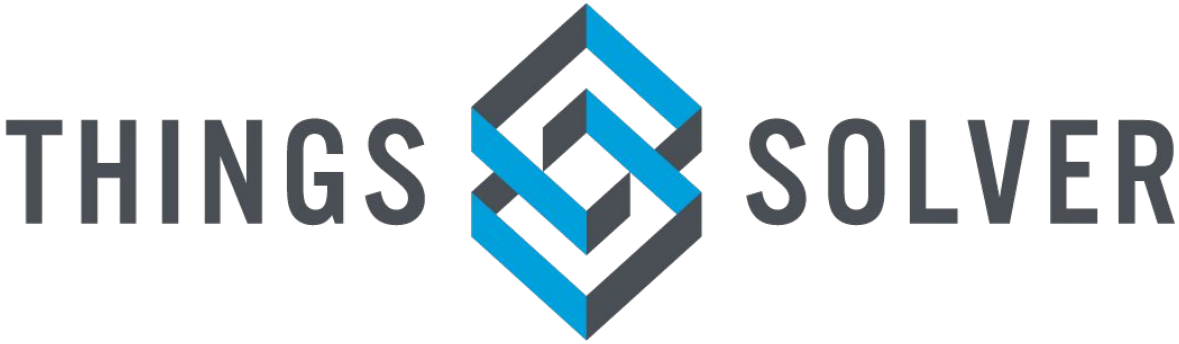


ENLIGHTEN YOUR DATA

Machine Learning Introduction

What is Data Science?

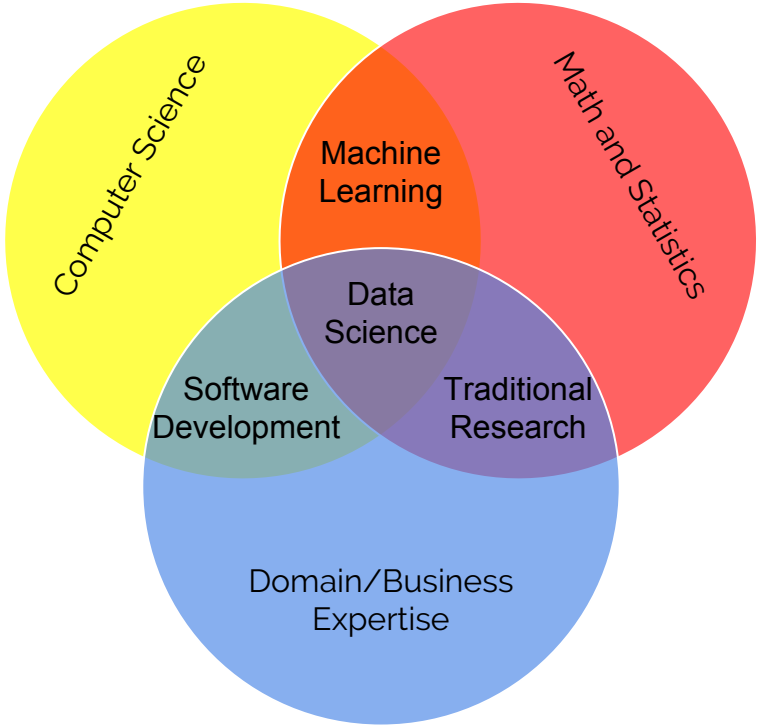
In essential, Data Science means **making sense about the world by using data**.

A buzzword typically used to describe the efforts of the companies and organizations in general to use the data to improve their performance and achieve goals.

Contrary to Data Analysts, the Data Scientists ask questions themselves driven by knowing which business goals are most important and how the data can be used to achieve certain goals for the organization. The communication is bottom up.

Data Science allows:

- Empowering management and officers to make better decisions with quantifiable, data-driven evidence, and testing these decisions.
- Directing the actions based on trends which in turn help in defining goals
- Challenging the staff to adopt best practices and focus on issues that matter.
- Identifying opportunities and target audiences.
- ...



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 tools e.g. Flare, D3.js, Tableau

MarketingDistillery.com is a group of practitioners in the area of e-commerce marketing. Our fields of expertise include: marketing strategy and optimization; customer tracking and on-site analytics; predictive analytics and econometrics; data warehousing and big data systems; marketing channel insights in Paid Search, SEO, Social, CRM and brand.

Marketing DISTILLERY

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 tools e.g. Flare, D3.js, Tableau

MarketingDistillery.com is a group of practitioners in the area of e-commerce marketing. Our fields of expertise include: marketing strategy and optimization; customer tracking and on-site analytics; predictive analytics and econometrics; data warehousing and big data systems; marketing channel insights in Paid Search, SEO, Social, CRM and brand.

Marketing DISTILLERY

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

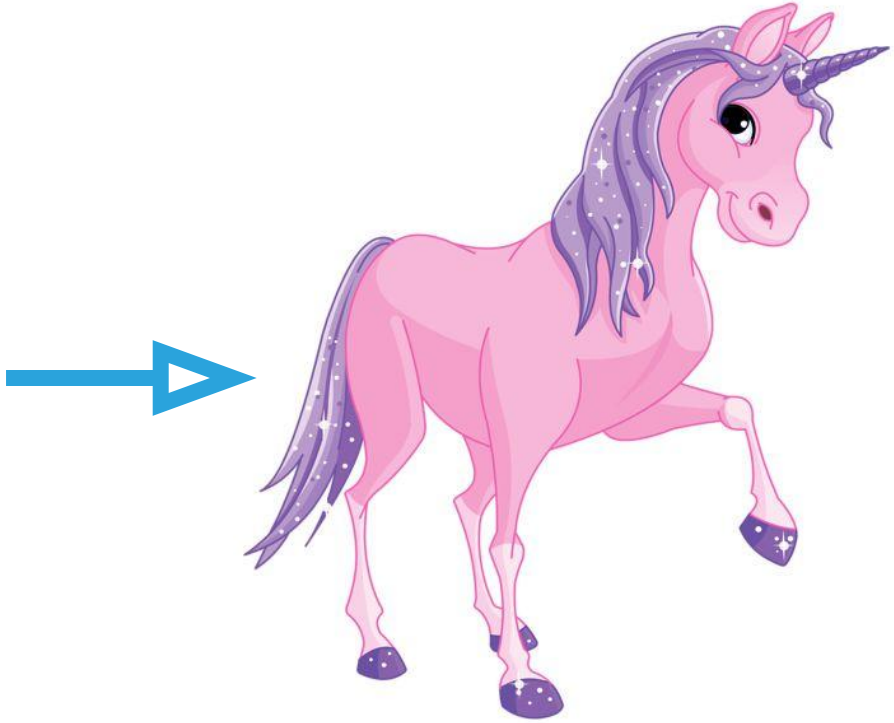
- Passionate about the business
- Curious about data
- Influence without authority
- Hacker mindset
- Problem solver
- Strategic, proactive, creative, innovative and collaborative

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 tools e.g. Flare, D3.js, Tableau

MarketingDistillery.com is a group of practitioners in the area of e-commerce marketing. Our fields of expertise include: marketing strategy and optimization; customer tracking and on-site analytics; predictive analytics and econometrics; data warehousing and big data systems; marketing channel insights in Paid Search, SEO, Social, CRM and brand.

Marketing DISTILLERY



THINGS SOLVER

Importance of Machine Learning

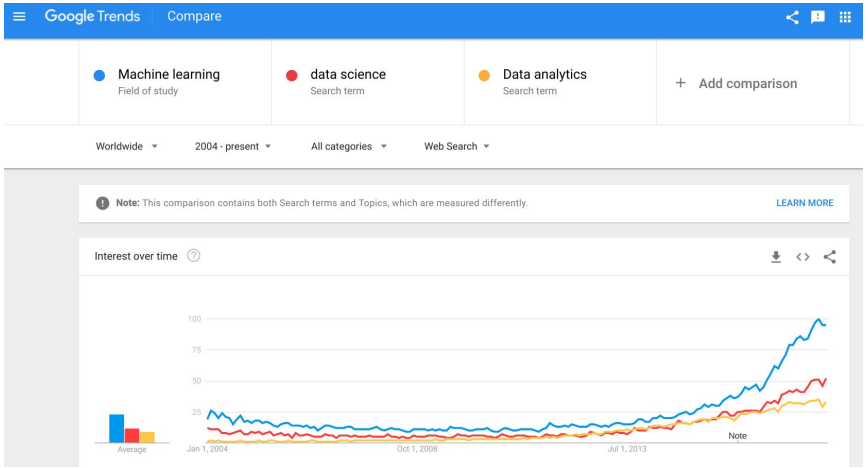
Trying to make a computer smart enough to learn from the data it's fed so that after a point of time the computer is able to predict further data.

Machine learning is given so much importance because it helps in **predicting behaviour** and **recognising patterns** that humans with their limited capacity can not predict.

The output is the machine that's capable of catching hold of patterns and predicting future events based on it.

Machine learning has several very practical applications that drive the kind of real business results – such as time and money savings – that have the potential to dramatically impact the future of the organizations.

Machine learning has made dramatic improvements in the past few years, but we are still very far from reaching human performance. Many times, the machine needs the assistance of human to complete its task.



Practical Examples

Self-driving cars.

Cyber fraud detection.

Online recommendation engines.

Stock and market predictions.

Smart Cities.

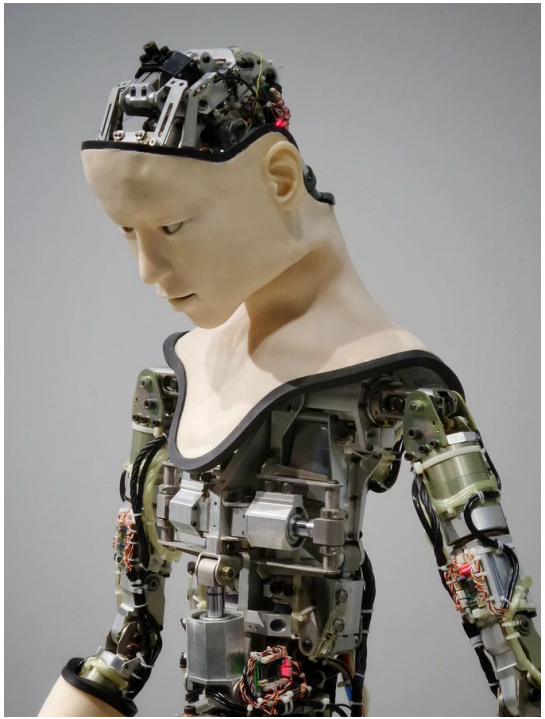
Personalised Healthcare.

Predictive Maintenance.

...

Artificial Intelligence

Artificial intelligence is the theory and development of computer systems **able** to perform tasks normally requiring human intelligence, such as: visual perception, speech recognition, decision-making and translation between languages.



Data Science ≠ Machine Learning ≠ Artificial Intelligence

The main differences:

- Data science produces **insights**
- Machine learning produces **predictions**
- Artificial intelligence produces **actions**

(David Robinson, Chief Data Scientist at DataCamp)


What are the benefits?

How various companies are using Data Science?

NETFLIX

Collaborative Filtering to recommend movies to users based on movies they have previously watched.

UBER

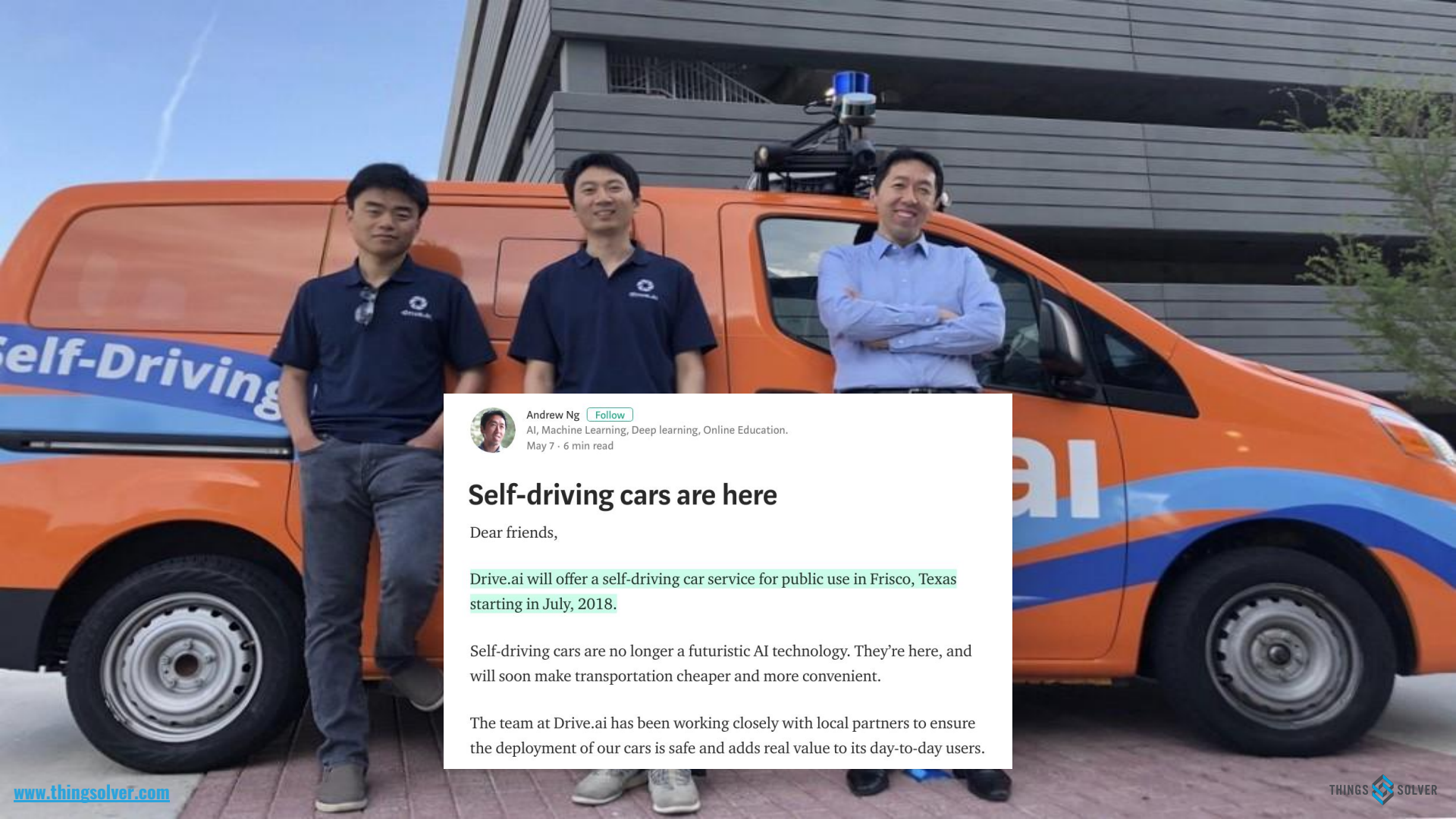
Implement models to follow the business logic behind how you make pricing decisions.

Dynamic fare model based on both geolocation and demand (for a ride) to help driver partners be more efficient.

Pornhub

Deploy facial recognition software that will detect 10,000 individual porn stars and tag them in videos.

As a second phase - using the software to identify the specific categories videos belong to.



Andrew Ng Follow
AI, Machine Learning, Deep learning, Online Education.
May 7 · 6 min read

Self-driving cars are here

Dear friends,

Drive.ai will offer a self-driving car service for public use in Frisco, Texas starting in July, 2018.

Self-driving cars are no longer a futuristic AI technology. They're here, and will soon make transportation cheaper and more convenient.

The team at Drive.ai has been working closely with local partners to ensure the deployment of our cars is safe and adds real value to its day-to-day users.

Coeus - Machine Learning for Retail

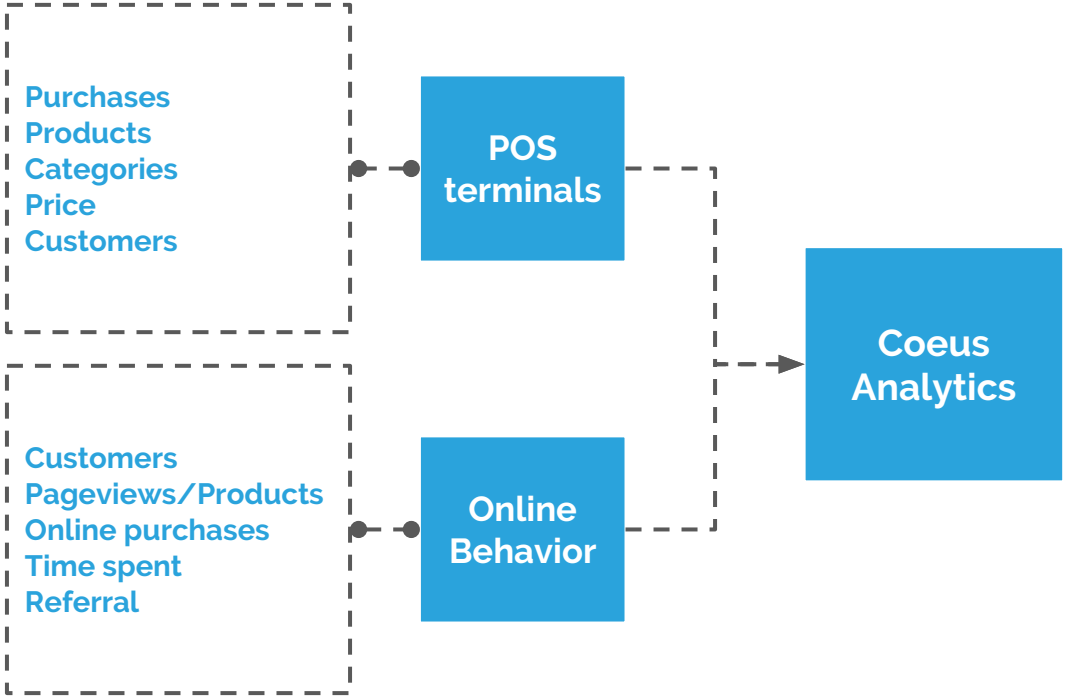
Coeus is an unique platform for scalable data streaming and processing, and advanced analytics, offering predictive models for:

- Propensity to purchase
- Association rules
- Customer Lifetime Value
- Recommendation system
- Customer segmentation

The goal of the project is to provide (across industries):

- Better understanding of the customer
- Loyal customer base and better offer towards the customer
- More engaged customers
- Better customer journey and experience

upsell | cross sell | customer analytics



Customer Segmentation

Group similar customers together based on their behavior (both online and offline) to facilitate:

- better campaign management
- identify the group of customers for the specific product
- segments: champions, loyal, recent, potentially loyal, hibernating, lost,...

Propensity to purchase

Predict whether the customer will make a purchase in the following period, which highly depends on frequency of purchases on a customer level.

Which customers are most likely to make a purchase in the next period, based on historical behavior and similar customers.

Customer Lifetime Value

Calculate how valuable the customer is to the company, and what is the expected income from the customer in the future period.

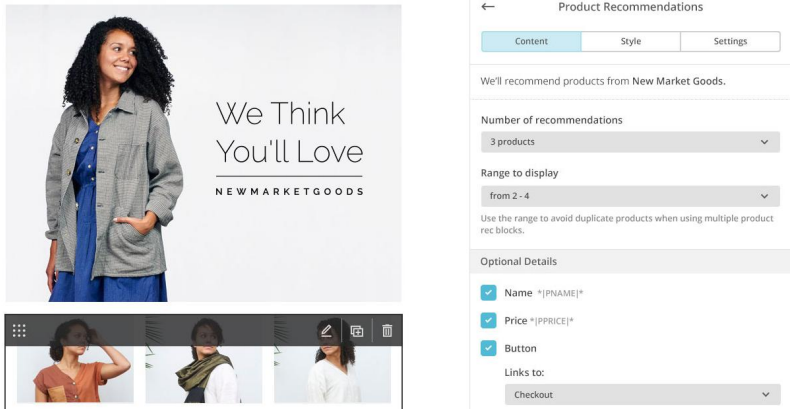
Recency - Frequency - Monetary value

Applied in each industry.

Recommendation System

Tailoring the next best offer towards the customer based on past behavior and similar customer purchases.

Online behavior (shop searches) can be used as an insight that the customer is interested in purchasing the product from the recommended category, and as a trigger to send the campaign and offer the recommended product.



Association rules

Discover interesting relations between variables.

Typically determine which products are usually purchased together.

As a technique can be applied to other problems, such as Predictive Maintenance.

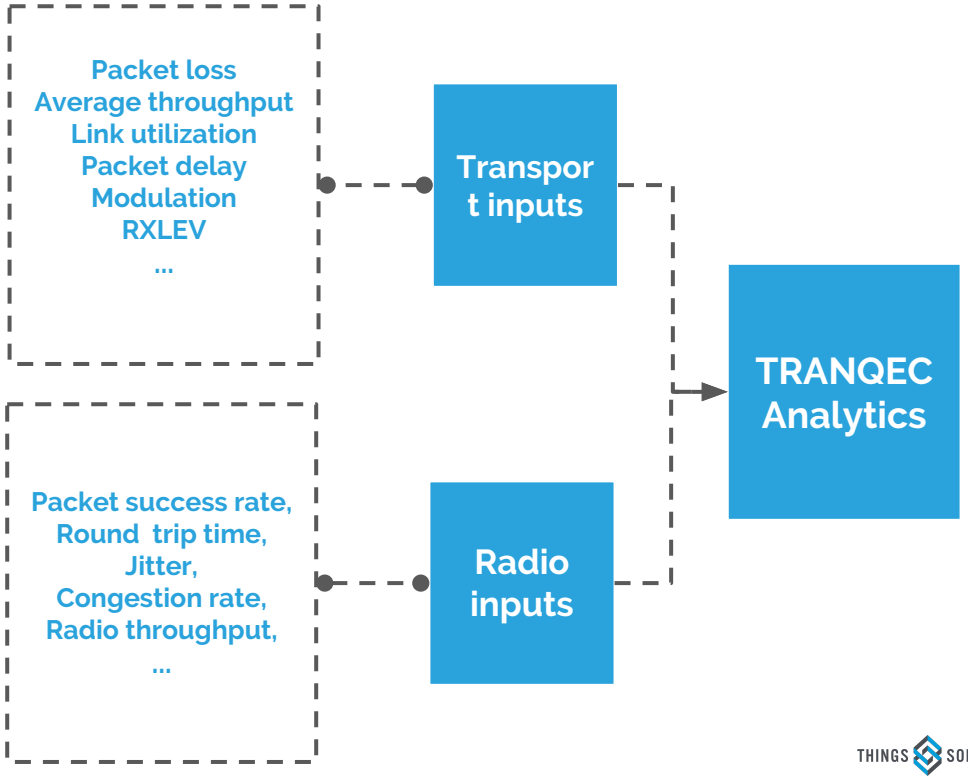
TRANQEC - Machine learning for Transport

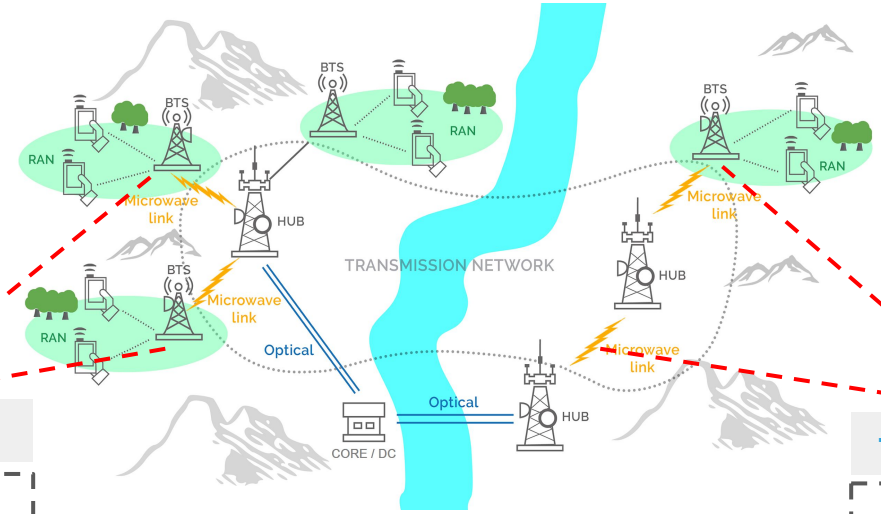
TRANQEC is a ML tool used for network transport analysis and problem detection, consisted of two main modules:

- Exploratory analysis
  - ❑ correlation analysis
  - ❑ association analysis
  - ❑ anomaly detection

- Predictive analysis
  - ❑ problem prediction
  - ❑ throughput short term forecast
  - ❑ throughput long term forecast

24/7 health check	problem prevention	costs optimization





Radio inputs

Packet success rate,
Round trip time,
Jitter,
Congestion rate,
Radio throughput,
Traffic intensity
...

Transport inputs

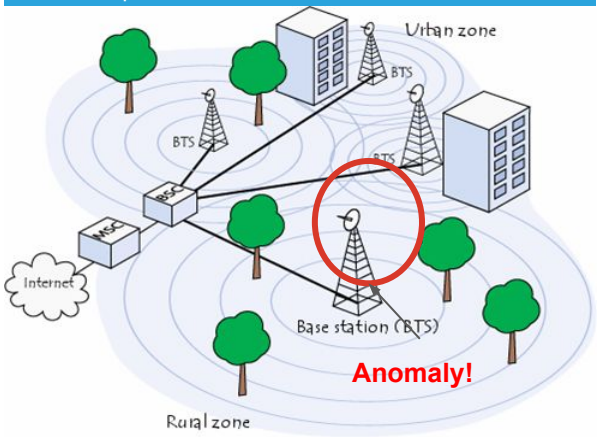
Packet loss
Average throughput
Link utilization
Packet delay
Link performances
...

TRANQEC Analytics

Problem Prediction

Predicting the problem occurrence on the site in the next period is crucial for transport and network optimization.

This module considers topology, performances and services, so it can efficiently determine whether there will be some link problem, and how widely it could spread.



Short term THP Forecast

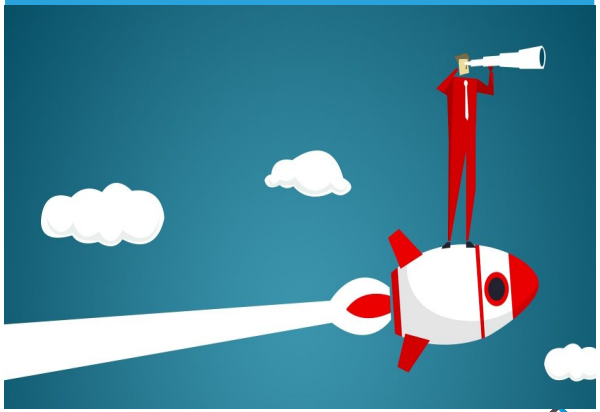
Short term forecast includes running a classical forecasting algorithm on a time series.

The forecasting period depends on the historical data available for training, but it is considered to be pretty accurate. The forecasting period length is around 3 to 6 weeks.

Long term THP Forecast

Long term forecast includes running a specially developed algorithm to obtain forecast for a longer period of time, such as 6-month, 2 years, 5 years, etc.

It makes predictions by analysing the whole network, and not the behaviour of only one cell at a time.



CAM: Customer Analytics Management

Online customer behavior provides important insights into customer behavior and preferences from different perspectives. Based on online customer behavior (aka customer activities on website) we can quite precisely determine that the customer is interested for some specific product or product category.

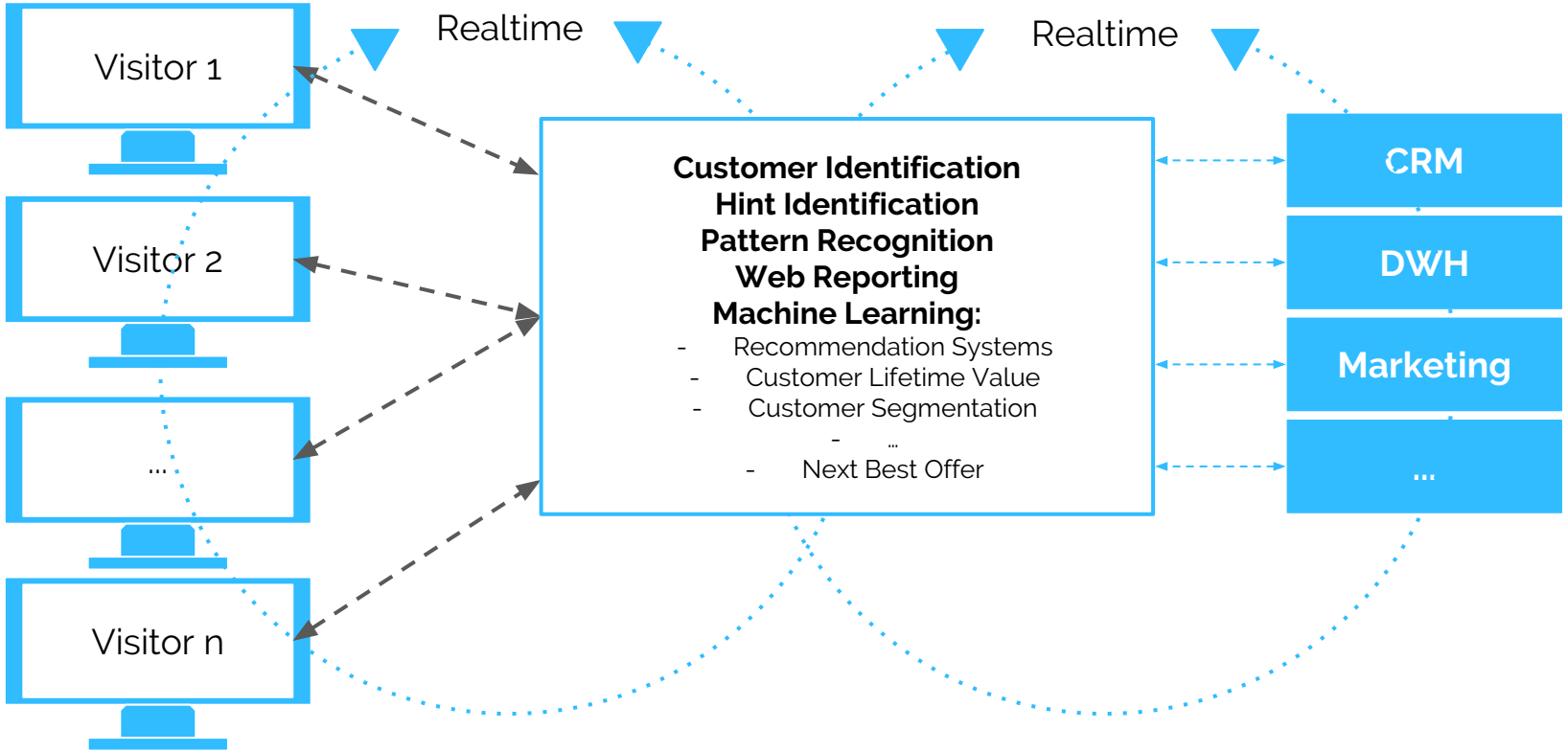
Variety of existing tools on the market provide Web analytics and reporting, but **catching specific event in a realtime is quite challenging with existing tools**. These particular events are making a difference in terms whether we will get to the client first.

Besides, getting raw data from existing platforms is either:

- Impossible
- Imprecise and hard to implement (ever tried getting raw data from Google Analytics? 😟)

The true power of online data is when combined with offline - in such a way there is a full picture about the customer behavior and preferences, and analytics is much more powerful..

CAM: Customer Analytics Management - Workflow



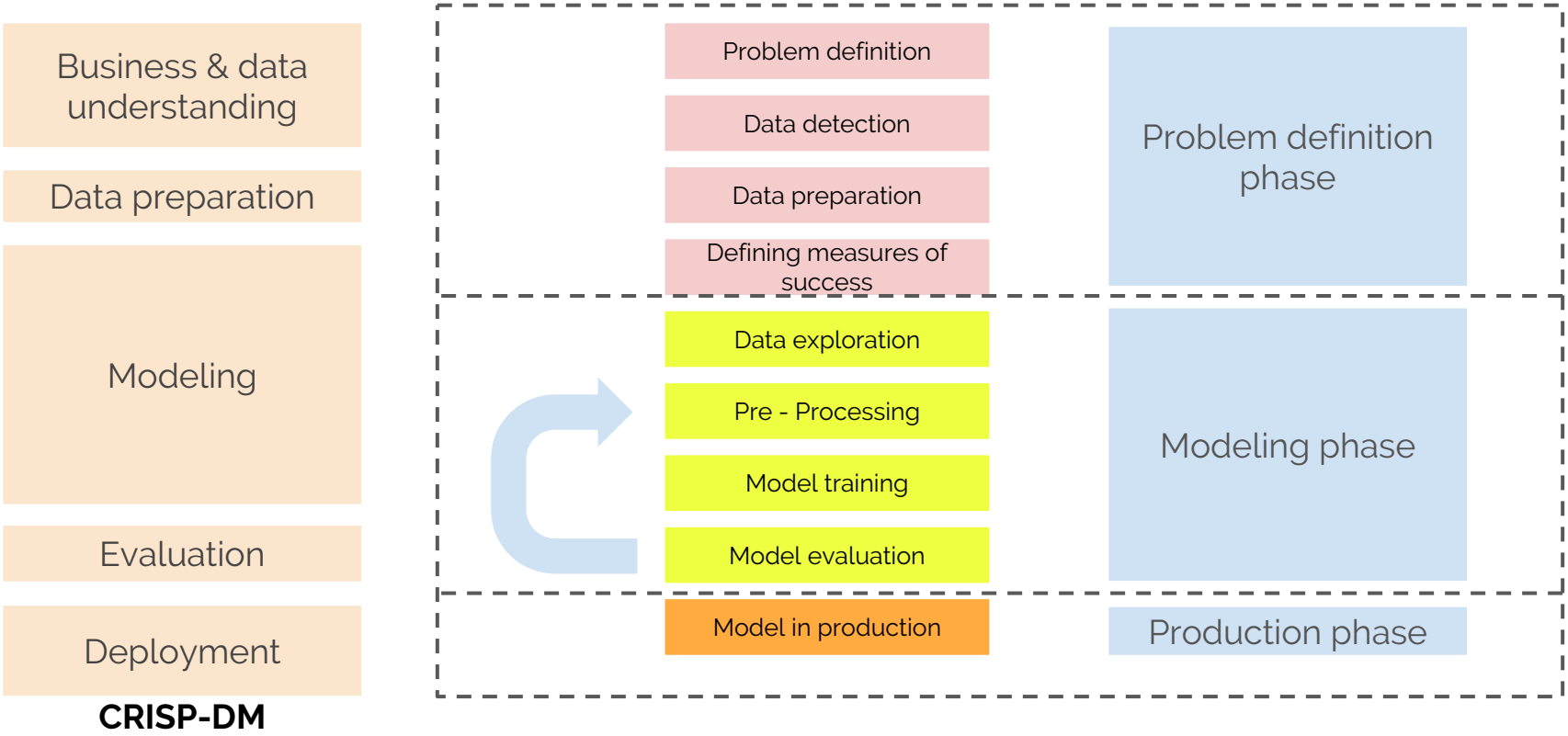
CAM: Customer Analytics Management - Example Scenario

A typical use case from telecommunication perspective:

1. A new visitor entered the web shop;
2. Customer is identified via cookie (personalized URL or login to the backend system);
3. Customer had 5 sessions in last 2 days with exploring new iPhone and available tariffs;
4. This information is registered (in CRM) in realtime;
5. Customer eligibility for this particular product is checked;
6. If eligible, a personalized offer or ad is placed to the client (on the same channel, during the same session);
7. Customer enters the renewal process.

** Similar scenario implemented in Banking.

Things Solver Data Science Workflow



Problem Definition

Problem definition is the crucial phase for a successful Machine Learning project (any project in general)..

Results of the most powerful and shiniest algorithms available will be meaningless if we are solving the wrong problem.

In this phase it is important to make the assumptions about the problem based on the domain knowledge, which are to be tested in the early phases of data exploration.

Step 1: What is the problem? Describe the problem informally and formally and list assumptions and similar problems.

Step 2: Why does the problem need to be solved? List your motivation for solving the problem, the benefits a solution provides and how the solution will be used.

Step 3: How would I solve the problem? Describe how the problem would be solved manually to flush domain knowledge.

- Dr. Jason Brownlee
  https://machinelearningmastery.com/how-to-define-your-machine-learning-problem/

Problem Definition

The step after defining the problem that is to be solved, is to identify the inputs for solving the problem, prepare the data and define the KPIs to measure the success of the projects.

Data detection	Data preparation	Define measures of success
Identifying the data inputs that are to be used to solve the problem and test the problem assumptions.	Prepare the data that is to be used in modeling phase in desired format, velocity and availability.	Define the numerical parameters (KPIs) that will measure the final result of the Machine Learning project.

Modeling

Modeling takes most of the development time and a pretty serious amount of patience.

Crucial prerequisite here is to prepare the data for future steps.

“Garbage in, garbage out”

Your analysis is as good as your data.

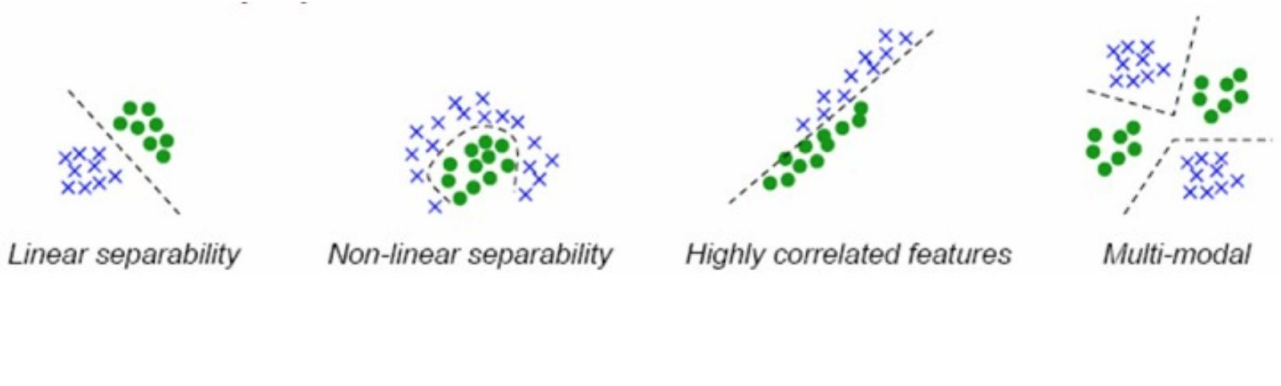
Step 1. What data am I working with? Understand the data. Try to gain maximum information from it and consider its constraints.

Step 2. What is the goal of this analysis? Define goals for each task in this phase. That would make the whole process easier and more efficient.

Step 3. Which models should I use? Define the flow of the modeling. If more models are to be used, define their performance measures in order to compare them.

Step 4. Is this the best I can get? Be patient and be prepared to re-train the model as many times as it is needed.

First step is data exploration...



Data Exploration

Statistics.

Variable exploration.

Univariate analysis.

Bivariate analysis.

Missing values treatment.

Outlier treatment.

Variable transformation.

Visualization.

Distribution plots.

Correlation plots.

Box plots.

Pre - Processing

Data Cleaning

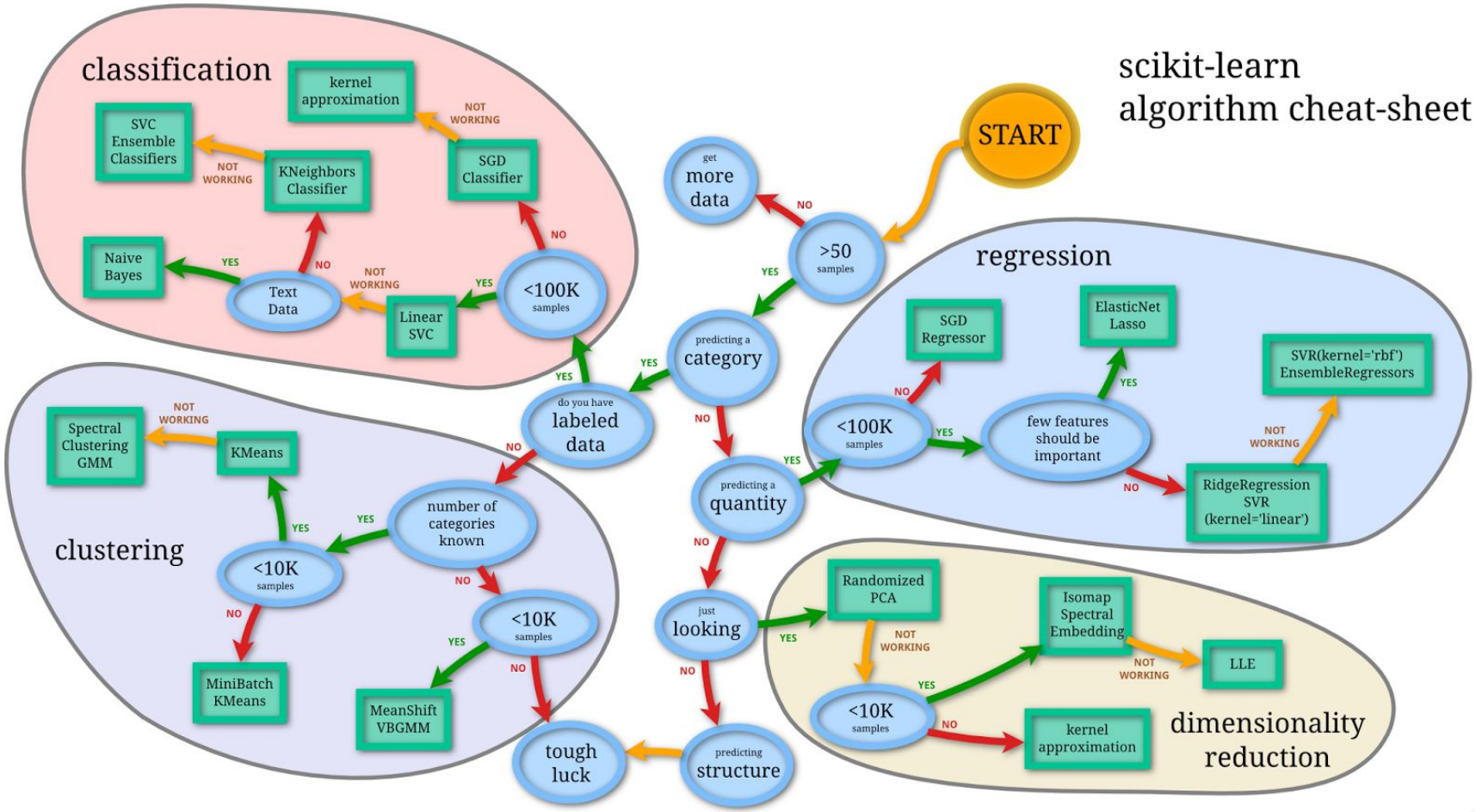
The process of detecting and correcting corrupt or inaccurate records. It includes identifying incomplete, incorrect, inaccurate or irrelevant parts of the data and then replacing, modifying, or deleting the dirty or coarse data.
Data comes from various systems and any many cases can be inaccurate.
Picking the right technique for handling inaccurate data is a quite sophisticated technique and requires domain expertise.

Feature Engineering

Create additional relevant features from the existing raw features in the data, and to increase the predictive power of the learning algorithm. The output of this phase enriches the model inputs.
Feature engineering is fundamental to the application of machine learning, and is both difficult and expensive.
Feature engineering represents identifying characteristics that might help when solving the problem.

Machine Learning Models

1. **Supervised models** (modeling labeled data) - during the training phase, a model learns to predict one of the possible outcome classes (or values) based on a set of input features
    a. classification
    b. regression
2. **Unsupervised models** (modeling unlabeled data) - during the training phase, a model learns behavior characteristics based on a set of input features
    a. clustering
    b. anomaly detection
    c. market basket analysis
3. **Semi-supervised models** (modeling partially labeled data) - during the training phase, a model learns one type of behavior (most likely normal or expected) based on a set of input features
    a. anomaly detection
    b. classification


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

Supervised Models

Classification models

1. Naive Bayes
2. Decision Trees
3. Logistic Regression
4. Random Forest
5. Neural Networks

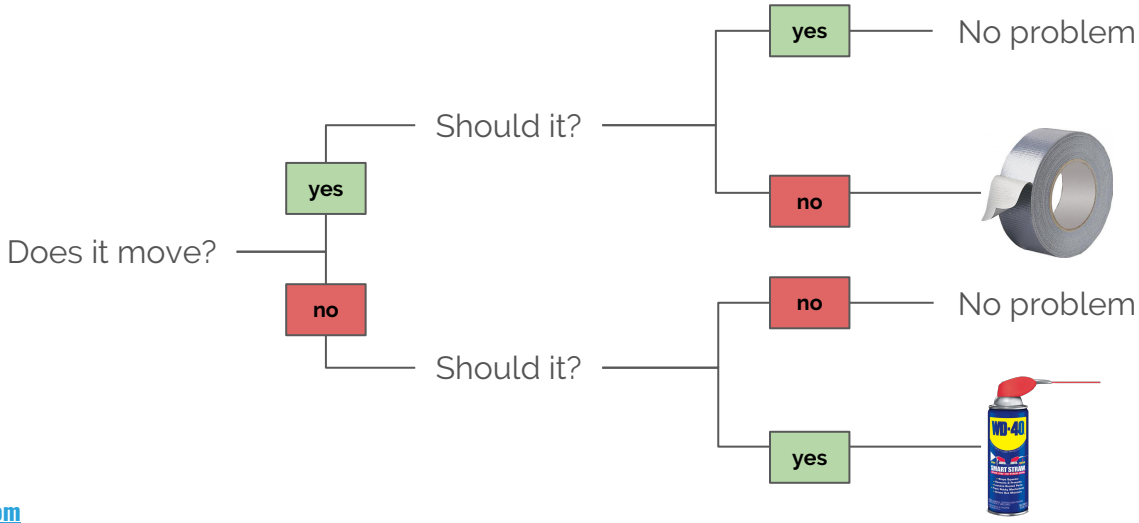
Regression models

1. Linear Regression
2. Random Forest Regressor
3. Neural Networks

Supervised models require LABELED DATA in order to work!

Decision Trees

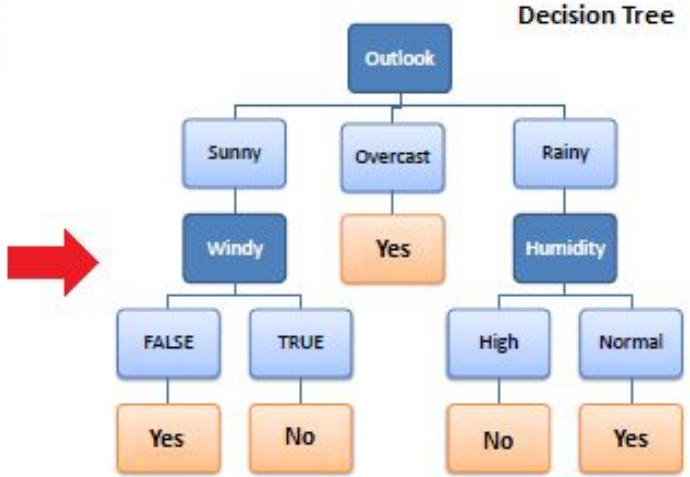
- a flowchart-like structure
- each internal **node** represents a "test" on an attribute (e.g. whether humidity is high or normal)
- each **branch** represents the outcome of the test
- each **leaf** node represents a class label (decision taken after computing all attributes)
- the paths from root to leaf represent **classification rules**



Decision Tree example

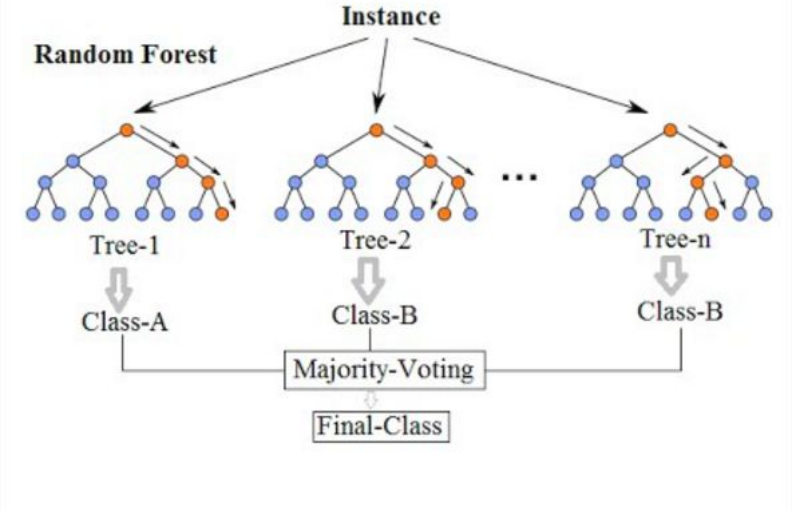
predictors | target

outlook	temperature	humidity	windy	play golf
Rainy	Hot	High	False	No
Rainy	Hot	High	True	No
Overcast	Hot	High	False	Yes
Sunny	Mild	High	False	Yes
Sunny	Cool	Normal	False	Yes
Sunny	Cool	Normal	True	No
Overcast	Cool	Normal	True	Yes
Rainy	Mild	High	False	No
Rainy	Cool	Normal	False	Yes
Sunny	Mild	Normal	False	Yes
Rainy	Mild	Normal	True	Yes
Overcast	Mild	High	True	Yes
Overcast	Hot	Normal	False	Yes
Sunny	Mild	High	True	No



Random Forest

- building an ensemble of decision trees
- the predicted class is determined by taking the most frequently predicted one, while regressed values are calculated as the average of all predictions
- excellent for controlling the over-fitting



Common Neural Networks Application

Image Processing/Visual Search. (**Ebay. Facebook.**)

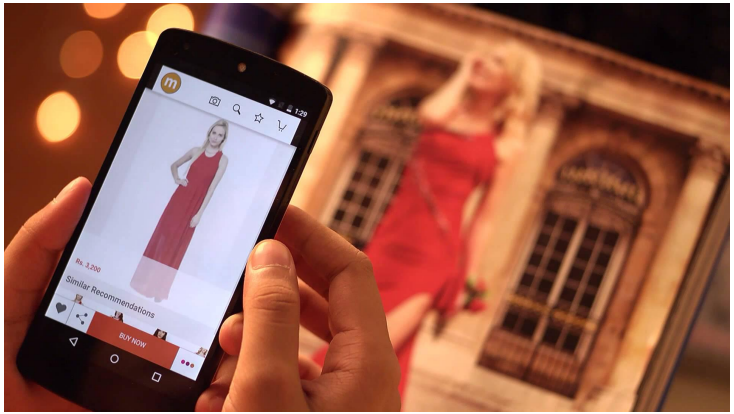
Computer Vision. (**NASA - Mars Exploration Rover. ESA ExoMars Rover.**)

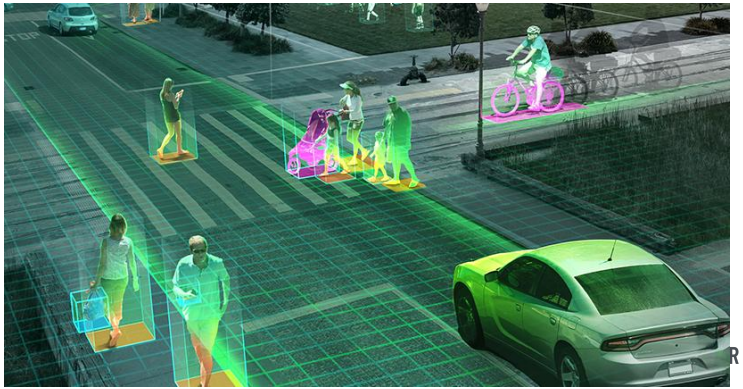
Speech Recognition. (**Google Assistant. Siri. Amazon Alexa. Cortana.**)

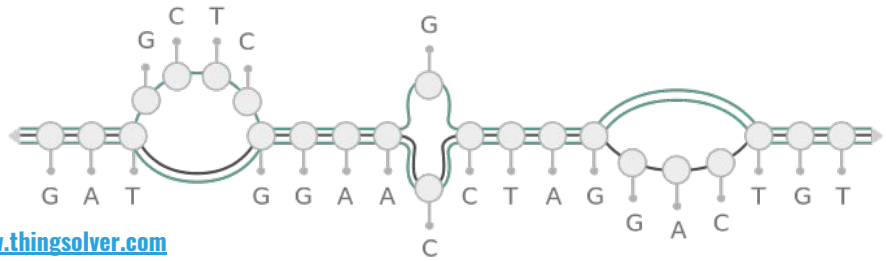
Vehicle control. (**Tesla. Boston Dynamic.**)

Game playing. (**Google Mind. OpenAI.**)

Healthcare (cancer research). (**Seven Bridges Genomics.**)







Unsupervised Models

Clustering models

1. k-Means
2. DBSCAN
3. Agglomerative clustering

Anomaly detection models

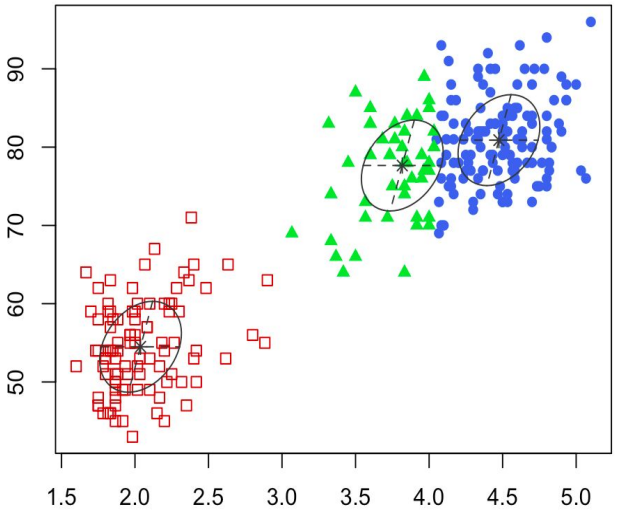
1. Isolation Forest
2. Autoencoders
3. Elliptic Envelope
4. One class SVM

Market basket analysis

1. Association rules

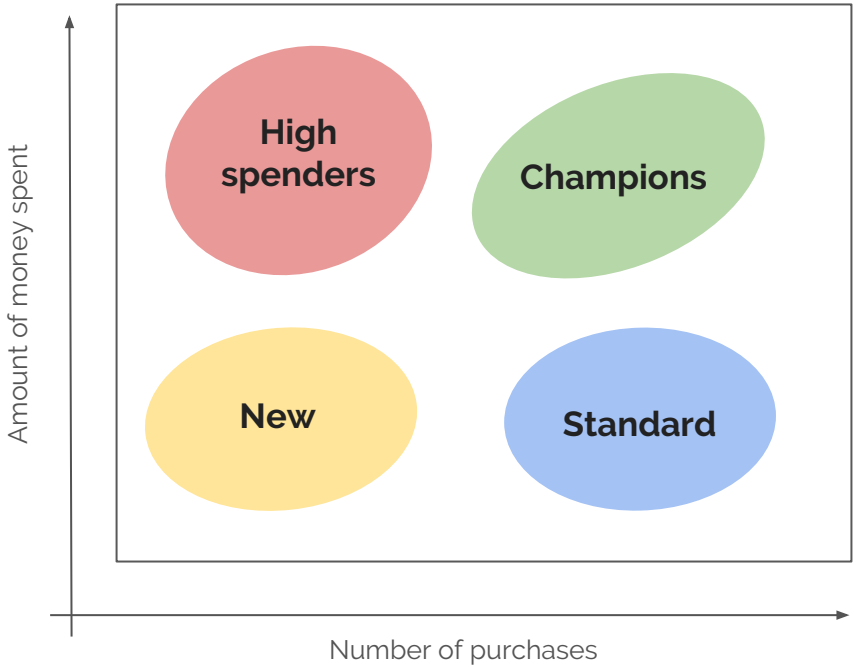
K-Means

- based on a set of input features, it groups instances with similar behavior and tries to separate those with a different one
- it requires initial number of clusters
- at the beginning, it randomly chooses centroids (cluster centers)
- in the training phase, the distance of each instance from each cluster centroid is calculated, and the instance is joined to the closest cluster
- cluster centroid is re-calculated as an average of all instances in a cluster
- the upper three steps are repeated until some criteria is met - maximal number of iterations reached/light or none cluster grouping improvement

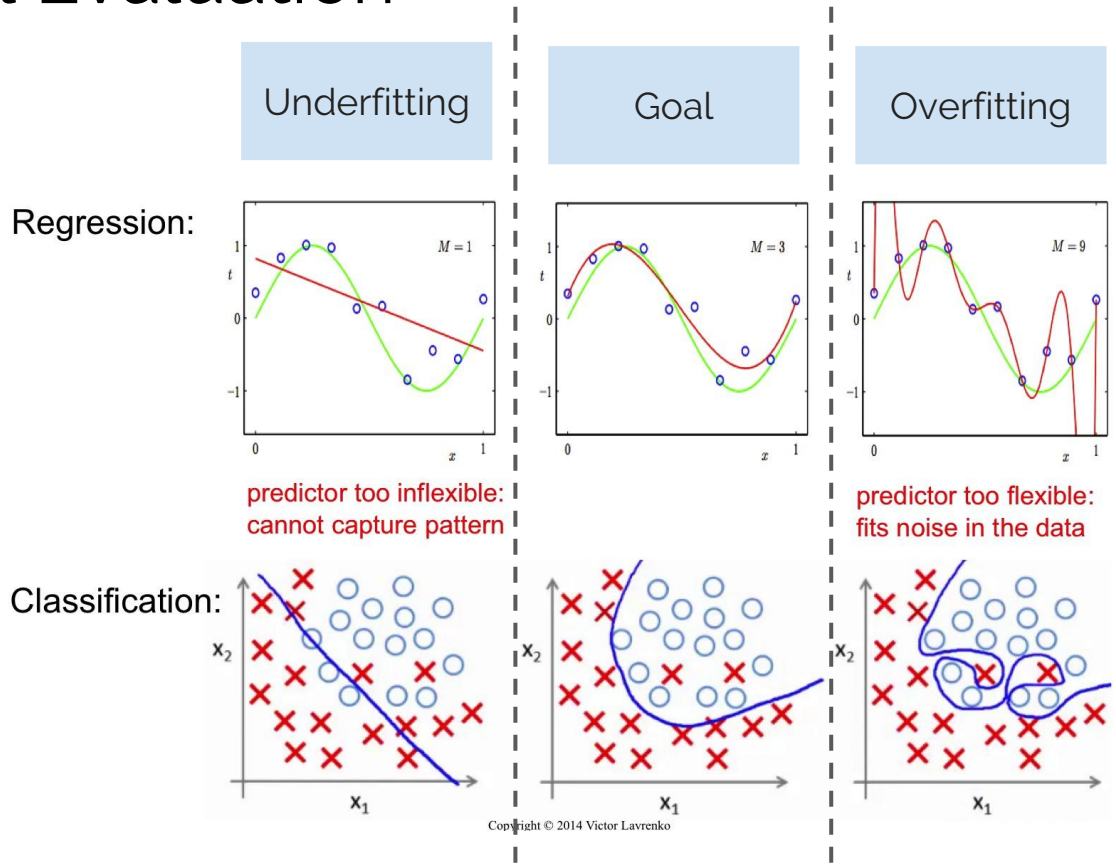


K-Means example

Identify customers segments based on their purchasing behaviour. Input data contains historical data for number of purchases and amount of spent money for each customer.



Model Evaluation



Big Data - Managing Data

"*Big data is high-volume and high-velocity and/or high-variety information assets that demand cost-effective, innovative forms of information processing that enable enhanced insight, decision-making, and process automation.*" - Gartner

Big Data is a term describing the approach how the data is being managed - putting massive volumes of structured and unstructured data and organizing in such a way that the analytics can be performed efficiently.

MANAGING DATA

Advanced Analytics/Data Science - Extract insights

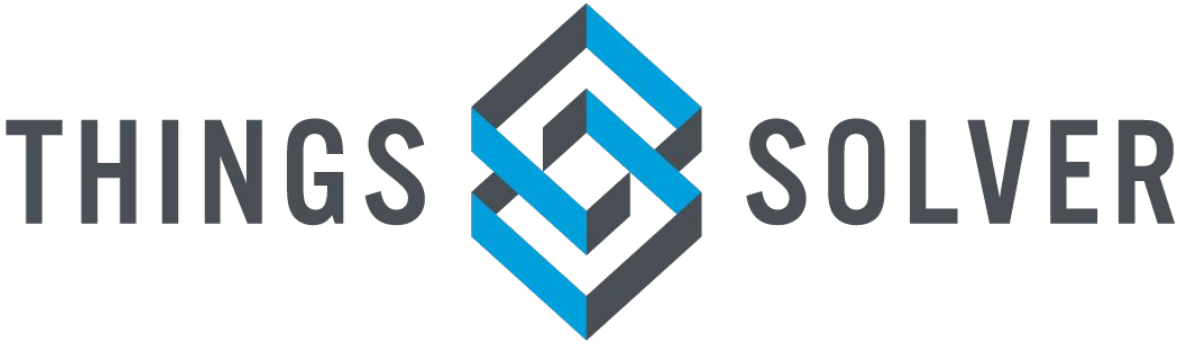
Analytics includes combining statistics, mathematics, programming, problem-solving, capturing data in ingenious ways, the ability to look at things differently to find patterns, along with the activities of cleansing, preparing, and aligning the data, in order to extract valuable insights from the data.

The term refers to both **small** data and **big** data.

Advanced Analytics utilizes the data managed by Big Data solution.

GETTING INSIGHTS FROM DATA

Q&A



ENLIGHTEN YOUR DATA