



How to Become a Data Scientist

Francesco Tisiot
Analytics Tech Lead

rittmanmead 
A DATA AND ANALYTICS COMPANY

A close-up photograph of a white bowl filled with spaghetti coated in a rich, dark red meat sauce. The bowl sits on a dark, textured surface, possibly a denim jacket, which is visible in the foreground and background.

Francesco Tisiot

Analytics Tech Lead

Verona, Italy



<http://ritt.md/ftisiot>



ft@rittmanmead.com



@FTisiot



Oracle ACE Director



rittmanmead

A DATA AND ANALYTICS COMPANY



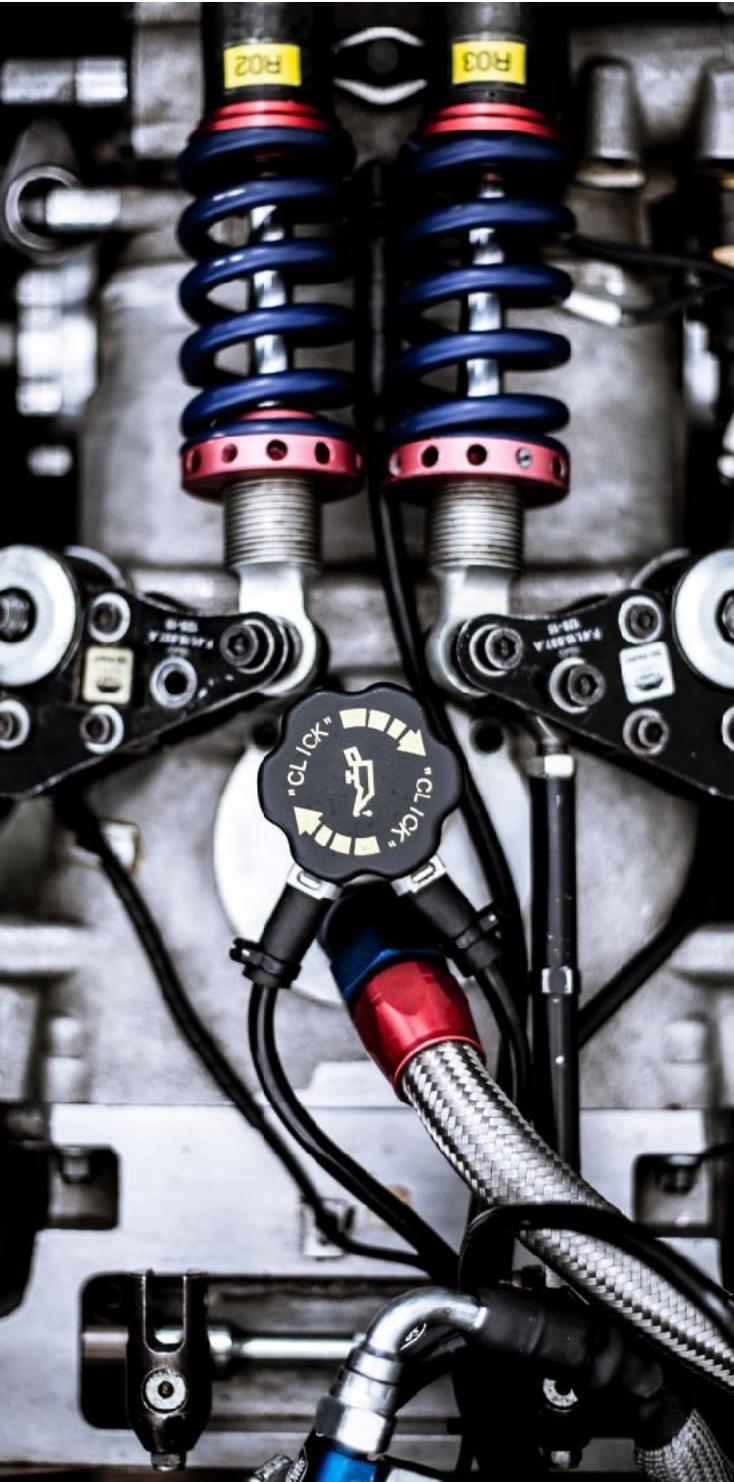
info@rittmanmead.com



www.rittmanmead.com



@rittmanmead



Data Engineering



Analytics



Data Science

Data Scientist





R

Python

Linear Algebra

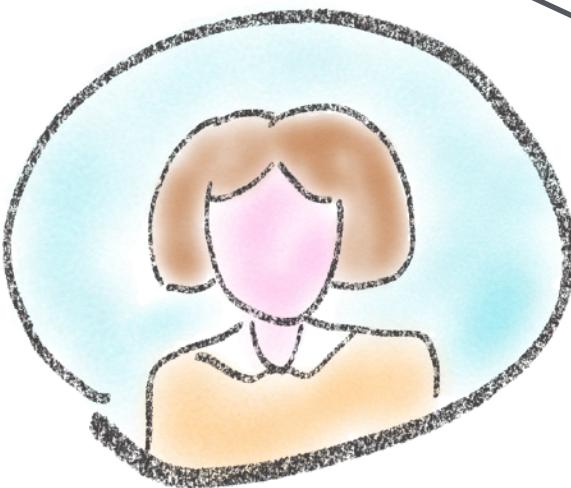
Statistics

Low Hanging Fruit Theory

A close-up photograph of a hand holding a bright orange fruit, likely a tangerine or orange, against a dark, out-of-focus background. The hand is positioned palm-up, with the fruit resting in the center. Large, vibrant green leaves are visible behind the hand, some partially obscuring the fruit. The lighting highlights the texture of the fruit's skin and the veins on the leaves.

Democratise
Data Science

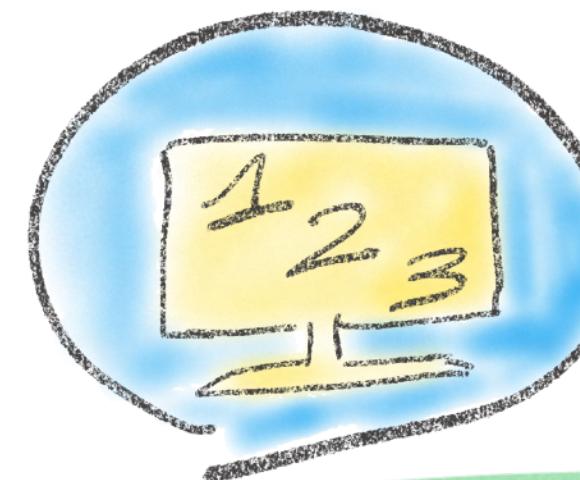
Basic Operations



What are the
Drivers for
My Sales?



Based on my
Experience
I can Guess....



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

Augmented Analytics

Basic Operations



Is he going to
accept the
Offer?

YES/NO

50%

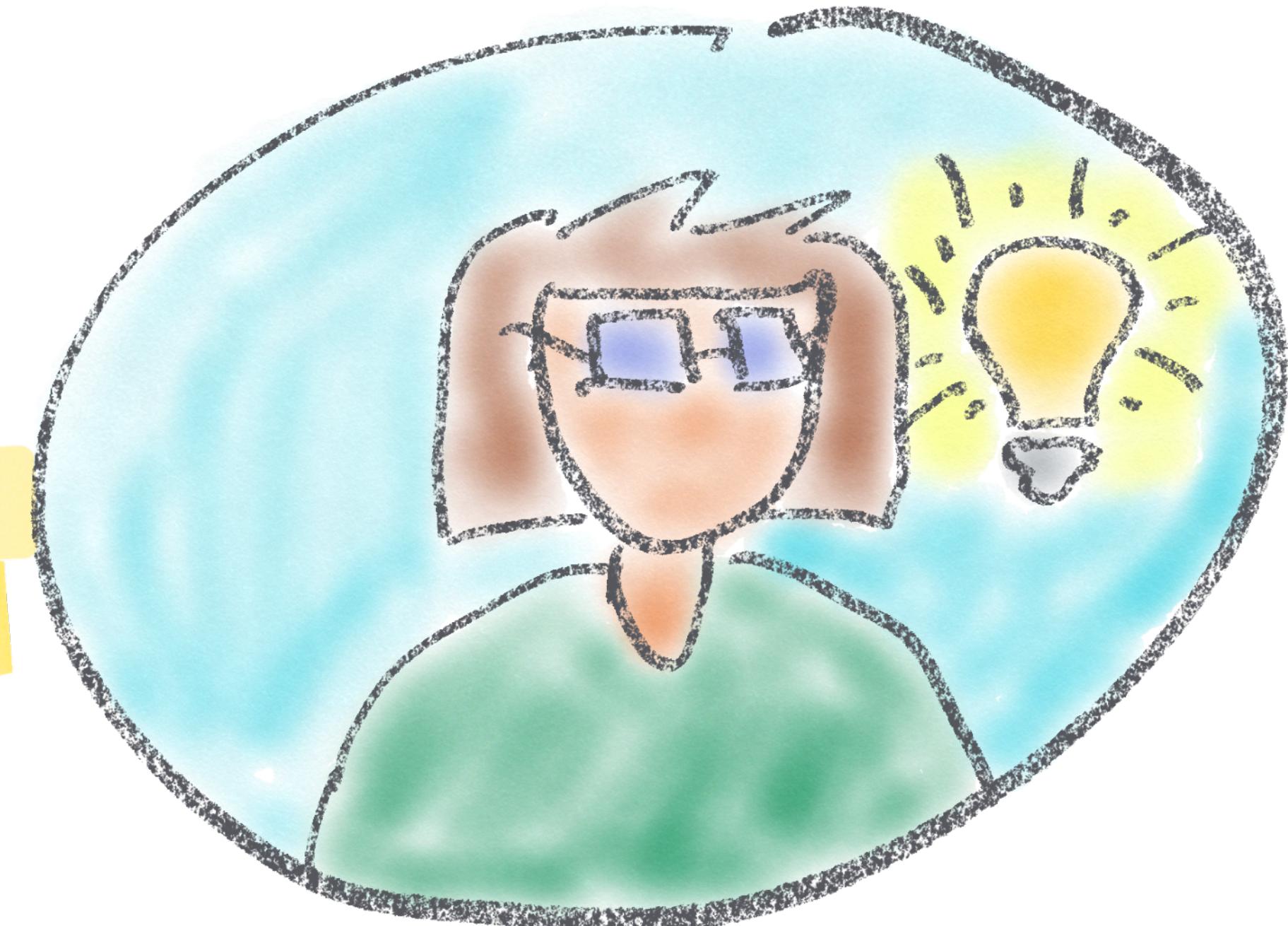
Basic ML
Model

70%

Oracle Analytics Cloud



How Do I Become Data Scientists with OAC?



$$c(\omega) = \frac{1}{\pi} \int_{-\infty}^{\infty} f(t) \cdot \sin(\omega t) dt$$

$f(t)$

$c(\omega)$



$$f(t) = \int_0^{\infty} a(\omega) \cdot \cos(\omega t) + b(\omega) \cdot \sin(\omega t)$$

$$a_0 = \frac{1}{T} \int_0^T f(t) dt$$

$$a_n = \frac{1}{T} \int_0^T f(t) \cdot \cos\left(\frac{n\pi t}{T}\right) dt$$

$$b_n = \frac{1}{T} \int_0^T f(t) \cdot \sin\left(\frac{n\pi t}{T}\right) dt$$

$$f(t) = a_0 + \sum_{n=1}^{\infty} \left(a_n \cdot \cos\left(\frac{n\pi t}{T}\right) + b_n \cdot \sin\left(\frac{n\pi t}{T}\right) \right)$$

$$c_n = \frac{1}{2\pi} \int_{-\infty}^{\infty} f(t) e^{-j\frac{n\pi t}{T}} dt$$

$$b(\omega) = \frac{1}{T} \int_0^{\infty} f(t) \cdot \sin(\omega t) dt$$

$$f(t) = \sum_{n=-\infty}^{\infty} c_n \cdot e^{j\frac{n\pi t}{T}}$$

$$c(\omega) = \int_{-\infty}^{\infty} f(t) \cdot e^{-j\frac{n\pi t}{T}} dt$$

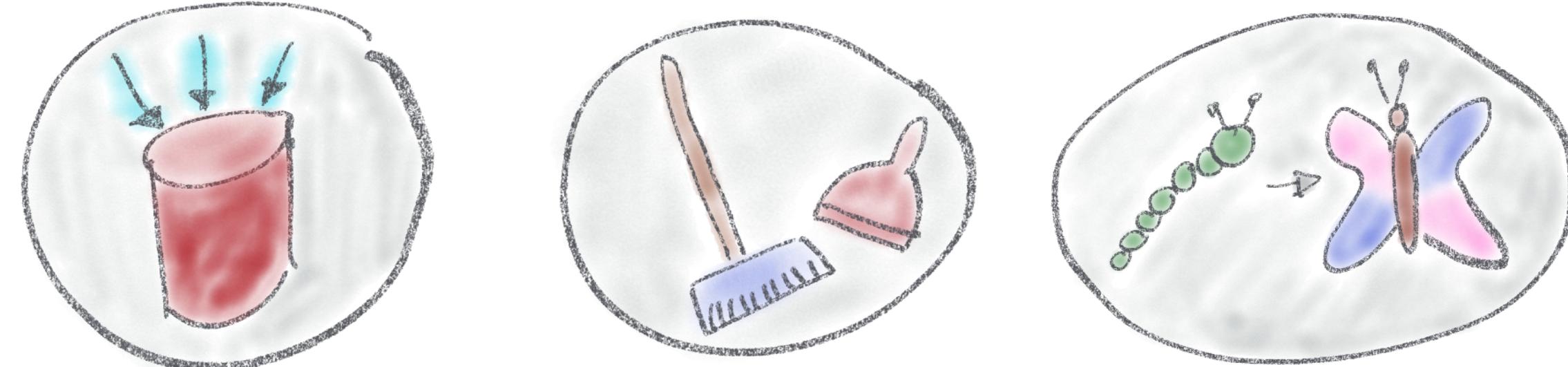
$$f(t) = \frac{1}{2\pi} \int_{-\infty}^{\infty} c(\omega) \cdot e^{j\omega t} d\omega$$



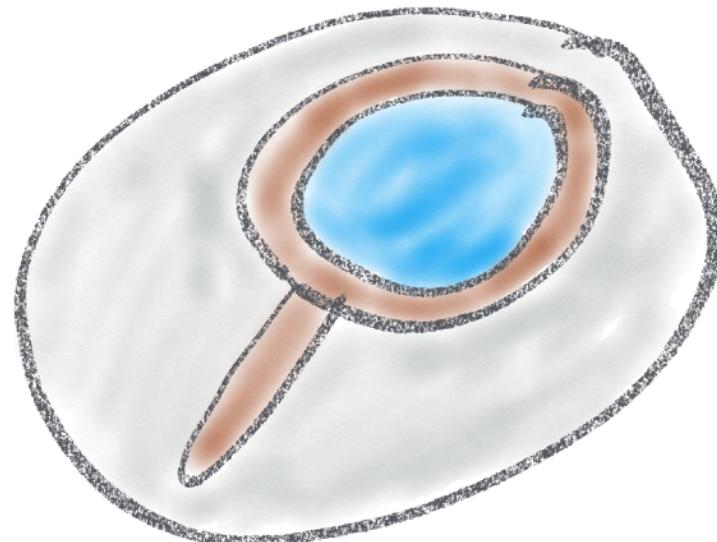
Define the Problem

6 Steps into Data Science

6



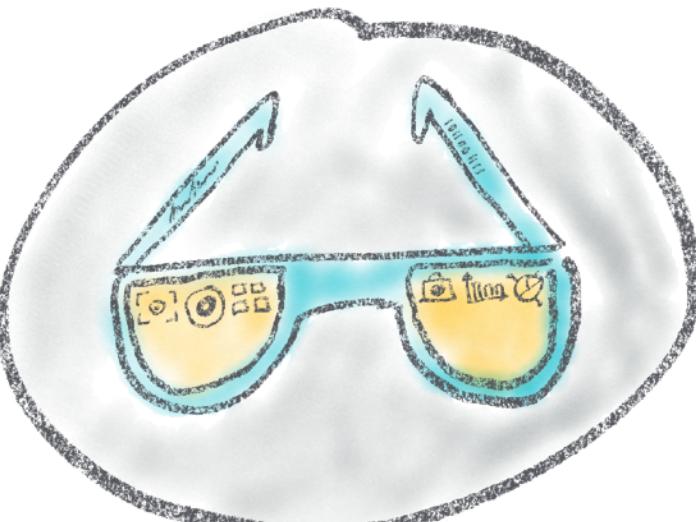
1. Connect



4. Analyse



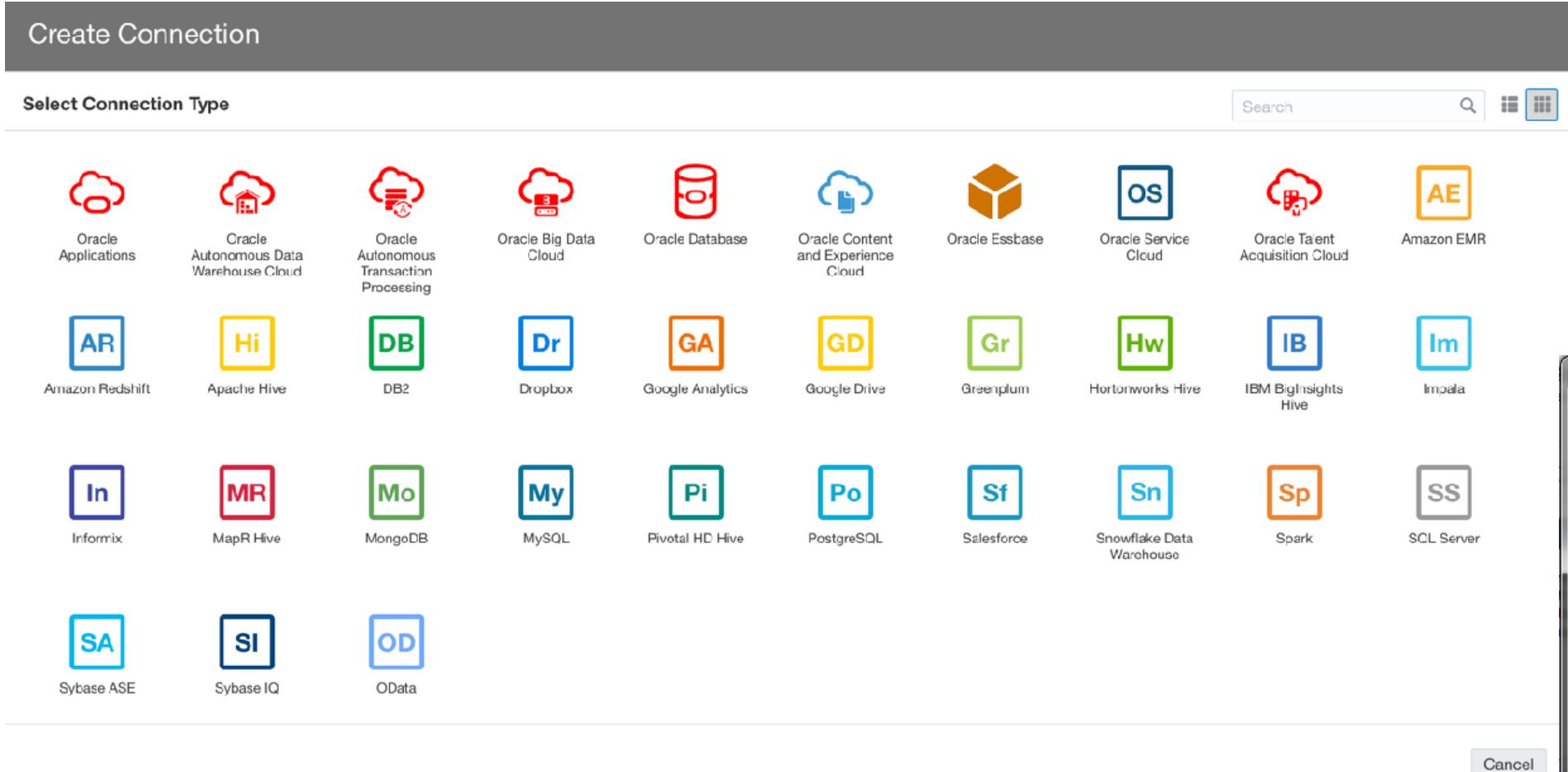
2. Clean
3. Transform
& Enrich



5. Transform
& Enrich

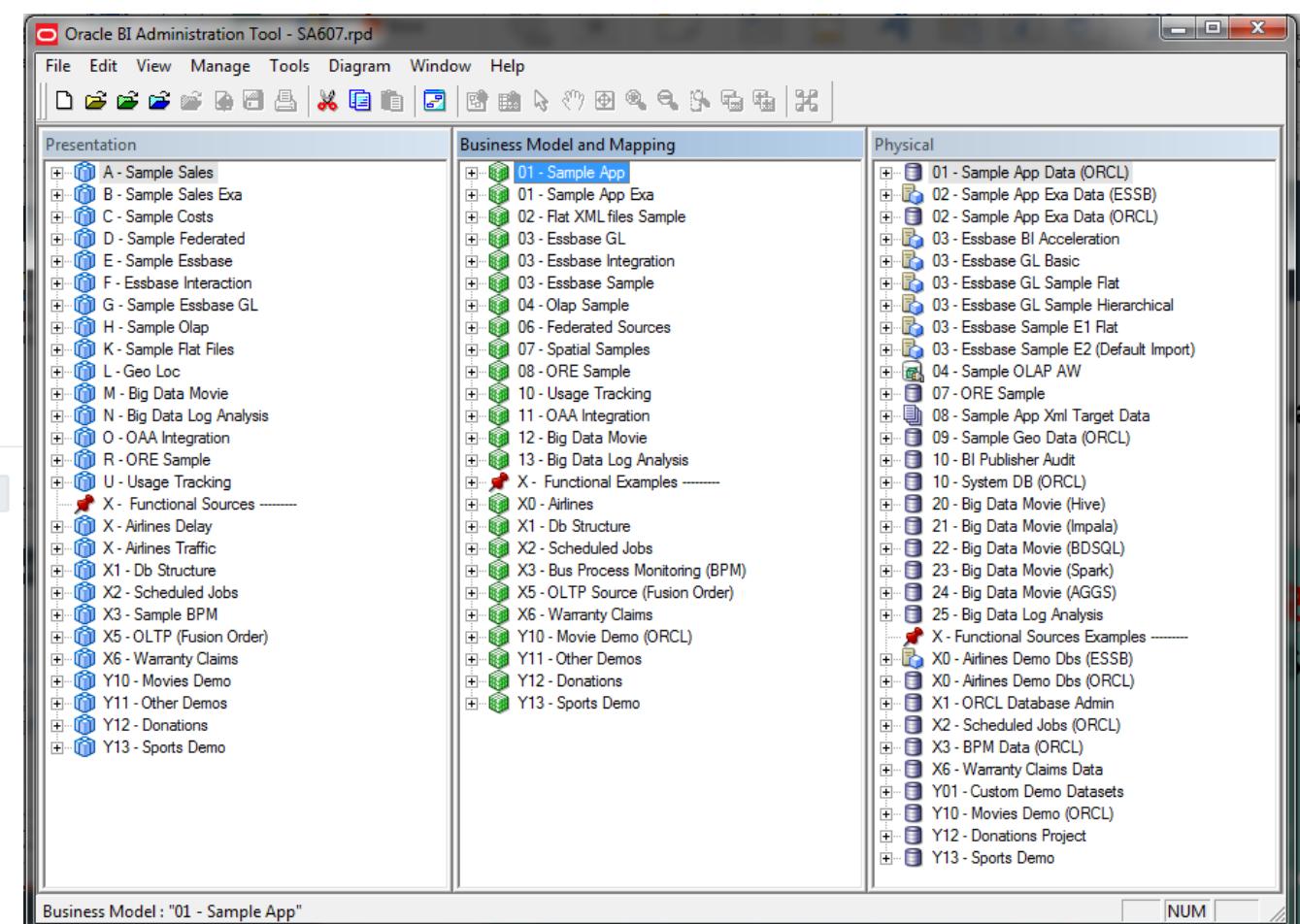
6. Predict

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

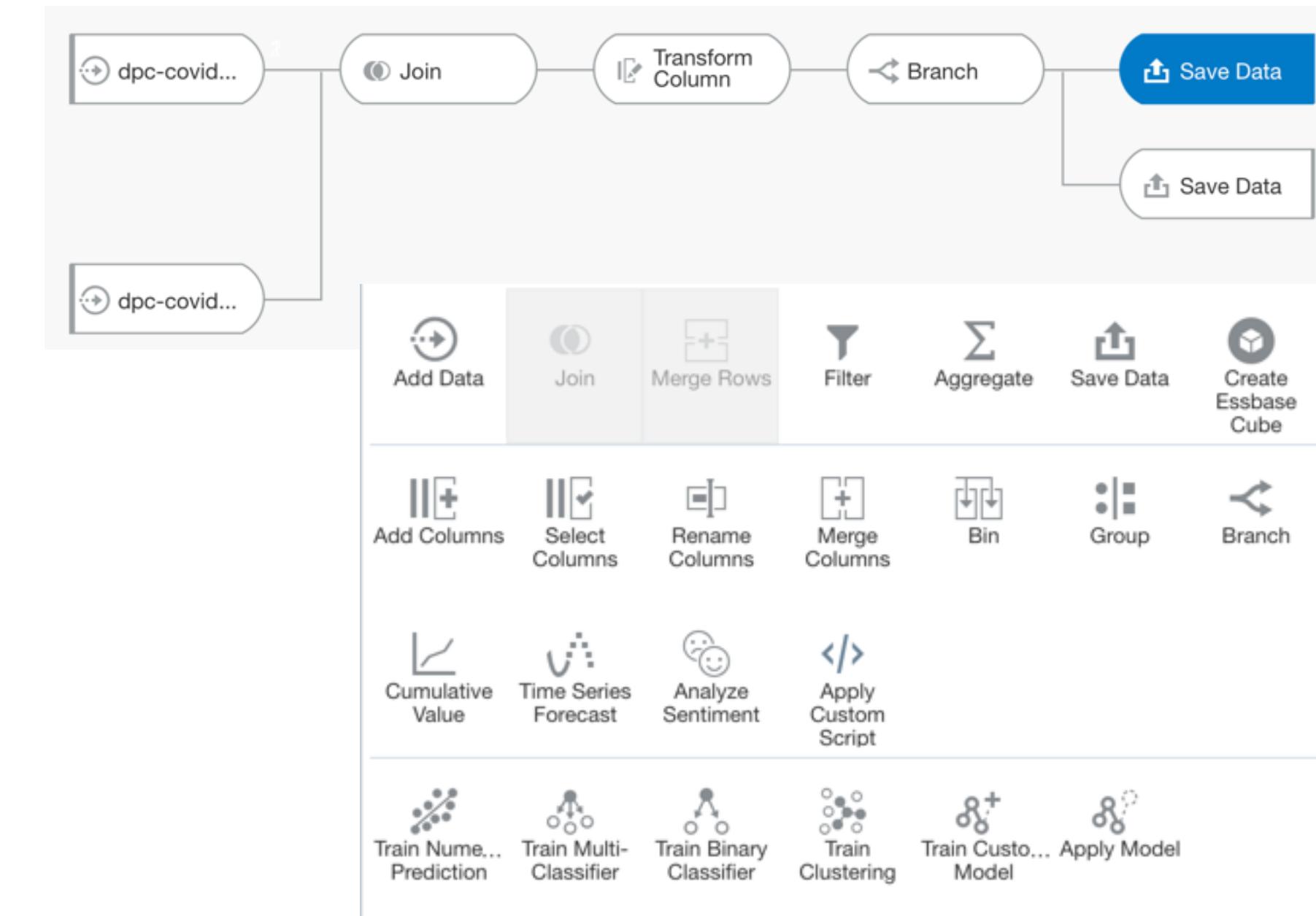
Train: 80%
Test: 20%

Train/Test Set Split

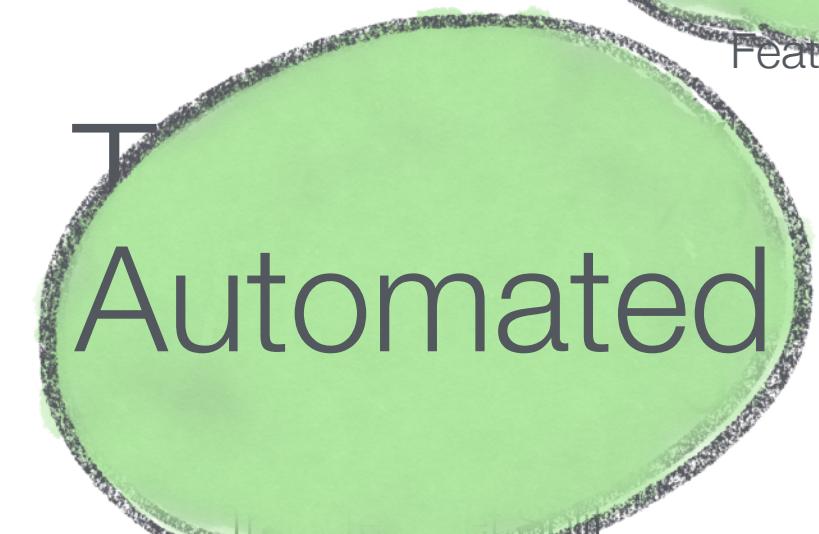
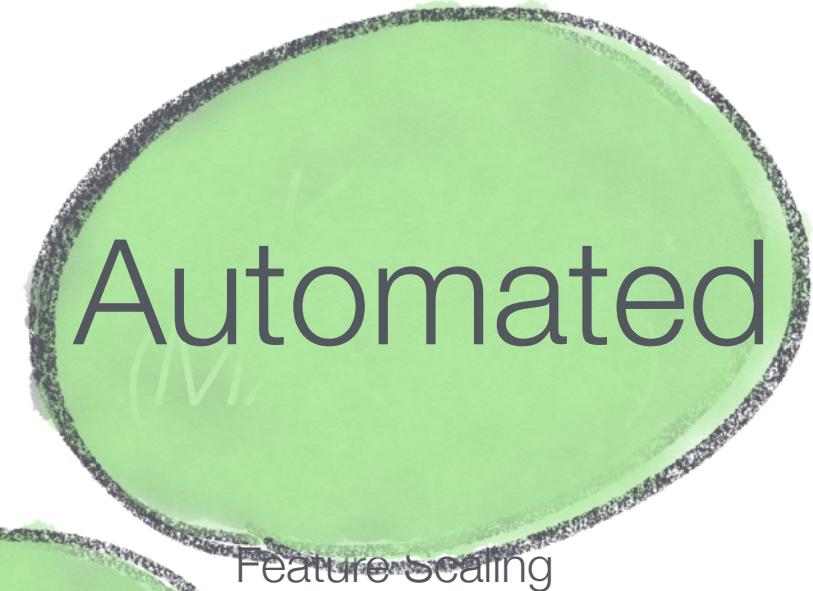
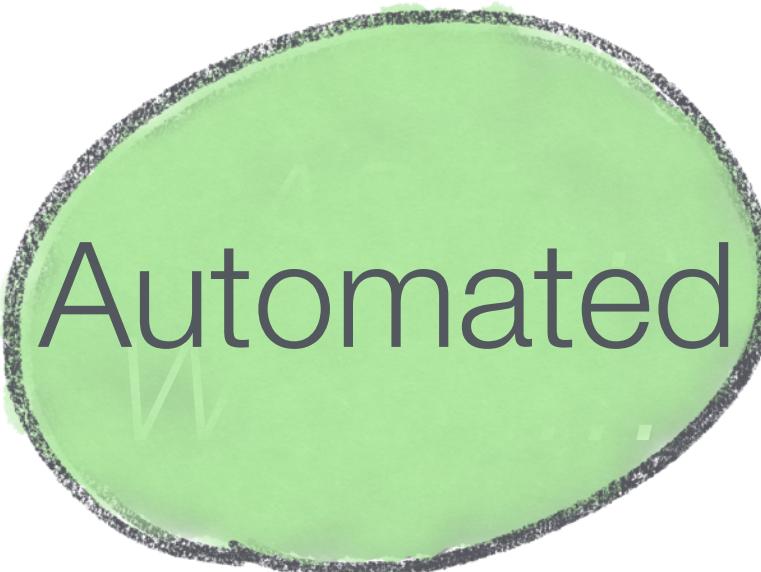
Clean... How?

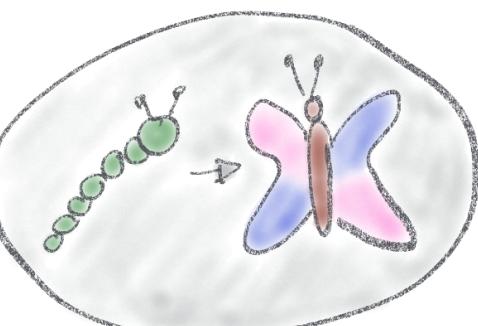
Data Flows

- Filter
- Aggregate
- Join
- Transform



Clean





Feature Engineering

Location -> ZIP Code
Additional Data Source?
Name -> Gender

2 Locations -> Distance
DataFlow
Day/Month/Year -> Date

Data Prep Recommendations

Preparation Script Enrichme... • Edit Formatted data ▾ Create Project

Add winemag-data_fir... Uploaded from winemag-data_fir...

Enrichment Insert • Enrich country with iso2

Enrichment Insert • Enrich country with iso_numeric

Results •

Apply Script

A country

A Id	A country	A country_country_name	A country_iso3	A description
1020	Germany	Germany	DEU	Pretty floral ...
4971	France	France	FRA	The chalk su...
1511	Austria	Austria	AUT	Ivy leaf and i...
4055	US	United States	USA	You'll taste t...
3641	US	United States	USA	Intense arom...
3852	US	United States	USA	From young ...
6	Spain	Spain	ESP	Slightly gritty...
1013	France	France	FRA	A dry wine w...
1122	US	United States	USA	Winemakers ...
9	US	United States	USA	The produce ...
2368	US	United States	USA	This densely ...
Treat As	Attribute	11	United States	From 18-year ...
Data Type	Text	4838	Portugal	This is a pov...
Aggregation	None	3814	Greece	Grapefruit, le...

< A country (16)

- Enrich country with iso2
- Enrich country with iso_numeric
- Enrich country with fips
- Enrich country with capital
- Enrich country with square_km
- Enrich country with population
- Enrich country with continent
- Enrich country with tld
- Enrich country with currency_abbr
- Enrich country with currency_name

16

Spatial Enrichment

Spatial Analysis Operations

All Filter Combine Transform Measure

Search X

Add a buffer of a specified distance
SDO_GEOM.SDO_BUFFER
More information

Create point in the middle of a shape
SDO_GEOM.SDO_CENTROID
More information

Create the area defined by the "rubber band" envelope of a shape
SDO_GEOM.SDO_CONVEXHULL
More information

Create the smallest circle that encloses a shape
SDO_GEOM.SDO_MBC
More information

Create the smallest box that encloses a shape
SDO_GEOM.SDO_MBR
More information

TEST_CUSTOMERS
TEST_SHOPS
TEST_SHOPS BUFFER
TEST_CUSTOMERS WITHIN BUFFER

© OpenMapTiles © OpenStreetMap contributors

Oracle Spatial Studio

<http://ritt.md/spatial-studio>

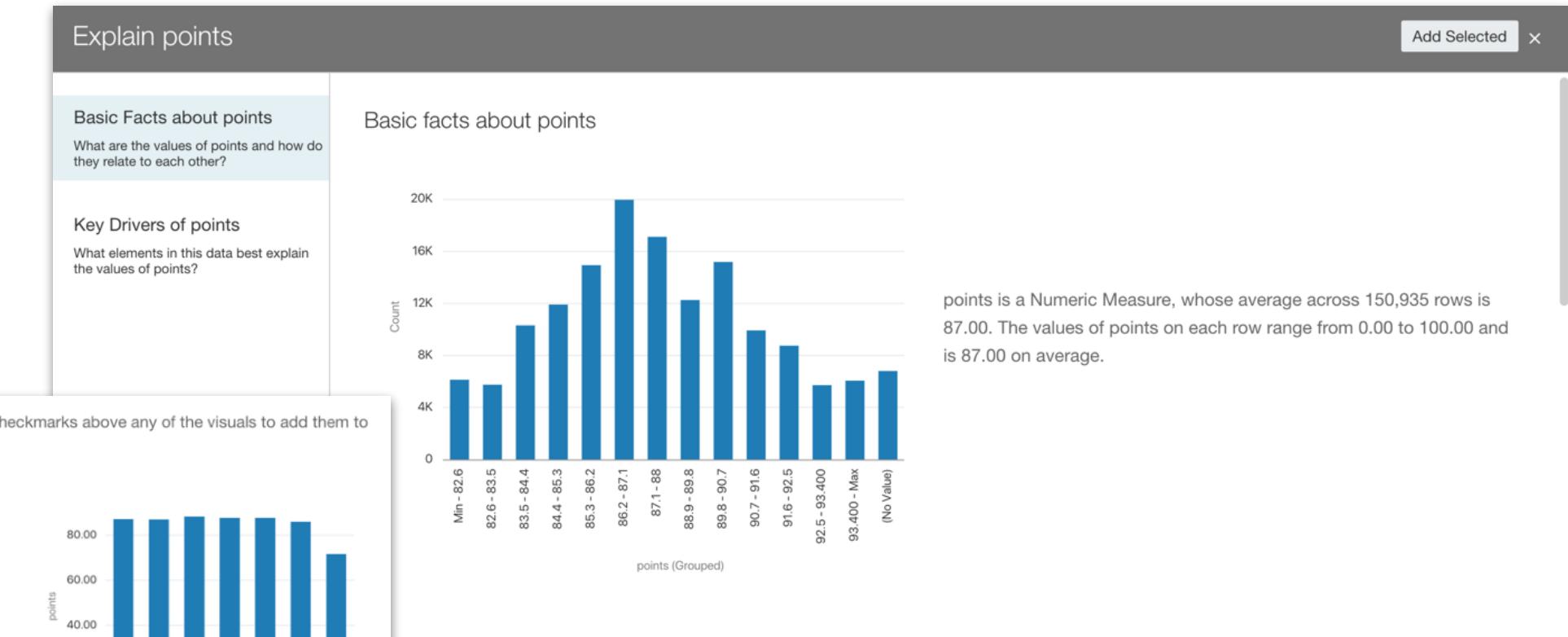
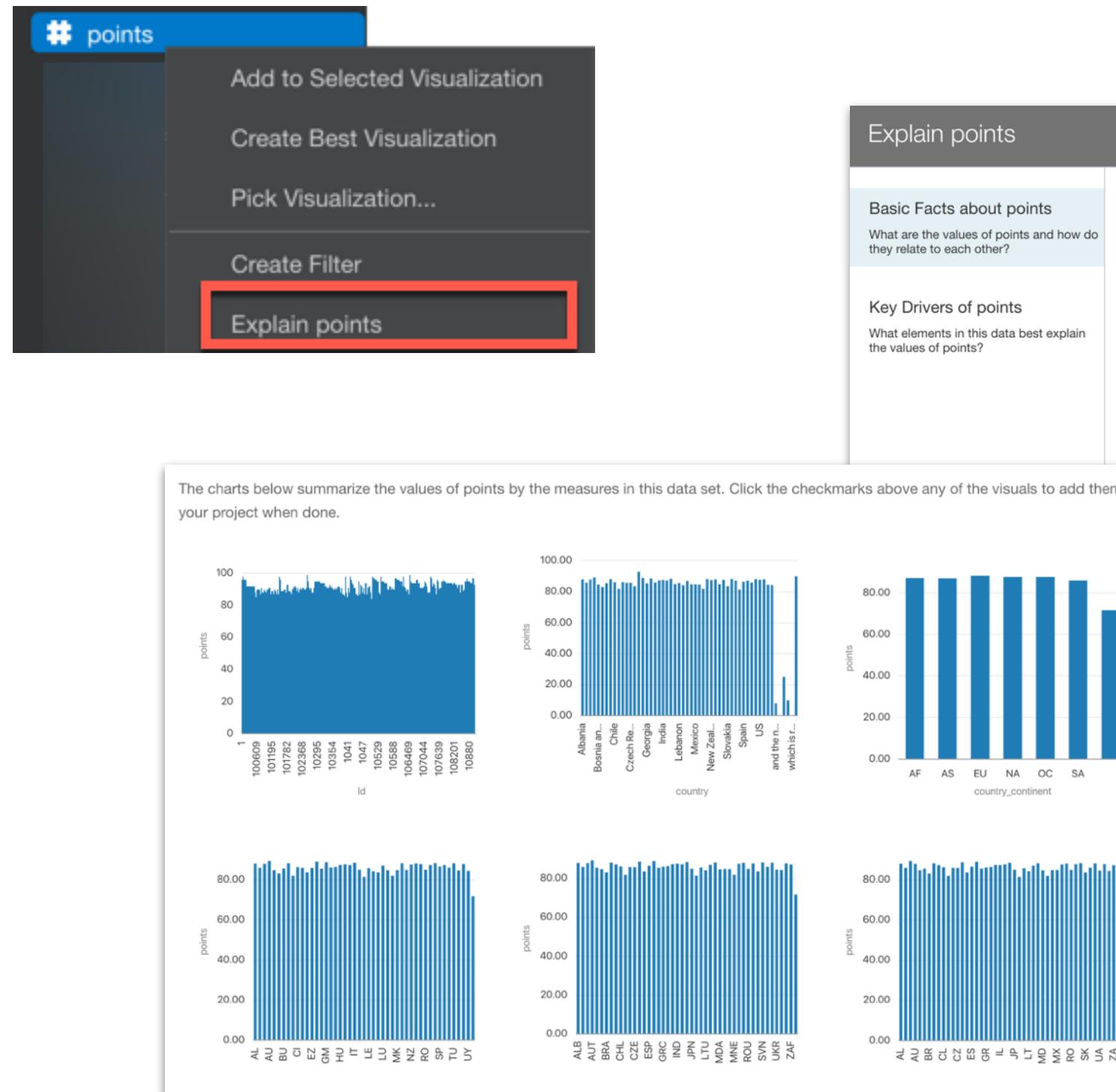
Data Overview

Results

Metadata ▾

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

Explain

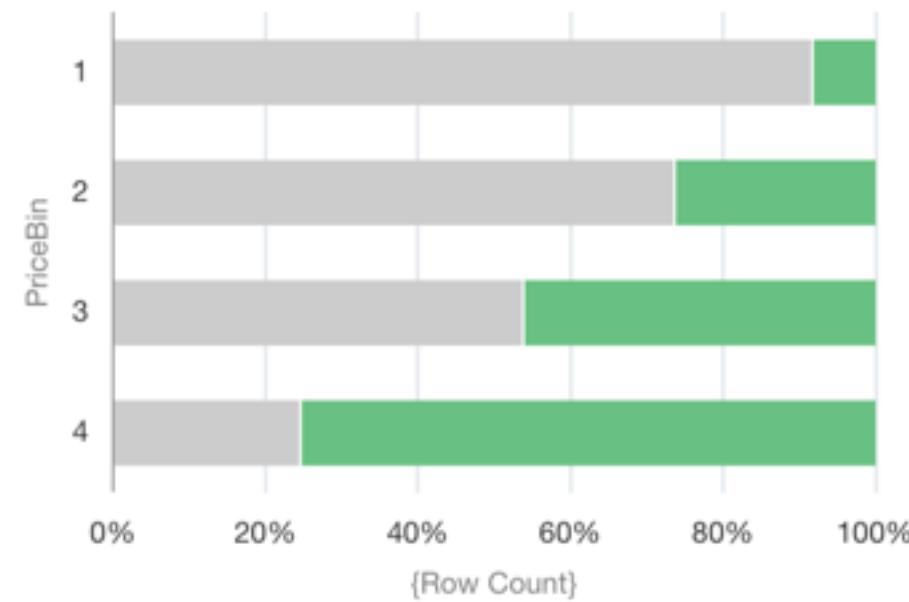


Explain - Key Drivers

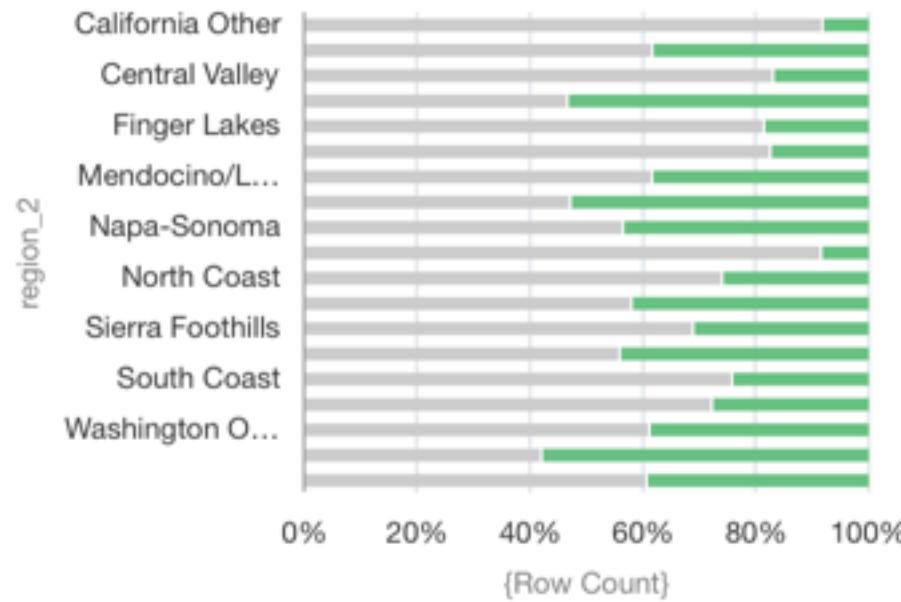
Key Drivers of Good or Bad Wine

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.



Good or Bad Wine: Good Other Good



Good or Bad Wine: Good Other Good

Natural Language Generation

Language Narrative ▾

Attributes

A Name

Values

Stamina

Crossing

Filters

Stamina, Crossing by Name

The data compares the Stamina with the Crossing for a total of 997 Names.

Focus on Stamina

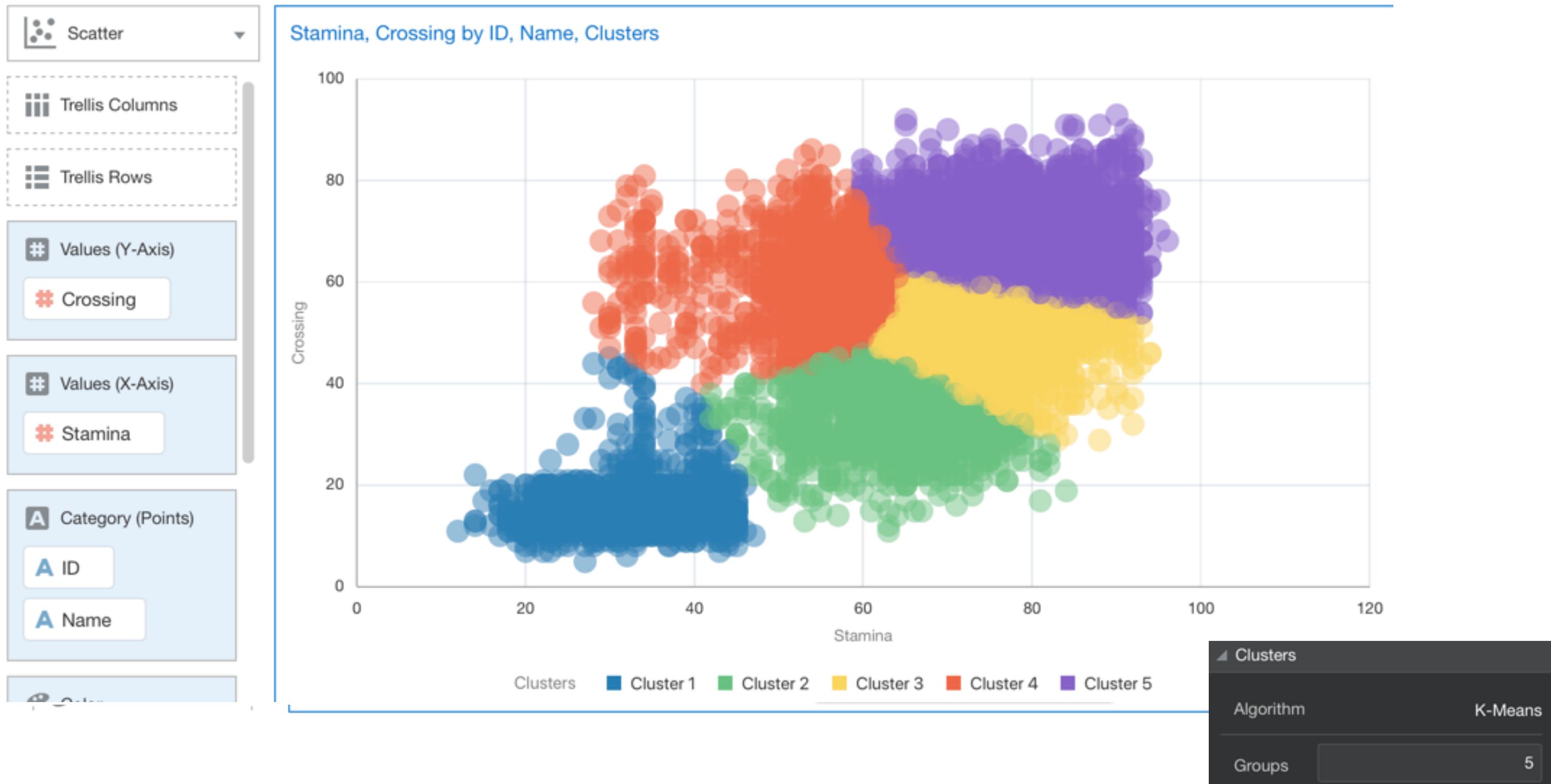
- When taken together, the 997 Names reach a total value of 66,234, an average of 66.43. The most frequent value is 68 and appears 35 times.
- The data was able to be divided into 4 distinct categories.
 - A. González is the biggest, with a Stamina of 313.
 - A. Ba and A. Correa are the next two in terms of Stamina, with 259 on average (0.78% of the total Stamina, about 0.39% each).
 - A. Al Khaibari, A. Majrashi and A. Castro are the next three in terms of Stamina, with 206.67 on average (0.94% of the total Stamina, about 0.31% each).
 - A. Mosquera, A. Diallo, A. Davies and 988 others finish the list, with 65.37 on average. This last group makes up the majority of Names (97.81% of the total Stamina, about 0.1% each).

Focus on Crossing

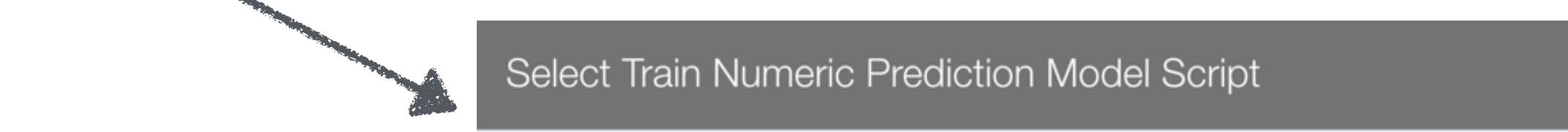
- When taken together, the 997 Names amount to a total value of 51,410, 51.57 on average. The most frequent value is 65 and appears 37 times.
- The data was able to be divided into 4 distinct categories.
 - A. González is the largest, with a Crossing of 263.
 - A. Castro is the second biggest, with a Crossing of 202.
 - A. Correa's numbers were not as high, but it is the third most important, with a Crossing of 172.
 - The remaining Names, A. Ba, A. Gómez, A. Majrashi and 991 others, finish the list, with 51.08 on average. Combined, this last group contains the majority of Names (98.76% of the total Crossing, approximately 0.1% each).

The comparison of two unordered measures is not yet available. Unordered means that the data is not in chronological order. The application will generate a separate analysis for each measure. Stay tuned, future releases will add functionalities for unordered dimensions.

Easy Models



DataFlow Train Model



Search



Train Numeric Prediction

Model Training Script [Linear Regression for model training](#)

* Target [Select a column](#)
target, the target(label) to learn/predict

Regression Method [Lasso](#)
Method for linear regression training.

Regularization Weight
Regularization Weight(L1 Ratio or L2 Ratio). Please enter 0 if it is Ordinary Least Squares linear regression.

Categorical Column Imputation [Most Frequent](#)
The mode method for categorical features to fill NA. Two options: most frequent and least frequent. Default is most frequent.

Numerical Column Imputation [Mean](#)
The mode method for numeric features to fill NA. Four options: mean, max, min, median. Default is mean.

Categorical Encoding Method [Indexer](#)
Encoding method.

Maximum Null Value Percent
Maximum Null Value Percent

Train Partition Percent
Train Partition Percent

Linear Regression for model training

Elastic Net Linear Regression for model training

Random Forest for Numeric model training

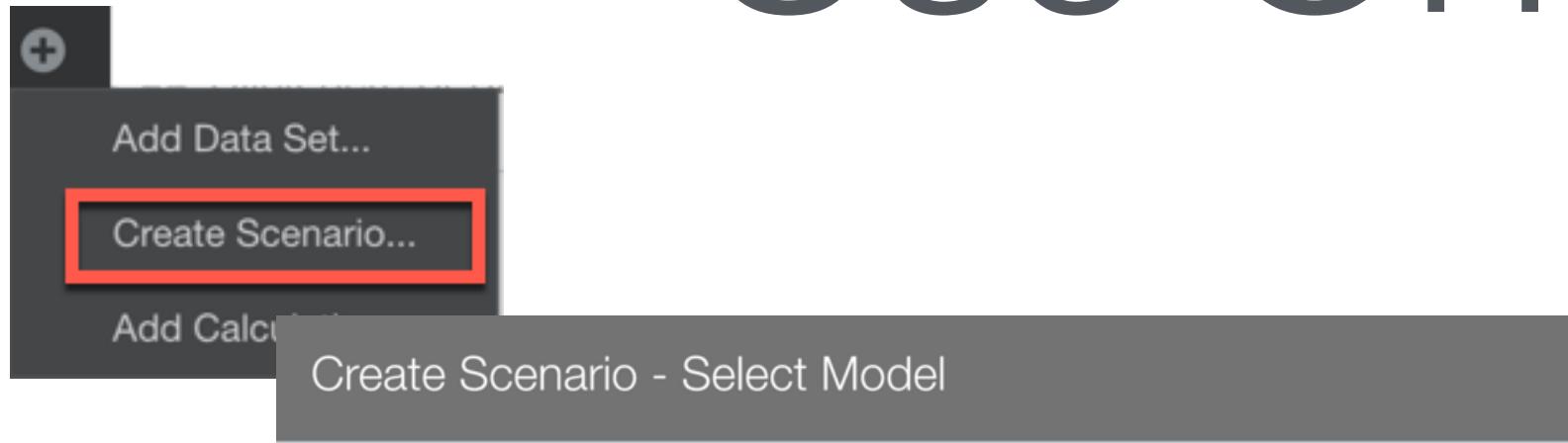
CART for Numeric Prediction training

Compare - Classification

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

		Predicted Value	
		Good	Bad
Real Value	Good		
	Bad		

Use On the Fly



Create Scenario - Select Model

Search grid icon grid icon

Edit Scenario - Map Your Data

Type Name

	BinaryCart2
	BinaryCart1
	BinaryLogistic1
	ELN1
	LR2
	LR1

Select which Data Set you want to use with the Model

Data Set

For each model input listed on the left, select a corresponding data element from your project

Model Input	Map To
bodypart	* bodypart
location	* location
player	* player
situation	* situation
is_goal	is_goal

* Required Fields

id_event by is_goal, is_goal Prediction
event_type: 1

Legend: 0.0 (blue), 1.0 (green), 1.0 (yellow)

Congratulations!



...You are now a Data Scientist!



Nearly
There



...But

Data Cleaning

Feature
Engineering

Model Creation &
Evaluation

Feature
Selection

70% > 50%

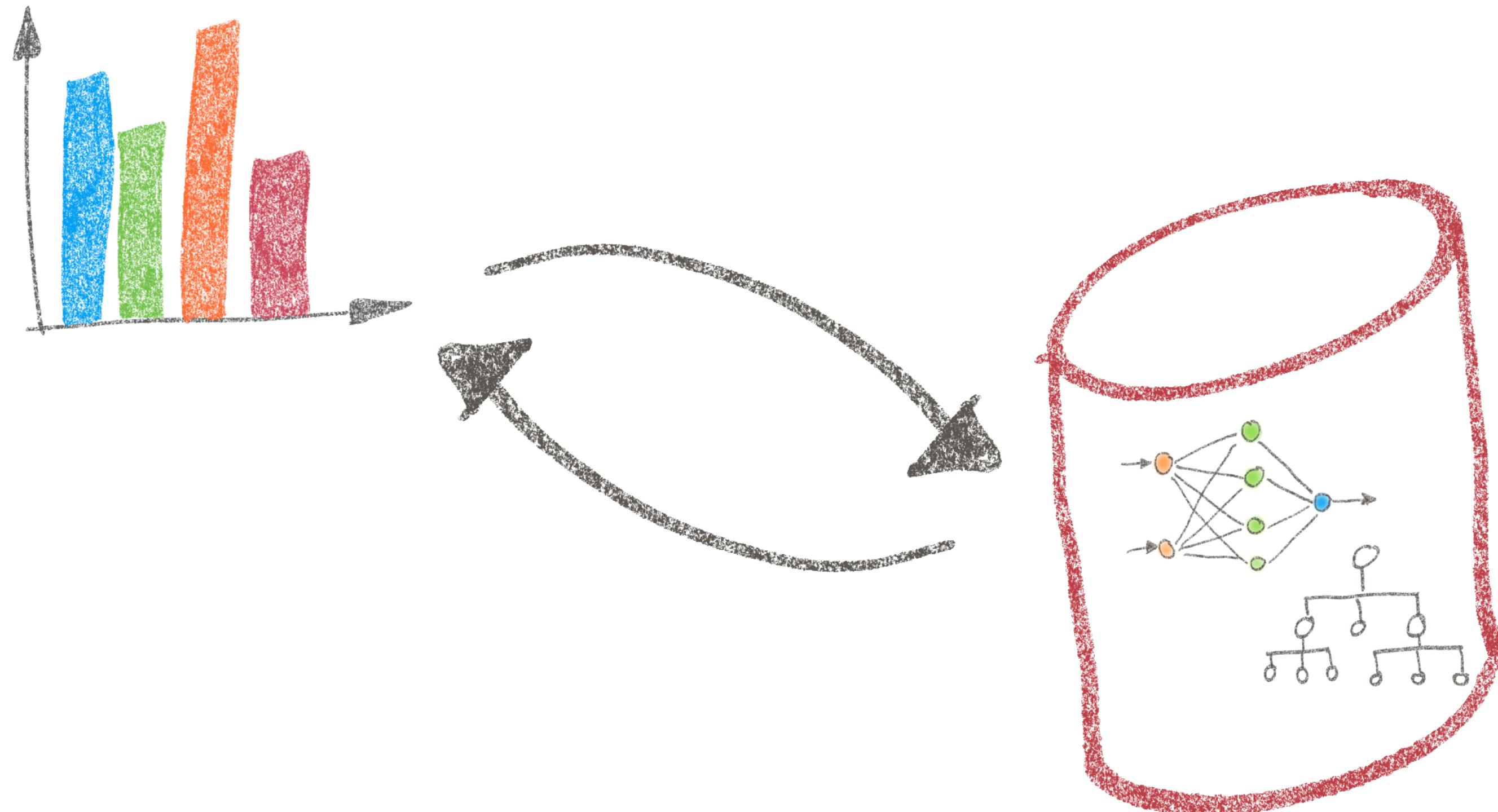
ML Production Deployment

Data Scientist

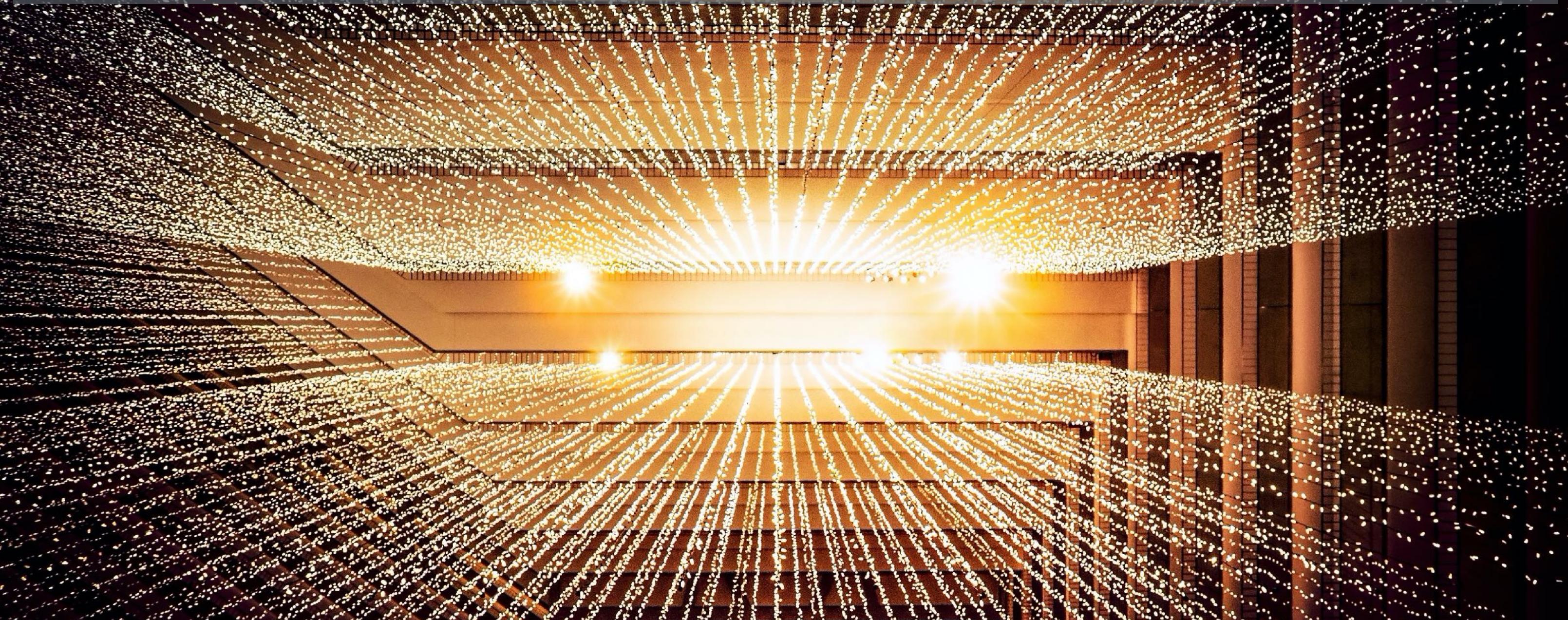
ML -> Data

Oracle Machine Learning

Access Oracle Machine Learning Models on OAC



Become a Data Scientist with OAC



<http://ritt.md/OAC-datascience>

ML in Action with OAC



Hello
Wine,
Goodbye
Problems

<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