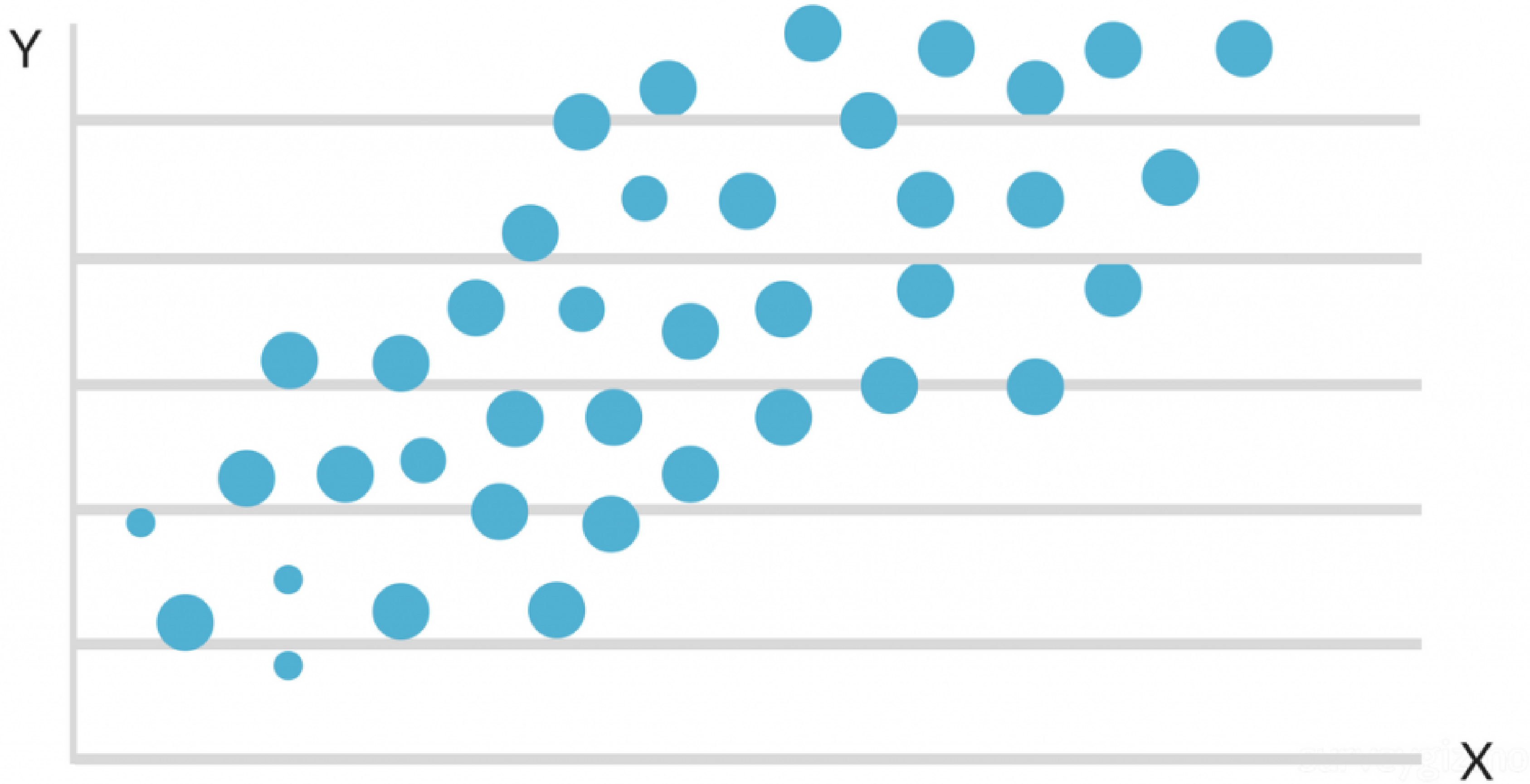
Regression models

PREPARING FOR STATISTICS INTERVIEW QUESTIONS IN PYTHON

Conor Dewey
Data Scientist, Squarespace

Getting started

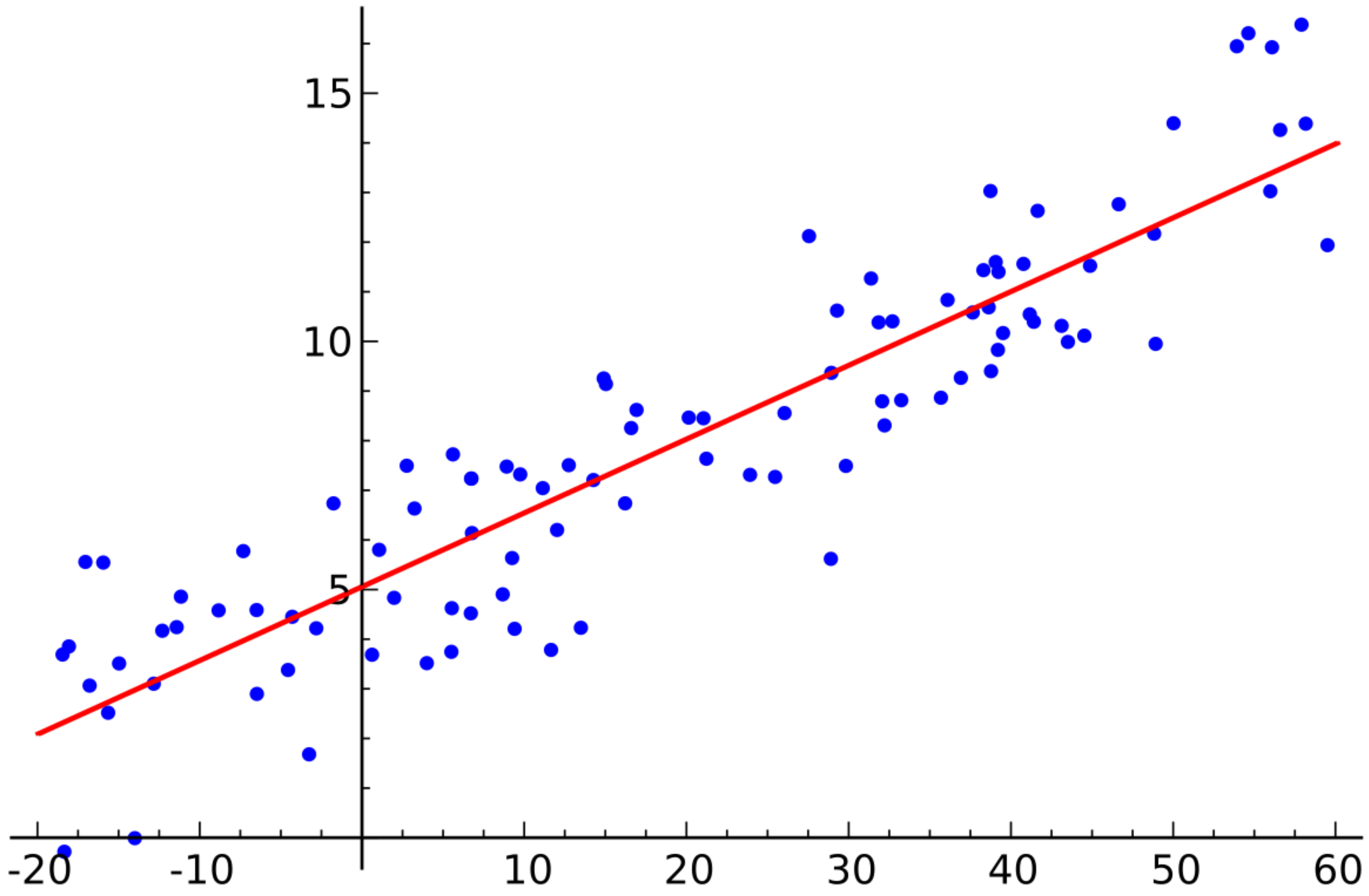


[1] Wikimedia

Assumptions

- Linear relationship
- Errors are normally distributed
- Homoscedasticity
- Independent observations

Linear regression



[1] Wikipedia

Linear regression

Population Y intercept

Population Slope Coefficient

Independent Variable

Random Error term

Dependent Variable

$$Y_i = \underbrace{\beta_0 + \beta_1 X_i}_{\text{Linear component}} + \underbrace{\varepsilon_i}_{\text{Random Error component}}$$

[1] ITS Surabaya

Example: linear regression

```
from sklearn.linear_model import LinearRegression
lm = LinearRegression()
lm.fit(X_train, y_train)
```

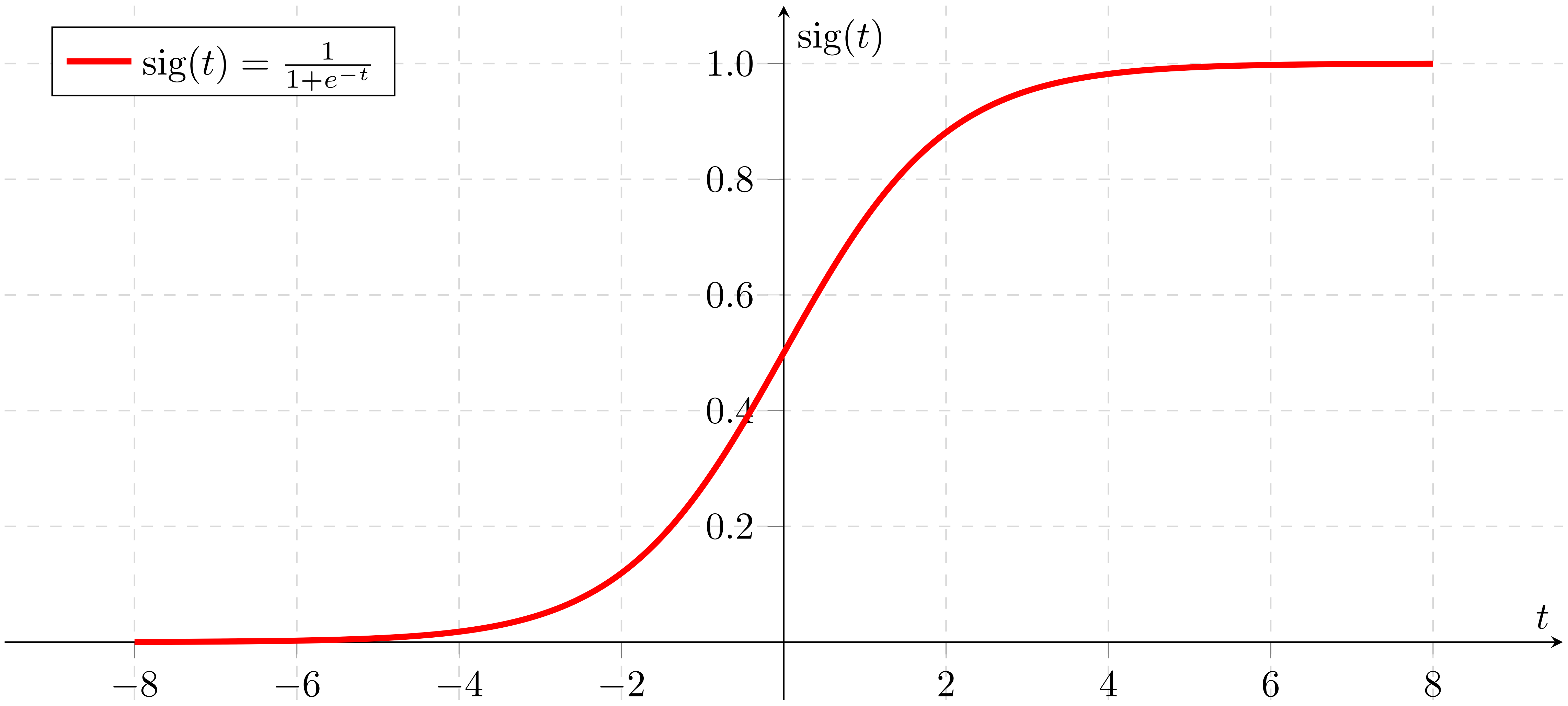
```
LinearRegression(copy_X=True, fit_intercept=True,
                 n_jobs=None, normalize=False)
```

Example: linear regression

```
coef = lm.coef_
print(coef)
```

```
[0.79086669]
```

Logistic regression



[1] Wikimedia

Logistic regression

$$f(x) = \frac{1}{1+e^{-(x)}}$$

Example: logistic regression

```
from sklearn.linear_model import LogisticRegression
clf = LogisticRegression(solver='lbfgs')
clf.fit(X_train, y_train)
```

```
LogisticRegression(C=1.0, class_weight=None,
                   dual=False, fit_intercept=True,
                   intercept_scaling=1,
                   max_iter=100, multi_class='warn',
                   n_jobs=None, penalty='l2',
                   random_state=None, solver='lbfgs',
                   tol=0.0001, verbose=0,
                   warm_start=False)
```

Example: logistic regression

```
coefs = clf.coef_
print(coefs)
```

```
[[0.4015177  3.85056451]]
```

```
accuracy = clf.score(X_test, y_test)
print(accuracy)
```

```
0.8583333333333333
```

Summary

- Review
- Assumptions
- Linear regression
- Logistic regression

Let's prepare for the interview!

PREPARING FOR STATISTICS INTERVIEW QUESTIONS IN PYTHON

Evaluating models

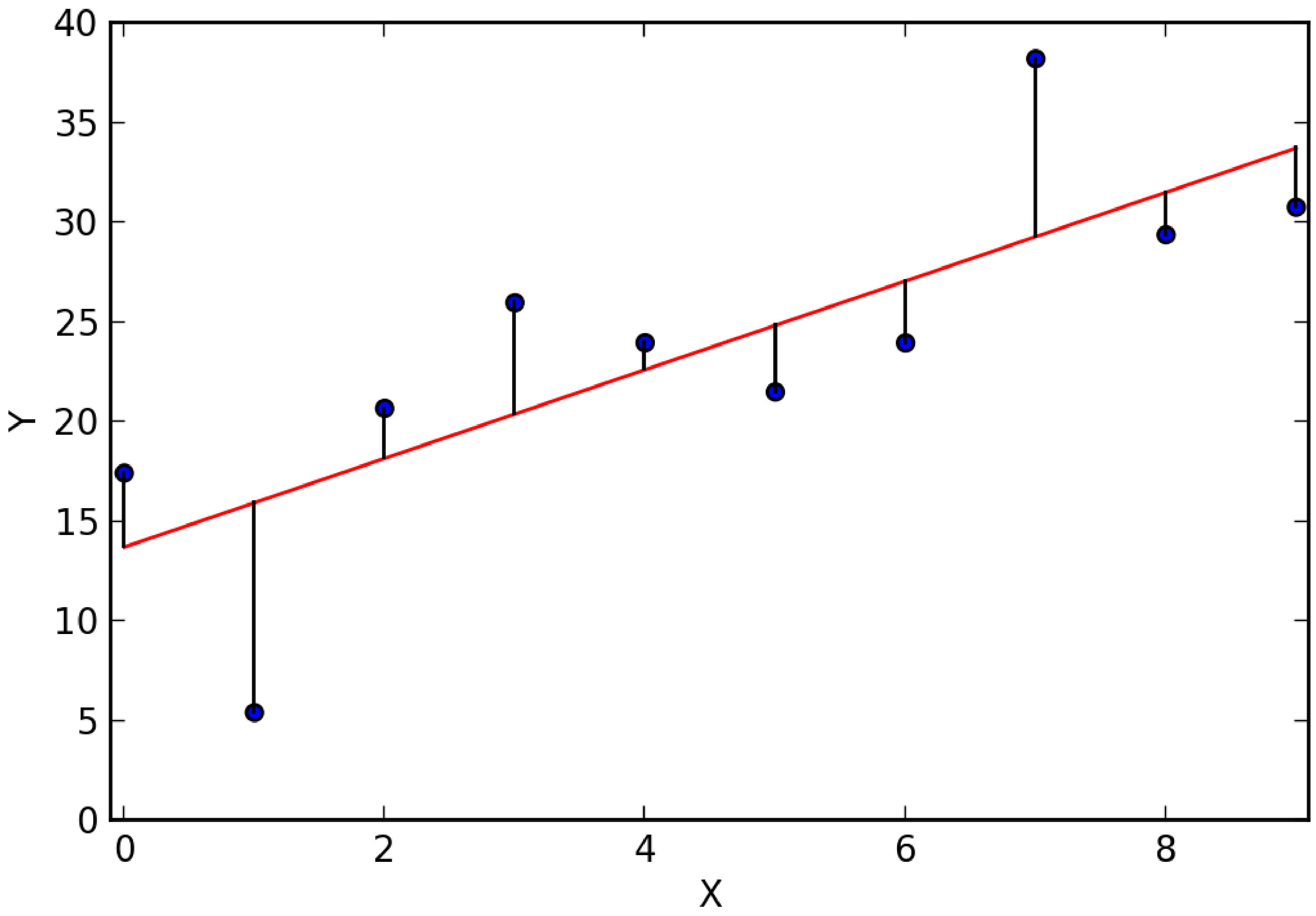
PREPARING FOR STATISTICS INTERVIEW QUESTIONS IN PYTHON

Conor Dewey
Data Scientist, Squarespace

Regression techniques

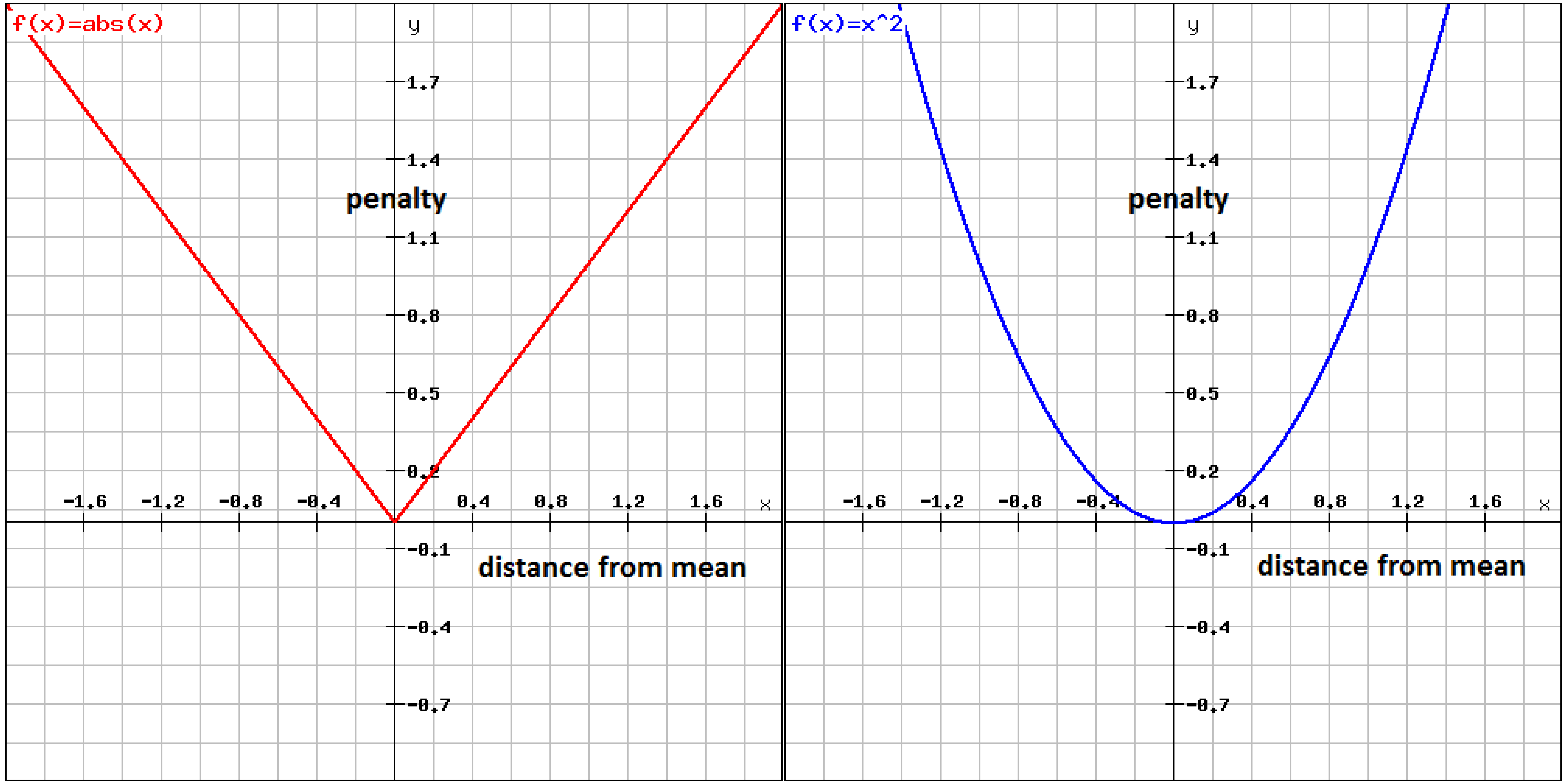
- R-squared
- Mean absolute error (MAE)
- Mean squared error (MSE)

R-squared



[1] Wikimedia

MAE vs. MSE



[1] Wikimedia

MAE vs. MSE

What are some differences you would expect in a model that minimizes squared error, versus a model that minimizes absolute error? In which cases would each error metric be appropriate?

[1] 120 Data Science Interview Questions

Classification techniques

- Precision
- Recall
- Confusion matrices

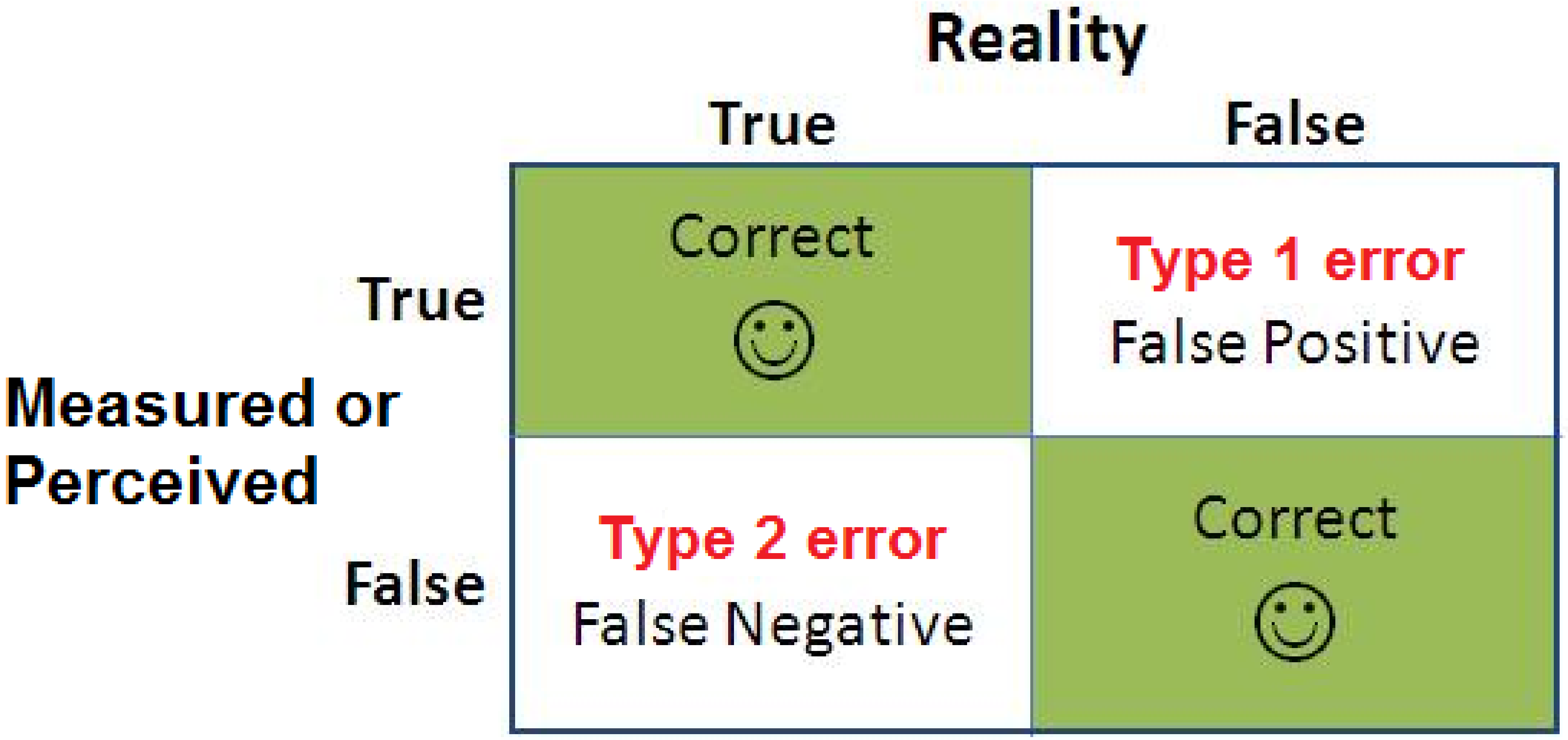
Precision

$$\text{Precision} = \frac{True\ Positive}{True\ Positive + False\ Positive}$$

Recall

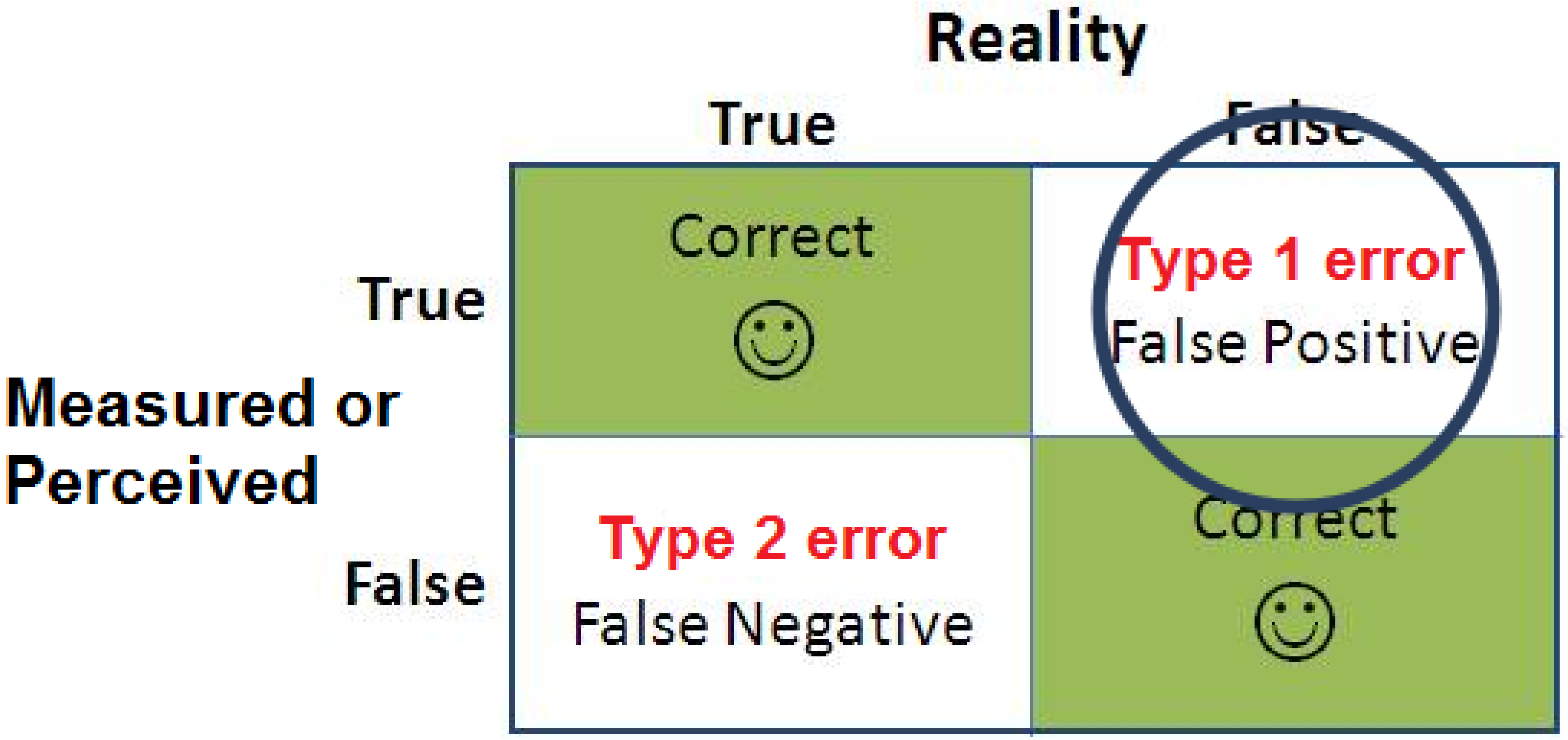
$$\text{Recall} = \frac{True\ Positive}{True\ Positive + False\ Negative}$$

Confusion matrix



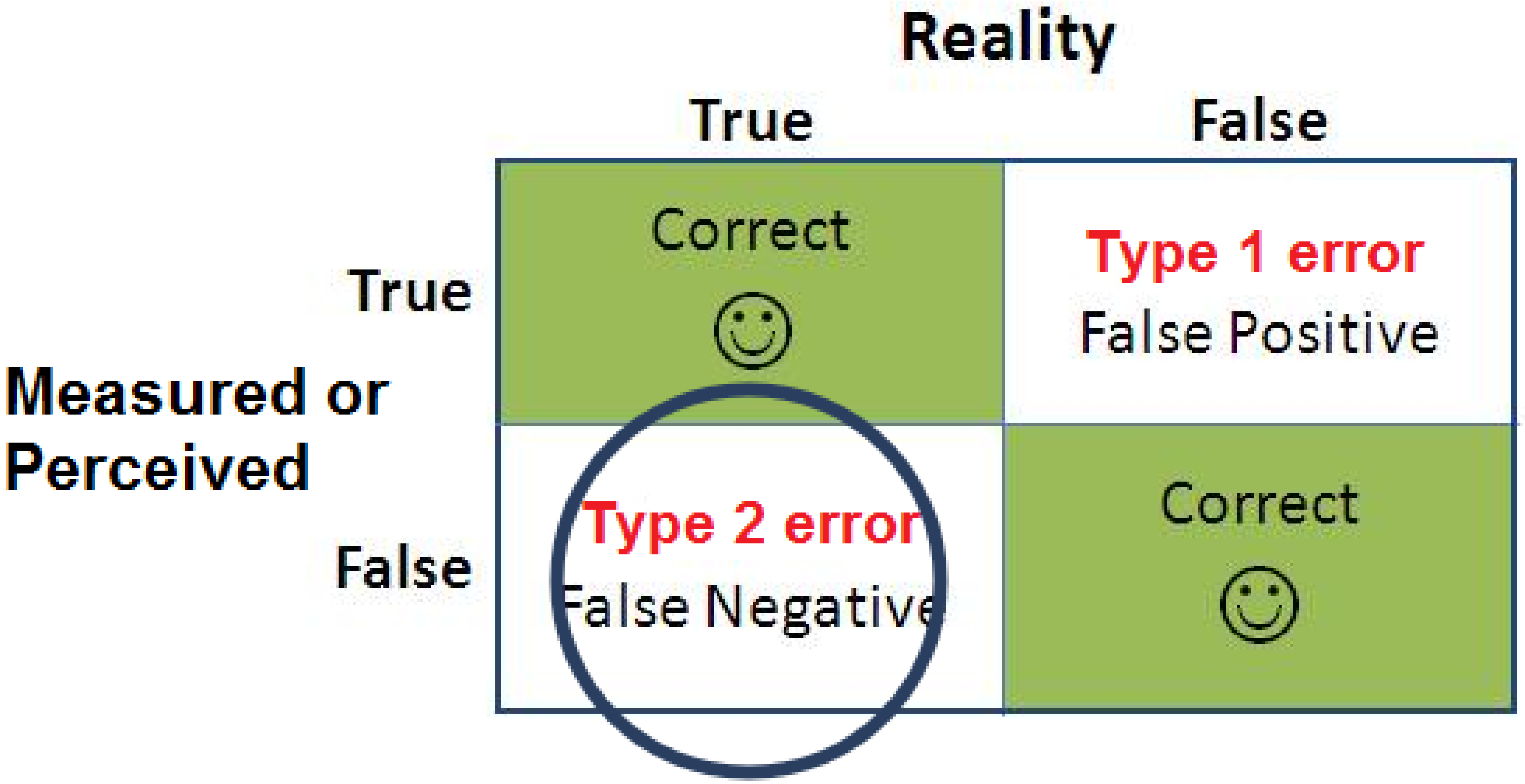
[1] AB Tasty

Confusion matrix



[1] AB Tasty

Confusion matrix



[1] AB Tasty

Summary

- R-squared
- Mean absolute error (MAE) vs. mean squared error (MSE)
- Precision and recall

Let's prepare for the interview!

PREPARING FOR STATISTICS INTERVIEW QUESTIONS IN PYTHON

Missing data and outliers

PREPARING FOR STATISTICS INTERVIEW QUESTIONS IN PYTHON

Conor Dewey
Data Scientist, Squarespace

Handling missing data

- Drop the whole row
- Impute missing values

Drop the whole row

```
df.dropna(inplace=True)
```

	Name	State	Gender	Score
0	George	Arizona	M	63
1	Andrea	Georgia	F	48
2	micheal	Newyork	M	56
3	maggie	Indiana	F	75
4	Ravi	Florida	M	NaN
5	Xien	California	M	77
6	Jalpa	NaN	NaN	NaN

Impute missing values

- Constant value
- Randomly selected record
- Mean, median, or mode
- Value estimated by another model

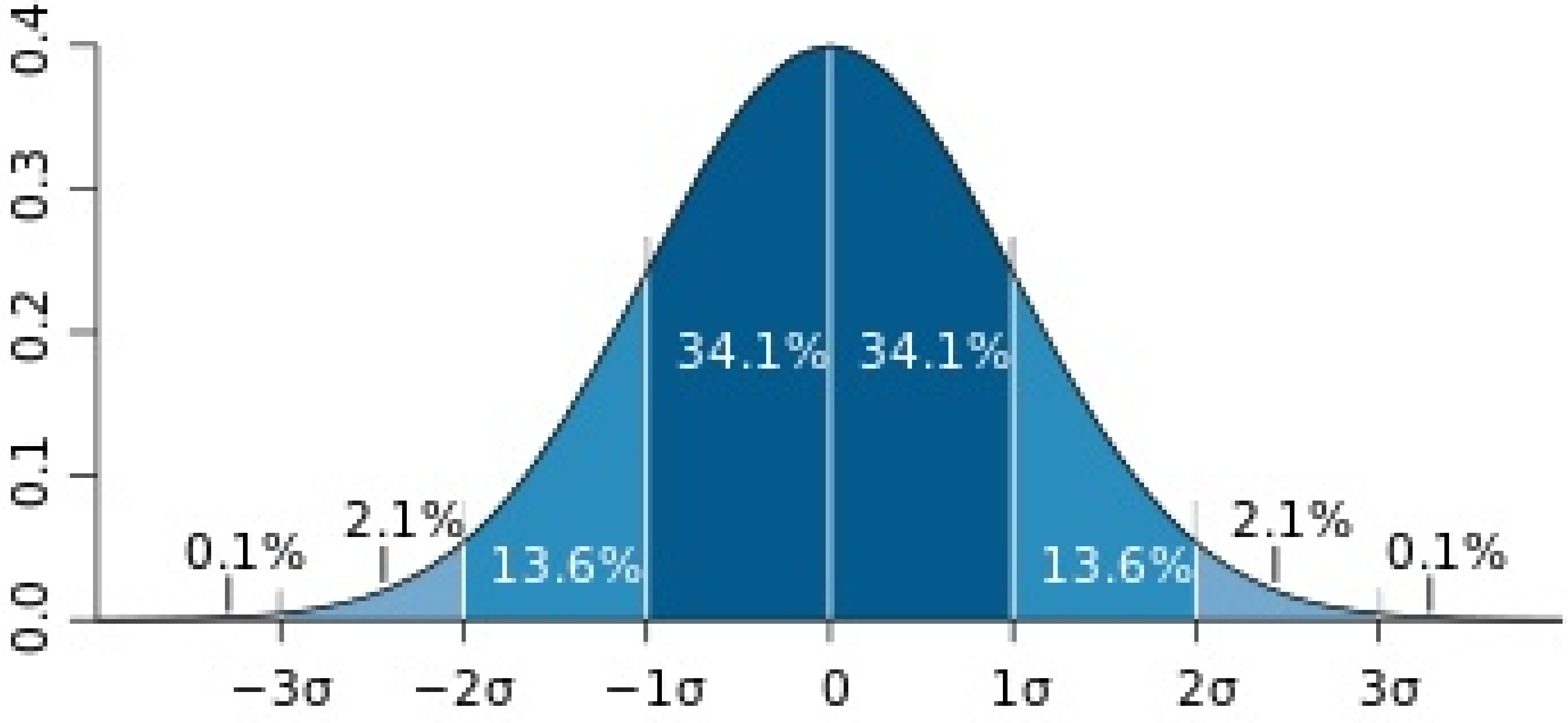
A few useful functions

- `isnull()`
- `dropna()`
- `fillna()`

Dealing with outliers

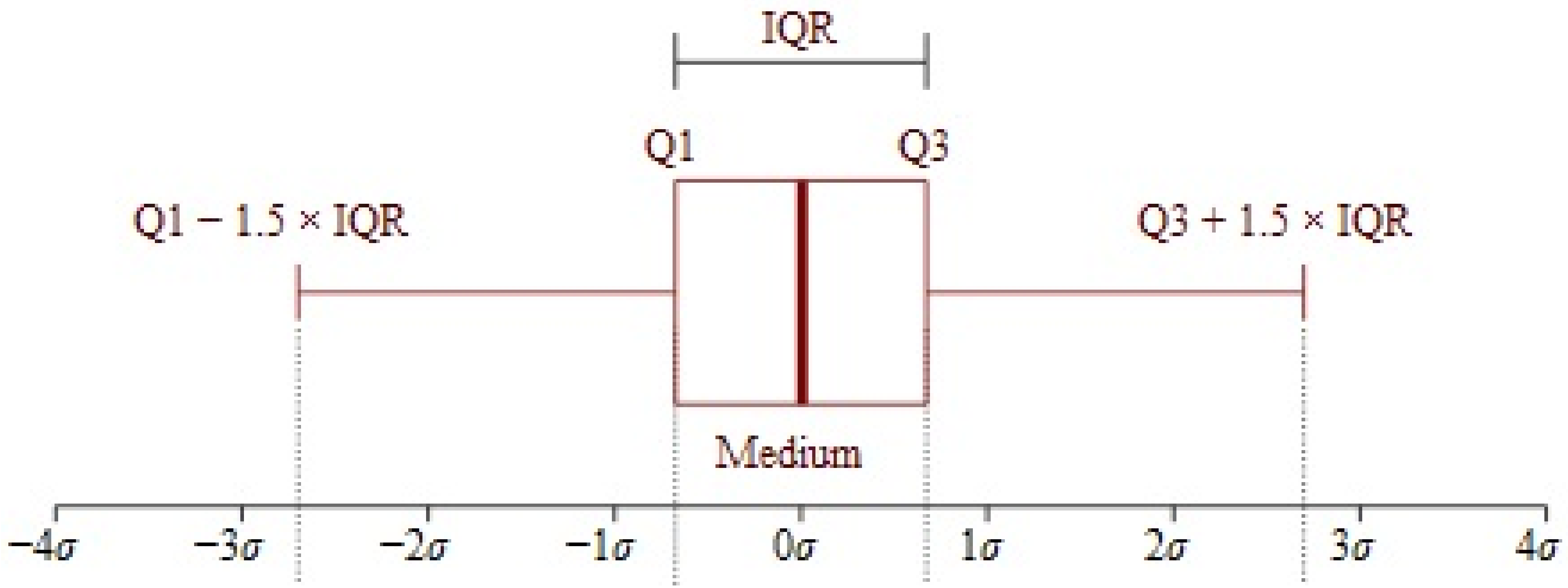
- Standard deviations
- Interquartile range (IQR)

Standard deviations



[1] Wikipedia

Interquartile range (IQR)



[1] Wikipedia

Summary

- Drop the whole row
- Impute missing values
- Standard deviations
- Interquartile range

Let's prepare for the interview!

PREPARING FOR STATISTICS INTERVIEW QUESTIONS IN PYTHON

Bias-variance tradeoff

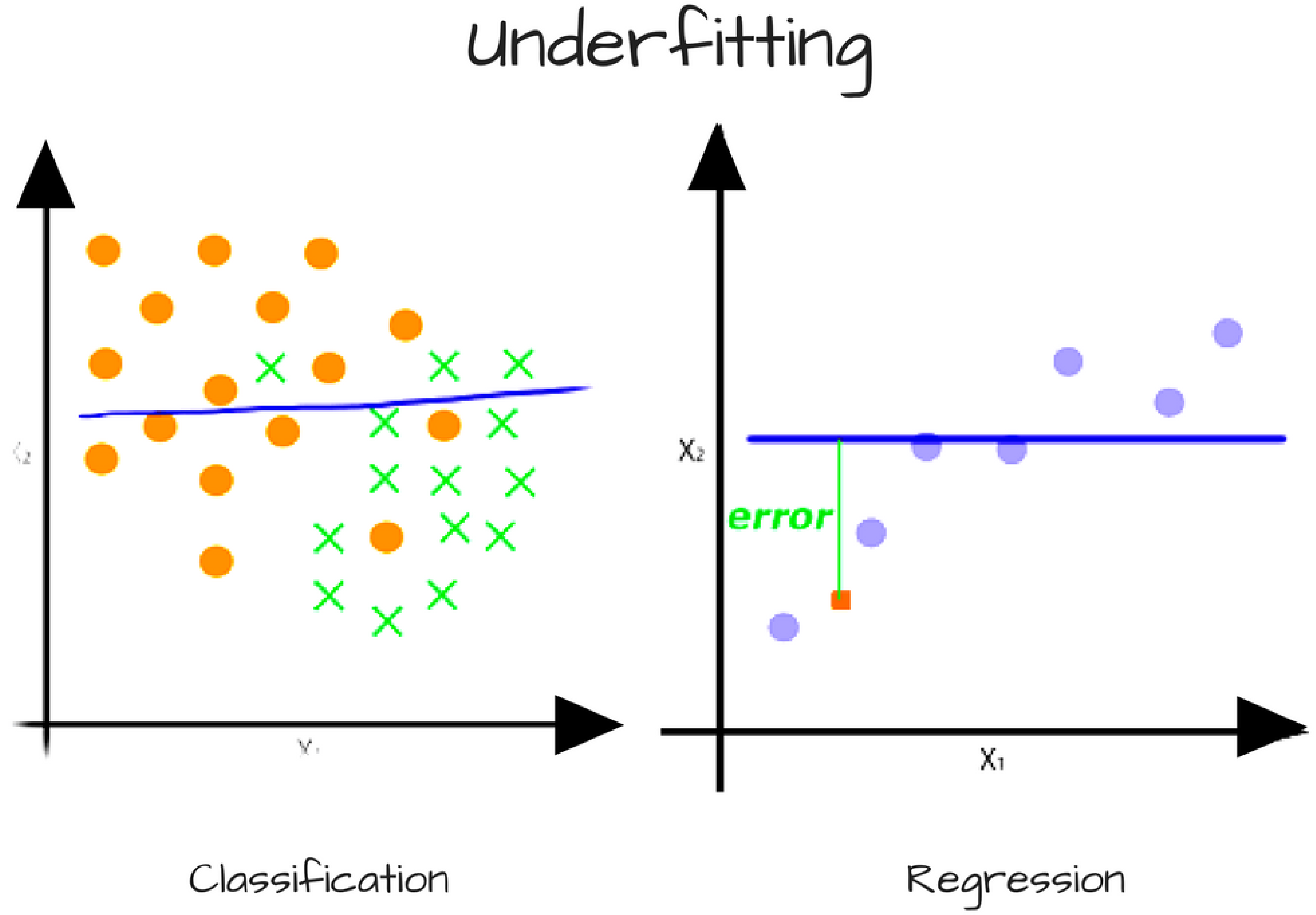
PREPARING FOR STATISTICS INTERVIEW QUESTIONS IN PYTHON

Conor Dewey
Data Scientist, Squarespace

Types of error

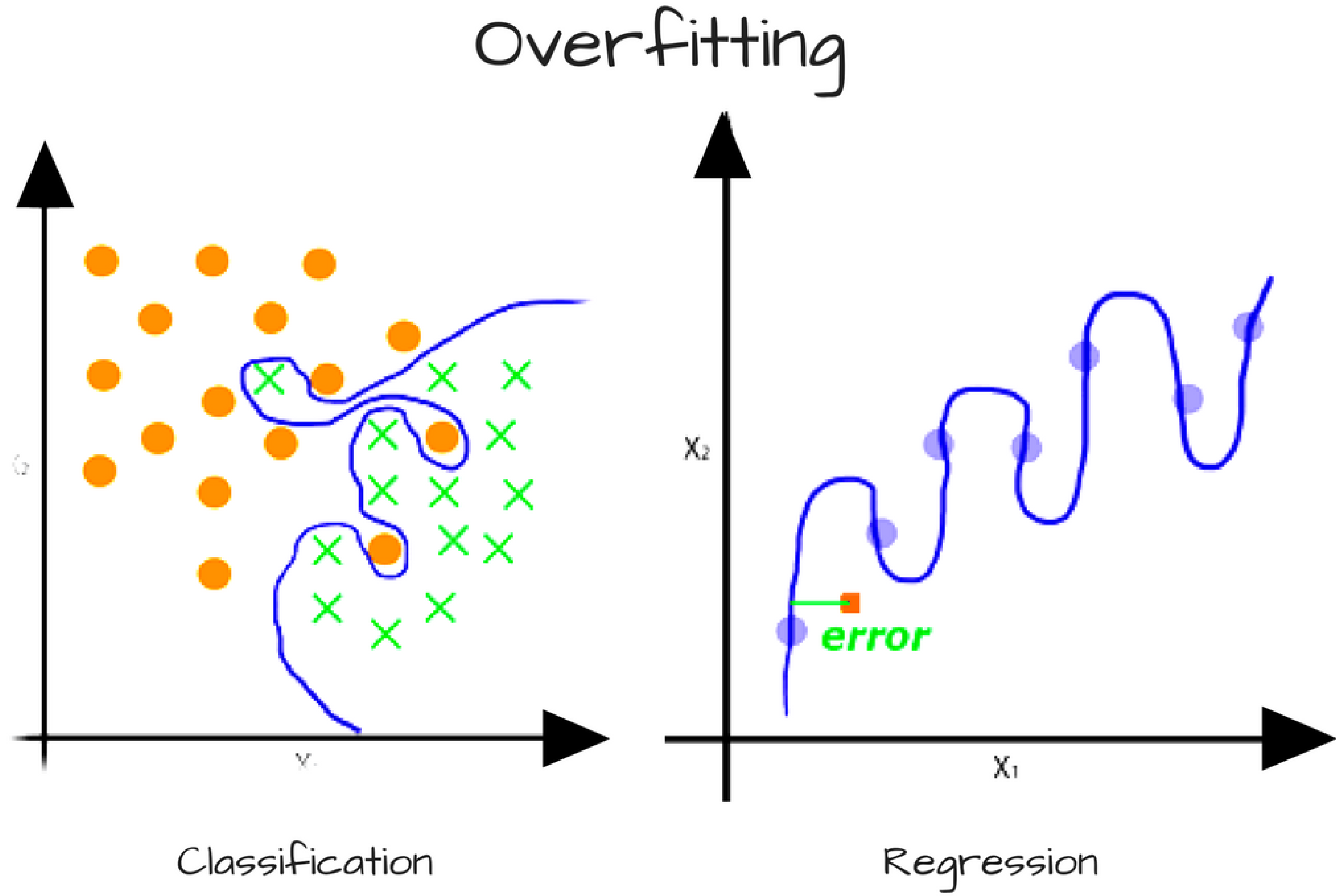
- Bias error
- Variance error
- Irreducible error

Bias error



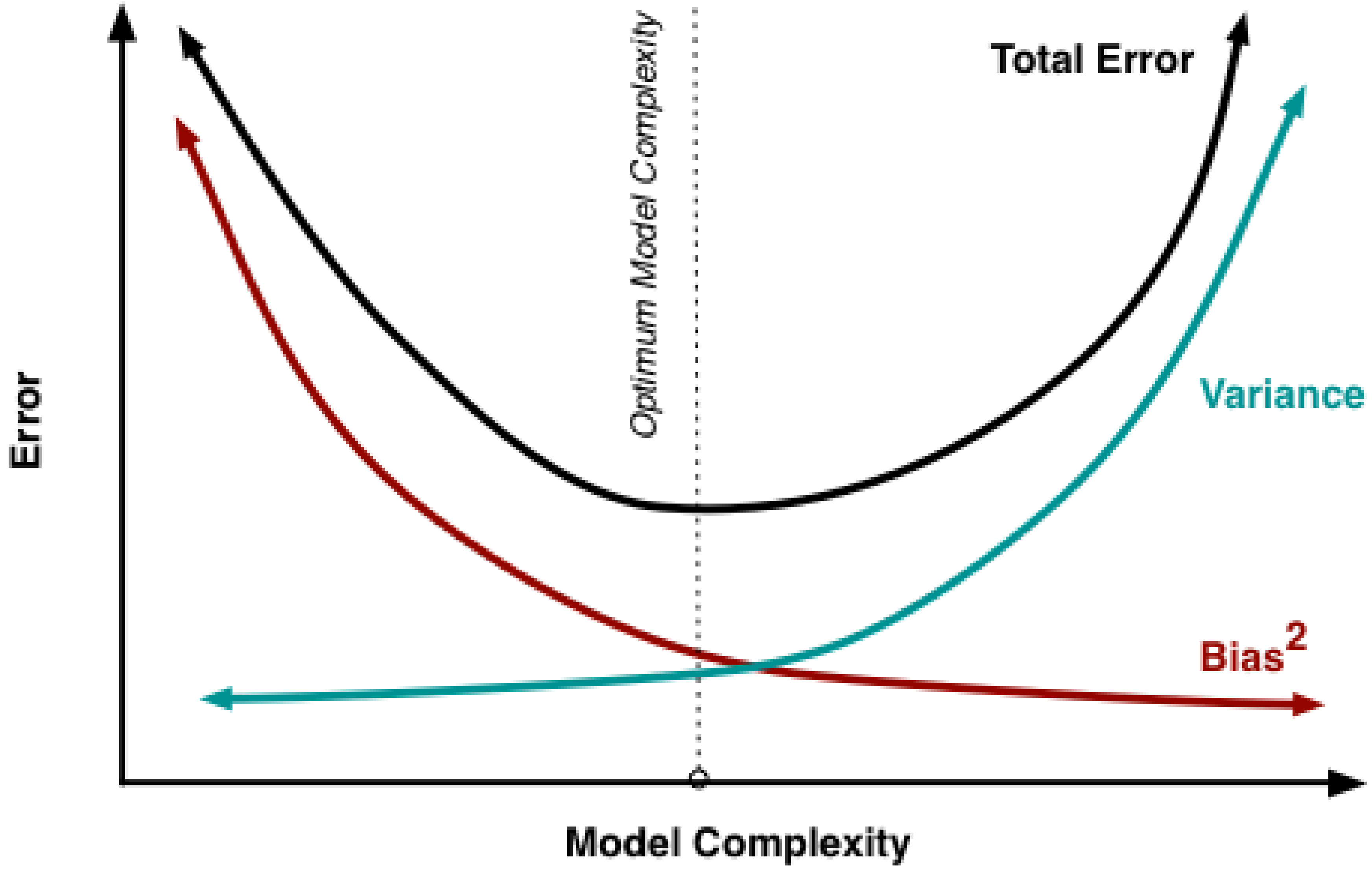
[1] How to Use Machine Learning to Predict the Quality of Wines

Variance error



[1] How to Use Machine Learning to Predict the Quality of Wines

Bias-variance tradeoff



[1] Scott Fortmann

Summary

- Types of error
- Bias error
- Variance error
- Bias-variance tradeoff

Let's prepare for the interview!

PREPARING FOR STATISTICS INTERVIEW QUESTIONS IN PYTHON