UNIVERSITY OF TORONTO

Clinical Text Data for Patient Phenotyping

Current Methods and Future Directions

Christopher Meaney

Biostatistician
Department of Family and Community Medicine
University of Toronto
christopher.meaney@utoronto.ca

October 16, 2025

UNIVERSITY OF TORONTO

Phenotyping Routinely Collected Health Data

- Phenotype: any observable (measurable) patient state/behavior/action.
- E.g. phenotype diabetes (labs, meds, billing, referral/consult, notes, etc.)
- Phenotypes derived from RCHD can be outcomes, covariates, incl/excl criterion, etc.
- Related areas of research:
  - Diagnostic accuracy studies.
  - Validation studies
  - Case-finding algorithms.
  - Code set engineering.

Goal of this talk:

Review of current and emerging **methods** for phenotyping routinely collected clinical text data.

- Out of scope for this talk:
  - Phenotype designs: cross-sectional, stratified sampling, advanced designs, etc.
  - Sample size and study planning.
  - Advanced issues: active learning, semi-sup, no labels, silver standard labels, etc.
  - Phenotype evaluation, imperfect phenotypes: misclassification, bias correction, etc.

UNIVERSITY OF TORONTO

Trends in EHR Phenotyping Studies

- Rule based methods popular. Interpretable. Fast. Challenging to code?
- Supervised ML gains popularity. Flexible. Effective. Black-box?
- Non-traditional designs considered to reduce time and cost of development/evaluation.
- Newer methods incorporate external knowledge bases and ontologies.
- Emphasis towards shareable phenotypes algorithms and/or modular/transportable pipelines.
  - PheKB, eMERGE, OHDSI, CALIBRE, etc.

Review Methods (Banda et al)

Primary method	Number of papers
Rule-based	19
Natural language processing	35
Standard machine learning	25
Learning with noisy data	11
Unsupervised phenotype discovery	11
Hybrid approaches	3
Collaborative frameworks	10
Total	89

Review Data Structures (Alzoubi et al)

Table 2. Categorise of feature extraction techniques.

Feature Extraction Method	Papers
Structured only	[18,35,38,90–96]
BoW only	[30,49]
Keyword search only	[29,54,97]
Concept extraction only	[31,59,69,72,73,75,98]
Structured + BoW	[42,50,52,99]
Structured + Keyword search	[6,16,17,32,55,58,60]
Structured + Concept extraction	[15,28,36,37,48,68,70,71,74,100]

Banda et al. (2018). Phenotpying: Rule Based vs. ML. Ann Rev Biomedical Data Sci.
Alzoubi et al. (2019). Auto Phenotyping EHRs: A Review. MDPI Electronics.

UNIVERSITY OF TORONTO

Content of Talk

- Rule-set mining using regular expressions and contextual modification.
- Supervised methods for phenotyping text data (document classification, NER, etc.).
- Unsupervised methods for discovering latent patient phenoptypes from text data.

Emerging Ideas

- LLMs and foundation models for extending/complementing phenotyping methods.

UNIVERSITY OF TORONTO

Rule Set Mining Using REGEX and Contextual Modification

- Regular expression: a sequence of characters specifying a match pattern in text.
- Contextual modification: negation, uncertainty, other experiencer, etc.

UNIVERSITY OF TORONTO

Example: Context Aware REGEX for COVID-19 Phenotyping

A Natural Language Processing System for National COVID-19 Surveillance in the US Department of Veterans Affairs

Alec B Chapman[1,2], Kelly S Peterson[1,2], Augie Turano[3], Tamára L Box[4], Katherine S Wallace[5], Makoto Jones[1,2]

[1] Veterans Affairs (VA) Salt Lake City Health Care System
[2] Division of Epidemiology, University of Utah
[3] VA Office of EHR Modernization
[4] VA Office of Clinical Systems Development and Evaluation (CSDE)
[5] VA Office of Biosurveillance, VA Central Office, Washington, DC

```python
from medspacy.ner import TargetRule

concept_tag_rules = {
    "coronavirus": [
        TargetRule(
            literal="coronavirus",
            category="COVID-19",
            pattern=[{"LOWER": {"REGEX": "coronavirus|hcov|ncov$"}}],
        ),
        TargetRule(
            literal="covid",
            category="COVID-19",
            pattern=[{"LOWER": {"REGEX": "^covid"}}],
        ),
        TargetRule(literal="Novel Coronavirus (COVID-19)", category="COVID-19"),
        TargetRule(literal="novel coronavirus", category="COVID-19"),
        TargetRule(
            literal="[{'LOWER': {'REGEX': '^covid-19'}}]",
            category="COVID-19",
            pattern=[{"LOWER": {"REGEX": "^covid-19"}}],
        ),
        TargetRule(
            literal="[{'LOWER': 'sars'}, {'LOWER': '-', 'OP': '?'}, {'LOWER': 'cov-2'}]",
            category="COVID-19",
            pattern=[{"LOWER": "sars"}, {"LOWER": "-", "OP": "?"}, {"LOWER": "cov-2"}],
        ),
        TargetRule(literal="2019-cov", category="COVID-19"),
        TargetRule(literal="cov2", category="COVID-19"),
        TargetRule(literal="ncov-19", category="COVID-19"),
        TargetRule(literal="novel coronavirus 2019", category="COVID-19"),
        TargetRule(literal="novel corona", category="COVID-19"),
        TargetRule(literal="covid-10", category="COVID-19"),
        TargetRule(literal="corona 2019", category="COVID-19"),
        TargetRule(literal="coronavirus 19", category="COVID-19"),
        TargetRule(literal="covd-19", category="COVID-19"),
        TargetRule(literal="COVID-19", category="COVID-19"),
        TargetRule(
            literal="COVID- [\s]+19",
            category="COVID-19",
            pattern=[
                {"LOWER": "covid-"},
                {"IS_SPACE": True, "OP": "+"},
                {"LOWER": {"REGEX": "19"}},
            ],
        ),
```

```python
In [38]: for i, doc in enumerate(docs):
             cov_bsv.visualize_doc(doc, document_id=i)
             print("__"*20)
```

Document Classification: POS

Document ID: 0

Patient presents to be tested for TEST COVID-19 COVID-19 . His wife OTHER_EXPERIENCER recently tested positive DEFINITE_POSITIVE_EXISTENCE for novel coronavirus COVID-19 . SARS-COV-2 COVID-19 results TEST came back positive DEFINITE_POSITIVE_EXISTENCE .

Document Classification: NEG

Document ID: 1

Lab results: << LABS >> SARS-COV-2 COVID-19 not detected NEGATED_EXISTENCE Diagnoses: << DIAGNOSES >> 1. COVID-19 COVID-19 testing TEST

Document Classification: POS

Document ID: 2

The patient was tested for TEST COVID-19 COVID-19 . Results are positive.

Document Classification: NEG

Document ID: 3

negative for NEGATED_EXISTENCE COVID-19 COVID-19 .

Document Classification: POS

Document ID: 4

https://github.com/abchapman93/VA_COVID-19_NLP_BSV

UNIVERSITY OF TORONTO

Example: Context Aware REGEX for COVID-19 Phenotyping/Monitoring

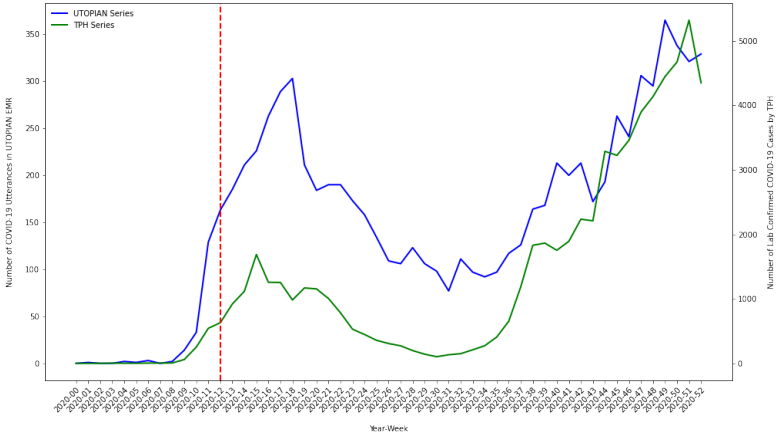
PLOS DIGITAL HEALTH

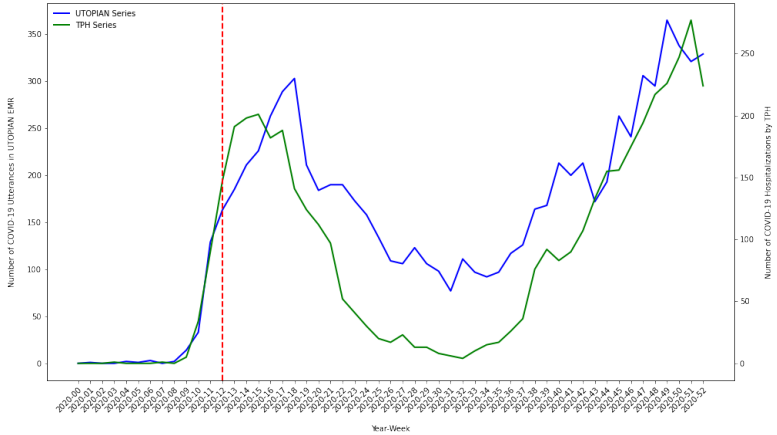
RESEARCH ARTICLE

Using Primary Care Clinical Text Data and Natural Language Processing to Identify Indicators of COVID-19 in Toronto, Canada

Christopher Meaney[1]*, **Rahim Moineddin**[1], **Sumeet Kalia**[1], **Babak Aliarzadeh**[1], **Michelle Greiver**[1,2]

1 Department of Family and Community Medicine, Faculty of Medicine, University of Toronto, Toronto, Canada, **2** North York Family Health Team, North York General Hospital, Toronto, Canada





UNIVERSITY OF TORONTO

Example: Using LLMs for Concept Set Generation at OHDSI

The "Minds Meet Machines" (MMM) Initiative

A Quality Improvement Project Evaluating Human and AI Workflows for OHDSI Concept Set Development.

Date: October 9, 2025 **Time:** 8:00 AM - 12:00 PM EDT **Location:** OHDSI Symposium, New Brunswick, NJ (and Virtual) **Sponsors:** Johns Hopkins University (JHU); OHDSI Community. **Principal Investigators:** Christopher Mecoli, MD (JHU); Gowtham A Rao, MD, PhD (OHDSI).

OHDSI Forums discussion

Overview

The MMM Initiative is a structured Quality Improvement (QI) project designed to evaluate and improve the internal OHDSI methodologies for phenotype development. We are utilizing mixed methods (quantitative and qualitative) to compare the performance (accuracy, completeness, and efficiency) of rigorous, human-led workflows against Generative AI (GenAI)-driven approaches for translating standardized clinical descriptions into concept sets.

Project Aims

1. **Primary:** Evaluate the performance of GenAI-driven approaches compared to human-led workflows.
2. **Secondary:** Quantify inter-human variability (Consensus Gain) and conduct a qualitative analysis of human reasoning and collaboration processes to identify best practices.

https://github.com/ohdsi-studies/MindMeetsMachines/

UNIVERSITY OF TORONTO

Supervised Methods for Phenotyping Clinical Text Data

- Phenoptying as a binary/multinomial classification problem (e.g. document classification).
- Phenotyping as multinomial classification problem at token level (e.g. BIO tagging + NER).
- Derive document level features (e.g. vector space models: DTM, TF-IDF).
- Derive token level features (e.g. morphology, position, neighbours, etc.).
- ML models: logistic LASSO, GBDTs, MLPs, random forests, etc.
- DNNs as feature learners: e.g. CNNs, LSTMs, transformers, LLMs, etc.
- Foundation models + prompt engineering for patient phenotype classification.

UNIVERSITY OF TORONTO

Example: NER for deidentifying sensitive entities from clinical text data

A COMPARATIVE EVALUATION OF TRANSFORMER MODELS FOR DE-IDENTIFICATION OF CLINICAL TEXT DATA

TECHNICAL REPORT

Christopher Meaney
Department of Family and Community Medicine
University of Toronto
Toronto, Ontario, Canada
christopher.meaney@utoronto.ca

Wali Hakimpour
Department of Family and Community Medicine
University of Toronto
Toronto, Ontario, Canada
ahmad.hakimpour@mail.utoronto.ca

Sumeet Kalia
Department of Family and Community Medicine
University of Toronto
Toronto, Ontario, Canada
sumeet.kalia@utoronto.ca

Rahim Moineddin
Department of Family and Community Medicine
University of Toronto
Toronto, Ontario, Canada
rahim.moineddin@utoronto.ca

April 15, 2022

Patient: [Patient Name]

MR#: [MRN]

Date of Admission: [Date]

Attending Oncologist: Dr. [Physician Name]

PCP: Dr. [Physician Name]

Chief complaint: [Age] year-old man with stage IV-B metastatic esophageal adenocarcinoma presents complaining of three days of constant dizziness.

History of Present Illness: Mr. [Patient Name] is a pleasant [Age] year-old man with a past medical history significant for metastatic esophageal adenocarcinoma s/p XRT and chemotherapy, CAD s/p LAD stent placement, and borderline DM who presents to the [Medical Center Name] complaining of 48-hour history of dizziness and subjective fevers.

UNIVERSITY OF TORONTO

Unsupervised Methods for Phenotyping Clinical Text Data

- Determine representation for text data (e.g. DTM, embeddings, etc.).
- Discover latent patterns/groups from high-dim text data (without labelled data).

UNIVERSITY OF TORONTO

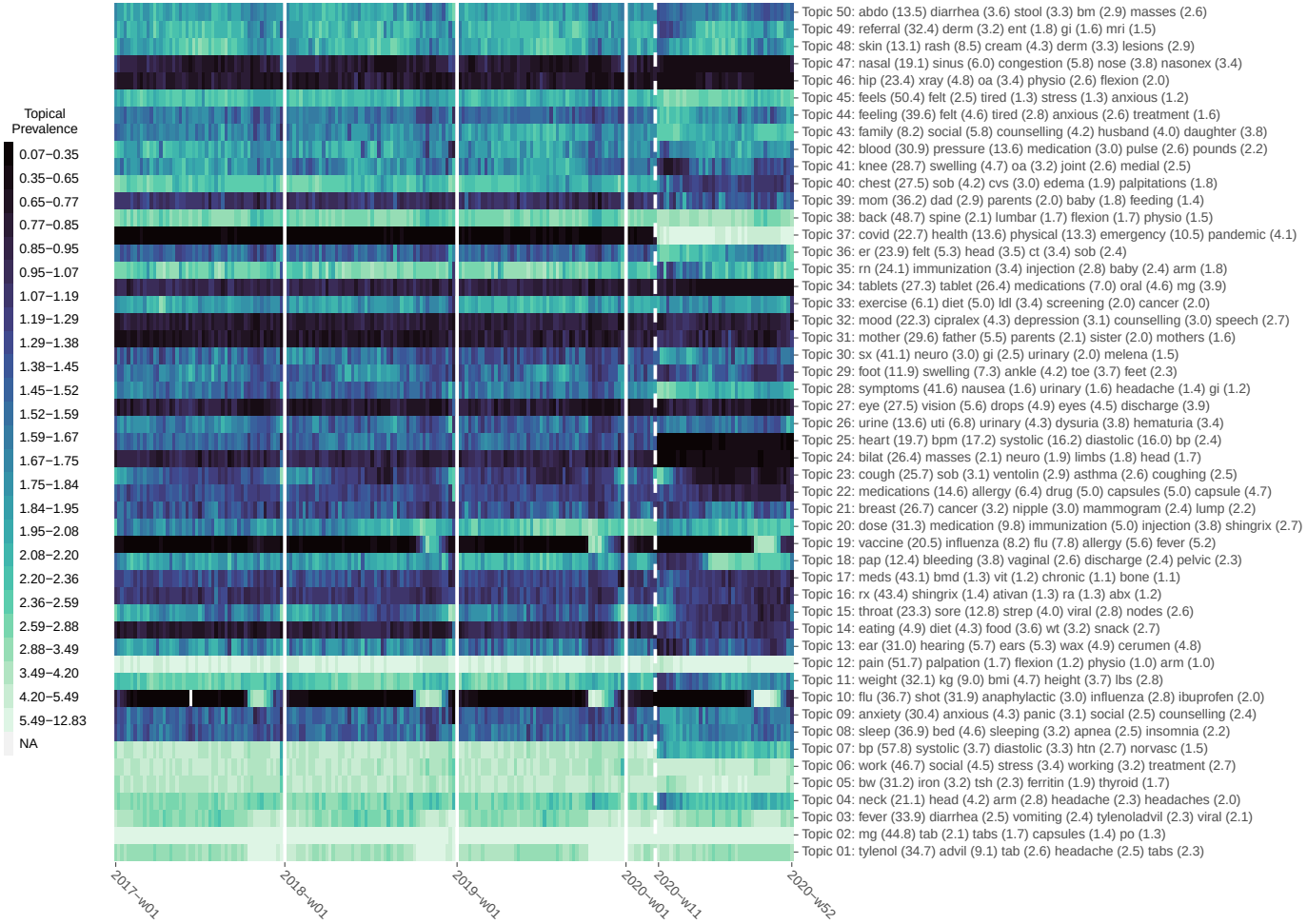
Example: NMF Temporal Topic Modelling

- Observe N*P dimensional DTM (X). Observe meta-data (z_i) for each note (i=1...N).
- Assume, for each note i=1...N we have $z_i \in (1 \ldots T)$. T discrete levels.
- Realize, each note i=1...N is associated with length K topical prevalence vector.
- For each stratifying factor (t=1...T), compute length-K mean topical prevalence vector.
- Resulting T*K dimensional MVTS object used for evaluating COVID-19 pandemic effects.

$$
\begin{array}{c}
z_i(t{=}1)\left\{\right. \\
z_i(t{=}2)\left\{\right. \\
z_i(t{=}3)\left\{\right. \\
z_i(t{=}T)\left\{\right.
\end{array}
\begin{bmatrix}
x_{1,1} & \cdots & \cdots & x_{1,P} \\
\vdots & \vdots & \vdots & \vdots \\
\vdots & \vdots & \vdots & \vdots \\
\vdots & \vdots & \vdots & \vdots \\
\vdots & \vdots & \vdots & \vdots \\
\vdots & \vdots & \vdots & \vdots \\
\vdots & \vdots & \vdots & \vdots \\
\vdots & \vdots & \vdots & \vdots \\
\vdots & \vdots & \vdots & \vdots \\
\vdots & \vdots & \vdots & \vdots \\
\vdots & \vdots & \vdots & \vdots \\
\vdots & \vdots & \vdots & \vdots \\
x_{N,1} & \cdots & \cdots & x_{N,P}
\end{bmatrix}
\approx
\begin{bmatrix}
\theta_{1,1} & \cdots & \theta_{1,K} \\
\vdots & \vdots & \vdots \\
\vdots & \vdots & \vdots \\
\vdots & \vdots & \vdots \\
\vdots & \vdots & \vdots \\
\vdots & \vdots & \vdots \\
\vdots & \vdots & \vdots \\
\vdots & \vdots & \vdots \\
\vdots & \vdots & \vdots \\
\vdots & \vdots & \vdots \\
\vdots & \vdots & \vdots \\
\vdots & \vdots & \vdots \\
\theta_{N,1} & \cdots & \theta_{N,K}
\end{bmatrix}
*
\begin{bmatrix}
\phi_{1,1} & \cdots & \cdots & \phi_{1,P} \\
\vdots & \vdots & \vdots & \vdots \\
\phi_{K,1} & \cdots & \cdots & \phi_{K,P}
\end{bmatrix}
$$

UNIVERSITY OF TORONTO

Example: NMF Topic Models Identify Pandemic Impacts on Primary Care

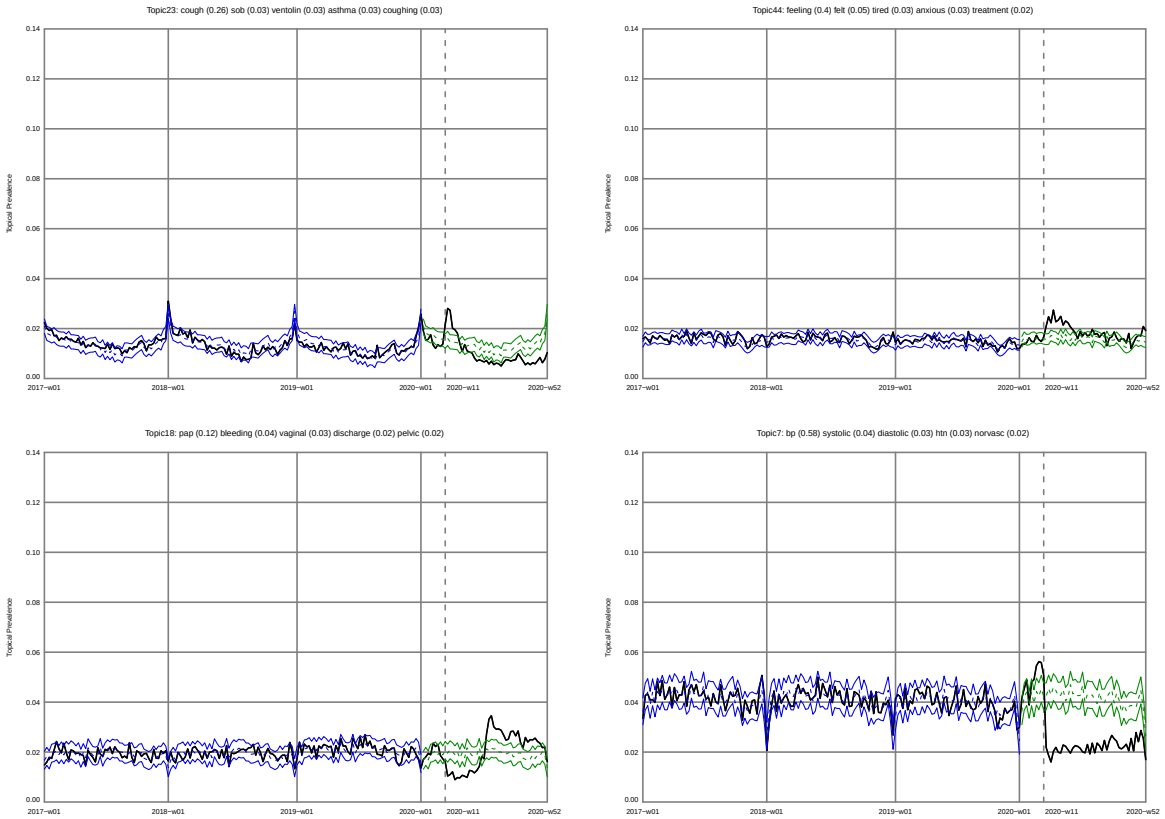


Meaney et al. (2021). NMF Topic Models for Monitoring COVID-19. JBI.

UNIVERSITY OF TORONTO

Example: NMF Topic Models Identify Pandemic Impacts on Primary Care

- Time series dynamic regression models. AR-1 auto-correlation structure.
- Estimation using 2017-2019 data; forecast 2020; compare predictions/observations.
- Select time series with altered dynamics: resp disease, mental health, screen/mgmt.



Meaney et al. (2021). NMF Topic Models for Monitoring COVID-19. JBI.

UNIVERSITY OF TORONTO

Example: LLMs as a Foundational Component in Modern Topic Models

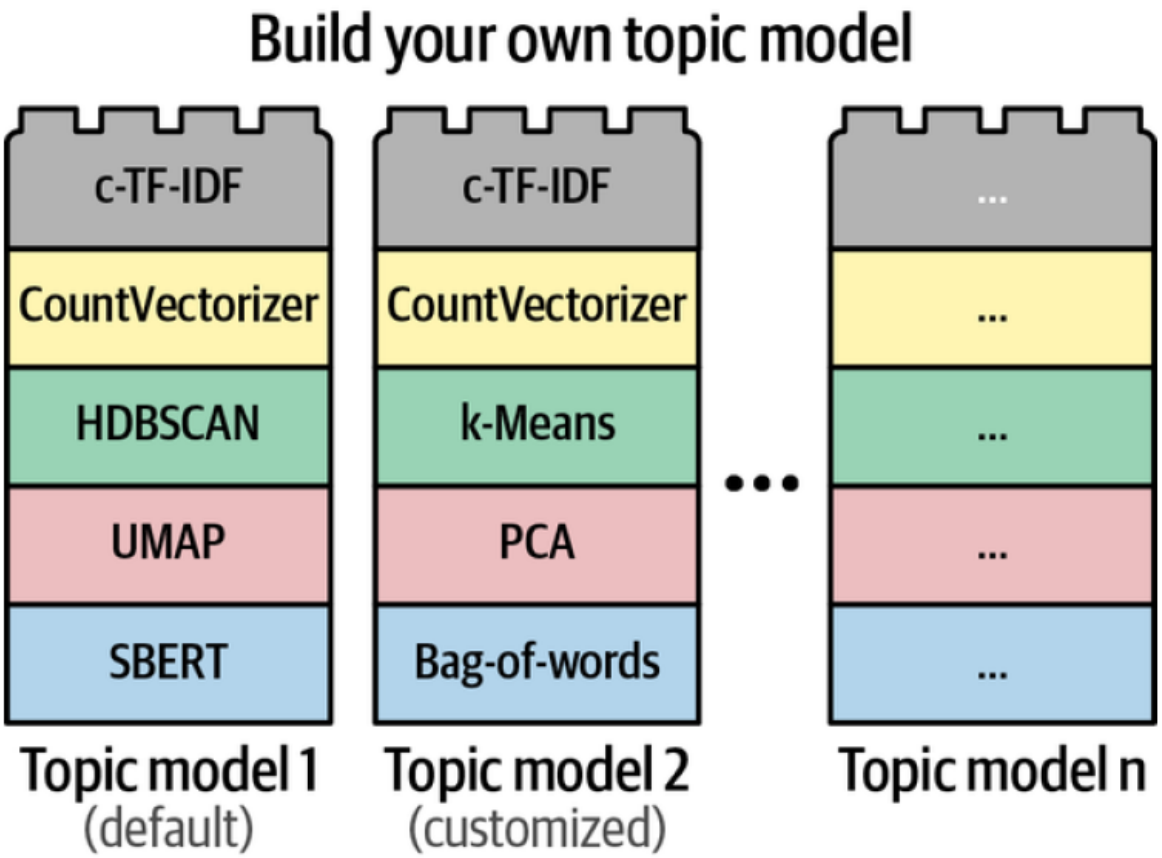


Figure 5-17. The modularity of BERTopic is a key component and allows you to build your own topic model however you want.

https://www.llm-book.com/

UNIVERSITY OF TORONTO

Reflecting on the Past and Future of Phenotyping Text Data

- Phenotyping a crucial element in working with routinely collected health data.
- Variety diverse methods exist for phenotyping clinical text data.
- Hybrid methodological pipelines have proven to be flexible/performant in practice.
- Foundation models + prompt engineering as emerging theme in phenotyping text data.
  - Instruction specificity.
  - Problem context.
  - Desired output/response format.
  - Persona, tone, audience, etc.
  - Data: zero-shot, few-shot, etc.
  - Chain of thought. Think slowly. Show steps.