VERITY Bioinformatics Mini-Course: Bioinformatics with Clinical Data

June 9, 2021

Katherine P. Liao, MD, MPH
Associate Professor of Medicine | Assistant Professor of Biomedical Informatics
Harvard Medical School
Director, VERITY Bioinformatics Core, Division of Rheumatology, Inflammation, and Immunity
Brigham and Women's Hospital

Course faculty

Tianrun Cai, MD
Associate Bioinformatician
Division of Rheumatology, Inflammation and Immunity,
Brigham and Women's Hospital (BWH)
Instructor in Medicine, Harvard Medical School (HMS)

Xu Shi, PhD
Assistant Professor Biostatistics
University of Michigan, School of Public Health

Tianxi Cai, ScD
John Rock Professor of Population & Translational Data Sciences, Harvard T.H. Chan School of Public Health
Professor of Biomedical Informatics, HMS
Director, Translational Data Science Center for a Learning Health System (CELEHS), HMS
Associate Director, VERITY Bioinformatics Core, BWH
Co-Director, Applied Bioinformatics Core, Veterans Affairs (VA) Boston

Katherine Liao, MD, MPH
Director, VERITY Bioinformatics Core, Division of Rheumatology, Inflammation, and Immunity, BWH
Associate Professor of Medicine & Assistant Professor of Biomedical Informatics, HMS
Co-Director, Applied Bioinformatics Core, VA Boston

Course contents

- Day 1 | June 9, 2021
  - Overview, Day 1
  - Introduction to natural language processing (NLP)
    - Unified Medical Language system and Building a Dictionary for Phenotyping
  - Overview of phenotyping electronic health record (EHR) data

- Day 2 | June 16, 2021
  - Overview, Day 2
  - Beyond phenotyping w/ ML and EHR
    - Machine Learning (ML) for clinical research studies w/ noisy EHR data
  - What is possible w/ NLP
    - Improve Chart Review Efficiency using NLP and ML
  - How to connect w/ the VERITY Bioinformatics Core

Bioinformatics

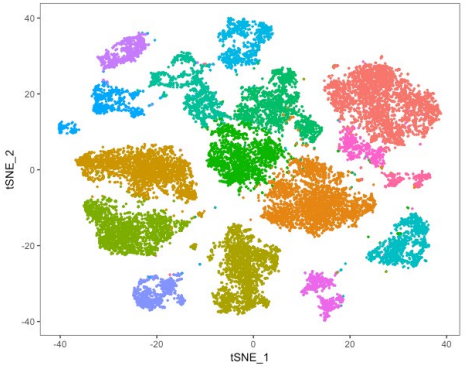
- Combines algorithms & approaches employed in computer science and statistics to analyze, understand, and hypothesize about large repositories of collected biologic data and knowledge

Raychaudhuri, Computational Text Analysis: For Functional Genomics and Bioinformatics, 2006

Bioinformatics

- Combines algorithms & approaches employed in computer science and statistics to analyze, understand, and hypothesize about large repositories of collected biologic data and knowledge

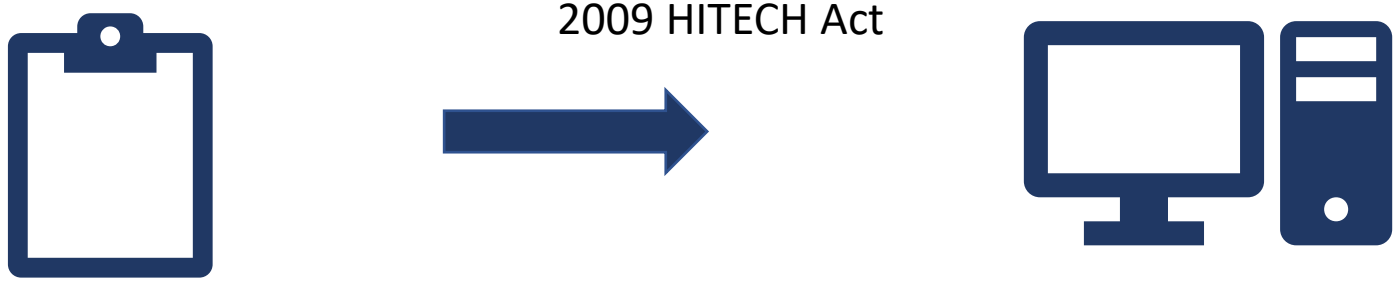
- Rate limiting step in biology
  - Data collection → data interpretation



- Genomics
  - Single gene → all genes in an organism → high throughput technologies

Raychaudhuri, Computational Text Analysis: For Functional Genomics and Bioinformatics, 2006

Bioinformatics for Clinical Data



Paper charts

Manual chart review to extract data

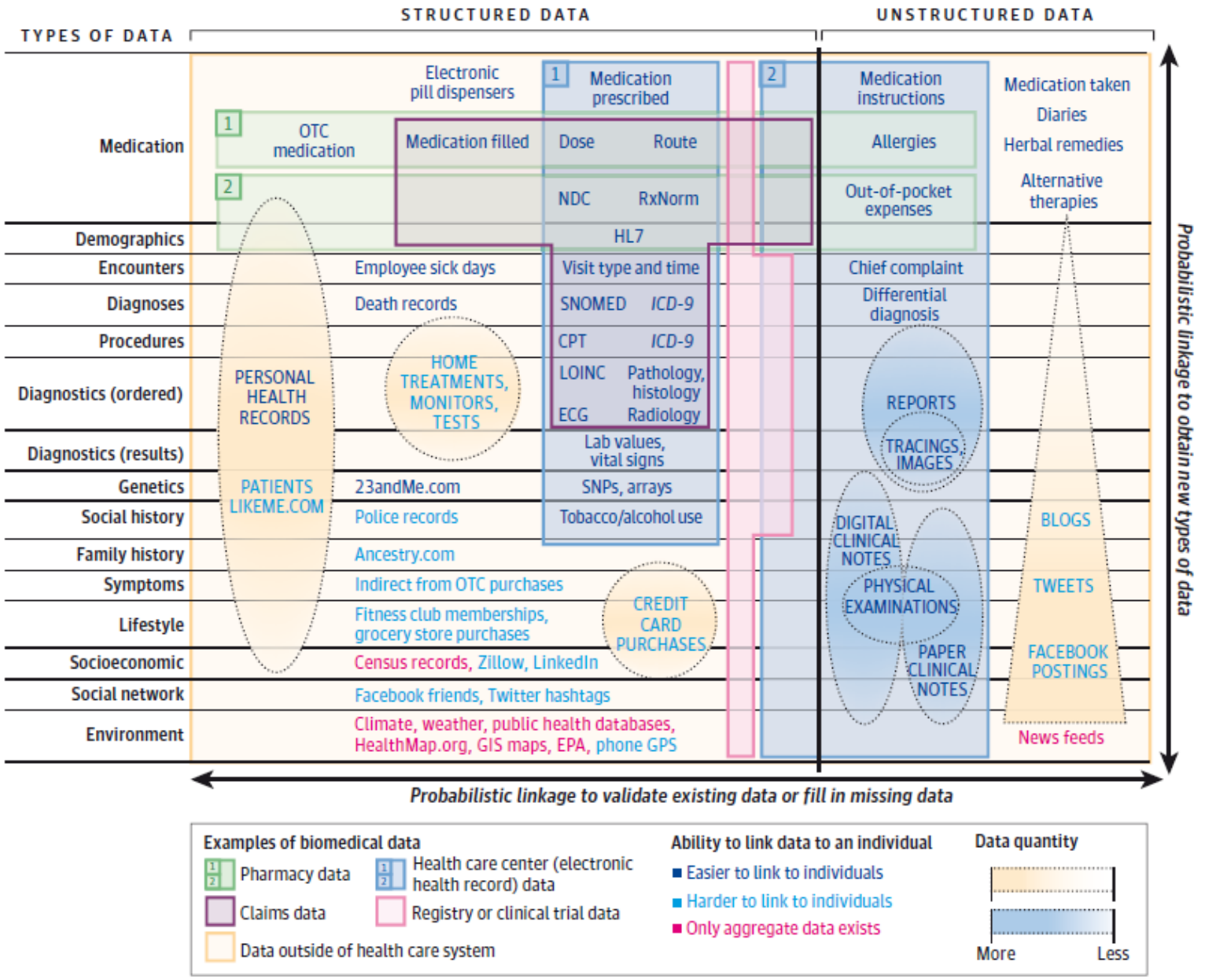
- Limits variables and outcomes for study
- Not feasible to study large populations

Electronic health records (EHR)

High volume data from clinical care

- Designed for billing, patient care
  - Suboptimal for research
- Enables studies in large populations

Figure. The Tapestry of Potentially High-Value Information Sources That May be Linked to an Individual for Use in Health Care



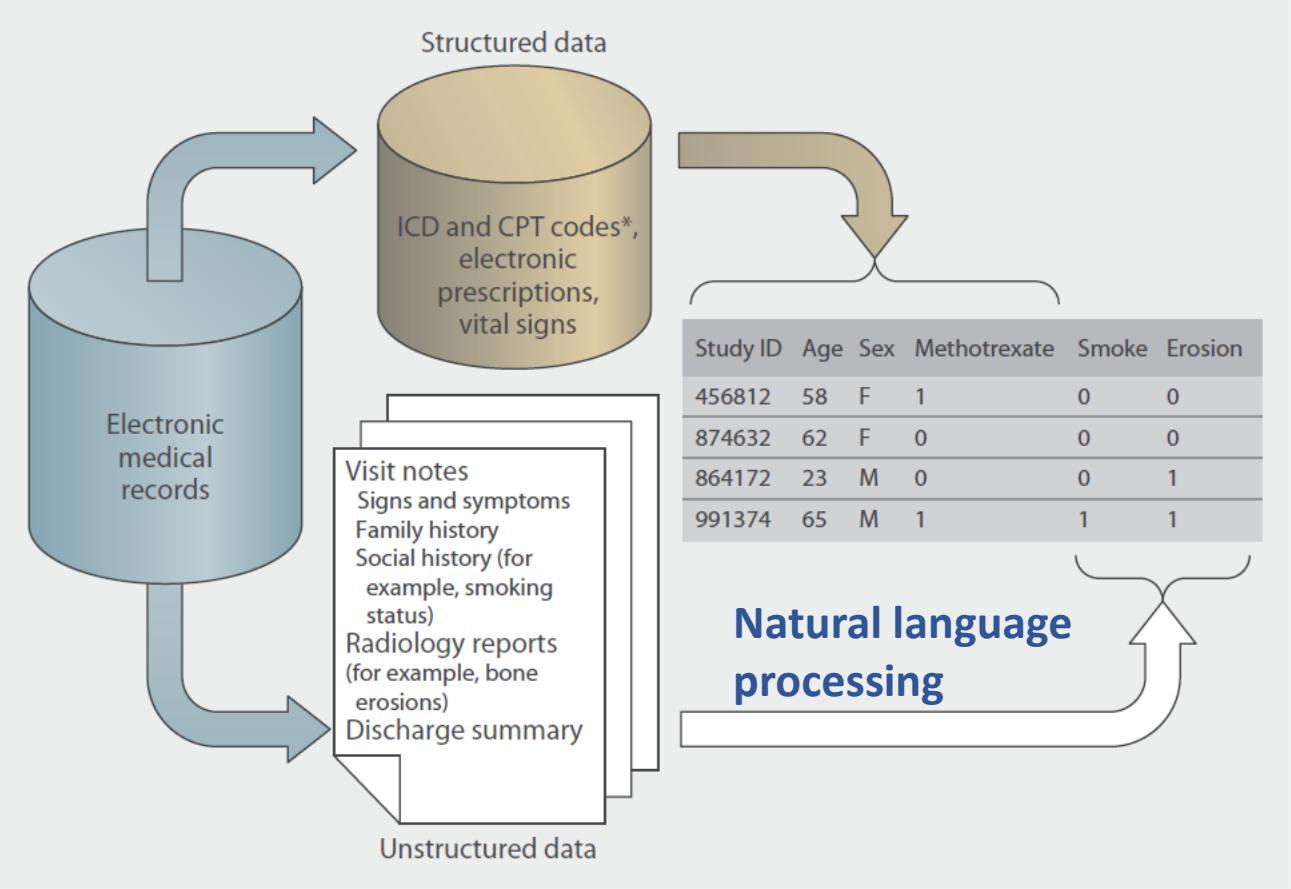
Weber, et al., JAMA 2014

Classifying RA using EHR

Model	PPV (95% CI)	Sensitivity (95% CI)
EHR RA algorithm (ICD +NLP)	94% (91-96)	63% (51-75)
≥3 ICD9 RA codes	56 (72-88)	80 (72-88)
≥1 RA ICD codes + ≥1 DMARD	45 (37-53)	66 (57-76)

Liao, Cai, et al., Arth Care & Res 2010

Types of EHR data



Liao, Cai, et al., BMJ 2015

NLP

- Enriches clinical data available

- Considerations
  - Which output from NLP will be useful for the algorithm
    - For RA, AM stiffness, joint swelling, hand pain, seropositivity, etc…
  - Organizing/streamlining the data
    - Shoulder pain, knee pain, hand pain = joint pain

NLP

- Enriches clinical data available

- Considerations
  - Which output from NLP will be useful for the algorithm
    - For RA, AM stiffness, joint swelling, hand pain, seropositivity, etc…
    - Approach: create a dictionary
  - Organizing/streamlining the data
    - Shoulder pain, knee pain, hand pain= joint pain
    - Approach: leverage existing resource from the National Library of Medicine, the Unified Medical Language System (UMLS)

Biomedical Big Data

- Digital objects with impact in basic, translational, clinical, social, behavioral, environmental, or informatics research questions

Primary	Secondary	Infrastructure
Imaging	Social media	Metadata
ICD	Search histories	Data standards
Genotypes	Cell phone data	Data analyses
Molecular	Geocode	Data processing

Adapted from https://datascience.nih.gov/bd2k

Biomedical Big Data

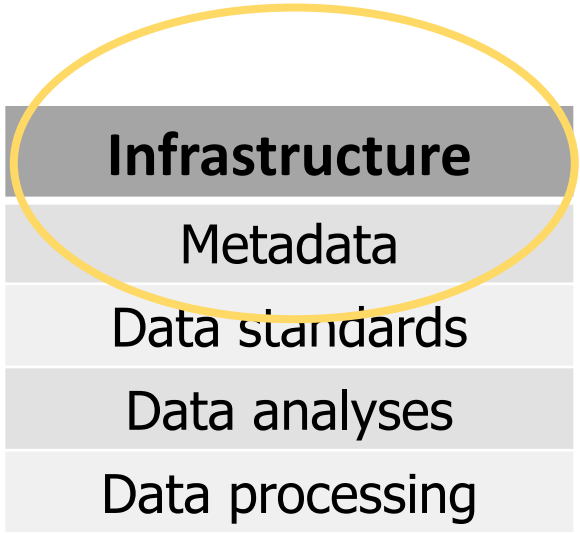
- Digital objects with impact in basic, translational, clinical, social, behavioral, environmental, or informatics research questions

Infrastructure
Metadata
Data standards
Data analyses
Data processing

Adapted from ht

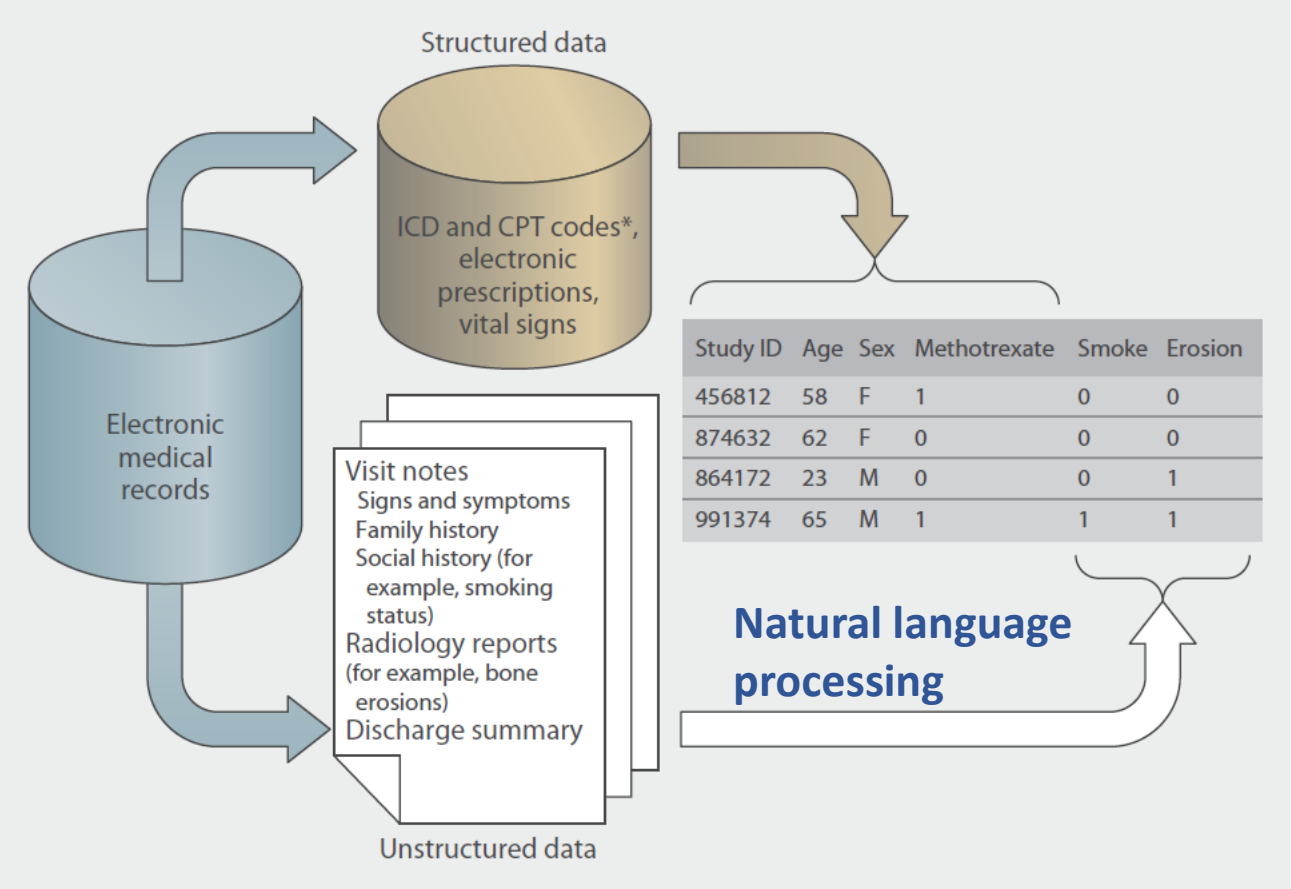
UMLS

- UMLS relates terms, e.g. shoulder pain or hip pain, with a general concept for “joint pain”
- Assigns concept unique identifier (CUI)
  - Joint pain, CUI C0003862



Adapted from https://datascience.nih.gov/bd2k

EHR data for phenotyping



Liao, Cai, et al., BMJ 2015

Thank you