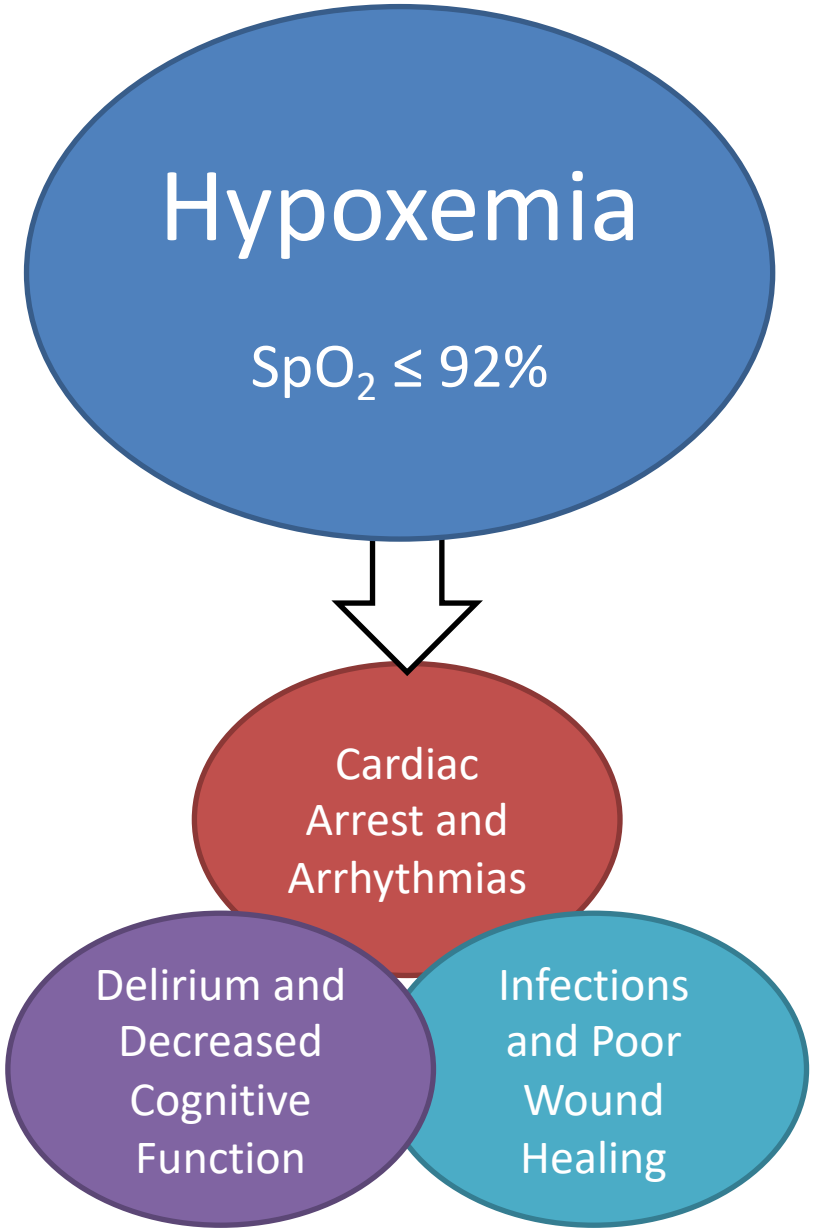
Machine Learning for Clinical Time Series

July 19, 2019

Block 4, Lecture 2
Applied Data Science
MMCi Term 4, 2019

Matthew Engelhard

Running Example: Predict Hypoxemia during Surgery



nature biomedical engineering

Article | Published: 10 October 2018

Explainable machine-learning predictions for the prevention of hypoxaemia during surgery

Scott M. Lundberg, Bala Nair, Monica S. Vavilala, Mayumi Horibe, Michael J. Eisses, Trevor Adams, David E. Liston, Daniel King-Wai Low, Shu-Fang Newman, Jerry Kim & Su-In Lee ✉

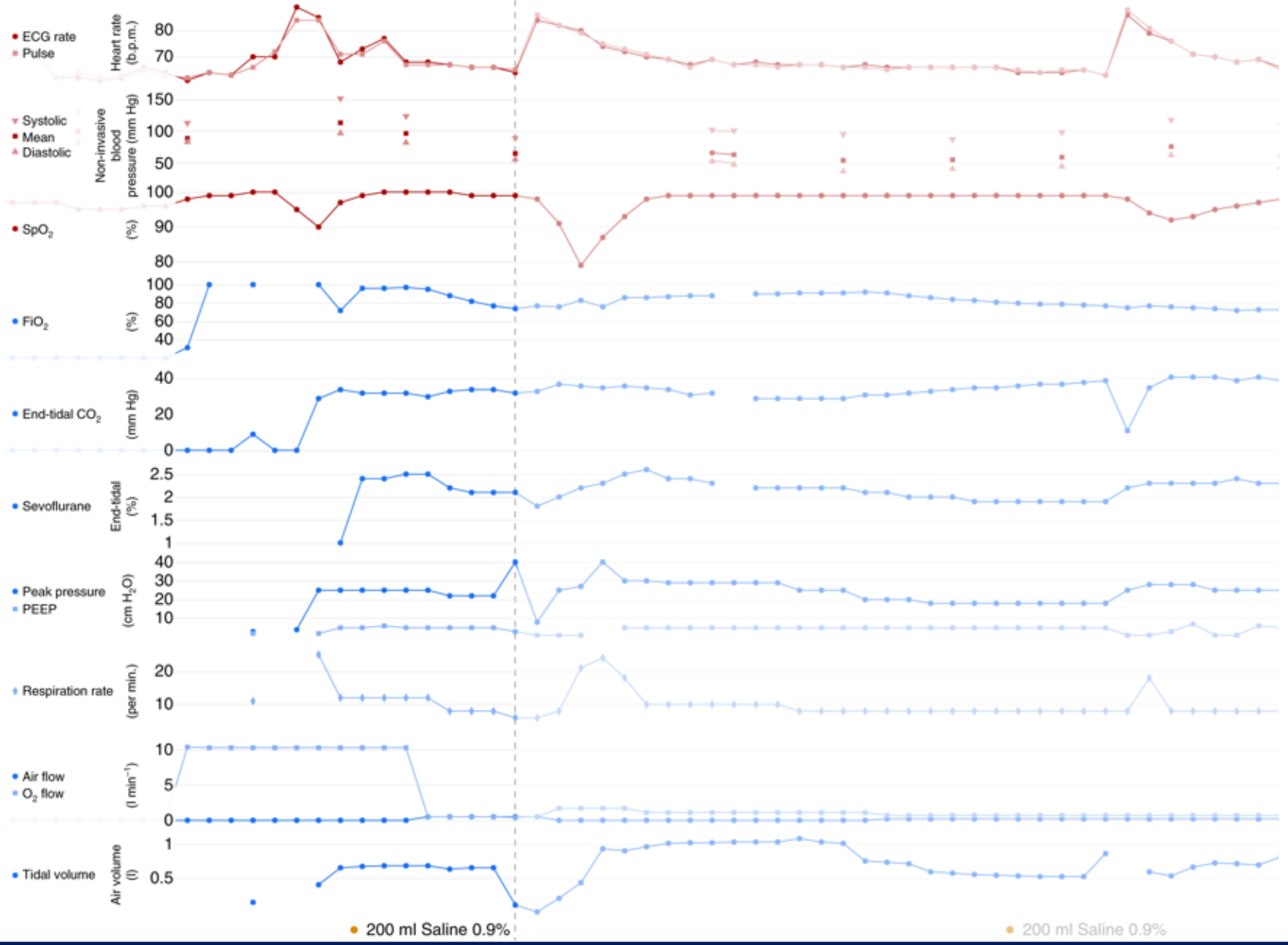
Nature Biomedical Engineering **2**, 749–760 (2018) | Download Citation ↓

Data: Anesthesia Information Management System (AIMS)

- 3,797 static extracted features for each patient from more than 20 original static sources

- an expanded superset of 3,905 real-time and static extracted features for each time point during anaesthesia care from the more than 20 original static sources as well as 45 different real-time data sources

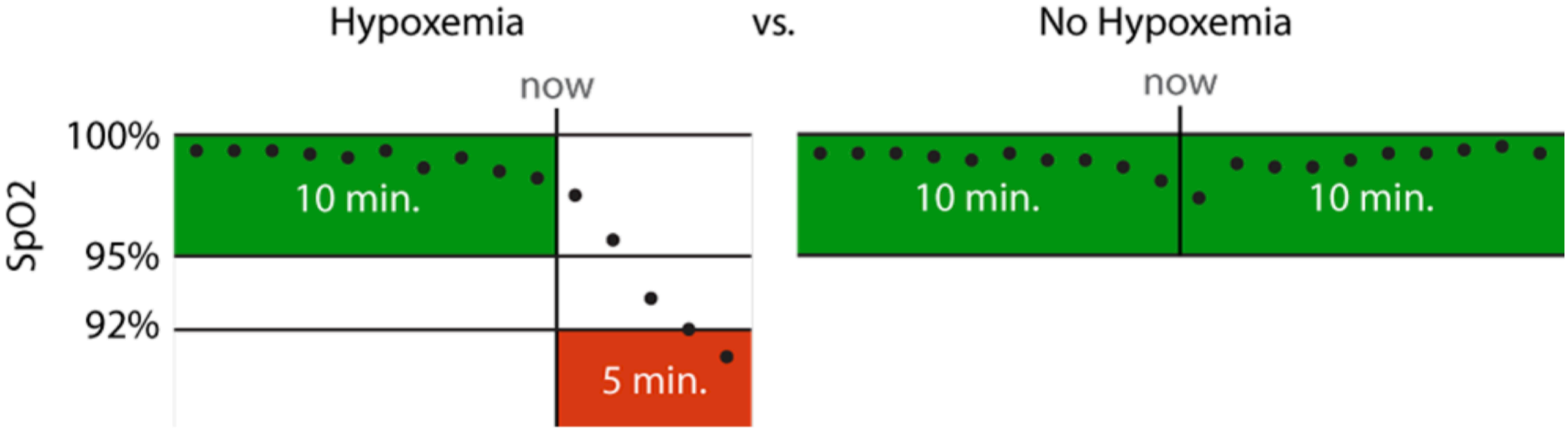
Element	Description
Patient	
Age	Age of patient
Sex	Sex of patient
ASA Physical Status	ASA physical status of patient before surgery
Height (cm)	Height of patient
Weight (kg)	Weight of patient
Patient class	Inpatient or outpatient
Procedure	
Procedure	Procedure description
Procedure code	Surgical procedure code
Billing codes	Anaesthesia Crosswalk/ procedure codes
Diagnosis	Diagnosis description
Diagnosis codes	ICD-9/10 codes
Location	Operating room location
Facility	Hospital facility
Emergency status	Whether case is an emergency or not – Y/N
Anaesthesia type	Type of anaesthesia
Case Events	
Anaesthesia Start	Time of Anaesthesia Start
In Room	Time of patient in room
Induction	Time of Induction start
Anaesthesia Ready	Time of Induction end or Anaesthesia Ready
Procedure Start	Time of Procedure start (incision)
Closing	Time of Closing
Procedure End	Time of Procedure End
Emergence	Time of start of emergence
Leave OR	Time of leave OR/ Transport to recovery
Anaesthesia End	Time of Anaesthesia End
Patient monitor/Ventilator Data	(Value, Time & Unit of measurement)
Heart Rate	Heart rate from ECG signal (Patient monitor)
O_2 Sat	O_2 saturation from pulse oximetry (Patient Monitor)
Pulse rate	Pulse rate from pulse oximetry (Patient Monitor)
NIBP – Sys	Cuff BP – systolic (Patient monitor)
NIBP – Dia	Cuff BP – diastolic (Patient monitor)
NIBP - Mean	Cuff BP – mean (Patient monitor)
Art BP - Sys	Arterial BP – systolic (Patient monitor)
Art BP - Dia	Arterial BP – diastolic (Patient monitor)
Art BP - Mean	Arterial BP – mean (Patient monitor)
PA - Sys	Pulmonary artery pressure – systolic (Patient monitor)
PA - Dia	Pulmonary artery pressure – diastolic (Patient monitor)
CVP	Central venous line pressure – mean (Patient monitor)
$EtCO_2$	End tidal CO_2 (Capnography) (Patient monitor)
Resp Rate	Measured respiration rate – capnography (Patient monitor)
FiO_2	Inspired O_2 (Patient monitor)
ET Sevo	End tidal Sevoflurane anaesthetic agent (Patient monitor)
ET Des	End tidal Desflurane anaesthetic agent (Patient monitor)
ET ISO	End tidal Isoflurane anaesthetic agent (Patient monitor)
ET N20	End tidal Nitrous oxide (Patient monitor)
BIS	Bispectral Index (Patient/BIS monitor)
SQI	Signal quality index of BIS (Patient/BIS monitor)
TEMP	Temperature (Patient monitor)
SvO_2	Mixed venous oxygenation (Patient monitor)
CCO	Continuous cardiac output (Patient monitor)
TV	Tidal volume (Patient monitor)
RATE	Ventilator rate setting (Ventilator)
PIP	Peak Inspiratory pressure (Ventilator)
PEEP	Positive End Expiration Pressure (Ventilator)
O_2 FLOW	O_2 flow rate (Ventilator)
Air FLOW	Air flow rate (Ventilator)
N_2O FLOW	N_2O Flow rate (Ventilator)
Cardiac Rhythm	Type of cardiac rhythm
Medications	Delivery information of medications
Time	Delivery time (start / end for infusion medications)
Drug Name	Drug Name
Dose	Drug dose
Dose Unit	Drug dose unit
Route	Drug route
Fluid totals	Fluid totals
Time	Time of fluid input or output
Fluid Name	Fluid name
Volume	Fluid volume recorded
Laboratory results	Intraoperative laboratory results
Time	Time of taking sample or Lab result time
Lab results	Lab result description
Sample type	Sample type – arterial/venous
Lab result	Result value
Unit	Unit of measurement
Notes	Attestations, AIMS Clinical notes, Note option selections
Time	Time associated with the note
Context code	Context code associated with a note or its selections
Note content	Note description field


ECG rate
Pulse
Heart rate (b.p.m.)
80
70
Systolic
Mean
Diastolic
Non-invasive blood pressure (mm Hg)
150
100
50
SpO2
(%)
100
90
80
FiO2
(%)
100
80
60
40
End-tidal CO2
(mm Hg)
40
20
0
Sevoflurane
End-tidal (%)
2.5
2
1.5
1
Peak pressure
PEEP
(cm H2O)
40
30
20
10
Respiration rate
(per min.)
20
10
Air flow
O2 flow
(l min−1)
10
5
0
Tidal volume
Air volume (l)
1
0.5
200 ml Saline 0.9%
200 ml Saline 0.9%

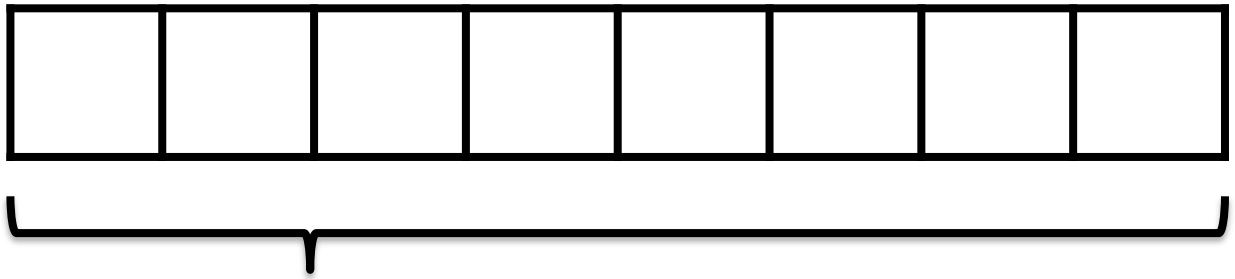
Prediction Tasks:

Initial Prediction: hypoxemia <u>at any time</u> based on static features

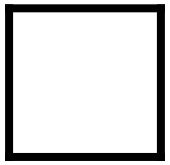
Real-time Prediction: hypoxemia <u>in the next 5 minutes</u> based on static features and real-time features collected up to that time point



Initial Prediction



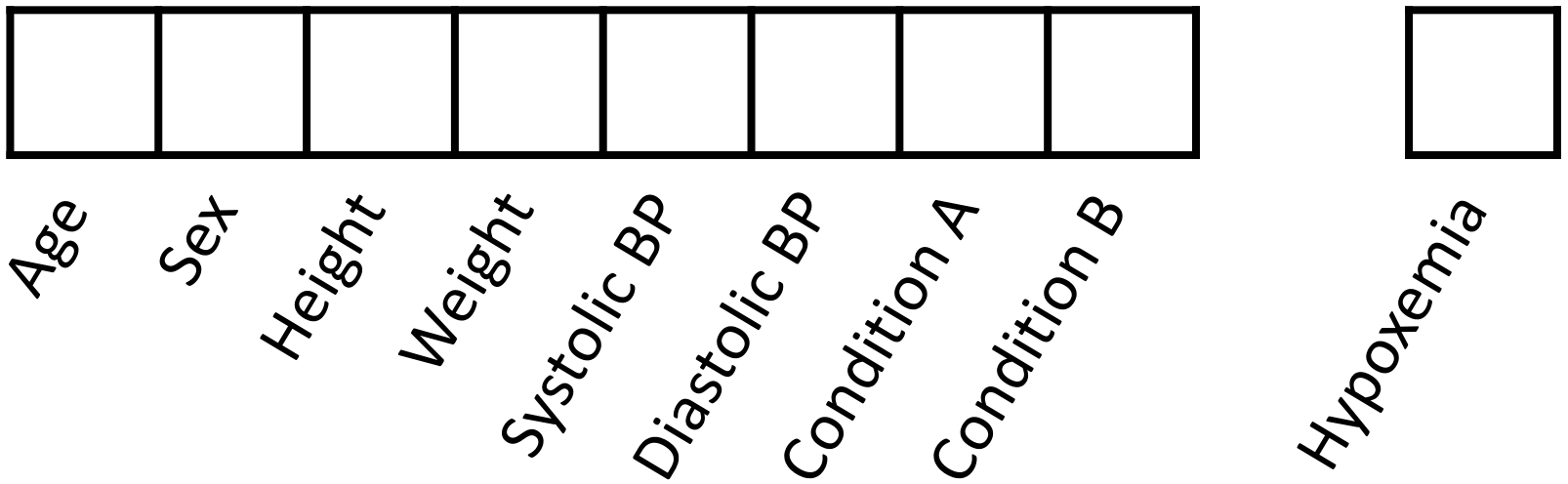
x, data/features for
a subject or patient



y, associated
value or label

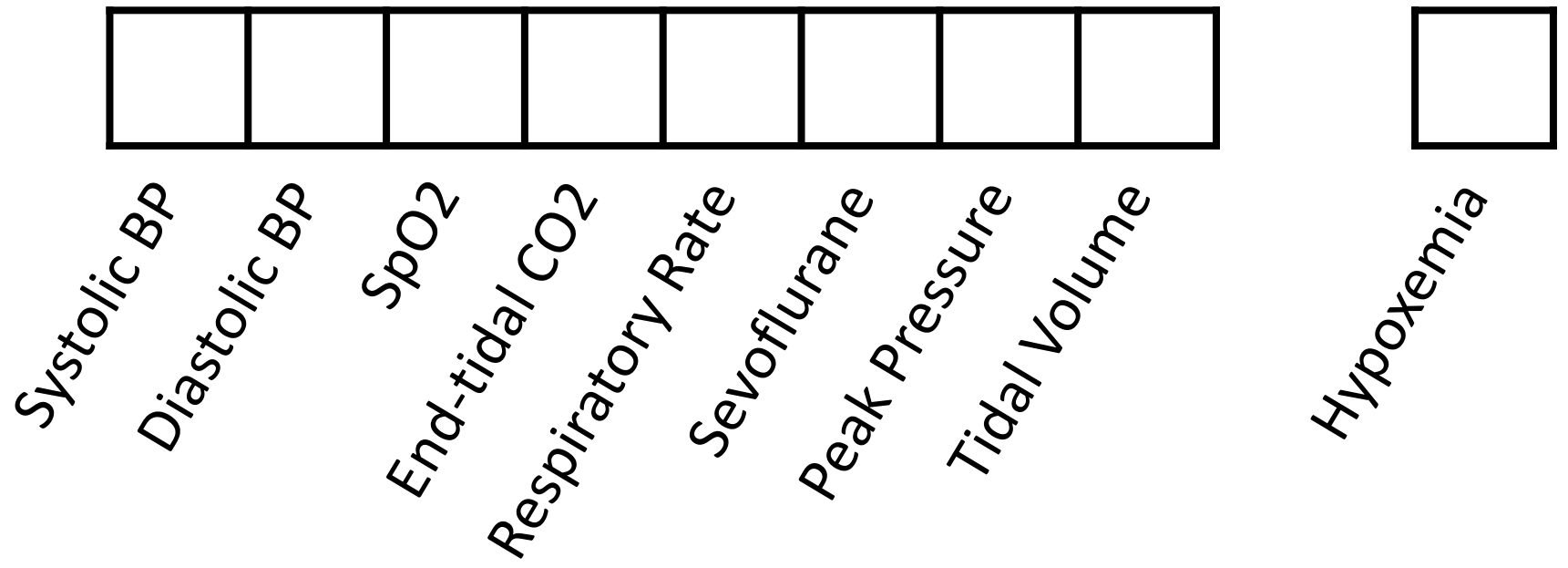
End goal: predict y from x

Initial Prediction



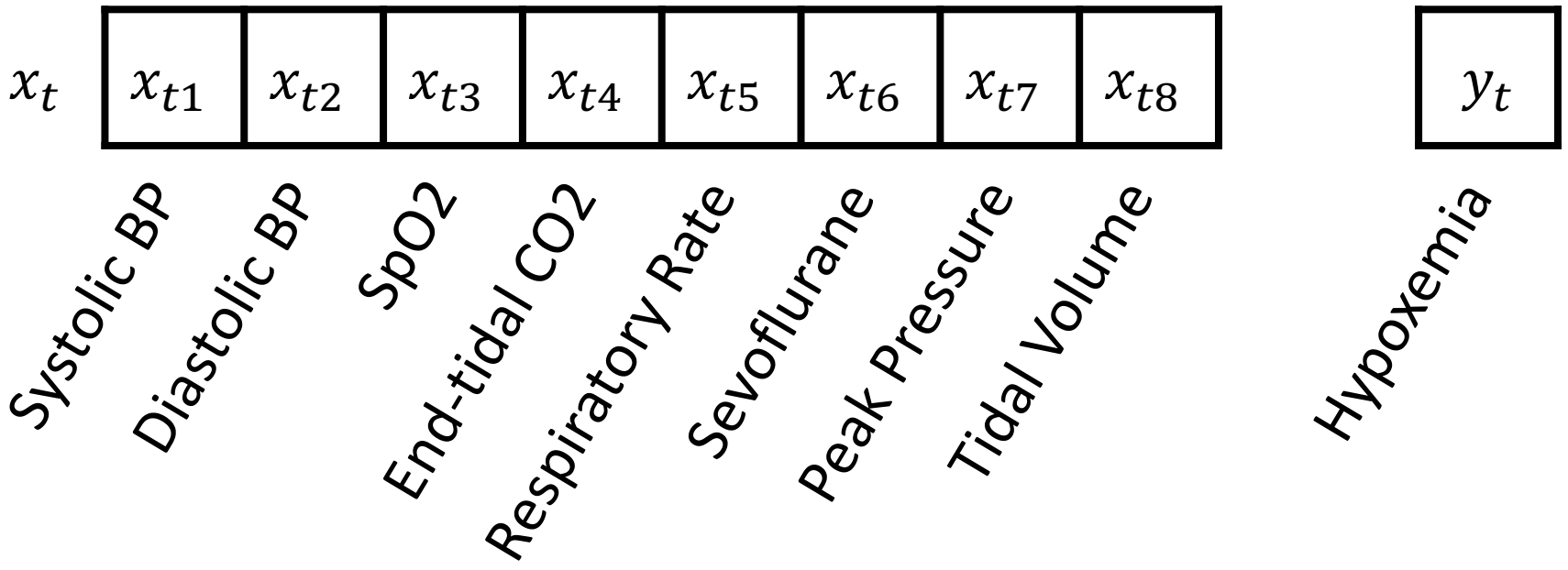
Goal: predict whether pt will become hypoxemic at any time during surgery

Real-time Prediction



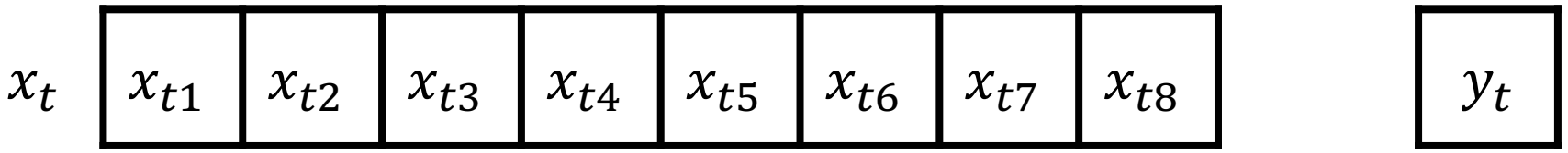
Goal: predict whether pt will become hypoxemic during the next 5 minutes

Real-time Prediction



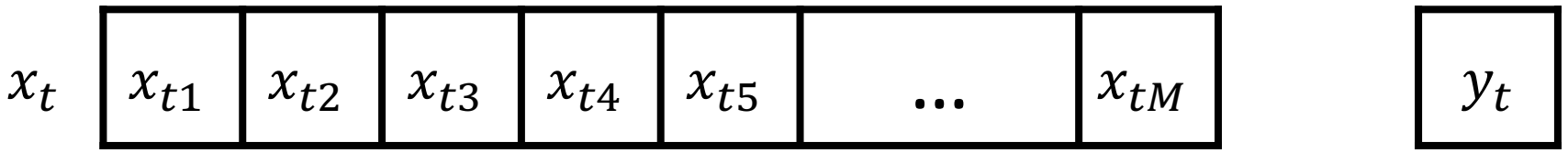
Goal: predict whether pt will become hypoxemic during the next 5 minutes

Real-time Prediction



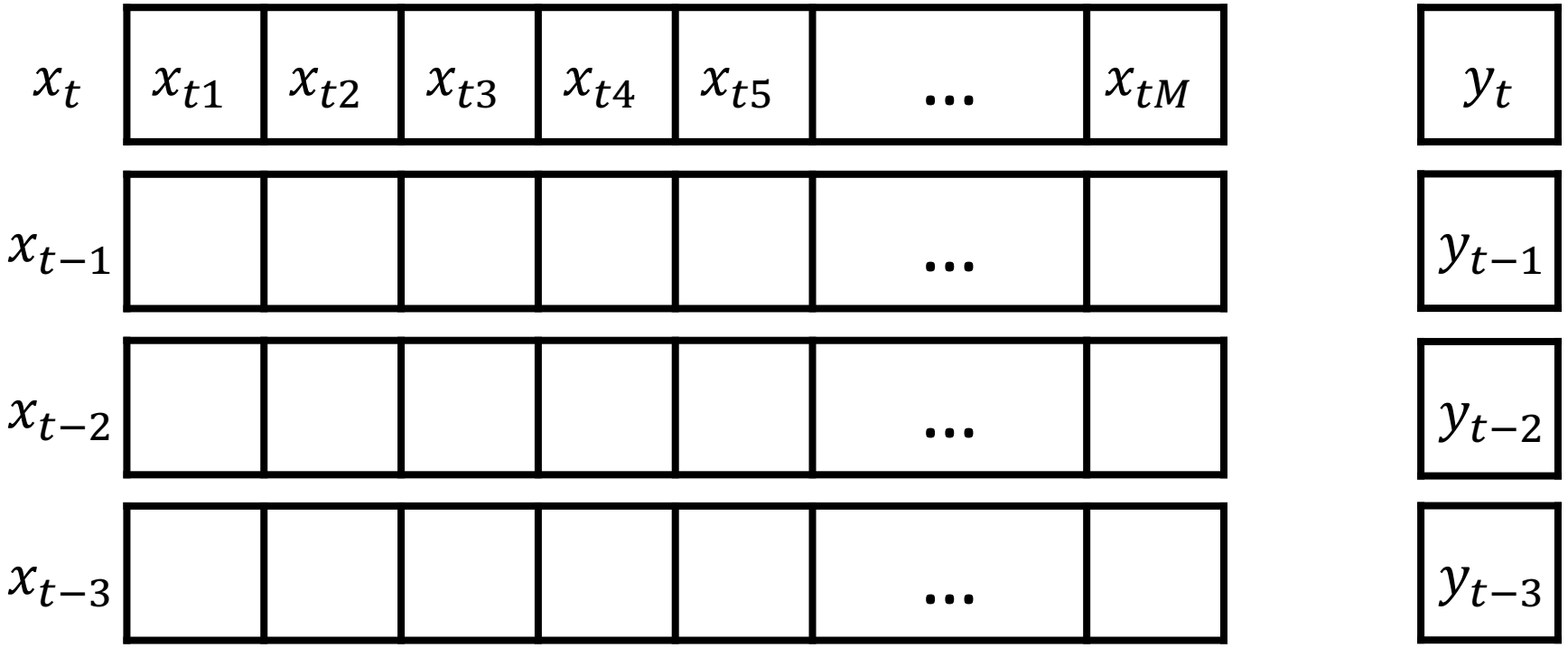
Goal: predict whether pt will become hypoxemic during the next 5 minutes

Real-time Prediction



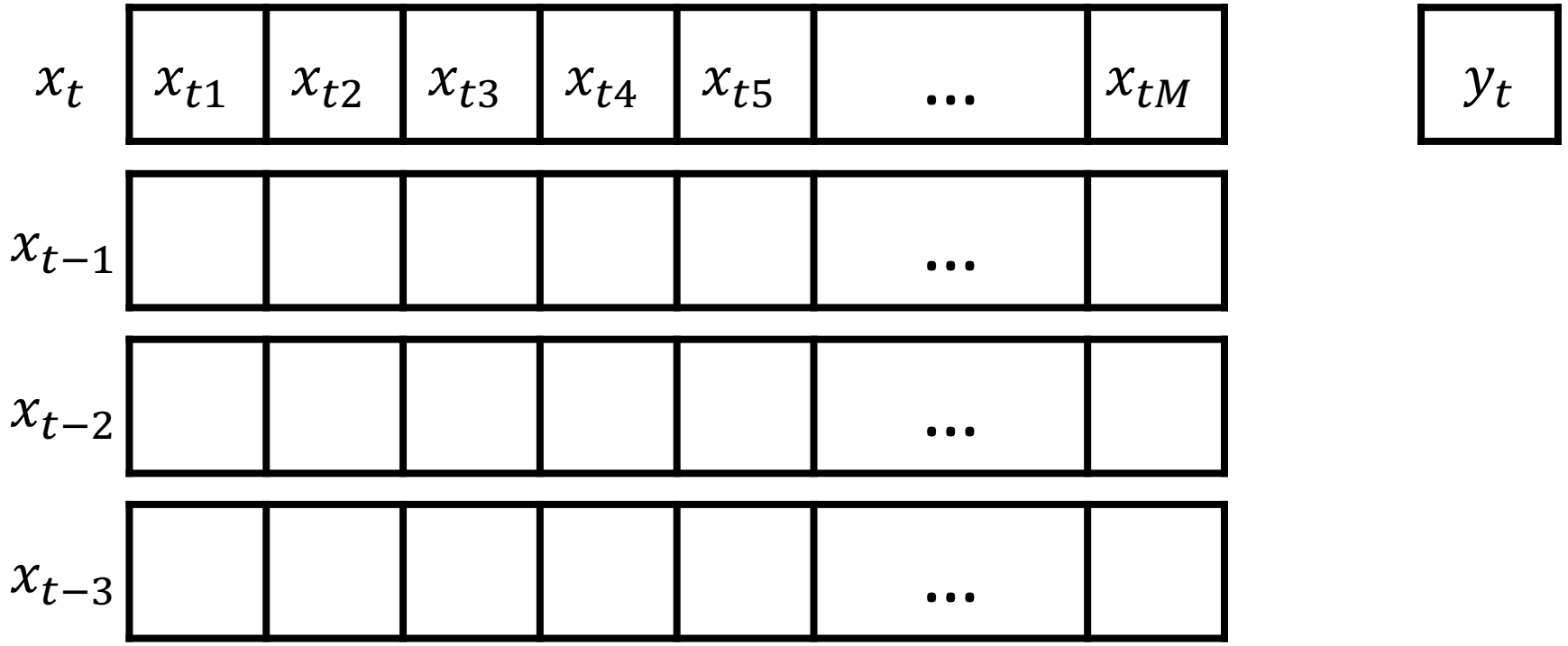
Goal: predict whether pt will become hypoxemic during the next 5 minutes

Real-time Prediction



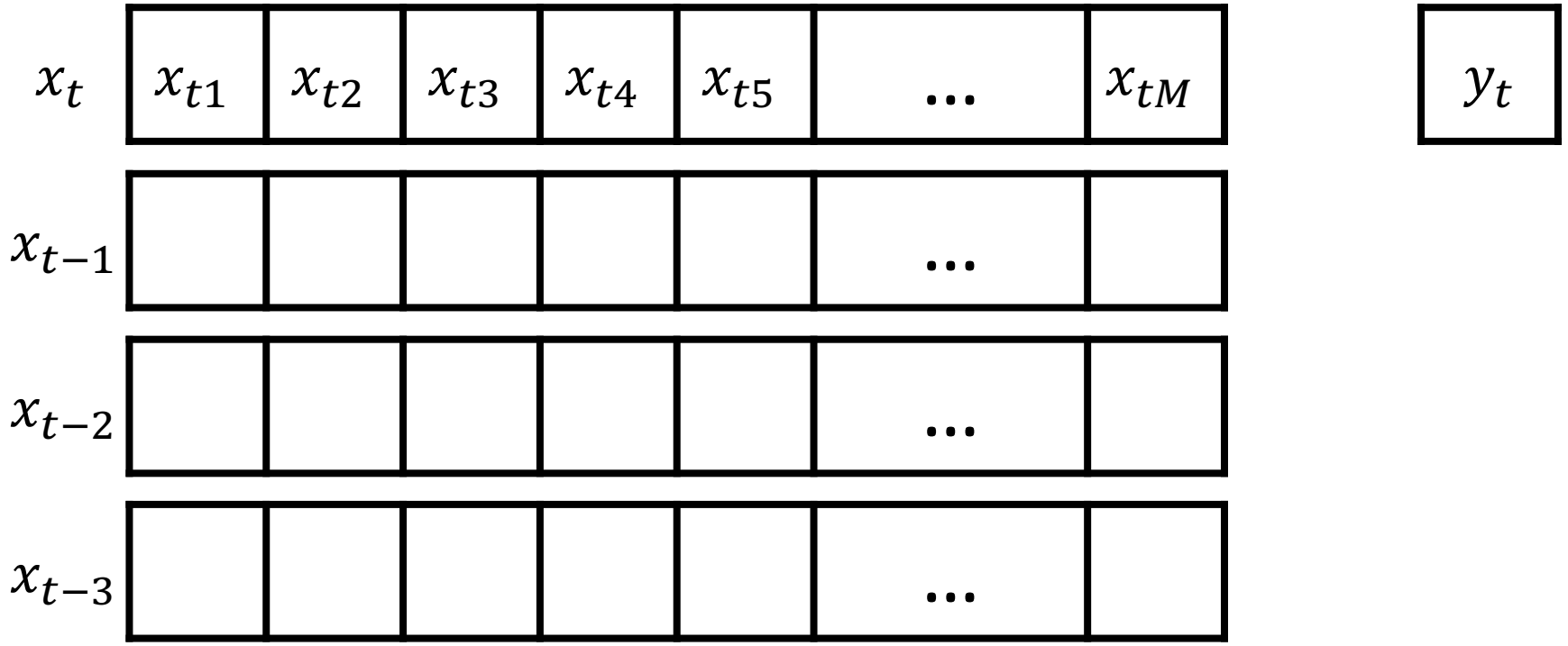
Goal: predict whether pt will become hypoxemic during the next 5 minutes

Real-time Prediction



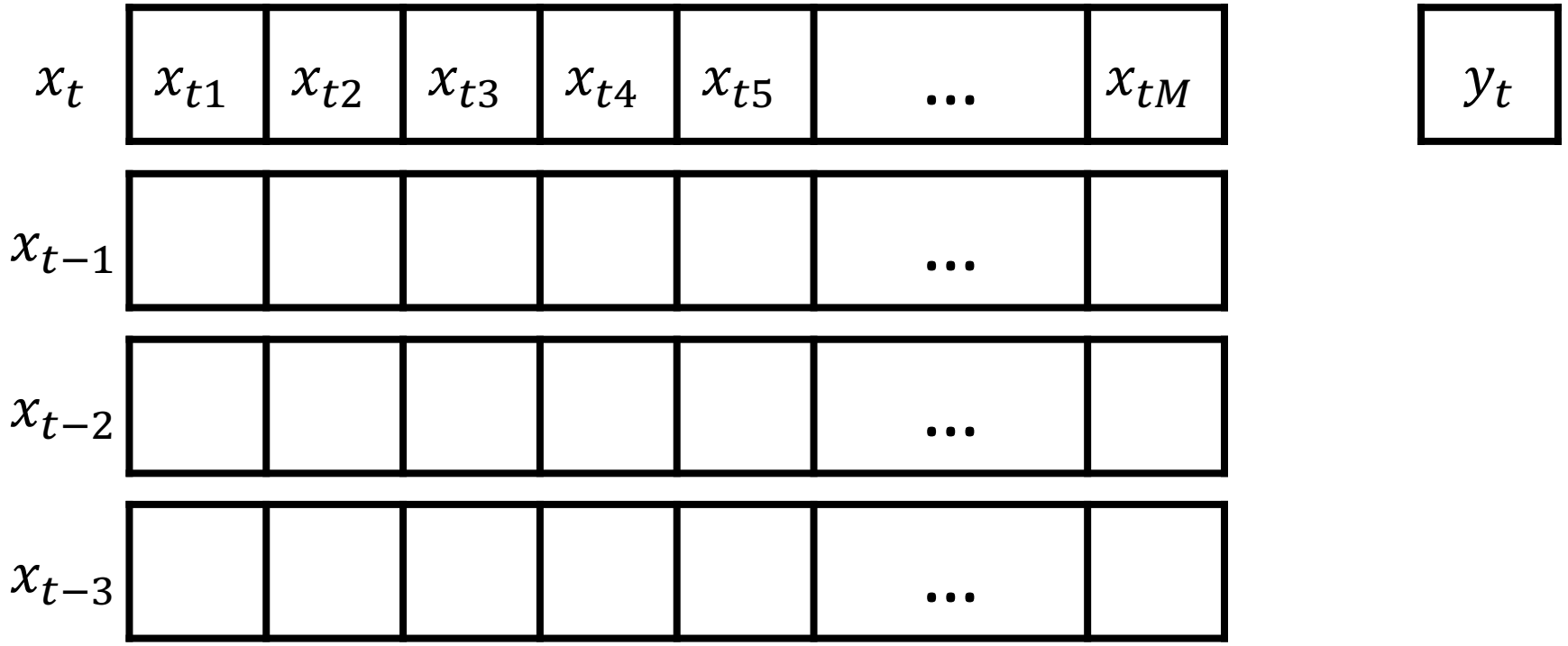
Goal: predict whether pt will become hypoxemic during the next 5 minutes

Goal: combine $x_t, \ldots, x_{t-\tau}$ into a fixed-length vector



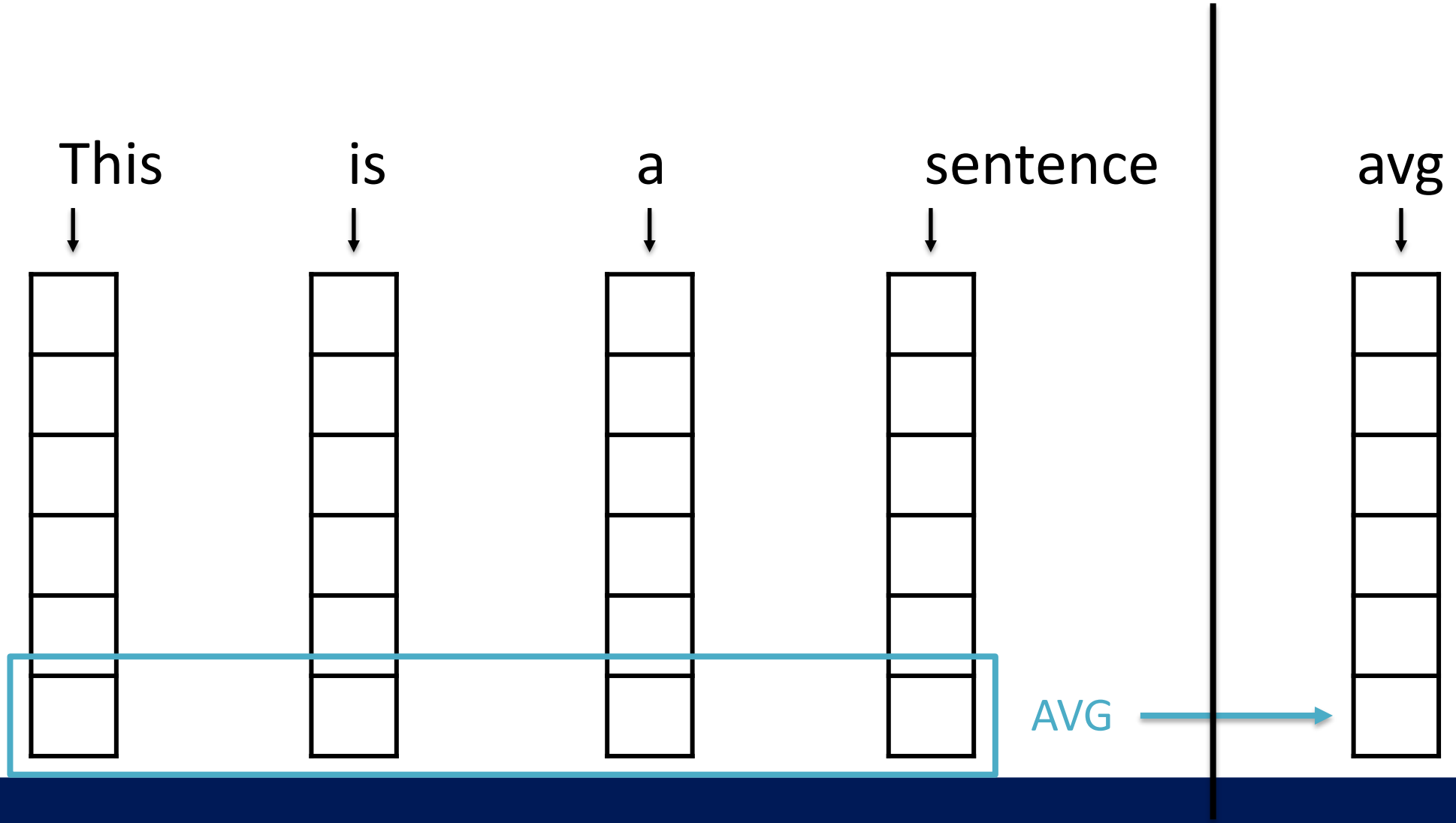
Goal: predict whether pt will become hypoxemic during the next 5 minutes

Goal: combine $x_t, \ldots, x_{t-\tau}$ into a fixed-length vector

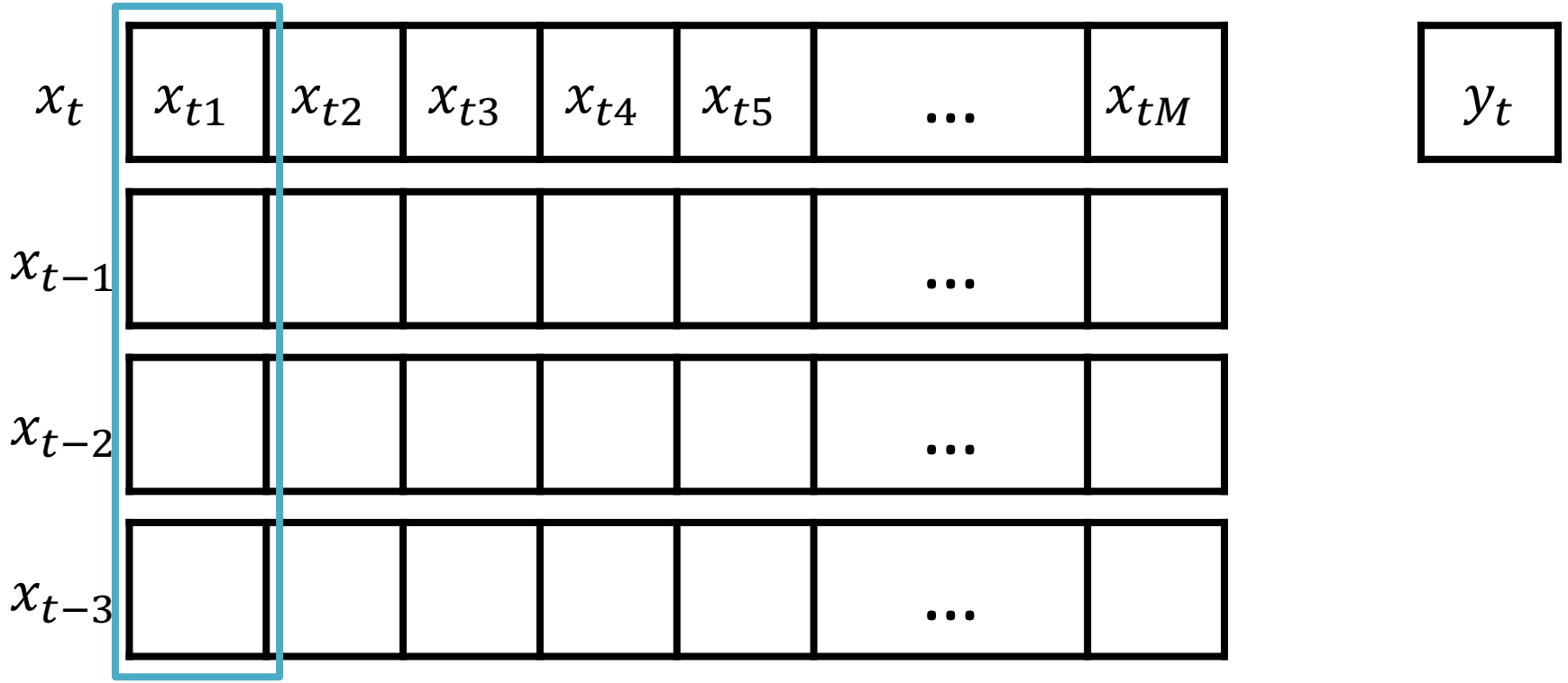


Goal: predict whether pt will become hypoxemic during the next 5 minutes

Recall: Convert variable-length sentence to fixed-length vector with max and average operations

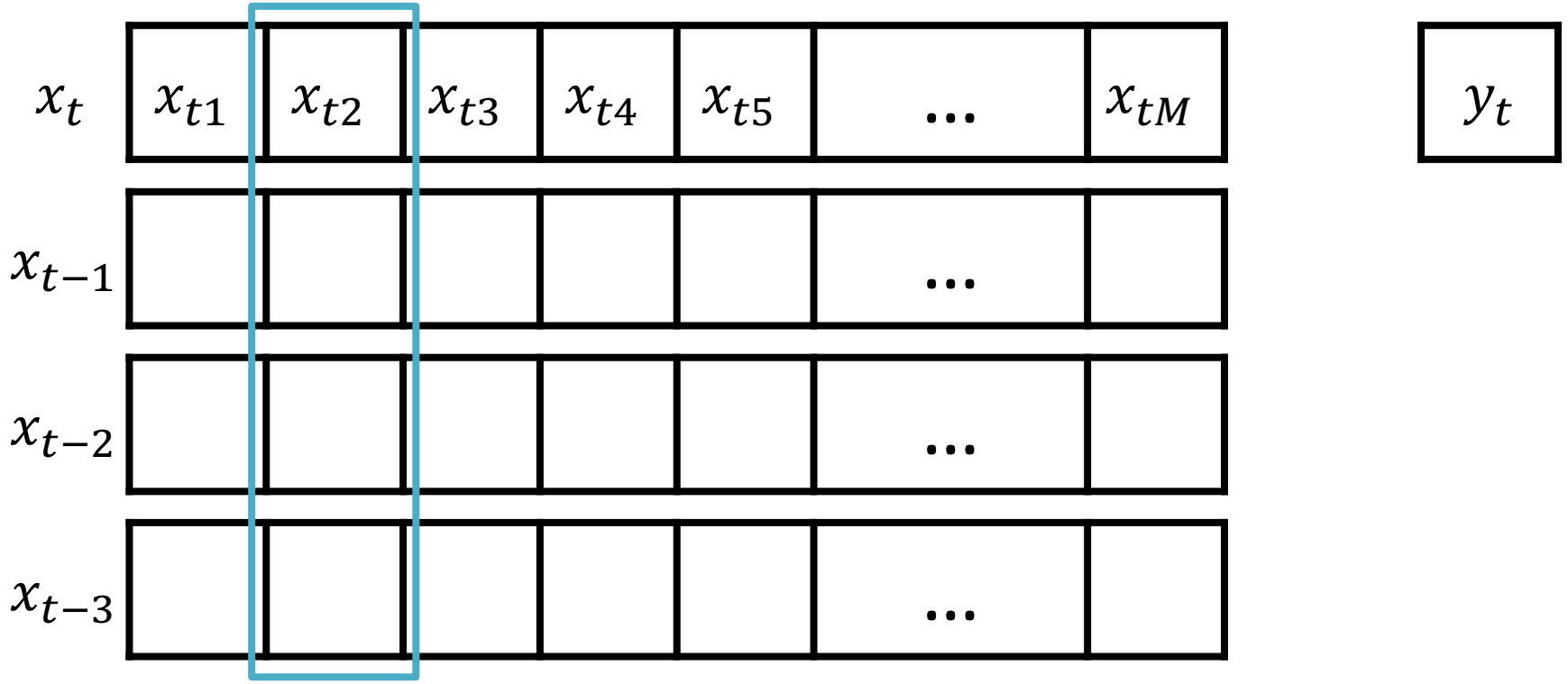


Goal: combine $x_t, \ldots, x_{t-\tau}$ into a fixed-length vector



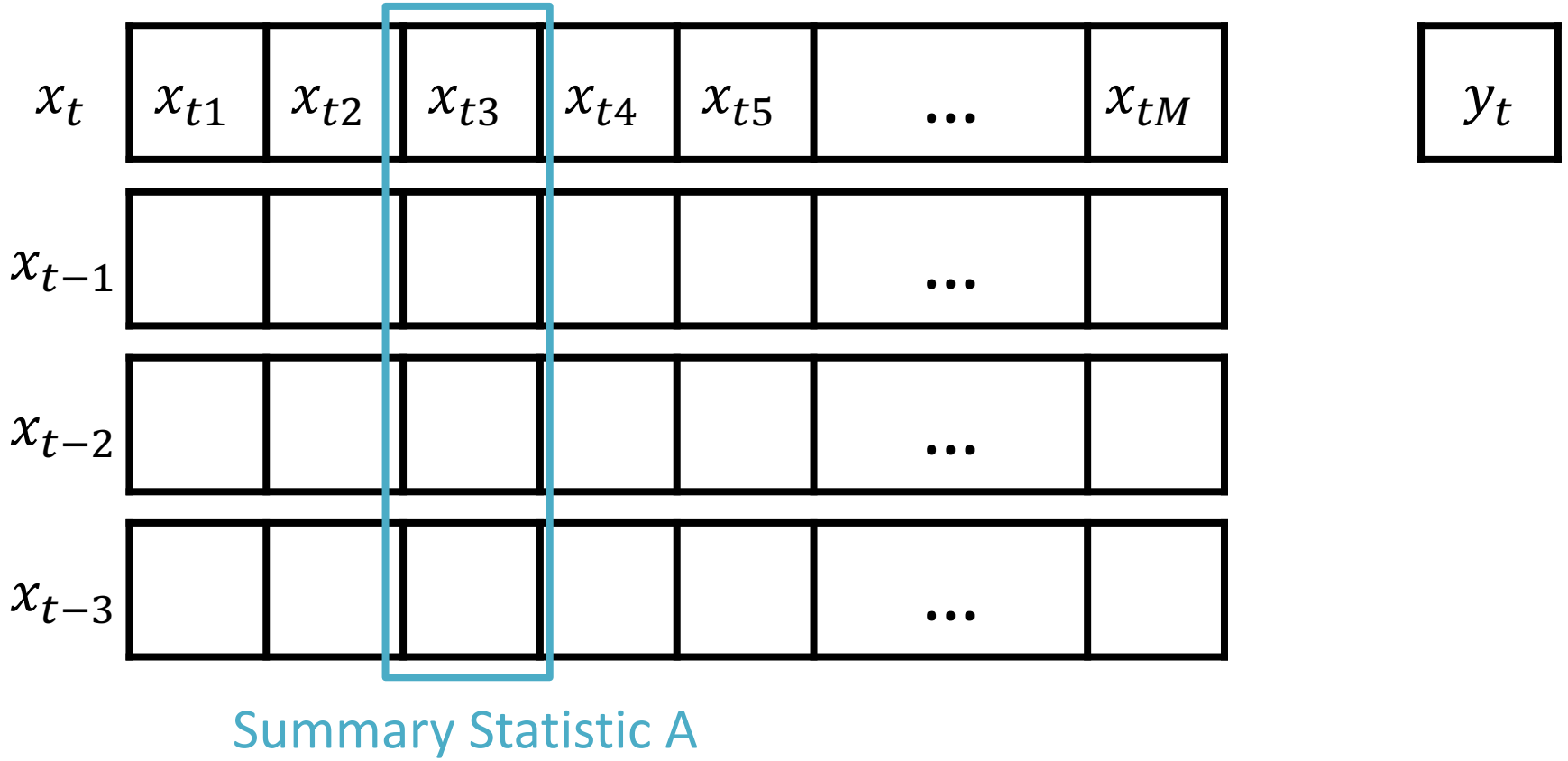
Goal: predict whether pt will become hypoxemic during the next 5 minutes

Goal: combine $x_t, \ldots, x_{t-\tau}$ into a fixed-length vector



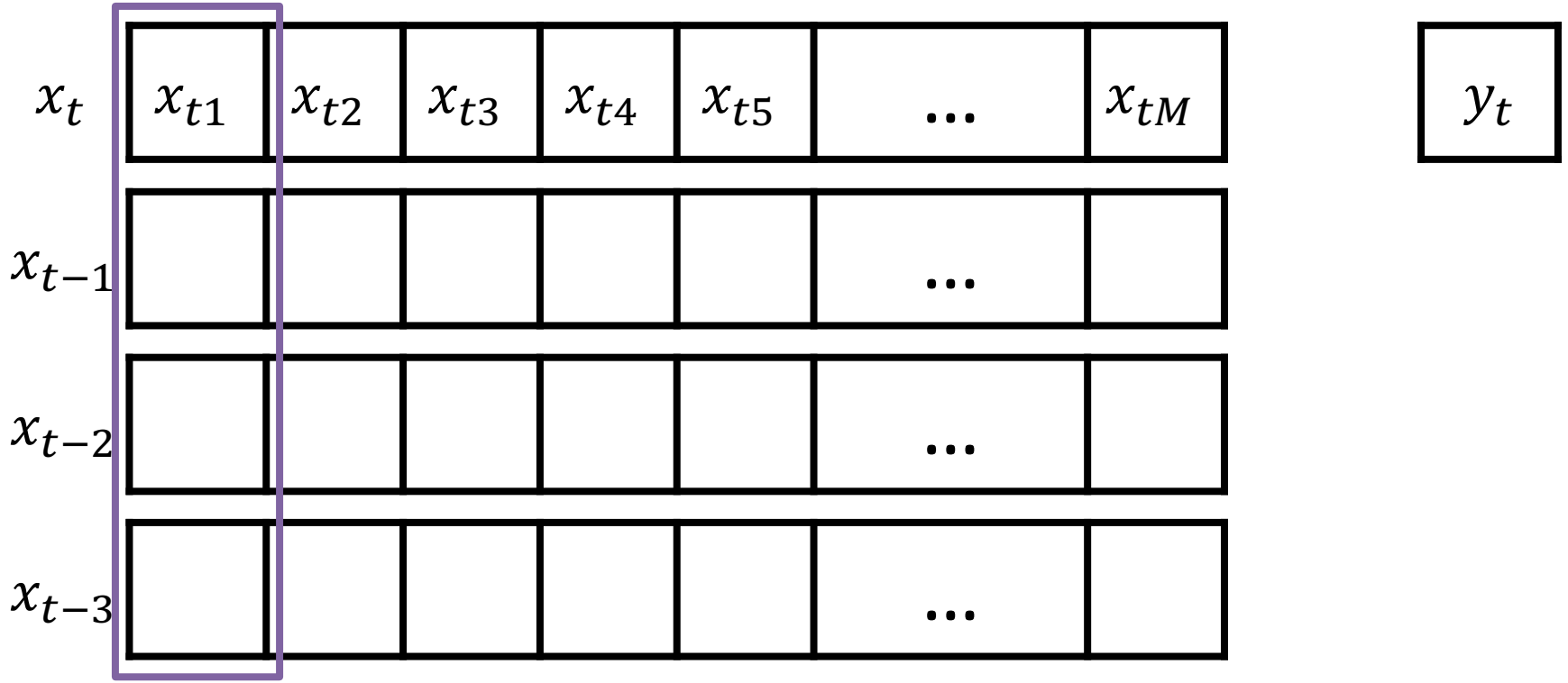
Goal: predict whether pt will become hypoxemic during the next 5 minutes

Goal: combine $x_t, \ldots, x_{t-\tau}$ into a fixed-length vector



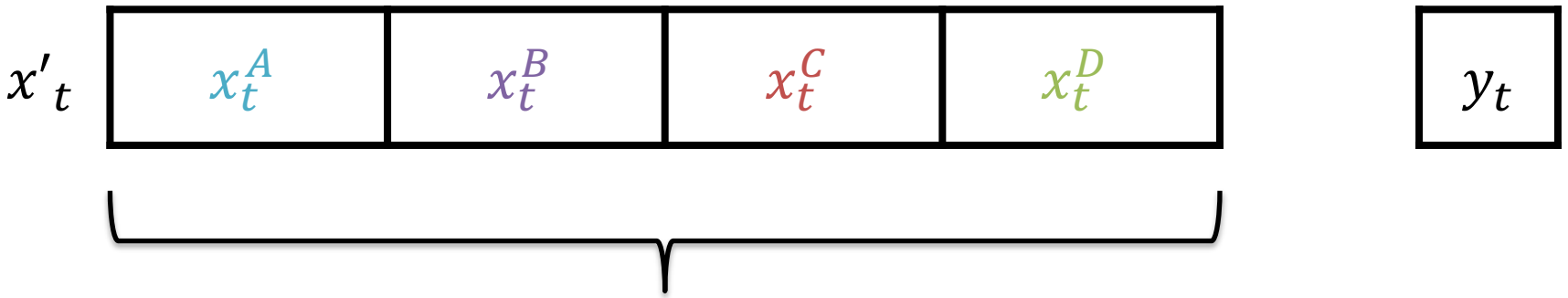
Goal: predict whether pt will become hypoxemic during the next 5 minutes

Goal: combine $x_t, \ldots, x_{t-\tau}$ into a fixed-length vector



Goal: predict whether pt will become hypoxemic during the next 5 minutes

Real-time Prediction



Vector of length $4M$ with summary stats A – D at time t

Goal: predict whether pt will become hypoxemic during the next 5 minutes

Summary Statistics

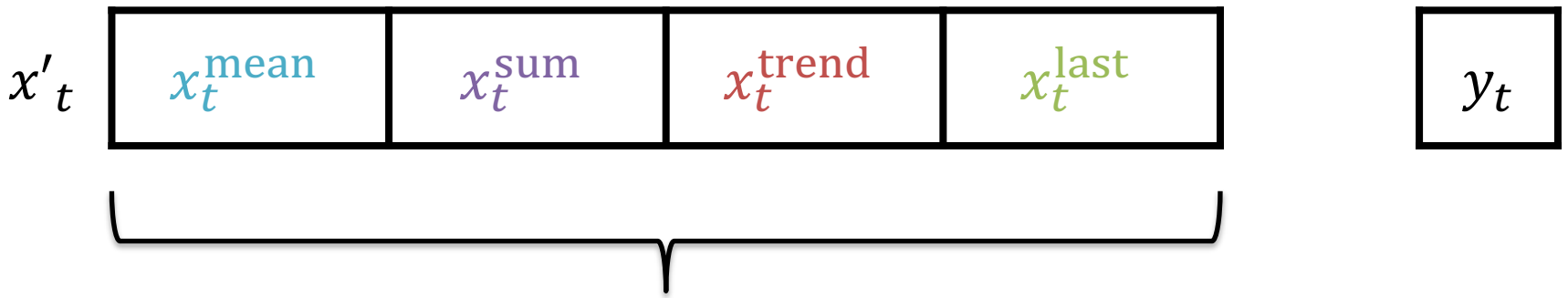
- Current value
- Average and/or extreme values
- Cumulative sums
- Recent trends
- Measurement frequency or time since last measurement

In the Hypoxemia Paper:

We summarized these unevenly sampled time-registered data into a fixed-length feature vector at any point in time using several complementary methods:

- **Last value**, which is zero before any data is recorded and the value of the data afterwards.
- **Exponentially decaying weighted average and variance** estimates using multiple decay rates (6 sec, 1 min, 5 min)
- **Exponentially decaying sum and a time since the last measurement**. For drug dose data (decay rates of 5 min and 1 hour)

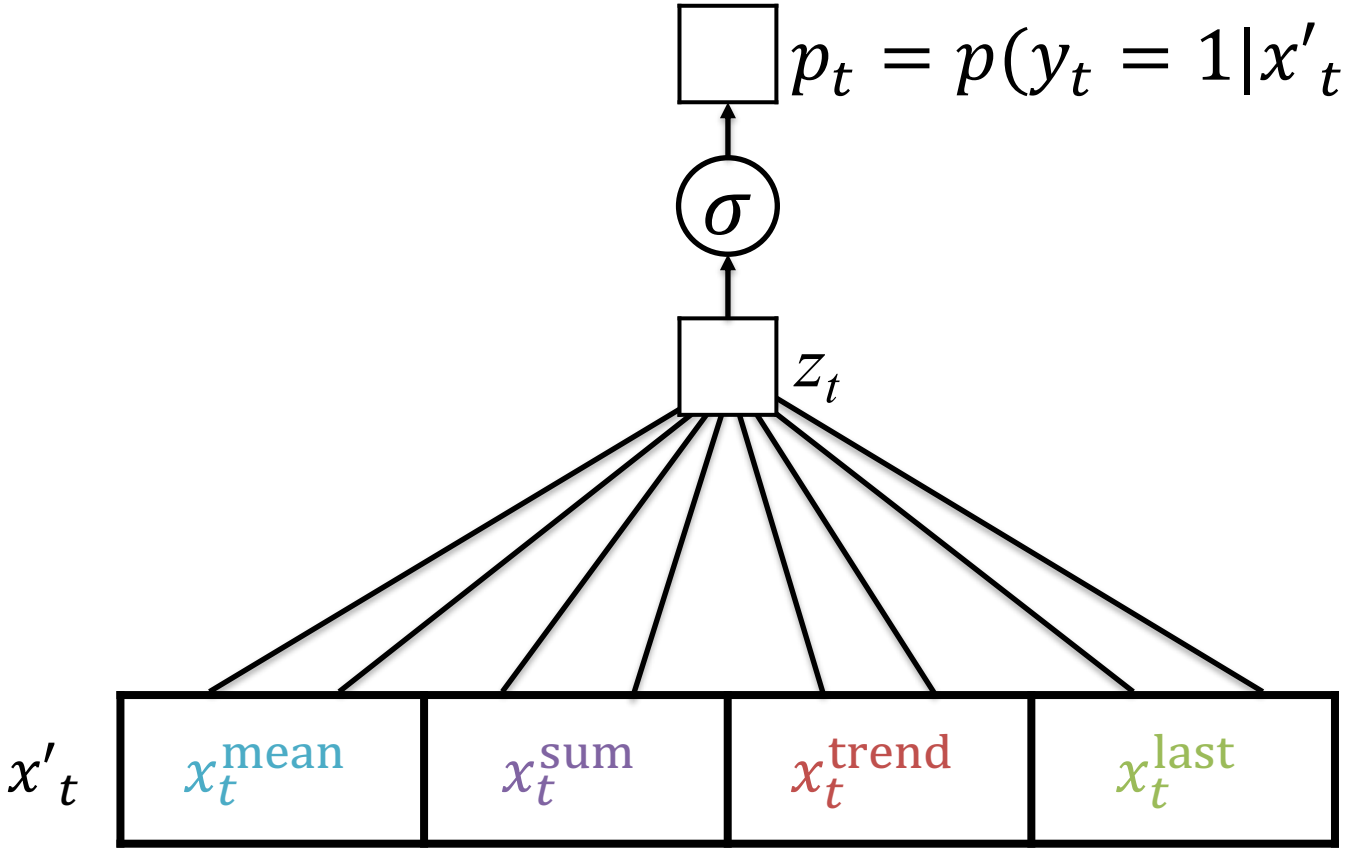
Real-time Prediction



Vector of length $4M$ with summary stats A – D at time t

Goal: predict whether pt will become hypoxemic during the next 5 minutes

Now we can use logistic regression (or any other classifier)



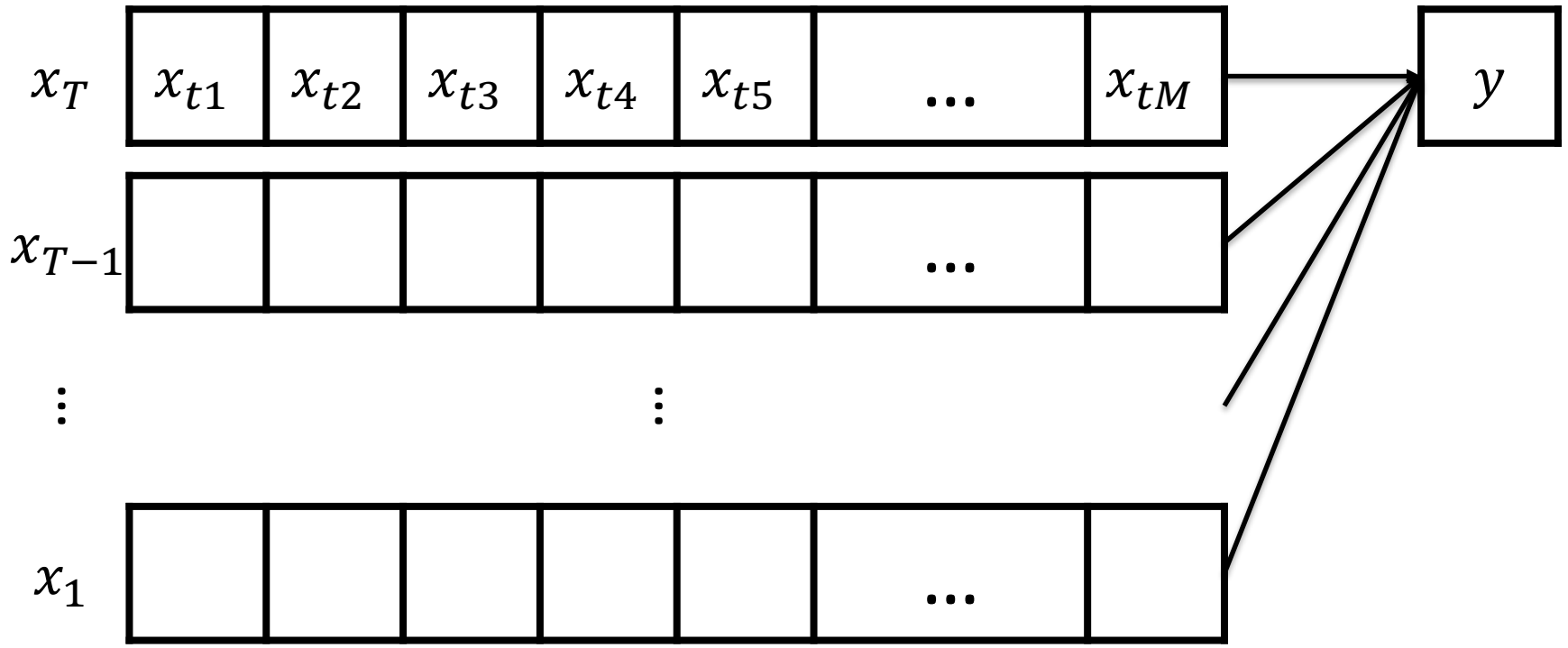
Contrast with text summary stats

- Cannot look at future values
- Window length is fixed

Challenges

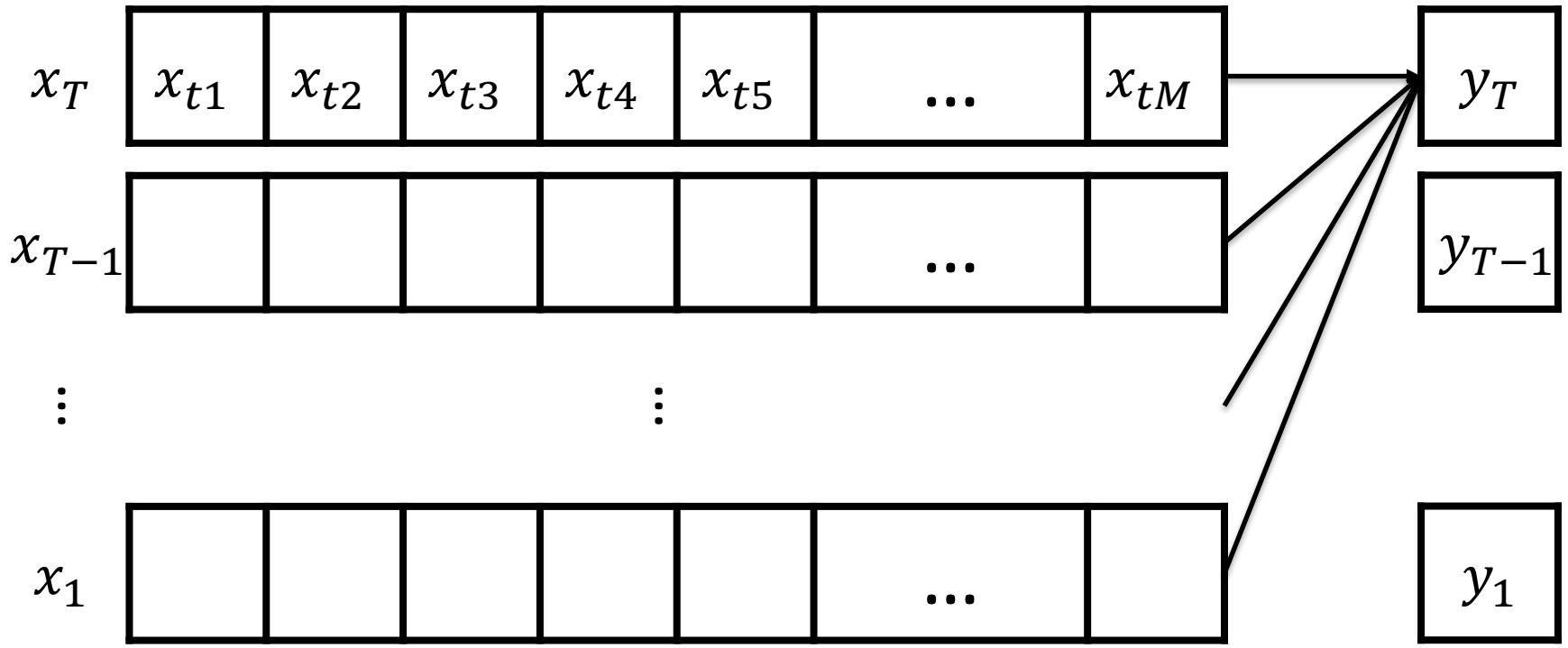
- Missing values
- Inconsistent or unequal measurement intervals
- Appropriate summary stats vary between measurement types

Prediction task A: one label per series



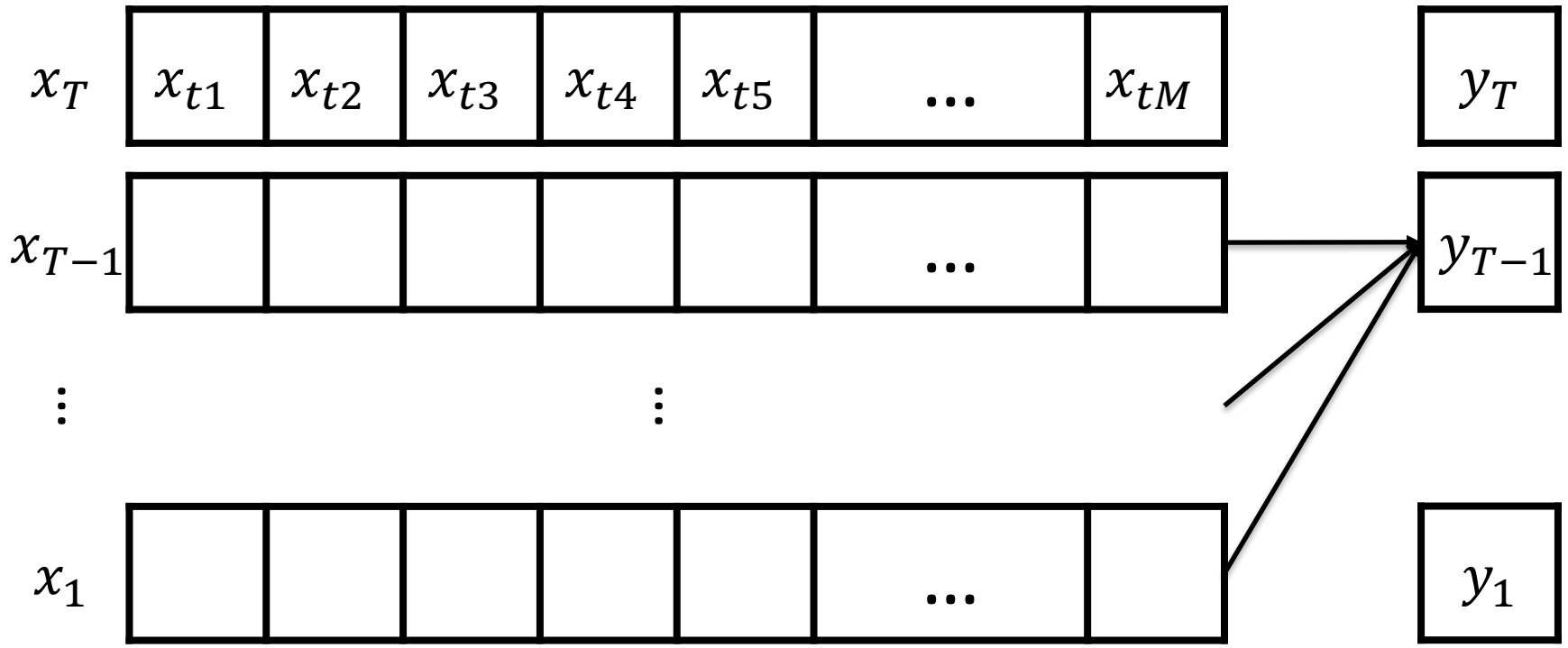
Goal: predict whether pt will become hypoxemic at any time during surgery

Prediction task B: one label per time step



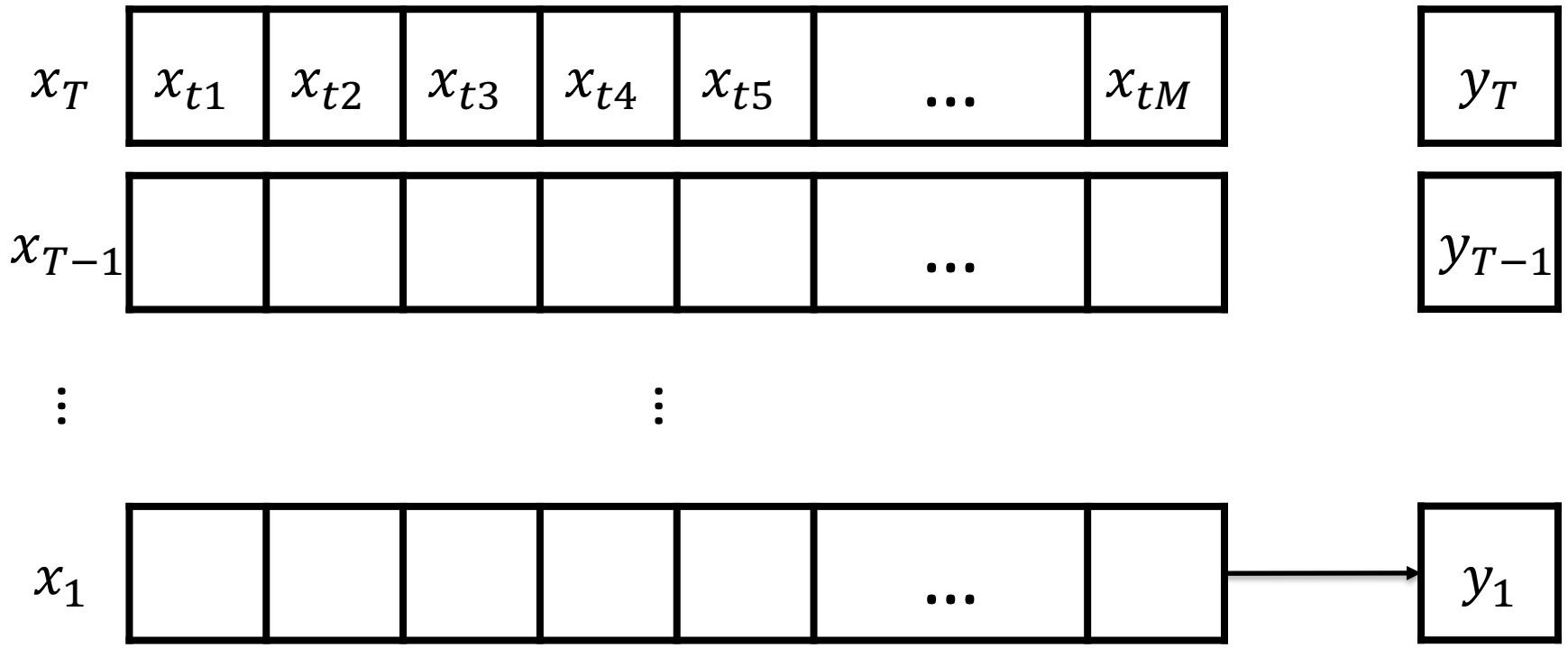
Goal: predict whether pt will become hypoxemic during the next 5 minutes

Prediction task B: one label per time step



Goal: predict whether pt will become hypoxemic during the next 5 minutes

Prediction task B: one label per time step



Goal: predict whether pt will become hypoxemic during the next 5 minutes

Hypoxemia Detection

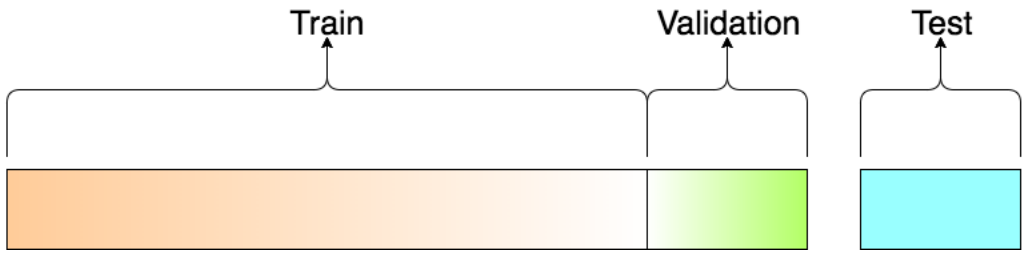
MODEL DEVELOPMENT & RESULTS

Test / Validation / Train

Initial prediction:
- Training: 42,420 procedures (each a single surgical case)
- Validation: 5,649 procedures
- Test: 5,057 procedures

Real-time prediction:
- Training: 8,087,476 per-minute time points
- Validation: 1,053,629 per-minute time points
- Test: 963,674 per-minute time points

All time points from the same procedure were included in the same sample set and no missing data imputation was performed.



"To ensure that there was no bias towards the final test set, the test data was initially compressed and left compressed until method development was completed."

Their Classifier: Gradient Boosting Machine

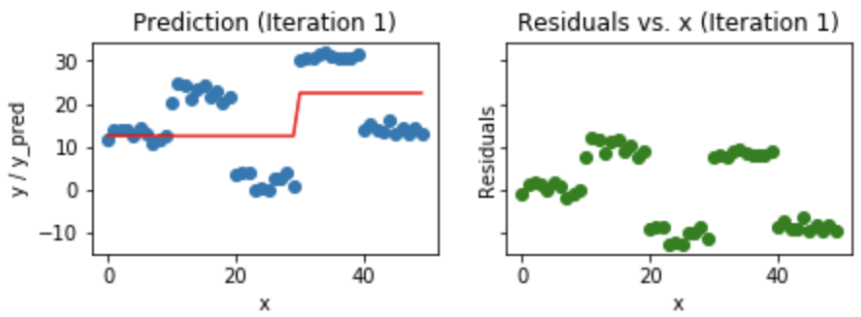
Step 1: Fit initial model $y = f_0(x) + \varepsilon_0$

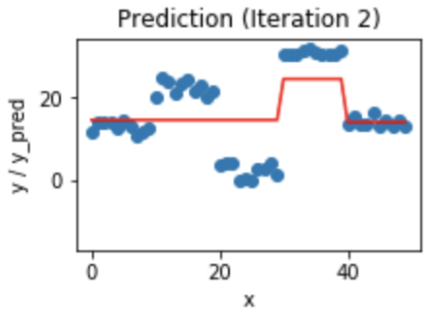
Step 2: Fit a second model to the residuals (or pseudo-residuals) $\varepsilon_0 = f_1(x) + \varepsilon_1$

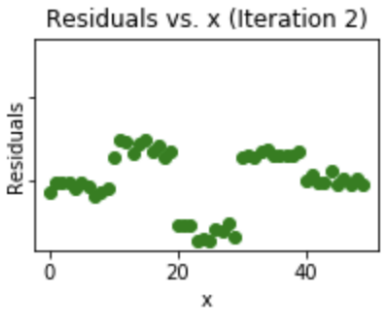
Step 3: Repeat (2)… $\hat{y} = \sum_{i=1}^{M} f_i(x)$

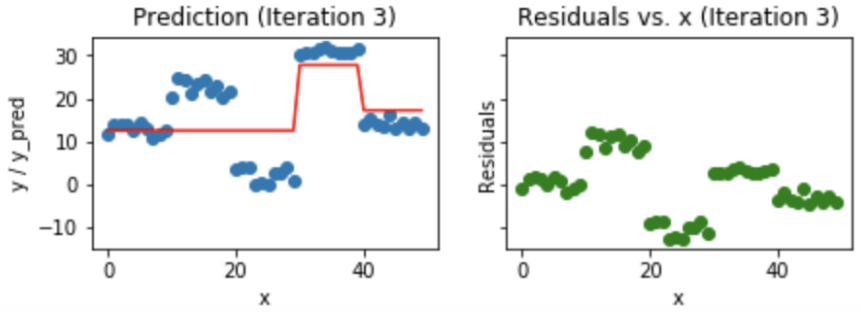
GBM for least squares regression

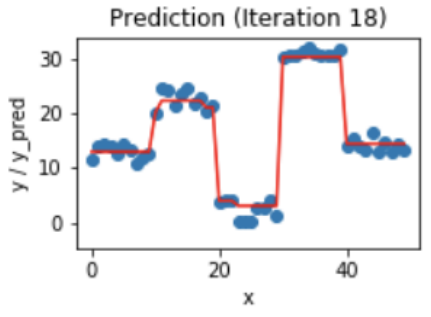
Their Model: Gradient Boosting Machine

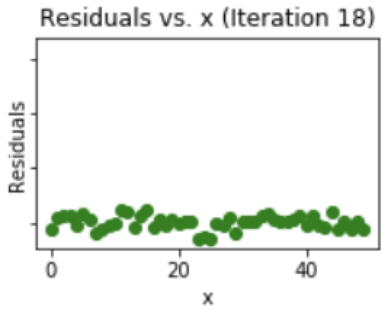


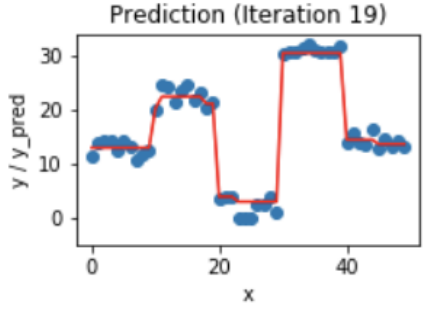


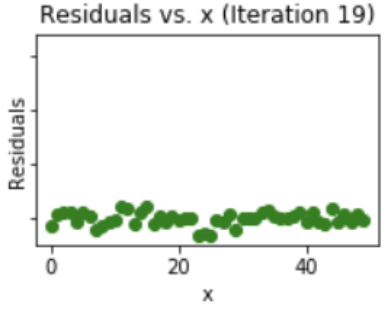


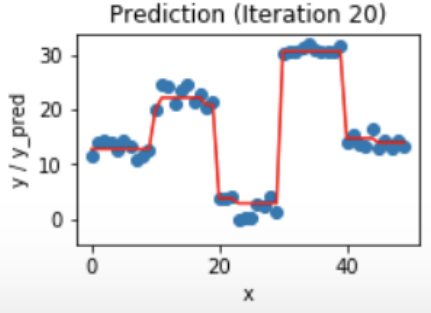


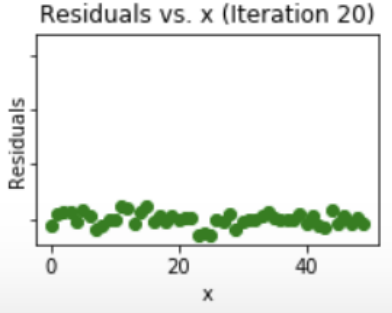




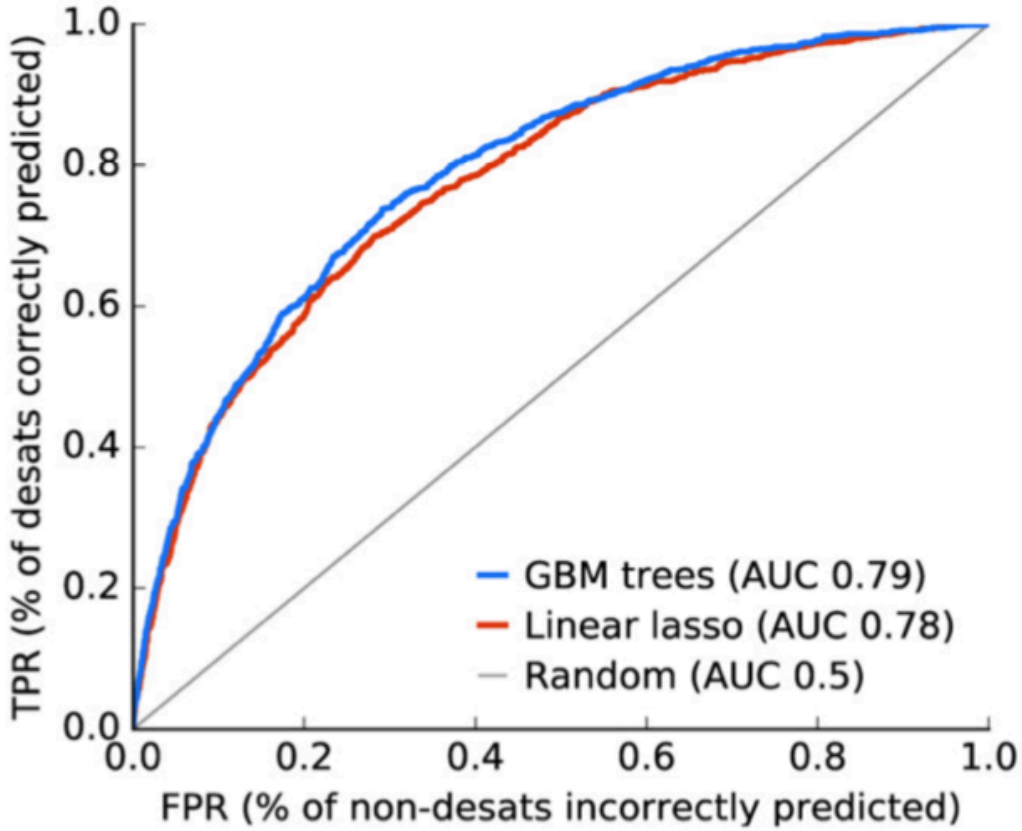




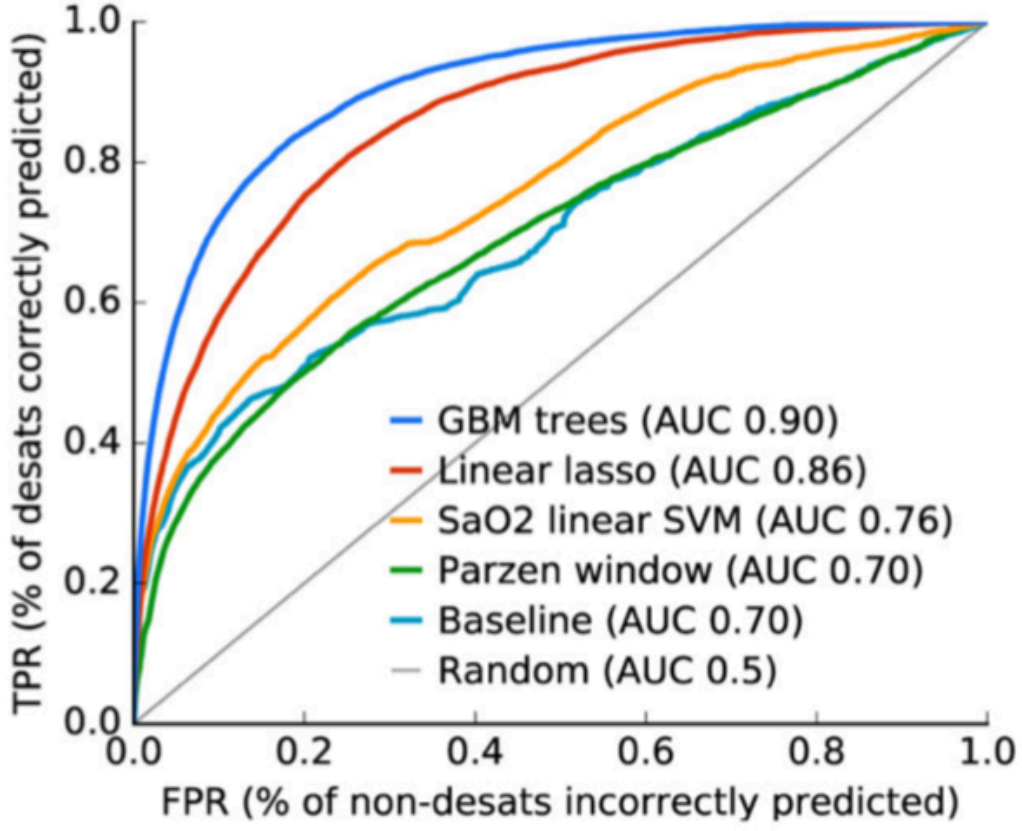




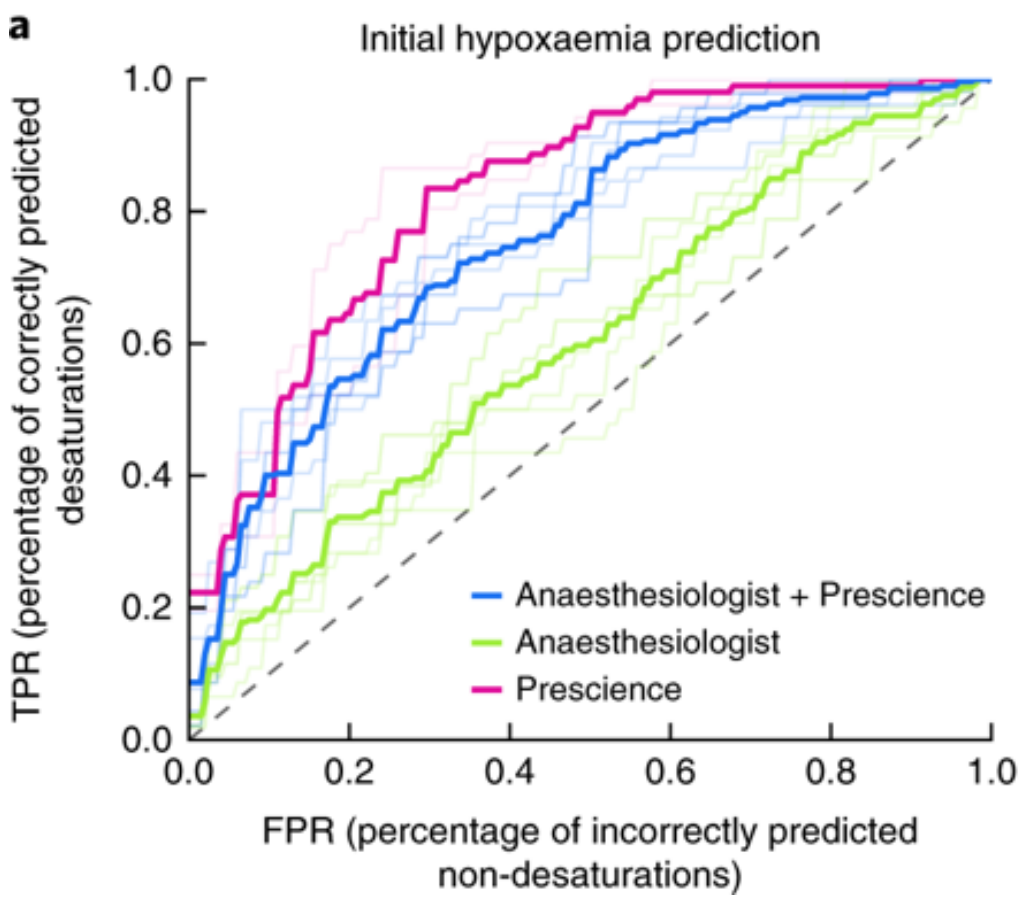
Comparison to Alternatives

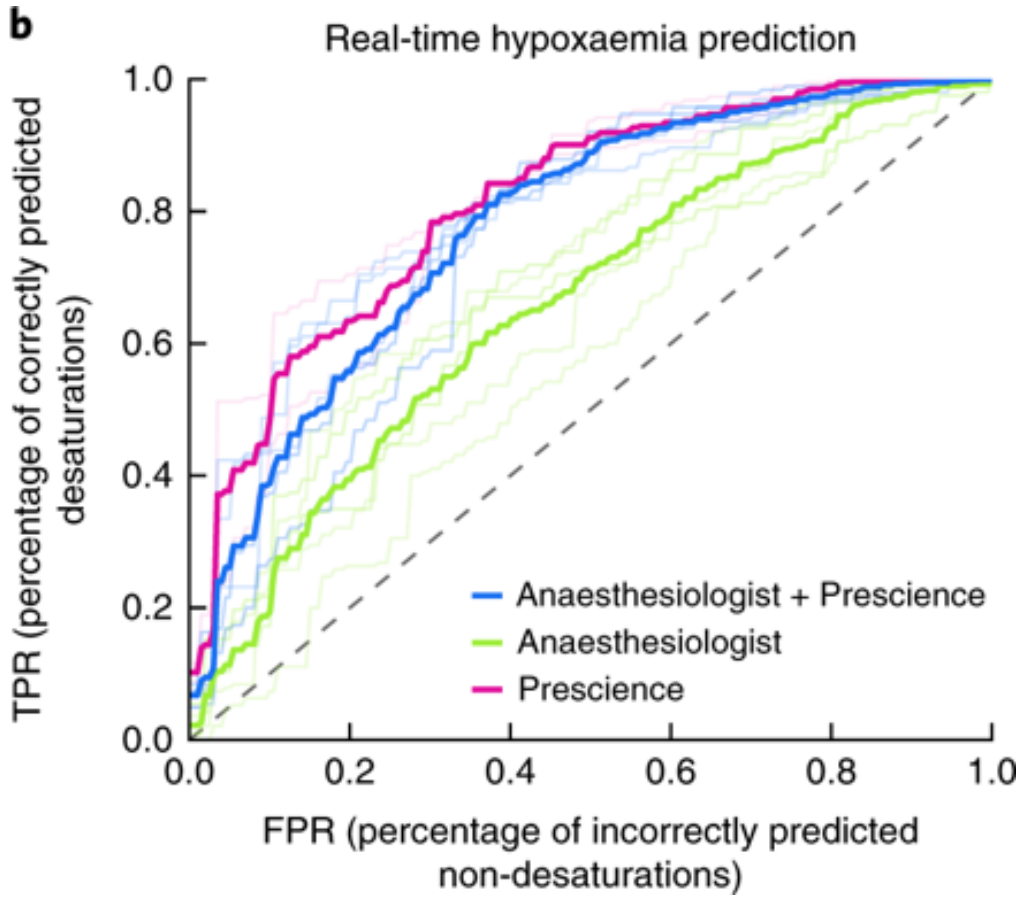


Initial Prediction



Real-time Prediction





Comparison to Experts

For initial risk prediction:

- Anaesthesiologists performed significantly better with Prescience (AUC = 0.76 versus 0.60; $P < 0.0001$)
- Prescience performed better in a direct comparison with anaesthesiologists (AUC = 0.83; $P < 0.0001$)

For intraoperative real-time (next 5 min) risk prediction:

- Anaesthesiologists (AUC = 0.66) again performed better with Prescience (AUC = 0.78; $P < 0.0001$)
- Prescience alone outperformed anaesthesiologists predictions (AUC = 0.81; $P < 0.0001$)

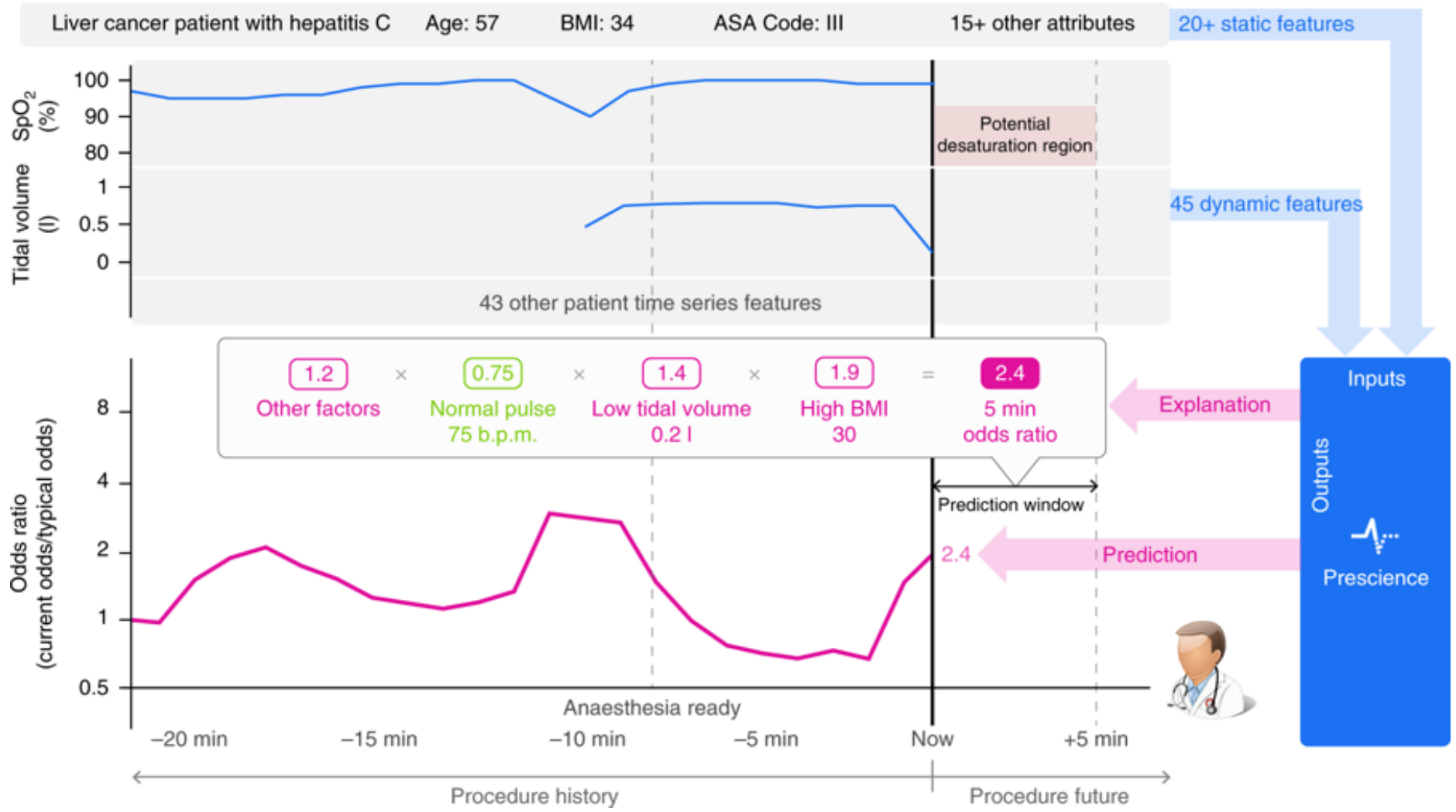
Is this a fair comparison?

- Training examples are episodes of hypoxemia that were not prevented during surgery

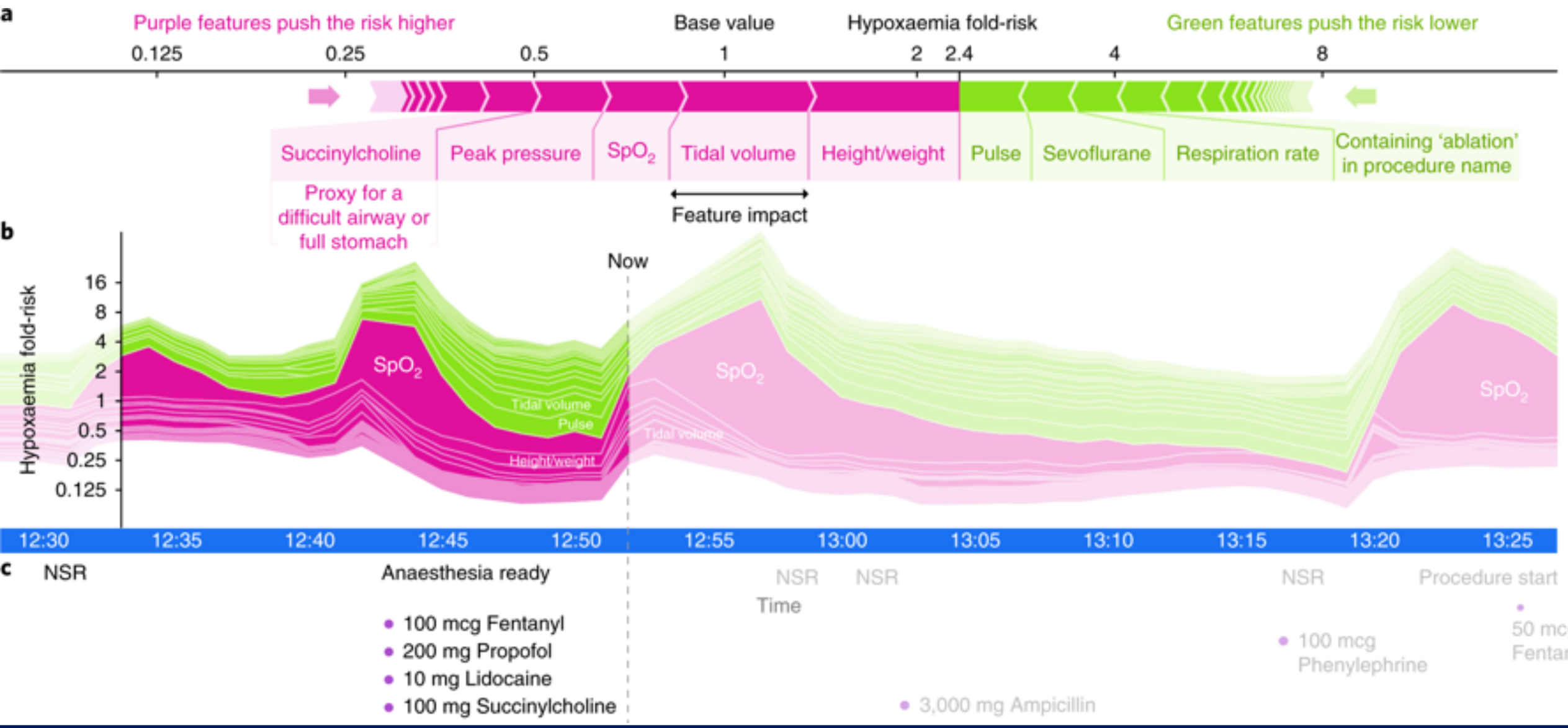
- Expert comparison:
  - the expert is predicting likelihood hypoxemic episodes, some of which were prevented
  - the model has learned to predict hypoxemic episodes that couldn't be avoided

INTERPRETABLE PREDICTIONS

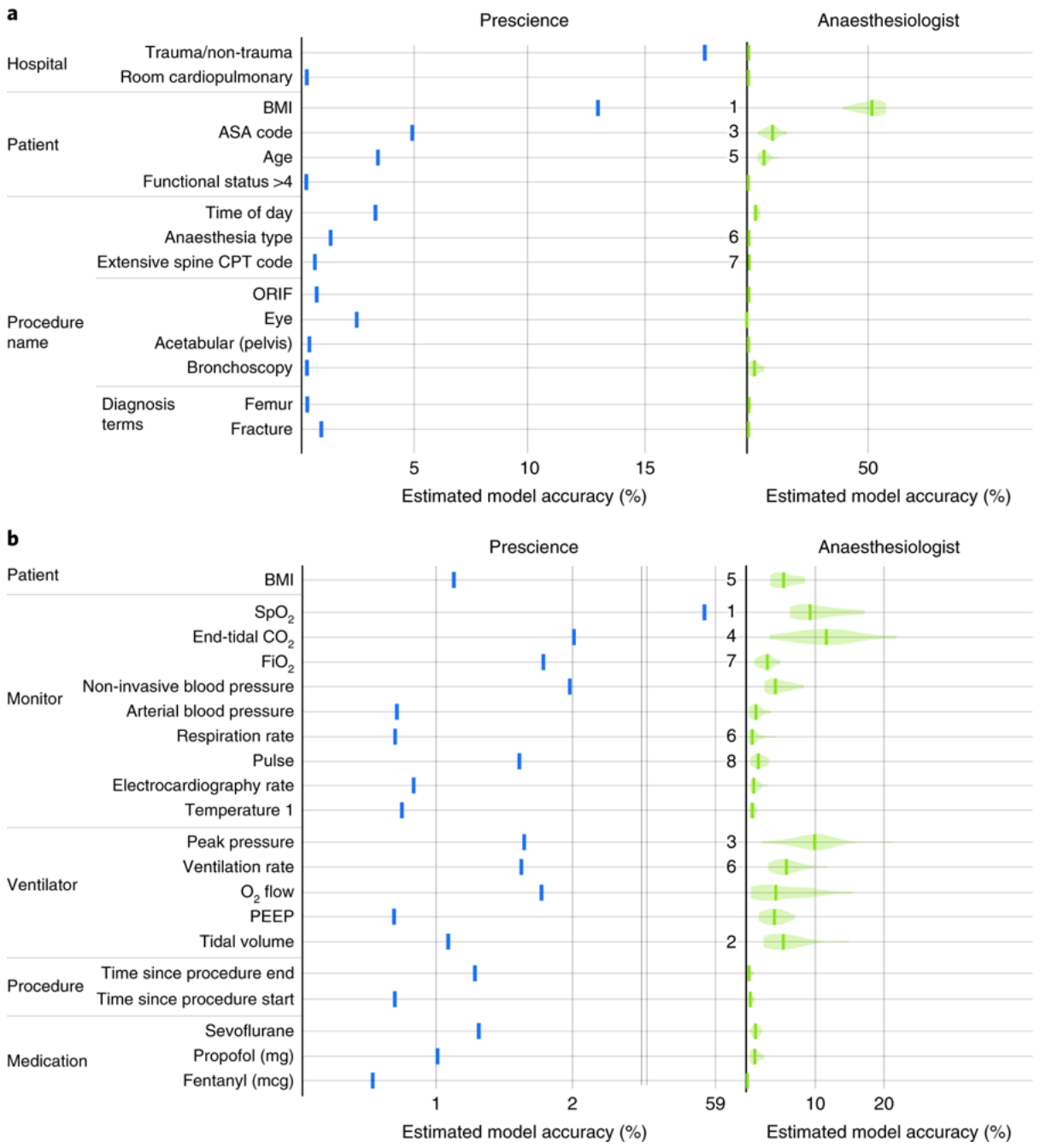
Highlight Feature Importance

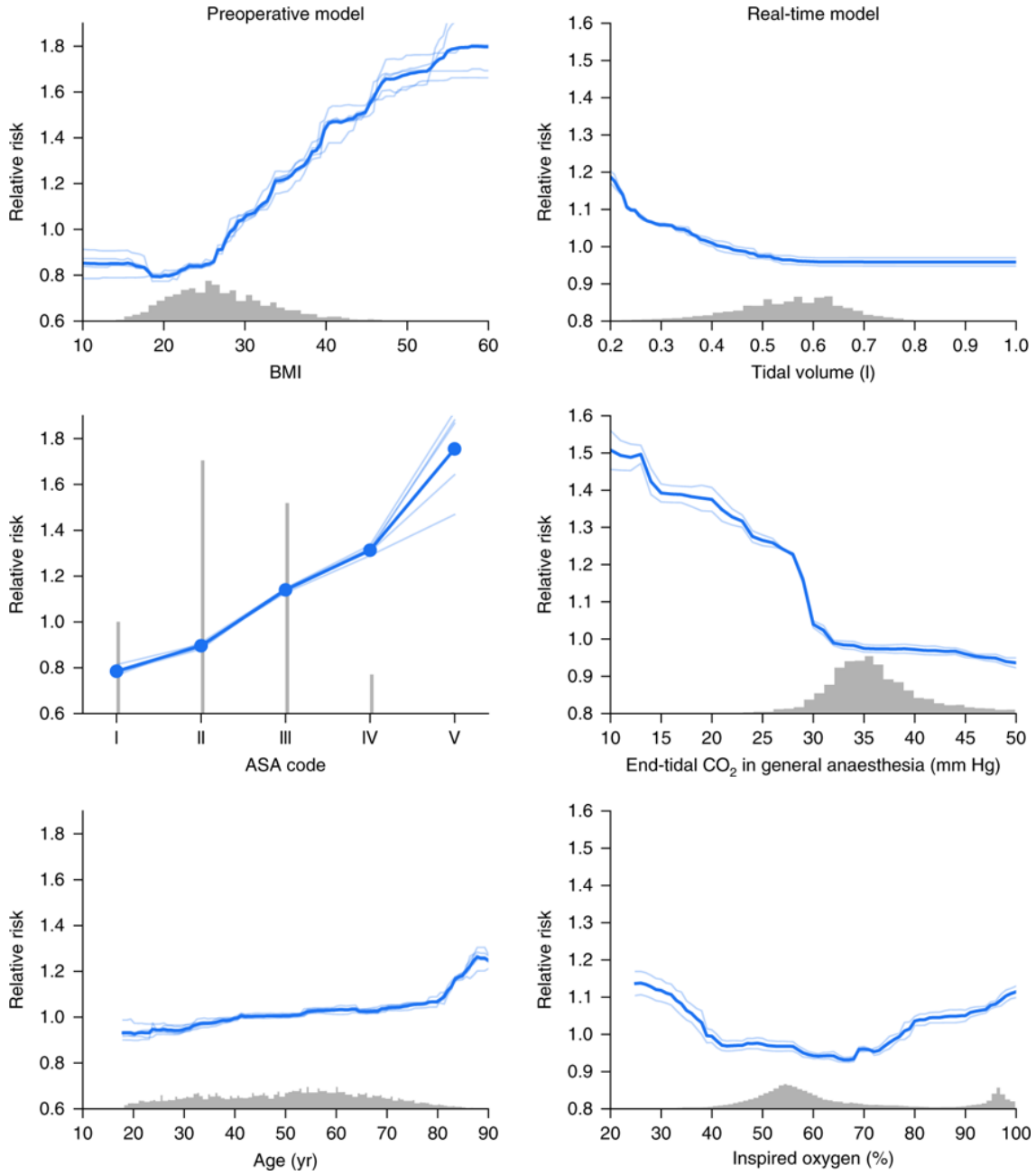


Cumulative Effect of Features



Model Predictions vs Physician Predictions



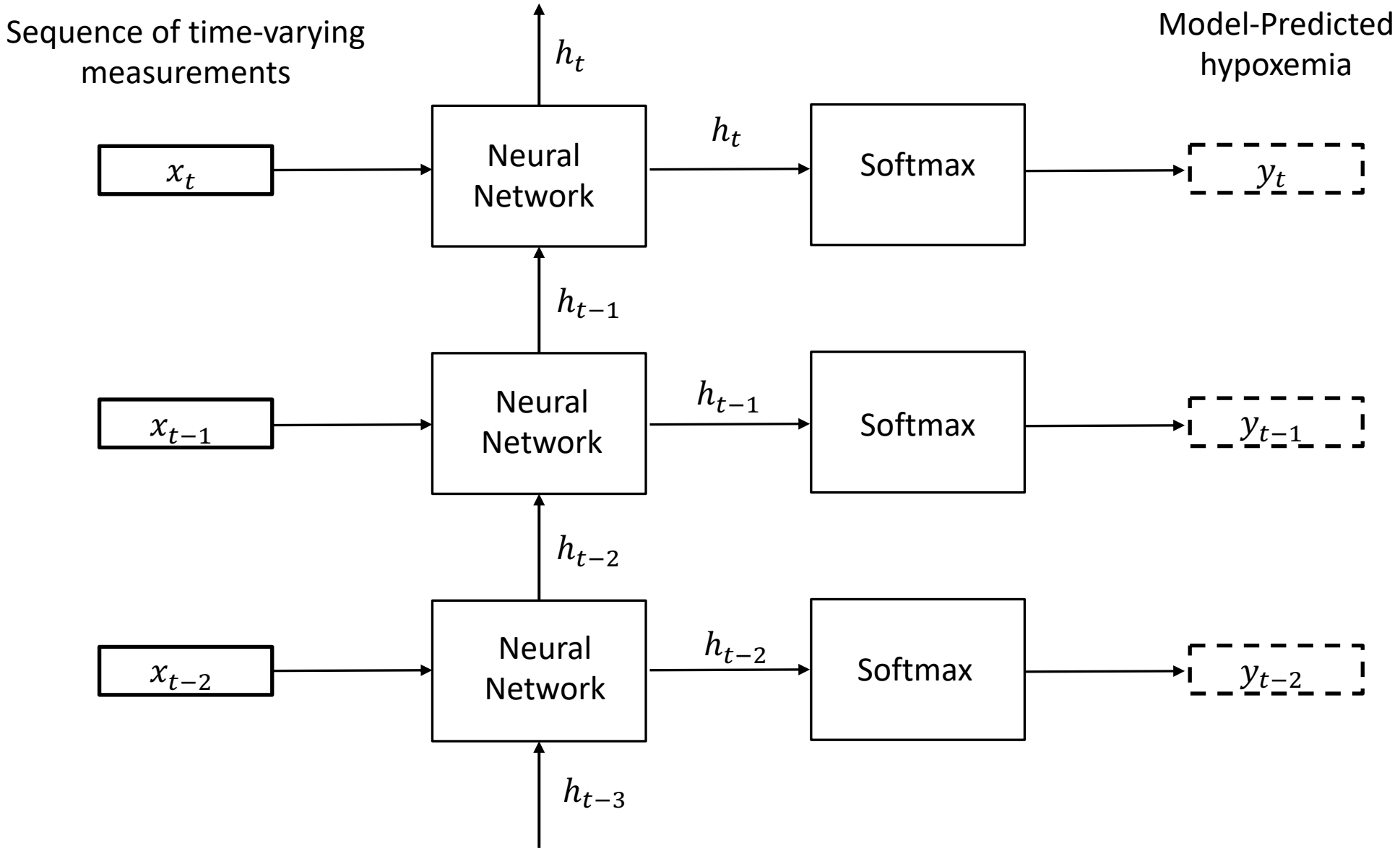


Sensitivity Analysis

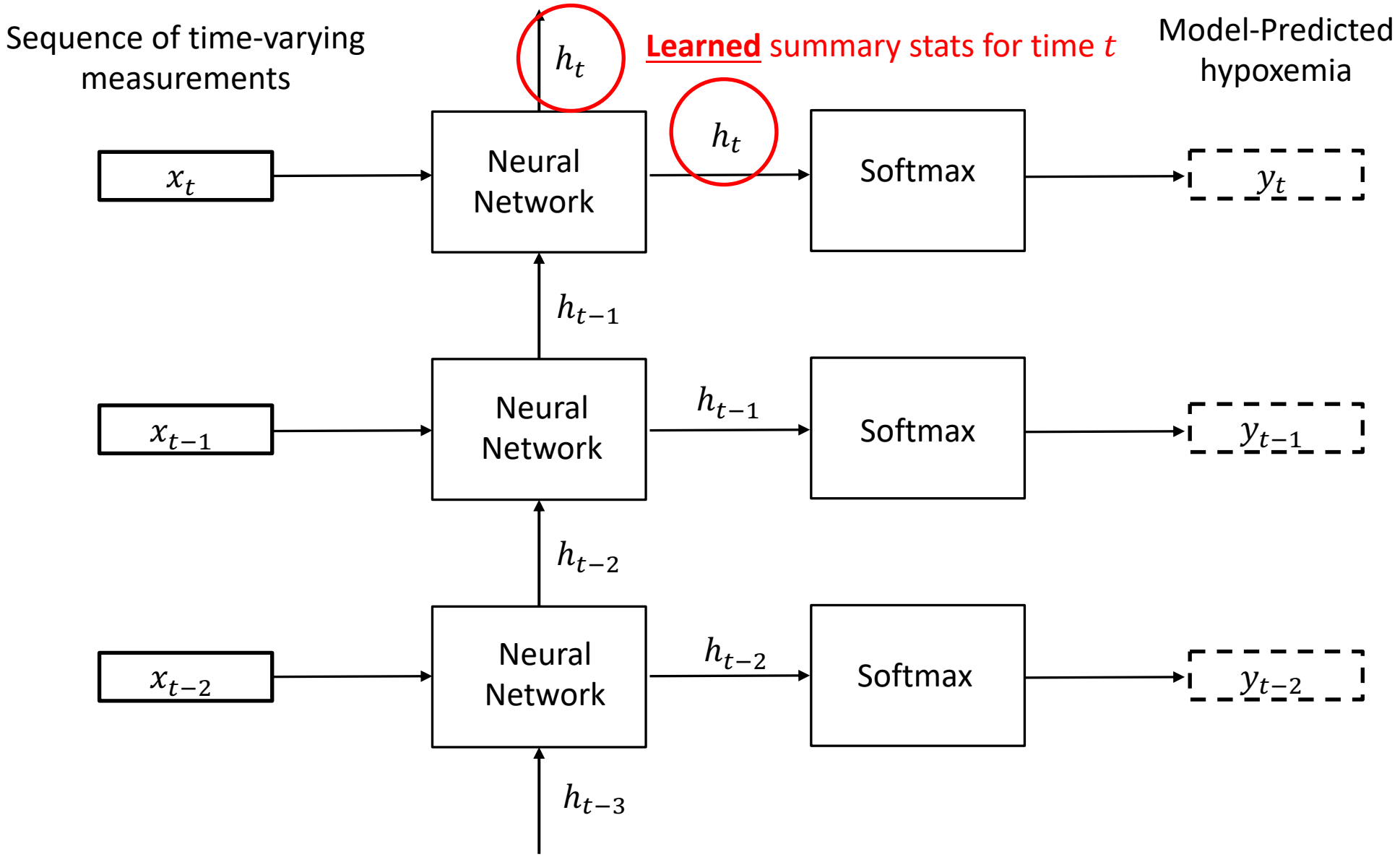
These partial dependence plots show the change in hypoxemia risk for all values of a given feature. The grey histograms on each plot show the distribution of values for that feature in the validation dataset. The lighter coloured lines represent model variability from bootstrap resampling of the training data, the dark lines represent the average of the bootstrap runs.

ALTERNATIVE APPROACH: RNN

Recurrent Neural Network



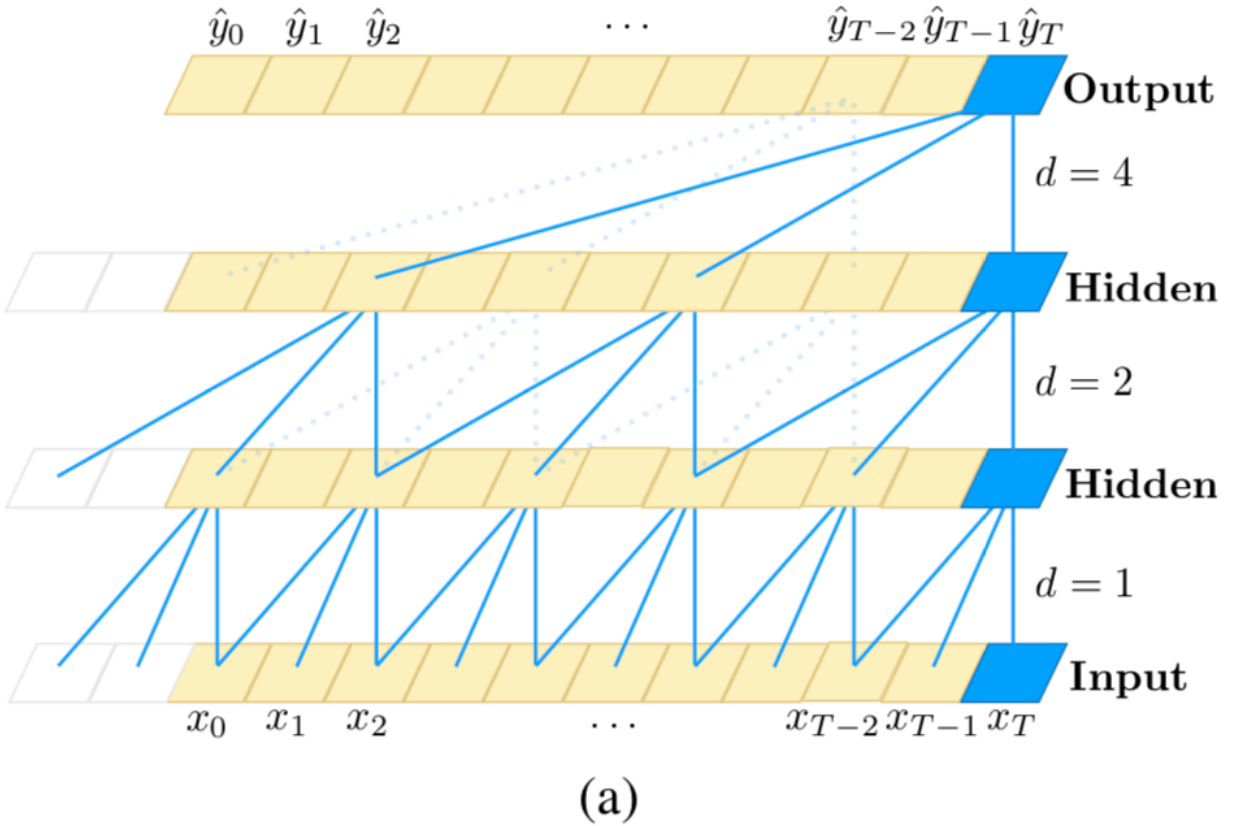
Recurrent Neural Network



RNN or CNN?

- Key advantage of RNN: able to handle variable-length sequences
- Key disadvantage of RNN: difficulty with long-term dependencies
- Recent work: CNNs that can handle variable-length sequences

Convolutional Architecture for Time Series (uses “dilated” convolutions)



(a)

A dilated causal convolution with dilation factors $d = 1, 2, 4$ and filter size $k = 3$.
The receptive field is able to cover all values from the input sequence.

CNNs for Sequences: (Bai et al, 2018)

An Empirical Evaluation of Generic Convolutional and Recurrent Networks for Sequence Modeling

Table 1. Evaluation of TCNs and recurrent architectures on synthetic stress tests, polyphonic music modeling, character-level language modeling, and word-level language modeling. The generic TCN architecture outperforms canonical recurrent networks across a comprehensive suite of tasks and datasets. Current state-of-the-art results are listed in the supplement. ^{h} means that higher is better. $^{\ell}$ means that lower is better.

Sequence Modeling Task	Model Size ($\approx$)	Models			
		LSTM	GRU	RNN	**TCN**
Seq. MNIST (accuracy^{h})	70K	87.2	96.2	21.5	**99.0**
Permuted MNIST (accuracy)	70K	85.7	87.3	25.3	**97.2**
Adding problem T=600 (loss$^{\ell}$)	70K	0.164	**5.3e-5**	0.177	**5.8e-5**
Copy memory T=1000 (loss)	16K	0.0204	0.0197	0.0202	**3.5e-5**
Music JSB Chorales (loss)	300K	8.45	8.43	8.91	**8.10**
Music Nottingham (loss)	1M	3.29	3.46	4.05	**3.07**
Word-level PTB (perplexity$^{\ell}$)	13M	**78.93**	92.48	114.50	88.68
Word-level Wiki-103 (perplexity)	-	48.4	-	-	**45.19**
Word-level LAMBADA (perplexity)	-	4186	-	14725	**1279**
Char-level PTB (bpc$^{\ell}$)	3M	1.36	1.37	1.48	**1.31**

Stacked RNN & Transformer

- Hierarchical architectures
- The latter is based on “self-attention”: determine which words modify a given word, and how