Data Science for the Rest of Us

Brandon Rohrer, Senior Data Scientist

Welcome to the webinar!
We'll start in a minute.

If you…

are wondering whether to hire a data scientist,
are thinking about becoming a data scientist,
want to surprise and impress your data scientist co-workers,
have a loved one that is a data scientist,

… I wrote this presentation for you.

TL;DR

Data science isn't magic,
but there are tricks
and secrets.

You can’t use just any data.

Recipe

Crispy crust
Marinara sauce
Fresh mozzarella
Thick cut pepperoni
Fresh tomato

Recipe

Data that is

- Relevant
- Connected
- Accurate
- Enough

and a Sharp Question

Irrelevant data

Price of milk ($/gal)	Red Sox batting avg.	Blood alcohol content (%)
3.79	.304	.03
3.45	.320	.09
4.06	.259	.01
3.89	.298	.05
4.12	.332	.13
3.92	.270	.06
3.23	.294	.10

Relevant data

Body mass (kg)	Margaritas	Blood alcohol content (%)
103	3	.03
67	5	.09
87	1	.01
52	2	.05
73	5	.13
79	3	.06
110	7	.10

[data points] [rows] [samples] [features] [columns] [attributes] [table] [database]

Irrelevant data

Price of milk ($/gal)	Red Sox batting avg.	Blood alcohol content (%)
3.79	.304	.03
3.45	.320	.09
4.06	.259	.01
3.89	.298	.05
4.12	.332	.13
3.92	.270	.06
3.23	.294	.10

Relevant data

Body mass (kg)	Margaritas	Blood alcohol content (%)
103	3	.03
67	5	.09
87	1	.01
52	2	.05
73	5	.13
79	3	.06
110	7	.10

[data points] [rows] [samples] [features] [columns] [attributes] [table] [database]

Irrelevant data

Price of milk ($/gal)	Red Sox batting avg.	Blood alcohol content (%)
3.79	.304	.03
3.45	.320	.09
4.06	.259	.01
3.89	.298	.05
4.12	.332	.13
3.92	.270	.06
3.23	.294	.10

Relevant data

Body mass (kg)	Margaritas	Blood alcohol content (%)
103	3	.03
67	5	.09
87	1	.01
52	2	.05
73	5	.13
79	3	.06
110	7	.10

[data points] [rows] [samples] [features] [columns] [attributes] [table] [database]

Disconnected data

Grill temperature (F)	Weight of beef patty (lb)	Burger rating (out of 10)
	.33	8.2
	.24	5.6
550		7.8
725	.45	9.4
600		8.2
625		6.8
	.49	4.2

Connected data

Grill temperature (F)	Weight of beef patty (lb)	Burger rating (out of 10)
575	.33	8.2
550	.24	5.6
550	.69	7.8
725	.45	9.4
600	.57	8.2
625	.36	6.8
550	.49	4.2

[missing values]

Disconnected data

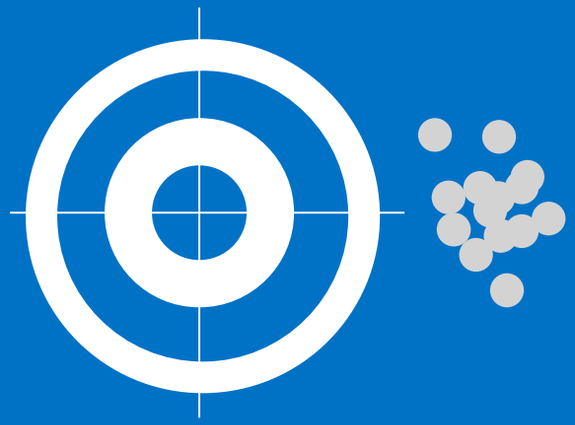
Grill temperature (F)	Weight of beef patty (lb)	Burger rating (out of 10)
	.33	8.2
	.24	5.6
550		7.8
725	.45	9.4
600		8.2
625		6.8
	.49	4.2

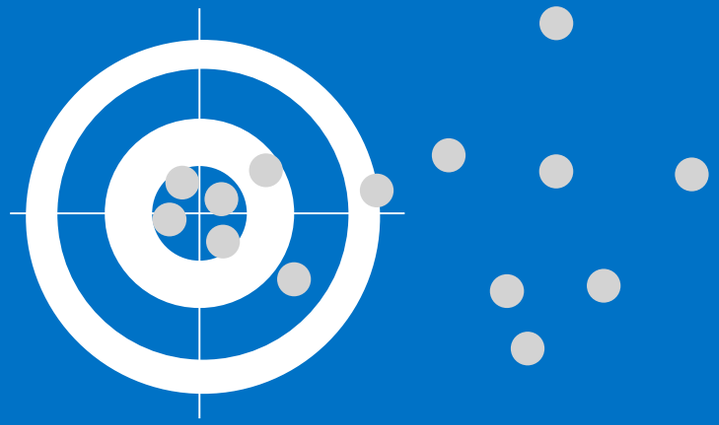
Connected data

Grill temperature (F)	Weight of beef patty (lb)	Burger rating (out of 10)
575	.33	8.2
550	.24	5.6
550	.69	7.8
725	.45	9.4
600	.57	8.2
625	.36	6.8
550	.49	4.2

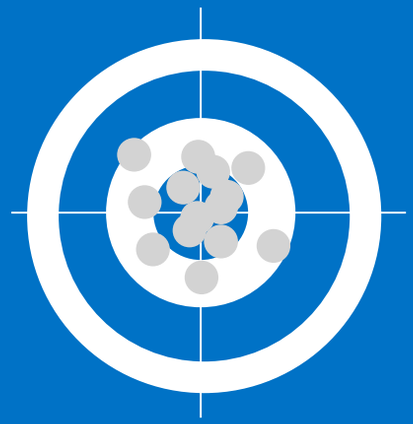
[missing values]

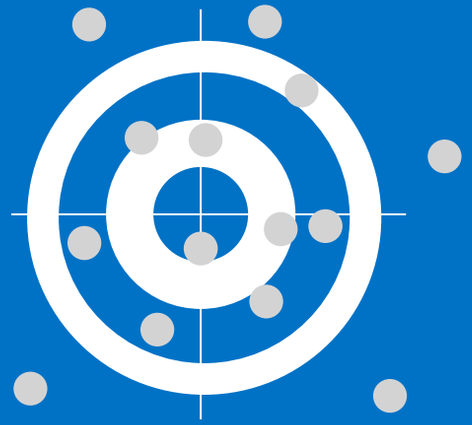
Inaccurate data





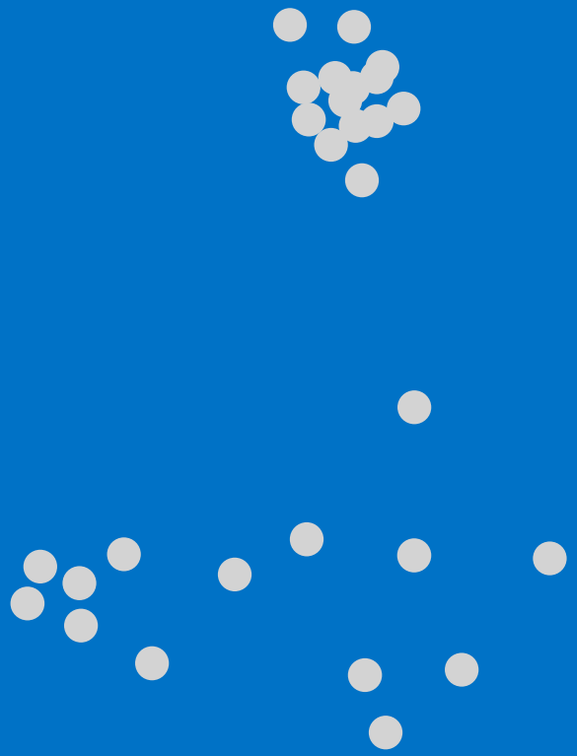
Accurate data



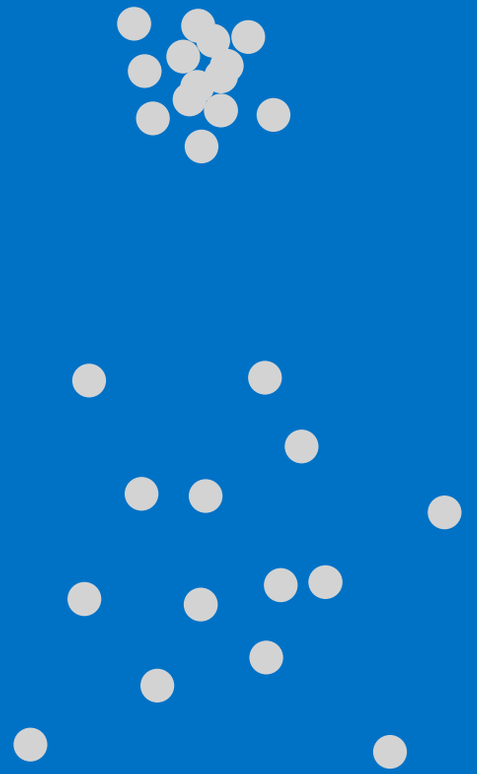


[precision] [accuracy] [bias]

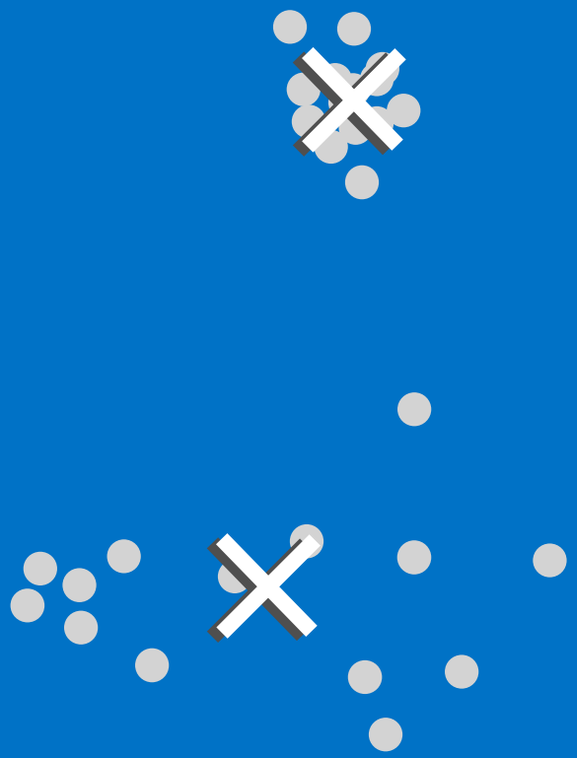
Inaccurate data



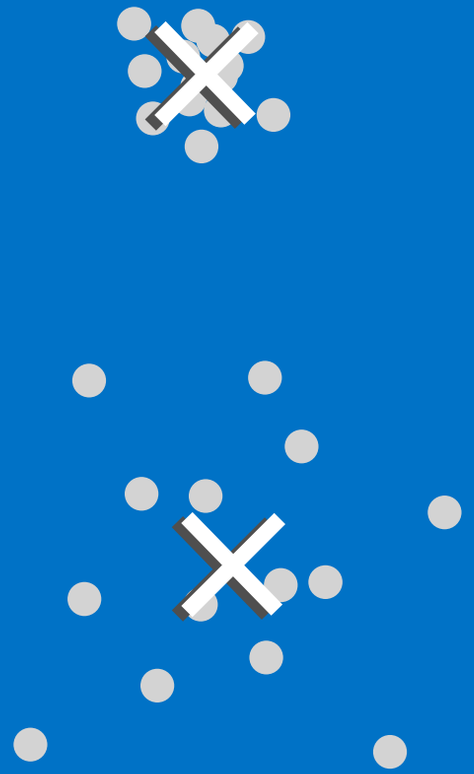
Accurate data



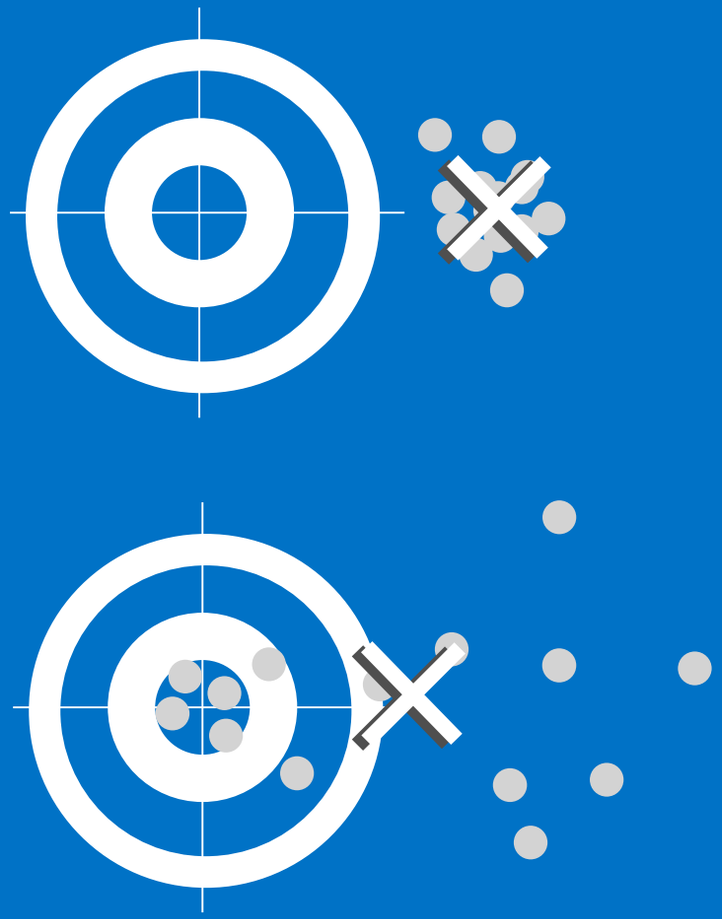
Inaccurate data



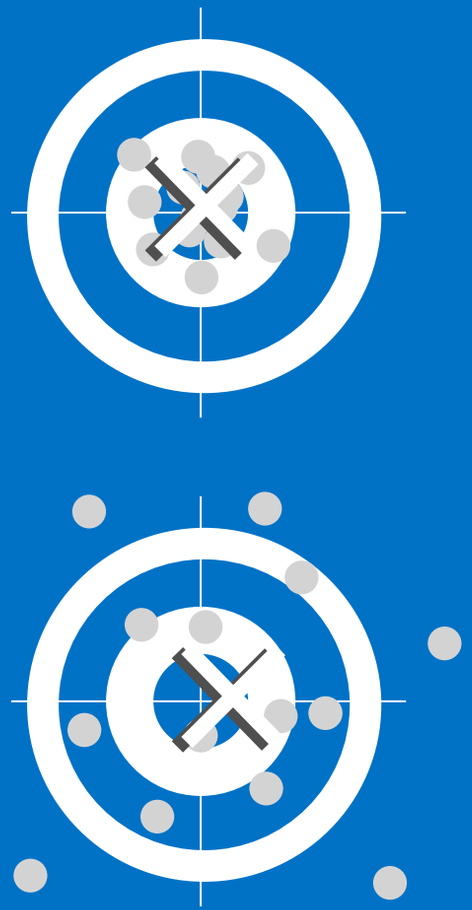
Accurate data



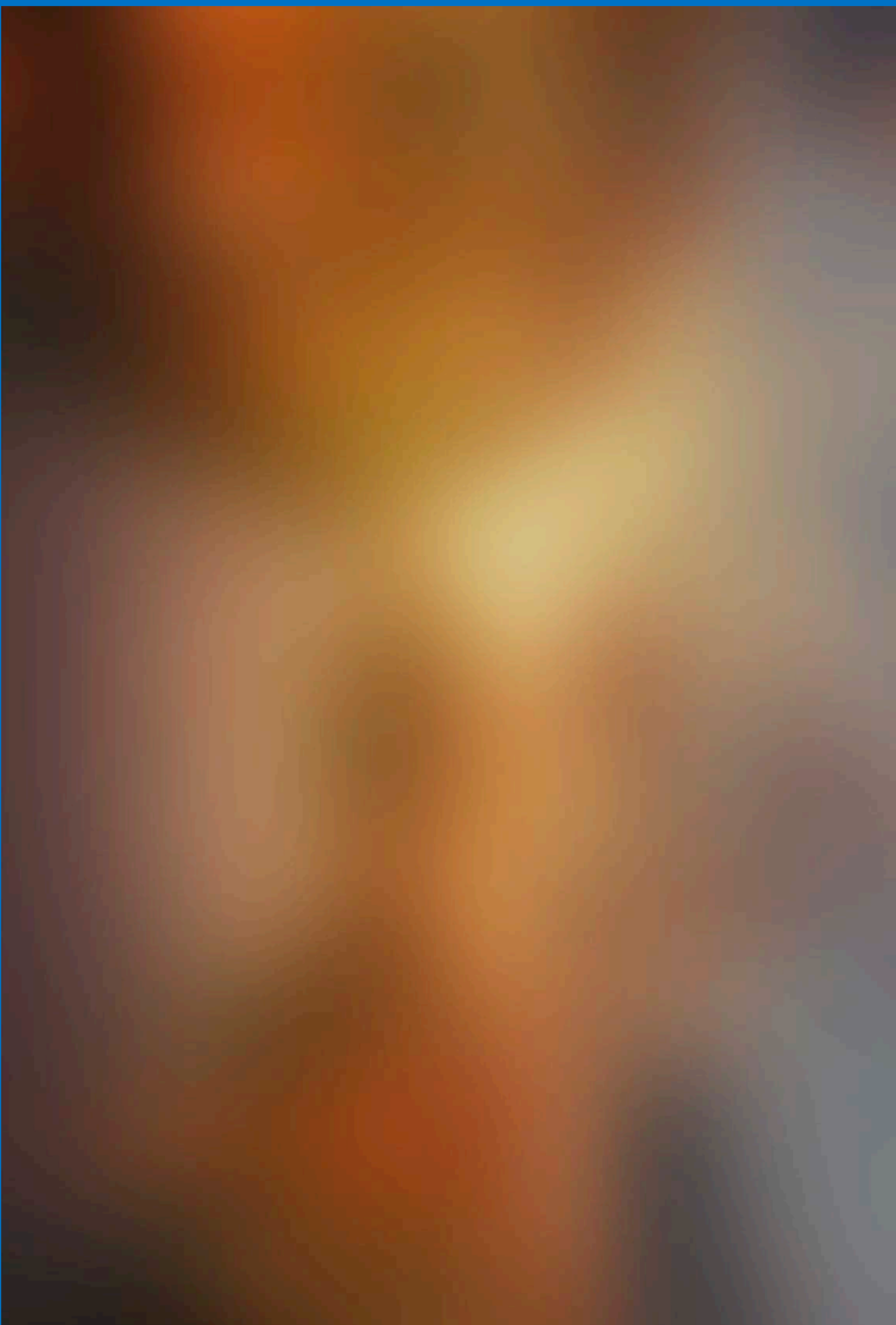
Inaccurate data



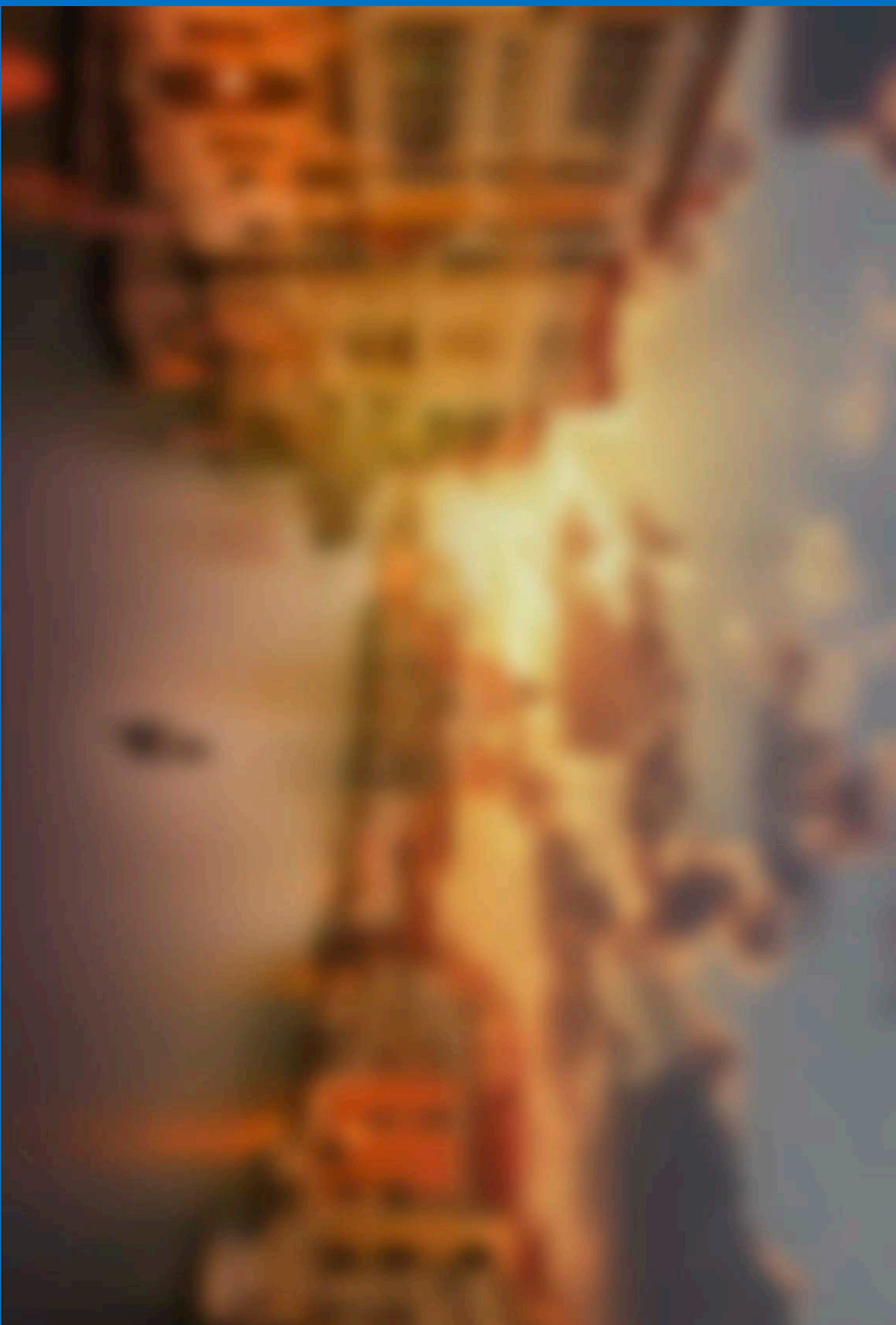
Accurate data



Not enough data

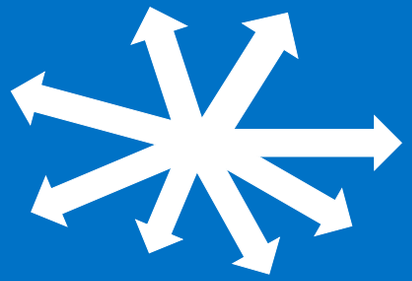


Barely enough data



Enough data

Vague questions vs. Sharp questions



Can't be answered with a name or a number

What can my data tell me about my business?

What should I do?

How can I increase my profits?

Can be answered with a name or a number.

How many Model Q Gizmos will I sell in Montreal during the third quarter?

Which car in my fleet is going to fail first?

Turn your data into a picture.

Mystery data

Two features
Four classes
1600 data points

[feature] [class] [histogram] [visualization]

rows 1600 columns 3

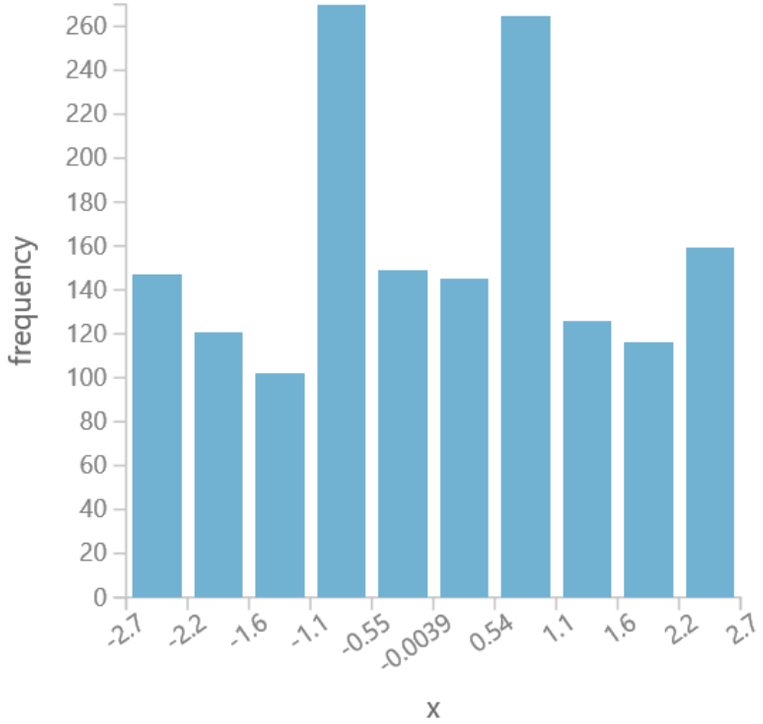
view as

x	y	l
0.756759	-1.373646	S
-1.071171	0.934892	L
0.301784	-1.447297	S
-0.071593	2.472344	C
0.470462	-1.388383	S
0.544586	-1.359058	S
0.689908	-1.280118	S
-0.317403	-1.500969	S
2.268651	0.537407	C
1.111744	-1.283004	S
-0.032994	-1.350037	S
-0.597814	-1.441323	S
1.141835	-2.325779	C
0.114728	-1.605493	S

x

Histogram

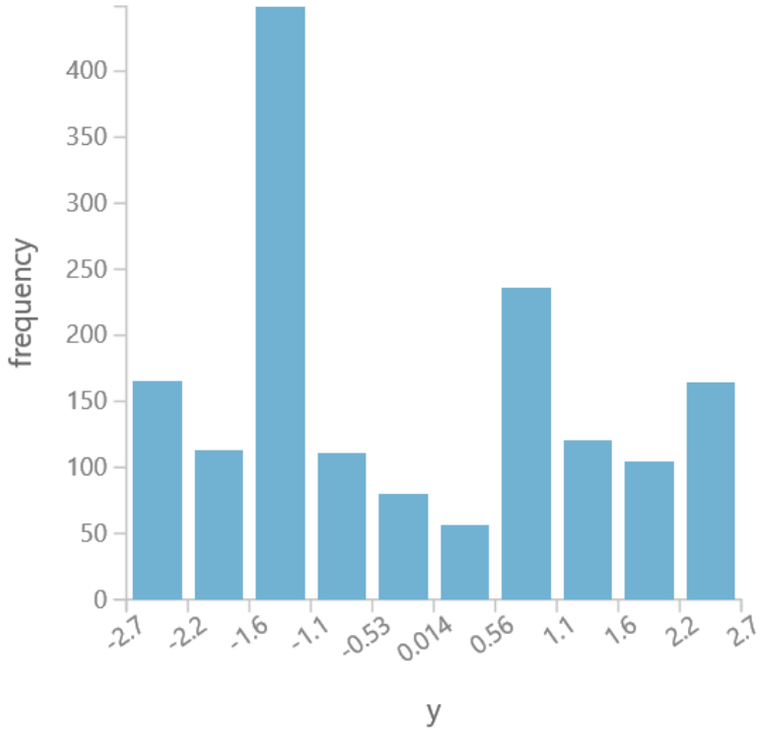
compare to None



y

Histogram

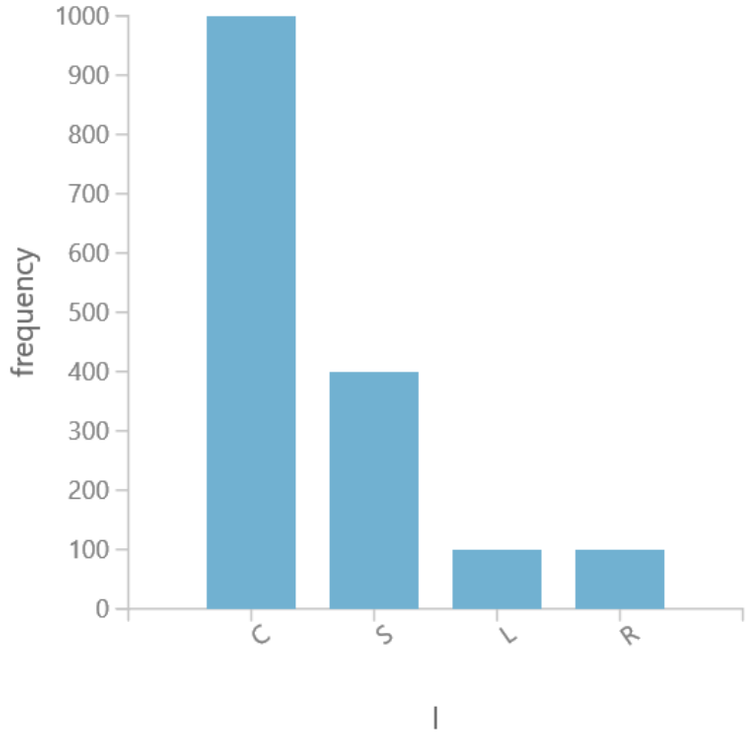
compare to None



l

Histogram

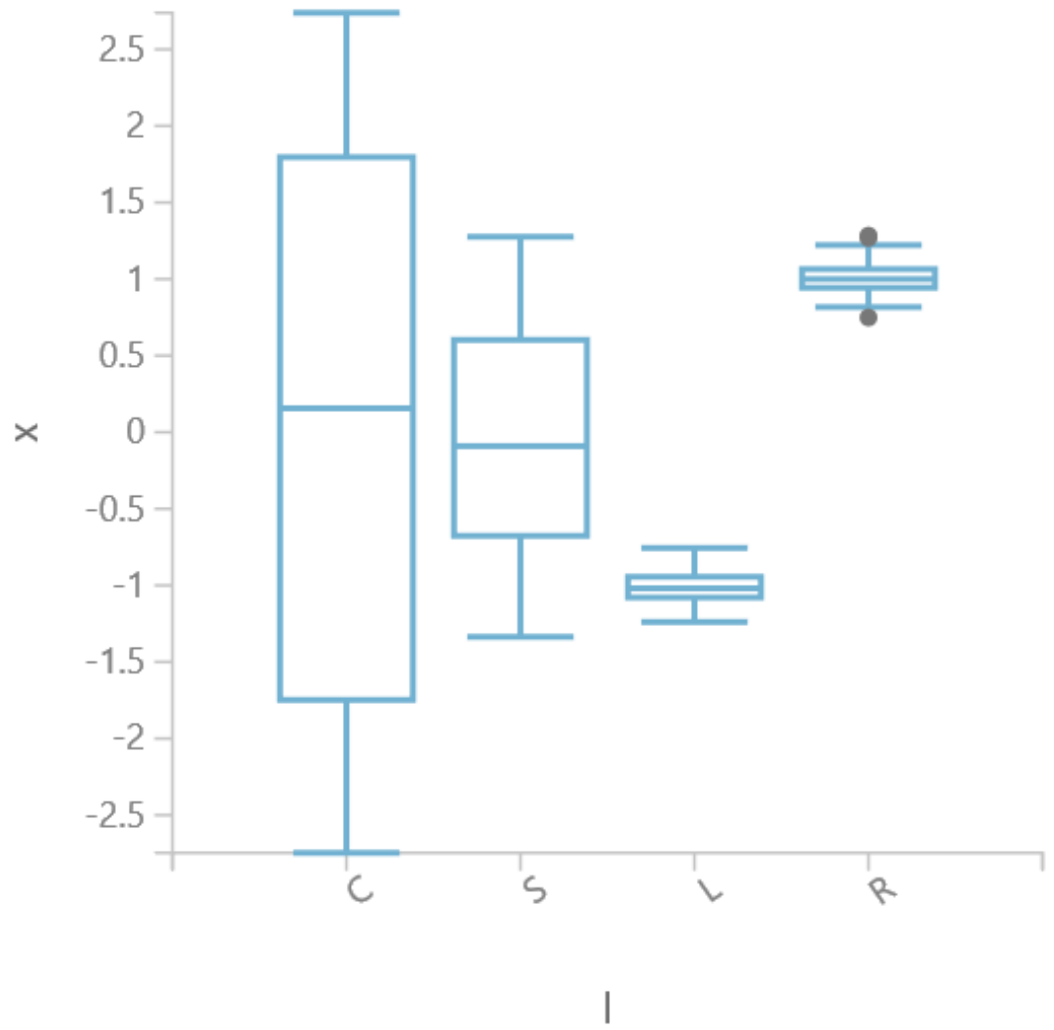
compare to None



x

MultiboxPlot

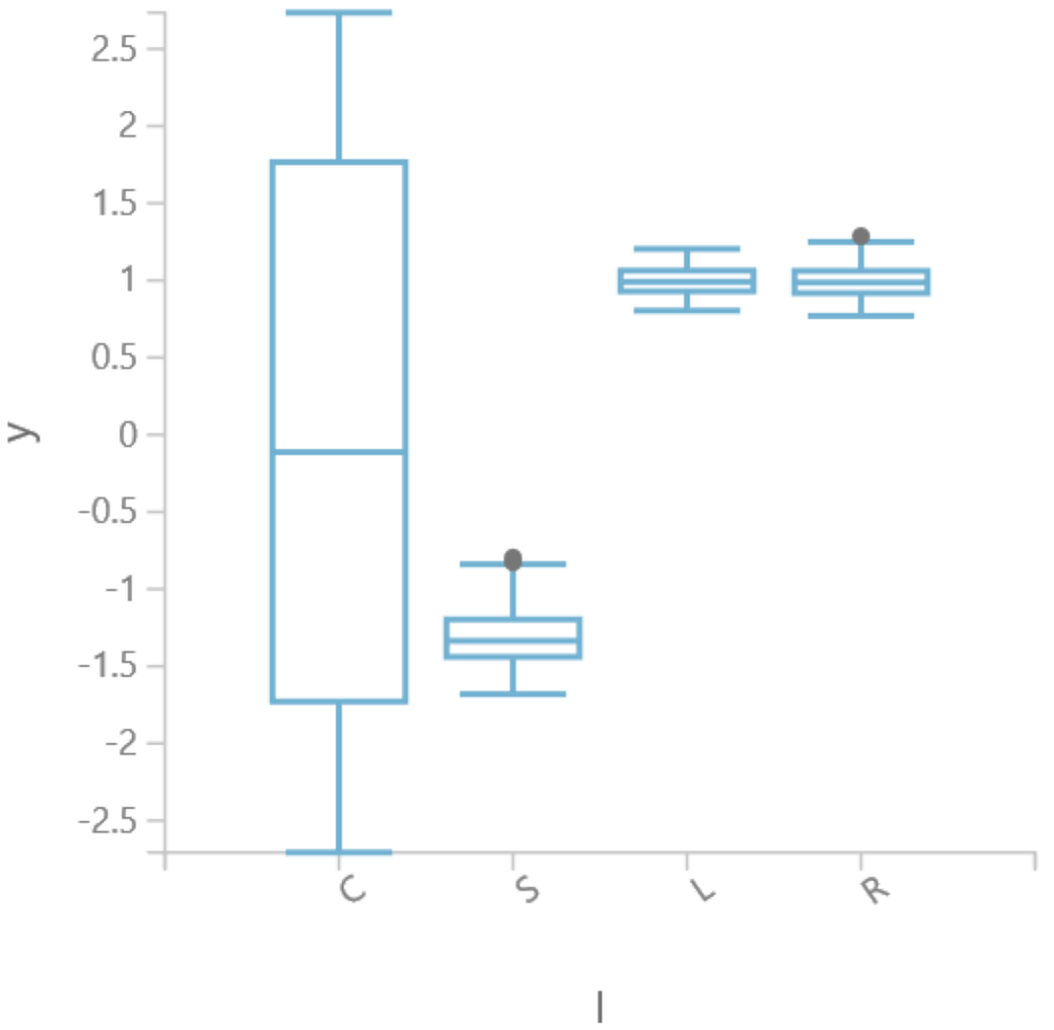
compare to l

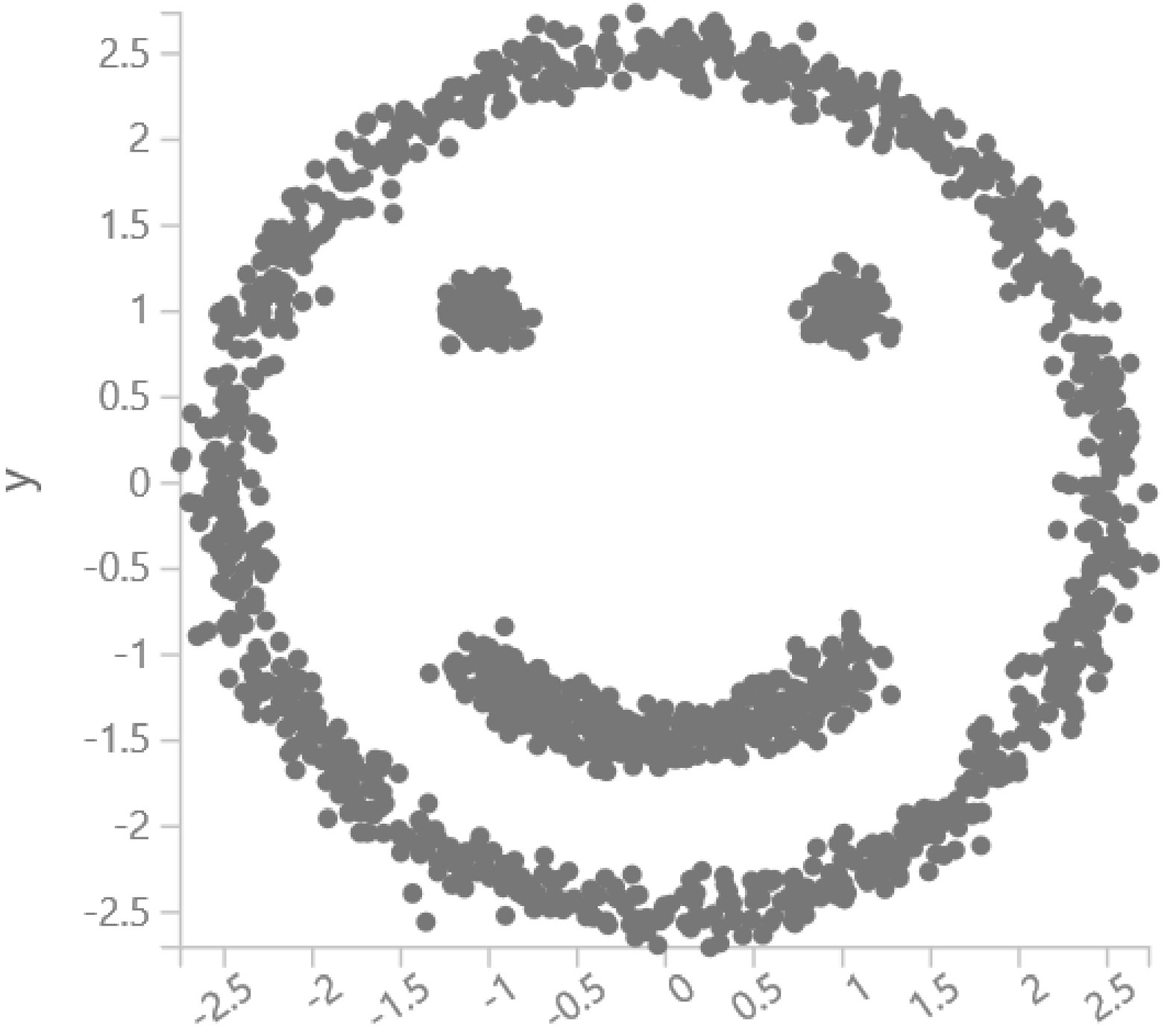


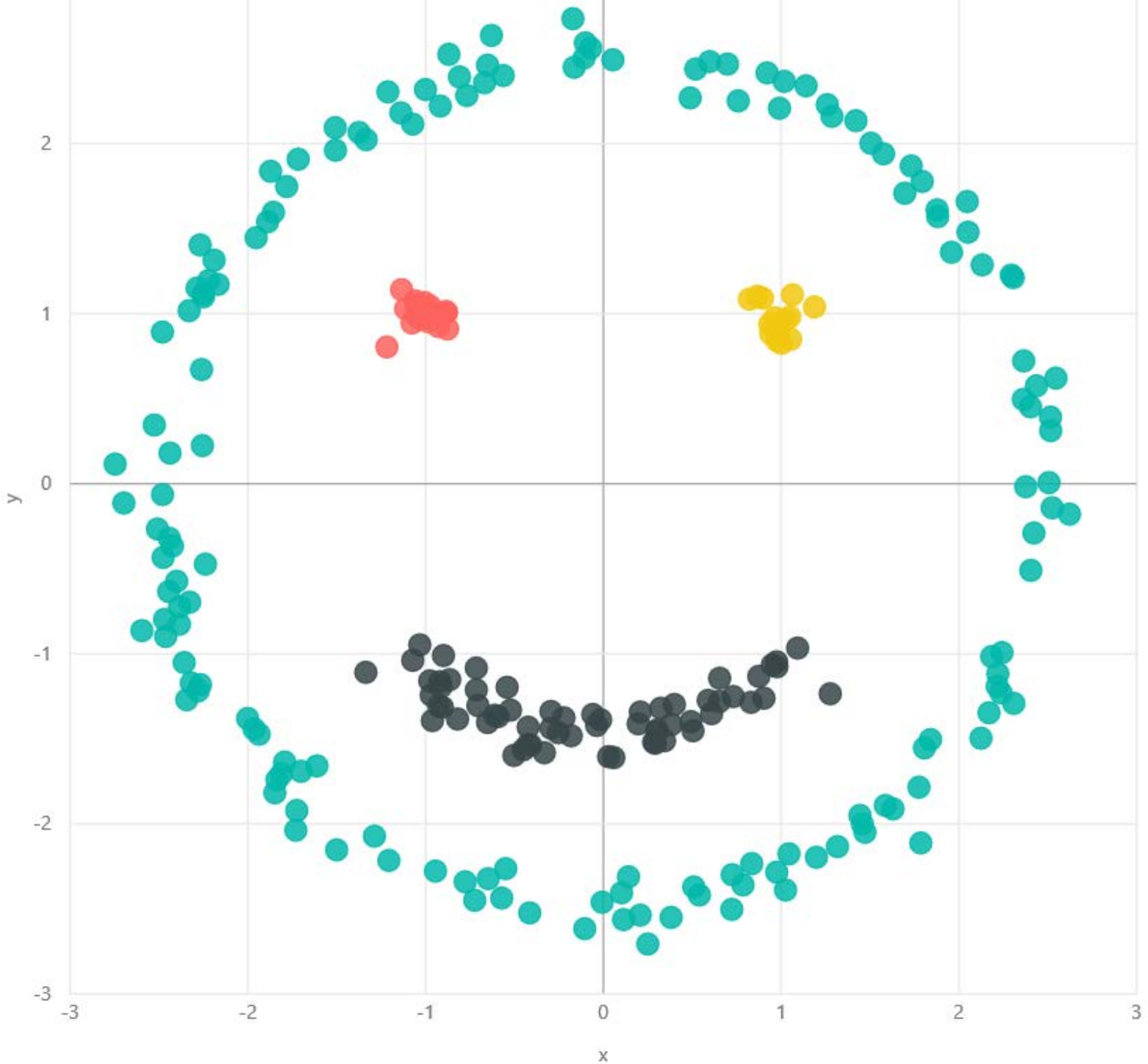
y

MultiboxPlot

compare to l



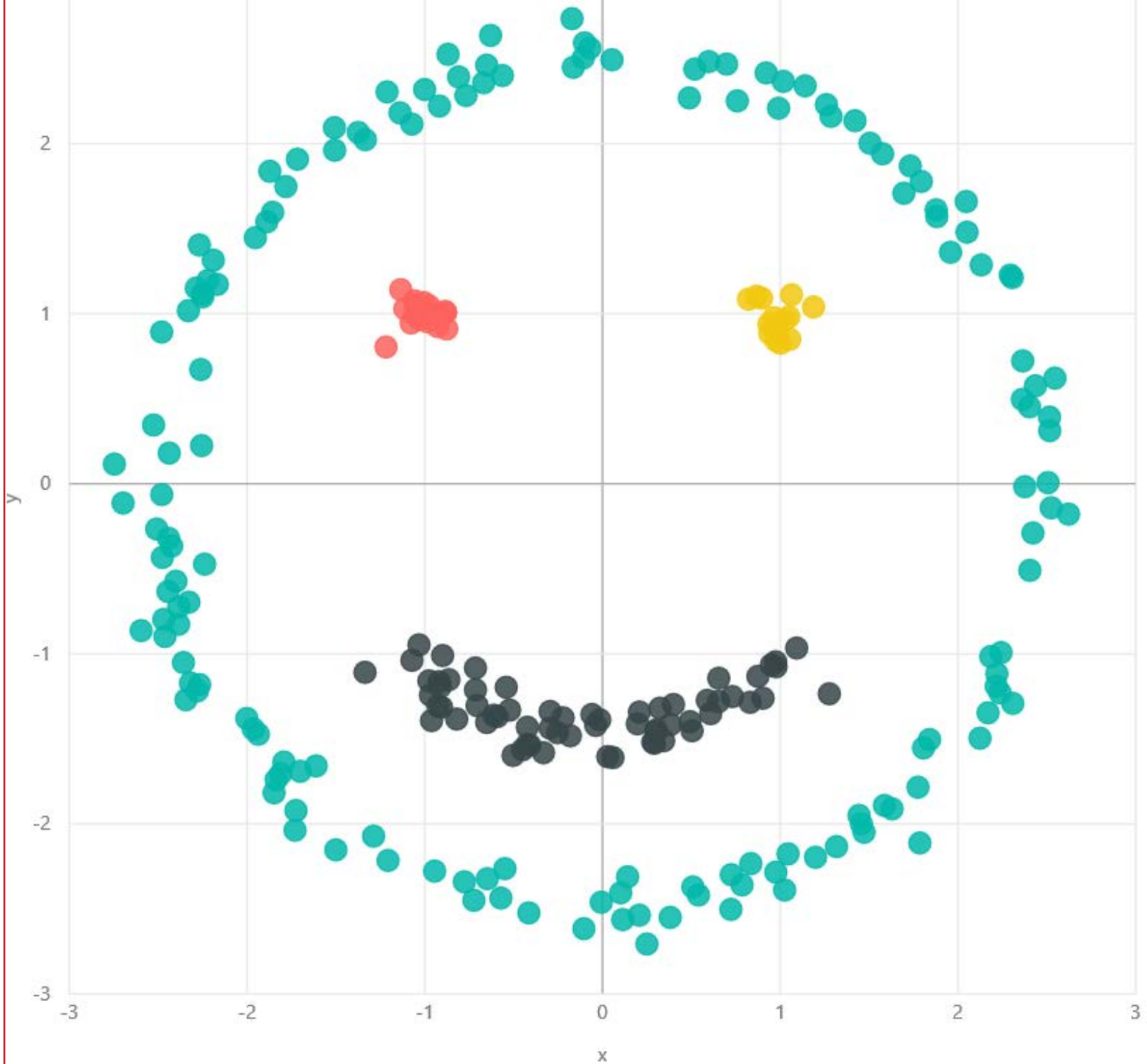

y
2.5
2
1.5
1
0.5
0
-0.5
-1
-1.5
-2
-2.5
-2.5
-2
-1.5
-1
-0.5
0
0.5
1
1.5
2
2.5


2
1
0
-1
-2
-3
y
-3
-2
-1
0
1
2
3
x

rows 1600

columns 3

x	y	l
0.756759	-1.373646	S
-1.071171	0.934892	L
0.301784	-1.447297	S
-0.071593	2.472344	C
0.470462	-1.388383	S
0.544586	-1.359058	S
0.689908	-1.280118	S
-0.317403	-1.500969	S
2.268651	0.537407	C
1.111744	-1.283004	S
-0.032994	-1.350037	S
-0.597814	-1.441323	S
1.141835	-2.325779	C
0.114728	-1.605493	S
-0.961781	-1.3926	S



Data science can only answer five questions.

1. How much / how many?
2. Which category?
3. Which groups?
4. Is it weird?
5. Which action?

[algorithm]

How much / how many?

What will the temperature be next Tuesday?

What will my fourth quarter sales in Portugal be?

How many new followers will I get next week?

[regression]

Which category?

Is this an image of a cat or a dog?

Which aircraft is causing this radar signature?

What is the topic of this news article?

[classification]

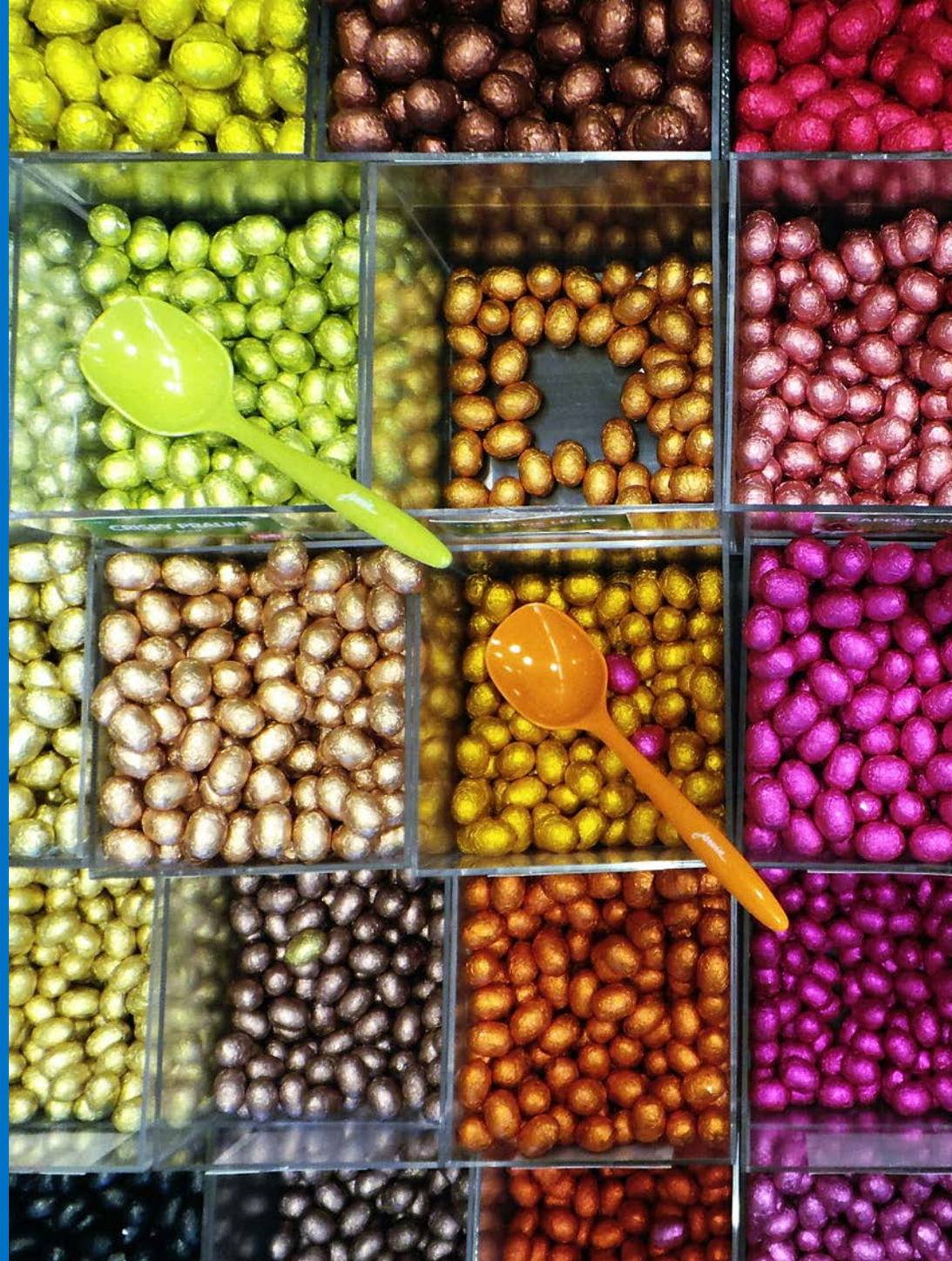
Which groups?

Which shoppers have similar tastes in produce?

Which viewers like the same kind of movies?

What is a natural way to break these documents into five topic groups?

[clustering] [recommendation]



Is this weird?

Is this pressure reading unusual?

Is this internet message typical?

Is this combination of purchases very different from what this customer has made in the past?

[anomaly detection]

Which action?

Should I raise or lower the temperature?

Should I vacuum the living room again or stay plugged in to my charging station?

Should I brake or accelerate in response to that yellow light?

[reinforcement learning]

Machine learning is simple.

Машинное
обучение
не действительно
сложной,
но
оно
иностранный
язык.

Machine
learning
isn’t actually
complicated,
but
it is
a foreign
language.

Machine
learning
isn’t actually
complicated,
but
it is
a foreign
language.

Машинное
обучение
не действительно
сложной,
но
оно
иностранный
язык.

language

язык

язык

yazeek

я	з	ы	к
ya	z	ee	k

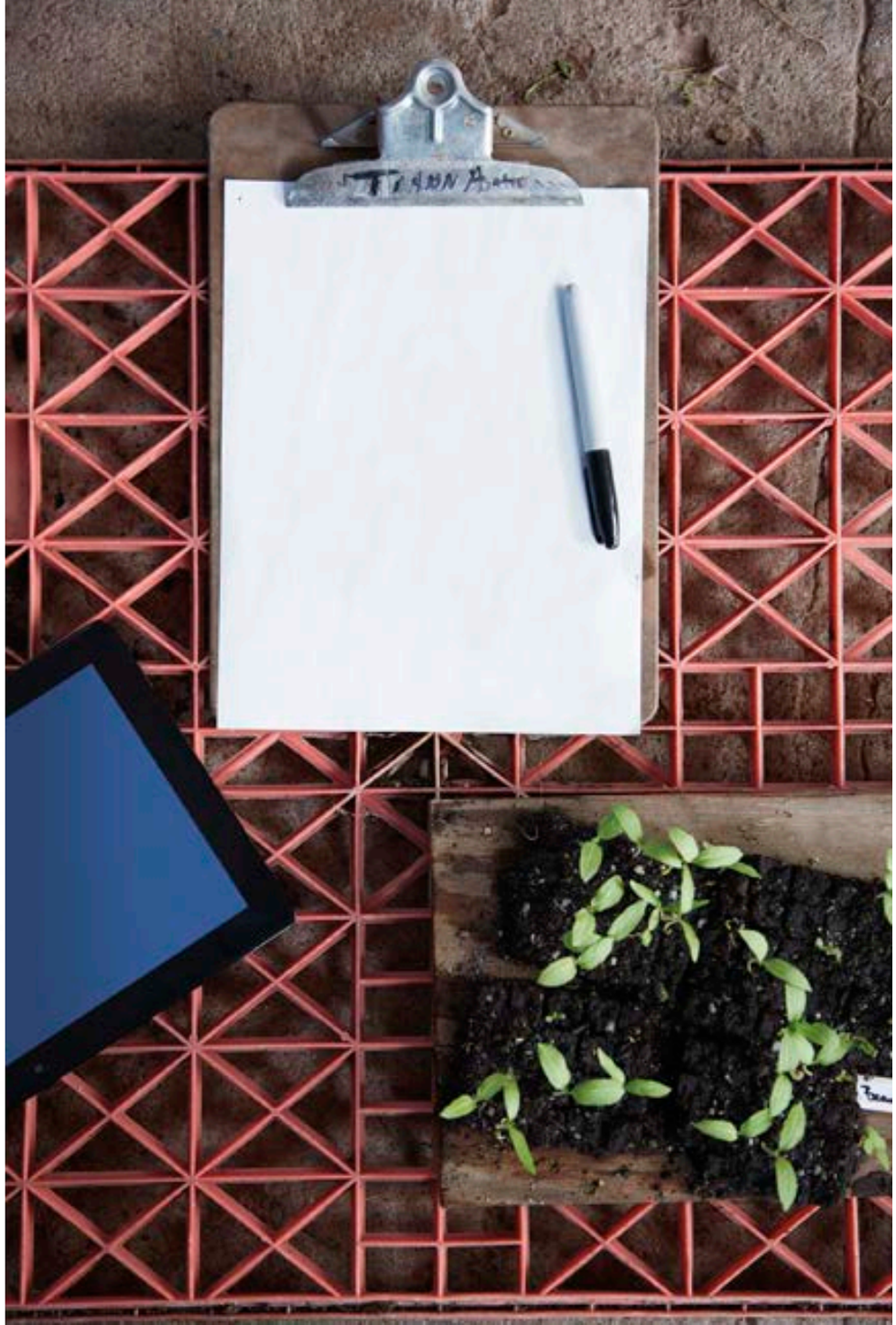
я

уа

Diamonds

Diamonds

carats	price
1.01	$7,366
.49	985
.31	544
1.51	9,140
.37	493
.73	3,011
1.53	11,413
.56	1,814
.41	876
.74	2,690
.63	1,991
.6	4,172
2.06	11,764
1.1	4,682
1.32	6,171
2.02	15,996
.34	695



Diamonds

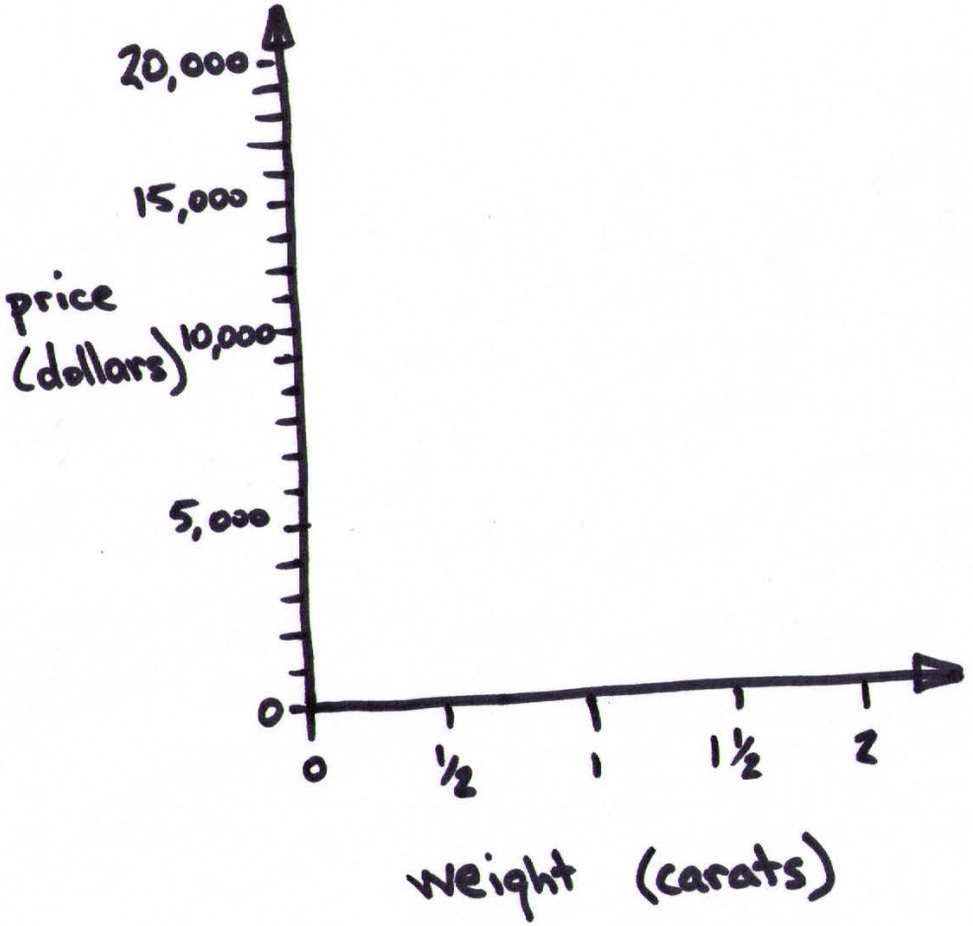
[number line]
[axis]

carats	price
1.01	$7,366
.49	985
.31	544
1.51	9,140
.37	493
.73	3,011
1.53	11,413
.56	1,814
.41	876
.74	2,690
.63	1,991
.6	4,172
2.06	11,764
1.1	4,682
1.32	6,171
2.02	15,996
.34	695

Diamonds

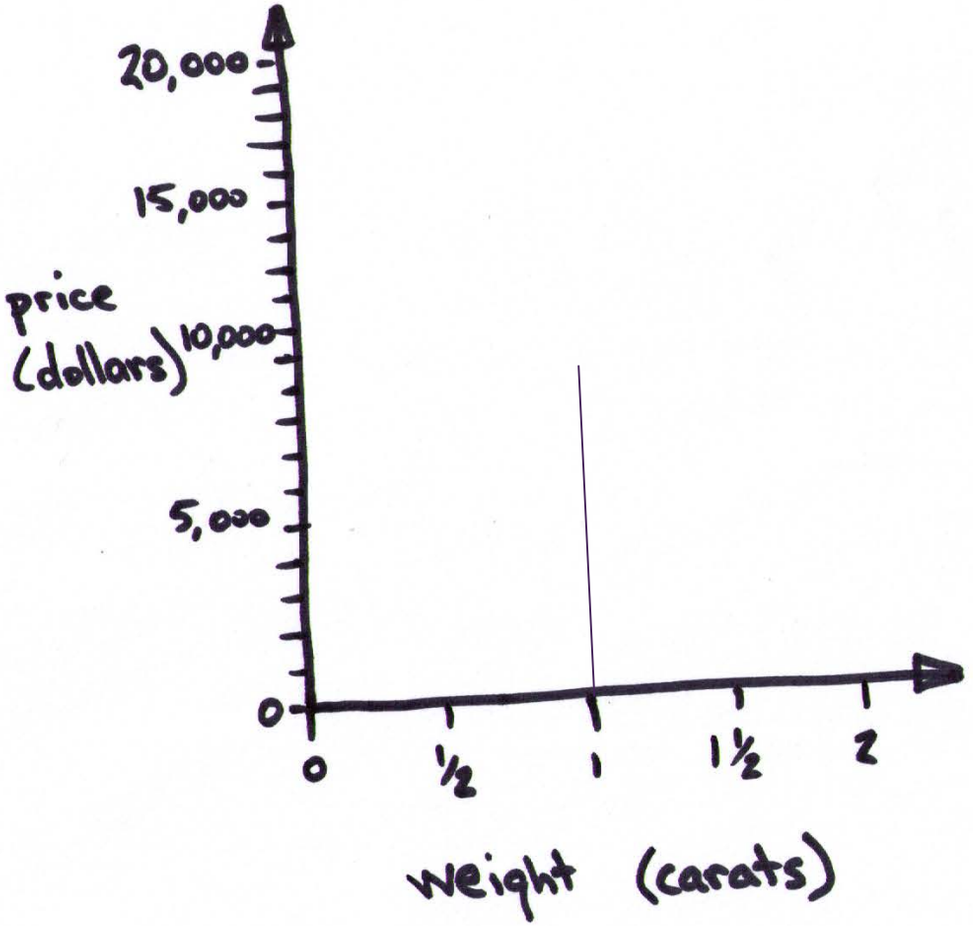
[axes]
[units]

carats	price
1.01	$7,366
.49	985
.31	544
1.51	9,140
.37	493
.73	3,011
1.53	11,413
.56	1,814
.41	876
.74	2,690
.63	1,991
.6	4,172
2.06	11,764
1.1	4,682
1.32	6,171
2.02	15,996
.34	695



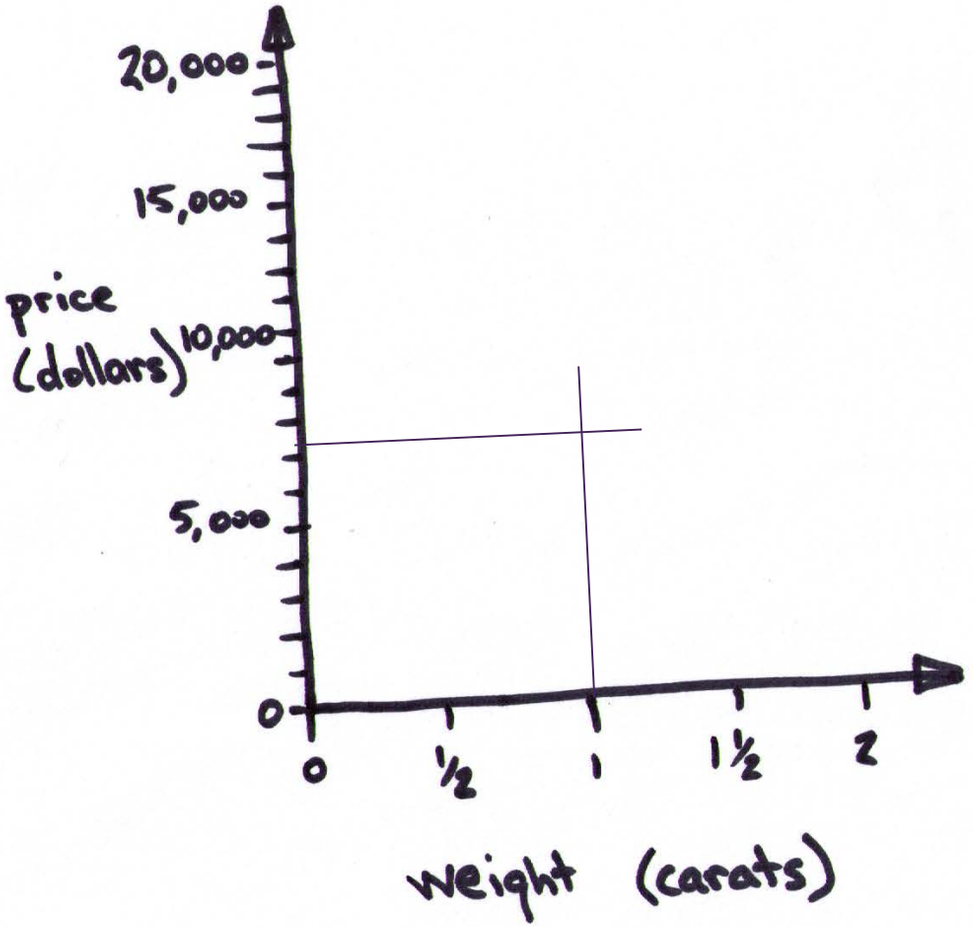
Diamonds

carats	price
1.01	$7,366
.49	985
.31	544
1.51	9,140
.37	493
.73	3,011
1.53	11,413
.56	1,814
.41	876
.74	2,690
.63	1,991
.6	4,172
2.06	11,764
1.1	4,682
1.32	6,171
2.02	15,996
.34	695



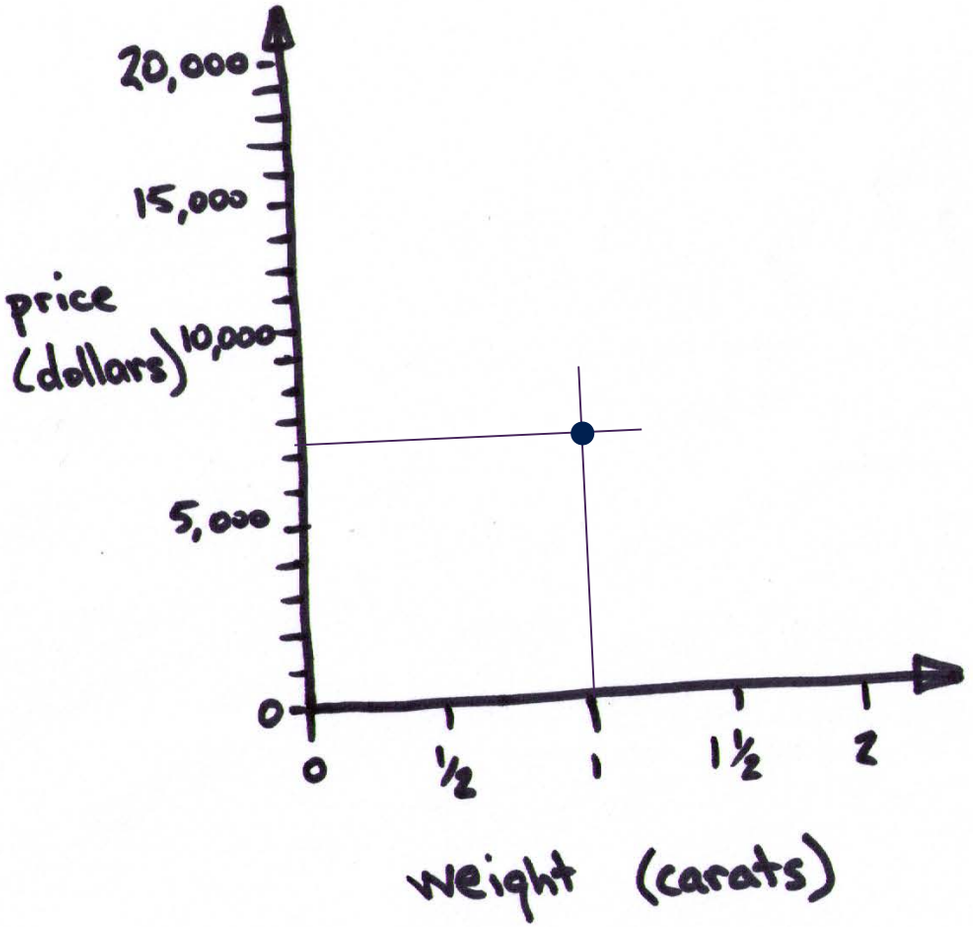
Diamonds

carats	price
1.01	$7,366
.49	985
.31	544
1.51	9,140
.37	493
.73	3,011
1.53	11,413
.56	1,814
.41	876
.74	2,690
.63	1,991
.6	4,172
2.06	11,764
1.1	4,682
1.32	6,171
2.02	15,996
.34	695



Diamonds

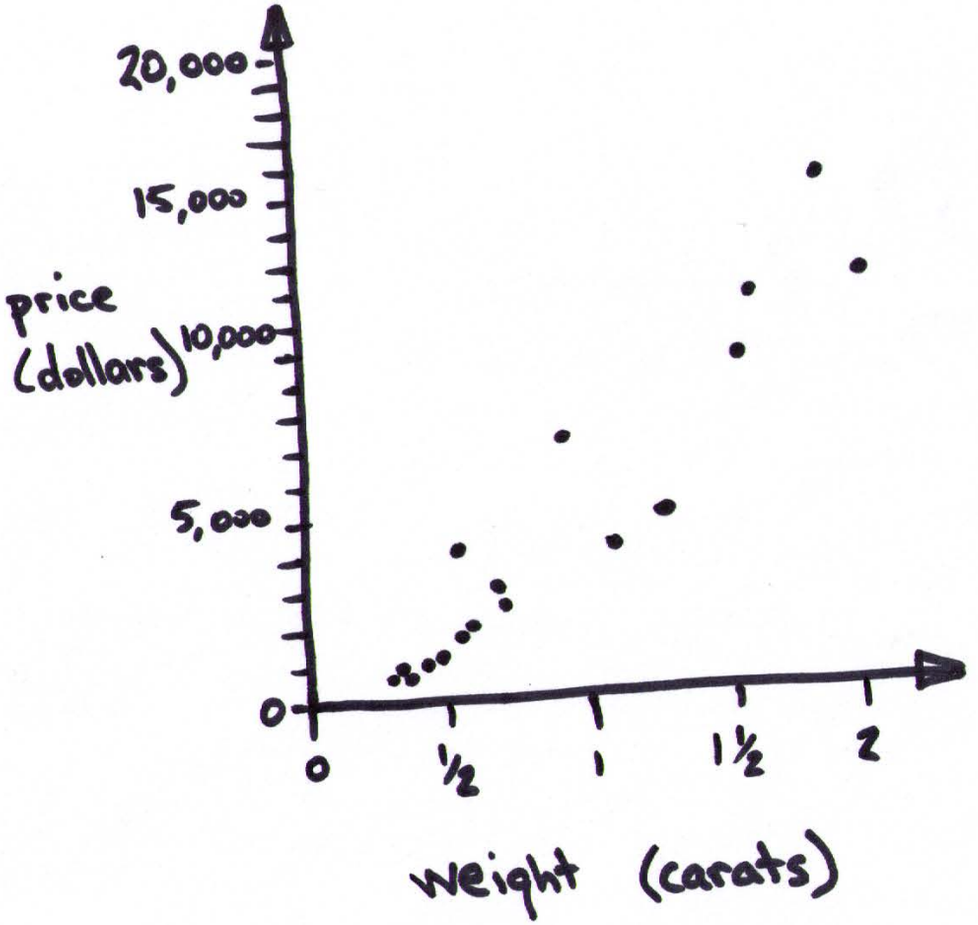
carats	price
1.01	$7,366
.49	985
.31	544
1.51	9,140
.37	493
.73	3,011
1.53	11,413
.56	1,814
.41	876
.74	2,690
.63	1,991
.6	4,172
2.06	11,764
1.1	4,682
1.32	6,171
2.02	15,996
.34	695



Diamonds

[plot]
[scatter plot]

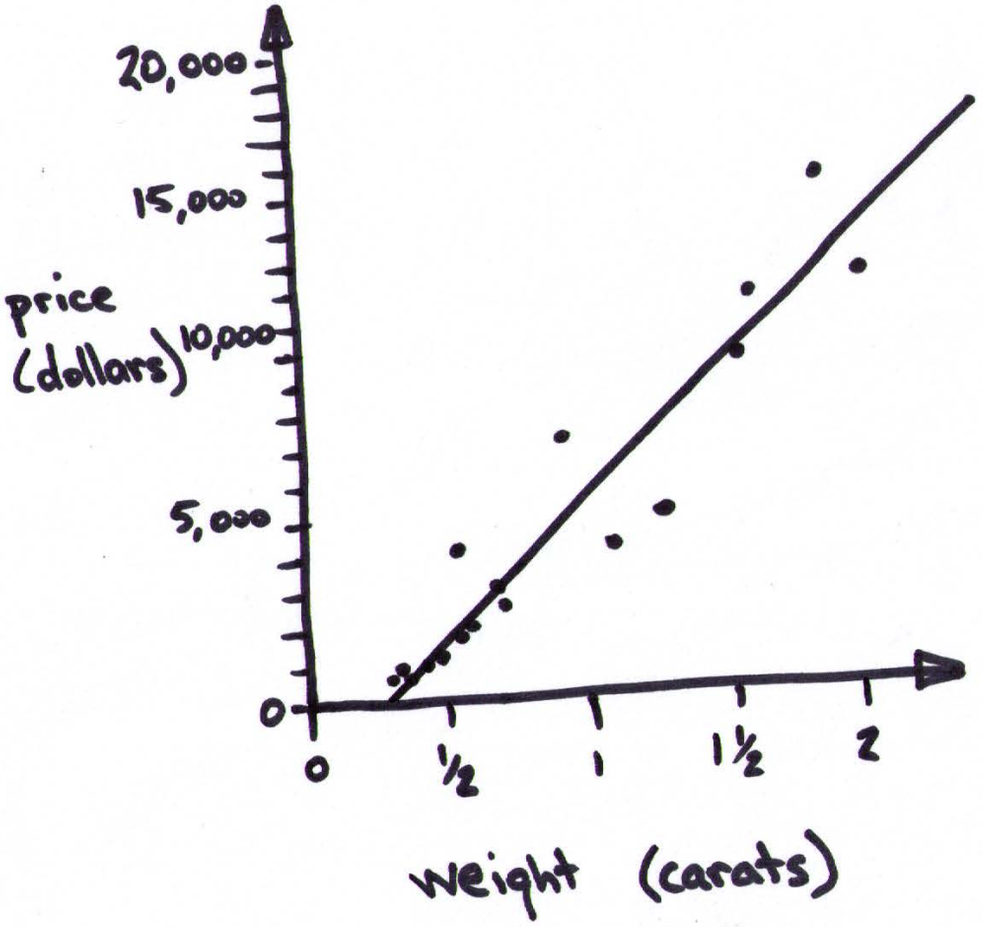
carats	price
1.01	$7,366
.49	985
.31	544
1.51	9,140
.37	493
.73	3,011
1.53	11,413
.56	1,814
.41	876
.74	2,690
.63	1,991
.6	4,172
2.06	11,764
1.1	4,682
1.32	6,171
2.02	15,996
.34	695



Diamonds

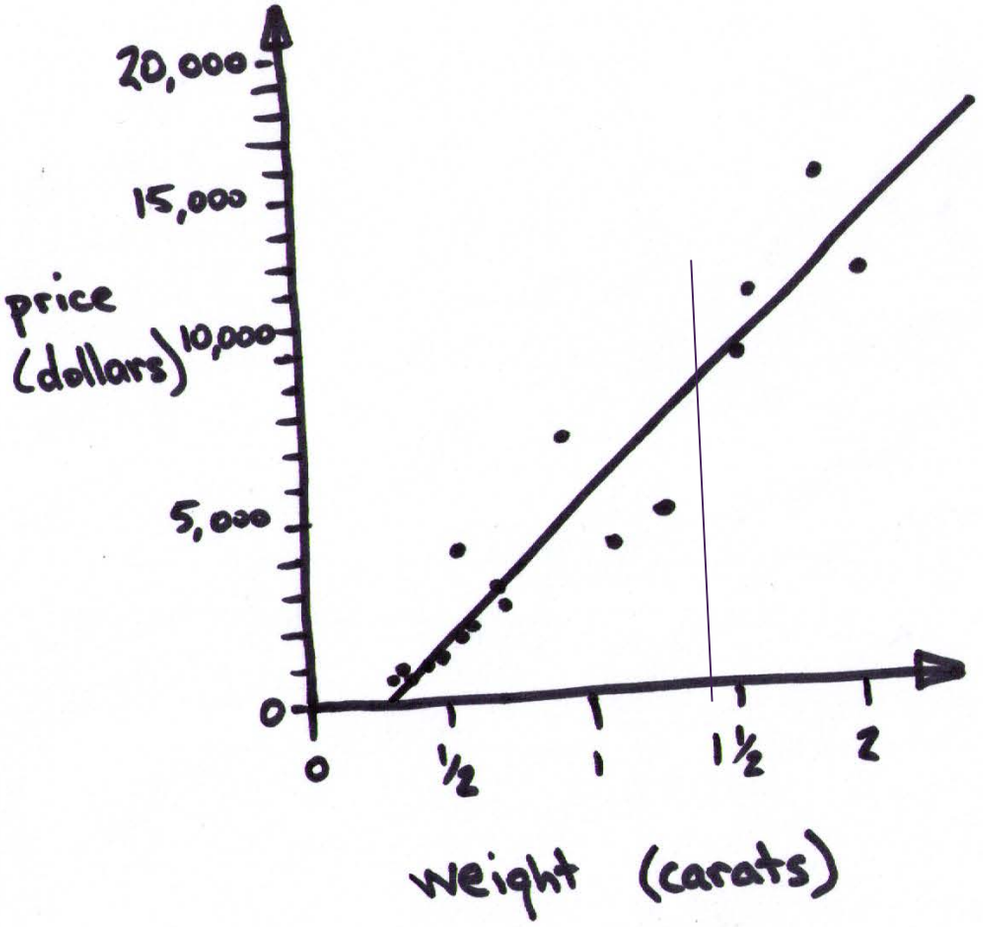
[modeling]
[linear regression]
[variance]
[noise]

carats	price
1.01	$7,366
.49	985
.31	544
1.51	9,140
.37	493
.73	3,011
1.53	11,413
.56	1,814
.41	876
.74	2,690
.63	1,991
.6	4,172
2.06	11,764
1.1	4,682
1.32	6,171
2.02	15,996
.34	695



Diamonds

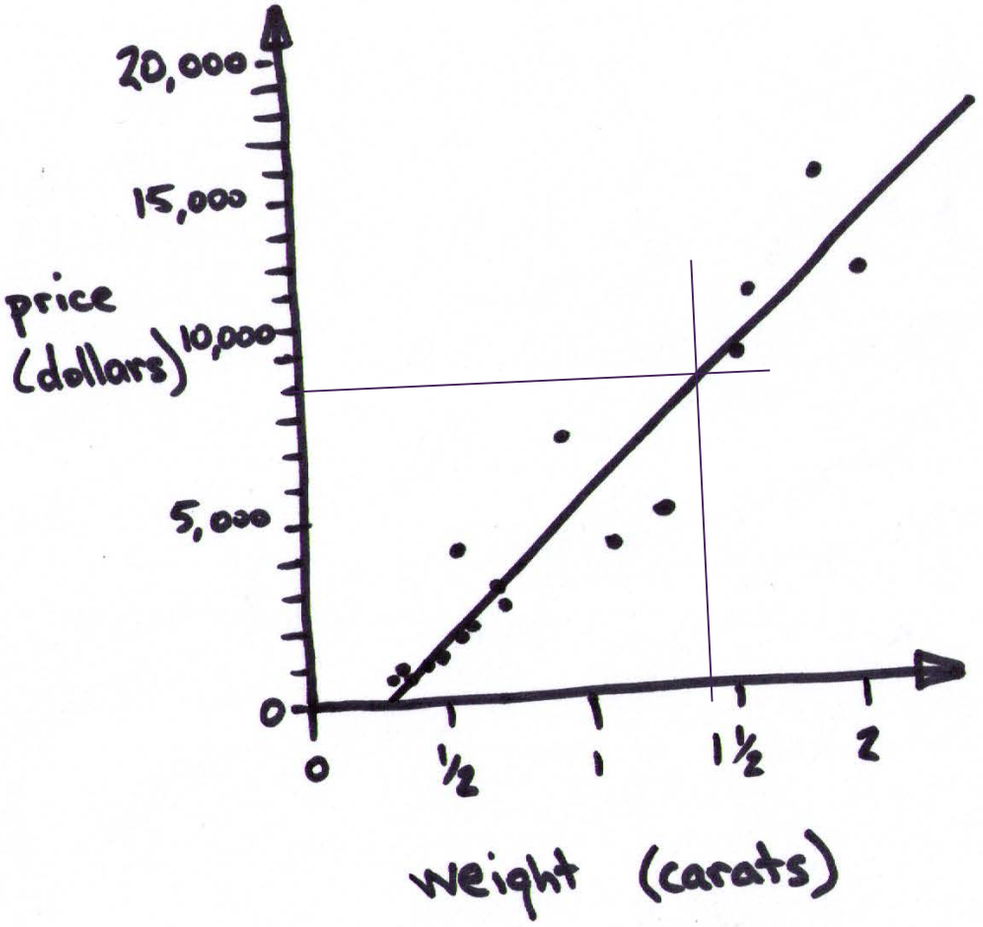
carats	price
1.01	$7,366
.49	985
.31	544
1.51	9,140
.37	493
.73	3,011
1.53	11,413
.56	1,814
.41	876
.74	2,690
.63	1,991
.6	4,172
2.06	11,764
1.1	4,682
1.32	6,171
2.02	15,996
.34	695



Diamonds

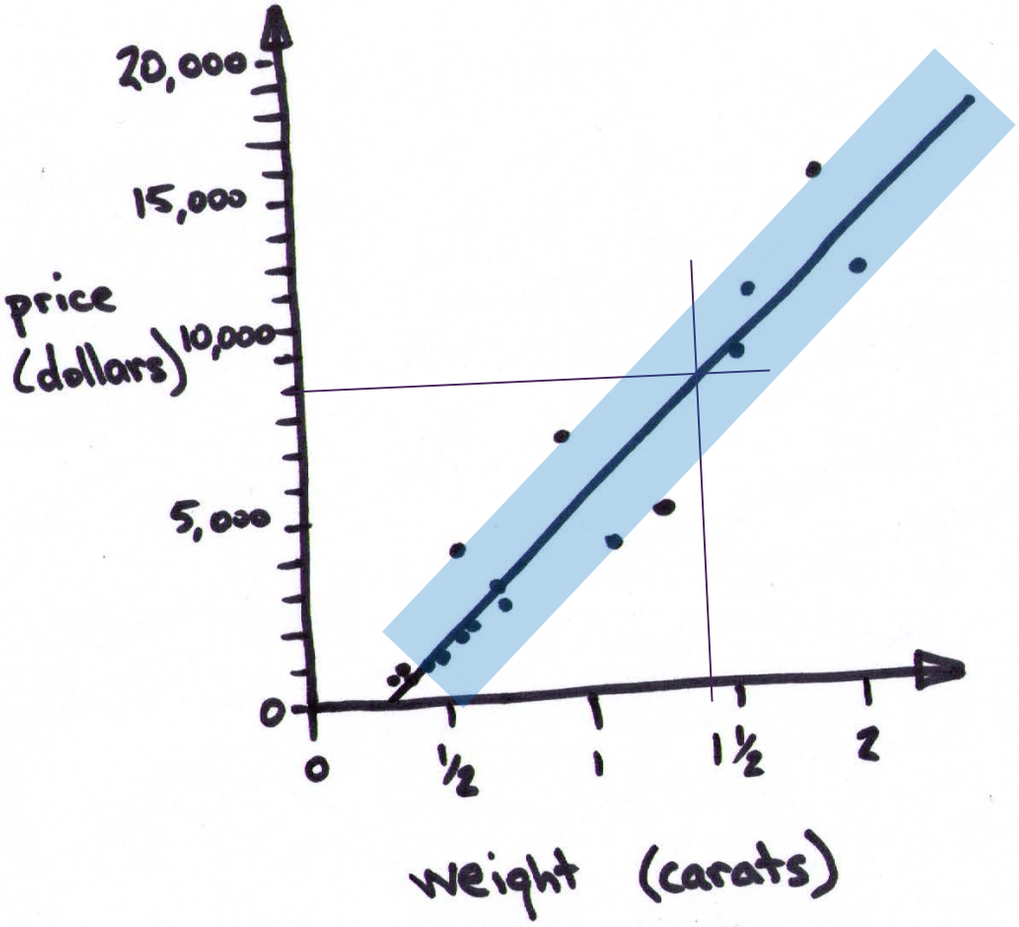
[prediction]

carats	price
1.01	$7,366
.49	985
.31	544
1.51	9,140
.37	493
.73	3,011
1.53	11,413
.56	1,814
.41	876
.74	2,690
.63	1,991
.6	4,172
2.06	11,764
1.1	4,682
1.32	6,171
2.02	15,996
.34	695



Diamonds

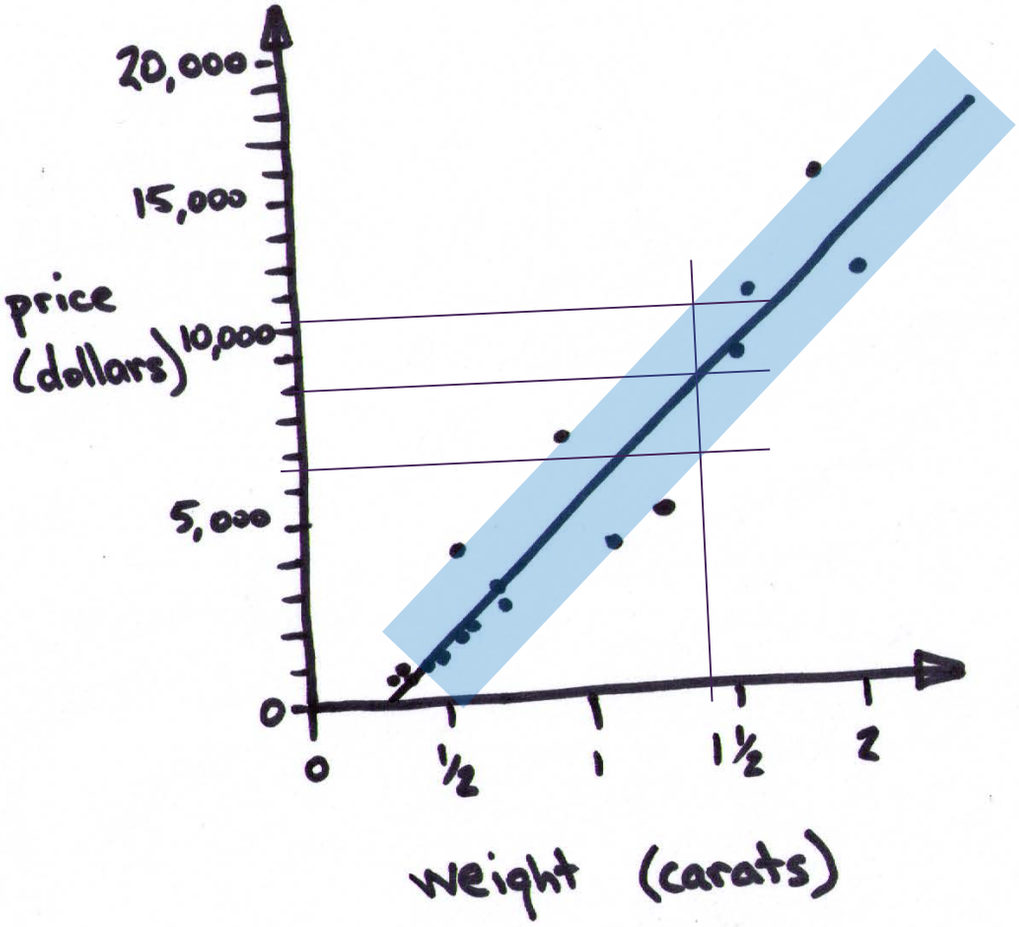
carats	price
1.01	$7,366
.49	985
.31	544
1.51	9,140
.37	493
.73	3,011
1.53	11,413
.56	1,814
.41	876
.74	2,690
.63	1,991
.6	4,172
2.06	11,764
1.1	4,682
1.32	6,171
2.02	15,996
.34	695



Diamonds

[confidence interval]

carats	price
1.01	$7,366
.49	985
.31	544
1.51	9,140
.37	493
.73	3,011
1.53	11,413
.56	1,814
.41	876
.74	2,690
.63	1,991
.6	4,172
2.06	11,764
1.1	4,682
1.32	6,171
2.02	15,996
.34	695



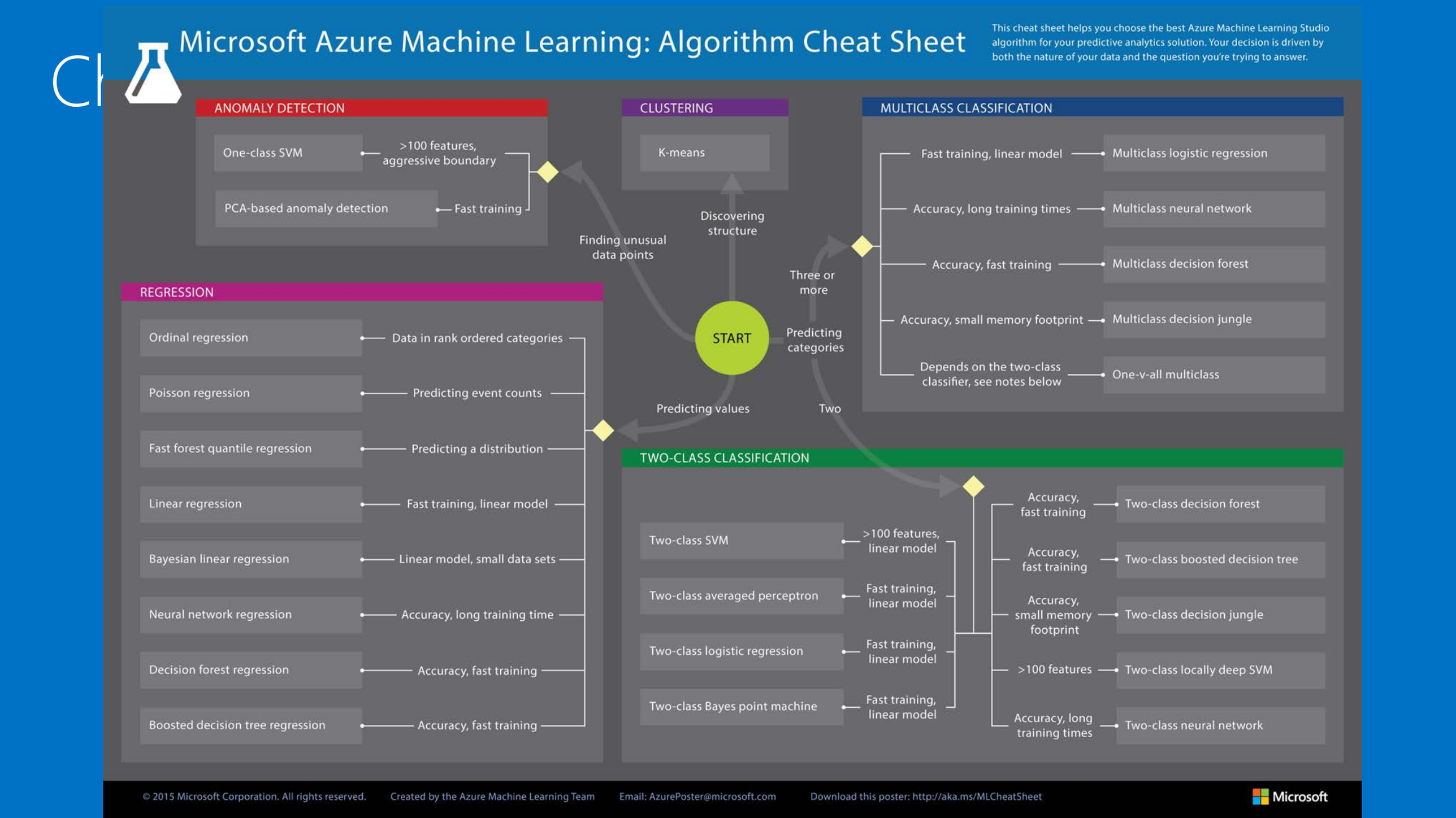
Congratulations!

We built a machine learning model.
We used linear regression.
We made a prediction, complete with a confidence interval.
(We didn't need math or computers.)

There are lots of right ways to do it.

Microsoft Azure Machine Learning: Algorithm Cheat Sheet

This cheat sheet helps you choose the best Azure Machine Learning Studio algorithm for your predictive analytics solution. Your decision is driven by both the nature of your data and the question you're trying to answer.



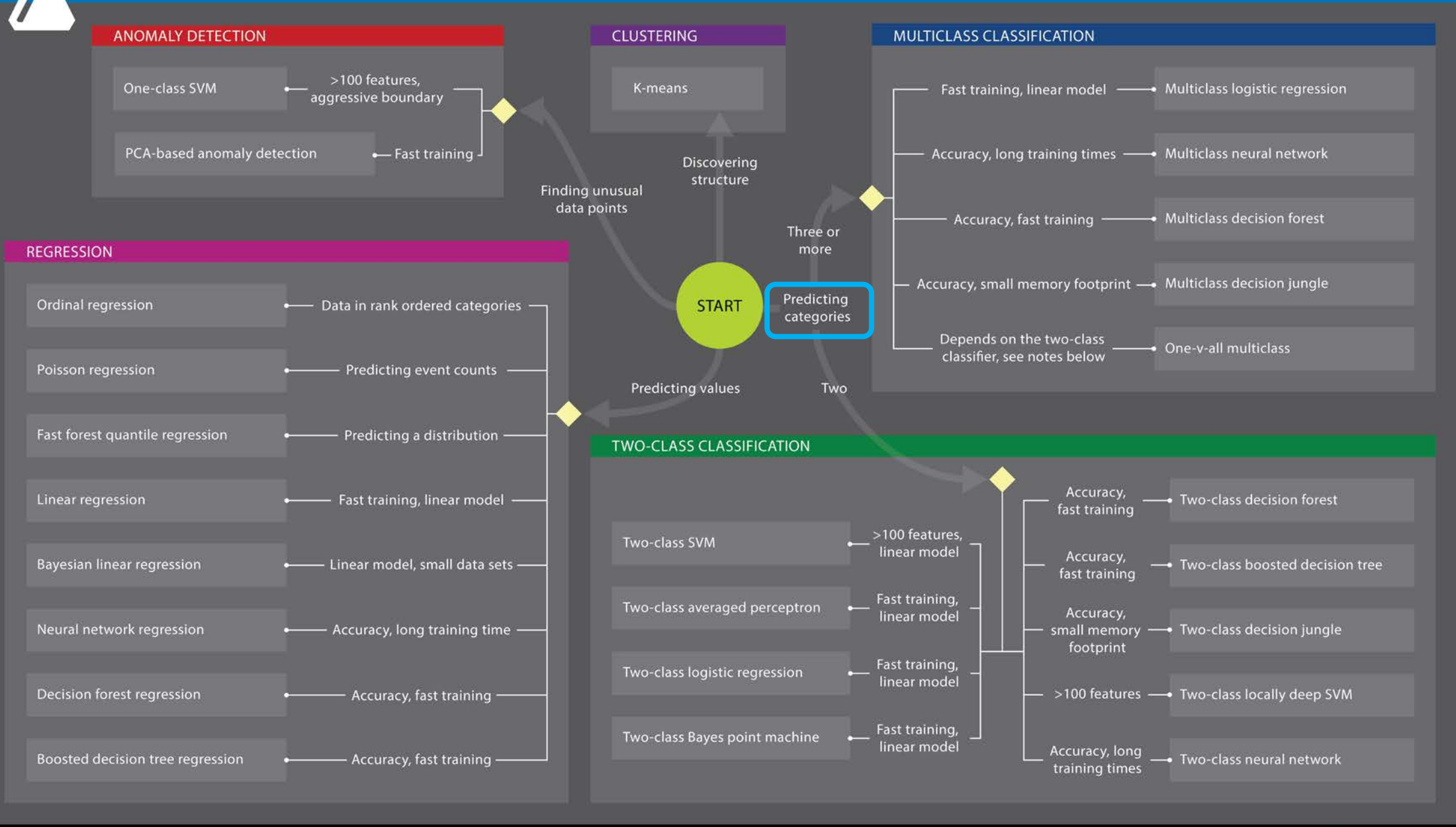
Created by the Azure Machine Learning Team Email: AzurePoster@microsoft.com

Download this poster: http://aka.ms/MLCheatSheet

Microsoft

Microsoft Azure Machine Learning: Algorithm Cheat Sheet

This cheat sheet helps you choose the best Azure Machine Learning Studio algorithm for your predictive analytics solution. Your decision is driven by both the nature of your data and the question you're trying to answer.



Created by the Azure Machine Learning Team

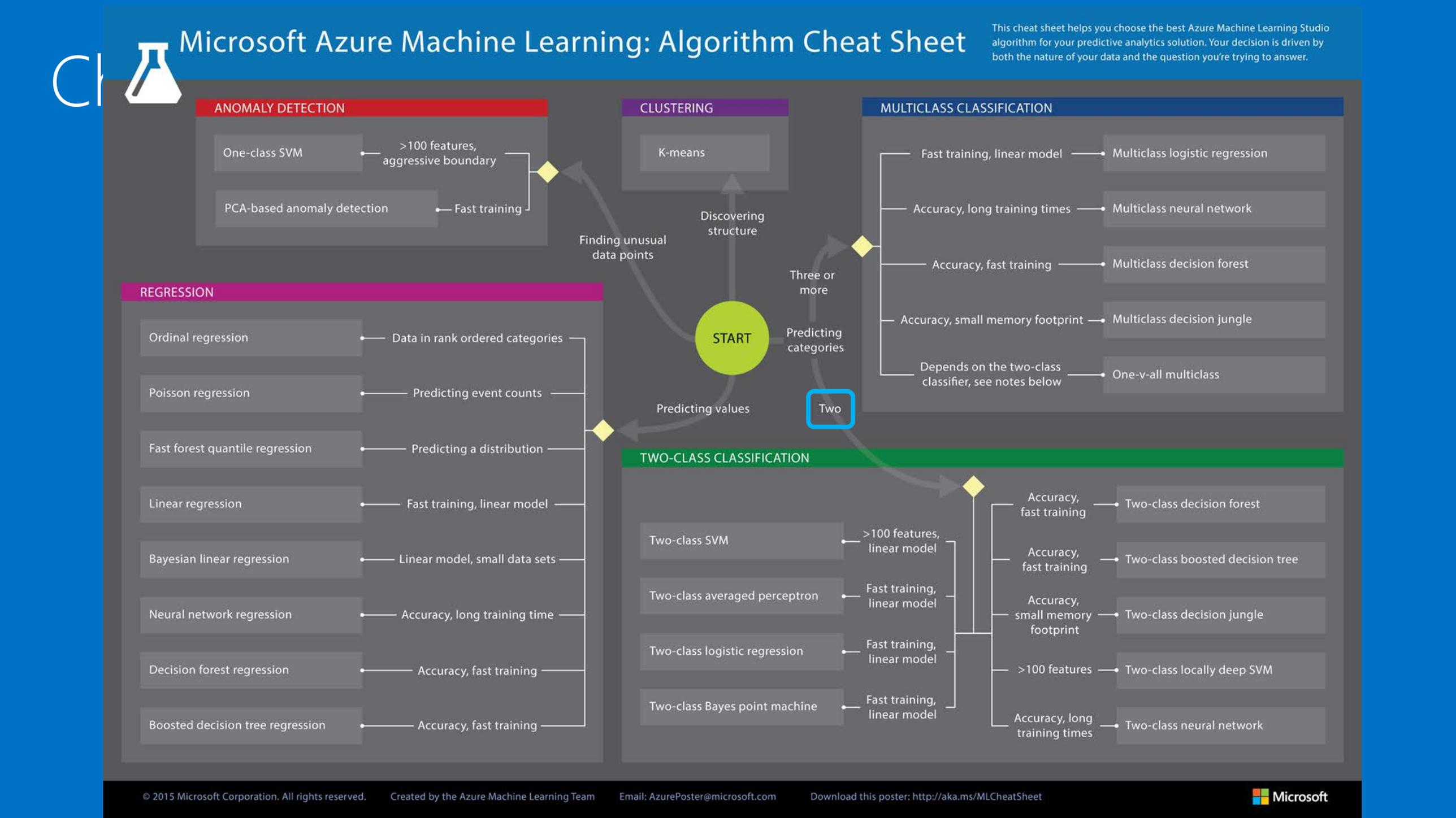
Email: AzurePoster@microsoft.com

Download this poster: http://aka.ms/MLCheatSheet

Microsoft

Microsoft Azure Machine Learning: Algorithm Cheat Sheet

This cheat sheet helps you choose the best Azure Machine Learning Studio algorithm for your predictive analytics solution. Your decision is driven by both the nature of your data and the question you're trying to answer.



Created by the Azure Machine Learning Team

Email: AzurePoster@microsoft.com

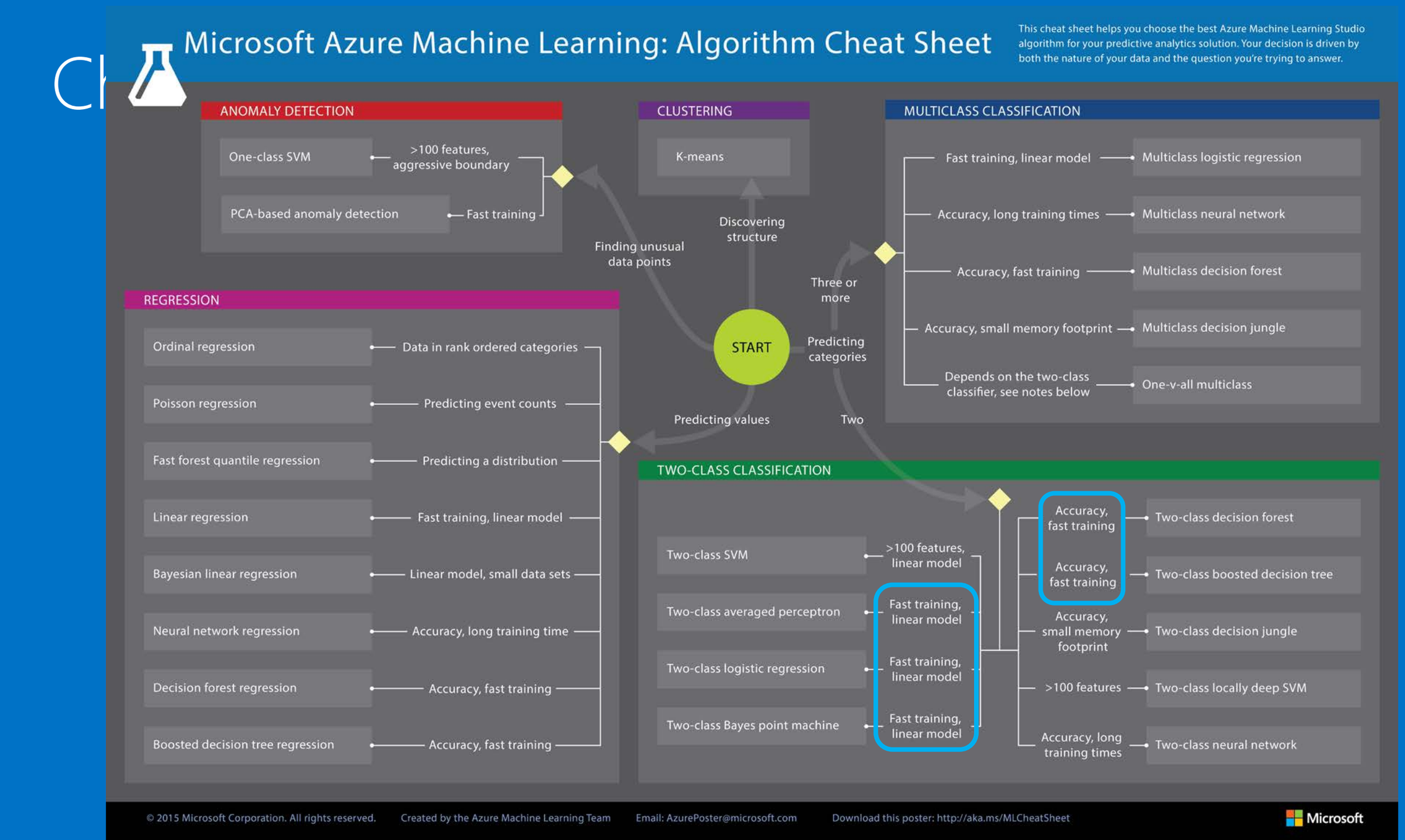
Download this poster: http://aka.ms/MLCheatSheet

Microsoft

Cl

Microsoft Azure Machine Learning: Algorithm Cheat Sheet

This cheat sheet helps you choose the best Azure Machine Learning Studio algorithm for your predictive analytics solution. Your decision is driven by both the nature of your data and the question you're trying to answer.

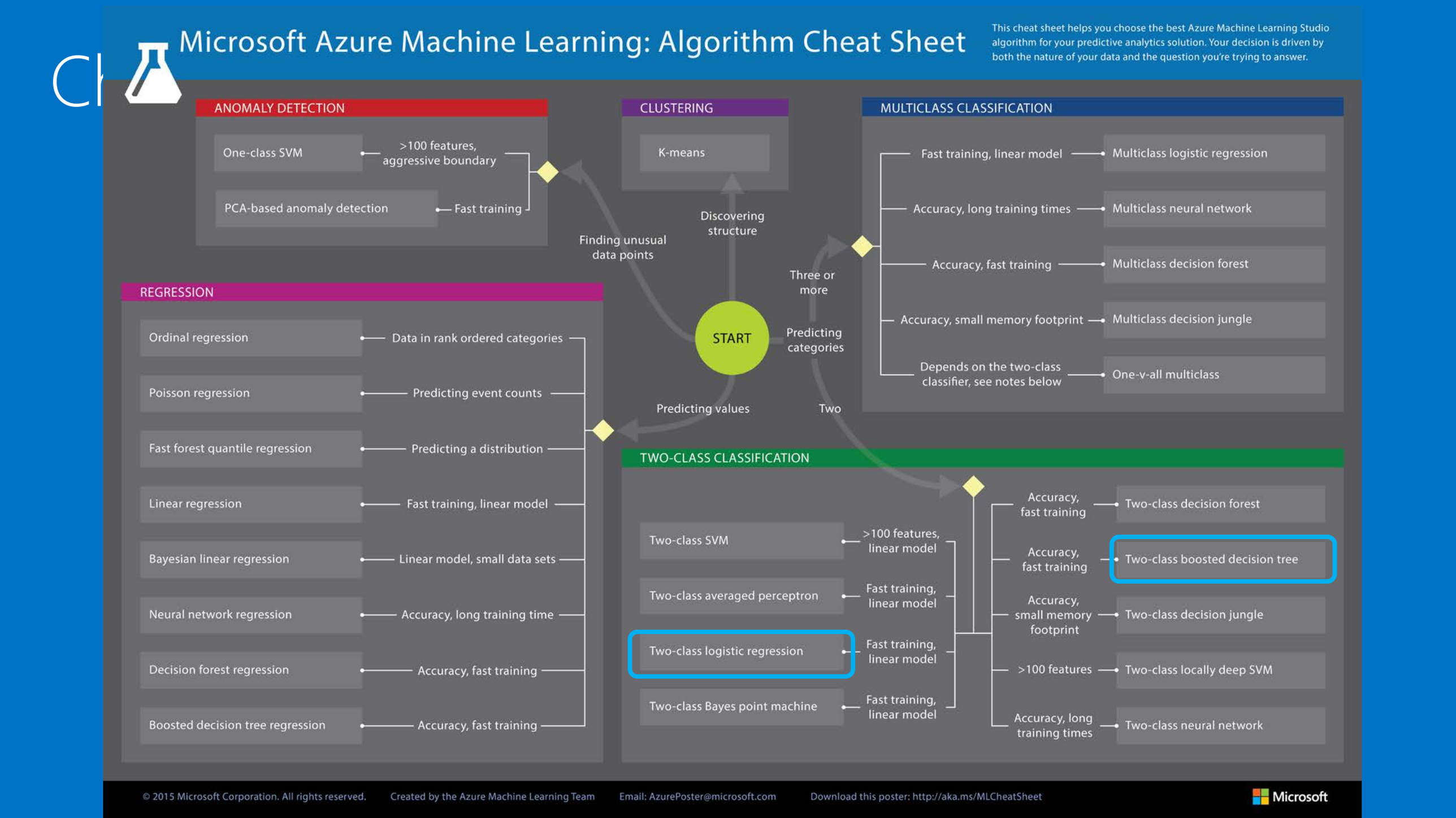


Created by the Azure Machine Learning Team Email: AzurePoster@microsoft.com Download this poster: http://aka.ms/MLCheatSheet

Microsoft

Microsoft Azure Machine Learning: Algorithm Cheat Sheet

This cheat sheet helps you choose the best Azure Machine Learning Studio algorithm for your predictive analytics solution. Your decision is driven by both the nature of your data and the question you're trying to answer.



Created by the Azure Machine Learning Team Email: AzurePoster@microsoft.com

Download this poster: http://aka.ms/MLCheatSheet

Microsoft

You can make other people do your work for you.

Gallery

Cortana Analytics Galler

gallery.cortanaanalytics.com

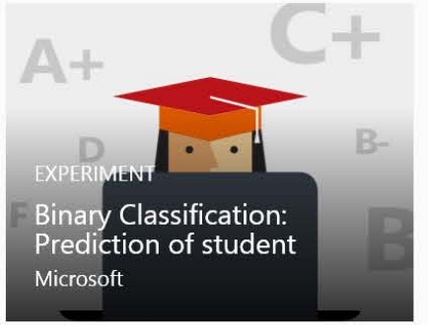
Cortana Analytics Gallery

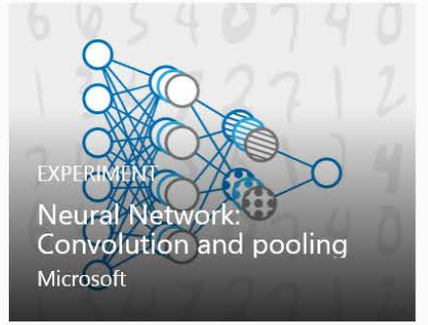
Search by name, algorithm or tags

Sign in

Browse all | Experiments | Machine Learning APIs | Tutorials | Collections

Cortana Analytics Gallery enables our growing community of developers and data scientists to share their analytics solutions. Learn how to contribute.



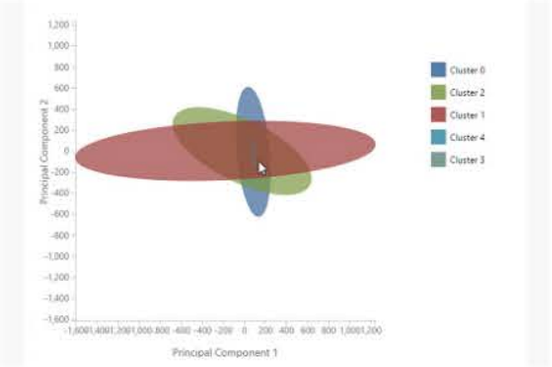


Recently added

See all

EXPERIMENT

Clustering sweep: diabetes dataset



This experiment demonstrates how to use the new Sweep Clustering module to select the best number of centroids an...

K-Means Clustering

154 87 6 days ago

Jeannine Takaki

EXPERIMENT

HARMAN ANALYTICS : Promotion Effectiveness

This collection of experiment demonstrates a model to estimate sales lift due to promotion in retail stores acr...

Linear Regression

39 4 yesterday

Shailendra Kumar

EXPERIMENT

HARMAN ANALYTICS: F3 Media Churn Analysis NN ...

Churn model using NN and SVM for 'F3: Fall Footbal Fanatica' app with demographic breakdown for customer...

Two-Class Neural Network, Two-Class Support Vector Machine

152 57 13 days ago

Dennis Sprous

EXPERIMENT

Pass Summit 2015 Data Storytelling with R AzureM...

How can we use technology to help the organization make data-driven decision-making part of its organizational DNA, ...

Two-Class Logistic Regression

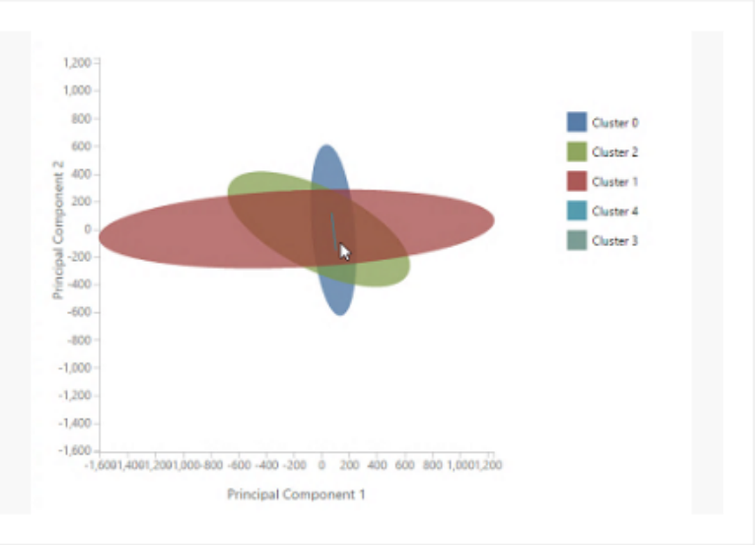
82 61 7 days ago

Jen Stirrup

EXPERIMENT

Clustering sweep: diabetes dataset

Jeannine Takaki • published on October 28, 2015



Open in Studio

+ Add to Collection

154 views

87 downloads

Tweet Share

ALGORITHMS

K-Means Clustering

Summary

This experiment demonstrates how to use the new Sweep Clustering module to select the best number of centroids and optimize other parameters

Description

Summary

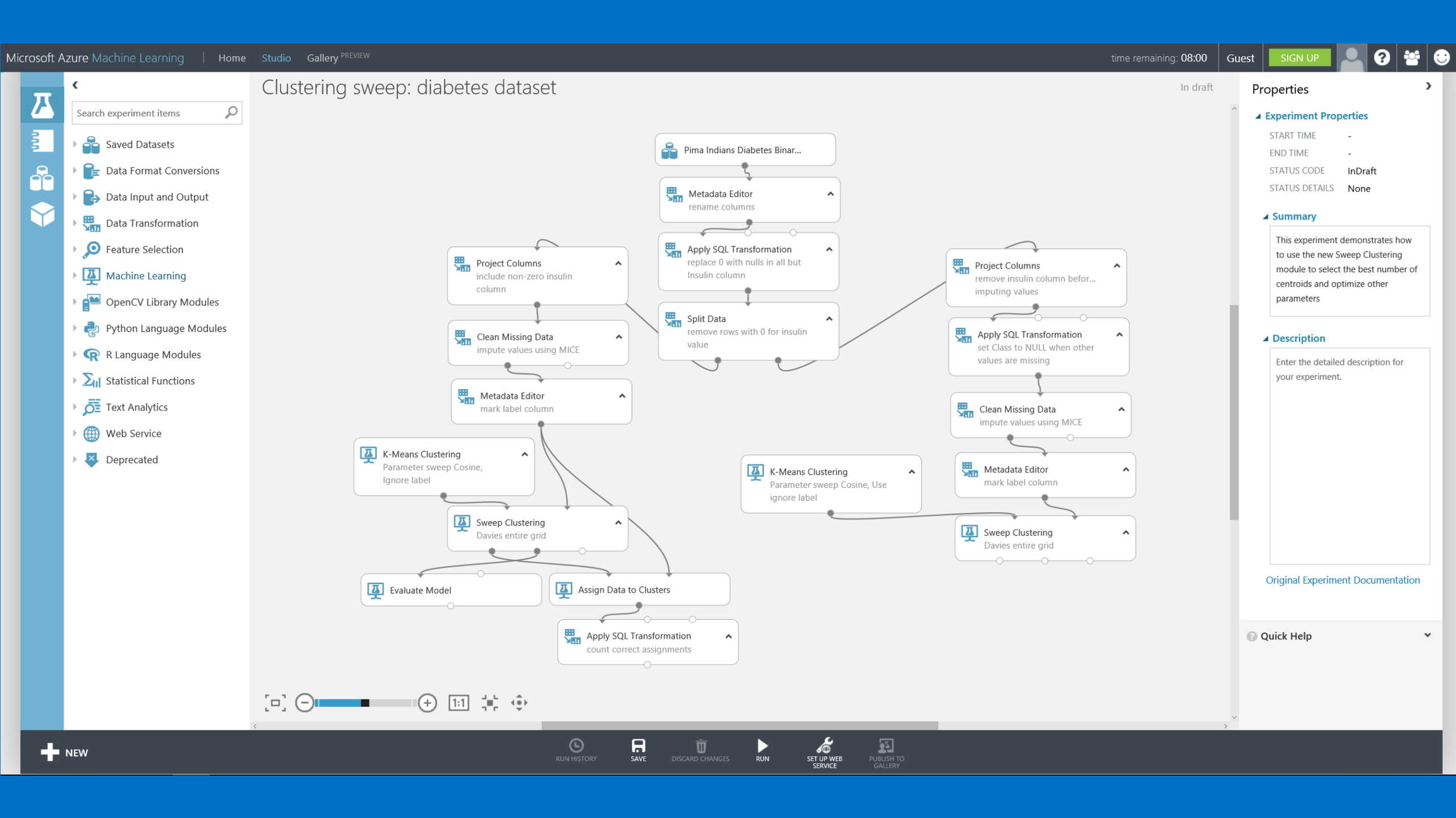
This experiment uses a parameter sweep with the K-means clustering algorithm to select the best number of clusters and the best initial centroid, based on a clustering metric you select. In the experiment, the original dataset is divided into two parts, to handle missing data and to compare the effect of variables on clustering. The results of the models are compared using a principal components graph. The experiment also demonstrates how you can use the Sweep Clustering module to fill in values for a label column.

Understanding the Data

The dataset contains 768 rows, from a larger dataset that studies the incidence of diabetes among different populations.

The following clinical values are included, which are often used in diagnosing diabetes.

Triceps skin fold measurements (TSF) and **body mass index (BMI)** are thought to be correlated with


Microsoft Azure Machine Learning
Home
Studio
Gallery PREVIEW
time remaining: 08:00
Guest
SIGN UP
Clustering sweep: diabetes dataset
In draft
Search experiment items
Saved Datasets
Data Format Conversions
Data Input and Output
Data Transformation
Feature Selection
Machine Learning
OpenCV Library Modules
Python Language Modules
R Language Modules
Statistical Functions
Text Analytics
Web Service
Deprecated
Pima Indians Diabetes Binar...
Metadata Editor
rename columns
Apply SQL Transformation
replace 0 with nulls in all but Insulin column
Project Columns
include non-zero insulin column
Split Data
remove rows with 0 for insulin value
Clean Missing Data
impute values using MICE
Metadata Editor
mark label column
K-Means Clustering
Parameter sweep Cosine, Ignore label
Sweep Clustering
Davies entire grid
Evaluate Model
Assign Data to Clusters
Apply SQL Transformation
count correct assignments
Project Columns
remove insulin column befor... imputing values
Apply SQL Transformation
set Class to NULL when other values are missing
Clean Missing Data
impute values using MICE
Metadata Editor
mark label column
K-Means Clustering
Parameter sweep Cosine, Use ignore label
Sweep Clustering
Davies entire grid
1:1
Properties
Experiment Properties
START TIME -
END TIME -
STATUS CODE InDraft
STATUS DETAILS None
Summary
This experiment demonstrates how to use the new Sweep Clustering module to select the best number of centroids and optimize other parameters
Description
Enter the detailed description for your experiment.
Original Experiment Documentation
Quick Help
NEW
RUN HISTORY
SAVE
DISCARD CHANGES
RUN
SET UP WEB SERVICE
PUBLISH TO GALLERY

EXPERIMENT

Methods for handling missing values

Brandon Rohrer • published on September 28, 2015

edit

Summary

This experiment illustrates a variety methods for handling missing data on a sample data set.

Description

Real world data is usually missing values, which trip up a lot of machine learning algorithms. There are lots of tricks for dealing with these, but you have to be careful. The way in which you fill them can change the result dramatically. Being explicit and thoughtful about how you handle missing values will get you the very best results.

I've illustrated a large handful of approaches to missing values here in a fake data set. The data shows a group of employees, some of their personal data, and some data regarding an upcoming office party. In every case, knowing what the data means is the most important part of handling it well.

Column 0	age	years_seniority	income	parking_space	attending_party	entree	pets	emergency_contact
Tony	48	27		1	5	shrimp		Pepper
Donald	67	25	86	10	2	beef		Jane
Henry	69	21	95	6	1	chicken	62	Janet
Janet	62	21	110	3	1	beef		Henry
Nick		17		4				
Bruce	37	14	63		1	veggie		NA
Steve	83		77	7	1	chicken		n/a
Clint	27	9	118	9		shrimp	3	None
Wanda	19	7	52	2	2	shrimp		empty
Natasha	26	4	162	5	3			-
Carol		3	127	11	1	veggie	1	""
Mandy	44	2	68	8	1	chicken		null

Open in Studio

Added to Collection

112 views

9 downloads

Tweet Share

TAGS

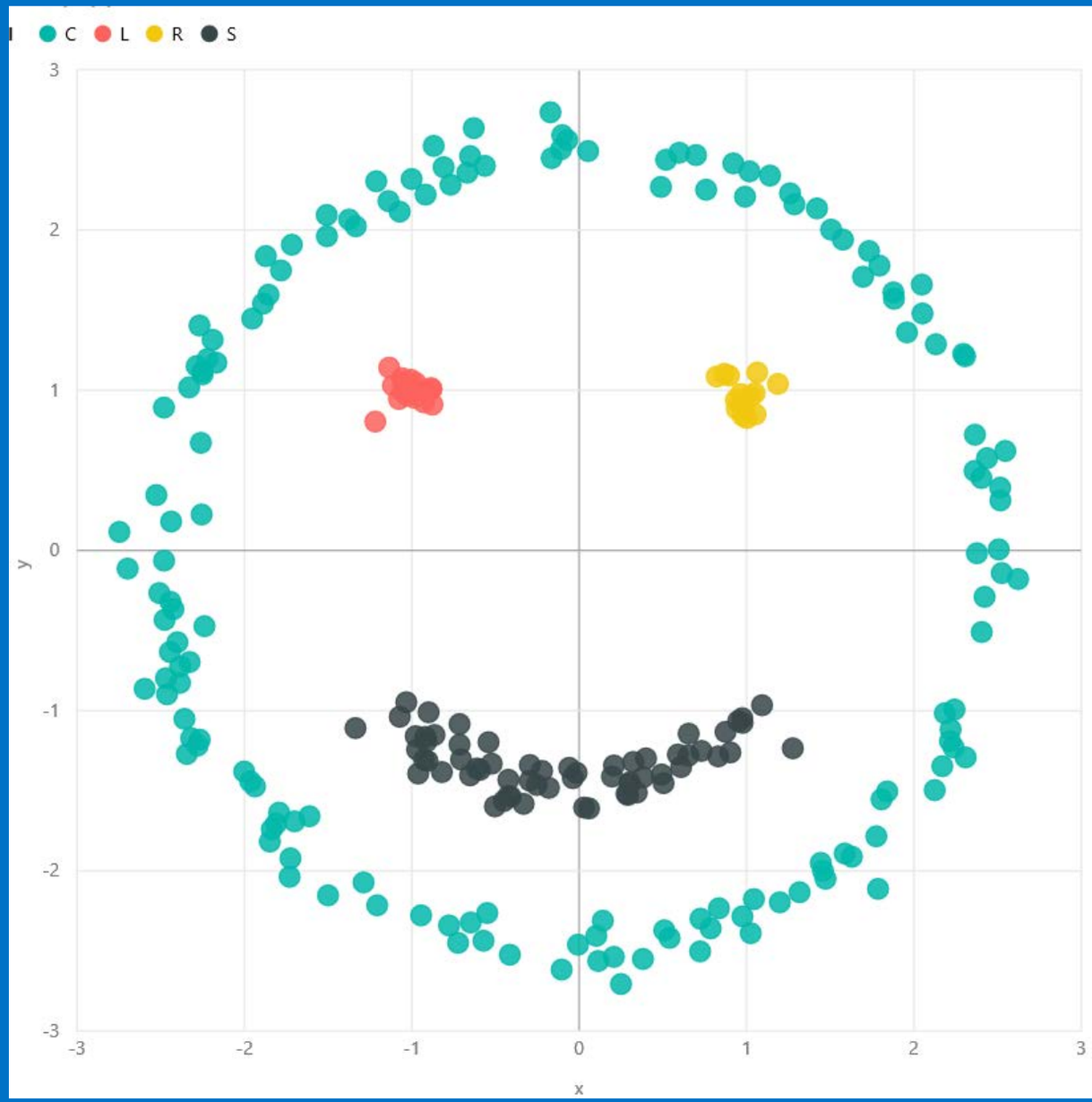
Secrets of data science revealed

Secrets of data science revealed

1. You can’t use just any data.

Secrets of data science revealed

1. You can’t use just any data.
2. Turn your data into a picture.



Secrets of data science revealed

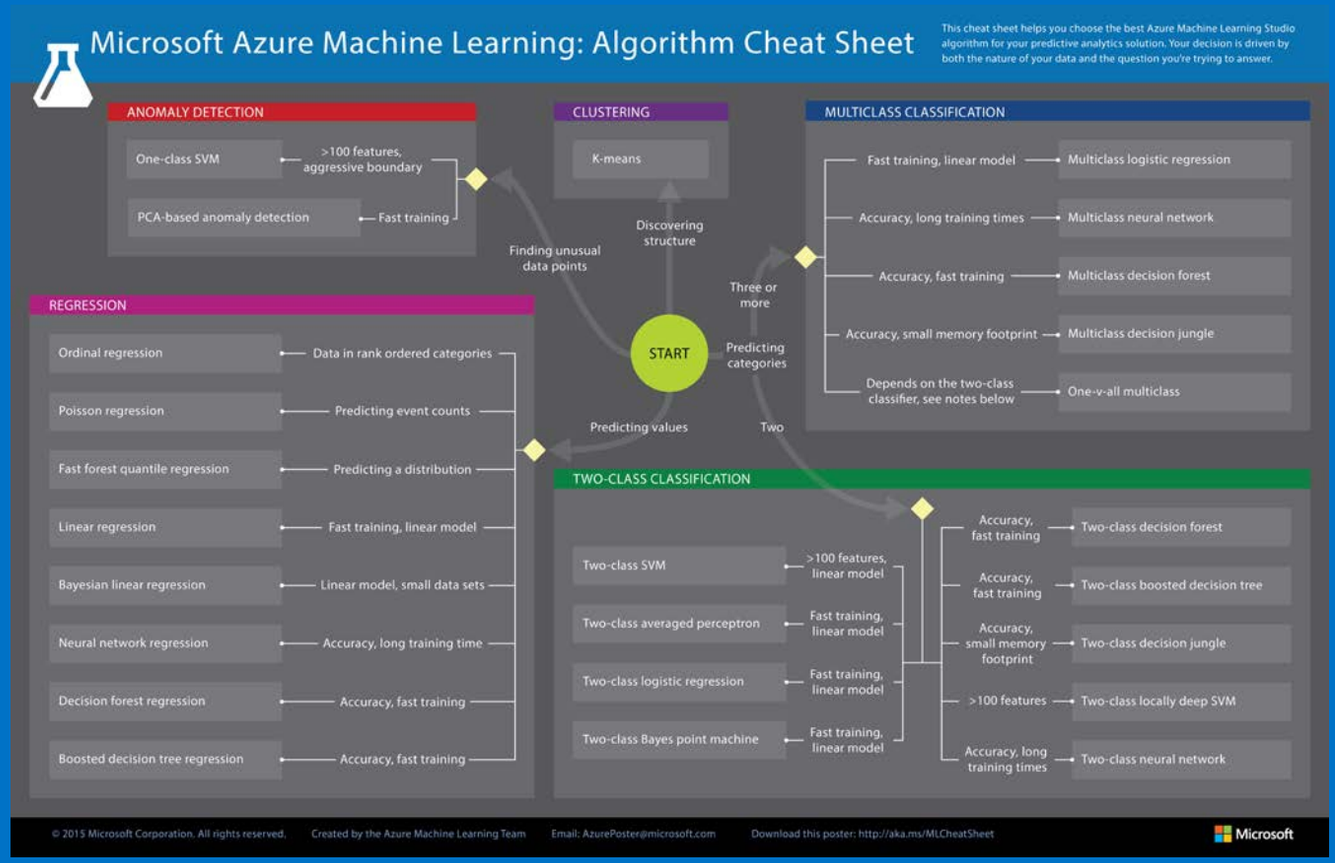
1. You can’t use just any data.
2. Turn your data into a picture.
3. Data science can only answer five questions.

Secrets of data science revealed

1. You can’t use just any data.
2. Turn your data into a picture.
3. Data science can only answer five questions.
4. Machine learning is simple.

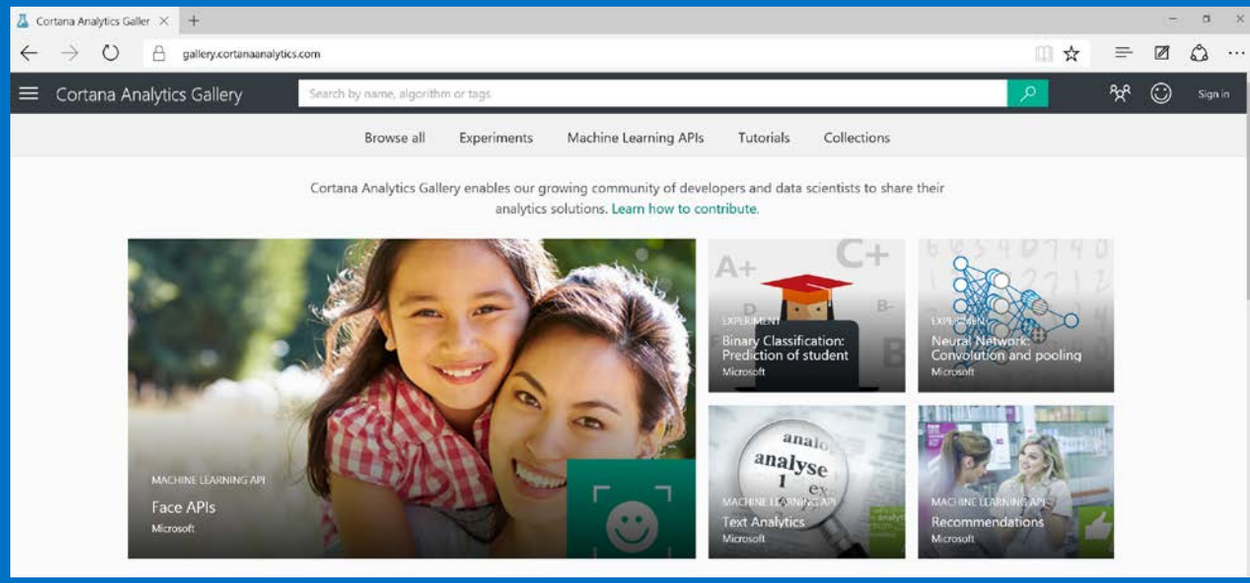
Secrets of data science revealed

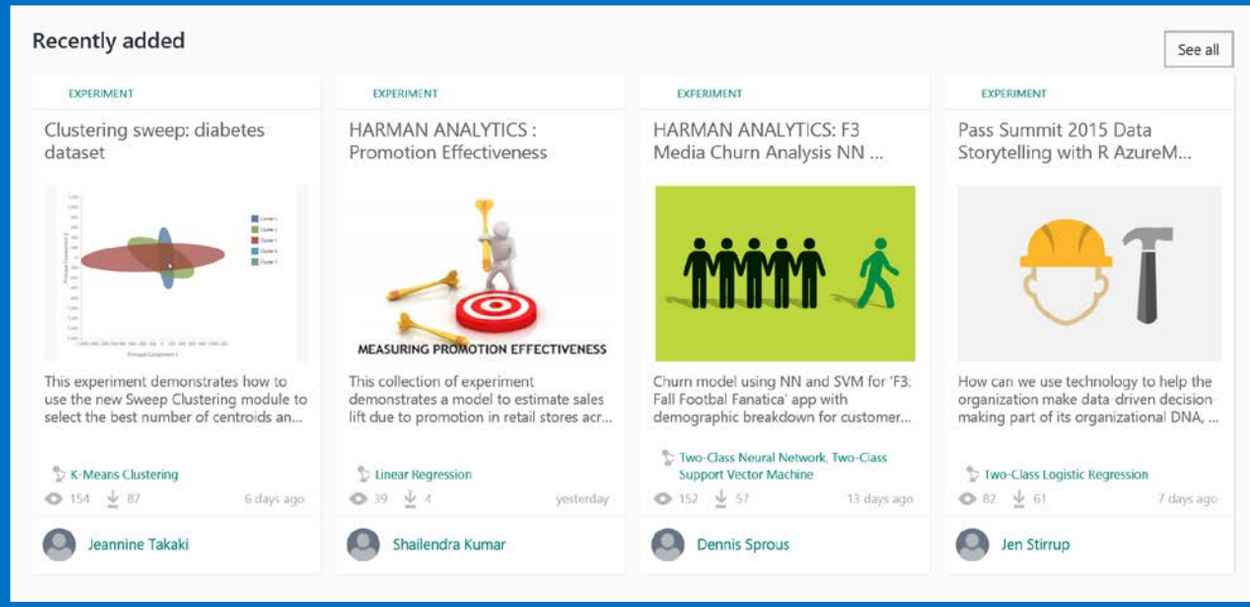
1. You can't use just any data.
2. Turn your data into a picture.
3. Data science can only answer five questions.
4. Machine learning is simple.
5. There are lots of right ways to do it.



Secrets of data science revealed

1. You can’t use just any data.
2. Turn your data into a picture.
3. Data science can only answer five questions.
4. Machine learning is simple.
5. There are lots of right ways to do it.
6. You can make other people do your work for you.





Resources

1. You can't use just any data.	http://blogs.technet.com/b/machinelearning/archive/2015/08/26/what-can-data-science-do-for-me.aspx
2. Turn your data into a picture.	https://gallery.cortanaanalytics.com/Experiment/Data-exploration-through-visualization-1
3. Data science can only answer five questions.	http://blogs.technet.com/b/machinelearning/archive/2015/08/27/what-types-of-questions-can-data-science-answer.aspx http://blogs.technet.com/b/machinelearning/archive/2015/09/01/which-algorithm-family-can-answer-my-question.aspx
4. Machine learning is simple.	Unfortunately, you will have to learn math and a computer language to do this one.
5. There are lots of right ways to do it.	https://azure.microsoft.com/en-us/documentation/articles/machine-learning-algorithm-cheat-sheet/ https://azure.microsoft.com/en-us/documentation/articles/machine-learning-algorithm-choice/
6. You can make other people do your work for you.	https://gallery.cortanaanalytics.com/

Thanks for listening!

Thoughts to share? Questions?
Connect with me offline:
Brandon Rohrer on LinkedIn
@_brohrer_ on Twitter
brohrer@microsoft.com

Special thanks to Diane Rohrer for image and layout design.

Microsoft