Microsoft Azure Machine Learning

INSTRUCTOR INFORMATION

Loy Vanich
084 007 5544
Line ID: laployv
laploy@gmaill.com
www.laploy.com

REPOSITORY AND NOTE SHARE

Repository
github.com/laploy/ML

Note Share
gist.github.com/laploy
(Azure ML Note for student)

INTRODUCTION TO AZURE ML

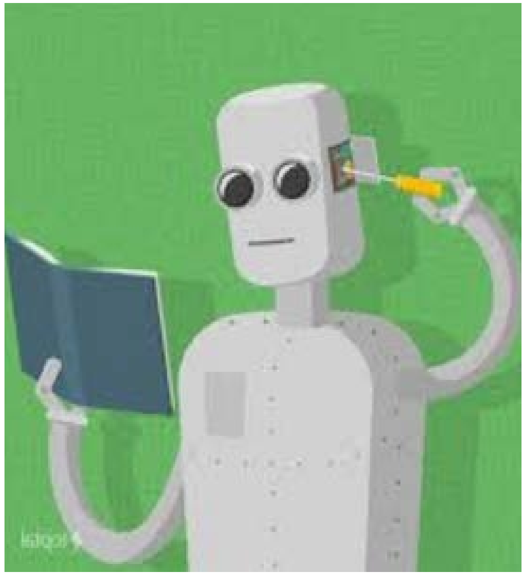
What Bill have to say?

“A breakthrough in machine learning would be worth ten Microsofts,”

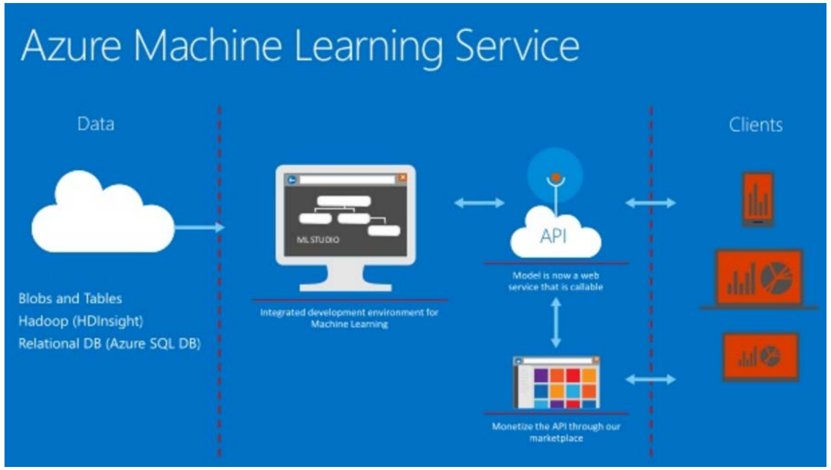
Bill Gates

What is Machine Learning?

- Subfield of computer science
- Ability to learn without being explicitly programmed
- Algorithms that can learn from and make predictions on data
- Closely related to computational statistics
- Focuses on prediction-making
- Sometimes mixed with Data Mining (DM)
- Used complex models and algorithms
- Predictive analytics



What is Azure ML?


Azure Machine Learning Service
Data
Blobs and Tables
Hadoop (HDInsight)
Relational DB (Azure SQL DB)
ML STUDIO
Integrated development environment for Machine Learning
API
Model is now a web service that is callable
Monetize the API through our marketplace
Clients

- Pronounce = Air-Cher
- Part of Azure Could platform
- Part of Cortana Intelligence Suite
- Platform as a service (PaaS)
- Create systems that improve with experience
- Help turning data into software
- Help training with huge volumes of data
- Used to predict certain patterns, trends, and outcomes
- Predictive analytics is the underlying technology
- Use the past to predict the future

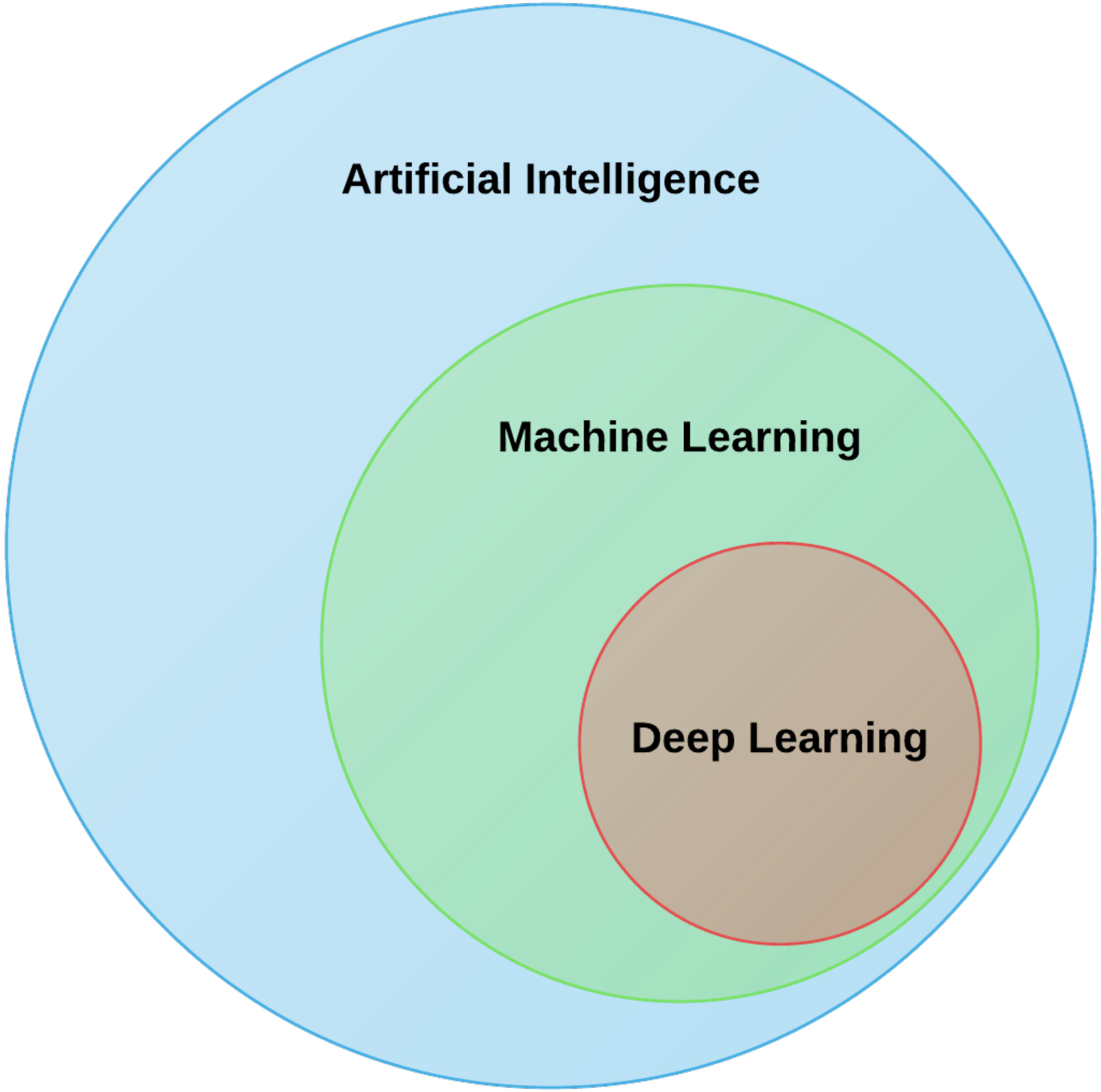
What is Azure ML Studio?

Azure Machine Learning Studio

What is Azure ML Studio?

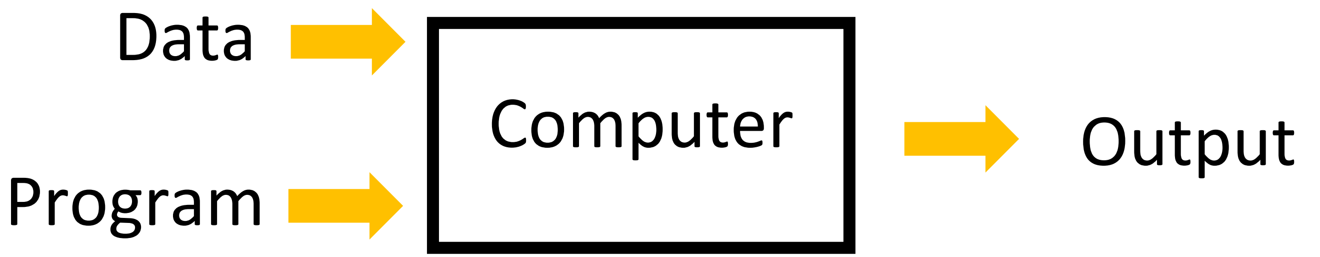
- Is a collaborative tool
- Drag-and-drop
- Use to build, test, and deploy predictive analytics solutions
- Publishes models as web services
- Easily be consumed by custom apps or BI tools
- Is where data science, predictive analytics, cloud resources, and your data meet.

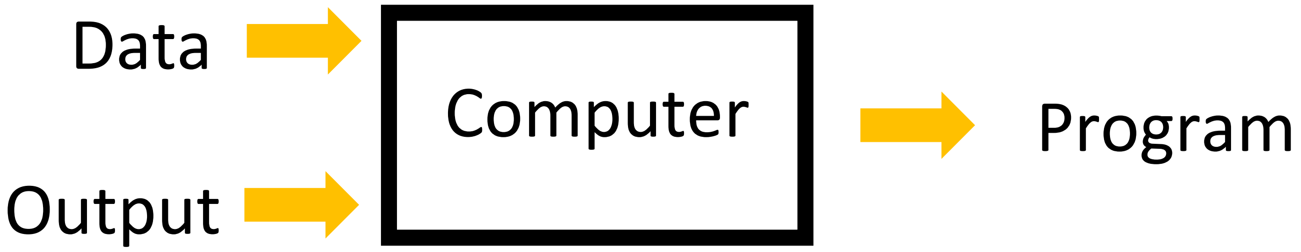
Artificial Intelligence VS AI



ML paradigm

Traditional Programming



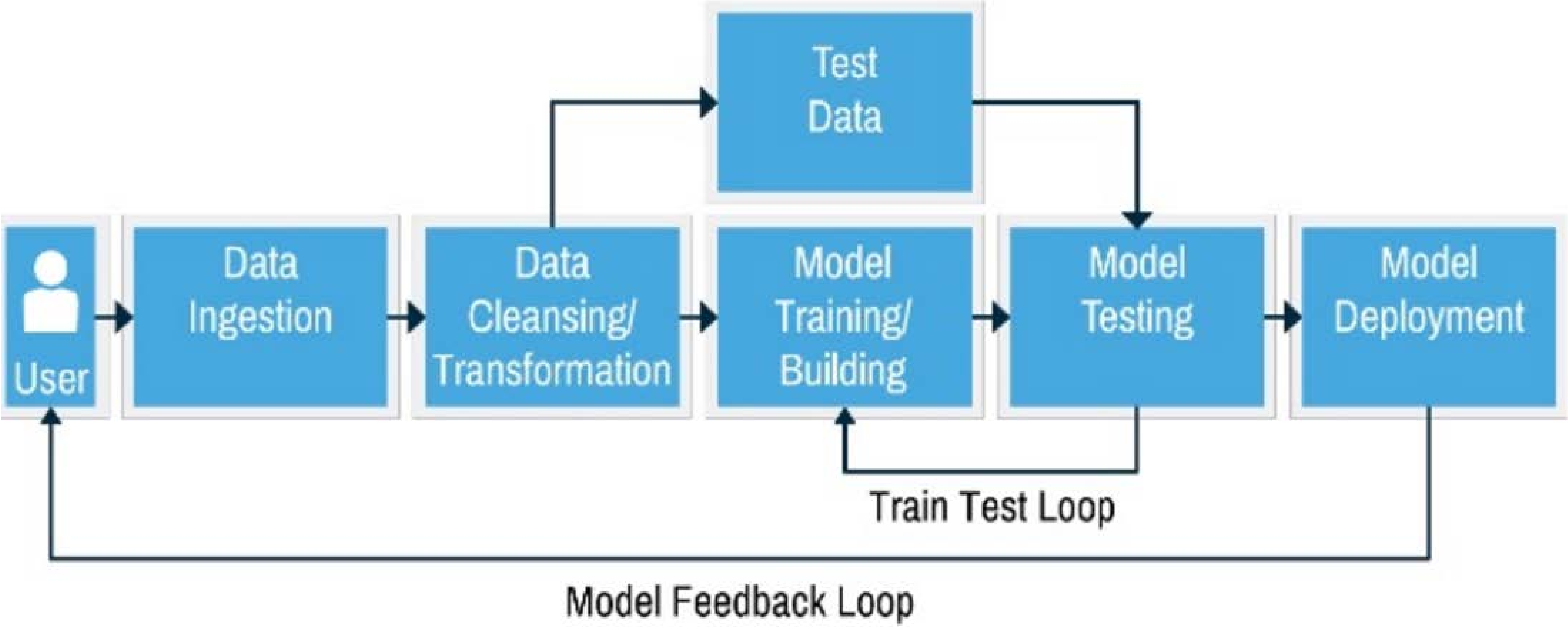


Everyday examples of predictive analytics

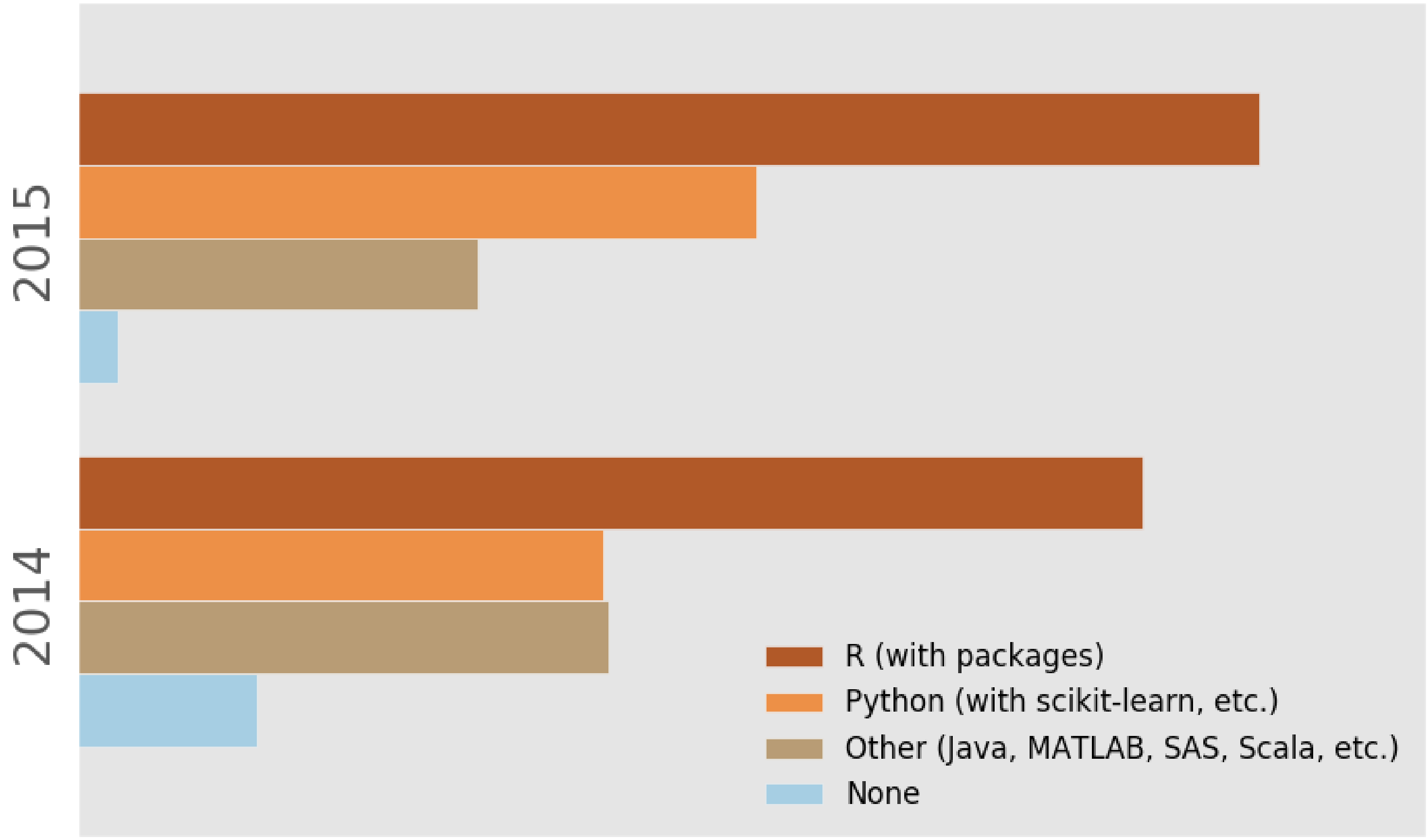
Everyday examples of predictive analytics

- Spam/junk email filters
- Mortgage applications
- Pattern recognition
- Life insurance
- Medical insurance
- Liability/property insurance
- Credit card fraud detection
- Airline flights
- Web search
- Predictive maintenance
- Health care

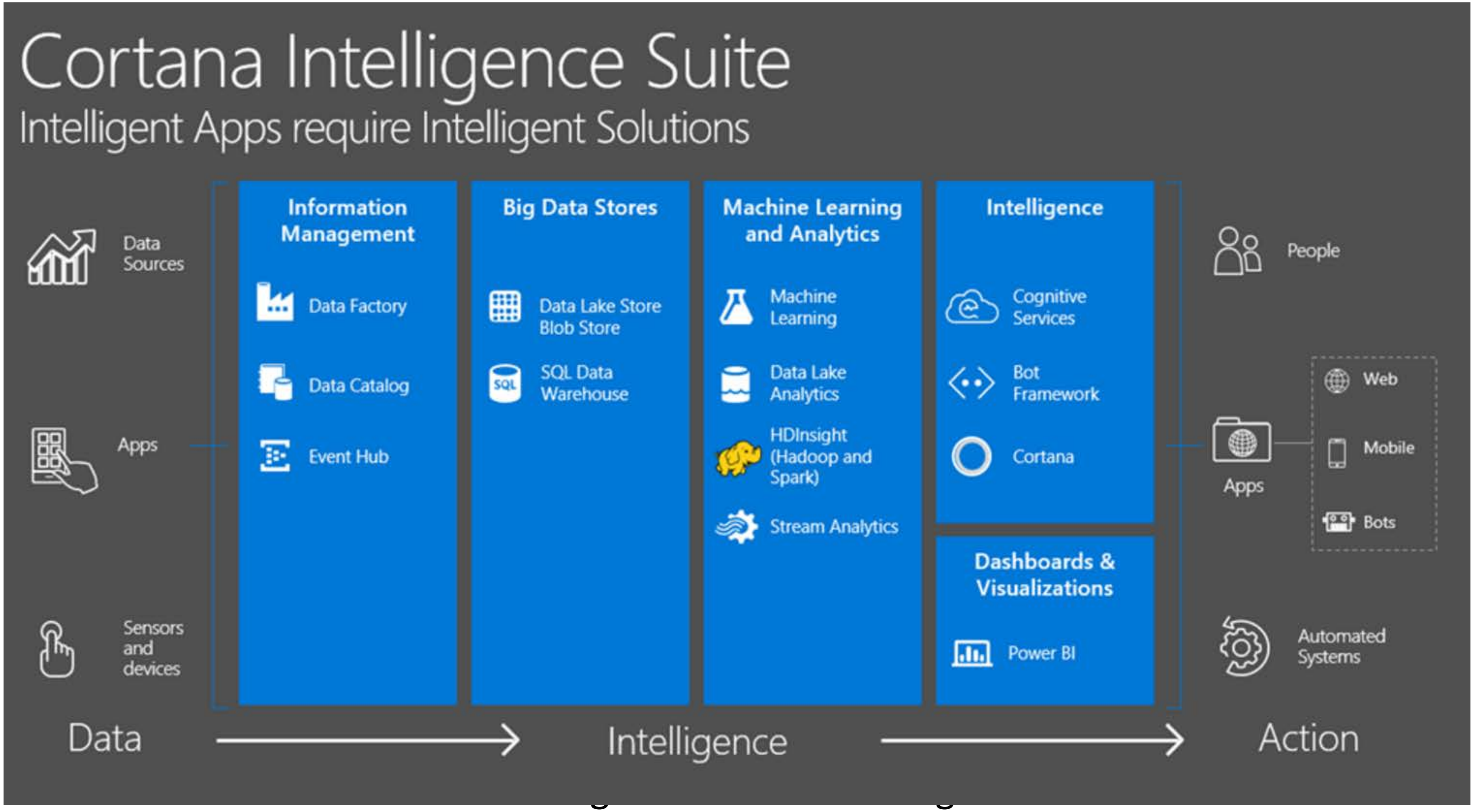
Machine Learning work flow



Machine learning language



Cortana Intelligence Suite (CIS)



Cortana Intelligence Suite

Intelligent Apps require Intelligent Solutions

Data Sources

Apps

Sensors and devices

Information Management

Data Factory

Data Catalog

Event Hub

Big Data Stores

Data Lake Store Blob Store

SQL Data Warehouse

Machine Learning and Analytics

Machine Learning

Data Lake Analytics

HDInsight (Hadoop and Spark)

Stream Analytics

Intelligence

Cognitive Services

Bot Framework

Cortana

Dashboards & Visualizations

Power BI

People

Apps

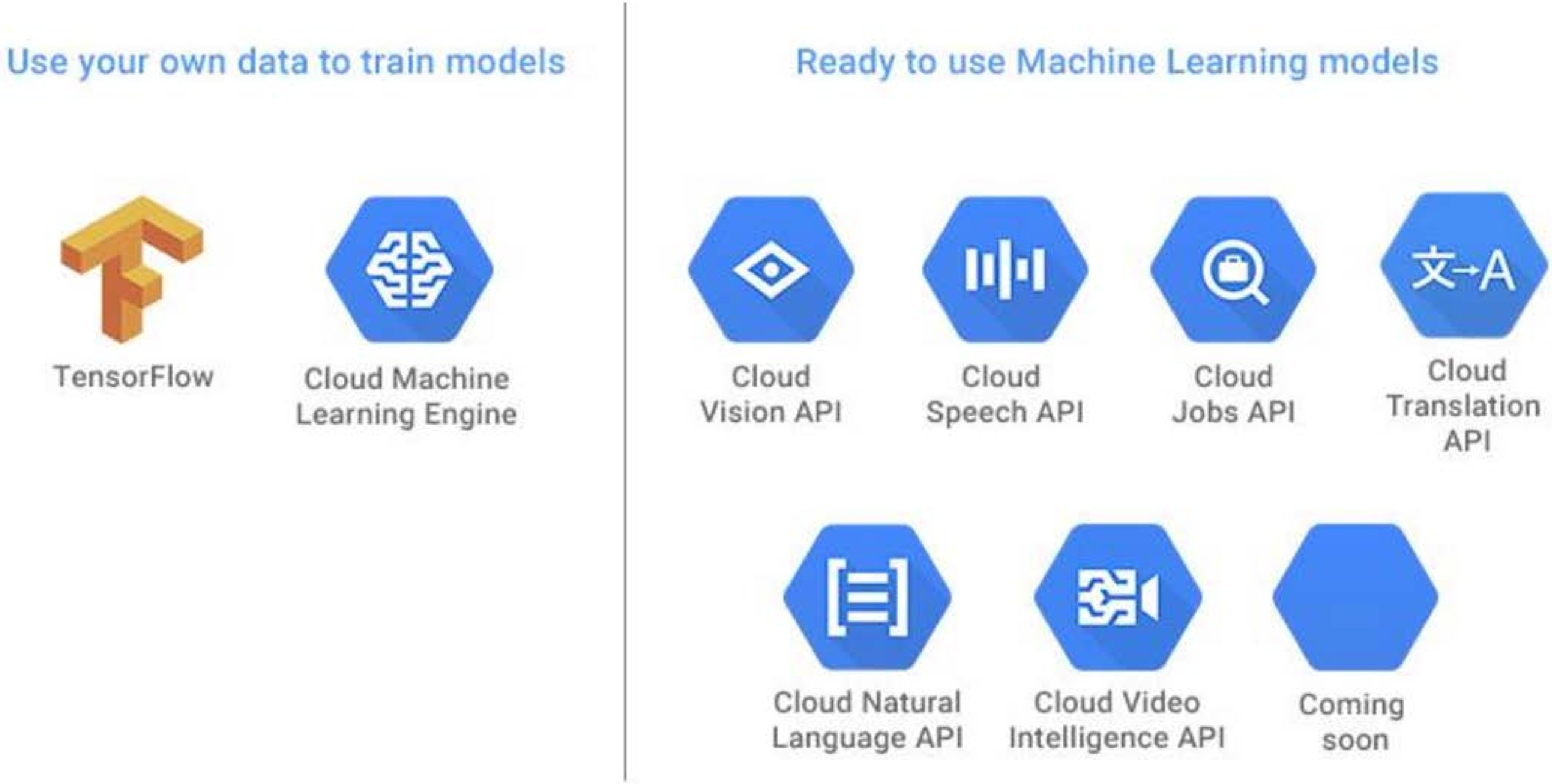
Web

Mobile

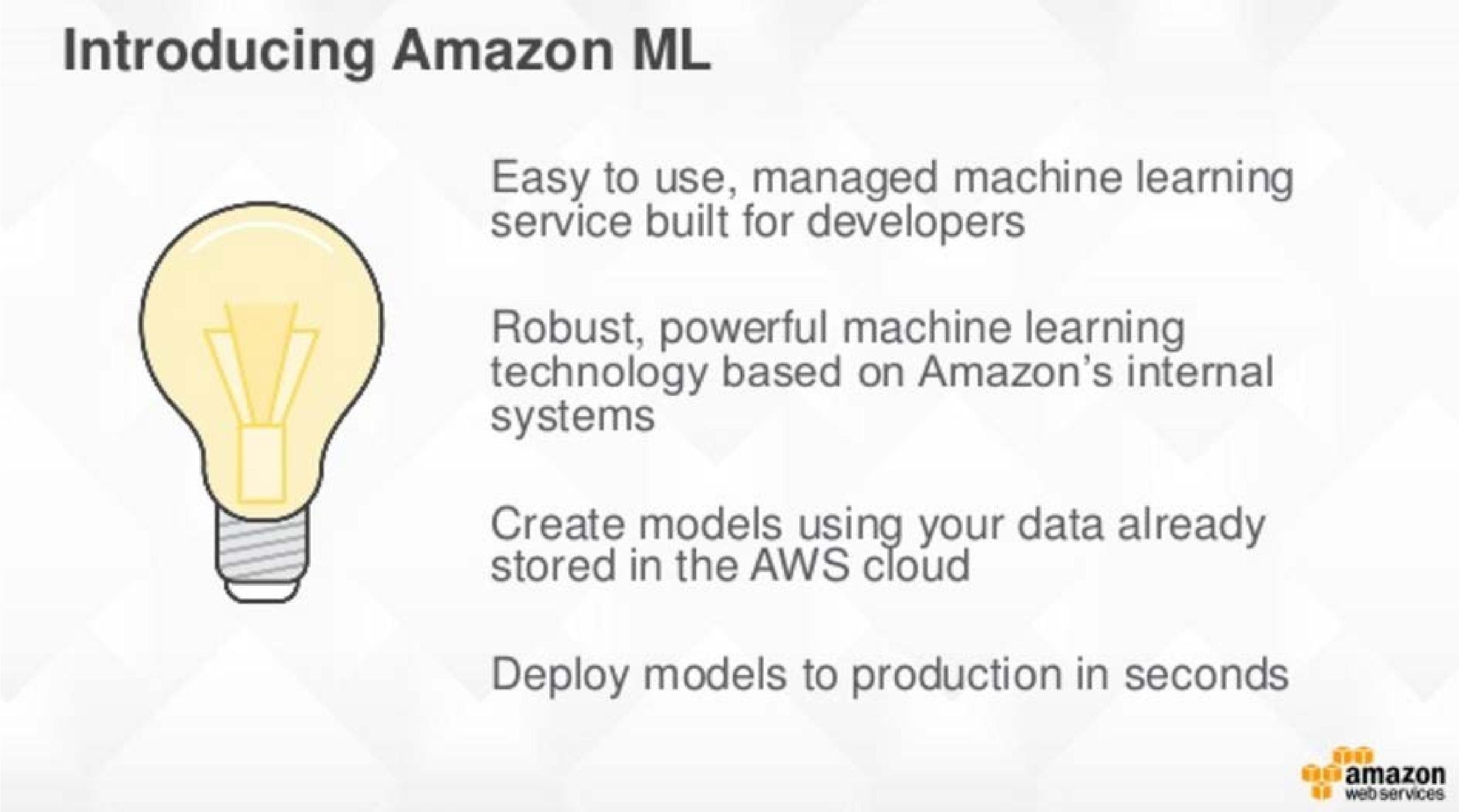
Bots

Automated Systems

Data → Intelligence → Action


Use your own data to train models
TensorFlow
Cloud Machine Learning Engine
Ready to use Machine Learning models
Cloud Vision API
Cloud Speech API
Cloud Jobs API
文→A
Cloud Translation API
Cloud Natural Language API
Cloud Video Intelligence API
Coming soon

Amazon Machine Learning



Introducing Amazon ML

Easy to use, managed machine learning service built for developers

Robust, powerful machine learning technology based on Amazon's internal systems

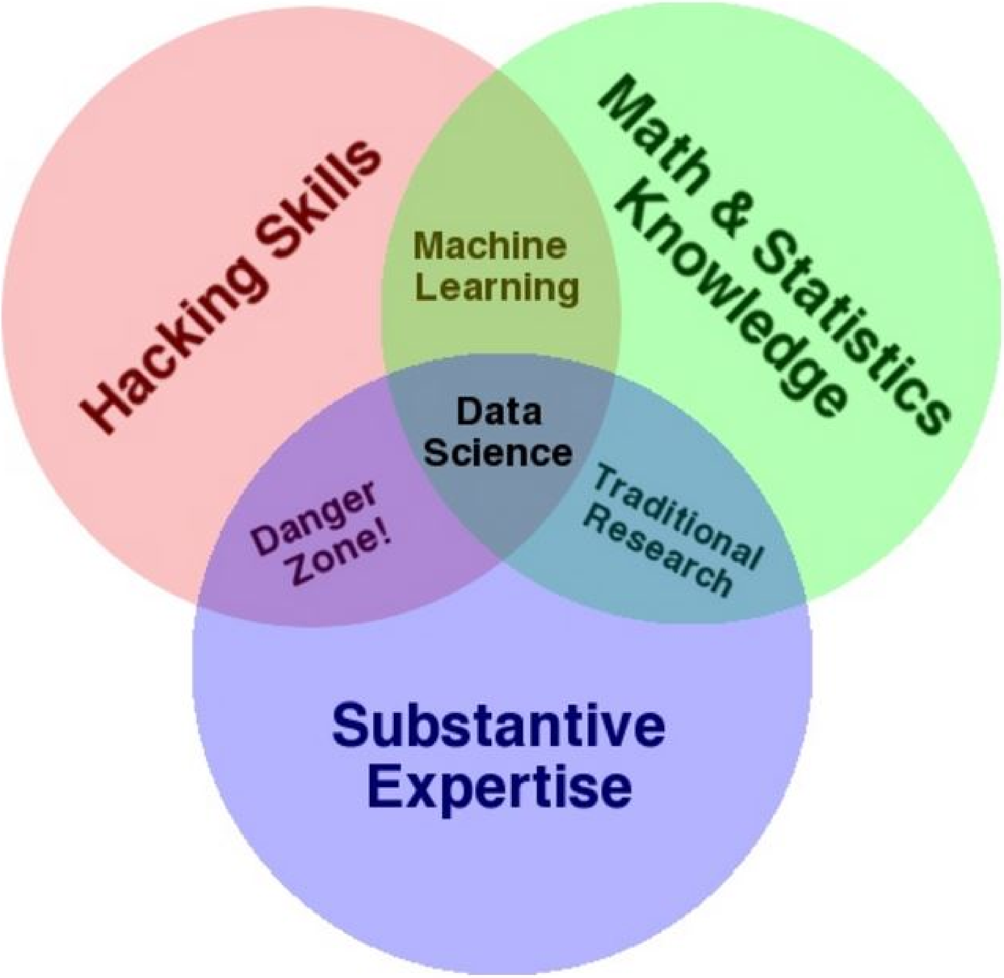
Create models using your data already stored in the AWS cloud

Deploy models to production in seconds

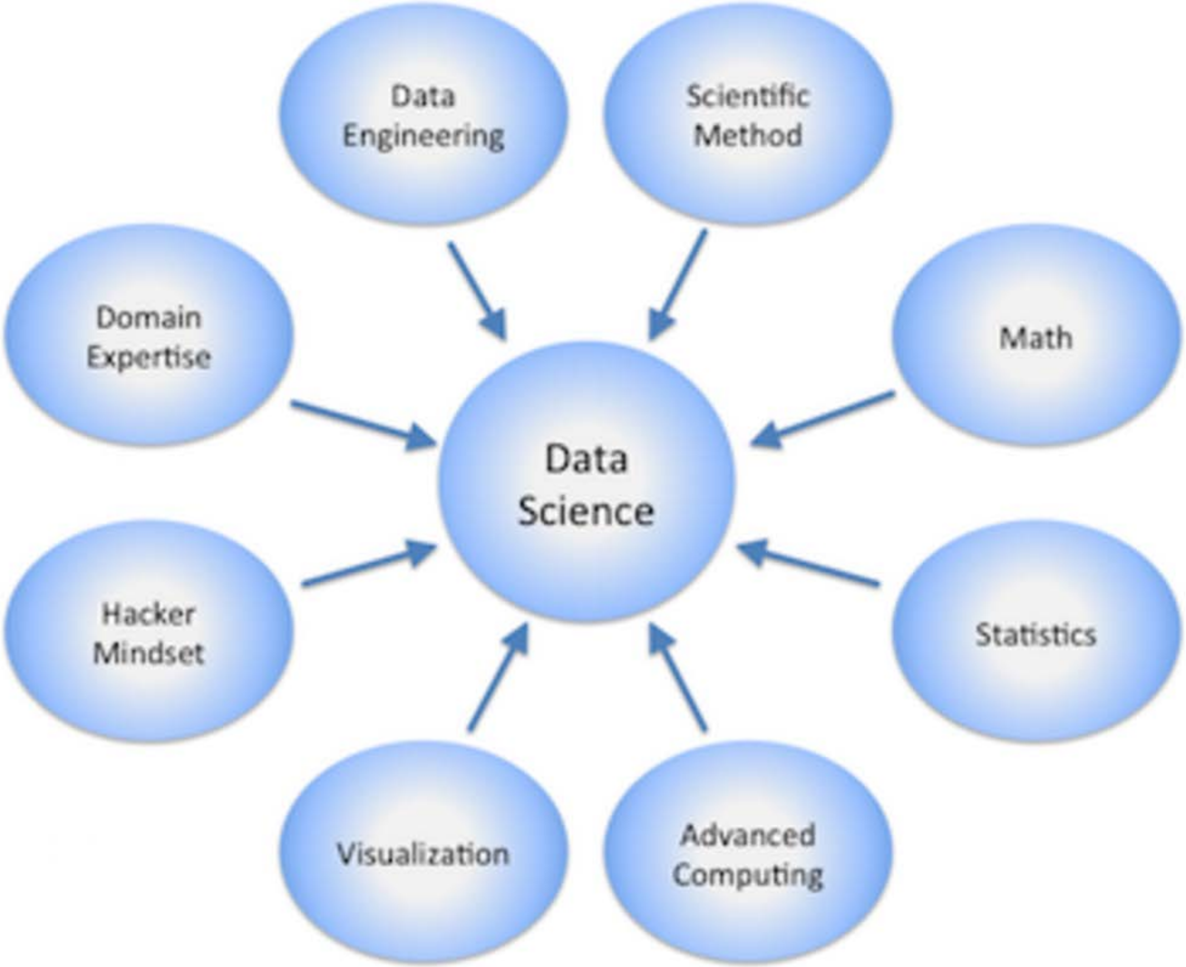
amazon
web services

DATA SCIENCE INTRODUCTION

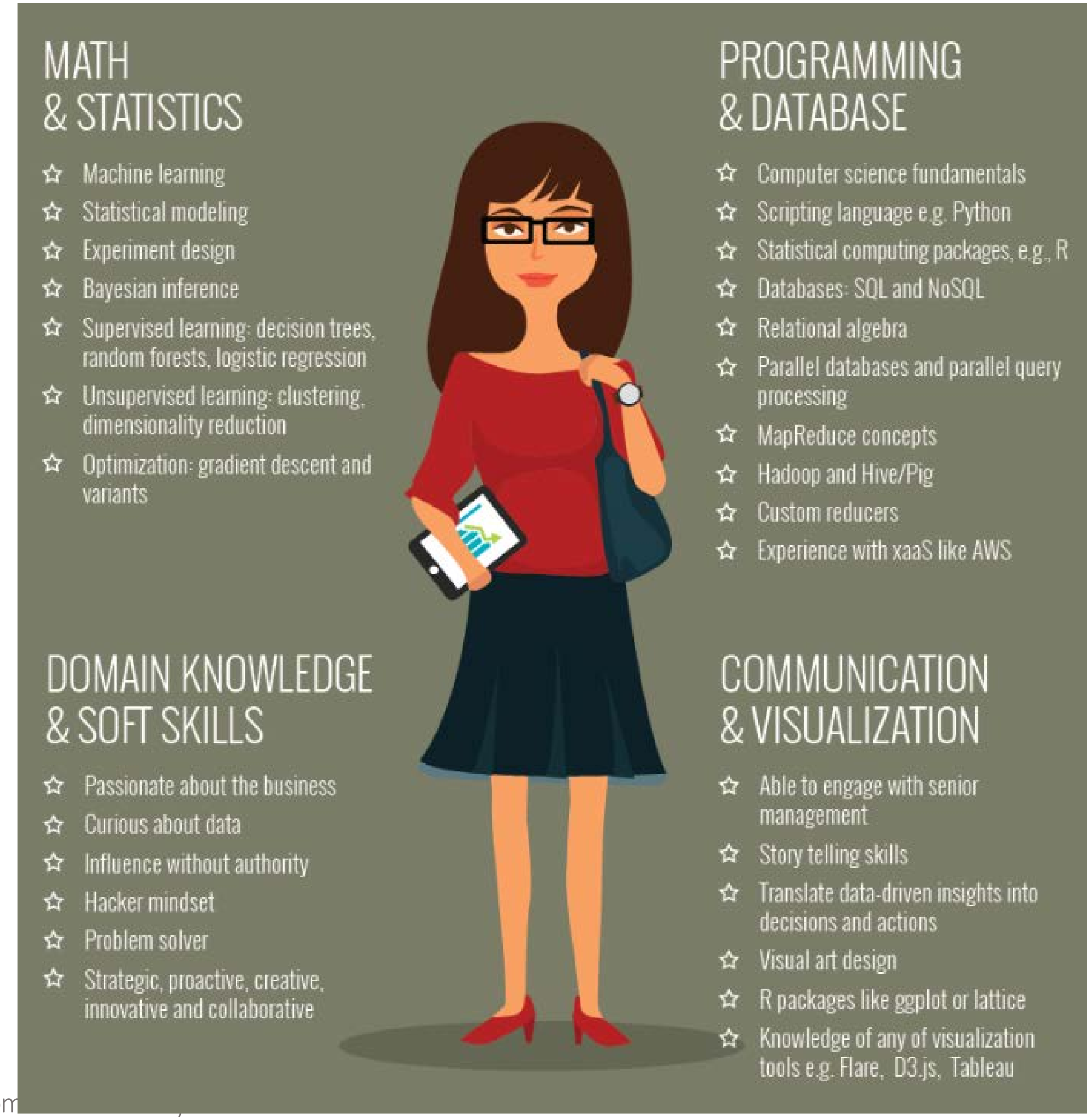
What is Data Science



What is data science?



Data scientist



MATH & STATISTICS

- Machine learning
- Statistical modeling
- Experiment design
- Bayesian inference
- Supervised learning: decision trees, random forests, logistic regression
- Unsupervised learning: clustering, dimensionality reduction
- Optimization: gradient descent and variants

PROGRAMMING & DATABASE

- Computer science fundamentals
- Scripting language e.g. Python
- Statistical computing packages, e.g., R
- Databases: SQL and NoSQL
- Relational algebra
- Parallel databases and parallel query processing
- MapReduce concepts
- Hadoop and Hive/Pig
- Custom reducers
- Experience with xaaS like AWS

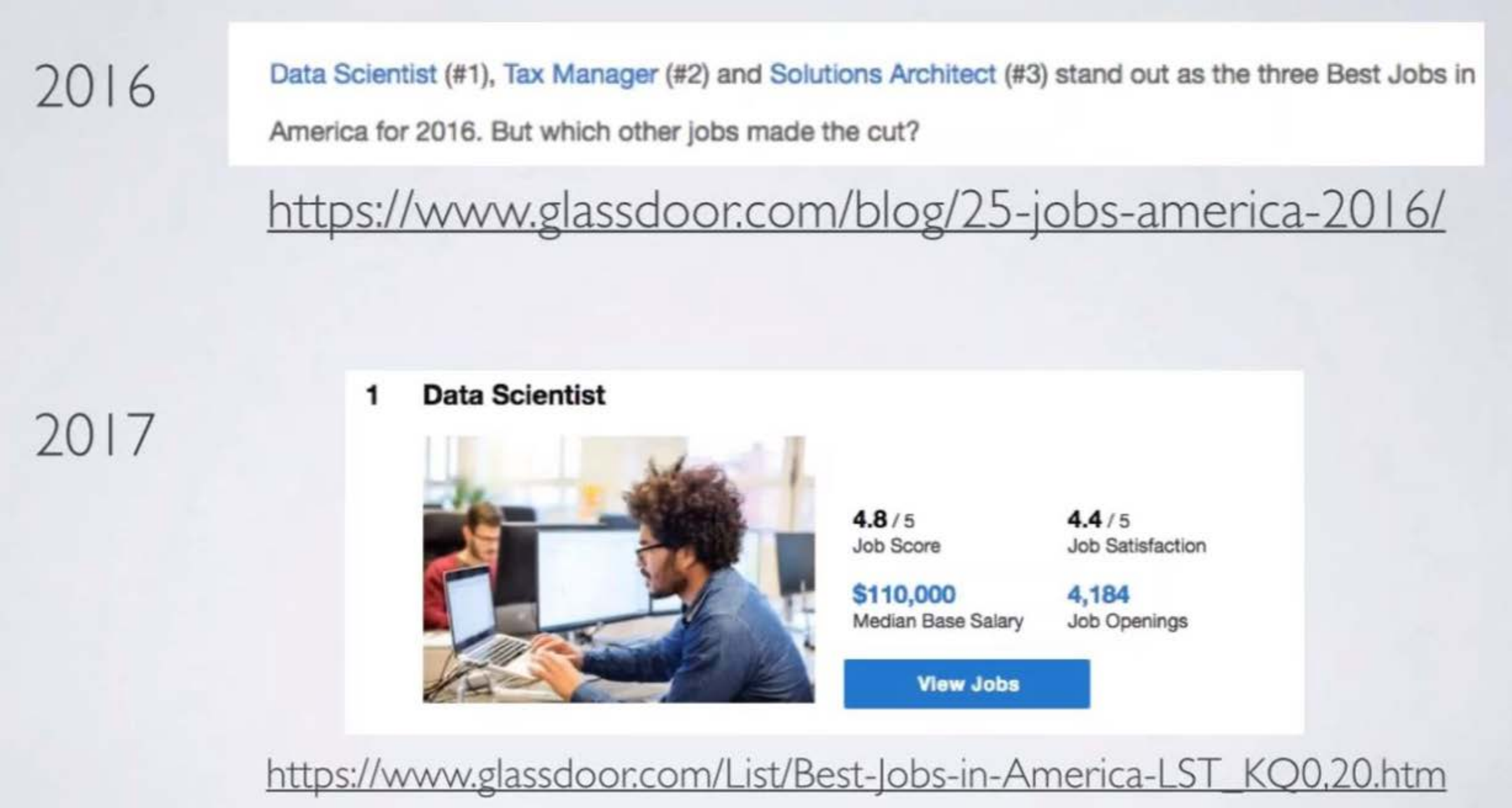
DOMAIN KNOWLEDGE & SOFT SKILLS

- Passionate about the business
- Curious about data
- Influence without authority
- Hacker mindset
- Problem solver
- Strategic, proactive, creative, innovative and collaborative

COMMUNICATION & VISUALIZATION

- Able to engage with senior management
- Story telling skills
- Translate data-driven insights into decisions and actions
- Visual art design
- R packages like ggplot or lattice
- Knowledge of any of visualization tools e.g. Flare, D3.js, Tableau

Glassdoor best job in 2016 - 2017



2016

Data Scientist (#1), Tax Manager (#2) and Solutions Architect (#3) stand out as the three Best Jobs in America for 2016. But which other jobs made the cut?

https://www.glassdoor.com/blog/25-jobs-america-2016/

2017

1 Data Scientist

4.8 / 5
Job Score

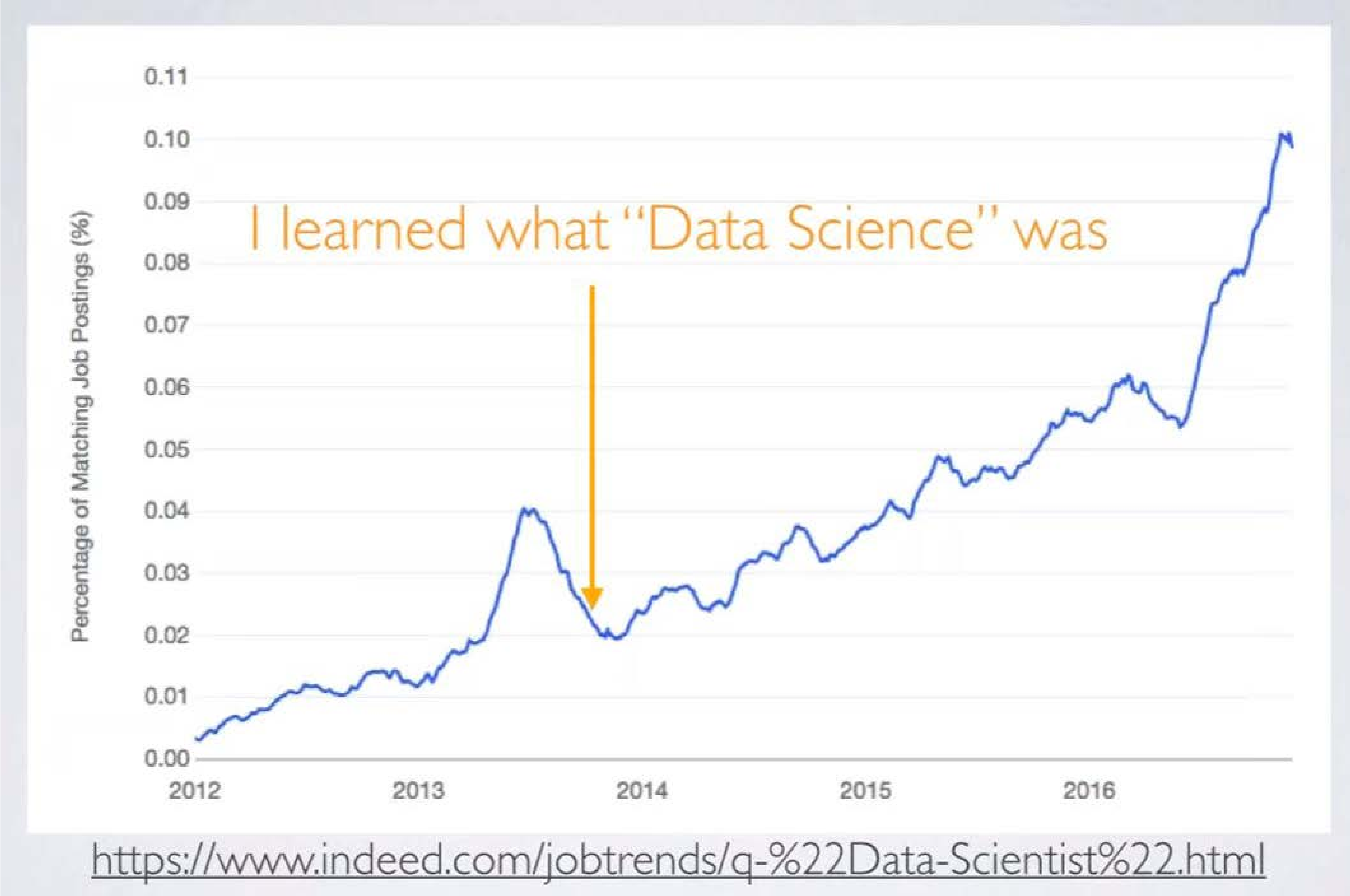
4.4 / 5
Job Satisfaction

$110,000
Median Base Salary

4,184
Job Openings

https://www.glassdoor.com/List/Best-Jobs-in-America-LST_KQ0,20.htm

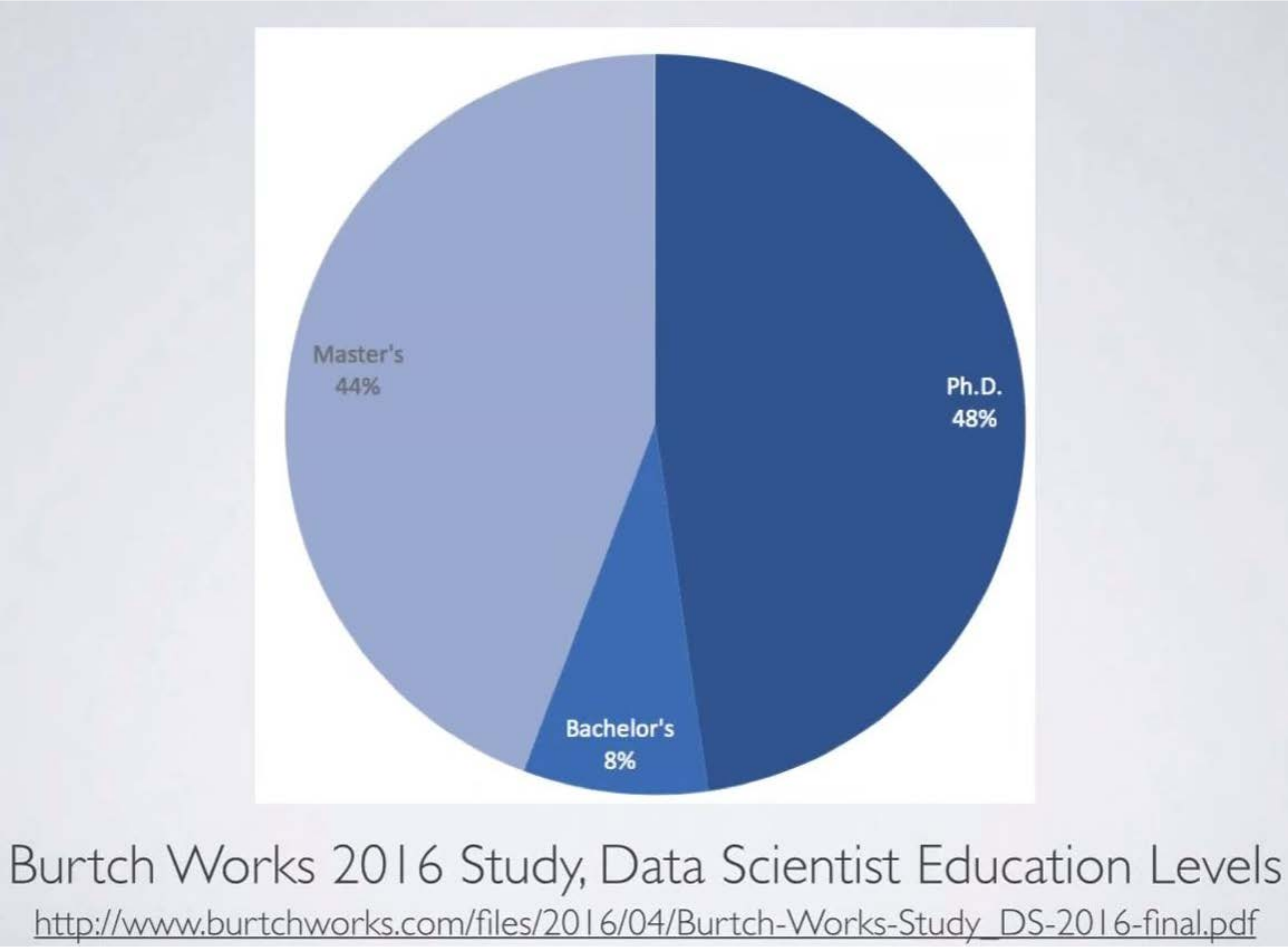
Data science job trend



I learned what "Data Science" was

https://www.indeed.com/jobtrends/q-%22Data-Scientist%22.html

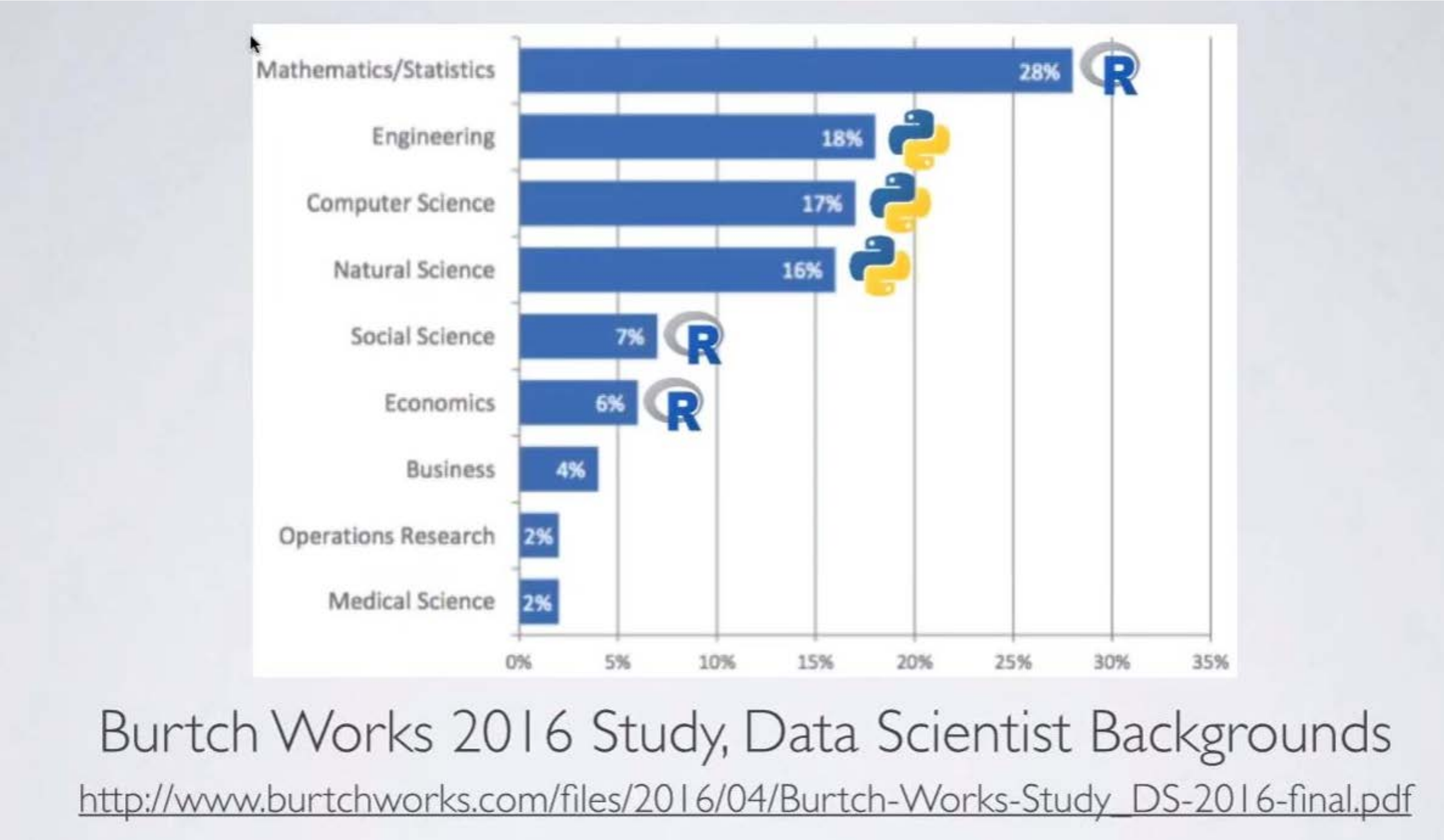
Data Scientist education levels



Burtch Works 2016 Study, Data Scientist Education Levels

http://www.burtchworks.com/files/2016/04/Burtch-Works-Study_DS-2016-final.pdf

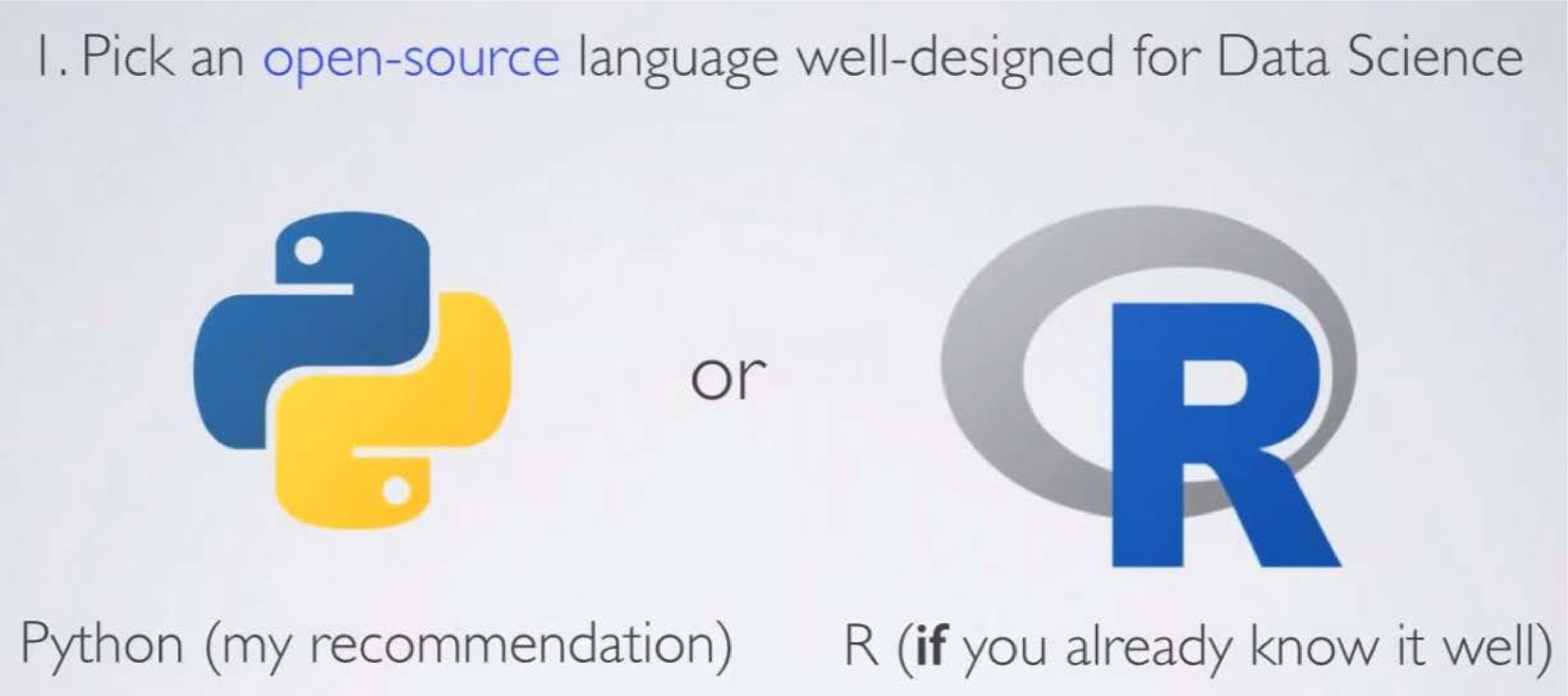
Data science backgrounds



Burtch Works 2016 Study, Data Scientist Backgrounds

http://www.burtchworks.com/files/2016/04/Burtch-Works-Study_DS-2016-final.pdf

Key topic to learn



1. Pick an open-source language well-designed for Data Science

Python (my recommendation) or R (**if** you already know it well)

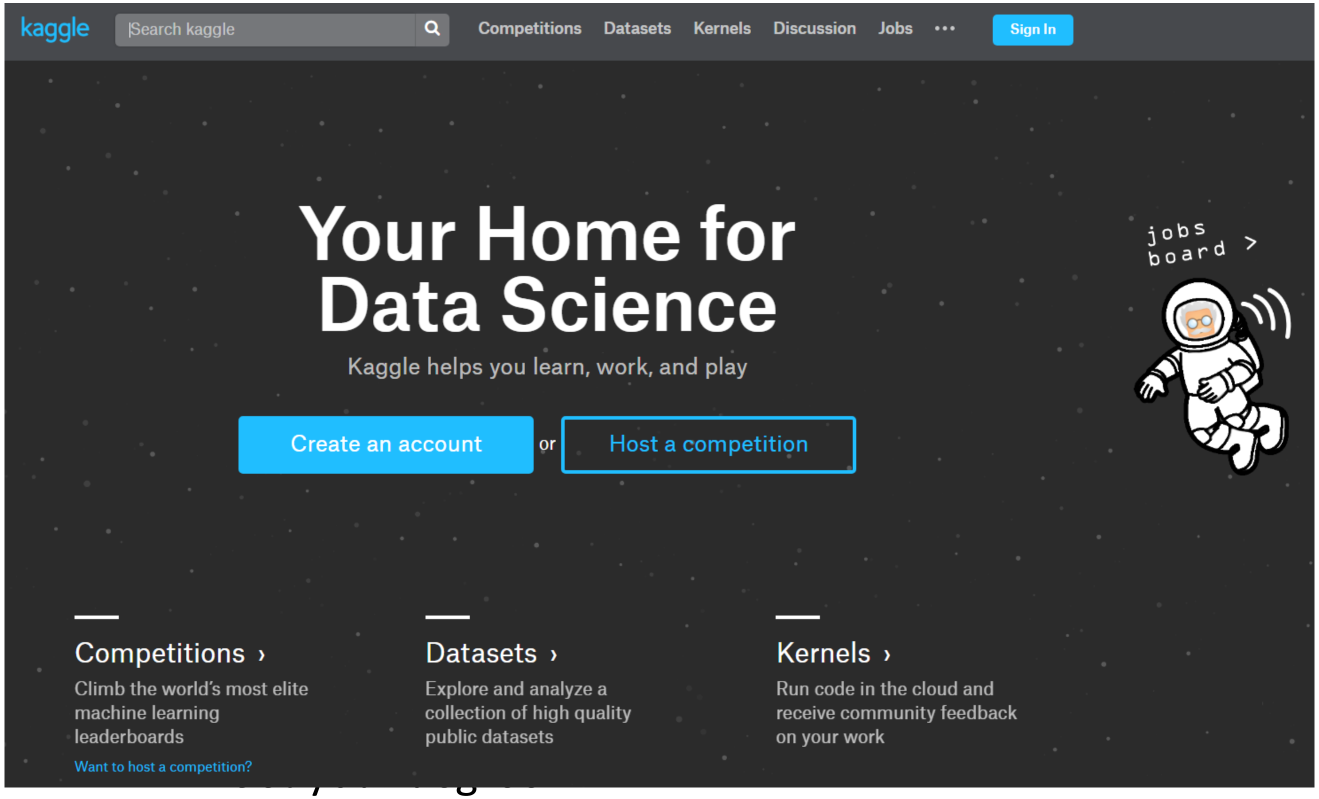
Learn Python library stack

SciPy
NumPy
matplotlib
IP[y]: IPython
Interactive Computing
pandas
$y_{it} = \beta' x_{it} + \mu_i + \epsilon_{it}$
scikit
learn

Go kaggle

- There are countless strategies that can be applied to any predictive modelling
- It is impossible to know at the outset which technique or analyst will be most effective
- Compete to produce the best models

kaggle™



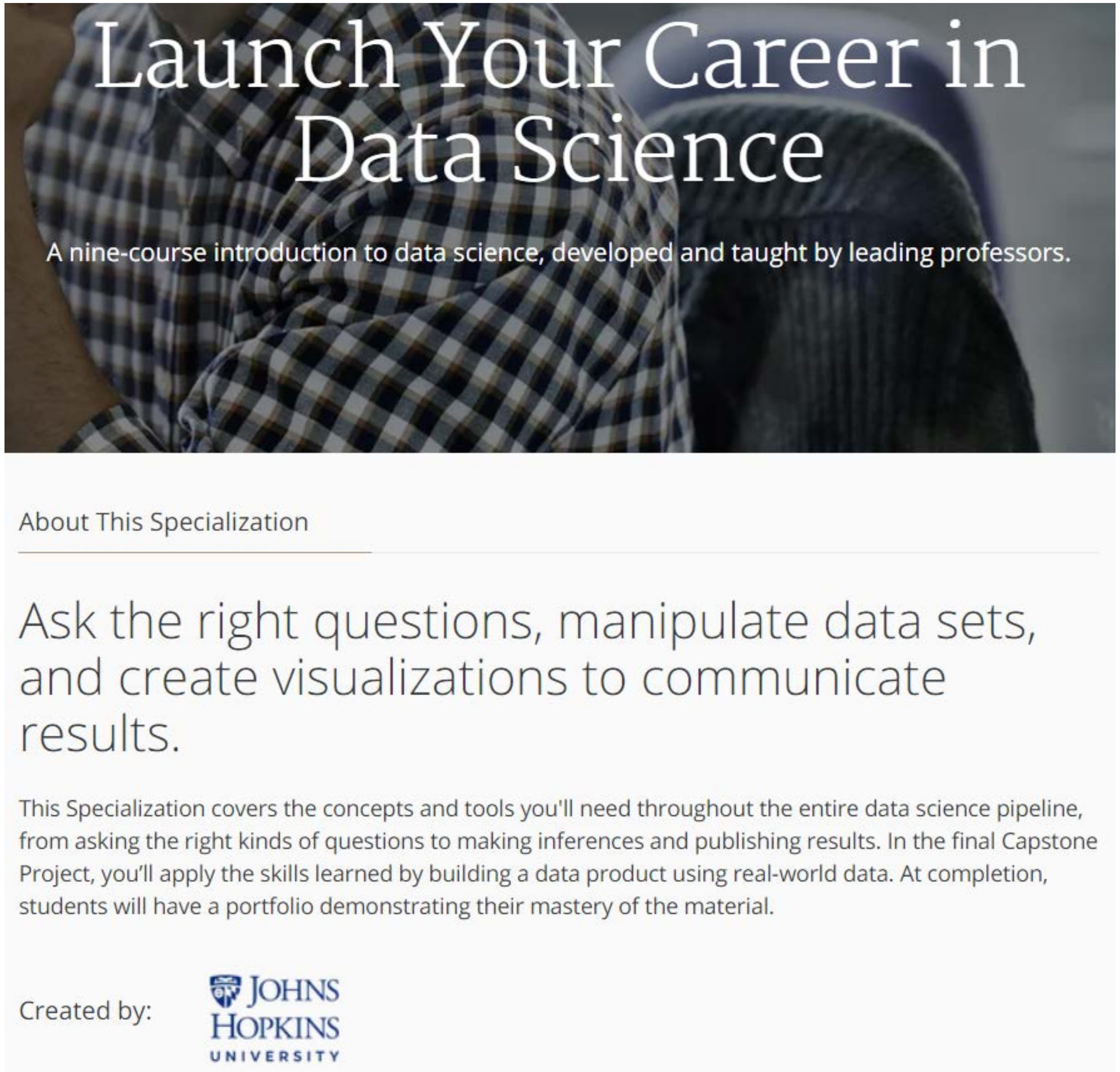
Your Home for Data Science

Kaggle helps you learn, work, and play

Competitions ›
Climb the world's most elite machine learning leaderboards
Want to host a competition?

Datasets ›
Explore and analyze a collection of high quality public datasets

Kernels ›
Run code in the cloud and receive community feedback on your work



Launch Your Career in Data Science

A nine-course introduction to data science, developed and taught by leading professors.

About This Specialization

Ask the right questions, manipulate data sets, and create visualizations to communicate results.

This Specialization covers the concepts and tools you'll need throughout the entire data science pipeline, from asking the right kinds of questions to making inferences and publishing results. In the final Capstone Project, you'll apply the skills learned by building a data product using real-world data. At completion, students will have a portfolio demonstrating their mastery of the material.

Created by: Johns Hopkins University

How is a Machine Learning Engineer different form a Data Scientist?

Data Scientist	ML Engineer
Trained to be strong in Data	Trained to be strong in Coding
R, Python, MATLAB	C++, Java, C#
Data treatment	Coding
Evaluate ML algorithm	Change algorithm to code
Evaluate ML module	Create ML module

DATA SCIENCE BASIC

In this session

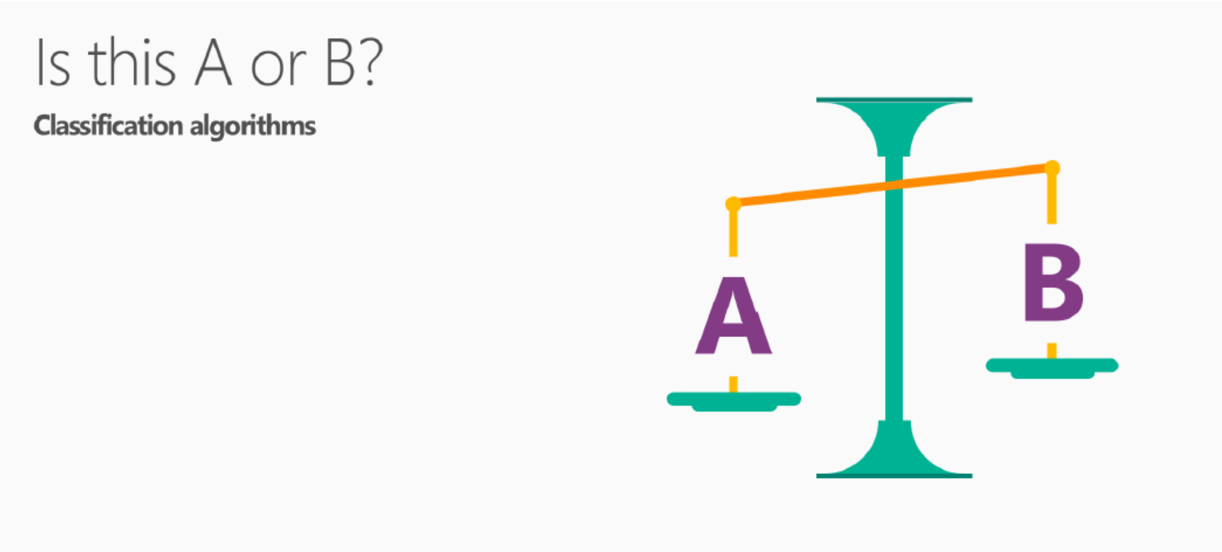
- The 5 questions data science answers
- Is your data ready for data science
- Ask a question you can answer with data
- Predict an answer with a simple model

The 5 questions data science answers

The 5 questions data science answers

- Is this A or B?
- Is this weird?
- How much – or – How many?
- How is this organized?
- What should I do next?

The 5 questions data science answers

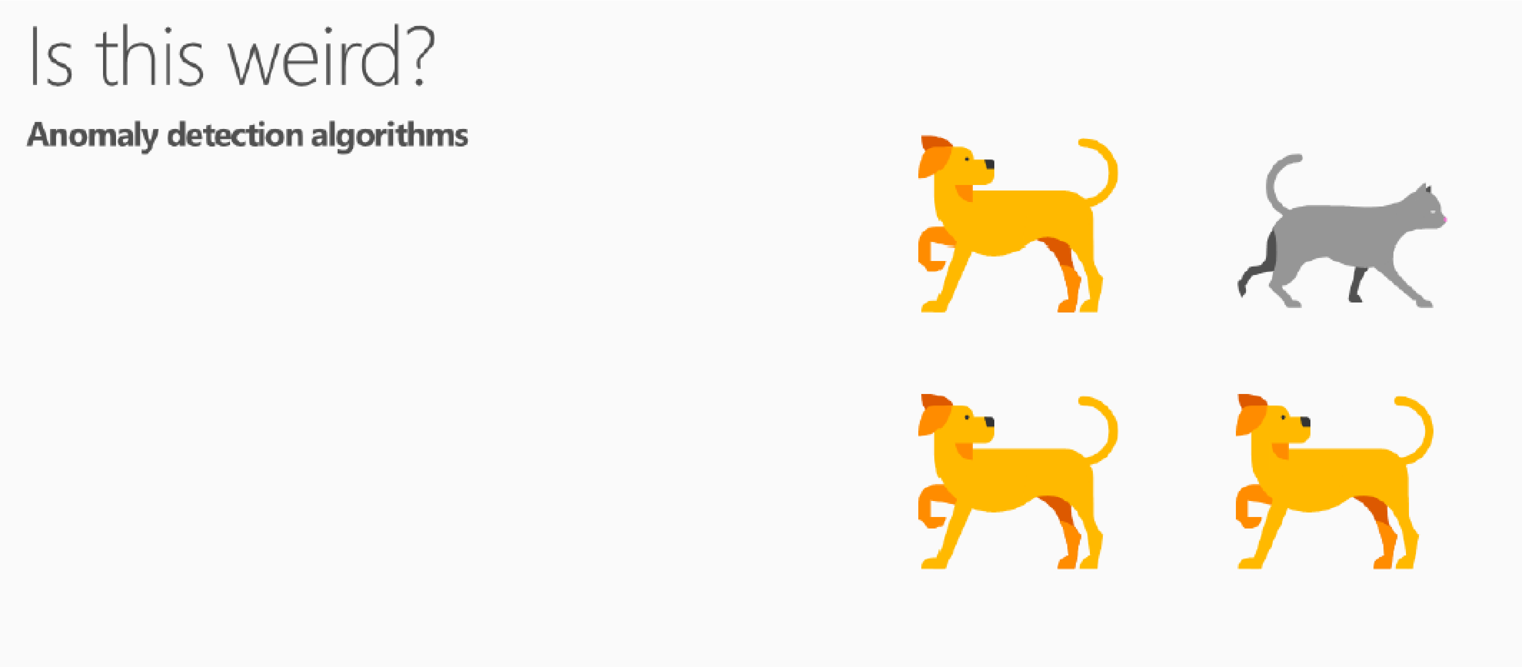


Is this A or B?

Classification algorithms

- Will this tire fail in the next 1,000 miles: Yes or no?
- Which brings in more customers: a $5 coupon or a 25% discount?
- Can also be more than two options: Is this A or B or C or D, etc.?
- Classification algorithms: helps choosing the most likely one.

The 5 questions data science answers

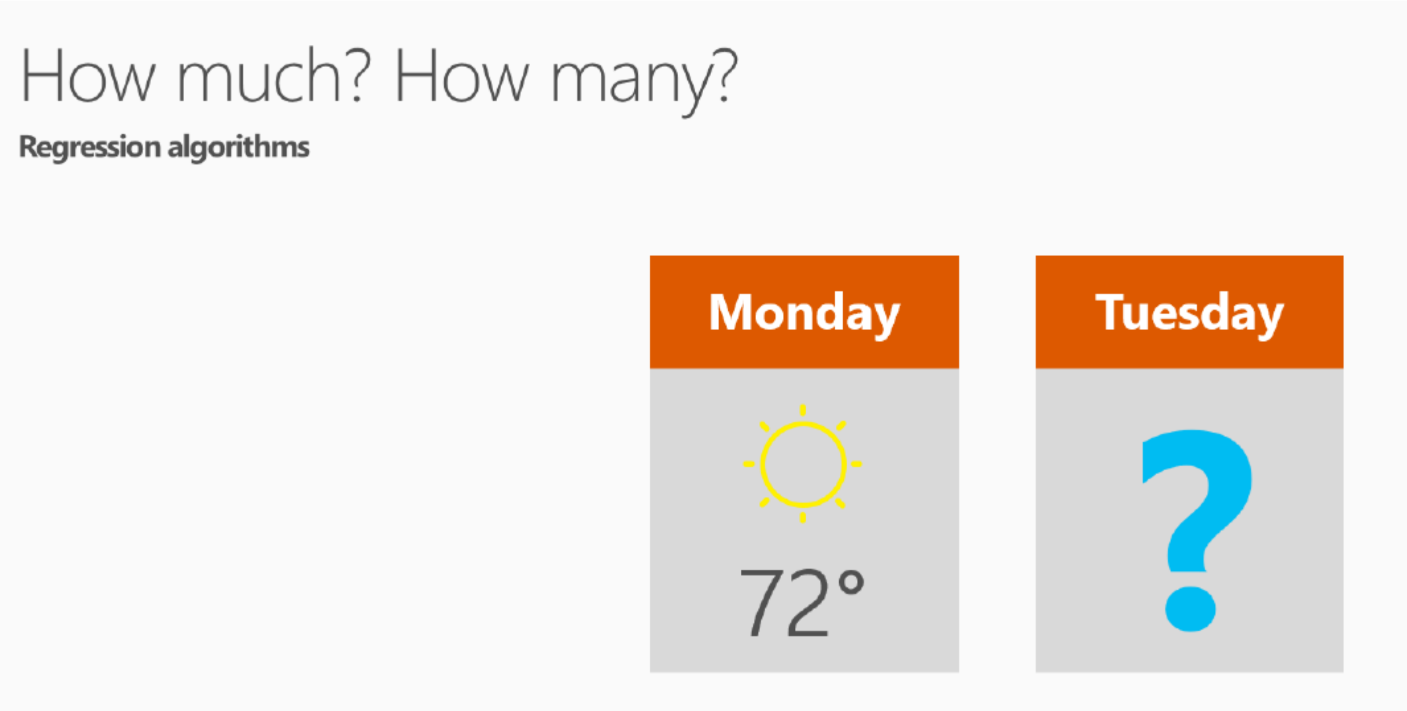


Is this weird?

Anomaly detection algorithms

- Is this pressure gauge reading normal?
- Is this message from the internet typical?
- Credit card purchase pattern normal?
- Anomaly algorithms: Detect unexpected or unusual events or behaviors

The 5 questions data science answers

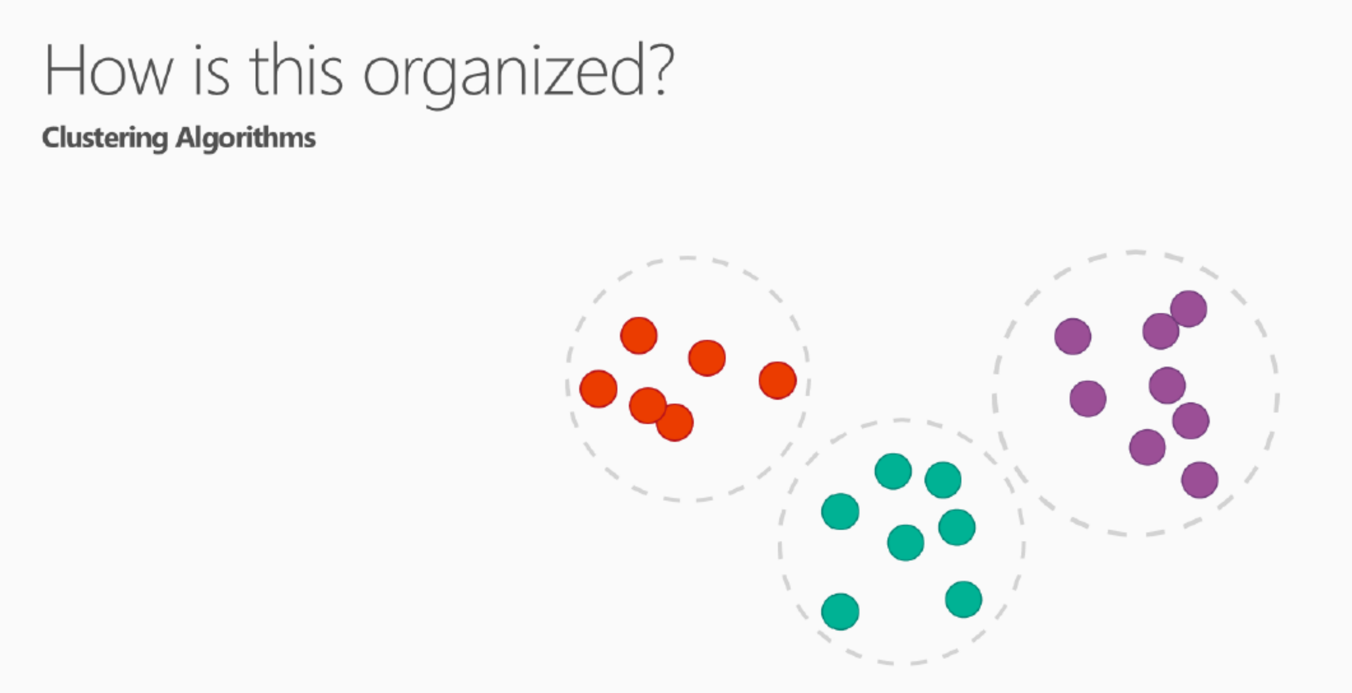


How much? How many?

Regression algorithms

- What will the temperature be next Tuesday?
- What will my fourth quarter sales be?
- Regression algorithms: Good for question that asks for a number

The 5 questions data science answers

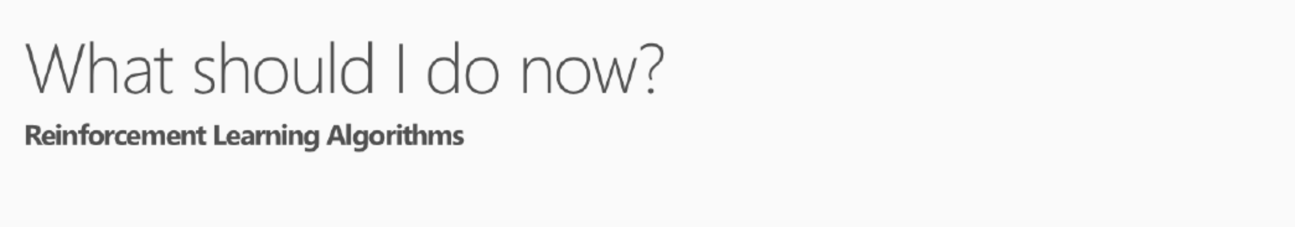


How is this organized?

Clustering Algorithms

- Which viewers like the same types of movies?
- Which printer models fail the same way?
- Clustering algorithms: helps arranging data into groups
- Understanding how data is organized, helps predict behaviors and events.

The 5 questions data science answers



What should I do now?

Reinforcement Learning Algorithms

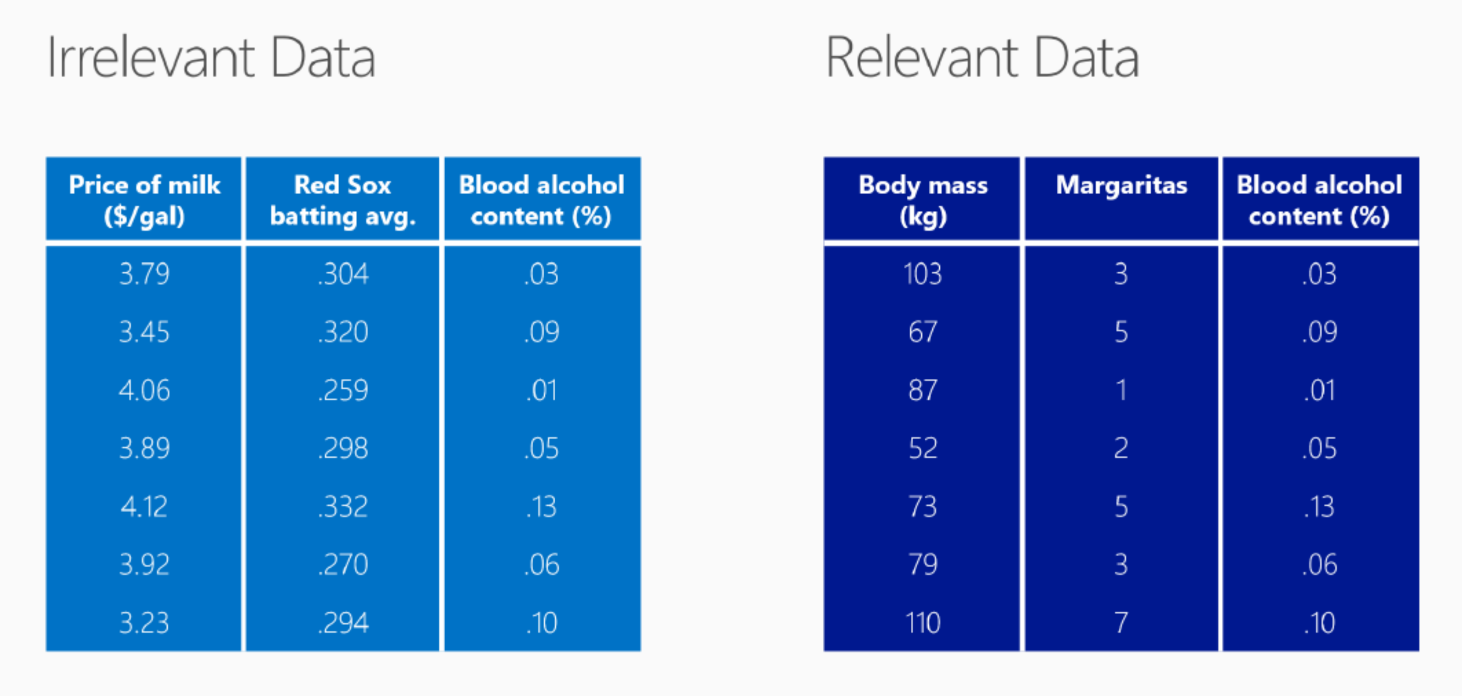
- Adjust the temperature or leave it where it is?
- At a yellow light, brake or accelerate?
- Keep vacuuming, or go back to the charging station?
- Reinforcement learning algorithms: the brains of rats and humans respond to punishment and rewards. learning from trial and error.

Is your data ready for data science

We need data that is:

- Relevant
- Connected
- Accurate
- Enough to work with

Relevant



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

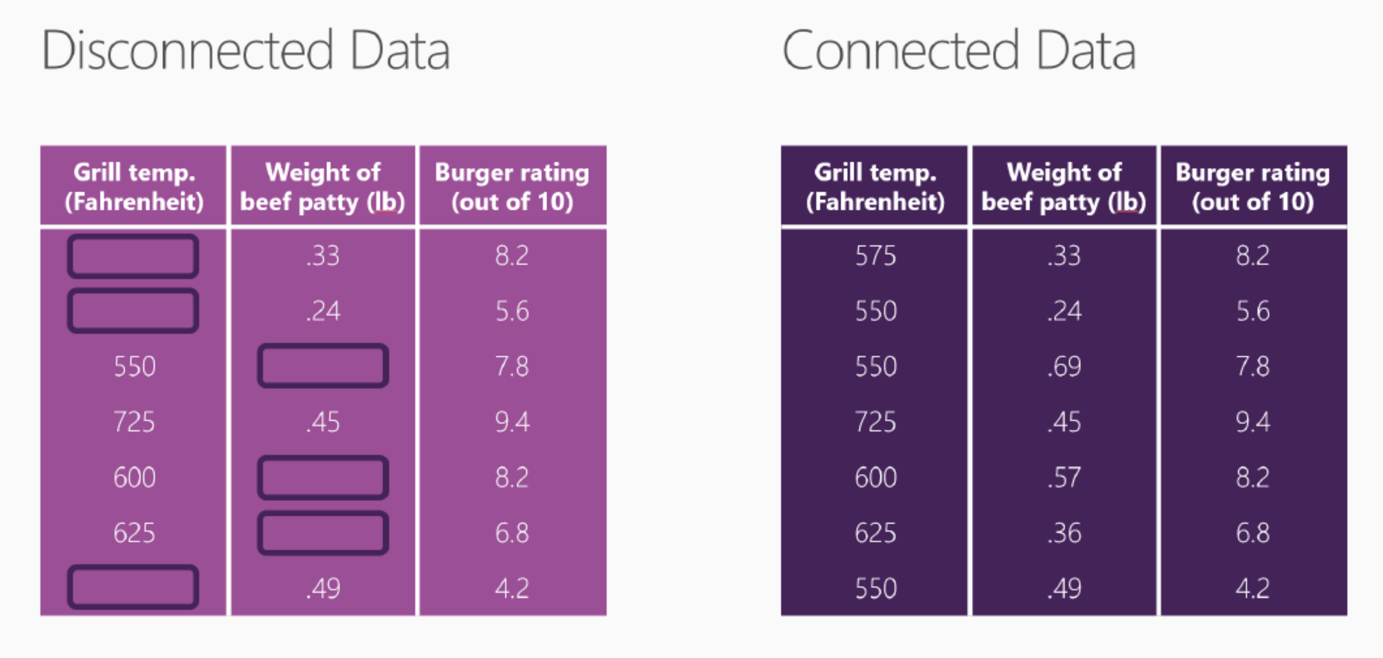
Relevant Data

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

- We need to know Blood alcohol content %
- Price of milk and Red Sox are irrelevant

Is your data ready for data science

Connected



Disconnected Data

Grill temp. (Fahrenheit)	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

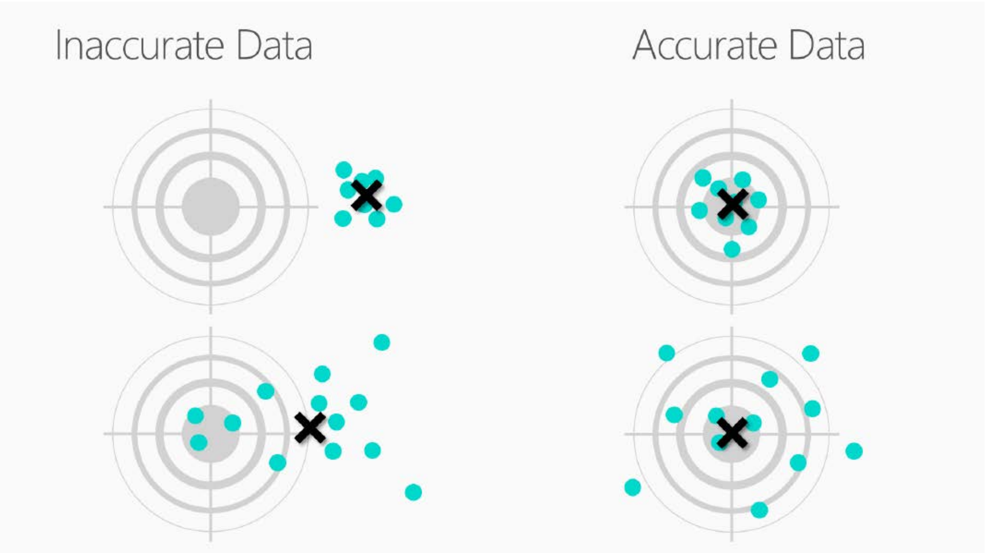
Connected Data

Grill temp. (Fahrenheit)	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

- Quality of hamburgers
- But notice the gaps in the table on the left
- It's common to have holes like this

Is your data ready for data science

Accurate



- Top left: precise=yes/accurate=no
- Button left: precise=no/accurate=no
- Top right: precise=yes/accurate=yes
- Button right: precise=no/accurate=yes

Is your data ready for data science

Enough to work with

We need enough data to work with

1. Not enough data: can not make decision
2. Barely enough data: can make basic decision (Is it somewhere I might want to visit? It looks bright, that looks like clean water – yes, that's where I'm going on vacation.)
3. Enough data: can make detailed decision (Now I can look at the three hotels on the left bank. You know, I really like the architectural features of the one in the foreground. I'll stay there, on the third floor.)

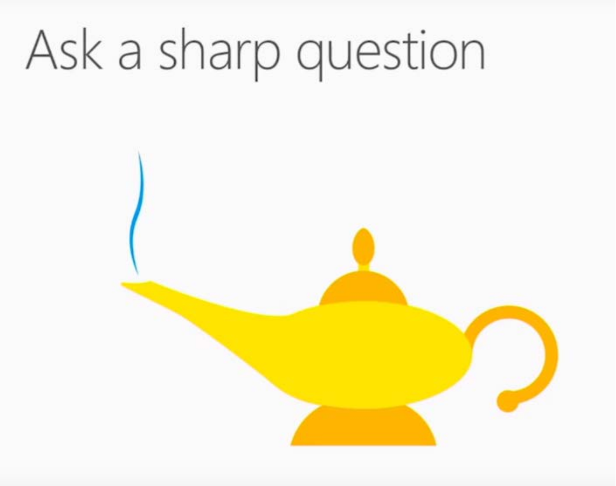
Ask a question you can answer with data

Ask a question you can answer with data

- Sharp question is the Key
- Know Target Data
- Reword the question

Ask a question you can answer with data

Sharp question is the Key


Ask a sharp question

- Asking a sharp question is the most important
- ML is a mischievous genie
- "What's going to happen with my stock?", the genie might answer, "The price will change"
- "What will my stock's sale price be next week?", the genie can't help but give you a specific answer and predict a sale price

Ask a question you can answer with data

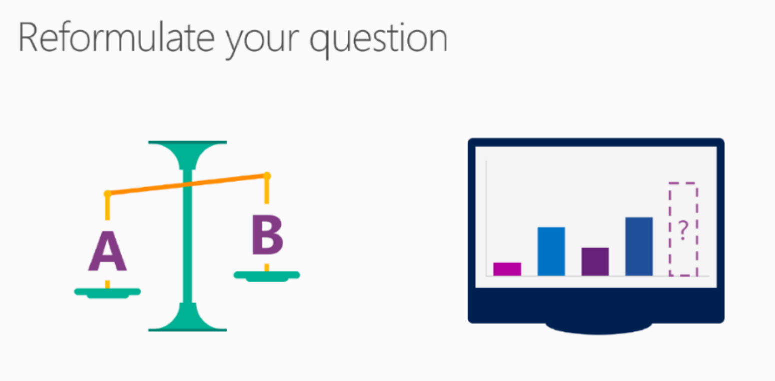
Know Target Data

Examples of the answer: Target data

- Target data = what we are trying to predict
- Must have examples of the target in data.
- Question = "What will my stock's sale price be next week?" Target = stock price history.
- Question = "Which car in my fleet is going to fail first?" Target = previous failures data.

Ask a question you can answer with data

Reword the question



Reformulate your question

- Question dictates the algorithm
- "Is this data point A or B?" = classification
- "How much?" or "How many?" = regression
- "Which news story is the most interesting to this reader?"
- Algorithm = classification A or B or C or D; difficult
- Reword = "How interesting is each story on this list to this reader?"
- Give each article a numerical score
- Identify the highest-scoring article; easy
- Above example change classification question into a regression question

Predict an answer with a simple model

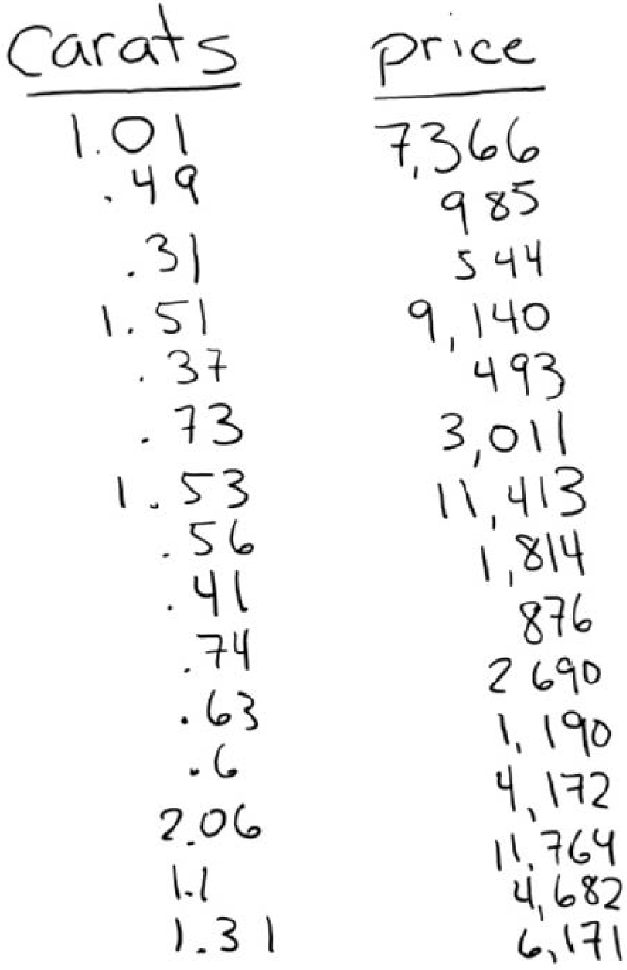
Predict an answer with a simple model

- Collect relevant, accurate, connected, enough data
- Ask a sharp question
- Plot the existing data
- Draw the model through the data points
- Use the model to find the answer
- Create a confidence interval

Predict an answer with a simple model

Collect relevant, accurate, connected, enough data

- I want to how much 1.35 carat diamond will cost
- Go to jewelry store
- Write down the price of all of the diamonds
- List has two columns
- Each column has a different attribute
- Weight in carats and price
- Each row is a single data point
- Data that represents a single diamond.
- This is a small data set; a table



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	2690
.63	1,190
.6	4,172
2.06	11,764
1.1	4,682
1.31	6,171

Predict an answer with a simple model

This data set meets our criteria for quality:

- Relevant: weight is definitely related to price
- Accurate: we double-checked the prices that we write down
- Connected: there are no blank spaces in either of these columns
- Enough data: to answer our question

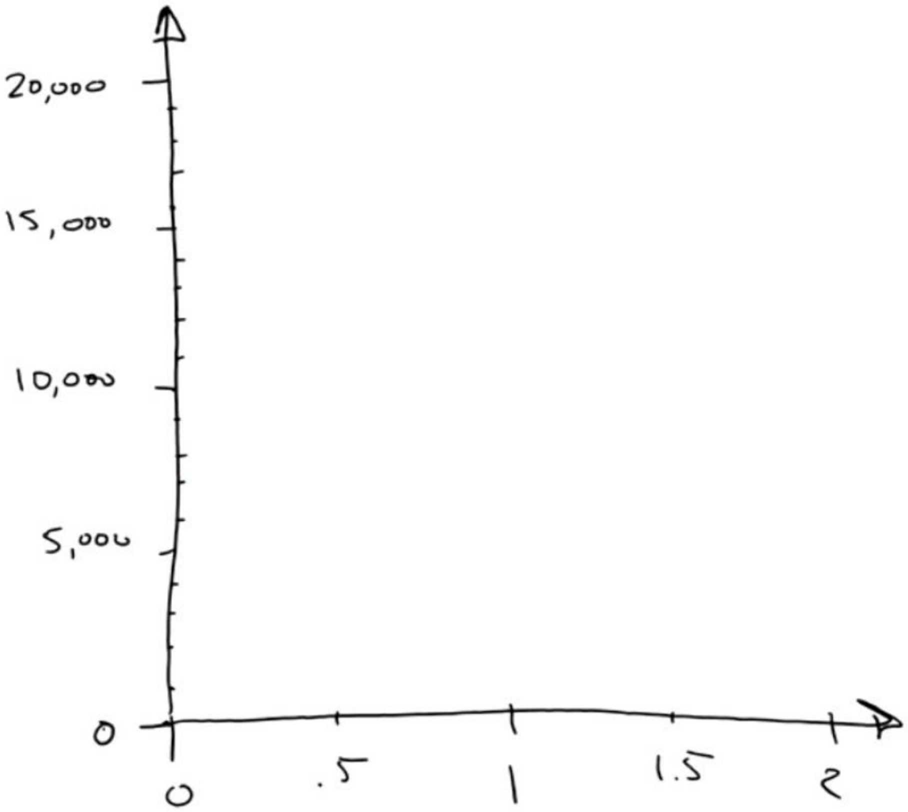
Ask a sharp question

- How much will it cost to buy a 1.35 carat diamond?
- Our list doesn't have a 1.35 carat diamond
- Use the rest of our data to get an answer to the question

Predict an answer with a simple model

Draw axis

- Draw a horizontal number line, called an axis, to chart the weights
- The range of the weights is 0 to 2
- Line covers that range and put ticks for each half carat
- Draw a vertical axis to record the price and connect it to the horizontal weight axis
- This will be in units of dollars
- Now we have a set of coordinate axes.

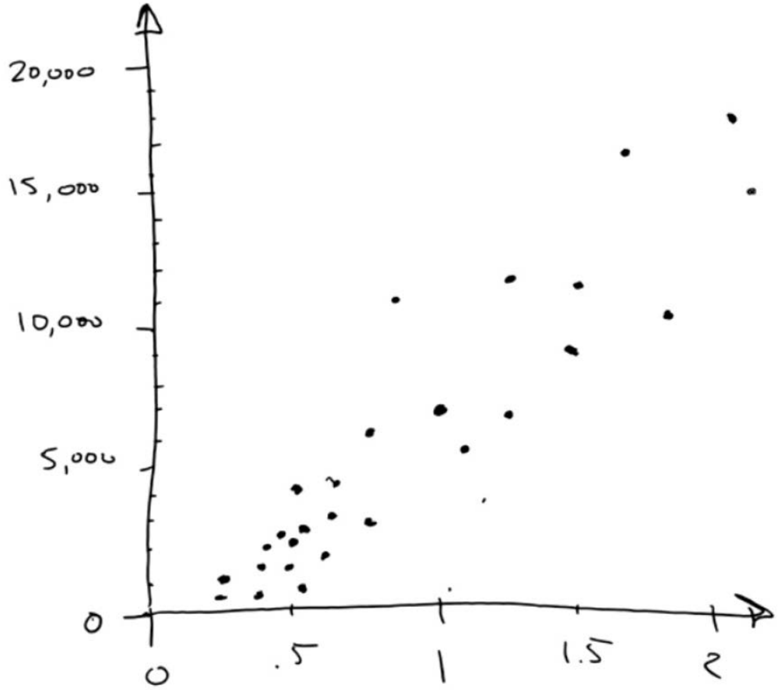


Predict an answer with a simple model

Plot the existing data

Carats	price
1.01	7,366
.49	985
.31	544

- Make a scatter plot
- Great way to visualize numerical data sets
- For the first data point, we eyeball a vertical line at 1.01 carats. Then, we eyeball a horizontal line at $7,366. Where they meet, we draw a dot
- This represents our first diamond.
- Now we go through each diamond on this list and do the same thing.
- We get a bunch of dots, one for each diamond



Predict an answer with a si

Draw the model through the data points

20,000
15,000

- Look at the dots and squint, the collection looks like a fat, fuzzy line
- Draw a straight line through it
- This a model
- Model = cartoon
- The cartoon is wrong
- But, it’s a useful simplification
- The line doesn't go through all the data points.
- It has some noise or variance
- But, it's a useful simplification
- Question = How much? regression
- we're using a straight line, linear regression

Predict an answer with a simple model

Use the model to find the answer

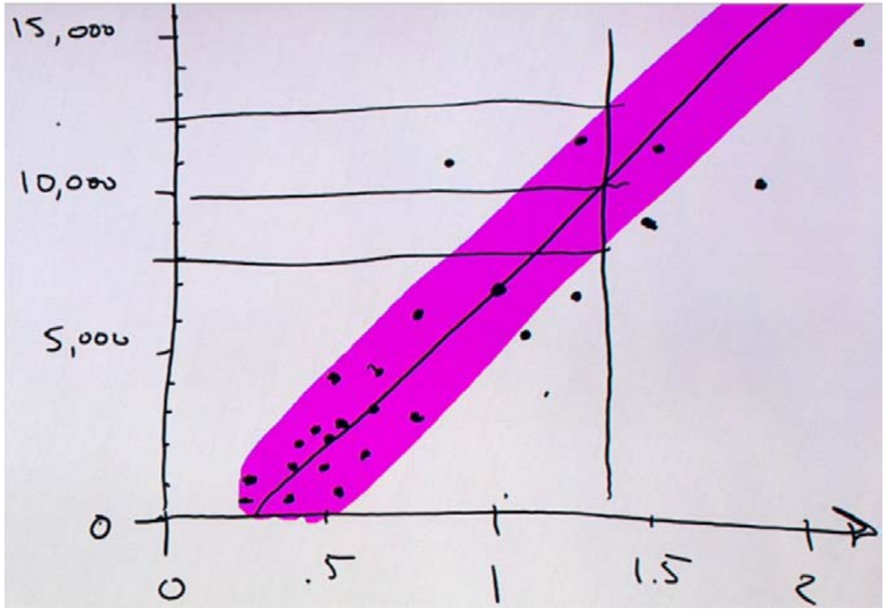
20,000
15,000

- How much will a 1.35 carat diamond cost?
- Look at 1.35 carats
- Draw a vertical line
- Draw at horizontal line to the dollar axis
- It hits right at 10,000
- Answer = about $10,000

Predict an answer with a simple model

Create a confidence interval

- How precise this prediction is?
- Is it a lot higher or lower?
- Draw an envelope around the regression line
- that includes most of the dots.
- This envelope is called our confidence interval
- We're pretty confident that prices fall within this envelope, because in the past most of them have.
- We can draw two more horizontal lines from where the 1.35 carat line crosses the top and the bottom of that envelope.
- The price of a 1.35 carat diamond is about $10,000 - but it might be as low as $8,000 and it might be as high as $12,000



We're done, with no math or computers!!

- We did what data scientists get paid to do, and we did it just by drawing:
- We asked a question that we could answer with data
- We built a model using linear regression
- We made a prediction, complete with a confidence interval

And we didn't use math or computers to do it.

Predict an answer with a simple model

Now if we'd had more information, like...

- the cut of the diamond
- color variations (how close the diamond is to being white)
- the number of inclusions in the diamond

...then we would have had more columns. In that case, math becomes helpful. If you have more than two columns, it's hard to draw dots on paper. The math lets you fit that line or that plane to your data very nicely.

$$a = \frac{n\sum_{i=1}^{n} x_i y_i - \sum_{i=1}^{n} x_i \sum_{i=1}^{n} y_i}{n\sum_{i=1}^{n} x_i^2 - (\sum_{i=1}^{n} x_i)^2}$$

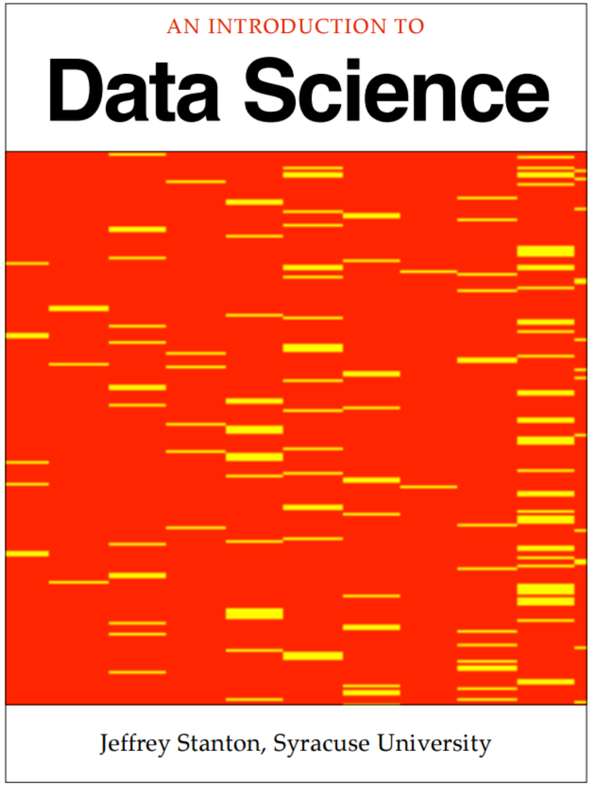
Also, if instead of just a handful of diamonds, we had two thousand or two million, then you can do that work much faster with a computer.

$$b = \frac{1}{n}(\sum_{i=1}^{n} y_i - a\sum_{i=1}^{n} x_i)$$

More information

More information on Data Science Basic

An introduction to Data Science: Jeffrey Stanton
https://ischool.syr.edu/media/documents/2012/3/DataScienceBook1_1.pdf

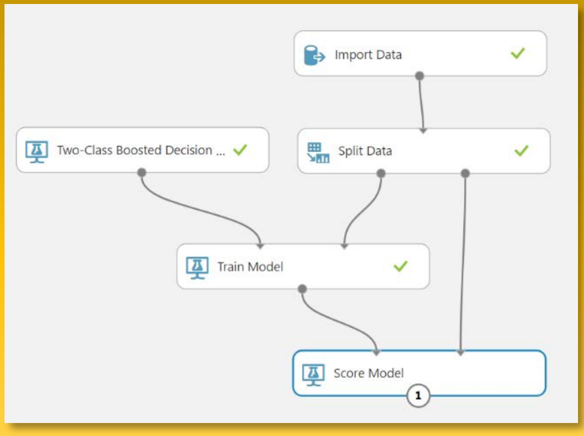


AN INTRODUCTION TO

Data Science

Jeffrey Stanton, Syracuse University

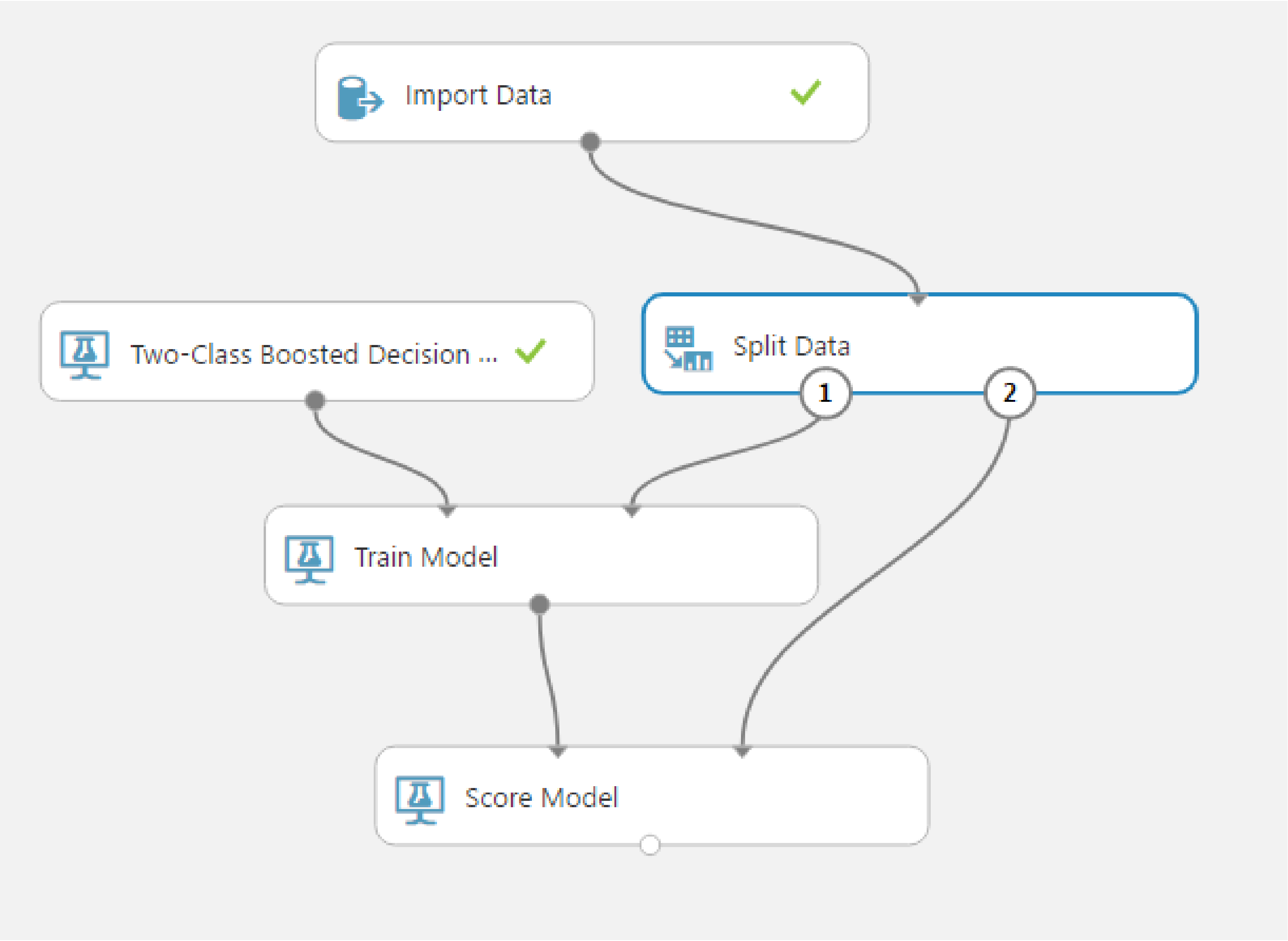
FIRST EXPERIMENT



In this session

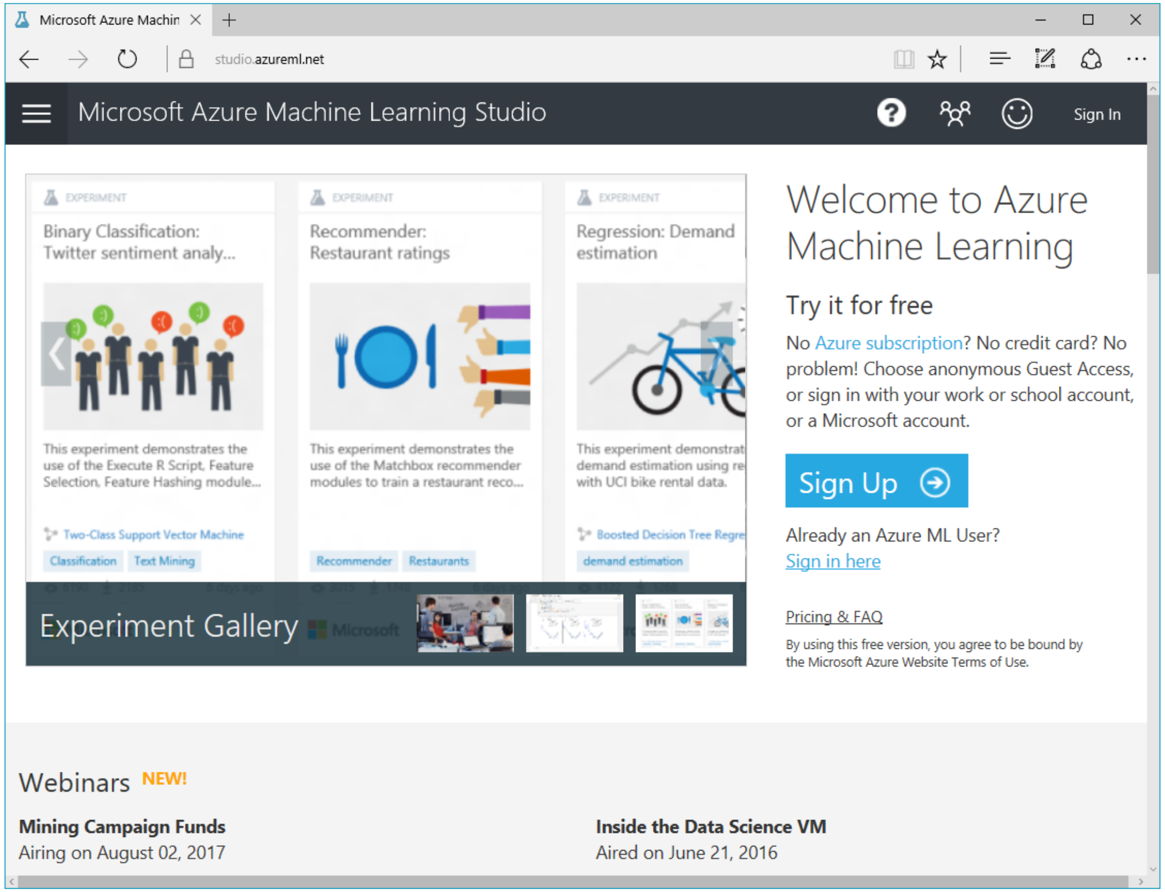
- Sing Up FREE Azure ML Studio Subscription
- Create Azure ML Studio workspace
- Train, Test, Evaluate for Binary Classification
- Import census income dataset
- Create a new Azure Machine Learning experiment
- Train and evaluate a prediction model
- Type of datasets

First experiment model



Sing Up FREE Azure ML Studio Subscription

https://studio.azureml.net/



Microsoft Azure Machine Learning Studio

Sign In

EXPERIMENT

Binary Classification: Twitter sentiment analy...

This experiment demonstrates the use of the Execute R Script, Feature Selection, Feature Hashing module...

Two-Class Support Vector Machine

Classification Text Mining

EXPERIMENT

Recommender: Restaurant ratings

This experiment demonstrates the use of the Matchbox recommender modules to train a restaurant reco...

Recommender Restaurants

EXPERIMENT

Regression: Demand estimation

This experiment demonstrat demand estimation using re with UCI bike rental data.

Boosted Decision Tree Regre

demand estimation

Experiment Gallery

Welcome to Azure Machine Learning

Try it for free

No Azure subscription? No credit card? No problem! Choose anonymous Guest Access, or sign in with your work or school account, or a Microsoft account.

Sign Up

Already an Azure ML User?
Sign in here

Pricing & FAQ

By using this free version, you agree to be bound by the Microsoft Azure Website Terms of Use.

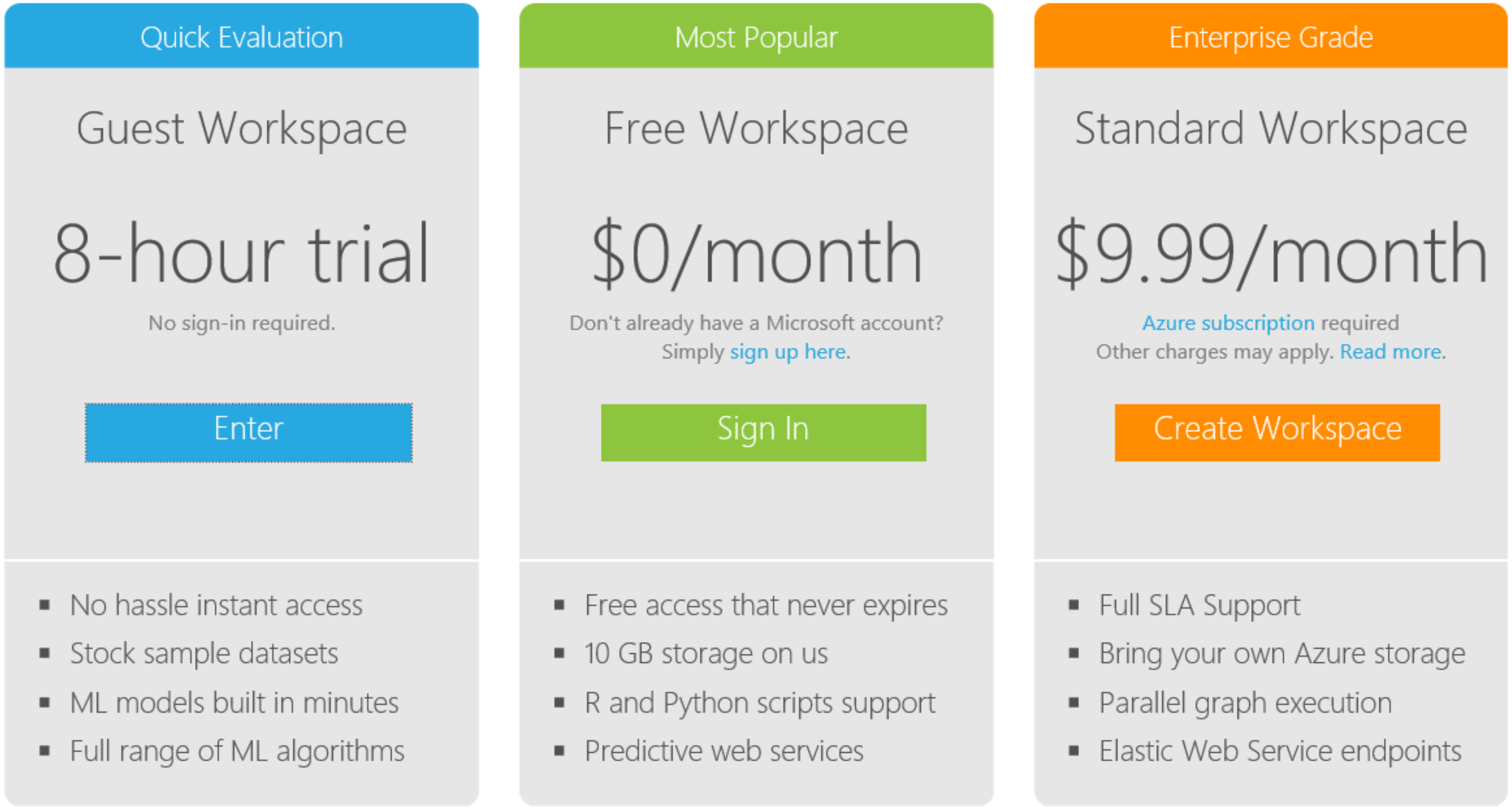
Webinars NEW!

Mining Campaign Funds
Airing on August 02, 2017

Inside the Data Science VM
Aired on June 21, 2016

Sing Up FREE Azure ML Studio Subscription

Free Workspace -> sign up here



Quick Evaluation

Guest Workspace

8-hour trial

No sign-in required.

Enter

- No hassle instant access
- Stock sample datasets
- ML models built in minutes
- Full range of ML algorithms

Most Popular

Free Workspace

$0/month

Don't already have a Microsoft account? Simply sign up here.

Sign In

- Free access that never expires
- 10 GB storage on us
- R and Python scripts support
- Predictive web services

Enterprise Grade

Standard Workspace

$9.99/month

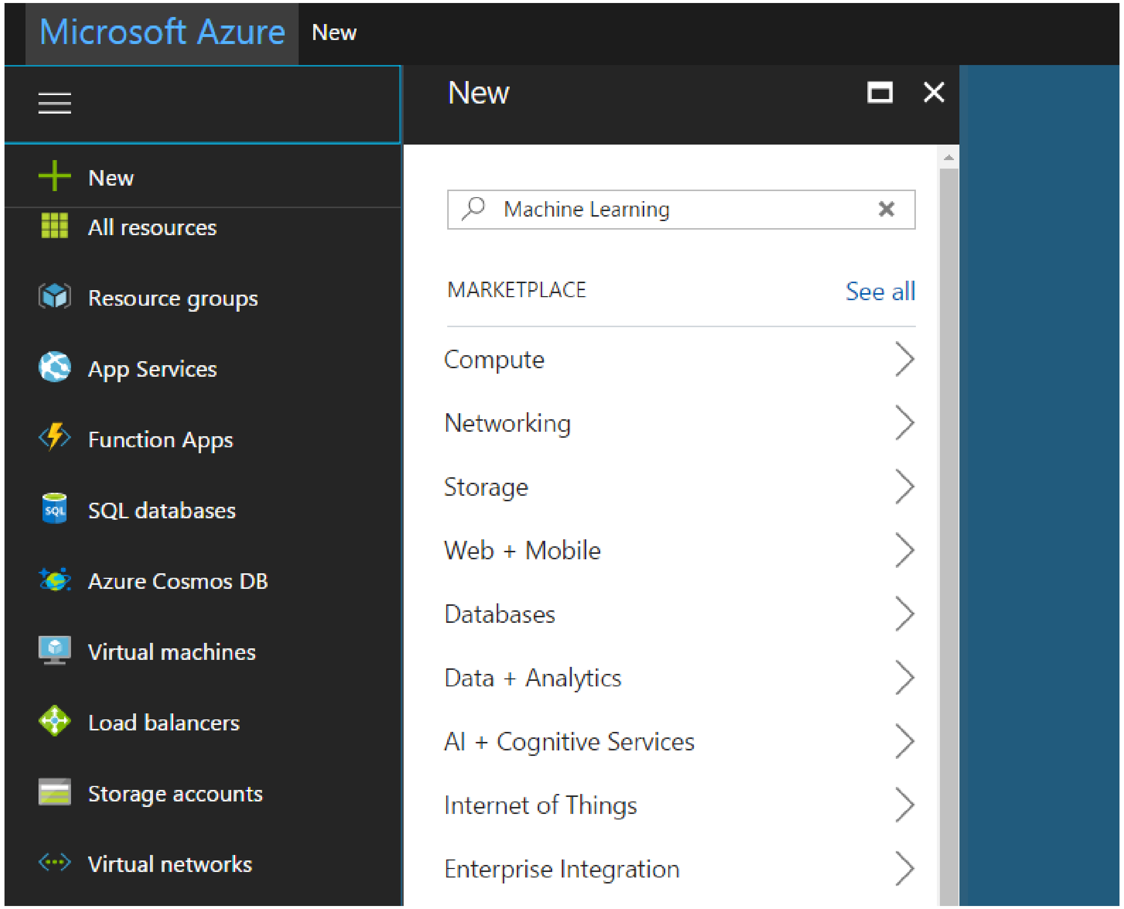
Azure subscription required
Other charges may apply. Read more.

Create Workspace

- Full SLA Support
- Bring your own Azure storage
- Parallel graph execution
- Elastic Web Service endpoints

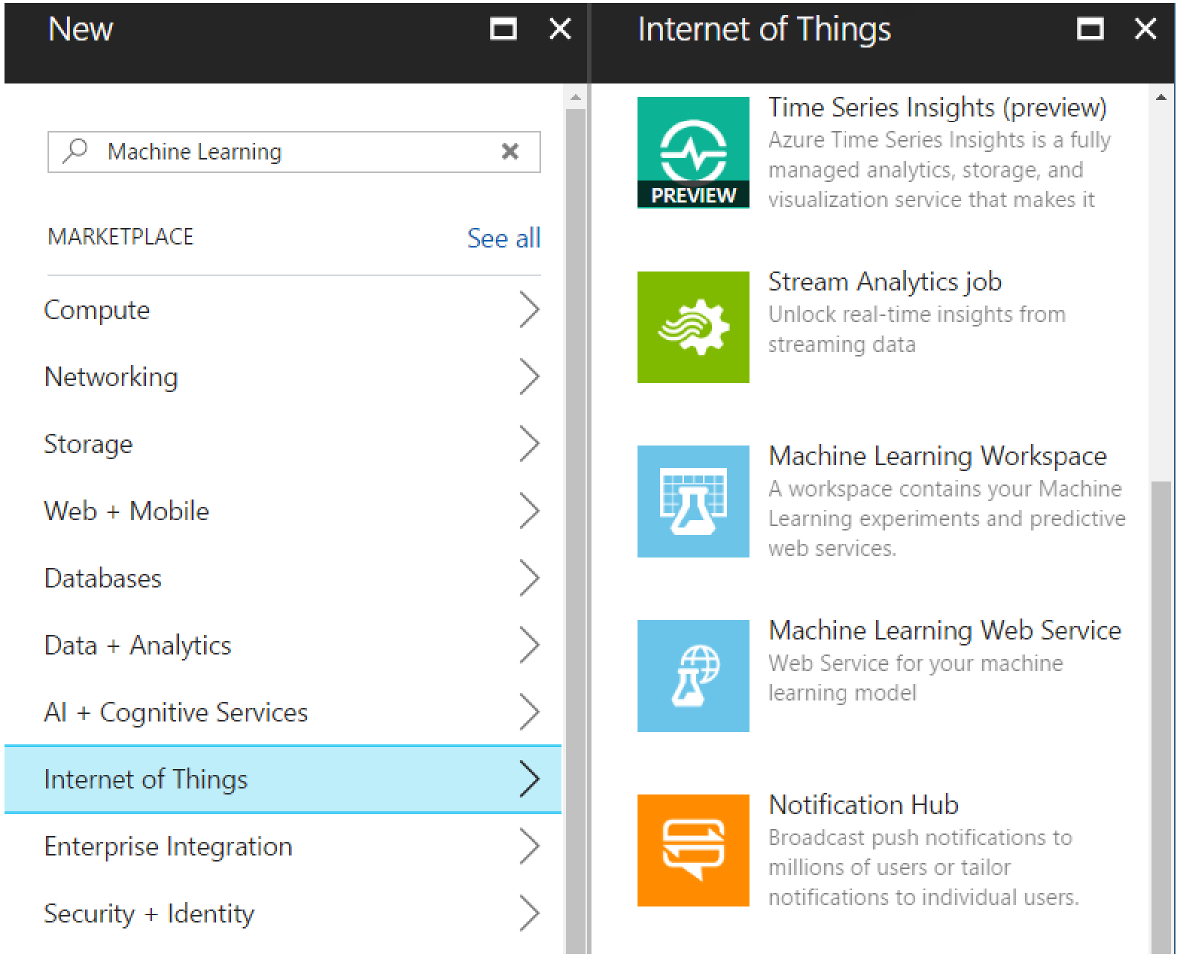
Create Azure ML Studio workspace

1. Go to the Azure portal https://portal.azure.com
2. Click +New



Create Azure ML Studio workspace

3. Select Internet of Things, click Machine Learning Workspace, then click Create



New

Machine Learning

MARKETPLACE — See all

- Compute
- Networking
- Storage
- Web + Mobile
- Databases
- Data + Analytics
- AI + Cognitive Services
- Internet of Things
- Enterprise Integration
- Security + Identity

Internet of Things

Time Series Insights (preview) — PREVIEW
Azure Time Series Insights is a fully managed analytics, storage, and visualization service that makes it

Stream Analytics job
Unlock real-time insights from streaming data

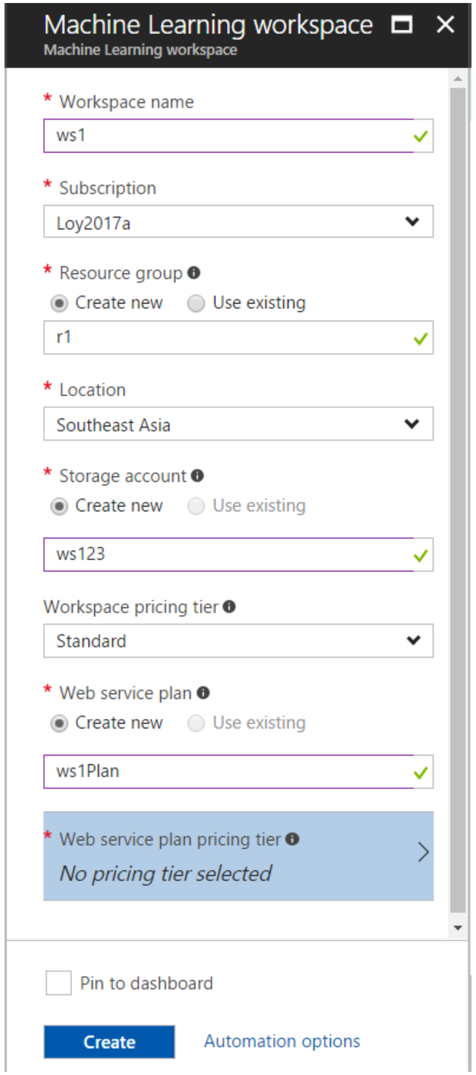
Machine Learning Workspace
A workspace contains your Machine Learning experiments and predictive web services.

Machine Learning Web Service
Web Service for your machine learning model

Notification Hub
Broadcast push notifications to millions of users or tailor notifications to individual users.

Create Azure ML Studio workspace

4. Workspace name = ws1
5. Subscription = defult
6. Resource group = Create new: rs1
7. Location = Southeast Asia
8. Storage account = Create new: names1
9. Workspace pricing tier = Standard
10. Web service plan = Create new: ws1Plan



Machine Learning workspace

Machine Learning workspace

* Workspace name: ws1 ✓

* Subscription: Loy2017a

* Resource group: ● Create new ○ Use existing

r1 ✓

* Location: Southeast Asia

* Storage account: ● Create new ○ Use existing

ws123 ✓

Workspace pricing tier: Standard

* Web service plan: ● Create new ○ Use existing

ws1Plan ✓

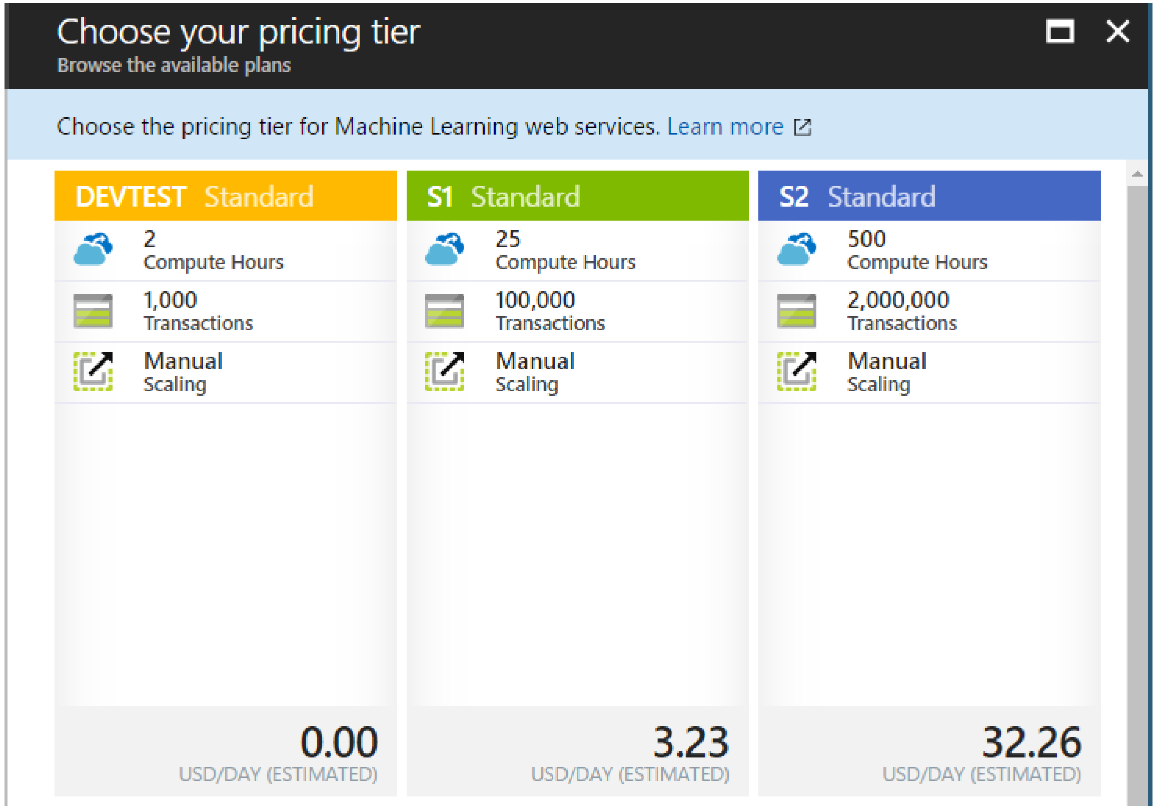
* Web service plan pricing tier

No pricing tier selected

□ Pin to dashboard

Create Azure ML Studio workspace

11. Click No pricing tier selected
12. Click DEVTEST
13. Click Pin to dashboard
14. Click Create



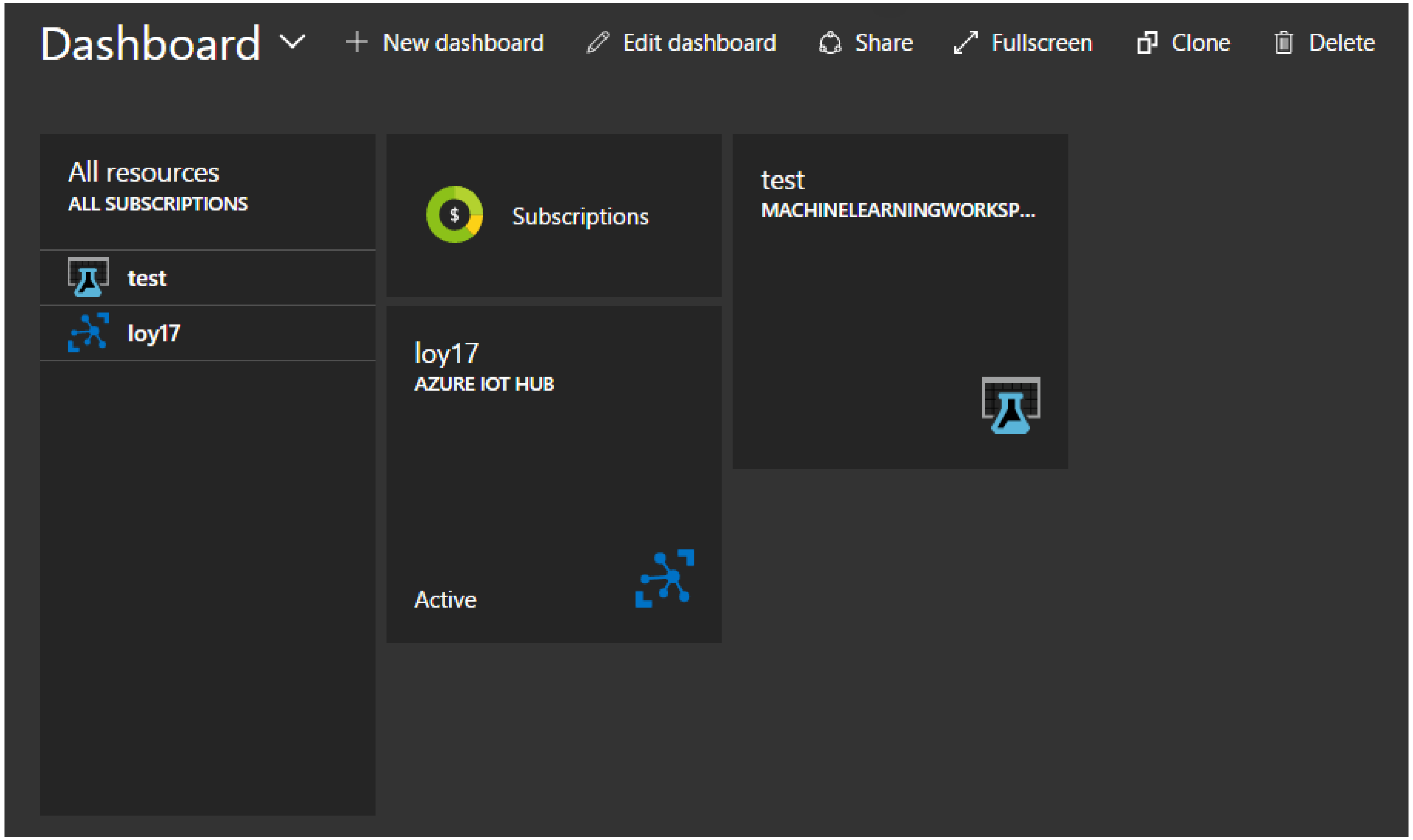
Choose your pricing tier

Browse the available plans

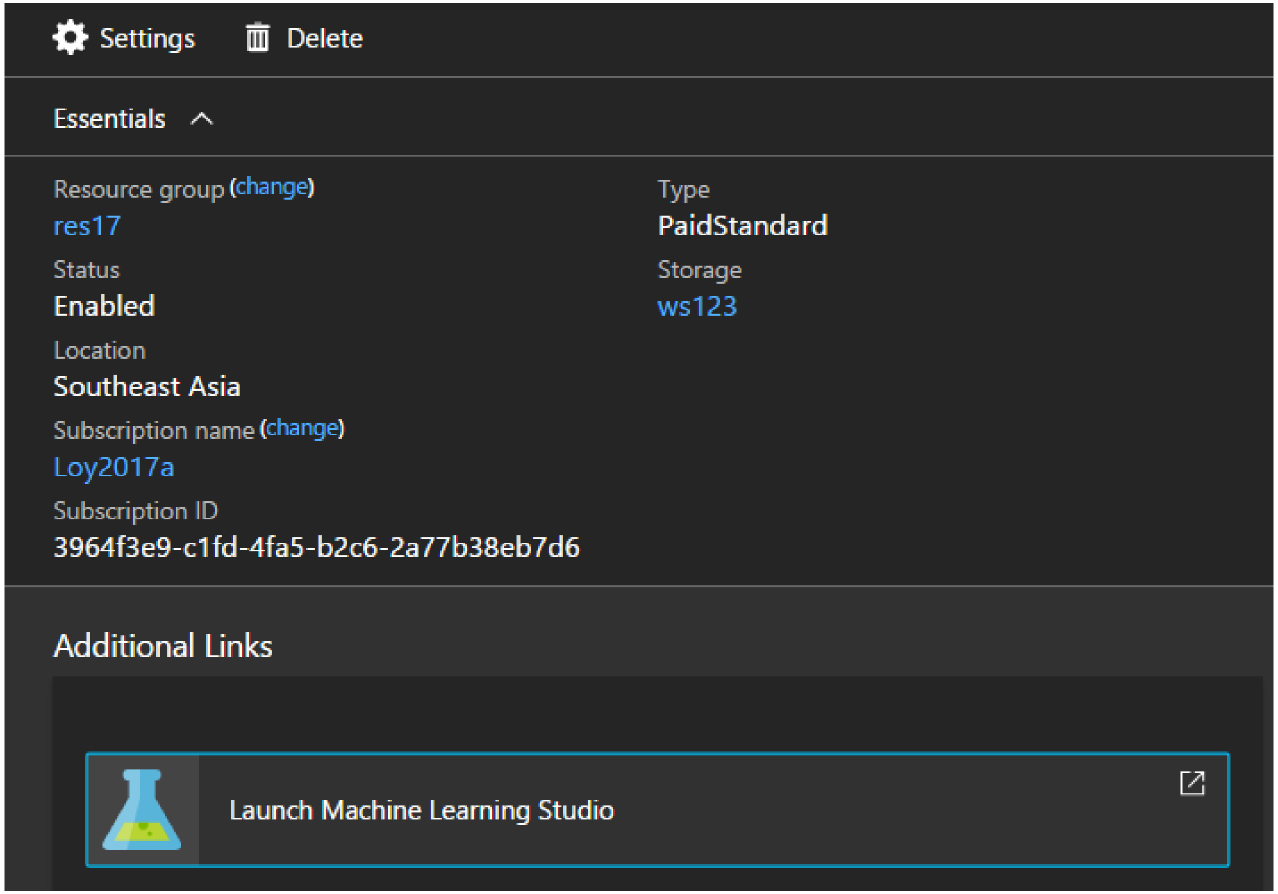
Choose the pricing tier for Machine Learning web services. Learn more

DEVTEST Standard	S1 Standard	S2 Standard
2 Compute Hours	25 Compute Hours	500 Compute Hours
1,000 Transactions	100,000 Transactions	2,000,000 Transactions
Manual Scaling	Manual Scaling	Manual Scaling
0.00 USD/DAY (ESTIMATED)	3.23 USD/DAY (ESTIMATED)	32.26 USD/DAY (ESTIMATED)

15. Click at Machine Learning workgroup on dashboard



16. Click Launch Machine Learning Studio



Resource group (change)
res17

Status
Enabled

Location
Southeast Asia

Subscription name (change)
Loy2017a

Subscription ID
3964f3e9-c1fd-4fa5-b2c6-2a77b38eb7d6

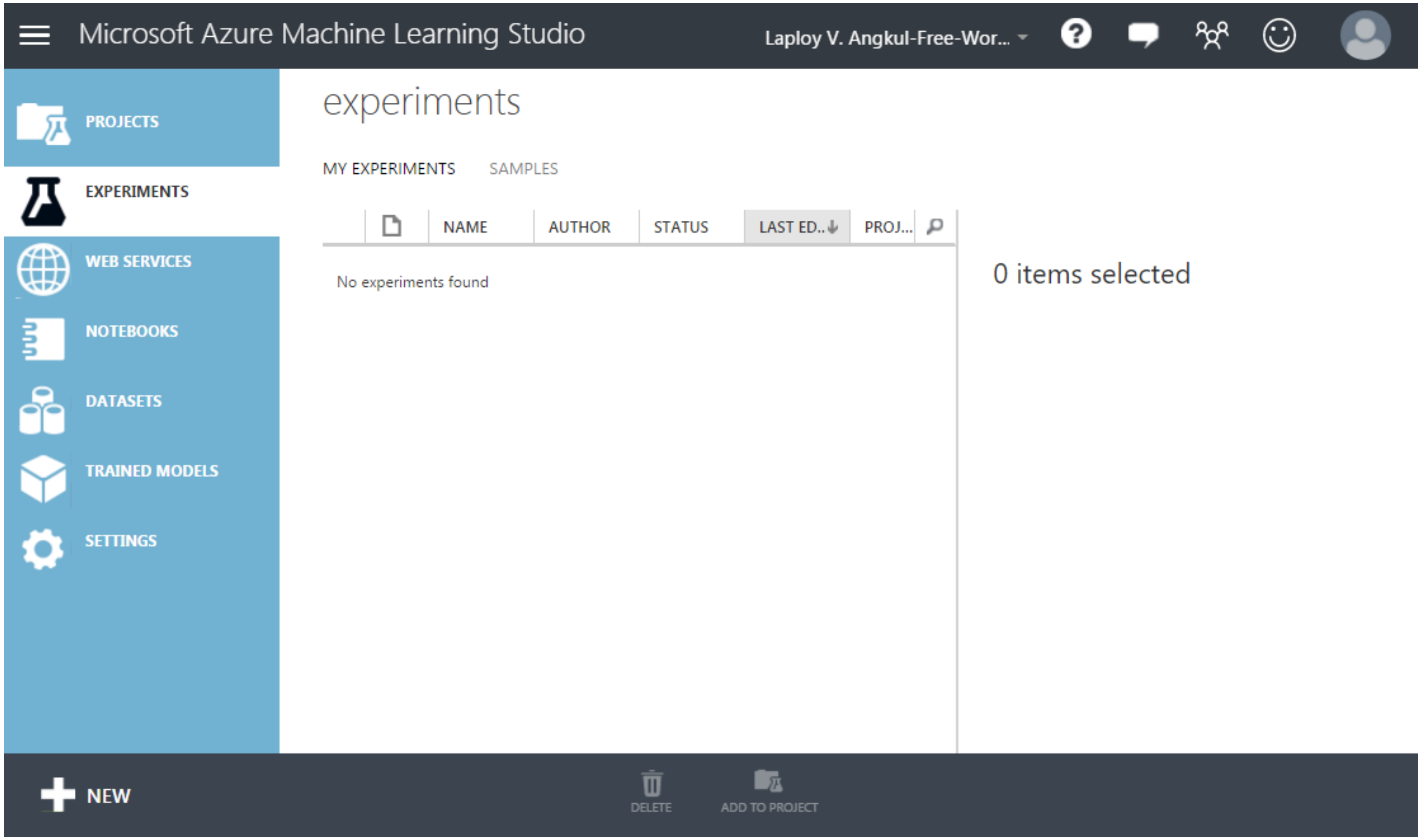
Type
PaidStandard

Storage
ws123

Additional Links

Create Azure ML Studio workspace

Blank, new ML Studio workspace



No experiments found

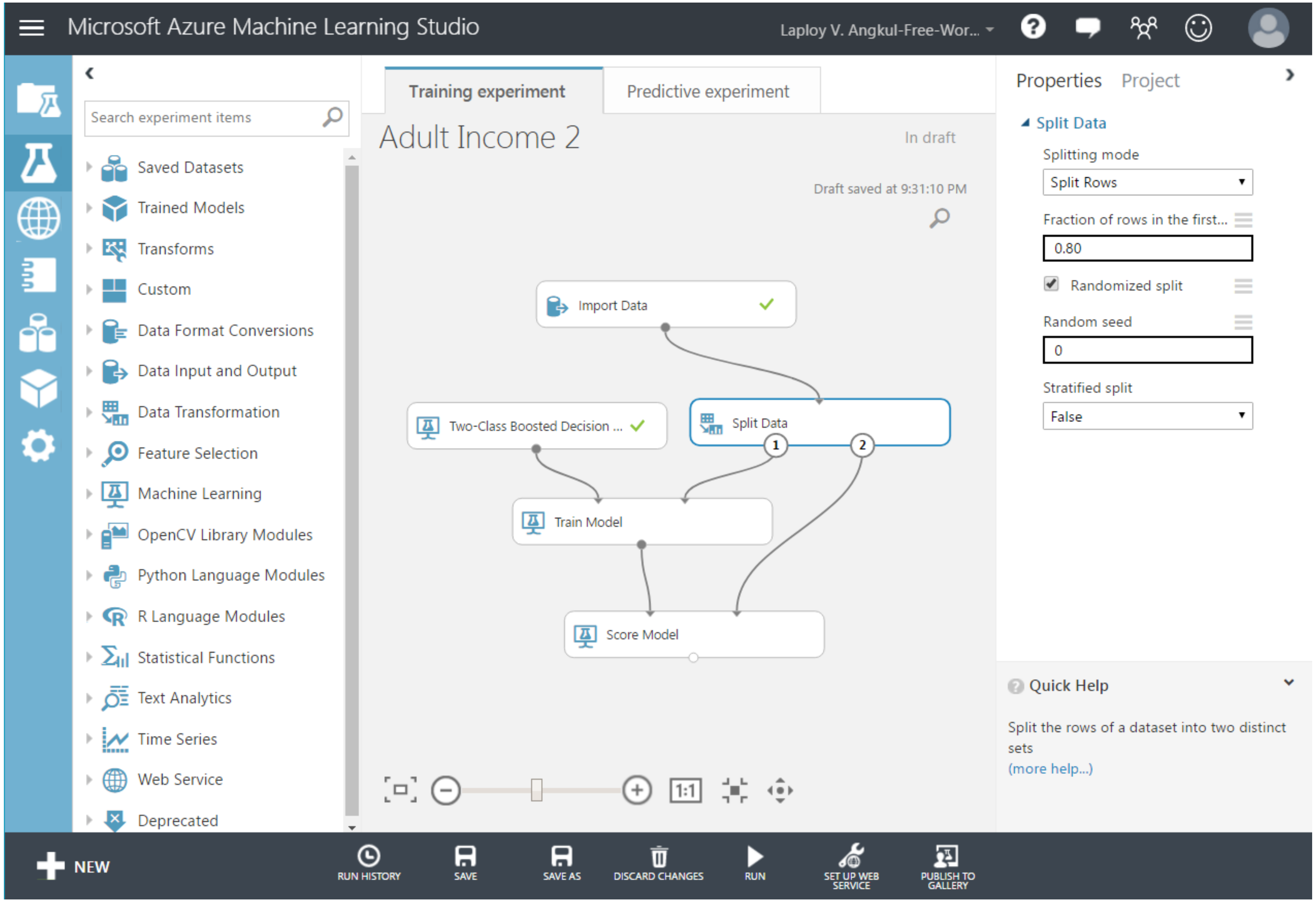
0 items selected

Train, Test, Evaluate for Binary Classification

Predicting whether a person’s income exceeds $50,000 per year based on his demographics or census data

1. Download, prepare, and upload a census income dataset.
2. Create a new Azure Machine Learning experiment.
3. Train and evaluate a prediction model.

The overall workflow of the experiment



Split Data

Splitting mode: Split Rows

Fraction of rows in the first...: 0.80

☑ Randomized split

Random seed: 0

Stratified split: False

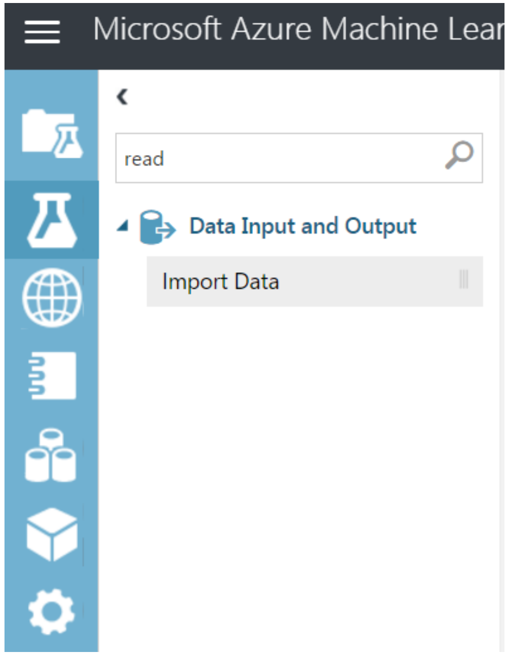
Quick Help

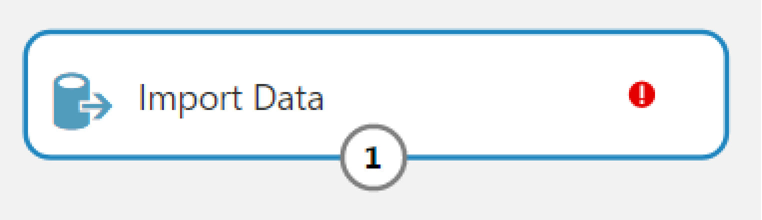
Split the rows of a dataset into two distinct sets

(more help...)

Train, Test, Evaluate for Binary Classification

- Create New blank experiment. Name = Adult Income 1
- Click Data Input and Output
- Drag & drop Import Data

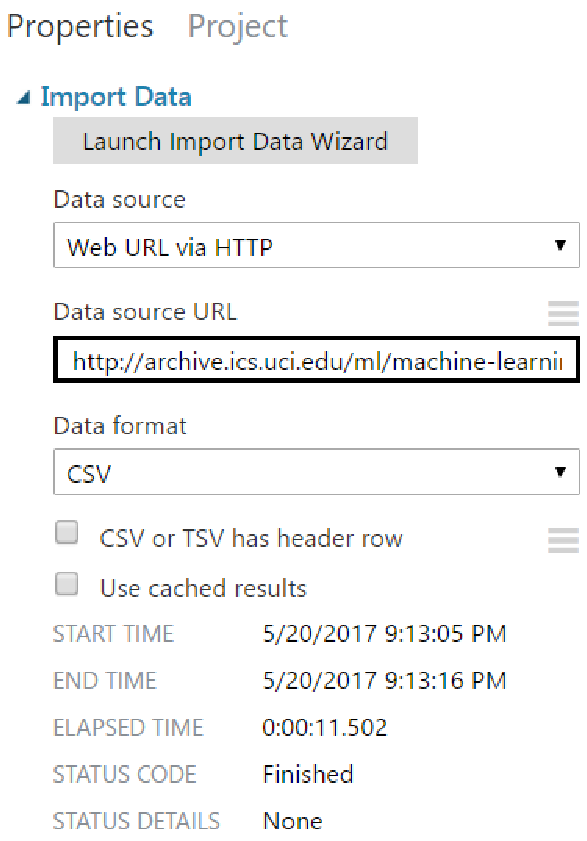




Train, Test, Evaluate for Binary Classification

Configure Import data module:

- Data source = Web URL via HTTP
- Data source URL = http://archive.ics.uci.edu/ml/machine-learning-databases/adult/adult.data
- Data format = CSV
- Run experiment



Data source

Web URL via HTTP ▾

Data source URL

http://archive.ics.uci.edu/ml/machine-learni

Data format

CSV ▾

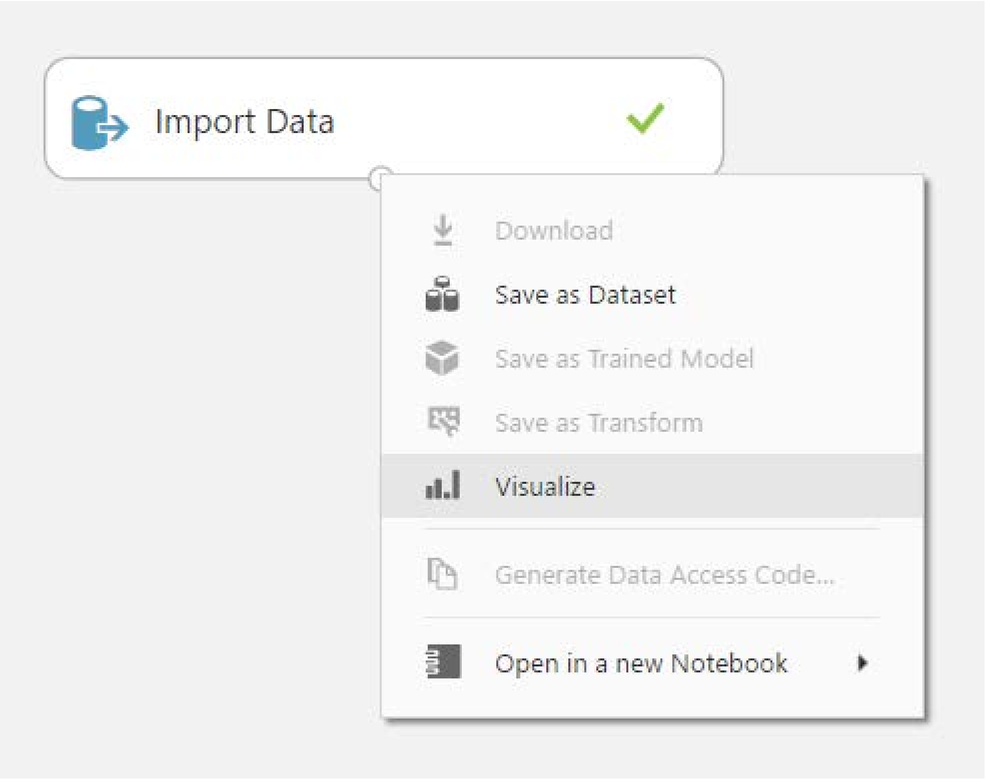
□ CSV or TSV has header row

□ Use cached results

START TIME	5/20/2017 9:13:05 PM
END TIME	5/20/2017 9:13:16 PM
ELAPSED TIME	0:00:11.502
STATUS CODE	Finished
STATUS DETAILS	None

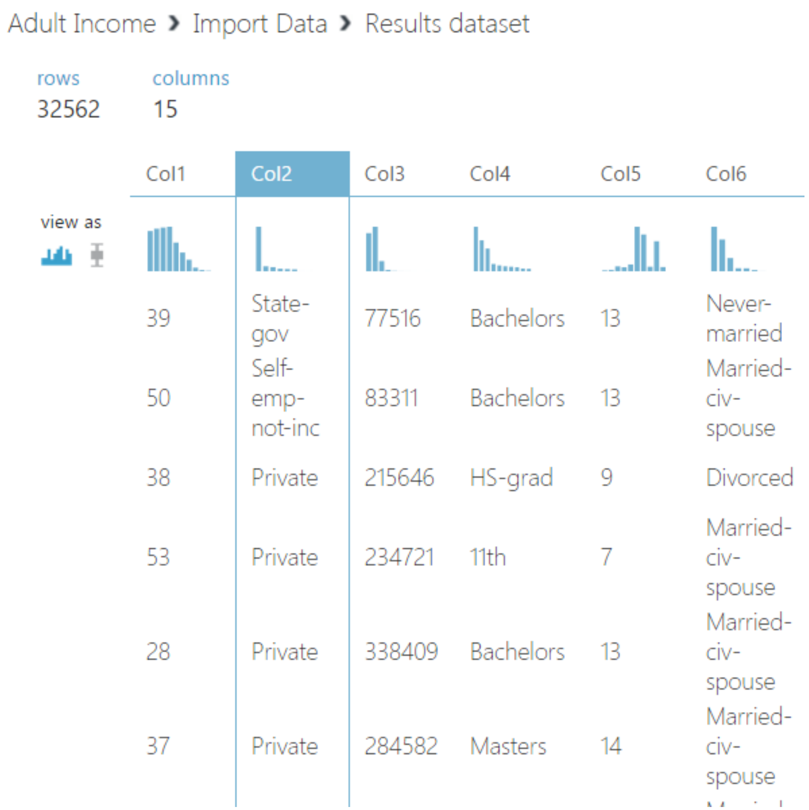
Train, Test, Evaluate for Binary Classification

- Right click at the output of Import Data
- Click Visualize



Train, Test, Evaluate for Binary Classification

- Click on Col2
- Look at Statistics and Histogram

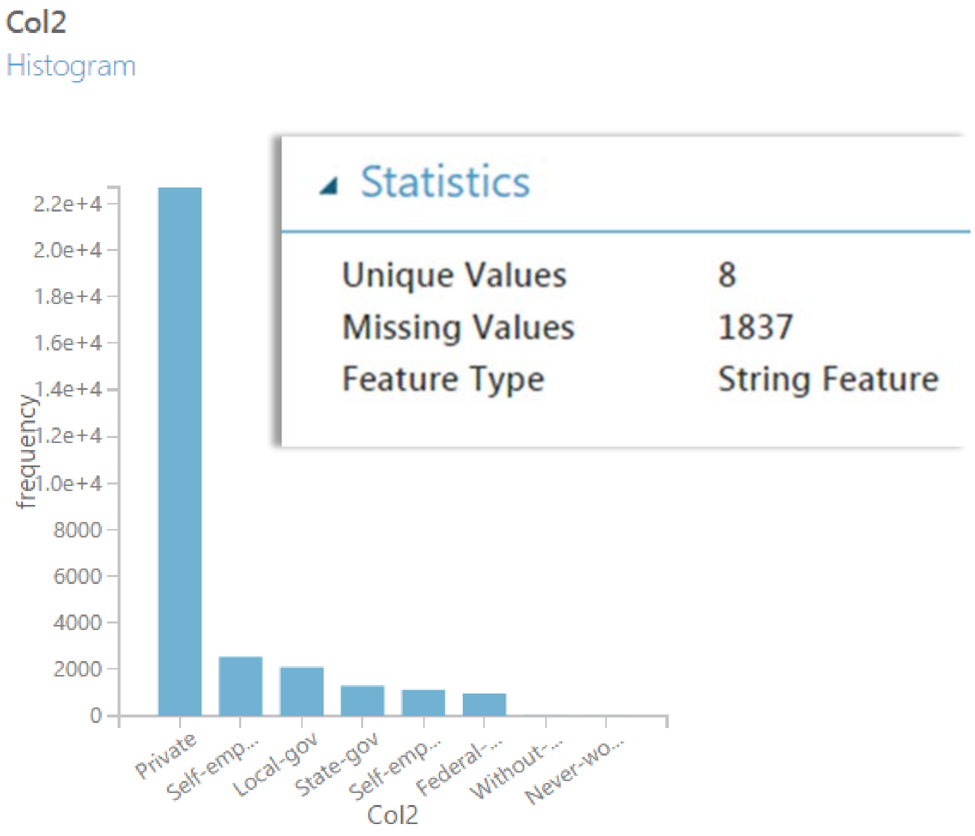


Adult Income › Import Data › Results dataset

rows	columns
32562	15

view as

Col1	Col2	Col3	Col4	Col5	Col6
39	State-gov	77516	Bachelors	13	Never-married
50	Self-emp-not-inc	83311	Bachelors	13	Married-civ-spouse
38	Private	215646	HS-grad	9	Divorced
53	Private	234721	11th	7	Married-civ-spouse
28	Private	338409	Bachelors	13	Married-civ-spouse
37	Private	284582	Masters	14	Married-civ-spouse



Col2

Histogram

Statistics

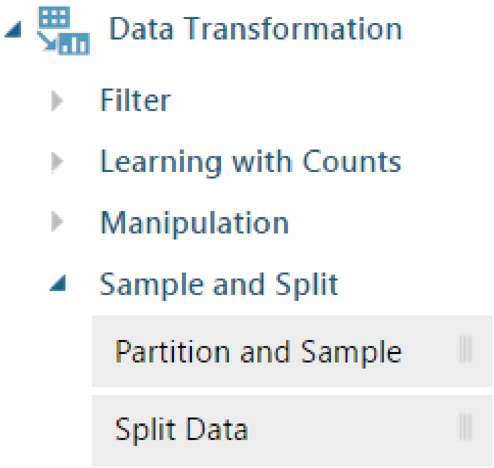
Unique Values	8
Missing Values	1837
Feature Type	String Feature

Split up the dataset

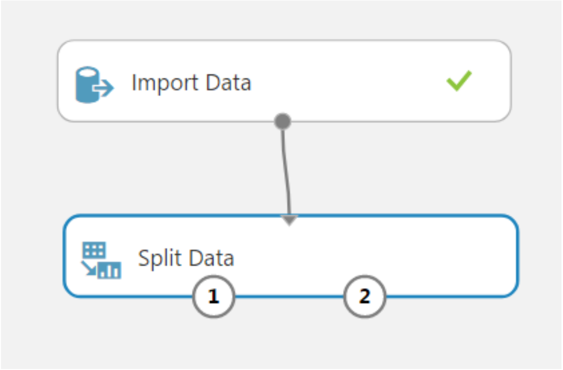
- Training data: This grouping is used for creating our new predictive model based on the inherent patterns found in the historical data via the ML algorithm we use for the solution.
- Validation data: This grouping is used for testing the new predictive model against known outcomes to determine accuracy and probabilities.

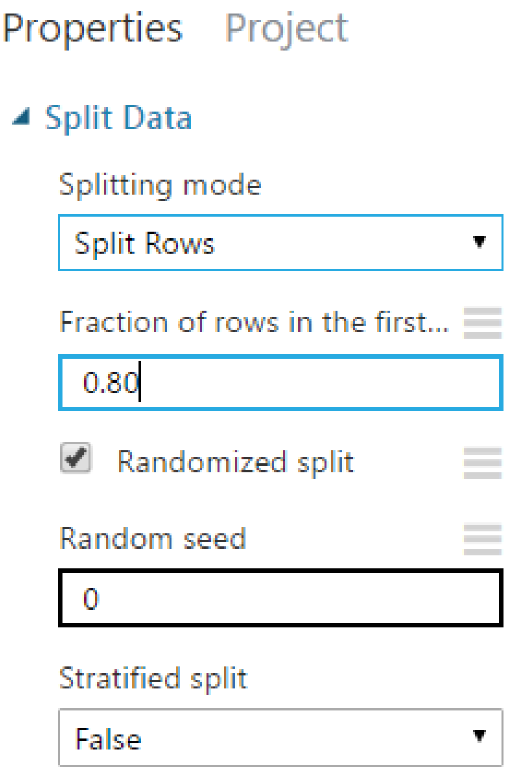
Add Split Data:

- Click Data Transformation
- Click Sample and Split
- Drag & drop Split Data module into canvas
- Connect Import Data to Split Data
- Set properties Fraction of row to 0.80



- Data Transformation
  - Filter
  - Learning with Counts
  - Manipulation
  - Sample and Split
    - Partition and Sample
    - Split Data





Properties Project

Split Data

Splitting mode: Split Rows

Fraction of rows in the first...: 0.80

☑ Randomized split

Random seed: 0

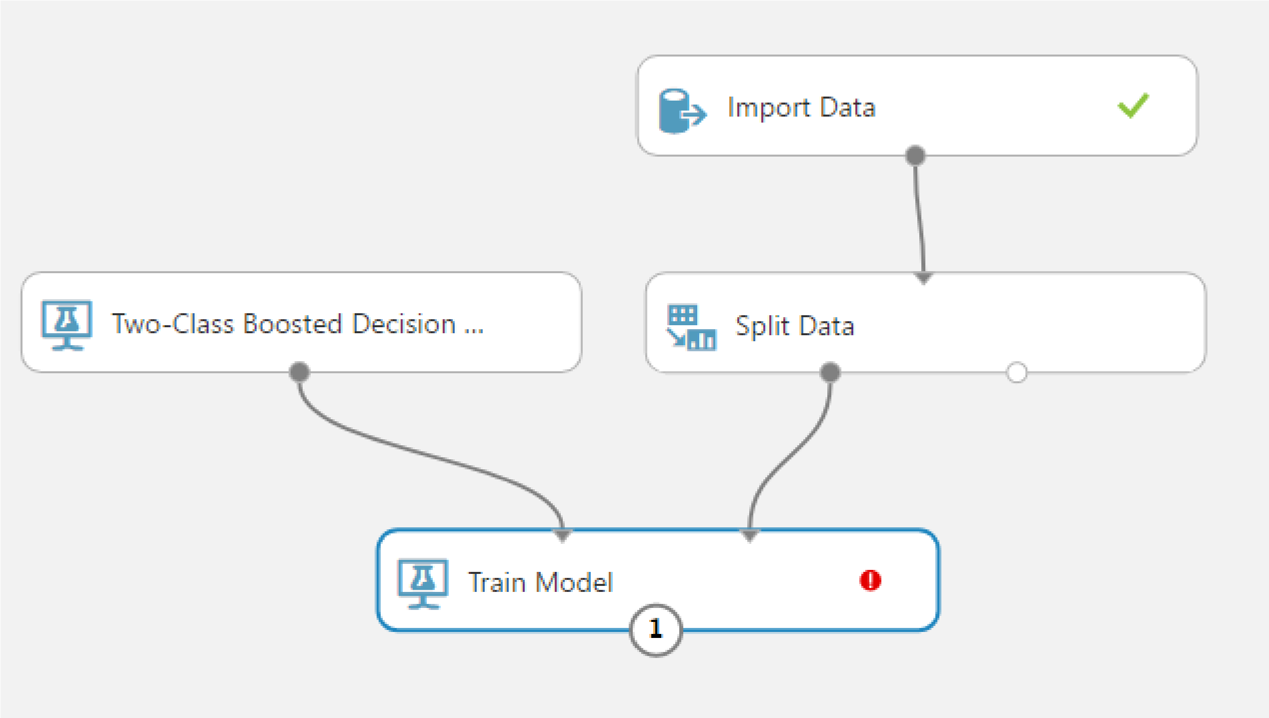
Stratified split: False

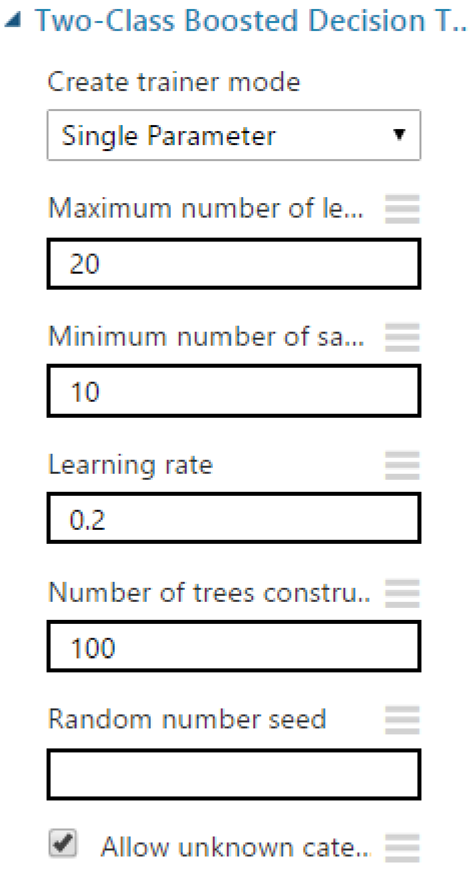
Train, Test, Evaluate for Binary Classification

Add Two-Class Boosted Decision Tree and Train Model

Connect Two-Class Boosted Decision Tree to Train Model

Connect Split Data to Train Model





◢ Two-Class Boosted Decision T..

Create trainer mode: Single Parameter ▾

Maximum number of le...: 20

Minimum number of sa...: 10

Learning rate: 0.2

Number of trees constru..: 100

Random number seed:

☑ Allow unknown cate..

Train, Test, Evaluate for Binary Classification

Click Train Model

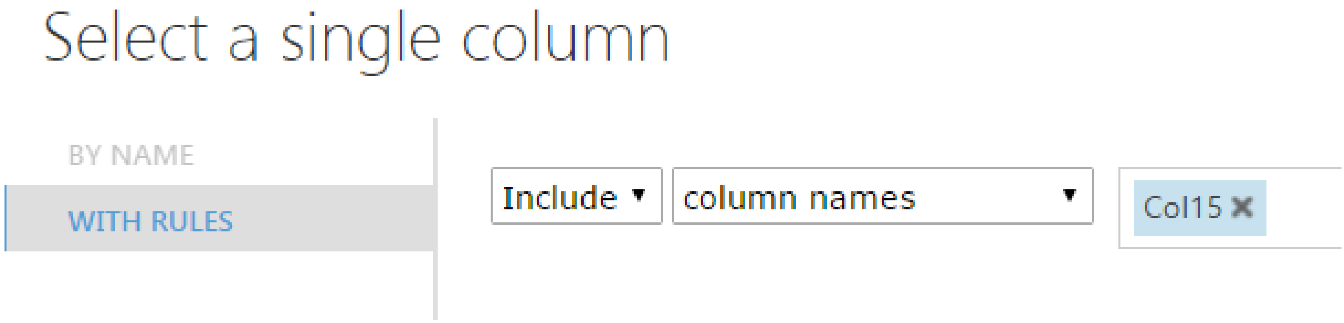
Click Launch column selector

Include col15

Click ✓

Save

Run



Select a single column

Include ▾ column names ▾ Col15 ✖

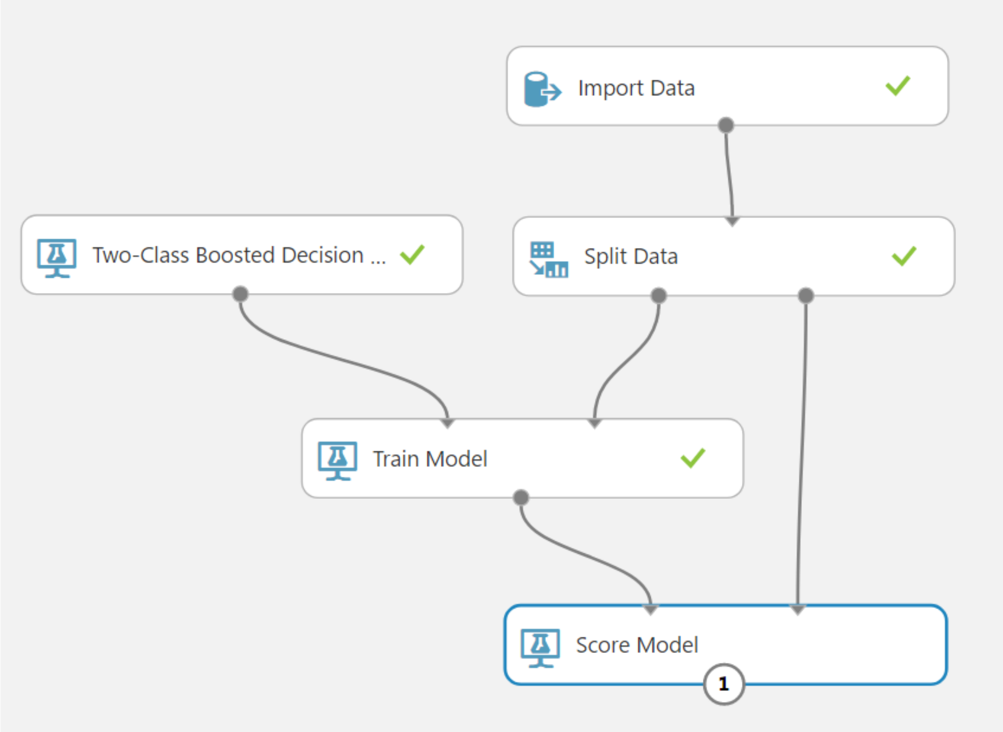
Score the model

Score the model:

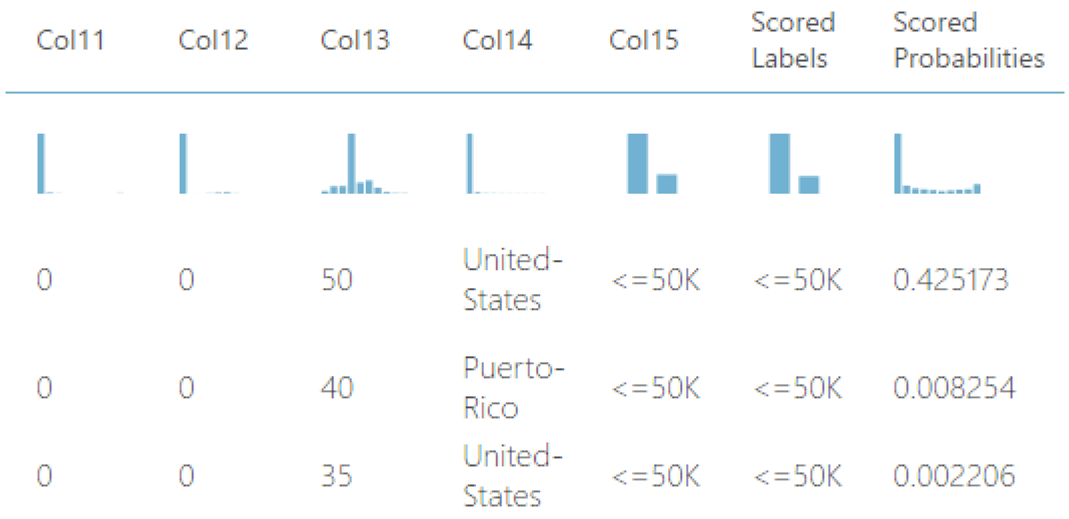
Add Score Model to canvas

Connect Score Model to Train and Split model

Run



Visualize the model results



Col11	Col12	Col13	Col14	Col15	Scored Labels	Scored Probabilities
0	0	50	United-States	<=50K	<=50K	0.425173
0	0	40	Puerto-Rico	<=50K	<=50K	0.008254
0	0	35	United-States	<=50K	<=50K	0.002206

Visualize the model results:

Visualize output of Score Model

Scored Labels This column denotes the model's prediction for this row of the dataset.

Scored Probabilities This column denotes the numerical probability (or the likelihood) of whether the income level for this row exceeds $50,000.

Type of datasets

Training set

- A set of examples used for learning
- Where the answer value is known.

Validation set

- A set of examples data
- Used to tune the architecture of a classifier
- And estimate the error

Test set

- Use to test the performances of a classifier
- Never used during the training process
- Give estimate of error

More Information