From Logistic Regression to the Multilayer Perceptron

May 24, 2019

Lecture 1, Applied Data Science
MMCi Term 4, 2019

Matthew Engelhard

Overview of Applied Data Science Course

- We will focus on current data science methods and their applications
  - What are data science, machine learning, and artificial intelligence; and how do they differ from statistics?
  - How do these techniques work, and what kinds of problems can they solve?
  - How can we develop our own data science models or projects?
- Study algorithms that *learn* from data to make predictions or decisions

Neural networks are state-of-the-art for *many* applications

- They are **not new**—many of the techniques go back decades
- Recent resurgence due to *amazing* performance on benchmark tasks
- One key task was the ImageNet Challenge
  - Want to recognize what is in an image (1 of 1000 categories)
  - Have ~1 million example images
  - Very relevant for things such as image search
- Example images are shown on the right, with predicted categories beneath each image

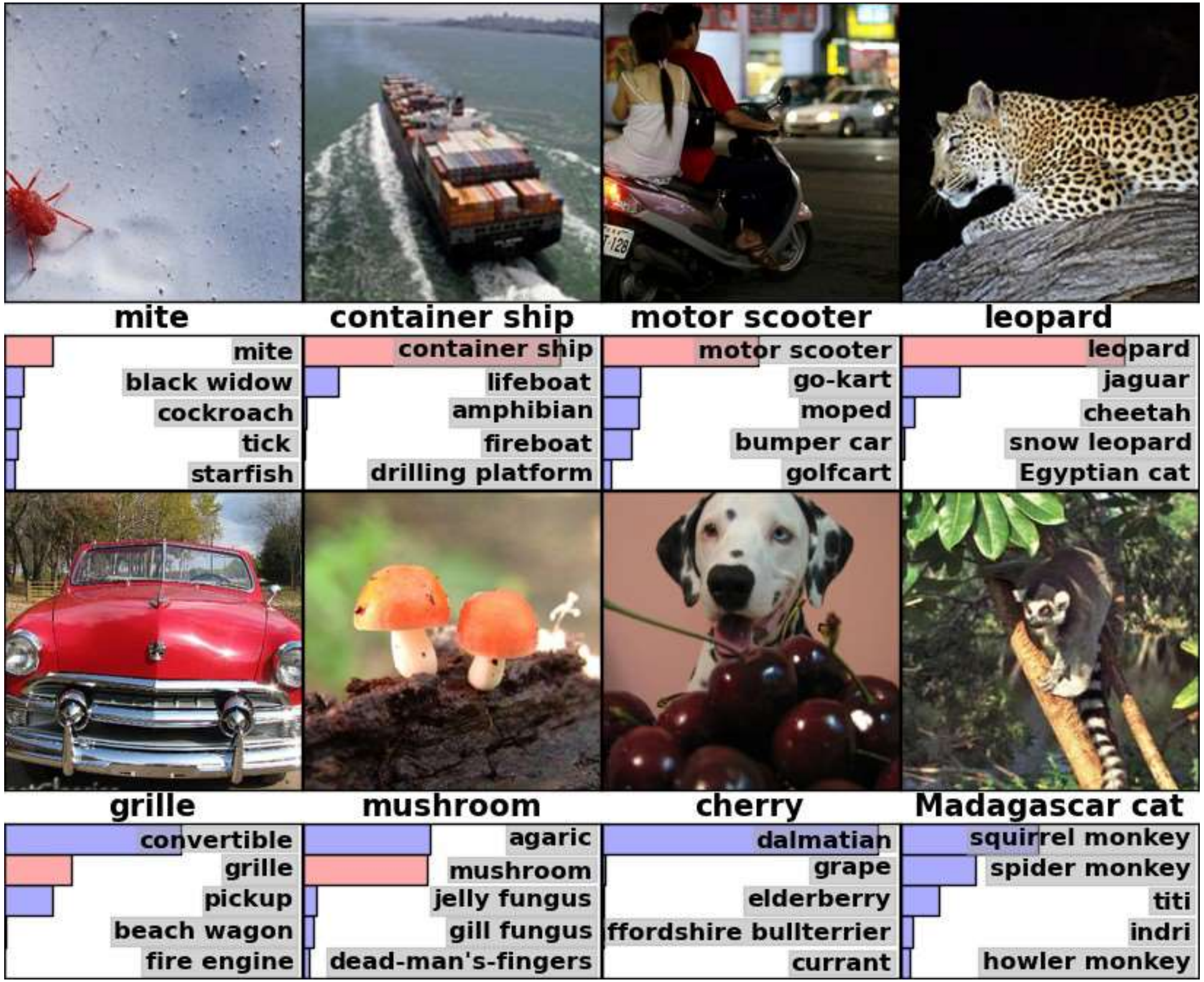
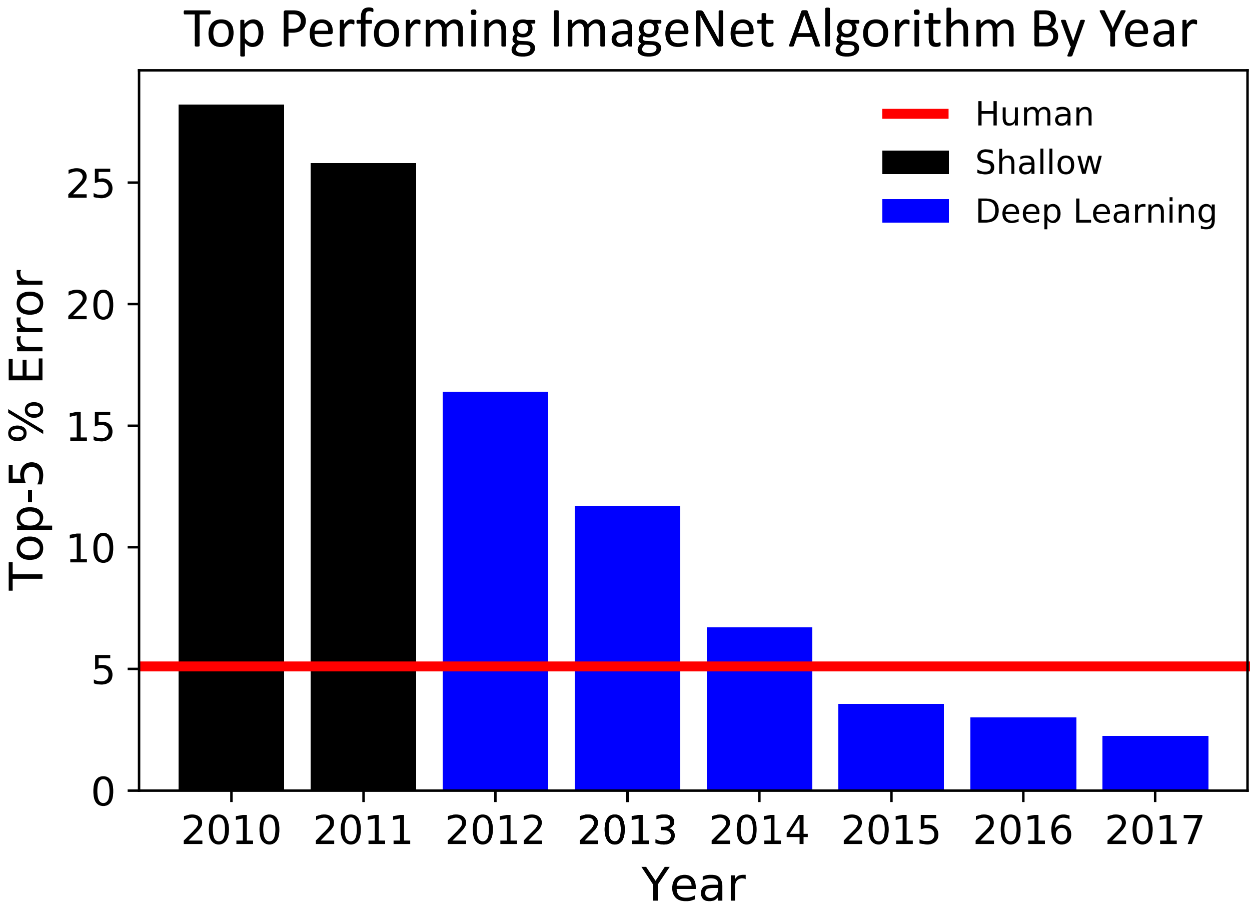


Figure from Krizhevsky et al 2012

Machine Learning can surpass human performance

- For ImageNet:
  - Deep Learning was a *huge* jump forward
  - State-of-the-art systems **significantly outperform humans** on the same task

- These use “Convolutional Neural Networks,” which you will learn about in block 2

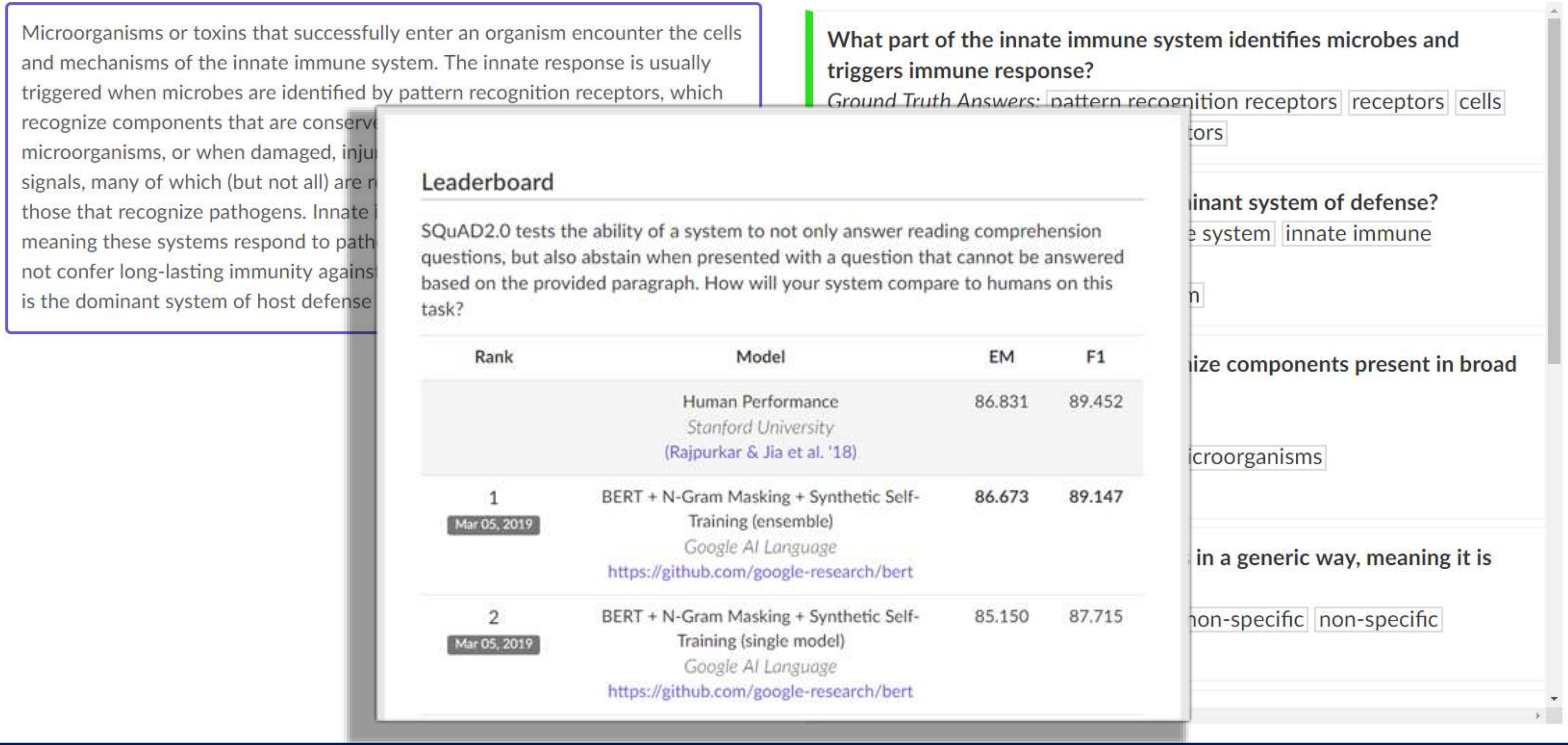


Machine Learning beats human performance in many tasks

- Famously, Google DeepMind trained “AlphaGo” to beat the world champion Go player (a complex game)
- AlphaGo uses Deep Reinforcement Learning (learned by repeatedly playing the game), to be covered in block 4
- Many other examples:
  - Voice Recognition
  - Object Detection
  - Text Translation
  - *Etc.*

nature
THE INTERNATIONAL WEEKLY JOURNAL OF SCIENCE
At last – a computer program that can beat a champion Go player PAGE 484
ALL SYSTEMS GO

Deep Learning is Approaching Human Performance in Language Understanding Tasks



Microorganisms or toxins that successfully enter an organism encounter the cells and mechanisms of the innate immune system. The innate response is usually triggered when microbes are identified by pattern recognition receptors, which
recognize components that are conserv
microorganisms, or when damaged, inju
signals, many of which (but not all) are r
those that recognize pathogens. Innate i
meaning these systems respond to path
not confer long-lasting immunity agains
is the dominant system of host defense

What part of the innate immune system identifies microbes and triggers immune response?

Ground Truth Answers: pattern recognition receptors | receptors | cells

tors

inant system of defense?

e system | innate immune

n

ize components present in broad

icroorganisms

in a generic way, meaning it is

non-specific | non-specific

Leaderboard

SQuAD2.0 tests the ability of a system to not only answer reading comprehension questions, but also abstain when presented with a question that cannot be answered based on the provided paragraph. How will your system compare to humans on this task?

Rank	Model	EM	F1
	Human Performance *Stanford University* (Rajpurkar & Jia et al. '18)	86.831	89.452
1 Mar 05, 2019	BERT + N-Gram Masking + Synthetic Self-Training (ensemble) *Google AI Language* https://github.com/google-research/bert	**86.673**	**89.147**
2 Mar 05, 2019	BERT + N-Gram Masking + Synthetic Self-Training (single model) *Google AI Language* https://github.com/google-research/bert	85.150	87.715

SYSTEM PROMPT (HUMAN-WRITTEN)

A train carriage containing controlled nuclear materials was stolen in Cincinnati today. Its whereabouts are unknown.

MODEL COMPLETION (MACHINE-WRITTEN, FIRST TRY)

The incident occurred on the downtown train line, which runs from Covington and Ashland stations.

In an email to Ohio news outlets, the U.S. Department of Energy said it is working with the Federal Railroad Administration to find the thief.

“The theft of this nuclear material will have significant negative consequences on public and environmental health, our workforce and the economy of our nation,” said Tom Hicks, the U.S. Energy Secretary, in a statement. “Our top priority is to secure the theft and ensure it doesn’t happen again.”

The stolen material was taken from the University of Cincinnati’s Research Triangle Park nuclear research site, according to a news release from Department officials. The Nuclear Regulatory Commission did not immediately release any information.

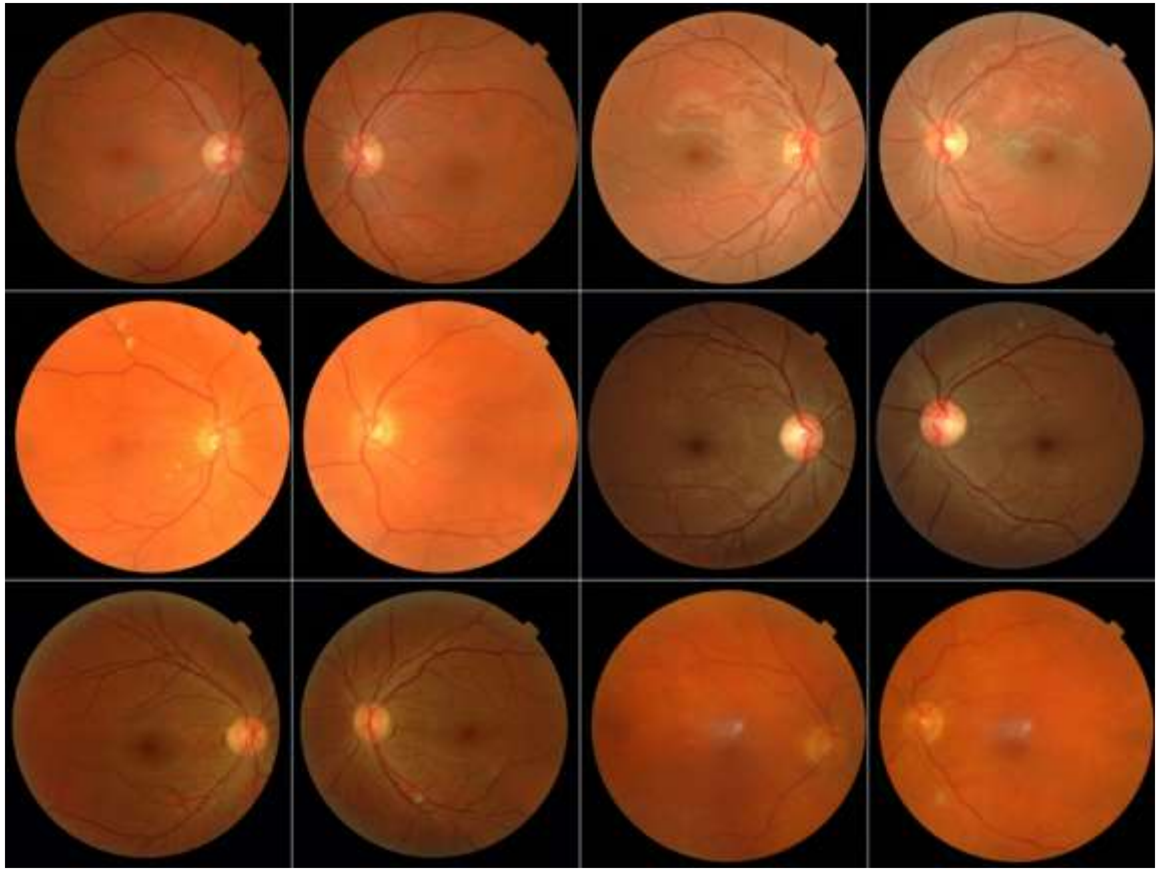
According to the release, the U.S. Department of Energy’s Office of Nuclear Material Safety and Security is leading that team’s investigation.

“The safety of people, the environment and the nation’s nuclear stockpile is our highest priority,” Hicks said. “We will get to the bottom of this and make no excuses.

“Better Language Models and Their Implications”
2/14/19
OPENAI

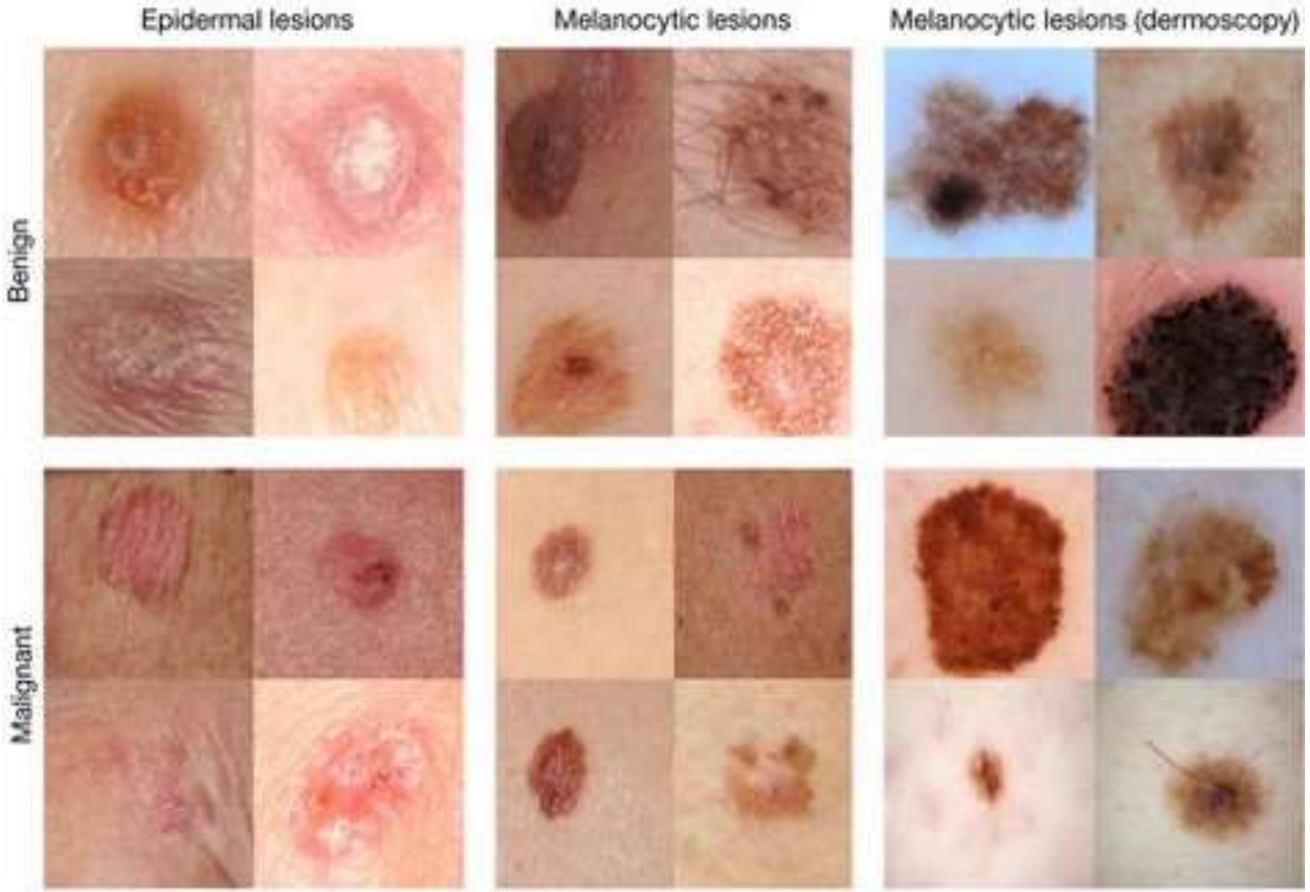
DATA SCIENCE IN MEDICINE

Deep learning-based diagnostics for medical images exceed expert performance



Improved Automated Detection of Diabetic Retinopathy

Invest. Ophthalmol. Vis. Sci.. 2016;57(13):5200-5206. doi:10.1167/iovs.16-19964



Dermatologist-level classification of skin cancer

Nature volume 542, pages 115–118 (02 February 2017)

Natural language processing models are beginning to make an impact

Classification of radiology reports using neural attention models, *IJCNN 2017*

Mass effect from extradural hemorrhage
https://radiopaedia.org

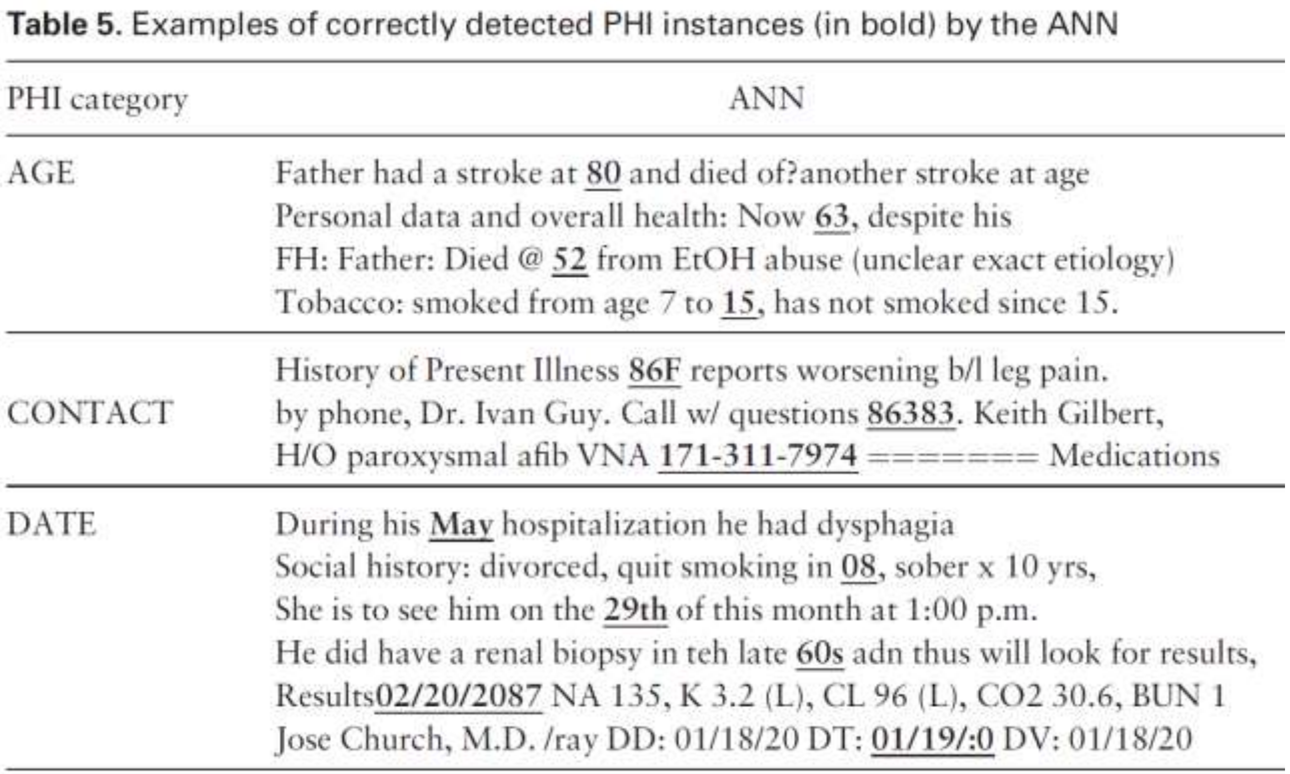


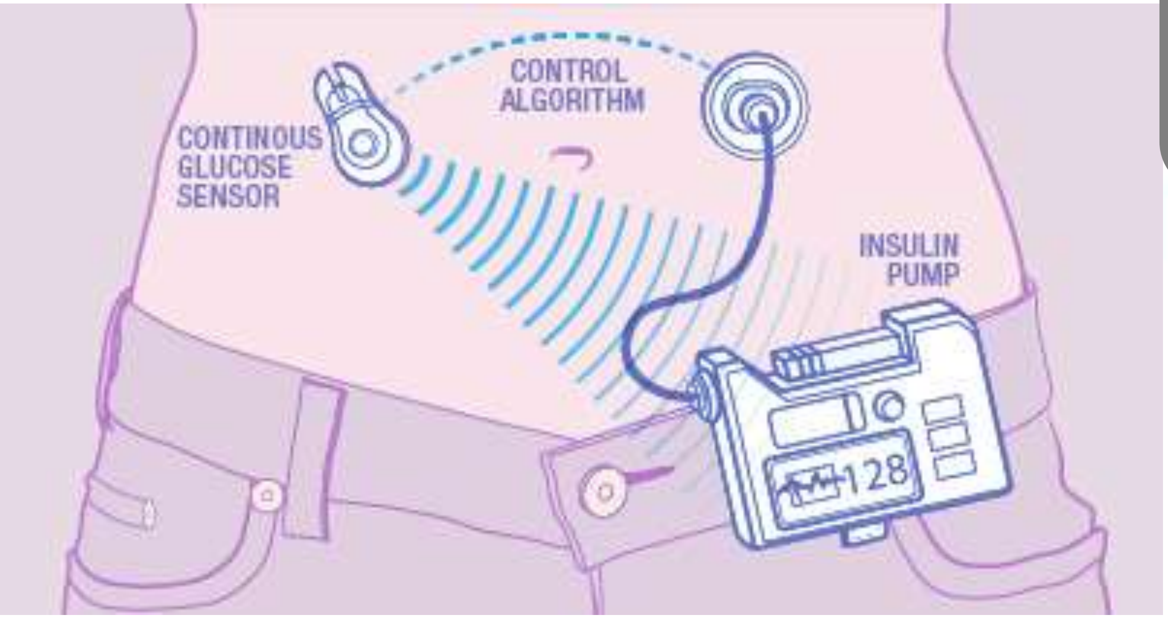
Table 5. Examples of correctly detected PHI instances (in bold) by the ANN

PHI category	ANN
AGE	Father had a stroke at **80** and died of?another stroke at age Personal data and overall health: Now **63**, despite his FH: Father: Died @ **52** from EtOH abuse (unclear exact etiology) Tobacco: smoked from age 7 to **15**, has not smoked since 15.
CONTACT	History of Present Illness **86F** reports worsening b/l leg pain. by phone, Dr. Ivan Guy. Call w/ questions **86383**. Keith Gilbert, H/O paroxysmal afib VNA **171-311-7974** ======= Medications
DATE	During his **May** hospitalization he had dysphagia Social history: divorced, quit smoking in **08**, sober x 10 yrs, She is to see him on the **29th** of this month at 1:00 p.m. He did have a renal biopsy in teh late **60s** adn thus will look for results, Results**02/20/2087** NA 135, K 3.2 (L), CL 96 (L), CO2 30.6, BUN 1 Jose Church, M.D. /ray DD: 01/18/20 DT: **01/19/:0** DV: 01/18/20

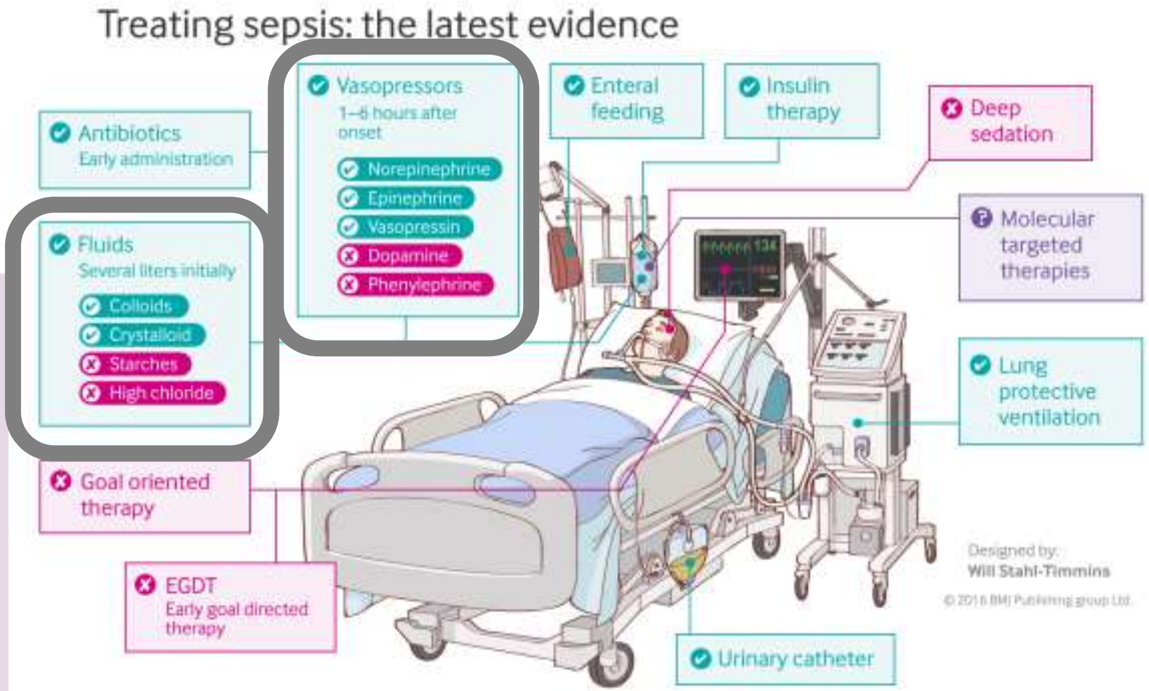
De-identification of patient notes with recurrent neural networks
JAMIA 24(3), 2017, 596–606

Sequential decision-making algorithms can also exceed human performance

Closed-loop blood glucose control (“artificial pancreas”)



https://www.mayo.edu/research/labs/artificial-pancreas/overview



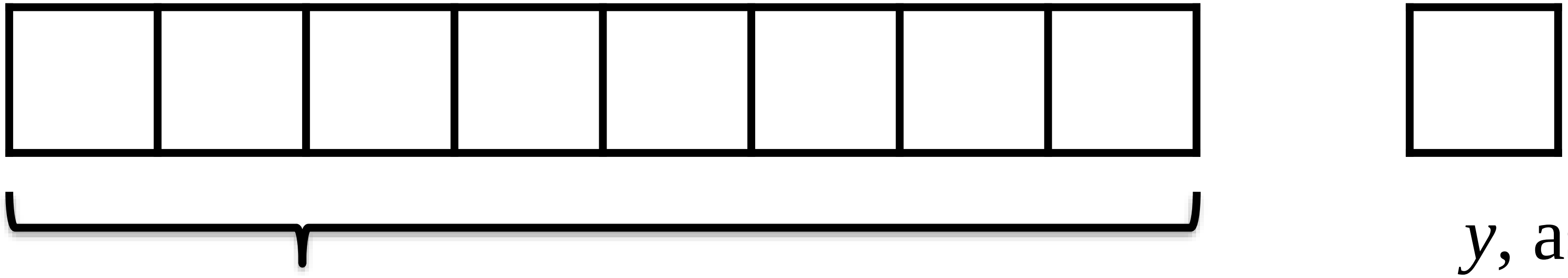
Fluid and vasopressor administration for sepsis treatment

Gotts JE, Matthay MA. Sepsis: pathophysiology and clinical management. bmj. 2016 May 23;353(i1585).

Begin with a simple model, then add complexity

A “SHALLOW” NETWORK: LOGISTIC REGRESSION

First: What is a Predictive Model?

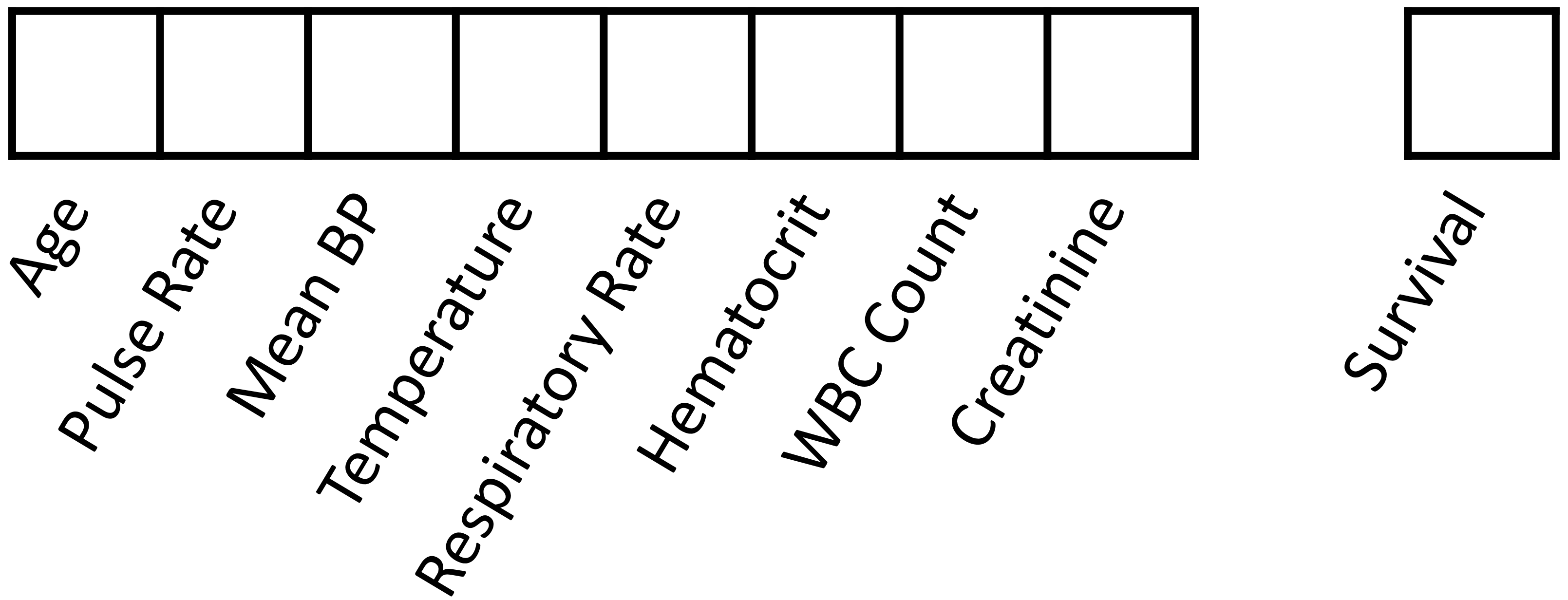


x, data/features for
a subject or patient

y, associated
value or label

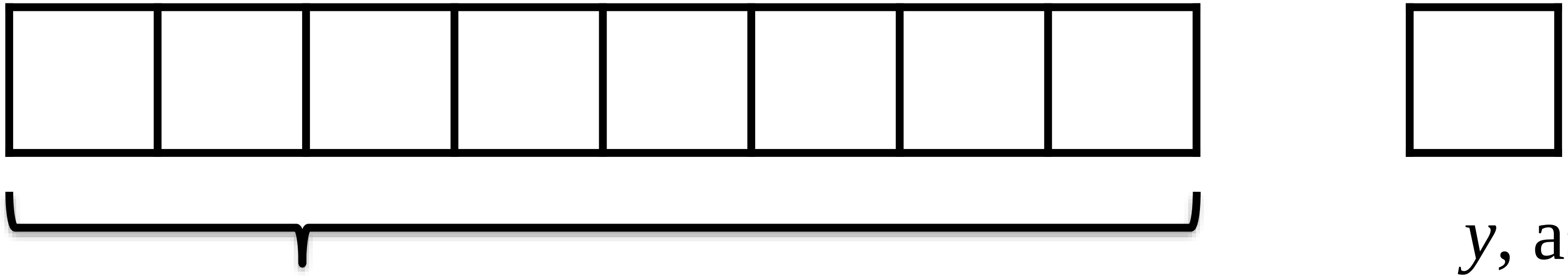
End goal: predict *y* from *x*

ICU Mortality: APACHE III



End goal: predict odds of hospital mortality

Learning a Predictive Model from Labeled Data

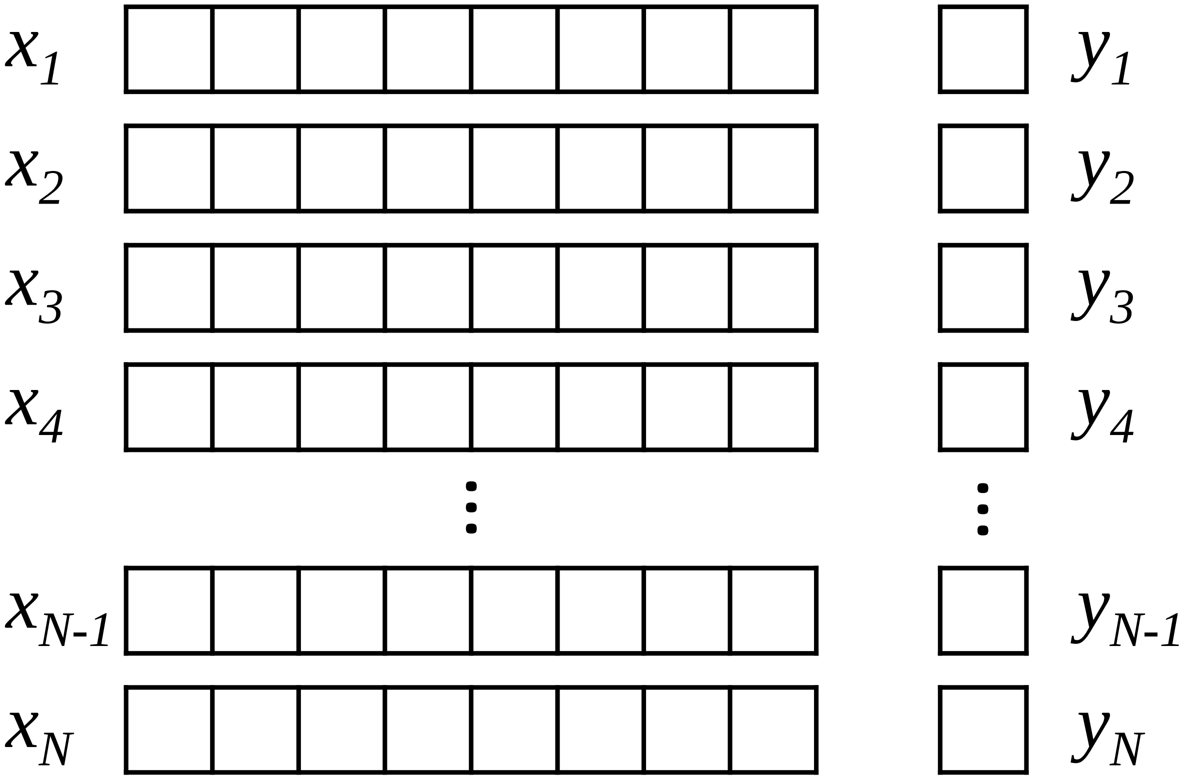


x, data/features for a subject or patient

y, associated value or label

The learning process: find the equation that best predicts *y* based on *x*

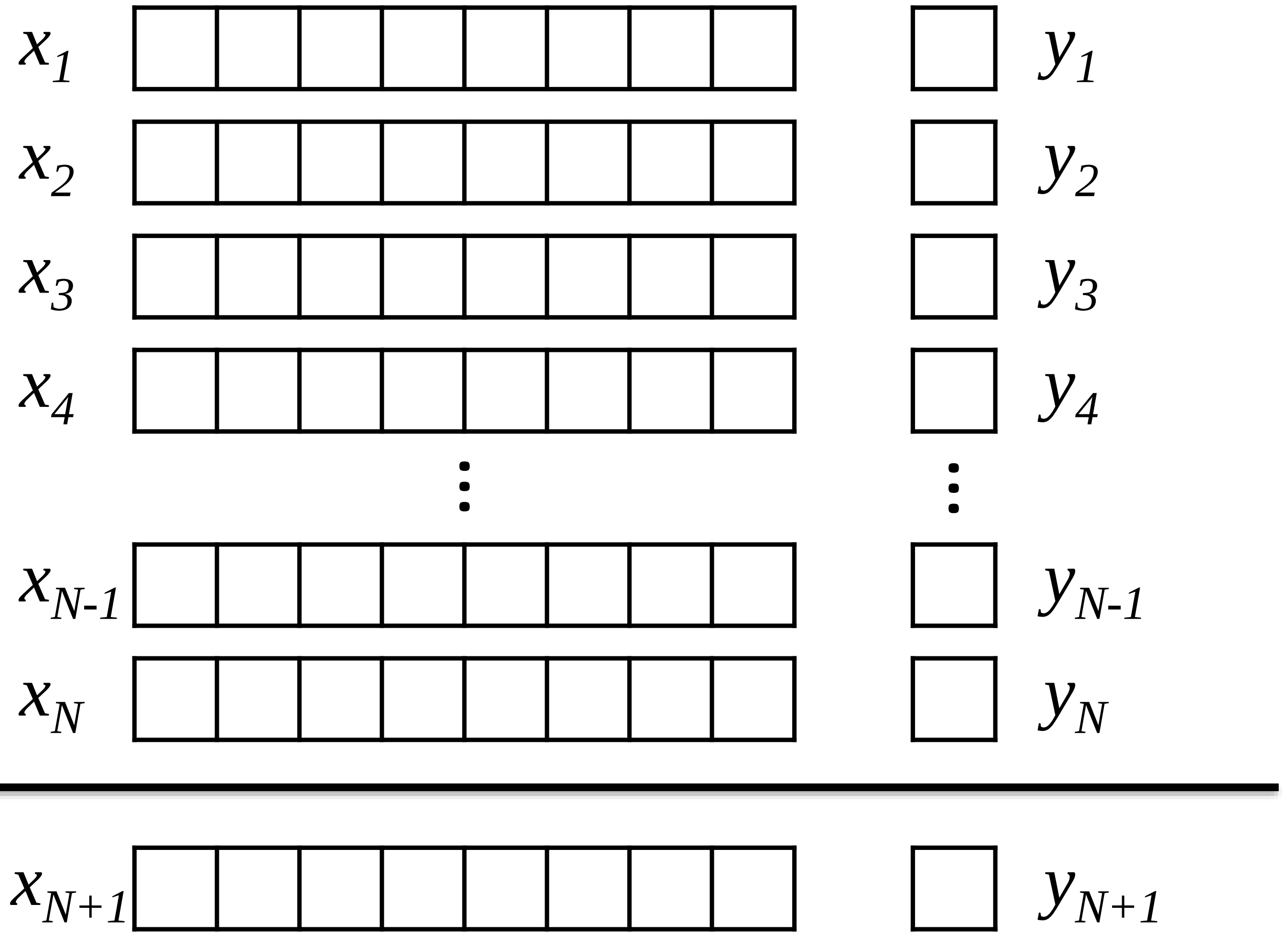
Training Set (Historical Data)



Find an equation that predicts y based on x across the training set

We'll begin by supposing y is binary
(i.e. $y \in \{0, 1\}$)

Making Predictions for New x

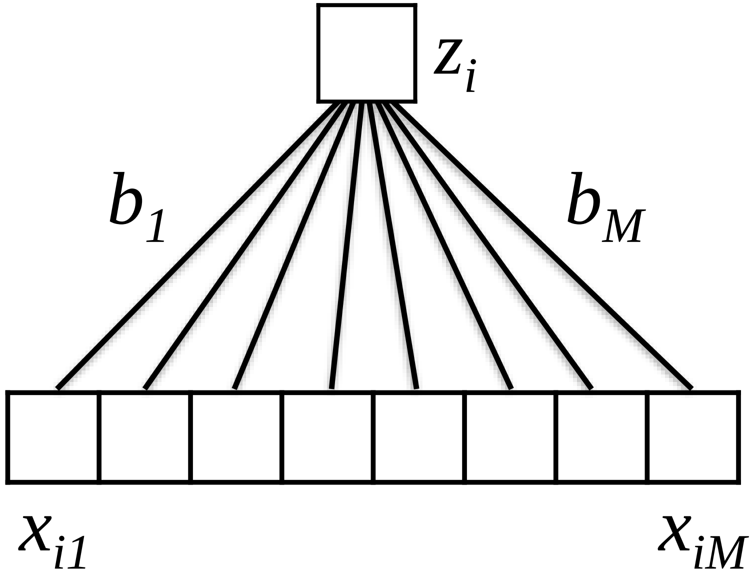


Find an equation that predicts y based on x across the training set

We'll begin by supposing y is binary (i.e. $y \in \{0, 1\}$)

<- Learn to predict new y

Linear Predictive Model



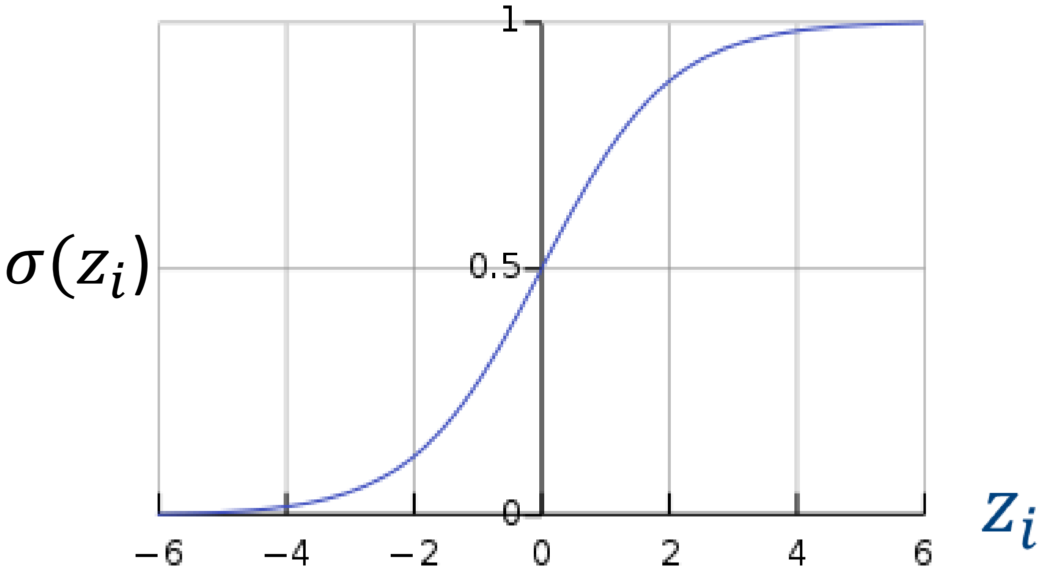
$$z_i = b_1 x_{i1} + b_2 x_{i2} + \cdots + b_M x_{iM}$$

Convert to a Probability

$$z_i = b_1 x_{i1} + b_2 x_{i2} + \cdots + b_M x_{iM} + \cdots + b_0$$

Extra Constant
(i.e. intercept)
(i.e. bias)

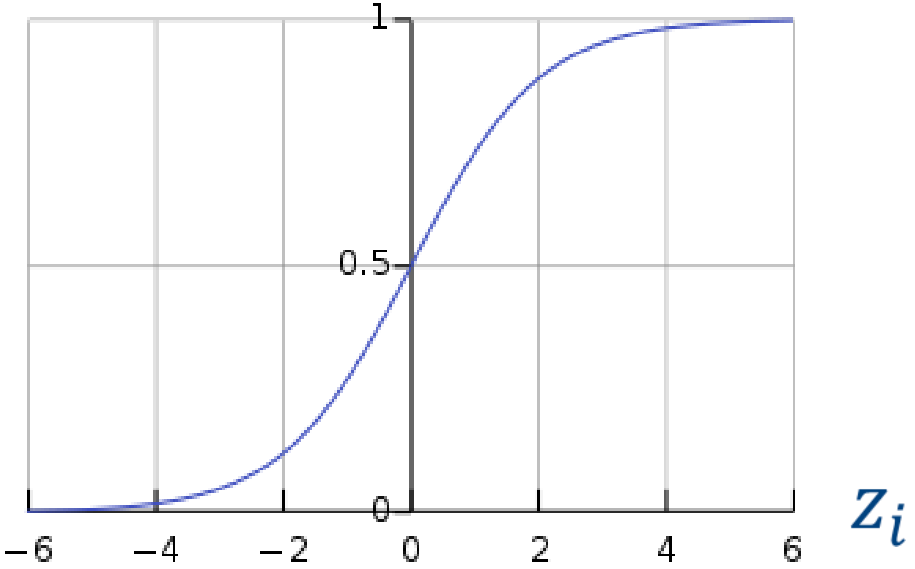
$$p(y_i = 1 | x_i) = \sigma(z_i)$$



Convert to a Probability

$$z_i = b_1 x_{i1} + b_2 x_{i2} + \cdots + b_M x_{iM} + \cdots + b_0$$

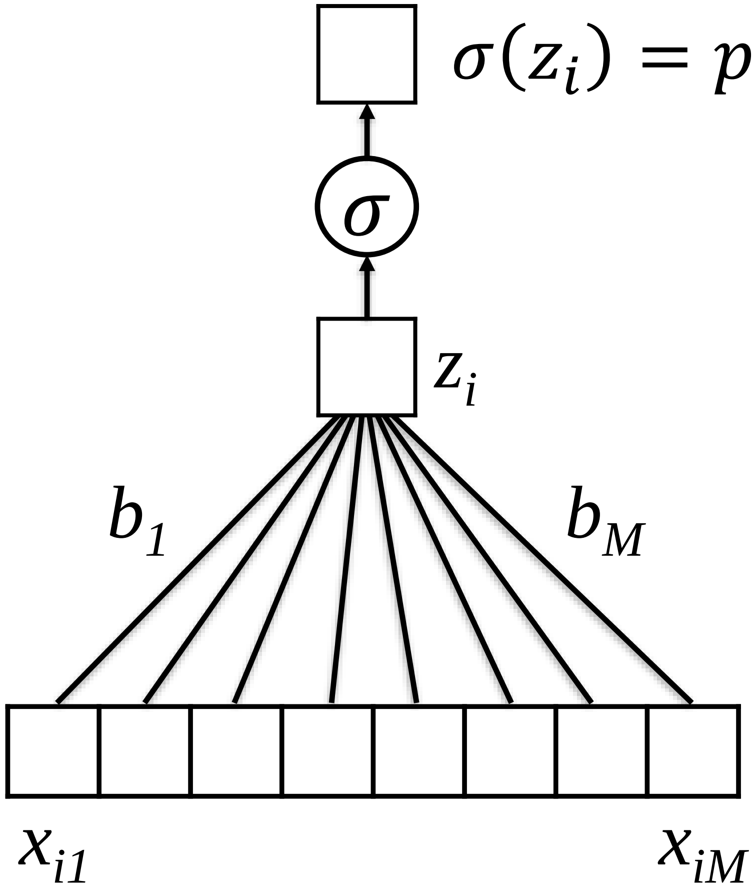
$$p(y_i = 1|x_i) = \sigma(z_i) = \frac{\exp(z_i)}{1+\exp(z_i)}$$



- ❑ Large and positive z_i indicates that event $y_i = 1$ is likely

- ❑ Large and negative z_i indicates that event $y_i = 0$ is likely

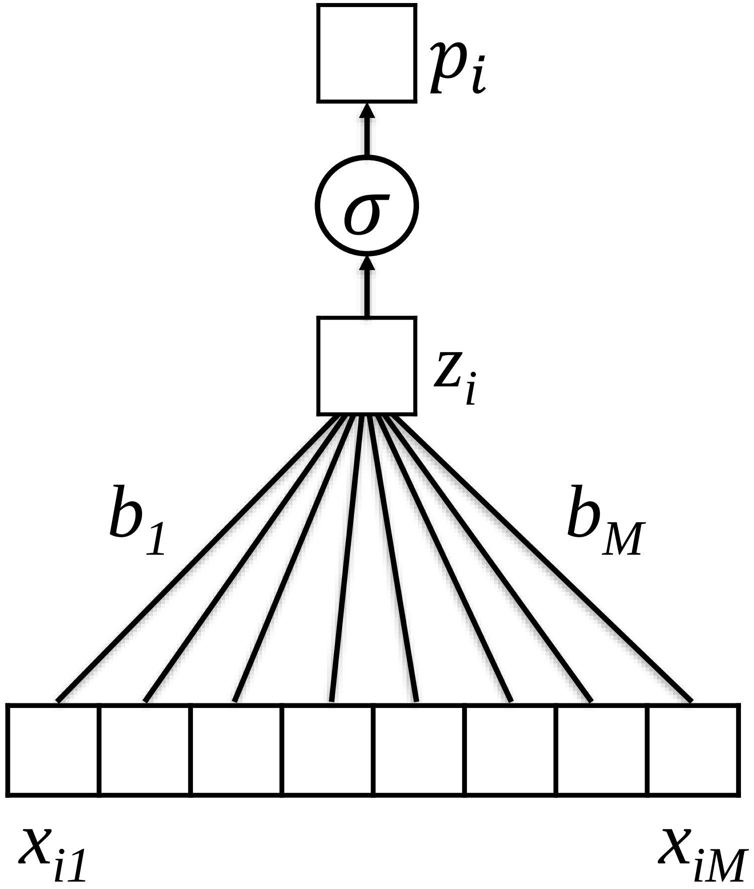
Logistic Regression



$$\sigma(z_i) = p(y_i = 1|x_i)$$

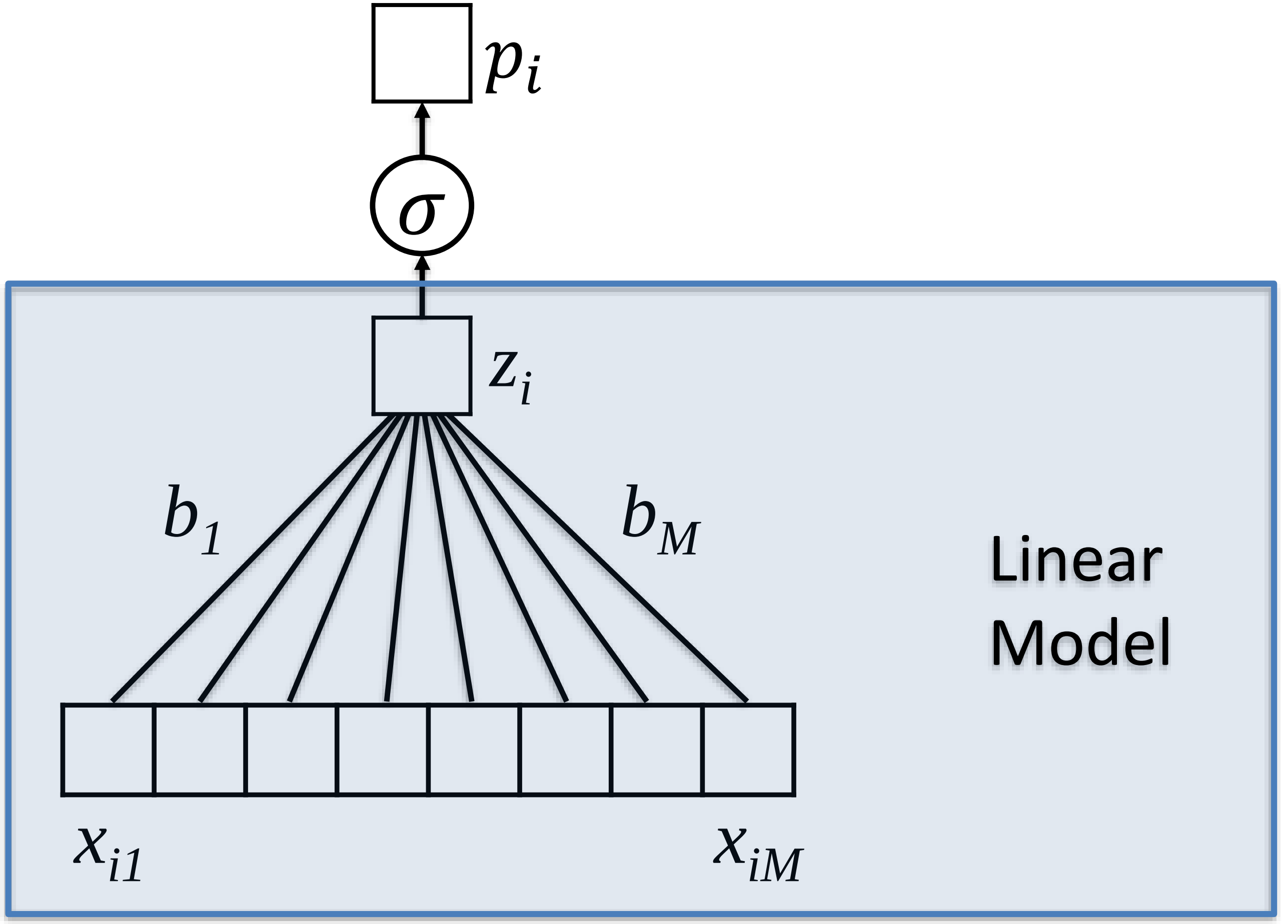
$$p(y_i = 1|x_i) = \sigma(b_1 x_{i1} + b_2 x_{i2} + \cdots + b_M x_{iM})$$

Logistic Regression



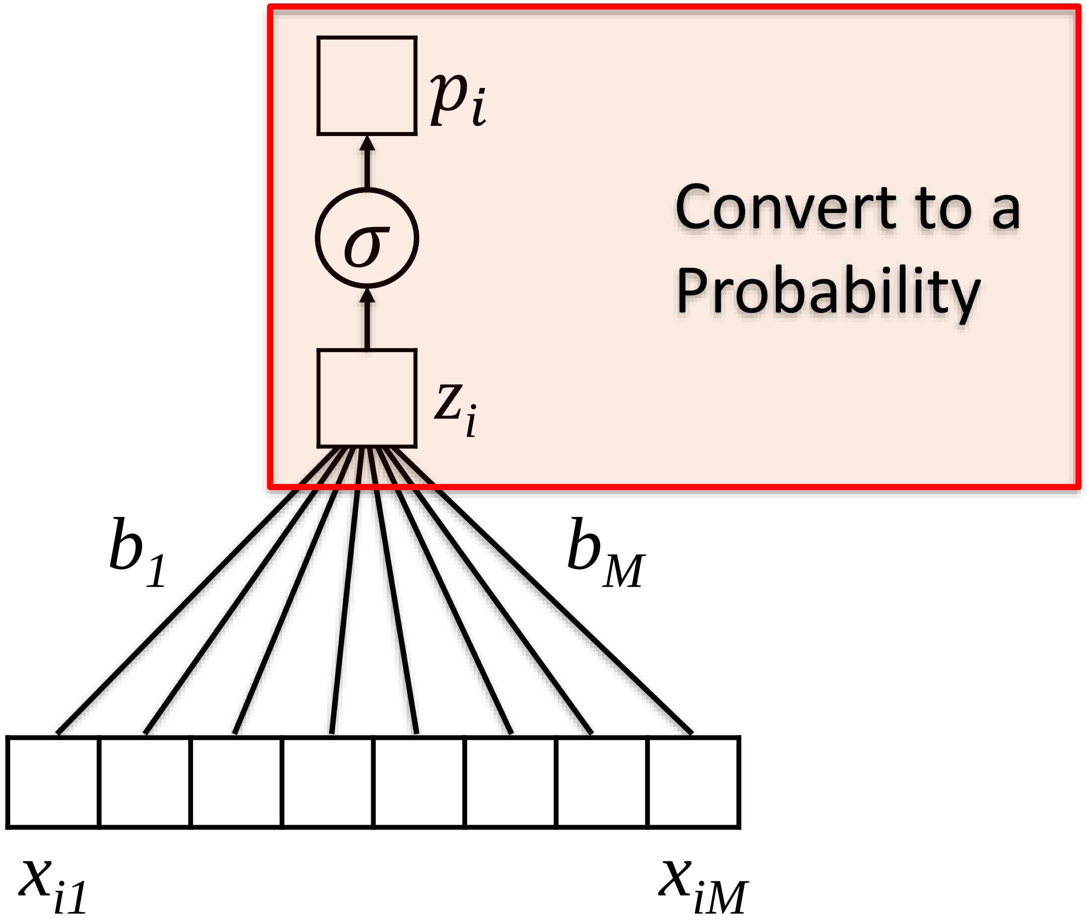
$$p_i = \sigma(b_1 x_{i1} + b_2 x_{i2} + \cdots + b_M x_{iM})$$

Logistic Regression



$$p_i = \sigma(b_1 x_{i1} + b_2 x_{i2} + \cdots + b_M x_{iM})$$

Logistic Regression



Convert to a Probability

$$p_i = \sigma(b_1 x_{i1} + b_2 x_{i2} + \cdots + b_M x_{iM})$$

Illustrative Example

ICU MORTALITY PREDICTION

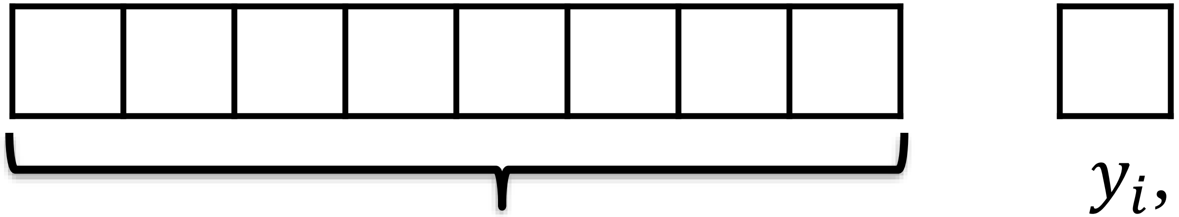
Example: ICU Mortality Prediction

- Outcome:

$$y_i = \begin{cases} 1, \text{patient } i \text{ dies} \\ 0, \text{patient } i \text{ lives} \end{cases}$$

- Features: On admission, what is patient i’s

$$\{\text{age}, \text{sex}, \text{temperature}, \text{blood pressure}, \dots\}$$



y_i, did patient i die

x_i, features for patient i

Example: ICU Mortality Prediction

- Outcome:

$$y_i = \begin{cases} 1, \text{patient } i \text{ dies} \\ 0, \text{patient } i \text{ lives} \end{cases}$$

- Features: On admission, what is patient i's:

$$\{1\text{: age}, 2\text{: sex}, 3\text{: temperature}, 4\text{: blood pressure} \ldots\}$$

$$z_i = b_1 \underset{\substack{\nearrow \\ \text{Age}}}{x_{i1}} + b_2 x_{i2} + \cdots + b_M \underset{\substack{\uparrow \\ \text{Blood Pressure}}}{x_{iM}} + b_0$$

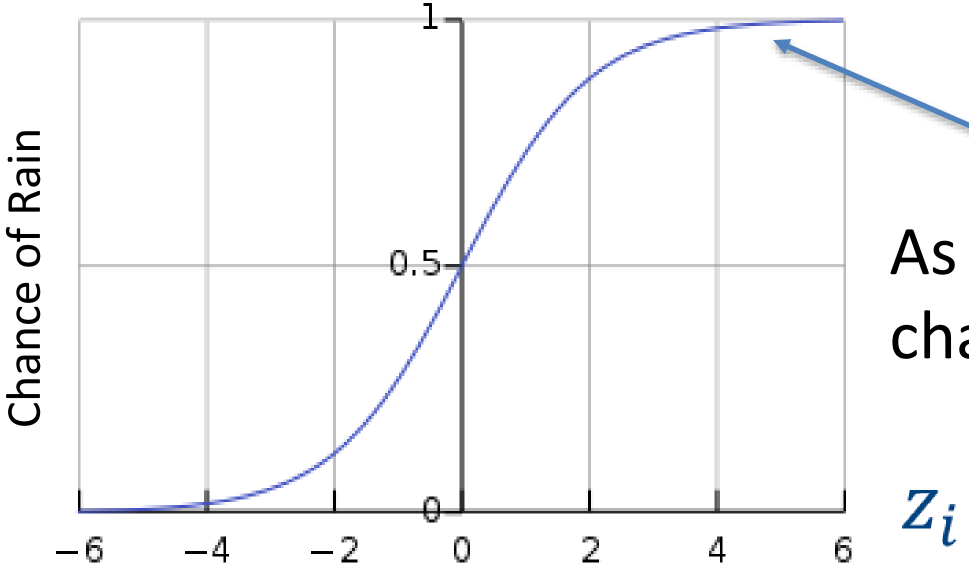
- If increased age increases odds of mortality, b_1 should be positive

Impact on the Sigmoid Function

$$z_i = b_1 x_{i1} + b_2 x_{i2} + \cdots + b_M x_{iM} + b_0$$

Age

$$p(y_i = 1 | x_i) = \sigma(z_i)$$



As the value z_i increases, the chance of mortality increases

Building the Training Set

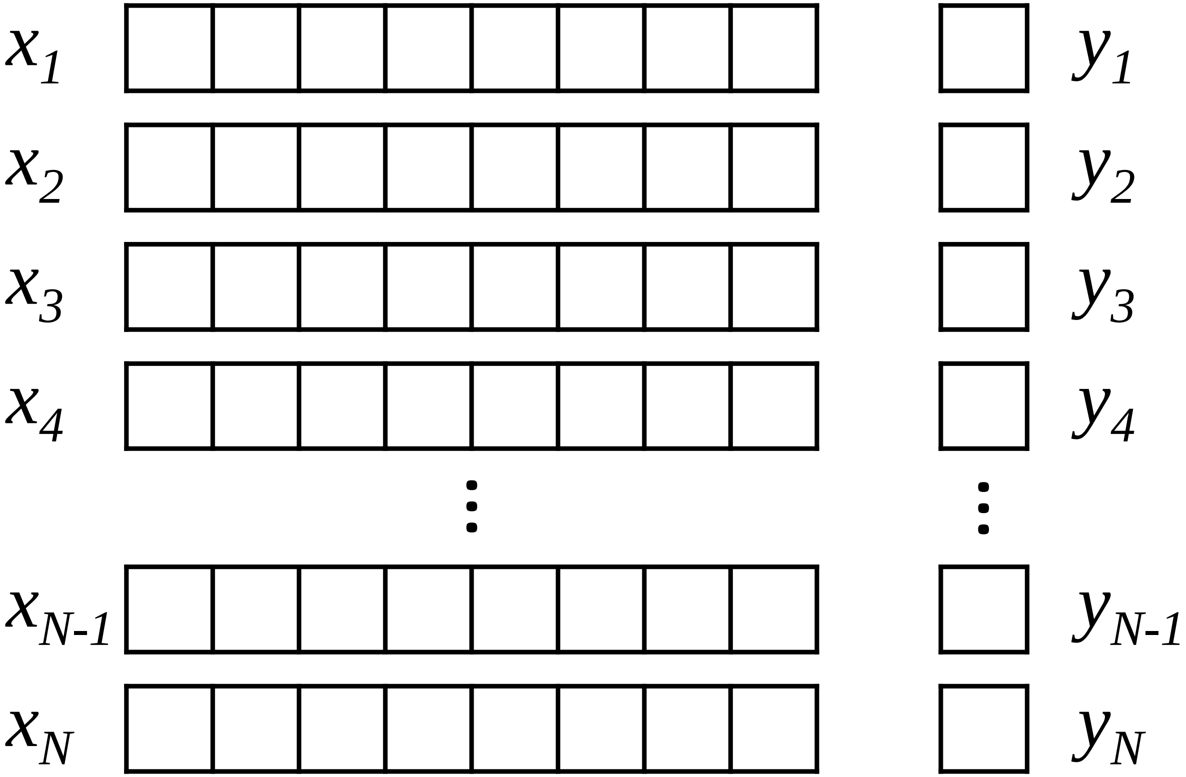
- We want to learn the model parameters

  $b = (b_0, \ldots, b_M)$

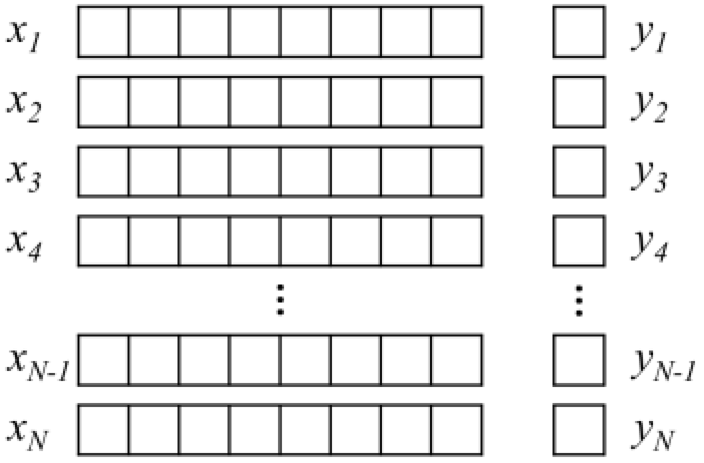
- This requires *training data*; we will find the *b* that match it best

- Record data from N patients
  - Capture features:

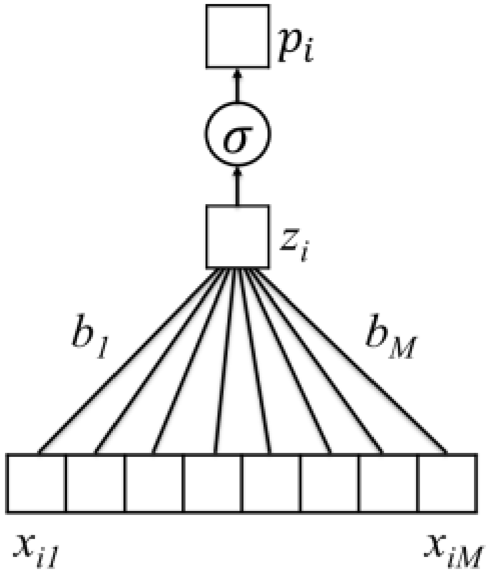
    {age, sex, temp, BP, …}
  - Did they survive?



Learning Model Parameters

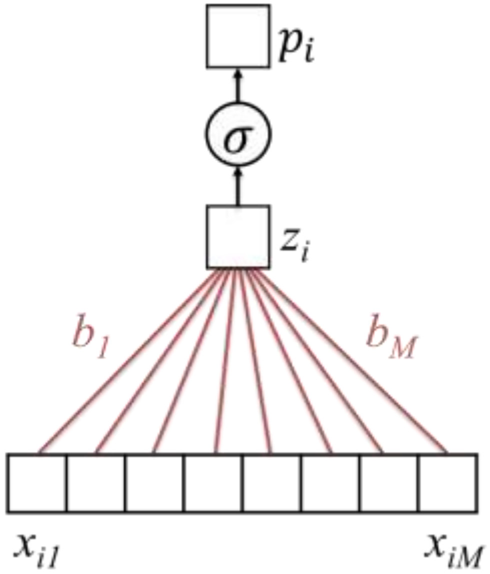


Training Set



$$p_i = \sigma(b_0 + b_1 x_{i1} + b_2 x_{i2} + \cdots + b_M x_{iM})$$

Untrained Logistic Regression Model (or “Network”)

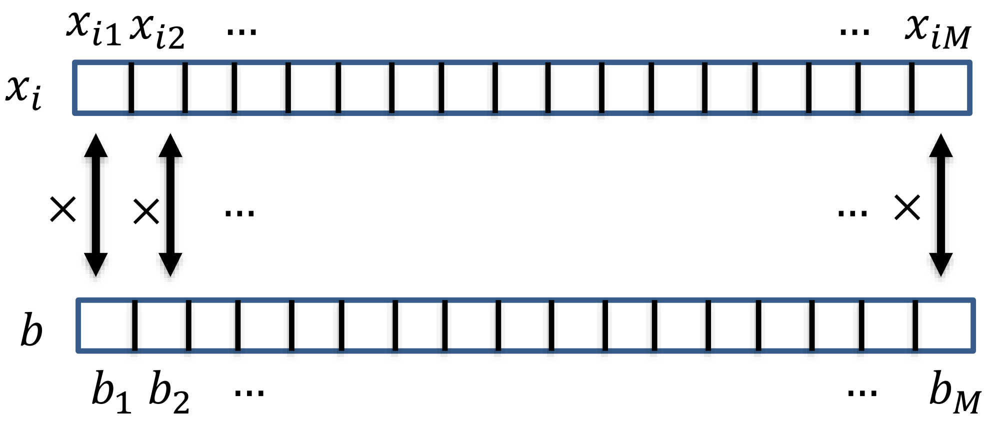


$$b = (b_0, \ldots b_M)$$

Trained Model (with learned parameters)

Simplifying our Notation…

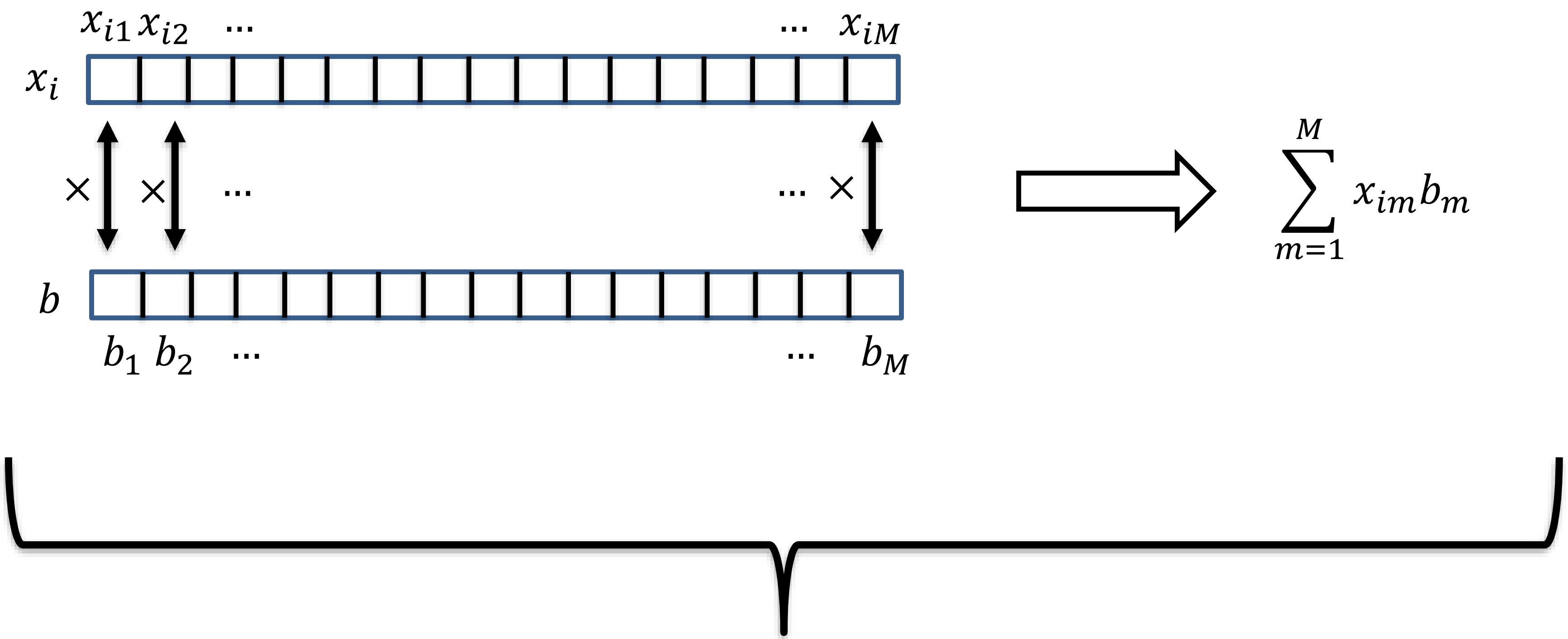
$$z_i = b_1 x_{i1} + b_2 x_{i2} + \cdots + b_M x_{iM}$$



$$\sum_{m=1}^{M} x_{im} b_m$$

Simplifying our Notation…

$$z_i = b_1 x_{i1} + b_2 x_{i2} + \cdots + b_M x_{iM}$$



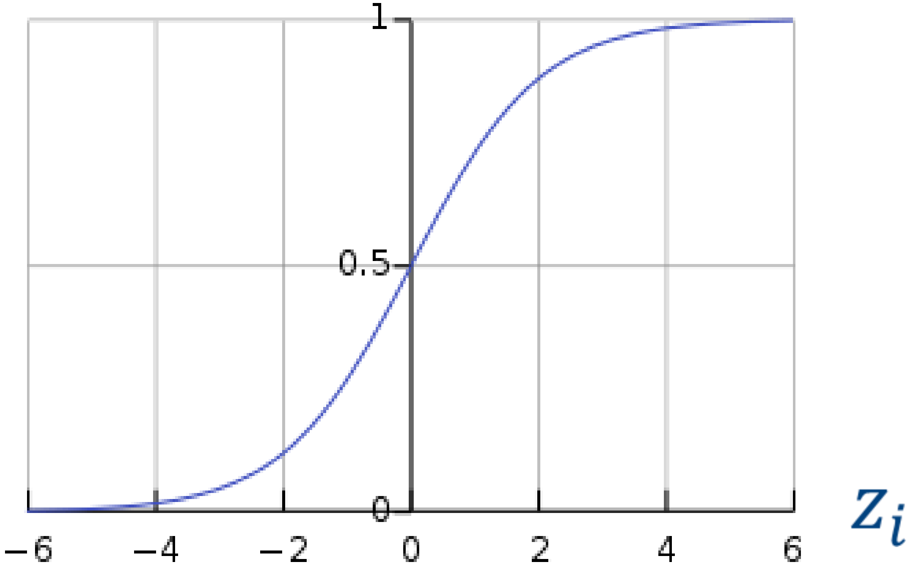
Compact Notation: $x_i \odot b$ (or “inner product”)

Interpretation of Logistic Regression

$$z_i = b_0 + b_1 x_{i1} + b_2 x_{i2} + b_M x_{iM}$$

$$= b_0 + x_i \odot b$$

$$p(y_i = 1|x_i) = \sigma(z_i)$$



- ❑ May think of vector b as a template or filter (will visualize to make clear)
- ❑ If x_i is aligned/matched with b, then $x_i \odot b$ will be large
- ❑ The parameter b_0 is a bias to correct for class prevalence

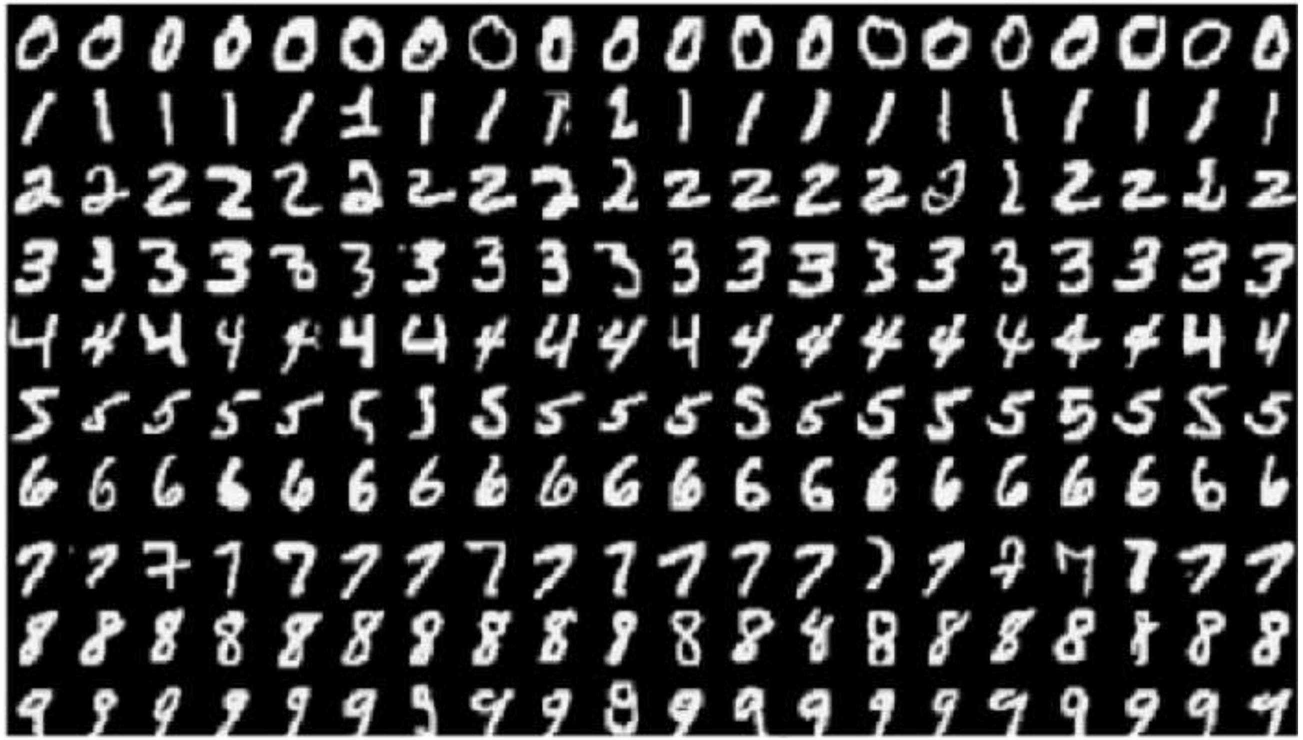
A visual example:

RECOGNIZING HANDWRITTEN DIGITS

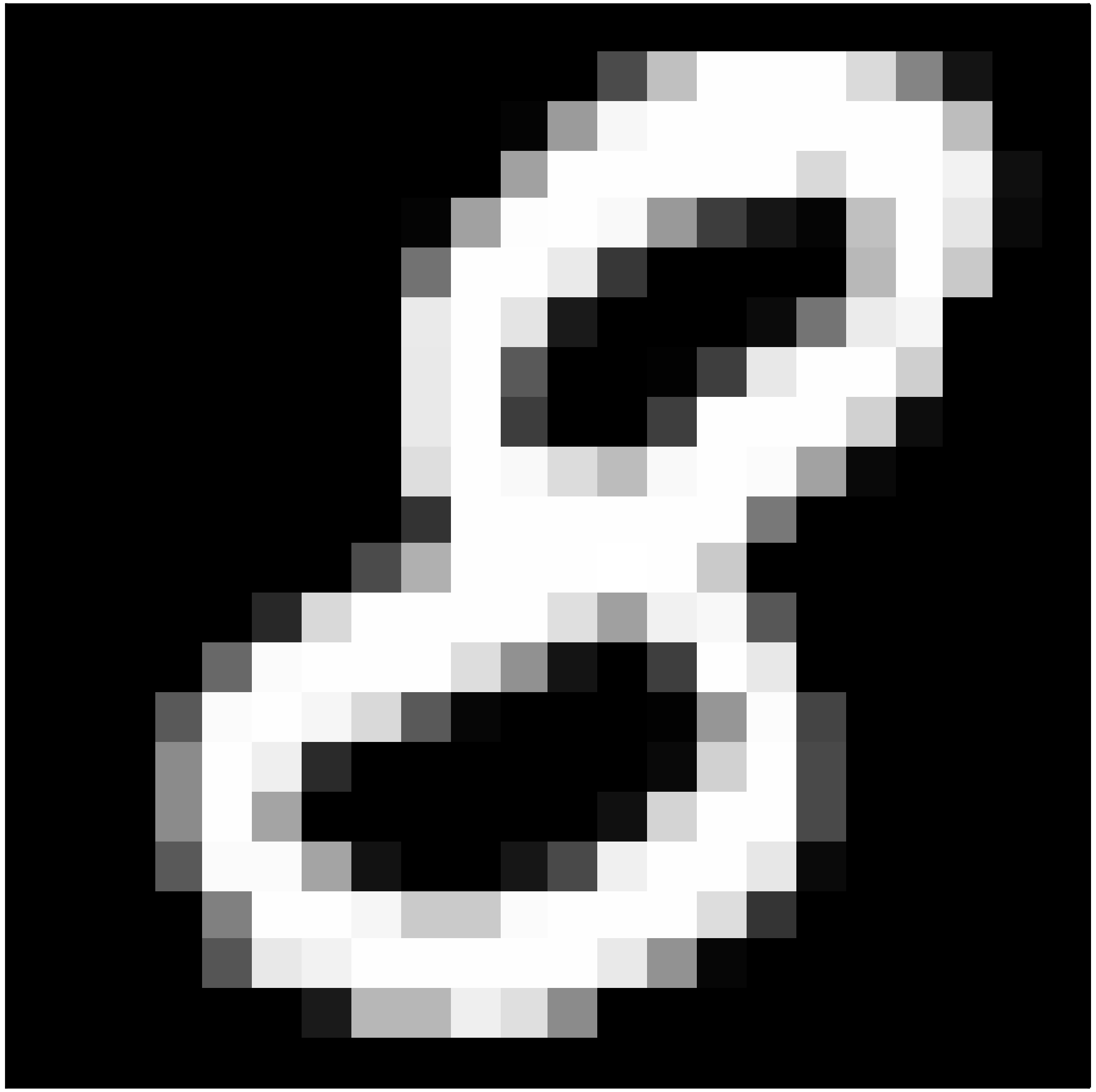
The MNIST Dataset

- The Modified National Institute of Standards and Technology (MNIST) contains pictures of handwritten digits (0,1,2,…)

- Want to be able to tell what digit each image is (*e.g.*, optical character recognition)



Images are Encoded as Numbers



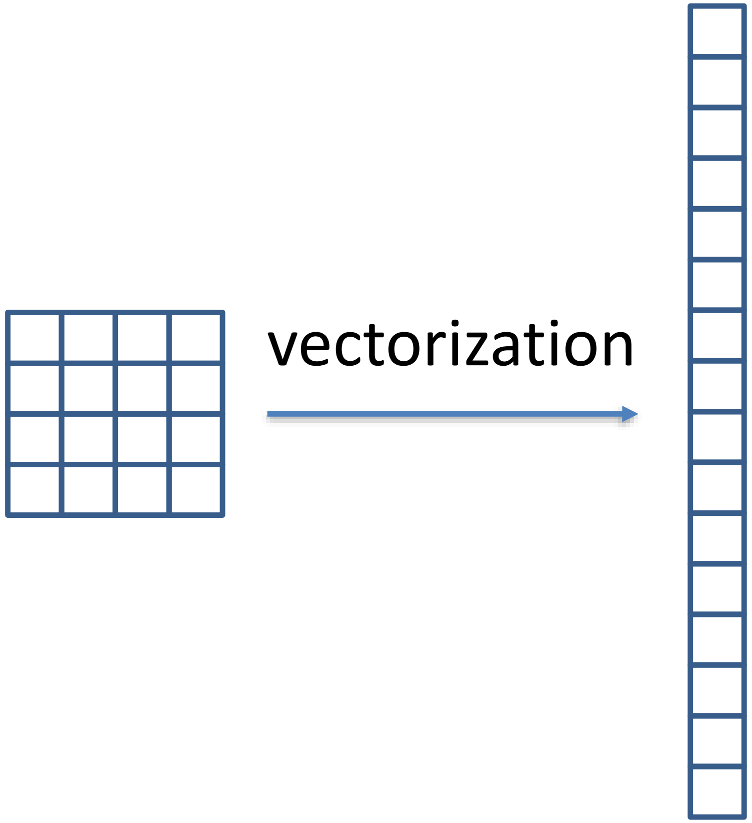
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0	0	118	200	223	155	155	23	0	0	0	0	0
0	0	0	0	0	0	0	0	6	124	214	253	253	253	253	254	229	213	67	0	0	0
0	0	0	0	0	0	0	43	198	253	254	253	247	175	175	236	253	253	108	0	0	0
0	0	0	0	0	0	9	212	253	253	224	58	19	0	0	16	139	247	247	71	0	0
0	0	0	0	0	0	58	254	254	186	14	0	0	0	0	0	0	139	254	118	0	0
0	0	0	0	0	0	58	253	183	8	0	0	0	0	0	0	36	222	253	118	0	0
0	0	0	0	0	0	54	250	128	2	0	0	0	6	71	192	237	253	247	71	0	0
0	0	0	0	0	0	0	213	253	50	0	18	123	198	253	254	253	247	85	0	0	0
0	0	0	0	0	0	0	69	241	227	136	200	253	253	253	254	192	34	0	0	0	0
0	0	0	0	0	0	0	0	175	254	256	254	254	254	149	59	0	0	0	0	0	0
0	0	0	0	0	0	0	101	253	253	254	253	253	253	42	0	0	0	0	0	0	0
0	0	0	0	0	8	138	247	253	243	159	196	243	253	199	0	0	0	0	0	0	0
0	0	0	0	12	183	253	253	253	50	0	0	49	253	214	0	0	0	0	0	0	0
0	0	0	0	180	254	253	213	50	2	0	0	71	253	214	0	0	0	0	0	0	0
0	0	0	0	234	217	97	10	0	0	0	23	207	254	215	0	0	0	0	0	0	0
0	0	0	174	253	156	0	0	0	0	45	215	253	253	95	0	0	0	0	0	0	0
0	0	9	210	253	163	5	19	49	130	244	253	251	137	4	0	0	0	0	0	0	0
0	0	13	229	253	254	192	253	253	253	254	253	137	0	0	0	0	0	0	0	0	0
0	0	0	160	253	254	253	253	253	253	239	132	4	0	0	0	0	0	0	0	0	0
0	0	0	18	112	194	254	254	254	163	59	0	0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0

Vectorization

- We will start talking about deep learning *without* using the structure of the image

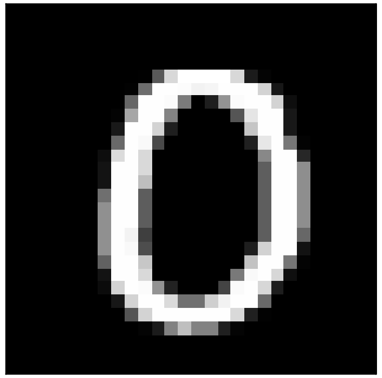
- Later, in block 2, we will consider how to take advantage of this structure

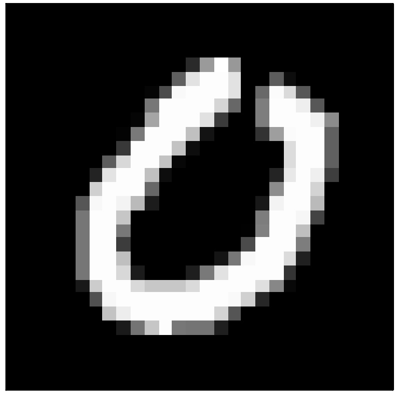
- To convert an image into an unstructured set of numbers, we *vectorize* (or *flatten*) it

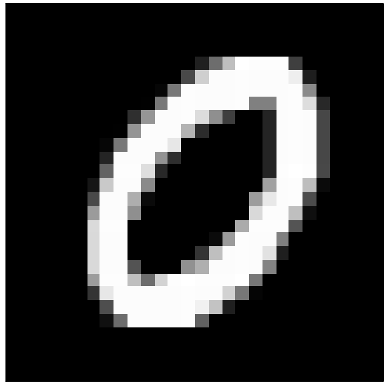


Start With The Binary Case

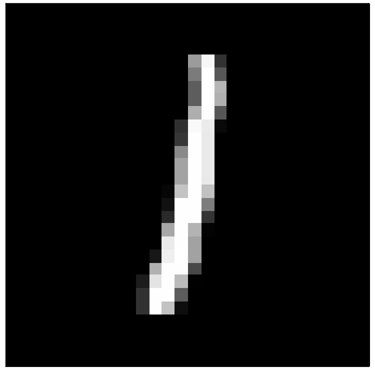
Zeros

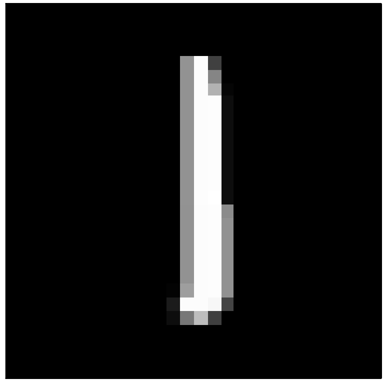


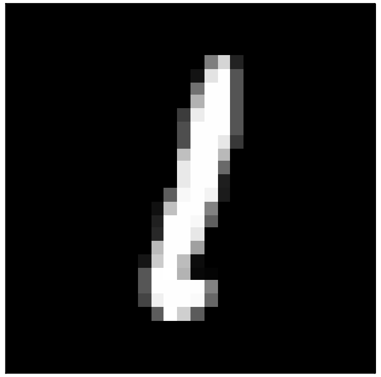




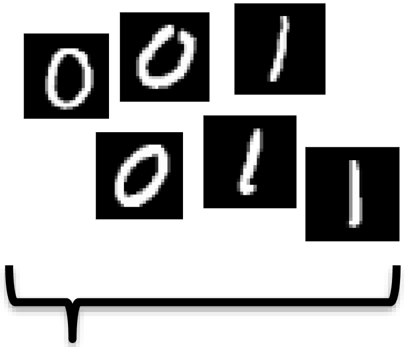
Ones





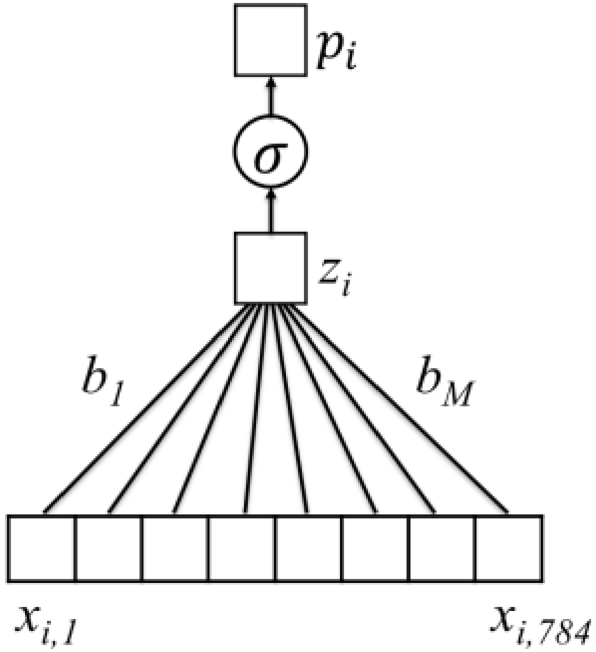


Learning on MNIST



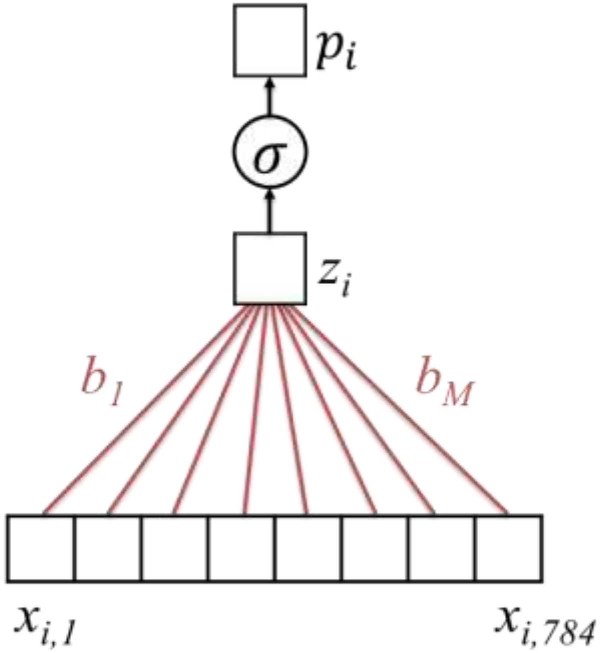
Training set:
28 x 28 images

Vectorize



$p_i = \sigma(b_0 + b_1 x_{i1} + b_2 x_{i2} + \cdots + b_M x_{iM})$

Untrained Logistic Regression Model (or “Network”)

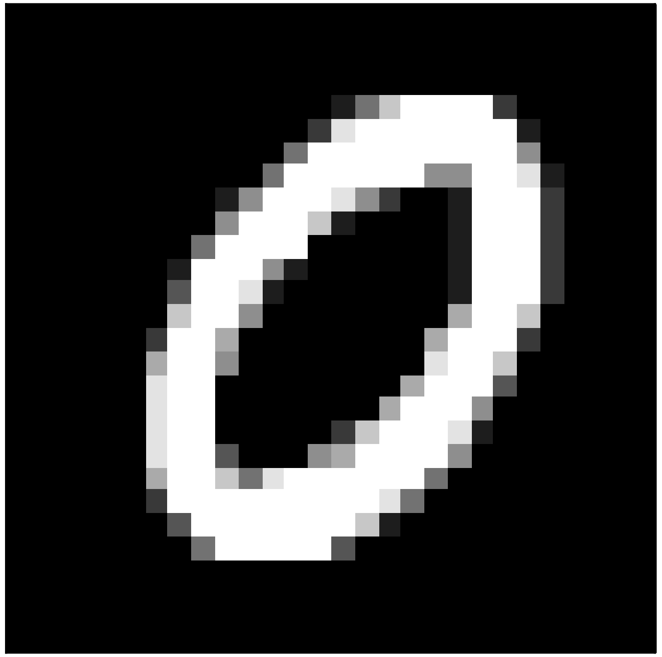


$b = (b_0, \dots b_M)$

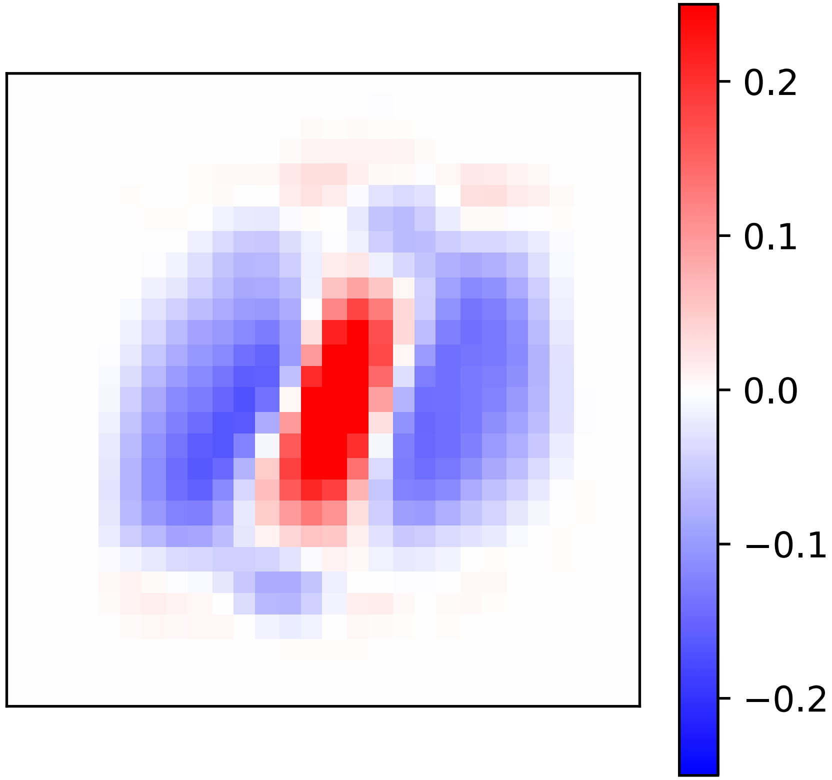
Trained Model (with learned parameters)

Zooming in on 0/1

$\sigma($



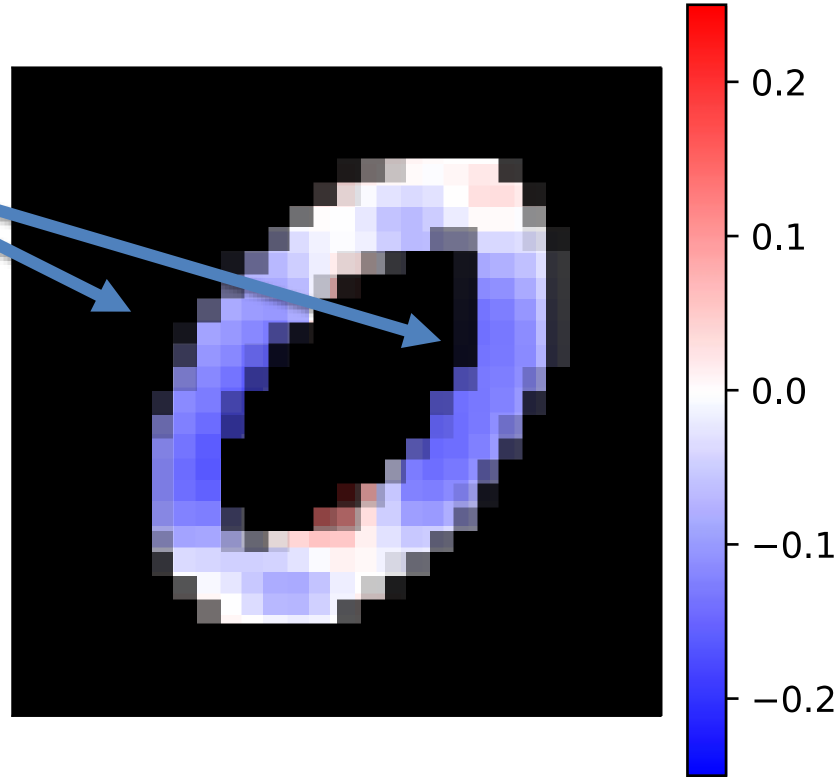
$\odot$



$)$

Zooming in on 0/1

$\sigma($

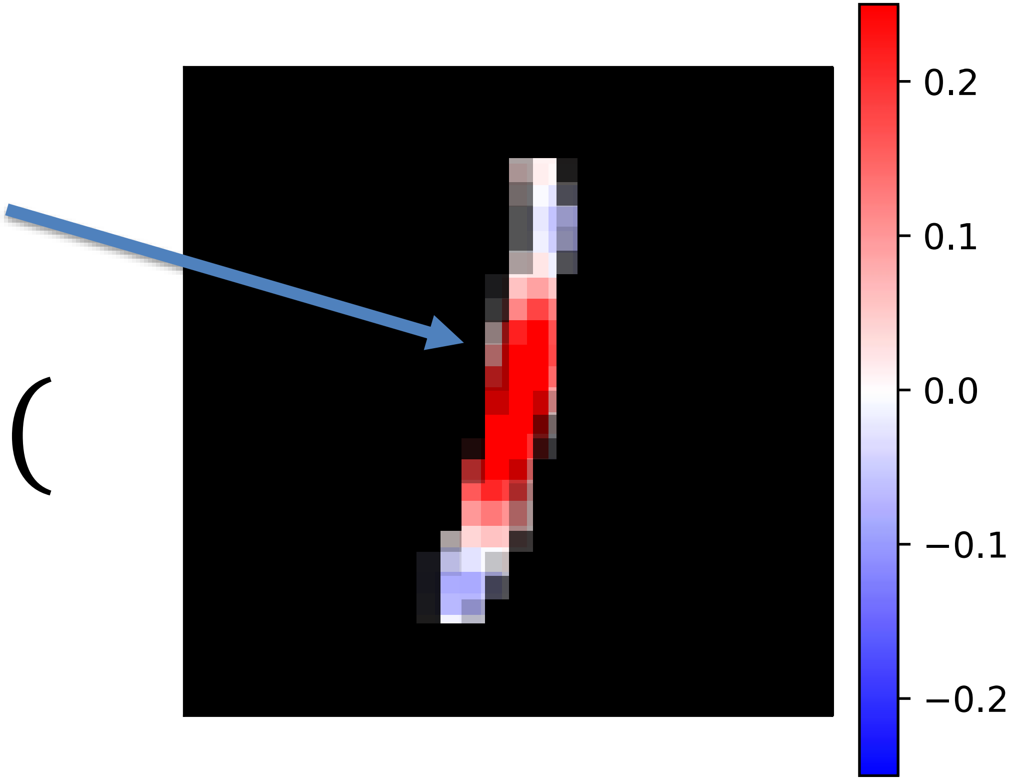


$) = 0.006$

Negative Sections

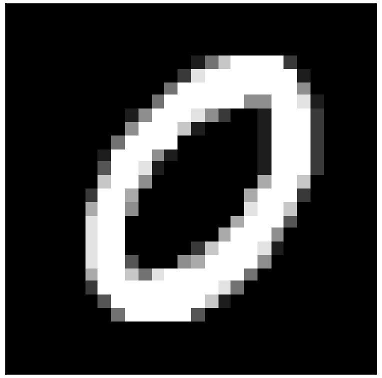
Zooming in on 0/1

Positive Section

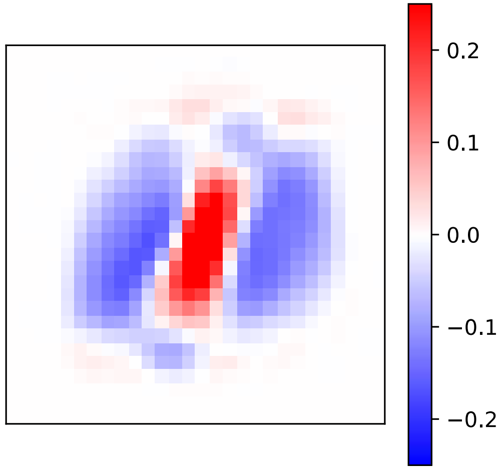
$\sigma($  $) = .991$

Learned Weights for 0/1

$\sigma($



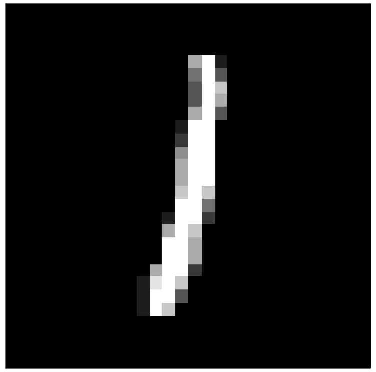
$\odot$



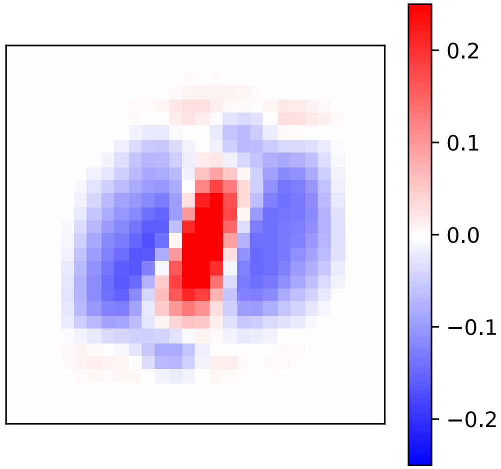
$) \; = \; .006$

We think that this is a “zero” (.6% chance it is a “one”)

$\sigma($



$\odot$



$) \; = \; .991$

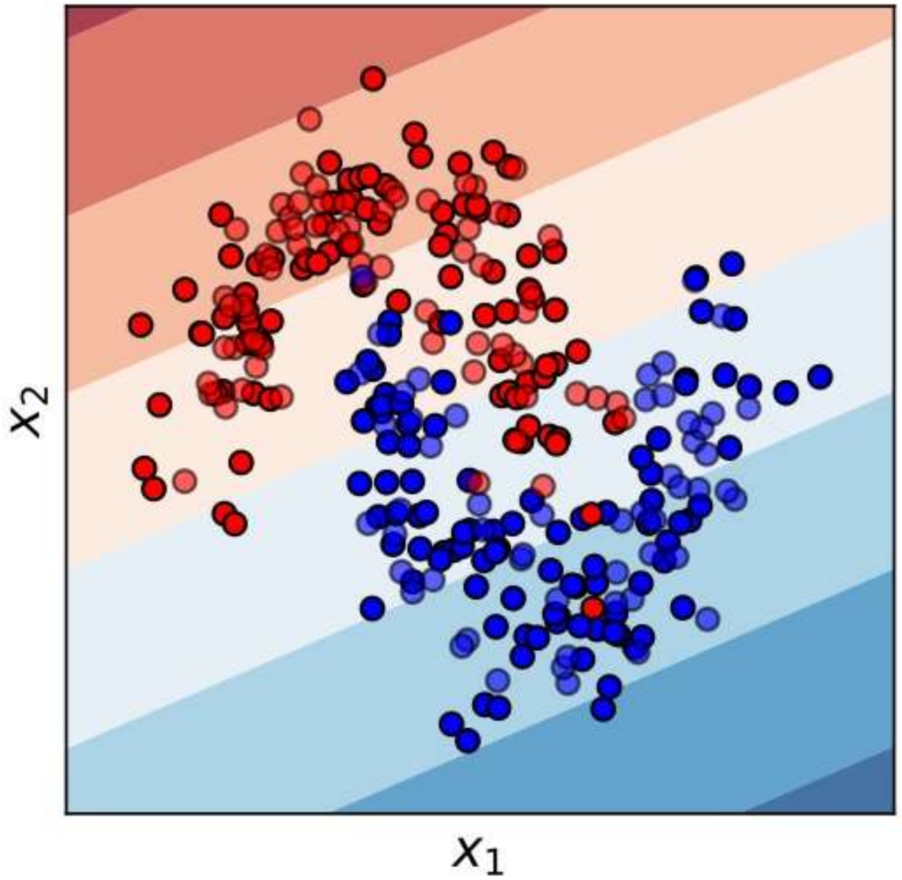
We think that this is a “one” (99.1% chance it is a “one”)

From Shallow to Deep Learning

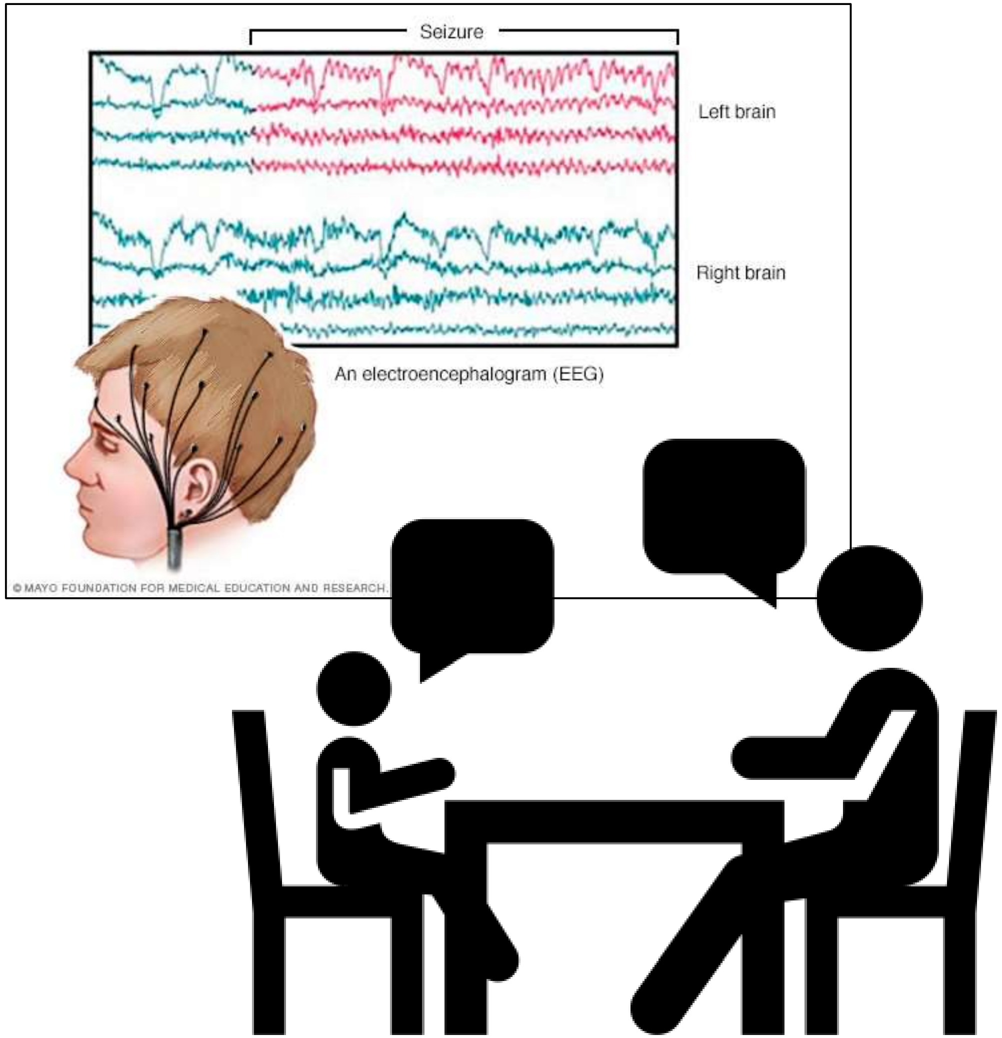
GENERALIZING LOGISTIC REGRESSION

Logistic Regression is a “Linear” Classifier

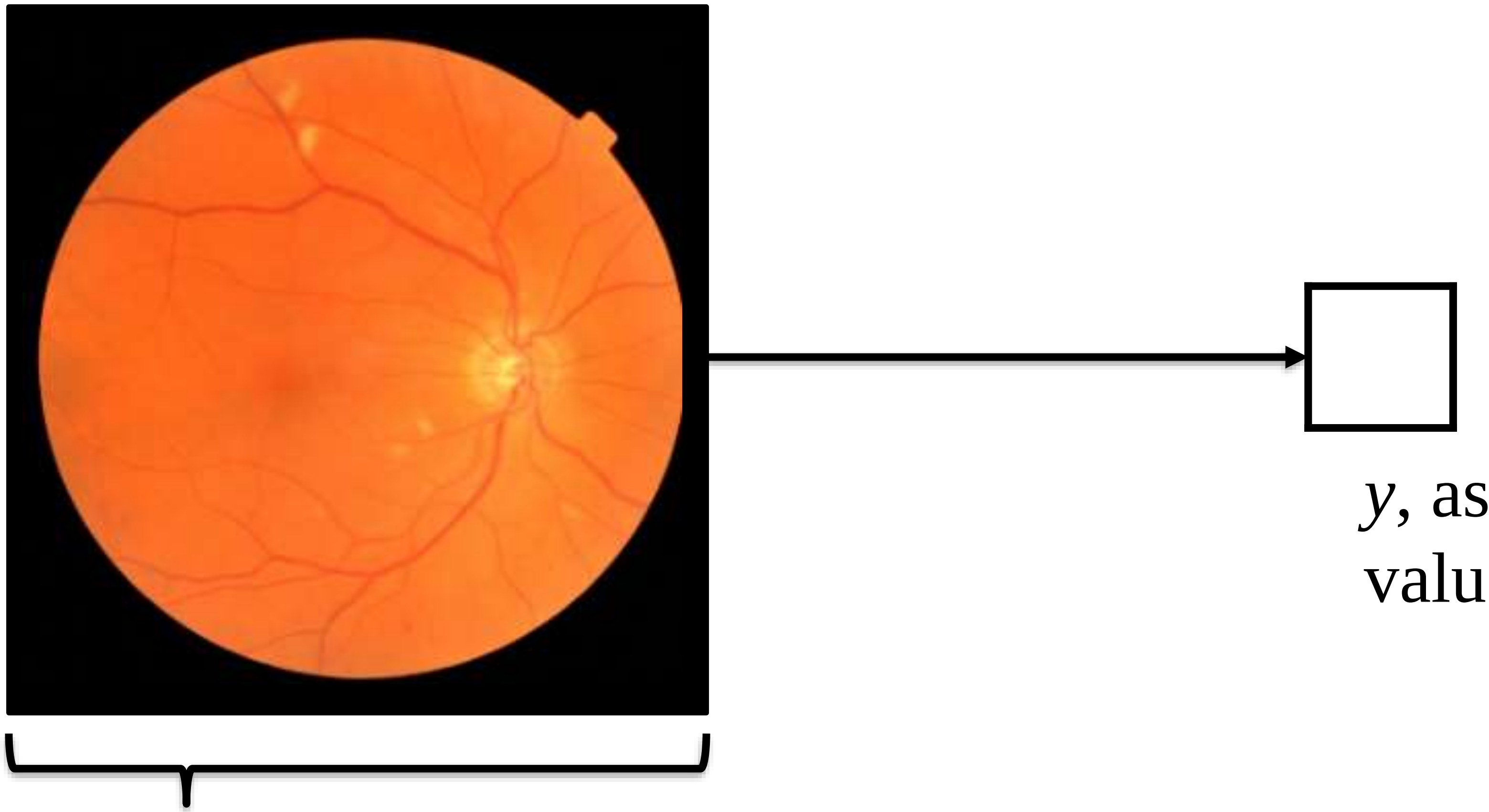
- A “generalized linear model”
- Can only split data by linear trends



Spatial, Temporal, and/or Semantic Structure


Seizure
Left brain
Right brain
An electroencephalogram (EEG)
© MAYO FOUNDATION FOR MEDICAL EDUCATION AND RESEARCH.

What Do Individual Pixels Tell Us about *y*?



x, data/features for a subject or patient

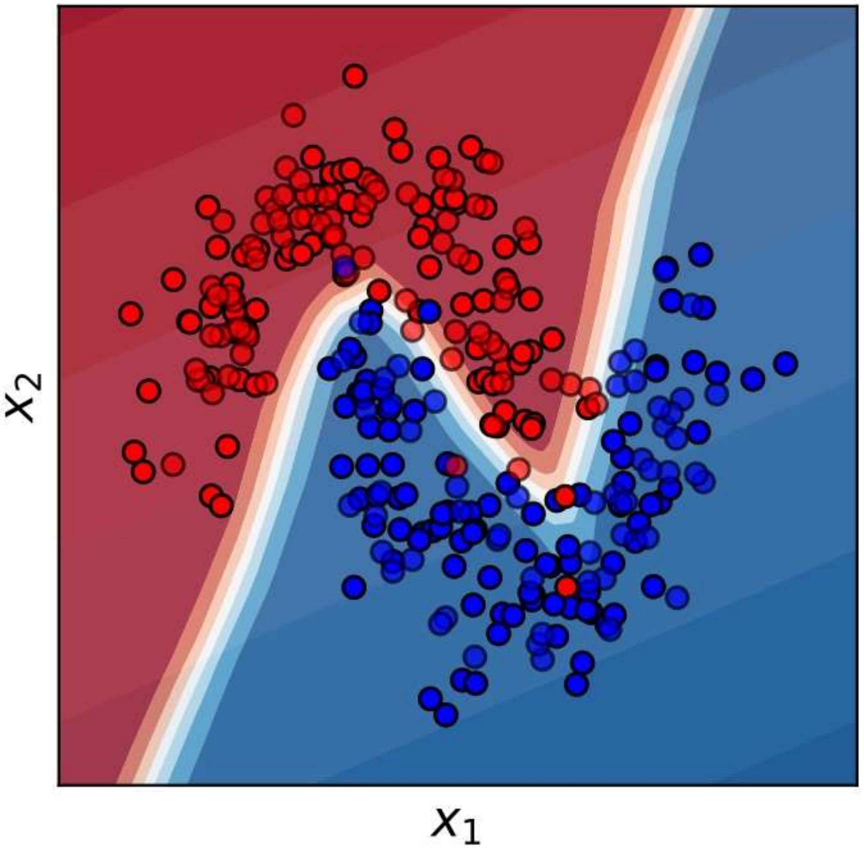
y, associated value or label

End goal: predict *y* from *x*

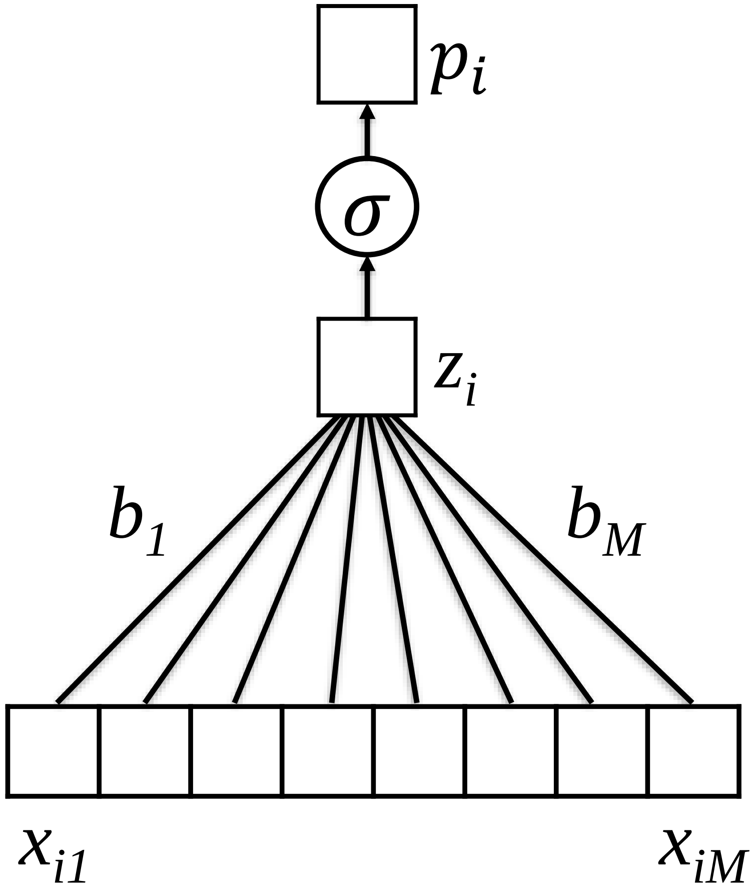
We need more flexible, non-linear classifiers

- Many ways to achieve this…

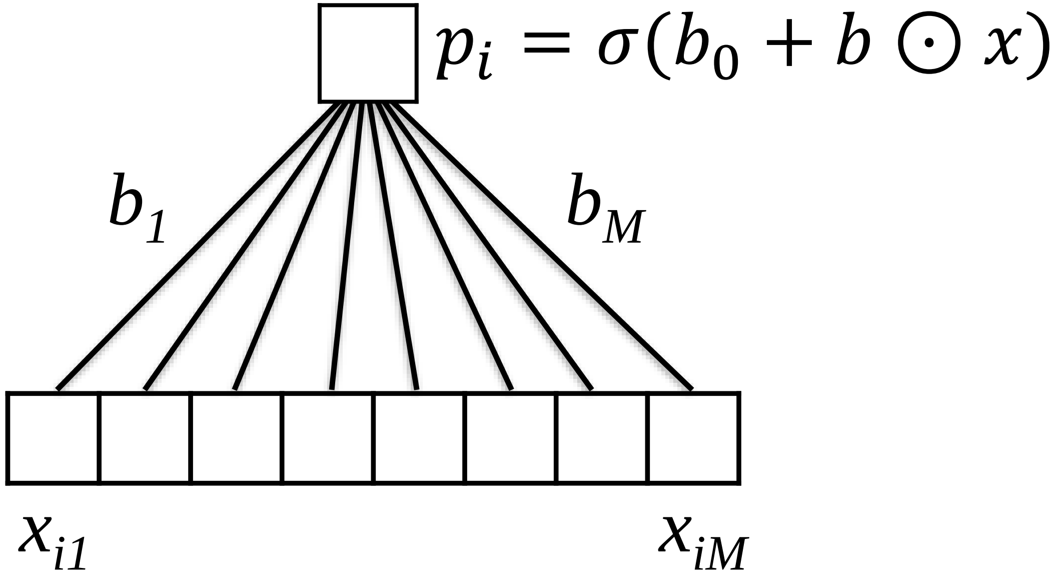
- One of them is to “extend” logistic regression to form a multilayer perceptron, i.e. a neural network



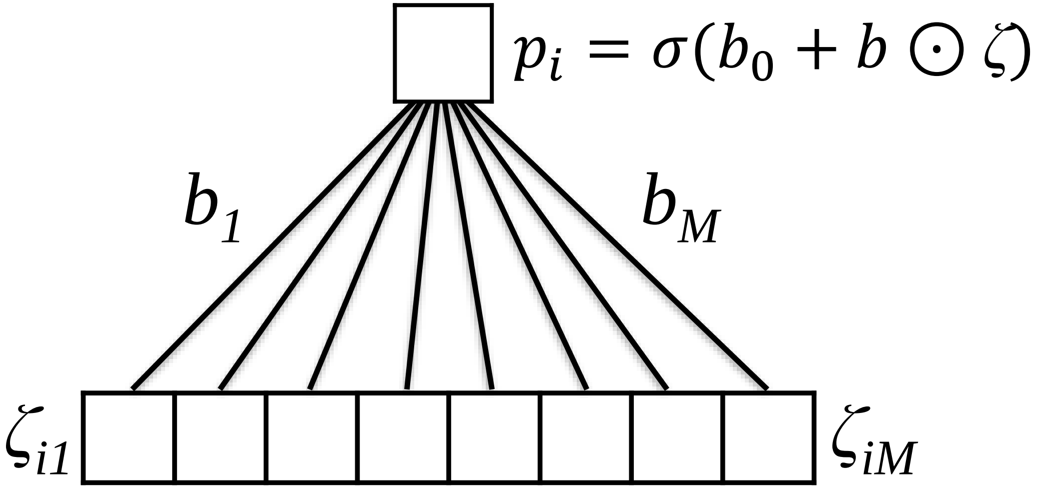
How can we modify logistic regression to learn complex, nonlinear relationships?



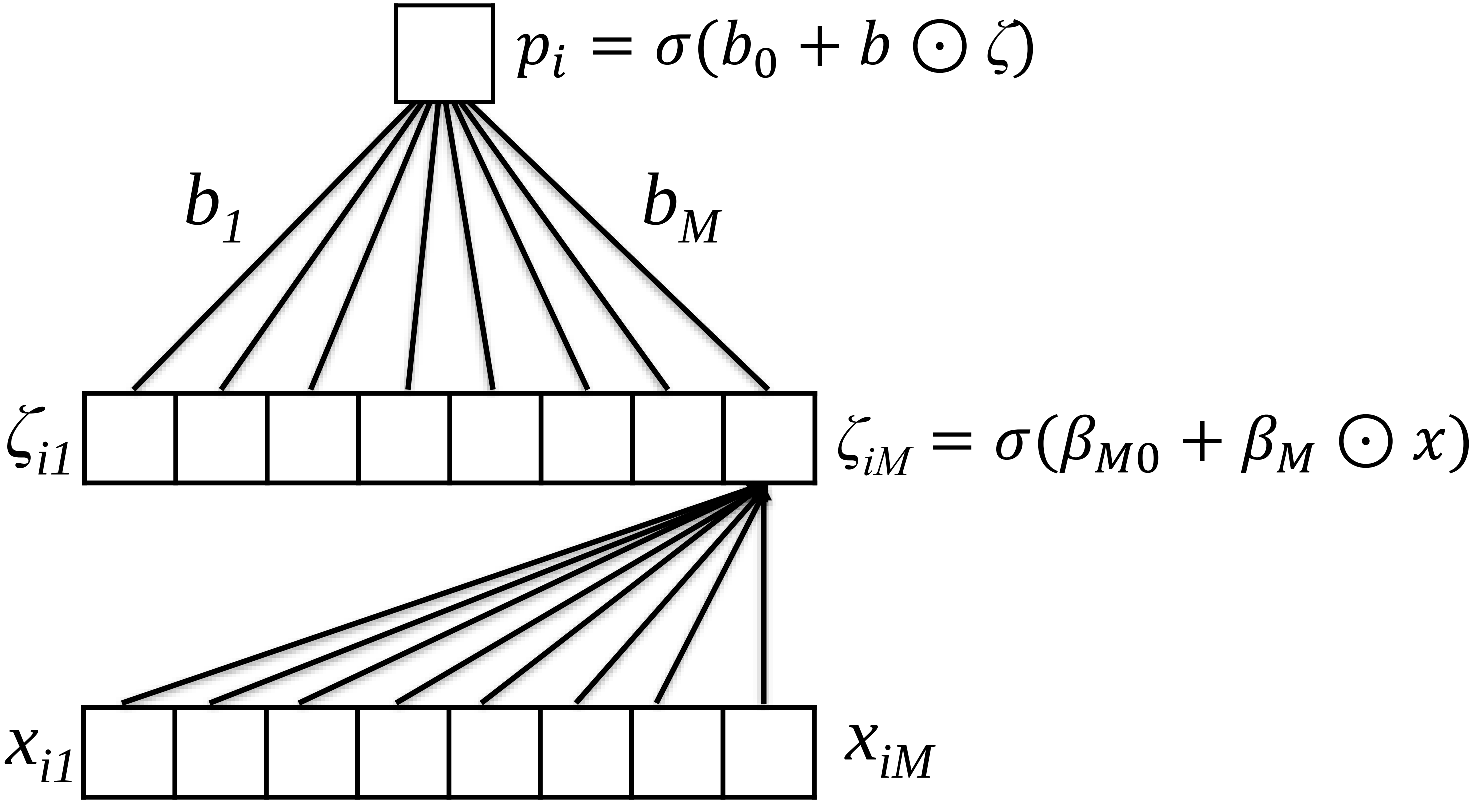
How can we modify logistic regression to learn complex, nonlinear relationships?



$$p_i = \sigma(b_0 + b \odot x)$$

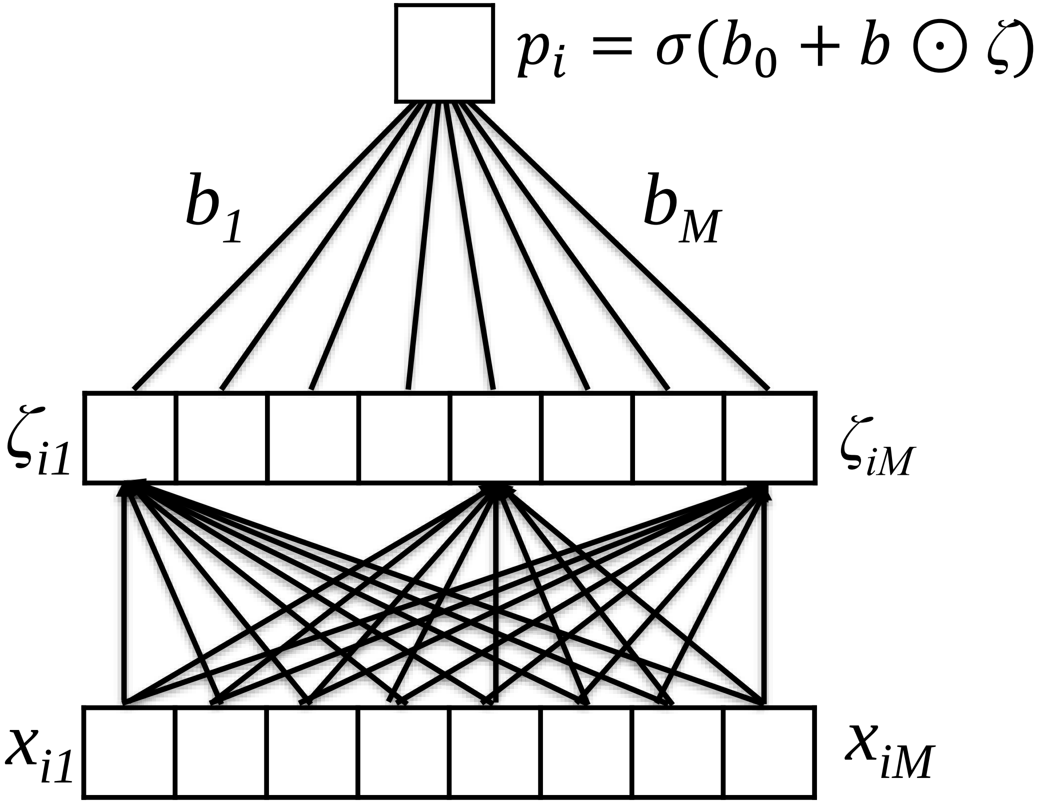


- Instead of predicting p_i directly from our feature vector x, introduce a vector of “latent” features ζ (zeta) that we will use to predict p_i


$p_i = \sigma(b_0 + b \odot \zeta)$
b_1
b_M
ζ_{i1}
$\zeta_{iM} = \sigma(\beta_{M0} + \beta_M \odot x)$
x_{i1}
x_{iM}

- Instead of predicting p_i directly from our feature vector x, introduce a vector of “latent” features ζ (zeta) that we will use to predict p_i

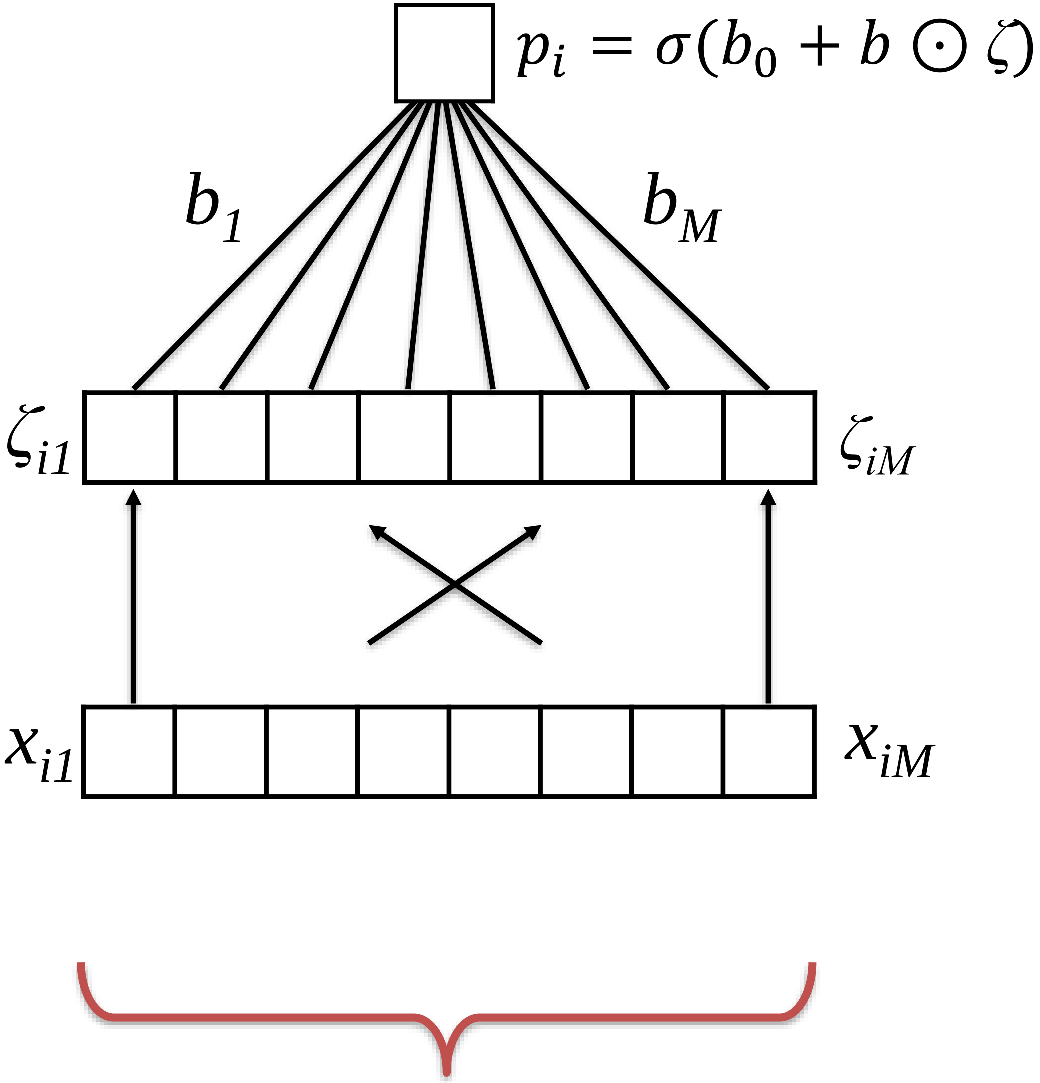
- Individual elements of ζ will themselves be the output of a logistic-regression-like model based on x



- Instead of predicting p_i directly from our feature vector x, introduce a vector of “latent” features ζ (zeta) that we will use to predict p_i

- Individual elements of ζ will themselves be the output of a logistic-regression-like model based on x

- Since this is true for all elements of ζ, *x* and ζ are said to be “fully connected”

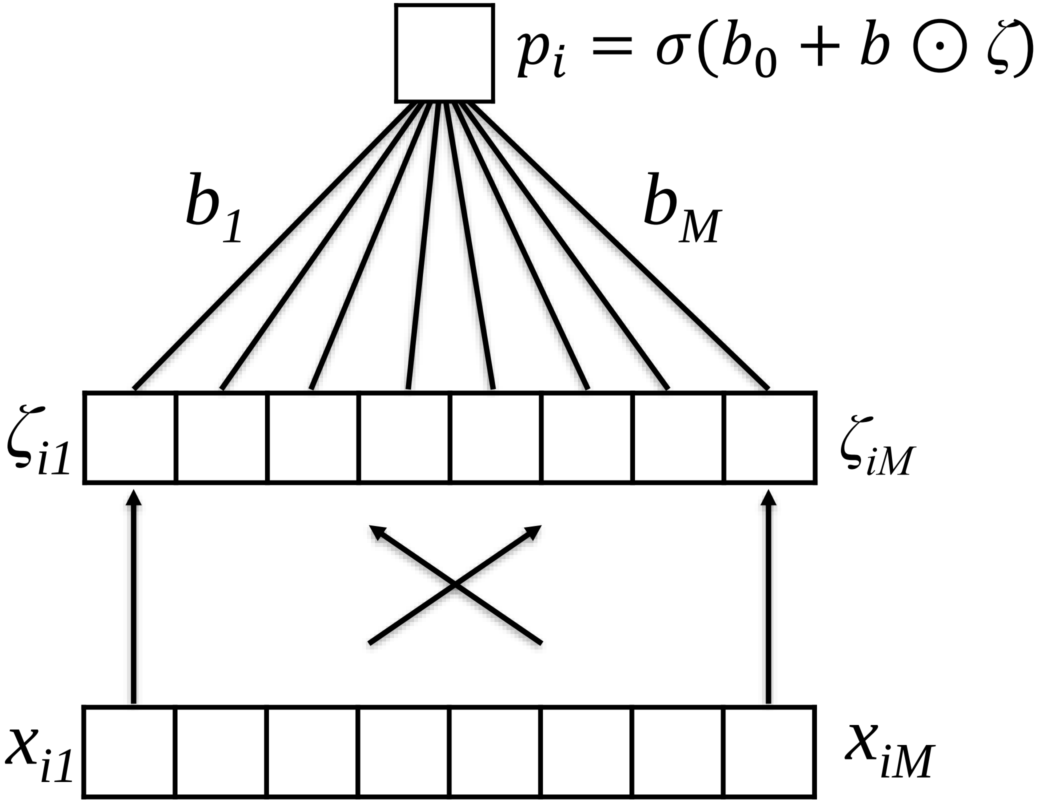


Simplified notation for fully connected layers

- Instead of predicting p_i directly from our feature vector x, introduce a vector of “latent” features ζ (zeta) that we will use to predict p_i

- Individual elements of ζ will themselves be the output of a logistic-regression-like model based on x

- Since this is true for all elements of ζ, x and ζ are said to be “fully connected”

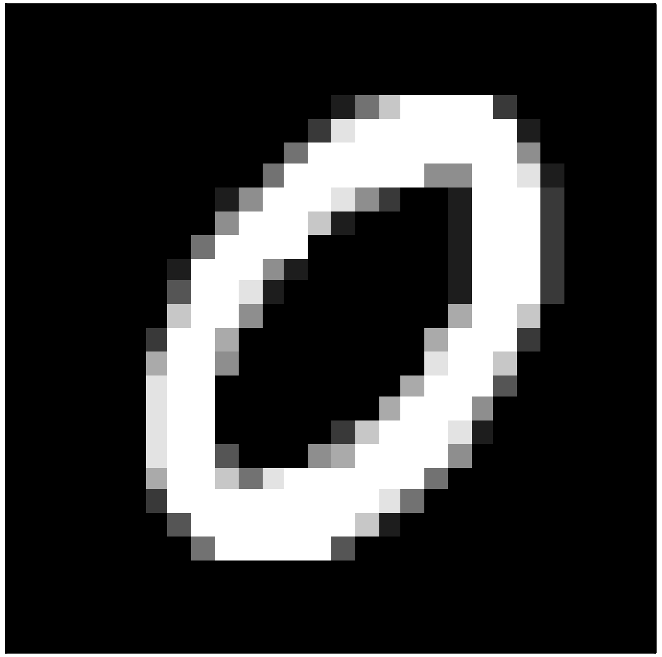


Since they are neither an input nor an output, the features ζ are said to be a “hidden” layer

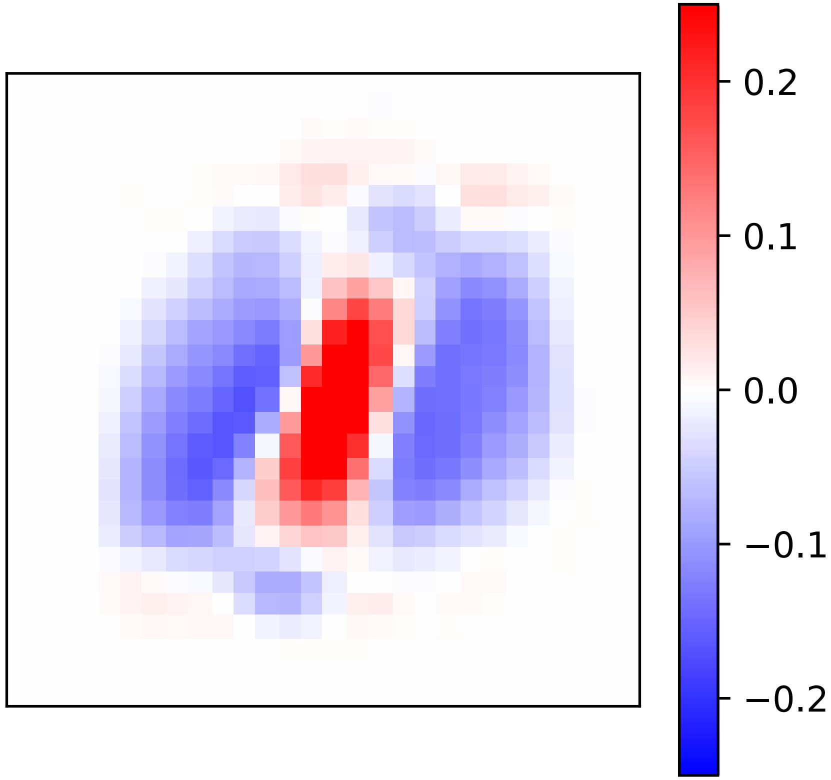
- Instead of predicting p_i directly from our feature vector x, introduce a vector of “latent” features ζ (zeta) that we will use to predict p_i
- Individual elements of ζ will themselves be the output of a logistic-regression-like model based on x
- Since this is true for all elements of ζ, x and ζ are said to be “fully connected”

Why Limit Ourselves to Only One Filter?

$\sigma($

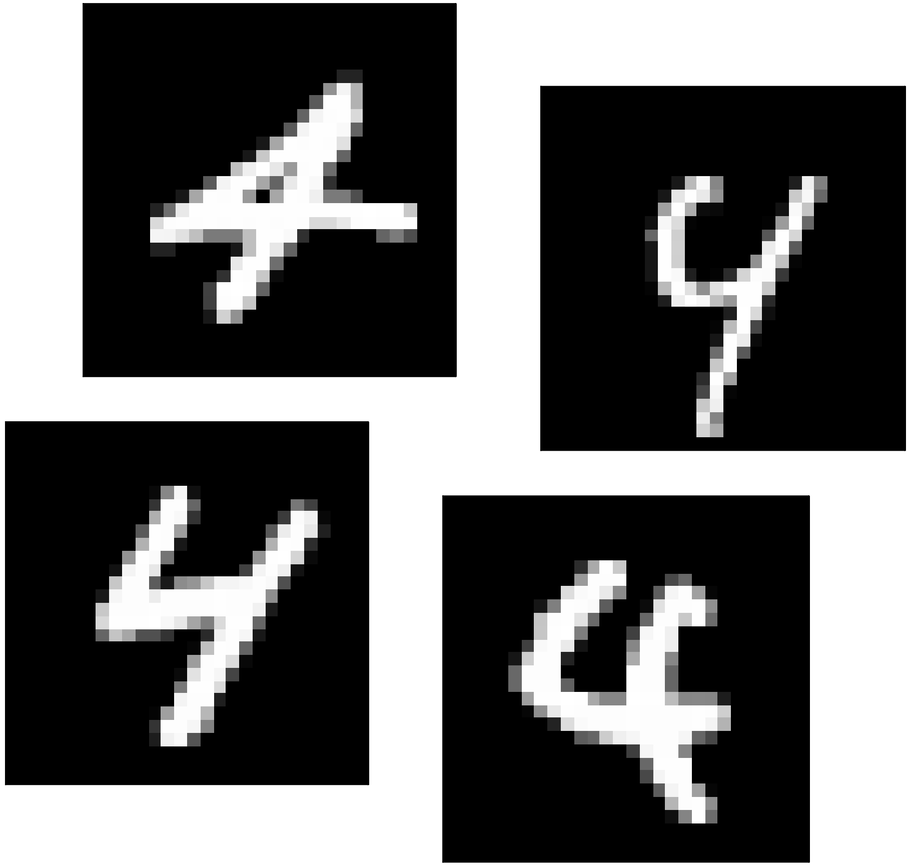


$\times$

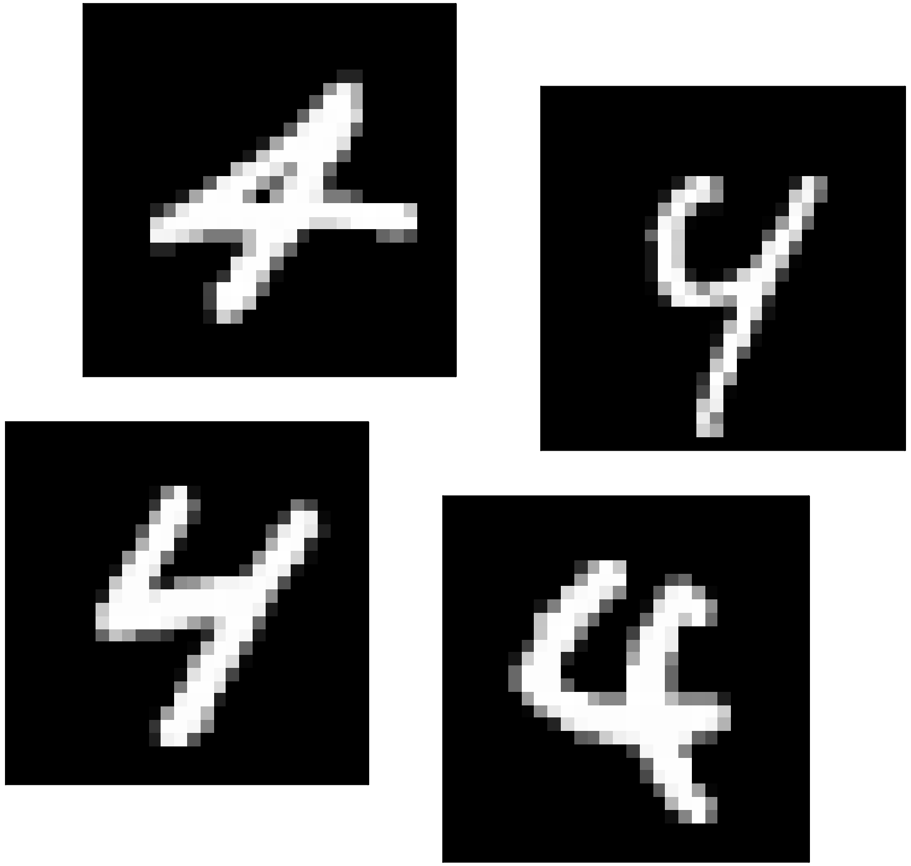


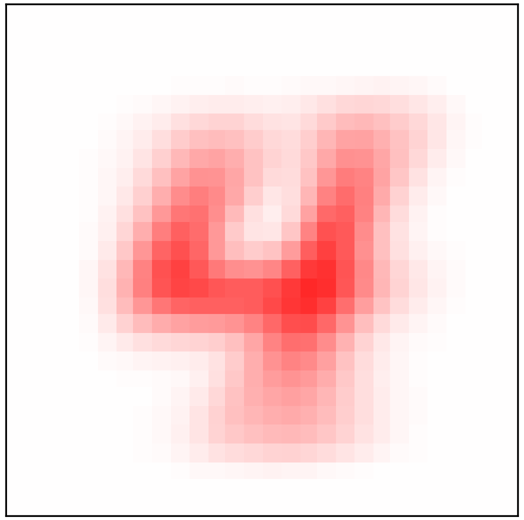
$)$

Return to MNIST:
Many ways of writing “4”



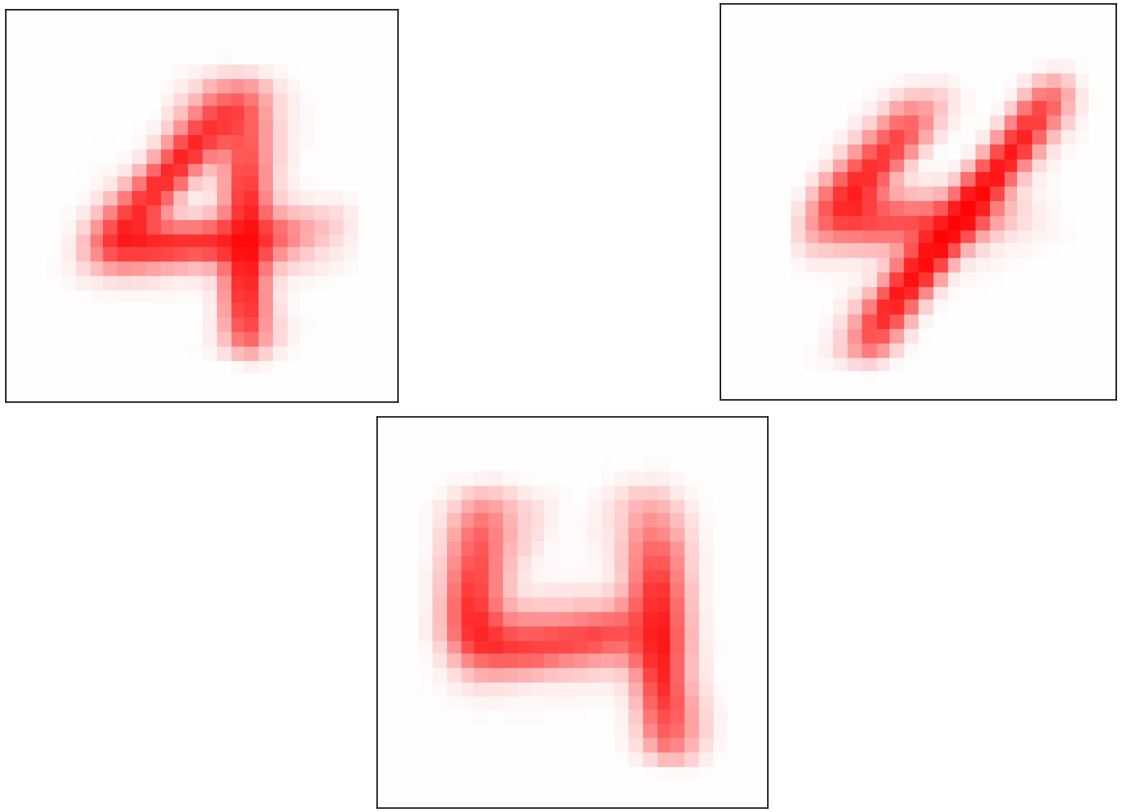
Return to MNIST:
Many ways of writing “4”



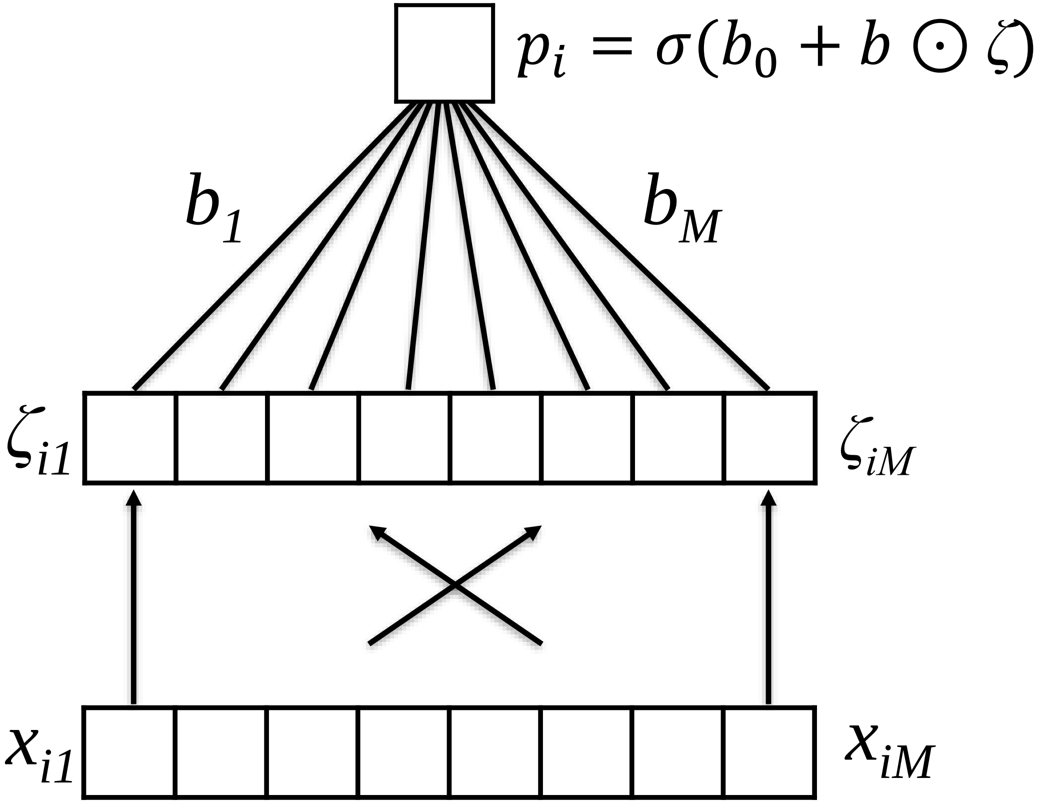


Single Filter (e.g. Logistic Regression/ “Shallow Learning”) only uses one filter, looks for the average shape

Return to MNIST:
Many ways of writing “4”

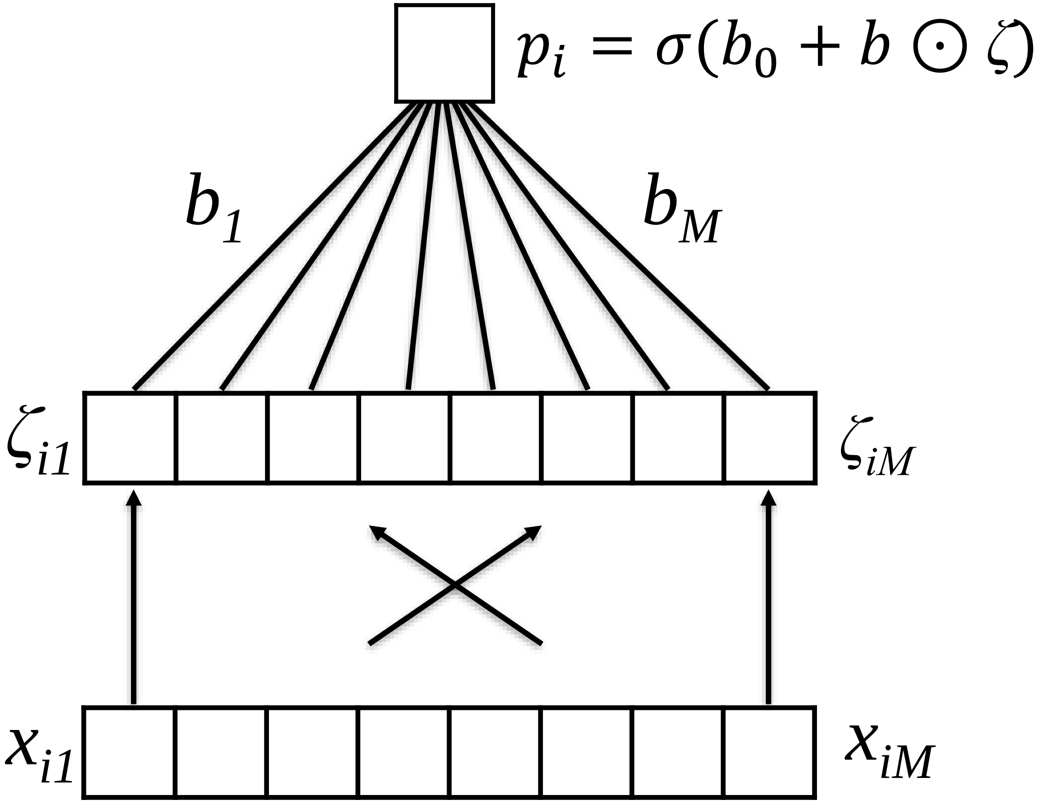


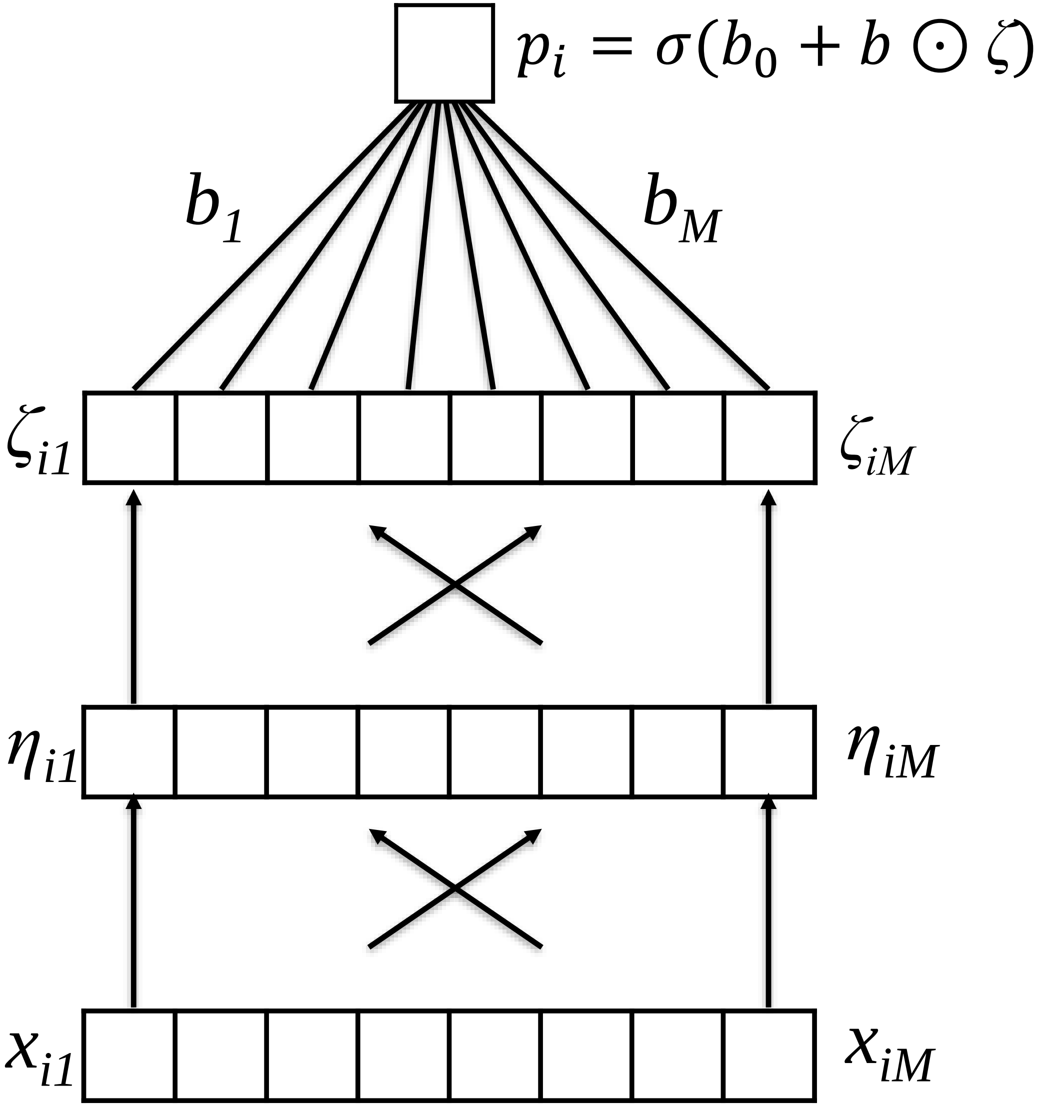
Multiple filters can look for *subtypes* indicative of different ways of writing “4”



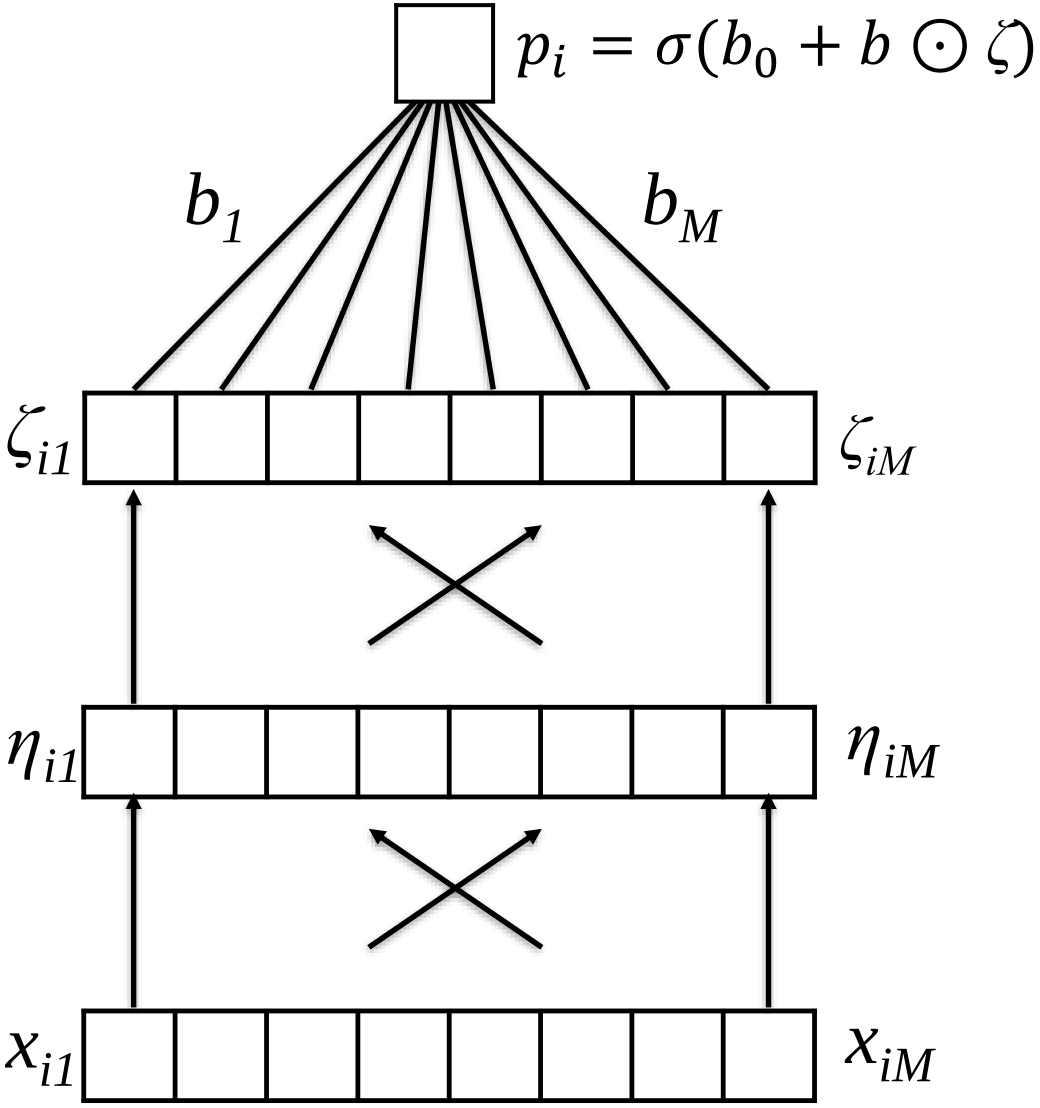
- Each element of ζ_i can be viewed as the output of a single filter applied to x_i
- We then perform logistic regression on the vector of these filter outputs

Extended Logistic Regression





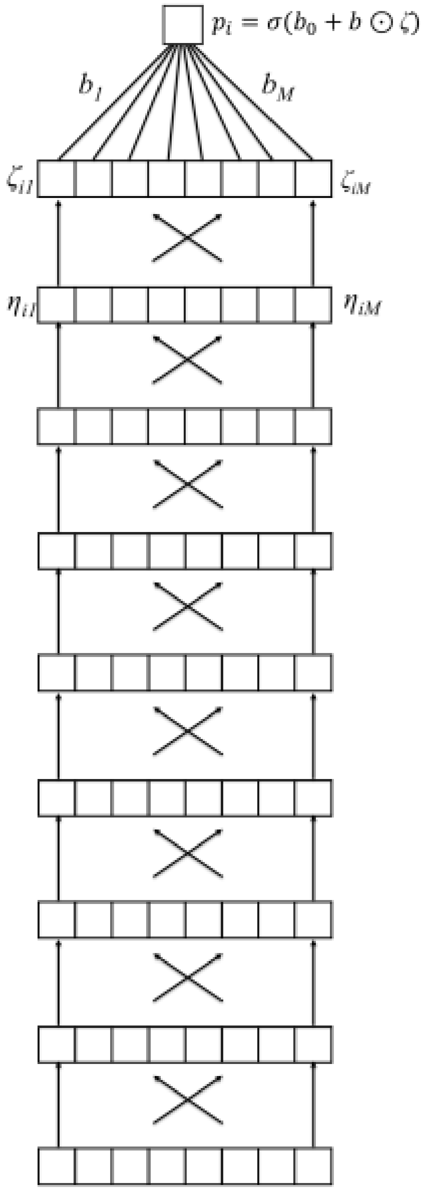
By adding layers, we build a hierarchy of features



Multilayer Perceptron

(i.e. neural network)

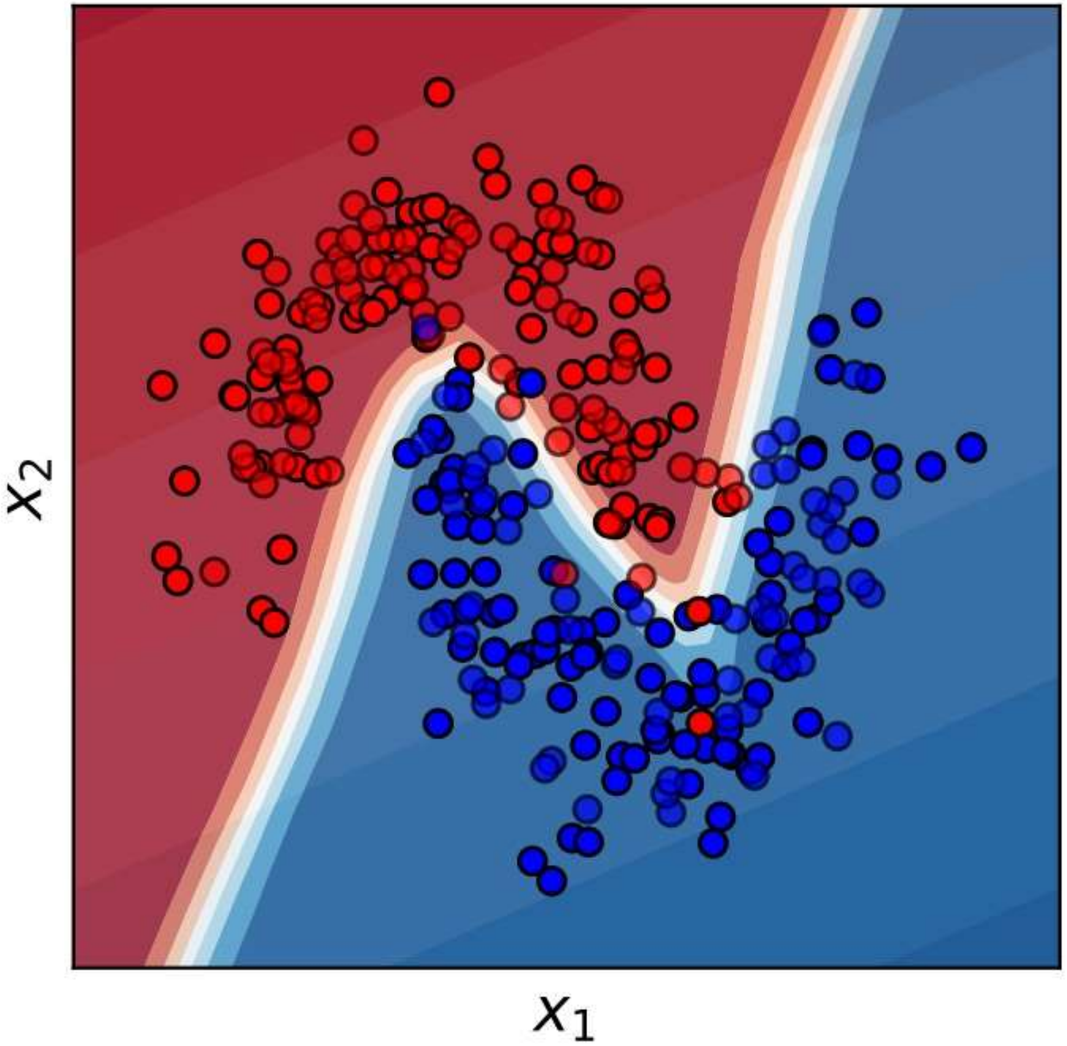
with 2 hidden layers



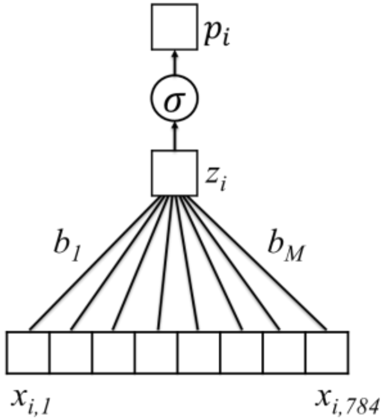
Deep Learning:

many hidden layers

Learn Highly Non-Linear Classification Surfaces

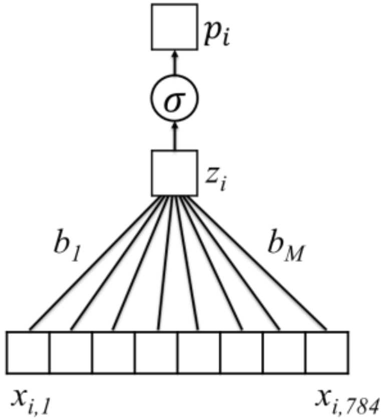


Does this work with MNIST?

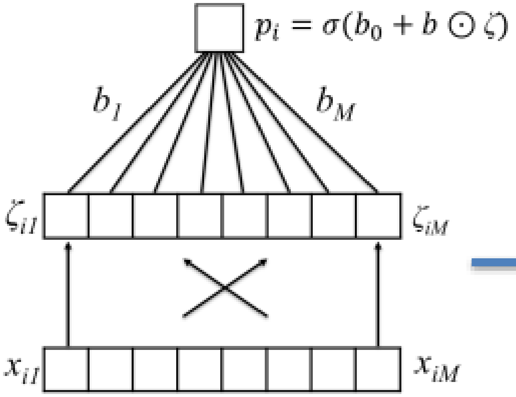


~91% Accurate

Does this work with MNIST?



~91% Accurate



~96% Accurate