ALMA MATER STUDIORUM
UNIVERSITÀ DI BOLOGNA

FEDORA

Open Physics Hub

Epistemological implications of different methodological approaches in textual data analysis

ESERA conference – 29th August, 2023

ESERA 2023 Cappadocia

Martina Caramaschi(1) - Andrea Zanellati(2) - Olivia Levrini(1)

(1) Department of Physics and Astronomy - University of Bologna

(2) Department of Computer Science - University of Bologna

Presentation outline

- Introduction
- Research focus: textual data in SER
- Methodological issue
- Aims and research question
- Method
- Results
- Evaluation and future developments

Introduction

Socio-economic fields:

- Last two decades
- **Big Data era** (Klašnja-Milićević et al., 2017)
  - Large amount
  - Great variety
  - Data Science
  - New technological tools

ALMA MATER STUDIORUM
UNIVERSITÀ DI BOLOGNA

Introduction

Educational field:

- Datafication (Jarke & Breiter, 2019)
- E-learning platforms (e.g. MOOCs)
- Collection of data
  - On all levels
  - About all processes

Introduction

Socio-economic fields:

- Big Data era (Klašnja-Milićević et al., 2017)
  - Large amount
  - Variety
  - New tool for analysis

Educational field:

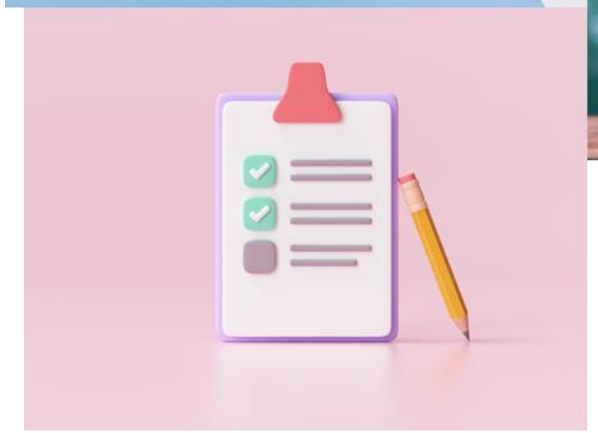
- Datafication (Jarke & Breiter, 2019)
- E-learning platforms (e.g. MOOCs)
- Collection of data
  - On all levels
  - About all processes

Data-intensive computational methods for data analysis

Focus of our study

Textual data in Science Education Research (Fesler, 2019)

- ➢ Data features:
  - Medium-sized volume
  - Context-specific
- ➢ Data sources:
  - Interviews
  - Textbooks
  - Written comments
  - Written tasks
  - Questionnaire open answers

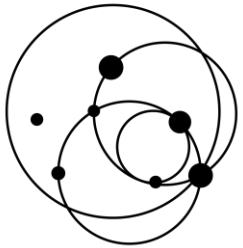


Our dataset

Individual essays in Italian language (Barelli et al., 2022)

- Projects: PLS & FEDORA
- Collection years: 2018-2021
- Students age: 17-19 y. o.
- Essays quantity: > 220
- Essays length: 1 page each (around 400 words)

Piano Lauree Scientifiche

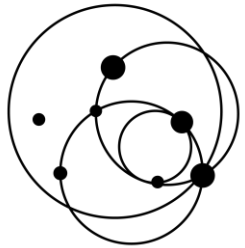


Project aim: to foster youngsters' ways of coping with images of the future through science education

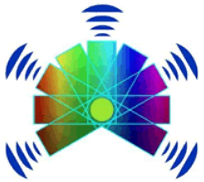
ALMA MATER STUDIORUM
UNIVERSITÀ DI BOLOGNA

Our dataset

Individual essays in Italian language (Barelli et al., 2022)



- Projects: PLS & FEDORA
- Collection years: 2018-2021
- Students age: 17-19 y. o.
- Essays quantity: > 220
- Essays length: 1 page each (around 400 words)



- Essay task: "*your future life in 2040*"
  - Where they imagine to live
  - Kind of life
  - Types of problems (in their daily life, community and society)
  - Possibilities and new opportunities
  - Objects, the city and the environment
  - Their social life

ALMA MATER STUDIORUM
UNIVERSITÀ DI BOLOGNA

Research issue

Textual data in SER **+** **Digitalisation**

- Data features:
  - Medium-sized volume
  - Context-specific

Data science techniques availability

Reflection on methods

?

Traditional Qualitative method

Vs

Data-intensive method

Our aim

Developing an epistemological and methodological reflection

- Textual data in SER (medium volume, context-specific)
- Qualitative and data-intensive computational methods

…NOT a mere technical reflection to establish which method is the best…

Research Question

When analyzing a corpus characterized as above, how is the extracted knowledge characterised by the choice of a qualitative or a computational data-intensive method?

Method

Case study

- "Essays dataset" analysis
- Two separate methods

Methods comparison

- Effort to produce the knowledge
- Quality of the knowledge achieved
- Method limits and strengths

Case study: data analysis with method #1

***Reflexive* Thematic Analysis** (qualitative method) (Braun & Clarke, 2019)

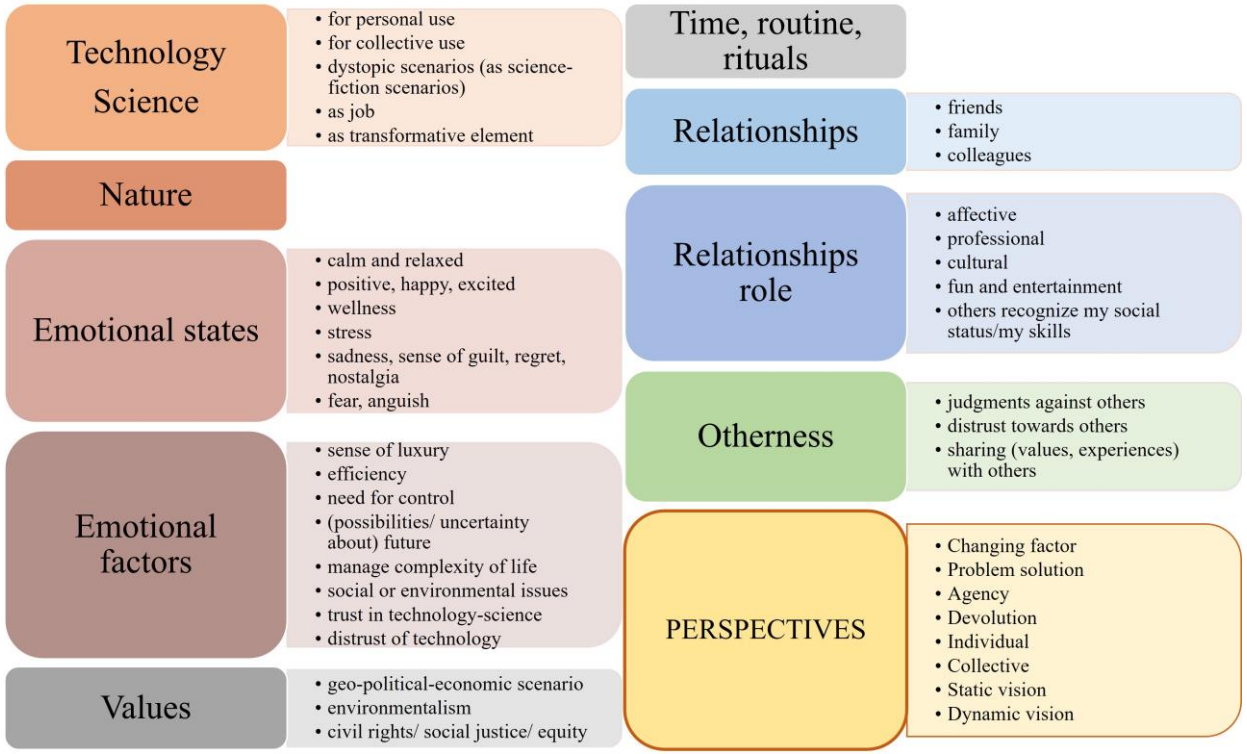
- Bottom-up process
- Immersion into the data
- Triangulation phases
- Sub-themes refinement
- Sub-themes clustering

} Finding shared themes among analysers

Case study: results

Thematic Analysis

- "The big matrix"
  - 9 main themes
  - Several sub-themes
  - Different perspectives
- Deep researchers' immersion and familiarity with data

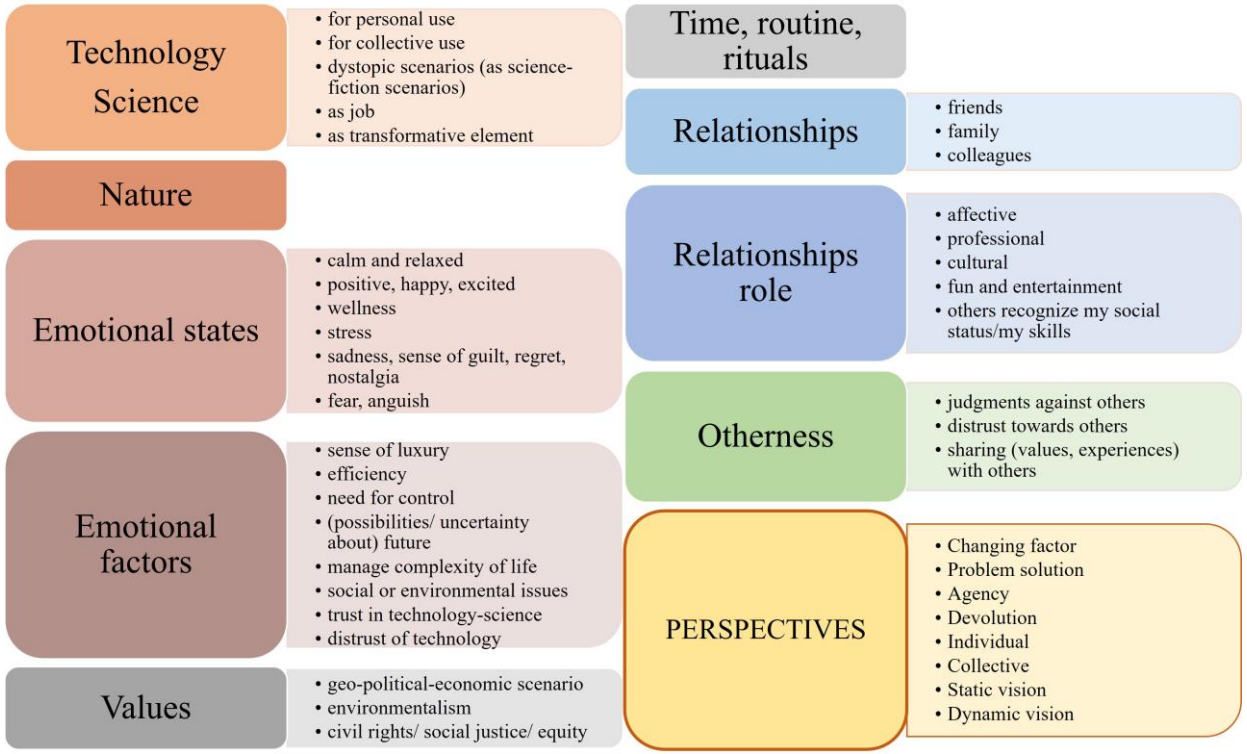


Themes and sub-themes found using a Thematic Analysis of the students' essays

Case study: limits

Thematic Analysis

- "The big matrix"
- Time-consuming: 1/4 essays analysed
- Loop of refinement of themes → No saturation
- No theoretical pattern



Themes and sub-themes found using a Thematic Analysis of the students' essays

Case study: limits

Thematic Analysis

- "The big matrix"
- Time-consuming: 1/4 essays analysed
- Loop of refinement of themes → No saturation
- No theoretical pattern

Technology Science	• for personal use • for collective use • dystopic scenarios (as science-fiction scenarios) • as job • as transformative element
Nature	
Emotional states	• calm and relaxed • positive, happy, excited • wellness • stress • sadness, sense of guilt, regret, nostalgia • fear, anguish
Emotional factors	• sense of luxury • efficiency • need for control • (possibilities/ uncertainty about) future • manage complexity of life • social or environmental issues • trust in technology-science • distrust of technology
Values	• geo-political-economic scenario • environmentalism • civil rights/ social justice/ equity
Time, routine, rituals	
Relationships	• friends • family • colleagues
Relationships role	• affective • professional • cultural • fun and entertainment • others recognize my social status/my skills
Otherness	• judgments against others • distrust towards others • sharing (values, experiences) with others
PERSPECTIVES	• Changing factor • Problem solution • Agency • Devolution • Individual • Collective • Static vision • Dynamic vision

Themes and sub-themes found using a Thematic Analysis of the students' essays

Exit from the loop?

1. Step back to look at data from a new distance
2. Searching for new lenses of analysis and top-down analysis
3. Awareness of theoretical framework (polarization)

Case study: data analysis with method #2

Latent Semantic Analysis (Blei, 2011)

- Topic modeling method
- Data-intensive → bottom-up process
- Mathematical method: singular values decomposition
- Semi-automatic computational method:
  - Customised pre-processing steps (lemmatization, tokenization, stopword removal, extra stopwords definition and removal)
  - LSA algorithm
  - Mathematical criteria to determine the number of topics

Case study: results and limits

Latent Semantic Analysis

- 7 lists of correlated words (patterns)
- Each list represents a topic contained into the essays
- Too generic lists
  - Difficult interpretation
- Sensitive to small task changes
- All essays included in the analysis

Technology and developments

'technology', 'issue', 'want', 'technological', 'developement'…

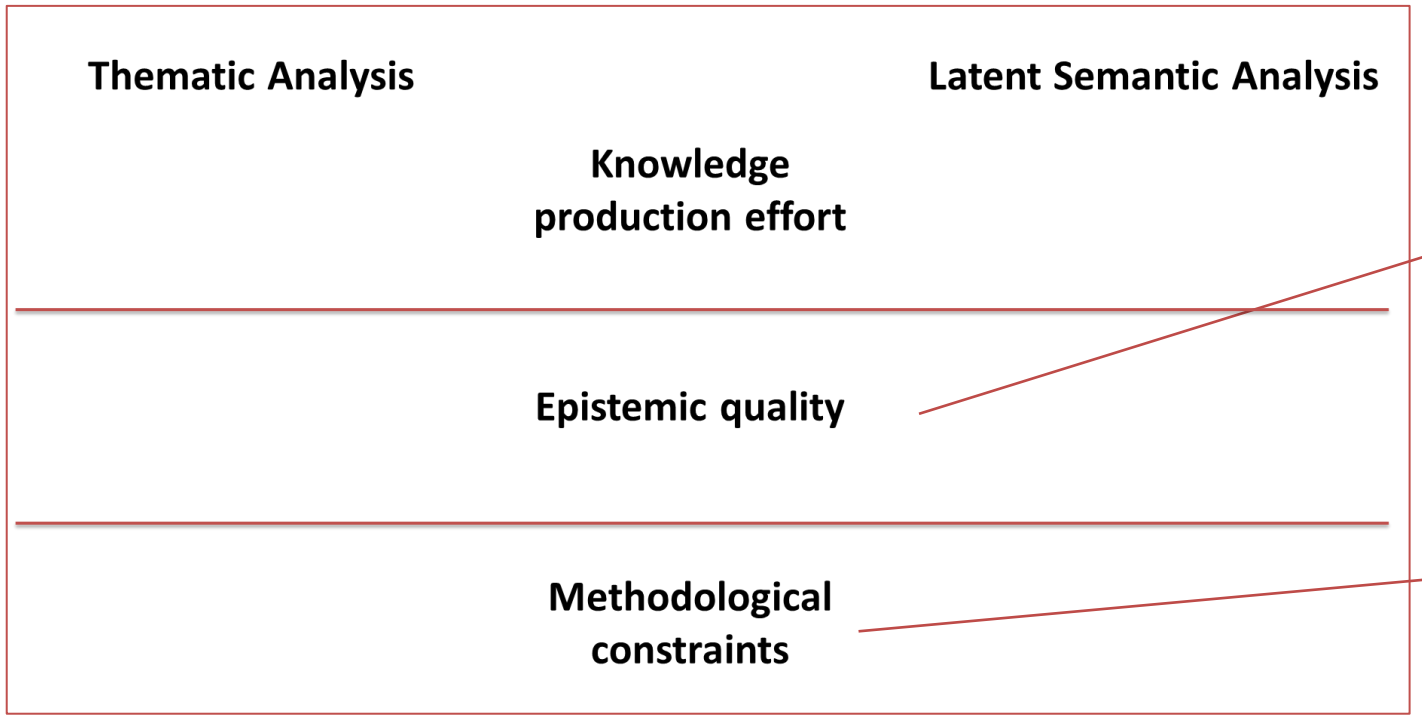
Wishes - hopes

'imagine', 'hope', 'want', 'be', 'live', 'pleasure', 'life', 'future'…

Examples of lists of words (translated from italian) for two different topics

Case studies comparative analysis

- Grid of analysis elaboration
  - Lenses of analysis choice to address our RQ
- Separate methods description with respect to the grid lenses
- Critical evaluation and comparison



- Originality
- Interpretability
- Generalizability
- Applicability to research or teaching

- Benefits
- Limits

Results

Thematic Analysis

- Iterations
- Triangulations
- Nine analysers

Knowledge production effort

Latent Semantic Analysis

- Pre-processing
- Computational skills
- One analyser

Epistemic quality

Methodological constraints

Results

Thematic Analysis

- Iterations
- Triangulations
- Nine analysers

Knowledge production effort

Latent Semantic Analysis

- Pre-processing
- Computational skills
- One analyser

Epistemic quality

Methodological constraints

Results

Thematic Analysis		Latent Semantic Analysis
• Iterations • Triangulations • Nine analysers	**Knowledge production effort**	• Pre-processing • Computational skills • One analyser
Good Interpretability Generalisability Low Originality Applicability	**Epistemic quality**	Good Generalisability Interpretability Low Originality** Applicability
	Methodological constraints	

Results

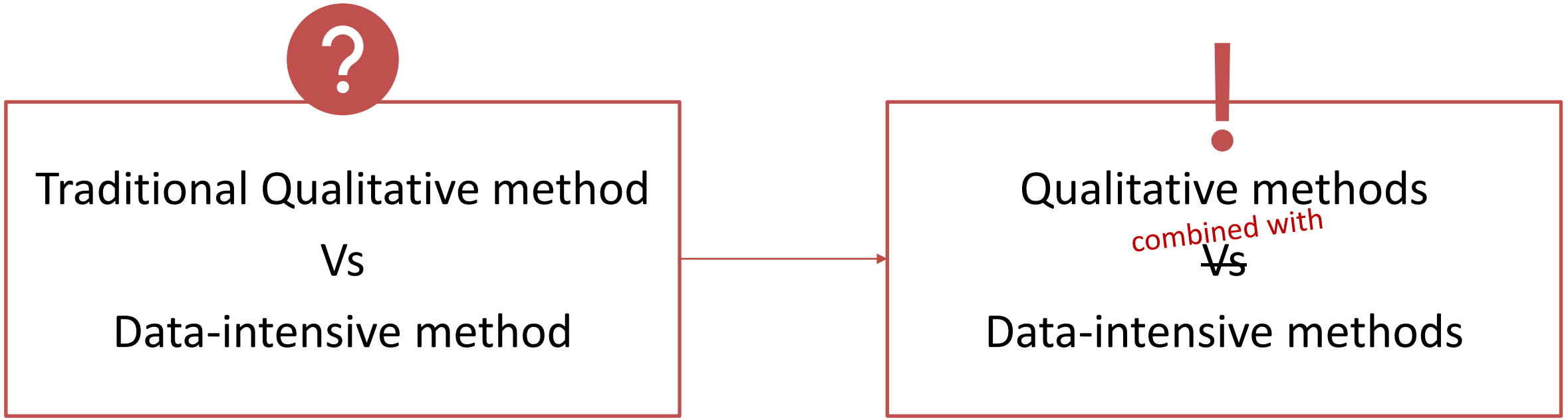
Thematic Analysis		Latent Semantic Analysis
• Iterations • Triangulations • Nine analysers	**Knowledge production effort**	• Pre-processing • Computational skills • One analyser
Good Interpretability Generalisability Low Originality Applicability	**Epistemic quality**	Good Generalisability Interpretability Low Originality** Applicability
PRO Deep immersion in the data CONS Time-consuming	**Methodological constraints**	PRO Scalability CONS Suits better with larger dataset

Final evaluation

Thematic Analysis		Latent Semantic Analysis
✓	Knowledge production effort	✓
COMPLEMENTARY	Epistemic quality	COMPLEMENTARY
	Methodological quality	

Possible future directions of work

Combination of the strengths → Mixed methods (Johnson & Onwuegbuzie, 2004)



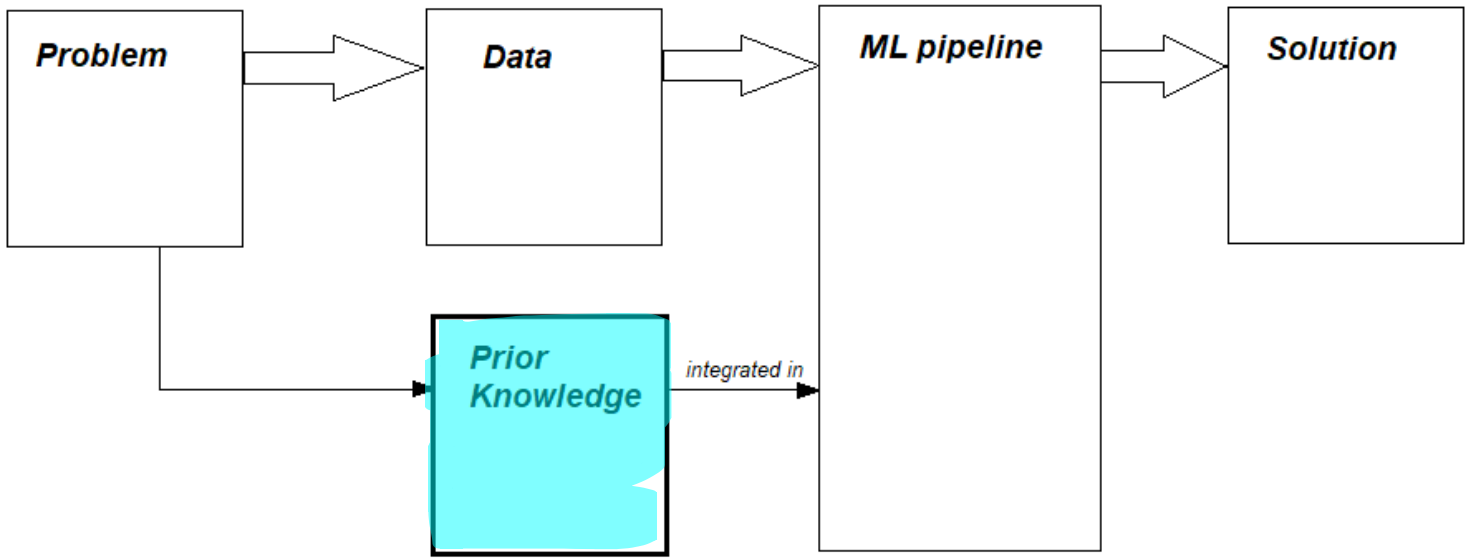
Possible future works

Expert contribution

+

Computational methods

→ Informed Machine Learning (Von Rueden et al., 2021)



References (1/2)

Barelli, E., Tasquier, G., Caramaschi, M., Satanassi, S., Fantini, P., Branchetti, L., & Levrini, O. (2022). Making sense of youth futures narratives: Recognition of emerging tensions in students' imagination of the future. Frontiers in Education, 7, 1-17. doi: 10.3389/feduc.2022.911052

Blei, D. (2011). Introduction to Probabilistic Topic Models. Communications of the ACM, 55, 77-84. doi: 10.1145/2133806.2133826

Braun, V. & Clarke, V. (2019). Reflecting on reflexive thematic analysis. Qualitative Research in Sport, Exercise and Health, 11:4, 589-597. doi: 10.18608/jla.2015.22.2

Fesler, L., Dee, T., Baker, R., & Evans, B. (2019). Text as Data Methods for Education Research. Journal of Research on Educational Effectiveness, 12:4, 707-727. doi: 10.1080/19345747.2019.1634168

Johnson, R., & Onwuegbuzie, A. (2004). Mixed Methods Research: A Research Paradigm Whose Time Has Come. Educational researcher, 33, 14. doi: 10.3102/0013189X033007014

Jarke, J., & Breiter, A. (2019). Editorial: the datafication of education. Learning, Media and Technology, 44, 1, 1-6. doi: 10.1080/17439884.2019.1573833

References (2/2)

Klašnja-Milićević, A., Ivanovic, A., & Budimac, M. (2017). Data science in education: big data and learning analytics. Computer Applications in Engineering Education, 25, 1068–1078. doi: 10.1002/cae.21844

Landauer, T. Foltz, P.W., & Laham, D. (1998). Introduction to Latent Semantic Analysis. Discourse Processes, 25, 259–284. doi: 10.1080/01638539809545028.

Von Rueden, L., Mayer, S., Beckh, K., Georgiev, B., Giesselbach, S., Heese, R., ... & Schuecker, J. (2021). Informed Machine Learning–A taxonomy and survey of integrating prior knowledge into learning systems. IEEE Transactions on Knowledge and Data Engineering, 35(1), 614-633.

ALMA MATER STUDIORUM
UNIVERSITÀ DI BOLOGNA

Thanks for your attention

Martina Caramaschi, Andrea Zanellati, Olivia Levrini

Department of Physics and Astronomy - University of Bologna

martina.caramaschi2@unibo.it

olivia.levrnini2@unibo.it