women
in
data

facebook

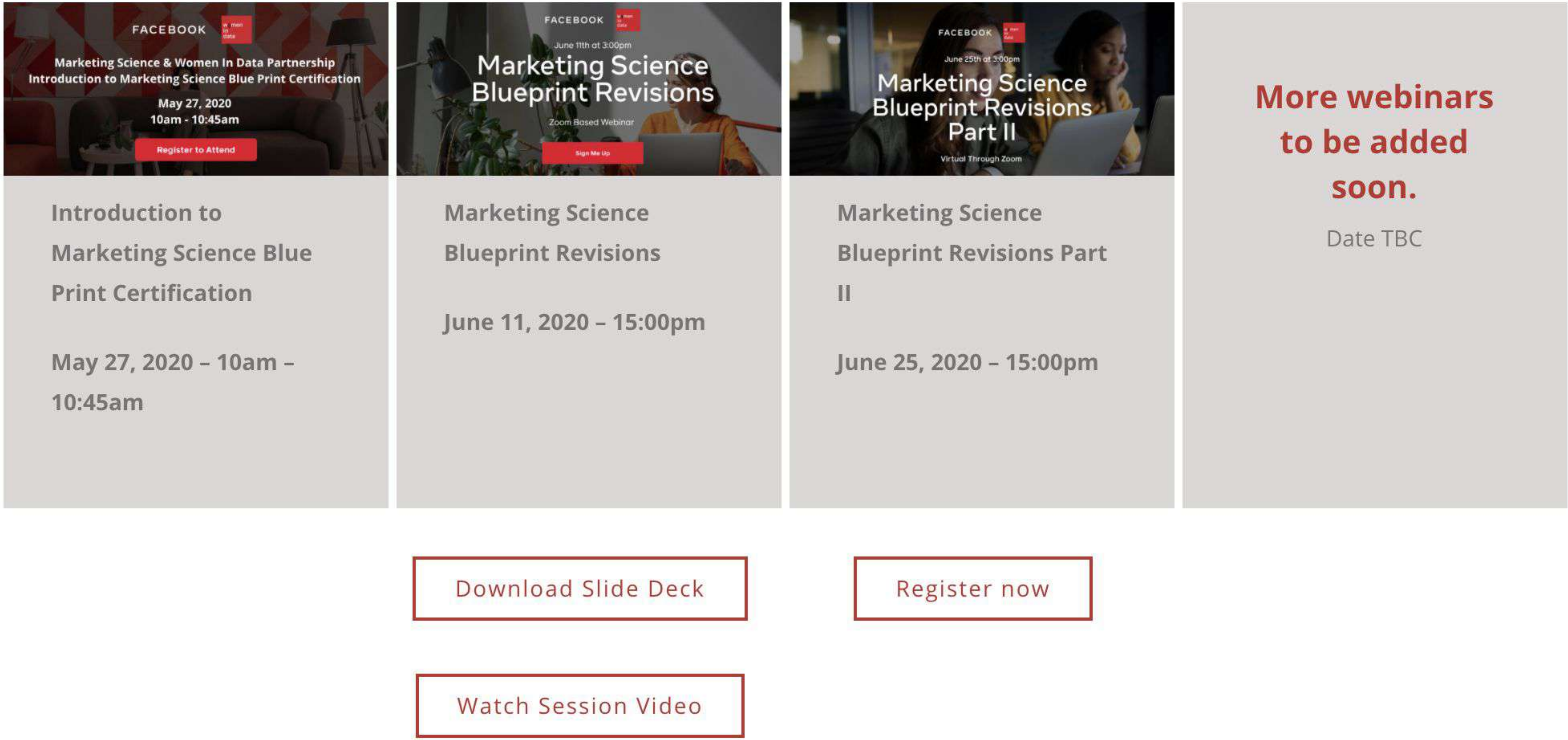
Facebook Marketing Science Revision for Blueprint Certification

WiD | Facebook

June 25, 2020

WEBINAR SCHEDULE

https://womenindata.co.uk/facebook-marketing-science-certification/



FACEBOOK

Marketing Science & Women In Data Partnership
Introduction to Marketing Science Blue Print Certification

May 27, 2020
10am - 10:45am

Register to Attend

Introduction to Marketing Science Blue Print Certification

May 27, 2020 – 10am – 10:45am

FACEBOOK

June 11th at 3:00pm

Marketing Science Blueprint Revisions

Zoom Based Webinar

Sign Me Up

Marketing Science Blueprint Revisions

June 11, 2020 – 15:00pm

FACEBOOK

June 25th at 3:00pm

Marketing Science Blueprint Revisions Part II

Virtual Through Zoom

Marketing Science Blueprint Revisions Part II

June 25, 2020 – 15:00pm

More webinars to be added soon.

Date TBC

Download Slide Deck

Register now

Watch Session Video

PREVIOUSLY:
Assess
Hypothesise

TODAY:

Recommend Measurement Solutions

Perform an Analysis:

- Analyze results from Facebook’s measurement tools
- Reconcile outputs from different sources
- Statistics and visualization methods
- Extract & manipulate data: SQL basics

Recommend Measurement Solutions

Which Measurement Solution?

MMM

An advertiser wants to cut its marketing budget by 10% and uses MMM to decide where to direct the cuts.

FB Attribution

An advertiser can track their consumer journey and **attribute incremental value** to all of their media touchpoints, allowing them to optimise budgets **across publishers and tactics**.

FB Conversion Lift

An advertiser wants to understand which of its targeting audiences generates the **greatest incremental ROAS**.

FB Brand Lift

An advertiser wants to understand which tactics result in the greatest **incremental lift in awareness** of its new line extension.

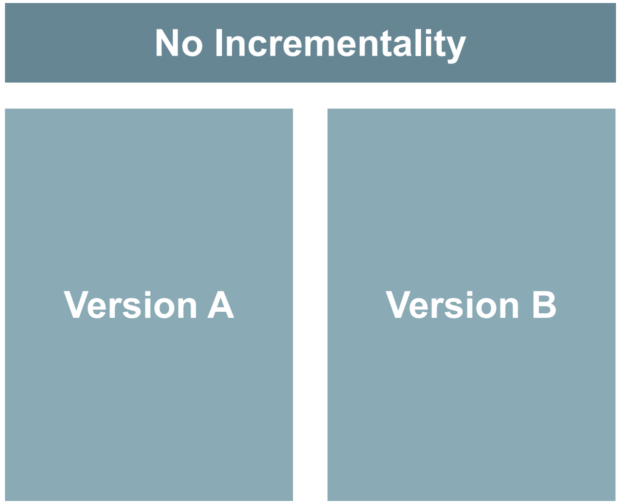
A/B Testing

Which creative execution (for example) is more effective?

Understand the test methodology

A/B Tests

Which ad set has better results?



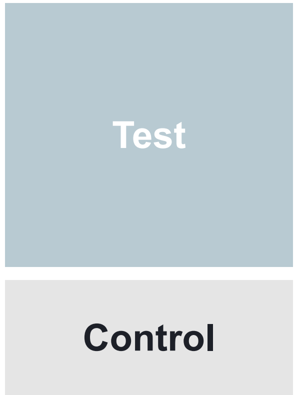
No Incrementality

Version A

Version B

Single Cell Lift Test

How effective is the Campaign at driving incremental results?

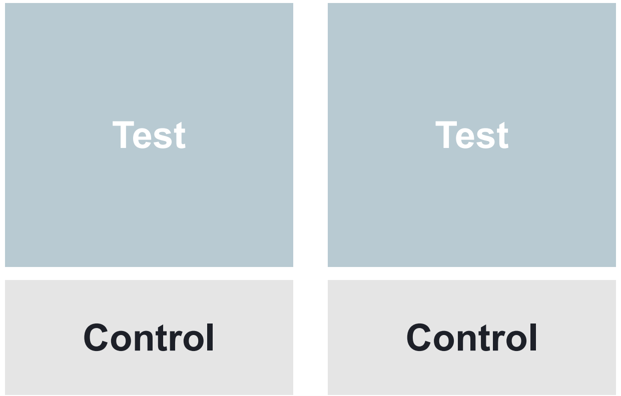


Test

Control

Multi Cell Lift Test

What campaigns / elements are most effective at driving results?



Test

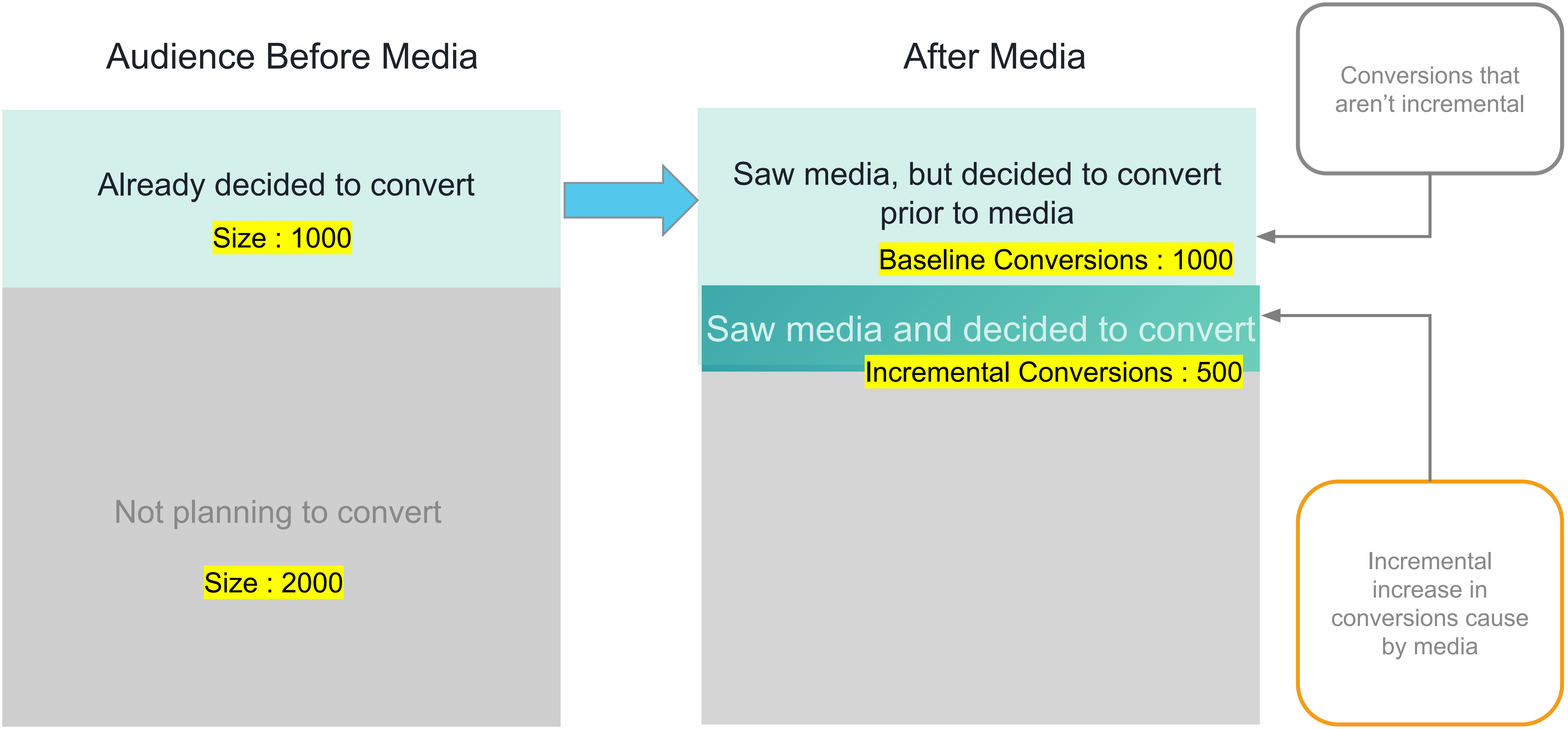
Test

Control

Control

Analyze Results

How is incremental lift calculated?



Audience Before Media

Already decided to convert

Size : 1000

Not planning to convert

Size : 2000

After Media

Saw media, but decided to convert prior to media

Baseline Conversions : 1000

Saw media and decided to convert

Incremental Conversions : 500

Conversions that aren't incremental

Incremental increase in conversions cause by media

What action can I take from the results?

If positive and statistically significant:

- Continue running strategy
- If you want to scale, increase budget and re-test
- Explore variables to A/B test
- Calibrate attribution (adjust attribution model to the one that matches lift results closest)

If flat or not statistically significant:

- Adjust strategy (consider optimizing creative) and re-run the test while also tracking upper-funnel conversion events
- Reference the test setup checklist for campaign and measurement best practices

Wait until the end of the study to evaluate results
>=90% chance is a reliable result
Test and control groups combined need at least 100 converters before we can show your lift results

Further revision on recommending measurement solutions and interpreting results

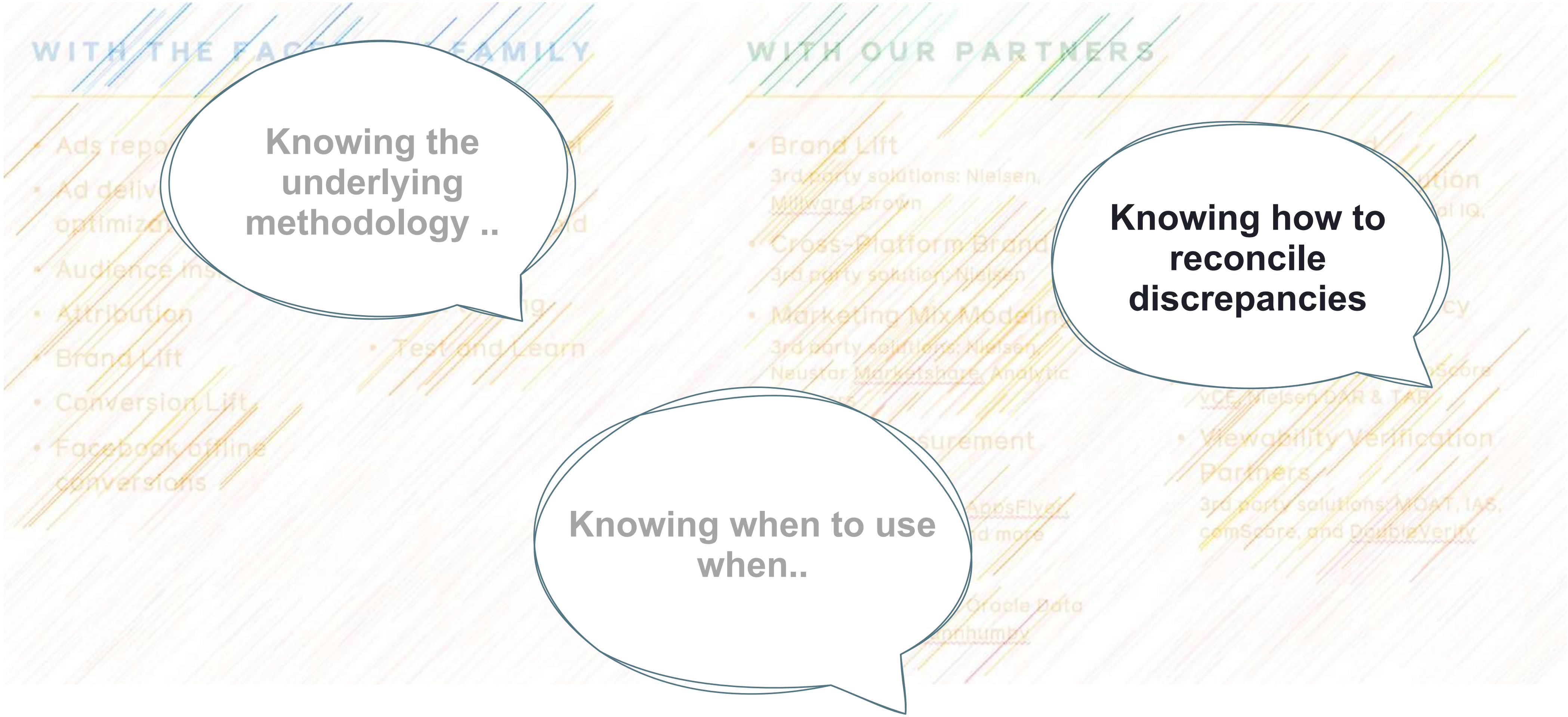
Recommend Measurement Solutions

Analyze Measurement Data to Extract Key Insights:

- Understand Measurement Tools and Data Availability
- Evaluate Different Measurement Methodologies
- Reconcile Results From Different Measurement Solutions
- Provide Data-Driven Recommendations

Reconciling Outputs from Different Sources

How to use different measurement solutions together but effectively?



Knowing the underlying methodology ..

Knowing how to reconcile discrepancies

Knowing when to use when..

WITH THE FACE FAMILY

- Ads repo
- Ad deliv optimiza
- Audience ins
- Attribution
- Brand Lift
- Conversion Lift
- Facebook offline conversions

- Test and Learn

WITH OUR PARTNERS

- Brand Lift
  3rd party solutions: Nielsen, Millward Brown
- Cross-Platform Brand
  3rd party solution: Nielsen
- Marketing Mix Modelin
  3rd party solutions: Nielsen, Neustar Marketshare, Analytic

surement

AppsFlyer, d more

Oracle Data nnhumby

tion al IQ,

cy

Score vCE, Nielsen DAR & TAR

- Viewability Verification Partners
  3rd party solutions: MOAT, IAS, comScore, and DoubleVerify

Before you reconcile , ask these questions

What data inputs are used?

Is the data captured accurately?

Time frames and conversion windows

What measurement methods are being used?

Observational vs. Experimental

Lift Methodology

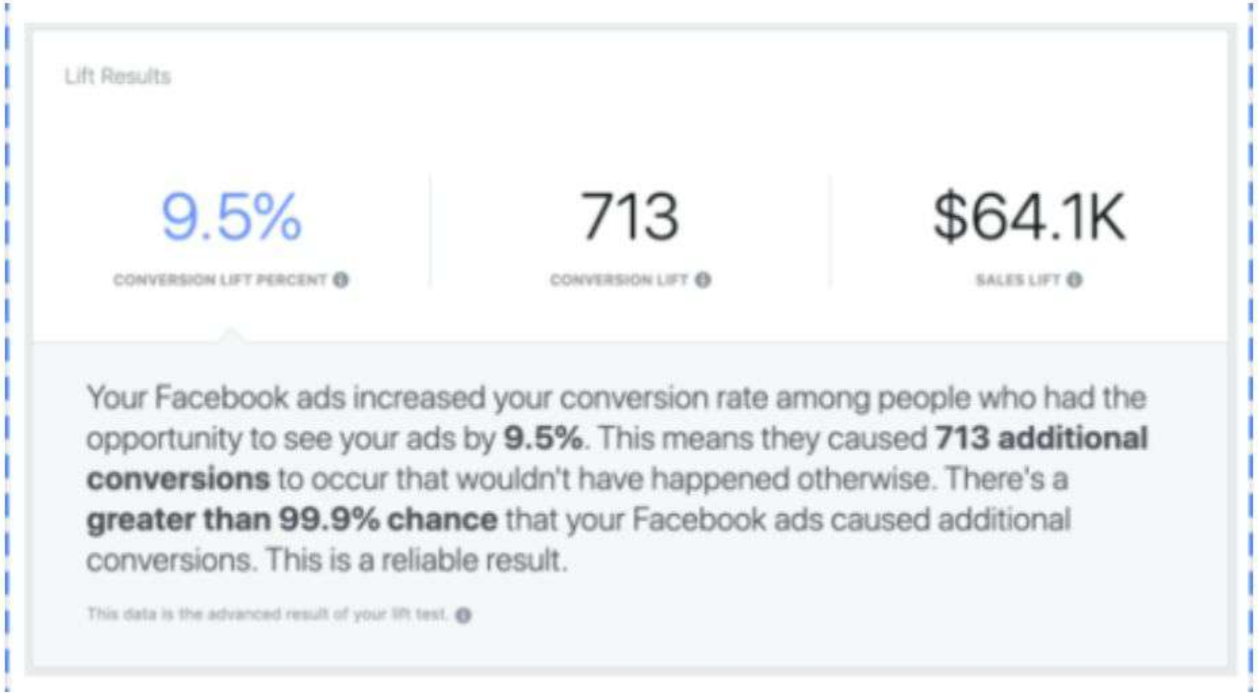
Example Scenario - Feroldis e-commerce fashion business

Conversion Lift Results

Ads Manager

Feroldis

Lift Test Result



Lift Results

9.5%
CONVERSION LIFT PERCENT ⓘ

713
CONVERSION LIFT ⓘ

$64.1K
SALES LIFT ⓘ

Your Facebook ads increased your conversion rate among people who had the opportunity to see your ads by **9.5%**. This means they caused **713 additional conversions** to occur that wouldn't have happened otherwise. There's a **greater than 99.9% chance** that your Facebook ads caused additional conversions. This is a reliable result.

This data is the advanced result of your lift test. ⓘ

Ads Manager

923 Conversions showing in Ads Manager for the month of January?

Parameters to reconcile

Data Inputs

Both Ads Manager and Conversion Lift use pixel fired event data

Time Frames

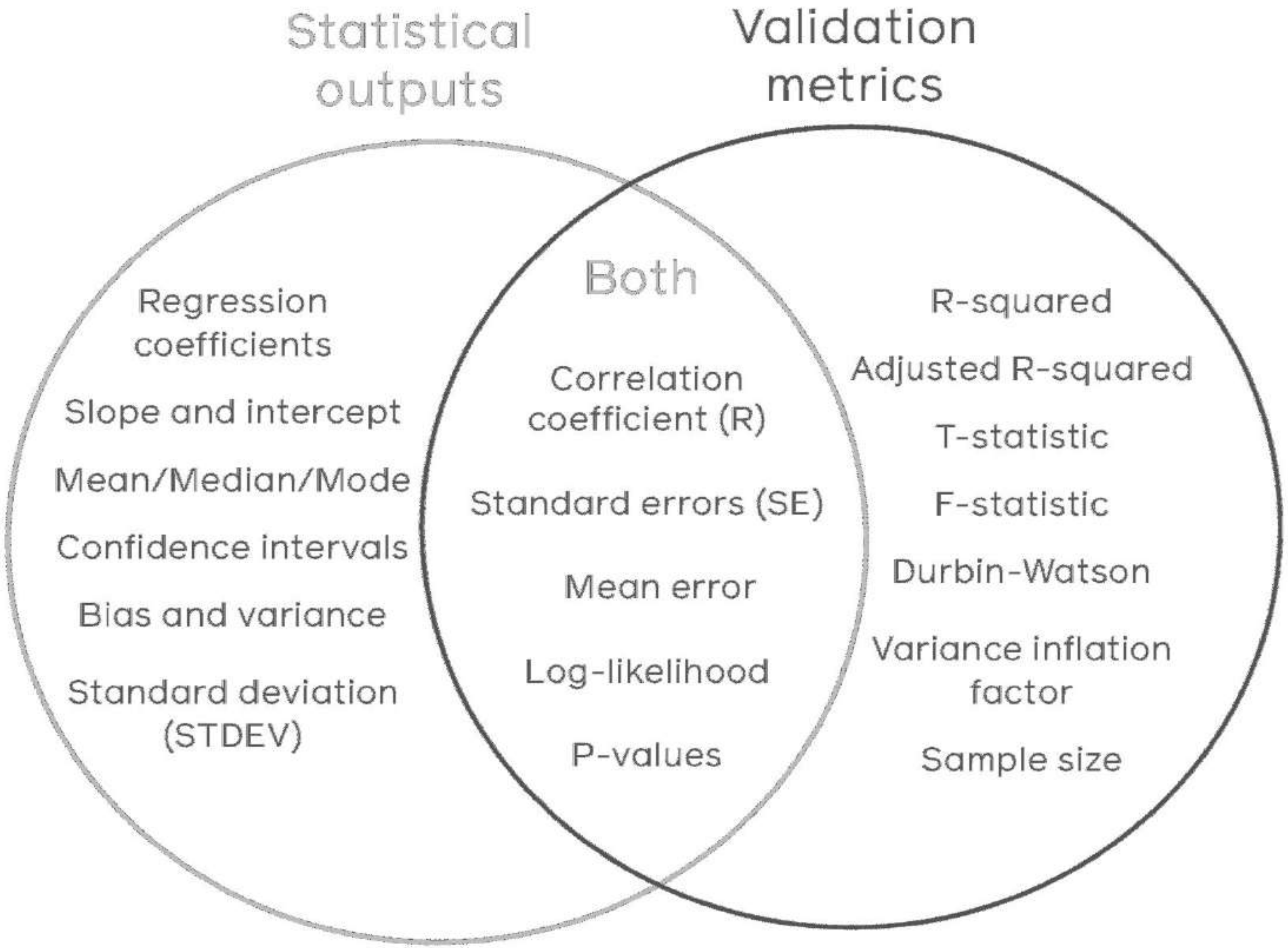
Limited to same time window Attribution is same
1 day after view
28 days after click

Measurement Methods

Experimental vs. Observational

Lift only provides incremental metrics whereas ads manager reports all

Statistics and Analytical Reference



Statistical outputs

Validation metrics

Regression coefficients

Slope and intercept

Mean/Median/Mode

Confidence intervals

Bias and variance

Standard deviation (STDEV)

Both

Correlation coefficient (R)

Standard errors (SE)

Mean error

Log-likelihood

P-values

R-squared

Adjusted R-squared

T-statistic

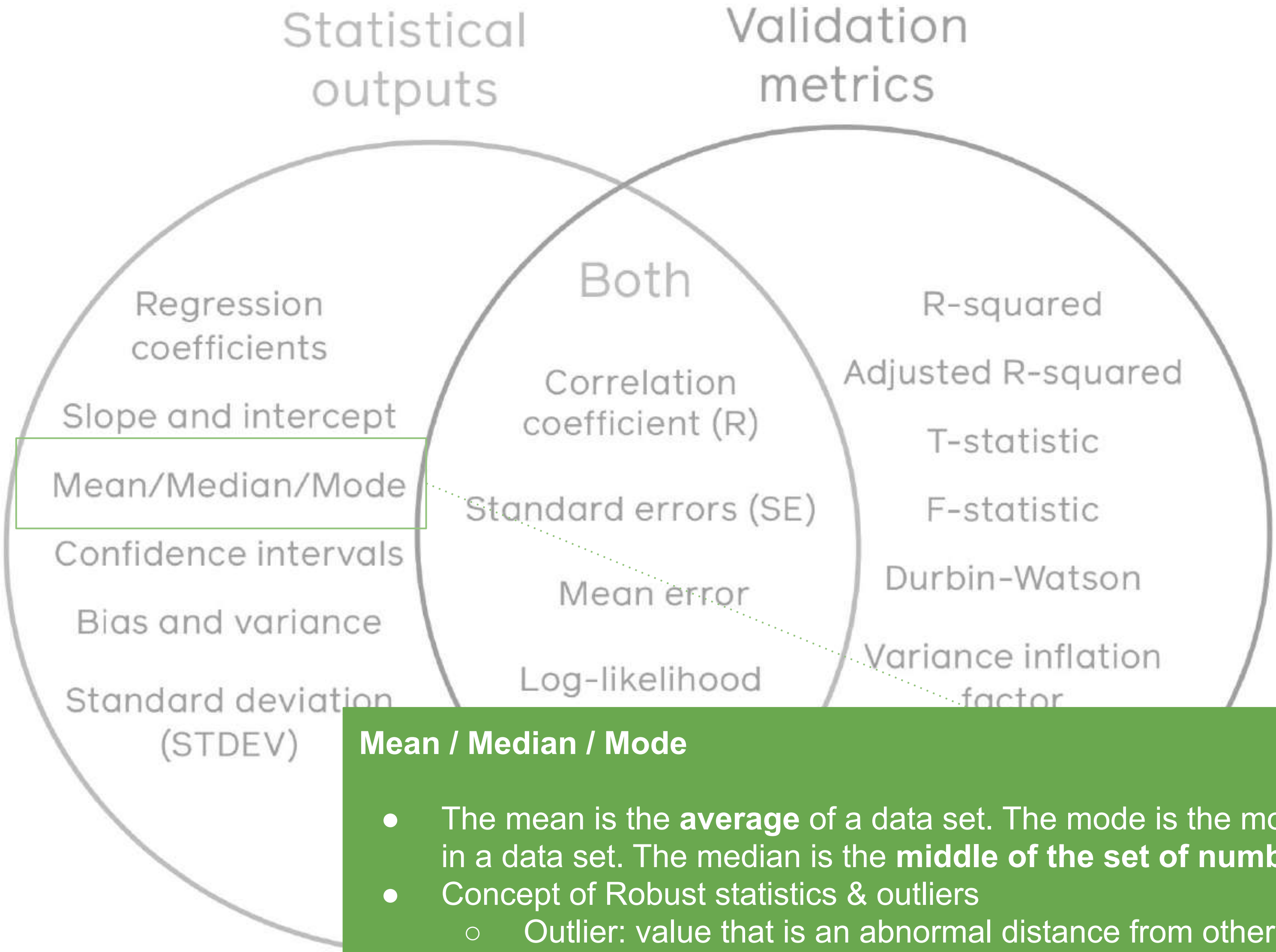
F-statistic

Durbin-Watson

Variance inflation factor

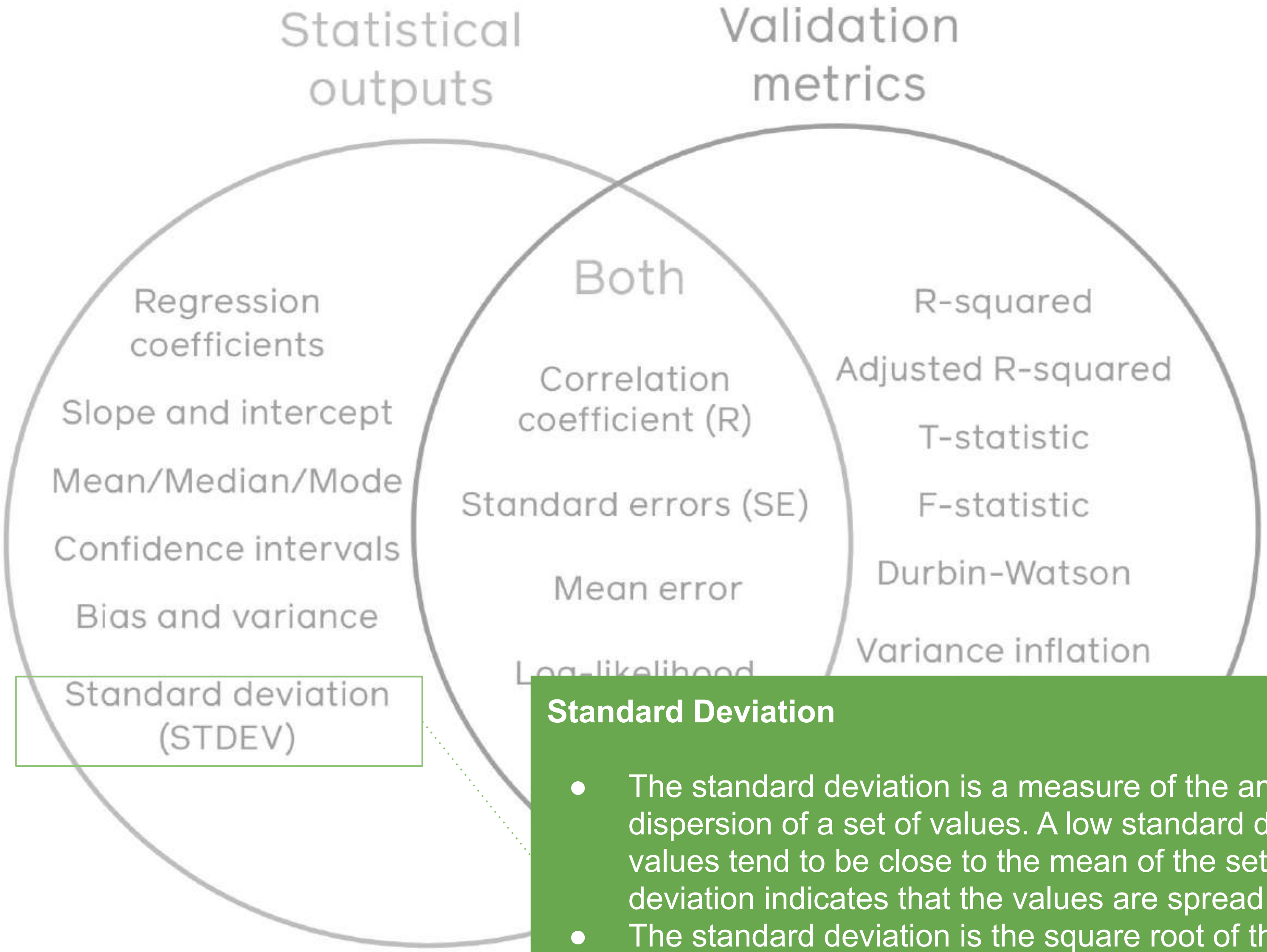
Sample size

GO DEEP ON THIS USING THE LINKS TO KHAN ACADEMY OR ANY SOURCE OF STATISTICAL KNOWLEDGE OF YOUR CHOICE



Mean / Median / Mode

- The mean is the **average** of a data set. The mode is the most **common** number in a data set. The median is the **middle of the set of numbers**
- Concept of Robust statistics & outliers
  - Outlier: value that is an abnormal distance from other values
  - The mean is very susceptible to outliers (non-robust)
  - While the median is not affected by outliers (robust)



Statistical outputs

Validation metrics

Both

Regression coefficients

Slope and intercept

Mean/Median/Mode

Confidence intervals

Bias and variance

Standard deviation (STDEV)

Correlation coefficient (R)

Standard errors (SE)

Mean error

Log-likelihood

R-squared

Adjusted R-squared

T-statistic

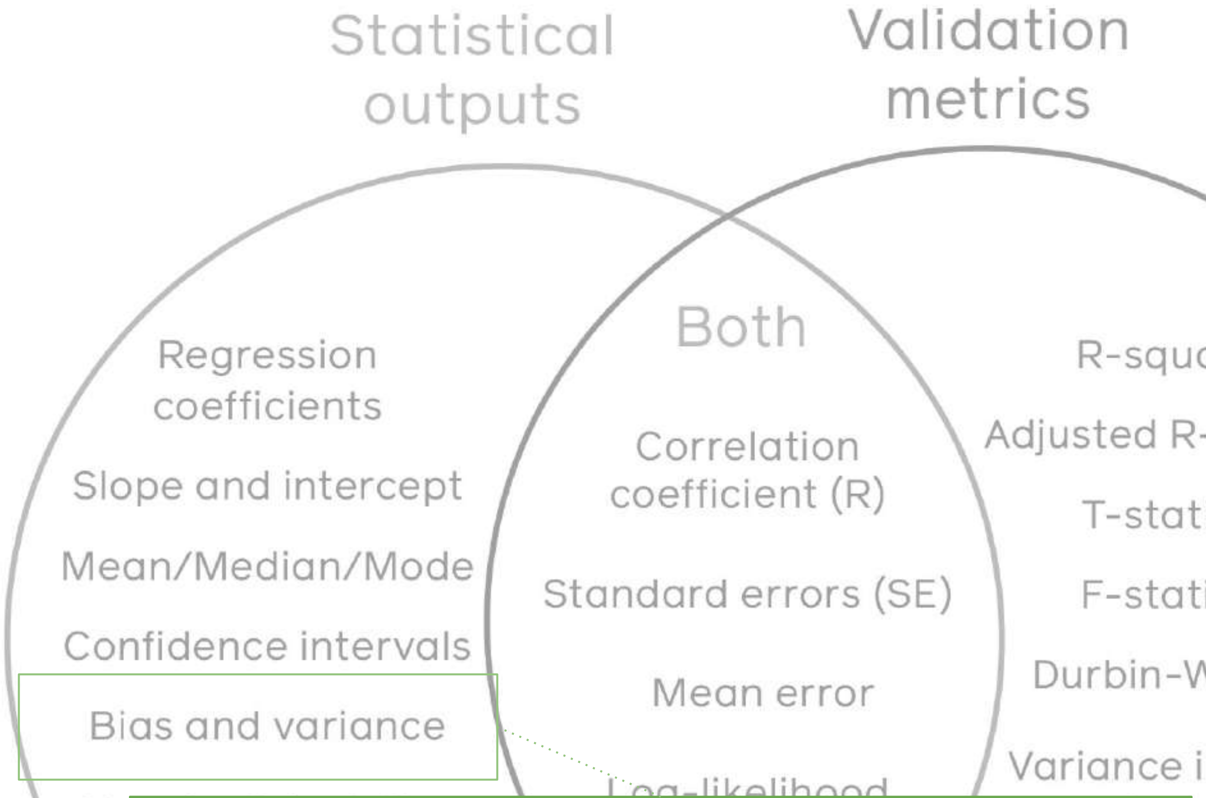
F-statistic

Durbin-Watson

Variance inflation

Standard Deviation

- The standard deviation is a measure of the amount of variation or dispersion of a set of values. A low standard deviation indicates that the values tend to be close to the mean of the set, while a high standard deviation indicates that the values are spread out over a wider range
- The standard deviation is the square root of the variance



Bias & Variance

- Bias: is error from erroneous assumptions in the underlying model – in a high bias situation, you are systematically off.
- Variance is the expectation of the squared deviation of a random variable from its mean. It measures how far a set of numbers are spread out from their average value.

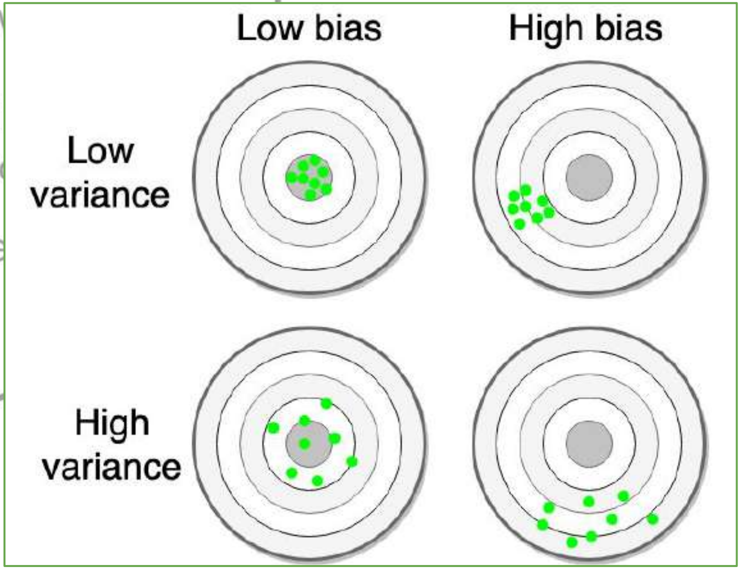
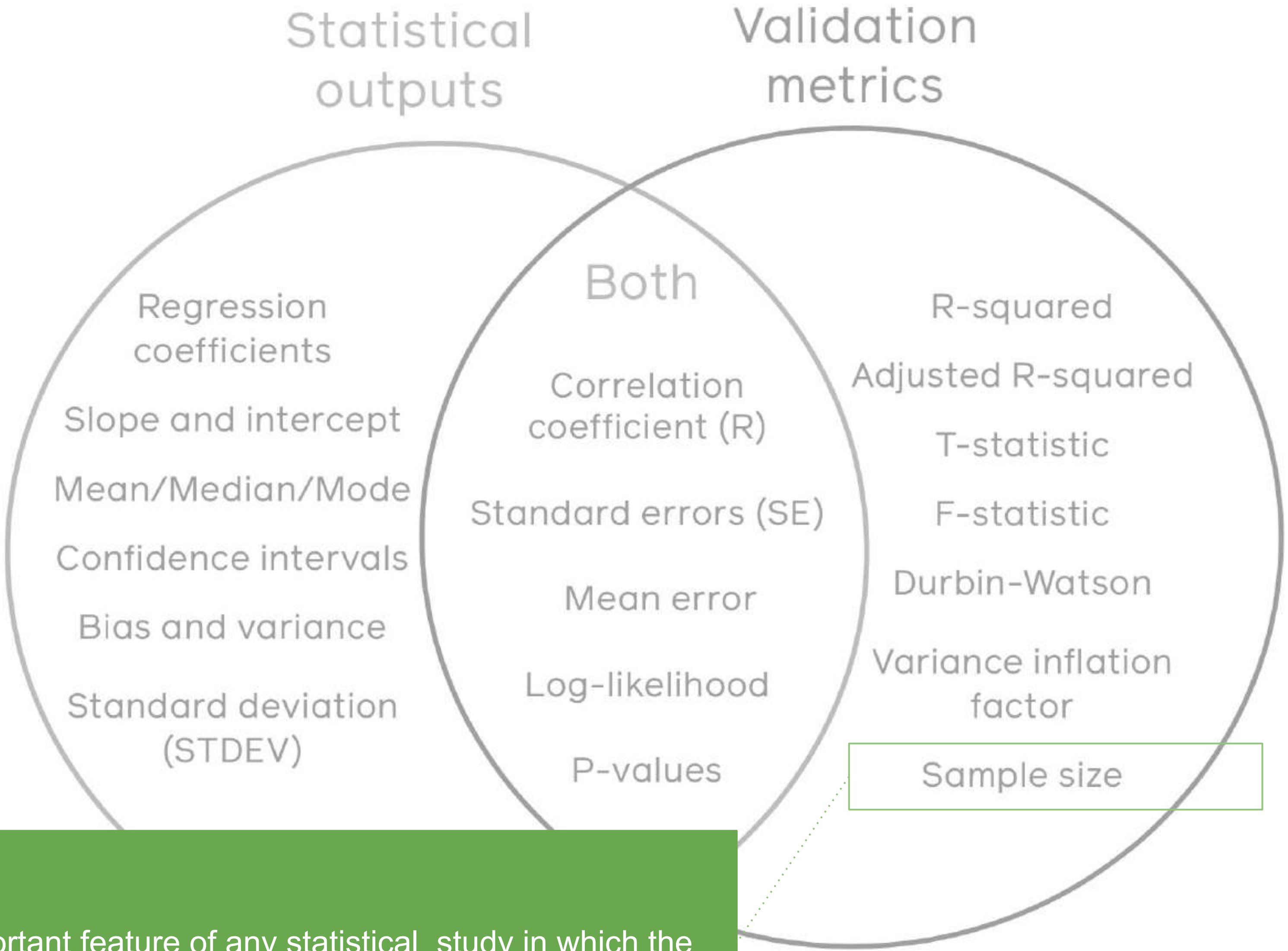


Image Source:
https://www.machinelearningtutorial.net/2017/01/26/the-bias-variance-tradeoff/



Statistical outputs

Regression coefficients

Slope and intercept

Mean/Median/Mode

Confidence intervals

Bias and variance

Standard deviation (STDEV)

Both

Correlation coefficient (R)

Standard errors (SE)

Mean error

Log-likelihood

P-values

Validation metrics

R-squared

Adjusted R-squared

T-statistic

F-statistic

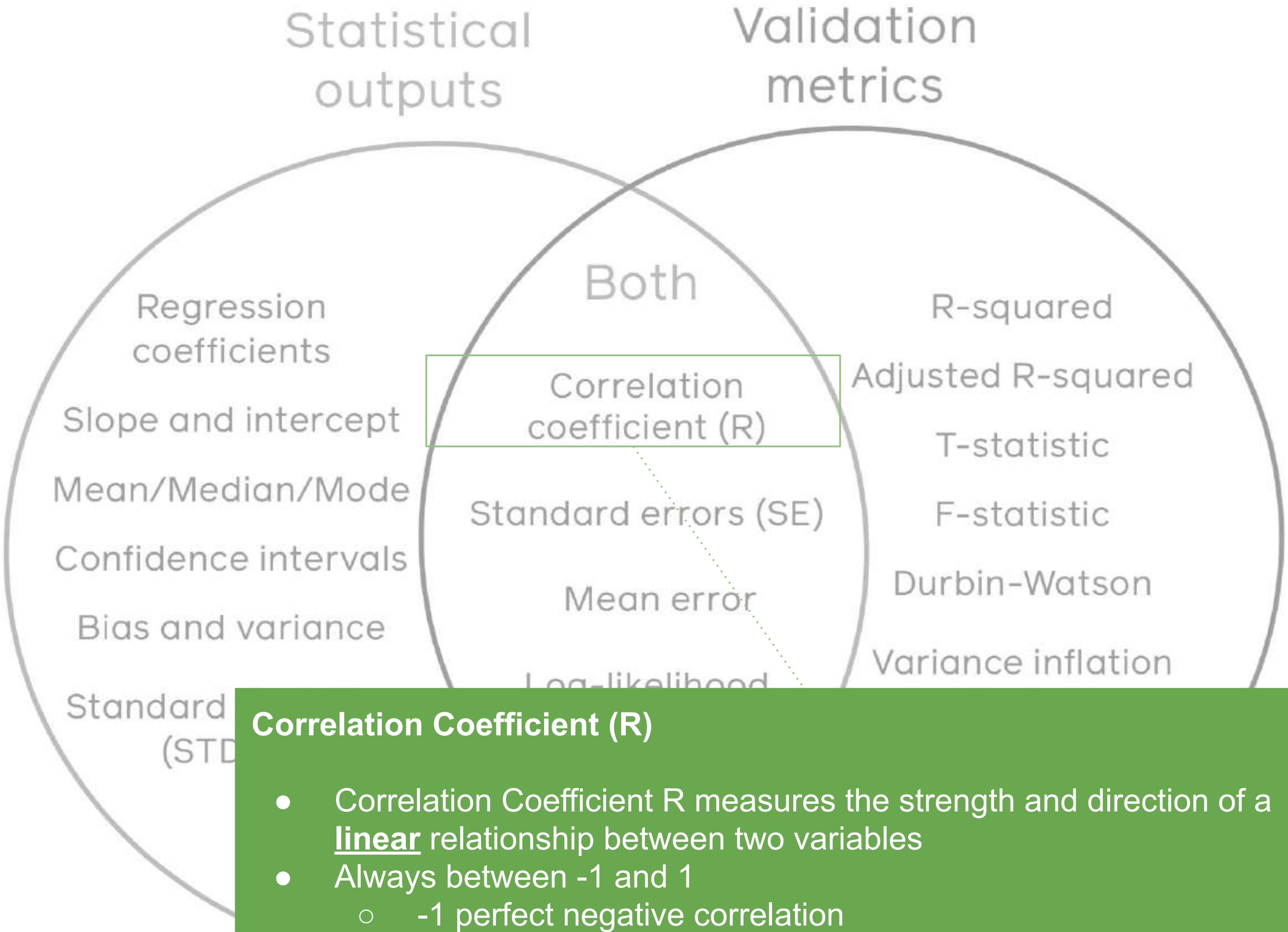
Durbin-Watson

Variance inflation factor

Sample size

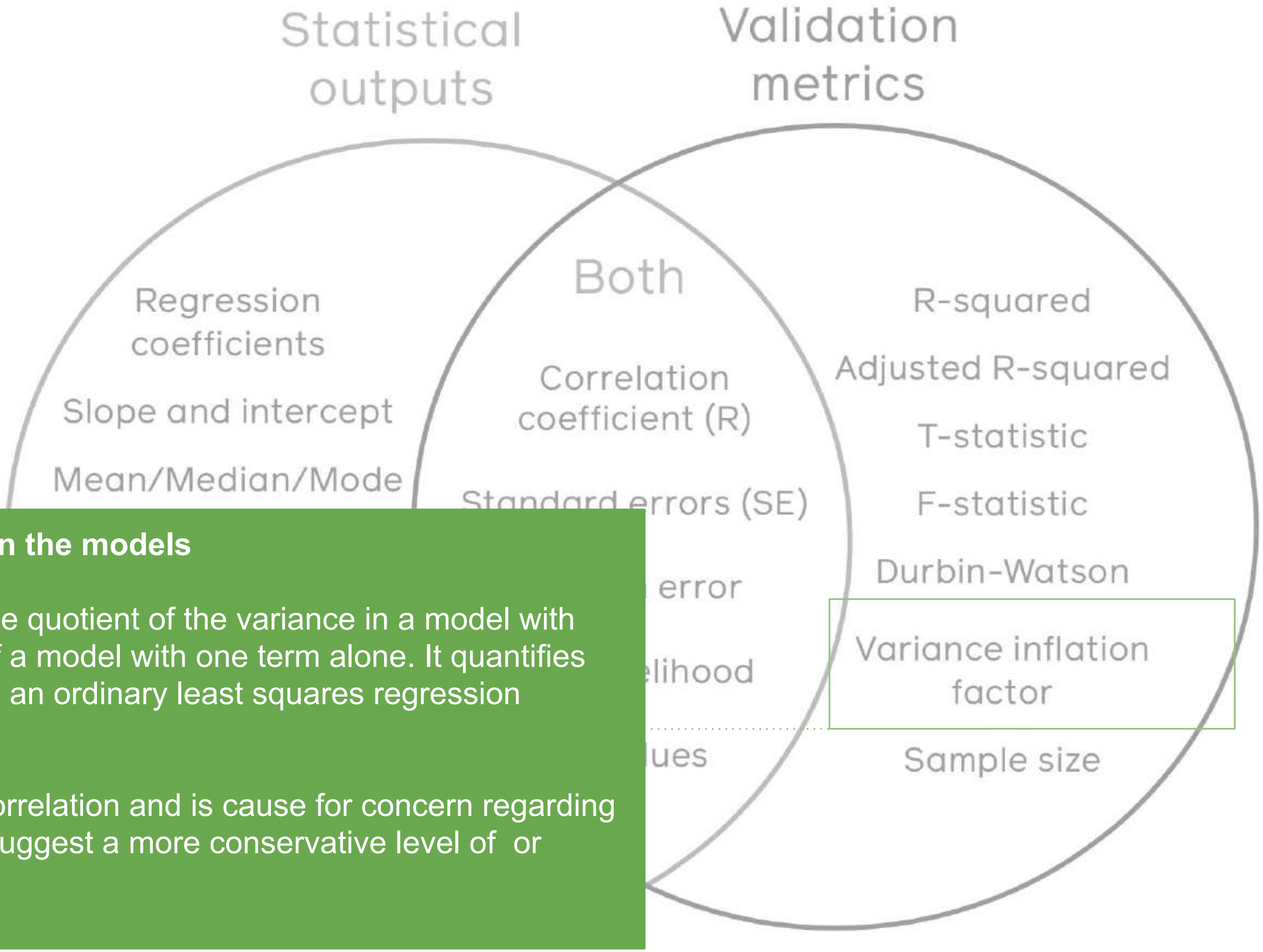
Sample size in statistics:

The sample size is an important feature of any statistical study in which the goal is to make inferences - and the need for it to offer sufficient **statistical power.**



Correlation Coefficient (R)

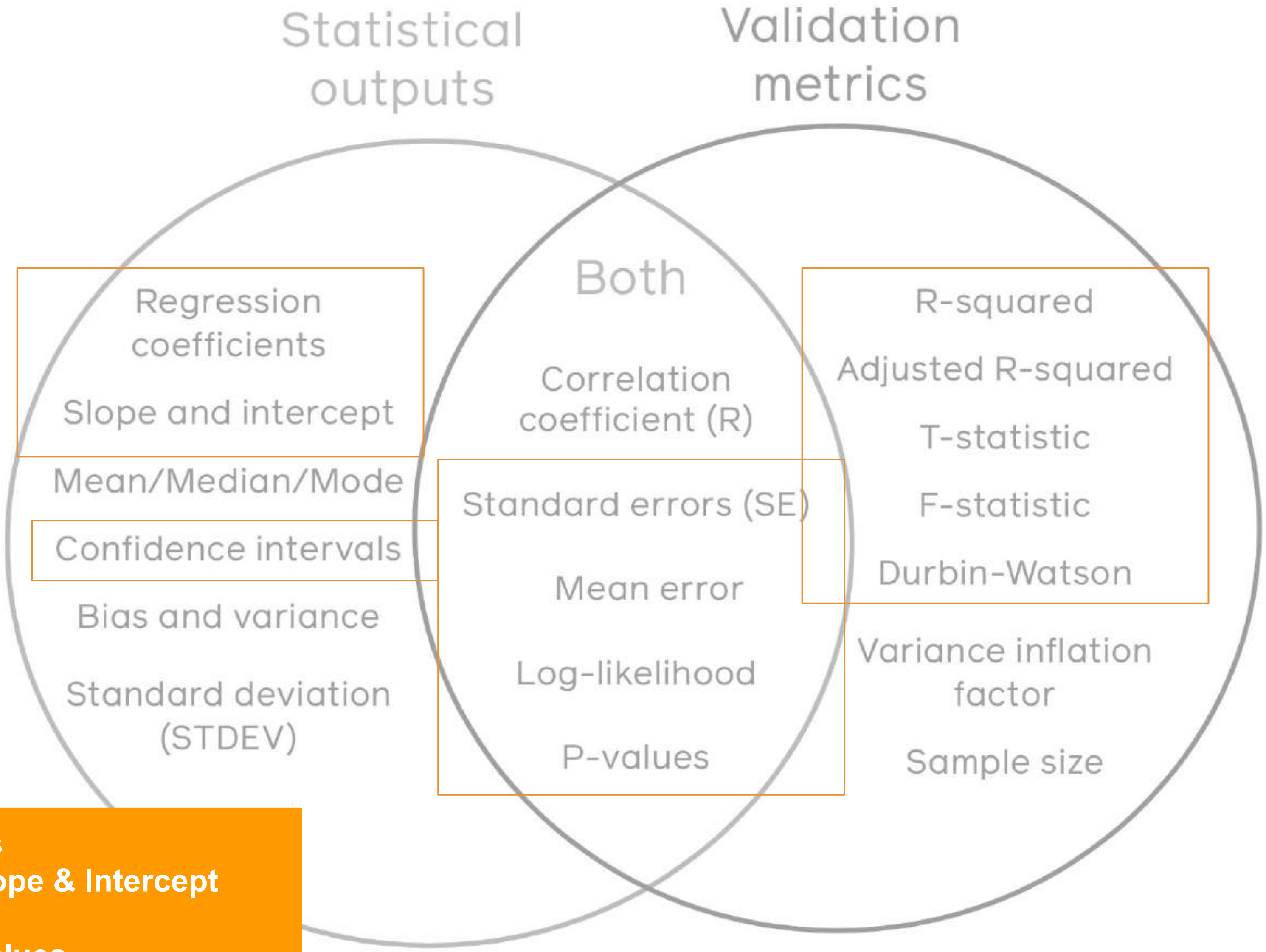
- Correlation Coefficient R measures the strength and direction of a **linear** relationship between two variables
- Always between -1 and 1
  - -1 perfect negative correlation
  - 0 no correlation
  - 1 perfect positive correlation
- Also R-squared = R x R = Square of Correlation



VIF - test for Multicollinearity in the models

The variance inflation factor is the quotient of the variance in a model with multiple terms by the variance of a model with one term alone. It quantifies the severity of multicollinearity in an ordinary least squares regression analysis.

A VIF above 10 indicates high correlation and is cause for concern regarding multicollinearity. Some authors suggest a more conservative level of or above.



Statistical outputs

Validation metrics

Regression coefficients

Slope and intercept

Mean/Median/Mode

Confidence intervals

Bias and variance

Standard deviation (STDEV)

Both

Correlation coefficient (R)

Standard errors (SE)

Mean error

Log-likelihood

P-values

R-squared

Adjusted R-squared

T-statistic

F-statistic

Durbin-Watson

Variance inflation factor

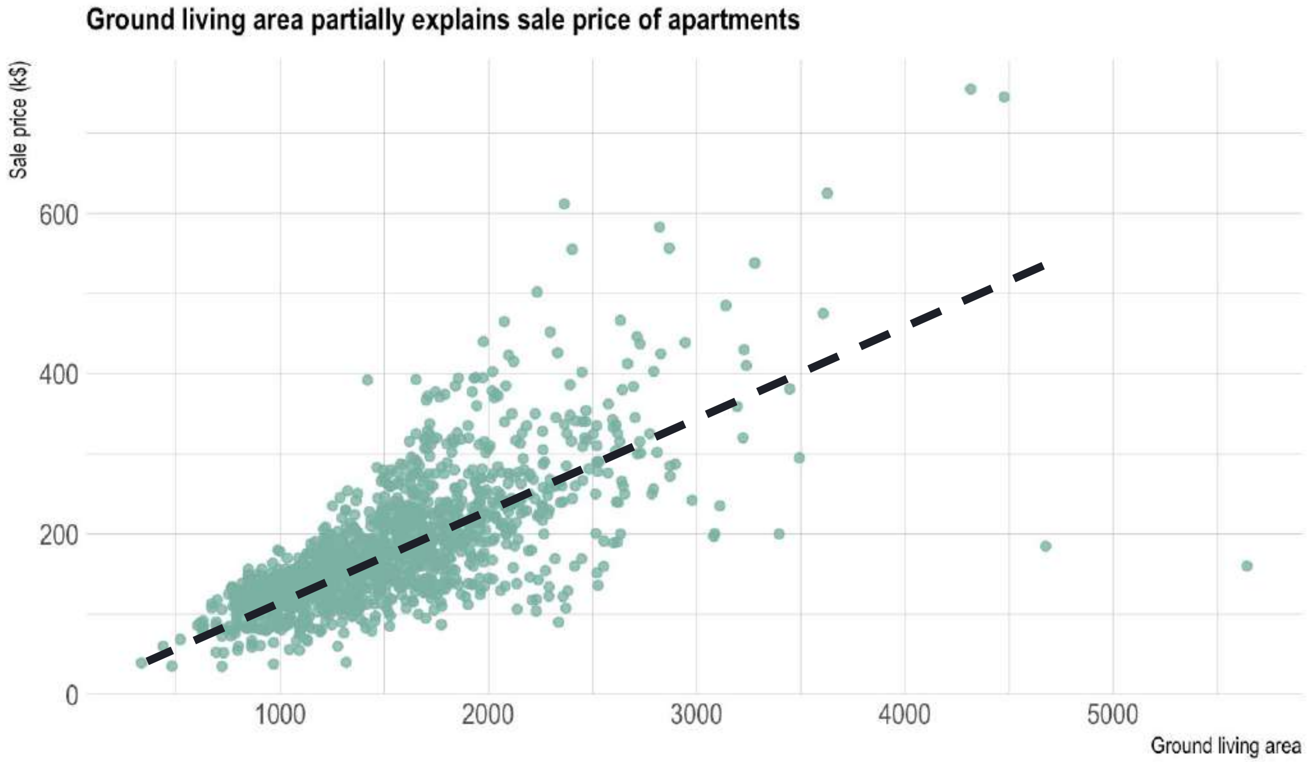
Sample size

Addressing on the next slides

- **Regression outputs/ Slope & Intercept**
- **Confidence intervals**
- **SE, Log Likelihood, P-values**
- **R-sq, Adjusted R-sq, T-stat, F-stat and DW**

Regression?

Y



X

Why do we need regression?

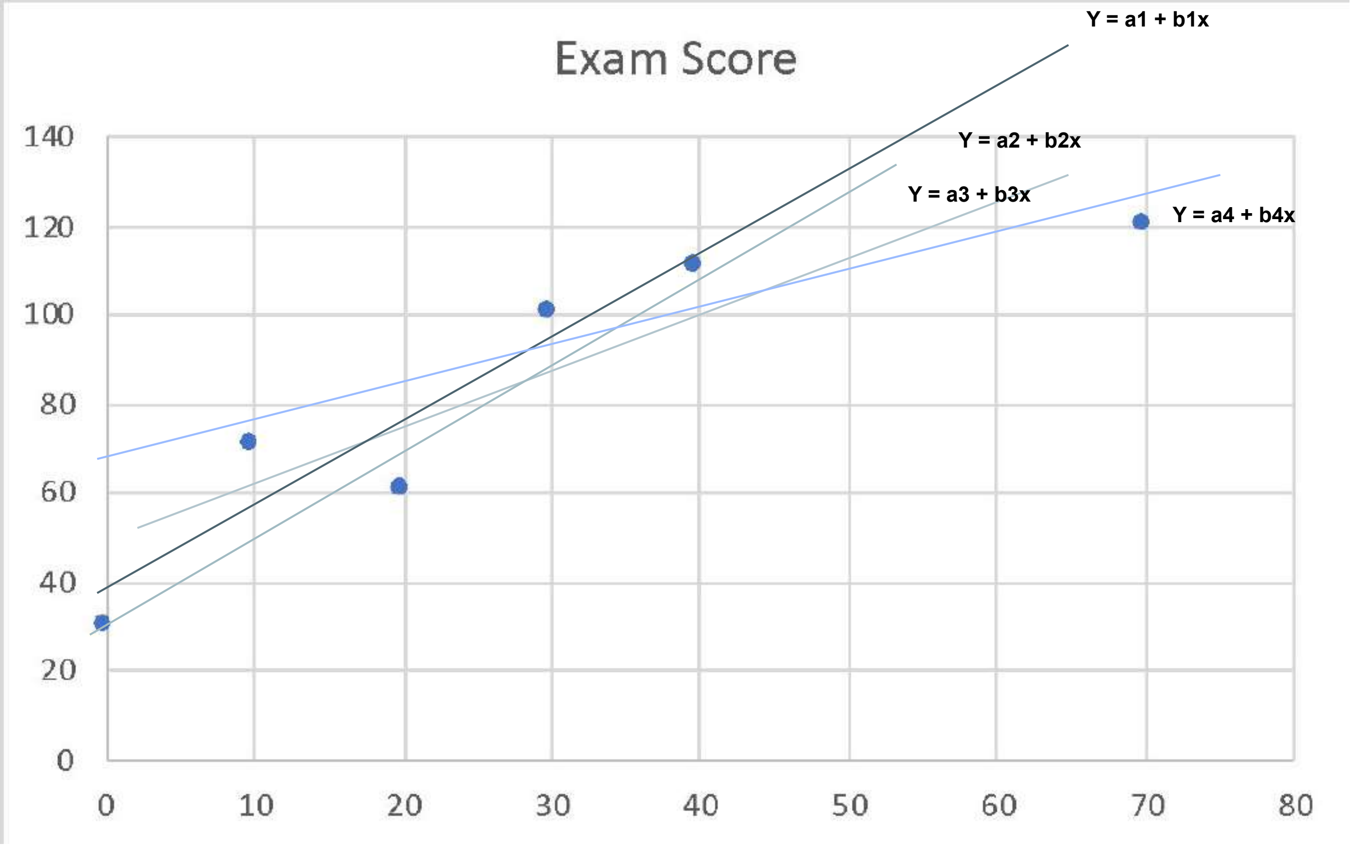
1. To determine if a significant relationship exists between X, (X2 & X3) and Y
2. To describe the nature of the relationship
3. To assess the degree of accuracy of the description or prediction achieved
4. In case of multiple predictors, one must also determine the relative importance of these predictors

Multivariate Statistical Analysis by Sam Kash Kachigan

So what is the best model or best fit?

Multiple lines can fit this data.

Which one is the best?



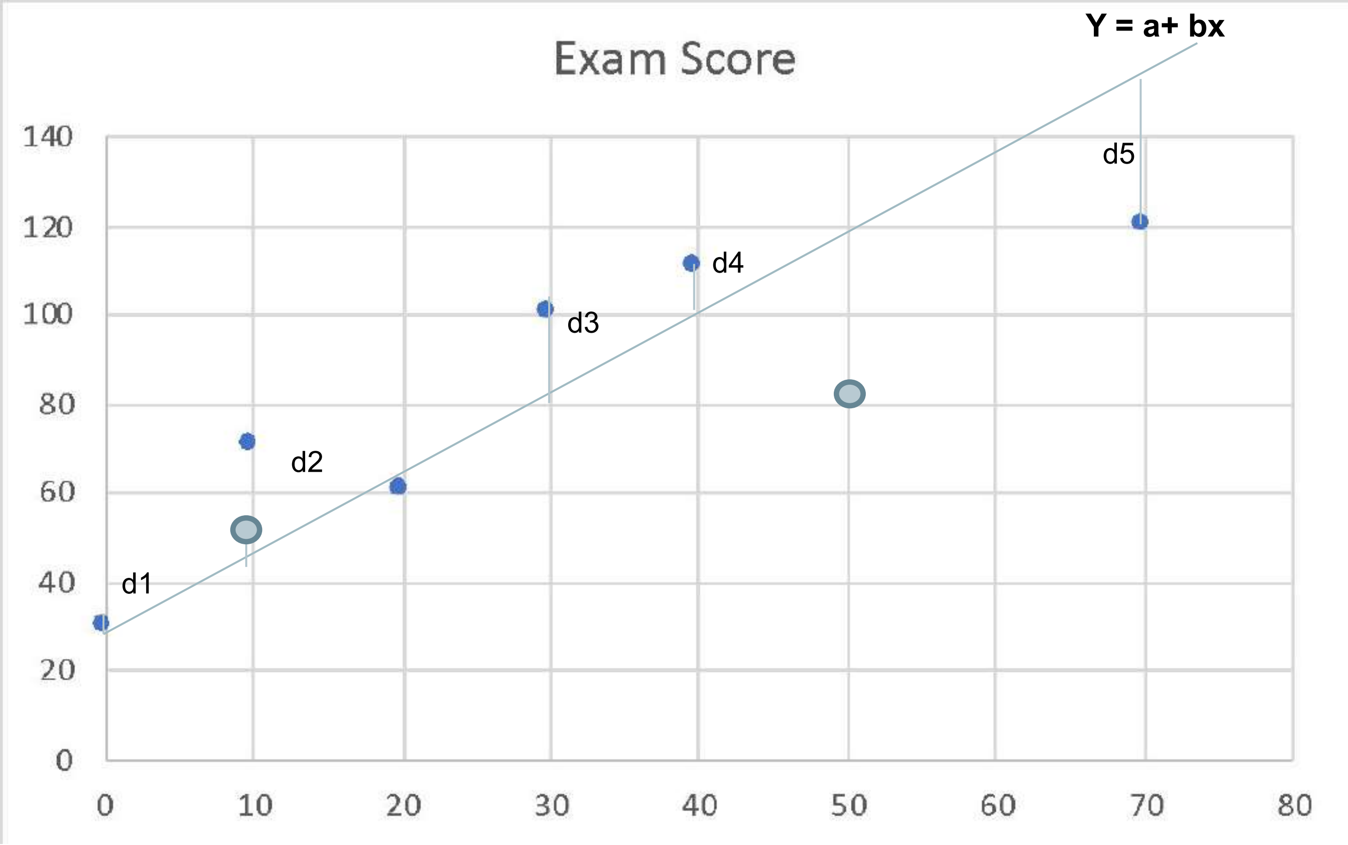
https://towardsdatascience.com/probability-concepts-explained-maximum-likelihood-estimation-c7b4342fdbb1

We need to look for the line that minimizes the error of fit the most

The square of sum total of error between actual value and predicted value of y is called residual error.

$$Residual\ Error = \sum_{y=1}^{n} (y,_{predicted} - y,_{actual})^2$$

Our task while modelling is to identify the right value of a and b, the coefficients of X and Y that minimized the this error.



https://www.quality-control-plan.com/StatGuide/mulreg_alts.htm

Example : Understanding Regression Outputs

A media agency analyst ran a regression model to understand the relationship between advertiser media and the competitive media on a sales KPI.
The following results were obtained from the regression software:

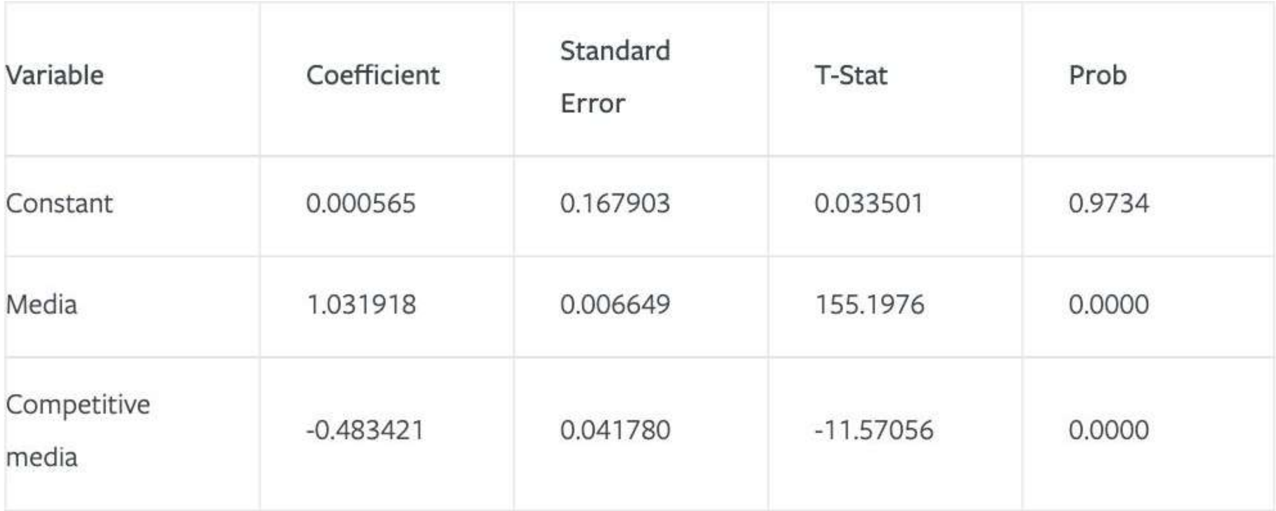
Output:

- Dependent Variable: LOG(SALES)
- Method: Least Squares
- Sample: 01-2016 52-2018
- Included Observations: 150

Statistics:

- R-squared: 0.99853
- Adjusted R-squared: 0.998515
- S.E of regression: 0.01685
- Log-likelihood: 121.4304
- Durbin-Watson: 0.63313
- Akaike info criterion: -5.263574
- Schwartz criterion: -5.143130
- F-Stat: 14979.05
- P(F-statistic): 0.00000

What is the correct interpretation of the results?



Variable	Coefficient	Standard Error	T-Stat	Prob
Constant	0.000565	0.167903	0.033501	0.9734
Media	1.031918	0.006649	155.1976	0.0000
Competitive media	-0.483421	0.041780	-11.57056	0.0000

Example : Understanding Regression Outputs

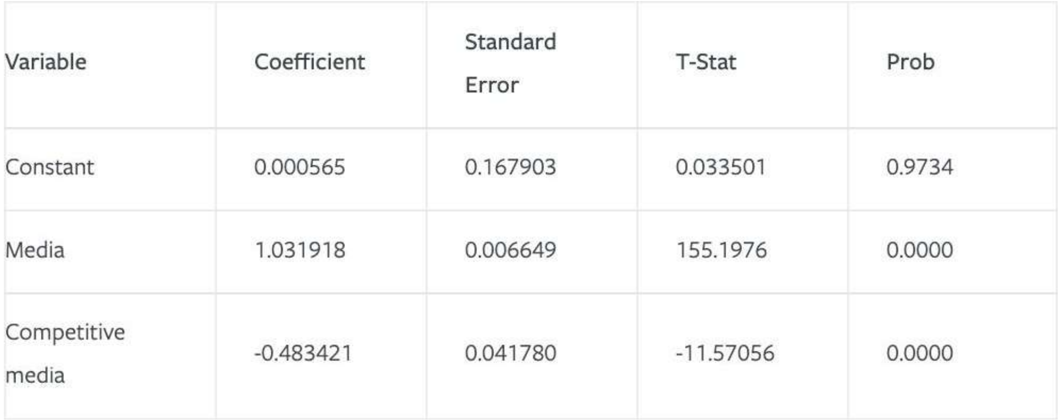
Output:

- Dependent Variable: LOG(SALES)
- Method: Least Squares
- Sample: 01-2016 52-2018
- Included Observations: 150

Statistics:

- R-squared: 0.99853
- Adjusted R-squared: 0.998515
- S.E of regression: 0.01685
- Log-likelihood: 121.4304
- Durbin-Watson: 0.63313
- Akaike info criterion: -5.263574
- Schwartz criterion: -5.143130
- F-Stat: 14979.05
- P(F-statistic): 0.00000

What is the correct interpretation of the results?



Variable	Coefficient	Standard Error	T-Stat	Prob
Constant	0.000565	0.167903	0.033501	0.9734
Media	1.031918	0.006649	155.1976	0.0000
Competitive media	-0.483421	0.041780	-11.57056	0.0000

R-Squared: the proportion of the variance for a dependent variable that's explained by an independent variable or variables in a regression model. Is it also equal squared correlation.

Adjusted R-Squared: modified version of R-squared that has been adjusted for the number of predictors in the model. Always Adj R2 <= R2

Example : Understanding Regression Outputs

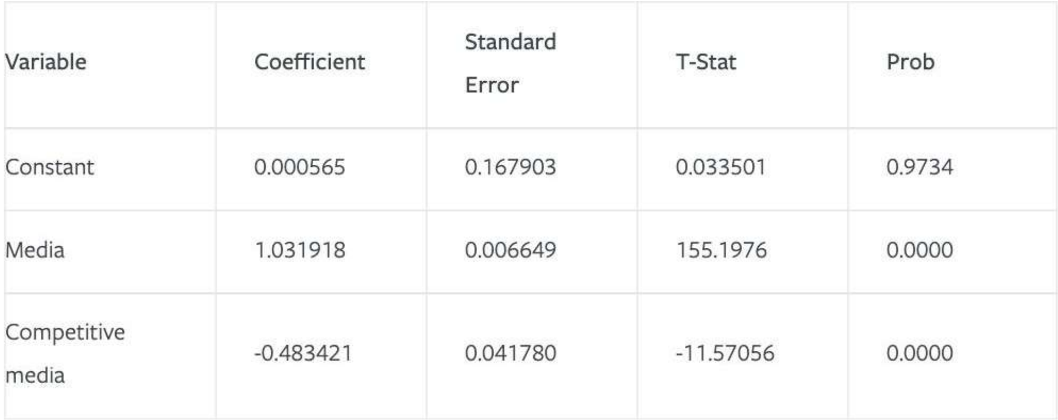
Output:

- Dependent Variable: LOG(SALES)
- Method: Least Squares
- Sample: 01-2016 52-2018
- Included Observations: 150

Statistics:

- R-squared: 0.99853
- Adjusted R-squared: 0.998515
- S.E of regression: 0.01685
- Log-likelihood: 121.4304
- Durbin-Watson: 0.63313
- Akaike info criterion: -5.263574
- Schwartz criterion: -5.143130
- F-Stat: 14979.05
- P(F-statistic): 0.00000

What is the correct interpretation of the results?



Variable	Coefficient	Standard Error	T-Stat	Prob
Constant	0.000565	0.167903	0.033501	0.9734
Media	1.031918	0.006649	155.1976	0.0000
Competitive media	-0.483421	0.041780	-11.57056	0.0000

SE of regression: standard error of the regression provides the absolute measure of the typical distance that the data points fall from the regression line

Log-likelihood: used to compare between models, Log Likelihood value is a measure of goodness of fit for any model. Higher the value, better is the model

Example : Understanding Regression Outputs

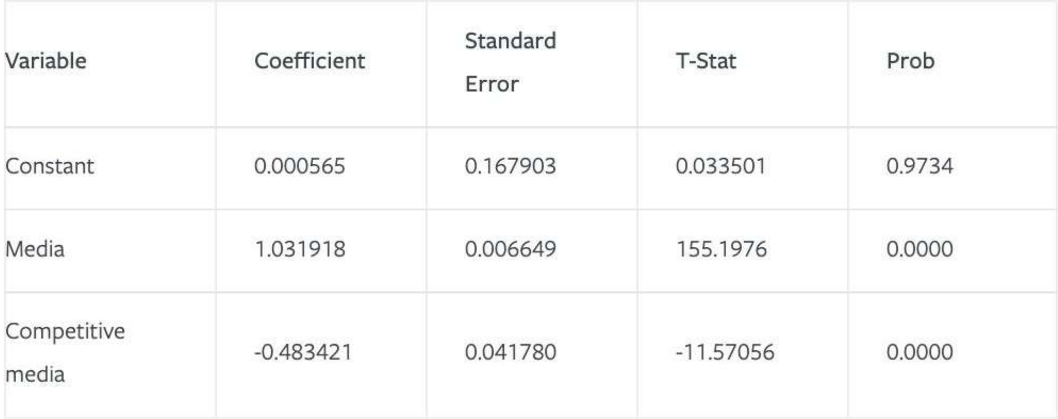
Output:

- Dependent Variable: LOG(SALES)
- Method: Least Squares
- Sample: 01-2016 52-2018
- Included Observations: 150

Statistics:

- R-squared: 0.99853
- Adjusted R-squared: 0.998515
- S.E of regression: 0.01685
- Log-likelihood: 121.4304
- Durbin-Watson: 0.63313
- Akaike info criterion: -5.263574
- Schwartz criterion: -5.143130
- F-Stat: 14979.05
- P(F-statistic): 0.00000

What is the correct interpretation of the results?



Variable	Coefficient	Standard Error	T-Stat	Prob
Constant	0.000565	0.167903	0.033501	0.9734
Media	1.031918	0.006649	155.1976	0.0000
Competitive media	-0.483421	0.041780	-11.57056	0.0000

Durbin-Watson: is a test for Autocorrelation. D-W statistic will always have a value between 0 and 4. A value of 2.0 means that there is no autocorrelation detected in the sample. Values from 0 to less than 2 indicate positive autocorrelation and values from from 2 to 4 indicate negative autocorrelation.

This example has relatively strong positive autocorrelation

Example : Understanding Regression Outputs

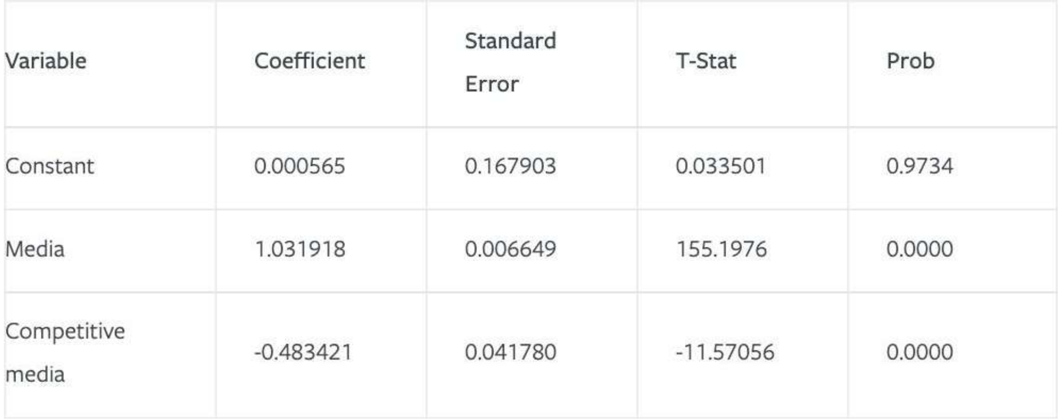
Output:

- Dependent Variable: LOG(SALES)
- Method: Least Squares
- Sample: 01-2016 52-2018
- Included Observations: 150

Statistics:

- R-squared: 0.99853
- Adjusted R-squared: 0.998515
- S.E of regression: 0.01685
- Log-likelihood: 121.4304
- Durbin-Watson: 0.63313
- Akaike info criterion: -5.263574
- Schwartz criterion: -5.143130
- F-Stat: 14979.05
- P(F-statistic): 0.00000

What is the correct interpretation of the results?



Variable	Coefficient	Standard Error	T-Stat	Prob
Constant	0.000565	0.167903	0.033501	0.9734
Media	1.031918	0.006649	155.1976	0.0000
Competitive media	-0.483421	0.041780	-11.57056	0.0000

Akaike & Schwartz criterions: Akaike information criterion (AIC) is an estimator of out-of-sample prediction error and thereby relative quality of statistical models for a given set of data. A lower AIC value indicates a better fit.

Schwartz criterions is also called Bayesian information criterion (BIC) and has similar function to AIC. The model with the lowest BIC is preferred.

Example : Understanding Regression Outputs

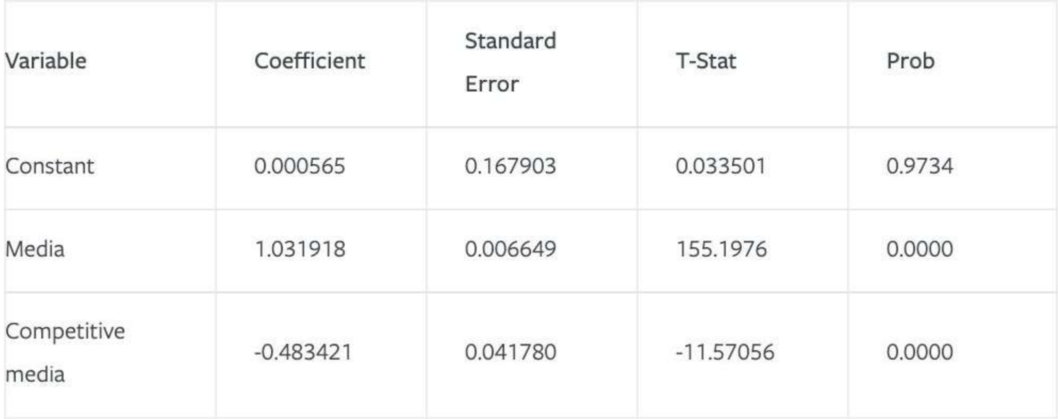
Output:

- Dependent Variable: LOG(SALES)
- Method: Least Squares
- Sample: 01-2016 52-2018
- Included Observations: 150

Statistics:

- R-squared: 0.99853
- Adjusted R-squared: 0.998515
- S.E of regression: 0.01685
- Log-likelihood: 121.4304
- Durbin-Watson: 0.63313
- Akaike info criterion: -5.263574
- Schwartz criterion: -5.143130
- F-Stat: 14979.05
- P(F-statistic): 0.00000

What is the correct interpretation of the results?



Variable	Coefficient	Standard Error	T-Stat	Prob
Constant	0.000565	0.167903	0.033501	0.9734
Media	1.031918	0.006649	155.1976	0.0000
Competitive media	-0.483421	0.041780	-11.57056	0.0000

F-stat of regression: The F-test of overall significance indicates whether your linear regression model provides a better fit to the data than a model that contains no independent variables.

If the overall equation is significant it must be greater than 10 as a rule of thumb

Example : Understanding Regression Outputs

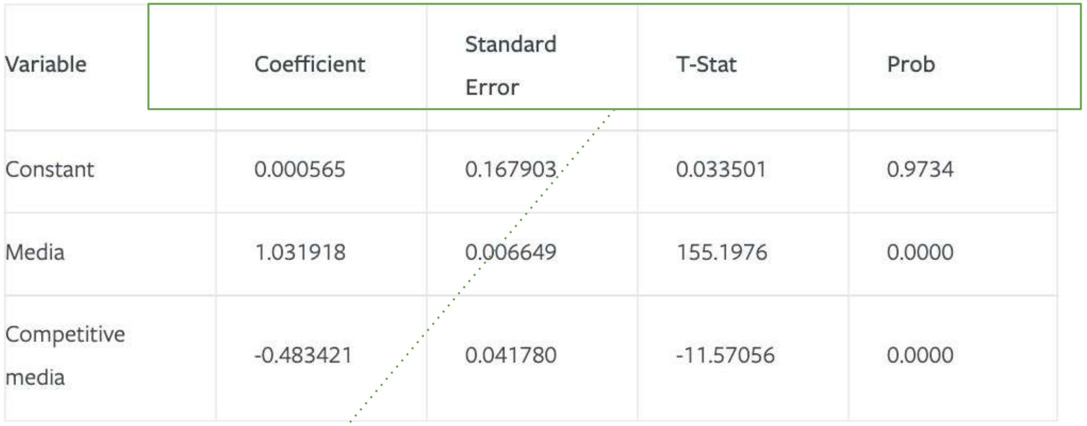
Output:

- Dependent Variable: LOG(SALES)
- Method: Least Squares
- Sample: 01-2016 52-2018
- Included Observations: 150

Statistics:

- R-squared: 0.99853
- Adjusted R-squared: 0.998515
- S.E of regression: 0.01685
- Log-likelihood: 121.4304
- Durbin-Watson: 0.63313
- Akaike info criterion: -5.263574
- Schwartz criterion: -5.143130
- F-Stat: 14979.05
- P(F-statistic): 0.00000

What is the correct interpretation of the results?



Variable	Coefficient	Standard Error	T-Stat	Prob
Constant	0.000565	0.167903	0.033501	0.9734
Media	1.031918	0.006649	155.1976	0.0000
Competitive media	-0.483421	0.041780	-11.57056	0.0000

Coefficient: coefficients are the values that multiply the predictor values. The sign of each coefficient indicates the direction of the relationship between a predictor variable and the dependent variable
SE: The standard deviation of an estimate. The standard error of the coefficient measures how precisely the model estimates the coefficient's unknown value.
T-stat = Coefficient / SE
T-stat & Prob: It is standard practice to use the coefficient t-stats & p-values to decide whether to include variables in the final mode

Example : Understanding Regression Outputs

A media agency analyst ran a regression model to understand the relationship between advertiser media and the competitive media on a sales KPI.
The following results were obtained from the regression software:

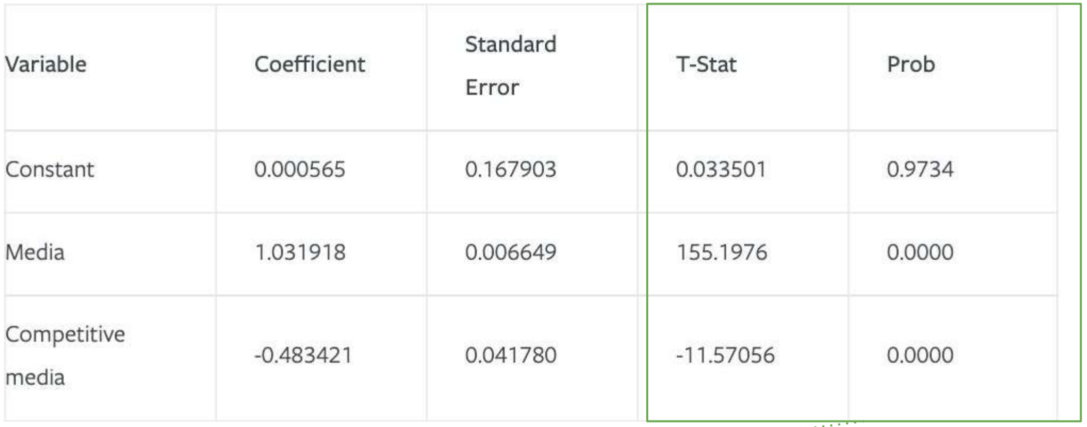
Output:

- Dependent Variable: LOG(SALES)
- Method: Least Squares
- Sample: 01-2016 52-2018
- Included Observations: 150

Statistics:

- R-squared: 0.99853
- Adjusted R-squared: 0.998515
- S.E of regression: 0.01685
- Log-likelihood: 121.4304
- Durbin-Watson: 0.63313
- Akaike info criterion: -5.263574
- Schwartz criterion: -5.143130
- F-Stat: 14979.05
- P(F-statistic): 0.00000

What is the correct interpretation of the results?



Variable	Coefficient	Standard Error	T-Stat	Prob
Constant	0.000565	0.167903	0.033501	0.9734
Media	1.031918	0.006649	155.1976	0.0000
Competitive media	-0.483421	0.041780	-11.57056	0.0000

With >100 degrees of freedom T-Stats over 1.8 are 95% stat sig. Always look at the p-value (Prob).

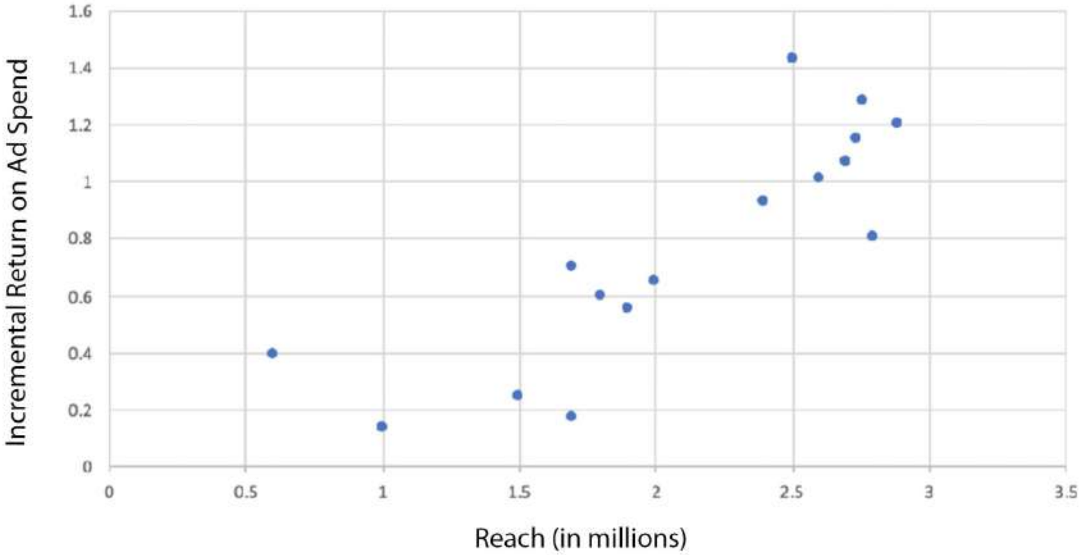
Differentiating between causality and correlation

An ecommerce company wants to understand the impact of reach on their incremental ROAS results when running randomized control trial experiments using their ads.

Refer to the chart.

What conclusion should the analytics team make ir respect to these findings?

- These results are correlative, not causal
- There is not a discernible relationship between reach and ROAS
- Reach should be optimized for 2.5 million unique users because that drove the highest incremental ROAS
- Higher reach causes higher ROAS

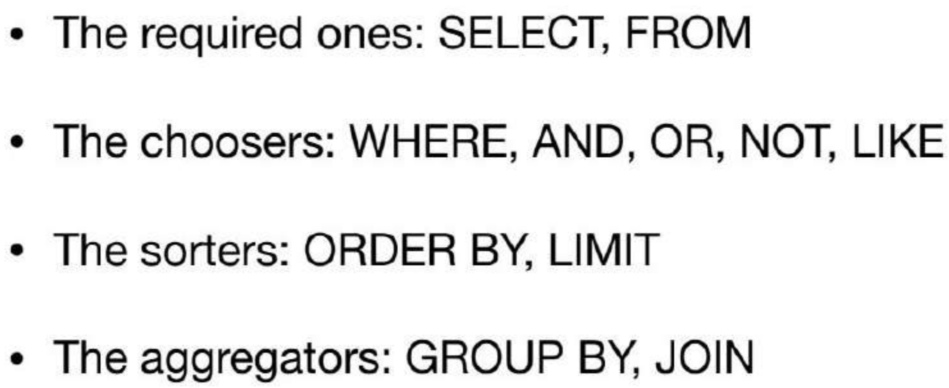


Extract & manipulate data: SQL basics

https://www.khanacademy.org/computing/computer-programming/sql

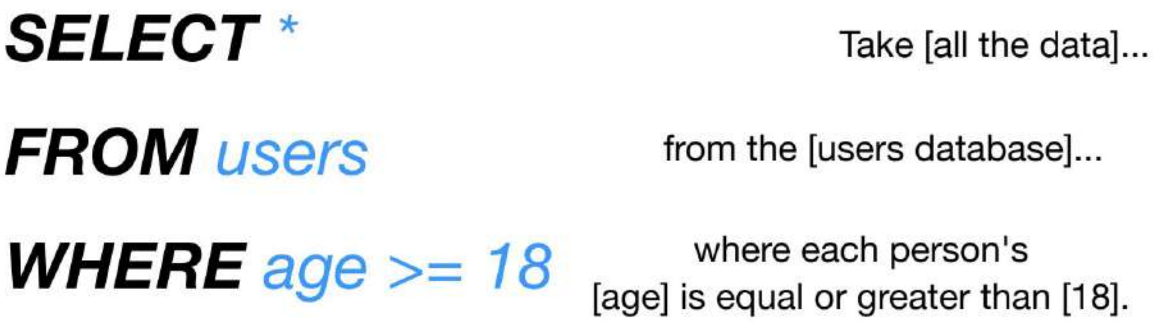
https://www.w3schools.com/sql/default.asp - with interactive exercises

Basic Query Structure

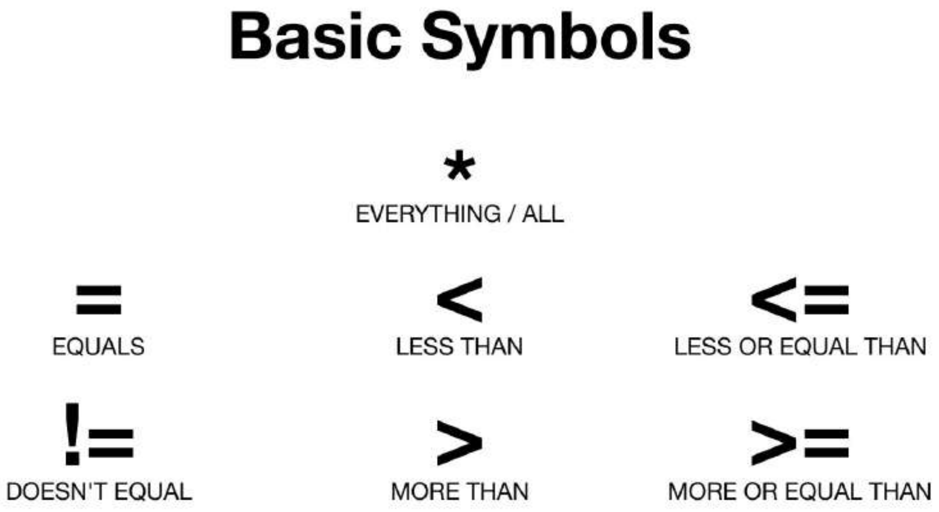


- The required ones: SELECT, FROM
- The choosers: WHERE, AND, OR, NOT, LIKE
- The sorters: ORDER BY, LIMIT
- The aggregators: GROUP BY, JOIN

For example:



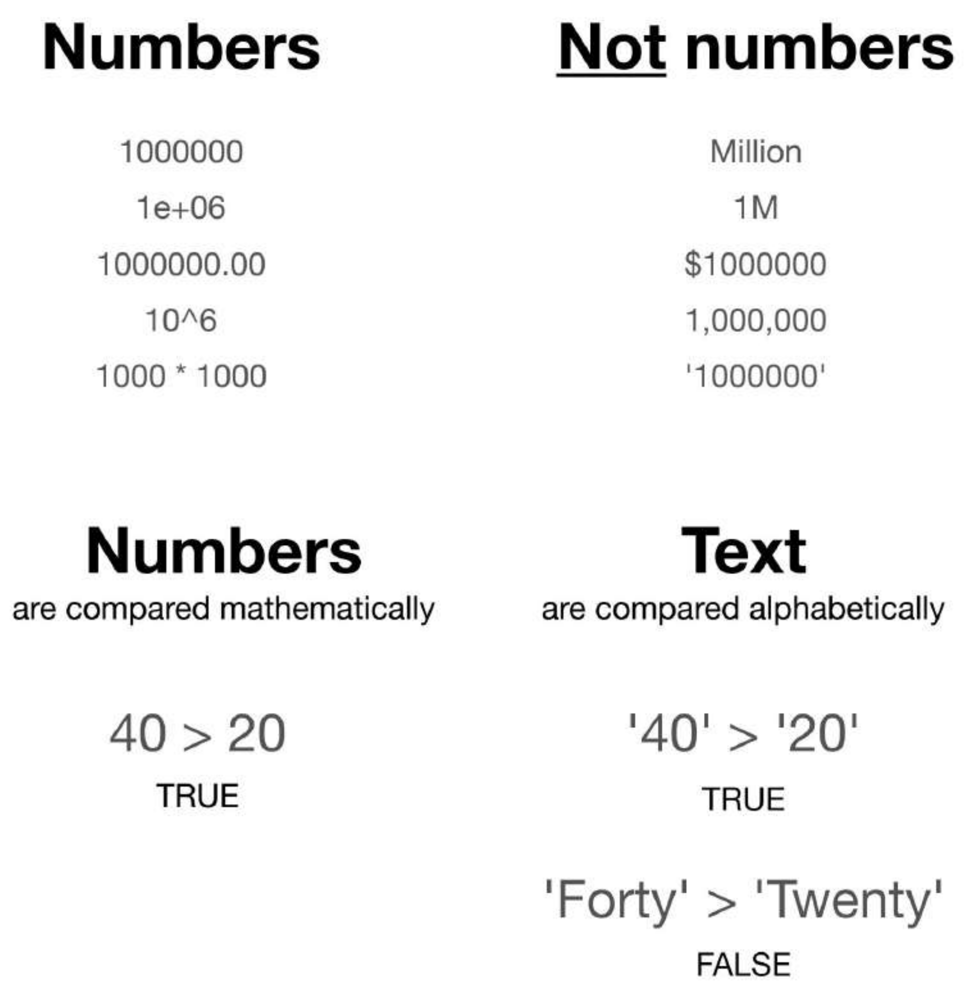
SELECT *	Take [all the data]...
FROM users	from the [users database]...
WHERE age >= 18	where each person's [age] is equal or greater than [18].



Basic Symbols

	* EVERYTHING / ALL	
= EQUALS	< LESS THAN	<= LESS OR EQUAL THAN
!= DOESN'T EQUAL	> MORE THAN	>= MORE OR EQUAL THAN

Numbers or text?



Numbers	Not numbers
1000000	Million
1e+06	1M
1000000.00	$1000000
10^6	1,000,000
1000 * 1000	'1000000'

Numbers

are compared mathematically

40 > 20

TRUE

Text

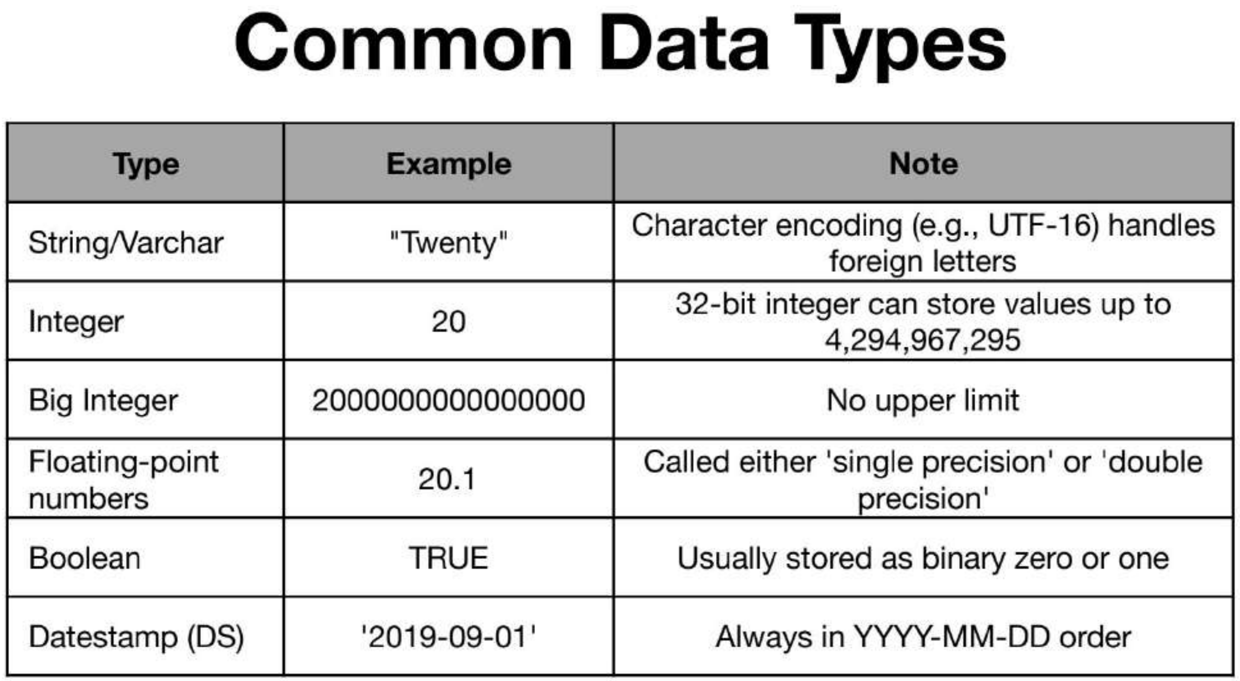
are compared alphabetically

'40' > '20'

TRUE

'Forty' > 'Twenty'

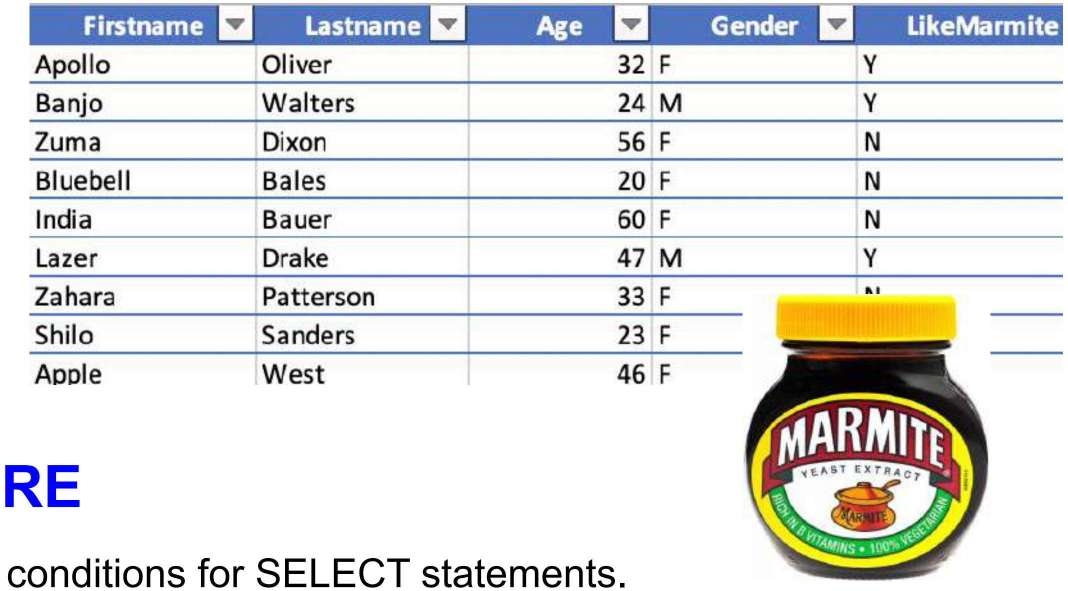
FALSE



Common Data Types

Type	Example	Note
String/Varchar	"Twenty"	Character encoding (e.g., UTF-16) handles foreign letters
Integer	20	32-bit integer can store values up to 4,294,967,295
Big Integer	2000000000000000	No upper limit
Floating-point numbers	20.1	Called either 'single precision' or 'double precision'
Boolean	TRUE	Usually stored as binary zero or one
Datestamp (DS)	'2019-09-01'	Always in YYYY-MM-DD order

SELECT - FROM - WHERE



Firstname	Lastname	Age	Gender	LikeMarmite
Apollo	Oliver	32	F	Y
Banjo	Walters	24	M	Y
Zuma	Dixon	56	F	N
Bluebell	Bales	20	F	N
India	Bauer	60	F	N
Lazer	Drake	47	M	Y
Zahara	Patterson	33	F	N
Shilo	Sanders	23	F	
Apple	West	46	F	

SELECT / FROM

Get data from a table in a database. Result: a new data table!

- **SELECT** [something] **FROM** [somewhere]
  - [something]: specific column(s) in a data table, or * (meaning all the columns)
  - [somewhere]: the name of the data table that contains the above columns
- Example: *get the first names and ages from the my_team table*

  SELECT firstname, age
  FROM my_team

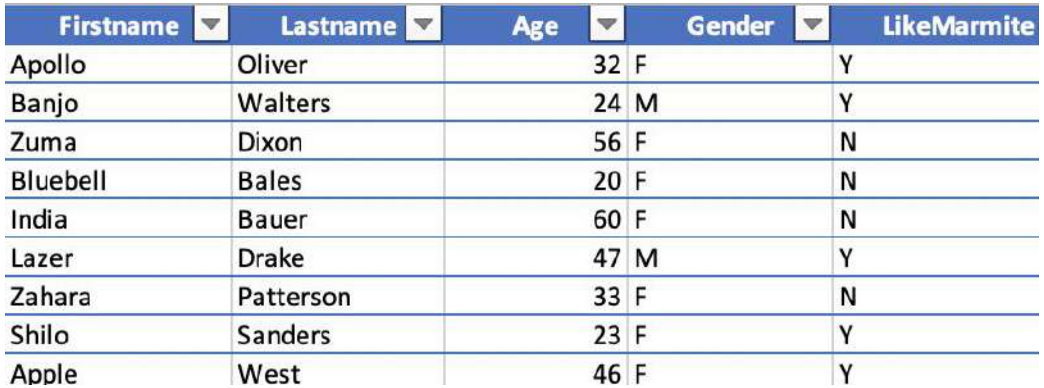
WHERE

Specify conditions for SELECT statements.

- **SELECT** [something] **FROM** [somewhere] **WHERE** some_condition
- Example: *get the first names and ages from the my_team table, of everyone who is female and likes Marmite*

  SELECT firstname, age
  FROM my_team
  WHERE gender = ‘F’ **AND** likemarmite = ‘Y’
- Use **AND**, **OR** to specify more than one condition

Sorters



Firstname	Lastname	Age	Gender	LikeMarmite
Apollo	Oliver	32	F	Y
Banjo	Walters	24	M	Y
Zuma	Dixon	56	F	N
Bluebell	Bales	20	F	N
India	Bauer	60	F	N
Lazer	Drake	47	M	Y
Zahara	Patterson	33	F	N
Shilo	Sanders	23	F	Y
Apple	West	46	F	Y

ORDER BY

Arrange the results in a certain order

- Goes after **FROM** & **WHERE** at the end of the query
- Looks like:
  - **ORDER BY** column_name
- Can include **ASC** or **DESC**
- Example:

SELECT firstname, age
FROM my_team
WHERE gender = ‘F’
ORDER BY lastname **DESC**

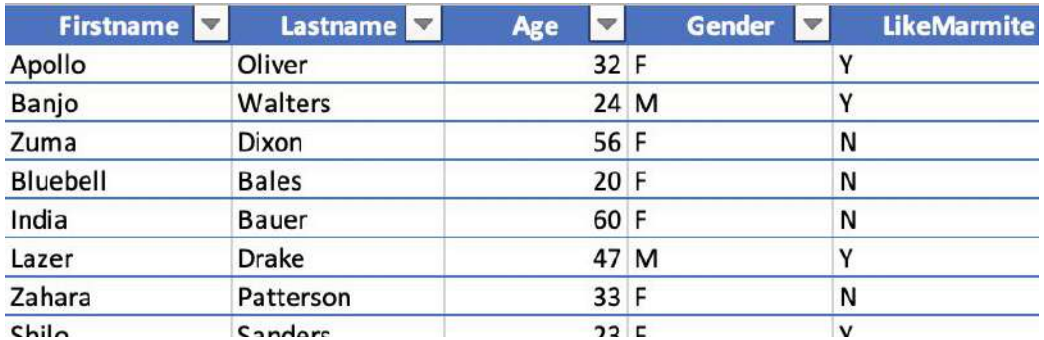
LIMIT

For queries that return very long results, the LIMIT clause restricts the number of rows in the result set

- Goes after **FROM** & **WHERE** at the end of the query
- Looks like:
  - **LIMIT** number
- Example:

SELECT firstname, age
FROM my_team
WHERE gender = ‘F’
LIMIT 100

Aggregators



Firstname	Lastname	Age	Gender	LikeMarmite
Apollo	Oliver	32	F	Y
Banjo	Walters	24	M	Y
Zuma	Dixon	56	F	N
Bluebell	Bales	20	F	N
India	Bauer	60	F	N
Lazer	Drake	47	M	Y
Zahara	Patterson	33	F	N
Shilo	Sanders	23	F	Y

COUNT

Returns the *number* of input values

- Goes after **SELECT**
- Looks like:
  - **COUNT** (#)
  - input_value either * or column name, or a number
  - Note: COUNT (*) = count all, COUNT (column_name) = count only non-null values
- Example: *How many people in my team like Marmite?*

```
SELECT COUNT (*)
FROM my_team
WHERE likemarmite = ‘Y’
```

GROUP BY

Divides the output of a SELECT statement into groups of rows containing matching values.

- Goes at the end of **SELECT** queries
- Looks like:
  - **GROUP BY** column_name
  - Note: column_name must also appear after SELECT
- Example: *First names of all my team members grouped by gender*

```
SELECT firstname, gender,
FROM my_team
WHERE likemarmite = ‘Y’
GROUP BY gender
```

MATH

Returns a functional result of input values, based on mathematical function

- Functions include: **SUM**, **AVG**, **MIN**, **MAX** etc.
- Often goes after **SELECT**
- Looks like:
  - **AVG** (column_name), **MIN** (column_name) etc.
- Example: *What is the average age of my team members per gender?*

```
SELECT gender, AVG (age) AS
avg_age
FROM my_team
WHERE likemarmite = ‘Y’
GROUP BY gender
```

rename a column (alias)

Others

DISTINCT

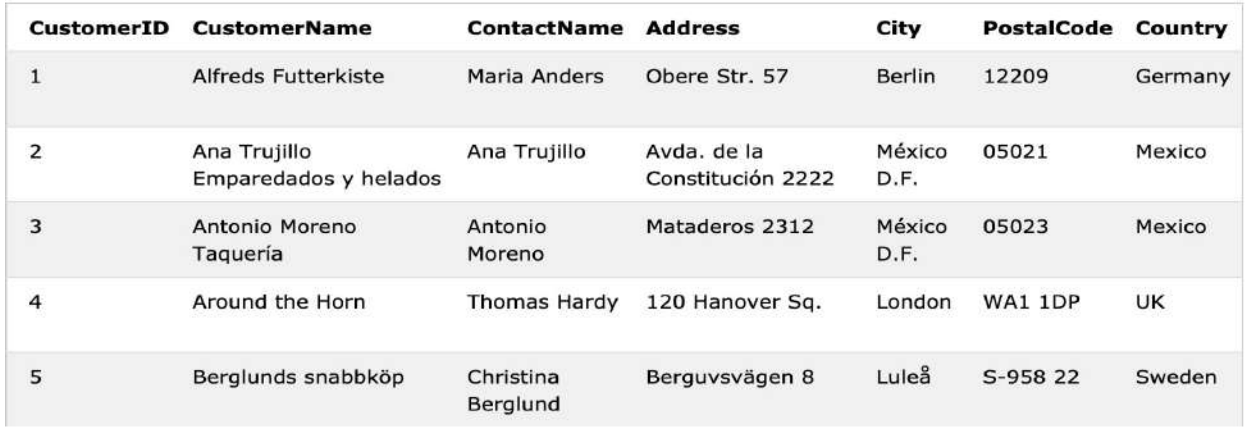
Remove duplicates from results, return only distinct (different) values

- Goes after **SELECT**
- Looks like:
    - **SELECT DISTINCT** something **FROM** somewhere
- Example: *list all countries where I have customers:*

  SELECT DISTINCT country
  FROM customer_table

- Example: *In how many countries do I have customers?*

  SELECT COUNT (DISTINCT Country)
  FROM customer_table



CustomerID	CustomerName	ContactName	Address	City	PostalCode	Country
1	Alfreds Futterkiste	Maria Anders	Obere Str. 57	Berlin	12209	Germany
2	Ana Trujillo Emparedados y helados	Ana Trujillo	Avda. de la Constitución 2222	México D.F.	05021	Mexico
3	Antonio Moreno Taquería	Antonio Moreno	Mataderos 2312	México D.F.	05023	Mexico
4	Around the Horn	Thomas Hardy	120 Hanover Sq.	London	WA1 1DP	UK
5	Berglunds snabbköp	Christina Berglund	Berguvsvägen 8	Luleå	S-958 22	Sweden

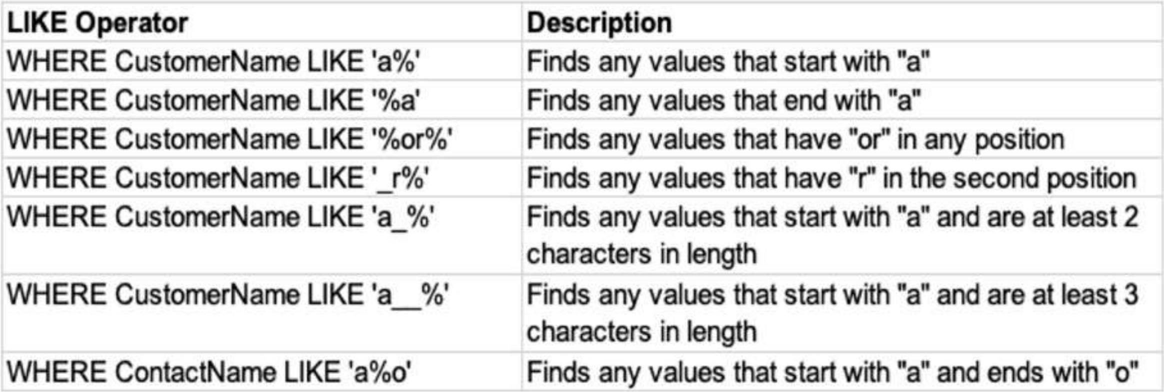
LIKE

Used in a WHERE clause to search for a specified text pattern (% is used to match any characters)

- Looks like:

  SELECT column1, column2 **FROM** table_name
  WHERE columnN **LIKE** pattern

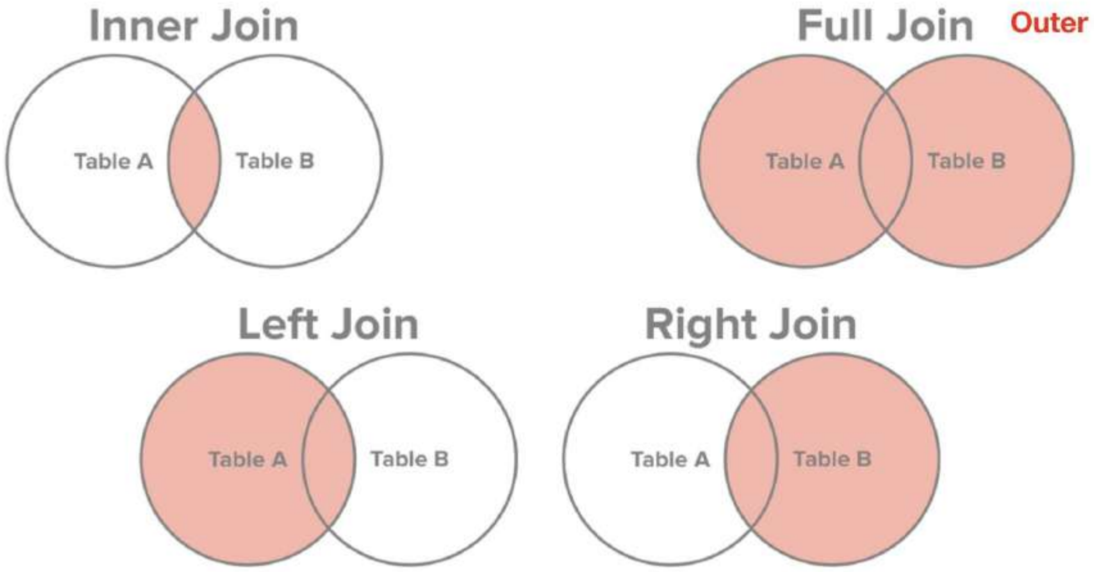
- Two wildcards often used with the LIKE operator:
    - **%** zero, one, or multiple characters
    - **_** a single character



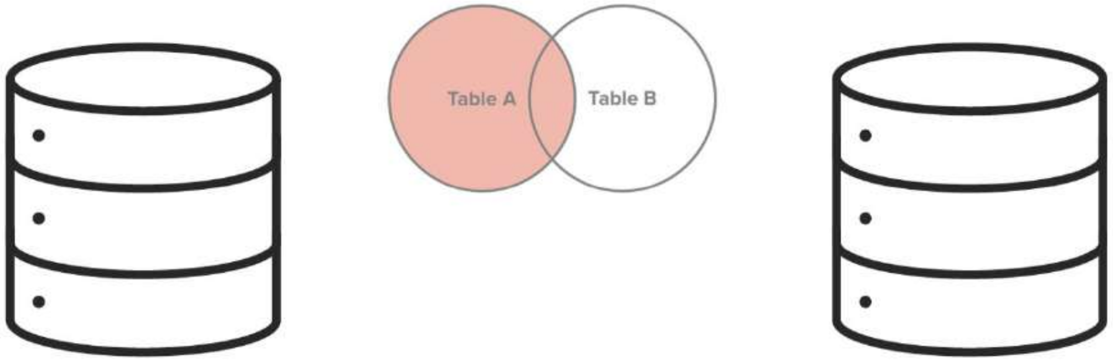
LIKE Operator	Description
WHERE CustomerName LIKE 'a%'	Finds any values that start with "a"
WHERE CustomerName LIKE '%a'	Finds any values that end with "a"
WHERE CustomerName LIKE '%or%'	Finds any values that have "or" in any position
WHERE CustomerName LIKE '_r%'	Finds any values that have "r" in the second position
WHERE CustomerName LIKE 'a_%'	Finds any values that start with "a" and are at least 2 characters in length
WHERE CustomerName LIKE 'a__%'	Finds any values that start with "a" and are at least 3 characters in length
WHERE ContactName LIKE 'a%o'	Finds any values that start with "a" and ends with "o"

JOINS

Most commonly used JOINS



Example: LEFT JOIN



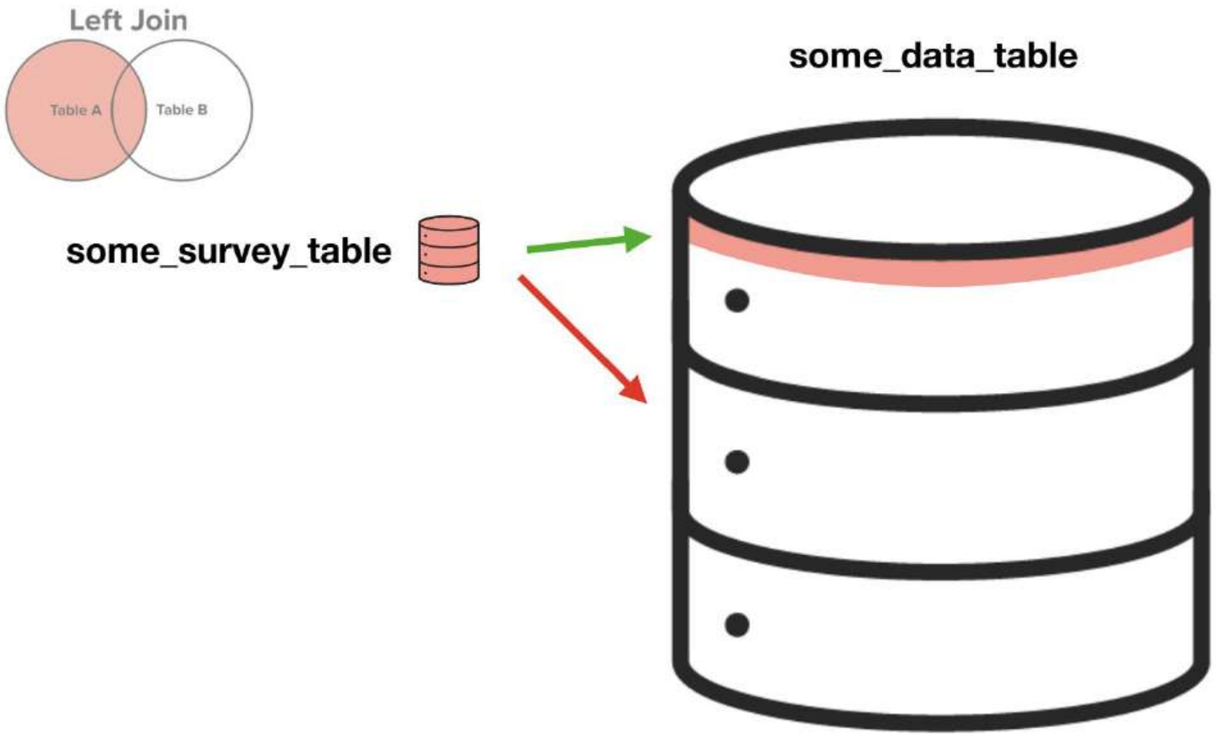
some_survey_table

- has maybe 1,000 - 100,000 rows
- you want to keep it all
- you care about JOINing data to augment survey responses

some_data_table

- can have 100 million+ rows
- you only want data for your survey takers
- you don't care about the rest

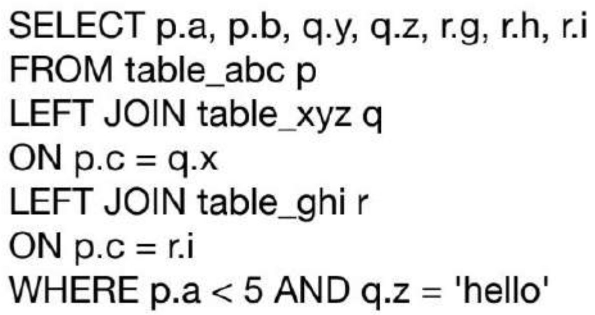
Example: LEFT JOIN



- JOIN is used after FROM clause; denotes a second table
- The ON statement shows which column to match between tables
- Example:

```
SELECT a, b, y, z
FROM table_abc
LEFT JOIN table_xyz
ON table_abc.c = table_xyz.x
WHERE a < 5 AND z = ‘hello’
```

- Example using multiple tables:



```
SELECT p.a, p.b, q.y, q.z, r.g, r.h, r.i
FROM table_abc p
LEFT JOIN table_xyz q
ON p.c = q.x
LEFT JOIN table_ghi r
ON p.c = r.i
WHERE p.a < 5 AND q.z = 'hello'
```

Questions?