Lecture 7: Strategies, Challenges and the Black Box

BIODS388/BIOMED388

Anuj Pareek MD PhD, Mars Huang PhD Student

10/18/2020

Outline

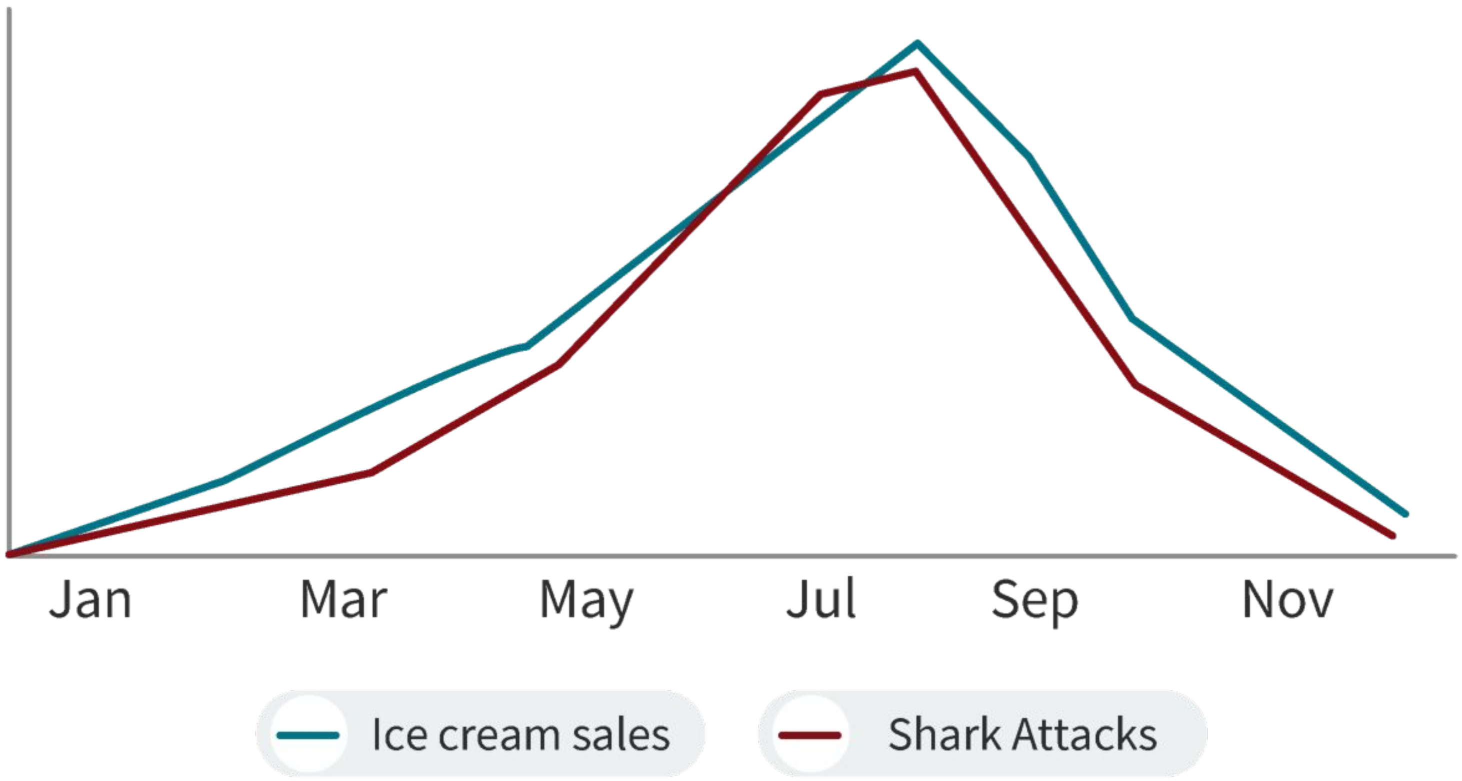
1. **Correlation vs. Causation**
2. Splitting your data
3. Underfitting vs overfitting
4. Strategies to address underfitting and overfitting

Some definitions

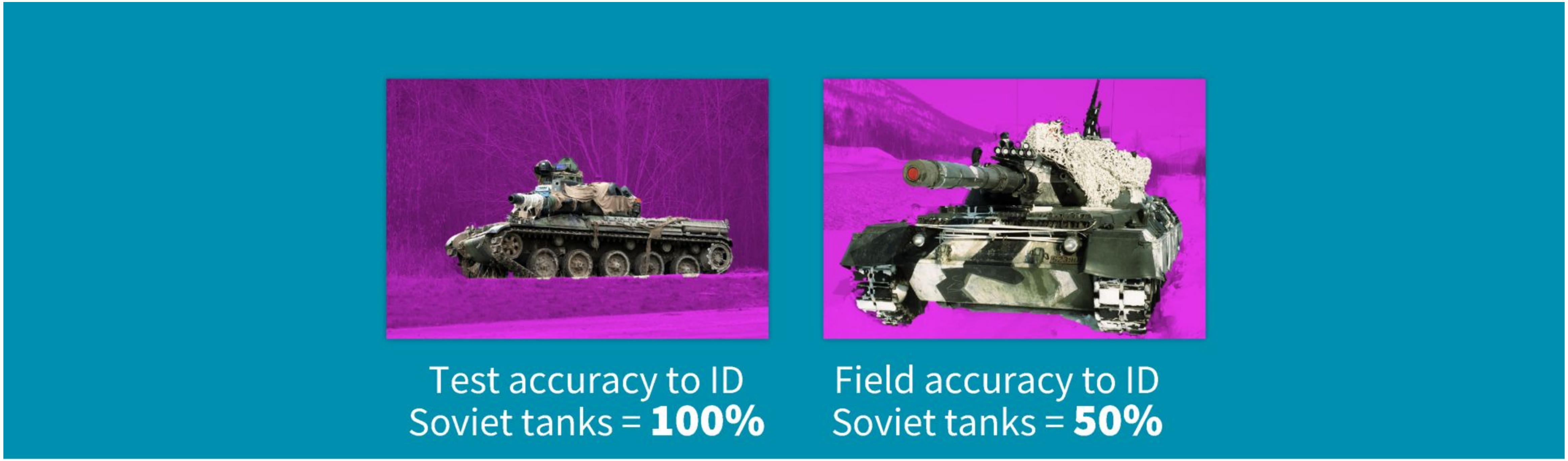
- **Correlation**: The degree to which two events/variables are (linearly) related. The relationship can be causal or non-causal.

- **Causation**: One event/variable directly influences the other event/variable.

CORRELATION IS NOT CAUSATION!

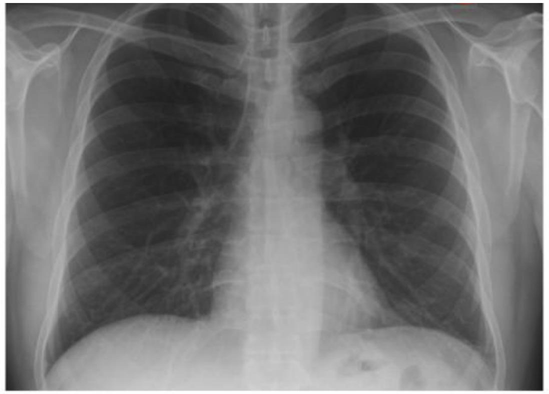


Both ice cream sales and shark attacks increase when the weather is hot and sunny, but they are not caused by other (they are caused by good weather, with lots of people at the beach, both eating ice cream and having a swim in the sea)

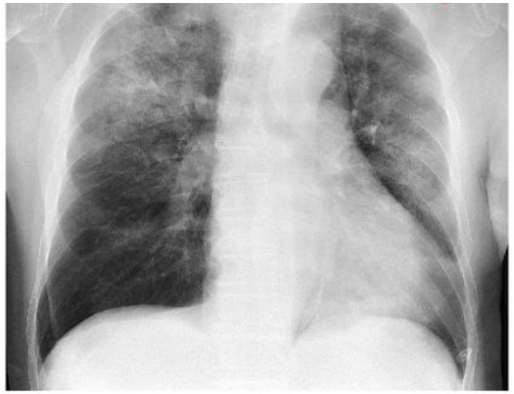


Test accuracy to ID
Soviet tanks = **100%**

Field accuracy to ID
Soviet tanks = **50%**

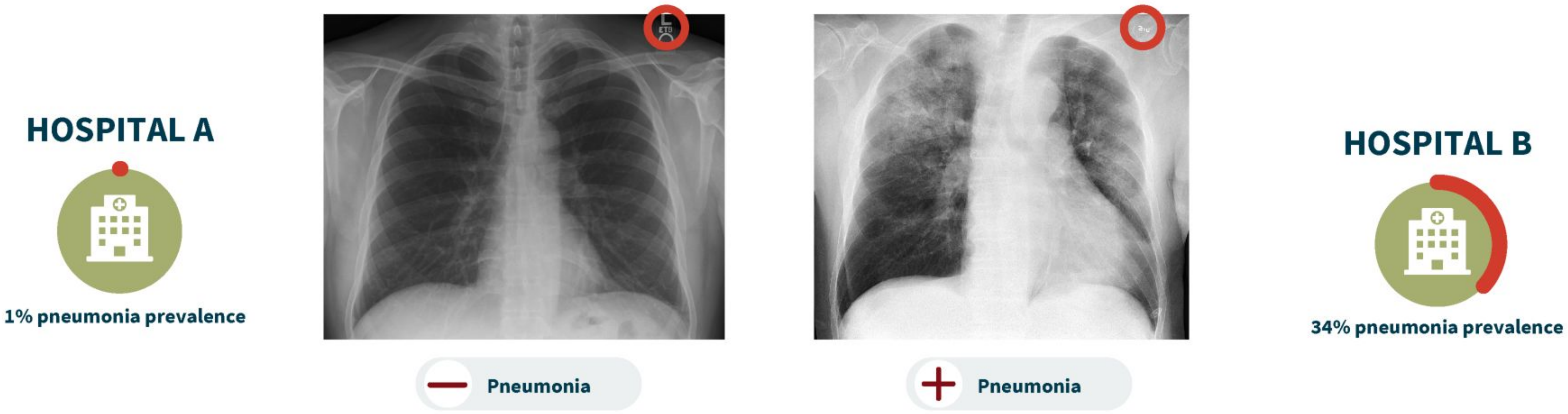


− Pneumonia



+ Pneumonia

What went wrong?



HOSPITAL A

1% pneumonia prevalence

− Pneumonia

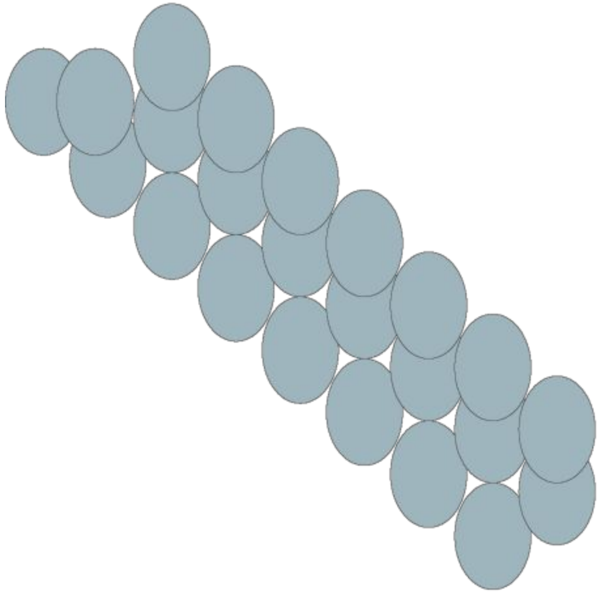
+ Pneumonia

HOSPITAL B

34% pneumonia prevalence

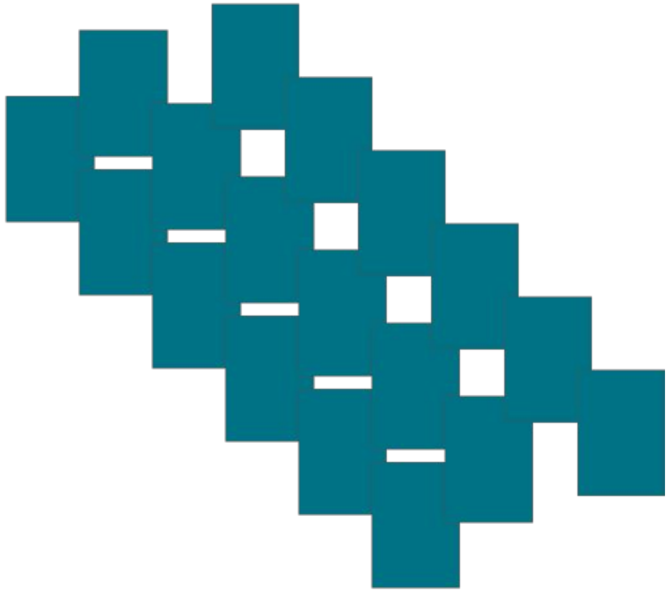
- Can you think of a scenario where a model that focuses on medically irrelevant correlations to make accurate predictions could be useful instead of useless?

- How would that be possible?



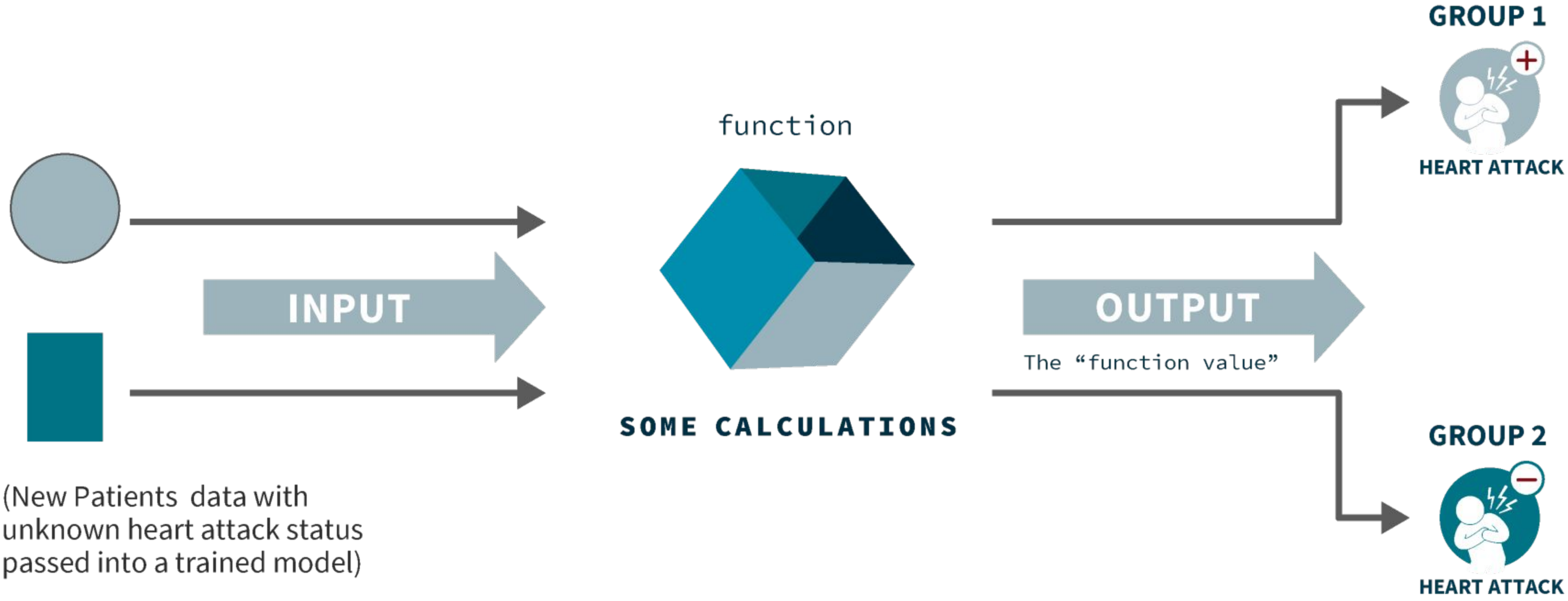
GROUP 1

HEART ATTACK



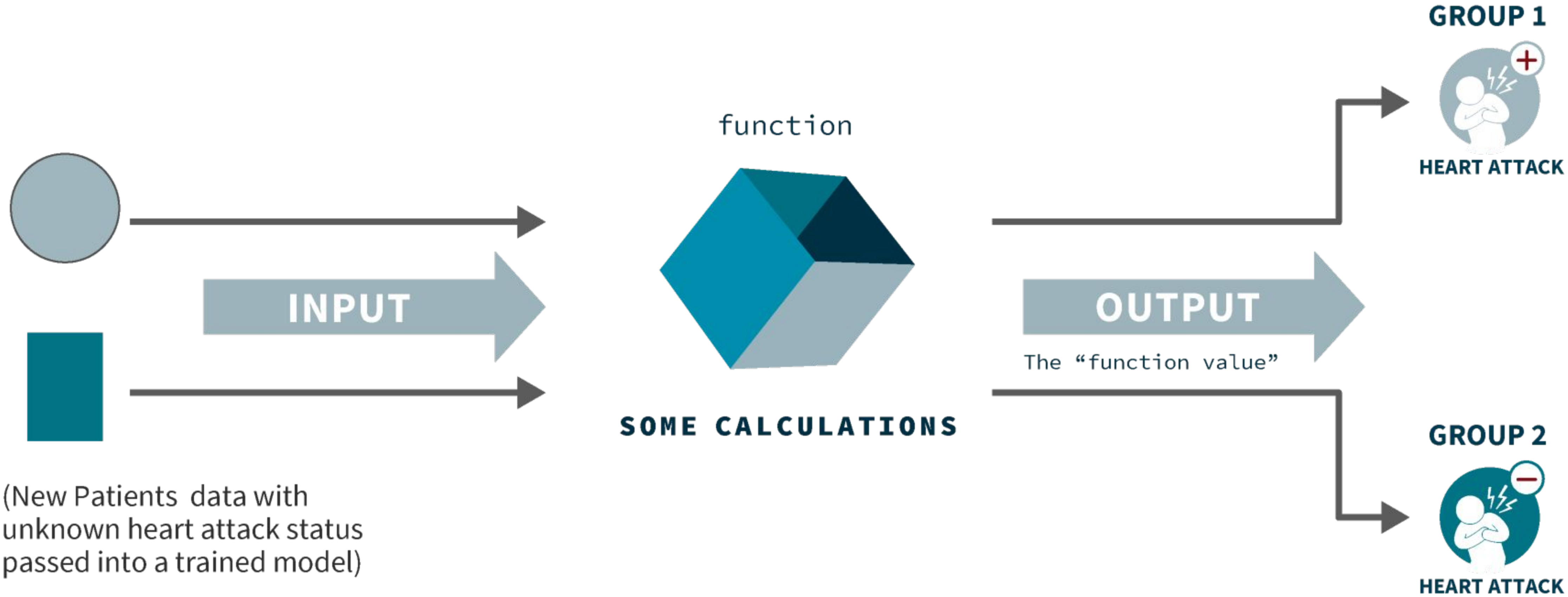
GROUP 2

HEART ATTACK



(New Patients data with unknown heart attack status passed into a trained model)

(The new model correctly classified them - but used the correlation of patient hair color)



INPUT

function

SOME CALCULATIONS

OUTPUT

The “function value”

GROUP 1

HEART ATTACK

GROUP 2

HEART ATTACK

(New Patients data with unknown heart attack status passed into a trained model)

(The new model correctly classified them - but used the correlation of patient hair color)

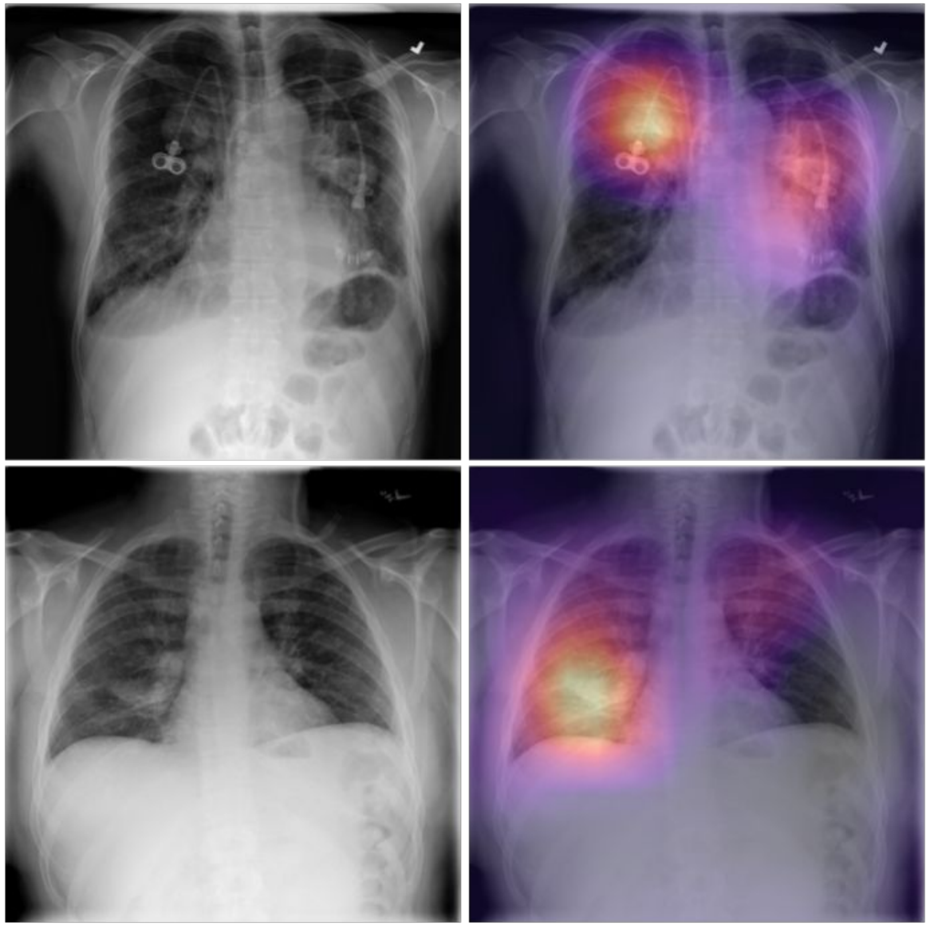
If this model was used for population health management planning rather than medical diagnosis or intervention the correlation (if accurate) of hair color and future heart attack risk can lead to a useful model in this context of risk planning or budget adjustment

- **Take home point**: In the properly constructed context an accurate model does not have to have causal output to be useful!

- Question: Can you think of ways to figure out whether the model is relying solely on spurious correlations rather than causative factors? (Perhaps you have seen examples already)

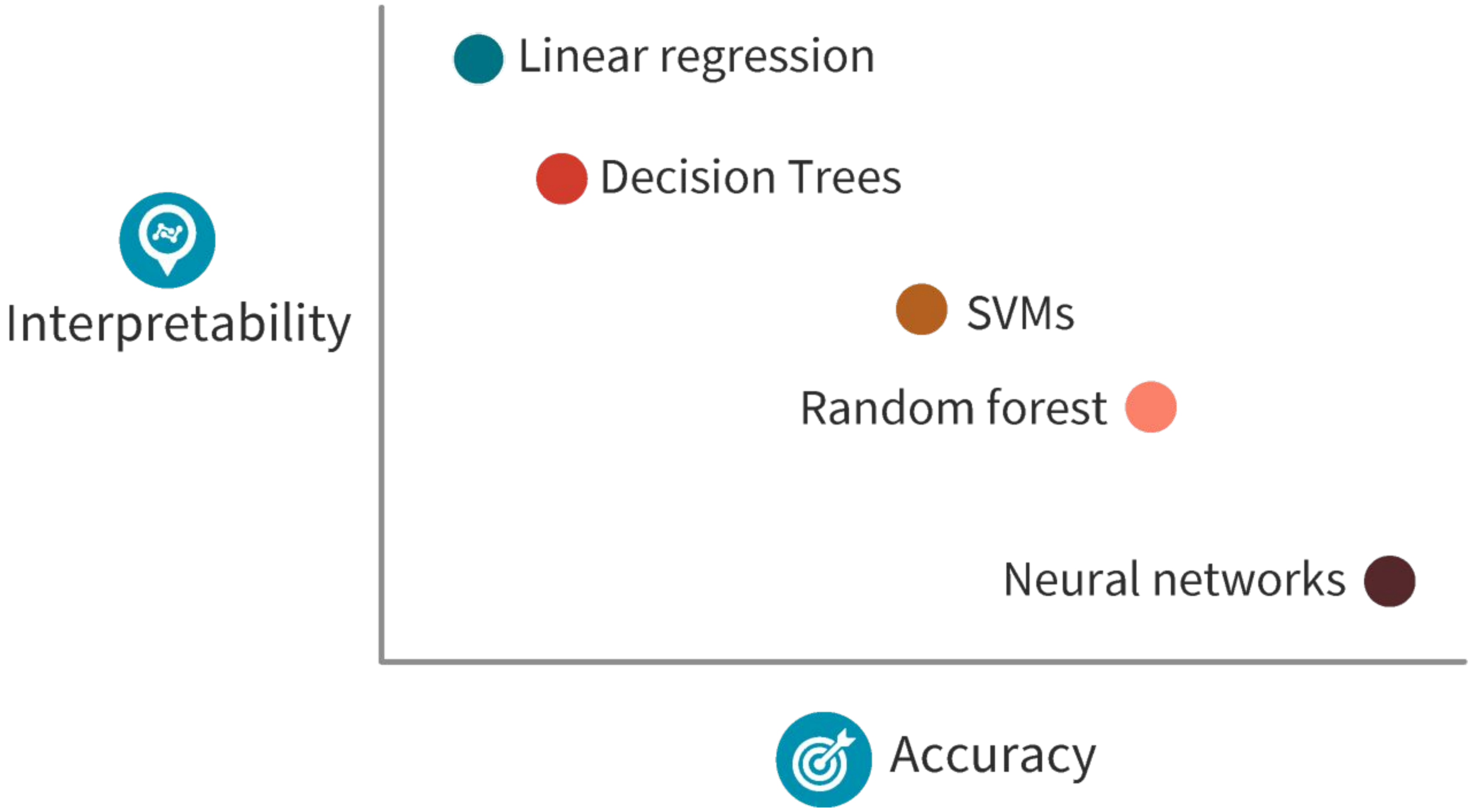
- ML models lack common sense, and therefore we need teams with domain experts!

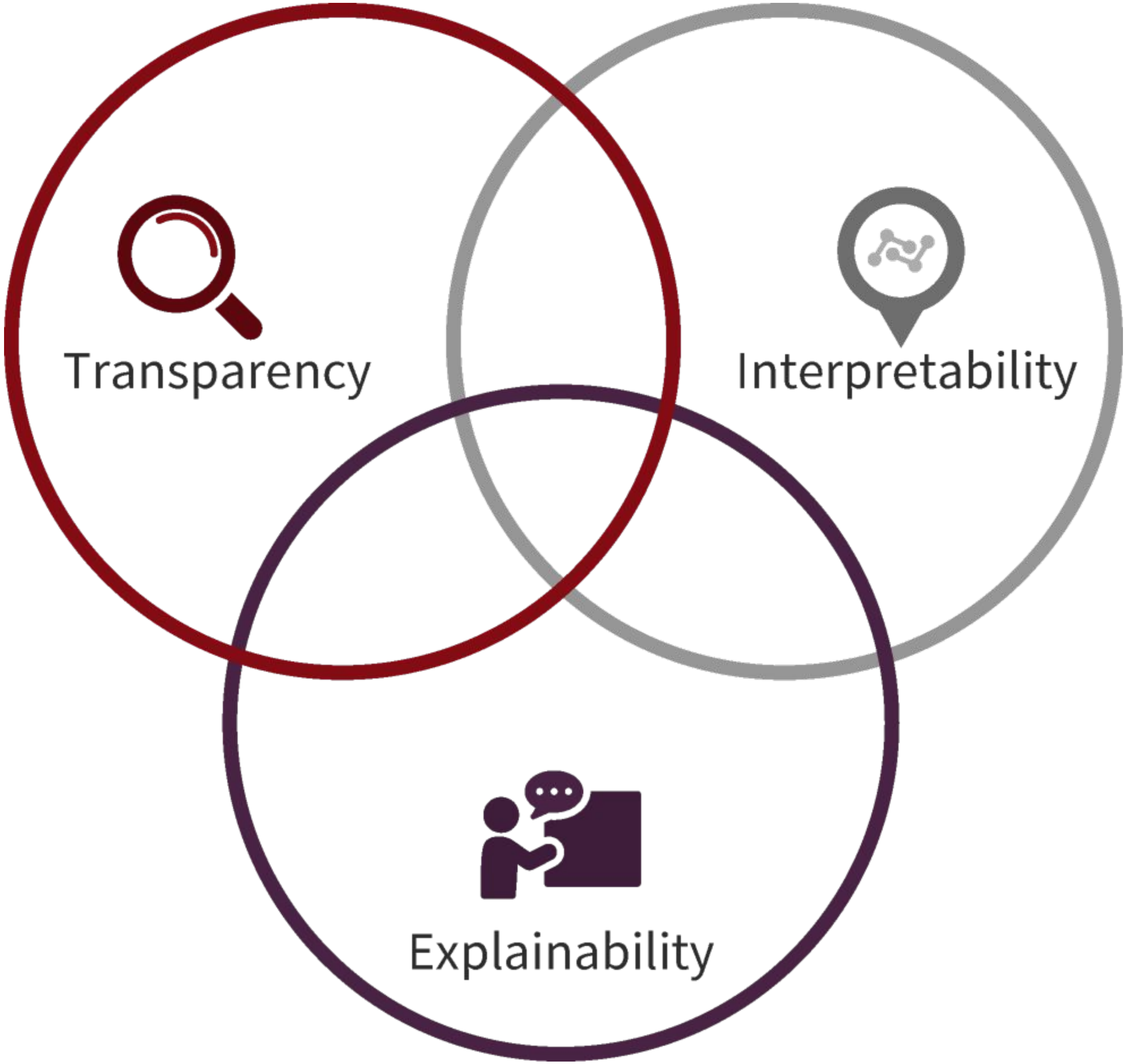
- Strategies include multi-disciplinary teams reviewing false positive and false negative cases predicted by the model, and testing the model on external datasets, to try to gain insight into causal vs. correlative features learned by the model

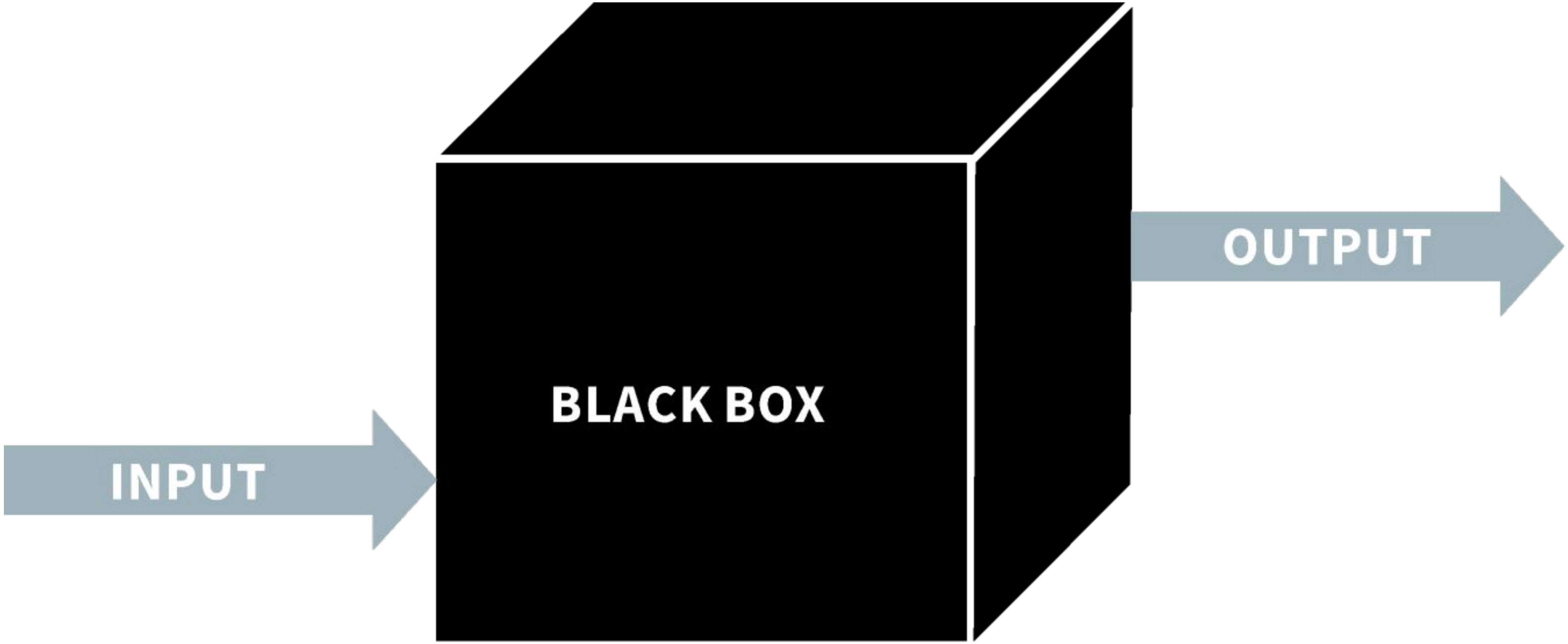


Rajpurkar, Irvin et al. CheXNeXt: Deep learning for chest radiograph diagnosis

TENSION BETWEEN BLACK BOX AND INTERPRETABLE ALGORITHMS



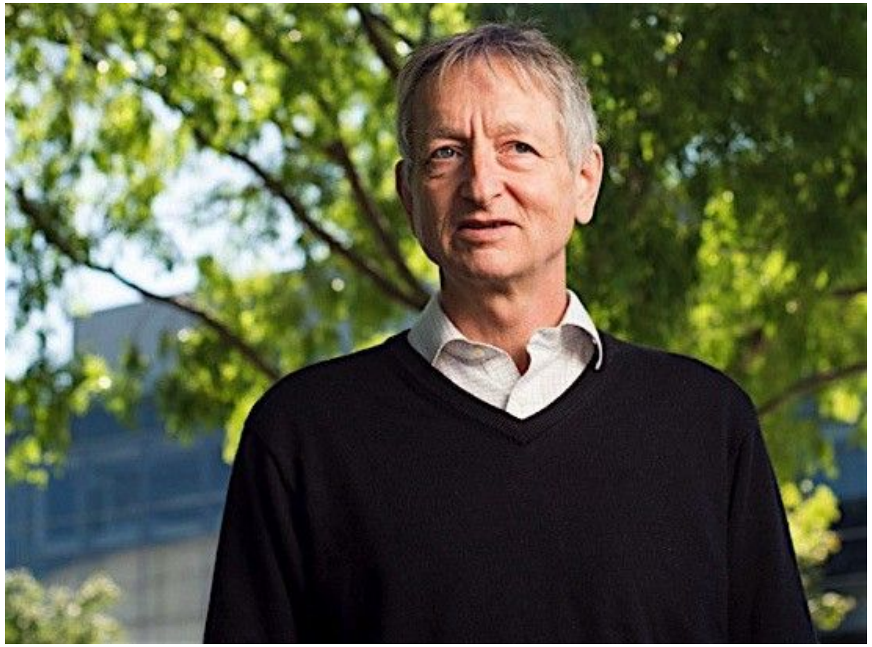

Transparency
Interpretability
Explainability



Internal behavior of the code is unknown

- *“Clinicians and regulators should not insist on explainability because people can't explain how they work for most of the things they do”*

(Geoffrey Hinton)



Machine Learning Model Explainability

- **Intrinsic**

  Intrinsic interpretability is simply referring to models, often simple models, that are self-explanatory from the start.

- **Post-hoc explainability**

  Post-hoc interpretability, which is used to understand decisions by complex models that do not have prescriptive declarative knowledge representations or features

LACE Index

- The LACE index predicts 30-day hospital readmission risk and is calculated using the following 4 intuitive and transparent feature inputs:

1. Length of current admission
2. Admission acuity
3. Patient comorbidities
4. No. of emergency department visits in the past 6 months.

L - Length of Stay

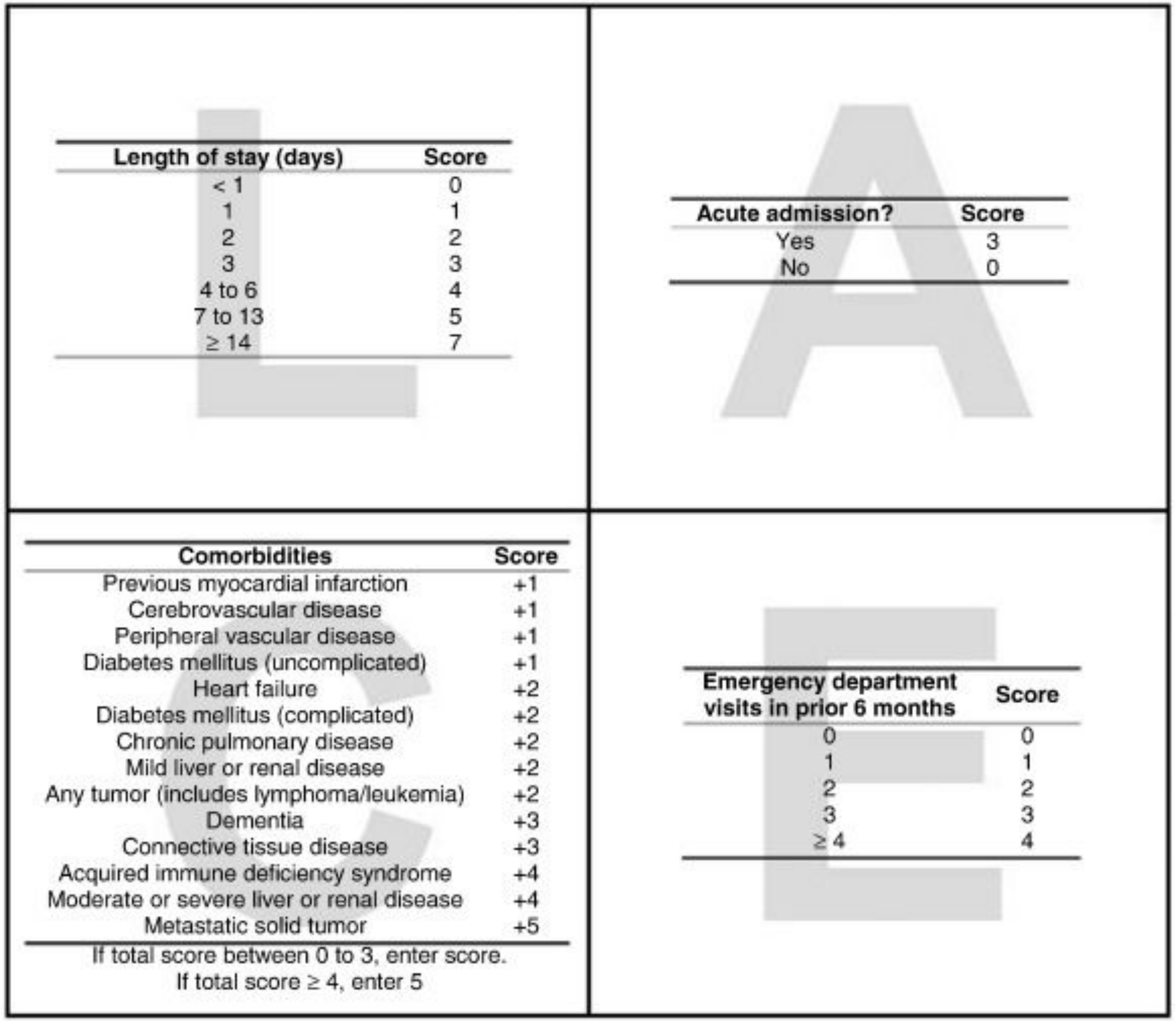
A - Acuity of Admission

C - Comorbidities

E - Emergency room visits

LACE Index

- What kind of interpretability does the LACE index have?



Length of stay (days)	Score
< 1	0
1	1
2	2
3	3
4 to 6	4
7 to 13	5
≥ 14	7

Acute admission?	Score
Yes	3
No	0

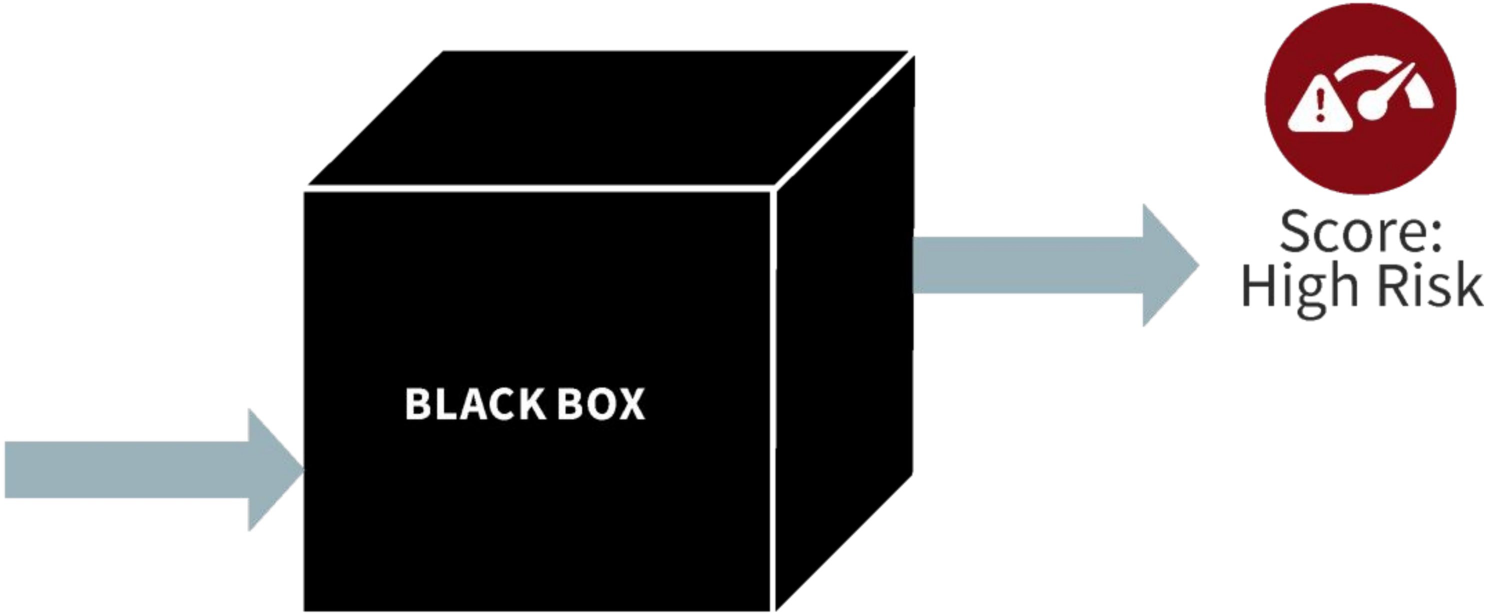
Comorbidities	Score
Previous myocardial infarction	+1
Cerebrovascular disease	+1
Peripheral vascular disease	+1
Diabetes mellitus (uncomplicated)	+1
Heart failure	+2
Diabetes mellitus (complicated)	+2
Chronic pulmonary disease	+2
Mild liver or renal disease	+2
Any tumor (includes lymphoma/leukemia)	+2
Dementia	+3
Connective tissue disease	+3
Acquired immune deficiency syndrome	+4
Moderate or severe liver or renal disease	+4
Metastatic solid tumor	+5

If total score between 0 to 3, enter score.
If total score ≥ 4, enter 5

Emergency department visits in prior 6 months	Score
0	0
1	1
2	2
3	3
≥ 4	4

Clinical input features

Glasgow onsciousness scale (GCS)
Systolic Blood Pressure (SBP) (mmHg)
Pulse Rate (Beat/minute)
Respiratory rate
Oral temperature (^{o}C)
O_2 Saturation (%)
Arterial HCO_3 (mM); Normal: 22-26 mM
Serum CO_2 Pressure; Normal: 35-45 mmHg
Arterial pH (7.35-7.45)
Serum Potassium (K) (meq/l); Normal:3.5-5
Serum Sodium (Na)(meq/l); Normal: 135-150
Hematocrite (%)
WBC Count (per mm)
Hemoglobin (g/dl)
Blood glucose level at admission (mg.ml)
(70-110mg/dl)
Serum Calcium (mg/dl); Normal 8-10 mg/dl
Serum Magnesium (mg/dl); Normal: 1.8-3 mg/dl
Alanine aminotransferase (ALT) (U/L) (7-56 U/l)
Asparte aminotransferase (AST) (U/L)(5-35 U/L)
Total Bilirubing (mg/dl); Normal: 0.2-1.3 mg/dl
Serum creatinin (mg/dl)
Blood Urea nitroger (mg/dl)



BLACK BOX

Score:
High Risk

Some key messages

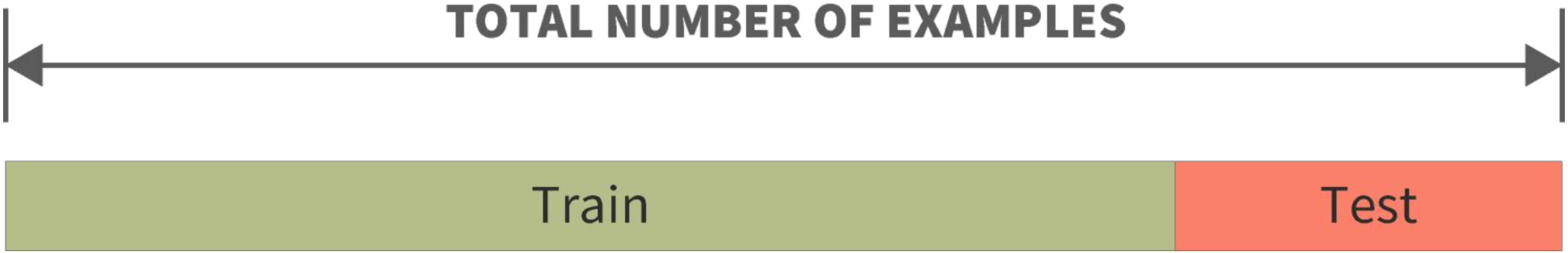
- Both black box and transparent model performance should be evaluated against existing standards of care on real-world data to evaluate effectiveness in their specific patient population.

- Black Box models (low model interpretability) are especially important to evaluate with empirical pilot testing. Preferably on prospective data, external data and potentially in a trial setting

- Clinicians should be educated on the benefits, risks, and limitations of a given clinical model based on the evaluation metrics.

Outline

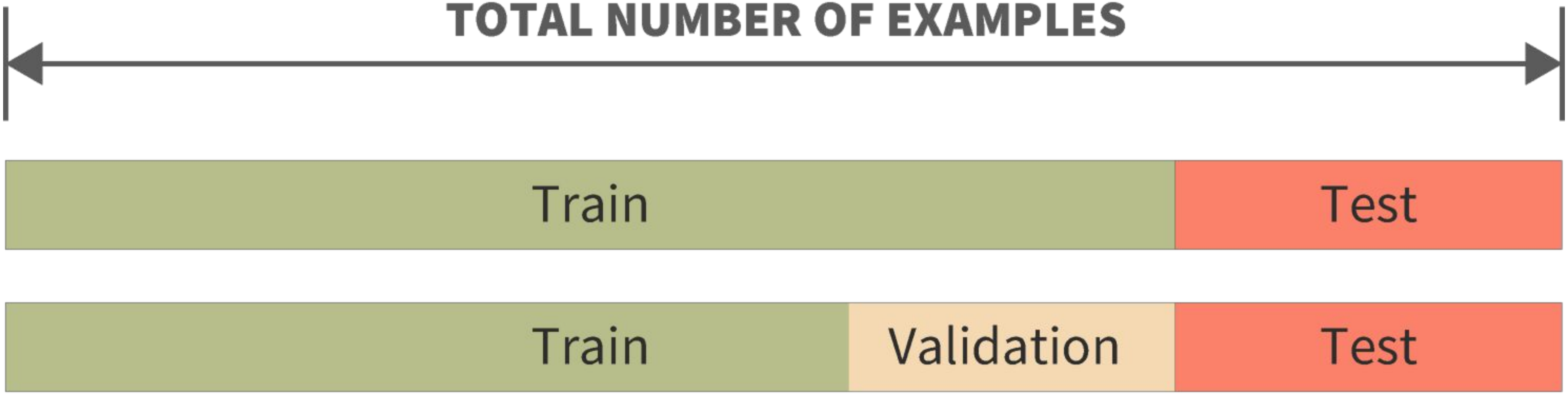
1. Correlation vs. Causation
2. **Splitting your data**
3. Underfitting vs overfitting
4. Strategies to address underfitting and overfitting



TOTAL NUMBER OF EXAMPLES

Train

Test



TOTAL NUMBER OF EXAMPLES

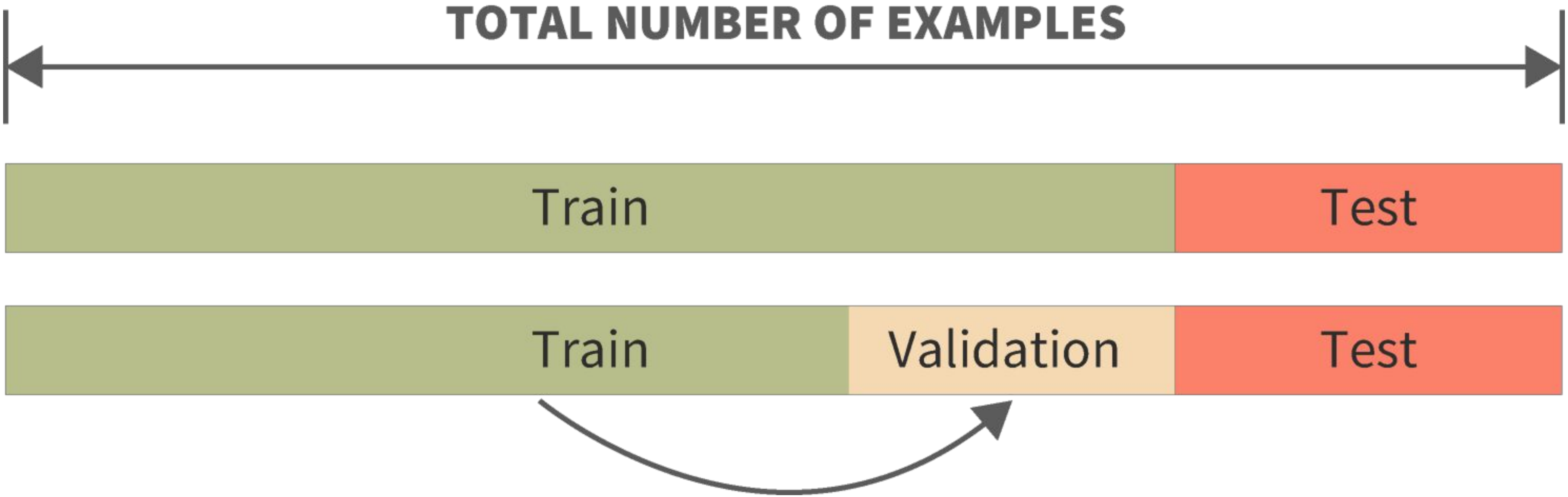
Train

Test

Train

Validation

Test



TOTAL NUMBER OF EXAMPLES

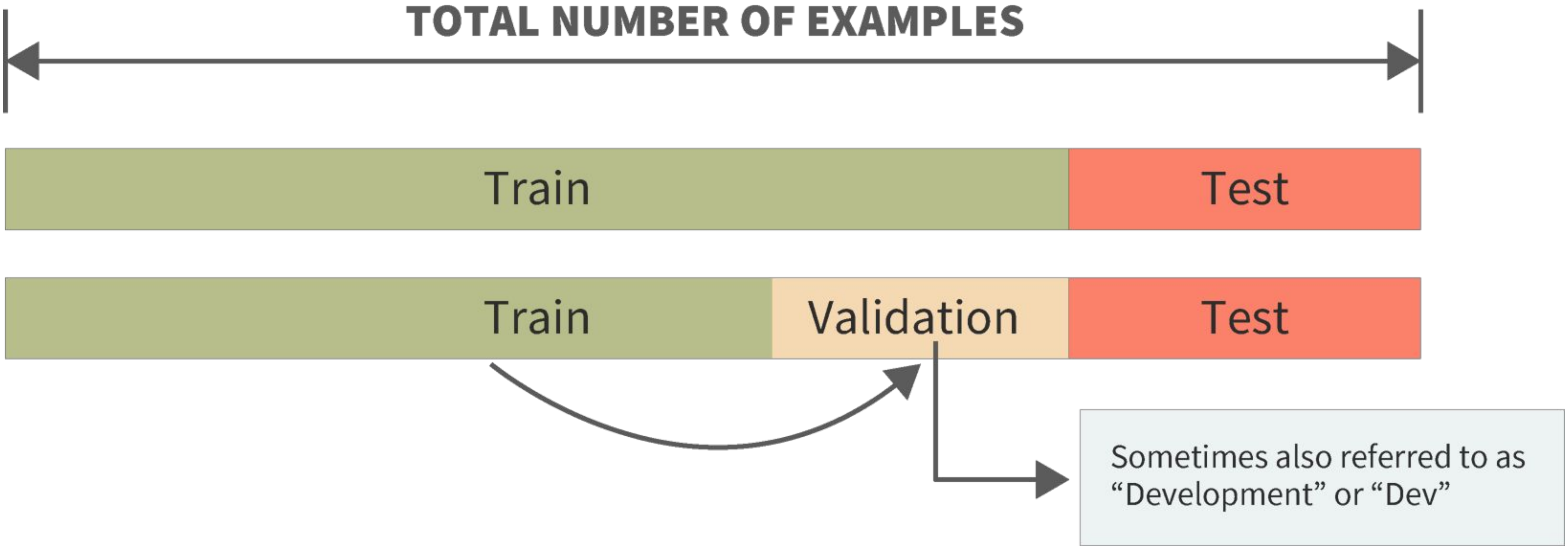
Train

Test

Train

Validation

Test



TOTAL NUMBER OF EXAMPLES

Train

Test

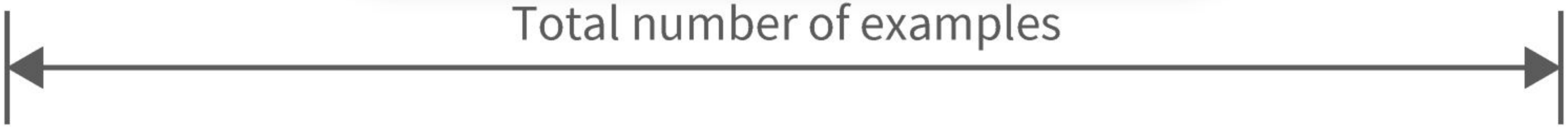
Train

Validation

Test

Sometimes also referred to as “Development” or “Dev”

K-FOLD CROSS VALIDATION

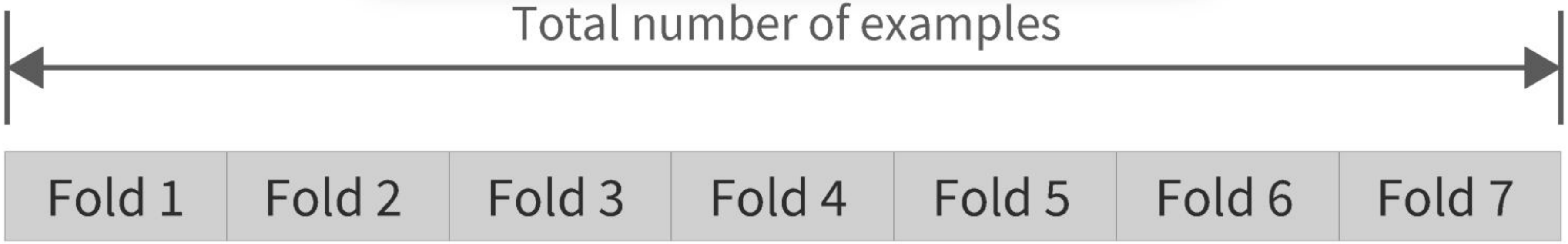


Total number of examples

IMPORTANT NOTE:

Cross-fold validation can lead to models learning from the test set and is a major flaw in many research papers. Instead, to tune your model, nested cross fold validation should be used.

K-FOLD CROSS VALIDATION



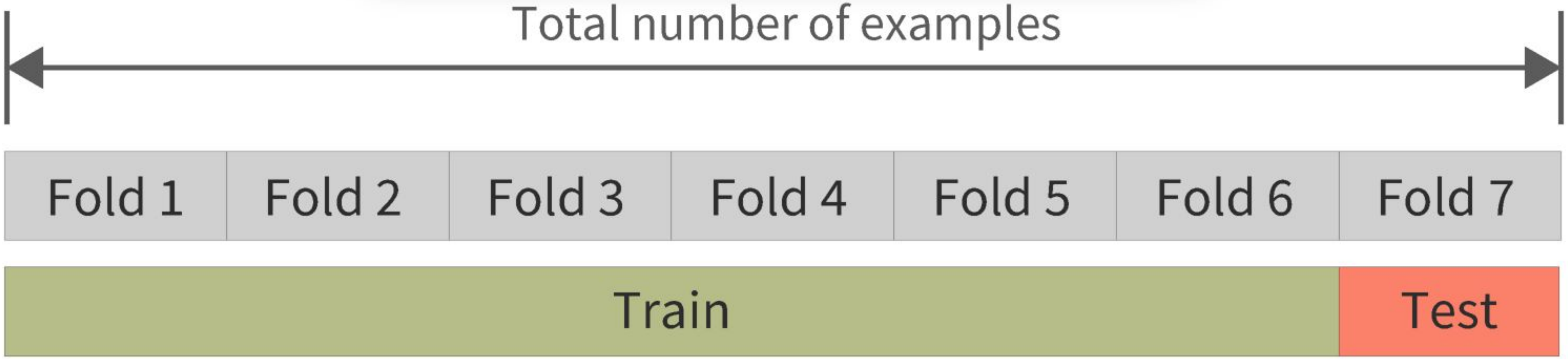
Total number of examples

Fold 1	Fold 2	Fold 3	Fold 4	Fold 5	Fold 6	Fold 7

IMPORTANT NOTE:

Cross-fold validation can lead to models learning from the test set and is a major flaw in many research papers. Instead, to tune your model, nested cross fold validation should be used.

K-FOLD CROSS VALIDATION



Total number of examples

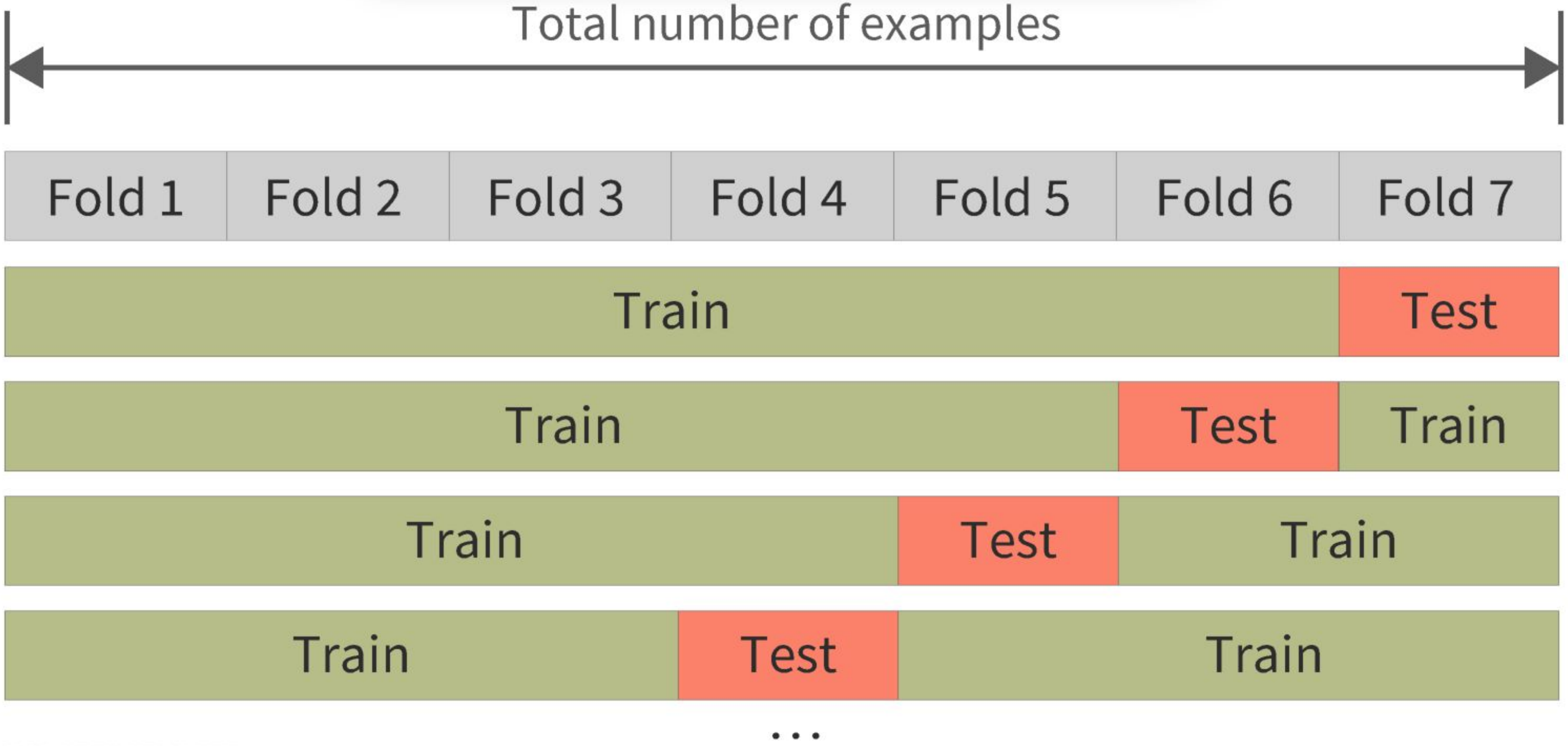
Fold 1 | Fold 2 | Fold 3 | Fold 4 | Fold 5 | Fold 6 | Fold 7

Train | Test

IMPORTANT NOTE:

Cross-fold validation can lead to models learning from the test set and is a major flaw in many research papers. Instead, to tune your model, nested cross fold validation should be used.

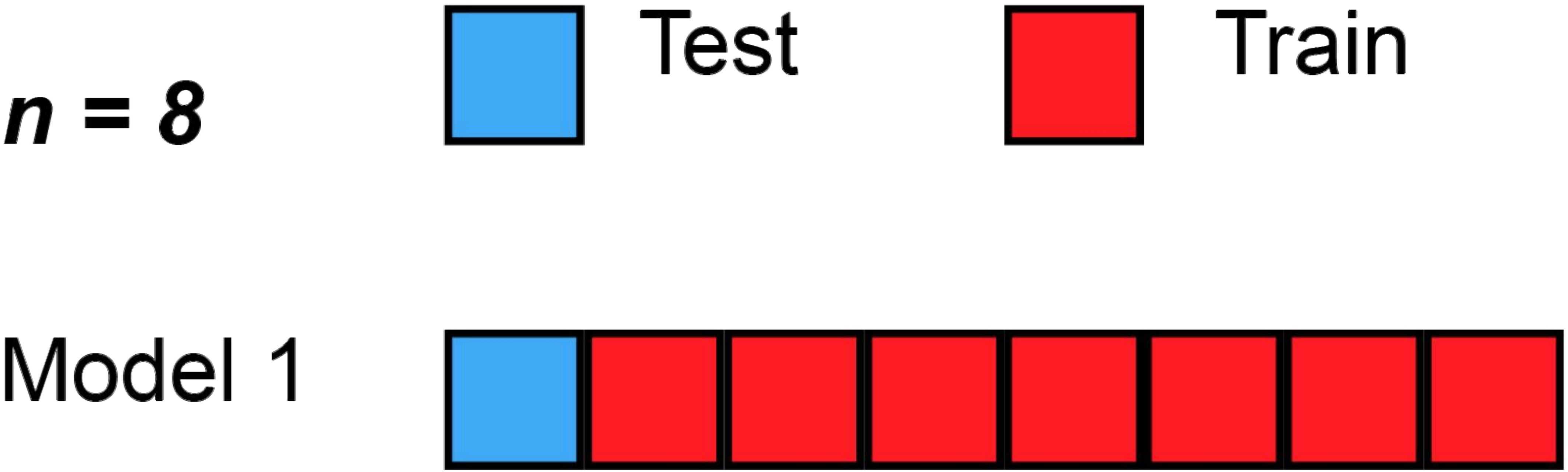
K-FOLD CROSS VALIDATION



IMPORTANT NOTE:

Cross-fold validation can lead to models learning from the test set and is a major flaw in many research papers. Instead, to tune your model, nested cross fold validation should be used.

Leave One Out Cross Validation



n = 8

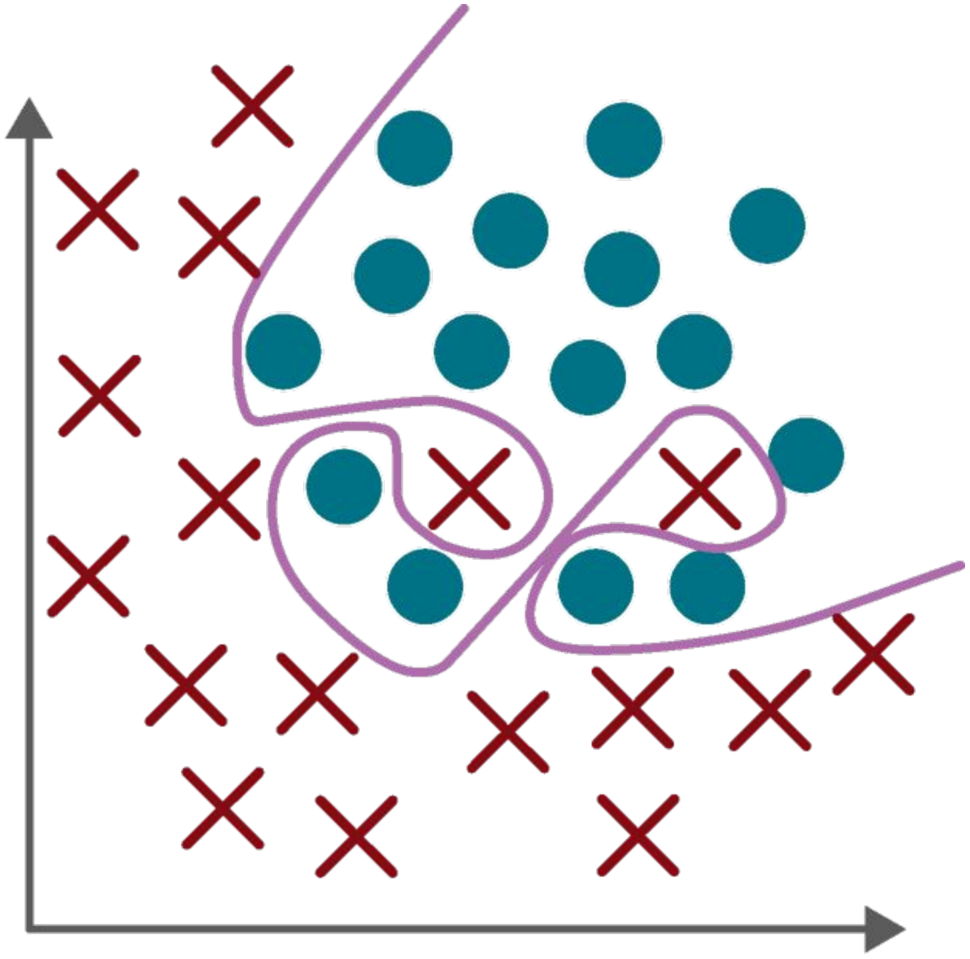
🟦 Test

🟥 Train

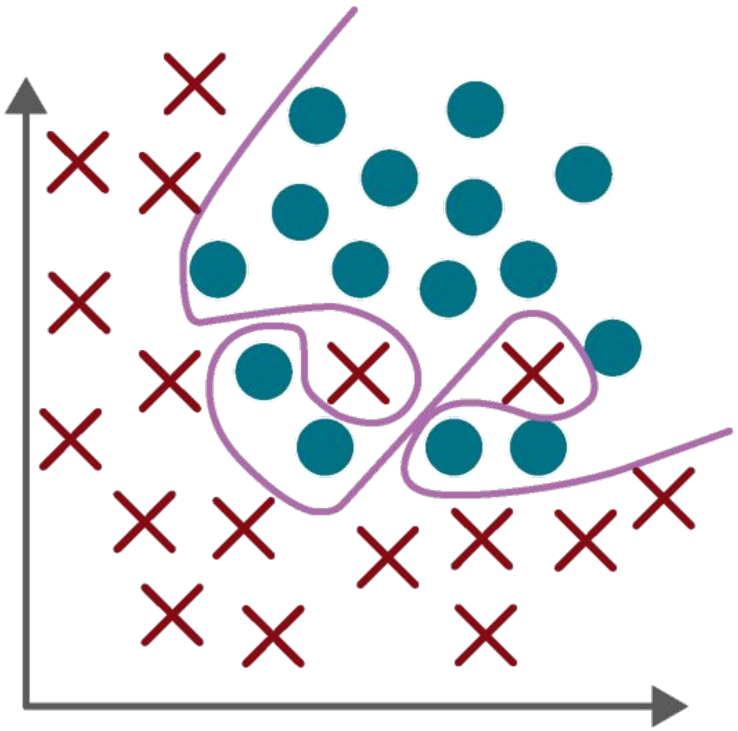
Model 1

Outline

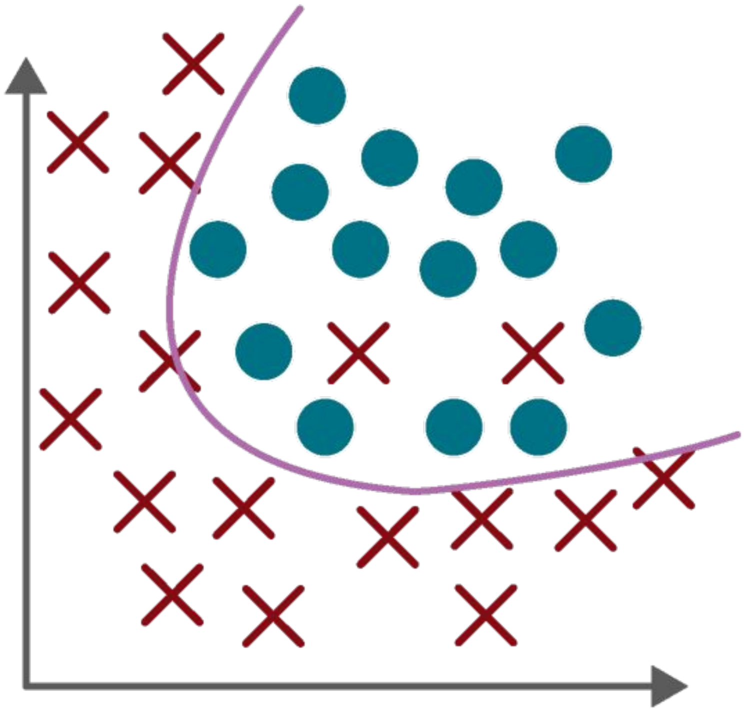
1. Correlation vs. Causation
2. Splitting your data
3. **Underfitting vs overfitting**
4. Strategies to address underfitting and overfitting



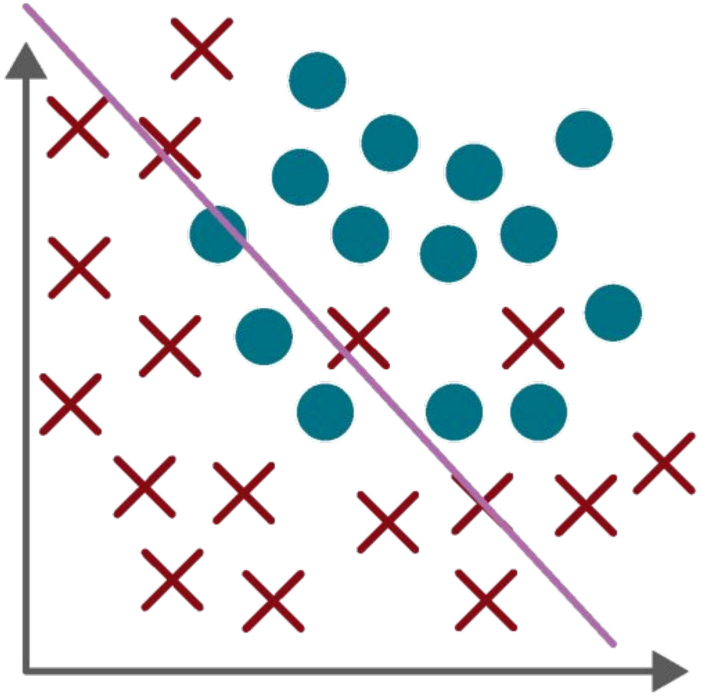
Over-fitting: fitting too strongly to the training data



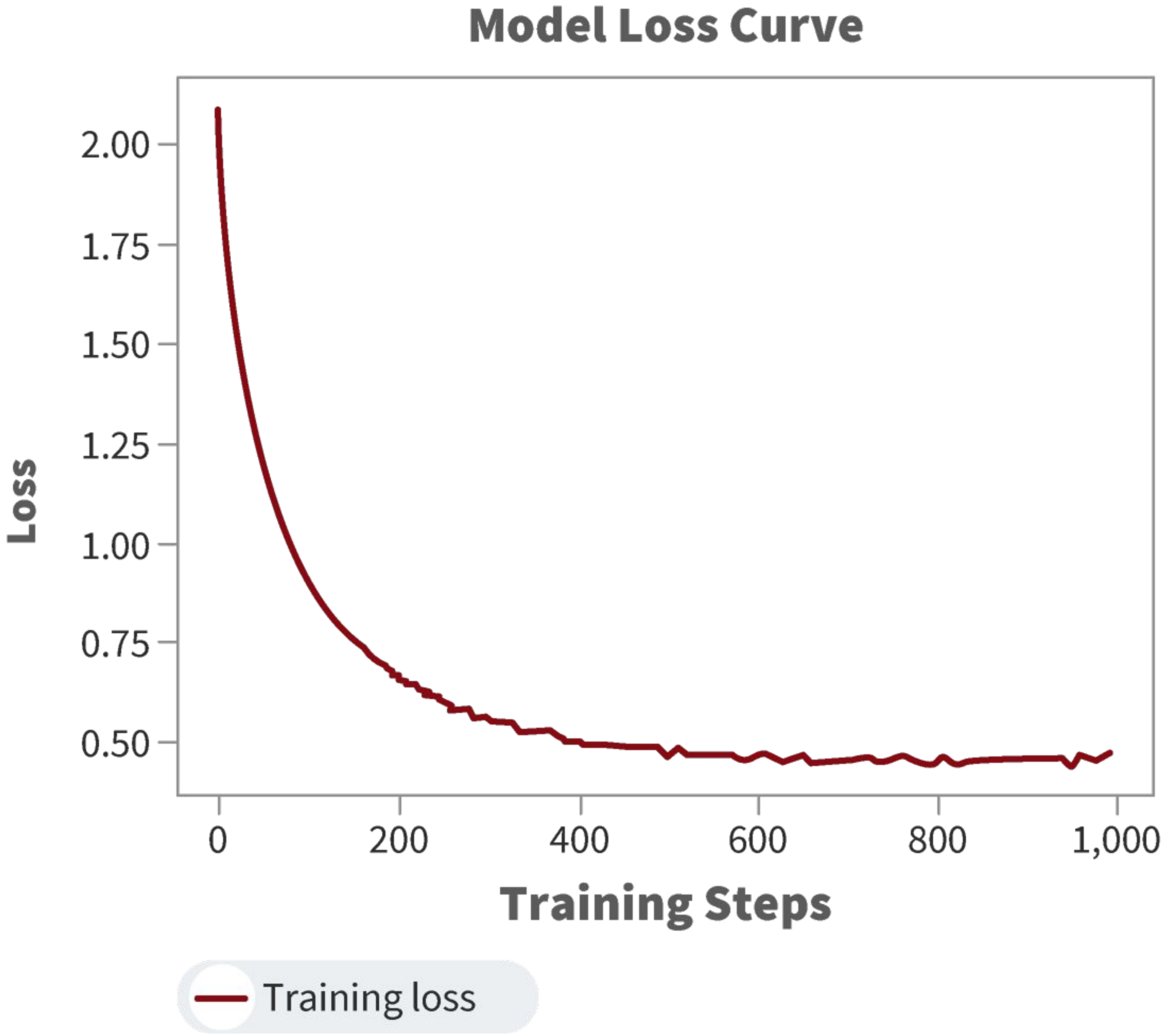
Over-fitting: fitting too strongly to the training data

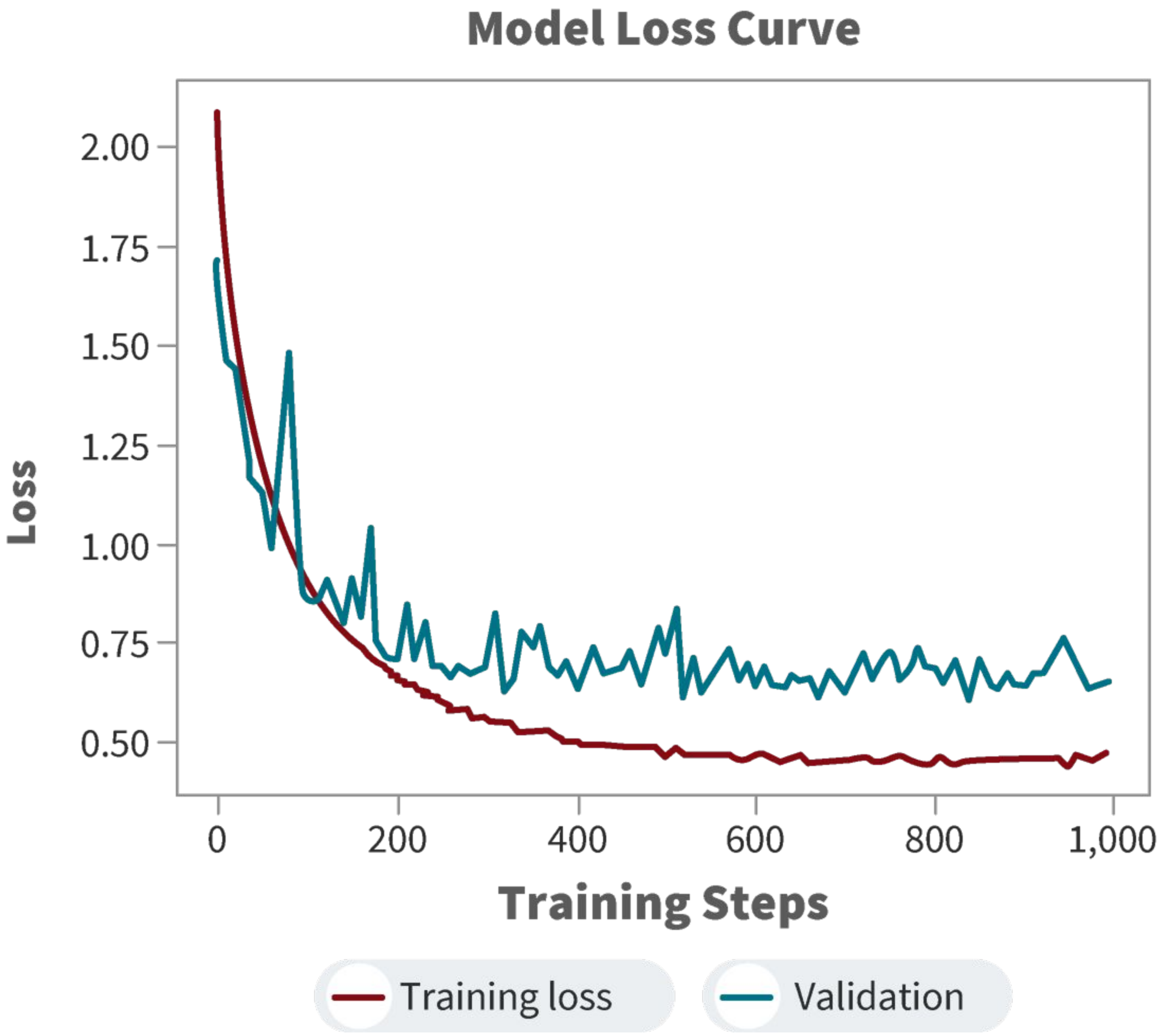


Appropriate-fitting

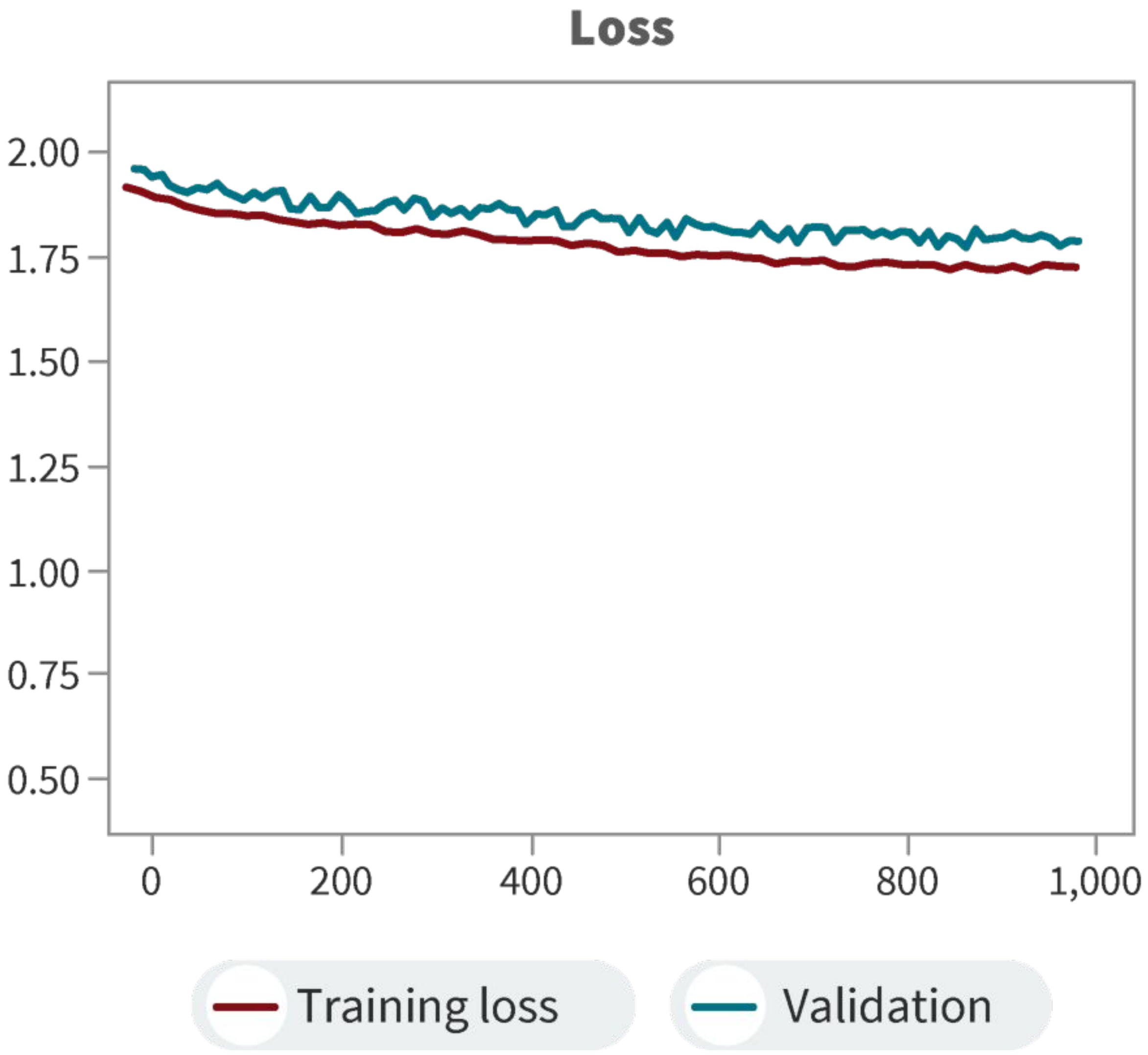


Under-fitting: unable to capture underlying trend of data

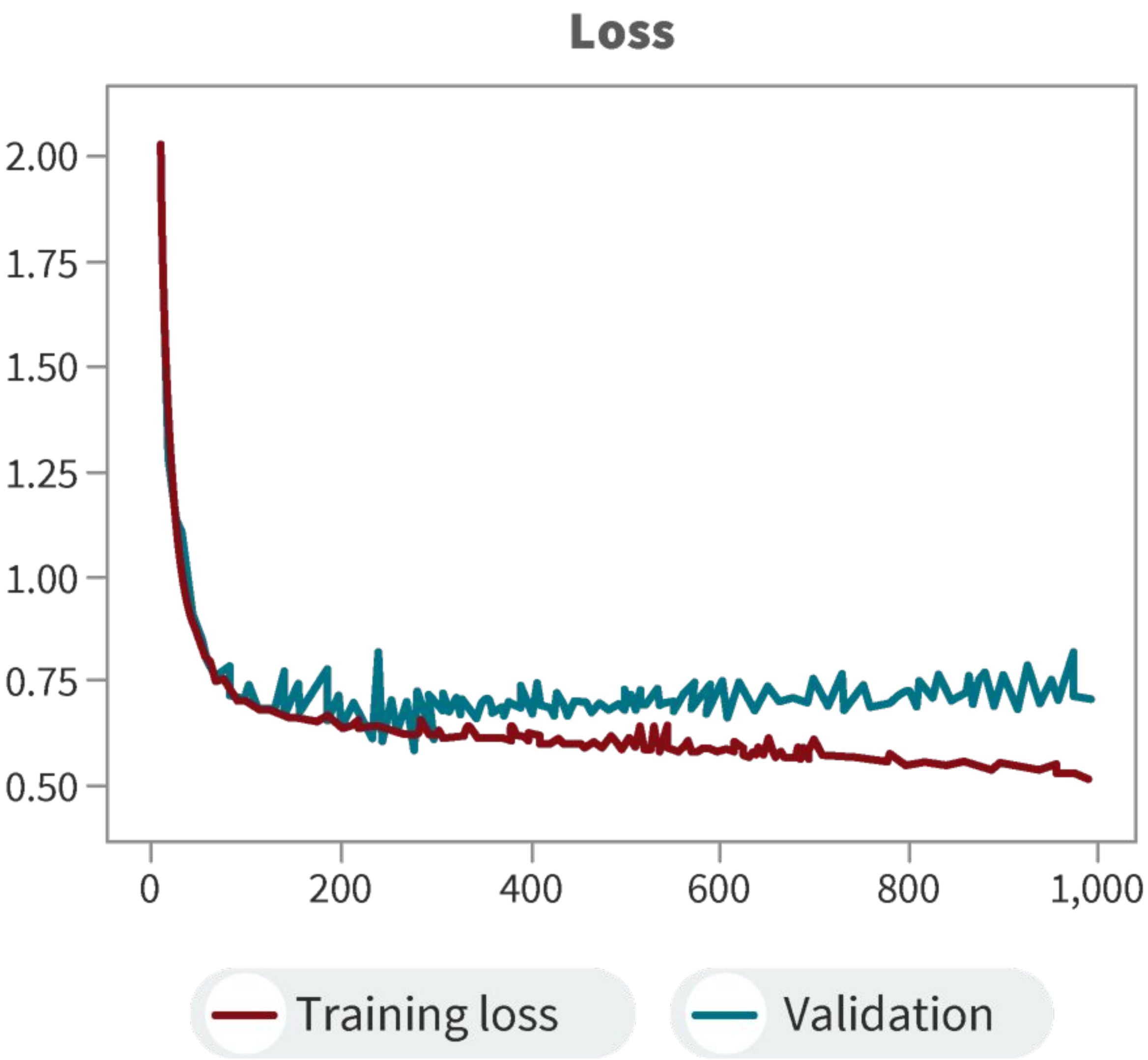

Model Loss Curve
2.00
1.75
1.50
1.25
1.00
0.75
0.50
Loss
0
200
400
600
800
1,000
Training Steps
Training loss


Model Loss Curve
Loss
2.00
1.75
1.50
1.25
1.00
0.75
0.50
0
200
400
600
800
1,000
Training Steps
Training loss
Validation

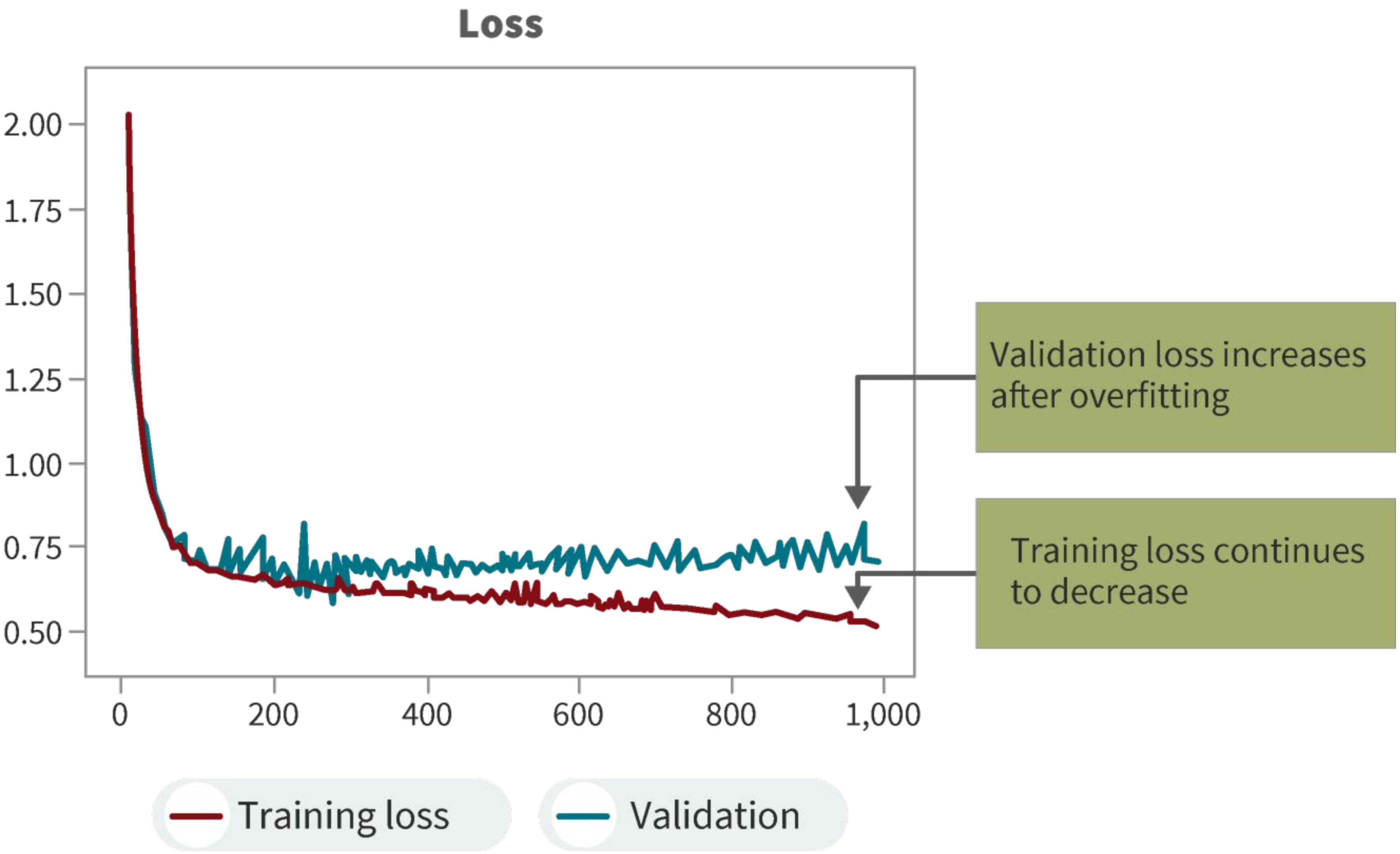
UNDERFITTING



OVERFITTING



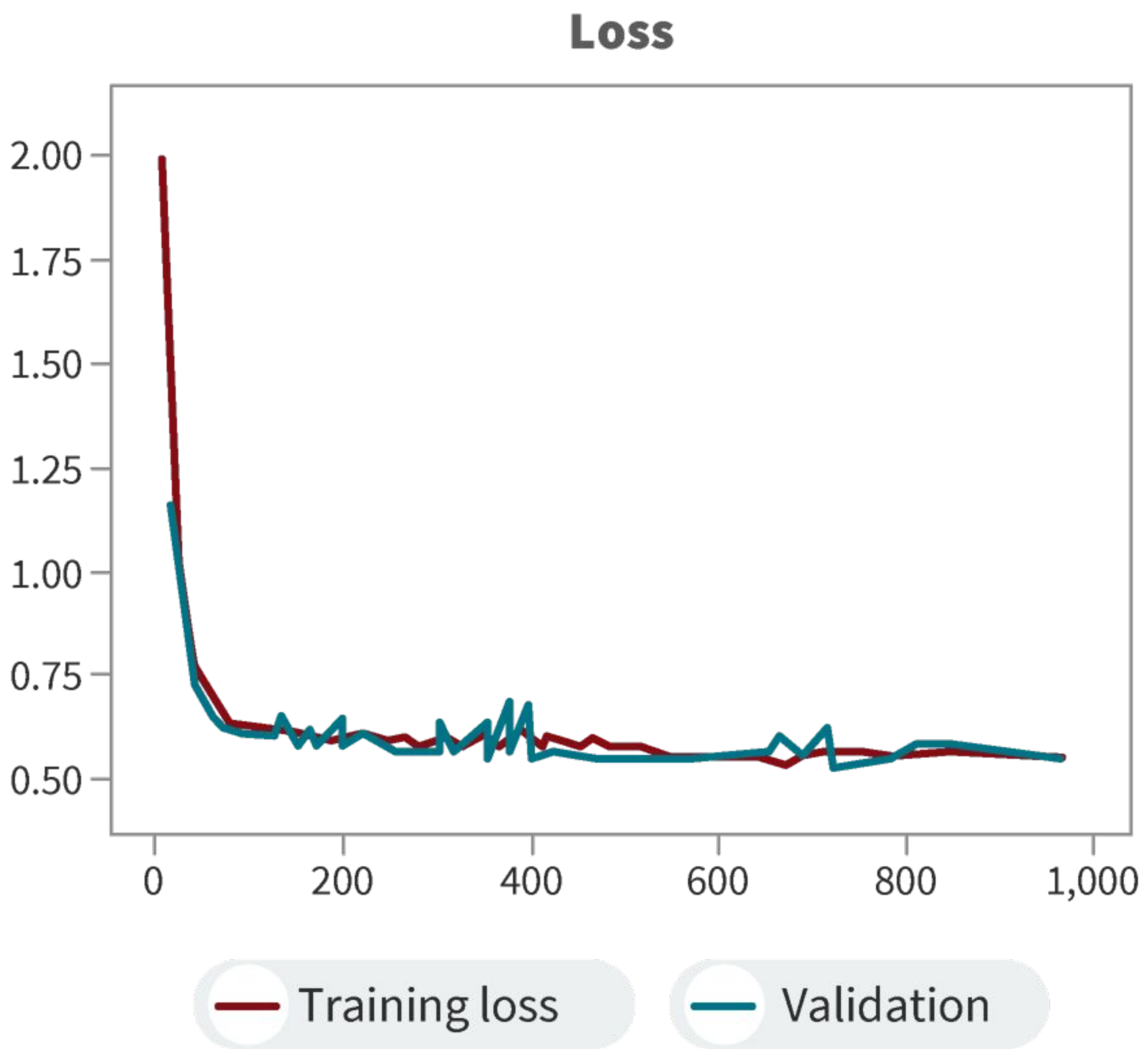
OVERFITTING

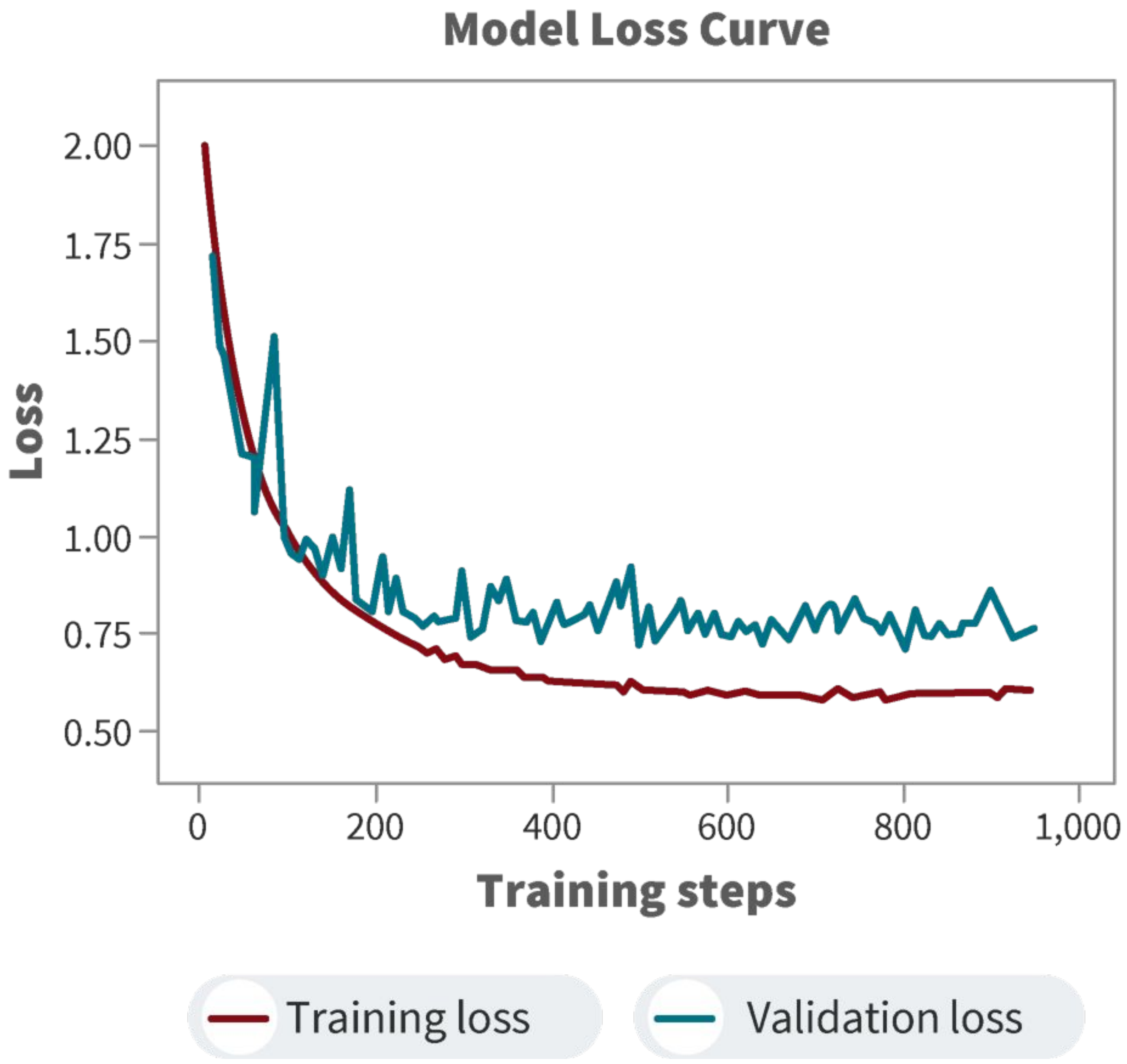


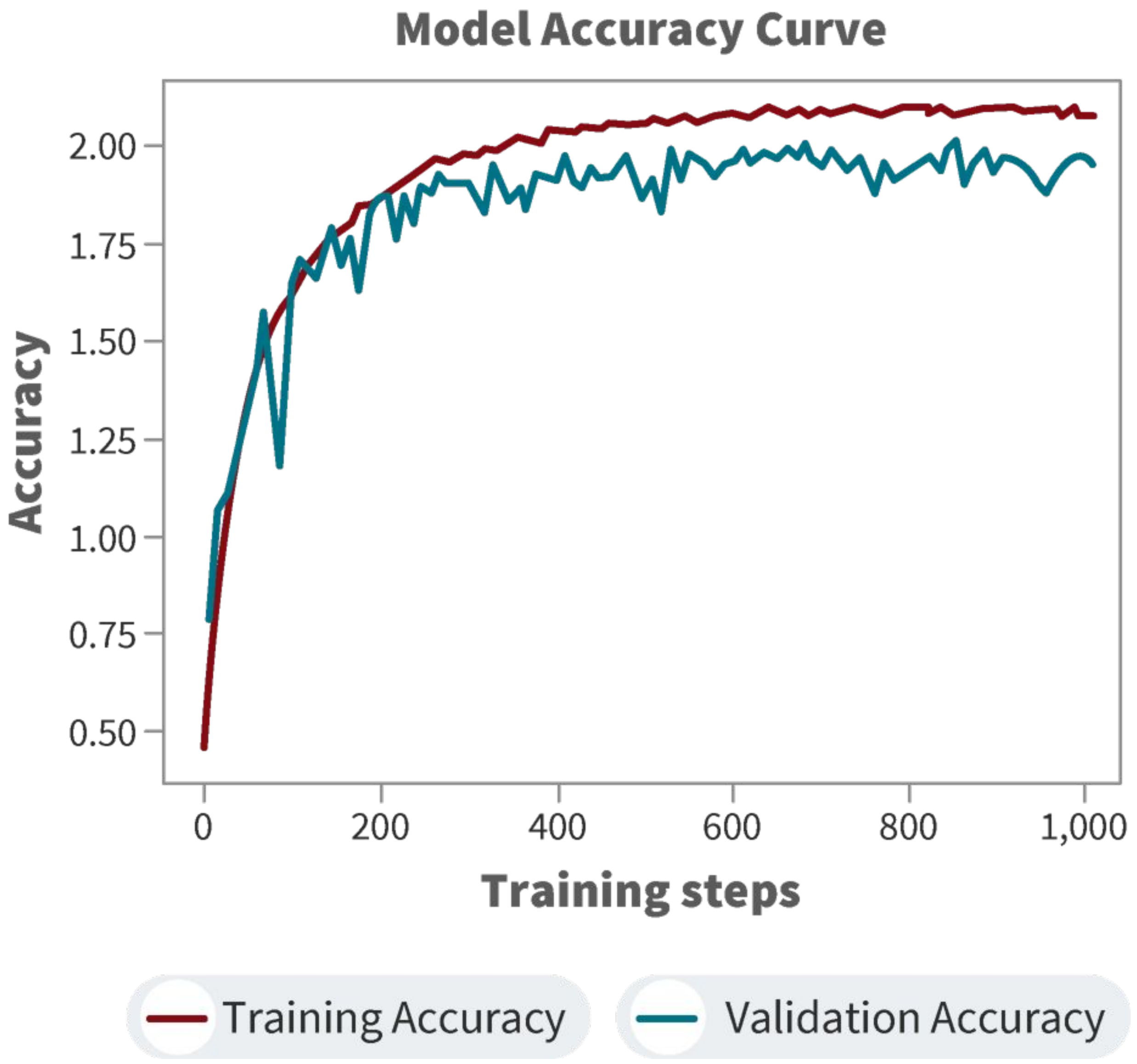
Validation loss increases after overfitting

Training loss continues to decrease

GOOD FIT

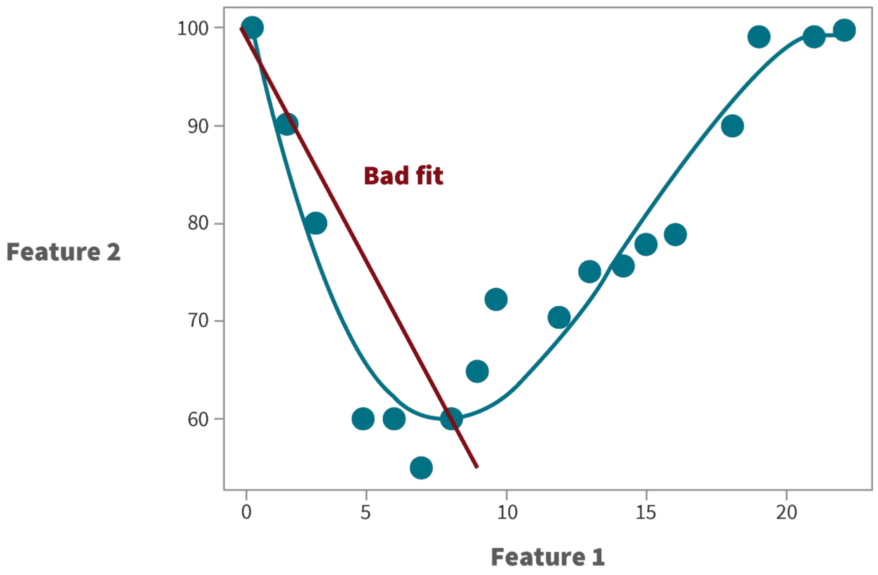

Loss
2.00
1.75
1.50
1.25
1.00
0.75
0.50
0
200
400
600
800
1,000
Training loss
Validation

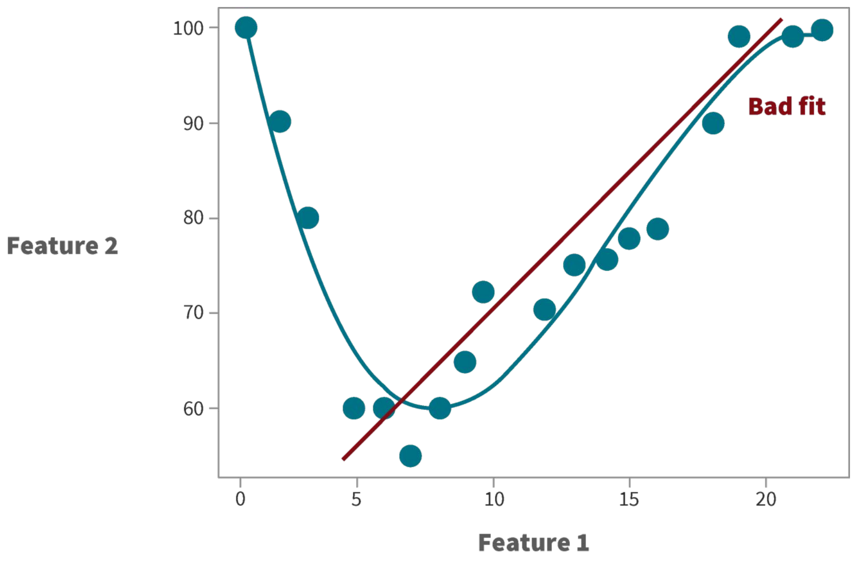

Model Loss Curve
Loss
2.00
1.75
1.50
1.25
1.00
0.75
0.50
0
200
400
600
800
1,000
Training steps
Training loss
Validation loss

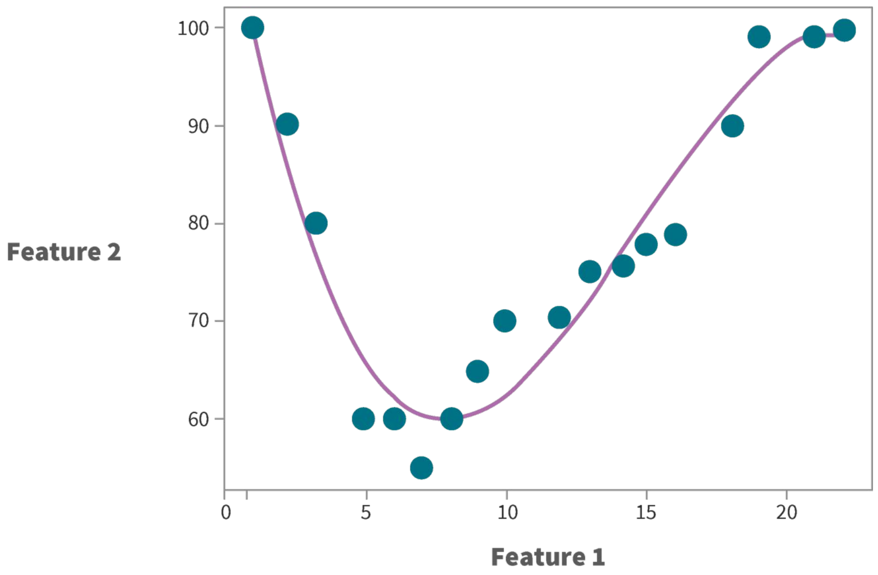

Model Accuracy Curve
Accuracy
2.00
1.75
1.50
1.25
1.00
0.75
0.50
0
200
400
600
800
1,000
Training steps
Training Accuracy
Validation Accuracy

Outline

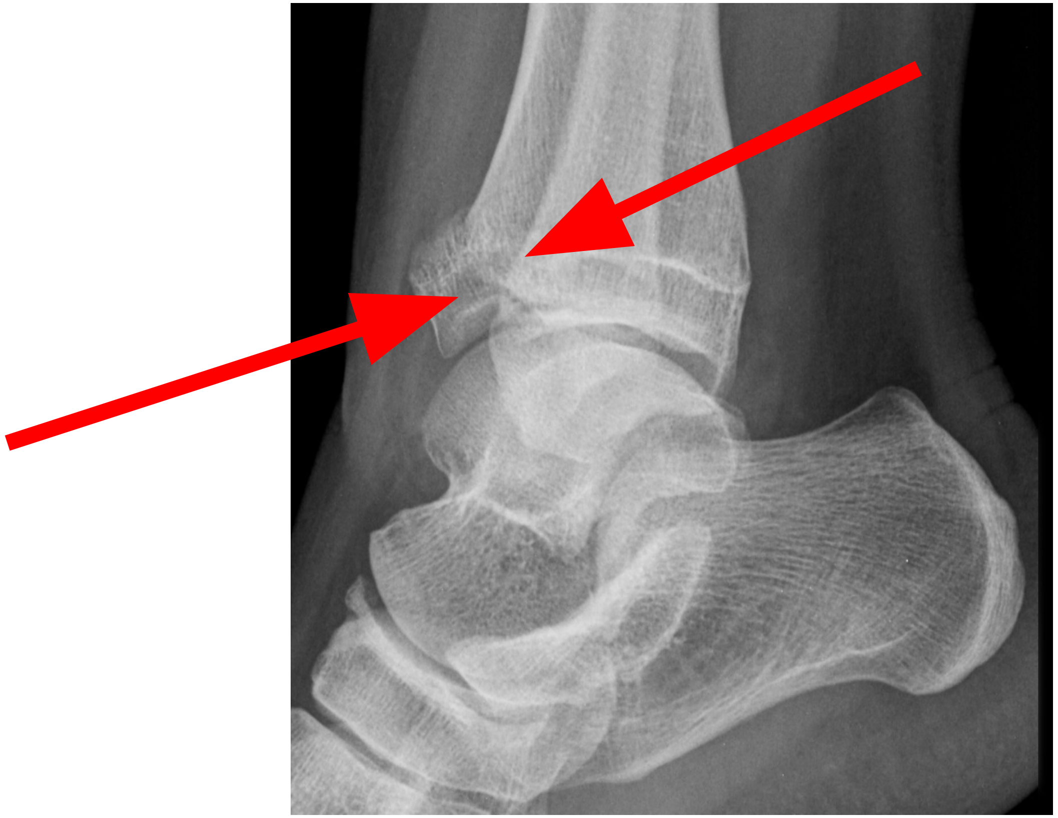
1. Correlation vs. Causation
2. Splitting your data
3. Underfitting vs overfitting
4. **Strategies to address underfitting and overfitting**


Feature 2
100
90
80
70
60
Bad fit
0
5
10
15
20
Feature 1

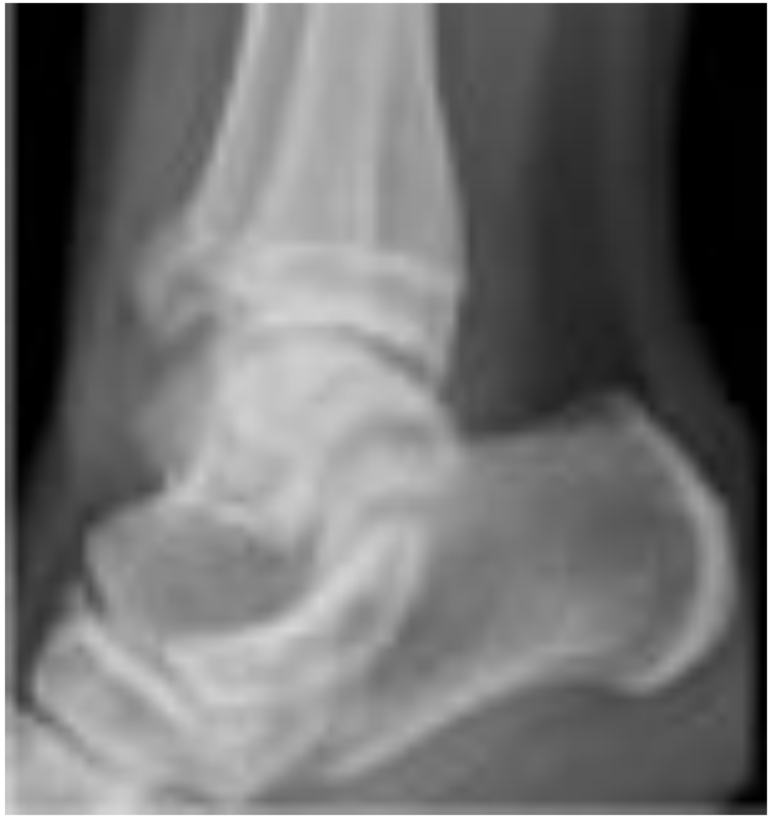

Feature 2
100
90
80
70
60
Bad fit
0
5
10
15
20
Feature 1


Feature 2
100
90
80
70
60
0
5
10
15
20
Feature 1

Debugging tip: sanity check to make sure training data includes the information required to make the right decisions



Full Resolution



Downsampled

Strategies to address underfitting

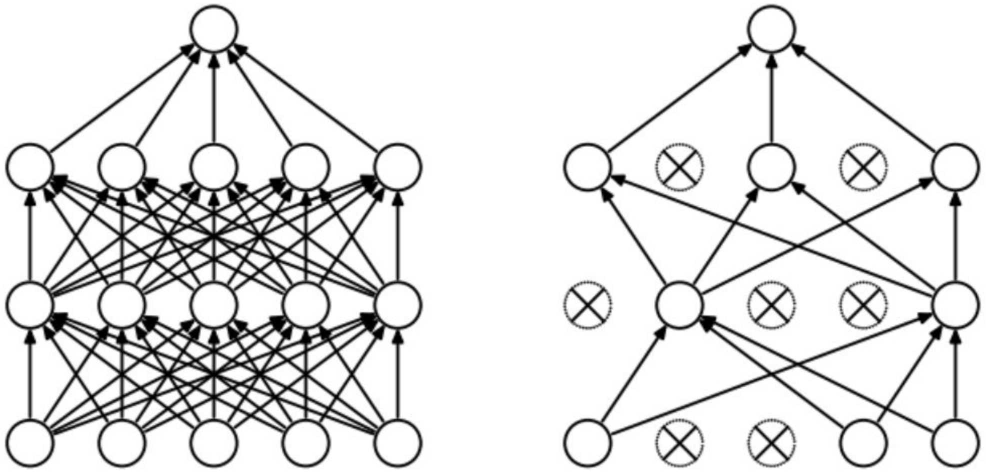
- Train your model for more time
- Increase the capacity of your model (use a neural network or make it bigger and deeper)

Strategies to address overfitting

- **Weight decay (i.e. L1/L2 regularization):** penalizing the model through the loss function for using too many or too much of its parameters

Strategies to address overfitting

- **Weight decay (i.e. L1/L2 regularization):** penalizing the model through the loss function for using too many or too much of its parameters
- **Dropout:** randomly setting parameter values to zero during model training, such that the model has to build in redundancy and cannot be as complex



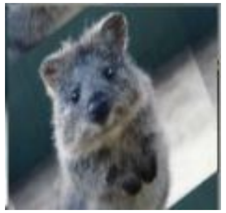
(a) Standard Neural Net

(b) After applying dropout.

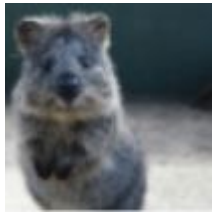
Strategies to address overfitting

- **Weight decay (i.e. L1/L2 regularization):** penalizing the model through the loss function for using too many or too much of its parameters
- **Dropout:** randomly setting parameter values to zero during model training, such that the model has to build in redundancy and cannot be as complex
- **Data augmentation:** randomly warping or transforming samples in the training set to encourage the model to learn more generalizable features.

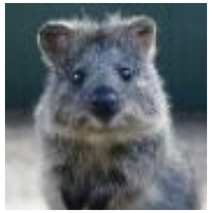
Original



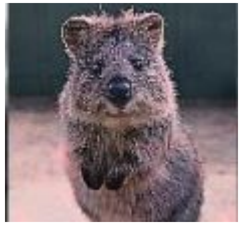
Rotation



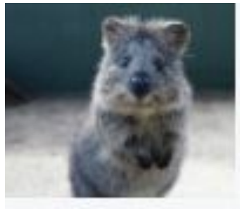
Random crops



Resizing



Color / brightness



Flipping

Summary

Today we covered:

- Correlation vs. Causation
- Splitting your data
- Underfitting vs overfitting

Coming up next time: **Data considerations for clinical machine learning**

Outline

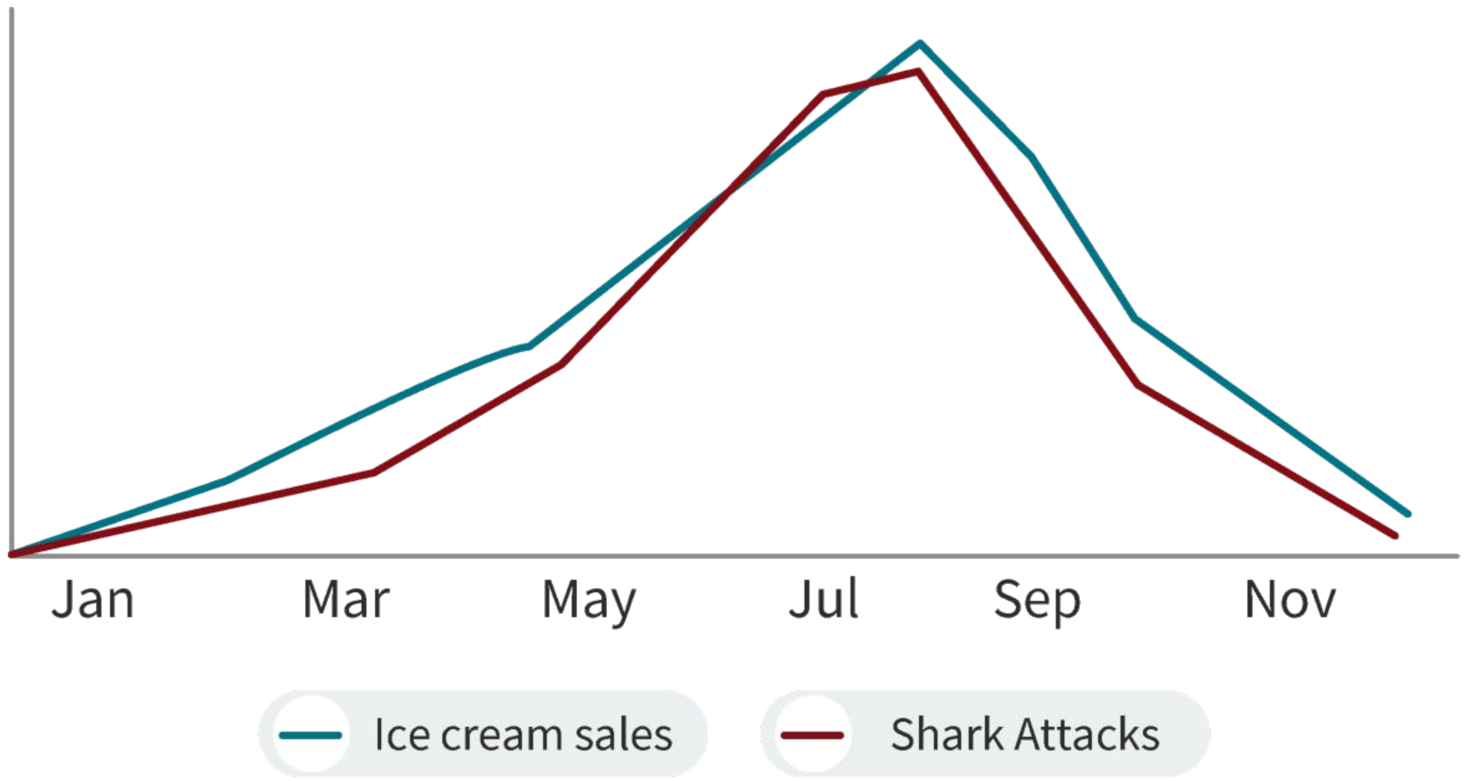
1. Splitting your data
2. Underfitting vs overfitting
3. Strategies to address underfitting and overfitting
4. **Correlation vs. Causation**

Some definitions

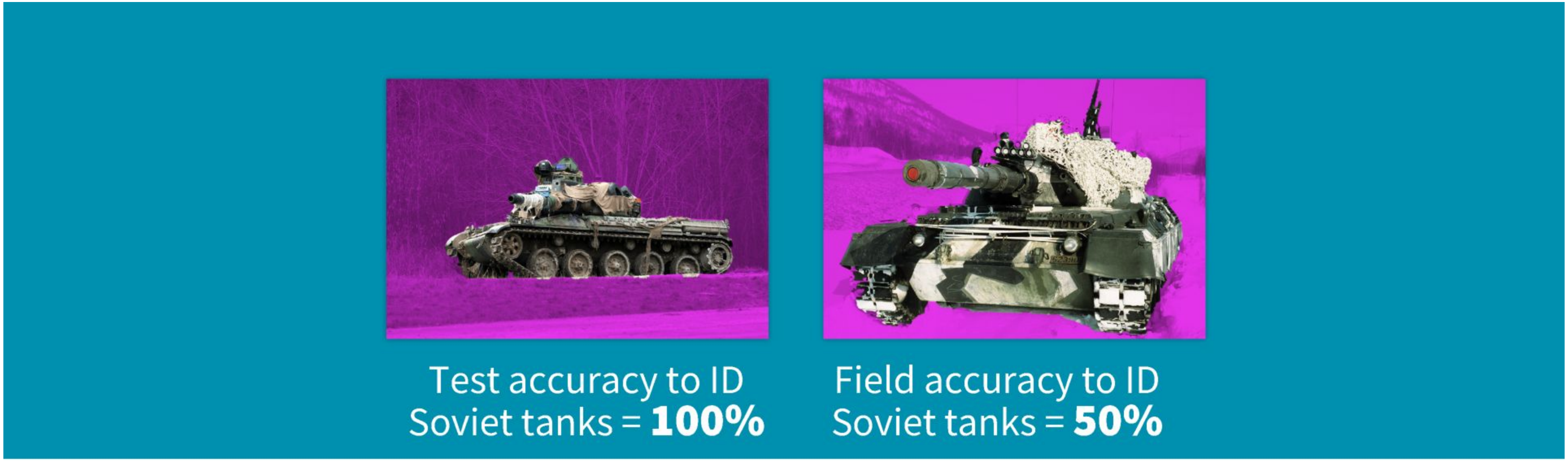
- **Correlation**: The degree to which two events/variables are (linearly) related. The relationship can be causal or non-causal.

- **Causation**: One event/variable directly influences the other event/variable.

CORRELATION IS NOT CAUSATION!

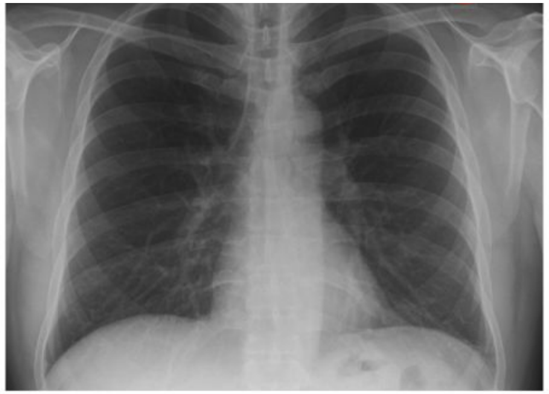


Both ice cream sales and shark attacks increase when the weather is hot and sunny, but they are not caused by other (they are caused by good weather, with lots of people at the beach, both eating ice cream and having a swim in the sea)

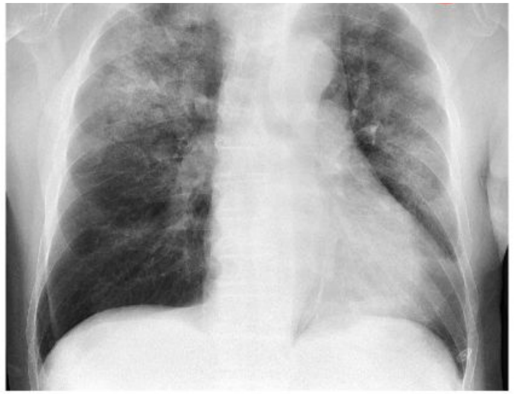


Test accuracy to ID Soviet tanks = **100%**

Field accuracy to ID Soviet tanks = **50%**



− Pneumonia

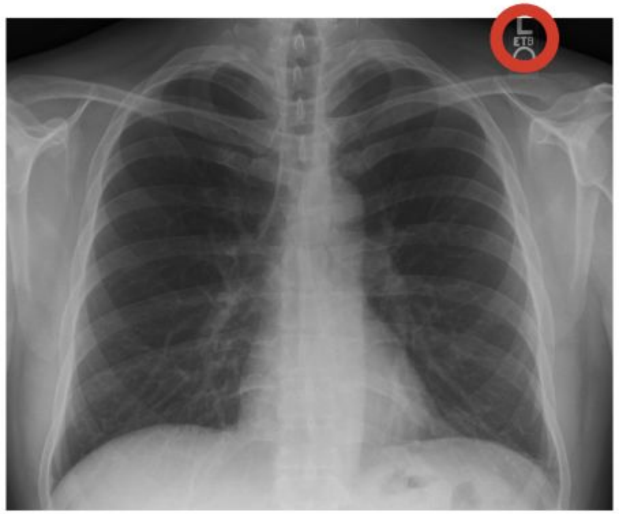


+ Pneumonia

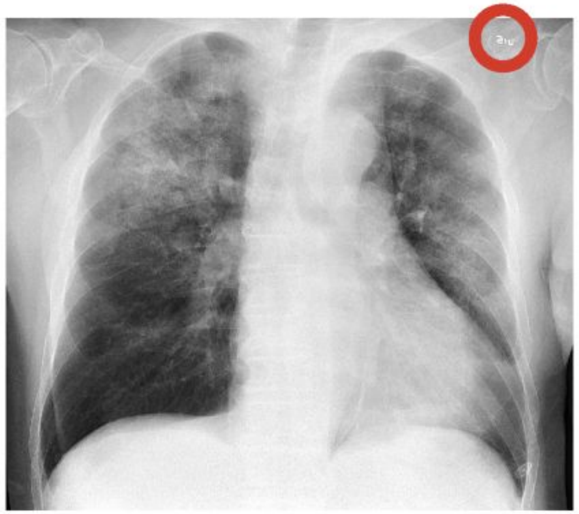
What went wrong?

HOSPITAL A

1% pneumonia prevalence



− Pneumonia



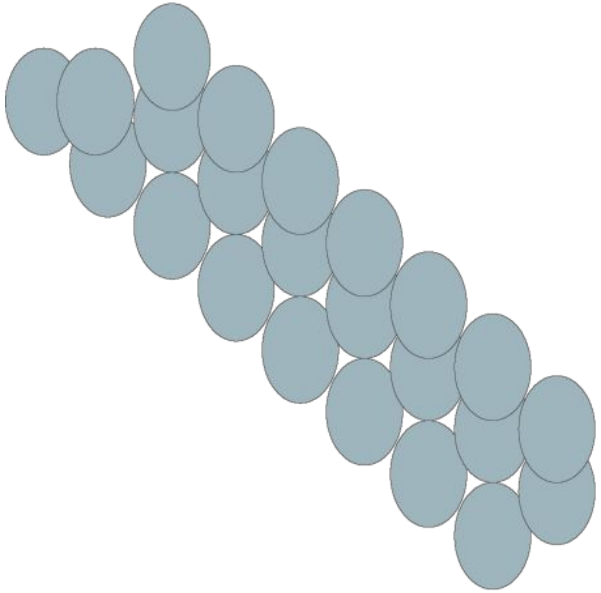
+ Pneumonia

HOSPITAL B

34% pneumonia prevalence

- Can you think of a scenario where a model that focuses on medically irrelevant correlations to make accurate predictions could be useful instead of useless?

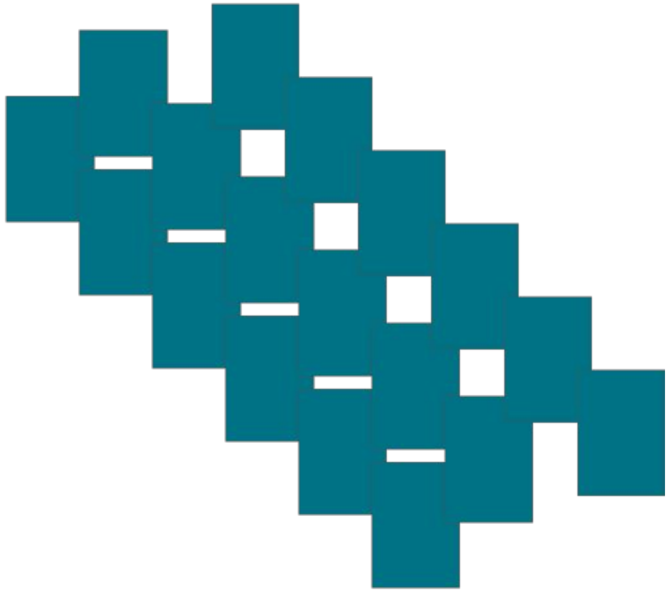
- How would that be possible?



GROUP 1

+

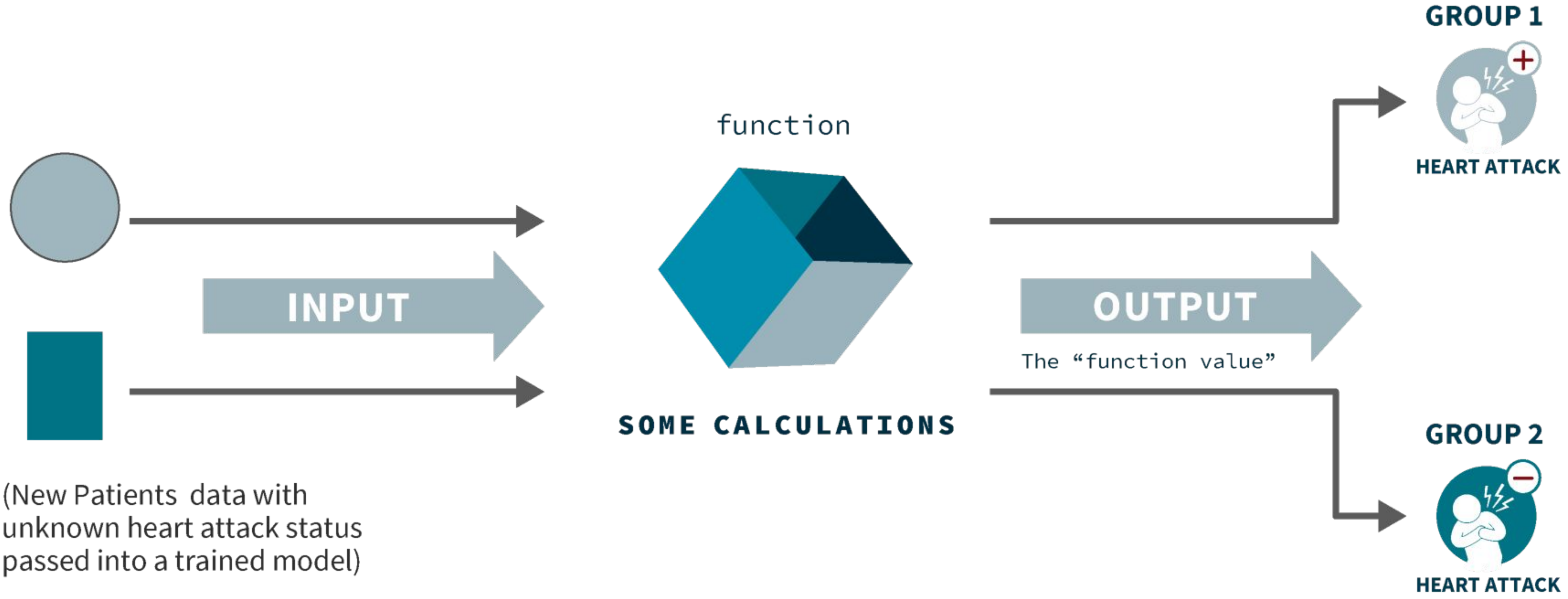
HEART ATTACK



GROUP 2

−

HEART ATTACK



INPUT

function

SOME CALCULATIONS

OUTPUT

The “function value”

GROUP 1

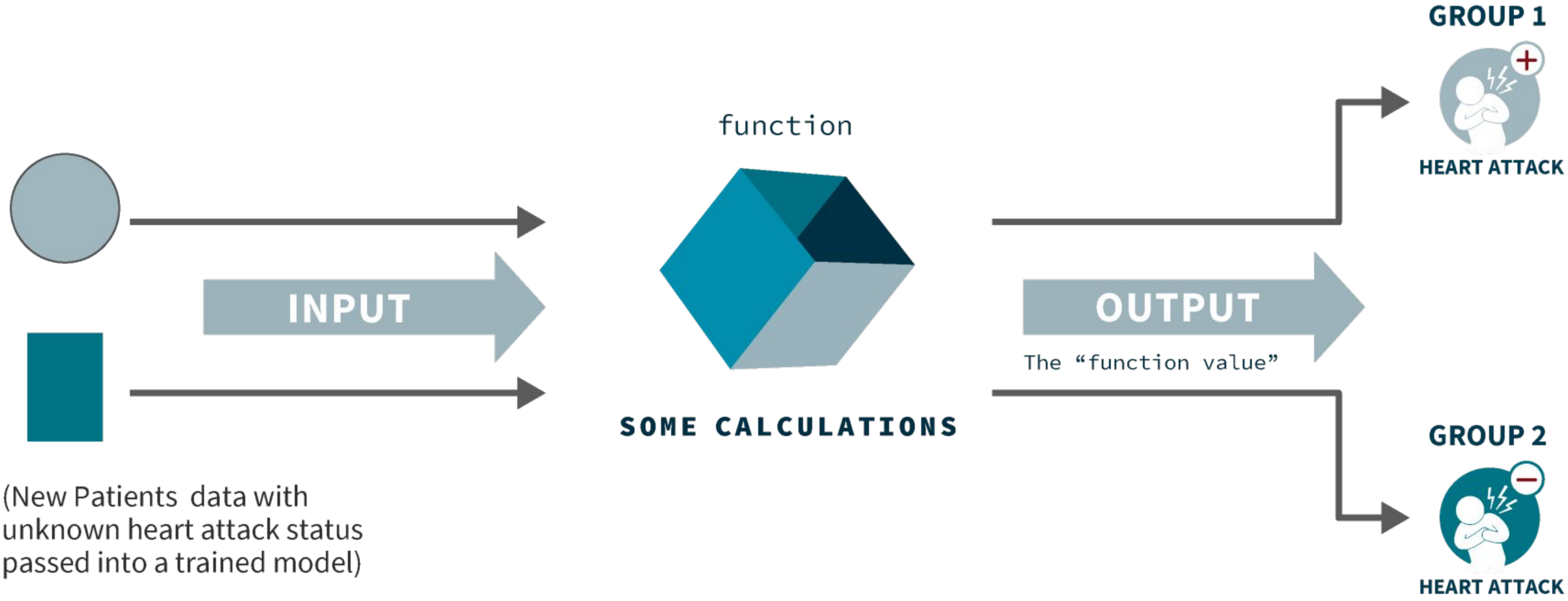
HEART ATTACK

GROUP 2

HEART ATTACK

(New Patients data with unknown heart attack status passed into a trained model)

(The new model correctly classified them - but used the correlation of patient hair color)



(New Patients data with unknown heart attack status passed into a trained model)

(The new model correctly classified them - but used the correlation of patient hair color)

If this model was used for population health management planning rather than medical diagnosis or intervention the correlation (if accurate) of hair color and future heart attack risk can lead to a useful model in this context of risk planning or budget adjustment

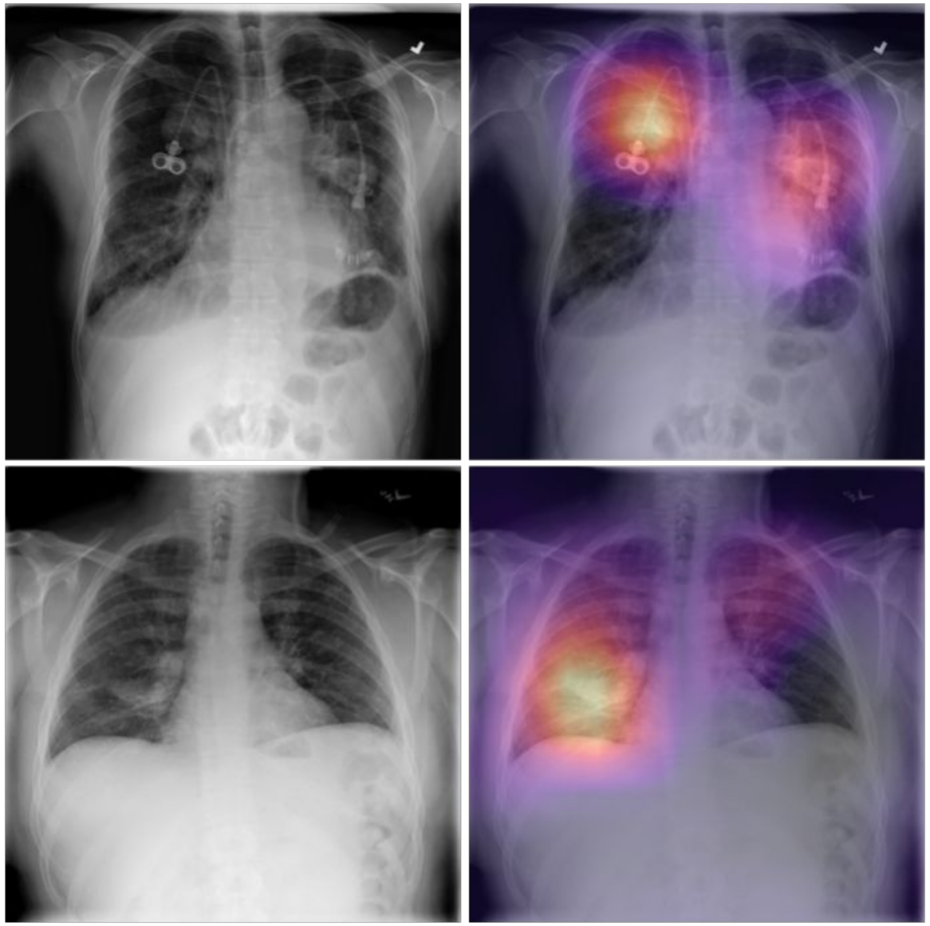
- **Take home point**: In the properly constructed context an accurate model does not have to have causal output to be useful!

- Question: Can you think of ways to figure out whether the model is relying solely on spurious correlations rather than causative factors?

- Perhaps you have seen examples in the literature?

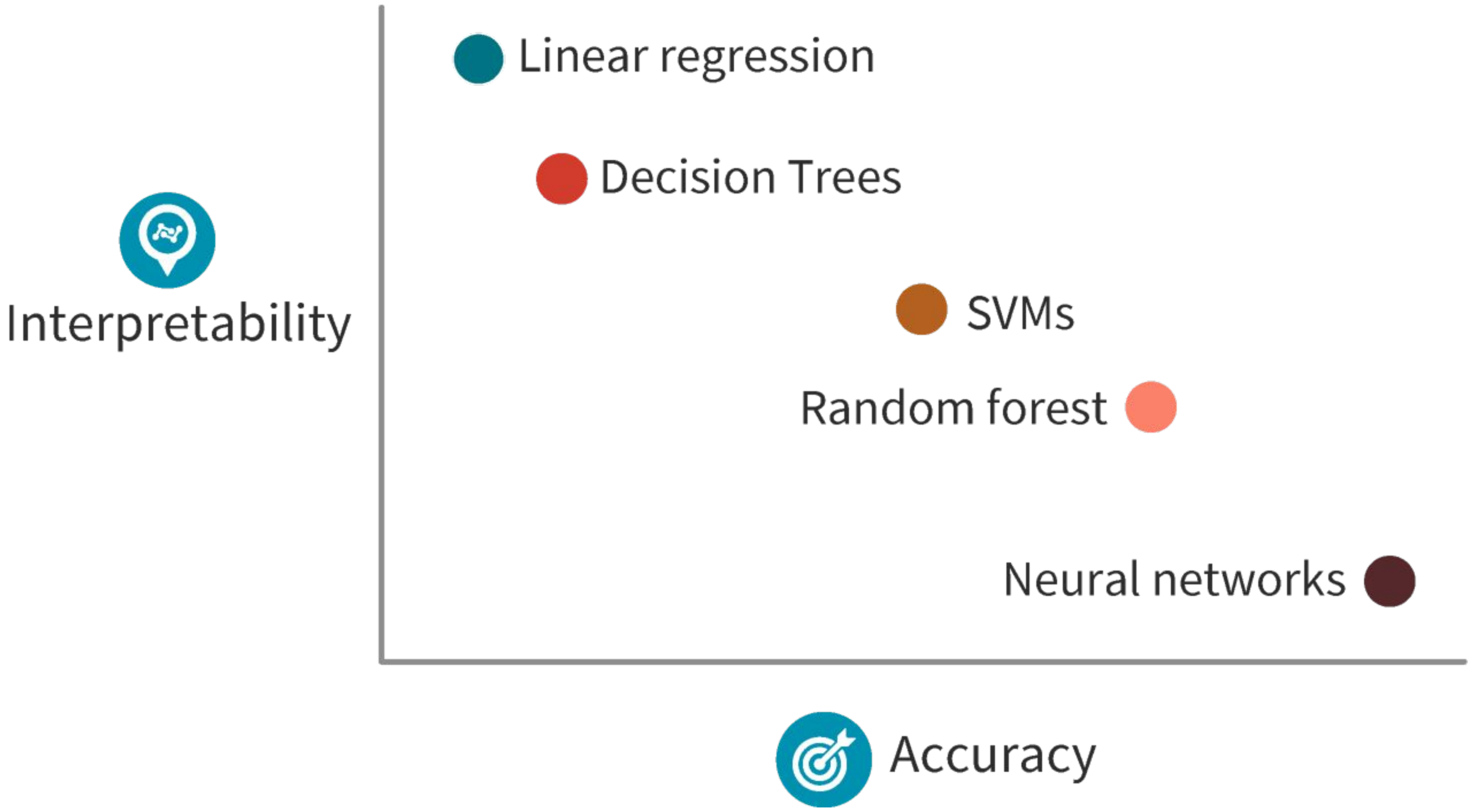
- ML models lack common sense (why we need teams with domain experts!)

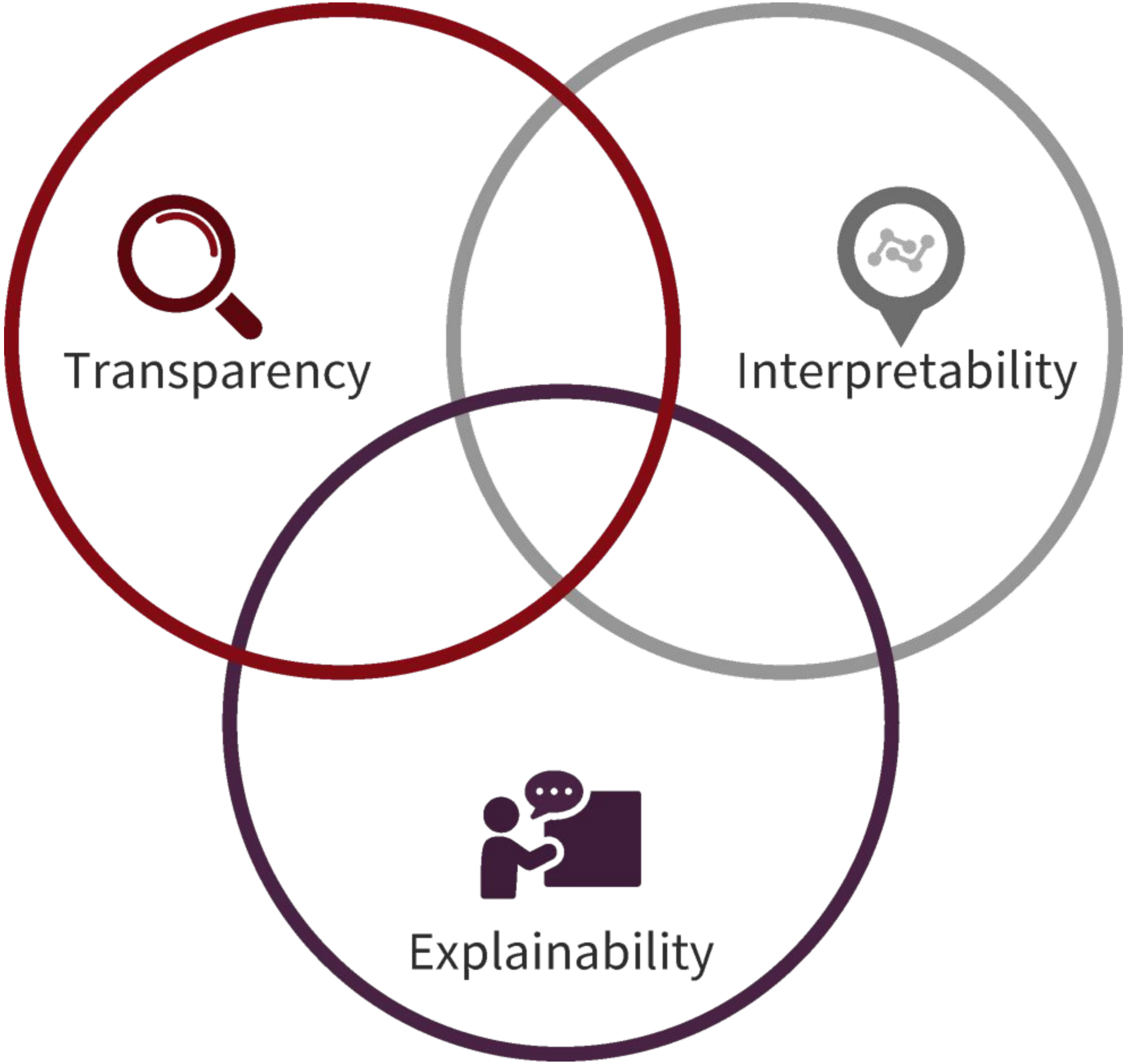
- Strategies include multi-disciplinary teams reviewing false positive and false negative cases predicted by the model, and testing the model on external datasets, to try to gain insight into causal vs. correlative features learned by the model

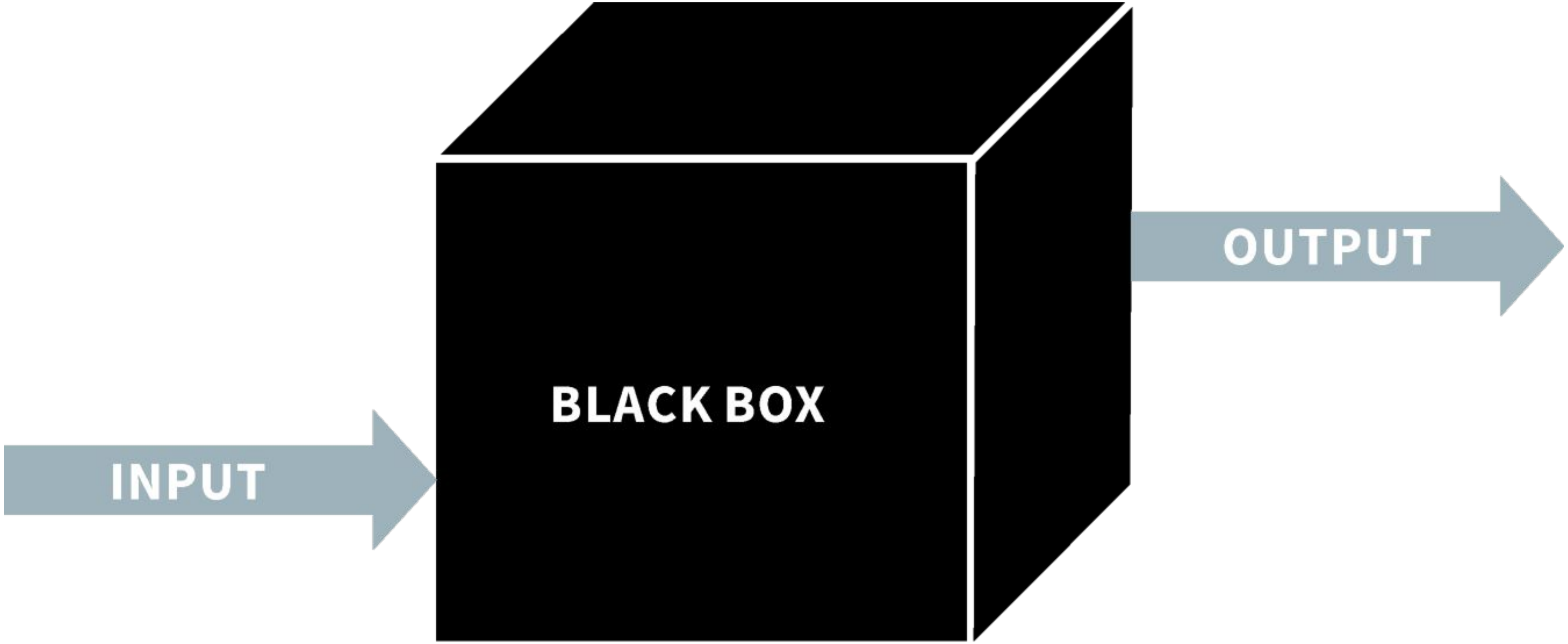


Rajpurkar, Irvin et al. CheXNeXt: Deep learning for chest radiograph diagnosis

TENSION BETWEEN BLACK BOX AND INTERPRETABLE ALGORITHMS



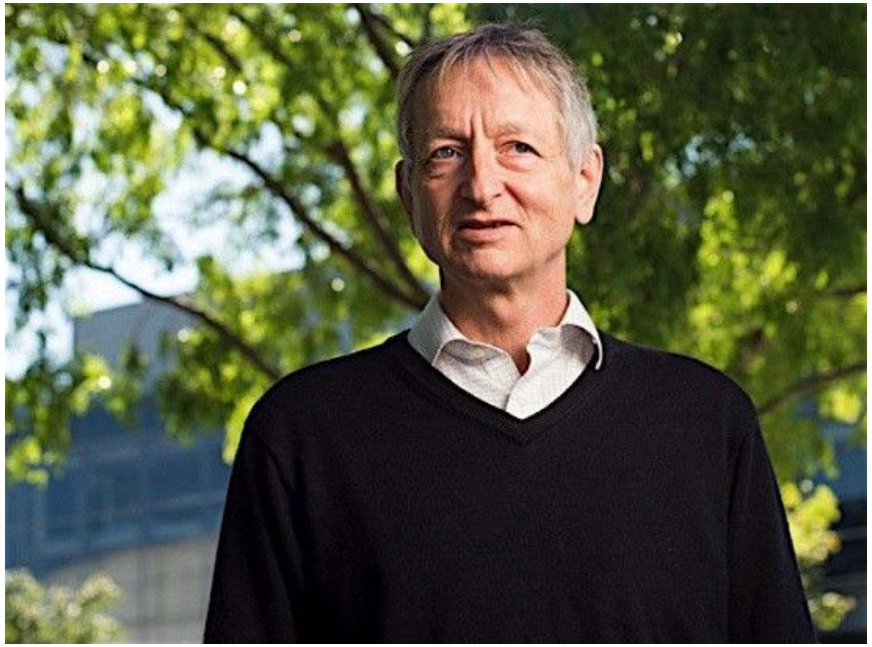

Transparency
Interpretability
Explainability



Internal behavior of the code is unknown

- *“Clinicians and regulators should not insist on explainability because people can't explain how they work for most of the things they do”*

(Geoffrey Hinton)



Machine Learning Model Explainability

- **Intrinsic**

  Intrinsic interpretability is simply referring to models, often simple models, that are self-explanatory from the start.

- **Post-hoc explainability**

  Post-hoc interpretability, which is used to understand decisions by complex models that do not have prescriptive declarative knowledge representations or features

LACE Index

- The LACE index predicts 30-day hospital readmission risk and is calculated using the following 4 intuitive and transparent feature inputs:

1. Length of current admission
2. Admission acuity
3. Patient comorbidities
4. No. of emergency department visits in the past 6 months.

L - Length of Stay

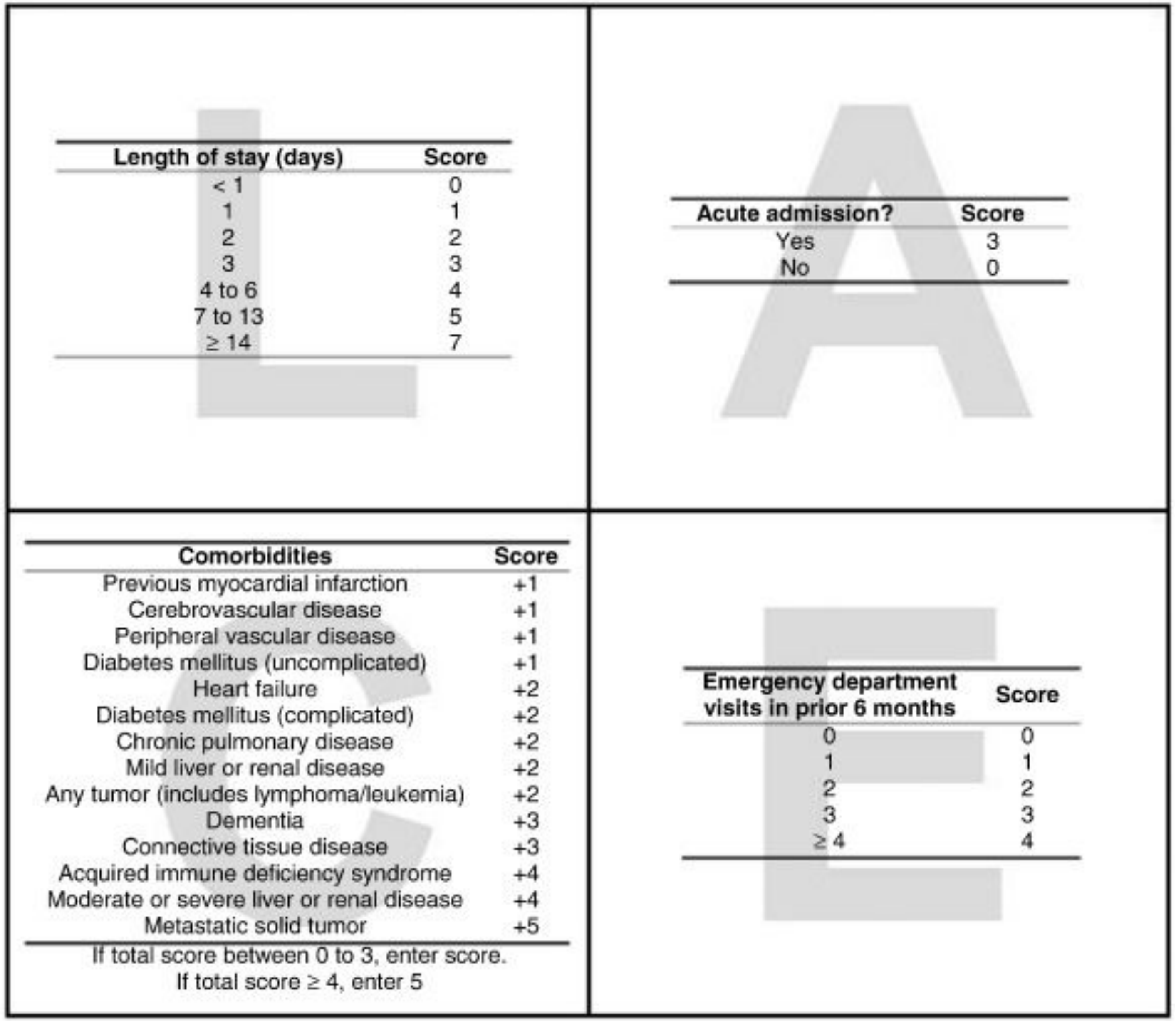
A - Acuity of Admission

C - Comorbidities

E - Emergency room visits

LACE Index

- What kind of interpretability does the LACE index have?



Length of stay (days)	Score
< 1	0
1	1
2	2
3	3
4 to 6	4
7 to 13	5
≥ 14	7

Acute admission?	Score
Yes	3
No	0

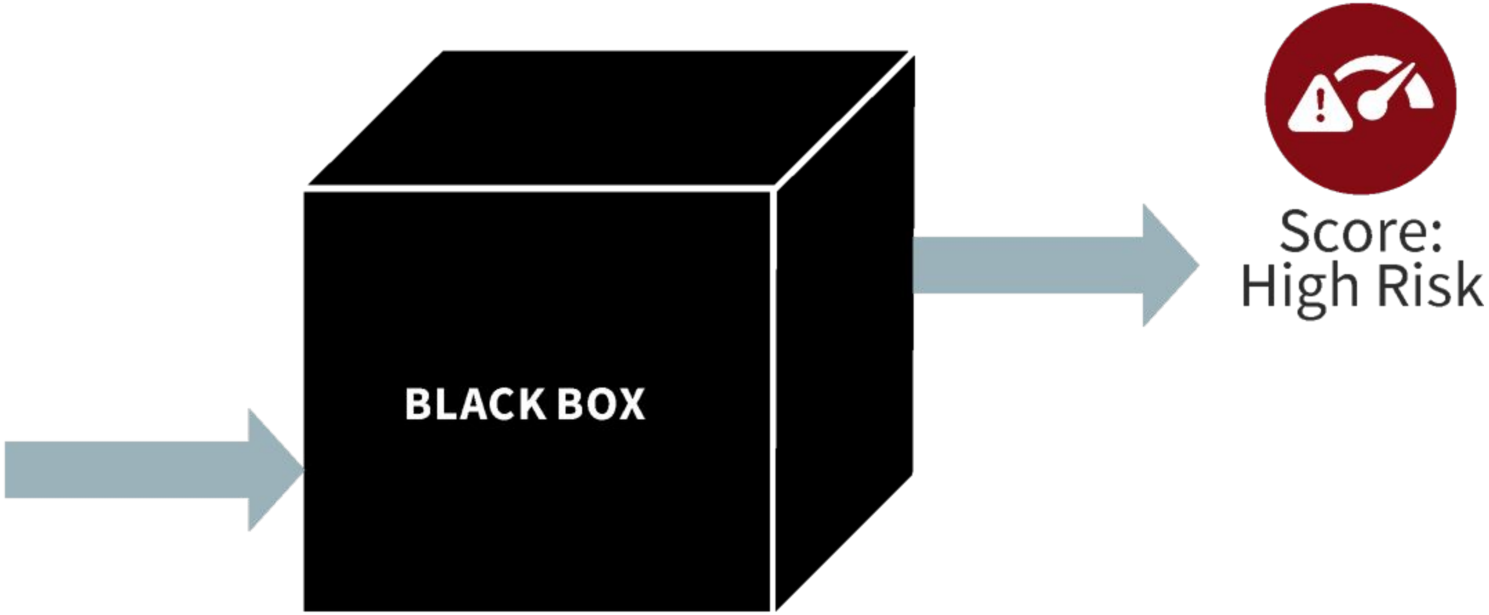
Comorbidities	Score
Previous myocardial infarction	+1
Cerebrovascular disease	+1
Peripheral vascular disease	+1
Diabetes mellitus (uncomplicated)	+1
Heart failure	+2
Diabetes mellitus (complicated)	+2
Chronic pulmonary disease	+2
Mild liver or renal disease	+2
Any tumor (includes lymphoma/leukemia)	+2
Dementia	+3
Connective tissue disease	+3
Acquired immune deficiency syndrome	+4
Moderate or severe liver or renal disease	+4
Metastatic solid tumor	+5

If total score between 0 to 3, enter score.
If total score ≥ 4, enter 5

Emergency department visits in prior 6 months	Score
0	0
1	1
2	2
3	3
≥ 4	4

Clinical input features

Glasgow onsciousness scale (GCS)
Systolic Blood Pressure (SBP) (mmHg)
Pulse Rate (Beat/minute)
Respiratory rate
Oral temperature (°C)
O_2 Saturation (%)
Arterial HCO_3 (mM); Normal: 22-26 mM
Serum CO_2 Pressure; Normal: 35-45 mmHg
Arterial pH (7.35-7.45)
Serum Potassium (K) (meq/l); Normal:3.5-5
Serum Sodium (Na)(meq/l); Normal: 135-150
Hematocrite (%)
WBC Count (per mm)
Hemoglobin (g/dl)
Blood glucose level at admission (mg.ml) (70-110mg/dl)
Serum Calcium (mg/dl); Normal 8-10 mg/dl
Serum Magnesium (mg/dl); Normal: 1.8-3 mg/dl
Alanine aminotransferase (ALT) (U/L) (7-56 U/l)
Asparte aminotransferase (AST) (U/L)(5-35 U/L)
Total Bilirubing (mg/dl); Normal: 0.2-1.3 mg/dl
Serum creatinin (mg/dl)
Blood Urea nitroger (mg/dl)



Some key messages

- Both black box and transparent model performance should both be evaluated against existing standards of care on real-world data to evaluate effectiveness in their specific patient population.

- Black Box models (low model interpretability) are especially important to evaluate with empirical pilot testing; example prospective data, external data and then maybe a trial

- Clinicians should be educated on the benefits, risks, and limitations of a given clinical model based on the evaluation metrics.