



H₂O.ai

<http://www.h2o.ai>

Machine Learning

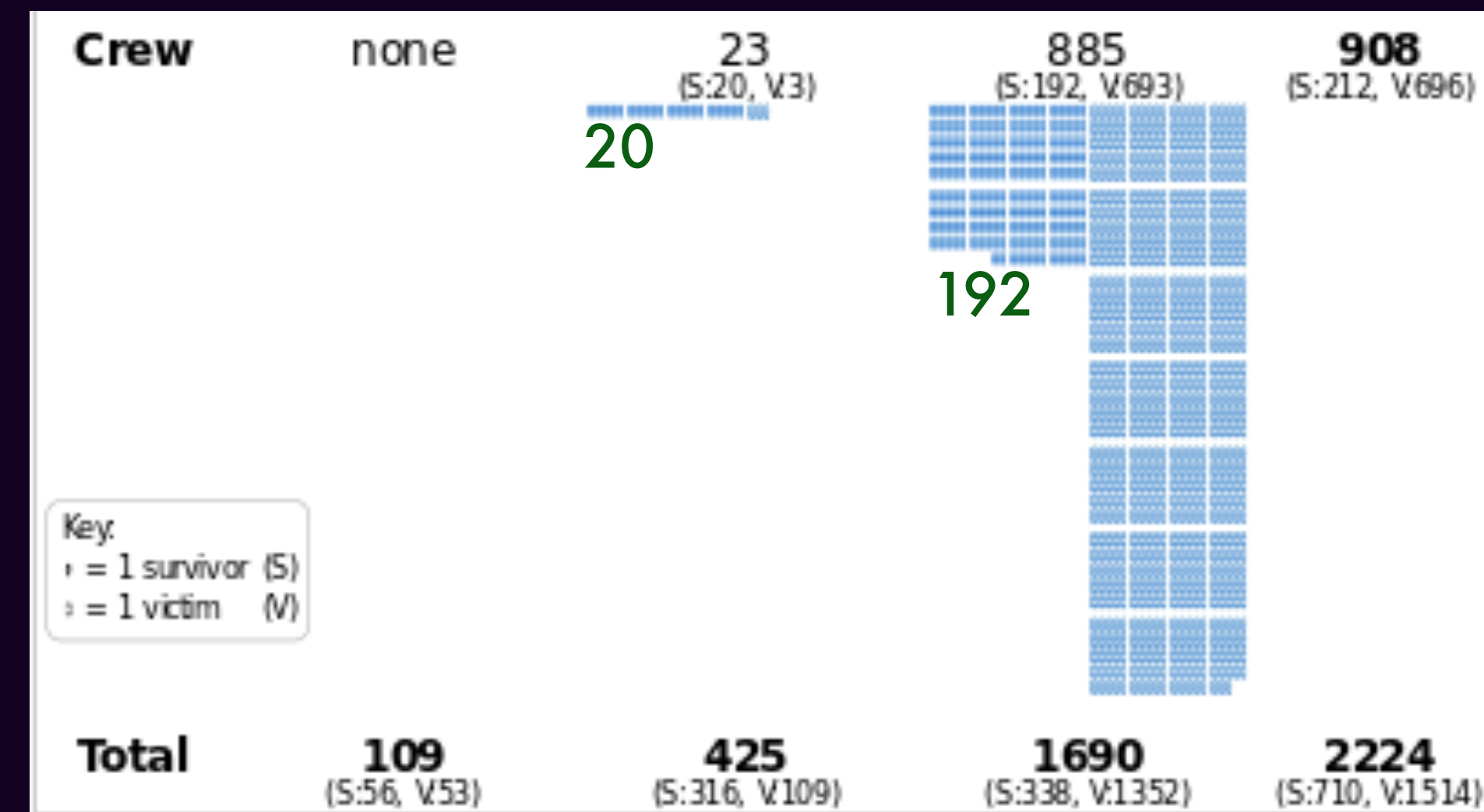
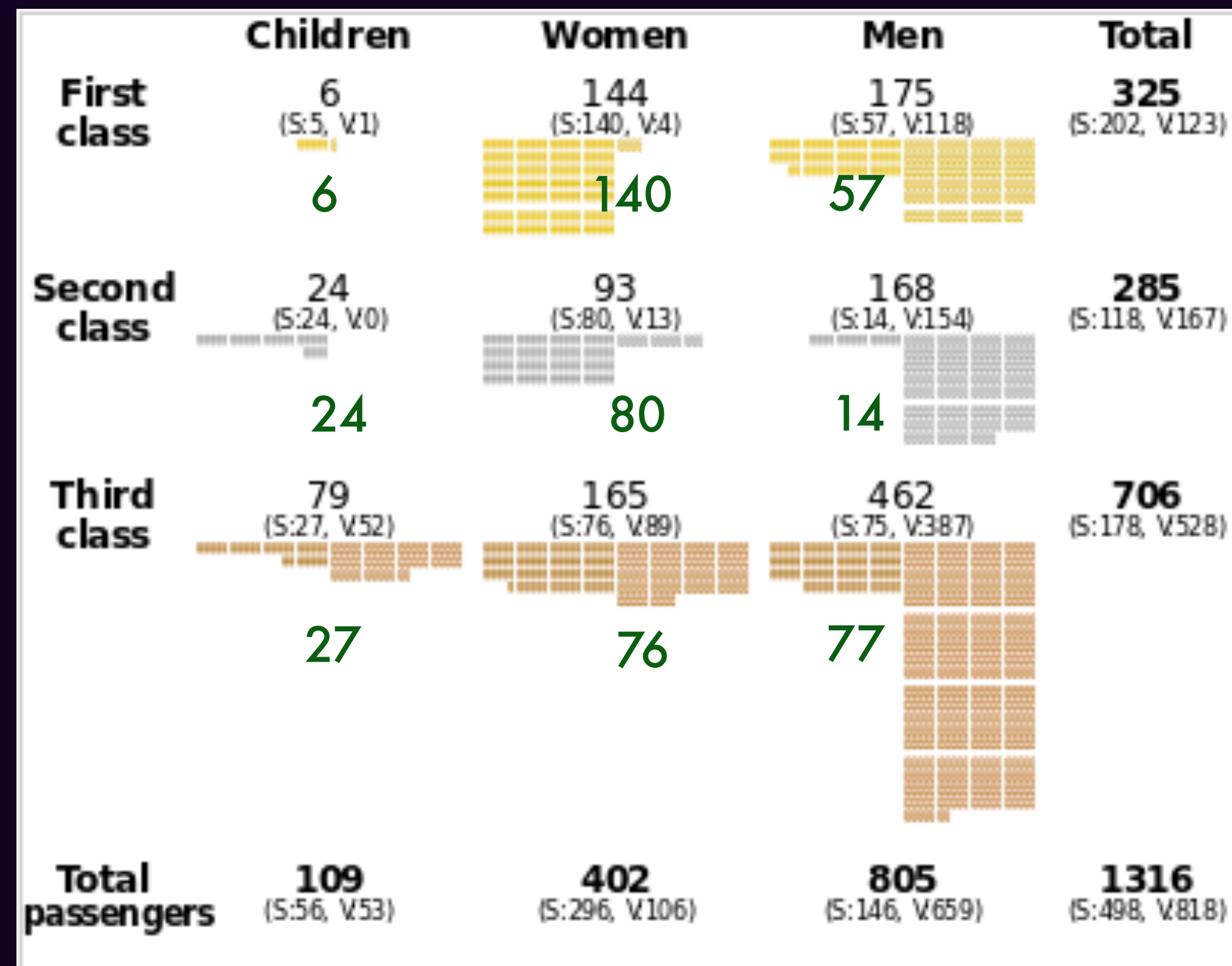
Devoxx France 2016
Claude Falguière
@cfalguiere

H2O

Machine learning et modèles prédictifs

importer et *parser* des sources
manipuler les *dataframes*
ajuster un *modèle prédictif*
calculer une *prédiction*
sauver les modèles et les réutiliser

Données Titanic



Source Wikipedia

Données
connues



Ajustement d'un
Modèle Prédictif



Nouvelles
données



Calcul d'une
Prédiction

Données
connues

Jeu d'entraînement



Entraînement d'un
modèle prédictif



Modèle
potentiel



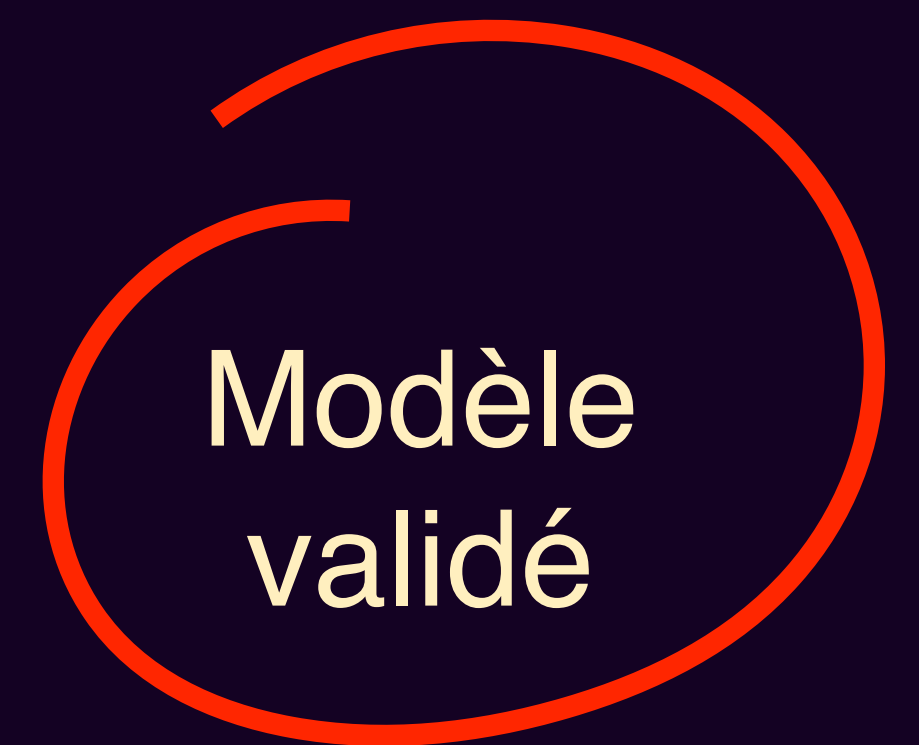
Jeu de validation



Prédiction et
vérification



Modèle
validé



Nouvelles
données



***Modèle
validé***

+

Jeu de travail

->

Prédiction

Données

Class (x_1)	Age (x_2)	Sex (x_3)	Survived (x_4)
1	42	1	1
3	26	1	0
2	33	2	1
2	6	1	1

Modèle

Par exemple

déterminer p_0, p_1, p_2, p_3 pour que

$$p_0 + p_1x_1 + p_2x_2 + p_3x_3 = x_4$$

Machine Learning

- ★ fixer p_0, p_1, p_2, p_3
- ★ calculer x'_4
- ★ évaluer l'écart avec x_4
(Loss function)
- ★ Adapter p_0, p_1, p_2, p_3 pour
minimiser l'écart
- ★ itérer

Données

Modèle

Prédiction

C	A	Sx	Su
1	42	1	1
3	26	1	0
2	33	2	1
2	6	1	1



p_0
 p_1
 p_2
 p_3



Su'
1
0
1
1

Données

C	A	Sx	Su
1	42	1	1
3	26	1	0
2	33	2	1
2	6	1	1

Modèle

p_0
 p_1
 p_2
 p_3
 p_0
 p_1
 p_2
 p_3
 p_0
 p_1
 p_2
 p_3

p_0	p_0	p_3
p_0	p_0	p_3
p_0	p_0	p_3

Prédiction

p_0
 p_1
 p_2
 p_3

Su'
1
0
1
1

Réseau de neurones

H2O Flow

L'ensemble des cellules constitue un Flow

The screenshot shows the H2O Flow web interface. At the top, there is a navigation bar with tabs: Flow, Cell, Data, Model, Score, Admin, and Help. The 'Data' tab is currently selected. Below the navigation bar, the main area is titled 'Untitled Flow'. On the left side of this area, there is a toolbar with icons for file operations (new, open, save, etc.) and a list of cells. One cell is highlighted with a yellow background and labeled 'CS'. To the right of the cell list, there is a text input field labeled 'Expression...'. A red arrow points from the 'Expression...' field to a grey callout box. Another red arrow points from the 'Admin' tab in the navigation bar to a second grey callout box. A third red arrow points from the 'Data' tab to a third grey callout box.

Menu pour générer les cellules et autres actions

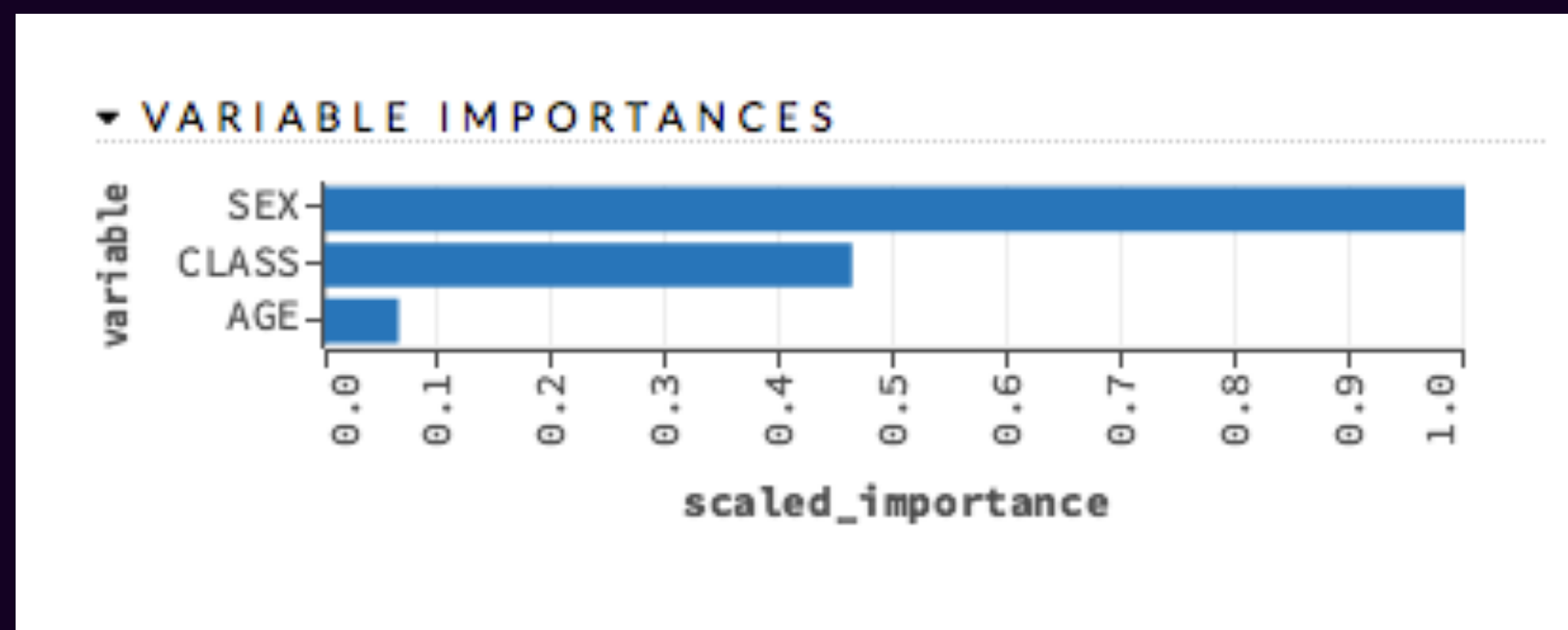
Les cellules contiennent

- ★ Une instruction H2O
- ★ ou du texte
- ★ ou du code R ou Python

Démo

Jeu préparé

classe, sexe, adulte/enfant
avec équipage

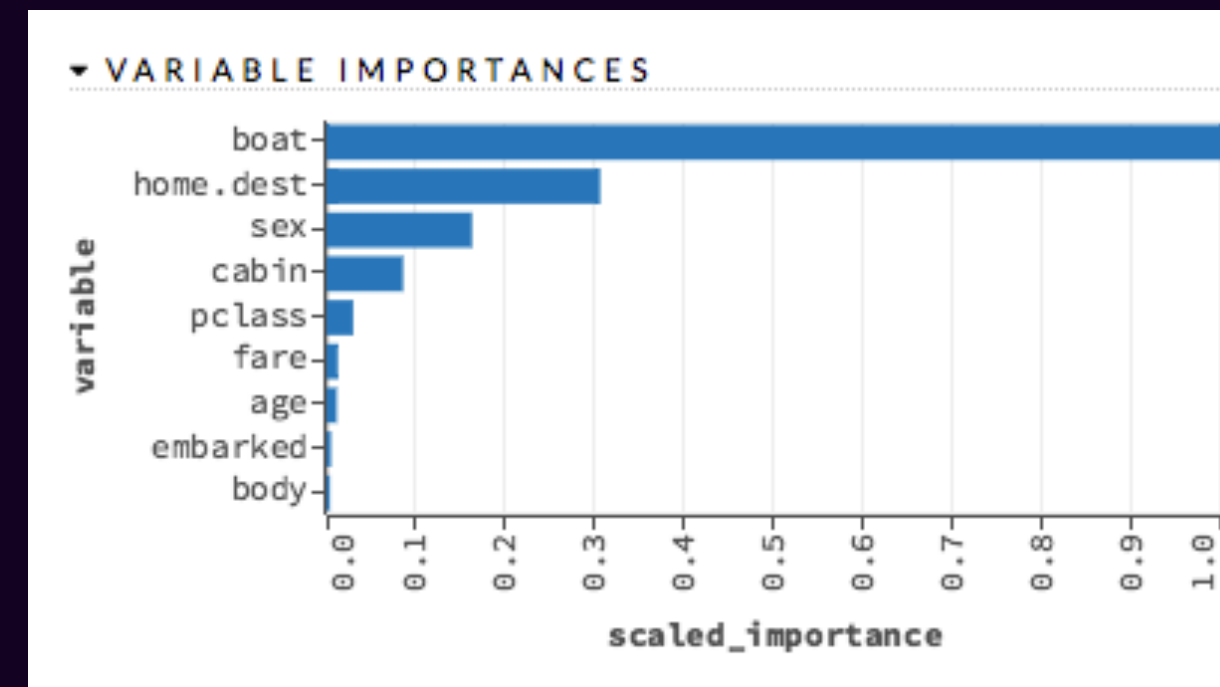


Distributed Random Forest

```
predictions •
MSE 0.155393
r2 0.289416
mean_residual_deviance 0.155393
```

Jeu détaillé

avec en plus âge, poids, bateau, destination
passagers seulement



Distributed Random Forest

```
predictions •
MSE 0.086517
r2 0.633509
logloss 0.398138
AUC 0.951124
Gini 0.902247
```

Deep Learning

```
predictions •
MSE 0.024767
r2 0.895086
logloss 0.101934
AUC 0.993252
Gini 0.986504
```


Algorithmes dans H2O

Supervised learning

On a un jeu de données dont on connaît les réponses et on veut une formule pour estimer la réponse sur d'autres jeux de données

Generalized Linear Models (GLM): Provides flexible generalization of ordinary linear regression for response variables with error distribution models other than a Gaussian (normal) distribution. GLM unifies various other statistical models, including Poisson, linear, logistic, and others when using ℓ_1 and ℓ_2 regularization.

Distributed Random Forest: Averages multiple decision trees, each created on different random samples of rows and columns. It is easy to use, non-linear, and provides feedback on the importance of each predictor in the model, making it one of the most robust algorithms for noisy data.

Gradient Boosting (GBM): Produces a prediction model in the form of an ensemble of weak prediction models. It builds the model in a stage-wise fashion and is generalized by allowing an arbitrary differentiable loss function. It is one of the most powerful methods available today.

Deep Learning: Models high-level abstractions in data by using non-linear transformations in a layer-by-layer method. Deep learning is an example of supervised learning, which can use unlabeled data that other algorithms cannot.

Naïve Bayes: Generates a probabilistic classifier that assumes the value of a particular feature is unrelated to the presence or absence of any other feature, given the class variable. It is often used in text categorization.

Algorithmes dans H2O

Unsupervised learning

On recherche une formule permettant de définir des groupes d'observations se ressemblant ou suivant le même pattern.

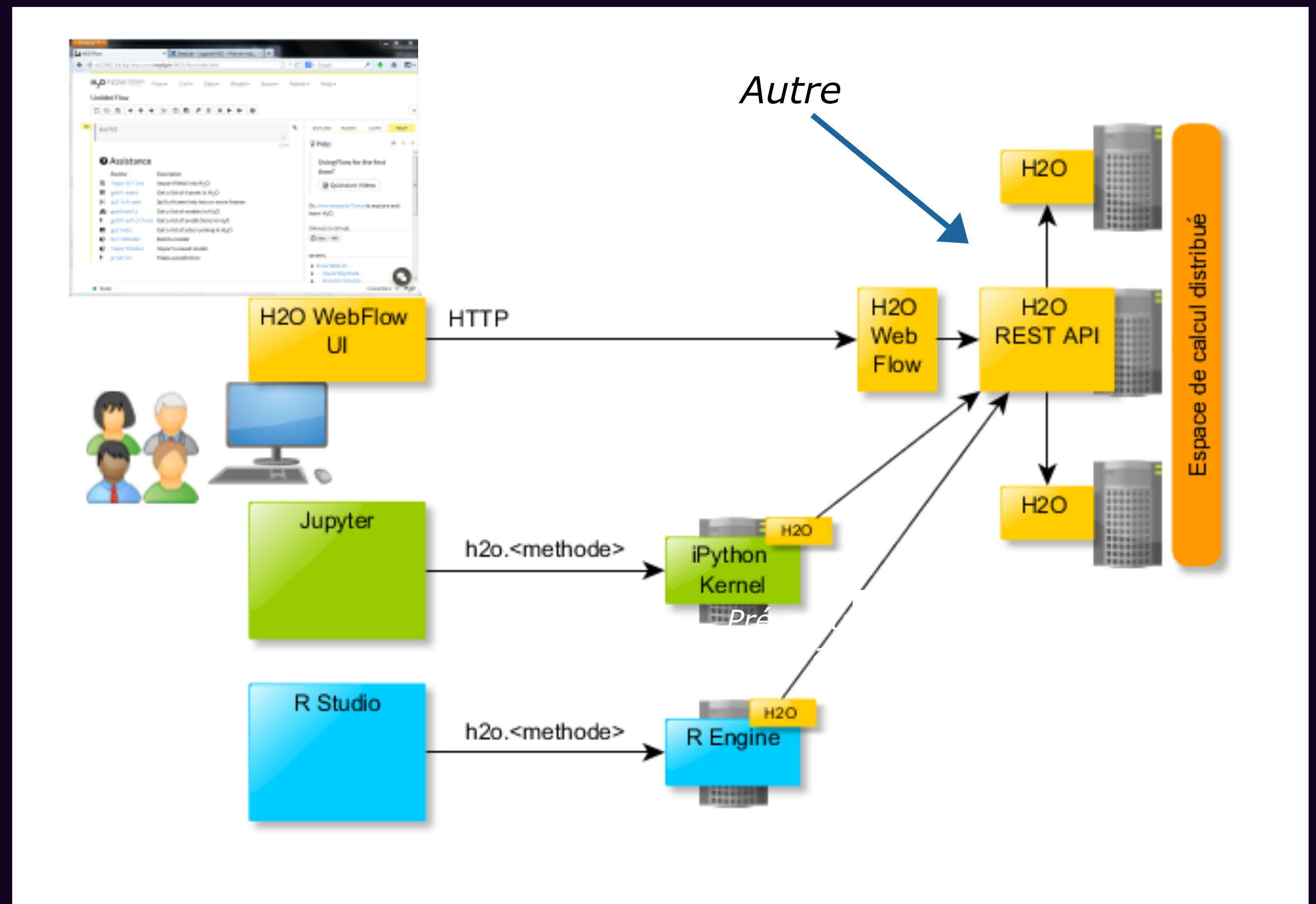
K-Means: Reveals groups or clusters of data points for segmentation. It clusters observations into k -number of points with the nearest mean.

Principal Component Analysis (PCA): The algorithm is carried out on a set of possibly collinear features and performs a transformation to produce a new set of uncorrelated features.

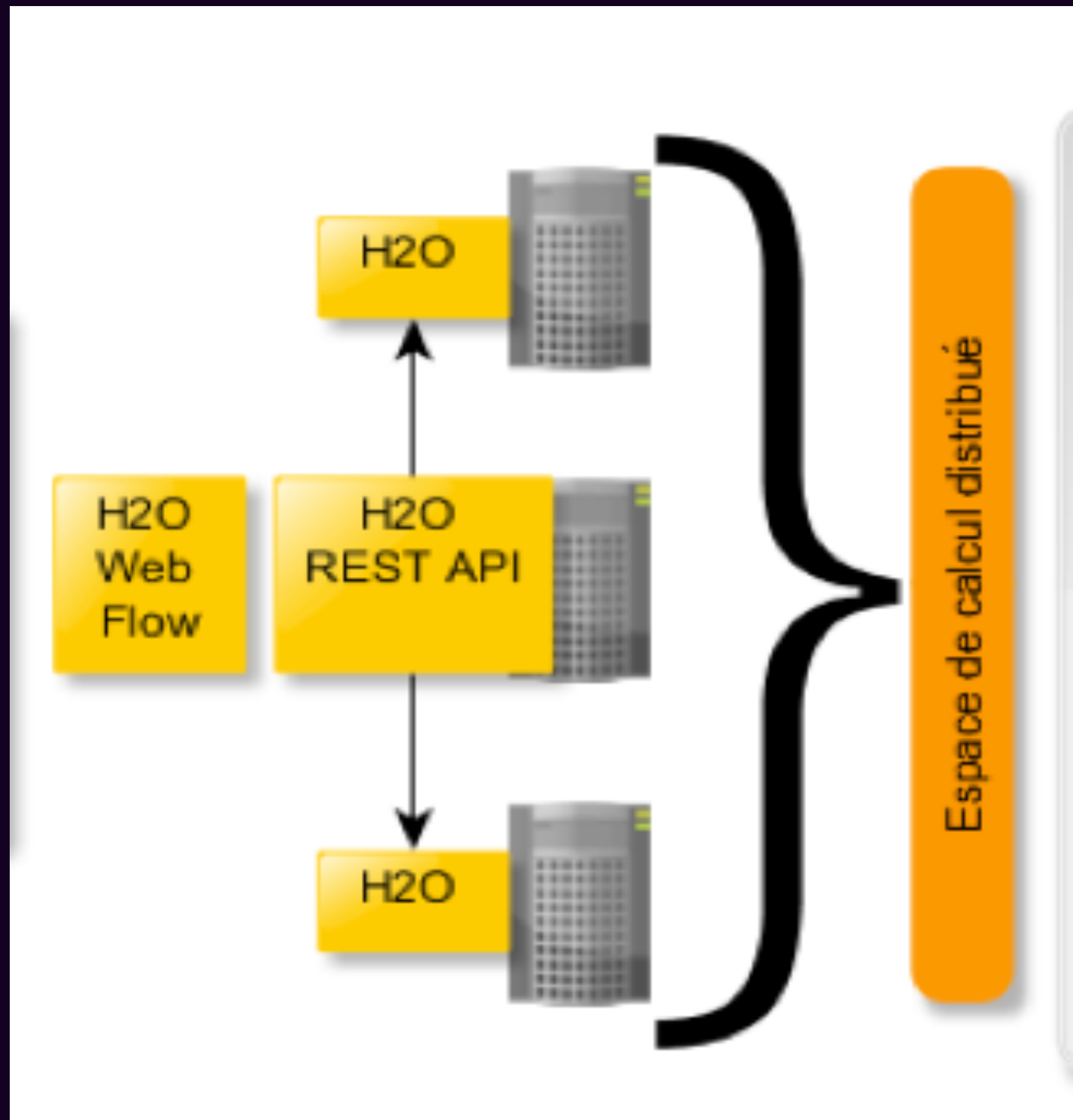
Anomaly Detection: Identifies the outliers in your data by invoking the deep learning autoencoder, a powerful pattern recognition model.

Architectures

- ★ Flow
- ★ API REST
- ★ Languages Bindings R et Python
- ★ Tout en mémoire
- ★ Data = 4 x RAM



Clusters



- ★ Peut constituer son propre cluster
- ★ Ou fonctionner sur des clusters Spark ou Hadoop existants
- ★ Versions spécifiques Hadoop
- ★ Sparkling Water pour Spark
- ★ Même mode de fonctionnement

APIs

```
In [92]: from h2o.estimators.deeplearning import H2ODeepLearningEstimator

model = H2ODeepLearningEstimator()           # default DL setup

# ajuste les paramètres du modèle - attention les colonnes sont base 0
model.train(x=train.names[1:4], y=train.names[5],
            training_frame=train, validation_frame=valid) # pass a validation frame in

model                                         # display the model summary

model.show()                                # equivalent to the above
```

```
In [105]: prediction = model.predict(test2)
#model.model_performance(test2)
```


Conclusion

Conclusion

***Prise en
main facile***

H2O Flow est attractif
Utilisation très facile du cluster Hadoop
Essayer des modèles très rapidement

***Compléter
avec l'API***

Combiner les modèles
Intégrer la préparation des données
Industrialiser

Merci
Avez vous des
questions ?