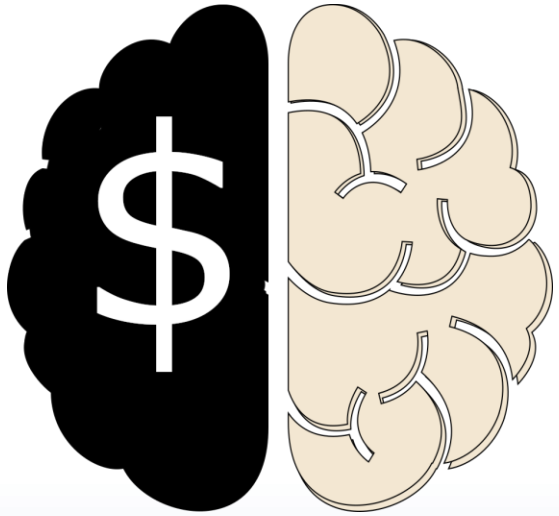
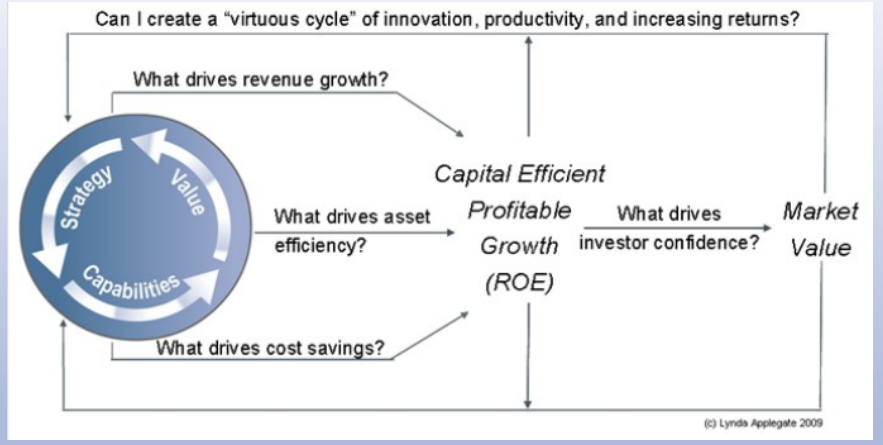
Business Intelligence & Big Data Analytics

ANANG SYARIFUDIN A., ST, MTI

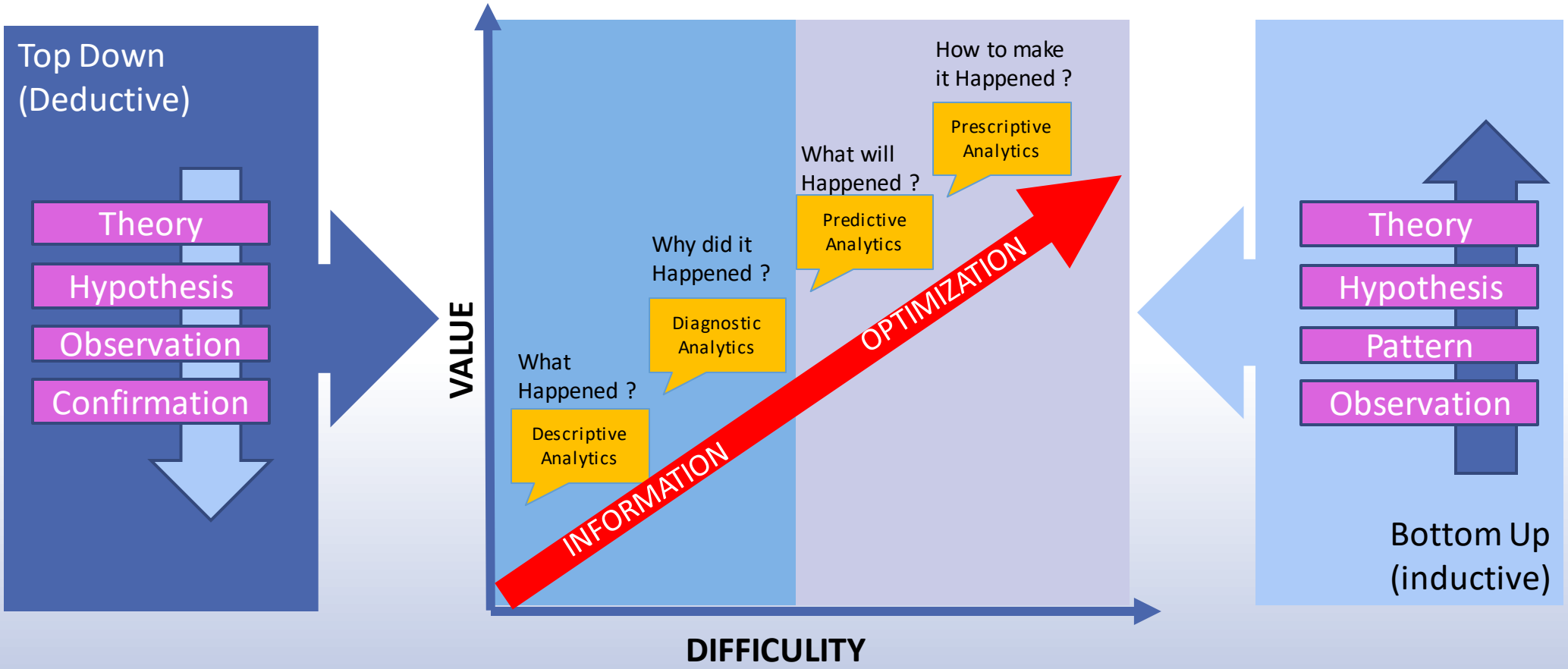
Business and Data Analytics



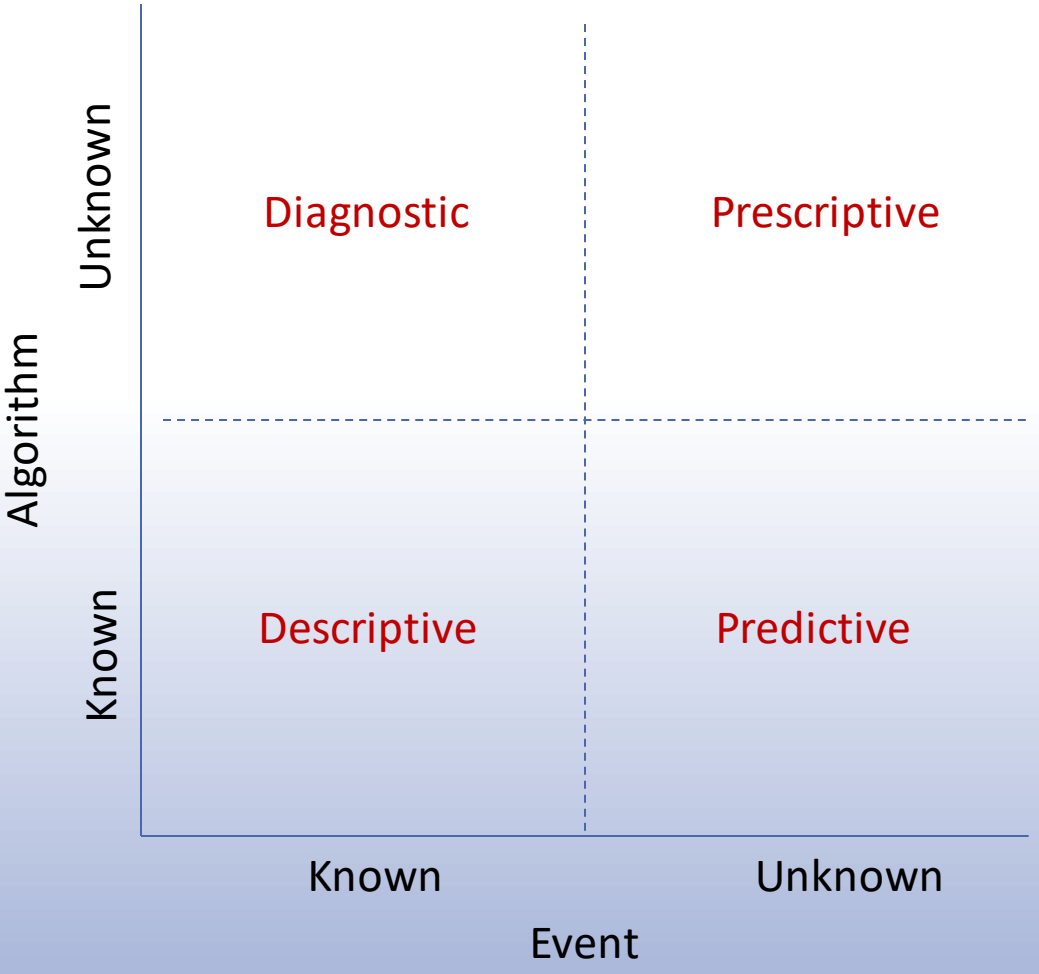
"Big Data are high-volume, high-velocity, and/or high-variety information assets that require new forms of processing to enabe enhanced decision making, insight discovery and process optimization" (Gartner 2012)

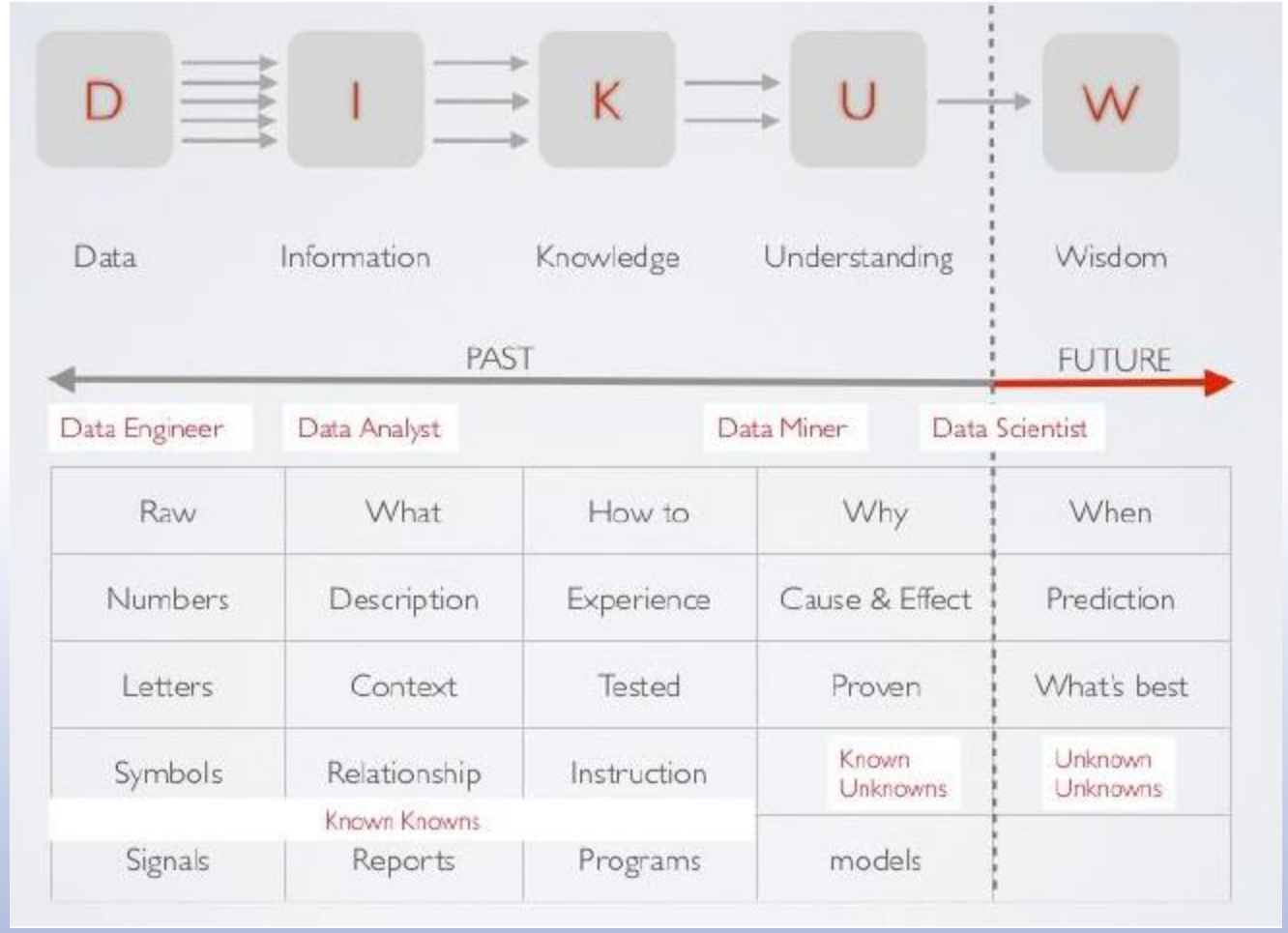


Big Data Analytics Evolution

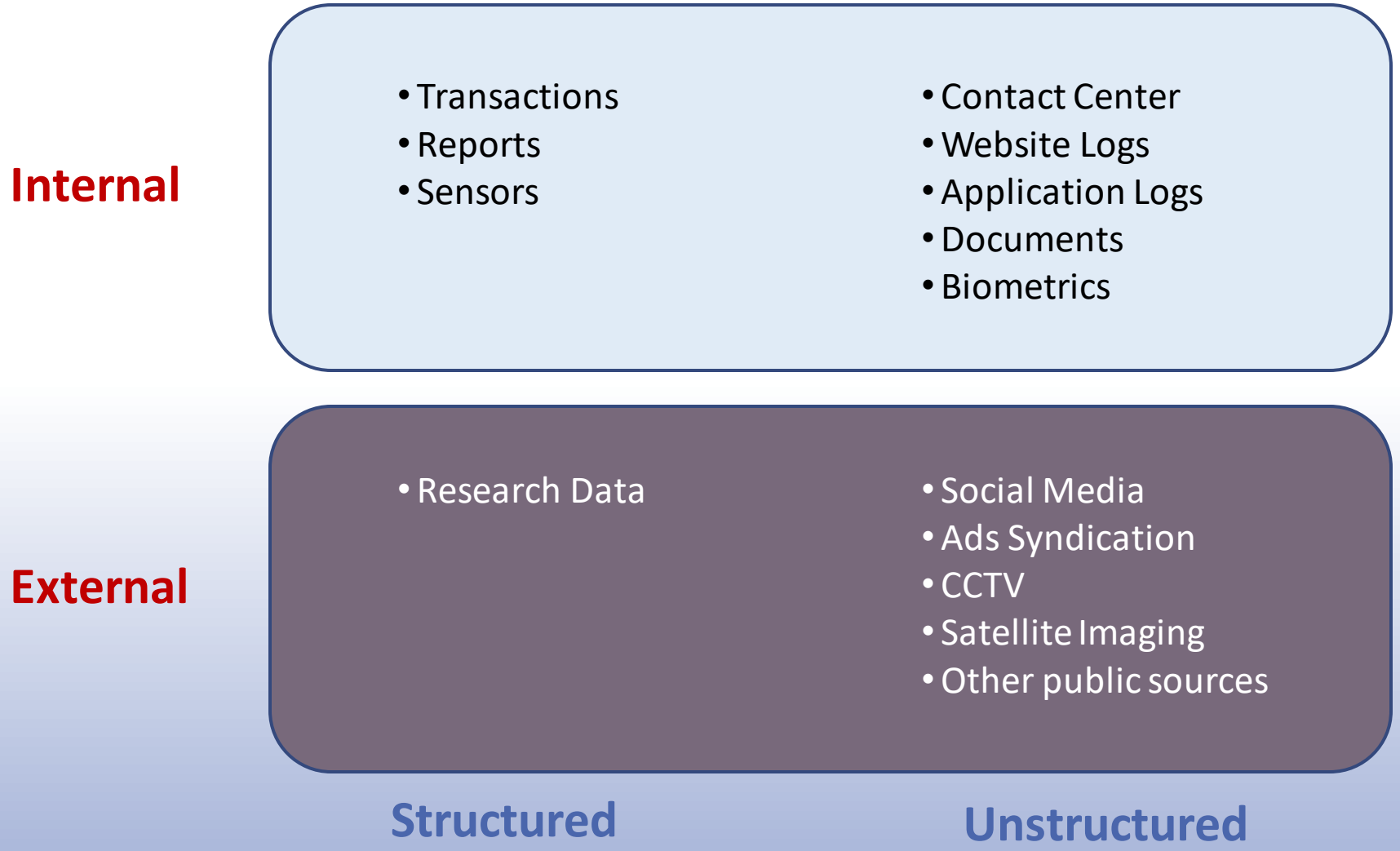


Known & Unknown

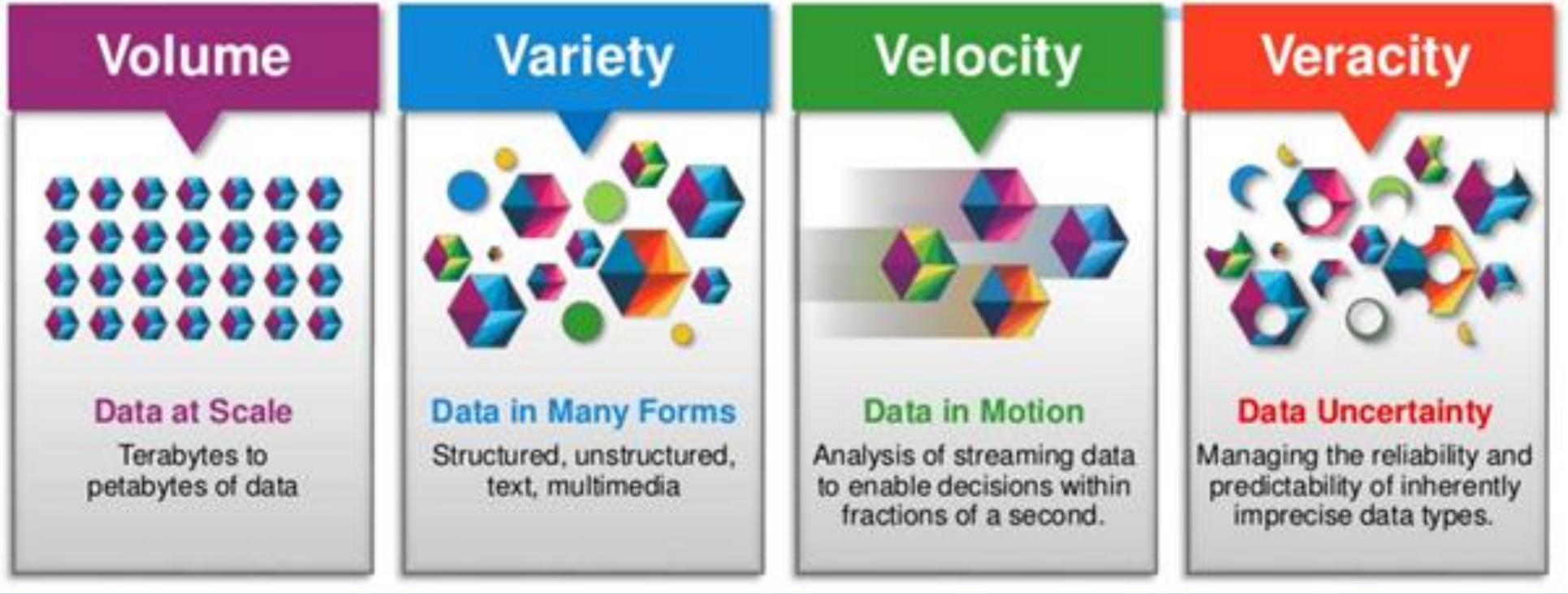




Data Source


Internal
• Transactions
• Reports
• Sensors
• Contact Center
• Website Logs
• Application Logs
• Documents
• Biometrics
External
• Research Data
• Social Media
• Ads Syndication
• CCTV
• Satellite Imaging
• Other public sources
Structured
Unstructured

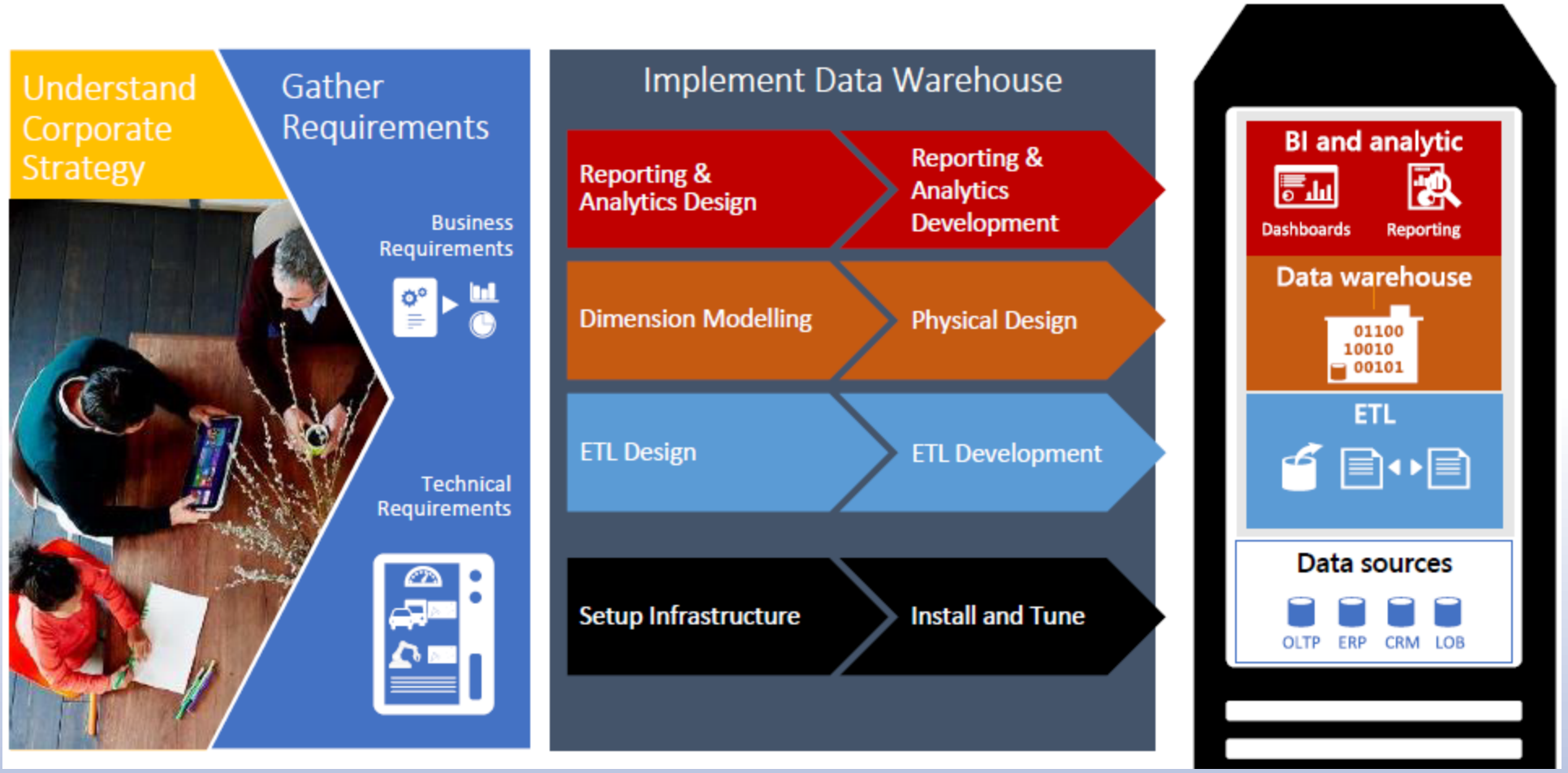
Big Data Characteristic



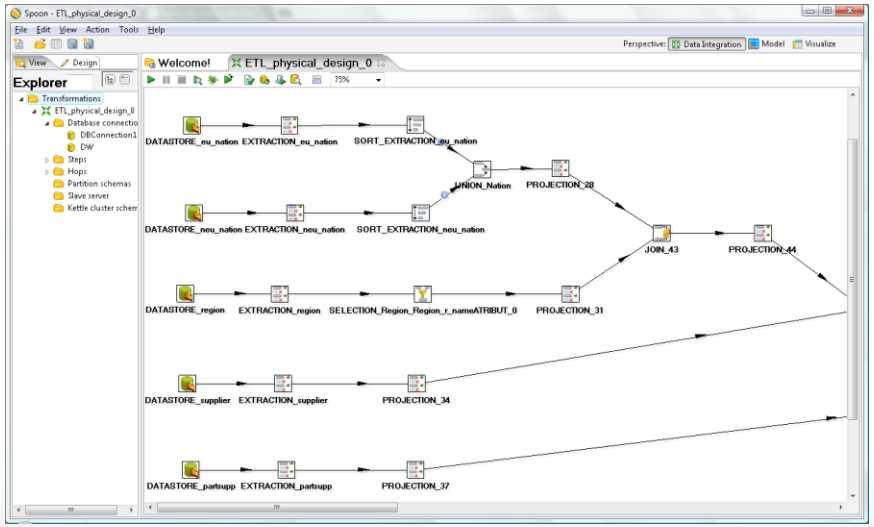
Descriptive Analytics

Describe or summarize past data, find the hindsight, monitoring KPI, answering question of “What is happened ?”

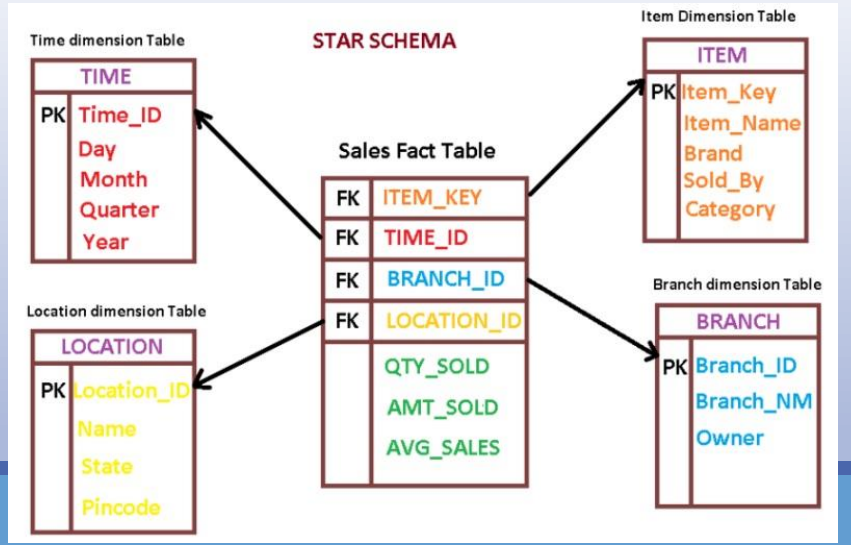
Datawarehousing



Datawarehouse & ETL

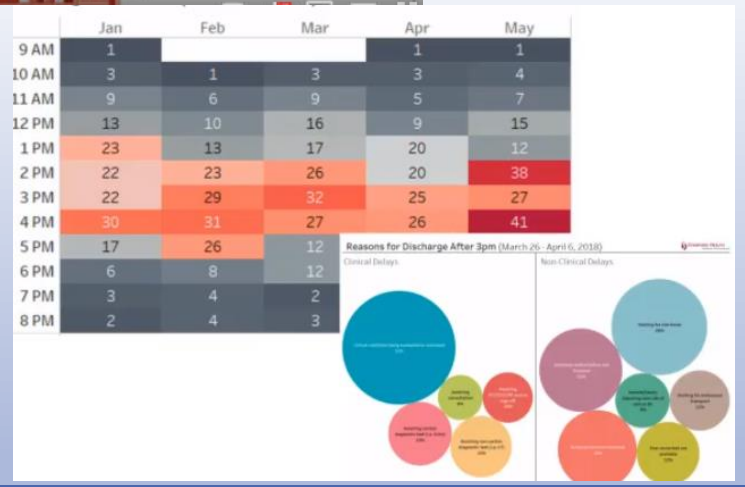


t Numt	Client Name	Street Address	City	State	Zip Code	Amoumt Pai
	Ann Toney PC Attourney	PO Box 1022	Meeker	CO	81641	$40.00
	Borchard Kent A. Att.	335 6th St #1	Meeker	CO	81641	$50.00
	Brooks Laurie J Appraiser	889 Main Street	Meeker	CO	81641	$250.00
V	Coulter Aviation	921 Market Street	Meeker	CO	81641	$50.00
I	Meeker Airport	921 Market Street	Meeker	CO	81641	$40.00
O	Meeker Collision Center	43904 Hwy 13	Meeker	CO	81641	$40.00
U	Northwest Auto	485 Market Street	Meeker	CO	81641	$50.00
C	Rosken LLC Accountant	592 Main St Suite 1	Meeker	CO	81641	$40.00
O	Rocky Mountain Bowstrings	696 Main Street	Meeker	CO	81641	$50.00
R	Zagar-Brown Trina K Att.	685 Main Street Suite 5	Meeker	CO	81641	$150.00



Diagnostic Analytics

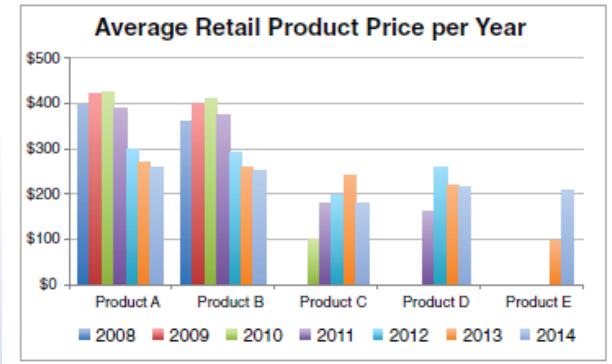
Advance Analytics with capability to: detect anomalies, drill-down/discovery information and/or find causality relations which can answer question: “Why did it happened ?”

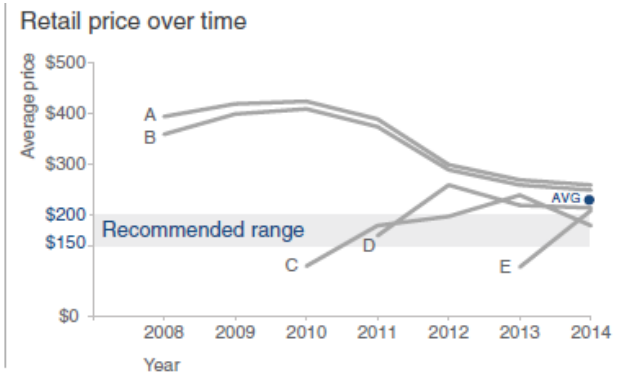


Business Intelligence Challenge

- Choosing & utilizing right database

- Effectively using right ETL tool





- Visualize & story telling with data effectively

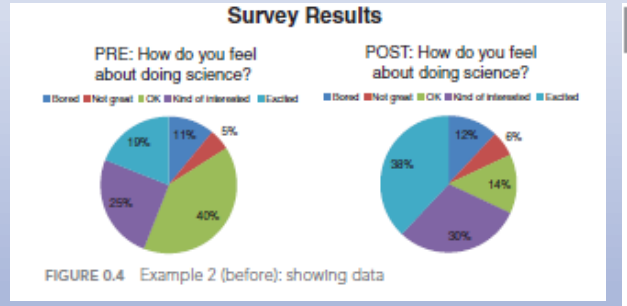


FIGURE 0.4 Example 2 (before): showing data

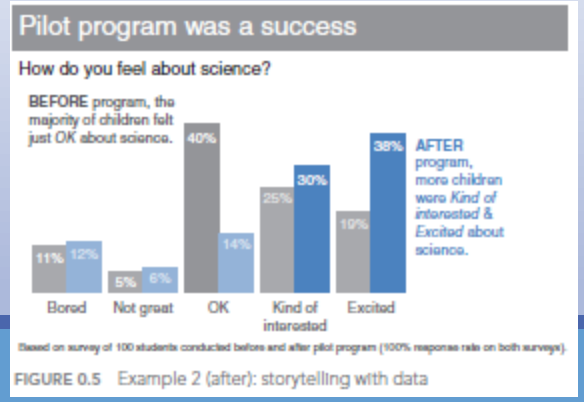
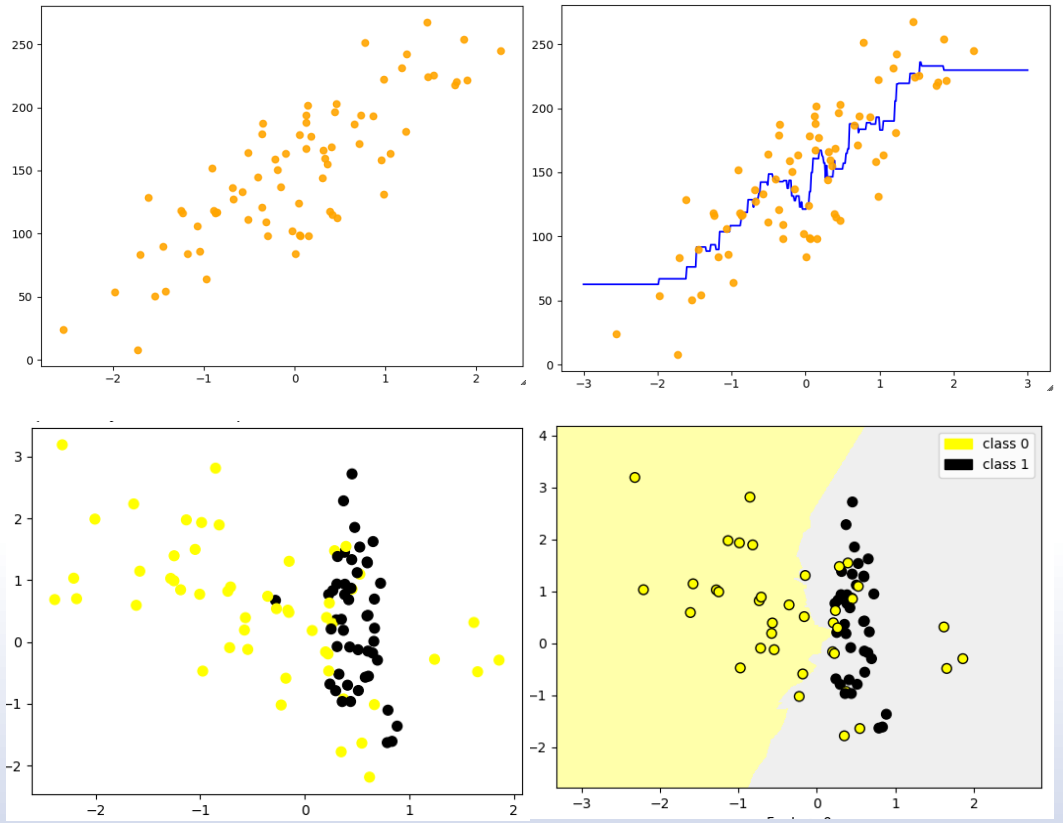
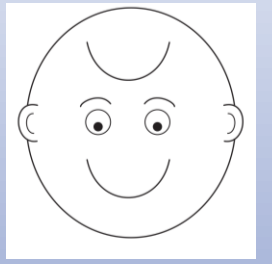


FIGURE 0.5 Example 2 (after): storytelling with data

Predictive Analytics

Predictive analytics is the use of data, statistical algorithms and machine learning techniques to identify the likelihood of future outcomes based on historical data. The goal is to go beyond knowing what has happened to providing a best assessment of what will happen in the future.





Machine Learning

Computer algorithms which learn concepts and make subsequent predictions in the presence of data without explicitly being programmed to understand those data.

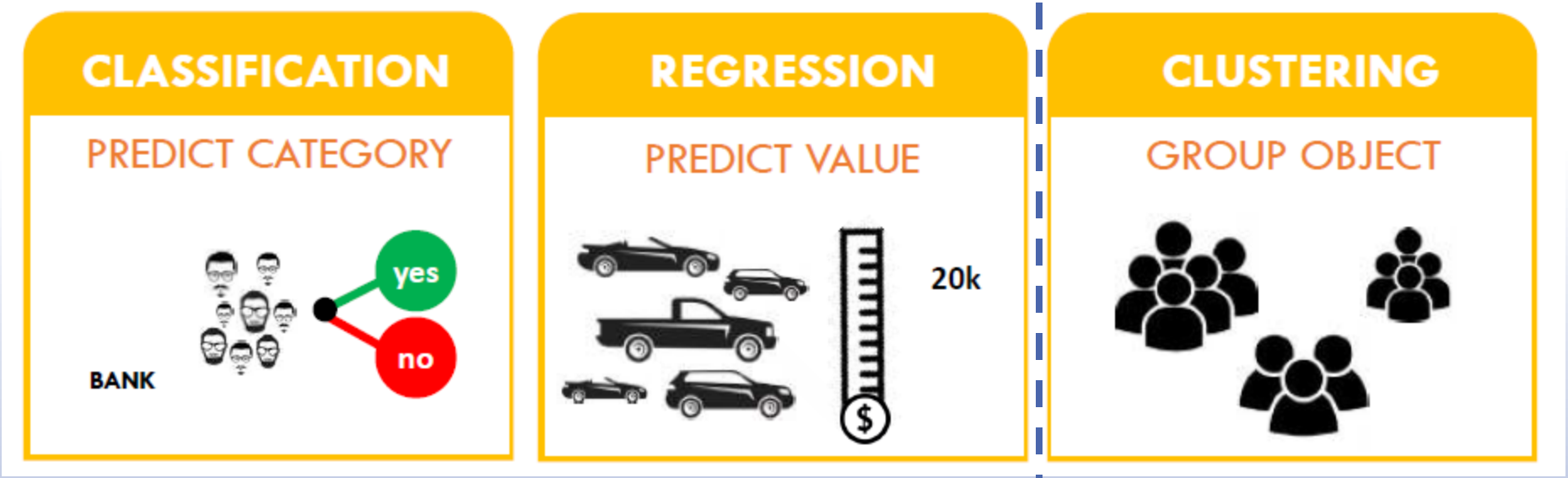
```
for i in range(train):
        y_predict[i] = coef1*train[i]['x1']+coef2*train[i]['x2']+C
```

```
model =KNeighborsRegressor(k=10).fit(X_train,y_train)
Y_predict = model.predict(X_test)
```

Machine Learning Application

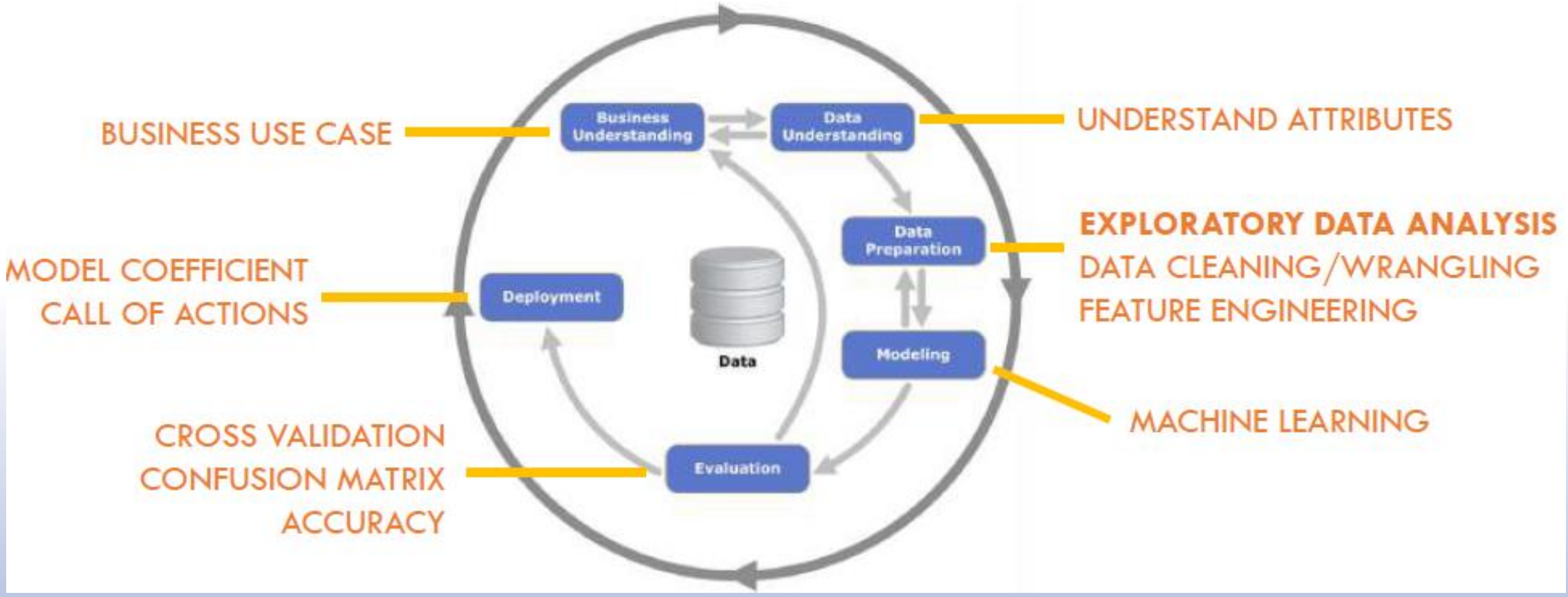
Supervised Learning

Unsupervised Learning

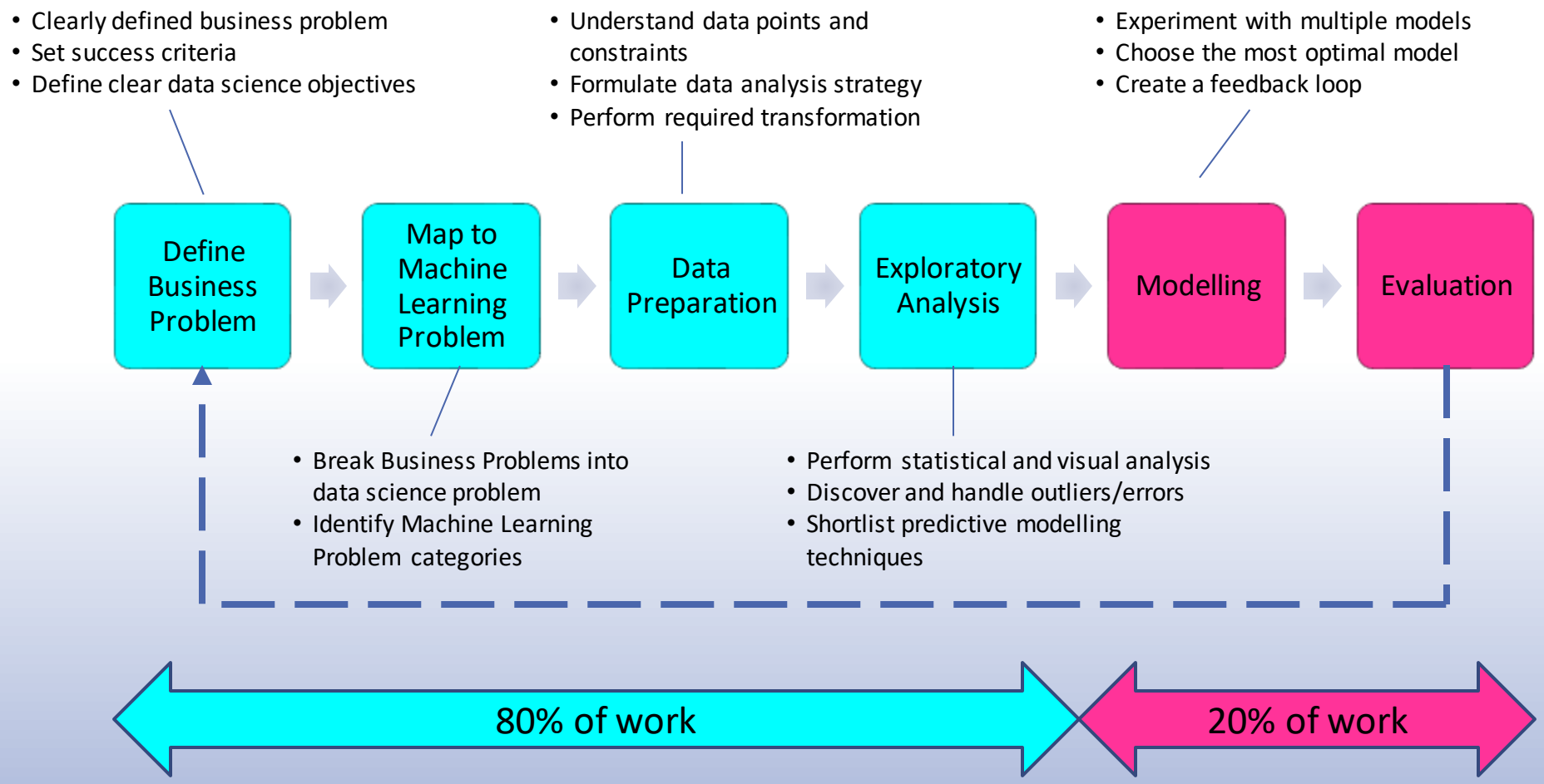


CRISP – DM

Cross Industry Standard Process– Data Mining



Practical & Main Stages



Business Case Example

A Company wants to increase next year profit by 10%

#1 This business problem can be break into detailed business questions:

- How to increase revenue ?
  - How to acquire new customer by 20 %?
  - How to increase revenue generated from existing customers (ARPU increase 15%) ?
- How to reduce lost ?
  - How to reduce production cost by 30% ?
  - How to prevent customer churn from 5% to 3% ?

#2 Translate business questions into machine learning tasks:

- Identify factors which leads into customer churn
- Detect customer that has high propensity to churn
- Identify factors which leads customer buy services
- Identity customer which can be target for up-sell campaign
- Etc..

#3 Data Preparation

One we have defined the business problem and decomposed into machine learning tasks, **we need to dive deeper into the data.**

- Covers activities to **construct final dataset (data for modelling)** from various type of data.
- Likely to be **performed multiple times, iterative and not in any prescribed order**.
- **Key things to note is the source of the data, quality of the data, data bias**, etc.
- Task included: **record selection, feature selection, cleaning, data transformation.**

cust_id	yr_mth	voice_duration	data_usage
234234	201801	40	5000
234234	201802	45	4500
234234	201803	60	3000
234237	201801	40	NULL
234237	201802	70	NULL
234237	201803	30	NULL

date_id	cust_id	problem_cat	problem_text
20180304	234237	drop call	telpon putus-putus
20180318	234237	sim error	tidak bisa konek

3/12/2019 17:03	234234	login
3/13/2019 18:07	234235	check balance
3/14/2019 17:03	234236	reload
3/15/2019 17:03	234237	login

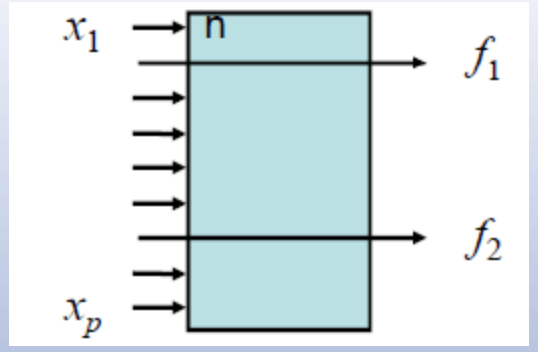
Denormalization

cust_id	voice_duration_m1	voice_duration_m2	voice_duration_m3	data_usage_m1	data_usage_m2	data_usage_m3	no_complain	app_access
234234	40	45	60	5000	4500	3000	NULL	5
234237	40	70	30	0	0	0	2	NULL

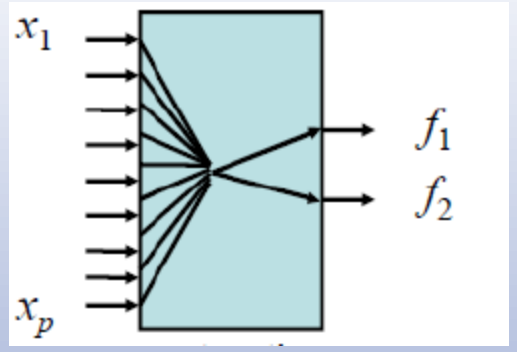
#4 Exploratory Data Analysis

EDA objectives

- Extract information/pattern from data
- Suggest hypotheses about causes of observed phenomena
- Give idea of features predicted power
- Identify the need for feature engineering: feature selection & feature extraction
- Provide basis for further data collection



selection



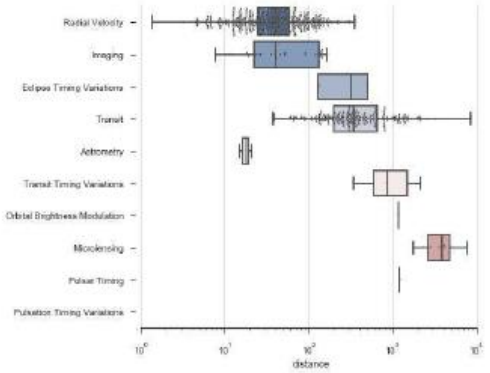
extraction

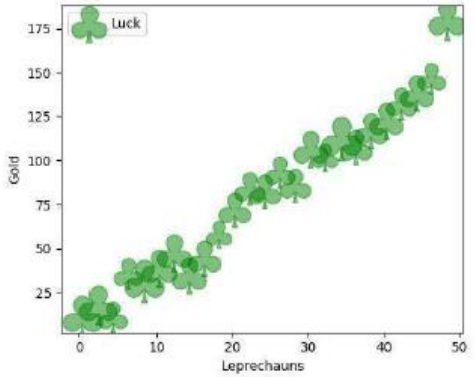
Exploratory VS Explanatory Analysis

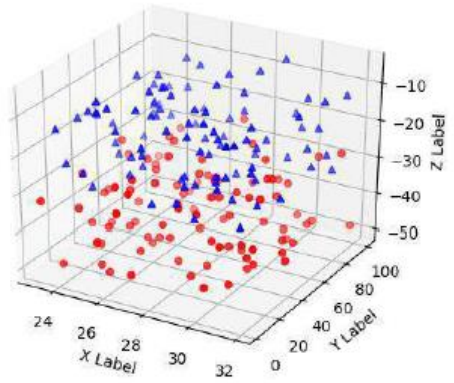
	Exploratory	Explanatory
Purpose	To Explore Understand the data Figure out what might be noteworthy *Opening hundreds of oysters* 	To explain Communicate via data Communicate what is noteworthy *Find perhaps only 2 pearls* 
Visualization	○ Best done using data with high level of granularity. ○ Possible presence of noise, but oversimply could end up missing information. ○ Not editorially driven. ○ Emphasis is discovering many stories in the visual. ○ May not even be sure what story is there in the data.	○ Low level data granularity, aggregated, summary. ○ Editorially driven, craft the visual with care to bring out the story most clearly. ○ Taking into account who's the audience, the background. ○ Emphasis is communicate analysis/data exploration result. ○ Defined and clear story.

Translate data into visual medium can help quickly identify features, trends or anomalous outliers. Most *Exploratory Data Analysis (EDA)* are graphical in nature.

Exploratory Data Visualization



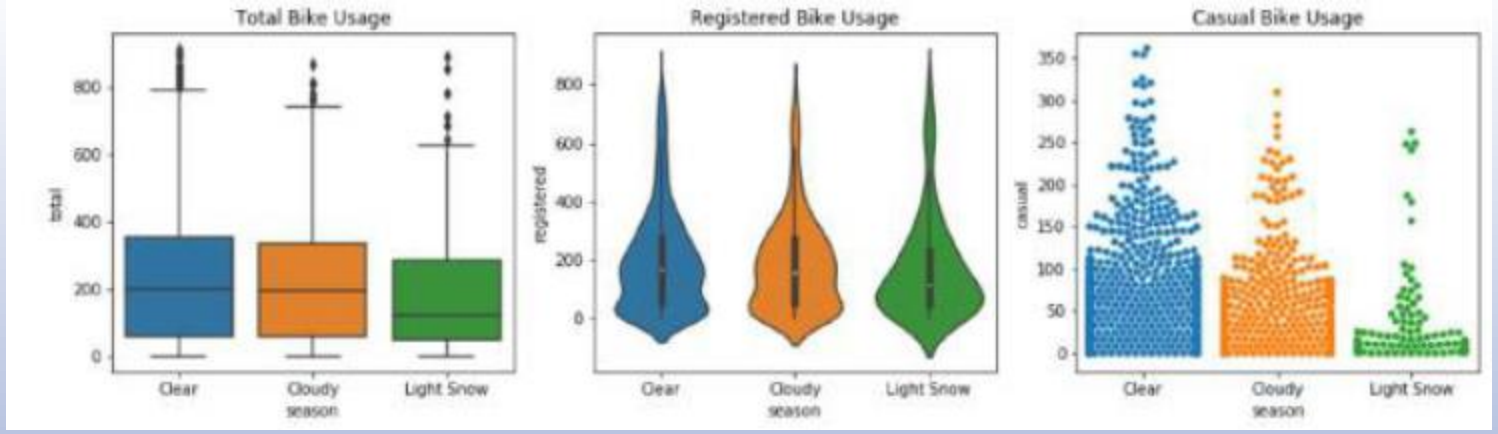




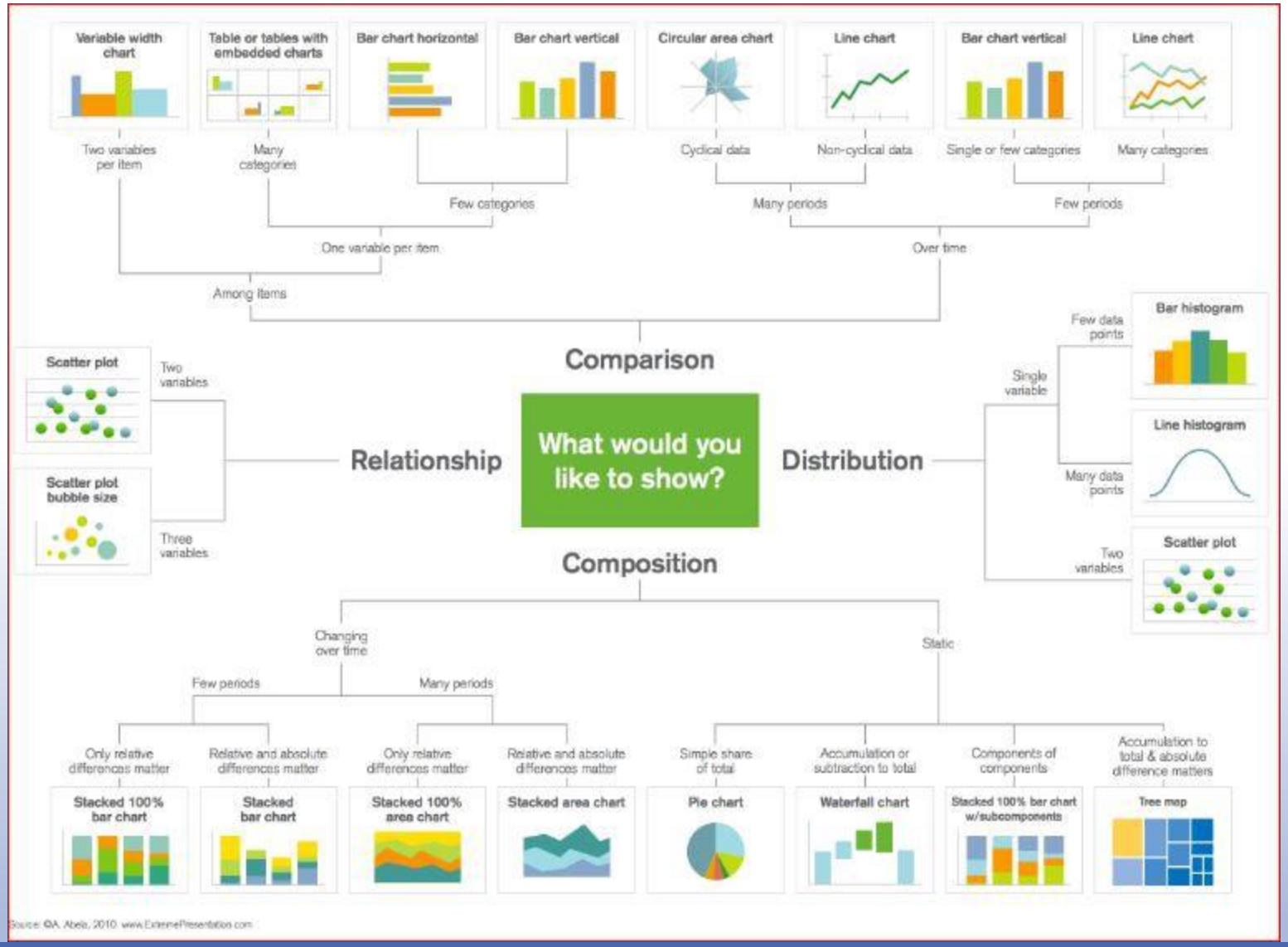
Univariate

Bivariate

Multivariate



Which Visualization, When ?

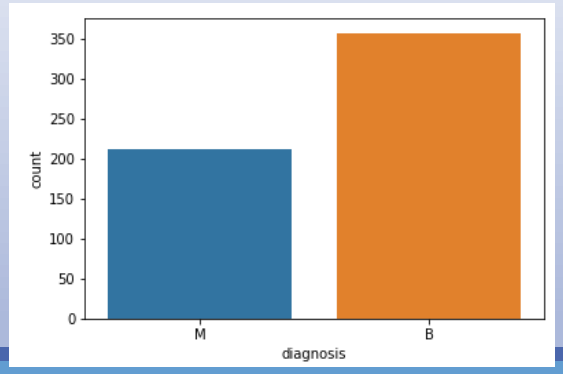


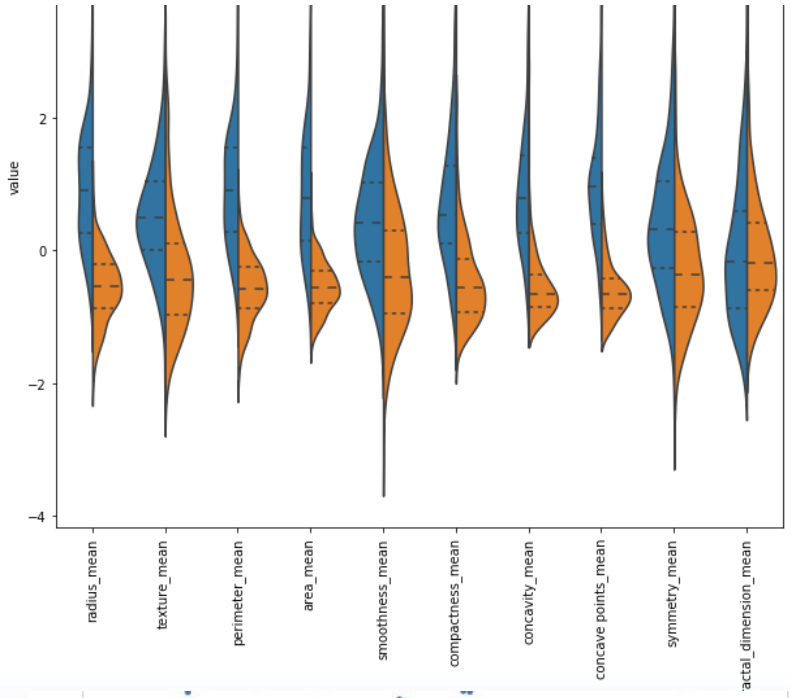
EDA Example

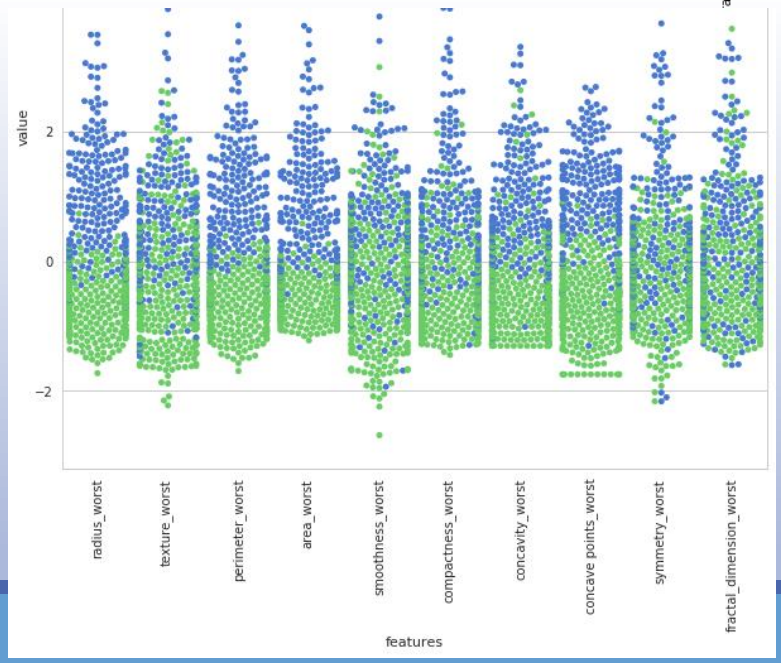
https://www.kaggle.com/kanncaa1/feature-selection-and-data-visualization

	id	diagnosis	radius_mean	texture_mean	perimeter_mean	area_mean	smoothness_mean	compactness_mean	con
0	842302	M	17.99	10.38	122.80	1001.0	0.11840	0.27760	0.3(
1	842517	M	20.57	17.77	132.90	1326.0	0.08474	0.07864	0.0
2	84300903	M	19.69	21.25	130.00	1203.0	0.10960	0.15990	0.1
3	84348301	M	11.42	20.38	77.58	386.1	0.14250	0.28390	0.2
4	84358402	M	20.29	14.34	135.10	1297.0	0.10030	0.13280	0.1

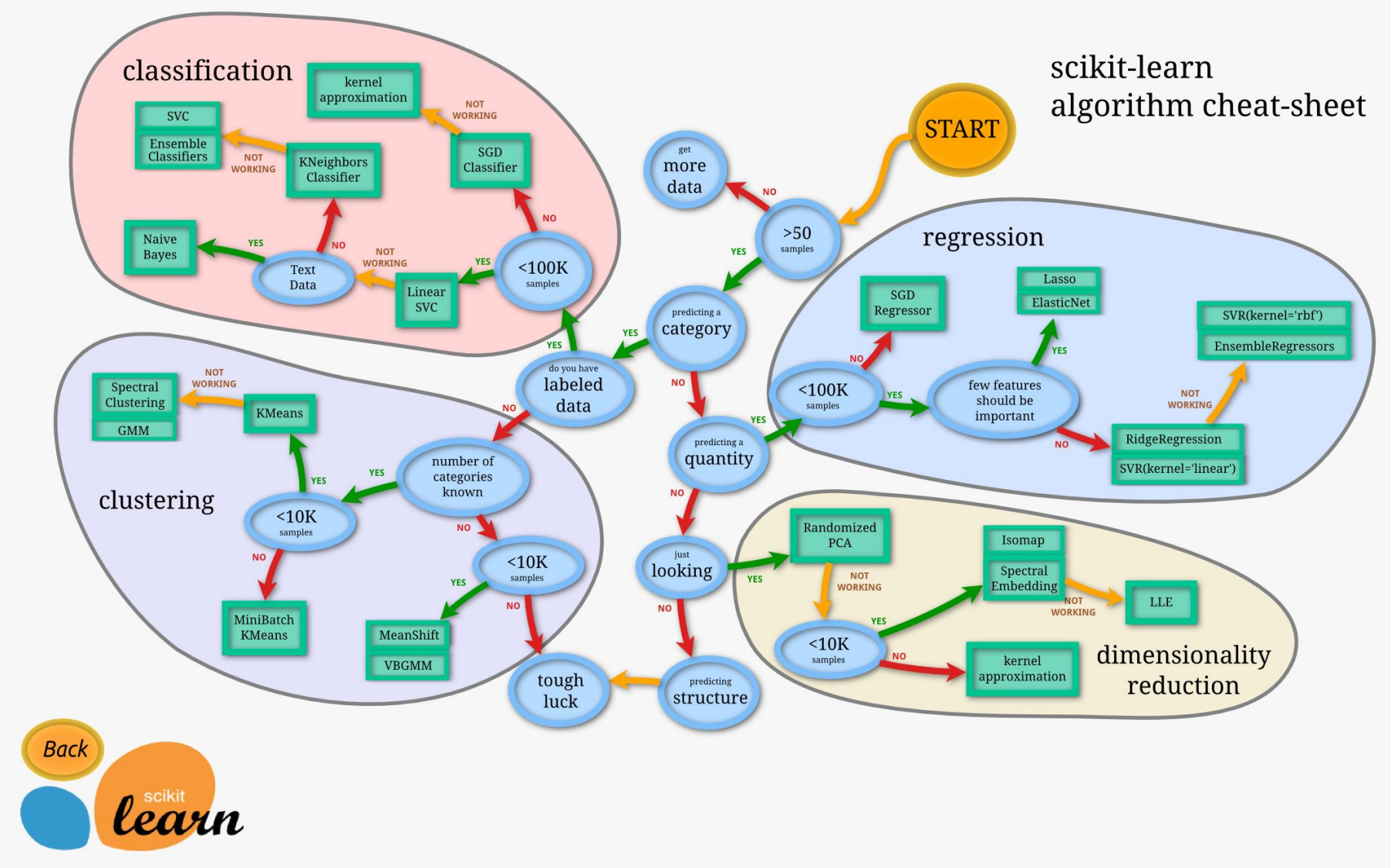
	radius_mean	texture_mean	perimeter_mean	area_mean	smoothness_mean	compactness_mean	concavity_mean	conc poin
count	569.000000	569.000000	569.000000	569.000000	569.000000	569.000000	569.000000	569.
mean	14.127292	19.289649	91.969033	654.889104	0.096360	0.104341	0.088799	0.04
std	3.524049	4.301036	24.298981	351.914129	0.014064	0.052813	0.079720	0.03
min	6.981000	9.710000	43.790000	143.500000	0.052630	0.019380	0.000000	0.00
25%	11.700000	16.170000	75.170000	420.300000	0.086370	0.064920	0.029560	0.02
50%	13.370000	18.840000	86.240000	551.100000	0.095870	0.092630	0.061540	0.03
75%	15.780000	21.800000	104.100000	782.700000	0.105300	0.130400	0.130700	0.07
max	28.110000	39.280000	188.500000	2501.000000	0.163400	0.345400	0.426800	0.20



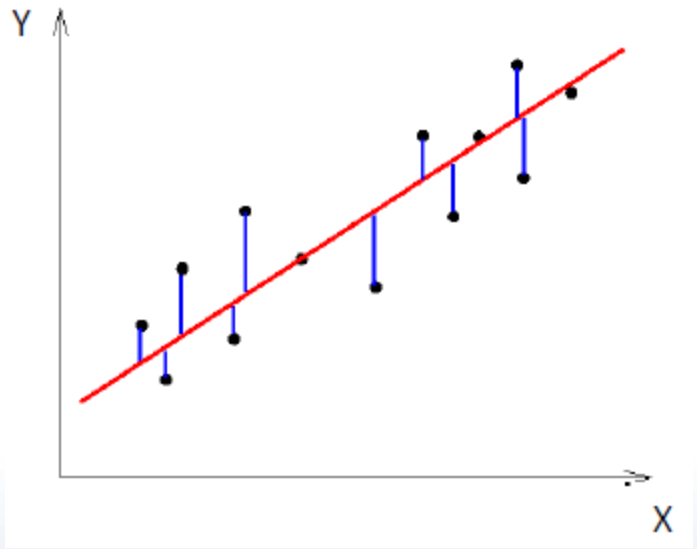




#5 Modelling

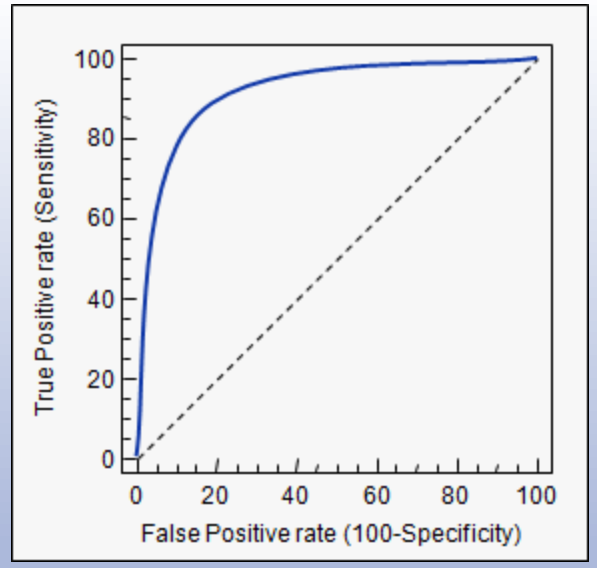

scikit-learn
algorithm cheat-sheet
START
classification
kernel approximation
NOT WORKING
SVC
Ensemble Classifiers
NOT WORKING
KNeighbors Classifier
SGD Classifier
NO
Naive Bayes
YES
Text Data
NO
NOT WORKING
Linear SVC
YES
<100K samples
NO
get more data
NO
>50 samples
YES
predicting a category
YES
do you have labeled data
YES
NO
regression
SGD Regressor
Lasso
ElasticNet
SVR(kernel='rbf')
EnsembleRegressors
NO
<100K samples
YES
few features should be important
YES
NOT WORKING
NO
RidgeRegression
SVR(kernel='linear')
clustering
Spectral Clustering
NOT WORKING
GMM
KMeans
YES
number of categories known
YES
NO
<10K samples
NO
MiniBatch KMeans
NO
<10K samples
YES
MeanShift
VBGMM
NO
tough luck
predicting quantity
YES
NO
just looking
YES
NO
predicting structure
Randomized PCA
YES
NOT WORKING
Isomap
Spectral Embedding
NOT WORKING
LLE
YES
<10K samples
NO
kernel approximation
dimensionality reduction
Back
scikit learn

#6 Evaluation



Predicted Value (predicted by the test)		Actual Value (as confirmed by experiment): positives	Actual Value (as confirmed by experiment): negatives
	positives	**TP** True Positive	**FP** False Positive
	negatives	**FN** False Negative	**TN** True Negative

$$\text{MSE} = \frac{1}{n}\sum_{i=1}^{n}(\hat{Y}_i - Y_i)^2$$



Accuracy Vs Recall Vs Sensitivity

True negative	TN = 400	FP = 7	
True positive	FN = 17	TP = 26	
	Predicted negative	Predicted positive	$N = 450$

$$\text{Accuracy} = \frac{TN+TP}{TN+TP+FN+FP}$$

$$= \frac{400+26}{400+26+17+7}$$

$$= 0.95$$

$$\text{Recall} = \frac{TP}{TP+FN}$$

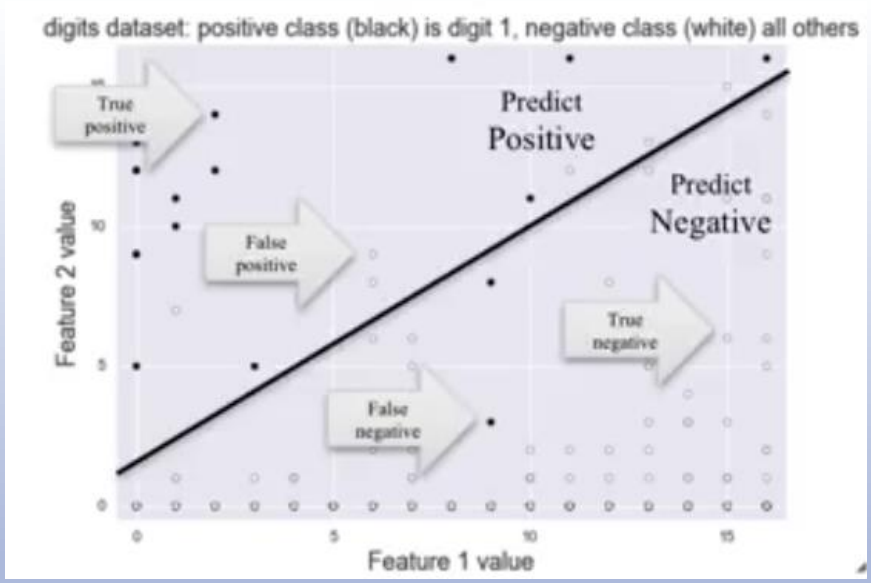
$$= \frac{26}{26+17}$$

$$= 0.60$$

$$\text{Precision} = \frac{TP}{TP+FP}$$

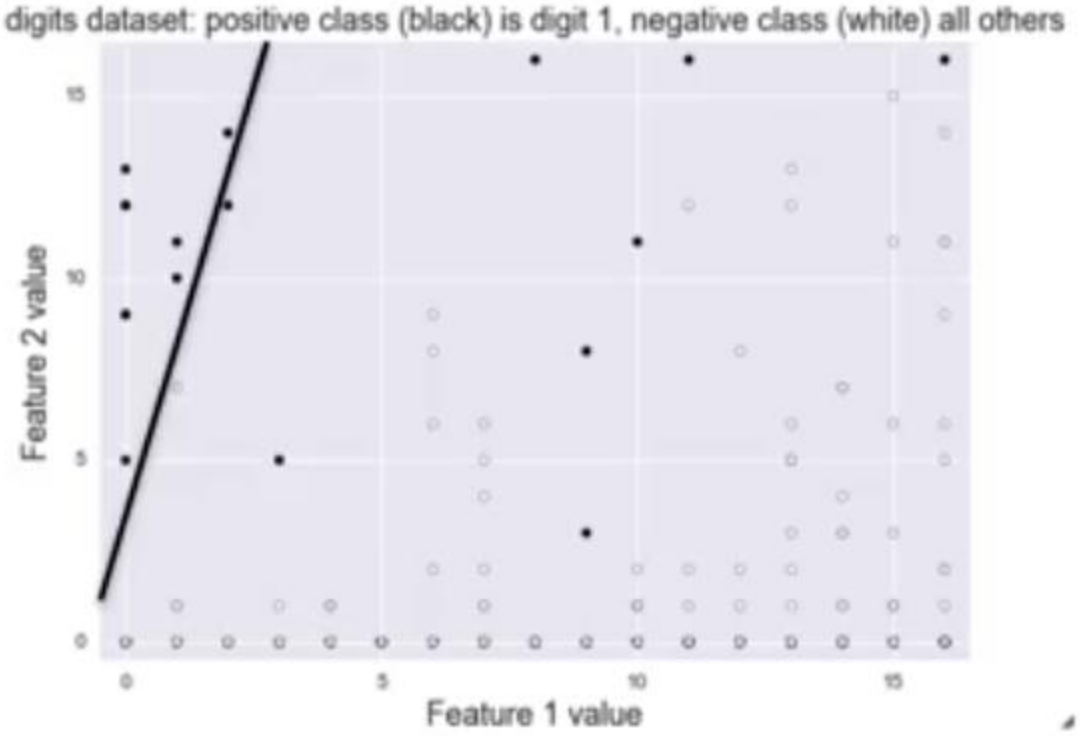
$$= \frac{26}{26+7}$$

$$= 0.79$$



Recall is also known as:
- True Positive Rate (TPR)
- Sensitivity
- Probability of detection

High Precision, Lower Recall



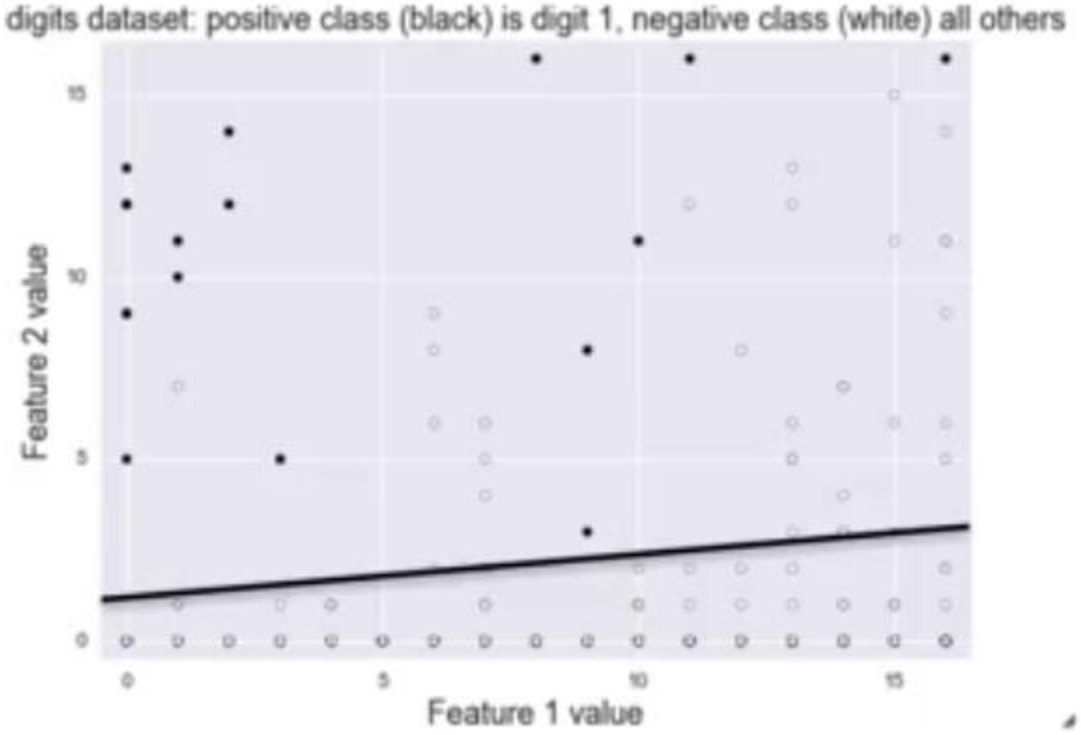
TN = 435	FP = 0
FN = 8	TP = 7

$$\text{Precision} = \frac{TP}{TP+FP} = \frac{7}{7} = 1.00$$

$$\text{Recall} = \frac{TP}{TP+FN} = \frac{7}{15} = 0.47$$

Example: Search Engine classification

Low Precision, High Recall



TN = 408	FP = 27
FN = 0	TP = 15

$$\text{Precision} = \frac{TP}{TP+FP} = \frac{15}{42} = 0.36$$

$$\text{Recall} = \frac{TP}{TP+FN} = \frac{15}{15} = 1.00$$

Example: Tumor detection

Modelling Example

	texture_mean	area_mean	smoothness_mean	concavity_mean	symmetry_mean	fractal_dimension_mean	texture_se	area_se	smoothness_se	cor
count	171.000000	171.000000	171.000000	171.000000	171.000000	171.000000	171.000000	171.000000	171.000000	1
mean	19.593333	643.526901	0.097288	0.090495	0.183149	0.063109	1.247638	39.731351	0.007011	
std	4.494926	335.928618	0.014734	0.084152	0.028621	0.006612	0.620910	36.515284	0.002734	
min	10.380000	143.500000	0.052630	0.000000	0.106000	0.050240	0.362800	8.205000	0.001713	
25%	16.255000	407.250000	0.087600	0.029530	0.163450	0.058215	0.828200	17.005000	0.005251	
50%	19.110000	552.400000	0.097800	0.059880	0.179800	0.061840	1.111000	23.290000	0.006248	
75%	22.425000	757.750000	0.106450	0.135950	0.197300	0.066660	1.486500	47.060000	0.008116	
max	31.120000	1878.000000	0.137100	0.426400	0.290600	0.082430	4.885000	199.700000	0.016040	

```
data.columns
```

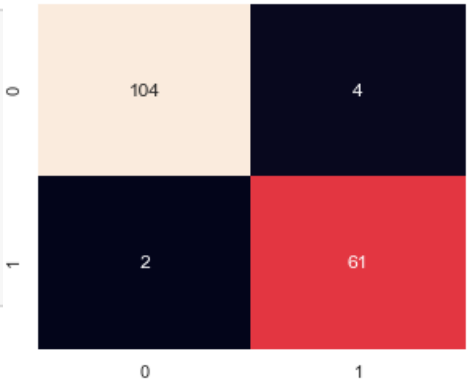
```
Index(['id', 'diagnosis', 'radius_mean', 'texture_mean', 'perimeter_mean',
       'area_mean', 'smoothness_mean', 'compactness_mean', 'concavity_mean',
       'concave points_mean', 'symmetry_mean', 'fractal_dimension_mean',
       'radius_se', 'texture_se', 'perimeter_se', 'area_se', 'smoothness_se',
       'compactness_se', 'concavity_se', 'concave points_se', 'symmetry_se',
       'fractal_dimension_se', 'radius_worst', 'texture_worst',
       'perimeter_worst', 'area_worst', 'smoothness_worst',
       'compactness_worst', 'concavity_worst', 'concave points_worst',
       'symmetry_worst', 'fractal_dimension_worst', 'Unnamed: 32'],
      dtype='object')
```

```
x_test.columns
```

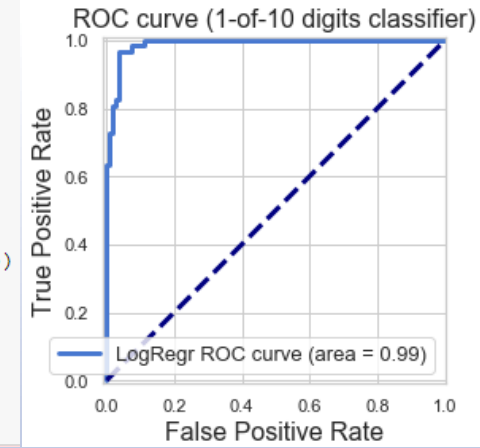
```
Index(['texture_mean', 'area_mean', 'smoothness_mean', 'concavity_mean',
       'symmetry_mean', 'fractal_dimension_mean', 'texture_se', 'area_se',
       'smoothness_se', 'concavity_se', 'symmetry_se', 'fractal_dimension_se',
       'smoothness_worst', 'concavity_worst', 'symmetry_worst',
       'fractal_dimension_worst'],
      dtype='object')
```

```
from sklearn.linear_model import LogisticRegression

lr = LogisticRegression().fit(x_train, y_train)

ac = accuracy_score(y_test,lr.predict(x_test))
print('Accuracy is: ',ac)
cm = confusion_matrix(y_test,lr.predict(x_test))
sns.heatmap(cm,annot=True,fmt="d")
```

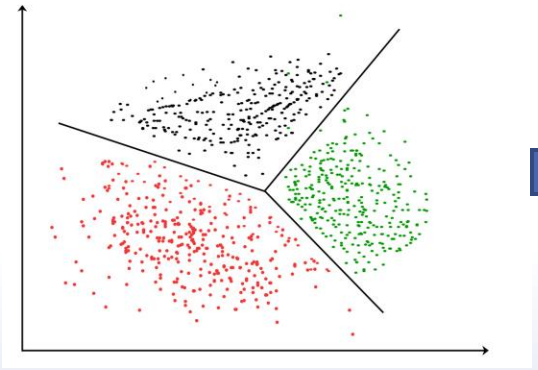
```
Accuracy is:  0.9649122807017544
```

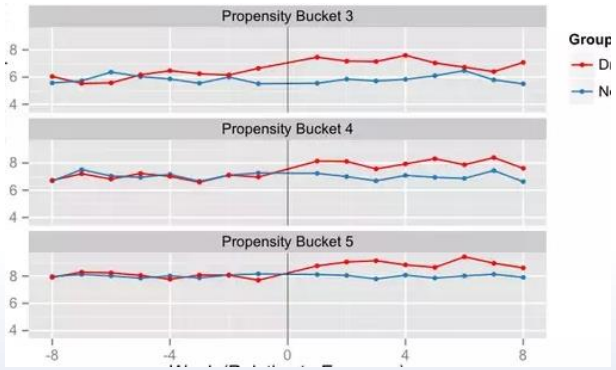



```
from sklearn.metrics import roc_curve, auc

y_score = lr.decision_function(x_test)

y_test_bin = [ 1 if x=='M' else 0 for x in y_test]

fpr, tpr, _ = roc_curve(y_test_bin, y_score)
roc_auc = auc(fpr, tpr)

plt.figure()
plt.xlim([-0.01, 1.00])
plt.ylim([-0.01, 1.01])
plt.plot(fpr, tpr, lw=3, label='LogRegr ROC curve (area = {:0.2f})'.format(roc_auc))
plt.xlabel('False Positive Rate', fontsize=16)
plt.ylabel('True Positive Rate', fontsize=16)
plt.title('ROC curve (1-of-10 digits classifier)', fontsize=16)
plt.legend(loc='lower right', fontsize=13)
plt.plot([0, 1], [0, 1], color='navy', lw=3, linestyle='--')
plt.axes().set_aspect('equal')
plt.show()
```

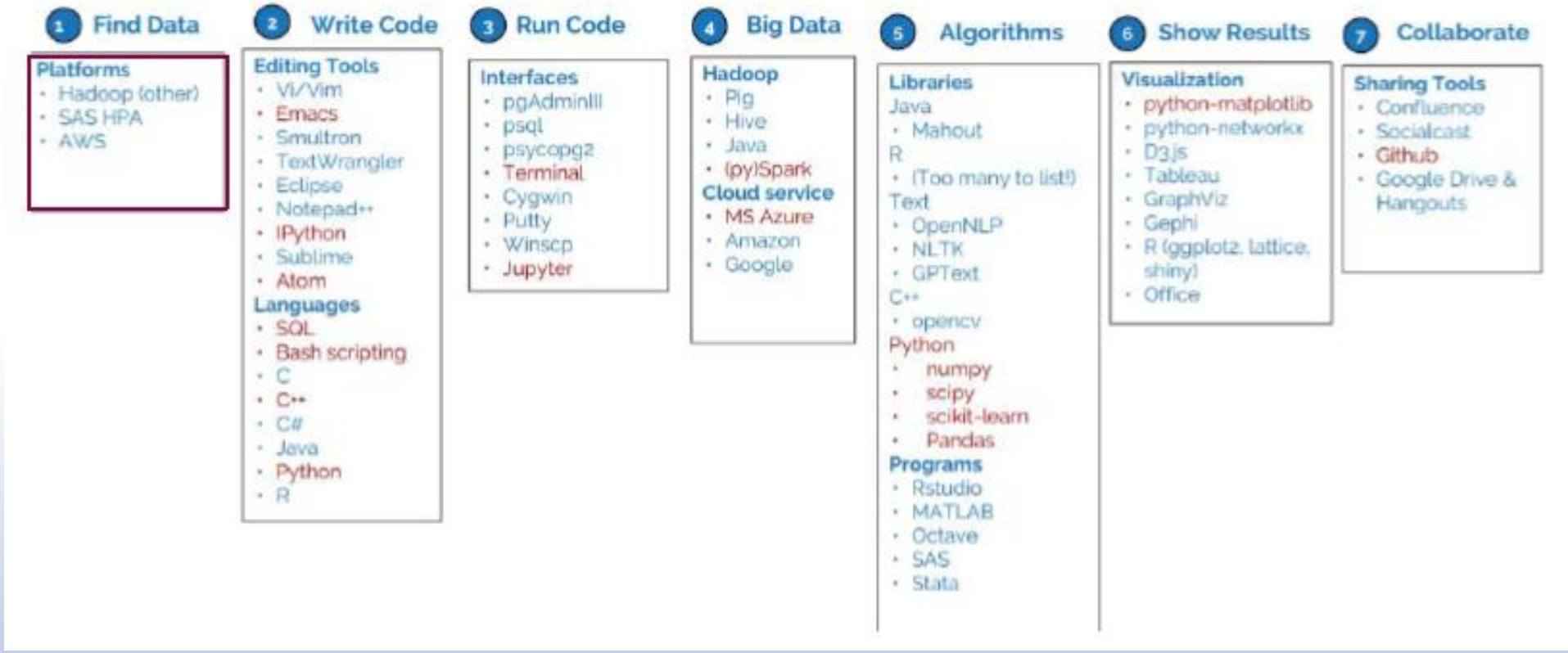


Prescriptive Analytics

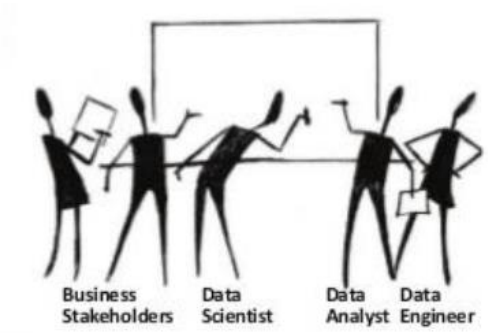


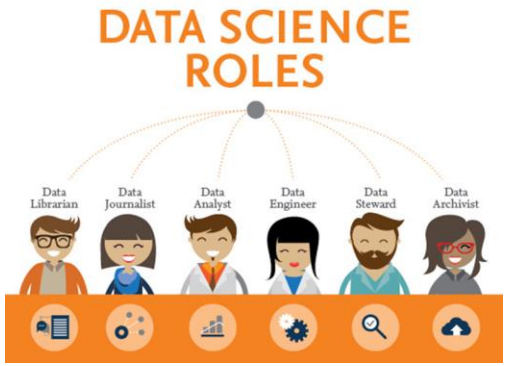


Data Science Tools



Data Science Team





Work	
Data scientists work on... • Data Analysis • Statistics • Machine Learning • Data Mining • Statistical modeling • Research • Algorithms • Analytics • Programming	**Data engineers** work on... • Data Warehousing • ETL • Databases • Business Intelligence
Tools	
Data scientists use... • Matlab • SaS	**Data engineers** use... • Oracle • Hadoop • Microsoft SQL Server • MySQL • Hive
Languages	
Data scientists code in: • R • Python • LaTeX • C++	**Data engineers** code in... • Java • Unix • Javascript • Linux • SQL

Data Science Required Skillsets