

From theoretical physics to data science consulting

Nicolas Thiébaut

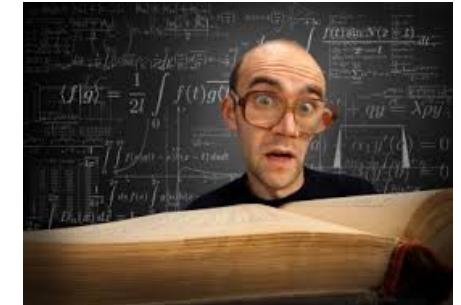
From science to data science

Centre de Recherches Interdisciplinaire

31/03/2017

What is this presentation about?

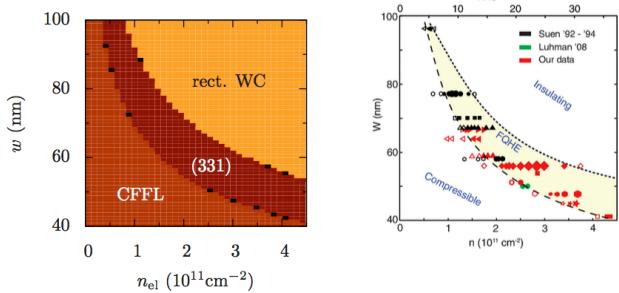
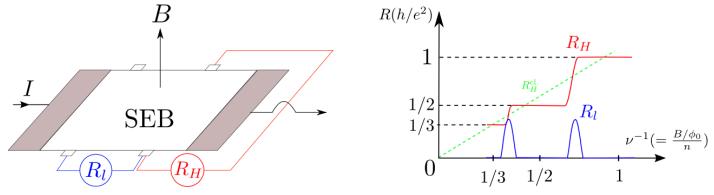
- Curriculum and transition to data science
 - Early years and PhD in physics
 - Reorientation thoughts
 - Learning data science
- Two years as a data science consultant
 - Experiences
 - Typical day
- Data scientist skill set
- Evolution of data science



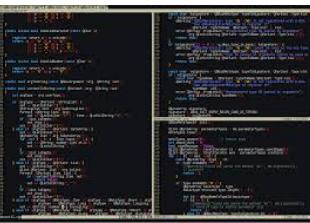


- Taste for physics, among many other topics!
- Magistère Fundamental Physics (Saclay University)
- Master 2 Condensed Matter Physics (ICFP)
- PhD: Theoretical Physics (fractional quantum Hall effect and Wigner crystals)

PhD: Fractional quantum Hall effect in the bilayer and the wide quantum well



- Deals with Fractional Quantum Hall effect and Wigner crystals (strongly correlated systems)
- 2 PhD advisors:
 - Nicolas Regnault (ENS/Princeton) : numerics
 - Mark Goerbig (LPS at Saclay): analytics
- Summer schools (Les Houches, Cargese)), conferences (Chamonix), talks (Florida), lab visit (Indiana)
- Opening experiences: teaching, courses (journalism)



ANNEXE D: CORRECTIONS À L'ÉNERGIE DU CRISTAL DE WIGNER DANS AU RECOUVRISSEMENT DES ÉTATS VERTUS

La densité de ce test est donnée par

$$\rho(\mathbf{R}) = \left(\frac{\pi^2}{3} \ln(n/\mu_0) \right)^{1/2} \int_{\mathbb{R}^3} \Psi_{\mathbf{R}}(\mathbf{r}, \mathbf{R}, \mathbf{R}') d\mathbf{r}$$

$$= \frac{1}{1 - e^{-\frac{1}{2} \ln(n/\mu_0)}} \int_{\mathbb{R}^3} \left[e^{-\frac{1}{2} \ln(n/\mu_0)} e^{-\frac{1}{2} \mathbf{R}^2} e^{-\frac{1}{2} \mathbf{R}'^2} - 2\cos\left(\frac{1}{2}\mathbf{R} \cdot \mathbf{R}'\right) e^{-\frac{1}{2} \mathbf{R}^2} e^{-\frac{1}{2} \mathbf{R}'^2} \right] d\mathbf{r}$$

$$= \frac{1}{2\pi \left(1 - e^{-\frac{1}{2} \ln(n/\mu_0)} \right)} \left[e^{-(\mathbf{R} + \mathbf{R}')^2} + e^{-(\mathbf{R} - \mathbf{R}')^2} - 2\cos\left(\frac{1}{2}\mathbf{R} \cdot \mathbf{R}'\right) \right]$$

$$= \frac{1}{2\pi \left(1 - e^{-\frac{1}{2} \ln(n/\mu_0)} \right)} \left[2e^{-(\mathbf{R} + \mathbf{R}')^2} - 2\cos\left(\frac{1}{2}\mathbf{R} \cdot \mathbf{R}'\right) \right]$$

Alors en remettant

$$\rho_0(\mathbf{R}) = -\frac{1}{1 - e^{-\frac{1}{2} \ln(n/\mu_0)}} \left[e^{-\frac{1}{2} \mathbf{R}^2} + e^{-\frac{1}{2} \mathbf{R}'^2} \right] \quad (D.8)$$

et

$$\delta\rho(\mathbf{R}) = -\frac{1}{1 - e^{-\frac{1}{2} \ln(n/\mu_0)}} \cos\left(\frac{1}{2}\mathbf{R} \cdot \mathbf{R}'\right) e^{-\frac{1}{2} (\mathbf{R} + \mathbf{R}')^2} \quad (D.9)$$

On a $\rho(\mathbf{R}) = \rho_0(\mathbf{R}) + \delta\rho(\mathbf{R})$. Alors on calcule l'énergie d'interaction de cet état dans l'approximation de Hartree-Fock (exprimé à l'aide des équations précédentes)

$$\delta E[\mathbf{R}] = -\frac{1}{2} \int_{\mathbb{R}^3} \int_{\mathbb{R}^3} \rho_0(\mathbf{R}') \rho_0(\mathbf{R}) e^{-\frac{1}{2} (\mathbf{R} + \mathbf{R}')^2} d\mathbf{R}' d\mathbf{R} \quad (D.10)$$

$$= e^{-\frac{1}{2} \ln(n/\mu_0)} e^{-\frac{1}{2} \mathbf{R}^2} e^{-\frac{1}{2} \mathbf{R}'^2} \cos\left(\frac{1}{2}\mathbf{R} \cdot \mathbf{R}'\right) \quad (D.11)$$

$$\delta E[\mathbf{R}] = -\frac{1}{1 - e^{-\frac{1}{2} \ln(n/\mu_0)}} \int_{\mathbb{R}^3} \int_{\mathbb{R}^3} \delta\rho(\mathbf{R}') \delta\rho(\mathbf{R}) e^{-\frac{1}{2} (\mathbf{R} + \mathbf{R}')^2} d\mathbf{R}' d\mathbf{R} \quad (D.12)$$

$$= -2e^{-\frac{1}{2} \ln(n/\mu_0)} e^{-\frac{1}{2} \mathbf{R}^2} \cos\left(\frac{1}{2}\mathbf{R} \cdot \mathbf{R}'\right) \quad (D.13)$$

La polarisation à la force de deux électrons locaux en les positions arbitrées \mathbf{R}_1 et \mathbf{R}_2

$$\Psi_{\mathbf{R}_1, \mathbf{R}_2}(\mathbf{r}_1, \mathbf{r}_2) = \langle \Psi_{\mathbf{R}_1}(\mathbf{r}_1) | \Psi_{\mathbf{R}_2}(\mathbf{r}_2) \rangle \quad (D.14)$$

$$= \frac{1}{1 - e^{-\frac{1}{2} \ln(n/\mu_0)}} \left[e^{-\frac{1}{2} \mathbf{R}_1^2} e^{-\frac{1}{2} \mathbf{R}_2^2} e^{-\frac{1}{2} (\mathbf{R}_1 - \mathbf{R}_2)^2} \right] \quad (D.15)$$

et $\delta\mathbf{R} = (\mathbf{R}_1 - \mathbf{R}_2)$ est la distance entre les deux centres des deux électrons, donc

$$\rho_{\delta\mathbf{R}} = -\frac{1}{1 - e^{-\frac{1}{2} \ln(n/\mu_0)}} \int_{\mathbb{R}^3} \int_{\mathbb{R}^3} \delta\rho(\mathbf{R}') \delta\rho(\mathbf{R}) e^{-\frac{1}{2} (\mathbf{R} + \mathbf{R}')^2} d\mathbf{R}' d\mathbf{R} \quad (D.16)$$

Le $\delta\mathbf{R} = (\mathbf{R}_1 - \mathbf{R}_2)$ est le vecteur de la distance entre les deux centres des deux électrons. Cette densité est tracée sur la figure D.3 pour un état à deux électrons localisés en \mathbf{R}_1 et \mathbf{R}_2 .

La transformée de Fourier est donnée par $\hat{\rho} = \hat{\rho}_1 + \hat{\rho}_2$ où

$$\hat{\rho}_{\delta\mathbf{R}}(q) = \frac{1}{1 - e^{-\frac{1}{2} \ln(n/\mu_0)}} \sum_{\mathbf{k}_1, \mathbf{k}_2} e^{i\mathbf{q} \cdot (\mathbf{k}_1 - \mathbf{k}_2)} \quad (D.17)$$

Reorientation



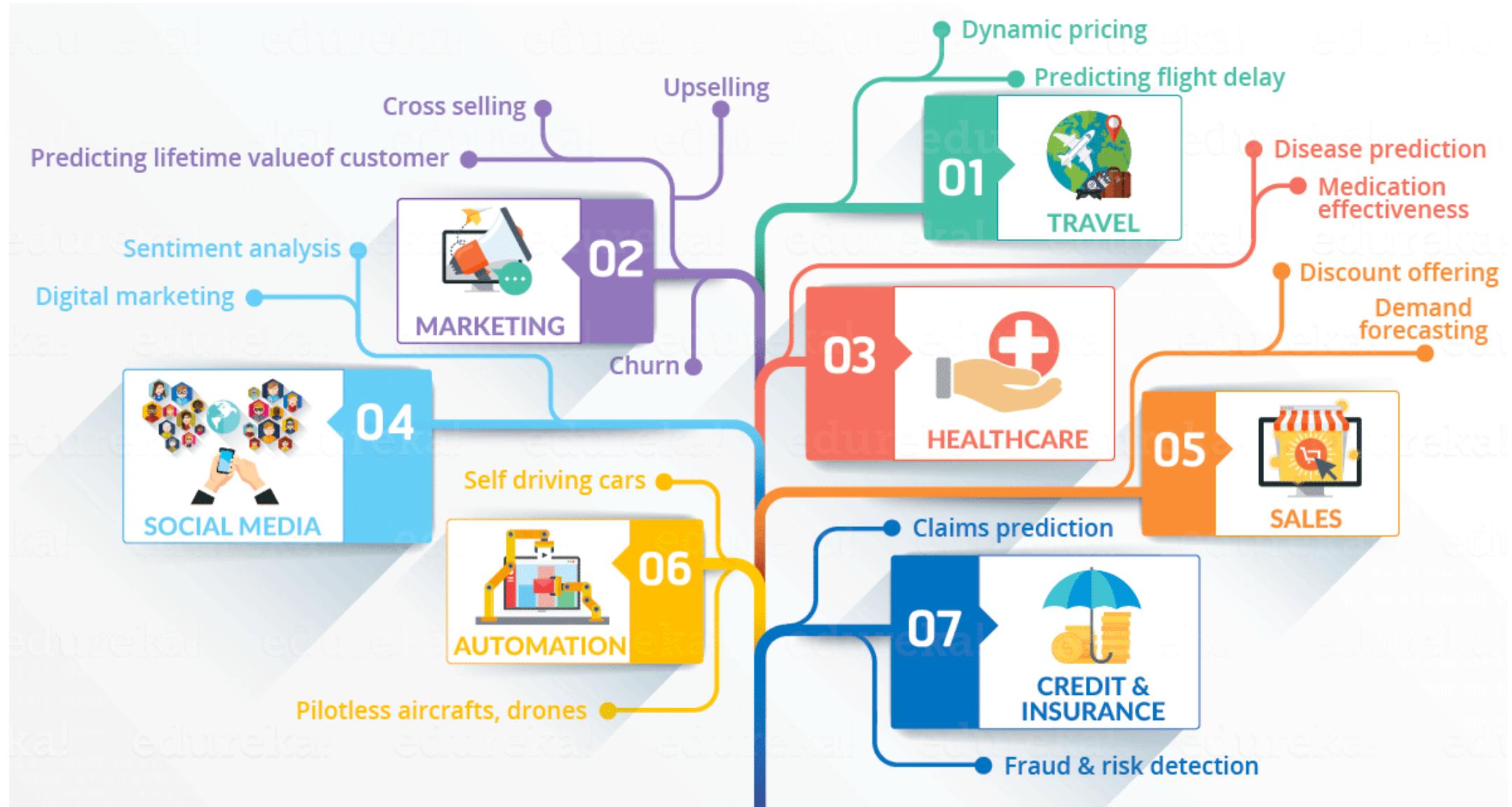
- Online courses (MOOCs):
 - Introduction to Machine Learning by Andrew NG (Coursera)
 - Les cours d'Hugo Larochelle de l'université de Sherbrooke (Youtube)
 - Cours d'Udacity
- Books:
 - The Elements of Statistical Learning by Trevor Hastie
 - Advanced Machine Learning by Hilary Mason
- Online data science competition
 - Datascience.net
 - Kaggle

2 years as a data science consultant

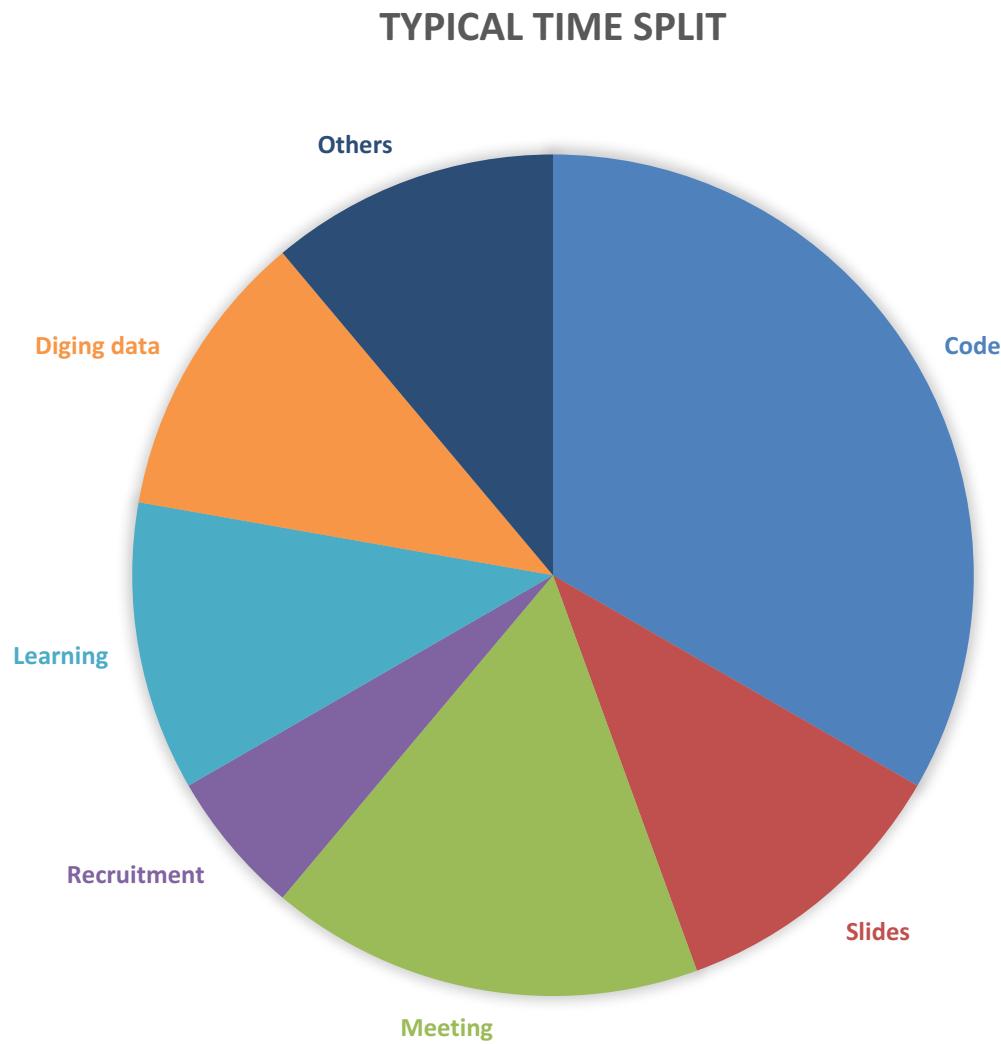
What I've been doing

- Bank: scraping (2 months)
- Insurance: optimize a call center (7 months)
- Music industry (6 months)
- Telecommunication (1 month)
- Fraud detection (10 days!)
- Hospital (5 months)
- Chatbot (2 months)
- Fire brigade (now)
- + lectures, R&D, recruitment, presentations, meetups, ...

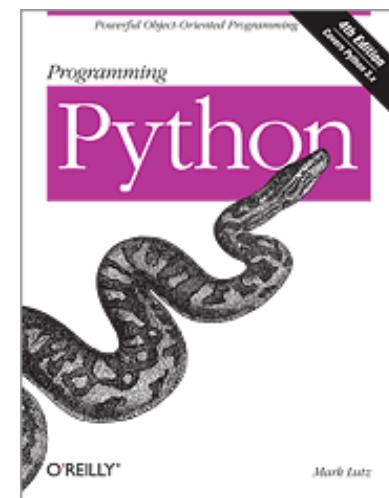
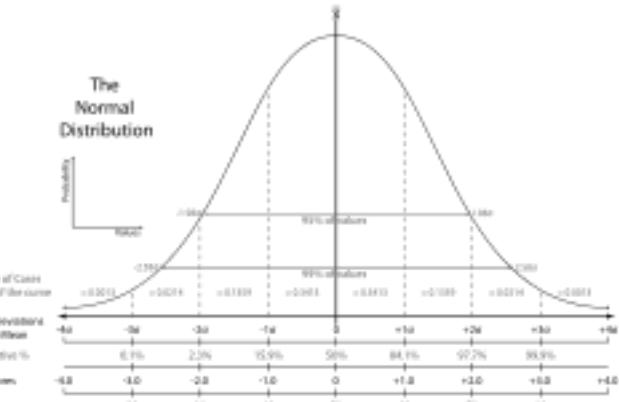
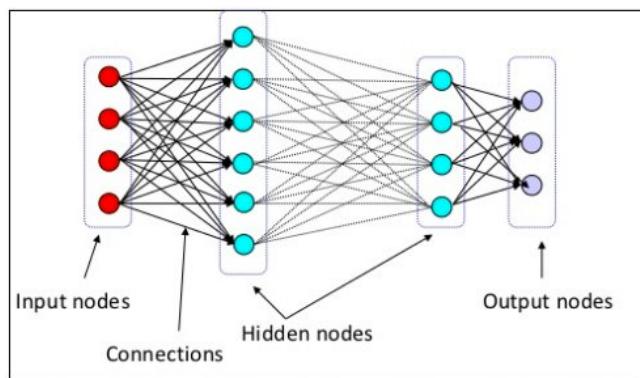
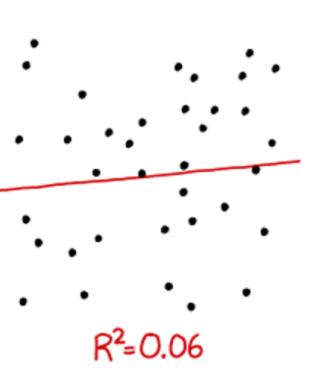
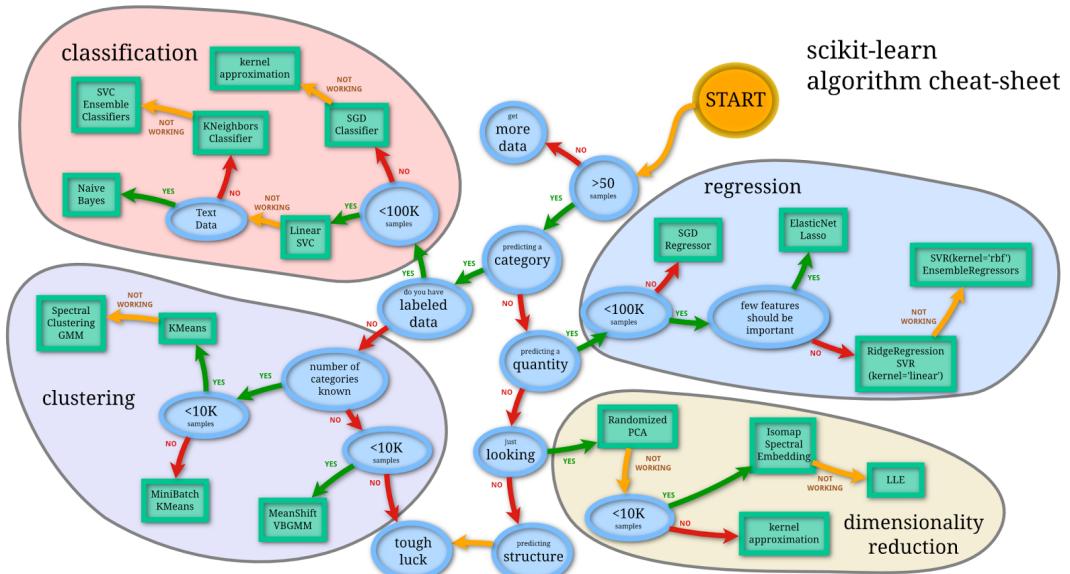
Data science use cases



Typical time split

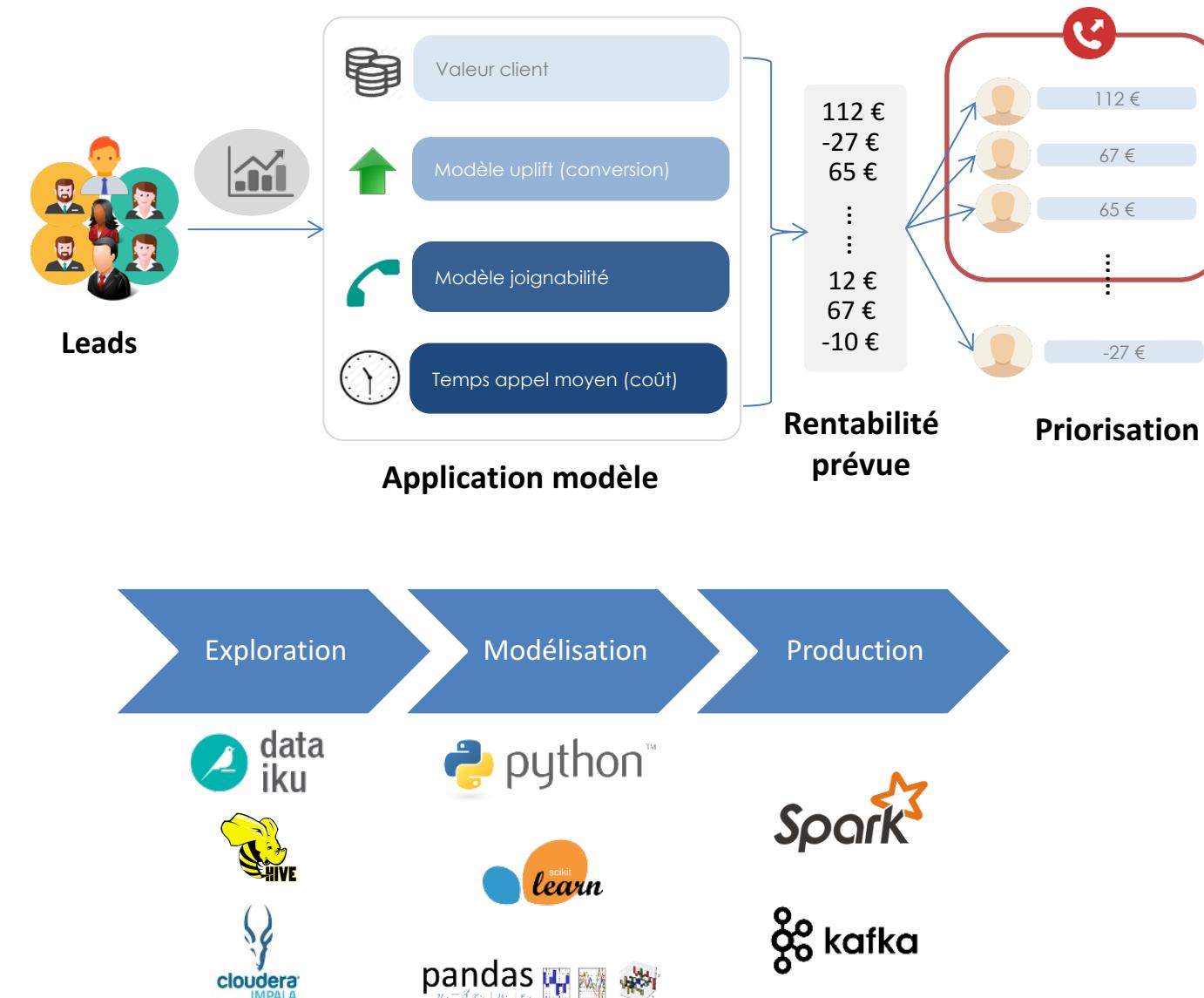


The data scientist tool box



Grande assurance française

Optimisation de centres d'appels téléphoniques



Contexte

Une assurance en ligne dispose de plusieurs centres d'appels téléphoniques, dont elle souhaite prioriser les tâches. Un même centre d'appel gère les appels entrants des clients et les appels commerciaux qui font suite à l'édition d'un devis en ligne. Une meilleure qualification des leads grâce à un modèle de machine learning permet de mieux prioriser les appels commerciaux, tout en améliorant la qualité de service sur les appels entrants.

Mission

Quantmetry a accompagné son client dans :

- Intégration et préparation des données téléphoniques, contractuelles et online
- Statistiques descriptives des appels et contrats
- Construction de modèles de prédiction de joignabilité, de conversion et d'upselling
- Construction de l'algorithme de priorisation à partir des prédictions

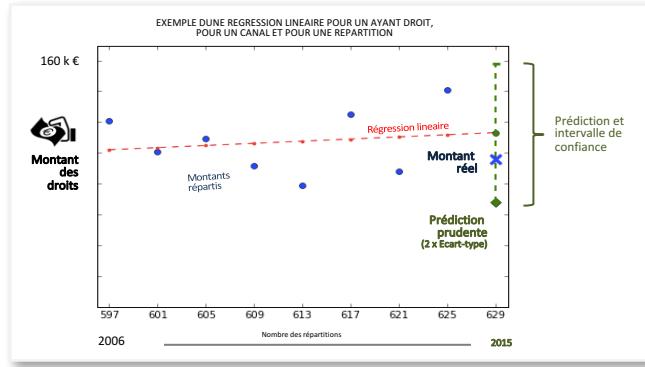
Charges : 140 JH

Durée : de juillet 2015 à janvier 2016

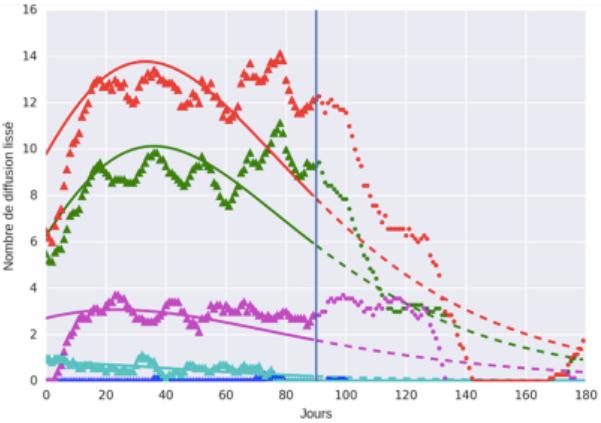
Environnement : Dataïku, Hive, Python, Spark, Kafka

Société de gestion de droits d'auteurs

Prédiction de l'usage des œuvres



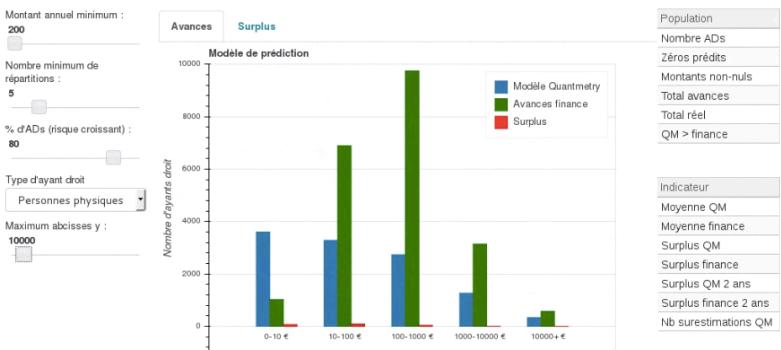
Prédiction directe des montants de droits perçus



Prédiction du nombre de diffusions d'une œuvre

MODÈLE DE PRÉDICTION DE DROITS

Avances prédites pour les ayants droit. Prédictions pour 2014 basées sur 2009-2013 inclus.



Outil de visualisation et d'analyse du modèle prédictif

Contexte

Notre client prend en charge la collecte et la répartition des droits dus aux auteurs, compositeurs et éditeurs de musique qui sont ses membres. Il souhaite anticiper le montants des droits générés par ses ayants droit en utilisant des modèles prédictifs, notamment pour proposer un système d'avance.

Solution proposée

Quantmetry a accompagné son client dans :

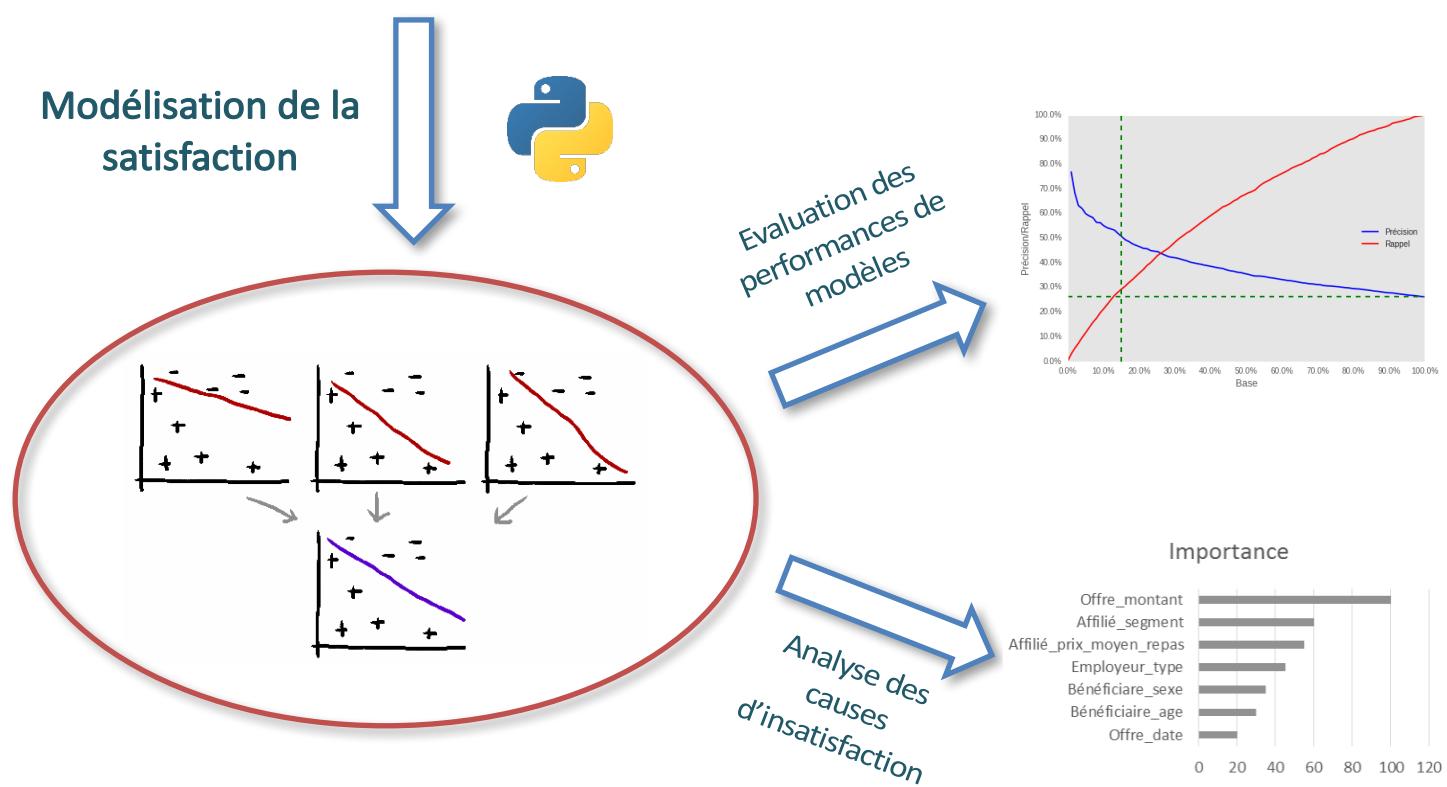
- La prédition des montants droits perçus par un modèle averse à la surestimation
- La segmentation des supports de diffusion des œuvres en fonction du profil de diffusion type
- La prédition du volume de diffusion des œuvres
- La réalisation d'un outil de visualisation et d'analyse des montants perçus par les ayants droit et des prédictions du modèle

Technologies

Python (scikit-learn, statsmodels, pandas, matplotlib)

Fournisseur d'Accès Internet

Prédiction d'insatisfaction des abonnés



Contexte

Notre client possède une quantité importante de données techniques sur le fonctionnement de ses boîtiers d'accès internet, ainsi que sur ses abonnés. Les interruptions de connexion et les ralentissements de débit peuvent mener à la résiliation de l'abonnement. Pour cibler les clients insatisfaits notre client a réalisé un sondage grâce auquel un modèle d'insatisfaction peut être construit.

Le client souhaite se doter d'un outil de prédiction de l'insatisfaction des clients vis-à-vis de leur connexion internet.

Mission

Quantmetry a accompagné cet acteur dans :

- La création et l'implémentation d'un ensemble de modèles de prédiction de l'insatisfaction des abonnés
- L'analyse des causes d'insatisfaction découvertes par les modèles
- L'optimisation fine de la performance des modèles par des méthodes d'assemblage de modèles variés

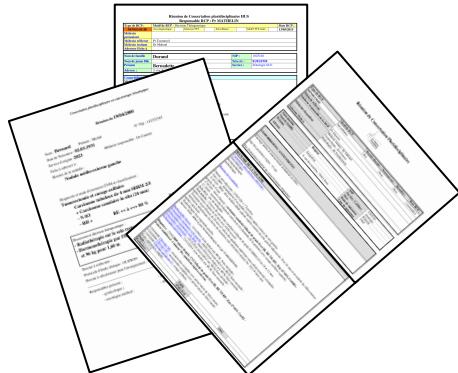
Durée : de février à avril 2016

Environnement : Python (pandas, sklearn, theano)

Analyse de fiches médicales de patientes atteintes du cancer du sein

Senometry
Data Science for Breast Health

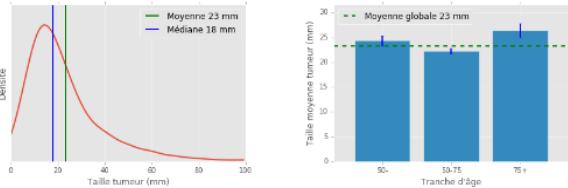
Fiches médicales



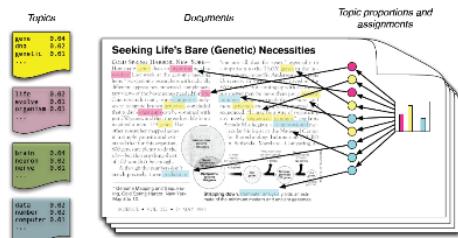
Constitution
d'un corpus

hystique
papillaire
métaglycique
adénoïde
séretoire
canalaire
médullaire
invasif
Carcinome mucineux
lobulaire CCI
tubuleux

Analyse statistique



Recherche de corrélations



Contexte

Les hôpitaux publics accumulent de nombreuses informations sur les patients qu'ils traitent. Une grande partie de ces informations est stockée dans des fichiers Word sous forme de texte non-structuré.

La structuration de ces fiches représente un enjeu important pour l'amélioration du fonctionnement d'un service, et plus généralement pour la recherche médicale. En effet, l'analyse de ces fiches par des méthodes de Traitement Automatique des Langues permet des études statistiques sur des cohortes très importantes, permettant par exemple de déterminer l'influence d'un traitement donné sur l'évolution d'une pathologie.

Mission

Quantmetry a accompagné cet acteur dans :

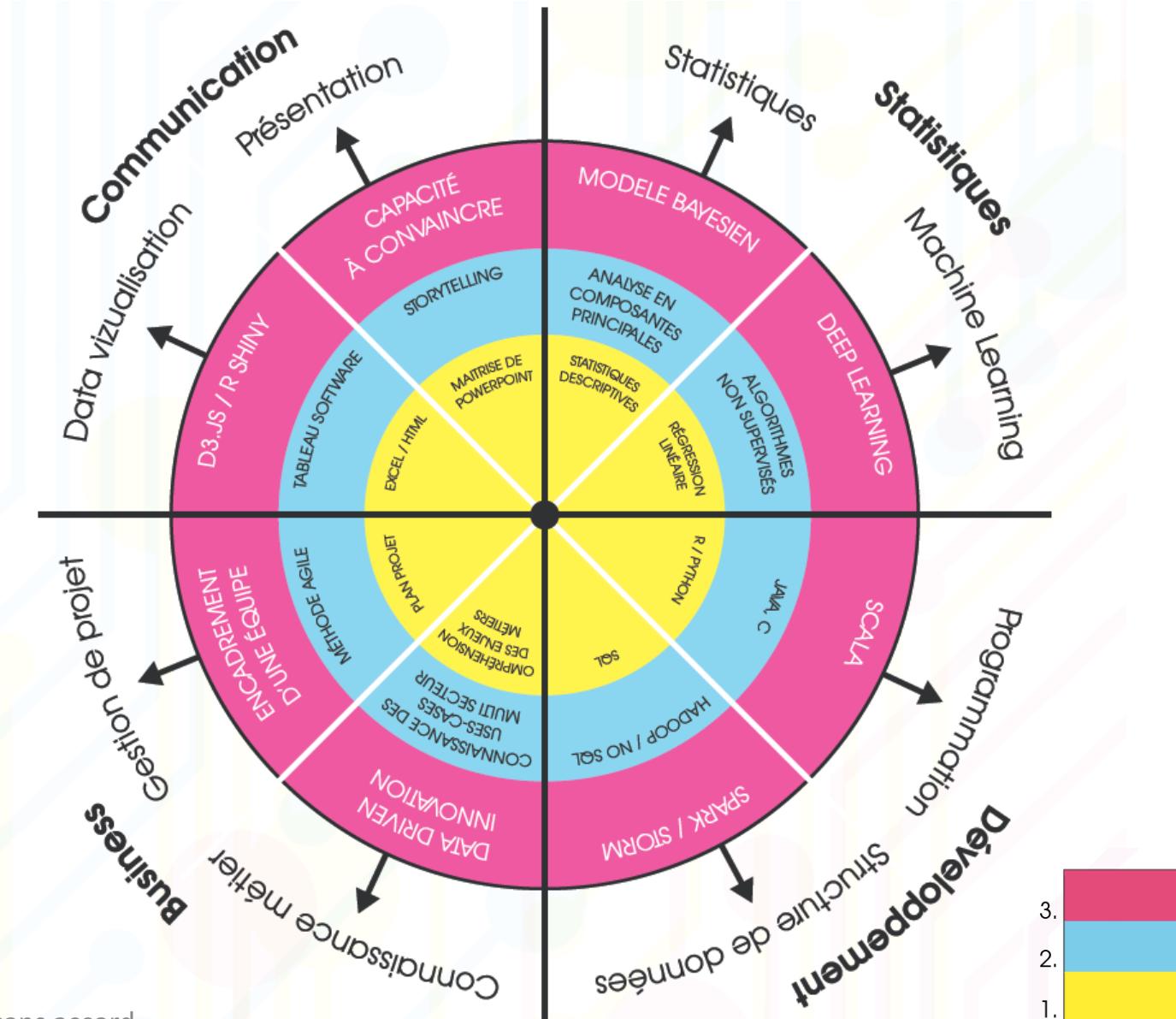
- L'anonymisation et la dé-identification des données
- La structuration des fiches médicales par extraction d'entités nommées
- L'analyse statistique de la cohorte de patientes
- La recherche de corrélation par des méthodes de Traitement Automatique des Langues

Durée : d'avril à juin 2016

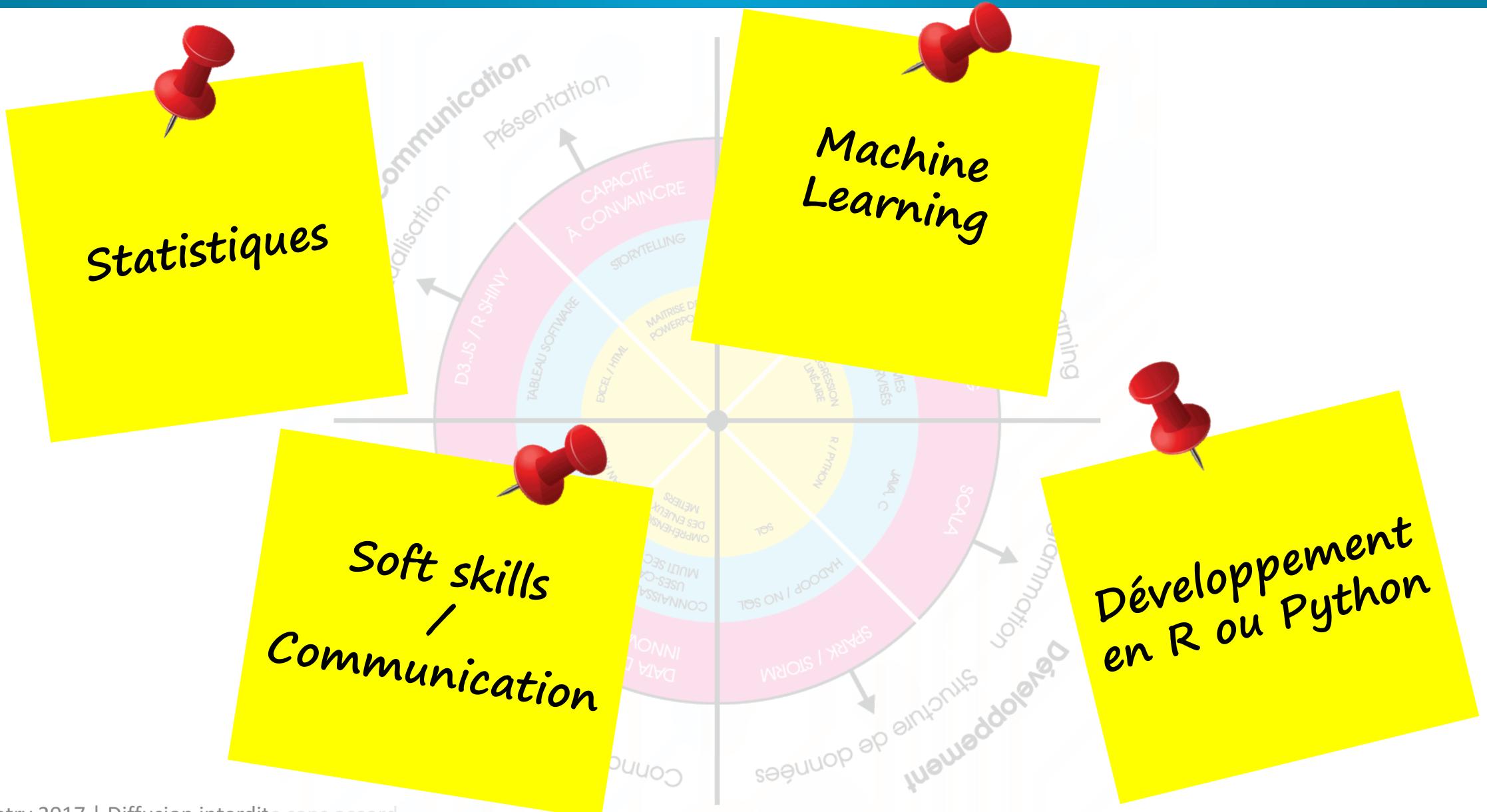
Environnement : Python, NLTK, Gensim

Data science skills

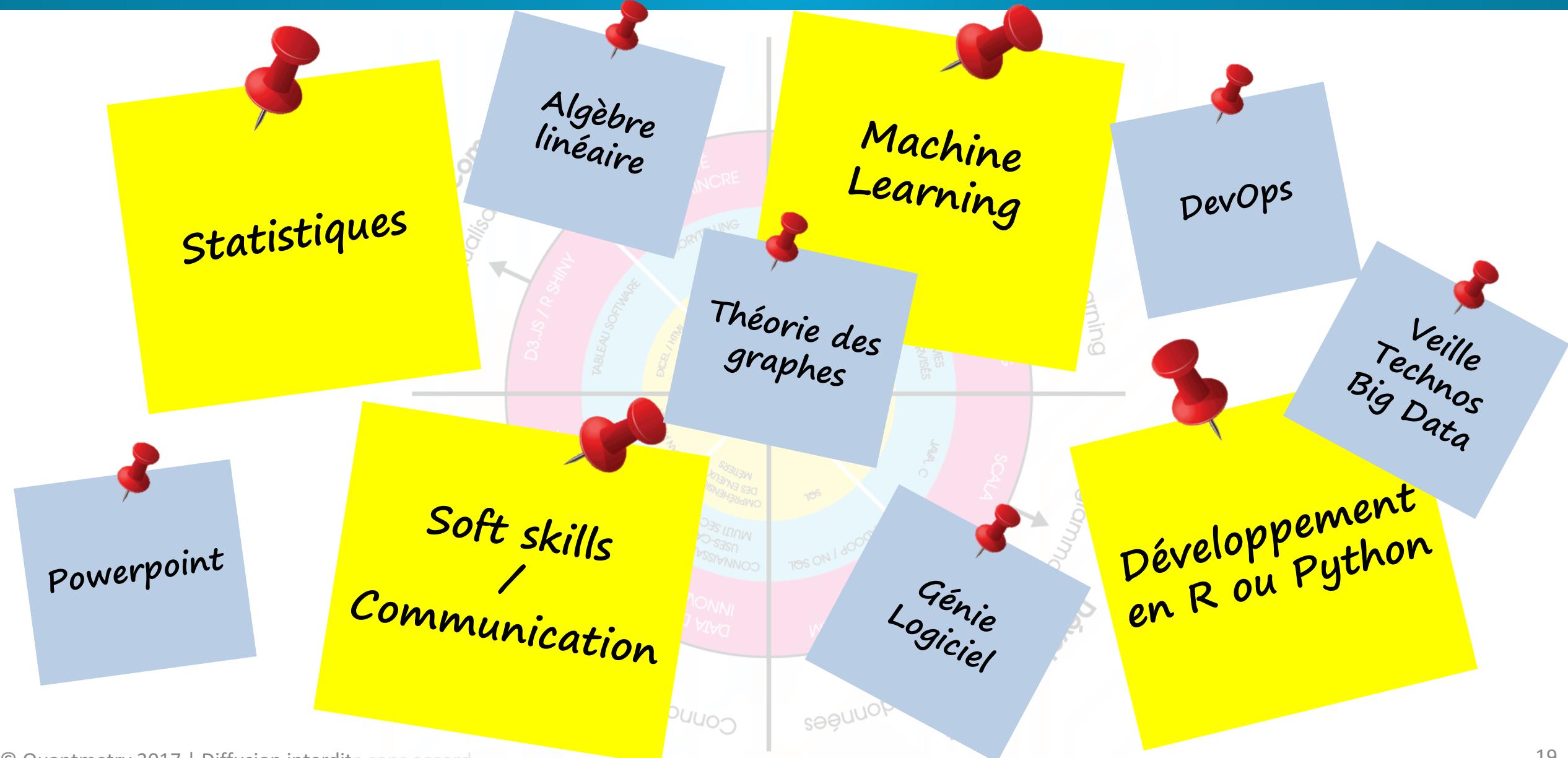
Data science profiles



Data science profiles



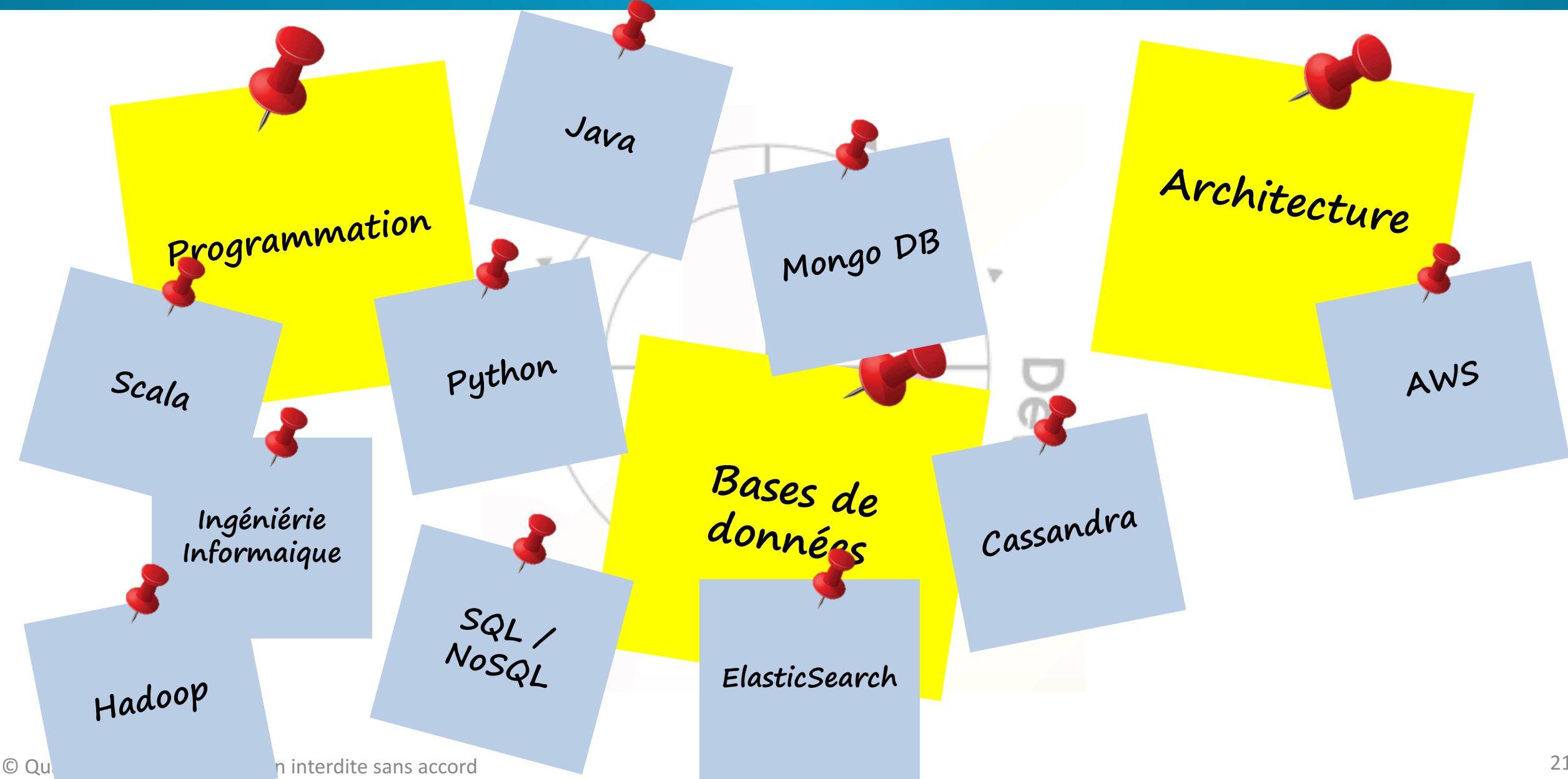
Data science profiles



Data Engineer & Data Architect



Data Engineer & Data Architect



Useful things learnt during PhD

- Overcome frustration!
- Modelling instinct
- Writing
- Computer tricks
- Presentation skills

A few tips

My 2 cents

The screenshot shows the homepage of the Paris Machine Learning Applications Group on Meetup.com. The header features the Meetup logo and navigation links for 'Create a Meetup', 'Invite', and 'Get the app'. The main content area is titled 'Paris Machine Learning Applications Group' and includes a red banner for an event at Devvoxx. The event details are: 'Paris Machine Learning @ Devvoxx', 'Thursday, April 6, 2017 7:00 PM', 'Devvoxx - Palais des congrès 2 place de la Porte Maillot, Paris (map)'. Below this, there's a 'Programme' section and an 'Inscription' section. On the right, there's a sidebar for 'Are you going?' with 'Yes' and 'No' buttons, showing '203 going'. It lists two organizers: Franck Bardol and Sébastien T., along with a third person, SUN quan.

The screenshot shows the 'Competitions' page on Kaggle. The top navigation bar includes 'kaggle', a search bar, and links for 'Competitions', 'Datasets', 'Kernels', 'Discussion', and 'Jobs'. There are 'Sign Up' and 'Log In' buttons. The main heading is 'Competitions'. Below it, there's a summary: '11 active competitions'. A table lists three featured competitions: 'Data Science Bowl 2017' (prize \$1,000,000, 1,831 teams), 'The Nature Conservancy Fisheries Monitoring' (prize \$150,000, 2,219 teams), and 'Intel & MobileODT Cervical Cancer Screening' (prize \$100,000, 216 teams). Each entry includes a thumbnail, a brief description, and a 'Featured' badge.

- Go to meet-ups
- Do not send applications everywhere!
- Learn less learn better
- Try Kaggle!
- Learn code for real
- Make sure you are motivated for the job