The Return of the Knowledge Scientist

An integrative curriculum to unlock the data-driven organization

Daniel Kapitan, Joran Lokkerbol, George Fletcher, Paul Groth, Juan Sequeda

01 September 2022

Introduction

In our appeal for the return of the Knowledge Scientist (Fletcher, Groth, and Sequeda (2020)) we have argued the need to improve reliance on data to unlock data-driven organization. In our view, to 'rely on data' can be unpacked into the following five properties:

1. Reliable data is clean data.
2. Reliable data is grounded in shared meaning spaces.
3. Reliable data is data in context. Clean data with shared meaning is not enough.
4. Reliable data is data accessible in a standardized format.
5. Reliable data is maintained.

Fletcher, George, Paul Groth, and Juan Sequeda. 2020. "Knowledge Scientists: Unlocking the Data-Driven Organization." arXiv. https://doi.org/10.48550/arXiv.2004.07917.

We have argued that the lost art of knowledge identification, knowledge elicitation and knowledge specification needs to be reinstated in industry to complement current data science practices. We have called for a return of the knowledge scientist in organizations, development of new courses for educating knowledge scientists and research to study the tripartite relationship between data/knowledge models, their corresponding query languages and the people both using and producing reliable data.

In this paper, we outline an integrative curriculum for a post-graduate course on knowledge science, thereby rising to our own call. Given that a knowledge scientists needs a diverse set of skills, ranging from data engineering, knowledge acquisition, machine learning and human-computer interactions, we evaluate the different frameworks and concepts that are commonly used for teaching each of these discplines. We subsequently propose how these frameworks can conceptually integrated into a cohesive and logical curriculum. Finally, we discuss the feasibility of the proposed curriculum by reflecting on nascent knowledge science practice in Dutch government and healthcare organizations. These reflections are based on our experience with Professional Education at Jheronimus Academy of Data Science (JADS PE) and our work in advising organizations on unlocking the data-driven organization, particularly for government and healthcare.

Overview of relevant frameworks

Our first challenge in defining a knowledge science curriculum is to arrive at a practical and meaningful scope of which existing domains and frameworks i.e. subjects are to be included. Table 1 lists the subjects and corresponding frameworks that we have included in our analysis.

Domain	Framework and/or key textbook	Relevant topics
Enterprise Architecture	TOGAF	Business architecture, Data architecture
Data Management	DAMA-DMBOK	Virtually all 10 topics of DMBOK 2 wheel (see figure 1)
Knowledge Graphs	hogan2021knowledge	Foundational theory of graphs as an introduction to its specific application in e.g. fact-based modeling, property graphs etc.
Information modeling	...	Encompassed generic tools and techniques (layered models, ontologies) and domain-specific application (buildings, accounting, healthcare)
Data engineering	reis2022fundamentals	Data processing pipelines, ETL/ELT, data warehousing

Table 1: Domains and frameworks included in the analysis.

Current trends and challenges in data-drive innovation

The importance of these properties of reliable data is similar to Ng's several the appeal for a more data-centric approach to AI ("Data-Centric AI Resource Hub," n.d.). Data-Centric AI (DCAI) represents the recent transition from focusing on modeling to the underlying data used to train and evaluate models. Increasingly, common model architectures have begun to dominate a wide range of tasks, and predictable scaling rules have emerged. While building and using datasets has been critical to these successes, the endeavor is often artisanal – painstaking and expensive. The community lacks high productivity and efficient open engineering tools to make building, maintaining, and evaluating datasets easier, cheaper, and more repeatable. The DCAI movement aims to address this lack of tooling, best practices, and infrastructure for managing data in modern ML systems (**neurips2021?**).

"Data-Centric AI Resource Hub." n.d. *Data-Centric AI Resource Hub*. https://datacentricai.org/.

neurips2021

Another recent and relevant development pertains to upcoming standard to ensure data is accessible in a standardized format. Explain:

- relates to linked data field of research
- recent advances to operationalize this via FAIR principles and relate to the FAIR Digital Object Framework

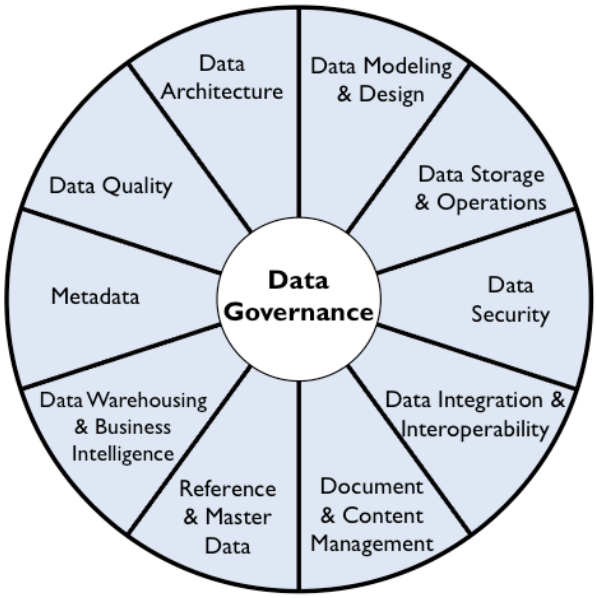

Figure 1: Wheel of the Data Management Body of Knowledge (DMBOK), second edition

How did we come to this, anyway?

- long-standing promises from linked data, knowledge engineering which didn't quite pay-off yet, each with their own 'winter'
- competing standards, no overarching end-to-end workflow yet. In theory UML2.0 open standaard, but transferring models between different UML tools is painful. Archimate is a better open standard, but less suitable for data modeling. Following MIM we want to include attributes in conceptual data model. From another perspective, linked data formats (RDF, Turtles) are expressive in capturing semantic and conceptual levels, but still a gap to translate that in the toolchain
- Historically, datawarehousing has been dominated by the dimensional modeling pattern by Kimball. Although concept of data vault has been around for a while, this approach is difficult to implement. Very labour intensive, new generation of code generation tools. But still now integrated workflow to go from semnatic –> conceptual –> logical –> physical

The return of the knowledge scientist: new knowledge science practices in industry

Work of Ronald Damhof for Dutch government ornganizations

Summarize Damhofs work here. Wetsanalyse by Lokin.

FAIR and FHIR in healthcare

Summarize current work in healthcare. Seems like ontologies and semantic models are finally picking up in mainstream. E.g. work Allen Institute for SciSpaCy

The return of the knowledge scientist: a revised curriculum

Elements of curriculum

- Methods for semantic knowledge engineering: CommonKADS, Cogniam, PROTEGE
- Methods for conceptual knowledge engineering:
  - Linked Data, UML2.0 …
  - FDOF, …
- Methods for logical layer:
  - Data vault
  - Graph databases: SPARQL, new GQL standard