Les aspects techniques et non techniques des données

LES PRINCIPES FONDAMENTAUX DE LA SCIENCE DES DONNÉES

1. Les aspects techniques et non techniques des données

Les compétences quantitatives

Contexte extra-universitaire :

- appliquer des **méthodes quantitatives** à des problèmes (d'affaires) afin d'obtenir des **informations exploitables**
- il est difficile d'avoir une expertise dans **tous les** domaines des mathématiques, des statistiques, de l'informatique, de la science des données, de l'ingénierie des données, etc.

Avec un diplôme en maths/stats, par exemple :

- **expertise** dans 2-3 domaines
- **compréhension décente** des disciplines connexes
- **connaissances de base** du reste domaines

La flexibilité est une alliée, le perfectionnisme… c'est un peu moins évident

Les compétences quantitatives

Suggestions :

- **suivre les tendances**
- devenir **compétent dans vos domaines non spécialisés**
- savoir **où trouver des renseignements**

Dans de nombreux cas (70 % ?), seuls les fondements (2e -3e années de cours obligatoires à uOttawa) suffisent pour répondre aux besoins du gouvernement et de l'industrie.

Focus : assurez-vous de bien **comprendre** les bases et les tremplins.

Dans les autres cas, des connaissances plus sophistiquées sont nécessaires.

Les compétences quantitatives

- échantillonnage et collecte des données
- traitement et nettoyage des données
- visualisation des données
- modélisation mathématique
- méthodes statistiques
- analyse de régression
- modèles de file d'attente
- apprentissage machine
- apprentissage profond
- apprentissage par renforcement
- modélisation stochastique
- optimisation et recherche opérationnelle
- analyse de survie
- analyse bayésienne des données
- détection des anomalies
- réduction de la dimension
- extraction et la prévision des tendances
- cryptographie et théorie du codage
- conception des expériences
- théorie des graphes et des réseaux
- traitement du langage naturel
- etc.

Les logiciels et les outils

Le travail quantitatif moderne requiert généralement de la **programmation** (ou l'utilisation de logiciels de type pointer-cliquer, à tout le moins).

Mais les langages de programmation **vont et viennent.**

Il est important de comprendre non seulement la syntaxe d'un langage particulier, mais aussi le fonctionnement des langages informatiques et de l'infrastructure informatique en général.

ATTENTION : ne vous laissez pas entraîner dans les rivalités de programmation ... tout est plus ou moins équivalent sur le plan fonctionnel !

Les logiciels et les outils

Programmation

- Python, R, C/C++/C#, Perl, Julia, regexps (, Visual Basic ?), Java, Ruby, etc.

Gestion des bases de données

- SQL et variantes, ArangoDB, MongoDB, Redis, Amazon DynamoDB (, Access ?), Big Query, Redshift, Synapse, etc.

Visualisation des données

- ggplot2, seaborn, plot.ly, Power BI, Tableau, D3.js, Google Data Studio, logiciels spécialisés, etc.

Simulations, analyse statistique, analyse des données, apprentissage automatique

- tidyverse, scikit-learn, numpy, pandas, scipy, MATLAB, Simulink, SAS, SPSS, STATA (, Excel ?), Visio, TensorFlow, keras, Spark, Scala, etc.

Mise en page et rapports

- LaTeX, R Markdown, Adobe Illustrator, GIMP (, Word ?, PowerPoint ?), etc.

Les logiciels et les outils

Q : À StatCan, R ou SAS ?

R : StatCan est dans une lente période de transition. L'Agence est mieux équipée pour SAS (avec des options "Big Data", comme SAS Grid).

R n'est pas aussi idéal pour les gros fichiers (par exemple, les données de recensement), il n'est donc pas une option dans de tels cas car il est encore trop lent (à moins que vous ne disposiez de serveurs très puissants). Mais nous préférerions utiliser les paquets R, c'est donc un dilemme.

TL;DR : R est notre avenir, mais SAS est encore très présent. En période de transition, **les analystes qui connaissent les deux sont mieux placés**.

L'approche des “I” multiples

La compétence (ou l'expertise) technique et quantitative est **nécessaire** pour faire un bon travail quantitatif *dans le monde réel*, mais elle **n'est pas suffisante** - les solutions optimales dans le monde réel ne sont pas toujours les solutions académiques ou analytiques optimales.

C'est peut-être la plus grande surprise pour les nouveaux gradés universitaires.

Ce qui fonctionne pour une personne, un projet, etc. peut ne pas fonctionner pour un autre - **méfiez-vous de la tyrannie des succès précédents** !

L'objectif du travail quantitatif inclus la livraison d'**analyses/produits utiles**.

L'approche des “I” multiples

- **intuition**
  compréhension du contexte
- **initiative**
  établir un plan d'analyse
- **innovation**
  nouvelles façons d'obtenir des résultats, au besoin
- **assurance (“insurance”)**
  essayer plusieurs approches
- **interprétabilité**
  fournir des résultats explicables
- **utilité (“insights”)**
  fournir des résultats exploitables
- **intégrité**
  rester fidèle aux objectifs et aux résultats
- **indépendance**
  auto-apprentissage et auto-enseignement
- **interactions**
  des analyses solides grâce au travail d'équipe
- **intérêt**
  trouver des résultats intéressants
- **intangibles**
  penser "en dehors de la boîte" ;
- **curiosité (“inquisitiveness”)**
  ne pas se contenter de poser les mêmes questions à plusieurs reprises

L'approche des “I” multiples

Les analystes ne sont pas seulement jaugés sur leur savoir-faire technique, mais aussi sur leur capacité à **contribuer positivement** au lieu de travail :

- communication
- travail en équipe et capacités multidisciplinaires
- les subtilités sociales et la flexibilité
- intérêts non techniques

Les employeurs choisissent rarement des robots lorsque des êtres humains sont disponibles ; les parties prenantes sont plus susceptibles d'accepter les recommandations quantitatives provenant d’**analystes bien équilibrés**.

Vous devez également évaluer les éventuels employeurs/clients sur ces axes.

Rôles et responsabilités

Une analyste de données ou une scientifique de données (au **singulier**) a peu de chances d'obtenir des résultats significatifs – il y a trop de parties mobiles.

Les projets réussis nécessitent des **équipes** de personnel hautement qualifié qui comprennent les **données**, le **contexte** et les **défis**.

La taille de l'équipe peut varier de quelques personnes à plusieurs dizaines ; il est généralement plus facile de gérer des équipes plus petites (de 1 à 4 membres, par exemple, avec des **chevauchements de rôles**).

Experts du domaine

- font autorité dans un domaine ou un sujet particulier
- guider l'équipe en cas de complications inattendues et de lacunes dans les connaissances

Rôles et responsabilités

Chefs de projet / chefs d'équipe

- comprendre suffisamment le processus pour reconnaître si ce qui est fait a du sens
- fournir des estimations réalistes du temps et des efforts nécessaires à la réalisation des tâches
- agir en tant qu'intermédiaire entre l'équipe et les clients/partenaires
- responsable des livrables du projet.

Traducteurs de données

- avoir une bonne maîtrise des données et du dictionnaire de données
- aider les experts de domaines à transmettre le contexte sous-jacent à l'équipe de science des données

Ingénieurs en données / Spécialistes en bases de données

- travailler avec les clients et les parties prenantes pour acquérir des sources de données utilisables
- peuvent participer aux analyses, mais ne sont pas nécessairement des spécialistes.

Rôles et responsabilités

Analystes de données

- nettoyer et traiter les données
- préparer les visualisations initiales
- avoir une compréhension décente des méthodes quantitatives (au maximum 1 domaine d'expertise)
- effectuer des analyses préliminaires

Scientifiques des données

- travailler avec des données traitées pour construire des modèles sophistiqués
- concentrez-vous sur des informations exploitables
- avoir une bonne compréhension des algorithmes/méthodes quantitatives (2 ou 3 domaines d'expertise)
- peuvent les appliquer à une variété de scénarios de données
- on peut compter sur vous pour rattraper rapidement les nouvelles matières.

Rôles et responsabilités

Ingénieurs en informatique

- concevoir et réaliser des systèmes informatiques et des pipelines
- participent au développement de logiciels et au déploiement de solutions de science des données.

Spécialistes en assurance qualité/contrôle de la qualité AI/ML

- concevoir des plans d'essai pour les solutions qui mettent en œuvre des modèles AI/ML
- aider l'équipe à déterminer si les modèles sont capables d'apprendre

Spécialistes en communication

- communiquer des informations exploitables aux gestionnaires, aux analystes politiques, aux décideurs et aux parties prenantes.
- peuvent participer aux analyses, mais pas nécessairement des spécialistes (souvent des traducteurs de données)
- se tenir au courant des comptes rendus populaires des résultats et développements quantitatifs

Aide-mémoire

1. Les solutions commerciales ne sont pas toujours des solutions académiques.
2. Les données ne soutiennent pas toujours les espoirs et les besoins des parties prenantes.
3. Une communication opportune est essentielle – à l'externe comme à l'interne.
4. Les scientifiques des données doivent être flexibles (dans la limite du raisonnable), et capables d'apprendre quelque chose de nouveau, rapidement.
5. Tous les problèmes ne doivent pas faire appel la science des données.
6. Nous devons tirer des leçons des bonnes comme des mauvaises expériences.

Aide-mémoire

7. Gérez les projets et les attentes.
8. Maintenez un équilibre sain entre vie professionnelle et vie privée.
9. Respectez les parties prenantes, le projet, les méthodes et l'équipe.
10. L’analyse ne consiste pas à montrer à quel point nous sommes intelligents, mais à savoir comment nous pouvons fournir des informations exploitables.
11. Lorsque ce que le client veut est impossible, proposez des alternatives.
12. "Il n'y a pas de repas gratuit."

Lectures conseillées

Les aspects techniques et non techniques des données

Data Understanding, Data Analysis, Data Science
Volume 2: Fundamentals of Data Insight

13. Non-Technical Aspects of Quantitative and Data Work
    13.1 First Principles
    - The “Multiple I” Approach
    - Roles and Responsibilities
    - Analysis Cheatsheet

    13.3 Lessons Learned

Exercices

Les aspects techniques et non techniques des données

1. Parmi les compétences quantitatives présentées dans cette section, quelles sont celles que vous possédez ? Lesquelles vous intéressent ? Lesquelles envisagez-vous d'apprendre ?

2. Parmi les compétencesinformatiques présentées dans cette section, quelles sont celles que vous possédez ? Lesquelles vous intéressent ? Lesquelles envisagez-vous d'apprendre ?

3. Quel rôle en matière de données occupez-vous dans votre organisation ? Pour quel rôle pensez-vous être le mieux placé actuellement ? Quel est le rôle auquel vous aspirez ?

4. Avez-vous rencontré les leçons de l'aide-mémoire dans votre travail ? En avez-vous rencontré d'autres ?