
Recall


Precision
0.0
0.2
0.4
0.6
0.8
1.0
0.0
0.2
0.4
0.6
0.8
1.0
Recall

2021-2022
Information et complexité

2020-2021
Représentations parcimonieuses

2019-2020
Modèles multi-échelles et réseaux de neurones convolutifs

2018-2019
L'apprentissage par réseaux de neurones profonds

2017-2018
L'apprentissage face à la malédiction de la grande dimension

L'apprentissage face à la malédiction de la grande dimension
Présentation

17 janvier 2018 ~ 09:30 ~ 11:0...
Cours
Cartographie des sciences des données
Stéphane Mallat

17 janvier 2018 ~ 11:15 ~ 12:3...
Séminaire
Présentation des challenges 2018 (1)
Stéphane Mallat

24 janvier 2018 ~ 09:30 ~ 11:0...
Cours
Compromis Biais-Complexité
Stéphane Mallat

24 janvier 2018 ~ 11:15 ~ 12:3...
Séminaire
Présentation des challenges 2018 (2)
Stéphane Mallat

# L'apprentissage face à la malédiction de la grande dimension

Accéder aux notes de cours

Les sciences des données ont pour objectif « d'extraire de la connaissance » de données numériques, avec des algorithmes. Les applications sont considérables, pour stocker, analyser et valoriser les masses de données : images, sons, textes, mesures physiques ou données d'Internet. On distingue deux types de problèmes : la prédiction et la modélisation. Les prédictions sont faites par des algorithmes d'apprentissage statistique, qui sont à l'origine du renouveau de l'intelligence artificielle. Un modèle décrit la variabilité des données et permet d'en générer des nouvelles. Les mathématiques ont ici pour but de comprendre sous quelles conditions il est possible d'apprendre et donc de généraliser, ou de construire des modèles, tandis que l'informatique a pour objectif de développer des algorithmes qui résolvent ces problèmes.

Le premier cours de la chaire pose le cadre mathématique et algorithmique de ce domaine, en dégageant les questions et techniques importantes pour l'apprentissage. La difficulté principale de la prédiction ou de la modélisation vient du grand nombre de variables des données, souvent plus d'un million, à l'instar du nombre de pixels d'une image. Cette grande dimension génère une explosion combinatoire des possibilités de prédiction ou de modélisation. On fait face à cette malédiction de la grande dimension avec des algorithmes qui utilisent de l'information *a priori* sur certaines régularités du problème. Le cours introduit des outils mathématiques et

https://www.college-de-france.fr/site/stephane-mallat/course-2017-2018.htm

# Challenge Data

ENS ÉCOLE NORMALE SUPÉRIEURE | PSL RESEARCH UNIVERSITY PARIS | INSTITUT Louis Bachelier | COLLÈGE DE FRANCE 1530

https://challengedata.ens.fr/