Intro to R - Basics - 1

One should look for what is and not what he thinks should be – Albert Einstein

Basics: Topic introduction

In this part of the course, we will cover the following concepts:

- Overview of Data Science and its tools
- R and RStudio as tools in data analysis and their features
- Basic calculations in R

Warm up

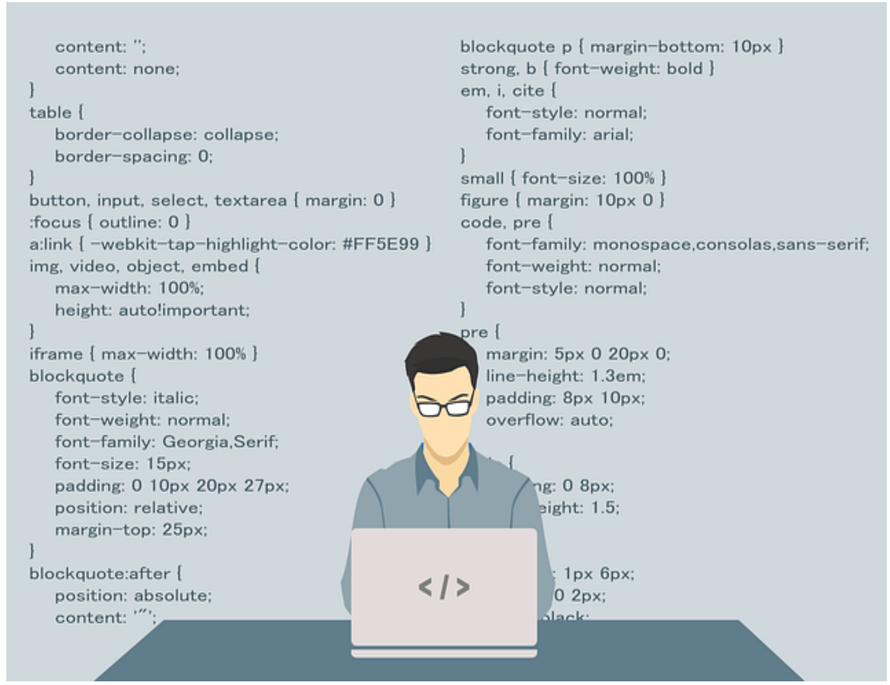
- Before we get underway, let's get to know one another a little with an ice-breaker question
- What everyday task that you perform do you most wish a computer program could do instead?
- Share your responses in the virtual chat

Module completion checklist

Objective	Complete
Discuss how programming is used across industries and define core functions of data scientist	
Identify stages of the data science control cycle	

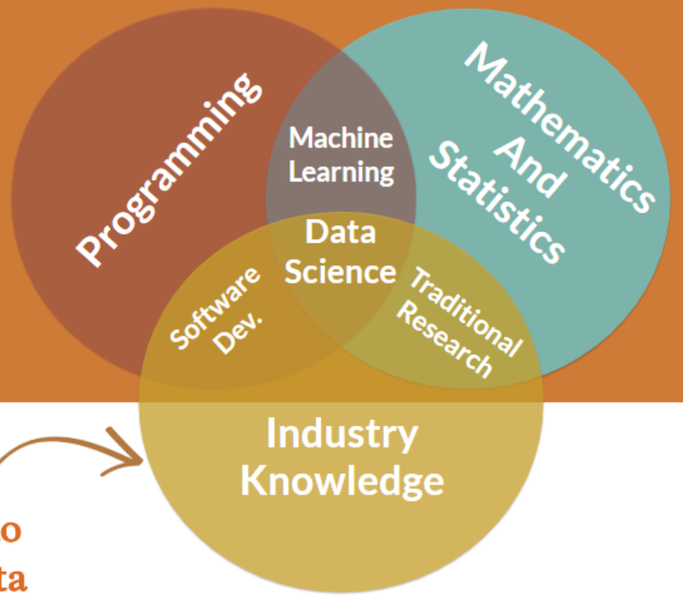
Why are we learning to program?

- Programming is becoming a universal skill, like making Word documents or PowerPoint presentations
- Writing a program in code allows you to perform the same operation over and over on a large scale
- The quality of your code will allow your work to be portable and reproducible
- Programming is a **mindset shift** that asks you to think through your problem in step-wise fashion, from initial state to end state



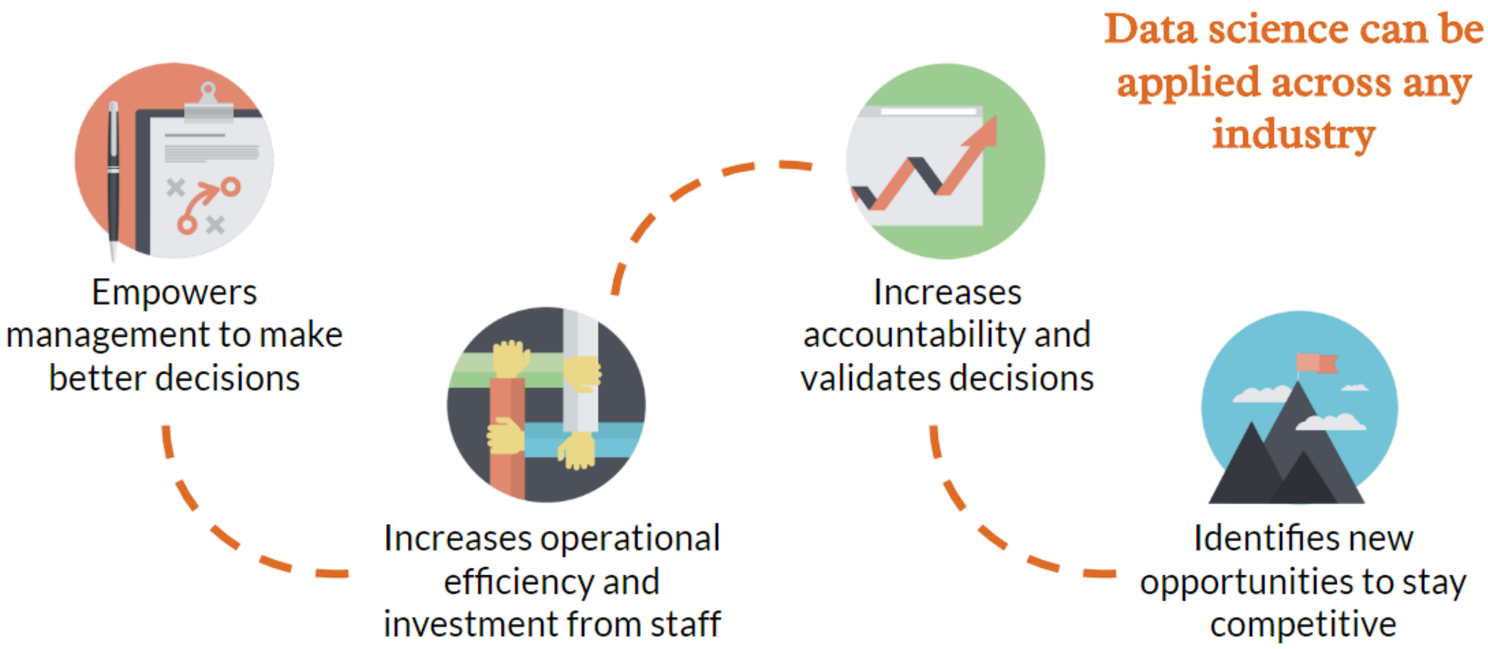
What is data science?

- Data science applies the scientific method to analyzing data
- It lies at the intersection of several disciplines
- It draws on industry knowledge that makes the analysis of Big Data possible

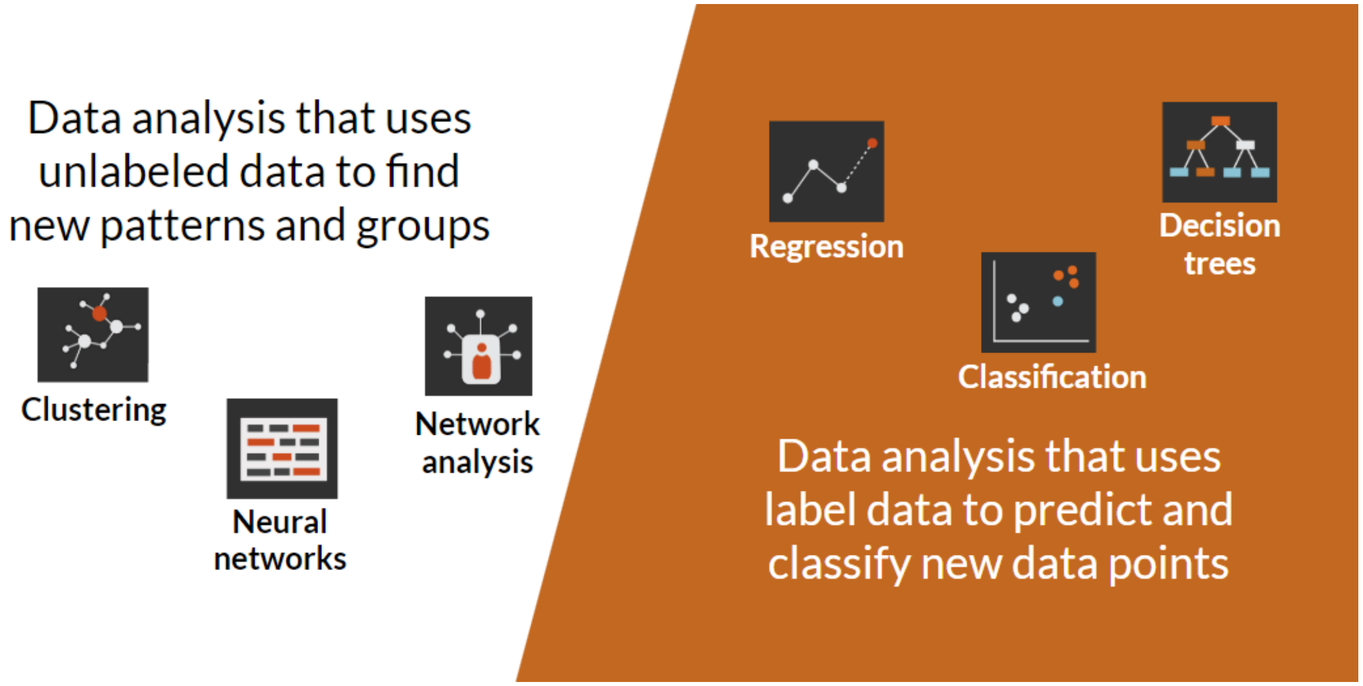


Industry knowledge is essential to knowing what to look for when exploring data

What can data science do?



Review: unsupervised vs supervised machine learning



What machines cannot do

Machines cannot:

- Understand context
- Think through a problem
- Ask the right questions
- Select the right tools
- Interpret results

But you can program machines to:

- Perform calculations quickly
- Automate repetitive tasks
- Follow pre-defined rules
- Visualize data

Data science tools are only useful when you have people who can use them effectively

What is a data scientist?

An analyst who can:

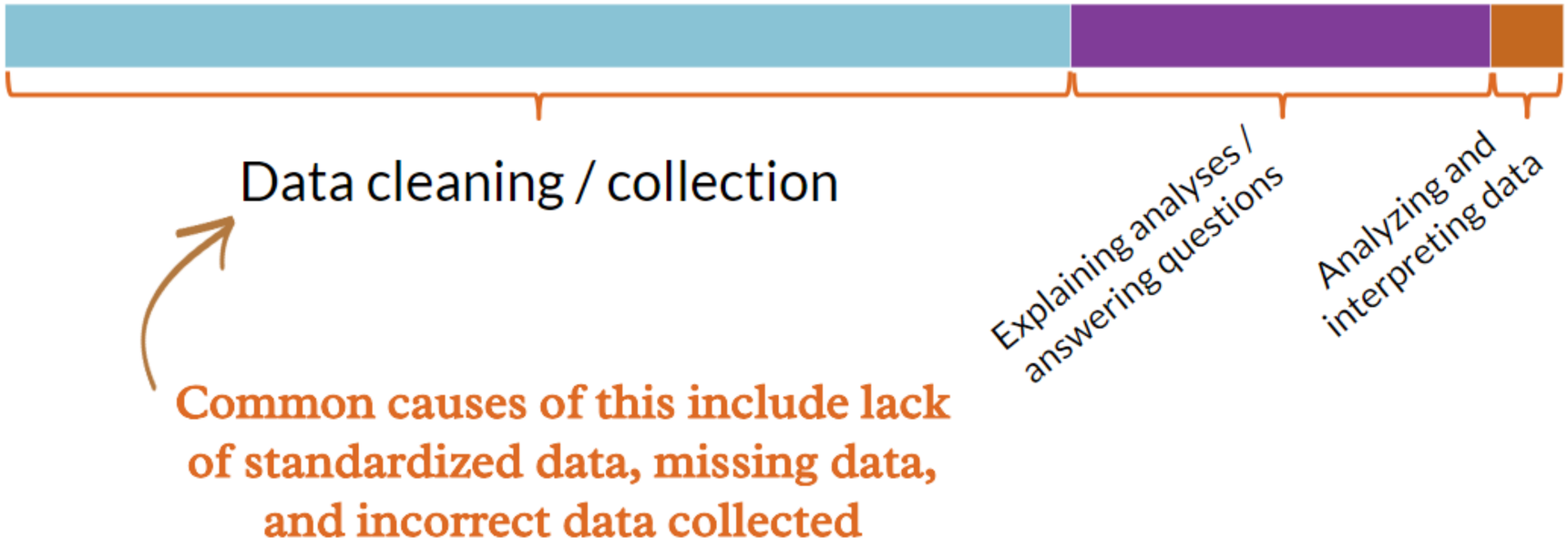
1. **Pose** the right question
2. **Wrangle** the data (gather, clean, and sample data to get a suitable data set)
3. **Manage** the data for easy access by the organization
4. **Explore** the data to generate a hypothesis
5. **Make predictions** using statistical methods such as regression and classification
6. **Communicate** the results using visualizations, presentations, and products




How people think data scientists spend time

- Which of the steps in the image **actually** takes the most time?

How data scientists actually spend time



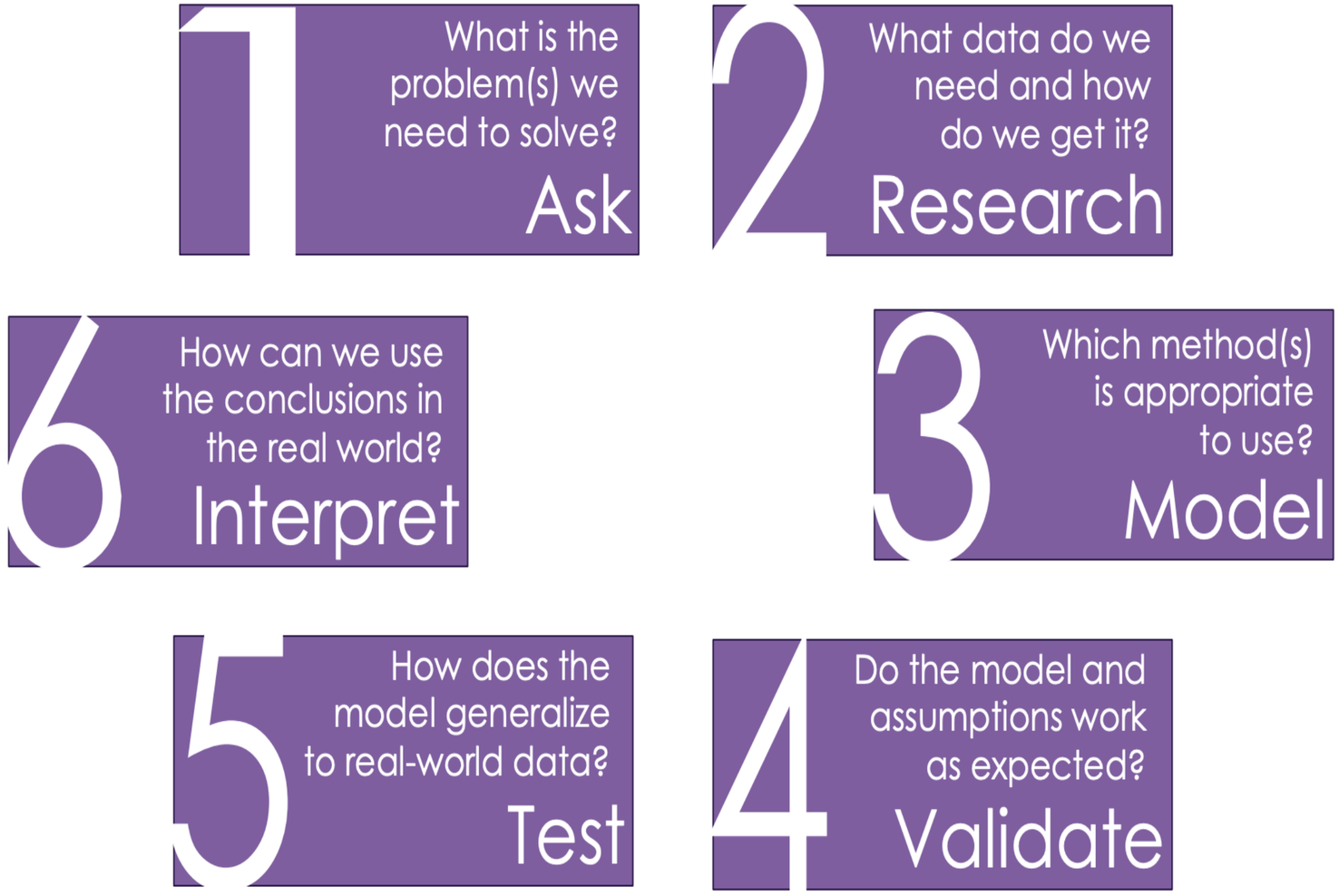
Module completion checklist

Objective	Complete
Discuss how programming is used across industries and define core functions of data scientist	✔
Identify stages of the data science control cycle	

Data science control cycle: framework for data

- Data scientists usually follow a specific protocol or standard for working with data
- The data science control cycle is a modified version of the scientific method
- The cycle involves working with data from start to finish
  - Asking the right questions
  - Collecting usable data
  - Understanding the data you're studying
  - Optimizing model performance

Data Science Control Cycle (DSCC)



DSCC: SMART questions

SPECIFIC

- How are you framing the question?
- What specific variables?

MEASURABLE

- What metrics are you using?
- What is the success criteria?

ACHIEVABLE

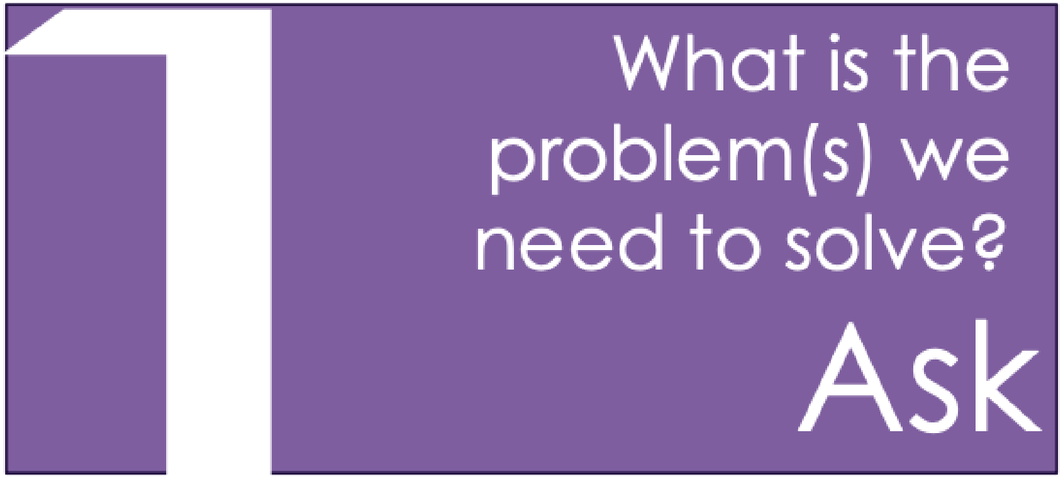
- Scope your analysis well
- Use data that is available to you

RELEVANT

- Who will use this analysis?
- Is it interesting or usable?

TIMEBOUND

- Reference time frame of analysis
- If predicting, in next year? next month? ever?



DSCC: SMART questions (cont'd)

- Some of the SMART questions to ask during the initial requirements gathering phase would be:
  - Who're the users of the analysis outcomes and how do they intend to apply them? (inclusive of board members, sales personnel, customers, staff, auditors, etc.)
  - What are the questions that might arise from the audience concerning our analysis? (such as the capacity to narrow down to crucial segments, examine data trends over time, delve into particulars, etc.)
  - In what sequence should they be ranked to maximize the obtained benefits?
  - What are the reports that currently exist? What changes might be made to existing reports?

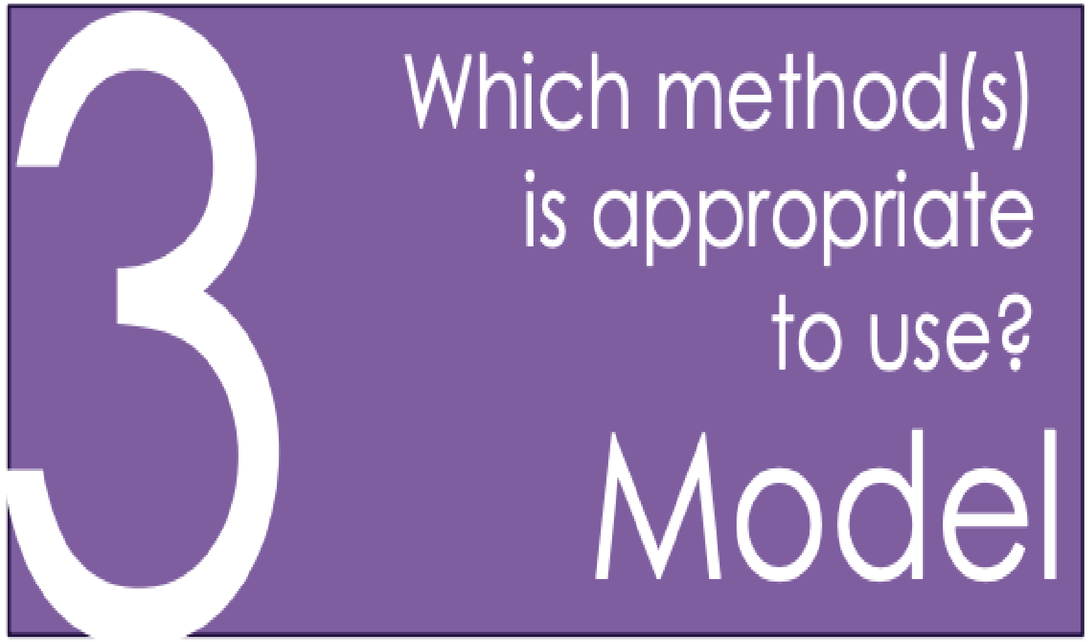
DSCC: SMART questions (cont'd)

- Examples of NON-SMART questions would be:
  - Questions about long-term organizational strategy if the project is purely tactical or short-term in nature
  - Asking about proprietary and third party data sources when it's already known that they would not be useful on the project

DSCC: research

- Data is key to quality results
- **Garbage in - garbage out** is the famous programming mantra that holds true for data science as well
- It should always be on your mind when working with any dataset
- The dataset can be obtained from the client or the organization you're working with or can be gathered online based on the requirements established
- The suitability and quality of your data should never be overlooked

DSCC: modeling

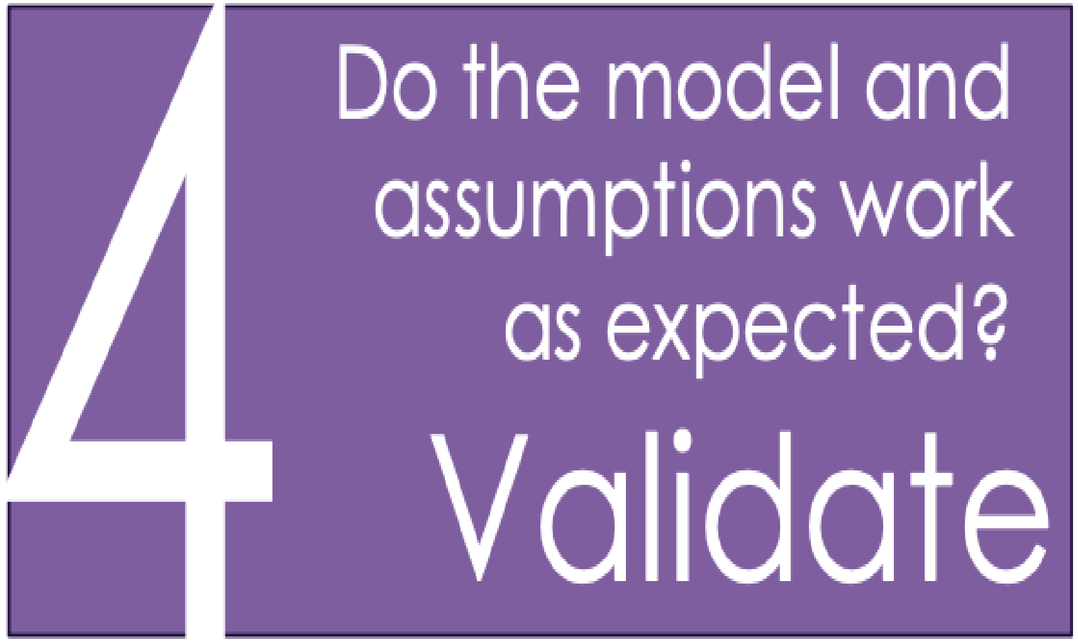


- A model is a replica of a real thing
- Select a model that simulates the real-life situation, or that suits your problem/data, in the closest possible way
- In the context of data science, there are different kinds of models available based on the use cases focused on supervised and unsupervised learning

$$\min_{w,b,\varepsilon} \frac{1}{2}\sum_{p=1}^{d} w_p^2 + \gamma \sum_{i=1}^{n} \varepsilon_i$$

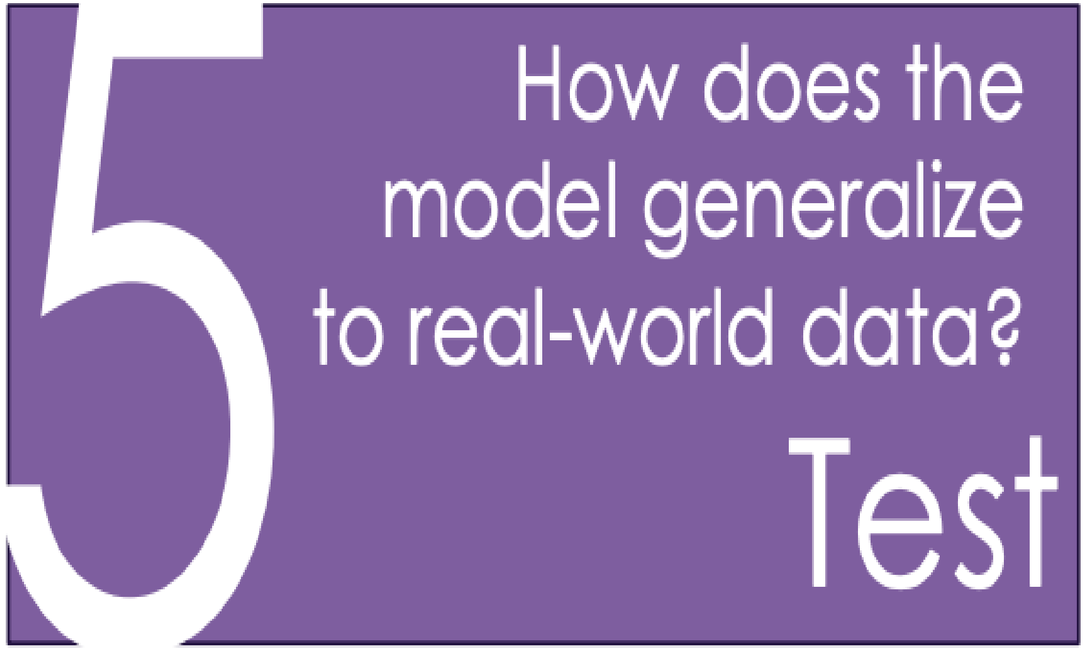
$$\text{subject to} \begin{cases} y_i \left[\sum_{p=1}^{d} w_p x_i^p + b \right] \geq 1 - \varepsilon_i, \forall i = 1, \ldots, n \\ \varepsilon_i \geq 0 \qquad \forall i = 1, \ldots, n \end{cases}$$

DSCC: validating



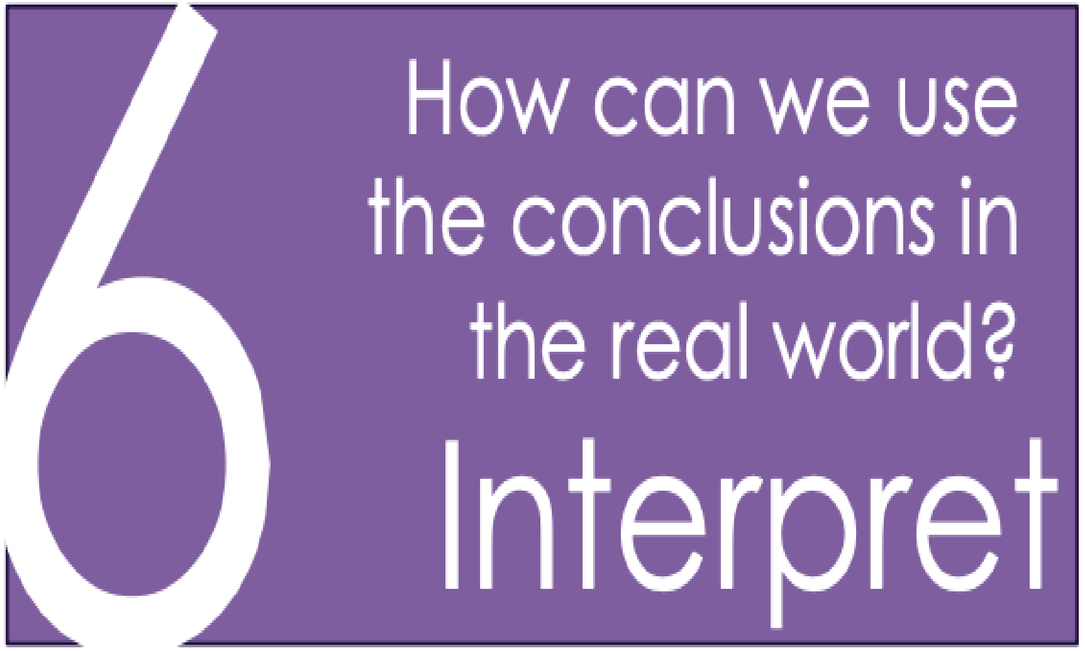
- We all have prior knowledge that sometimes makes us pre-conditioned to make incorrect assumptions
- Don’t let your extensive experience get in the way, always start as if you know nothing about the problem
- In the context of data science, we would need to validate if our model is working as expected and is predicting well on the unseen data which is not used for training phase

DSCC: testing



- Test your validated model with real-world data before delivering results to stakeholders
- Adjust the model if necessary
- In the context of data science, we would need to validate if our model is working as expected and is predicting well on the unseen or real-world data which is not used during training and validation phase
- Have you ever made incorrect assumptions about a problem you were trying to solve?

DSCC: step 6



- Interpreting the results is just as important as the results themselves
- Use your best judgment and expertise to deliver relevant information supported by the data to stakeholders
- Make your conclusions actionable, so that stakeholders know what next steps to take from looking at your results

Why use R?

- De facto standard among professional statisticians
- Comparable and often superior in power to commercial products (SAS, SPSS, Stata)
- Available for the Windows, Mac, and Linux operating systems
- General-purpose programming language useful for automating analysis
- Create dynamic graphics and visualizations
- Large user community connected through ***www.r-bloggers.com***
- Over 12,000 packages and rising to run data analyses contributed by user base

R compared to Excel

	R	Excel
Data capacity	R can read files as big as several gigabytes and trillions of data points; only limitation is your RAM	Excel can't read more than 1,048,576 rows and 16,384 columns (2013 version), files over ~300 megabytes can be very slow to work with
Customization	Can create custom visualizations through code, very flexible	Drop down menus limit ability to manipulate charts and graphs
Analyzing data	Powerful, pre-built packages that speed up work flow	Less flexible built-in analytic abilities that can be augmented by macros
Modeling	Data analysis and statistical models	Complex financial and accounting models
Seeing data	Built-in spreadsheet viewer	Easy to use spreadsheet interface
Usability	Direct commands similar to Excel if-statements	Keyboard shortcuts and slower point-and-click functionality

Knowledge check

Module completion checklist

Objective	Complete
Discuss how programming is used across industries and define core functions of data scientist	✔
Identify stages of the data science control cycle	✔

Congratulations on completing this module!