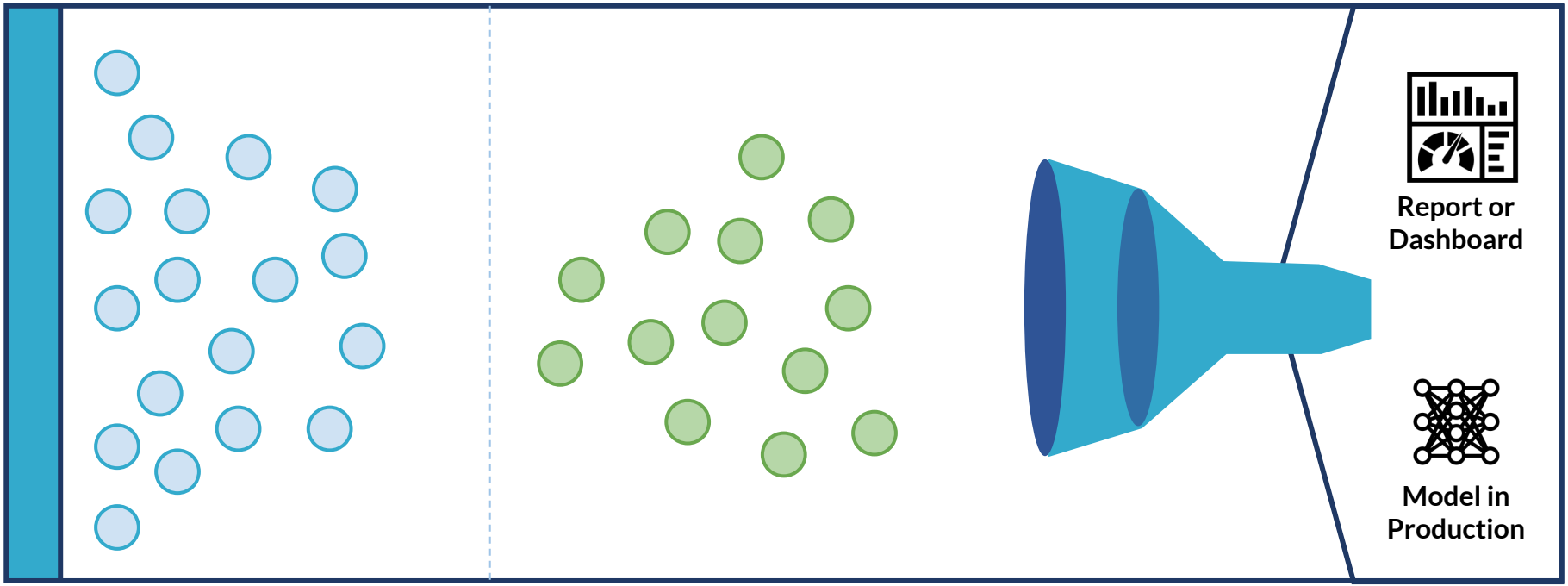
Live training:
Cleaning Data in Python

About me

ADEL NEHME

Content Developer

The data science workflow

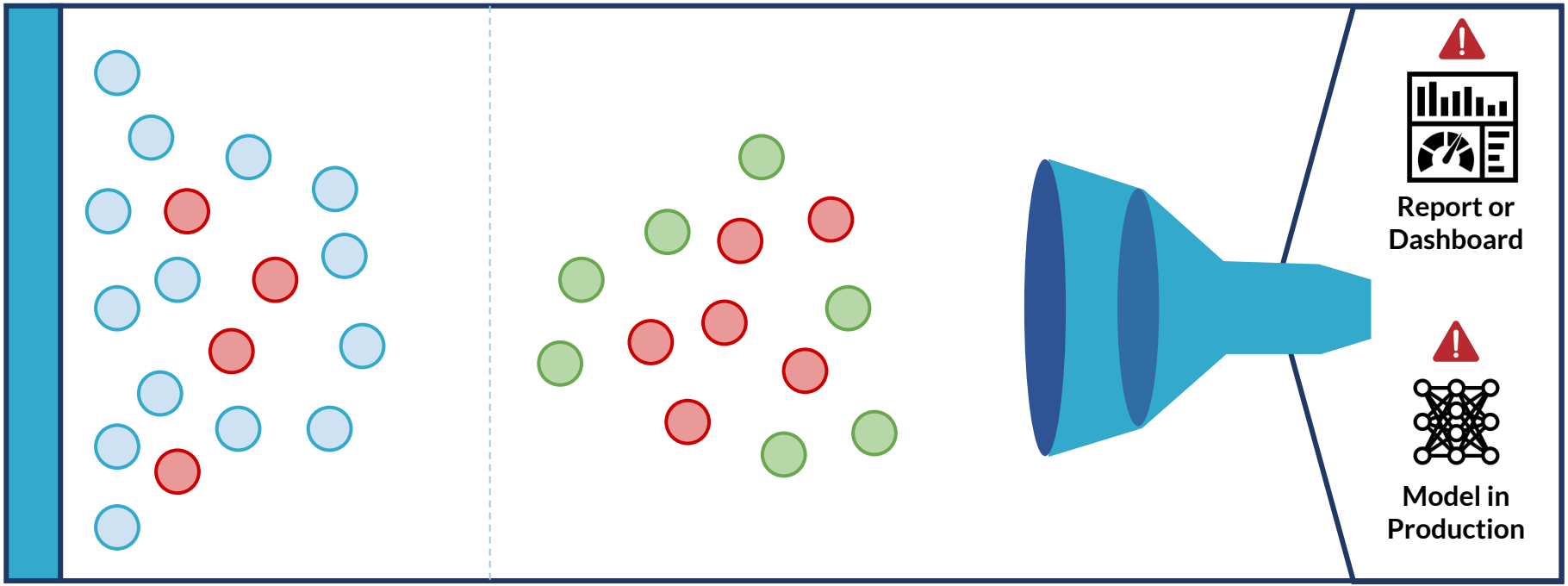


Access
Data

Explore and
Process Data

Extract
Insights

The data science workflow

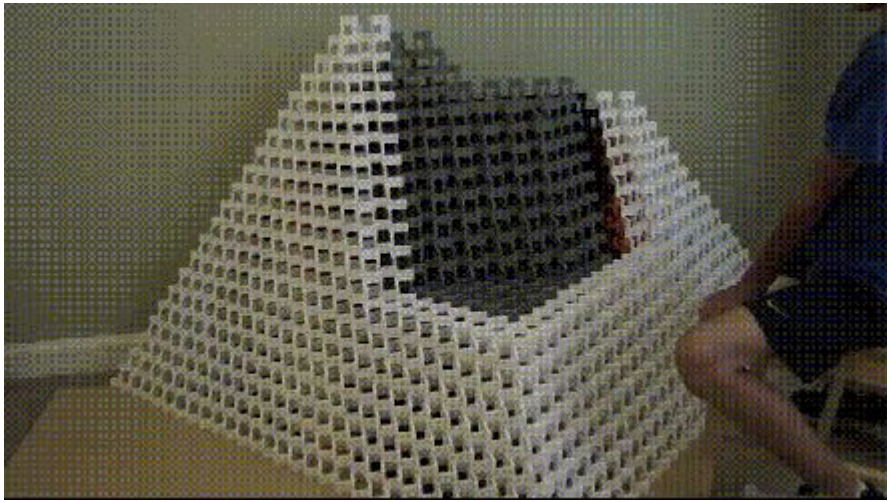


Access
Data

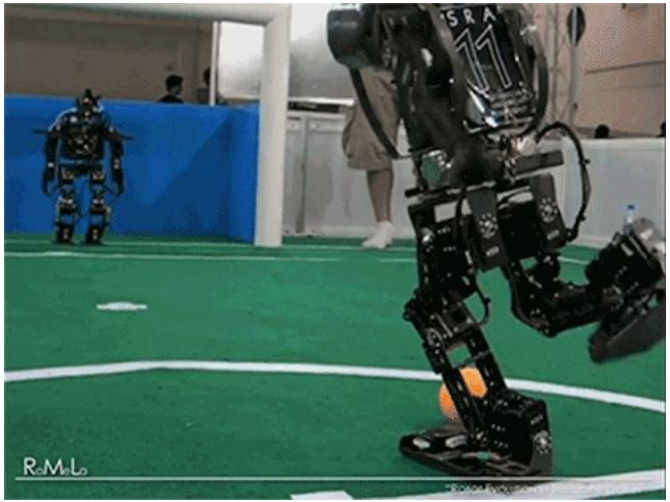
Explore and
Process Data

Extract
Insights

Why do we have dirty data?



Human error



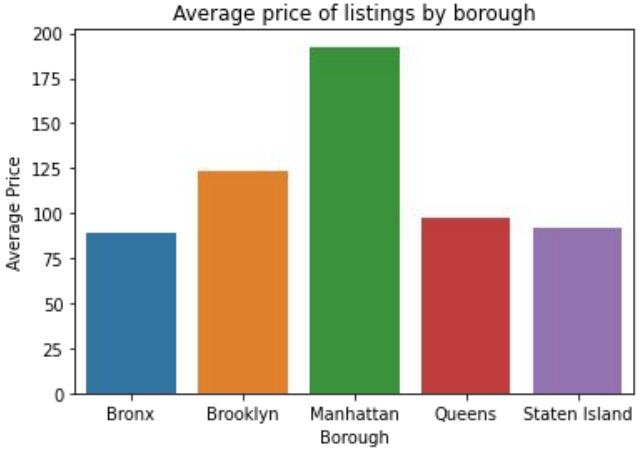
Technical error

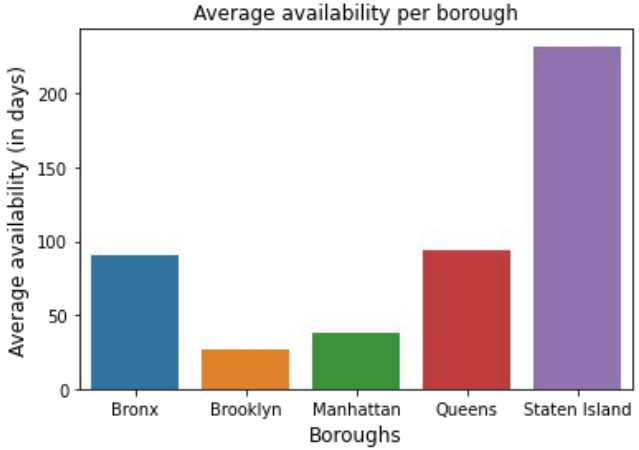
The dataset

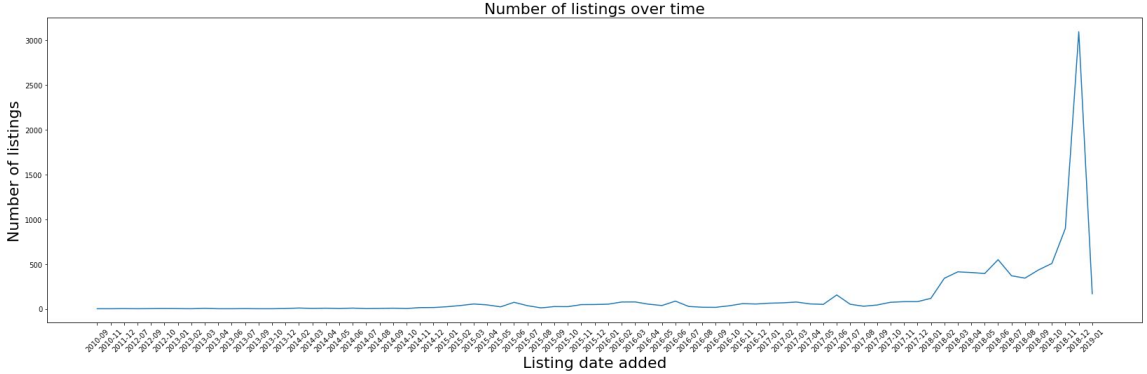
Airbnb data
listing_id: *Unique identifier for a listing*
Name: *Description used for a listing*
Host_id: *Unique identifier for each host*
Host_name: *Name of host*
Neighbourhood_full: *Burrough and neighbourhood*
Coordinates: *Latitude, Longitude*
Room_type: *Type of room*
Price: *Price per night*
Number of reviews: *Number of reviews so far*
Last_review: *Date of last review*
Reviews_per_month: *# of reviews per month*
Availability_365: *Days available per year*
Rating: *Average rating (0 to 5)*
Number_of_stays: *Number of stays so far*
5_stars: *Percentage of ratings that is 5_stars*
Listing_added*: Date listing added to site*

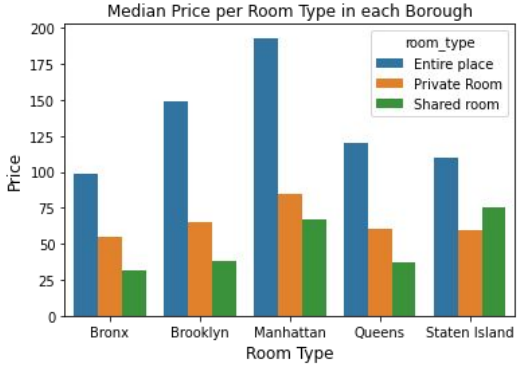
Airbnb data *featuring listings in New York*

The end result









Technologies

pandas

$y_{it} = \beta' x_{it} + \mu_i + \epsilon_{it}$

Popular open source data analysis tool for tabular data

matplotlib

Open source plotting library for 2-D visualizations

Seaborn

Open source plotting library built on top of matplotlib

Technologies

NumPy

Popular open source computing tool for arrays

missingno

Open source plotting library for missing data

datetime

Package for easy date data manipulation

Technologies

jupyter applayout_example Last Checkpoint: 20 hours ago (autosaved)

Logout

File Edit View Insert Cell Kernel Widgets Help Trusted Python 3

```
                      icon='backward',
                      layout=Layout(width='80%',
                                    height='30%'))
next_button = Button(description="Next",
                     icon='forward',
                     layout=Layout(width='80%',
                                   height='30%'))
footer = HTML("Filename: {}".format(image_file))

AppLayout(header=header,
          left_sidebar=prev_button,
          center=image,
          right_sidebar=next_button,
          footer=footer,
          grid_gap='20px',
          justify_items='center',
          align_items='center')
```

Simple Image Viewer

Prev

Next

Filename: images/cat.jpg

❗❗Requires a gmail account to edit❗❗

Session outline

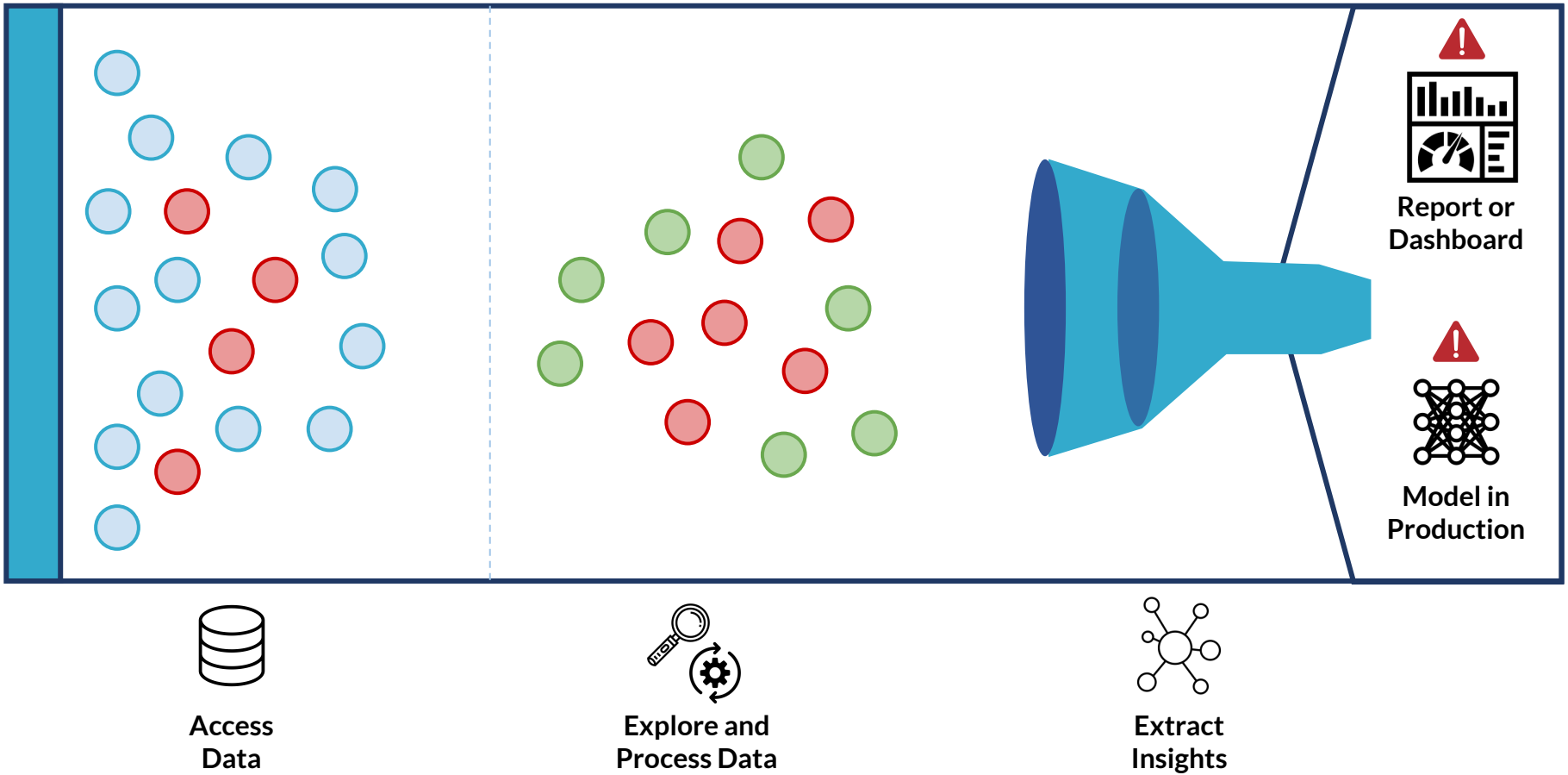
1. *Introduction*
2. *Importing our dataset*
3. *Diagnosing our data problems*
4. *Q&A*
5. *Our to do list*
6. *Data cleaning*
7. *Q&A*
8. *Recap & closing notes*
9. *Take home question*

Notebook

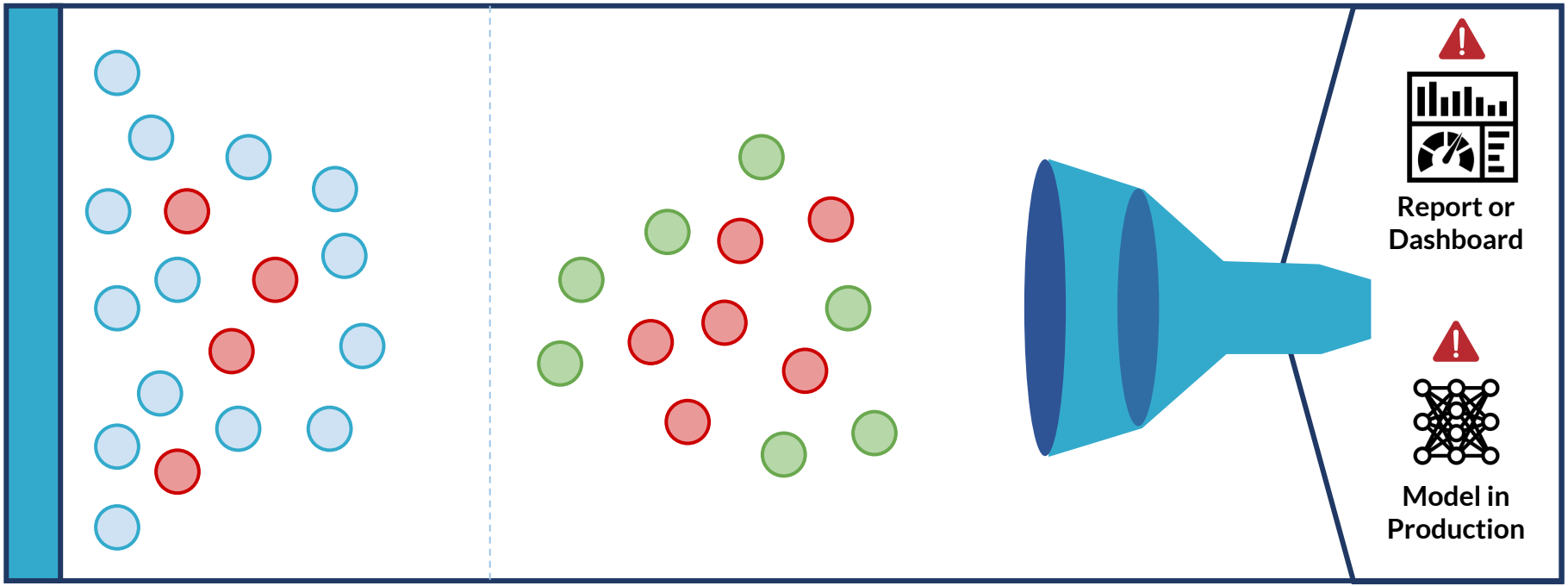
Session outline

1. *Introduction*
2. *Importing our dataset*
3. *Diagnosing our data problems*
4. *Q&A*
5. *Our to do list*
6. *Data cleaning*
7. *Q&A*
8. *Recap & closing notes*
9. *Take home question*

The data science workflow - revisited



The data science workflow - revisited

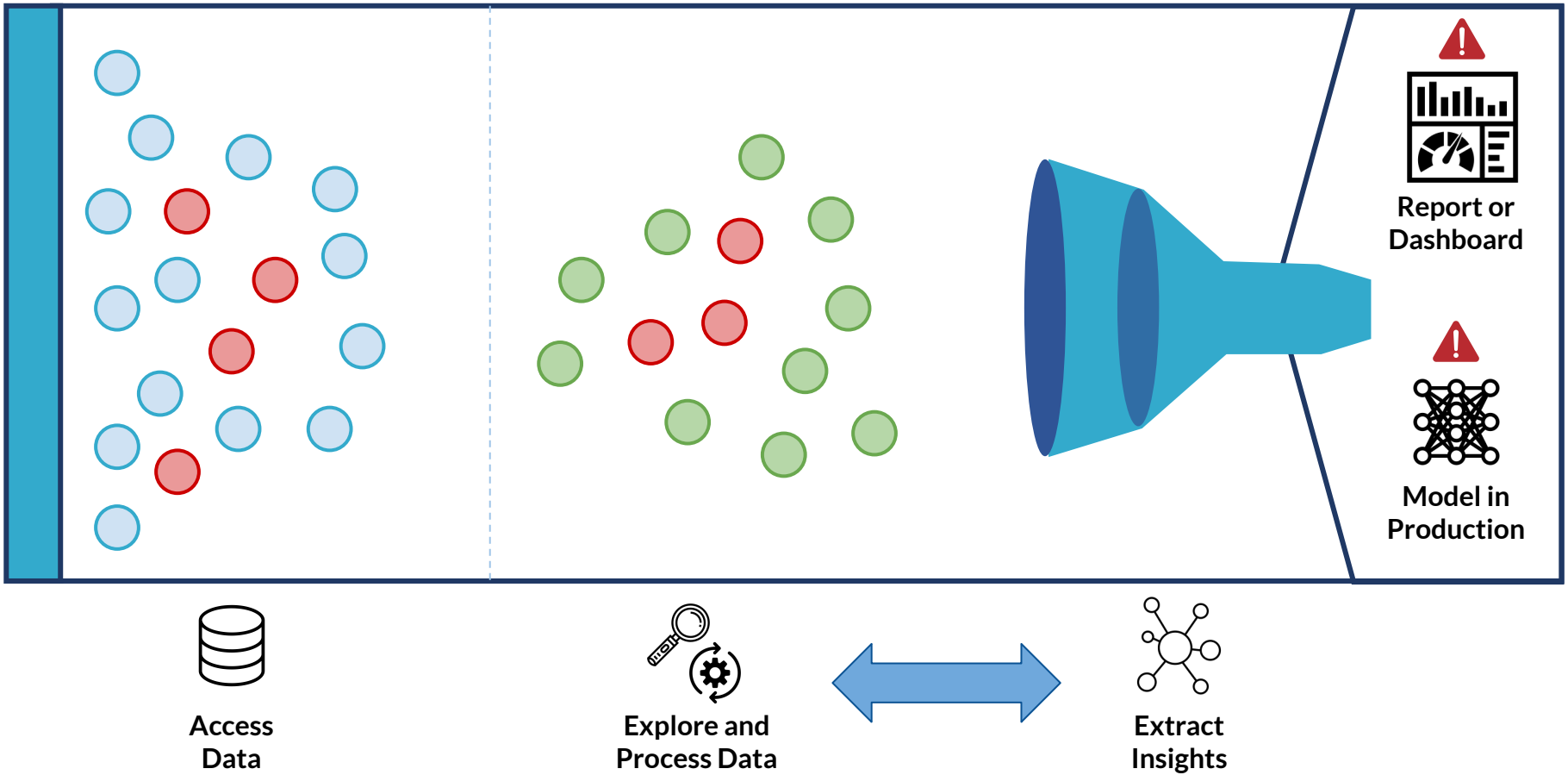


Access
Data

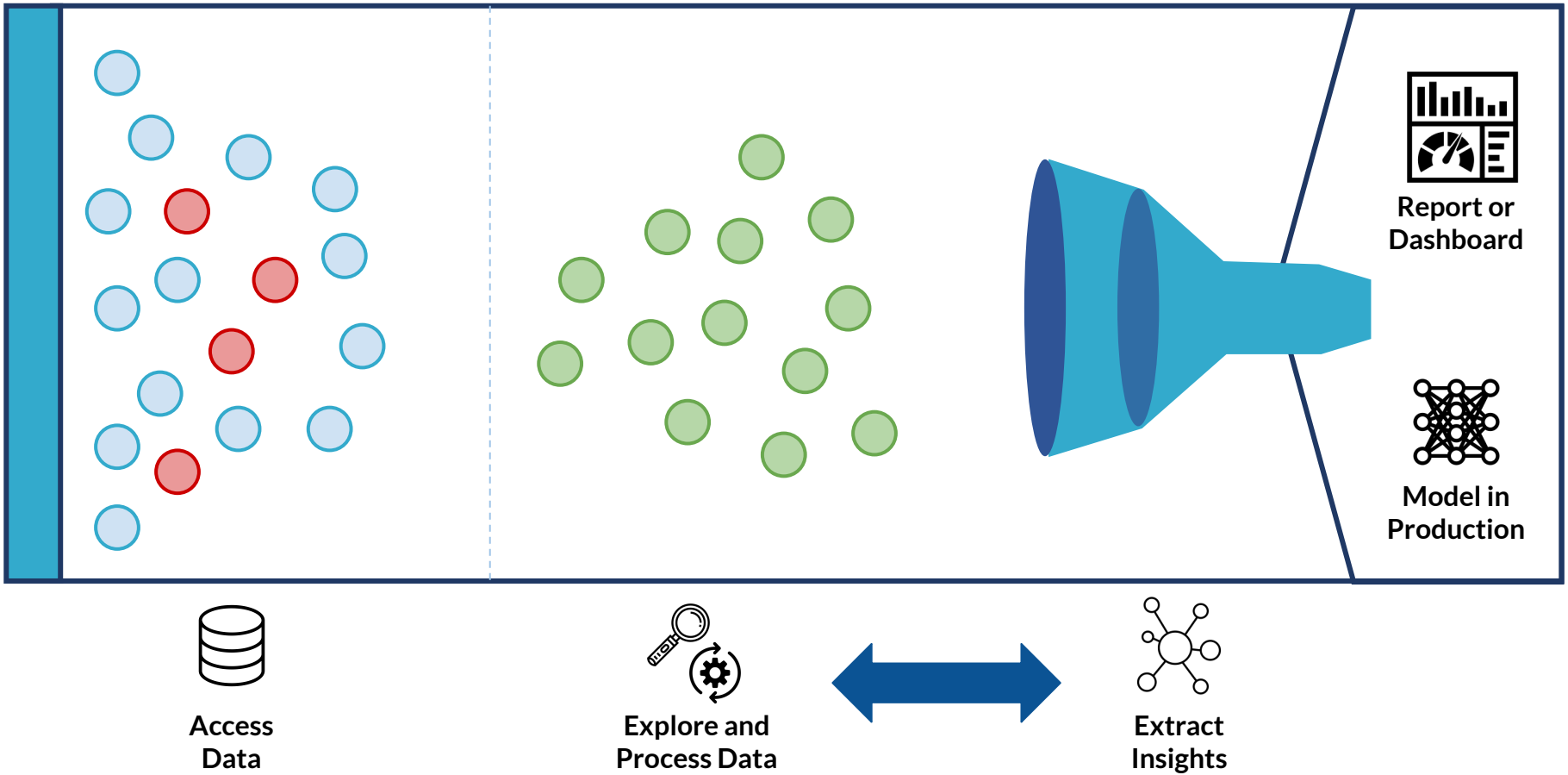
Explore and
Process Data

Extract
Insights

The data science workflow - revisited



The data science workflow - revisited



Coming soon

Check out our upcoming webinars!

Live training: Data Viz with ggplot2

Bring your data to life with the power of data viz with ggplot2, and answer real-world data questions. This session will run for three hours, allowing you time to really immerse yourself in the subject, and includes short breaks and opportunities to ask the expert questions throughout the training.

Wednesday, April 15, 2020, 11 AM EST, 4 PM BST

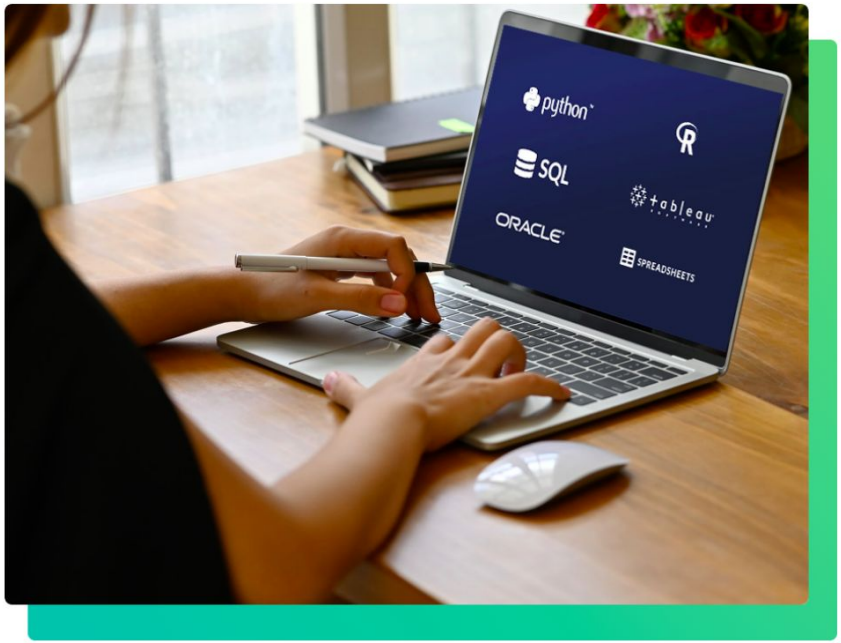
***Register** here*

DCVirtual: Webinar week

Join DataCamp and industry thought leaders for a week-long virtual conference on all things data science. Learn how to roll out a data strategy that includes online training, how to measure the success of your data science initiatives, and hear from experts about how AI and machine learning are impacting industries like finance and healthcare.

Monday, April 20 to Friday, April 24, 2020, 11 AM EST

Registration Link Coming Soon



DataCamp for Enterprise: What's New in Q2 2020

DataCamp for Enterprise: What's New in Q2 2020

Discover what's new in Q2 2020 for DataCamp Enterprise plans.

Wednesday, April 29, 2020, 11 AM EST, 4 PM BST



Register here

Take home question

Pick one of the following:

1) *What is the average price of listings by borough? Visualize your results with a bar plot!*
2) *What is the average availability in days of listings by borough? Visualize your results with a bar plot!*
3) *What is the median price per room type in each borough? Visualize your results with a bar plot!*
4) *Visualize the number of listings over time.*

Functions that should/could be used:

- `.groupby()` and `.agg()`
- `sns.barplot(x = , y = , hue = , data = )`
- `sns.lineplot(x = , y = , data = )`
- `.dt.strftime()` for extracting specific dates from a `datetime` column

Bonus points *if* you finish more than one question

Submission details:

- Share with us a code snippet with your output on LinkedIn, Twitter or Facebook
- Tag us on `@DataCamp` with the hashtag `#datacamplive`

Recap of the functions used

Diagnosis functions	Description
import pandas as pd	Imports the pandas package with the alias pd
.head()	Prints the header of a DataFrame
.dtypes	Gets the data types of each column in a DataFrame
.info()	Returns a # observations, data types and missing values per column
.describe()	Returns statistical distribution of numeric value in a DataFrame
.isna().sum()	Returns # of missing values per column
sns.distplot()	Plots distribution of one variable
msno.matrix()	Visualizes missingness matrix
msno.barplot()	Visualizes missingness barplot
.duplicated(subset = , keep =)	Lets you find duplicates in a DataFrame based on all or subset of columns

Treatment functions	Description
.str.replace("", "")	Replaces one string with another for each row of a str column
.str.split("", expand = True)	Splits a string column into two based on input
.astype()	Converts a column to a datatype of choice
pd.to_datetime()	Converts a date column to datetime
.str.lower()	Lowercases each row in a str column
.str.strip("")	Removes a pattern from each row of an str column
.replace()	Replace values for others in a column
.fillna()	Fills missing values of a column with a value of your choice
.drop_duplicates()	Drops duplicates