# Proposal for a full-time Chair in Data Science and Engineering

Stijn Vansummeren, Esteban Zimanyi

February 25, 2019

## 1 Integration into existing teaching programs

From a teaching viewpoint, the position is solicited to:
(1) ensure the organisation of the courses previously taught by Toon Calders within the context of the Erasmus Mundus joint master in Big Data Management and Analytics (BDMA) as well as the Master Civil Engineer in Computer Science; and
(2) strengthen the expertise in the organisation of the specializing Master in Data Science and Big Data, which is co-organized by the EPB, the faculty of sciences, and the Solvay Brussels school of economics and management.
This comprises, in particular, the following courses.

- *INFOH420 Business Process Management*
  (5 ECTS, 24h Theory, 36h Ex).
  Offered in:
  - MA-IRIFB Master : ingénieur civil en informatique, Big Data Management and Analytics (Erasmus Mundus)
  - MA-IRIFS Master : ingénieur civil en informatique
  - MA-IREM Master: ingénieur civil électromécanicien, Gestion et technologies.
  - MA-INFO Master en sciences informatiques
- *INFOH423 Data Mining*
  (5 ECTS, 24h Theory, 12h Ex, 24h Tp).
  Offered in:
  - MA-IRIFB Master : ingénieur civil en informatique, Big Data Management and Analytics (Erasmus Mundus)
  - MA-IRIFS Master : ingénieur civil en informatique
  - MS-BGDA Master de spécialisation en science des données, Big data
  - MA-INFO Master en sciences informatiques
- *INFOH600 Computing Foundations of Data Sciences*
  (5 ECTS, 24h Theory, 12h Ex, 12h Tp).
  Offered in:
  - MS-BGDA Master de spécialisation en science des données, Big data

| | 2016-2017 | 2017-2018 | 2018-2019 |
|---|---|---|---|
| INFOH420 Business Process Management | 25 | 55 | 65 |
| INFOH423 Data Mining | N/A | 35 | 55 |
| INFOH600 Computing Foundations of Data Sciences | N/A | N/A | 20 |

Table 1: Overview of student registrations per course, 2016-2019.

Table 1 summarizes the number of students registered for these courses in the past 3 years. Note that INFOH600 is a new course that has started this academic year, and that INFOH423 started in 2017–2018.

It is important to stress that the organisation of the courses INFOH420 and INFOH423 is a **contractual obligation of the ULB** in the the EU-funded Erasmus Mundus Joint master program in Big Data Management and Analytics (BDMA). The position is solicited to ensure the human resources necessary to meeting this obligation.

Besides the courses directly concerned by the position, the topics will reinforce already existing courses, in particular INFOH501 Pattern recognition and image analysis given by Olivier DEBEIR & Christine DECAESTECKER given to the Master ingénieur civil en informatique and ingénieur civil biomédical.

## 2 Strategic character of the proposed research

**Introduction to the research domain** Data drives the modern world. Public and private organisations in all sectors face an avalanche of digital data. According to market analysis firm McKinsey, "*Data is now a critical corporate asset. It comes from the web, billions of phones, sensors, payment systems, cameras, and a huge array of other sources–and its value is tied to its ultimate use. While data itself will become increasingly commoditized, value is likely to accrue to the owners of scarce data, to players that aggregate data in unique ways, and especially to providers of valuable analytics*".[1] The European Political Strategy Centre (EPSC) confirms "*Data is rapidly becoming the lifeblood of the global economy. It represents a key new type of economic asset. Those that know how to use it have a decisive competitive advantage in this interconnected world, through raising performance, offering more user-centric products and services, fostering innovation—often leaving decades-old competitors behind. [. . . ] Data analytics will soon be indispensable to any economic activity and decision-making process, both public and private*".[2]

The data deluge and need for data analysis is not only prevalent in industry and the global economy, but also in science. Indeed, already in 2007 Turing award winner Jim Gray envisioned "data-driven science" as a "fourth paradigm" of science that uses the computational analysis of large data as a primary scientific method to complement the standard three paradigms of empirical, theoretical, and computational research.[3]

Unlocking value from raw data is hard, however: before any newly acquired data becomes useful, it must be preprocessed, integrated with existing data, cleaned from errors, appro-

[1] https://www.mckinsey.com/business-functions/mckinsey-analytics/our-insights/the-age-of-analytics-competing-in-a-data-driven-world

[2] https://ec.europa.eu/epsc/sites/epsc/files/strategic_note_issue_21.pdf

[3] S. Tansley; K. Michele Tolle (2009). The Fourth Paradigm: Data-intensive Scientific Discovery. Microsoft Research. ISBN 978-0-9825442-0-4.

priately stored, and prepared for analysis. Further, analysis is no longer restricted to simple querying or data mining, it increasingly requires exploration, recommendation, explanation, and visualisation in addition to novel insights from machine learning.

A typical data value creation chain encompasses multiple disciplines and people with different roles, of which Data Science and Data Engineering are the two most prominent. **Data Science (DS)** is the scientific "*interdisciplinary field that uses scientific methods, processes, algorithms and systems to extract knowledge and insights from data in various forms, both structured and unstructured.*"[4] In other words, data scientists focus on extracting meaning and insight from data through analytics. They are supported in their activities by **Data Engineering** (DE). Data engineers "*design and build the data ecosystem that is essential to analytics. Data engineers are responsible for the databases, data pipelines, and data services that are prerequisites to data analysis and data science.*"[5] In setting up these pipelines and functionalities that conform the ecosystem, data engineers are often faced with the challenges posed by extreme characteristics of **Big Data (BD)**, defined by the Oxford English Dictionary as "*data of a very large size, typically to the extent that its manipulation and management present significant logistical challenges.*"

The solicited chair is hence positioned in the strategic research domain of data science and data engineering.

**Relevance, cross-disciplinarity and potential impact.** While attention to the importance of data is not new, the predicted economic value that can be extracted from the availability of these data remains as of yet largely unrealized. According to McKinsey, "*Most companies are capturing only a fraction of the potential value from data and analytics. [...] manufacturing, the public sector, and health care have captured less than 30 percent of the potential value we highlighted five years ago.*"[1] This is particularly true in Europe, which is "*lagging behind in embracing the digital and data revolution.*"[1,2]. The solicited chair will contribute to improving the knowledge triangle between education; research and technology; and innovation, required to improve this situation.

We stress that the application domain of the chair is inherently cross-disciplinary: while the research domain is centered in the research pole *information technologies and intelligent systems*, data-driven innovation is being applied in Energy, Finance, Health, Transport, Security and Tourism. To illustrate, the OECD reports that, due to data-driven innovation, 380 megatonnes of $CO_2$ emissions may be saved worldwide in transport and logistics, while the utility sector may see a $CO_2$ reduction of more than 2 gigatonnes.[6]

We also stress that the problems of data science and data engineering are far from solved, as confirmed by scientific literature[7] and leading laboratories in industry and academia[8]. By

[4] V. Dhar. *Data science and prediction.* Commun. ACM 56 (12), 2013.

[5] https://www.eckerson.com/articles/data-engineering-coming-of-age

[6] *Exploring Data-Driven Innovation as a New Source of Growth – mapping the policy issues raised by "Big Data"*, OECD, 2013.

[7] E.g.: L. Cao. *Data Science: A Comprehensive Overview.* ACM Comp. Surveys, 50(3), 2017; A. L'Heureux, et al. *Machine Learning With Big Data: Challenges and Approaches.* IEEE Access, 5, 2017; A. Doan, et al. *Toward a System Building Agenda for Data Integration (and Data Science).* IEEE Data Eng. Bull., 41(2), 2018.

[8] E.g., IBM: http://www.research.ibm.com/client-programs/accelerated-discovery-lab/index.shtml, Microsoft: https://www.microsoft.com/en-us/research/research-area/data-management-analysis-visualization, Facebook: https://research.fb.com/category/data-science, Xerox: http://www.xrci.xerox.com/data-analytics/text-and-graph-analytics, Stanford

opening a chair in the research field of data science and data engineering, the EPB will hence strengthen its leadership in the strategic area. (See also Section 3 for a discussion of existing research at the EPB in this respect, and the targeted complementarity).

## 3 Integration in the Ecole Polytechnique

The chair is to be integrated in the laboratory for Web & Information Technologies (WIT) at the Dept. of Computer & Decision Engineering (CoDE) of the EPB.

WIT's research focuses on (big) data management and business intelligence. In this context, the two professors of the lab (S. Vansummeren and E. Zimanyi) have built expertise in data warehousing, business intellingence, and spatio-temporal data management (E. Zimanyi), as well as information extraction, big data systems, and foundations of data management (S. Vansummeren).

Collectively, the CoDE department has a track record of cutting-edge research in big data processing (H. Bersini, S. Vansummeren), data management (S. Vansummeren, E. Zimanyi), business intelligence (H. Bersini, E. Zimanyi), swarm robotics (M. Birattari, M. Dorigo), metaheuristics and optimization (M. Birattari, T. Stutzle) and decision engineering (Y. De Smet) and is well-equipped for research in these areas.

The position is solicited to complement the existing expertise in the department and school with expertise in large-scale data science and data engineering, focusing on topics such as: data mining, process mining, text mining, predictive data analytics, deep learning, data quality and data fusion, natural language processing, data integration and interoperability, advanced query processing (e.g. approximate query processing), and large-scale data analysis systems, and/or data visualisation.

Depending on the exact expertise of the selected candidate, synergies are possible with the LISA lab of the EPB, which is active in deep learning for image analysis (C. Decaestecker, O. Debeir) and the MLG group at the Computer Science dept. at the faculty of Sciences, which is active in machine learning and behavioral intelligence research (G. Bontempi).

In general, Data Science is a transversal discipline needed in many application domains. In particular, it is gaining importance throughout the engineering disciplines as witnessed, for example, by specialized master programs "Data Science in Engineering" offered across engineering schools in Europe[9]. As such, we expect that the chair can establish strong research links with many other laboratories of the Ecole polytechnique.

## 4 Research equipment

Research in data science and data engineering requires appropriate computational infrastructure for storing, processing, and analyzing large-scale datasets. The following infrastructure available in the EPB could be used by the solicited chair to kick-start the research activities:

- Within the WIT lab at CoDE: high end workstations (20 cores, 2.4TB storage, 128GB RAM) and a compute cluster (9 compute nodes; total 72 cores, 4.5TB storage, 216GB RAM).

(Many related projects): `http://infolab.stanford.edu/db_pages/projects.html`, UC Berkeley: `https://amplab.cs.berkeley.edu/projects/mlbase`, MIT: `https://www.csail.mit.edu/research/data-civilizer`

[9] E.g., `https://studiegids.tue.nl/opleidingen/graduate-school/special-masters-tracks/data-science-in-engineering/`

- Within the IRIDIA lab at CoDE: compute cluster (512 cores, 32TB storage, 2TB RAM)
- Within the LISA department: a new computer room with 18 high-end workstations equipped with 1080TI NVIDIA GPU cards (accessible to students for practical labs and MFEs)
- At the ULB level: a replacement of the VEGA compute cluster is planned for the beginning of 2020. This replacement, dubbed VEGA2, will have support for Big Data compute software. Currently, VEGA2 is planned to consist of 2500 cores, 1 PB of storage, and 10 TB of RAM.

## 5 Contractual research opportunities

CoDE and LISA have a track record of attracting research funding related to computer science, big data, and business intelligence, as summarized below. The solicited chair is expected to further advance this track record, thereby strengthening the EPBs leadership.

**At the EU level.**

- Two ERC grants in swarm robotics at CoDE, awarded to M. Dorigo ("E-Swarm: Engineering Swarm Intelligence Systems") and M. Birattari ("DEMIURGE: automatic design of robot swarms").
- Erasmus Mundus Joint Master in Big Data Management and Analytics (BDMA) at CoDE.
- Erasmus Mundus Joint Doctorate "Information Technologies for Business Intelligence Doctoral College" (IT4BI-DC) at CoDE.
- Erasmus+ Joint Master Programme "Information Technologies for Business Intelligence" (IT4BI) at CoDE.

**At the national level.**

- FNRS: 4 FNRS research mandates 4 FNRS research projects, and 2 FRIA funded in the last 6 years at CoDE. One FNRS project related to deep learning in medical imaging ongoing at LISA.
- Innoviris (Brussels Region): 13 projects funded in the last 6 years at CoDE, 2 projects related to deep learning at LISA.
- Walloon Region: 6 projects funded in the last 6 years at CoDE, 1 project related to deep learning in medical imaging at LISA.

**Contract research with industry.** CoDE is currently working on industrial projects in the field of mobility (with Joyn Joyn, STIB), IOT (with Degetel), healthcare (CluePoints, Kantify), and Big Data (Dataa, Wavemaker).

## 6 Targeted opening of the position

To ensure continuity of the courses previously taught by Toon Calders within the context of the Erasmus Mundus Joint master in Big Data Management and Analytics (BDMA), we ask that that position is opened with priority, such that the selected candidate can start on October 1, 2020.

# 7 Support

This proposal is supported by:

- The entire CoDE department: H. Bersini, M. Birattari, Y. De Smet, M. Dorigo, T. Stutzle, S. Vansummeren, E. Zimanyi.
- The LISA department: O. Debeir