TU 257 – Fundamentals of Data Science

Data Analytics

Lab 2 – Data Understanding

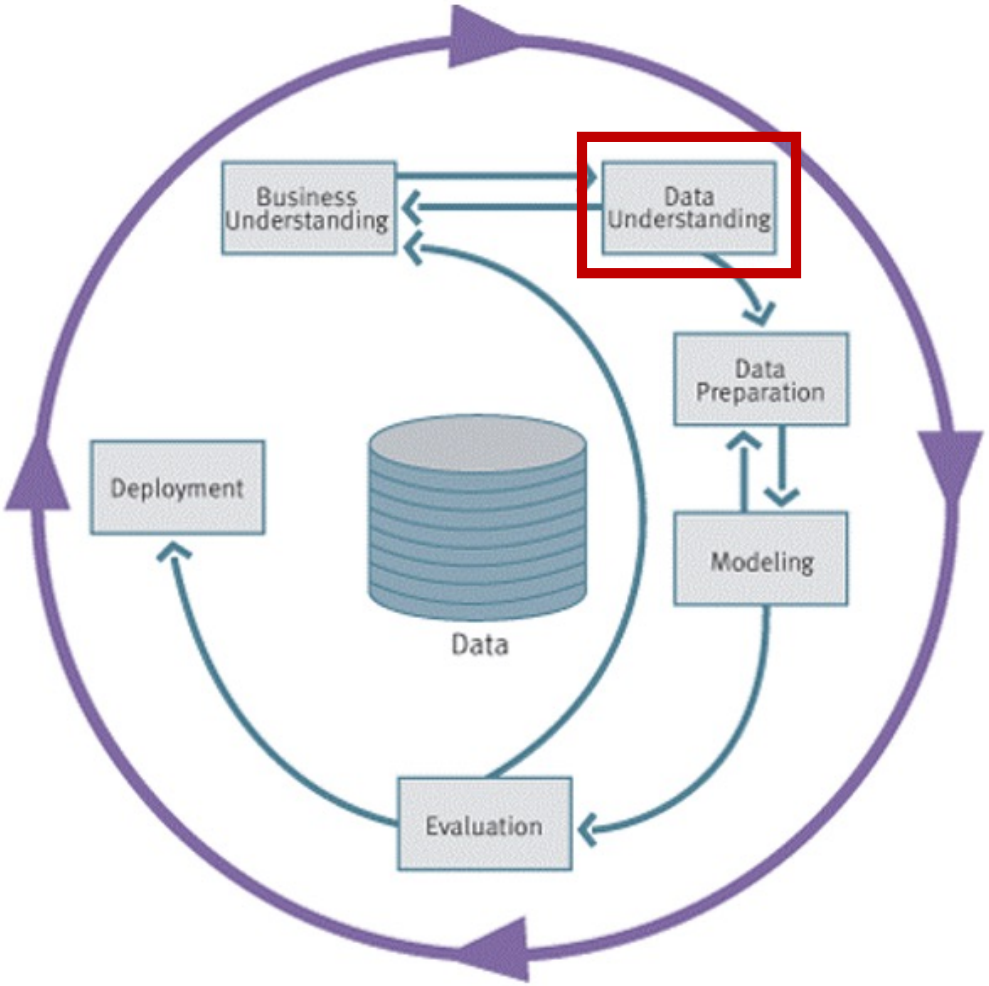
Brendan Tierney

Agenda

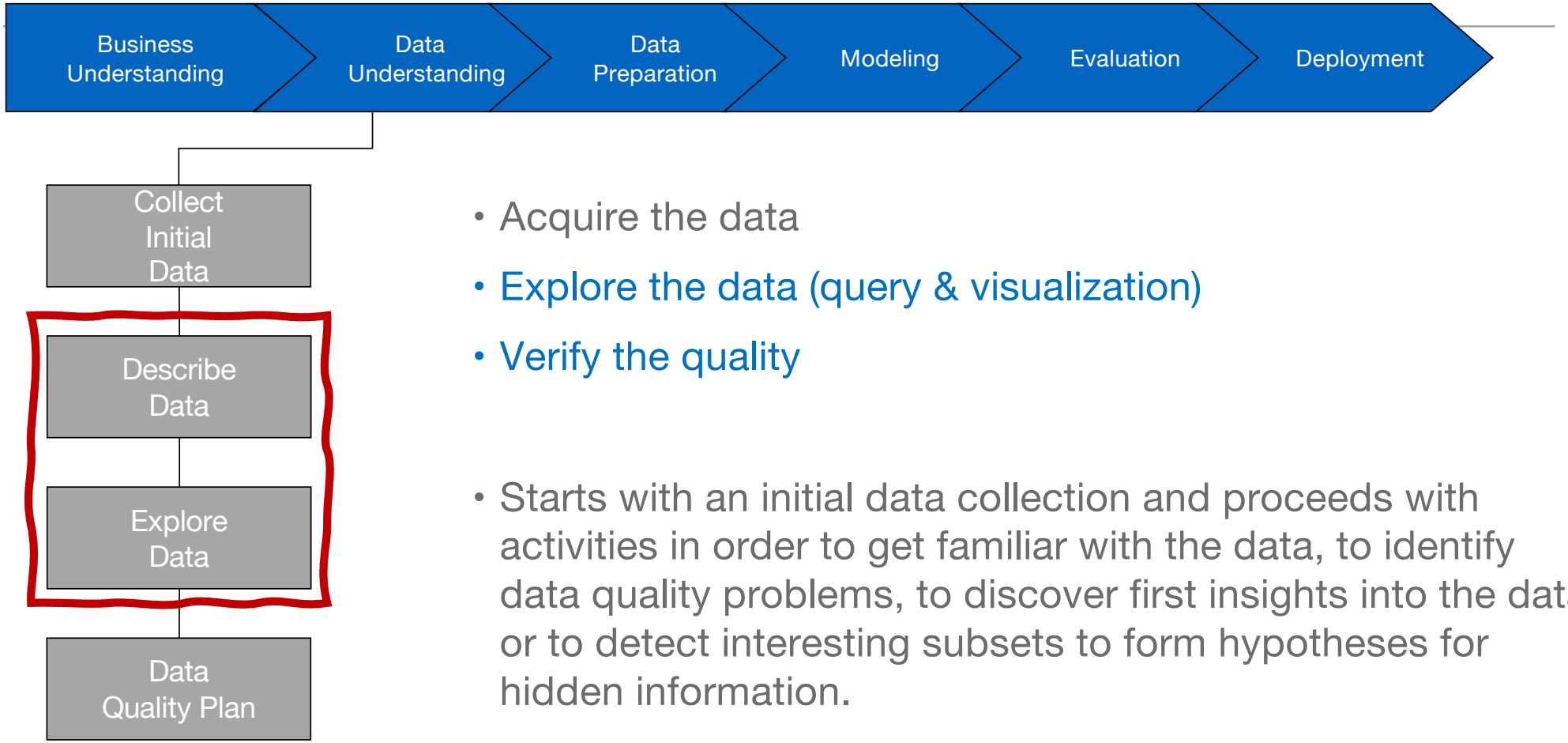
- CRISP-DM – Data Understanding

- Exercise 1 - Data Understanding & Exploring – Automated
  - Demo

- Exercise 2 - Data Understanding & Exploring – Manually with code
  - Demo


Business Understanding
Data Understanding
Data Preparation
Modeling
Evaluation
Deployment
Data

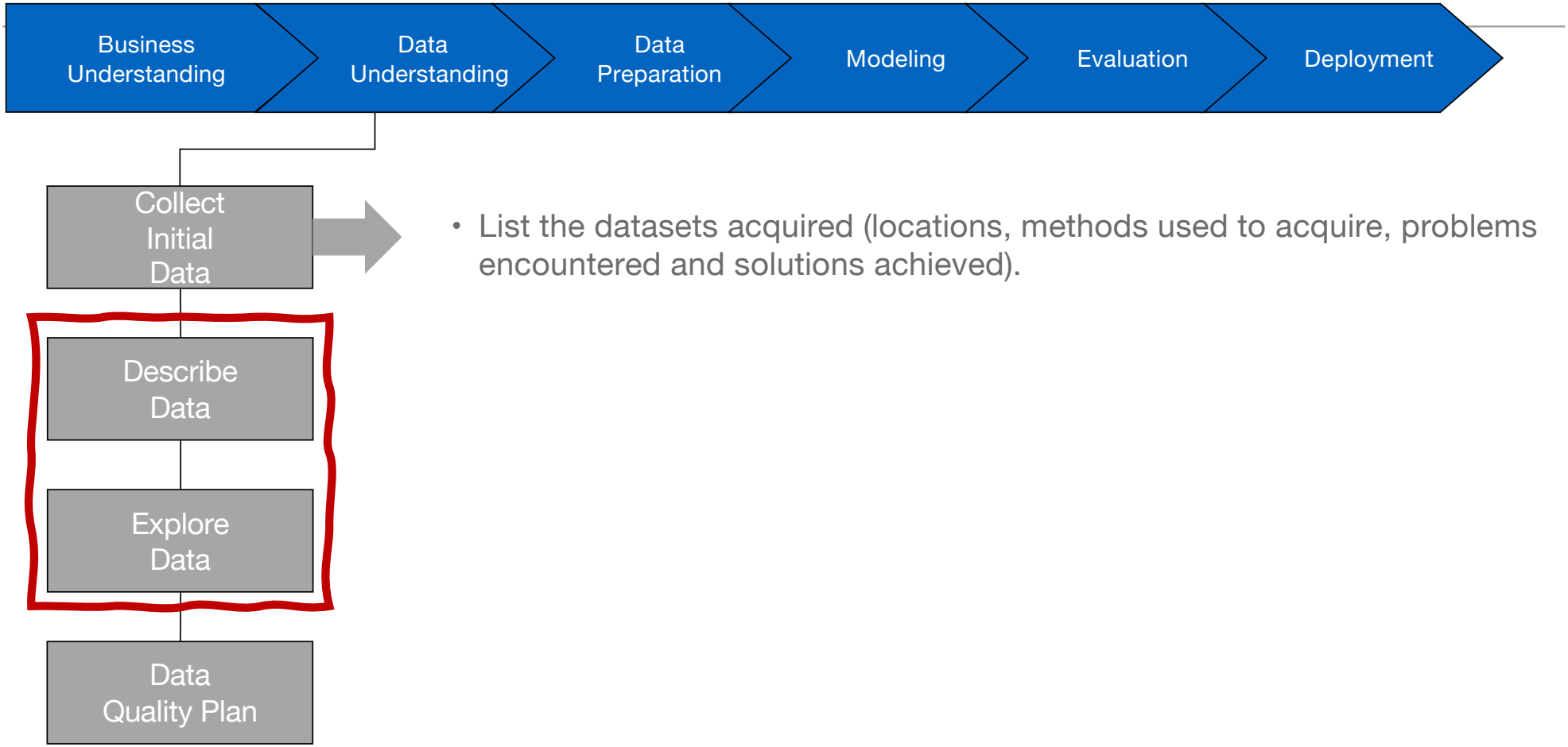
Phase 2 – Data Understanding



- Acquire the data
- Explore the data (query & visualization)
- Verify the quality

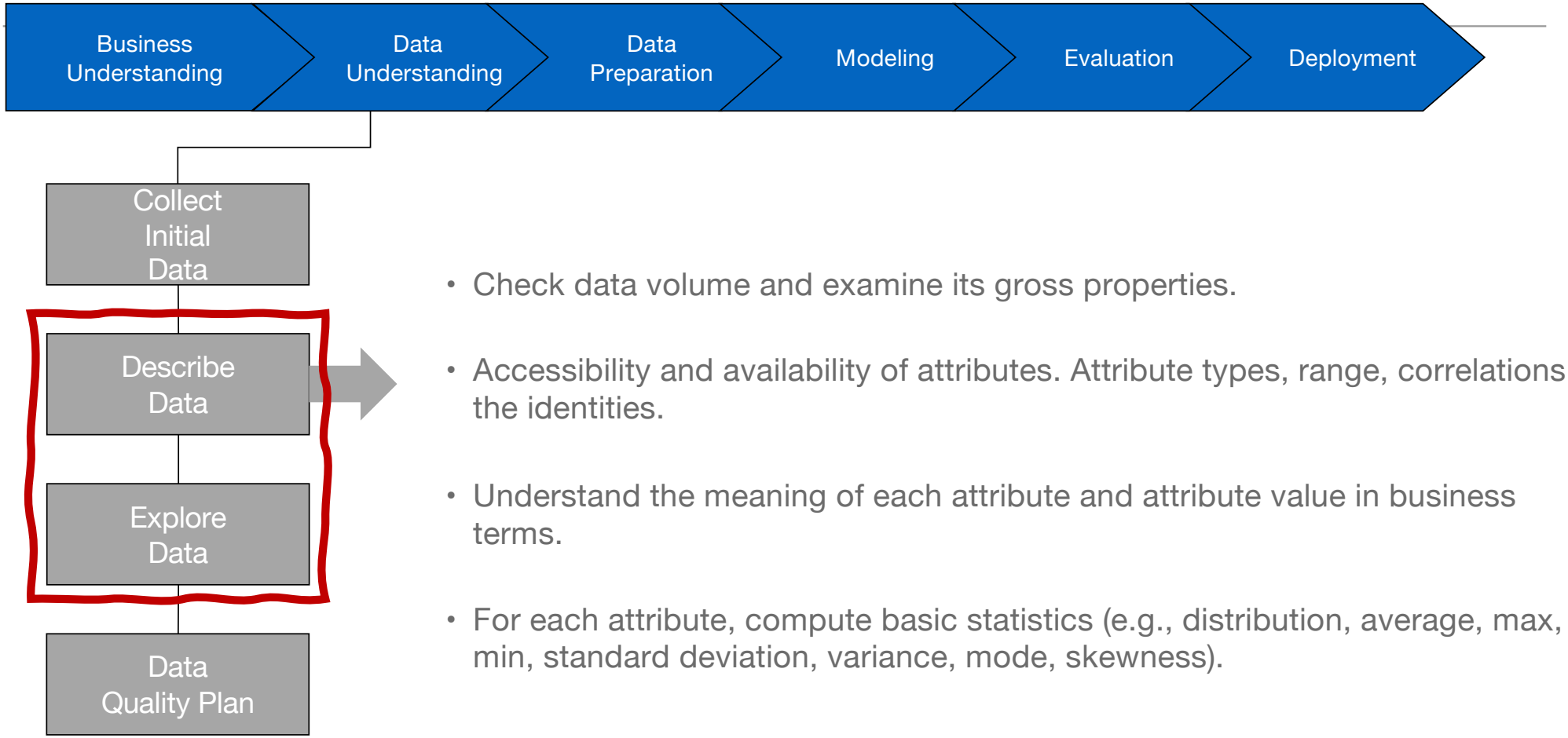
- Starts with an initial data collection and proceeds with activities in order to get familiar with the data, to identify data quality problems, to discover first insights into the data or to detect interesting subsets to form hypotheses for hidden information.

Phase 2 – Data Understanding



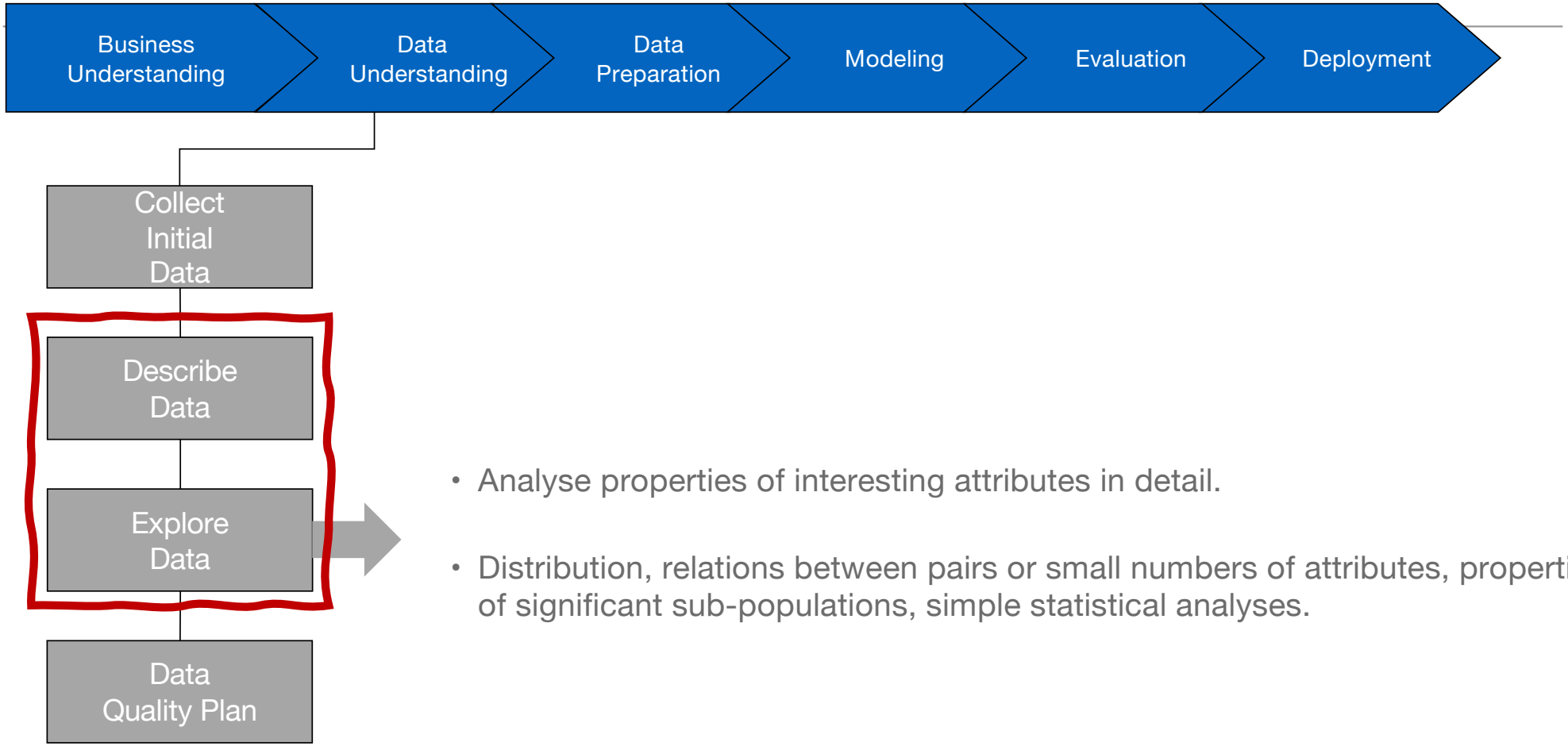
- List the datasets acquired (locations, methods used to acquire, problems encountered and solutions achieved).

Phase 2 – Data Understanding



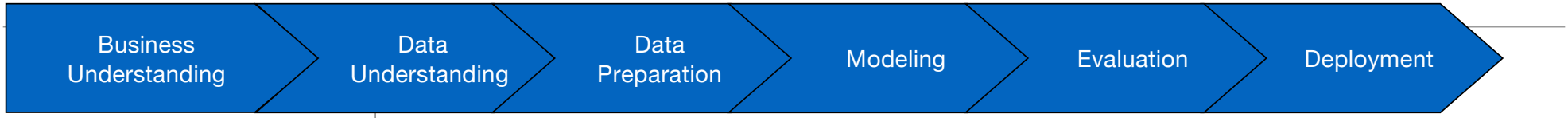
- Check data volume and examine its gross properties.
- Accessibility and availability of attributes. Attribute types, range, correlations, the identities.
- Understand the meaning of each attribute and attribute value in business terms.
- For each attribute, compute basic statistics (e.g., distribution, average, max, min, standard deviation, variance, mode, skewness).

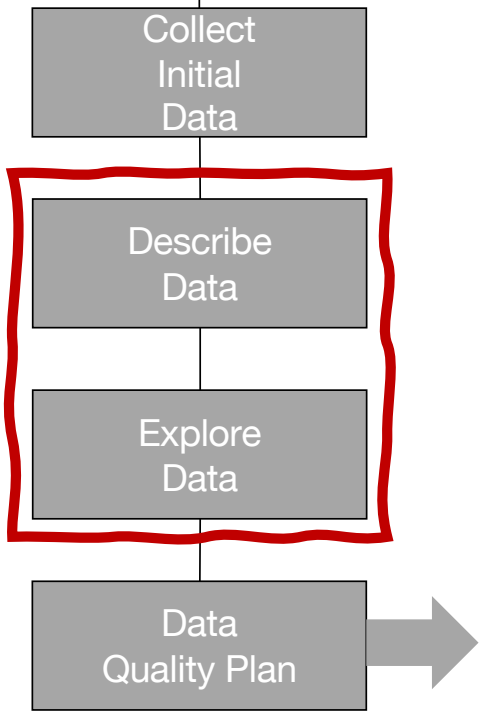
Phase 2 – Data Understanding



- Analyse properties of interesting attributes in detail.
- Distribution, relations between pairs or small numbers of attributes, properties of significant sub-populations, simple statistical analyses.

Phase 2 – Data Understanding





- Identify special values and catalogue their meaning.
- Does it cover all the cases required? Does it contain errors and how common are they?
- Identify missing attributes and blank fields. Meaning of missing data.
- Do the meanings of attributes and contained values fit together?
- Check spelling of values (e.g., same value but sometime beginning with a lower case letter, sometimes with an upper case letter).
- Check for plausibility of values, e.g. all fields have the same or nearly the same values.

Documenting your Work

- Why do we document our work?
- What is different about Data Science / AI / ML Projects?
  - Iterative
  - Gaps between phases and Iterations
  - People move on
- Traditionally CRISP-DM give a framework for Documenting each Phase of the project
  - What was done
  - Why it was done that way
  - Results and Outcomes
- Can compare Changes and Results with each iteration

Exercise 1 - Data Understanding & Exploring – Automated

Install new Library / Package

- See video for a Demonstration of how to install in Anconda
- Install in Virtual Environment created in Data Wrangling module
- Can also use the base environment – but some people have difficulties with this
- The instructions that follow will work for most people
- For others, here are things to try
  - What environment are you installing into? Maybe use the Environment created for Data Wrangling
  - If that doesn't work, try using the conda command – see example on webpage
  - If that doesn't work, try using the pip command – see example

Trouble Shooting

Python Libraries needed for this Lab

`pandas`

`ydata-profiling`

`ipywidgets`

Exercise 1-1 - Data Understanding & Exploring – Automated

- Data Wrangling Module – Lots of different ways to load data into Python & Explore it.
- First step with any Analytics project is to get a basic understanding of what data you have
- In this Exercise we will use a pre-built Python library to generate a report
- Library = ydata_profiling
  - `pip3 install ydata_profiling`
  - `pip3 install ipywidgets  #this library is also needed`
  - If Notebook is already open, Restart the Kernel
    - Menu -> Kernel -> Restart
- ydata_profiling website, examples, documentation
- https://github.com/ydataai/ydata-profiling

Library was previously called pandas-profiling

Exercise 1-1 - Data Understanding & Exploring – Automated

- Simple to do with only a few lines of Python Code
- Check out this post for an example.
  - https://oralytics.com/2019/11/25/data-profiling-in-python/
  - Blog uses the previous version called pandas-profiling
  - They are the same libraries, they have different names (old vs new)
- Follow the steps in this post,
  - Install the library
  - Create a Notebook
  - Download & Open the dataset (see module webpage for data set link)
  - Download the data set (take note of where you saved it
  - Profile the data
  - Inspect what is generated
  - How useful is the generated information
    - Check each menu and what these contain
  - Save the Profiling Report to a file (see library website for examples)

```
In [6]: #Make sure to install 'ydata_profiling' library before running the following
        #see Lab Notes

        #import ypandas_profiling as pp
        from ydata_profiling import ProfileReport

        profile = ProfileReport(df2, title="Profiling Report")
        profile
```

Summarize dataset: 100% 28/28 [00:06<00:00, 4.37it/s, Completed]

Generate report structure: 100% 1/1 [00:05<00:00, 5.06s/it]

Render HTML: 100% 1/1 [00:00<00:00, 1.09it/s]

Profiling Report | Overview | Variables | Interactions | Correlations | Missing values | Sample

Overview

Overview | Alerts 7 | Reproduction

Dataset statistics

Number of variables	9
Number of observations	891
Missing cells	866
Missing cells (%)	10.8%
Duplicate rows	10
Duplicate rows (%)	1.1%
Total size in memory	62.8 KiB
Average record size in memory	72.1 B

Variable types

Categorical	4
Numeric	3
Text	2

Variables

Select Columns

Common errors / problems

- Look back at the slide for installing the library/package – Trouble Shooting

- If this cell generates an error about not being able to find or locate the file/directory

```
In [2]: import pandas as pd

        #Change this next command to the location of train.csv on your Computer
        df = pd.read_csv("/Users/brendan.tierney/Dropbox/4-Datasets/titanic/train.csv")
        #df = pd.read_csv("C:\Studies\TU257\DataAnalytics\Week2\train.csv")
        df.head(8)
```

  - You need to change the location of the file to where your dataset is located
  - Check the spelling
  - Check the use of the / versus \
  - Check you have the double quotes at start and end
  - Did you type something incorrectly

Demo

Exercise 1-2 - Data Understanding & Exploring – Automated

- You have now completed an example of Automated Data Profiling

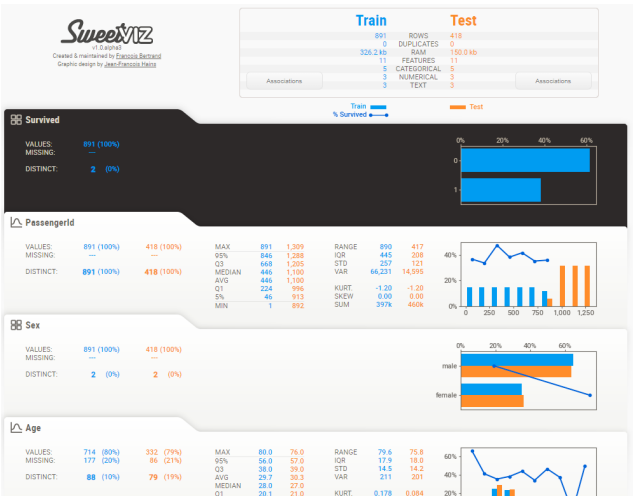
- Do you know of any other datasets?

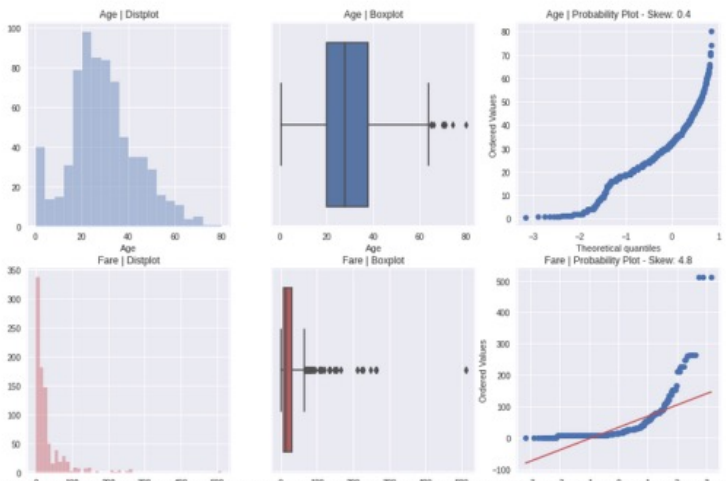
- Try using the Automated Data Profiling on a different datasets.
  - Use datasets from your other modules
  - Check out the UCI Dataset Repository
    - https://archive.ics.uci.edu/ml/index.php
  - Check out this list of Datasets Repositories
    - https://oralytics.com/2019/04/18/data-sets-for-analytics/

Exercise 1-3 - Data Understanding & Exploring – Automated

[Optional]

- You’ve tried one automated data profiling library
- Now try some others
- Check out this for other Python Data Profiling Libraries
  - https://oralytics.com/2022/04/04/python-data-profiling-libraries/





number

	missing	complete rate	mean	sd	p0	p25	p75	p100	hist
sepal_length	0	1.0	0.85	0.99	0.0	0.0	1.0	3.0	
sepal_width	1	0.92	3.32	0.32	2.9	3.08	3.52	3.9	
petal_length	0	1.0	1.46	0.1	1.3	1.4	1.5	1.7	
petal_width	0	1.0	0.21	0.08	0.1	0.2	0.2	0.4	
rand_flower	0	1.0	-0.05	0.11	-0.2	-0.13	0.03	0.11	

category

	missing	complete rate	ordered	unique
class	0	1.0	False	2
location	1	0.92	False	5

datetime

	missing	complete rate	first	last	frequency
date	0	1.0	2018-01-31 00:00:00	2019-01-31 00:00:00	M
date_no_freq	1	0.92	1992-01-05 00:00:00	2023-03-04 00:00:00	None

string

Automated Analytics

- Only gives you more data
- Doesn’t tell you the answers to anything
- It might give you some insights to the data
- Can be a useful starting point
- You can start building upon this, define and write code to explore the data more

- Apply your business/problem understanding to what you learned from the data

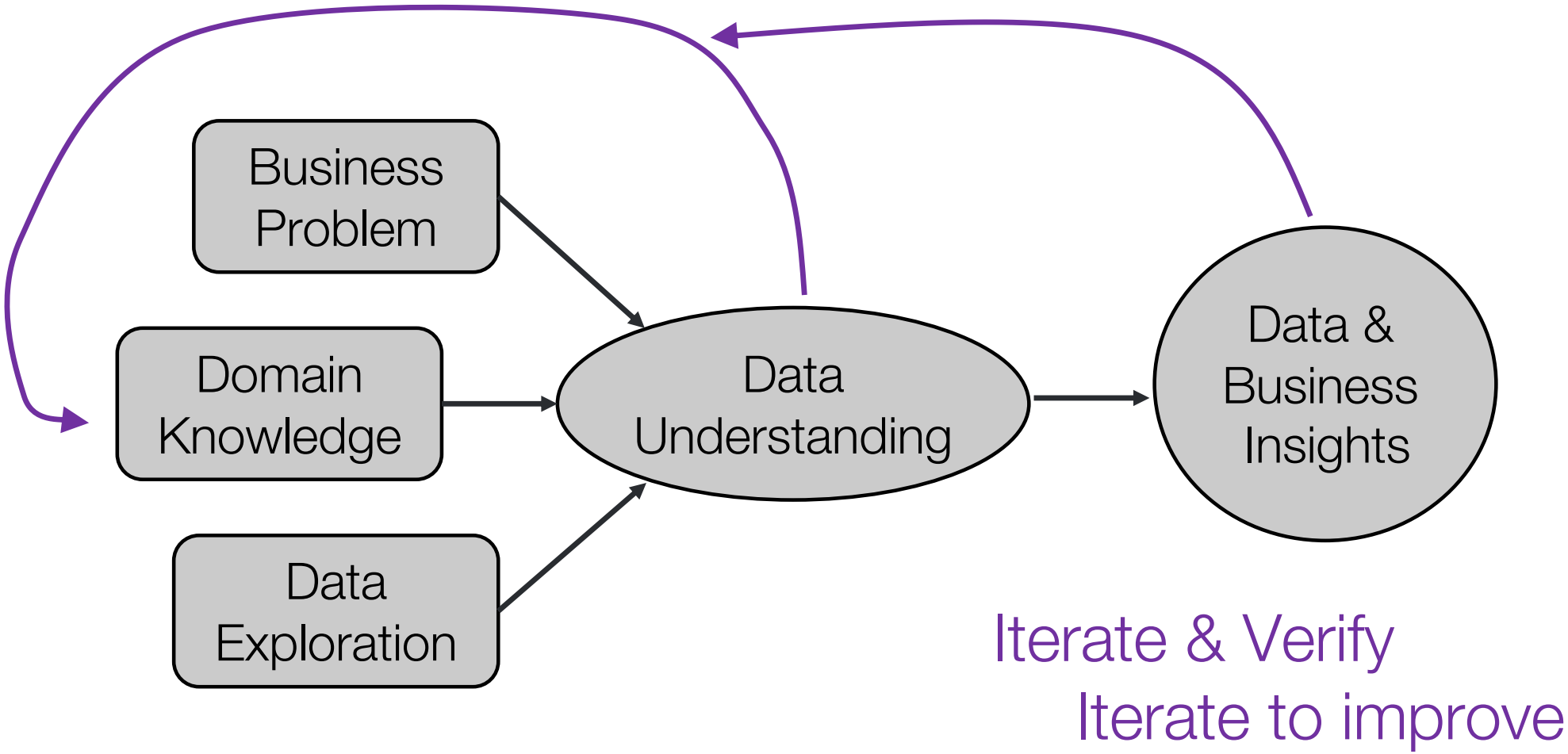
  => Valuable Insights – Comes with Experience, Can be challenging

Exercise 2 - Data Understanding & Exploring – Manually with code

Exercise 2 - Data Understanding & Exploring – Manually with code

- In Exercise 1, we used an Automated way to profile the data
- This can be useful to get an initial view of the data, but these rarely give the full picture
- We will have to write Code or use other tools.
  - Lots of Data Analytics tools out there, where you don't have to write code
  - Use them when you can. It will be quicker to find most of the things you are looking for
- Data Understanding (CRISP-DM)
  - Data Exploration
  - Data Description
  - Explorative Data Analysis
  - Data Profiling
  - …

Exercise 2 - Data Understanding & Exploring – Manually with code



Exercise 2 - Data Understanding & Exploring – Manually with code

- Python Panda dataframes is your main tool for storing, processing and analysing data
- Can use other Python libraries to analyse data in Pandas dataframes
- RTFM
  - https://pandas.pydata.org/docs/
  - https://pandas.pydata.org/docs/user_guide/index.html
  - Check the documentation before googling for an answer
    - Google makes you Stupid!
- Install

```
pip3 install pandas
```

Additional Learning Resouces
10 Minutes to Pandas
Pandas Tutorials

Exercise 2 - Data Understanding & Exploring – Manually with code

- Following the Notebook for Exercise 2 (see module webpage)
- Download the data set from Module Webpage to your local machine
    - Change the notebook to load the data from Your Computer
- Run the Cells
- See what is produced
- Can you understand what each code segment is doing
- Check out the Pandas documentation
- Can you expand the example code segments to do different analysis on the data

Exercise 2 - Data Understanding & Exploring – Manually with code

[Optional]

- You’ve tried one Notebook analysing data
- Now try another dataset

- Try comparing the Automatic Library and manually coding on each dataset.
  - What are the limits of the Automated library?
  - What are the types of things manual coding gives you?
  - Could you use a mixture of the two?

Any Questions ?

What Now/Next ?

Complete **all** Lab Exercises before Next Week

Pick another dataset and complete same/similar tasks with it

Alternative way to install a Conda Package

- Run the conda command in Powershell window

- To find the correct command, got to

  - Anaconda Conda Forge website
    https://anaconda.org/conda-forge
  - Or go to Google, enter Anaconda and Package name
    e.g. Anaconda ydata_profiling

  - Search for name of Package

  - See Installers section