

AIM 2: Artificial Intelligence in Medicine II

Harvard - BMIF 203 and BMI 702, Spring 2025

Lecture 1: Course overview, Introduction to NLP, NLP in clinical settings,
Medical terminology challenges, Concept extraction from EHRs,
Clinical trial matching



HARVARD
MEDICAL SCHOOL

Marinka Zitnik
marinka@hms.harvard.edu

Outline for today's class



1. Overview of this course

2. What makes biomedical data unique

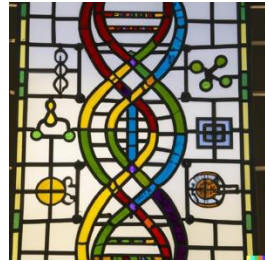
3. Introduction to distributed language representations

4. Introduction to NLP in clinical settings

What will you learn in this course?

- **Key data modalities**

- Clinical data
- Networks, graphs, and multimodal datasets
- Language and text
- Images



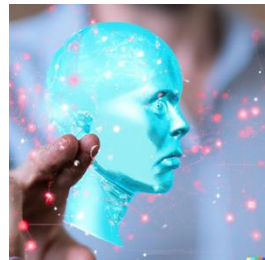
- **Cutting-edge algorithmic principles underlying AI**

- Self-supervised learning and transfer learning
- Large-scale pre-training and efficient fine-tuning
- Multimodal learning
- Generative AI



- **Broader impacts:**

- Model evaluation, benchmarking, and deployment
- Privacy, safety, and copyright issues of AI



Course staff

- **Marinka Zitnik** (Instructor)
 - Biomedical Informatics at HMS
 - Kempner Institute at Harvard University
 - Broad Institute of Harvard and MIT
 - <https://zitniklab.hms.harvard.edu>

- **Grey Kuling** (Curriculum Fellow)
 - Curriculum Fellow in Medical AI



Course staff

- **Yasha Ektefaie**

- PhD student in BIG program
- yasha_ektefaie@g.harvard.edu



- **Yepeng Huang**

- PhD student in BBS program
- yepeng@fas.harvard.edu



- **Courtney A Shearer**

- PhD student in SSQB program
- courtney.shearer@gmail.com

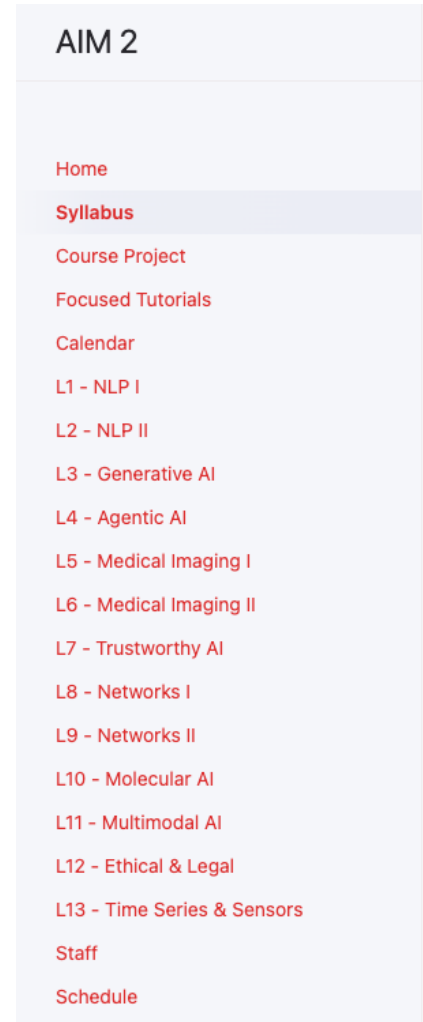


Dates, times and format

- **Course website:**
 - <https://zitniklab.hms.harvard.edu/AIM2>
 - BMIF 203 runs jointly with BMI 702. Refer to <https://canvas.harvard.edu/courses/151093>
- **Tuesdays, 2:00 PM – 4:00 PM ET**
 - No class or assignments due: Week of March 17
- **Location:**
 - TMEC 227 (except week 1 in room 128)
- **Office hours:**
 - Mon, 12-1pm (Zitnik)
 - Mon, 4-5pm (Shearer)
 - Thu, 1-2pm (Ektefaie)
 - Thu, 2-3pm (Huang)

Key components of this course

- Weekly lectures
- Focused tutorials
- Research project
- Weekly reading assessments



A screenshot of a course navigation menu for AIM 2. The menu is displayed in a light blue sidebar with a white background for the text. The title 'AIM 2' is at the top. Below it, a list of navigation items is shown, with 'Syllabus' highlighted in a darker blue background. The items are: Home, Syllabus, Course Project, Focused Tutorials, Calendar, L1 - NLP I, L2 - NLP II, L3 - Generative AI, L4 - Agentic AI, L5 - Medical Imaging I, L6 - Medical Imaging II, L7 - Trustworthy AI, L8 - Networks I, L9 - Networks II, L10 - Molecular AI, L11 - Multimodal AI, L12 - Ethical & Legal, L13 - Time Series & Sensors, Staff, and Schedule.

AIM 2
Home
Syllabus
Course Project
Focused Tutorials
Calendar
L1 - NLP I
L2 - NLP II
L3 - Generative AI
L4 - Agentic AI
L5 - Medical Imaging I
L6 - Medical Imaging II
L7 - Trustworthy AI
L8 - Networks I
L9 - Networks II
L10 - Molecular AI
L11 - Multimodal AI
L12 - Ethical & Legal
L13 - Time Series & Sensors
Staff
Schedule

Focused tutorials

- Practical tutorials are designed to give you hands-on experience applying AI techniques to real-world healthcare problems
- Core AI applications: NLP, medical image analysis, graph neural networks, generative models, LLMs, biological and clinical foundation models

[Tutorial 1: NLP in Medicine](#)

[Tutorial 2: Generative AI in Medicine](#)

[Tutorial 3: Medical Image Analysis](#)

[Tutorial 4: AI in Genomics](#)

[Tutorial 5: Biomolecular Structure Modeling with AlphaFold3, Boltz-1, and Chai-1 Foundation Models](#)

[Tutorial 6: Protein Language Models for Clinical Variant Effect Prediction](#)

[Tutorial 7: Modeling Single-Cell Perturbations with Foundation Models](#)

Research projects

- Research project:
 - Identify a medical question aligned with your area of interest
 - Identify one or more dataset to study the question
 - Develop, apply or adapt one or more AI models for the dataset
 - Run experiments, benchmark models, share findings and results
- Project proposal (due in week 3)
- Mid-term project presentation (week 7)
- Final presentations and report (week 13)
- Form groups:
 - BMIF 203: Groups of size 1-2 students
 - BMI 702: Groups of size 2-3 students
- We will provide Google Colab subscriptions
- Check out our project ideas and open medical datasets

https://zitniklab.hms.harvard.edu/AIM2/course_project/

Weekly reading assessments

- Weekly quizzes based on ~2 medical AI papers
- These readings are essential for building a strong understanding of the concepts we will discuss in class
- Quizzes are graded on completion, so if you submit thoughtful responses, you will receive full credit
- You will also receive model answers to compare with your own, helping you check your understanding of course materials

Grading

Component	Percentage	Description
Project Proposal	5%	A 2-page proposal outlining your project's research question, methodology, dataset, and contingency plans, evaluated for clarity and feasibility. A third page is allowed for figures and tables. Unlimited space for references.
Peer-Reviewed Feedback on Proposal	5%	Constructive feedback is provided to peers, following the criteria for effective research review.
Midterm Project Presentation	10%	A presentation summarizing your progress, baseline results, and challenges. Assessed for clarity, engagement, and preparedness for feedback. Presentation file submitted through Canvas.
Final Project Report	50%	A comprehensive, NeurIPS-style report detailing your research question, methods, results, and conclusions. Assessed for depth, accuracy, and insights.
Final Project Presentation	13%	A conference-style presentation summarizing your project's outcomes, strengths, and limitations. Evaluated on clarity, organization, and professionalism. Presentation file submitted through Canvas.
Focused Tutorials and Lectures	5%	Participation in hands-on tutorials and lectures, demonstrating engagement with coding and model application exercises.
Weekly Reading Assessments	12%	Completion of weekly assessments following assigned readings, ensuring ongoing engagement and comprehension. 1 point per quiz; there is no quiz for Lecture 1.

We Want You to Succeed!

You are welcome to visit our office hours and talk with us. We know graduate school can be stressful and we want help you succeed

Course culture and attendance

- **Course culture and collaboration:**
 - Students taking this course come from diverse backgrounds
 - All members of this course are expected to treat each other with courtesy and respect
 - You can collaborate with others but we ask that you write your solutions individually in your own words
- **Attendance:**
 - We ask students to attend all classes
 - You are encouraged to attend focused tutorials. We expect that students will attend at least some of them

Policies

- **We support using LLMs, genAI and coding copilots:**
 - **Responsibility for content:** Students who use LLMs and generative AI tools in their assignments take full responsibility for the content they submit
 - **Acknowledgment of AI use:** Clearly acknowledge any use of LLMs, specifying the nature and extent of assistance received from AI. Make sure to perform critical thinking, analysis, and synthesis of information
 - **Ethical use and originality:** Follow the principles of academic Do not use AI to plagiarize, misrepresent original work, or fabricate data
 - **Instructor discretion:** We may specify assignments where LLMs and generative AI use is encouraged or prohibited

Outline for today's class



1. Overview of this course

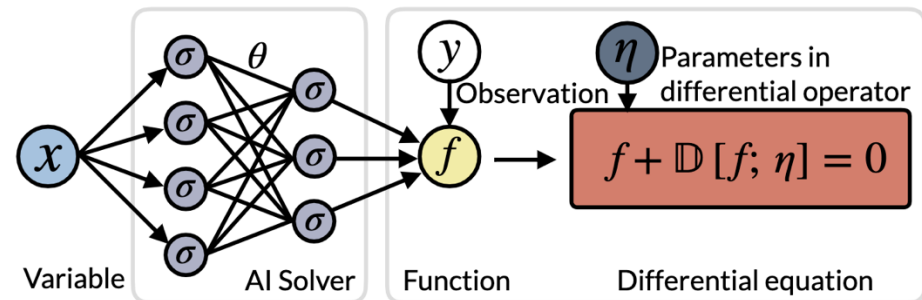
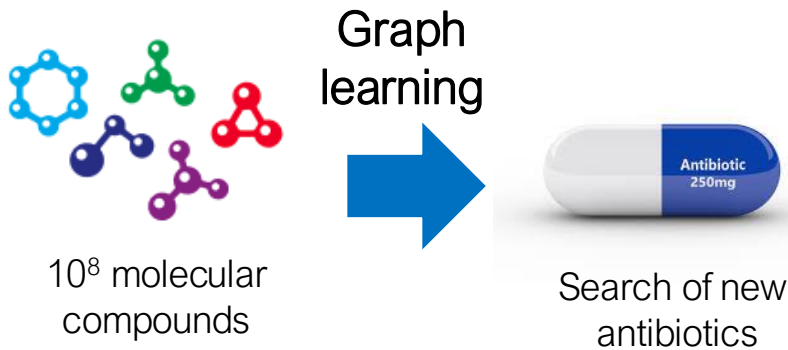
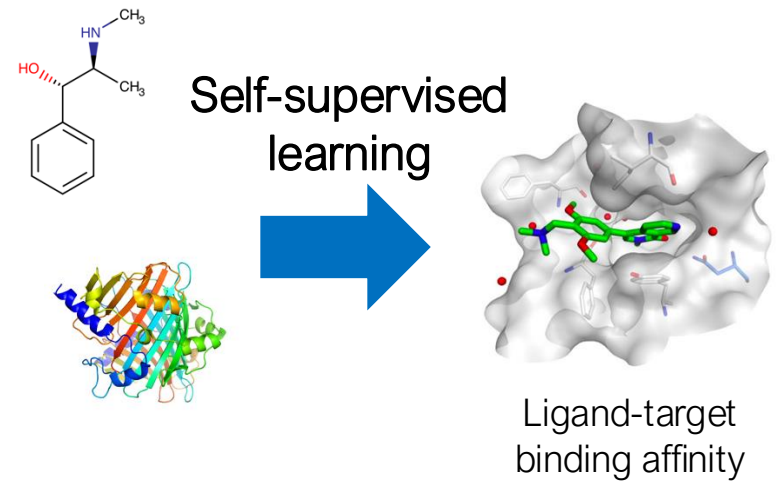
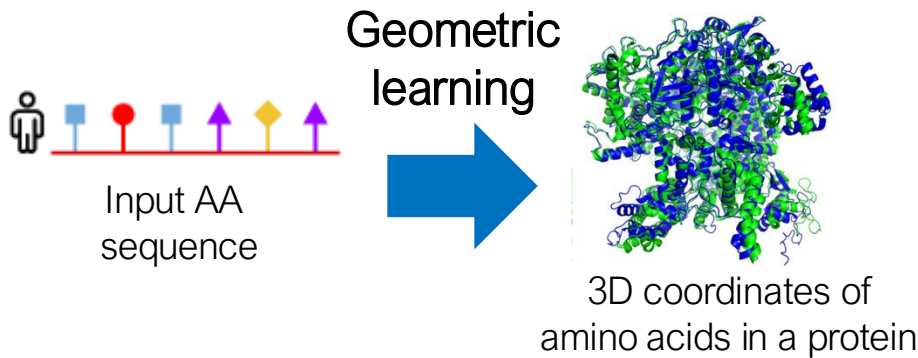


2. What makes biomedical data unique

3. Introduction to distributed language representations

4. Introduction to NLP in clinical settings

AI in medicine

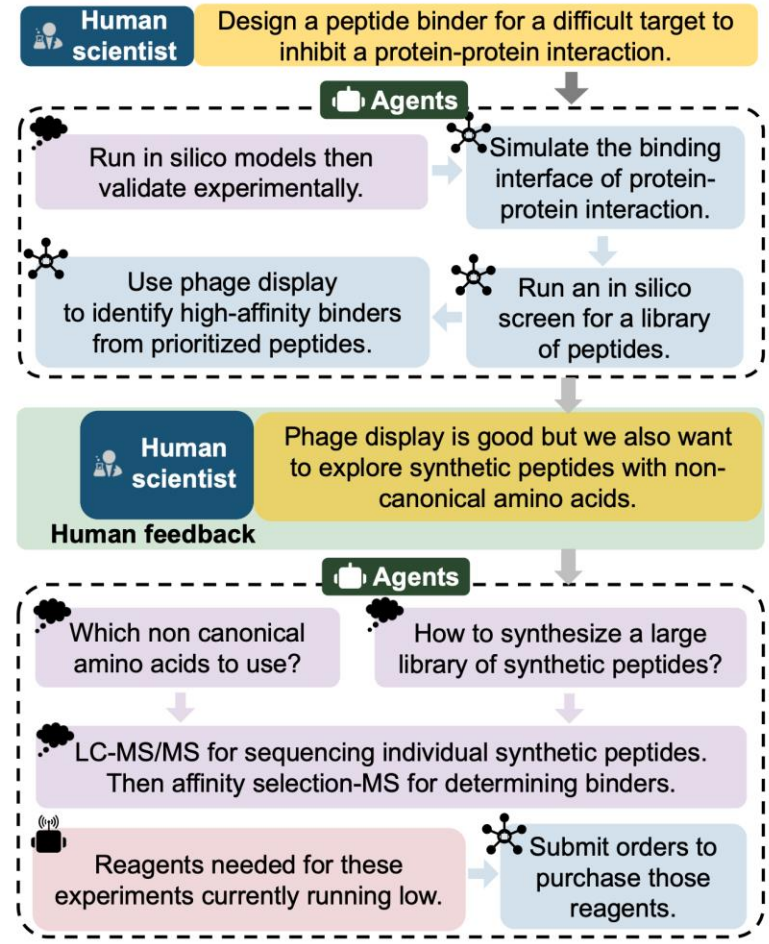
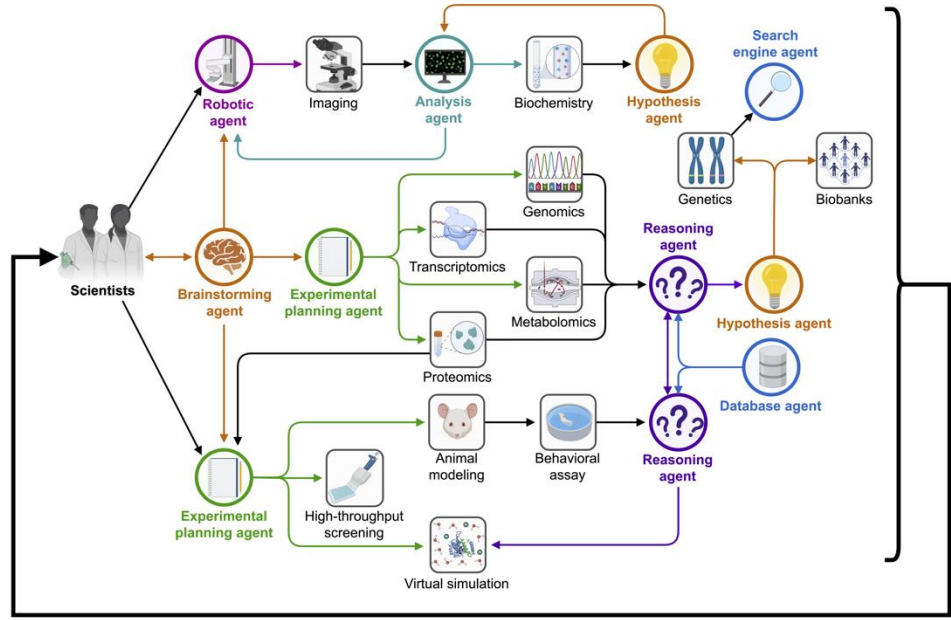


New fundamental results in molecular dynamics

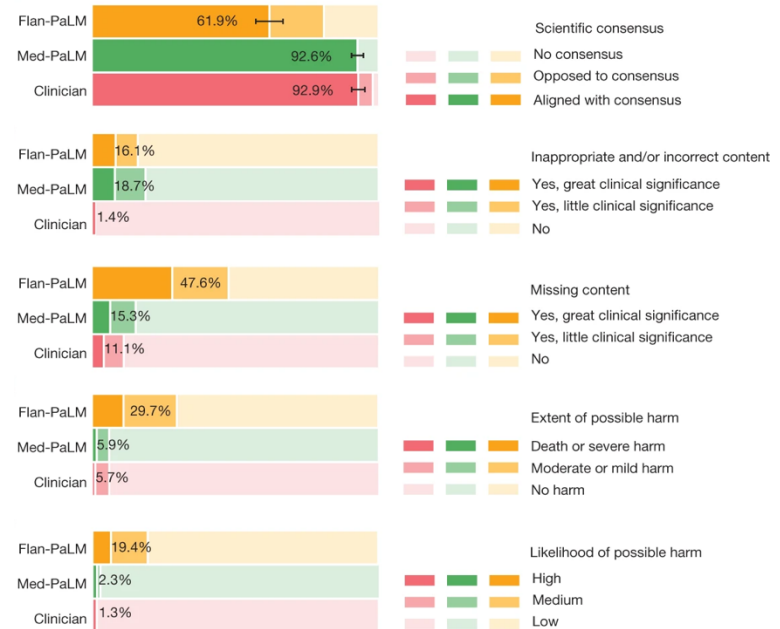
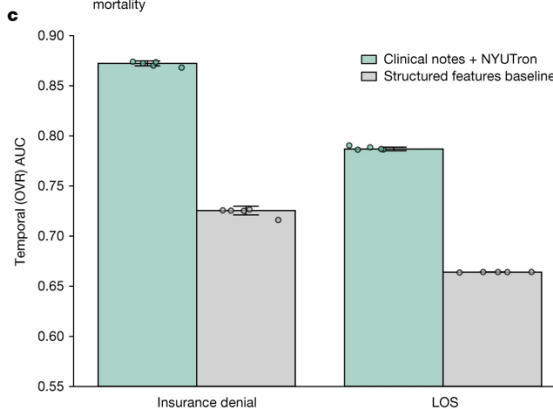
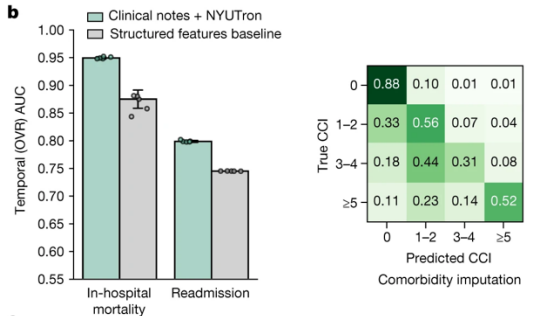
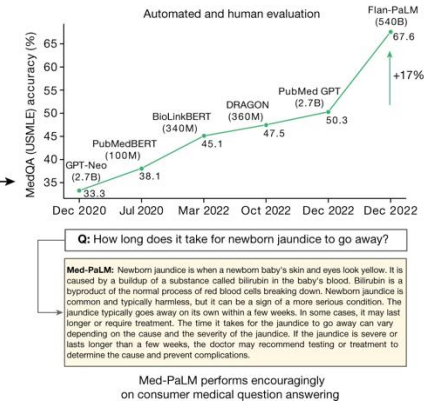
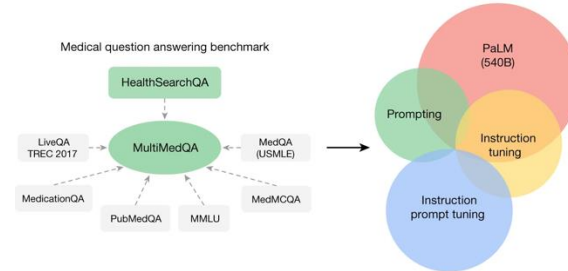
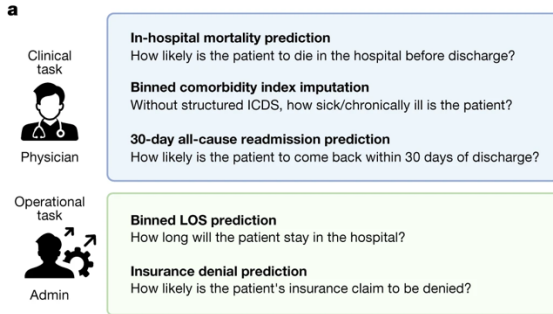
“AI scientists” as generative AI agents

A long-standing ambition for biomedical AI is the development of AI systems that can make major discoveries with the potential to be worthy of a Nobel Prize—fulfilling the Nobel Turing Challenge

d. Reasoning with feedback for alternative experimental approach



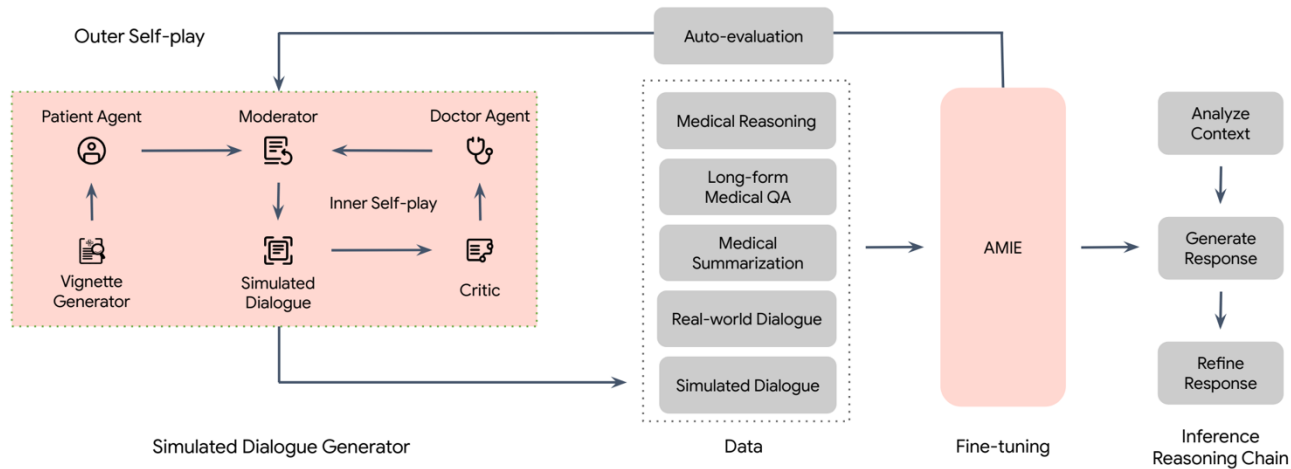
AI in healthcare



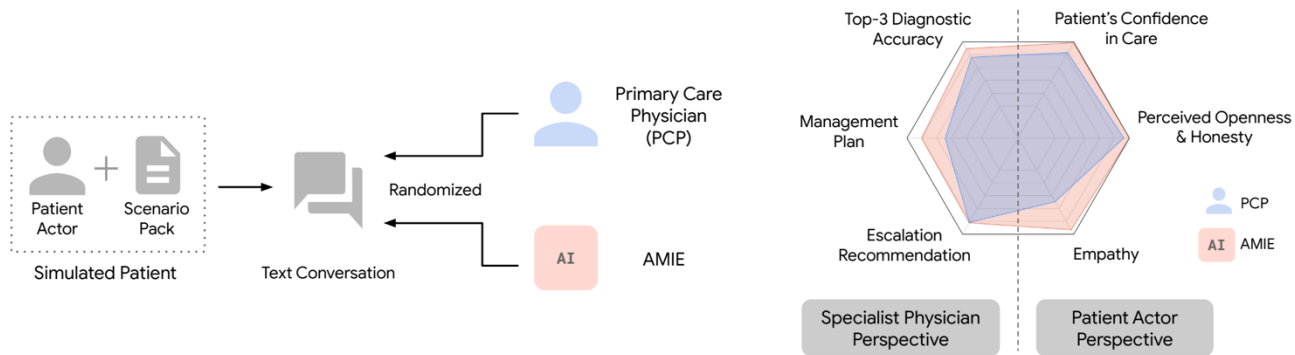
Health system-scale language models are all-purpose prediction engines, *Nature* 2023

Large language models encode clinical knowledge, *Nature* 2023

“AI doctors”: Conversational medical AI optimized for diagnostic dialogue



AMIE System Design

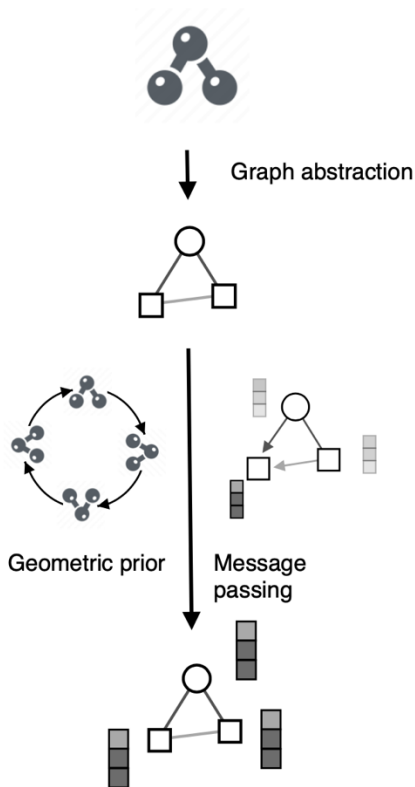


Randomized Study Design for Remote Objective Structured Clinical Examination (OSCE)

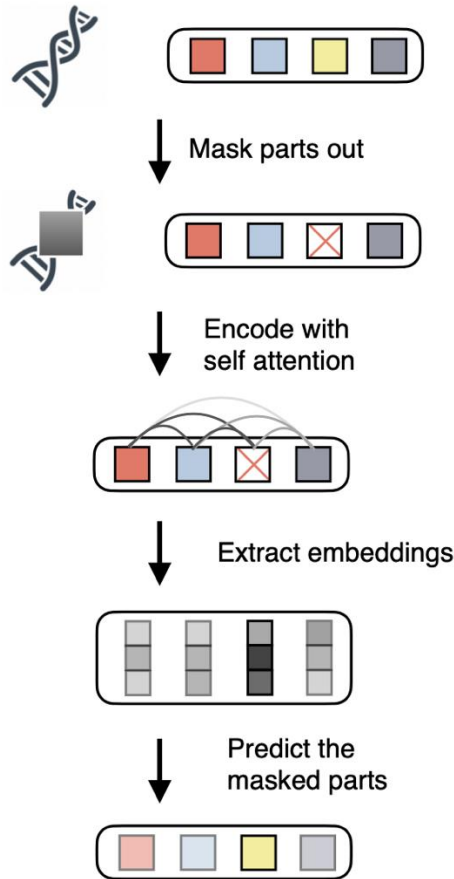
AMIE Outperforms PCPs on Multiple Evaluation Axes for Diagnostic Dialogue

Key algorithmic advances

Geometric learning

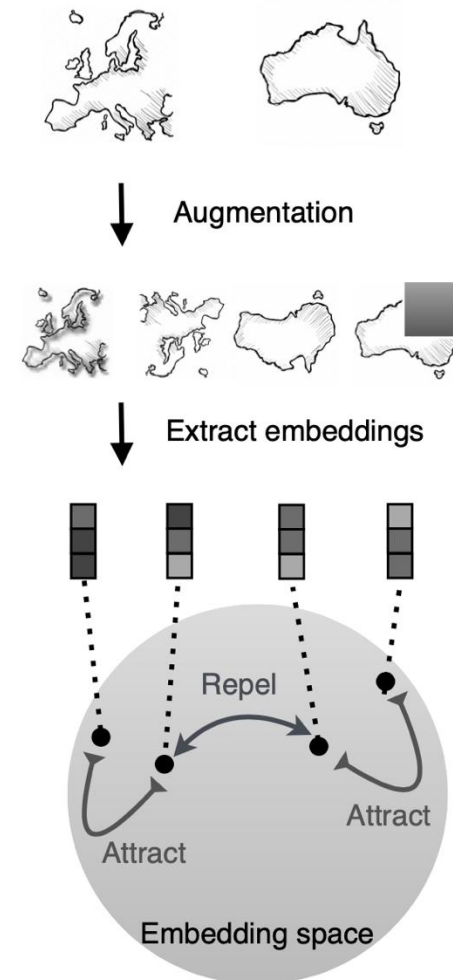


Self-supervised learning



    Labeled parts  Masked parts

Generative AI



What makes biomedical data so different?

- Life or death decisions
 - Need **robust** algorithms
 - Checks and balances built into ML deployment
 - (Also arises in other applications of AI such as autonomous driving)
 - Need **fair** and **accountable** algorithms
- Many questions are about **unsupervised learning**
 - Discovering disease subtypes, or answering question such as “characterize the types of people that are highly likely to be readmitted to the hospital”?
- Many of the questions we want to answer are **causal**
 - Naïve use of supervised machine learning is insufficient

What makes biomedical data so different?

- ML models are increasingly deployed in real-world applications and implemented in clinical settings:
 - It is critical to ensure that these models are behaving responsibly and are trustworthy
- Accuracy alone is no longer enough
- Auxiliary criteria are important:
 - **Explainable predictions and interpretable models**
 - **Fair and non-discriminatory predictions**
 - **Privacy-preserving, causal, and robust predictions**
- This broad area is known as **trustworthy ML**



High-stakes decisions

What makes biomedical data so different?

- Very little labeled data
- Recent breakthroughs in AI depended on lots of labeled data!

Large, diverse data
(+ large models)



Broad generalization



Russakovsky et al. '14

GPT-2

Radford et al. '19

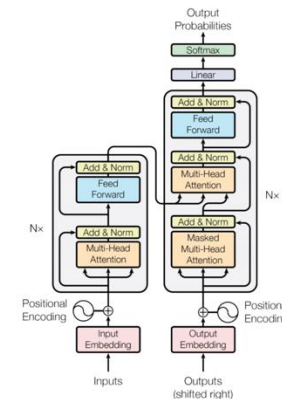


Figure 1: The Transformer - model architecture.

Vaswani et al. '18

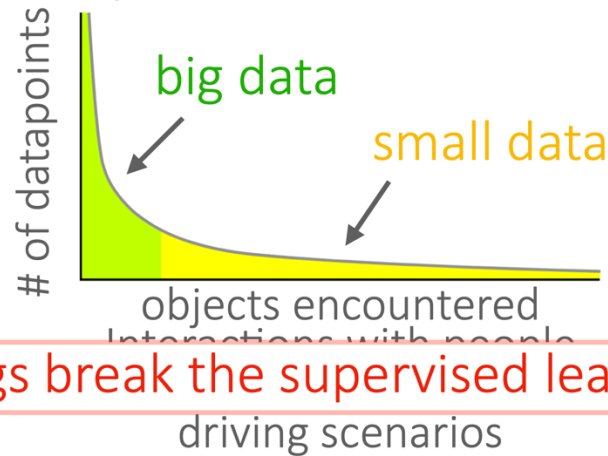
What if you don't have a large dataset?

medical imaging robotics personalized education,
translation for rare languages recommendations

What if you want a general-purpose AI system in the real world?

Need to continuously adapt and learn on the job.
Learning each thing from scratch won't cut it.

What if your data has a long tail?



These settings break the supervised learning paradigm.

What makes biomedical data so different?

- Very little labeled data
 - Motivates **semi-supervised and self-supervised learning**
- Sometimes small numbers of samples (e.g., a rare disease)
 - **Learn as much as possible from other data** (e.g., from healthy patients)
 - Model the problem **carefully**
- Lots of **missing data, varying time intervals, censored labels**

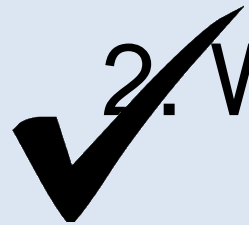
What makes biomedical data so different?

- Difficulty of **de-identifying** data:
 - Need for **data sharing agreements** and **sensitivity**
- Difficulty of **deploying ML**:
 - Commercial electronic health record software **is difficult to modify**
 - Data are often in siloes; everyone recognizes need for **interoperability**, but slow progress
 - **Rigorous testing and iteration** are needed
- Difficulty of **correcting for biases and inequities**:
 - Consideration of ethical and legal issues
 - Health data on which algorithms are trained are likely to be influenced by **many facets of social inequality**

Outline for today's class



1. Overview of this course



2. What makes biomedical data unique



3. Introduction to distributed language representations

4. Introduction to NLP in clinical settings

Distributed word representations

“apple” is a **polysemic** word...



🔍 grow an apple

🔍 buy an apple

Distributed word representations

... whose **particular meaning** is resolved via **sentence context**



🔍 grow an apple

🔍 grow an apple **tree**

🔍 grow an apple **tree from seed**

🔍 grow an apple **tree in a pot**

🔍 grow an apple **tree indoors**



🔍 buy an apple|

🔍 buy an apple **watch**

🔍 buy an apple **gift card**

🔍 buy an apple **tv**

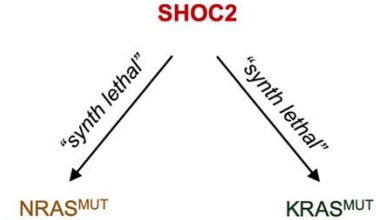
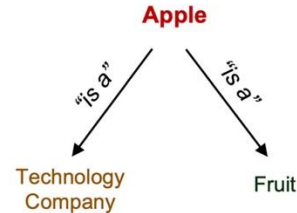


Distributional hypothesis

The **Distributional hypothesis** is that words that occur in the same contexts tend to have similar meanings (Harris, 1954).

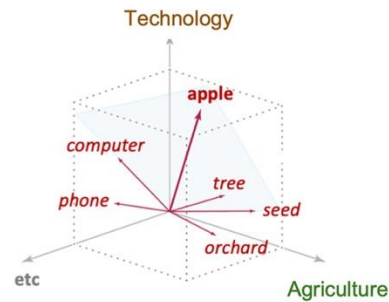
The underlying idea that "a word is characterized by the company it keeps" was popularized by Firth (1957).

Distributional hypothesis



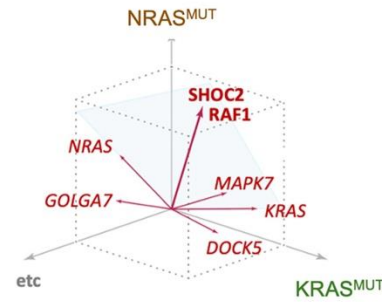
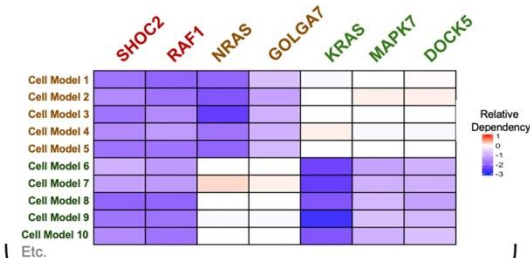
B Distributional Hypothesis of Word Meaning

1. I mainly use my **Apple** **iPhone** to make **phone** calls.
 2. The **Apple** MacBook Pro is a **computer** with a powerful **processor**.
 3. I use an **Apple** **computer** to write **emails** and create **documents**.
 4. I picked a red **apple** from the **tree** in the backyard.
 5. The planted **seeds** in the **orchard** produced several **apple trees**.
 6. **Apples** are my favorite type of **fruit**.
- Etc.



Word Polysemy

Distributional Hypothesis of Gene Function



Gene Pleiotropy

Intuition

“Probability of a sentence” = how likely is it to occur in natural language

Example 1: Syntax and grammatical properties

$$p(\textit{the cat purrs}) > p(\textit{cat purrs the})$$

Example 2: Semantic properties

$$p(\textit{the cat purrs}) > p(\textit{the cat smokes})$$

What about the probability of "the Archaeopteryx winged jaggedly amidst foliage"?

Probabilistic model of language

Probability model:

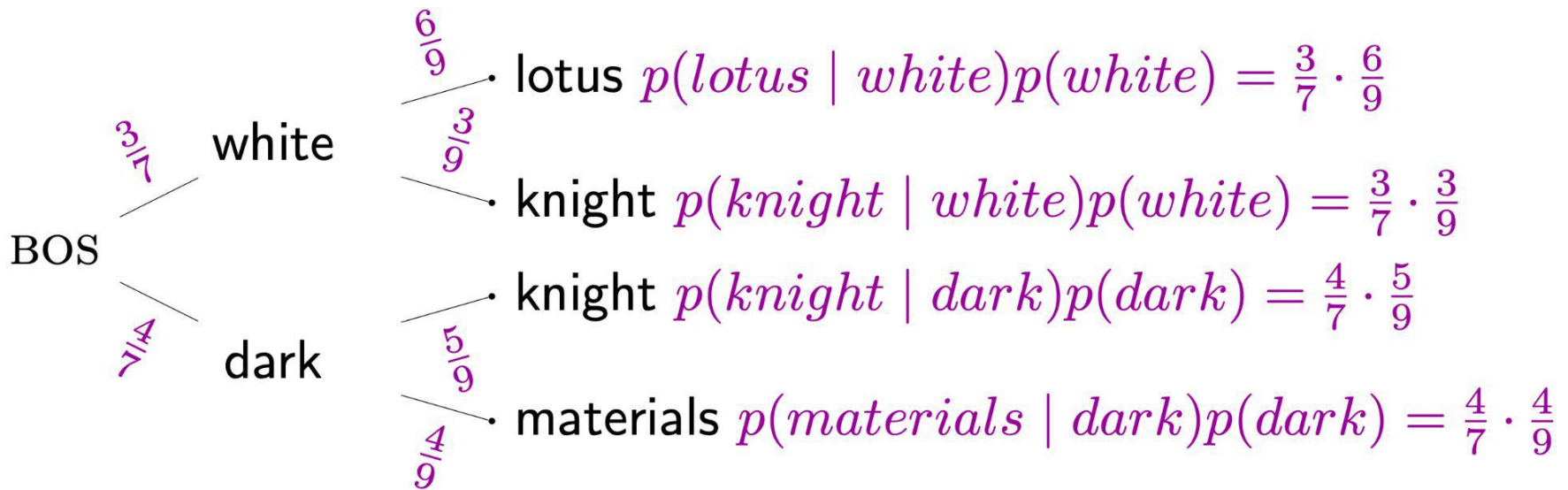
1. $p(L) = 1$

2. $p(\bigcup_{i=1}^n \mathcal{E}_i) = \sum_{i=1}^n p(\mathcal{E}_i)$ if $\mathcal{E}_1, \mathcal{E}_2, \dots$ is a countable sequence of disjoint sets of $\mathcal{P}(L)$, the power set (=set of all subsets) of L .

3. (Conditional probability)
$$p(\mathbf{x}) = p(x_0) \prod_{i=1}^L p(x_i | x_1, \dots, x_{i-1})$$

$$\log p(\mathbf{x}) = \log p(x_0) \sum_{i=1}^L \log p(x_i | x_1, \dots, x_{i-1})$$

Example



Estimation

We assume there is some true p^* which we estimate/approximate with a (parametric) estimator) \hat{p} which is an element of $\{p_\theta \mid \theta \in \Theta\}$.

This is done by learning from data $\mathcal{D} = \{\mathbf{x}_1, \dots, \mathbf{x}_n\} \subseteq L$, e.g. by minimizing some loss:

$$\hat{\theta} \triangleq \arg \min_{\theta \in \Theta} \ell(\theta, \theta^*)$$

Since the optimal model is unknown, we use the data as an estimate:

$$p_{\theta^*} \approx \frac{1}{|\mathcal{D}|} \sum_{i=1}^{\mathcal{D}} \delta_{\mathbf{x}_i}(\mathbf{x}) \quad \delta_{\mathbf{x}_i}(\mathbf{x}) \triangleq \begin{cases} 1 & \text{if } \mathbf{x}_i = \mathbf{x} \\ 0 & \text{else} \end{cases}$$

How to learn a model from the data?

A suitable loss function is the KL-Divergence (divergence between prob. distributions):

$$\text{KL}(p_{\theta^*}, p_{\hat{\theta}}) \triangleq \underbrace{-\sum_{\mathbf{x} \in L} p_{\theta^*}(\mathbf{x}) \log p_{\hat{\theta}}(\mathbf{x})}_{\text{Cross-Entropy}} + \underbrace{p_{\theta^*}(\mathbf{x}) \log p_{\theta^*}(\mathbf{x})}_{-H(p_{\theta^*})}$$

constant wrt. model param.

Justification:

From Information Theory: Measures the excess number of bits we pay by encoding our data with a sub-optimal model. The optimum is just the entropy (Shannon, 1948).

N-Gram models

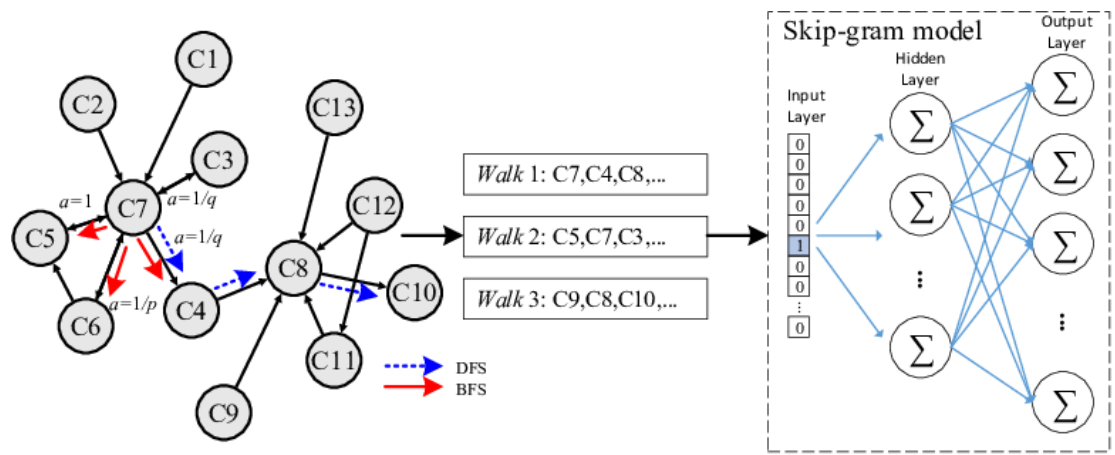
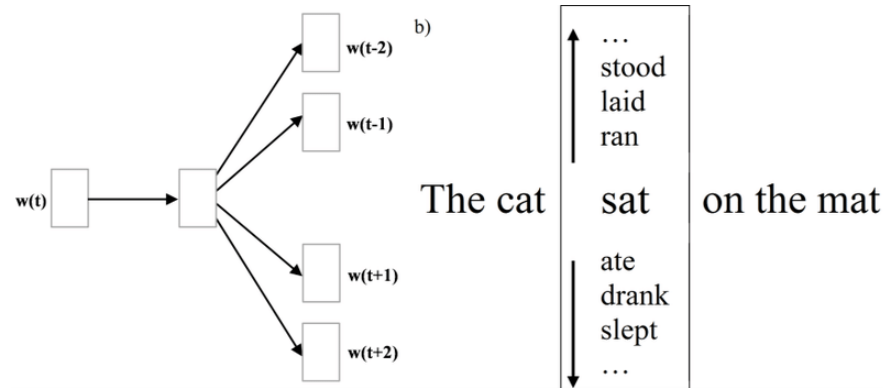
We can obtain a very simple form for $\{p_\theta \mid \theta \in \Theta\}$ by making the Markov assumption:

$$\begin{aligned} p(\mathbf{x}) &= p(x_1, \dots, x_n) \\ &= p(x_n | x_1, x_2, \dots, x_{n-1}) p(x_{n-1} | x_1, x_2, \dots, x_{n-2}) \dots p(x_1) \\ &\approx p(x_n | x_{n-2}, x_{n-1}) p(x_{n-1} | x_{n-3}, x_{n-2}) \dots p(x_1) \end{aligned}$$

This is a tri-gram model (history of two). Assumes all of these are equal:

- $p(\text{slept} | \text{the cat})$
- $p(\text{slept} | \text{after lunch the cat})$
- $p(\text{slept} | \text{the dog chased the cat})$
- $p(\text{slept} | \text{except for the cat})$

Word2vec, node2vec, sentence2vec, and many others



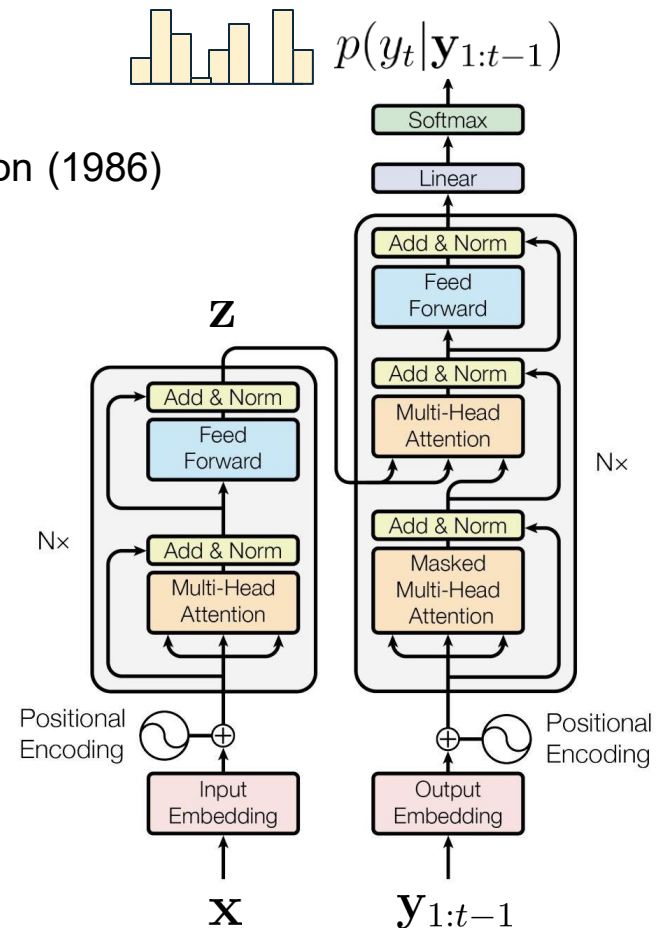
Transformer architecture

Probably the most influential ML paper since Backpropagation (1986)

→ Over 150K citations since 2017

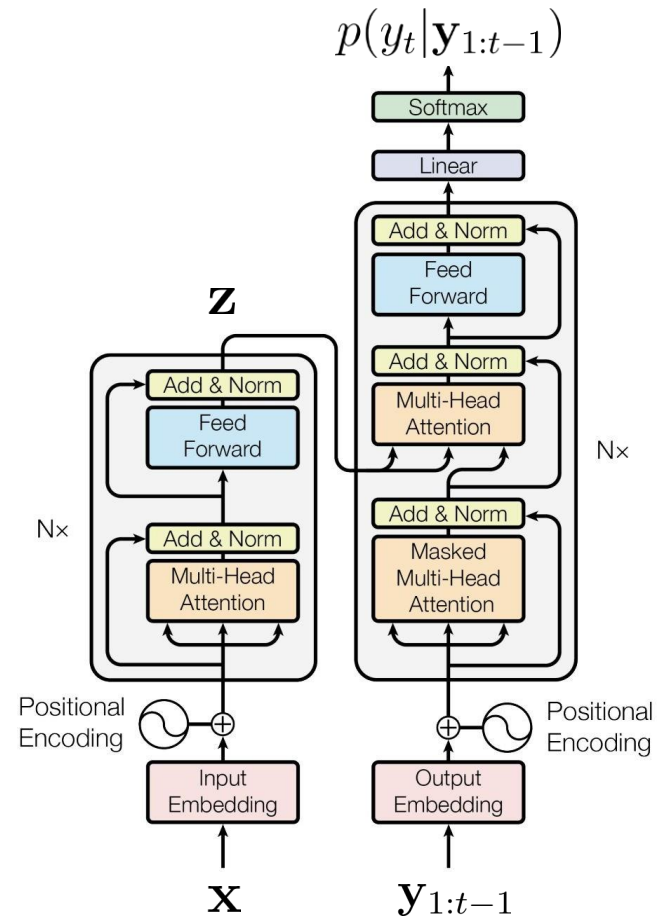
→ Essentially replaced RNNs for most purposes

A simple sequence to sequence model mapping an input (x_1, \dots, x_n) (tokenized and "embedded") into a continuous representation $\mathbf{z} = (z_1, \dots, z_n)$ based on which the decoder produces (y_1, \dots, y_m) autoregressively, i.e. one symbol at a time.



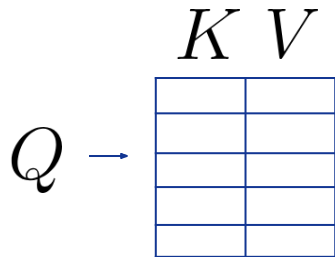
Transformer building block

1. Attention Mechanism
2. Position Encodings
3. Residual connections + Normalization



Attention mechanism

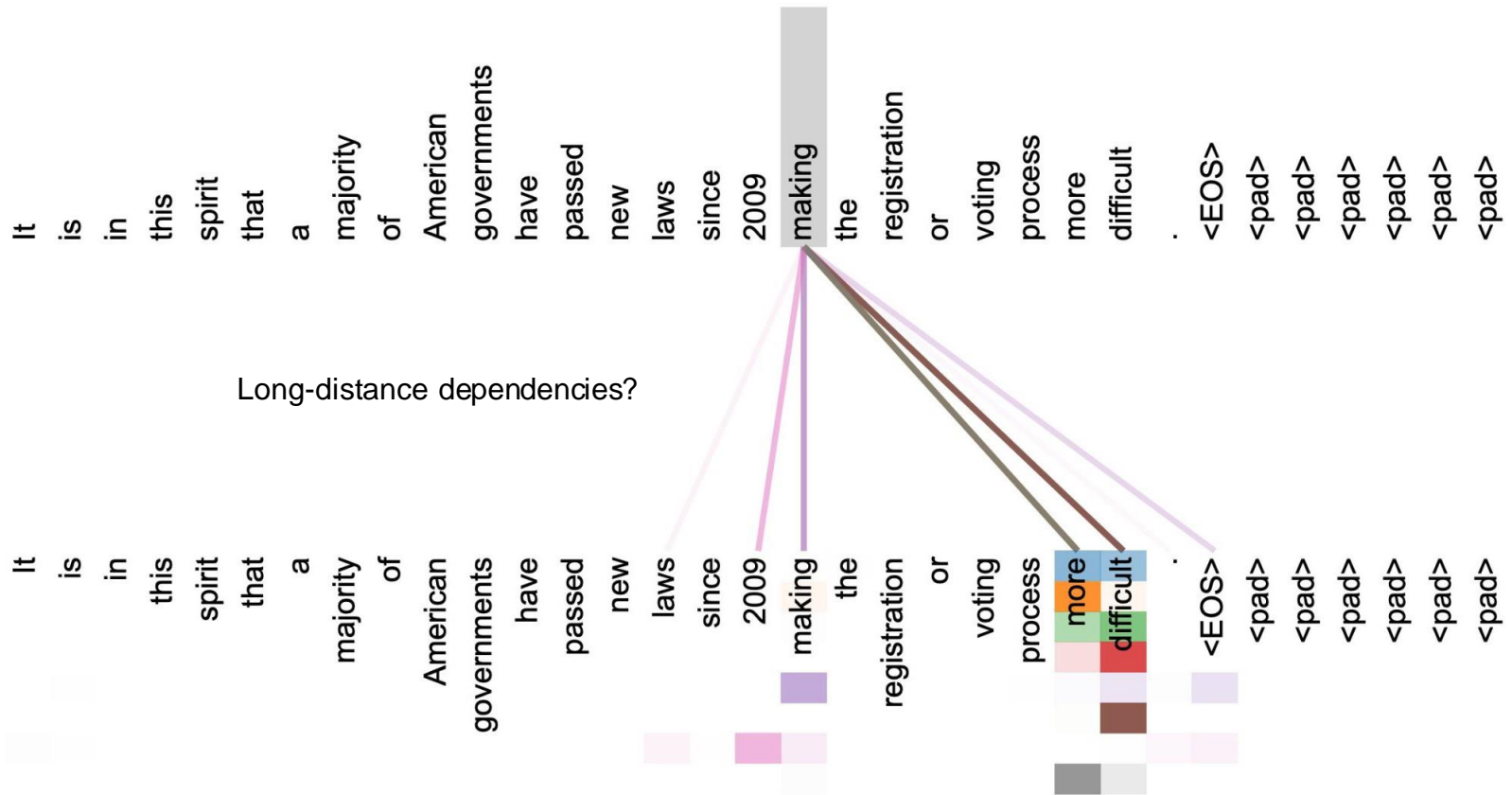
$$H = \text{Attention}(QW^Q, KW^K, VW^V)$$



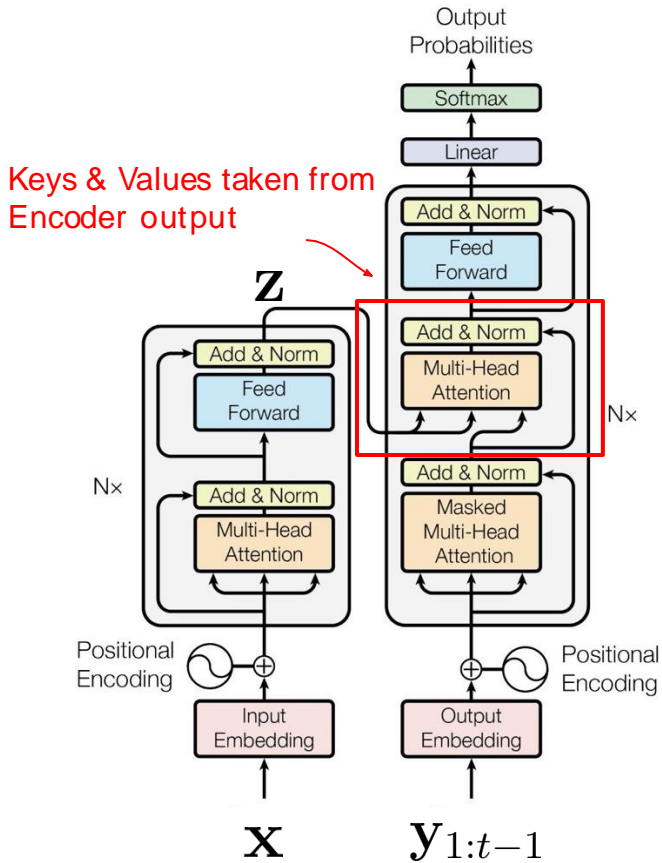
Think of this as a soft "look-up" operation in an associative memory using dot-products as a similarity measure.

$$W^Q \in \mathbb{R}^{d_{\text{model}} \times d_k}, W^K \in \mathbb{R}^{d_{\text{model}} \times d_k}, W^V \in \mathbb{R}^{d_{\text{model}} \times d_v}$$

Dot-product attention mechanism



Encoder-decoder transformer



$$H_i^{(l)} = \text{Attention}(QW_i^Q, K_iW^K, V_iW^V)$$

$$Q = Y^{(l-1)}$$

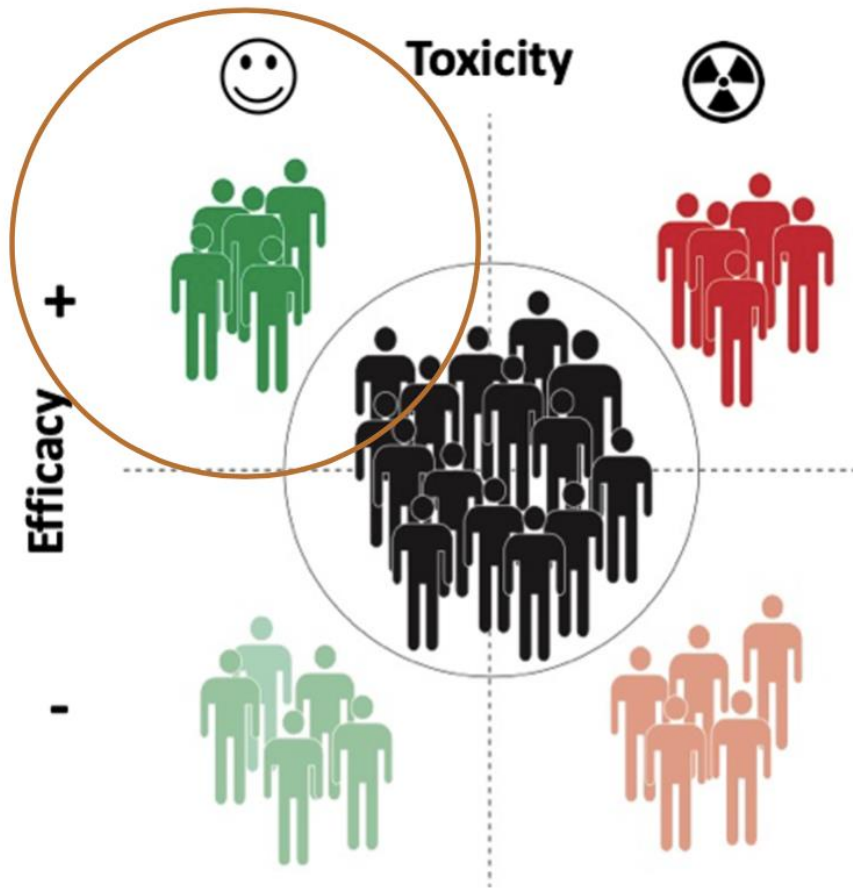
$$K = V = Z^{(l-1)}$$

Encoder- Decoder

Outline for today's class

- ✓ 1. Overview of this course
- ✓ 2. What makes biomedical data unique
- ✓ 3. Introduction to distributed language representations
- ✎ 4. Introduction to NLP in clinical settings

Precision medicine goals

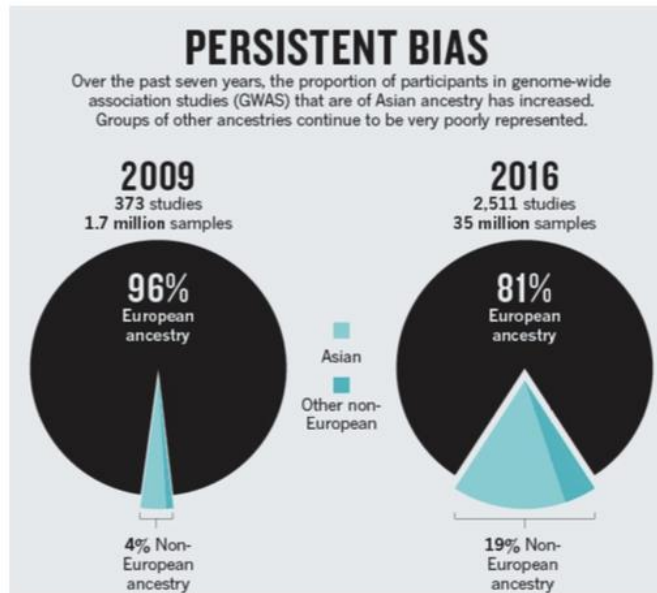


<http://hitconsultant.net/2014/04/03/infographic-the-rise-of-personalized-medicine/>

Why are these goals relevant?

Problem: Underrepresentation in clinical research

Genomics



Popejoy and Fullerton, *Nature*, 2016

Clinical Trials

Participation in Cancer Clinical Trials Race-, Sex-, and Age-Based Disparities

Table 1. Participants in National Cancer Institute Cooperative Group Breast, Colorectal, Lung, or Prostate Cancer Therapeutic Trials, 1996-2002 (N = 75 215)*

Characteristic	Trial Participants, No. (%)	Proportion of Incident Cancer Patients, %†	Proportion of US Population, %‡
Race/ethnicity			
White non-Hispanic	64 355 (85.6)	83.1	75.7
Hispanic	2292 (3.1)	3.8	9.1
Black	6882 (9.2)	10.9	10.8
Asian/Pacific Islander	1446 (1.9)	2.0	3.8
American Indian/Alaskan Native	240 (0.3)	0.2	0.7

Murthy et al., *JAMA*, 2004.



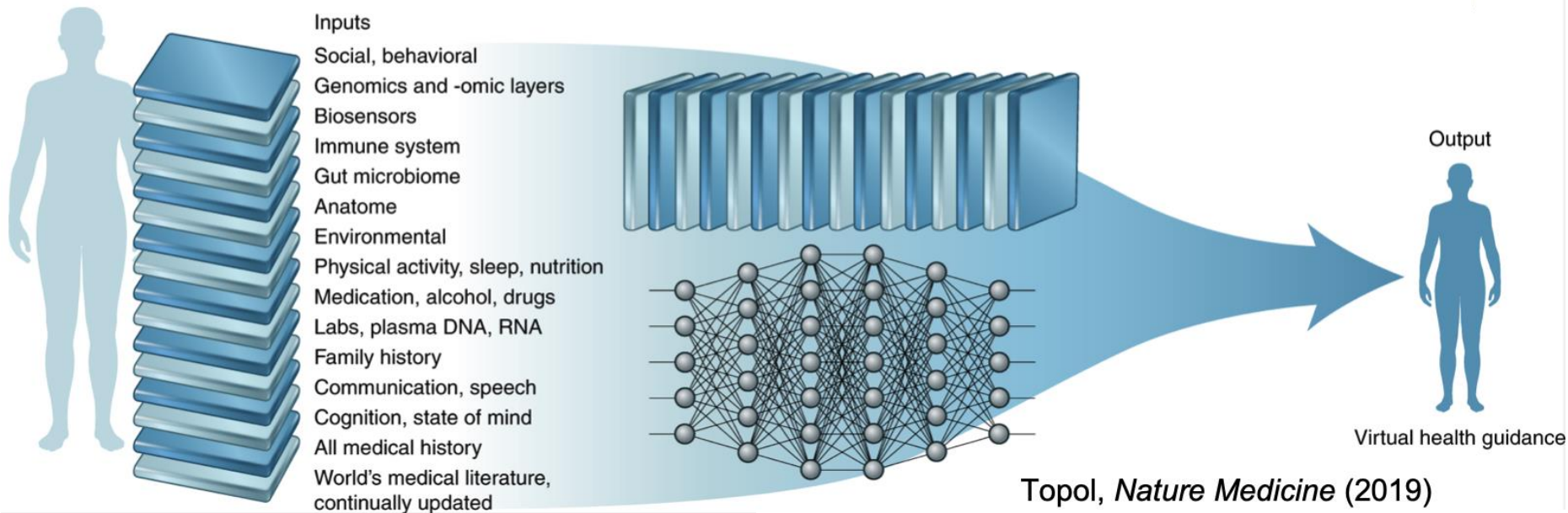
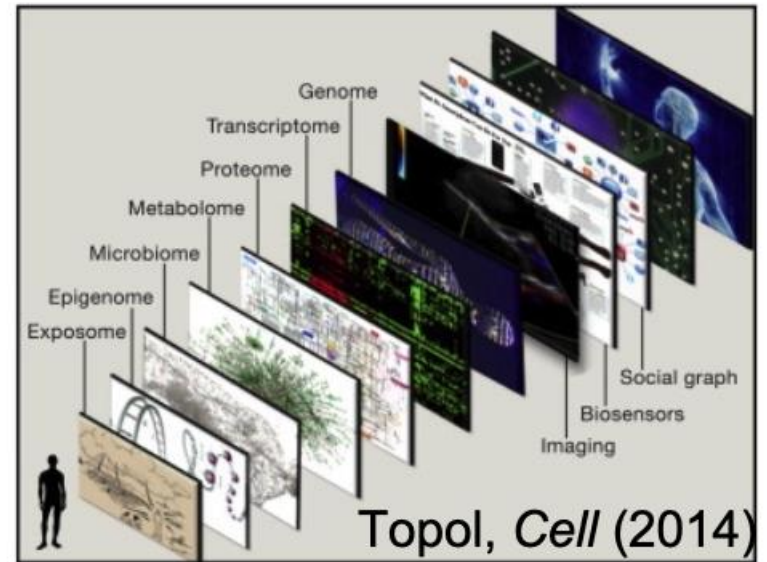
inferential gap



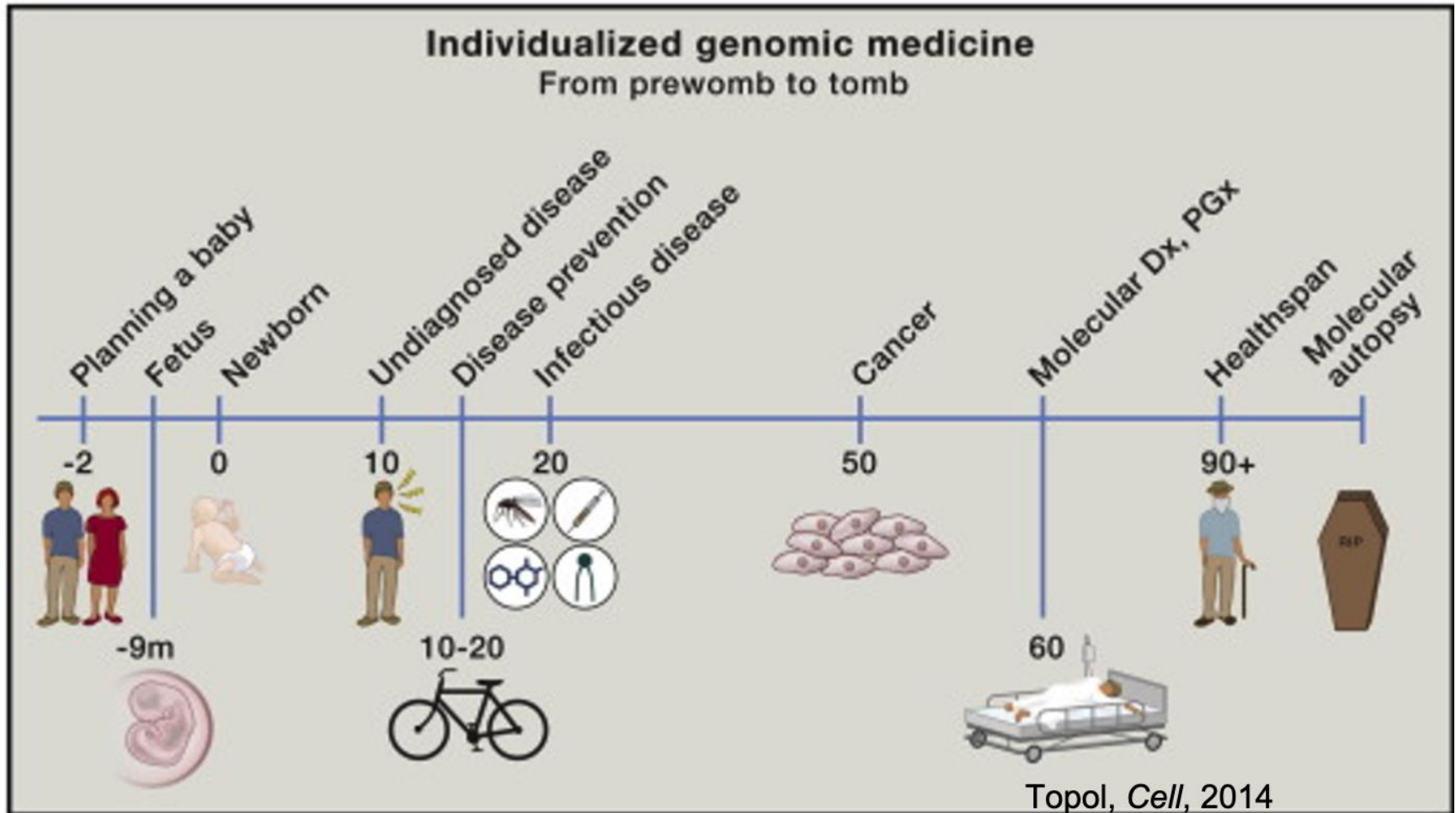
Most clinical decisions involve bridging the **inferential gap**: Clinicians are required to “fill in” where they lack knowledge or where no knowledge yet exists:

- Misdiagnoses, medical errors, prescription errors, surgical errors, under-treatments, over-treatments, unnecessary lab tests can be due to inferential gaps
- Late diagnosis of cancer can be due to the inferential gaps at the primary care
- Crisis caused by misuse, underuse, or overuse of antibiotics is in part due to serious inferential gaps

Precision medicine requires a multi-level understanding of health and disease...

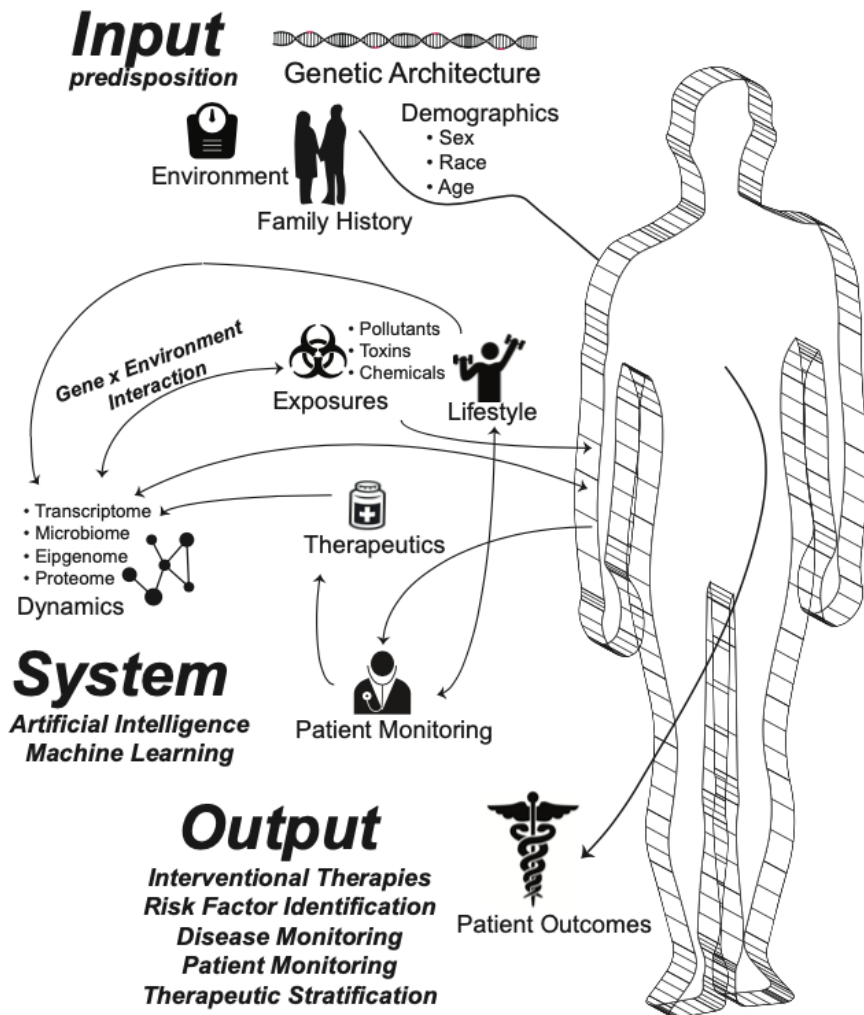


...and understanding how health and disease states evolve



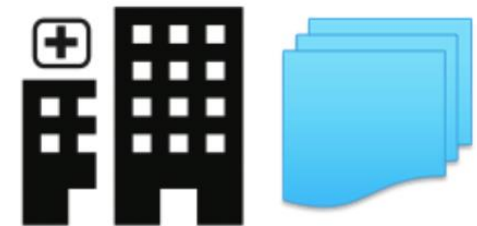
This all-encompassing dataset does not exist...

“The Quantified Self”



... but real-world data can serve as proxy

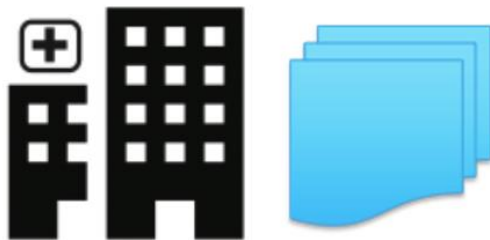
Electronic Health Records



Next: How are electronic health records used for research?

Electronic health records

- The digitized paper charts
- The underlying goal/purpose of EHRs is **billing/infrastructure**
- Contains any data collected during an individual's interaction with a medical system
- Different software vendors (e.g., EPIC, Cerner)



Data type examples:

Clinical

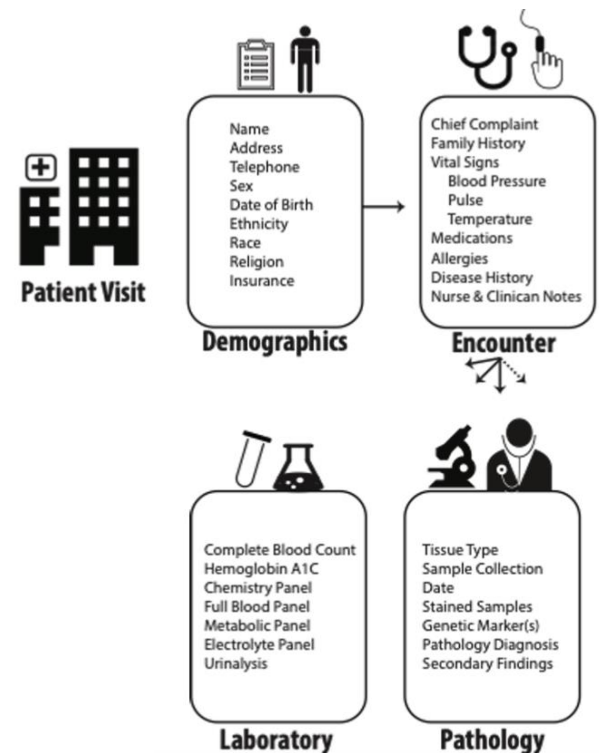
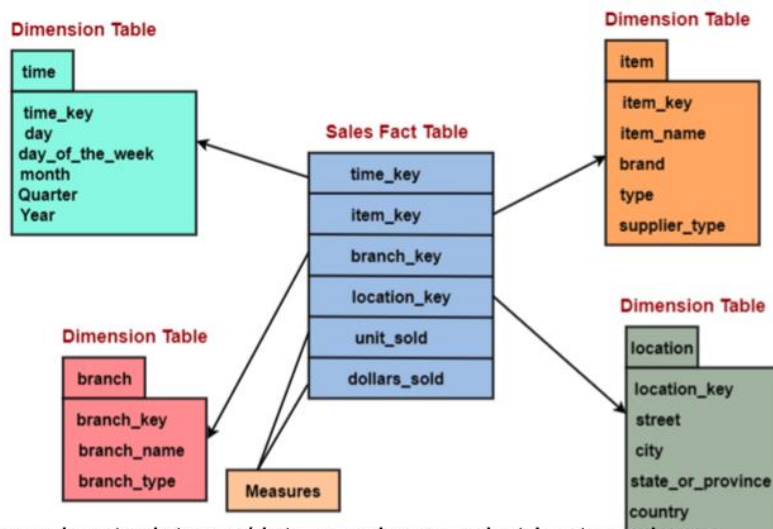
- Diagnoses
- Procedures
- Lab test results
- Imaging
- Medications
- Notes

Non-clinical

- Demographics
- Insurance
- Location
- Lifestyle

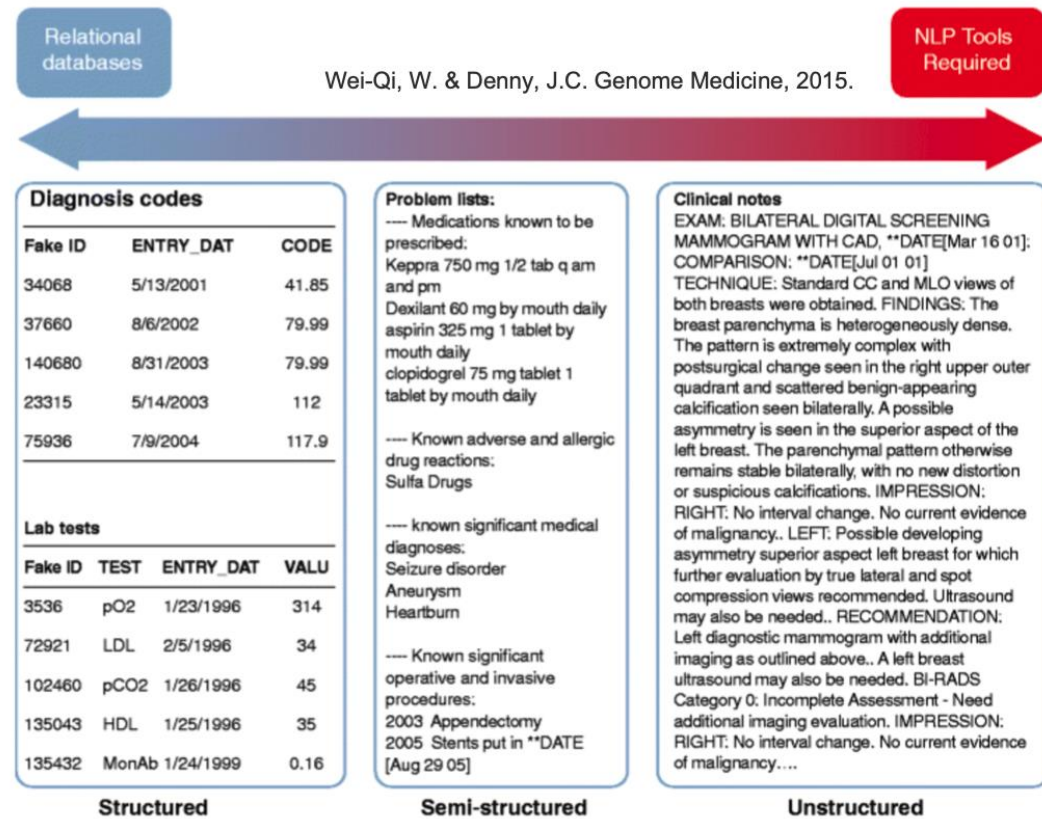
EHR data types and formats

- Made available by data warehouses
- Are often *encounter-based*
- Typically separated by modality (e.g., demographics table, lab table)
- Often in star-schema format



EHR data structure

- **Structured:** labs, medications, etc.
- **Semi-structured:** smartforms, radiology impressions, echo reports
- **Unstructured:** clinical notes
- **Note:** It does not have all data!

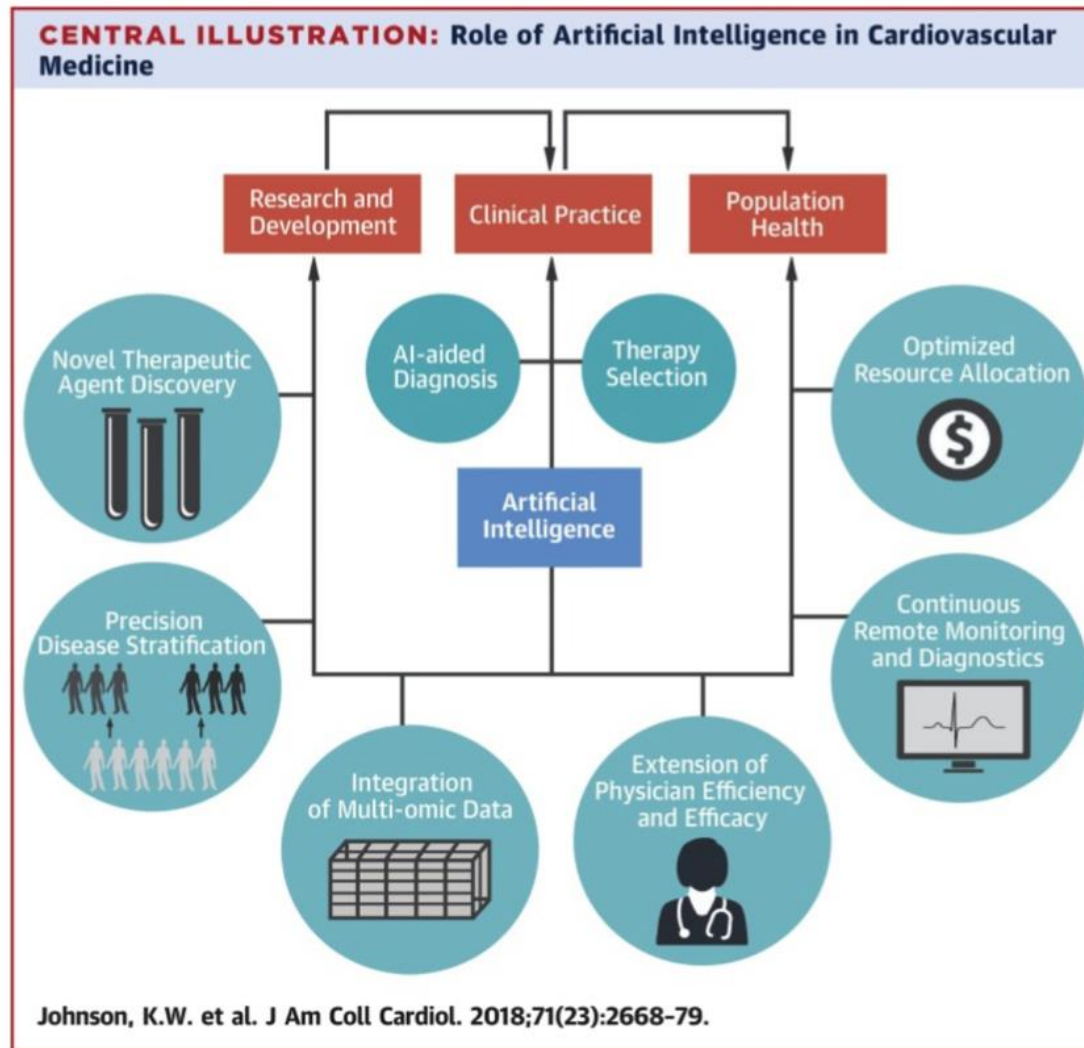


Types of research using EHRs?

- Characterize co-morbidities & epidemiological trends
- Identify disease sub-phenotypes
- Identify unknown drug adverse events
- Find symptom clusters
- Predict medication response
- Anticipate disease flare-ups
- Guide triage decisions
- Track treatment progression and sequelae
- Couple with other patient data modalities: genetics, images, notes, biosignals, etc.

+ countless more...

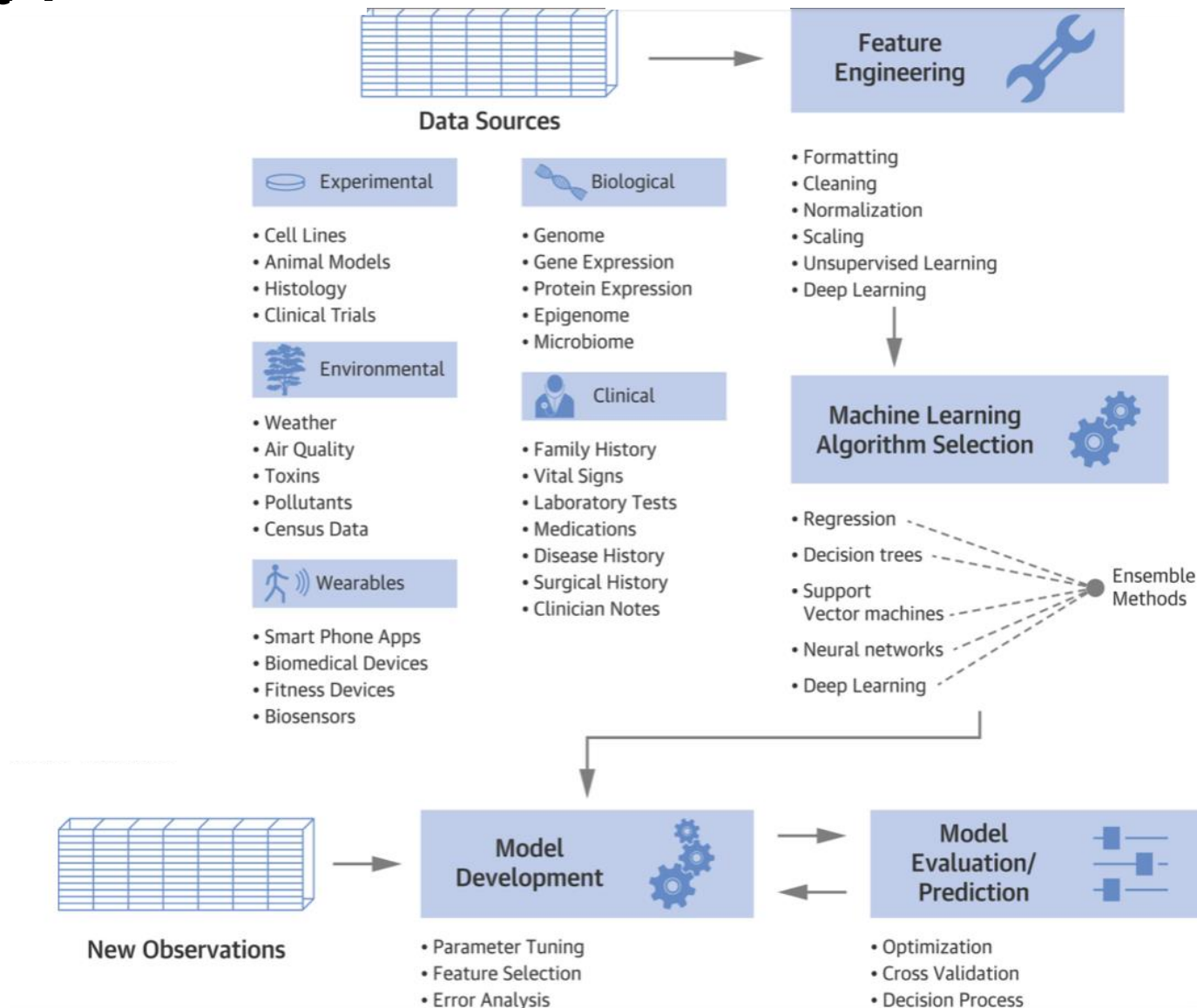
Goals of ML for healthcare using EHR



Typical ML workflow for EHR data

- Gather (identify relevant feature)
- QC values (wrong unit?)
- Check for/address missingness
- Phenotype and design cohort
- Define outcome (label) and study period
- Use relevant ML techniques
- Pre-process data to fit the ML technique
- Refine and repeat

Typical ML workflow for EHR data



Johnson et al., JACC, 2018

What is a disease?

- A disease is not easily defined in EHRs!
- Many ways in which a disease can be represented (and often wrong)
- Phenotyping algorithms and standardized concepts to **the rescue**: accurately identify patients with a specific observable trait from imperfect EHR data



How well do various data types define a disease? (1/3)

- **Goal:** Evaluate phenotyping performance of major EHRs
 - Diagnosis codes
 - Primary notes
 - Medication lists
- **Approach:**
 - Select ten diseases: atrial fibrillation, Alzheimer's disease, breast cancer, gout, human immunodeficiency virus infection, multiple sclerosis, Parkinson's disease, rheumatoid arthritis, and T1D/T2D
 - For each disease, classify patients into seven categories based on the presence of evidence for disease in a) diagnosis codes, b) primary notes, and c) specific medications
 - For each disease, select 175 patients for **manual chart review**
 - Use review results to estimate **positive predictive value (PPV)** for each EHR data type alone and in combination

How well do various data types define a disease? (2/3)

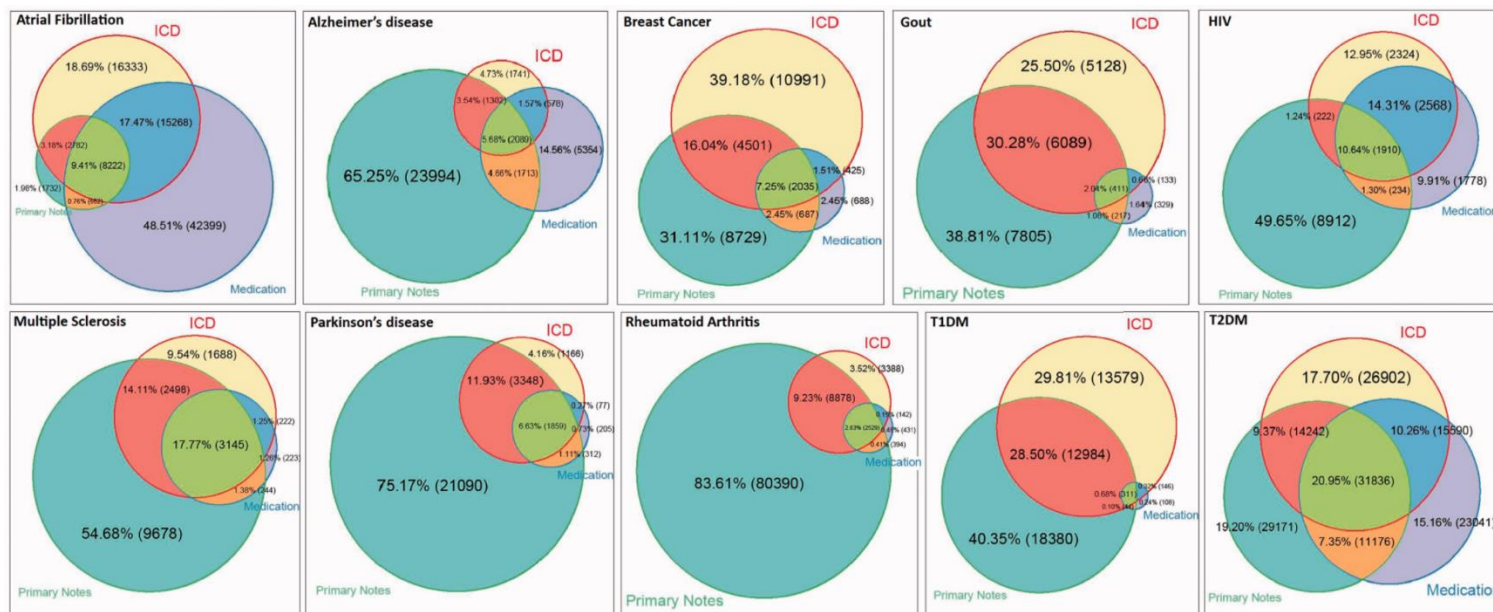
- PPV is the ratio of patients that truly have the disease according to manual chart review to all patients who had been identified as having the disease
- PPVs on single data types were inadequate for accurate phenotyping (0.06–0.71)
- Using two or more ICD codes improved the average PPV to 0.84

Positive prediction values of various categories based on chart review results

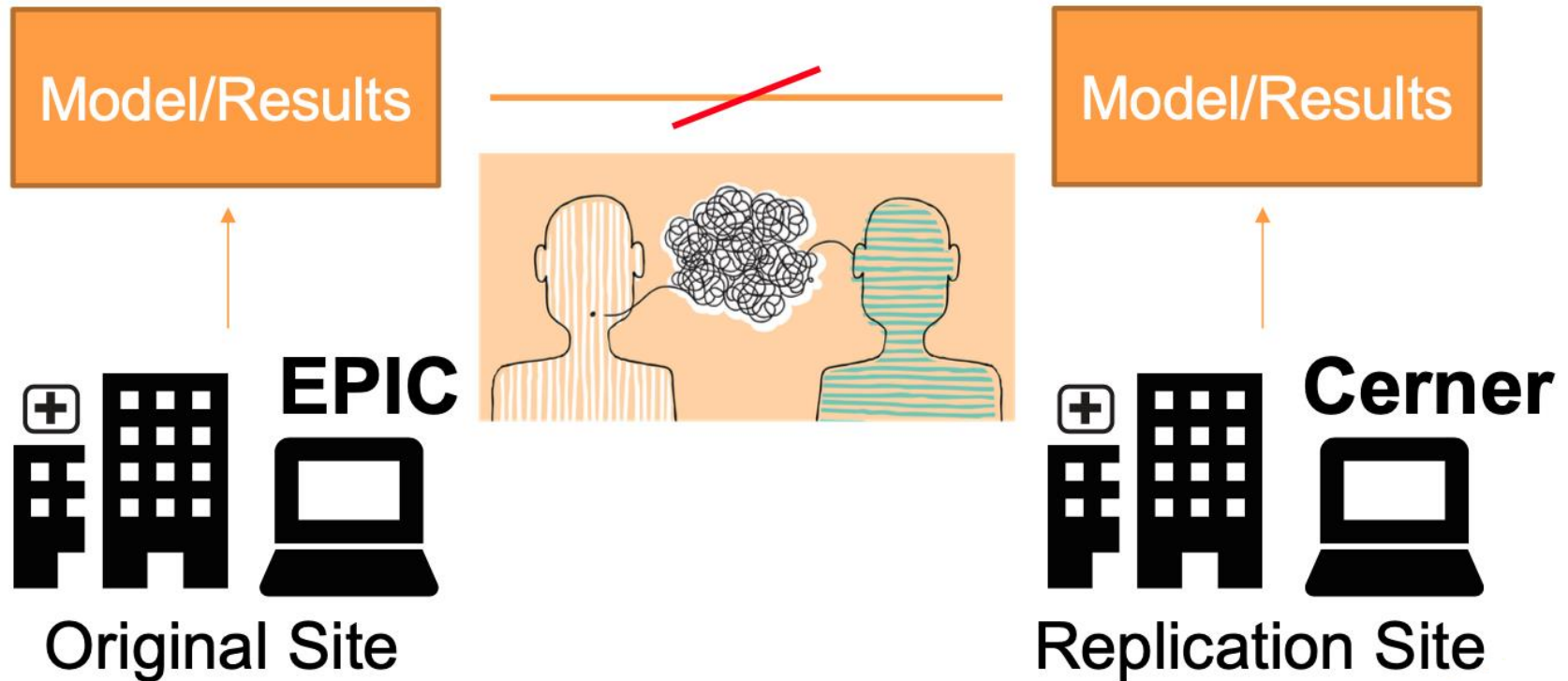
Disease	ICD-9 Only	PN Only	Meds Only	ICD-9+Meds	ICD-9+PN	Meds+PN	ICD-9+both	ICD-9	Meds	PN	≥2 ICD-9s	≥2 Components
AFIB	0.52	0.72	0.08	0.72	1.00	1.00	1.00	0.72	0.35	0.96	0.88	0.84
Alzheimer's	0.28	0.20	0.00	0.80	0.88	0.92	0.88	0.69	0.40	0.32	0.74	0.88
Breast CA	0.12	0.72	0.04	0.88	0.96	1.00	1.00	0.45	0.81	0.84	1.00	0.97
Gout	0.56	0.84	0.00	0.92	1.00	1.00	1.00	0.81	0.69	0.91	0.93	1.00
HIV	0.52	0.00	0.00	0.92	0.84	0.88	1.00	0.81	0.69	0.20	0.89	0.95
MS	0.20	0.08	0.12	0.88	0.88	0.88	1.00	0.78	0.93	0.41	0.86	0.94
Parkinson	0.48	0.16	0.04	0.84	1.00	0.88	0.96	0.89	0.87	0.33	0.94	0.98
RA	0.36	0.20	0.00	0.64	0.76	0.88	0.84	0.68	0.73	0.27	0.77	0.78
T1DM	0.28	0.12	0.04	0.16	0.92	0.84	0.76	0.59	0.49	0.45	0.62	0.91
T2DM	0.36	0.68	0.24	0.60	0.80	1.00	0.84	0.65	0.65	0.80	0.73	0.81
Average	0.37	0.37	0.06	0.74	0.90	0.93	0.93	0.71	0.66	0.55	0.84	0.91
Standard Deviation	0.15	0.32	0.08	0.23	0.09	0.06	0.09	0.13	0.20	0.29	0.12	0.08

How well do various data types define a disease? (3/3)

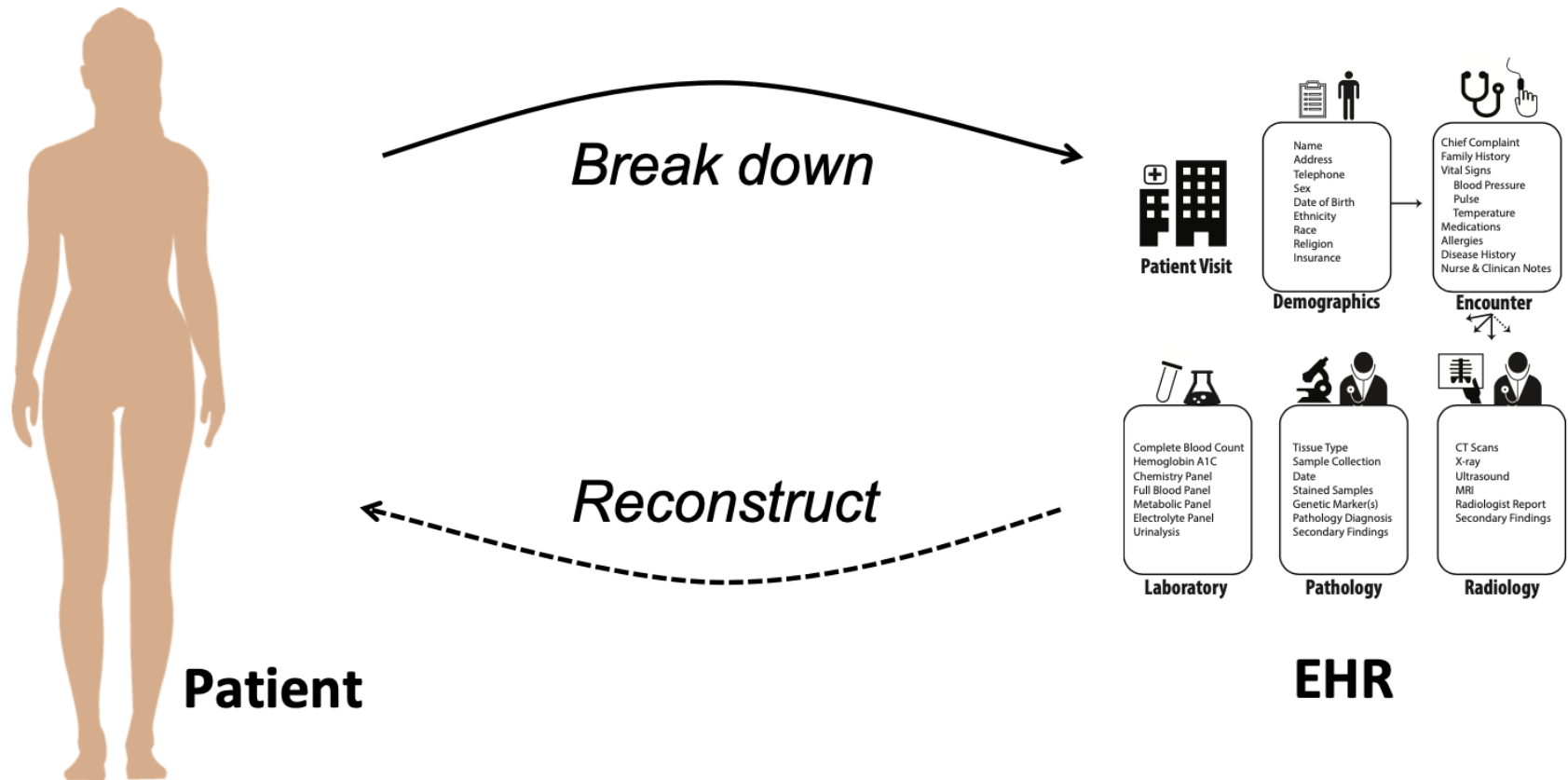
- Multiple data types provide a more consistent and higher performance than a single one
- **Use multiple EHR data types for disease phenotyping**



External replication is necessary but not easy to facilitate



It is challenging to capture health state from EHR



ML models can learn the wrong information

RESEARCH

OPEN ACCESS

Biases in electronic health record data due to processes within the healthcare system: retrospective observational study

Denis Agniel,¹ Isaac S Kohane,^{1,2} Griffin M Weber^{1,3}

RESULTS

The presence of a laboratory test order, regardless of any other information about the test result, has a significant association ($P < 0.001$) with the odds of survival in 233 of 272 (86%) tests. Data about the timing of when laboratory tests were ordered were more accurate than the test results in predicting survival in 118 of 174 tests (68%).

CONCLUSIONS

Healthcare processes must be addressed and accounted for in analysis of observational health data.

Without careful consideration to context, EHR data are unsuitable for many research questions. However, if explicitly modeled, the same processes that make EHR data complex can be leveraged to gain insight into patients' state of health.

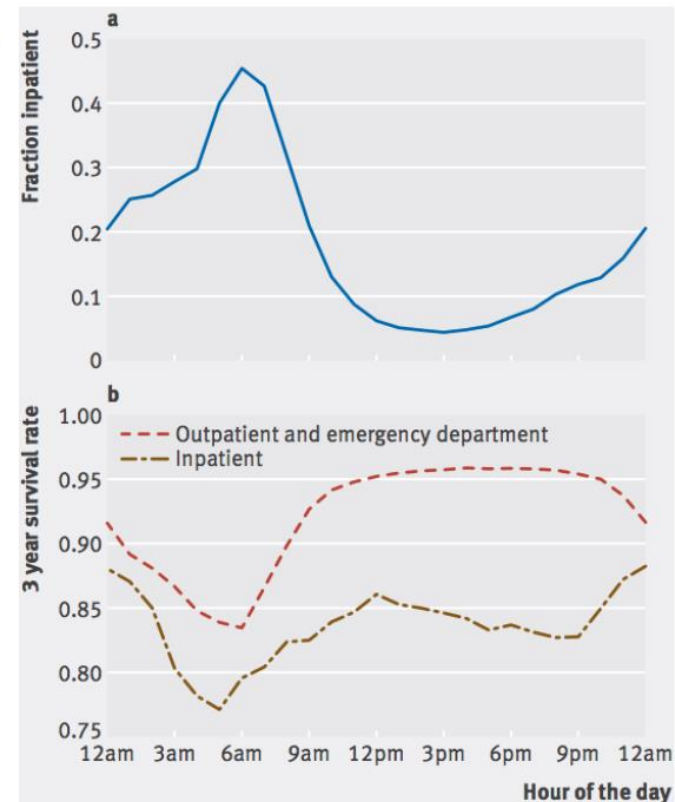


Fig 4 | White blood cell count by hour of the day. Note that (b) was smoothed using a three point running average

ML models can “cheat” (1/3)

- **Objective:** Hip fractures are a leading cause of death and disability among older adults
 - Most commonly missed diagnosis on pelvic radiographs
 - Delayed diagnosis leads to higher cost & worse outcomes

Deep learning predicts hip fracture using confounding patient and healthcare variables

[Marcus A. Badgeley](#), [John R. Zech](#), [Luke Oakden-Rayner](#), [Benjamin S. Glicksberg](#), [Manway Liu](#), [William Gale](#), [Michael V. McConnell](#), [Bethany Percha](#), [Thomas M. Snyder](#) & [Joel T. Dudley](#) ✉

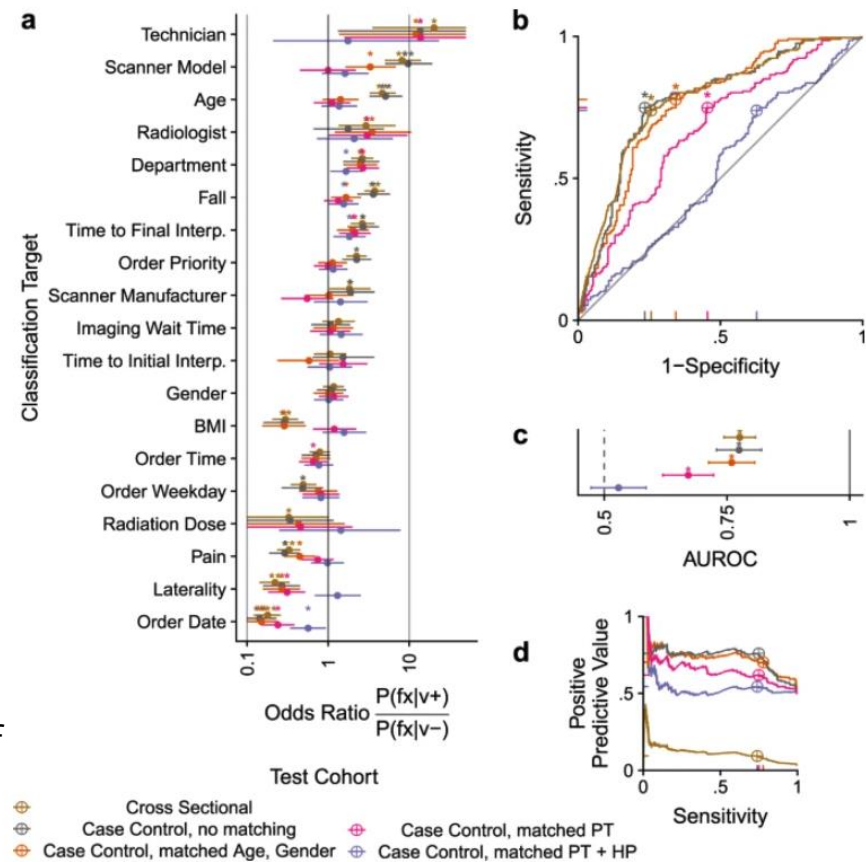
- **Data:** Collect 23,602 hip radiographs from 9,024 patients, patient and hospital process EHR data:
 - Prevalence of fracture is 3% (779/23,602)
 - Patients with fractures were more likely to report a recent fall and less likely to report pain
 - Features: image (**IMG**), disease (fracture) class, 5 patient (**PT**) features, 14 hospital process (**HP**) features

ML models can “cheat” (2/3)

- **ML model:** Train a neural network on radiographs to classify fracture
- **Results:** Fracture is predicted:
 - Moderately well from the **IMG** data alone (AUC = 0.78)
 - Better when combining **IMG + PT** (AUC = 0.86)
 - Better when combining **IMG + PT + HP** (AUC = 0.91)
- **Follow-up analysis:**
 - Test ML model whether it can **directly detect fracture** versus **indirectly predict fracture by detecting confounding variables associated with fracture**
 - On a test set with fracture risk balanced across PT and HP variables, fracture detector is no better than random (AUC = 0.52)

ML models can “cheat” (3/3)

- On test set with fracture risk balanced across PT and HP features, **fracture detector is no better than random (AUC=0.52)**
- Confounding variable (e.g., time since prior lab order, or which scanner in a hospital is used to acquire a radiograph) is associated with both:
 - **Explanatory variable** (acuity of a patient’s illness, or a patient’s clinically predicted risk of fracture)
 - **Outcome** (mortality, or the likelihood of a radiograph’s pixels containing patterns suggestive of fracture)

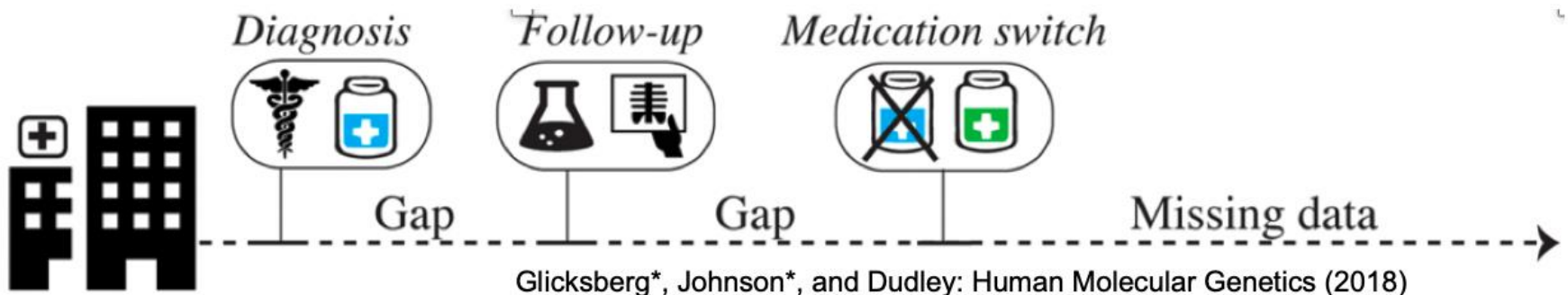


Limitations & biases of EHR

- Diseases are not easily defined in EHRs!
- External replication is not easy to facilitate
- It is challenging to capture health state from EHR
- ML algorithms can learn the wrong information
- ML algorithms can “cheat”
- ML algorithms can fail on other patient populations
- Biased real-world data can lead to real-world consequences

Fine print of using EHRs

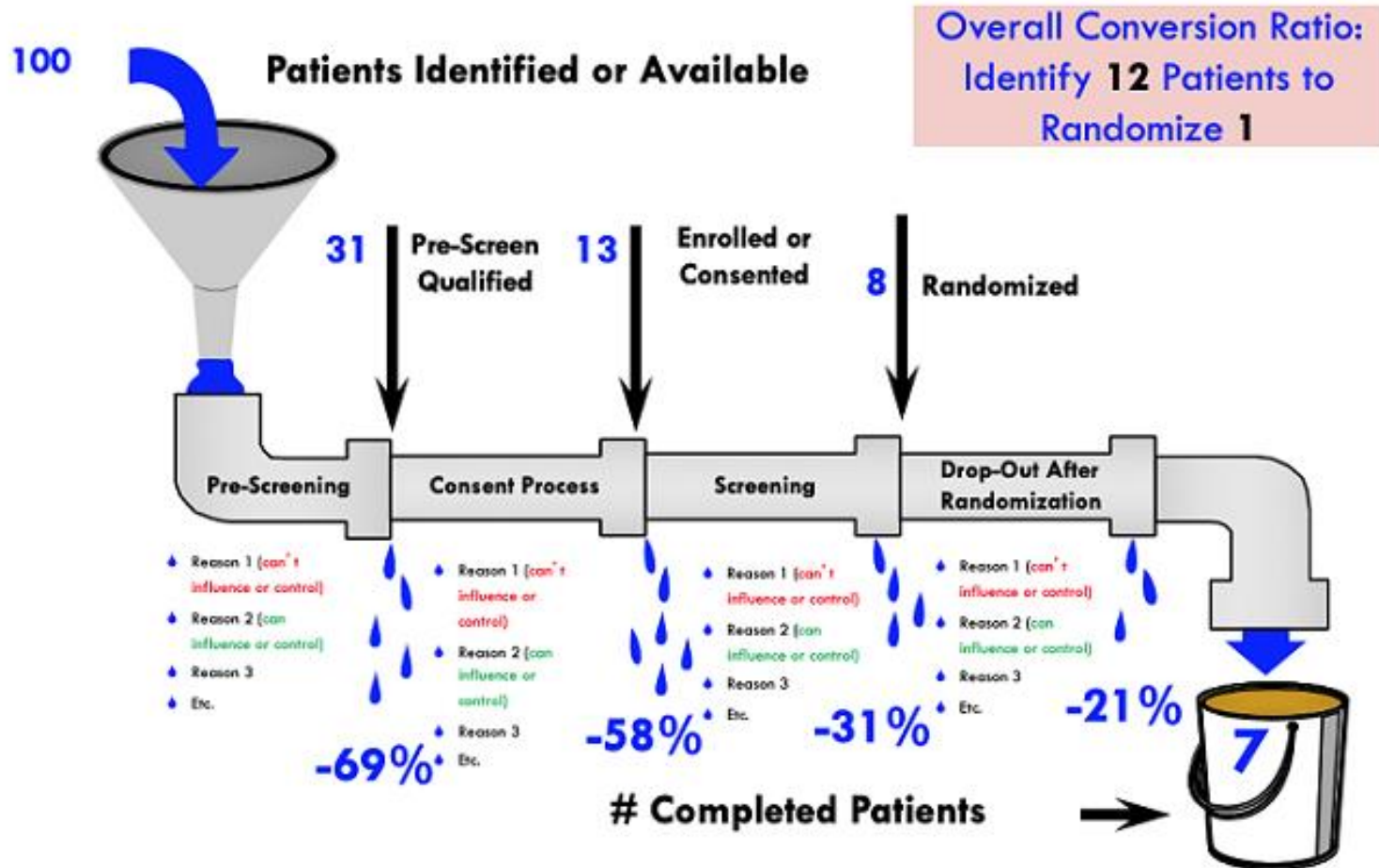
- In USA (and elsewhere), the healthcare is fragmented and EHRs do not extend beyond specific health system
- EHRs capture only data that is entered and how it is entered: “Garbage in, garbage out”
- EHR systems are messy, redundant, incomplete, heterogenous, erroneous, etc.
- Interfacing with EHR data is challenging and requires domain expertise
- Biases are propagated through!
- Poorly encoded key information: i.e., social determinants of health
- The “missing phenome”



Understanding recruitment of patients to clinical trials

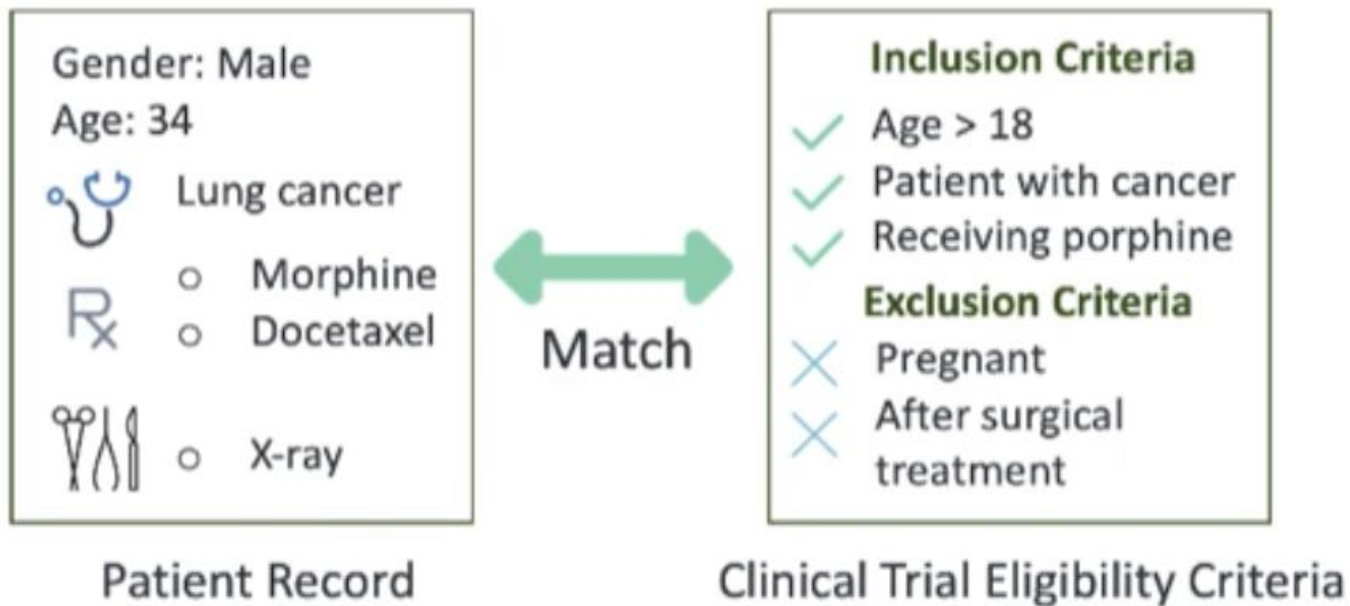
- Nearly 80% of all clinical studies fail to finish on time, and 20% of those delayed are for six months or more
- 85% of clinical trials fail to retain enough patients
- The average dropout rate across all clinical trials is around 30%
- Over two-thirds of sites fail to meet original patient enrollment for a given trial
- Up to 50% of sites enroll one or no patients in their studies

“Leaky pipe” framework for understanding patient recruitment



What is patient-trial matching?

Goal: Find qualified patients for a clinical trial given patient data and trial eligibility criteria (EC) described as both inclusion and exclusion criteria



Patient data can come from longitudinal EHRs or screening or surveys

Challenges of patient-trial matching

1. Varying concept granularity

- Eligibility criteria encode general diseases
- EHRs use specific medical codes

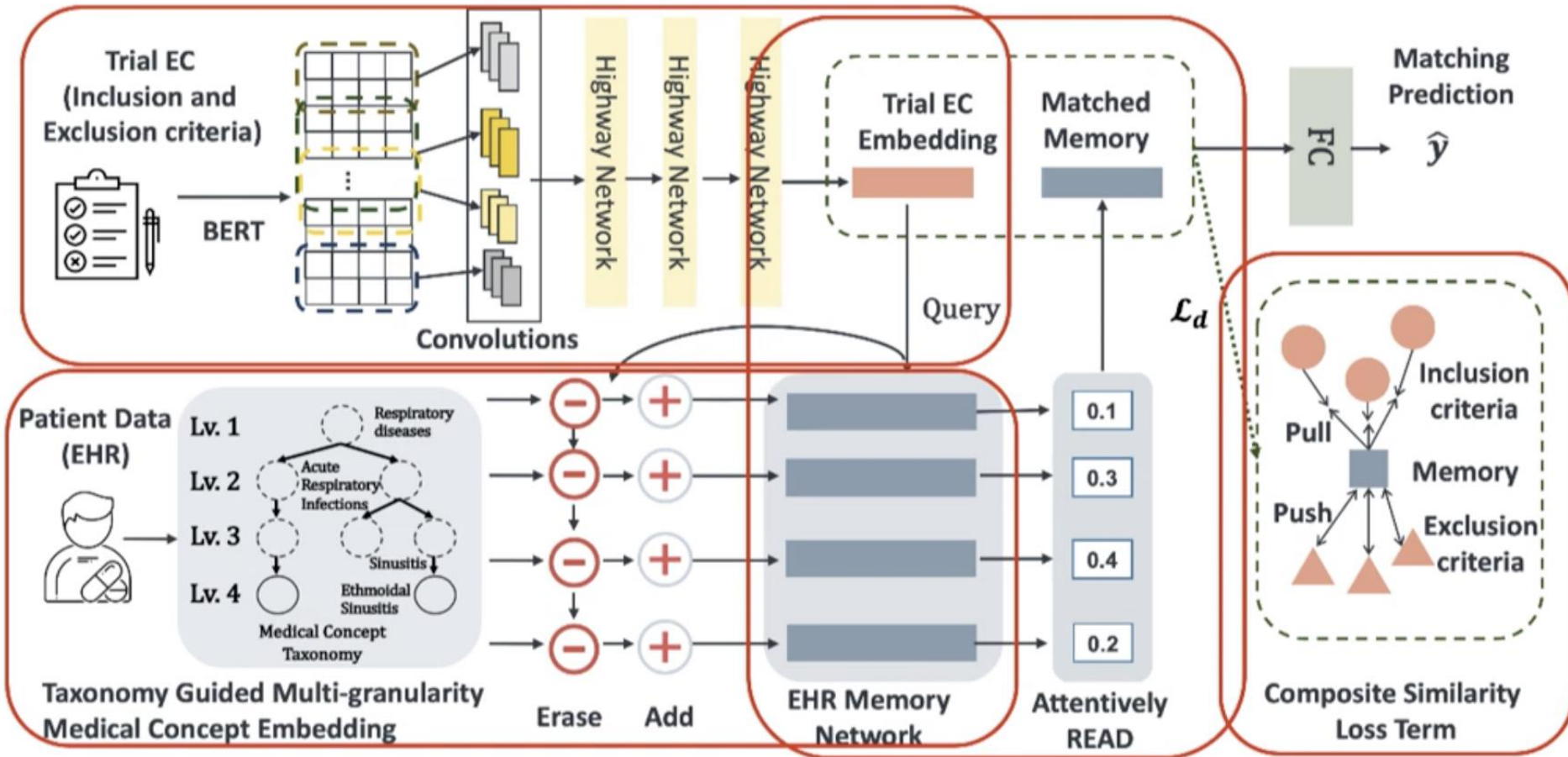
2. Many-to-many matching

- Every patient might enroll in more than one trial and vice versa
- Aligning patient embeddings to multiple trial embeddings can confuse the embedder

3. Handling explicit inclusion/exclusion criteria

- Criteria describe desired and unwanted characteristics of target patients

COMPOSE: Method overview (1/6)



COMPOSE: Method overview (2/6)

- Use BERT to learn contextual embeddings for EC sentence $[w_1, \dots, w_N]$

$$\tilde{c} = [\tilde{w}_1, \dots, \tilde{w}_N] = \text{BERT}([w_1, \dots, w_N])$$

- Use different kernel sizes to capture different granularity semantics

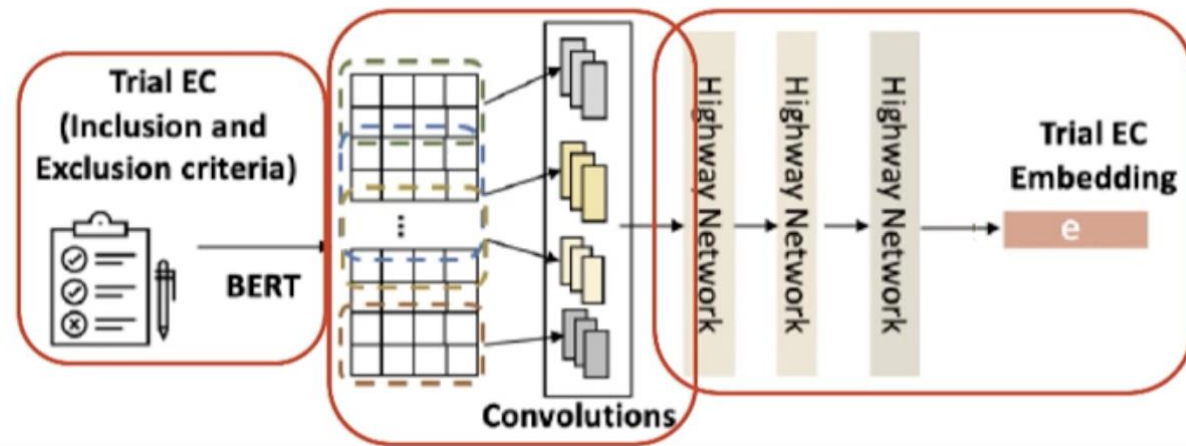
$$x = [\text{Conv}(\tilde{c}, k_1), \text{Conv}(\tilde{c}, k_2), \text{Conv}(\tilde{c}, k_3), \text{Conv}(\tilde{c}, k_4)]$$

- Use highway network and max pooling to obtain the final EC embedding

$$u = \sigma(\text{Conv}(x, k))$$

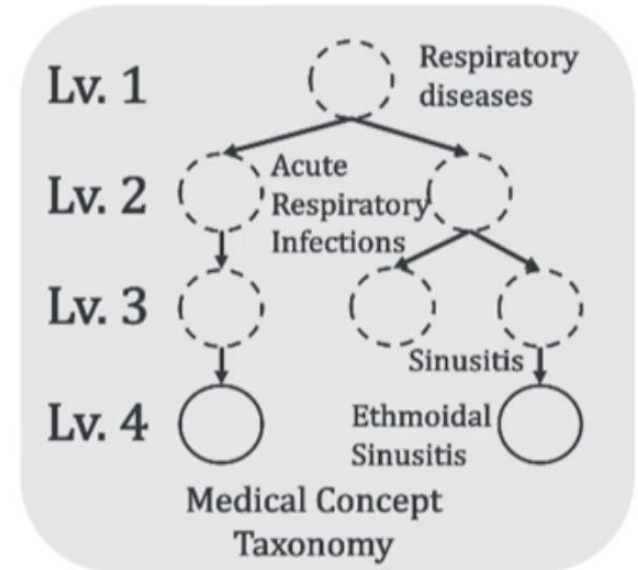
$$v = u \cdot \text{Conv}(x, k) + x \cdot (1 - u)$$

$$e = \text{MaxPool}(v)$$



Taxonomy guided patient embeddings (3/6)

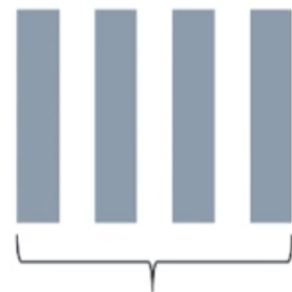
- Use medical concept taxonomy to divide each concept into four levels
 - the Uniform System of Classification (USC)
- Three memory networks to store diagnosis, medications and procedures



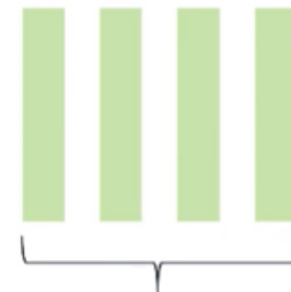
Lv. 1 Lv. 2 Lv. 3 Lv. 4

$$m = [m_D, m_O, m_P]$$

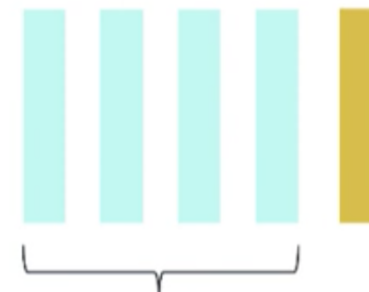
$$= [m_D^1, \dots, m_D^4, m_O^1, \dots, m_O^4, m_P^1, \dots, m_P^4]$$



Diagnosis



Medication



Procedure

Demographic

Taxonomy guided patient embeddings (4/6)

- Augment medical codes with textual description:
 - Code 692.9 -> "Contact dermatitis and other eczema"

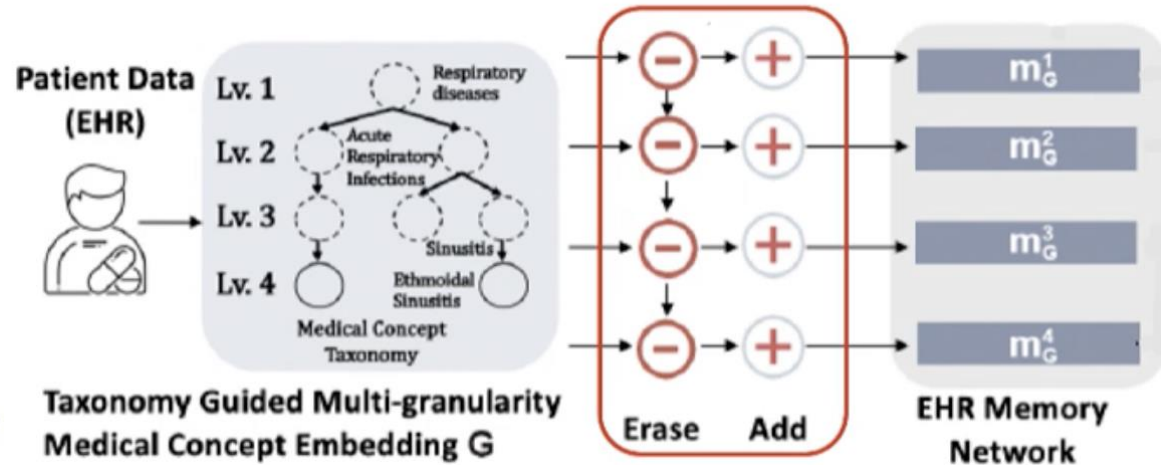
$$\tilde{g}_t = \text{MaxPool}(\text{BERT}([w_1, \dots, w_L]))$$

- Update memories at each visit
 - Erase-followed-by-add:

$$\text{erase}_t = \sigma(W_e \tilde{g}_t^k + b_e),$$

$$\text{add}_t = \text{tanh}(W_a \tilde{g}_t^k + b_a)$$

$$m_G^k \leftarrow m_G^k \odot (1 - \text{erase}_t) + \text{add}_t$$

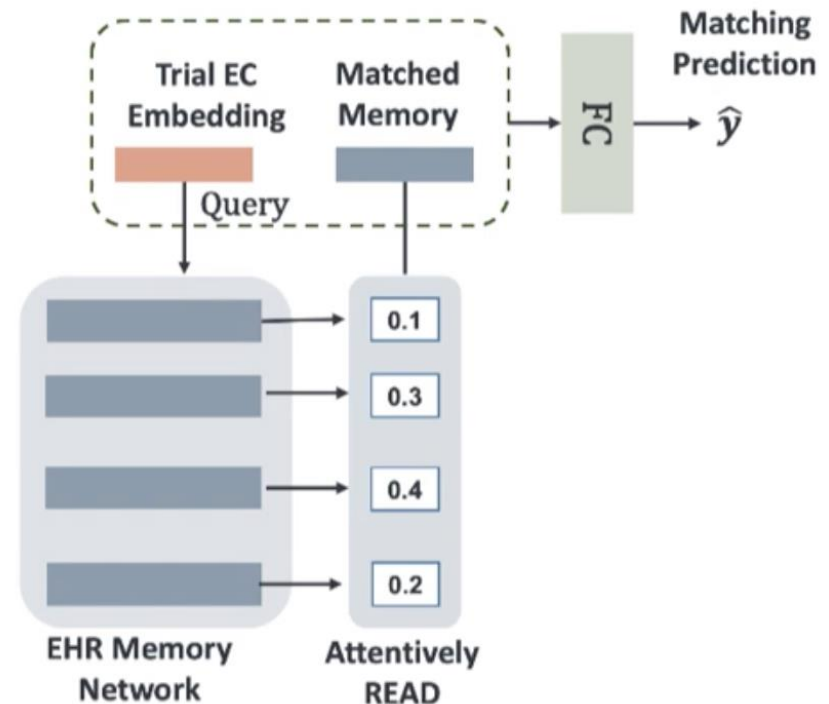


COMPOSE: Method overview (5/6)

- Let each EC correspond to the sub-memories
- Attentional matching
 - Trial EC embedding -> Query
 - Matched memory -> Response

$$a_{k,G} = \frac{\exp(\mathbf{m}_G^k \text{T} \text{MLP}(\mathbf{e}))}{\sum_{x \in \{\mathcal{D}, \mathcal{O}, \mathcal{P}\}} \sum_{i=1}^4 \exp(\mathbf{m}_x^i \text{T} \text{MLP}(\mathbf{e}))}$$

$$\tilde{\mathbf{m}} = \sum_{x \in \{\mathcal{D}, \mathcal{O}, \mathcal{P}\}} \sum_{i=1}^4 a_{i,x} \mathbf{m}_x^i$$



COMPOSE: Method overview (6/6)

- Classification loss:

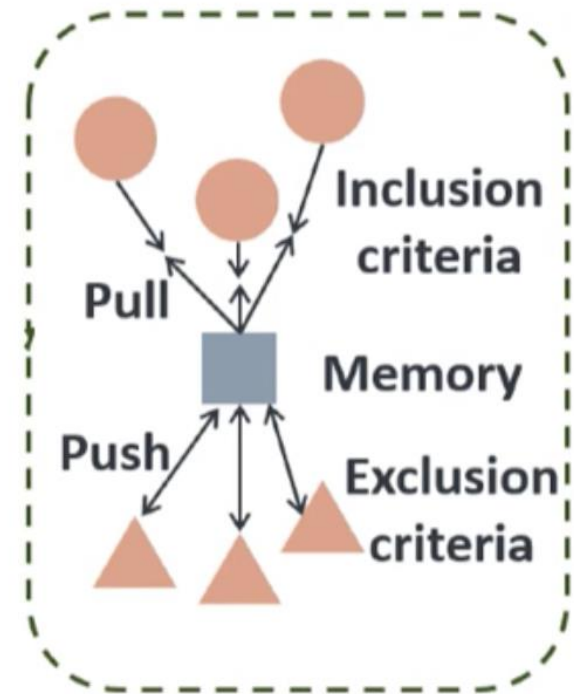
$$\mathcal{L}_c = -(\mathbf{y}^T \log(\hat{\mathbf{y}}) + (1 - \mathbf{y})^T \log(1 - \hat{\mathbf{y}}))$$

- Inclusion/Exclusion loss:

$$\mathcal{L}_d = \begin{cases} \underline{1 - d(e, \tilde{m}_I)}, & \text{if } e \text{ is } e_I \\ \underline{\max(0, d(e, \tilde{m}_E) - \alpha)}, & \text{if } e \text{ is } e_E \end{cases} \geq \alpha$$

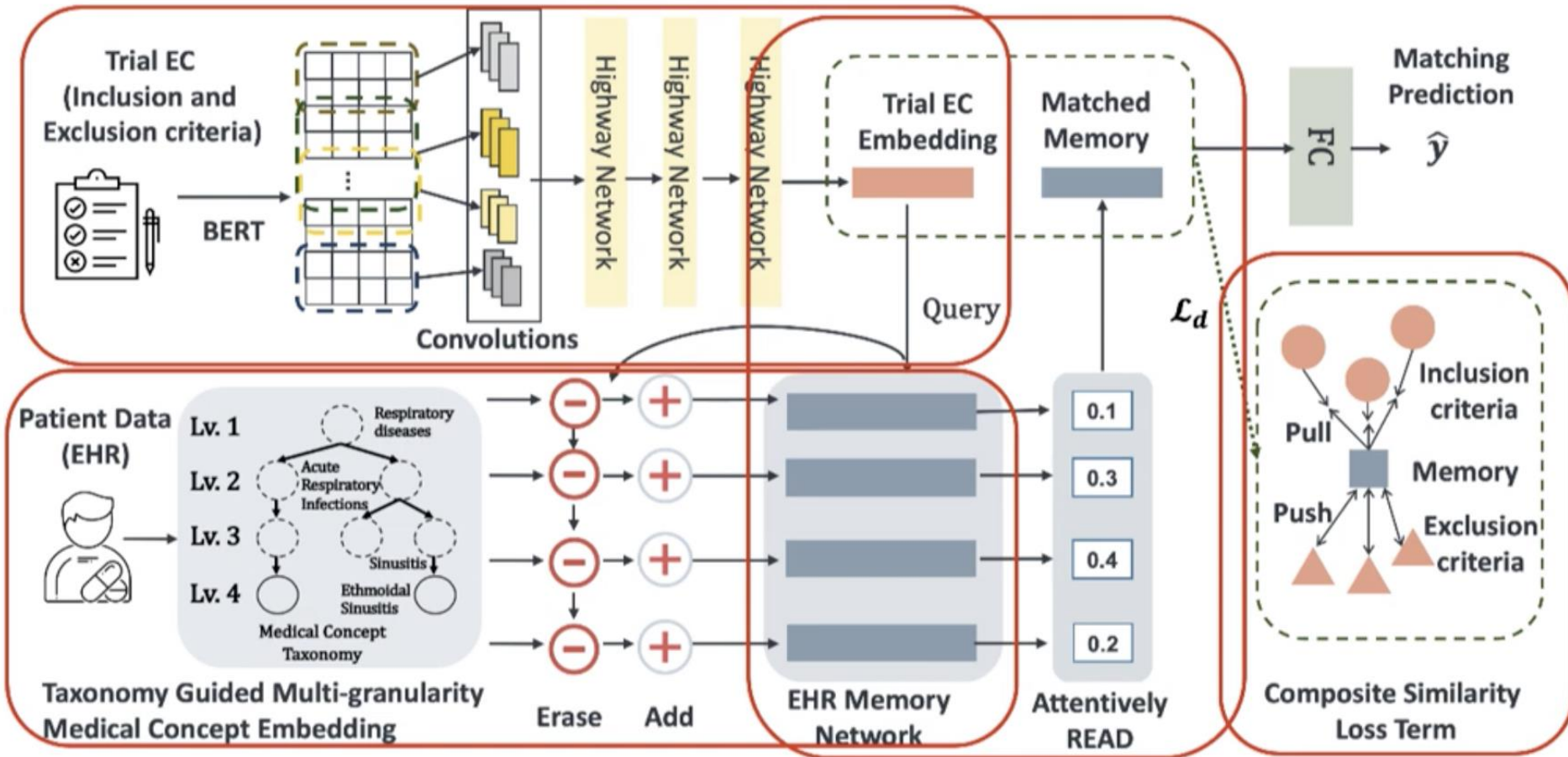
- Final loss:

$$\mathcal{L} = \mathcal{L}_c + \mathcal{L}_d$$



**Composite Similarity
Loss Term**

COMPOSE: Patient-trial matching



Experimental setup: Data

- **Clinical trials:**
 - 590 trials from publicly available data source (clinicaltrials.gov)
 - 12,445 criteria-level EC statements
- **Patient EHR dataset:**
 - 83,731 patients from 2012 to 2018

Results: Criteria-level matching

	Model	Accuracy	AUROC	AUPRC
Baselines	LSTM+GloVe	0.722±0.010	0.789±0.009	0.784±0.009
	LSTM+BERT	0.834±0.008	0.845±0.007	0.840±0.007
	DeepEnroll	0.869±0.012	0.936±0.013	0.947±0.011
Reduced	COMPOSE-MN	0.899±0.012	0.955±0.013	0.960±0.010
	COMPOSE-Highway	0.912±0.007	0.965±0.007	0.967±0.009
	COMPOSE- \mathcal{L}_d	0.939±0.010	0.976±0.009	0.973±0.007
Proposed	COMPOSE	0.945±0.008	0.980±0.007	0.979±0.008

Model	Phase I	Phase II	Phase III
LSTM+GloVe	0.0008	0.5865	0.3743
LSTM+BERT	0.0025	0.6045	0.4862
Criteria2Query	0.3025	0.6433	0.5870
DeepEnroll	0.2034	0.7493	0.6329
COMPOSE	0.5189	0.8939	0.8005

Results: Trial-level matching

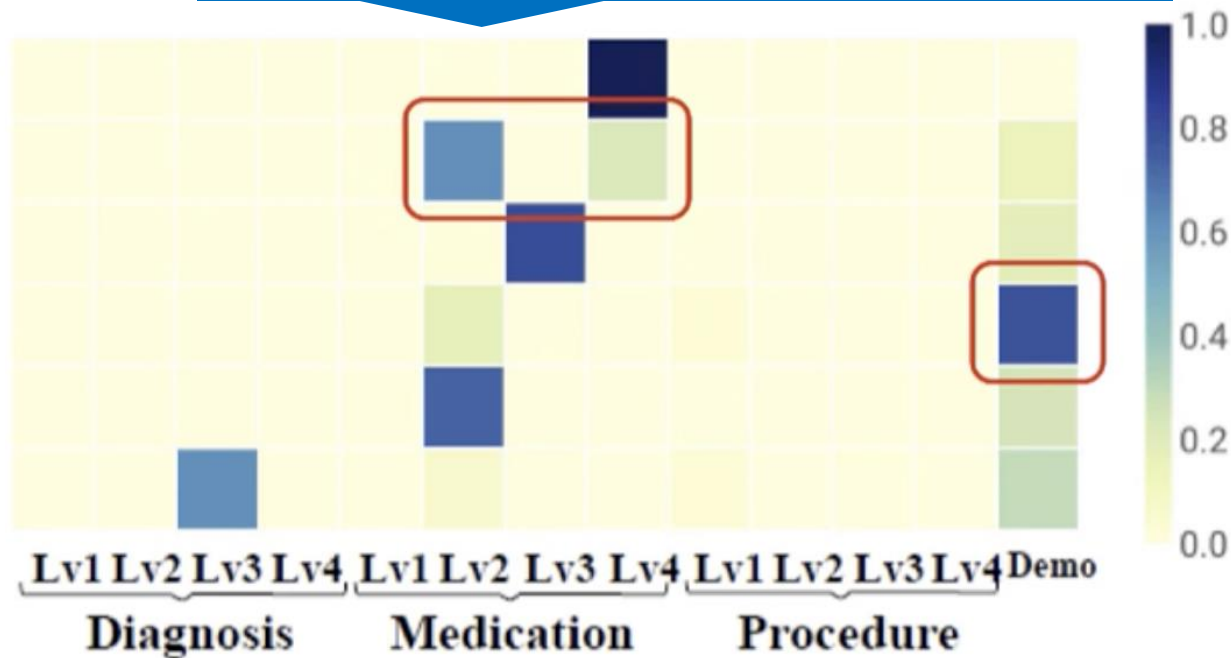
	Model	Accuracy
Baselines	LSTM+GloVe	0.4294±0.010
	LSTM+BERT	0.5460±0.008
	Criteria2Query	0.6147±-
	DeepEnroll	0.6737±0.021
Reduced	COMPOSE-MN	0.7833±0.011
	COMPOSE-Highway	0.8102±0.009
	COMPOSE- \mathcal{L}_d	0.8212±0.010
Proposed	COMPOSE	0.8373±0.012

Model	Chronic Diseases	Oncology	Rare Diseases
LSTM+GloVe	0.1793	0.0000	0.0000
LSTM+BERT	0.2062	0.0000	0.0000
Criteria2Query	0.5103	0.2722	0.2292
DeepEnroll	0.3345	0.0000	0.0000
COMPOSE	0.5931	0.6370	0.6875

Trial on Cabozantinib, which treats grade IV astrocytic tumors

Attention weights on the memory slots for the Cabozantinib trial for treating grade IV astrocytic tumors

1. received temozolomide therapy
2. receiving warfarin (or other coumarin derivatives)
3. acute intracranial/
intratumoral hemorrhage.
4. pregnant or breast-feeding
5. serious intercurrent illness
6. inherited bleeding diathesis or
coagulopathy

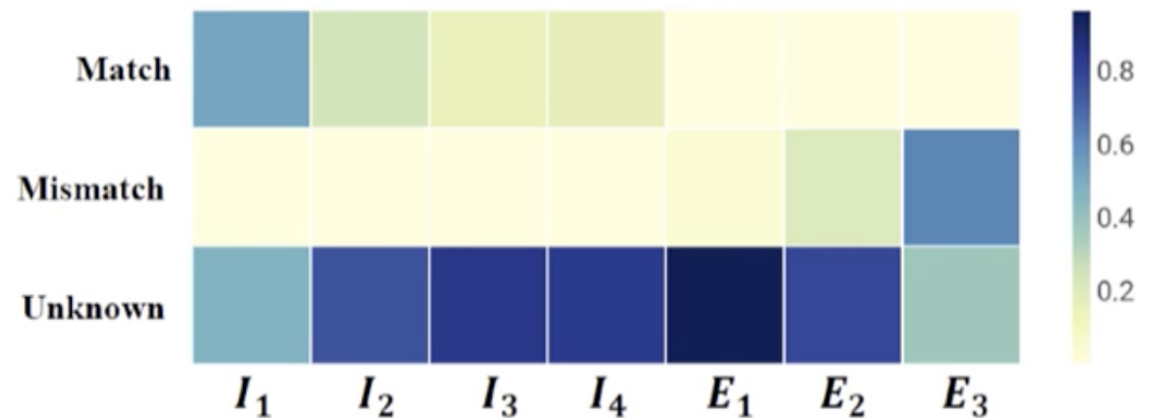


COMPOSE successfully matches this trial (94% matching) while all baselines fail (< 50% matching)

Trial for early-stage non-small cell lung cancer

#NCT02998528

- I_1 : Early stage IB-IIIa, operable non-small cell lung cancer, confirmed in tissue
- I_2 : Lung function capacity capable of tolerating the proposed lung surgery
- I_3 : Eastern Cooperative Oncology Group (ECOG) Performance Status of 0-1
- I_4 : Available tissue of primary lung tumor
- E_1 : Presence of locally advanced, inoperable or metastatic disease
- E_2 : Participants with active, known or suspected autoimmune disease
- E_3 : Prior treatment with any drug that targets T cell co-stimulations pathways (such as checkpoint inhibitors)



Inclusion criteria are denoted as I_i and exclusion criteria as E_j

An example of a trial for which it is difficult to find matching patients. All models achieve a lower than 50% accuracy score for this trial. Shown are prediction results for COMPOSE and a case patient. The results show that COMPOSE successfully matches I_1 and E_3 to the patient but classifies other ECs to unknown

Outline for today's class

- ✓ 1. Overview of this course
- ✓ 2. What makes biomedical data unique
- ✓ 3. Introduction to distributed language representations
- ✓ 4. Introduction to NLP in clinical settings