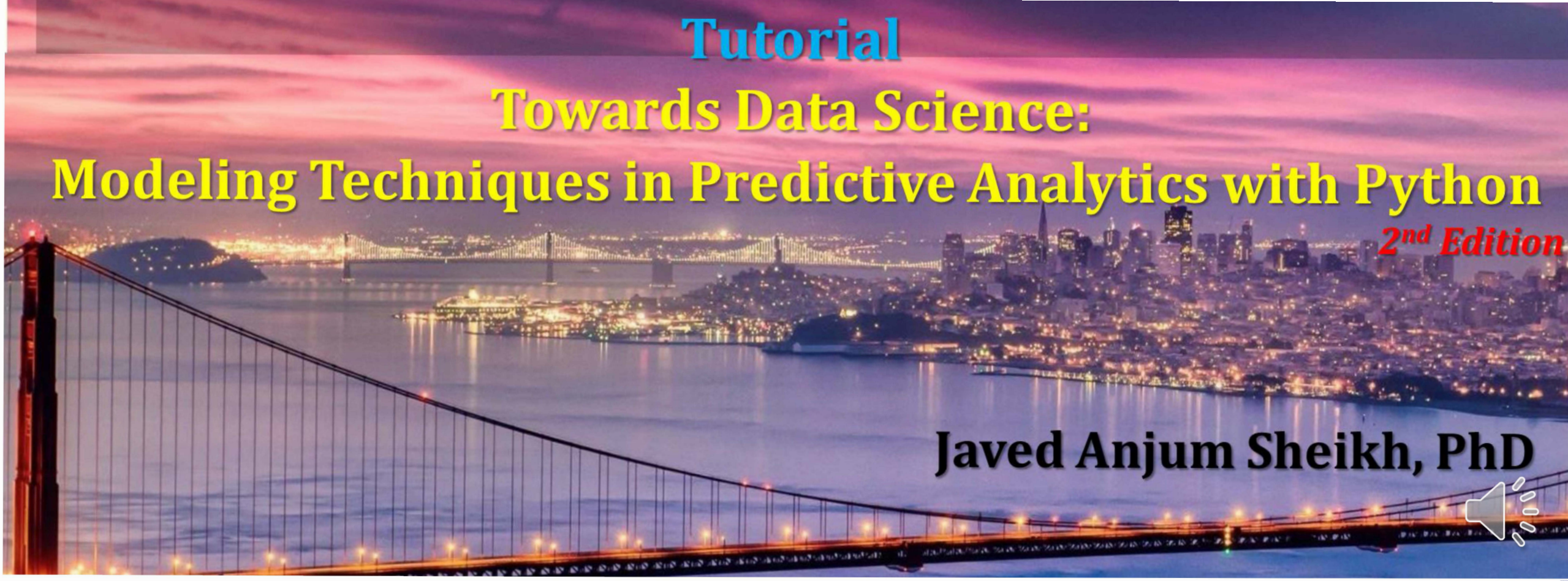
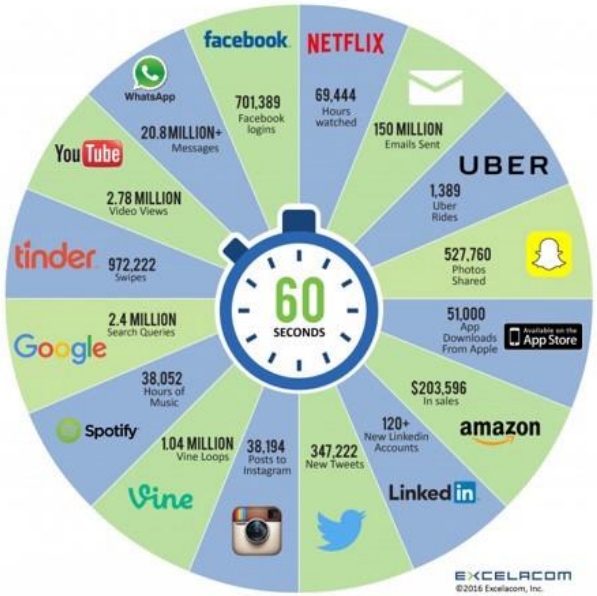

Tutorial
Towards Data Science:
Modeling Techniques in Predictive Analytics with Python
2nd Edition
Javed Anjum Sheikh, PhD

Agenda

- Big Data
- Data Science
- Analytics
- Predictive Analytics
- Data Science/Python
- Python Code

What Happens in an **Internet Minute?**

2016 What happens in an INTERNET MINUTE?



2017 This Is What Happens In An Internet Minute

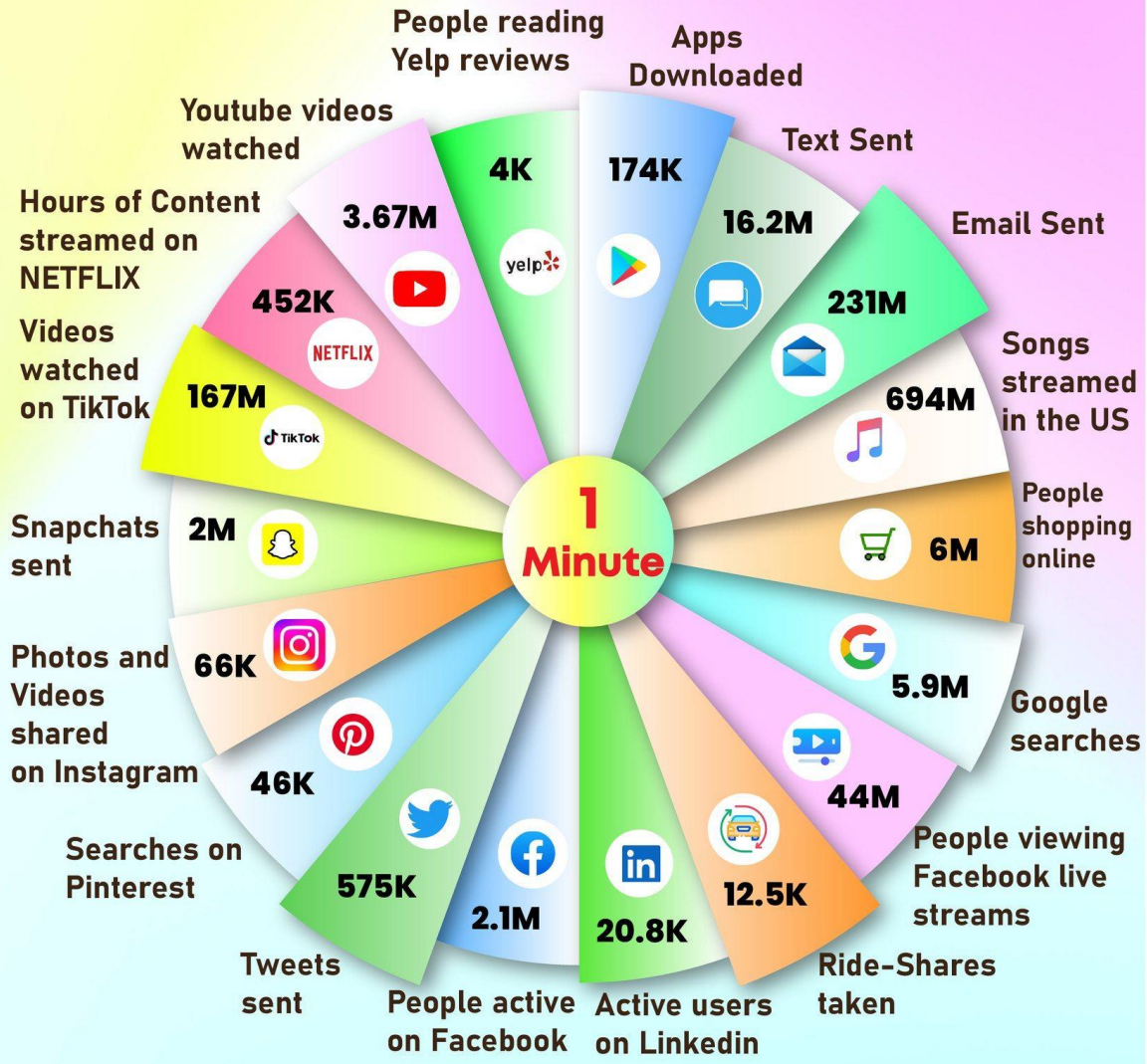
2018 This Is What Happens In An Internet Minute

2019 This Is What Happens In An Internet Minute

2020 This Is What Happens In An Internet Minute

2021 This Is What Happens In An Internet Minute

A Minute on the Internet in 2022

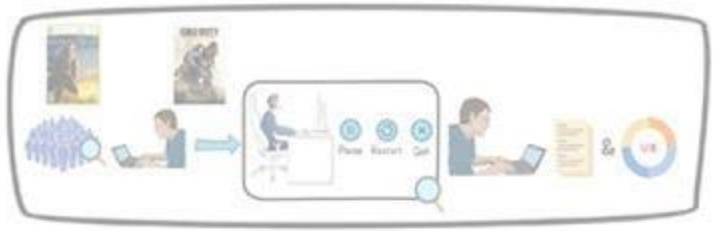


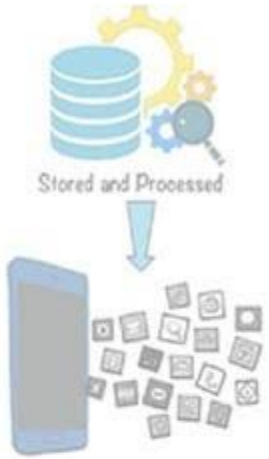
Source: LocaliQ

RankingRoyals

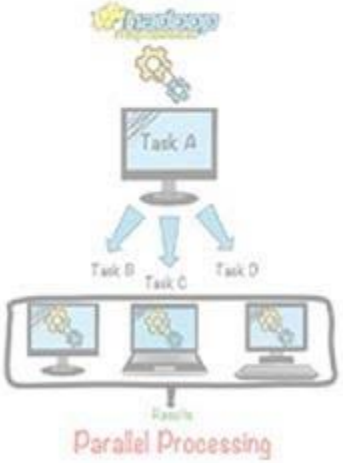
THE INTERNET IN 2023 EVERY MINUTE

Created by: eDiscovery Today & LTMG





What is Big Data?



Hurricane Sandy in 2012

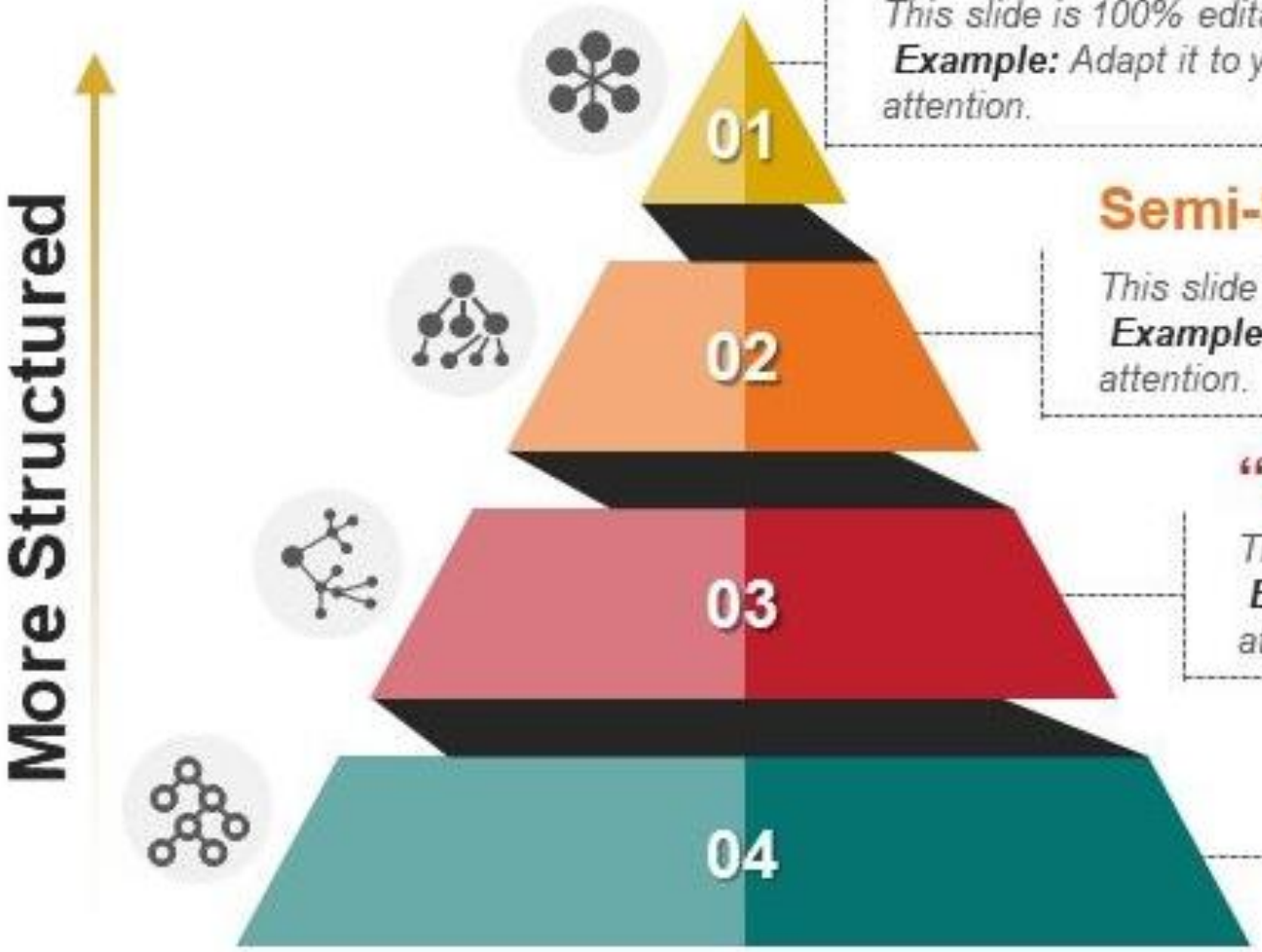
simplilearn

BIG DATA Key Activities

Store

Process

Access



Structured

This slide is 100% editable.
***Example:** Adapt it to your needs and capture your audience's attention.*

Semi-Structured

This slide is 100% editable.
***Example:** Adapt it to your needs and capture your audience's attention.*

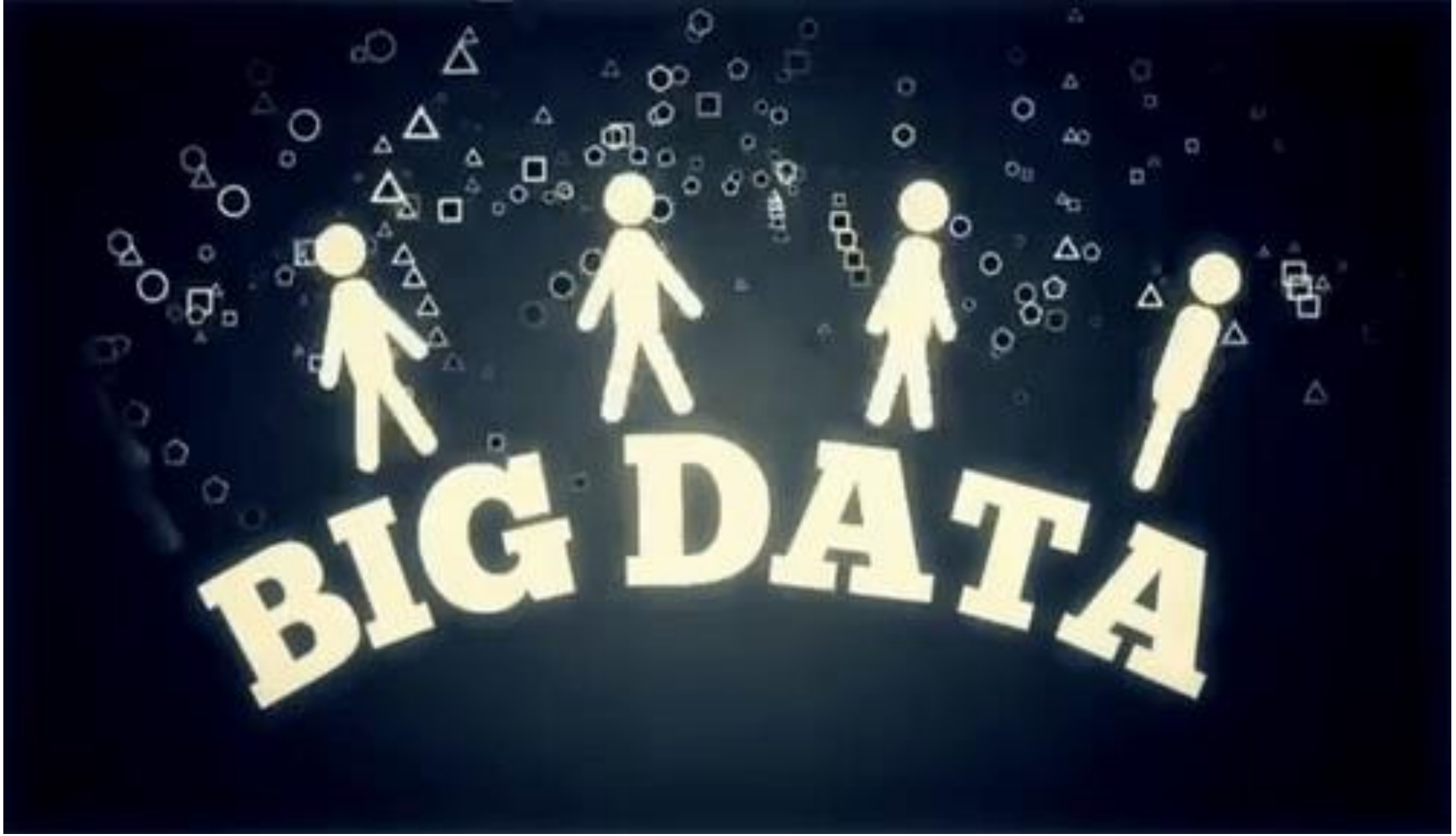
"Quasi" Structured

This slide is 100% editable.
***Example:** Adapt it to your needs and capture your audience's attention.*

Unstructured

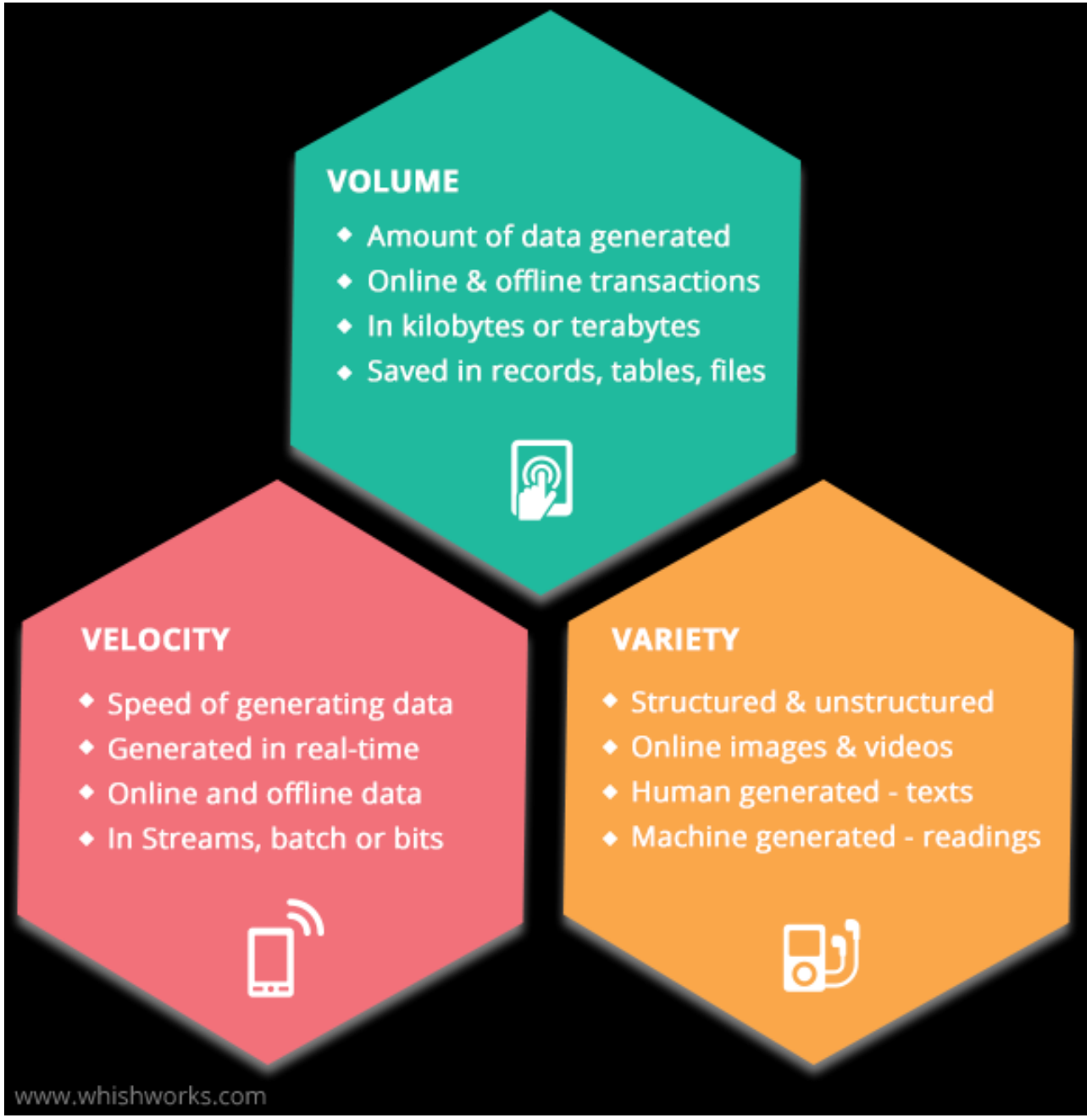
This slide is 100% editable.
***Example:** Adapt it to your needs and capture your audience's attention.*

Types of Big Data



How Many *"V's"* in Big Data?

https://morioh.com/p/5f328640dec8


VOLUME
Amount of data generated
Online & offline transactions
In kilobytes or terabytes
Saved in records, tables, files
VELOCITY
Speed of generating data
Generated in real-time
Online and offline data
In Streams, batch or bits
VARIETY
Structured & unstructured
Online images & videos
Human generated - texts
Machine generated - readings
www.whishworks.com

Volume

Data at Scale

Terabytes to petabytes of data

Variety

Data in Many Forms

Structured, unstructured, text, multimedia

Velocity

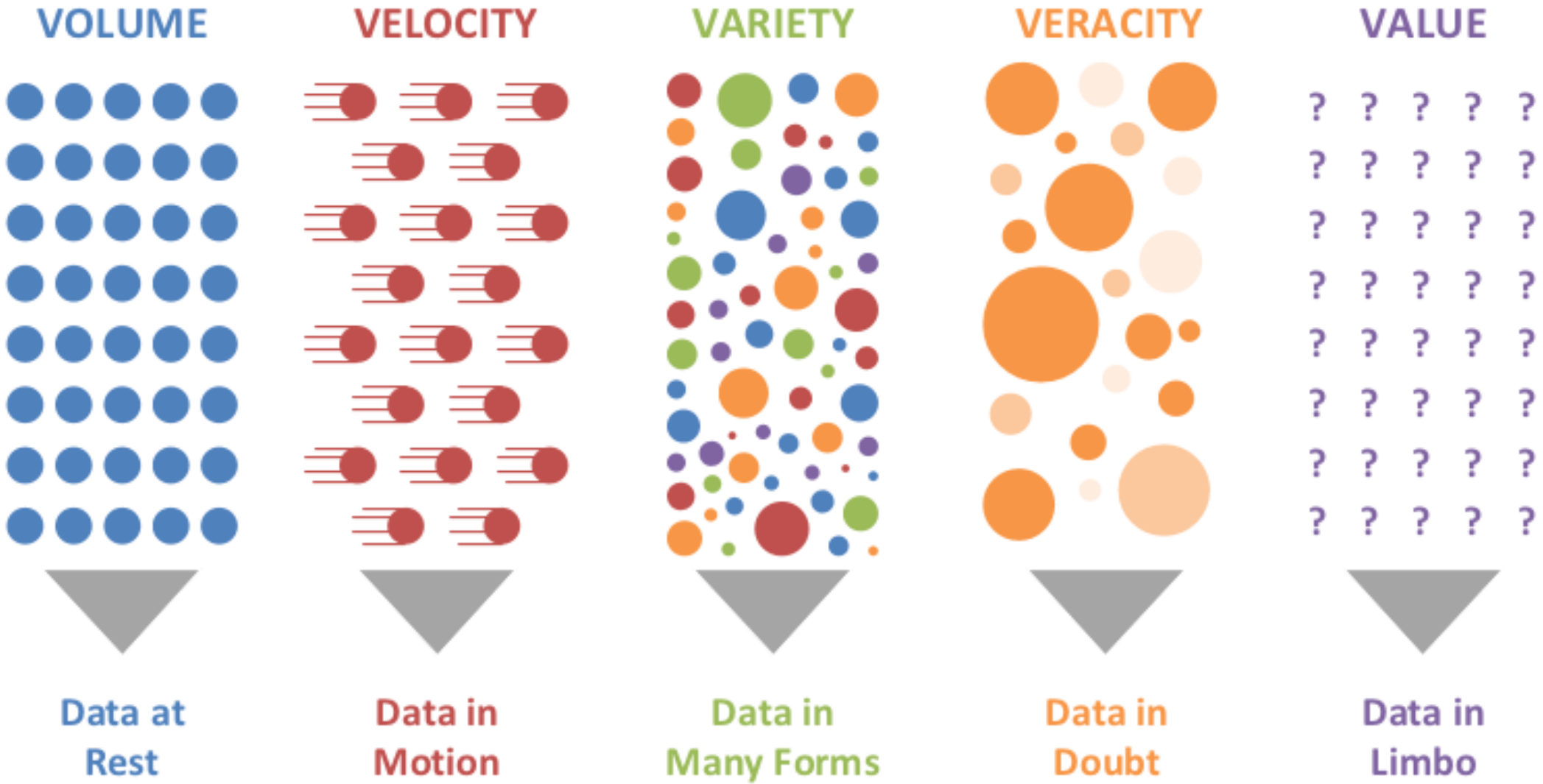
Data in Motion

Analysis of streaming data to enable decisions within fractions of a second.

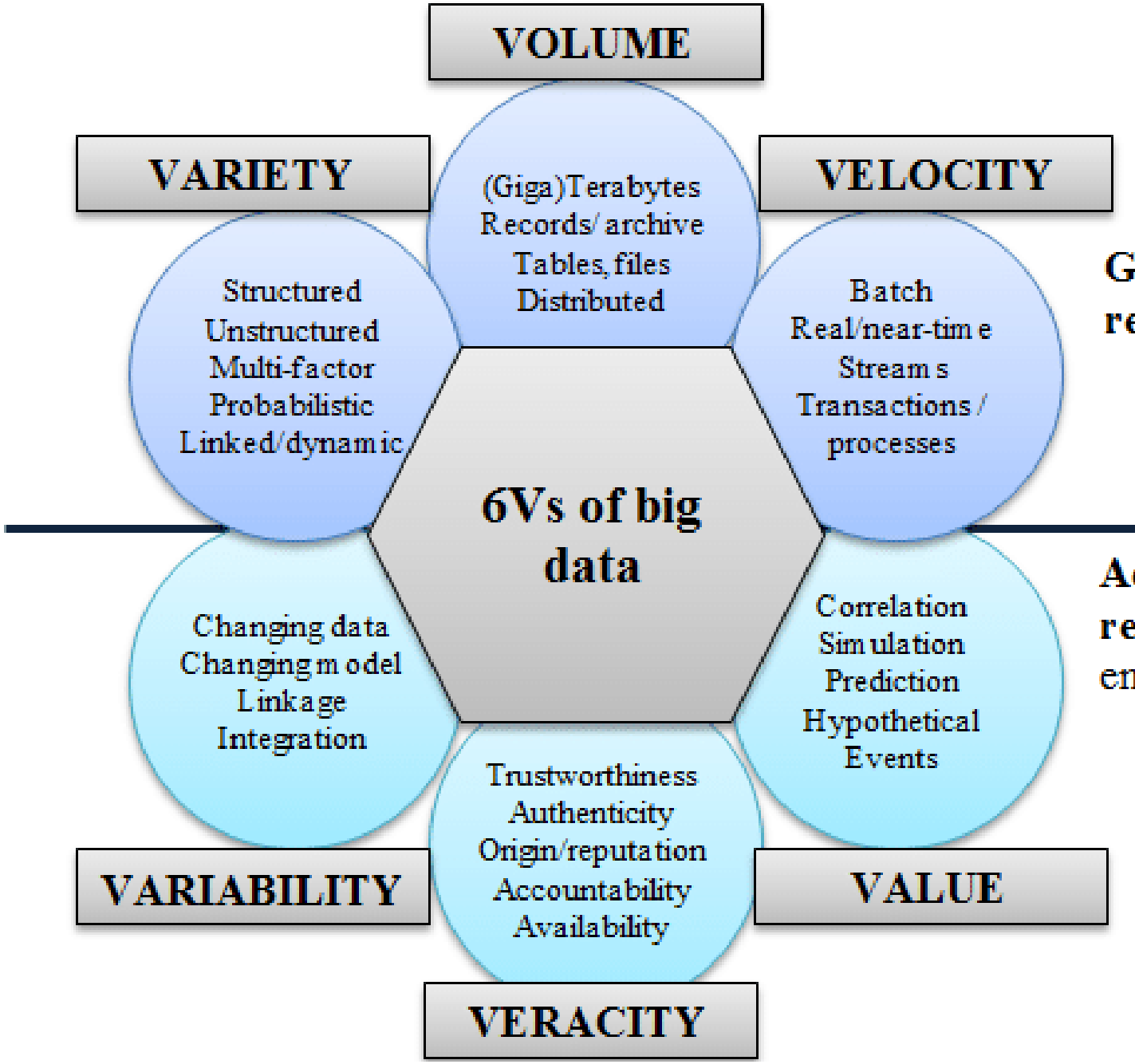
Veracity

Data Uncertainty

Managing the reliability and predictability of inherently imprecise data types.



evariant
MOVING HEALTHCARE AHEAD™



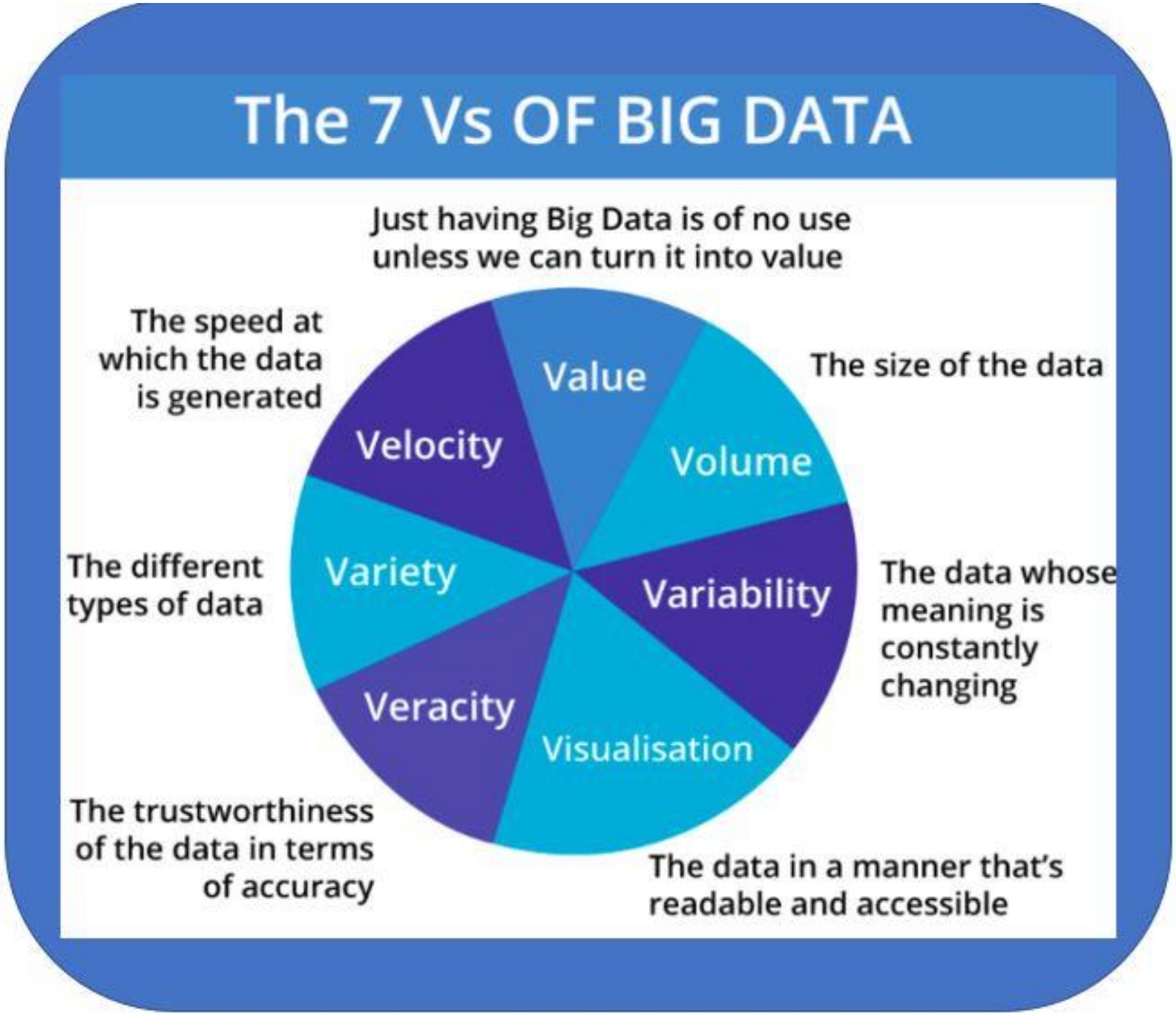
Generic big data requirements:

- variety,
- volume,
- velocity.

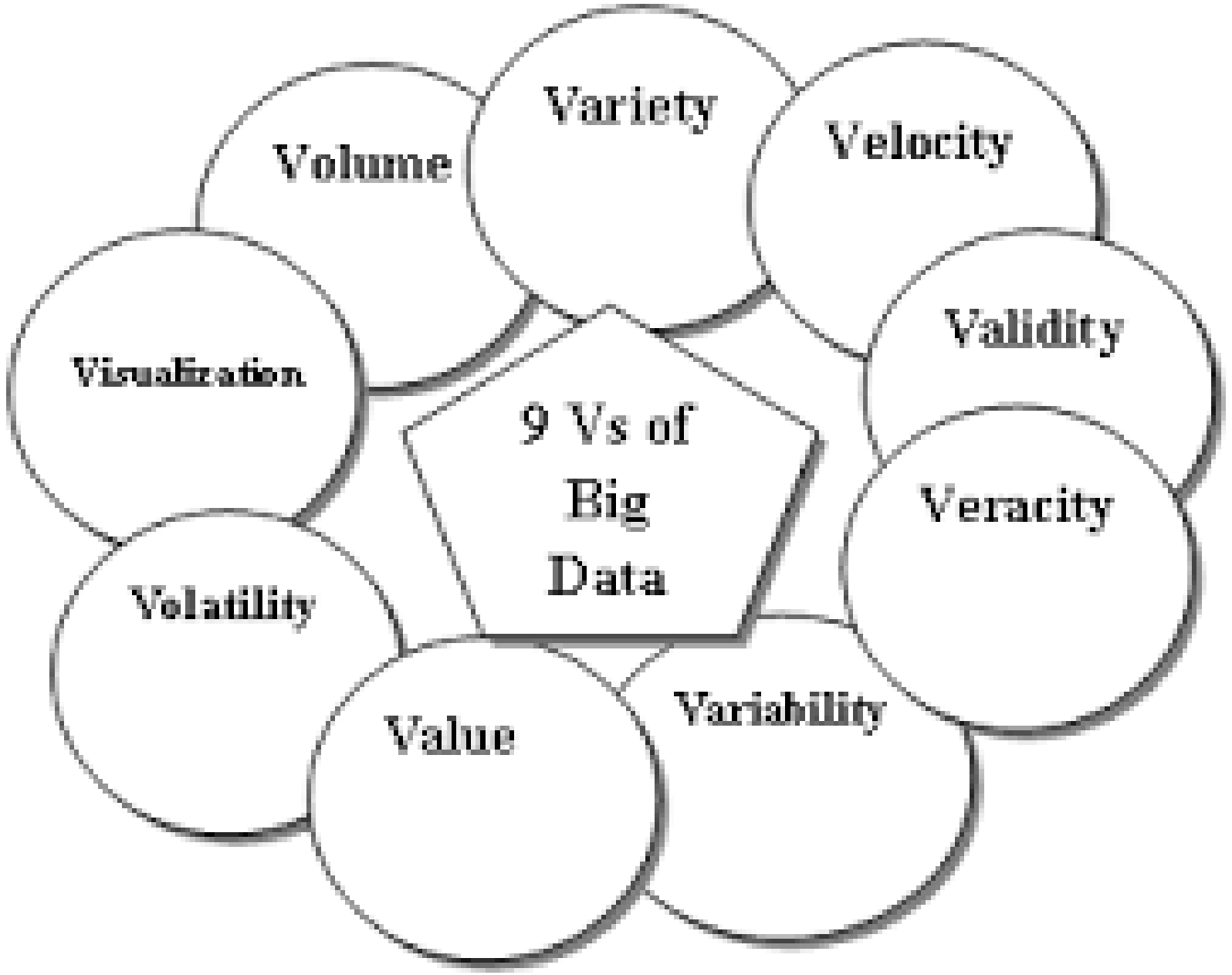
Acquired big data requirements (after entering system):

- variability,
- veracity,
- value.

The 7 Vs OF BIG DATA




01
VOLUME
Can you find the information you are looking for?
02
VALUE
Can you find it when you most need it?
03
VERACITY
Are you dealing with information or disinformation?
04
VISUALISATION
Can you make sense at a glance? Does it trigger a decision?
05
VARIETY
Is a picture worth a thousand words in 70 languages? Is your information balanced?
06
VELOCITY
Information gains momentum and crises & opportunities evolve in real time. How is outlook for today?
07
VISCOSITY
Does it stick with you? Does it call for action?
08
VIRALITY
AHA to-go? Does it convey a message that can be pasted into a presentation or Instgrammed?
BIG DATA
with 8 V's



https://inpressco.com/wp-content/uploads/2017/11/Paper44277-280.pdf


Value
Can you find the information you are looking for ?
Virality
Aha to - go? Does it convey a message that can be pasted into a presentation or instagrammed?
Venue
Distributed Heterogeneous Data from Multiple Platforms
Variability
Dynamic, Evolving Behaviour in Data Source
Big Data
With 10 V's
Volume
Can you find it when you most need it ?
Viscosity
Does it stick with you ? Does it call for action?
Veracity
Are you dealing with information or disinformation ?
Visualisation
Can you make sense at a glance? Does it trigger a decision?
Velocity
information gains momentum and crises & opportunities enolve in real time Hoe is outlook for today ?
Variety
is a picture worth a thousand words
languages? IS your information balanced?

Sources of Big Data

- Social Data
- Machine-Generated Data
- Transactional Data

WHAT IS DATA SCIENCE?

simplilearn

DATA PREPARATION

DATA CLEANING

INCONSISTENT DATATYPES

MISSPELLED ATTRIBUTES

MISSING AND DUPLICATE VALUES

TRANSFORMATION

talend

Informatica

EXPLORATORY DATA ANALYSIS

DEFINES AND REFINES

THE SELECTION OF FEATURE

VARIABLES THAT WILL BE USED

IN THE MODEL DEVELOPMENT

DATA MODELING

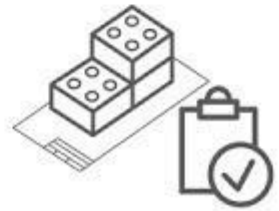
KNN

NAIVE BAYES

DECISION TREE

VISUALIZATION AND COMMUNICATION

DEPLOYS AN



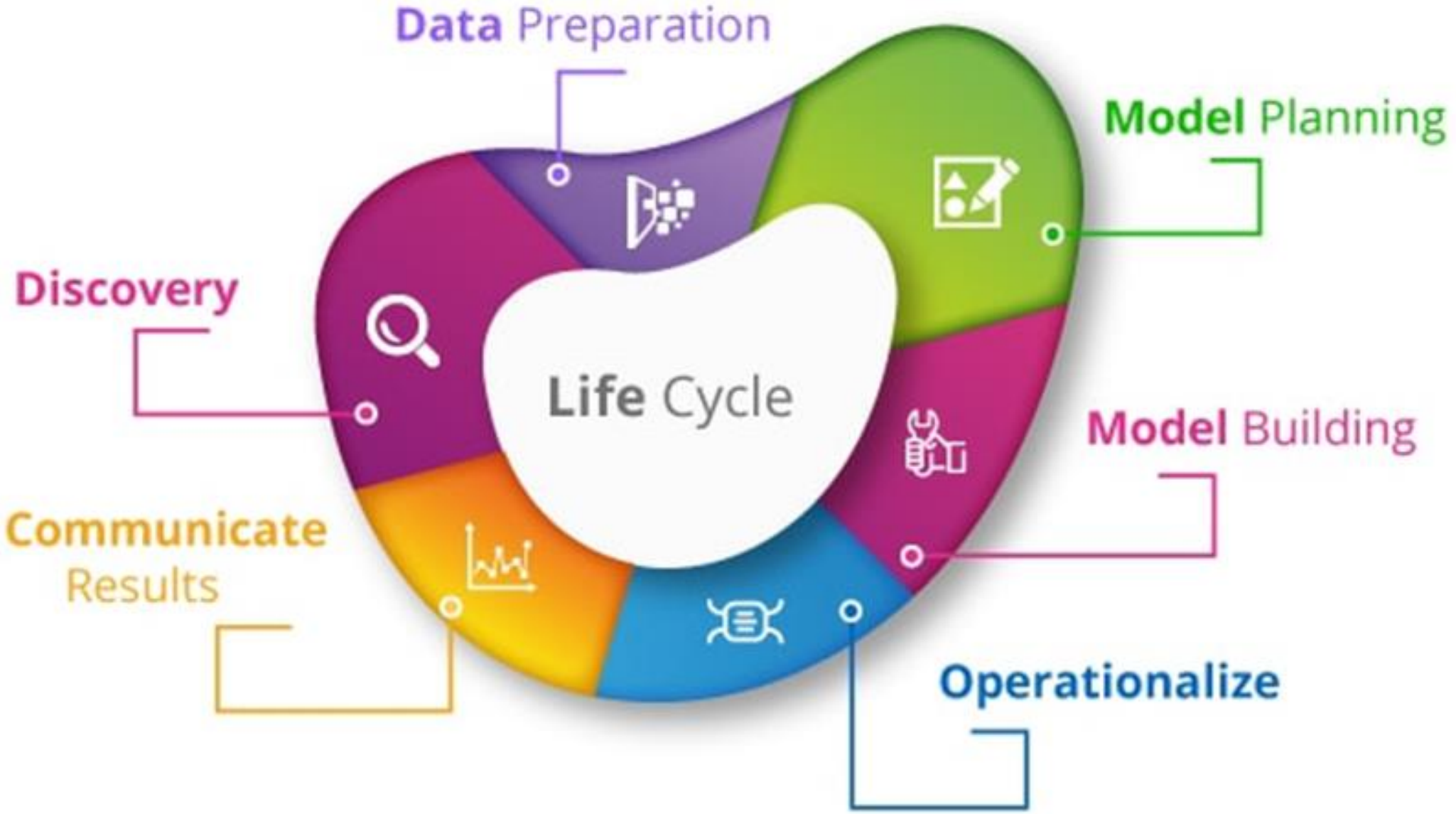
DATA ACQUISITION

- WEB SERVERS
- LOGS
- DATABASES
- APIS
- ONLINE REPOSITORIES

WHY?....WHY?....WHY?....

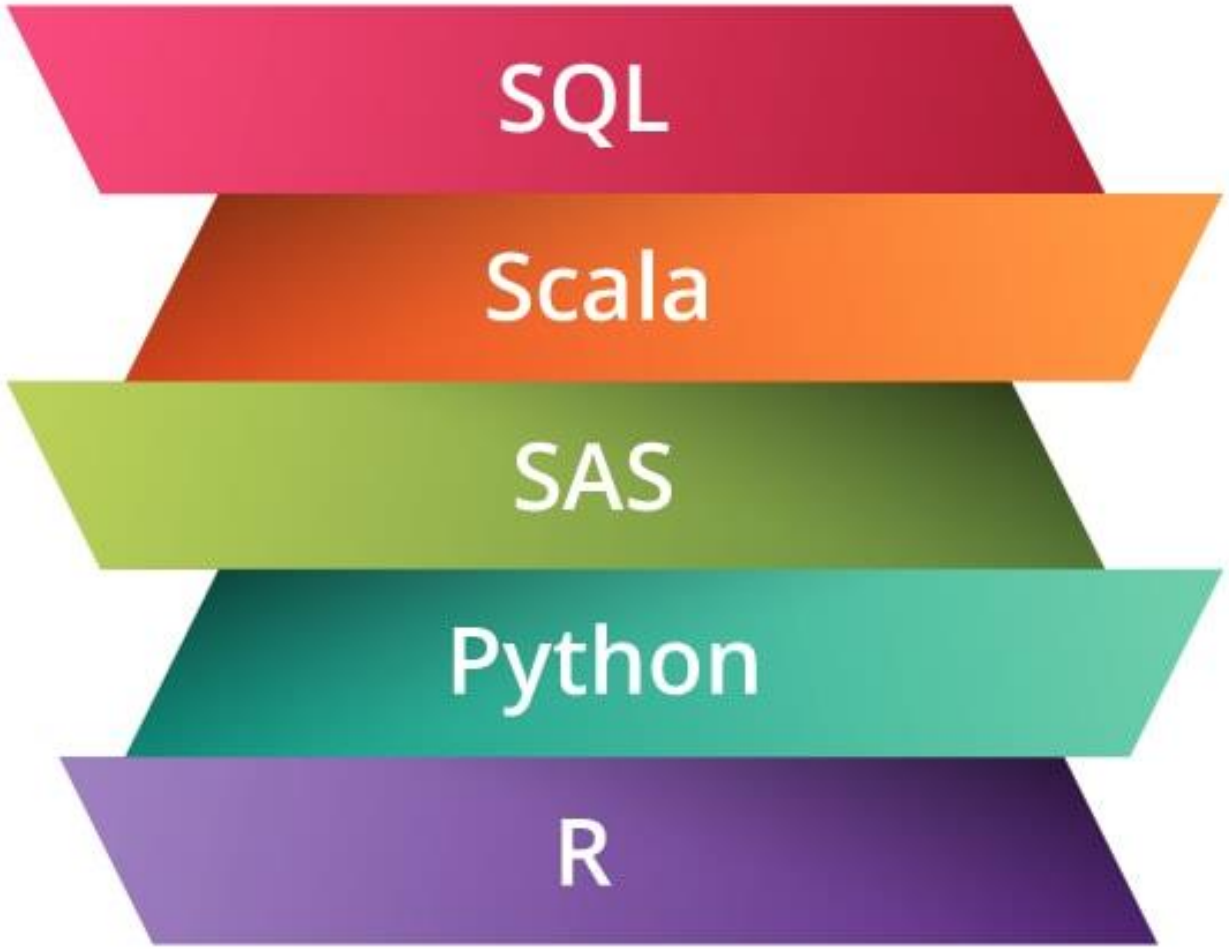
https://morioh.com/p/5f328640dec8

The lifecycle of Data Science



Advantage of Data Science

Top 5 programming languages in Data Science



Basis	Big Data	Data Science
Meaning	revolves around the huge volumes of data which cannot be handled using the conventional data analysis method	skewed towards the scientific approach of interpreting the data and retrieves the information from a given data set
Concept	scientific techniques to process data, extract information and interpret results which help in the decision-making process	obtained with big data is heterogeneous that indicates a diversified data set which has to be pre-cleaned and sorted before running analytics on them
Formation	data filtering, preparation, and analysis	Internet users/ traffic, live feeds, and data generated from system logs
Application areas	Telecommunication, financial service, health and sports, research and development, and security and law enforcement	Internet search, digital advertisements, text-to-speech recognition, risk detection, and other activities
Approach	used by businesses to track their presence in the market which helps them develop agility and gain a competitive advantage over others	uses mathematics and statistics extensively along with programming skills to develop a model to test the hypothesis and make decisions in the business

Big Data Vs Data Science

Factors	Big Data	Data Science
Concept	Handling large data	Analyzing data
Responsibility	Process huge volumes of data and generate insights	Understand pattern within data and make decisions
Industry	E-commerce, security services, telecommunication	Sales, image recognition, advertisement, risk analytics
Tools	Hadoop, Spark, Flink	SAS, R, Python

What is Data Analytics?

https://morioh.com/p/5f328640dec8

What Is Big Data Analytics?

WHAT IS DATA SCIENCE?	WHAT IS DATA ANALYTICS?	WHAT IS BIG DATA?
Data Science is a field that refers to the collective processes, theories, concepts, tools and technologies that enable the review, analysis and extraction of valuable knowledge and information from raw data.	**Data Analytics (DA)** is the process of examining data sets in order to draw conclusions about the information they contain, increasingly with the aid of specialized systems & software.	**Big Data** refers to voluminous amounts of structured or unstructured data that organizations can potentially mine & analyze for business gains.
	APPLICATION AREAS	
1. Digital advertisements 2. Internet Research 3. Recommender System 4. Image/Speech Recognition	1. Gaming 2. Travel 3. Energy Management 4. Healthcare	1. Communication 2. Retail 3. Financial services 4. Education
	TOOLS & LANGUAGES	
1. Python 2. SAS 3. SQL	1. R 2. Tableau Public 3. Apache Spark	1. Hadoop 2. NoSQL 3. Hive

Differentiation	Big Data	Data Science	Data Analytics
Definition	Unprocessed data sets of humongous volumes	Science of cleaning, preparing and aligning the data for analysis using statistical and mathematical models	It is related to examining raw data which is required to provide conclusive information
Applications	Financial services	Delivery of better search results on the internet	Gaining efficiency in the Healthcare
	Fraud analytics	Digital advertisements from display banners to finding the appropriate prospects	Optimization of buying experience through mobile and social media data analysis.
	Communication industry to retain and expand the consumer base	The recommender system to help in the user experience	Collection of data in the gaming industry
	Brick-and-mortar and online retailer for better customer service		Energy management
Skill requirements	In order to become a big data professional, the following skills are required:	A data scientist must highlight a profile that has the following skills:	Following skills are necessary to become a data analyst:
	Analytical skills	Education with a Master's degree in either data analysis, statistics, or mathematics	Programming skills
	Creativity	Knowledge of SAS or R or SPSS	Statistical and mathematics skills
	Mathematics and statistical skills	Knowledge in coding on Python and Hadoop	Machine learning skills and certificates
	Basic computation knowledge	Working efficiency with unstructured data	Data visualization skills
	Computer science and business skills		Analytical skills Data wrangling skills

Analysis looks backwards over time.

Analytics look forward to model the future or predict a result

https://datascientistinsights.com/2015/06/19/coopetition-a-comparison-of-data-analytics-data-sciences/

Data Science vs Data Analytics

	Data Science	Data Analytics
SKILLSET	• Data Modelling • Predictive Analytics • Advanced Statistics • Engineering/Programming	• BI Tools • Intermediate Statistics • Solid Programming Skills • Regular Expression (SQL)
SCOPE	Macro	Micro
EXPLORATION	• Search Engine Exploration • Machine Learning • Artificial Intelligence • Big data - Often Unstructured	• Data Visualization Techniques • Designing Principles • Big Data - Mostly Structured
GOALS	Discover New Questions to Drive Innovation	Use Existing Information to Uncover Actionable Data

Use of Big Data in Data Analytics

Solving Common Data Challenges

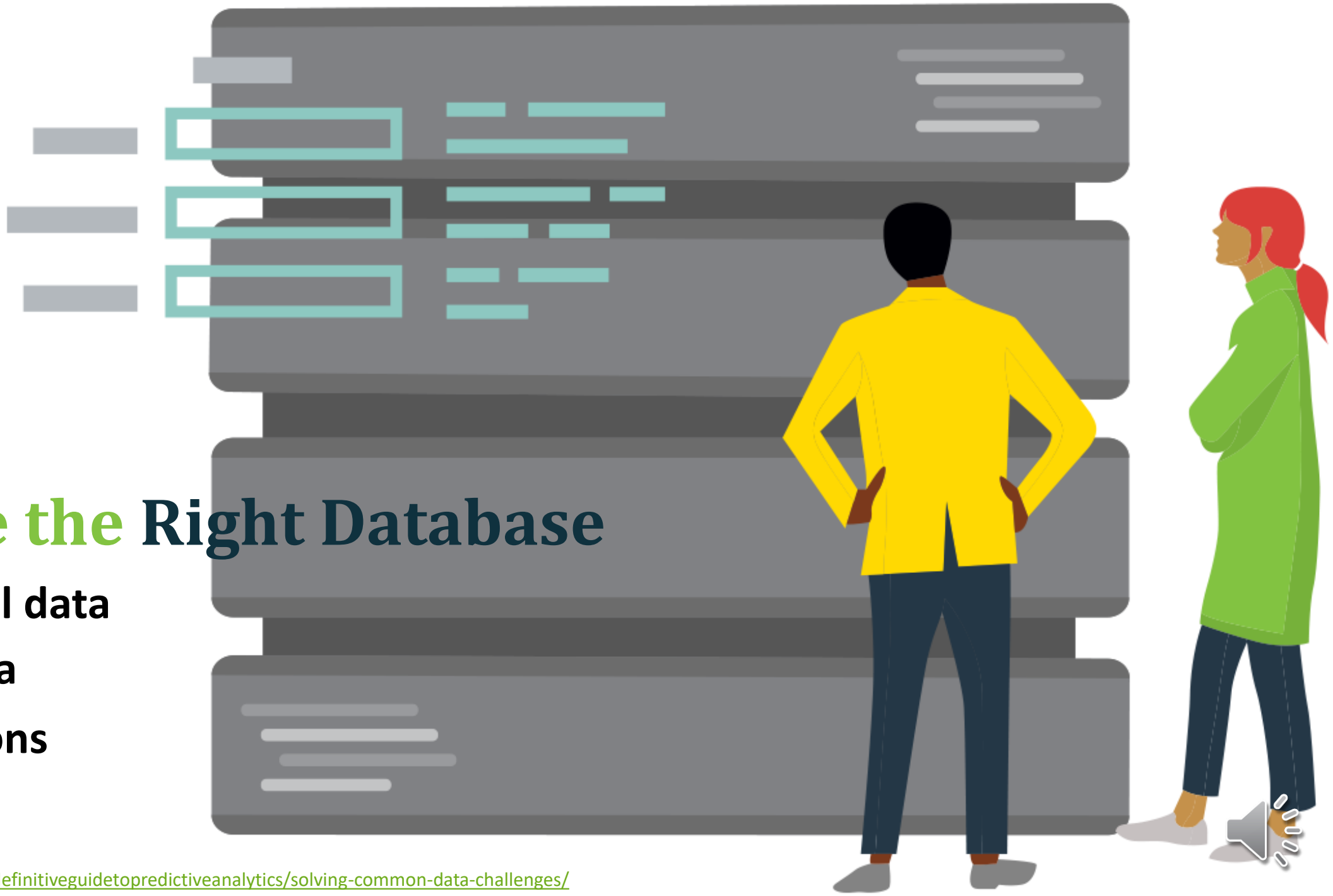
Find the Data You Need
Choose the Right Database
Practice Database Hygiene
Cleanse Your Data
Avoid Bias in Your Data and Models
Validate your model is working and establish a performance baseline
Know When It's Time to Refresh
Boost Predictive Performance Over Time
How to Adapt to Changing Business Requirements

Find the Data You Need

https://www.logianalytics.com/definitiveguidetopredictiveanalytics/solving-common-data-challenges/

Choose the Right Database

- Historical data
- New data
- Predictions



https://www.logianalytics.com/definitiveguidetopredictiveanalytics/solving-common-data-challenges/

Practice Database Hygiene

https://www.logianalytics.com/definitiveguidetopredictiveanalytics/solving-common-data-challenges/

Cleanse Your Data

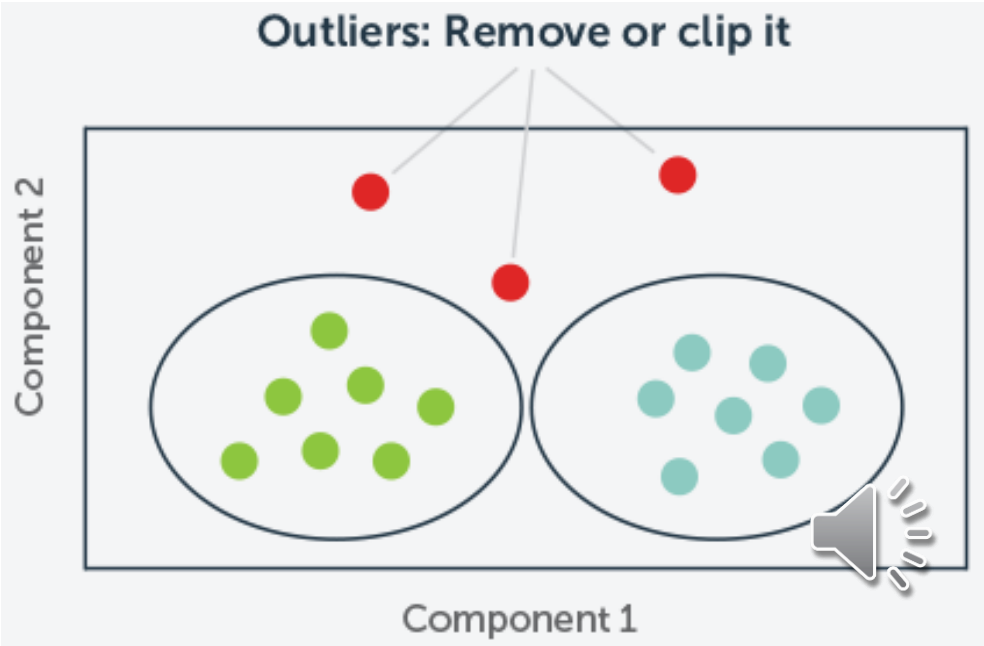
1. Missing values

Row no	State	Salary	Yrs of Experience
1	NY	57400	Mid
2	TX		Entry
3	NJ	90000	High
4	VT	36900	Entry
5	TX		Mid
6	CA	76600	High
7	NY	85000	High
8	CA		Entry
9	CT	45000	Entry

Missing Values

1. Delete the rows
2. Replace with mean value
3. Predict the value using other columns

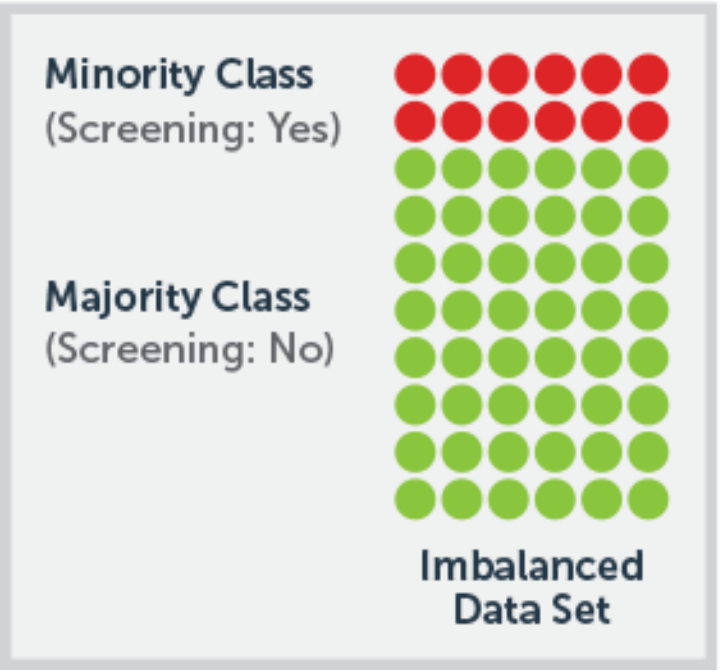
2. Outliers

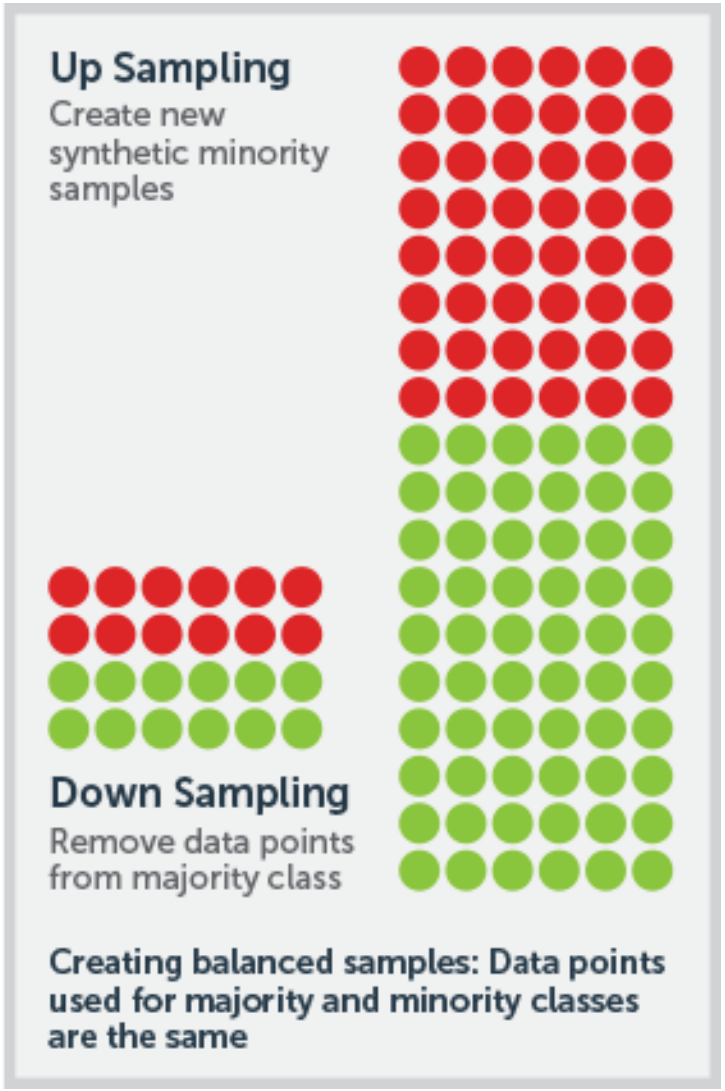


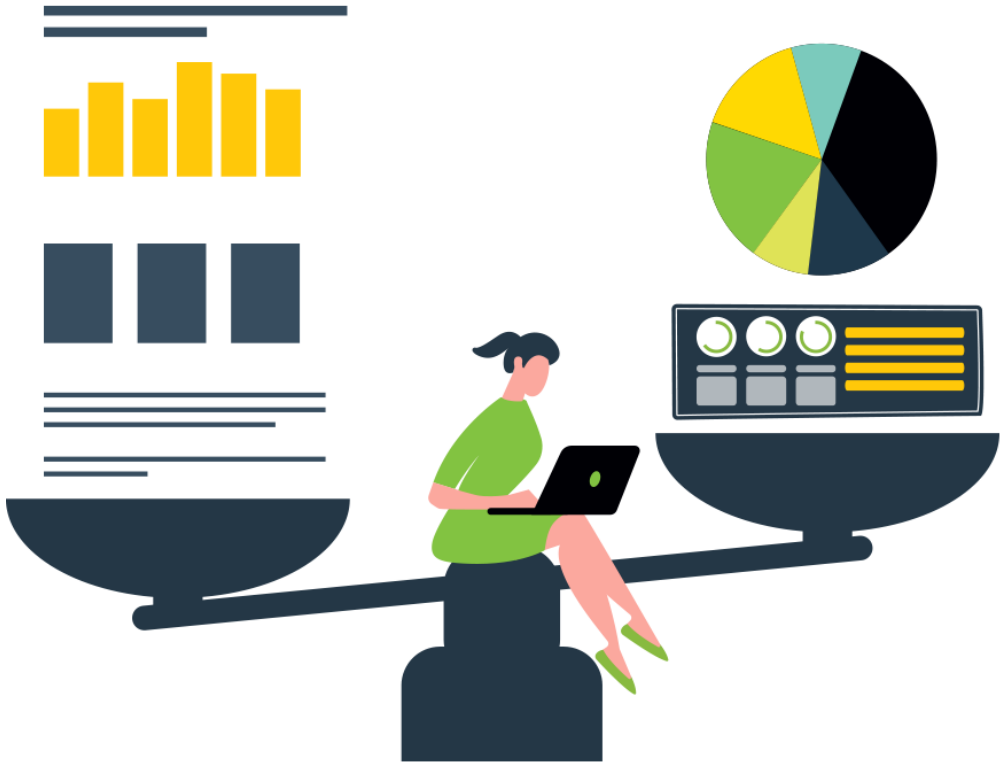
https://www.logianalytics.com/definitiveguidetopredictiveanalytics/solving-common-data-challenges/

Avoid Bias in Your Data and Models

1. Data Bias
2. Selection bias







https://www.logianalytics.com/definitiveguidetopredictiveanalytics/solving-common-data-challenges/

Validate your model is working and establish a performance baseline

Accuracy

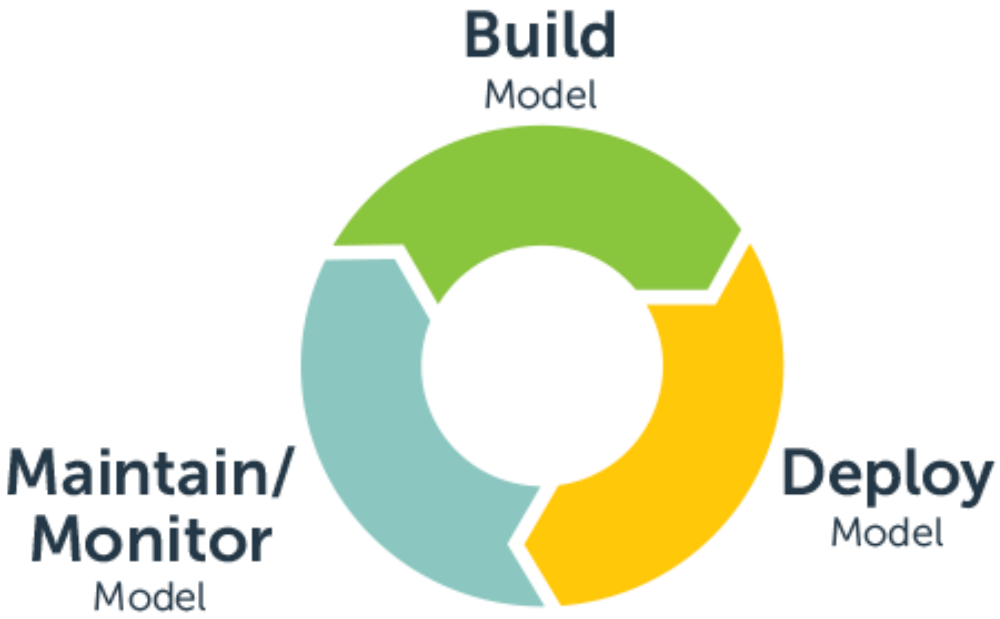
Watch for Imbalanced Data

https://www.logianalytics.com/definitiveguidetopredictiveanalytics/solving-common-data-challenges/

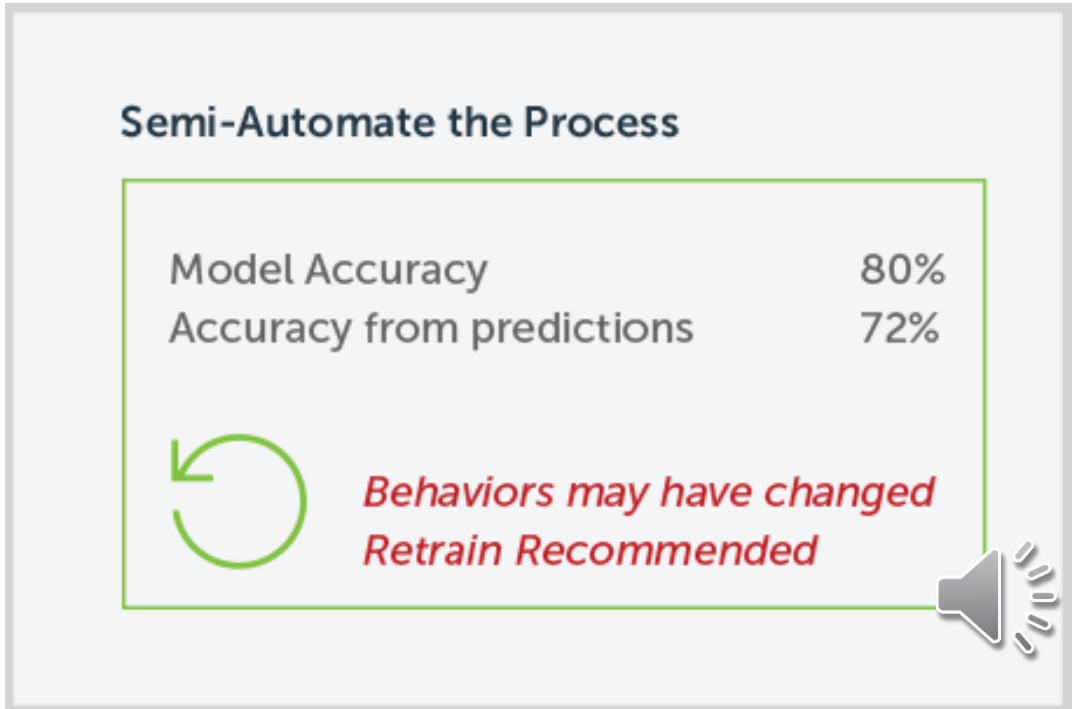
Know When It's Time to Refresh

Seasonal.

Measurement-based.

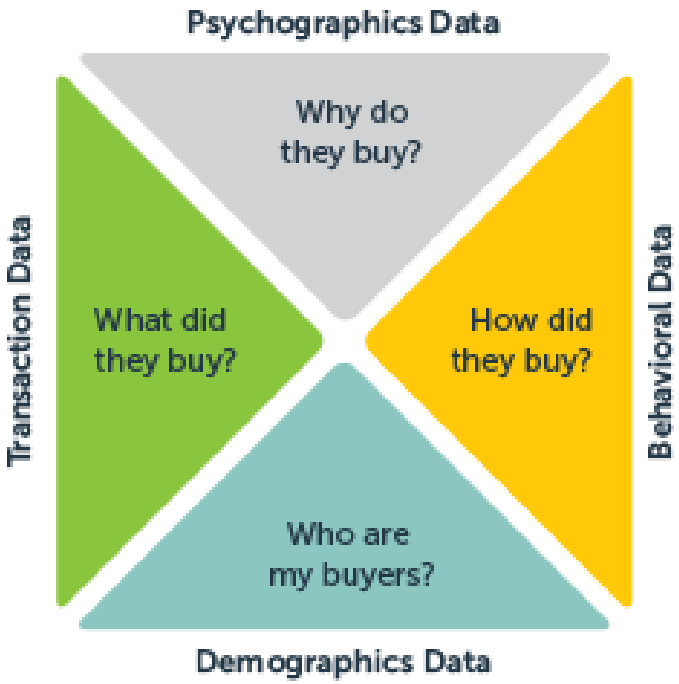


https://www.logianalytics.com/definitiveguidetopredictiveanalytics/solving-common-data-challenges/



Boost Predictive Performance Over Time

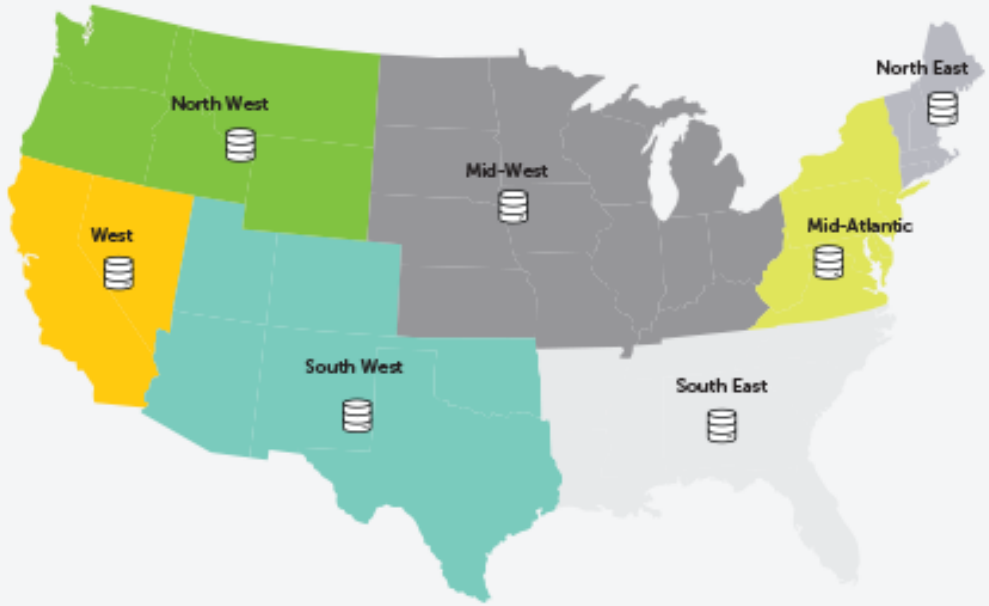
Data Variety
Model Refinement Based on
Additional Performance Metrics



https://www.logianalytics.com/definitiveguidetopredictiveanalytics/solving-common-data-challenges/

How to Adapt to Changing Business Requirements

One model for entire USA or models per region to cater regional differences and only retrain for any changes in that region without affecting models in other regions



Data Analytics: Overview

Qualitative and quantitative data
Structured and unstructured data
The data analysis process
Types of data analytics
Data analytics trends

https://www.itproportal.com/features/six-ways-predictive-data-analytics-are-reshaping-marketing/

Qualitative and quantitative data

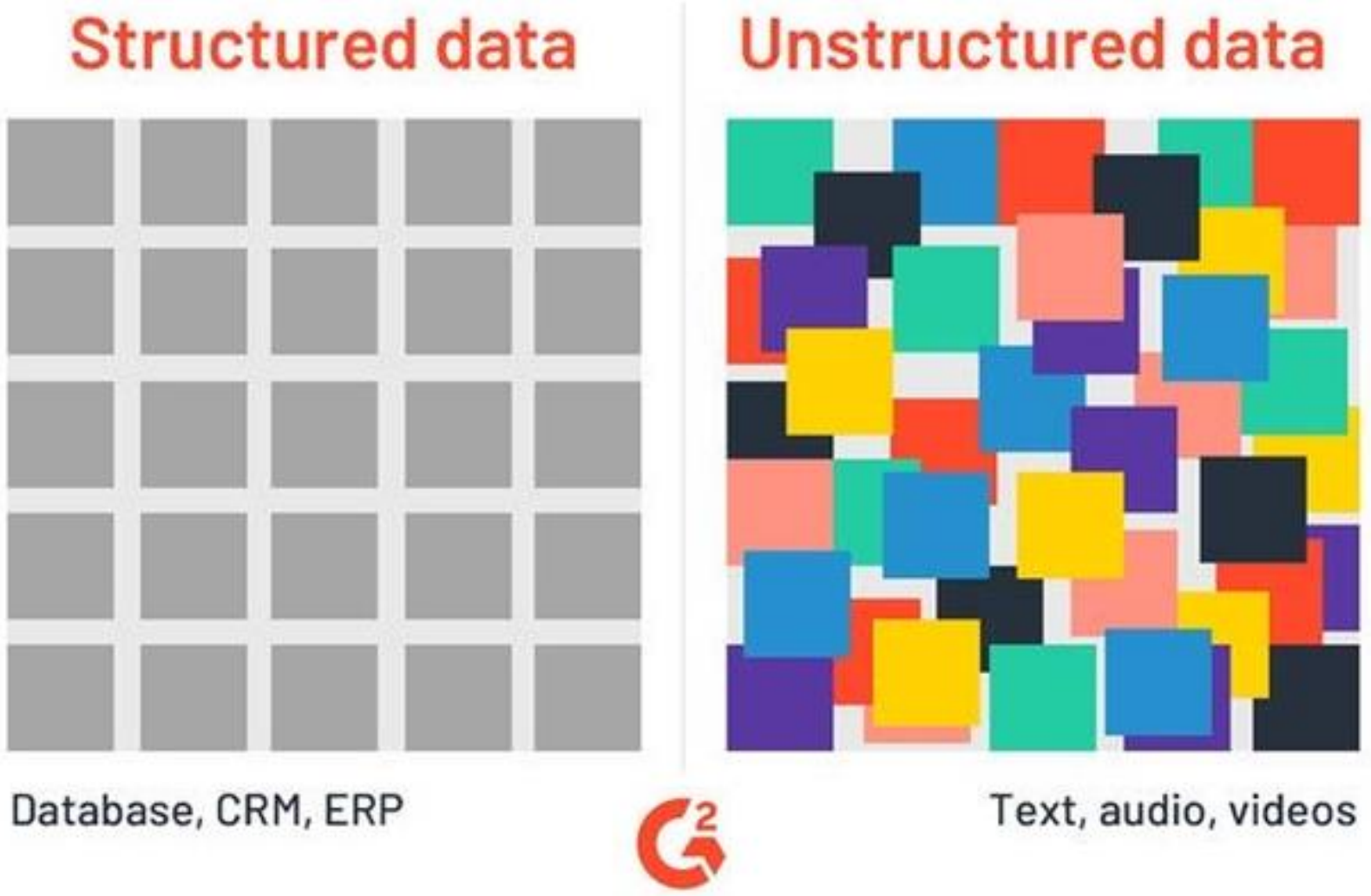
What is qualitative data?
This bookcase...

- Is made of wood
- Was built in Italy
- Is deep brown
- Has golden knobs
- Smells like oak
- Has a smooth finish

What is quantitative data?
This bookcase...

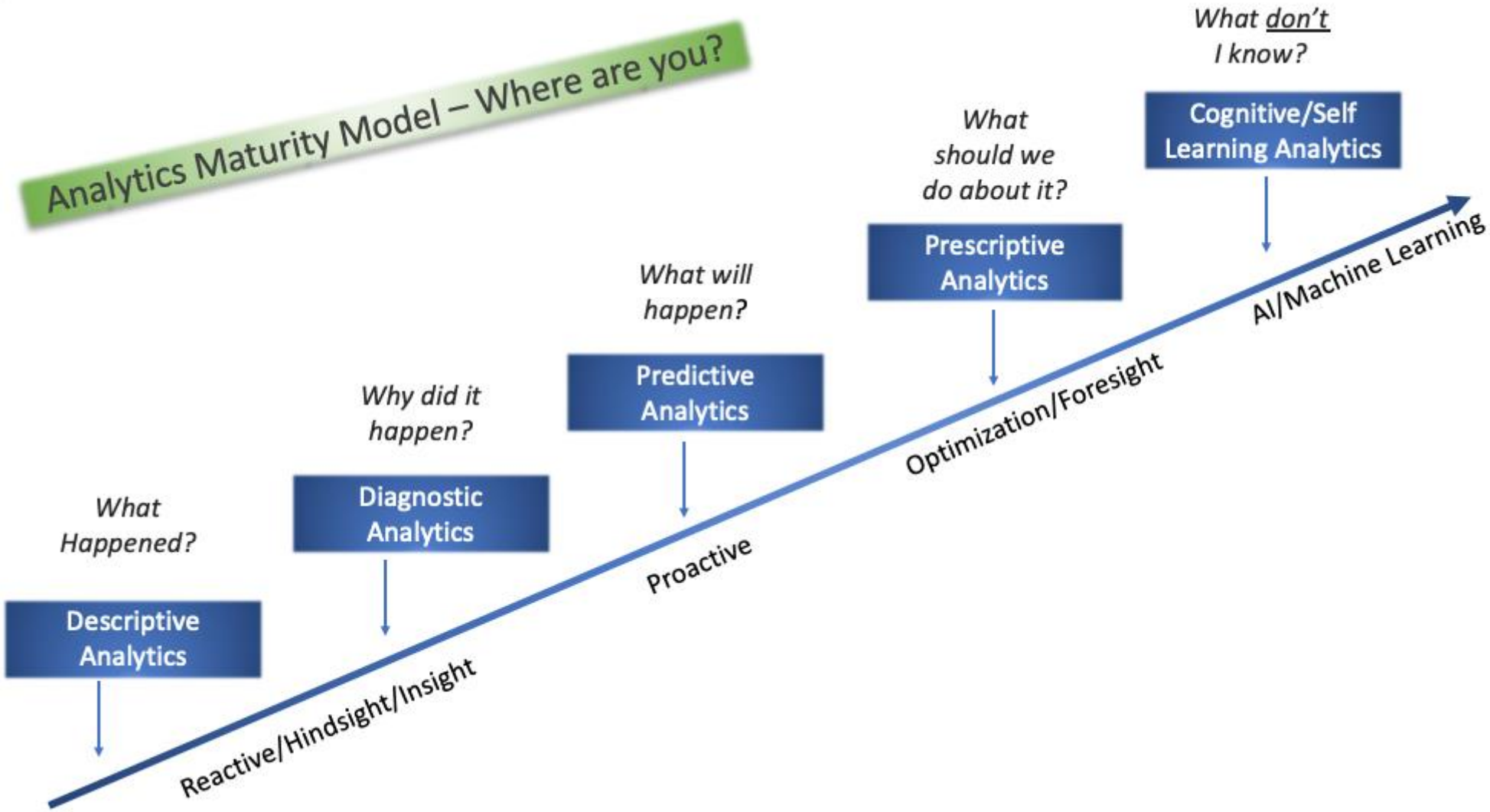
- Is 3 feet tall
- Weighs 100 pounds
- Has 15 books on it
- Has 3 shelves
- Has 2 cabinets
- Sells for $1500

Structured and unstructured data



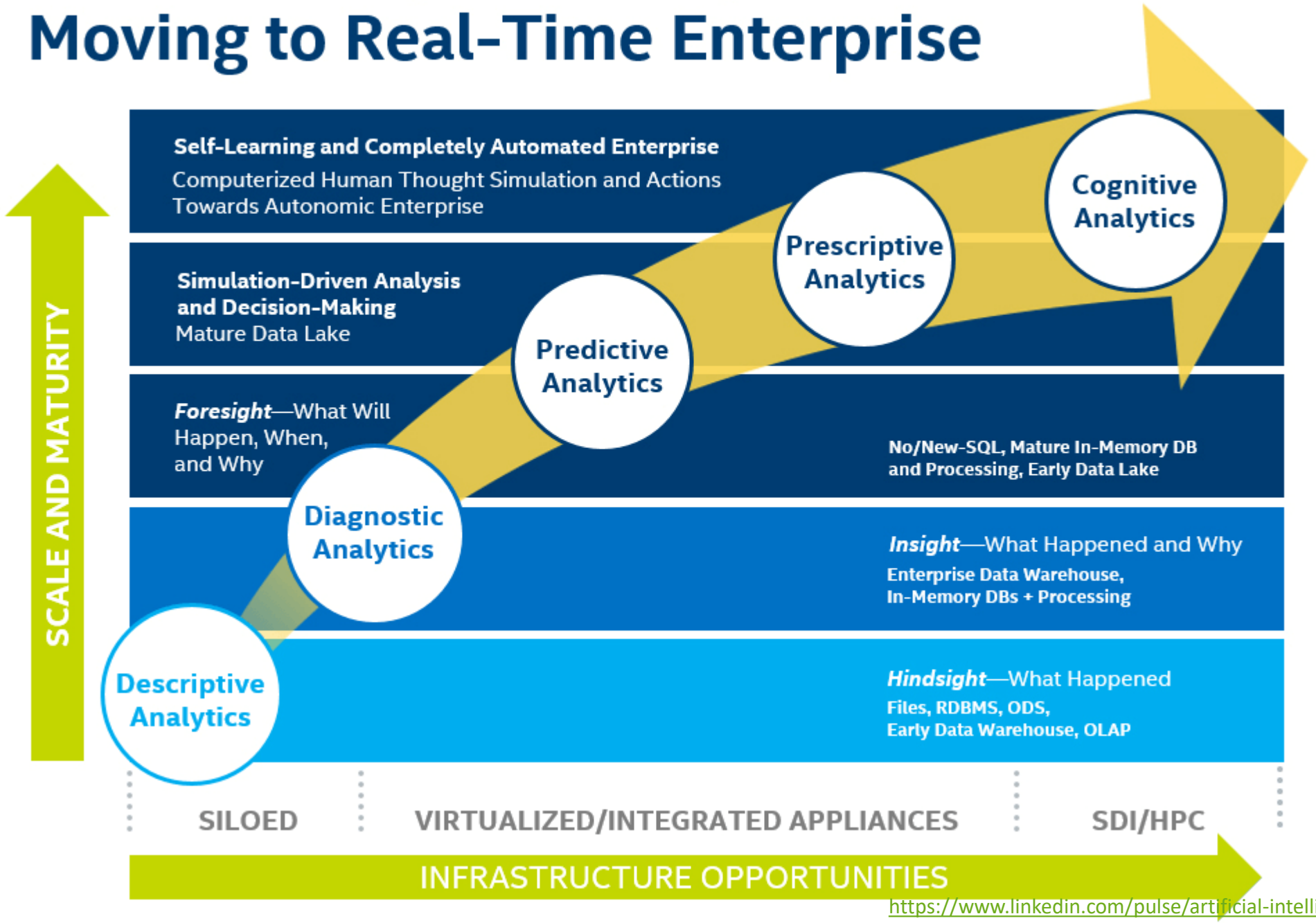
The data analysis process

Types of data analytics

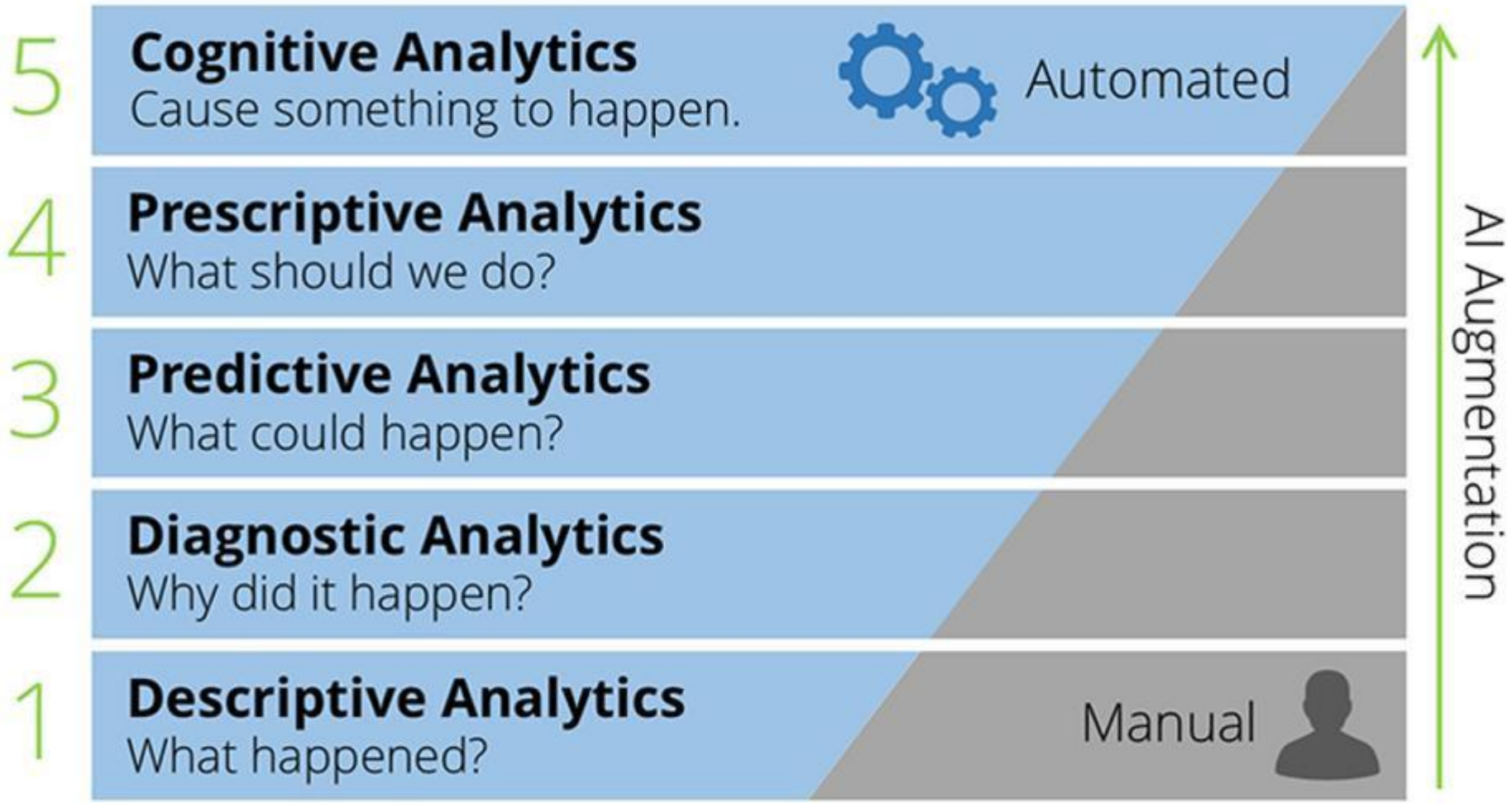


https://www.ecapitaladvisors.com/blog/analytics-maturity/

Advanced Analytics Maturity Path:
Moving to Real-Time Enterprise



https://www.linkedin.com/pulse/artificial-intelligence-data-analytics-samir-el-masri-phd/

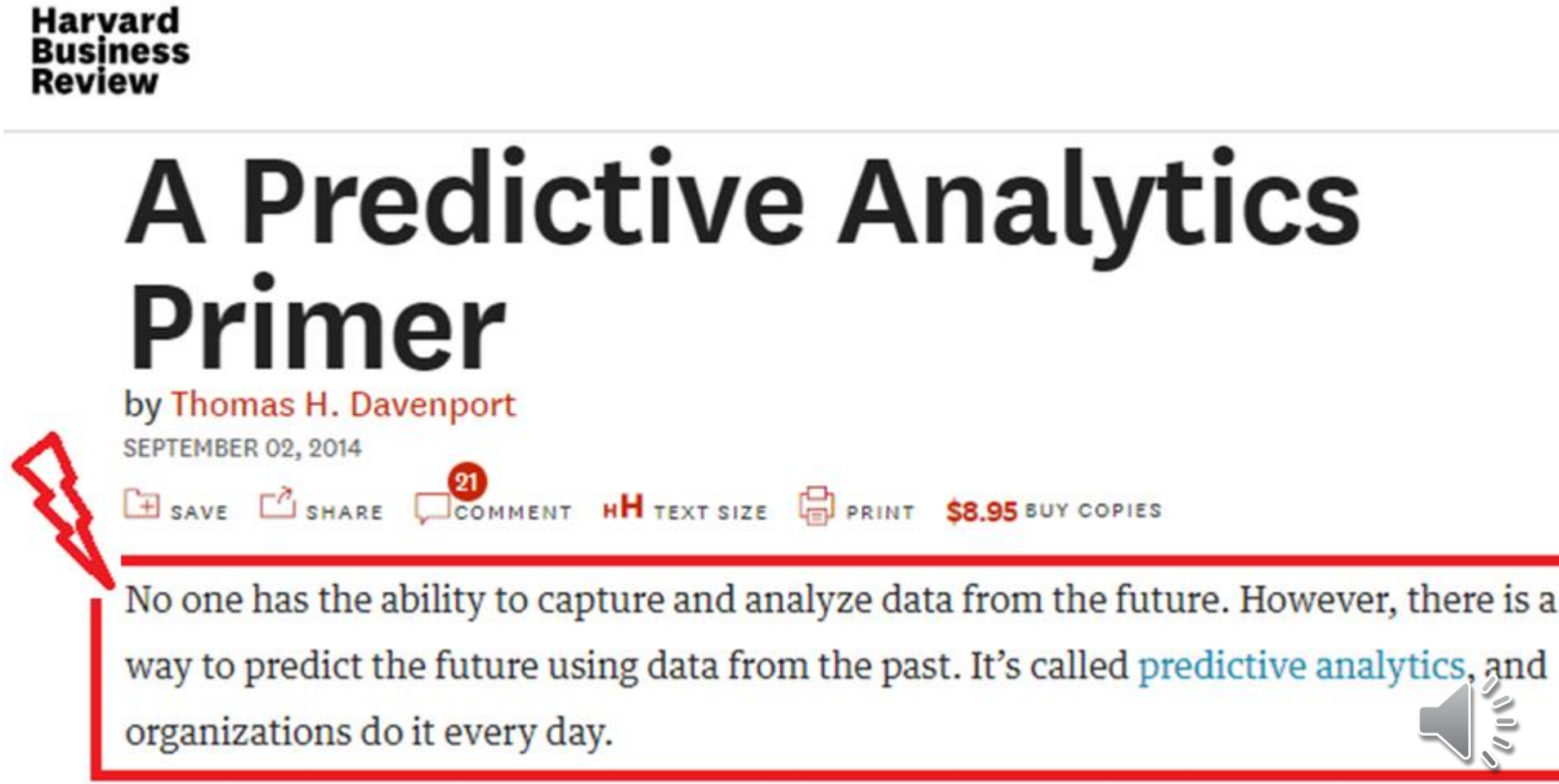


https://www.forbes.com/sites/brentdykes/2017/01/11/crawl-with-analytics-before-running-with-artificial-intelligence

Predictive Analytics

- what the future holds (to a certain degree)
- show a variety of possible outcomes

Application



Why is predictive analytics important?

- Lead generation
- Enhanced marketing efforts, targeted to specific customers
- Identification of future trends
- Identification of growth opportunities
- Reduction of customer churn
- Improved content marketing and distribution


LEAD GENERATION
PROMOTION
CONSUMER
CHANNEL
STRATEGY
TRAFFIC
POTENTIAL
INFUENCE

Targeted Profiling of Customers

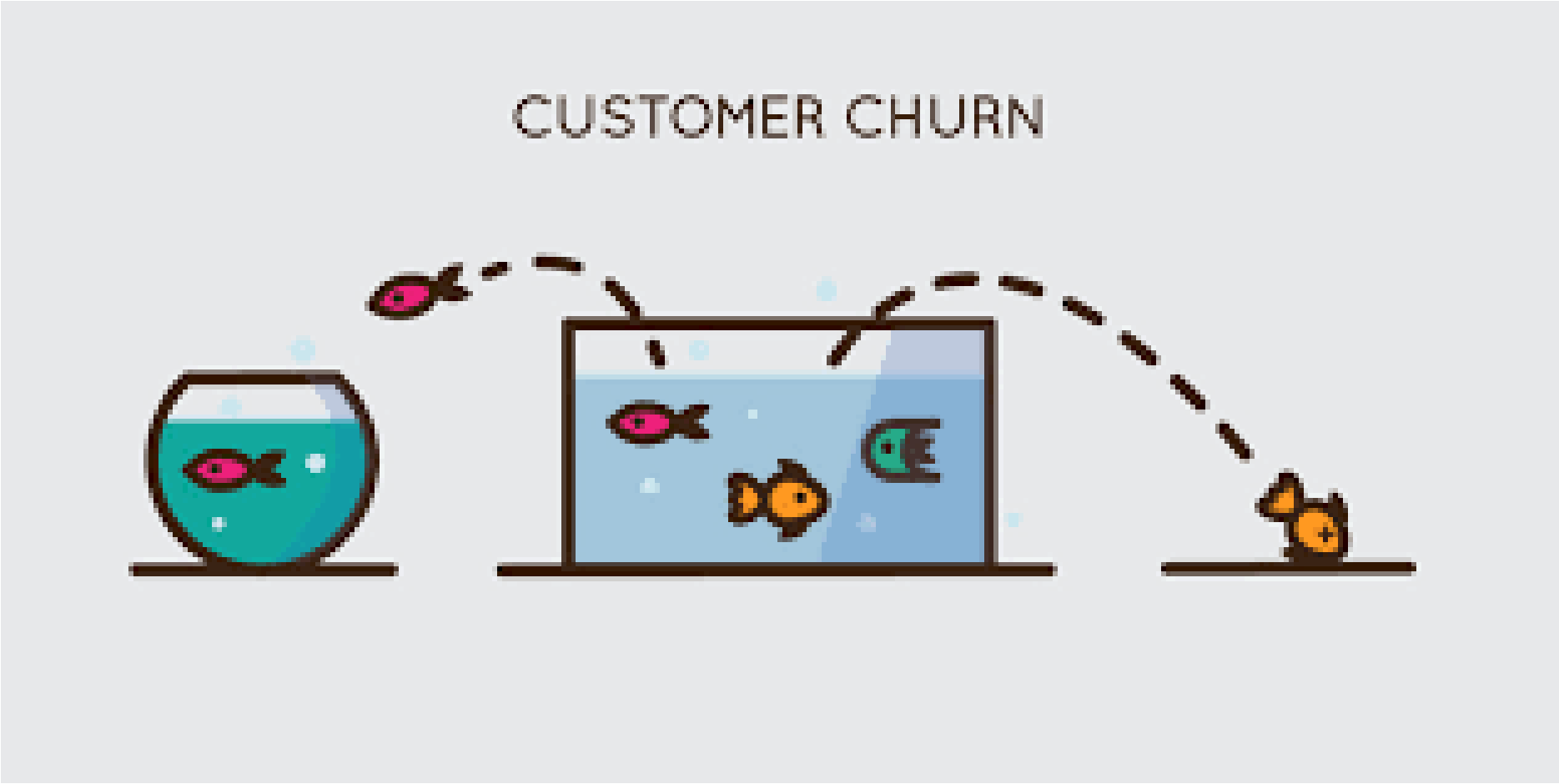
Improved Lead Scoring

Segmentation for Nurture Campaigns

Improved Content Distribution

Accurate Prediction of Lifetime Value

More Insight to Reduce Churn



Enhanced Upsell/cross-sell Opportunities

Improved Determination of Product Fit

Determining the Optimal Campaign Channels & Content

Identify New Trends and Growth Opportunities

Examples of predictive analysis applications

amazon

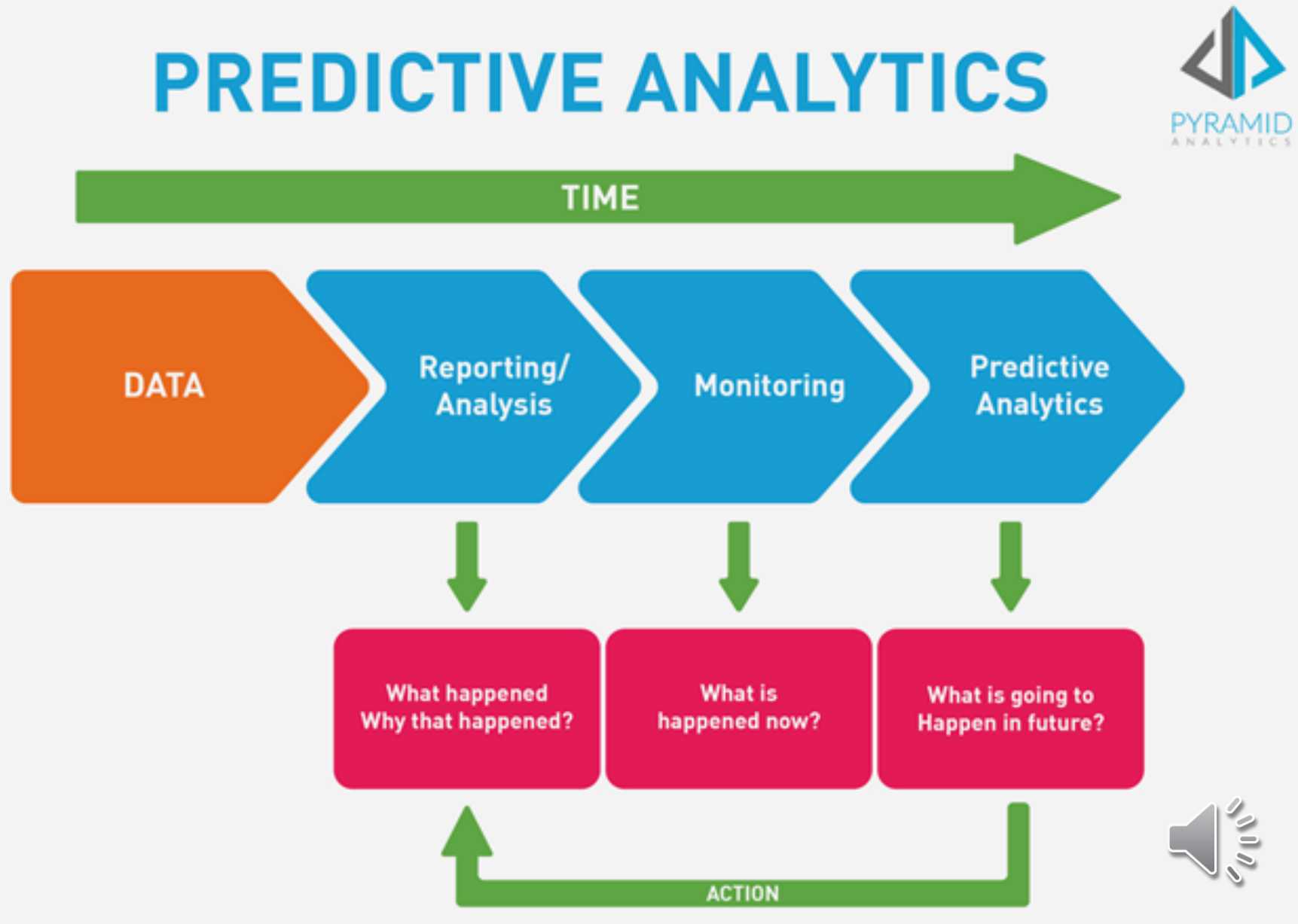
NETFLIX

Process of predictive analytics

1. Project Definition
2. Data Collection
3. Data Analysis and Statistics
4. Modeling
5. Deployment/Integration
6. Model Monitoring

Methodologies used in predictive analytics

Logistic Regression
Decision Trees
Time Series Analysis
Text Analytics



Five Industry Examples of Predictive Analytics

Healthcare

Manufacturing

Finance

Insurance

SaaS

Improving Patient Outcomes

Problem:
Benefits:
Data to Analyze:
Actions to Take:

Healthcare

Predictive Maintenance

Manufacturing

Problem:

Benefits:

Data to Analyze:

Actions to Take:

Sensor 1	Sensor 2	Sensor 3	Sensor 4	Sensor 5	Sensor 6		State
0.0051	0.035	0.001750	0.004	1.575	0.173250	.....	HEALTHY
0.0051	0.035	0.001750	0.004	1.575	0.173250		OK
.....							

Predicting Late Payments

Finance

Problem:

Benefits:

Data to Analyze:

Actions to Take:

Preventing Fraud

Insurance

Problem:

Benefits:

Data to Analyze:

Actions to Take:

Reducing Customer Churn

SaaS

Problem:

Benefits:

Data to Analyze:

Actions to Take:

Seven Steps to Start Your PA Project

1. **Identify a Problem to Solve**
2. **Select and Prepare Your Data**
3. **Involve Others**
4. **Run Your Predictive Analytics Models**
5. **Close the Gap Between Insights and Actions**
6. **Build Prototypes**
7. **Iterate Regularly**

Identify a Problem to Solve

PADS (Performance Analytics Decision Support) Framework

Preventing Problems:
Assisting Humans:
Detecting Problems:
Streamlining Services:

Requirements for Predictive Analytics Problems

Loss Prevention
Increase Happiness
Improve Processes

Select & Prepare Your Data

Data Volume:

Data Scalability:

Involve Others

Run Your Predictive Analytics Models

1. Classification Model:
2. Clustering Model:
3. Forecast Model:
4. Outliers Model:
5. Time Series Model:

Close the Gap Between Insights & Actions

Build Prototypes

Iterate Regularly

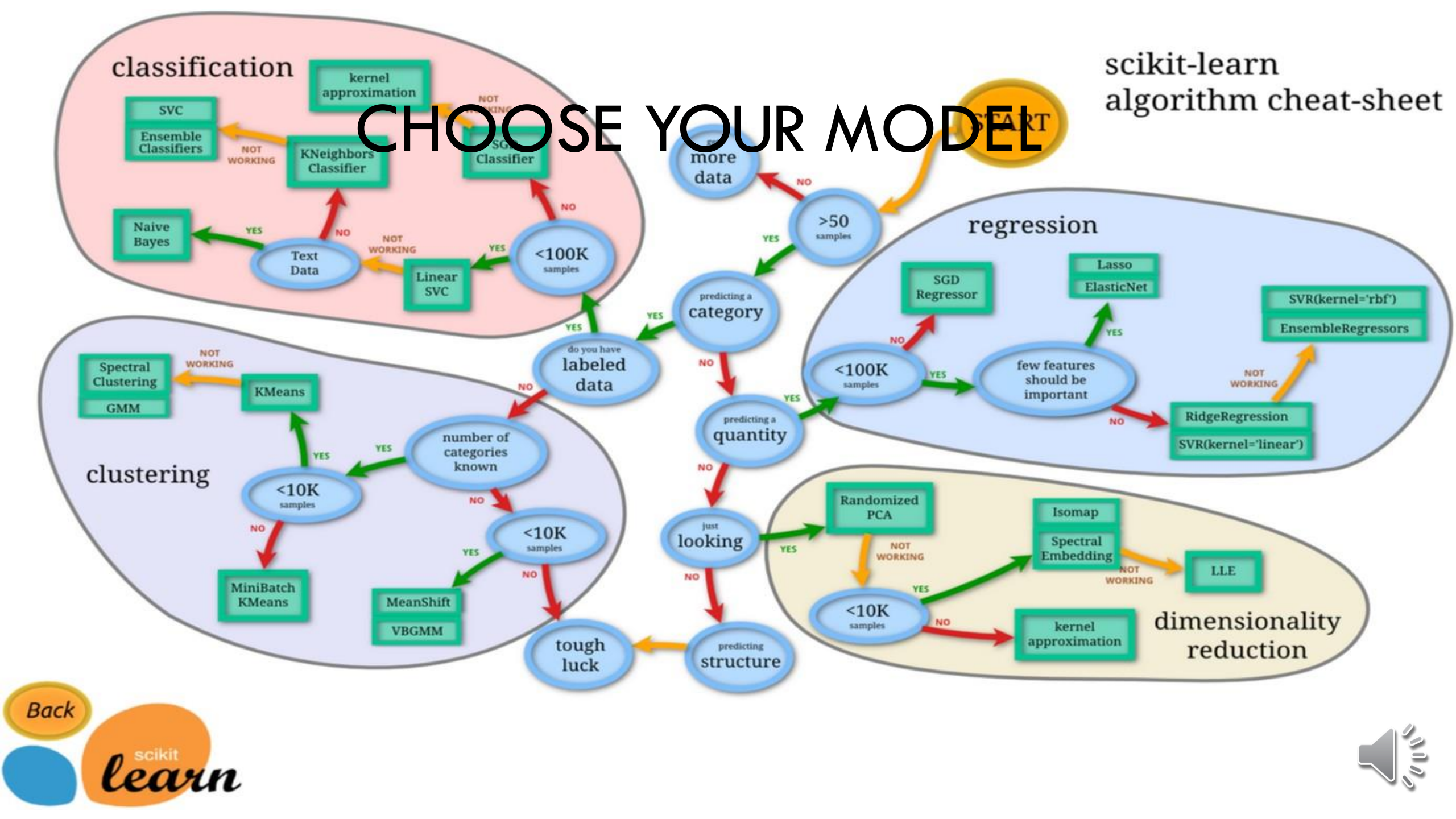
Data Analysis Techniques

1. Techniques based on Mathematics and Statistics
2. Techniques based on Artificial Intelligence and Machine Learning
3. Techniques based on Visualization and Graphs

Techniques based on Mathematics and Statistics

Descriptive Analysis
Dispersion Analysis:
Regression Analysis:
Factor Analysis:
Discriminant Analysis:
Time Series Analysis:

CHOOSE YOUR MODEL



Techniques based on AI and ML

Artificial Neural Networks:

Decision Trees:

Evolutionary Programming:

Fuzzy Logic:

Techniques based on Visualization and Graphs

Column Chart, Bar Chart:

Line Chart:

Area Chart:

Pie Chart:

Funnel Chart:

Word Cloud Chart:

Gantt Chart:

Radar Chart:

Scatter Plot:

Bubble Chart:

Gauge:

Frame Diagram:

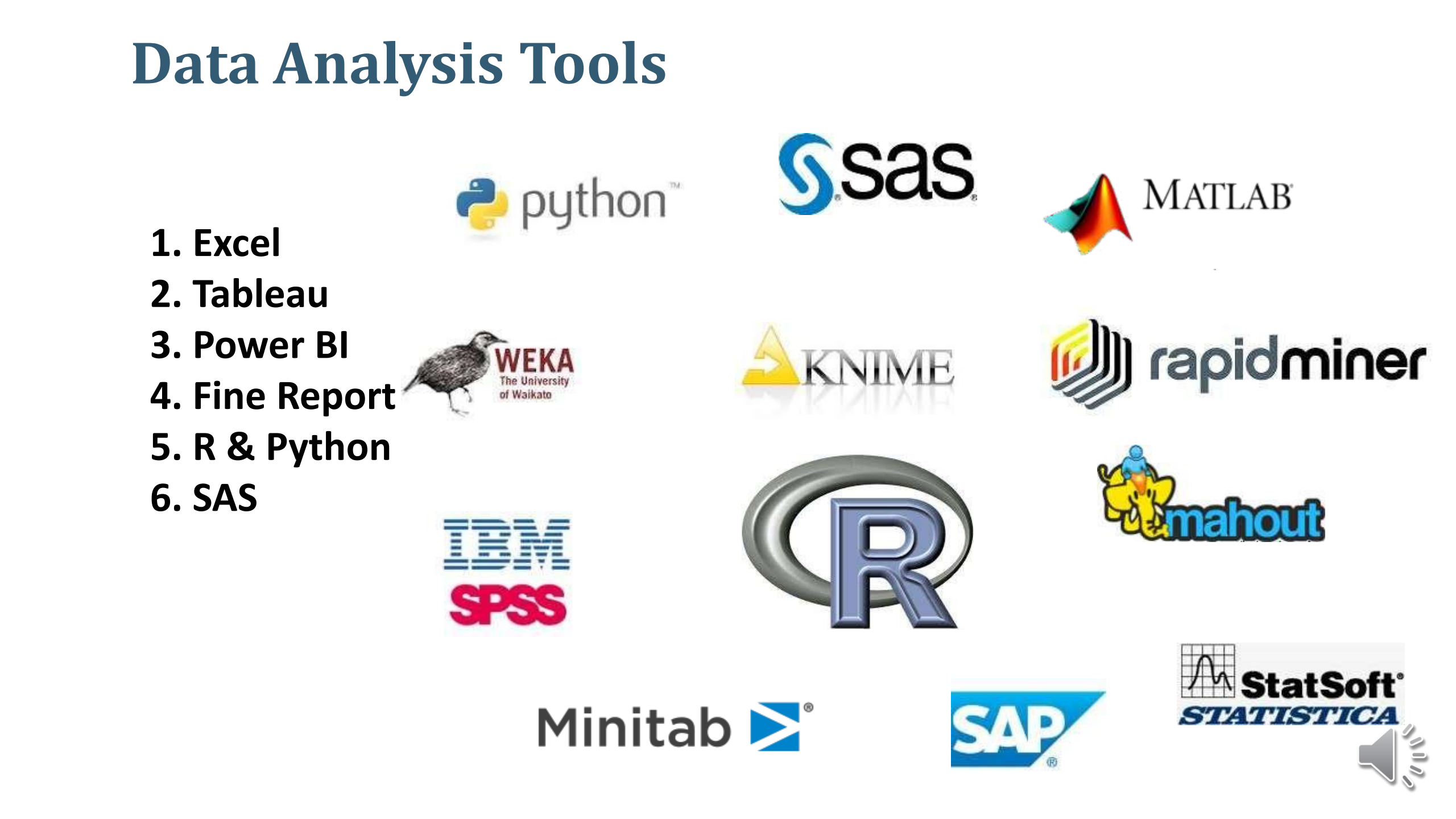
Rectangular Tree Diagram:

Map

- **Regional Map:**
- **Point Map:**
- **Flow Map:**
- **Heat Map**:

Data Analysis Tools

1. Excel
2. Tableau
3. Power BI
4. Fine Report
5. R & Python
6. SAS



Python Packages for Data Analysis

https://www.kdnuggets.com/2016/04/datacamp-learning-python-data-analysis-data-science.html

https://learn.datacamp.com/courses/tech:python

Python Code of the predictive modeling tasks

Load dataset

```
import pandas as pd

import numpy as np

from sklearn.preprocessing import LabelEncoder

import random

from sklearn.ensemble import RandomForestClassifier

from sklearn.ensemble import GradientBoostingClassifier
```

Identifying missing values

```
fullData.isnull().any()#Will return the feature with True or False,True means have missing value else Fa

lse
```

```
#Impute numerical missing values with mean

fullData[num_cols] = fullData[num_cols].fillna(fullData[num_cols].mean(),inplace=True)
```

Impute missing values

```
#Impute categorical missing values with -9999

fullData[cat_cols] = fullData[cat_cols].fillna(value = -9999)
```

Checking correlation and visualization

```
import seaborn as sns
import matplotlib.pyplot as plt
%matplotlib inline
corr = df.corr()
sns.heatmap(corr,
            xticklabels=corr.columns,
            yticklabels=corr.columns)
```

Training and test data split

```
from sklearn.cross_validation import train_test_split

train, test = train_test_split(df1, test_size = 0.4)
train = train.reset_index(drop=True)
test = test.reset_index(drop=True)

features_train = train[list(vif['Features'])]
label_train = train['target']
features_test = test[list(vif['Features'])]
label_test = test['target']
```

Predictive models on training data

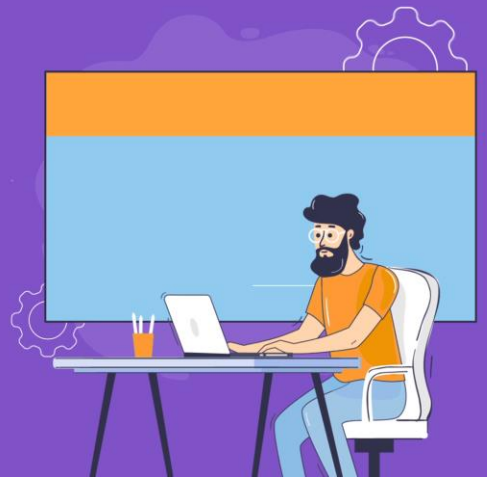
```
from sklearn.ensemble import RandomForestClassifier
clf = RandomForestClassifier()

clf.fit(features_train,label_train)

pred_train = clf.predict(features_train)
pred_test = clf.predict(features_test)

from sklearn.metrics import accuracy_score
accuracy_train = accuracy_score(pred_train,label_train)
accuracy_test = accuracy_score(pred_test,label_test)

from sklearn import metrics
fpr, tpr, _ = metrics.roc_curve(np.array(label_train),
clf.predict_proba(features_train)[:,1])
auc_train = metrics.auc(fpr,tpr)

fpr, tpr, _ = metrics.roc_curve(np.array(label_test),
clf.predict_proba(features_test)[:,1])
auc_test = metrics.auc(fpr,tpr)
```

References

https://www.1to1media.com/data-analytics/analysis-vs-analytics-whats-difference
https://theappsolutions.com/blog/development/what-is-big-data-analytics/
https://www.pyramidanalytics.com/blog/details/report-understanding-predictive-analytics
https://www.valamis.com/hub/descriptive-analytics
https://www.branex.ca/blog/what-is-big-data-analytics-why-is-matters-to-your-business/
https://www.gangboard.com/blog/what-is-big-data-analytics
https://learn.g2.com/what-is-data-analytics
https://melsatar.blog/2017/07/30/the-evolution-of-analytics/
http://managedcommunityanalytics.com/what-is-analytics/
https://becominghuman.ai/lets-talk-about-advanced-analytics-a-brief-look-at-artificial-intelligence-bf1c7a7d3f96
https://datafloq.com/read/the-four-types-of-data-analytics/3903
https://nicoleparmar.com/moving-from-descriptive-to-predictive-and-prescriptive-analytics/
https://www.kdnuggets.com/2017/07/4-types-data-analytics.html
https://www.edureka.co/blog/what-is-data-analytics/
https://medium.com/analytics-for-humans/a-comprehensive-guide-to-predictive-analytics-d1eb688f37dd
https://www.logianalytics.com/definitiveguidetopredictiveanalytics/introduction-to-predictive-analytics/
https://www.gangboard.com/blog/what-is-data-science/
https://hackr.io/blog/what-is-data-analysis-methods-techniques-tools
https://www.bbva.com/en/five-vs-big-data/
https://www.freepik.com/premium-vector/big-data-analytics-informed-decisions_2067385.htm

https://www.youtube.com/watch?v=cUw3DsDpQCE
ttps://www.how2shout.com/tools/top-open-source-data-analysis-software.html
http://ideal.ece.utexas.edu/courses/ee380l_ese/ppt/Research_Paper_Presentation_Pandas_Moshiul_Arefin.pdf
https://users.cs.fiu.edu/~giri/
https://www.analyticsvidhya.com/blog/2016/01/complete-tutorial-learn-data-science-python-scratch-2/
https://medium.com/my-data-camp-journey/predictive-analysis-in-python-97ca5b64e97f
https://medium.com/datadriveninvestor/a-simple-guide-to-creating-predictive-models-in-python-part-1-8e3ddc3d7008
https://ift6758.github.io/
https://www.andrew.cmu.edu/user/georgech/95-865/
https://www.ecapitaladvisors.com/blog/analytics-maturity/
https://www.logianalytics.com/definitiveguidetopredictiveanalytics/7-steps-to-start-your-predictive-analytics-project/
https://learn.g2.com/what-is-data-analytics#trends
https://data-flair.training/blogs/data-analytics-tutorial/
https://www.omnisci.com/technical-glossary/predictive-analytics

THANKS
FOR WATCHING



Sometimes questions are more important than answers

javed.sheikh@uskt.edu.pk
as8699666@gmail.com

WhatsApp +923348699666
https://www.researchgate.net/profile/Javed_Sheikh4
https://www.linkedin.com/in/javedanjumsheikh