Working with Multi-Modal Health Data

Matthew Engelhard

Working with categorical variables

One-hot / Cell coding
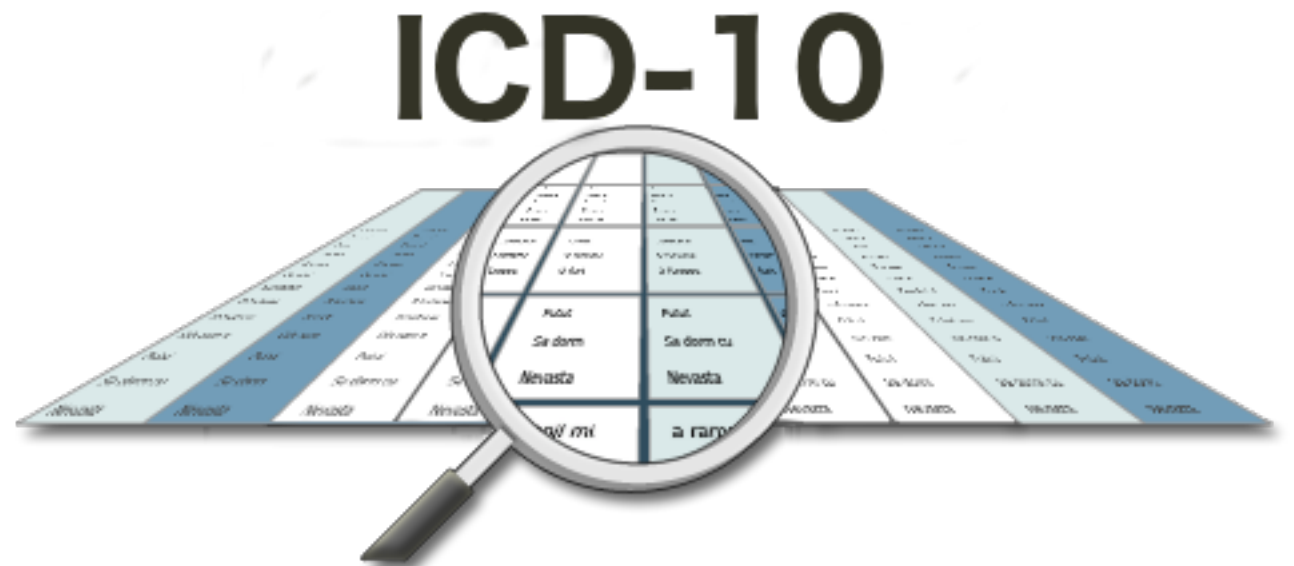
Embeddings

Cell / "One-hot" encoding

- Familiar from linear models: e.g. 0 for male, 1 for female
- Requires # features equal to # categories

- <u>Always categorical</u>: diagnosis codes, procedure codes
- <u>Reduce to categorical</u>: look at which labs are collected, ignore values

- Consider similarity to bag of words models…

Very large number of categories

- Option 1: don't worry about it
- Option 2: group manually or using existing software (e.g. clinical classification software)
- Option 3: use or create embeddings
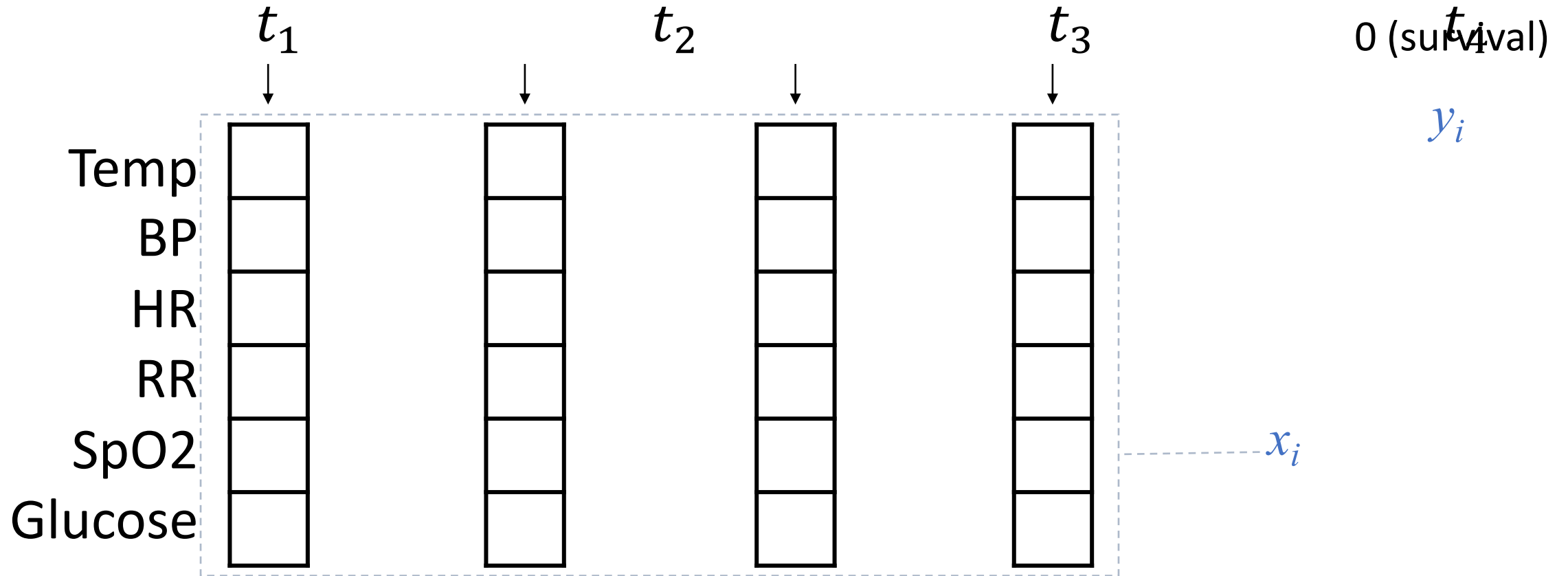 - pre-train
  - learn directly

Measurements over time

- Option 1: Count number of occurrences (+ log transform)
- Option 2: If we have embeddings, we can aggregate (like VSWEM)
- Option 3: Use a sequential model (covered in the next lecture)

- How do we incorporate information about the TIMESTAMP??
  - Positional embedding
  - Temporal embedding
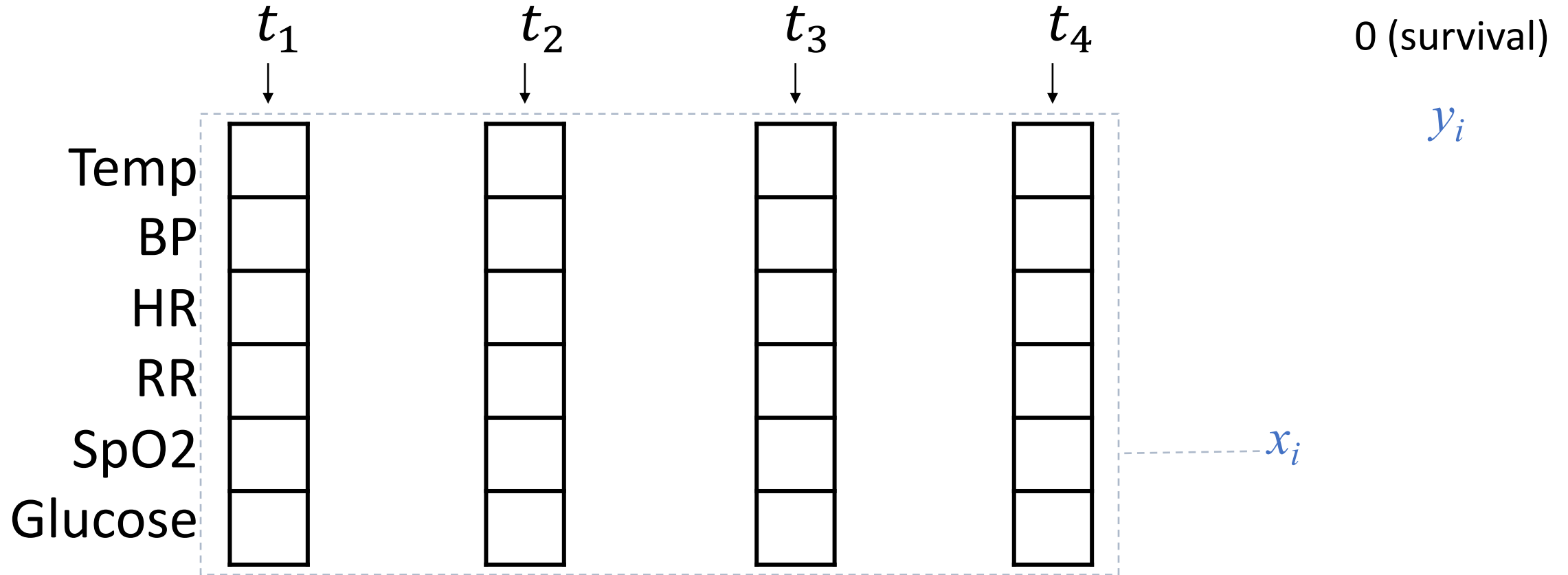
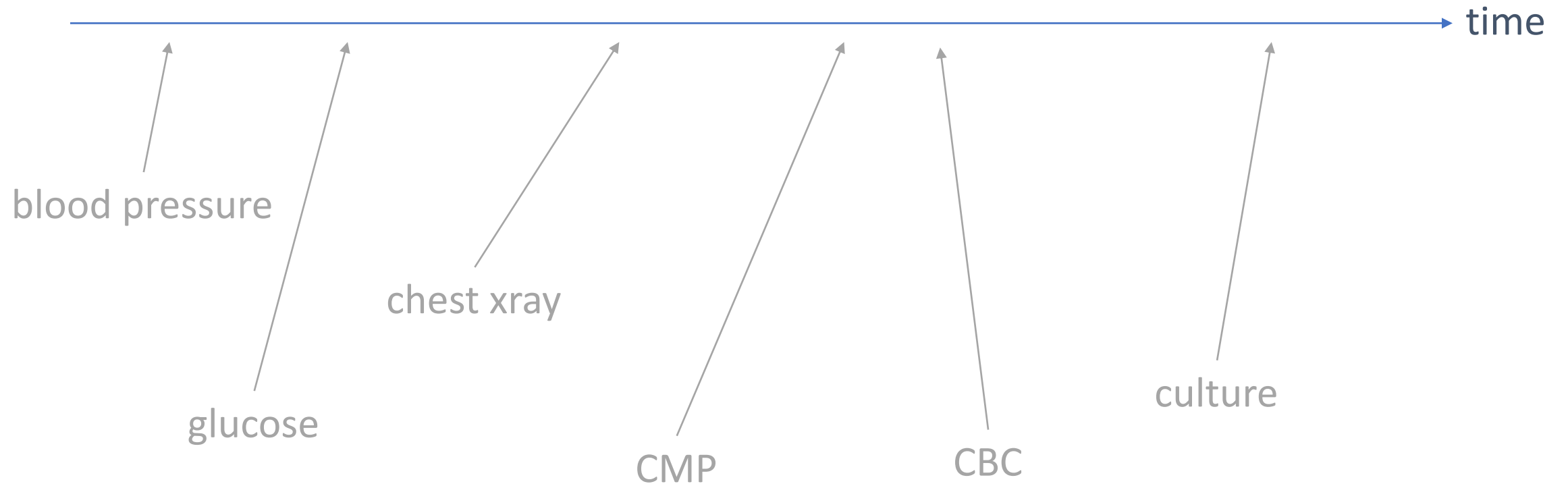# Working with irregularly spaced measurements
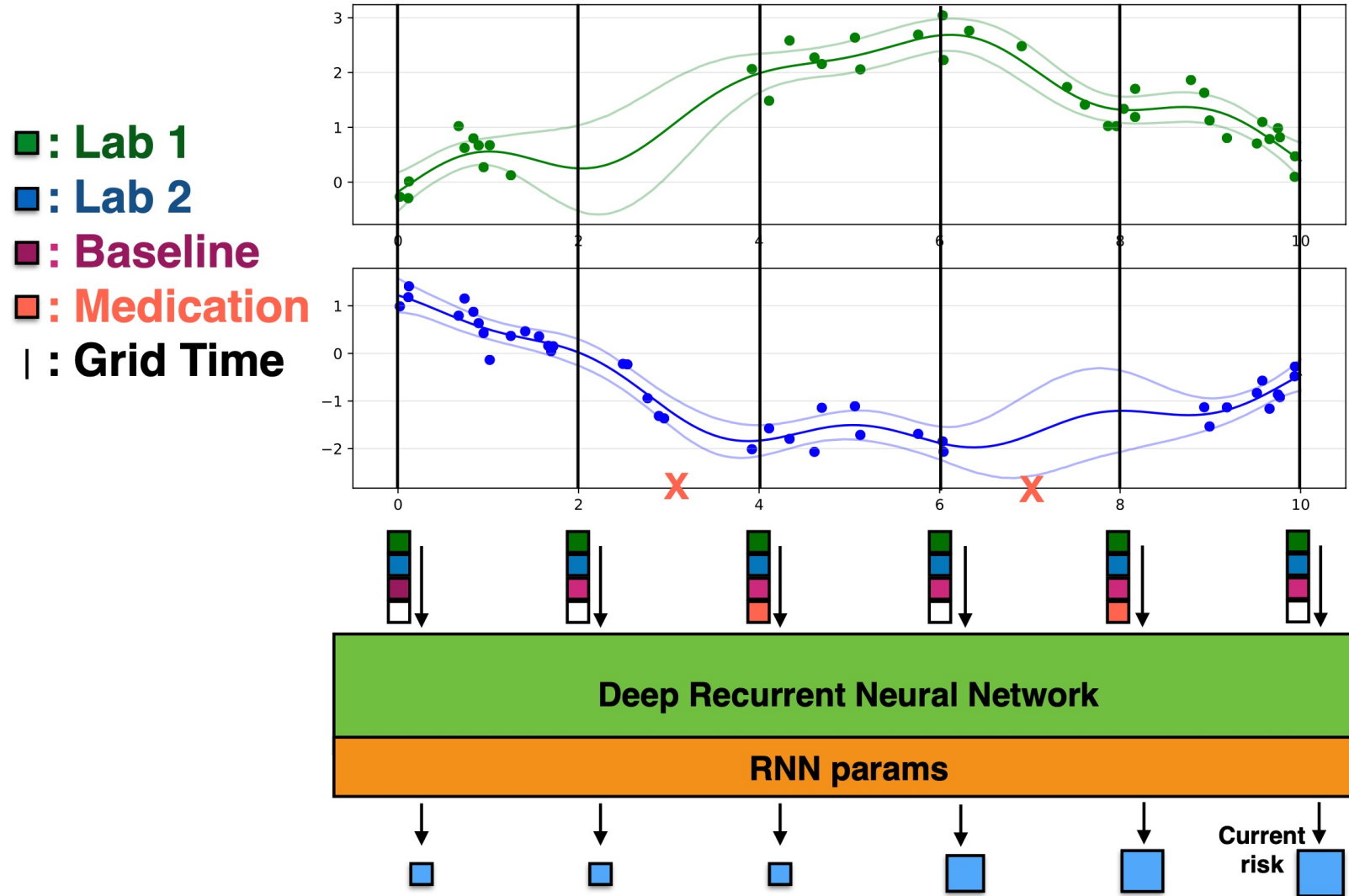
Aggregation

Imputation

Measurements in the ICU

t_1 t_2 t_3 t_4 0 (survival)

y_i

Temp
BP
HR
RR
SpO2
Glucose

x_i

Measurements in the ICU

Measurements on the Wards…

time

blood pressure

glucose

chest xray

CMP

CBC

culture

This is a major difficulty!

DIHI Sepsis Watch

: Lab 1
: Lab 2
: Baseline
: Medication
| : Grid Time

<- Use GP regression to predict measurements at regular intervals

Deep Recurrent Neural Network

RNN params

<- Predict sepsis risk using an RNN

Current risk

Simplest Method…

time

blood pressure

glucose

chest xray

CMP

CBC

culture
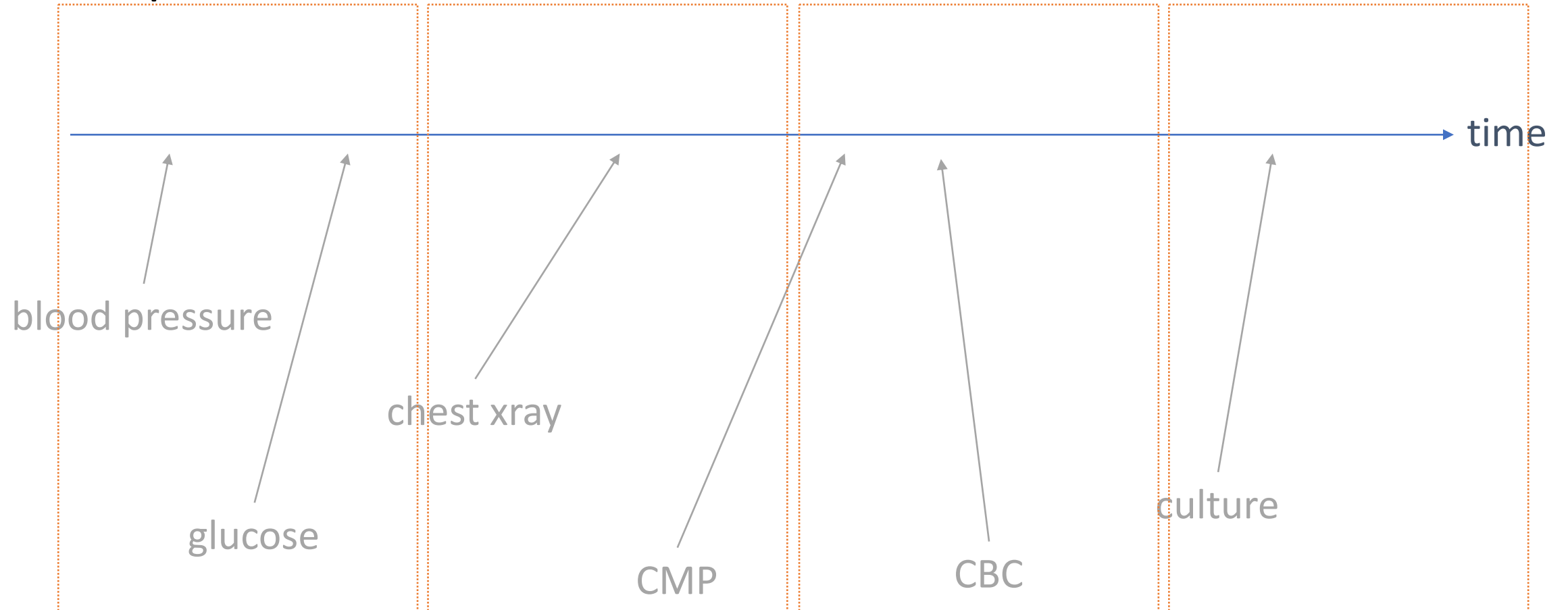
aggregate in hour 1   aggregate in hour 2   aggregate in hour 2   aggregate in hour 4…

In the EHR, measurements are highly "sparse"

- Many more missing measurements than non-missing

- Consider diagnosis codes, procedure codes, uncommon labs, etc

- We want to learn from these measurements, but most patients don't have them

- Choose an aggregation window to reduce this, but nevertheless a great deal of imputation is required

Making predictions from multi-modal data
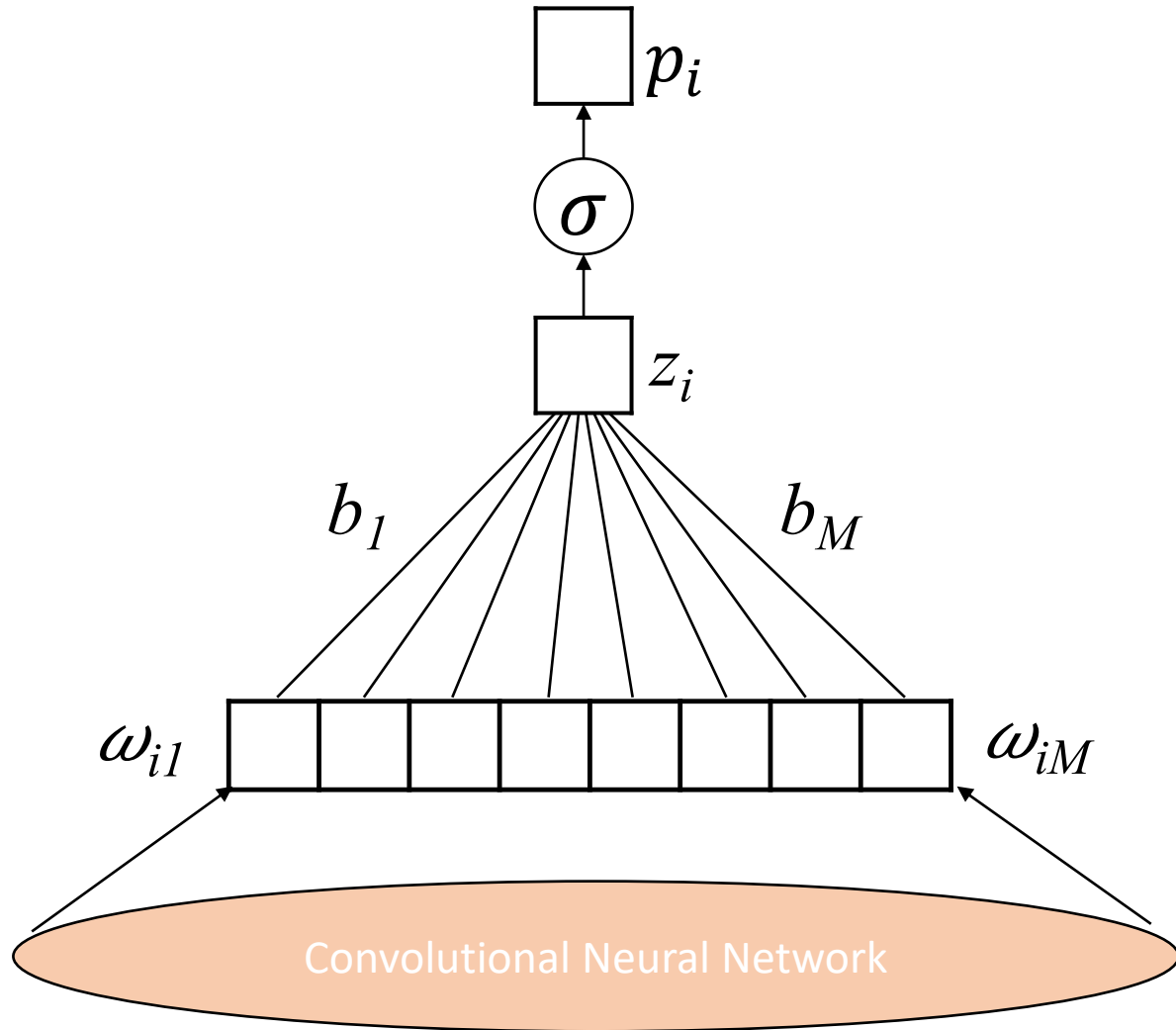
"Data fusion"

Logistic Regression
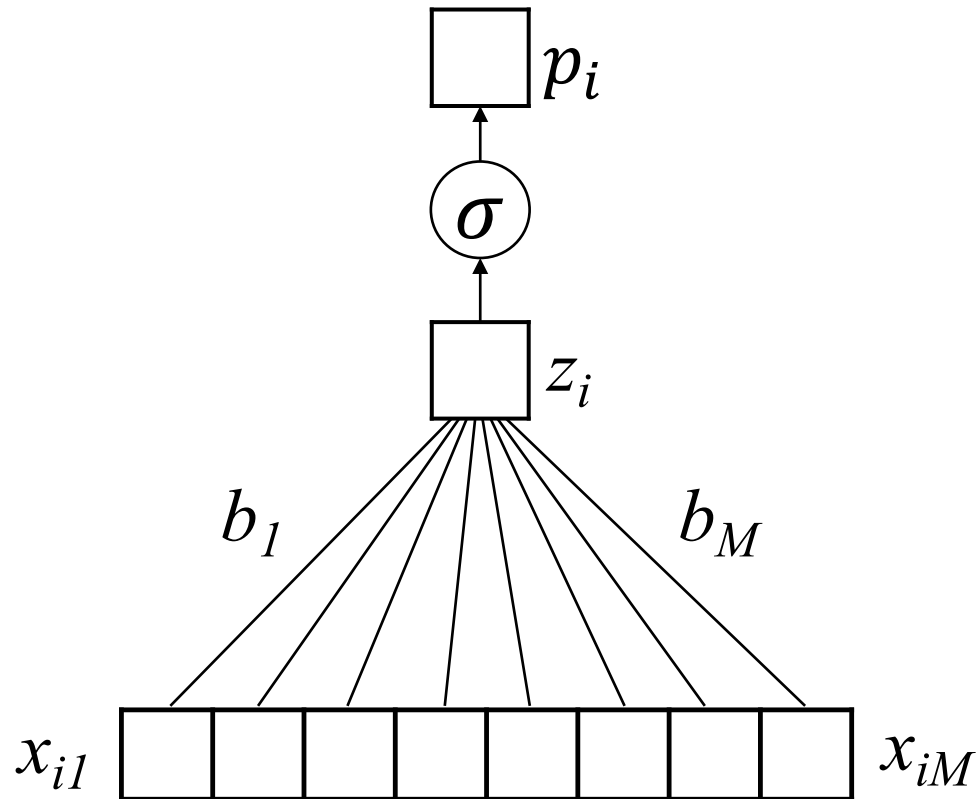
x_i is a vector of predictor variables

We predict logit(p_i)
(the log-odds)
Directly from x_i

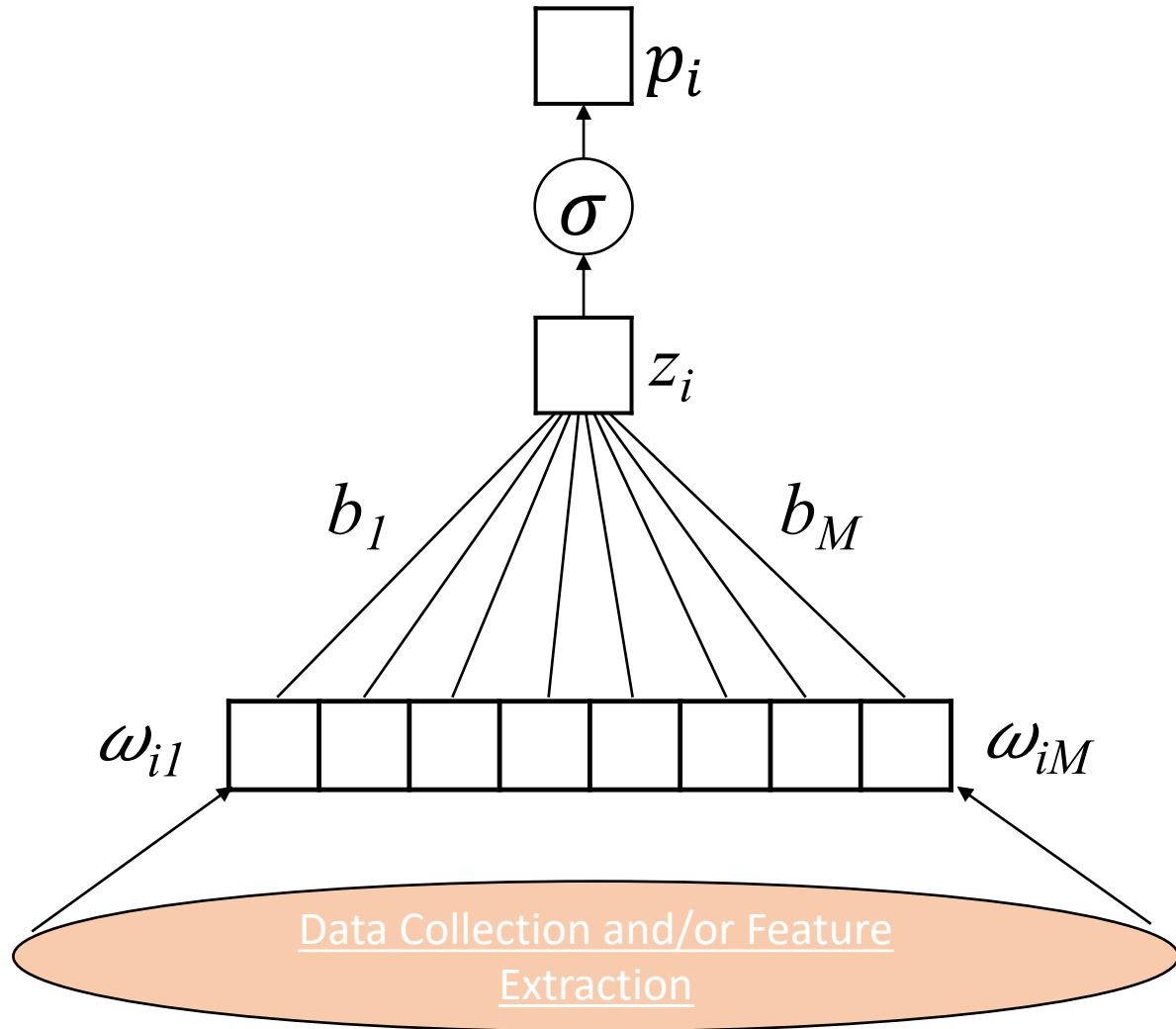No intermediate or hierarchical features

Convolutional Neural Network

ω_i is a vector of high-level features extracted by the CNN

We predict logit(p_i) (the log-odds) based on ω_i

No intermediate or hierarchical features

Any Modality

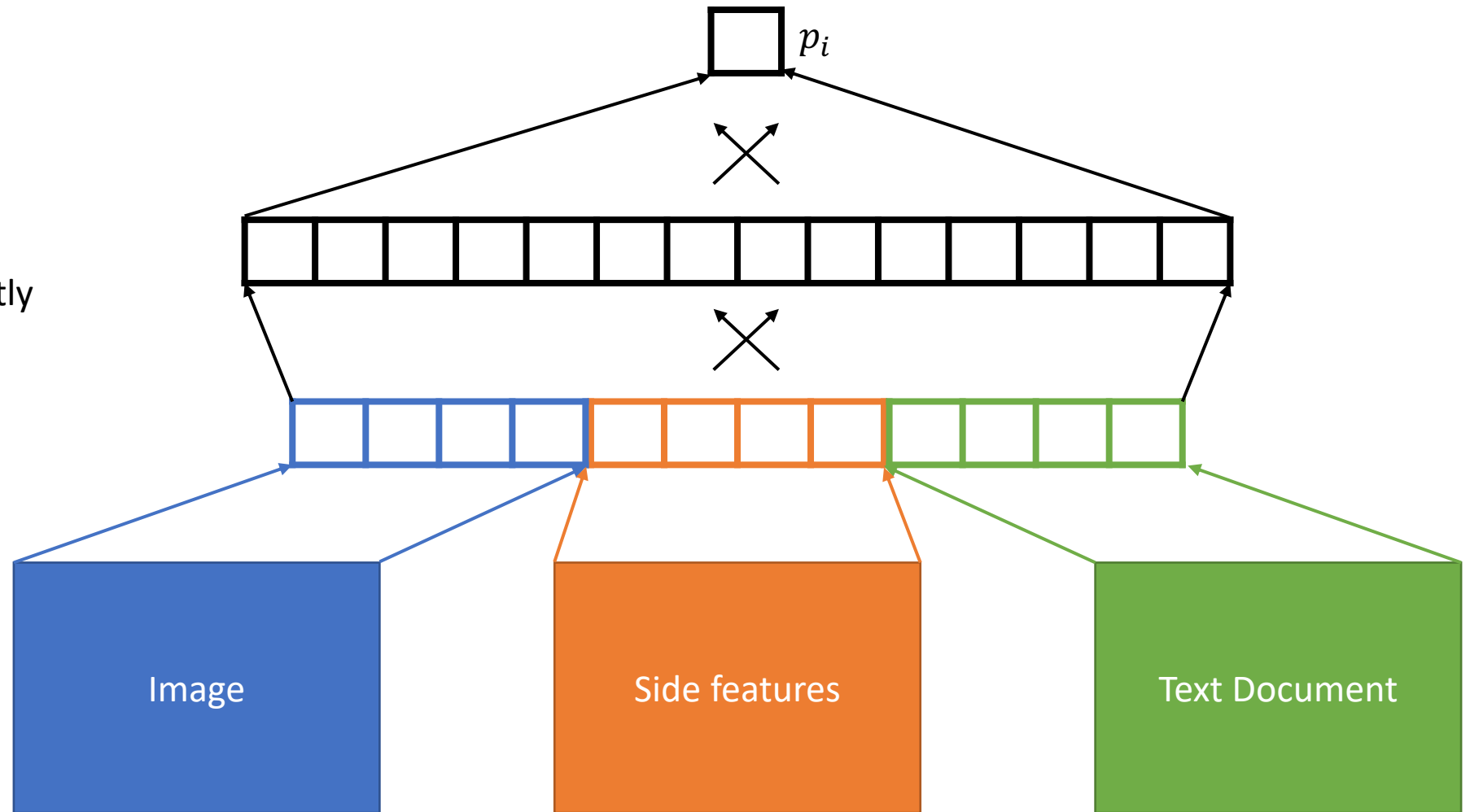We predict logit(p_i)
(the log-odds)
based on ω_i

Which may be observed predictors, or
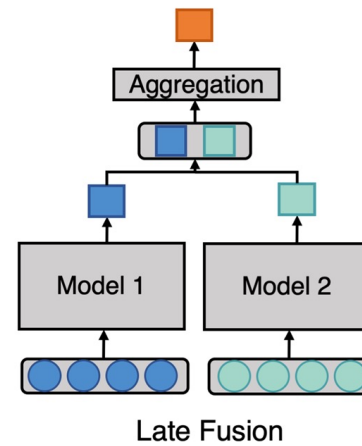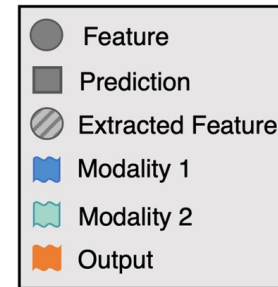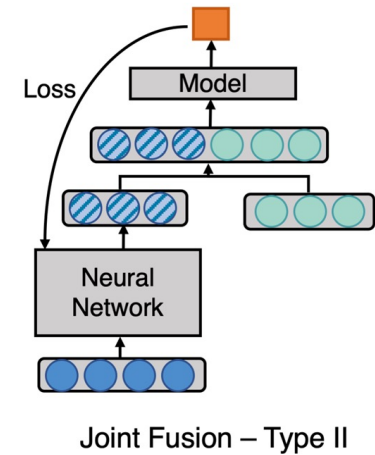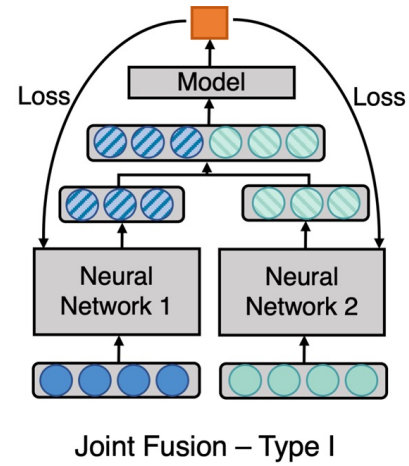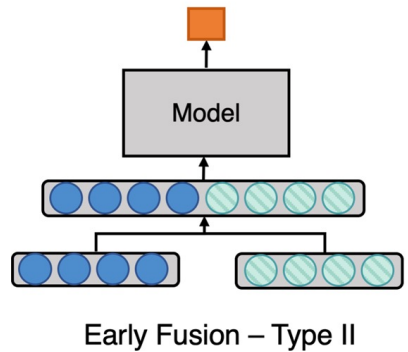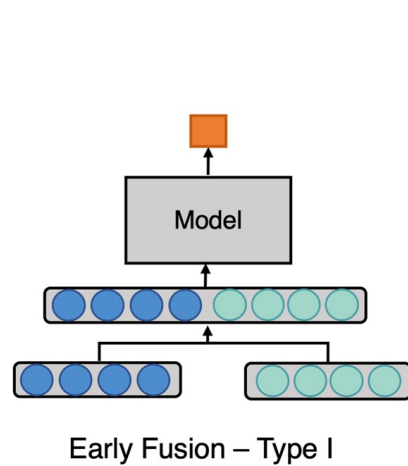it may be extracted by a deep NN

Combining Modalities

1. Train separate models for each modality. This is more important for images, text, and other high-dimensional and/or structured data than for tabular data.

2. Create a single model by concatenating the highest-level features (used to predict logit(p_i)) across all modalities.

3. Consider adding one more fully-connected layer; this allows you to take advantage of interactions between features from different modalities.

4. Train the final layer(s)

5. Consider fine-tuning the entire model

Combining Modalities

1. Train independently
2. Concatenate
3. Train together

Early Fusion – Type I

Joint Fusion – Type I

Early Fusion – Type II

Joint Fusion – Type II

Late Fusion

Feature

Prediction

Extracted Feature

Modality 1

Modality 2

Output

Huang, SC., Pareek, A., Seyyedi, S. *et al.* Fusion of medical imaging and electronic health records using deep learning: a systematic review and implementation guidelines. *npj Digit. Med.* **3,** 136 (2020).

Conclusions

- In healthcare, we commonly have the following challenges:
    1. Categorical variables with a large number of categories (i.e. high cardinality)
    2. Sparse measurements
    3. Multiple modalities we need to integrate

- Always start with the simplest method

- Often, aggregating measurements/features performs quite well

- Once our simple baseline is established, we can consider more complex models (e.g. RNNs for sequences)