Training Data Scientists : a Few Challenges

Gilbert Saporta
CEDRIC- CNAM,
292 rue Saint Martin, F-75003 Paris

http://cedric.cnam.fr/~saporta

1. Which skills for Data Scientists?

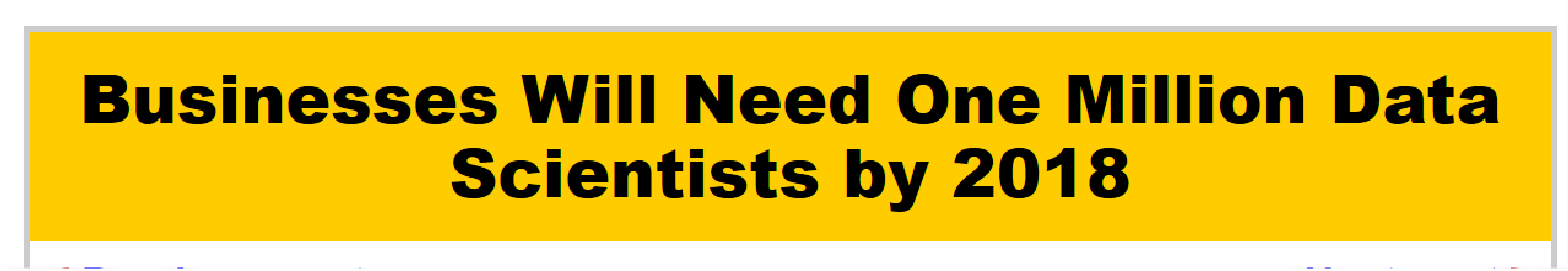
- A "Data Scientist" is a professional who uses scientific methods to liberate and create meaning from raw data - somebody who can play with data, spot trends and learn truths few others know. (the Data Science Association)
- A Data Scientist is a person who is better at statistics than any software engineer and better at software engineering than any statistician (Donoho)

A rare bird with three sets of skills

- **technical skills** : Mastery of statistical modeling methods (fundamental, main models, learning techniques, articial intelligence) and associated software. Knowledge of the computer tools required to reconcile and process large data from multiple and heterogeneous sources.
- To have the **meaning and the feeling of data**: like the ability to select potentially relevant data, be more concerned by checking the robustness and operational value of the results than to use a new statistical model or the latest software, show creativity in the combination of methods.
- **Business skills**: understanding of the company and its issues. Communication capability. Concern for ethics and social responsibility.

2.Shortage of talents

- Kdnuggets (jan.2016)

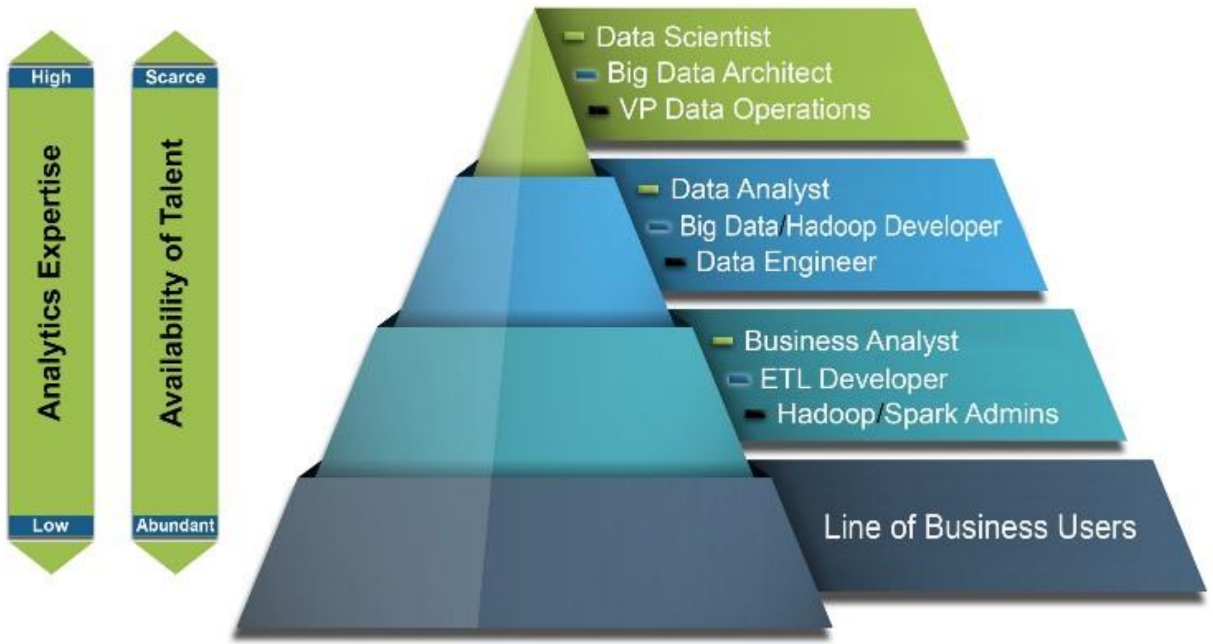


Businesses Will Need One Million Data Scientists by 2018

- There will be a shortage of talent necessary for organizations to take advantage of big data. **By 2018, the United States alone could face a shortage of 140,000 to 190,000 people with deep analytical skills** as well as 1.5 million managers and analysts with the know-how to use the analysis of big data to make effective decisions. (**McKinsey**)

- **International Data Corporation** (IDC) predicts a need by 2018 for **181,000 people with deep analytical skills**, and a requirement five times that number for jobs with the need for data management and interpretation skills.

Addressing the Big Data Skills Gap



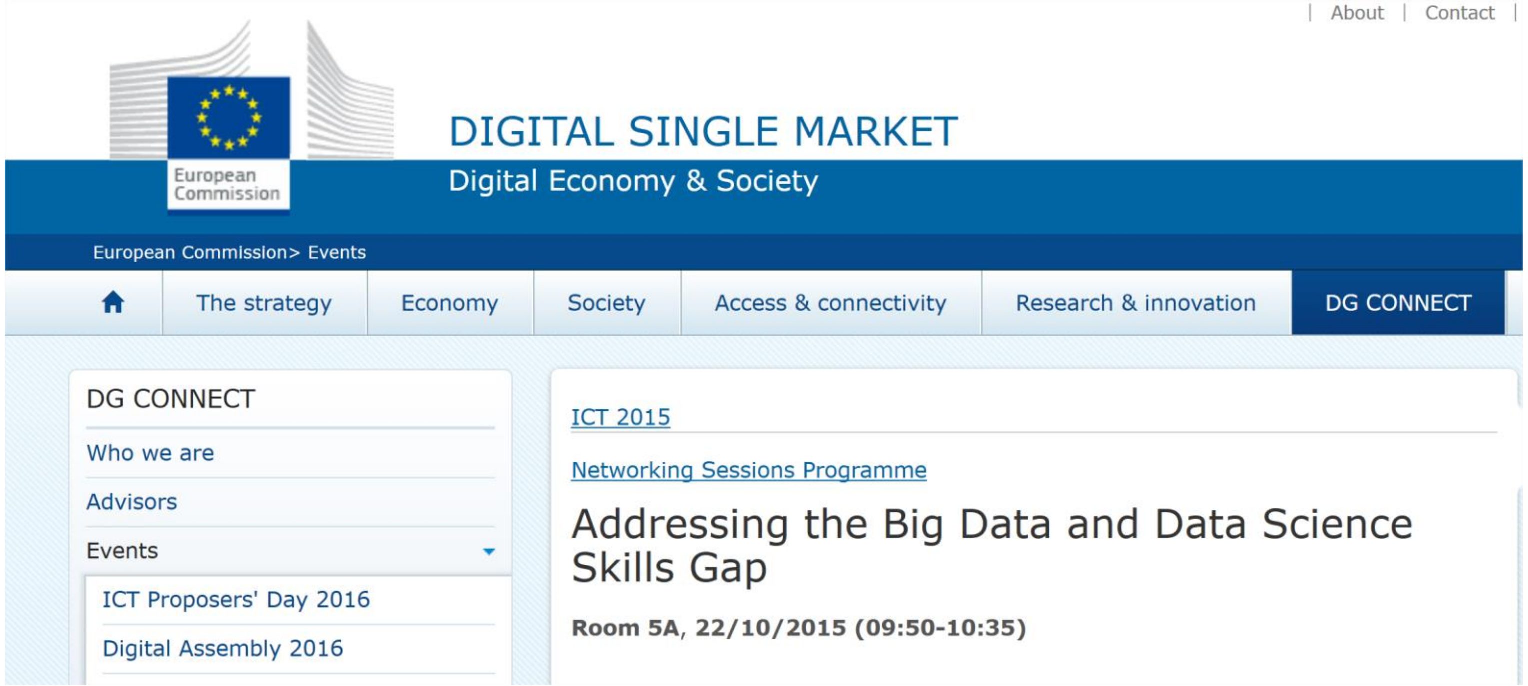
Roles: Data Analysis - Development - Operations

- According to her estimates, there were about **6.1 million data workers** in the EU, which represents 3% of total EU employment, most being high profile professionals, managers in the field of data innovation.
- She then revealed that there was **a skill gap of 7.5 percent** of total demand, in other words vacancies in 2014, between the supply of data workers in Europe and the demand for data workers, and that that represents a threat to the European data industry. **It became apparent that the most relevant gap concerned data scientist skills.**
- “Data scientists” were defined as people with special technical, mathematical tools specialised for designing tools and applications for data and there are about 100 000 -150 000 in Europe of these highly specialised scientists.

Gabriella Cattaneo, IDC European Government Consulting
European Data Forum 2015, Luxembourg, 17 November 2015

- over the past 5 years the demand for Big Data staff in the UK alone had risen tenfold and currently 77% of Big Data roles are hard to fill. It is also estimated that there will be a 160% increase in demand for Big Data specialists between **2013-2020 to 346,000 new jobs**.

(https://ec.europa.eu/digital-single-market/events/cf/ict2015/item-display.cfm?id=15762)



ICT 2015

Networking Sessions Programme

Addressing the Big Data and Data Science Skills Gap

Room 5A, 22/10/2015 (09:50-10:35)

Big data skills gap, according to IBM report

Written by Tony Parish 11/05/2016

"One of the biggest issues is going to be the gap in skills.
Getting the skills required to analyse and manage all of this data is going to be difficult."
"By 2020 we will have one million unfilled jobs in the IT sector.
Primarily because the skills we have today aren't the right skills for the future.
The future is more about the business understanding and the data understanding."

http://www.ignite.digital/big-data-skills-gap-ibm-report/

3.Impacts on curricula

- Necessary update
- More information technologies
  - Distributed and parallel computing
- More Machine Learning
- Less big software, more free environments:
  - R, Python, ScikitLearn, Spark etc.

4.Initial training is not enough

- Opening (or converting) masters in Data Sciences will not fill the Data scientists skills gap quickly enough
- A large number, but not all, of these "new“ masters are just revamping of existing masters in statistics.

The Big Data Analytics War: IBM Watson v Kaggle? (Bernard Marr, 2016)

- It’s **humans** versus the **machines**
- on one hand the crowd-sourced trained data scientists of **Kaggle** with its 150,000 members, ready to solve problems
- IBM believes that it can offer a solution to the skills shortage in big data. “With a cognitive system like **Watson** you just bring your question – or if you don’t have a question you just upload your data and Watson can look at it and infer what you might want to know.”

https://www.linkedin.com/pulse/big-data-analytics-war-ibm-watson-v-kaggle-bernard-marr

And B.Marr concludes:

The winner is…?

- Right now, the humans and the machines seem to be tied.
- For one thing **companies** and organizations need to access data analytics wherever they can get it. Kaggle is not yet large enough, and Watson not yet ubiquitous enough that every company can just pick one.
- In the short and medium term, a combination will probably be the winner. AI is a nice supplement to the efforts of the data scientists today.
- But in the end, **I predict AI will win**. Assuming that tools like Watson continues to improve as it has in the past, it will become easier and more efficient to use, democratizing data science to anyone who can phrase a question.

- **An other solution: continuing education (lifelong learning) and learning on the job**

Our opinion is that a large part of the solution has to be found in continuous education (or life long learning) and learning on the job of statisticians and computer scientists already employed: **the troops are there!**

See French government roadmap (Bourdoncle et al., 2014)

http://www.economie.gouv.fr/files/files/PDF/Feuille-de-route_big-data151214.pdf

An experience at CNAM

- Founded in 1794 by Abbé Grégoire in order to “perfect national industry”.
- CNAM is the only public institution in France that offers adults of all ages living in many different places the tools they need to move forward.
- Nearly 70,000 employees, students and job-seekers participate in CNAM's educational programmes every year
  - 150 educational centres in France and a presence in many places around the world.
  - Most CNAM students register without informing their employers thereby avoiding pressure and facilitating their mobility to jobs in other enterprises.

- CNAM opened in 2014 a professional certificate in Big Data Analytics (27 ECTS), for statisticians and IT engineers (already holding a master’s degree) who wish to evolve or retrain in the Big Data field.
- 3 courses (180 hours)
  - *Data warehousing and Data mining* preprocessing, missing data, machine learning algorithms
  - *Distributed and documentary Data Bases* management of documentary data, non-structured or semi-structured, NoSQL systems and distributed computing
  - *Mining and visualizing massive data* recommender systems, graphs and social networks analysis
- A final project on professional or open data
  - build a movie recommender system
  - build a prediction model for the satisfaction of members of a community of readers and the associated recommender system (book-crossing)
  - detect communities from the Facebook user database
  - build a semantic search engine on articles from Le Monde newspaper
  - ..

- Organization
  - Courses taught outside working hours, during evening classes or through distance teaching.
  - Course materials are available to students on a learning platform.
  - Low registration fees (~400 €) in accordance with the social vocation of CNAM

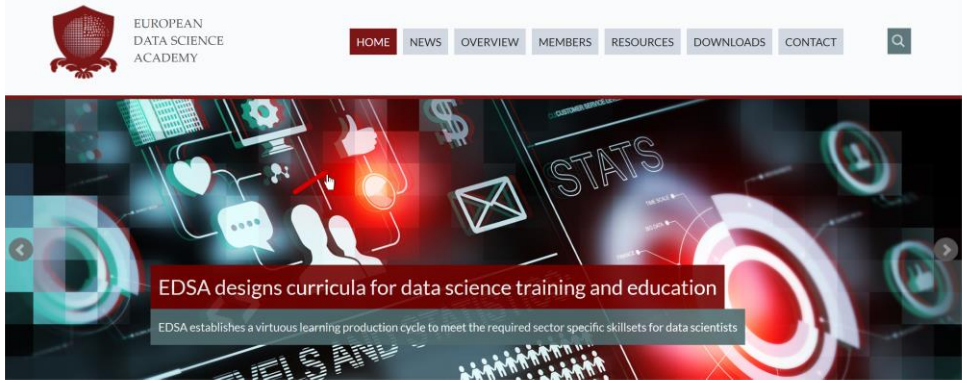
- More than 100 students with 2 main profiles
  - young professionals under 30
    - graduates with a Master's Degree in Computer Science
    - graduates with a Master in statistics (theoretical) looking for more applied training and computer skills.
  - computer scientists aged 35-40 and over with little knowledge in statistics and data analysis, who wish to anticipate the evolution of their professions

- Perspectives
  - Opening the certificate in the form of intensive courses during the day
  - Deploy it in our network
  - Add lectures on ethical and privacy issues and techniques that preserve confidentiality.

- Other experiences
  - Diploma of university (DU) “Big Data Analyst” at Paris-Descartes University 150 hours for people with two years of higher education

5.Other issues

- On line courses, MOOC : validation by credits, diplomas
- Need for cooperation and mutualisation of efforts
  - Cf. "European Data Science Academy", a Horizon 2020 project



EDSA designs curricula for data science training and education

EDSA establishes a virtuous learning production cycle to meet the required sector specific skillsets for data scientists

- Accreditation: a role for learned societies?

Federation of European National Statistical Societies FENStatS

enbis

DATA SCIENCE EuADS European Association for Data Science

Thanks!