Understanding Applied Data Science

Dzahar Mansor
NTO, Microsoft Malaysia
dmansor@microsoft.com
https://www.linkedin.com/in/dzahar-mansor

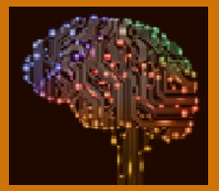
4IR – Economic Impact

United Nations Conference on Trade and Development

DIGITAL ECONOMY REPORT 2019

Value Creation and Capture:
Implications for Developing Countries

https://unctad.org/en/pages/PublicationWebflyer.aspx?publicationid=2466



Artificial Intelligence

Democratized Digital Platforms

Powering 4IR

Internet Enabled – Digital Resiliency

Social

Mobility

IoT

Connected devices

Workplace mobility

Laws and regulations

Big data

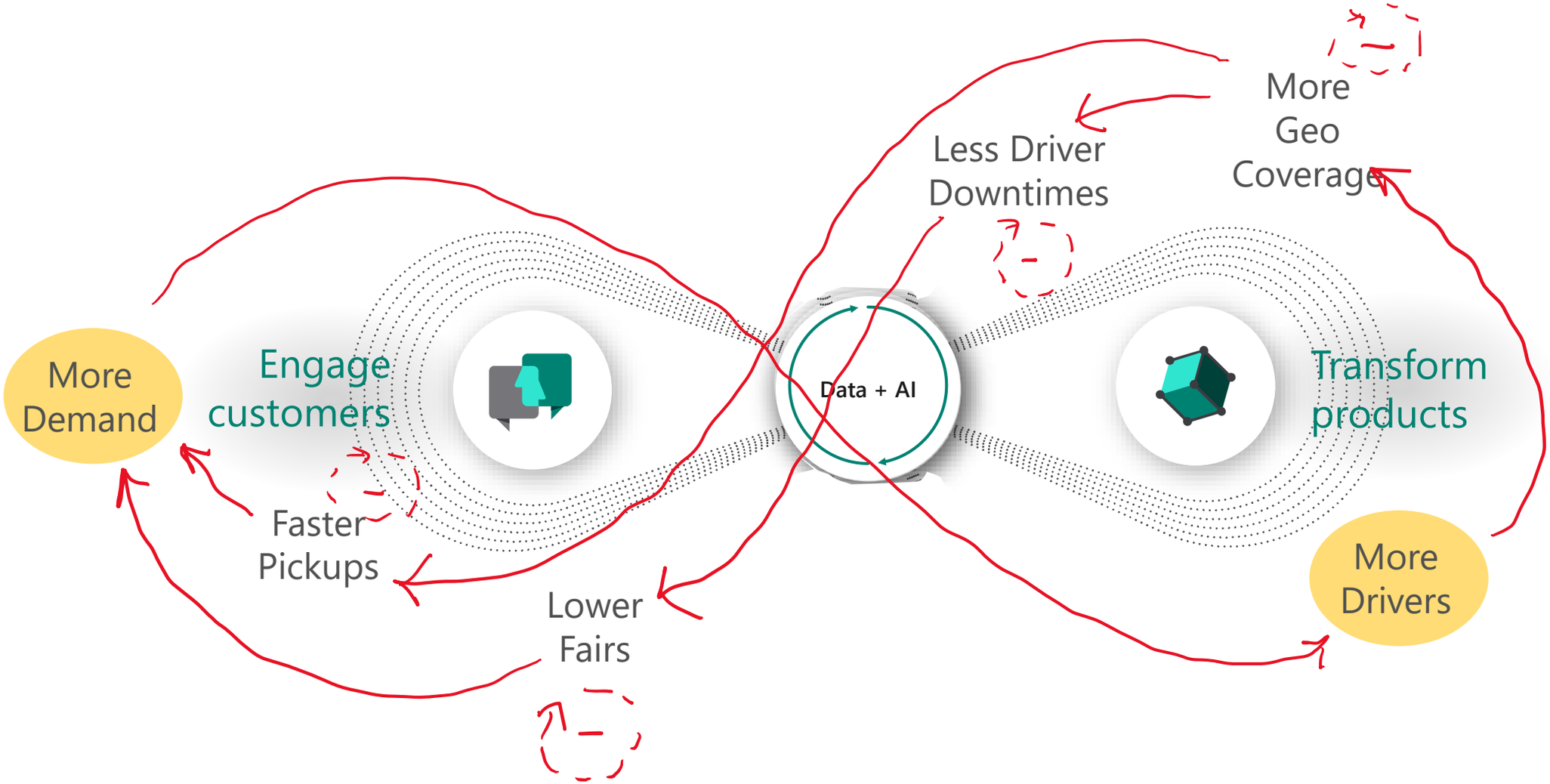
Hyperscale Commercial Cloud

ISO 17788:2014:

- Broad network access
- Measured service.
- Multi-tenancy.
- On-demand self-service
- Rapid elasticity and scalability.
- Resource pooling.

GRAB – leveraging on Democratized Platform

Pillars for Digital Transformation



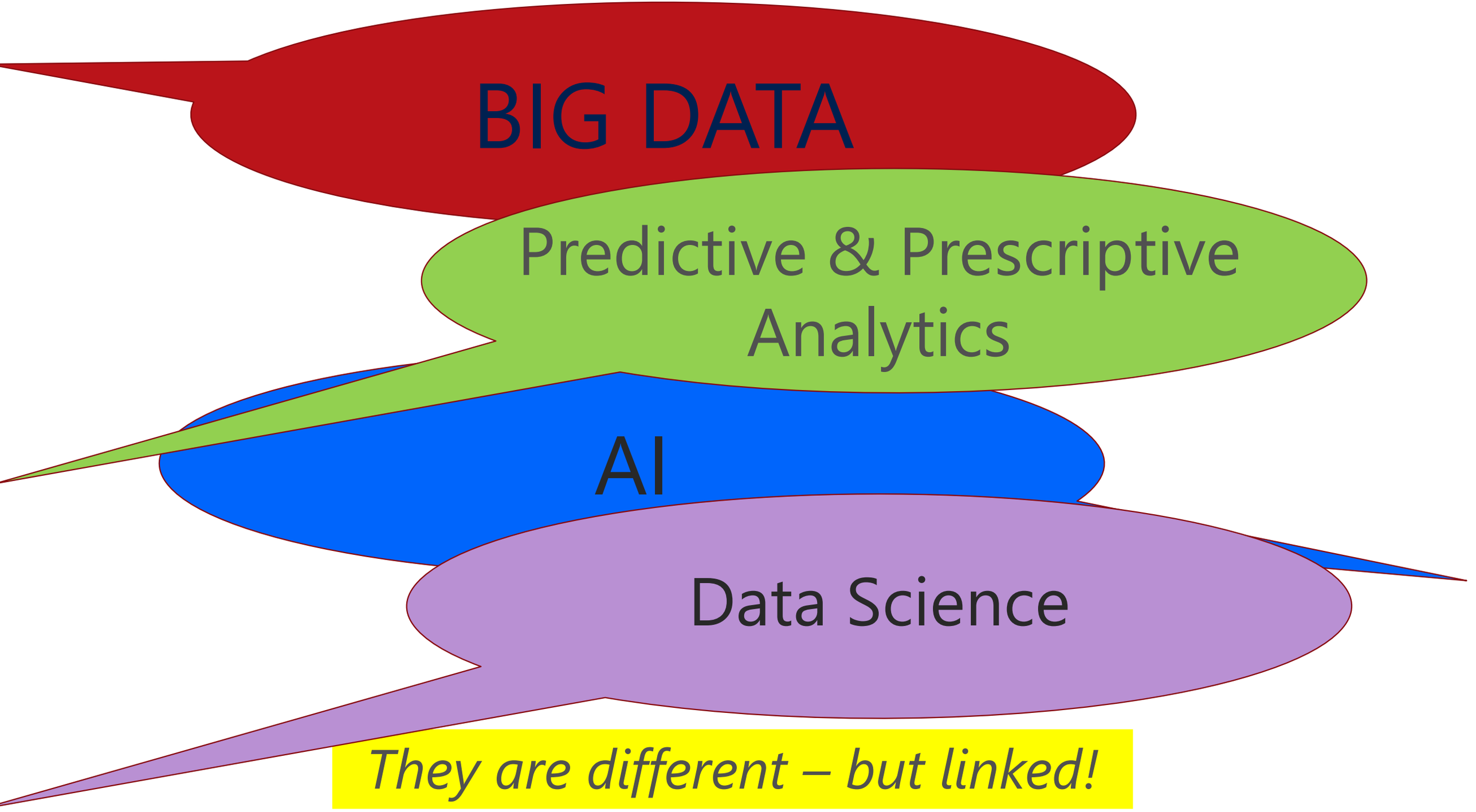
"Platform Revolution", G.G.Parker et al, WW Norton & Co.

“Data creates opportunity. It's the oil of the 21st century. Whoever has the right data will ultimately win.”

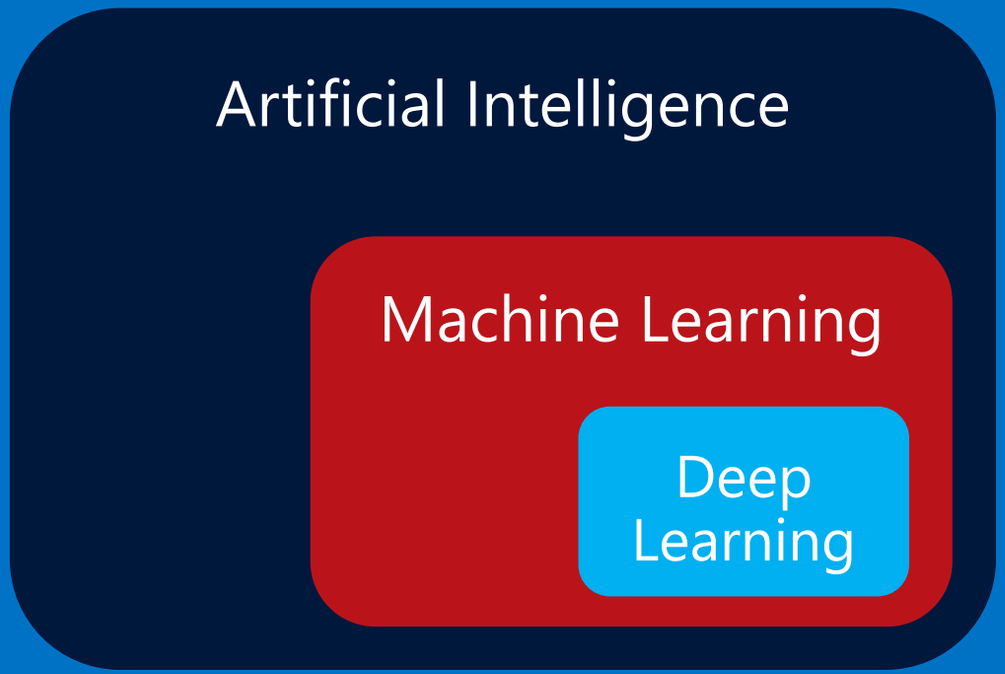
44 ZETTABYTES
2020 OF DATA

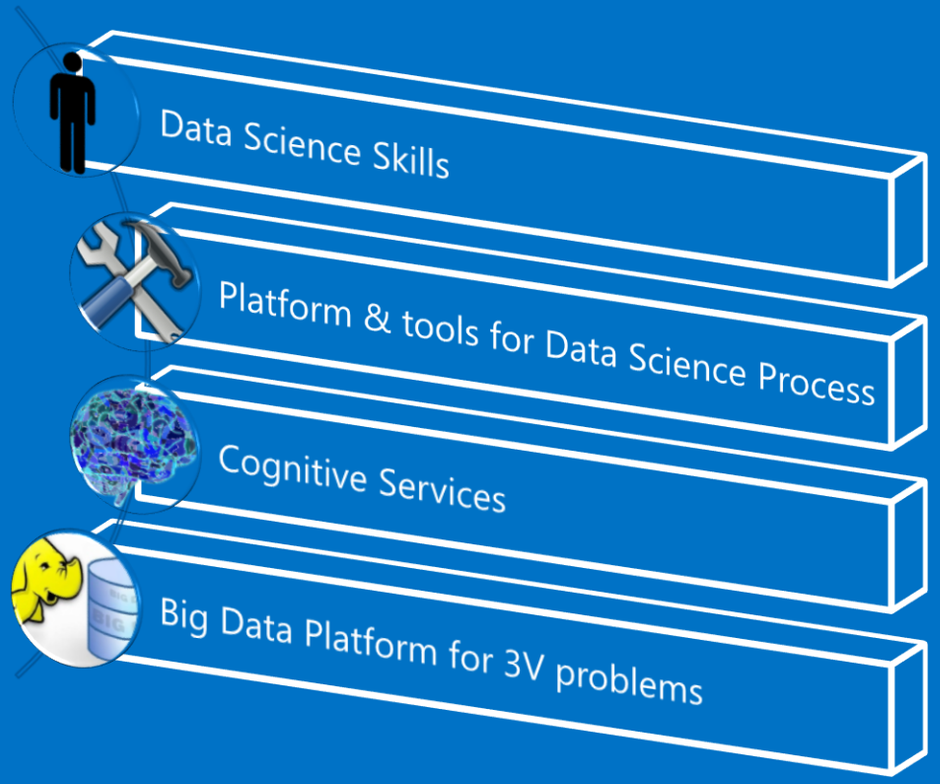
Peter Sondergaard, senior vice president of Global Research at Gartner
https://www.nyse.com/network/article/Gartner-Whats-Next

1 ZETTABYTE = 1 B TERABYTES


BIG DATA
Predictive & Prescriptive Analytics
AI
Data Science
They are different – but linked!

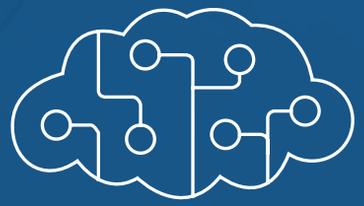
What is needed to do AI and Advanced Analytics





The Progression of ML & AI

Artificial Intelligence



Machine Learning

Deep Learning

1950 1960 1970 1980 1990 2000 2010 2018 ...

Mechanical

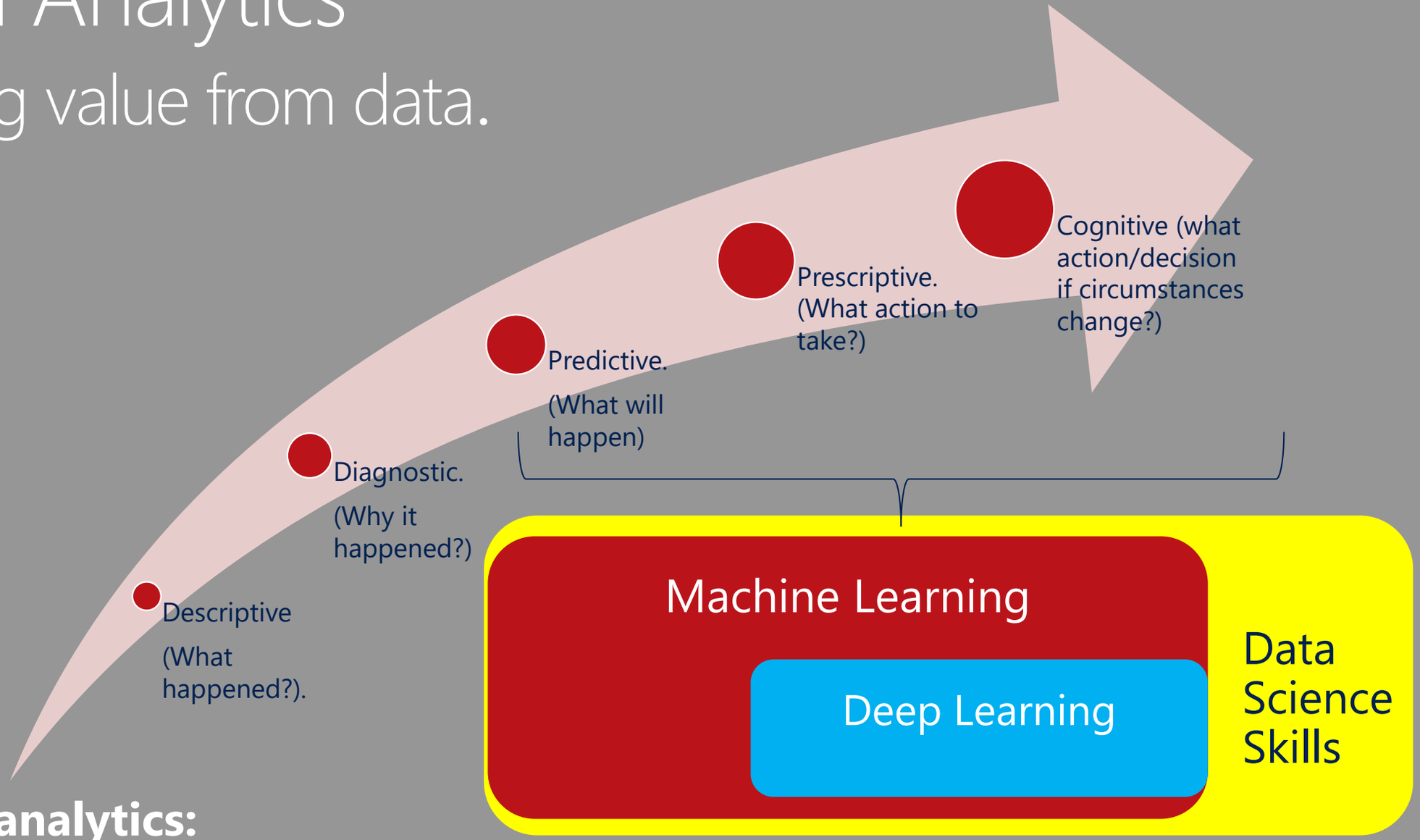
Electronics

Computers

Cloud + AI

Shift in Analytics

Extracting value from data.



Data analytics:

Refers to qualitative and quantitative techniques and processes used to enhance productivity and business gain.

Data Scientist is in High Demand

50 Best Jobs in America

This report ranks jobs according to each job's Glassdoor Job Score, determined by combining three factors: number of job openings, salary, and overall job satisfaction rating.

Employers: Want to recruit better in 2017? Find out how.

United States | 2017 | 11k Shares

1 Data Scientist

4.8 / 5
Job Score

4.4 / 5
Job Satisfaction

$110,000
Median Base Salary

4,184
Job Openings

View Jobs

Wikipedia 2015

https://en.wikipedia.org/wiki/Data_science

Data Science is the extraction of knowledge from large volumes of data that are structured or unstructured.

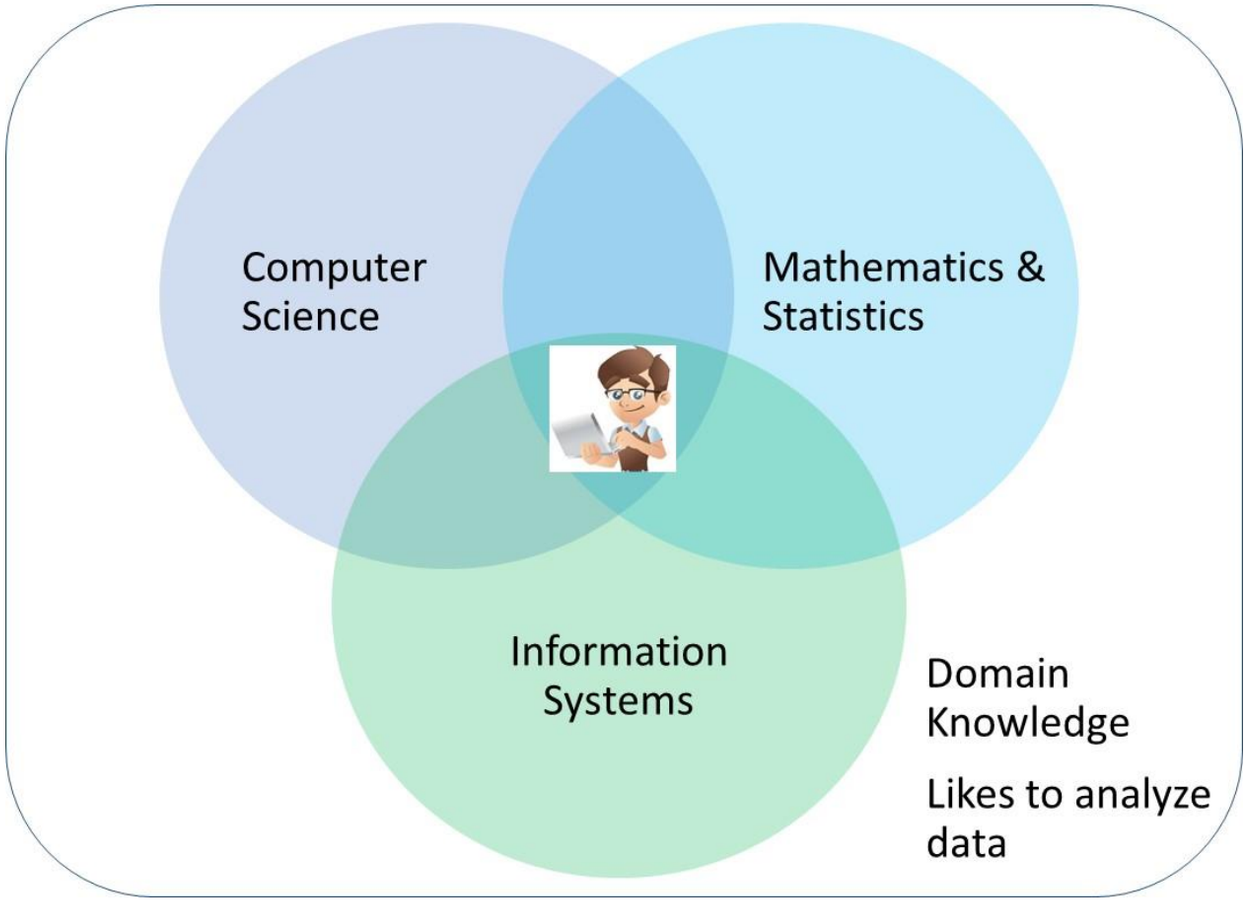
Wikipedia 21-02-2017

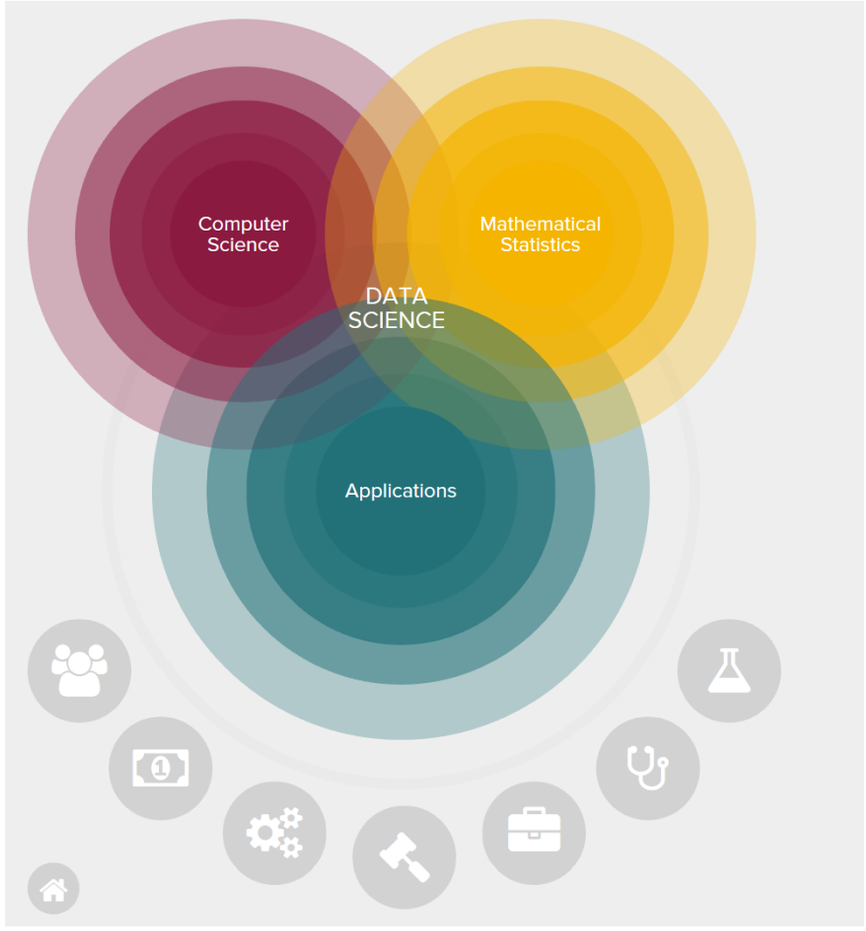
https://en.wikipedia.org/wiki/Data_science

Data science, also known as data-driven science, is an interdisciplinary field about scientific methods, processes and systems to extract knowledge or insights from data in various forms, either structured or unstructured, similar to Knowledge Discovery in Databases (KDD).

General consensus of Data Science

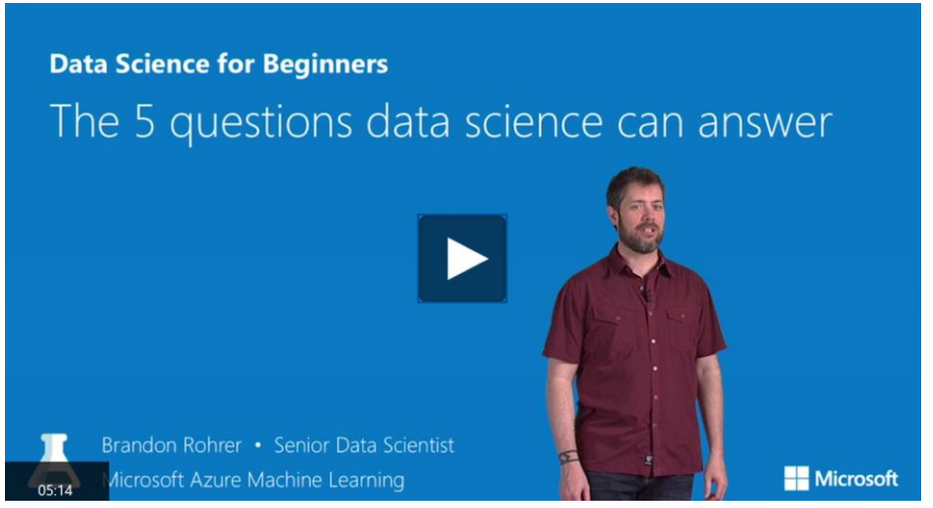
http://datascience.nyu.edu/what-is-data-science/





- Applied Field, Multi-disciplinary, Science and Mathematics based.
- “data scientist is a person who is better at statistics than any programmer and better at programming than any statistician.”

Data Science Activities – industry / practitioner view



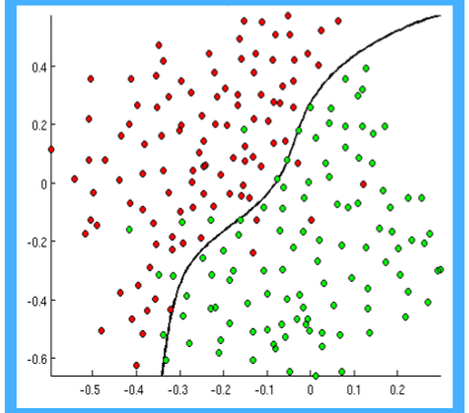
https://azure.microsoft.com/en-us/resources/videos/data-science-for-beginners-series-the-5-questions-data-science-answers/

Primary goal: To answer 5 types of questions using data – but to achieve and benefit from this, they also need to:

- Prepare data to answer the 5 questions (Data Engineering, Feature Engineering).
- To extract value on the answers from these 5 question (operationalizing data):
  - Predicting and forecasting.
  - AI/Intelligence/Cognitive applications.

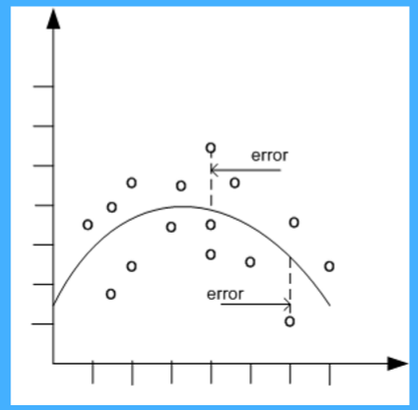
Advanced Analytics Questions Answered by "Data Science"

Classification



Is it A or B?
e.g Will the HDD fail next month? Yes/No.

Regression (Forecasting)

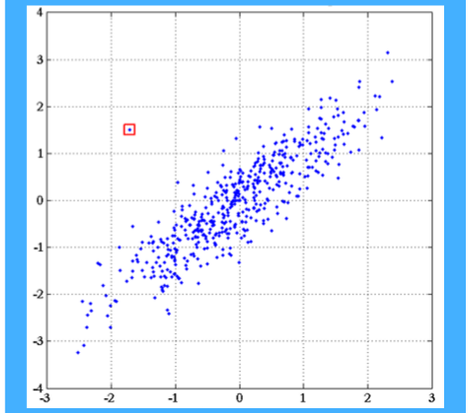


How much/how many?
e.g. what is the temperature next Tuesday?

Reinforcement Learning

Which option?
e.g. do the car stop or go on an orange light.

Anomaly Detection



Is it weird?
e.g. Fraud Detection.

Clustering

Which groups? *e.g. which viewers like the same type of movie.*

Machine Learning *algorithms are used in Advanced Analytics by Data Scientists.*

Mapping CRISP DM Process

70% of time spent here

Business Understanding

Getting more data

- Asking the sharp question.
- Determining the base ML algorithms.
- Sourcing and Selecting data sources.
- Determining the features and label.

Data Understanding

Collecting Data.

Describing the data.

Exploring & Visualizing Data

- Summary statistics.
- Univariate plotting.
  - Histogram.
  - Box Plot.
  - Histograms (categorical).
- Plotting 2 Variables.
  - Scatter plots.
- Scatter plot matrices.
- Faceting scatter plots.
- Conditional histogram plots.

Verifying Data Quality.

Data Preparation

Joining data into a single table.

- SQL.
- Programming R, Python.

Feature Selection – numeric features

- Features correlated with label.
- Remove redundant collinear features
- Feature importance.

Cleansing:

- Missing value.
- Removing Repeated value.

Handling Error and Outliers:

- Censor
- Trim.
- Interpolate.
- Substitute.

Transforming data:

- Transform string to categories.
- Group/aggregate categories.
- Quantizing continuous variables.
- Convert to indicator values.
- Rename features if required.
- Text Frequency Hashing.

Scale numeric features:

- Z-Score scaling.
- Min-max scaling.
- De-trend data.

Feature Engineering:

- Feature Creation:
  - Log of features
  - Feature difference/Differencing.

Modeling

Classification (A or B)

- Logistic regression.
- Two-Class Boosted Decision Trees.
- Two-Class Decision Forest.
- Two-Class Neural Network.
- Two-Class [Locally-Deep] SVM
- K-Nearest Neighbors.

Regression, Forecasting (How many)

- Linear regression.
- Boosted Decision Tree Regression.
- Forest regression.
- NN Regression.
- ARIMA
- K-Nearest Neighbors.

Neural Net (Reinforcement Learning)

Clustering (Which Group)

- K-Means.
- Hierarchical Agglomerative Clustering.
- Recommenders.

Evaluation

Classification (A or B)

- Accuracy.
- AOC.
- Confusion matrix.
- Recall.
- Feature interpretation.
- Tuning/Sweeping Model Parameters.
- Cross validation.
- Nested cross validation to tune parameters.
- Comparison between models.

Regression (How many)

- Root Mean Square Error.
- Standard Error (ARMA).
- Coefficient of determination.
- Residual Visualization:
  - Time series plots of actual versus predicted.
  - Box-Plot
  - Time series plots of residuals (line, histograms) .
- Regularization.
- Tuning/Sweeping Model Parameters.
- Cross validation.
- Nested cross validation to tune parameters.
- Comparison between models.

Clustering (Which Group)

- Maximal distance to cluster center.
- Principal Component Analysis (PCA)

Deployment

Deployment :

- Deployment lifecycle.
- Web Services APIs
- Web Interface.
- [Deployment coding.]

Advanced Analytics:

- Business Intelligence.
- Predictive Analytics.

Artificial Intelligence:

- AI Applications.
- Cognitive Services.

Business Understanding

Asking the Sharp Questions, Understanding the Data Sources

Dzahar Mansor
NTO, Microsoft Malaysia

https://www.youtube.com/watch?v=tKa0zDDDaQk&feature=youtu.be

Mapping CRISP DM Process

Business Understanding

Getting more data

- Asking the sharp question.
- Determining the base ML algorithms.
- Sourcing and Selecting data sources.
- Determining the features and label.

Data Understanding

Importing Data.

Exploring & Visualizing Data

- Univariate plotting.
  - Histogram.
  - Box Plot.
  - Histograms (categorical).
- Plotting 2 Variables.
  - Scatter plots.
- Scatter plot matrices.
- Faceting scatter plots.
- Conditional histogram plots.

Joining data into a single table.

- SQL.
- Programming R, Python.

Feature Selection – numeric features

- Features correlated with label.
- Remove redundant collinear features

Data Preparation

Cleansing:

- Missing value.
- Removing Repeated value.

Handling Error and Outliers:

- Censor
- Trim.
- Interpolate.
- Substitute.

Transforming data:

- Transform string to categories.
- Group/aggregate categories.
- Quantizing continuous variables.
- Convert to indicator values.
- Rename features if required.
- Text Frequency Hashing.

Scale numeric features:

- Z-Score scaling.
- Min-max scaling.
- De-trend data.

Feature Engineering:

- Feature Creation:
  - Log of features
  - Feature difference/Differencing.
  - Add features.
  - Feature multiplication
- Refining Feature Selection
  - Removing marginal impact features.

Handle imbalance categorical labels for classification.

- Coding R/Python.
- SMOTE

Modeling

Classification (A or B)

- Logistic regression.
- Two-Class Boosted Decision Trees.
- Two-Class Decision Forest.
- Two-Class Neural Network.
- Two-Class [Locally-Deep] SVM

Regression, Forecasting (How many)

- Linear regression.
- Boosted Decision Tree Regression.
- Forest regression.
- NN Regression.
- ARIMA

Neural Net (Reinforcement Learning)

Clustering (Which Group)

- K-Means.
- Hierarchical Agglomerative Clustering.
- Recommenders.

Evaluation

Classification (A or B)

- Accuracy.
- AOC.
- Confusion matrix.
- Recall.
- Feature interpretation.
- Sweeping Model Parameters.
- Cross validation.
- Nested cross validation to tune parameters.
- Comparison between models.

Regression (How many)

- Root Mean Square Error.
- Standard Error (ARMA).
- Coefficient of determination.
- Residual Visualization:
  - Time series plots of actual versus predicted.
  - Box-Plot
  - Time series plots of residuals (line, histograms) .
- Regularization.
- Cross validation.
- Nested cross validation to tune parameters.
- Comparison between models.

Clustering (Which Group)

- Maximal distance to cluster center.
- Principal Component Analysis (PCA)

Deployment

Deployment :

- Deployment lifecycle.
- Web Services APIs
- Web Interface.
- [Deployment coding.]

Advanced Analytics:

- Business Intelligence.
- Predictive Analytics.

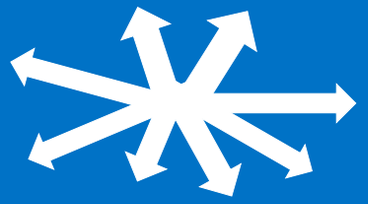
Artificial Intelligence:

- AI Applications.
- Cognitive Services.

Sourcing and Selecting the Data Sources

- Asking the sharp question.
- Determining the ML algorithm(s)
- Sourcing the data
- Selecting the data sources.

Vague questions vs. Sharp questions



Doesn't have to be answered with a name or a number

What can my data tell me about my business?

Is the car good?

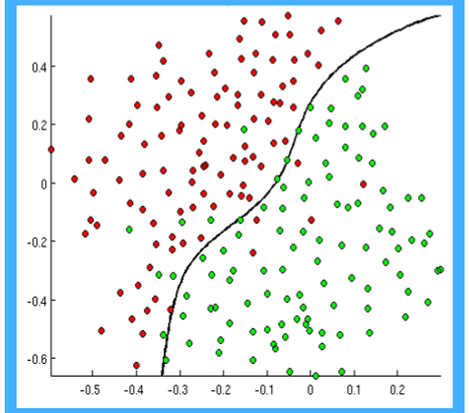
Must be answered with a name/category or a number.

What will be my projected revenue next quarter?

How many months before the car develops problems?

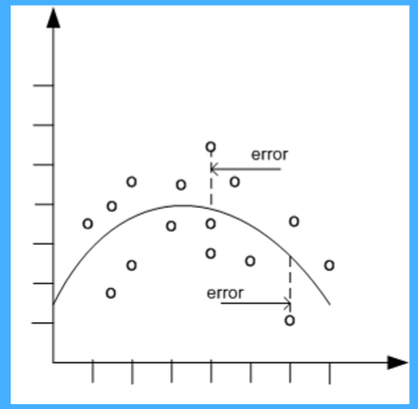
Advanced Analytics Questions Answered by "Data Science"

Classification



Is it A or B?
e.g Will the HDD fail next month? Yes/No.

Regression (Forecasting)

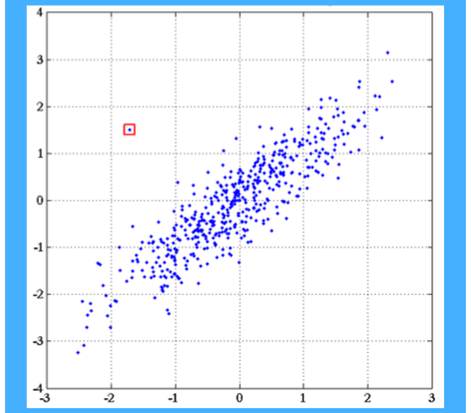


How much/how many?
e.g. what is the temperature next Tuesday?

Reinforcement Learning

Which option?
e.g. do the car stop or go on an orange light.

Anomaly Detection



Is it weird?
e.g. Fraud Detection.

Clustering

Which groups? *e.g. which viewers like the same type of movie.*

Sourcing and Selecting the Data Sources

- Enterprise data repository:
  - Transactional Databases.
  - Archives.
  - Data warehouse, Data Lake.
  - Files – log files, spreadsheet etc.
- Real time data:
  - Sensors.
  - Cameras.
- Data from the internet.

Selection of Data

What will my stock price be next week?

Database

Date	My stock price

Exel

Date	DowJ	NQDXUSBT

128

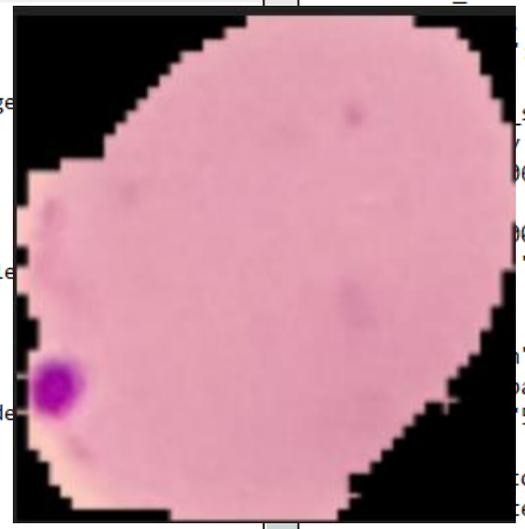
HTML

```
!DOCTYPE HTML>
<html lang="en">

<!-- To use your snippet, just add the following code in your page
<!-- {% include header %} -->


<head>
  <meta charset="utf-8">
  <meta name="viewport" content="width=device-width, initial-scale
  <meta http-equiv="X-UA-Compatible" content="IE=edge" />

  <div class="row no-gutters">
        <div class="col-md-4">
          <table class="m-0 table table-striped border-right borde
            <tbody>
              <tr>
                <th class="w-50" scope="row">Stock Code</th>
                <td class="w-50">8583</td>
              </tr>
              <tr>
                <th scope="row">Change</th>
                <td class="bold text-success">
                  +0.195
                </td>
```



JSON

```
{
   "entities":
   {
      "user_mentions":[],

      :[]

      screen_name":null,
      stocks fluctuation. Time to reconsider strategy.",
      0621131198173184",

      03Ca href=\"http:\/\/twitter.com\/#!\/download\/iphone\",
      Thu Jun 07 06:36:05 +0000 2012",

      ",
      ackground_image_url":"http:\/\/a0.twimg.com\/images\/themes
      568286862",

      ount":65,
      ext_color":"333333",
      "location":"Stark Industries"
   },
   "retweeted":false,
   "id":210621131198173184,
   "coordinates":null,"geo":null
}
```

128

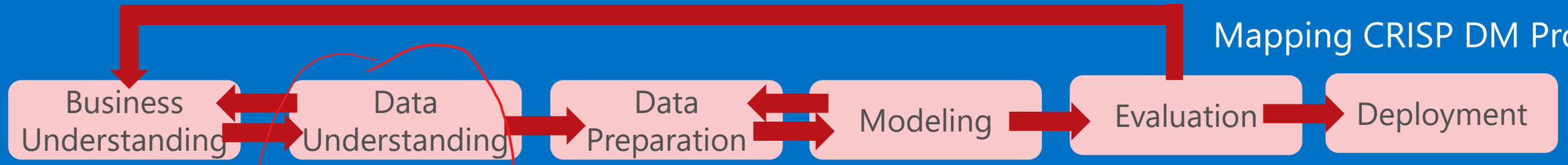
Understanding Data

End to end Data Science Activities

Dzahar Mansor
NTO, Microsoft Malaysia

https://www.youtube.com/watch?v=tKa0zDDDaQk&feature=youtu.be

Mapping CRISP DM Process



Business Understanding

Getting more data

- Asking the sharp question.
- Determining the base ML algorithms.
- Sourcing and Selecting data sources.
- Determining the features and label.

Data Understanding

Importing Data.

Exploring & Visualizing Data

- Univariate plotting.
  - Histogram.
  - Box Plot.
  - Histograms (categorical).
- Plotting 2 Variables.
  - Scatter plots.
- Scatter plot matrices.
- Faceting scatter plots.
- Conditional histogram plots.

Joining data into a single table.

- SQL.
- Programming R, Python.

Feature Selection – numeric features

- Features correlated with label.
- Remove redundant collinear features

Data Preparation

Cleansing:

- Missing value.
- Removing Repeated value.

Handling Error and Outliers:

- Censor
- Trim.
- Interpolate.
- Substitute.

Transforming data:

- Transform string to categories.
- Group/aggregate categories.
- Quantizing continuous variables.
- Convert to indicator values.
- Rename features if required.

Scale numeric features:

- Z-Score scaling.
- Min-max scaling.
- De-trend data.

Feature Engineering:

- Feature Creation:
  - Log of features
  - Feature difference.
  - Add features.
  - Feature multiplication
- Refining Feature Selection
  - Removing marginal impact features.

Handle imbalance categorical labels for classification.

- Coding R/Python.
- SMOTE

Modeling

Classification (A or B)

- Logistic regression.
- Two-Class Boosted Decision Trees.
- Two-Class Decision Forest.
- Two-Class Neural Network.
- Two-Class [Locally-Deep] SVM

Regression (How many)

- Linear regression.
- Boosted Decision Tree Regression.
- Forest regression.
- Neural Network Regression.

Neural Net (Reinforcement Learning)

Clustering (Which Group)

- K-Means.
- Hierarchical Agglomerative Clustering.

Evaluation

Classification (A or B)

- Accuracy.
- AOC.
- Confusion matrix.
- Recall.
- Cross validation.
- Nested cross validation to tune parameters.
- Comparison between models.

Regression (How many)

- Root Mean Square Error.
- Coefficient of determination.
- Residual Visualization:
  - Time series plots of actual versus predicted.
  - Box-Plot
  - Time series plots of residuals (line, histograms) .
- Cross validation.
- Nested cross validation to tune parameters.
- Comparison between models.

Clustering (Which Group)

- Maximal distance to cluster center.
- Principal Component Analysis (PCA)

Model Improvement:

- Select best performing model.
- Identifying marginal impact features.
- Select tuned model parameters.

Deployment

Deployment :

- Deployment lifecycle.
- Web Services APIs
- Web Interface.
- [Deployment coding.]

Advanced Analytics:

- Business Intelligence.
- Predictive Analytics.

Artificial Intelligence:

- AI Applications.
- Cognitive Services.

Importing and Understanding Data

- Data from various sources are combined into one table.
  - The column (variable) that we want to predict the answer to the sharp question is called the *label*.
  - The other columns (variable) are called the *features*.
- 2 variable types :
  - Numeric
  - Categorical
- Data Visualization
- Feature selection.

Joining Data into one Table

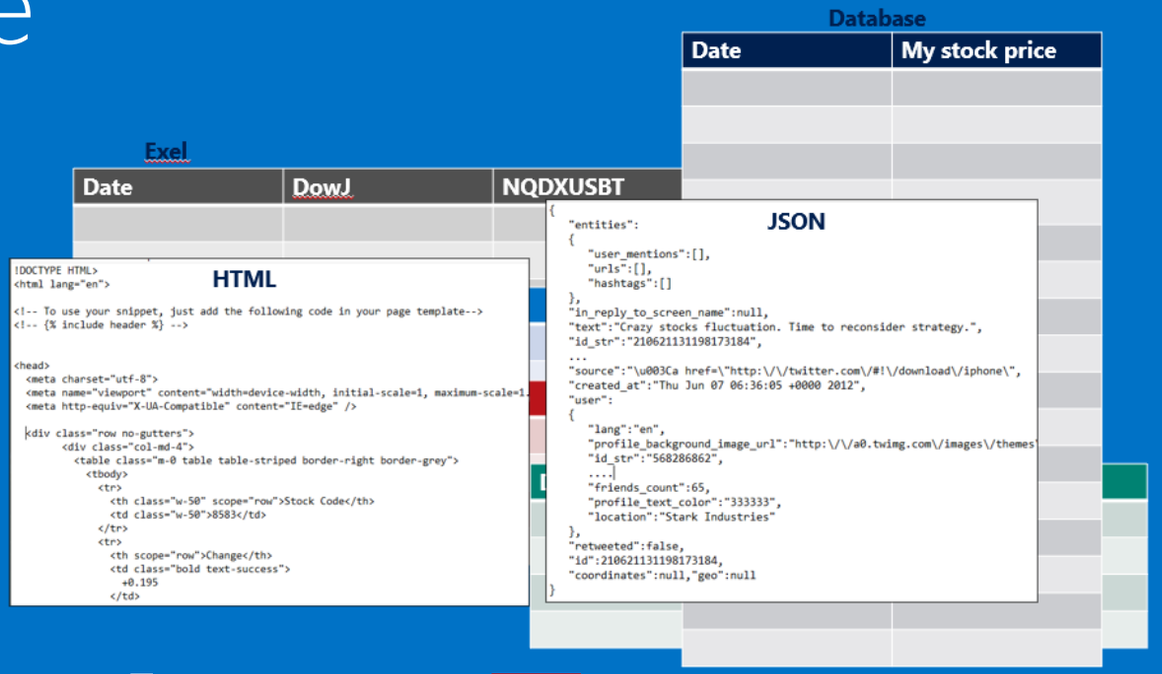
Importing Data.

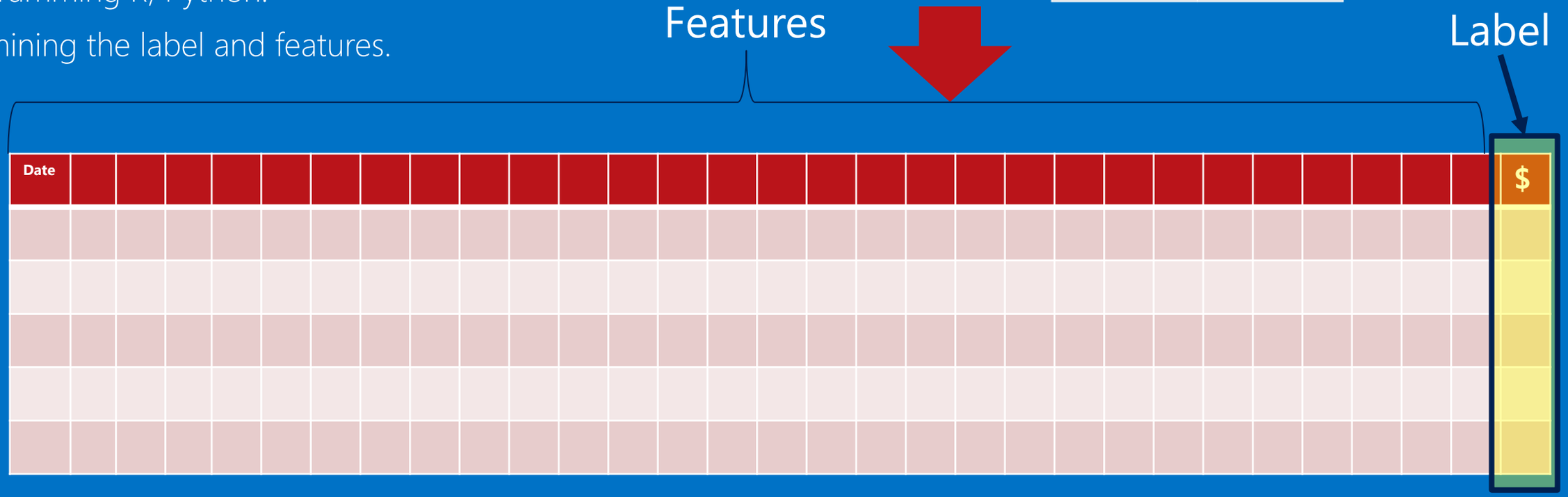
- Database.
- File transfer (Excel).
- Data feed (CSV)
- Twitter Feed (JSON)

Joining data into a single table.

- SQL.
- Programming R, Python.

Determining the label and features.





Types of features and labels

Numerical (Number)

Amount : 38.3 degrees

Count : 39 pizzas

Money : $1,387

Pixel brightness : 232/255

Sound intensity : .64

Categorical (Names)

Type : Shih Tzu

Variety : Caramel latte

ID : Air Force One

Model number : R2-D2

Category : Chocolate

Text : “Best. Show. Ever. <3”

Some categorical features look numeric

Phone number : 847-5609

Zip code : 90210

ID number : 007

Serial number : 100000184573

Credit card number : 5738-7539-9898-0023

Social security number : 627-42-0932

Some numeric features look like categories

Place : first, second, third
Size : small, medium, large
Side : left, middle, right
Time zone : Pacific, Mountain, Central, Eastern
Train stops : Kendall, Central, Harvard, Porter

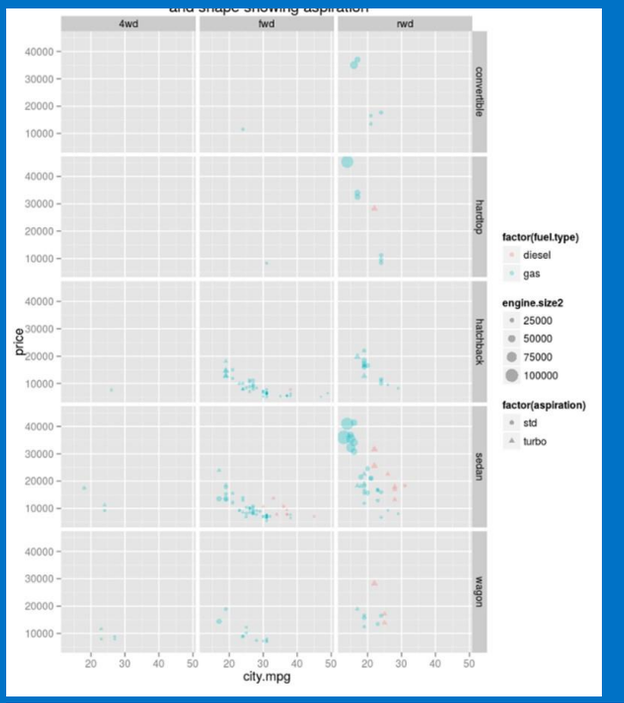
Data Understanding – Visualizing the Data in Table

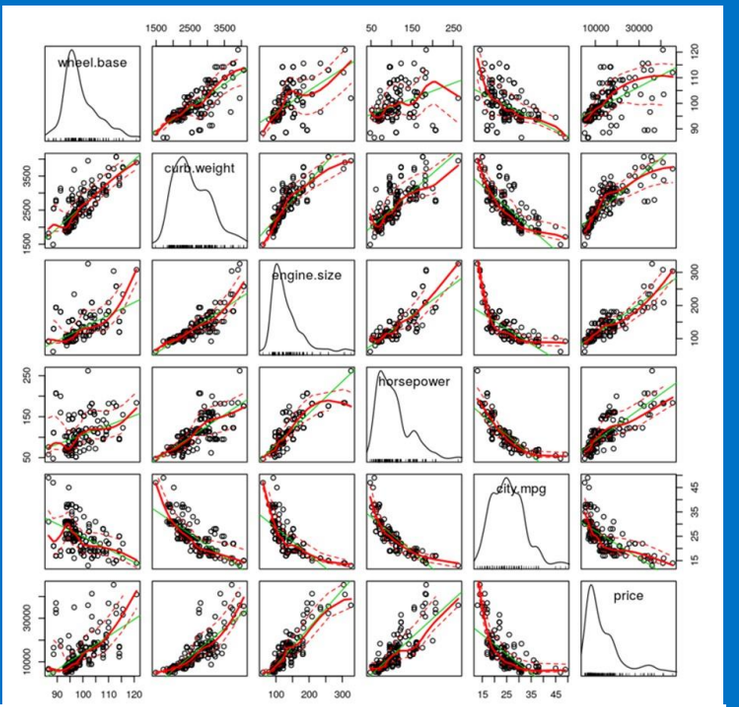
Exploring & Visualizing Data

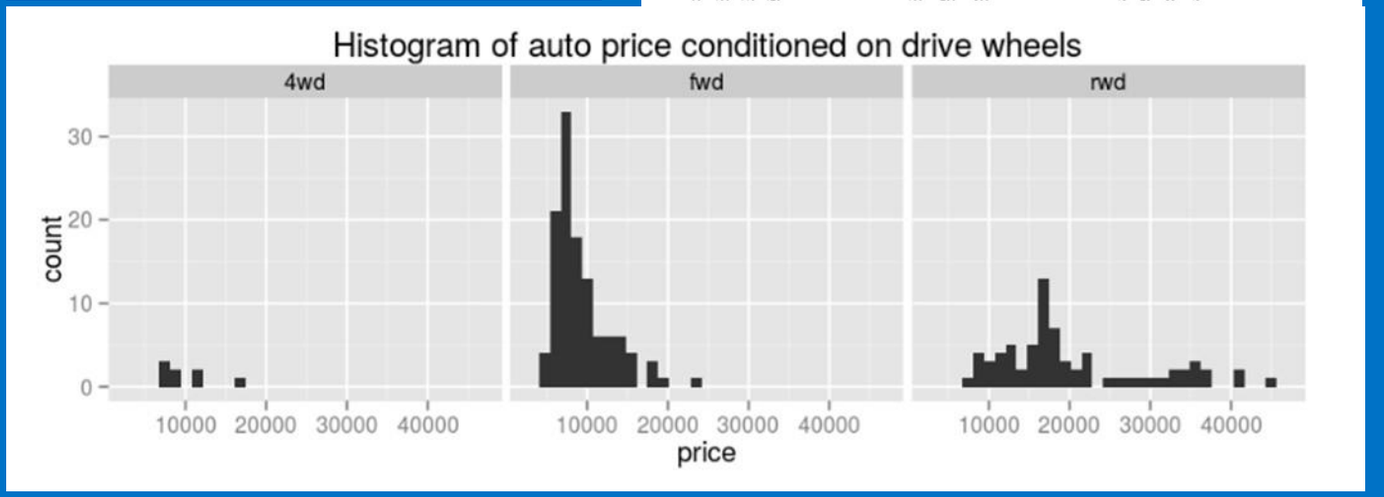
- Summary Statistics.
- Univariate plotting.
  - **Histogram.**
  - **Box Plot.**
  - **Histograms (categorical).**
- Plotting 2 Variables.
  - **Scatter plots.**
- Scatter plot matrices.
- Faceting scatter plots.
- Conditional histogram plots.

(Initial) Feature Selection – numeric features

- Features correlated with label.
- Remove redundant collinear features.
- Feature importance evaluation.





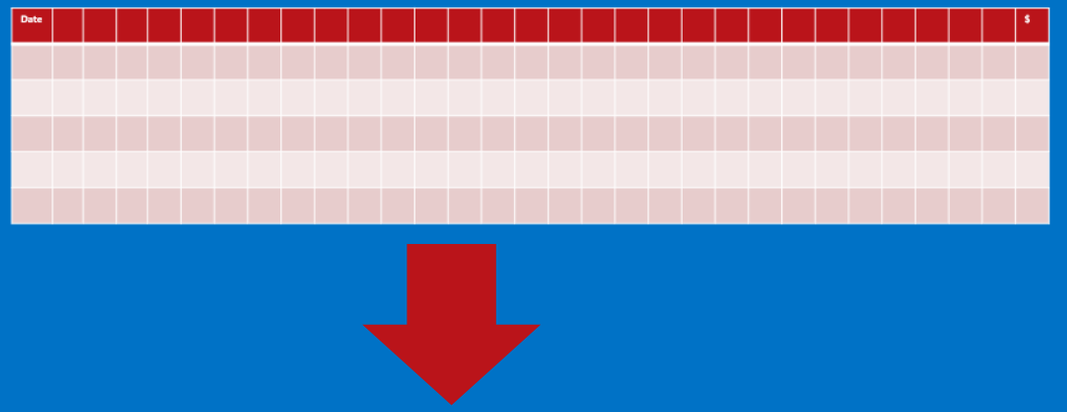


Data Understanding – Select Features and Label

Two approaches for seature selection:

1. Greedy backward selection :
   Start with all features, remove feature that hurts power least, repeat until criteria met.

2. Greedy forward selection
   Start with no features, add feature that improves model, iterate until criteria is met.



Stock price	Date	Day of week	P-score	Dow Jones	Daily Sales	New users	Date Last Press Release	Date Last Product Release
57.3	“5/21”	“Tue”		17,245	2.6M	2647	5/18	2/15
58.8	“05/22”	“Wed”	“H”	17,289	3.2M		5/19	2/15
56.9	“5-23”	“Thu”		17,115	2.1M	2120	5/20	2/15
57.4	“5/24”	“Fri”		17,278	2.8M	2901	5/21	2/15

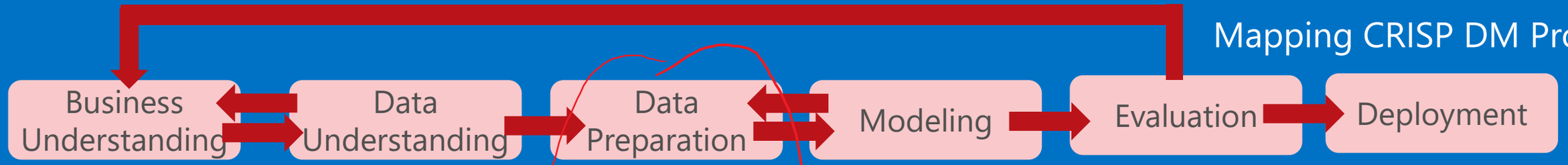
Data Preparation

End to end Data Science Activities

Dzahar Mansor
NTO, Microsoft Malaysia

https://www.youtube.com/watch?v=tKa0zDDDaQk&feature=youtu.be

Mapping CRISP DM Process



Business Understanding

Getting more data

- Asking the sharp question.
- Determining the base ML algorithms.
- Sourcing and Selecting data sources.
- Determining the features and label.

Data Understanding

Importing Data.

Exploring & Visualizing Data

- Univariate plotting.
  - Histogram.
  - Box Plot.
  - Histograms (categorical).
- Plotting 2 Variables.
  - Scatter plots.
- Scatter plot matrices.
- Faceting scatter plots.
- Conditional histogram plots.

Joining data into a single table.

- SQL.
- Programming R, Python.

Feature Selection – numeric features

- Features correlated with label.
- Remove redundant collinear features

Data Preparation

Cleansing:

- Missing value.
- Removing Repeated value.

Handling Error and Outliers:

- Censor
- Trim.
- Interpolate.
- Substitute.

Transforming data:

- Transform string to categories.
- Group/aggregate categories.
- Quantizing continuous variables.
- Convert to indicator values.
- Rename features if required.

Scale numeric features:

- Z-Score scaling.
- Min-max scaling.
- De-trend data.

Feature Engineering:

- Feature Creation:
  - Log of features
  - Feature difference.
  - Add features.
  - Feature multiplication
- Refining Feature Selection
  - Removing marginal impact features.

Handle imbalance categorical labels for classification.

- Coding R/Python.
- SMOTE

Modeling

Classification (A or B)

- Logistic regression.
- Two-Class Boosted Decision Trees.
- Two-Class Decision Forest.
- Two-Class Neural Network.
- Two-Class [Locally-Deep] SVM

Regression (How many)

- Linear regression.
- Boosted Decision Tree Regression.
- Forest regression.
- Neural Network Regression.

Neural Net (Reinforcement Learning)

Clustering (Which Group)

- K-Means.
- Hierarchical Agglomerative Clustering.

Evaluation

Classification (A or B)

- Accuracy.
- AOC.
- Confusion matrix.
- Recall.
- Cross validation.
- Nested cross validation to tune parameters.
- Comparison between models.

Regression (How many)

- Root Mean Square Error.
- Coefficient of determination.
- Residual Visualization:
  - Time series plots of actual versus predicted.
  - Box-Plot
  - Time series plots of residuals (line, histograms) .
- Cross validation.
- Nested cross validation to tune parameters.
- Comparison between models.

Clustering (Which Group)

- Maximal distance to cluster center.
- Principal Component Analysis (PCA)

Model Improvement:

- Select best performing model.
- Identifying marginal impact features.
- Select tuned model parameters.

Deployment

Deployment :

- Deployment lifecycle.
- Web Services APIs
- Web Interface.
- [Deployment coding.]

Advanced Analytics:

- Business Intelligence.
- Predictive Analytics.

Artificial Intelligence:

- AI Applications.
- Cognitive Services.

Data Preparation – Data Cleansing

- Sparse columns/rows.
- Dealing with missing data.
- Badly formatted data

Stock price	Date	Day of week	P-score	Dow Jones	Daily Sales	New users	Date Last Press Release	Date Last Product Release
57.3	“5/21”	“Tue”		17,245	2.6M	2647	“5/18”	“2/15”
58.8	“05/22”	“Wed”	“H”	17,289	3.2M		“5/19”	“2/16”
56.9	“5-23”	“Thu”		17,115	2.1M	2120	“5/20”	“2/17”
57.4	“5/24”	“Fri”		17,278	2.8M	2901	“5/21”	“2/18”

Stock price	Date	Day of week	P-score	Dow Jones	Daily Sales	New users	Date Last Press Release	Date Last Product Release
57.3	“5/21”	“Tue”		17,245	2.6M	2647	“5/18”	“2/15”
58.8	“05/22”	“Wed”	“H”	17,289	3.2M		“5/19”	“2/16”
56.9	“5-23”	“Thu”		17,115	2.1M	2120	“5/20”	“2/17”
57.4	“5/24”	“Fri”		17,278	2.8M	2901	“5/21”	“2/18”

Data Preparation – Prepare Features and Label

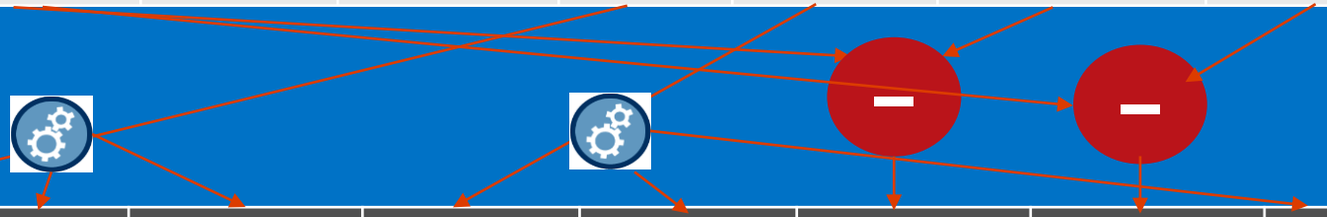
- Converting string (e.g dates) to numbers.
- Converting strings to categories.

Stock price	Date	Day of week	Dow Jones	Daily Sales	New users	Date Last Press Release	Date Last Product Release
57.3	1558396800000	C(Tue)	17,245	2.6M	2647	1558137600000	1550188800000
58.8	1558483200000	C(Wed)	17,289	3.2M	2742	1558224000000	1550275200000
56.9	1558569600000	C(Thu)	17,115	2.1M	2120	1558310400000	1550361600000
57.4	1558656000000	C(Fri)	17,278	2.8M	2901	1558396800000	1550448000000

Data Preparation – Feature Engineering

- Convert dates # days from.
- Calculate daily data to monthly and quarterly.
- Calculate daily new users to monthly, quarterly and total users.
- Remove redundant cols

Stock price	Date	Day of week	Dow Jones	Daily Sales	New users	Date Last Press Release	Date Last Product Release
57.3	1558396800000	C(Tue)	17,245	2.6M	2647	1558137600000	1550188800000
58.8	1558483200000	C(Wed)	17,289	3.2M	2742	1558224000000	1550275200000
56.9	1558569600000	C(Thu)	17,115	2.1M	2120	1558310400000	1550361600000
57.4	1558656000000	C(Fri)	17,278	2.8M	2901	1558396800000	1550448000000



Stock price	Date	Day of week	Dow Jones	Last month sales	Last quarter sales	Market share	New users last month	New users last quarter	Days since press release	Days since product release	Total users
57.3	1558396800000	C(Tue)	17245	68.8M	211.2M	23.1%	63,522	195,322	3	96	2.49M
58.8	1558483200000	C(Wed)	17289	68.8M	211.2M	23.1%	63,522	195,322	4	97	2.49M
56.9	1558569600000	C(Thu)	17115	68.8M	211.2M	23.1%	63,522	195,322	5	98	2.49M
57.4	1558656000000	C(Fri)	17278	68.8M	211.2M	23.1%	63,522	195,322	6	99	2.49M

Data Preparation

- Normalize numeric features (min-max)

Stock price	Date	Day of week	Dow Jones	Last month sales	Last quarter sales	Market share	New users last month	New users last quarter	Days since press release	Days since product release	Total users
57.3	1558396800000	C(Tue)	17245	68.8M	211.2M	23.1%	63,522	195,322	3	96	2.49M
58.8	1558483200000	C(Wed)	17289	68.8M	211.2M	23.1%	63,522	195,322	4	97	2.49M
56.9	1558569600000	C(Thu)	17115	68.8M	211.2M	23.1%	63,522	195,322	5	98	2.49M
57.4	1558656000000	C(Fri)	17278	68.8M	211.2M	23.1%	63,522	195,322	6	99	2.49M

Stock price	Date	Day of week	Dow Jones	Last month sales	Last quarter sales	Market share	New users last month	New users last quarter	Days since press release	Days since product release	Total users
57.3	C(1558396800000)	C(Tue)	0.7945	0.7621	0.7816	0.231	0.6923	0.74901	0.3	0.480	0.7328
58.8	C(1558483200000)	C(Wed)	0.7213	0.7621	0.7816	0.231	0.6923	0.74901	0.4	0.485	0.7328
56.9	C(1558569600000)	C(Thu)	0.6980	0.7621	0.7816	0.231	0.6923	0.74901	0.5	0.490	0.7328
57.4	C(1558656000000)	C(Fri)	0.8369	0.7621	0.7816	0.231	0.6923	0.74901	0.6	0.495	0.7328

Data Preparation

End to end Data Science Activities

Dzahar Mansor
NTO, Microsoft Malaysia

https://www.youtube.com/watch?v=tKa0zDDDaQk&feature=youtu.be

Mapping CRISP DM Process

Business Understanding

Getting more data

- Asking the sharp question.
- Determining the base ML algorithms.
- Sourcing and Selecting data sources.
- Determining the features and label.

Data Understanding

Importing Data.

Exploring & Visualizing Data

- Univariate plotting.
  - Histogram.
  - Box Plot.
  - Histograms (categorical).
- Plotting 2 Variables.
  - Scatter plots.
- Scatter plot matrices.
- Faceting scatter plots.
- Conditional histogram plots.

Joining data into a single table.

- SQL.
- Programming R, Python.

Feature Selection – numeric features

- Features correlated with label.
- Remove redundant collinear features

Data Preparation

Cleansing:

- Missing value.
- Removing Repeated value.

Handling Error and Outliers:

- Censor
- Trim.
- Interpolate.
- Substitute.

Transforming data:

- Transform string to categories.
- Group/aggregate categories.
- Quantizing continuous variables.
- Convert to indicator values.
- Rename features if required.

Scale numeric features:

- Z-Score scaling.
- Min-max scaling.
- De-trend data.

Feature Engineering:

- Feature Creation:
  - Log of features
  - Feature difference.
  - Add features.
  - Feature multiplication
- Refining Feature Selection
  - Removing marginal impact features.

Handle imbalance categorical labels for classification.

- Coding R/Python.
- SMOTE

Modeling

Classification (A or B)

- Logistic regression.
- Two-Class Boosted Decision Trees.
- Two-Class Decision Forest.
- Two-Class Neural Network.
- Two-Class [Locally-Deep] SVM

Regression (How many)

- Linear regression.
- Boosted Decision Tree Regression.
- Forest regression.
- Neural Network Regression.

Neural Net (Reinforcement Learning)

Clustering (Which Group)

- K-Means.
- Hierarchical Agglomerative Clustering.

Evaluation

Classification (A or B)

- Accuracy.
- AOC.
- Confusion matrix.
- Recall.
- Cross validation.
- Nested cross validation to tune parameters.
- Comparison between models.

Regression (How many)

- Root Mean Square Error.
- Coefficient of determination.
- Residual Visualization:
  - Time series plots of actual versus predicted.
  - Box-Plot
  - Time series plots of residuals (line, histograms) .
- Cross validation.
- Nested cross validation to tune parameters.
- Comparison between models.

Clustering (Which Group)

- Maximal distance to cluster center.
- Principal Component Analysis (PCA)

Model Improvement:

- Select best performing model.
- Identifying marginal impact features.
- Select tuned model parameters.

Deployment

Deployment :

- Deployment lifecycle.
- Web Services APIs
- Web Interface.
- [Deployment coding.]

Advanced Analytics:

- Business Intelligence.
- Predictive Analytics.

Artificial Intelligence:

- AI Applications.
- Cognitive Services.

Sharp questions

What is the stock price tomorrow?

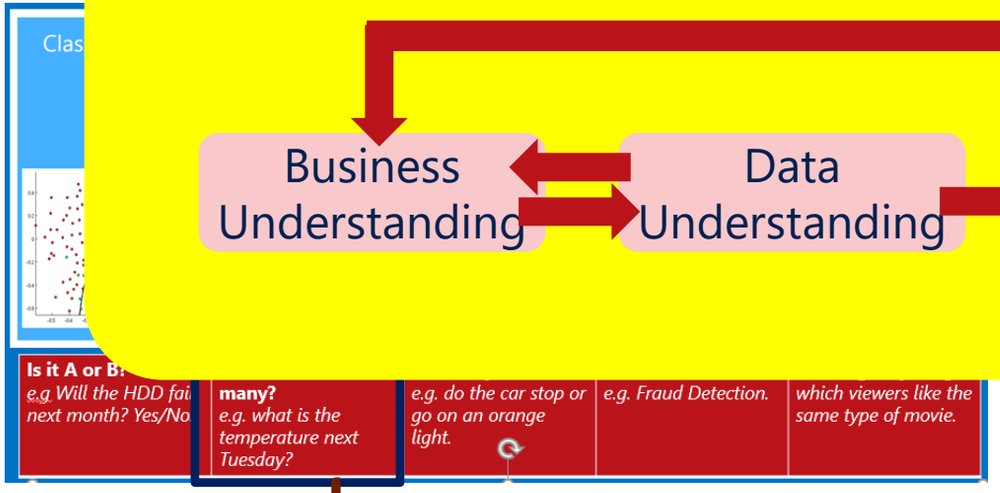
Stock price	Date	Day of week	Dow Jones	Last month sales	Last quarter sales	Market share	New users last month	New users last quarter	Days since press release	Days since product release	Total users
57.3	5/21	Tue	17,245	68.8M	211.2M	23.1%	63,522	195,322	3	96	2.49M
58.8	5/22	Wed	17,289	68.8M	211.2M	23.1%	63,522	195,322	4	97	2.49M
56.9	5/23	Thu	17,115	68.8M	211.2M	23.1%	63,522	195,322	5	98	2.49M
57.4	5/24	Fri	17,278	68.8M	211.2M	23.1%	63,522	195,322	6	99	2.49M

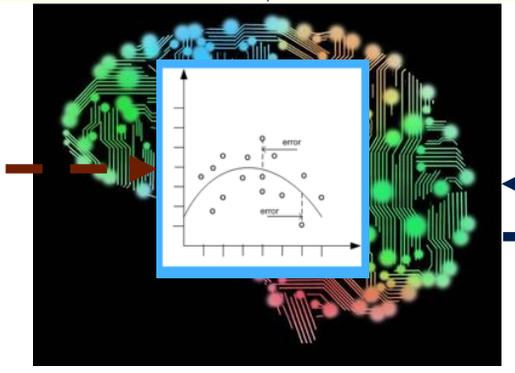
• • • • • • •

57.3	5/21	Tue	17,245	68.8M	211.2M	23.1%	63,522	195,322	3	96	2.49M
58.8	5/22	Wed	17,289	68.8M	211.2M	23.1%	63,522	195,322	4	97	2.49M
56.9	5/23	Thu	17,115	68.8M	211.2M	23.1%	63,522	195,322	5	98	2.49M
57.4	5/24	Fri	17,278	68.8M	211.2M	23.1%	63,522	195,322	6	99	2.49M

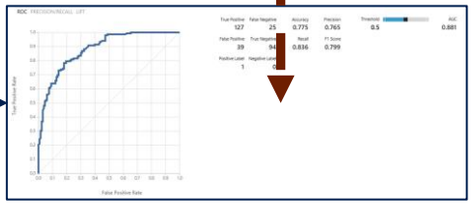
Many Iterations until the target accuracy is achieved

Business Understanding ⇄ Data Understanding → Data Preparation ⇄ Modeling → Evaluation





Evaluate

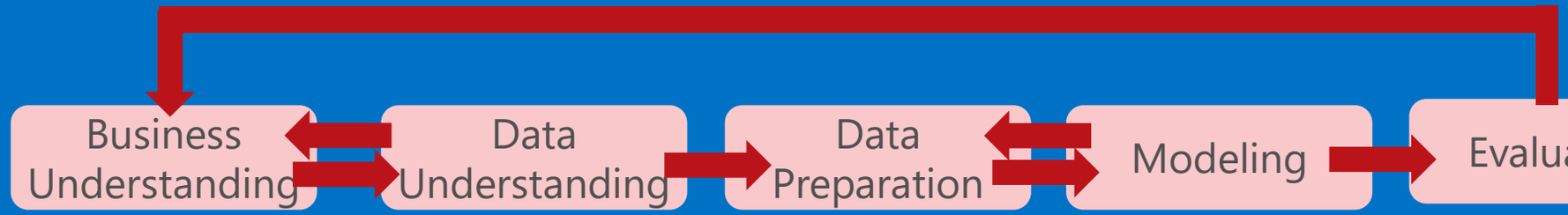


Statistics (e.g RMSE)

Regression (How many)

- Linear regression?
- Boosted Decision Tree Regression?
- Forest regression?
- Neural Network Regression.?

Mapping CRISP DM Process



Business Understanding

Getting more data

- Asking the sharp question.
- Determining the base ML algorithms.
- Sourcing and Selecting data sources.
- Determining the features and label.

Data Understanding

Importing Data.

Exploring & Visualizing Data

- Univariate plotting.
  - Histogram.
  - Box Plot.
  - Histograms (categorical).
- Plotting 2 Variables.
  - Scatter plots.
- Scatter plot matrices.
- Faceting scatter plots.
- Conditional histogram plots.

Joining data into a single table.

- SQL.
- Programming R, Python.

Feature Selection – numeric features

- Features correlated with label.
- Remove redundant collinear features

Data Preparation

Cleansing:

- Missing value.
- Removing Repeated value.

Handling Error and Outliers:

- Censor
- Trim.
- Interpolate.
- Substitute.

Transforming data:

- Transform string to categories.
- Group/aggregate categories.
- Quantizing continuous variables.
- Convert to indicator values.
- Rename features if required.

Scale numeric features:

- Z-Score scaling.
- Min-max scaling.
- De-trend data.

Feature Engineering:

- Feature Creation:
  - Log of features
  - Feature difference.
  - Add features.
  - Feature multiplication
- Refining Feature Selection
  - Removing marginal impact features.

Handle imbalance categorical labels for classification.

- Coding R/Python.
- SMOTE

Modeling

Classification (A or B)

- Logistic regression.
- Two-Class Boosted Decision Trees.
- Two-Class Decision Forest.
- Two-Class Neural Network.
- Two-Class [Locally-Deep] SVM

Regression (How many)

- Linear regression.
- Boosted Decision Tree Regression.
- Forest regression.
- Neural Network Regression.

Neural Net (Reinforcement Learning)

Clustering (Which Group)

- K-Means.
- Hierarchical Agglomerative Clustering.

Evaluation

Classification (A or B)

- Accuracy.
- AOC.
- Confusion matrix.
- Recall.
- Cross validation.
- Nested cross validation to tune parameters.
- Comparison between models.

Regression (How many)

- Root Mean Square Error.
- Coefficient of determination.
- Residual Visualization:
  - Time series plots of actual versus predicted.
  - Box-Plot
  - Time series plots of residuals (line, histograms) .
- Cross validation.
- Nested cross validation to tune parameters.
- Comparison between models.

Clustering (Which Group)

- Maximal distance to cluster center.
- Principal Component Analysis (PCA)

Model Improvement:

- Select best performing model.
- Identifying marginal impact features.
- Select tuned model parameters.

Deployment

Deployment :

- Deployment lifecycle.
- Web Services APIs
- Web Interface.
- [Deployment coding.]

Advanced Analytics:

- Business Intelligence.
- Predictive Analytics.

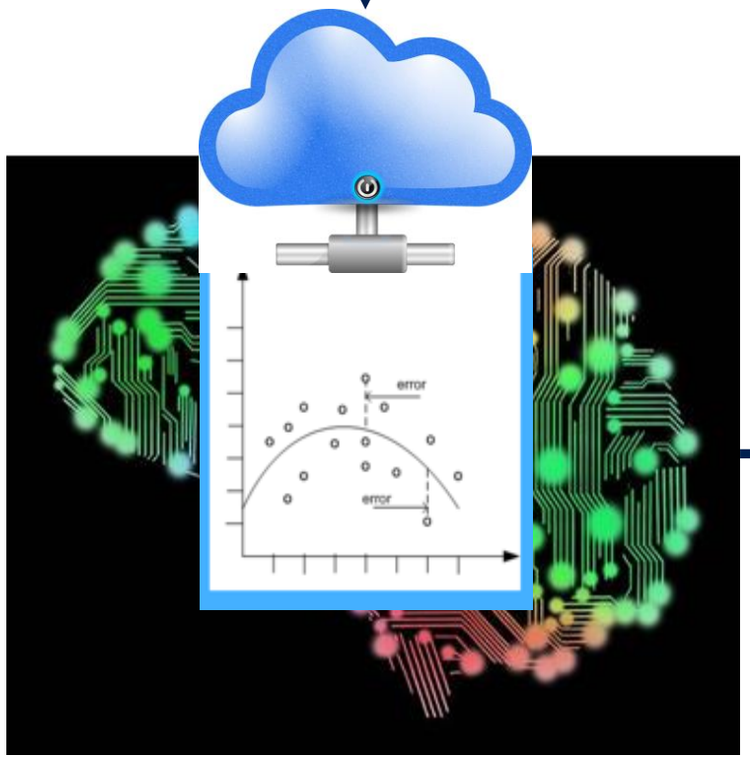
Artificial Intelligence:

- AI Applications.
- Cognitive Services.

Stock price	Date	Day of week	Dow Jones	Last month sales	Last quarter sales	Market share	New users last month	New users last quarter	Days since press release	Days since product release	Total users
?	5/27	Mon	16,935	68.8M	211.2M	23.1%	63,522	195,322	9	102	2.50M

Develop App

Deploy Web Service



Trained Model

Stock price
55.3

Tools

End to end Data Science Activities

Dzahar Mansor
NTO, Microsoft Malaysia

https://www.youtube.com/watch?v=tKa0zDDDaQk&feature=youtu.be

Mapping Generic Tools to CRISP DM Process

RDBMS

Data warehouse/Datamart platforms

BI Cubes

Big Data Storage (Data Lake, HDFS)

Storage Platforms

Legend:

General purpose

Big Data

Custom

ETL, ELT, EII, EAI

Big Data ETL , Streaming Application

Data Management Tools

Custom

ML

Stats Pkg

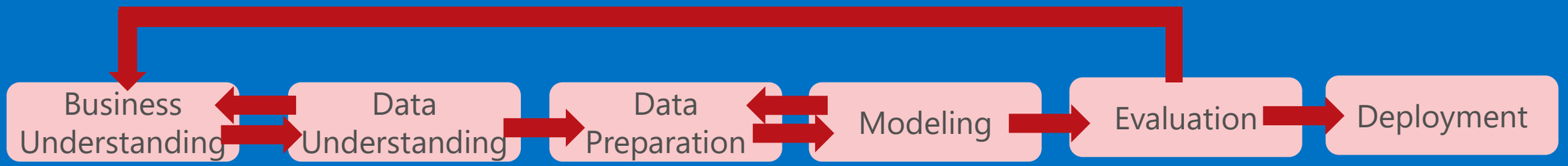
Data Mining Tools

Custom

BI Dashboard

Middleware

BI & Web Services



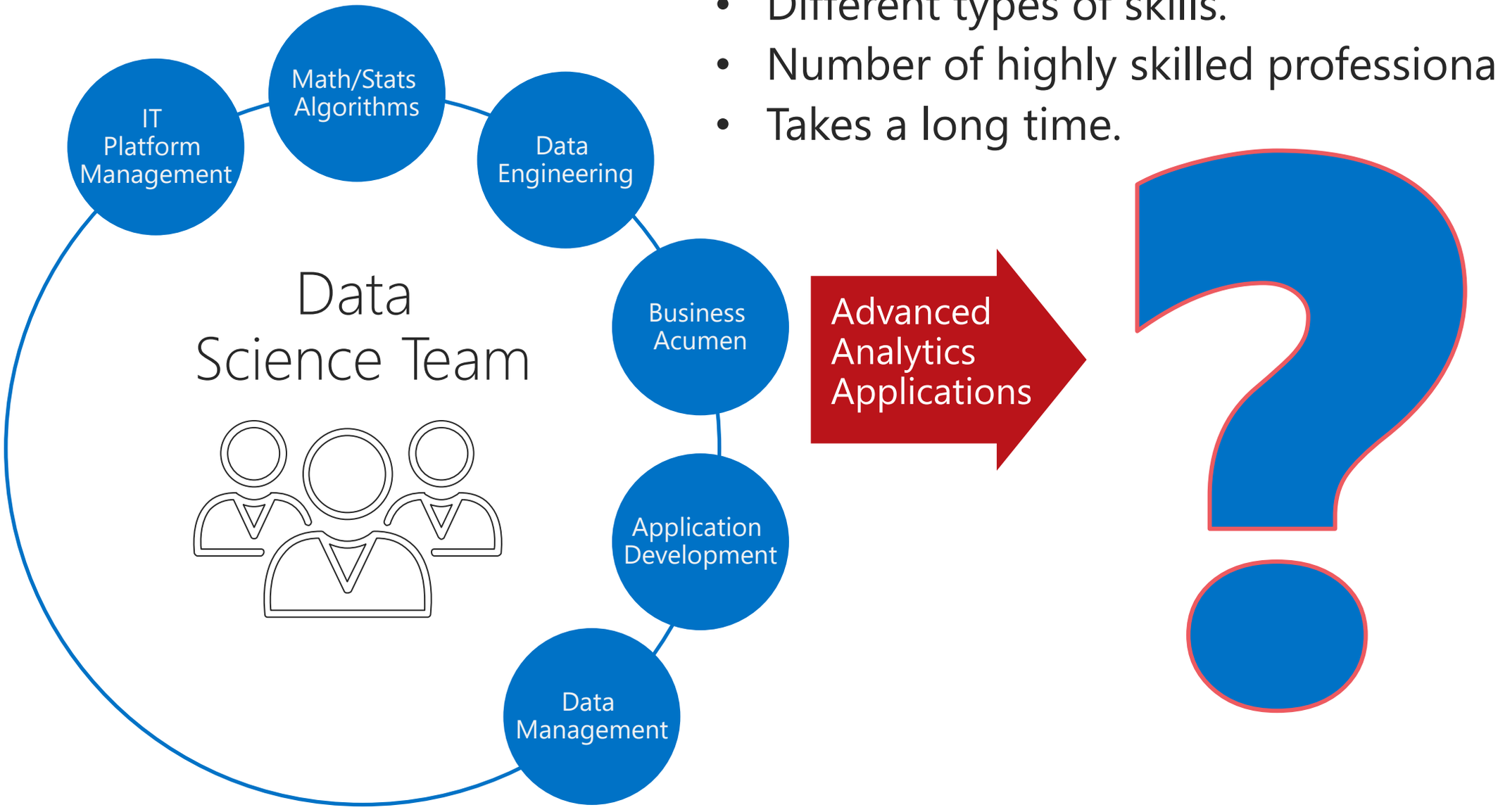
Democratizing Data Science

https://docs.microsoft.com/en-us/azure/machine-learning/preview/overview-what-is-azure-ml

Dzahar Mansor
dmansor@microsoft.com

Data Science Team for Advanced Analytics - Challenges

- Different types of skills.
- Number of highly skilled professionals.
- Takes a long time.



Understanding the role of Mathematics

- **Given data x_1, x_2, …., x_{t-1} predict x_t**
- **Or $P(x_t \mid x_1, x_2, \ldots, x_{t-1})$**

Machine Learning

- Supervised learning.
- Unsupervised learning.
- Reinforced learning

$$P(y|x) = \frac{P(x|y)P(y)}{P(x)}$$

$$P(x|m) = \int P(x|\theta, m)P(\theta|m)d\theta$$

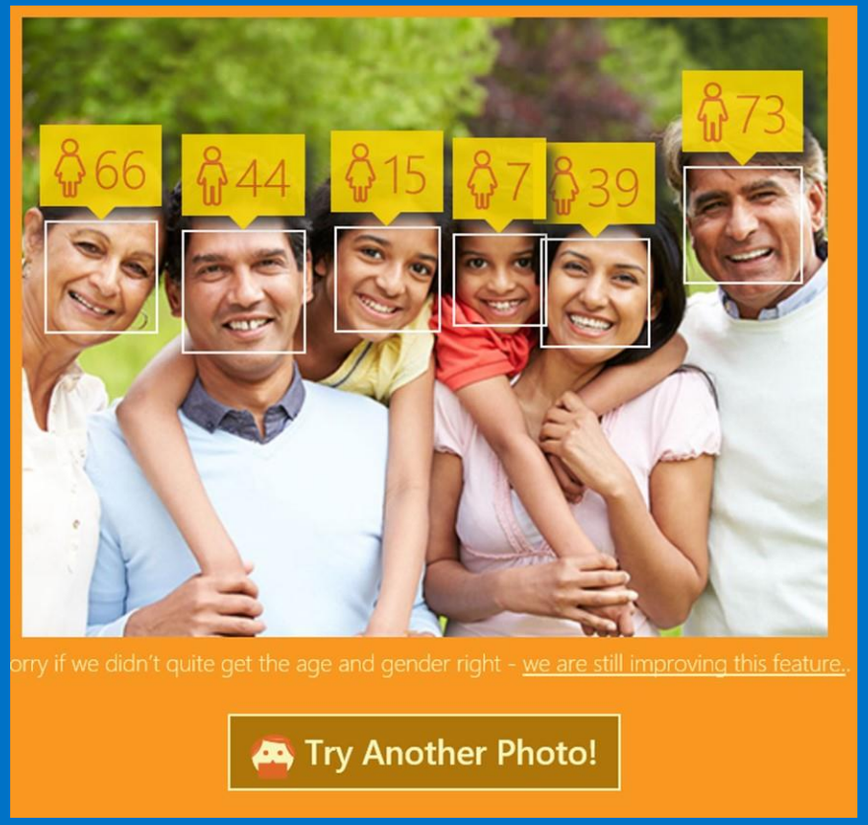
$$\hat{\theta}_{\mathrm{MAP}} = \arg\max_{\theta} P(\theta|\mathcal{D}, m) = \arg\max_{\theta} \left[\log P(\theta|m) + \sum_n \log P(x_n|\theta, m)\right]$$

$$\hat{\theta}_{\mathrm{ML}} = \arg\max_{\theta} P(\mathcal{D}|\theta, m) = \arg\max_{\theta} \sum_n \log P(x_n|\theta, m)$$

Selection or combining following mathematical modelling:

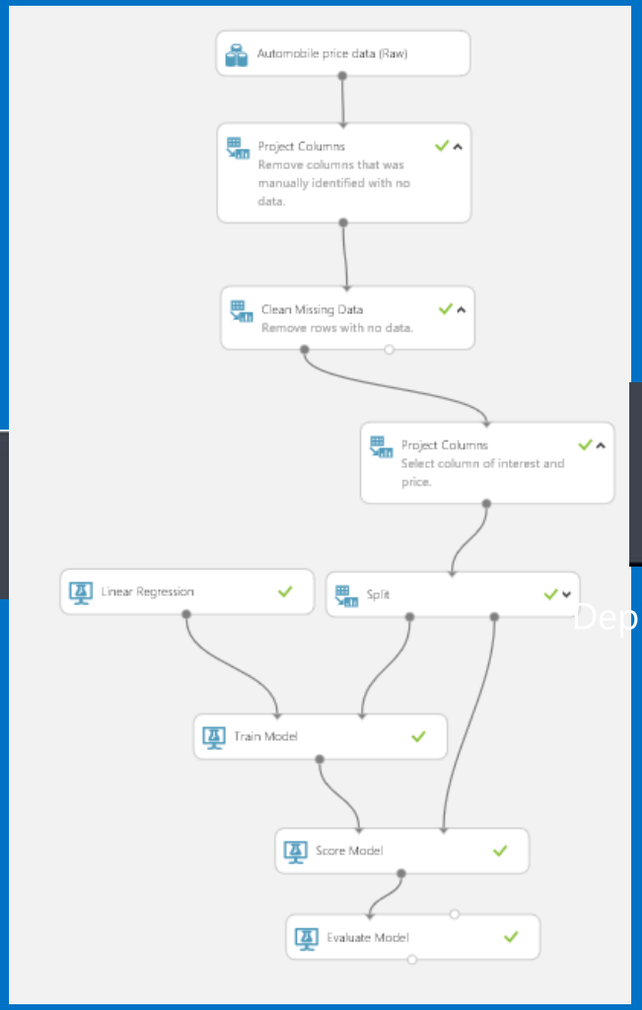
- Latent variable models (e.g. factor analysis, K-means etc).
- EM algorithms (ML estimation).
- Modelling time series.
- Nonlinear, factorial models & hierarchical models.
- Graphical models.

Example : How-Old.net



- Went viral : 50 Million Users in 7 Days, 1.2M users/hour.

Machine Learning Development tools



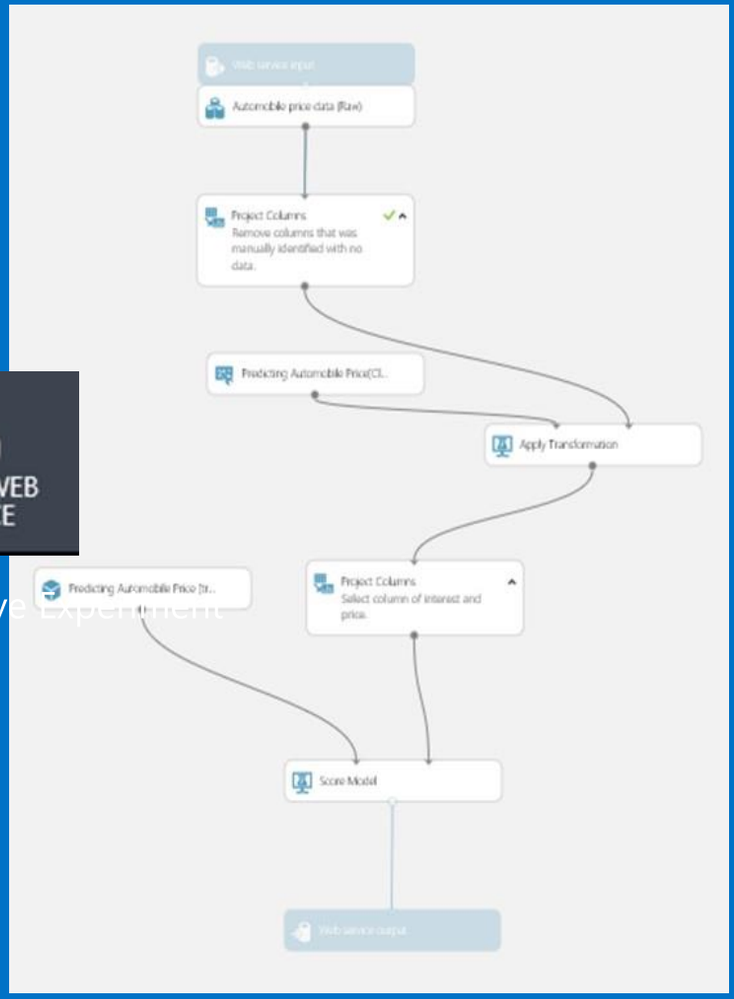
SAVE AS

SET UP WEB SERVICE

Deploy Web Service

SET UP WEB SERVICE

Update Predictive Experiment



Cognitive Services

Vision

- Computer vision
- Face
- Emotion
- Video

Speech

- Speaker recognition
- Speech
- Custom Recognition

Language

- Text analysis
- Bing speller
- Web language model
- Linguistic analysis
- Language understanding
- Translator

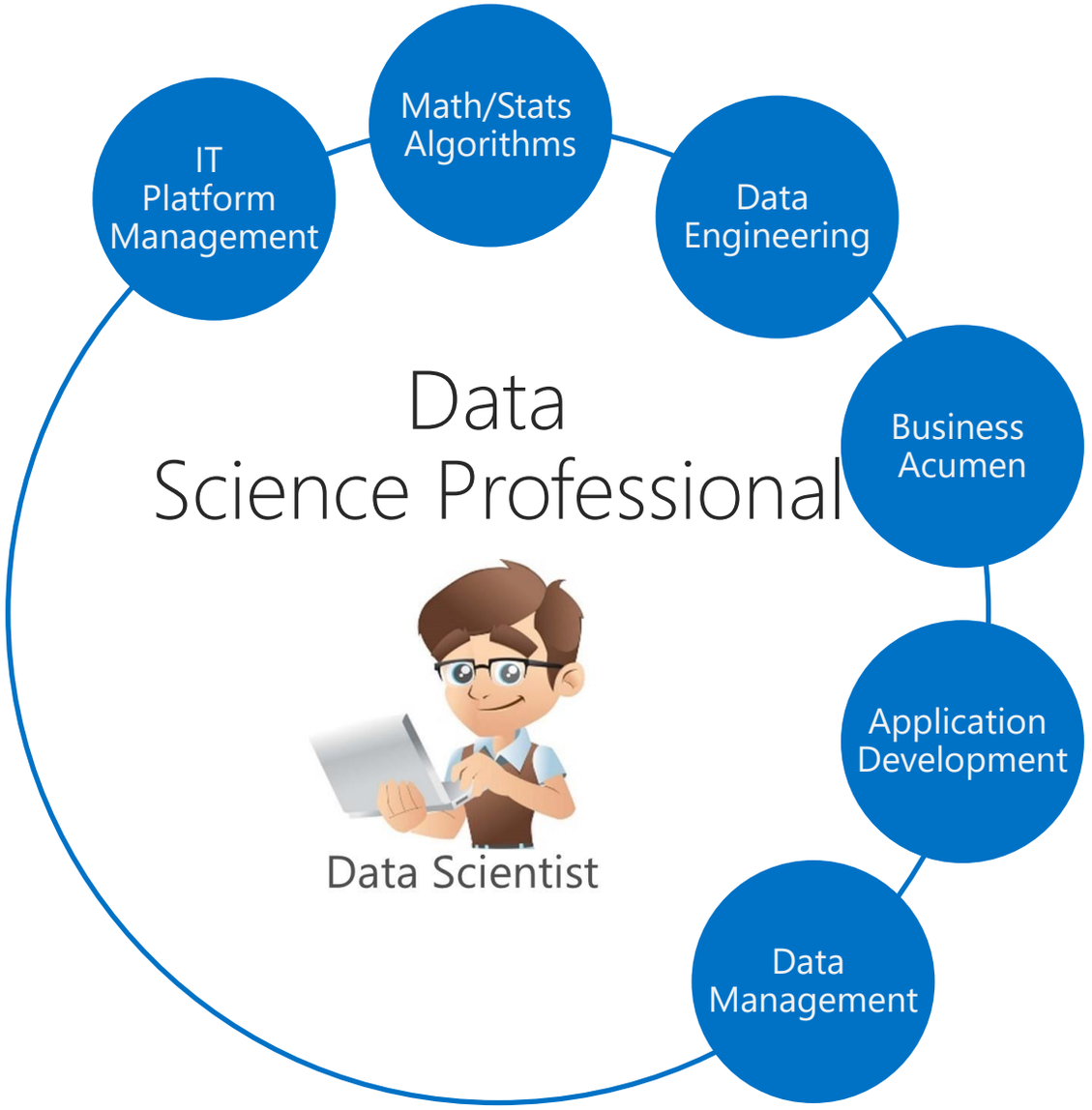
Knowledge

- Academic knowledge
- Entity linking services
- Knowledge exploration service
- Recommendations

Search

- Bing search API
- Bing image search API
- Bing video search API
- Bing news search API
- Bing auto suggestions API

Democratizing Data Science



Advanced Analytics Applications

Data Dividend at lower cost and faster turn-around

Why Azure ML

- Easy to use.
- Quick to deploy production solutions as web services.
- Model runs in a highly scalable hyperscale cloud.
- Secure cloud environment for data and code.
- Powerful, efficient built-in algorithms.
- Extensible – SQL, Python and R.
- Incorporates Jupyter Notebook – R & Python.

Azure ML Free Tier Account

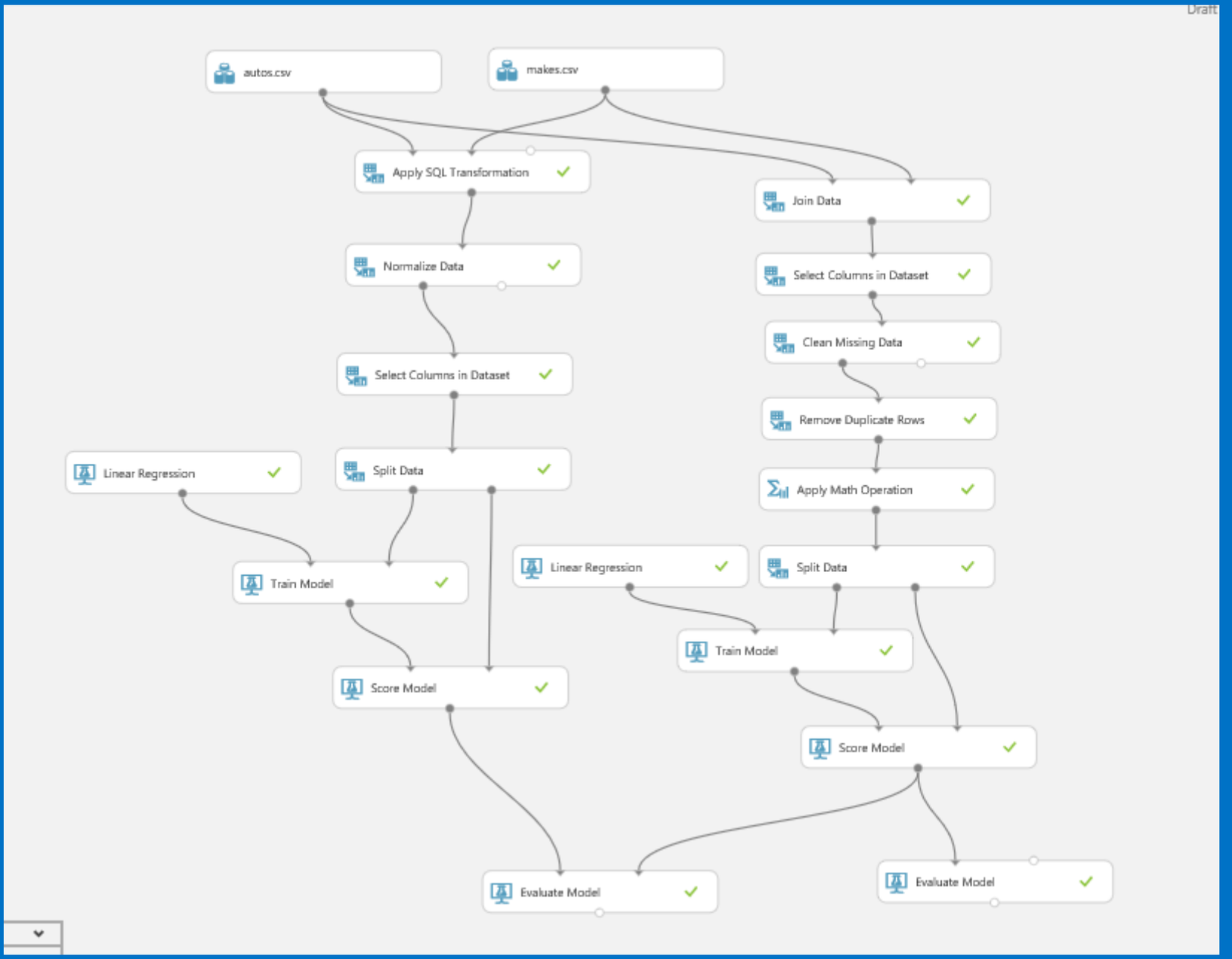
- Free Tier Account.
- Unlimited time, with restricted priority.
- Paid account provides full performance.

https://studio.azureml.net

Hands on lab :
https://1drv.ms/u/s!Ah-e-4zJ5AEccalhapMEb_ePMvA?e=THbNey
Lab.zip

Demo

https://1drv.ms/u/s!AscnSCoIDa9fdN5eMWqpmzb7USM?e=d1eob3
Lab.zip



Microsoft AI: Amplifying Human Ingenuity

to empower every person and every organization on the planet to achieve more (watch)

Dr Dzahar Mansor
National Technology Officer
Microsoft Malaysia
dmansor@microsoft.com
www.linkedin.com/in/dzahar-mansor