Building data capacity for PCOR: Role of NLP in data infrastructure

V.G.Vinod Vydiswaran

vgvinodv@umich.edu

Assistant Professor of Learning Health Sciences

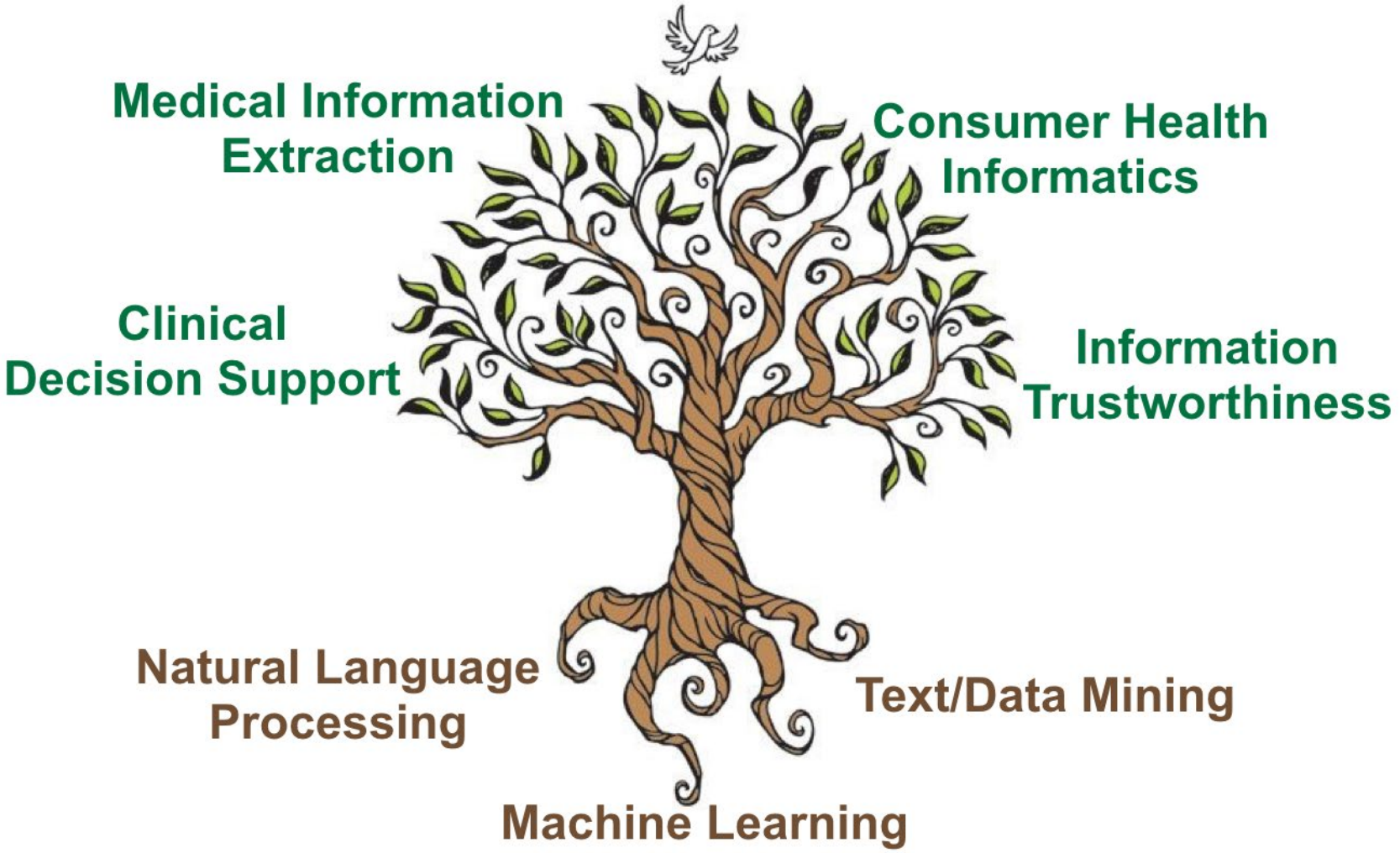
Assistant Professor of Information

May 24th, 2021

Brief Bio

- Assistant Professor (’15–’21); Associate Professor (’21–)
  - Department of Learning Health Sciences, Medical School
  - School of Information
- Ph.D. in Computer Science (2013), University of Illinois at Urbana-Champaign
- M.Tech. (2004), Indian Institute of Technology Bombay
- Member of LHSNet, a PCORI Clinical Data Research Network

My research interests



Research focuses on **transforming data into actionable knowledge** while addressing issues of context, scale, quality, and impact.

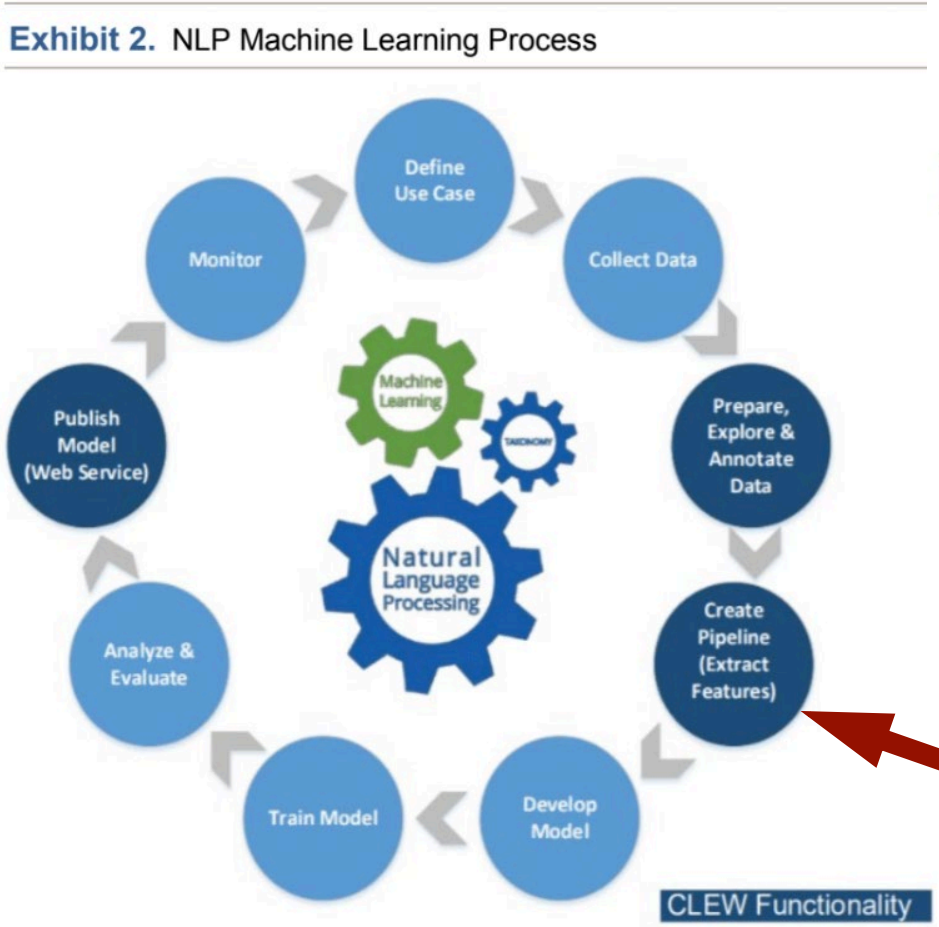
Making PCOR data infrastructure more useful for research : My two cents

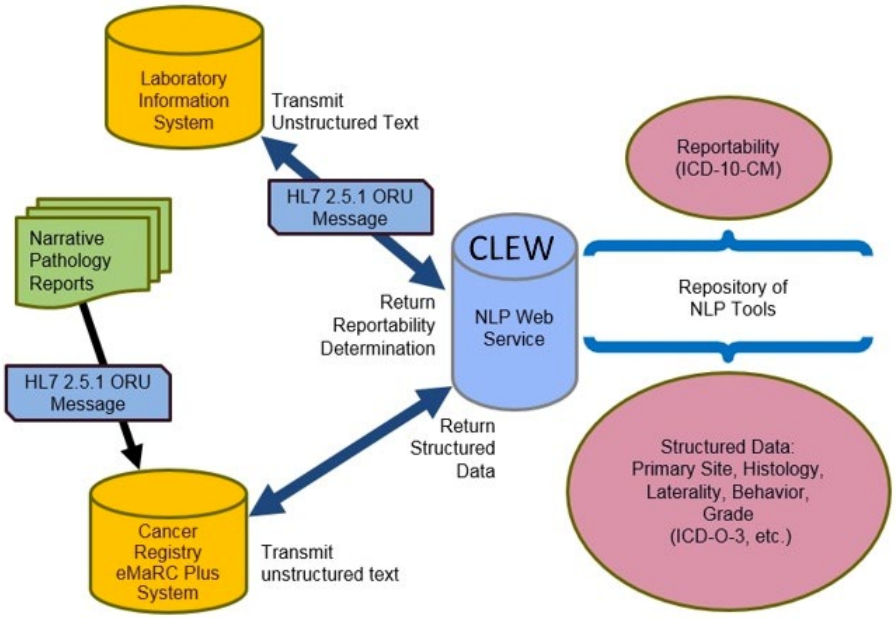
- 1. Informatics infrastructure that includes clinical notes
- 2. Computable phenotypes as knowledge objects
- 3. Looking beyond EHRs for health data

Insights from LHSNet

- Patient-Centered Network of Learning Health Systems (LHSNet)
  - PCORI-supported Clinical Data Research Network (CDRN)
  - Across nine organizations , incl. 6 health systems
- Common Data Model for structured data (demographics, lab, …)
  - Generally successful in launching studies across multiple sites
  - Does not include text data
- **1. Informatics Infrastructure for processing textual clinical notes**
  - Extending Common Data Models to include textual components
  - Extracting clinically-relevant information from free text

Extends prior efforts





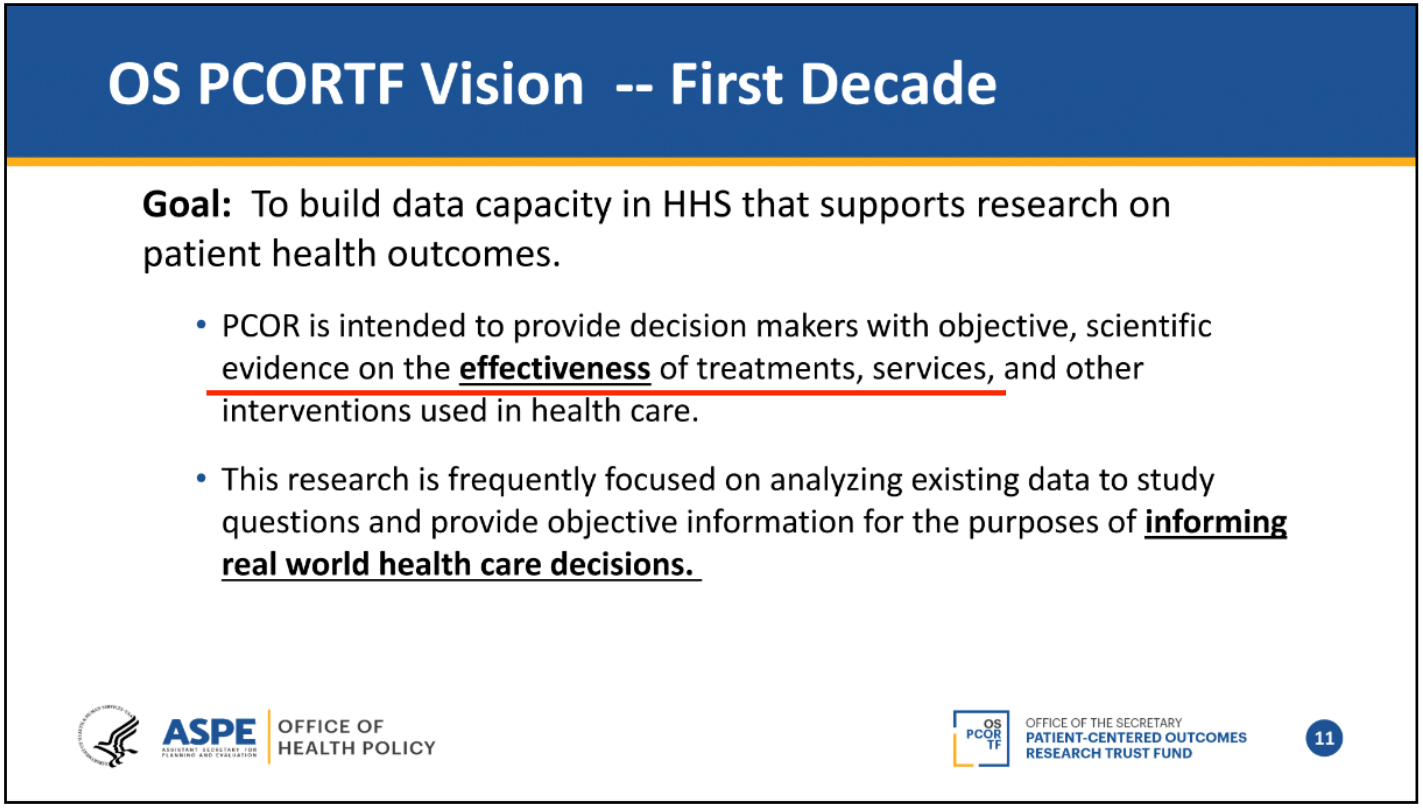
Clinical Language Engineering Workbench (CLEW)

- Developed as part of CDC's National Program of Cancer Registries (NPCR)

Not only features, but also concepts & relationships between concepts

CLEW: https://www.cdc.gov/cancer/npcr/informatics/nlp-workbench/pilot-cancer.htm

2. Computable phenotypes as knowledge objects



- Standardized definitions
- Computational tools, including clinical NLP tools for processing text
  - Knowledge objects, supporting/expanding the MCBK initiatives

Incorporating patient-reported outcomes

- Typical data elements in computable phenotypes
  - ICD-9/ICD-10 codes, CPT codes
  - medications
  - key terms and phrases, frequency of mentions
- Novel features from patient-reported outcomes
  - Symptoms, medication response, and adverse events in telephone notes
  - Medication refill requests through web portal requests
  - Care provider information, especially for patients unable to independently manage their health care needs
- Patient-reported outcomes are often in text form in EHRs

Prior work: Patient-provided text

- Self-reporting behavior about oral anticancer agent toxicity in clinical notes[1]
  - Self-reported notes captured in telephone encounters, emails, and messages via portal
  - 23.5% of OAA clinical notes were telephone encounter notes

- Patient-reported outcomes in Crohn's Disease
  - Extract patient-provided information regarding Crohn's Disease symptoms, medication response, and adverse events using email and telephone notes stored in electronic medical records
  - Funded by Helmsley Foundation, 2021—2024 (PI: Vydiswaran, Stidham)

[1] Jiang Y, **Vydiswaran VGV**, Eun YL, Joo H, Zheng A, Harris MR. Feasibility of Identifying Oral Anticancer Agent Toxicity Self-Reporting and Management Advice from Clinical Notes. *AMIA Annu Symp Proc* 2018.

3. Looking beyond EHRs for health data

Adverse drug events in social media

- Health related communication has increased significantly online
- Laypersons share health behaviors freely on social media
  - including use of drugs and associated adverse effects

I wonder if the milk and honey thing work for heartbirn cause Tums ain't shit for me

- If understanding clinical text is hard, consumer-generated text is harder!
  - Grammatical structure might not hold, typos abundant
  - Not "natural" language: use of short text language
    - new acronyms, abbreviations ==> disambiguation challenges

Baby im ur insulin if u need ur medicine. I'll be ur prescription, come & take a dose of me. Once u get a dose of me, baby i'm ur medicine.

“Social” determinants of health

- Exploring community health information in geo-located social media

- Analyzing social media through the lens of
  - communities: affluent vs. disadvantaged neighborhoods
  - demographics: e.g. “BlackTwitter”
  - patient cohorts: e.g. smoking cessation patient groups

- Linking patients to spatiotemporal factors
  - Environmental / spatial factors such as air pollution, neighborhood walkability, rurality, and “food deserts” affects health, exercise, and diet
  - Social media provides a non-traditional way to learn about health information exposure, sentiment, and “patient-reported” outcomes

PCOR data infrastructure for health social media

- Social media can augment information in FAERS and VAERS
- NLP to the rescue, for parsing text reports in FAERS/VAERS, and also on social media

Tweet count	Drug mentioned	Total		Mental		Sleep		Pain		Tiredness		Nausea		Sweat		Itch	
133	Benadryl	3	3,1	X	X			X	X		X	X	X				
119	Adderall	6	5	X	X			X	X	X	X	X	X	X	X	X	
108	Ibuprofen	5	2,1	X	X		X	X	X	X		X				X	
74	Xanax	5	4	X	X	X	X	X	X	X	X	X					
73	Acetaminophen	5	4,2		X		X	X	X	X	X	X	X	X	X	X	
64	Vyvanse	6	5	X	X	X	X	X	X	X	X	X		X	X		
34	Codeine	6	5	X	X	X	X	X	X			X	X	X		X	X
17	Morphine	6	3	X	X	X	X	X	X			X		X		X	
16	Ambien	5	2	X	X			X	X	X		X				X	
12	Concerta	7	4	X		X	X	X	X	X	X	X	X	X		X	

- MedlinePlus - Twitter

Example 1

this benadryl is making me tired

Example 2

I wish benadryl did not make me so tired

. . .

Example 12

apparently benadryl makes me tired

Key takeaways

- 1. Need for an enhanced **informatics infrastructure** for processing textual clinical notes
- 2. Treating **computable phenotypes as knowledge objects**, and incorporating patient-reported outcomes derived using NLP
- 3. **Health-related social media** can augment existing population health and SDOH efforts

Questions & Comments: vgvinodv@umich.edu