The Return of the Knowledge Scientist

A revised curriculum to pay off the data debt of organizations and democratize data

Daniel Kapitan, Joran Lokkerbol, George Fletcher

8th July 2022

Table of contents

Introduction

Building on our knowledge science manifesto (Fletcher, Groth, and Sequeda 2020) we expand our view on why such a role is required, how nacent knowledge science practices are currently conducted in different domains, and what a revised curriculum looks like. First we sketch current challenges in data-driven

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

innovation: what are key challenges, both from academic research and industry practices. Second, we analyze two cases of nascent knowledge science practises. By examining these practices we aim to distill generic patterns to inform how a revised knowledge science curriculum should look like. Finally, we lay out the revised knowledge science curriculum.

Current trends and challenges in data-drive innovation

As laid out in Fletcher, Groth, and Sequeda (2020) there is a clear need to improve reliance on data to unlock data-driven organization, unpacking it in 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.

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

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