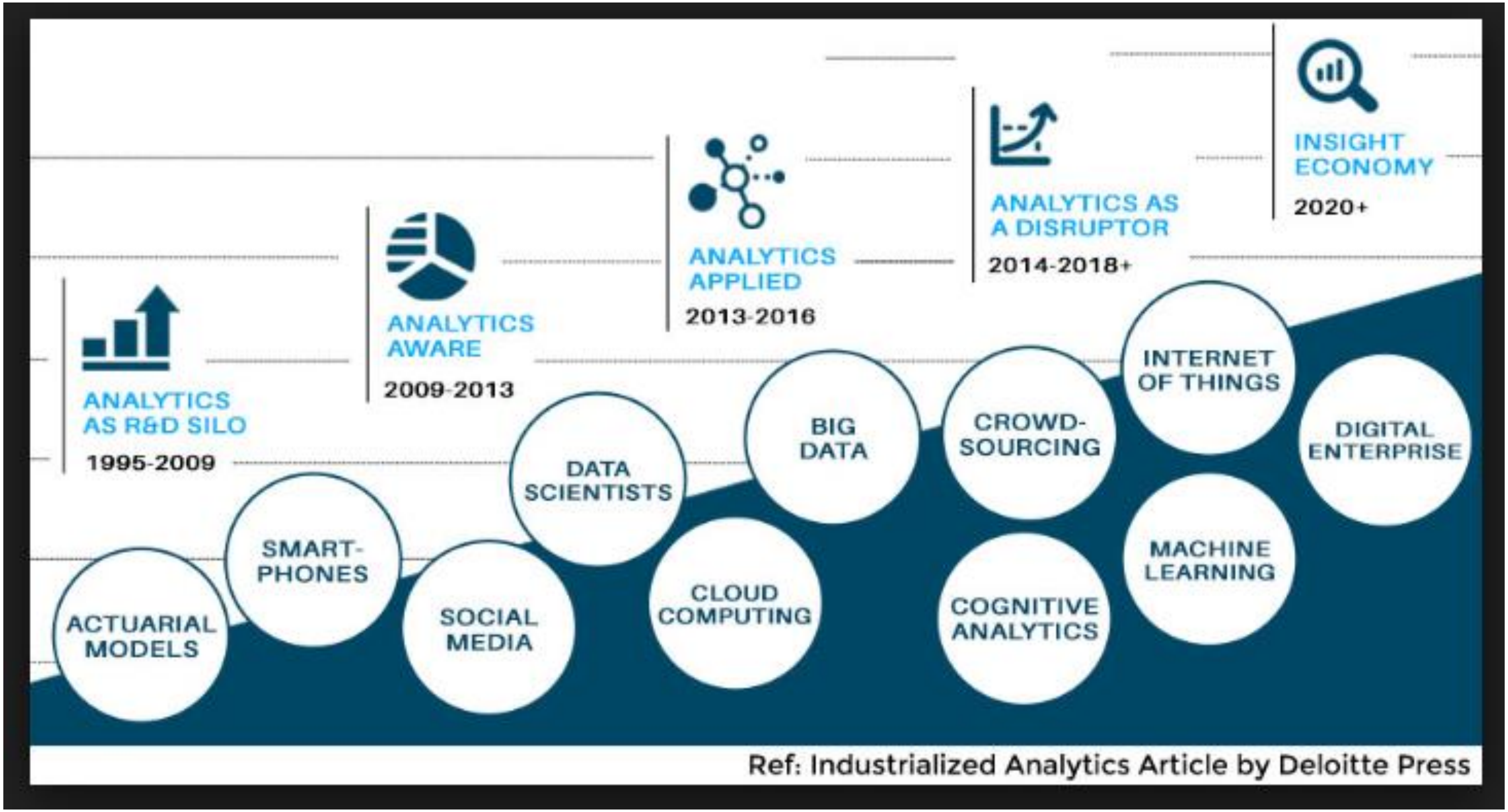
Discovering value in information
Analytics Landscape Journey

Dr.Satish Raghavendran PhD (US)

24 July 2020

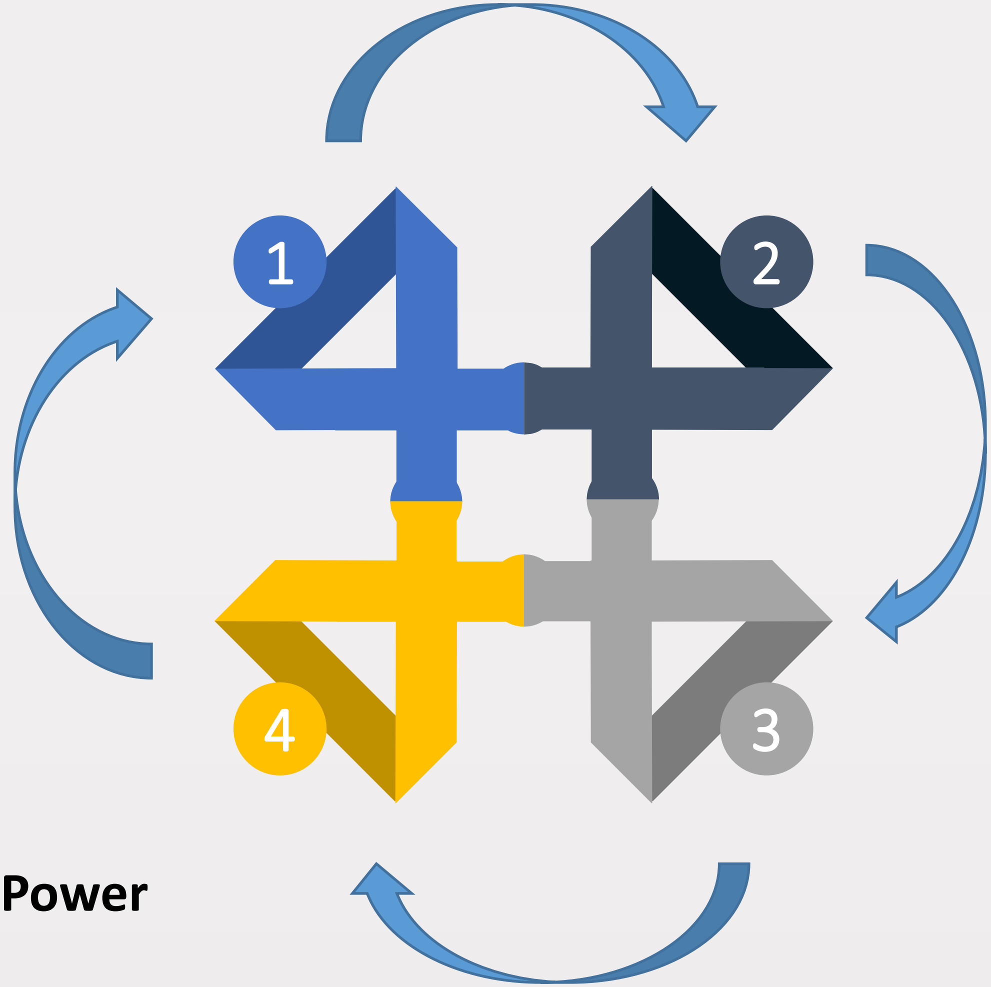
How did Analytics Evolve?



It is an evolving science, but not something new

Analytics is the practice of using data to drive business strategy and performance. It includes a range of approaches and solutions from looking backward to evaluate what happened in the past to forward-looking scenario planning and predictive modeling.

Analytics Drivers – Virtuous Cycle



1 **Digitization**

2 **Globalization**

3 **Real-Time Decisioning**

4 **Algorithms & Computational Power**

DEMYSTIFYING THE ANALYTICS MARKET

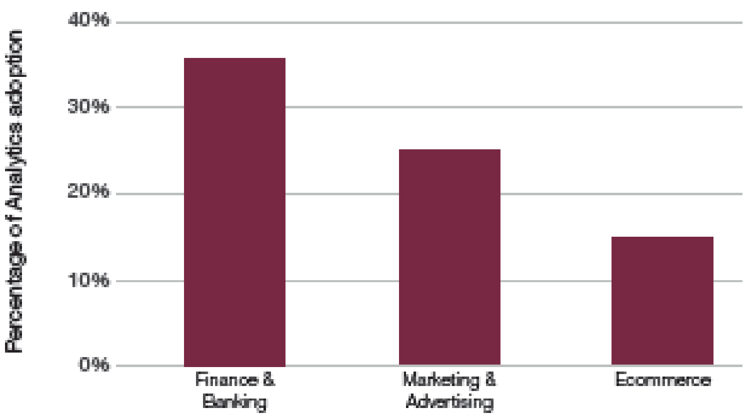
- The cumulative analytics market in India stands at $30 Billion
- Outsourcing is the main driver of revenue for Indian vendors accounting for $27 billion in revenue
- The domestic analytics market stands at $3.03 billion in size and is expected to double by 2025

Cumulative analytics market in India stands at **$30 Billion**	Outsourcing is the main driver of revenue amounting at **$27 Billion**	Domestic Analytics market in India stands at **$3.03 Billion**

SECTOR TYPE FOR ANALYTICS INDUSTRY

The analytics adoption is led by India's BFSI sector which has also broadened its portfolio by piloting several key AI and ML-related initiatives. After finance and banking, marketing and advertising sector is the second biggest adopter of analytics.

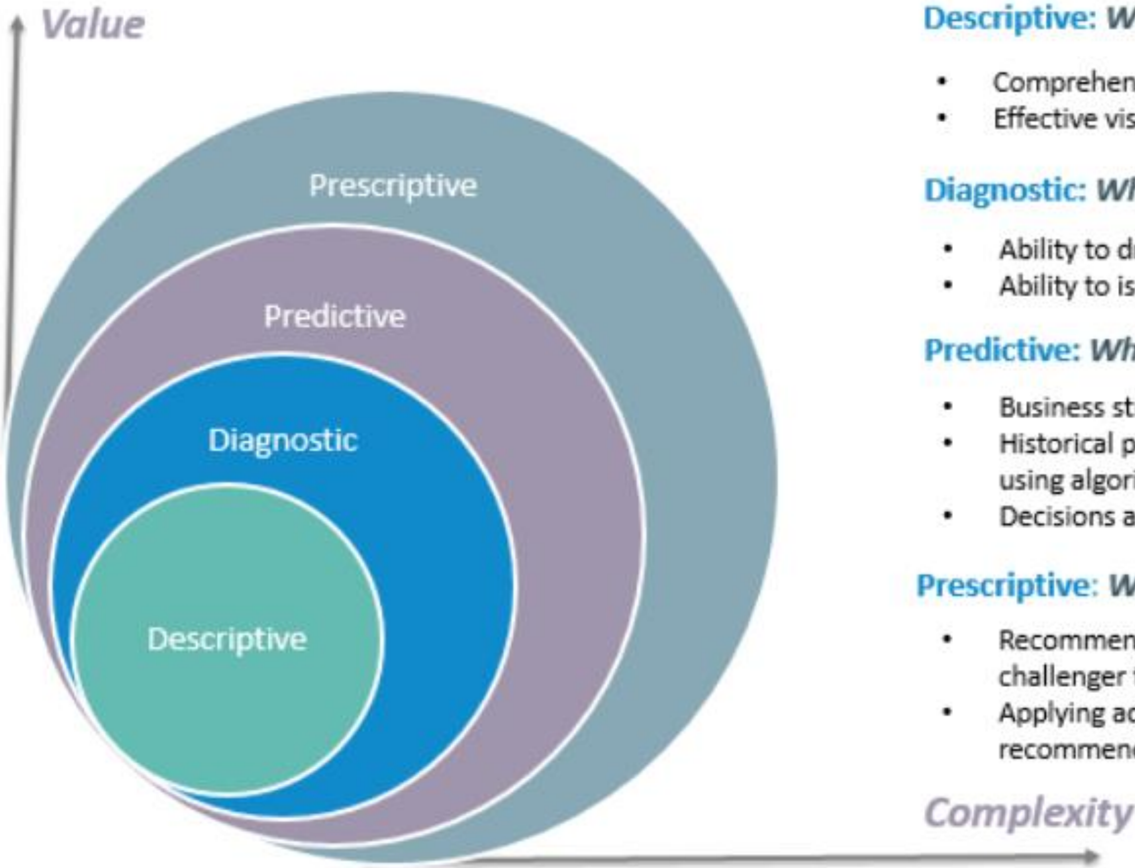
- In terms of Sector type, Finance & Banking continues to be the largest sector being served by analytics in India. Overall, 36% of analytics market size in India comes from Finance & Banking
- Marketing & advertising comes second at 25%, followed by E-commerce sector at 15% of analytics revenues in India



PERCENTAGE OF ANALYTICS ADOPTION ACROSS INDUSTRIES

Data Analytics Taxonomy

4 types of Data Analytics



What is the data telling you?

Descriptive: ***What's happening in my business?***

- Comprehensive, accurate and live data
- Effective visualisation

Diagnostic: ***Why is it happening?***

- Ability to drill down to the root-cause
- Ability to isolate all confounding information

Predictive: ***What's likely to happen?***

- Business strategies have remained fairly consistent over time
- Historical patterns being used to predict specific outcomes using algorithms
- Decisions are automated using algorithms and technology

Prescriptive: ***What do I need to do?***

- Recommended actions and strategies based on champion / challenger testing strategy outcomes
- Applying advanced analytical techniques to make specific recommendations

Principa
www.principa.co.za

Business challenges which make Analytics a hot topic

The level of sophistication around "analytics-informed" decisions required to compete is increasing exponentially

Enterprise, social, mobile data mash-up

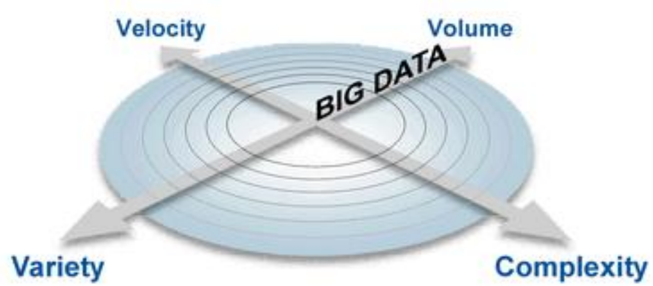
Dimensions of Analytics

Foresight / Insight / Hindsight

Predictive and Prescriptive
- Optimization algorithms
- Simulation and modeling
- Quantitative analyses
- Advanced forecasting

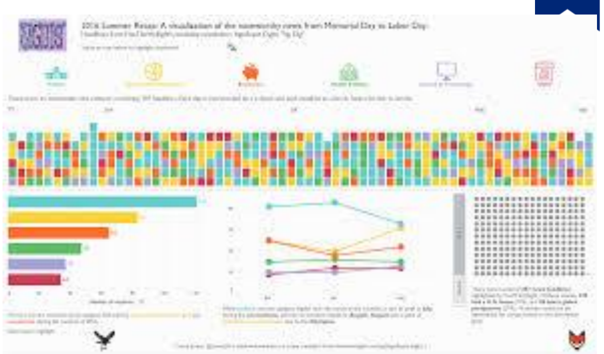
Descriptive
- Role-based performance metrics
- Exceptions and alerts
- Slice and dice queries and drill downs
- Management reporting
- Enterprise data management

Increasing business advantage and sophistication



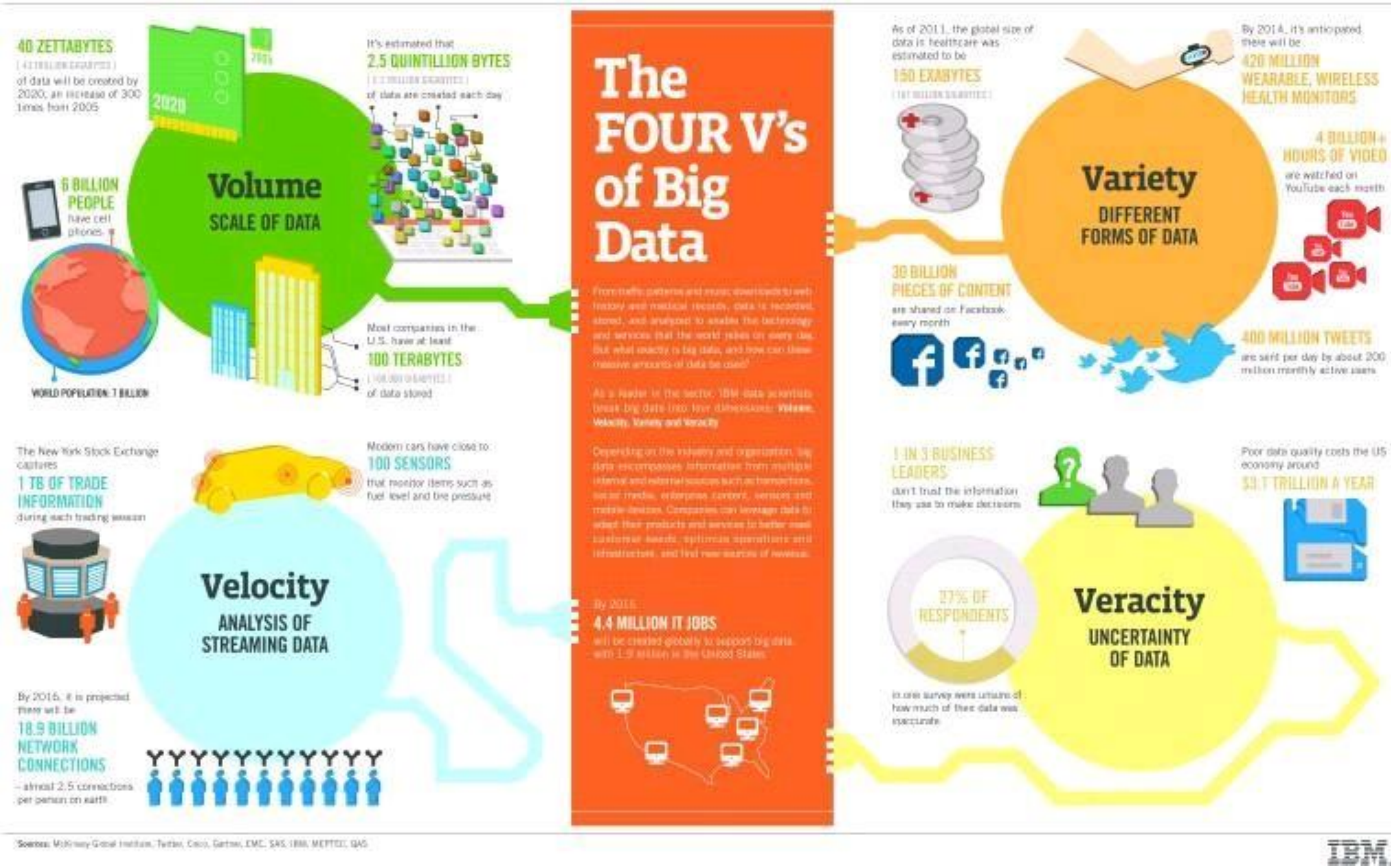
Enhanced "What if" Analysis

Interactive graphics and dashboards

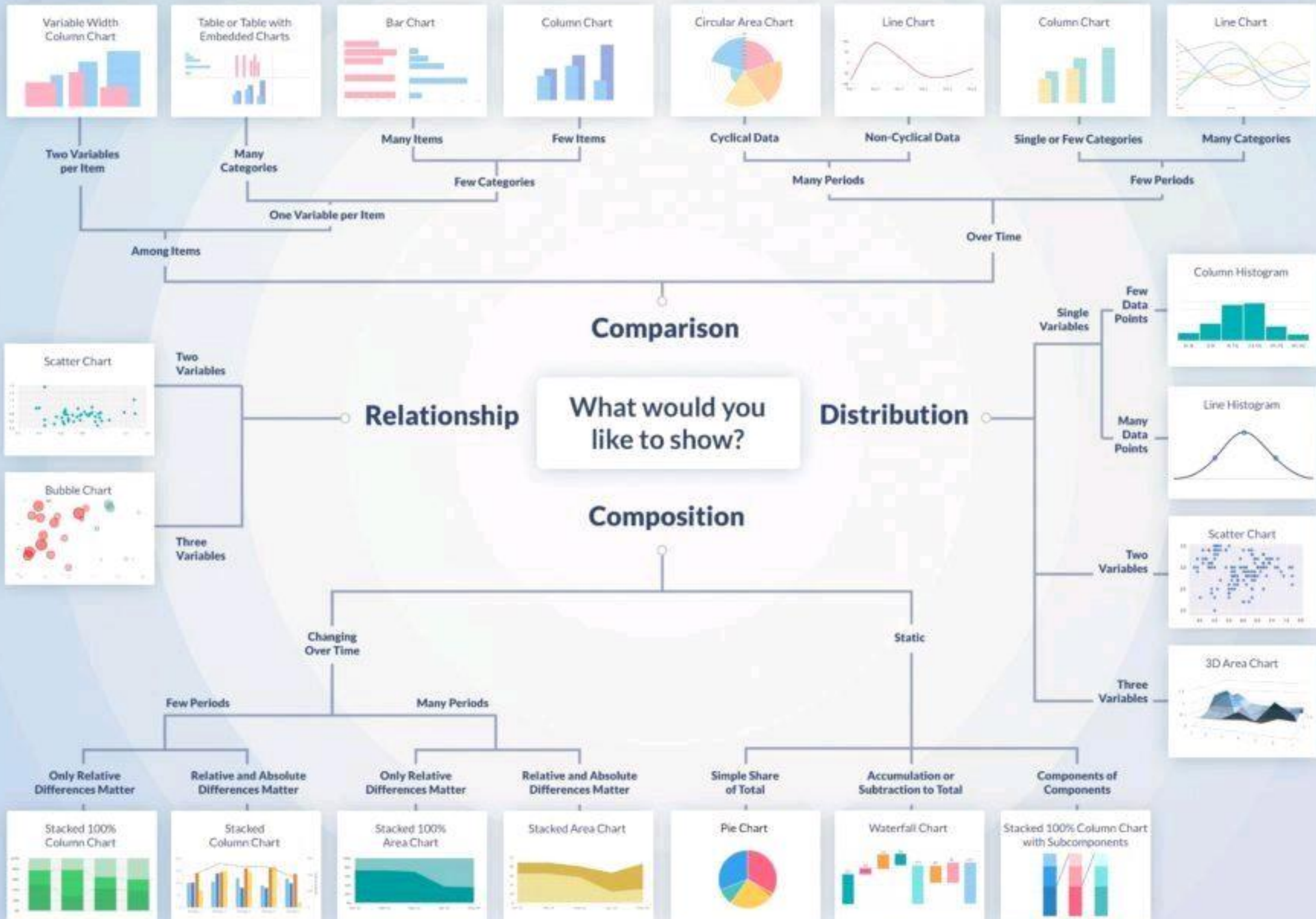


"Real-time" reporting

The reality is that it is a struggle to get the right information to the right place in the right form at the right time.


The FOUR V's of Big Data
Volume
SCALE OF DATA
40 ZETTABYTES
of data will be created by 2020, an increase of 300 times from 2005
It's estimated that
2.5 QUINTILLION BYTES
of data are created each day
6 BILLION PEOPLE
have cell phones
WORLD POPULATION: 7 BILLION
Most companies in the U.S. have at least
100 TERABYTES
of data stored
Velocity
ANALYSIS OF STREAMING DATA
The New York Stock Exchange captures
1 TB OF TRADE INFORMATION
during each trading session
Modern cars have close to
100 SENSORS
that monitor items such as fuel level and tire pressure
By 2016, it is projected there will be
18.9 BILLION NETWORK CONNECTIONS
- almost 2.5 connections per person on earth
By 2015
4.4 MILLION IT JOBS
Variety
DIFFERENT FORMS OF DATA
As of 2011, the global size of data in healthcare was estimated to be
150 EXABYTES
By 2014, it's anticipated there will be
420 MILLION WEARABLE, WIRELESS HEALTH MONITORS
4 BILLION+ HOURS OF VIDEO
are watched on YouTube each month
30 BILLION PIECES OF CONTENT
are shared on Facebook every month
400 MILLION TWEETS
are sent per day by about 200 million monthly active users
Veracity
UNCERTAINTY OF DATA
1 IN 3 BUSINESS LEADERS
don't trust the information they use to make decisions
27% OF RESPONDENTS
in one survey were unsure of how much of their data was inaccurate
Poor data quality costs the US economy around
$3.1 TRILLION A YEAR
IBM

Guided Visualizations for Charts and Graphs



Industry challenges to drive journey towards smart reporting

The six common impediments for transformation of reporting practice to a next generation operating model

*Gartner "the average financial impact of **poor data quality** and reporting on organizations is **$9.7M per year**"*

3 **Redundancies** due to lack of visibility on the effective usage of the reports

4 **Lack of compelling user experience** and limited actionable insights

2 **Disparate data sources,** a challenge for creation a golden source

5 **Inadequate data governance** mechanism, leading to poor data and reporting quality

1 **Legacy IT platforms** that are often non-synchronized

6 **Lack of futuristic analytics platform** to accommodate data driven decision making

Impact of Ineffective Reporting

What are the major consequences of an ineffective insights?

Delayed / incorrect strategic decisions made from bad data and reporting could potentially cost the business in millions

Data-rich but insight-poor reporting.
Missed opportunity in exploiting emerging growth areas, lost revenue due to slower responsiveness

Increased risk of SLA and non- compliance issues. The average cost for organizations that experience non-compliance problems is $14.82 million.

With high volume, velocity, and variety of data coming from multiple sources, many lack "single source of truth" for reporting to multiple stakeholders.

Why BI solutions are superior to traditional spreadsheets

	Traditional Spreadsheets	BI Solution
Huge Capacity	• Cumbersome to open large data files. Some programs do not even allow loading or viewing datasets bigger than a given number of rows or megabytes. • Preparation, modification, and transformation of data – manual, error prone	• Enable impressive data compression. • Easily combine, concatenate, and merge various datasets • Analytical and statistical models can be generated that include different dimensions and data attributes like date-time combinations.
Data Connectivity	• Do not allow directly connecting other tools like twitter, Facebook, or google analytics, often require third-party plugins	• Can combine data into a reportable model with connector functions dedicated to different data handlers and tools ex. Python, SQL • Instant visualizations and analyses that can be refreshed in real time.
Automated Reporting	• Manual data collection and refresh	• Able to create scheduled, strategic reports and/or deliver the updated information directly to your users, contractors, and clients
Visualizations	• Visualizations	• Intuitive UX features with self-service capabilities. Often offer drag-and-drop functionalities and other handy utilities to modify specific visuals • Complex visualizations involving geodata, maps, and timeframes

Why BI solutions are superior to traditional spreadsheets

	Traditional Spreadsheets	BI Solution
Mobility	• Difficult and ineffective to use on mobile devices	• Monitoring KPIs in mobile BI dashboards is a far better solution, especially in terms of viewing and visualizing spreadsheets. • Mobile versions often preserve most desktop functionalities around analysis and statistical interpretation. • Some BI solutions have dedicated mobile apps with access to downloadable dashboards for offline work
Collaboration, Distribution, and Publishing	• Files need to be manually shared or uploaded to a virtual drive • Risk of version conflicts, especially problematic when working with big data	• Allow you to publish reports that are accessible to your collaborators. They can comment on your KPIs and visualizations based on the context of each view. • Analytical models do not need to be rebuilt or reset; everything can be adjusted on the go.
Scalability and Performance	• Excel slows down dramatically when you reach approximately one million rows.	• Adjusted for millions of rows of data and run their own memory load optimization engines • Designed to store and calculate data either on desktops or in a cloud environment
Security and Monitoring	• Do not have their own security gateways	• Have many security layers controlling and monitoring users’ activity and access rights • Simple to implement security protocols like LDAP or active directory authentication.

Comparison of BI Tools

tableau

Market Leader

- Advanced interactive visual exploration which requires no programming experience
- Can connect to nearly any data-source
- Allows advanced statistical analysis and forecasting
- User friendly with online support readily available
- For a simple interface and beautiful visualizations

QlikView

Intermediary

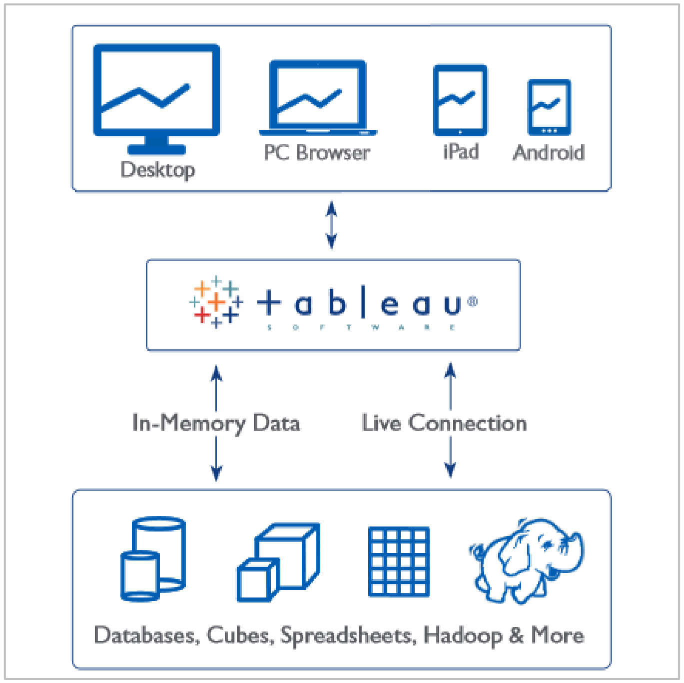
- Allow basic statistical analysis and forecasting
- More technical and scripting required (e.g.., map view) – SQL scripts
- Highly customizable charts
- Implementation speed is better
- Lead in terms of architecture, security and admin tasks
- For more extensive data transformation and analysis for the enterprise level

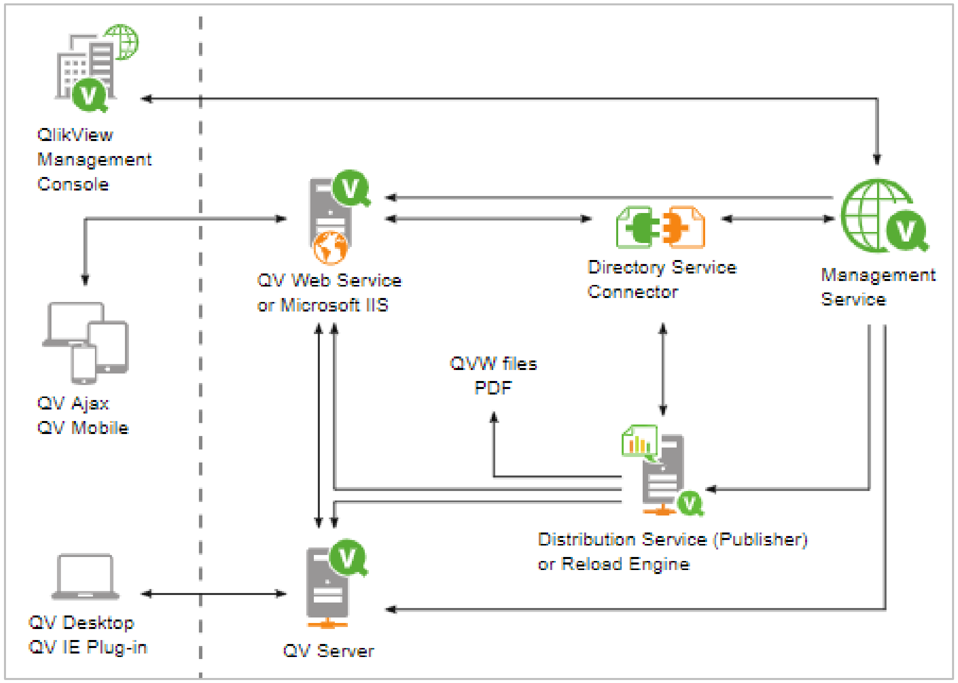
Microsoft Power BI

BI for the Masses

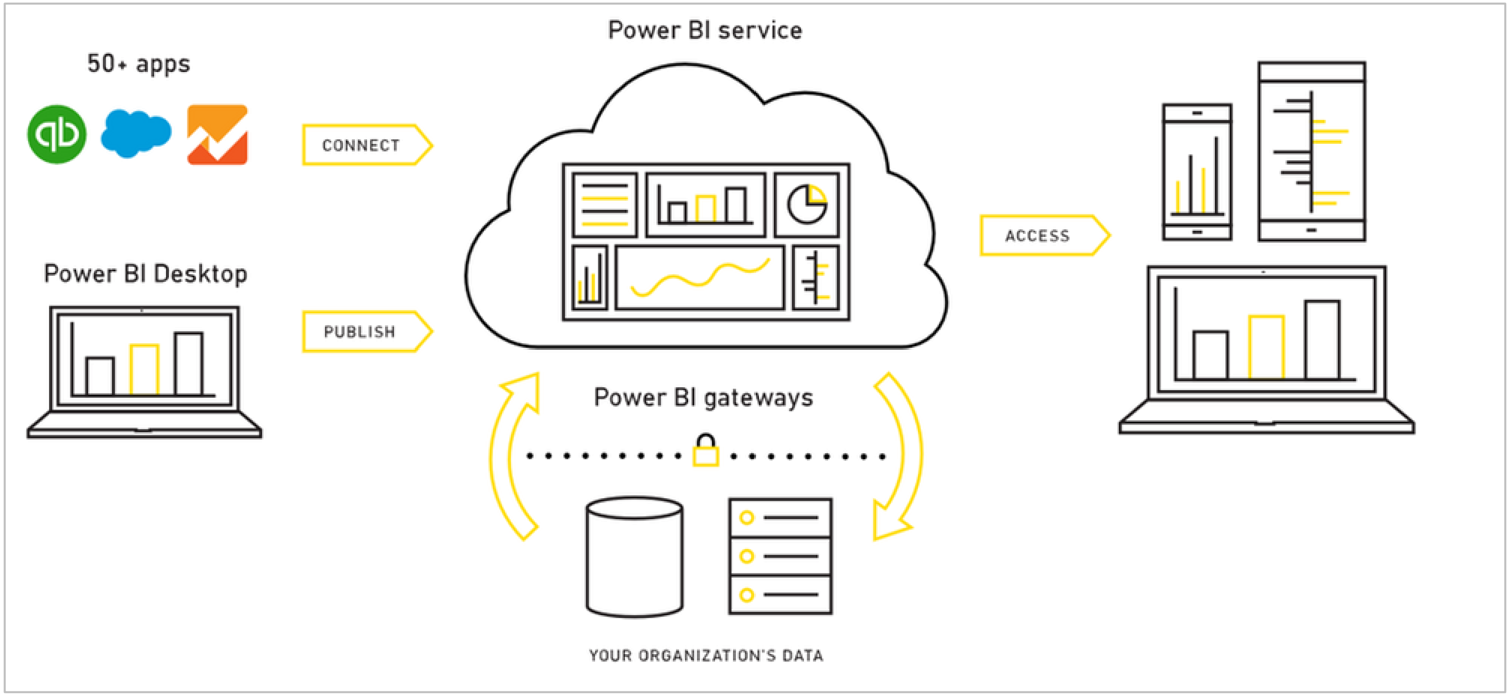
- Power BI Gallery provides ready to download apps/visualizations
- Uses DAX functions that are similar to Excel functions
- Cloud-based service
- Monthly Updates with bug-fixes or new releases
- Slower performance in comparison
- For quick insights and greater usability

Tableau Technical Architecture | QlikView Technical Architecture



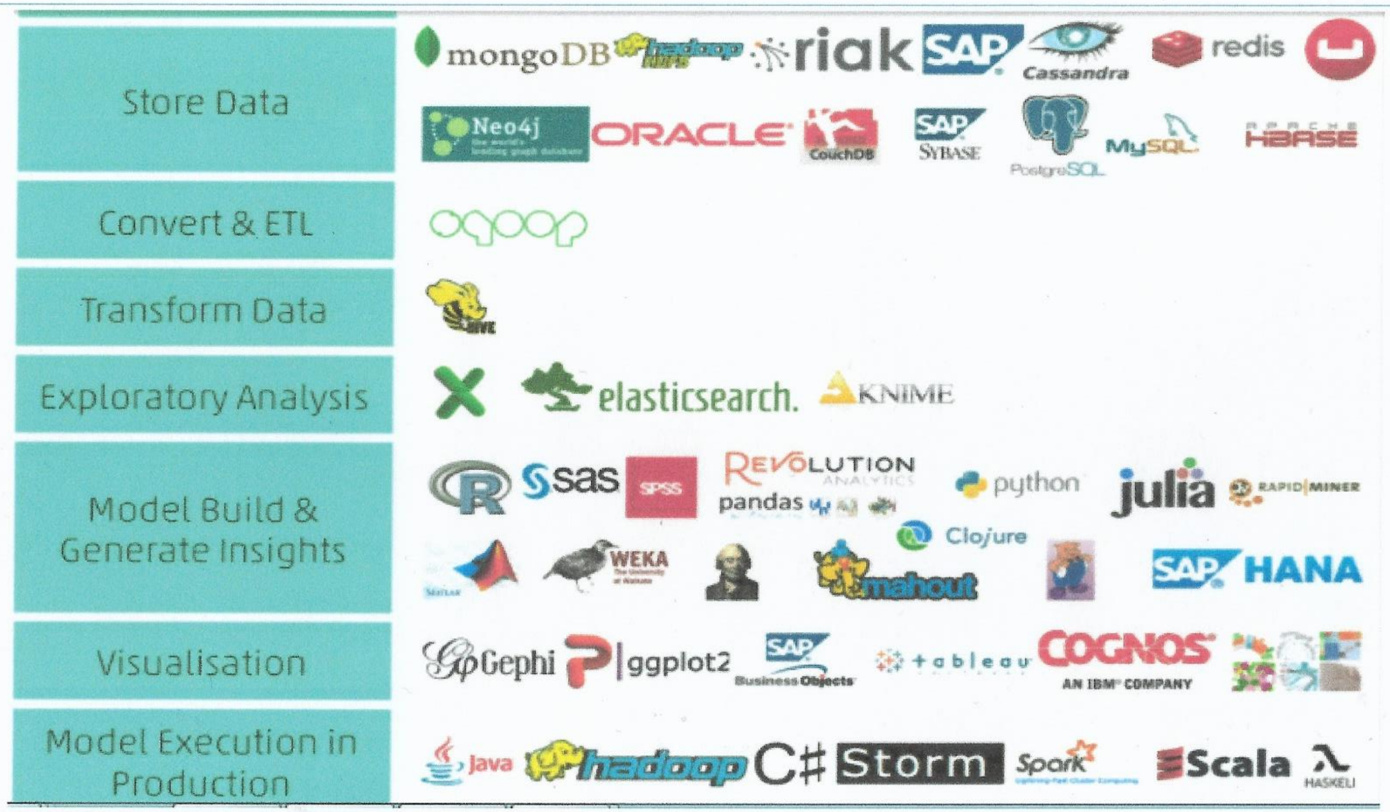


Power BI Technical Architecture



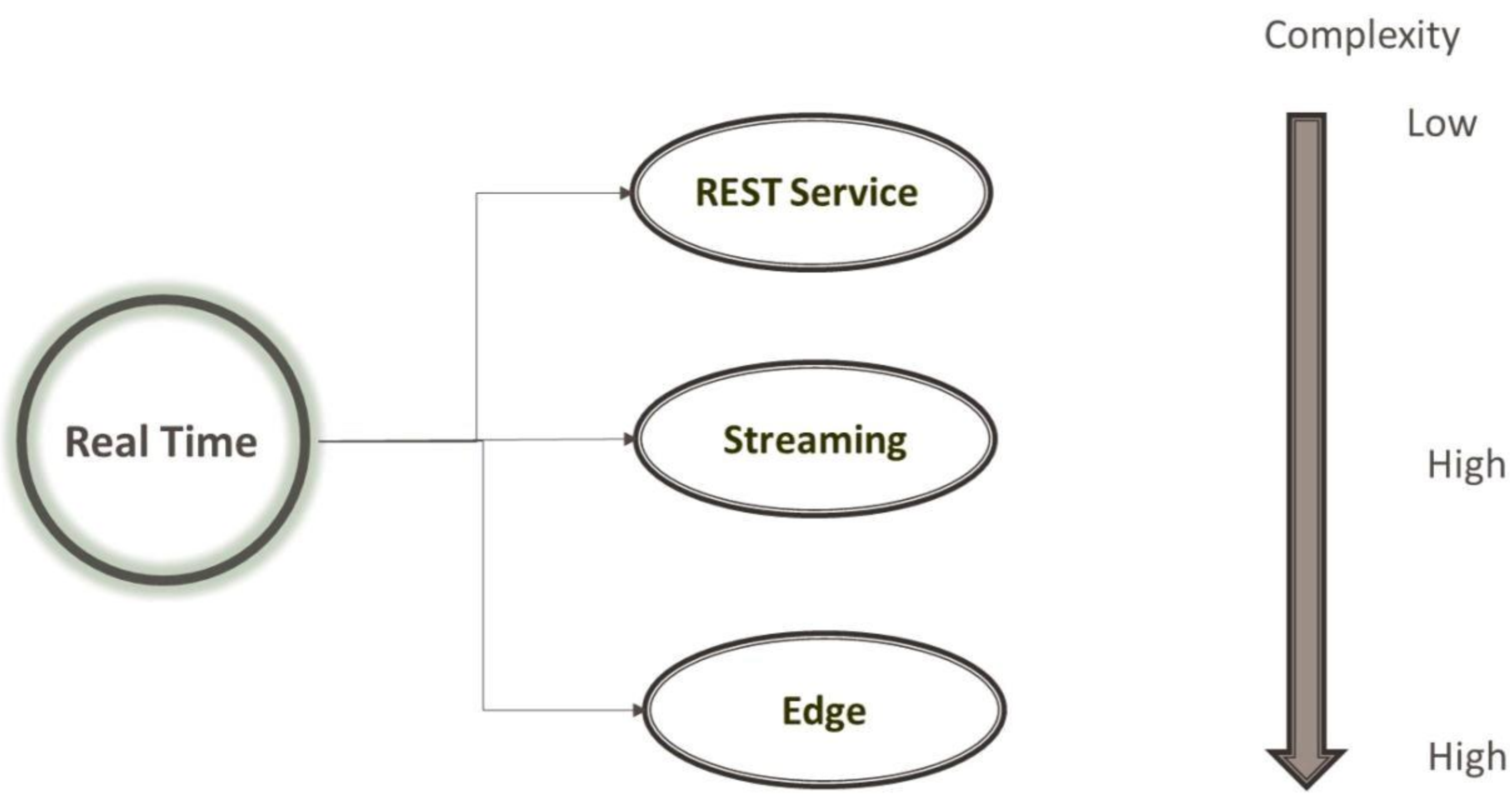
Analytics Ecosystem

Analytics Ecosystem

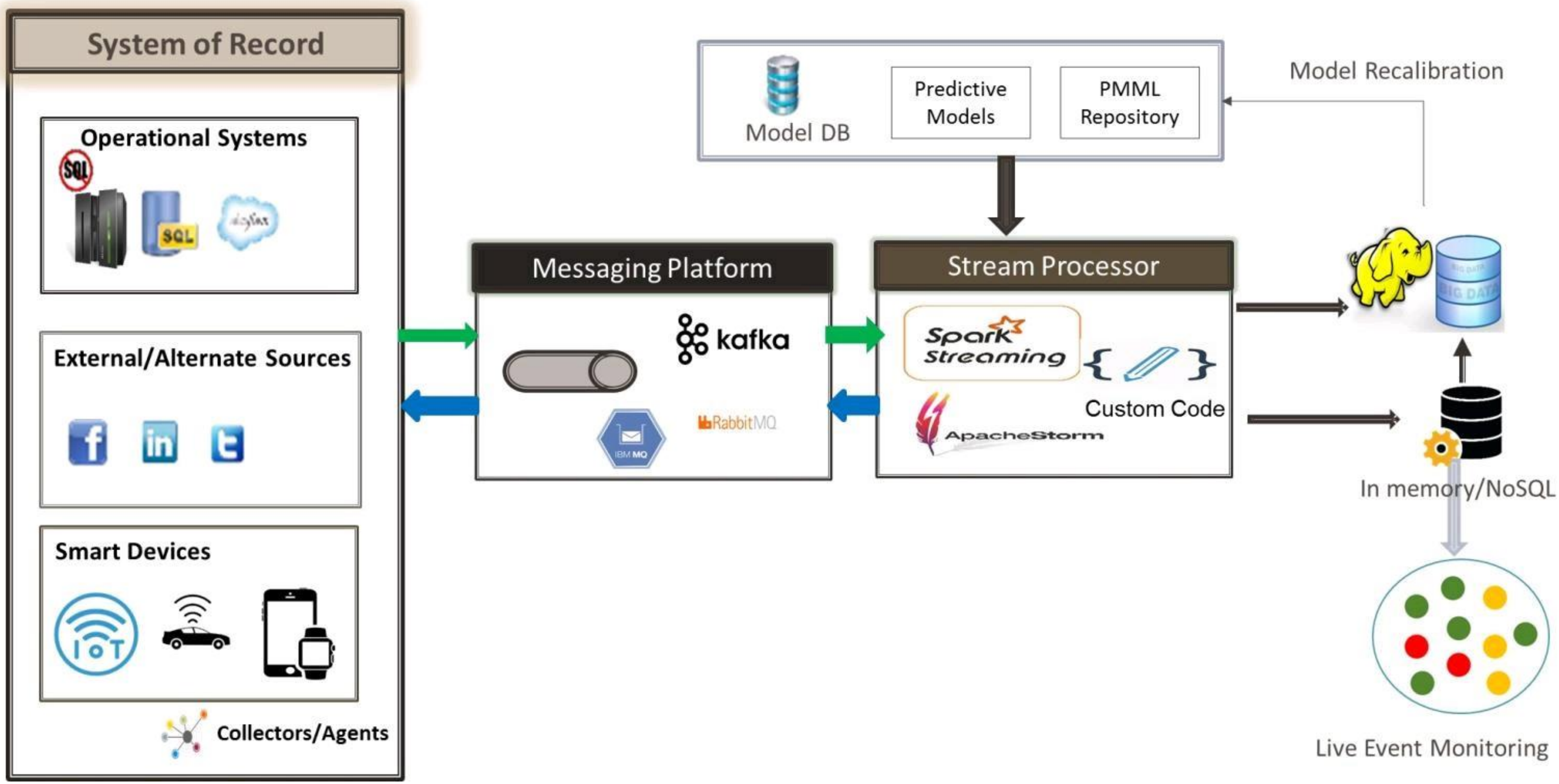


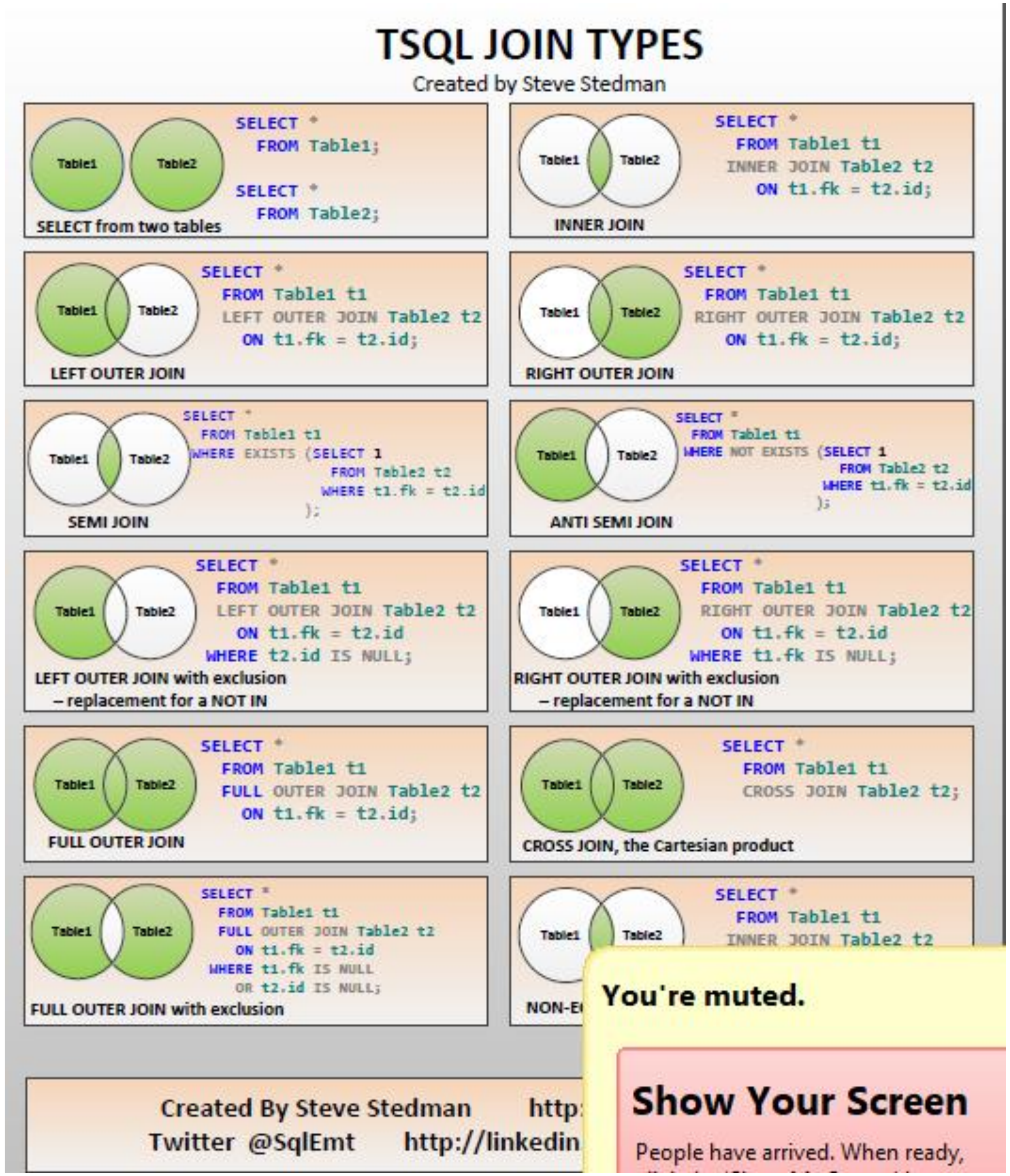
https://becomingadatascientist.wordpress.com/2013/07/26/choosing-a-data-science-technology-stack-w-survey/

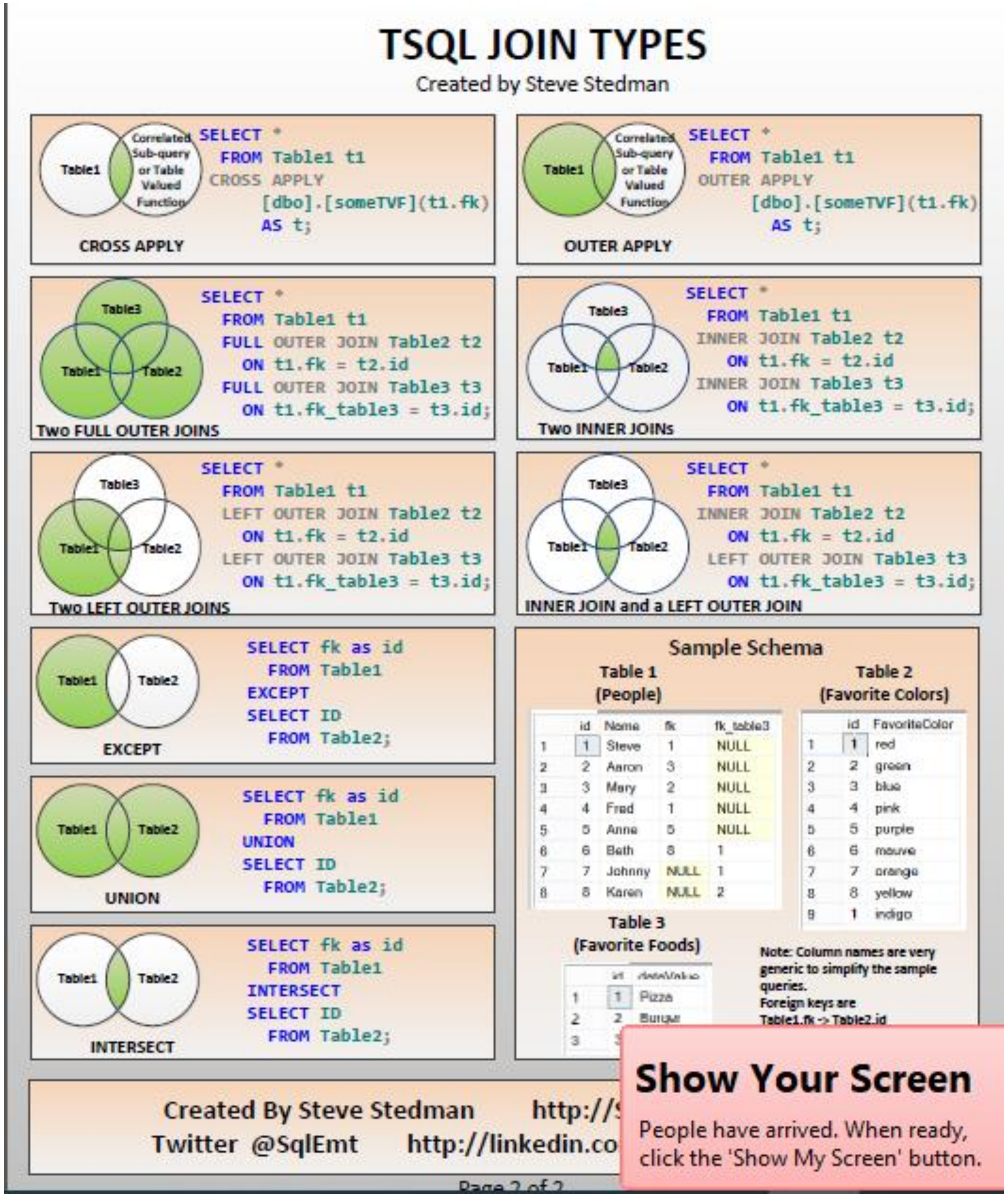
Real Time Deployment Patterns



Real Time Architecture




TSQL JOIN TYPES
Created by Steve Stedman
Table1
Table2
SELECT * FROM Table1;
SELECT * FROM Table2;
SELECT from two tables
SELECT * FROM Table1 t1 INNER JOIN Table2 t2 ON t1.fk = t2.id;
INNER JOIN
SELECT * FROM Table1 t1 LEFT OUTER JOIN Table2 t2 ON t1.fk = t2.id;
LEFT OUTER JOIN
SELECT * FROM Table1 t1 RIGHT OUTER JOIN Table2 t2 ON t1.fk = t2.id;
RIGHT OUTER JOIN
SELECT * FROM Table1 t1 WHERE EXISTS (SELECT 1 FROM Table2 t2 WHERE t1.fk = t2.id);
SEMI JOIN
SELECT * FROM Table1 t1 WHERE NOT EXISTS (SELECT 1 FROM Table2 t2 WHERE t1.fk = t2.id);
ANTI SEMI JOIN
SELECT * FROM Table1 t1 LEFT OUTER JOIN Table2 t2 ON t1.fk = t2.id WHERE t2.id IS NULL;
LEFT OUTER JOIN with exclusion – replacement for a NOT IN
SELECT * FROM Table1 t1 RIGHT OUTER JOIN Table2 t2 ON t1.fk = t2.id WHERE t1.fk IS NULL;
RIGHT OUTER JOIN with exclusion – replacement for a NOT IN
SELECT * FROM Table1 t1 FULL OUTER JOIN Table2 t2 ON t1.fk = t2.id;
FULL OUTER JOIN
SELECT * FROM Table1 t1 CROSS JOIN Table2 t2;
CROSS JOIN, the Cartesian product
SELECT * FROM Table1 t1 FULL OUTER JOIN Table2 t2 ON t1.fk = t2.id WHERE t1.fk IS NULL OR t2.id IS NULL;
FULL OUTER JOIN with exclusion
SELECT * FROM Table1 t1 INNER JOIN Table2 t2
NON-E
You're muted.
Show Your Screen
People have arrived. When ready,
Created By Steve Stedman
http:
Twitter @SqlEmt
http://linkedin



TSQL JOIN TYPES

Created by Steve Stedman


```
SELECT *
  FROM Table1 t1
 CROSS APPLY
      [dbo].[someTVF](t1.fk)
        AS t;
```



```
SELECT *
  FROM Table1 t1
 OUTER APPLY
      [dbo].[someTVF](t1.fk)
        AS t;
```



```
SELECT *
  FROM Table1 t1
  FULL OUTER JOIN Table2 t2
    ON t1.fk = t2.id
  FULL OUTER JOIN Table3 t3
    ON t1.fk_table3 = t3.id;
```



```
SELECT *
  FROM Table1 t1
 INNER JOIN Table2 t2
    ON t1.fk = t2.id
 INNER JOIN Table3 t3
    ON t1.fk_table3 = t3.id;
```



```
SELECT *
  FROM Table1 t1
  LEFT OUTER JOIN Table2 t2
    ON t1.fk = t2.id
  LEFT OUTER JOIN Table3 t3
    ON t1.fk_table3 = t3.id;
```



```
SELECT *
  FROM Table1 t1
 INNER JOIN Table2 t2
    ON t1.fk = t2.id
  LEFT OUTER JOIN Table3 t3
    ON t1.fk_table3 = t3.id;
```



```
SELECT fk as id
  FROM Table1
EXCEPT
SELECT ID
  FROM Table2;
```



```
SELECT fk as id
  FROM Table1
UNION
SELECT ID
  FROM Table2;
```



```
SELECT fk as id
  FROM Table1
INTERSECT
SELECT ID
  FROM Table2;
```

Sample Schema

Table 1 (People)

	id	Name	fk	fk_table3
1	1	Steve	1	NULL
2	2	Aaron	3	NULL
3	3	Mary	2	NULL
4	4	Fred	1	NULL
5	5	Anne	5	NULL
6	6	Beth	8	1
7	7	Johnny	NULL	1
8	8	Karen	NULL	2

Table 2 (Favorite Colors)

	id	FavoriteColor
1	1	red
2	2	green
3	3	blue
4	4	pink
5	5	purple
6	6	mauve
7	7	orange
8	8	yellow
9	1	indigo

Table 3 (Favorite Foods)

	id	[illegible]
1	1	Pizza
2	2	Burger
3	[illegible]	[illegible]

Note: Column names are very generic to simplify the sample queries.
Foreign keys are
Table1.fk -> Table2.id

Created By Steve Stedman http://S
Twitter @SqlEmt http://linkedin.co

Show Your Screen

People have arrived. When ready, click the 'Show My Screen' button.

Analytics Approaches: A Taxonomy

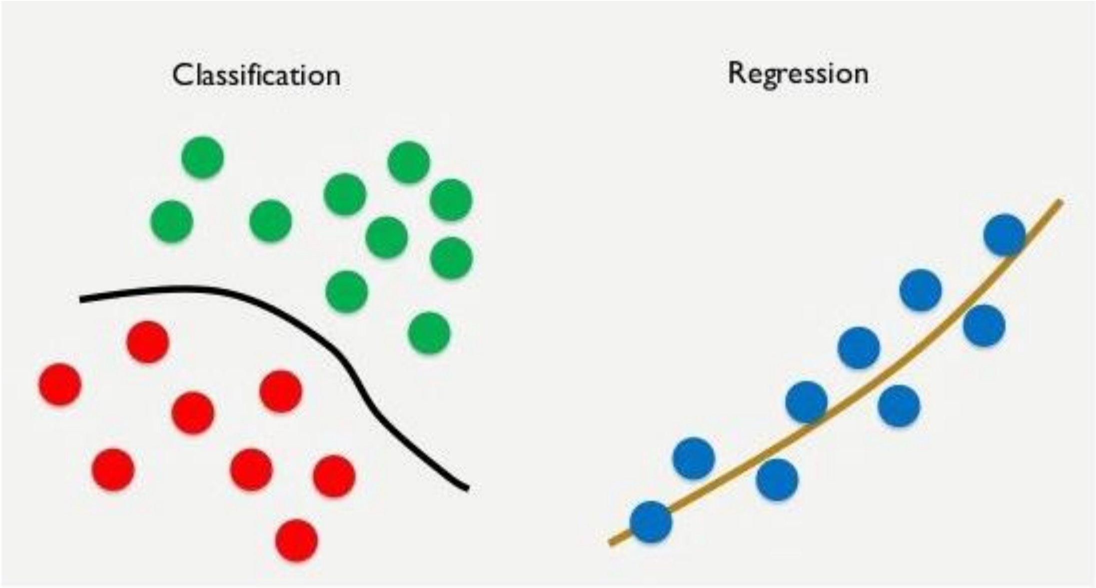
Supervised learning:

Supervised learning means where you will teach or train the machine using labeled data. Labeled data means where the answer is already known. So here our system learn by predicting the value. Then it does an accuracy check by using a cost function, to check how close the prediction was to actual output.

Suppose you have provided a data set consisting of bikes and cars. Now
You need to train the machine on how to classify all the different images. You can train it like this:
If there is 2 number of wheels and 1 headlight on the front it will be labeled as a bike.
If there is 4 number of wheels and 2 headlights on the front it will be labeled as a car.

Since your machine has already learned the things, it needs to use that knowledge. The machine will classify the Image regarding the presence or absence of a number of wheels and number of Headlights and would label the image name as Bike.

Supervised learning is typically done in the context of Classification and Regression.



Analytics Approaches: A Taxonomy

Unsupervised learning

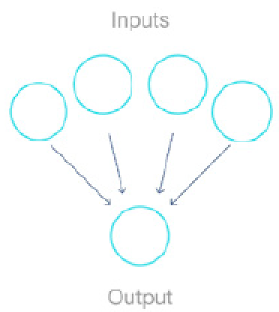
Unlike supervised learning, In this, the result is not known, we approach with little or No knowledge of what the result would be, the machine is expected to find the hidden patterns and structure in un-labelled data on their own.

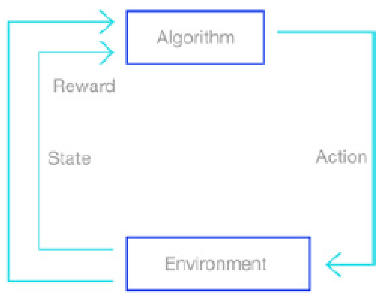
- ❑ Clustering means segregating or dividing a data set into a number of groups such that data set in the same groups are more similar than those in other groups. In simple words, the aim is to separate groups with similar traits and assign them into clusters.
- ❑ Association is about discovering some interesting relationships between variables in large databases. For example, people that buy a new house also tend to buy new furniture. It discovers the probability of the co-occurrence of items in a collection.

- ❑ Clustering is about grouping data points according to their similarities while Association is about discovering some relationships between the attributes of those data points.

https://sebastianraschka.com/Articles/2014_intro_supervised_learning.html
https://chatbotsmagazine.com/lets-know-supervised-and-unsupervised-in-an-easy-way-9168363e06ab

Major Types of Machine Learning





	Supervised learning	Unsupervised learning	Reinforcement learning
Definition	Uses training data and feedback from humans to learn the relationship of given inputs to a given output	Explores input data without being given an explicit output variable	Learns to perform a task simply by trying to maximize rewards it receives for its actions
Example	Ex. How the inputs "time of year" and "interest rates" predict housing prices	Explores customer demographic data to identify patterns	Maximizes points it receives for increasing returns of an investment portfolio
When to use	You know how to classify the input data and the type of behavior you want to predict, but you need the algorithm to calculate it for you on new data	You do not know how to classify the data, and you want the algorithm to find patterns and classify the data for you	Not much training data is available, the ideal end state cannot be clearly defined ; or the only way to learn about the environment is to interact with it

Supervised Learning – Algorithms and Sample Use Cases

Algorithms		Sample Business Use Cases
Linear Regression	Highly interpretable, standard method for modeling the past relationship between independent input variables and dependent output variables (which can have an infinite number of values) to help predict future values of the output variables	• Understand product-sales drivers such as competition prices, distribution, advertisement, etc.. • Optimize price points and estimate product-price elasticities
Logistic Regression	Extension of linear regression that's used for classification tasks, meaning the output variable is binary (e.g., only black or white) rather than continuous (e.g., an infinite list of potential colors)	• Classify customers based on how likely they are to repay a loan • Predict if a skin lesion is benign or malignant based on its characteristics (size, shape, color, etc.)
Linear/Quadratic Discriminant Analysis	Upgrades a logistic regression to deal with nonlinear problems—those in which changes to the value of input variables do not result in proportional changes to the output variables	• Predict client churn • Predict a sales lead's likelihood of closing Algorithms Sample business use cases
Decision Tree	Highly interpretable classification or regression model that splits data-feature values into branches at decision nodes (e.g., if a feature is a color, each possible color becomes a new branch) until a final decision output is made	• Provide a decision framework for hiring new employees • Understand product attributes that make a product most likely to be purchased

Supervised Learning – Algorithms and Sample Use Cases

Algorithms		Sample Business Use Cases
Naive Bayes	Classification technique that applies Bayes theorem, which allows the probability of an event to be calculated based on knowledge of factors that might affect that event (e.g., email containing the word "money," has high probability of being spam)	• Analyze sentiment to assess product perception in the market • Create classifiers to filter spam emails Supervised learning: Algorithms and sample business use cases
Support Vector Machine	A technique that's typically used for classification but can be transformed to perform regression. It draws an optimal division between classes (as wide as possible). It also can be quickly generalized to solve nonlinear problems	• Predict how many patients a hospital will need to serve in a time period • Predict how likely someone is to click on an online ad
Random Forest	Classification or regression model that improves the accuracy of a simple decision tree by generating multiple decision trees and taking a majority vote of them to predict the output, which is a continuous variable (e.g., age) for a regression problem and a discrete variable (e.g., either black, white, or red) for classification	• Predict call volume in call centers for staffing decisions • Predict power usage in an electrical distribution grid
Adaboost	Classification or regression technique that uses a multitude of models to come up with a decision but weighs them based on their accuracy in predicting the outcome	• Detect fraudulent activity in credit-card transactions. Achieves lower accuracy than deep learning • Simple, low-cost way to classify images (e.g., recognize land usage from satellite images for climate-change models). Achieves lower accuracy than deep learning

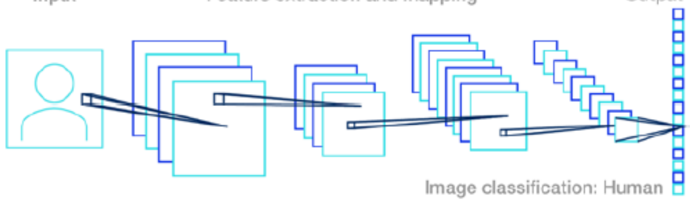
Supervised Learning – Algorithms and Sample Use Cases

Algorithms		Sample Business Use Cases
Gradient-boosting Trees	Classification or regression technique that generates decision trees sequentially, where each tree focuses on correcting the errors coming from the previous tree model. The final output is a combination of the results from all trees	• Forecast product demand and inventory levels • Predict the price of cars based on their characteristics (e.g., age and mileage)
Simple Neural Network	Model in which artificial neurons (software-based calculators) make up three layers (an input layer, a hidden layer where calculations take place, and an output layer) that can be used to classify data or find the relationship between variables in regression problems	• Predict the probability that a patient joins a healthcare program

Unsupervised Learning – Algorithms and Sample Use Cases

Algorithms		Sample Business Use Cases
K-means clustering	Puts data into a number of groups (k) that each contain data with similar characteristics (as determined by the model, not in advance by humans)	• Segment customers into groups by distinct characteristics (e.g., age group)— for instance, to better assign marketing campaigns or prevent churn
Gaussian mixture model	A generalization of k-means clustering that provides more flexibility in the size and shape of groups (clusters)	• Segment customers to better assign marketing campaigns using less-distinct customer characteristics (e.g., product preferences) • Segment employees based on likelihood of attrition
Hierarchical clustering	Splits or aggregates clusters along a hierarchical tree to form a classification system	• Cluster loyalty-card customers into progressively more micro segmented groups • Inform product usage/development by grouping customers mentioning keywords in social-media data
Recommender system	Often uses cluster behavior prediction to identify the important data necessary for making a recommendation .	• Recommend what movies consumers should view based on preferences of other customers with similar attributes • Recommend news articles a reader might want to read based on the article she or he is reading

Major Types of Deep Learning

	Convolutional neural network	Recurrent neural network
		
Definition	A multilayered neural network with a special architecture designed to extract increasingly complex features of the data at each layer to determine the output	A multilayered neural network that can store information in context nodes, allowing it to learn data sequences and output a number or another sequence
When to use	When you have an unstructured data set (e.g., images) and you need to infer information from it	When you are working with time-series data or sequences (e.g., audio recordings or text)

DATA SCIENCE
Main Formulas for Machine Learning

Naïve Bayes

$$P(a|c) = \frac{P(c|a).P(a)}{P(c)}$$

$$Prob = \Pi P(a|c)$$

K Nearest Neighbor

$$D(x_i, x_j) = \sqrt{(x_i - x_j)^2 + (y_i - y_j)^2}$$

Support Vector Machines

$$f(x) = \text{sign}[\lambda.y.K(x_i \cdot x_j)]$$

$$K(x_i \cdot x_j) = \sqrt{\frac{\sqrt{(x_i - x_j)^2 + (y_i - y_j)^2}}{width}}$$

$$\lambda \rightarrow \nabla L = 0$$

$$y = 1 \wedge y = -1$$

Perceptron

$$f(x) = sign\left[\sum_{i=1}^{n} w_j x_{ij}\right]$$

Neural Networks

$$f(x) = w_0 + K.\sum_{i=1}^{n} w_i x_i$$

Backpropagation

$$\Delta w_{ij}(n) = \eta \delta j x_{ij} + \alpha \Delta w_{ij}(n-1)$$

Gradient Descent

$$\theta_{ji} = \theta_j - \alpha \sum_{i=1}^{n} (h(x_i) - y).x_i$$

Linear Regression

$$f(x) = \sum_{i=1}^{n} m_i x_i + b$$

Principal Components Analysis

$$x_j = x_i - \bar{x}$$

$$Eingenvector = Engeinvalue.[x_i \ldots x_n]$$

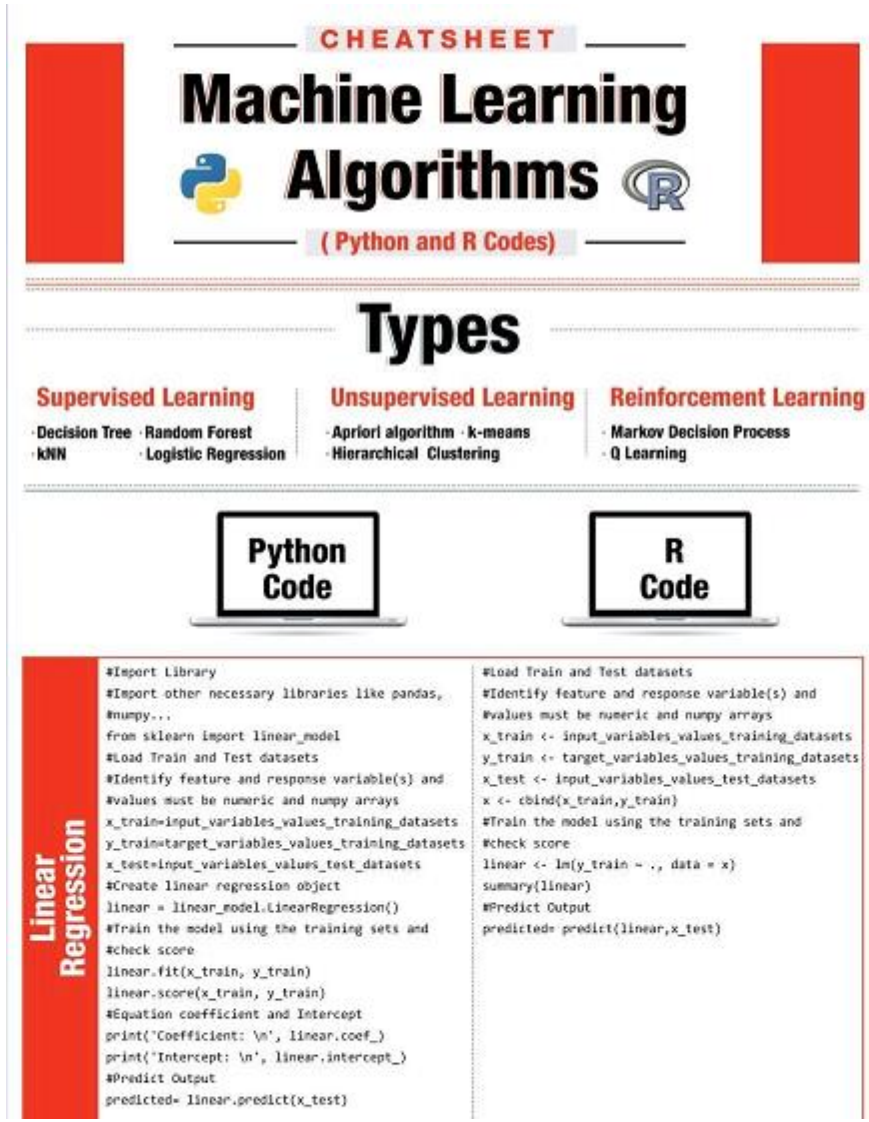
$$f(x) = Eigenvector^T.[x_{j1} \ldots x_{jn}]$$

Logistic Regression

$$Odds\ Ratio = log\left(\frac{P(a|c)}{1 - P(a|c)}\right)$$

$$Prob(y = 1) = \frac{1}{1 + e^{-\theta(\sum_{i=1}^{n} m_i x_i + b)}}$$

Rubens Zimbres



CHEATSHEET

Machine Learning Algorithms

(Python and R Codes)

Types

Supervised Learning

- Decision Tree
- Random Forest
- kNN
- Logistic Regression

Unsupervised Learning

- Apriori algorithm
- k-means
- Hierarchical Clustering

Reinforcement Learning

- Markov Decision Process
- Q Learning

Python Code | **R Code**

Linear Regression

Python Code:

```python
#Import Library
#Import other necessary libraries like pandas,
#numpy...
from sklearn import linear_model
#Load Train and Test datasets
#Identify feature and response variable(s) and
#values must be numeric and numpy arrays
x_train=input_variables_values_training_datasets
y_train=target_variables_values_training_datasets
x_test=input_variables_values_test_datasets
#Create linear regression object
linear = linear_model.LinearRegression()
#Train the model using the training sets and
#check score
linear.fit(x_train, y_train)
linear.score(x_train, y_train)
#Equation coefficient and Intercept
print('Coefficient: \n', linear.coef_)
print('Intercept: \n', linear.intercept_)
#Predict Output
predicted= linear.predict(x_test)
```

R Code:

```r
#Load Train and Test datasets
#Identify feature and response variable(s) and
#values must be numeric and numpy arrays
x_train <- input_variables_values_training_datasets
y_train <- target_variables_values_training_datasets
x_test <- input_variables_values_test_datasets
x <- cbind(x_train,y_train)
#Train the model using the training sets and
#check score
linear <- lm(y_train ~ ., data = x)
summary(linear)
#Predict Output
predicted= predict(linear,x_test)
```

Logistic Regression

```python
#Import Library
from sklearn.linear_model import LogisticRegression
#Assumed you have, X (predictor) and Y (target)
#for training data set and x_test(predictor)
#of test_dataset
#Create logistic regression object
model = LogisticRegression()
#Train the model using the training sets
#and check score
model.fit(X, y)
model.score(X, y)
#Equation coefficient and Intercept
print('Coefficient: \n', model.coef_)
print('Intercept: \n', model.intercept_)
#Predict Output
predicted= model.predict(x_test)
```

```r
x <- cbind(x_train,y_train)
#Train the model using the training sets and check
#score
logistic <- glm(y_train ~ ., data = x,family='binomial')
summary(logistic)
#Predict Output
predicted= predict(logistic,x_test)
```

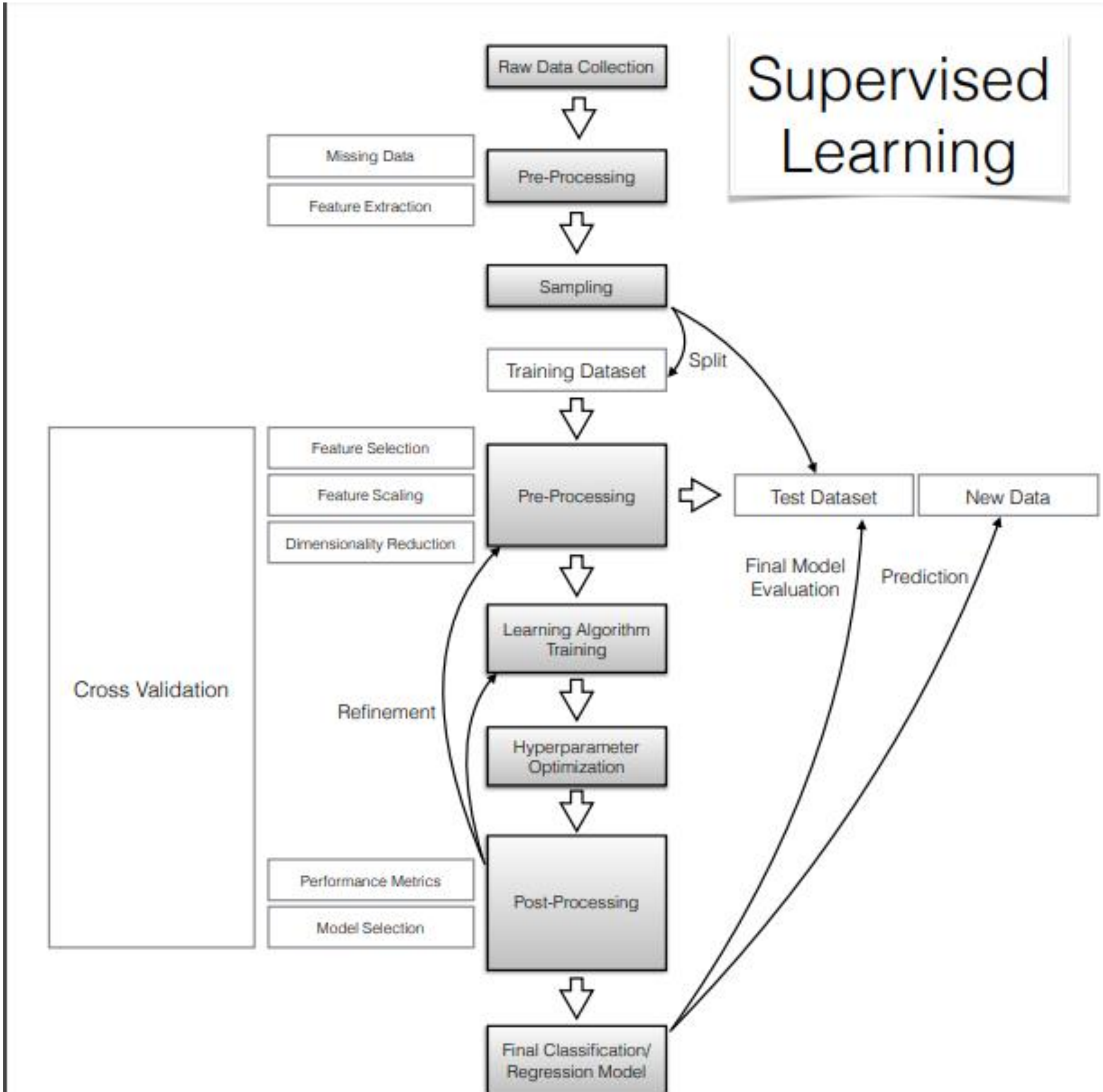
Method	Uses	Concerns
Decision Trees	• Trees handle outliers and missing observations well. • High Interpretability. • Interactions considered automatically, but implicitly. • Ensemble trees e.g. random forest and gradient boosting have the ability to increase prediction accuracy and decreases overfitting to some extent.	• Overfitting. • Unstable with a small dataset. • Unstable with outliers and noisy data. • Careful parameter tuning required.
Penalized Regression	• Supervised regression or classification. • Modeling linear or non- linear occurrence by specifying interactions terms. • Parsimonious model. • When interpretability is important.	• Standardization needed. • Careful parameter tuning required. • Treat missing and outliers values prior to algorithm implementation.
K-Mean	• Unsupervised clustering. • Finding similar observations to form the groups in a dataset without labels.	• Standardization needed. • Finding an optimal number of *K*. • Sensitive to missing values and outliers.
Hierarchical Clustering	• Unsupervised clustering. • Create a known number of overlapping clusters of different sizes.	• Standardization needed. • An optimal number of clusters. • Sensitive to missing values and outliers. • Curse of dimensionality.
Support Vector Machines (SVM)	• Modeling linear and non- linear occurrence by using linear and non-linear kernels. • Capture much more complex relationships between observations.	• Low interpretability. • Computationally intensive. • Missing values and outliers. • Standardization needed. • Parameter tuning.
Neural Networks	• Pattern recognition in images, videos etc. • Unsupervised feature extraction. • Supervised regression or classification. • Anomaly detection with autoencoder networks.	• Low interpretability. • Missing values and outliers. • Standardization needed. • Parameter tuning. • Computationally intensive.

Analytics Approaches: A Taxonomy

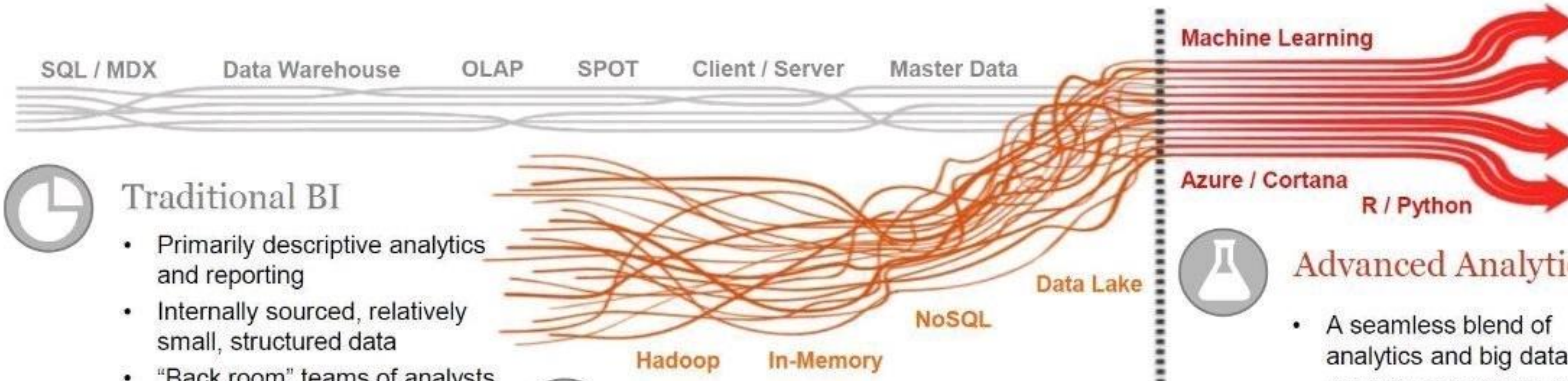
Supervised Learning Algorithms	**Learning Task**
Decision Trees	Classification
K Nearest Neighbours	Classification
Logistic Regression	Classification
Naive Bayes	Classification
Linear Regression	Numerical Prediction
Regression Tree	Numerical Prediction
Model Trees	Numerical Prediction
Support Vector Machines	Classification & Numerical Prediction
Neural Networks	Classification & Numerical Prediction
Unsupervised Learning Algorithms	
Association Rules	Pattern Detection
K-means Clustering	Clustering
Meta-Learning Algorithms	
Bagging	Classification & Numerical Prediction
Boosting	Classification & Numerical Prediction
Random Forests	Classification & Numerical Prediction

https://sebastianraschka.com/Articles/2014_intro_supervised_learning.html
https://chatbotsmagazine.com/lets-know-supervised-and-unsupervised-in-an-easy-way-9168363e06ab

Analytics Approaches: A Taxonomy



The field of Data & Analytics is evolving fast, driven by the large amount o data available as well as technological quantum leaps



Traditional BI

- Primarily descriptive analytics and reporting
- Internally sourced, relatively small, structured data
- "Back room" teams of analysts
- Mainly informational / processual value
- On premise systems

Big Data

- Complex, large, unstructured data sources
- Data streams and "real time" processing
- New analytical and computational capabilities

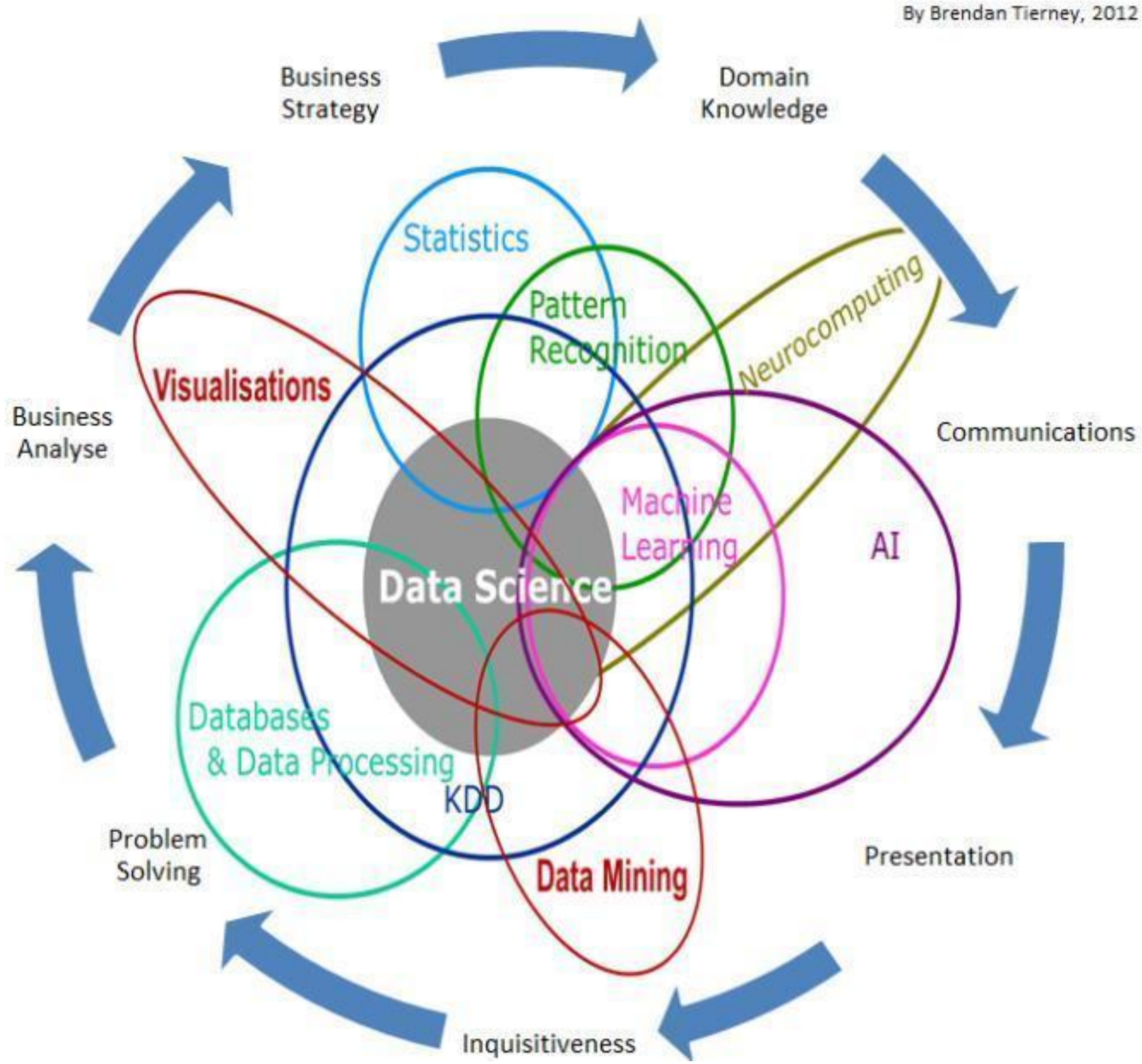
Advanced Analytics

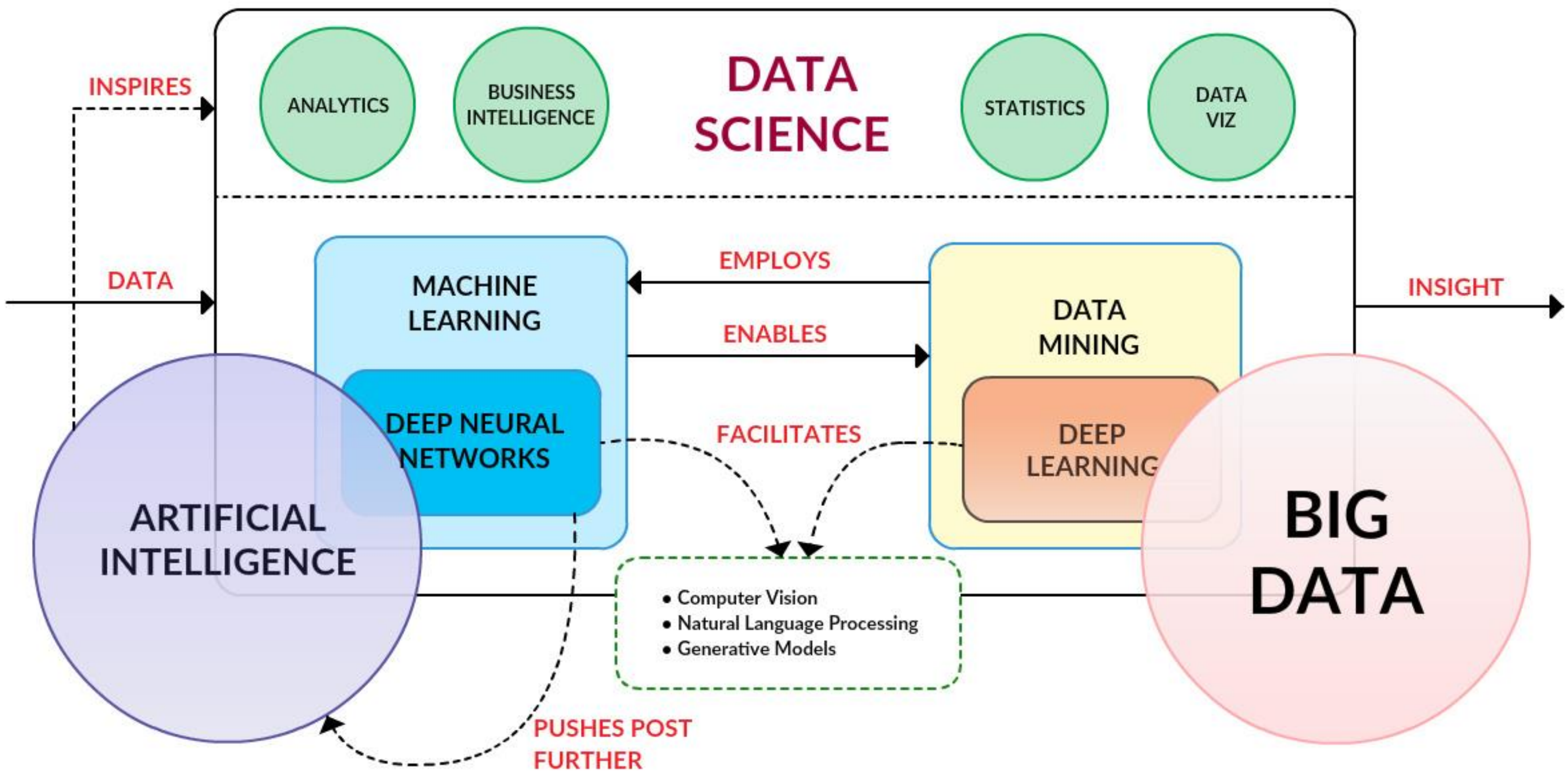
- A seamless blend of analytics and big data
- Analytics integral to running the business, strategic asse
- Rapid and agile insight delivery
- Analytical tools available at point of decision
- Cultural evolution embeds analytics into decision and operational processes

source pwc via @mikequin

JOB ROLES IN ANALYTICS

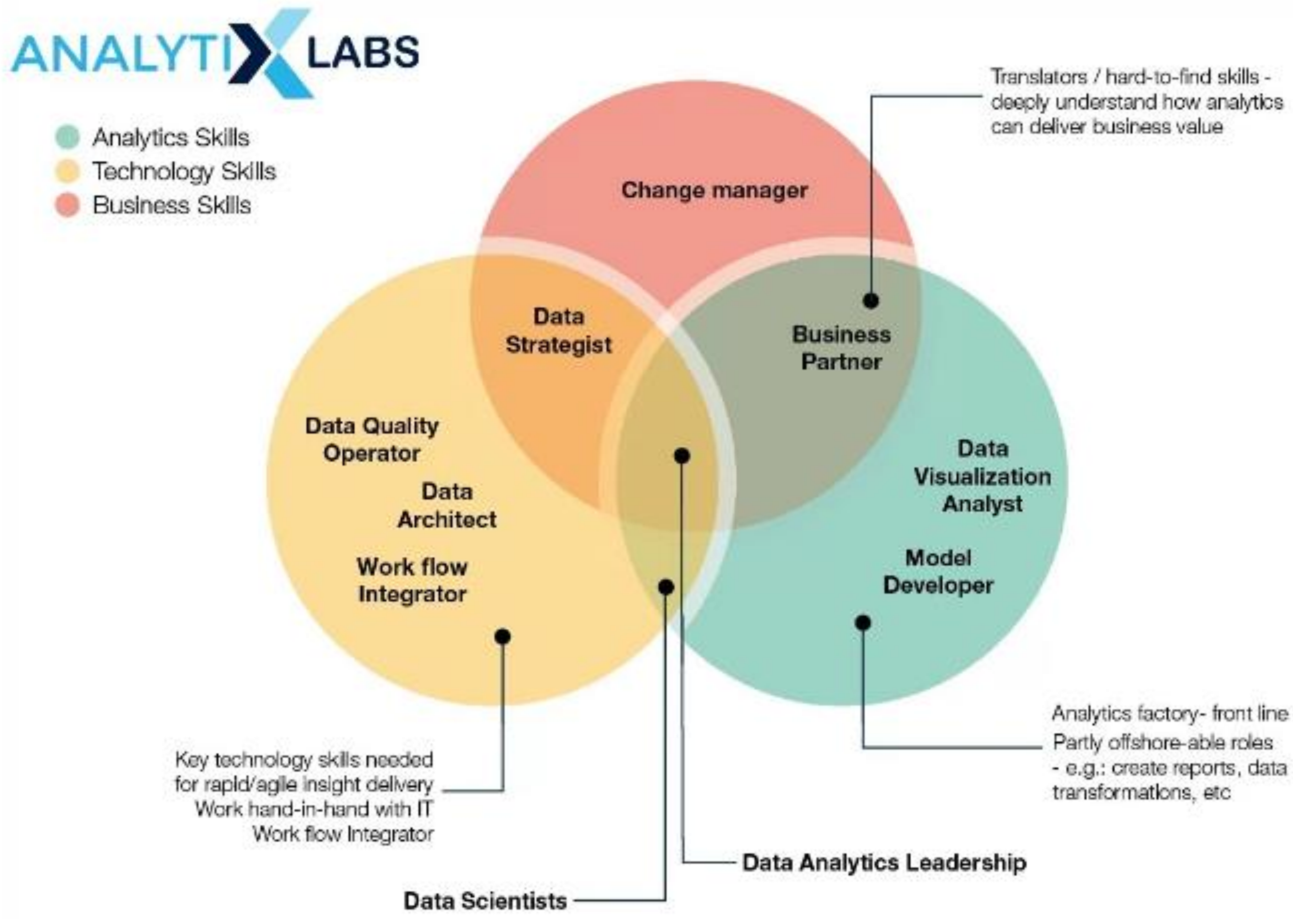
Data Science Is Multidisciplinary



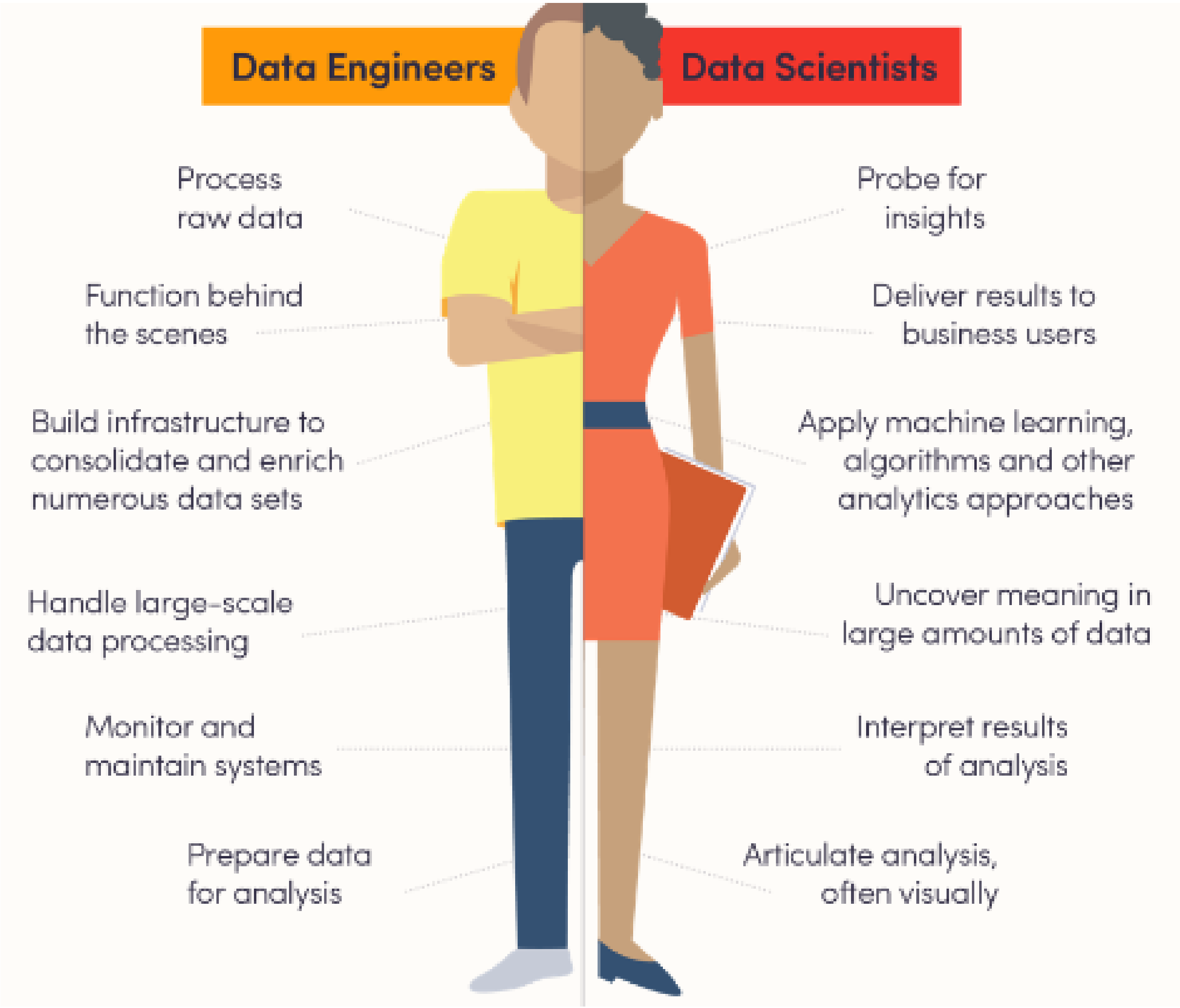


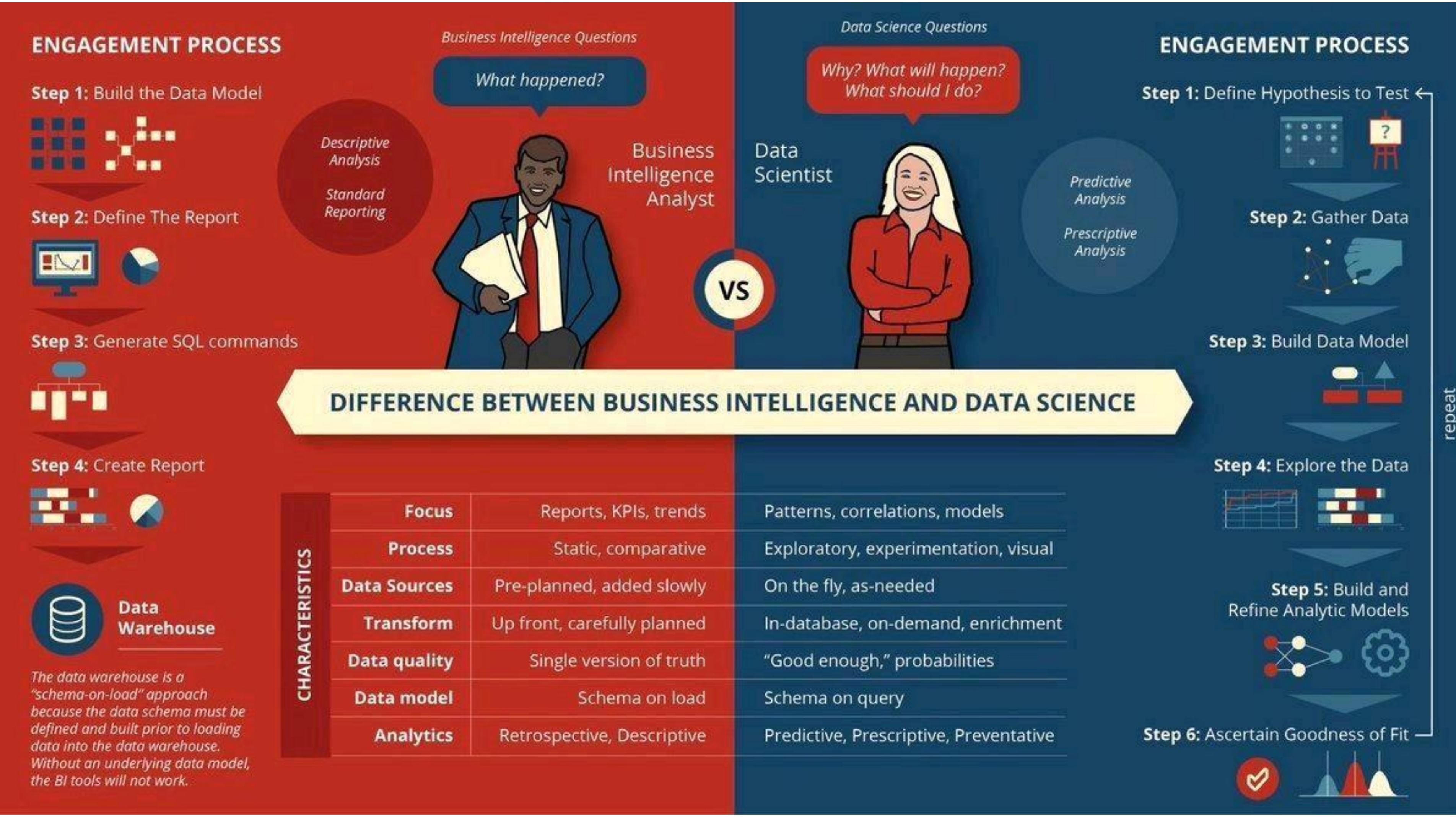
DATA SCIENCE
ANALYTICS
BUSINESS INTELLIGENCE
STATISTICS
DATA VIZ
INSPIRES
DATA
INSIGHT
MACHINE LEARNING
DEEP NEURAL NETWORKS
EMPLOYS
ENABLES
FACILITATES
DATA MINING
DEEP LEARNING
ARTIFICIAL INTELLIGENCE
BIG DATA
• Computer Vision
• Natural Language Processing
• Generative Models
PUSHES POST FURTHER

Analytics : Job Roles

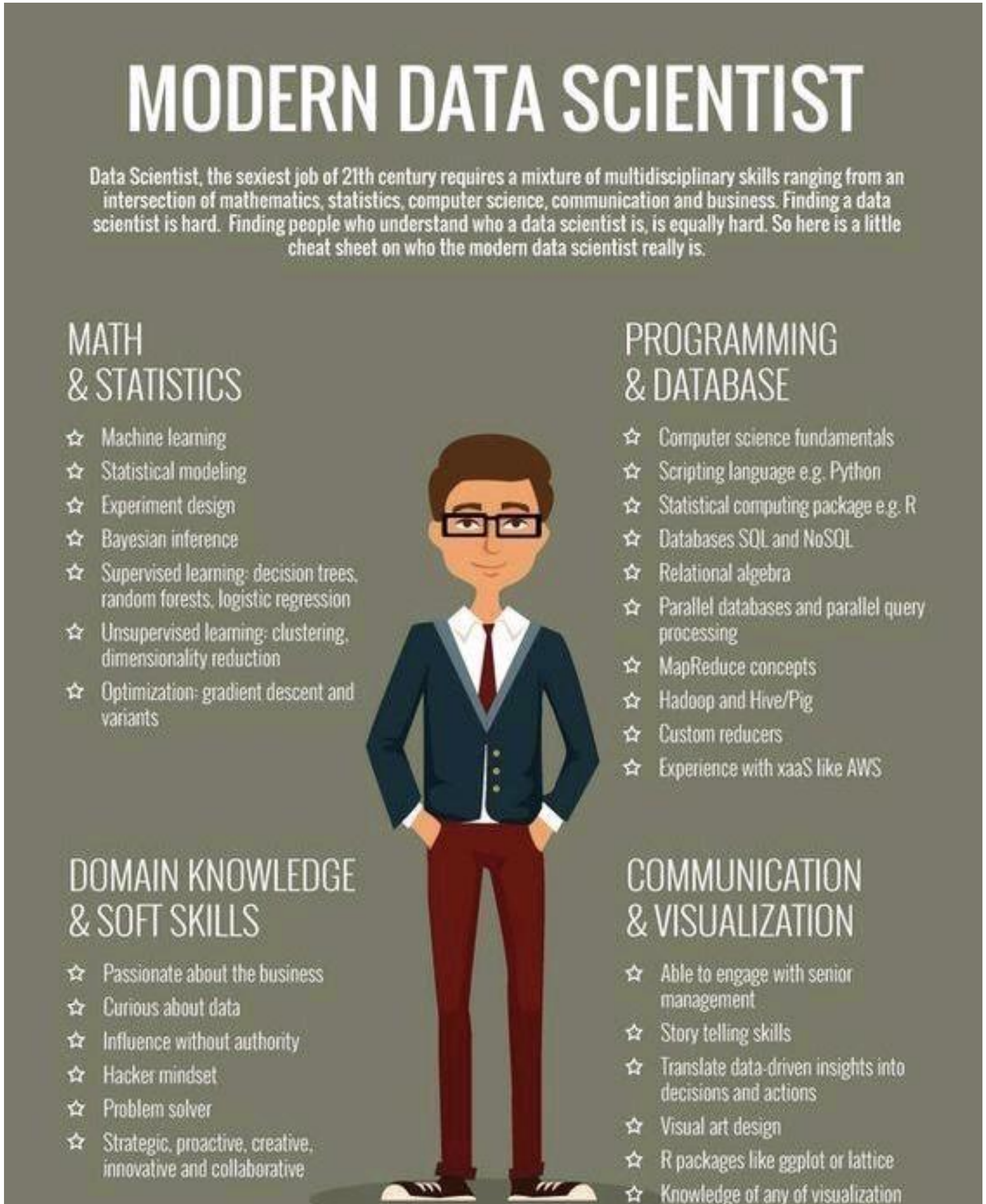


Analytics : Job Roles





CHARACTERISTICS	Business Intelligence	Data Science
Focus	Reports, KPIs, trends	Patterns, correlations, models
Process	Static, comparative	Exploratory, experimentation, visual
Data Sources	Pre-planned, added slowly	On the fly, as-needed
Transform	Up front, carefully planned	In-database, on-demand, enrichment
Data quality	Single version of truth	"Good enough," probabilities
Data model	Schema on load	Schema on query
Analytics	Retrospective, Descriptive	Predictive, Prescriptive, Preventative


MODERN DATA SCIENTIST
Data Scientist, the sexiest job of 21th century requires a mixture of multidisciplinary skills ranging from an intersection of mathematics, statistics, computer science, communication and business. Finding a data scientist is hard. Finding people who understand who a data scientist is, is equally hard. So here is a little cheat sheet on who the modern data scientist really is.
MATH & STATISTICS
Machine learning
Statistical modeling
Experiment design
Bayesian inference
Supervised learning: decision trees, random forests, logistic regression
Unsupervised learning: clustering, dimensionality reduction
Optimization: gradient descent and variants
PROGRAMMING & DATABASE
Computer science fundamentals
Scripting language e.g. Python
Statistical computing package e.g. R
Databases SQL and NoSQL
Relational algebra
Parallel databases and parallel query processing
MapReduce concepts
Hadoop and Hive/Pig
Custom reducers
Experience with xaaS like AWS
DOMAIN KNOWLEDGE & SOFT SKILLS
Passionate about the business
Curious about data
Influence without authority
Hacker mindset
Problem solver
Strategic, proactive, creative, innovative and collaborative
COMMUNICATION & VISUALIZATION
Able to engage with senior management
Story telling skills
Translate data-driven insights into decisions and actions
Visual art design
R packages like ggplot or lattice
Knowledge of any of visualization

Analytics : Data Scientists - Types

One type of data scientist creates output for humans to consume, in the form of product and strategy recommendations. They are decision scientists. The other creates output for machines to consume like models, training data, and algorithms. They are modeling scientists.

1. **Data science for humans:** the consumers of the output are decision makers like executives, product managers, designers, or clinicians. They want to draw conclusions from data in order to make decisions such as which content to license, which sales lead to follow, which medicine is less likely to cause an allergic reaction, which webpage design will lead to more engagement or more purchases, which marketing email will yield higher revenue, or which specific part of a product user experience is suboptimal and needs attention. These data scientists design, define, and implement metrics, run and interpret experiments, create dashboards, draw causal inferences, and generate recommendations from modeling and measurement.
2. **Data science for machines:** here the consumers of the output are computers which consume data in the form of training data, models, and algorithms. Examples of the work products of these data scientists ar recommendation systems which recommend what shirt a customer might like or what medicine a physician should consider prescribing based on a designed optimization function, such as optimizing fo customer clicks or for minimizing readmission rates to the hospital. Depending on the engineering background of these data scientists, these work products are either deployed directly to the production system, or if they are prototypes they are handed off to software engineers to help implement, optimize and scale them.

Analytics : Data Science Operations Roles

- **Data infrastructure:** data ingestion, availability, operations, access, and running environments to support workflows of data scientists. e.g. running Kafka and a Hadoop cluster
- **Data engineering:** determination of data schemas needed to support measurement and modeling needs, and data cleansing, aggregation, ETL, dataset management
- **Data quality and data governance:** tools, processes, guidelines to ensure data is correct, gated and monitored, documented, standardized. This includes tools for data lineage and data security.
- **Data analytics engineering:** enabling data scientists focused on analytics to scale via analytics applications for internal use, e.g. analytics software libraries, productizing workflows, and analytic microservices.
- **Data-product product manager:** creating products for internal customers to use within their workflow, to enable incorporation of measurement created by data scientists. Examples include: a portal to read out results of A/B tests, a failure analysis tool, or a dashboard that enables self serve data and root cause diagnosing of changes to metrics or model performance.

Analytics Translators: Job Roles

To understand more about what translators are, it's important to first understand what they aren't. Translators are neither data architects nor data engineers. They're not even necessarily dedicated analytics professionals, and they don't possess deep technical expertise in programming or modelling.

Instead, translators play a critical role in bridging the technical expertise of data engineers and data scientists with the operational expertise of marketing, supply chain, manufacturing, risk, and other frontline managers. In their role, translators help ensure that the deep insights generated through sophisticated analytics translate into impact at scale in an organization. By 2026, the McKinsey Global Institute estimates that demand for translators in the United States alone may reach two to four million.

The wide range of responsibilities — leader, communicator, project manager, industry expert — inherent in the translator role makes the following skills essential:

Domain knowledge
General technical fluency
Project Management Skills
Entrepreneurial Spirit

Step 1. Identifying and prioritizing business use cases
Translator role: Works with business-unit leaders to identify and prioritize problems that analytics is suited to solve.

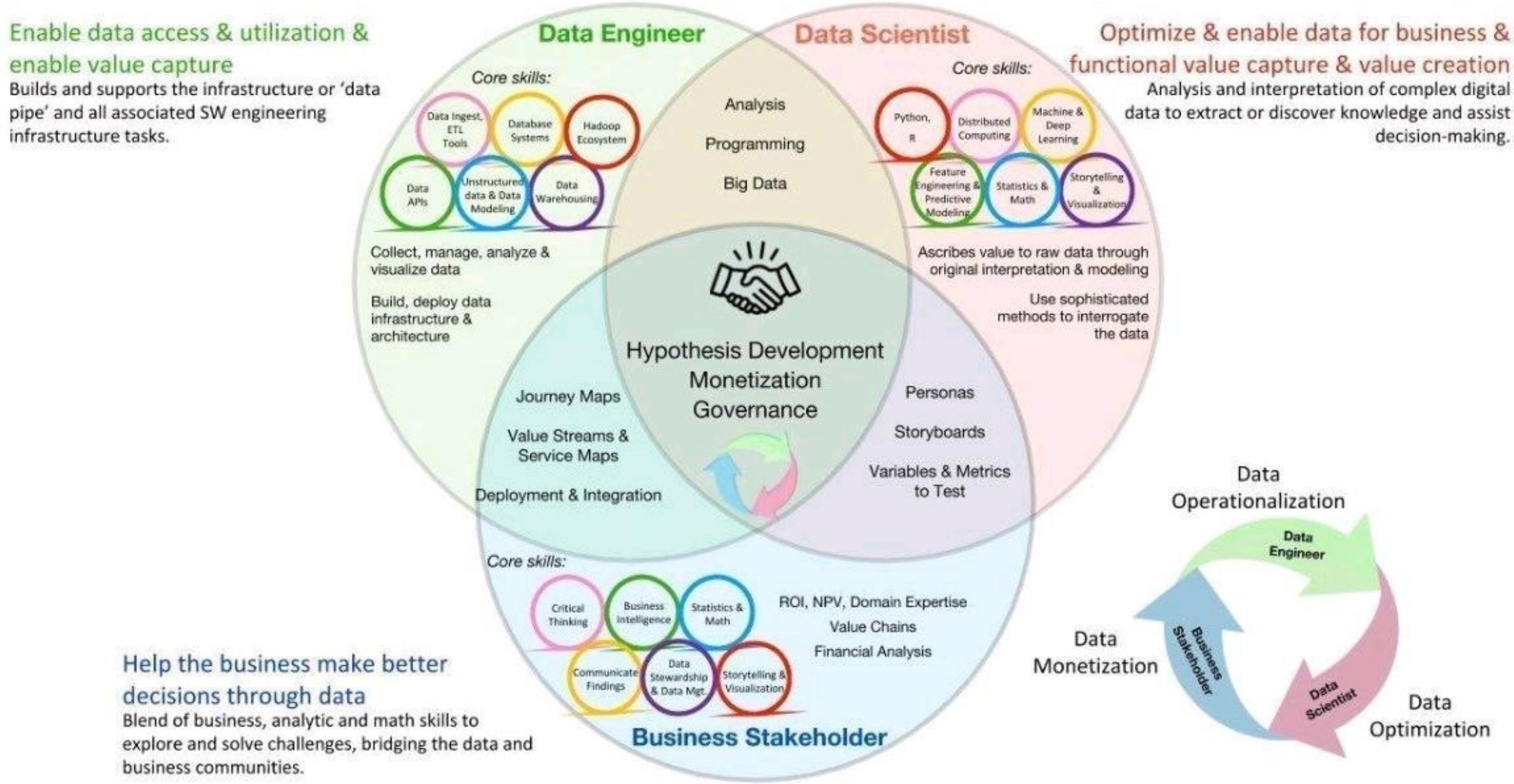
Step 2: Collecting and preparing data
Translator role: Helps identify the business data needed to produce the most useful insights.

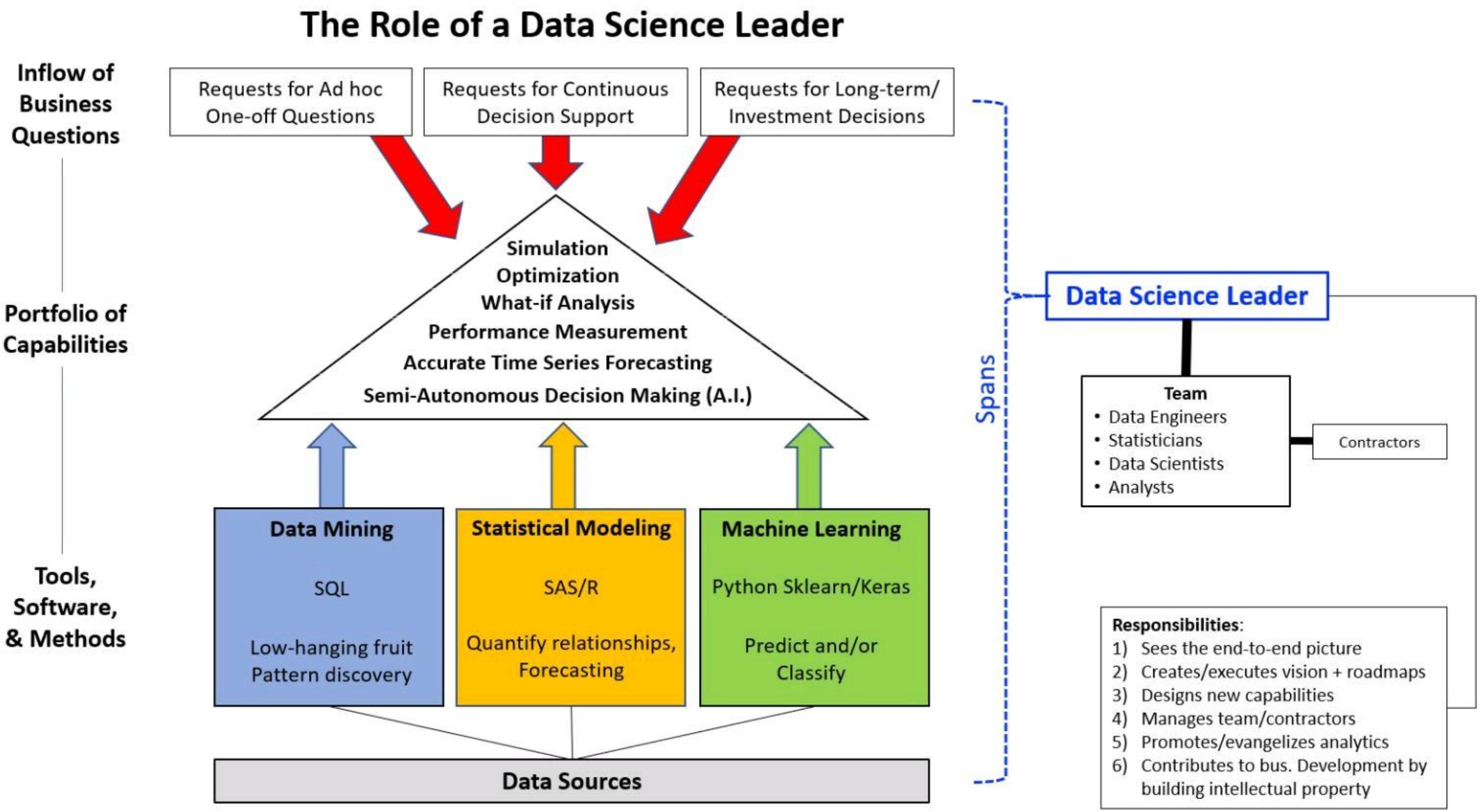
Step 3: Building the analytics engine
Translator role: Ensures the solution solves the business problem in the most efficient and interpretable form for business users.

Step 4: Validating and deriving business implications
Translator role: Synthesizes complex analytics-derived insights into easy-to-understand, actionable recommendations that business users can easily extract and execute on.

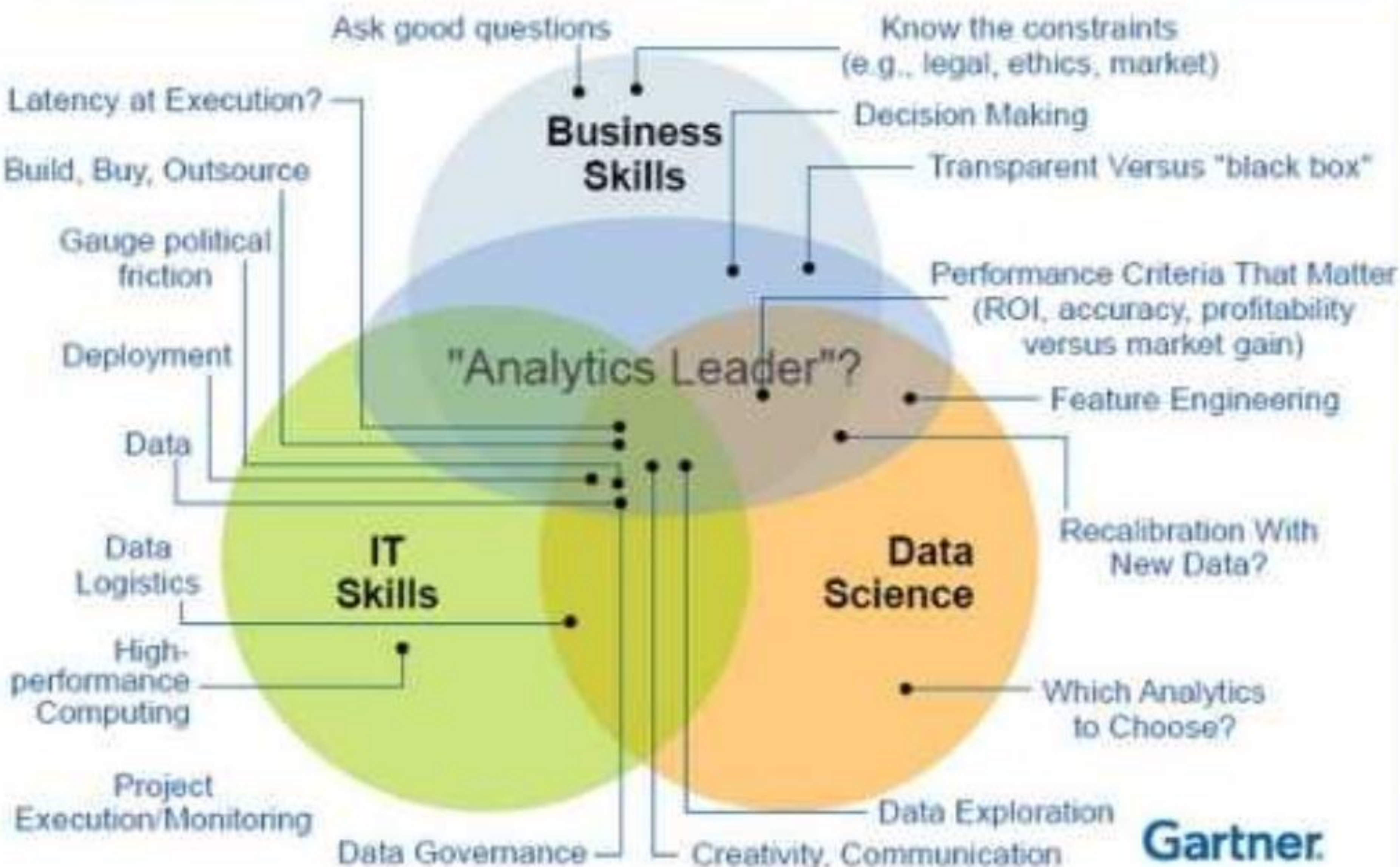
Step 5: Implementing the solution and executing on insights
Translator role: Drives adoption among business users.

Data Science Roles & How They Interact




The Role of a Data Science Leader
Inflow of Business Questions
Requests for Ad hoc One-off Questions
Requests for Continuous Decision Support
Requests for Long-term/ Investment Decisions
Portfolio of Capabilities
Simulation
Optimization
What-if Analysis
Performance Measurement
Accurate Time Series Forecasting
Semi-Autonomous Decision Making (A.I.)
Tools, Software, & Methods
Data Mining
SQL
Low-hanging fruit
Pattern discovery
Statistical Modeling
SAS/R
Quantify relationships, Forecasting
Machine Learning
Python Sklearn/Keras
Predict and/or Classify
Data Sources
Spans
Data Science Leader
Team
• Data Engineers
• Statisticians
• Data Scientists
• Analysts
Contractors
Responsibilities:
1) Sees the end-to-end picture
2) Creates/executes vision + roadmaps
3) Designs new capabilities
4) Manages team/contractors
5) Promotes/evangelizes analytics
6) Contributes to bus. Development by building intellectual property

Driving the Success of Data Science Solutions: Skills, Roles and Responsibilities ...



Core Talents for Communicating Data

Here are the ways that various talents are involved as a data science project proceeds from gathering data to developing insight to presenting to stakeholders

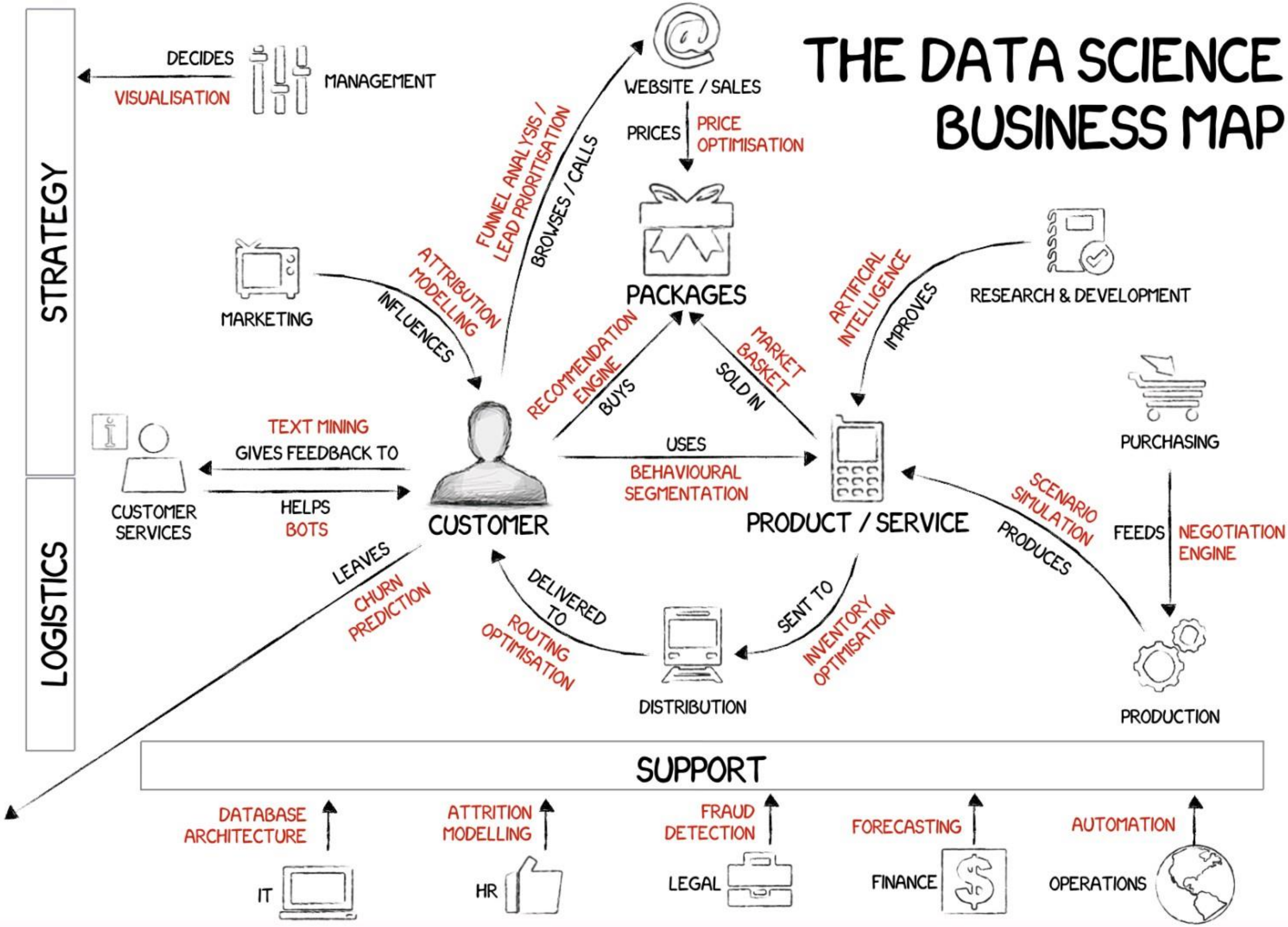
TALENT	TASKS	SKILLS	LEADS	SUPPORTS
Project Management	▪ Manage creation of team, timeline and schedules ▪ Marshal resources ▪ Troubleshoot	▪ Organization ▪ Methodology (such as Scrum) ▪ People management	▪ During creation of a data science operation ▪ During creation and execution of a project	▪ Ongoing data science operations
Data Wrangling	▪ Find, clean and structure data ▪ Develop and implement data and visualization systems, algorithms and models ▪ Develop templates and systems for repeatable processes	▪ Coding ▪ Statistics ▪ Systems architecture	▪ Early in a data team's existence ▪ Early in a project's development	▪ During routine data analysis, hypothesis testing and visual exploration of data
Data Analysis	▪ Develop and test hypotheses on data and data models ▪ Find patterns and useful trends to inform business decisions	▪ Statistics ▪ Scientific method ▪ Critical thinking ▪ Technical and nontechnical communication	▪ During routine data analysis, project design, hypothesis testing and visual exploration of data	▪ Early in a data team's existence ▪ Early in project development ▪ During visual communication development and presentations to lay audiences

Core Talents for Communicating Data

Here are the ways that various talents are involved as a data science project proceeds from gathering data to developing insight to presenting to stakeholders

TALENT	TASKS	SKILLS	LEADS	SUPPORTS
Subject Expertise	▪ Define business goals ▪ Develop and test hypotheses ▪ Develop nontechnical communication	▪ Functional knowledge ▪ Critical thinking ▪ Strategy development ▪ Nontechnical communication	▪ During project design, hypothesis testing and visual exploration of data ▪ During communication to nontechnical audiences	▪ Early in a data team's existence ▪ During visualization and design process
Design	▪ Develop visual communication and presentations ▪ Create templates and styles for repeatable visualization	▪ Information design ▪ Presentation design ▪ Design thinking ▪ Persuasive communication	▪ During data visualization and the creation of presentation and visual systems (templating)	▪ During visual iteration and prototyping
Storytelling	▪ Develop stories from data and visuals ▪ Help construct presentations in story format ▪ Present to nontechnical audiences	▪ Information design ▪ Writing and editing ▪ Presenting ▪ Persuasive communication	▪ During creation of data visualization and presentations ▪ During presentation to nontechnical audiences	▪ During visual iteration and prototyping

https://applied-data.science

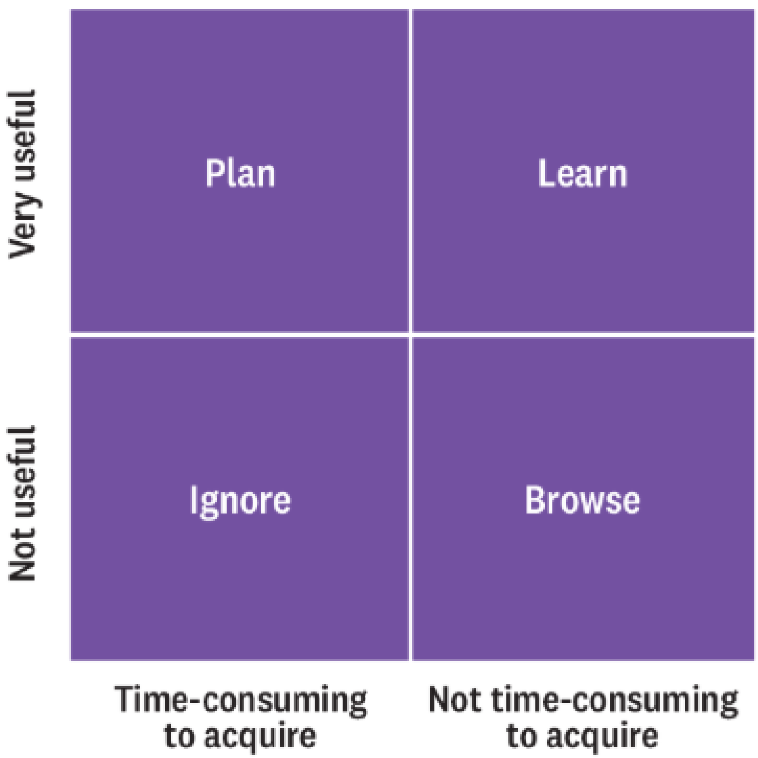


ANALYTICS LEARNING PATHWAYS

Analytics: Learning Pathways

Which Data Skills Should You Learn First?

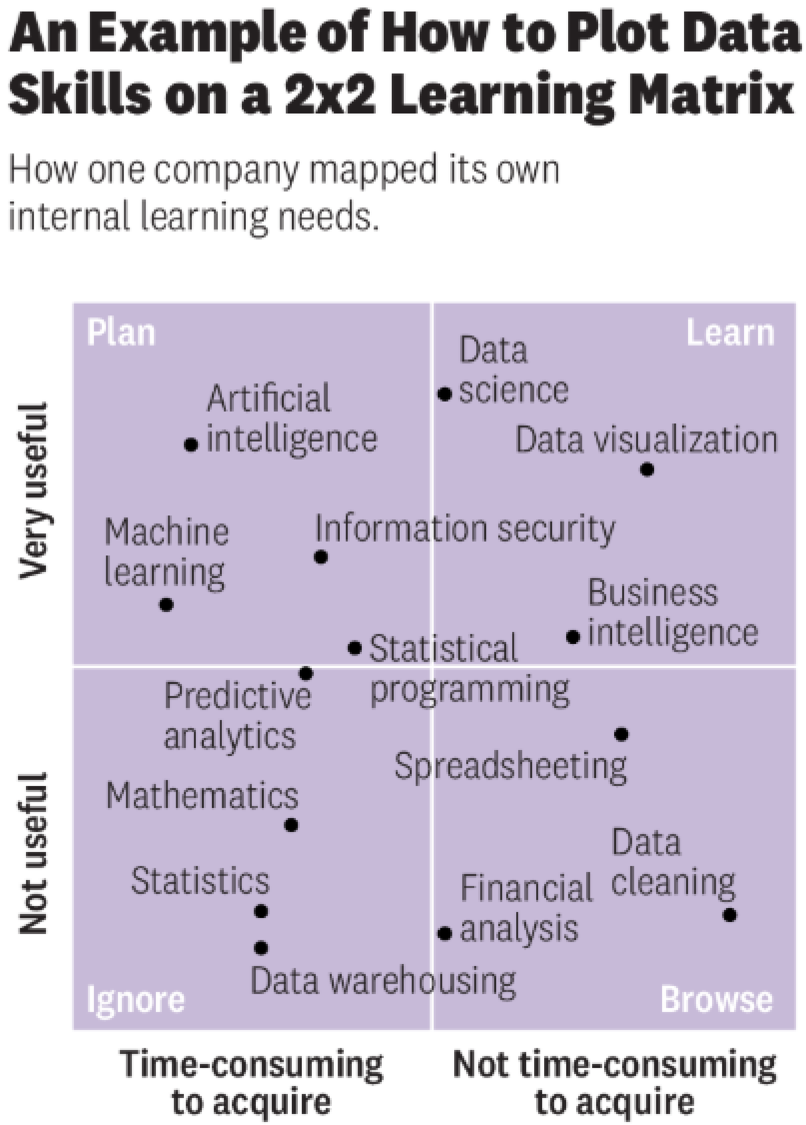
Make the most of your limited learning time.



- **Learn:** high utility, low time-to-learn. This is low hanging fruit that will add value for you and your team quickly.
- **Plan:** high utility, high time-to-learn. While this is valuable, acquiring this skill will mean prioritizing it ahead of other learning and activities. You need to be sure that it's worth the investment.
- **Browse:** low utility, low time-to-learn. You don't need this now, but it's easy to acquire so stay aware in case its utility increases.
- **Ignore:** low utility, high time-to-learn. You don't have the time for this.

https://hbr.org/2018/10/prioritize-which-data-skills-your-company-needs-with-this-2x2-matrix

Analytics: Learning Pathways



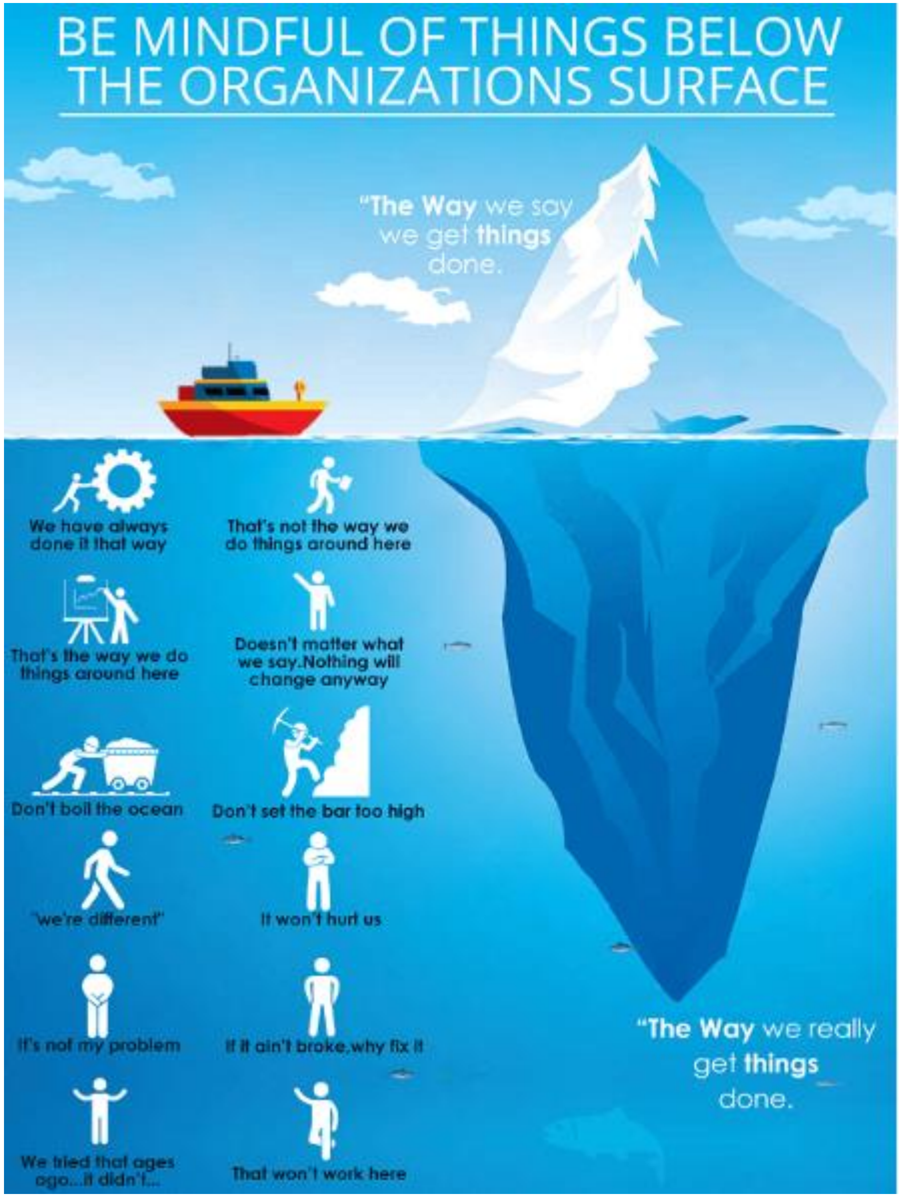
https://hbr.org/2018/10/prioritize-which-data-skills-your-company-needs-with-this-2x2-matrix

Analytic: Job Roles Descriptions

Adobe Acrobat
Document

ANALYTICS – COMPATIBLE ORGANIZATIONAL CULTURE

Tip of Iceberg



Fostering the right organizational culture

Align Innovation & Business Strategy

- Ensure that all stakeholders have visibility on the strategy/vision of company
- This ensures that innovation efforts are targeted towards initiatives that the organization supports, and enhances their chances of success

Reduce Fear of Failure

- Propagate an operating structure where a senior level sponsor provides approvals at every stage of the NDP value chain
- This enables project teams to try different ideas as it provides a sense of "joint responsibility" in case of failures

Foster Collaboration

- Create small start-up like teams that focus on testing new ideas
- Smaller teams enhance collaboration/communication and can therefore predict upstream challenges

Drive Openness to New Ideas

- Eliminate resistance to external ideas/functional fixedness
- Companies need to develop a culture where they constantly validate internal NPD initiatives against external stimulus to ensure the success of innovation programs

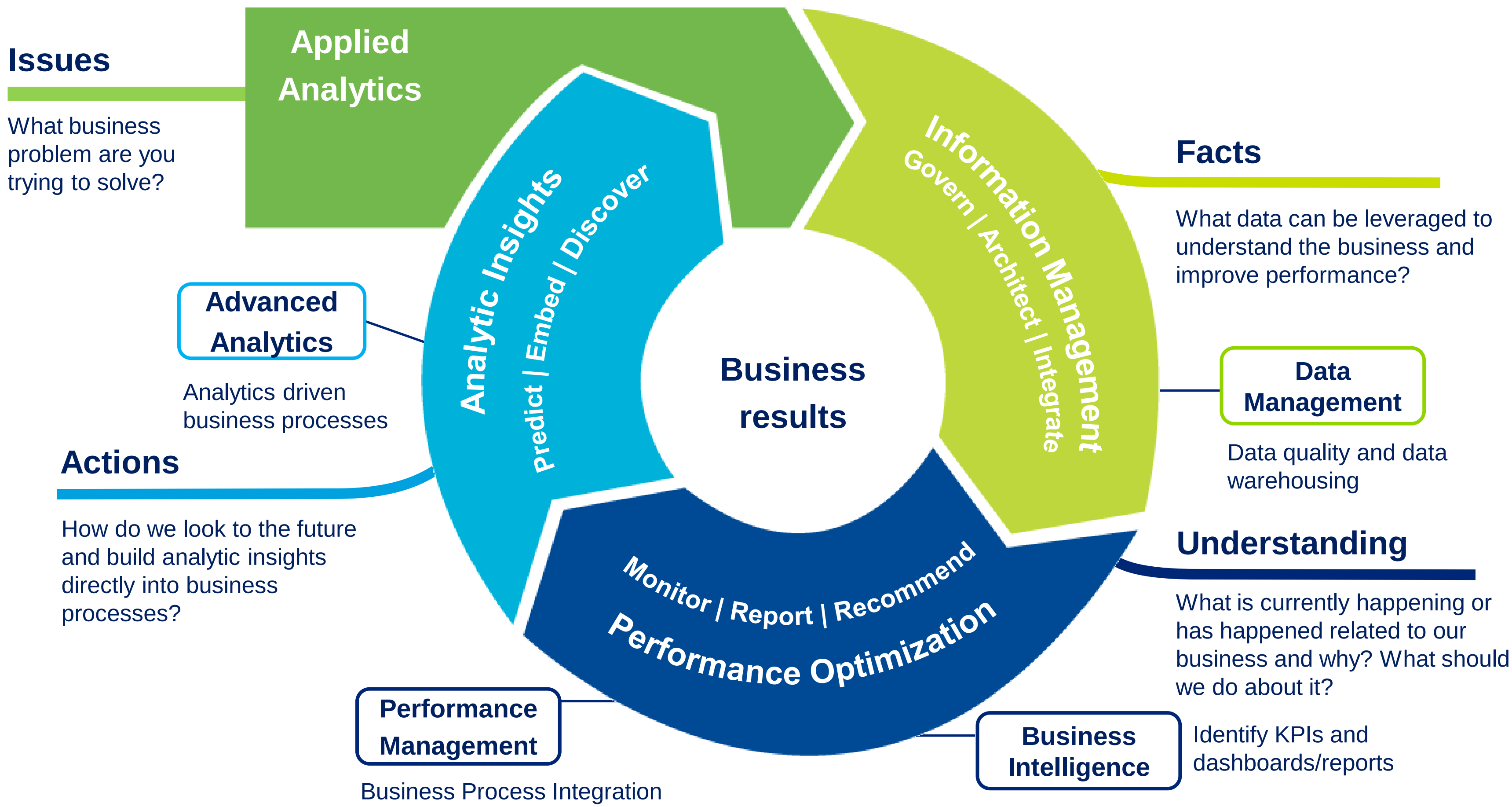
Key Success Factors

- Learning must occur in multiple ways
- The environment needs to be safe to fail
- Fearless environment
- Collaborative
- Involves diversity and many skills
- Results from trial and error
- Driven by inquiry
- Expanded by imagination
- Has to be fun

CRAFTING ANALYTICS VISION AND STRATEGY

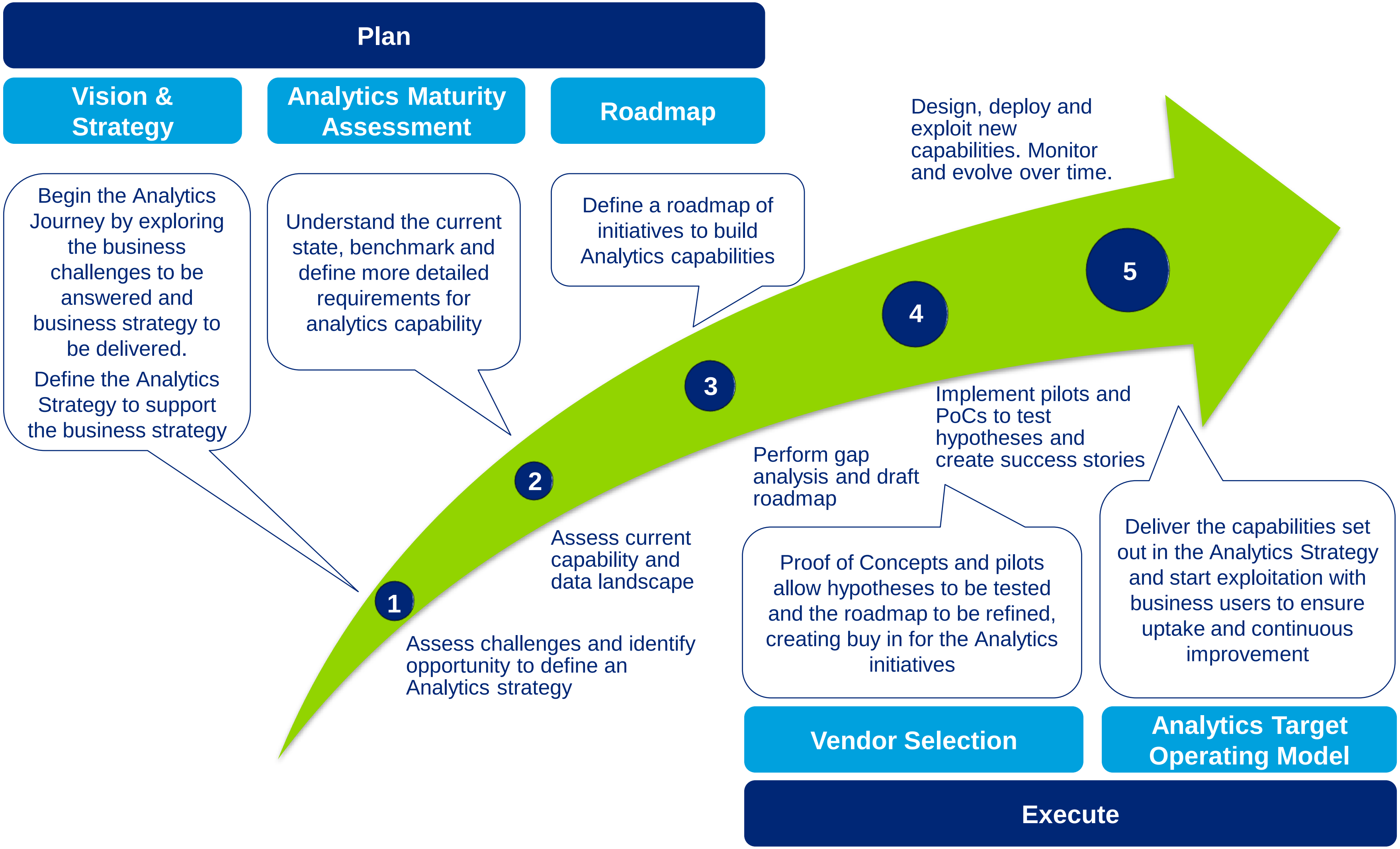
Analytics: Key capabilities to provide Analytics Vision

Starting anywhere in the analytics cycle, an organization can immediately begin to address its specific needs. Analytics comprises the skills, technologies, applications, and practices to assist organizations continuously gain insights to drive business outcomes.



Visualising the Analytics journey

There are five key stages on the Analytics journey



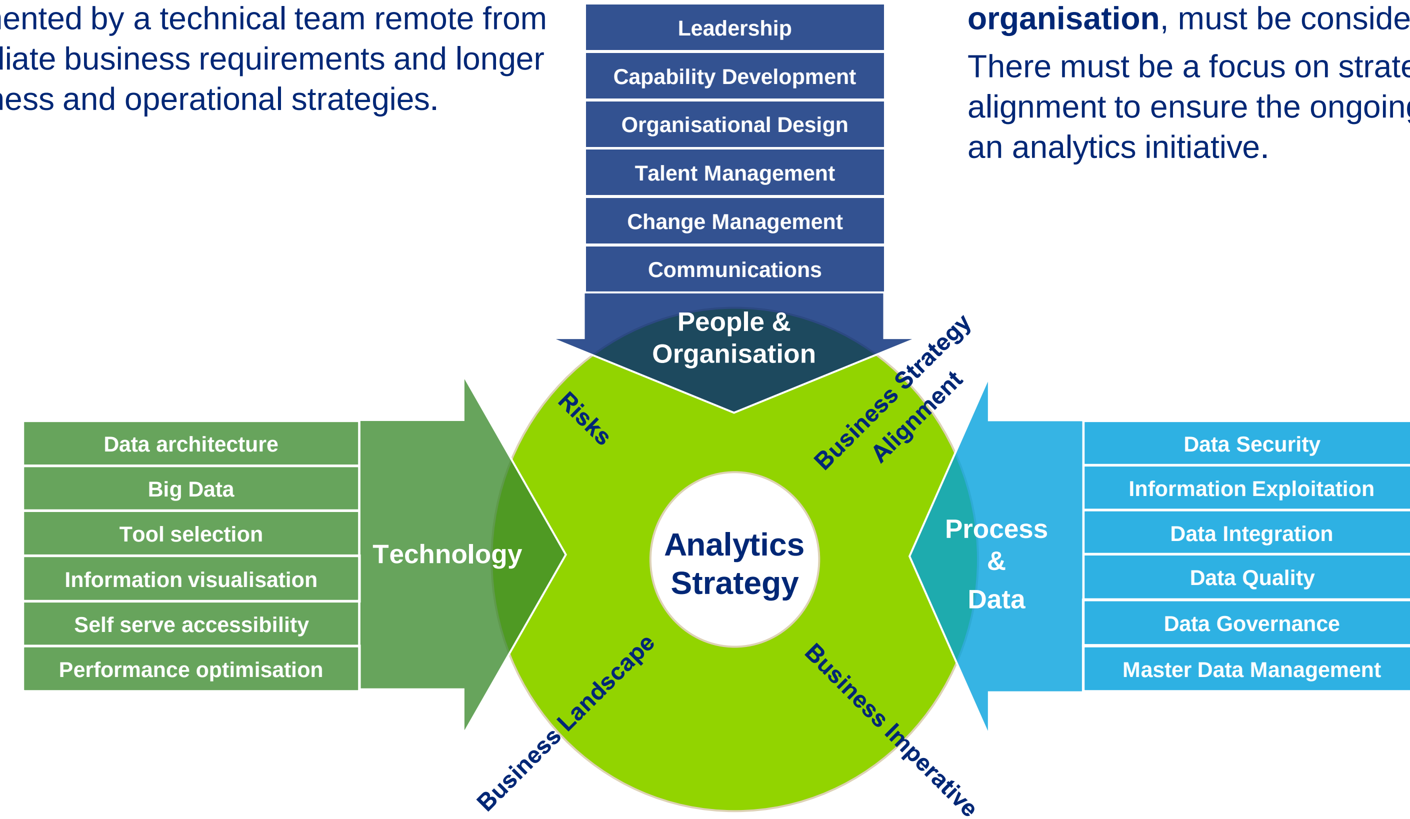
Structuring an Analytics strategy

A cohesive strategy requires consideration of technology, process & data and people & organisation

We often find that organisations consider their Analytics strategy as a **technology** problem, to be implemented by a technical team remote from the immediate business requirements and longer term business and operational strategies.

We believe that all areas of an organisation, including **process & data** and **people & organisation**, must be considered.

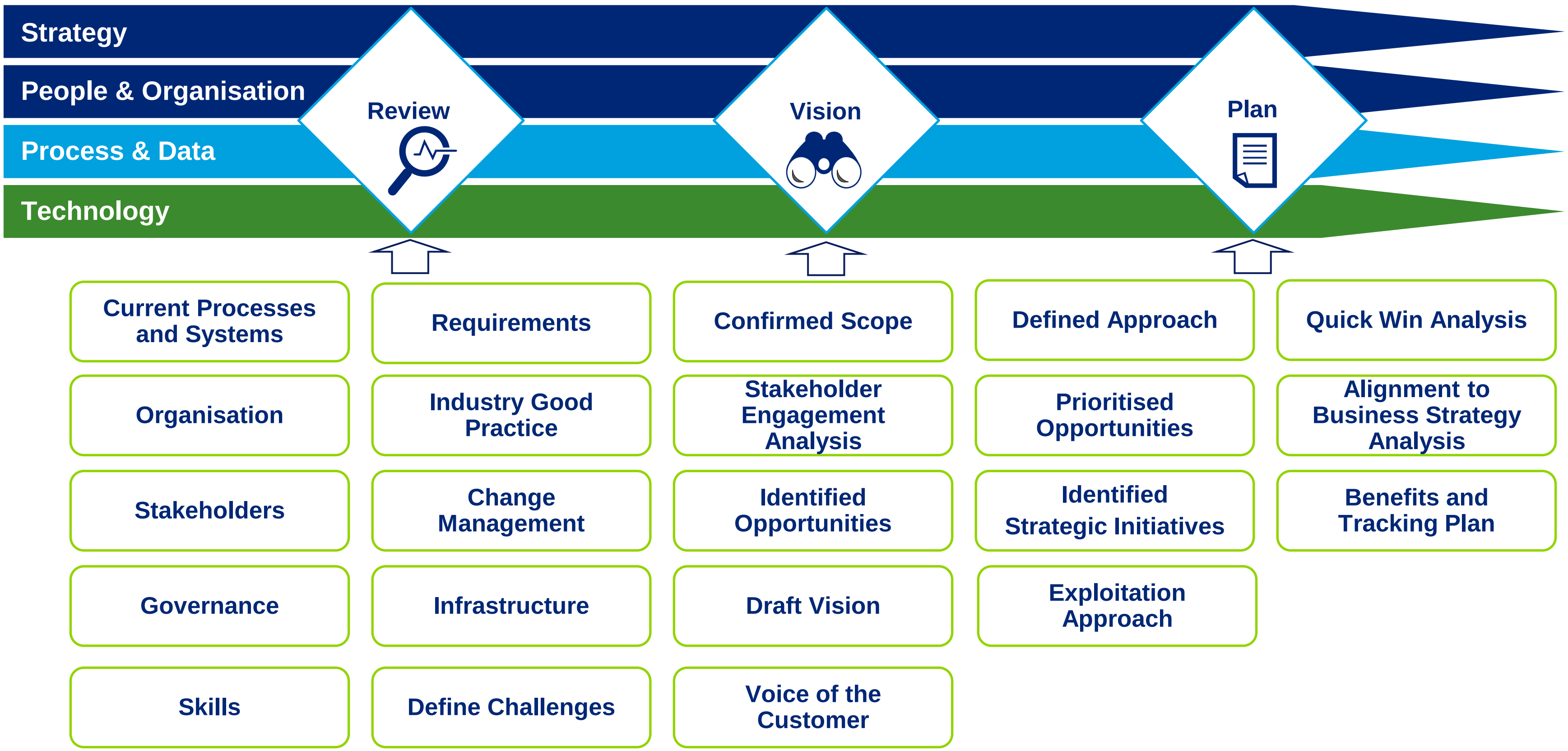
There must be a focus on strategic alignment to ensure the ongoing success of an analytics initiative.



The Analytics Target Operating Model (ATOM) uses these dimensions to integrate Analytics within an organisation.

Our recommended approach to formulate the strategy

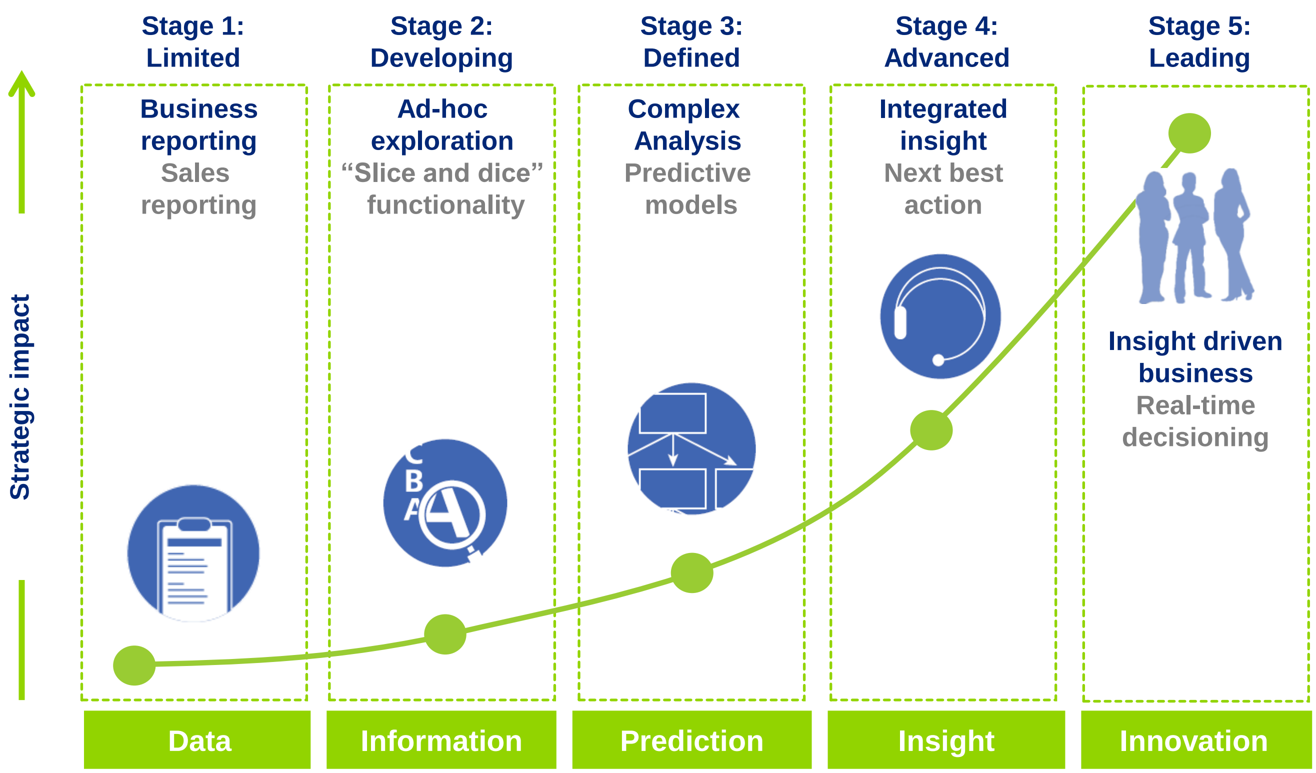
Following the three steps depicted below, current capability will be assessed against vision and future requirements to define a series of strategic and tactical activities. These, in turn, will be incorporated within the roadmap for developing Analytics, to solve business challenges, within the organisation.



Analytics maturity assessment

Provides an approach to categorise and evaluate key aspects of Analytics maturity in an organisation

A scoring system is derived from a series of interviews with stakeholders from across the organisation. The exercise is conducted using 5 dimensions: strategy, people, process, technology and data. Additionally, these scores can be cross-referenced against the components of the Analytics pyramid: Data Management, Business Intelligence, Performance Management and Advanced Analytics.



Strategy

Degree to which Analytics is integral to strategy development, decision-making, and execution

People

The extent to which there is a critical mass of personnel recruited, trained and incentivised to apply analytic techniques

Process

The level to which Analytics and analytic approaches are embedded in core business processes

Technology

The sophistication and proliferation of Analytics tools and technologies

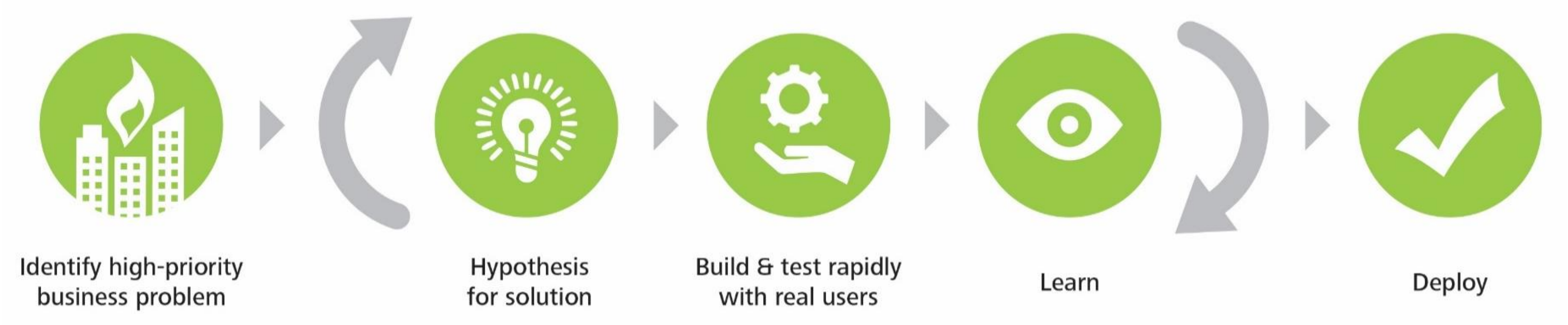
Data

The richness, availability, quality and governance of data across business functions

Test hypotheses and create success stories

Experimenting and evaluating will help create buy in

- The next step on the Analytics Journey, once the current state and the target state have been defined, is to begin to bridge the gap between the two.
- However, an ambitious roll out plan which implements all desired Analytics capabilities simultaneously involves significant risks, both from a technical and change management perspective.
- Instead, by prioritising the Analytics initiatives leading to deliver the strategy, and first testing them in a controlled environment, it will be possible to demonstrate the benefits and outcomes from Analytics without exposing the organisation to unnecessary risks.
- In a small scale experiment, specific hypotheses can be tested, and the results evaluated, allowing the original ideas to be refined and redrawn to better achieve the specific business aims.



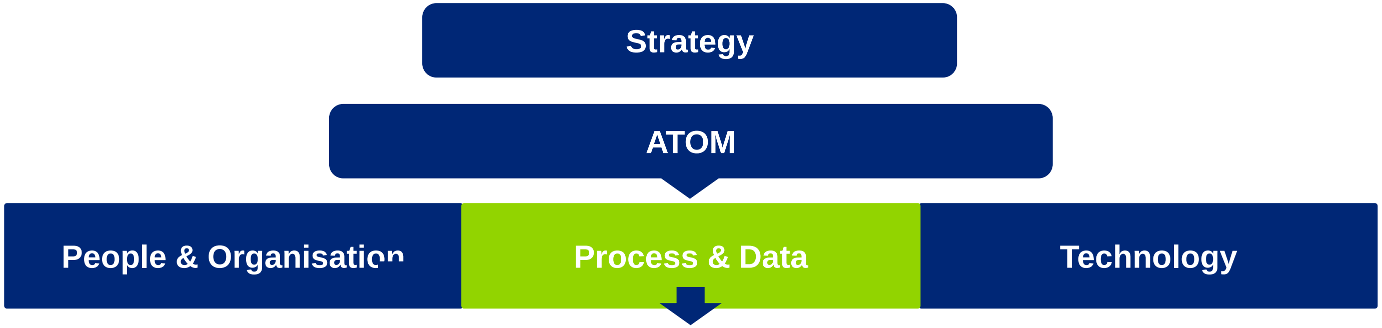
Design, deploy and exploit capability

Step five is the implementation of Analytics within an organisation

Having decided on the priorities for implementation of the new Analytics capabilities, and tested them extensively in a controlled environment, the final phase of the Analytics Journey is to design and build the necessary systems and technology to enable full implementation.

Areas of consideration for this final phase are:

- Identify people, process and technology capabilities required to deliver your strategy
- Identify known use cases
- Decide how to procure and implement what you need
- Prioritise and implement successful, high value initiatives in production
- The key areas of an organisation for implementing Analytics, introduced through the concept of an Analytics Target Operating Model (ATOM), are highlighted below and will be expanded upon in the subsequent sections of this chapter.



Once the infrastructure for Analytics is in place, and data has been analysed to produce the insight which will drive the business forward, it is important to continuously measure both the output, and the model used to produce the output, in order to refine and improve how Analytics is used and delivered.

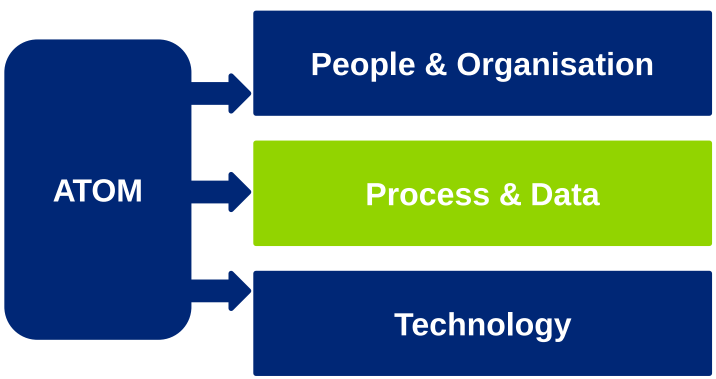
Analytics Target Operating Model (ATOM)

Define how to run and manage analytics in day to day operations

Analytics offers the opportunity to transform how an organisation operates, but to achieve the maximum benefit, the organisation's operations will need to be adapted to accommodate the development of the Analytics capability.

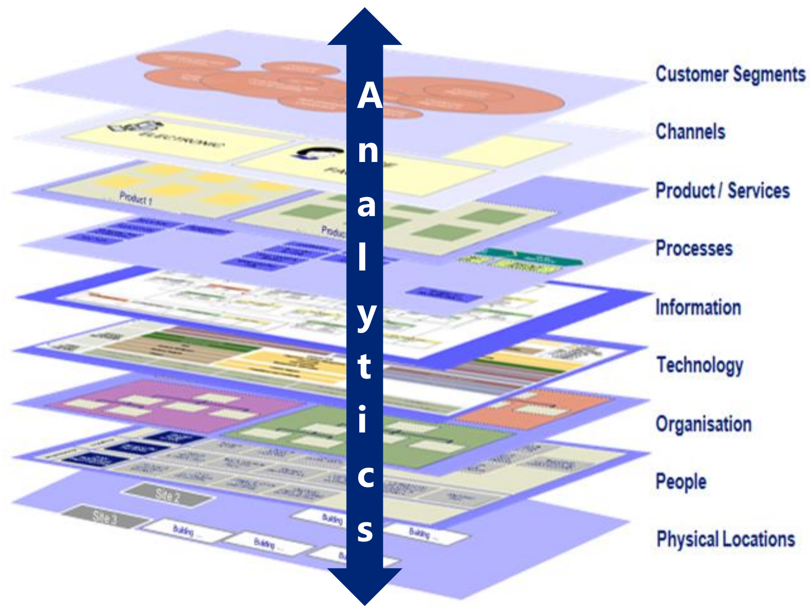
The effects of developing an Analytics capability can necessitate changes to the organisation's operating model, including

- Strategic process re-engineering
- Tactical process improvements
- Operational improvements
- Technology renewal
- People/culture changes



The Analytics roadmap is designed based on the ATOM analysis, where each change to the model must be defined in terms of initiatives which will achieve it.

By structuring the firm to accommodate Analytics, through the ATOM, the benefits of well positioned, well supported and well executed Analytics can resonate throughout all the different layers of the organisation.



Analytics Platform

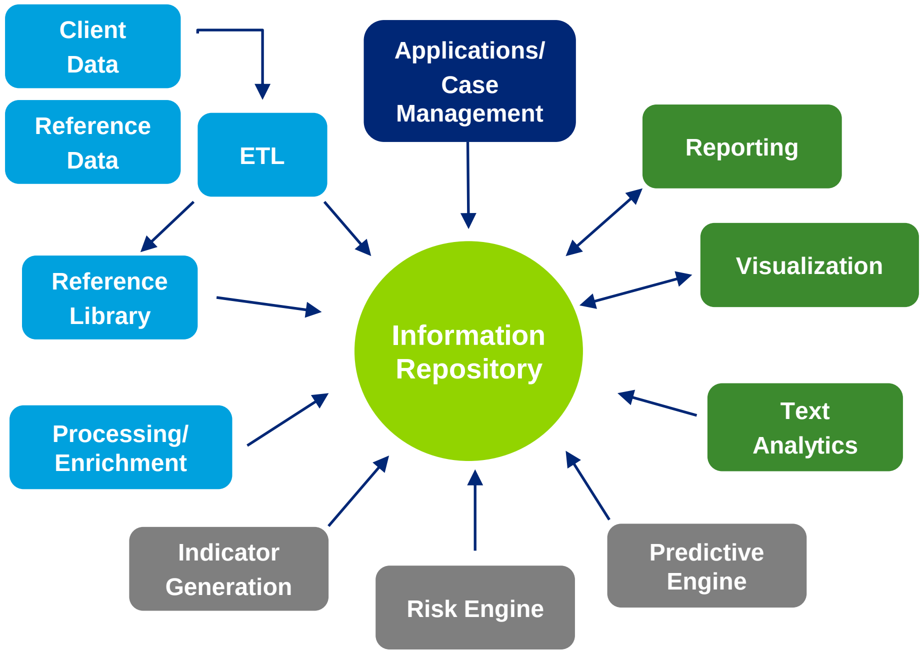
Key features

Data preparation

- Standardized connectors for data extraction, transformation, and loading (ETL) process
- Common data model facilitating global consistency
- Built-in reference library of third-party data sources (e.g., PEP lists)

Data enrichment

- Enhances an entity's existing data
- Geocoding — assigns longitude and latitude coordinates to data records for analysis
- Text matching — used for entity resolution and de-duplication
- Validation — confirm mailing addresses, taxpayer IDs, etc.



Data reporting

- Interactive case management and customizable workflow
- Advanced, integrated third-party data visualization and interactive reporting tools
- Customizable web-based interfaces for tailored investigation needs and fact sets

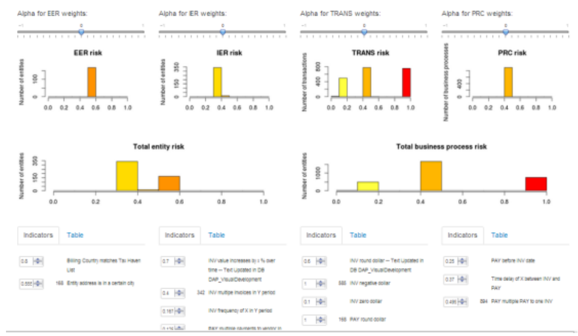
Data analysis

- Proprietary, multi-process risk scoring engine
- Hundreds of pre-programmed indicators derived from actual investigations and research
- Analytic capabilities for predictive, text, geospatial, & social network analyses

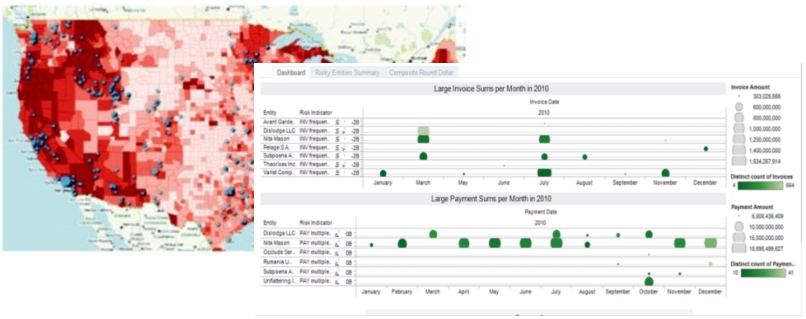
Analytics Platform

Interactive, flexible, visual, integrated, and powerful

Fine tune indicators to create optimal risk profiles

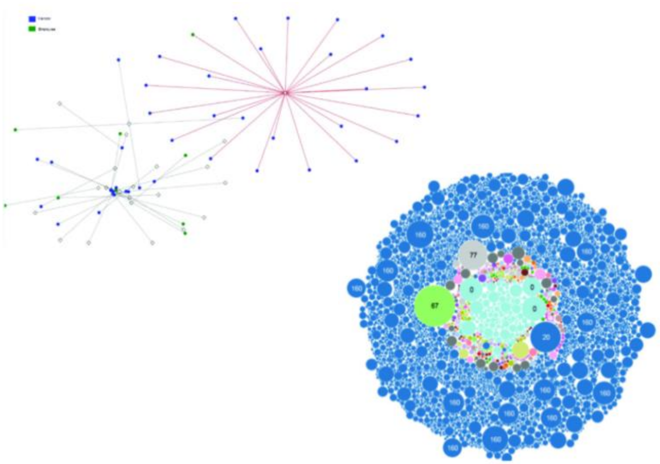


Intuitive and highly digestible visuals



Risk based review and prioritization

Advanced analytics to derive insights and uncover interrelationships



69

Value delivered as repeatable, embedded predictions and actions

Run Better

Driving process agility and intelligence

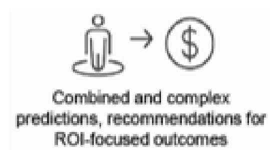
RUN, DESCRIBE AND DIAGNOSE

Measure and Report: when, what, where, how – know why and be alerted

Run Different

Find new insights to drive new outcomes

PREDICT AND PRESCRIBE

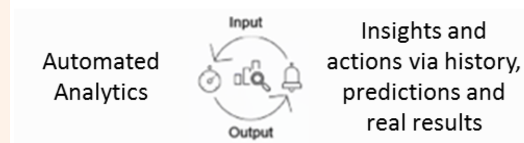


Know what happens next with new actions

Run Digital

Integrating automated analytics into human process

EMBEDDED INTELLIGENCE



Real time predict, act, improve and repeat

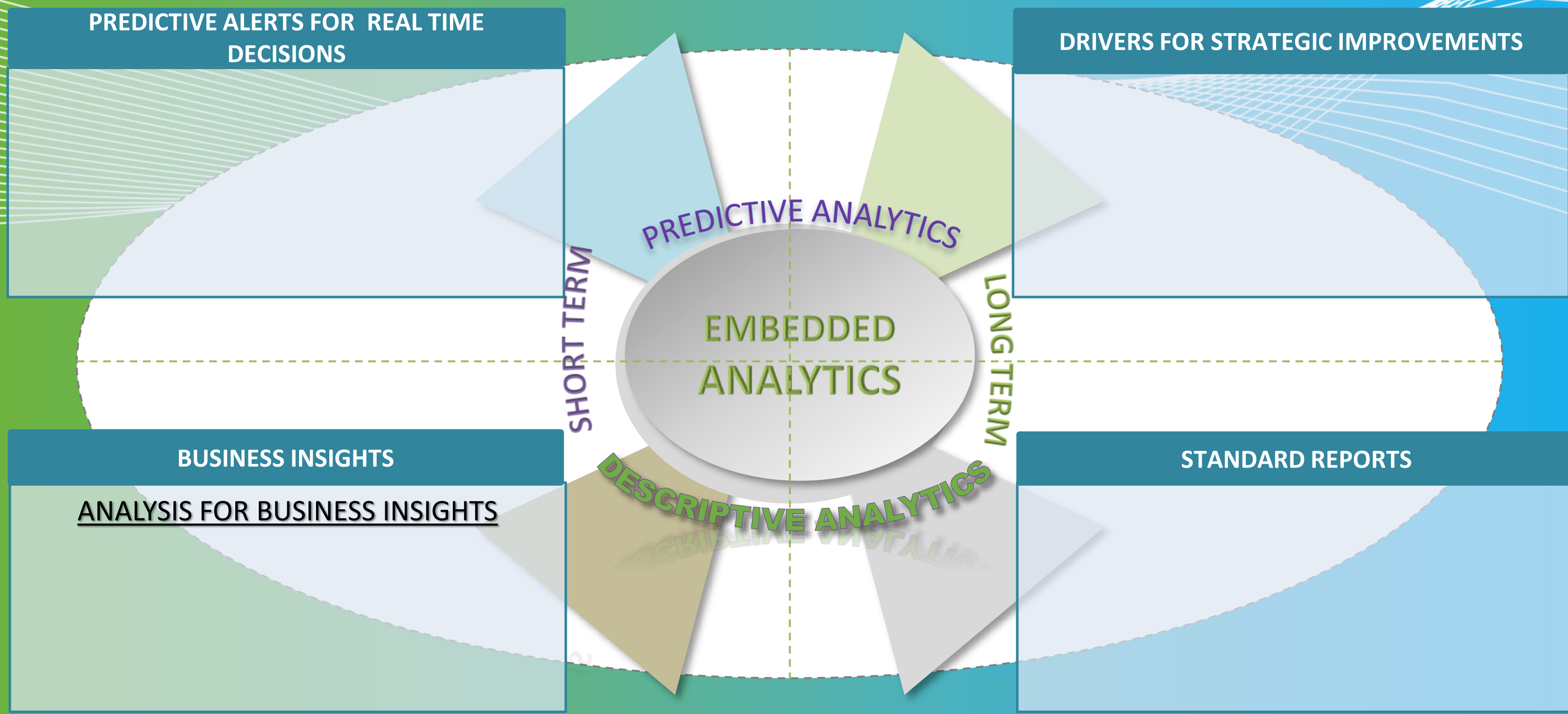
IMPROVED BUSINESS PROCESS, DECREASED COST AND/OR INCREASED REVENUE

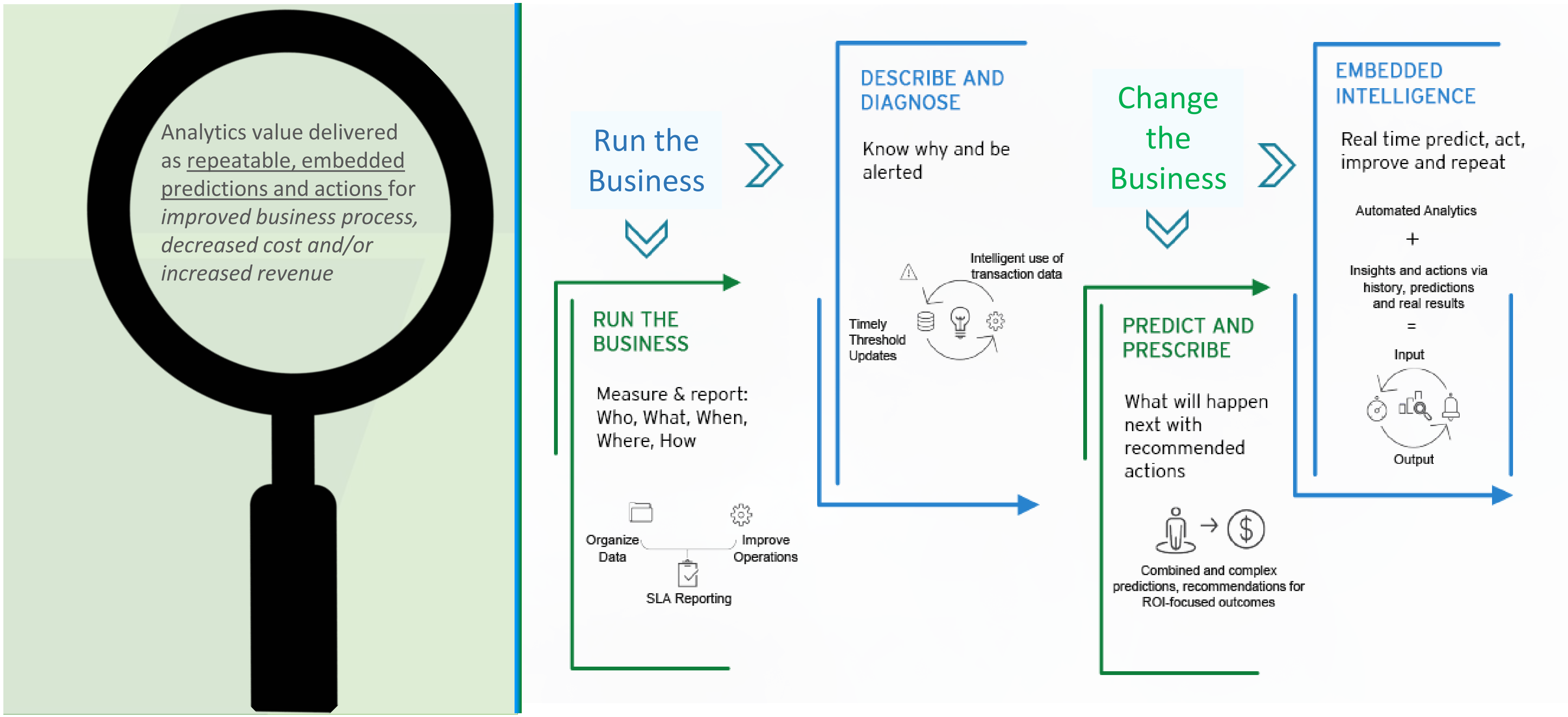
Reduce Cost, Prevent Leakage, Improve TAT, Quality, Compliance, e.g. FCR, Directory Accuracy

Improve CSAT, MTM, Star Ratings,
Close Gaps in Care, Reduce Interest Payments on Claims

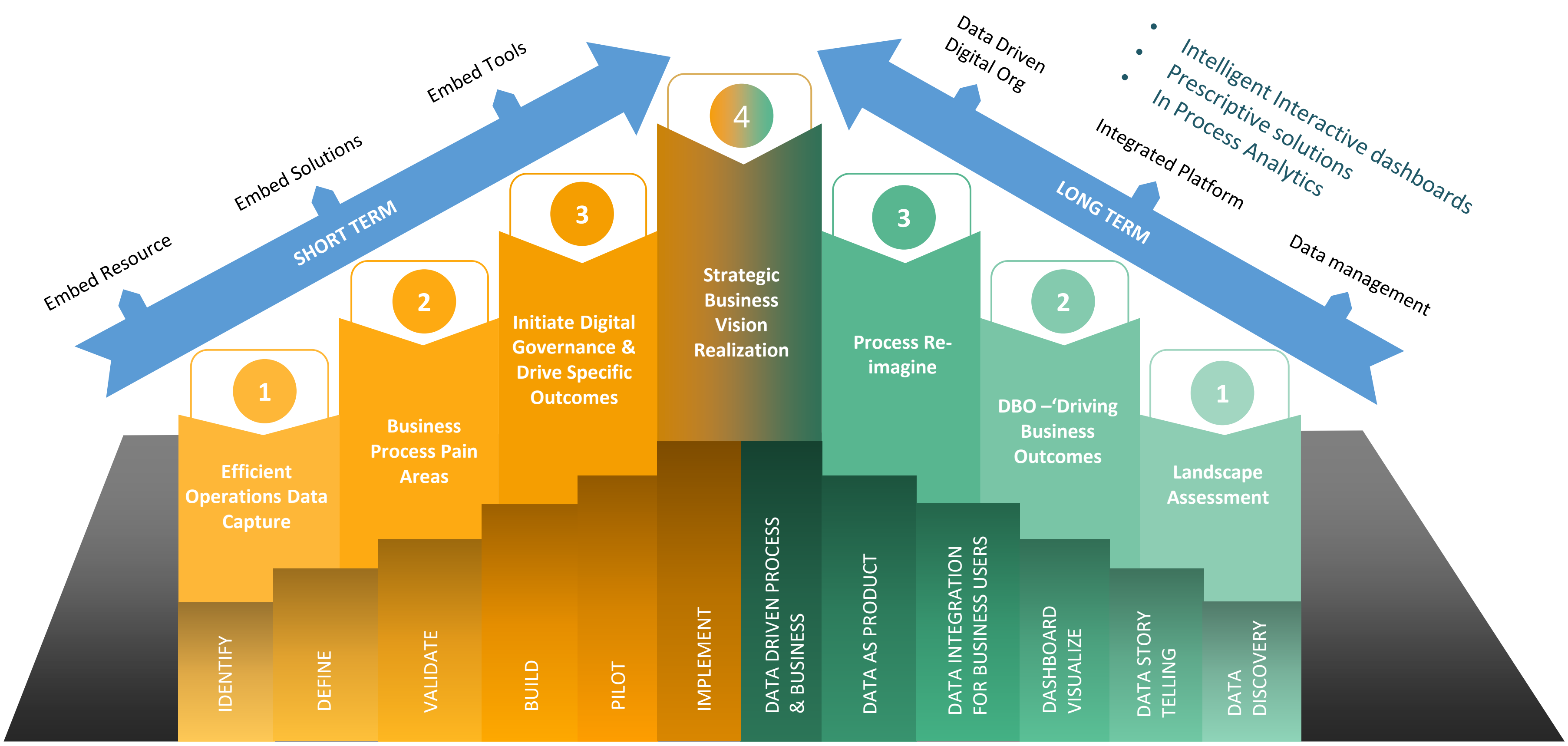
Enhanced Customer Experience and Increase Sentiments

How Operations consumes Analytics Output?

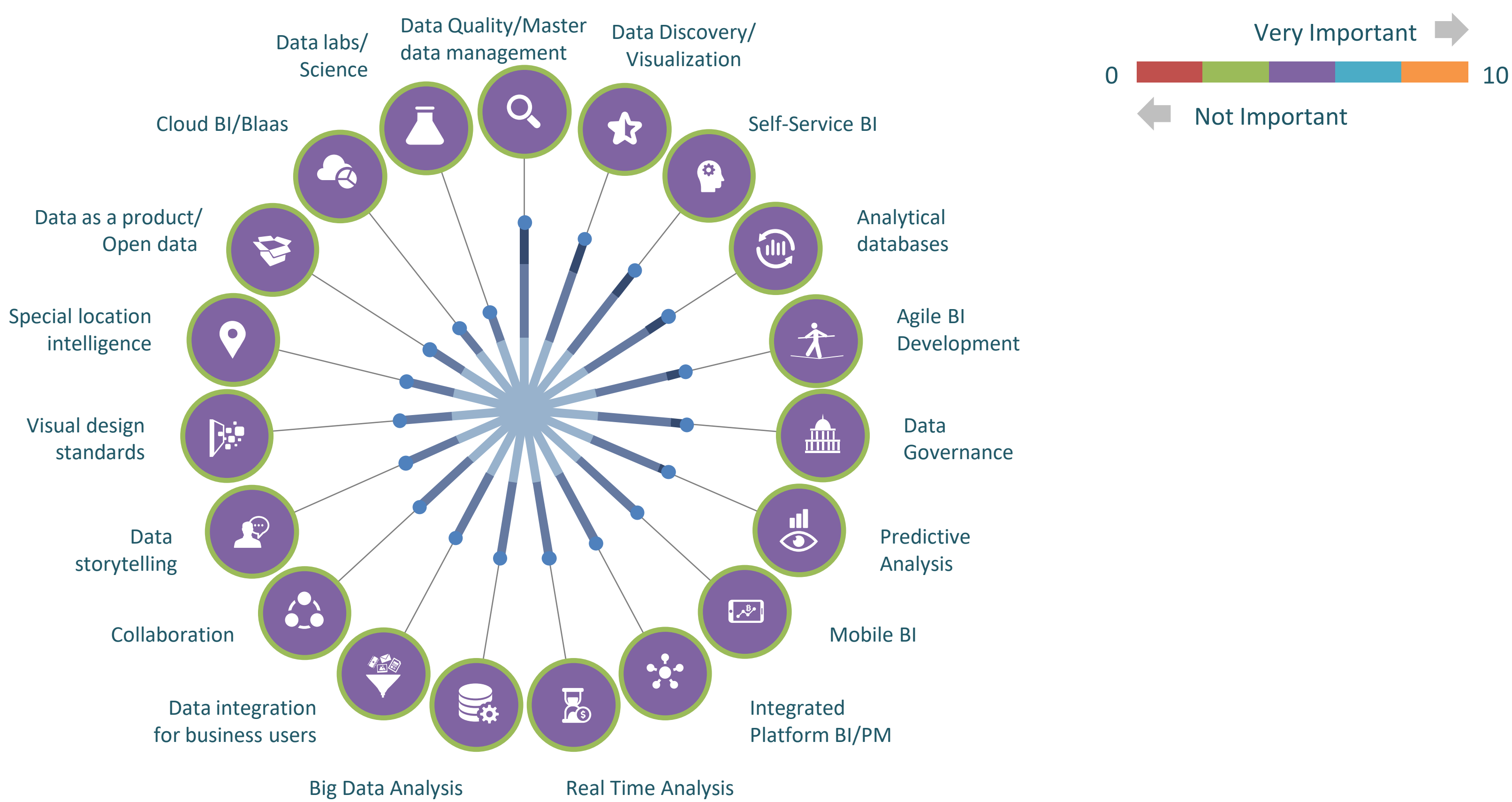



Analytics value delivered as repeatable, embedded predictions and actions for improved business process, decreased cost and/or increased revenue
Run the Business
RUN THE BUSINESS
Measure & report: Who, What, When, Where, How
Organize Data
Improve Operations
SLA Reporting
DESCRIBE AND DIAGNOSE
Know why and be alerted
Intelligent use of transaction data
Timely Threshold Updates
Change the Business
PREDICT AND PRESCRIBE
What will happen next with recommended actions
Combined and complex predictions, recommendations for ROI-focused outcomes
EMBEDDED INTELLIGENCE
Real time predict, act, improve and repeat
Automated Analytics
+
Insights and actions via history, predictions and real results
=
Input
Output

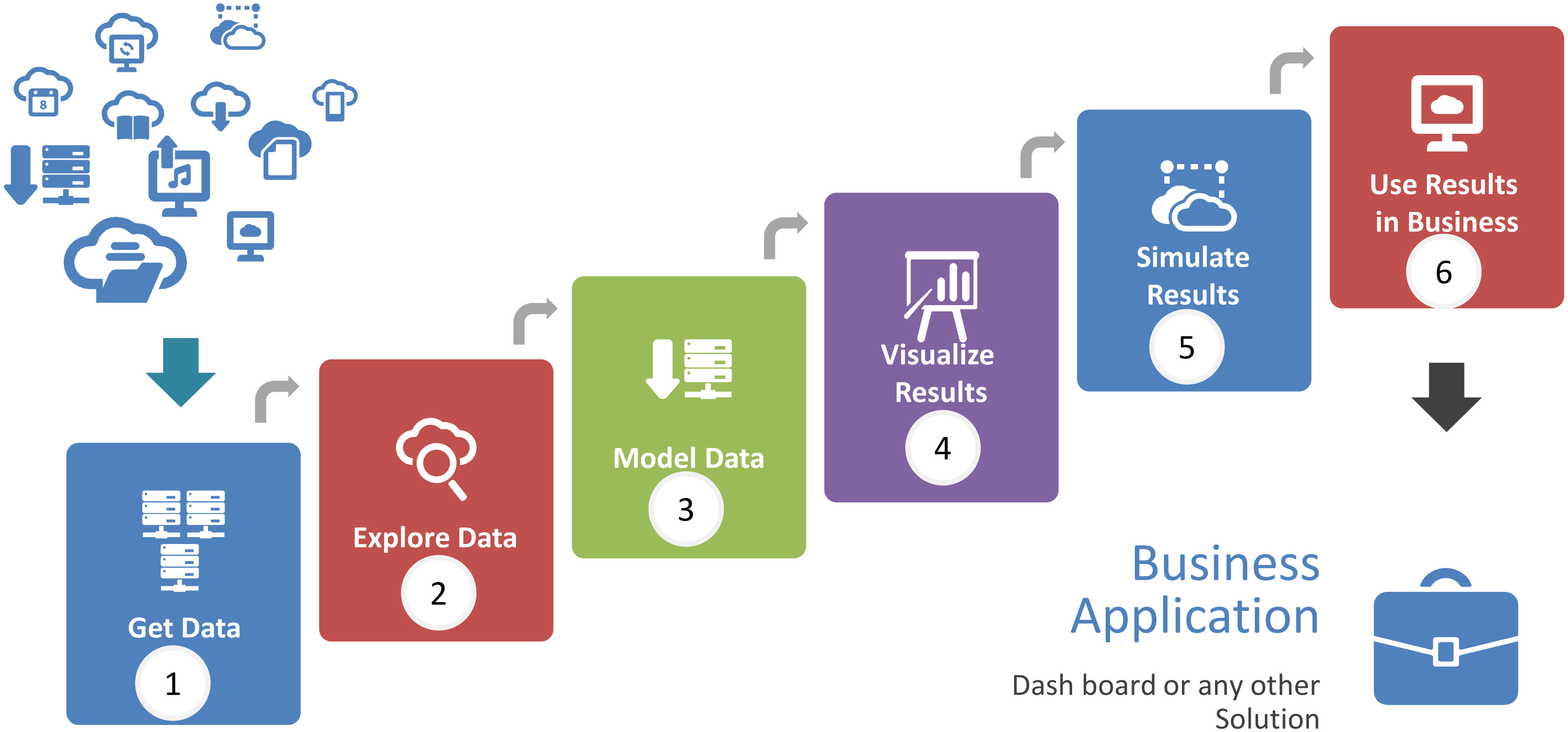
Analytics Journey on the ground



Analytics Ecosystem



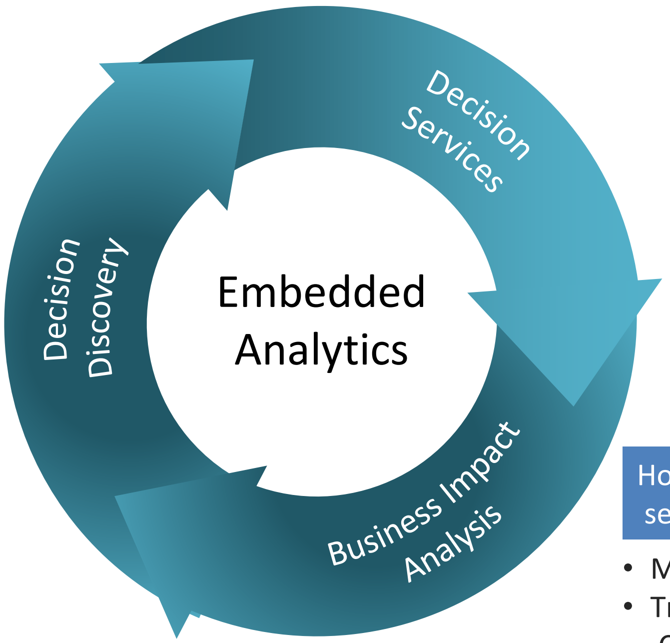
Analytics Process



Embedded Analytics

How is it discovered?

- Identify Critical Process
- Map Data requirements related to process.
- Descriptive & Diagnostic Analytics



How services are delivered?

- Predictive Analytics based on Discovery.
- Enable Intelligent Decisions for Agent
- Enable Process Automation
- Enable Process Redesign

How effective are services delivered?

- Measure model Quality & Efficiency
- Translate Quality & Efficiency to impact by TAT , Quality ,CSAT & other business outcomes
- Business Sign off on the impact value
- Input for Decision Discovery for continuous improvement

Embedded Analytics

1 Discovery

- Identify Critical Business Process
- Map Data requirements related to process.
- Descriptive & Diagnostic Analytics

2 Develop Model

Develop Predictive Models based on discovery
Descriptive analytics provides input for Process Automation.

3 Model Usage

Agents enable to make Intelligent Decisions using Predictive analytics
Automation of Process based on Analytics insight

5 Business Impact

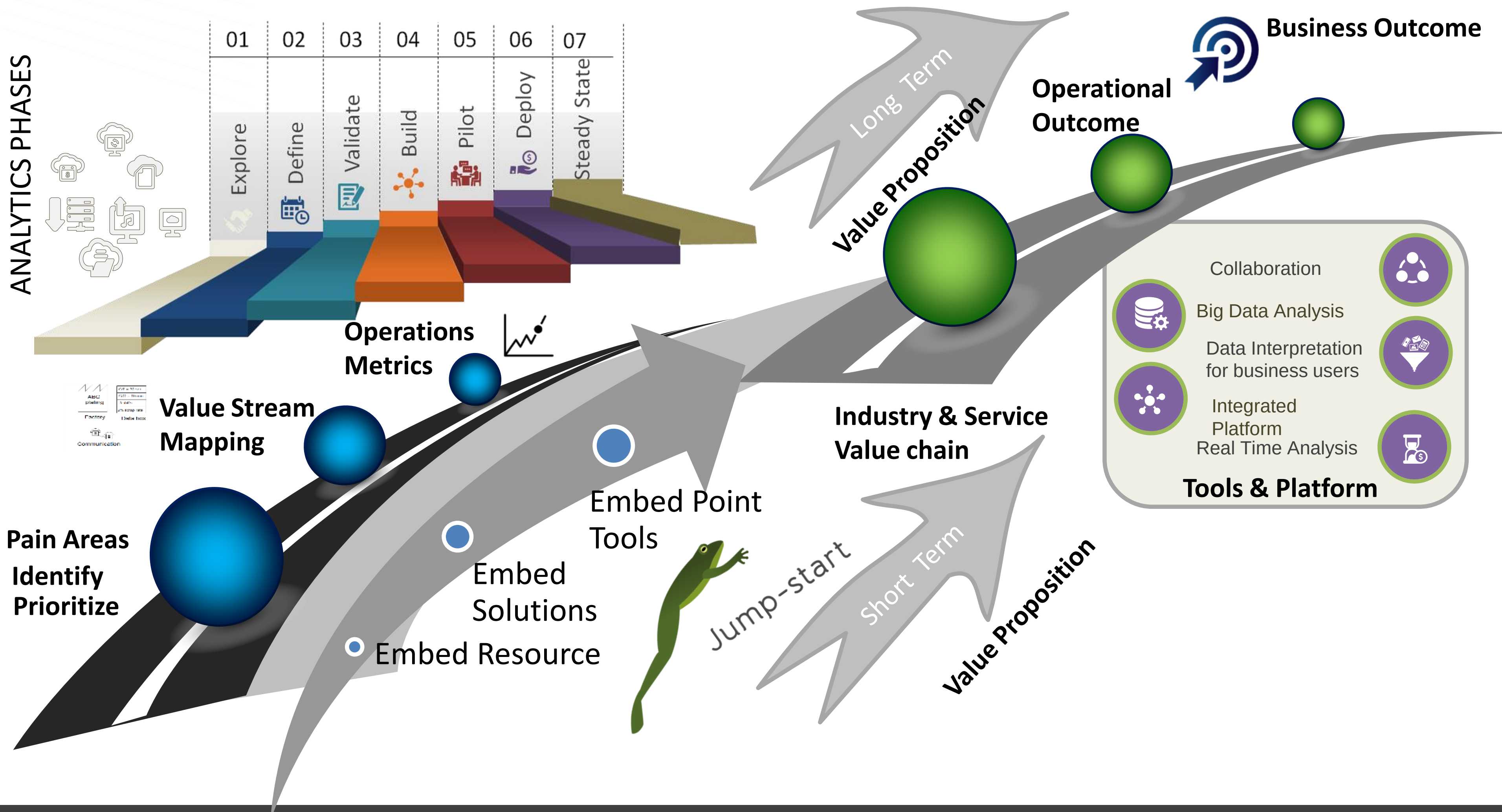
Translate Efficiency to measurable business impact.
Business user buy in & Signoff of on impact
Explore and improve new process

4 Measure Success

Measure the Model Quality
Measure the Process TAT & other SLA improvement
Measure Automation performance & success

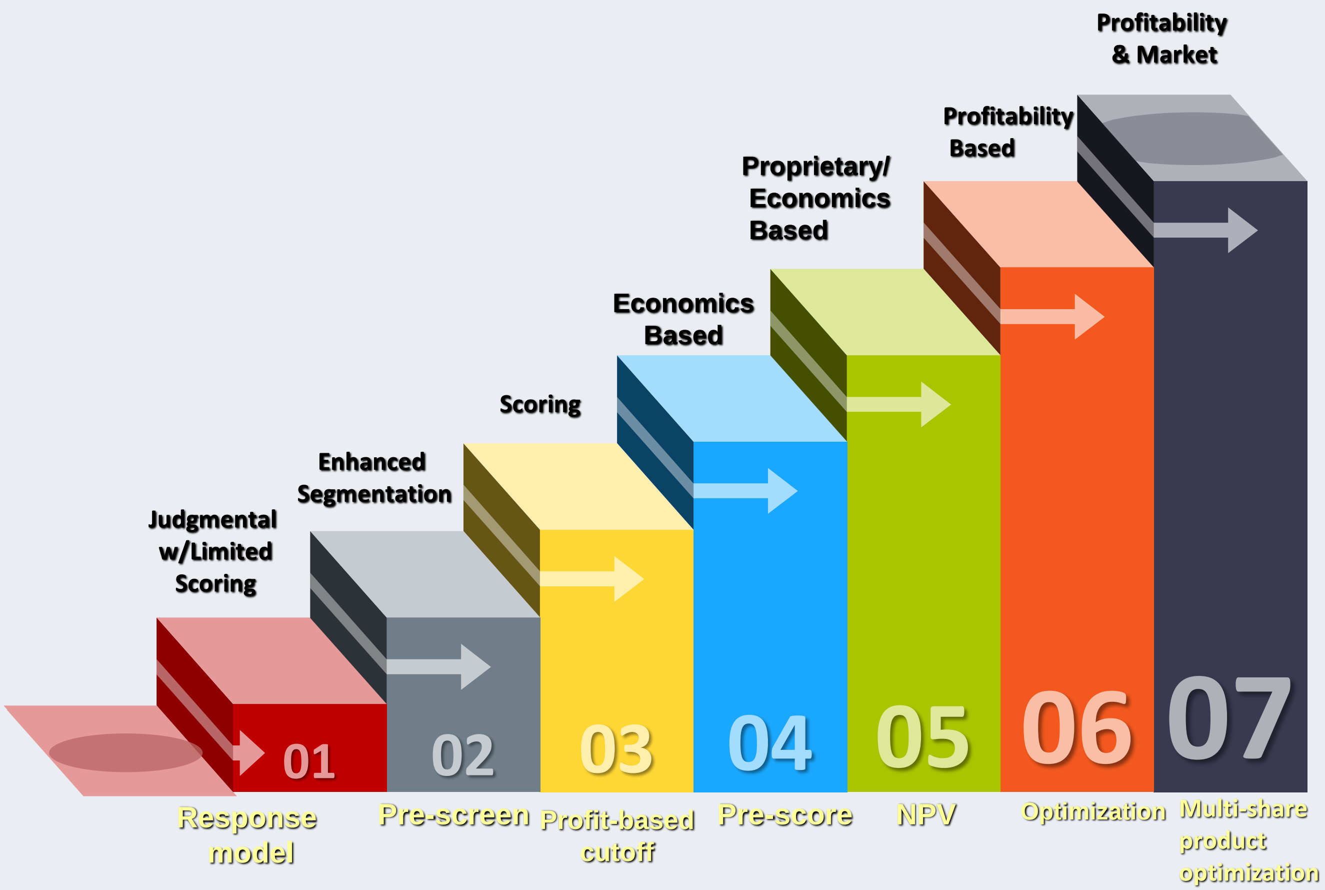
Embedded Analytics driving Business Impact

Strategy Driven Implementation



Credit Decisioning – Analytics Evolution

- ✓ Credit granting decisions were solely based on judgmental approach (5C's :Character, capital, collateral, capacity, conditions)
- ✓ The arrival of credit cards in late 1960's provided the impetus to automate consumer lending decisions.
- ✓ The financial institutions realized the value of applying credit scoring algorithms leveraging availability of credit bureau information and faster computing power
- ✓ Acceptance of credit scoring was further promoted by passing of Equal Credit Opportunity Acts 1975 and 1976



Evolution of Acquisition Decision Support System

Thank You