Què és un data scientist ?

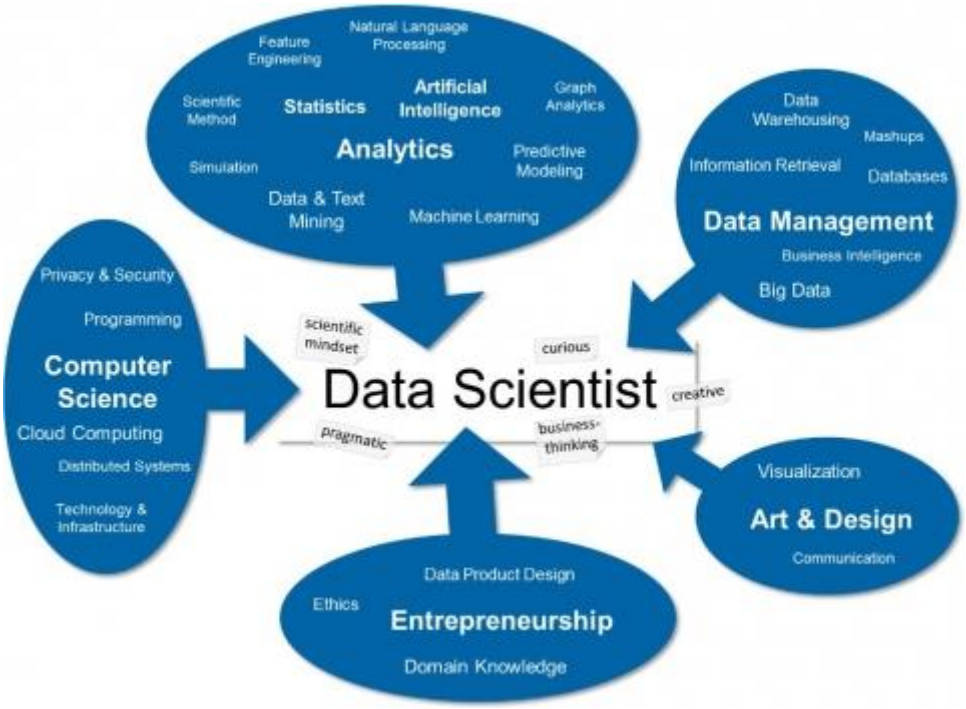

Figura extreta de l'article [Applied Data Science in Europe]

Un Data Scientist és un expert en Data Science (Ciència de dades), la seva feina consisteix en extreure coneixement a partir de les dades per poder respondre les preguntes que se li formulen.

Què és la "ciència de dades" ?

Aquesta "ciència de les dades", nascuda del mètode científic, és l'evolució del que fins ara es coneixia com a Analista de dades, però a diferència d'aquest que només es dedicava a analitzar fonts de dades d'una única font, el Data Scientist ha d'explorar i analitzar dades de múltiples fonts, sovint immenses (conegudes com a Big Data), i que poden tenir formats molt diferents. A més a més, ha de tenir una forta visió de negoci per ser capaç d'extreure i transmetre recomanacions als responsables de negoci de la seva empresa.

Aquests conjunts de dades poden provenir de les dades generades per tot tipus de dispositius electrònics (com un mòbil, tot tipus de sensors, seqüenciadors de genoma,....), xarxes socials, dades mèdiques, pàgines web,... i afecten de manera molt significativa la recerca actual en molts camps com les ciències biològiques, la informàtica mèdica, la salut, les ciències socials, per citar-ne només alguns.

Quin procés segueix un data scientist ?

El procés que segueix un Data Scientist per a respondre a les qüestions que se li plantegen es poden resumir en aquests 5 passos:

- Extreure les dades, independentment de la seva font (webs, csv, logs, apis, etc.) i del seu volum (Big Data o Small Data).
- Netejar les dades, per eliminar allò que distorsiona les mateixes.
- Processar les dades usant diferents mètodes estadístics (inferència estadística, models de regressió, proves d'hipòtesis, etc.).
- Dissenyar nous testos o experiments en cas necessari.
- Visualitzar i presentar gràficament les dades.

Què s'espera d'un Data Scientist ?

El que s'espera d'un Data Scientist és que no només sigui capaç d'abordar un problema d'explotació de dades des del punt de vista d'anàlisi, sinó que també tingui les aptituds necessàries per cobrir l'etapa de gestió de dades. Així, l'objectiu d'un perfil d'aquest tipus és apropar dos mons (el de gestió i anàlisi de dades), que fins ara havien pogut existir separats, però que a causa dels nous requisits de volum, de varietat de dades i de velocitat en l'explotació d'aquestes (i.e., les tres V's de la definició estàndard del terme Big Data), s'ha tornat imprescindible dur a terme aquesta explotació a través d'un perfil combinat, i que a més a més, també entengui el negoci per tal de dirigir aquesta explotació cap a resultats que puguin ser d'interès per a la companyia.

Quin perfil ha de tenir un Data Scientist ?

El perfil del Data Scientist, és en certa manera, com una poció màgica, requereix com a ingredients principals habilitats avançades en informàtica, matemàtiques/estadística, aprenentatge automàtic, passió per les dades, saber manegar grans volums de dades, curiositat, capacitat de comunicar el coneixement que hem extret de les dades, visió de negoci, etc...

Com ja intuïu, cal aprendre moltes coses, ja que la "ciència de dades" és multi disciplinar, i és una especialització alhora exigent i avançada, però la combinació és molt potent i difícil de trobar, pot ser és per això que la revista Harvard Business Review la va definir com la feina més Sexy del segle 21.

En el diagrama que encapçala l'article, extret de l'article Applied Data Science in Europe publicat a la Zurich University of Applied Sciences i al blog d'un del seus autors, en Thilo Stadelmann , es detallen les diferents habilitats que hauria de tenir un bon Data Scientist.

Quins reptes podem abordar ?

Per citar només un exemple, un dels reptes de les tecnologies actuals de Big Data i Data Science és la seva aplicació en l'anàlisi de la quantitat ingent d'informació genòmica de què disposem, i que serveix per estudiar malalties com per exemple el càncer.

Actualment existeixen un munt de fonts de dades obertes (Open Data) que podem analitzar, com per exemple, les de l'ajuntament de Barcelona (http://opendata.bcn.cat/opendata/ca) o si volem big data, les del Projecte del Genoma del Càncer Pediàtric de la Universitat de Washington, de l'Hospital Infantil St. Jude, que ha posat a disposició de tothom les dades complertes del genoma del càncer humà (http://www.pediatriccancergenomeproject.org/site/).

Si el tema us motiva, podeu participar en diferents reptes de Data Science, com per exemple: Identificar signes de retinopatia diabètica en imatges de l'ull. Aquest i d'altres reptes , es publiquen, per exemple a la pàgina https://www.kaggle.com/competitions , on si sou bons, podeu aconseguir unes bones recompenses.

Altres reptes més globals, per citar-ne un, de les tecnologies actuals de Big Data i Data Science, és la seva aplicació en l'anàlisi de la quantitat ingent d'informació genòmica de què disposem, i que serveix per estudiar malalties com per exemple el càncer.

Penseu que els humans, que tenim 23 parells de cromosomes, cadascun es compon per uns 3.200 milions de parells de bases d'ADN que contenen uns 20.000 - 25.000 gens. Determinar quina combinació d'aquests gens són significatius per certes malalties obre la porta a pensar que pot ser algun dia tindrem una medicina personalitzada.

Com en puc aprendre ?

Una bona manera d'aprendre Data Science, és mitjançant l'especialització a la plataforma de MOOCs (cursos on-line) Coursera, des d'on s'ofereixen els nou cursos que componen aquesta especialització de manera gratuïta.

A l'inLAb FIB fa molts anys que treballem en l'Anàlisi de dades, sobretot en els àmbits de modelització, simulació i optimització. Amb l'aparició de les tecnologies per tractar grans volums de dades (Big data) ara disposem d'eines molt potents que complementen aquest àmbit.

-