MACHINE LEARNING FOR HEALTHCARE

6.S897, HST.S53

Lecture 10: Interpretability of machine learning models

Prof. David Sontag

MIT EECS, CSAIL, IMES

(Thanks to Zack Lipton for many of the slides)

Massachusetts Institute of Technology

Outline of today's class

1. **The mythos of model interpretability in health care**
2. Learning intelligible models
3. Post-hoc interpretability

What is interpretability?

- Many papers make axiomatic claims that some model is interpretable and therefore preferable
- But what interpretability is and precisely what desiderata it serves are seldom defined

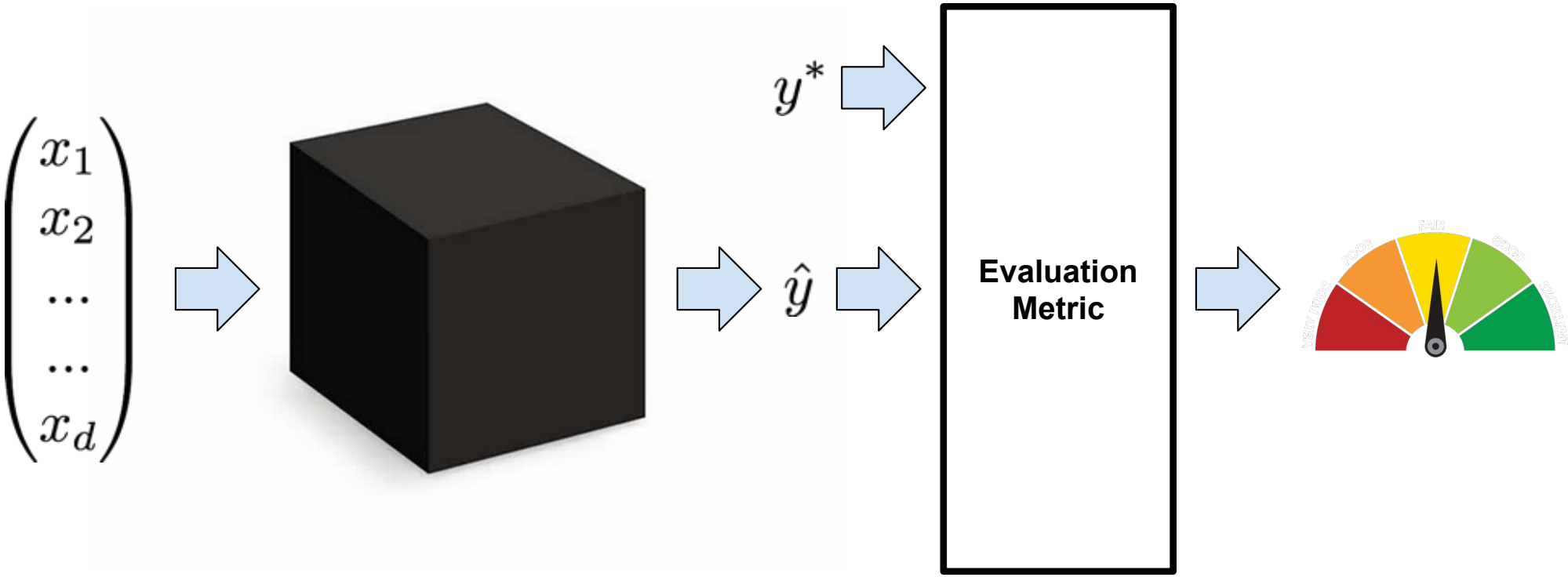
(Slide credit: Zachary Lipton)

Inconsistent definitions

- Papers use the words ***interpretable, explainable, intelligible, transparent,*** and ***understandable***, both interchangeably (within papers) and inconsistently (across papers)
- One common thread, however, is that interpretability is something other than performance

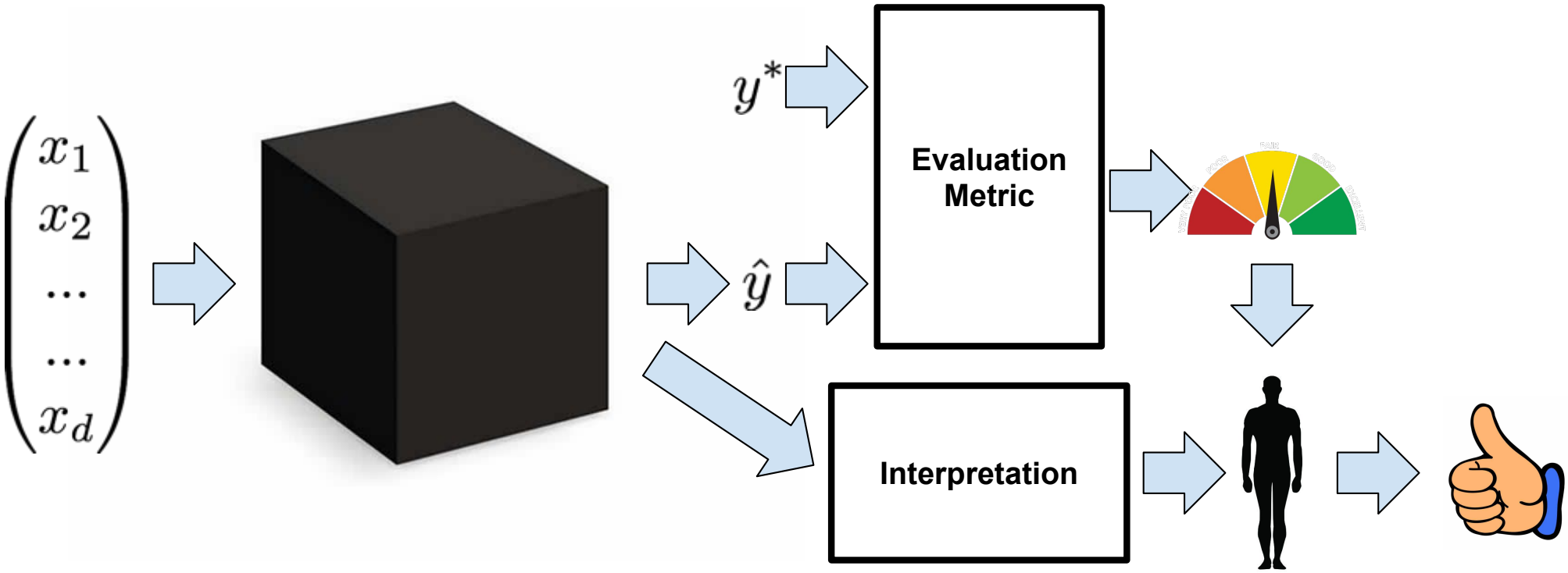
(Slide credit: Zachary Lipton)

We want good models



(Slide credit: Zachary Lipton)

We also want interpretable models



The human wants something the metric doesn't. But, what?

(Slide credit: Zachary Lipton)

Trust

- Does the model *know* when it's uncertain?
- Does the model make same mistakes as human? (e.g., would we be happy delegating decision making authority?)
- Are we *comfortable* with the model?

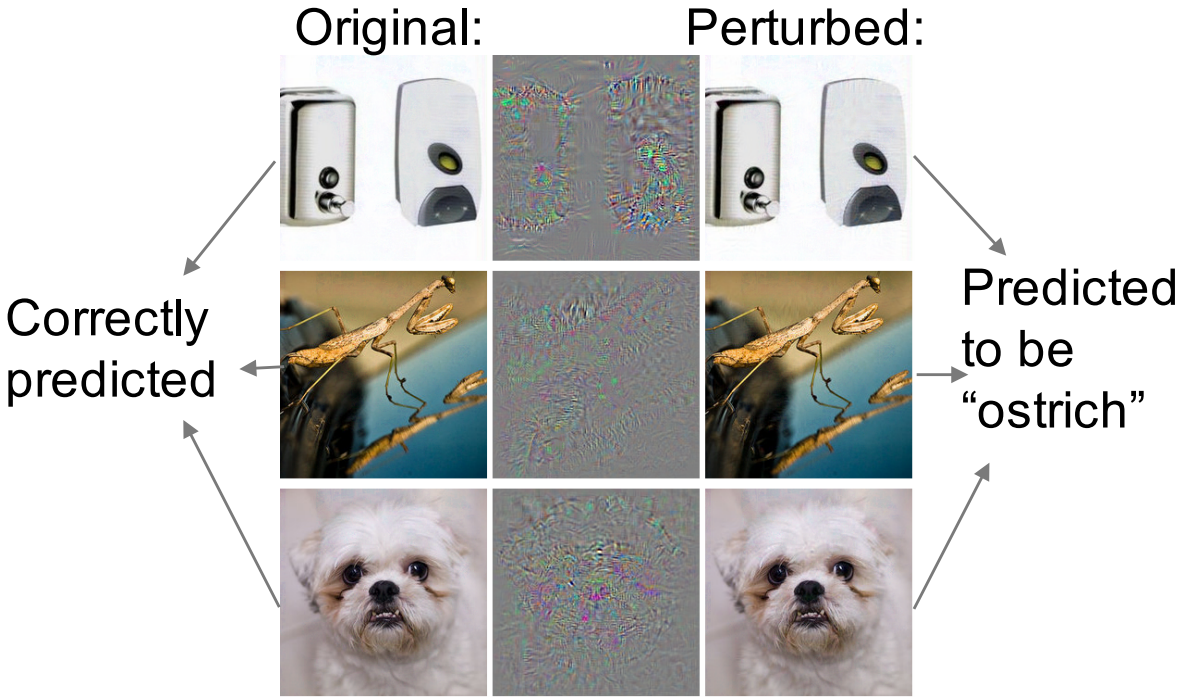
(Slide credit: Zachary Lipton)

Trust: can you fool the classifier?

- Example from Szegedy et al., “Intriguing properties of neural networks”, ICLR 2014
- Small perturbations of image do not affect visual semantics, but *do* affect classifications using neural networks

Minimize $\|r\|_2$ subject to:

1. $f(x+r) = l$
2. $x + r \in [0,1]^m$



Causality

- We may want models to tell us something about the natural world
- Supervised models are trained simply to make predictions, but often used to take actions
- Caruana (2015) shows a mortality predictor (for use in triage) that assigns lower risk to asthma patients
- Naïve interpretations can be misleading

(Slide credit: Zachary Lipton)

Causality: reminder from Lecture 3

- Why one *might* interpret weights learned by linear model causally:

$$Y_t(x) = \beta^T x + \gamma \cdot t + \epsilon_t$$
$$\mathbb{E}[\epsilon_t] = 0$$

$$ATE = \mathbb{E}[Y_1(x) - Y_0(x)] = \gamma$$

- Here we care about γ, not about $Y_t(x)$
  Identification, not prediction
- Danger: all bets are off with model misspecification

Causality: reminder from Lecture 3

- Suppose true data generating process, $x \in \mathbb{R}$:
  $Y_t(x) = \beta x + \gamma \cdot t + \delta \cdot x^2$
  $ATE = \mathbb{E}[Y_1 - Y_0] = \gamma$
- Hypothesized linear model (misspecified):
  $\hat{Y}_t(x) = \hat{\beta} x + \hat{\gamma} \cdot t$

$$\hat{\gamma} = \gamma + \delta \frac{\mathbb{E}[xt]\mathbb{E}[x^2] - \mathbb{E}[t^2]\mathbb{E}[x^2 t]}{\mathbb{E}[xt]^2 - \mathbb{E}[x^2]\mathbb{E}[t^2]}$$

The sign of the weight can flip from negative to positive (and vice-versa)!

Transferability

- The idealized training setups often differ from the real world
  - E.g., data leakage, errors in outcome definition from observational data

- Real problem may be non-stationary, noisier, etc.

- Want sanity-checks that the model doesn't depend on weaknesses in setup

(Slide credit: Zachary Lipton)

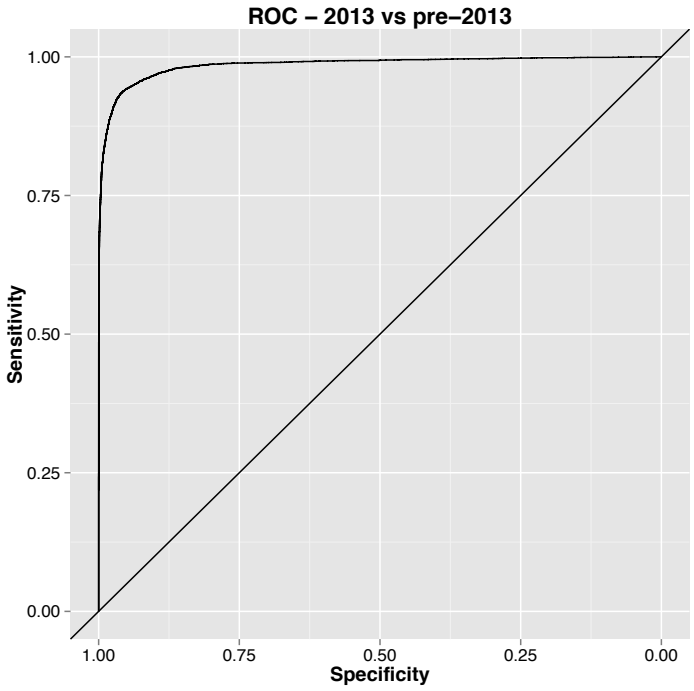
Transferability: non-stationary

- Data created during health care is from a non-stationary process due to changes in:
  - Medical science
  - Incentives & regulations
  - Business processes

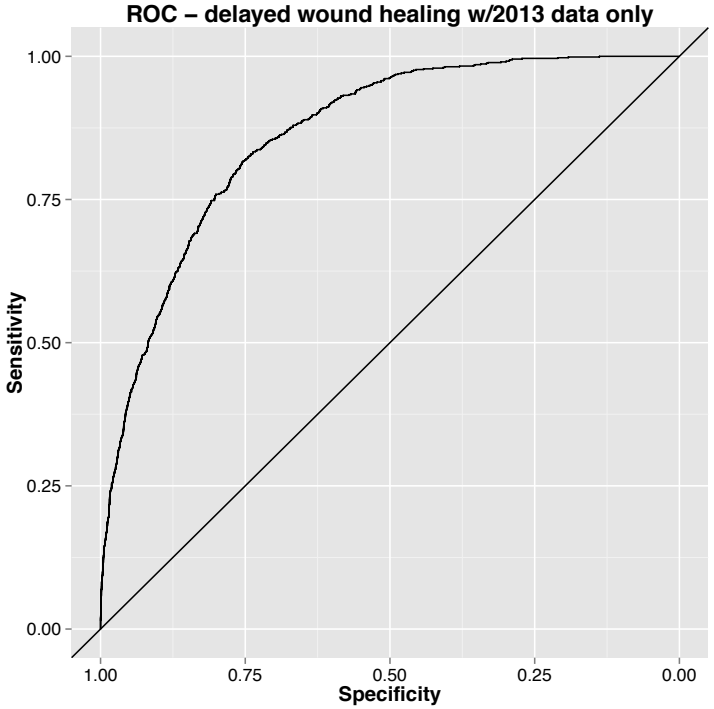
(Slide credit: Ken Jung)

Transferability: non-stationary

- Testing for covariate shift (wound healing):



Distinguish 2013 from pre-2013

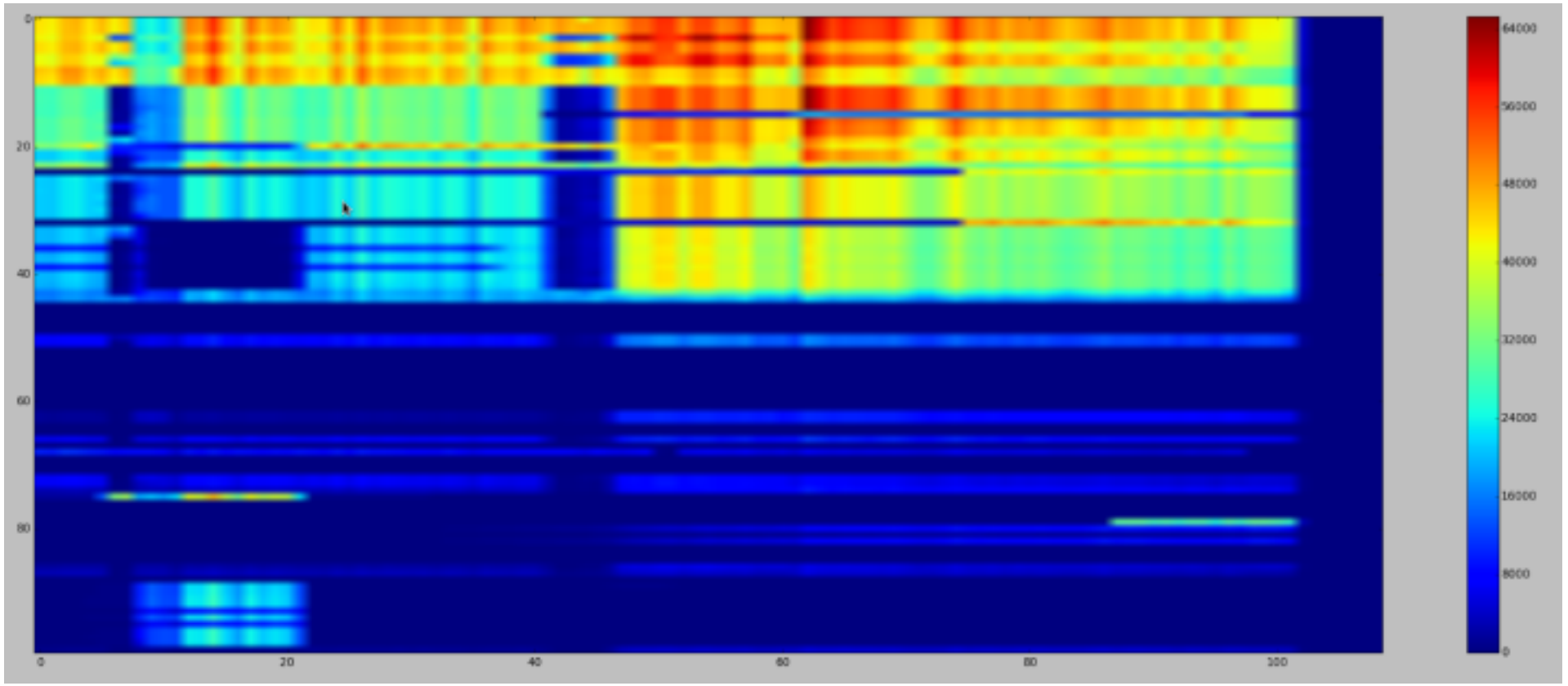


Distinguish first 2/3 of 2013 from last 1/3 of 2013

(Slide credit: Ken Jung)

Transferability: non-stationary

Top 100 lab measurements over time



Time (in months, from 1/2005 up to 1/2014)

Case study on transferability: Framingham CHD risk score

- Many ML models are trained in one place and deployed more broadly
- **Example:** Framingham coronary heart disease (CHD) risk score
  - Model based on 6 major risk factors: age, BP, smoking, diabetes, total cholesterol (TC), and high-density lipoprotein cholesterol (HDL-C)

[Wilson et al., Circulation, 1998]

CHD score sheet for men using TC or LDL-C categories.

Step 1

Age		
Years	LDL Pts	Chol Pts
30-34	-1	[-1]
35-39	0	[0]
40-44	1	[1]
45-49	2	[2]
50-54	3	[3]
55-59	4	[4]
60-64	5	[5]
65-69	6	[6]
70-74	7	[7]

Step 2

LDL - C		
(mg/dl)	(mmol/L)	LDL Pts
<100	<2.59	-3
100-129	2.60-3.36	0
130-159	3.37-4.14	0
160-190	4.15-4.92	1
≥190	≥4.92	2

Cholesterol		
(mg/dl)	(mmol/L)	Chol Pts
<160	<4.14	[-3]
160-199	4.15-5.17	[0]
200-239	5.18-6.21	[1]
240-279	6.22-7.24	[2]
≥280	≥7.25	[3]

Step 3

HDL - C			
(mg/dl)	(mmol/L)	LDL Pts	Chol Pts
<35	<0.90	2	[2]
35-44	0.91-1.16	1	[1]
45-49	1.17-1.29	0	[0]
50-59	1.30-1.55	0	[0]
≥60	≥1.56	-1	[-2]

Step 4

Blood Pressure					
Systolic (mm Hg)	Diastolic (mm Hg) <80	80-84	85-89	90-99	≥100
<120	0 [0] pts				
120-129		0 [0] pts			
130-139			1 [1] pts		
140-159				2 [2] pts	
≥160					3 [3] pts

Note: When systolic and diastolic pressures provide different estimates for point scores, use the higher number

Step 5

Diabetes		
	LDL Pts	Chol Pts
No	0	[0]
Yes	2	[2]

Step 6

Smoker		
	LDL Pts	Chol Pts
No	0	[0]
Yes	2	[2]

(sum from steps 1-6)

Step 7

Adding up the points	
Age	_____
LDL-C or Chol	_____
HDL - C	
Blood Pressure	_____
Diabetes	_____
Smoker	_____
Point total	_____

Key	
Color	Relative Risk
green	Very low
white	Low
yellow	Moderate
rose	High
red	Very high

(determine CHD risk from point total)

Step 8

CHD Risk			
LDL Pts Total	10 Yr CHD Risk	Chol Pts Total	10 Yr CHD Risk
<-3	1%		
-2	2%		
-1	2%	[<-1]	[2%]
0	3%	[0]	[3%]
1	4%	[1]	[3%]
2	4%	[2]	[4%]
3	6%	[3]	[5%]
4	7%	[4]	[7%]
5	9%	[5]	[8%]
6	11%	[6]	[10%]
7	14%	[7]	[13%]
8	18%	[8]	[16%]
9	22%	[9]	[20%]
10	27%	[10]	[25%]
11	33%	[11]	[31%]
12	40%	[12]	[37%]
13	47%	[13]	[45%]
≥14	≥56%	[≥14]	[≥53%]

(compare to average person your age)

Step 9

Comparative Risk			
Age (years)	Average 10 Yr CHD Risk	Average 10 Yr Hard* CHD Risk	Low** 10 Yr CHD Risk
30-34	3%	1%	2%
35-39	5%	4%	3%
40-44	7%	4%	4%
45-49	11%	8%	4%
50-54	14%	10%	6%
55-59	16%	13%	7%
60-64	21%	20%	9%
65-69	25%	22%	11%
70-74	30%	25%	14%

* Hard CHD events exclude angina pectoris

** Low risk was calculated for a person the same age, optimal blood pressure, LDL-C 100-129 mg/dL or cholesterol 160-199 mg/dl, HDL-C 45 mg/dL for men or 55 mg/dL for women, non-smoker, no diabetes

Risk estimates were derived from the experience of the Framingham Heart Study, a predominantly Caucasian population in Massachusetts, USA

Peter W. F. Wilson et al. Circulation. 1998;97:1837-1847

American Heart Association.

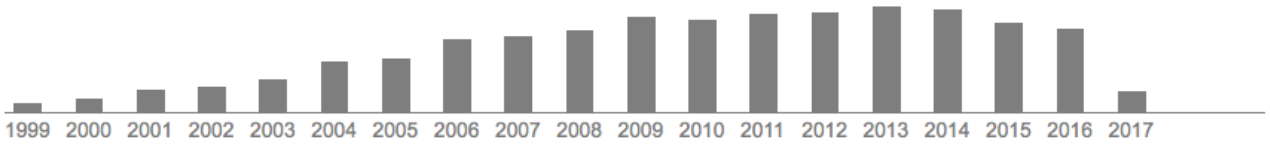
Case study on transferability: Framingham CHD risk score

- Many ML models are trained in one place and deployed more broadly
- **Example:** Framingham coronary heart disease (CHD) risk score

Prediction of coronary heart disease using risk factor categories

[HTML] from ahajournals.org
Full text - MIT Libraries

Authors	Peter WF Wilson, Ralph B D'Agostino, Daniel Levy, Albert M Belanger, Halit Silbershatz, William B Kannel
Publication date	1998/5/1
Journal	Circulation
Volume	97
Issue	18
Pages	1837-1847
Publisher	Lippincott Williams & Wilkins
Description	Background—The objective of this study was to examine the association of Joint National Committee (JNC-V) blood pressure and National Cholesterol Education Program (NCEP) cholesterol categories with coronary heart disease (CHD) risk, to incorporate them into coronary prediction algorithms, and to compare the discrimination properties of this approach with other noncategorical prediction functions. Methods and Results—This work was designed as a prospective, single-center study in the setting of a community-based ...
Total citations	Cited by 8422



Case study on transferability: Framingham CHD risk score

- Many ML models are trained in one place and deployed more broadly
- **Example:** Framingham coronary heart disease (CHD) risk score
  - 99% of Framingham participants are of European descent
  - How well does it generalize to a Chinese population?

- C-statistic (=AUC on censored data) on Chinese population is 0.705/0.742 (M/F)

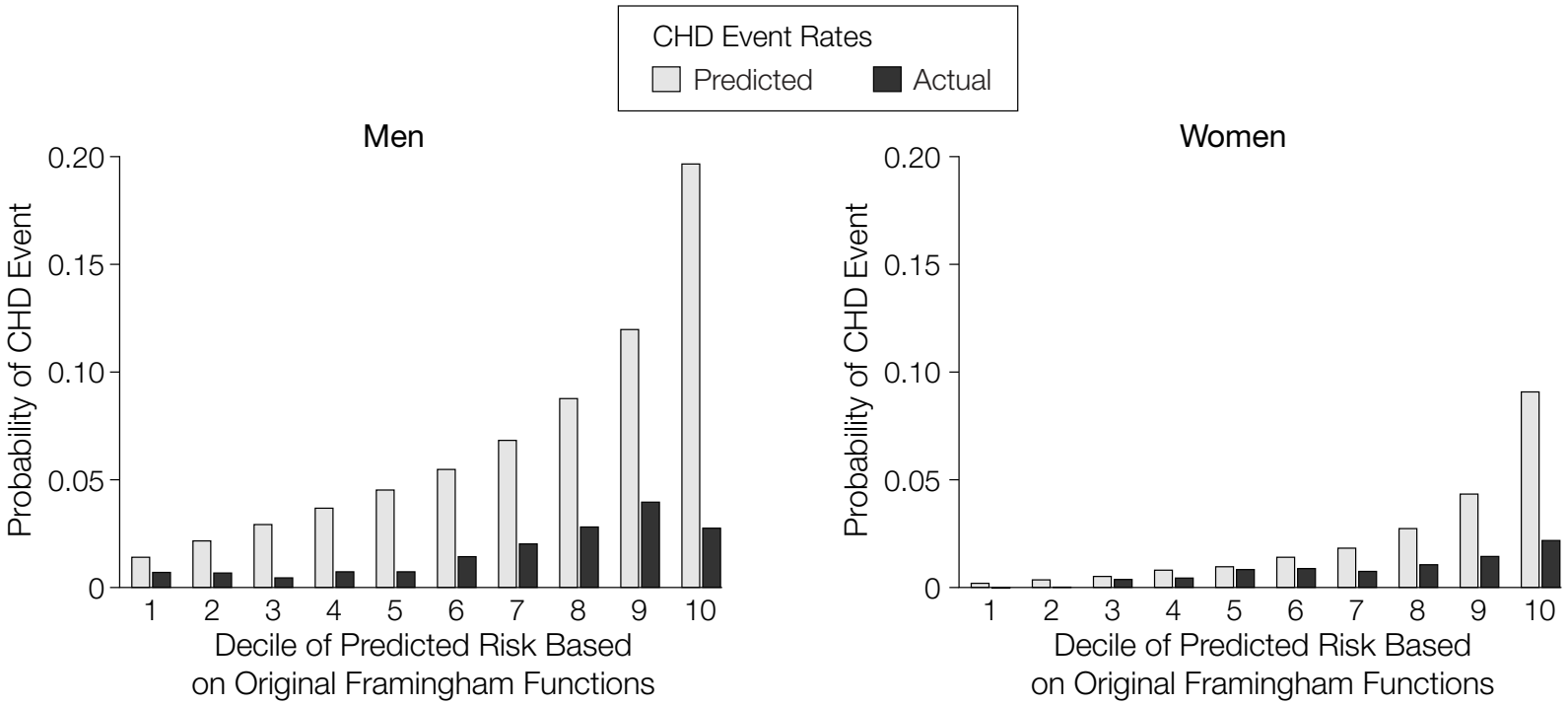
- What else should we look at?

[Liu et al., JAMA '04]

Case study on transferability: Framingham CHD risk score

- **Example:** Framingham coronary heart disease (CHD) risk score (directly applied to Chinese population)

Figure 2. Ten-Year Prediction of CHD Events in CMCS Men and Women Using the Original Framingham Functions



[Liu et al., JAMA '04]

Case study on transferability: Framingham CHD risk score

- Many ML models are trained in one place and deployed more broadly
- **Example:** Framingham coronary heart disease (CHD) risk score
  - 99% of Framingham participants are of European descent
  - How well does it generalize to a Chinese population?

- C-statistic (=AUC on censored data) 0.705/0.742 (M/F)
- Re-fit using local data only slightly improves C-statistic (=AUC on censored data), to 0.736/0.759 (M/F)

[Liu et al., JAMA '04]

Case study on transferability: Framingham CHD risk score

- **Example:** Framingham coronary heart disease (CHD) risk score (re-fit to Chinese population)

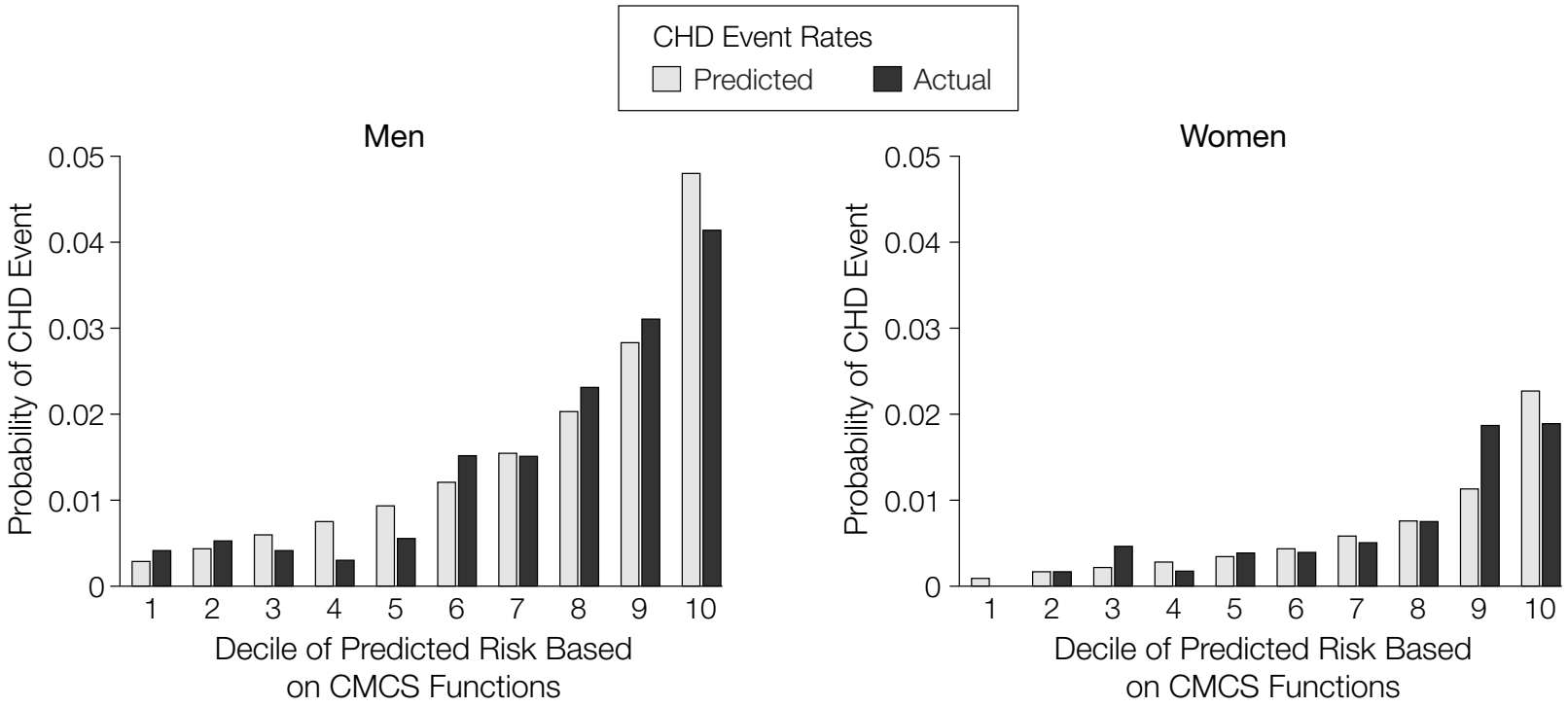
	CMCS	Framingham*
Risk Factors	β	β
Age	0.07	0.05
Age squared	NA	NA
Blood pressure		
Optimal	−0.51	0.09
Normal		
High normal	0.21	0.42
Stage 1 hypertension	0.33	0.66
Stage 2-4 hypertension	0.77	0.90
TC, mg/dL		
<160	−0.51	−0.38
160-199		
200-239	0.07	0.57
240-279	0.32	0.74
≥280	0.52	0.83
HDL-C, mg/dL		
<35	−0.25	0.61
35-44	0.01	0.37
45-49		
50-59	−0.07	0.00
≥60	−0.40	−0.46
Diabetes	0.09	0.53
Smoking	0.62	0.73

[Liu et al., JAMA '04]

Case study on transferability: Framingham CHD risk score

- **Example:** Framingham coronary heart disease (CHD) risk score (re-fit to Chinese population)

Figure 1. Ten-Year Prediction of CHD Events in CMCS Men and Women Using the CMCS Functions



[Liu et al., JAMA '04]

KEY QUESTION TO THINK ABOUT

How robust are your models to changes in the data?

Informativeness

- We may train a model to make a decision

- But it's real purpose is usually to aid a person in making a decision

- Thus an interpretation may be valuable for the extra bits it carries

  I.e., ability to integrate model output with human prior beliefs

(Slide credit: Zachary Lipton)

DISCUSS

What are examples where informativeness may be important for clinical decision making?

DISCUSS

Where does interpretability show up in your projects?

Outline of today's class

1. The mythos of model interpretability in health care
2. **Learning intelligible models**
3. Post-hoc interpretability

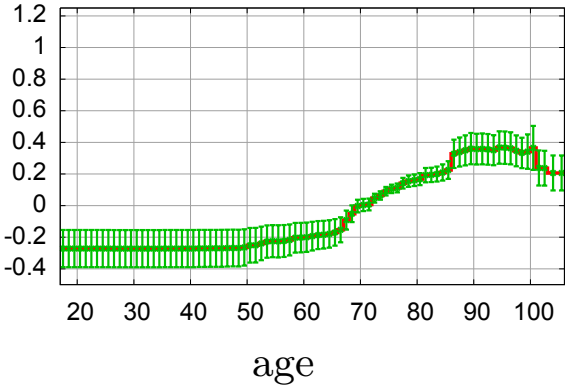
Generalized additive models (GAMs)

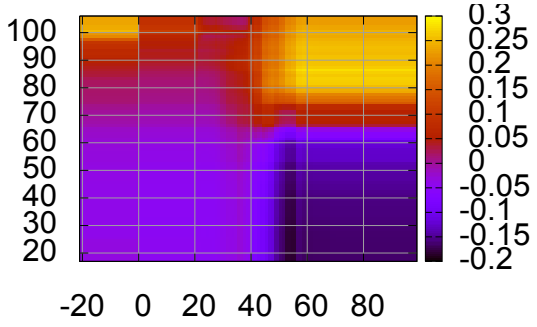
- GAMs with pairwise interactions have the form:

$$g(E[y]) = \beta_0 + \sum_j f_j(x_j) + \sum_{i \neq j} f_{ij}(x_i, x_j)$$

- *g* is the link function (e.g. logistic, for binary data), and *E[f] = 0*.

Model	Pneumonia	Readmission
Logistic Regression	0.8432	0.7523
GAM	0.8542	0.7795
GA^2M	0.8576	0.7833
Random Forests	0.8460	0.7671
LogitBoost	0.8493	0.7835





[Caruana et al., KDD '15]

Falling rule lists

- Ordered list of if-then rules where:
  1. It is a decision list, i.e. order matters
  2. Probability of outcome decreases monotonically

	Conditions		Probability	Support
IF	IrregularShape AND Age $\geq$ 60	THEN malignancy risk is	85.22%	230
ELSE IF	SpiculatedMargin AND Age $\geq$ 45	THEN malignancy risk is	78.13%	64
ELSE IF	IllDefinedMargin AND Age $\geq$ 60	THEN malignancy risk is	69.23%	39
EI		EN malignancy risk is	63.40%	153
EI		EN malignancy risk is	39.68%	63
EI		EN malignancy risk is	26.09%	46
EI		EN malignancy risk is	10.38%	366

r mammographic mass dataset.

Method	Mean AUROC (STD)
FRL	.80 (.02)
NF_FRL	.75 (.02)
NF_GRD	.75 (.02)
RF	.79 (.03)
SVM	.62 (.06)
Logreg	.82 (.02)
Cart	.52 (.01)

Table 3: AUROC values for readmission data

[Wang & Rudin, AISTATS '15]

Supersparse linear integer models

- Learn **linear** model where:
  1. Coefficients are all integer
  2. As sparse as possible

Training objective:

$$\min_{\boldsymbol{\lambda}} \quad \frac{1}{N}\sum_{i=1}^{N} \mathbb{1}\left[y_i \boldsymbol{\lambda}^T \boldsymbol{x}_i \leq 0\right] + C_0 \|\boldsymbol{\lambda}\|_0 + \epsilon \|\boldsymbol{\lambda}\|_1$$

$$\text{s.t.} \quad \boldsymbol{\lambda} \in \mathcal{L}.$$

PREDICT PATIENT HAS OBSTRUCTIVE SLEEP APNEA IF SCORE > 1

1.	$age \geq 60$	4 points		.....
2.	*hypertension*	4 points	+	.....
3.	*body mass index* ≥ 30	2 points	+	.....
4.	*body mass index* ≥ 40	2 points	+	.....
5.	*female*	-6 points	+	.....
	ADD POINTS FROM ROWS 1 – 5	**SCORE**	=	.....

[Ustun & Rudin, ML '16]

Neural attention

Motivation

- Complex (neural) models come at the cost of interpretability
- Applications often need interpretable justifications — rationales.

this beer pours ridiculously clear with tons of carbonation that forms a rather impressive rocky head that settles slowly into a fairly dense layer of foam. this is a real good lookin' beer, unfortunately it gets worse from here ... first, the aroma is kind of bubblegum-like and grainy. next, the taste is sweet and grainy with an unpleasant bitterness in the finish. ... ... overall, the fat weasel is good for a fairly cheap buzz, but only if you like your beer grainy and bitter .

review with rationales

Ratings

Look: 5 stars

Aroma: 2 stars

(Slide credit: Tao Lei)

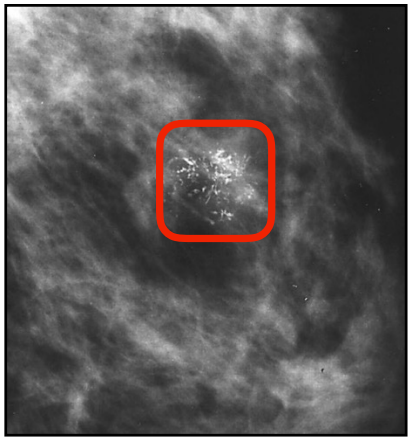
[Lei et al., EMNLP '16]

Neural attention

Motivation

- Complex (neural) models come at the cost of interpretability
- Applications often need interpretable justifications — rationales.

There is no evidence of extranodal extension.
BREAST (RIGHT), EXCISIONAL BIOPSY:
INVASIVE DUCTAL CARCINOMA (SEE TABLE #1). DUCTAL CARCINOMA IN-SITU, GRADE 1. ATYPICAL DUCTAL HYPERPLASIA. LOBULAR NEOPLASIA (ATYPICAL LOBULAR HYPERPLASIA). TABLE OF PATHOLOGICAL FINDINGS #1 INVASIVE CARCINOMA
... ...



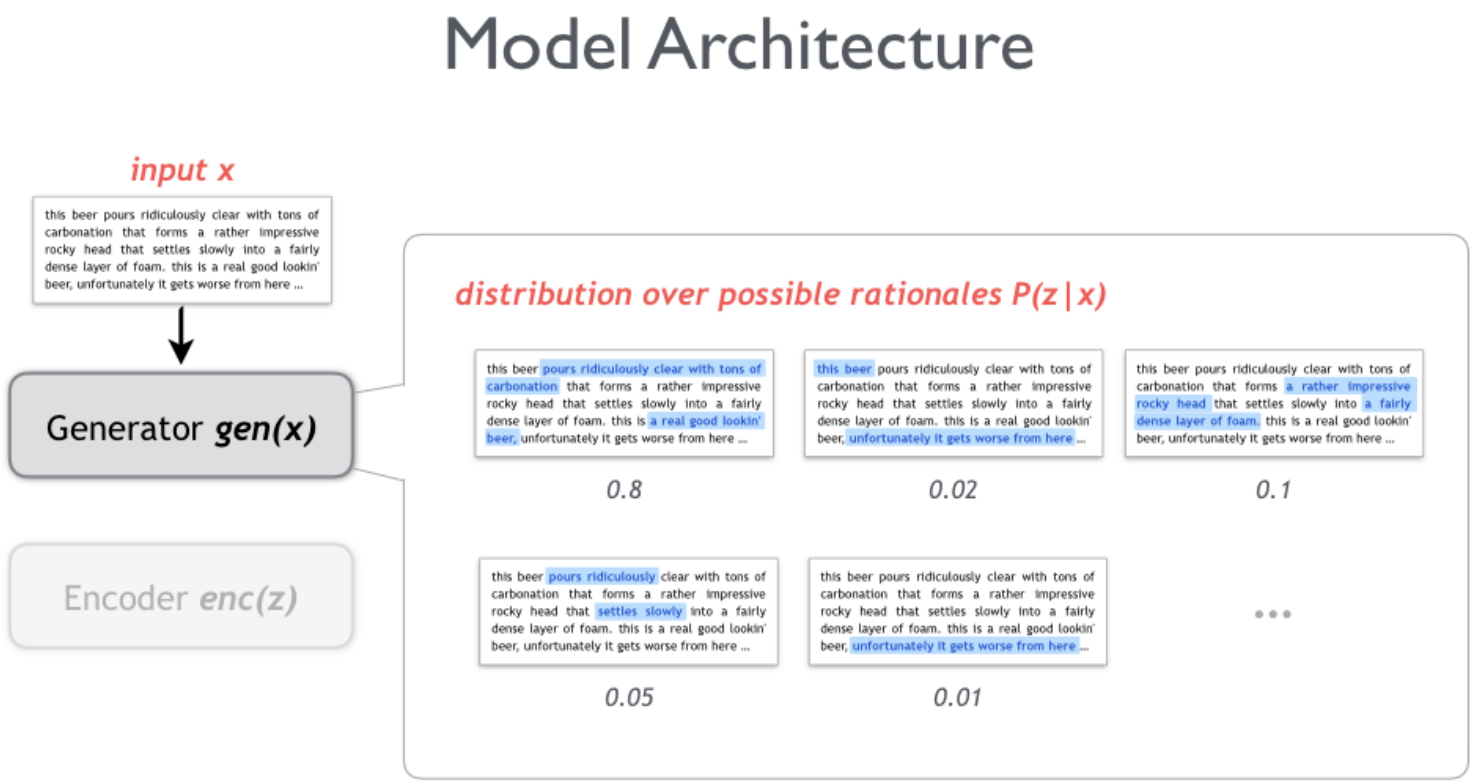
prediction: high risk of recurring cancer

Doctors won't trust machines, unless evidence is provided

(Slide credit: Tao Lei)

[Lei et al., EMNLP '16]

Neural attention

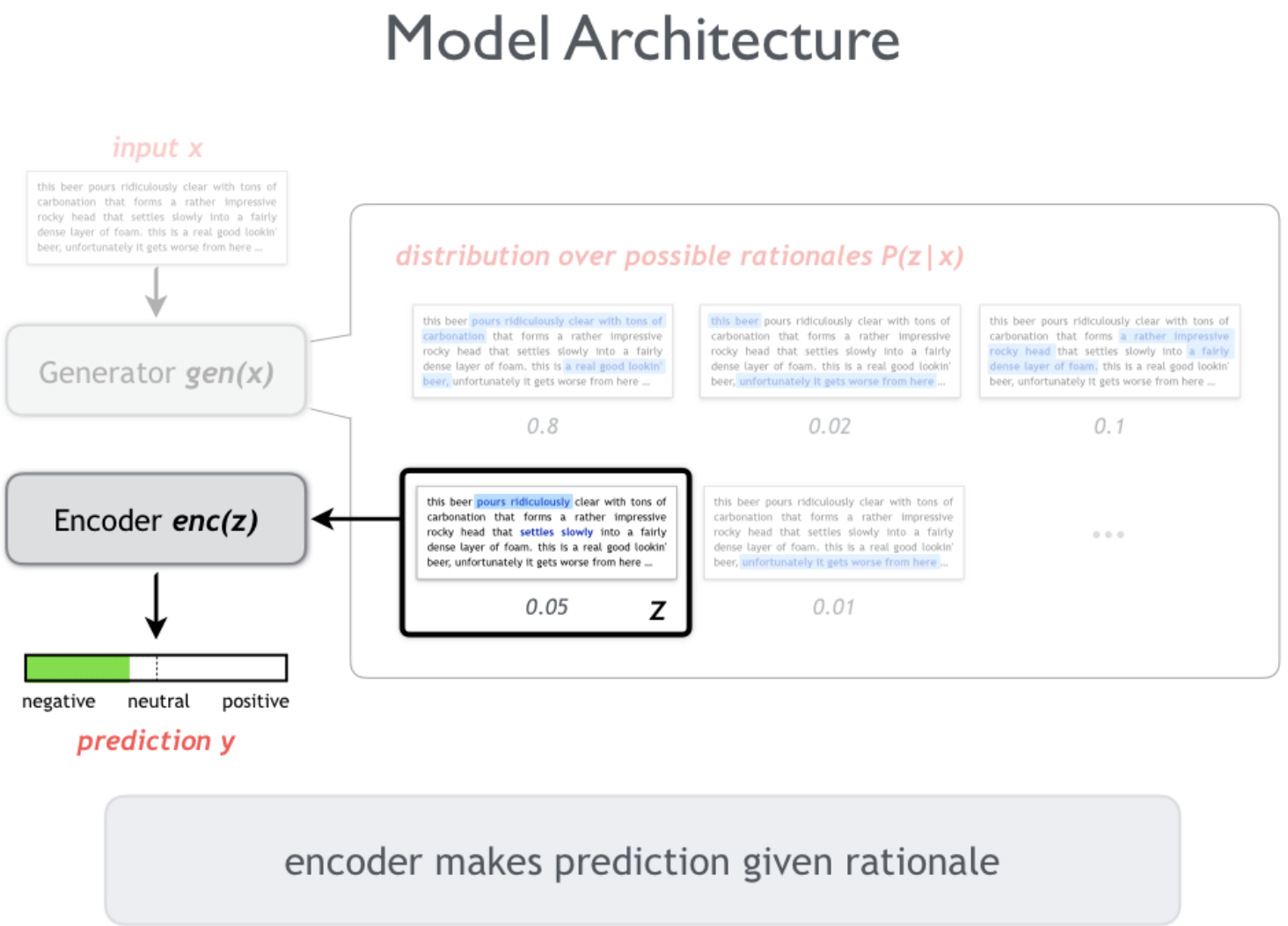


generator specifies the distribution of rationales

(Slide credit: Tao Lei)

[Lei et al., EMNLP '16]

Neural attention



(Slide credit: Tao Lei)

[Lei et al., EMNLP '16]

Neural attention

Evaluation: Parsing Pathology Report

Dataset: patients' pathology reports from hospitals such as MGH

Task: check if a disease/symptom is positive in text

binary classification for each category

Statistics: several thousand report for each category

pathology report is long (>1000 words) but structured

Model: use CNNs fro ***gen()*** and ***enc()***

(Slide credit: Tao Lei)

[Lei et al., EMNLP '16]

Neural attention

Evaluation: Parsing Pathology Report

Category:		*F-score:*
IDC	*Accession Number* <unk> *Report Status* Final *Type* Surgical Pathology … *Pathology Report:* LEFT BREAST ULTRASOUND GUIDED CORE NEEDLE BIOPSIES … **INVASIVE DUCTAL CARCINOMA poorly differentiated modified Bloom** Richardson grade III III measuring at least 0 7cm in this limited specimen Central hyalinization is present within the tumor mass but no necrosis is noted No lymphovascular invasion is identified No in situ carcinoma is present Special studies were performed at an outside institution with the following results not reviewed ESTROGEN RECEPTOR NEGATIVE PROGESTERONE RECEPTOR NEGATIVE …	98%
LCIS	… **Extensive** LCIS DCIS **Invasive** carcinoma of left breast FINAL DIAGNOSIS BREAST **LEFT LOBULAR CARCINOMA IN SITU PRESENT** ADJACENT TO PREVIOUS BIOPSY SITE SEE NOTE CHRONIC INFLAMMATION ORGANIZING HEMORRHAGE AND FAT NECROSIS BIOPSY SITE NOTE There is a second area of focal lobular carcinoma in situ noted with pagetoid spread into ducts No vascular invasion is seen The margins are free of tumor No tumor seen in 14 lymph nodes examined BREAST left breast is a <unk> gram 25 x 28 x 6cm left …	97%
LVI	FINAL DIAGNOSIS BREAST RIGHT EXCISIONAL BIOPSY INVASIVE DUCTAL CARCINOMA DUCTAL CARCINOMA IN SITU SEE TABLE 1 MULTIPLE LEVELS EXAMINED TABLE OF PATHOLOGICAL FINDINGS 1 INVASIVE CARCINOMA Tumor size <unk> X <unk> X 1 3cm Grade 2 **Lymphatic vessel invasion Present Blood vessel invasion Not identified** Margin of invasive carcinoma Invasive carcinoma extends to less than 0 2cm from the inferior margin of the specimen in one focus Location of ductal carcinoma in situ …	84%

(Slide credit: Tao Lei)

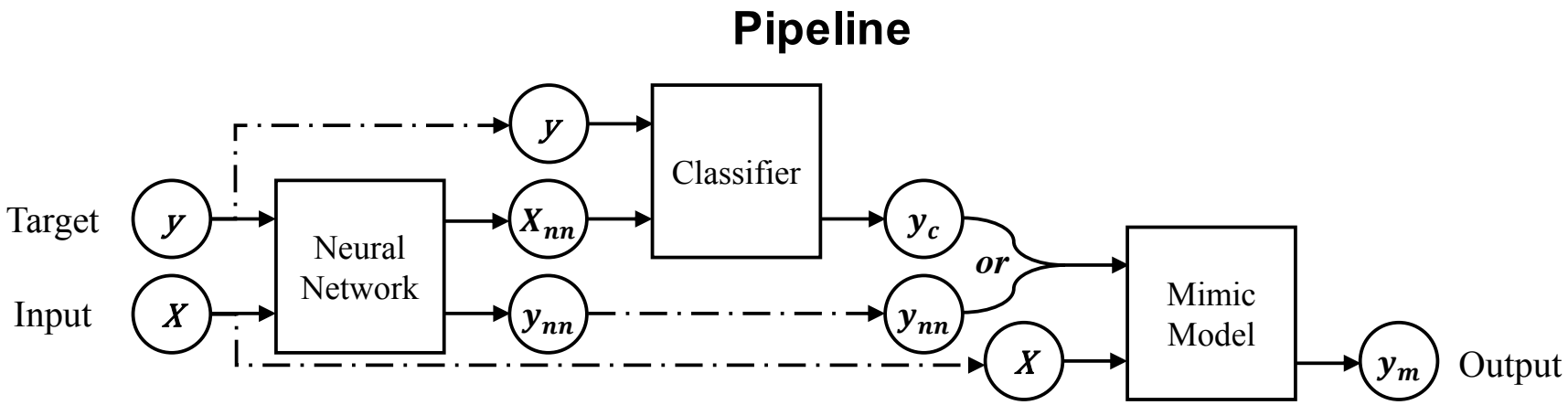
[Lei et al., EMNLP '16]

Outline of today's class

1. The mythos of model interpretability in health care
2. Learning intelligible models
3. **Post-hoc interpretability**

Compiling to a simpler model

- **Key idea:** use complex model (e.g. neural network) to train, then compile to a simpler model

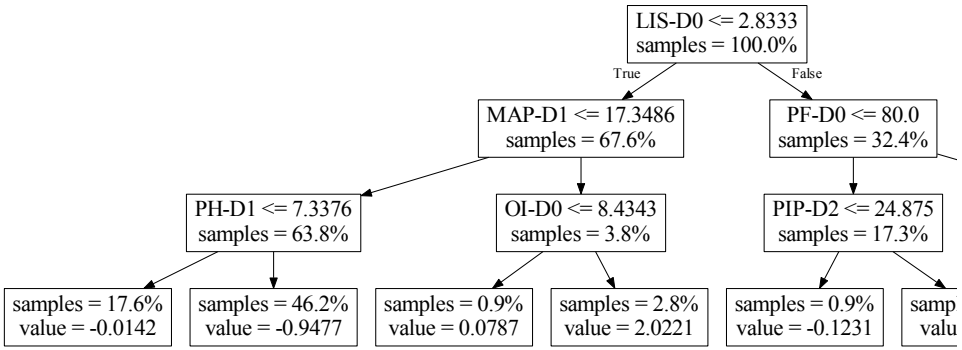


[Che et al., arXiv:1512.03542, '15]

Compiling to a simpler model

- **Key idea:** use complex model (e.g. neural network) to train, then compile to a simpler model

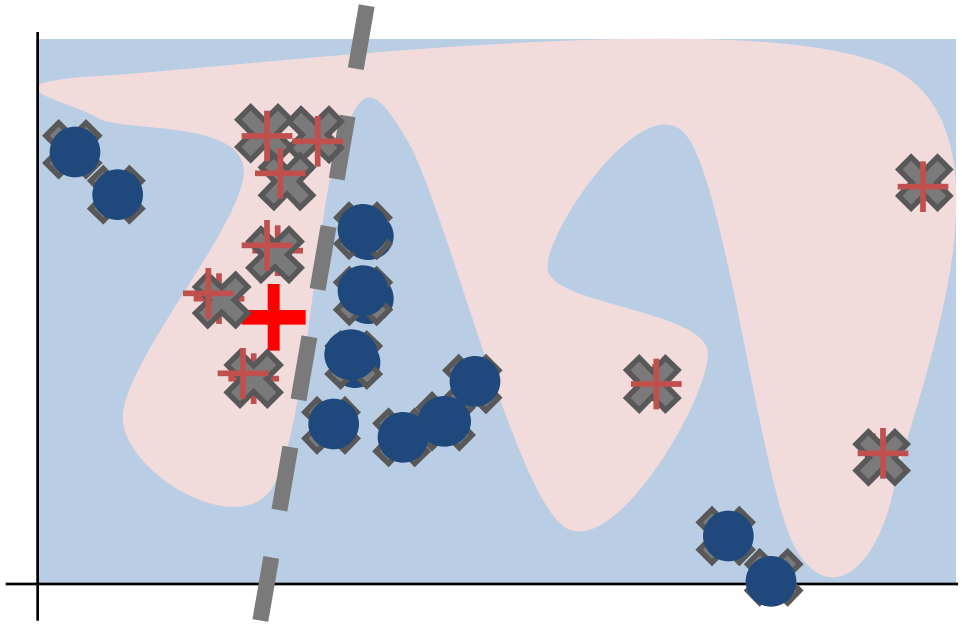
Method		Task			
		MOR		VFD	
		AUC	AUC(std)	AUC	AUC(std)
Baseline	SVM	0.6431	0.059	0.7248	0.056
	LR	0.6888	0.068	0.7602	0.053
	DT	0.5965	0.081	0.6024	0.044
	GBT	0.7233	0.065	0.7630	0.051
NN-based	DNN	0.7288	0.084	0.7756	0.053
	SDA	0.7313	0.083	0.7211	0.051
	LSTM	**0.7726**	0.062	0.7720	0.061
	LR-DNN	0.7300	0.084	0.7759	0.052
	LR-SDA	0.7459	0.068	**0.7818**	0.051
	LR-LSTM	0.7658	0.063	0.7665	0.063
Mimic	GBTmimic-DNN	0.7574	0.064	**0.7835**	0.054
	GBTmimic-SDA	0.7382	0.084	0.7194	0.049
	GBTmimic-LSTM	**0.7668**	0.059	0.7357	0.054
	GBTmimic-LR-DNN	**0.7673**	0.070	**0.7862**	0.058
	GBTmimic-LR-SDA	**0.7793**	0.066	**0.7818**	0.049
	GBTmimic-LR-LSTM	0.7555	0.067	0.7524	0.060



[Che et al., arXiv:1512.03542, '15]

LIME: *Local Interpretable Model-Agnostic Explanations*

1. Sample points around x_i
2. Use complex model to predict labels for each sample
3. Weigh samples according to distance to x_i
4. Learn new simple model on weighted samples
5. Use simple model to explain



(Slide credit: Marco Tulio Ribeiro)

[Ribeiro et al., KDD '16]