

---

# CONCEPT MINING FOR DATA MANAGEMENT (AN INTRODUCTION)

Swarupa Hardikar

HAN University of Applied Sciences / Radboud University

1st year PhD candidate

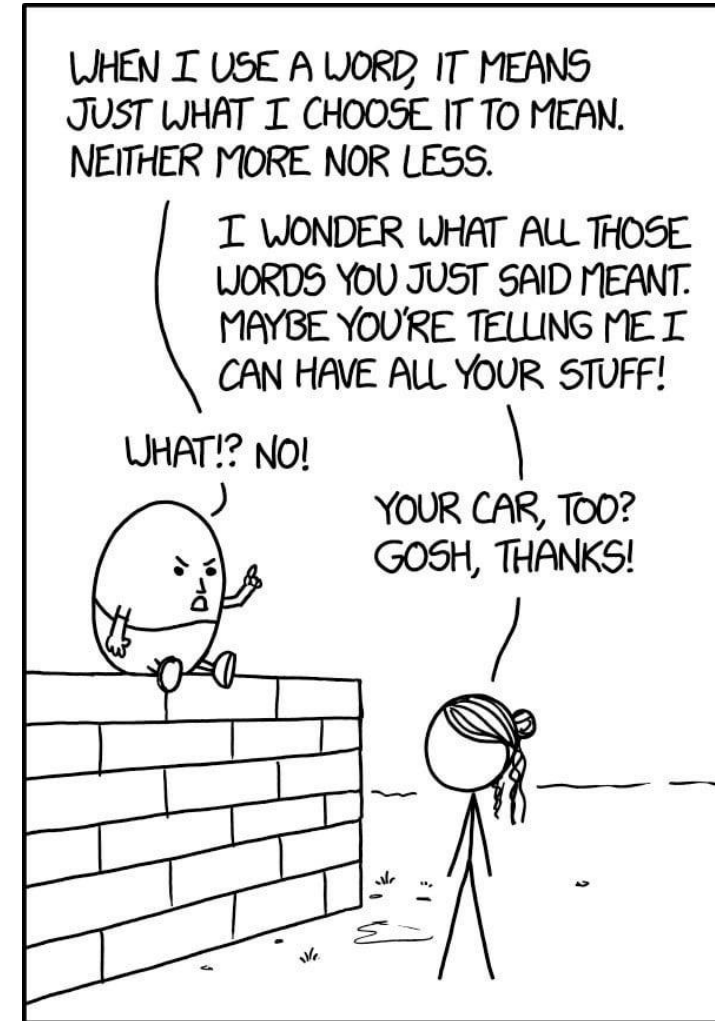
Promoter: Arjen de Vries

Supervisors: Stijn Hoppenbrouwers (HAN), Maya Sappelli (HAN)

---

# THE ISSUE: DATA SEMANTICS

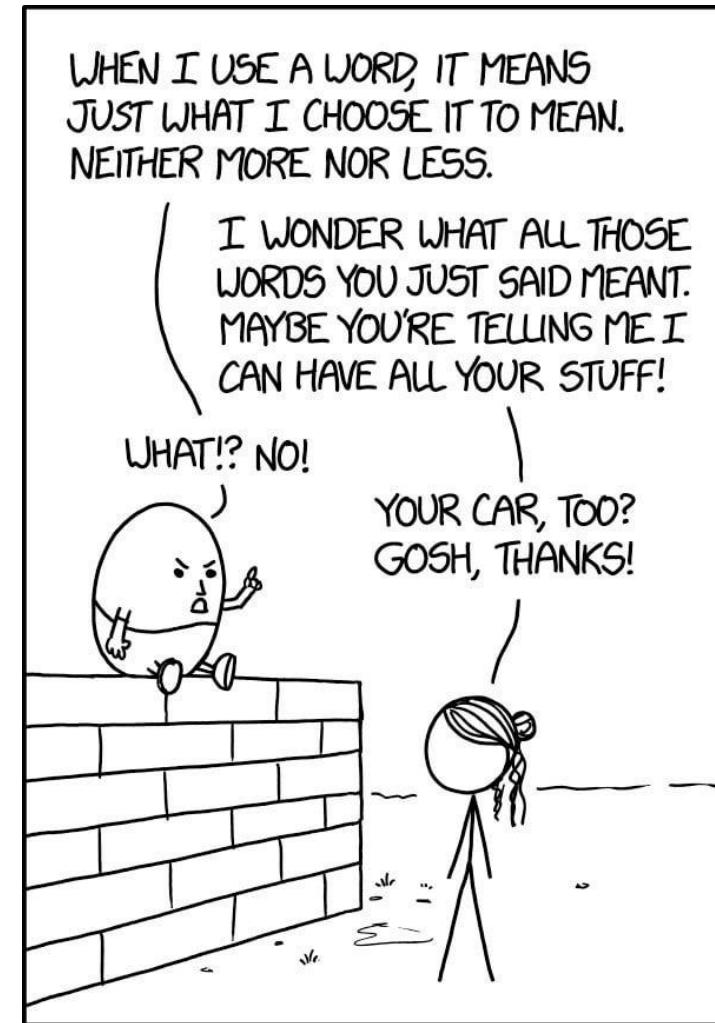
I.e., data meaning



---

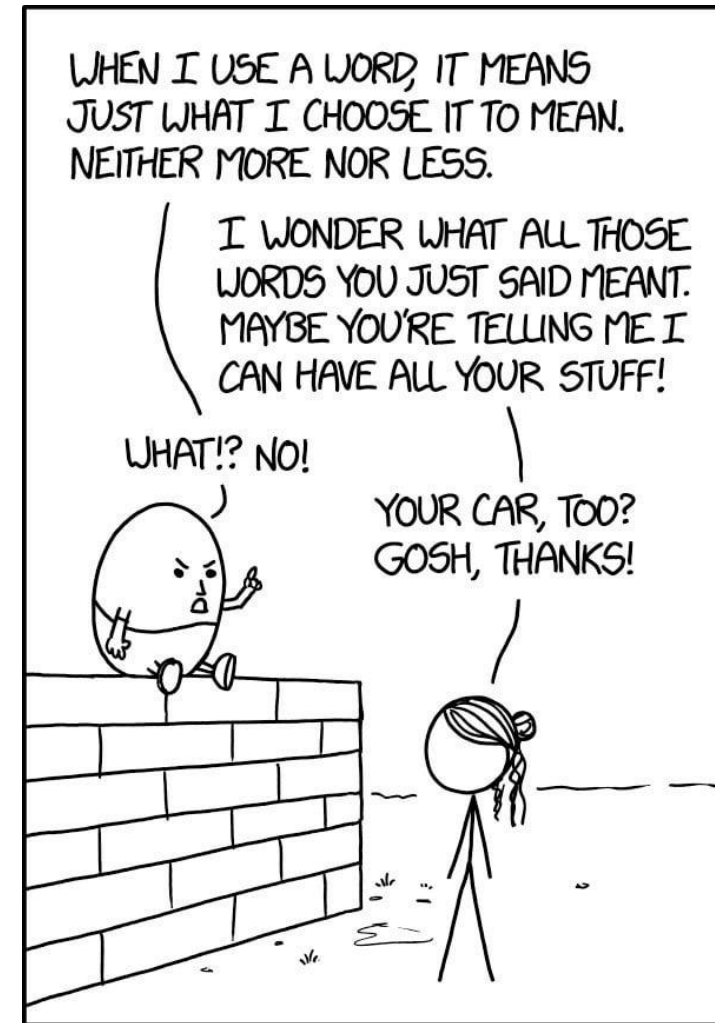
# THE ISSUE: DATA SEMANTICS

- people within different organisational domains use words/terms differently – say, a model
- This is prevalent especially in the case of domain-specific terminology



# THE ISSUE: DATA SEMANTICS

- The problem: the loss of informational value due to misinterpretation of data
- The need: to be able to compare the **intended meaning** of the data owner with the **perceived interpretation** of the data user



---

# HOW DO WE APPROACH THE SEMANTIC GAP?

This is where "concept mining" may come in

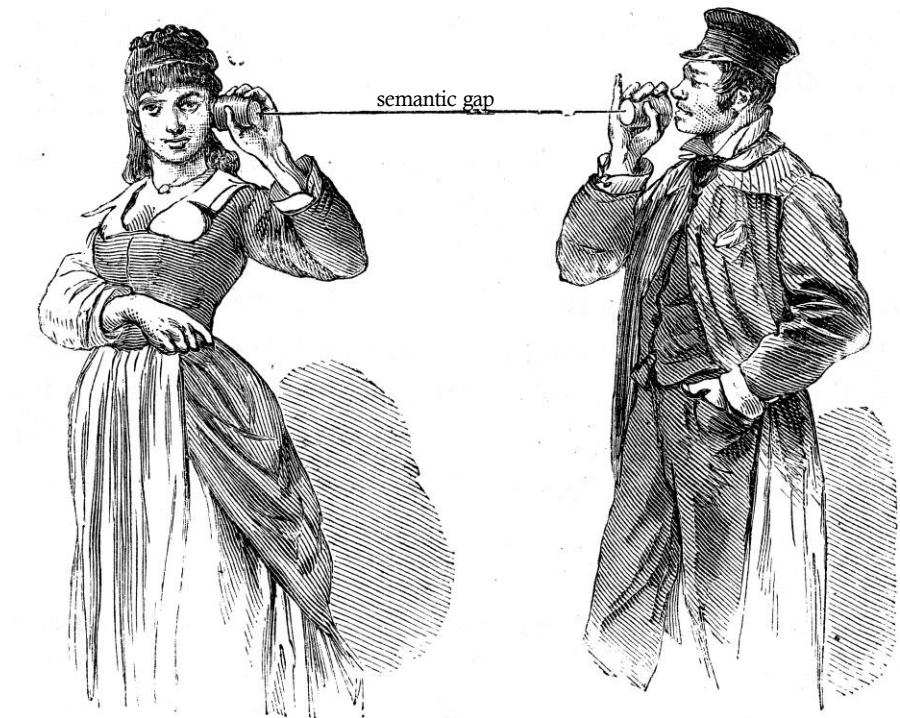


Fig. 7.

---

# A POSSIBLE SOLUTION: CONCEPT MINING (BUT WHAT IS CONCEPT MINING?)

- Concept mining is an approach using
- corpora of unstructured or semi-structured domain data,
- which can be explored and analysed
- to systematically extend or produce
- a pluriform **lexical asset** containing
- information on the **concepts of interest**
- and how they are **perceived by data users**.



Fig. 7.

---

# RESEARCH QUESTION

"How can a concept mining system be designed, applied, and evaluated to reconcile the intended meaning of the author with the perceived meaning of the reader, in order to prevent the loss of informational value caused by the misinterpretation of domain-specific data?"

Sub-questions:

1. What functionality and components would a concept mining system require to support the (semi-)automatic formulation of domain-specific lexical assets?
2. What frameworks and metrics can be used to evaluate concept mining systems in generating domain-specific lexical assets?
3. How effectively does such a concept mining system support data management professionals in practice?

---

# ATTACK AREAS

Things the production/extension of the lexical asset aims to address, using text mining and other techniques:

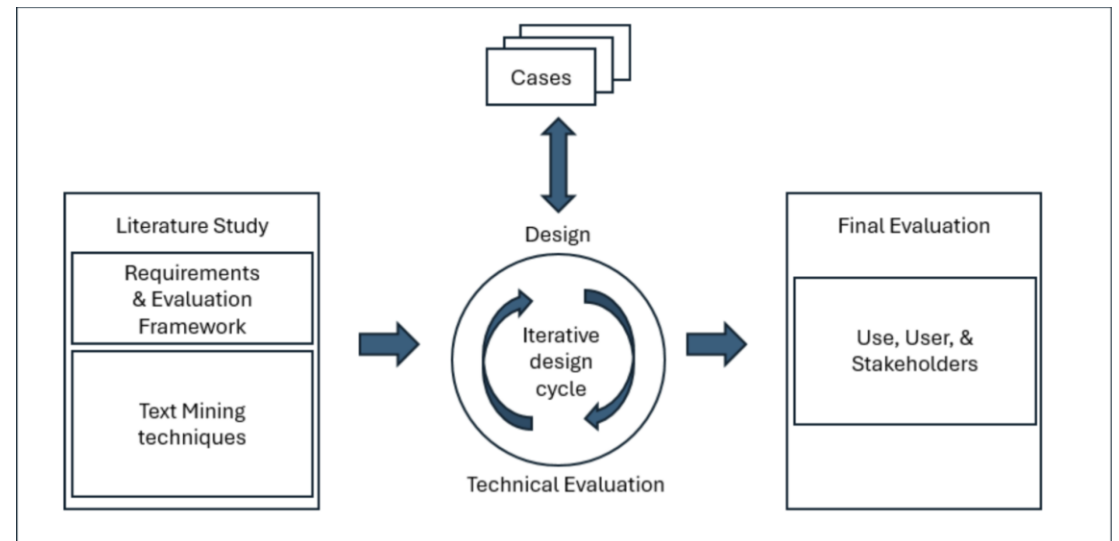
- Homonyms/synonyms widely in use within organisations
- Definitions/descriptions
- Entity-relationship structures



---

# METHOD: DESIGN SCIENCE RESEARCH

- Building an artifact – a toolbox – to enable concept mining
- This consists of:
  - Searching for and identifying potential techniques to satisfy the requirements
  - Applying and subsequently testing these techniques on the data,
  - Finally: evaluating the system quantitatively and qualitatively



---

# STUDY 1: ABBREVIATION DISAMBIGUATION

---

# STUDY 1: ABBREVIATION DISAMBIGUATION

FYI, per our QBR, the KPI delta on our OKRs is trending south, so let's circle back EOD and loop in the SME from the PMO. If we don't get alignment ASAP, we risk missing the SLA and that'll trigger a CYA fire drill with the CFO and the CEO—possibly even an FYI to the BOD. BTW, our CRM flagged a VIP FYI from a key MQL turned SQL who wants a demo ROI ASAP, so ping the AE/SE duo and prep the deck. TL;DR: synergy or chaos. LMK.

---

# STUDY 1: ABBREVIATION DISAMBIGUATION

**For Your Information** (FYI), per our **Quarterly Business Review** (QBR), the **Key Performance Indicator** (KPI) delta on our **Objectives and Key Results** (OKRs) is trending south, so let's circle back by **End of Day** (EOD) and loop in the **Subject Matter Expert** (SME) from the **Project Management Office** (PMO). If we don't get alignment **As Soon As Possible** (ASAP), we risk missing the **Service-Level Agreement** (SLA), and that'll trigger a **Cover Your Assets** (CYA) fire drill with the **Chief Financial Officer** (CFO) and the **Chief Executive Officer** (CEO)—possibly even an FYI to the **Board of Directors** (BOD). **By The Way** (BTW), our **Customer Relationship Management** system (CRM) flagged a **Very Important Person** (VIP) FYI from a key **Marketing Qualified Lead** (MQL) turned **Sales Qualified Lead** (SQL) who wants a demo **Return on Investment** (ROI) ASAP, so ping the **Account Executive** (AE)/**Sales Engineer** (SE) duo and prep the presentation deck. **Too Long; Didn't Read** (TL;DR): synergy or chaos. **Let Me Know** (LMK).

---

# USE CASE: ABBREVIATION DISAMBIGUATION WITH ALLIANDER

- Extract abbreviations from the company's internal documents and compare them against a **provided list of abbreviations** in use
- Classify the detected abbreviations as internal (company-specific) or domain-specific or global (widely used across industries and domains)
- A multilingual/low resource problem: Documentation is in English, Dutch and Frisian

---

# USE CASE: WORK DONE SO FAR

- Annotation of 300+ abbreviations according to usage, language and existence of corresponding Wikipedia page
- Test set of 15 sentences containing abbreviations (internal, domain-specific and global)
  - Tested primarily with MS Copilot (can search the internet)
  - Global examples can typically be identified
  - Domain-specific - somewhat
  - Internal...typically not

---

# EXAMPLE

## GPLK: GEPANTSEERDE PAPIER LOODKABEL

De GPLK was oud en in zeer slechte staat. Dit is een risico dat volgens de CROW 500 in de ontwerpfase moet worden vastgesteld bij het lokaliseren van de bestaande kabels en leidingen, inclusief het maken van proefsleuven. De slechte status van de kabel moet zijn vastgelegd in de risico-inventarisatie en het maatregelenplan.

In the above statement, based on the context, what is the full-form of 'GPLK'?



In the context of the CROW 500 guidelines, 'GPLK' stands for **Gas, Petroleum, and Liquid Kerosene** pipelines. These guidelines focus on preventing excavation damage to underground cables and pipelines, and ensuring safety during construction and maintenance activities <sup>1</sup> <sup>2</sup>.

Is there anything else you need help with regarding CROW 500 or related safety procedures?

---

# INITIAL EXPLORATION: ANGLES

1. LLM-aided disambiguation
  - Analysing if metadata input aids this
2. A comparison of abbreviation detection techniques
3. Data management given abbreviation lists
  - Analysis of the evolution + increasing frequency of terms



---

## ALSO IN THE WORKS

Interviews with data managers

Literature review



---

# SUGGESTIONS, FEEDBACK?

