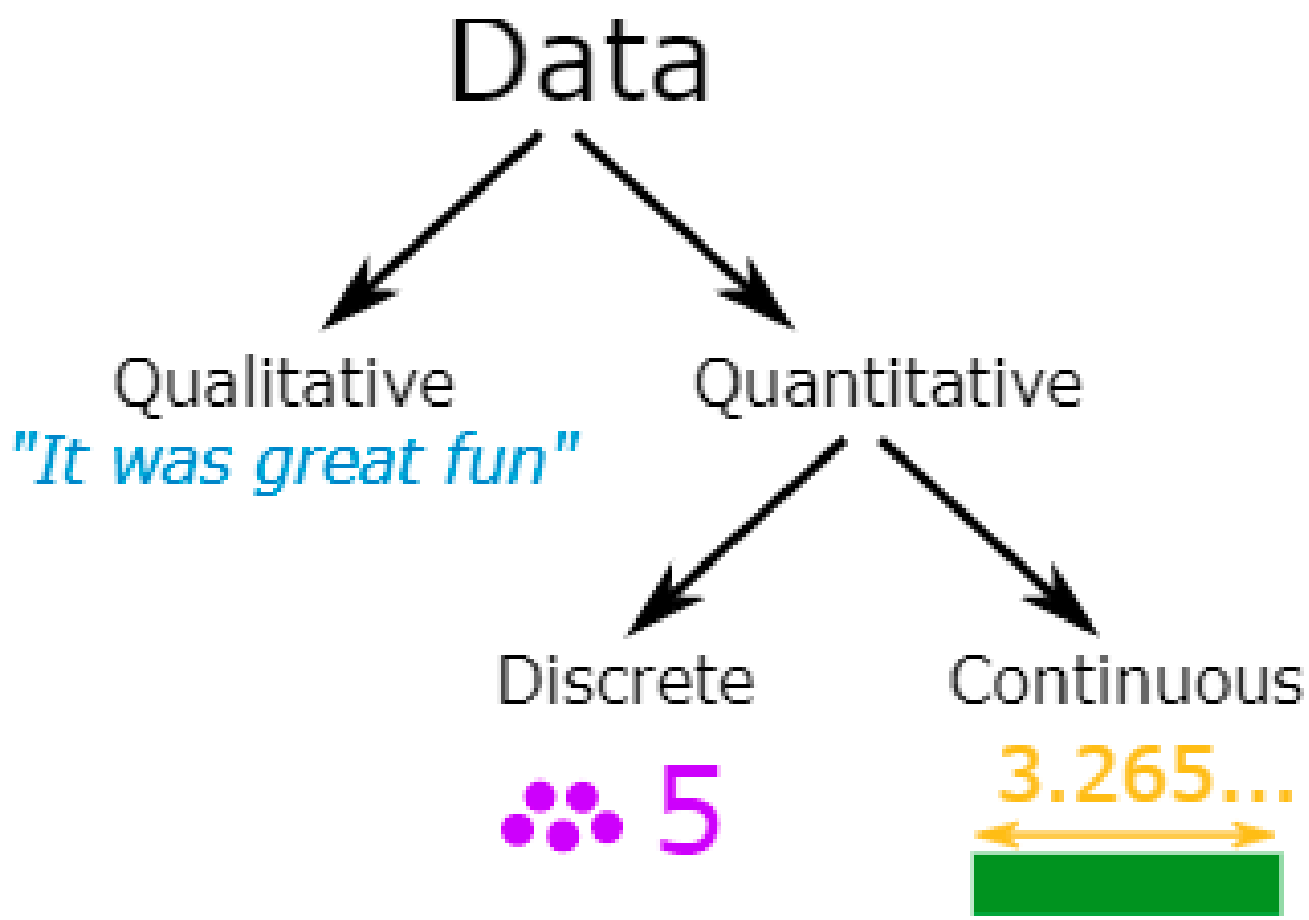
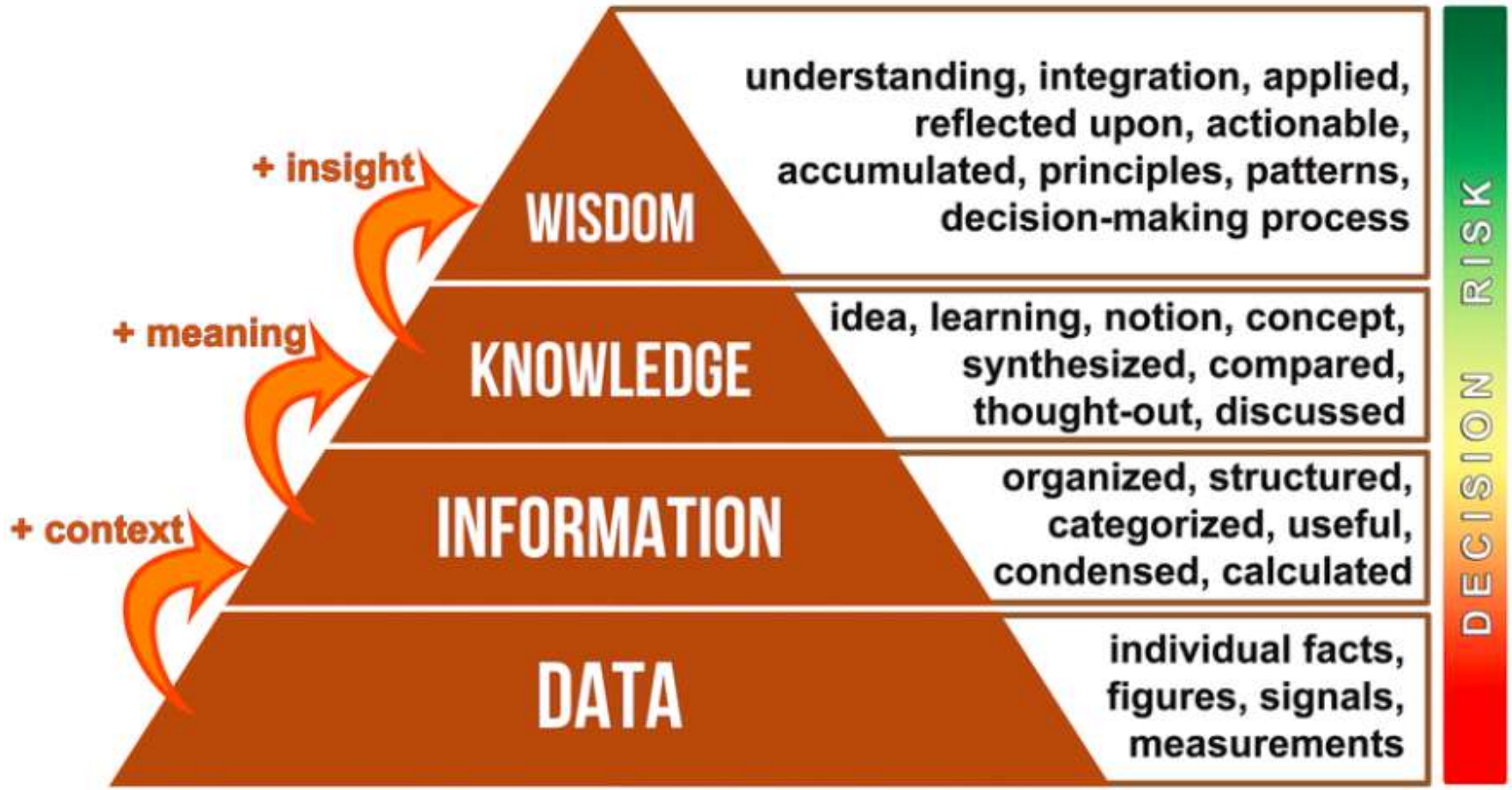
What is data?

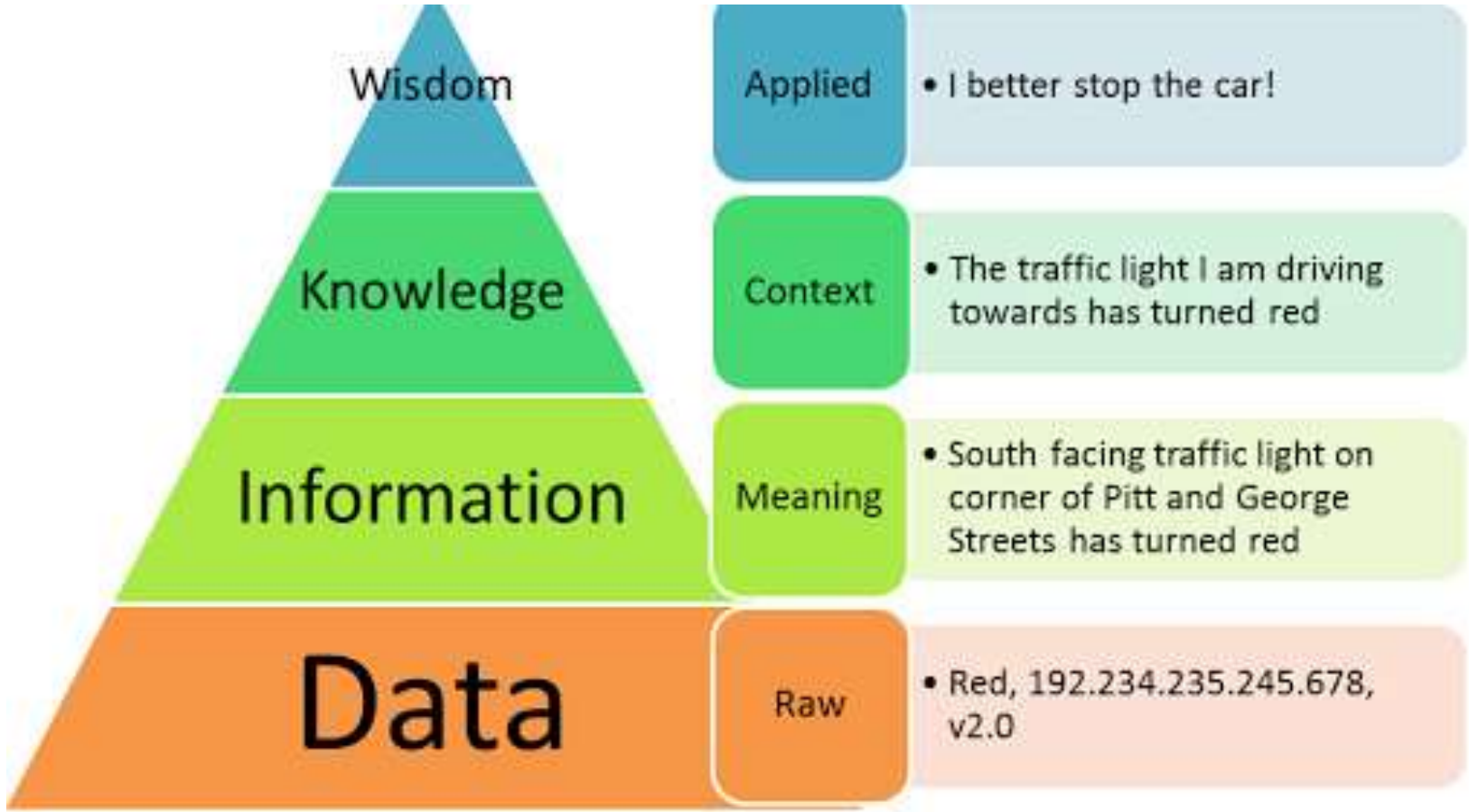
- Data is a collection of facts, such as numbers, words, measurements, observations or just description of things.
- Data is plain facts. The word “data” is plural for “datum.” When data is processed, organized, structured or presented in a given context so as to make it useful, its called Information.
- Data itself is fairly useless, but when this data is interpreted and processed to determine its true meaning, it becomes useful and can be named as Information.

Qualitative vs Quantitative Data

- Data can be qualitative or quantitative.
- Qualitative data is descriptive information (it describes something)
- Quantitative data is numerical information (numbers)
  - Discrete data can only take certain values (like whole numbers)
  - Continuous data can take any value (within a range)

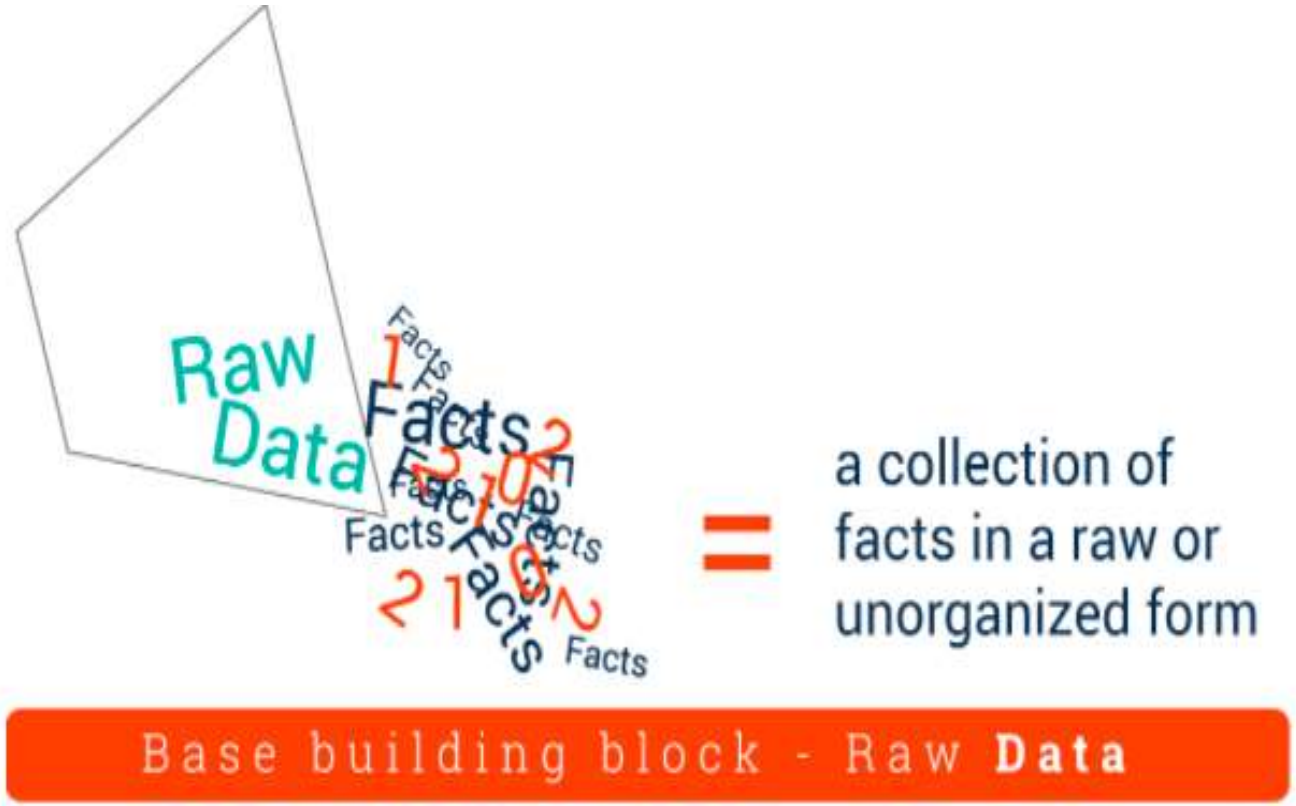



+ insight
+ meaning
+ context
WISDOM
KNOWLEDGE
INFORMATION
DATA
understanding, integration, applied, reflected upon, actionable, accumulated, principles, patterns, decision-making process
idea, learning, notion, concept, synthesized, compared, thought-out, discussed
organized, structured, categorized, useful, condensed, calculated
individual facts, figures, signals, measurements
DECISION RISK


Wisdom
Knowledge
Information
Data
Applied
• I better stop the car!
Context
• The traffic light I am driving towards has turned red
Meaning
• South facing traffic light on corner of Pitt and George Streets has turned red
Raw
• Red, 192.234.235.245.678, v2.0

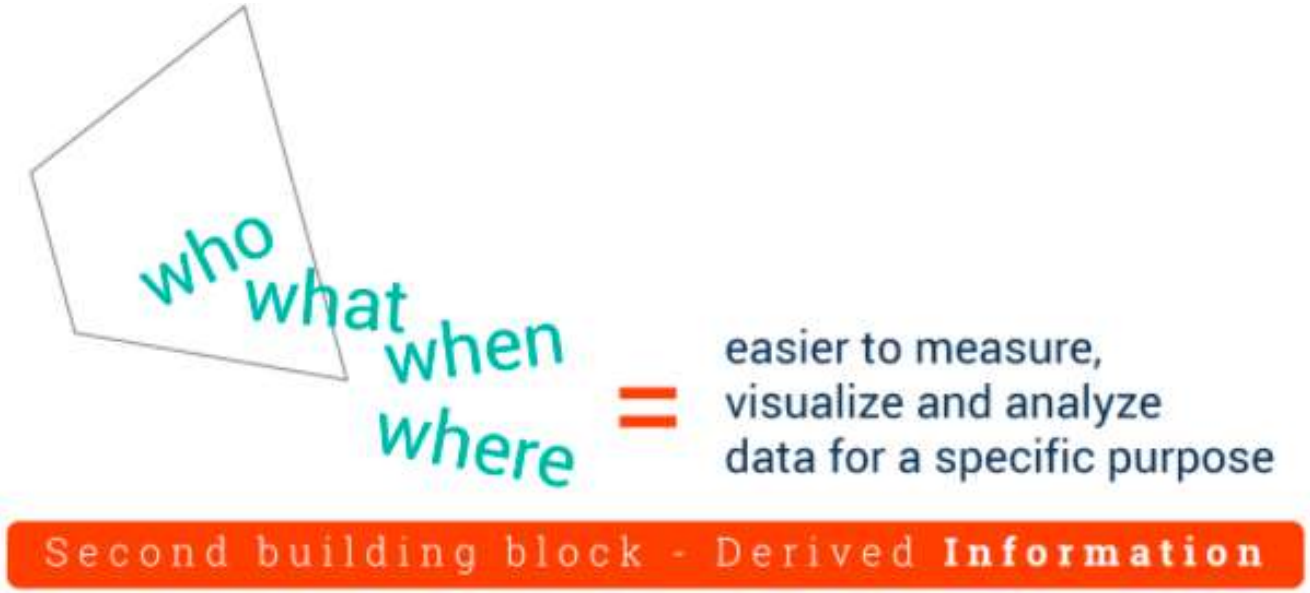
Data

- Data is a collection of facts in a raw or unorganized form such as numbers or characters.
- However, without context, data can mean little. For example, *12012012* is just a sequence of numbers without apparent importance. But if we view it in the context of ‘this is a date’, we can easily recognize *12th of January, 2012*. By adding context and value to the numbers, they now have more meaning.
- In this way, we have transformed the raw sequence of numbers into-INFORMATION



Information

- This is data that has been “cleaned” of errors and further processed in a way that makes it easier to measure, visualize and analyze for a specific purpose.
- Depending on this purpose, data processing can involve different operations such as combining different sets of data (aggregation), ensuring that the collected data is relevant and accurate (validation), etc.
- By asking relevant questions about ‘who’, ‘what’, ‘when’, ‘where’, etc., we can derive valuable information from the data and make it more useful for us.
- But when we get to the question of ‘how’, this is what makes the leap from information to-KNOWLEDGE



Knowledge

- “How” is the information, derived from the collected data, relevant to our goals?
- When we don’t just view information as a description of collected facts, but also understand how to apply it to achieve our goals, we turn it into knowledge.
- But only when we use the knowledge and insights gained from the information to take proactive decisions, we can say that we have reached the final – WISDOM

Wisdom

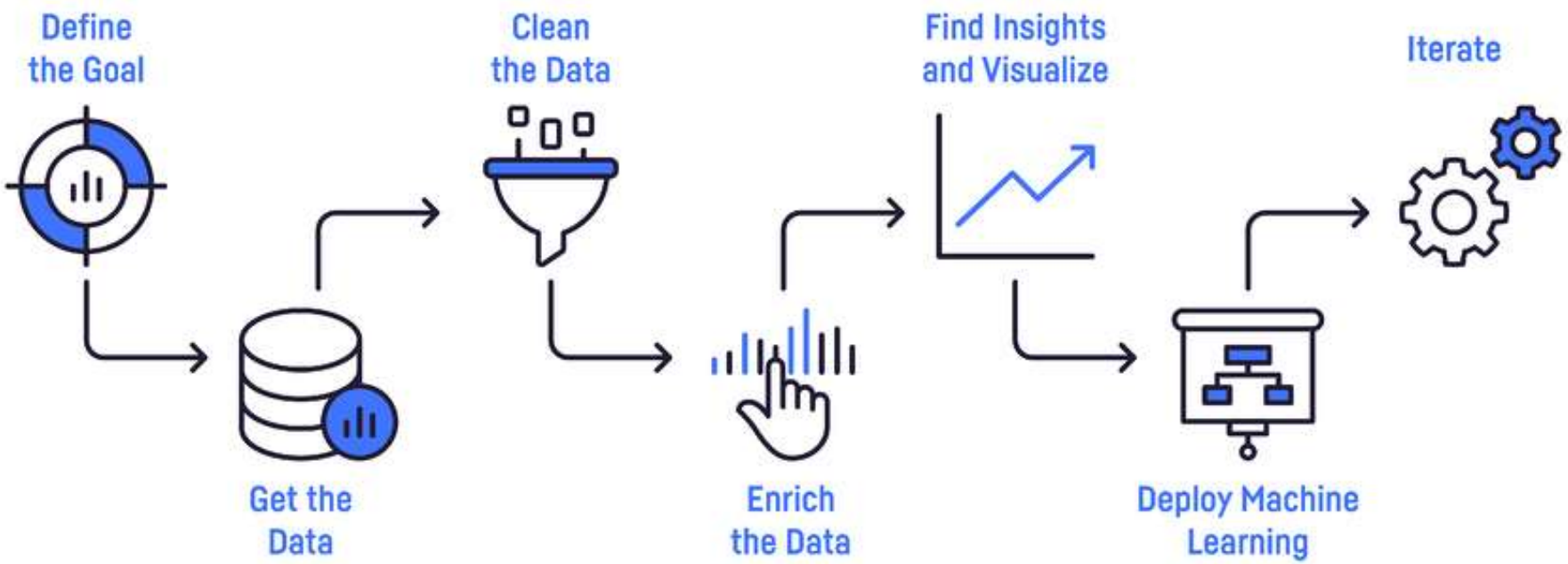
- Wisdom is the top of the DIKW hierarchy and to get there, we must answer questions such as ‘why do something’ and ‘what is best’. In other words, wisdom is knowledge applied in action.

- Datum (an observation): 12
- Data: a collection of observations.
- Information: 12 degrees Fahrenheit
- Knowledge: 12 degrees Fahrenheit, today, at 7:30 AM, in Summerville, Oregon
- Wisdom: I need to put on a coat when I feed the horses.

1. You must be familiar of Amazon or flipkart reviews which we give when we buy products .This reviews which they recieve is what you say DATA. They have terrabytes of data like this .
2. This data when processed to some meaningful form like how many female / male like a particular product or likely to buy a product .This is what you call is INFORMATION .
3. And this information when used to apply in business like during marriage season (dec-march) people are most likely to buy ethnic wear so they fill up the stock and prepare for this which help them to grow their business reducing the overall cost for them. This is what you call knowledge .

Click to add text

Steps in Data Analytics



Step 1: Define the goal

- To Understand the business
- Find an Interesting Topic
- Your project must be the answer to a clear organizational need so you should always concentrate on the overall scope and objective of the topic.
- For instance, if you are interested in Healthcare Analytics, there are many topics you can try-Lung cancer classification based on gene expression levels, EEG based emotion recognition in music listening, Breast cancer detection using anomaly classification.
- Simply downloading a cool open dataset is not enough. In order to have motivation, direction, and purpose, you have to identify a clear objective of what you want to do with data: a concrete question to answer, a product to build, etc.

Step 2: Get the data

- Obtain and Understand Data
- There are many online data sources where you can get free data sets to use in your project. Some amazing data repositories- Kaggle, Google Cloud Public Datasets, Data.gov, and websites containing academic papers with datasets. Websites such as Facebook and Twitter allow users to connect to their web servers and access their data.
- You can use their Web API to crawl their data.

contd…

- Here are a few ways to get yourself some usable data:
- Connect to a database: Ask your data and IT teams for the data that's available or open up your private database and start digging through it to understand what information your company has been collecting.
- Use APIs: Think of the APIs to all the tools your company's been using. You have to work on getting these all set up so you can use those email open and click stats, the information your sales team put in Pipedrive or Salesforce, the support ticket somebody submitted, etc.
- Look for open data: Kaggle, Google Cloud Public Datasets, Data.gov

Step 3: Clean the data

- Data Preparation step: to transform the raw data in a useful and efficient format.
- After obtaining data the next step is exploring and cleaning data. When going through the data sets, look for missing data, duplicate data, different spelling errors, or even the data that doesn't make sense logically. To organize your data you can use different tools –R, Python, Tableau, Spark, etc.
- To perform any analytical activity on any data it needs to be in a structured format.
- You have to verify if data types in data are compatible or not? Are there missing values or outliers? Are there any naturally occurring discrepancies or errors that should be corrected before fitting the data into a model? Do you need to create dummy variables for categorical variables? Will you need all the variables in the data set?

Step 4: Enrich your data

- Now that you have clean data, it’s time to manipulate it in order to get the most value out of it. You should start the data enrichment phase of the project by joining all your different sources and group logs to narrow your data down to the essential features. One example of that is to enrich your data by creating time-based features, such as:
  - Extracting date components (month, hour, day of the week, week of the year, etc.)
  - Calculating differences between date columns
  - Flagging national holidays

Step 5: Find insights and Visualization

- Data Visualization is a graphical representation of information and data. By using visual elements like charts, graphs, and maps, data visualization tools provide a quick and effective way to communicate and illustrate your conclusions.
- This is a good time to start exploring it by building graphs. When you’re dealing with large volumes of data, visualization is the best way to explore and communicate your findings.
- Graphs are way to enrich your dataset and develop more interesting features.
- The tricky part here is to be able to dig into your graphs at any time and answer any question someone would have about a given insight.

Step 6: Deploy Machine learning

- Machine learning algorithms can help you go a step further into getting insights and predicting future trends.
- You can use regression for predicting future values, and classification to identify, and clustering to group values. For model performance measurement, precision, recall, F1-score can be used in classification.
- To perform the tasks above, you will need certain technical skills and tools like Python or R. If you are using Python, you need to know how to use Numpy, Matplotlib, Sci-Kit learn, and Pandas. If you are using R, you should know GGplot2,CARET, or data exploration. For handling bigger data sets you are required to have skills in Hadoop, Spark.

Step 7: Iterate

- One of the biggest mistakes that people make with regard to machine learning is thinking that once a model is built and goes live, it will continue working as normal indefinitely. On the contrary, models will actually degrade in quality over time if they’re not continuously improved and fed new data.
- Ironically, in order to successfully complete your first data project, you need to recognize that your model will never be fully “complete." In order for it to remain useful and accurate, you need to constantly reevaluate, retrain it, and develop new features.