

INTRO TO DATA SCIENCE USING AZURE ML STUDIO

Laura Da Silva, Founder of Da Silva Advanced Analytics,
Founder of IWDS and Microsoft MVP in AI
@lauradatasci @wogisci

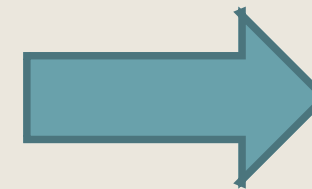
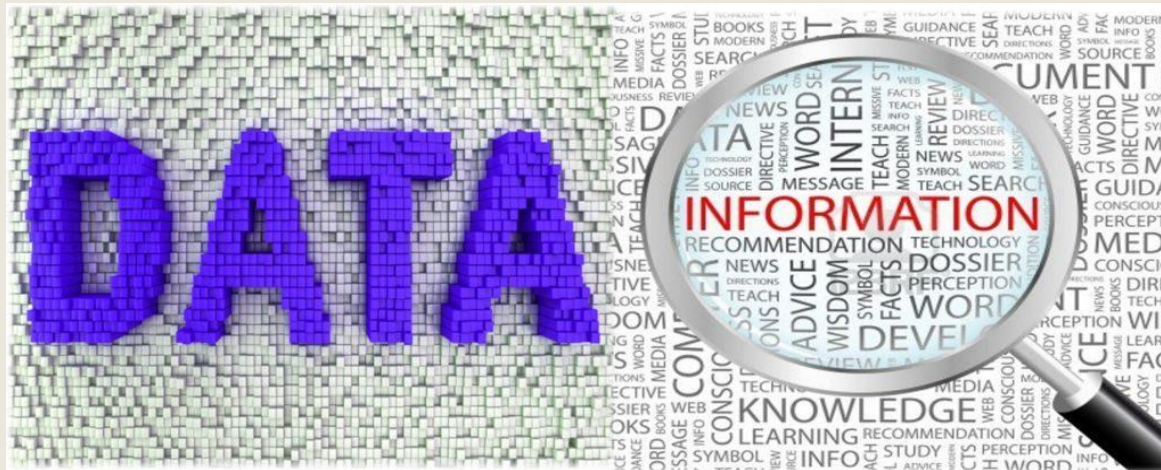
Topics for today:

- Intro to data science
- Data science applications
- Intro to machine learning
- Data science life cycle
- Getting started with Azure ML Studio

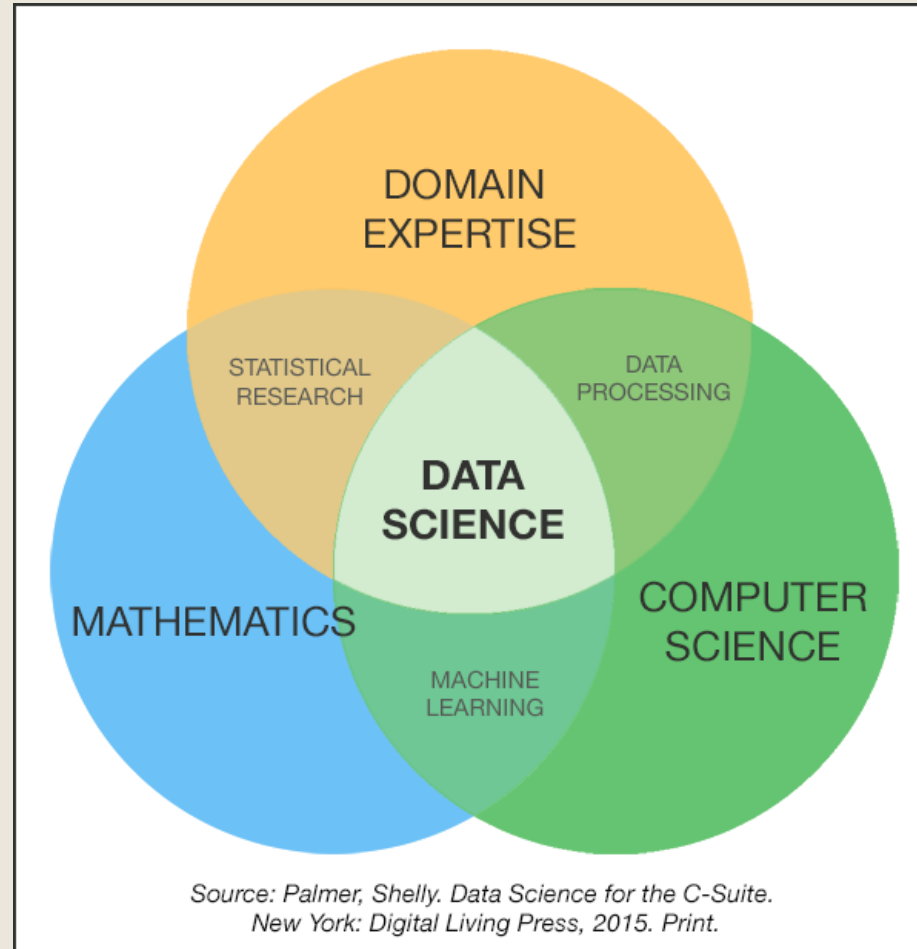
INTRO TO DATA SCIENCE



What is Data Science?



What to learn to become a Data Scientist?



MODERN DATA SCIENTIST

Data Scientist, the sexiest job of the 21st century, requires a mixture of multidisciplinary skills ranging from an intersection of mathematics, statistics, computer science, communication and business. Finding a data scientist is hard. Finding people who understand who a data scientist is, is equally hard. So here is a little cheat sheet on who the modern data scientist really is.

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



What skills are expected from a Data Scientist?

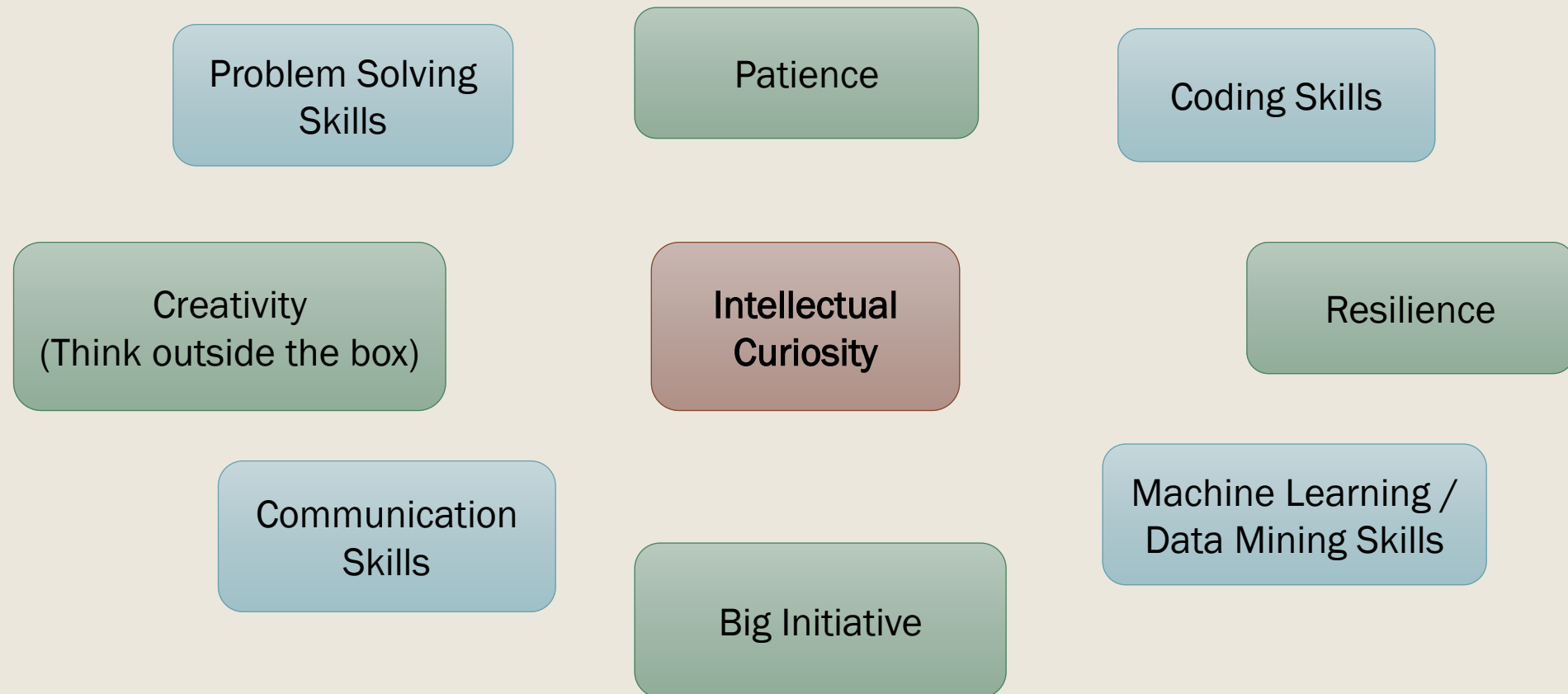
Experience coding
(SQL, R, Python)

Machine Learning
(Maths & Statistics)

Curious about
data

Good communicator:
Translate data-driven
insights into
decisions and actions

General skills for becoming a Data Scientist



Because in your every day you will be ...

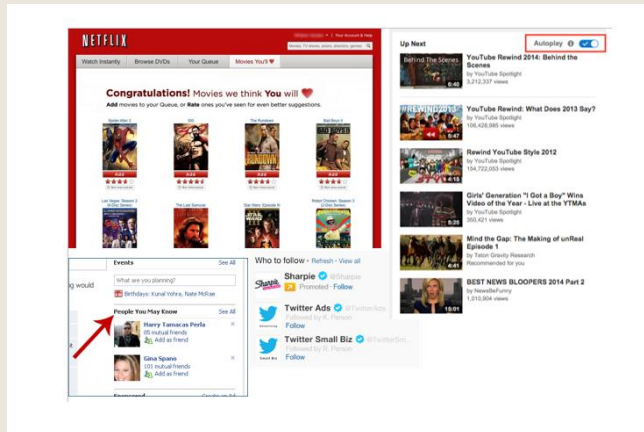
- *Solving problems* for getting the best solution for your client/company
- *Understanding data* and business objectives
- *Transforming data* to give it a coherent and useful form
- Implementing new *algorithms* to make predictions on data
- Studying and using *models* to understand specific data
- Using feedback to learn how to improve your solution or make better predictions

DATA SCIENCE APPLICATIONS



Data Science Applications

Current Trends



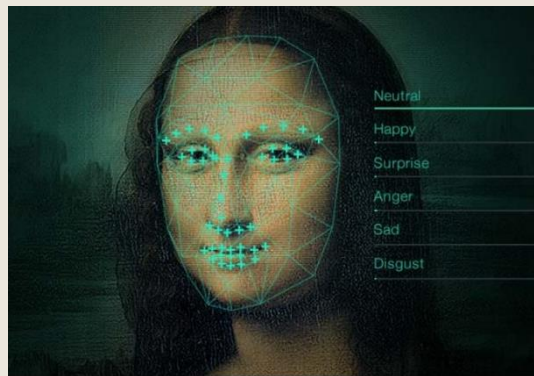
Recommender Systems



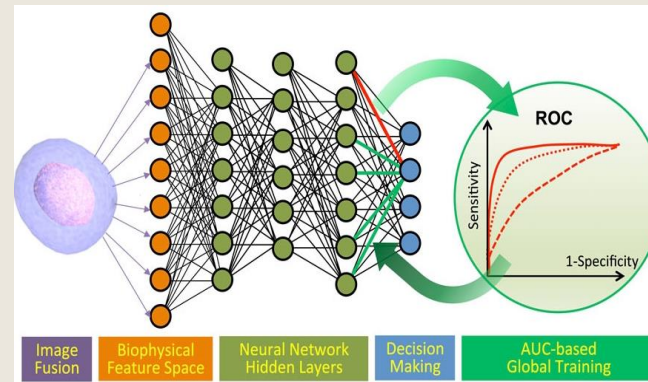
Speech recognition



Fraud detection



Facial identification/recognition



Health (detection of cancer)



Self Driving cars



Smart houses

Data Science Applications



<https://youtu.be/hRUvWdUZyF0>

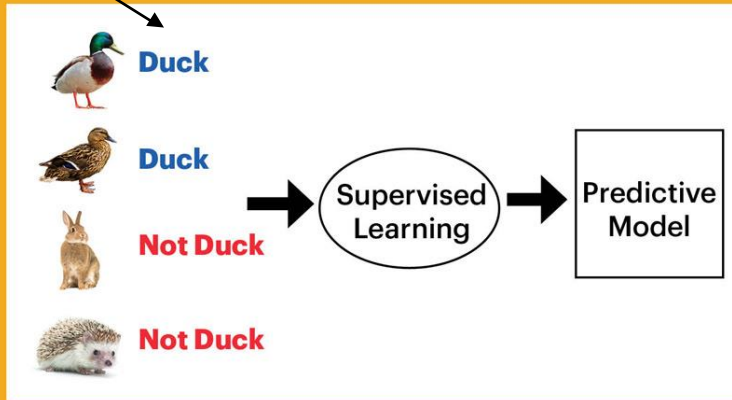
INTRO TO MACHINE LEARNING



Intro to Machine learning

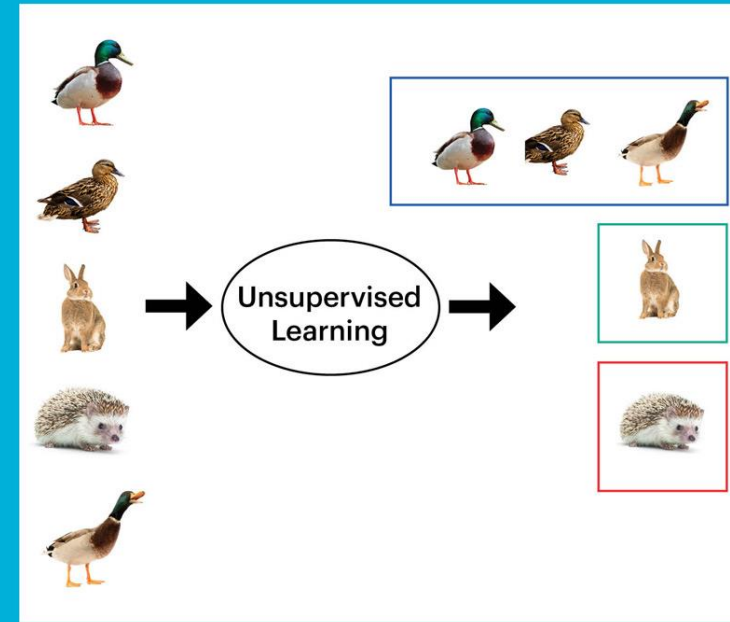
With labels

Supervised Learning (Classification Algorithm)

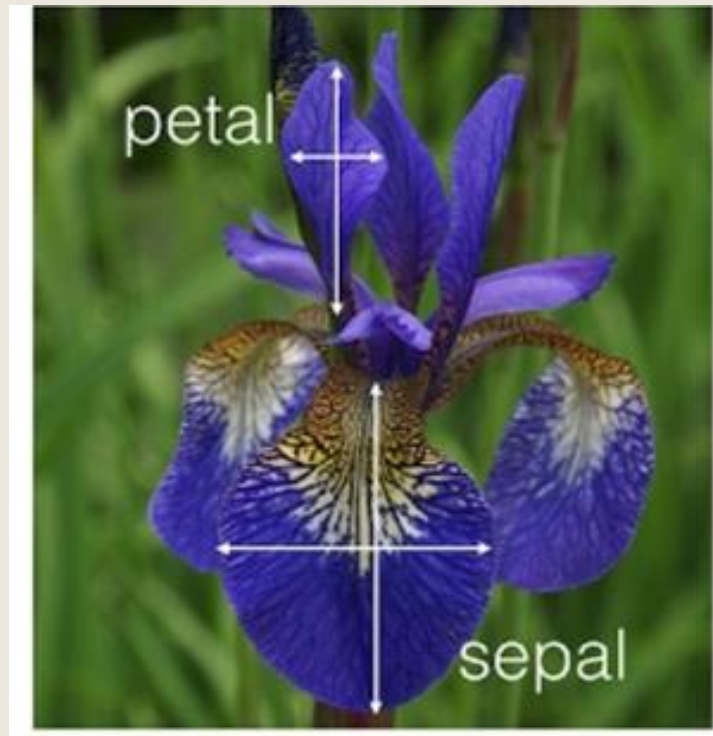


Unsupervised Learning (Clustering Algorithm)

Without labels



Example: Iris flower



Features				Labels
Sepal length	Sepal width	Petal length	Petal width	Species
5.1	3.5	1.4	0.2	Iris setosa
4.9	3.0	1.4	0.2	Iris setosa
7.0	3.2	4.7	1.4	Iris versicolor
6.4	3.2	4.5	1.5	Iris versicolor
6.3	3.3	6.0	2.5	Iris virginica
5.8	3.3	6.0	2.5	Iris virginica



Iris Versicolor



Iris Setosa



Iris Virginica

Unsupervised Learning

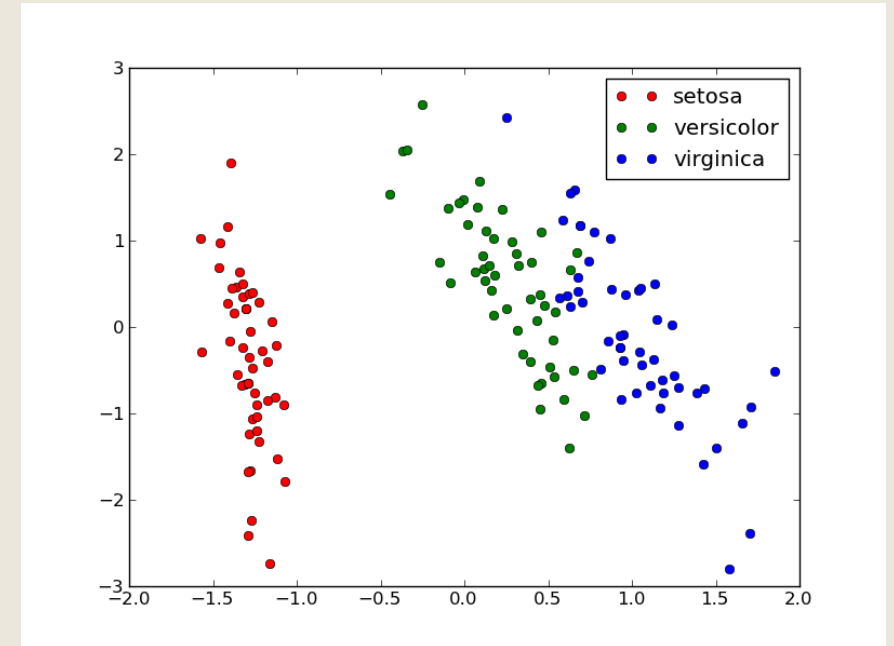
Definition

- You give the input data (X) and no corresponding output variables (labels).

Techniques

- Clustering: you want to discover the inherent groupings in the data.
- Association: you want to discover rules that describe large portions of your data.

Features				Labels
Sepal length	Sepal width	Petal length	Petal width	Species
5.1	3.5	1.4	0.2	Iris setosa
4.9	3.0	1.4	0.2	Iris setosa
7.0	3.2	4.7	1.4	Iris versicolor
6.4	3.2	4.5	1.5	Iris versicolor
6.3	3.3	6.0	2.5	Iris virginica
5.8	3.3	6.0	2.5	Iris virginica



Supervised Learning

Definition

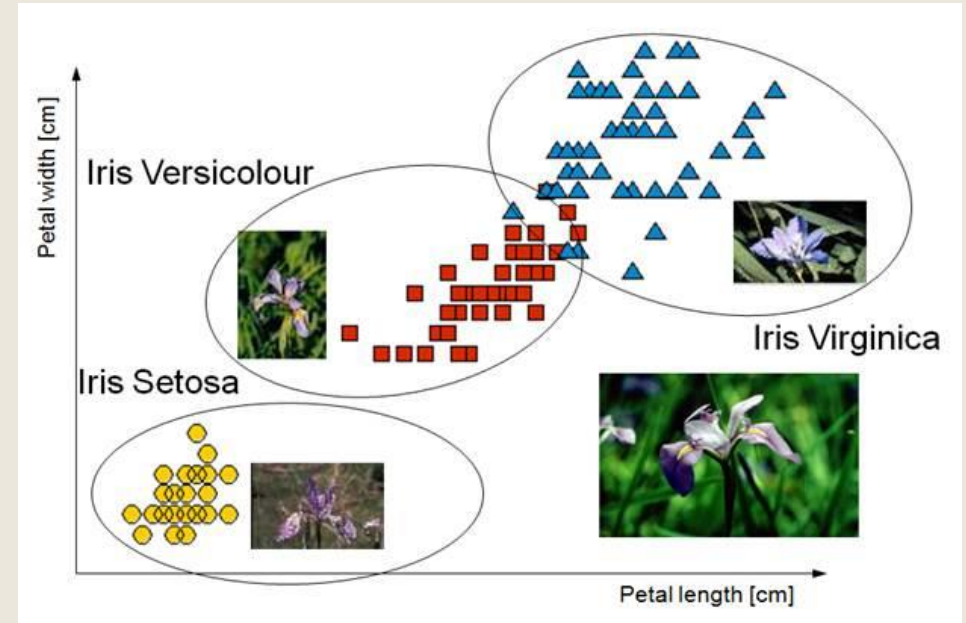
- You give the input data (X) and an output variable (Y) (labels), and you use an algorithm to learn the mapping function from the input to the output.

$$Y = f(X)$$

Techniques

- Classification: you want to classify a new input value.
- Regression: you want to get a function that fit really well the data so you can predict the future.

Features				Labels
Sepal length	Sepal width	Petal length	Petal width	Species
5.1	3.5	1.4	0.2	Iris setosa
4.9	3.0	1.4	0.2	Iris setosa
7.0	3.2	4.7	1.4	Iris versicolor
6.4	3.2	4.5	1.5	Iris versicolor
6.3	3.3	6.0	2.5	Iris virginica
5.8	3.3	6.0	2.5	Iris virginica



DATA SCIENCE LIFE CYCLE



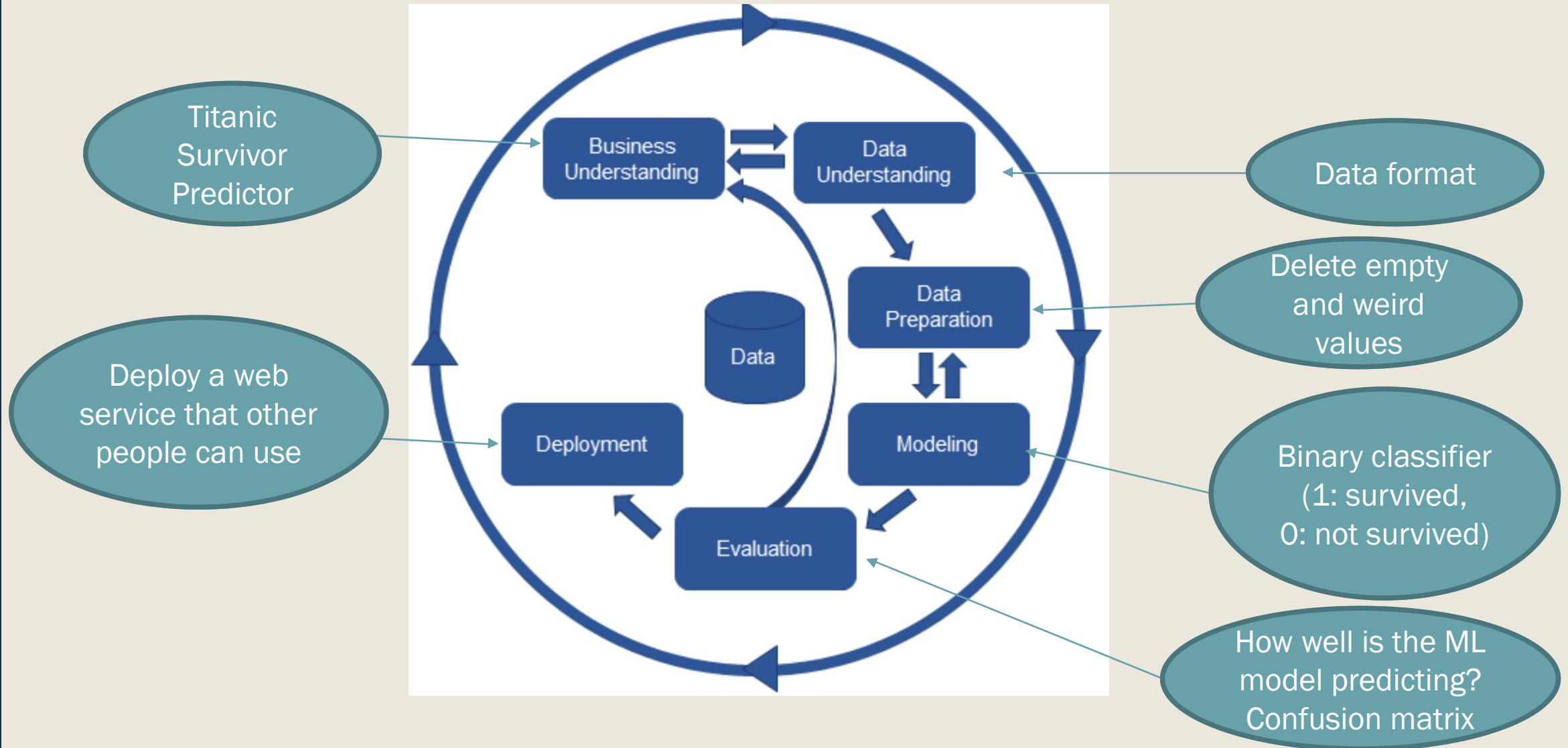
Supervised learning example: Titanic survivor predictor

The titanic survivor predictor example will predict the probability of surviving in the Titanic based on data about passengers on the Titanic.

1. Let's explore the dataset
2. Let's follow the instructions together to get our predictions



Data Science cycle



The ORIGINAL Dataset

The original dataset usually has to be transformed in order to be used

	PassengerId	Survived	Pclass	Name	Sex	Age	SibSp	Parch	Ticket	Fare	Cabin	Embarked
37	37	1	3	Mamee, Mr. Hanna	male	NA	0	0	2677	7.2292		C
38	38	0	3	Cann, Mr. Ernest Charles	male	21.00	0	0	A/5. 2152	8.0500		S
39	39	0	3	Vander Planke, Miss. Augusta Maria	female	18.00	2	0	345764	18.0000		S
40	40	1	3	Nicola-Yarred, Miss. Jamila	female	14.00	1	0	2651	11.2417		C
41	41	0	3	Ahlin, Mrs. Johan (Johanna Persdotter Larsson)	female	40.00	1	0	7546	9.4750		S
42	42	0	2	Turpin, Mrs. William John Robert (Dorothy Ann Wonna...	female	27.00	1	0	11668	21.0000		S
43	43	0	3	Kraeff, Mr. Theodor	male	NA	0	0	349253	7.8958		C
44	44	1	2	Laroche, Miss. Simonne Marie Anne Andree	female	3.00	1	0	SC/Paris 2123	41.5792		C
45	45	1	3	Devaney, Miss. Margaret Delia	female	18.00	0	0	330958	7.8792		Q
46	46	0	3	Rogers, Mr. William John	male	NA	0	0	S.C./A.4. 23567	8.0500		S
47	47	0	3	Lennon, Mr. Denis	male	NA	1	0	370371	15.5000		Q
48	48	1	3	O'Driscoll, Miss. Bridget	female	NA	0	0	14311	7.7500		Q
49	49	0	3	Samaan, Mr. Youssef	male	NA	2	0	2662	21.6792		C
50	50	0	3	Arnold-Franchi, Mrs. Josef (Josefine Franchi)	female	18.00	1	0	349237	17.8000		S
51	51	0	3	Panula, Master. Juha Niilo	male	7.00	4	1	3101295	39.6875		S
52	52	0	3	Nosworthy, Mr. Richard Cater	male	21.00	0	0	A/4. 39886	7.8000		S
53	53	1	1	Harper, Mrs. Henry Sleeper (Myna Haxtun)	female	49.00	1	0	PC 17572	76.7292	D33	C

Data Dictionary

Variable Name	Description
Survived	Survived (1) or died (0)
Pclass	Passenger's class
Name	Passenger's name
Sex	Passenger's sex
Age	Passenger's age
SibSp	Number of siblings/spouses aboard
Parch	Number of parents/children aboard
Ticket	Ticket number
Fare	Fare
Cabin	Cabin
Embarked	Port of embarkation

Final dataset

Attributes or features

Observations

A	B	C	D	E	F	G
PassengerClass	Gender	Age	SiblingSpouse	ParentChild	FarePrice	PortEmbarkation
1	male	22	1	1	7.25	S
1	female	38	1	1	71.28	C
2	female	40	3	0	34.5	S
1	male	12	0	0	51.6	C
1	female	3	1	2	18.4	Q
2	male	45	1	0	11.3	C
1	female	23	2	2	16.7	Q
3	male	16	0	0	30.1	C
1	female	21	2	1	20.3	S
2	male	58	1	1	18.8	C

Numerical Data

Categorical Data

Splitting the data for training and testing

70%

Class	Sex	Age	SibSp	Parch	Fare	Embarked	FamilySize	Survived

30%

Class	Sex	Age	SibSp	Parch	Fare	Embarked	FamilySize

Model evaluation: the confusion matrix

		prediction outcome		
		p	n	total
actual value	p'	True Positive	False Negative	P'
	n'	False Positive	True Negative	N'
total		P	N	

$n=165$		Predicted: NO	Predicted: YES
Actual: NO		50	10
Actual: YES		5	100

AZURE ML STUDIO



Create your first experiment in Azure Machine Learning Studio

Sign in Azure Machine Learning Studio: <https://studio.azureml.net/>

Let's follow the tutorial of creating an experiment in Azure ML for Income prediction.

Details about the dataset can be found at:
<https://archive.ics.uci.edu/ml/datasets/adult>

The training experiment

The screenshot displays the Microsoft Azure Machine Learning Studio interface. The top navigation bar shows the user 'Laura da Silva-Free-Works...' and various utility icons. The left sidebar contains a search bar and a categorized list of experiment items: Saved Datasets, Samples, Trained Models, Data Format Conversions, Data Input and Output, Data Transformation (with sub-items like Filter, Learning with Counts, and Manipulation), Sample and Split (with sub-items like Partition and Sample and Split Data), Scale and Reduce, Feature Selection, Machine Learning (with sub-items like Evaluate, Cross Validate Model, Evaluate Model, and Evaluate Recommen...), and Initialize Model (with sub-item Anomaly Detection). The main workspace is titled 'Income Prediction' and shows a workflow diagram with the following steps: 'Adult Census Income Binary...' (data source), 'Split Data' (data transformation), 'Two-Class Boosted Decision...' (model), 'Train Model' (training), 'Score Model' (evaluation), and 'Evaluate Model' (evaluation). All steps are marked with a green checkmark, indicating they are completed. The status 'Finished running' is shown in the top right of the workspace. The right sidebar contains the 'Properties' and 'Project' tabs, with 'Experiment Properties' showing 'START TIME' (8/17/2018 11:28:52 ...), 'END TIME' (8/17/2018 11:29:17 ...), 'STATUS CODE' (Finished), and 'STATUS DETAILS' (None). Below this is a 'Summary' section with a text input field and a 'Description' section with another text input field. At the bottom, a 'Quick Help' section is visible. The bottom navigation bar includes icons for 'NEW', 'RUN HISTORY', 'SAVE', 'SAVE AS', 'DISCARD CHANGES', 'RUN', 'SET UP WEB SERVICE', and 'PUBLISH TO GALLERY'.

Microsoft Azure Machine Learning Studio

Laura da Silva-Free-Works...

Training experiment Predictive experiment

Income Prediction

Finished running ✓

Search experiment items

Saved Datasets

- Samples
- Trained Models
- Data Format Conversions
- Data Input and Output
- Data Transformation
 - Filter
 - Learning with Counts
 - Manipulation
- Sample and Split
 - Partition and Sample
 - Split Data
- Scale and Reduce
- Feature Selection
- Machine Learning
 - Evaluate
 - Cross Validate Model
 - Evaluate Model
 - Evaluate Recommen...
 - Initialize Model
 - Anomaly Detection

Adult Census Income Binary...

Split Data

Two-Class Boosted Decision...

Train Model

Score Model

Evaluate Model

Experiment Properties

START TIME 8/17/2018 11:28:52 ...

END TIME 8/17/2018 11:29:17 ...

STATUS CODE Finished

STATUS DETAILS None

Summary

Enter a few sentences describing your experiment (up to 140 characters).

Description

Enter the detailed description for your experiment.

Quick Help

NEW

RUN HISTORY

SAVE

SAVE AS

DISCARD CHANGES

RUN

SET UP WEB SERVICE

PUBLISH TO GALLERY

The predictive experiment

The screenshot displays the Microsoft Azure Machine Learning Studio interface. The top navigation bar includes the studio name, a user profile, and various utility icons. The left sidebar contains a search bar and a categorized list of experiment items. The main workspace shows a predictive experiment titled 'Income Prediction [Predictive Exp.]' with a 'Finished running' status. The workflow diagram includes nodes for data input, model scoring, and web service output. The right sidebar provides details about the experiment's properties, summary, and description.

Microsoft Azure Machine Learning Studio

Laura da Silva-Free-Works...

Training experiment Predictive experiment

Income Prediction [Predictive Exp.] Finished running ✓

Search experiment items

- Saved Datasets
 - Samples
 - Trained Models
 - Data Format Conversions
 - Data Input and Output
- Data Transformation
 - Filter
 - Learning with Counts
 - Manipulation
- Sample and Split
 - Partition and Sample
 - Split Data
- Scale and Reduce
- Feature Selection
- Machine Learning
 - Evaluate
 - Cross Validate Model
 - Evaluate Model
 - Evaluate Recommen...
 - Initialize Model
 - Anomaly Detection

Adult Census Income Binary...

Web service input

Income Prediction [trained ...]

Score Model ✓

Web service output

Properties Project

Experiment Properties

START TIME	8/17/2018 11:42:01 ...
END TIME	8/17/2018 11:42:05 ...
STATUS CODE	Finished
STATUS DETAILS	None

[Go to web service](#)

Summary

Enter a few sentences describing your experiment (up to 140 characters).

Description

Enter the detailed description for your experiment.

Quick Help

NEW RUN HISTORY SAVE SAVE AS DISCARD CHANGES RUN DEPLOY WEB SERVICE PUBLISH TO GALLERY

The web service details

Microsoft Azure Machine Learning Studio

Laura da Silva-Free-Works...

?

income prediction [predictive exp.]

DASHBOARD CONFIGURATION

General [New Web Services Experience](#) preview

Published experiment

[View snapshot](#) [View latest](#)

Description

No description provided for this web service.

API key

5zczsHCxfLeJd/VjVS+6MdrKKip14jsl/XVdZDMTQJLi8zCc8i0/An/akk5upG/p2chF91IV2cynaOR+Ti+A==

Default Endpoint

API HELP PAGE	TEST	APPS	LAST UPDATED
REQUEST/RESPONSE	Test Test preview	Excel 2013 or later Excel 2010 or earlier workbook	8/17/2018 11:42:34 AM
BATCH EXECUTION	Test preview	Excel 2013 or later workbook	8/17/2018 11:42:34 AM

NEW

DELETE

Challenges

Looking at examples you will find in the Gallery (<https://gallery.azure.ai/>):

- Create the experiment for the Titanic survivor predictor.
- Create the experiment for Sentiment analysis using messages from Twitter.
- Create an experiment for movie recommendation.