



Dataiku

Data Science Studio

CHERIF DIALLO
CAMILLE SCHIFFLER
ISSAM CHARNI

SOMMAIRE

- **Présentation**
- **Installation**
- **Fonctionnalités**
- **Méthodes Statistiques**
- **Evaluation**
- **Démonstration**

SOMMAIRE

- **Présentation**
- **Installation**
- **Fonctionnalités**
- **Méthodes Statistiques**
- **Evaluation**
- **Démonstration**

PRESENTATION

- Création : 2013
- Logiciel : Data Science Studio
- Capgemini, Blablacar, Chronopost, Venteprivée ...
- Objectif 2019 : 10 millions de CA



PRESENTATION

- Version gratuite et version payante

	Free	Entreprise
Visual and code lab and transformations	●	●
Standard connectors (Files, MySQL, PostgreSQL)	●	●
Big data connectors and engines (Analytics databases, Hadoop, Spark)		●
Production (automation, monitoring, real-time API)		●
Multi-user (collaboration, security, sharing)		●
Complete feature list		TRY FREE FOR 2 WEEKS
		Then rollback to Community Edition

SOMMAIRE

- **Présentation**
- **Installation**
- **Fonctionnalités**
- **Méthodes Statistiques**
- **Evaluation**
- **Démonstration**

INSTALLATION

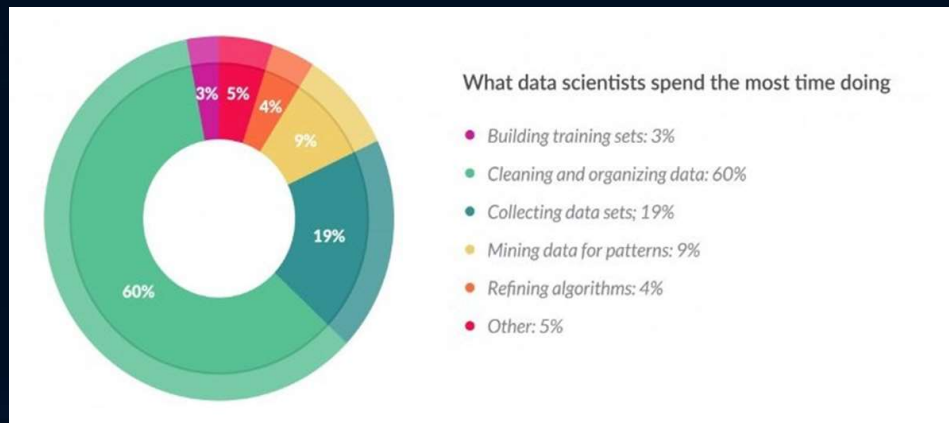
- Non hébergé
 - ✓ Linux
 - ✓ Mac OS
 - ✓ Windows
 - VirtualBox / Vmware
- Cloud
 - ✓ Microsoft Azure
 - ✓ Amazon Ec2

SOMMAIRE

- **Présentation**
- **Installation**
- **Fonctionnalités**
- **Méthodes Statistiques**
- **Critiques**
- **Démonstration**

FONCTIONNALITES

- Processus d'analyse



PREPARE &
ENRICH



MODEL &
PREDICT



DEPLOY & RUN

FONCTIONNALITES

- Modes opératoires



FONCTIONNALITES



- Workflow
- Dataset
- Analyze
- Dashboard (payant)

FONCTIONNALITES



R	Python	SQL	Javascript	SparkR
Hive	Pig	Shell	PySpark	SparkSQL

SOMMAIRE

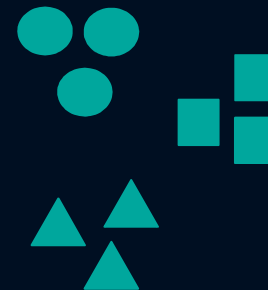
- Présentation
- Installation
- Fonctionnalités
- Méthodes Statistiques
- Evaluation
- Démonstration

METHODES STATISTIQUES

PREDICTION



CLUSTERING



METHODES STATISTIQUES

PREDICTION	CLUSTERING
Régression linéaire	Kmeans
Régression logistique	Méthode de WARD
Arbre de décision	DBScan
Forêt aléatoire	Mélange de mélanges Gaussiens (GMM)
Gradient Boosting	Spectral clustering

SOMMAIRE

- Présentation
- Installation
- Fonctionnalités
- Méthodes Statistiques
- Evaluation
- Démonstration

EVALUATION

- **Avantages**

- Sources de données
- Simple d'utilisation
- Permet le travail collaboratif (*Version payante*)
- Possibilité de programmer (Python, R ...)
- Possibilité de faire une analyse complète

- **Inconvénients**

- Nécessite des ressources importantes (Microsoft)
- Programmation
- Accent sur la prédiction, clustering (faible)
- Version payante

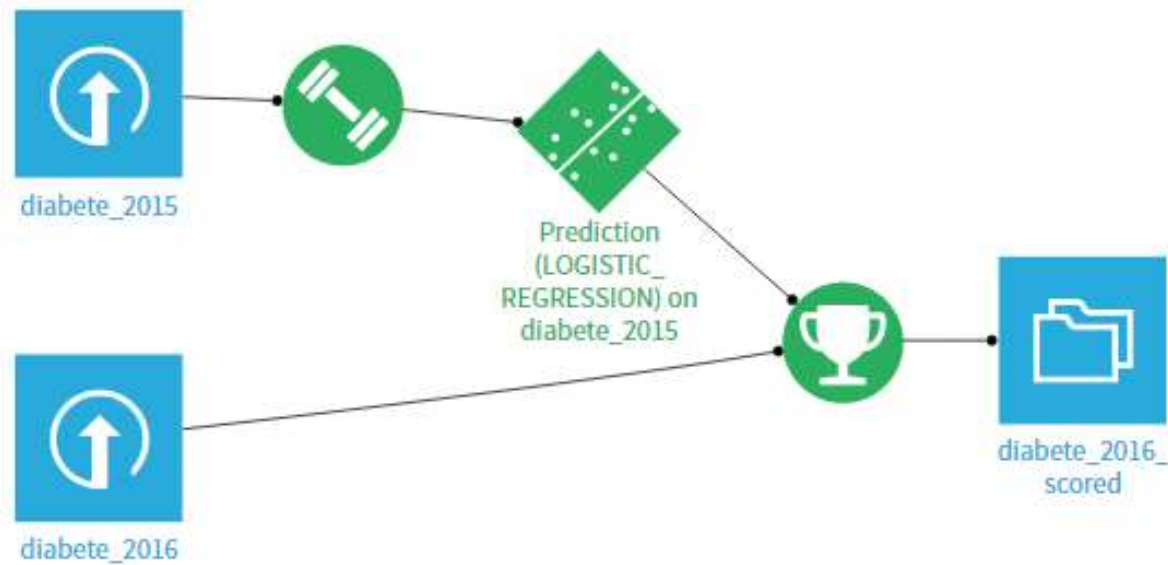
EVALUATION

- Retour d'expérience
 - Documentations, tutoriels
 - Support réactif
 - Lenteur
 - Utilisation

SOMMAIRE

- **Présentation**
- **Installation**
- **Fonctionnalités**
- **Méthodes Statistiques**
- **Evaluation**
- **Démonstration**

DEMONSTRATION



A VOTRE ECOUTE



REFERENCES

- www.Dataiku.com
- http://www.decideo.fr/Dataiku-veut-seduire-les-equipes-d-analystes-de-donnees_a8787.html
- http://www.lesechos.fr/20/02/2014/LesEchos/21631-098-ECH_dataiku-lit-l-avenir-dans-les-donnees.htm
- Youtube (Dataiku)