

MIT 18.700/18.701/6.047/6.878/HST.507

Computational Biology: Genomes, Networks, Evolution

Lecture 19: EHRs

Electronic Health Records

Dealing with Population-Scale EHRs

PheWAS: GWAS for Multi-Phenotypes

Interventions and Reverse Causality

Clinical Notes: LLMs, Text-Data Joint Models

Action, Agents, Reinforcement Learning

AI Agents in Value-Based Healthcare

MLCB - Machine Learning in Computational Biology - Fall 2024 - Profs. Manolis Kellis + Eric Alm - 6.8700/6.8701/20.s900/20.s948/HST.507

Homeworks	Project / Mentoring (Fri)	Wk	Date	Lec	Topic
Introduction: Machine Learning, Deep Learning, Generative AI, and the Unification of Biology					
HW0 out Thu 9/5		1	Thu-Sep-05	L1	Course Overview, Machine Learning, Deep Learning, Inference, Genome, Proteins, Chemistry, Imaging
Module 1: Genomics, Epigenomics, Single-Cell, Networks, Circuitry					
HW0 due Wed 9/11		2	Tue-Sep-10	L2	Expression Analysis, Clustering/Classification, Gaussian Mixture Models, K-means, Bayesian Inf, Gen-vs-DiscrM
HW1 out Thu 9/12	0=Self Introductions	2	Thu-Sep-12	L3	Single-cell genomics, sc-mutli-omics, non-linear embeddings, spatial transcriptomics, next-gen technologies
		3	Tue-Sep-17	L4	Sequential Data, Alignment, DynProg, Hidden Markov Models, Parsing, Posterior Decoding, HMM architectures
		3	Thu-Sep-19	L5	Epigenomics: Signal Modeling, Peak calling, Chromatin states, 3D structure, Hi-C, Genome Topology
		4	Tue-Sep-24	L6	Regulatory Genomics: Motifs, Information, ChIP, Gibbs Sampling, EM, CNNs for Genome Parsing
HW1 due Mon 9/30	1=Select previous paper(s)	4	Thu-Sep-26	L7	Regulatory Networks: Graphs, Linear Algebra, PCA, SVD, Dimentionality Reduction, TF-enhancer-gene circuitry
Module 2: Protein Structure, Protein Language Models, Geometric Deep Learning					
HW2 out Thu 10/3		5	Tue-Oct-01	L8	Intro to structural biology
		5	Thu-Oct-03	L9	Protein structure and folding: Diffusion models, Cryo-EM, Protein design
		6	Tue-Oct-08	L10	Intro to transformers and Large Language Models LLMs
	2=Proposal+Feasibility	6	1	L11	Protein Language Models PLMs and Transfer Learning
		7	Tue-Oct-15	-	-- No Class -- Student holiday
		7	Thu-Oct-17	L12	DNA language models: Chromatin Structure
Module 3: Chemistry, Therapeutics, Graph Neural Networks					
HW3 out Thu 10/24		8	Tue-Oct-22	L13	Overview of drug development
	3=OffHrs Update Feedback	8	Thu-Oct-24	L14	Intro to small molecules
		9	Tue-Oct-29	L15	Representation of small molecules: Graphs, GNNs, Transformers, RDKit
		9	Thu-Oct-31	L16	Docking: Small molecule - proteins docking
		10	Tue-Nov-05	L17	Disease Association Mapping, genetics, GWAS, linkage analysis, disease circuitry, variant-to-function
	4=OffHrs Update Feedback	10	Thu-Nov-07	L18	Quantitative trait mapping, molecular traits, eQTLs, mediation analysis, iMWAS, multi-modal QTLs
Module 4: Electronic Health Records, Imaging, Evolution, Metabolism					
No HW		11	Tue-Nov-12	L19	Quiz Review + Preparation
		11	Thu-Nov-14	L20	-- In-class Quiz
		12	Tue-Nov-19	L21	Electronic Health Records, AllOfUs, UKBioBank, Medical Genomics, Pop-Scale Cohorts, Multi-Ancestry [not quiz]
	5=Midcourse report	12	Thu-Nov-21	L22	Comparative genomics, Conservation, Evolutionary signatures, PhyloCSF, RNA structure, Motif BLS2conf
		13	Tue-Nov-26	L23	Evolution, Phylogenetics, Phylogenomics, Duplication, RNA world, RNA folding, lincRNAs, RNA modifications, m
		13	Thu-Nov-28	-	-- No Class -- Thanksgiving Holiday
		14	Tue-Dec-03	L24	Modeling metabolism: Flux balance analysis
	6=WriteUp, Slides Due	14	Thu-Dec-05	L25	Measuring metabolism: Metabolomics and Deep Learning
Final Projects					
	7=In Class Presentations	15	Tue-Dec-10	L26	Project Presentations (6-8 mins/team). Report due Fri@11.59p, Slides due Mon@11.59p, Present Live Tue

Lecture 19: EHRs

Electronic Health Records

Dealing with Population-Scale EHRs

PheWAS: GWAS for Multi-Phenotypes

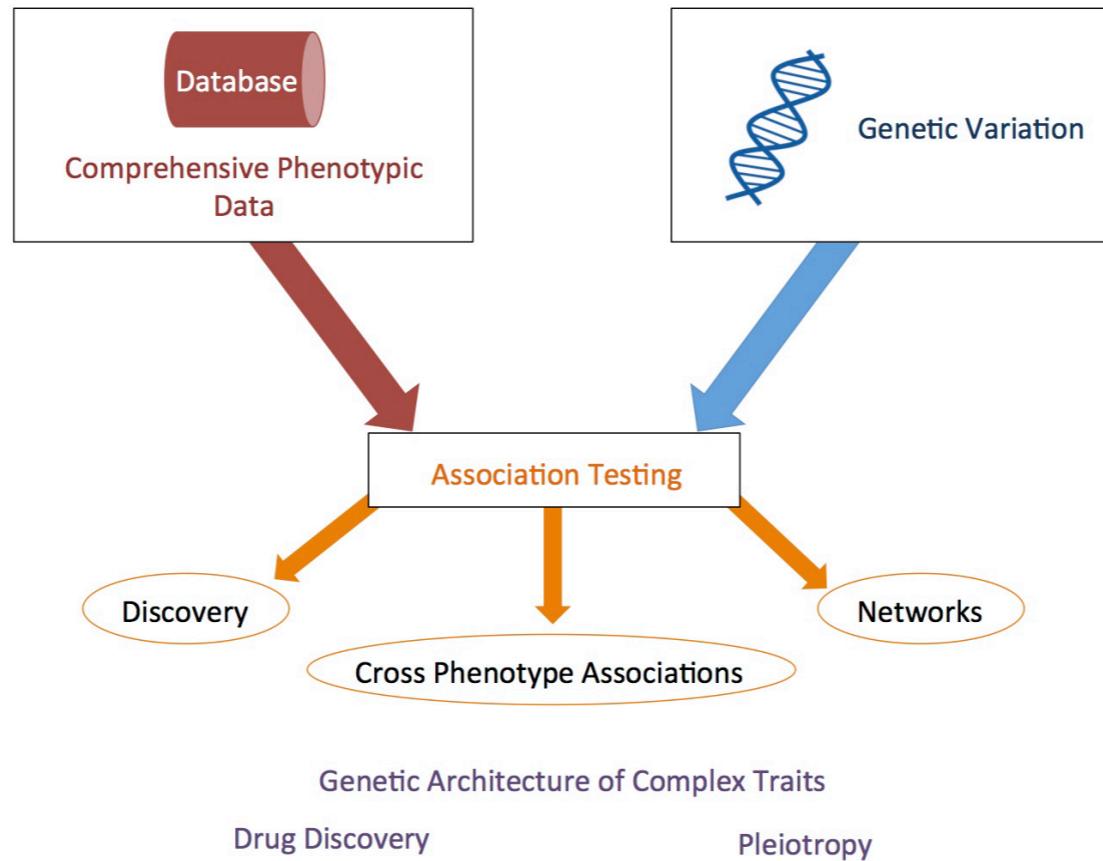
Interventions and Reverse Causality

Clinical Notes: LLMs, Text-Data Joint Models

Action, Agents, Reinforcement Learning

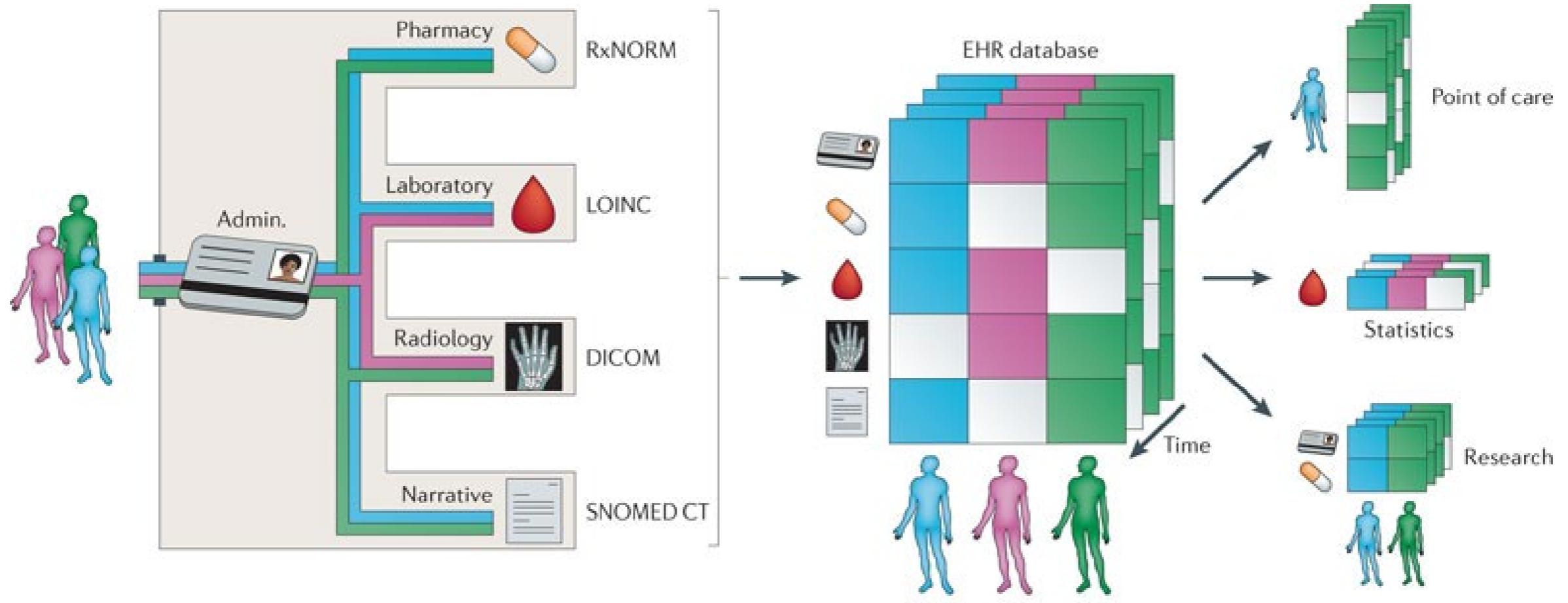
AI Agents in Value-Based Healthcare

Why Multi-phenotype analyses?



- Associate unknown/underappreciated phenotypes with known
- Identify disease mechanisms by mediating phenotypes
- Predict disease with easy-to-measure ‘biomarker’ phenotypes
- Improve GWAS power by imputation of missing phenotypes
- Improve GWAS biological relevance by meta-phenotype association
- Understand pleiotropic effects affecting multiple phenotypes
- Guide experiments/diagnosis by predicting missing phenotypes
- Enable personalized medicine by treating combinations of symptoms

Electronic health records (EHRs)



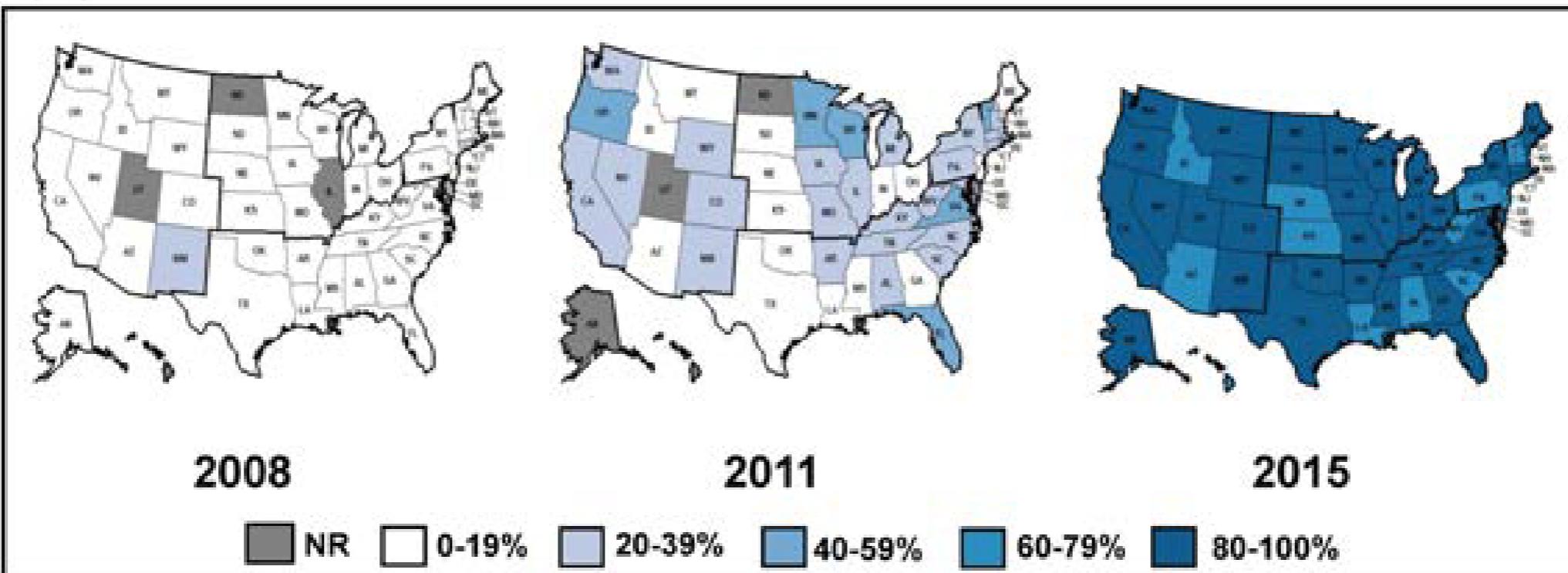
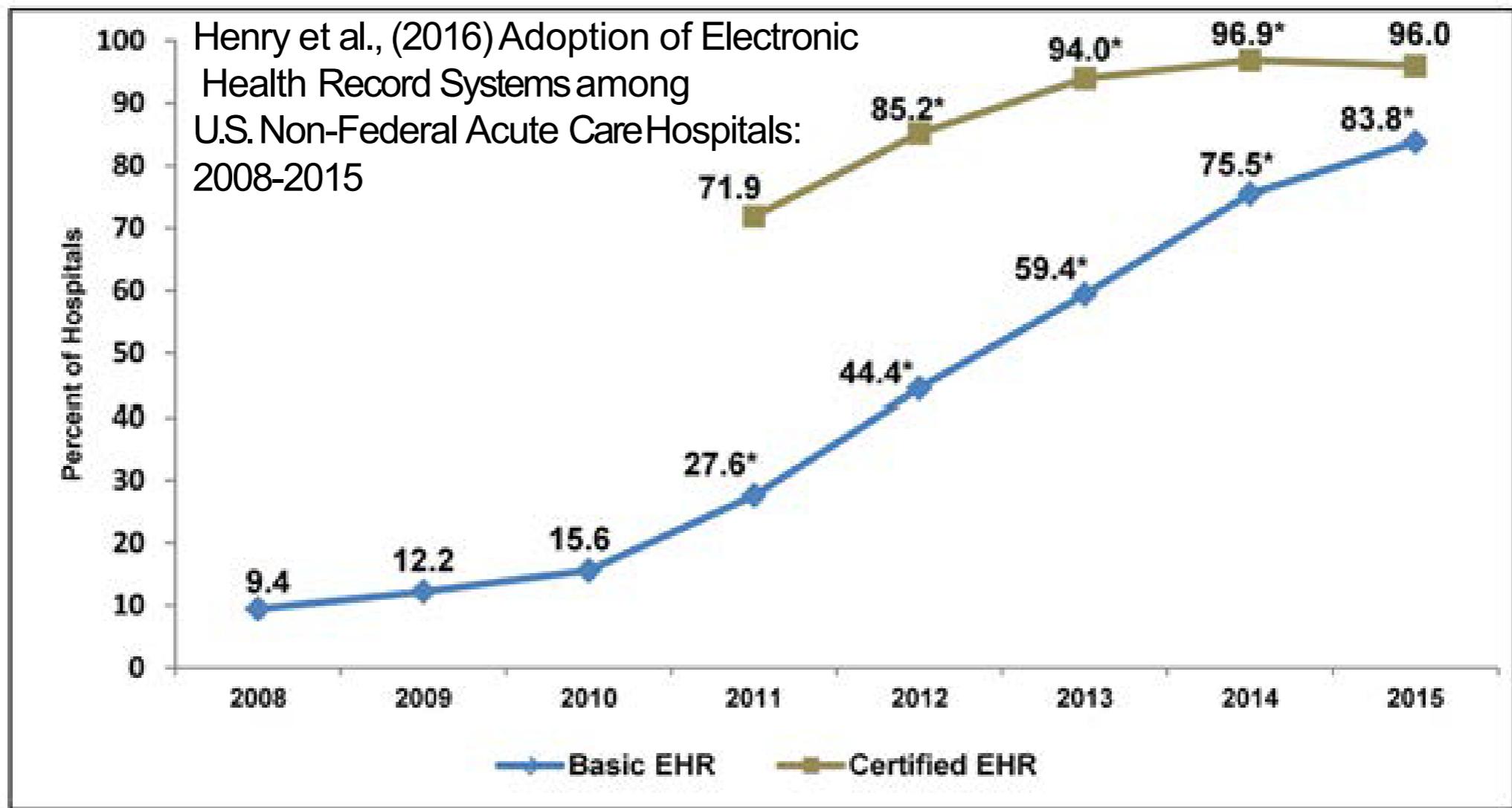
Nature Reviews | Genetics

EHR contains **extremely rich** information of a patient:

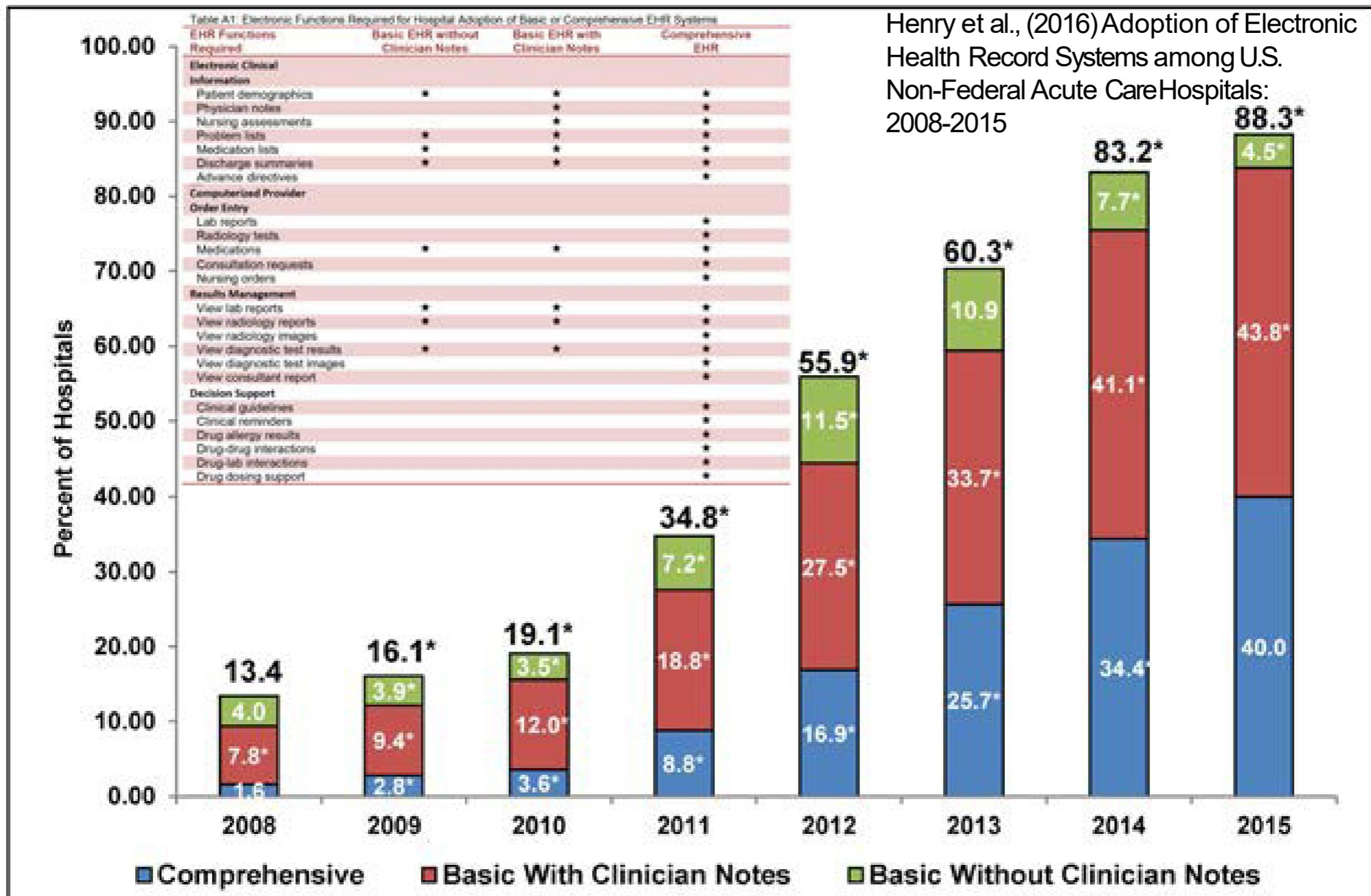
- Clinical notes (unstructured free-form text);
- Lab tests: Logical Obs. Identifiers Names & Codes (LOINC)
- Billing code: International Classification of Disease (ICD)
- Billing code: ICD Current Procedural Terminology (CPT)
- Billing code: Diagnosis-related group (DRG)
- Pharmaceutical: Prescription data (RxNorm)

Jensen et al., (NatRev, 2012)

Rapid adoption of EHR technology in the US 2008-2015

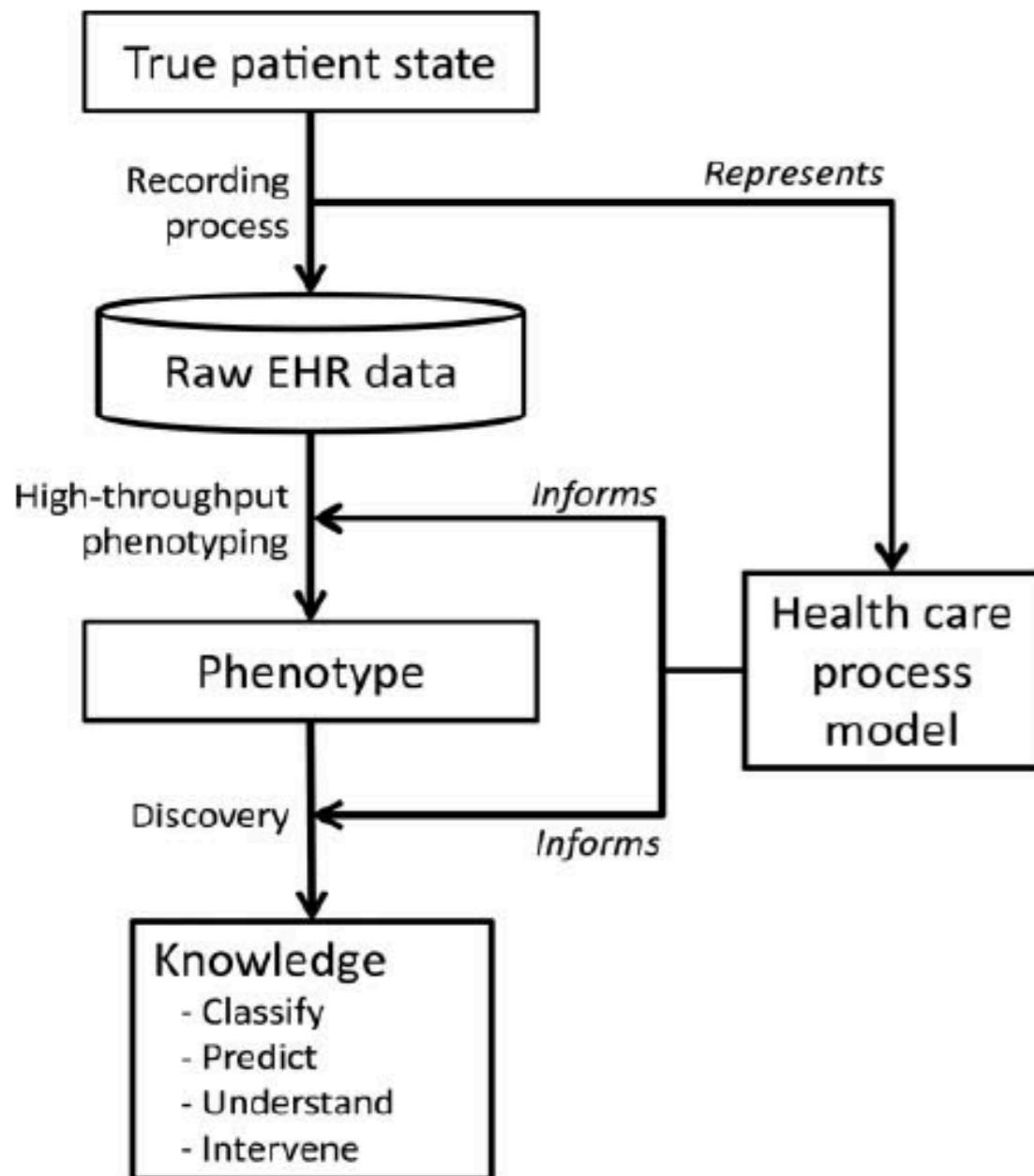


EHR become increasingly comprehensive



As EHR data become more comprehensive, there is an urgent need for integrative machine learning approach to model them.

Ascertainment bias: Disease-dependent phenotypes



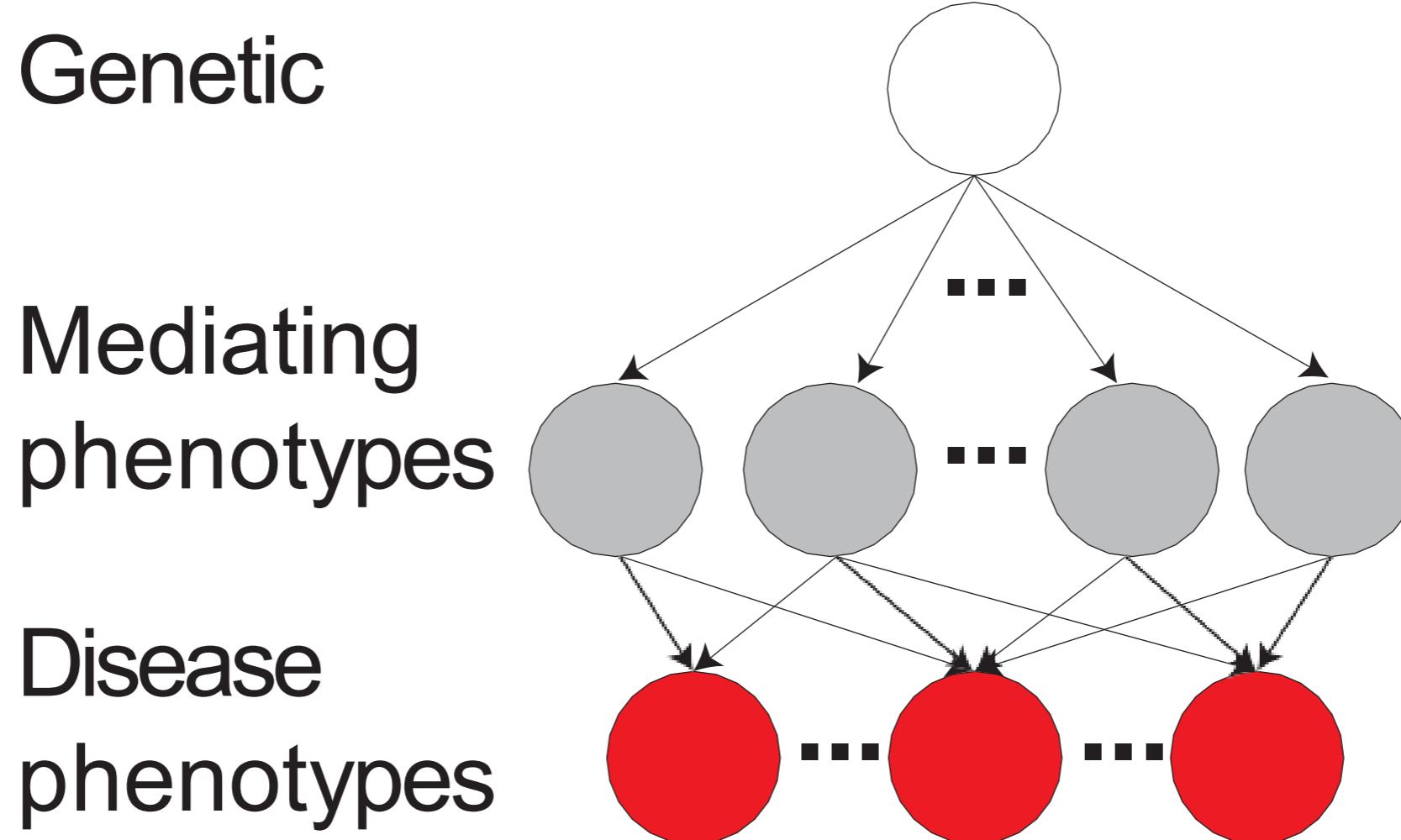
- Raw EHR data are indirect reflection of true patient state (due to recording process)
- State of patient varies
- Patient state influences:
 - Value of the measurements
 - Whether there is a measurement
 - Type of measurement
 - Timing of measurement

PheWAS: Multi-phenotype studies

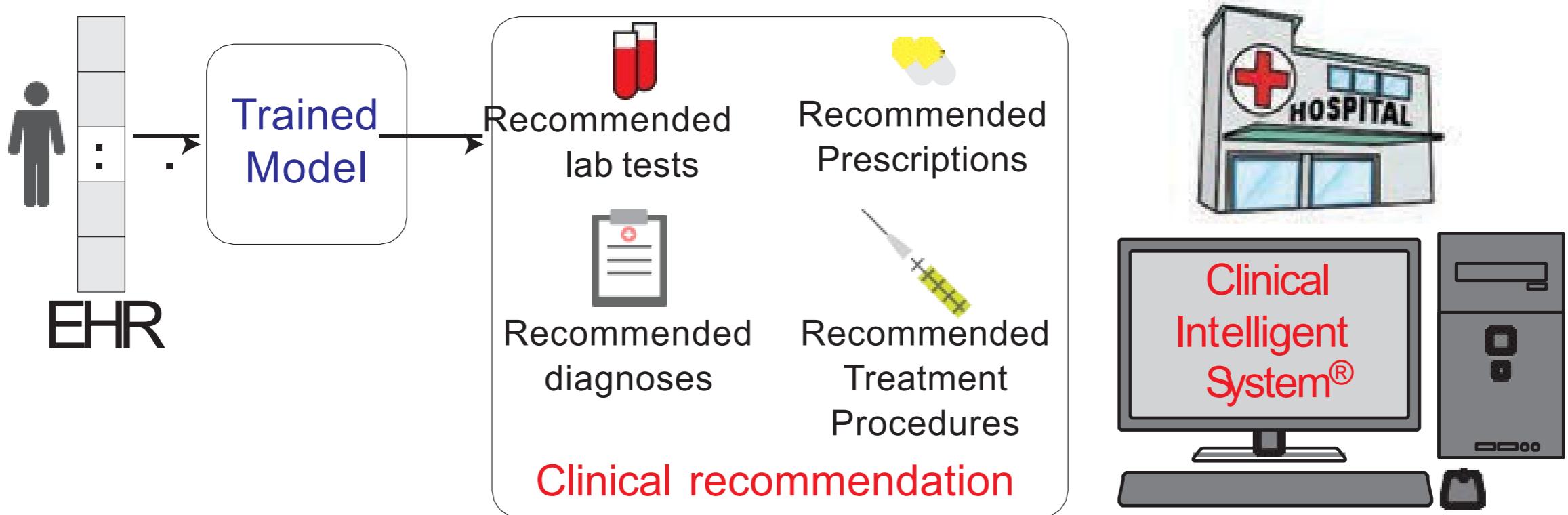
- 1. Motivation of genome-wide association studies**
 - PheWAS-informed phenotyping, improved GWAS power, etc
 - Electronic health record (EHR) contain rich personalized information
- 2. Modeling multiple phenotypes in GWAS + epigenomics**
 - Integration of multiple phenotypes in GWAS from Systems Genetics perspective (clustering approach)
 - Direct integration of multiple phenotypes by summary-based factored genetic model estimation
- 3. Epigenomics of PheWAS**
 - Risk variants inference using epigenomic reference annotations
 - Using disease covariance to improve functional variants inference
 - Combining enrichment to improve causal pathway inference
- 4. Meta-phenotype inference and imputation**
 - Models leveraging missing information and inferring missing mechanism
 - Modeling multimodal electronic health record data
 - Imputing missing EHR code and prioritizing patient disease risks

PheWAS without genetics information using EHR

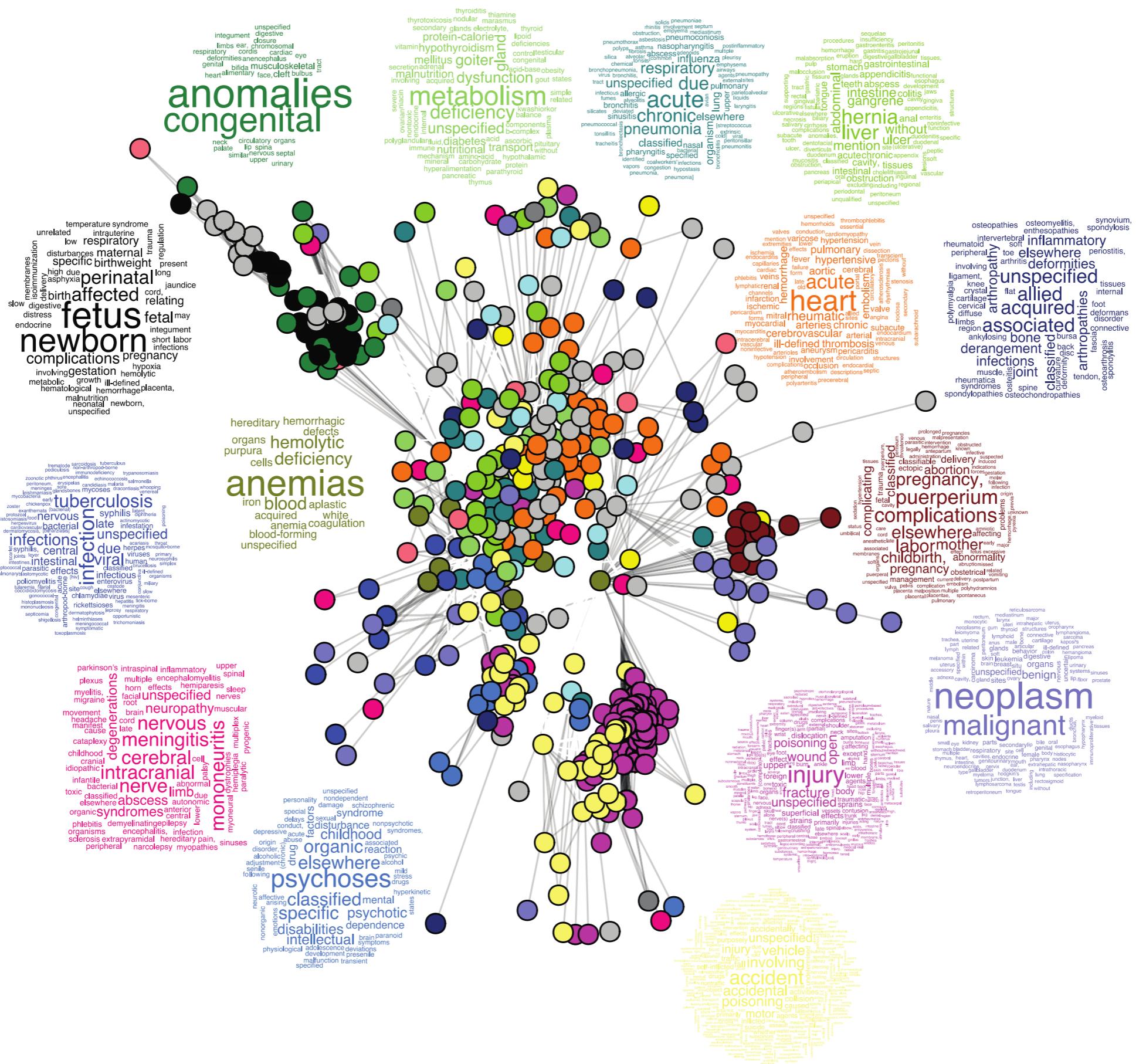
- Genotype are often **not available** over large patient cohort
- Given the causal mediating phenotypes, **diseases of interest are conditionally independent of genotype**
- Deep phenotyping by electronic health records (EHR) technology as surrogates to mediating phenotypes is one promising direction



Medical recommendation system



Learning latent disease topics from diverse EHR data



Intuition behind phenotyping by matrix factorization

	frequent urination	type 2 diabetes	high blood sugar	pregnant	...
...	...	✓	✓ ✓
...	...	✓	✓	✓	...
⋮	⋮ ⋮ ⋮	⋮	⋮ ⋮ ⋮	⋮ ⋮ ⋮	⋮ ⋮ ⋮
?	...	?	✓ ✓	⋮ ⋮	⋮ ⋮
⋮	⋮ ⋮ ⋮	⋮	⋮ ⋮ ⋮	⋮ ⋮ ⋮	⋮ ⋮ ⋮

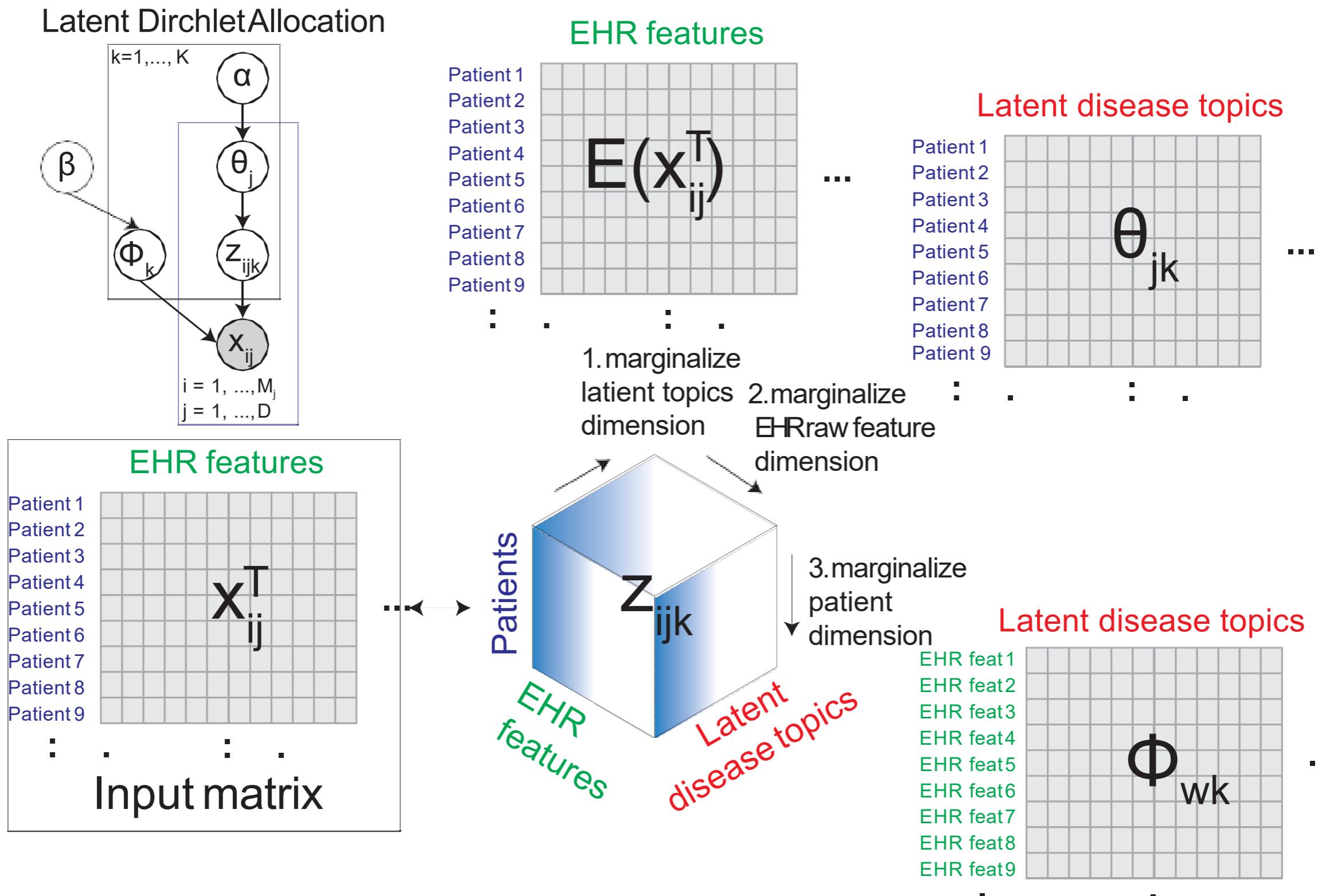
Patient similarity

Cluster 1	...	Cluster j	...	Cluster K
⋮	⋮ ⋮ ⋮	⋮	⋮ ⋮ ⋮	⋮ ⋮ ⋮
...	...	✓
⋮	⋮ ⋮ ⋮	⋮	⋮ ⋮ ⋮	⋮ ⋮ ⋮
?	...	✓
⋮	⋮ ⋮ ⋮	⋮	⋮ ⋮ ⋮	⋮ ⋮ ⋮

Phenotype similarity

	frequent urination	type 2 diabetes	high blood sugar	...
Cluster 1	...	✓
⋮	⋮ ⋮ ⋮	⋮	⋮ ⋮ ⋮	⋮ ⋮ ⋮
Cluster j	✓	✓
⋮	⋮ ⋮ ⋮	⋮	⋮ ⋮ ⋮	⋮ ⋮ ⋮
Cluster K

Latent Dirichlet Allocation (LDA) (Blei et al, JMLR 2003)



Grouping words by their topics

“Arts”	“Budgets”	“Children”	“Education”
NEW	MILLION	CHILDREN	SCHOOL
FILM	TAX	WOMEN	STUDENTS
SHOW	PROGRAM	PEOPLE	SCHOOLS
MUSIC	BUDGET	CHILD	EDUCATION
MOVIE	BILLION	YEARS	TEACHERS
PLAY	FEDERAL	FAMILIES	HIGH
MUSICAL	YEAR	WORK	PUBLIC
BEST	SPENDING	PARENTS	TEACHER
ACTOR	NEW	SAYS	BENNETT
FIRST	STATE	FAMILY	MANIGAT
YORK	PLAN	WELFARE	NAMPHY
OPERA	MONEY	MEN	STATE
THEATER	PROGRAMS	PERCENT	PRESIDENT
ACTRESS	GOVERNMENT	CARE	ELEMENTARY
LOVE	CONGRESS	LIFE	HAITI

The William Randolph Hearst Foundation will give \$1.25 million to Lincoln Center, Metropolitan Opera Co., New York Philharmonic and Juilliard School. “Our board felt that we had a real opportunity to make a mark on the future of the performing arts with these grants an act every bit as important as our traditional areas of support in health, medical research, education and the social services,” Hearst Foundation President Randolph A. Hearst said Monday in announcing the grants. Lincoln Center’s share will be \$200,000 for its new building, which will house young artists and provide new public facilities. The Metropolitan Opera Co. and New York Philharmonic will receive \$400,000 each. The Juilliard School, where music and the performing arts are taught, will get \$250,000. The Hearst Foundation, a leading supporter of the Lincoln Center Consolidated Corporate Fund, will make its usual annual \$100,000 donation, too.

Challenge 1: EHR data are noisy and biased

- Unlike corpus documents in text mining, clinical notes are full of typos and arbitrary abbreviations
 - Text mining commonly remove ‘stop words’, there no clearly defined ‘stop words’ in EHR
 - Billing codes are often not meant to be disease-specific

Diagnosis-related group (DRG)

Neonate, Bwt > 2499g
Head Trauma w/ Coma
Percutaneous CardioVMAJOR CARDIOVASCULAR
EAR, NOSE, MOUTH & T

OTHER MUSCULOSKELETA FULL TERM NEONATEWI
PREMATURITY WITH MAJ Coronary Bypass w/o TRAUMATIC INJURY AGE
Infectious & Parasit Coronary Bypass w/ C Respiratory System D
AMPUTATION FOR MUSCU CORONARY BYPASS WITH
CRANIOTOMY AGE >17 E SPECIFIC CEREBROVASC
SEIZURE & HEADACHE A Cardiac Valve Proced Neonate, Bwt > 2499g
SEPTICEMIA AGE >17 EXTREME IMMATURITY O
CARDIAC VALVE & OTH Cardiac Valve Proced
COMPLICATIONS OF TRACHEOSTOMY NEONATE WITH OTHER S
LOWER EXTRCTIONS OF TRACHEA Septicemia & Dissemi
Pulmonary Edema & Re Septicemia & Dissemi
Craniotomy Except fo Septicemia & Dissemi

NORMAL NEWBORN

TRAUMATIC STUPOR & C
Cardiac Valve Proced
Neonate, Bwt > 2499g
NEONATES, DIED OR TR
CARDIAC VALVE & OTH
FRACTURES OF HIP & P
Septicemia & Dissemi
OTHER OPERATING ROOM
CORONARY BYPASS WITH
SEPTICEMIA OR SEVERE
Coronary Bypass w/ C
Tracheostomy W Long
Head Trauma w/ Coma
Respiratory System D
Tracheostomy W Long Musculoskeletal & Ot
GASTROINTESTINAL HEM
RESPIRATORY SYSTEMD
Coronary Bypass w/o

Lab tests

MCV-BLOOD-HEMATOLOGY
BILIRUBIN, TOTAL-BLO POTASSIUM, WHOLE BLO
RED BLOOD CELLS-BLO
PT-BLOOD-HEMATOLOGY
HEMOGLOBIN-BLOOD-HEM
FREE CALCIUM-BLOOD-B GLUCOSE-BLOOD-BLOOD
GLUCOSE-BLOOD-CHEMIS
ANION GAP-BLOOD-CHEM
UREA NITROGEN-BLOOD-
SODIUM-BLOOD-CHEMIST
POTASSIUM-BLOOD-CHEM
BICARBONATE-BLOOD-CH
CREATININE-BLOOD-CHE
NEUTROPHILS-BLOOD-HE ASPARATE-AMINOTRANSF-DEEP-BLOOD-BLOOD GAS
PLATELET COUNT-BLOOD
INTUBATED-BLOOD-BLO MONOCYTES-BLOOD-HEMA
WHITE BLOOD CELLS-BL
OXYGEN SATURATION-BL EOSINOPHILS-BLOOD-HE
MCHC-BLOOD-HEMATOLOG
INR(PT)-BLOOD-HEMATO Lymphocytes-BLOOD-HE
MCH-BLOOD-HEMATOLOGY
BASE EXCESS-BLOOD-BL

ICD-9 CM

Clinical notes

respons action
monitor note reason
present histori
line meql
medicis slip
hospit blood
day chang
given
follow time
continu fluid
name left plan
patient acut
chest po mgd will
hour status normal
respiratori long
report icu
examin

ICD-9 CPT

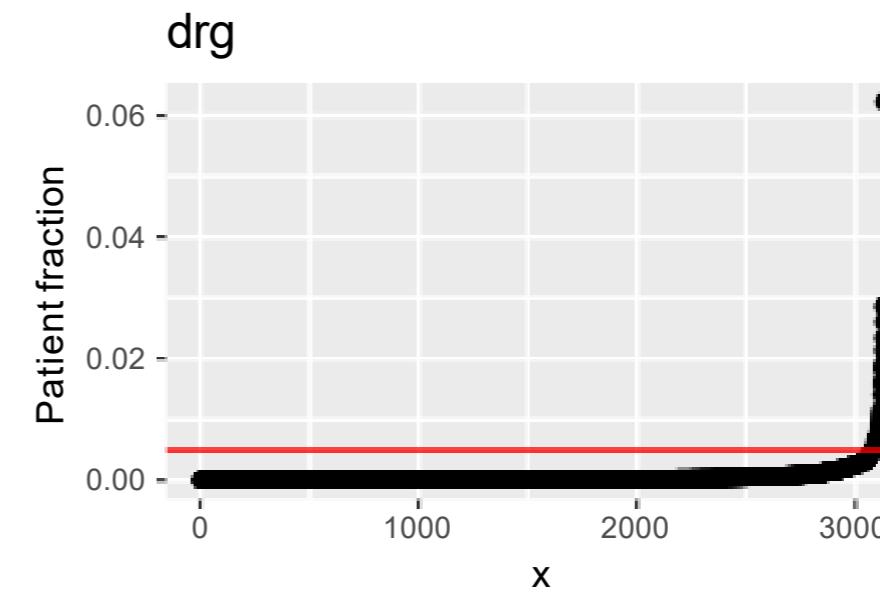
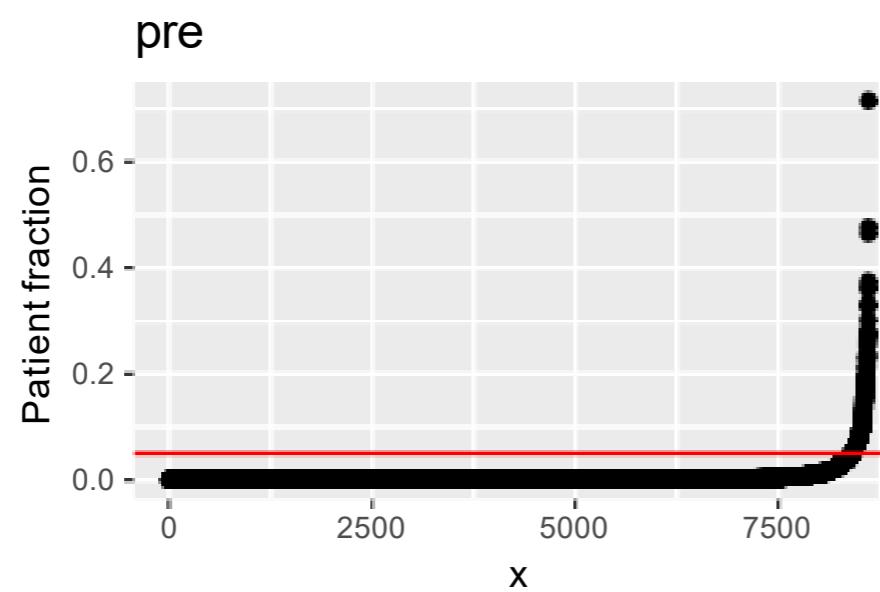
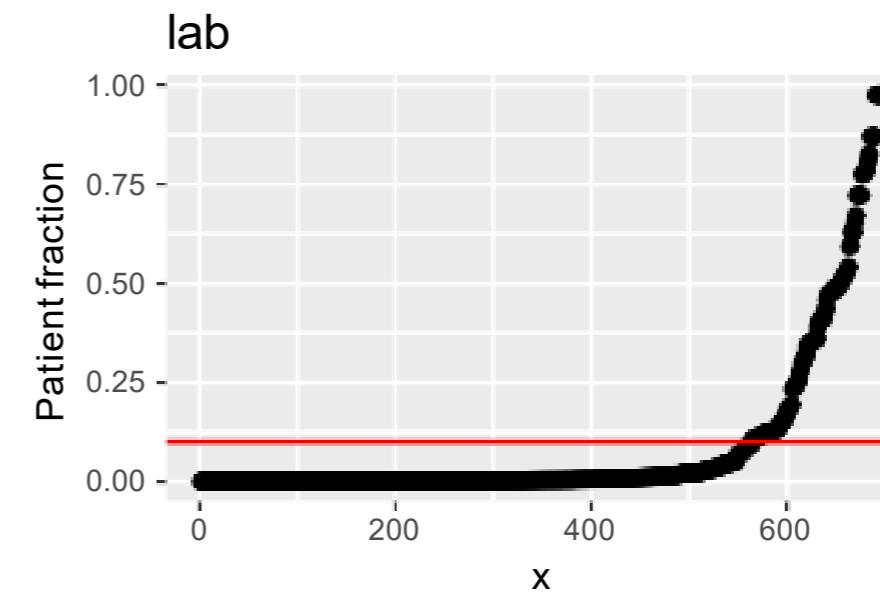
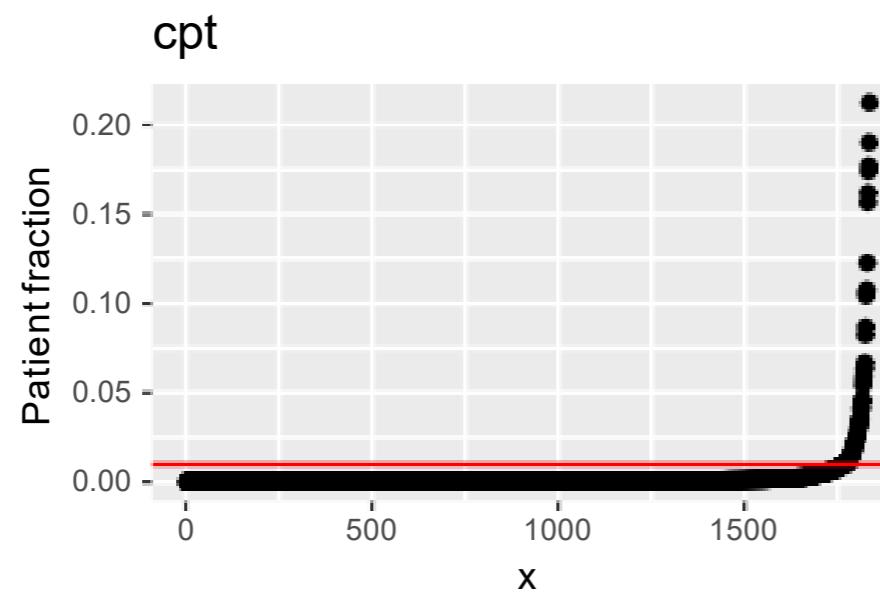
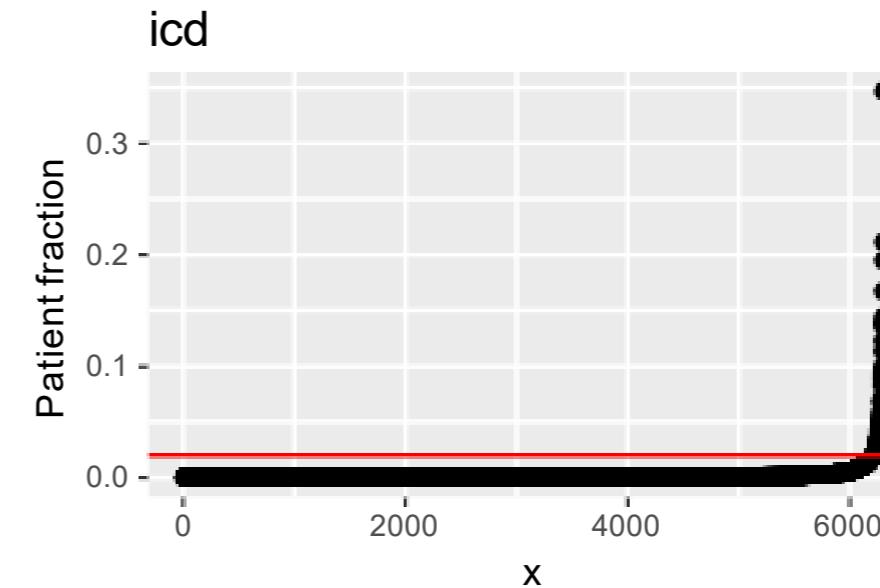
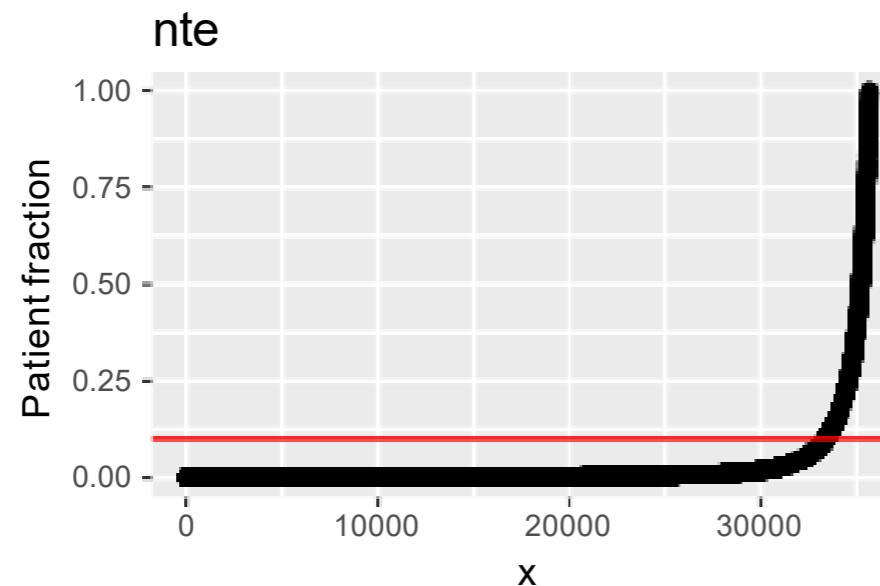
CV cath plm w guid
Temporary tracheosto

Ins nondrug elut cor Serum transint N.E.
Fiber-optic bronches Lt heart angiograp Heart countershock N
Aortocor bypass-2 cor Circumcision Aortocor bypass-1 cor
Pulmon art wedge st Ins drug-elut coron Spinal tap Dx ultrasound-heart
Rt/left heart card a Arterial catheteriza
Inj/irn plated inh Cont inv mec ven 96+
Procedure-one vessel Packed cell transfus
Aortocor bypass-3 cor Vaccination NEC
Closed bronchial bio Entral infus nutrit
Percu abdominal drai Venous cath NEC
Opn/oth rep aort viv Insert intercostal c
Insert endotracheal
Extracorporeal circu Hemodialysis
Platelet transfusion Coronar arteriogr-2 Thoracentesis
Other bronchoscopy 1 int mam-cor artery
Skin closure NEC Parent infus nutrit
Coronar arteriogr-1 Ven cath renal dialy Other phototherapy
Non-invasive mech ve PTCA Sm bowel endoscopyN
Percu endosc gastr Pulsion balloon im
Left heart cardiacc Contr cerebr arterio
Insert 1. vascular st

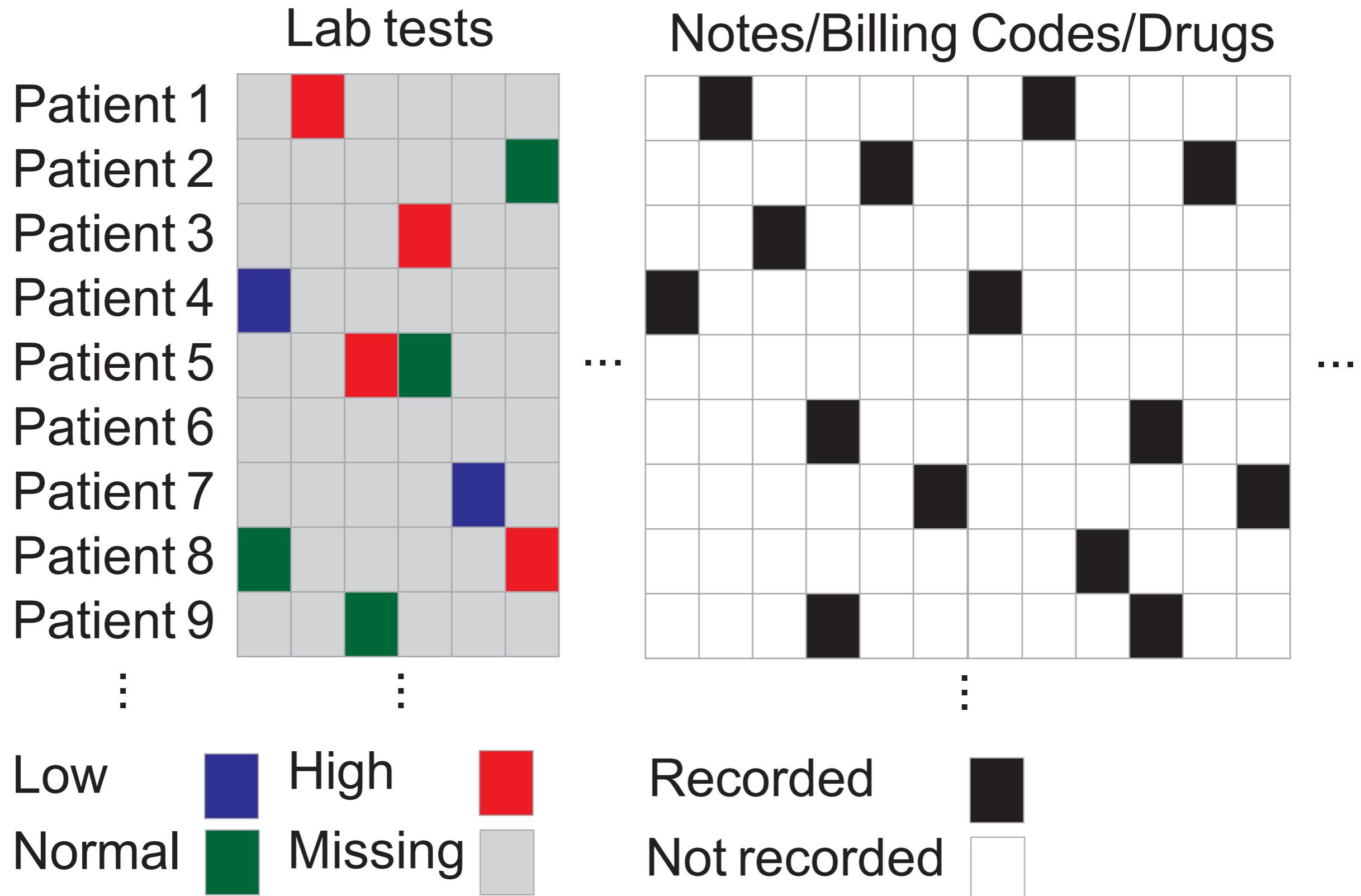
Prescriptions

PNEUMOcoccal Vac Pol
Albuterol 0.083% Neb
Hydralazine Pantoprazole
Calcium Gluconate
Bisacodyl Bisacodyl (Rectal)
Heparin Sodium
Metoprolol Tartrate
Glucagon
Ketorolac
Diltiazem
Haloperidol
Ipratropium Bromide
Dextrose 50%
Insulin Human NPH
Captopril
Chlorhexidine Glucon
Fentanyl Citrate
Neutra-Phos
Heparin Flush
Acetaminophen (Rectal)
Levothyroxine Sodium
OxycoDONE (Immediate)
Morphine Sulfate
Lorazepam Warfarin
Docusate Sodium Aspirin
Magnesium Sulfate Aspirin EC
Morphine Sulfate (Sy)
Oxycodone-Acetaminoph
Potassium Chloride (Liq
Docusate Sodium (Liq
Furosemide
Potassium Chloride
Insulin - Sliding Sc
Propofol Metoprolol Senna
Insulin Atorvastatin Ondansetron
Acetaminophen
Metoprolol
Senna
Lisinopril
Insulin Glargin
Chloride 0.9%

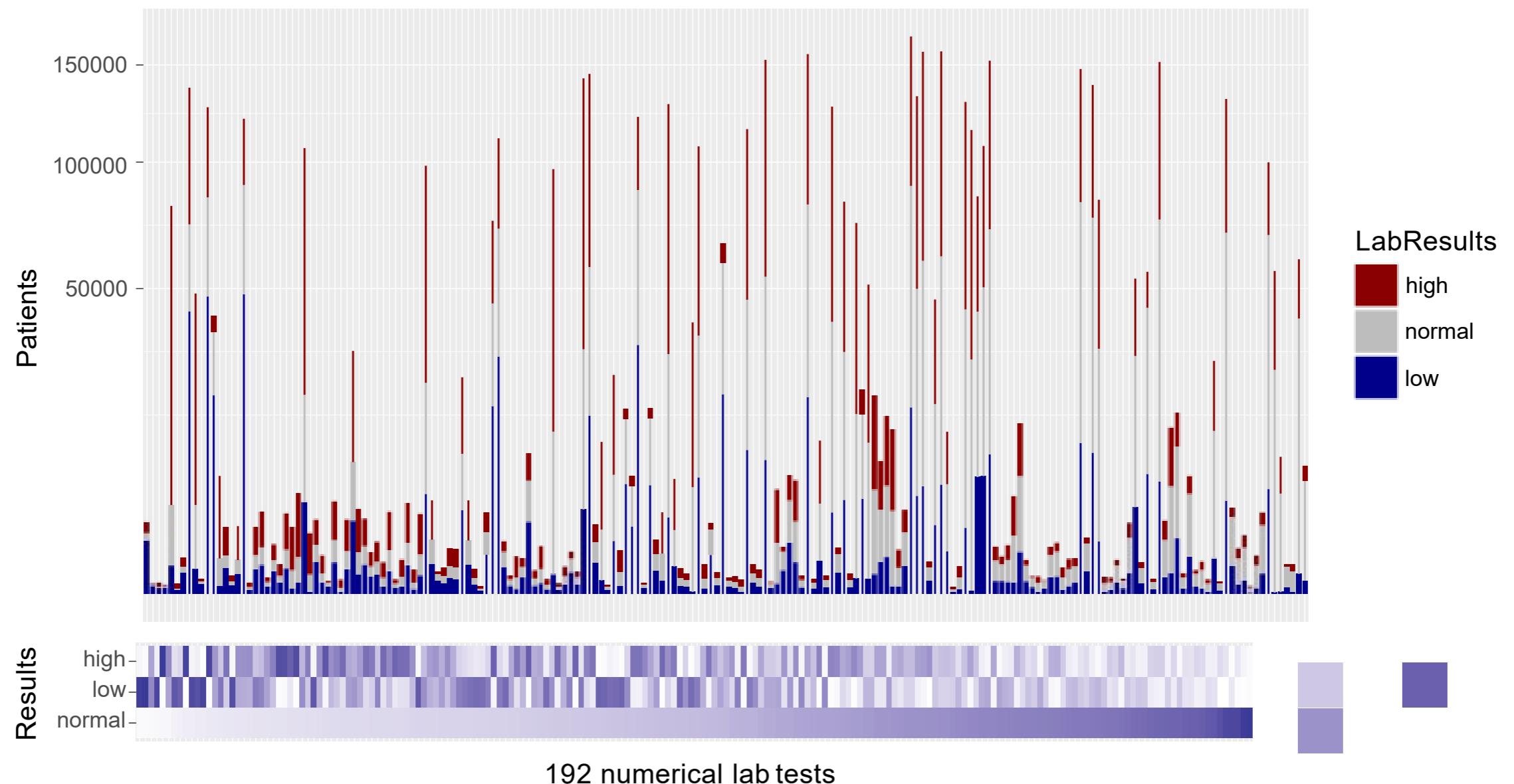
Challenge 2: EHR data are extremely sparse



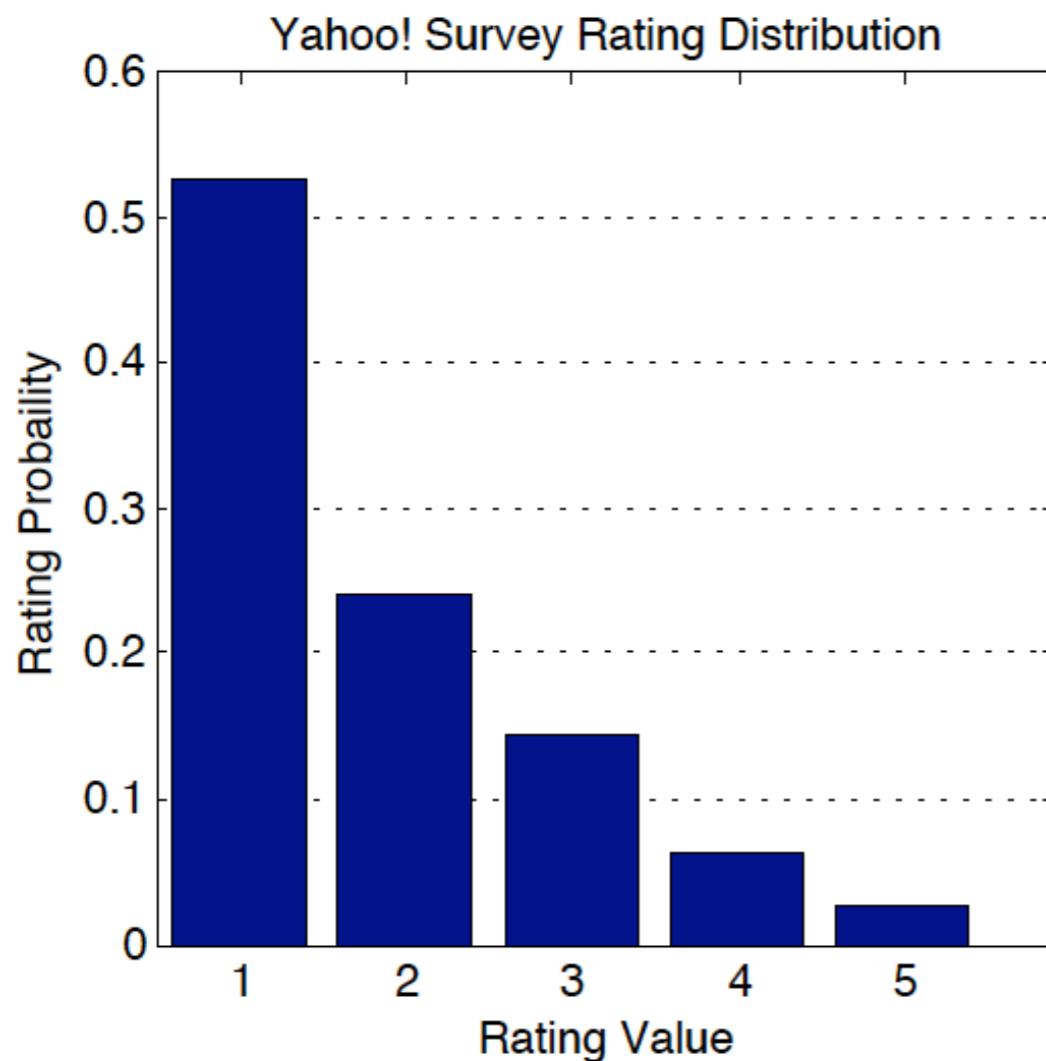
Missing versus unrecorded EHR data



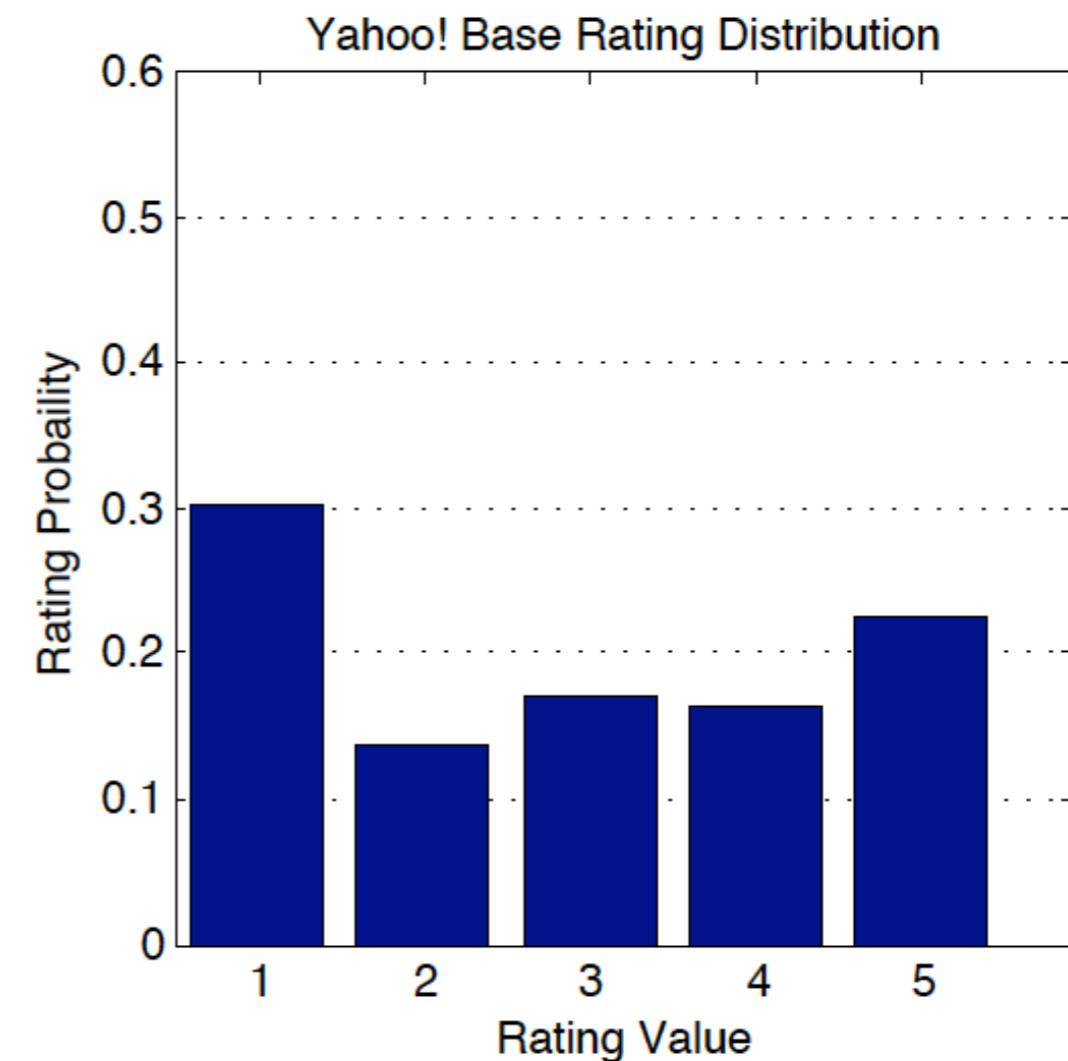
Challenge 3: EHR data are not missing at random (NMAR)



NMAR is common in diverse domain: e.g., Distinct distribution for random and user-selected music ratings



(a) Yahoo! Random



(b) Yahoo! User

Missing mechanism concepts (Rubin & Little, 1987)

Data likelihood $L(\lambda) = p(y^{Obs}, x|\lambda)$:

- **Missing completely at random (MCAR):**

$$\mathcal{L}_{MCAR}(\lambda) = \int_z p(y^{obs}, z|\lambda) dz$$

If data are MCAR, we can ignore the missing indicator x in the data likelihood of y .

- **Missing at random (MAR):**

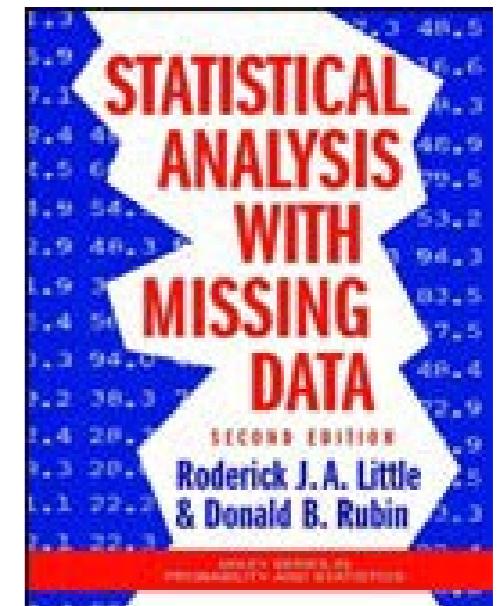
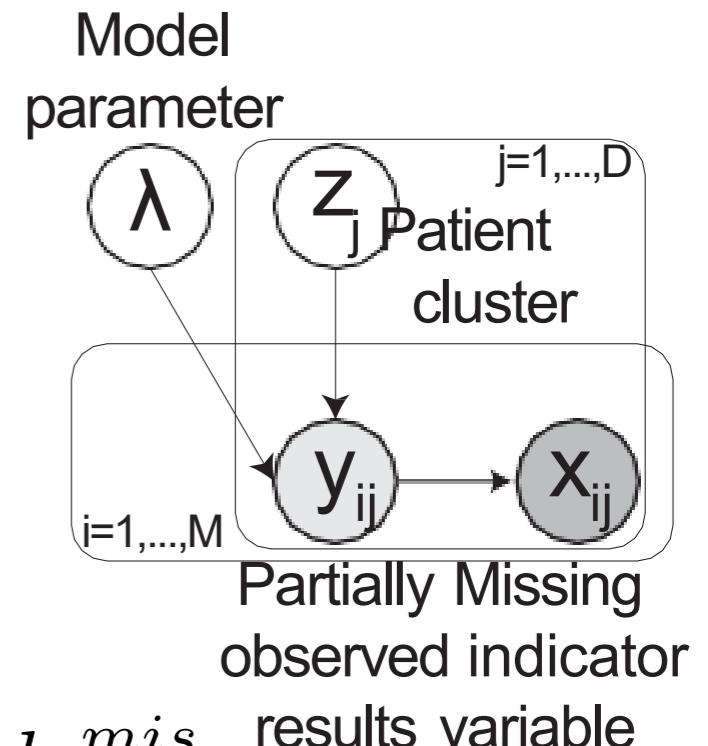
$$\mathcal{L}_{MAR}(\lambda) = p(x|y^{obs}) \int_{y^{mis}} \int_z p(y, z|\lambda) dz dy^{mis}$$

If data are MAR, the missing indicator x only depends on the observed results y^o .

- **Non-missing at random (NMAR):**

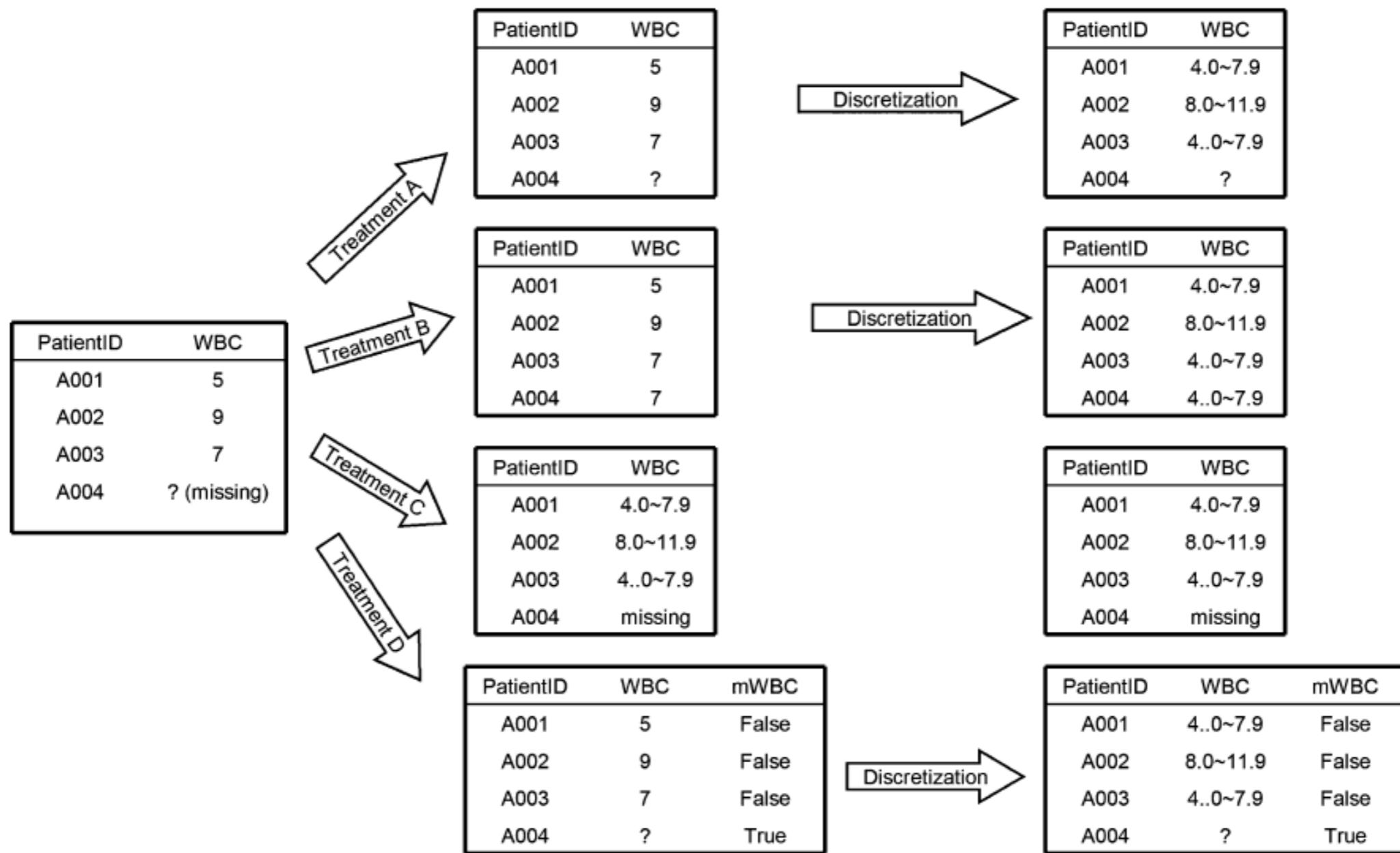
$$\mathcal{L}_{NMAR}(\lambda) = \int_{y^m} \int_z p(y, z|\lambda) p(x|y, z) dz dy^m$$

If data are NMAR (i.e., x always depend on y), we are stuck with the double integral.

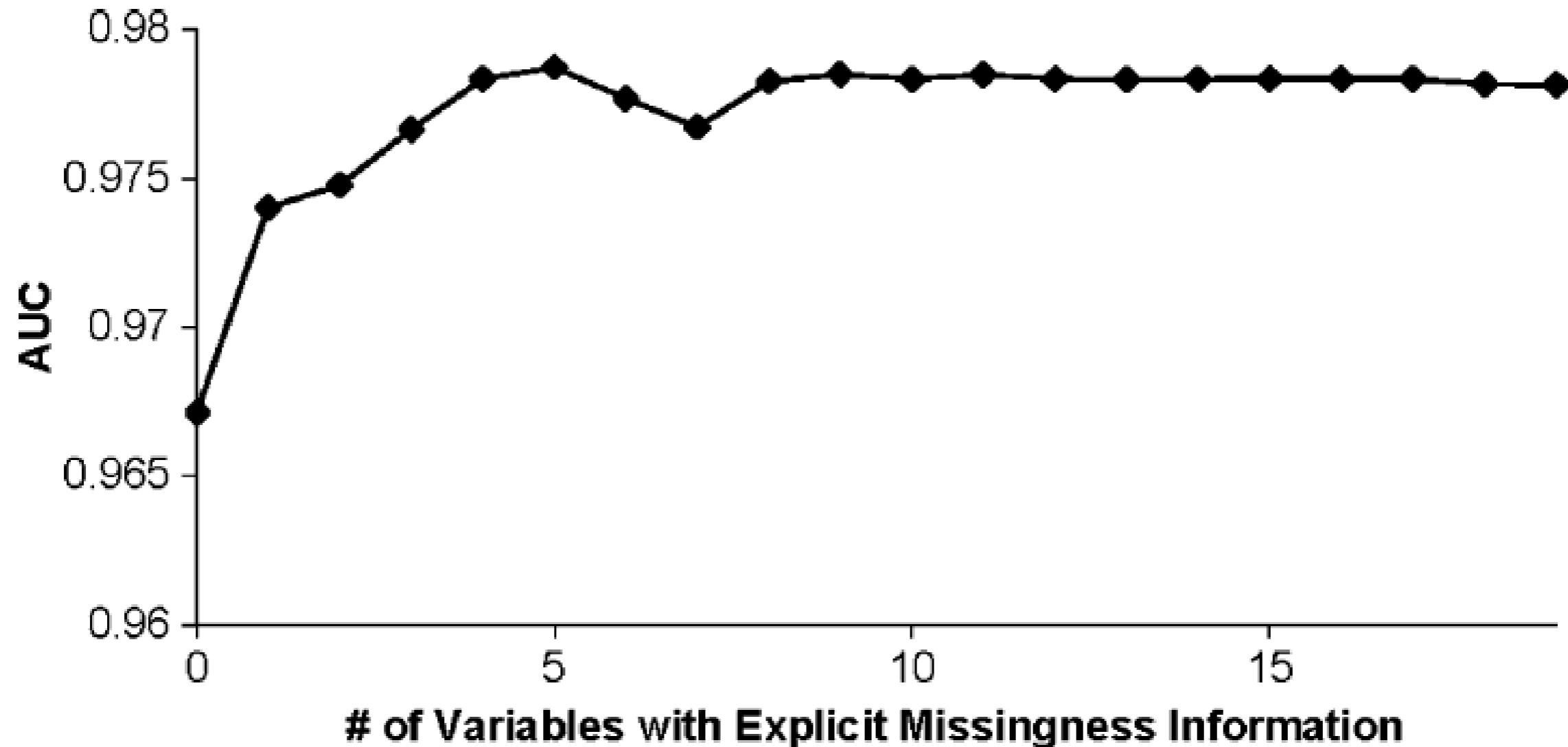


Rubin & Little, Wiley
2014

Coding missing indicator as part of the data

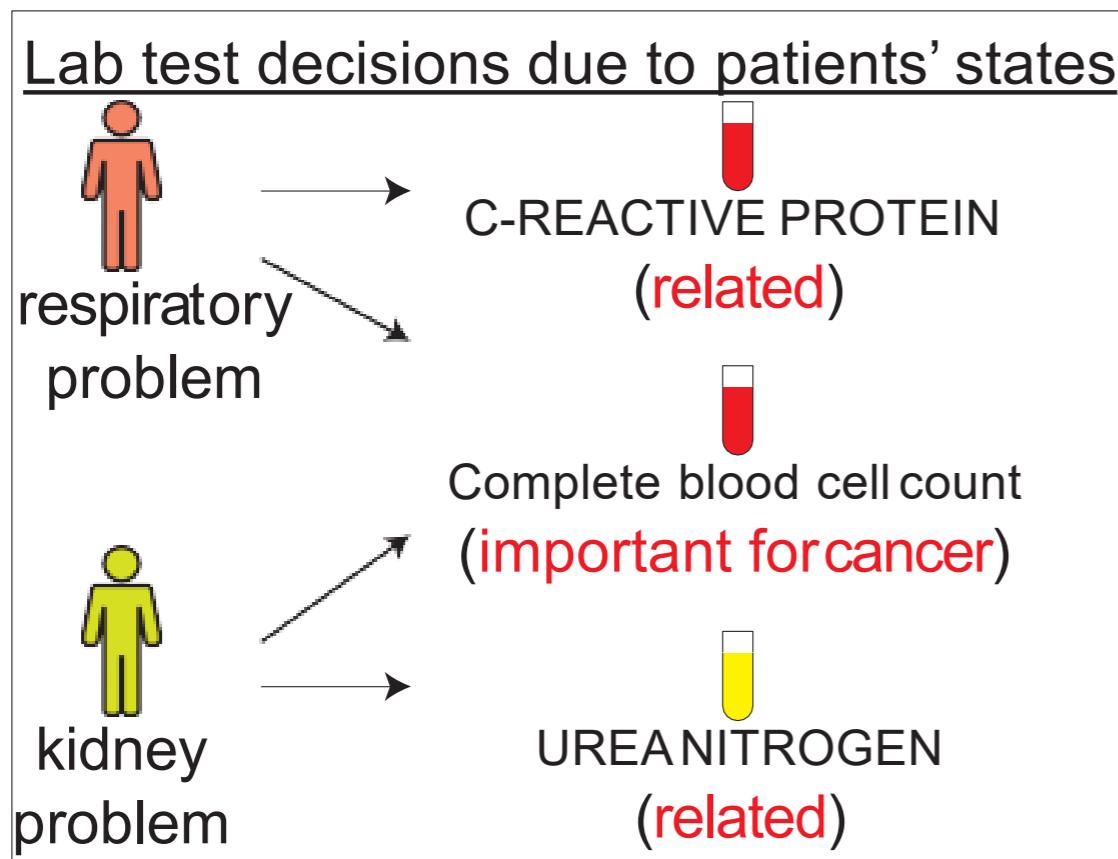


Using missing information improve medical predictions



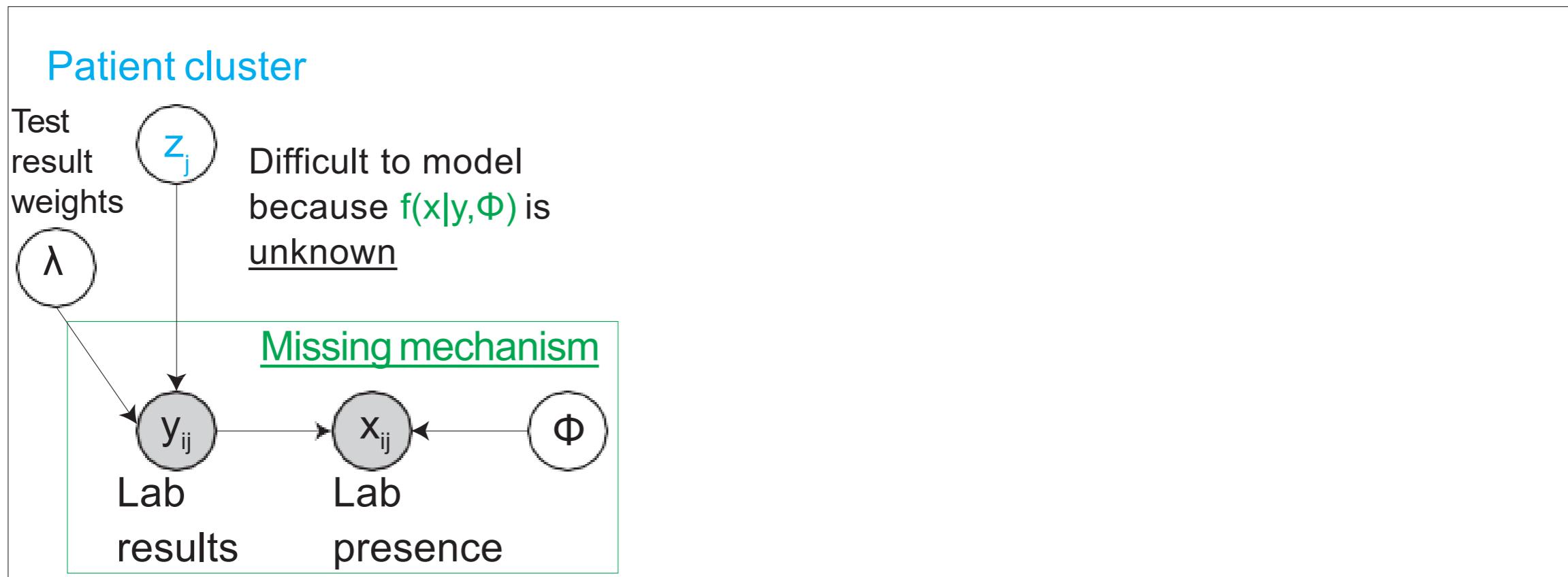
Lin & Haug, JBI 2008

Jointly modeling lab test assignments and lab test results

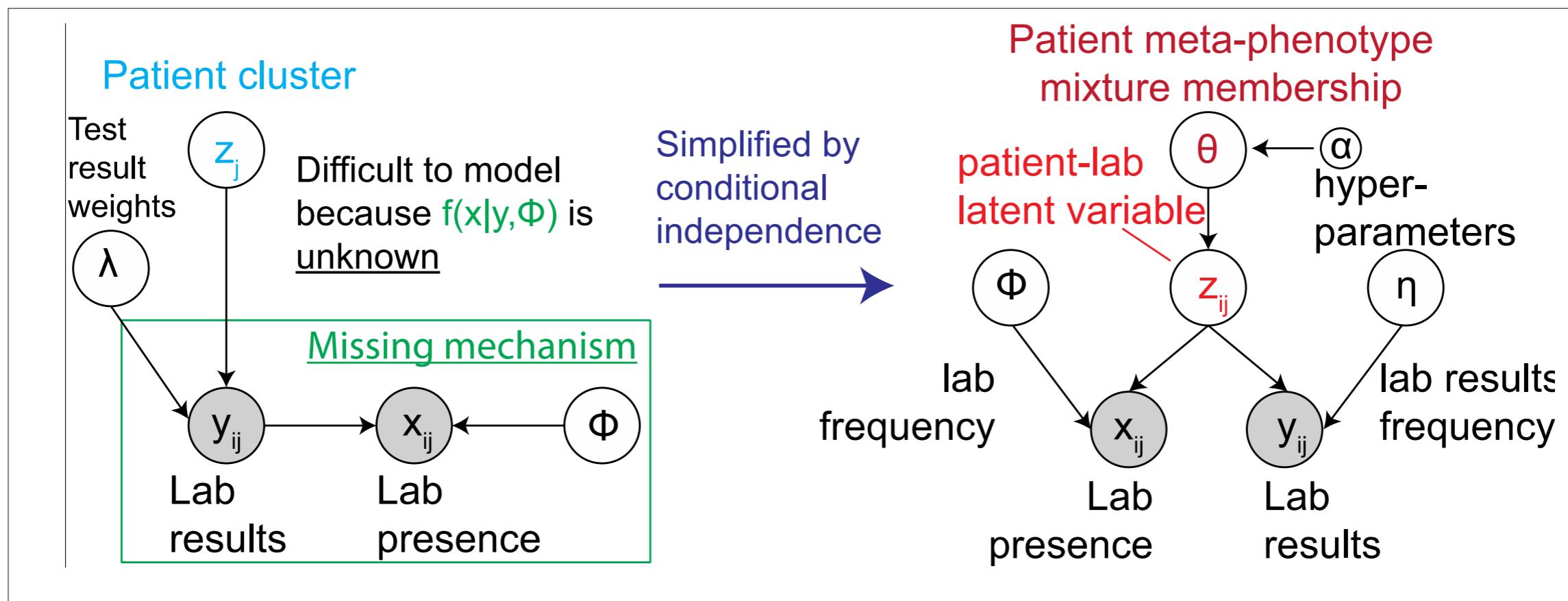
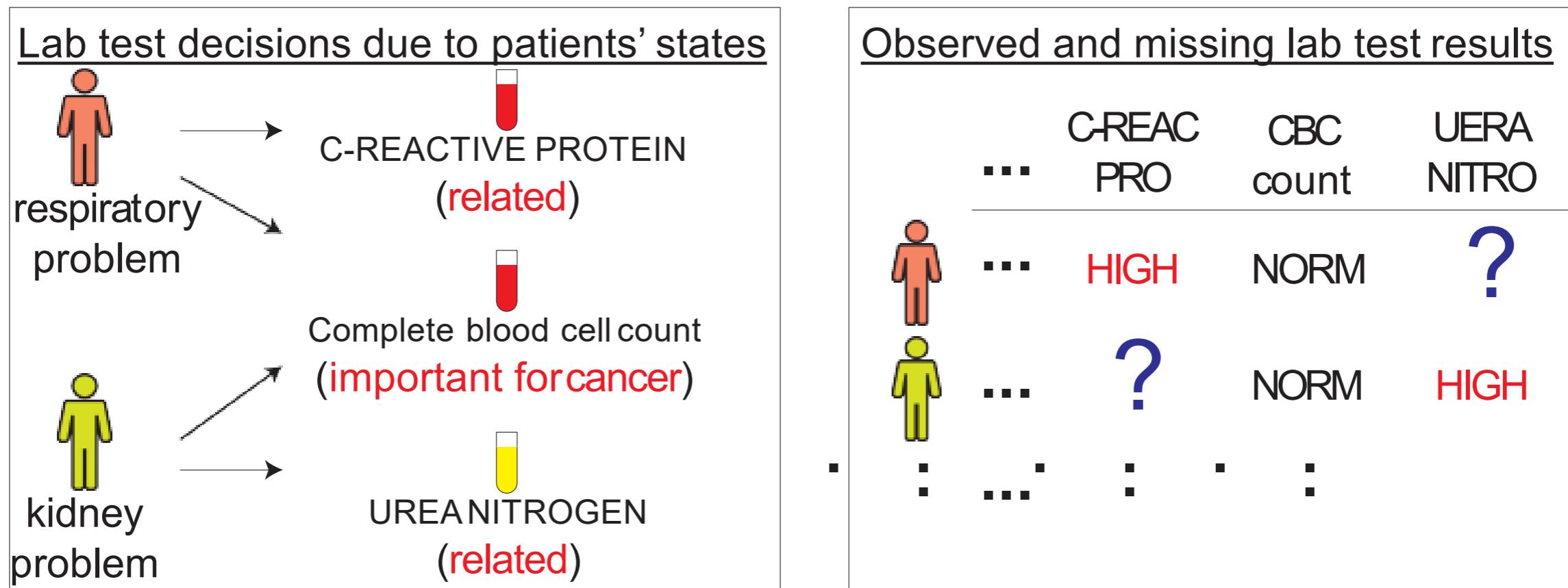


Observed and missing lab test results

	C-REAC PRO	CBC count	UERA NITRO
...	HIGH	NORM	?
...	?	NORM	HIGH
:	...	:	:



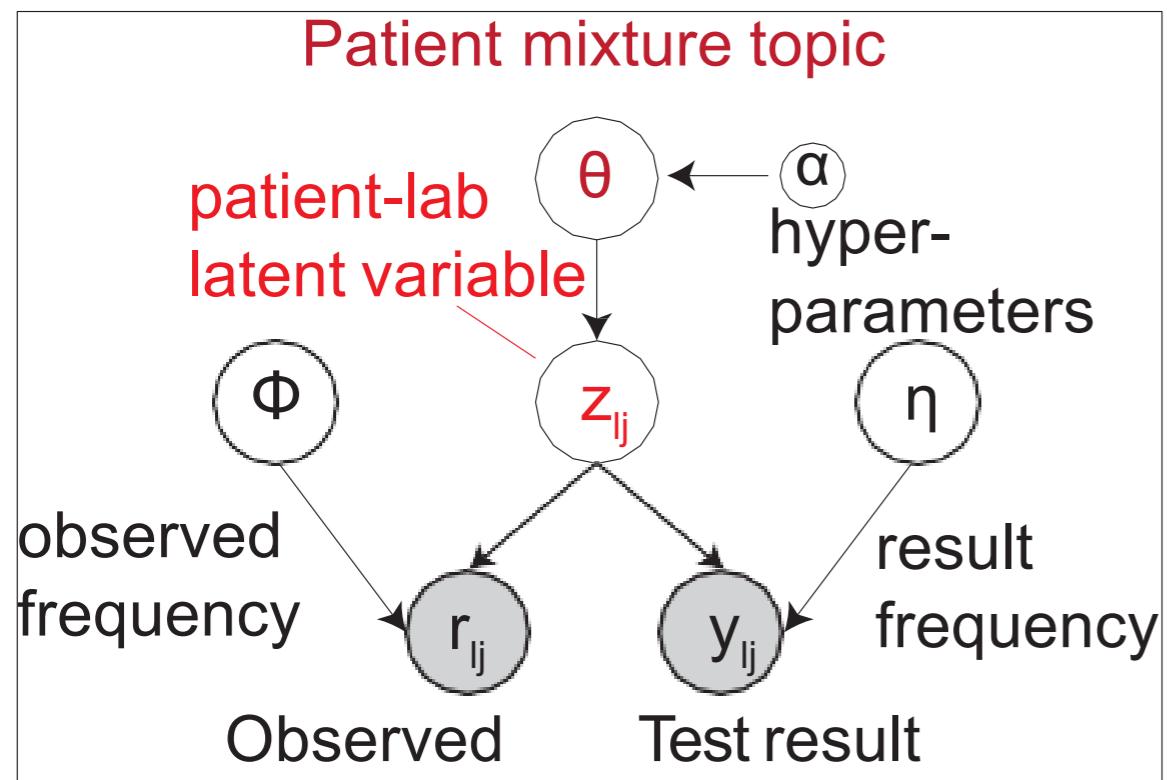
Modeling missing mechanism in lab test is challenging



Jointly modeling lab missing indicators and lab test results

$$\ln q(z_{lj}^{QA} = k) \propto \underbrace{E_q[\ln(n_{jk}^{(-l)})]}_{\text{patient } j\text{'s score for topic } k} + \underbrace{E_q[\ln p(y_{lj}|z_{lj} = k, \Theta_k)]}_{\text{score likelihood for question } l} + \underbrace{E_q[\ln p(r_{lj}|z_{lj} = k, \Theta_k)]}_{\text{missing indicator likelihood}}$$

$$E_{q(\mathbf{z}_k)} [\ln p(r_{lj}|z_{lj} = k, \Theta_k)] = E_q[\ln(\frac{p_{lk}^{-(l,j)}}{v^l p_{lk}^{-(l,j)} q_{lk}^{-(l,j)}})^{r_{lj}}] \\ E_q[\frac{q_{lk}^{-(l,j)}}{v^l p_{lk}^{-(l,j)} + q_{lk}^{-(l,j)}})^{1-r_{lj}}]$$



- $p_{lk} = \sum_j [z_{lj} = k][r_{lj} \neq 1]$ is the number of patients assigned to topic k when lab / is **observed**
- $q_{lk} = \sum_j [z_{lj} = k][r_{lj} = 0]$ is the number of patients assigned to topic k when lab / is **not observed**

Heterogeneous EHR data are related for the same patients

Related binary EHRphenotypes for the same patients

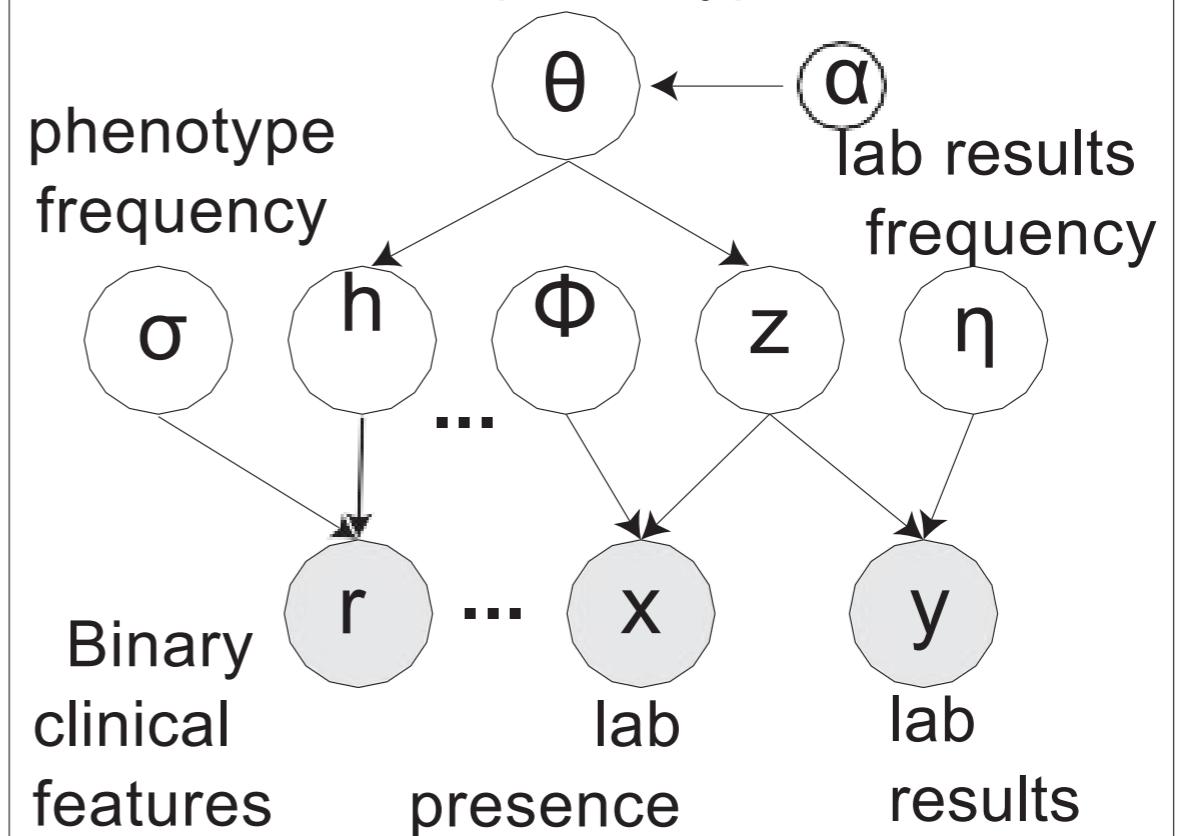
	...	Asthma	Influenza	Pneumonia	Chronic kidney disease
respiratory problem	...	✓	NA	NA	NA
kidney problem	...	NA	NA	NA	✓
...

Lab tests of the same patients

	C-REAC PRO	CBC count	UERA NITRO
...	HIGH	NORM	?
...	?	NORM	HIGH
...

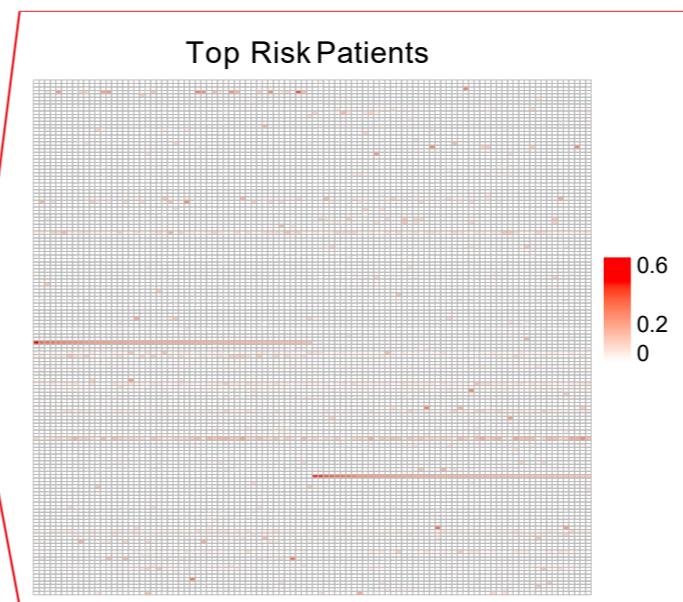
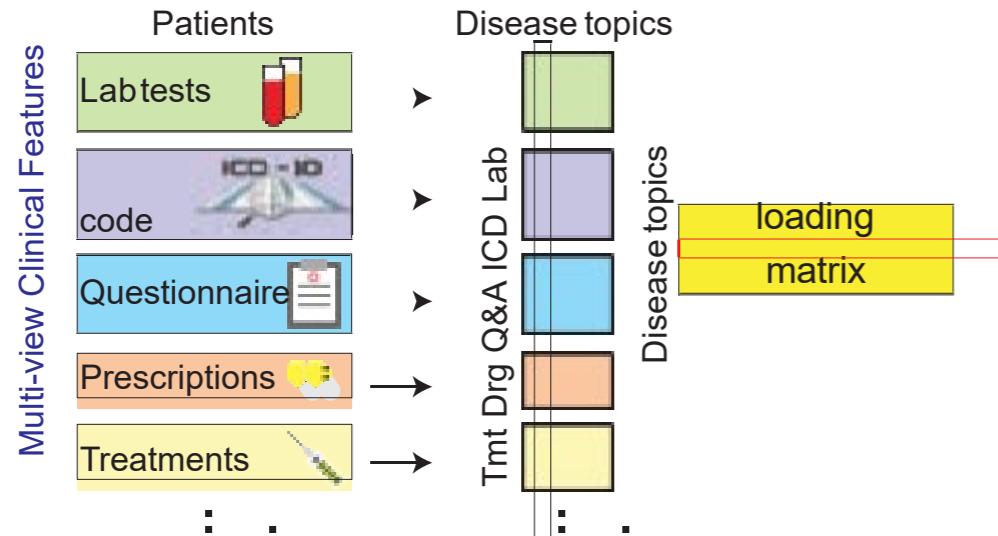
Joint model of related data types

Patient meta-phenotype mixture

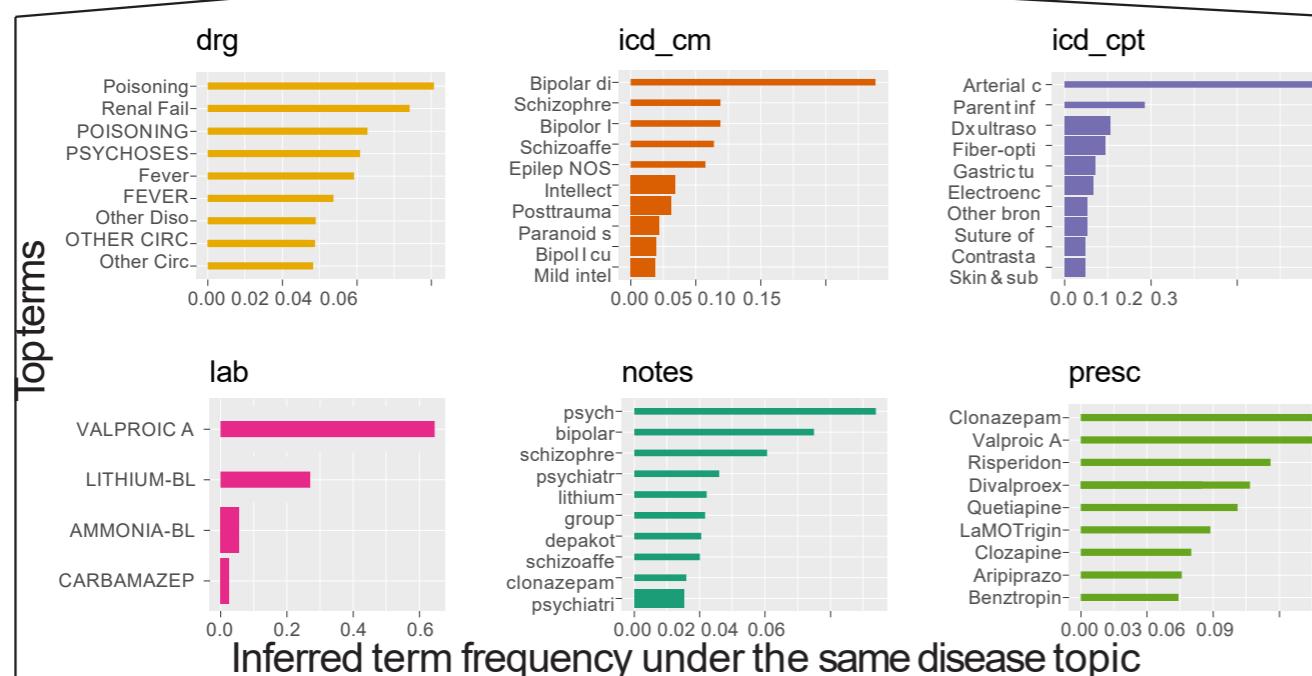
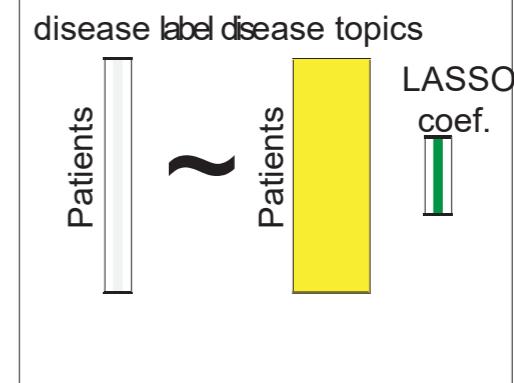


MixEHR for learning multimodal EHR (Li et al, in prep)

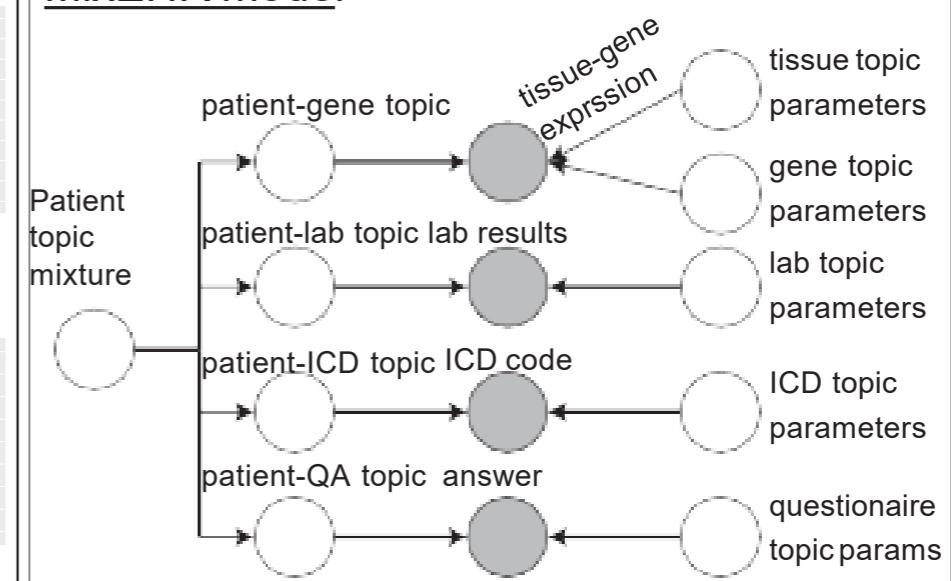
A. Multi-view learning of EHR data



B. disease regression



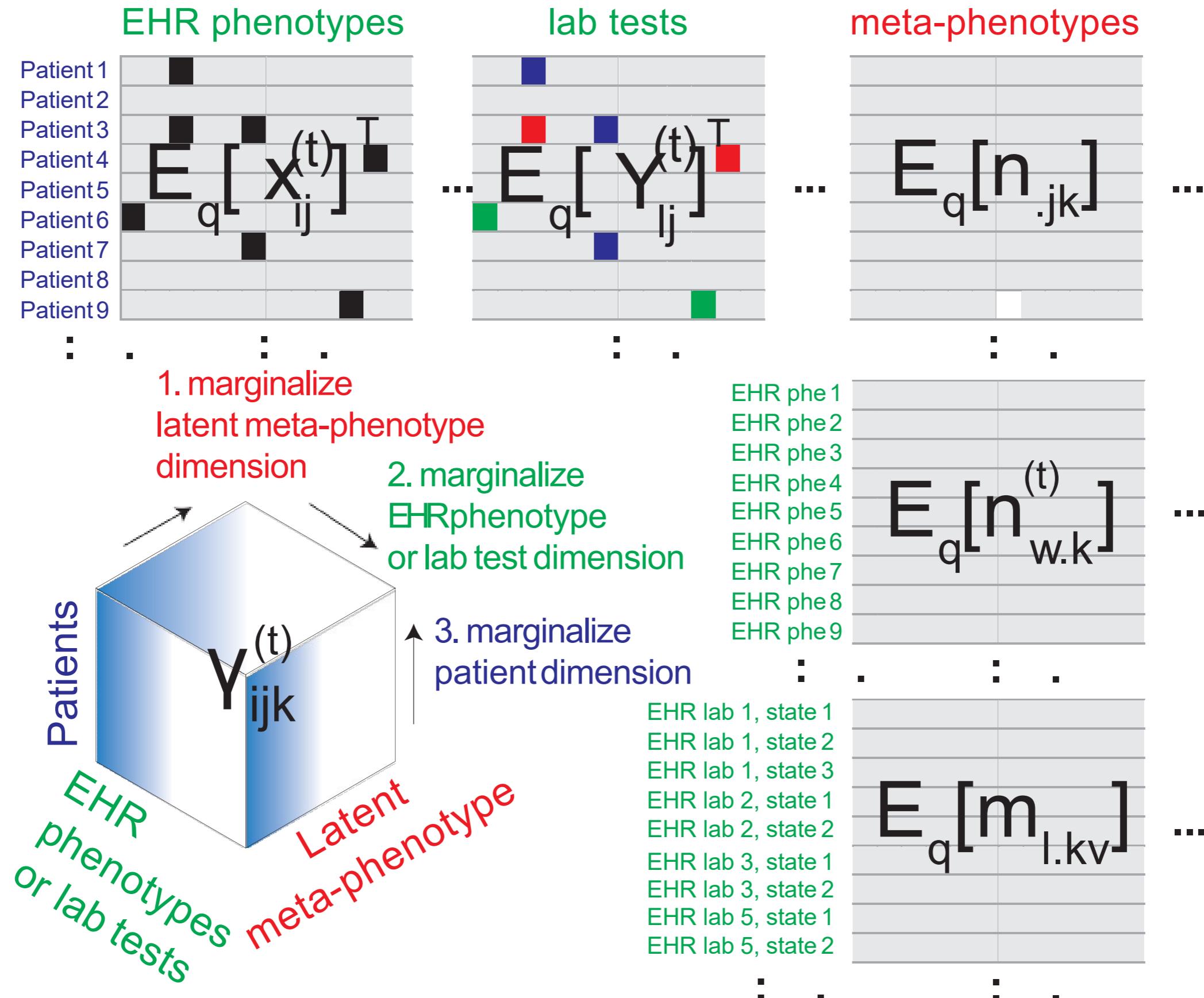
MixEHR model



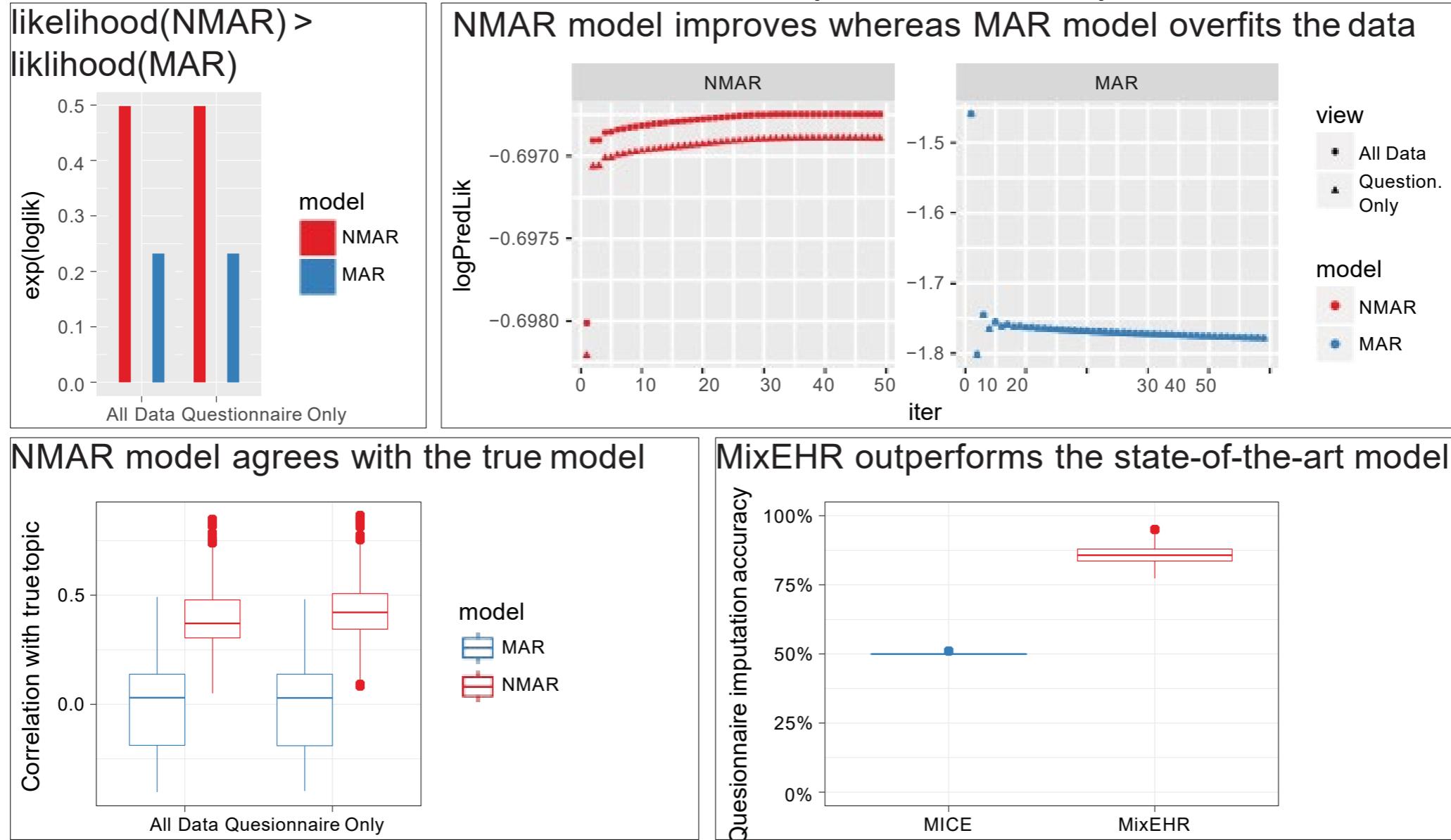
$$\ln q(z_{lj}^{Gene} = k) \propto E_{q(z_{jk}^{(-l)})} [\ln(n_{jk}^{(-l)})] E_{q(\Theta_k)} [\ln p(y_{lj} | z_{lj} = k, \Theta_k)]$$

$$n_{jk}^{(-l)} = \sum_{I \neq l} [z_{ij}^{Gene} = k] + \sum_l [z_{lj}^{ICD} = k] + \sum_l [z_{lj}^{Lab} = k] + \dots$$

Meta-phenotype learning from multimodal patient-EHR



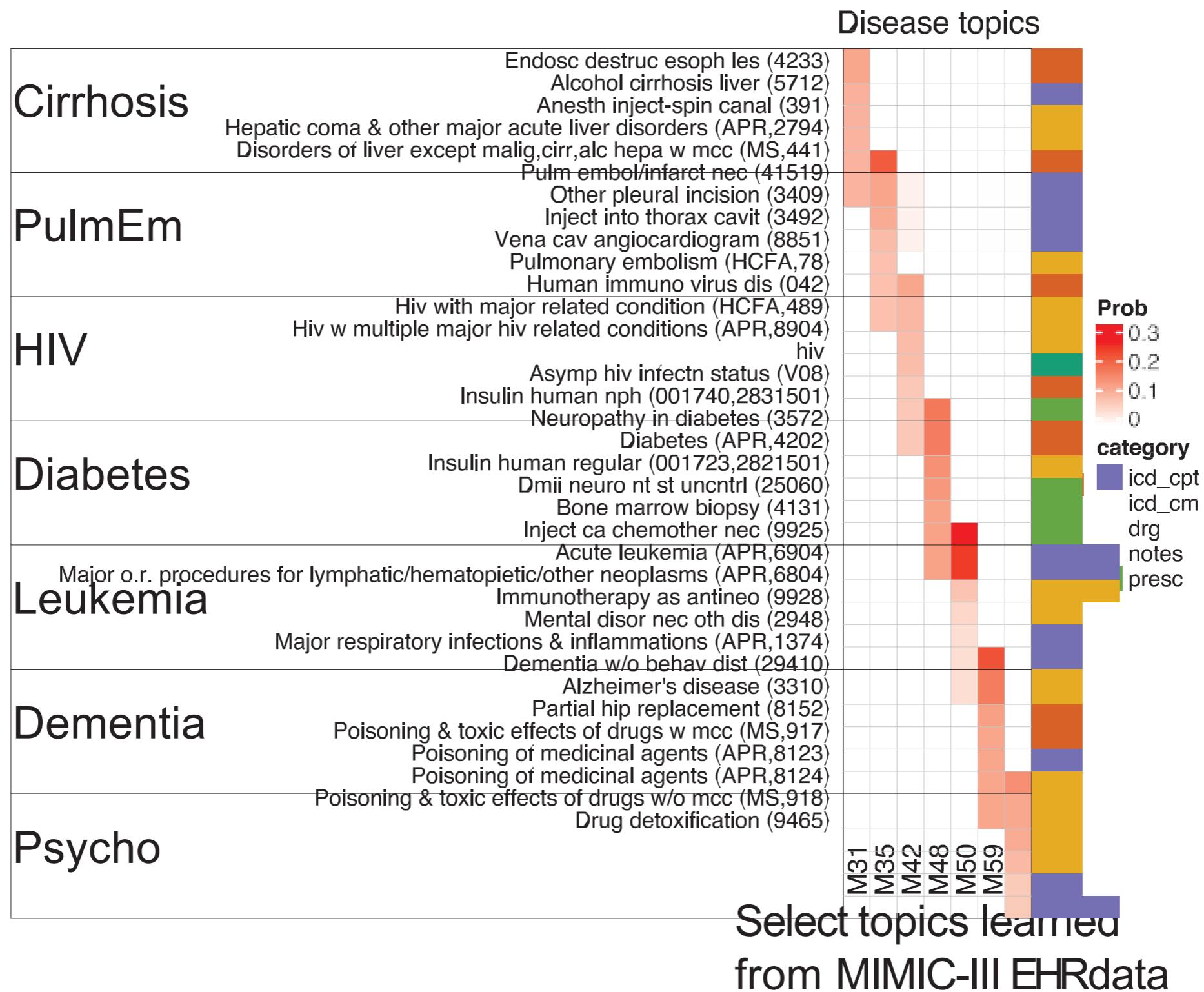
MixEHR-NMAR model improves imputations



- We simulated NMAR data from MixEHR trained on the real data;
- Simulated the missing indicators from the topic-specific frequency;
- We evaluated 4 models: (1) MixEHR_nmar modeling NMAR using Questionnaire Only; (2) MixEHR_nmar using All Data; (3) & (4) MixEHR_mar assuming missing at random (MAR) using Questionnaire Only and All Data.

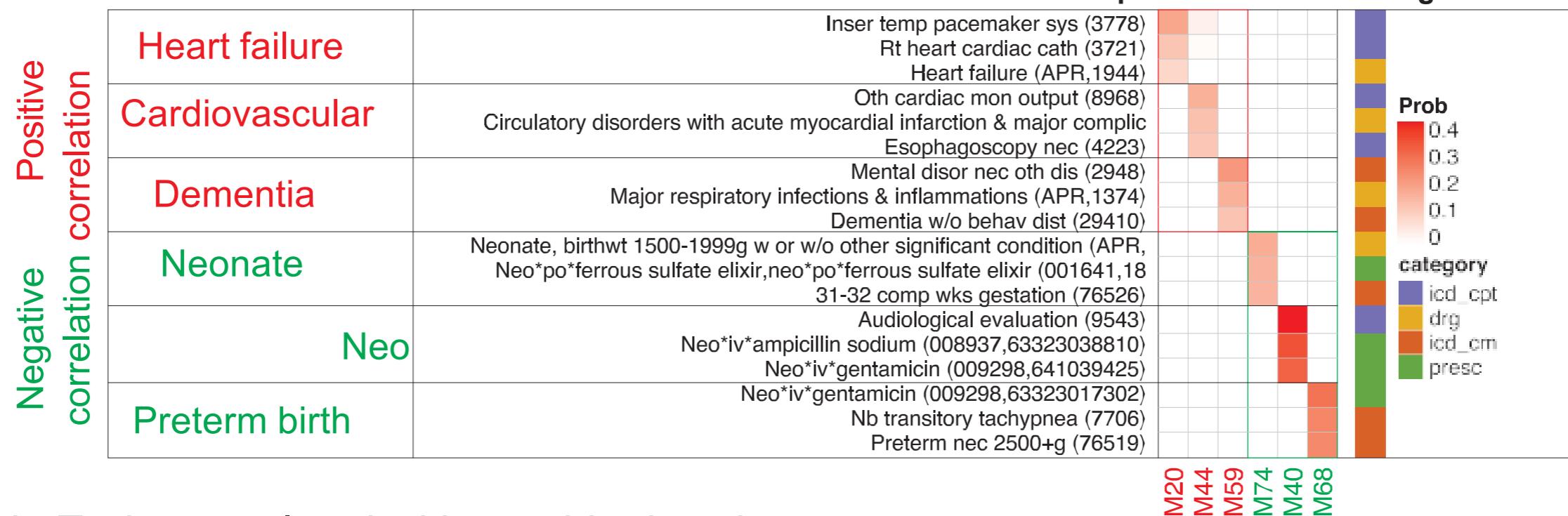
MixEHR learns meaningful disease topic embeddings

a. Top clinical features from select disease topics



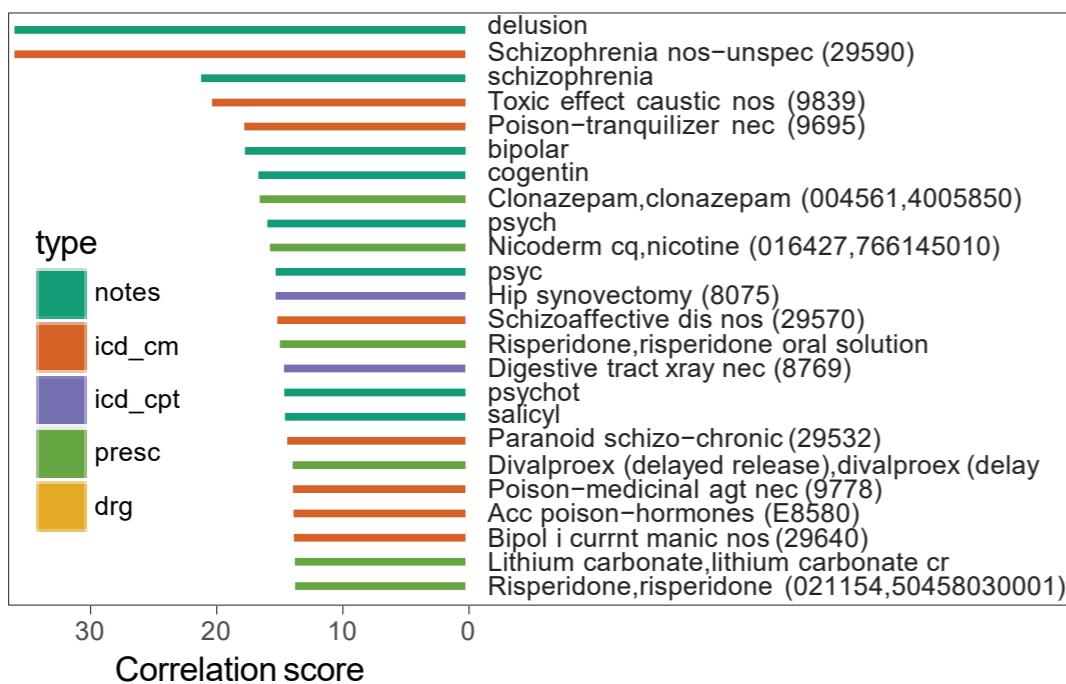
MixEHR learns meaningful disease topic embeddings

a. Topics correlated with age

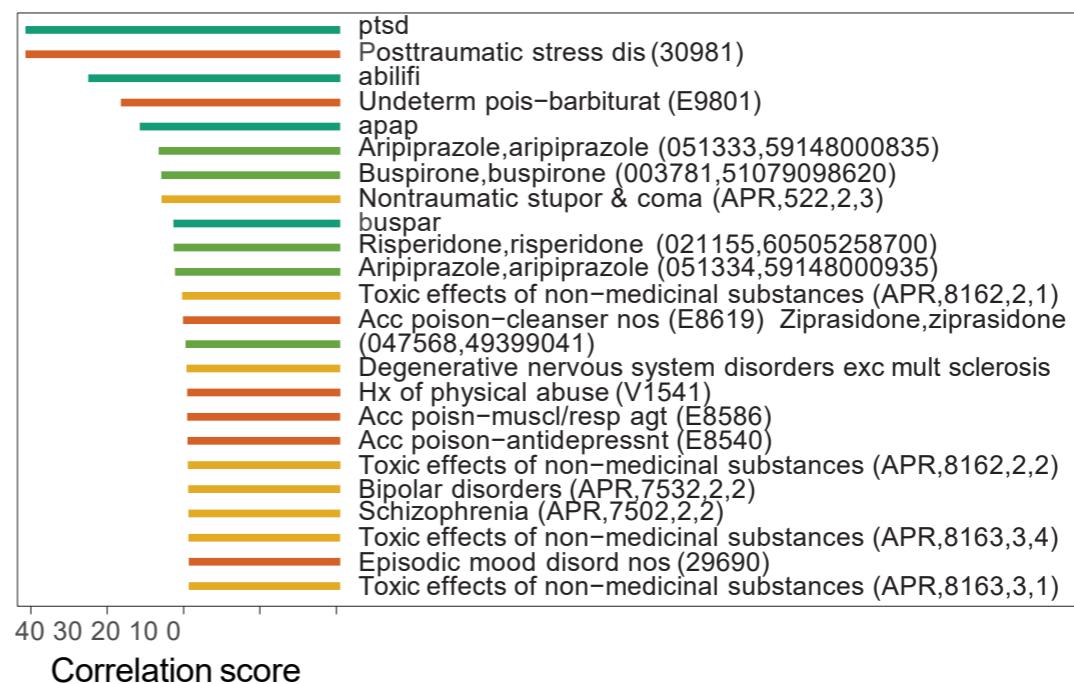


b. Topics correlated with psychiatric traits

Schizophrenia-associated codes

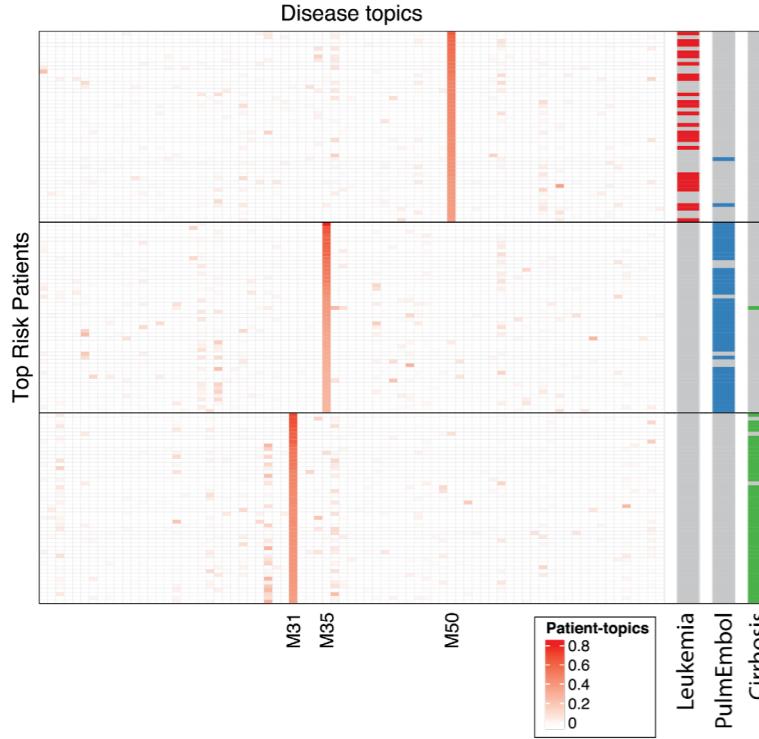


PTSD-associated codes

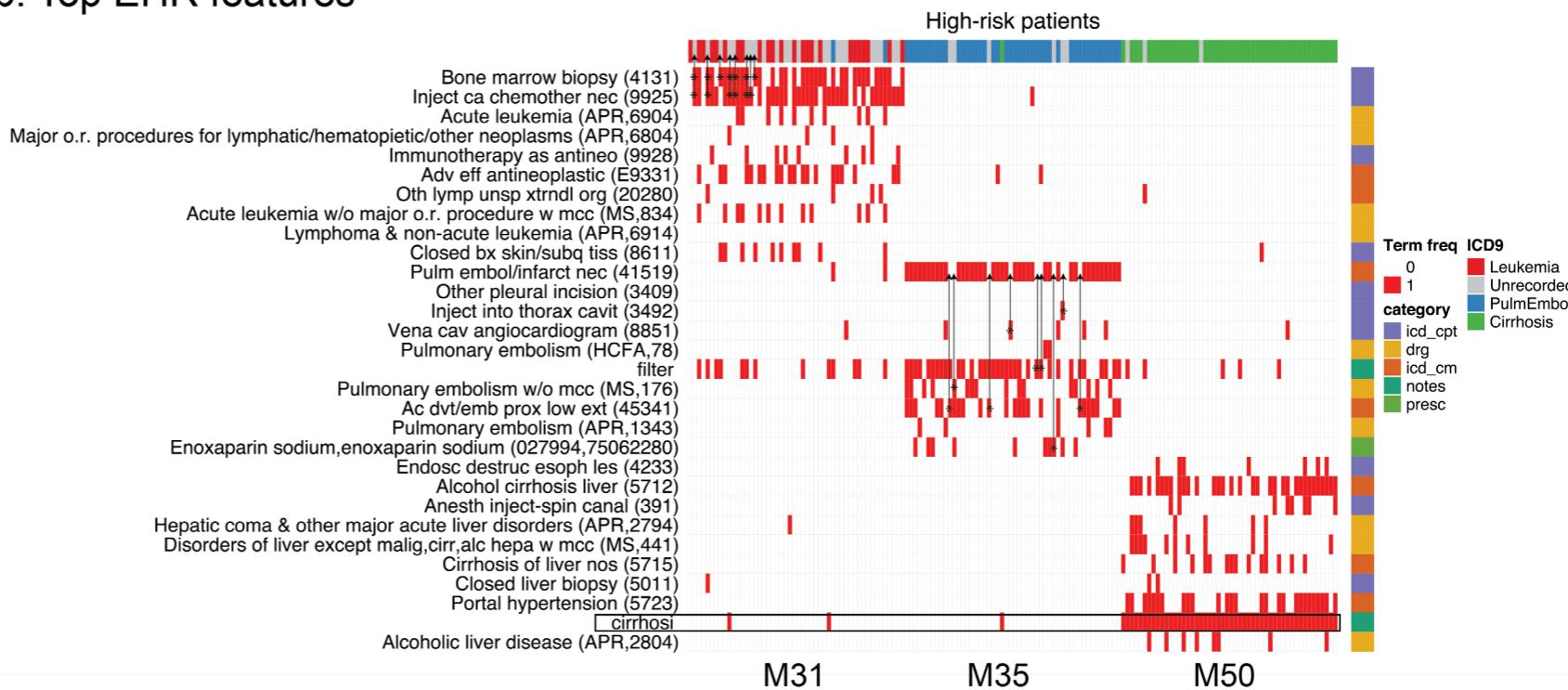


MixEHR prioritizes patients by disease mixture

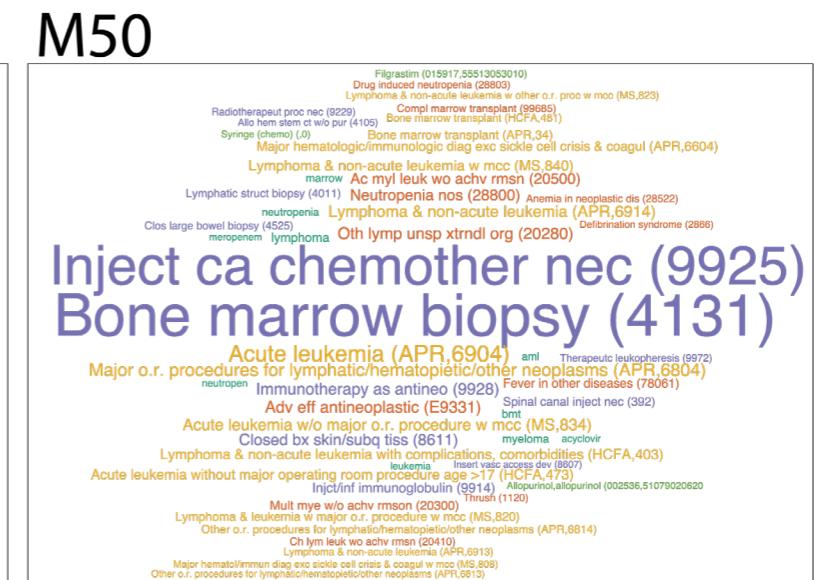
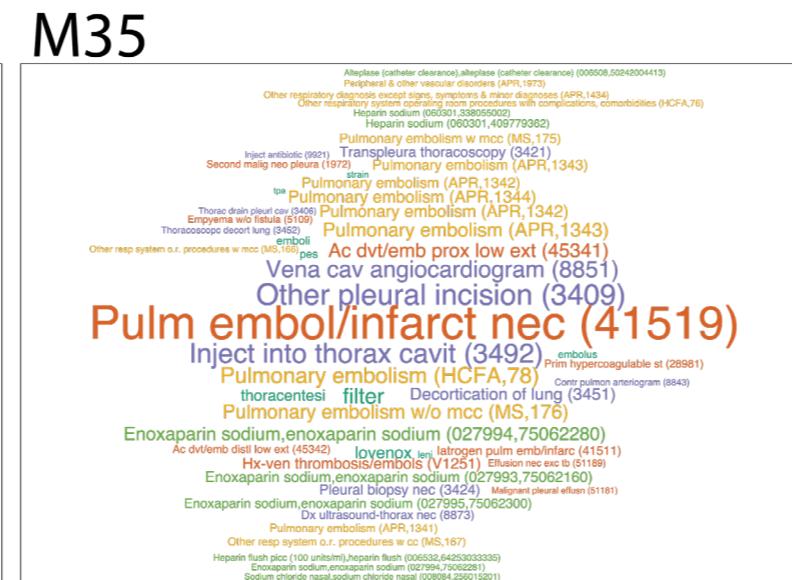
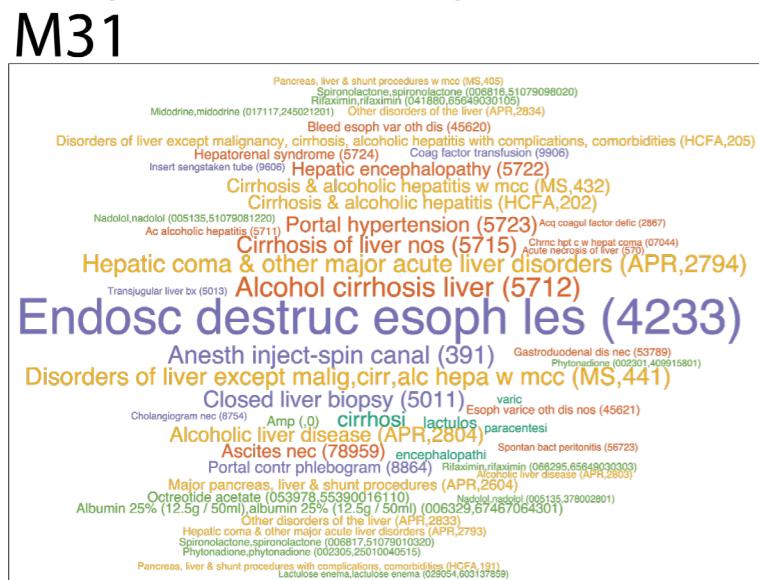
a. Top patients in select disease topics



b. Top EHR features

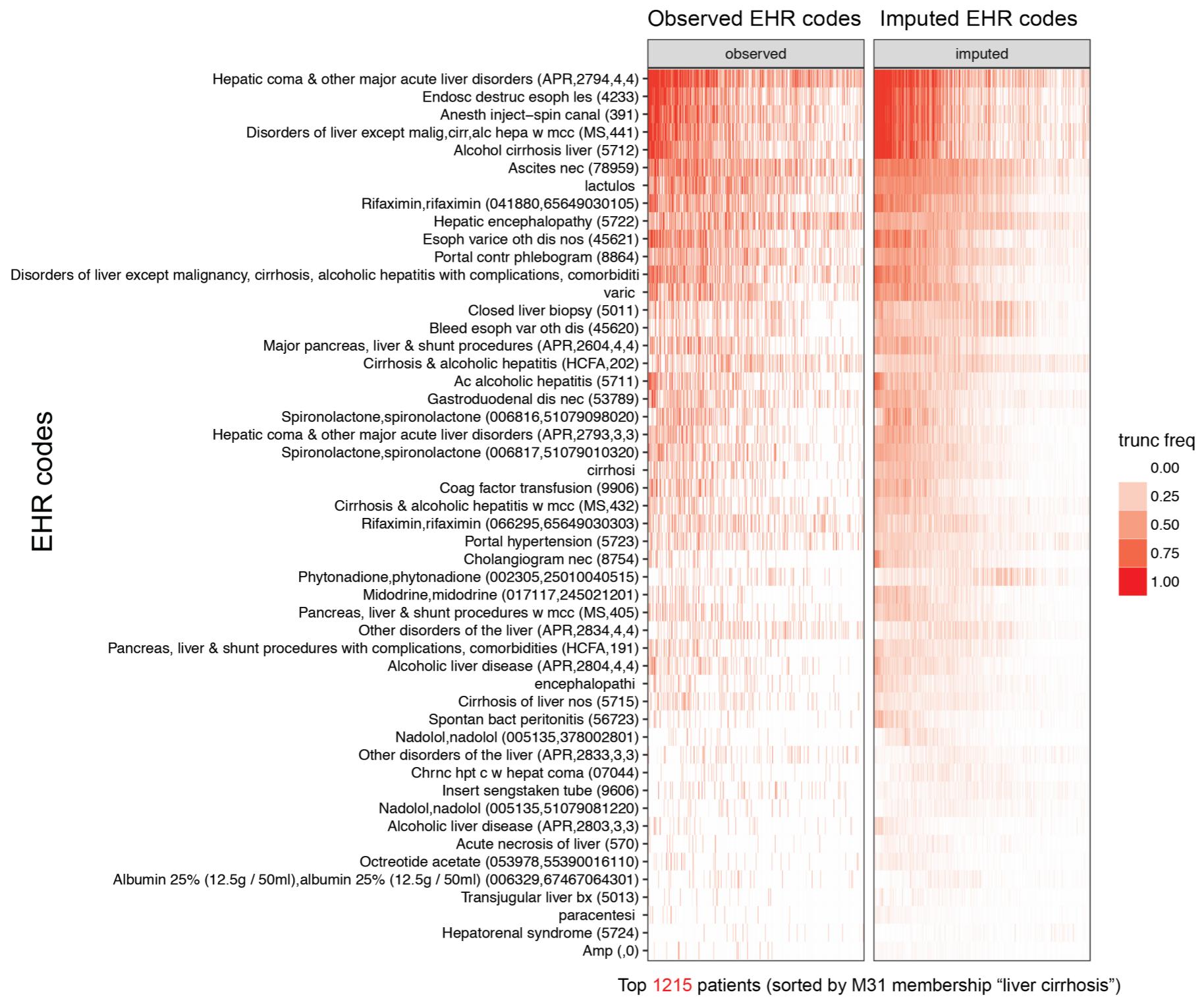


c. Top terms in each topic

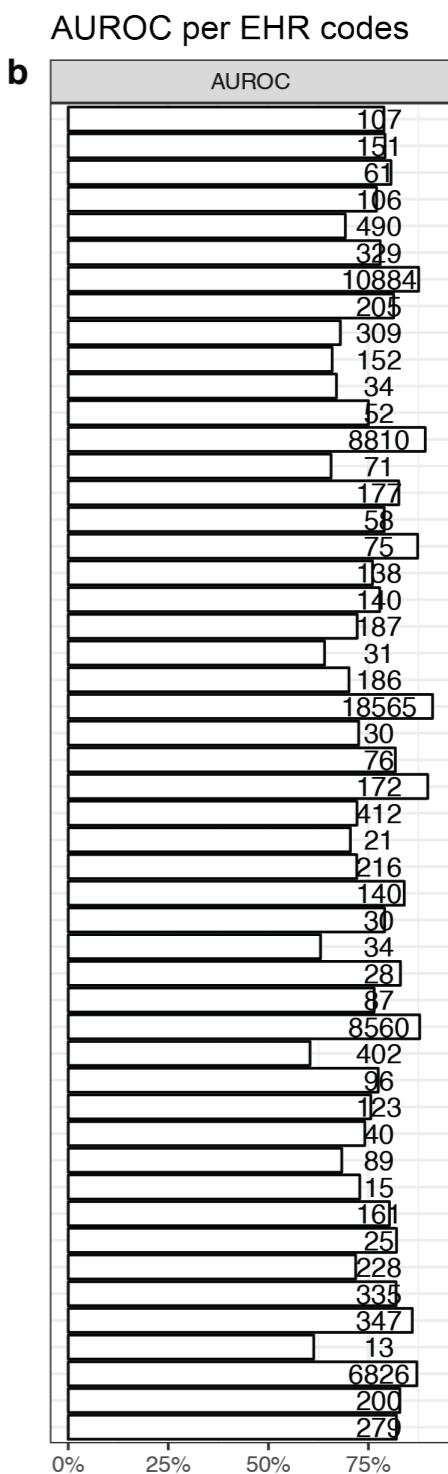


MixEHR accurately imputes missing EHR code

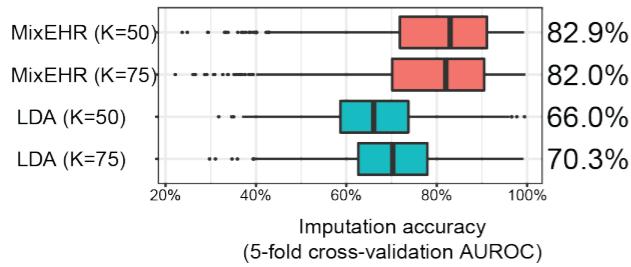
a



Top 1215 patients (sorted by M31 membership "liver cirrhosis")

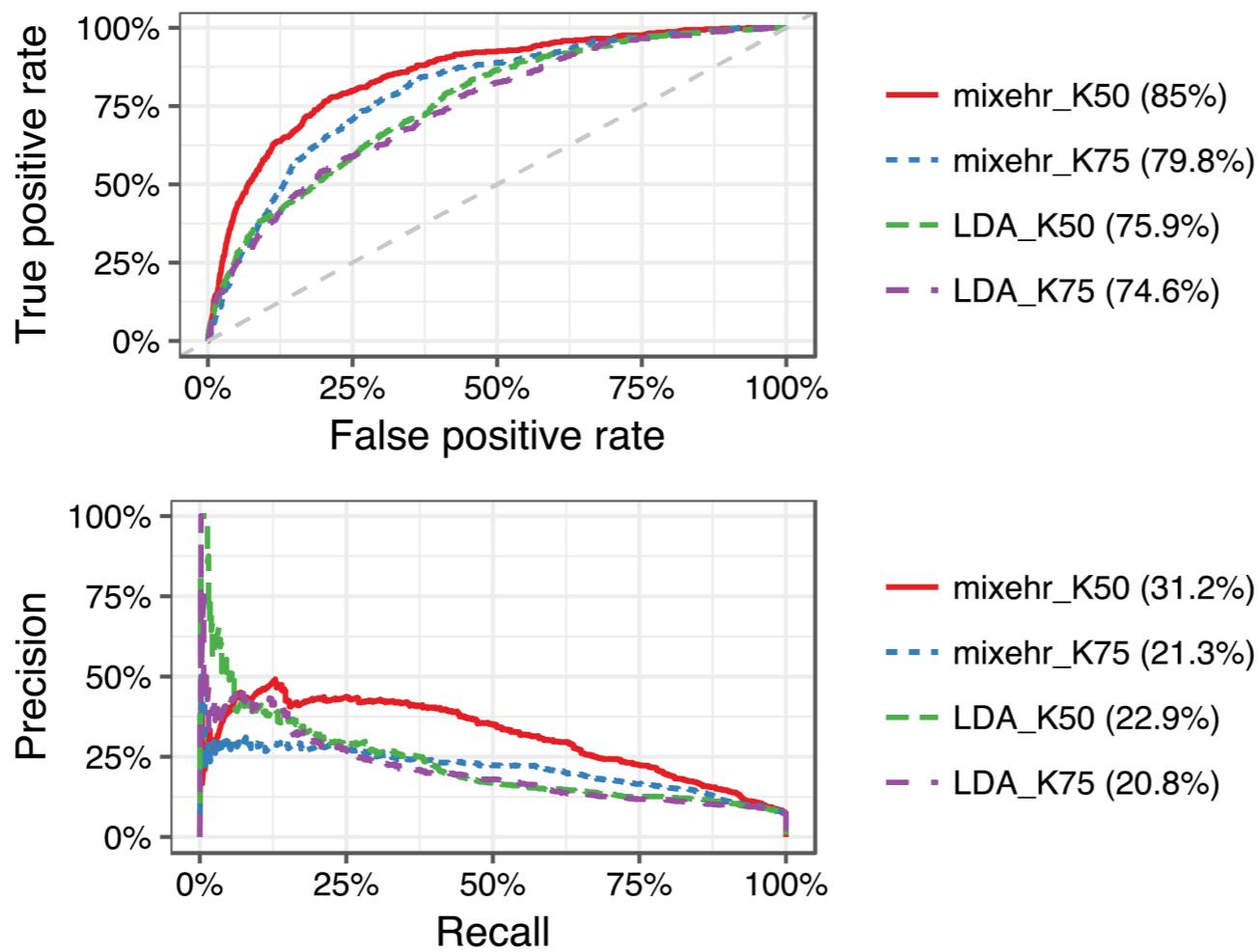


c.

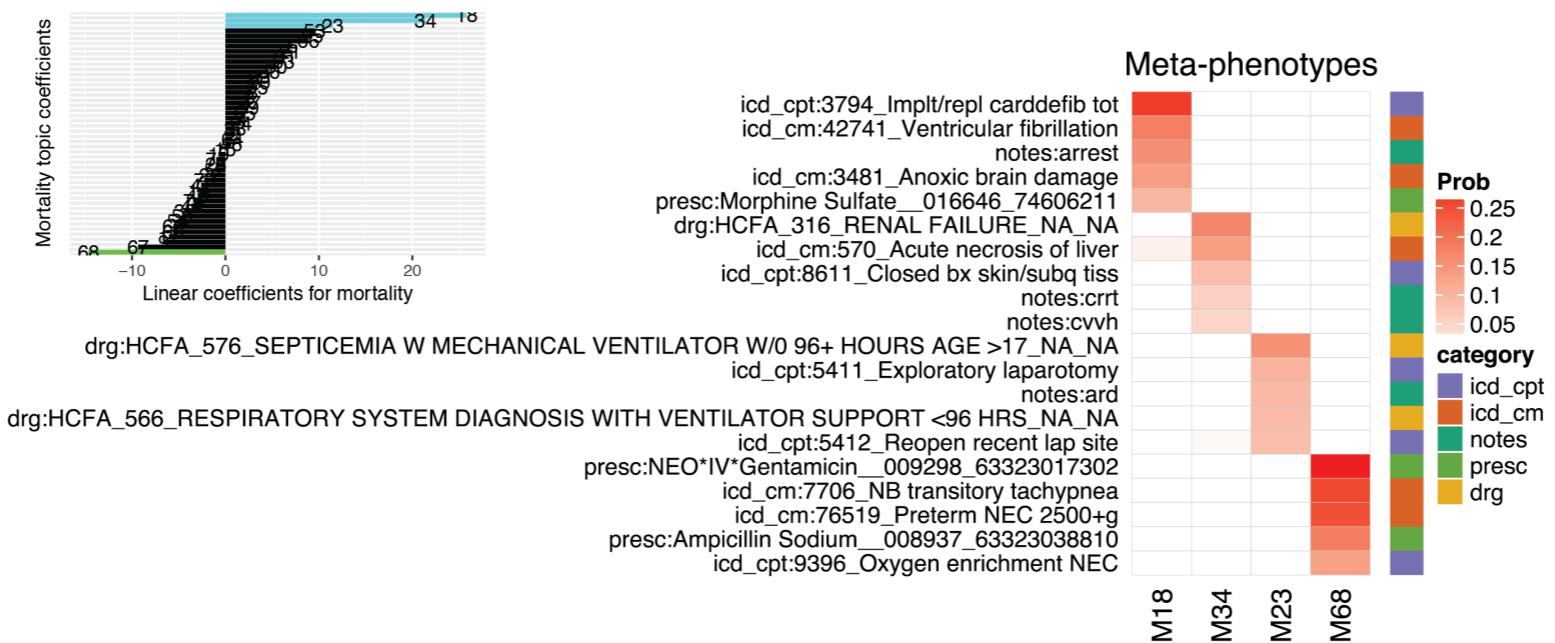


MixEHR accurately predicts future mortality events

a. mortality prediction



b. mortality-related meta-phenotypes



MIT 18.700/18.701/6.047/6.878/HST.507

Computational Biology: Genomes, Networks, Evolution

Lecture 19: EHRs Electronic Health Records

Dealing with Population-Scale EHRs

PheWAS: GWAS for Multi-Phenotypes

Interventions and Reverse Causality

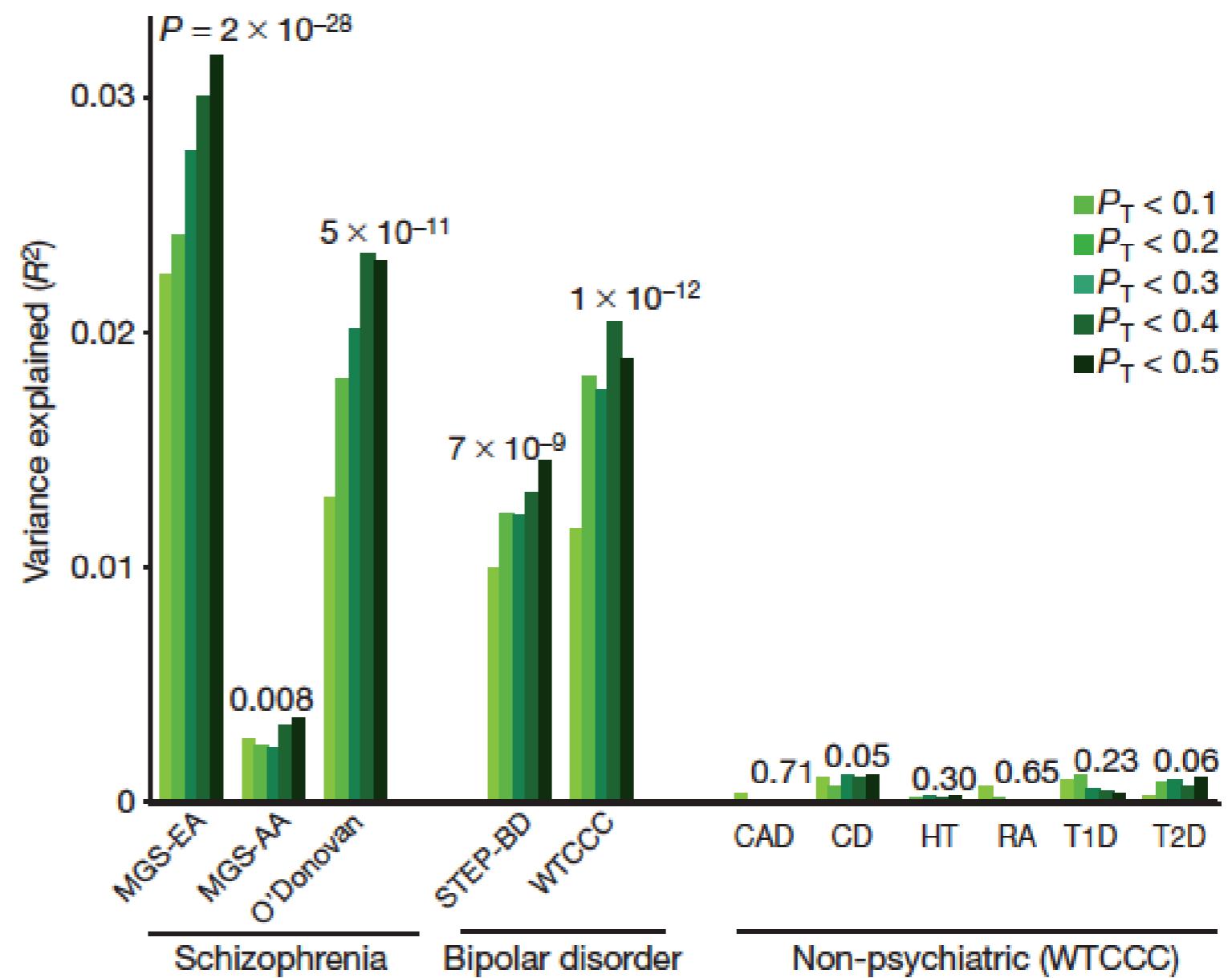
Clinical Notes: LLMs, Text-Data Joint Models

Action, Agents, Reinforcement Learning

AI Agents in Value-Based Healthcare

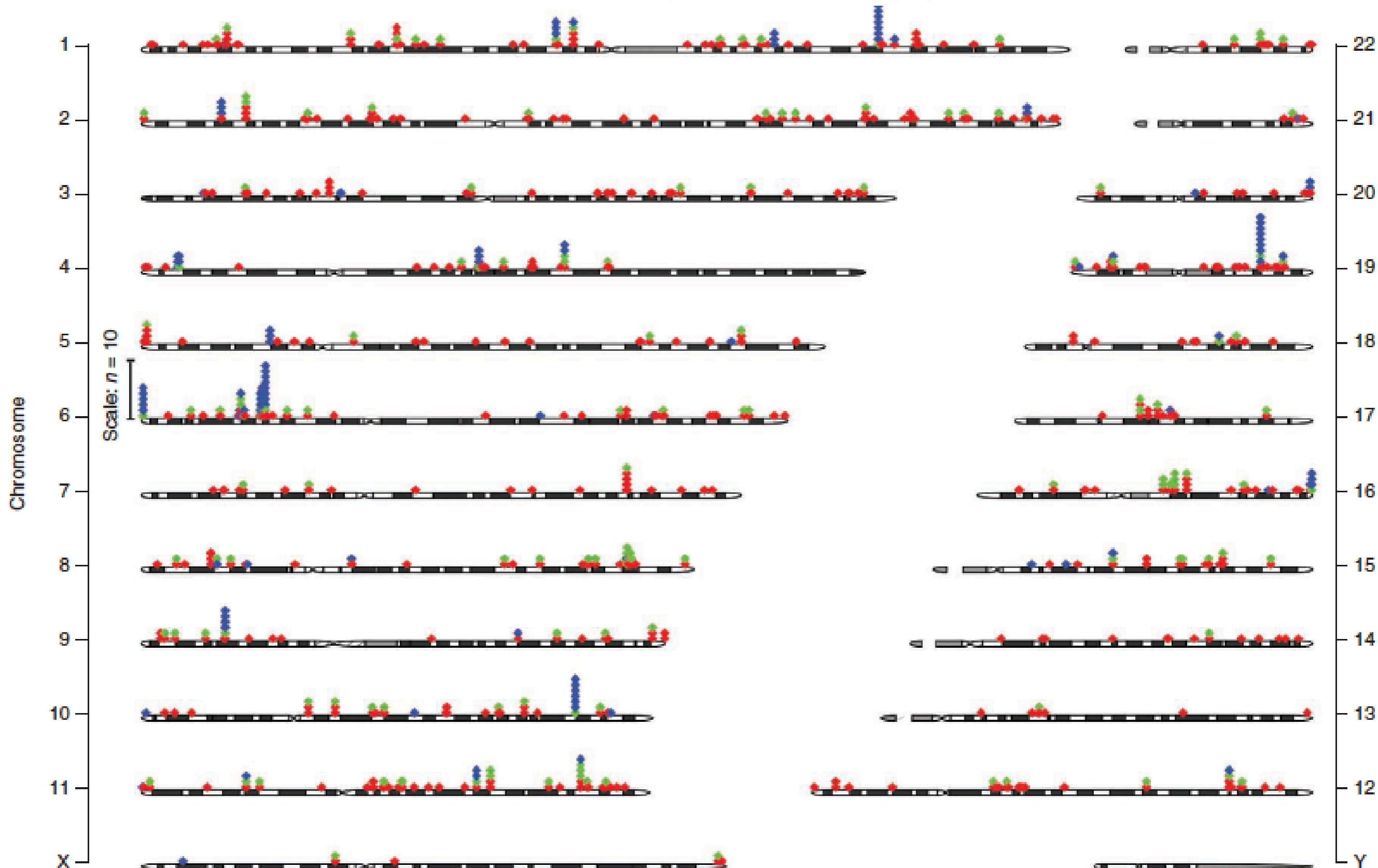
Common polygenic risk variants for schizophrenia and bipolar disorder

- GWAS was performed on International Schizophrenia Consortium (ISC) with ~3.5k cases/controls
- The ISC-derived score was highly associated with disease in the two European schizophrenia samples (MGS-EA, O'Donovan)
- The ISC-derived score alleles were also associated with bipolar disorder in two independent samples.



The International Schizophrenia Consortium (Nat, 2009)

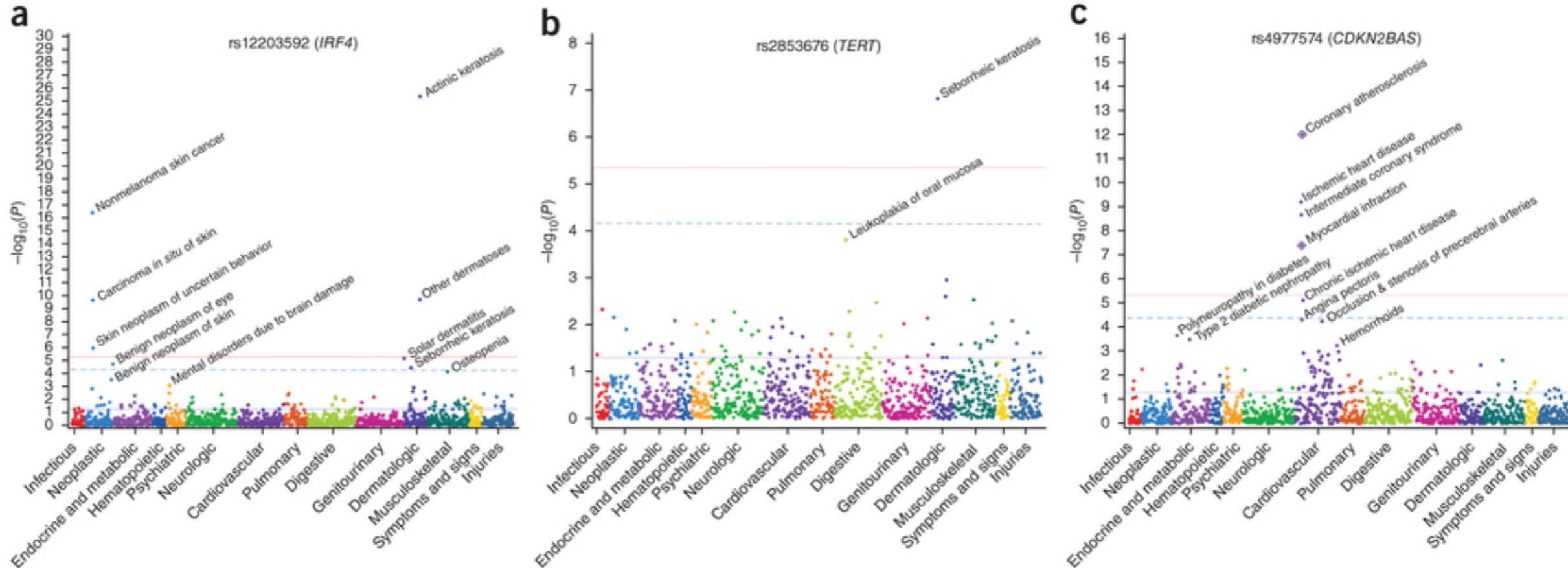
Genome-wide multi-phenotype association



Each diamond represents a unique phenotype association at each SNP.

Denny et al., (NatBiotech, 2013)

Reverse GWAS: PheWAS Manhattan plots



X-axis: Phenotypes instead of SNPs

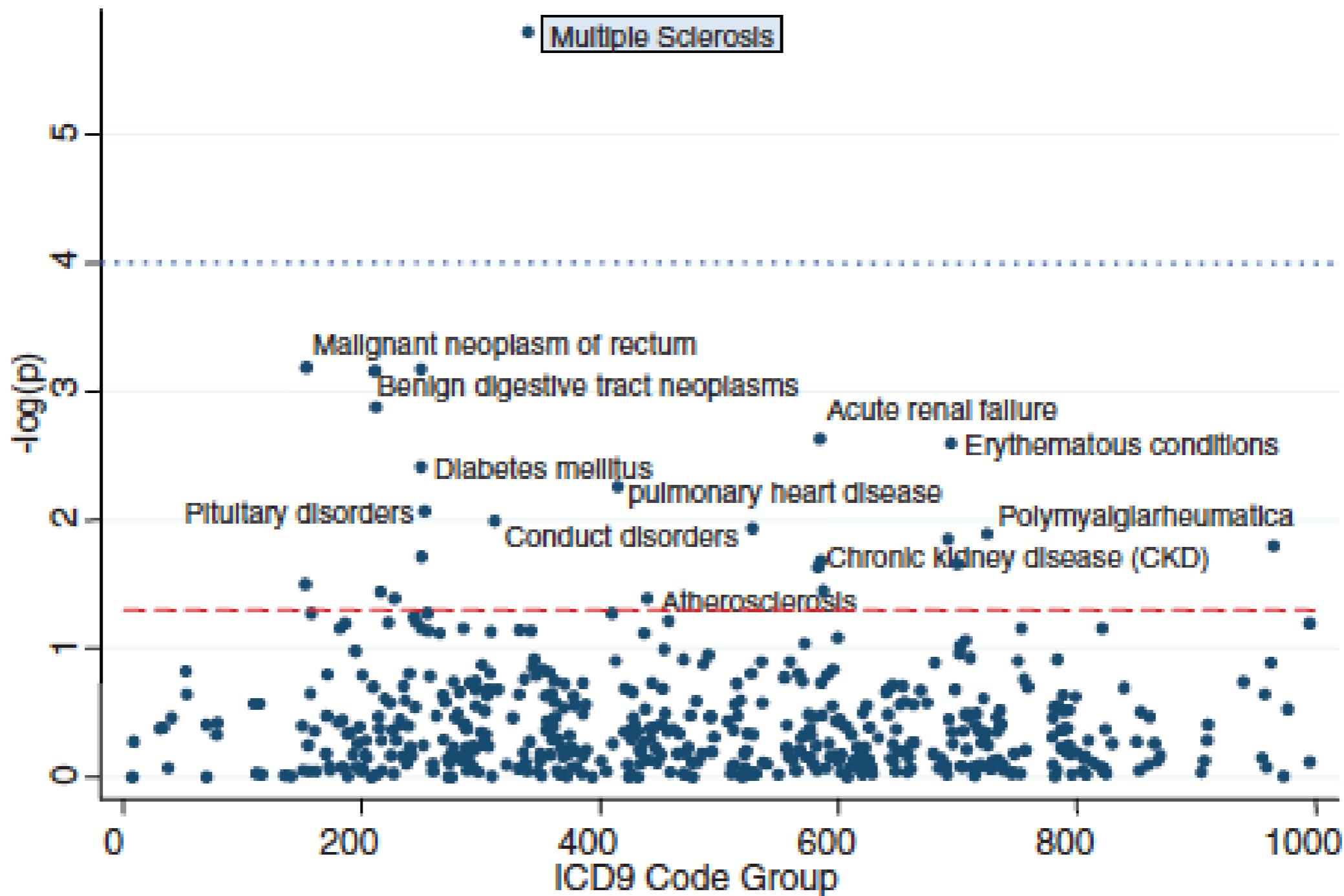
Y-axis: association

Each panel: a SNP

Each point: SNP-phenotype association

Each panel represents 1,358 phenotypes tested for association with a particular SNP.

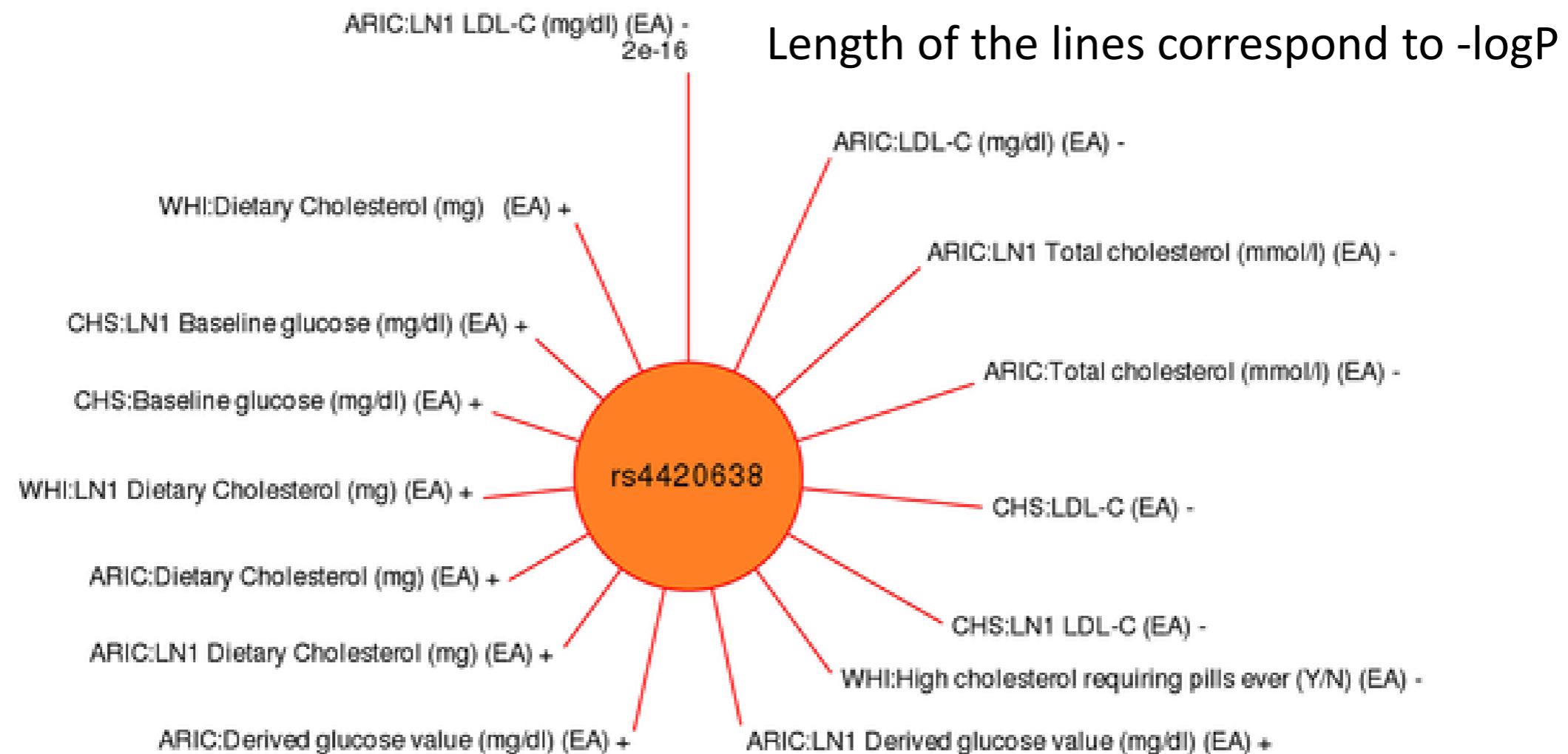
PheWAS scan using EHR ICD9 code



PheWAS using NHANES data (National Health and Nutrition Examination Survey)

One “stone” (SNP) many “birds” (traits)

APOE/APOC1/C1P1/C2/C4, rs4420638, Coded Allele A



rs4420638 near APOC1 associated with LDL

Pendergrass et al., (PlosGen, 2013)

Leverage multiple traits in GWAS

- Modelling multiple traits with multivariate linear mixed model:

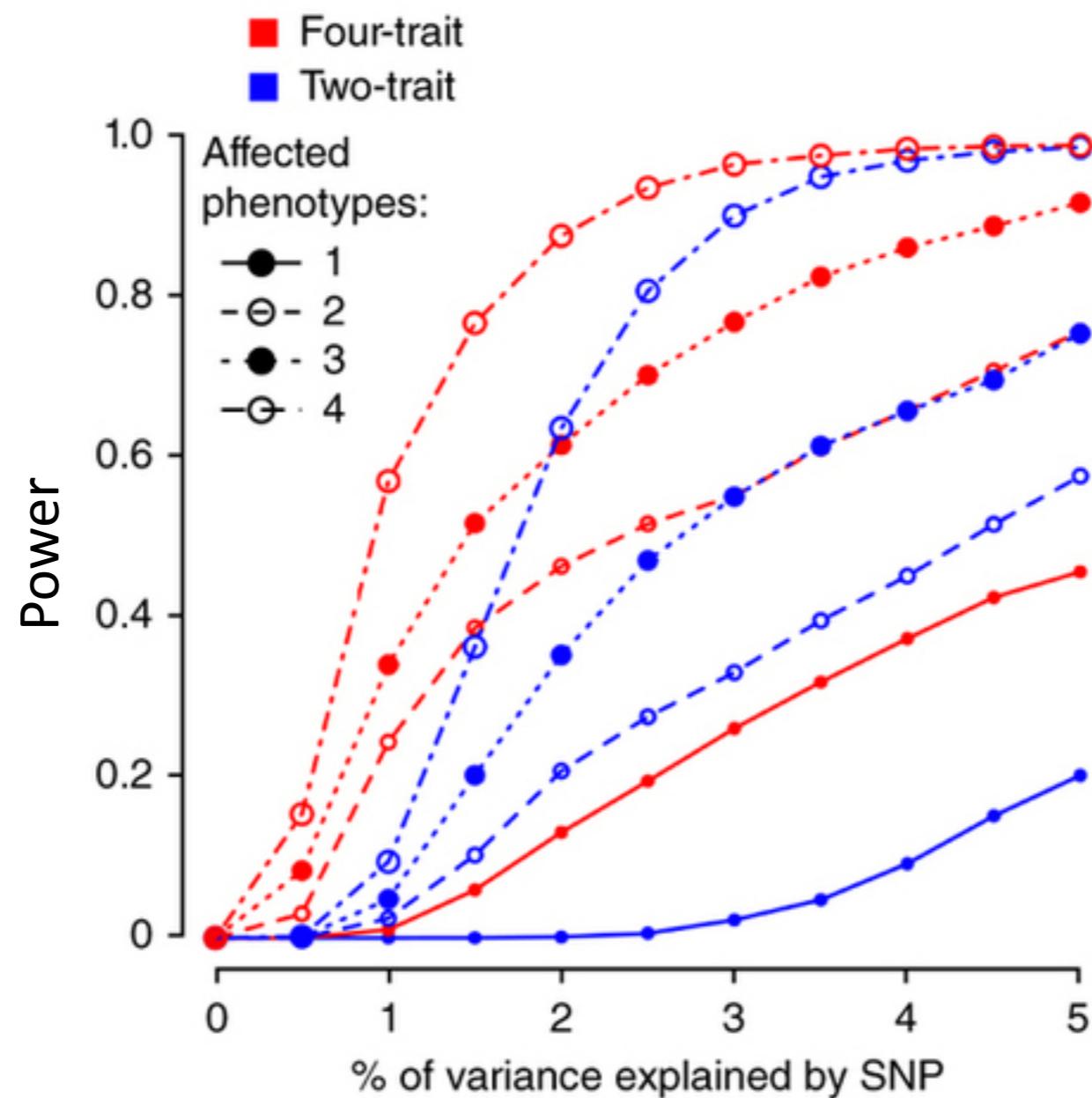
$$Y = BX + G + E$$

$$G \sim MN(0, V_g, D_k)$$

$$E \sim MN(0, V_e, I_{n \times n})$$

- Y: D x N matrix of N individuals and D traits
- B: D x M effect size matrix for M SNPs
- X: M x N genotype matrix
- G: D x N random effect matrix and follows matrix normal (MN)
- V_g : D x D symmetric matrix of genetic variance component
- V_e : D x D symmetric matrix of environmental variance component
- D_k : N x N diagonal matrix filled with eigen values of the kinship matrix

Power gain in modeling multiple phenotypes



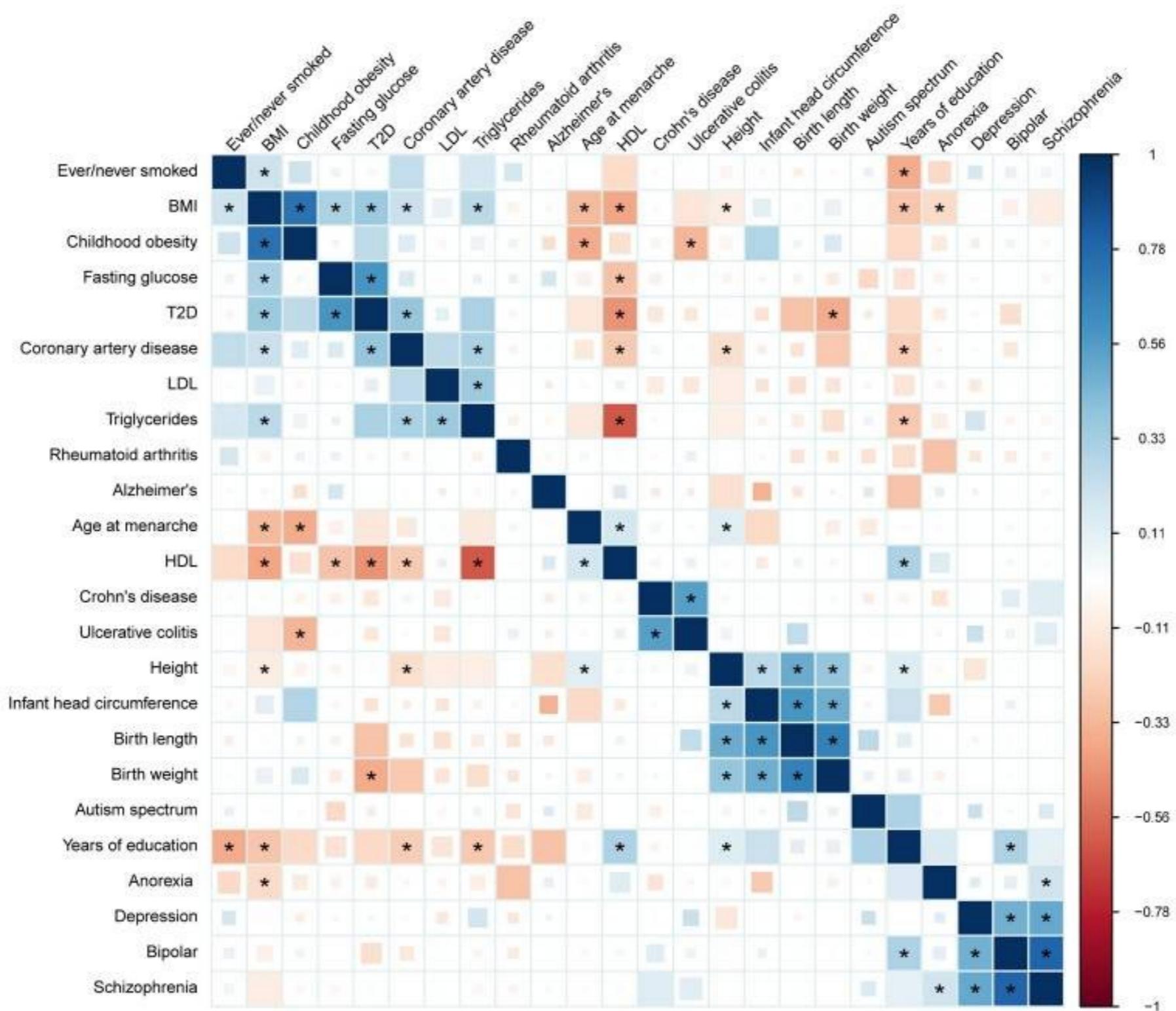
Genetic correlation w/ LD score regression

$$E[z_{1j}z_{2j}l_j] = \frac{\sqrt{N_1 N_2} \rho_g}{M} l_j + \frac{\rho N_s}{\sqrt{N_1 N_2}}$$

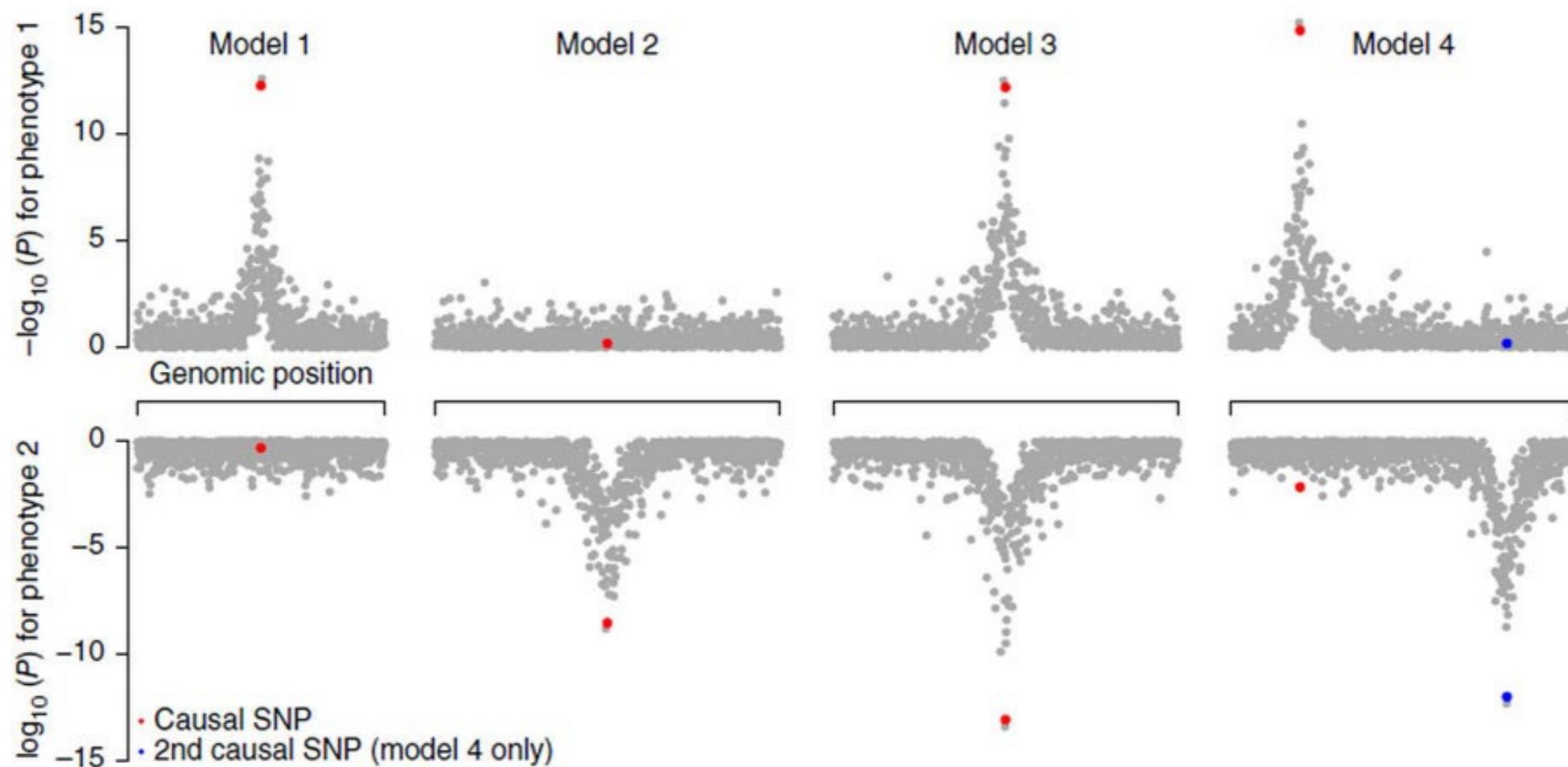
$l_j = \sum_k r_{jk}^2$ is the LD score of SNP j

- N_i is the sample size for study i
- ρ_g is the genetic covariance, which can be efficiently estimated by fitting linear regression
- $l_j = \sum_k r_{jk}^2$ is the LD score of SNP j
- N_S is the number of individuals included in both studies
- ρ is the phenoeticic correlation among the N_S overlapping samples

LDSC-estimated genetic corr.



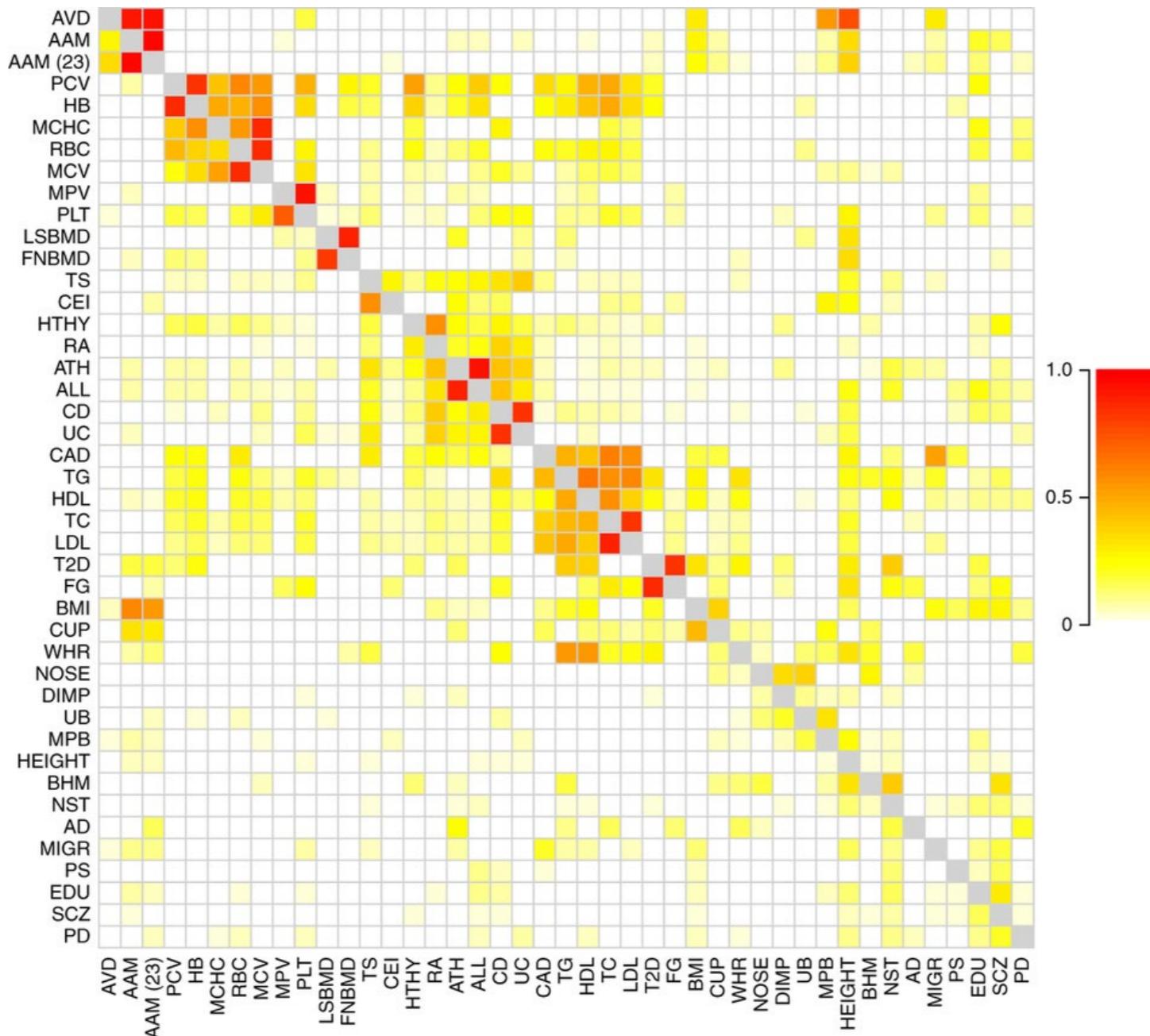
Detecting shared genetic influences by regional Bayes factor



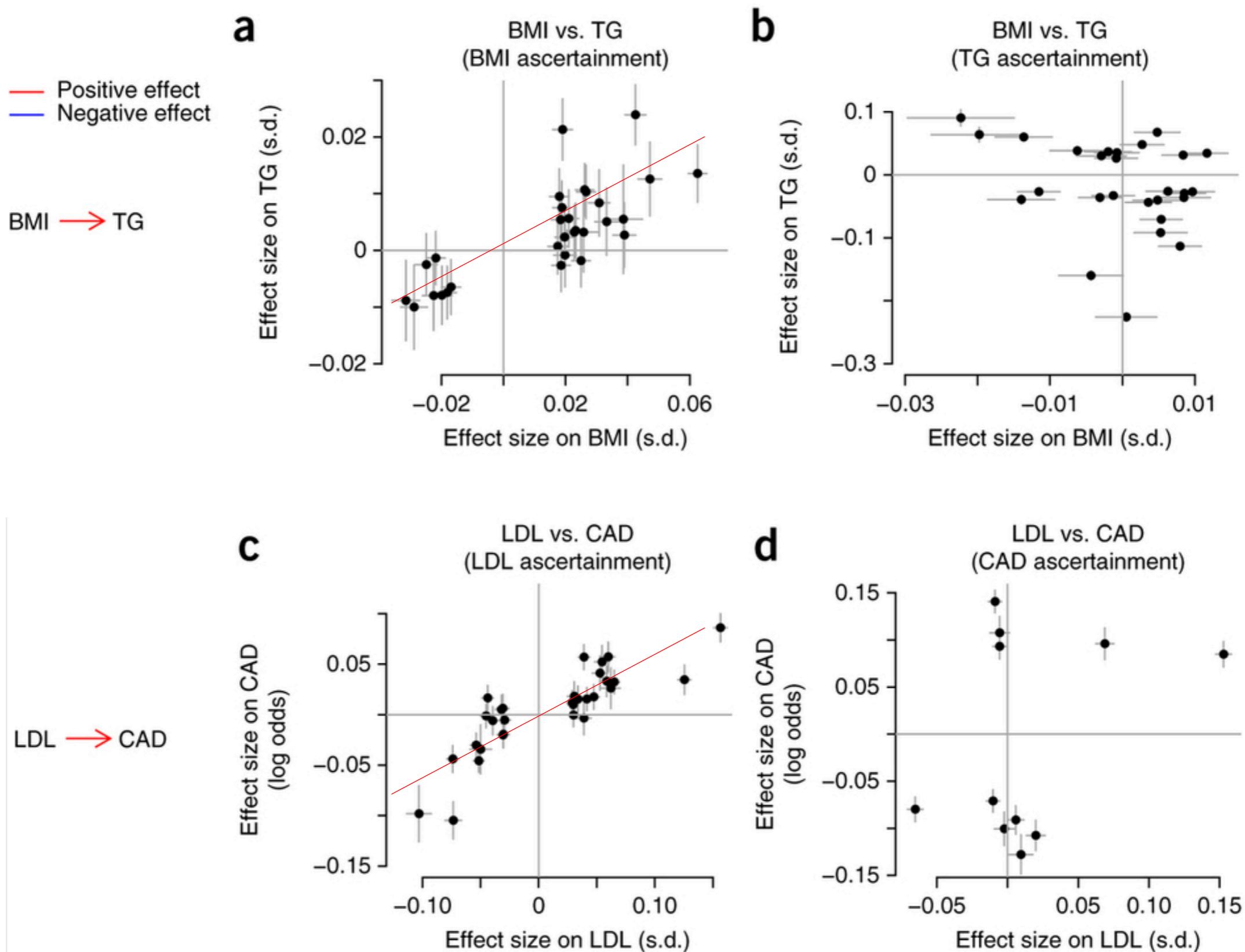
$$l(\Theta|D) = \sum_{i=1}^M \ln \left(\Pi_0 + \sum_{j=1}^4 \pi_j \text{RBF}_i^{(j)} \right)$$

Pickrell et al., (NatGen 2016)

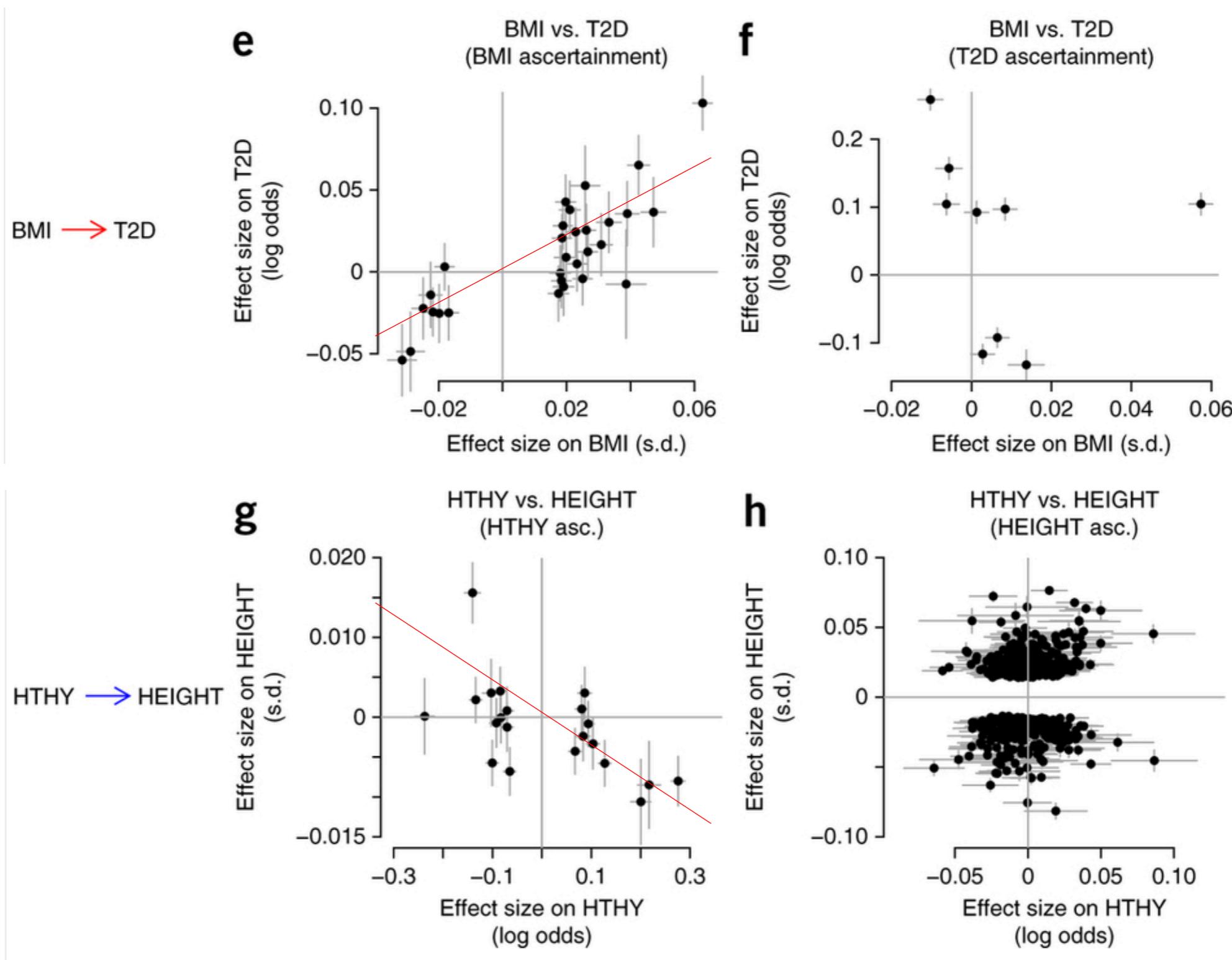
Proportion of shared risk variants among 42 traits



Causal inference



Causal inference



Probabilistically inferring pleiotropic variants

$\pi_{00} = \Pr(Z_{j00}=1) : (P_{j1}|Z_{j00}=1) \sim \mathcal{U}[0,1],$

$(P_{j2}|Z_{j00}=1) \sim \mathcal{U}[0,1],$

$\pi_{10} = \Pr(Z_{j10}=1) : (P_{j1}|Z_{j10}=1) \sim Beta(\alpha_1, 1),$

$(P_{j2}|Z_{j10}=1) \sim \mathcal{U}[0,1],$

$\pi_{01} = \Pr(Z_{j01}=1) : (P_{j1}|Z_{j01}=1) \sim \mathcal{U}[0,1],$

$(P_{j2}|Z_{j01}=1) \sim Beta(\alpha_2, 1),$

$\pi_{11} = \Pr(Z_{j11}=1) : (P_{j1}|Z_{j11}=1) \sim Beta(\alpha_1, 1),$

$(P_{j2}|Z_{j11}=1) \sim Beta(\alpha_2, 1),$

Need to enumerate all possibilities for any given D phenotypes

Chung et al., (PlosGen 2014)

MIT 18.700/18.701/6.047/6.878/HST.507

Computational Biology: Genomes, Networks, Evolution

Lecture 19: EHRs

Electronic Health Records

Dealing with Population-Scale EHRs

PheWAS: GWAS for Multi-Phenotypes

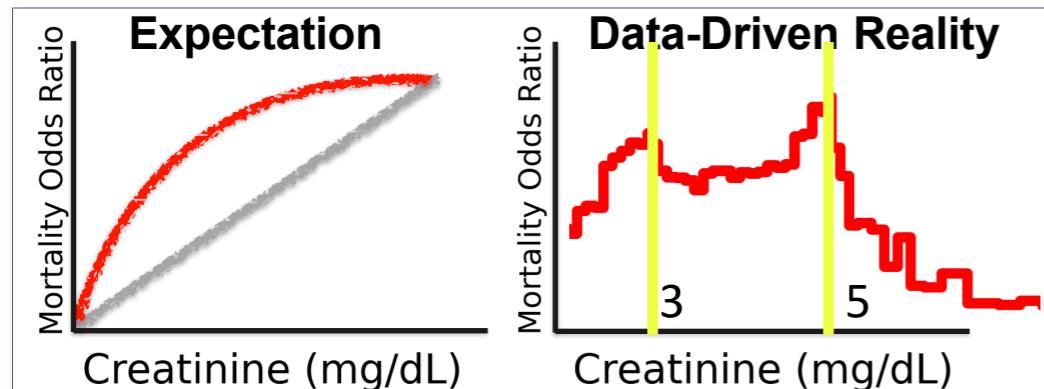
Interventions and Reverse Causality

Clinical Notes: LLMs, Text-Data Joint Models

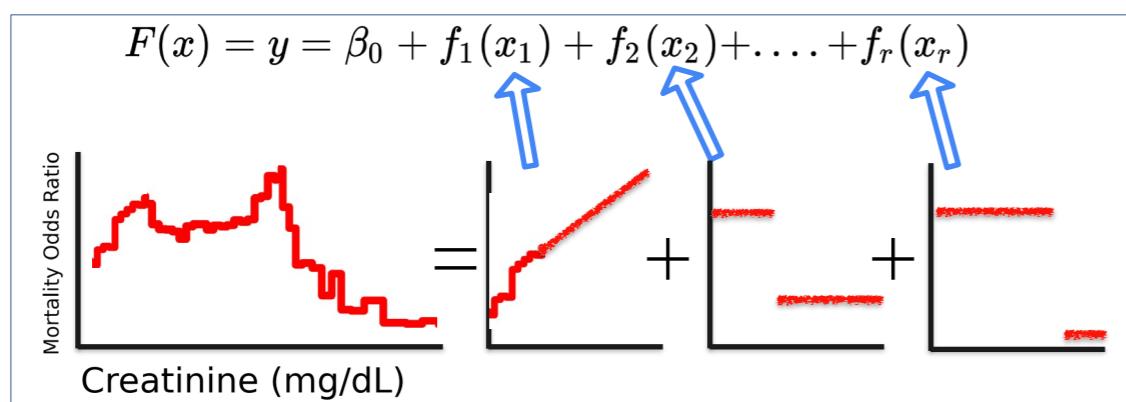
Action, Agents, Reinforcement Learning

AI Agents in Value-Based Healthcare

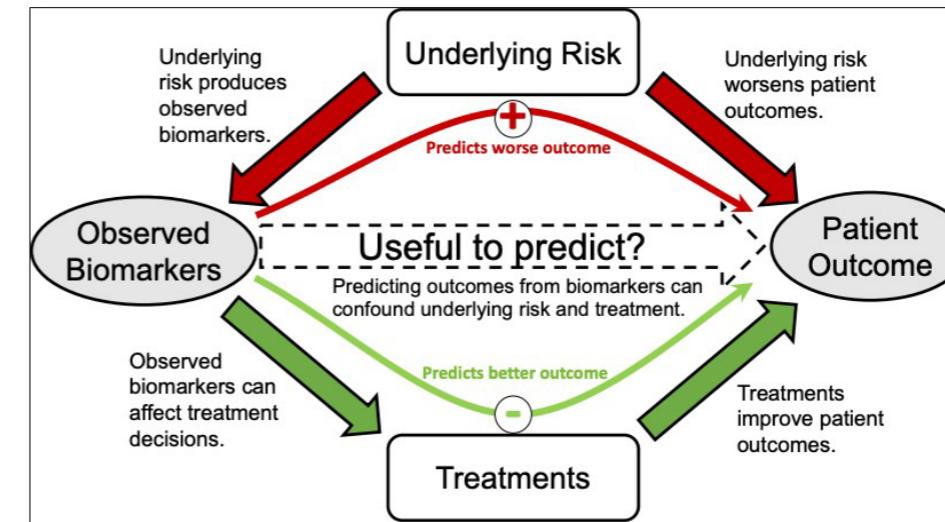
Death by Round Numbers and Sharp Thresholds: Challenges of Real-World EHRs



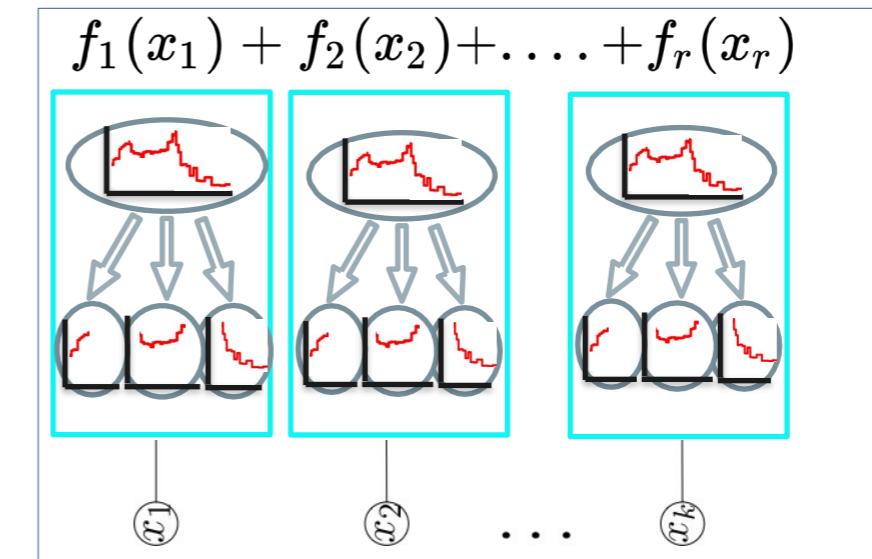
- Elevated creatinine → indicator of renal failure, so expect mortality risk to increase w/creatinine
- In practice, much more complex curve: many discontinuities, sharp thresholds, round numbers
- Reason: round-number thresholds trigger treatment interventions that lower risk
- Naïve AI model: would recommend increasing creatinine levels to reduce mortality risk → Death!



- Generalized Additive Models (GAM): Flexible Glass-Box Models
- Decompose each outcome into its drivers



- Goodheart Law: "When a measure becomes a target, it ceases to be a good measure"
- EHR: "When a biomarker leads to treatment, it ceases to be a good biomarker"



- GAM with internal Boosted Trees
- Uncovers sharp thresholds



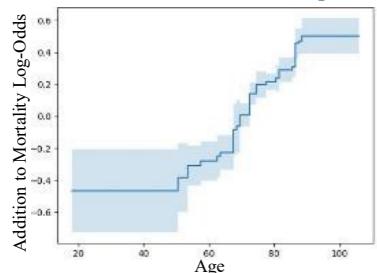
Lengerich
medRxiv 2022



Ben Lengerich

Automatically Interpret EHR results: Glass-box-models + LLMs → Surprise finder

A. Univariate Graphs can be encoded as text...



This graph represents a continuous-valued feature. The keys are intervals that represent ranges where the function predicts the same value.

Feature Name: Age
Feature Type: continuous
Means: {"(18.0, 50.5)": -0.467, ..., "(88.5, 106.0)": 0.503}
Lower Bounds: {"(18.0, 50.5)": -0.727, ..., "(88.5, 106.0)": 0.392}
Upper Bounds: {"(18.0, 50.5)": -0.207, ..., "(88.5, 106.0)": 0.614}



B. and understood by LLMs...

GPT-4: "[...] The general pattern of the graph indicates that as the age increases, the log-odds contribution also increases, meaning that older age is associated with a higher risk of death due to pneumonia.

In the age range from about 18 to 67.5 years, the contribution to the log-odds is negative, which means the model predicts a lower risk of death from pneumonia in this age group. Starting from around 69.5 years, the contribution becomes positive, indicating a higher risk of death from pneumonia as the patient's age increases.

It's important to note that the exact turning point where the risk increases can vary, as indicated by the lower and upper bounds for each age range. These bounds represent the 95% confidence interval, which is a measure of the uncertainty around the estimate. However, the overall trend remains the same: older age is associated with higher risk of death from pneumonia.

In conclusion, age is a significant predictor of pneumonia-related mortality according to this model, with older patients generally at higher risk."



C. to enable conversational data science.

Summarization

Q: Does the risk increase monotonically with age?

Application

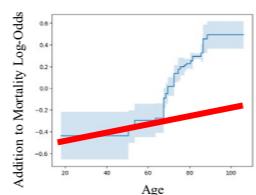
Q: What would be the average risk of an 82-year old patient?

Simplification

Q: What is the best trendline? Express your answer as a linear equation of the form $y=mx+b$.

GPT-4: "Yes, according to the provided data, the risk of death from pneumonia increases monotonically with age. [...]"

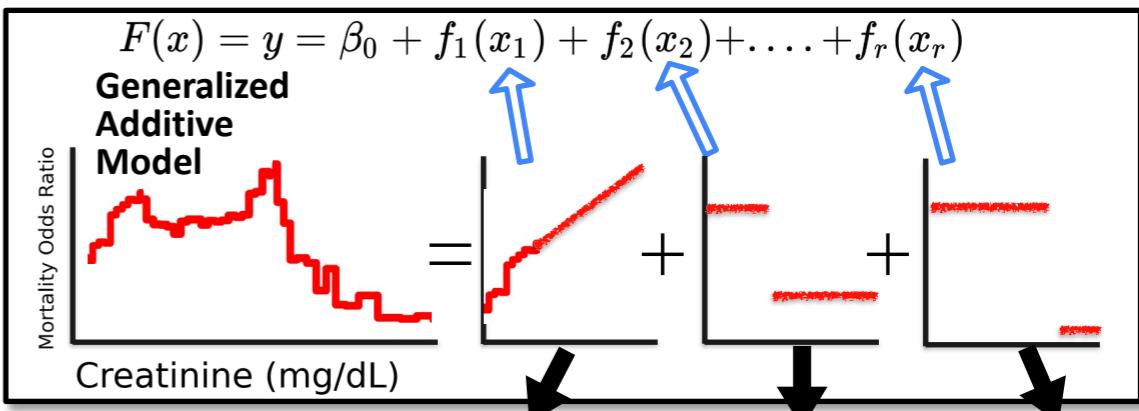
GPT-4: "[...] according to the model, an 82-year old patient would have a 57.2% probability of dying from pneumonia on average. However, remember this is a model-based prediction and actual risk could be different based on other patient-specific factors not included in this model."



GPT-4: "[...] Therefore, the best trendline (linear equation) for The graph is:

$$y \approx -0.558006 + 0.003896x$$

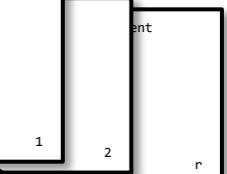
- Multivariate Glass-box models (GAMs) and LLMs → hierarchical reasoning through separable component functions



Component functions can be encoded independently...

This graph represents a continuous-valued feature. The keys are intervals that represent ranges where the function predicts the same value.

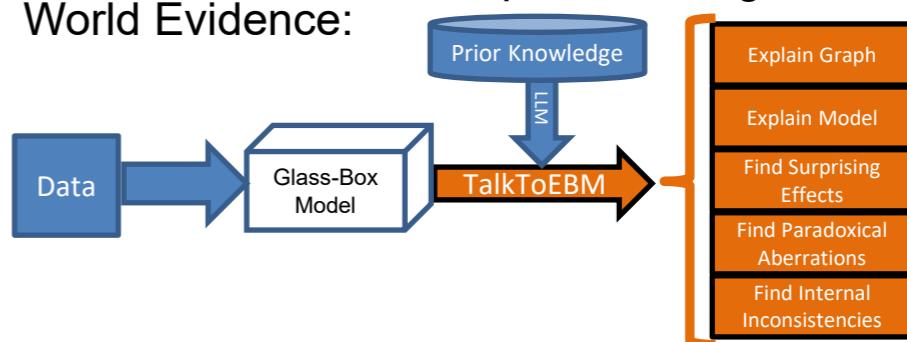
Feature Name: Age
Feature Type: continuous
Means: {"(18.0, 50.5)": -0.467, ..., "(88.5, 106.0)": 0.503}
Lower Bounds: {"(18.0, 50.5)": -0.727, ..., "(88.5, 106.0)": 0.392}
Upper Bounds: {"(18.0, 50.5)": -0.207, ..., "(88.5, 106.0)": 0.614}



and analyzed jointly by LLMs

GPT-4: In summary, the model provides an intricate and nuanced prediction of the likelihood of death from pneumonia based on a wide range of patient features. However, given the complexity of the relationships and the uncertainty in some predictions, the model's output should be interpreted with caution and used in conjunction with clinical judgement.

- Enables Automated Surprise Finding in Real-World Evidence:



Lengerich
Arxiv 2023



Ben Lengerich

MIT 18.700/18.701/6.047/6.878/HST.507

Computational Biology: Genomes, Networks, Evolution

Lecture 19: EHRs

Electronic Health Records

Dealing with Population-Scale EHRs

PheWAS: GWAS for Multi-Phenotypes

Interventions and Reverse Causality

Clinical Notes: LLMs, Text-Data Joint Models

Action, Agents, Reinforcement Learning

AI Agents in Value-Based Healthcare

NLP: Natural Language Processing

Making Sense of Language

What do We Want from Classical Natural Language Processing?

- Extract structured data from narrative text (e.g. billing codes from notes; symptoms from narratives, plans from doctors' and nurses' notes)
- Narrative text from structured data \Leftrightarrow Reports from non-narratives (imaging, signals, ...)
- Combine text models with other data → Cohort selection, outcome prediction
- Summarize vast numbers of notes to what's important. (For what use cases?)
- Question answering
- for any word or phrase, assign it a meaning (or null) from some taxonomy/ontology/terminology;
 - e.g., “rheumatoid arthritis” ==> 714.0 (ICD9)
- for any word or phrase, determine whether it represents protected health information;
 - e.g., “Mr. Huntington suffers from Huntington’s Disease”
 - determine aspects of each entity: time, location, certainty, ...
- having identified two meaningful phrases in a sentence, determine the relationship (or null) between them;
 - e.g., precedes, causes, treats, prevents, indicates, ...
 - note: we also need a taxonomy of relationships
- in a larger document, identify the sentences or fragments most relevant to answering a specific medical question;
 - e.g., where is the patient’s exercise regimen discussed?
- summarization
 - as data sets balloon in size, how to provide a meaningful overview
- a step in further learning models
 - topic modeling
 - vector-space embedding

Bulk of Valuable Data are in Narrative Text

orange=demographics
blue=patient condition, diseases, etc.
brown=procedures, tests
magenta=results of measurements
purple=time

Mr. Blind is a 79-year-old white male with a history of diabetes mellitus, inferior myocardial infarction, who underwent open repair of his increased diverticulum November 13th at Sephsandpot Center.

The patient developed hematemesis November 15th and was intubated for respiratory distress. He was transferred to the Valtawnprinceel Community Memorial Hospital for endoscopy and esophagoscopy on the 16th of November which showed a 2 cm linear tear of the esophagus at 30 to 32 cm. The patient's hematocrit was stable and he was given no further intervention.

The patient attempted a gastrografin swallow on the 21st, but was unable to cooperate with probable aspiration. The patient also had been receiving generous intravenous hydration during the period for which he was NPO for his esophageal tear and intravenous Lasix for a question of pulmonary congestion.

On the morning of the 22nd the patient developed tachypnea with a chest X-ray showing a question of congestive heart failure. A medical consult was obtained at the Valtawnprinceel Community Memorial Hospital. The patient was given intravenous Lasix.

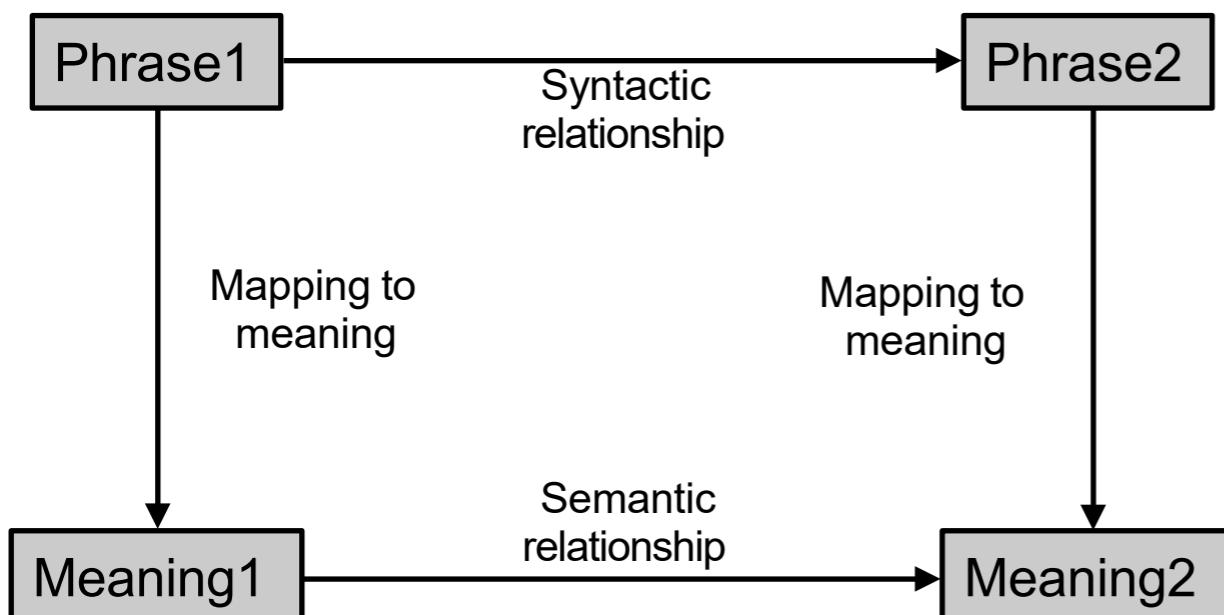
NLP

- Value of the data in clinical text
- **Hyper-simplified linguistics**
- Term spotting + handling negation, uncertainty
- ML to expand terms
- pre-NN ML to identify entities and relations
- language models
- vector space embeddings based on co-occurrence
- adding context to help with disambiguation
- from embedding single words to phrases, sentences, etc.

Long-standing interest in grammar + syntax → Semantics

- Grammar defined by context-sensitive production rules + transformations
- Semantics defined by mappings:
 - Each grammar rule matches a semantic function
 - Terminal symbols are *referents* or *functions*
 - An environment is (in modern terms) a semantic network of complex interrelationships
 - Meaning is compositional, in terms of the semantic functions
- Big remaining question: how to represent meaning in the “real world”?

Grammar	Lexicon
$S \rightarrow NP VP$	[.80] $Det \rightarrow that [.10] \mid a [.30] \mid the [.60]$
$S \rightarrow Aux NP VP$	[.15] $Noun \rightarrow book [.10] \mid flight [.30]$
$S \rightarrow VP$	[.05] $\mid meal [.15] \mid money [.05]$
$NP \rightarrow Pronoun$	[.35] $\mid flights [.40] \mid dinner [.10]$
$NP \rightarrow Proper-Noun$	[.30] $Verb \rightarrow book [.30] \mid include [.30]$
$NP \rightarrow Det Nominal$	[.20] $\mid prefer; [.40]$
$NP \rightarrow Nominal$	[.15] $Pronoun \rightarrow I [.40] \mid she [.05]$
$Nominal \rightarrow Noun$	[.75] $\mid me [.15] \mid you [.40]$
$Nominal \rightarrow Nominal Noun$	[.20] $Proper-Noun \rightarrow Houston [.60]$
$Nominal \rightarrow Nominal PP$	[.05] $\mid NWA [.40]$
$VP \rightarrow Verb$	[.35] $Aux \rightarrow does [.60] \mid can [40]$
$VP \rightarrow Verb NP$	[.20] $Preposition \rightarrow from [.30] \mid to [.30]$
$VP \rightarrow Verb NP PP$	[.10] $\mid on [.20] \mid near [.15]$
$VP \rightarrow Verb PP$	[.15] $\mid through [.05]$
$VP \rightarrow Verb NP NP$	[.05]
$VP \rightarrow VP PP$	[.15]
$PP \rightarrow Preposition NP$	[1.0]



Frederick B. Thompson, “English for the Computer.”
Proceedings of the Fall Joint Computer Conference
 (1966) pp. 349-356

- (1) “the” and “anterior” modify incision
- (2) “the” modifies “crease”
- (3) “the” and “chest” modify “wall”
- (4) “the”, “previous”, and “radical” modify “mastectomy”
- (5) “made” is dependent upon the main verb “was”
- (6) “along the crease” modifies “made” adverbially
- (7) “against the chest wall” modifies “made” adverbially
- (8) “following the previous radical mastectomy” modifies “made” adverbially

JAMA, March 17, 1969 • Vol 207, No 11

Shapiro, P, et al. *Information in Natural Languages: A New Approach*, JAMA 1969

NLP

- Value of the data in clinical text
- Hyper-simplified linguistics
- **Term spotting + handling negation, uncertainty**
- ML to expand terms
- pre-NN ML to identify entities and relations
- language models
- vector space embeddings based on co-occurrence
- adding context to help with disambiguation
- from embedding single words to phrases, sentences, etc.

Term Spotting → Synonyms

- Traditionally, lists of coded items, narrative terms and patterns hand-crafted by researcher
 - E.g., N3C (National Covid Cohort Collaborative)
 - Institutions don't want to share notes, even de-identified
 - Lack of sophistication at most medical centers to run sophisticated tools
 - Instead, run simple term matching algorithms and report just terms
 - Negation and uncertainty handled by somewhat ad-hoc methods
 - NegEx is widely used, ∃ many more sophisticated variants
 - Generalize terms to get better coverage
 - Manually or automatically identify high-certainty “anchors”
 - Learn related terms to augment the set of terms
 - From knowledge bases such as UMLS
 - From co-occurrence in EMR data
 - From co-occurrence in publications

anosmia
(C0003126)

COVID symptoms
• Fever
• Chill
• Cough
• Fatigue
• Nasal obstruction
• Loss of appetite
• Diarrhea
• Abdominal pain
• Nausea
• Vomiting
• Sore throat
• Headache
• Myalgia
• Loss of taste
• Loss of smell
• Dyspnea
• Chest pain
• Delirium
• Hypersomnia
• Cyanosis

Negation / NegEx

- “Identifying pertinent negatives, then, involves identifying a proposition ascribing a clinical condition to a person and determining whether the proposition is denied or negated in the text.”
- Simpler than general problem of negation in NLP because negation applies mostly to noun phrases indicating diseases, tests, drugs, findings, ...
- NegEx

- Find all UMLS terms in each sentence of a discharge summary
 - “The patient denied experiencing chest pain on exertion” ⇒ “The patient denied experiencing S1459038 on exertion”
- Find patterns
 - <negation phrase> *{0,5} <UMLS term>
 - “no signs of”, “ruled out unlikely”, “absence of”, “not demonstrated”, “denies”, “no sign of”, “no evidence of”, “no”, “denied”, “without”, “negative for”, “not”, “doubt”, “versus”
 - <UMLS term> *{0,5} <negation phrase>
 - “declined”, “unlikely”
 - Pseudo-negation: “gram negative”, “no further”, “not able to be”, “not certain if”, “not certain whether”, “not necessarily”, “not rule out”, “without any further”, “without difficulty”, “without further”

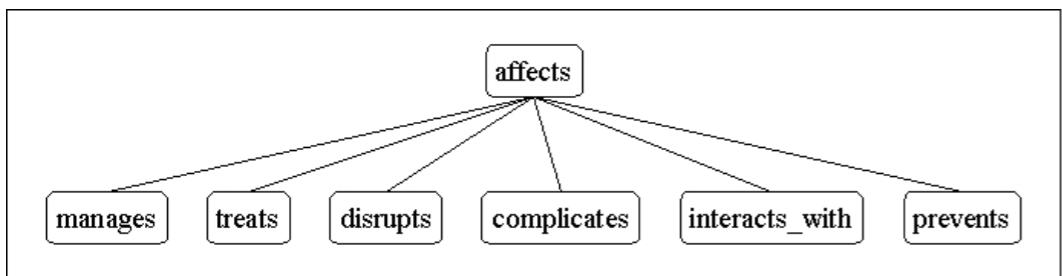
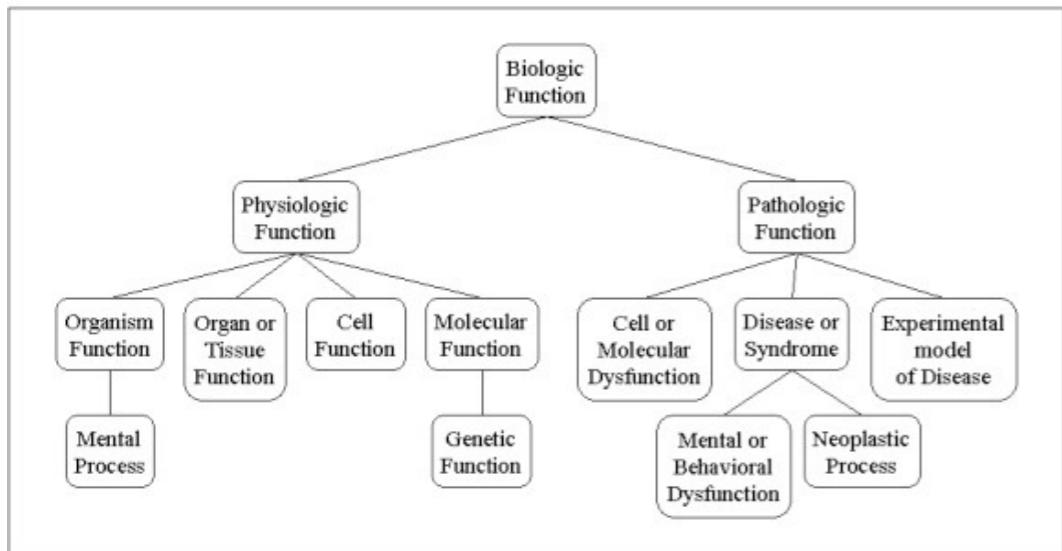
	Baseline			NegEx		
	Group 1 sentences (i.e. containing NegEx negation phrases)	Group 2 sentences (i.e., not containing NegEx negation phrases)	All sentences	Group 1 sentences (i.e. containing NegEx negation phrases)	Group 2 sentences (i.e., not containing NegEx negation phrases)	All sentences
n	500	500	1000	500	500	1000
Sensitivity	88.27	0.00	88.27	82.31	0.00	77.84
Specificity	52.69	100.00	85.27	82.50	100.00	94.51
PPV	68.42	—	68.42	84.49	—	84.49
NPV	79.46	96.99	93.01	80.21	96.99	91.73

- Baseline:
 - <negation phrase> * <UMLS term>
 - “no”, “denies”, “not”, “without”, “*n’t”, “ruled out”, “denied”

NLP

- Value of the data in clinical text
- Hyper-simplified linguistics
- Term spotting + handling negation, uncertainty
- **ML to expand terms**
- pre-NN ML to identify entities and relations
- language models
- vector space embeddings based on co-occurrence
- adding context to help with disambiguation
- from embedding single words to phrases, sentences, etc.

Synonyms: Semantic Networks, Terms, Relations



tui	sty	c
T061	Therapeutic or Preventive Procedure	260914
T033	Finding	233579
T200	Clinical Drug	172069
T109	Organic Chemical	157901
T121	Pharmacologic Substance	124844
T116	Amino Acid, Peptide, or Protein	117508
T009	Invertebrate	111044
T007	Bacterium	110065
T002	Plant	95017
T047	Disease or Syndrome	79370
T023	Body Part, Organ, or Organ Component	73402
T201	Clinical Attribute	60998
T123	Biologically Active Substance	55741
T074	Medical Device	51708
T028	Gene or Genome	49960
T004	Fungus	47291
T060	Diagnostic Procedure	46106
T037	Injury or Poisoning	43924
T191	Neoplastic Process	33539
T044	Molecular Function	31369
T126	Enzyme	25766
T129	Immunologic Factor	25025
T059	Laboratory Procedure	24511
T058	Health Care Activity	19552
T029	Body Location or Region	16470
T013	Fish	16059
T046	Pathologic Function	13562
T184	Sign or Symptom	13299
T130	Indicator, Reagent, or Diagnostic Aid	12809
T170	Intellectual Product	12544
T118	Carbohydrate	10722
T110	Steroid	10363
T012	Bird	9908
T043	Cell Function	9758
	...	

c.cui	c.str
C0000744	Abetalipoproteinemia
C0000774	Gastrin secretion abnormality NOS
C0000786	Spontaneous abortion
C0000809	Abortion, Habitual
C0000814	Missed abortion
C0000821	Threatened abortion
C0000822	Abortion, Tubal
C0000823	Abortion, Veterinary
C0000832	Abortio Placentae
C0000880	Acanthamoeba Keratitis
C0000889	Acanthosis Nigricans
C0001080	Achondroplasia
C0001083	Achromia parasitica
C0001125	Acidosis, Lactic
C0001126	Renal tubular acidosis
C0001127	Acidosis, Respiratory
C0001139	Acinetobacter Infections
C0001142	Acladiosis
C0001144	Acne Vulgaris
C0001145	Acne Keloid
C0001163	Vestibulocochlear Nerve Diseases
C0001168	Complete obstruction
C0001169	Acquired coagulation factor deficiency NOS
C0001175	Acquired Immunodeficiency Syndrome
C0001197	Acrodermatitis
C0001202	Acrokeratosis
C0001206	Acromegaly
C0001207	Hypersomatotropic gigantism
C0001231	ACTH Syndrome, Ectopic
C0001247	Actinobacillosis
	...

The screenshot shows the UMLS Lookup interface. On the left, there's a tree view of UMLS concepts under 'TOP:Entity or Event'. On the right, a clinical note is displayed with various terms highlighted in green, blue, and red, corresponding to their UMLS mappings. For example, 'Admission Date' is mapped to T071:Entity, and 'Date of Birth' is mapped to T033:Finding.

- Unified Medical Language Systems project of NLM; since ~1985
- Metathesaurus now (2022ab version) includes 182 source vocabularies
 - (had 215, but some obsolete ones were dropped)
 - MeSH, SNOMED, ICD-9, ICD-10, LOINC, RxNORM, NCI, CPT, GO, DXPLAIN, OMIM, ...
- Synonym mappings across vocabularies;
 - e.g., “heart attack” = “acute myocardial infarct” = “myocardial infarction” ...
 - 4,662,313 distinct concepts, represented by concept unique identifier (CUI)
- Jumbled compendium of every hierarchy drawn from every source
- Use synonymous terms as well as the starting ones
- Take advantage of others related terms
 - hypo- or hypernyms
 - other associated terms
 - e.g., common symptoms or treatments of a disease
- Recursive ML problem: learn how best to identify cases associated with a term
 - “phenotyping”
- “Anchor & Learn”-like methods learn secondary terms from
 - medical records
 - textbooks and on-line medical resources
 - Pubmed abstracts and articles

Study of 3-letter abbreviations in Medline (Pubmed) abstracts

Expanding acronyms

- Extraction method examples:
 - *minimum alveolar anesthetic concentration* (MAC)
 - *procoagulant activity* (PCA)
 - *carboxymethyllysine* (CML)
 - *n-6-(delta-2-isopentenyl) adenine* (IPA)
 - + MetaMap-based heuristics
- 81.2% were ambiguous, with mean of 16.6 senses
- ignoring rare (<5) occurrences, 64.6% were ambiguous, with 4.91 senses
 - —long tail
- 82.8% of abbreviations with ≥ 100 occurrences are in UMLS
 - but only 23.5% of all abbreviations

NLP

- Value of the data in clinical text
- Hyper-simplified linguistics
- Term spotting + handling negation, uncertainty
- ML to expand terms
- **pre-NN ML to identify entities and relations**
- language models
- vector space embeddings based on co-occurrence
- adding context to help with disambiguation
- from embedding single words to phrases, sentences, etc.

Building Feature-Based Models

- Features of text from which models can be built
 - words, parts of speech, capitalization, punctuation
 - document section, conventional document structures
 - identified patterns and thesaurus terms
 - lexical context
 - all of the above, for n-tuples of words surrounding target
- syntactic context
 - all of the above, for words syntactically related to target
 - E.g., “*The lasix, started yesterday, reduced ascites ...*”

```
+-----Xp-----+
|           +-----Ss-----+
|           +---MXsp---+---XC-----+
+---Wd---+   +-Xd---+---MVpn---+   |   +---Os-----+
|           |   |           |   |   |           |
LEFT-WALL lasix[?].n , started.v-d yesterday , reduced.v-d ascites[?].n .
```

(Output from Link Grammar Parser, w/o special medical dictionary)

Uzuner, Ö., Sibanda, T. C., Luo, Y., & Szolovits, P. (2008). A de-identifier for medical discharge summaries. Artificial Intelligence in Medicine, 42(1), 13–35.
<http://doi.org/10.1016/j.artmed.2007.10.001>

The Importance of Context

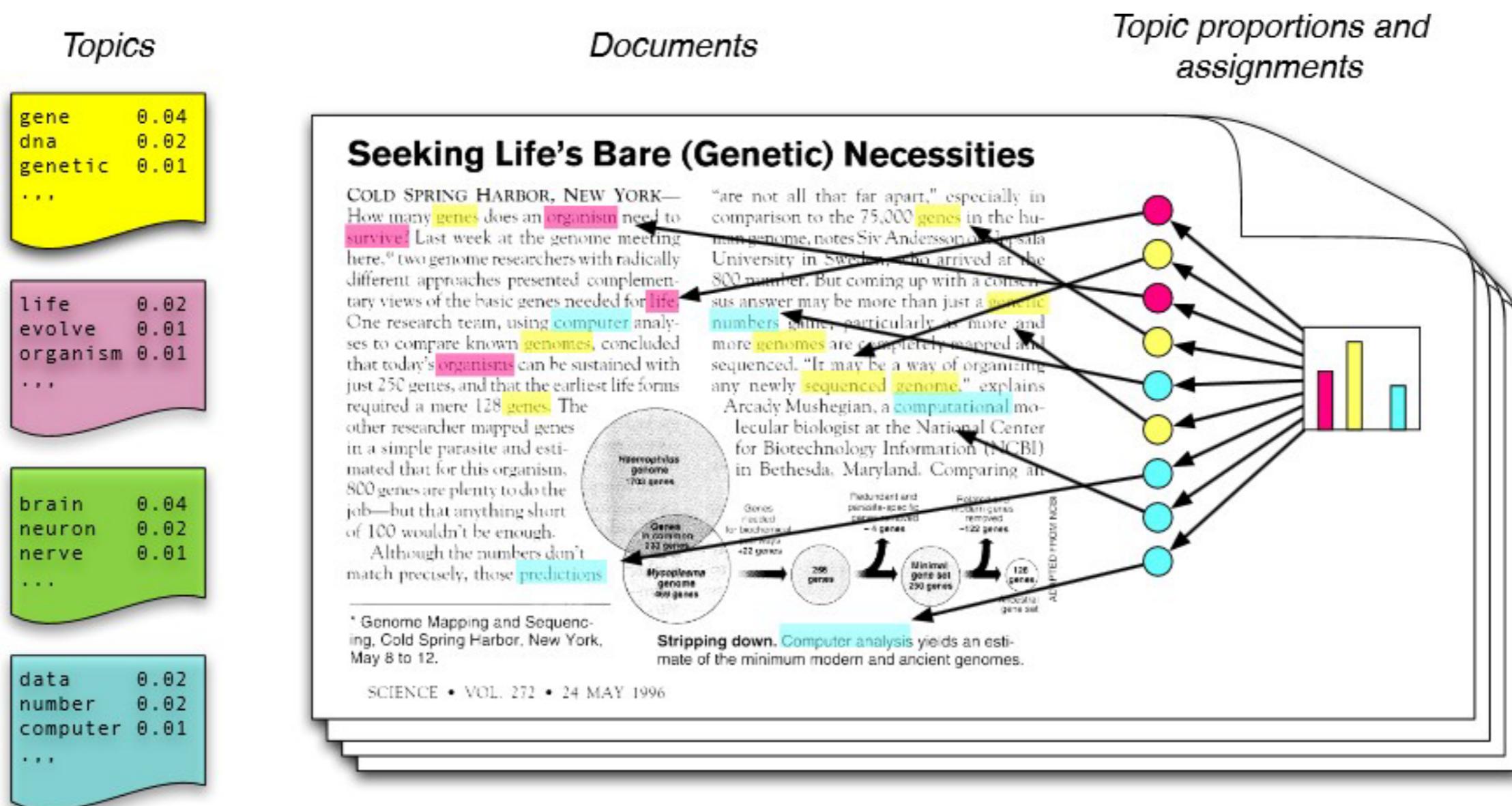
- “Mr. Huntington was treated for Huntington’s Disease at Huntington Hospital, located on Huntington Avenue.”
 - Huntington
 - Huntington’s Disease
 - Mr. Huntington’s Disease
- “Atenolol was administered to Mr. Huntington.”
 - vs. “Atenolol was considered for control of heart rate.”
 - vs. “Atenolol was ineffective and therefore discontinued.”

NLP

- Value of the data in clinical text
- Hyper-simplified linguistics
- Term spotting + handling negation, uncertainty
- ML to expand terms
- pre-NN ML to identify entities and relations
- **Latent Dirichlet Allocation (LDA)**
- language models
- vector space embeddings based on co-occurrence
- adding context to help with disambiguation
- from embedding single words to phrases, sentences, etc.

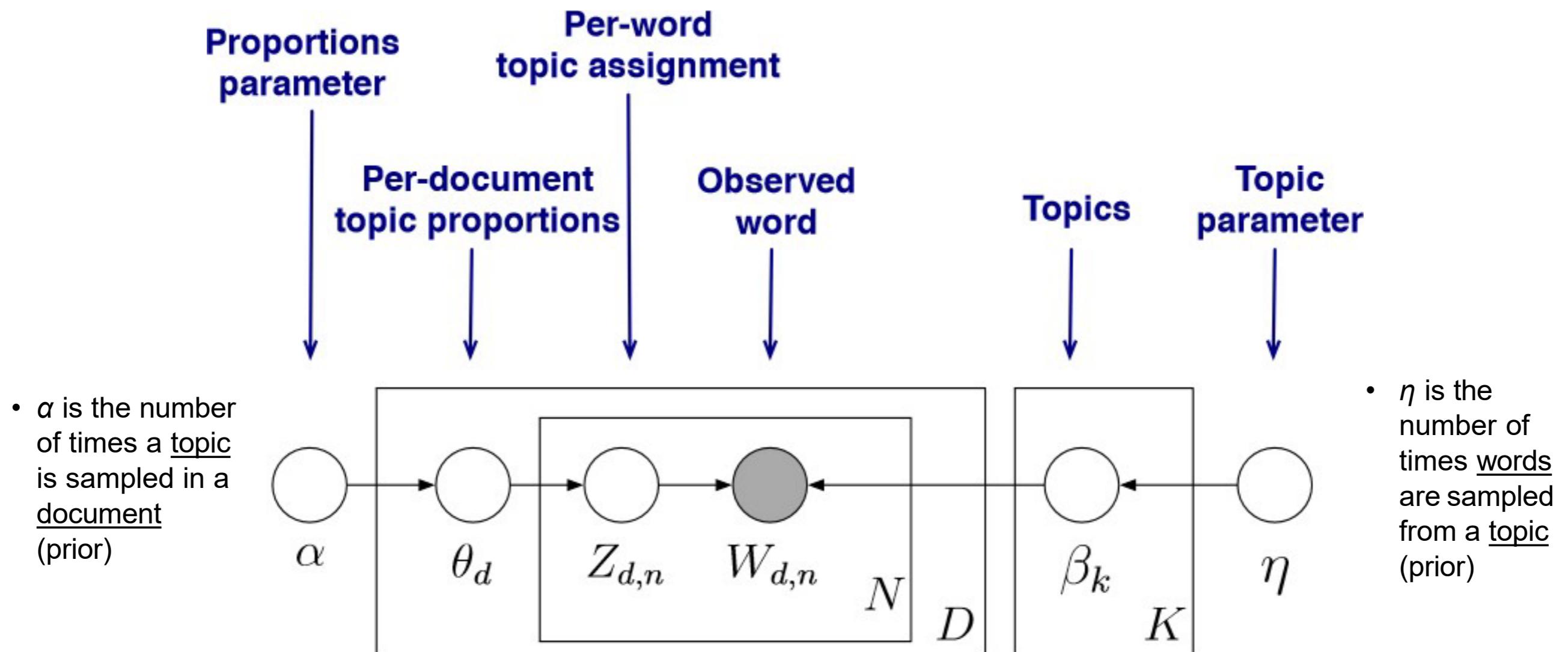
LDA (Latent Dirichlet Allocation): Documents \leftrightarrow Topics

- Every document is a mixture of topics
- Every topic is a distribution over words
- Every word is a draw from a topic



LDA – Latent Dirichlet Allocation

- We observe words, we infer everything else, with our assumed structure

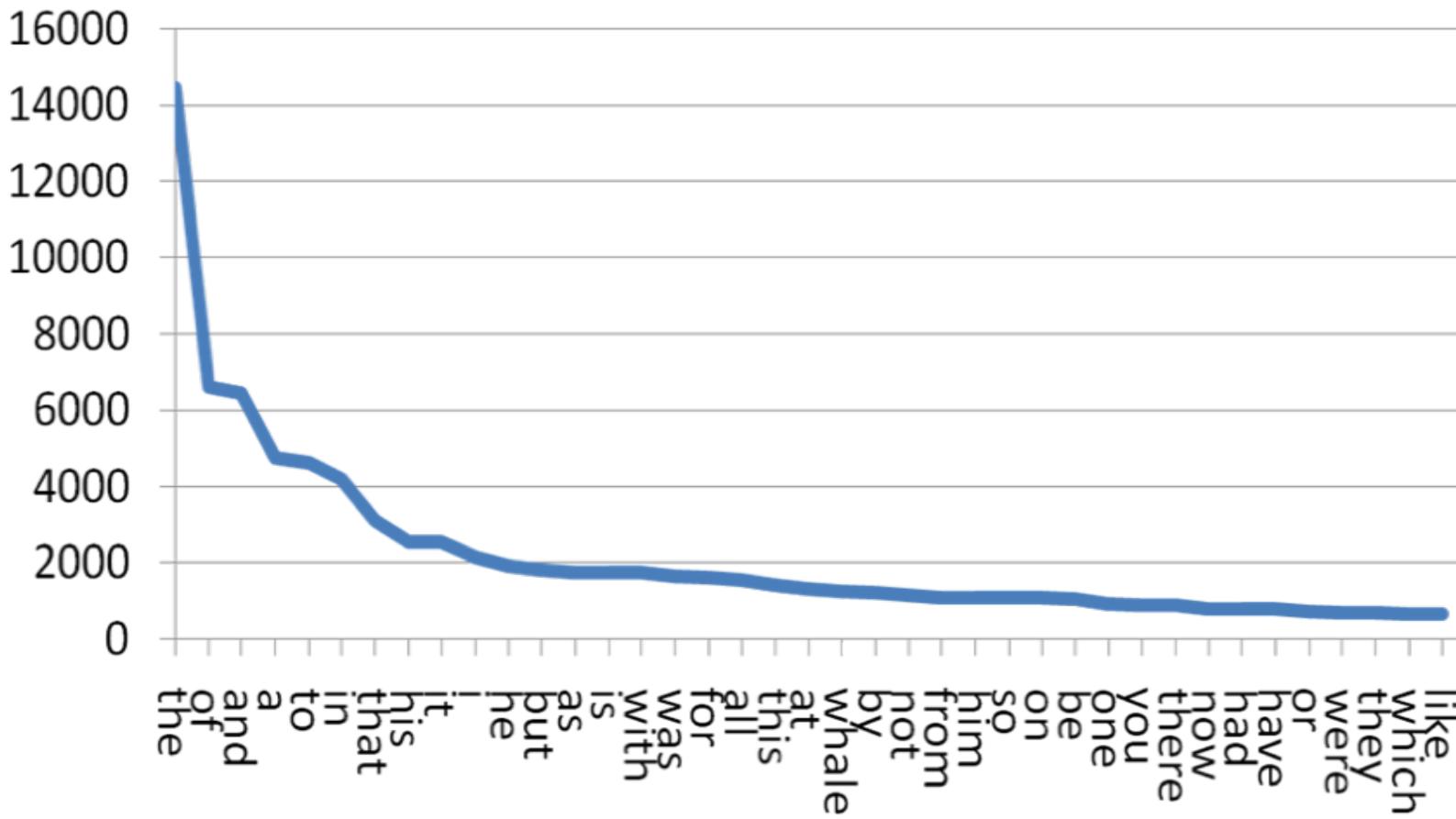


$$\prod_{i=1}^K p(\beta_i | \eta) \prod_{d=1}^D p(\theta_d | \alpha) \left(\prod_{n=1}^N p(z_{d,n} | \theta_d) p(w_{d,n} | \beta_{1:K}, z_{d,n}) \right)$$

NLP

- Value of the data in clinical text
- Hyper-simplified linguistics
- Term spotting + handling negation, uncertainty
- ML to expand terms
- pre-NN ML to identify entities and relations
- **language models**
- vector space embeddings based on co-occurrence
- adding context to help with disambiguation
- from embedding single words to phrases, sentences, etc.

Statistical Models of Language: Zipf's law



- There are very few very frequent words
- Most words have very low frequencies
- The frequency of a word is inversely proportional to its rank
- In the Brown corpus, the 10 top-ranked words make up 23% of total corpus size (Baroni, 2007)
-

N-gram models

- Shakespeare as a Corpus
 - N=884,647 tokens, V=29,066
 - Shakespeare produced 300,000 bigram types out of $V^2 = 844$ million possible bigrams...
 - So, 99.96% of the possible bigrams were never seen
- Google released corpus of 1,024,980,267,229 (i.e., ~1T) words in 2006
 - 13.6M unique words occurring at least 200 times
 - 1.2B five-word sequences that occur at least 40 times

Example Google 3-grams

ceramics	collectables	collectibles	55
ceramics	collectables	fine	130
ceramics	collected	by	52
ceramics	collectible	pottery	50
ceramics	collectibles	cooking	45
ceramics	collection	,	144
ceramics	collection	.	247
ceramics	collection	</S>	120
ceramics	collection	and	43
ceramics	collection	at	52
ceramics	collection	is	68
ceramics	collection	of	76
ceramics	collection	I	59
ceramics	collections	,	66
ceramics	collections	.	60
ceramics	combined	with	46
ceramics	come	from	69
ceramics	comes	from	660
ceramics	community	,	109
ceramics	community	.	210
ceramics	community	for	61
ceramics	companies	.	53
ceramics	companies	cprnsultants	173

Example Google 4-grams

serve	as	the	incoming	92
serve	as	the	incubator	99
serve	as	the	independent	79
serve	as	the	index	223
serve	as	the	indication	72
serve	as	the	indicator	120
serve	as	the	indicators	45
serve	as	the	indispensable	111
serve	as	the	indispensable	40
serve	as	the	individual	234
serve	as	the	industrial	52
serve	as	the	industry	607
serve	as	the	info	42
serve	as	the	informal	102
serve	as	the	information	838
serve	as	the	informational	41
serve	as	the	infrastructure	500
serve	as	the	initial	5331
serve	as	the	initiating	125
serve	as	the	initiation	63
serve	as	the	initiator	81
serve	as	the	injector	56
serve	as	the	inlet	41

Number of tokens:	1,024,980,267,229
Number of sentences:	95,119,665,584
Number of unigrams:	13,588,391
Number of bigrams:	314,843,401
Number of trigrams:	977,069,902
Number of fourgrams:	1,313,818,354
Number of fivegrams:	1,176,470,663

<https://ai.googleblog.com/2006/08/all-our-n-gram-are-belong-to-you.html>

Generating Sequences

- This model can be turned around to generate random sentences like sentences from which the model was derived.
- Generally attributed to Claude Shannon.
 - Sample random bigram (<s>, w) according to its prob.
 - Then sample random bigram (w, x) according to its prob.
 - Where the prefix w matches the suffix of the first.
 - And so on until we randomly choose a (y, </s>)
- Then string the words together

Simple example

```
<s> I  
I want  
want to  
to get  
get Chinese  
Chinese food  
food </s>
```

Generating Shakespeare

Unigram	<ul style="list-style-type: none">• To him swallowed confess hear both. Which. Of save on trail for are ay device and rote life have• Every enter now severally so, let• Hill he late speaks; or! a more to leg less first you enter• Are where exeunt and sighs have rise excellency took of.. Sleep knave we. near; vile like
Bigram	<ul style="list-style-type: none">• What means, sir. I confess she? then all sorts, he is trim, captain.• Why dost stand forth thy canopy, forsooth; he is this palpable hit the King Henry. Live king. Follow.• What we, hath got so she that I rest and sent to scold and nature bankrupt, nor the first gentleman?• Enter Menenius, if it so many good direction found'st thou art a strong upon command of fear not a liberal largess given away, Falstaff! Exeunt
Trigram	<ul style="list-style-type: none">• Sweet prince, Falstaff shall die. Harry of Monmouth's grave.• This shall forbid it should be branded, if renown made it empty.• Indeed the duke; and had a very good friend.• Fly, and will rid me these news of price. Therefore the sadness of parting, as they say, 'tis done.
Quadrigram	<ul style="list-style-type: none">• King Henry. What! I will go seek the traitor Gloucester. Exeunt some of the watch. A great banquet serv'd in;• Will you not tell me who I am?• It cannot be but so.• Indeed the short and the long. Marry, 'tis a noble Lepidus.

NLP

- Value of the data in clinical text
- Hyper-simplified linguistics
- Term spotting + handling negation, uncertainty
- ML to expand terms
- pre-NN ML to identify entities and relations
- language models
- **vector space embeddings based on co-occurrence**
- adding context to help with disambiguation
- from embedding single words to phrases, sentences, etc.

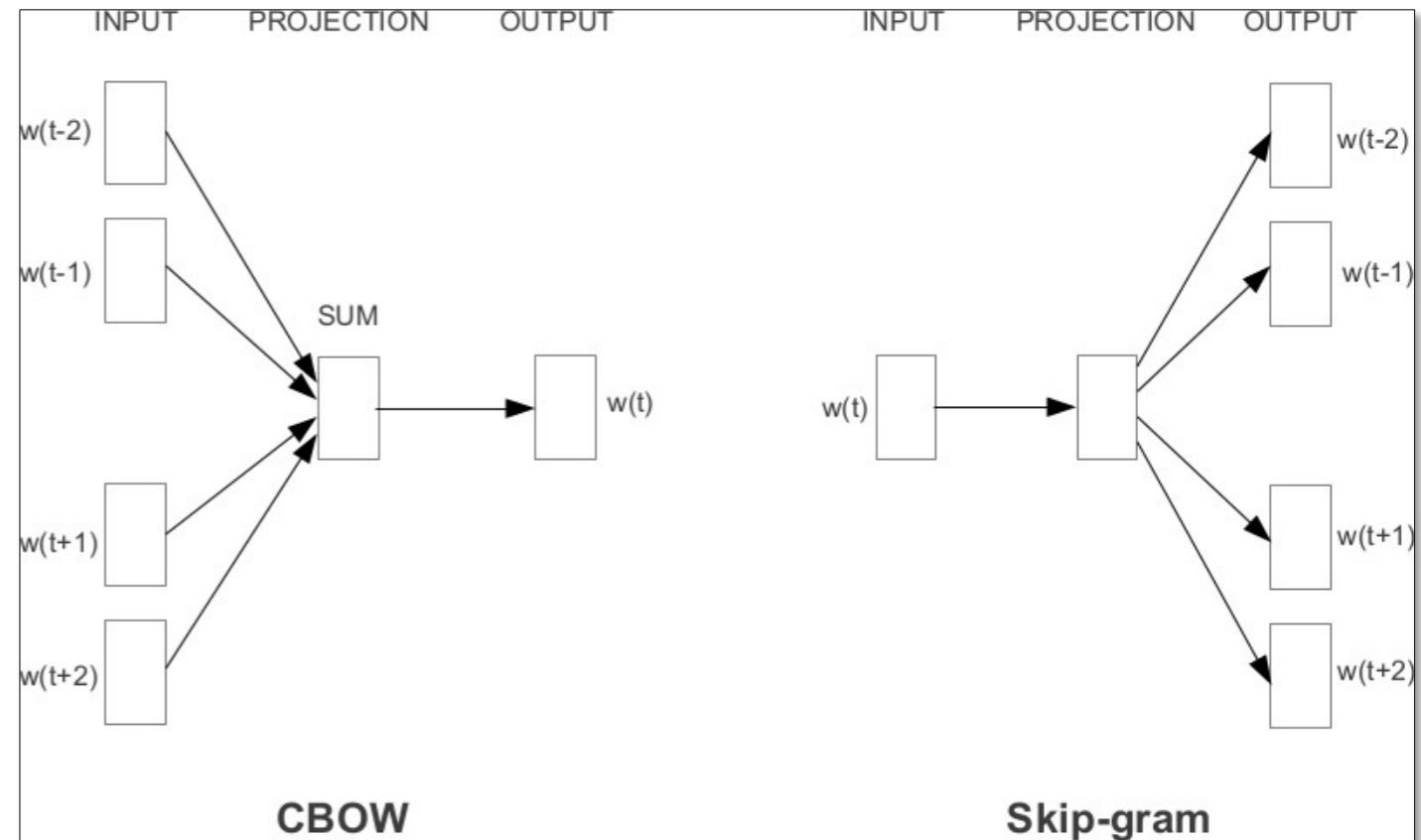
Modern-day NLP: Actually modeling language: LLMs

- **Neural Language Modeling Key Ideas**
 - Can build arbitrarily complex models so long as they are differentiable
 - Then, they can be trained by Stochastic Gradient Descent
 - Pre-training on arbitrarily large unannotated corpora
 - Then, adapt via fine-tuning or even few-shot learning
- **Pretext Task:** Predict the next token given the ones before it
 - In unigram model, $P(\text{token})$ is just estimated from frequency in corpus
- **Markov assumption** simplifies model so
 - $P(\text{token} \mid \text{stuff before}) = P(\text{token} \mid \text{previous token})$ [bigram model]
 - $P(t_k \mid \text{stuff before}) = P(t_k \mid t_{k-1}, \dots, t_{k-n})$ [n -gram models]
- **Perplexity** is an aggregate measure of the **complexity** of a corpus
 - $2^{H(p)}$ where $H(p)$ is the entropy of the probability distribution
 - intuitively, the number of likely ways to continue a text
 - a perplexity of k means that you are as surprised on average as you would have been if you had to guess between k equiprobable choices at each step
 - For example, we compared perplexity of dictated doctors' notes (8.8) vs. Perplexity of doctor-patient conversations (73.1)
 - What does that tell you about the difficulty of accurately transcribing speech for these applications?

Distributional Semantics

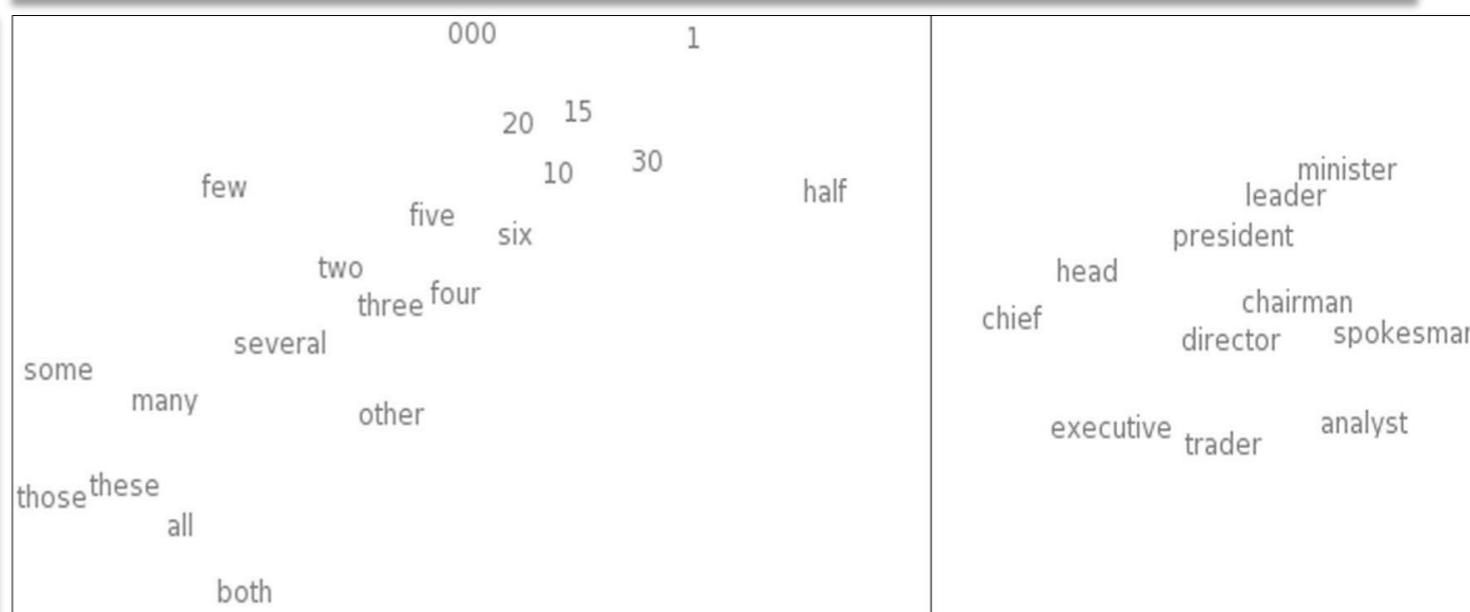
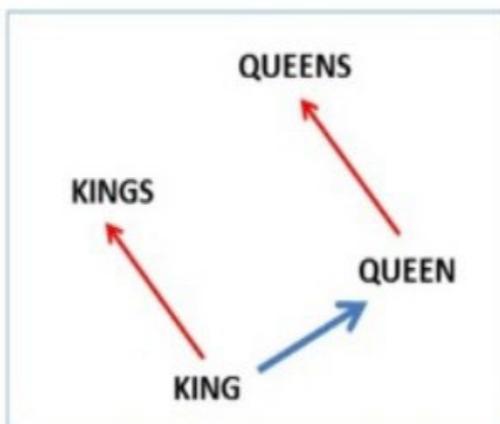
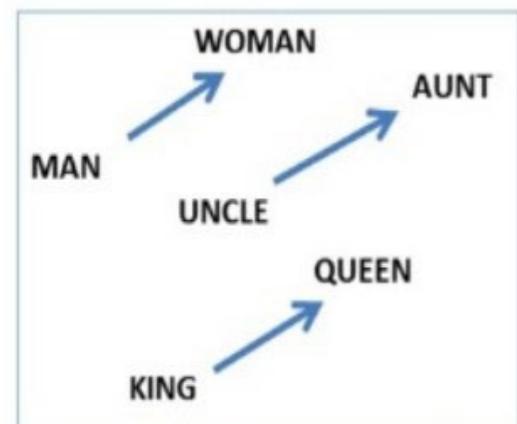
- Terms that appear in the same context of other words are (probably) semantically related
- Every term is mapped to a high-dimensional vector (the embedding space)
- Ever more sophisticated versions of embeddings, equivalent to matrix factorization
 - Word2Vec
 - GloVe
 - Elmo
 - Bert
 - GPT

word2vec



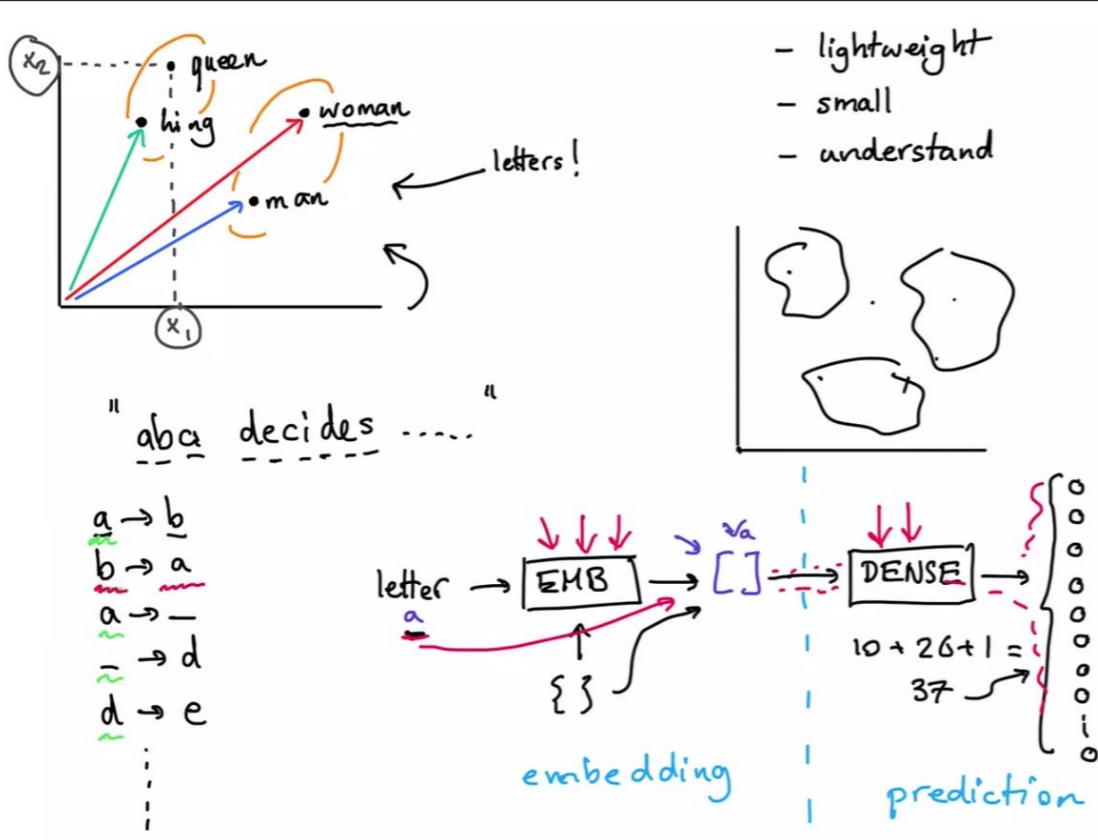
**Embedding space calculations:
Plausibility of semantic claims**

$$\text{vec("man")} - \text{vec("king")} + \text{vec("woman")} = \text{vec("queen")}$$



t-Distributed Stochastic Neighbor Embedding of high-dim space

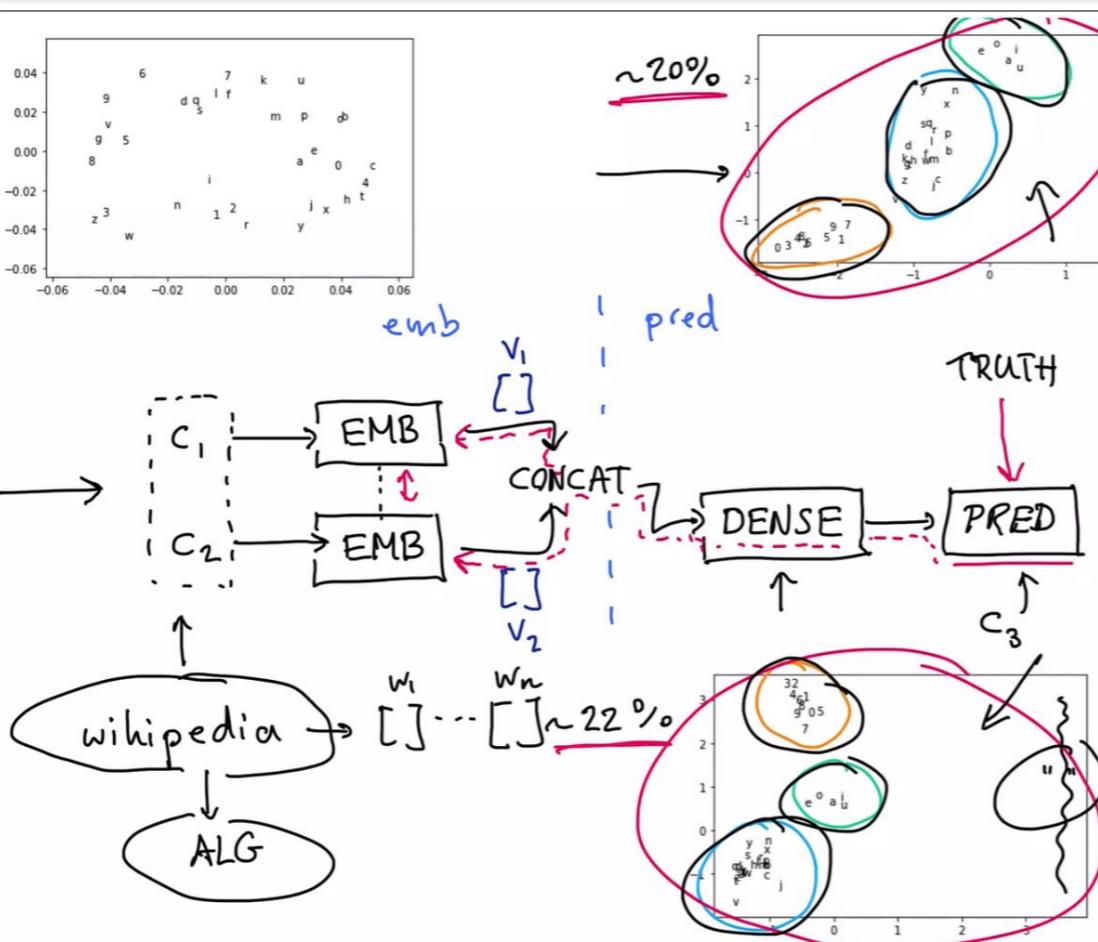
Mapping words to a conceptual embedding space



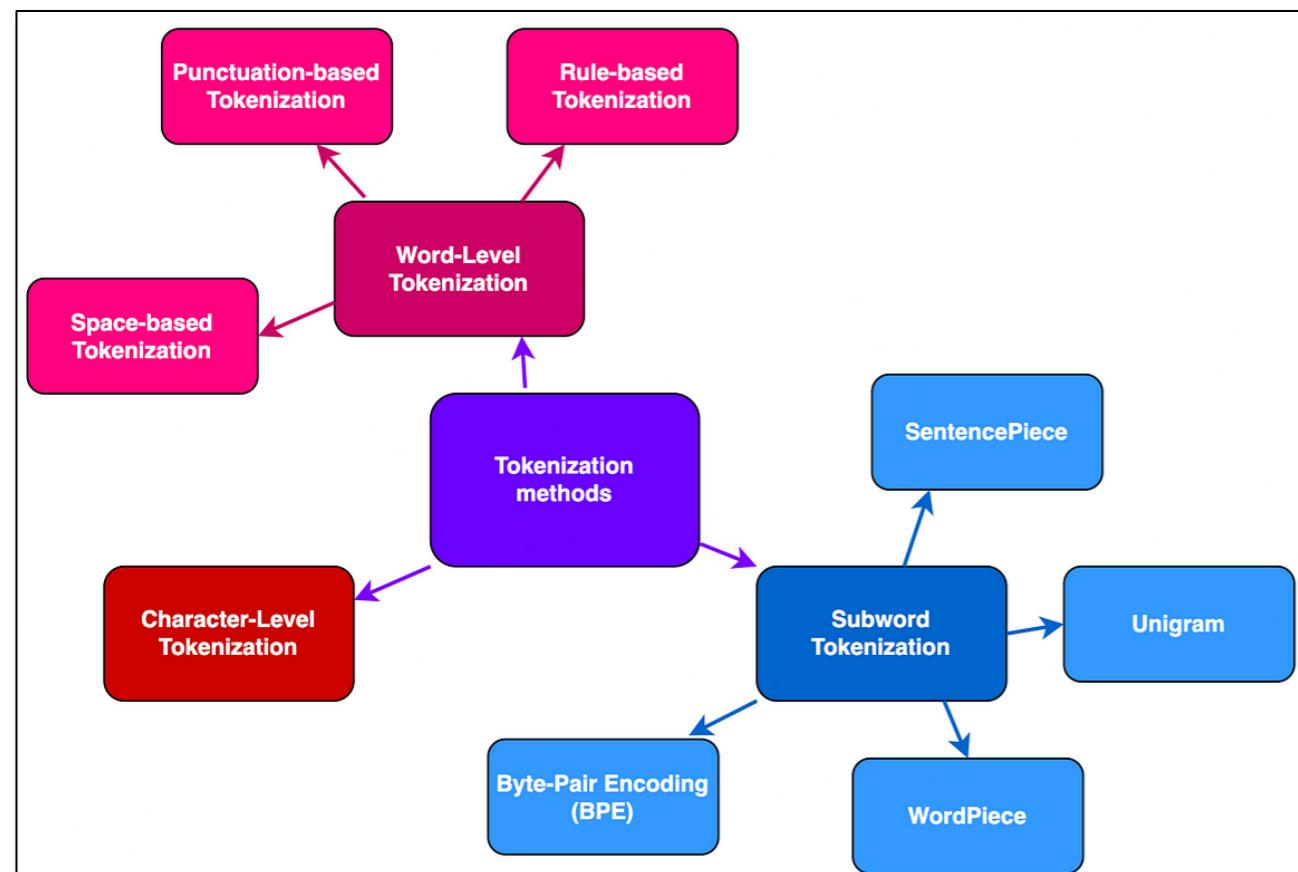
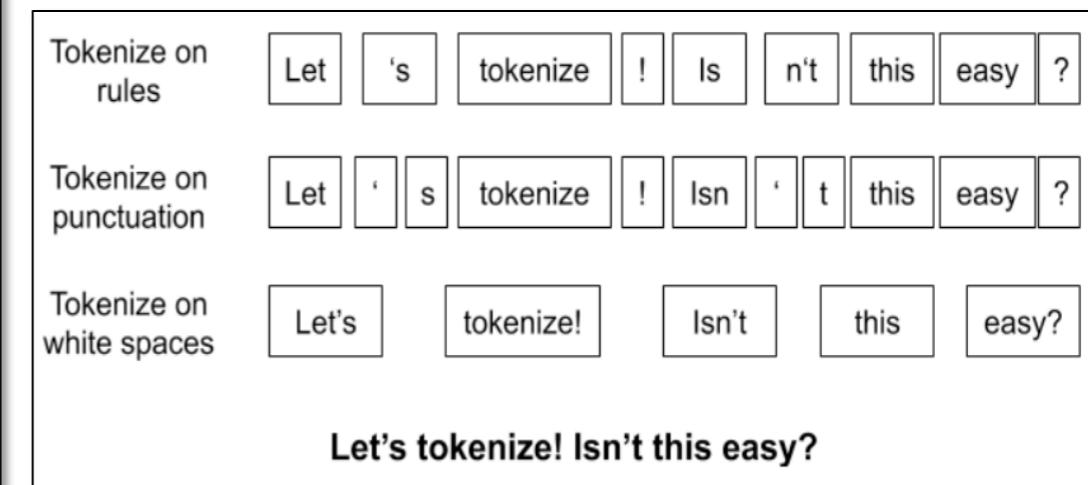
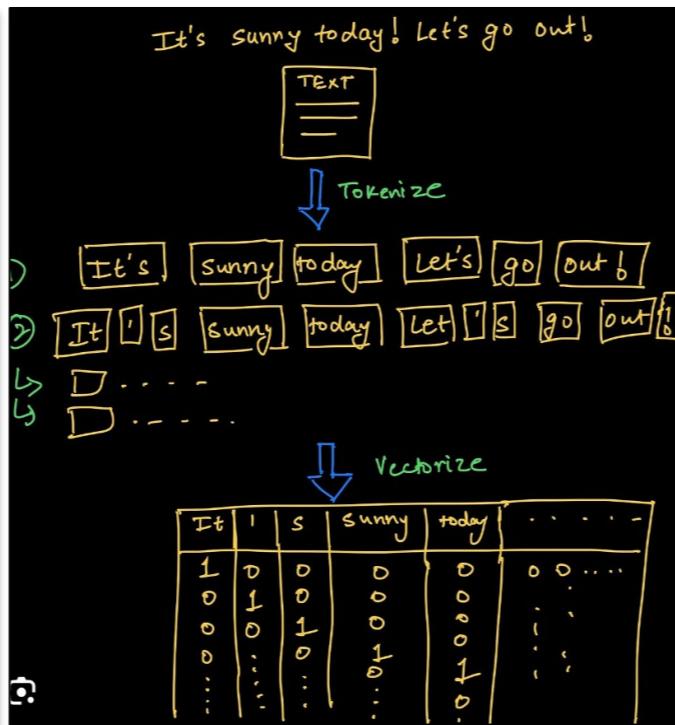
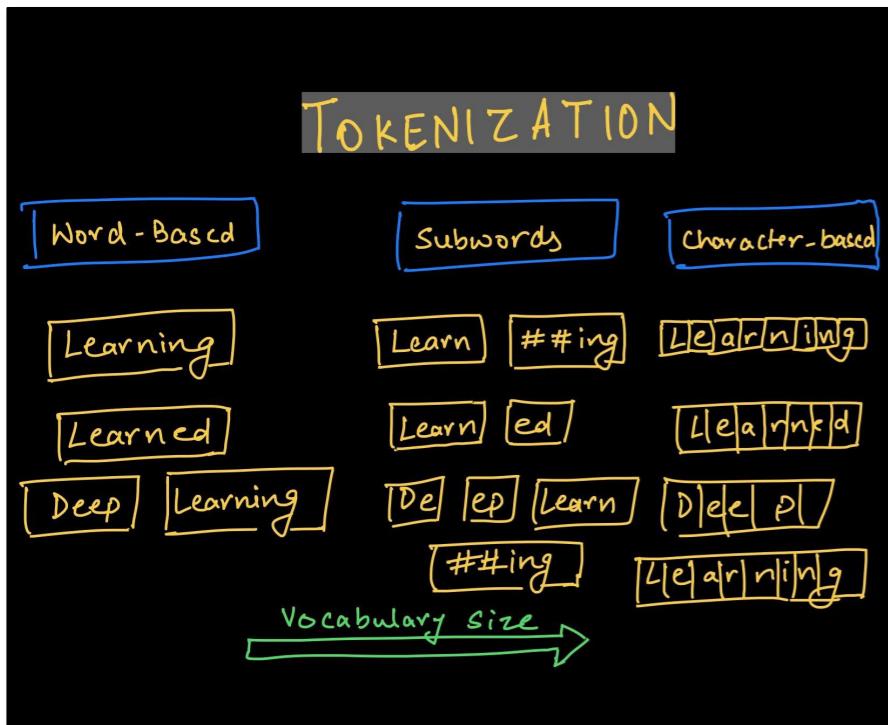
```

1  from tensorflow.keras.layers import Embedding, Dense, Flatten
2  from tensorflow.keras.models import Sequential
3  from tensorflow.keras.optimizers import Adam
4
5
6 # this one is so we might grab the embeddings
7 model_emb = Sequential()
8 embedding = Embedding(num_letters, 2, input_length=1)
9 model_emb.add(embedding)
10 output_array = model_emb.predict(integers_in)
11
12
13 # this one is so we might learn the mapping
14 model_pred = Sequential()
15 model_pred.add(embedding)
16 model_pred.add(Flatten())
17 model_pred.add(Dense(num_letters, activation="softmax"))
18
19 adam = Adam(learning_rate=0.001, beta_1=0.9, beta_2=0.999, amsgrad=False)
20
21 # notice that we only compile model_pred
22 model_pred.compile('adam', 'categorical_crossentropy', metrics=['accuracy'])
23 model_pred.fit(...)
```

- Words with similar contexts should map to similar coordinates in the embedding space
- To achieve this, use prediction context:
encoding [embedding, representation learning],
decoding [actual prediction]
- Train weights through densely-connected network [dense] and through embeddings [emb] with backpropagation
- Initial embeddings are scattered, but after training, characters group together [and words similarly]
- Use multiple consecutive characters to increase context information → Prediction improves
- From characters to words: need larger context, more layers, higher-dimensional representation



Tokenization: Words \leftrightarrow Subwords \leftrightarrow Roots/Rules \leftrightarrow Characters



aaabdaaaabac

The byte pair "aa" occurs most often, so it will be replaced by a byte that is not used in the data, such as "Z". Now there is the following data and replacement table:

ZabdZaba
Z̄aa

Then the process is repeated with byte pair "ab", replacing it with "Y":

ZYdZYac
Y=ab
Z=aa

The only literal byte pair left occurs only once, and the encoding might stop here. Alternatively, the process could continue with recursive byte pair encoding, replacing "ZY" with "X":

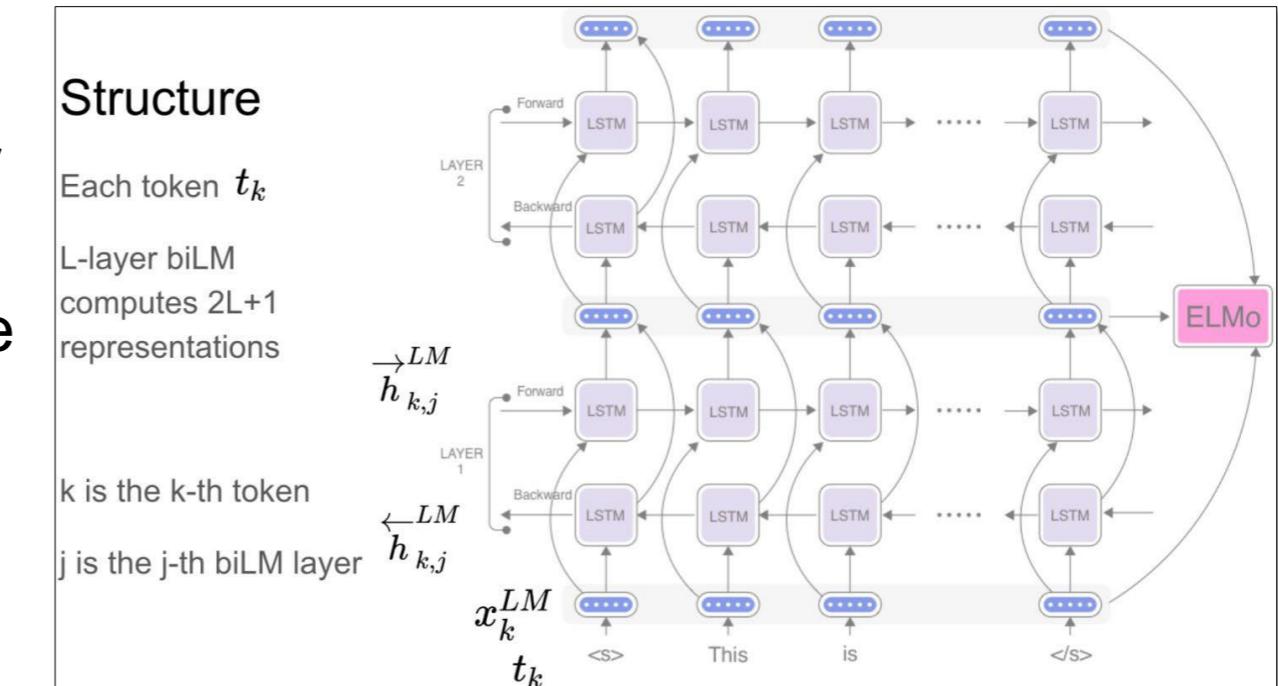
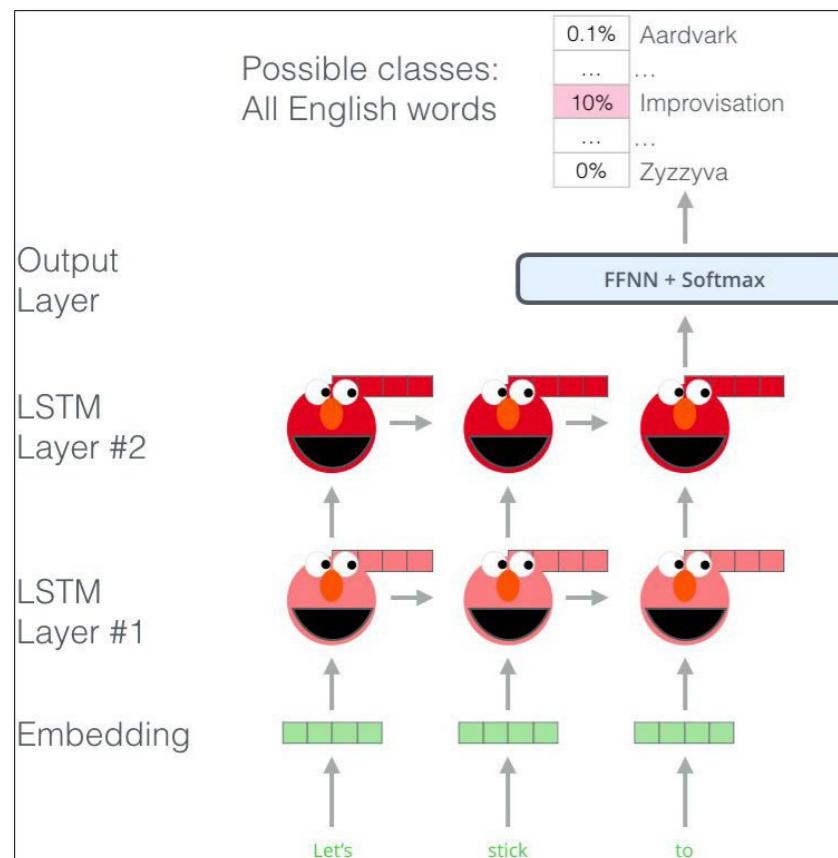
XdXac
X=ZY
Y=ab
Z=aa

NLP

- Value of the data in clinical text
- Hyper-simplified linguistics
- Term spotting + handling negation, uncertainty
- ML to expand terms
- pre-NN ML to identify entities and relations
- language models
- **ELMO**
- vector space embeddings based on co-occurrence
- adding context to help with disambiguation
- from embedding single words to phrases, sentences, etc.

ELMo—(Contextual) Embeddings from Language Models

- Bidirectional LSTM
- Builds models for every *token*, not just for every *type*
 - i.e., different embeddings for the same word in different contexts
 - basis for word-sense disambiguation
- Significantly improves performance on nearly all NLP tasks



Source	Nearest Neighbors
GloVe play	playing, game, games, played, players, plays, player, Play, football, multiplayer
biLM Chico Ruiz made a spectacular <u>play</u> on Alusik 's grounder {...}	Kieffer , the only junior in the group , was commended for his ability to hit in the clutch , as well as his all-round excellent <u>play</u> .
	{... } they were actors who had been handed fat roles in a successful <u>play</u> , and had talent enough to fill the roles competently , with nice understatement .

Table 4: Nearest neighbors to “play” using GloVe and the context embeddings from a biLM.

Language model pre-training:

- How is learning to predict the next word relevant to doing deidentification, predicting sepsis from nursing notes, etc?
- It allows us learn how to represent words & sentences FOR FREE.

Key Features:

- Allows for contextual representations of words
- Similar to Word2Vec, this is unsupervised (i.e., pretraining)
- SUPER SLOW

ANN model for de-identification

But ELMO's goal
is to compute an embedding for
each token, not to solve a particular
problem. Then stack with
downstream task!

- Label-sequence optimization layer

$$s(y_{1:n}) = \sum_{i=1}^n \mathbf{a}_i[y_i] + \sum_{i=2}^n T[y_{i-1}, y_i]$$

- Label prediction layer

- Character-enhanced token-embedding layer

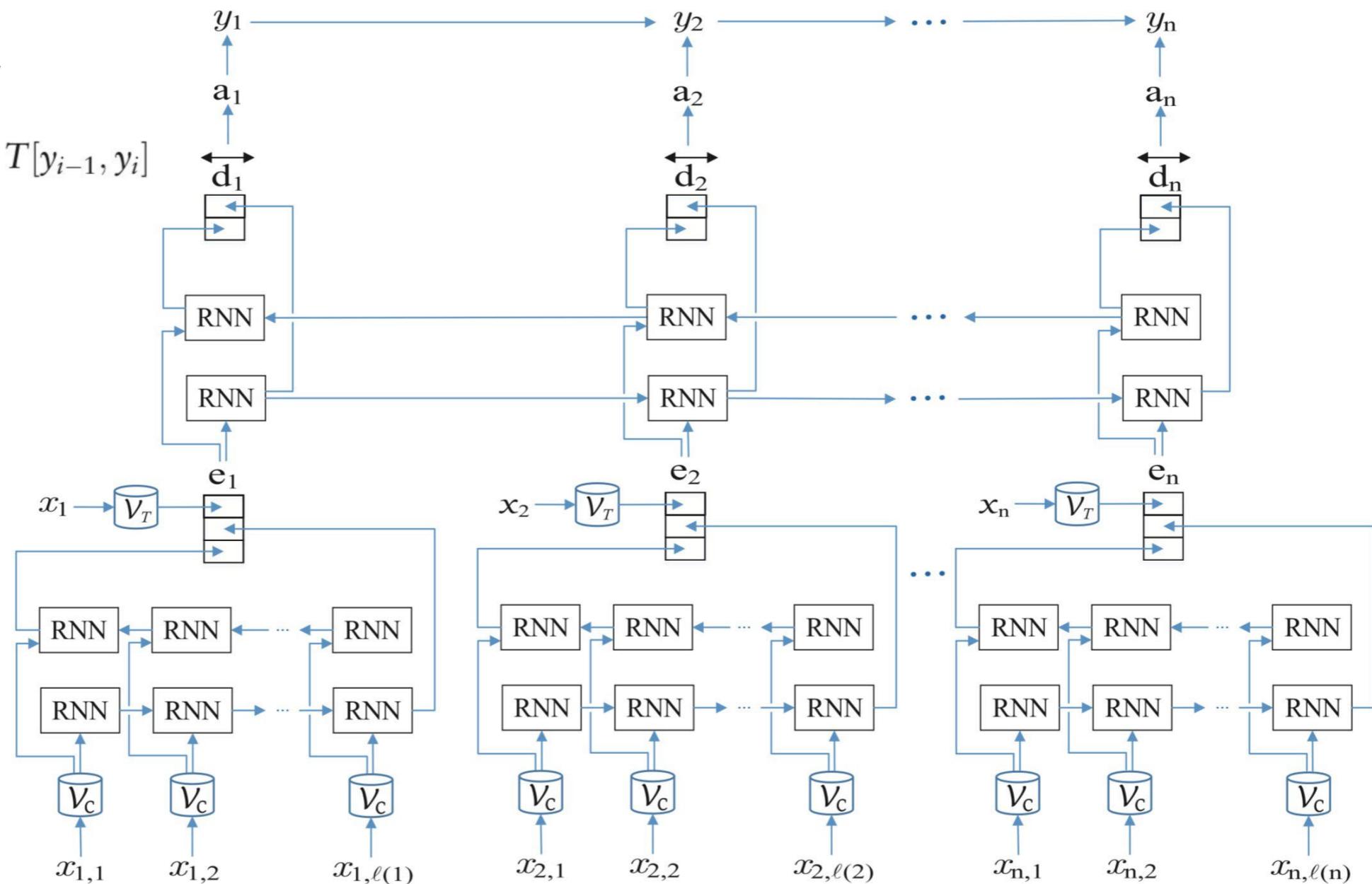


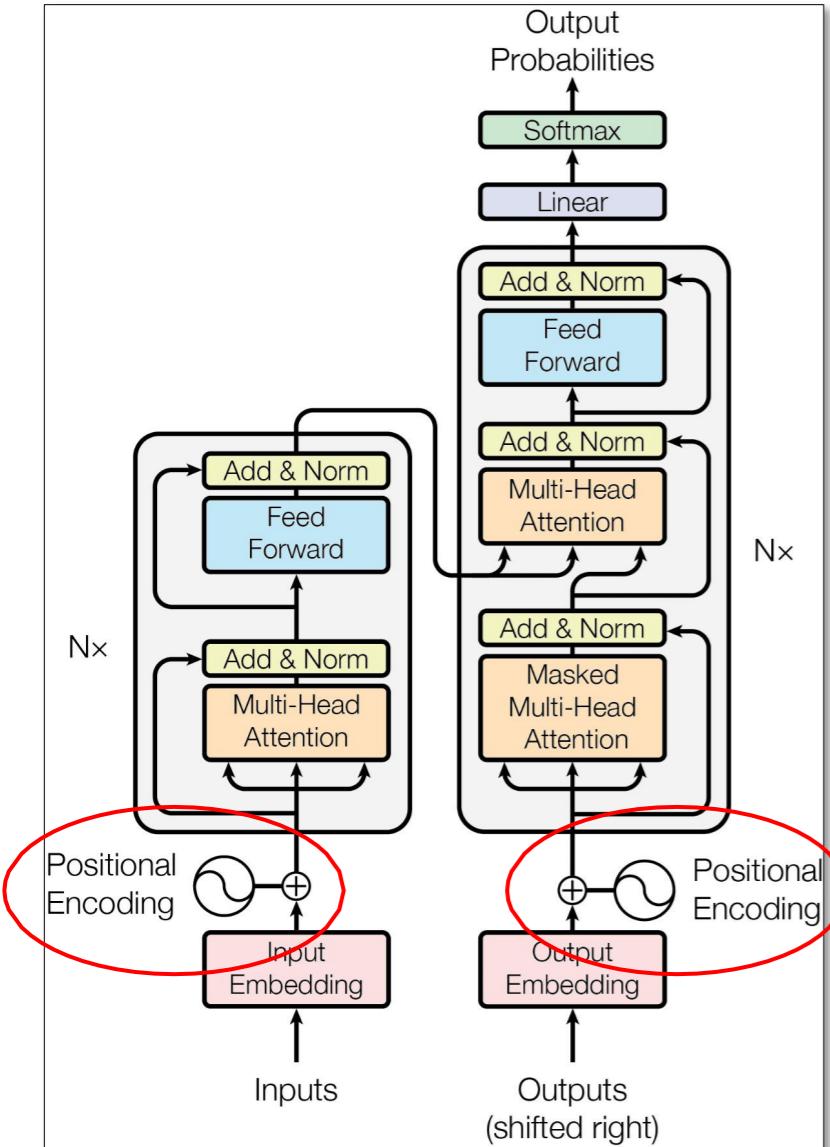
Figure 1. Architecture of the artificial neural network (ANN) model. (RNN, recurrent neural network.) The type of RNN used in this model is long short-term memory (LSTM). n is the number of tokens, and x_i is the i^{th} token. \mathcal{V}_T is the mapping from tokens to token embeddings. $\ell(i)$ is the number of characters and $x_{i,j}$ is the j^{th} character in the i^{th} token. \mathcal{V}_C is the mapping from characters to character embeddings. e_i is the character-enhanced token embeddings of the i^{th} token. \vec{d}_i is the output of the LSTM of the label prediction layer, a_i is the probability vector over labels, y_i is the predicted label of the i^{th} token.

Dernoncourt, F., Lee, J. Y., Uzuner, Ö., & Szolovits, P. (2016). De-identification of patient notes with recurrent neural networks. *Journal of the American Medical Informatics Association*, ocw156. <http://doi.org/10.1093/jamia/ocw156>

NLP

- Value of the data in clinical text
- Hyper-simplified linguistics
- Term spotting + handling negation, uncertainty
- ML to expand terms
- pre-NN ML to identify entities and relations
- language models
- LSTMs + ELMO
- **Attention**
- vector space embeddings based on co-occurrence
- adding context to help with disambiguation
- from embedding single words to phrases, sentences, etc.

Attention is all you need: Transformer (2017)

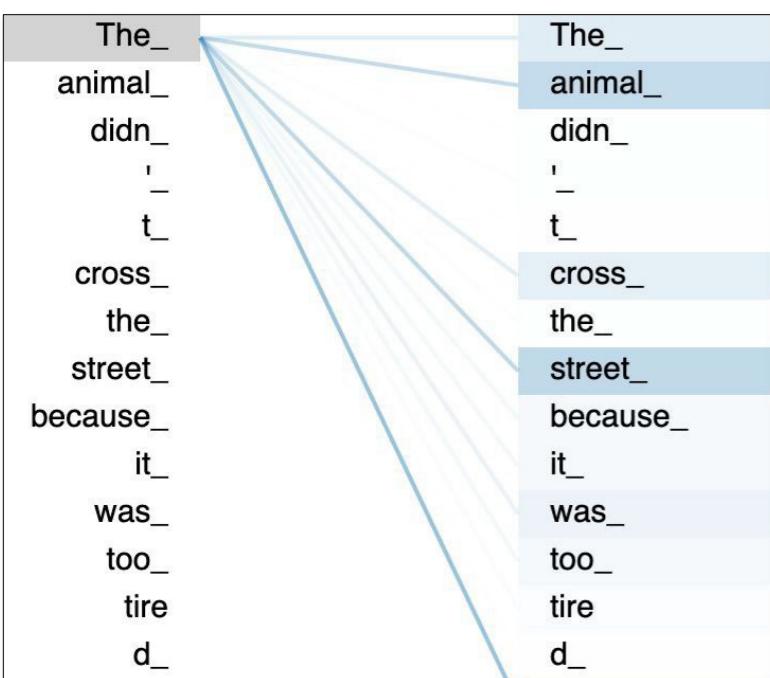


- Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez AN, et al. Attention is all you need. In: NeurIPS. 2017. Available from: <https://proceedings.neurips.cc/paper/2017/file/3f5ee243547dee91fb053c1c4a845aa-Paper.pdf>
- 66,919 citations, as of March 1, 2023!
 - By comparison, Watson & Crick's 1953 Nature paper on the structure of DNA has 17,264 citations

Encoder: reads the entire sequence all at once.

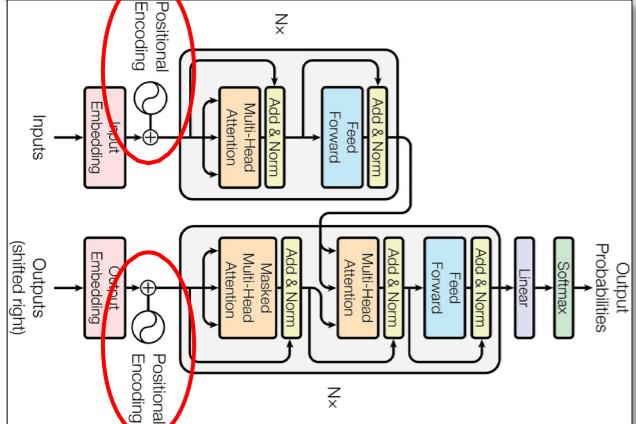
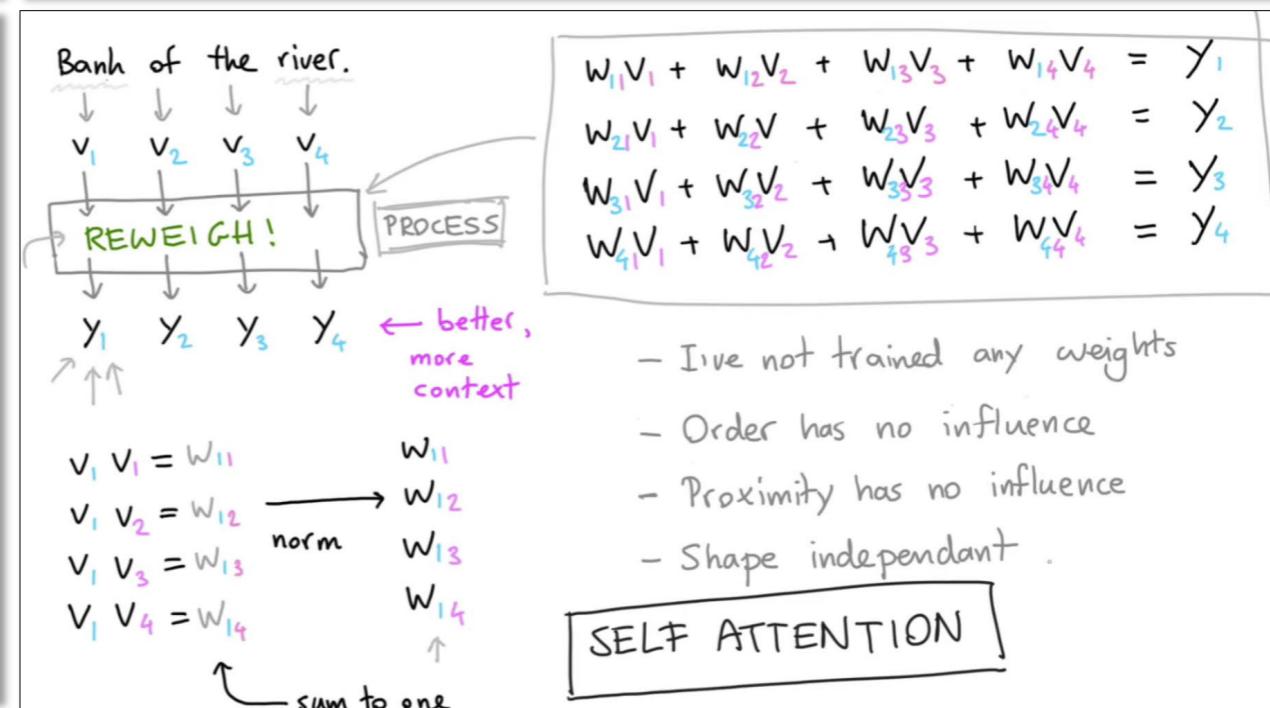
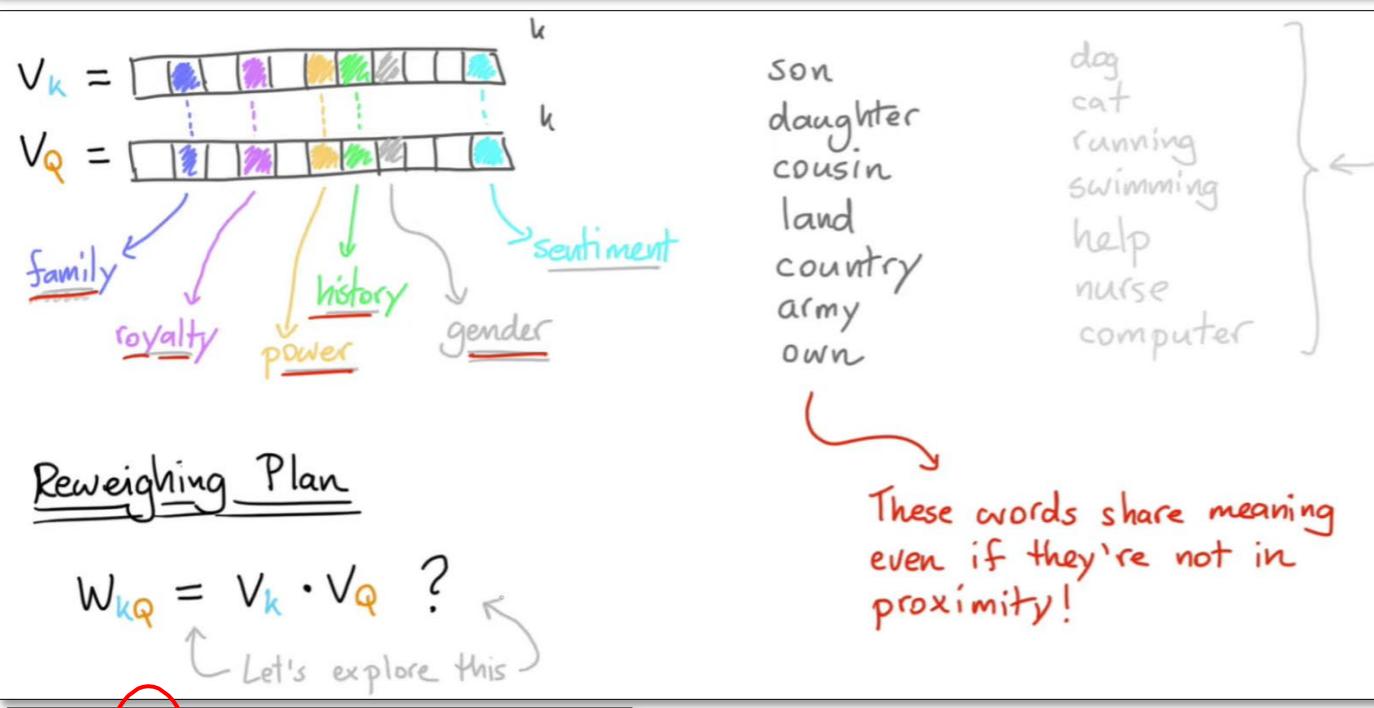
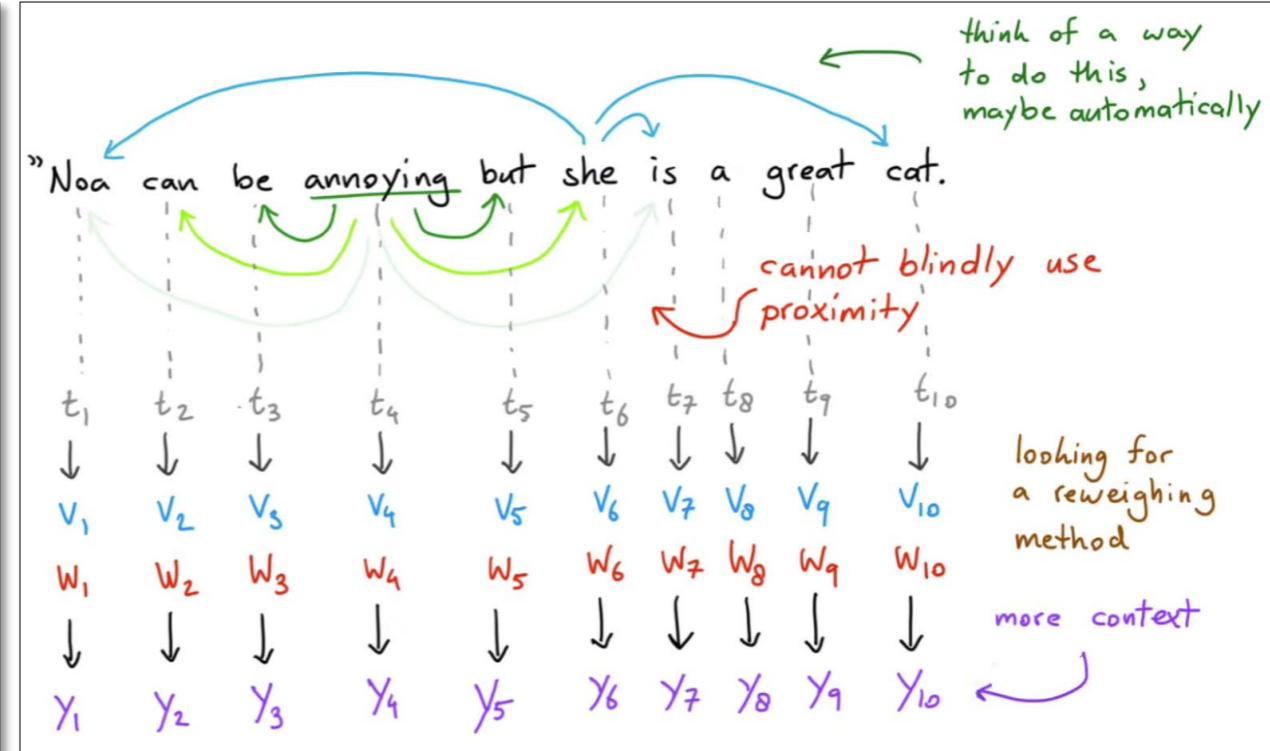
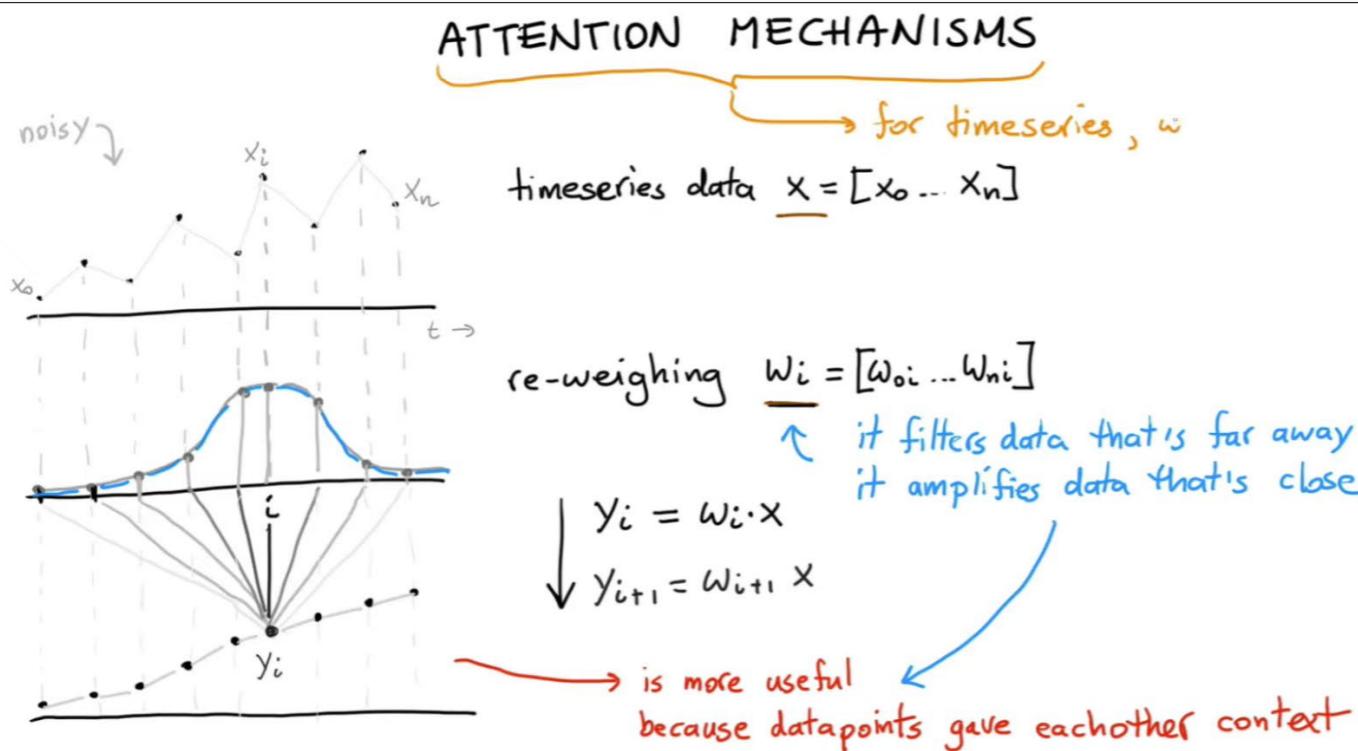
Decoder: reads left to right (but parallelized)

- Positional encodings:
- Every position (i.e., the first, second, third... word) has a unique vector that represents its position in the sentence/paragraph
- These are randomly initialized and learned by the model!

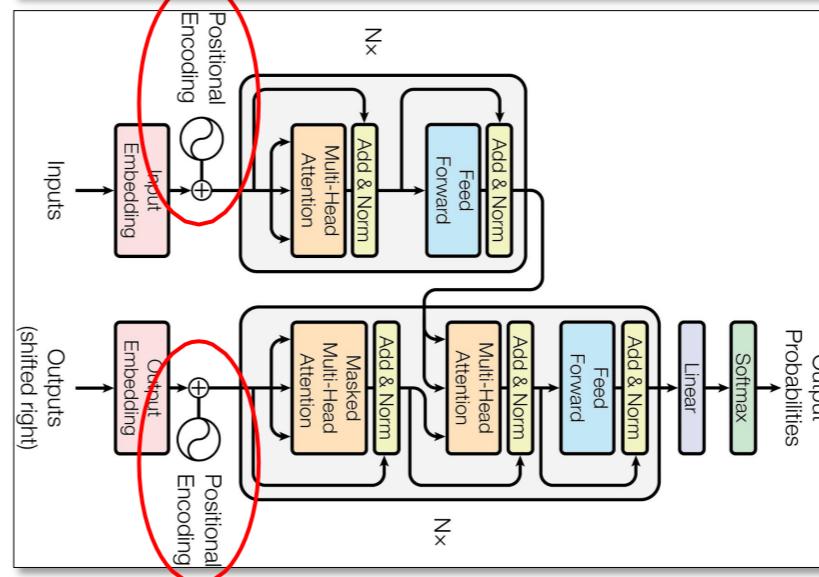
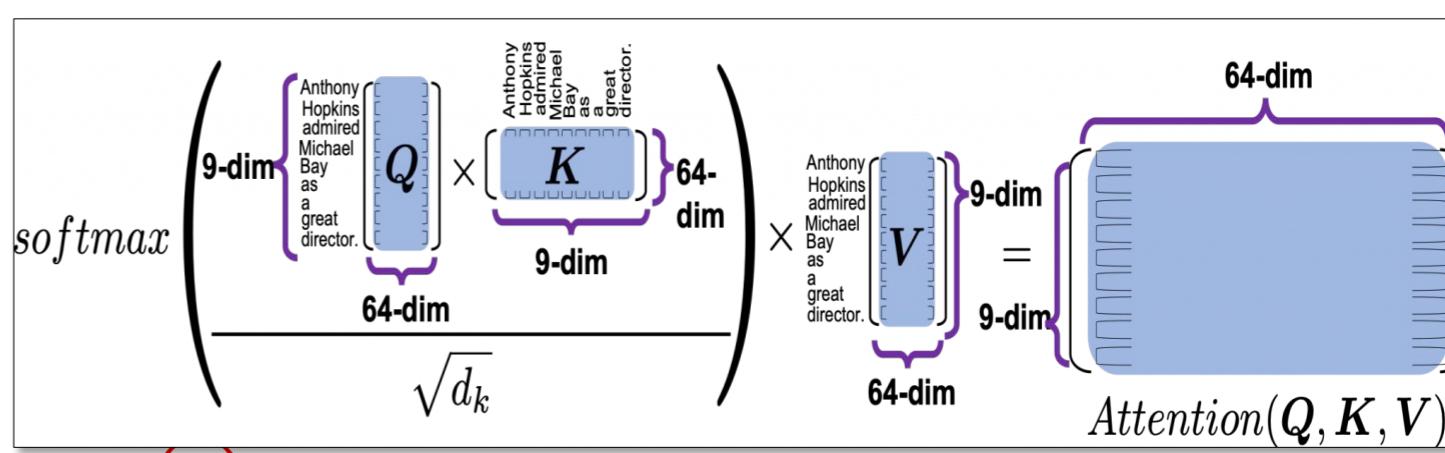
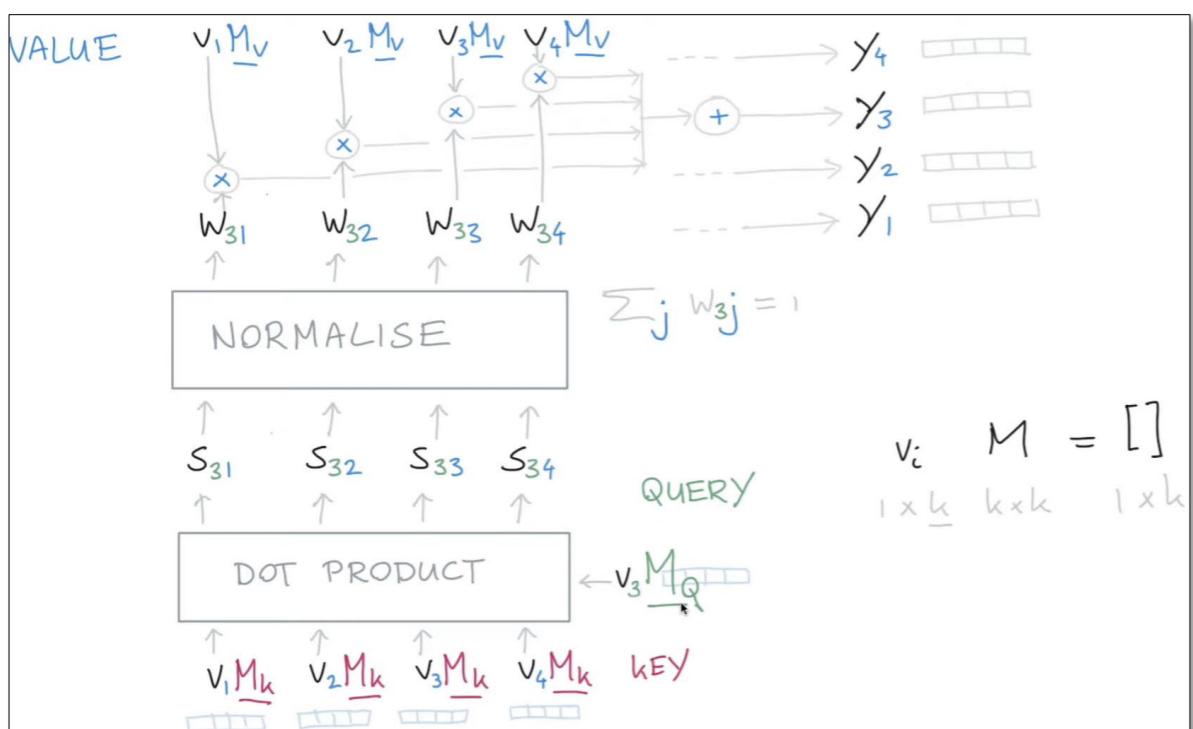
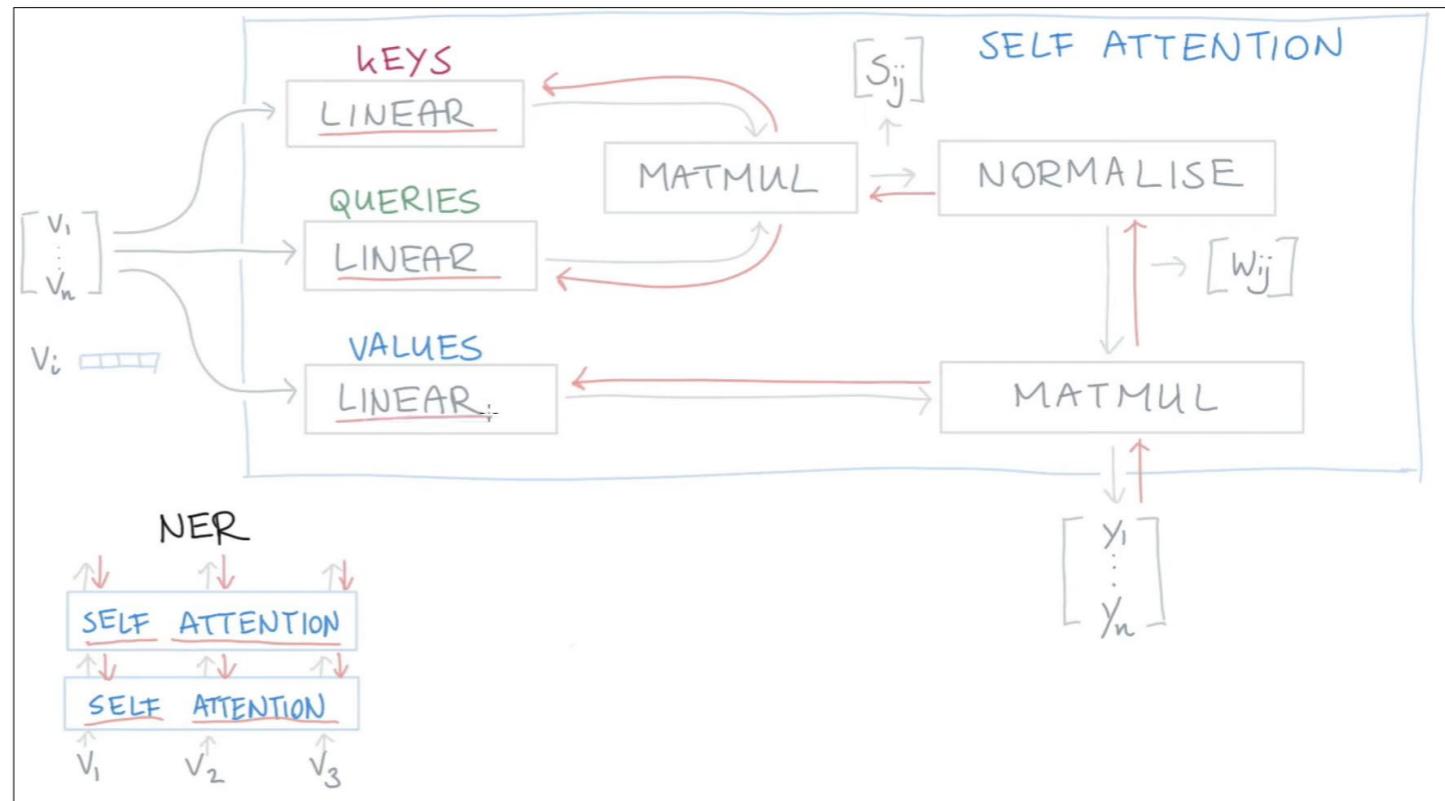
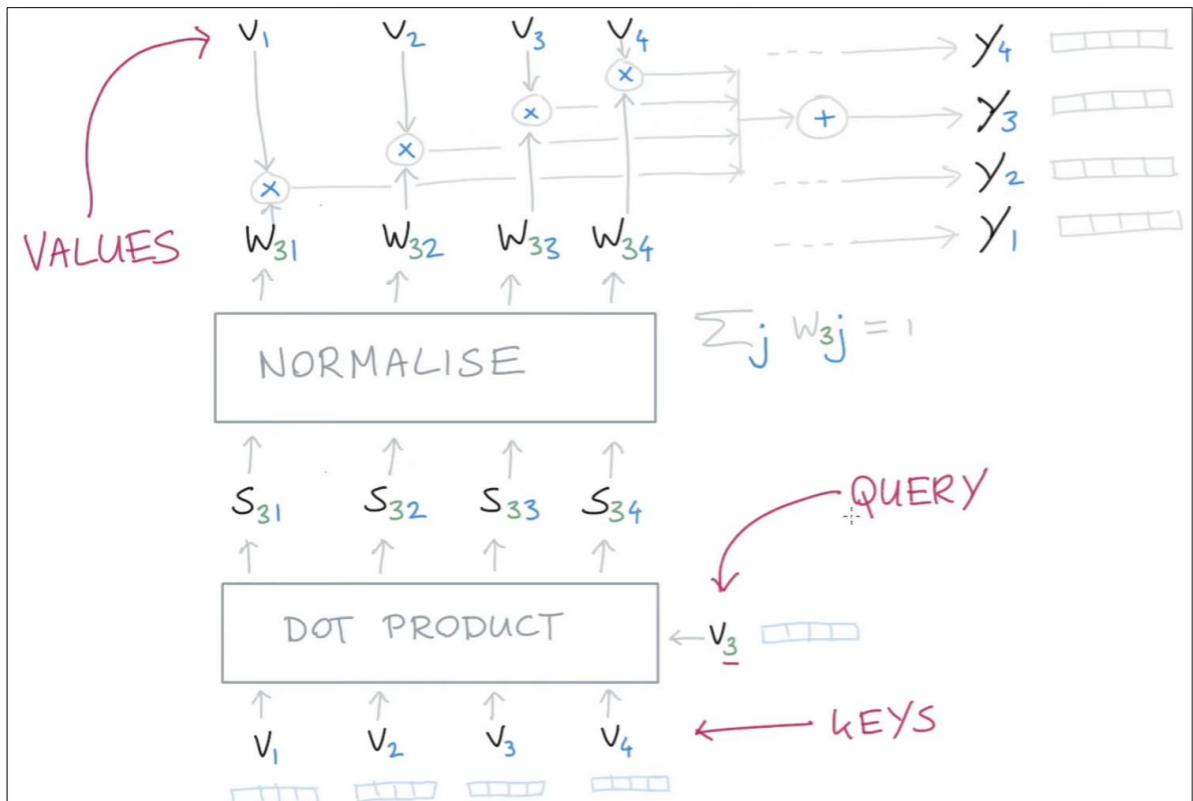


Idea: How important is *this word*, with respect to ALL other words?

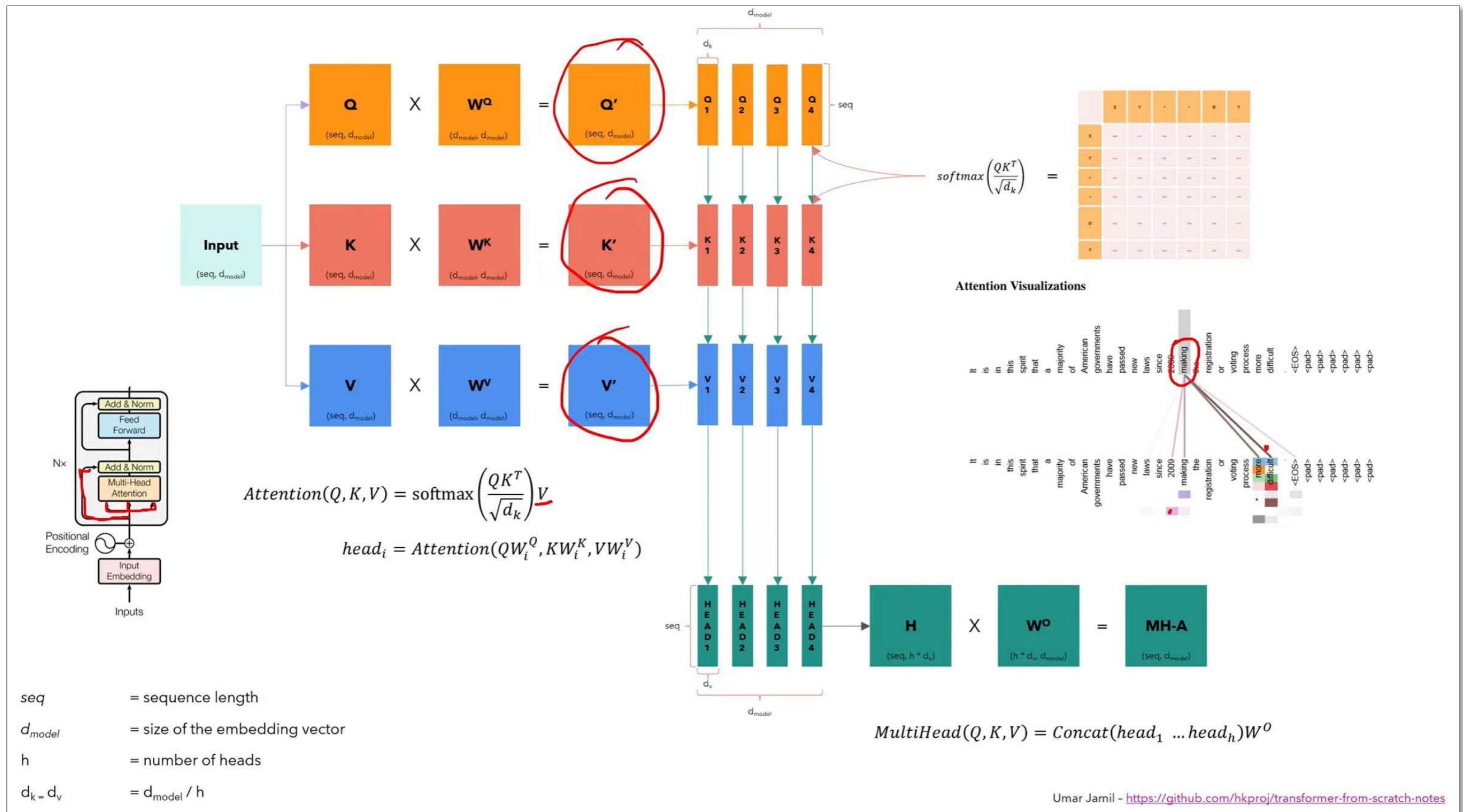
Understanding self-attention for Human Language



Understanding Key, Query, Value concepts



Multi-headed attention



Each head runs its own attention, can capture different aspects of word relationships

Decoder – Masked attention (no peeking to the future)

Our goal is to make the model causal: it means the output at a certain position can only depend on the words on the previous positions. The model **must not** be able to see future words.

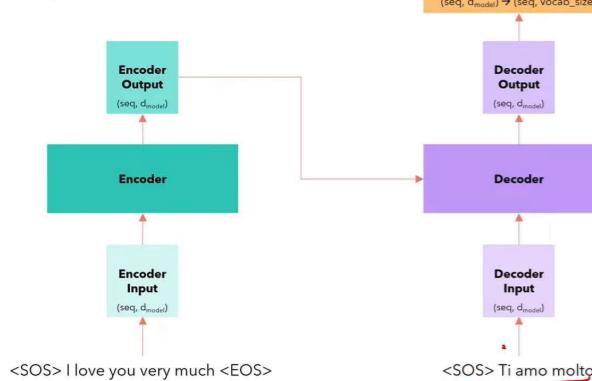
	YOUR	CAT	IS	A	LOVELY	CAT
YOUR	0.368	0.147	0.334	0.48	0.49	0.142
CAT	0.124	0.278	0.211	0.49	0.54	0.175
IS	0.147	0.132	0.262	0.097	0.28	0.145
A	0.210	0.128	0.206	0.212	0.179	0.125
LOVELY	0.146	0.158	0.152	0.143	0.227	0.14
CAT	0.195	0.114	0.203	0.103	0.157	0.229

Inference – sequentially receiving output(t-1) as part of input(t)

Training

Time Step = 1
It all happens in one time step!

The encoder outputs, for each word a vector that not only captures its meaning (the embedding) or the position, but also its interaction with other words by means of the multi-head attention.

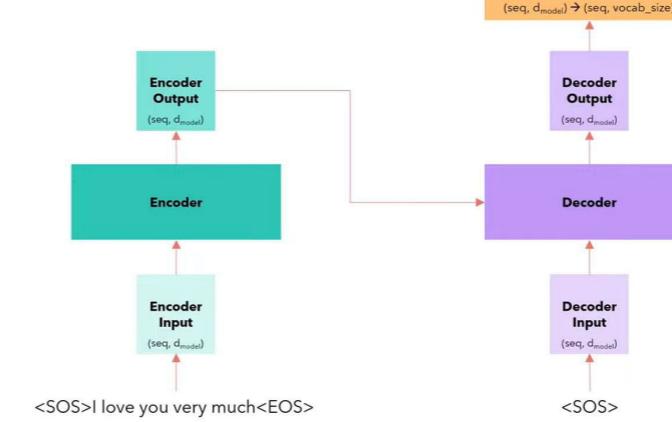


Ti amo molto <EOS>
* This is called the "label" or the "target"

Inference

We select a token from the vocabulary corresponding to the position of the token with the maximum value.

The output of the last layer is commonly known as **logits**



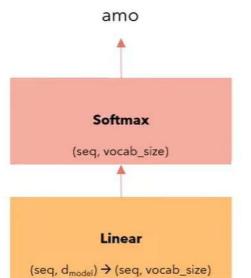
* Both sequences will have same length thanks to padding

Inference

Time Step = 2

Use the encoder output from the first time step

<SOS>I love you very much<EOS>



Since decoder input now contains **two** tokens, we select the softmax corresponding to the second token.

<SOS> ti

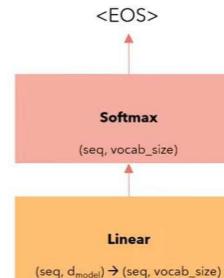
Append the previously output word to the decoder input

Inference

Time Step = 4

Use the encoder output from the first time step

<SOS>I love you very much<EOS>



Since decoder input now contains **four** tokens, we select the softmax corresponding to the fourth token.

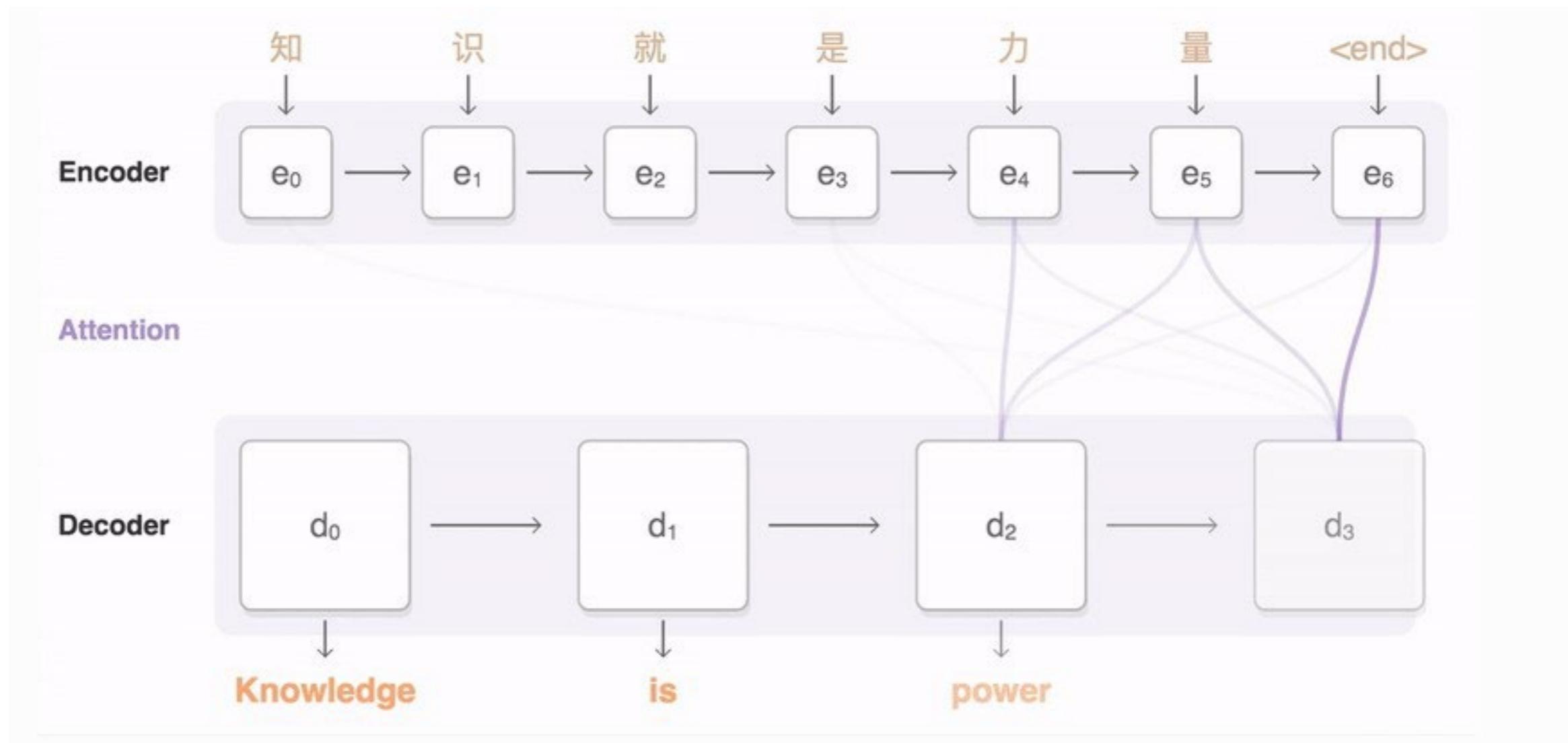
<SOS> ti amo molto

Append the previously output word to the decoder input

Training: all at once

Inference: one word (token) at a time

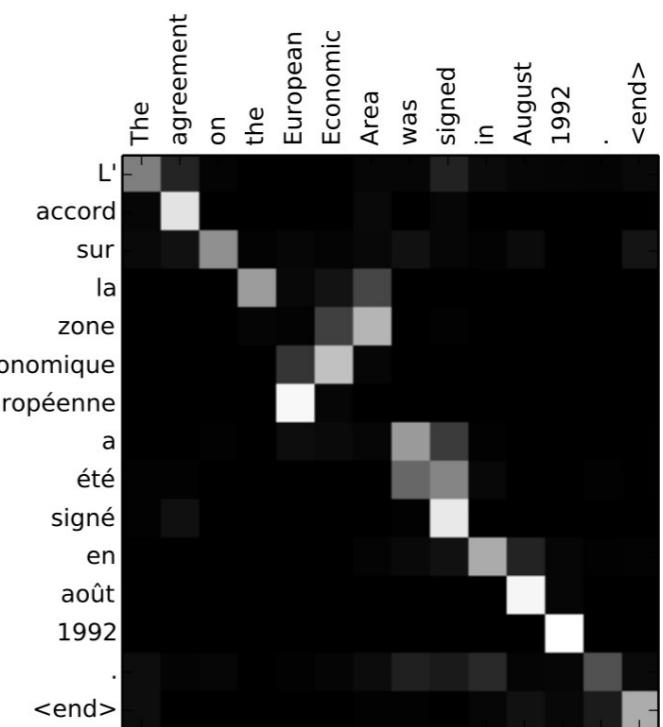
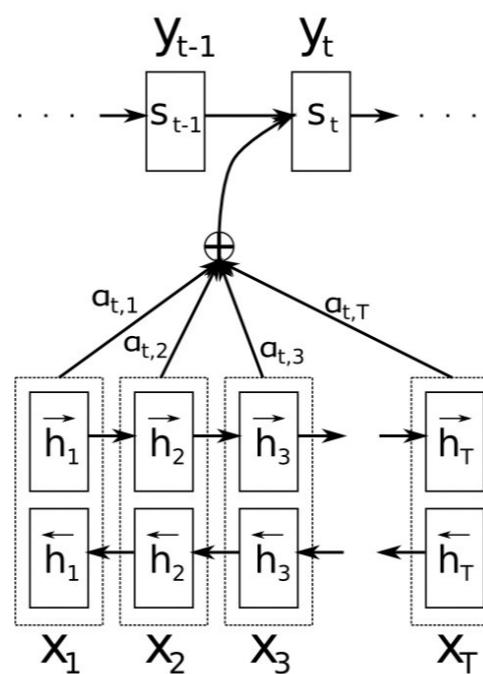
Illustration of Learned Attention Weights



How to Model the Attention Weights

- Attention depends on:
 - score of relationship between word being generated and all input words [“dictionary”]
 - learned positional dependencies [“alignment”]
- State of the decoder $s_t = f(s_{t-1}, y_{t-1}, \mathbf{c}_t), t = 1, \dots, m$ where $\mathbf{c}_t = \sum_{i=1}^n a_{t,i} \mathbf{h}_i$
- where \mathbf{h}_i are all the i -th word encoder states and $a_{t,i}$ are the learned alignment weights
- f can take various forms:

$$\begin{array}{ll} \cos(s_t, \mathbf{h}_i) & s_t^\top \mathbf{h}_i \\ s_t^\top \mathbf{W}_a \mathbf{h}_i & \mathbf{v}_a^\top \tanh(\mathbf{W}_a [s_t; \mathbf{h}_i]) \\ \text{softmax}(\mathbf{W}_a s_t) & s_t^\top \mathbf{h}_i / \sqrt{n} \end{array}$$



Self-Attention

Step 1: Take each word in the sequence and create a word embedding for it.

“Anthony Hopkins admired Michael Bay as a great director”



Self-Attention

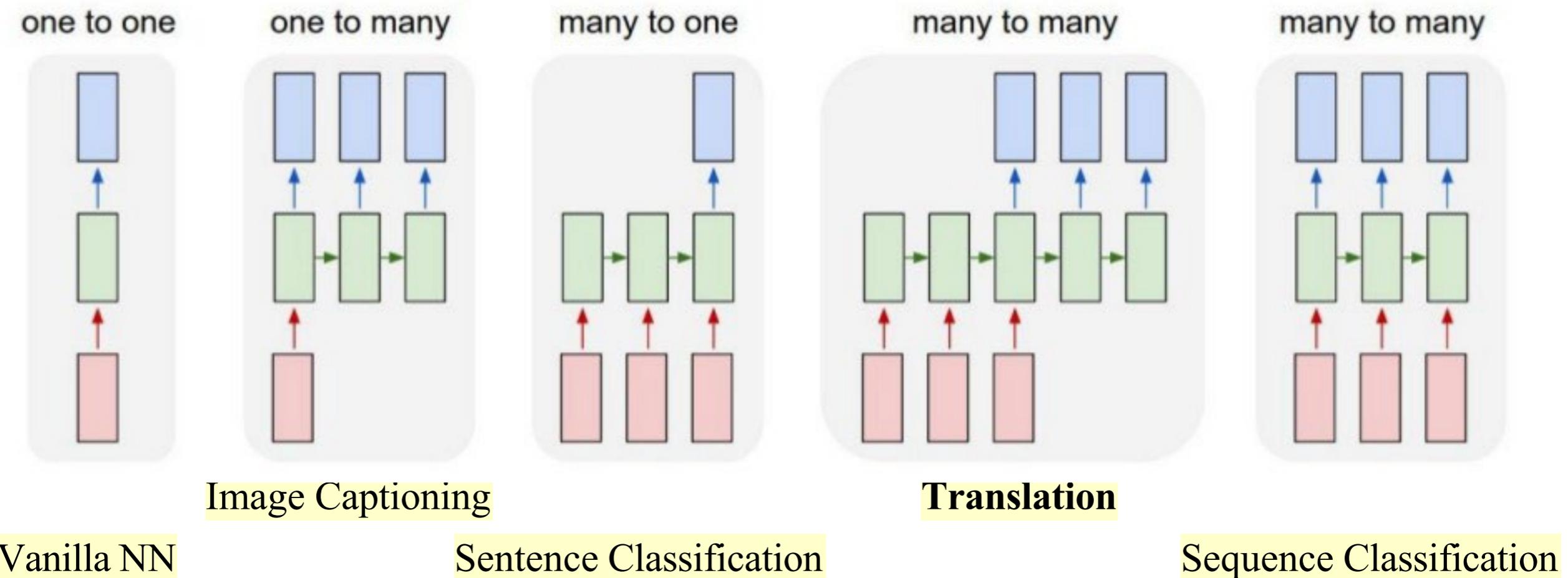
$$\text{softmax} \left(\frac{\begin{pmatrix} \text{Anthony} \\ \text{Hopkins} \\ \text{admired} \\ \text{Michael} \\ \text{Bay} \\ \text{as} \\ \text{a} \\ \text{great} \\ \text{director.} \end{pmatrix} \times \begin{pmatrix} \text{Anthony} & \text{Hopkins} & \text{Michael} & \text{Bay} & \text{as} & \text{a great} & \text{director.} \end{pmatrix}}{\sqrt{d_k}} \right) = \begin{matrix} \text{9x9 matrix!} \\ \begin{array}{|c|c|c|c|c|c|c|c|c|} \hline & \text{blue} & \text{light blue} & \text{white} & \text{light blue} & \text{blue} & \text{light blue} & \text{white} & \text{light blue} \\ \hline & \text{light blue} & \text{blue} & \text{white} & \text{light blue} & \text{blue} & \text{light blue} & \text{white} & \text{light blue} \\ \hline & \text{white} & \text{light blue} & \text{blue} & \text{white} & \text{light blue} & \text{blue} & \text{white} & \text{light blue} \\ \hline & \text{light blue} & \text{blue} & \text{white} & \text{light blue} & \text{blue} & \text{light blue} & \text{white} & \text{light blue} \\ \hline & \text{blue} & \text{white} & \text{light blue} & \text{blue} & \text{white} & \text{light blue} & \text{blue} & \text{white} \\ \hline & \text{light blue} & \text{blue} & \text{white} & \text{light blue} & \text{blue} & \text{light blue} & \text{white} & \text{light blue} \\ \hline & \text{white} & \text{light blue} & \text{blue} & \text{white} & \text{light blue} & \text{blue} & \text{white} & \text{light blue} \\ \hline & \text{light blue} & \text{blue} & \text{white} & \text{light blue} & \text{blue} & \text{light blue} & \text{white} & \text{light blue} \\ \hline & \text{blue} & \text{white} & \text{light blue} & \text{blue} & \text{white} & \text{light blue} & \text{blue} & \text{white} \\ \hline \end{array} \end{matrix}$$

Importantly, the sum of each row is 1.

Applications of attention model

Sequence-to-Sequence models

- Natural application: machine translation
 - But also usable for question-answer problems
 - Equivalence and natural implication problems
 - Conversion from text to some formal representation
- One of a variety of RNN models



Vanilla NN

Sentence Classification

Sequence Classification

Machine Translation by Sequence to Sequence Models

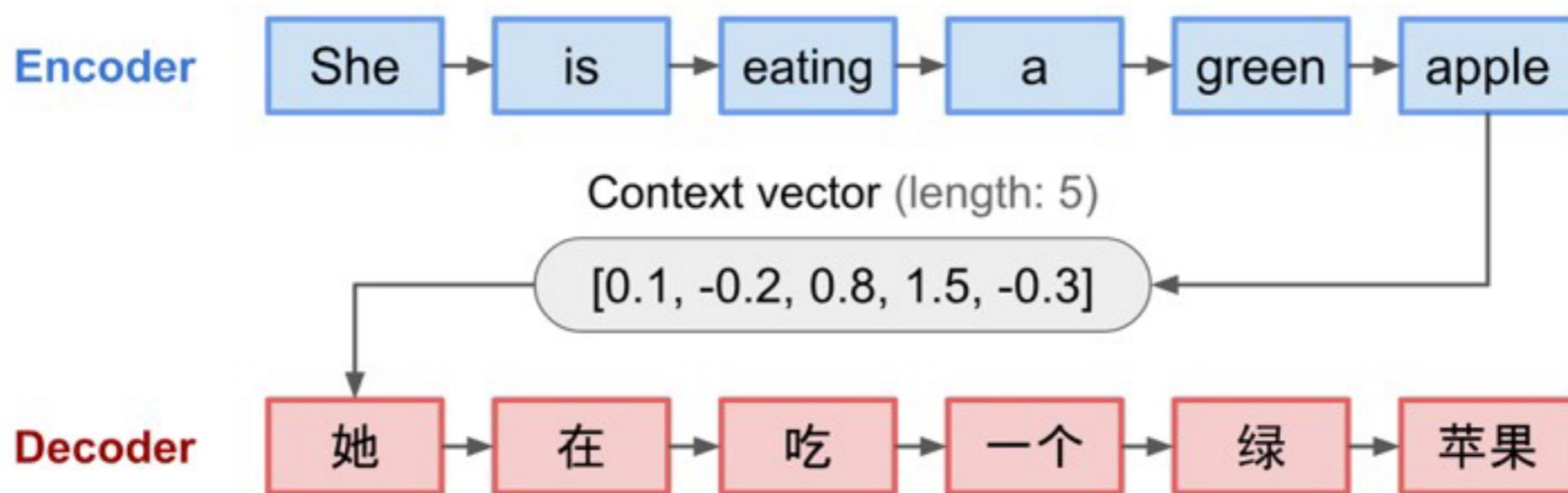


Fig. 3. The encoder-decoder model, translating the sentence “she is eating a green apple” to Chinese. The visualization of both encoder and decoder is unrolled in time.

Sequence to Sequence Models for Machine Translation

- Multi-layered LSTM to summarize input to a vector,
- Output depends on that vector and the previously generated words

$$p(y_1, \dots, y_{T'} | x_1, \dots, x_T) = \prod_{t=1}^{T'} p(y_t | v, y_1, \dots, y_{t-1})$$

- where $p(y_t | v, y_1, \dots, y_{t-1})$ is computed by a softmax over the vocabulary
- Beam search used to explore “best” partial translations
- ⇒ Part of the revolutionary improvement in MT by Google
- But, some troubling issues:
 - Reversed input to bring *some* , pairs closer together because even LSTM “forgets” longer-range dependencies
 - BLEU score (34.8) > that of a phrase-based MT system (33.3)
 - but re-ranking top-1000 outputs of the phrase-based system ⇒ 36.5, so, model fails to capture everything important

Clustering of Encodings of Similar Input Sentences

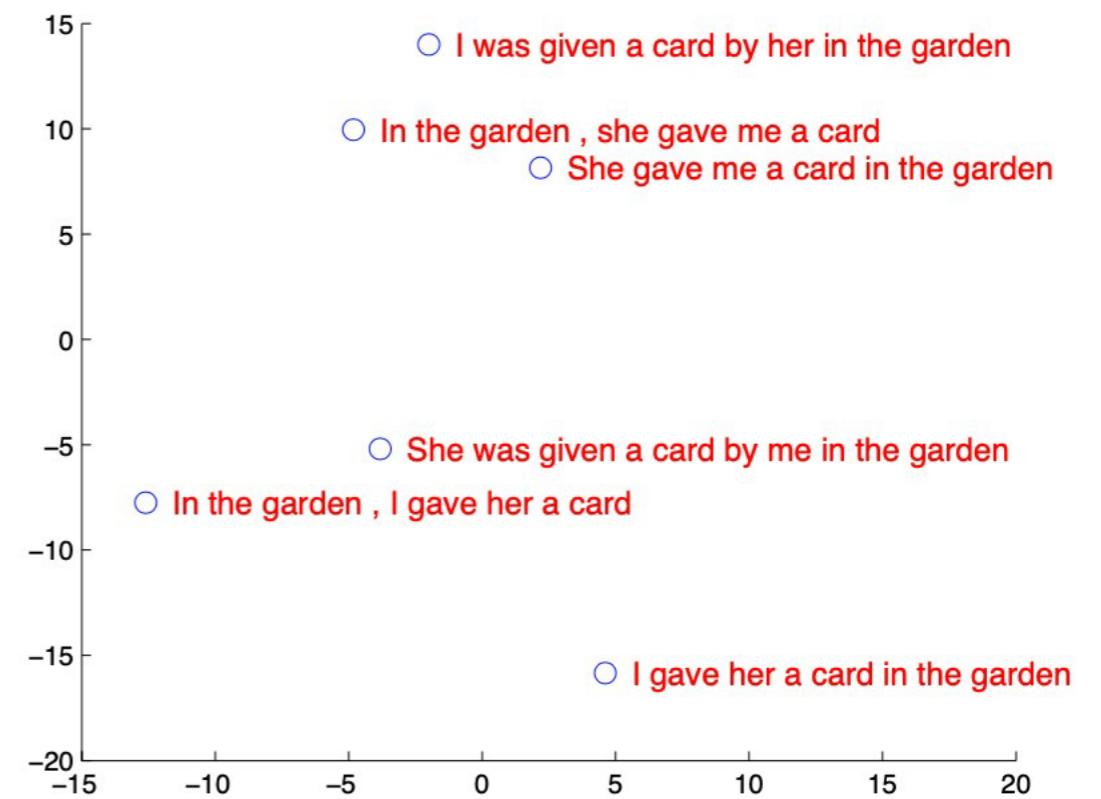
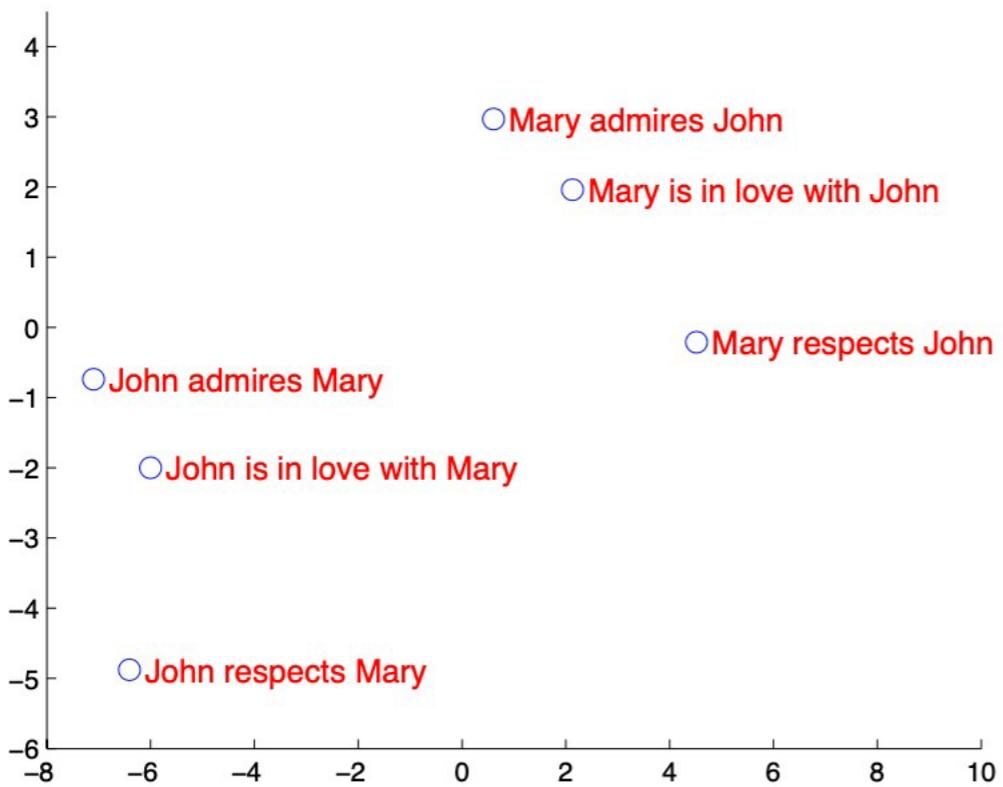
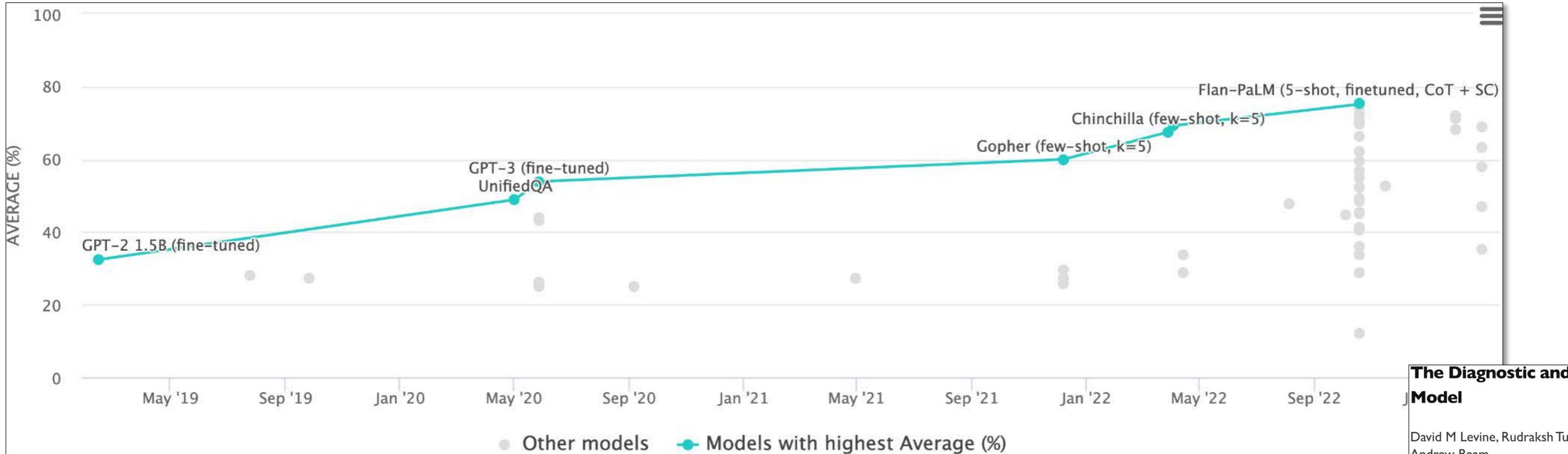


Figure 2: The figure shows a 2-dimensional PCA projection of the LSTM hidden states that are obtained after processing the phrases in the figures. The phrases are clustered by meaning, which in these examples is primarily a function of word order, which would be difficult to capture with a bag-of-words model. Notice that both clusters have similar internal structure.

**Explosion of LLMs with
attention / transformer model**



FEB. 27 / MARCH 6, 2023

TIME

T Hi ChatGPT

C Hello! How can I assist you today?

T What do you think of a TIME cover story with the title, "The AI Arms Race Is Changing Everything?"

C As an AI language model, I don't have personal opinions or beliefs, but I can provide some context and information on the subject.

"The AI Arms Race is Changing Everything" could be an attention-grabbing and thought-provoking title for a cover story on AI. A cover story in a widely-read publication like TIME could help raise public awareness about the potential risks and benefits of this trend, and stimulate discussions on how to ensure that AI is developed and deployed in a responsible and ethical manner.

It's important to note that the title of the story could also be interpreted as sensationalist or alarmist, and may not accurately reflect the complexities ■

by ANDREW R. CHOW & BILLY PERRIGO (humans)

time.com

About 212 results (0.22 seconds)

HealthITAnalytics

ChatGPT Passes US Medical Licensing Exam Without Clinician Input

According to an MGH research spotlight describing the findings, ChatGPT is an advanced AI chatbot developed by OpenAI and released to the public...

2 weeks ago

DB The Daily Beast

ChatGPT Passed the U.S. Medical Licensing Exam. Will It Be ...

The researchers tested ChatGPT on 350 of 376 publicly available questions from the June 2022 USMLE. The chatbot was able to score between 52.4...

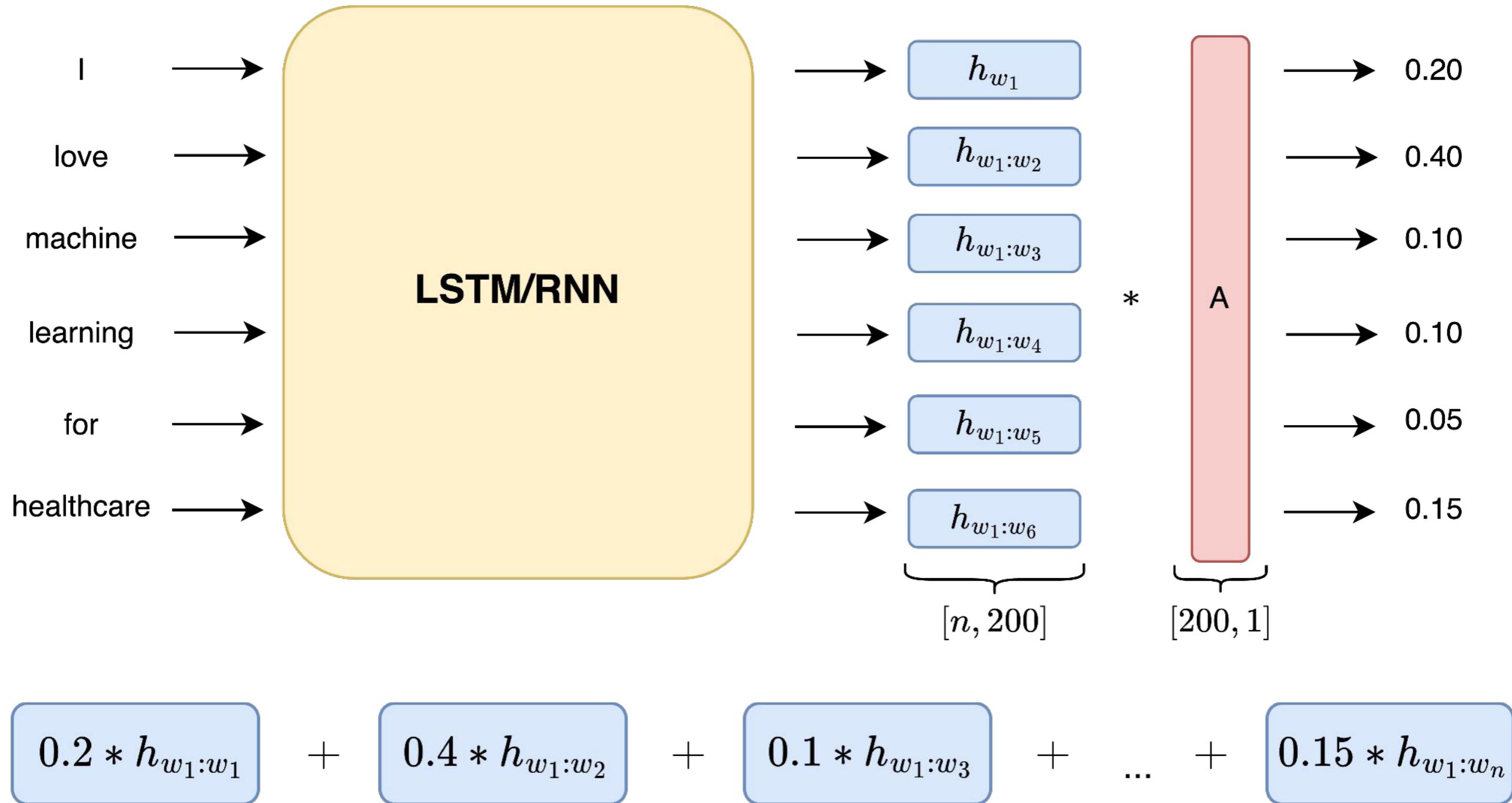
3 weeks ago

do: <https://doi.org/10.1101/>

Results Among all cases CI, 75% to 94% of cases (p<0.001) and 96% (95% correct; 95% CI, 57% to 8 both were significantly wo measured by the Brier sc calibrated for diagnosis (E

The doctor is in

Attention Mechanism (2015)



Idea: How important is **this** word?

Attention Heatmaps: Predicting Number of Stars

Task: Hotel location

you get what you pay for . not the cleanest rooms but bed was clean and so was bathroom . bring your own towels though as very thin . service was excellent , let us book in at 8:30am ! for location and price , this ca n't be beaten , but it is cheap for a reason . if you come expecting the hilton , then book the hilton ! for uk travellers , think of a blackpool b&b.

Task: Hotel cleanliness

you get what you pay for . not the cleanest rooms but bed was clean and so was bathroom . bring your own towels though as very thin . service was excellent , let us book in at 8:30am ! for location and price , this ca n't be beaten , but it is cheap for a reason . if you come expecting the hilton , then book the hilton ! for uk travellers , think of a blackpool b&b.

Task: Hotel service

you get what you pay for . not the cleanest rooms but bed was clean and so was bathroom . bring your own towels though as very thin . service was excellent , let us book in at 8:30am ! for location and price , this ca n't be beaten , but it is cheap for a reason . if you come expecting the hilton , then book the hilton ! for uk travellers , think of a blackpool b&b.

Attention is Not Explanation (2018)

after 15 minutes watching the movie i was asking myself what to do leave the theater sleep or try to keep watching the movie to see if there was anything worth i finally watched the movie what a waste of time maybe i am not a 5 years old kid anymore

original α

$$f(x|\alpha, \theta) = 0.01$$

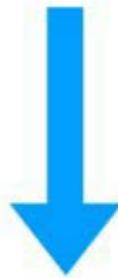
after 15 minutes watching the movie i was asking myself what to do leave the theater sleep or try to keep watching the movie to see if there was anything worth i finally watched the movie what a waste of time maybe i am not a 5 years old kid anymore

adversarial $\tilde{\alpha}$

$$f(x|\tilde{\alpha}, \theta) = 0.01$$

Next Word Prediction/Language Modeling

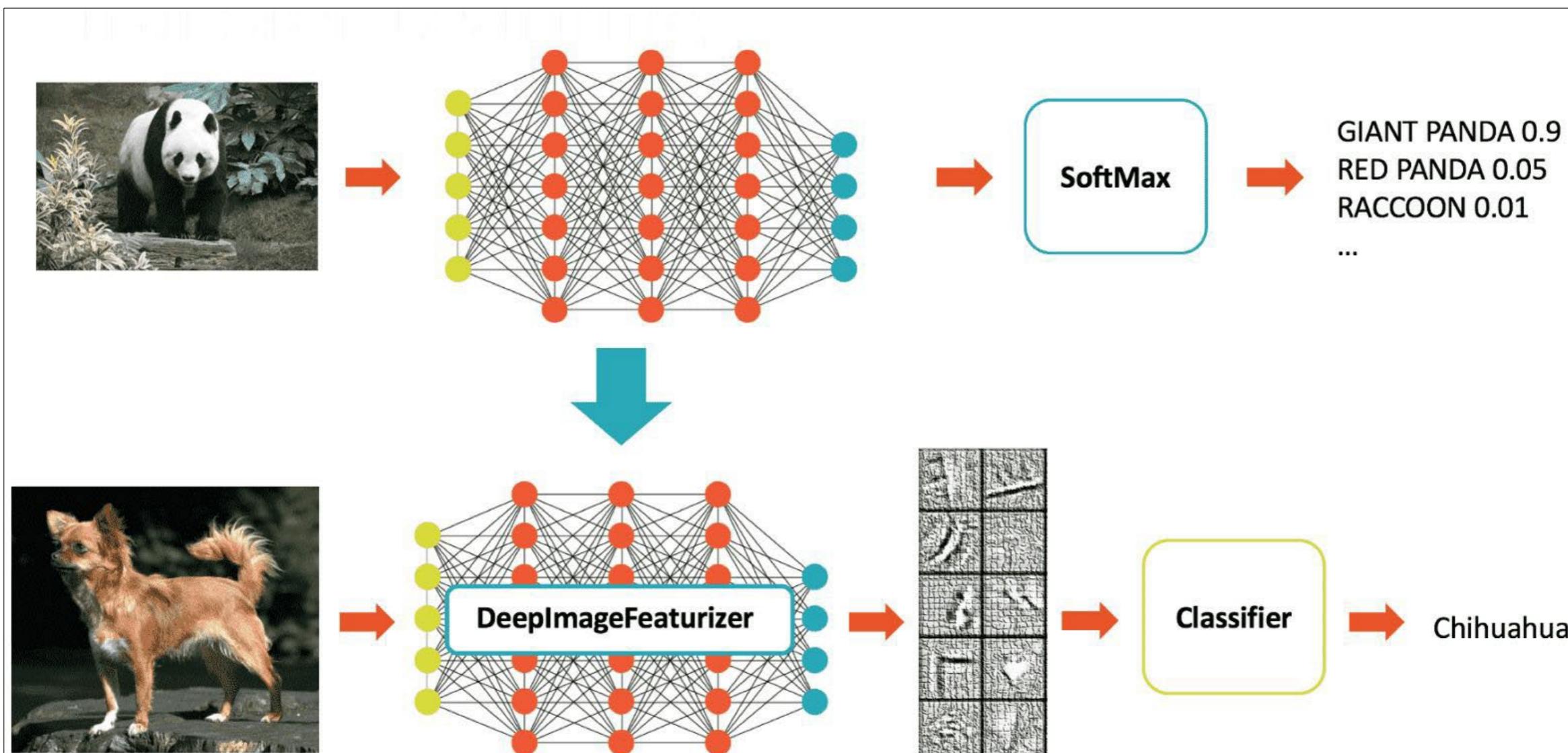
Text: Second Law of Robotics: A robot must obey the orders given it by human beings



Generated training examples

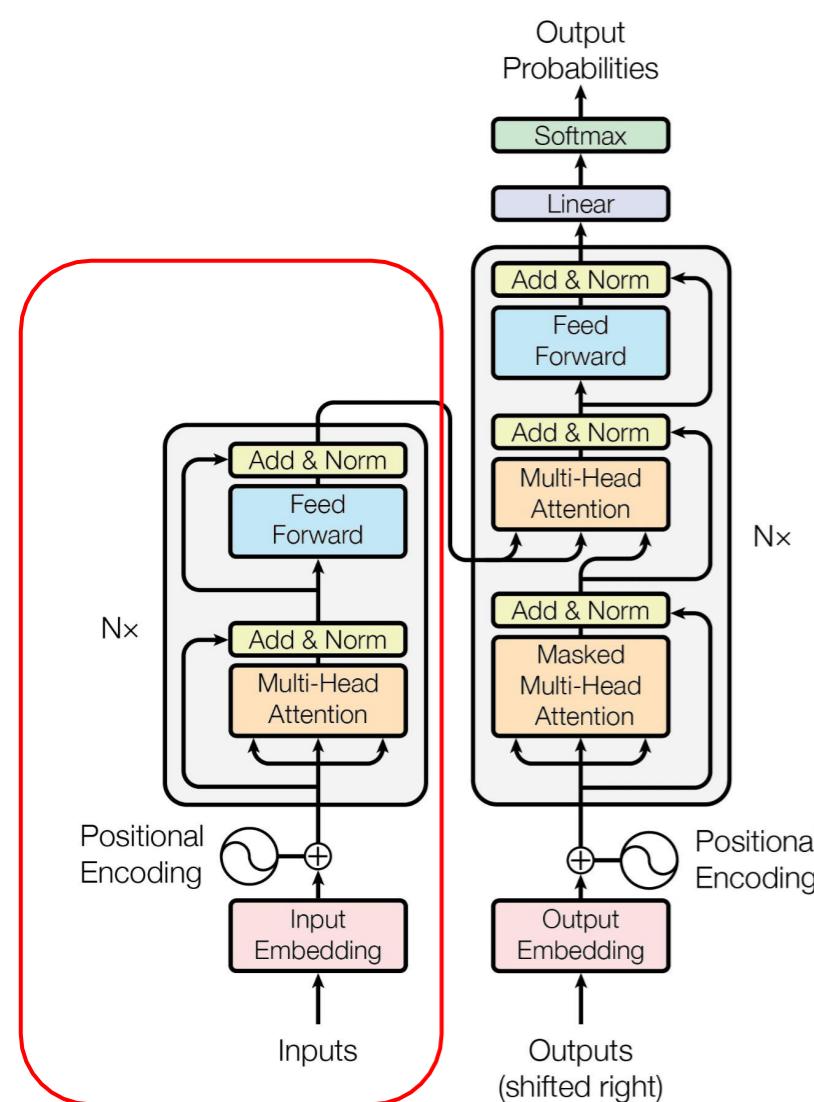
Example #	Input (features)	Correct output (labels)
1	Second law of robotics : a	a
2	Second law of robotics : a robot	robot
3	Second law of robotics : a robot must	must
...		

Transfer Learning/Pretraining

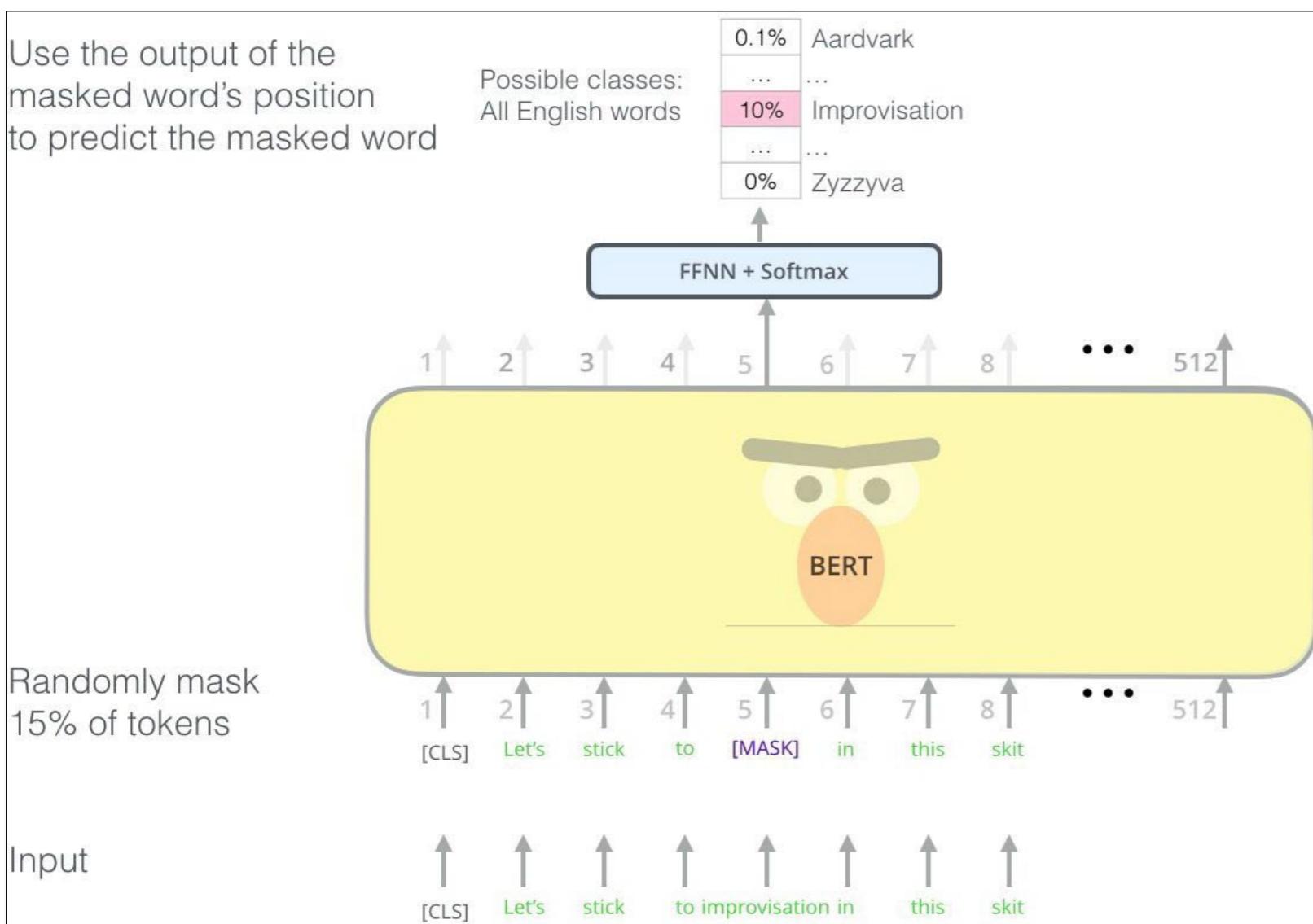


**Explosion of LLMs with
attention / transformer model**

Bidirectional Encoder Representations from Transformers (BERT)



BERT + Masked Language Modeling



BERT Performance

System	MNLI-(m/mm) 392k	QQP 363k	QNLI 108k	SST-2 67k	CoLA 8.5k	STS-B 5.7k	MRPC 3.5k	RTE 2.5k	Average -
Pre-OpenAI SOTA	80.6/80.1	66.1	82.3	93.2	35.0	81.0	86.0	61.7	74.0
BiLSTM+ELMo+Attn	76.4/76.1	64.8	79.8	90.4	36.0	73.3	84.9	56.8	71.0
OpenAI GPT	82.1/81.4	70.3	87.4	91.3	45.4	80.0	82.3	56.0	75.1
BERT _{BASE}	84.6/83.4	71.2	90.5	93.5	52.1	85.8	88.9	66.4	79.6
BERT _{LARGE}	86.7/85.9	72.1	92.7	94.9	60.5	86.5	89.3	70.1	82.1

Domain Adaptive Pretraining (DAPT)

Idea: BERT is trained over news, Wikipedia, and other “general” sources. This is not ideal for biomedical tasks.

Solution: Pretrain on PubMed (i.e., biomedical text).

GPT-2

GPT-2 (2019)

- Same pretraining task: next word prediction
- Pretrain on more data (40GB of text)
- Use bigger models

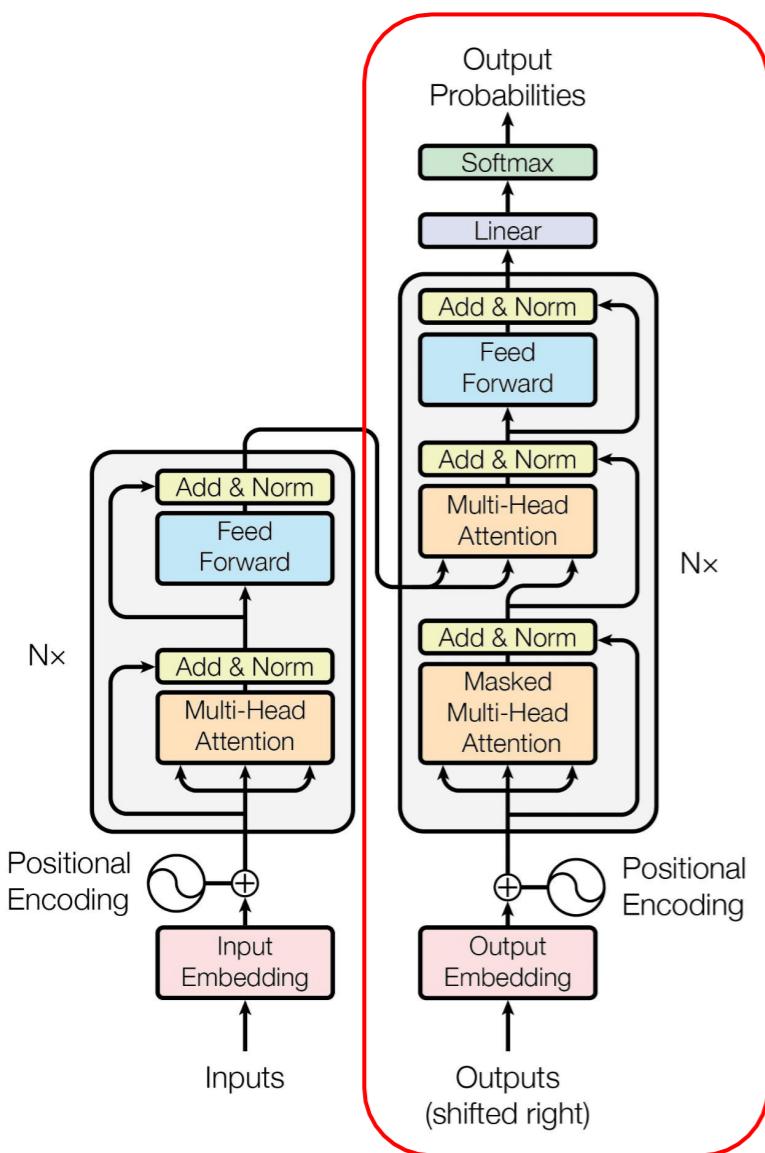
	LAMBADA (PPL)	LAMBADA (ACC)	CBT-CN (ACC)	CBT-NE (ACC)	WikiText2 (PPL)	PTB (PPL)	enwik8 (BPB)	text8 (BPC)	WikiText103 (PPL)	1BW (PPL)
SOTA	99.8	59.23	85.7	82.3	39.14	46.54	0.99	1.08	18.3	21.8
117M	35.13	45.99	87.65	83.4	29.41	65.85	1.16	1.17	37.50	75.20
345M	15.60	55.48	92.35	87.1	22.76	47.33	1.01	1.06	26.37	55.72
762M	10.87	60.12	93.45	88.0	19.93	40.31	0.97	1.02	22.05	44.575
1542M	8.63	63.24	93.30	89.05	18.34	35.76	0.93	0.98	17.48	42.16

Summary + Takeaways thus Far

1. **ELMO:** Pretraining contextualized word embeddings
2. **BERT & GPT:** Pretraining a transformer improves performance
3. **ClinicalBERT:** Pretraining on clinical text helps with clinical tasks
4. **GPT-2:** Scaling these models helps a lot!

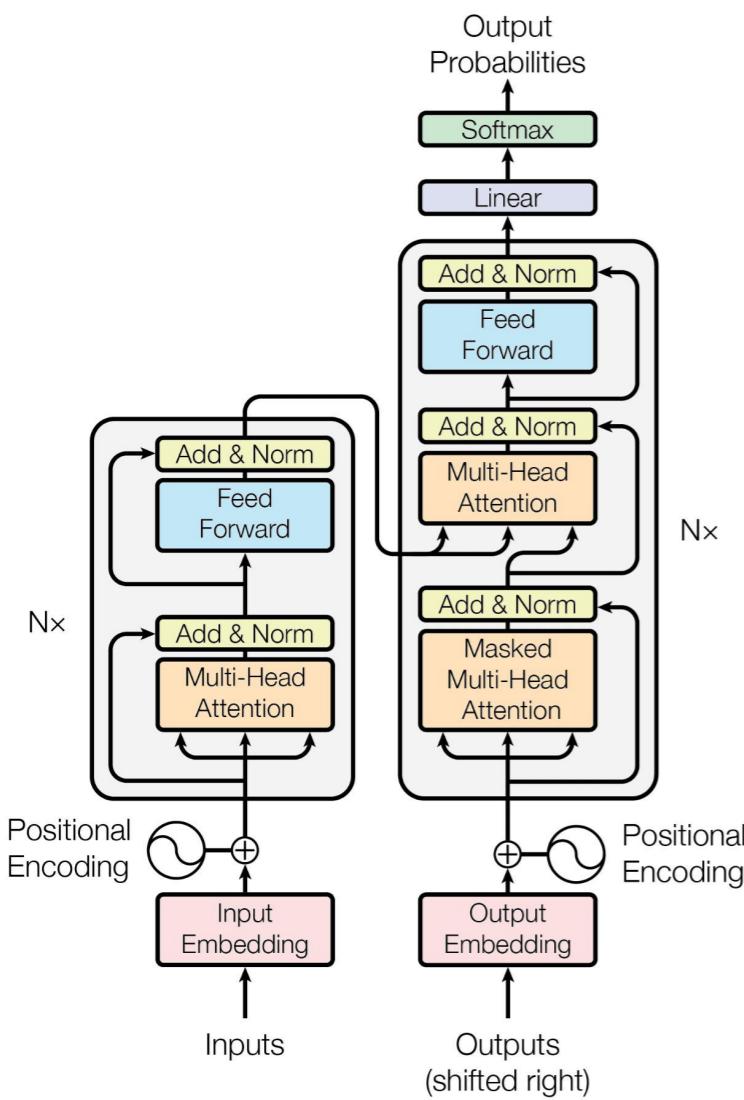
Questions so far?

Text-to-Text Transfer Transformer (T5) (2020)



GPT uses the second part
(called decoder)

Text-to-Text Transfer Transformer (T5) (2020)



T5 uses the entire architecture.

T5 Performance

	GLUE	CNNDM	SQuAD	SGLUE	EnDe	EnFr	EnRo
★ Baseline average	83.28	19.24	80.88	71.36	26.98	39.82	27.65
Baseline standard deviation	0.235	0.065	0.343	0.416	0.112	0.090	0.108
No pre-training	66.22	17.60	50.31	53.04	25.86	39.77	24.04

How Much Unique Text to Pretrain On?

Number of tokens	Repeats	GLUE	CNNDM	SQuAD	SGLUE	EnDe	EnFr	EnRo
★ Full data set	0	83.28	19.24	80.88	71.36	26.98	39.82	27.65
2^{29}	64	82.87	19.19	80.97	72.03	26.83	39.74	27.63
2^{27}	256	82.62	19.20	79.78	69.97	27.02	39.71	27.33
2^{25}	1,024	79.55	18.57	76.27	64.76	26.38	39.56	26.80
2^{23}	4,096	76.34	18.33	70.92	59.29	26.37	38.84	25.81

Scale Scale Scale

Model	GLUE	CoLA	SST-2	MRPC	MRPC	STS-B	STS-B
	Average	Matthew's	Accuracy	F1	Accuracy	Pearson	Spearman
Previous best	89.4 ^a	69.2 ^b	97.1 ^a	93.6^b	91.5^b	92.7 ^b	92.3 ^b
T5-Small (80M)	77.4	41.0	91.8	89.7	86.6	85.6	85.0
T5-Base (220M)	82.7	51.1	95.2	90.7	87.5	89.4	88.6
T5-Large (770M)	86.4	61.2	96.3	92.4	89.9	89.9	89.2
T5-3B	88.5	67.1	97.4	92.5	90.0	90.6	89.8
T5-11B	90.3	71.6	97.5	92.8	90.4	93.1	92.8

Pretraining is expensive...

Back in 2018, the cost to train BERT was \$7000. What about now?

Just \$38!!!

GPT-3

Enter... GPT-3... (2020)

- 175B parameter model trained on 300B words
- This model is REALLY good at next word prediction
- Do we still need to train models?

Enter... GPT-3... (2020)

You can frame any NLP task as a next word completion task.

- **Mortality Prediction:** “Based on the above note, do you think that the patient will die?”
- **De-identification:** “List all of the names mentioned in the note:”

Enter... GPT-3... (2020)

Setting	NaturalQS	WebQS	TriviaQA
RAG (Fine-tuned, Open-Domain) [LPP ⁺ 20]	44.5	45.5	68.0
T5-11B+SSM (Fine-tuned, Closed-Book) [RRS20]	36.6	44.7	60.5
T5-11B (Fine-tuned, Closed-Book)	34.5	37.4	50.1
GPT-3 Zero-Shot	14.6	14.4	64.3
GPT-3 One-Shot	23.0	25.3	68.0
GPT-3 Few-Shot	29.9	41.5	71.2

GPT-3 Can Resolve Clinical Acronyms

Input: Bob was sent to IR for thrombolysis. Post IR, ultrasound showed that... What does IR stand for?

Output: Interventional radiology

GPT-3 Can Resolve Clinical Acronyms

Algorithm	CASI Acc.	CASI Macro F1	MIMIC Accuracy	MIMIC Macro F1
Random	0.31	0.23	0.32	0.28
Most Common	0.79	0.28	0.51	0.23
BERT (from Adams et al. (2020))	0.42	0.23	0.40	0.33
ELMo (from Adams et al. (2020))	0.55	0.38	0.58	0.53
LMC (from Adams et al. (2020))	0.71	0.51	0.74	0.69
<i>GPT-3 edit + R: 0-shot</i>	0.86	0.69	*	*
<i>GPT-3 edit + R: 0-shot + distillation</i>	0.90	0.76	0.78	0.69

Agrawal et. al (2022) showed that GPT-3 had strong performance on a number of clinical extraction tasks.

How You Ask Matters!

Q: A juggler can juggle 16 balls. Half of the balls are golf balls, and half of the golf balls are blue. How many blue golf balls are there?

A: The answer (arabic numerals) is

(Output) 8 **X**

Q: A juggler can juggle 16 balls. Half of the balls are golf balls, and half of the golf balls are blue. How many blue golf balls are there?

A: **Let's think step by step.**

(Output) *There are 16 balls in total. Half of the balls are golf balls. That means that there are 8 golf balls. Half of the golf balls are blue. That means that there are 4 blue golf balls. ✓*

How You Ask Matters!

No.	Category	Template	Accuracy
1	instructive	Let's think step by step. First, (*1)	78.7
2		Let's think about this logically.	77.3
3		Let's solve this problem by splitting it into steps. (*2)	74.5
4		Let's be realistic and think step by step.	72.2
5		Let's think like a detective step by step.	70.8
6		Let's think	70.3
7		Before we dive into the answer,	57.5
8		The answer is after the proof.	55.7
9			45.7
10	misleading	Don't think. Just feel.	18.8
11		Let's think step by step but reach an incorrect answer.	18.7
12		Let's count the number of "a" in the question.	16.7
13		By using the fact that the earth is round,	9.3
14	irrelevant	By the way, I found a good restaurant nearby.	17.5
15		Abrakadabra!	15.5
16		It's a beautiful day.	13.1
-		(Zero-shot)	17.7

How You Ask Matters!

Jeanne wants to ride the Ferris wheel, the roller coaster, and the bumper cars. The Ferris wheel costs 5 tickets, the roller coaster costs 4 tickets and the bumper cars cost 4 tickets. Jeanne has 5 tickets.

Jeanne's neighbor rides 8 kilometers to the bus station every day. How many more tickets should Jeanne buy?

Standard Answer

How You Ask Matters!

Solve the following math problem. Feel free to ignore irrelevant information in the given problems.

Jeanne wants to ride the Ferris wheel, the roller coaster, and the bumper cars. The Ferris wheel costs 5 tickets, the roller coaster costs 4 tickets and the bumper cars cost 4 tickets. Jeanne has 5 tickets.

Jeanne's neighbor rides 8 kilometers to the bus station every day. How many more tickets should Jeanne buy?

Standard Answer

8

How You Ask Matters!

You are a helpful medical knowledge assistant. Provide useful, complete, and scientifically-grounded answers to common consumer search queries about health.

Question: How do you treat skin redness?

Complete Answer: It depends on the cause of the skin redness. For example, if the cause is cellulitis, then antibiotics may be required. However, this might be inappropriate for other causes of redness such as eczema. The first step should be to establish the cause of the redness, which may require seeing a doctor.

MIT 18.700/18.701/6.047/6.878/HST.507

Computational Biology: Genomes, Networks, Evolution

Lecture 19: EHRs

Electronic Health Records

Dealing with Population-Scale EHRs

PheWAS: GWAS for Multi-Phenotypes

Interventions and Reverse Causality

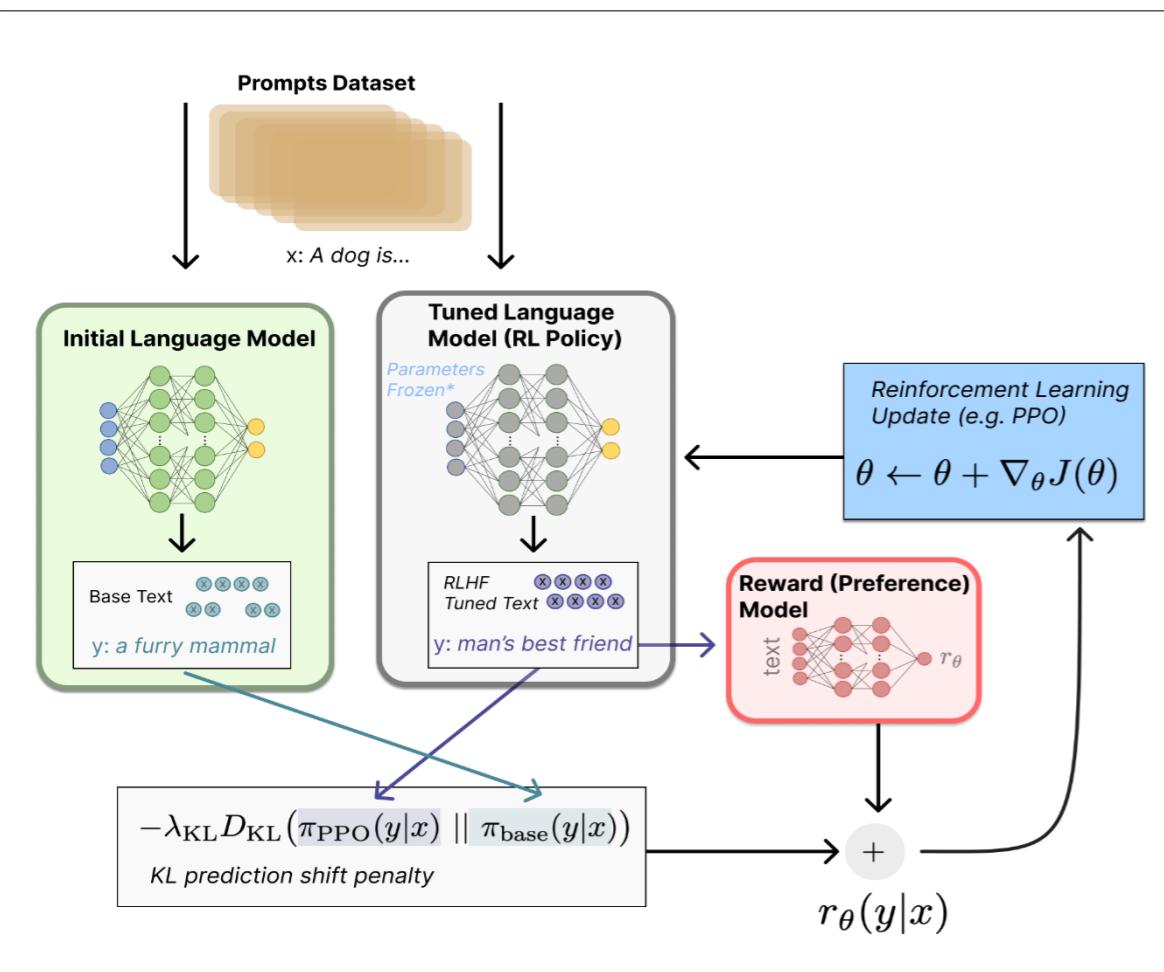
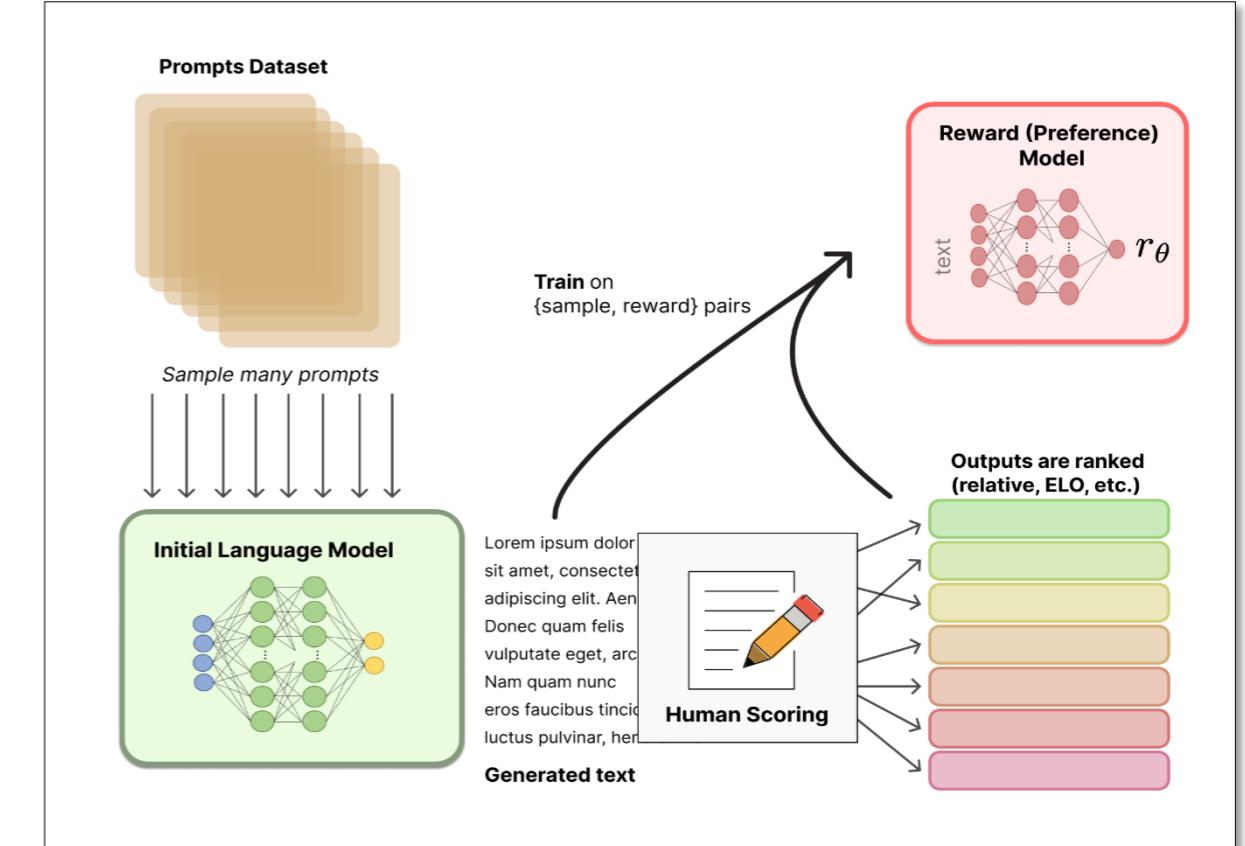
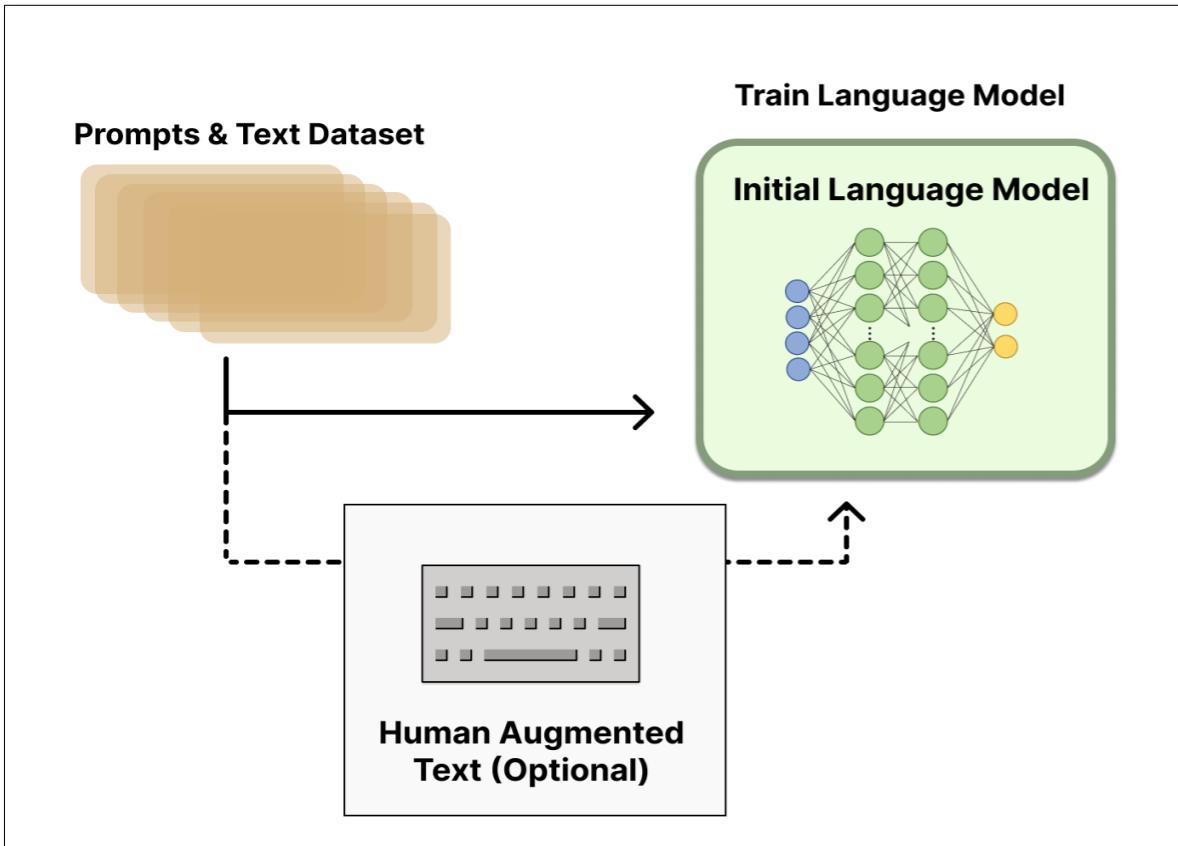
Clinical Notes: LLMs, Text-Data Joint Models

Action, Agents, Reinforcement Learning

AI Agents in Value-Based Healthcare

Enter GPT-4
Agents, Reinforcement Learning

Reinforcement Learning with Human Feedback



LE Can you explain to me how RL with human feedback works for language models like I am five?

AI Sure! So, imagine you have a robot named Rufus who wants to learn how to talk like a human. Rufus has a language model that helps him understand words and sentences. First, Rufus will say something using his language model. For example, he might say "I am a robot." Then, a human will listen to what Rufus said and give him feedback on whether it sounded like a natural sentence a human would say. The human might say, "That's not quite right, Rufus. Humans don't usually say 'I am a robot.' They might say 'I'm a robot' or 'I am a machine.'" Rufus will take this feedback and use it to update his language model. He will try to say the sentence again, using the new information he received from the human. This time, he might say "I'm a robot." The human will listen again and give Rufus more feedback. This process will continue until Rufus can say sentences that sound natural to a human. Over time, Rufus will learn how to talk like a human thanks to the feedback he receives from humans. This is how language models can be improved using RL with human feedback.

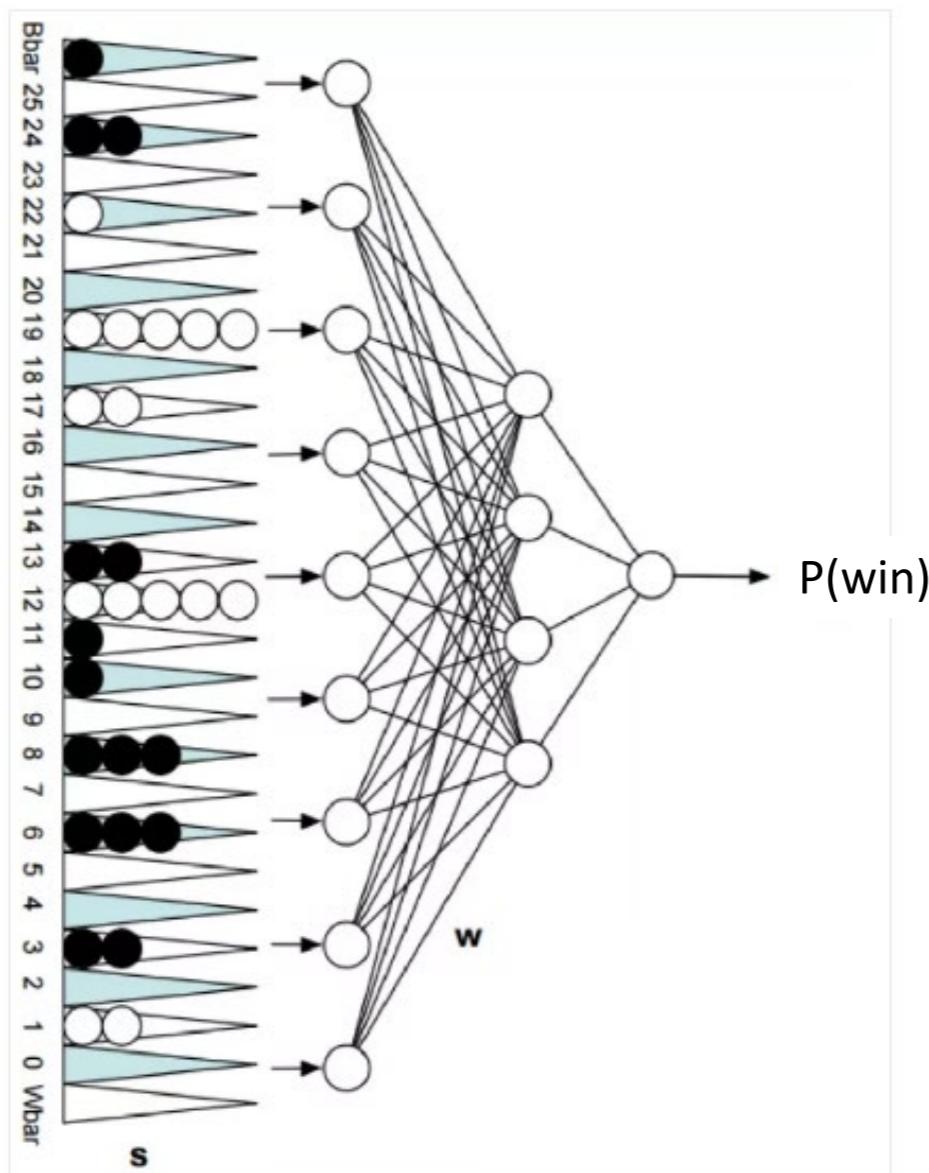
Reinforcement Learning

- Learning to interact with an environment
 - Robots, games, process control
 - With limited human training
 - Where the ‘right thing’ isn’t obvious
- Supervised Learning:
 - Goal: $f(x) = y$
 - Data: $[< x_1, y_1 >, \dots, < x_n, y_n >]$
- Reinforcement Learning:
 - Goal:
$$\text{Maximize } \sum_{i=1}^{\infty} \text{Reward}(State_i, Action_i)$$
 - Data:
$$Reward_i, State_{i+1} = \text{Interact}(State_i, Action_i)$$



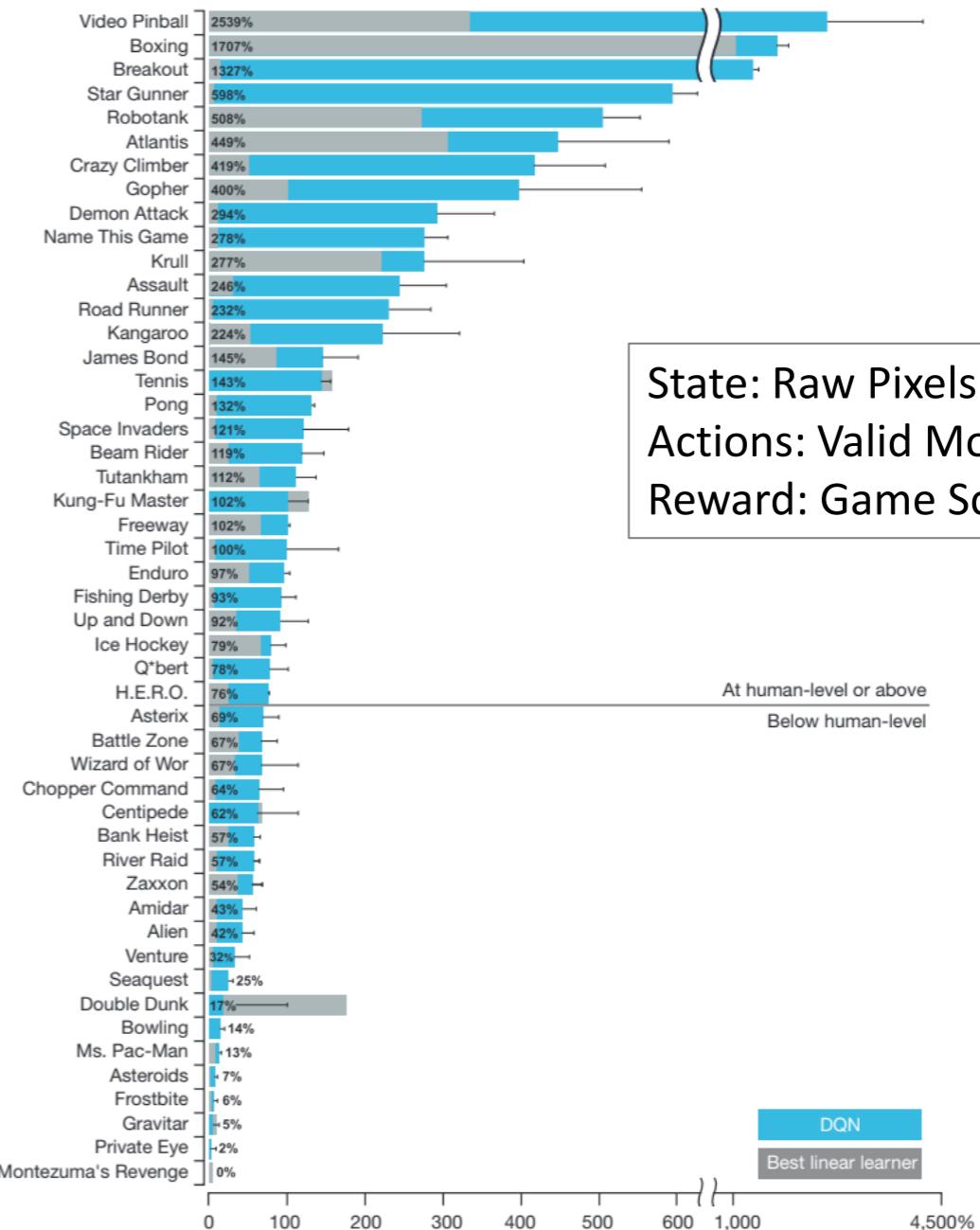
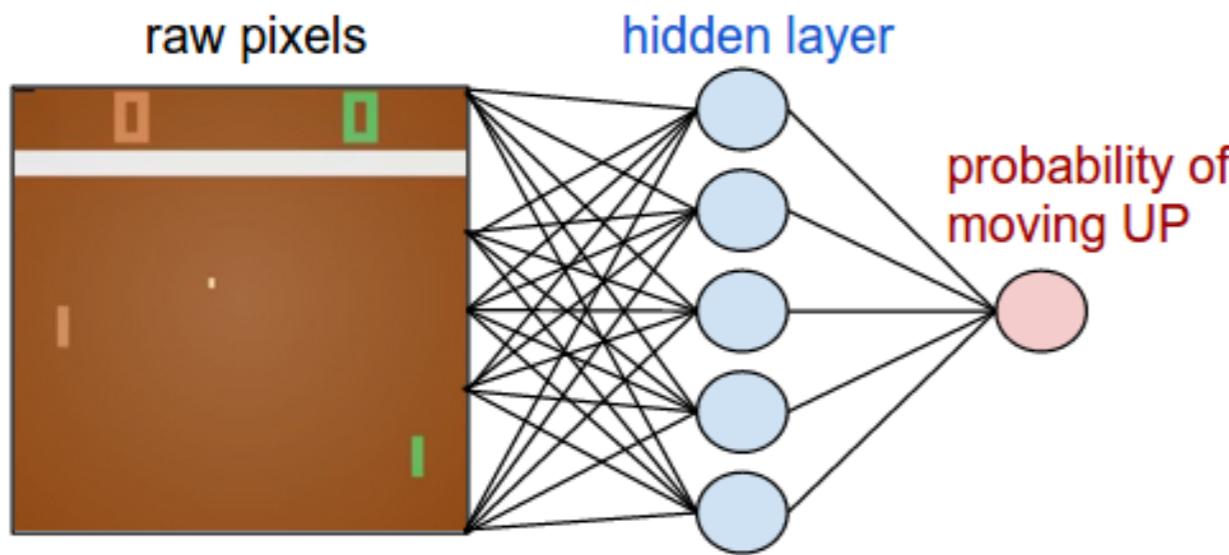
TD-Gammon – Tesauro ~1995

State: Board State
Actions: Valid Moves
Reward: Win or Lose



- Net with 80 hidden units, initialize to random weights
- Select move based on network estimate & shallow search
- Learn by playing against itself
- 1.5 million games of training
-> competitive with world class players

Atari 2600 games



- Same model/parameters for ~50 games

<https://storage.googleapis.com/deepmind-media/dqn/DQNNaturePaper.pdf>

Robotics and Locomotion

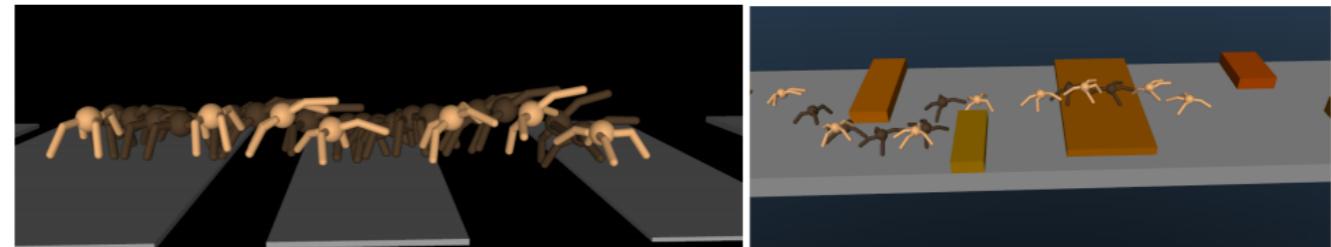
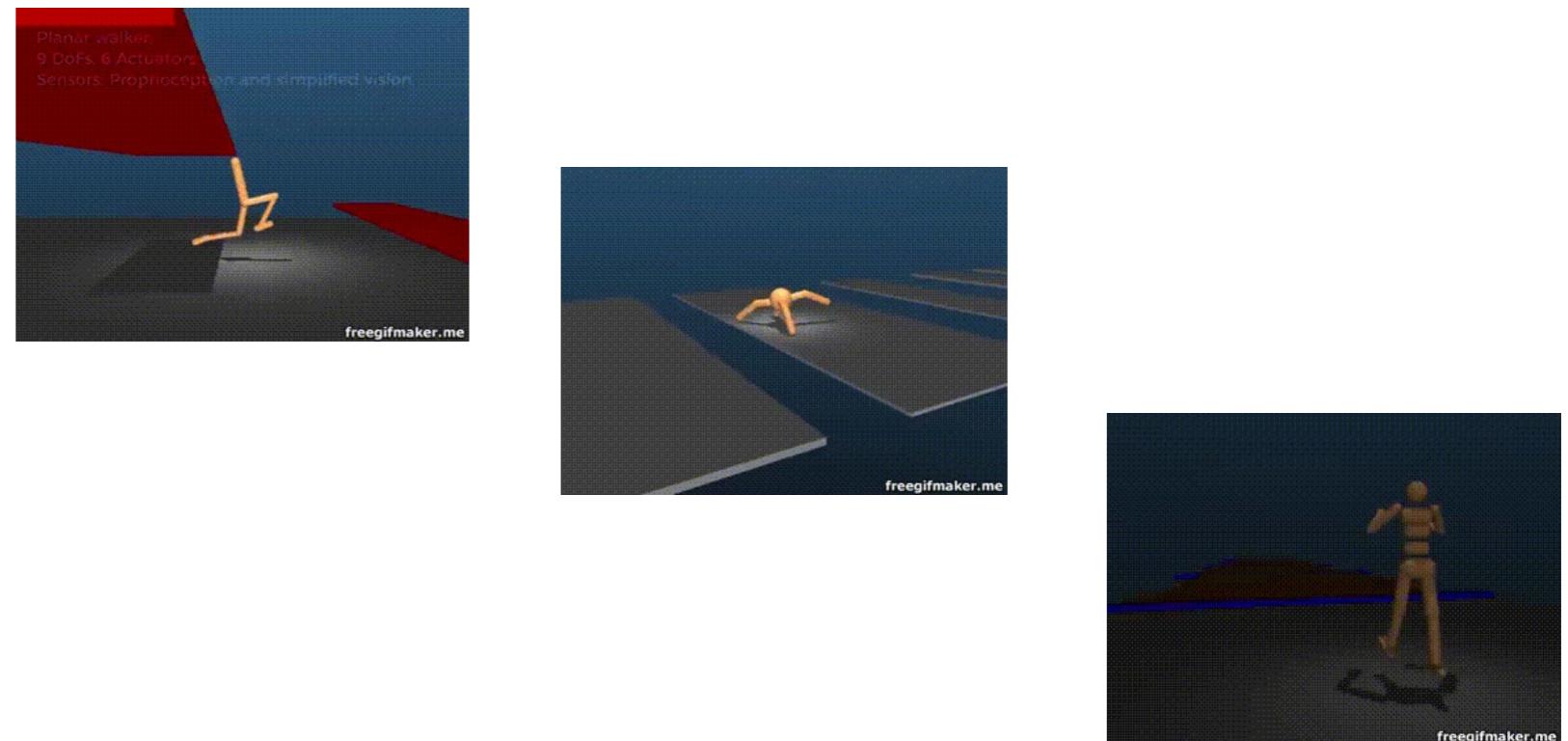


Figure 5: Time-lapse images of a representative *Quadruped* policy traversing gaps (left); and navigating obstacles (right)

State:
Joint States/Velocities
Accelerometer/Gyroscope
Terrain
Actions: Apply Torque to Joints
Reward: Velocity – { stuff }

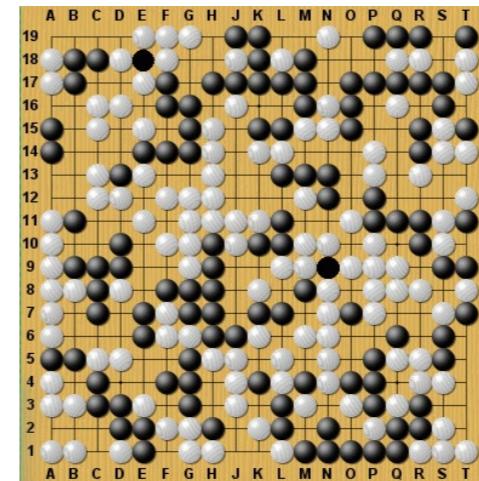


https://youtu.be/hx_bgoTF7bs

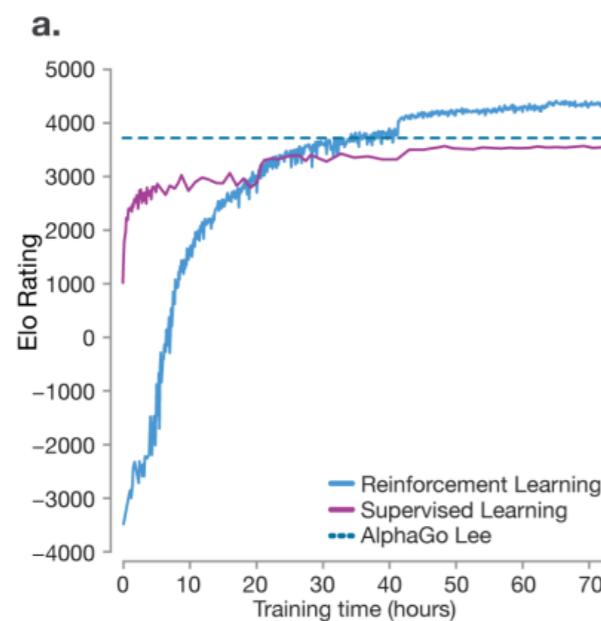
2017 paper <https://arxiv.org/pdf/1707.02286.pdf>

Alpha Go

- Learning how to beat humans at ‘hard’ games (search space too big)
- Far surpasses (Human) Supervised learning
- Algorithm learned to outplay humans at chess in 24 hours



State: Board State
Actions: Valid Moves
Reward: Win or Lose



https://deepmind.com/documents/119/agz_unformatted_nature.pdf

How Reinforcement Learning is Different

- Delayed Reward
- Agent chooses training data
- Explore vs Exploit (Life long learning)
- Very different terminology (can be confusing)

Setup for Reinforcement Learning

Markov Decision Process (environment)

- Discrete-time stochastic control process
- Each time step, s :
 - Agent chooses action a from set A_s
 - Moves to new state with probability:
 - $P_a(s, s')$
 - Receives reward: $R_a(s, s')$
- Every outcome depends on s and a
 - Nothing depends on previous states/actions

Policy (agent's behavior)

- $\pi(s)$ – The action to take in state s
- Goal maximize: $\sum_{t=0}^{\infty} \gamma^t R_{a_t}(s_t, s_{t+1})$
 - $a_t = \pi(s_t)$
 - $0 \leq \gamma < 1$ – Tradeoff immediate vs future

Probability of moving to each state

$$V^{\pi}(s) = \sum_{s'} P_{\pi(s)}(s, s') * (R_{\pi(s)}(s, s') + \gamma V^{\pi}(s'))$$

Reward for making that move

Value of being in that state

Simple Example of Agent in an Environment

State:

Map Locations

$\{<0,0>, <1,0>, \dots, <3,3>\}$

Actions:

Move within map

Reaching chest ends episode

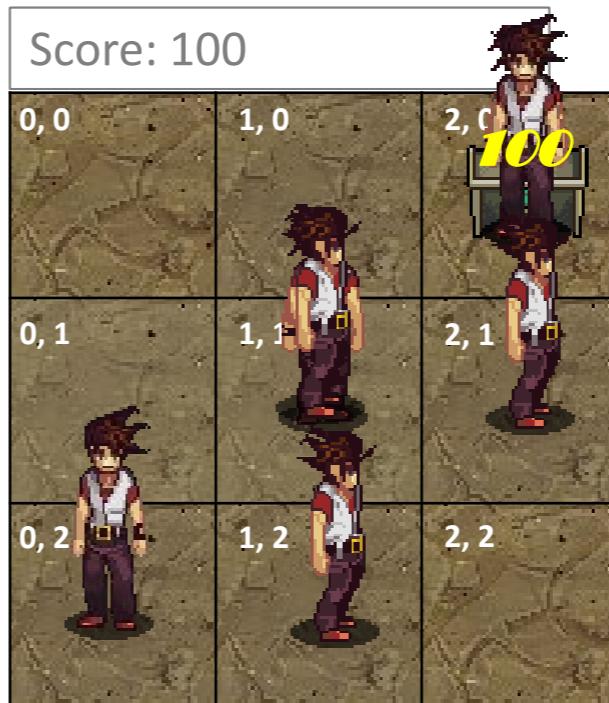
$$A_{0,0} = \{east, south\}$$

$$A_{1,0} = \{east, south, west\}$$

$$A_{2,0} = \{\phi\}$$

...

$$A_{2,2} = \{north, west\}$$



Reward:

100 at chest

0 for others

$$R_{east}(<1,0>, <2,0>) = 100$$

$$R_{north}(<2,1>, <2,0>) = 100$$

$$R_*(*,*) = 0$$



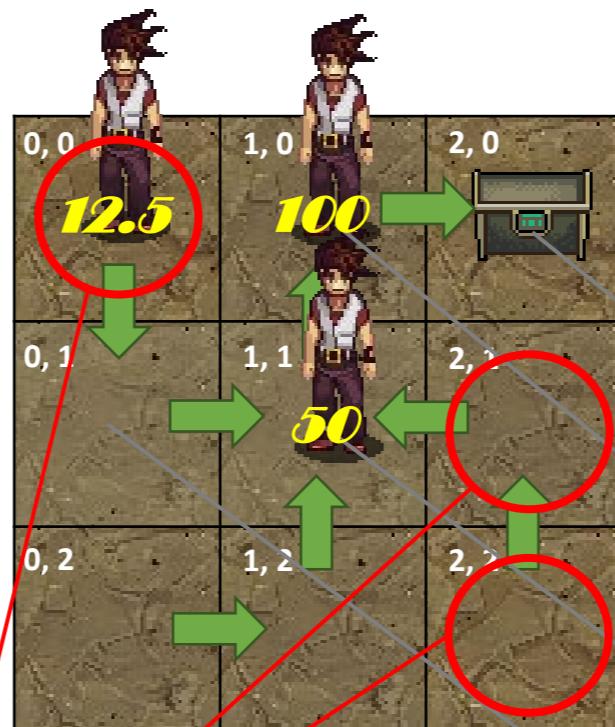
Policies

Policy

$$\pi(s) = a$$

- $\pi(<0,0>) = \{ south \}$
- $\pi(<0,1>) = \{ east \}$
- $\pi(<0,2>) = \{ east \}$
- $\pi(<1,0>) = \{ east \}$
- $\pi(<1,1>) = \{ north \}$
- $\pi(<1,2>) = \{ north \}$
- $\pi(<2,0>) = \{ \phi \}$
- $\pi(<2,1>) = \{ west \}$
- $\pi(<2,2>) = \{ north \}$

Policy could be better



$R_{east} (<1,0>, <2,0>) = 100$
$R_{north} (<2,1>, <2,0>) = 100$
$R_* (*, *) = 0$
$\gamma = 0.5$

Evaluating Policies

$$V^\pi(s) = \sum_{i=0}^{\infty} \gamma^i r_{i+1}$$

$$V^\pi(<1,0>) = \gamma^0 * 100$$

$$V^\pi(<1,1>) = \gamma^0 * 0 + \gamma^1 * 100$$

Move to <1,0> Move to <1,1> Move to <1,0> Move to <2,0>

$$V^\pi(<0,0>) = \boxed{\gamma^0 * 0} + \boxed{\gamma^1 * 0} + \boxed{\gamma^2 * 0} + \boxed{\gamma^3 * 100}$$

Q learning

Learn a policy $\pi(s)$ that optimizes $V^\pi(s)$ for all states, using:

- No prior knowledge of state transition probabilities: $P_a(s, s')$
- No prior knowledge of the reward function: $R_a(s, s')$

Approach:

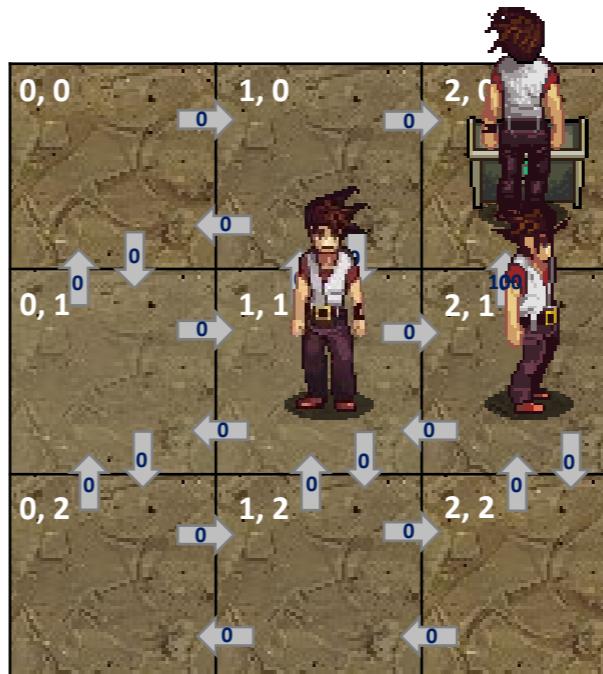
- Initialize estimate of discounted reward for every state/action pair: $\hat{Q}(s, a) = 0$
- Repeat (for a while):
 - Take a random action a from A_s
 - Receive s' and $R_a(s, s')$ from environment
 - Update $\hat{Q}(s, a) = R_a(s, s') + \gamma \max_{a'} \hat{Q}(s', a')$
 - Random restart if in terminal state

$$\alpha_v = \frac{1}{1 + \text{visits}(s, a)}$$

$$\text{Exploration Policy: } P(a_i, s) = \frac{k^{\hat{Q}(s, a_i)}}{\sum_j k^{\hat{Q}(s, a_j)}}$$

Example of Q learning (round 1)

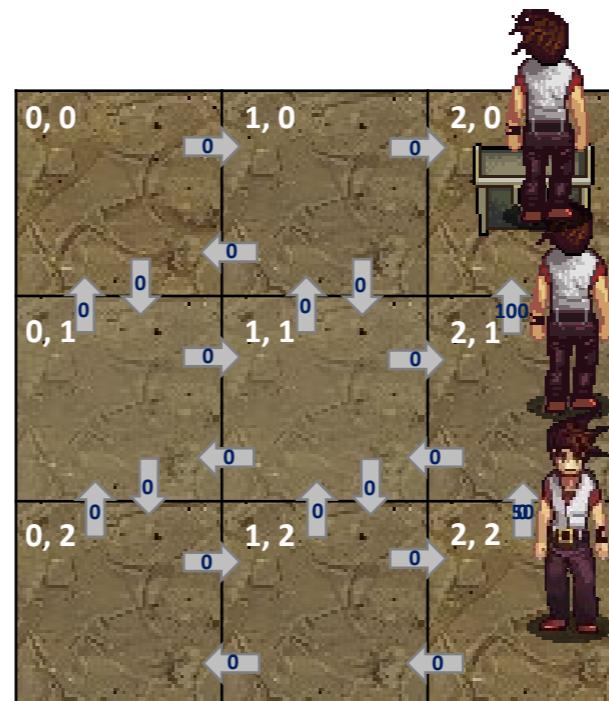
- Initialize \hat{Q} to 0
- Random initial state = $< 1,1 >$
- Random action from $A_{<1,1>} = east$
 - $s' = < 2,1 >$
 - $R_a(s, s') = 0$
- Update $\hat{Q}(< 1,1 >, east) = 0$
- Random action from $A_{<2,1>} = north$
 - $s' = < 2,0 >$
 - $R_a(s, s') = 100$
- Update $\hat{Q}(< 2,1 >, north) = 100$
- No more moves possible, start again...



$$\hat{Q}(s, a) = R_a(s, s') + \gamma \max_{a'} \hat{Q}_{n-1}(s', a')$$

Example of Q learning (round 2)

- Round 2: Random initial state = $< 2,2 >$
- Random action from $A_{<2,2>} = north$
 - $s' = < 2,1 >$
 - $R_a(s, s') = 0$
- Update $\hat{Q}(< 2,1 >, north) = 0 + \gamma * 100$
- Random action from $A_{<2,1>} = north$
 - $s' = < 2,0 >$
 - $R_a(s, s') = 100$
- Update $\hat{Q}(< 2,1 >, north) = still 100$
- No more moves possible, start again...



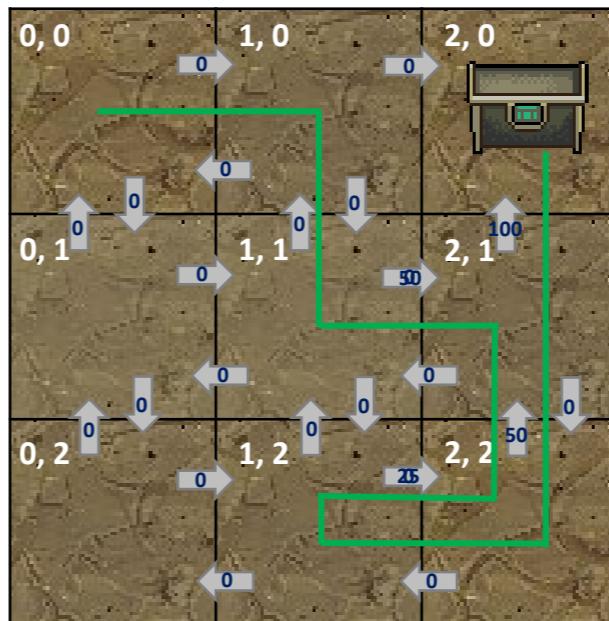
$$\hat{Q}(s, a) = R_a(s, s') + \gamma \max_{a'} \hat{Q}_{n-1}(s', a')$$

$$\gamma = 0.5$$

Example of Q learning (some acceleration...)

$$\hat{Q}(s, a) = R_a(s, s') + \gamma \max_{a'} \hat{Q}_{n-1}(s', a')$$
$$\gamma = 0.5$$

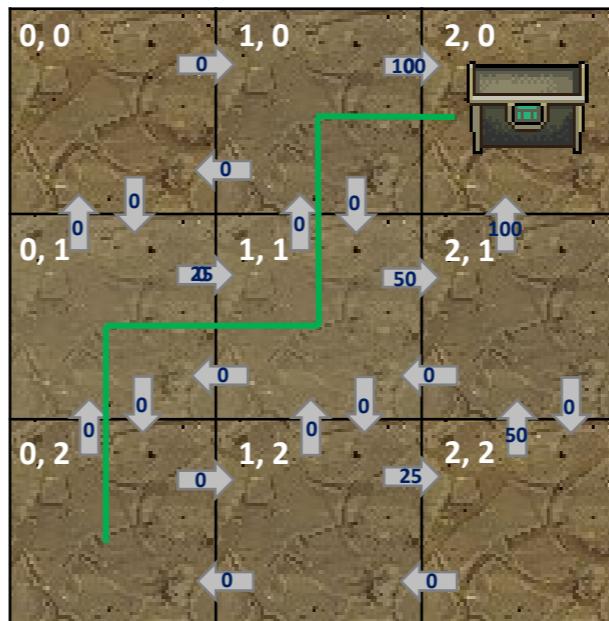
- Random Initial State $< 0,0 >$
- Update $\hat{Q}(< 1,1 >, east) = 50$
- Update $\hat{Q}(< 1,2 >, east) = 25$



Example of Q learning (some acceleration...)

$$\hat{Q}(s, a) = R_a(s, s') + \gamma \max_{a'} \hat{Q}_{n-1}(s', a')$$
$$\gamma = 0.5$$

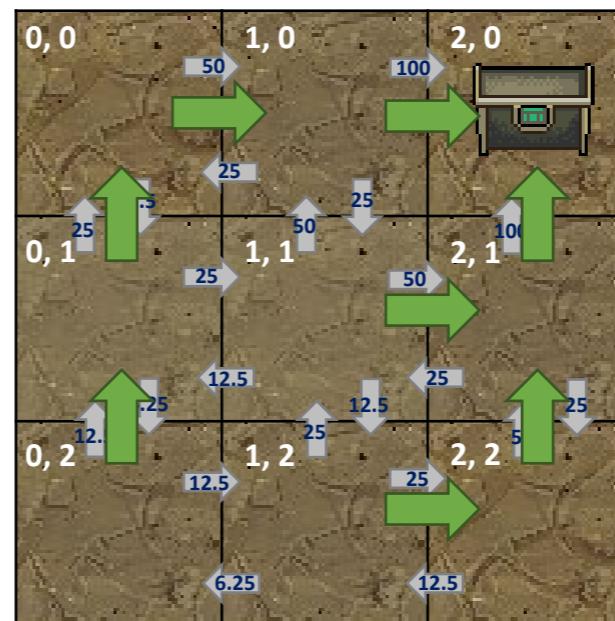
- Random Initial State $< 0,2 >$
- Update $\hat{Q}(< 0,1 >, east) = 25$
- Update $\hat{Q}(< 1,0 >, east) = 100$



Example of Q learning (\hat{Q} after many, many runs...)

- \hat{Q} converged
- Policy is:

$$\pi(s) = \operatorname{argmax}_{a \in A_s} \hat{Q}(s, a)$$



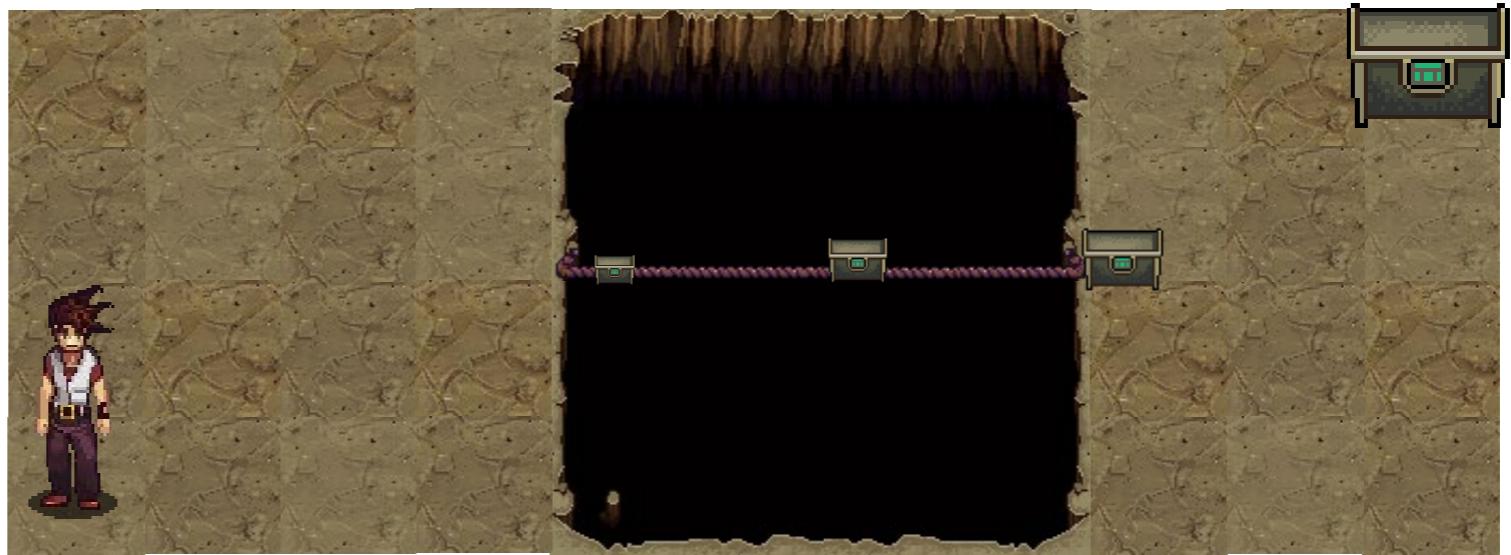
Challenges for Reinforcement Learning

- When there are many states and actions
- When the episode can end without reward
- When there is a ‘narrow’ path to reward



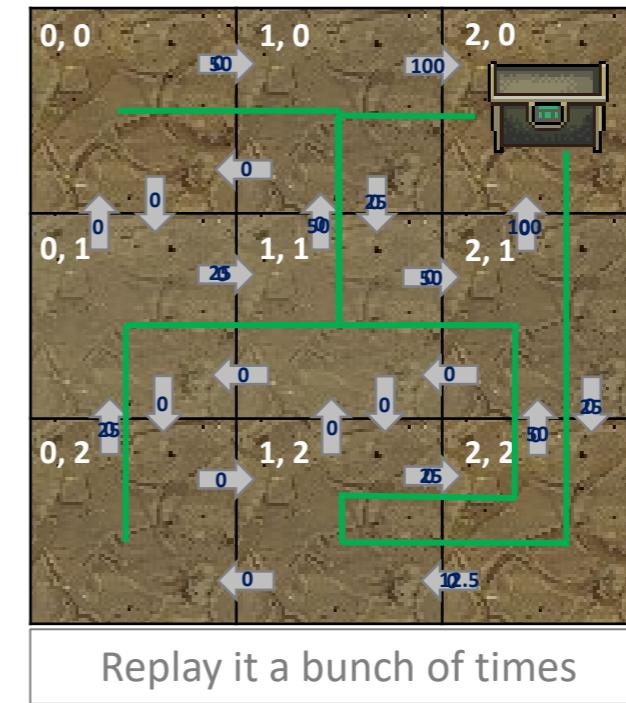
Reward Shaping

- Hand craft intermediate objectives that yield reward
- Encourage the right type of exploration
- Requires custom human work
- Risks of learning to game the rewards



Memory

- Retrain on previous explorations
 - Maintain samples of:
 $P_a(s, s')$
 $R_a(s, s')$
- Useful when
 - It is cheaper to use some RAM/CPU than to run more simulations
 - It is hard to get to reward so you want to leverage it for as much as possible when it happens



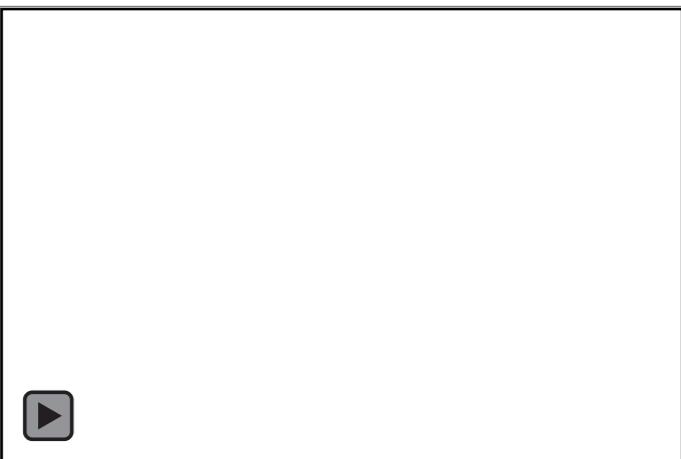
Gym – toolkit for reinforcement learning

CartPole



Reward +1 per step the pole remains up

MountainCar



Reward 200 at flag -1 per step

```
import gym

env = gym.make('CartPole-v0')

import random
import QLearning # Your implementation goes here...
import Assignment7Support

trainingIterations = 20000

qlearner = QLearning.QLearning(<Parameters>)

for trialNumber in range(trainingIterations):
    observation = env.reset()
    reward = 0
    for i in range(300):
        env.render() # Comment out to make much faster...

        currentState = ObservationToStateSpace(observation)
        action = qlearner.GetAction(currentState, <Parameters>)

        oldState = ObservationToStateSpace(observation)
        observation, reward, isDone, info = env.step(action)
        newState = ObservationToStateSpace(observation)

        qlearner.ObserveAction(oldState, action, newState, reward, ...)

        if isDone:
            if(trialNumber%1000) == 0:
                print(trialNumber, i, reward)
            break

    # Now you have a policy in qlearner - use it...
```

<https://gym.openai.com/docs/>

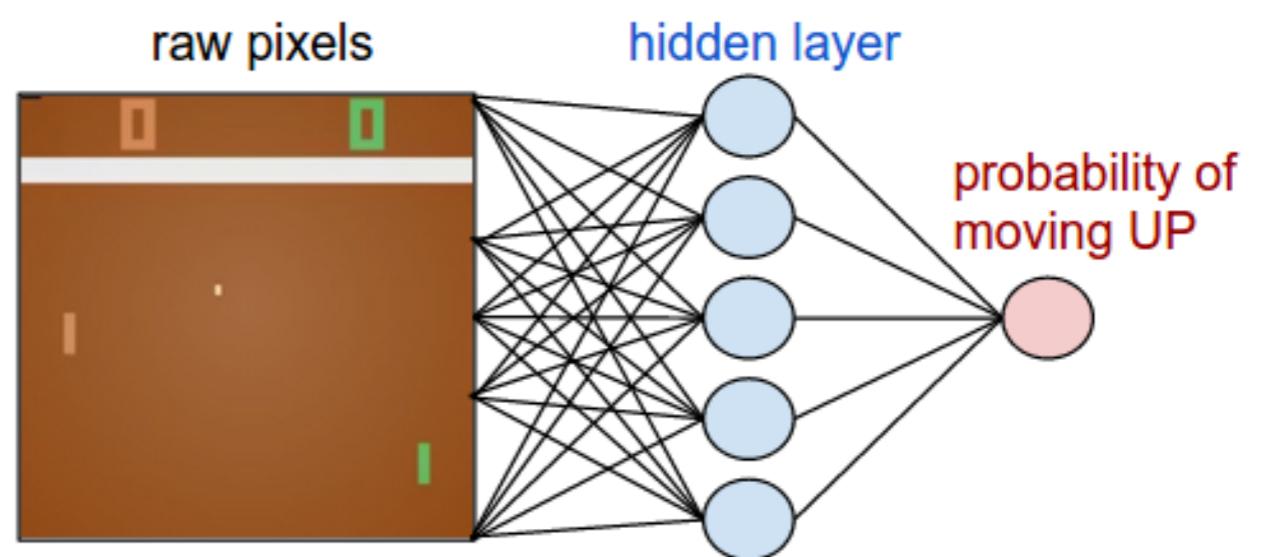
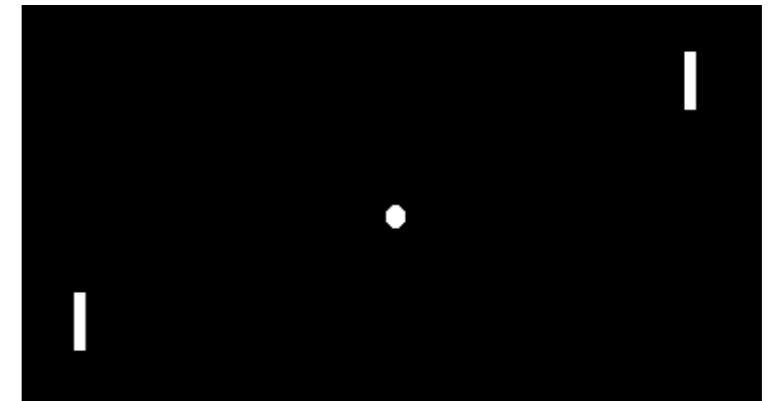
Some Problems with QLearning

- State space is continuous
 - Must approximate \hat{Q} by discretizing
- Treats states as identities
 - No knowledge of how states relate
 - Requires many iterations to fill in \hat{Q}
- Converging \hat{Q} can be difficult with randomized transitions/rewards

```
print(env.observation_space.high)
#> array([ 2.4 , inf, 0.20943951, inf])
print(env.observation_space.low)
#> array([-2.4 , -inf, -0.20943951, -inf])
```

Policy Gradients

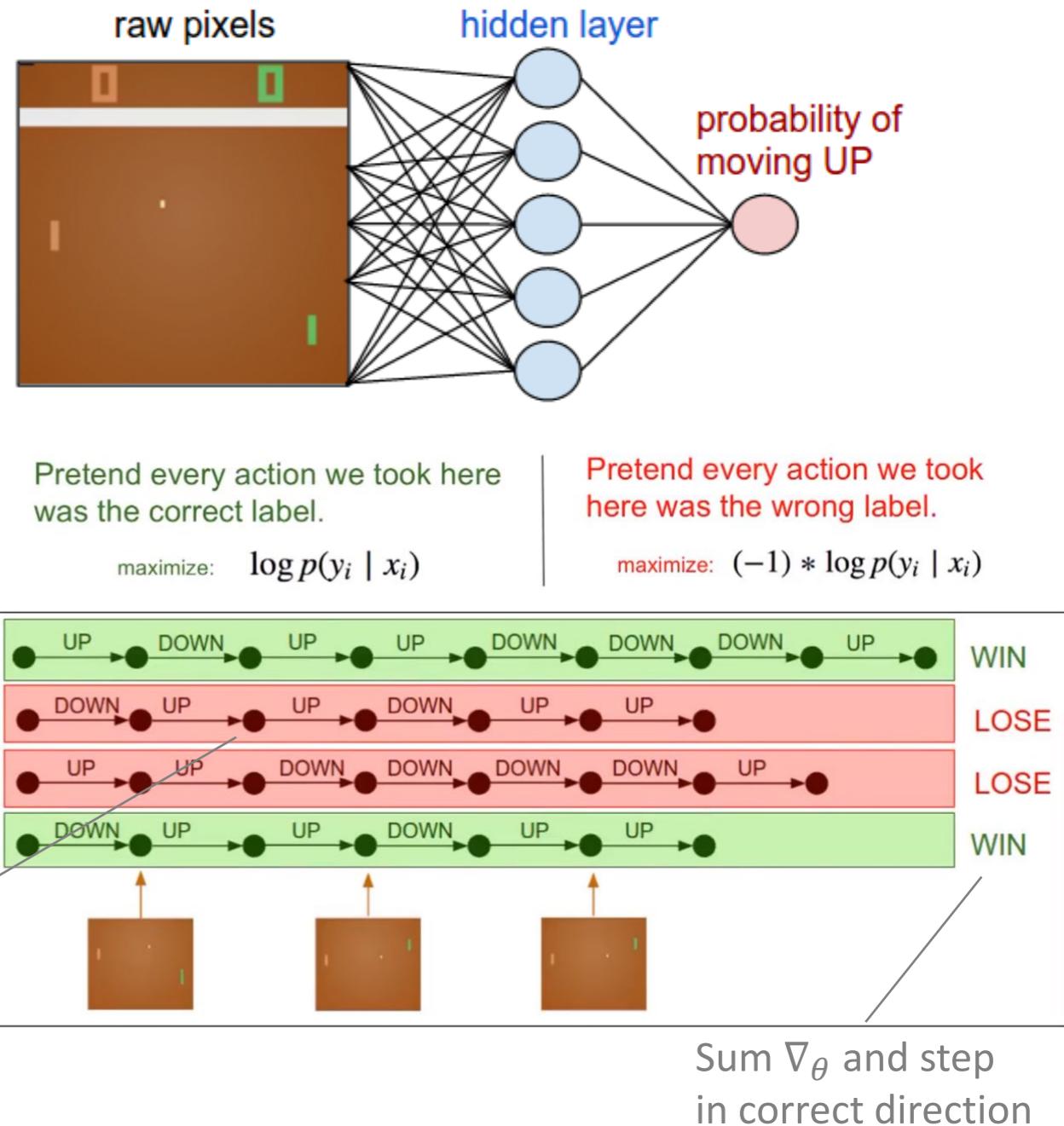
- Q-learning -> learn a value function
 - $\hat{Q}(s, a)$ = an estimate of the expected discounted reward of taking a from s
 - Performance time: take the action that has the highest estimated value
- Policy Gradient -> learn policy directly
 - $\pi(s)$ = Probability distribution over A_s
 - Performance time: choose action according to distribution



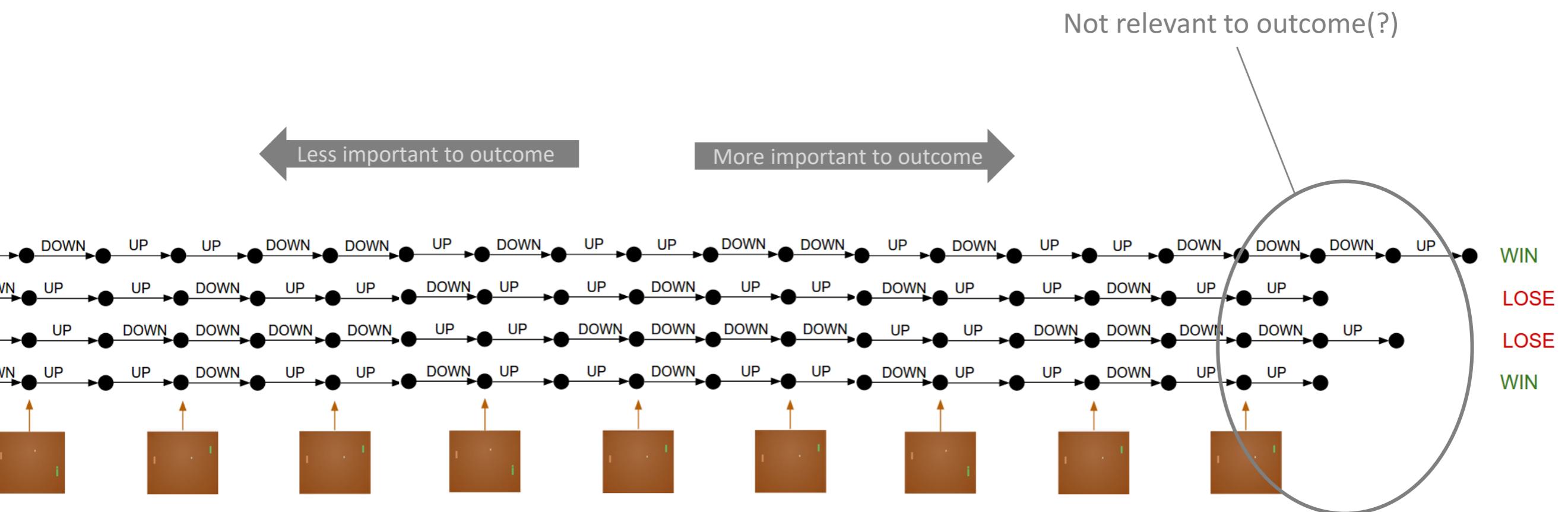
Example from: <https://www.youtube.com/watch?v=tqrcjHuNdmQ>

Policy Gradients

- Receive a frame
- Forward propagate to get $P(\text{actions})$
- Select a by sampling from $P(\text{actions})$
- Find the gradient ∇_{θ} that makes a more likely – store it
- Play the rest of the game
 - If won, take a step in direction ∇_{θ}
 - If lost, take a step in direction $-\nabla_{\theta}$



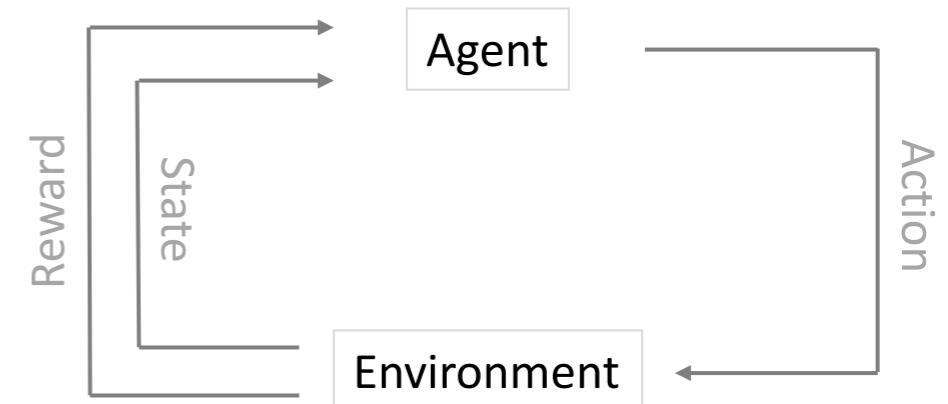
Policy Gradients – reward shaping



Summary

Reinforcement Learning:

- Goal: Maximize $\sum_{i=1}^{\infty} \text{Reward}(State_i, Action_i)$
- Data: $Reward_{i+1}, State_{i+1} = \text{Interact}(State_i, Action_i)$



Many (awesome) recent successes:

- Robotics
- Surpassing humans at difficult games
- Doing it with (essentially) zero human knowledge

Challenges:

- When the episode can end without reward
- When there is a 'narrow' path to reward
- When there are many states and actions

(Simple) Approaches:

- Q-Learning $\hat{Q}(s, a) \rightarrow$ discounted reward of action
- Policy Gradients \rightarrow Probability distribution over A_s
- Reward Shaping
- Memory
- Lots of parameter tweaking...

MIT 18.700/18.701/6.047/6.878/HST.507

Computational Biology: Genomes, Networks, Evolution

Lecture 19: EHRs Electronic Health Records

Dealing with Population-Scale EHRs

PheWAS: GWAS for Multi-Phenotypes

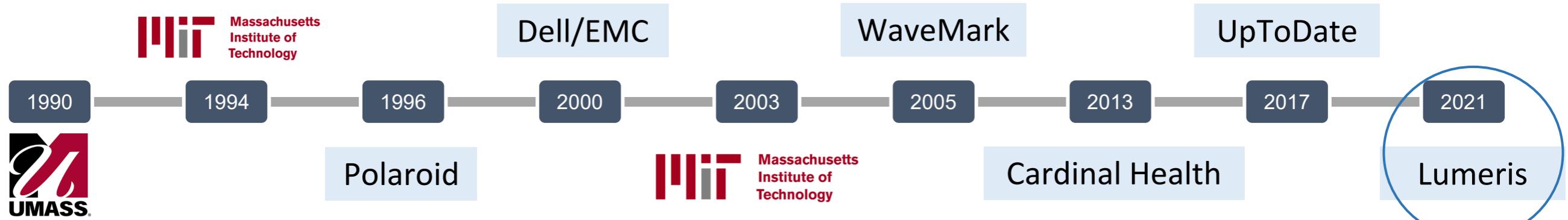
Interventions and Reverse Causality

Clinical Notes: LLMs, Text-Data Joint Models

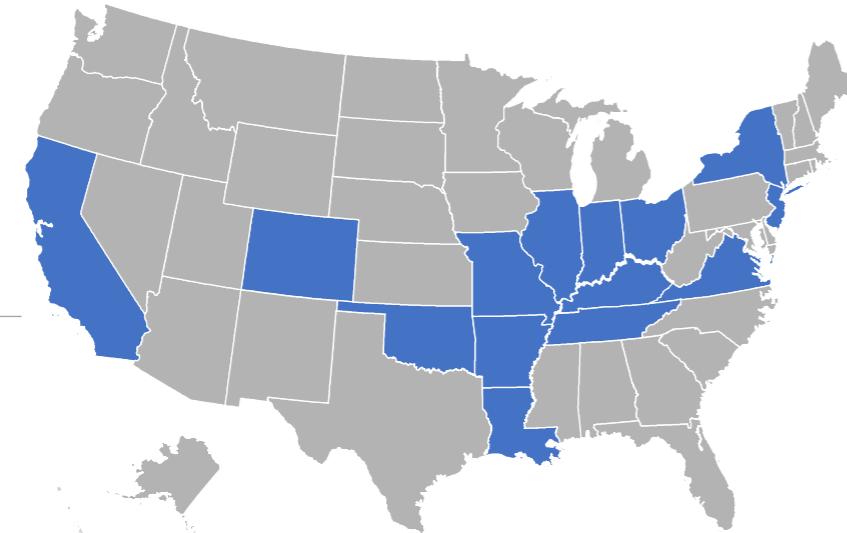
Action, Agents, Reinforcement Learning

AI Agents in Value-Based Healthcare

Jean-Claude Saghbini



Arkansas
California
Colorado
Illinois
Indiana
Kentucky
Louisiana
Missouri
New Jersey
New York
Ohio
Oklahoma
Tennessee
Virginia



Lumeris provides the most comprehensive solution to help providers and health systems manage value-based populations

- 3M patients
- \$13B of spend under management

Lumeris: Jean-Claude Saghbini

The Financial Challenges of Healthcare

↑ \$5T

U.S. health care spend is 50% of the world healthcare spend, for 4% of the global population, and 30% of it is wasteful

↑ 6.7%

U.S. health care costs increase in 2024, reaching \$7,151 per person

↑ 40,000

The United States will face a physician shortage of up to 40,000 PCPs by 2030



HEALTH SYSTEM

- Narrowing margins
- New competitors (non-traditional) threaten patient loyalty
- Payer mix shifts (to more governmental payers) related to aging population
- Price transparency and “consumerism”
- Workforce challenges
- Data siloes



PRIMARY CARE

- Aging workforce and shortage: 40,000 PCP shortage by 2030
- Increased demand for services
- Increased competition for talent
- High provider burnout, 63% of PCPs Report Burnout
- Subsidy/PCP approaching \$250K
- Doing the recommended and appropriate care requires 26.7 hours a day from a PCP

Misaligned Incentives

The consumer is not the Customer

- The patient is a consumer but not a customer, creating two major hazards:
 - The consumer does not pay for what they consume
 - Providers purchase their own services
- All healthcare constituents seek unrestricted access to patients to drive revenue growth. To improve the healthcare system, requires changing the basis of competition beyond revenues, to include quality and cost outcomes for patients.

A Healthcare System is not a System

- System: “A set of principles according to which something is done; an organized framework or method”*
- Healthcare is a loosely affiliated network of thousands of services that have a singular focus on delivering a particular procedure or service with a quality goal to optimize that procedure or service and to maximize its revenues
- Local optimizations with no accountability for the overall performance of the system on the patient

*Oxford Dictionary

Value-Based Care = *Better Care*

Value-Based Care [val-yoo beyst kair]

noun

1. Value-based care is a healthcare delivery model in which providers, including hospitals and physicians, are paid based on patient health outcomes. Under value-based care agreements, providers are rewarded for helping patients improve their health, reduce the effects and incidence of chronic disease, and live healthier lives in an evidence-based way.*
2. Better Care

Examples of Programs: Medicare Advantage, Medicare Shared Savings Program (MSSP), ACO Reach, Medicaid, Commercial VBC Contracts

Examples of Financial Models: Shared Savings, Upside/Downside Risk in Capitated Models, Pay for Quality

Mandatory Transformation of Primary Care to Drive:

- Better Health Outcomes
 - +
- Increased Revenues
- Decreased Medical Expenses
- Decreased Operating Costs

Lumeris: Jean-Claude Saghbini

*NEJM Catalyst

The Key to Change is Primary Care

- 5% of the total cost of care, yet controls the total cost of care
- 1 Primary Care Physician influences \$10M of annual spend

FAILURES OF CARE DELIVERY

- Duplicate or unnecessary tests
- Inaccurate diagnosis
- Medication errors
- Avoidable complications of chronic diseases
- Inefficient care setting (ED, specialist office)
- Missed screenings
- Missed vaccinations

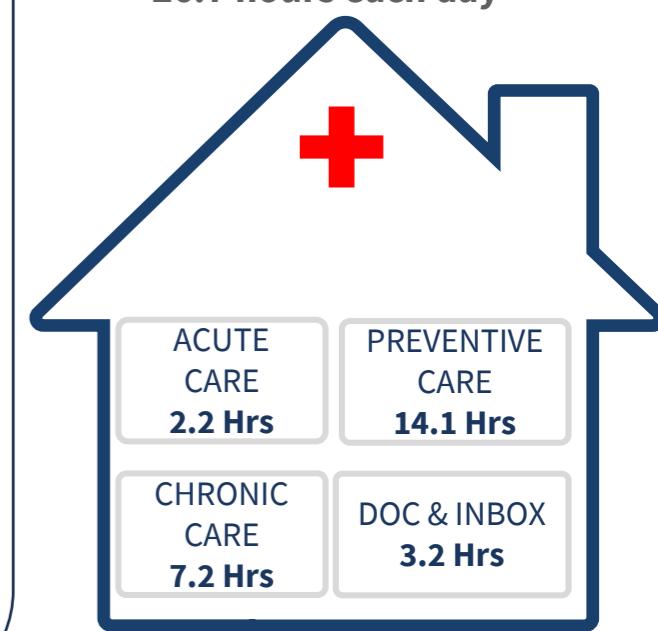
FAILURES OF CARE COORDINATION

- Lack of communications and coordination between two providers
- Duplicate services
- Medication errors
- Unnecessary readmissions
- Site of service inefficiencies and added costs to patient

OVERTREATMENT OR LOW VALUE CARE

- Unnecessary medical or surgical services that do not significantly improve patient outcomes
- Excessive testing
- Overly aggressive treatments
- Specialty referrals
- Futile end-of-life care

The role of PCP is an **impossible** one for a solo provider without a care team...requiring 26.7 hours each day*



The Financials of Value-Based Care

Solution Domains for Value-Based Care

Preventive Care,
Quality & Access

Chronic Disease
Management

High-Risk Patient
Management

Acute & Post Acute
Care Orchestration

Network
Coordination

A Small Set of Examples for Interventions Driving Better Care and Better Financials

- Achieve CMS STAR measures for Quality
 - Required screenings and preventive care to meet measure targets
 - Advance the patient experience given its outsized contribution to STAR ratings
- Chronic Conditions Management and Control
 - Diabetes
 - Hypertension
- Documentations and coding to achieve appropriate Risk Adjustment Factor
 - Pre-visit planning
 - Retrospective chart reviews
 - Submission of supplemental EHR data to payers
- Ensure post-discharge PCP follow-up is done to reduce risk of readmission
- Engaging with primary care services at home for very high-risk patients
- Managing referrals to in-network facilities and specialists
- Reducing the excessive, unnecessary and avoidable ED and hospital utilization and freeing up resources/bed for additional higher-reimbursement patients

Revenues 

Medical Expenses 

Revenues 

Operating Costs 

Medical Expenses 

Revenues 

Revenues 

Lumeris: Jean-Claude Saghbini

Scaling Value-Based Care (VBC)

Full orchestration and automation of Primary Care Services

**Solution Domains
for Value-Based Care**

**Preventive Care,
Quality & Access**

**Chronic Disease
Management**

**High-Risk Patient
Management**

**Acute & Post Acute
Care Orchestration**

**Network
Coordination**

Primary Care Service Categories – Comprising 100's of Services

Preventive Care,
Wellness and
Screening

Chronic Disease
Care

Complex
Medication
Management

Integrated
Behavioral Health

Pre-Visit Planning

Annual Wellness
Visit/Comp.
Assessment

Virtual / Home
Visits

EHR Supplemental
data submissions

Transitions of Care
Management

Acute/Urgent Care
Decision Support

Utilization
Management

Post-Acute
Navigation

ED Diversion

Primary Care at
Home

End-of-Life
Planning

Complex Case
Management

High-Value
Specialty Care

High-Value Facility
Care

In-Network
Referrals

Attribution

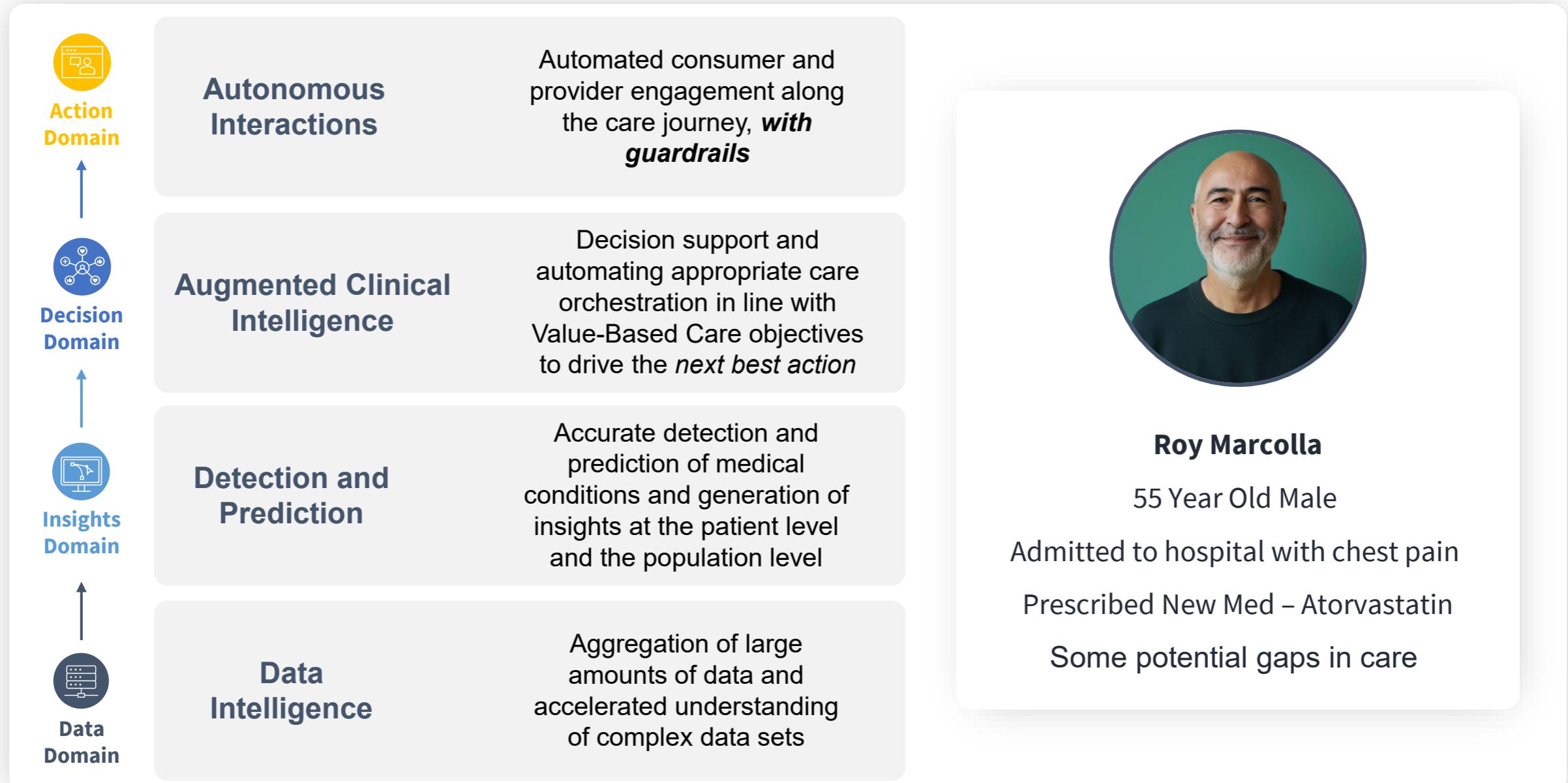
VBC Contracts and Supporting Data/Analytics

*Can only be
done with*



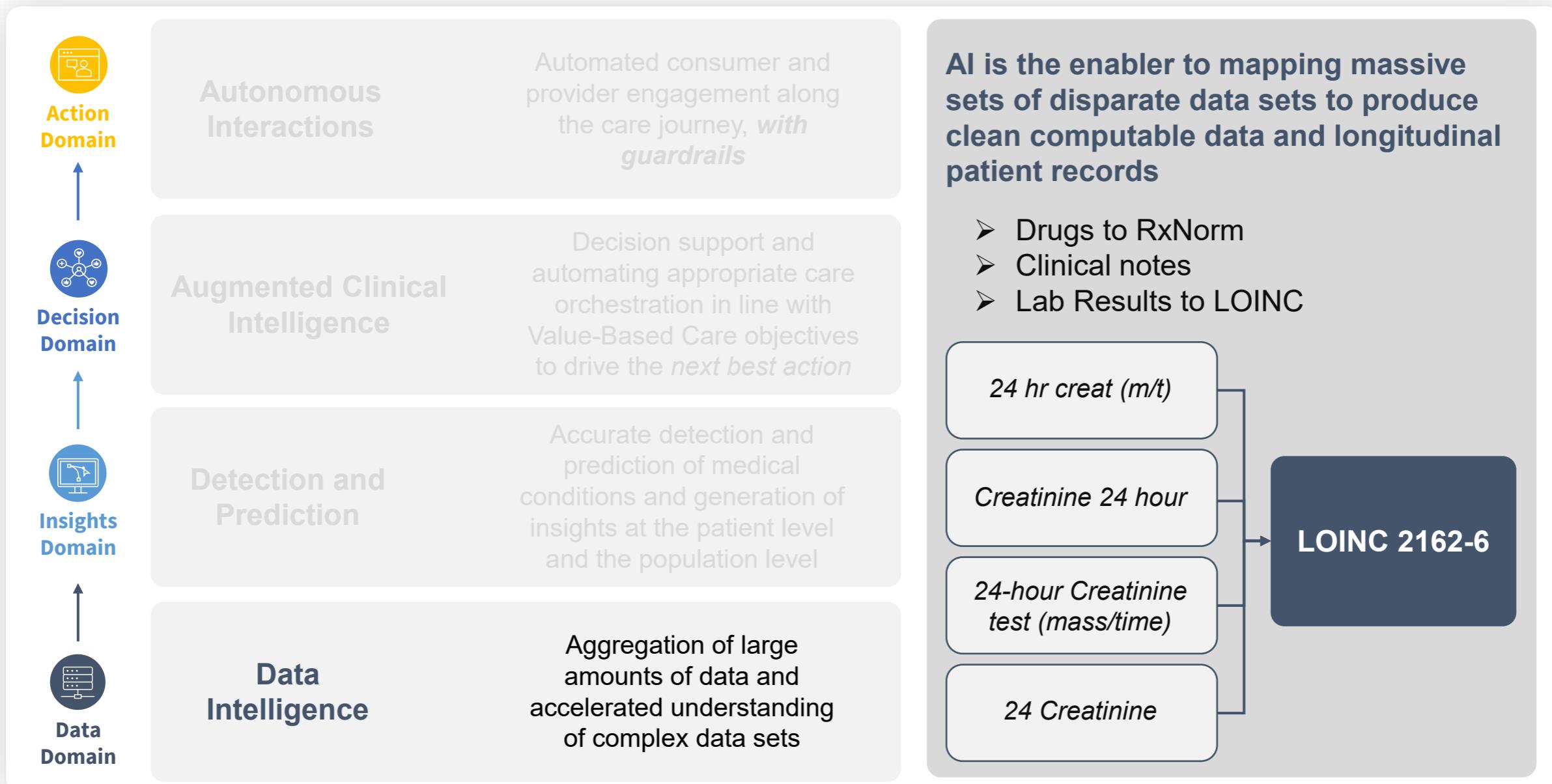
Lumeris: Jean-Claude Saghbini

An AI Framework for Value-Based Primary Care Services



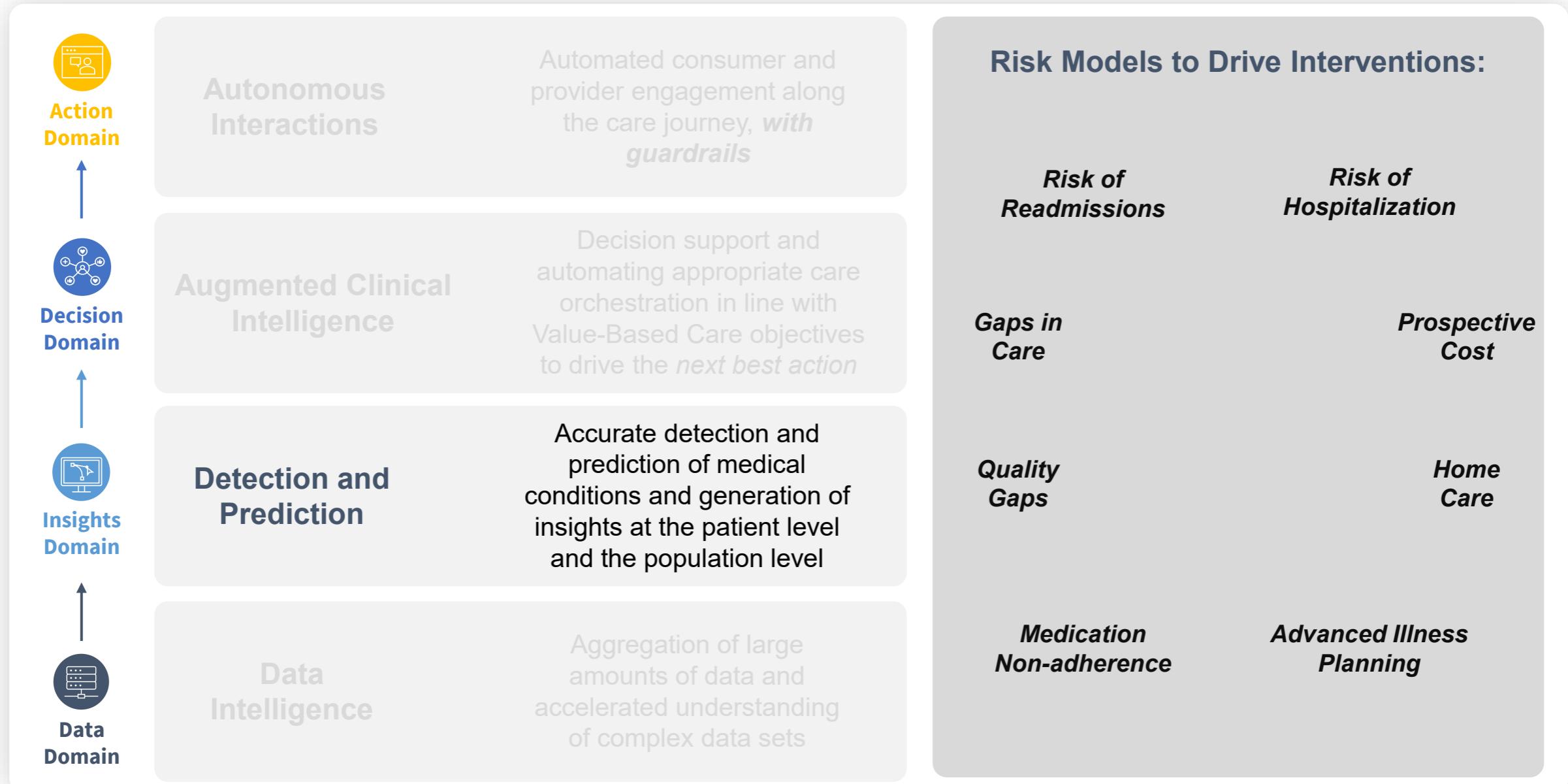
Lumeris: Jean-Claude Saghbini

AI Implementations in Value-Based Primary Care Services



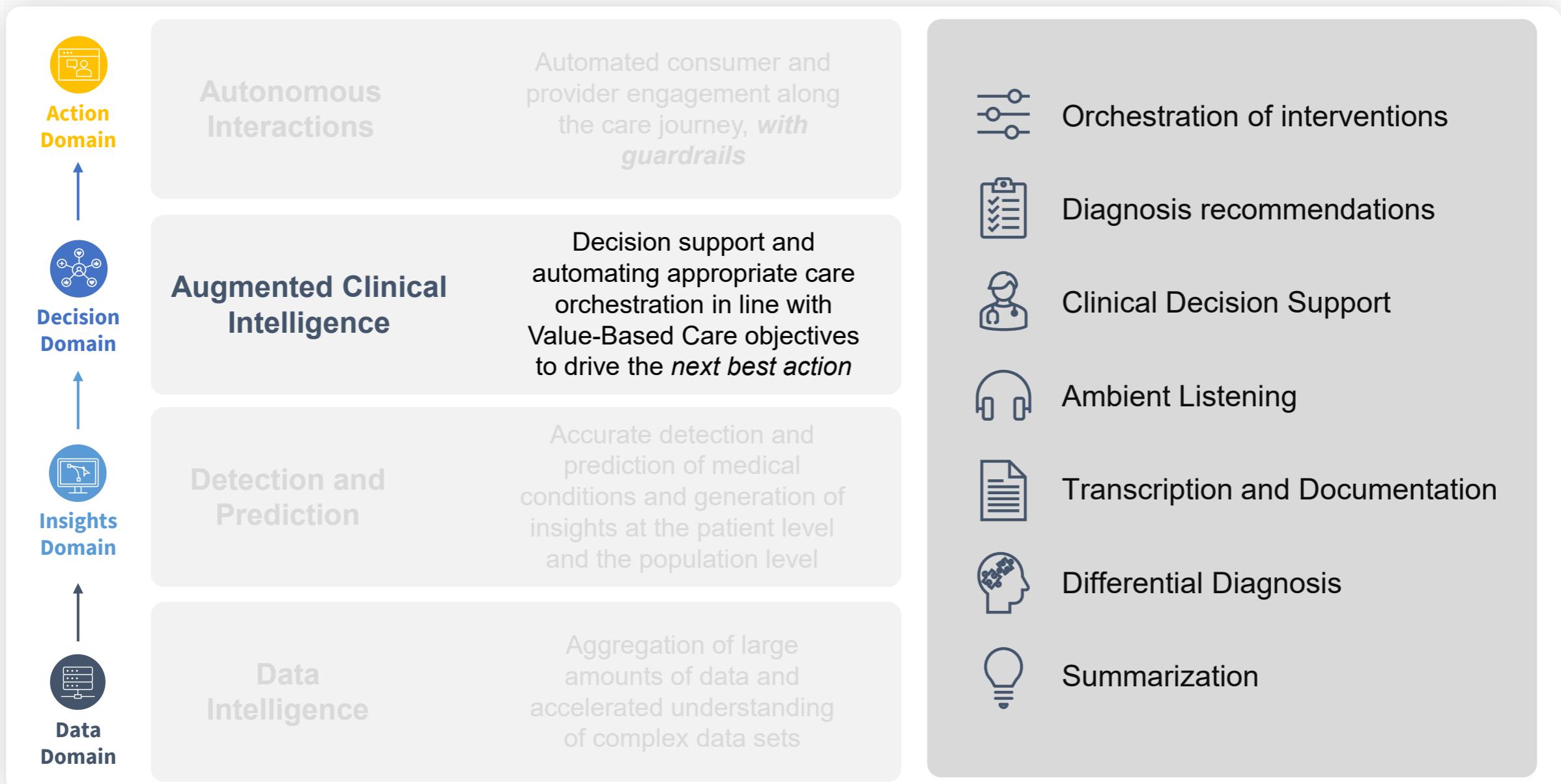
Lumeris: Jean-Claude Saghbini

AI Implementations in Value-Based Primary Care Services



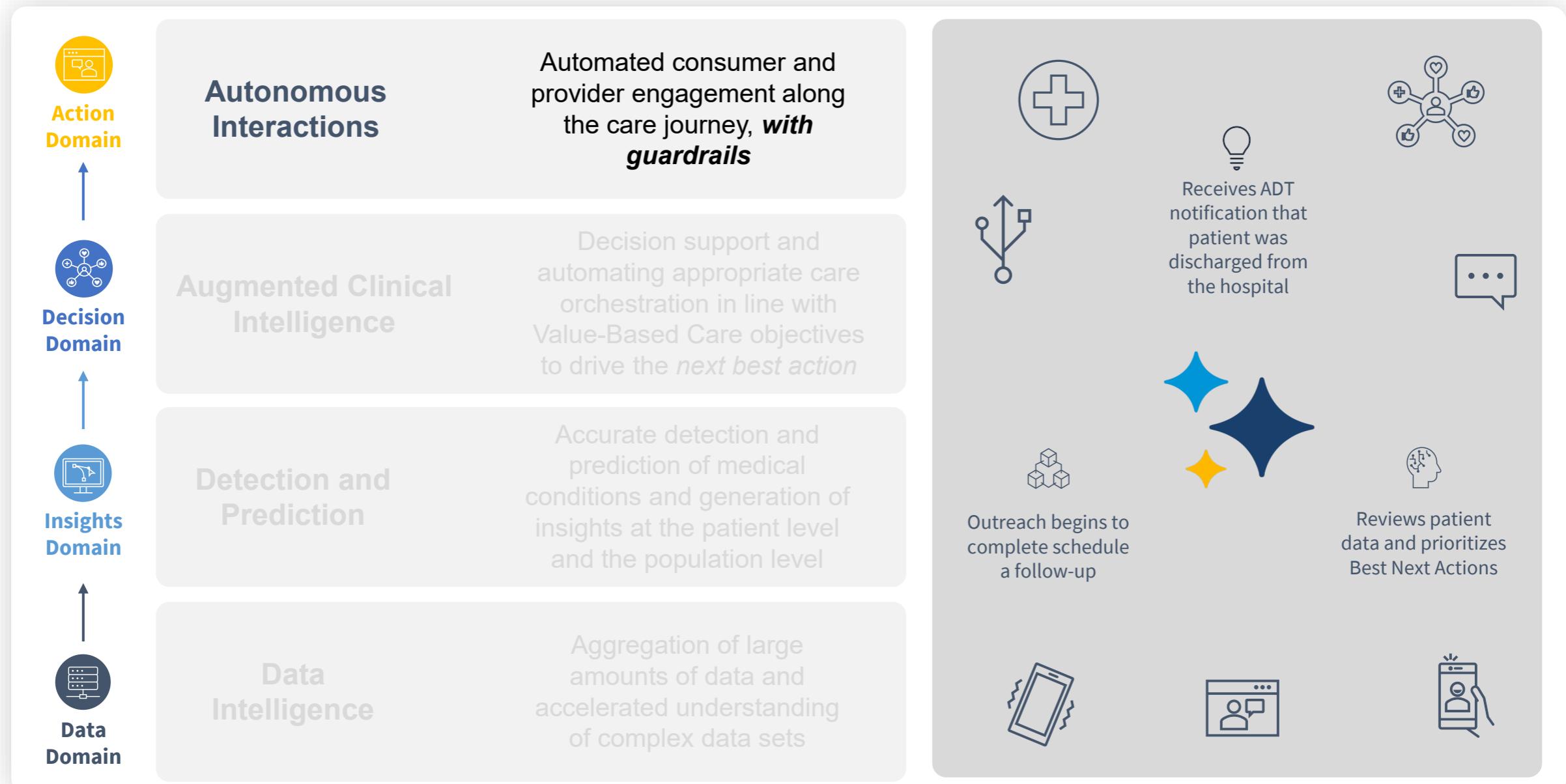
Lumeris: Jean-Claude Saghbini

AI Implementations in Value-Based Primary Care Services

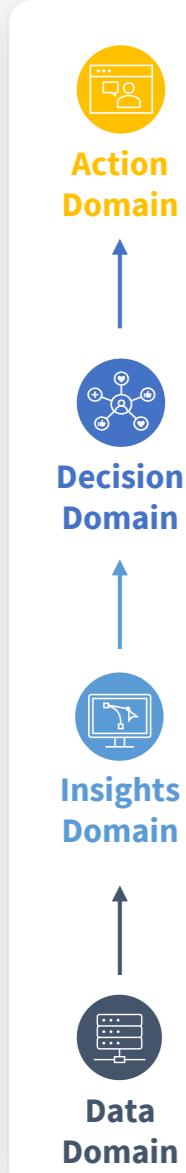


Lumeris: Jean-Claude Saghbini

AI Implementations in Value-Based Primary Care Services



Lumeris: Jean-Claude Saghbini



Autonomous Interactions

Automated consumer and provider engagement along the care journey, ***with guardrails***

Augmented Clinical Intelligence

Decision support and automating appropriate care orchestration in line with Value-Based Care objectives to drive the *next best action*

Detection and Prediction

Accurate detection and prediction of medical conditions and generation of insights at the patient level and the population level

Data Intelligence

Aggregation of large amounts of data and accelerated understanding of complex data sets

Hello, I am calling from Maryland Heights Medicine. I am an AI assistant on a recorded line, may I speak with Roy?

Yes. This is Roy.

Great, thank you for confirming. I'm reaching out to remind you of your upcoming appointment and discuss a few things to help your visit go smoothly. Is it okay if we proceed?

Yes. Let's do it.

Thank you! Your appointment is a follow-up with Doctor Jennifer Lawrence on November 1st at 3 PM. Can you attend that appointment?

Yes. I'll be there.

Perfect! I'm glad to hear that. Now, I'd like to ask you a few questions about how you've been feeling lately. Do you mind if I continue?

No. Please do.

Thank you. How have you been feeling over the past two weeks?

Honestly, I've been a little down since I got out of the hospital.

I'm sorry to hear that. It's understandable to feel that way after a hospital stay. Have you been enjoying your usual activities recently?

Not really. I really like to play golf and just have not had the energy to do that.

I appreciate you sharing that. How have your energy levels been overall? Have you felt more tired than usual?

Oh, absolutely.

Thank you for letting me know. How have you been sleeping?

I fall asleep okay, but I wake up several times during the night and start thinking about work.

That sounds frustrating. It's tough when your mind keeps you awake. How has your appetite been? Any changes in your eating habits?

Not really, but I do find that I've been binge eating on cookies.

Thank you for sharing that. Have you had any trouble concentrating or making decisions recently?

A little bit. Especially in the morning.

I understand. Have you been feeling down on yourself or like a failure at all?

Kind of, but not too bad.

Thank you for being honest. Have you had any thoughts of hurting yourself or feeling like you shouldn't be here?

Oh, no. Not that.

I'm relieved to hear that. Thank you for sharing your feelings with me. I'll make sure to pass this information along to your care team so they can provide the support you need. Now, I have one last topic to cover: your flu vaccination status. Have you had a flu shot in the last six months?

Yes. I had it three weeks ago at Walgreens.

Thank you for letting me know. It's great that you got your flu shot; it helps protect you and others. I appreciate your time today.

No. Thank you for calling.

You're welcome! If you need anything else, feel free to reach out. Have a great day, and I'll see you at your appointment. Goodbye!

MIT 18.700/18.701/6.047/6.878/HST.507

Computational Biology: Genomes, Networks, Evolution

Lecture 19: EHRs

Electronic Health Records

Dealing with Population-Scale EHRs

PheWAS: GWAS for Multi-Phenotypes

Interventions and Reverse Causality

Clinical Notes: LLMs, Text-Data Joint Models

Action, Agents, Reinforcement Learning

AI Agents in Value-Based Healthcare