How to Become a Data Scientist

Francesco Tisiot
Analytics Tech Lead

rittmanmead
A DATA AND ANALYTICS COMPANY

Francesco Tisiot

Analytics Tech Lead

Verona, Italy

http://ritt.md/ftisiot

ft@rittmanmead.com

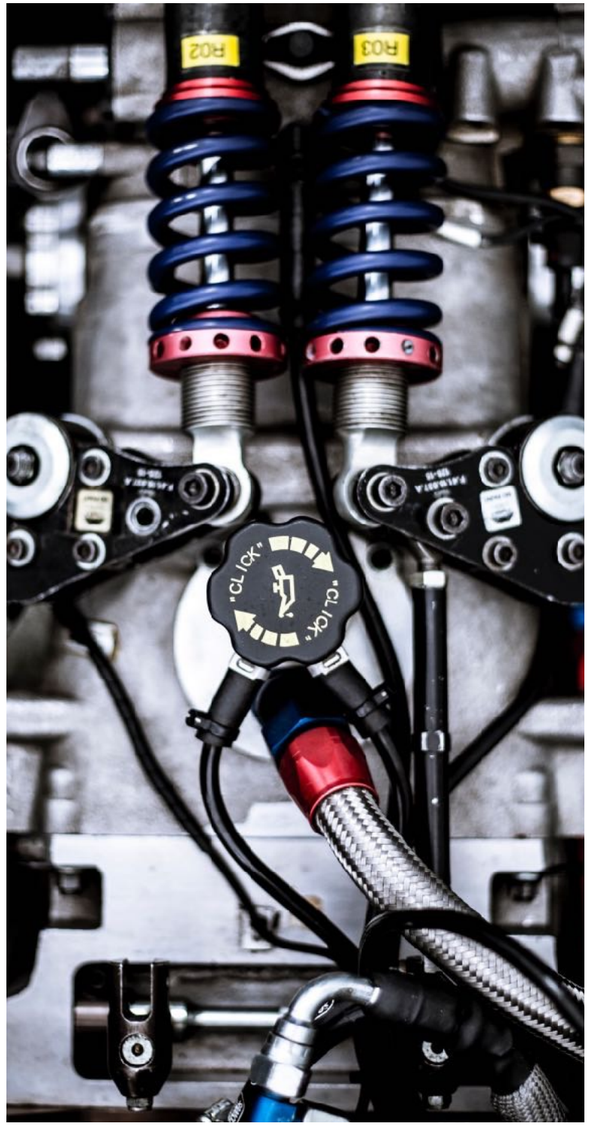
@FTisiot

Oracle ACE Director

rittmanmead

A DATA AND ANALYTICS COMPANY

info@rittmanmead.com
www.rittmanmead.com
@rittmanmead



Data Engineering

Analytics

Data Science

Data Scientist


R
Python
Linear Algebra
Statistics

Low Hanging Fruit Theory

Democratise
Data Science

Basic Operations

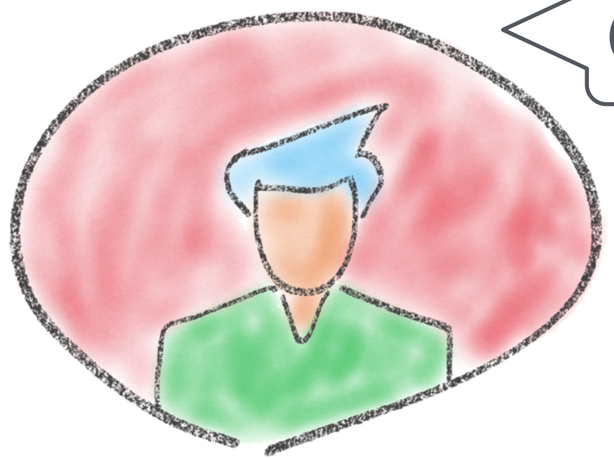
Based on my
Experience
I can Guess….

What are the
Drivers for
My Sales?

The **Statistically Significant Drivers** for Sales Are …

Augmented Analytics

Basic Operations



YES/NO

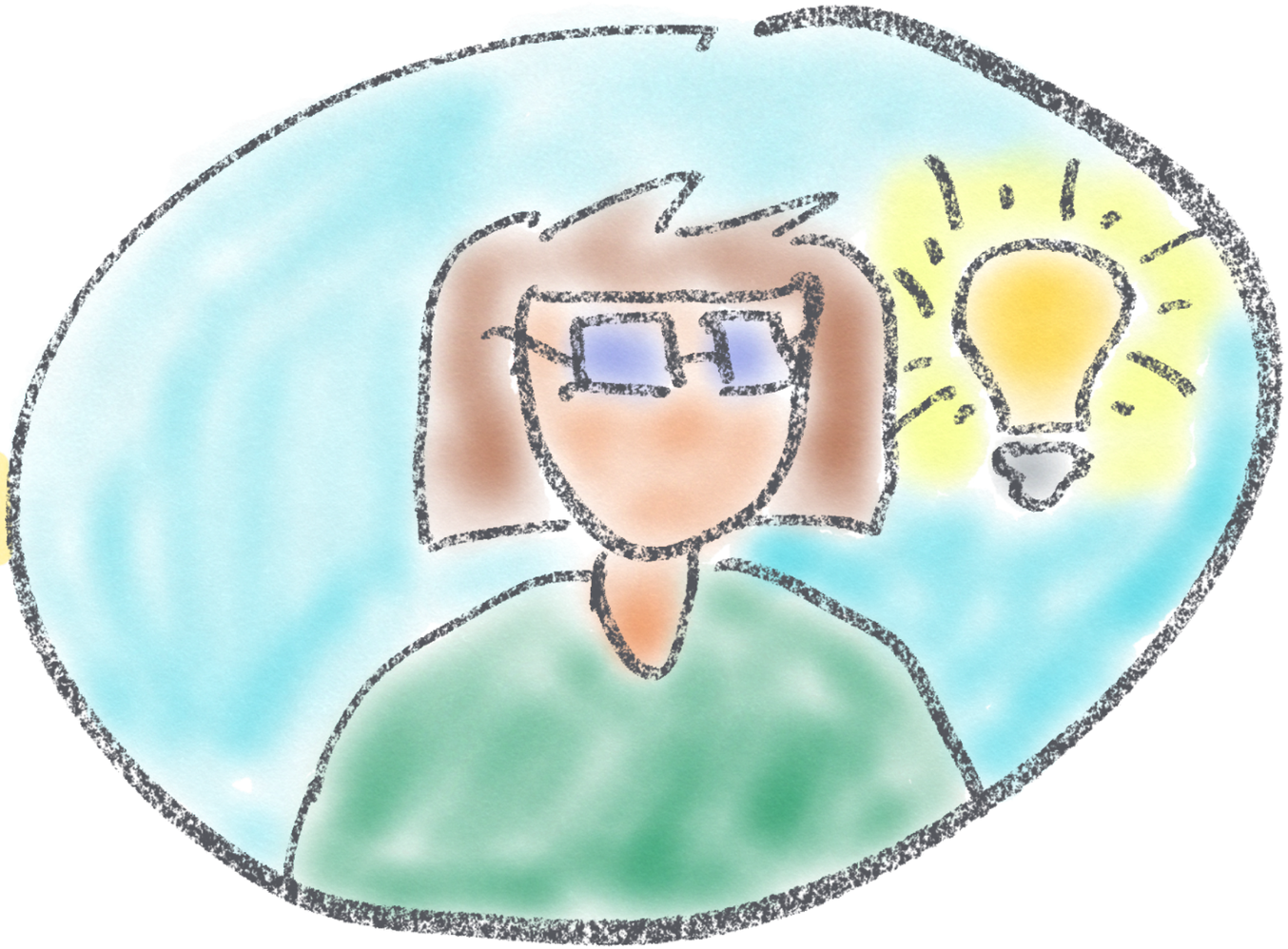
50%

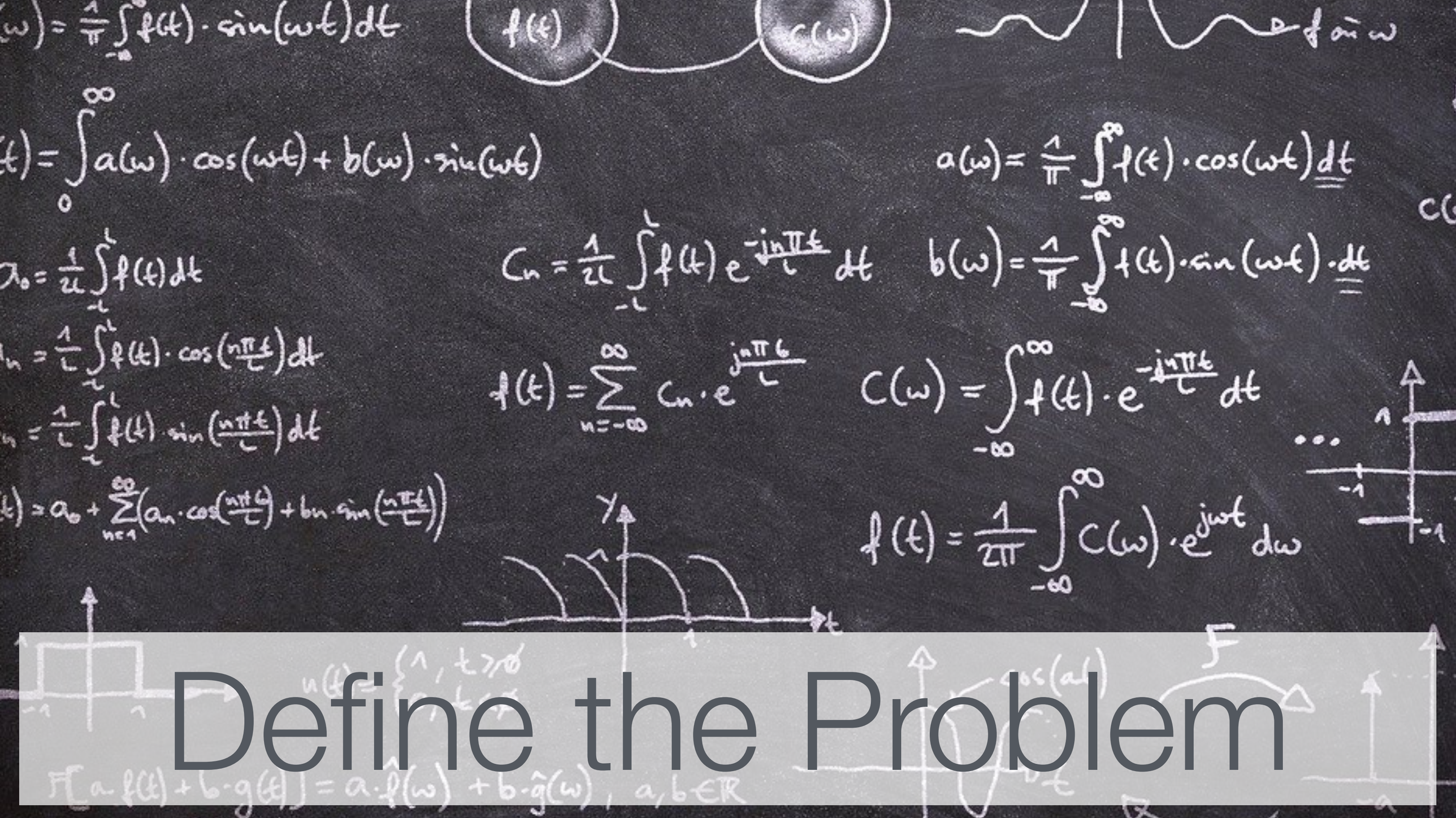
Basic ML Model

70%

Oracle Analytics Cloud

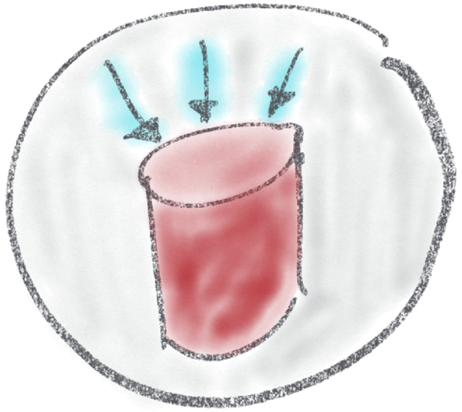
How Do I Become Data Scientists with OAC?



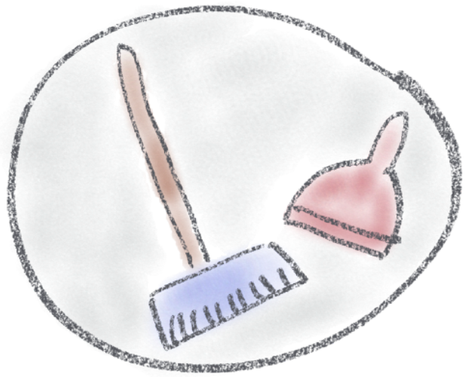


Define the Problem

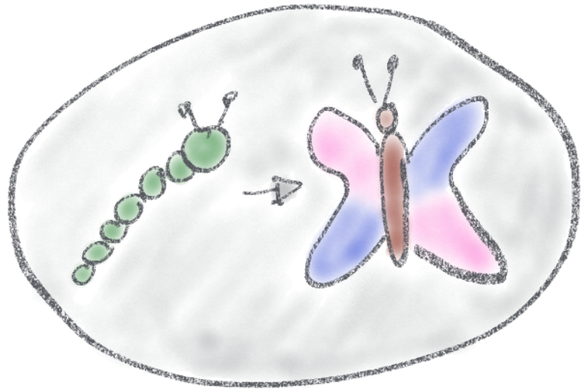
6 Steps into Data Science



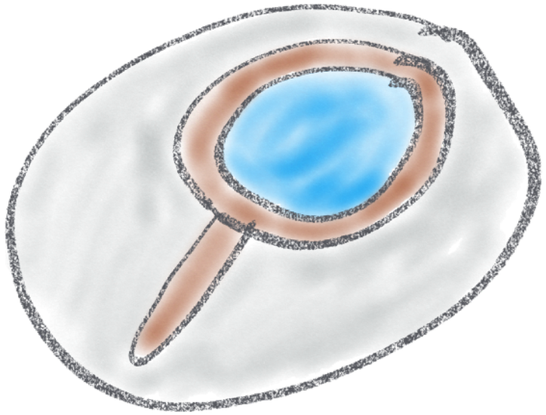
1. Connect



2. Clean

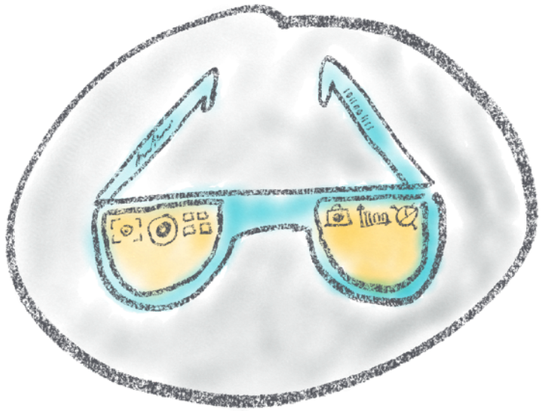


3. Transform & Enrich



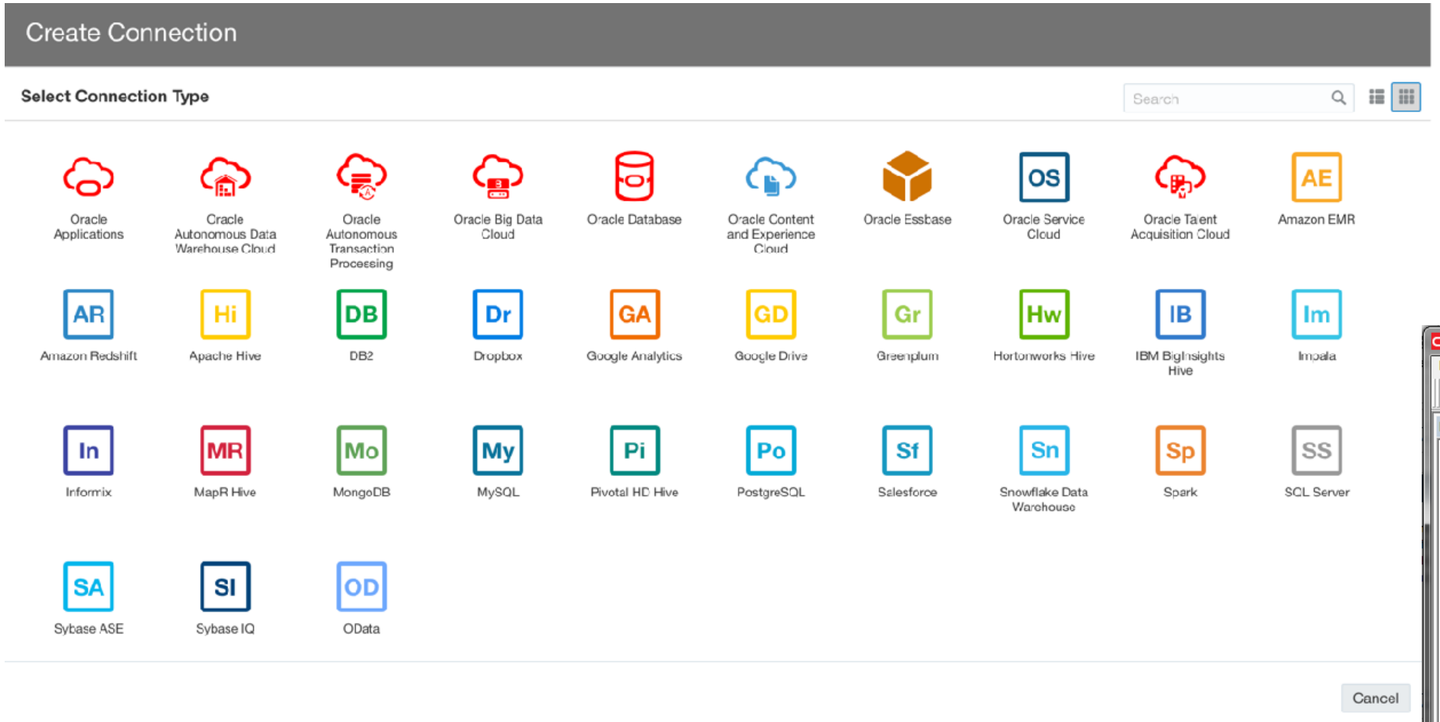
4. Analyse

5. Transform & Enrich



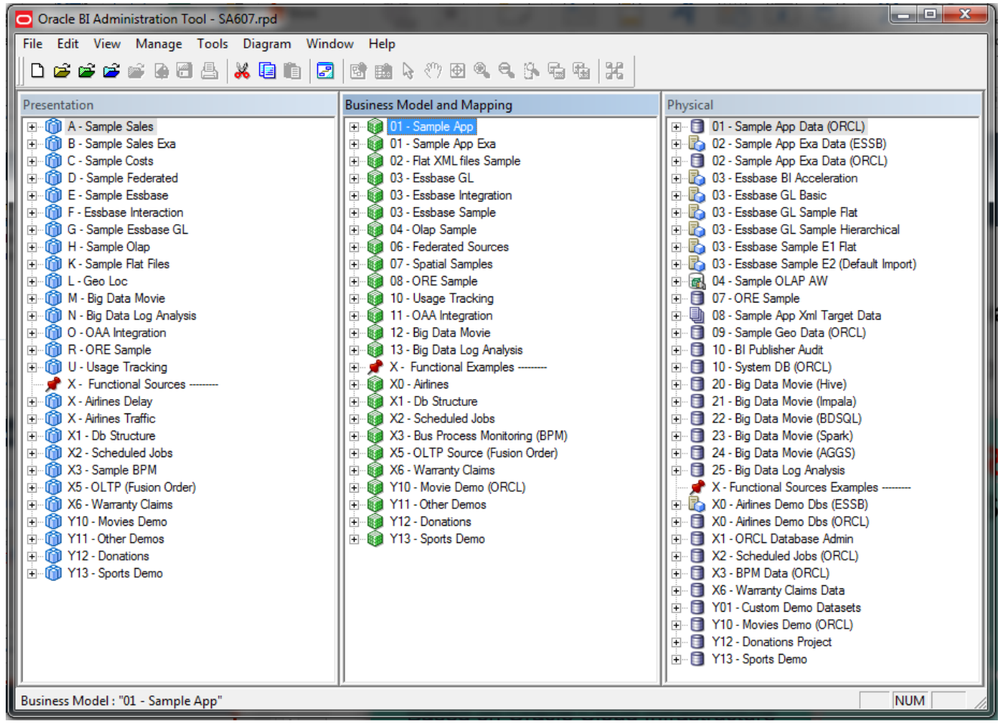
6. Predict

Connection Options



External
Data Sources

Pre-Defined
Data Models



Clean

N/A

Missing Values

Mark <> MArk

Wrong Values

City
“Rome”

Irrelevant Observations

Col1 -> Name

Labelling Columns

Role: CIO
Salary:500 K$

Handling Outliers

0-200k
0-1

Feature Scaling

Of Clicks

Aggregation

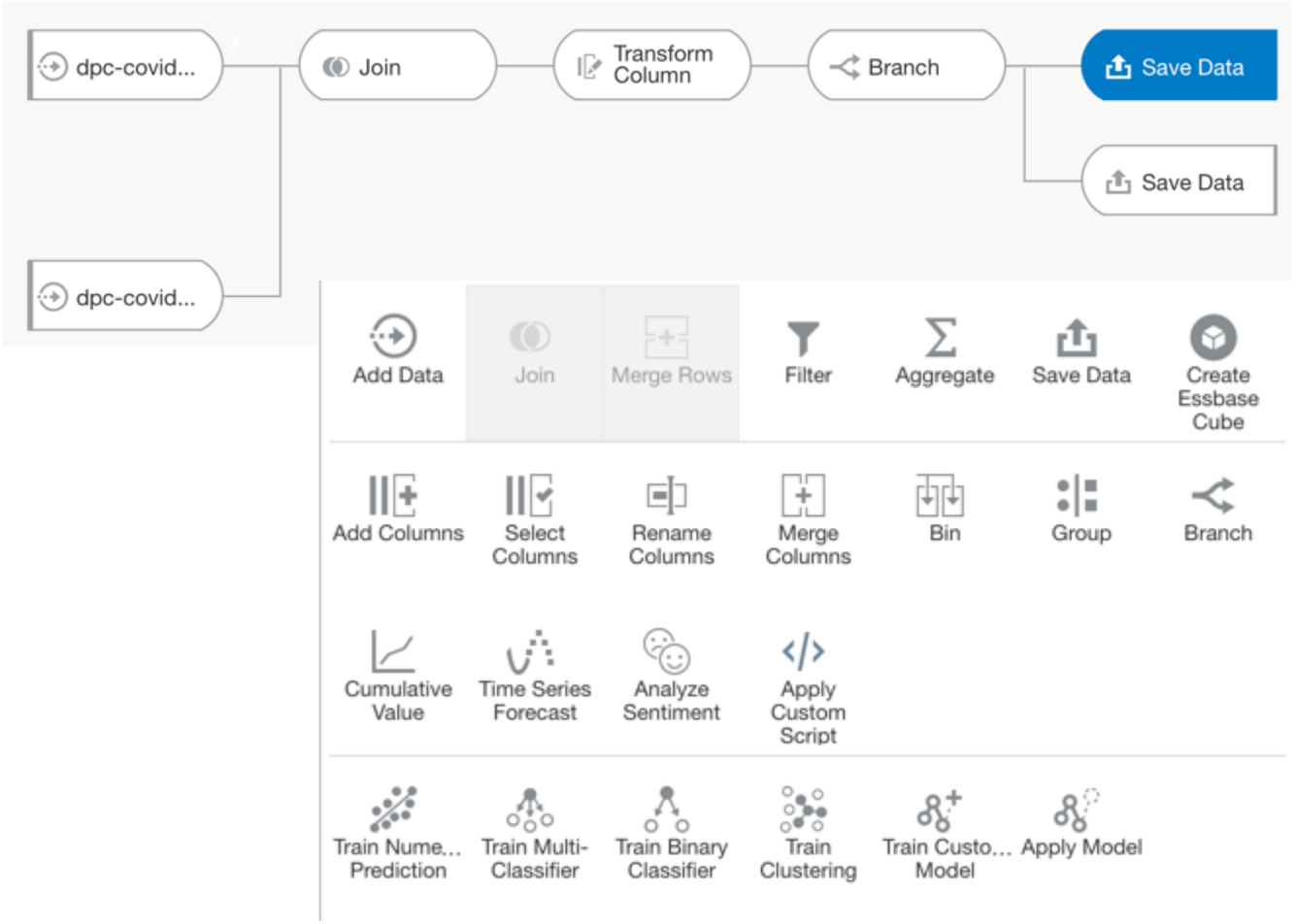
Train: 80%
Test: 20%

Train/Test Set Split

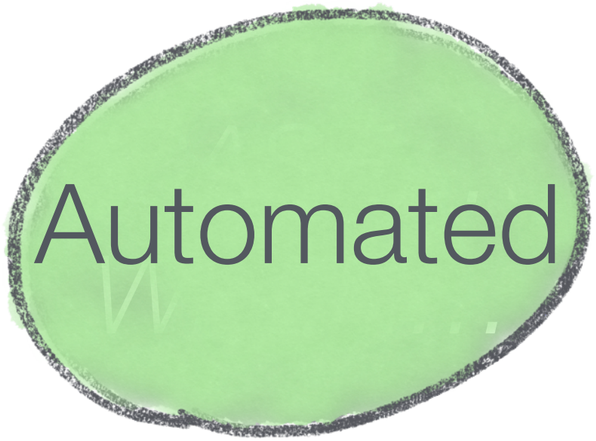
Clean…How?

Data Flows

- Filter
- Aggregate
- Join
- Transform



Clean

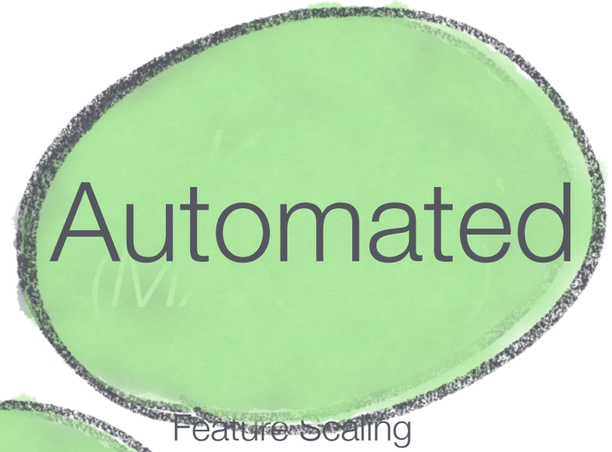


Wrong Cases

Irrelevant Observations

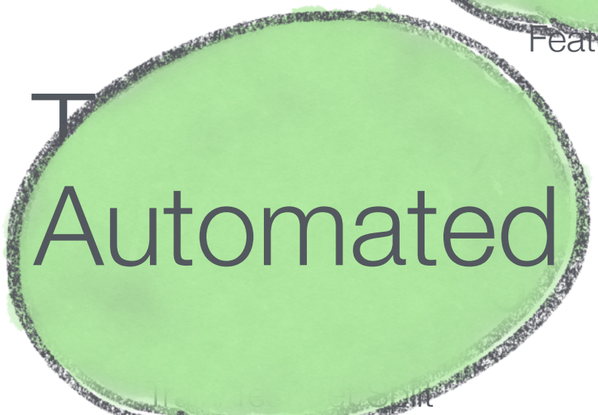
Labelling Columns

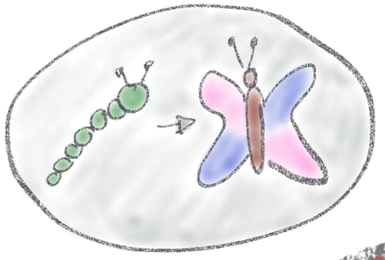
Handling Outliers



Feature Scaling

Aggregation





Feature Engineering

Location -> ZIP Code

Additional Data Source?

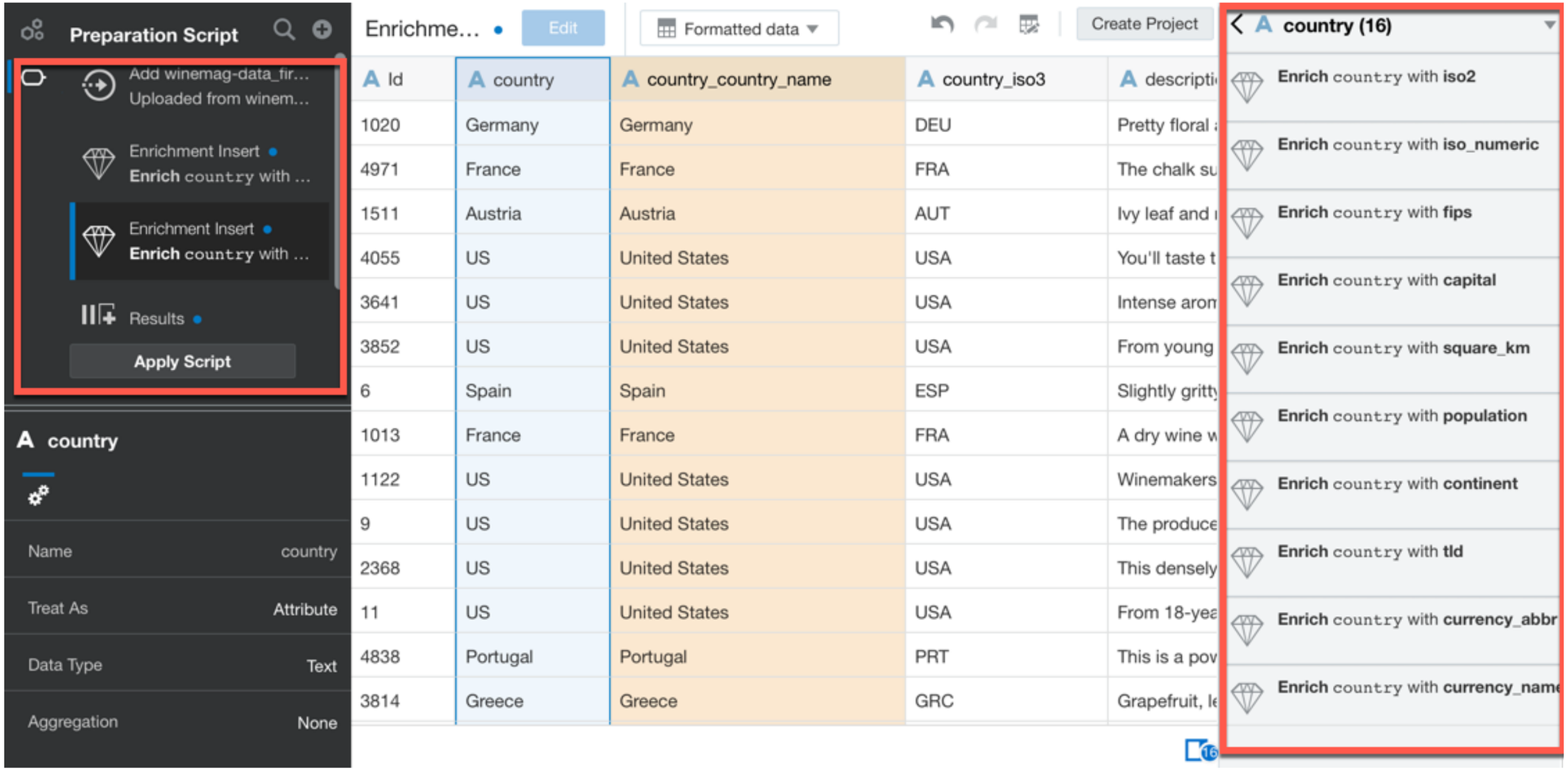
Name -> Gender

2 Locations -> Distance

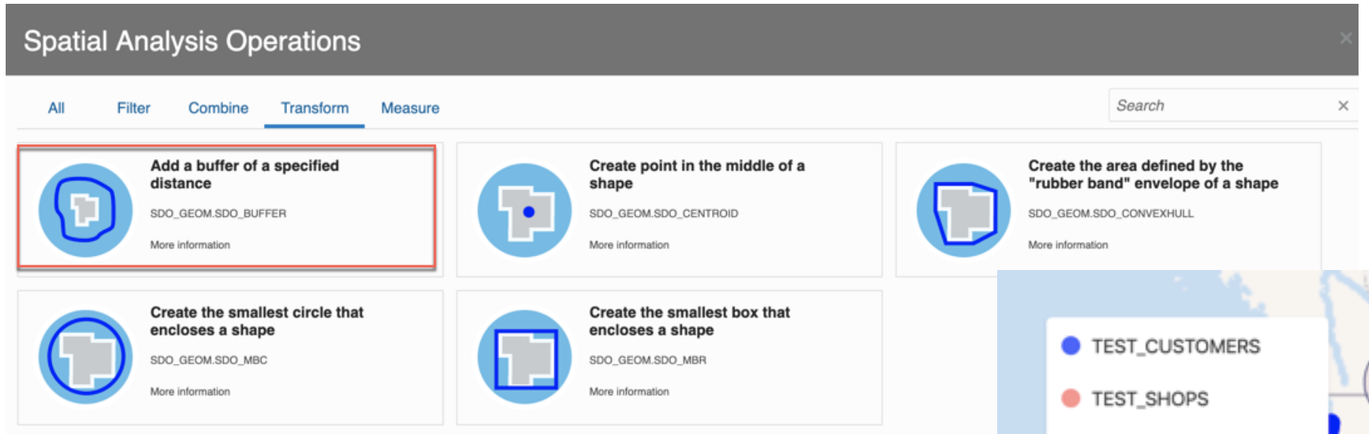
DataFlow

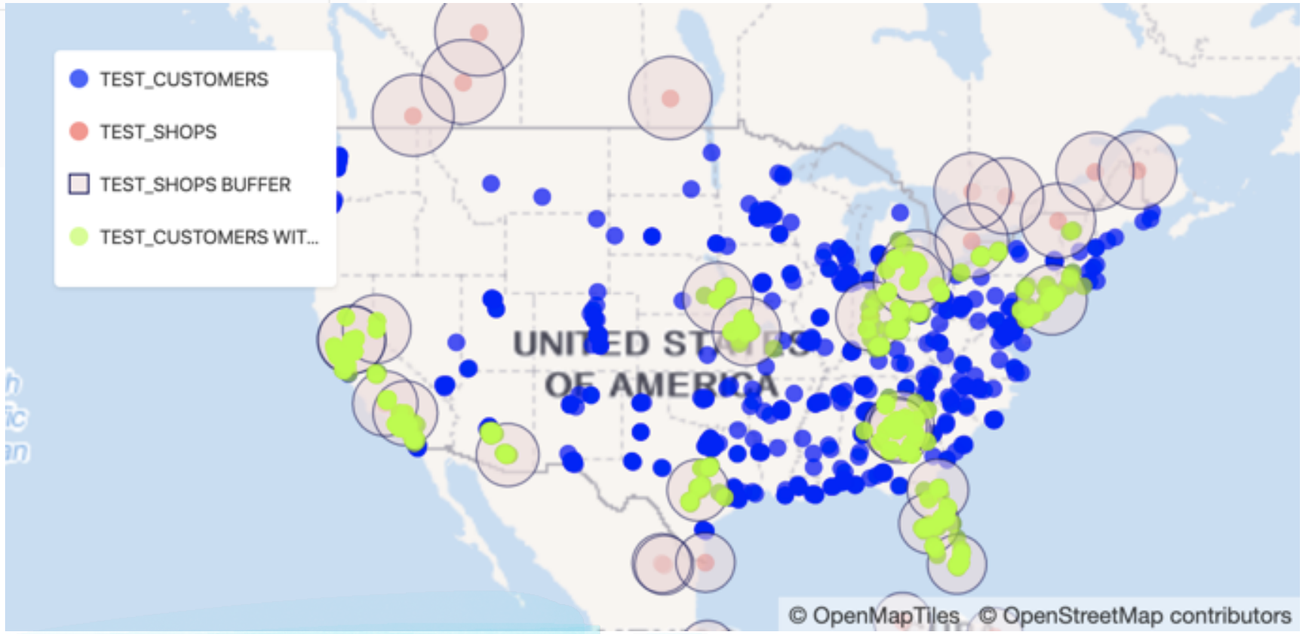
Day/Month/Year -> Date

Data Prep Recommendations



Spatial Enrichment





Oracle Spatial Studio

http://ritt.md/spatial-studio

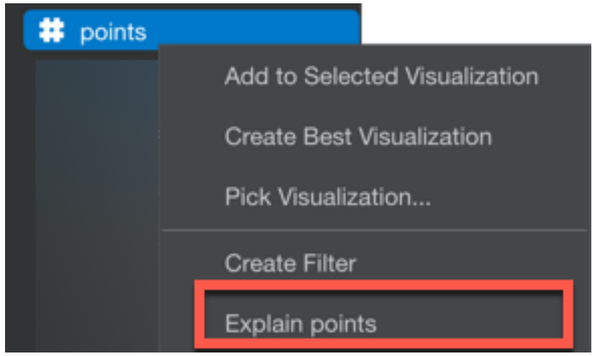
Data Overview

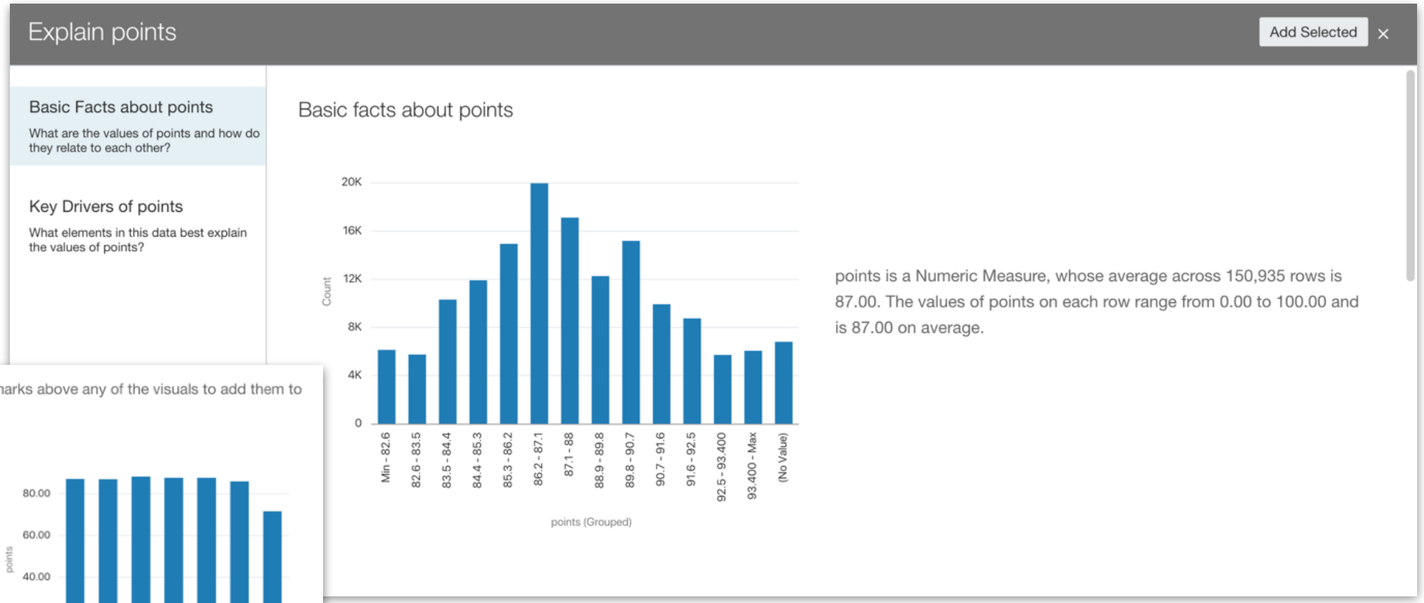
Results

Metadata

Data Element	Data Type	Treat As	Aggregation	Sample Values
Id	varchar(80)	Attribute	none	1470; 817; 1028; 632; 3689; 4148; 2576; 963; 4979; 281
country	varchar(137)	Attribute	none	US; France; Italy; Spain; Portugal; Germany; Argentina; Chile; Austria; Greece
country_continent	varchar(4000)	Attribute	none	NA; EU
country_fips	varchar(4000)	Attribute	none	US; IT; FR
country_iso3	varchar(4000)	Attribute	none	USA; FRA; ITA
country_iso_numeric	number	Measure	sum	840; 380; 250
country_iso2	varchar(4000)	Attribute	none	US; IT; FR
description	varchar(1247)	Attribute	none	This elegant wine combines subtle nutmeg and cardamom aromas with crisp app…
designation	varchar(122)	Attribute	none	Reserve; Estate; Reserva; Riserva; Estate Bottled; Vieilles Vignes; Crianza; Classic…
points	number	Measure	sum	90; 89; 88; 87; 91; 86; 92; 93; 85; 94
price	varchar(15)	Attribute	none	25; 20; 40; 18; 60; 30; 28; 35; 50; 15
province	varchar(53)	Attribute	none	California; Oregon; Bordeaux; Tuscany; Piedmont; Washington; Northern Spain; M…
region_1	varchar(75)	Attribute	none	Willamette Valley; Napa Valley; Barolo; Brunello di Montalcino; Russian River Valle…
region_2	varchar(35)	Attribute	none	Central Coast; Sonoma; Willamette Valley; Napa; Columbia Valley; Mendocino/La…
variety	varchar(53)	Attribute	none	Pinot Noir; Chardonnay; Bordeaux-style Red Blend; Cabernet Sauvignon; Red Ble…
winery	varchar(84)	Attribute	none	Tarara; Heron Hill; Byron; Bergstr√∂m; Herdade do Rocim; Rusack; Sarah's Viney…

Explain


points
Add to Selected Visualization
Create Best Visualization
Pick Visualization...
Create Filter
Explain points


Explain points
Add Selected
Basic Facts about points
What are the values of points and how do they relate to each other?
Key Drivers of points
What elements in this data best explain the values of points?
Basic facts about points
Count
20K
16K
12K
8K
4K
0
Min - 82.6
82.6 - 83.5
83.5 - 84.4
84.4 - 85.3
85.3 - 86.2
86.2 - 87.1
87.1 - 88
88.9 - 89.8
89.8 - 90.7
90.7 - 91.6
91.6 - 92.5
92.5 - 93.400
93.400 - Max
(No Value)
points (Grouped)
points is a Numeric Measure, whose average across 150,935 rows is 87.00. The values of points on each row range from 0.00 to 100.00 and is 87.00 on average.

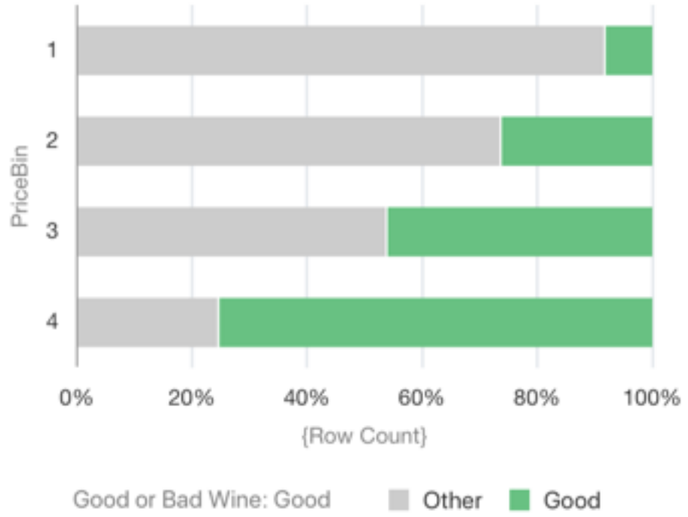

The charts below summarize the values of points by the measures in this data set. Click the checkmarks above any of the visuals to add them to your project when done.
points
id
country
country_continent
AF
AS
EU
NA
OC
SA

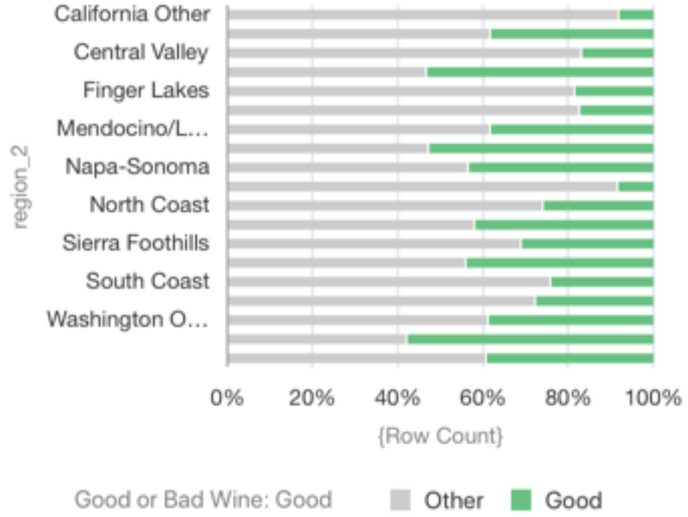
Explain - Key Drivers

Key Drivers of Good or Bad Wine

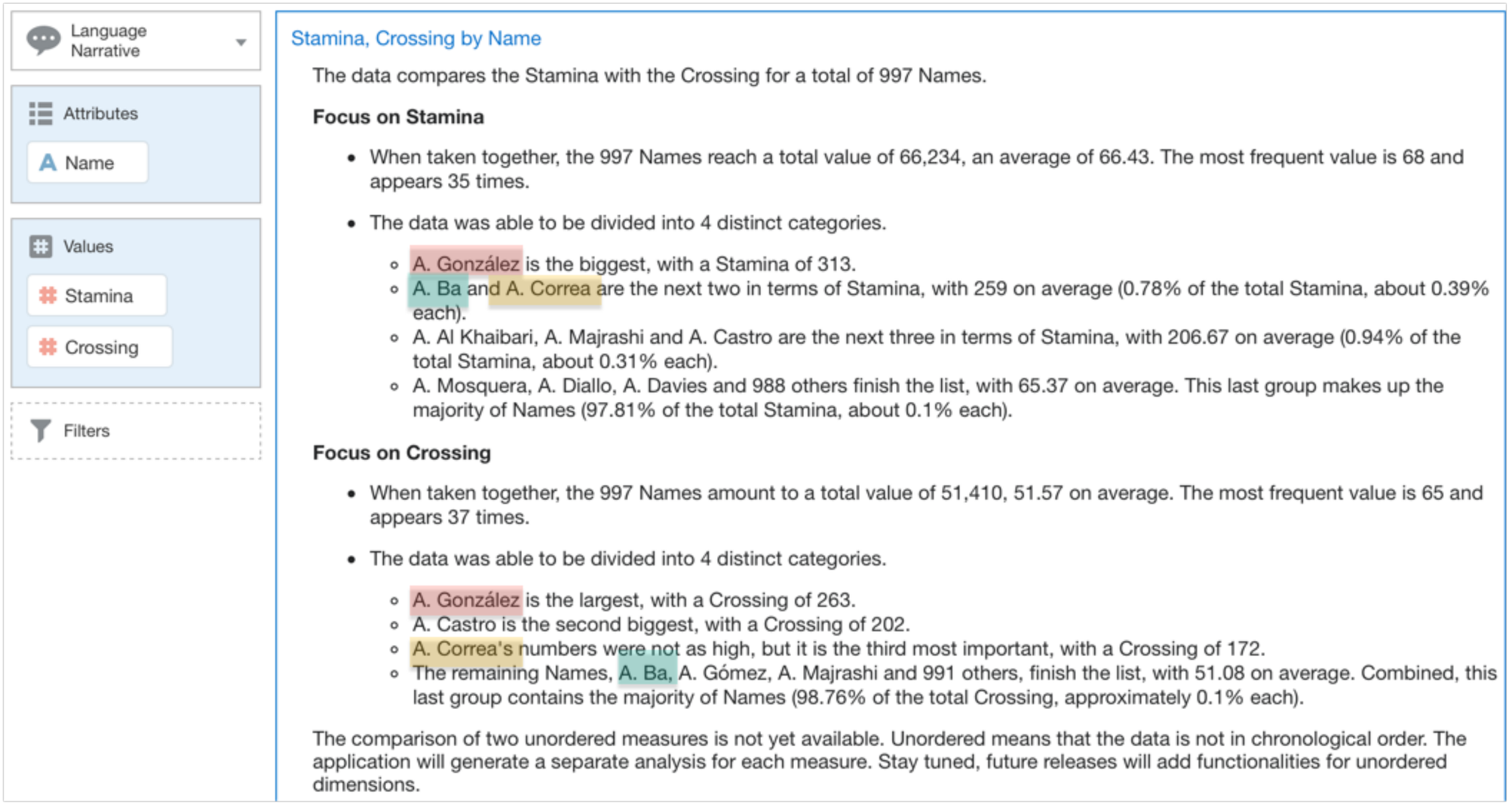
Based on Good or Bad Wine: Good the 2 attributes that are most strongly correlated are **PriceBin, region_2**

The charts below show the distribution of Good or Bad Wine values across each of the key drivers. Click the checkmarks above any of the visuals to add them to your project when done.

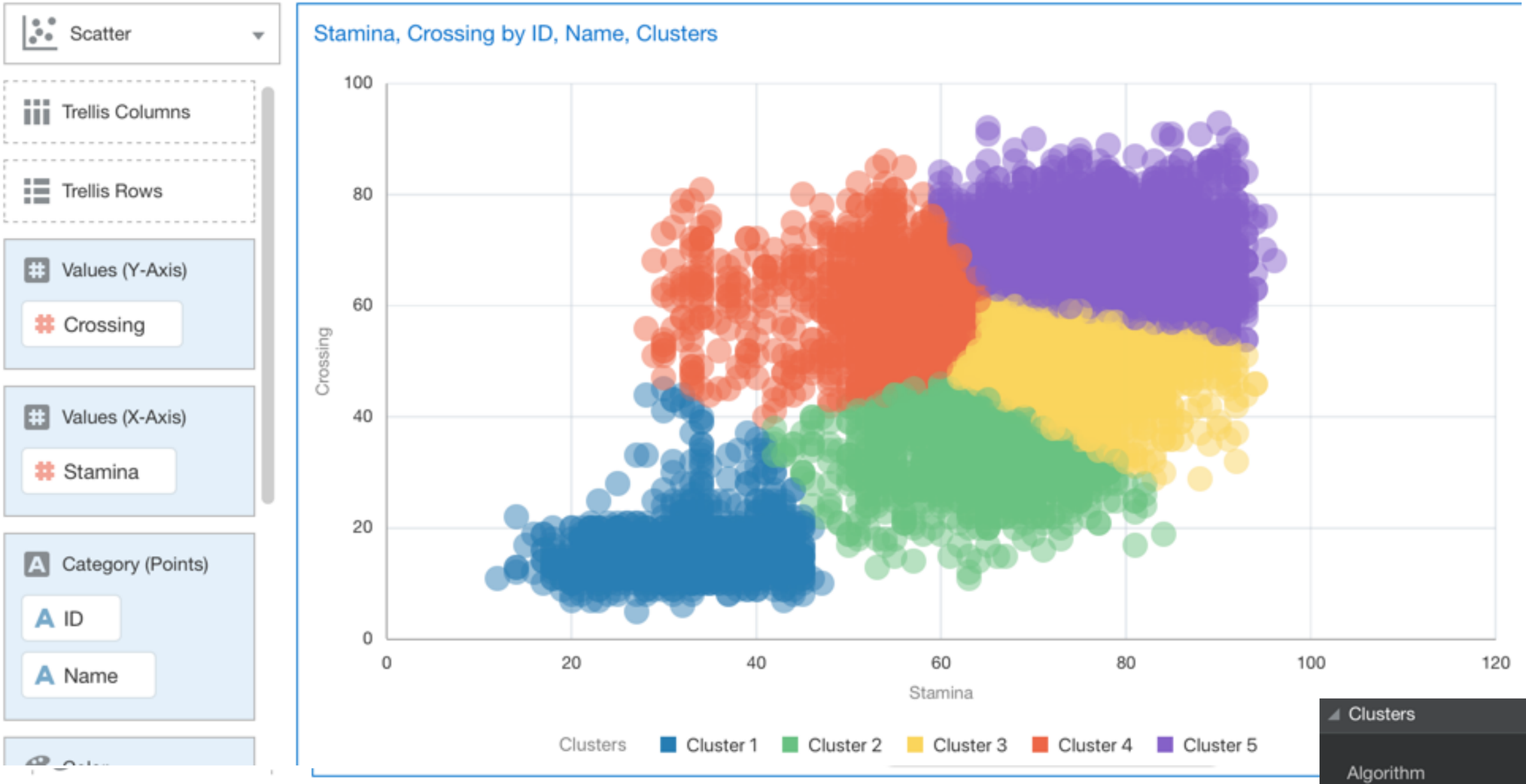




Natural Language Generation



Easy Models


Scatter
Trellis Columns
Trellis Rows
Values (Y-Axis)
Crossing
Values (X-Axis)
Stamina
Category (Points)
ID
Name
Stamina, Crossing by ID, Name, Clusters
100
80
60
40
20
0
Crossing
0
20
40
60
80
100
120
Stamina
Clusters
Cluster 1
Cluster 2
Cluster 3
Cluster 4
Cluster 5
Clusters
Algorithm
K-Means
Groups
5

DataFlow Train Model

Compare - Classification

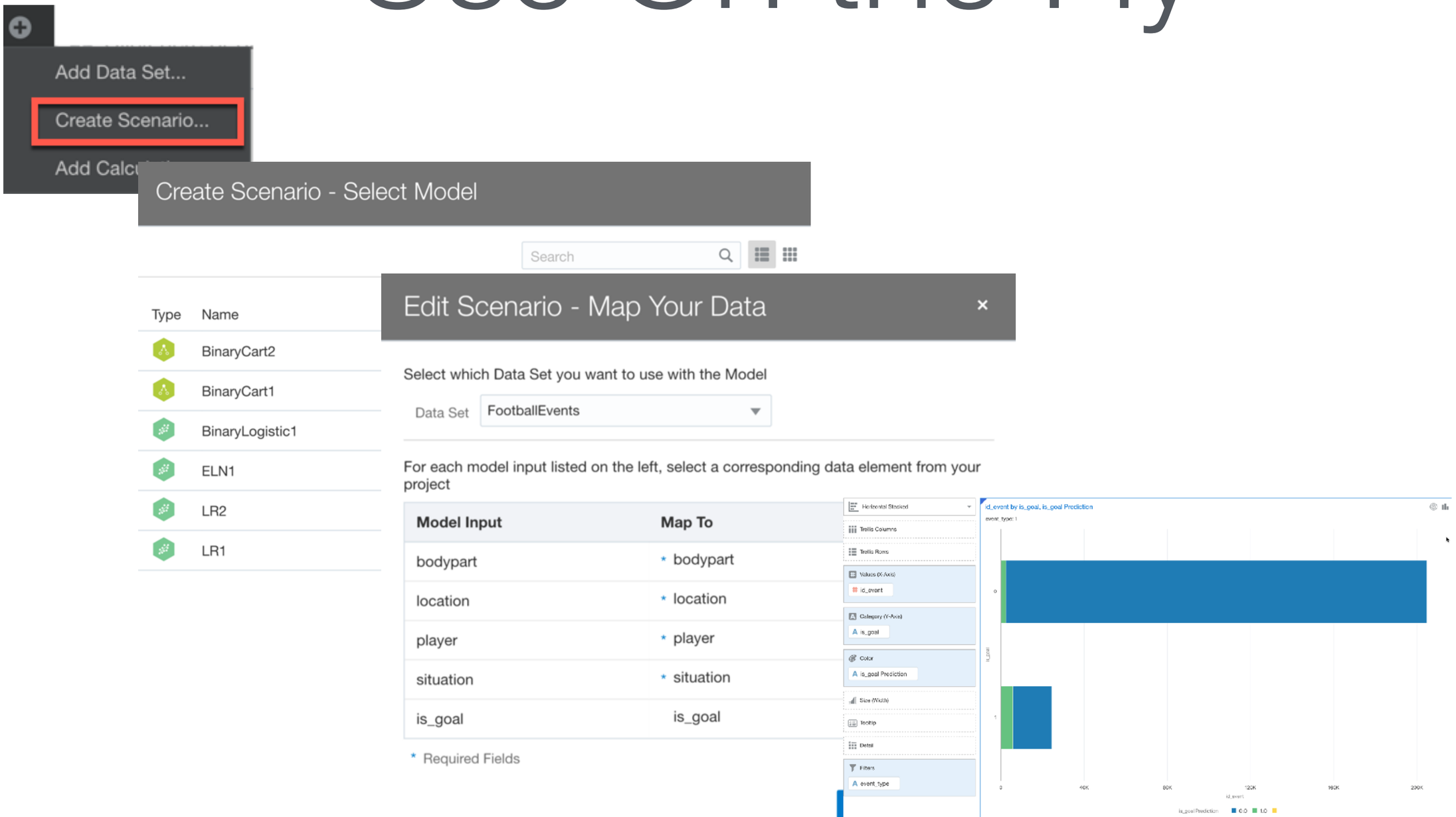
Predicted Values

Actual Values	0.0	1.0	Total
0.0	40439	471	40910 (90%)
1.0	3761	866	4627 (10%)
Total	44200 (97%)	1337 (3%)	45537 (100%)

Predicted Value

Real Value	Good	Bad
Good		
Bad		

Use On the Fly



Congratulations!

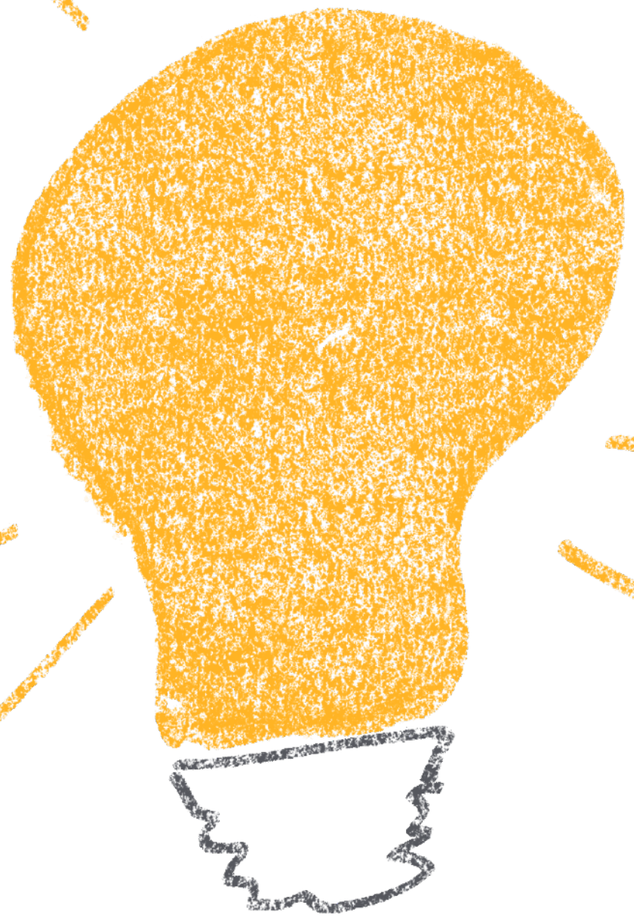
…You are now a Data Scientist!

Nearly There

...But

Data Cleaning

Feature Engineering

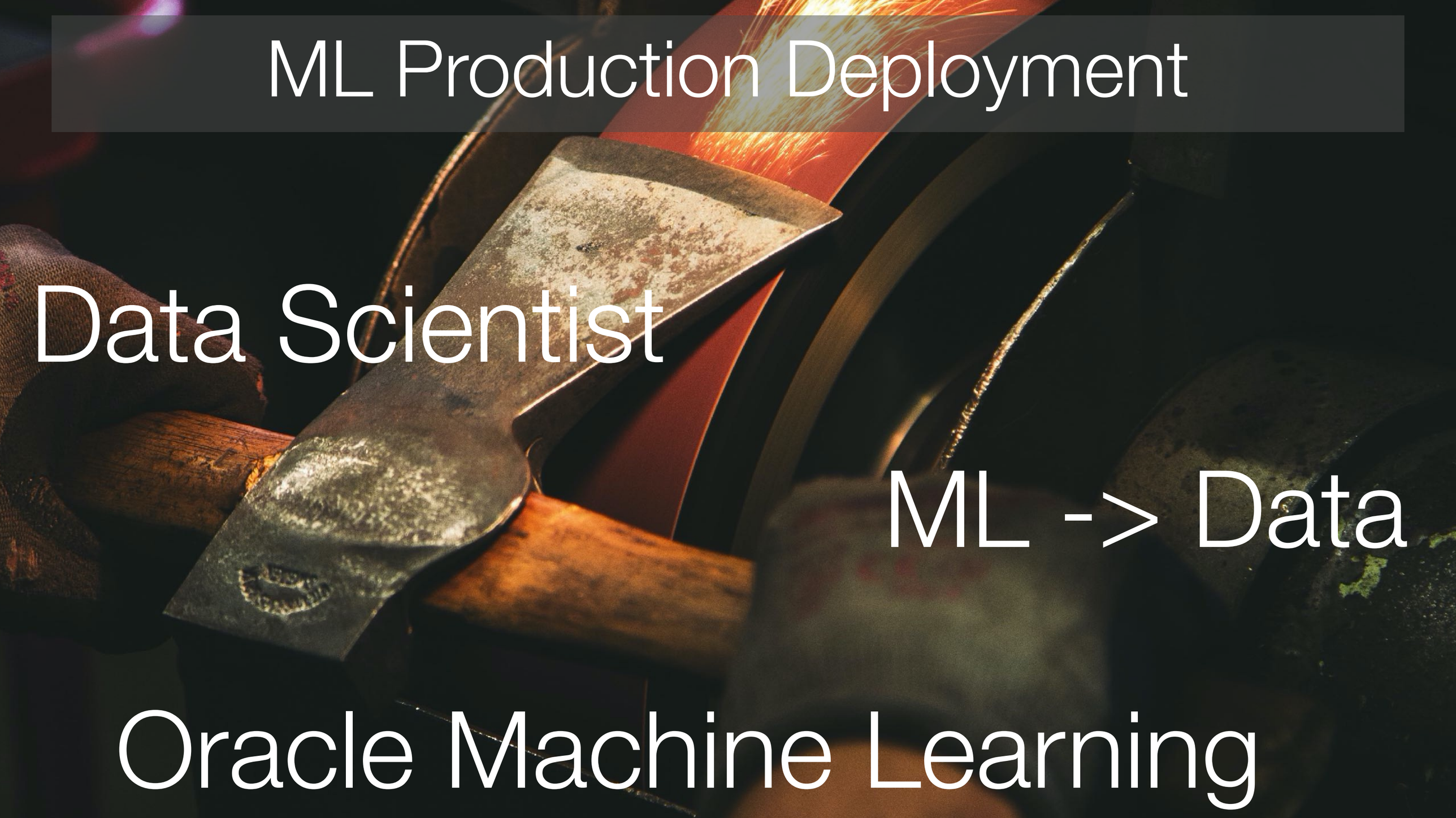


Model Creation & Evaluation

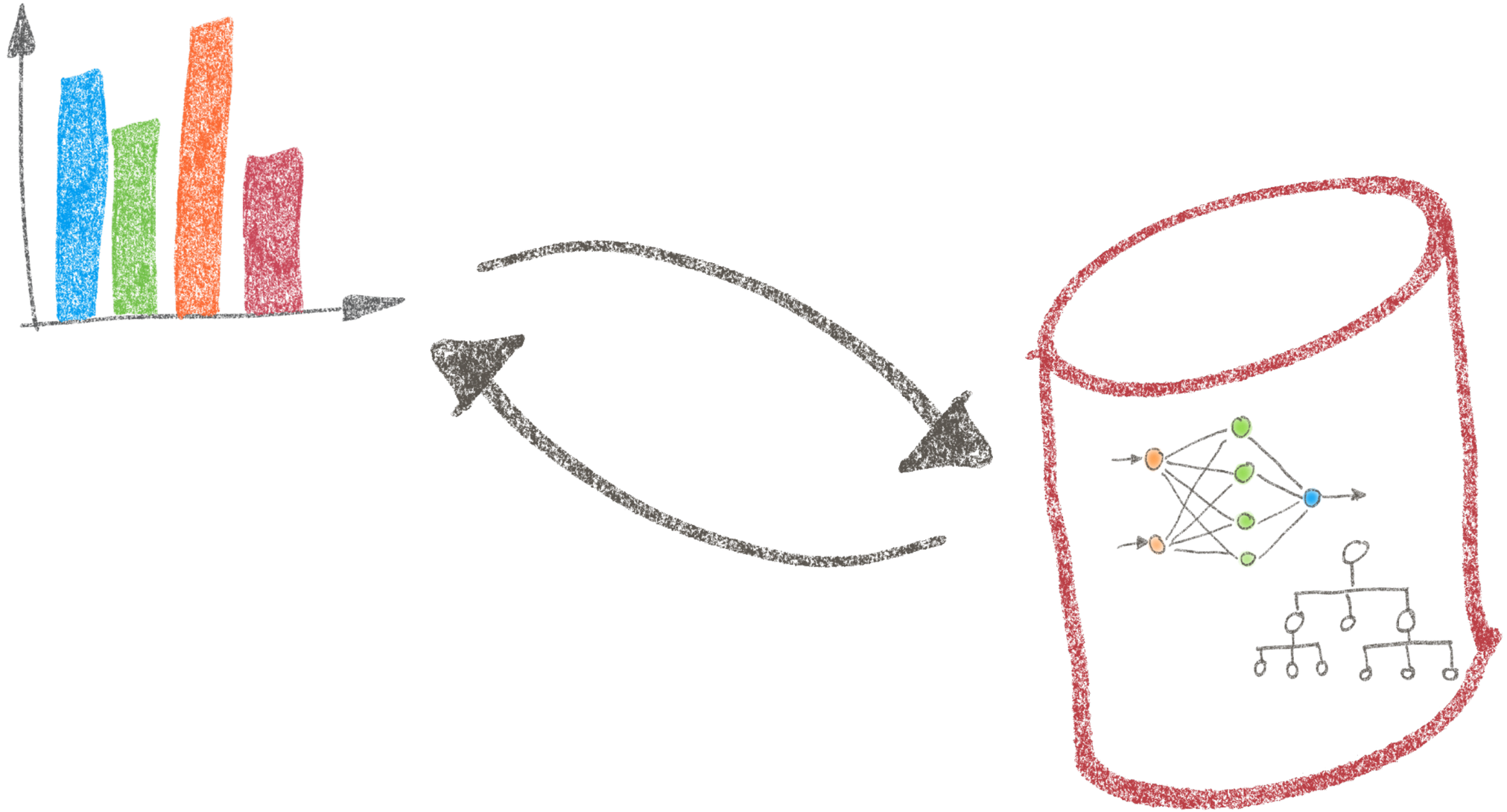
Feature Selection

70% > 50%

ML Production Deployment



Access Oracle Machine Learning Models on OAC



Become a Data Scientist with OAC

http://ritt.md/OAC-datascience

ML in Action with OAC

http://ritt.md/OAC-ML-Video

Insights Lab

https://www.rittmanmead.com/insight-lab/

How to Become a Data Scientist

Francesco Tisiot
Analytics Tech Lead

rittmanmead
A DATA AND ANALYTICS COMPANY