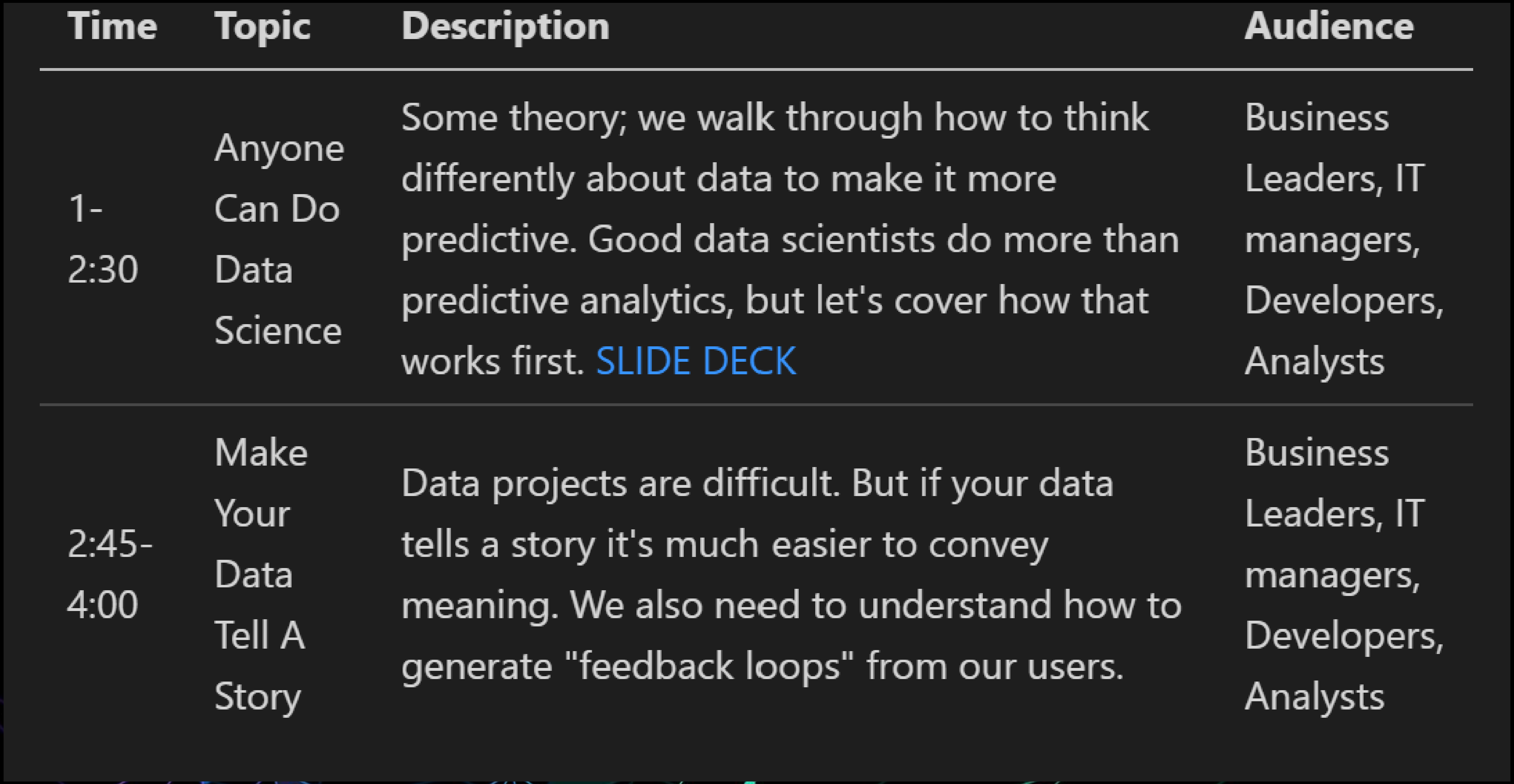
What You Need to Know About the Technology, Science, Process, and Practice of Data Science

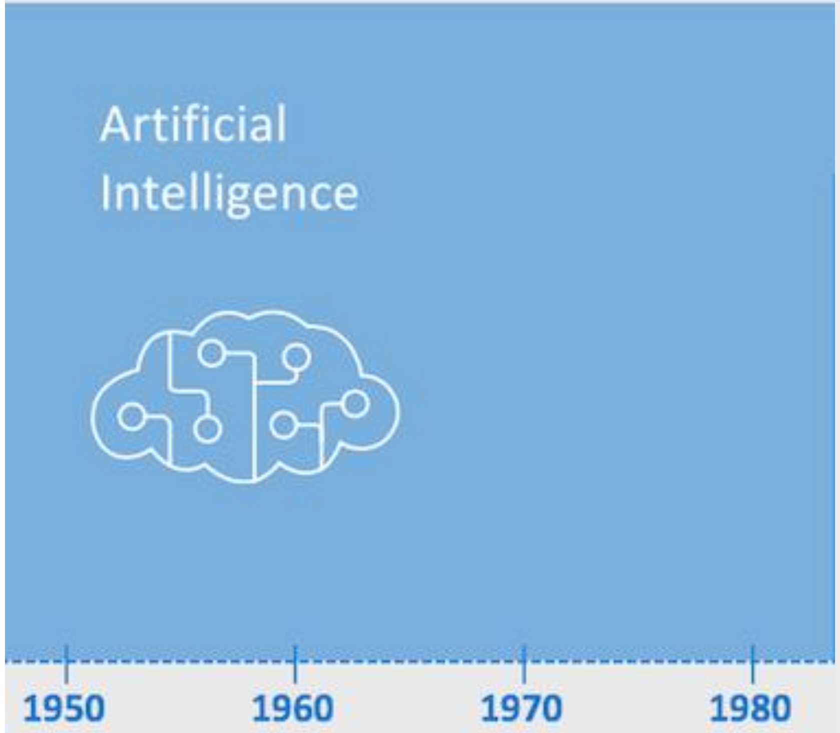


Time	Topic	Description	Audience
1-2:30	Anyone Can Do Data Science	Some theory; we walk through how to think differently about data to make it more predictive. Good data scientists do more than predictive analytics, but let's cover how that works first. SLIDE DECK	Business Leaders, IT managers, Developers, Analysts
2:45-4:00	Make Your Data Tell A Story	Data projects are difficult. But if your data tells a story it's much easier to convey meaning. We also need to understand how to generate "feedback loops" from our users.	Business Leaders, IT managers, Developers, Analysts

Why AI now?

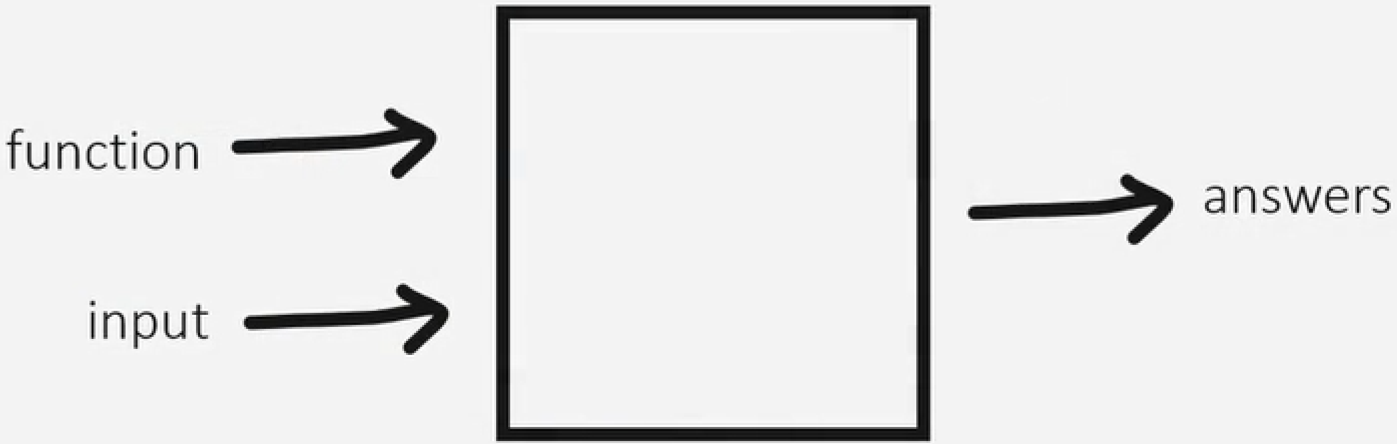
- More data
- More compute
- Innovation in algorithms, tools, and frameworks

AI, Machine Learning and Deep Learning

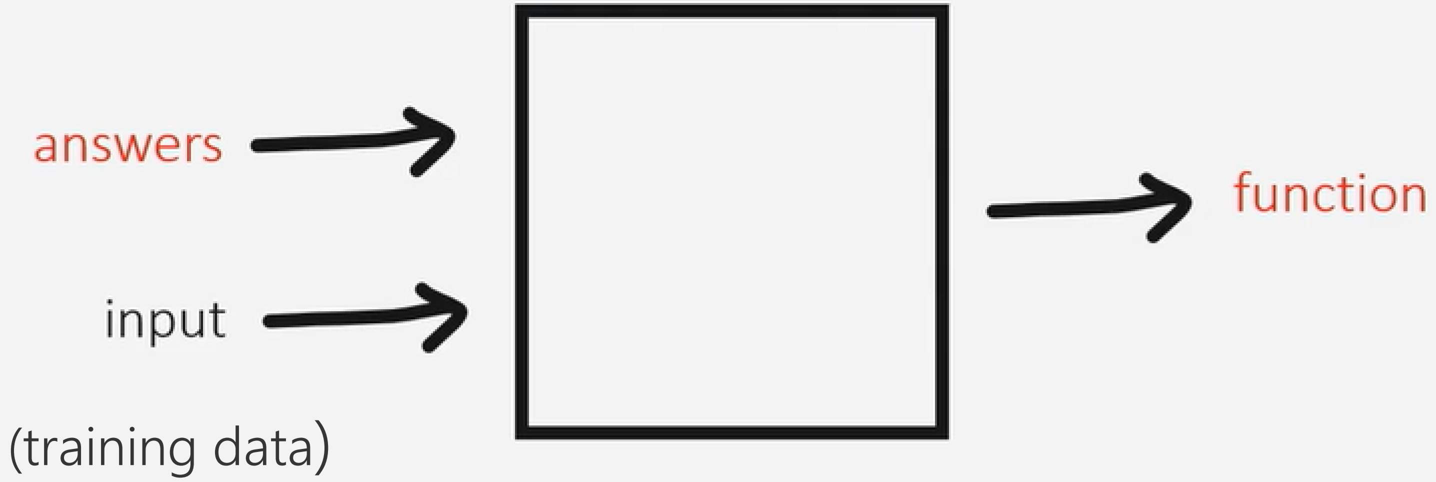


Artificial Intelligence

programming



machine learning



answers

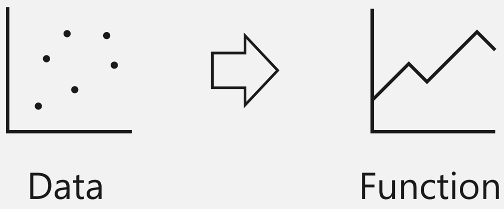
input

(training data)

function

Building models

What is a model?



A model is a function, with its parameters learned from data

How is it created

Machine learning is using a variety of algorithms and techniques to learn the right parameters

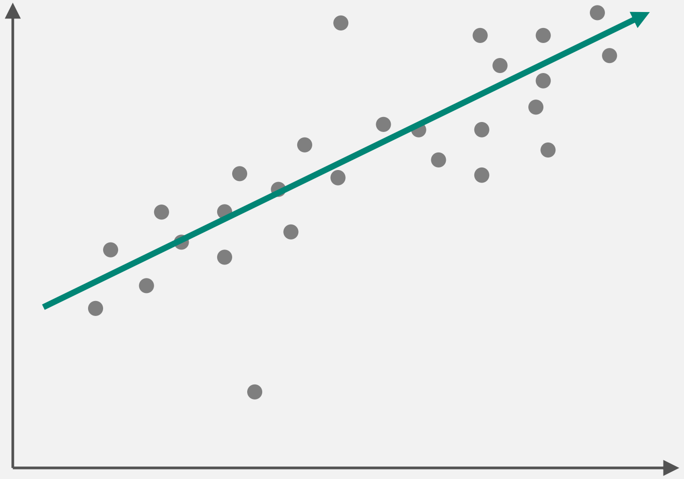
What languages are used

python™

Majority of ML is done in Python and R, using frameworks like scikit-learn

What does a model do?

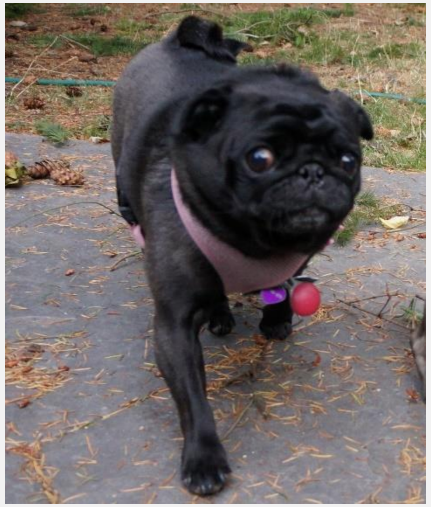
Regression



$F(x) = mx + b$

Learn m & b from the data

Classification



Dog .78

Cat .04

Tiger .001

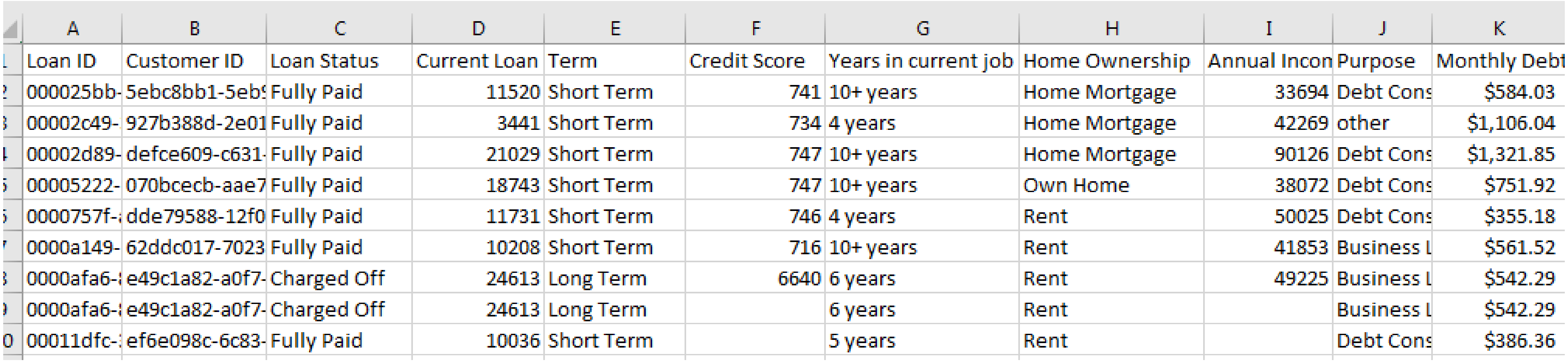
Clustering

- Segmenting customers
- Arranging articles into categories
- Discovering similar items

Questions so far?

Understanding the “Shape” of Data
Thinking Like a Data Scientist

Will this loan be charged off?



	A	B	C	D	E	F	G	H	I	J	K
1	Loan ID	Customer ID	Loan Status	Current Loan	Term	Credit Score	Years in current job	Home Ownership	Annual Incon	Purpose	Monthly Debt
2	000025bb-	5ebc8bb1-5eb9	Fully Paid	11520	Short Term	741	10+ years	Home Mortgage	33694	Debt Cons	$584.03
3	00002c49-	927b388d-2e01	Fully Paid	3441	Short Term	734	4 years	Home Mortgage	42269	other	$1,106.04
4	00002d89-	defce609-c631-	Fully Paid	21029	Short Term	747	10+ years	Home Mortgage	90126	Debt Cons	$1,321.85
5	00005222-	070bcecb-aae7	Fully Paid	18743	Short Term	747	10+ years	Own Home	38072	Debt Cons	$751.92
6	0000757f-	dde79588-12f0	Fully Paid	11731	Short Term	746	4 years	Rent	50025	Debt Cons	$355.18
7	0000a149-	62ddc017-7023	Fully Paid	10208	Short Term	716	10+ years	Rent	41853	Business L	$561.52
8	0000afa6-	e49c1a82-a0f7-	Charged Off	24613	Long Term	6640	6 years	Rent	49225	Business L	$542.29
9	0000afa6-	e49c1a82-a0f7-	Charged Off	24613	Long Term		6 years	Rent		Business L	$542.29
10	00011dfc-	ef6e098c-6c83-	Fully Paid	10036	Short Term		5 years	Rent		Debt Cons	$386.36

Terminology

Training Data : A set of samples (table of data)

Testing Data: A set of samples (training data) set aside to test your model

Features: Individual columns in our data set. These might be used to help make our prediction, or not.

Factors: aka features

Categorical Features: features with a known domain of values

independent variables: aka features

Feature Engineering: manipulating existing data to make it more meaningful

very similar to ETL

Data Wrangling/Munging: ETL

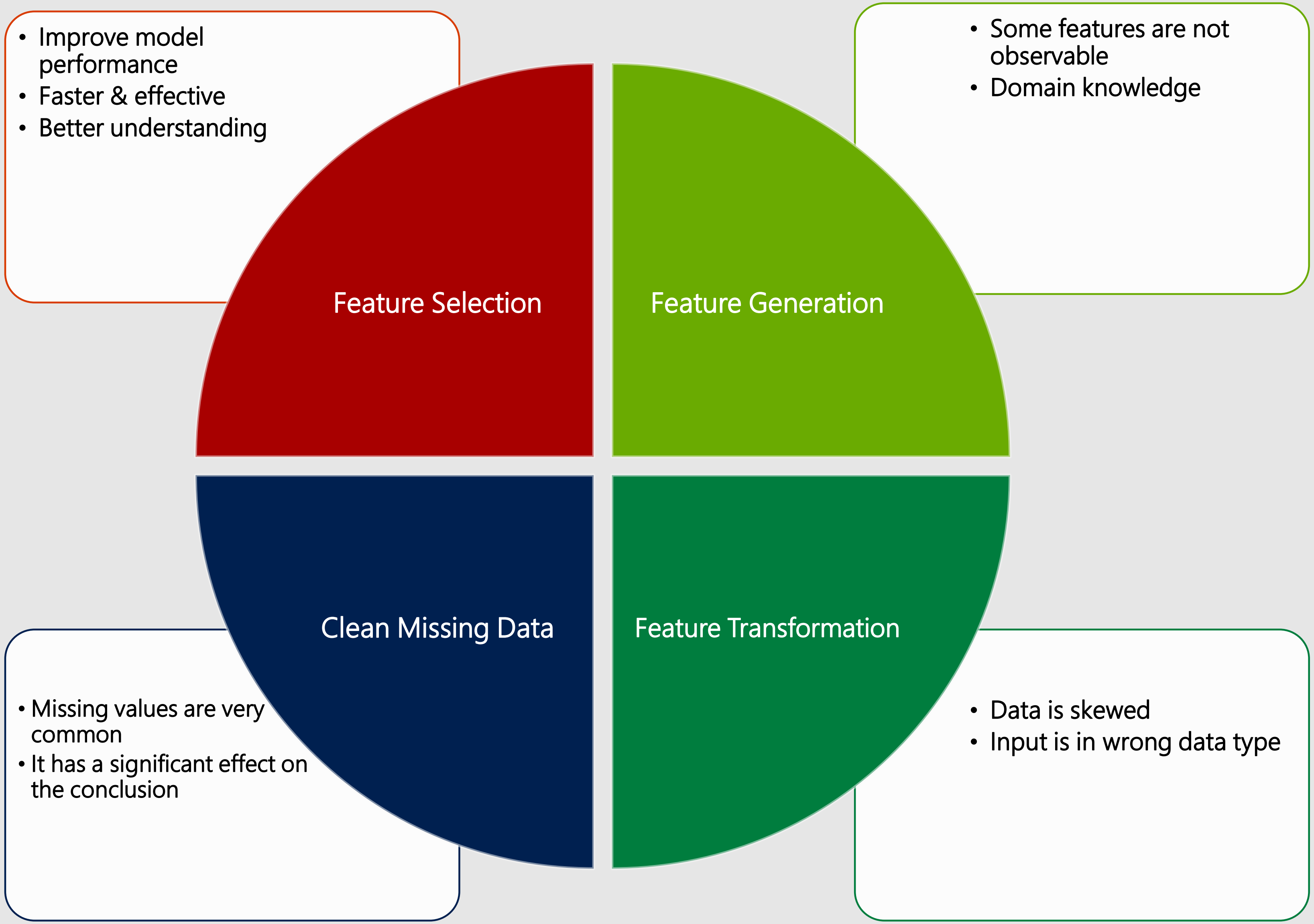
Data Dredging: make the data fit the hypothesis (don't do this)

Label: Historical outcome or result related to a set of samples.

What you are trying to predict.

aka, "the target" or "the dependent variable", or "response"

Feature Engineering Tasks



Feature Selection

- Improve model performance
- Faster & effective
- Better understanding

Feature Generation

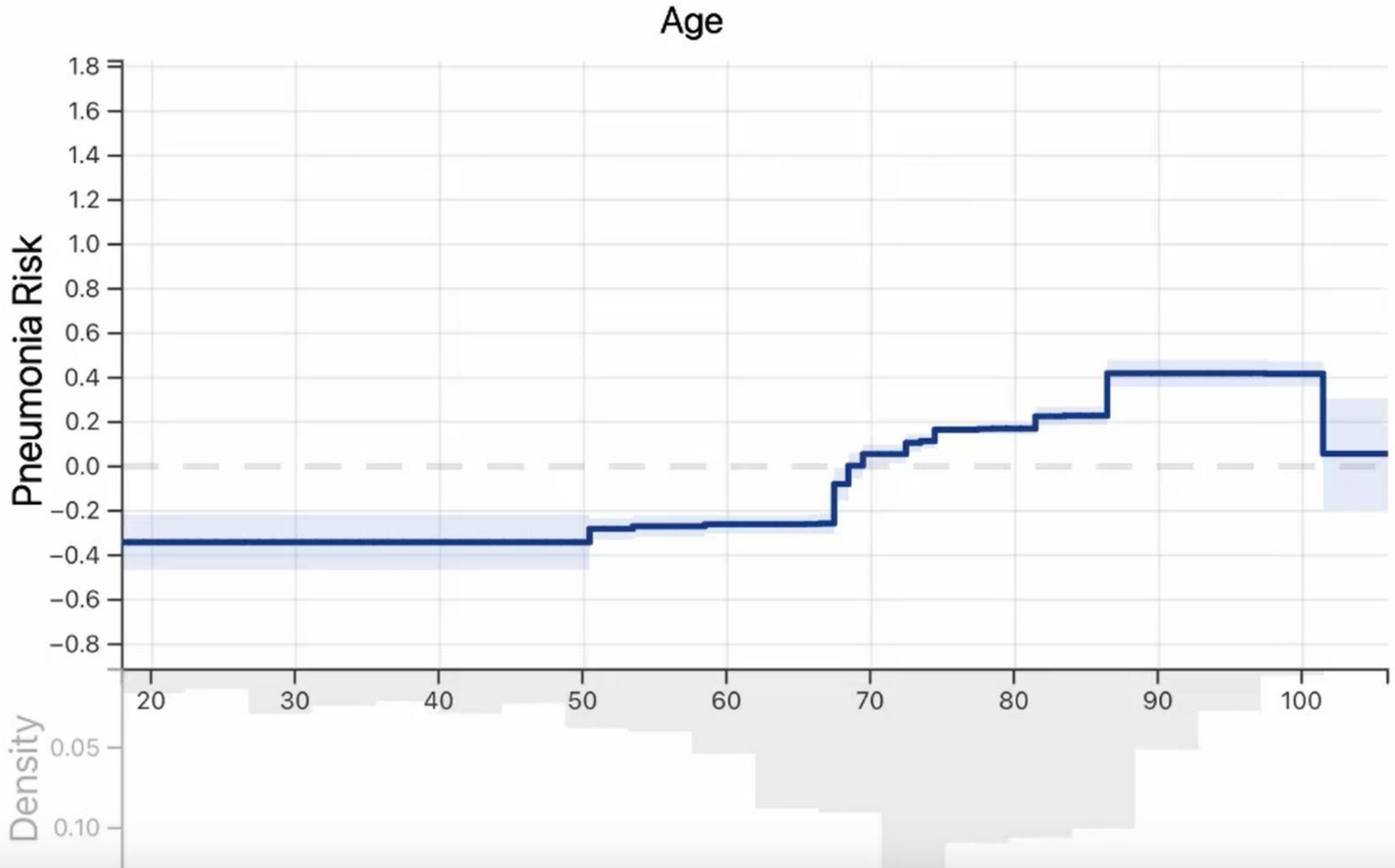
- Some features are not observable
- Domain knowledge

Clean Missing Data

- Missing values are very common
- It has a significant effect on the conclusion

Feature Transformation

- Data is skewed
- Input is in wrong data type


Age
Pneumonia Risk
1.8
1.6
1.4
1.2
1.0
0.8
0.6
0.4
0.2
0.0
−0.2
−0.4
−0.6
−0.8
20
30
40
50
60
70
80
90
100
Density
0.05
0.10

Terminology

Learner: Machine learning algorithm

Supervised Learning: we know the label ("will this loan chargeoff?")

you "train" these models with training and test data

Unsupervised Learning: we don't know the label ("Recommendation engines")

the learner tries to find patterns

ex: anomaly detection, clustering

Parameter / Hyper-parameter

Correlation vs Causation: ML *cannot* tell what caused what

Leakage

you accidentally used the response to predict the response

(you used the answer to predict the answer)

ex: you use GPA to predict if a student will fail

these are often "masked" or "derived" and are hard to find

happens possibly when your "relative influence" gets to high (70%?)

Terminology

Overfitting/Underfitting a model

the more you try to make your model perfect, the more you risk fitting the model to the training data and the model begins to memorize the data it has seen.

If it walks like a duck and talks like a duck, it's a duck

It is a duck if, and only if, it walks and quacks exactly in the ways I have personally observed from ducks in Pennsylvania. Since I've never observed ducks in Australia, that may look, walk and quack differently from ducks in PA…in that case, they aren't ducks at all.

If it walks on two legs and emits shrill, nasal-y, high pitched noises, it's a duck. Therefore, Fran Drescher is a duck.

" Feature engineering is the most important but underrated step of machine learning."

Better features are better than better algorithms…

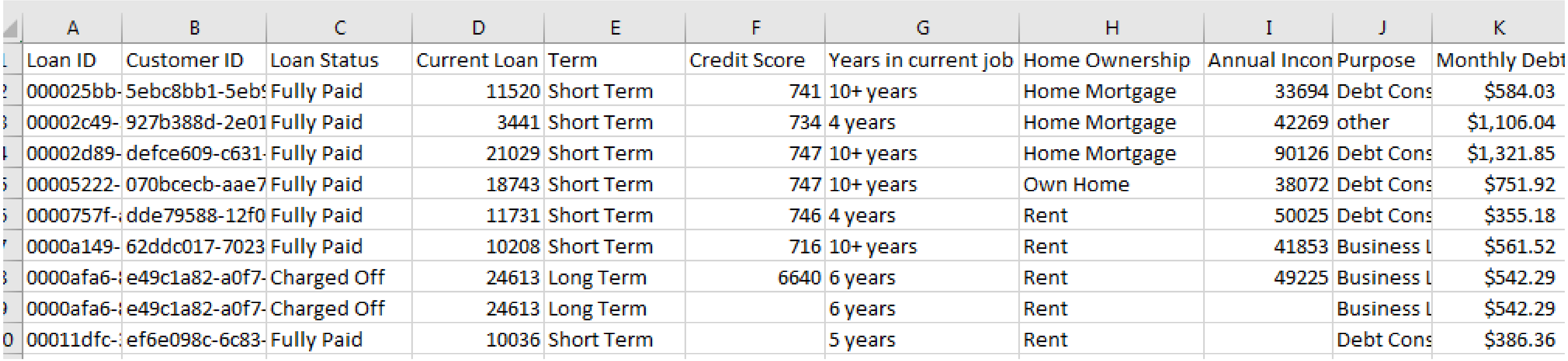
Better features are better than more data…

More data is better than better algorithms…

Questions so far?

Understanding the “Shape” of Data
Thinking Like a Data Scientist

Will this loan be charged off?



	A	B	C	D	E	F	G	H	I	J	K
1	Loan ID	Customer ID	Loan Status	Current Loan	Term	Credit Score	Years in current job	Home Ownership	Annual Incon	Purpose	Monthly Debt
2	000025bb-	5ebc8bb1-5eb!	Fully Paid	11520	Short Term	741	10+ years	Home Mortgage	33694	Debt Cons	$584.03
3	00002c49-	927b388d-2e01	Fully Paid	3441	Short Term	734	4 years	Home Mortgage	42269	other	$1,106.04
4	00002d89-	defce609-c631-	Fully Paid	21029	Short Term	747	10+ years	Home Mortgage	90126	Debt Cons	$1,321.85
5	00005222-	070bcecb-aae7	Fully Paid	18743	Short Term	747	10+ years	Own Home	38072	Debt Cons	$751.92
6	0000757f-;	dde79588-12f0	Fully Paid	11731	Short Term	746	4 years	Rent	50025	Debt Cons	$355.18
7	0000a149-	62ddc017-7023	Fully Paid	10208	Short Term	716	10+ years	Rent	41853	Business L	$561.52
8	0000afa6-!	e49c1a82-a0f7-	Charged Off	24613	Long Term	6640	6 years	Rent	49225	Business L	$542.29
9	0000afa6-!	e49c1a82-a0f7-	Charged Off	24613	Long Term		6 years	Rent		Business L	$542.29
10	00011dfc-:	ef6e098c-6c83-	Fully Paid	10036	Short Term		5 years	Rent		Debt Cons	$386.36

Feature Engineering Best Practices – Handling Continuous Numerical Values

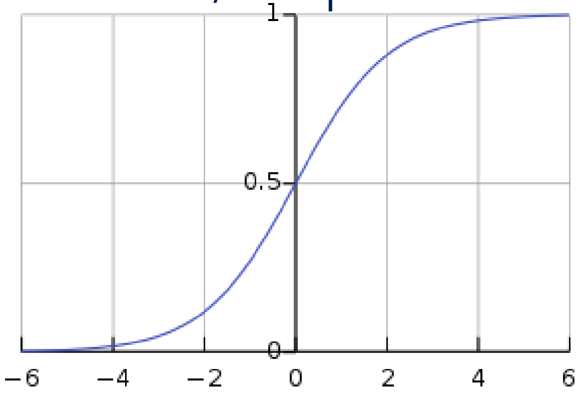
Categorical features are always better

the algorithm isn't wasting time trying to determine if there are meaningful patterns to continuous numeric data

if possible:

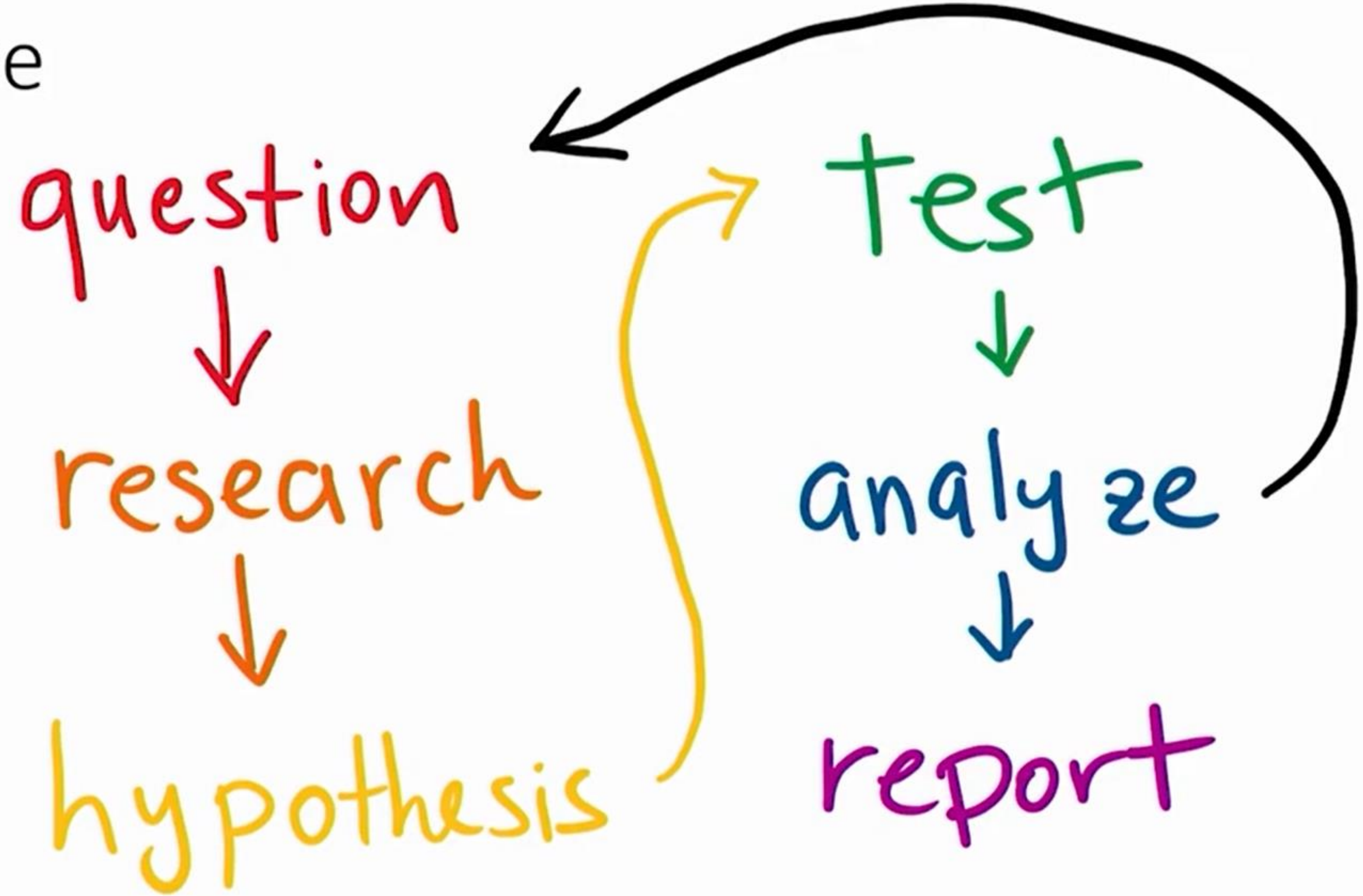
- make a continuous variable a discrete variable
- then use "banding"
- then use one-hot encoding

Regardless, when using a continuous variable, "squash" it with a sigmoid function.

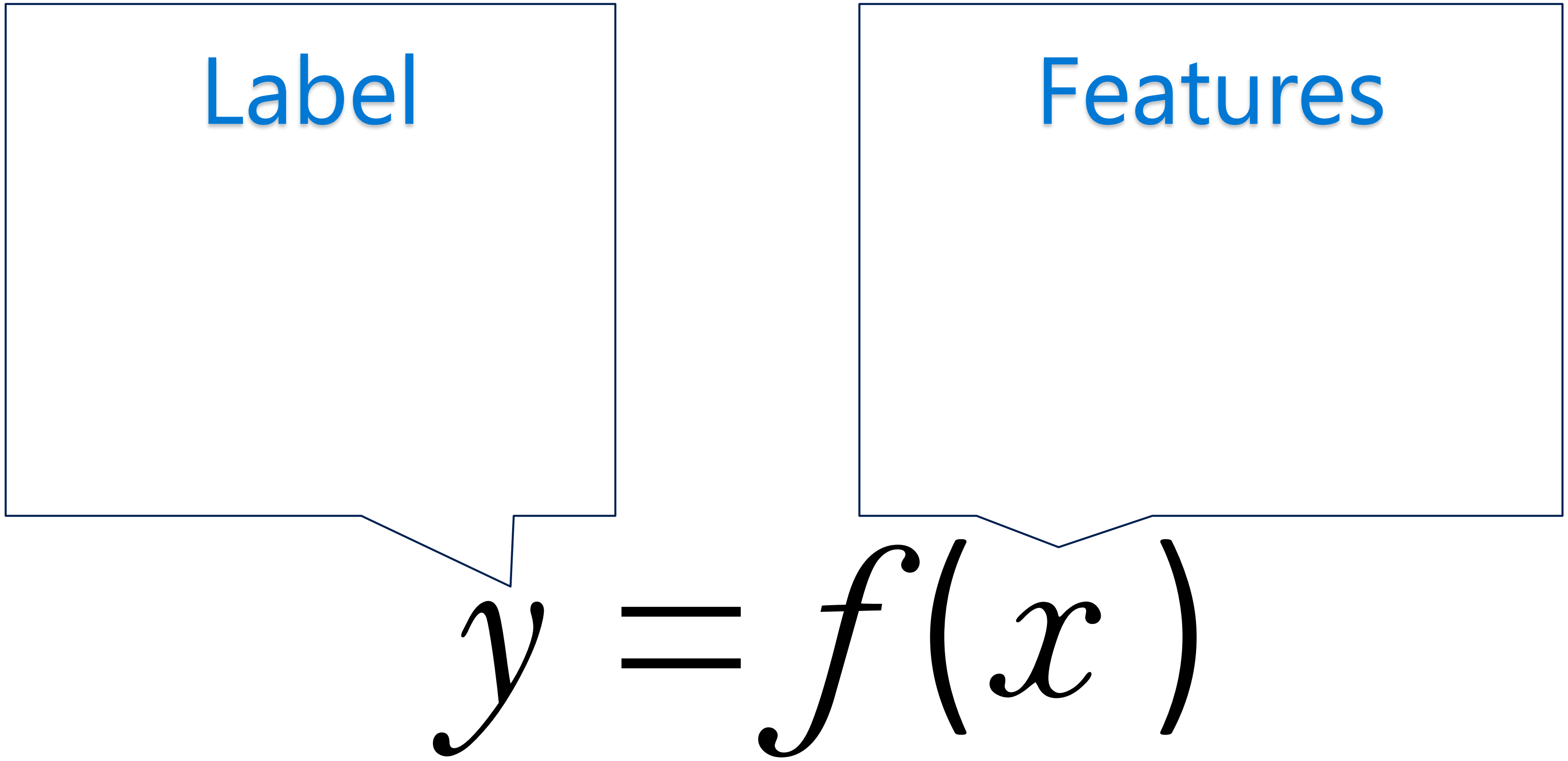


Why are we doing this?

science



Deep Learning, Convolutional Neural Networks, and “AI”



Label

Features

$$y = f(x)$$

Species

0 (“Setosa”)

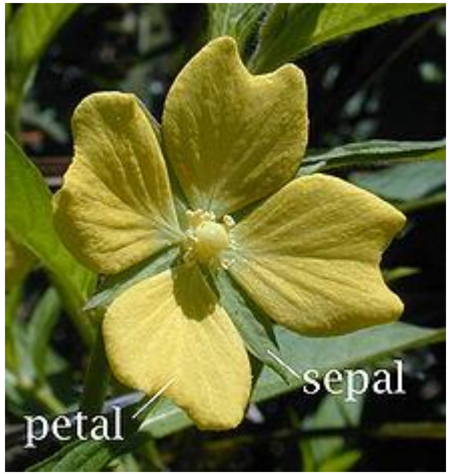
[✓ 1 (“Versicolor”)]

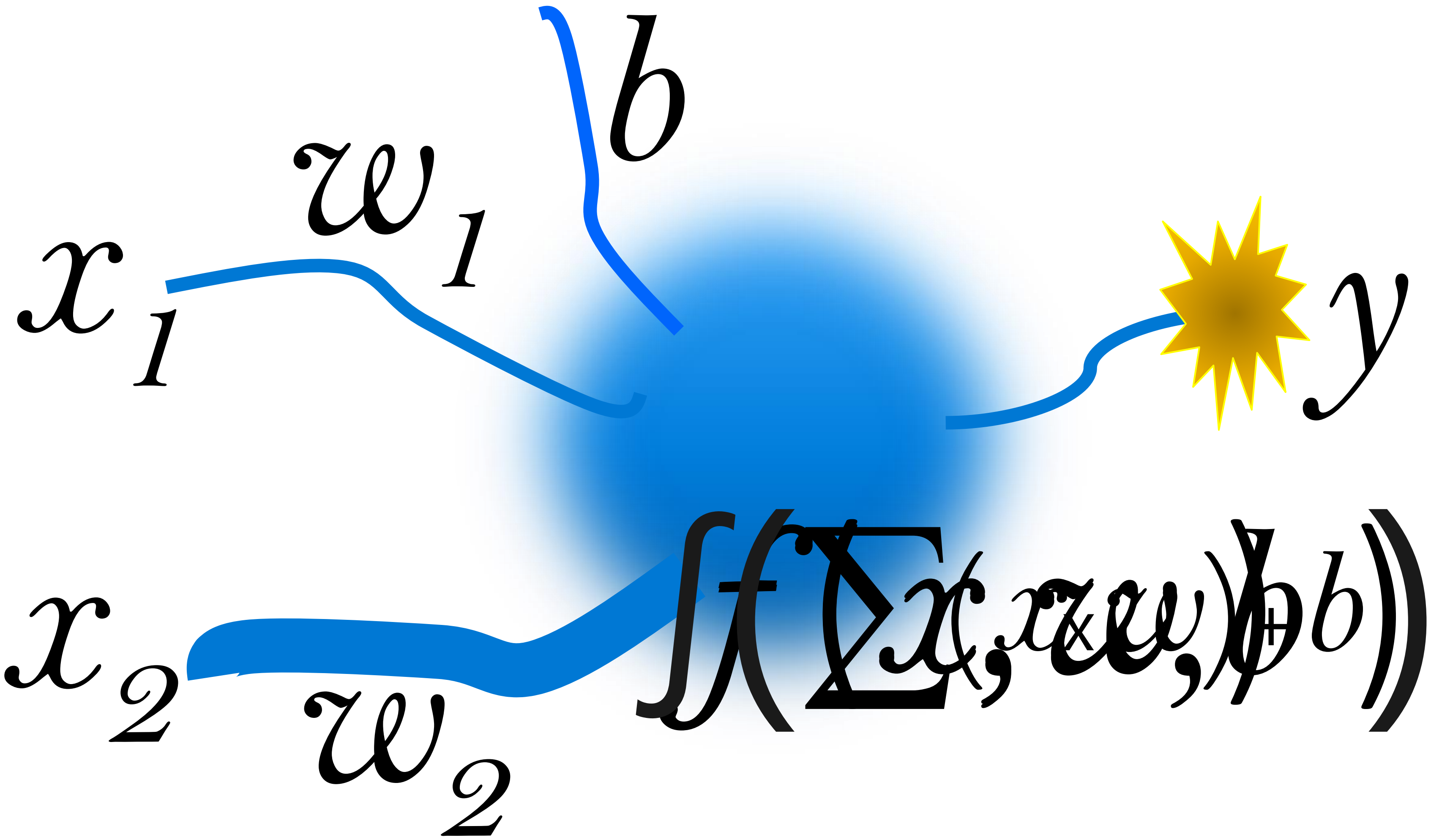
2 (“Virginica”)

Sepal / Petal Measurements

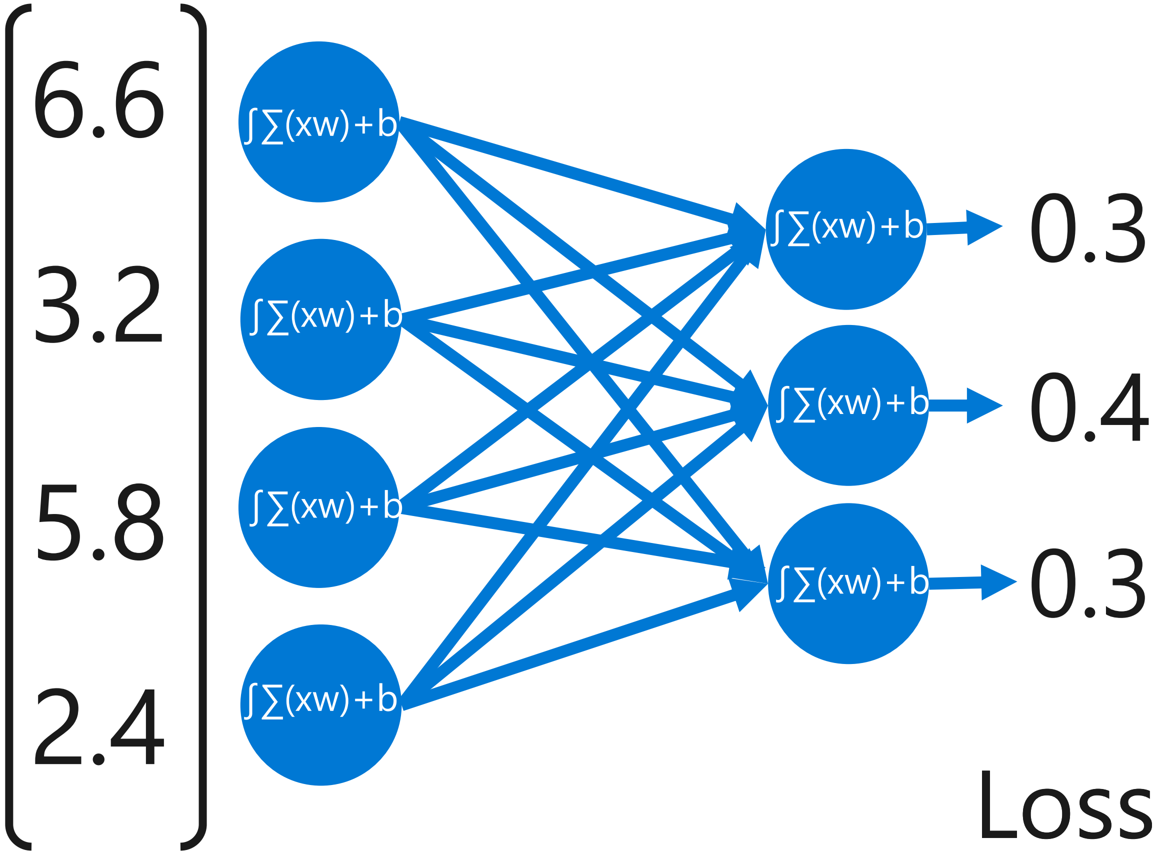
[6.6, 3.2, 5.8, 2.4]

$$y = f(\vec{x}) = [0, 1, 0]$$




b
w_1
x_1
y
x_2
w_2
$f(\sum(x, w) + b)$

Neural Network (Perceptron)

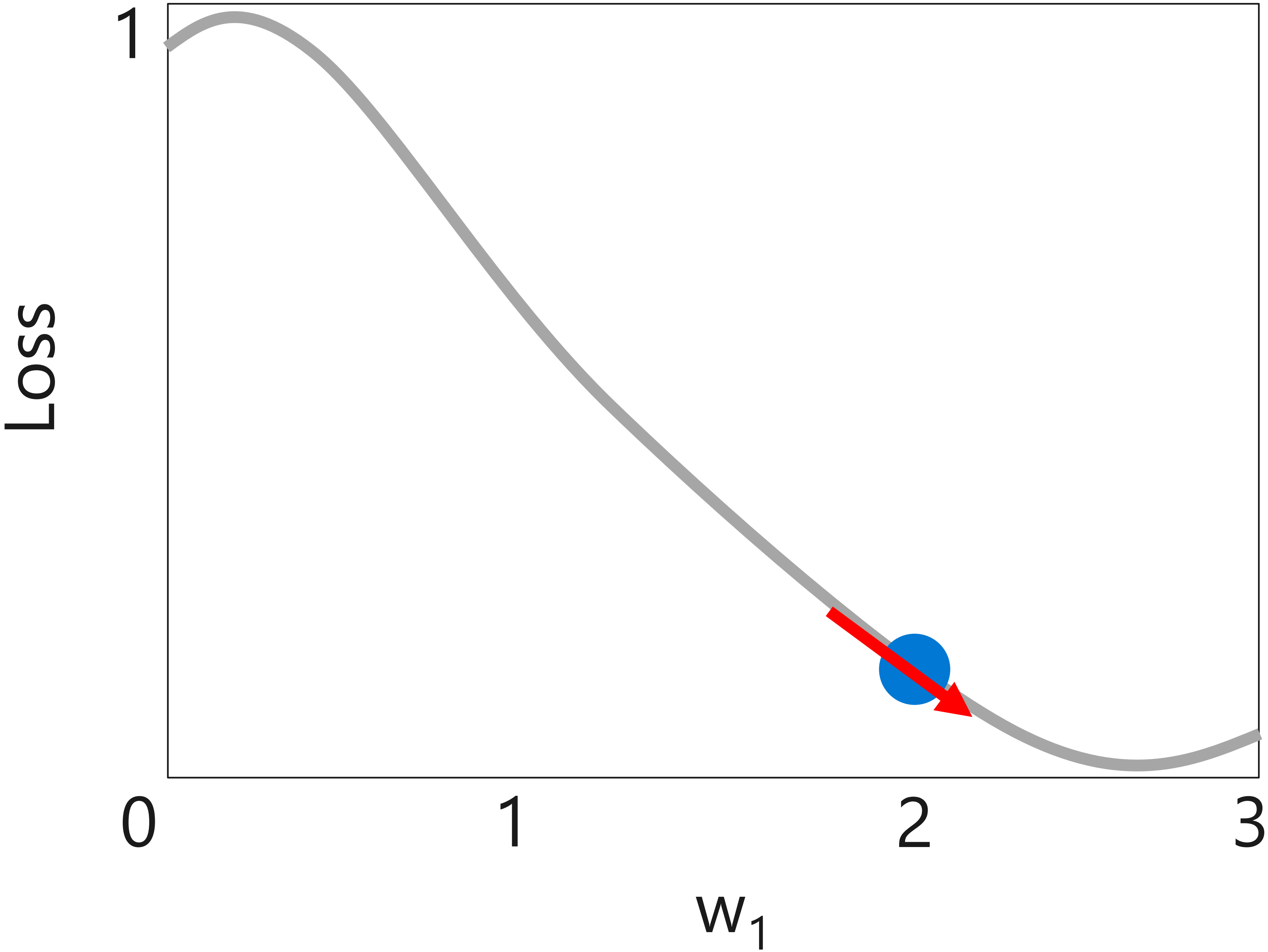


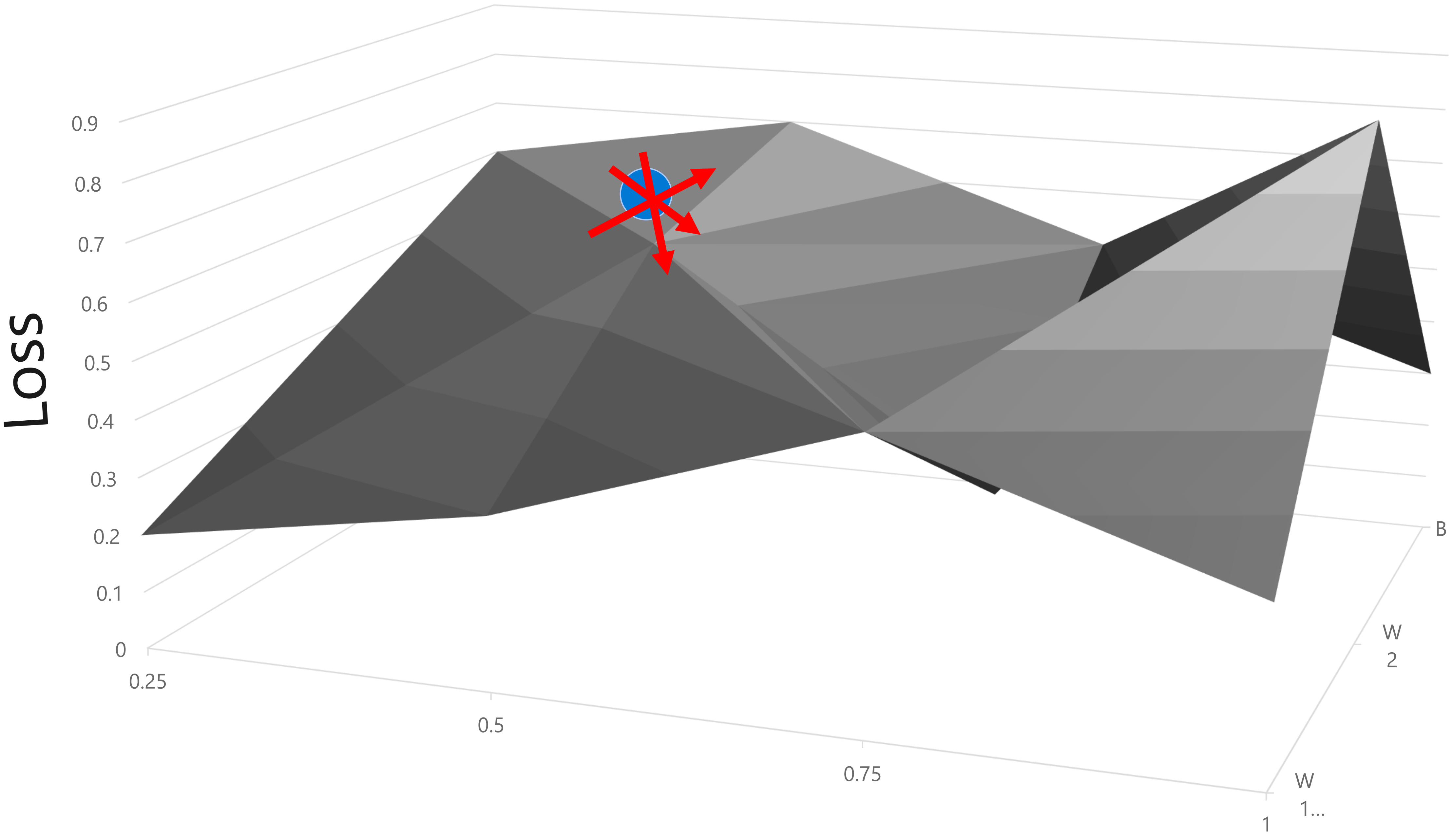
= [0.3, 0.4, 0.3]

[0.0, 1.0, 0.0]

Loss = $\mu(0.3, 0.6, 0.3)^2$

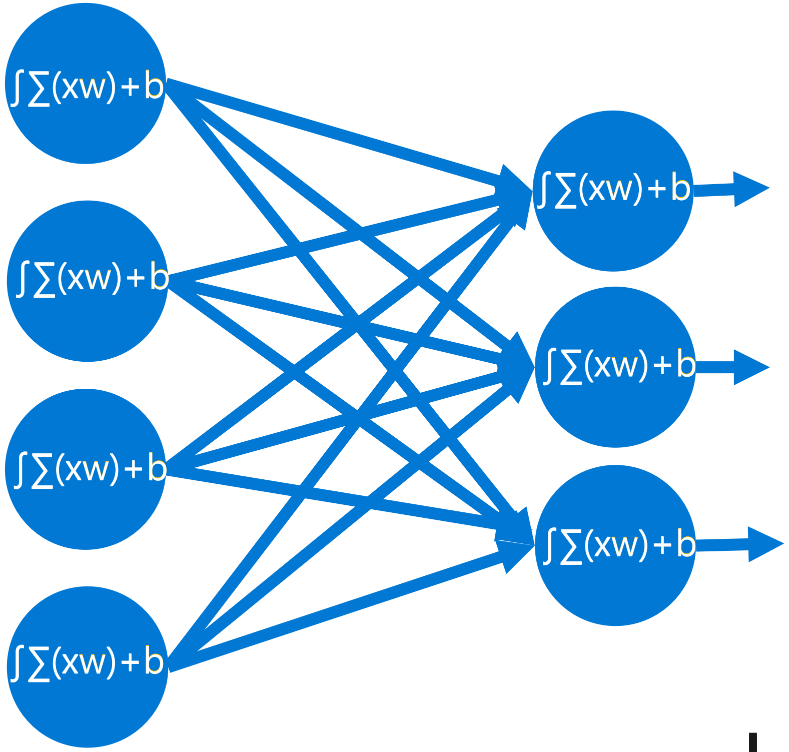
0.16


1
Loss
0
1
2
3
w_1


Loss
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
0.25
0.5
0.75
1
B
W
2
W
1...

Backpropagation

$$\begin{bmatrix} 6.6 \\ 3.2 \\ 5.8 \\ 2.4 \end{bmatrix}$$



0.2

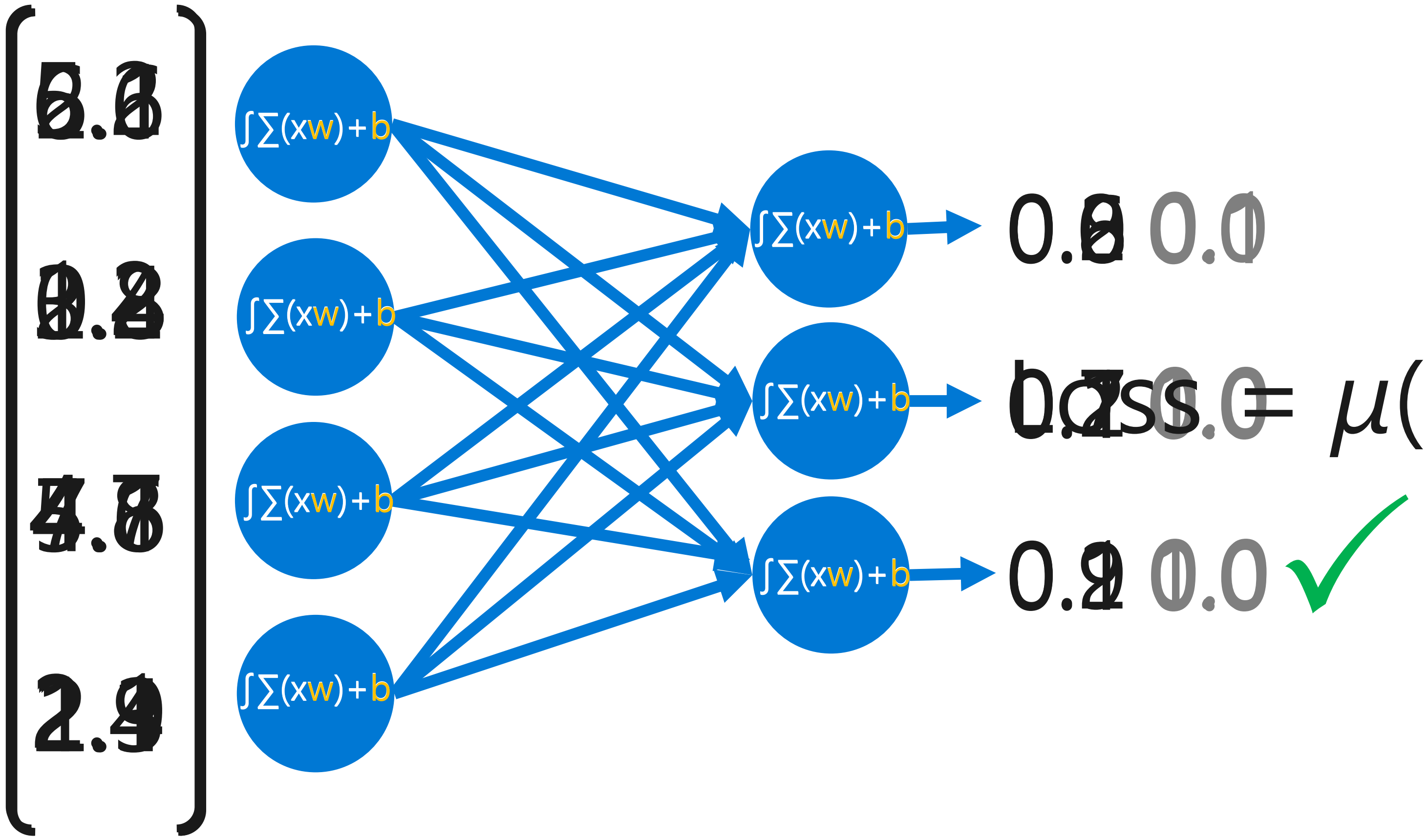
0.7 = [0.2, 0.7, 0.1]

[0.0, 1.0, 0.0]

0.1

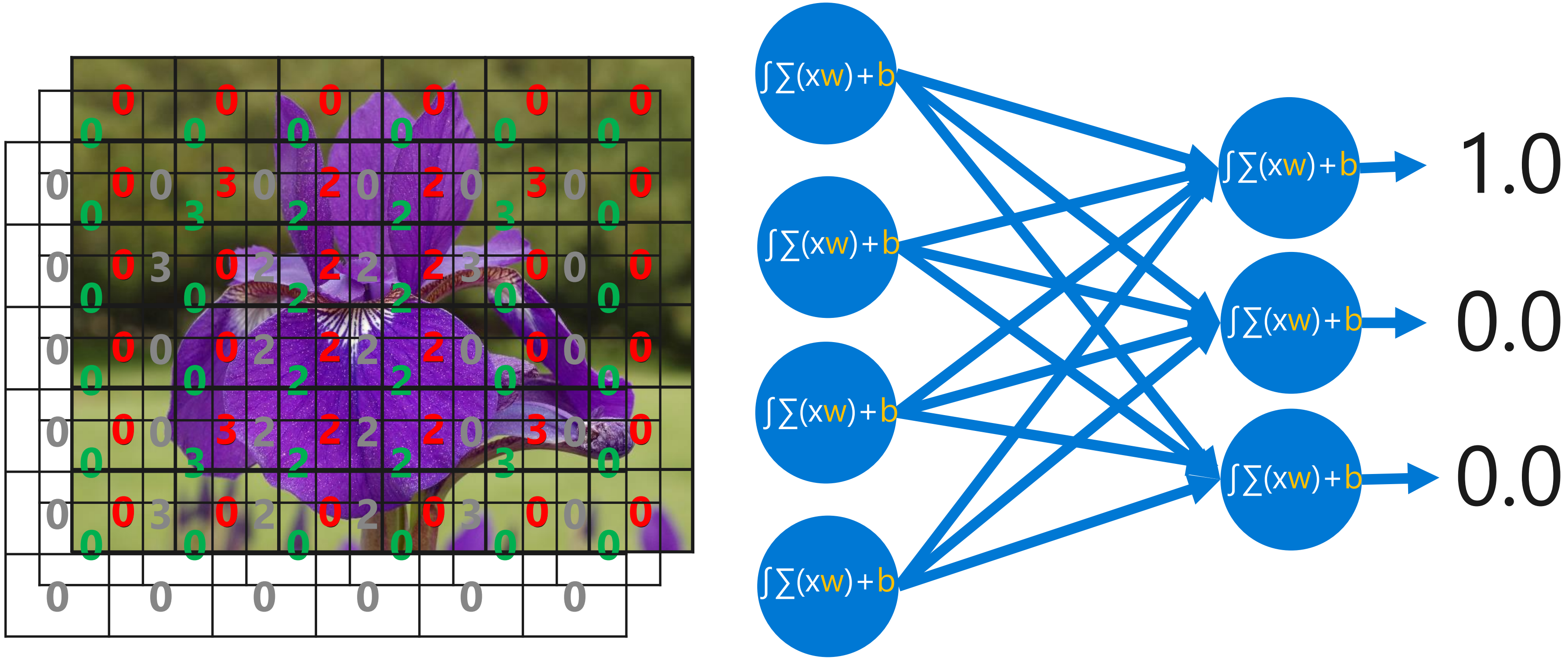
Loss = $\mu(0.2, 0.3, 0.1)^2$

Backpropagation

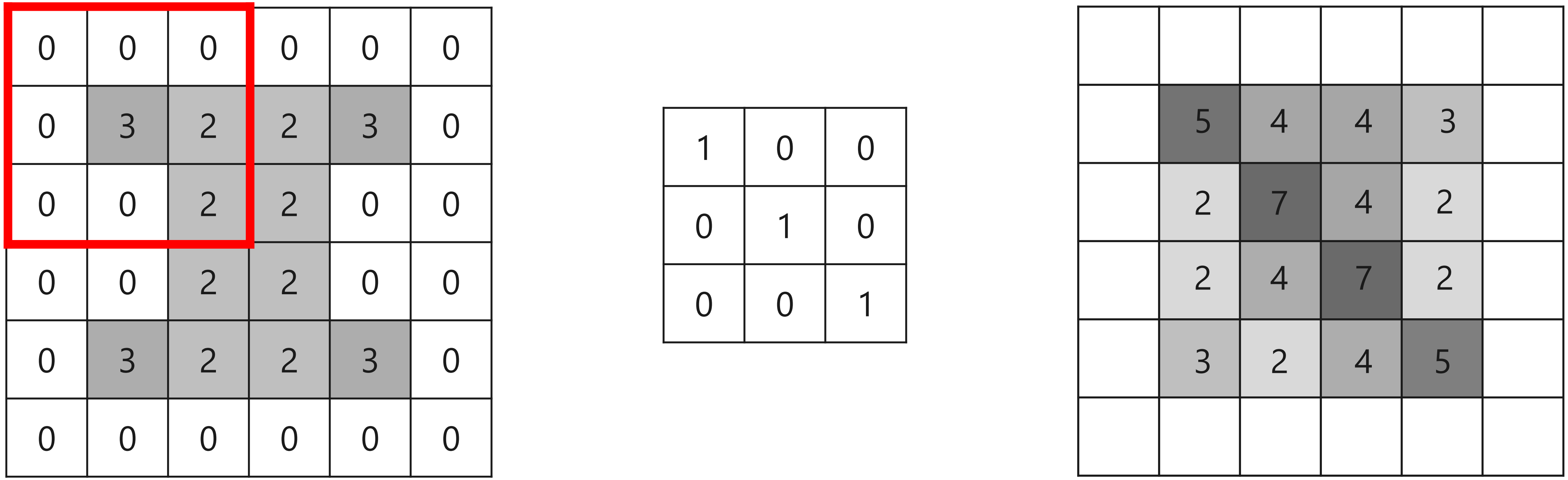


$$loss = \mu([c_1, c_2, c_3, \ldots])^2$$

$$\begin{array}{ccccc} x & & w & & b \\ \begin{pmatrix} 6.6 & 6.1 & 2.1 \\ 3.2 & 2.4 & 0.8 \\ 5.8 & 4.6 & 7.1 \\ 2.4 & 1.9 & 2.3 \end{pmatrix} & \text{x} & \begin{bmatrix} 0.6 \\ 0.3 \\ 0.4 \\ 0.9 \end{bmatrix} & + & \begin{bmatrix} 0.1 \\ 0.2 \\ 0.1 \\ 0.3 \end{bmatrix} \end{array}$$


∫∑(xw)+b
∫∑(xw)+b
∫∑(xw)+b
∫∑(xw)+b
∫∑(xw)+b
∫∑(xw)+b
∫∑(xw)+b
1.0
0.0
0.0

Convolution

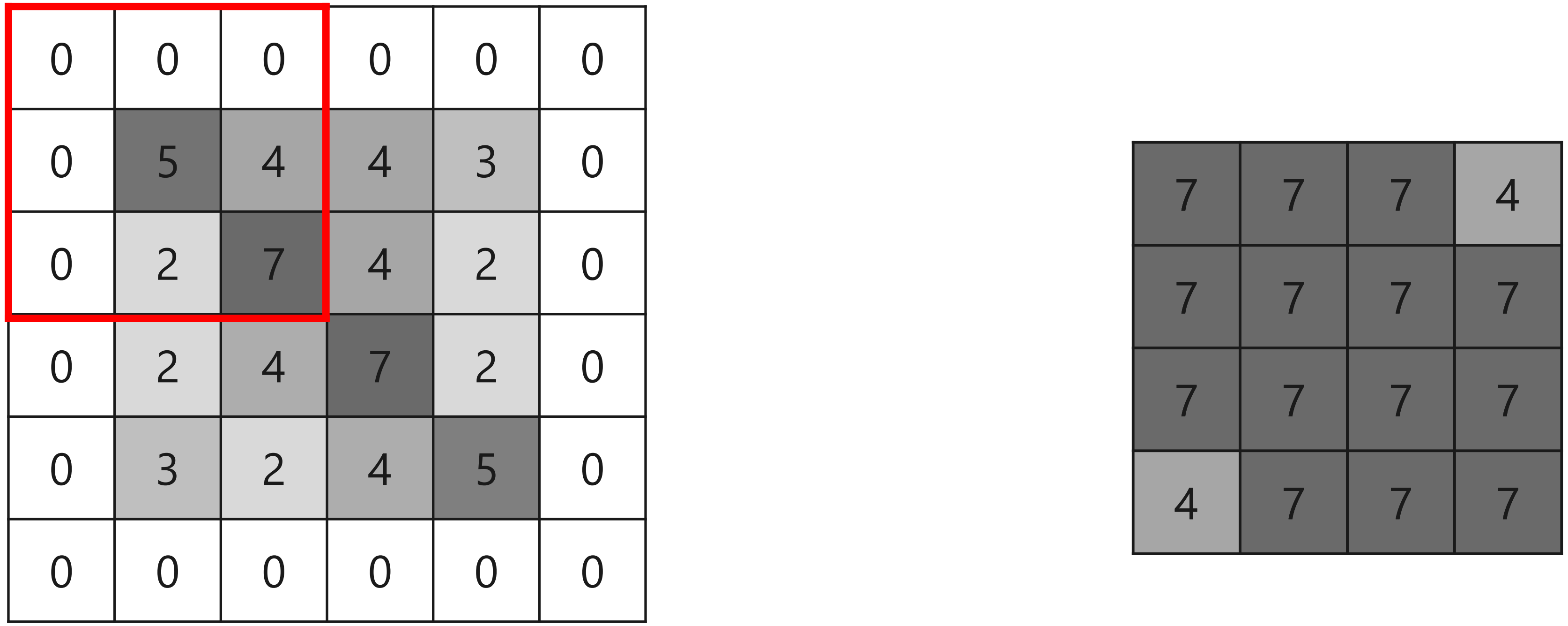


0	0	0	0	0	0
0	3	2	2	3	0
0	0	2	2	0	0
0	0	2	2	0	0
0	3	2	2	3	0
0	0	0	0	0	0

1	0	0
0	1	0
0	0	1

	5	4	4	3	
	2	7	4	2	
	2	4	7	2	
	3	2	4	5	

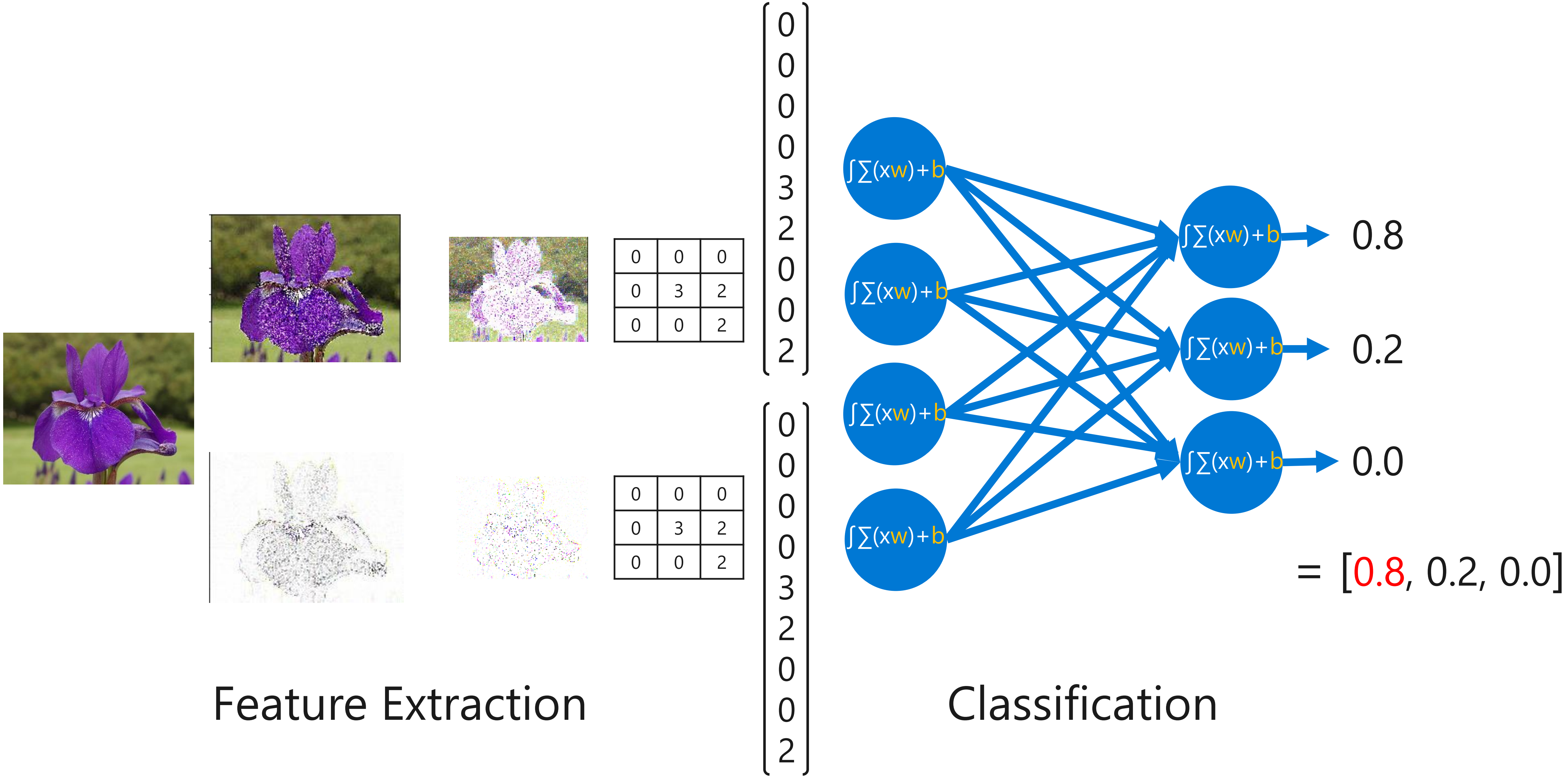
Pooling (Downsampling)



0	0	0	0	0	0
0	5	4	4	3	0
0	2	7	4	2	0
0	2	4	7	2	0
0	3	2	4	5	0
0	0	0	0	0	0

7	7	7	4
7	7	7	7
7	7	7	7
4	7	7	7

Convolutional Neural Network



0	0	0
0	3	2
0	0	2

0	0	0
0	3	2
0	0	2

$$\begin{bmatrix}0\\0\\0\\0\\3\\2\\0\\0\\2\end{bmatrix}$$

$$\begin{bmatrix}0\\0\\0\\0\\3\\2\\0\\0\\2\end{bmatrix}$$

$\int\sum(xw)+b$

$\int\sum(xw)+b$

$\int\sum(xw)+b$

$\int\sum(xw)+b$

$\int\sum(xw)+b$ 0.8

$\int\sum(xw)+b$ 0.2

$\int\sum(xw)+b$ 0.0

= [0.8, 0.2, 0.0]

Feature Extraction

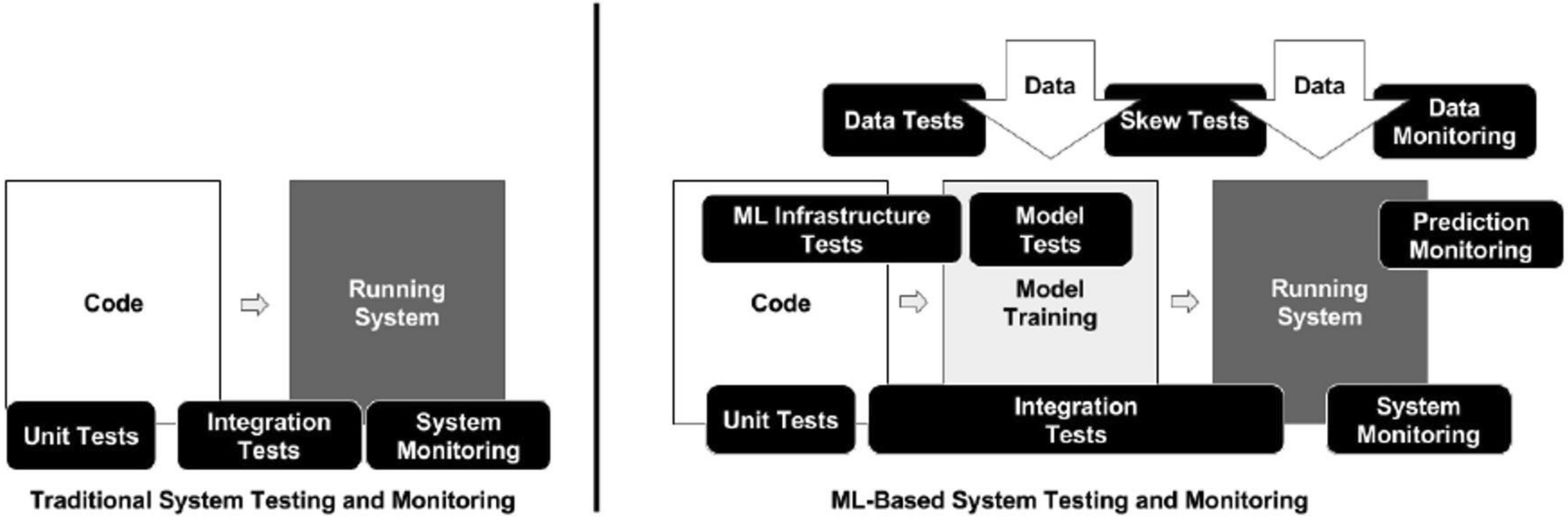
Classification

Questions so far?

Machine Learning in the real world

- Machine Learning
  - 2016-> All about Frameworks (TensorFlow, CNTK, etc..)
  - 2017 -> Training at Scale (GPUs, etc.)
  - 2018 -> MLOps (AIOps)
  - 2019 -> automl
  - 2020 ->citizen data scientists?
  - 2021 ->GPT-3

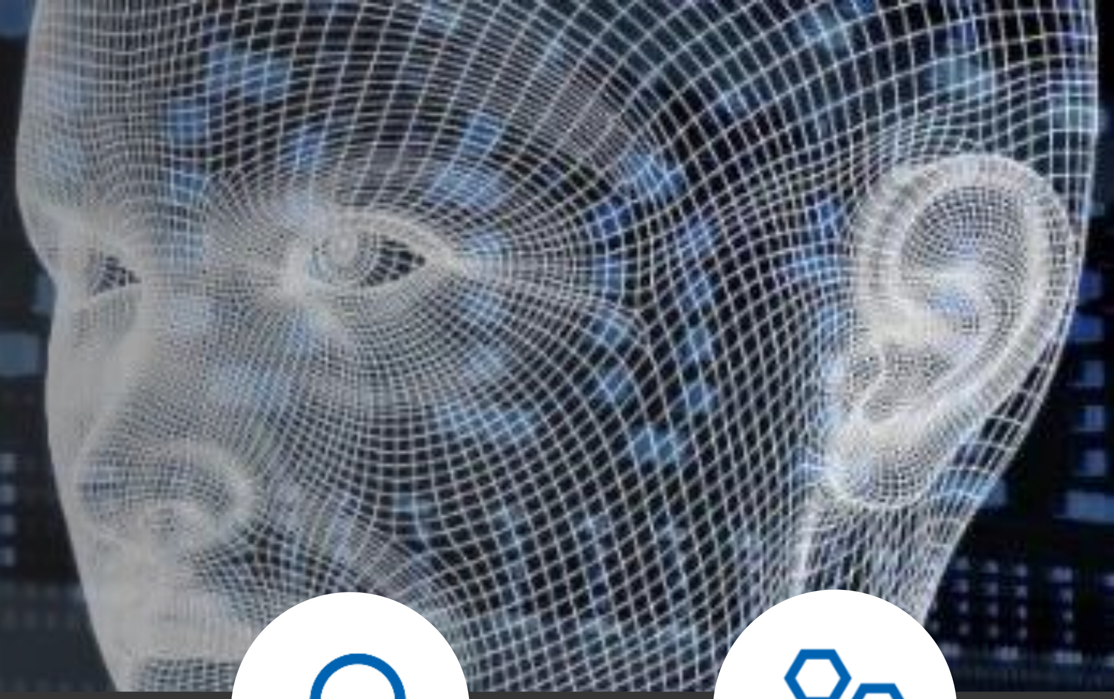
Traditional v/s AI application



Source: Google AI Paper "What's your ML test score? A rubric for ML production systems"

Azure Cognitive Services

Give your apps a human side



Vision

From objects to faces and feelings, enable your apps to analyze still images and video.

Speech

Speak to and hear your users, compensating for environmental noise.

Use with **Language** for max results.

Language

Analyze text to extract user feeling and intent.

Extract knowledge from existing sources and use it to seed chat bots.

Translate between 60+ languages and growing.

Search

Access billions of web pages, images, videos, and news with the power of Bing.

Knowledge

Preview the newest capabilities from analyzing time series to personalization over reinforcement learning.