



DATA SCIENCE BOOTCAMP

MODULE 1

■ Day 1

- Module 1 – Introduction to Data Science
- Module 2 – Machine Learning concepts
- Module 3 – Ingestion, Preparation & EDA

■ Day 2

- Module 4 - Creating Models on AZURE ML STUDIO
- Module 5 – Deploying AZURE ML Models

■ Day 3

- Module 6 – Data Science Competition

Module 1 Outline



Microsoft



- Overview of Data Science
- Introduction to Big data
- Big data Platform
- Data Analytics
 - **Descriptive Analytics**
 - **Diagnostic Analytics**
 - **Predictive Analytics**
 - **Prescriptive Analytics**
- Azure ML Studio

Overview of Data Science



Microsoft



- Data science incorporates varying elements and builds on techniques and theories from many fields, including mathematics, statistics, data engineering, pattern recognition and learning, advanced computing, visualization, uncertainty modeling, data warehousing, and high performance computing with the goal of extracting meaning from data and creating data products.
- A practitioner of data science is called a data scientist.

What is a Data Scientist ?



Microsoft



- "A data scientist is somebody who is inquisitive, who can stare at data and spot trends."
- The data scientist will sift through all incoming data with the goal of discovering a previously hidden insight, which in turn can provide a competitive advantage or address a pressing business problem.

Introduction to Big Data



Microsoft



- Big Data refers to datasets whose size is beyond the ability of typical database software tools to capture, store, manage and analyze.
- The demand for deep analytical positions in a big world could exceed the supply being produced on current trends by 140K to 190K positions.
- A need for 1.5 million additional managers and analysts in the US who can ask the right questions and consume the results of the analysis of big data effectively.

The seven V's sum it up pretty well – Volume, Velocity, Variety, Variability, Veracity, Visualization, and Value.

- ✓ **Volume**: Volume is how much data we have.
- ✓ **Velocity**: Velocity is the measure of how fast the data is coming in.
- ✓ **Variety**: More sources of data means more varieties of data in different formats.

- ✓ **Variability**: Variability is different from variety. It means that the meaning of the data is constantly changing.
- ✓ **Veracity**: Veracity is all about making sure the data is accurate.
- ✓ **Visualization**: Visualization is critical in today's world. Using charts and graphs to visualize large amounts of complex data is much more effective in conveying meaning than spreadsheets and reports chock-full of numbers and formulas.
- ✓ **Value**: Value is the end game. After addressing volume, velocity, variety, variability, veracity, and visualization – which takes a lot of time, effort and resources – you want to be sure your organization is getting value from the data.



Top 10 Commercial Hadoop Platforms Vendor

1. Cloudera
2. Microsoft HDInsight
3. Amazon Web Services
4. Hortonworks
5. MapR
6. IBM
7. Intel Distribution for Apache Hadoop
8. Datastax Enterprise Analytics
9. Teradata Enterprise Access for Hadoop
10. Pivotal HD

Big data Platform (Cont.)



Big Data Platform Component	
Apache Hadoop	Apache Oozie
Apache Hbase	Apache Hive
Apache Avro	Kite SDK
Apache ZooKeeper	Apache Crunch
Apache Impala	HUE
Apache DataFu	Apache Parquet
Apache Mahout	Apache Spark
Apache Sentry	Apache Ranger
Apache Sqoop	Apache Pig
Apache Flume	Apache Kafka

Data Analytics

- Data analytics is a process of inspecting, cleansing, transforming, and modeling data with the goal of discovering useful information, suggesting conclusions, and supporting decision-making

Four types of Data Analytics



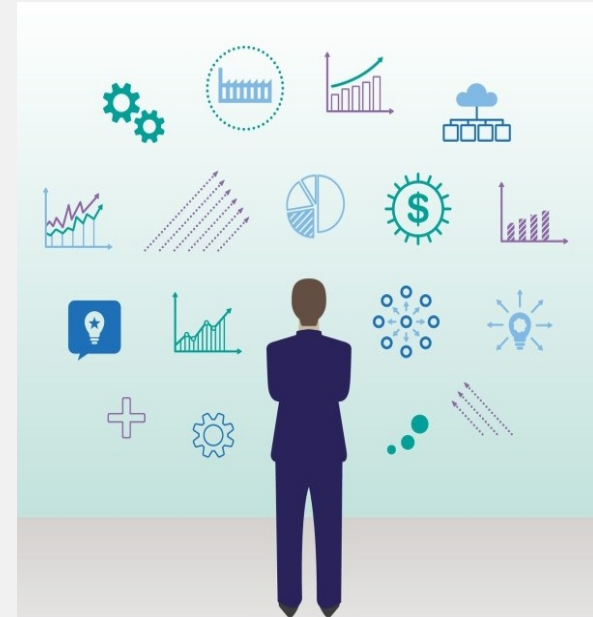
Microsoft



Analytic Excellence Leads to Better Decisions

Descriptive Analytics : What happened?

- This is the most common of all forms. In business it provides the analyst a view of key metrics and measures within the business.
- An examples of this could be a monthly profit and loss statement. Similarly, an analyst could have data on a large population of customers. Understanding demographic information on their customers (e.g. 30% of our customers are self-employed) would be categorized as "descriptive analytics". Utilizing effective visualization tools enhances the message of descriptive analytics.



Diagnostic: Why is it happening?

- This is the next step of complexity in data analytics is diagnostic analytics. On assessment of the descriptive data, diagnostic analytical tools will empower an analyst to drill down and in so doing isolate the root-cause of a problem.
- Well-designed business information (BI) dashboards incorporating reading of time-series data (i.e. data over multiple successive points in time) and featuring filters and drill down capability allow for such analysis.



Predictive: What is likely to happen?

- Predictive analytics is all about forecasting. Whether it's the likelihood of an event happening in future, forecasting a quantifiable amount or estimating a point in time at which something might happen – these are all done through predictive models.
- Predictive models typically utilize a variety of variable data to make the prediction. The variability of the component data will have a relationship with what it is likely to predict (e.g. the older a person, the more susceptible they are to a heart-attack – we would say that age has a linear correlation with heart-attack risk). These data are then compiled together into a score or prediction.
- In a world of great uncertainty, being able to predict allows one to make better decisions. Predictive models are some of the most important utilized across a number of fields.



Data Analytics



Microsoft



Prescriptive: What do I need to do?

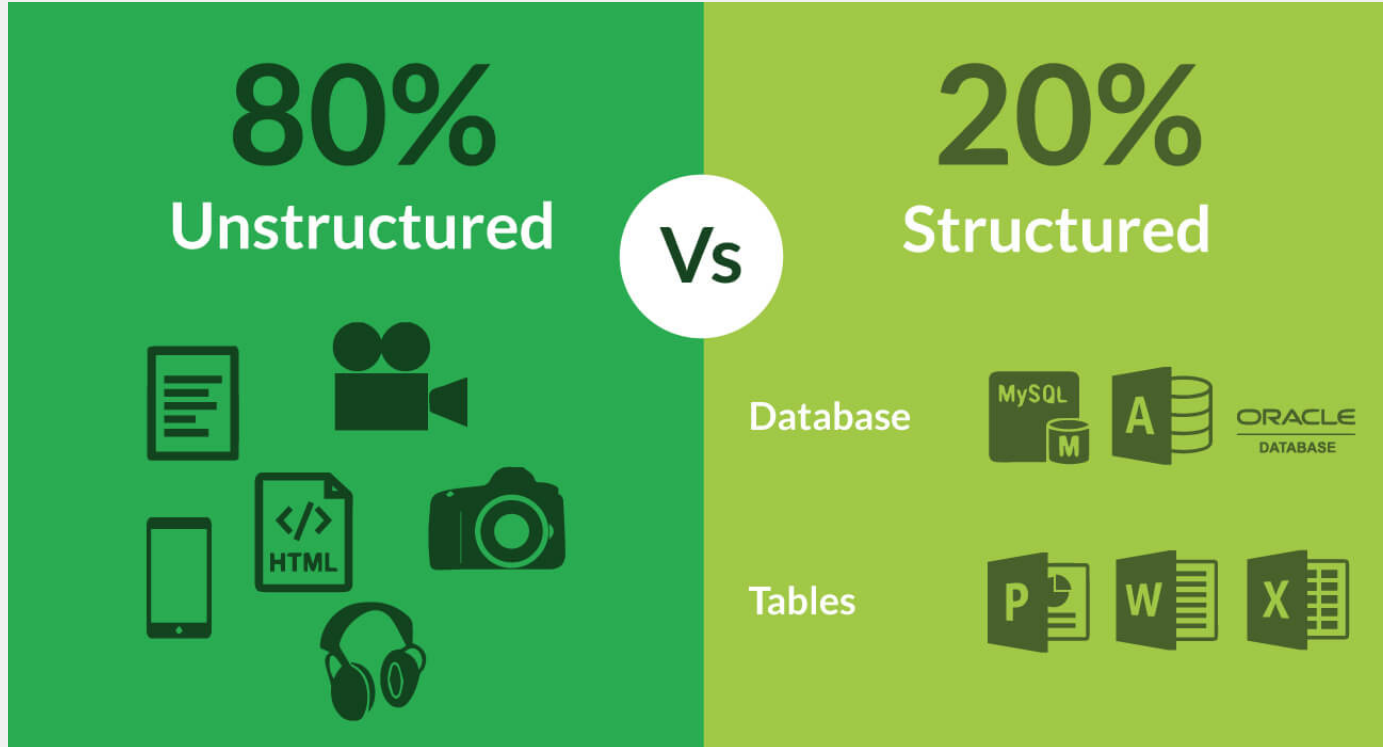
- The next step up in terms of value and complexity is the prescriptive model. The prescriptive model utilizes an understanding of what has happened, why it has happened and a variety of "what-might-happen" analysis to help the user determine the best course of action to take.
- A good example of this is a traffic application helping you choose the best route home and taking into account the distance of each route, the speed at which one can travel on each road and, crucially, the current traffic constraints.





Analytics Example	
Grouping items by similarity	Clustering
Discovering relationships between items	Association rules
Determining relationship between outcome and the input variables	Regression
Analyzing text data to find trending terms, sentiment analysis, document classification, etc.	Text analytics
Assigning label/class to records	Classification

Structured vs. Non-Structured Data



Structured vs. Non-Structured Data



Microsoft



- Most business databases contain structured data consisting of well-defined fields with numeric or alpha-numeric values.
- Examples of semi-structured data are electronic images of business documents, medical reports, executive summaries, etc. The majority of web documents also fall in this category.
- An example of unstructured data is a video recorded by a surveillance camera in a departmental store. This form of data generally requires extensive processing to extract and structure the information contained in it.

Structured vs. Non-Structured Data (Cont'd)



Microsoft



- Structured data is often referred to as traditional data, while the semi-structured and unstructured data are lumped together as non-traditional data.
- Most of the current data mining methods and commercial tools are applied to traditional data.

Introduction To Azure ML Studio



Microsoft



- Log on to <https://studio.azureml.net/>
- If you don't have an account so browse to sign up **Free Workspace** option with your existing Microsoft account.

Introduction To Azure ML Studio



Microsoft



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

Introduction To Azure ML Studio



Microsoft



Microsoft Azure Machine Learning Studio

Inseyab-Consulting

PROJECTS

EXPERIMENTS

WEB SERVICES

NOTEBOOKS

DATASETS

TRAINED MODELS

SETTINGS

experiments

MY EXPERIMENTS SAMPLES

	NAME	AUTHOR	STATUS	LAST EDITED	PROJECT
	Module 2 Reg...	AzureML Team	Draft	8/27/2017 5:06...	None
<input type="checkbox"/>	Flight Delay Pr...	demo	Draft	8/21/2017 10:3...	None
<input type="checkbox"/>	Experiment cr...	demo	Finished	8/21/2017 2:49...	None
<input type="checkbox"/>	Module 2 - Bi...	demo	Draft	8/20/2017 2:34...	None
<input type="checkbox"/>	Credit Risk Pro...	demo	Draft	8/15/2017 4:32...	None
<input type="checkbox"/>	Consumer Beh...	demo	Draft	8/14/2017 9:27...	None
<input type="checkbox"/>	Consumer Pay...	demo	Draft	8/11/2017 5:45...	None

Introduction To Azure ML Studio



Microsoft



- Click on +NEW tab and create a blank new experiment
- Explore!






Introduction To Azure ML Studio





Microsoft




NEW

-  DATASET
-  MODULE
-  PROJECT
PREVIEW
-  EXPERIMENT
-  NOTEBOOK
PREVIEW


 Search experiment templates


 Microsoft Samples




Blank Experiment

Experiment
Tutorial



[VIEW MORE IN GALLERY](#) 

Sample 1: Download dataset from UCI: Adult 2 class dataset



END OF MODULE 1
